Abrasives

Jean-Claude Menard, Federation of European Producers of Abrasives, 75041 Paris, France

Newman W. Thibault, Norton Co., Worcester, Massachusetts 01606, United States

1. Introduction 1
2. Natural Abrasives 1
2.1. Quartz 1
2.2. Garnet 2
2.3. Corundum and Emery 2
2.4. Diamond 2
3. Manufactured Abrasives 3
3.1. Fused Aluminum Oxides 3
3.1.1. Raw Materials 3
3.1.2. Furnace Designs 3
3.1.3. Fused Alumina Types 3
3.2. Fused Zirconia – Aluminas 5
3.3. Sintered Aluminas 5
3.4. Silicon Carbide 6
3.5. Boron Carbide 6
3.6. Diamond 7
3.7. Cubic Boron Nitride 7
4. Manufacture and Testing of Sized Grains 8
5. Physical and Chemical Properties . 11
5.1. Hardness 11
5.2. Grain Strength or Fracture Toughness 11
5.3. Resistance to Attrition or Plastic Flow 12
6. Loose-Grain Applications 12
7. Bonded Abrasive Products 13
7.1. Vitrified Bonds 13
7.2. Organic Bonds 14
7.3. Metal Bonds 15
8. Coated Abrasive Products 15
8.1. Components 15
8.2. Production 16
8.3. Applications 16
9. International Quality Specifications 16
10. Safety 16
11. References 18

1. Introduction

An abrasive has been defined as "any of a wide variety of natural or manufactured substances used to grind, wear down, rub away, smooth, scour, clean or polish, often combined with a binder to make grinding wheels or affixed with glue to the surface of paper or cloth" [1].

Although this broad definition could include such products as plastic pads, metal wool, pumice, and various tools employing large single diamonds, as in bits for oil-well drilling, this article is limited to granular abrasives as hard as, or harder than, quartz (crystalline silicon dioxide) with grit sizes from about 5 mm average diameter down to 1 µm or smaller. The abrasives can be used in their loose, bonded (grinding wheel), or coated (sandpaper) forms.

Abrasives have applications that are so extensive and varied that only the major ones are mentioned in this article. For further information see [2–12]. In addition major manufacturers of abrasives, bonded and coated products, and related trade organizations have their own brochures, reprints of published articles, etc., available on almost any aspect of the application of their products. Most of them, as well as some universities and miscellaneous organizations, also have Internet sites.

2. Natural Abrasives

2.1. Quartz

Quartz [*14808-60-7*] is found worldwide. Major applications include the abrasive for common flint sandpaper used industrially for the finishing of nonmetallic substances, such as leather and felt, and also for miscellaneous household applications. Other uses are in sandblasting, in lapping of soft materials, as the cutting medium in wire sawing of soft stone, such as marble, and in scouring compounds. The use of this abrasive has declined greatly because inhalation of free silica dust may lead to silicosis. Quartz has been replaced to a considerable extent by manufactured abrasives, such as fused aluminum oxide and silicon carbide, as well as by garnet, glass, and slags.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_001

2.2. Garnet

Mineralogically, the term garnet refers to a group of minerals with similar crystal structure but varying considerably in chemical composition. Most abrasive garnet is almandite [*1302-62-1*], an iron aluminum silicate, $Fe_3Al_2(SiO_4)_3$.

Almost all of the garnet used for high-performance coated abrasives is mined in the Adirondack area of New York State. The ore is crushed and beneficiated and the resulting garnet is separated into various grit sizes and heat treated. This particular garnet is unique, as the mineral fractures to produce grains having sharp, continuous edges that are optimum for use as a coated abrasive. Also its hardness is at the upper end of the range for garnets. The coated product is used for finish sanding of wood in furniture and woodworking plants.

Other U.S. operations in New York, Idaho, and Maine, and those elsewhere in the world, produce garnet of inferior quality for coated applications, but which is used extensively for sandblasting, metal lapping, and the fine grinding of glass.

Estimated 1996 world production in 10^3 t/a are [49]:

United States	68.2
India	15
Australia	40
China	15
Others	16.8
Total	155

Approximately 30 % of U.S. production was used for water-filtration applications.

2.3. Corundum and Emery

Corundum [*1302-74-5*] is a naturally occurring aluminum oxide mineral of high purity. Estimated production in 1990 in 10^3 t/a was [50]:

South America	0.25
Europe + CIS	10
Africa	20
Asia (India)	2
World total	32.25

The major use of corundum is for the loose-grain grinding and polishing of optical components.

Emery [*12415-34-8*] consists predominantly of corundum but also contains varying amounts of other minerals, particularly magnetite [*1309-38-2*] and hematite [*1317-60-8*]. High-quality emery has been produced on the Greek island of Naxos for at least 2000 years and also is mined in Turkey. Major uses include polishing metals with the coated products and incorporation of the abrasive in nonslip floors, pavements, and stair treads.

The term "emery" has also been applied to a rock mined near Peekskill in New York State. Also known as spinel emery, it contains little or no corundum. However, because of its content of spinel [*1302-67-6*], a mineral with hardness between quartz and corundum, this "emery" has utility in less demanding applications. Very little, however, is used for coated abrasives.

Estimated world emery production in 1990 was 20 000 t/a and has decreased sharply since 1983.

2.4. Diamond

Diamonds [*7782-40-3*] apparently were found first in India and Borneo over 2000 years ago, but the discovery of them in South Africa in 1866 began an era of great expansion that is still continuing (→ Carbon).

Diamonds may be classified as gem quality, high-grade industrial, and bort, the borderlines between grades overlapping and varying with demand. Bort contains inclusions, microcracks, and other flaws, giving rise to translucency or opacity, off-color, etc. The material therefore is used almost exclusively in grain and powder forms. Most natural diamonds are single crystals, but carbonado [*12414-51-6*], a type found in Brazil, and ballas, found in Brazil and Africa, are polycrystalline in structure. They are especially tough because of the absence of well-developed cleavage planes, and are particularly useful for heavy-duty applications, such as truing and dressing grinding wheels.

World production of natural industrial diamonds in 1996 in 10^6 carats, including bort, which accounts for over 85 %, was:

Congo (Kinshasa)	15
Russia	9.2
Botswana	5
Republic of South Africa	6
Australia	23.1
All others	3.3
Total	61.6

Australia is the worlds largest producer since the startup of the Argyle alluvial operation in Western Australia in January 1983.

Major uses of diamonds are discussed in Section 3.6.

3. Manufactured Abrasives

3.1. Fused Aluminum Oxides

Differences in the composition of natural corundum and emery gave rise to variability in grinding wheel manufacture and in end-use performance. Therefore attempts were made to produce more consistent synthetic corundum abrasives.

Although some early work was carried out in Europe, the first commercially successful product resulted from the experiments of C. B. Jacobs, C. M. Hall, and A. C. Higgins in the United States at the end of the 19th century and the beginning of the 20th [13]. This technology was the foundation for a variety of fused aluminum oxide [*1344-28-1*] abrasives (→ Aluminum Oxide).

3.1.1. Raw Materials

Bauxite [*1318-16-7*] is used directly for the production of some fused alumina types, and indirectly, through the Bayer process, for other varieties. Largely for economic reasons, bauxites used for the production of fused aluminum oxide abrasives are of relatively high Al_2O_3 content. Materials from various sources are usually blended, a typical specification being:

Al_2O_3* = 82 % minimum
SiO_2 = 8 % maximum
Fe_2O_3 = 8 % maximum
TiO_2 = 4 % maximum

* determined by subtracting total SiO_2, Fe_2O_3, and TiO_2 from 100 %.

Abrasive-grade bauxite consumption in the United States and Canada was quoted as 197×10^3 t in 1994 and 133×10^3 t in 1995, both on a dry basis. This is less than 2 % of the total consumption of bauxite, 90 % of which was used for the production of aluminum.

The alumina from the indirect Bayer process contains about 99 % Al_2O_3. The type commonly used for fused alumina production contains approximately 0.4 – 0.5 % Na_2O, the remainder being small amounts of SiO_2, Fe_2O_3, and water.

3.1.2. Furnace Designs

Two general types of furnace are employed, one resulting in a batch process and the other in a continuous one.

The first, generally known as the *Higgins furnace*, consists of an open-ended, slightly tapered cylinder of sheet steel, smaller diameter up, resting on a hearth protected by a carbon coating [14]. Fusion of the bauxite is accomplished using an electric arc. Water continuously flows down the outside of the shell, keeping it from melting during the process. After fusion is complete, the product is cooled for several days, resulting in a coarsely crystalline material. This is sorted and rough crushed to yield crude abrasive about 50 mm in size.

Continuous furnaces are patterned after the batch type except that they are generally much larger and arranged for tilting and tapping the molten products into suitable containers [15], [16]. For economic reasons most fused aluminas being produced in similar large-capacity casting furnaces where fusion may continue uninterruptedly for many days or even months.

3.1.3. Fused Alumina Types

The fused alumina produced in largest tonnage is known as *regular aluminum oxide*. The raw materials are calcined bauxite, carbon usually in the form of coke, and iron borings, a typical mass ratio being 80 : 15 : 5. Because the relative ease of reduction of the bauxite feed to metal in the fusion process with carbon is $Fe_2O_3 > SiO_2 > TiO_2 > Al_2O_3$, an abrasive of considerably higher Al_2O_3 content than that of the furnace feed can be attained. During fusion two immiscible liquids exist. The upper one becomes the abrasive on crystallization, and the lower one becomes a byproduct, ferrosilicon.

Regular aluminum oxide is brown and has a chemical composition of approximately 95 %

Al_2O_3, 1.5 % SiO_2, less than 0.5 % Fe_2O_3, and 3 % TiO_2. Because of very slow cooling in the Higgins furnace, the crystals of alumina are coarse, averaging 10 – 15 mm in diameter. If the fused material is cast, the upper liquid must be poured into very large molds holding at least 5 t of product to insure similar slow cooling and coarse crystallization.

A second brown type is *semifriable fused alumina*. It is purer than regular, containing approximately 97 % Al_2O_3, 0.5 % SiO_2, less than 0.5 % Fe_2O_3, and 2 % TiO_2. Both batch and continuous processes are used to produce this variety.

Because the regular and semifriable types are close in composition, the present trend is toward a single brown fused alumina of intermediate composition, produced in casting furnaces. The relative friabilities are adjusted by crushing and processing methods to give different grain shapes, the more equidimensional being the regular and the sharper, less blocky being the semifriable.

In general, the major use of fused alumina abrasives is for processing high-tensile-strength ferrous metals, such as carbon steels. The regular kind is employed in loose-grain form in blasting, lapping, and barrel finishing, in bonded form mostly in nonprecision operations, such as cutting off and rough grinding in organic-bonded products. Semifriable alumina is more commonly used in vitrified-bonded products for precision grinding operations, and on coated belts and discs.

A third type, similar in composition and fusion method to the above, is known as *microcrystalline*. In order to produce an abrasive with very small alumina crystals, rapid cooling of the melt is required. This is effected by pouring it into small molds or onto pans to produce slabs. The resulting abrasive has higher intrinsic strength than the others. It formerly was used extensively in heat-treated form for very heavy duty grinding wheel applications, such as the conditioning of steel slabs and billets using resin-bonded products. For the most part this abrasive has been superseded by fused zirconia – alumina of the ZS type (Section 3.2), or by sintered alumina of the 76 Alundum type (Section 3.3).

White fused alumina is produced by direct fusion of Bayer process alumina [14]. Either the batch or the continuous casting process may be employed. Because the raw material is of high purity, only fusion and crystallization take place. The resulting abrasive has about the same chemical composition as the alumina. However, a small amount of the soda is volatilized, giving rise to pores in the resulting product. Sometimes small amounts of chromium oxide [*1308-38-9*] are added to the furnace feed, producing a pink or ruby-colored abrasive. Similarly, a small addition of vanadium oxide [*1314-34-7*] results in an emerald green color. Although claims have been made for the value of such additions, controlled grinding tests have failed to detect positively any advantages. The major use for the white abrasive is in vitrified-bonded wheels for toolroom grinding or precision grinding of heat-sensitive steels.

Monocrystalline aluminum oxide is a very high purity abrasive, produced directly from bauxite in a single-stage fusion [17]. The preferred method employs a furnace feed consisting of bauxite, pyrite (FeS_2) [*1309-36-0*] or sulfur [*7704-34-9*], carbon, and iron borings. When subjected to fusion in the Higgins batch furnace, two immiscible liquids are formed as with regular aluminum oxide. However, in the present case very slow cooling of the upper liquid results in essentially pure, individual crystals of Al_2O_3 in a matrix of sulfides. After the pig has cooled and been crushed, the matrix is removed chemically and mechanically. This treatment releases alumina crystals in the range of sizes required by the industry.

Little or no crushing is needed, so in this respect this abrasive is entirely different from all the other types of fused aluminas in use today. Chemical analysis approximates 99.2 % Al_2O_3, 0.6 % TiO_2, 0.2 % impurities. This abrasive also may be produced in a casting furnace provided that the fusion is poured into a large receptable capable of holding many tons of product to permit slow cooling of the melt. The size distribution of the resulting abrasive crystals can be varied by changes in manufacturing parameters. Surprisingly the monocrystalline type also may be produced from a mixture of Bayer process alumina, elemental sulfur, and carbon [18]. In this process only one liquid is present in the furnace during fusion, but otherwise the sequence of operations is similar, and the resulting product is the same as that produced directly from bauxite. This premium-priced abrasive is found to be more cost efficient than others, such as the

semifriable type, in many precision operations using vitrified-bonded wheels.

Production. World production capacity for fused aluminum oxide is listed in Table 1.

Table 1. World production capacity (10^3 t/a) for fused aluminum oxide in 1996 and 1997 (estimated) [49]

	1996	1997
USA and Canada	220	220
Australia	75	50
Austria	60	60
Brazil	100	100
China	450	500
France	45	45
Germany	150	150
India	20	20
Japan	55	55
Others	125	100
World total (rounded)	1300	1300

Trade Names: Alundum, Aloxite, etc.

3.2. Fused Zirconia – Aluminas

The most outstanding commercial development in the area of alumina-containing abrasives in the last two decades was the invention of fused zirconia – alumina products. For the applications in which they are most useful, they result in substantially greater efficiencies (lower total grinding costs) than the fused aluminas previously employed. United States and Canadian production for 1980 was 17.2×10^3 t, amounting to 11 % of all fused aluminas, including the zirconia – aluminas. No production data were available for 1981 or 1983; however, 1982 production was quoted as 7.3×10^3 t or 6 % on the same basis as the 11 % for 1980. This reduction was caused by the severe depression in those segments of the metal-working industry in which zirconia – alumina abrasives are most applicable.

These abrasives also are produced in Japan, France, Austria, Federal Republic of Germany, and possibly other countries. However, production data are not available for them.

Zirconia Content 25 wt %. Although the basic patent specified an abrasive alloy consisting essentially of zirconium oxide [*1314-23-4*] and aluminum oxide, with the content of the former between 10 and 60 wt %, the resulting commercial product typically contains about 25 % [19]. The alumina may be derived from bauxite or from the Bayer process, and the zirconia from the mineral zircon (a zirconium silicate) [*14940-68-2*], baddeleyite (a naturally occurring mineral high in ZrO_2) [*12036-23-6*], or fused zirconium oxide. Zirconia and alumina are fused by electric-arc casting furnaces to give primary crystals of alumina with diameters no greater than about 300 µm but typically less than 30 µm in a matrix of a eutectic of alumina and zirconia. The resulting abrasive is dense and very tough. The grains dull slowly during heavy-duty grinding operations, as with resin-bonded wheels on floorstand machines in foundries. They also have a longer life on high-speed, high-pressure grinders used for conditioning billets and slabs, particularly those made of alloy steels or of stainless steel, where better surface finish is required than produced by the sintered alumina of the 76 Alundum type (Chap. 3.3.).

Trade Names: ZF-ZS Alundum, BX, BZ, etc.

Zirconia Content 40 wt %. The raw materials and fusion methods are similar to those discussed above; however, in this case the melt must be solidified very rapidly in order to produce extremely small crystals [20]. In addition to a relatively slow dulling rate, these grains tend to fracture to a greater extent than the 25 % analogue and do so in such a manner as to provide new cutting edges rather than being shed from the grinding wheel or coated abrasive product. Major applications include resin-bonded portable wheels, cones and plugs, cloth-backed resin-bonded belts, and fiber-backed discs for grinding castings of steel, gray and ductile iron, and stainless steel alloys. In some cases mixtures of this abrasive with the less expensive fused aluminas or with silicon carbide are used for specific purposes; for example, the latter blend can efficiently clean and grind castings having burned-on sand.

Trade Names: NZ Alundum, etc.

3.3. Sintered Aluminas

From Bauxite. Calcined bauxites are ground to a fine powder, mixed with an organic binder, such as grease, extruded through an orifice, cut, and fired statically or in a rotary furnace to temperatures between 1400 and 1600 °C [21]. This

results in limited recrystallization of the alumina, increased density, and great strength of the particles, which have the desired grain size without being crushed. The major use of this abrasive is the conditioning of stainless steel billets and slabs using hot-pressed resin-bonded wheels.

Trade Names: 76 Alundum, SO 200, etc.

From Alumina Gels. A colloidal dispersion of alumina monohydrate [*12252-67-4*] and of modifying oxides, such as zirconia and/or magnesia [*1309-48-4*], is formed, and the resulting gel dried to produce chunks. These are crushed and air fired (sintered) to above 1250 °C, resulting in considerable shrinkage and formation of a dense, sharp, and extremely tough abrasive [22], [23]. These expensive abrasives are now widely used in industry as they significantly outperform fused alumina grains for high-precision grinding and creep-feed grinding. They are mostly used in vitrified wheels, but specific operations requiring organic-bonded wheels, such as flute grinding, surfacing, and calibrating with disk grinders, also use sol – gel aluminas. Coated abrasives are also produced with these grains for foundry snagging with belts and fiber disks and for other work on stainless steel and high-resistance alloys. As new types of sol – gel aluminas are developed and production volume grows, they are progressively replacing fused aluminas in many operations.

Trade Names: Cubitron, SG, TG, etc. Often referred to with the generic name "ceramic abrasives".

3.4. Silicon Carbide

The discovery of silicon carbide [*409-21-2*] by Edward G. Acheson in 1891 was indeed epoch-making because it was the first synthetic abrasive invented and commercialized, and signaled the gradual decline in the use of natural abrasives (→ Silicon Carbide). In an attempt to make diamonds by an electric-arc heating process, Acheson produced a very few hard crystals. Knowing that they were not diamond but thinking that they were a combination of carbon and corundum, he called them "carborundum," a term that remained even after the crystals were determined to be silicon carbide. Carborundum became the name of the company he organized to exploit the discovery. Abrasive SiC is produced today in furnaces similar to the one he patented in 1893 [24]. Silica sand plus carbon in the form of coke, or low-ash coal in the approximate mass ratio of 60 : 40 is heated in a troughlike electric resistance furnace to produce SiC according to the equation:

$$SiO_2 + 3\,C \longrightarrow SiC + 2\,CO$$

Several grades of this abrasive are produced. The most common black or gray type is used in bonded and coated form for grinding low-tensile, nonferrous metals such as aluminum, brass, copper; some cast irons; and nonmetallics, such as glass, stone, concrete, ceramics, and refractories. A higher purity green variety is employed for those operations, such as rough grinding of cemented carbide tools with vitrified-bonded wheels, in which this abrasive's more friable characteristic and the light green color are helpful. A type lower in purity than that used for bonded and coated products has utility in loose-grain applications, such as wiresawing and lapping.

World production capacity for silicon carbide abrasives is listed in Table 2.

Table 2. World production capacity (10^3 t/a) for silicon carbide abrasives in 1996 and 1997 (estimated) [49]

	1996	1997
USA and Canada	90	90
Argentina	5	5
Brazil	43	43
China	450	450
France	16	16
Germany	36	36
India	5	5
Japan	90	90
Mexico	60	60
Norway	80	80
Venezuela	40	40
Others	185	185
World total (rounded)	1100	1100

Trade Names: Carborundum, Crystolon, etc.

3.5. Boron Carbide

Abrasive boron carbide, B_4C [*12069-32-8*], is produced in electric resistance furnaces from a charge of high-purity boron oxide glass and high-purity coke according to the equation [25]:

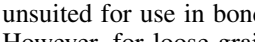

$$2\,B_2O_3 + 7\,C \longrightarrow B_4C + 6\,CO$$

As explained in Section 5.3, this abrasive is unsuited for use in bonded or coated products. However, for loose-grain applications, such as the lapping of cemented carbides and other hard materials, it is used alone or mixed with silicon carbide.

Trade Names: Norbide, etc.

3.6. Diamond

Commercial availability of manufactured diamond [*7782-40-3*] is the most important abrasive development in the 20th century (→ Carbon). Not only did it eliminate occasional shortages of natural diamond grain, but also it made possible a great variety of diamonds with vastly different characteristics, thereby increasing diamond efficiency and consequent use.

Commercial manufacture of diamond requires high pressure (5.0 – 6.5 GPa) and high temperature (> 1400 °C) [26]. Depending upon manufacturing parameters, a wide range of properties may be obtained, from very friable composite structures of small diamond crystals to stronger single crystals containing various defects and inclusions, to very strong, almost flawless crystals, equidimensional in shape, with well-developed crystal faces. One leading producer offers no fewer than seven different basic types. Over the years the major manufacturers have gradually been able to produce economically high-quality single diamonds of larger and larger size, now up to 20 – 25 mesh (approximately 700 – 850 µm in diameter). No doubt they will continue to extend availability further in the coarse direction.

The two major producers further improved the efficiency of diamonds for use in resin-bonded products by introducing metal-clad varieties in 1966 [27]. These coatings are believed to improve wheel efficiency by controlling abrasive breakdown, and by acting as a heat sink between the abrasive and the bond, retarding deterioration of the latter from heat generated during grinding. The most commonly employed coatings are nickel-based (30, 55, or 60 wt %) or copper (50 wt %). Although used mostly with manufactured diamonds, the 55 % nickel coating is also available on natural diamonds.

World production of synthetic diamond is listed in Table 3.

In general, diamonds, both natural and manufactured, are used for lapping carbides and other hard materials, and for grinding, drilling, and sawing cemented carbides and a wide variety of nonmetallics, such as plastics, glass, stone, concrete, ceramics, refractories, and such electronic materials as silicon and quartz. Silicon carbide also can be used for these same materials. However, bonded diamond products, in spite of their higher cost, often are more efficient, resulting in lower total costs per unit of work performed. This is particularly true in cemented carbide grinding, where diamond, now nearly always one of the manufactured types, is used almost exclusively except for offhand roughing with vitrified-bonded green silicon carbide wheels. Because of the affinity between diamond (carbon) and iron, diamonds are not economical for grinding ferrous metals except those containing relatively large amounts of hard constituents, such as certain vanadium steels.

The cost of synthetic diamond has greatly decreased since the mid-1980s, and this has allowed it to be used in nonprecision applications such as cutting masonry materials, stones and bricks with steel disks on hand-held grinders. The disks have laser-welded, metal-bonded diamond segments, and their low cost and long life have allowed them to replace most conventional silicon carbide cutting-off disks.

3.7. Cubic Boron Nitride

The second, and eventually possibly the most important development in the abrasive area this century, was the invention [28] of cubic boron nitride (CBN) [*10043-11-5*], first introduced commercially in 1969 (→ Boron Carbide, Boron Nitride, and Metal Borides).

This abrasive may be made in any of the high-pressure, high-temperature apparatuses used for diamond production. By using different raw materials and manufacturing conditions, a variety of products having a range of friabilities may be produced. Color ranges from almost colorless through shades of yellow, red, and black. As with diamonds, the producers supply CBN in sized grains and flours. Metal cladding of this abrasive also has improved its efficiency in resin-bonded

Table 3. Estimated world production of synthetic diamond (10^3 carat)

	1992	1993	1994	1995	1996
Belarus	30 000	30 000	25 000	25 000	25 000
China	15 000	15 000	15 000	15 000	15 000
Czech Republic	10 000 ∗	5 000	5 000	5 000	5 000
France	3 500	3 500	3 500	3 000	3 000
Greece	750	1 000	1 000	1 000	750
Ireland	60 000	65 000	65 000	60 000	60 000
Japan	30 000	32 000	32 000	32 000	32 000
Poland	320	98	271	256	250
Romania	3 000	5 000	5 000	5 000	5 000
Russia	80 000	80 000	80 000	80 000	80 000
Slovakia		5 000	5 000	5 000	5 000
South Africa	60 000	60 000	60 000	60 000	60 000
Sweden	25 000	25 000	25 000	25 000	25 000
Ukraine	10 000	10 000	8 000	8 000	8 000
USA	90 000	103 000	104 000	115 000	114 000
Total	418 000	440 000	434 000	440 000	439 000

∗ Czechoslovakia

products [29]. A 60 % nickel-based coating is used commonly.

Initial successful applications of CBN were for high-performance grinding of difficult to grind ferrous metals where use of fused alumina abrasives results in high rates of wheel wear per unit of material removed from the workpiece and in rapid dulling of the abrasive grains. Therefore machine productivity is low and grinding costs high, as is the liability of metallurgical damage to the work. Because of its hardness (inferior only to diamond), strength, low coefficient of friction during operation, and thermal and chemical stability to well over 1000 °C, CBN wheels, although expensive, increase productivity. This is because of much lower wheel wear, improved workpiece integrity, and fewer rejections and reworks. Now, with the development of computer-controlled, high-production systems especially designed for use with this abrasive, its application is being extended to more common, less difficult to grind ferrous metals. This has resulted in increased productivity for such operations as camshaft grinding in the automobile industry.

In 1981 one producer introduced a CBN family (*trade names:* Borazon 550, 560, 570) with sized grain that is polycrystalline and extremely tough. In certain applications resinoid- and vitrified-bonded wheels containing this abrasive have shown lower wheel wear than similar products containing the usual monocrystalline CBN. However, the expense of the abrasive has been a problem in establishing lower overall grinding costs in many cases.

Production of CBN by region is summarized in Table 4.

Table 4. Production of CBN by region (10^6 carat) [51]

	1992	1993	1998
Pacific Rim	12.5	13.2	18.8
United States	10.9	13.1	37.7
Europe	9.9	10.9	19.5
Total	33.3	37.2	76.1

Trade Names: Borazon, Amber Boron Nitride, etc.

4. Manufacture and Testing of Sized Grains

As mentioned in Chapter 3, some abrasive types are produced directly to size for use in bonded or on coated products. Other abrasives, such as the fused aluminum oxides, must be crushed, sized, and otherwise treated before use.

Crushing is accomplished by a variety of means, depending on the shape and other characteristics of the desired product. Jaw crushers, hammer mills, roll mills (in which precrushed crude is passed through sets of alloy steel rolls), and ball or rod mills commonly are employed. If exceptionally equidimensional particles are desired, the grain may be mulled by the use of heavy steel rollers working the grain in a revolving pan.

To produce specific grit sizes, screening is used for the coarser ones (macrogrits; >50 µm

Table 5. Standards for sizing of abrasive grains

Type of abrasive	Screen sizes (macrogrits)	Subsieve sizes (microgrits)
All except diamond and CBN for :		
Grinding wheel and general industrial usage	ANSI B74.12−1992 [30]	ANSI B74.10−1977 (R1992)
	FEPA 42/93	FEPA 42/93
	ISO 8486, part 1 (1996)	ISO 8486, part 2 (1996)
Coated Applications	ANSI B74.18−1996	ANSI B74.18−1996
	FEPA 43/93	FEPA 43/93
	ISO 6344, parts 1 and 2 (1998)	ISO 6344, part 3 (1998)
Diamond, CBN, all uses	ANSI B74.16−1995	ANSI B74.20−1997
	FEPA 61/97	FEPA 60/77 (under revision)

in diameter). Finer sizes (micogrits) are separated by various elutriation methods. Centrifuging often is employed for the finest grains.

For checking the sizing of abrasive grits, the standards shown in Table 5 are used.

Conventional (i.e., non-diamond and non-CBN) abrasive grits are sized according to standards that assign each size a number and define the mean size and the proportion and dimensions of coarser and finer sizes that can be present. The standards usually differ for bonded and coated abrasives, as the optimum performance of the products requires different grain shapes and a different size distribution for each grit number. For example, FEPA (Federation of European Producers of Abrasive Products) has issued two standards: FEPA 42/93, which defines the F series for bonded abrasives, and FEPA 43/93, which defines the P series for coated abrasives. Similarly, ISO has issued ISO 8486 for bonded and ISO 6344 for coated abrasives. The grit numbers and the mean particle sizes in micrometers in the F series (aluminum oxide and silicon carbide grains for grinding wheels and other bonded abrasives are as follows:

Macrogrits	
F4	4890
F5	4125
F6	3460
F7	2900
F8	2460
F10	2085
F12	1765
F14	1470
F16	1230
F20	1040
F22	885
F24	745
F30	625
F36	525
F40	438
F46	370
F54	310
F60	260
F70	218
F80	185
F90	154
F100	129
F120	109
F150	82
F180	69
F200	58
Microgrits	
F230	55.7 ± 3
F240	47.5 ± 2
F280	39.9 ± 1.5
F320	32.8 ± 1.5
F360	26.7 ± 1.5
F400	21.4 ± 1
F500	17.1 ± 1
F600	13.7 ± 1
F800	11 ± 1
F1000	9.1 ± 0.8
F1200	7.6 ± 0.5

The grit numbers and the mean particle sizes in micrometers in the P series (aluminum oxide and silicon carbide grains for coated abrasives are as follows:

Macrogrits	
P12	1815
P16	1324
P20	1000
P24	764
P30	642
P36	538
P40	425
P50	336
P60	269
P80	201
P100	162
P120	125
P150	100
P180	82
P220	68
Microgrits	
P240	58.5 ± 2
P280	52.2 ± 2
P320	46.2 ± 1.5
P360	40.5 ± 1.5
P400	35 ± 1.5
P500	30.2 ± 1.5
P600	25.8 ± 1
P800	21.8 ± 1
P1000	18.3 ± 1
P1200	15.3 ± 1
P1500	12.6 ± 1
P2000	10.3 ± 0.8
P2500	8.4 ± 0.5

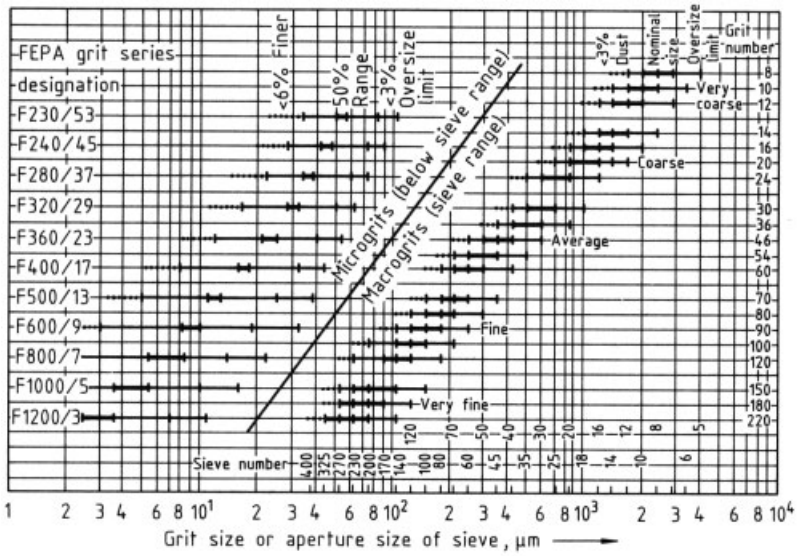

Figure 1. Example of size gradations of the FEPA standard F series for bonded abrasive grits

The approximate relationship between the FEPA and other national diamond and CBN grit size designations is summarized in Table 6.

Table 6. Approximate relationship between the FEPA and other national diamond and CBN grit size designations

FEPA grit designation ∗ (approx. mean size, µm)	ASTM 11	BS 1987
Narrow-range grades		
1181	16/18	14/16
1001	18/20	16/18
851	20/25	18/22
711	25/30	22/25
601	30/35	25/30
501	35/40	30/36
426	40/45	36/444
356	45/50	44/52
301	50/60	52/60
251	60/70	60/72
213	70/80	72/85
181	80/100	85/100
151	100/120	100/120
126	120/140	120/150
107	140/170	150/170
91	170/200	170/200
76	200/230	200/240
64	230/270	240/300
54	270/325	300/350
46	325/400	350/400
Wide-range grades		
1182	16/20	14/18
852	20/30	18/25
602	30/40	25/36
502	35/45	30/44
427	40/50	36/52
252	60/80	60/85

∗ The grit designation is prefixed with a "D" to denote diamond and a "B" to denote CBN.

Figure 1 shows the relationship between grit size designations, sieve numbers, aperture size of sieves, and micrometer dimensions of particles for both screen-size and subscreen-size grains according to FEPA standards [31].

In practice various manufacturers use other equipment for checking abrasive sizing, such as the Coulter counter.

The major suppliers of subsieve diamond and cubic boron nitride products offer their own series, many items of which are extremely well classified into the very narrow ranges demanded by the trade. One supplier, for example, offers 17 sizes from 40 – 80 µm down to 0 – 0.5 µm.

After sizing, grits are washed, processed further, and tested using the methods in Table 7.

Instrumental techniques such as atomic absorption, emission spectroscopy, and X-ray fluorescence are used in practice for chemical analyses. Eventually new standards based on such methods will be forthcoming.

Strength of abrasive grains containing glass as a minor impurity, as in regular fused alumina, can be increased significantly by heat treatment in air to 1250 – 1350 °C. The glass migrates to the surface, sealing microcracks, and thus repairing damage caused by the crushing operation. Alternatively, the grain may be coated with a glass frit that matures at 800 – 1000 °C. Such treatments increase abrasive efficiency in heavy-

Table 7. Test methods for abrasive grain properties

Application	Test method	Comments
Presence of magnetic particles	ANSI B74.19 – 1990 (R1995)	
Measure of coating weight for diamond and CBN grains	FEPA 62/93	
Relative strength of saw diamond grits	FEPA 63/93	
Degree of capillarity	ANSI B74.5 – 1964 (revised 1995)	capillarity is increased by heating to 500 – 600 °C
	ISO 9137 (1990)	used for conventional abrasives
	FEPA 44/93, part 3	used for conventional abrasives
Grain shape	ANSI B74.4 – 1992 (R1997)	equidimensional shapes pack to a higher bulk density than flat or slivery ones
	ANSI B74.17 – 1973 (revised 1993)	this method is used for testing diamond and CBN
Bulk density	ISO 9136, parts 1 and 2 (1989)	used for conventional abrasives
	FEPA 46/93, part 2	
Grain strength (friability)	ANSI B74.8 – 1987	ball mill test (see Section 5.2 for interpretation of strength data)
Chemical analysis (classical wet methods)	FEPA 46/93	used for fused aluminum oxides
	ISO 9285 (1995)	used for fused aluminum oxides
	ANSI B74.14 – 1992	used for fused aluminum oxides
	ISO 9286 (1995)	used for silicon carbides
	FEPA 45/93	used for silicon carbides
	ANSI B74.15 – 1992	used for silicon carbides
Sampling and splitting	ISO 9138 (1993)	used for conventional abrasives
	FEPA 44/93, part 1	used for conventional abrasives
Test-sieving machines	ISO 9284 (1992)	used for conventional abrasives

duty applications, such as in the conditioning of steel slabs and billets with resinoid wheels.

Sometimes grits are treated with a liquid silicone resin before being incorporated into resin-bonded products. Silicon carbide grain may be subjected to froth flotation to remove free carbon and/or treated with a sodium hydroxide solution to remove free silicon. In order to insure vitrified-bonded products free from iron spotting, fused alumina grains, especially the white variety, may be acid-treated to remove ferrosilicon or tramp steel not eliminated by magnetic separation.

5. Physical and Chemical Properties

5.1. Hardness

The hardness level required of an abrasive obviously depends upon the specific use. However, the vast majority of industrial applications necessitates abrasives at least as hard as quartz.

A common means for determining hardness is by use of the Knoop indenter, a gem-quality diamond carefully lapped to the shape of an elongated pyramid. This device indents the material to be tested under controlled-load conditions [32]. The applied gram load must be specified, e.g., K_{100}, because it can cause considerable variation in the resulting Knoop hardness number. Values of K_{100} for common abrasives compared with the Mohs' hardness scale generally used by mineralogists are:

	Knoop, K_{100}	Mohs'
Quartz (silica)	820	7
Spinel (magnesium aluminate)	1270	
Garnet (almandite from Adirondacks, U.S.)	1360	8
Fused zirconia-alumina (NZ Alundum type)	1600	
Fused alumina (white type)	2050	9
Silicon carbide	2480	
Boron carbide	2800	
Cubic boron nitride	4700	
Diamond	7000 – 8000	10

5.2. Grain Strength or Fracture Toughness

Strengths resulting from three different properties are important and usually act together during an abrasive operation. First is the grain shape, the blocky or equidimensional one being referred to as a strong shape, that with many flakes and slivers as a weak shape. Second, grains containing microcracks and other flaws produced by crushing tend to be weaker than those produced directly to size without subsequent comminution, assuming that the grains are otherwise of about the same overall shape. Third, is the intrinsic strength of the body of the abrasive itself.

Although this is difficult to measure in finished grain form because of the effect of shape and crushing, relative intrinsic strengths may be estimated. If the difference between abrasives is large, this is done by means of data derived from ball milling (method of ANSI B 74.8−1977) or by crushing or blasting tests, provided that the difference in grain shape has at least been minimized and crushing histories are similar.

The fracture strength or toughness of ceramics is determined by studying the cracking produced by indenting polished sections with a Vickers indenter. The load is such that fractures are produced at the corners of the indentations. The fracture toughness (K_c) may be calculated by measuring the average crack length and the diagonals of the indentations when Young's modulus and the microhardness are known [33]. This method of measuring intrinsic strength or toughness of abrasives is destined to replace the other procedures because it is a direct one.

5.3. Resistance to Attrition or Plastic Flow

When abrasive boron carbide was first produced in the early 1930s, its hardness was greater than that of any known material except diamond. Its effectiveness as an abrasive in loosegrain lapping operations, such as on cemented carbides, correlated well with relative hardnesses, being greatly superior to silicon carbide and fused aluminas, but inferior to diamond. For that reason boron carbide was excepted to have extensive applications in bonded products, replacing those older, softer abrasives. However, extensive grinding tests proved boron carbide to be completely unsuitable in any such operation because the wheels quickly became dull. The major reasons appear to be oxidation of the boron carbide, and the reaction or diffusion between the abrasive and the material being ground. A similar explanation appears to apply to the relatively poor performance of diamond when grinding common ferrous metals. Temperatures at the interface may approach the melting point of steel. Under these conditions both the abrasive and piece being worked are greatly softened, with reaction or diffusion between the two greatly accelerated. The result is attritious wear or plastic flow, leading to the development of polished flats on the abrasive grains and so the loss of much or all of their stock-removal capabilities [34–36].

For loose-grain lapping operations, hardness and strength are the most important abrasive properties. However, stock removal with fixed abrasives is a complicated process involving the interaction of hardness, strength, and chemical properties of both the abrasive and the workpiece. In addition the conditions of operation also must be considered.

6. Loose-Grain Applications

Blasting. In this operation, screen-size abrasives, such as quartz sand, garnet, fused alumina, and silicon carbide, impact the work material by means of compressed air, centrifugal force, or pressurized water. Blasting is used to descale or otherwise clean and deburr metal parts, clean buildings, and carve letters and designs on stone.

Wire Sawing. Endless, multistrand, twisted wire under tension is used to carry a slurry of water and abrasive for sawing blocks or slabs of stone such as limestone, marble, and granite. Quartz sand may be used to cut the softest stones, but fused alumina or silicon carbide is required for the harder granites and sandstones. This method is used both in the quarrying process itself and to shape and slice blocks removed from the quarry.

Barrel Finishing. Metal parts requiring cleaning, deburring, and/or refining of surface finish are tumbled in a slowly rotating barrel with water, acid or alkaline cleaning compounds, and suitable abrasives. Depending upon the work material and finishing requirements, the abrasives may be natural ones, such as emery or crushed granite, or, more commonly, manufactured abrasives, such as fused aluminas. The latter may be used as such, or they may be bonded with vitrified or organic materials into specific shapes, such as triangles, stars, or pins. A more efficient finishing method, involves vibrating the materials in bowls or tubs.

Lapping. Loose abrasives in a vehicle of water, soluble oil, kerosine, greases, etc., can be used to fine grind flat, cylindrical, or other surfaces. Laps of cast iron and other metals are used, and the abrasive grains may become embedded to some degree in these during the operation. The choice of abrasive depends upon the nature of

the workpiece. Harder abrasives, such as boron carbide and diamond, are most suitable for lapping cemented carbides, whereas quartz, garnet, and emery may be used on relatively soft materials. Gear lapping is accomplished by feeding an abrasive slurry between the parts as they revolve.

Buffing. Abrasives, such as fused aluminas and silicon carbide of subsieve size, are bonded with greases or waxes into cakes or sticks. These are applied dry to the face of rotating resilient wheels made of such materials as felt and other cloth types. The bond melts during the buffing operation so the abrasive is not fixed. A small amount of material is removed from the work, usually a metal; but, more importantly, a lustrous, satin, or mirror-like finish is produced.

7. Bonded Abrasive Products

These are rigid or only slightly flexible bodies, such as grinding wheels, which normally have more than a single layer of abrasive grains bonded with glass (vitrified bonds), organic materials, or metals. The products are manufactured in a wide variety of shapes and sizes, including wheels (mounted and unmounted), segments, bricks, sticks, etc. For details see ANSI B74.2–1974, Specifications for Shapes and Sizes of Grinding Wheels. . . . The specifications for diamond and CBN wheels are given in ANSI B74.3–1974 (revised 1980), and for marking of abrasive products in ANSI B 74.13–1977. Considering possible variations in types of abrasives, grit sizes, bond types, structure of the bodies (variations in volume percentages of abrasive, bond, and pores), and size and shape of the bodies themselves, it is understandable that major manufacturers produce several hundred thousand varieties of grinding wheels alone.

7.1. Vitrified Bonds

Abrasives bonded with glass include silicon carbide, diamond, cubic boron nitride, and all of the fused alumina types.

With fused alumina and silicon carbide, a typical manufacturing method involves coating the abrasive grain with premixed temporary and permanent bonding ingredients such as feldspar, clay, fluxes, or frits, molding the mix to a predetermined mass : volume ratio, drying it, altering the shape in the green state by shaving, if necessary, and finally maturing the bond by firing in a kiln, usually at 850 – 1250 °C or above. After it cools, a wheel may be sided and faced, and the hole reamed or bushed.

Products are characterized by volume percentages of abrasive, bond, and pores (Fig. 2). Softer grades contain higher pore volumes, and lower structure numbers have higher grain volumes. The final product is inspected by measuring mass : volume, modulus of elasticity, and resistance to penetration by a rotating chisel or by a blast of a known volume of sand or other abrasive under known pressure. The pores of the wheel may be filled with sulfur or waxes to improve grinding action by retarding loading of the wheel face with swarf during operation. A composite wheel can be made with a stronger center portion (finer grit abrasive, harder grade) to increase the overall strength of the body. Also the portion adjacent to the hole can be treated with a liquid epoxy resin for the same purpose.

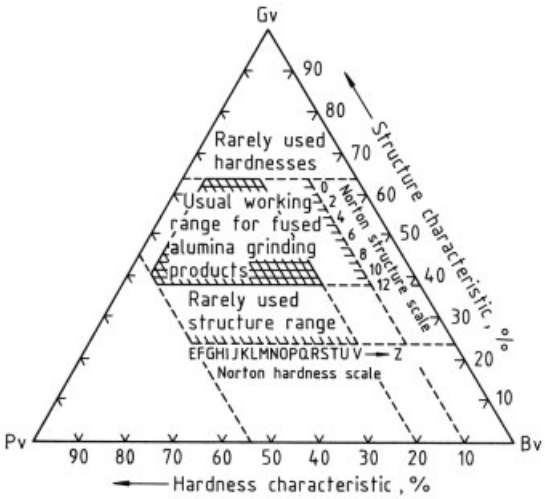

Figure 2. Relationship of hardness to structure in bonded abrasives
Bv) Bond volume; Gv) Grain volume; Pv) Pore volume

Because of the high cost of diamond and cubic boron nitride (CBN), only a small portion of the body contains these abrasives, the remainder being the core or "preform". A "green" prepressed ceramic body is first formed, placed in

a mold, and the abrasive-bond mix packed between preform and mold assembly. Bonds commonly employed are of the borosilicate type from frits or raw materials of the same general composition. After pressing and stripping from the mold, the product is fired to 900 – 1000 °C. Because diamond oxidizes at such temperatures, firing is usually in nitrogen; CBN may be fired in air.

Content of diamond and CBN in bonded products is based on volume percentages, 100 concentration being 25 vol % and others being proportional. For diamond-containing items, concentrations are almost always in the range of 25 – 100, whereas for CBN the range is 50 – 200. Average porosities are lower and the items more durable than those containing fused aluminas and silicon carbide.

For the most part, vitrified bonds are used for precision operations, such as surface, internal, and cylindrical grinding, where close dimensional tolerances are required.

7.2. Organic Bonds

Phenol – Formaldehyde Polymers. Most organic bonds are of this type. They have gradually replaced vitrified types in rough grinding applications because the wheels are stronger and more shock resistant and so can be operated safely at higher speeds with resulting greater efficiency. Diamond and all of the manufactured abrasives described in Chapter 3 are available in this bond type.

For abrasives other than diamond and CBN, a common cold-pressed manufacturing method involves wetting the abrasive with furfural [*98-01-1*] or liquid phenolic resin followed by coating it with a premixed blend of powdered phenol – formaldehyde resin and fillers or grinding aids, such as pyrite (FeS_2), cryolite [*15096-52-3*], or potassium tetrafluoroborate [*14075-53-7*]. Molding is similar to that for vitrified products. Curing is carried out in an oven at 140 – 200 °C, or dielectrically. Very low porosity products may be produced by hot pressing at 160 – 175 °C followed by oven curing, as with the cold-pressed products.

When a softer grade action is desired, the phenolic resins may be modified with epoxies, rubbers, or other thermoplastics. Some resin-bonded products, such as thin wheels for cutting-off operations and portable wheels for offhand grinding, are reinforced by molding sheets of woven fiberglass onto the sides, within the body, or both.

Resin-bonded alumina and silicon carbide products are used for precision-grinding, roll-grinding, centerless-grinding, and, most commonly, for rough-grinding operations where dimensional tolerances and finish are less critical, as in offhand grinding of rough castings and the conditioning of steel billets and slabs.

Most diamond and CBN products are hot pressed, and preforms are employed, as with the vitrified-bonded analogs. Uses are mentioned in Sections 3.6 and 3.7.

Rubber. Natural or synthetic rubbers or combinations are milled between rolls to break down the fibers, after which the abrasive grain, fillers, and sulfur for vulcanization are added. After being mixed, the batch is calendered to the required thickness, cut to shape, and heated to 150 – 175 °C to vulcanize the rubber. Depending on the amount of sulfur, type of rubber, and variety and amount of fillers, the product may range from soft and resilient to hard.

Because of strength and resiliency, rubber cutoff wheels, particularly thin ones, give accurate cuts with good surface finish and little burring in wet-grinding operations. Another application is the grinding of ball bearing races and centerless feed wheels.

Shellac. Shellac is a natural polymer prepared by heating and filtering the secretion of the lac insect, a parasite found on tress in India and surrounding countries. A common wheel-making process involves coating the abrasive with shellac and hot pressing the mixture in steel molds. The mix also may be calendered into thin sheets, from which wheels are cut and cured at 150 – 175 °C. Another method involves moistening the abrasive with a shellac solvent, adding powdered shellac, mixing, cold pressing, and postcuring. Shellac wheels exhibit a considerable degree of thermoplasticity, giving rise to a soft grinding action with a distinct polishing characteristic. They are used in some wet, light grinding operations, particularly for finishing steel rolls.

Polyimide Polymers. For limited, specialized applications, such as the grinding of flutes on carbide drills, and edge-grinding of carbide inserts on certain types of machines, diamonds in a polyimide bond have proved to be advantageous [37]. This particular polymer has considerably higher resistance to thermal degradation than the phenol – formaldehydes.

7.3. Metal Bonds

The abrasive most often bonded with metal is manufactured diamond, but the use of cubic boron nitride (CBN) in metal bonds is expected to increase very significantly.

Three types of metal-bonded products are made:

1) Those in which the abrasive zone is bonded directly to the core by a heating process;
2) Those in which segments or rims are produced and then attached to the core or steel blade afterward;
3) Those bonded by electroplating.

For the first type, a core or preform is placed in a mold, and the abrasive-metal mix added and then pressed. After being stripped from the mold, the body usually is sintered to maximum density, or it may be sintered to controlled porosity followed by infiltration with a liquid metal, such as a silver solder. Alternatively, the product may be hot pressed in a graphite mold. Bond compositions vary greatly. Commonly used ones include bronzes, various cobalt and nickel alloys, steels, and cemented carbides. Maturing temperatures vary from 500 °C for bronzes to 1200 °C for cemented carbides. Firing is in neutral or reducing atmospheres. In general softer bonds are used with hard, dense work materials, whereas harder ones are used to grind relatively soft but abrasive materials. For example, bronze bonds commonly are used on dense alumina, cemented carbides, and quartz crystals, whereas carbide bonds often are used to groove concrete highways and cut sandstones.

In the case of rims and segments, similar bonds, manufacturing methods, and firing temperatures are employed. In the production of diamond blades for use in the construction industry, segments have been attached to the steel center by brazing. However, laser welding, introduced in 1982, has permitted a much higher bonding strength, virtually eliminating loss of segments because of weakening of the joint from the heat generated during the cutting operations.

Electroplated products normally have a rigid core, a nickel bond, and a single layer of either diamond or CBN. Examples are: diamond-coated mounted points and discs used by dentists, relatively inexpensive diamond-coated wheels for offhand sharpening of carbide tools, and wheels of complicated shapes coated with diamond or CBN used for form grinding of workpieces, where great precision can be attained because of little or no tool wear.

8. Coated Abrasive Products

In coated abrasives a single layer of abrasive grains is bonded with an adhesive to flexible or semirigid backings.

8.1. Components

Abrasives include quartz, known in the trade as flint; garnet; emery, usually the Turkish variety; fused aluminas of the regular, semifriable, white, and pink types; sintered aluminas made from gels; fused zirconia – alumina of the NZ Alundum variety; silicon carbide, black and green types; and diamond and CBN, for very limited applications.

Backings include *Paper* of different weights: A for fine hand sanding with sheets, C for medium hand operations or machine finishing, E for heavy machine grinding.

Cloth: woven cotton or polyester, stitch-bonded polyester. They must be specially treated before being coated with abrasives on cloth-finishing lines to give them the appropriate mechanical characteristics. Different weights are used, depending on the operation: J for flexible products, X or Y for coarse grits or use on powerful machines.

Combinations of paper plus cloth: used when limited backing stretching and high resistance to tearing are necessary: wide belts for high power machines, drum sanders.

Vulcanized fiber made of several layers of cotton-based paper gelatinized with zinc chloride and vulcanized together. Different thicknesses are used, from 0.4 to 0.75 mm, depending upon the required stiffness and strength. Main use is fiber disks for portable machines for rough grinding and semi-finishing.

Polyester films for microfinishing and polishing of mechanical, electronic and optic components with very fine P series grits and micron or sub-micron sizes Abrasive grains are aluminium oxide, silicon carbide, diamond and special materials such as alumina flour and chromium, iron, and cerium oxides.

Adhesives used are high-quality hide glue; phenol – or urea – formaldehyde resins; and polyurethane or epoxy-based varnishes. These adhesives may contain mineral fillers to modify the physical properties of the bond, or to aid the operations in other ways (see Section 7.2).

8.2. Production

A strip of backing material up to 1.5 m or more in width is passed into a making machine where a thin film of bond, known as the maker coat, is applied. Sized abrasive is fed onto it by gravity, or an electrostatic field is used to orient the longer dimension of the grain perpendicular to the backing. The coating may be closed, in which the abrasive entirely covers the adhesive, or open, in which 50 – 75 % of the surface is covered. The item is dried at about 60 °C after which another layer of adhesive, known as the sizing coat, is applied to secure more firmly the abrasive grains. This second coat may be of the same composition as the first, or it may be different, e.g., resin used over glue. Then the bond is dried further and cured at about 150 °C. The coated abrasive is then coiled into a large roll known as a jumbo.

For most uses a further operation is required to improve product flexibility. This involves breaking the bond in a controlled manner. In single flexing, cracks are developed at right angles to the length of the strip. In double flexing, there are two series of cracks at about 45° to the length. Triple flexing is a combination of the other two types. Because of disturbance to the abrasive bond layer by such operations, only the minimum amount needed to satisfy end-use conditions is employed.

8.3. Applications

The jumbo rolls are cut and converted into many different shapes: narrow rolls, sheets, disks, endless belts 10 to 3000 mm wide and 250 to 7000 mm long, flap wheels and flap disks, spirally wound cones and pencils. These shapes are used by hand or on sanding machines in practically all industries for roughing, finishing, and polishing of wood, sheet and structural metal, mechanical parts, weldings, lacquers, glass, plastic, rubber, plaster, semiconductors, etc.

9. International Quality Specifications

Because of the mature nature of the abrasive industry and the use of various specifications, truly international standards have been slow in developing.

However, the International Organization for Standardization (ISO) has been active in establishing specifications relating to bonded and coated products, and, to a lesser extent, to the abrasive grain itself. This organization has issued ISO standards, some 21 of which relate to those items [38]. Eleven of them cover designations and dimensions of various coated products, nine of bonded items, including diamond and CBN, and one relates to grit sizes of diamond and CBN.

National standards exist in the United States (ANSI) [30], Japan, Japanese Industrial Standards (JIS) [42], and Europe (FEPA) [31].

10. Safety

Bonded abrasive products are not indestructible. Strengths vary with such factors as the type of bond, the grit size of the abrasive, and the structure and grade of the product required to efficiently perform a particular grinding operation. Therefore, the user must take special precautions to see that these products are properly handled,

mounted correctly on the machine, and not operated at excessive speed. The latter is the most frequent cause of wheel breakage, leading to damage of equipment and possible personal injury. Most wheels are marked for maximum operating speeds or instructions are packaged with the items.

"Safety Requirements for the Use, Care and Protection of Abrasive Wheels," ANSI B 7.1–1988, 106 p., is particularly useful because it includes sections on definitions of the various types of grinding operations, and on wheel shapes with their limitations, and handling and storage of abrasive products. General conditions of machines, safety guards, flanges, proper methods for mounting, general operating rules, and, most importantly, standard and special maximum operating speeds are covered also.

"Safety Requirements for the Construction, Care and Use of Grinding Machines," ANSI B11.9–1975, 71 p., is likewise helpful because it contains sections on design of machines, guards and flanges; on operating risks, and on the responsibilities of manufacturer, employer, and employee in connection with the care and use of equipment.

An EC standard concerning the safety of grinding machines is in preparation (EN 13218).

Japanese Industrial Standards (JIS), R6240–1972 and R6241–1972 [42], as well as the European Safety Code, FEPA standard 12/87, are similar to ANSI B7.1 but vary in details. Requirements in Germany (VBG 49) are more stringent than in most other countries because high-speed wheels must carry a certificate with certifying that they have been tested by the manufacturer in accordance with the specifics of ZH 1/670. Independent tests also are carried out to assure compliance with ZH 1/670.

EC standards concerning the safety of bonded, coated, and diamond/CBN abrasives are in preparation [52].

Another potential risk involves the effect of breathing dusts generated during abrasive operations, such as grinding and polishing or blasting, particularly when done in a dry operation.

"Ventilation Control of Grinding, Polishing and Buffing Operations," ANSI Z 43.1–1966, 21 p., is of particular interest because it covers exhaust hoods and enclosures, with minimum exhaust volumes specified for various operations and wheel dimensions. Drawings of suitable equipment are included. Similarly, "Ventilation and Safe Practices of Abrasive Blasting Operations," ANSI Z9.4–1979, 12 p., includes sections on dust risks, equipment, and operational procedures. It contains information on minimum air volumes for blast cleaning rooms occupied by blasters.

Most abrasive products are inert, producing dusts classified as inert or nuisance types when the work material also is inert. When the dust is not inert, not only must the amount to which the operator is exposed be known, but also its composition. Most of the dust generated from dry grinding and coated-abrasive operations is from the work material, with lesser amounts derived from the abrasive products themselves. Some abrasive products incorporate active fillers or grinding aids containing sulfur or fluorine compounds. Use of coated abrasives containing quartz (free silica) as the abrasive may require special controls. Only when both the amount and the composition of the airborne dust are known can a determination be made as to whether the requirements of the particular jurisdiction (city, country, etc.) are being met, and the health of the operator safeguarded.

For further details see "Fundamentals of Industrial Hygiene" [43], which relates specifically to requirements in the United States but which should be of interest worldwide. Particularly valuable are the appendices and their revisions [44–47]. For requirements in Europe (MAK), see [48].

With respect to disposal of wastes from abrasive products, the major concern is related to the chemical composition of the swarf because of its finely divided nature and consequently large surface area, which accelerates solubility and chemical reactions. Because most of the grinding debris originates from the work material, primary interest resides in its composition, solubility, and toxicity, with secondary emphasis on those properties of the bonded or coated abrasive used in the operation. Knowledge of such factors is required to determine the method of disposal that will not have an adverse impact on the environment.

Disposal of organic-bonded wheels and coated abrasives is also becoming a problem as they release phenol when leached by rain water in open landfills.

11. References

1. P. B. Gove (ed.): *Webster's Third New International Dictionary,* Merriam-Webster, Inc., Springfield, Mass. 1981.
2. T. J. Drozda, C. Wick (eds.): *Tool and Manufacturing Engineers Handbook,* 4th ed., **vol. 1,** Machining, Chapter 11, Grinding, 130 pages, Society of Manufacturing Engineers, Dearborn, Mich.1983, see especially pages 11 – 1 to 11 – 15, 11 – 49 to 11 – 130.
3. *Coated Abrasives – Modern Tool of Industry,* 1st ed., Coated Abrasives Manufacturers' Institute, Cleveland, Ohio 1982, esp. p. 80 – 426.
4. R. Williams (ed.): *Machining Hard Materials,* 1st ed., Society of Manufacturing Engineers, Dearborn, Mich. 1982, p. 131 – 243.
5. R. L. McKee: *Machining with Abrasives,* Van Nostrand Reinhold Co., New York 1982, esp. 1 – 36, 123 – 304.
6. P. Daniel (ed.): *Advances in Ultrahard Materials Applications Technology,* **vol. 1,** DeBeers Industrial Diamond Division, Ascot, England 1982, esp. 16 – 71, 92 – 103.
7. *Ultrahard Materials in Industry, Grinding Metals with Abrasive Boron Nitride,* DeBeers Industrial Diamond Division, Ascot, England 1982, p. 1 – 63.
8. W. Burkart, K. Schmotz: *Grinding and Polishing Theory and Practice,* 1st ed. (English), Portcullis Press, Redhill, England 1981, esp. 47 – 239.
9. F. T. Farago: *Abrasive Methods Engineering,* Industrial Press, Inc., New York 1980, vol. 1, 366 p., vol. 2, 508 p.
10. F. Hughes: *Diamond Grinding of Metals,* 2nd ed., Industrial Diamond Information Bureau, Ascot, England 1978, esp. 39 – 290.
11. K. B. Lewis, W. F. Schleicher: *The Grinding Wheel,* 3rd ed., Grinding Wheel Institute, Cleveland, Ohio 1976, esp. 36 – 463.
12. P. Daniel (ed.): *Industrial Diamond Review,* DeBeers Industrial Diamond Division, Ascot, England. (Includes articles on the application of diamond and CBN as well as a section devoted to abstracts relating to various aspects of diamond, cubic boron nitride and other hard materials), published1940 – present.
13. V. L. Eardley-Wilmot: "Artificial Abrasives and Manufactured Abrasive Products and Their Uses," *Abrasives,* Canada Dept. Mines, no. 699, part 4, Ottawa 1929.
14. *Ind. Miner. (London)* **149** (Feb. 1980) 55 – 57.
15. Norton Co., US 2 426 643, 1947 (R. R. Ridgway).
16. Norton Co., US 2 579 885, 1951 (J. A. Upper).
17. Norton Co., US 2 003 867, 1935 (R. R. Ridgway).
18. Norton Co., US 3 216 794, 1965 (S. J. Roschuk).
19. Norton Co., US 3 181 939, 1965 (D. W. Marshall, S. J. Roschuk, N. W. Thibault).
20. Norton Co., US 3 891 408, 1975 (R. A. Rowse, G. R. Watson).
21. Norton Co., US 3 079 243, 1963 (H. F. G. Ueltz).
22. 3 M Co., US 4 314 827, 1982 (M. A. Leitheiser, H. G. Sowman).
23. Kennecott Corp., GB 2 099 012 A, 1982 (R. J. Seider, A. P. Gerk).
24. Carborundum Co., US 492 767, 1893 (E. G. Acheson).
25. Norton Co., US 1 897 214, 1933 (R. R. Ridgway).
26. General Electric Co., US 2 941 241 to 248 inclusive, and 250 to 252 incl., 1960 (various G. E. Co. personnel). US 2 947 608 to 611 incl., 1960 (H. T. Hall, 608; H. M. Strong, 609; Hall, Strong, 610; F. P. Bundy, 611).
27. Industrial Distributors (1946) Ltd., US 3 902 873, 1975 (F. H. Hughes). ASEA, US 3 904 391, 1975, US 3 957 461, 1976 (O. Lindstrom, E. Lundblad).
28. General Electric Co., US 2 947 617, 1960 (R. H. Wentorf, Jr.).
29. General Electric Co., US 3 645 706, 1972 (H. P. Bovenkerk, W. A. Berecki).
30. ANSI, standards available from American National Standards Institute, Inc., 1430 Broadway, New York, NY.
31. *FEPA, Fédération Européenne des Fabricants de Produits Abrasifs, Standards,*20, Avenue Reille, F-75014, Paris, France; Internet: http://www.fepa-abrasives.org.
32. N. W. Thibault, H. L. Nyquist, *Trans. Am. Soc. Met.* **38** (1947) 271 – 330.
33. A. G. Evans, E. A. Charles, *J. Am. Ceram. Soc.* **59** (1976) 371 – 372.
34. T. N. Loladze, G. V. Bokuchava, G. E. Davidova in J. H. Westbrook, H. Conrad (eds.): *The Science of Hardness Testing and Its Research Applications,* Am. Soc. for Metals, Metals Park, Ohio 1973, p. 251 – 257.
35. T. N. Loladze, G. V. Bokuchava, G. E. Davidova in J. H. Westbrook, H. Conrad (eds.): *The Science of Hardness Testing and Its Research Applications,* Am. Soc. for Metals, Metals Park, Ohio 1973, p. 495 – 502.

36. L. Coes, Jr.: *Abrasives,* Springer Verlag, New York-Wien 1971, p. 154 – 163.
37. E. I. du Pont, US 3 179 631, 1965 (A. L. Endrey).
38. "Tools", *ISO Standards Handbook 6*, 1st ed., ISO Central Secretariat, Case postale 56, CH-1211 Geneva 20, Switzerland 1980, p. 81 – 86, 121 – 133, 164 – 185, 289 – 292, 330 – 334, 400 – 405, 417, 465, 519 – 521, 589 – 591, 663, 683 – 703.
39. *Industry and Trade Summary: Abrasives,* U.S. Dept. Commerce, International Trade Commission, Washington, D.C.1995, 30 pages.
40. P. Harben, *Ind. Miner. (London)* **134** (Nov. 1978) 62.
41. T. Dickson, *Ind. Miner. (London)* **159** (Dec. 1980) 70.
42. *JIS Standards,* available from ANSI, American National Standards Institute, Inc., 1430 Broadway, New York, NY.
43. B. A. Plog (ed.): *Fundamentals of Industrial Hygiene,* 4th ed., National Safety Council, Itasca, Ill., 1995, esp. pp. 175 – 182, 456 – 459, 538 – 539, 574 – 576.
44. American Conference of Governmental Industrial Hygienists. (ACGIH) (ed.): *Threshold Limit Values for Chemical Substances 1998 – 99,* (TLV), Cincinnati, Ohio 1998, esp. pp. 15 – 86.
45. OSHA Safety & Health Standards (29 CFR 1910), OSHA 2206, U.S. Department of Labor Occupational Safety & Health Administration, Washington, D.C., July 1997, esp. Subpart G, Section 1910.94: Ventilation.
46. RTECS, US Department of Health and Human Services, Cincinnati, 1997.
47. OSHA hazard communication standard (29 CFR 1910–1200) and material safety data sheets available from chemical manufacturers.
48. Deutsche Forschungsgemeinschaft (ed.): *Maximum Concentrations at the Workplace and Biological Tolerance Values for Working Materials 1995 (MAK),* VCH Verlagsgesellschaft, Weinheim 1995.
49. US Geological Survey, Mineral Commodity Summaries, Jan. 1988.
50. Mineral Facts and Problems, US Bureau of Mines Bulletin, 675 (1985)
51. F. J. Kuzler: *Hard and Superhard Materials–World Markets, Applications, and Opportunities: 1993 –1998 Analysis* , World Information Technologies, Northport, NY 1993.
52. EN 12413: Safety of Bonded Abrasives (1998); EN 13236: Safety of Diamond/CBN Abrasives (in preparation); Safety of Coated Abrasives (in preparation).

ABS → Polystyrene and Styrene Copolymers

Acaricides

Franz Müller (formerly Novartis Crop Protection AG, Basel), Allschwil, Switzerland (Chaps. 2, 3, 4)
Hans Peter Streibert, Novartis Crop Protection AG, Basel, Switzerland (Chap. 1; Chaps. 2 and 3 in part)
Saleem Farooq, Novartis Crop Protection AG, Basel, Switzerland (Chap. 1; Chaps. 2 and 3 in part)

1. Phytophagous Mites and Their Control . . . 1
1.1. Possible Reasons for the Mite Problem 1
1.2. Mite Species of Economic Importance 1
1.3. Possibilities for Mite Control . . . 3
1.4. Mite Resistance . . . 3
2. Older Acaricides . . . 3
3. Insecticides with Acaricidal Activity . 4
3.1. Organophosphates and (Oxime)Carbamates . . . 4
3.2. Pyrethroids . . . 11
3.3. Formamidines . . . 13
3.4. Nitrophenyl Esters . . . 13
3.5. Sulfonic Acid Esters . . . 14
3.6. Halogenated Benzhydrol Derivatives . 14
3.7. Organometallic Compounds . . . 15
3.8. Other Compounds . . . 15
4. Toxicology . . . 20
4.1. Organophosphates and (Oxime)Carbamates . . . 20
4.2. Pyrethroids . . . 30
4.3. Formamidines . . . 31
4.4. Nitrophenyl Esters . . . 32
4.5. Sulfonic Acid Ester . . . 32
4.6. Halogenated Benzhydrol Derivatives . 33
4.7. Organometallic Compounds . . . 33
4.8. Other Compounds . . . 34
5. References . . . 38

1. Phytophagous Mites and Their Control

Mites belong to the phylum Arthropoda, the class Arachnida, and the order Acarina. They have developed an astonishing variety of feeding habits. Some mites are predators, while others feed on detritus in soil or water. The economically important species are parasitic on either animals (e.g., ticks, scab mites) or plants (e.g., spider mites, eriophyid mites). Phytophagous mites are found in all parts of the world on practically all crop plants and may cause considerable damage. Approximately 8 – 10 % (annually ca. $ 400 × 10^6) of the total insecticide-acaricide market is spent on the control of phytophagous mites.

1.1. Possible Reasons for the Mite Problem

In recent decades, attacks by mites on food crops and fibers have increased so drastically that in many situations mite control has become an important feature of crop protection. Reasons for this increase of mites are manifold and not yet completely understood. Intensified agricultural production methods, such as monocultures, extensive use of fertilizers, improved irrigation systems, and cultural practices such as pruning have improved the vigor and nutritional value of the plants. However, these conditions also are beneficial to the development of phytophagous mites and may result in higher rates of reproduction. Furthermore, the indiscriminate use of pesticides in the past may have reduced or even eliminated the populations of natural enemies of the mites (predatory mites and beneficial insects) thereby encouraging the development of certain phytophagous mite species. In addition, some insecticides (e.g., DDT, Carbaryl) used for the control of insect pests, may promote the reproduction of phytophagous mites [6–8].

1.2. Mite Species of Economic Importance

The economically important species of the phytophagous mites are as follows:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_017

Tetranychidae – Spider mites

Tetranychus urticae (Koch), Two-spotted mite, widespread on fruit and on grapes, vegetables, cotton, hops, roses, ornamentals, and greenhouse cultures.

Tetranychus cinnabarinus (Boisduval), carmine mite, widely spread in the warmer climatic regions on cotton, fruit cultures, ornamentals, and in greenhouses.

Tetranychus kanzawai (Kishida), tea red spider mite, found on tea in Asia.

Tetranychus mcdanieli (McGregor), McDaniel mite, found on fruit on the west coast of the United States, as are other, different *Tetranychus* species.

Panonychus ulmi (Koch), European red mite, widespread on fruit (apples, peaches, pears, prunes, plums) and on grapes.

Panonychus citri (McGregor), citrus red mite, found in citrus cultures worldwide.

Tenuipalpidae – False spider mites

Various *Brevipalpus* species, flat mites, found on citrus and different subtropical fruits, tea, greenhouse cultures.

Tarsonemidae – Soft-bodied mites

Hemitaronemus latus (Banks), citrus silver mite, found on cotton, citrus, attacks a wide range of agricultural crops and ornamental and indigenous plants in tropical and subtropical regions.

Eriophyidae – Gall mites

Phyllocoptruta oleivora (Ashmead), citrus rust mite, found on citrus.

Aceria sheldoni (Ewing), citrus bud mite, found on citrus.

Life cycles and behavioral and morphologic characteristics of mites vary greatly. The description of life history given here concentrates on the agriculturally important Tetranychidae family.

The *spider mites* are 0.3 – 0.5 mm long, eggshaped, eight-legged animals covered with hairlike appendages. These hairs (setae) are immobile and are sense organs. Spider mites live in dense colonies, mainly on the underside of the leaves. In such colonies, usually all the development stages are present: eggs, larvae, and nymphs, as well as adults of both sexes.

Many spider mite species produce protective webbing composed of almost invisible strands of silk; these webs produce a favorable microclimate on the leaf surface. Characteristic of spider mites is their enormous reproductive potential: a female of the European red mite (*Panonychus ulmi*) produces 14 – 30 eggs during its lifetime, the female two-spotted mite as many as 70 – 120. The tiny, pale green eggs are usually deposited on the underside of the leaves within the webbing [9–11].

The eggs hatch after 3 – 4 days in the case of *Tetranychus* species, whereas in the European red mites the first instar larva appears after 6 – 20 days. All the spider mites go through five stages in their life cycle: egg, six-legged larva, eight-legged protonymph, deutonymph, and adult. These stages are separated by resting stages followed by molts. The time for the whole development depends strongly upon the temperature, humidity, and the host plant. Under favorable conditions, the postembryonal development in Tetranychidae is completed within 8 – 14 days.

Orchard mites usually develop six to eight generations per season, depending on the areas where they occur. In greenhouses and in warm climates, the two-spotted mite may develop up to 30 generations per vegetation period, whereas for *Panonychus* species only 6 – 20 generations can be expected.

The climatic conditions during the summer months have a strong influence on the population build-up during the following spring. Most spider mites overwinter as bright orange adult females within bark crevices or in plant debris on the ground. The European red mite is an exception, overwintering as a winter egg on twigs.

Many weed species are excellent host plants from which the spider mites may move on to crop plants, either by crawling or by ballooning (transport by wind). All mites live on the sap of the plant, piercing the leaf tissue with the two sharp lances attached to the mouth and thus puncturing the cells in the leaf epidermis. The leaves then become speckled, later turn brown, and, if the attack is severe, may drop. As a consequence, photosynthesis and respiration are impaired and this may affect the number of flower buds in the following season. A mass attack of phytophagous mites on deciduous fruit may result in yield losses of up to 40 % [12].

The *gall mites* (Eriophyidae) are very tiny rodshaped mites with threadlike appendages. Gall mites are very host specific and have high reproductive potential. They destroy buds,

leaves, and fruit, and they also transmit viruses. They may cause serious economic damage, especially in citrus fruits.

After a heavy attack of citrus rust mites (*Phyllocoptruta oleivora*), the fruit turns reddish brown and shrinks because of the loss of water. The citrus bud mite (*Aceria sheldoni*) causes malformation of citrus fruits.

1.3. Possibilities for Mite Control

Chemical treatment can be aimed either at overwintering eggs (e.g., those of *Panonychus ulmi*) or at the mobile stages of the mites. Winter eggs often are treated with mineral oils. For the effective control of mites during the vegetative period, a number of chemicals are available. However, compounds that are effective against all development stages and have a long-term residual effect are preferred.

The compounds used should be well tolerated by the plants, and the effects on environment and beneficial insects should be minimal. The population of these natural enemies, which could control mite populations effectively, has been reduced severely or eliminated in most orchards, vineyards, and citrus groves as a result of the use of broadspectrum pesticides aimed at targets other than mites.

Biological control of phytophagous mites by releasing their natural enemies, such as predatory bugs or coccinellids, has been successful in greenhouses or in plastic tunnels. However, only limited success has been achieved under field conditions [13–15].

Integrated pest management, which makes use of a variety of possible techniques, including chemicals, to keep the pest population below the economic threshold, is far more promising than simple biological control. In integrated pest management systems, chemicals are used only when the population density reaches a certain level (economic threshold) and if possible, selective acaricides are chosen so as not to harm beneficial insects and predatory mites.

1.4. Mite Resistance

In many areas, mites have developed resistance to certain chemicals: their control therefore has become problematic. Because of the large number of generations per season, selection may occur very rapidly in a mite population. Resistance is induced by repeated application of the same or closely related chemicals. To compensate for this effect, higher dosages are needed, but this is not a long-term solution to the resistance problem. New compounds with new modes of action must continue to be developed to replace the older ones, that have decreased in effectiveness.

New control concepts, such as integrated pest management, have made it possible to reduce the number of acaricide treatments in many situations, and this has slowed development of resistance in mite populations.

2. Older Acaricides

Products introduced at the beginning of the pesticide era still find limited use as acaricides in various parts of the world.

Nitrophenols , in combination with mineral oil, were used to fight mites in orchards by eliminating their overwintering eggs. However, the phytotoxicity of these compounds has been a drawback [16, p. 527]. For a review of nitrophenols with acaricidal activity see [17, p. 2] and [18, p. 537].

Sulfur [*7704-34-9*], used mainly as a protective fungicide to control powdery mildews, is effective against the mobile stages of various mite species. It is used in combination with other fungicides and insecticides mainly in vineyards [17, p. 2].

Azobenzene [*103-33-3*] has been used as a fumigant in greenhouses against insects and mites, especially against eggs [17, p. 2].

Binaparcryl [*485-31-4*] has been used as a nonsystemic acaricide, mainly against all stages of spider mites and powdery mildew of apples, citrus fruits, cotton, etc. [19, p. 73].

Chlorfenson [*80-33-1*] has been used as a nonsystemic acaricide with long residual ovicidal activity. It is effective against mites of citrus and other fruits [19, p. 150].

Tetrasul [*2227-13-6*] has been used as a nonsystemic acaricide, particularly suitable for the control of various phytophagous mites which hibernate in winter egg form. It has been used on vegetables and fruits [19, p. 790].

Chlorpropylate [*5836-10-2*] has been used as a nonsystemic contact acaricide on cotton, fruits, and ornamentals [19, p. 169].

Aldoxycarb [*1646-88-4*] is a systemic insecticide and nematicide, and is a potent cholinesterase inhibitor. It has been used on tobacco and as a cotton seed dressing [19, p. 9].

Chlorobenzilate [*510-15-6*] is a nonsystemic acaricide with little insecticidal activity. It has been used against phytophagous mites on fruits and vegetables [19, p. 162].

3. Insecticides with Acaricidal Activity

3.1. Organophosphates and (Oxime)Carbamates

Organophosphates and carbamates, used as broad-spectrum insecticides but also exhibiting acaricidal action, were able to contain the mite problem for some time. However, development of mite resistance toward these compounds has made them less and less effective as acaricides.

Aldicarb [*116-06-3*], 2-methyl-2-(methylthio)propanal *O*-methylcarbamoyloxime, $C_7H_{14}N_2O_2S$, M_r 190.3, *mp* 98 – 100 °C, forms colorless crystals which are practically insoluble in heptane and mineral oils, moderately soluble in water, and soluble in most organic solvents [20, p. 26].

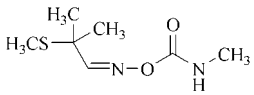

Aldicarb is produced by reaction of isobutene with nitrosyl chloride followed by reaction of the product with sodium methyl sulfide; the resulting oxime is then converted into the carbamate with methyl isocyanate [21].

Aldicarb (announced in 1965) is a cholinesterase inhibitor which is metabolically activated to sulfoxide. It is applied to soil to control chewing and sucking insects, spider mites, and nematodes in glasshouse and outdoor ornamentals, vegetables, sugar beet, peanuts, fruits, etc.

Trade Names. Temik (Rhône-Poulenc); Sanacarb (Sanachem).

Chlorfenvinphos [*470-90-6*], 2-chloro-1-(2,4-dichlorophenyl)vinyl diethyl phosphate, $C_{12}H_{14}Cl_3O_4P$, M_r 359.6, *mp* – 23 to – 29 °C, *bp* 167 – 170 °C, is a colorless liquid which is sparingly soluble in water, and miscible with most organic solvents [20, p. 211].

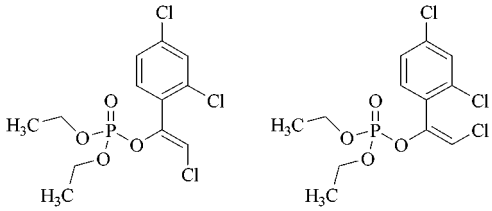

(*Z*) (*E*)

Chlorfenvinphos is produced by reaction of 2,4-dichloroacetophenone with triethyl phosphite [22].

Chlorfenvinphos (announced in 1952) is used for soil application to control root flies, root worms, and insects in vegetables and fruit flies in maize.

Trade Names. Birlane (Cyanamid); Apachlor (Rhône-Poulenc).

Chlorpyrifos-methyl [*5598-13-0*], *O,O*-dimethyl *O*-3,5,6-trichloro-2-pyridyl phosphorothioate, $C_7H_7C_{13}NO_3PS$, M_r 322.5, *mp* 45.5 – 46.5 °C, consists of colorless crystals which are sparingly soluble in water but soluble in acetone, benzene, chloroform, hexane, and methanol [20, p. 237].

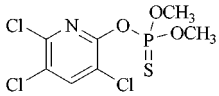

Chlorpyrifos-methyl is produced by reaction of 3,5,6-trichloro-2-pyridinol with dimethylthiophosphoryl chloride [23].

Chlorpyrifos-methyl is used to control aquatic larvae, flies, household pests, mosquitoes, pests in stored grain, and various foliar crop pests.

Trade Name. Reldan (DowElanco); Pyriban (Aimco).

Diazinon [*333-41-5*], *O,O*-diethyl *O*-2-isoproyl-6-methylpyrimidine-4-yl phosphorothioate, $C_{12}H_{21}N_2O_3PS$, M_r 304.3, *bp*

83 – 84 °C (26.6 mPa), is a clear colorless liquid which is sparingly soluble in water but completely miscible with common organic solvents [20, p. 354].

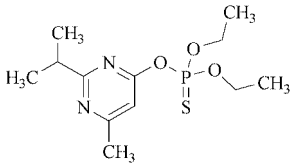

Diazinon is produced by condensation of isobutyramidine with acetoacetate to yield the intermediate 2-isopropyl-4-methylpyrimidine, which is transformed into the final product by treatment with diethylthiophosphoric acid [24].

Diazinon (announced in 1953) is a nonsystemic insecticide/acaricide with contact, stomach, and respiratory action. It is used to control sucking and chewing insects and mites on a variety of crops.

Trade Names. Basudin (Novartis); Dianon (Nippon Kayaku); Knox-out (Elf Atochem).

Dicrotophos [*141-66-2*], dimethyl (*E*)-2-dimethylcarbamoyl-1-methylvinyl phosphate, $C_8H_{16}NO_5P$, M_r 237.2, *bp* 460 °C, is a yellowish liquid which is completely miscible with water and common organic solvents [20, p. 382].

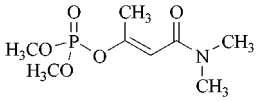

(*E*)

Dicrotophos is produced by reaction of trimethyl phosphite with 2-chloro-*N,N*-dimethyl-3-oxobutyramide [25].

Dicrotophos (announced in 1965) is used to control sucking, chewing, and boring insects and mites in cotton, coffee, rice, sugar cane, and tobacco.

Trade Names. Bidrin (Cyanamid); Dicron (Hui Kwang).

Dimethoate [*60-51-5*], *O,O*-dimethyl *S*-(*N*-methylcarbamoyl)methyl phosphorodithioate, $C_5H_{12}NO_3PS_2$, M_r 229.2, *mp* 51 – 52 °C, consists of colorless crystals which are moderately soluble in water, soluble in alcohols, benzene, chloroform, dichloromethane, ketones, and toluene [20, p. 550].

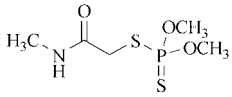

Dimethoate is produced by reaction of the sodium salt of *O,O*-dimethyldithiophosphoric acid with *N*-methylchloroacetamide [26].

Dimethoate (announced in 1948) is effective against houseflies and Diptera of medical importance.

Trade Names. Cygon, Roxion (Wilbur-Ellis); Perfekthion (BASF); Champ (Searle India); Danadim (Cheminova); Robgor (Ramcides).

Disulfoton [*298-04-4*], *O,O*-diethyl *S*-(2-ethylthio)ethyl phosphorodithioate, $C_8H_{19}O_2PS_3$, M_r 274.4, *mp* < −25, *bp* 128 °C, is a colorless oil which is sparingly soluble in water but readily miscible with common organic solvents [20, p. 438].

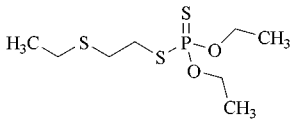

Disulfoton is produced by reaction of the sodium salt of diethyldithiophosphoric acid with ethylmercaptoethyl chloride [27].

Disulfoton (announced in 1952) is used to control aphids, thrips, mealybugs, other sucking insects, and spider mites in potatoes, vegetables, cereals and other crops.

Trade Names. Disyston (Bayer); Solvirex, Fremin AL (Novartis).

EPN [*2104-64-5*], *O*-ethyl *O*-4-nitrophenyl phenylphosphonothioate, $C_{14}H_{14}NO_4PS$, M_r 323.3, *mp* 34.5 °C, consists of yellow crystals which are practically insoluble in water but soluble in common organic solvents [20, p. 464].

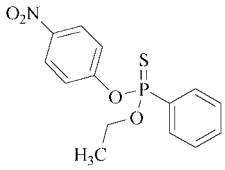

EPN is produced by reaction of benzene with phosphorus trichloride in the presence of aluminum chloride and treatment of the resulting

product with phosphenyl chloride to give the thiophosphonic acid, which is converted to ethyl phenyl thiophosphonate chloride by treatment with ethanol, followed by condensation of the product with the sodium salt of *p*-nitrophenol [28].

EPN (announced in 1948) is used against a broad range of Lepidoptera larvae, especially ballworms and *Alabama argillacea* in cotton, *Chilo* spp. in rice, and other leaf-eating larvae in fruit and vegetables.
Trade Name. EPN (Nisson).

Ethion [*563-12-2*], *O,O,O′,O′*-tetraethyl *S,S′*-methylene bis(phosphorodithioate), $C_9H_{22}O_4P_2S_4$, M_r 384.5, *bp* 164–165 °C (40 Pa), is a colorless to amber liquid which is practically insoluble in water but miscible with most common organic solvents [20, p. 480].

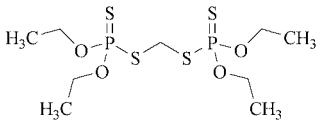

Ethion is produced by reaction of diethyldithiophosphoric acid with formaldehyde in the presence of sulfuric acid [29].

Ethion (announced in 1957) is used to control spider mites, aphids, scale insects, thrips, and lepidopterous larvae in fruits, vegetables, and turf.
Trade Names. Ethiol, Rhodocide (Rhône-Poulenc); Tafethion (Rallis); Dhanumix (Dhanuka).

Fenothiocarb [*62850-32-2*], *S*-4-phenoxybutyl dimethylthiocarbamate, $C_{13}H_{19}NO_2S$, M_r 253.4, *mp* 40–41 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in cyclohexanone, acetonitrile, acetone, xylene, methanol, and hexane [20, p. 517].

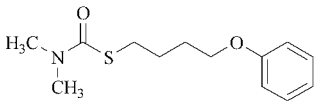

Fenothiocarb is produced by reaction of the sodium salt of *N,N*-dimethylcarbamothioate with phenoxybutylchloride, which is obtained by reaction of an excess of dichlorobutane with phenol in the presence of potassium hydroxide [30].

Fenothiocarb (announced in 1985) is a nonsystemic acaricide used to control eggs and young stages of *Panonychus citri*, *Panonychus ulmi*, and other *Panonychus* spp.
Trade Name. Panocan (Kumiai).

Formothion [*2540-82-1*], *O,O*-dimethyl *S*-[formyl(methyl)carbomylmethyl] phosphorodithioate, $C_6H_{12}NO_4PS_2$, M_r 257.3, *mp* 25–26 °C, is a pale yellow viscous liquid or crystalline mass which is moderately soluble in water but completely miscible with common organic solvents [20, p. 625].

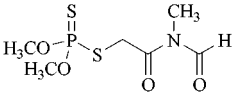

Formothion is produced by reaction of the sodium salt of *O,O*-dimethyldithiophosphoric acid with *N*-methyl *N*-formyl carbamoylmethyl chloride [31].

Formothion (announced in 1960) is used against a wide range of sucking and mining insects, such as Aphididae, bugs, Cicadellidae, Cocidae, as well as against some chewing insects and spider mites on a variety of field crops, fruit trees, citrus and other tropical fruit, cotton, ornamentals, rice, tobacco, and vegetables.
Trade Name. Anthio (Novartis).

Mecarbam [*2595-54-2*], *O,O*-diethyl *S*-(*N*-ethoxycarbonyl-*N*-methylcarbamoylmethyl) phosphorodithioate, $C_{10}H_{20}NO_5PS_2$, M_r 329.4, *bp* 144 °C (2.7 Pa), is a pale yellow to brown oil which is sparingly soluble in water, soluble in aliphatic hydrocarbons, and miscible with alcohols, esters, and ketones [20, p. 774].

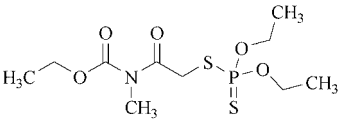

Mecarbam is produced by reaction of the sodium salt of *O,O*-diethyldithiophosphoric acid with *N*-ethoxycarbonyl-*N*-methylcarbamoylmethyl chloride [32].

Mecarbam (announced in 1961) possesses slight systemic properties, contact and stomach action, and long residual activity. It is used to control aphids, suckers, whitefly, scale insects, mealybugs and red spider mites on a variety of crops; and leaf hoppers, plant hoppers, and miners on rice.
Trade Name. Murfatox (Efthymiadis).

Methacrifos [*62610-77-9*], (*E*)-*O*-2-methoxycarbonylprop-1-enyl *O,O*-dimethyl phosphorothioate, $C_7H_{13}O_5PS$, M_r 240.2, *bp* 90 °C (1.3 Pa), is a colorless liquid which is sparingly soluble in water and miscible with many organic solvents [20, p. 806].

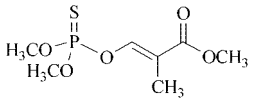

Methacrifos is produced by reaction of 2-hydroxymethylenepropionic acid methyl ester with dimethylthiophosphoryl chloride [33].

Methacrifos (announced in 1977) is an insecticide and acaricide with respiratory, contact, and stomach action. It is used for control of arthropod pests in stored products by incorporation or by surface treatment.
Trade Name. Damfin (Novartis).

Methamidophos [*10265-92-6*], *O,S*-dimethylphosporamidothioate, $C_2H_8NO_2PS$, M_r 141.1, *mp* 44.9 °C, consists of colorless crystals which are highly soluble in water, isopropanol, and dichloromethane, and moderately soluble in hexane and toluene [20, p. 808].

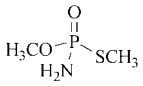

Methamidophos is prepared by isomerization of *O,O*-dimethylthiophosphamidate [34].

Methamidophos (announced in 1970) is a systemic insecticide and acaricide with contact and stomach action. It is used for control of chewing and sucking insects and spider mites on ornamentals, vegetables, and fruits.
Trade Names. Monitor (Bayer, Tomen, Valent); Tamaron (Bayer); Metaphos (Eftymiadis); Patrole (Pruductos OSA).

Methidathion [*950-37-8*], *O,O*-dimethyl *S*-2,3-dihydro-5-methoxy-2-oxo-1,3,4-thiadiazol-yl methyl phosphorodithioate, $C_6H_{11}N_2O_4PS_3$, M_r 302.3, *mp* 39 – 40 °C, consists of colorless crystals which are sparingly soluble in water, moderately soluble in hexane and *n*-octanol, and soluble in ethanol, acetone, and toluene [20, p. 811].

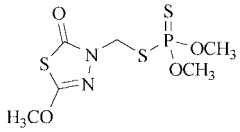

Methidathion is produced by reaction of 2-methoxy-1,3,4-thiadiazol-5(4*H*)-one with dimethyldithiophosphoric acid via the *N*-chloromethyl derivative or in a one-step reaction in the presence of formaldehyde [35].

Methidathion (announced in 1965) is a non-systemic insecticide and acaricide with contact and stomach action that is used against a wide range of sucking and chewing insects and spider mites in many crops.
Trade Names. Supracide (Novartis); Suprathion (Makhteshim-Agan).

Methiocarb [*2032-65-7*], (3,5-dimethyl-4-methylthio)phenyl methylcarbamate, $C_{11}H_{25}NO_2S$, M_r 225.3, *mp* 119 °C, consists of colorless crystals which are sparingly soluble in water, moderately soluble in hexane, and soluble in dichloromethane and isopropanol [20, p. 813].

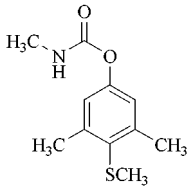

Methiocarb is produced by reaction of 4-methyl-3,5-dimethylphenol with chloroacetoacetate [36].

Methiocarb (announced in 1962) is used for control of slugs and snails in a wide range of agricultural applications: broad-range control of Lipidoptera, Coleoptera, Piptera, Thysanoptera, and Homoptera in vegetables, fruits, oilseed rape, and ornamentals.

Trade Names. Draza, Mesurol (Bayer).

Methomyl [*16752-77-5*], *S*-methyl-*N*-(methylcarbamoyloxy) thioacetimidate, $C_5H_{10}N_2O_2S$, M_r 162.2, *mp* 78 – 79 °C, consists of colorless crystals which are fairly soluble in water and highly soluble in methanol, ethanol, acetone, isopropanol [20, p. 815].

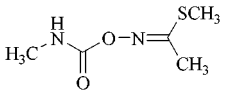

Methomyl is produced by chlorination of acetaldoxime and conversion of the resulting α-chlorooxime with sodium methylmercaptide [37].

Methomyl (announced in 1968) controls a wide range of insects and spider mites in fruits, vines, olives, hops, vegetables, and ornamentals.
Trade Names. Lannate (Du Pont); Methavin (Rhône-Poulenc); Methosan (Sanachem); Nudrin (Cyanamid).

Mevinphos [*26718-65-0*], 2-methoxycarbonyl-1-methylvinyl dimethyl phosphate, $C_7H_{13}O_6P$, M_r 224.1, *bp* 21 °C (*E* isomer), 6.9 °C (*Z* isomer), is a colorless liquid which is completely miscible with water and many organic solvents [20, p. 844].

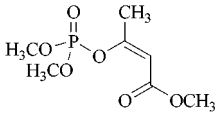

(*Z*)

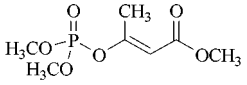

(*E*)

Mevinphos is produced by reaction of trimethyl phosphite with chloroacetoacetate [38].

Mevinphos (announced in 1953) is used for control of chewing and sucking insects and spiter mites on a wide range of crops.
Trade Names. Phosdrin (Cyanamid, Amvac); Duraphos (Amvac); Mevindrin (Hui Kwang).

Monocrotophos [*2157-98-4*], dimethyl (*E*)-1-methyl-2-(methylcarbamoyl)vinyl phosphate, $C_7H_{14}NO_5P$, M_r 223.2, *mp* 54 – 55 °C, consists of colorless, hygroscopic crystals which are soluble in water, methanol, acetone, and *n*-octanol [20, p. 849].

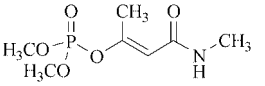

Monocrotophos is produced by reaction of trimethylphosphite with chloroacetoacetic acid methylamide in the presence of a base [39].

Monocrotophos (announced in 1959) is used for control of a wide range of pests on cotton, rice, maize, vegetables, and ornamentals.
Trade Names. Azodrin (Cyanamid); Nuvacron (Novartis); Apadrin (Rhône-Poulenc); Balwan (Rallis); Monodhan (Dhanuka).

Omethoate [*1113-02-6*], *O,O*-dimethyl *S*-methylcarbamoylmethyl phosphorothioate, $C_5H_{12}NO_4PS$, M_r 213.2, *mp* – 28 °C (decomp.), *bp* ca. 135 °C, is a colorless liquid which is readily soluble in water, alcohols, acetone, and many hydrocarbons, slightly soluble in diethyl ether, and almost insoluble in petroleum ether [20, p. 896].

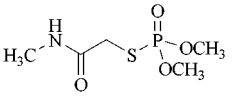

Omethoate is produced by reaction of *O,O*-dimethylphosphorylmercaptoacetic acid with methyl isocyanate [40].

Omethoate (announced in 1959) is used for control of spider mites, aphids, beetles, caterpillars, scale insects, thrips, and other pests on fruit, hops, cereal, rice, ornamentals, and other crops.
Trade Name. Folimat (Bayer).

Oxamyl [*23135-22-0*], *N,N*-dimethyl-2-methylcarbamoylimino-2-(methylthio)acetamide, $C_7H_{13}N_3O_3S$, M_r 219.3, *mp* 100 – 102 °C, consists of colorless crystals which are readily soluble in water, methanol, ethanol, acetone, and fairly soluble in toluene [20, p. 909].

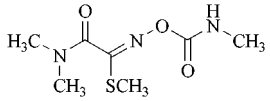

Oxamyl is produced by chlorination of the oxime of methylglycolate, reaction with methanethiol and alkali, and conversion to the carbamate with methyl isocyanate [41].

Oxamyl (announced in 1968) is used for control of chewing and sucking insects, spider mites, and nematodes in ornamentals, vegetables, potatoes, and other crops.

Trade Name. Vydate (DU Pont).

Phenthoate [*2597-03-7*], *O,O*-dimethyl *S*-(α-carboethoxy)phenylmethyl phosphorodithioate, $C_{12}H_{17}O_4PS_2$, M_r 320.4, *mp* 186 – 187 °C, consists of colorless crystals which are slightly soluble in water but readily soluble in many organic solvents [20, p. 952].

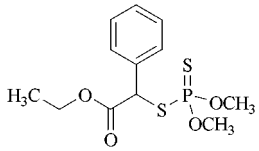

Phenthoate is produced by reaction of the sodium salt of *O,O*-dimethyldithiophosphonic acid with phenylbromoethyl acetate [42].

Phenthoate (announced in 1955) is used for control of aphids, scale insects, jassids, lacebugs, etc., in cereals, maize, rice, coffee, sunflowers, sugar cane, and other crops.

Trade Names. Elsan (Nissan); Cidial (Isagro); Aimsan (Aimco).

Phorate [*298-02-2*], *O,O*-diethyl *S*-(2-ethylthio)methyl phosphorodithioate, $C_7H_{17}O_2PS_3$, M_r 260.4, *mp* < -15 °C, *bp* 118 – 120 °C (0.1 kPa), is a colorless liquid which is sparingly soluble in water but miscible with alcohols, ketones, ethers, and esters [20, p. 959].

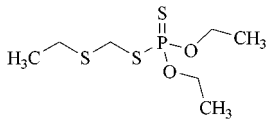

Phorate is produced by reaction of *O,O*-diethyldithiophosphoric acid with ethanethiol and formaldehyde [43].

Phorate (announced in 1948) is a systemic insecticide and acaricide used for control of Agromyzidae, Aleyrodidae, Aphididae, Chrysomelidae, Noctuidae, Pyralidae, Tetranychidae, and certain nematodes in a variety of crops.

Trade Names. Thimet (Cyanamid); Ramcides (Kunurai); Umet (United Phosphorus).

Phosalone [*2310-17-0*], *O,O*-diethyl *S*-(6-chloro-2,3-dihydro-2-oxobenzoxazol-3-yl)methyl phosphorodithioate, $C_{12}H_{15}ClNO_4PS_2$, M_r 367.8, *mp* 42 – 48 °C, consists of colorless crystals which are sparingly soluble in water, fairly soluble in hexane, and readily soluble in many organic solvents [20, p. 961].

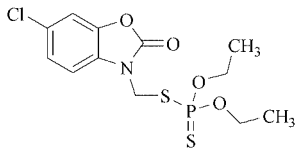

Phosalone is produced by treating *O,O*-diethyldithiophosphoric acid with *N*-chloromethyl-5-chlorobenzoxazolone [44].

Phosalone (announced in 1963) is used as a nonsystemic acaricide and insecticide, primarly in pome and stone fruit trees against Coleoptera, Homoptera, Lepidoptera and Thysanoptera.

Trade Name. Zolone (Rhône-Poulenc).

Phosmet [*732-11-6*], *O,O*-dimethyl *S*-(*N*-phthalimidomethyl) phosphorodithioate, $C_{11}H_{12}NO_4PS_2$, M_r 317.3, *mp* 72 – 72.7 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in acetone, toluene, xylene, and methanol [20, p. 963].

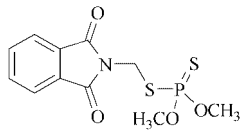

Phosmet is produced by reaction of *N*-chloromethylphthalimide with dimethyldithiophosphoric acid [45].

Phosmet (announced in 1961) is used for control of lepidopterous larvae, aphids, suckers,

fruit flies, and spider mites on pome fruit, stone fruit, citrus fruit, ornamentals, and vines.
Trade Names. Prolate (Gowan); Fosdan (General Quimica); Inovat (Productos OSA).

Phosphamidon [*13171-21-6*], *O,O*-dimethyl *O*-(2-chloro-2-diethylcarbamoyl-1-methyl)vinyl phosphate, $C_{10}H_{19}ClNO_5P$, M_r 299.7, *bp* 162 °C (2 kPa), is a pale yellow liquid which is miscible with water and many organic solvents with the exception of aliphatic hydrocarbons [20, p. 965].

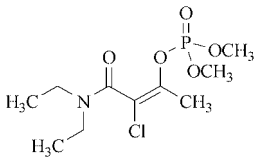

(*E*)

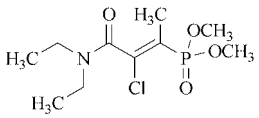

(*Z*)

Phosphamidon is produced by reaction of trimethylphosphite with α,α-dichloroacetic acid diethylamide [46].

Phosphamidon (announced in 1956) is used for control of sucking, chewing, and boring insects, and spider mites on a wide range of crops.
Trade Names. Dimecron (Novartis); Rilan (Rallis); Kinadon (United Phosphorus).

Pirimiphos-methyl [*29232-93-7*], *O,O*-dimethyl *O*-2-diethylamino-6-methylpyrimidin-4-yl phosphorothioate, $C_{11}H_{20}N_3O_3PS$, M_r 305.3, *bp* 15 – 18 °C, is a straw-colored liquid which is sparingly soluble in water but miscible with most organic solvents [20, p. 988].

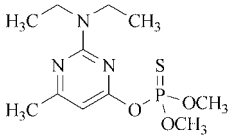

Pirimiphos-methyl is produced by condensation of diethylguanidine with ethyl acetate, followed by reaction with dimethylphosphoryl chloride [47].

Pirimiphos-methyl (announced in 1966) is used for control of a wide range of insects and mites in warehouses, stored grain, and animal houses, as well as industrial and domestic premises.
Trade Names. Actellic (Zeneca); Actellifog (Hortochem).

Profenofos [*41198-08-7*], *O*-4-bromo-2-chlorophenyl *O*-ethyl *S*-propyl phosphorothioate, $C_{11}H_{15}BrClO_3PS$, M_r 373.6, *bp* 100 °C (1.8 Pa), is a pale yellow liquid which is sparingly soluble in water but readily miscible with most organic solvents [20, p. 1006].

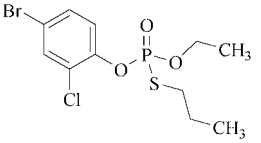

Profenofos is produced by reaction of *O*-(4-bromo-2-chlorophenyl) *O,O*-diethyl thiophosphate with propylbromide and triethylenediamine in acetonitrole [48].

Profenofos (announced in 1975) is used for control of insects, particularly Lepidoptera, and mites on cotton, maize, sugar beet, soya beans, potatoes, vegetables, tobacco, and other crops.
Trade Names.Curacron (Novartis); Sanofos (Sanonda).

Quinalphos [*13593-03-8*], *O,O*-diethyl *O*-quinoxalin-2-yl phosphorothioate, $C_{12}H_{15}N_2O_3PS$, M_r 298.3, *mp* 31 – 32 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in many organic solvents [20, p. 1077].

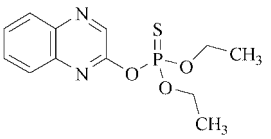

Quinalphos is produced by cyclization of *o*-phenylenediamine with chloroacetic acid, oxidation of the resulting 2-hydroxyquinoxaline and reaction with diethylthiophosphoryl chloride [49].

Quinalphos (announced in 1969) is used to control many insect pests such as Lepidoptera, Coleoptera and Diptera on fruit trees, cotton, vegetables, peanuts, and many other crops.
Trade Names. Ekalux (Novartis); Hubelux (Sanonda); Starlux (Shaw Wallace).

Sulfotepp [*3689-24-5*], *O,O,O′,O′*-tetraethyl dithiopyrophosphate, $C_8H_{20}O_5P_2S_2$, M_r 322.3, *bp* 136 – 139 °C (266 Pa), is a pale yellow liquid which is sparingly soluble in water but miscible with most organic solvents [20, p. 1131].

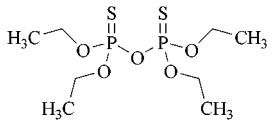

Sulfotepp is produced by reaction of diethylthionophosphoric acid chloride with water in the presence of pyridine and potassium carbonate [50].

Sulfotepp (announced in 1944) is used for fumigation control of aphids, thrips, spider mites, and whiteflies on glasshouse crops, and sciarid flies and phorid flies on mushrooms.
Trade Name. (Bayer).

Thiometon [*640-15-3*], *O,O*-dimethyl *S*-(2-ethylthio)ethyl phosphorodithioate, $C_6H_{15}O_2PS_2$, M_r 246.3, *bp* 110 °C (13 Pa), is a colorless liquid which is sparingly soluble in water but readily soluble in common organic solvents [20, p. 1199].

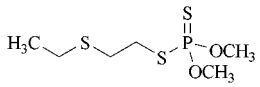

Thiometon is produced by reaction of the sodium salt of *O,O*-dimethyldithiophosphoric acid with ethylmercaptochloroethanol [51].

Thiometon (announced in 1952) is used for control of sucking insects, bryobia mites, tetranychid mites, and fruit-tree red spider mites on a wide range of crops.
Trade Name. Mavrik (Novartis).

Triazophos [*24017-47-8*], *O,O*-diethyl *O*-1-phenyl-1*H*-1,2,4-triazol-3-yl phosphorothioate, $C_{12}H_{16}N_3O_3PS$, M_r 313.3, *mp* −2 to −5 °C, is a pale yellow oil which is sparingly soluble in water and readily soluble in acetone, dichloromethane, methanol, isopropanol, ethyl acetate [20, p. 1226].

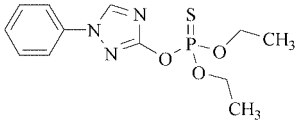

Triazophos is produced by reaction of 1-phenyl-3-hydroxy-1*H*-1,2,4-triazole, suspended in acetone, with diethoxythiophosphoryl chloride in the presence of triethylamine [52].

Triazophos (announced in 1970) is used for control of aphids, thrips, midges, beetles, lepidopterous larvae, and other soil insects, spider mites and other species of mites in ornamentals, fruit trees, vines, bananas, and many other crops.
Trade Names. Hostthion, Trilka, Spark (AgrEvo); Try (Sanonda).

Vamidothion [*2275-23-2*],*O,O*-dimethyl *S*-2-(1-methylcarbamoylethylthio)ethyl phosphorodithioate, $C_8H_{18}NO_4PS_2$, M_r 287.3, *mp* 43 °C, consists of colorless crystals which are readily soluble in water, benzene, toluene, acetonitrile, dichloromethane, cyclohexanone, and chloroform and sparingly soluble in cyclohexane, petroleum ether [20, p. 1263].

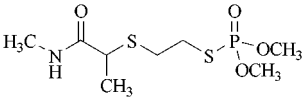

Vamidothion is produced by reaction of *O,O*-dimethylphosphoric acid chloride with the sodium salt of *N*-methylcarbamylethylmercaptoethanethiol [53].

Vamidothion (announced in 1962) is a systemic insecticide and acaricide that provides persistent control of *Eriosoma lanigerum* and other piercing and sucking Homoptera in cotton, hops, pome and stone fruit, and rice.
Trade Name. Kilval (Rhône-Poulenc).

3.2. Pyrethroids

Some pyrethroids, with their toxicity to insects, show acaricidal activity initially, but later the

mite population often increases [54]. It is improbable that pyrethroids will be able to provide long-term solutions to mite problems.

Acrinathrin [*101007-06-1*], (*S*)-α-cyano-3-phenoxybenzyl-(*Z*)-(1*R*,3*S*)-2,2-dimethyl-3-[2-(2,2,2-trifluoro-1-trifluoromethylethoxycarbonyl)vinyl]cyclopropanecarboxylate, $C_{26}H_{21}F_6NO_5$, M_r 541.4, *mp* 81.5 °C, consists of colorless crystals which are sparingly soluble in water, and moderately to highly soluble in most organic solvents [20, p. 16].

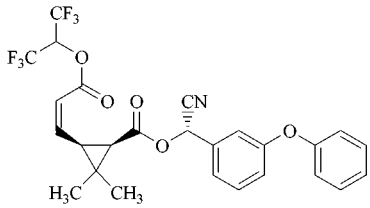

Acrimathrin is produced by esterification of 2,2-dimethyl-3-[3-oxo-3-{2,2,2-trifluoro-1-(trifluoromethyl)ethoxy}-1-propenyl]cyclopropanecarboxylic acid with cyano(3-phenoxyphenyl)methanol [55].

Acrimathrin (announced in 1983) acts on the central nervous system and is a contact and stomach acaricide/insecticide. It is effective against a wide range of phytophagous mites on vegetables, fruits, hops, ornamentals, vines, tobacco, and other crops.

Trade Name. Rufast (AgrEvo).

Bifenthrin [*82657-04-3*], 2-methylbiphenyl-3-methyl-(*Z*)-(1*RS*,3*RS*)-3-(2-chloro-3,3,3-trifluoroprop-1-enyl)-2,2-dimethylcyclopropanecarboxylate, $C_{23}H_{22}ClF_3O_2$, M_r 422.9, *mp* 51 – 56 °C, is a viscous liquid, crystalline or waxy product which is sparingly soluble in water, slightly soluble in heptane, methanol, soluble in acetone, chloroform, diethyl ether, and toluene [20, p. 118].

Bifenthrin is produced by reaction of 3-(2-chloro-3,3,3-trifluoro-1-propenyl)-2,2-dimethylcyclopropanecarboxylic acid with 3-(bromomethylene-2-methyl)biphenyl in the presence of potassium hydroxide [56].

Bifenthrin (announced in 1984) is effective against a broad range of foliar pests. It also controls some species of Acarina on cereals, citrus, cotton, fruits, grapes, and ornamentals.

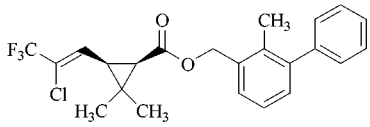

(*Z*)-(1*R*)-*cis*-

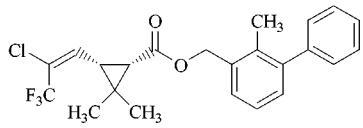

(*Z*)-(1*S*)-*cis*-

Trade Name. Talstar (FMC).

Fenpropathrin [*64257-84-7*], (*R,S*)-α-cyano-3-phenoxybenzyl 2,2,3,3-tetramethylcyclopropanecarboxylate, $C_{22}H_{23}NO_3$, M_r 349.4, *mp* 45 – 50 °C, is a yellow-brown solid which is sparingly soluble in water and soluble in xylene, cyclohexanone, and methanol [20, p. 524].

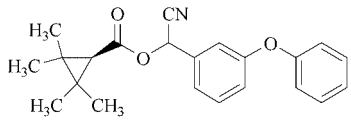

Fenpropathrin is produced by reaction of α-cyano-3-phenoxybenzyl alcohol with 2,2,3,3-tertamethylcyclopropanecarboxylic acid chloride [57].

Fenpropathrin (announced in 1981) is used to control many species of mites and insects on fruits, vegetables, field crops, ornamentals, and cotton.

Trade Names. Danitol (Valent/Sumitomo); Rody (Sumitomo/Cyanamid).

Fenvalerate [*51630-58-1*], (*R,S*)-α-cyano-3-phenoxybenzyl-(*R,S*)-2-(4-chlorophenyl)-3-methylbutyrate, $C_{25}H_{22}ClNO_3$, M_r 419.9, *mp* 39.5 – 53.7 °C, is a viscous yellow or brown liquid product, sometimes partly crystalline at room temperature, which is practically insoluble in water, moderately soluble in hexane and methanol, and soluble in xylene [20, p. 539].

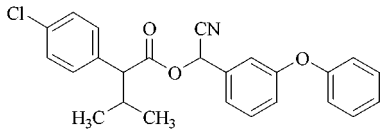

Fenvalerate is produced by reaction of α-cyano-3-phenoxybenzyl alcohol with 2-(4-chlorophenyl)-3-methylbutyric acid chloride [58].

Fenvalerate (announced in 1974) is used to control a wide range of pests, including those resistant to organochlorine, organophosphorous, and carbamate insecticides, in fruits, vines, olives, hops, nuts, vegetables, cotton, and oilseed rape.

Trade Names. Sumicidin (Sumitomo); Sanvalerate (Sanachem); Shasicidin (Sanonda).

Halfenprox [*111872-58-3*], 2-(4-bromodifluoromethoxyphenyl)-2-methylpropyl-3-phenoxybenzyl ether, $C_{24}H_{23}BrF_2O_3$, M_r 477.3, *bp* 291 °C, is a colorless clear liquid which is sparingly soluble in water [20, p. 654].

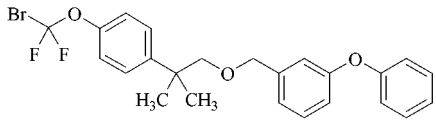

Halfenprox is produced by reaction of 3-phenoxybenzyl-2-(4-hydroxyphenyl)-2-methyl propyl ether and the potassium salt of *tert*-butanol with difluorodibromomethane in 1,3-dimethyl-2-imidazolidinone [59].

Halfenprox (announced in 1987) is a broad-spectrum acaricide with contact activity, it controls all stages of red spider mites, fruit tree red mites, two-spotted spider mites, and rust mites on citrus fruit, vines, tea, ornamentals, and other crops.

Trade Names. Anniverse, Sirbon (Mitsui Toatsu).

3.3. Formamidines

Products derived from the formamidines are used both as acaricides and as effective and selective insecticides. This class of compounds acts mainly as an ovicide on mites.

Amitraz [*33089-61-1*], N^1-methyl-N^2-(2,4-xylyl)-N^1-[*N*-(2,4-xylyl)-formimidoyl] formamidine, $C_{19}H_{23}N_3$, M_r 293.4, *mp* 86 – 87 °C, colorless needles, has low solubility in water (ca. 1 mg/L at 20 °C) but is soluble in acetone and toluene [16, p. 15].

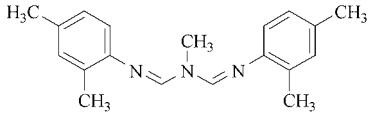

Amitraz is produced by the reaction of 2,4-xylidine, ethylorthoformate, and methylamine [60].

Amitraz is active against a wide range of mites and against all stages of phytophagous mites [61].

Trade Names. Taktic, Mitac (AgrEvo, NOR-AM); ByeBye (Agriphar); Edrizer (Siapa); Sender (Sanonda).

Formetanate [*22259-30-9*], 3-(dimethylaminomethyleneamino)phenyl *N*-methylcarbamate, $C_{11}H_{15}N_3O_2$, M_r 221.3, *mp* 102 °C, yellow crystals, negligible vapor pressure, has a solubility in water of < 1 g/L at 20 °C and is soluble in acetone, chloroform, and methanol [17, p. 5], [16, p. 284].

Formetanate is produced by the reaction of 3-dimethylaminomethyleneaminophenol with methyl isocyanate [62].

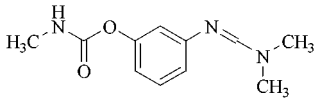

Formetanate is active against the mobile stages of fruit tree spider mites and is recommended also for use in greenhouses.

Trade Names. Dicarzol (AgrEvo); Carzol (NOR-AM).

3.4. Nitrophenyl Esters

These compounds are effective as fungicides against powdery mildews and as acaricides. The compounds are nonsystemic and are less phytotoxic than the corresponding phenols.

Dinobuton [*973-21-7*], 2-*sec*-butyl-4,6-dinitrophenyl isopropyl carbonate, $C_{14}H_{18}N_2O_7$, M_r 326.3, *mp* 60 – 62 °C, yellow crystals, negligible vapor pressure at room temperature, is insoluble in water but soluble in most organic solvents [16, p. 207], [17, p. 3], [18, p. 540].

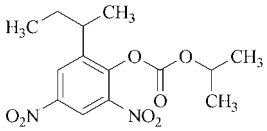

Dinobuton is produced by esterification of 2-*sec*-butyl-4,6-dinitrophenol with isopropyl chloroformate [63].

Dinobuton is used against red spider mites in greenhouses and in the field.

Trade Names. Acrex (Keno Gard AB, Murphy, Wacker); Dessin (Union Carbide).

Dinocap [*39300-45-3*], 2-(1-methylheptyl)-4,6-dinitrophenyl crotonate, $C_{18}H_{24}N_2O_6$, M_r 364.4, *bp* 138 – 140 °C at 6.7×10^{-5} kPa, dark brown liquid, very low vapor pressure at 20 °C, is insoluble in water but soluble in most organic solvents [19, p. 305].

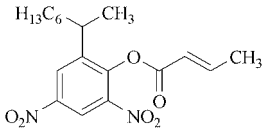

Dinocap is produced by nitration of *o*-(1-methylheptyl)phenol followed by esterification with crotonic acid [64]. The industrial product is a mixture of isomers [65] and is analyzed by the Kjeldahl method [66].

Trade Name. Acrex (Efthymiadis).

3.5. Sulfonic Acid Esters

Compounds in this class show mainly contact activity and have long residual effects without any adverse phytotoxic effects.

Tetradifon [*116-29-0*], 4-chlorophenyl-2,4,5-trichlorophenyl sulfone $C_{12}H_6Cl_4O_2S$, M_r 356.0, *mp* 146.5 – 147.5 °C, colorless crystals, vapor pressure 3.2×10^{-11} kPa at 20 °C, is slightly soluble in water, has low solubility in acetone and alcohols, but is more soluble in aromatic hydrocarbons, chloroform, and dioxane [16, p. 505], [17, p. 4], [18, p. 543].

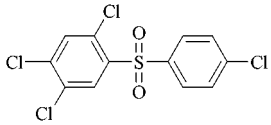

Tetradifon is produced by the Friedel – Crafts reaction of 2,4,5-trichlorobenzenesulfonyl chloride and chlorobenzene [67].

Tetradifon is recommended for use on tree fruit, citrus, grapes, vegetables, ornamental plants, and nursery stock.

Trade Names. Tedion V-18 (Uniroyal); Duracide (Endura); Tedone (Siapa).

3.6. Halogenated Benzhydrol Derivatives

These selective contact acaricides show activity toward a large variety of phytophagous mites. They are active against all postembryonal stages and eggs and are further characterized by their nonsystemic and good residual activity. Despite their resemblance to DDT they show no insecticidal activity.

Bromopropylate [*18181-80-1*], isopropyl-4,4′-dibromobenzilate $C_{17}H_{16}Br_2O_3$, M_r 428.1, *mp* 76 – 78 °C, colorless crystals, vapor pressure 6.8×10^{-9} kPa at 20 °C, is almost insoluble in water (< 10 mg/L at 20 °C) but soluble in most organic solvents [16, p. 54], [17, p. 5].

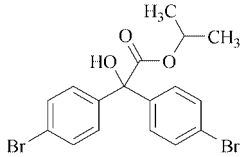

Bromopropylate is produced by esterification of 4,4′-dibromobenzilic acid with 2-propanol [68].

Bromopropylate is used against mites on citrus, grapes, cotton, vegetables, soybeans, apple and stone fruit, and ornamentals.

Trade Name. Neoron (Novartis).

Dicofol [*115-32-2*], 2,2,2-trichloro-1,1-bis(4-chlorophenyl)ethanol $C_{14}H_9Cl_5O$, M_r 370.5, *mp* 78.5 – 79.5 °C, colorless crystals, very low vapor pressure at 20 °C, is practically insoluble in water but soluble in most organic solvents [16, p. 183], [17, p. 4], [18, p. 534].

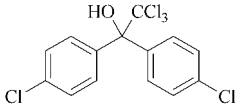

Dicofol is produced by chlorination of chlorfenethol [69] or by partial hydrolysis of 1,1,1,2-tetrachloro-2,2-bis(4-chlorophenyl)ethane with 80 – 90 % formic acid in the presence of a sulfonic acid at 120 – 125 °C [70].

Dicofol has been used successfully against spider mites on citrus, cotton, deciduous fruit, and other crops.

Trade Names. Kelthane (Rohm & Haas); Acarin, Mitigan (Makhteshim-Agan); Cecudifol (Ceguisa).

3.7. Organometallic Compounds

Chemicals of this class, all tin compounds, have opened a new era in mite control and are among the most important acaricides in use. They are characterized by their nonsystemic and excellent residual effectiveness. They control the mobile stages of a large variety of mites and show little phytotoxicity.

Azocyclotin [*41083-11-8*], 1-tricyclohexyltin-1*H*-1,2,4-triazole, $C_{20}H_{35}N_3Sn$, *M* 436.2, *mp* 218.8 °C, colorless crystals, vapor pressure $< 5\times10^{-6}$ kPa at 25 °C, is practically insoluble in water, and slightly soluble in cyclohexane, dichloromethane, toluene, and 2-propanol [19, p. 44].

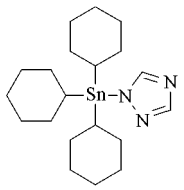

Azocyclotin is produced by the reaction of tricyclohexyltin chloride with the sodium salt of 1,2,4-triazole or by the reaction of 1,2,4-triazole with bis-tricyclohexyltin oxide [71].

Azocyclotin shows activity against a broad spectrum of spider mites [71].

Trade Name. Peropal (Bayer)

Cyhexatin [*13121-70-5*], tricyclohexyltin hydroxide, $(C_6H_{12})_3SnOH$, $C_{18}H_{34}OSn$, M_r 385.2, *mp* 195 – 198 °C, colorless crystals, vapor pressure negligible at 25 °C, is practically insoluble in water (< 1 mg/L at 25 °C), soluble in acetone, chloroform, and methanol, but otherwise has low solubility in most organic solvents [17, p. 6], [18, p. 142].

Cyhexatin is produced by the Grignard reaction of cyclohexylmagnesium chloride with tin tetrachloride followed by hydrolysis with aqueous sodium hydroxide [72].

Cyhexatin is an effective acaricide against a wide range of mites, mainly in deciduous fruit. It acts is by contact against the mobile stages [57].

Trade Names. Acarotine (Siapra); Pennstyl (Elf Atochem); Aramol (Diachem).

Fenbutatin oxide [*13356-08-6*], bis[tris(2-methyl-2-phenylpropyl)tin] oxide, $C_{60}H_{78}OSn_2$, M_r 1053, *mp* 138 – 139 °C, colorless crystalline powder, is insoluble in water, but soluble in acetone, dichloromethane, and benzene [16, p. 259].

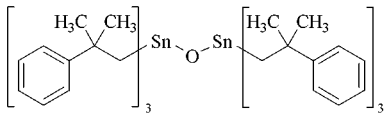

Fenbutatin oxide is produced by the treatment of tris(2-methyl-2-phenylpropyl)tin chloride with sodium hydroxide [74].

Fenbutatin oxide gives effective control of mites on citrus, fruit, vines, and ornamental plants.

Trade Names. Osaden, Torque (Cyanamid); Vendex (Cyanamid, DuPont).

3.8. Other Compounds

Abamectin [*7151-41-2*], a mixture containing at least 80 % avermectin B_{1a} ($C_{48}H_{72}O_{14}$, M_r 873.1) and not more than 20 % avermectin B_{1b} ($C_{47}H_{70}O_{14}$, M_r 860.1), *mp* 150 – 155 °C,

consists of colorless to pale yellow crystals which are sparingly soluble in water, fairly soluble in chloroform, ethanol, methanol, *n*-butanol, and cyclohexane, and soluble in toluene, acetone, and isopropanol [20, p. 3].

Abamectin is obtained from fermentation of *Streptomyces overmitilis* [75].

Abamectin (announced in 1981) acts by stimulating the release of γ-aminobutyric acid, an inhibitory neurotransmitter. It is an insecticide and acaricide with contact and stomach action that is used to control motile stages of mites, leaf miners, suckers, and other pests, on vegetables, fruits, and other crops.

Trade Names. Dynamec (Merck); Abacide (Mauget).

Benzoximate [*29104-30-1*], 3-chloro-α-ethoxyimino-2,6-dimethoxybenzyl benzoate, $C_{18}H_{18}ClNO_5$, M_r 363.8, *mp* 73 °C, colorless crystalline solid, is practically insoluble in water but soluble in benzene, hexane, xylene, and dimethylformamide [16, p. 35].

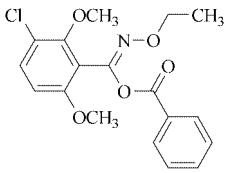

Benzoximate is produced by reaction of ethyl-3-chloro-2,6-dimethoxyhydroximate with benzoyl chloride in the presence of a base [76].

Benzoximate is a nonsystemic acaricide for use against mites on apples and citrus fruits.

Trade Names. Acarmate (Sipcam); Artaban (Procida); Citrazon (Nippon Soda).

Buprofezin [*69327-76-0*], 2-*tert*-butylimino-3-isopropyl-5-phenyl-1,3,5-thiazinan-4-one, $C_{16}H_{23}N_3OS$, M_r 305.4, *mp* 104.5 – 105.5 °C, consists of colorless crystals which are sparingly soluble in water but soluble in most organic solvents [20, p. 157].

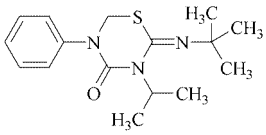

Buprofezin is prepared by cyclocondensation of *N*-phenyl-*N′*-*tert*-butylthiourea with *N*-methylenechloride *N*-carbonylchloride-2-propylamine [77].

Buprofezin (announced in 1981) is probably a chitin synthesis and prostaglandin inhibitor. It inhibits moulting of nymphs and larvae and also suppresses oviposition by adults. It is effective against Cicadellidae and Homoptera in rice, Cicadellidae in potatoes, and Aleyrodidae in citrus, cotton, and vegetables.

Trade Name. Applaud (Nihon Nohyaku).

Chinomethionate [*2439-01-2*], quinomethionate, 6-methyl-[1,3]dithiolo[4,5-*b*]quinoxalin-2-one, $C_{10}H_6N_2OS_2$, M_r 234.3, *mp* 171 – 172 °C, yellow crystals, vapor pressure 2.7×10^{-8} kPa at 20 °C, is practically insoluble in water but soluble in dimethylformamide, cyclohexanone, and petroleum oils [16, p. 463], [17, p. 6], [18, p. 561].

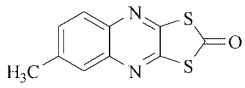

Chinomethionate is produced by the reaction of carbonyl chloride with 6-methyl-2,3-quinoxalinedithiol in aqueous alkaline solution [79].

Chinomethionate is a nonsystemic acaricide useful against mites on fruits, vegetables, and ornamental plants [80].

Trade Name. Morestan (Bayer)

Clofentezine [*74115-24-5*], 3,6-bis-(2-chlorophenyl)-1,2,4,5-tetrazine, $C_{14}H_8Cl_2N_4$, M_r 303.1, *mp* 182.3 °C, forms magenta crystals which are sparingly soluble in water, moderately soluble in hexane, ethanol, and fairly soluble in dichloromethane, and acetone [20, p. 254].

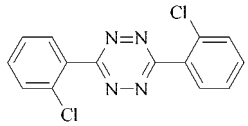

Clofentezine is obtained by treatment of 2-chlorobenzoylchloride with hydrazine [81].

Clofentezine (announced in 1981) is a specific acaricide with contact action and prolonged residual activity. It inhibits embryo development. It is used to control eggs and young motile

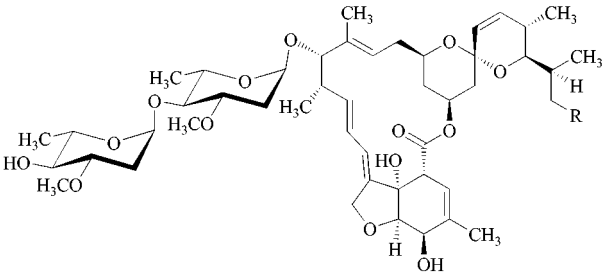

R = CH_3: Abamectin B_{1a}

R = H : Abamectin B_{1b}

stages of *Panonychus ulmi* and *Tetranychus* spp. on pome fruit, stone fruit, nuts, vines, and hops.
Trade Name. Apollo (AgrEvo).

Cymizaol [*61676-87-7*], *N*-2,3-dihydro-3-methyl-1,3-thiazol-2-ylidene-2,4-xylidine, $C_{12}H_{14}N_2S$, M_r 218.3, *mp* 44 °C, forms colorless crystals which are sparingly soluble in water but soluble in benzene, methanol, dichloromethane, and hexane [20, p. 197].

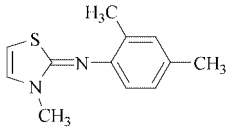

Cymizaol is produced by cyclocondensation of *N*-2,4-dimethylphenyl-*N′*-methylthiourea with chloroacetaldehyde [78].

Cymizaol (announced in 1977) is used in dips or as a spray to control all Acarina species, including strains resistant to organochlorines, organophosphates, and carbamates.
Trade Name. Tifatol (Novartis).

Dienochlor [*2227-17-0*], perchlorobi(2,4-cyclopentadienyl), $C_{10}Cl_{10}$, M_r 474.6, *mp* 122 – 123 °C, tan crystalline solid, vapor pressure 10^{-6} kPa at 25 °C, is insoluble in water, slightly soluble in ethanol, acetone, and aliphatic hydrocarbons, and moderately soluble in aromatic hydrocarbons [16, p. 186], [18, p. 529].

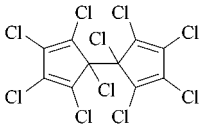

Dienochlor is produced by the catalytic reduction of hexachlorocyclopentadiene [82], [83].

Dienochlor is used mainly for control of mites on ornamentals and may act as an inhibitor of oviposition.
Trade Name. Pentac (Novartis).

Endosulfan [*115-29-7*], 6,7,8,9,10,10-hexachloro-1,5,5a,6,9,9a-hexahydro-6,9-methano-2,4,3-benzodioxathiepine 3-oxide, $C_9H_6Cl_6O_3S$, M_r 406.9, *mp* > 80 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in ethyl acetate, dichloromethane, toluene, ethanol, and hexane [20, p. 459].

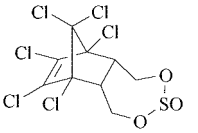

Endosulfan is produced by reaction of 1,2,3,4,7,7-hexachloro-5,6-dihydroxymethylbicylo[2,2,1]-hept-5-ene, obtained by condensation of *cis*-1,4-dihydroxybut-2-ene with hexachlorocyclopentadiene, with thionyl chloride [84].

Endosulfan (announced in 1956) acts as an antagonist of the GABA receptor chloride channel complex, and is used to control sucking, chewing, and boring insects and mites on a broad range of crops.
Trade Names. Fan (FMC); Thiodan (AgrEvo); Thionex (Makhteshim-Agan).

Fenazaquin [*120928-09-8*], 4-*tert*-butylphenylmethyl quinazolin-4-yl ether, $C_{20}H_{22}N_2O$, M_r 306.4, *mp* 77.5 – 80 °C, consists of colorless crystals which are practically insoluble in water but soluble in many organic solvents [20, p. 506].

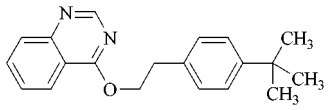

Fenazaquin is produced by reaction of 4-hydroxyquinazoline with phosphite and chlorine, and treating the resulting haloquinazoline with 4-*tert*-butylphenylethanol [85].

Fenazaquin (announced in 1992) affects the metabolism by inhibiting the mitochondrial electron transport chain by binding with complex I at coenzyme Q. It is a contact acaricide with good knockdown activity on motile forms, as well as true ovicidal activity, preventing eclosion of mite eggs. It is used against *Eutetranychus*, *Panonychus* and *Tetranychus* spp., and *Brevipalpus phoenics* in almonds, apples, citrus, cotton, grapes, and ornamentals.
Trade Name. Magister (DowElanco).

Fenpyroximate [*111812-58-9*], *tert*-butyl (*E*)-α-(1,3-dimethyl-5-phenoxypyrazol-4-yl-methyleneaminooxy)-*p*-toluate, $C_{24}H_{27}N_3O_4$, M_r 421.5, *mp* 101.1 – 102.4 °C, is a colorless crystalline powder which is practically insoluble in water, moderately soluble in methanol, and highly soluble in acetone, dichloromethane, chloroform, and THF [20, p. 530].

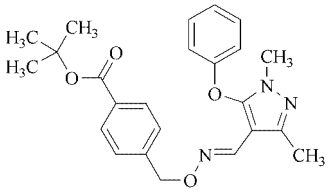

Fenpyroximate is produced by reaction of *tert*-butyl 4-(bromomethyl)benzoate with 1,3-dimethyl-5-phenoxypyrazole-4-carboxaldehyde oxime in the presence of potassium hydroxide in dimethylsulfoxide [86].

Fenpyroximate (announced in 1990) inhibits mitochondrial electron transport at complex I; it possesses quick knockdown activity against larvae, nymphs and adults. It is used to control important phytophagous mites such as Tetranychidae, Tarsonemidae, and other pests on citrus, apples, pears, grapes, tea, cherries, and melons.
Trade Names. Danitorante, Danitron (Nihon Noyaku).

Flucycloxuron [*94050-52-9*], 1-[α-(4-chloro-α-cyclopropylbenzylideneamino-oxo)-*p*-tolyl]-3-(2,6-difluorobenzoyl)urea, $C_{25}H_{20}ClF_2N_3O_3$, M_r483.9, *mp* 143.6 °C, consists of off-white to yellow crystals which are insoluble in water, sparingly soluble in cyclohexane, xylene, ethanol, and soluble in *N*-methylpyrrolidone [20, p. 562].

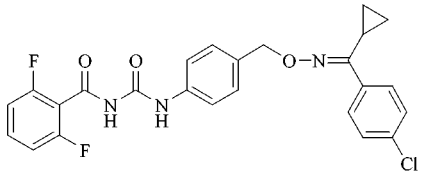

Flucycloxuron is produced by reaction of 4-chlorocyclopropylphenone with hydroxylamine to yield the corresponding oxime, which is treated with 4-nitrobenzylbromide, followed by reduction to give the corresponding amine, which is further treated with 2,6-difluorobenzoyl isocyanate [87].

Flucycloxuron (announced in 1988) is a chitin synthesis inhibitor that is only active against egg and larval stages (does not affect adult mites and insects). It is used on a variety of crops.
Trade Name. Andalin (Uniroyal).

Flufenoxuron [*101463-69-8*], 1-[-4-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-fluorophenyl]-3-(2,6-difluorobenzoyl)urea, $C_{21}H_{11}ClF_6N_2O_3$, M_r 488.8, *mp* 169 – 172 °C, is a colorless crystalline solid which is practically insoluble in water and moderately soluble in acetone, xylene, dichloromethane [20, p. 568].

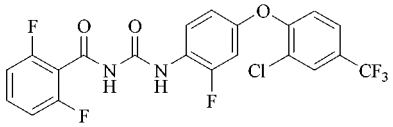

Flufenoxuron is produced by reaction of 2-fluoro-4-hydroxyaniline in potassium hydroxide-containing dimethyl sulfoxide with 3,4-dichlorotrifluoromethyltoluene to give the corresponding phenoxyaniline, which is treated with 2,6-difluorophenyl isocyanate [88].

Flufenoxuron is a chitin synthesis inhibitor and insect and acarid growth regulator with contact and stomach action. It is used for control of immature stages of many phytophagous mites and insect pests on pome fruits, vines, citrus fruits, tea, cotton, maize, vegetables, ornamentals, etc.

Trade Name. Cascade (Cyanamid).

Hexythiazox [*78587-05-0*], (4*RS*,5*RS*)-5-(4-chlorophenyl)-*N*-cyclohexyl-4-methyl-2-oxo-1,3-thiazolidine-3-carboxamide, $C_{17}H_{21}ClN_2O_2S$, M_r 352.9, *mp* 108.0 – 108.5 °C, consists of colorless crystals which are sparingly soluble in water, fairly soluble in hexane, acetonitrile, and soluble in chloroform, methanol, and acetone [20, p. 679].

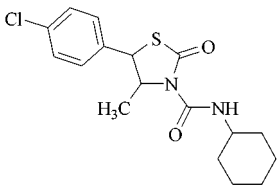

Hexythiazox is produced by treatment of *trans*-4-methyl-5-(4-chlorophenyl)-2-thiazolidinone with phenylisocyanate [89].

Hexythiazox (announced in 1982) possesses ovicidal, larvicidal, and nymiphicidal activity. It is used against eggs and larvae of many phytophagous mites on fruit, vines, vegetables, and cotton.

Trade Names. Cesar (AgrEvo); Zeldox (Zeneca); Ordoval (BASF); Matacar (Sipcam); Hexygon (Gowen).

Lufenuron [*103055-07-8*], (*RS*)-1-[2,5-dichloro-4-(1,1,2,3,3,3-hexafluoropropoxy)phenyl] 3-(2,6-difluorobenzoyl)urea, $C_{17}H_8Cl_2F_8N_2O_3$, M_r 511.2, *mp* 164.7 – 167.7 °C, forms colorless crystals which are practically insoluble in water, moderately soluble in ethanol, toluene, *n*-octanol, and *n*-hexane, and readily soluble in acetone [20, p. 753].

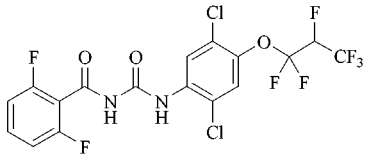

Lufenuron is produced by condensation of 2,5-dichloro-4-(1,1,2,3,3,3-hexafluoropropyloxy)aniline with 2,4-difluorobenzyl isocyanate in anhydrous toluene [90].

Lufenuron inhibits chitin synthesis and acts mostly by ingestion; larvae are unable to molt and also cease feeding. It is used for control of Lepidoptera and Coleoptera larvae on cotton, maize, and vegetables and citrus whiteflies and rust on citrus fruit.

Trade Name. Match (Novartis).

Propargite [*2312-35-8*], 2-(4-*tert*-butylphenoxy)cyclohexyl 12-propynyl sulfite, $C_{29}H_{26}O_4S$, M_r 350.5; the technical product is a dark, viscous oil, d_4^{25} 1.085 – 1.115, practically insoluble in water but soluble in most organic solvents [16, p. 447], [17, p. 4], [18, p. 560].

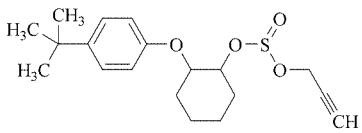

Proargite is produced by the reaction of 4-*tert*-butylphenol with 1,2-epoxycyclohexane, followed by treatment with thionyl chloride to give the chlorosulfinate, then with 2-propynol (propargyl alcohol) in the presence of a base [91].

Propargite is used mainly for the control of mites on fruit and other crops.

Trade Names. Omite (Uniroyal); Redador (Productos OSA).

Pyridaben [*96489-71-3*], 2-*tert*-butyl-5-(4-*tert*-butylbenzylthio)-4-chloropyridazin-3-(2*H*)-one, $C_{19}H_{25}ClN_2OS$, M_r 364.9, *mp* 111 – 112 °C, consists of colorless crystals which are practically insoluble in water, moderately soluble in ethanol, hexane, and *n*-octanol, and readily soluble in acetone, benzene, xylene, and cyclohexane [20, p. 1061].

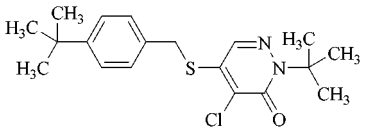

Pyridaben is produced by reaction of mucochloric acid with 4-*tert*-butylmercaptotoluene, followed by cyclocondensation of the resulting malealdehydic acid derivative with *N*-*tert*-butylhydrazine [92].

Pyridaben (announced in 1988) is active against all developing stages, especially against the larval and nymph stages, and is used for control of Acari, Aleyrodidae, Aphididae, Cicadellidae and Thysanoptera on field crops.

Trade Name. Sanmite (Nissan).

Tebufenpyrad [*119168-77-3*], *N*-(4-*tert*-butylbenzyl)-4-chloro-3-ethyl-1-methylpyrazole-5-carboxamide, $C_{18}H_{24}ClN_3O$, M_r 333.8, *mp* 61 – 61 °C, consists of colorless crystals which are sparingly soluble in water but soluble in acetone, methanol, chloroform, acetonitrile, hexane, and benzene [20, p. 1148].

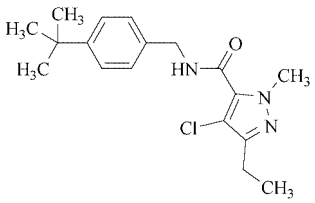

Tebufenpyrad is produced by treatment of 4-chloro-1-methyl-3-ethyl-pyrazole-5-carboxylic acid with thionyl chloride under reflux, followed by reaction of the resulting acid chloride with 4-*tert*-butylbenzylamine [93].

Tebufenpyrad (announced in 1988) is a mitochondrial respiration inhibitor that acts in the electron transport chain at site I. It is a nonsystemic acaricide that acts by contact and ingestion. It is used for control of all stages of *Tetranychus*, *Panonychus*, *Origonychus*, and *Eotetranychus* spp. on top fruit, vines, citrus, vegetables, hops, ornamentals, melons, and cotton.

Trade Names. Comanché, Masai, Oscar (Cyanamid); Pyranica (Mitsubishi).

4. Toxicology

The following summary gives an overview of the toxicological profile of various acaricides. The data were assembled by searching the literature and databanks available to the public. Because most of these compounds are proprietary products, many of the toxicological data are not published. Information in well-documented monographs, (e.g., WHO/FAO), in which the original data from the companies are available, were preferred to other published data, (e.g., handbooks), because the studies submitted to WHO/FAO were performed according to internationally accepted test protocols and good laboratory practices. The values presented should be viewed as a broad estimate of human toxicity as limited by strain, number of animal tested, study design, and applicability of the biological model.

4.1. Organophosphates and (Oxime)Carbamates

Toxic Action. Cumulative inhibition of cholinesterase causes toxic effects at doses not effective acutely. Delayed neurotoxicity leading to demyelination is observed with compounds that inhibit neurotoxic esterase. Carbamates act similarly, but reactivation of esterase occurs more rapidly [94, p. 531, p. 539].

Aldicarb [*116-06-3*]

LD_{50} (rat, oral)	0.62 – 1.23	mg/kg	[95]
LD_{50} (mouse, oral)	0.58 – 1.5	mg/kg	[95]
LD_{50} (guinea pig, oral)	1.0	mg/kg	[95]
LD_{50} (rabbit, oral)	1.3	mg/kg	[95]
LD_{50} (chicken)	9.0	mg/kg	[95]
LD_{50} (rat, dermal)	3.2 to > 10	mg/kg	[95]
LD_{50} (rabbit, dermal)	5.0 to > 10	mg/kg	[95]
Rabbit: nonirritating to eyes and skin			[95]
Guinea pig: not a skin sensitizer			[95]
Nonmutagenic			[95]
NOEL/NOAEL:			
Reproduction.			
Rat (feed):	0.3	$mg\,kg^{-1}\,d^{-1}$	[96, p. 33]
Teratogenicity.			
Rat (gavage):	0.125	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 33]
	0.25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 33]
Chronic.			
Rat (93-day feed):	0.1	$mg\,kg^{-1}\,d^{-1}$	[96, p. 33]
Mouse (103-week feed):	0.9	$mg\,kg^{-1}\,d^{-1}$	[96, p. 33]
Dog (one-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[96, p. 33]
Human.			
(single oral dose):	0.025	$mg\,kg^{-1}\,d^{-1}$ (cholinesterase)	[96, p. 33]

ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [96, p. 33]

Chlorfenvinphos [*470-90-6*]

LD_{50} (rat, oral)	9.6 – 39	mg/kg	[97, p. 15]
LD_{50} (mouse, oral)	117 – 200	mg/kg	[97, p. 15]
LD_{50} (rabbit, oral)	300 – 1000	mg/kg	[97, p. 15]
LD_{50} (dog, oral)	> 12000	mg/kg	[97, p. 15]
LD_{50} (guinea pig, oral)	125 – 250	mg/kg	[97, p. 15]
LD_{50} (rat, dermal)	30 – 108	mg/kg	[97, p. 15]
LD_{50} (rabbit, dermal)	412 – 4700	mg/kg	[97, p. 15]
Rabbit: slightly irritating to skin			[97, p. 15]
Rabbit: nonirritating to eyes			[97, p. 15]
Nonmutagenic			[97, p. 15]
NOEL/NOAEL:			
Reproduction.			
Rat (two-generation feed):	0.05	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[97, p. 15]
Teratogenicity.			
Rat (gavage):	1.0	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[97, p. 15]
	7.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[97, p. 15]
Rabbit (gavage):	< 25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[97, p. 15]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[97, p. 15]
Chronic.			
Rat (2-year feed):	0.15	$mg\,kg^{-1}\,d^{-1}$	[97, p. 15]
Mouse (2-year feed):	3.7	$mg\,kg^{-1}\,d^{-1}$	[97, p. 15]
Dog (one-year feed):	2.8	$mg\,kg^{-1}\,d^{-1}$	[97, p. 15]

ADI: $0.0005\,mg\,kg^{-1}\,d^{-1}$ [97, p. 15]

Chlorpyrifos-methyl [*5598-13-0*]

LD_{50} (rat, oral)	2680 – 3069	mg/kg	[98, p. 115]
LD_{50} (mouse, oral)	1122 – 2440	mg/kg	[98, p. 115]
LD_{50} (rabbit, oral)	ca. 2000	mg/kg	[98, p. 115]
LD_{50} (chick, oral)	> 7950	mg/kg	[98, p. 115]
LD_{50} (rat, dermal)	> 2000	mg/kg	[97, p. 15]
LD_{50} (mouse, dermal)	> 2856	mg/kg	[98, p. 115]
LC_{50} (rat, 4-h inhalation)	0.76	mg/L air	[98, p. 115]
Rabbit: nonirritating to skin and eyes			[98, p. 115]
Guinea pig: not a skin sensitizer			[98, p. 115]
Nonmutagenic			[98, p. 115]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	3	$mg\,kg^{-1}\,d^{-1}$	[98, p. 115]
Teratogenicity.			
Rat (gavage):	250	$mg\,kg^{-1}\,d^{-1}$	[98, p. 115]
Rabbit (gavage):	1.0	$mg\,kg^{-1}\,d^{-1}$	[96, p. 111]
Chronic.			
Rat (2-year feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[96, p. 111]
Mouse (78-week feed):	3.9	$mg\,kg^{-1}\,d^{-1}$	[98, p. 115]
Dog (2-year feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[98, p. 115]
Human.			
(4-week capsule):	0.1	$mg\,kg^{-1}\,d^{-1}$	[98, p. 115]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [96, p. 111]

Diazinon [*333-41-51*]

LD_{50} (rat, oral)	300 – 1012	mg/kg	[99, p. 57]
LD_{50} (mouse, oral)	187	mg/kg	[99, p. 57]
LD_{50} (rat, dermal)	> 2150	mg/kg	[99, p. 57]
LD_{50} (rabbit, dermal)	> 20	mg/kg	[99, p. 57]
LC_{50} (rat, 4-h inhalation)	2.3	mg/L air	[99, p. 57]
Rabbit: nonirritating to eyes and skin			[99, p. 57]
Guinea pig: not a skin sensitizer			[99, p. 57]
Nonmutagenic			[99, p. 57]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[99, p. 57]
Teratogenicity.			
Rat (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[99, p. 57]
Rabbit (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 57]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 57]
Chronic.			
Rat (99-week feed):	0.07	$mg\,kg^{-1}\,d^{-1}$	[99, p. 57]
Dog (one-year feed):	0.02	$mg\,kg^{-1}\,d^{-1}$	[99, p. 57]
Human.			
(34 – 36-day capsule):	0.025	$mg\,kg^{-1}\,d^{-1}$	[99, p. 57]

ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$ [99, p. 57]

Dicrotophos [*141-66-2*]

LD_{50} (rat, oral)	17 – 22	mg/kg	[20, p. 382]
LD_{50} (mouse, oral)	15	mg/kg	[20, p. 382]
LD_{50} (rat, dermal)	110 – 180	mg/kg	[20, p. 382]
LD_{50} (rabbit, dermal)	224	mg/kg	[20, p. 382]
LC_{50} (rat, 4-h inhalation)	0.09	mg/L air	[20, p. 382]
Rabbit: slightly irritating to skin and eyes			[20, p. 382]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	0.02	$mg\,kg^{-1}\,d^{-1}$	[20, p. 382]
Chronic.			
Rat (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[20, p. 382]
Dog (2-year feed):	0.04	$mg\,kg^{-1}\,d^{-1}$	[20, p. 382]

Dimethoate [*60-51-5*]

LD_{50} (rat, oral)	600 – 700	mg/kg	[100, p. 1015]
LD_{50} (mouse, oral)	60	mg/kg	[100, p. 1015]
LD_{50} (rabbit, oral)	500	mg/kg	[100, p. 1015]
LD_{50} (cat, oral)	100	mg/kg	[100, p. 1015]
LD_{50} (guinea pig, oral)	550	mg/kg	[100, p. 1015]
LD_{50} (horse, oral)	> 50	mg/kg	[101]
LD_{50} (sheep, oral)	80	mg/kg	[101]
LD_{50} (cattle, oral)	70	mg/kg	[101]
LD_{50} (hen, oral)	50	mg/kg	[101]
LD_{50} (pheasant)	15 – 20	mg/kg	[101]
LD_{50} (rat, dermal)	100 – 443	mg/kg	[100, p. 1015]
LD_{50} (rabbit, dermal)	1500	mg/kg	[100, p. 1015]
LC_{50} (rat, 4-h inhalation)	> 0.2	mg/L air	[100, p. 1015]
Rabbit: nonirritating to skin and eyes			[100, p. 1015]
Guinea pig: not a skin sensitizer			[102, p. 79]
Nonmutagenic			[102, p. 79]
NOEL/NOAEL:			
Teratogenicity.			
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$	[102, p. 79]
Chronic.			
Rat (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[102, p. 79]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [102, p. 79]

Disulfoton [*298-04-4*]

LD_{50} (rat, oral)	1.9 – 9.6	mg/kg	[103]
LD_{50} (mouse, oral)	7.0 – 28	mg/kg	[103]
LD_{50} (dog, oral)	ca. 5	mg/kg	[103]
LD_{50} (rat, dermal)	15.9 – 22.6	mg/kg	[103]
LD_{50} (mouse, dermal)	35	mg/kg	[103]
LC_{50} (rat, 4-h inhalation)	0.015	mg/L air	[103]
Guinea pig: not a skin sensitizer			[103]
Nonmutagenic			[103]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.15	$mg\,kg^{-1}\,d^{-1}$ (parental)	[103]
	0.05	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[103]
Teratogenicity.			
Rat (gavage):	0.3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[103]
	0.1	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[103]
Rabbit (gavage):	0.3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[103]
	1.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[103]
Chronic.			
Rat (2-year feed):	0.06	$mg\,kg^{-1}\,d^{-1}$	[103]
Mouse (99-week feed):	0.55	$mg\,kg^{-1}\,d^{-1}$	[103]
Dog (2-year feed):	0.03	$mg\,kg^{-1}\,d^{-1}$	[103]
Human.			
	0.01	$mg\,kg^{-1}\,d^{-1}$	[103]

ADI: $0.0003\,mg\,kg^{-1}\,d^{-1}$ [103]

EPN [*2104-64-5*]

LD_{50} (rat, oral)	24 – 36	mg/kg	[20, p. 464]
LD_{50} (mouse, oral)	60 – 95	mg/kg	[20, p. 464]
LD_{50} (rat, dermal)	538 – 2850	mg/kg	[20, p. 464]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.15	$mg\,kg^{-1}\,d^{-1}$	[20, p. 464]
Mouse (78-week feed):	0.7 – 3.6	$mg\,kg^{-1}\,d^{-1}$	[20, p. 464]

Ethion [*563-12-2*]

LD_{50} (rat, oral)	47	mg/kg	[20, p. 480]
LD_{50} (mouse, oral)	40 – 45	mg/kg	[20, p. 480]
LD_{50} (guinea pig, oral)	40 – 48	mg/kg	[20, p. 480]
LD_{50} (rabbit, dermal)	1084	mg/kg	[20, p. 480]
LC_{50} (rat, 4-h inhalation)	0.45	mg/L air	[20, p. 480]
NOEL/NOAEL:			
Teratogenicity.			
Rat (gavage):	0.06	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[104, p. 65]
	0.2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[104, p. 65]
Chronic.			
Rat (2-year feed):	0.3	$mg\,kg^{-1}\,d^{-1}$	[20, p. 480]
Dog (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[20, p. 480]
Human.			
	0.15	$mg\,kg^{-1}\,d^{-1}$	[104, p. 65]

ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$ [20, p. 480]

Fenothiocarb [*62850-32-2*]

LD_{50} (rat, oral)	1150 – 1200	mg/kg	[20, p. 517]
LD_{50} (mouse, oral)	4875 – 7000	mg/kg	[20, p. 517]
LD_{50} (mouse, dermal)	> 8000	mg/kg	[20, p. 517]
LD_{50} (rat, dermal)	2075 – 2425	mg/kg	[20, p. 517]
LC_{50} (rat, 4-h inhalation)	> 0.79	mg/L air	[20, p. 517]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	1.86 – 1.94	$mg\,kg^{-1}\,d^{-1}$	[20, p. 517]
Dog (one-year feed):	1.5 – 3.0	$mg\,kg^{-1}\,d^{-1}$	[20, p. 517]

ADI: $0.0075\,mg\,kg^{-1}\,d^{-1}$ [20, p. 517]

Formothion [*2540-82-1*]

LD_{50} (rat, oral)	365 – 500	mg/kg	[20, p. 625]
LD_{50} (mouse, oral)	190 – 195	mg/kg	[20, p. 625]
LD_{50} (rabbit, oral)	570	mg/kg	[20, p. 625]
LD_{50} (cat, oral)	210	mg/kg	[20, p. 625]
LD_{50} (rat, dermal)	> 1000	mg/kg	[20, p. 625]
LC_{50} (rat, 4-h inhalation)	3.2	mg/L air	[20, p. 625]
Rabbit: slightly irritating to skin			[20, p. 625]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	4.0	$mg\,kg^{-1}\,d^{-1}$	[20, p. 625]
Dog (2-year feed):	2.0	$mg\,kg^{-1}\,d^{-1}$	[20, p. 625]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [20, p. 625]

Mecarbam [*2591-54-2*]

LD_{50} (rat, oral)	36 – 53	mg/kg	[20, p. 774]
LD_{50} (mouse, oral)	106	mg/kg	[20, p. 774]
LD_{50} (rat, dermal)	> 1220	mg/kg	[20, p. 774]
LC_{50} (rat, 6-h inhalation)	0.7	mg/L air	[20, p. 774]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.21	$mg\,kg^{-1}\,d^{-1}$	[105, p. 89]
Dog (one-year feed):	0.35	$mg\,kg^{-1}\,d^{-1}$	[105, p. 89]

ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$ [105, p. 89]

Methacrifos [*62610-77-9*]

LD_{50} (rat, oral)	678	mg/kg	[106, p. 276]
LD_{50} (mouse, oral)	58 – 66	mg/kg	[106, p. 276]
LD_{50} (dog, oral)	573	mg/kg	[106, p. 276]
LD_{50} (chicken, oral)	101	mg/kg	[106, p. 276]
LD_{50} (rat, dermal)	> 3100	mg/kg	[106, p. 276]
LD_{50} (rabbit, dermal)	2732	mg/kg	[106, p. 276]
LC_{50} (rat, 6-h inhalation)	2.5	mg/L air	[106, p. 276]
Rabbit: mildly irritating to skin			[106, p. 276]
Rabbit: nonirritating to eyes			[106, p. 276]
Guinea pig: not a skin sensitizer			[106, p. 276]
Nonmutagenic			[106, p. 276]

NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[104, p. 101]
Chronic.			
Rat (2-year feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[104, p. 101]
Mouse (21-22-month feed):	15	$mg\,kg^{-1}\,d^{-1}$	[104, p. 101]
Dog (26-week feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[104, p. 101]
Human.			
	0.06	$mg\,kg^{-1}\,d^{-1}$	[104, p. 101]

ADI: $0.006\,mg\,kg^{-1}\,d^{-1}$ [104, p. 101]

Methamidophos [*10265-92-6*]

LD_{50} (rat, oral)	∼ 20	mg/kg	[104, p. 105]
LD_{50} (rabbit, oral)	10 – 30	mg/kg	[20, p. 808]
LD_{50} (dog, oral)	10 – 30	mg/kg	[20, p. 808]
LD_{50} (cat, oral)	10 – 30	mg/kg	[20, p. 808]
LD_{50} (guinea pig, oral)	30 – 50	mg/kg	[20, p. 808]
LD_{50} (rat, dermal)	130	mg/kg	[20, p. 808]
LC_{50} (rat, 4-h inhalation)	0.2	mg/L air	[20, p. 808]
Rabbit: nonirritating to skin			[20, p. 808]
Rabbit: slightly irritating to eyes			[20, p. 808]
Guinea pig: not a skin sensitizer			[104, p. 105]
Nonmutagenic			[104, p. 195]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.1	$mg\,kg^{-1}\,d^{-1}$	[104, p. 105]
Mouse (2-year feed):	0.7	$mg\,kg^{-1}\,d^{-1}$	[20, p. 808]
Dog (one-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[104, p. 105]
Neurotoxicity.			
Hen:	0.3	$mg\,kg^{-1}\,d^{-1}$	[104, p. 105]
Human.			
(21-day):	0.04	$mg\,kg^{-1}\,d^{-1}$	[104, p. 105]

ADI: $0.004\,mg\,kg^{-1}\,d^{-1}$ [104, p. 105]

Methidathion [*950-37-8*]

LD_{50} (rat, oral)	26 – 43.8	mg/kg	[104, p. 233]
LD_{50} (rat, oral)	26	mg/kg	[104, p. 233]
LD_{50} (mouse, oral)	17	mg/kg	[104, p. 233]
LD_{50} (hamster, oral)	30	mg/kg	[104, p. 233]
LD_{50} (guinea pig, oral)	25	mg/kg	[104, p. 233]
LD_{50} (rabbit, oral)	80	mg/kg	[104, p. 233]
LD_{50} (dog, oral)	200	mg/kg	[104, p. 233]
LD_{50} (chicken, oral)	80	mg/kg	[104, p. 233]
LD_{50} (rat, dermal)	1546 – 1663	mg/kg	[104, p. 233]
Rabbit: irritating to eyes			[104, p. 233]
Rabbit: very slightly irritating to skin			[104, p. 233]
Guinea pig: not a skin sensitizer			[104, p. 105]
Nonmutagenic			[104, p. 233]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.43	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[104, p. 233]
Teratogenicity.			
Rat (gavage):	1.0	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[104, p. 233]
	2.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[104, p. 233]
Rabbit (gavage):	6	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[104, p. 233]
	12	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[104, p. 233]
Chronic.			
Rat (2-year feed):	0.16	$mg\,kg^{-1}\,d^{-1}$	[104, p. 233]
Mouse (23-month feed):	1.4	$mg\,kg^{-1}\,d^{-1}$	[104, p. 233]
Dog (one-year feed):	0.1	$mg\,kg^{-1}\,d^{-1}$	[104, p. 233]
Human.			
	0.11	$mg\,kg^{-1}\,d^{-1}$	[104, p. 233]

ADI: $0.001\,mg\,kg^{-1}\,d^{-1}$ [104, p. 233]

Methiocarb [*2032-65-7*]

LD_{50} (rat, oral)	13 – 135	mg/kg	[100, p. 95]
LD_{50} (mouse, oral)	52 – 58	mg/kg	[20, p. 813]
LD_{50} (guinea pig, oral)	14 – 100	mg/kg	[100, p. 95]
LD_{50} (dog, oral)	10 – 25	mg/kg	[100, p. 95]
LD_{50} (rat, dermal)	> 300 – > 5000	mg/kg	[100, p. 95]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[100, p. 95]
LC_{50} (rat, 4-h inhalation)	0.32	mg/L air	[100, p. 95]
Rabbit: nonirritating to skin and eyes			[20, p. 813]
Nonmutagenic			[100, p. 1161]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	1.3	$mg\,kg^{-1}\,d^{-1}$	[107, p. 31]
Dog (2-year feed):	0.13	$mg\,kg^{-1}\,d^{-1}$	[107, p. 31]

ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$ [107, p. 31]

Methomyl [*16752-77-5*]

LD_{50} (rat, oral)	17 – 45	mg/kg	[105, p. 97]
LD_{50} (rabbit, oral)	30	mg/kg	[105, p. 97]
LD_{50} (chicken, oral)	28	mg/kg	[105, p. 97]
LD_{50} (dog, oral)	20	mg/kg	[105, p. 97]
LD_{50} (chicken, oral)	28	mg/kg	[105, p. 97]
LD_{50} (rat, dermal)	> 1000	mg/kg	[105, p. 97]
LC_{50} (rat, 4-h inhalation)	0.30	mg/L air	[105, p. 97]
Rabbit: irritating to eyes			[20, p. 815]
Guinea pig: not a skin sensitizer			[20, p. 815]
Nonmutagenic			[105, p. 97]
NOEL/NOAEL:			
Reproduction.			
Rat (feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[105, p. 97]
Teratogenicity.			
Rat (feed):	8.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[105, p. 97]
	34	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[105, p. 97]
Rabbit (gavage):	6	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[105, p. 97]
	16	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[105, p. 97]
Chronic.			
Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[105, p. 97]
Mouse (2-year feed):	8.7	$mg\,kg^{-1}\,d^{-1}$	[105, p. 97]
Dog (2-year feed):	3.1	$mg\,kg^{-1}\,d^{-1}$	[105, p. 97]
Human.			
Cholinesterase (in vitro):	$I_{50} = 265$	µmol/L	[108, p. 157]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [105, p. 97]

Mevinphos [*26718-65-03*]

LD_{50} (rat, oral)	3.7 – 7	mg/kg	[100, p. 1006]
LD_{50} (mouse, oral)	4.3 – 12.9	mg/kg	[100, p. 1006]
LD_{50} (rat, dermal)	4.2 – 4.7	mg/kg	[100, p. 1006]
LD_{50} (mouse, dermal)	40 – 50.2	mg/kg	[100, p. 1006]
LD_{50} (rabbit, dermal)	16 – 33	mg/kg	[20, p. 844]
LC_{50} (rat, one-h inhalation)	0.125	mg/L air	[20, p. 844]
Rabbit: mildly irritating to eyes and skin			[20, p. 844]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.02	$mg\,kg^{-1}\,d^{-1}$	[108, p. 385]
Dog (2-year capsule):	0.025	$mg\,kg^{-1}\,d^{-1}$	[108, p. 385]
Human.			
	0.014	$mg\,kg^{-1}\,d^{-1}$	[108, p. 385]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [108, p. 385]

Monocrotophos [*6923-22-4*]

LD_{50} (rat, oral)	14 – 23	mg/kg	[108, p. 423]
LD_{50} (mouse, oral)	15	mg/kg	[108, p. 423]
LD_{50} (rat, dermal)	112 – 126	mg/kg	[20, p. 849]
LD_{50} (rabbit, dermal)	130 – 250	mg/kg	[20, p. 849]
LC_{50} (rat, 4-h inhalation)	0.08	mg/L air	[20, p. 849]
Rabbit: nonirritating to skin and eyes			[20, p. 849]
Mutagenicity: ambiguous			[99, p. 333]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	0.05	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[99, p. 333]
Teratogenicity.			
Rat (gavage):	0.3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 333]
	2.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 333]
Rabbit (gavage):	1.0	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 333]
	3.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 333]
Chronic.			
Rat (2-year feed):	0.005	$mg\,kg^{-1}\,d^{-1}$	[109, p. 78]
Mouse (2-year feed):	< 0.15	$mg\,kg^{-1}\,d^{-1}$	[109, p. 78]
Dog (2-year feed):	0.0125	$mg\,kg^{-1}\,d^{-1}$	[108, p. 423]
Human.			
(oral, 30-day):	0.006	$mg\,kg^{-1}\,d^{-1}$	[99, p. 333]

ADI: $0.0006\,mg\,kg^{-1}\,d^{-1}$ [99, p. 333]

Omethoate [*1113-02-6*]

LD_{50} (rat, oral)	28 – 68	mg/kg	[97, p. 151]
LD_{50} (mouse, oral)	27 – 36	mg/kg	[97, p. 151]
LD_{50} (rabbit, oral)	50	mg/kg	[97, p. 151]
LD_{50} (cat, oral)	50	mg/kg	[97, p. 151]
LD_{50} (guinea pig, oral)	100	mg/g	[97, p. 151]
LD_{50} (hen, oral)	125	mg/kg	[97, p. 151]
LD_{50} (rat, dermal)	200	mg/kg	[97, p. 151]
LC_{50} (rat, 4-h inhalation)	ca. 0.3	mg/L air	[20, p. 896]
Rabbit: nonirritating to eyes and skin			[110, p. 117]
NOEL/NOAEL:			
Chronic.			
Mouse (2-year feed):	2.1 – 3.1	$mg\,kg^{-1}\,d^{-1}$	[110, p. 117]
Dog (one-year gavage):	0.025	$mg\,kg^{-1}\,d^{-1}$	[110, p. 117]
Rat (3-month feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[110, p. 117]

ADI: $0.0003\,mg\,kg^{-1}\,d^{-1}$ [110, p. 117]

Oxamyl [*23135-22-0*]

LD_{50} (rat, oral)	2.5 – 16	mg/kg	[111]
LD_{50} (mouse, oral)	2.3 – 3.3	mg/kg	[111]
LD_{50} (guinea pig, oral)	7.1	mg/kg	[111]
LD_{50} (rat, dermal)	> 1200	mg/kg	[111]
LD_{50} (rabbit, dermal)	130	mg/g	[106, p. 310]
LC_{50} (rat, 1-h inhalation)	0.12 – 0.17	mg/L air	[111]
Rabbit: mildly irritating to skin and eyes			[111]
Guinea pig: not a skin sensitizer			[111]
Nonmutagenic			[106, p. 310]
NOEL/NOAEL:			
Teratogenicity.			
Rabbit (gavage):	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[112, p. 59]
	2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[112, p. 59]
Rat (feed):	ca. 4	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[106, p. 310]
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[106, p. 310]
Reproduction.			
Rat (3-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[106, p. 310]
	2.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[106, p. 310]
Chronic.			
Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[112, p. 59]
Mouse (2-year feed):	3.6	$mg\,kg^{-1}\,d^{-1}$	[112, p. 59]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[112, p. 59]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [112, p. 59]

Phenthoate [*2597-03-7*]

LD_{50} (rat, oral)	245 – 440	mg/kg	[106, p. 404]
LD_{50} (mouse, oral)	360 – 840	mg/kg	[106, p. 404]
LD_{50} (dog, oral)	> 500	mg/kg	[106, p. 404]
LD_{50} (guinea pig, oral)	377	mg/kg	[106, p. 404]
LD_{50} (rabbit, oral)	ca. 210	mg/g	[106, p. 404]
LD_{50} (chicken, oral)	ca. 255	mg/kg	[106, p. 404]
LD_{50} (rat, dermal)	2100	mg/kg	[106, p. 404]
LD_{50} (rabbit, dermal)	1830	mg/kg	[106, p. 404]
LC_{50} (rat, 4-h inhalation)	> 0.8	mg/L air	[20, p. 952]
Rabbit: nonirritating to eyes and skin			[106, p. 404]
Nonmutagenic			[106, p. 404]

NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[106, p. 404]
Teratogenicity.			
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[106, p. 404]
Chronic.			
Rat (20-month feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[106, p. 404]
Mouse (18-month feed):	4.5	$mg\,kg^{-1}\,d^{-1}$	[112, p. 64]
Dog (2-year feed):	0.29	$mg\,kg^{-1}\,d^{-1}$	[106, p. 404]

ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [112, p. 64]

Phorate [*298-02-2*]

LD_{50} (rat, oral)	1.9 – 10	mg/kg	[113, p. 101]
LD_{50} (mouse, oral)	3 – 11	mg/kg	[113, p. 101]
LD_{50} (rat, dermal)	3.9 – 9.3	mg/kg	[113, p. 101]
LD_{50} (rabbit, dermal)	5.2	mg/kg	[113, p. 101]
LC_{50} (rat, 4-h inhalation)	0.06	mg/L air (male)	[113, p. 101]
	0.011	mg/L air (female)	[113, p. 101]
Nonmutagenic			[113, p. 101]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.17	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[113, p. 101]
Mouse (3-generation feed):	0.23	$mg\,kg^{-1}\,d^{-1}$	[113, p. 101]
Teratogenicity.			
Rat (gavage):	0.3	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[113, p. 101]
Rabbit (gavage):	0.15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[113, p. 101]
	1.2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[113, p. 101]
Chronic.			
Rat (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[113, p. 101]
Mouse (18-month feed):	0.45	$mg\,kg^{-1}\,d^{-1}$	[113, p. 101]
Dog (one-year capsule):	0.05	$mg\,kg^{-1}\,d^{-1}$	[113, p. 101]

ADI: $0.0005\,mg\,kg^{-1}\,d^{-1}$ [113, p. 101]

Phosalone [*2310-17-0*]

LD_{50} (rat, oral)	110 – 130	mg/kg	[99, p. 347]
LD_{50} (mouse, oral)	93 – 205	mg/kg	[99, p. 347]
LD_{50} (guinea pig, oral)	82 – 150	mg/kg	[99, p. 347]
LD_{50} (rat, dermal)	700 to > 2560	mg/kg	[99, p. 347]
LD_{50} (rabbit, dermal)	1530	mg/kg	[99, p. 347]
LC_{50} (rat, 4-h inhalation)	64	mg/L air	[108, p. 493]
Nonmutagenic			[99, p. 347]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[99, p. 347]
Teratogenicity.			
Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[99, p. 347]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[99, p. 347]
Chronic.			
Rat (2-year feed):	0.2	$mg\,kg^{-1}\,d^{-1}$	[99, p. 347]
Mouse (2-year feed):	23 – 31	$mg\,kg^{-1}\,d^{-1}$	[99, p. 347]
Dog (107-week feed):	5	$mg\,kg^{-1}\,d^{-1}$	[99, p. 347]

ADI: $0.001\,mg\,kg^{-1}\,d^{-1}$ [99, p. 347]

Phosmet [*732-11-6*]

LD_{50} (rat, oral)	100 – 300	mg/kg	[113, p. 119]
LD_{50} (mouse, oral)	20 – 50	mg/kg	[113, p. 119]
LD_{50} (guinea pig, oral)	200	mg/kg	[113, p. 119]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[113, p. 119]
LC_{50} (rat, 4-h inhalation)	0.15	mg/L air	[113, p. 119]
Rabbit: nonirritating to skin			[113, p. 119]
Rabbit: mildly irritating to eyes			[113, p. 119]
Mutagenicity: ambiguous			[113, p. 119]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	1.3	$mg\,kg^{-1}\,d^{-1}$ (parental)	[113, p. 119]
	5.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[113, p. 119]
Teratogenicity.			
Rat (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[113, p. 119]
	15	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[113, p. 119]
Rabbit (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[113, p. 119]
	2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[113, p. 119]
Chronic.			
Rat (2-year feed):	1.8	$mg\,kg^{-1}\,d^{-1}$	[113, p. 119]
Mouse (2-year feed):	4.0	$mg\,kg^{-1}\,d^{-1}$	[113, p. 119]
Dog (2-year feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[113, p. 119]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [113, p. 119]

Phosphamidon [*13171-21-6*]

LD_{50} (rat, oral)	24	mg/kg	[100, p. 1031]
LD_{50} (mouse, oral)	6.5 – 220	mg/kg	[100, p. 1031]
LD_{50} (rat, dermal)	107 – 143	mg/kg	[100, p. 1031]
LC_{50} (rat, 4-h inhalation)	0.18	mg/L air	[20, p. 963]
Rabbit: slightly irritating to skin			[20, p. 963]
Rabbit: moderately irritating to eyes			[20, p. 963]
Nonmutagenic			[105, p. 169]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	ca. 0.3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[105, p. 169]
Teratogenicity.			
Rat (gavage):	0.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[105, p. 169]
	2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[105, p. 169]
Rabbit (gavage):	3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[105, p. 169]
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[105, p. 169]
Chronic.			
Rat (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[105, p. 169]
Dog (one-year capsule):	0.1	$mg\,kg^{-1}\,d^{-1}$	[105, p. 169]

ADI: $0.0005\,mg\,kg^{-1}\,d^{-1}$ [105, p. 169]

Pirimiphos-methyl [*29232-93-7*]

LD_{50} (rat, oral)	1667 – 1861	mg/kg	[96, p. 325]
LD_{50} (mouse, oral)	1180	mg/kg	[96, p. 325]
LD_{50} (rabbit, oral)	1150 – 2300	mg/kg	[96, p. 325]
LD_{50} (guinea pig, oral)	1000 – 2000	mg/kg	[96, p. 325]
LD_{50} (cat, oral)	575 – 1150	mg/kg	[96, p. 325]
LD_{50} (dog, oral)	>1500	mg/kg	[96, p. 325]
LD_{50} (hen, oral)	30 – 60	mg/kg	[96, p. 325]
LD_{50} (quail, oral)	ca. 140	mg/kg	[96, p. 325]
LD_{50} (rat, dermal)	>4592	mg/kg	[20, p. 988]
LC_{50} (rat, 4-h inhalation)	>5.04	mg/L air	[20, p. 988]
Rabbit: slightly irritating to eyes and skin			[20, p. 980]
Guinea pig: not a skin sensitizer			[20, p. 988]
Nonmutagenic			[96, p. 325]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[96, p. 325]
Teratogenicity.			
Rat (gavage):	15	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[96, p. 325]
Rabbit (gavage):	<1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 325]
	16	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 325]
Chronic.			
Rat (2-year feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[96, p. 325]
Mouse (80-week feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[96, p. 325]
Dog (2-year capsule):	2	$mg\,kg^{-1}\,d^{-1}$	[96, p. 325]
Human.			
Oral (28-56-day):	0.25	$mg\,kg^{-1}\,d^{-1}$	[96, p. 325]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [96, p. 325]

Profenofos [*41198-08-7*]

LD_{50} (rat, oral)	358 – 502	mg/kg	[104, p. 123]
LD_{50} (mouse, oral)	298	mg/kg	[104, p. 123]
LD_{50} (Chinese hamster, oral)	153	mg/kg	[104, p. 123]
LD_{50} (rabbit, oral)	700	mg/kg	[104, p. 123]
LD_{50} (dog, oral)	>500	mg/kg	[104, p. 123]
LD_{50} (rat, dermal)	3300	mg/kg	[104, p. 123]
LD_{50} (rabbit, dermal)	475 – 2540	mg/kg	[104, p. 123]
LC_{50} (rat, 4-h inhalation)	3 – 3.6	mg/L air	[104, p. 123]
Rabbit: moderately irritating to eyes			[104, p. 123]
Rabbit: irritating to skin			[104, p. 123]
Nonmutagenic			[104, p. 123]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	1.0	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[104, p. 123]
Teratogenicity.			
Rat (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[104, p. 123]
	18	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[104, p. 123]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[104, p. 123]
	175	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[104, p. 123]
Chronic.			
Rat (2-year feed):	5.7	$mg\,kg^{-1}\,d^{-1}$	[104, p. 123]
Mouse (85-96-week feed):	5.8	$mg\,kg^{-1}\,d^{-1}$	[104, p. 123]
Dog (6-month feed):	2.9	$mg\,kg^{-1}\,d^{-1}$	[104, p. 123]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [104, p. 123]

Quinalphos [*13593-03-8*]

LD_{50} (rat, oral)	51 – 56	mg/kg	[114]
LD_{50} (mouse, oral)	55 – 59	mg/kg	[114]
LD_{50} (rat, dermal)	850 – 940	mg/kg	[114]
LC_{50} (rat, 4-h inhalation)	0.71	mg/L air	[114]
Rabbit: nonirritating to eyes and skin			[114]
Guinea pig: not a skin sensitizer			[114]
Nonmutagenic			[114]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	70.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[114]
Teratogenicity.			
Rat (gavage):	5.0	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[114]
	12.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[114]
Rabbit (gavage):	4	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[114]
Chronic.			
Rat (2-year feed):	0.2 – 0.26	$mg\,kg^{-1}\,d^{-1}$	[114]
Mouse (18-month feed):	0.7	$mg\,kg^{-1}\,d^{-1}$	[114]
Dog (2-year feed):	0.012 – 0.014	$mg\,kg^{-1}\,d^{-1}$	[114]

Sulfotep [*3689-24-5*]

LD_{50} (rat, oral)	ca. 10	mg/kg	[20, p. 1131]
LD_{50} (rat, dermal)	262	mg/kg	[20, p. 1131]
LC_{50} (rat, 4-h inhalation)	0.05	mg/L air	[20, p. 1131]
Rabbit: nonirritating to skin and eyes			[20, p. 1131]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[20, p. 1131]
Mouse (2-year feed):	7	$mg\,kg^{-1}\,d^{-1}$	[20, p. 1131]
ADI: $0.001\,mg\,kg^{-1}\,d^{-1}$ [20, p. 1131]			

Thiometon [*640-15-3*]

LD_{50} (rat, oral)	100 – 125	mg/kg	[115, p. 397]
LD_{50} (mouse, oral)	62 – 66	mg/kg	[115, p. 397]
LD_{50} (guinea pig, oral)	261	mg/kg	[115, p. 397]
LD_{50} (rabbit, oral)	95	mg/kg	[115, p. 397]
LD_{50} (rat, dermal)	1429 – 1997	mg/kg	[115, p. 397]
LC_{50} (rat, 4-h inhalation)	> 1.93	mg/L air	[20, p. 1199]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	> 1	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[116, p. 503]
Chronic.			
Rat (2-year feed):	0.12	$mg\,kg^{-1}\,d^{-1}$	[116, p. 503]
Dog (2-year feed):	0.15	$mg\,kg^{-1}\,d^{-1}$	[116, p. 503]
ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [116, p. 503]			

Triazophos [*24017-47-8*]

LD_{50} (rat, oral)	57 – 59	mg/kg	[116, p. 395]
LD_{50} (mouse, oral)	29 – 31	mg/kg	[116, p. 395]
LD_{50} (rat, dermal)	2000	mg/kg	[116, p. 395]
LC_{50} (rat, 4-h inhalation)	0.53	mg/L air	[116, p. 395]
Rabbit: nonirritating to skin and eyes			[116, p. 395]
Guinea pig: not a skin sensitizer			[116, p. 395]
Nonmutagenic			[105, p. 187]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	2 – 3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[98, p. 303]
Teratogenicity.			
Rat (feed):	22	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[105, p. 187]
Rabbit (gavage):	4	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[105, p. 187]
Chronic.			
Rat (2-year feed):	0.17	$mg\,kg^{-1}\,d^{-1}$	[98, p. 303]
Mice (2-year feed):	4.5	$mg\,kg^{-1}\,d^{-1}$	[98, p. 303]
Dog (one-year feed):	0.12	$mg\,kg^{-1}\,d^{-1}$	[98, p. 303]
Human.			
(3-week oral):	0.0125	$mg\,kg^{-1}\,d^{-1}$	[98, p. 303]
ADI: $0.001\,mg\,kg^{-1}\,d^{-1}$ [99, p. 391]			

Vamidothion [*2275-23-2*]

LD_{50} (rat, oral)	64 – 105	mg/kg	[115, p. 453]
LD_{50} (mouse, oral)	64	mg/kg	[115, p. 453]
LD_{50} (guinea pig, oral)	85	mg/kg	[115, p. 453]
LD_{50} (rabbit, oral)	160	mg/kg	[115, p. 453]
LD_{50} (dog, oral)	110	mg/kg	[115, p. 453]
LD_{50} (mouse, dermal)	1460	mg/kg	[115, p. 453]
LD_{50} (rabbit, dermal)	1160	mg/kg	[115, p. 453]
LC_{50} (rat, 4-h inhalation)	1.73	mg/L air	[20, p. 1263]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation gavage):	0.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[117, p. 75]
Subchronic.			
Rat (3-month feed):	0.054	$mg\,kg^{-1}\,d^{-1}$	[117, p. 75]
Mouse (3-month feed):	0.137	$mg\,kg^{-1}\,d^{-1}$	[117, p. 75]

Chronic.			
Dog (one-year gavage):	0.5	$mg\,kg^{-1}\,d^{-1}$	[117, p. 75]
Human.			
(5-day oral):	0.08	$mg\,kg^{-1}\,d^{-1}$	[117, p. 75]

ADI: $0.008\,mg\,kg^{-1}\,d^{-1}$ [117, p. 75]

4.2. Pyrethroids

Toxic action. General nerve activation with convulsion, paralysis, and respiratory failure; *cis* isomers are metabolized poorly and have higher toxicity [94, p. 586].

Acrinathrin [*101007-06-1*]

LD_{50} (rat, oral)	>5000	mg/kg	[20, p. 16]
LD_{50} (mouse, oral)	>5000	mg/kg	[20, p. 16]
LD_{50} (rat, dermal)	>2000	mg/kg	[20, p. 16]
LC_{50} (rat, 4-h inhalation)	1.6	mg/L air	[20, p. 16]
Rabbit: nonirritating to skin and eyes			[20, p. 16]
Guinea pig: not a skin sensitizer			[20, p. 16]
Nonmutagenic			[20, p. 16]
NOEL/NOAEL:			
Subchronic.			
Rat (90-day feed):	2.4 – 3.1	$mg\,kg^{-1}\,d^{-1}$	[20, p. 16]
Chronic.			
Dog (one-year feed):	3.0	$mg\,kg^{-1}\,d^{-1}$	[20, p. 16]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [20, p. 16]

Bifenthrin [*82657-04-3*]

LD_{50} (rat, oral)	56	mg/kg	[96, p. 79]
LD_{50} (mouse, oral)	43	mg/kg	[96, p. 79]
LD_{50} (rabbit, dermal)	>2000	mg/kg	[96, p. 79]
Rabbit: nonirritating to skin			[96, p. 79]
Rabbit: slightly irritating to eyes			[96, p. 79]
Guinea pig: not a skin sensitizer			[96, p. 79]
Nonmutagenic			[96, p. 79]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	3	$mg\,kg^{-1}\,d^{-1}$ (parental)	[96, p. 79]
	5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 79]
Teratogenicity.			
Rat (gavage):	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 79]
	>2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 79]
Rabbit (gavage):	2.7	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 79]
	>8	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 79]
Chronic.			
Rat (2-year feed):	4	$mg\,kg^{-1}\,d^{-1}$ (males)	[96, p. 79]
	7.5	$mg\,kg^{-1}\,d^{-1}$ (females)	[96, p. 79]
Mouse (20-month feed):	7.6	$mg\,kg^{-1}\,d^{-1}$ (males)	[96, p. 79]
	37	$mg\,kg^{-1}\,d^{-1}$ (females)	[96, p. 79]
Dog (one-year capsule):	1.5	$mg\,kg^{-1}\,d^{-1}$ (females)	[96, p. 79]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [96, p. 79]

Fenpropathrin [*64257-84-7*]

LD_{50} (rat, oral)	49 – 164	mg/kg	[99, p. 233]
LD_{50} (mouse, oral)	58 – 154	mg/kg	[99, p. 233]
LD_{50} (rabbit, oral)	510 – 675	mg/kg	[99, p. 233]
LD_{50} (dog, oral)	>1000	mg/kg	[99, p. 233]
LD_{50} (rat, dermal)	870 – 1600	mg/kg	[99, p. 233]
LD_{50} (mouse, dermal)	740 – 920	mg/kg	[99, p. 233]
LD_{50} (rabbit, dermal)	510 – 675	mg/kg	[99, p. 233]
LC_{50} (rat, 2-h inhalation)	>96	mg/L air	[99, p. 233]
LC_{50} (mouse, ?-h inhalation)	43 – 100	mg/L air	[99, p. 233]
Rabbit: slightly irritating to eyes			[99, p. 233]
Rabbit: nonrritating to skin			[99, p. 233]
Guinea pig: not a skin sensitizer			[99, p. 233]
Nonmutagenic			[99, p. 233]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[99, p. 233]
Teratogenicity.			
Rat (gavage):	3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 233]
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 233]
Rabbit (gavage):	4	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 233]
	36	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 233]
Chronic.			
Rat (2-year feed):	7	$mg\,kg^{-1}\,d^{-1}$	[99, p. 233]
Mouse (2-year feed):	56	$mg\,kg^{-1}\,d^{-1}$	[99, p. 233]
Dog (one-year feed):	3	$mg\,kg^{-1}\,d^{-1}$	[99, p. 233]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [99, p. 233]

Fenvalerate [*51630-58-1*]

LD_{50} (rat, oral)	451	mg/kg (DMSO)	[118]
	> 3200	mg/kg (PEG/Water)	[118]
LD_{50} (mouse, oral)	100 – 300	mg/kg (DMSO)	[118]
	1002	mg/kg (PEG/Water)	[118]
LD_{50} (chinese hamster, oral)	83 – 98	mg/kg (DMSO)	[118]
LD_{50} (hen, oral)	> 1500	mg/kg	[118]
LD_{50} (rat, dermal)	5000	mg/kg	[118]
LD_{50} (rabbit, dermal)	1000 – 3200	mg/kg	[118]
LC_{50} (rat, 3-h inhalation)	> 101	mg/L air	[118]
LC_{50} (mouse, 3-h inhalation)	101	mg/L air	[118]
Rabbit: nonirritating to skin and eyes			[118]
Guinea pig: not a skin sensitizer			[118]
Nonmutagenic			[119, p. 299]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[118]
	2.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[118]
Teratogenicity.			
Mouse (gavage):	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[118]
	50	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[118]
Rabbit (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[118]
	50	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[118]
Chronic.			
Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[120]
Mouse (78-week feed):	11.9	$mg\,kg^{-1}\,d^{-1}$	[121]
Dog (90-day feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[119, p. 299]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [122, p. 28]

Halfenprox [*111872-58-3*]

LD_{50} (rat, oral)	132 – 159	mg/kg	[20, p. 654]
LD_{50} (mouse, oral)	121 – 149	mg/kg	[20, p. 654]
LD_{50} (rat, dermal)	2000	mg/kg	[20, p. 654]
LC_{50} (rat, 4-h inhalation)	0.36 – 1.38	mg/L air	[20, p. 654]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	1.4	$mg\,kg^{-1}\,d^{-1}$	[20, p. 654]

ADI: $0.014\,mg\,kg^{-1}\,d^{-1}$ [20, p. 654]

4.3. Formamidines

Amitraz [*33089-61-1*]

LD_{50} (rat, oral)	600	mg/kg	[106, p. 1]
LD_{50} (mouse, oral)	> 1600	mg/kg	[106, p. 1]
LD_{50} (guinea pig, oral)	400 – 800	mg/kg	[106, p. 1]
LD_{50} (rabbit, oral)	> 100	mg/kg	[106, p. 1]
LD_{50} (dog, oral)	100	mg/kg	[106, p. 1]
LD_{50} (babboon, oral)	100 – 250	mg/kg	[106, p. 1]
LD_{50} (rat, dermal)	> 600	mg/kg	[106, p. 1]
LD_{50} (rabbit, dermal)	> 200	mg/kg	[106, p. 1]
LC_{50} (rat, 6-h inhalation)	65	mg/L air	[106, p. 1]
Rabbit: nonirritating to skin and eyes			[106, p. 1]
Guinea pig: not a skin sensitizer			[106, p. 1]
Nonmutagenic			[106, p. 1]
NOEL/NOAEL:			
Reproduction.			
Rat (3-generation feed):	1	$mg\,kg^{-1}\,d^{-1}$	[106, p. 1]
Teratogenicity.			
Rat (gavage):	1	$mg\,kg^{-1}\,d^{-1}$	[106, p. 1]
Chronic.			
Rat (two-year feed):	3	$mg\,kg^{-1}\,d^{-1}$	[123, p. 14]
Mouse (80-week feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[123, p. 14]
Dog (2-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[123, p. 14]

ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [106, p. 1]

Formetanate [*22259-30-9*]

LD_{50} (rat, oral)	15 – 26	mg/kg	[100, p. 1159]
LD_{50} (mouse, oral)	13 – 25	mg/kg	[100, p. 1159]
LD_{50} (dog, oral)	19	mg/kg	[100, p. 1159]
LD_{50} (rabbit, dermal)	> 10200	mg/kg	[100, p. 1159]
LD_{50} (rat, dermal)	> 5600	mg/kg	[100, p. 1159]
LC_{50} (rat, 4-h inhalation)	0.28 – 2.9	mg/L air	[100, p. 1159]
Rabbit: irritating to eyes			[20, p. 623]
Guinea pig: skin sensitizer			[20, p. 623]
Nonmutagenic			[100, p. 1159]

NOEL/NOAEL:			
Teratogenicity.			
Not teratogenic			[100, p. 1159]
Reproduction.			
Rat (2-generation feed):	2.5	$g\,kg^{-1}\,d^{-1}$ (parental and developmental)	[106, p. 1]
Chronic.			
Rat (2-year feed):	0.52	$mg\,kg^{-1}\,d^{-1}$	[20, p. 623]
Mouse (2-year feed):	8.2	$mg\,kg^{-1}\,d^{-1}$	[20, p. 623]
Dog (one-year feed):	0.37	$mg\,kg^{-1}\,d^{-1}$	[20, p. 623]
ADI: 0.037 $mg\,kg^{-1}\,d^{-1}$ [20, p. 623]			

4.4. Nitrophenyl Esters

Toxic Action. Uncoupling of oxidative phosphorylation; cataract formation [94, p. 556].

Dinobuton [*973-21-7*]

LD_{50} (rat, oral)	140	mg/kg	[124]
LD_{50} (mouse, oral)	2540	mg/kg	[125]
LD_{50} (rat, dermal)	> 5000	mg/kg	[126]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	3 – 6	$mg\,kg^{-1}\,d^{-1}$	[127]
Dog (one-year feed):	4.5	$mg\,kg^{-1}\,d^{-1}$	[127]

Dinocap [*39300-45-31*]

LD_{50} (male rat, oral)	1581 – 2321	mg/kg	[128, p. 75]
LD_{50} (female rat, oral)	510 – 1493	mg/kg	[128, p. 75]
LD_{50} (mouse, oral)	150 – 180	mg/kg	[128, p. 75]
LD_{50} (rabbit, oral)	2000 – 3000	mg/kg	[128, p. 75]
LD_{50} (dog, oral)	100	mg/kg	[128, p. 75]
LD_{50} (rabbit, dermal)	> 4700		[128, p. 75]
LC_{50} (rat, 4-h inhalation)	0.36	mg/L air	[20, p. 426]
Rabbit: moderately to markedly irritating, to skin and eyes			[20, p. 426]
Nonmutagenic			[20, p. 426]
NOEL/NOAEL:			
Reproduction.			
Rat (feed):	6.4	$mg\,kg^{-1}\,d^{-1}$	[128, p. 75]
Teratogenicity.			
Mouse (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[128, p. 75]
	40	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[128, p. 75]
Rat (gavage):	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[128, p. 75]
Syrian hamster (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[128, p. 75]
	12.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[128, p. 75]
Rabbit (gavage):	0.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[128, p. 75]
Chronic.			
Rat (30-month feed):	6.4	$mg\,kg^{-1}\,d^{-1}$ (males)	[128, p. 75]
	8.1	$mg\,kg^{-1}\,d^{-1}$ (females)	[128, p. 75]
Dog (one-year feed):	0.4	$mg\,kg^{-1}\,d^{-1}$	[128, p. 75]
Eye cateracts in duck and rabbit: NOAEL (rabbit, gavage)	9	$mg\,kg^{-1}\,d^{-1}$	[128, p. 75]
ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$ [128, p. 75]			

4.5. Sulfonic Acid Ester

Tetradifon [*116-29-0*]

LD_{50} (rat, oral)	> 14700	mg/kg	[20, p. 1177]
LD_{50} (rabbit, dermal)	> 10000	mg/kg	[20, p. 1177]
LC_{50} (rat, 4-h inhalation)	> 3	mg/L air	[20, p. 1177]
Rabbit: nonirritating to skin			[20, p. 1177]
Rabbit: slightly irritating to eyes			[20, p. 1177]
Nonmutagenic			[20, p. 1177]
Not teratogenic in rat and rabbit			[20, p. 1177]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	ca. 10	$mg\,kg^{-1}\,d^{-1}$	[20, p. 1177]
Chronic.			
Rat (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$ (males)	[20, p. 1177]

4.6. Halogenated Benzhydrol Derivatives

Bromopropylate [*18181-80-1*]

LD_{50} (rat, oral)	> 5000	mg/kg	[99, p. 39]
LD_{50} (mouse, oral)	8000	mg/kg	[99, p. 39]
LD_{50} (rabbit, oral)	> 6000	mg/kg	[99, p. 39]
LD_{50} (rat, dermal)	> 4000	mg/kg	[99, p. 39]
LD_{50} (rabbit, dermal)	> 2020	mg/kg	[99, p. 39]
LC_{50} (rat, 4-h inhalation)	> 4458	mg/L air	[99, p. 39]
Rabbit: nonirritating to eyes and skin			[99, p. 39]
Guinea pig: not a skin sensitizer			[99, p. 39]
Nonmutagenic			[99, p. 39]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	9	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[99, p. 39]
Teratogenicity.			
Rat (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 39]
	700	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 39]
Rabbit (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[99, p. 39]
	120	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[99, p. 39]
Chronic.			
Rat (2-year feed):	3.7	$mg\,kg^{-1}\,d^{-1}$	[99, p. 39]
Mouse (2-year feed):	16	$mg\,kg^{-1}\,d^{-1}$	[99, p. 39]
Dog (one-year feed):	2.7	$mg\,kg^{-1}\,d^{-1}$	[99, p. 39]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [99, p. 39]

Dicofol [*115-32-3*]

LD_{50} (rat, oral)	587 – 595	mg/kg	[107, p. 137]
LD_{50} (mouse, oral)	669 – 675	mg/kg	[107, p. 137]
LD_{50} (rabbit, oral)	1820	mg/kg	[107, p. 137]
LD_{50} (dog, oral)	> 4000	mg/kg	[107, p. 137]
LD_{50} (rat, dermal)	> 5000	mg/kg	[107, p. 137]
LD_{50} (rabbit, dermal)	> 2500	mg/kg	[107, p. 137]
LC_{50} (rat, 4-h inhalation)	> 5	mg/L air	[107, p. 137]
Rabbit: nonirritating to eyes			[107, p. 137]
Rabbit: irritating to skin			[107, p. 137]
Guinea pig: not a skin sensitizer			[134]
Nonmutagenic			[107, p. 137]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.25	$mg\,kg^{-1}\,d^{-1}$ (parental)	[107, p. 137]
	2.1	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 137]
Teratogenicity.			
Rat (gavage):	0.25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[107, p. 137]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 137]
Rabbit (gavage):	0.4	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[107, p. 137]
	4	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 137]
Chronic.			
Rat (2-year feed):	0.22	$mg\,kg^{-1}\,d^{-1}$	[107, p. 137]
Mouse (78-week feed):	> 40	$mg\,kg^{-1}\,d^{-1}$	[107, p. 137]
Dog (one-year feed):	0.82	$mg\,kg^{-1}\,d^{-1}$	[107, p. 137]

ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$ [107, p. 137]

4.7. Organometallic Compounds

Toxic Action. Interference with lymphocyte function; induction of heme oxygenase; decreased liver hemoprotein level [94, p. 353].

Azocyclotin [*41083-11-8*]

LD_{50} (rat, oral)	> 150	mg/kg	[116, p. 50]
LD_{50} (rat, oral)	209 – 363	mg/kg	[20, p. 69]
LD_{50} (guinea pig, oral)	261	mg/kg	[20, p. 69]
LD_{50} (mouse, oral)	870 – 980	mg/kg	[20, p. 69]
LD_{50} (rat, dermal)	> 5000	mg/kg	[116, p. 50]
LC_{50} (rat, 4-h inhalation)	~ 0.02	mg/L air	[116, p. 50]
Rabbit: irritating to skin and eyes			[116, p. 50]
Guinea pig: not a skin sensitizer			[116, p. 50]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[129, p. 22]
Teratogenicity.			
Rabbit (gavage):	0.1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[129, p. 22]
	1.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[129, p. 22]
Chronic.			
Rat (2-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[129, p. 22]
Mouse (2-year feed):	2.14	$mg\,kg^{-1}\,d^{-1}$	[129, p. 22]
Dog (2-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[129, p. 22]

ADI: $0.007\,mg\,kg^{-1}\,d^{-1}$ [129, p. 22]

Cyhexatin [*13121-70-5*]

LD_{50} (rat, oral)	540	mg/kg	[20, p. 302]
LD_{50} (rabbit, oral)	500 – 1000	mg/kg	[20, p. 302]
LD_{50} (guinea pig, oral)	780	mg/kg	[20, p. 302]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[20, p. 302]
Rabbit: irritating to eyes			[20, p. 302]
Rabbit: nonirritating to skin			[20, p. 302]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	0.1	$mg\,kg^{-1}\,d^{-1}$ (parental)	[113, p. 65]
	0.7	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[113, p. 65]
Teratogenicity.			
Rabbit (gavage):	0.75	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[113, p. 65]
Chronic.			
Mouse (2-year feed):	3	$mg\,kg^{-1}\,d^{-1}$	[113, p. 65]
Dog (one-year feed):	0.75	$mg\,kg^{-1}\,d^{-1}$	[98, p. 129]

ADI: $0.007\,mg\,kg^{-1}\,d^{-1}$ [113, p. 65]

Fenbutatin oxide [*13356-08-6*]

LD_{50} (rat, oral)	4400	mg/kg	[96, p. 193]
LD_{50} (mouse, oral)	1450	mg/kg	[96, p. 193]
LD_{50} (dog, oral)	> 1500	mg/kg	[96, p. 193]
LD_{50} (rabbit, oral)	1500 – 3000	mg/kg	[96, p. 193]
LD_{50} (rat, dermal)	> 1000	mg/kg	[96, p. 193]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[96, p. 193]
LC_{50} (rat, 4-h inhalation)	0.072	mg/L air	[96, p. 193]
Rabbit: irritating to skin			[96, p. 193]
Rabbit: severely irritating to eyes			[96, p. 193]
Nonmutagenic			[96, p. 193]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	6.0	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[96, p. 193]
Teratogenicity.			
Rabbit (gavage):	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 193]
	30	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 193]
Rabbit (gavage):	1	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[96, p. 193]
Chronic.			
Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[96, p. 193]
Dog (2-year gavage):	15	$mg\,kg^{-1}\,d^{-1}$	[96, p. 193]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [96, p. 193]

4.8. Other Compounds

Abamectin [*71751-41-2*]

LD_{50} (rat, oral)	8.7 – 12.8	mg/kg	[96, p. 3]
LD_{50} (mouse, oral)	13.6 – 23.8	mg/kg	[96, p. 3]
LD_{50} (dog, oral)	ca. 8	mg/kg	[96, p. 3]
LD_{50} (monkey, oral)	> 24	mg/kg	[96, p. 3]
LD_{50} (rat, dermal)	> 330	mg/kg	[96, p. 3]
LD_{50} (rabbit, dermal)	> 1600	mg/kg	[96, p. 3]
Rabbit: nonirritating to skin			[96, p. 3]
Rabbit: very slightly irritating to eyes			[96, p. 3]
Guinea pig: not a skin sensitizer			[96, p. 3]
Nonmutagenic			[96, p. 3]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation gavage):	0.05	$mg\,kg^{-1}\,d^{-1}$ (parental)	[96, p. 3]
	0.12	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 3]
Teratogenicity.			
Mice (gavage):	0.05	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 3]
	0.2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 3]
Rat (gavage):	1.6	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[96, p. 3]
	< 1.6	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[96, p. 3]
Rabbit (gavage):	1.0	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[96, p. 3]
Chronic.			
Rat (> 2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[96, p. 3]
Mouse (94-week feed):	4	$mg\,kg^{-1}\,d^{-1}$	[96, p. 3]
Dog (52-week feed):	< 0.25	$mg\,kg^{-1}\,d^{-1}$	[96, p. 3]

ADI: $0.0001\,mg\,kg^{-1}\,d^{-1}$ [96, p. 3]

Benzoximate [*29104-30-1*]

LD_{50} (rat, oral)	5000 – 15000	mg/kg	[20, p. 113]
LD_{50} (mouse, oral)	12000 – 14500	mg/kg	[20, p. 113]
LD_{50} (rat, dermal)	> 15000	mg/kg	[20, p. 113]
LD_{50} (mouse, dermal)	> 15000	mg/kg	[20, p. 113]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	20	$mg\,kg^{-1}\,d^{-1}$	[20, p. 113]

Buprofezin [*69327-76-0*]

LD_{50} (rat, oral)	1635 – 2355	mg/kg	[98, p. 75]
LD_{50} (mouse, oral)	> 1000	mg/kg	[98, p. 75]
LD_{50} (hamster, oral)	> 10000	mg/kg	[98, p. 75]
LD_{50} (rabbit, oral)	> 5000	mg/kg	[98, p. 75]
LD_{50} (rat, dermal)	> 5000	mg/kg	[98, p. 75]
LC_{50} (rat, 4-h inhalation)	> 4.6	mg/L air	[98, p. 75]
Rabbit: mildly irritating to skin			[98, p. 75]
Rabbit: nonirritating to eyes			[98, p. 75]
Nonmutagenic			[98, p. 75]
NOEL/NOAEL:			
Reproduction.			
Rat (one-generation feed):	6.4	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[98, p. 75]
Teratogenicity.			
Rat (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[98, p. 75]
	166	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[98, p. 75]
Rabbit (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[98, p. 75]
Chronic.			
Rat (2-year feed):	0.9	$mg\,kg^{-1}\,d^{-1}$	[98, p. 75]
Mouse (2-year feed):	1.82	$mg\,kg^{-1}\,d^{-1}$	[98, p. 75]
Dog (2-year capsule):	2	$mg\,kg^{-1}\,d^{-1}$	[98, p. 75]
ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [98, p. 75]			

Cymizaol [*61676-87-7*]

LD_{50} (rat, oral)	758	mg/kg	[20, p. 197]
LD_{50} (rat, dermal)	> 3100	mg/kg	[20, p. 197]
LC_{50} (rat, 4-h inhalation)	2.8	mg/L air	[20, p. 197]
Rabbit: slightly irritating to skin and eyes			[20, p. 197]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[20, p. 197]
Mouse (2-year feed):	14	$mg\,kg^{-1}\,d^{-1}$	[20, p. 197]

Chinomethionate [*2439-01-2*]

LD_{50} (rat, oral)	1800 – 4800	mg/kg	[131]
LD_{50} (guinea pig, oral)	1500	mg/kg	[20, p. 198]
LD_{50} (rat, dermal)	> 5000	mg/kg	[20, p. 198]
LC_{50} (rat, 4-h inhalation)	3.0	mg/L air	[20, p. 198]
Rabbit: slightly irritating to skin and eyes			[20, p. 198]
Nonmutagenic			[102, p. 27]
NOEL/NOAEL:			
Reproduction.			
Rat (one-generation feed):	ca. 1	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[102, p. 27]
Teratogenicity.			
Rat (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[102, p. 27]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[102, p. 27]
Chronic.			
Rat (2-year feed):	0.6	$mg\,kg^{-1}\,d^{-1}$	[102, p. 27]
Mouse (21-month feed):	16 – 21	$mg\,kg^{-1}\,d^{-1}$	[102, p. 27]
Dog (one-year feed):	0.6	$mg\,kg^{-1}\,d^{-1}$	[102, p. 27]
ADI: 0.006 $mg\,kg^{-1}\,d^{-1}$ [102, p. 27]			

Clofentezine [*74115-24-5*]

LD_{50} (rat, oral)	> 3200	mg/kg	[115, p. 17]
LD_{50} (mouse, oral)	> 3200	mg/kg	[115, p. 17]
LD_{50} (hamster, oral)	> 3200	mg/kg	[115, p. 17]
LD_{50} (dog, oral)	> 2000	mg/kg	[115, p. 17]
LD_{50} (guinea pig, oral)	> 1500	mg/kg	[115, p. 17]
LD_{50} (rat, dermal)	> 1332	mg/kg	[115, p. 17]
LC_{50} (rat, 6-h inhalation)	> 9.08	mg/L air	[115, p. 17]
Rabbit: mildly irritating to skin and eyes			[115, p. 17]
Guinea pig: not a skin sensitizer			[115, p. 17]
Nonmutagenic			[115, p. 17]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	ca. 2.7	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[115, p. 17]
Teratogenicity.			
Rat (gavage):	320	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[115, p. 17]
	1280	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[115, p. 17]
Rabbit (gavage):	250	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[115, p. 17]
	1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[115, p. 17]
Chronic.			
Rat (118-week feed):	2	$mg\,kg^{-1}\,d^{-1}$	[115, p. 17]
Mouse (105-week feed):	5	$mg\,kg^{-1}\,d^{-1}$	[115, p. 17]
Dog (52-week feed):	1.2	$mg\,kg^{-1}\,d^{-1}$	[115, p. 17]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [115, p. 17]

Endusolfan [*115-29-7*]

LD_{50} (rat, oral)	70 – 110	mg/kg	[20, p. 459]
LD_{50} (dog, oral)	77	mg/kg	[20, p. 459]
LD_{50} (hamster, oral)	118	mg/kg	[132]
LD_{50} (rat, dermal)	500	mg/kg (female)	[20, p. 459]
	> 4000	mg/kg (male)	[20, p. 459]
LD_{50} (rabbit, dermal)	359	mg/kg	[20, p. 459]
LD_{50} (guinea pig, dermal)	1000	mg/kg	[132]
LC_{50} (rat, 4-h inhalation)	0.0345 – 0.913	mg/L air (male)	[107, p. 95]
Guinea pig: not a skin sensitizer			[107, p. 95]
Nonmutagenic			[107, p. 95]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	1.0	$mg\,kg^{-1}\,d^{-1}$ (parental)	[107, p. 95]
	6.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 95]
Teratogenicity.			
Rat (gavage):	0.66	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[107, p. 95]
	2.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 95]
Rabbit (gavage):	0.7	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[107, p. 95]
	1.8	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[107, p. 95]
Chronic.			
Rat (2-year feed):	0.6 – 0.7	$mg\,kg^{-1}\,d^{-1}$	[107, p. 95]
Mouse (2-year feed):	∼ 0.9	$mg\,kg^{-1}\,d^{-1}$	[107, p. 95]
Dog (one-year feed):	∼ 0.6	$mg\,kg^{-1}\,d^{-1}$	[107, p. 95]

ADI: $0.006\,mg\,kg^{-1}\,d^{-1}$ [107, p. 95]

Dienochlor [*2227-17-0*]

LD_{50} (rat, oral)	> 5000	mg/kg	[20, p. 385]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 385]
LD_{50} (rabbit, dermal)	> 3160	mg/kg	[20, p. 385]
LD_{50} (rat, 4-h inhalation)	0.08	mg/L air	[20, p. 385]
Rabbit: mildly irritating to eyes and skin			[20, p. 385]
Guinea pig: not a skin sensitizer			[20, p. 385]

Fenazaquin [*120928-09-8*]

LD_{50} (rat, oral)	134 – 138	mg/kg	[20, p. 506]
LD_{50} (mouse, oral)	1480 – 2449	mg/kg	[20, p. 506]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[20, p. 506]
LC_{50} (rat, 4-h inhalation)	1.9	mg/L air	[20, p. 506]
Rabbit: slightly irritating to eyes			[20, p. 506]
Rabbit: nonirritating to skin			[20, p. 506]
Guinea pig: not a skin sensitizer			[20, p. 506]
Nonmutagenic			[20, p. 506]
Not teratogenic or carcinogenic			[20, p. 506]

ADI: $0.005\,mg\,kg^{-1}\,d^{-1}$ [20, p. 506]

Fenpyroximate [*111812-58-9*]

LD_{50} (rat, oral)	245 – 480	mg/kg	[133]
LD_{50} (mouse, oral)	440 – 520	mg/kg	[133]
LD_{50} (rat, dermal)	> 4000	mg/kg (female)	[134]
LC_{50} (rat, 4-h inhalation)	0.0002 – 0.0003	mg/L air	[133]
Rabbit: nonirritating to skin			[134]
Rabbit: slightly irritating to eyes			[134]
Nonmutagenic			[133]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	2.0	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[133]
Teratogenicity.			
Rat (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[133]
Rabbit (gavage):	2.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[133]
	> 5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[133]
Subchronic.			
Rat (13-week feed):	1.3	$mg\,kg^{-1}\,d^{-1}$	[133]
Dog (13-week feed):	2	$mg\,kg^{-1}\,d^{-1}$	[133]
Chronic.			
Rat (2-year feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[133]
Mouse (18-month feed):	9.5	$mg\,kg^{-1}\,d^{-1}$	[133]
Dog (one-year feed):	5.0	$mg\,kg^{-1}\,d^{-1}$	[133]
ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [133]			

Flucycloxuron [*94050-52-9*]

LD_{50} (rat, oral)	> 5000	mg/kg	[20, p. 562]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 562]
LC_{50} (rat, 4-h inhalation)	3.3	mg/L air	[20, p. 562]
Rabbit: nonirritating to skin			[20, p. 562]
Rabbit: mildly irritating to eyes			[20, p. 562]
Nonmutagenic			[20, p. 562]
Nonteratogenic			[20, p. 562]
Noncarcinogenic			[20, p. 562]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	ca. 10	$mg\,kg^{-1}\,d^{-1}$	[20, p. 562]
Chronic.			
Rat (2-year feed):	ca. 6	$mg\,kg^{-1}\,d^{-1}$	[20, p. 562]

Flufenoxuron [*101463-69-8*]

LD_{50} (rat, oral)	> 3000	mg/kg	[20, p. 568]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 568]
LD_{50} (mouse, dermal)	> 2000	mg/kg	[20, p. 568]
LC_{50} (rat, 4-h inhalation)	5	mg/L air	[20, p. 568]
Rabbit: nonirritating to skin and eyes			[20, p. 568]
NOEL/NOAEL:			
Subchronic.			
Rat (90-day feed):	ca. 3.3	$mg\,kg^{-1}\,d^{-1}$	[20, p. 568]
Mouse (90-day feed):	ca. 7.1	$mg\,kg^{-1}\,d^{-1}$	[20, p. 568]
Dog (one-year feed):	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$	[20, p. 568]

Hezythiazox [*78587-05-0*]

LD_{50} (rat, oral)	> 5000	mg/kg	[98, p. 265]
LD_{50} (mouse, oral)	> 5000	mg/kg	[98, p. 265]
LD_{50} (dog, oral)	> 5000	mg/kg	[98, p. 265]
LD_{50} (rat, dermal)	> 50000	mg/kg	[98, p. 265]
LC_{50} (rat, 4-h inhalation)	> 2.0	mg/L air	[98, p. 265]
Rabbit: nonirritating to skin			[98, p. 265]
Rabbit: slightly irritating to eyes			[98, p. 265]
Guinea pig: not a skin sensitizer			[98, p. 265]
Nonmutagenic			[98, p. 265]
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	4.22 – 5.34	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[98, p. 265]
Teratogenicity.			
Rat (gavage):	240	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[98, p. 265]
Rabbit (gavage):	360	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[98, p. 265]
Chronic.			
Rat (2-year feed):	3.2 – 4.02	$mg\,kg^{-1}\,d^{-1}$	[98, p. 265]
Mouse (2-year feed):	< 6.7	$mg\,kg^{-1}\,d^{-1}$	[98, p. 265]
Dog (one-year feed):	2.87 – 3.17	$mg\,kg^{-1}\,d^{-1}$	[98, p. 265]
ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [98, p. 265]			

Lufenuron [*103055-0-8*]

LD_{50} (rat, oral)	> 2000	mg/kg	[20, p. 753]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 753]
LC_{50} (rat, 4-h inhalation)	> 2.35	mg/L air	[20, p. 753]
Rabbit: nonirritating to skin and eyes			[20, p. 753]
Guinea pig: not a skin sensitizer			[20, p. 753]
Nonmutagenic			[20, p. 753]
NOEL/NOAEL: *Chronic.*			
Rat (2-year feed):	2.0	$mg\,kg^{-1}\,d^{-1}$	[20, p. 753]

Propargite [*2312-35-8*]

LD_{50} (rat, oral)	~ 2800	mg/kg	[130, p. 403]
LD_{50} (rabbit, oral)	311	mg/kg	[130, p. 403]
LD_{50} (rabbit, dermal)	4000	mg/kg	[130, p. 403]
LC_{50} (rat, 4-h inhalation)	0.89	mg/L air	[130, p. 403]
Rabbit: severely irritating to skin and eyes			[130, p. 403]
Guinea pig: not a skin sensitizer (guinea pig)			[130, p. 403]
NOEL/NOAEL: *Reproduction.*			
Rat (3-generation feed):	15	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[130, p. 403]
Teratogenicity.			
Rat (gavage):	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[130, p. 403]
Chronic.			
Rat (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[130, p. 403]
Dog (one-year feed):	22	$mg\,kg^{-1}\,d^{-1}$	[130, p. 403]

ADI: 0.15 $mg\,kg^{-1}\,d^{-1}$ [130, p. 403]

Pyridaben [*96489-71-3*]

LD_{50} (rat, oral)	820 – 1350	mg/kg	[20, p. 1061]
LD_{50} (mouse, oral)	383 – 424	mg/kg	[20, p. 1061]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 1061]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[20, p. 1061]
LC_{50} (rat, 4-h inhalation)	0.62 – 0.66	mg/L air	[20, p. 1061]
Rabbit: nonirritating to skin and eyes			[20, p. 1061]
Guinea pig: not a skin sensitizer (guinea pig)			[20, p. 1061]
Nonmutagenic			[20, p. 1061]
NOEL/NOAEL: *Chronic.*			
Rat (2-year feed):	1.4	$mg\,kg^{-1}\,d^{-1}$	[20, p. 1061]
Mouse (78-week feed):	0.81	$mg\,kg^{-1}\,d^{-1}$	[20, p. 1061]

Tebufenpyrad [*119168-77-3*]

LD_{50} (rat, oral)	595 – 997	mg/kg	[20, p. 1148]
LD_{50} (mouse, oral)	210 – 224	mg/kg	[20, p. 1148]
LD_{50} (rat, dermal)	> 2000	mg/kg	[20, p. 1148]
LC_{50} (rat, 4-h inhalation)	2.66 – 3.1	mg/L air	[20, p. 1148]
Rabbit: nonirritating to skin			[20, p. 1148]
Rabbit: slightly irritating to eyes			
Guinea pig: not a skin sensitizer (guinea pig)			[20, p. 1148]
Nonmutagenic			[20, p. 1148]

5. References

General References

1. R. Bovey, M. Bagglioni, A. Bolay, E. Bovay, R. Corbaz, G. Mathys, A. Meylan, R. Murbach, F. Pelet, A. Savary, G. Trivelli: *La Défense des Plantes Cultivées,* 7th ed., Payot, Lausanne 1979.
2. R. Gasser, *Mitt. Schweiz. Entomol. Ges.* **24** (1957) 217.
3. T. H. Hughes: *Mites or the Acari,* University of London, The Athlone Press, London 1959.
4. J. E. Laing, *Acarologia* **11** (1969) 32.
5. A. E. Pritchard, E. W. Baker: *A Revision of the Spider Mite, Family Tetranychidae,* Pacific Coast Entomological Society, San Francisco 1955.

Specific References

6. F. Chaboussou, *Z. Angew. Zool.* **53** (1966) 257.
7. F. Löcher, *Z. Angew. Zool.* **45** (1958) 201.
8. V. Dittrich, P. Streibert, P. A. Bathe, *Environ. Entomol.* **3** (1974) 534.
9. A. McMurtry, C. B. Huffaker, M. Van de Vrie, *Hilgardia* **40** (1970) 331.
10. M. Van de Vrie, J. A. McMurtry, C. B. Huffaker, *Hilgardia* **41** (1972) 343.
11. H. Grob, *Mitt. Schweiz. Entomol. Ges.* **24** (1951) 163.
12. W. Kolbe, *Erwerbsobstbau* **10** (1968) 41.

13. D. A. Chant, *Can. Entomol.* **93** (1961) 437.
14. C. B. Huffaker, M. Van de Vrie, J. A. McMurtry, *Annu. Rev. Entomol.* **14** (1969) 125.
15. U. Gerson, M. Van de Vrie, *Proc. 4th Int. Congr. Acarology* 1974, 629 – 635.
16. C. R. Worthing: *The Pesticide Manual,* 6th ed., The British Crop Protection Council, Croydon, U.K., 1979.
17. *Ullmann,* 4th ed., **7,** 1.
18. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 1,** Springer, Berlin-Heidelberg-New York 1970.
19. C. D. S. Tomlin: *The Pesticide Manual,* 10th ed., British Crop Protection Council, Croydon 1994.
20. C. D. S. Tomlin: *The Pesticide Manual,* 11th ed., British Crop Protection Council, 1997.
21. Union Carbide, FR 1 377 474, 1965 (L. K. Payne, M. H. J. Weiden).
22. Shell, US 3 116 201, 1960 (R. R. Whestone, D. Harman).
23. Dow, FR 1 360 901, 1964 (R. H. Rigterink).
24. Novartis, DE 910 652, (H. Gysin, A. Margot).
25. Shell, US 2 956 073, 1960 (R. R. Whestone, D. Harman).
26. Cyanamid, US 2 494 283, 1948 (J. T. Cassady, E. I. Hoegberg).
27. Bayer, DE 917 668, 1952 (W. Lorenz).
28. Du Pont, US 2 503 390, 1952 (A. G. Jenkins).
29. Food Machenery and Chem. Corp., US 2 873 228, 1959 (J. L. Willard, J. F. Henahan).
30. H. Ogawa, *Jpn. Pest. Info* **1985,** 11.
31. Novartis, DE 1 148 806, 1960 (K. Lutz, M. Schuler).
32. M. Pinda, *Chem. and Ind.* **1961,** 324.
33. Novartis, DE 2 119 488, 1971 (E. Beriger).
34. Bayer, NL 6 508 556, 1966.
35. K. Rüfenach (Novartis), *Helv. Chim. Acta* (1968) 518.
36. Bayer, US 912 895, 1962 (E. Schegk, G. Schrader, K. Wedemeyer).
37. Du Pont, US 3 576 834, 1971 (J. B. Buchanan).
38. Shell, US 2 685 552, 1952 (a. R. Stiles).
39. Novartis, US 2 908 605, 1959 (E. Beriger, R. Sallmann).
40. Bayer, DE 1 251 304, 1962 (G. Oertl, A. Dörken).
41. Du Pont, US 3 530 220, 1968 (J. B. Buchanan).
42. Bayer, DE 1 011 416, 1955 (G. Schrader).
43. Cyanamid, US 2 586 655, 1948 (E. O. Hook, Ph. H. Moss).
44. Rhône-Poulenc, FR 1 482 025, 1966 (J. Metivier).
45. Stauffer, US 2 767 194, 1956 (L. W. Fancher).
46. Novartis, BE 552 284, 1955 (E. Beriger, R. Sallmann).
47. Zeneca, GB 1 019 227, 1966 (G. V. McHattie).
48. Novartis, DE 2 249 462, 1973 (E. Beriger, J. Drabek).
49. Novartis, NL 6 611 511, 1967.
50. Bayer, DE 848 812, 1950 (G. Schrader, R. Mühlmann).
51. Novartis, CH 319 579, 1953 (K. Lutz, O. Jucker).
52. AgrEvo, S. Afr. 68 08 471, 1968 (O. Scherer, H. Mildenberger).
53. Rhône-Poulenc, BE 575 106, 1961 (J. Metivier).
54. H. Riedl, S. S. Hoying, *J. Econ. Entomol.* **73** (1980) 117.
55. Roussel-UCLF, FR 2 486 073, 1982 (J. Martel, J. Tessier, A. Teche).
56. FMC, GB 2 085 005, 1982 (J. F. Engel).
57. Sumitomo, DE 2 231 312, 1973 (T. Matsuo et al.).
58. Sumitomo, DE 2 335 347, 1974.
59. Mitsui Zoatsu, DE 3 708 231, 1987 (S. Numaka et al.).
60. I. R. Harrison et al., *Pestic. Sci.* **3** (1972) 679; **4** (1973) 901.
61. I. R. Harrison, A. Kozlik, J. F. McCarthy, B. H. Palmer, D. M. Weighton, *Meded. Fac. Landbouwwet Rijksuniv. Gent.* **37** (1972) 765.
62. AgrEvo, DE 1 169 194, 1962 (H. Peissker, A. Jäger, W. Steinhausen, G. Boroschewski).
63. Murphy Chem., GB 1 019 451, 1961 (M. Pianka, D. J. Polton).
64. Rohm & Haas Co., US 2 526 660, 1946, (W. F. Hester, W. E. Craig).
65. R. D. Clifford, D. A. M. Watkins, D. Woodcock, *Chem. Ind. (London)* 1965, 1654.
66. U.S. Dept. Agriculture, Agricultural Research Division, Method 580-0.
67. N. V. Philips' Gloeilampenfabrieken, DE 1 023 761, 1958 (J. Meltzer, H. O. Huisman).
68. Agripat, CH 471 065, 1965 (K. Gubler).
69. S. Reuter, K. R. S. Ascher, *Experientia* **12** (1956) 316.
70. Rohm & Haas Co., US 2 812 280, 1954 (H. F. Wilson, E. Y. Shropshire, W. E. Craig).
71. I. Hamman, K. H. Büchel, K. Bungarz, L. Born, *Pflanzenschutz Nachr.* **31** (1978) 60.
72. E. Krause, R. Pohland, *Ber. Dtsch. Chem. Ges.* **57** (1924) 532.
73. W. E. Allison, A. E. Doty, J. L. Hardy, E. E. Kenaga, W. K. Whitney, *J. Econ. Entomol.* **61** (1968) 1254.

74. H. Zimmer, I. A. Hornberg, M. Jayawant, *J. Org. Chem.* **31** (1966) 3857.
75. J. Putter et al., *Exoerientia,* **37** (1981) 963.
76. K. K. Siccama, R. Heynderickx, M. Asada En, H. F. A. de Lange, *Meded. Fac. Landbouwwet. Rijksuniv. Gent.* **42** (1977) 1479.
77. Nihon Noyaku, DE 2 824 126, 1978 (K. Ikeda, H. Kanno, M. Yasui, T. Harada).
78. Novartis, DE 2 619 724, 1976 (D. Duerr, W. D. Traber).
79. Bayer, DE 1 088 965, 1958 (K. Sasse, R. Wegler, G. Unterstenhöfer).
80. K. Sasse, R. Wegler, G. Unterstenhöfer, F. Grewe, *Angew. Chem.* **72** (1960) 973.
81. Fisons, EP 5 912, 1979 (J. H. Persens).
82. Hooker Chemical Corp., US 2 908 723, 1959 (J. T. Rucker).
83. Hooker Chemical Corp., US 2 934 470, 1960 (J. T. Rucker).
84. AgrEvo, DE 1 015 797, 1954.
85. Lilly Eli, EP 380 264, 1950 (W. I. Glaisby).
86. Nihon Noyaku, EP 234 945, 1987 (H. R. M. F. Hamaguchi et al.).
87. Duphar, EP 117 320, 1984 (M. S. Brouwer, A. C. Grosscourt).
88. Shell, US 161 019, 1985 (M. Anderson).
89. Nippon, DE 3 037 105, 1981 (I. Iwakaki, M. Kaeriyama, N. Matsui).
90. Novartis, EP 179 022, 1986 .(J. Drabek, M. Boeger).
91. U.S. Rubber Co., US 3 272 854, 1963 (R. A. Covey, A. E. Smith, W. L. Hubbard).
92. Nissan, JP 62 67 076, 1987 (H. Suzuki, M. Kawashima, Z. Ogura).
93. Mitsubishi, EP 289 879, 1988 (I. Okada, S. Okui, Y. Takahashi, T. Fukuchi).
94. C. D. Klassen, M. O. Amdur, J. Doull (eds.): *Casarret and Doull's Toxicology,* 3rd ed., Macmillan Publ. Co., New York 1986 p. 556.
95. R. L. Baron, T. L. Merriam, *revs. Environm. Contaim. Toxicol.* **105** (1988) 1.
96. WHO/PCS/93.34 (1993).
97. WHO/PCS/95.2 (1995).
98. WHO/PCS/92.52 (1992).
99. WHO/PCS/94.4 (1994).
100. W. J. Hayes, E. R. Laws (eds.): *Handbook of Pesticide Toxicology,* Academic Press, New York 1991.
101. "Dimethoate", *WHO Environmental Health Criteria* **90** (1989).
102. Pesticide Residues in Food-1987, *FAO Plant Prod. Prot. Paper* **86/2** (1988).
103. WHO/PCS/92.52 (1992) 151.
104. WHO/PCS/91.47 (1991).
105. Pesticide Residues in Food-1986, *FAO Plant Prod. Prot. Paper* **78/2** (1987).
106. Pesticide Residues in Food-1980, *FAO Plant Prod. Prot. Paper* **26 Sup.** (1981).
107. Pesticide Residues in Food-1987, *FAO Plant Prod. Prot. Paper* **84** (1987).
108. *1972 Evaluations of some Pesticide Residues in Food,* WHO Pesticide Residues Series, No. **2** (1973).
109. *1975 Evaluations of some Pesticide Residues in Food,* WHO Pesticide Residues Series, No. **5** (1976).
110. Pesticide Residues in Food-1985, *FAO Plant Prod. Prot. Paper* **72/2** (1986).
111. G. L. Kennedy, *Fundam. Appl. Toxicol.* **6** (1986) 423.
112. Pesticide Residues in Food-1984, *FAO Plant Prod. Prot. Paper* **62** (1985).
113. WHO/PCS/95.2 (1995).
114. Regulatory Affairs Department, SDS Biotech K. K., *J. Pesticide Sci.* **16** (1991) 337.
115. *1973 Evaluations of some Pesticide Residues in Food,* WHO Pesticide Residues Series, No. **3** (1973).
116. Industrieverband Agrar e.V. Frankfurt (M.): "Physikalisch-chemische und toxikologische Daten",*Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmitteln,* 2nd ed., Frankfurt (Main) 1990, p. 50.
117. Pesticide Residues in Food-1988, *FAO Plant Prod. Prot. Paper* **93/2** (1989).
118. "Fenvalerate", *WHO Environmental Health Criteria* **95** (1990).
119. Pesticide Residues in Food-1979, *FAO Plant Prod. Prot. Paper* **20 Sup.** (1980).
120. C. M. Parker, D. R. Patterson, G. A. van Gelder, F. B. Gordon, M. G. Valerio, W. C. Hall, *J. Toxicol. Environm. Health* **13** (1984) 83.
121. Y. Okuno, S. Ito, T. Seki, T. Hiromori, M. Murakami, T. Kadota, J. Miyamoto, *J. Toxicol. Sci.* **11** (1986) 53.
122. Pesticide Residues in Food-1986, *FAO Plant Prod. Prot. Paper* **77** (1986).
123. Pesticide Residues in Food-1990, *FAO Plant Prod. Prot. Paper* **102** (1990).
124. T. B. Gaines, *Toxicol. Appl. Pharmacol.* **14** (1969) 515.
125. NIOSH Registry: Tracor Jitco. U.S. Government Printing Office, Washington, DC, 1982.
126. R. T. Meister (ed.): *Farm Chemical Handbook,* Meister Publishing Co., Willoughby, Ohio, 1980 – 1982.
127. H. Martin, C. R. Worthing (eds.): *Pesticide Manual,* 5th ed., British Crop Protection Council, Croydon1977.

128. Pesticide Residues in Food-1991, *FAO Plant Prod. Prot. Paper* **111** (1991).
129. Pesticide Residues in Food-1989, *FAO Plant Prod. Prot. Paper* **100/2** (1990).
130. Pesticide Residues in Food-1977, *FAO Plant Prod. Prot. Paper* **10 Sup.** (1978).
131. *1974 Evaluations of some Pesticide Residues in Food,* WHO Pesticides Residues Series No. **4** (1975).
132. "Endosulfan", *WHO Environmental Health Criteria* **40** (1984).
133. Pesticide Residues in Food-1995, *FAO Plant Prod. Prot. Paper* **133** (1996) 86.
134. T. Tsuda, J. Tamura, H. Ueda, I. Ichimoto, *J. Pesticide Sci.* **17** (1992) 261.

Accumulators → Batteries
Acenaphthene → Hydrocarbons

Acetaldehyde

Marc Eckert, Wacker Chemie AG, Werk Burghausen, Germany

Gerald Fleischmann, Wacker Chemie AG, Werk Burghausen, Germany (Chap. 1, 2, 3, 4, 5, 6, 7 and 8)

Reinhard Jira, Wacker Chemie AG, Werk Burghausen, Germany (Chap. 1, 2, 3, 4, 5, 6, 7 and 8)

Hermann M. Bolt, Institut für Arbeitsphysiologie an der Universität Dortmund, Dortmund, Germany (Chap. 9)

Klaus Golka, Institut für Arbeitsphysiologie an der Universität Dortmund, Dortmund, Germany (Chap. 9)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties and Uses 3
3.1. Addition Reactions 3
3.2. Derivatives of Aldol Addition 3
3.3. Reaction with Nitrogen Compounds 3
3.4. Oxidation 4
3.5. Reduction 4
3.6. Miscellaneous Reactions 4
3.7. Consumption 4
4. Production 4
4.1. Production from Ethanol 5
4.2. Production from Acetylene 6
4.3. Production from Ethylene 7
4.3.1. Direct Oxidation of Ethylene 7
4.3.2. Acetaldehyde as Byproduct 10
4.3.3. Isomerization of Ethylene Oxide . . . 10
4.4. Production from C_1 Sources 11
4.5. Production from Hydrocarbons . . 11
5. Quality and Analysis 11
6. Storage and Transportation 11
6.1. Storage 11
6.2. Transportation 12
6.3. Other Regulations 12
7. Economic Aspects 12
8. Polymers of Acetaldehyde 13
8.1. Paraldehyde 13
8.2. Metaldehyde 13
8.3. Polyacetaldehyde 14
9. Toxicology and Occupational Health 14
10. References 15

1. Introduction

Acetaldehyde (ethanal), CH_3CHO [*75-07-0*], was observed in 1774 by Scheele during reaction of black manganese dioxide and sulfuric acid with alcohol. Its constitution was explained in 1835 by Liebig who prepared pure acetaldehyde by oxidation of ethanol with chromic acid and designated this product "aldehyde," a contraction of the term "alcohol dehydrogenatus."

Acetaldehyde is a mobile, low-boiling, highly flammable liquid with a pungent odor. Because of its high chemical reactivity, acetaldehyde is an important intermediate in the production of acetic acid, acetic anhydride, ethyl acetate, peracetic acid, butanol, 2-ethylhexanol, pentaerythritol, chlorinated acetaldehydes (chloral), glyoxal, alkyl amines, pyridines, and other chemicals. The first commercial application was the production of acetone via acetic acid between 1914 and 1918 in Germany (Wacker-Chemie and Hoechst) and in Canada (Shawinigan).

Acetaldehyde is an intermediate in the metabolism of plant and animal organisms, in which it can be detected in small amounts. Larger amounts of acetaldehyde interfere with biological processes. As an intermediate in alcoholic fermentation processes it is present in small amounts in all alcoholic beverages, such as beer, wine, and spirits. Acetaldehyde also has been detected in plant juices and essential oils, roasted coffee, and tobacco smoke.

Commercial production processes include dehydrogenation or oxidation of ethanol, addition of water to acetylene, partial oxidation of hydrocarbons, and direct oxidation of ethylene. In the 1970s, the world capacity of this last process, the Wacker-Hoechst direct oxidation, increased to over 2×10^6 t/a. However, the impor-

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_031.pub2

tance of acetaldehyde as an organic intermediate is now steadily decreasing, because new processes for some acetaldehyde derivatives have been developed, such as the oxo process for butanol and 2-ethylhexanol and the Monsanto process for acetic acid. In the future, new processes for acetic anhydride (Halcon, Eastman, Hoechst), for vinyl acetate (Halcon), and for alkyl amines (from ethanol) will diminish the use of acetaldehyde as a starting material.

2. Physical Properties

Acetaldehyde, C_2H_4O, M_r 44.054, is a colorless liquid with a pungent, suffocating odor that is slightly fruity when diluted.

bp at 101.3 kPa	20.16 °C
mp	–123.5 °C
Critical temperature t_{crit}	181.5 °C
other values	187.8 °C [12], 195.7 °C [13]
Critical pressure p_{crit}	6.44 MPa
other values	5.54 MPa [12], 7.19 MPa [13]
Relative density	$d_4^t = 0.8045 - 0.001325 \cdot t$ (t in °C) [3]
Refractive index	$n_D^t = 1.34240 - 0.0005635 \cdot t$ (t in °C) [14]
Molar volume of the gas	
at 101.3 kPa and 20.16 °C	23.40 L/mol
at 25.0 °C	23.84 L/mol

For dependence on T (293.32 – 800 K) and p (0.1 – 30 MPa), see [15].

Specific volume of the vapor	
at 20.16 °C	0.531 m^3/kg
at 25.0 °C	0.541 m^3/kg
Vapor density (air = 1)	1.52

Vapor pressure

t, °C	−20	−0.27	5.17	14.76	50	100
p, kPa	16.4	43.3	67.6	82.0	279.4	1014.0

For further values between −60 and +180 °C, see [14]

Viscosity of liquid η	
at 9.5 °C	0.253 mPa · s
at 20 °C	0.21 mPa · s
Viscosity of vapor η	
at 25 °C	86×10^{-4} mPa · s

For further values between 35.0 and 77.8 °C and between 0.13 and 0.40 kPa, see [16].

Surface tension γ at 20 °C	21.2×10^{-2} mN cm^{-1}
Dipole moment (gas phase)	2.69 ± 2 % D [12]
Dielectric constant	
of liquid at 10 °C	21.8
of vapor at 20.16 °C, 101.3 kPa	1.0216
Heat capacity of liquid	
c_p (l) at 0 °C	2.18 J g^{-1} K^{-1}
at 20 °C	1.38 J g^{-1} K^{-1}

For further values between −80 °C (c_p = 1.24 J g^{-1} K^{-1}) and +120 °C (c_p = 1.50 J g^{-1} K^{-1}), see [17].

Heat capacity of vapor	
c_p (g) at 25 °C, 101.3 kPa	1.24 J g^{-1} K^{-1}

For dependence on temperature (nonlinear) between 0 °C (c_p = 1.17 J g^{-1} K^{-1}) and 1000 °C (c_p = 2.64 J g^{-1} K^{-1}), see [17].

c_p/c_v (= $\varkappa$) at 30 °C, 101.3 kPa	1.145 [18]
Thermal conductivity	
of liquid at 20 °C	0.174 J m^{-1} s^{-1} K^{-1}; for more values, see [19]
of vapor at 25 °C	1.09×10^{-2} J m^{-1} s^{-1} K^{-1} for further values, see [20].
Cubic expansion coefficient per K (0 – 20 °C)	0.00169
Heat of combustion of liquid at constant p	1168.79 (1166.4 [12]) kJ/mol
Heat of solution in water (infinite dilution)	17 906 J/mol
Latent heat of fusion	3246.3 J/mol
Latent heat of vaporization at 20.2 °C	25.73 kJ/mol
other values	27.2 [21], 30.41, 27.71 [12], 26.11 [22] kJ/mol

For dependence on temperature (nonlinear) between -80 °C (32.46 kJ/mol) and 182 °C (0 kJ/mol), see [17].

Heat of formation ΔH from the elements at 25 °C for gaseous acetaldehyde −166.47 (−166.4 [21]) kJ/mol

For dependence of heat of formation for gaseous and liquid acetaldehyde, and enthalpy of vaporization on temperature up to 800 K and 30 MPa, see [15].

Gibbs free energy of formation ΔG from elements	
at 25 °C for gaseous acetaldehyde	−133.81 kJ/mol
other values	133.72 [12], 132.9 [21] kJ/mol
Entropy for gaseous acetaldehyde	
at 25 °C	265.9 J mol^{-1} K^{-1}
Entropy for liquid acetaldehyde	
at 20.16 °C	172.9 J mol^{-1} K^{-1}
Entropy of vaporization	
at 20.16 °C	91.57 J mol^{-1} K^{-1}
First ionization potential	10.5 eV
Dissociation constant at 0 °C	0.7×10^{-14} mol/L ($H_3CCHO^- H_2CCHO + H^+$)

For the second virial coefficient of the equation of state for gaseous acetaldehyde at 31 °C, 66 °C, and 85 °C, see [23].

Acetaldehyde is completely miscible with water and most organic solvents. It forms no azeotrope with water, methanol, ethanol, acetone, acetic acid, or benzene. Binary azeotropes are formed with butane (*bp* −7 °C, 84 wt % of butane) and diethyl ether (*bp* 18.9 °C, 23.5 wt % of ether).

Other Physical Data. Compressibility and viscosity at higher pressure are given in [24], vapor pressure of aqueous acetaldehyde solutions in [25]. For solubility of carbon dioxide, acetylene, and nitrogen in acetaldehyde, see [11]; for freezing points of aqueous acetaldehyde solutions, see [11]; for vapor – liquid equilibria of binary systems of acetaldehyde with water, ethanol, acetic acid, and ethylene oxide, see [26, pp. 392, 561, 565, and 570], with vinyl acetate, see [27].

Safety Data. Flash point (Abel – Pensky; DIN 51 755; ASTM 56 – 70) −20 °C (−40 °C according to the safety regulations of the Berufsgenossenschaft der Chemischen Industrie, Federal Republic of Germany). Ignition temperature (DIN 51 794; ASTM D 2155 – 66) 140 °C; for ignition retardation when injected into a hot air stream, see [28]. Explosive limits in air: 4 – 57 vol %; for influence of pressure on explosive limits, see [29].

3. Chemical Properties and Uses

Acetaldehyde is a highly reactive compound showing all of the typical aldehyde reactions as well as those of an alkyl group in which hydrogen atoms are activated by the carbonyl group in the α position. When heated above 420 °C acetaldehyde decomposes into methane and carbon monoxide.

3.1. Addition Reactions

With water, acetaldehyde forms an unstable hydrate; isolable solid hydrates are known only with chlorinated acetaldehydes. *Alcohols* add to acetaldehyde giving hemiacetals, which form acetals (→ Aldehydes, Aliphatic and Araliphatic) with additional alcohol in the presence of acids by removal of water. Diols give cyclic acetals; for example, 2-methyl-1,3-dioxolane is obtained from ethylene glycol and acetaldehyde, and 2-methyl-1,3-dioxane from 1,3-propanediol.

Aqueous *sodium bisulfite* solution and acetaldehyde give a crystalline adduct from which acetaldehyde can be liberated. Dry *ammonia* forms crystalline acetaldehyde ammonia. Acetaldehyde and *hydrocyanic acid* react to give lactonitrile (α-hydroxypropionitrile), a possible intermediate in acrylonitrile production [30].

Acetaldehyde reacts with *acetic anhydride* to give ethylidene diacetate, an intermediate in the vinyl acetate process of Celanese Corp. [31] (→ Vinyl Esters).

3.2. Derivatives of Aldol Addition

Two molecules of acetaldehyde combine in the presence of alkaline catalysts or dilute acids at room temperature or with moderate heating to form acetaldol [*107-89-1*], $CH_3CH(OH)CH_2CHO$. At increased temperatures, water is cleaved easily from this acetaldol, forming crotonaldehyde (→ Aldehydes, Aliphatic and Araliphatic). Further condensation under more stringent conditions to form aldehyde resins (e.g., synthetic shellac) now has no industrial importance.

Urea and acetaldehyde condense in the presence of H_2SO_4 to form crotonylidenediurea (6-methyl-4-ureidohexahydropyrimidin-2-one [*1129-42-6*]), which is used as a long-term nitrogen fertilizer (→ Fertilizers).

Acetaldehyde is also an intermediate in the butadiene synthesis starting from acetylene and proceeding via acetaldol and its hydrogenation product, 1,3-butanediol [32]. This process was introduced around 1918 and is still carried out on a commercial scale in some Eastern European countries.

Acrolein is obtained by aldol condensation of acetaldehyde and formaldehyde and subsequent water elimination, analogous to the formation of crotonaldehyde. This method is also without commercial importance today, whereas the production of pentaerythritol from acetaldehyde and a fourfold amount of formaldehyde in the presence of $Ca(OH)_2$ or NaOH is very important industrially (→ Alcohols, Polyhydric).

3.3. Reaction with Nitrogen Compounds

With primary amines, Schiff bases, $CH_3CH{=}NR$, are formed. Nitrogen compounds such as hydroxylamine, hydrazine, phenylhydrazine, and semicarbazide react with acetaldehyde to give easily crystallizable compounds

that are used for the analytical determination and characterization of aldehydes (semicarbazone, *mp* 162 – 163 °C; *p*-nitrophenylhydrazone, *mp* 128.5 °C; 2,4-dinitrophenylhydrazone, *mp* 168 °C; oxime, *mp* 47 °C). Many other aldehydes and ketones can be characterized in the same way because their analogous derivatives generally have sharp melting points.

The synthesis of pyridine and pyridine derivatives is of increasing importance. 5-Ethyl-2-methylpyridine is obtained in the presence of fluoride ions by the reaction of aqueous ammonia with acetaldehyde (or with paraldehyde, which slowly releases the monomer). In the added presence of formaldehyde or acrolein, mixtures of pyridine and alkylpyridines form (→ Pyridine and Pyridine Derivatives).

3.4. Oxidation

The major part of the acetaldehyde produced commercially is used for manufacturing acetic acid by oxidation with oxygen or air (→ Acetic Acid). Acetaldehyde monoperacetate is formed as an intermediate and decomposes into peracetic acid and acetaldehyde at elevated temperatures and in the presence of catalytic amounts of iron or cobalt salts. In the presence of Mn^{2+} salts, acetic acid is obtained from acetaldehyde monoperacetate, and in the presence of Co^{2+} and Cu^{2+} salts, acetic anhydride can be formed.

Oxidation with nitric acid gives glyoxal (→Glyoxal and → Glyoxylic Acid). Halogenated acetaldehydes are prepared by halogenation.

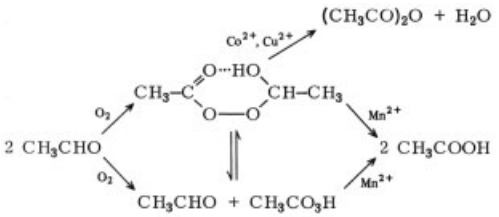

Mono-, di-, and trichloroacetaldehydes (→ Chloroacetaldehydes) and tribromoacetaldehyde (bromal) are useful for producing insecticides (e.g., DDT, DDD), pharmaceuticals, and dyes.

3.5. Reduction

Acetaldehyde is hydrogenated readily to ethanol. Prior to 1939, that is, before petrochemically produced ethylene became available in Europe, this reaction was used industrially to produce ethanol from acetaldehyde and, therefore, from acetylene.

Mono-, di-, and triethylamine [*75-04-7*], [*109-89-7*], [*121-44-8*] can be produced from acetaldehyde, ammonia, and hydrogen in the presence of a hydrogenation catalyst [33] (→ Amines, Aliphatic).

3.6. Miscellaneous Reactions

The Tishchenko reaction of acetaldehyde gives the commercially important solvent ethyl acetate (→ Acetic Acid); it is catalyzed by aluminum alcoholate.

As a "radical trapping agent," acetaldehyde is used to control chain length in the polymerization of vinyl compounds.

Oligomers of acetaldehyde are treated in Chapter 8.

3.7. Consumption

The consumption of acetaldehyde has changed during the last few years. Since 1993 in the USA, acetaldehyde is no longer used for the production of acetic acid, butanol, or 2-ethylhexanol, which are now produced by other routes (Table 1). The consumption of acetaldehyde for some other chemicals like peracetic acid or pyridine bases is increasing.

4. Production

Raw materials that have been used for the production of acetaldehyde are:

1) Ethanol from fermentation of carbohydrates or from hydration of ethylene
2) Acetylene
3) Ethylene
4) Lower hydrocarbons
5) Carbon monoxide and hydrogen
6) Methanol

Table 1. Consumption of acetaldehyde (10^3 t) in 2003 [56]

Product	USA	Mexico	W. Europe	Japan	Total
Acetic acid/acetic anhydride		11	89	47	147
Acetate esters	35	8	54	224	321
Pentaerythritol	26		43	11	80
Pyridine and pyridine bases	73		10	*	83
Peracetic acid	23			*	23
1,3-Butylene glycol	14			*	14
Others	5	3	10	80	98
Total	176	22	206	362	766

* Included in others (glyoxal/glyoxalic acid, crotonaldehyde, lactic acid, *n*-butanol, 2-ethylhexanol).

The economy of the commercial processes depends essentially upon prices and the availability of raw materials. In highly industrialized countries maintaining high prices for ethanol by fiscal measures or where petrochemical ethanol was not available, as in Germany or Japan before 1939, acetylene was the favored starting material for acetaldehyde. The acetylene process is still operated in some Eastern European countries and also by companies where cheap acetylene is available. Petrochemically produced ethanol, however, was the favored raw material in the other countries, whereas ethanol made by fermentation was and still is used on a small scale in countries with less chemical industry.

In Western countries, including Japan, all these processes have now been almost completely replaced by the direct oxidation process developed in the late 1950s by Wacker-Chemie and Hoechst. This is because ethylene is available at a lower price than acetylene.

Even the two-stage processes using ethanol from ethylene as starting material are no longer competitive because of the decreasing importance of acetaldehyde as an organic intermediate (see Chap. 7).

Generally, all processes based on acetylene, ethylene, and ethanol are more selective than the oxidation of saturated hydrocarbons. This is because, in the latter case, other oxidation products are formed in addition to acetaldehyde. Because of the great expense of separating the product mixture, such processes are economical only in large units and when all main and secondary products obtained in the process are utilized.

4.1. Production from Ethanol

For the production of acetaldehyde, ethanol can either be dehydrogenated or oxidized in the presence of oxygen. Between 1918 and 1939, dehydrogenation took precedence over oxidation because of the simultaneous production of hydrogen. Later, however, the catalytic vapor-phase oxidation of ethanol became the preferred process, probably because of the long catalyst life and the possibility of recovering energy.

Dehydrogenation of Ethanol. In the first work on ethanol dehydrogenation, published in 1886, ethanol was passed through glass tubes at 260 °C.

$$CH_3CH_2OH\ (l) \rightarrow CH_3CHO\ (l) + H_2\ (g)$$
$$\Delta H = +82.5\ \text{kJ/mol}$$

Improved yields are obtained in the presence of catalysts such as platinum, copper, or oxides of zinc, nickel, or cobalt. In later patents, zinc and chromium catalysts [34], oxides of rare earth metals [35], and mixtures of copper and chromium oxides [36] have been reported. The lowest amounts of decomposition products are obtained using copper catalysts. Frequent regeneration of the catalysts is required, however.

Process Description. Ethanol vapor is passed at 260 – 290 °C over a catalyst consisting of copper sponge or copper activated with chromium oxide in a tubular reactor [37]. A conversion of 25 – 50 % per run is obtained. By washing with alcohol and water, acetaldehyde and ethanol are separated from the exhaust gas, which is mainly hydrogen. Pure acetaldehyde is obtained by distillation; the ethanol is separated from water and higher-boiling products by

distillation and flows back to the reactor. The final acetaldehyde yield is ca. 90 %. Byproducts include butyric acid, crotonaldehyde, and ethyl acetate.

Oxidation of Ethanol. Oxidation of ethanol is the oldest and the best laboratory method for preparing acetaldehyde. In the commercial process, ethanol is oxidized catalytically with oxygen (or air) in the vapor phase.

$$CH_3CH_2OH\,(g) + 1/2O_2\,(g) \rightarrow CH_3CHO\,(l) + H_2O\,(l) \qquad \Delta H = -242.0\,\mathrm{kJ/mol}$$

Copper, silver, and their oxides or alloys are the most frequently used catalysts [38].

For an example of a simultaneous oxidation – dehydrogenation process, see [39].

Veba-Chemie Process (Fig. 1). Ethanol is mixed with air and passed over a silver catalyst at 500 – 650 °C (c). The temperature depends on the ratio of alcohol to air and the flow rate of the gas through the catalyst. Alcohol conversion varies between 50 and 70 % and the yield is between 97 and 99 % depending on the reaction conditions. Acetaldehyde and unconverted alcohol are removed from the waste gas by washing with cold alcohol (e) and separated by fractional distillation (h); after concentration the alcohol returns to the reactor. Heat formed in the reaction is utilized for steam production using a waste-heat recovery system immediately after the reaction zone.

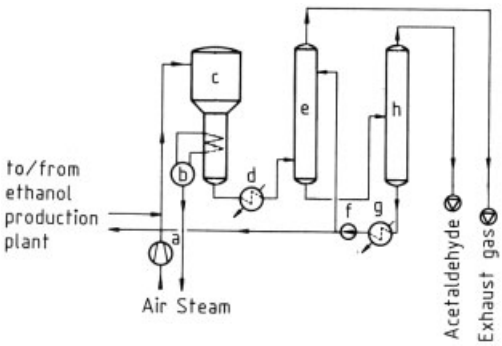

Figure 1. Acetaldehyde production by the Veba-Chemie process
a) Air compressor; b) Heat recovery system; c) Reactor; d) Cooler; e) Waste-gas scrubber; f) Washing-alcohol and return pump; g) Cooler; h) Acetaldehyde rectification

The waste gas consists mainly of nitrogen, hydrogen, methane, carbon monoxide and carbon dioxide; it is burned as lean gas with low calorific value in steam generators. Small amounts of acetic acid are obtained as a byproduct.

4.2. Production from Acetylene

The most important catalysts for the industrial water addition (hydration) are mercury compounds:

$$C_2H_2\,(g) + H_2O\,(l) \xrightarrow[H_2SO_4]{Hg^{2+}} CH_3CHO\,(l)\ \Delta H = -138.2\,\mathrm{kJ/mol}$$

This method only succeeds industrially when the polymerization and condensation products of acetaldehyde formed in the acid medium are eliminated. To achieve this, the Consortium für elektrochemische Industrie in 1912 proposed a process using excess acetylene at an elevated temperature and removing the acetaldehyde product immediately from the reaction liquid. At the same time, the heat of reaction is removed by distilling an appropriate amount of water. Secondary reactions, such as the oxidation of acetaldehyde to form acetic acid and carbon dioxide, result in reduction of Hg^{2+} to metallic mercury. In Western countries, acetaldehyde production from acetylene has now been discontinued.

Wet Oxidation Process (Hoechst). The wet oxidation process avoided direct handling of the toxic mercury compounds. It was operated, among others, by Wacker-Chemie until the changeover to ethylene as the starting material in 1962.

In this method, iron(III) sulfate is added to reoxidize the mercury metal to the mercury(II) salt, thus ensuring sufficient concentrations of active catalyst. The acetylene reacts at 90 – 95 °C with the aqueous catalyst solution; between 30 and 50 % of the injected acetylene reacts in one run. The gas emerging from the reactor is cooled; mainly water and traces of mercury are separated and returned to the reactor. Acetaldehyde and water are condensed in additional coolers and the acetaldehyde finally is washed

out with water from the cycle gas which has been cooled to 25 – 30 °C. An 8 – 10 % aqueous acetaldehyde solution is obtained. Nitrogen is introduced with the feed gas while carbon dioxide is formed as a byproduct; to avoid excessive accumulation, these gases are removed by withdrawing a small stream of the cycle gas. Iron(II) sulfate is formed in the reaction and is oxidized in a separate reactor with 30 % nitric acid at 95 °C. Pure acetaldehyde is obtained by fractional distillation of the aqueous solution at about 200 kPa. For further details of this process, see [40].

Chisso Process [41]**.** The Chisso process also uses sulfuric acid/mercury sulfate solution as a catalyst. The acetylene reacts completely with the catalyst solution at 68 – 78 °C and a gauge pressure of 140 kPa. A combination of pressure and vacuum process stages at low temperature and without excess acetylene is used; pure acetaldehyde can be isolated and distilled by utilizing the heat of the reaction. As in the Hoechst process, the catalyst can be regenerated with nitric acid. Production of acetaldehyde by this method was discontinued at Chisso Corp. more than a decade ago.

Production via Vinyl Ether. Reppe at BASF developed the process using vinyl ether [40]; it was operated in a pilot plant between 1939 and 1945. The use of toxic mercury compounds is avoided altogether. Methanol is added to acetylene at 150 – 160 °C and 1600 kPa in the presence of potassium hydroxide to form methyl vinyl ether [*107-25-5*]. The methyl vinyl ether is then hydrolyzed with dilute acid:

$$C_2H_2 + CH_3OH \xrightarrow{KOH} CH_3OCH{=}CH_2 \xrightarrow{H_2O,\,H^+} CH_3OH + CH_3CHO$$

Production via Ethylidene Diacetate. Addition of acetic acid to acetylene in the presence of mercury(II) salts yields ethylidene diacetate [*542-10-9*], $CH_3CH(OCOCH_3)_2$, which decomposes into acetaldehyde and acetic anhydride at 130 – 145 °C in the presence of acid catalysts (e.g., $ZnCl_2$). This process was developed by the Societe Chimique des Usines du Rhône on an industrial scale in 1914 but is now without importance.

4.3. Production from Ethylene

Ethylene is now the most important starting material for the production of acetaldehyde. Most of the present capacity works by the direct oxidation of ethylene (Wacker process).

4.3.1. Direct Oxidation of Ethylene

This process was developed between 1957 and 1959 by Wacker-Chemie and Hoechst [42]. Formally, the reaction proceeds as follows:

$$C_2H_4 + 1/2O_2 \rightarrow CH_3CHO \quad \Delta H = -244\ \mathrm{kJ/mol}$$

An aqueous solution of $PdCl_2$ and $CuCl_2$ is used as catalyst. Acetaldehyde formation had already been observed in the reaction between ethylene and aqueous palladium chloride. This reaction is almost quantitative:

$$C_2H_4 + PdCl_2 + H_2O \rightarrow CH_3CHO + Pd + 2HCl$$

In the Wacker-Hoechst process, metallic palladium is reoxidized by $CuCl_2$, which is then regenerated with oxygen:

$$Pd + 2CuCl_2 \rightarrow PdCl_2 + 2CuCl$$

$$2CuCl + 1/2O_2 + 2HCl \rightarrow 2CuCl_2 + H_2O$$

Therefore only a very small amount of $PdCl_2$ is required for the conversion of ethylene. The reaction of ethylene with palladium chloride is the rate-determining step.

One- and two-stage versions of the process are on stream. In the *one-stage method*, an ethylene – oxygen mixture reacts with the catalyst solution. During the reaction a stationary state is established in which "reaction" (formation of acetaldehyde and reduction of $CuCl_2$) and "oxidation" (reoxidation of CuCl) proceed at the same rate. This stationary state is determined by the degree of oxidation of the catalyst, as expressed by the ratio $c_{Cu^{2+}} / (c_{Cu^{2+}} + c_{Cu^{+}})$. In the *two-stage process* the reaction is carried out with ethylene and then with oxygen in two separate reactors. The catalyst solution is alternately reduced and oxidized. At the same time the degree of oxidation of the catalyst changes alternately. Air is used instead of pure oxygen for the catalyst oxidation.

Reaction Mechanism. The first step of the reaction is the complexation of ethylene to give a palladium ethylene complex

$$[PdCl_4]^{2-} + H_2C = CH_2 \rightleftharpoons [(C_2H_4)\,PdCl_3]^- + Cl^-$$

Kinetic studies of this reaction show that chloride ions have a inhibiting effect which is explained by the following substitution of a further chloride ligand by a water molecule:

$$[(C_2H_4)\,PdCl_3]^- + H_2O \rightleftharpoons [(C_2H_4)\,PdCl_2\,(H_2O)] + Cl^-$$

Dissociation of hydrogen ions explains the inhibiting effect of acids:

$$[(C_2H_4)\,PdCl_2\,(H_2O)] \rightleftharpoons [(C_2H_4)\,PdCl_2\,(OH)]^- + H^+$$

It is assumed that the hydroxyl complex has the *trans* geometry. Evidence for the *trans – cis* isomerisation of this complex was revealed by a detailed kinetic study of the reaction. π-Bonded ethylene ligands in the *trans* position weaken the metal – chlorine bonds, so that the chloro ligand can be easily substituted by a water molecule from which a hydrogen ion dissociates.

$$trans\text{-}[(C_2H_4)\,PdCl_2\,(OH)]^- + H_2O \rightleftharpoons [(C_2H_4)\,PdCl(OH)_2]^- + H^+ + Cl^-$$

The OH^- ligand in the *trans* position is replaced by a Cl^- ion, leading to a *cis* complex:

$$[(C_2H_4)\,PdCl(OH)_2]^- + H^+ + Cl^- \rightleftharpoons cis\text{-}[(C_2H_4)\,PdCl_2\,(OH)]^- + H_2O$$

The next reaction step is the formation of a σ-bonded hydroxyethylpalladium species. This reaction has been regarded as a *cis* ligand insertion reaction in which the OH^- ligand attacks the π-bonded olefin:

$$cis\text{-}[(C_2H_4)\,PdCl_2\,(OH)]^- \rightleftharpoons [HOCH_2CH_2PdCl_2]^-$$

Hydride transfer to give an α-hydroxyethylpalladium complex is followed by reductive elimination, which is the rate-determining step:

$$[HOCH_2CH_2PdCl_2]^- \rightleftharpoons [CH_3CH\,(OH)\,PdCl_2]^-$$

$$[CH_3CH\,(OH)\,PdCl_2]^- \rightarrow CH_3CHO + Pd + H^+ + 2Cl^-$$

For a detailed description, see [43]. The rate of reaction can be given by the following equation [44]:

$$\frac{-d\,c_{C_2H_4}}{d\,t} = k\,\frac{c_{PdCl_4^{2-}} \cdot c_{C_2H_4}}{c_{H^+} \cdot c_{Cl^-}^2}$$

The rate of reaction is diminished by the acid formed in the reduction of palladium chloride. This can be prevented by buffering the acid with basic copper salts (copper oxychloride, copper acetate). Reformation of the basic copper salts takes place during catalyst oxidation.

One-Stage Process (Fig. 2). Ethylene and oxygen are charged into the lower part of the reaction tower (a); the catalyst is circulated via the separating vessel (b) by the airlift principle and thoroughly mixed with the gas. Reaction conditions are about 130 °C and 400 kPa. An acetaldehyde – water vapor mixture, together with unconverted gas, is withdrawn from the separating vessel; from this mixture the reaction products are separated by cooling (c) and washing with water (d); unconverted gas is returned to the reactor. A small portion is discharged from the cycle gas as exhaust gas to prevent accumulation of inert gases in the cycle gas; these inert gases are either introduced as contamination of the feed gas (nitrogen, inert hydrocarbons) or formed as byproducts (carbon dioxide). A partial stream of catalyst is heated to 160 °C (m) to decompose byproducts that have accumulated in the catalyst.

Crude acetaldehyde obtained during washing of the reaction products is distilled in two stages. The first stage (g) is an extractive distillation with water in which lights ends having lower boiling points than acetaldehyde (chloromethane, chloroethane, and carbon dioxide) are separated at the top, while water and higher-boiling byproducts, such as acetic acid, crotonaldehyde, or chlorinated acetaldehydes, are withdrawn together with acetaldehyde at the bottom. In the second column (i) acetaldehyde is purified by fractional distillation.

Two-Stage Process (Fig. 3). Tubular reactors (a), (d) are used for both "reaction" and "oxidation". The gases react almost completely in the presence of the catalyst. Reaction of ethylene takes place at 105 – 110 °C and 900 – 1000 kPa. Catalyst solution containing acetaldehyde is then expanded in a flash tower (b) by reducing the pressure to atmospheric level. An acetaldehyde – water vapor mixture distills overhead while catalyst is sent via the pump (c) to the oxidation reactor (d), in which it reacts with oxy-

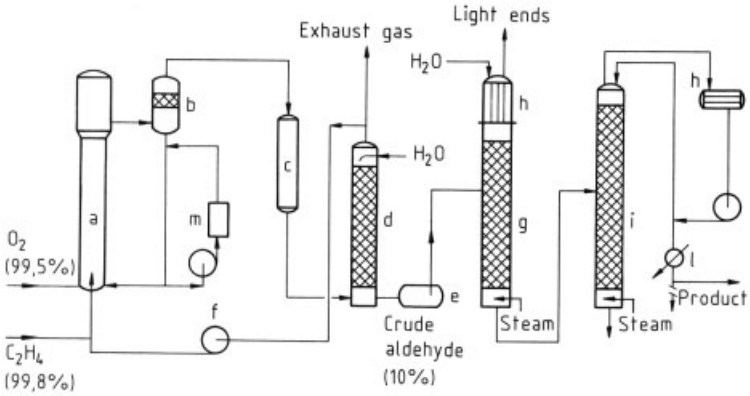

Figure 2. One-stage process
a) Reactor; b) Separating vessel; c) Cooler; d) Scrubber; e) Crude aldehyde tank; f) Cycle-gas compressor; g) Light-ends distillation; h) Condensers; i) Purification column; l) Product cooler; m) Regeneration

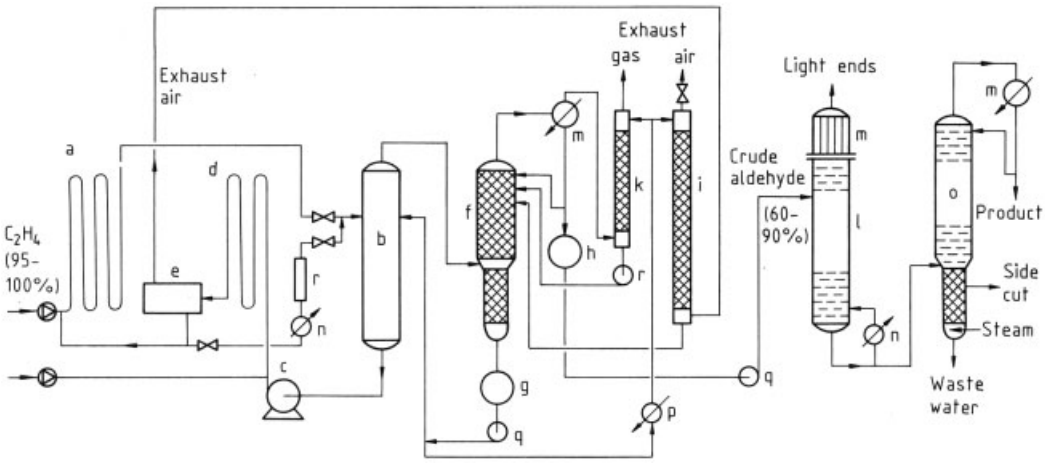

Figure 3. Two-stage process
a) Reactor; b) Flash tower; c) Catalyst pump; d) Oxidation reactor; e) Exhaust-air separator; f) Crude-aldehyde column; g) Process-water tank; h) Crude-aldehyde container; i) Exhaust-air scrubber; k) Exhaust-gas scrubber; l) Light-ends distillation; m) Condensers; n) Heater; o) Purification column; p) Cooler; q) Pumps; r) Regeneration

gen at about 1000 kPa. As oxidation and reaction are carried out separately, no high-purity starting gas is required. Generally, air is used instead of oxygen. Oxygen conversion is almost complete; the exhaust air from (e) can be used as inert gas for plant use. The oxidized catalyst solution separated from exhaust air in the separator (e) is reused for the reaction with ethylene in (a).

Acetaldehyde – water vapor mixture from the flash tower (b) is preconcentrated in column (f) to 60 – 90 % acetaldehyde by utilizing the heat of reaction. Process water discharged at the bottom of (f) is returned to the flash tower to maintain a constant catalyst concentration. A portion of the process water is used for scrubbing exhaust air (nitrogen from the "oxidation") in (i) and exhaust gas (inert gas from the "reaction") in (k) free of acetaldehyde. Scrubber water then flows to the crude aldehyde column (f).

A two-stage distillation of the crude acetaldehyde follows. In the first stage (l), low-boiling substances, such as chloromethane, chloroethane and carbon dioxide, are separated. In the second stage (o), water and higher-boiling byproducts, such as chlorinated acetaldehydes and acetic acid, are removed from acetaldehyde, and the latter is obtained in pure form overhead. Chlorinated acetaldehydes become concentrated within the column as medium-boiling substances and are discharged laterally. From

this mixture, monochloroacetaldehyde can be obtained as the hemihydrate. Residual byproducts can be returned to the catalyst for oxidative decomposition. This oxidative self-purification is supported by thermal treatment of a partial stream of catalyst at about 160 – 165 °C (regeneration, r).

When gas mixtures obtained in naphtha cracking processes are used as raw material, conventional towers are used as reactors instead of coiled pipes; So far, these processes have not been developed industrially. Such mixtures contain 30 – 40 % ethylene in addition to inert hydrocarbons and hydrogen [45].

Comparison of the Two Methods. In both one- and two-stage processes the acetaldehyde yield is about 95 % and the production costs are virtually the same. The advantage of using dilute gases in the two-stage method is balanced by higher investment costs. Both methods yield chlorinated hydrocarbons, chlorinated acetaldehydes, and acetic acid as byproducts. Generally, the choice of method is governed by the raw material and energy situations as well as by the availability of oxygen at a reasonable price.

Balance of Reaction and Side Products. The yield in both type of process is nearly the same. The balance of the two stage process is as follows:

100 parts of ethylene gives:

95 parts acetaldehyde
1.9 parts chlorinated aldehydes
1.1 parts unconverted ethylene
0.8 parts carbon dioxide
0.7 parts acetic acid
0.1 parts chloromethane
0.1 parts ethyl chloride
0.3 parts ethane, methane, crotonaldehyde and other minor side products

The chlorinated aldehydes consist of chloroacetaldehyde, dichloroacetaldehyde, trichloroacetaldehyde, and 2-chloro-2-butenal.

Process Variant. An interesting variant of the process, although so far of no technical importance, uses glycol as the reaction medium. The cyclic acetal of acetaldehyde, namely the easily hydrolyzable 2-methyl-1,3-dioloxane, is obtained. An advantage of this method is the high rate of reaction [46].

Construction Materials. During process development, serious problems have been caused by the extremely corrosive aqueous $CuCl_2 - PdCl_2$ solution. These problems have been solved in the two-stage process either by constructing parts in contact with the catalyst solution entirely from titanium or by lining those parts with the metal. In the one-stage process, the reactor is lined with acid-proof ceramic material, the tubing is made of titanium, and certain other parts are of tantalum.

Waste Air. The waste air from the oxidation process contains small amounts of unconverted ethylene, some acetaldehyde, and side products from the reaction such as ethane, chloromethane, chloroethane, and methane. In Germany the waste air must be purified of these side products to meet the criteria of TA-Luft [47]. The byproducts are oxidized over a chromium oxide catalyst. The hydrogen chloride generated is removed by washing, so that the waste air contains mainly carbon dioxide.

Wastewater. Side products of the oxidation process that enter the wastewater are acetic acid, crotonaldehyde and chlorinated aldehydes. Some of the chlorinated aldehydes are highly toxic and show high antimicrobial activity. Therefore they must be treated before entering the wastewater plant to render them biologically degradable. Cleavage of organic chlorine by alkaline hydrolysis is a possible method. If the chlorinated compounds can not be destroyed the wastewater has to be incinerated.

4.3.2. Acetaldehyde as Byproduct

Acetaldehyde is also formed in the production of vinyl acetate from ethylene or acetylene (→ Vinyl Esters). It is separated by distillation and is normally converted to acetic acid for reuse. In one version of the method starting from ethylene, vinyl acetate and acetaldehyde are obtained in a molar ratio of 1 : 1. This makes the process nearly self-sufficient in acetic acid.

4.3.3. Isomerization of Ethylene Oxide

Research was carried out on this process [48] before adopting the direct oxidation of ethylene.

Catalysts were Al_2O_3, SiO_2, and acid salts of mineral acids, such as sulfuric acid, phosphoric acid, or molybdic acid. Yields of 90 – 95 % have been reported, but the process has not gained industrial importance.

4.4. Production from C_1 Sources

Since the increases in oil price in 1973/74 and 1977, C_1 material has gained interest as a feedstock for organic chemicals and as a substitute for petrochemicals. However, for acetaldehyde production, C_1 material seems to be of minor importance because most of the classical acetaldehyde derivatives can be made from C_1 sources.

Production Directly from Synthesis Gas. Acetaldehyde is formed with low selectivity and a yield of ca. 30 % from synthesis gas, together with acetic acid, ethanol, and saturated hydrocarbons, mainly methane. Catalysts are cobalt and rhodium compounds activated by iodine compounds or magnesium chloride and supported on a silicate carrier [49]. There is so far no industrial use of this process.

Production via Methanol, Methyl Acetate, or Acetic Anhydride. *Hydroformylation of methanol* with CO/H_2 has been well known since the discovery of the oxo process [50]. It takes place in the presence of hydroformylation catalysts, such as cobalt, nickel, and iron salts (e.g., $CoBr_2$, CoI_2) or the corresponding metal carbonyls, at increased temperature (180 – 200 °C) and high pressure (30 – 40 MPa). An acetaldehyde selectivity of 80 % or more has been claimed using an iron – cobalt carbonyl or alternatively a cobalt – nickel catalyst in the presence of tertiary amines, phosphines, or nitriles as the catalyst [51, 52].

Similarly, high selectivity has been claimed for the *hydrocarbonylation of methyl acetate* with palladium or rhodium catalysts in the presence of tertiary phosphines and iodine compounds as well as cobalt – ruthenium catalysts in the presence of methyl and sodium iodides [53].

Some patents describe the formation of acetaldehyde by *reduction of acetic anhydride* with hydrogen over palladium or platinum on a carrier at low pressures and moderate temperatures [54]. If the importance of acetaldehyde as an organic intermediate were to decrease in the future as outlined in Chapter 8, this method might be of some interest for the economical production of small quantities of acetaldehyde.

4.5. Production from Hydrocarbons

Acetaldehyde is a byproduct of the production of acrolein, acrylic acid, and propene oxide from propene. It is also formed in the oxidation of saturated hydrocarbons (e.g., propane or butane) in the gas phase as operated by Celanese in the United States [55] (→ Acetic Acid).

5. Quality and Analysis

High demands on acetaldehyde purity generally are made; a typical specification is:

Color	practically colorless
Acetaldehyde	more than 99.5 wt %
Acid (as acetic acid)	less than 0.1 wt %
Water	less than 0.02 wt %
Chlorine	less than 30 mg/kg
Dry residue	less than 10 mg/kg

The *acid content* is determined directly by titration, *water content* using the Karl Fischer reagent or empirically from the cloud point of a carbon disulfide – acetaldehyde mixture, and *chlorine content* (mostly in the form of organic chlorine compounds) by combustion in a hydrogen stream and determination of the hydrochloric acid in the condensate.

6. Storage and Transportation

6.1. Storage

For storage of acetaldehyde, the national regulations must be observed. In *Germany* the ordinance on operational safety (Betriebssicherheitsverordnung) applies, under which acetaldehyde is classified as highly inflammable liquid (F+). In the *USA*, loading and storage of acetaldehyde are governed by the EPA.

In *Japan* the Fire Defense Law and the supplement "Cabinet Order for Control of Dangerous Articles" applies. According to these regulations, outer storage tanks must not be made of

copper, magnesium, silver, mercury, or alloys of these metals. They must be equipped with cooling facilities in order to keep the temperature below 15 °C, and also with an inert-gas sealing system.

For safety data, see Chapter 2.

6.2. Transportation

International Regulations. Acetaldehyde is classified as a flammable liquid. Transportation is governed by:

IMDG Code (sea transport): class 3, UN 1089, packing group I
RID (rail): class 3, UN 1089, packing group I
ADR (road): class 3, UN 1089, packing group I
ADNR (inland waterways), class 3, UN 1089, packing group I
IATA-DGR (air transportation): class 3, UN 1089, packing group I, PAC forbidden, CAC packing instruction 304, 30 L max.

National Regulations. *Germany:* GGVSE (road and rail); GGVSee (sea). *USA:* Regulations of the U.S. Department of Transportation: CFR 49.

Types and sizes of packaging and containers are recommended by the regulations mentioned above.

6.3. Other Regulations

In *Germany*, any accidental release of acetaldehyde into the air, water, or soil must be reported to the appropriate authorities according to the ordinance on operational safety (Betriebssicherheitsverordnung) or the regulations on units for storing, filling, and moving materials dangerous to water supplies (Hazardous Incident Ordinance, Störfallverordnung, 2005).

In the *United States*, a similar regulation applies according to the Comprehensive Environmental Response, Compensation, and Liability Act of 1980. Incidents must be reported to the National Response Center for Water Pollution.

In *Japan*, acetaldehyde is specified as an Offensive Odor Material by the Offensive Odor Control Law. Under this law a district of dense population is specified, for which the concentration of the material in the air is limited. The maximum level of acetaldehyde allowed at the boundary of the factory or place where acetaldehyde is handled is 0.5 ppm.

Environmental problems are dealt with by the respective national laws (see above). If acetaldehyde is highly diluted with water it can easily be degraded biologically. In higher concentrations it kills bacterial flora.

7. Economic Aspects

Today the most important production process worldwide is the direct oxidation of ethylene. In Western Europe there is also some capacity for the production of acetaldehyde by oxidation of ethanol (7 %) and hydration of acetylene (17 %). In South America the entire capacity for production of acetaldehyde is based on the oxidation of ethanol.

However, the demand for acetaldehyde worldwide has continued to decrease primarily as a result of less consumption for acetic acid manufacture, as the industry continues to move toward the more efficient methanol carbonylation process. All manufacture of acetic acid from acetaldehyde in North America has been discontinued, and in Europe significant capacity for this transformation has been permanently shut down. Acetaldehyde use for acetic acid manufacture in Asia continues but is under pressure and will decline because of the establishment of methanol carbonylation technology. Since 1995, some 360 000 t of acetaldehyde capacity has been shut down in Western Europe, and Mexico has eliminated its acetaldehyde capacity. New acetaldehyde capacity has been installed in China, but this is the only area where new capacity has been added. Further carbonylation capacity will be added between 2006 and 2008, but afterward the demand for acetaldehyde for acetic acid will eventually be phased out. Capacity in China is expected to be around 860 000 t by 2008 and is expected to remain at those levels for at least five more years.

Table 2 gives production data for acetaldehyde.

Important producers and their production capacities (10^3 t) are listed in the following [56]:

Table 2. Production of acetaldehyde (10^3 t) [56]

Year	USA	W. Europe	Mexico	Japan	China	Total
1990	283	603	190	384		1460
1995	111	668	271	395	400	1845
2000	155	370	113	401	400	1439
2003	142	212	57	362	500	1273

Celanese Chemicals Europe GmbH, Germany	120
Eastman Chemical Company, USA	225
ECROS, SA, Spain	90
Japan Aldehyde Company Ltd., Japan	69
Jilin Chemical Industrial Company, China	180
Kyowa Yuka Company Ltd., Japan	61
Showa Denko K.K., Japan	300
Sinopec Yangzi Petrochemical Co., China	78 (400 planned)
Wacker Chemie AG, Germany	65

8. Polymers of Acetaldehyde

8.1. Paraldehyde

Paraldehyde, 2,4,6-trimethyl-1,3,5-trioxane [*123-63-7*], $C_6H_{12}O_3$, M_r 132.161, is a cyclic trimer of acetaldehyde:

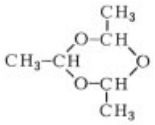

Properties. Paraldehyde is colorless and has an ethereal, penetrating odor.

bp	124.35 °C
mp	12.54 °C
Critical temperature t_{crit}	290 °C
Solubility in 100 g water at 13 °C	12 g
at 75 °C	5.8 g
In the solid state, four different crystal forms exist; transition points: 230.3 K, 147.5 K, 142.7 K.	
Paraldehyde is miscible with most organic solvents.	
Density d_4^{20}	0.9923
Refractive index n_D^{20}	1.4049
Viscosity at 20 °C	1.31 mPa · s
Heat of combustion at constant pressure	3405 kJ/mol
Heat capacity c_p at 25 °C	1.947 J g^{-1} K^{-1}
Entropy (l) at 25 °C	2.190 J g^{-1} K^{-1}
Free energy (l) at 25 °C	276.4 J/g
Heat of vaporization	41.4 kJ/mol
Latent heat of melting	104.75 J/g
Heat of formation from acetaldehyde(calculated from combustion enthalpies)	−113.0 kJ/mol
The equilibrium 3 acetaldehyde ⇌ paraldehyde is 94.3 % on the paraldehyde side at 150 °C.	

Production. Paraldehyde is produced from acetaldehyde in the presence of acid catalysts, such as sulfuric acid, phosphoric acid, hydrochloric acid, or acid cation exchangers. In the homogeneous reaction, acetaldehyde is added, with stirring and cooling, to paraldehyde containing a small amount of sulfuric acid. After the addition is completed, stirring is continued for some time to establish the equilibrium; the sulfuric acid is exactly neutralized with a sodium salt, such as sodium acetate, sodium carbonate, or sodium bicarbonate; the reaction mixture is separated into acetaldehyde, water, and paraldehyde by fractional distillation [57].

For continuous production, liquid acetaldehyde at 15 – 20 °C or acetaldehyde vapor at 40 – 50 °C is passed over an acid cation exchanger [58]. Conversion is greater than 90 %. Acetaldehyde and paraldehyde are separated by distillation. For depolymerization, acetaldehyde is slowly distilled off in the presence of acid catalysts. Paraldehyde also can be decomposed in the gas phase. Catalysts are HCl, HBr, H_3PO_4, or cation exchangers. The reaction is first order. Other catalysts described in the literature are Al_2O_3, SiO_2, $ZnSO_4$, and $MgSO_4$ [59].

Uses. Paraldehyde is used in chemical synthesis as a source of acetaldehyde whereby resin formation and other secondary reactions are largely eliminated. Such synthetic reactions are, for instance, used for the production of pyridines and chlorination of chloral. Between 1939 and 1945 paraldehyde was used as a motor fuel.

8.2. Metaldehyde

Metaldehyde [*9002-91-9*], $C_8H_{16}O_4$, M_r 176.214, is the cyclic tetramer of acetaldehyde:

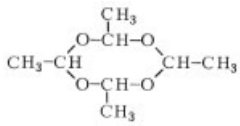

Properties. Metaldehyde forms tetragonal prisms, *mp* (closed capillary) 246.2 °C, sublimation temperature (decomp.) 115 °C, heat of combustion at constant volume 3370 kJ/mol.

Metaldehyde is insoluble in water, acetone, acetic acid, and carbon disulfide.

Depolymerization of metaldehyde to acetaldehyde begins at 80 °C, and is complete above 200 °C. Depolymerization takes place faster and at lower temperatures in the presence of acid catalysts, such as dilute H_2SO_4 or H_3PO_4. Metaldehyde does not show the typical acetaldehyde reactions. It is stabilized by ammonium carbonate or other weakly basic compounds which neutralize acidic potential catalysts.

Production. Metaldehyde is obtained in addition to large amounts of paraldehyde during polymerization of acetaldehyde in the presence of HBr and alkaline earth metal bromides, such as $CaBr_2$, at temperatures below 0 °C. However, yields are scarcely higher than 8 %. Yields of 14 – 20 % have been reported when working in the presence of 7 – 15 % of an aliphatic or cyclic ether at 0 – 20 °C [60]. Insoluble metaldehyde is filtered out. Acetaldehyde is then distilled from the filtrate following depolymerization of the paraldehyde and is returned to the polymerization. Recycling of the large amounts of acetaldehyde results in losses that increase the process costs.

Uses. Metaldehyde in pellet form is marketed as a dry fuel (Meta). Mixed with a bait, metaldehyde is used as a molluscicide.

8.3. Polyacetaldehyde

Polyacetaldehyde [*9002-91-9*] is a high-molecular-mass polymer with an acetal structure (polyoxymethylene structure):

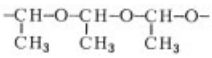

By using cationic initiators, mainly an amorphous polymer is obtained. Temperatures below −40 °C are preferred in this case. Above −30 °C, mainly paraldehyde and metaldehyde are produced. The initiator activity also depends on the solvent used. Suitable initiators include H_3PO_4 in ether and pentane, as well as HCl, HNO_3, CF_3COOH, $AlCl_3$ in ether, and particularly BF_3 in liquid ethylene [61]. Al_2O_3 and SiO_2 also seem to be good initiators [62].

The polymer has a rubber-like consistency and is soluble in common organic solvents. It depolymerizes at room temperature, liberating acetaldehyde. It evaporates completely within a few days or weeks. Acidic compounds accelerate depolymerization, and amines (e.g., pyridine) stabilize polyacetaldehyde to a certain extent. A complete stabilization (as, for instance, in the case of polyformaldehyde) has not yet been achieved, so the polymer is still of no practical importance.

Copolymers with propionaldehyde, butyraldehyde, and allylacetaldehyde also have been produced [63]. Crystalline, isotactic polymers have been obtained at low temperatures (for example, −75 °C) by using anionic initiators [64]. Suitable initiators are alkali metal alkoxides, alkali metals, or metal alkyls in hydrocarbon solvents. The products are insoluble in common organic solvents but have an acetal structure like the amorphous polymers' [65]. Polymerization of acetaldehyde to poly(vinyl alcohol), which in contrast to polyacetaldehyde has a pure carbon backbone, has not yet been achieved [66].

9. Toxicology and Occupational Health

Acetaldehyde. At higher concentrations (up to 1000 ppm), acetaldehyde irritates the mucous membranes. The perception limit of acetaldehyde in air is in the range between 0.07 and 0.25 ppm [67, 68]. At such concentrations the fruity odor of acetaldehyde is apparent. Conjunctival irritations have been observed after a 15-min exposure to concentrations of 25 and 50 ppm [69], but transient conjunctivitis and irritation of the respiratory tract have been reported after exposure to 200 ppm acetaldehyde for 15 min [69, 70]. The penetrating odor, the low perception limit, and the irritation that acetaldehyde causes, give an effective warning so that no serious cases of acute intoxication with pure acetaldehyde have been reported. Acute acetaldehyde intoxication can also be observed following combined ingestion of disulfiram (Antabuse) and ethanol [73]. In animal experiments at high concentrations (3000 – 20 000 ppm), pulmonary edema

and a narcotic effect become evident. The clinical course is similar to alcohol intoxication. Death occurs by breath paralysis or — with retardation — by pulmonary edema.

For rats, the LC_{50} (30 min inhalation) is 20 500 ppm [71]. No studies on subchronic or chronic toxicity of acetaldehyde in humans are available. Investigations of sister chromatid exchange in cell cultures [72] and in human lymphocytes [74] and studies of single- and double-strand breaks in human lymphocytes incubated with acetaldehyde [75] have revealed mutagenic effects of acetaldehyde. Long-term exposures of Syrian golden hamsters to concentrations in the range 1650 – 2500 ppm have resulted in inflammatory hyperplastic and metaplastic alterations of the upper respiratory tract with an increase in carcinomas of the nasal mucosa and the larynx [76]. Male and female Wistar rats were exposed to aldehyde concentrations of 0 ppm, 750 ppm, 1500 ppm for 6 h daily, 5 d per week for 52 weeks. In the highest dose group, the initial aldehyde concentration of 3000 ppm was reduced to 1000 ppm in the course of the study due to toxic effects. The dose-dependent effects found were increased mortality in all dose groups and delayed growth in the middle- and highest-dose group. Adenocarcinomas were observed at all three investigated concentrations. An increased rate of squamous epithelial carcinomas was only seen at 1500 ppm or more. Histological signs of irritation were observed in the larynx region in most animals from the medium- and high-dose groups [77]. Concomitant exposure to acetaldehyde also considerably increases the number of tracheal carcinomas induced by instillation of benzo[*a*]pyrene [72]. This suggests that chronic tissue injury is a prerequisite for tumor formation by acetaldehyde. Tumors probably do not develop if doses are not sufficient to cause tissue necrosis. In male Wistar rats exposed to 150 ppm or 500 ppm for 6 h per day, 5 d per week for 4 weeks, morphological changes in the olfactory epithelium were observed in the high-dose group [78]. However, long-term toxicity data at lower exposure concentrations are not yet available.

Currently, the TLV is 25 ppm (STEL/ceiling value) [79], and the MAK is 50 ppm [80]; the latter value is preliminary. At 50 ppm acetaldehyde, no irritation or local tissue damage in the nasal mucosa is observed. Because the mechanism of action is assumed to be analogous to that of formaldehyde, acetaldehyde is regarded as a suspected carcinogen [80].

When taken up by the organism, acetaldehyde is metabolized rapidly in the liver to acetic acid. Only a small proportion is exhaled unchanged. After intravenous injection, the halflife in the blood is approximately 90 s [81].

Paraldehyde acts as a sedative with few side effects. Ingested paraldehyde partly is metabolized to carbon dioxide and water and partly is exhaled unchanged. It generates an unpleasant odor in the expired air and therefore is not much used.

Metaldehyde decomposes slowly to acetaldehyde in the presence of acids, so ingestion may cause irritation of the gastric mucosa with vomiting. As characteristic signs of a metaldehyde intoxication, especially in children, heavy convulsions (sometimes lasting several days) have been reported, with lethal outcomes, after ingestion of several grams of metaldehyde [82]. For these reasons, those molluscicides and solid fuels which contain metaldehyde must be kept away from children.

10. References

General References

1. *Beilstein* **1** 594; **1**, 1st suppl., 1, 321; **1**, 2nd suppl., 1, 654; **1**, 2nd suppl., 1, 654; **1**, 3rd suppl., 1, 2617; **1**, 4th suppl., 3094.
2. S. A. Miller (ed.): *Acetylene*, E. Benn, London 1965.
3. S. A. Miller (ed.): *Ethylene*, E. Benn, London 1969.
4. R. Jira in 3: pp. 639 and 650.
5. R. Sieber in 3: pp. 659 and 668.
6. R. Page in 3: p. 767.
7. W. Reppe: *Chemie und Technik der Acetylen-Druck-Reaktionen*, Verlag Chemie, Weinheim 1951.
8. *Review on Manufacturing Processes of Wacker, Lonza, IG-Farben factories Hüls, Gendorf, Schkopau and Ludwigshafen, Stickstoff AG, Knapsack*, BIOS Final Report 1049, item no. 22 (1946).
9. *Report on the IG Works Hüls and Gendorf*, FIAT Final Report 855 (1946).

10. *Acetaldehyde from Acetylene*, BIOS Final Report 370.
11. *Kirk-Othmer*, 3rd ed., **1**, 97 – 112.

Specific References

12. R. C. Weast, H. J. Astle (eds.): *CRC Handbook of Chemistry and Physics*, 60th ed., CRC Press, Boca Raton, Florida 1979 – 1980.
13. D'Ans-Lax, *Taschenbuch für Chemiker und Physiker*, 3rd ed., Springer Verlag, Berlin 1964.
14. Th. F. Smith, R. F. Bonner, *Ind. Eng. Chem.* **43** (1951) 1169.
15. T. R. Das, N. R. Kuloor, *J. Indian Inst. Sci.* **50** (1968) 45.
16. P. M. Craven, J. D. Lambert, *Proc. R. Soc. (London)* **205 A** (1951) 439 and 444.
17. R. W. Gallant, *Hydrocarbon Process.* **47** (1968) no. 5, 151.
18. A. K. Shaka, *Indian J. Phys.* **6** (1931) 449.
19. L. P. Filippov, *Vestn. Mosk. Univ.* **9**, no. 12, *Ser. Fiz. Mat. Estestv. Nauk.* (1954) no. 8, 45 – 48.
20. J. D. Lambert, E. N. Staines, S. D. Woods, *Proc. R. Soc. (London)* **200 A** (1950) 262.
21. *Landolt-Börnstein*, 6th ed., vol. **2**, part 4, Springer Verlag, Berlin 1961.
22. M. M. Brazhnikov, A. D. Peshchenko, O. V. Ral'ko, *Zh. Prikl. Khim (Leningrad)* **49** (1976) no. 5, 1041;
Chem. Abstr. **85** (1976) 45 851.
23. J. D. Lambert, G. A. H. Roberts, J. S. Rowlington, V. J. Wilkinson, *Proc. R. Soc. (London)* **200 A** (1960) 262.
24. P. M. Chaudhuri, R. A. Stanger, G. P. Mathur, *J. Chem. Eng. Data* **13** (1968) 9 – 11.
25. A. A. Dobrinskaya, V. G. Markovich, M. B. Neiman, *Isv. Akad. Nauk. S.S.S.R., Otd. Khim. Nauk.* 1953, 434 – 441;
Bull. Acad. Sci. U.S.S.R. Div. Chem. Sci. (Engl. Transl.) 1953, 391 – 398;
Chem. Abstr. **49** (1955) 4378.
26. *Landholt-Börnstein*, 6th ed., vol. **2**, part 2, Springer Verlag, Berlin 1961.
27. G. O. Morrison, T. P. G. Shaw, *Trans. Electrochem. Soc.* **63** (1933) 425.
28. B. P. Mullins, *Fuel* **32** (1953) 481.
29. F. C. Mitchell, H. C. Vernon, *Chem. Met. Eng.* **44** (1937) 733.
30. K. Sennewald, *Erdöl Kohle* **12** (1959) 364.
31. *Petroleum Refiner* **40** (1961) no. 11, 308.
32. *Ullmann*, 4th ed., **9**, 6.
33. *Ullmann*, 4th ed., **7**, 374.
34. SU 287 919, 1970.
35. Heavy Minerals Co., US 2 884 460, 1955 (V.I. Komarevsky).
36. Knapsack-Griesheim, DE 1 097 969, 1954 (W. Opitz, W. Urbanski);
DE 1 108 200, 1955 (W. Opitz, W. Urbanski).
37. W. L. Faith, D. B. Keyes, R. L. Clarks: *Industrial Chemicals*, 3rd ed., J. Wiley & Sons, New York 1965, p. 2.
38. Shell Development Co., US 2 883 426, 1957 (W. Brackman).
Eastman Kodak Co., US 3 106 581, 1963 (S.D. Neely).
Veba-Chemie, DE 1 913 311, 1969 (W. Ester, W. Hoitmann).
39. *Petroleum Refiner* **36** (1957) 249.
40. *Ullmann*, 3rd ed., **3**, 4.
41. K. Kon, T. Igarashi, *Ind. Eng. Chem.* **48** (1956) 1258.
42. Consortium für elektrochemische Industrie, DE 1 049 845, 1957 (W. Hafner, J. Smidt, R. Jira, R. Rüttinger, J. Sedlmeier).
DE 1 061 767, 1957 (W. Hafner, J. Smidt, R. Jira, J. Sedlmeier);
DE 1 080 994, 1957 (W. Hefner, J. Smidt, R. Jira).
J. Smidt, W. Hafner, R. Jira, J. Sedlmeier, R. Sieber, R. Rüttinger, H. Kojer, *Angew. Chem.* **71** (1959) 176;
ibid. **74** (1962);
Angew. Chem. Int. Ed. Engl. **1** (1962) 80;
Chem. Ind. 1962, 54.
43. R. Jira: "Oxidations," in W. A. Herrmann (ed.): *Applied Homogeneous Catalysis with Organometallic Compounds*, vol. **1** VCH, Weinheim 1996.
44. I. I. Moiseev, M. N. Vargaftik, Y. K. Sirkin, *Dokl. Akad. Nauk SSSR* **153** (1963) 140.
P. M. Henry, *J. Am. Chem. Soc.* **86** (1964) 3246.
45. Consortium für elektrochemische Industrie, DE 1 215 677, 1964 (W. Hefner, R. Jira, J. Smidt).
46. Lummus Co., BE 668 601, 1965.
W. G. Lloyd, *J. Org. Chem.* **34** (1969) 3949.
47. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft TA-Luft) vom 27.02.1986, GMBL 1986, S. 95.
48. Sicedison, FR 1 316 720, 1961 (J. Herzenberg, P. Gialtoni).
British Petroleum Co., FR 1 367 963, 1963.
Hoechst, US 3 067 256, 1959 (K. Fischer, K. Vester).

49. Rhône-Poulenc, EP 0 011 043, 1979 (J. Gauthier-Lafage, R. Perron).
Union Carbide, US 4 235 798, 1980 (W. J. Bartley, T. P. Wilson, P. C. Ellgen).
Hoechst, DE-OS 2 814 365, DE-OS 2 814 365, DE-OS 2 825 598, 1978 (H. J. Arpe, E.-J. Leupold, F. A. Wunder, H.-J. Schmidt).
50. Du Pont, US 2 457 204, 1946 (R. E. Brooks).
US 3 356 734, 1963 (M. Kuraishi, S. Asano, A. Takahashi).
51. Exxon Research and Eng. Co., EP 0 027 000, 1980 (G. Doyle).
Celanese Corp., US 4 201 868, 1980 (W. E. Slinkard).
Union Rheinische Braunkohlen Kraftstoff, DE-OS 2 913 677, 1979 (J. Korff, M. Fremery, J. Zimmermann).
52. Gulf Research & Development Co., US 4 239 704/5, 1980 (W. R. Pretzer, T. P. Kobylinsky, J. E. Bozik).
British Petroleum Co., EP 0 029 723, 1980 (M.T. Barlow).
53. Rhône-Poulenc, EP 0 046 128/9, 1981 (J. Gauthier-Lafage, R. Perron).
Halcon Research & Development Corp., US 4 302 611, 1981;
BE 890 376, 1982 (R. V. Porcelli).
54. Halcon SD Group, FR 8 118 437/8, 1981 (D. Moy).
Kuraray Co., EP 0 040 414, 1981 (S. Nakamura, M. Tamura).
55. *Ullmann*, 3rd ed., suppl. vol., p. 172.
56. B. Suresh, G. Toki: "Acetaldehyde", *Chemical Economics Handbook*, SRI International 2004.
57. Eastman Kodak Co., US 2 318 341, 1937 (B. Thompson).
58. Publicker Ind., US 2 479 559, 1947 (A. A. Dolnick et al.).
Melle-Bezons, DE-OS 1 927 827, 1969 (M. G. Gobron, M. M. Repper).
59. T. Kawaguchi, S. Hasegawa, *Tokyo Gakugei Daigaku* **21** (1969) 63.
60. Publicker Ind., US 2 426 961, 1944 (R. S. Wilder).
61. O. Vogl, *J. Polym. Sci., Part A* **2** (1964) 4591;
H. Staudinger, *Trans. Faraday Soc.* **32** (1936) 249;
H. A. Rigby, C. J. Danby, C. N. Hinshelwood, *J. Chem. Soc.* 1948, 234.
62. J. T. Furukawa, T. Saegusa, T. Tsuruta, H. Fujii, T. Tatana, *J. Polym. Sci.* **36** (1959) 546.
J. T. Furukawa, T. Saegusa, T. Tsuruta, H. Fujii, A. Kawazaki, *Makromol. Chem.* **33** (1960) 32.
Consortium für elektrochemische Industrie, DE 1 106 075, 1958 (J. Smidt, J. Sedlmeier).
63. Consortium für elektrochemische Industrie, DE 1 292 394, 1959 (J. Smidt, J. Sedlmeier).
64. Bridgestone Tire Co., FR 1 268 322 and FR 1 268 191, 1960 (J. Furukawa, T. Tsuruta, T. Saegusa, H. Fujii).
J. T. Furukawa, *Makromol. Chem.* **37** (1969) 149.
G. Natta, *Makromol. Chem.* **37** (1960) 156;
J. Polym. Sci. **51** (1960) 505.
65. O. Vogl, *J. Polym. Sci., Part A* **2** (1964) 4607.
66. T. Imoto, T. Matsubara, *J. Polym. Sci.* **56** (1962) 5.
67. C. P. McCord: *Odors, Physiology and Control*, McGraw-Hill, New York 1949.
68. W. Summer: *Odour Pollution of Air Causes and Control*, Chemical and Process Engineering Series, Leonard Hill, London 1971.
69. L. Silverman, H. F. Schulte, M. W. First, *J. Ind. Hyg. Toxicol.* **28** (1946) 265.
70. V. M. Sim, R. E. Pattle, *J. Am. Med. Assoc.* **165** (1957) 1908.
71. E. Skog, *Acta Pharmacol.* **6** (1950) 299.
72. G. Obe, H. J. Ristow, *Mutat. Res.* **58** (1978) 115.
73. J. Becker, H. Desel, H. Schuster, G. F. Kahl, *Ther. Umsch.* **52** (1995) 183.
74. G. Obe, R. Jonas, S. Schmidt, *Mutat. Res.* **174** (1986) 47.
75. N. P. Singh, A. Kahn, *Mutat. Res.* **337** (1995) 9.
76. V. J. Feron, A. Kruysse, R. A. Woutersen, *Eur. J. Cancer Clin. Oncol.* **18** (1982) 13.
77. R. A. Woutersen, L. M. Appelman, A. van Garderen-Hoetmer, V. J. Feron, *Toxicology* **41** (1986) 213.
78. L. M. Appelman et al., Report No. V84.382/140327, CIVO Institutes TNO, NL-3700AJ Zeist, The Netherlands, 1985.
79. *1997 TLVs and BEIs*. American Conference of Governmental Industrial Hygienists Inc., Cincinnati, Ohio 1997.
80. DFG Deutsche Forschungsgemeinschaft: *Occupational Toxicants: Critical Data for MAK Values and Classification of Carcinogens*, vol. **3**, Commission for the Investigation of Health Hazards of Chemical Compounds in the Work Area, VCH Verlagsgesellschaft, Weinheim 1992.
81. K. J. Freundt, *Naunyn-Schmiedeberg's Arch. Exp. Pathol. Pharmakol.* **260** (1968) 111;
Beitr. Gerichtl. Med. **27** (1970) 368.
82. O. R. Klimmer: *Pflanzenschutz- und Schädlingsbekämpfungsmittel, Abriß einer Toxikologie und Therapie von Vergiftungen*, Hundt-Verlag, Hattingen 1972.

Acetic Acid

Hosea Cheung, Celanese Ltd., Chemicals Division, Technical Center, Corpus Christi, Texas 78469, United States

Robin S. Tanke, Celanese Ltd., Chemicals Division, Technical Center, Corpus Christi, Texas 78469, United States

G. Paul Torrence, Celanese Ltd., Chemicals Division, Technical Center, Corpus Christi, Texas 78469, United States

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 3
4. Production 3
4.1. Carbonylation of Methanol 4
4.2. Direct Oxidation of Saturated Hydrocarbons 10
4.3. Acetaldehyde Process 13
4.4. Other Processes 15
4.5. Concentration and Purification . . 17
4.6. Construction Materials 17
5. Wastewater and Off-Gas Problems 17
6. Quality Specifications 18
7. Analysis 18
8. Storage, Transportation, and Customs Regulations 18
9. Uses 19
10. Derivatives 19
10.1. Salts 19
10.1.1. Aluminum Acetate 19
10.1.2. Ammonium Acetate 19
10.1.3. Alkali Metal Salts 21
10.2. Esters 21
10.2.1. Methyl Acetate 22
10.2.2. Ethyl Acetate 22
10.2.3. Butyl Acetate 22
10.2.4. 2-Ethylhexyl Acetate 23
10.2.5. Other Esters 23
10.3. Acetyl Chloride 23
10.4. Amides 23
10.4.1. Acetamide 23
10.4.2. *N,N*-Dimethylacetamide 24
10.5. Phenylacetic Acid 24
11. Economic Aspects 24
12. Toxicology and Occupational Health 25
13. References 26

1. Introduction

Acetic acid [*64-19-7*], CH_3COOH, M_r 60.05, is found in dilute solutions in many plant and animal systems. Vinegar, an aqueous solution containing about 4 – 12 % acetic acid, is produced by the fermentation of wine and has been known for more than 5000 years.

The major producers of acetic acid, accounting for ca. 70 % of total worldwide production, are the United States, Western Europe, and Japan. World capacity exceeds 7×10^6 t/a [1]. The largest end uses are in the manufacture of vinyl acetate [*108-05-4*] and acetic anhydride [*108-24-7*]. Vinyl acetate is used in the production of latex emulsion resins for applications in paints, adhesives, paper coatings, and textile treatment. Acetic anhydride is used in the manufacture of cellulose acetate textile fibers, cigarette filter tow, and cellulose plastics.

2. Physical Properties

Acetic acid is a clear, colorless, corrosive liquid that has a pungent odor and is a dangerous vesicant. It has a pK_a of 4.77 [2]. It melts at 16.75 °C [3] and boils at 117.9 °C [4] under 101.3 kPa [5]. It has a pungent vinegarlike odor. The detectable odor is as low as 1 ppm. The acid is combustible with a low flash point of 43 °C. The explosion limits of acetic acid vary from the upper explosion limit (UEL) of 16 % at 92 °C to the lower explosion limit (LEL) of 4 % at 59 °C. The liquid is usually available as glacial acetic acid with less than 1 wt % water and over 98 % purity. Besides water, the acid contains traces of impurities such as acetaldehyde, oxidized substances, iron, and chlorides.

Occasionally, the acid may be colored due to the presence of ethyl acetoacetate [*141-97-9*].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_045

The acetate is easily mistaken for formic acid because it reduces mercuric chloride. Traces of mercury may cause extensive corrosion by reaction with aluminum. Aluminum is a common material for containers to ship the acid [6].

Glacial acetic acid is very hydroscopic. The presence of 0.1 wt % water lowers the freezing point significantly [7]. Measuring the freezing point is a convenient way to evaluate acetic acid purity. This is shown in Table 1 [8].

Table 1. Freezing points for various acetic acid – water mixtures

wt % CH_3COOH	*fp*, °C	wt % CH_3COOH	*fp*, °C
100	16.75	96.8	11.48
99.6	15.84	96.4	10.83
99.2	15.12	96.0	10.17
98.8	14.49	93.5	7.1
98.4	13.86	80.6	− 7.4
98.0	13.25	50.6	−19.8
97.6	12.66	18.1	− 6.3
97.2	12.09		

Acetic acid forms azeotropes with many common solvents, such as benzene, pyridine, and dioxane. Acetic acid is miscible with water, ethanol, acetone, benzene, ether, and carbon tetrachloride. However, it is not soluble in CS_2 [2].

The physical properties of acetic acid are well documented, and their accuracy is important for commercial production. For instance, design and operation of distillation processes requires precise data. High-precision values provide a valuable asset to the chemical industry [9], [10].

The density of mixtures of acetic acid and water [11–13] is listed in Table 2. The density exhibits a maximum between 67 wt % and 87 wt %, corresponding to the monohydrate (77 wt % acetic acid). The density of pure acetic acid as a function of temperature is listed in Table 3 [14], [15].

Table 2. Densities of aqueous acetic acid solutions at 15 °C

wt % CH_3COOH	ϱ, g/cm^3	wt % CH_3COOH	ϱ, g/cm^3
1	1.007	60	1.0685
5	1.0067	70	1.0733
10	1.0142	80	1.0748
15	1.0214	90	1.0713
20	1.0284	95	1.0660
30	1.0412	97	1.0625
40	1.0523	99	1.0580
50	1.0615	100	1.0550

Table 3. Dependence of the density of pure acetic acid on temperature

t, °C	ϱ, g/cm^3	*t*, °C	ϱ, g/cm^3
20	1.0491	130	0.9235
26	1.0420	139	0.9119
34	1.0324	140	0.9092
52	1.0134	145	0.9030
60	1.0060	156	0.8889
63	1.0007	180	0.8555
75	0.9875	220	0.7941
85	0.9761	260	0.7136
97	0.9611	300	0.5950
100	0.9599	320	0.4615
107	0.9506	321	0.3506
117	0.9391		

Due to the difficulty in eliminating traces of water from acetic acid, the value for the boiling point varies from 391 to 392 K [10]. Careful studies prove that pure acetic acid boils at 391.10 K under 101.325 kPa [16]. The critical temperature and critical pressure are 594.45 K and 5785.7 kPa [3].

Precise data on vapor pressure of acetic acid are available from a regression equation (Eq. 1) [10], which covers the range from the normal boiling point to the critical point.

$$P = P_c \exp(A + B/T_r + C \ln T_r + DT_r^6) \quad (1)$$

where P is vapor pressure in kPa; T_r is reduced temperature T/T_c, T is temperature in K; T_c is 594.45 K, P_c is 5785.7 kPa, A is 10.08590, B is − 10.37932; C is − 3.87306, and D is 0.29342.

Table 4. Vapor pressure of pure acetic acid

t, °C	*p*, mbar	*t*, °C	*p*, mbar
0	4.7	150.0	2 461.1
10	8.5	160	3 160
20	15.7	170	4 041
30	26.5	180	5 091
40	45.3	190	6 333
50	74.9	200	7 813
60	117.7	210	9 612
70	182.8	220	11 733
80	269.4	230	14 249
90	390.4	240	17 057
100	555.3	250	20 210
110	776.7	260	23 854
118.2	1013	270	28 077
130.0	1386.5	280	32 801
140.0	1841.1		

The vapor pressure of pure acetic acid is given in Table 4 [17]. The density of the vapor corresponds to approximately twice the molecular

mass because of vapor-phase hydrogen bonding [8]. Hydrogen-bonded dimers and tetramers have both been proposed. Reports indicate that in the gas phase, the acid exists mainly in an equilibrium between monomer and dimer (Eq. 2) according to vapor density data [18], [19] molecular modeling, IR analysis [20], and gas-phase electron diffraction [21].

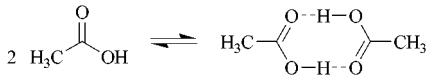

(2)

In the gas phase, monomic and dimeric acetic acid undergo extensive hydrolysis [22], [23]. In the liquid state, acetic acid equilibrates between monomer and dihydrated dimer or cyclic dimer (Eqs. 2 and 3) [24]. As the concentration of acetic acid increases, the equilibrium shifts to the right, favoring dimeric acetic acid. As the temperature is increased, the system shifts to the left, favoring the monomer. In addition, the acid may also form open-chain trimers and higher oligomers. However, at about 95 wt %, the acid exists mainly as a cyclic dimer and is no longer associated with water.

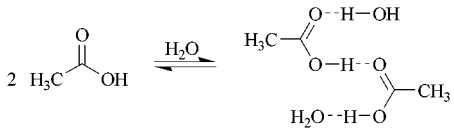

(3)

Up to 32 wt %, mixing acetic acid with water leads to evolution of heat. At higher acid concentrations heat is absorbed [25]. The measured values of the heat of mixing are consistent with the calculated values based on the dimers and tetramers described above. The aqueous mixture of acetic acid forms a eutectic mixture at − 26 °C. This eutectic mixture prevented earlier attempts to concentrate the acid by freezing. A way to obtain pure acid is to add urea or potassium acetate to the acid. Then, glacial acetic acid can be distilled.

Other physical properties of acetic acid are listed below.

Specific heat capacity	
Gaseous acid, c_p	1.110 J g^{-1} K^{-1} at 25 °C [33]
Liquid acid, c_p	2.043 J g^{-1} K^{-1} at 19.4 °C
Crystalline acid, c_p	1.470 J g^{-1} K^{-1} at 1.5 °C
	0.783 J g^{-1} K^{-1} at −175.8 °C
Heat of fusion	195.5 J/g [29]
Heat of combustion	−874.2 kJ/mol at 20 °C [30]
Cryoscopic constant	−269.56 °C [30]
Ebullioscopic constant	270.62 °C [30]
Melting point	6.76 °C [31]
Triple-point temperature	16.93 °C [19]
Boiling point	118.0 °C at 101.3 kPa [32]
Density	1265.85 kg/m^3 (solid at *mp*)
Viscosity	11.83 mPa · s at 20 °C [26]
	10.97 mPa · s at 25 °C [27]
	8.18 mPa · s at 40 °C [26]
	0.067 mPa · s at critical point [28]
Dielectric constant	6.194 at 20 °C [30]
Dipole moment	5.804 × 10^{-30} C · m [34]
Electrolyte conductivity	112.0 pS/m [33]
Refractive index n_D^{20}	1.372 [33]
Enthalpy of formation	
ΔH^0 (l, 25 °C)	−484.50 kJ/mol [35]
ΔH^0 (g, 25 °C)	−432.25 kJ/mol
Normal entropy	
$S°$ (l, 25 °C)	159.8 J mol^{-1} K^{-1} [35]
$S°$ (g, 25 °C)	282.5 J mol^{-1} K^{-1}
Flash point	43 °C (closed cup) [36]
	57 °C (open cup)
Autoignition point	465 °C [36]
Flammability	4.0 to 16.0 vol % in air [36]
Critical data	
p_c	5.786 MPa [3]
T_c	321.45 °C [3]

3. Chemical Properties

Many useful materials are made from acetic acid; several are discussed in Chapter 9. Acetate esters are formed by reaction of olefins or alcohols with acetic acid [37]. Acetamide is prepared by the thermal decomposition of ammonium acetate. Acetic acid can be converted to acetyl chloride with phosphorous trichloride or thionyl chloride.

Acetic acid is a raw material for a number of commercial processes. It can be converted to vinyl acetate with ethylene and oxygen (→ Vinyl Esters). Acetic acid is used in the manufacture of acetic anhydride (→ Acetic Anhydride) via ketene and in the production of chloroacetic acid (→ Chloroacetic Acids) using chlorine.

4. Production

Vinegar is still made by fermentation (→ Vinegar). However, the most important synthetic

routes to acetic acid are methanol carbonylation and liquid-phase oxidation of butane, naphtha, or acetaldehyde. Methanol carbonylation has been the method of choice for the past 25 years [38–40] and will likely remain the preferred route for large-scale production.

Several new technologies for producing acetic acid are being studied. Showa Denko may produce acetic acid by the gas-phase reaction of ethylene with oxygen over a supported palladium catalyst that contains a heteropolyacid or salt [41]. Numerous patents and publications discuss the production of acetic acid directly from ethane and oxygen. Production of acetic acid and acetate salts by microorganisms has also received considerable attention.

4.1. Carbonylation of Methanol

The manufacture of acetic acid from methanol [*67-56-1*] and carbon monoxide [*630-08-0*] at high temperature and high pressure was described by BASF as early as 1913 [42].

$$CH_3OH + CO \longrightarrow CH_3COOH \quad \Delta H = -138.6\,\text{kJ}$$

In 1941 Reppe at BASF demonstrated the efficiency of group VIII metal carbonyls as catalysts for carbonylation reactions, including hydroformylation [43], [44]. This led to the development of a high-pressure, high-temperature process (700 bar, 250 °C) with a cobalt iodide catalyst. The thrust of this work was to develop an acetic acid process not dependent on petroleum-based feedstocks. The current advantage of the methanol carbonylation route to acetic acid is the favorable raw material and energy costs. The synthesis gas raw material required for this process can be obtained from a variety of sources, which range from natural gas to coal. The cobalt-based carbonylation process was commercialized in 1960 by BASF in Ludwigshafen, Federal Republic of Germany [43], [45–47]. The initial capacity of 3600 t/a was expanded to 45 000 t/a by 1981 [48]. In 1966 Borden Chemical Co. started up a 45 000 t/a acetic acid unit in Geismar, Louisiana, United States, based on the BASF technology [43], [45]. The unit was expanded to 64 000 t/a by 1981 before it was shut down in 1982 [1], [48]. This unit was brought on stream again in 1988 for one year to meet acetic acid supply shortages in the United States.

Monsanto developed a low-pressure acetic acid process in the late 1960s with a rhodium iodide promoted catalyst system that demonstrated significantly higher activity and selectivity than the cobalt-based process. Methanol can be carbonylated even at atmospheric pressure with yields of 99 % and 90 % with respect to methanol and carbon monoxide, respectively [49]. This process was proven commercially in 1970 at Texas City, Texas. The initial plant capacity of 135 000 t/a has been expanded to 270 000 t/a since 1975 [1]. Operating conditions in the reactor are much milder (3 MPa and 180 °C) than in the BASF process [50]. Soon after the Monsanto process was commercialized, the BASF process became uncompetitive, so the Monsanto process is the preferred technology for grass-roots acetic acid units. Since the start-up of the Texas City plant by Monsanto, more than ten companies have licensed and operated this technology worldwide.

In 1978 at the Clear Lake Texas Plant, Celanese Chemical Company (now Celanese, Ltd.) was the first licensee to demonstrate commercially the rhodium-catalyzed Monsanto process. Initial capacity was 27 000 t/a [51]. In the early 1980s, Celanese developed a proprietary low-reaction-water rhodium-catalyzed methanol carbonylation process by modification of the original Monsanto high-reaction-water chemistry. Modifications to the catalyst system by the addition of inorganic iodide salts improved catalyst stability and activity significantly [52]. This technology improvement enabled the Clear Lake unit to expand by more than three times the original capacity to 900 000 t/a with minor capital cost [53].

In 1986, BP Chemicals purchased from Monsanto the high-reaction-water, low-pressure, rhodium-catalyzed methanol carbonylation technology and licensing rights. The acquisition of the technology did not include the improvements developed by Celanese Chemical Company.

Monsanto in the early 1960s also discovered that iridium, like rhodium, is an effective methanol carbonylation catalyst. This catalyst system has since been developed commercially by BP in the early 1990s and is known as the Cativa process [54]. This process was uti-

lized to convert the original rhodium-catalyzed methanol carbonylation plant in Texas City to an iridium-based process. Several advantages claimed by BP for the Cativa process over the conventional rhodium-catalyzed high reaction water carbonylation process include superior catalyst stability, operation at lower reaction water, and less liquid byproducts [55].

Chemistry and Reaction Conditions. The chemistry of the cobalt- (BASF), rhodium- (Monsanto and Celanese), and iridium-catalyzed (BP) processes is similar in requiring promotion by iodide, but the different kinetics indicate different rate-determining steps. In all three processes, two important catalytic cycles are common, one that involves the metal carbonyl catalyst and one that involves the iodide promoter [56], [57].

The cobalt-catalyzed *BASF process* uses cobalt(II) iodide [*15238-00-3*] for in situ generation of $[Co_2(CO)_8]$ and hydrogen iodide [*10034-85-2*]. Compared to other methanol carbonylation processes, severe conditions are required to give commercially acceptable reaction rates. The rate of reaction depends strongly on both the partial pressure of carbon monoxide and the methanol concentration. Acetic acid yields are 90 % based on methanol and 70 % based on carbon monoxide. A proposed mechanism for the iodide-promoted reaction is summarized in Figure 1 [43], [56], [57].

The generation of the active nucleophile, $[Co(CO)_4]^-$, can be considered a water-gas shift reaction [Eq. (4a); cf. Eq. (4b)] in which $[Co_2(CO)_8]$ is the catalyst and the hydrogen formed is dissociated via the hydridocarbonyl complex.

$$[Co_2(CO)_8] + H_2O + CO \longrightarrow 2\,[Co(H)(CO)_4] + CO_2 \quad (4a)$$

Subsequently, the methyl iodide formed from hydrogen iodide and methanol undergoes nucleophilic attack by the $[Co(CO)_4]^-$ anion. Iodide facilitates this reaction because it is a better leaving group than OH^-. The CH_3I reacts with a coordinatively saturated d^{10} complex, which is the preferred electron configuration of cobalt(I). Therefore, the methyl migration to form the acyl cobalt carbonyl complex, $[CH_3C(O)Co(CO)_3]$, is less favored than the same process for the rhodium(III) species. Once formed, this acyl intermediate cannot undergo simple reductive elimination of acetyl iodide because iodide is not coordinated to cobalt. Acetyl iodide is formed from the reaction of hydrogen iodide with the acyl complex to regenerate the cobalt carbonyl anion. Rapid hydrolysis of the acetyl iodide forms acetic acid and hydrogen iodide. All of the individual steps involved in the otherwise similar mechanism can be assumed to occur at a lower rate for cobalt than for rhodium and iridium. This explains the higher temperature needed for the BASF process. In addition, higher carbon monoxide partial pressures are required to stabilize the $[Co(CO)_4]^-$ complex at the higher reactor temperatures.

Byproducts in the BASF process are CH_4, CH_3CHO, C_2H_5OH, CO_2, C_2H_5COOH, alkyl acetates, and 2-ethyl-1-butanol [58–60]. About 3.5 % of the methanol reactant leaves the system as CH_4, 4.5 % as liquid byproducts, and 2 % is lost as off-gas. Some 10 % of the CO feed is converted to CO_2 by the water gas shift reaction (Eq. 4b).

$$CO + H_2O \longrightarrow CO_2 + H_2 \quad (4b)$$

The *Monsanto process* with rhodium carbonyl catalyst [*38255-39-9*] and iodide promoter operates under milder conditions than the BASF cobalt-catalyzed process. Methanol and carbon monoxide selectivities of greater than 99 % and 90 %, respectively, are obtained [49]. The system is not as sensitive to hydrogen as the BASF process, and therefore reduction products such as methane and propionic acid are comparatively insignificant.

The chemistry of the rhodium-catalyzed methanol carbonylation reaction has been studied in detail [57]. Kinetic studies show the reaction to be zero-order in carbon monoxide and methanol, and first-order in rhodium and iodide promoter. The carbonylation rate is strongly affected by the reaction media, but the overall kinetics are unaffected by the solvent, which suggests that it does not participate in the transition state of the rate-determining step [61–64]. Methanol carbonylation in polar solvents generally provides a rate enhancement, especially with the addition of protic solvents. An acetic acid/water solvent medium is preferred [65].

Many different rhodium compounds act as effective catalyst precursors for methanol carbonylation at common reaction temperatures of

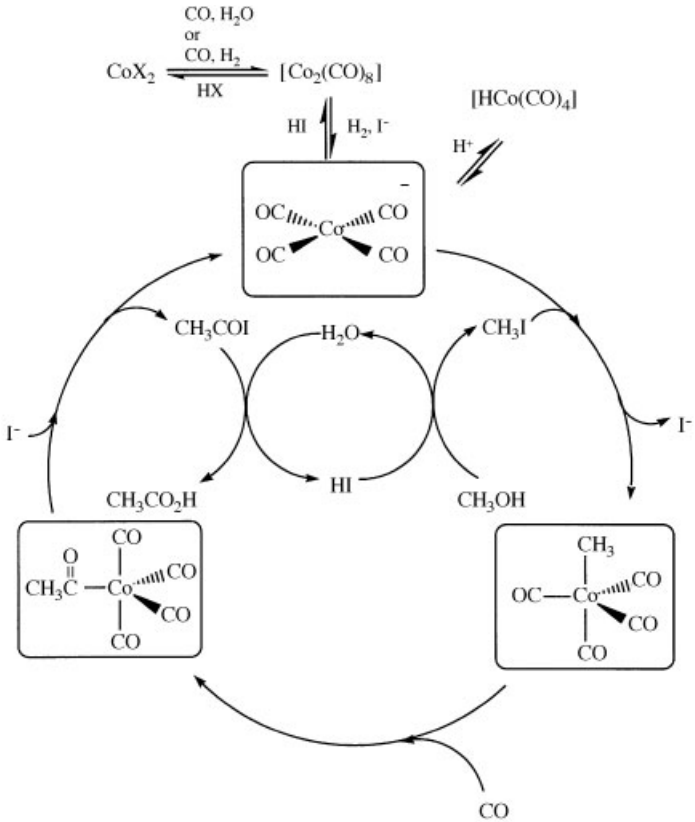

Figure 1. Reaction cycle proposed for the cobalt-catalyzed methanol carbonylation reaction (BASF process)

150 – 200 °C. The iodide promoter is normally methyl iodide, but other forms of iodide, such as hydrogen iodide or iodine, can be used without marked differences in reaction rates.

Spectroscopic investigations have shown that rhodium(III) halides can be reduced in aqueous or alcoholic media to $[Rh(CO)_2X_2]^-$ [50]. The reaction rate is independent of the rhodium precursor charged to the reaction provided adequate iodide, generally methyl iodide, and carbon monoxide are available. Under these conditions, $[Rh(CO)_2I_2]^-$ is the predominant rhodium species in the reaction solution. These observations strongly suggest the generation of $[Rh(CO)_2I_2]^-$ as the active catalyst species [50].

The catalytic cycle shown in Figure 2 is based on kinetic and spectroscopic studies [66–78]. The complex anion $[Rh(CO)_2I_2]^-$ reacts in the rate-determining step with methyl iodide by oxidative addition to form the transient methylrhodium(III) intermediate. Methyl migration gives the pentacoordinate acyl intermediate. The acyl intermediate eliminates acetyl iodide and regenerates $[Rh(CO)_2I_2]^-$. The acetyl iodide reacts with water to regenerate hydrogen iodide and produce acetic acid. Hydrogen iodide reacts with methanol to form methyl iodide. In this way both the original rhodium complex and methyl iodide promoter are regenerated.

In the 1980s *Celanese* made innovative improvements to the rhodium catalyst system of the Monsanto process with the implementation of the proprietary acetic acid optimization (AO) technology which incorporates catalyst co-promoters. Advantages of AO technology are enhanced carbonylation rates and increased carbon monoxide and methanol efficiencies at lower reaction-water concentrations [52]. These modifications represent the most significant development in rhodium-catalyzed methanol carbonylation since the development of the Monsanto process. The AO technology enabled the

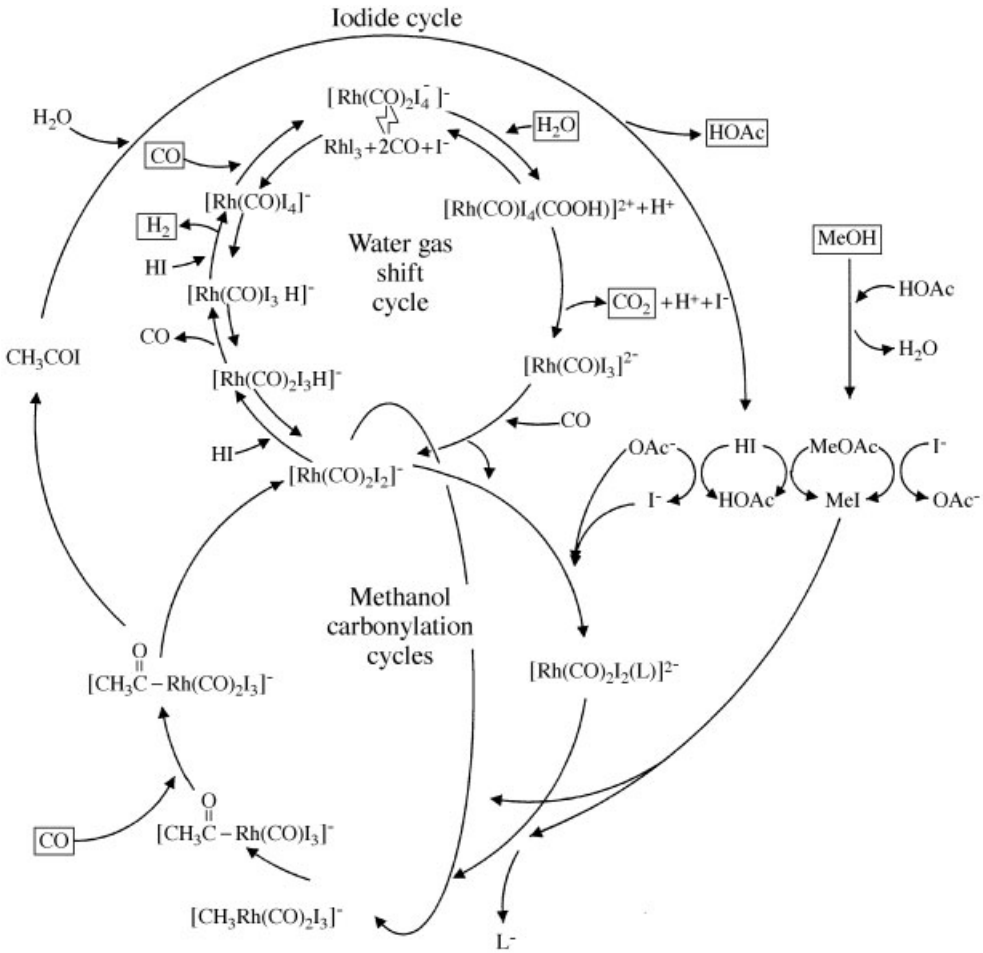

Figure 2. Reaction cycle proposed for the rhodium-catalyzed methanol carbonylation reaction (Monsanto process) and with inorganic iodide co-promotion (Celanese process)

Celanese Clear Lake Plant to more than triple unit capacity from 27 000 t/a since start-up in 1978 to 90 000 t/a acetic acid in 1998 with very low capital investment [51], [53].

The increased carbonylation rates at low reaction water concentrations are achieved by catalyst promotion with iodide and acetate anions [38], [70–72]. This unique enhancement in catalyst activity is due presumably to the generation of a strongly nucleophilic five-coordinate dianionic catalyst species, namely, $[Rh(CO)_2I_2(L)]^{2-}$ ($L = I^-$ or OAc^-), which is more active than $[Rh(CO)_2I_2]^-$ toward oxidative addition of methyl iodide. This additional reaction pathway is summarized in Figure 2.

The primary byproducts and major inefficiency with respect to carbon monoxide in the rhodium-catalyzed process are carbon dioxide and hydrogen via the water gas shift reaction (Eq. 4). The water gas shift reaction is also catalyzed by $[Rh(CO)_2I_2]^-$ [50], [79–81]. The overall two-step reaction is summarized in the following:

$$[Rh(CO)_2I_2]^- + 2\,HI \longrightarrow [Rh(CO)I_4]^- + H_2 + CO$$
$$[Rh(CO)I_4]^- + 2\,CO + H_2O \longrightarrow [Rh(CO)_2I_2]^- + CO_2 + 2\,HI$$
$$\overline{CO + H_2O \longrightarrow CO_2 + H_2}$$

The proposed kinetic pathway for the water gas shift reaction is shown in Figure 2, which emphasizes the interrelation between methanol

carbonylation and water gas shift reaction as catalyzed by $[Rh(CO)_2I_2]^-$. The addition of inorganic iodide co-promoters in the Celanese process reduces significantly the proportion of $[Rh(CO)_2I_2]^-$ that catalyzes the water gas shift reaction. As a result, the conversion of carbon monoxide and methanol to acetic acid is improved over the conventional Monsanto process, and the rate of methanol carbonylation is increased [38].

Compared to the cobalt-based process, byproducts derived from methanol such as CH_4, CH_3CHO, and C_2H_5COOH are formed in very small amounts, even in the presence of significant amounts of hydrogen in the carbon monoxide feed gas [82]. This low methanol inefficiency is associated with occurrence of the rate-determining step prior to formation of a organometallic compound, the short lifetime of the methylrhodium complex, and the rapid reductive elimination of the acylrhodium complex to form acetyl iodide, which is rapidly hydrolyzed to acetic acid and hydrogen iodide.

Similar to the rhodium-based carbonylation processes, the reaction chemistry of the BP *iridium-catalyzed methanol carbonylation* is well characterized [50], [56], [57], [78], [85–88]. The iridium-catalyzed reaction proceeds through a series of similar reaction pathways to the rhodium-catalyzed system but involves a different rate-determining step. The proposed rate-determining step is methyl migration to form the iridium acyl complex. This pathway involves the elimination of iodide and the subsequent addition of carbon monoxide. The direct dependence of the reaction rate on carbon monoxide concentration and the inhibiting effect of low concentations of iodide are consistent with the rate-determining step [55], [83]. The proposed catalytic scheme for the reaction is given in Figure 3 [55]. Model studies at 25 °C demonstrate that the oxidative addition of methyl iodide to iridium is about 120 – 150 times faster than for rhodium [55], [83]. However, methyl migration on iridium is 10^5 to 10^6 times slower than for rhodium [83]. The main byproducts of the iridium-catalyzed process are also carbon dioxide and hydrogen via the water gas shift reaction (Eq. 4b). The mechanism of this reaction is similar to the rhodium-catalyzed process [57]. The production of CH_4 derived from methanol is higher for iridium than for rhodium. This observation is consistent with the greater stability of the methyliridium(III) complex associated with the hydrogenation of the iridium – carbon bond [83].

Other transition-metal complexes have been investigated as promising catalysts for methanol carbonylation, in particular, nickel complexes by Halcon and Rhône-Poulenc [57], [89], [90].

BASF Process [91] (Fig. 4).

Carbon monoxide, methanol (containing up to 60 % dimethyl ether), catalyst recycle, catalyst makeup, and methyl iodide recycle (from the wash column) are sent to the high-pressure reactor (b; stainless steel lined with Hastelloy). Part of the relatively low heat of reaction is used to preheat the feed and the rest is ultimately dissipated through the reaction vent. The reaction product is cooled and sent to the high-pressure separator (d). The off-gas goes to the wash column (o) and the liquid is expanded to a pressure of 0.5 – 1.0 MPa in the intermediate-pressure separator (e). The gas released is also sent to the wash column; the liquid from the intermediate-pressure separator is sent to the expansion chamber (f). The gas from the chamber goes to the scrubber (p). The gas from the scrubber and the wash column is discarded as off-gas. Both scrubber and wash column use the methanol feed to recover methyl iodide and other iodine-containing volatile compounds; this methanolic solution is returned to the reactor. The off-gas composition in vol % is 65 – 75 CO, 15 – 20 CO_2, 3 – 5 CH_4, and the balance CH_3OH.

The raw acid from the expansion chamber contains 45 wt % acetic acid, 35 wt % water, and 20 wt % esters, mainly methyl acetate. The acid is purified in five distillation towers. The first column (h) degasses the crude product; the off-gas is sent to the scrubber column. The catalyst is then separated as a concentrated acetic acid solution by stripping the volatile components in the catalyst separation column (i). The acid is then dried by azeotropic distillation in the drying column (k). The overhead of the drying column contains acetic and formic acids, water, and byproducts that form an azeotrope with water. This overhead is a two-phase system that is separated in the chamber (g). Part of the organic phase, composed mainly of esters, is returned to column (k), where it functions as an azeotroping agent. The remainder of the organic phase

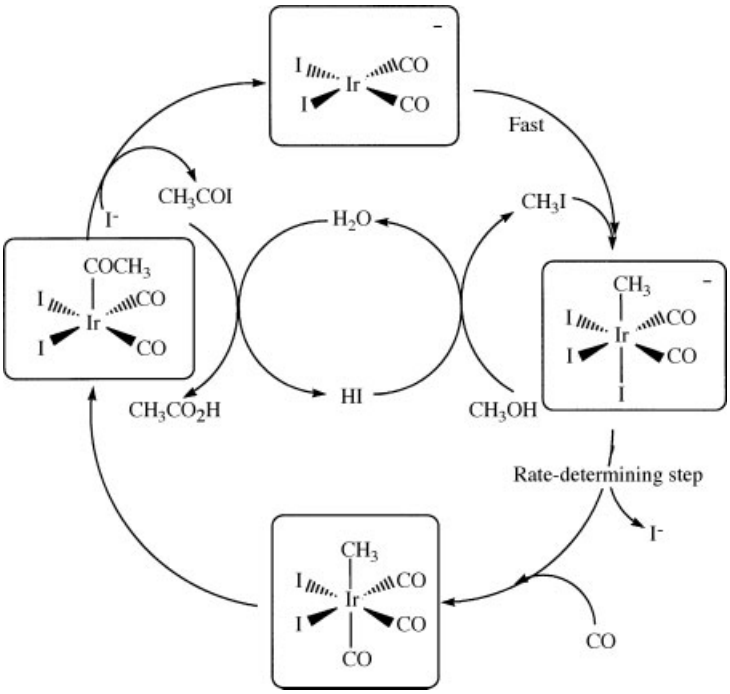

Figure 3. Reaction cycle proposed for the iridium-catalyzed methanol carbonylation reaction (Cativa process)

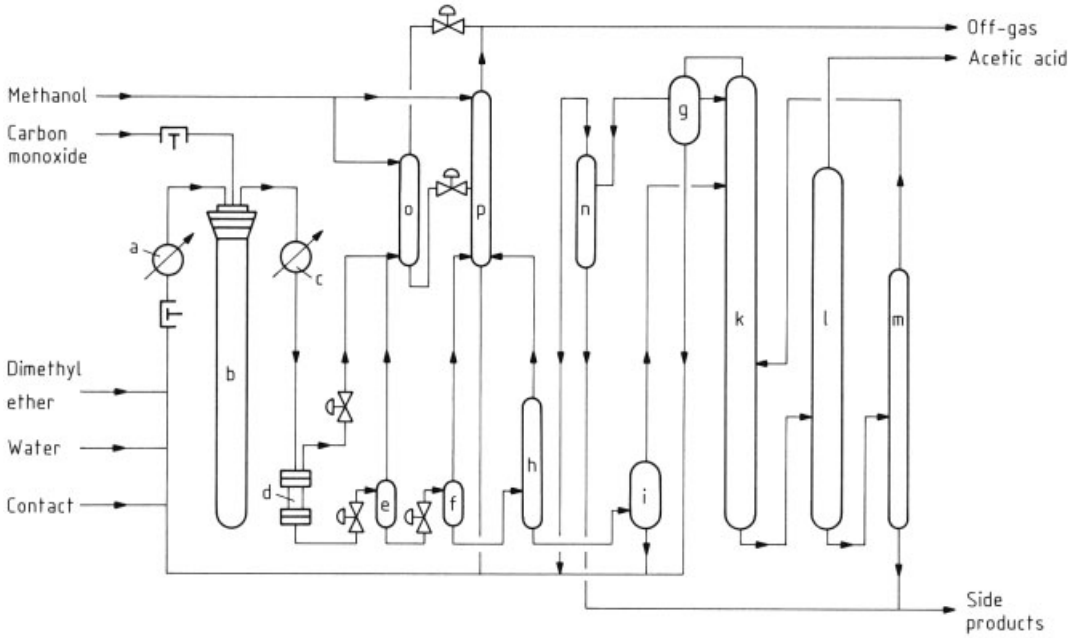

Figure 4. Production of acetic acid (BASF process)
a) Preheater; b) Reactor; c) Cooler; d) High-pressure separator; e) Intermediate pressure separator; f) Expansion chamber; g) Separation chamber; h) Degasser column; i) Catalyst separation column; k) Drying column; l) Pure acid column; m) Residue column; n) Auxiliary column; o) Wash column; p) Scrubbing column

is sent to the auxiliary column (n) where heavy ends are separated at the bottom of the column, and light esters from the overhead are recycled to the reactor. The aqueous phase and the catalyst solution are returned to the reactor. The base of the drying column is sent to a finishing column (l), in which pure acetic acid is taken overhead. The bottom stream of the finishing tower is sent to the residue column (m). The overhead of this residue column is sent back to the dehydration column. The bottom of the residue column contains about 50 wt % propionic acid, which can be recovered.

Monsanto Process [38], [83], [92] (Fig. 5).

Carbon monoxide and methanol are introduced continuously into a back-mixed liquid-phase reactor (a) at ca. 150 – 200 °C and 30 – 60 bar [38], [83], [92]. The noncondensable byproducts (CO_2, H_2, and CH_4) are vented from the reactor to control the carbon monoxide partial pressure in the reactor. The off-gas from the reactor and the purification sections of the process are combined and sent to a vent recovery system in which the light ends, including organic iodides such as methyl iodide, are scrubbed from the vent before the noncondensable gases are flared. The light ends from the vent recovery system are recycled to the reactor. The reactor solution is forwarded to the flasher (b) where the catalyst is separated as a residue stream from the crude acetic acid product and recycled to the reactor. The crude acetic acid, which contains methyl iodide, methyl acetate, and water, is taken overhead in (b) and sent to the light-ends column (c). The light components (methyl iodide, methyl acetate, and water) are recycled to the reactor as a two-phase overhead stream, while wet acetic acid is removed as a side stream from (c) and sent to the dehydration column (d). An aqueous acetic acid overhead stream from (c) is recycled to the reactor and a dry acetic acid product residue stream is forwarded to a heavy ends column (e). As a residue stream in (e), propionic acid, which is the major liquid byproduct of the process is removed with other higher boiling carboxylic acids. Product acetic acid is removed in (e) as a sidestream, and the overhead stream is recycled to the purification section of the process.

Since the Cativa process as demonstrated in the Sterling Plant at Texas City, Texas is a retrofit of the original Monsanto process, the overall general process is presumably the same, possibly with some modifications [54], [93], [94].

4.2. Direct Oxidation of Saturated Hydrocarbons

Liquid-phase oxidation (LPO) of aliphatic hydrocarbons was once practiced worldwide [95]. Due to competition from carbonylation technology, plants have reduced production by LPO significantly. The process changes depending on the availability of raw materials. Raw materials include *n*-butane and light naphtha [96]. In the United States and Canada, Celanese employs butane, while BP in the United Kingdom uses light naphtha to produce acetic acid [97]. Capacities of related production are shown in Table 5 [99].

Table 5. Butane liquid-phase oxidation processes

Company	Location	Acetic acid capacity, t/a
Celanese	Pampa, Texas	250 000 [43]
	Edmonton, Alberta, Canada	75 000 [99]
BP	Hull, England	210 000
AKZO Zout Chemie	Europoort, The Netherlands	110 000 [100]

Reaction Mechanism. Oxidation of hydrocarbons follows similar kinetics, both in the gas and liquid phases, especially in slightly polar solvents [98]. However, the mechanism of the reaction is very complicated. The reaction can be considered roughly as a radical chain reaction [99]. For example, the oxidation of butane proceeds through initiation, oxidation, propagation, and decomposition steps [96]. The initiation and propagation probably involve radicals abstracting hydrogen from a secondary carbon atom of butane. Subsequent reaction with oxygen yields hydroperoxides. These intermediates decomposes to produce acetic acid. Catalysts, agitation, and high temperature accelerate the decomposition.

Catalysts are not essential for LPO [100]. However, metal catalysts may influence the distribution of products, induction period, and operating temperature [101], [102].

In a simple mechanism, the first step of oxidation is the abstraction of a secondary hydrogen

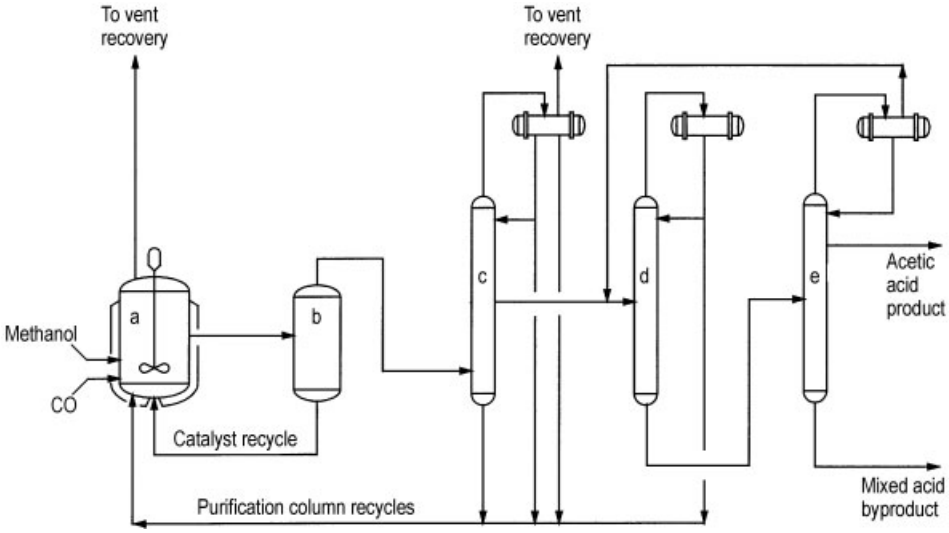

Figure 5. Production of acetic acid (Monsanto process)
a) Reactor; b) Flasher; c) Light-ends column; d) Dehydration column; e) Heavy-ends column

atom (Eq. 5) to give alkyl radicals. The oxygen in the solvent rapidly converts these radicals to *sec*-butylperoxy radicals [103] (Eq. 6). In other interpretations, oxygen is believed to directly react with one or two alkane molecules to form radicals (Eqs. 7 and 8) [104], [105], [106]. Initiation, especially with catalysts, affects the induction period significantly [107].

$$RH + In^{\bullet} \longrightarrow R^{\bullet} + HIN \tag{5}$$

$$R^{\bullet} + O_2 \longrightarrow ROO^{\bullet} \tag{6}$$

$$RH + O_2 \longrightarrow R^{\bullet} + HOO^{\bullet} \tag{7}$$

$$2\,RH + O_2 \longrightarrow 2\,R^{\bullet} + HOOH \tag{8}$$

$$ROO^{\bullet} + RH \longrightarrow ROOH + R^{\bullet} \tag{9}$$

$$ROOH \longrightarrow RO^{\bullet} + HO^{\bullet} \tag{10}$$

$$ROO^{\bullet} + ROO^{\bullet} \longrightarrow ROOOOR \tag{11}$$

$$ROO^{\bullet} + M^{n+} \longrightarrow ROO^{-} + H^{+} + M^{(n+1)+} \longrightarrow ROOH \tag{12}$$

$$ROOH + M^{n+} \longrightarrow RO^{\bullet} + OH^{-} + M^{(n+1)+} \tag{13}$$

$$ROOOOR \longrightarrow RO^{\bullet} + O_2 + RO^{\bullet} \tag{14}$$

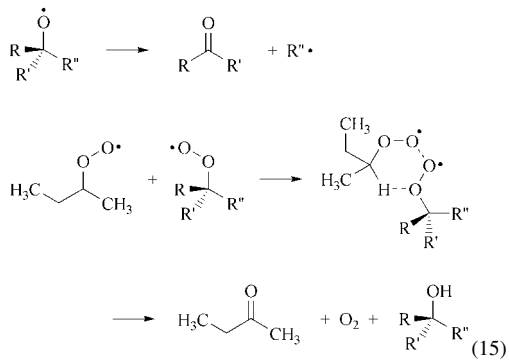

(15)

In general, peroxy radicals do not abstract hydrogen efficiently from other molecules, so the reaction medium contains an abundance of these radicals. Each radical abstracts hydrogen from an alkane to form a hydroperoxide and an alkyl radical (Eq. 9). Alkyl radicals propagate the chain (Eq. 6).

The hydroperoxide decomposes to give an alkoxy radical by thermolysis. The radical undergoes a bimolecular reaction to terminate the formation of radicals (Eq. 11). In the steady state, the termination of radicals is balanced by their production. This mechanism is too simple to explain the reaction in detail. In real systems, thermolysis may not be the only course of decomposition [108].

Alternatively, hydroperoxides may come from the complexation of metal catalysts with peroxy radicals (Eq. 12). Hydroperoxide radicals generate new radicals (Eqs. 10, 13). Therefore, the catalyst is important in maintaining a constant flux of radicals. However, in certain cases, catalysts may actually inhibit the decomposition [109].

Manganese and cobalt are common catalysts for LPO. These metals accelerate the decomposition of hydroperoxide. There is evidence that manganese may enhance the oxidation of ketone intermediates through a mechanism involving enols [110].

The peroxy radicals terminate by forming tetroxides (Eq. 11), which decompose to yield alkoxy radicals and oxygen (Eq. 14).

If the tetroxide has an α-hydrogen atom, the decomposition may follow a Russel mechanism [111]. The products are oxygen, a ketone, and an alcohol (Eq. 15). However, the mechanism is controversial. Recent reports both support [112] and challenge the mechanism [113].

Besides hydrogen abstraction, alkoxy radicals can participate in β-scission (Eq. 16). Products vary depending on the structure of the radicals. Primary, secondary, and tertiary alkoxy radicals yield alcohols, aldehydes, and ketones, respectively.

The mechanism of butane oxidation is complex. However, with detailed understanding of product distribution and rates, a mathematical model was developed [103].

About 25 % of the carbon of consumed butane appears as ethanol in the initial step as the first isolable non-peroxidic intermediate. The reaction probably involves decomposition of *sec*-butoxy radicals to form acetaldehyde and an ethyl radical (Eqs. 16, 17). Another source of acetaldehyde is the oxidation of ethanol. Acetaldehyde then reacts rapidly to produce acetic acid. Therefore, acetaldehyde is a major intermediate in butane LPO.

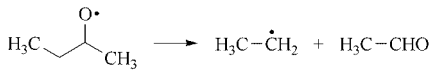

(16)

$$CH_3CH_2^{\bullet} + O_2 \longrightarrow CH_3CH_2OO^{\bullet}$$

$$2\,CH_3CH_2OO^{\bullet} \longrightarrow 2\,CH_3CH_2O^{\bullet} + O_2$$

$$CH_3CH_2O^{\bullet} + RH \longrightarrow CH_3CH_2OH + R^{\bullet} \qquad (17)$$

Besides acetaldehyde, 2-butanone or methyl ethyl ketone is another major byproduct. The ketone is the result of the termination of *sec*-butylperoxy radical by the Russel mechanism (Eq. 15).

Other impurities include propionic acid and butyric acid. *sec*-Butoxy radicals undergo β-scission to yield propionaldehyde. Oxidation of the aldehyde gives propionic acid. Butyric acid is derived from *n*-butyl radicals.

Higher paraffins are oxidized by similar mechanisms to that of butane. However, the products include shorter chain methyl ketones and difunctional intermediates. These are the results of intramolecular hydrogen abstraction. European countries usually produce acetic acid from naphtha since naphtha is cheap and available. However, the naphtha-based processes yield large amounts of impurities that increase the cost of purification.

Industrial Operation [6, 114–116] (Fig. 6). Air or oxygen-enriched air can be used as the oxidant. Multivalent metal ions, such as Mn, Co, Ni, and Cr, are used as catalysts. Some processes, however, are noncatalytic. Reaction conditions are 150 – 200 °C for a range of reaction pressures that include 5.6 MPa. The reaction pressure for naphtha oxidation is lower. The reaction solvent consists of acetic acid, varying amounts of intermediates, water, and dissolved hydrocarbons. Control of the water concentration below some maximum level appears to be critical.

The reaction section (a – d) of a liquid-phase oxidation unit consists of a sparged tubular reactor or column (a) and one or more phase separation vessels (b – d) for separating the gas and two liquid phases from the reactor. The hydrocarbon vapor exiting the reactor (a) can be recovered for recycling. Normally the reactor vapor is expanded through a turbine (b) to recover duty to compress air for use in the reactor. The consequent reduction in temperature condenses the hydrocarbon. The top organic layer from the phase separator (d) is rich in hydrocarbons and is recycled to the reactor. The bottom, aqueous layer, is distilled to recover the hydrocarbon for recycle.

The residual, hydrocarbon-free product consists of volatile, neutral oxygenated derivatives (aldehydes, ketones, esters, and alcohols),

water, volatile monocarboxylic acids (formic, acetic, propionic, and butyric from butane), and nonvolatile materials (difunctional acids, γ-butyrolactone, condensation products, catalyst residues, etc).

The volatile neutral substances can be recovered as mixtures or individually. They are used for derivatives, sold, or recycled to the reactor. Most of these components generate acetic acid on further oxidation.

The separation of water and formic acid from acetic acid involves several distillations (f). Water removal is the most difficult and costly step of the purification process. It is accomplished by azeotropic distillation with entrainment agents, such as ethers, or by extractive methods.

Formic acid is separated from the resulting anhydrous acetic acid usually by fractionation with an azeotroping agent. The remaining higher boiling acids are separated from acetic acid as a residue stream by fractionation.

In some cases, hydrocarbons such as heptane and isooctane can be added to the reactor to improve separation. These chemicals form a high-boiling azeotrope with formic acid.

In the case of naphtha oxidation, diacids such as succinic acids are isolated for sale [117]. The nonvolatile residue can be burned to recover energy.

4.3. Acetaldehyde Process

Oxidation of acetaldehyde represents a major process for making acetic acid [6], [8] (Eq. 18).

$$CH_3CHO + 1/2\,O_2 \longrightarrow CH_3COOH \qquad (18)$$

Although oxidation reactions are rather complex, a simplified view is presented here. The oxidation of acetaldehyde to acetic acid proceeds through a free-radical chain which produces peracetic acid as an intermediate.

Iniation:

(19)

Propagation:

(20)

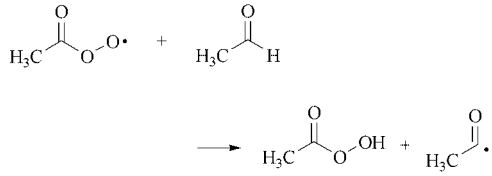

(21)

Peracetic acid reacts with acetaldehyde to generate acetaldehyde monoperacetate [*7416-48-0*]. The acetaldehyde monoperacetate decomposes efficiently to acetic acid by a hydride shift in a Baeyer – Villiger reaction. The methyl migration leads to the byproduct methyl formate [118]:

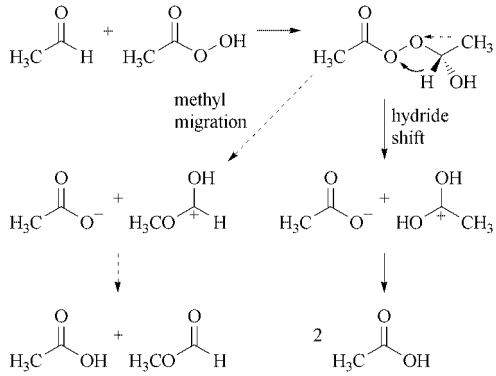

The alkyl migration becomes more pronounced with higher aldehydes, particularly aldehydes having a branch at the α-position.

Chain termination occurs primarily through bimolecular reactions of acetylperoxy radicals via an intermediate tetroxide (Eq. 22). [119].

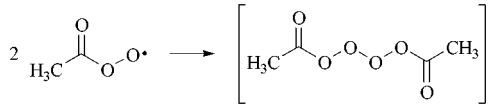

(22)

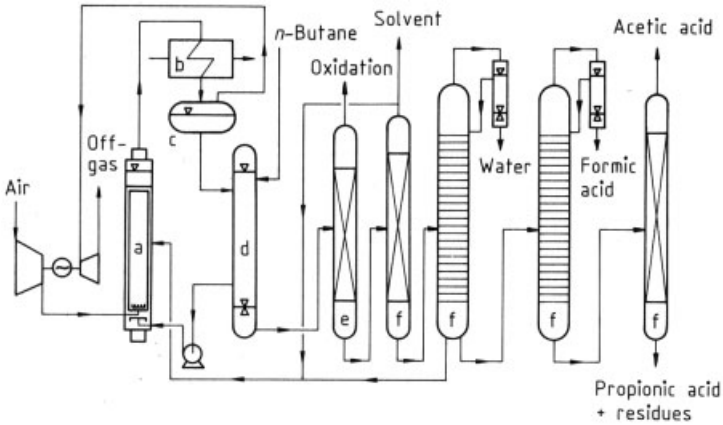

Figure 6. Oxidation of *n*-butane in the liquid phase (Chemische Werke Hüls process)
a) Reactor; b) Air cooler; c) Collector; d) Separation vessel; e) Pressure column; f) Distillation column

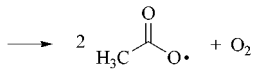

(23)

(24)

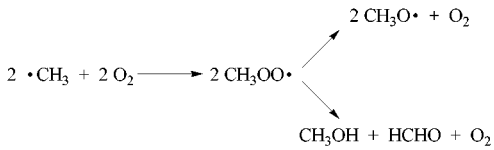

(25)

Equations (23)–(25) are the source of most of the carbon dioxide, methanol, formaldehyde, and formic acid byproducts. Uncatalyzed oxidation is efficient so long as the conversion of acetaldehyde is low and there is a significant concentration of aldehyde in the solvent. This keeps the steady state concentration of acetylperoxy radicals low and favors the Baeyer–Villiger reaction over the reactions (23)–(25). Special precautions must be taken in the uncatalyzed reaction to prevent the concentration of acetaldehyde monoperacetate from reaching explosive levels [6].

At low oxygen concentrations another free-radical decomposition reaction becomes important (Eq. 26).

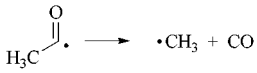

(26)

The rate of decarbonylation increases with increasing temperature. Decarbonylation becomes significant when insufficient oxygen is present to scavenge the acetyl radicals.

Catalysts can play several important roles in aldehyde oxidations [120]. Catalysts decompose peroxides and so minimize the explosion hazard (Eq. 27). In addition, manganese also reduces acetylperoxy radicals [121] directly to peroxy anions [122]:

(27)

(28)

Reaction with Mn^{2+} assists in suppressing the concentration of acetylperoxy radicals. The Mn^{3+} formed can generate the acyl radical for

the propagation step (Eq. 28), but does not contribute to inefficiency-generating reactions.

Manganese also greatly increases the rate of reaction of peracetic acid and acetaldehyde to produce acetic acid [123]. The reaction in the presence of manganese is first-order with respect to peracid, aldehyde, and manganese. In addition, the decomposition replenishes the supply of radicals. This is important since the oxidation requires a constant flux of radicals. Manganese ions increase both the rate and efficiency of oxidation.

Copper can interact synergistically with a manganese catalyst [124]. Manganese has some negative aspects associated with the fact that it greatly increases the reaction rate [125]. The increased reaction rate leads to oxygen starvation and an increasing steady state concentration of radicals. Both contribute to byproduct formation by decarbonylation and decarboxylation. However, Cu^{2+} can oxidize acetyl radicals very rapidly (Eq. 29) [126].

(29)

The acetylium ion can react as shown in Equation (30).

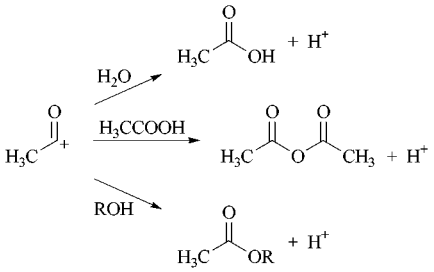

(30)

Peroxide or Mn^{3+} can reoxidize the Cu^{+}. Copper diverts a fraction of the reaction through a nonradical pathway and consequently provides a termination step that does not result in inefficiency.

Another effective catalyst is cobalt. At low concentration, this catalyst shortens the induction period [126]. The cobalt catalyst is regenerated by decomposing peracetic acid. In kinetic studies [127], CH_3CO_3 was the main free radical in the liquid phase. Other catalysts, for example phosphomolybdic acids [128], also catalyze acetaldehyde oxidation.

Additional byproducts from acetaldehyde oxidation include ethylidene diacetate, crotonic acid, and succinic acid.

Industrial Operation. A typical acetaldehyde oxidation unit is depicted in Figure 7. The reactor (a) is sparged with air or oxygen-enriched air. Temperatures are typically 60 – 80 °C with pressures of 0.3 – 1.0 MPa. The reaction mixture is circulated rapidly through an external heat exchanger to remove the heat of reaction. The vent gas is cooled and then scrubbed with recirculated crude product (which goes to the reactor) and finally with water (which goes to the aldehyde recovery column). The reactor product is fed to the aldehyde recovery column (b), from which the aldehyde is recycled, and then to a low-boilers column (c) in which methyl acetate is removed. The next column is the acetic acid finishing column (d), where water is removed overhead by azeotropic distillation and finished product comes off as a vapor sidestream. Yields are generally in excess of 90 % and purity is greater than 99 %.

4.4. Other Processes

Acetic Acid from Ethylene. Efforts to convert ethylene directly to acetic acid without going through an acetaldehyde isolation step have been of interest. Although the price of ethylene generally makes this process unattractive, some developments are noteworthy. Effective oxidation of ethylene to acetic acid has been demonstrated with metal oxides such as vanadium pentoxide. Seone et al. [129] demonstrated that the presence of palladium enhances etylene oxidation to acetic acid as temperatures as low as 230 °C. Showa Denko combined palladium with heteropolyacids and patented a catalyst for producing acetic acid from ethylene and oxygen. The catalyst contains palladium, an heteropolyacid, such as silicotungstic acid, and potassium tellurite or potassium selenite. The process is operated at 150 °C and 0.7 MPa. Although water is not consumed in the reaction, the presence of water enhances the selectivity to acetic acid and is recommended.

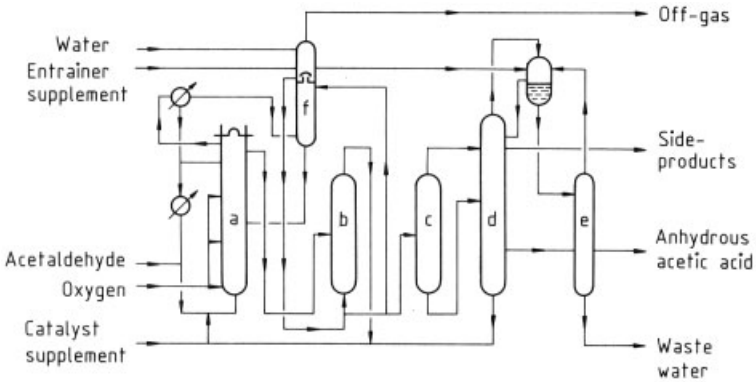

Figure 7. Oxidation of acetaldehyde to acetic acid
a) Reactor; b) Acetaldehyde column; c) Methyl acetate column; d) Finishing column; e) Column for recovering entrainer; f) Off-gas scrubber column

Acetic Acid from Ethane. Several groups have investigated catalysts for the conversion of ethane to acetic acid (Eq. 31).

$$C_2H_6 + 3/2\,O_2 \longrightarrow CH_3CO_2H + H_2O \qquad (31)$$

Union Carbide [130] developed the Ethoxene process for the production of ethylene from ethane and oxygen. The earliest catalysts consisted of molybdenum, vanadium, and niobium oxides and were very selective for ethylene at temperatures below 300 °C. Unfortunately, ethane conversions were low (about 10 %) due to inhibition by the product. Further development [131] focused on the coproduction of ethylene and acetic acid with a catalyst comprised of molybdenum, vanadium, niobium, calcium, and antimony. The addition of water enhances acetic acid formation. A later patent [132] discloses that the addition of water and an ethylene hydration catalyst improves selectivity to acetic acid.

Rhône-Poulenc [133], [134] has patented a process to make acetic acid from ethane with a vanadium oxide or vanadyl pyrophosphate supported on titanium dioxide. Tessier et al. [135] found that acetic acid production was favored over ethylene and carbon oxides at temperatures below 300 °C. Roy et al. [136] demonstrated that the addition of molybdenum enhanced acetic acid selectivity. Desorption of acetic acid from the catalyst surface was speculated to be the rate-determining step of the reaction.

BP Chemicals [137], [138] claims that the addition of rhenium to mixed metal oxide catalysts enhances selectivity to acetic acid from the oxidation of ethane and/or ethylene. Again, water is used to enhance selectivity to acetic acid. Standard Oil [139] claims good selectivity to acetic acid with a vanadyl pyrophosphate catalyst containing a transition metal. Hoechst [140] claims a catalyst containing molybdenum and palladium as effective for the production of acetic acid from ethane. An acetic acid selectivity of 84 % was attained at 250 °C and 7 bar with a 14-s residence time.

Although ethane is an inexpensive raw material and high selectivities to acetic acid have been achieved, it is unlikely that ethane oxidation will compete with methanol carbonylation in the near future. The oxygen concentration must be limited for safety reasons and therefore ethane conversion is limited per pass in the reactor. Although staged addition of oxygen is possible, product inhibition remains a problem and limits ethane conversion. Since the addition of water is needed to improve selectivity to acetic acid, water must be removed from the acetic acid, usually by extractive distillation.

Acetic Acid from Microorganisms Since about 10 000 B.C., aqueous solutions of acetic acid have been prepared from spoiled wine [141], [142]. Ethanol and sugar were the primary feedstocks for microorganism production of acetic acid, although biomass has been proposed [143]. The concentration of acetic acid in solution is limited by the ability of bacteria to thrive in low-pH solutions. Consequently,

researchers have focused on improving acetic acid productivity and lowering the pH of the bacterial culture. Mutant strains of *Clostridium thermoaceticum* [144] have been developed to produce acetic acid in solutions below pH 5. Improvement of acetic acid productivity of an *Acetobacter aceti* strain was achieved by amplification of the aldehyde dehydrogenase gene with a multicopy vector [145]. Bacterial production is gaining interest since it is a environmentally friendly process. Nevertheless, cost-effective acetic acid concentration and purification remain a challenge.

4.5. Concentration and Purification

In most commercial processes acetic acid is purified from an aqueous acetic acid solution by straight distillation methods that are economically viable provided the acetic acid is in high concentrations when initially isolated from the reaction system of the process. Other acetic acid purification techniques can be used, in particular, for dilute acetic acid such as byproduct streams of important industrial processes, e.g., cellulose acetate production (→ Cellulose Esters). Azeotropic distillation, solvent extraction, and extractive distillation can recover acetic acid from these aqueous streams. Direct distillation requires a significant number of plates and a high reflux ratio in the columns. In extractive distillation, the vapor streams of acetic acid and water are scrubbed with a high-boiling solvent that preferentially dissolves one of the components.

Azeotropic Distillation. In azeotropic distillation, a compound that decreases the boiling point of water is used. If the azeotropic agent is immiscible with water, it can be separated easily from water and recycled. Esters, ethers, benzene, and chlorinated hydrocarbons have been proposed as azeotropic agents [146].

Solvent Extraction. Acetic acid is removed from aqueous solutions (0.5 – 5 wt % acetic acid) by solvent extraction. The aqueous solution is contacted with an organic solvent, often containing an organic base, and then acetic acid is stripped from the base in a distillation process and separated from the cosolvent, usually an alkane, by phase separation. Consequently, it is important for the organic solvent to have a high affinity for acetic acid relative to water. Primary amines are generally not suitable extraction additives due to their water solubility. Secondary amines have high extraction coefficients [147] but can form amides during the acid recovery process. Tertiary amines such as trioctylamine [148] have extraction coefficients which are strongly dependent on amine concentration and the cosolvent used. Eyal and Canari [149] discuss how anion exchange, ion-pair formation, hydrogen-bond formation, and solvation all play important roles in amine-based extraction of acetic acid. In addition to amines, dibutyl phosphonate [150] and trioctyl phosphine oxide [151] have high extraction coefficients relative to alcohols, ethers, or esters.

The selection of a water-removal method depends on economics. Extraction is generally preferred at low concentrations of acetic acid.

4.6. Construction Materials [152], [153]

Little is known about construction materials for equipment in acetic acid processes. The information on materials is usually considered a trade secret. Liquid streams containing acetic acid corrode metals. The corrosion increases with increasing concentrations of acetic acid, halides, and formic acid. High temperatures (> 200 °C) also increase the corrosion rate. AISI 316 stainless steel [*12597-68-1*] is used for processes without reducing agents and halides. In the presence of reducing agents, AISI 321 stainless steel is necessary. Hastelloy B or C or other materials are required if halides are present. Aluminum [*7429-90-5*] and AISI 304 stainless steel are used in storage tanks, pumps, and piping at room temperature and high acid concentrations. AISI 316 stainless steel contains 16 – 18 wt % Cr, 10 – 14 wt % Ni and 2 – 3 wt % Mo. AISI 321 stainless steel has 17 – 19 wt % Cr, 9 – 12 wt % Ni and > 0.4 wt % Ti. AISI 304 stainless steel is comprised of 18 – 20 wt % Cr and 8 – 10.5 wt % Ni.

5. Wastewater and Off-Gas Problems [154]

Waste streams from acetic acid process usually contain residual acetic and formic acids.

These acids must be recycled, diluted, or treated chemically and then degraded biologically. European countries require neutralization or biological degradation of these waste streams [6], [8].

Acetic acid is corrosive and can damage human skin and internal organs. Acetic acid should be handled with care to avoid breathing vapors. If a spill occurs, neutralize soil and wash with soda ash or lime. The Clean Water Act (USA) and Superfund (USA) dictate that a reportable quality (RQ) of 2270 kg (5000 lb) [155].

In Germany [156], no more than 3 kg/h may be emitted in vent gases, with a maximum of 150 mg/m^3. The vent gas is scrubbed or chilled to eliminate the acids.

Under the same German regulations, the maximum emissions for methyl acetate, ethyl acetate, butyl acetate, and phenyl acetate are 300 mg/m^3. The off-gas is limited to 6 kg/h [8]. The recommended methods of removal of the acetic acid are thermal condensation, washing with nonvolatile organic solvent, physical absorption or burning.

In the United States, the United States Clean Air Act governs the amount of emissions. Each state also sets individual emission limits. The limits range from 0.25 mg/m^3 to 0.37 mg/m^3 (North Dakota), to 0.4 mg/m^3 (Virginia) to 0.5 mg/m^3 (Connecticut and South Dakota) to 0.595 mg/m^3 (North Carolina) [154].

6. Quality Specifications

There is essentially one commercial grade of acetic acid sold in the United States. Elsewhere, three or more grades are available differing as to specifications for heavy metals, chloride, sulfate, arenic, and iron [157]. Also, depending on specific end uses, additional product criteria must be considered. For example, in the manufacture of vinyl acetate, halides are important since they are a catalyst poison in vinyl acetate processes. One of the most critical specifications for all end uses of acetic acid is "permanganate time." The test for "permanganate time" is a semi-quantitative wet method for the measurement of oxidizable impurities. One manufacturer's sales specification is given Table 6 [76].

Table 6. Sales specifications for glacial acetic acid [76]

Sales specifications	Limits
Acetic acid, wt %	min. 99.85
Distillation range (including 118.1), °C	max. 1.0
Freezing point, °C	min. 16.35
Color, Pt – Co units	max. 10
Water content, wt %	max. 0.15
Reducing substances, as formic acid, wt %	max. 0.05
Aldehydes, as acetaldehyde, wt %	max. 0.05
Iron, ppm	max. 1
Heavy metals, as Pb, ppm	max. 0.05
Chlorides, ppm	max. 1
Sulfates, ppm	max. 1
Sulfurous acid, ppm	max. 1
Permanganate time	min. 2 h

7. Analysis

Gas chromatography is the usual technique for determining the acetic acid content of a volatile mixture. Packed columns can be used, but capillary columns are preferred. Carbowax 20M, terephthalic acid terminated [*41479-14-5*], is the preferred substrate.

8. Storage, Transportation, and Customs Regulations [8], [152], [159]

DOT regulations regard glacial acetic acid (99.7 %) as a hazard. The acid belongs to UN/DOT Hazard Class 8 and Shipping Group II [160]. Acetic acid should be stored in a cool, dry place away from contact with oxidizing materials. Containers should be kept closed and labeled as "Corrosive". Precautionary labels such as "Flammable" are recommended. Even when emptied, the containers may still retain acid residue. Wash the container with water. The emptied container should still have a warning label.

The acid can be stored and transported in containers lined with stainless steel, glass, or polyethylene. Another material for containers is aluminum. Aluminum forms a protective coating of aluminum oxide or aluminum acetate when exposed to acetic acid. The coating prevents further corrosion. Caution is warranted to prevent water and mercury contamination. Water and small amounts of mercury accelerate the

corrosion. Maintaining a high acid concentration and not using a mercury thermometer help to prevent catastrophic corrosion.

The limits for shipping concentrated acetic acid in passenger aircraft or railcar is 1 L, and 30 L in cargo aircraft. Acetic acid has a low melting point of 16 °C. Transporting the acid in cold weather requires the vessels to be heated. Storage containers, tank trucks, and pipes for concentrated acid should be equipped with heated coils. Storage vents also need to be heat traced to prevent the acid from crystallizing.

German Legal Requirement for Foodstuffs. Commercial vinegar exceeding 11 wt % acetic acid may be sold in closed containers. The material of the containers should be resistant to acetic acid. The container must be clearly marked with a label warning as "Handle with Care.", "Do Not Swallow Undiluted." Solutions containing more than 25 wt % acetic acid are not available to the public. Commercial dealers are allowed to handle the acid.

German Customs Requirements (Vinegar Law) [5]. The Vinegar Law, Paragraph 1, dictates that the transfer of synthetic vinegar in retail markets is taxable. The tax is known as the Vinegar Tax. Both imported and local products are subjected to the tax. With changing of vinegar price, the tax is adjusted.

Tax exempt status is granted for acetic acid used in commercial purposes. Another case for tax exempt status also deals with commercial uses where the acid is consumable by humans.

9. Uses

Acetic acid has a broad spectrum of applications. An outline of these is given in Figure 8 [1], [48]. Greater than 65 % of the acetic acid produced worldwide goes into polymers derived from vinyl acetate (→ Vinyl Esters), or cellulose (→ Cellulose Esters). Most of the poly(vinyl acetate) is used in paints and coatings or for making poly(vinyl alcohol) and plastics. Cellulose acetate is used to produce acetate fibers. Acetic acid and acetate esters are used extensively as solvents.

10. Derivatives

This section contains information about acetic acid derivatives that are not discussed in detail elsewhere in this encyclopedia (→ Acetic Anhydride, → Cellulose Esters, → Chloroacetic Acids, → Esters, Organic, → Vinyl Esters).

10.1. Salts

10.1.1. Aluminum Acetate

Three aluminum acetates are known, two basic and one neutral: $Al(OH)_2CO_2CH_3$ [*24261-30-1*], $Al(OH)(CO_2CH_3)_2$ [*142-03-0*], and $Al(CO_2CH_3)_3$ [*139-12-8*].

Properties. $Al(CO_2CH_3)_3$, M_r 204.1, is a white, water-soluble powder that decomposes at about 130 °C to give acetic anhydride along with basic aluminum acetates [161].

Production. Neutral aluminum acetate is made from aluminum metal and glacial acetic acid that contains 0.1 – 1 % acetic anhydride, which scavenges any water present [162]. The anhydrous triacetate cannot be made from aqueous solutions. The basic acetates can be prepared from $Al(OH)_3$ and aqueous acetic acid solutions [163].

Uses. Aluminum acetate is used as a dye mordant in the dyeing of fabrics.

10.1.2. Ammonium Acetate

Properties. Ammonium acetate [*631-61-8*]. $CH_3CO_2NH_4$, M_r 77.08, *mp* 114 °C, forms colorless, hygroscopic needles. The solubilities of this salt in 100 g of water or methyl alcohol are, respectively, 148 g (4 °C) and 7.9 g (15 °C).

Production. Ammonium acetate is manufactured by neutralizing acetic acid with ammonium carbonate or by passing ammonia gas into glacial acetic acid [114]. Acidic ammonium acetate, $CH_3CO_2NH_4 \cdot CH_3CO_2H$ [*25007-86-7*], is manufactured by dissolving the neutral salt in acetic acid.

Uses. Ammonium acetate is used in the manufacture of acetamide and as a diuretic and diaphoretic in medical applications. The wool industry also uses this salt as a dye mordant.

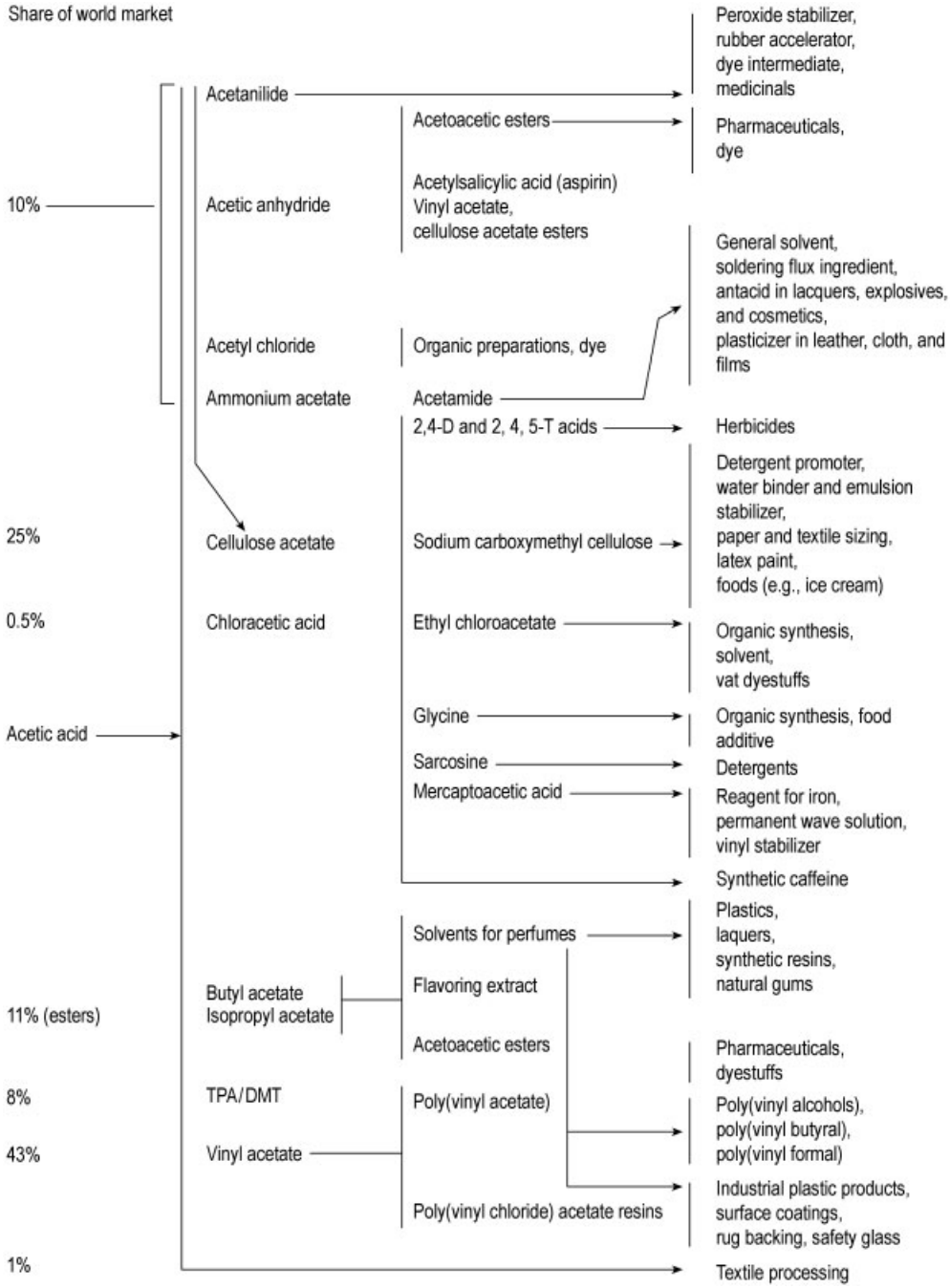

Figure 8. Uses of acetic acid

10.1.3. Alkali Metal Salts

Properties. All of the alkali metals (Li, Na, K, Rb, Cs) form acetates. The aqueous solubilities of these salts increase with increasing atomic mass of the alkali metal [164]. The potassium salt [*127-08-2*]. M_r 98.14, *mp* 292°C, and sodium salt [*127-09-3*], M_r 82.04, *mp* 324°C, are the most common. The former crystallizes from water as white columns. The latter can be purified by crystallization from acetic acid. Several acetic acid solvates of the alkali metal salts are known [163].

Production. Potassium and sodium acetates normally are manufactured from glacial acetic acid and the corresponding hydroxides. The salts also can be prepared from acetic acid on treatment with the metal, the metal carbonate, or the metal hydride [163].

Uses. Potassium acetate is used to purify penicillin. It is also used as a diuretic and as a catalyst to make polyurethane. The sodium salt is used frequently in water as a mild alkali.

10.2. Esters (→ Esters, Organic) [166–169]

Properties. Physical properties of acetate esters are listed in Table 7 [168].

Production. The reaction of an alcohol and an organic acid produces an ester in the presence of an acid catalyst [166]. The reaction, esterification, is the most common process for manufacturing acetate esters (Eq. 32).

$$ROH + CH_3COOH \rightleftharpoons CH_3COOR + H_2O \tag{32}$$

Other industrial processes include the reaction of acetic anhydride with alcohols and the Tishchenko reaction. Other methods for synthesizing acetate esters are the addition of acetic acid to olefins, carbonylation of acetic acid or anhydride, and reductive methods involving carboxylic acids or anhydride.

Commercial processes for esterification are usually continuous and are performed in the liquid phase. The reaction requires strong acids such as sulfuric acid or organic sulfonic acids. Another type of catalyst is ion-exchange resins. In contrast, gas-phase technology is available with sulfonated polymer [170] or supported phosphoric acid as catalyst [171].

Esterification involves an equilibrium between reactants and product (Eq. 33). Adding excess reactants or removal of water drive the equilibrium to favor the formation of ester. Azeotroping agents eliminate water and force the reaction to completion. Common agents are cyclohexane, benzene, and toluene.

$$K_{eq} = \frac{[CH_3COOR]\,[H_2O]}{[CH_3COOH]\,[ROH]} \tag{33}$$

Esterification with acetic anhydride offers an alternative route for making acetates. Due to its high reactivity, acetic anhydride readily forms esters with alcohols. The reaction is irreversible. The esterification goes to completion without eliminating products or water.

Additions of acetic acid to olefins present an alternative way to make acetates, especially for branched esters. Synthesis of these hindered esters by esterification is not efficient. The equilibrium does not favor the desirable products. Rhône-Poulenc uses this method to prepare isopropyl acetate from acetic acid and propylene [172]. Catalysts can be strong protic or Lewis acids, like sulfuric acid or $BF_3 \cdot OEt_2$ [173]. Ethyl acetate can also be produced directly from acetaldehyde.

Adding hydrogen and carbon monoxide to acetic acid or acetic anhydride can produce acetates. Alternatively, hydrogenation of acetic acid yields ethanol and ethyl acetate. Many studies relate to supported heterogeneous catalysts with molybdenum and group VIII metals (Pd, Rh, or Ru). Additionally, homogeneous catalysis with ruthenium can result in high selectivity [172].

Commercial processes also include the reaction of aldehdyes to make esters. The Tischenko reaction uses alkoxide catalysts. Typical catalysts are aluminum alkoxide, boric acid and magnesium or calcium aluminum alkoxides [173].

Uses. Applications of acetates depend on their physical properties such as boiling point and evaporation rate. Therefore, uses of acetates cover many products such as coatings, solvents for plastics, lacquers, resins, and gums.

Table 7. Physical properties of acetate esters

Ester	CAS no.	M_r	n_D^{20}	d_{20}^{20}	*bp*, °C	*fp*, °C	Flash pt., °C
Methyl acetate	[*141-78-6*]	74.1	1.36	0.93	57.0	−98.1	−10
Ethyl acetate	[*108-05-4*]	88.1	1.37	0.902	77.1	−83.6	− 4
Propyl acetate	[*109-60-4*]	102.13	1.38	0.887	101.6	−92.5	13
Isopropyl acetate	[*108-21-4*]	102.13	1.38	0.872	90.1	−73.4	2
Butyl acetate	[*123-86-4*]	116.16	1.39	0.882	126.0	−73.5	22
Isobutyl acetate	[*110-19-0*]	116.16	1.39	0.871	117.2	−98.6	18
sec-Butyl acetate	[*105-46-4*]	116.16	1.39	0.876	112.0		31
tert-Butyl acetate	[*540-88-5*]	116.16	1.39	0.867	97.0		
n-Pentyl acetate	[*628-63-7*]	130.18	1.40	0.876	149.2	−70.8	
2-Ethylhexyl acetate	[*103-09-3*]	172.26	1.42	0.873	199.3	−93.0	71

10.2.1. Methyl Acetate

Properties. Methyl acetate is colorless and has a pleasant odor. Methyl acetate forms azeotropes with water (3.5 wt % H_2O, *bp* 56.5 °C) and methanol (19 wt % MeOH, *bp* 54.0 °C). Other properties are listed in Table 7.

Production. Most commercially available methyl acetate is a byproduct in the manufacture of acetic acid. Another method is the esterification of methanol and acetic acid with sulfuric acid as catalyst. The product ester is removed as the methanol/methyl acetate azeotrope.

Uses. Methyl acetate has the lowest boiling point (55.8 – 58.2 °C) of all organic acetates. It evaporates about five times faster than butyl acetate [166]. Therefore, methyl acetate finds uses where rapid evaporation is necessary. It is employed as a process solvent for preparing cellulose nitrates. Transesterification to make esters uses methyl acetate as a reagent. In addition, the production of acetic anhydride requires the acetate.

10.2.2. Ethyl Acetate

Properties. Ethyl acetate has a pleasant odor. It forms azeotropes with water (8.2 wt % H_2O, *bp* 70.4 °C), ethanol (30.8 wt % EtOH, *bp* 71.8 °C), and methanol (44 wt % MeOH, bp 60.2 °C). With water and ethanol, the ester forms a ternary azeotrope (7.8 wt % H_2O, 9.0 wt % EtOH, *bp* 70.3 °C). Other properties are listed in Table 7.

Production. Ethyl acetate is made from ethanol [*64-17-5*] and acetic acid in batch or continuous processes. A major step in the continuous process is the removal of the ester. The distillation process separates the ester by taking advantage of the alcohol – ester – water azeotrope. After decanting and purification, the final yield of the ester can reach as high as 95 %.

Another means of making ethyl acetate is theTischenko reaction. Two moles of acetaldehyde can undergo bimolecular reaction to yield ethyl acetate with a metal alkoxide catalyst at low temperature (Eq. 34). A popular catalyst is aluminum alkoxide.

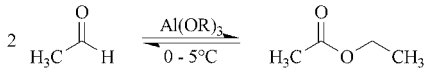

(34)

Uses. Ethyl acetate is a common solvent for coatings, adhesives, ink, cosmetics, and film base. Specialty chemicals and pharmaceutical industries use ethyl acetate as a solvent for synthesis. Coffee is occasionally decaffeinated with ethyl acetate in an extraction process.

10.2.3. Butyl Acetate

Properties. All four of the esters, *n*-butyl [*123-86-4*], *sec*-butyl [*105-46-4*], isobutyl [*110-19-0*], and *tert*-butyl acetate [*540-88-5*], are colorless liquids with pleasant odors. Azeotropic data for the butyl acetates are given in Tables 8 and 9 [5]. Other properties are listed in Table 7.

Table 8. Binary azeotropes with the butyl acetates

Component A	Component B	*bp*, °C	wt % B
n-Butyl acetate	Water	90.2	26.7
sec-Butyl acetate	Water	86.6	19.4
Isobutyl acetate	Water	87.4	–
n-Butyl acetate	*n*-Butanol	117.2	47
sec-Butyl acetate	*sec*-Butanol	99.6	86.3

Table 9. Ternary azeotropes with the butyl acetates

Component A	Component B	Component C	*bp*, °C	wt % B	wt % C
n-Butyl acetate	Water	*n*-Butanol	89.4	37.3	37.3
sec-Butyl acetate	Water	2-Butanol	86.0	23.0	45.0
Isobutyl acetate	Water	Isobutyl alcohol	86.8	30.4	23.1

Production. Esterification of butanol and acetic acid produces butyl acetate. The catalyst is sulfuric acid. The removal of water by azeotropic separation forces the reaction to completion. The acid catalyst is neutralized and the ester purified by distillation [8].

Uses. Butyl acetate is used primarily in coatings. It is a good solvent with low volatility. The lacquer and enamel industries employ butyl acetate as a solvent for making acylic polymers, vinyl resins, and nitrocellulose. Other industries utilizing the acetate as solvent include printing, photographic, adhesives, extraction and pharmaceutical.

10.2.4. 2-Ethylhexyl Acetate

2-Ethylhexyl acetate is prepared by esterification of 2-ethyl-1-hexanol with acetic acid with the aid of an acid catalyst. Distillation with a low-boiling azeotropic solvent removes water from the reaction. The ester is used in lacquers, in silk-screen inks, and in paints as a coalescing agent. Properties are listed in Table 7.

10.2.5. Other Esters

Properties of some other esters are given in Table 7. Most esters are colorless liquids with a mild fruity smell. The esters are used as solvents for making cellulose and elastomers [178].

10.3. Acetyl Chloride

Properties. Acetyl chloride [*75-36-5*] M_r 78.5, n_D^{20} 1.3871, d_4^{20} 1.1051, *bp* 51.8°C, *mp* −112.86 °C is a colorless, corrosive liquid. Acetyl chloride reacts with moist air to produce hydrogen chloride and acetic acid.

Production. The normal industrial method involves reaction of acetic anhydride [*108-24-7*] with anhydrous hydrogen chloride [*7647-01-0*] [179], [180]. On the laboratory scale, it can be produced from acetic acid and reagents such as thionyl chloride [*7719-09-7*], phosphorus trichloride [*7719-12-2*], or phosphorus pentachloride [*10026-13-8*] [181].

Acetyl chloride is normally consumed at the site of generation since transportation and storage are difficult. Glass, enamel, porcelain, clay, and polytetrafluoroethylene are suitable materials for handling acetyl chloride.

Uses. Acetyl chloride is an efficient acetylating agent for alcohols and amines to produce esters and amides. It is important in the synthesis of dyes and pharmaceuticals. Acetyl chloride is used in the Friedel – Craft acylation of benzene to yield acetophenone.

10.4. Amides

10.4.1. Acetamide

Properties. Acetamide [*60-35-5*], CH_3CONH_2, M_r 59.07, *mp* 81.5 °C, *bp* 222°C, d^{20} 1.161, forms deliquescent hexagonal crystals that are odorless when pure. Acetamide is soluble in water and low molecular mass alcohols.

Production. Acetamide can be produced by several methods. Ammonia reacts vigorously with acetyl halides or acetic anhydride to produce acetamide. Alkyl acetates also react with ammonia to yield acetamide. Dehydration of ammonium acetate is the standard procedure for preparation of acetamide. Finally, acetonitrile is hydrolyzed to yield acetamide in the presence of an acid or base catalyst. Metal catalysts [183], [184] and biocatalysts [185] are claimed to catalyze the hydrolysis of acetonitrile. This method of acetamide production is popular because acetonitrile is a byproduct of acrylonitrile production.

Uses. Acetamide is used as a solvent and as a plasticizer.

10.4.2. *N,N*-Dimethylacetamide

Properties. *N,N*-Dimethylacetamide, [*127-19-5*], $CH_3CON(CH_3)_2$, M_r 87.1, *mp* −20 °C, *bp* 165.5 °C, d^{20} 0.943, is a colorless liquid which is soluble in a wide variety of solvents.

Production. *N,N*-Dimethylacetamide is prepared by reaction of dimethylamine with acetic acid, acetic anhydride, or acetate esters. Heating dimethylamine acetate with [186], [187] or without [188] a catalyst affords *N,N*-dimethylacetamide. Reaction of dimethylamine with acetate esters requires a catalyst; sodium methoxide is typically used.

Uses. *N,N*-dimethylacetamide is an excellent solvent and often acts as a catalyst in halogenation, cyclization, and alkylation reactions. *N,N*-Dimethylacetamide is used in the spinning of polyacrylonitrile and as a solvent in the processing of other polymers. *N,N*-dimethylacetamide lithium chelate complexes intercalate cationic sites in layered silicates [189].

10.5. Phenylacetic Acid

Properties. Phenylacetic acid [*103-82-2*], $C_6H_5CH_2CO_2H$, M_r 136.14, *bp* 265.5 °C, *mp* 78 °C, $d^{79.8}$ 1.0809, forms white leaflets on vacuum distillation. It dissolves in hot water but is only slightly soluble in cold water.

Production. Hydrolysis of benzyl nitrile or carbonylation [190] of benzyl halides or alcohols produces phenylacetic acid [191].

Uses. Penylacetic acid is a starting material in the synthesis of synthetic perfumes and cosmetics [192]. It is used in the synthesis of penicillin G to stimulate benzylpenicillinase in *E. Coli*. Penicillin G, after extraction from the culture, is converted to 6-aminopenicillanic acid and phenylacetic acid.

11. Economic Aspects [1]

World acetic acid production between 1998 and 2006 is forecast to grow by an average of 3 – 4 % per annum, while world acetic acid capacity will grow by an average of 4 % per annum. In 1996 total world acetic acid capacity was ca. 7.1×10^6 t/a and total world production was ca. 5.5×10^6 t/a.

In 1996, North America was the region with the greatest acetic acid production at ca. 36 % of the world total. Western Europe was second with ca. 24 % followed by Japan at 15 % and Eastern Asia with 14 %. By 2006 the share of acetic acid production in North America, West Europe, and Japan is expected to drop to ca. 32 %, 17 %, and 11 %, respectively, while East Asian production will increase to 18 %. Consumption and production of acetic acid in Asia are forecast to grow strongly from 1998 to 2006.

The largest end uses of acetic acid are for the manufacture of vinyl acetate, acetic anhydride, acetate esters, monochloroacetic acid, and as a solvent in the production of dimethyl terephthalate and terephthalic acid. Figure 8 provides a detailed summary of acetic acid end uses. Vinyl acetate is by far the most significant end use for acetic acid. About 43 % of world production of acetic acid in 1996 was used in the manufacture of vinyl acetate. About 12 – 15 % of world acetic acid production is used for each of the remaining major end uses of acetic acid.

There is little doubt that the technology of the future is methanol carbonylation. In 1972, low-pressure methanol carbonylation in the United States accounted for only 10 % of the total acetic acid capacity, whereas in 1982 the share had increased to 40 %. By 1993, 46 % of the acetic acid was produced by methanol carbonylation. Over 90 % of all new acetic acid is manufactured by low-pressure methanol carbonylation technology.

BP Chemicals purchased the technology and license rights to the Monsanto methanol carbonylation process in 1986 and has continued to license in joint ventures this high-reaction-water process. Technology improvements have been made in the rhodium-catalyzed carbonylation process since this process was first implemented commercially [52]. Celanese in the early 1980s departed from the Monsanto high-water technology to develop a proprietary low-reaction-water process. BP also has developed a low-reaction-water technology based on iridium called Cativa

that has several advantages over the rhodium-catalyzed high-reaction-water process [55] originally developed by Monsanto. These developments, as well as others, will most likely ensure that in the future the low-pressure methanol carbonylation processes will remain the most economical routes to manufacture acetic acid.

Acetic acid from low-pressure methanol carbonylation is gradually replacing the more expensive acetic acid produced by the oxidation of acetaldehyde, butane, and naphtha and by high-pressure methanol carbonylation. The oxidation routes to acetic acid in most cases are not competitive with low-pressure methanol carbonylation. Acetaldehyde oxidation units are only economically viable provided an acetic acid producer has a captive source of acetaldehyde. Similarly, *n*-butane or naphtha oxidation processes are not competitive with low-pressure methanol carbonylation unless byproduct credits are considered. As a result, many high-cost acetic acid units have ceased operation, such as the acetaldehyde oxidation processes of Celanese in Clear Lake and Bay City Texas, and a similar process of Eastman Chemical in Kingsport, Texas. Additionally, the butane oxidation plant operated by Union Carbide at Brownsville, Texas, and the high-pressure methanol carbonylation plant of Borden in Geismar, Louisiana, have ceased operation.

12. Toxicology and Occupational Health

Acetic acid has a pK_a of 4.5. The acid is not as strong as mineral acids or formic acid. However, its corrosiveness justifies caution in handling.

Table 10 lists exposure limits for acetic acid and derivatives [194].

When humans ingest acetates, the esters are hydrolyzed readily to the corresponding alcohols and carboxylic acids. A study using radioactive labeling shows that the liver and brain cholesterol incorporate acetic acid [195]. No cumulative toxicity has been reported.

Dilute acid with 5 % acetic acid such as vinegar can cause irritation to human mucous membranes. Vinegar contains 3 – 6 % acid. Persons of older than two years can take as much as 2.1 g daily. However, the dilute acid damages skin more severely than other mineral acids because it is miscible with lipids.

Acetic acid irritates the eyes, nose, and throat above 10 ppm. Repeated exposure to 26 ppm for ten days causes adverse effects. Over 100 ppm, the acid may even damage organs as described above.

Concentrated acid can damage skin severely. Avoid exposure to glacial acetic acid or inhaling vapor. In case of contact, wash the exposed area with plenty of water. Therefore, handling of the acid requires caution.

Oral ingestion of the acid causes pain in the digestive tract and in the mouth. It may lead to vomiting, and respiratory and circulatory distress. If the person is conscious, wash the mouth with water and consult physicians.

Acetic acid has a low flash point of 39 °C [196]. It is readily combustible with an autoignition point of 516 °C. Fortunately, the acid lowers its flammability significantly when mixed with water. Beyond 56 %, the flash point is no longer a danger [197].

Organic acetates do not cause severe damage unless high exposure occurs. The TLVs vary from 5 to 400 ppm [198]. The LD_{50} for small mammals range from 0.4 to 16 g/kg. The toxicity of the esters decreases as the molecular mass increases from C_1 to C_4. The LD_{50} of isoamyl acetate (7.42 g/kg) is twice that of methyl acetate (3.7 g/kg). However, ingestion or absorbtion of methyl acetate poses a hazard since the ester is converted to methanol.

Acetates irritate the mucous membrane slightly. When ingested or absorbed, esters are readily hydrolyzed to the corresponding alcohols and carboxylic acids. Some volatile esters act as asphyxiants and narcotics.

Ethyl acetate has an LD_{50} of 4.97 g/kg in the rabbit. The toxicity of acetates is usually in the range of 100 – 250 ppm. Ethyl acetate has the highest PEL of 500 ppm. Humans can tolerate up to 400 ppm before nose and throat irritation.

n-Butyl acetate differs from its isomers. When tested with rats, *n*-butyl acetate has an oral LD_{50} of 14 mg/kg. For humans, irritation in the throat starts at an exposure of 200 ppm [194]. Severe eye and nose irritation occurs at 300 ppm. *sec*-Butyl acetate poses less hazard than its normal isomer, with less odor and less irritation. The lethal dose for *sec*-butyl acetate in rabbits is 4.8 mg/kg.

Table 10. Exposure limits to acetic acid and its derivatives [93], [94]

Substance	TLV				MAK	
	TWA		STEL			
	ppm	mg/m^3	ppm	mg/m^3	mL/m^3	mg/m^3
Acetic acid	10	25	15	37	10	25
n-Butyl acetate	150	713	200	950	200	950
sec-Butyl acetate	200	950			200	950
Dimethylacetamide (skin)	10	36			10	35
Ethyl acetate	400	1400			400	1400
Methyl acetate	200	606	250	757	200	610

Acetyl chloride is harmful if swallowed, inhaled, or absorbed through the skin. Target areas include the peripheral nervous system, sense organs, CNS (behavior), lungs, thorax and respiratory system. Acetyl chloride reacts with moist air to produce hydrogen chloride and acetic acid.

Amides are more hazardous than esters. Acetamide causes skin and eye irritation. It is also irritating to the mucous membranes and the upper respiratory tract. It is considered a carcinogen and has been shown to alter DNA in mouse and rat embryos. Methylacetamide has teratogenic and neoplastic potential. *N,N*-Dimethylacetamide causes hallucinogenic effects similar to those of LSD at doses of 400 mg/kg [199]. *N,N*-dimethylacetamide overexposure has led to reproductive disorders in laboratory animals. Target organs include the liver, kidney, and CNS. For humans, damage to the liver occurs at the exposure level of 20 – 25 ppm. The recommended TLV is 10 ppm without skin exposure.

Phenylacetic acid is irritating to the eyes, skin, mucous membranes, and upper respiratory system.

13. References

1. W. K. Johnson, P. Vau: "Acetic Acid", in *Chemical Economics Handbook,* SRI International, Menlo Park, CA 1996, 602.5000.
2. L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis,* John Wiley & Sons, New York 1995 p. 11.
3. A. P. Kudchadker, G. H. Alani, B. J. Zwolinski, *Chem. Rev.* **68** (1968) no. 6, 659 – 735.
4. D. R. Lide, H. P. R. Frederikse (eds.): *CRC Handbook of Chemistry and Physics,* 77th ed., CRC Press, Boca Raton, Florida, No. 100. "Acetic acid", Beilstein reference no. 4-02-00-000094; Merck Index, 10th ed., no. 47.
5. J. A. Ridduck, W. B. Bunger: *Organic Solvents: Physical Properties and Methods of Purification,* 3rd ed., Wiley-Interscience, New York 1970.
6. *Kirk-Othmer,* 4th ed., **vol. 1,** pp. 121 – 129.
7. D. D. Perrin, W. L. F. Armarego, D. R. Perrin: *Purification of Laboratory Chemicals,* Pergamon Press, New York 1966.
8. *Ullmann,* 5th ed., **vol. A1,** p. 45.
9. W. H. Seaton in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, p. 106.
10. W. H. Seaton, unpublished work for the Design Data Research Laboratory, Tennessee Eastman Co., Kingsport, TN, 1970 – 1974.
11. *International Critical Tables,* McGraw-Hill, New York 1926.
12. A. Ragni, G. Ferrari, P. Papoff, *Ann. Chim. (Rome)* **45** (1955) 960 – 969.
13. A. N. Campbell, E. M. Kartzmark, *Can. J. Res.* **28B** (1950) 161 – 169.
14. J. Timmermans, *Physico-chemical Constants of Pure Organic Compounds,* Elsevier, New York 1950.
15. A. W. Francis, *Chem. Eng. Sci.* **10** (1959) 37.
16. I. Brown, A. H. Ewald, *Austr. J. Sci. Res. Ser. A* **3** (1950) 306 – 323.
17. J. J. Jasper, *J. Phys. Chem. Ref. Data* **1** (1972) no. 4, 841 – 1009.
18. R. Barton, C. C. Hsu, *J. Chem. Eng. Data* **14** (1969) 184 – 187.
19. M. D. Taylor, *J. Amer. Chem. Soc.* **73** (1951) 315.
20. W. J. Weltner, *J. Amer. Chem. Soc.* **77** (1955) 3941 – 3950.
21. J. L. Derissen, *J. Mol. Struct.* **7** (1971) 67 – 80.
22. G. Arich, G. Tagliavini, *Ric. Sci.* **28** (1958) 2493.

23. S. D. Christian, H. E. Affsprung, C. Ling, *J. Chem. Soc.* **1965** (1965) 2378 – 2381.
24. R. Vilcu, E. Lecinescu, *Rev. Roum. Chim.* **14** (1969) 283 – 290.
25. C. Sandonnini, *(Atti. Accad. Naz. Lincei CL. Sci.) Fis. Mat. Nat. Rend.* **4** (1926) 63 – 68.
26. S. P. Miskidzh′yan, H. A. Trifonof, *Zh. Obshch. Khim.* **17** (1947) 1033.
27. M. Usanovich, L. N. Vasil′eve, *Zh. Obshch. Khim.* **16** (1946) 1202.
28. K. T. Abtiev, *Vestnik. Adad. Nauk Kazakh, S.S.R.* **15** (1959) no. 12, 72 – 77.
29. G. S. Park, K. K. Kelley, *J. Am. Chem. Soc.* **47** (1925) 2089.
30. J. J. Lagowski in: *The Chemistry of Nonaqueous Solvents,* **vol. 3,** Academic Press, New York 1970, p. 244.
31. F. H. MacDougall, *J. Am. Chem. Soc.* **58** (1936) 2586.
32. I. Brown, A. H. Ewald, *Aust. J. Sci. Res.-Series A-Phys. Sci.,* **3** (1950) 306 – 323.
33. J. A. Dean (ed.) in: *Lange′s Handbook of Chemistry,* 12th ed., McGraw-Hill, New York 1979, pp. 6 – 36.
34. R. D. Nelson, D. R. Lide, A. A. Maryott in: *Selected values of electric dipole moments in the gas phase,* National Reference Data Sheet, National Bureau of Standards, Bulletin no. 10, 1967.
35. D. D. Wagman et al., *J. Phys. Chem. Ref. Data* **11** (1982) Suppl. 2.
36. N. I. Sax: *Dangerous Properties of Industrial Materials,* 5th ed., Van Nostrand Reinhold Co., New York 1979, p. 333.
37. E. K. Euranto in S. Patai (ed.): *The Chemistry of Carboxylic Acids and Esters,* Interscience Publishers, London 1969, pp. 505 – 588.
38. M. Gauß, A. Seidel, P. Torrence, P. Heymanns in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **vol. 1,** VCH Publishers, New York 1996, pp. 104 – 138.
39. P. M. Matlis, A. Haynes, G. J. Sunley, M. J. Howard, *J. Chem. Soc., Dalton Trans.* 1996 2187 – 2196.
40. M. J. Howard, M. D. Jones, M. S. Roberts, S. A. Taylor, *Catal. Today* **18** (1994) 325 – 354.
41. Showa Denko K. K., WO 94/22803, 1994 (T. Suzuki, H. Yoshikawa, K. Abe, K. Sano).
42. R. P. Lowry, A. Aguilo, *Hydrocarbon Process.* **53** (1974) no. 11, 103 – 113.
43. H. Hohenschutz et al., *Hydrocarbon Process.* **45** (1966) no. 11, 141 – 143.
44. J. W. Reppe: *Acetylene Chemistry,* Charles A. Meyer & Co., Boston, Mass. 1949.
45. *Hydrocarbon Process.* **46** (1967) no. 11, 136.
46. *Hydrocarbon Process.* **52** (1973) no.11, 92.
47. N. V. Kutepow, W. Himmele, H. Hohenschutz, *Chem. Ing. Tech.* **37** (1965) 383 – 388.
48. S. F. Dickson, J. Bakker, A. Kitai: "Acetic Acid," in *Chemical Economics Handbook,* SRI International, Menlo Park, CA March 1982, rev. July 1982, 602.5020, 602.5021.
49. F. E. Paulik, J. F. Roth, *Chem. Commun.* (1968) 1578.
50. D. Forster in F. C. A. Stone, R. West (eds.): *Advances in Organometallic Chemistry* **vol. 17,** Academic Press, New York 1979, pp. 255 – 267.
51. *Chem. Eng.* 1978, June 5, 182.
52. Hoechst Celanese Corp., US 5001259, 1991.(B. L. Smith, G. P. Torrence, A. Aguilo, S. Alder).
53. *Chemical Market Reporter* 1997, Sept. 1, 3.
54. *Process Engineering,* 1996, July, p. 21.
55. D. J. Watson: "The Cativa Process for the Production of Acetic Acid", *15th Meeting of the North American Catalysis Society,* Chicago, Ill., May 18 – 22, 1997.
56. D. Forster, T. C. Singleton, *J. Mol. Catal.* **17** (1982) 299 – 314.
57. T. Dekleva, D. Forster, *Adv. Catalysis,* **34** (1986) 81 – 130.
58. D. L. King, K. K. Ushiba, T. E. Whyte, Jr., *Hydrocarbon Process.* **61** (1982) no. 11, 131 – 136.
59. P. Ellwood, *Chem. Eng. (N.Y.)* **76** (1969) no. 11, 148 – 150.
60. J. J. McKetta, W. A. Cummingham (eds.): *Encyclopedia of Chemical Processing Design,* Marcel Dekker, New York 1976, pp. 216 – 246.
61. J. Hjortkjaer, V. W. Jenson, *Ind. Eng. Chem., Prod. Res. Dev.* **15** (1976) 46 – 49.
62. J. Hjortkjaer, O. R. Jenson, *Ind. Eng. Chem., Prod. Res. Dev.* **16** (1977) 281 – 285.
63. T. Matsumoto, K. Mori, T. Mizorki, A. Ozaki, *Bull. Chem. Soc. Japan* **50** (1977) no. 9, 2337 – 2340.
64. S. B. Dake, D. S. Kolhe, R. V. Chaudhari, *Ind. Eng. Chem. Res.* **28** (1989) 1107 – 1110.
65. Monsanto, US 3 769 329, 1973 (F. E. Paulik, A. Hershman, W. R. Knox, J. F. Roth).
66. G. W. Adamson, J. J. Daly, D. Forster, *J. Organomet. Chem.* **71** (1974) C17 – C18.
67. A. G. Kent, B. E. Mann, C. P. Manuel, *J. Chem. Soc. Chem. Commun.* **1985** (1985) 728 – 729.
68. D. Forster, *Ann. N.Y. Acad. Sci.* **295** (1977) 79 – 82.

69. D. Forster, *J. Am. Chem. Soc.* **98** (1976) 846 – 848.
70. B. L. Smith, M. A. Murphy, G. T. Torrence, A. Aguilo, *Inorg Chim Acta* **101** (1985) L47 – L49.
71. M. A. Murphy, B. L. Smith, G. P. Torrence, A. Aguilo, *J. Organomet. Chem.* **303** (1986) 257 – 272.
72. M. A. Murphy, B. L. Smith, G. P. Torrence, A. Aguilo, *J. Mol. Catal.* **39** (1987) 115 – 136.
73. A. Haynes, B. E. Mann, D. J. Gulliver, G. E. Morris, P. M. Maitlis, *J. Am. Chem. Soc.* **113** (1991) 8567 – 8569.
74. A. Haynes, B. E. Mann, D. J. Gulliver, G. E. Morris, P. M. Maitlis, *J. Am. Chem. Soc.* **115** (1993) 4093 – 5000.
75. C. E. Hickey, P. M. Maitlis, *J. Chem. Soc. Chem. Commun.* (1984) 1609 – 1611.
76. A. Fulford, P. M. Maitlis, *J. Organomet. Chem.* **366** (1989) C20 – C22.
77. A. Fulford, C. E. Hickey, P. M. Maitlis, *J. Organomet. Chem.* **398** (1990) 311 – 323.
78. P. M. Maitlis, A. Haynes, G. J. Sunley, M. J. Howard, *J. Chem. Soc., Dalton Trans* 1996, no. 11, 2187 – 2196.
79. E. Baker, D. E. Hendrickson, R. Eisenberg, *J. Am. Chem. Soc.* **102** (1980) 1020 – 1027.
80. B. R. James, G. L. Rempel, *J. Chem. Soc. A* (1969) 78 – 84.
81. T. C. Singleton, L. J. Park, D. Forster, *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **24** (1979) 329 – 335.
82. D. Forster, *The Chemist* (1981) Jan., 7 – 10.
83. M. J. Howard, M. D. Jones, M. S. Roberts, S. A. Taylor, *Catal. Today* **18** (1993) 325 – 354.
84. M. Bassetti et al., *Gazz. Chim. Ital.* **122** (1992) 391 – 393.
85. P. R. Ellis et al., *Organometallics* **13** (1994) 3215 – 3226.
86. J. M. Pearson, A. Haynes, G. E. Morris, G. J. Sunley, P. M. Maitlis, *J. Chem. Soc., Chem. Commun.* 1995, no. 10, 1045 – 1046.
87. T. Ghaffar et al., *Chem. Commun.* 1998, no. 9, 1023 – 1024.
88. A. Haynes, J. M. Pearson, P. W. Vickers, J. P. H. Charmant, P. M. Maitlis, *Inorg. Chim. Acta.* **270** (1998) nos. 1 + 2, 382 – 391.
89. J. Gauthier-Lafaye, R. Perron: "Methanol et Carbonylation," Rhône-Poulenc Recherches, Courbevoie 1986 [English translation: Methanol Carbonylation, Editions Technip, Paris, and Rhône-Poulenc Recherches, Courbevoie 1987 p. 117 – 148.
90. J. R. Zoeller in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, pp. 35 – 51.
91. T. Grewer, A. Schmidt, *Chem. Ing. Tech.* **45** (1973) 1063 – 1067.
92. R. T. Eby, T. C. Singleton, *Appl. Ind. Catal.* **1** (1983) 275 – 296.
93. BP Chemicals Limited, EP 0 728 729, 1995 (S. J. Smith, R. J. Watt, B. L. Williams).
94. BP Chemicals Limited, US 5 750 007, 1998 (K. E. Clode, D. J. Watson, C. J. E. Vercauteren).
95. C. C. Hobbs in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **vol. 1,** VCH Verlagsgesellschaft, Weinheim 1996, p. 521.
96. *Kirk-Othmer,* 4th ed., **vol. 13,** pp. 682 – 717.
97. *Ethylene and Derivatives,* **vol. 1,** Wood Petrochemicals Program, SRI International, Menlo Park, CA1995, p. WORL-28.
98. S. W. Benson in R. R. Gould, F. R. Mayo (eds.): *Advances in Chemistry Series 76,* American Chemical Society, Washington, D.C. 1968, pp. 143 – 153.
99. G. Irick in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, p. 31.
100. D. L. Llord, P. L. Eve, P. D. Gammer, *Erdoel, Erdgas, Kohle* **109** (1993) no. 6, 266.
101. R. A. Sheldon, J. K. Kochi: *Metal-Catalyzed Oxidation of Organic Compounds,* Academic Press, New York 1981, p. 143.
102. R. S. Drago, *Coord. Chem. Rev.* **117** (1992) 185.
103. C. C. Hobbs et al., *Ind. Eng. Chem. Process Des. Der.* **11** (1972) 59.
104. S. W. Benson, P. S. Nangia, *Acc. Chem. Res.* **12** (1979) no. 7, 223 – 228.
105. R. A. Sheldon, J. K. Kochi: *Metal-Catalyzed Oxidation of Organic Compounds,* Academic Press, New York 1981, p. 20.
106. A. T. Betts, H. Uri in R. R. Gould, F. R. Mayo (eds.): *Advances in Chemistry Series 76,* American Chemical Society, Washington, D.C. 1968, pp. 161 – 181.
107. L. Sieg in W. Jost (ed.): *Low Temperature Oxidation,* Gordon and Breach, New York 1965, p. 222.
108. D. Swern (ed.): *Organic Peroxides,* John Wiley & Sons, New York 1970, p. 214.
109. J. F. Black, *J. Amer. Chem. Soc.* **100** (1978) 527.
110. J. Skriniarova, M. Hronec, J. Ilavsky, *Oxid. Commun.* **10** (1987) nos. 1 + 2, 51 – 68.
111. G. A. Russel, *J. Amer. Chem. Soc.* **79** (1957) 3871.

112. Q. J. Niu, G. D. Mendenhall, *J. Amer. Chem. Soc.* **114** (1992) no. 1, 165 – 172.
113. S. W. Benson, *Oxid. Commun.* **2** (1982) nos. 3 + 4, 169 – 172.
114. *Ullmann,* 4th ed., **vol. 11,** 57 – 64.
115. H. Hofermann, *Chem. Ing. Tech.* **36** (1964) no. 5, 422 – 429.
116. *Acetic Acid, report 37,* 37 A, SRI Internationals Process Economics Program, private report, Feb. 1968, March 1973.
117. T. Yamaguchi, *Jpn. Chem. Q.* **IV-I** (1968) 27 – 33.
118. J. Royer, M. Beugelmans-Verrier, *C. R. Hebd. Seances Acad. Sci. Ser. C* **272** (1971) 1818.
119. N. A. Cliton, R. A. Kenley, T. G. Traylor, *J. Am. Chem. Soc.* **97** (1975) 3746, 3752, 3757.
120. F. Colombo, *Quad. Ing. Chim. Ital.* **18** (1982) 68 – 84. *Chem. Abstr.* **98** 142586.
121. A. T. Fanning in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, pp. 18 – 21.
122. W. J. DeKlein, E. C. Kooyman, *J. Catal.* **4** (1965) 626.
123. G. C. Allen, A. Aguilo, *Adv. Chem. Ser.* **76** (1968) 363.
124. Celanese Corp., NL 8004 525, 1980.
125. D. R. Larkin, *J. Org. Chem.* **55** (1990) 1563.
126. R. A. Sheldon, J. K. Kochi: *Metal Catalyzed Oxidations of Organic Compounds,* Academic Press, New York 1981.
127. T.-C. Chou, F.-S. Lin, *Can. J. Chem.* **61** (1983) 1295 – 1300.
128. Standard Oil Co, US 4 408 071, 1983 (S. E. Pedersen, H. F. Hardman, L. F. Wagner).
129. J. L. Seone, P. Boutry, R. Montarnal, *J. Catal.* **63** (1980) 182 – 190.
130. E. M. Thorstein, T. P. Wilson, F. G. Young, P. H. Kasai, *J. Catal.* **52** (1978) 116 – 132.
131. Union Carbide Corporation, EP 0 216 264 A1, 1986 (R. M. Manyik, J. L. Brockwell, J. E. Kendall).
132. Union Carbide Corporation, EP 0294 845 A1, 1988 (J. H. McCain, S. W. Kaiser, G. L. O′Connor).
133. Rhône-Poulenc Chime, EP 0 627 401 A1, 1994 (F. Blaise, E. Bordes, M. Gubelman, L. Tessier).
134. Rhône-Poulenc Chimie, EP 0 479 692 A1, 1991 (P. Barthe, G. Blanchard).
135. L. Tessier, E. Bordes, M. Gubelmann-Bonneau, *Catal. Today* **24** (1995) 335 – 340.
136. M. Roy, M. Gubelmann-Bonneau, H. Ponceblanc, J.-C. Volta, *Catalysis Letters* **42** (1996) 93 – 97.
137. BP Chemicals Limited, EP 0 407 091 A1, 1990 (M. Kitson).
138. BP Chemicals Limited, EP 0 480 594 A2, 1991 (C. Hallet).
139. The Standard Oil Company, EP 0 518 548 A2, 1992 (P. R. Blum, M. A. Pepera).
140. Hoechst AG, WO 9805619 A1, 1998 (H. Borchert, U. Dingerdissen).
141. L. R. Partin, W. H. Heise in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, pp. 3 – 13.
142. H. Ebner, S. Sellmer, H. Follmann in H. J. Rehm, G. Reed (eds.): *Biotechnology,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1996, pp. 381 – 401.
143. The Texas A&M University System, WO 98/04729 A2, 1997 (M. Holtzapple, R. Davison, M. Loescher, C. Loeffler).
144. CPC International Inc., EP 0 120 370, 1984 (F. A. Keller, J. S. Ganoung, S. J. Luenser).
145. T. Beppu, *Antonie van Leeuwenhoek* **64** (1993) 121 – 135.
146. P. N. Lodal in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, pp. 65 – 69.
147. N. L. Ricker, J. N. Micheals, C. J. King, *J. Separ. Tech.* **1** (1979) 36 – 41.
148. F. Chen, H. Tanaka, Y. Naka, E. Oshima, *J. of Chem. Eng. of Japan* **22** (1989) 6 – 11.
149. A. M. Eyal, R. Canari, *Ind. Eng. Chem. Res.* **34** (1995) 1789 – 1798.
150. J. M. Wardell, C. J. King, *J. of Chem. and Eng. Data* **23** (1974) 144.
151. Dow Chemical Co., US 3 980 701, 1975 (R. R. Ginstead).
152. M. Sittig (ed.): *Handbook of Toxic and Hazardous Chemicals and Carcinogens,* 3rd ed., **vol. 1,** Noyes Publications, Park Ridge, NJ 1991, p. 22.
153. J. J. McKetta, W. A. Cunningham (eds.): *Encyclopedia of Chemical Process Design,* Marcel Dekker, New York 1976, pp. 216 – 246.
154. K. A. Roberts: *private communication,* Celanese Limited, Chemical Division, Dallas, TX, June 16, 1998.
155. *Material Safety Data Sheet for Glacial Acetic Acid,* Eastman Chemical Products, Inc., Kingsport, Tennessee, 1990.
156. American Conferences of Governmental Industrial Hygienists (ACGIH) (ed): *Threshold Limits Values,* Cincinnati, Ohio, 1982; cf. Deutsche Forschungsgemeinschaft, *Maximale Arbeitsplatzkonzentrationen,* Verlag Chemie, Weinheim 1983, p. 15.

157. Celanese Ltd., Celanese Chemicals Division, http://www.celanese.com
158. S. Nunn, private communication, Celanese Limited, Chemical Division, Dallas, TX, June 15, 1998.
159. R. E. Langa, K. L. Votoupal (eds.): *The Sigma-Aldrich Library of regulatory and Safety Data,* **vol. 1,** Sigma-Aldrich Chemical Company, Milwaukee, WI 1993, pp. 533 – 534.
160. New York Department of Health (ed.): *Hazardous Substances Fact Sheet,* Trenton, NJ Sept. 1985.
161. A. I. Grigor′ev, E. G. Pogudilova, A. V. Novoselova, *Zh. Neorg. Khim.* **10** (1965) 772.
162. A. G. Stickstoffdunger, US 2 141 477, 1937 (J. Losch).
163. J. C. Bailar, H. J. Emeleus, R. Nyholm, A. F. Trotman-Dickenson: *Comprehensive Inorganic Chemistry,* **vol. 1,** Pergamon Press, Oxford 1973, pp. 462 – 464.
164. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* **vol. 1,** Pt 1, Macmillan Publishing, New York 1963, pp. 107 – 212.
165. *Hoyben-Weyl,* 4th ed., **vol. 8,** p. 558.
166. J. S. Witzeman, V. H. Agreda in V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Marcel Dekker, New York 1993, pp. 257 – 284.
167. *Kirk-Othmer,* 4th ed., **vol. 9,** pp. 781 – 812.
168. S. Patai (ed.): *The Chemistry of Carboxylic Acids and Esters,* Wiley-Interscience, New York 1969.
169. R. W. Johnson, E. Fritz (eds.): *Fatty Acids in Industry,* Marcel Dekker, New York 1989.
170. Z. Gomzi, S. Zrncevic: *Croat. Chem. Acta* **53** (1980) no. 1, 25 – 32. *Chem. Abstr.,* **93** (13), 131701x.
171. H. P. Thahar, S. A. Puranik, M. G. Kher, *Fluid. Ses. Appl., C. R. Congr. Int.* (1972) 615 – 627. *Chem. Abstr.,* **87** (3), 22303z.
172. *Eur. Chem. News,* 1986, March 17, 6.
173. *Org. Synth.* **4** (1963) 261, 417.
174. U. Matteoli, G. Menchi, M. Bianchi, P. Frediani, F. Piacenti, *Gazz. Chim. Ital. Pt. A* **115** (1985) nos. 11, 12, 603 – 606. *Chem. Abstr.,* **104** (25), 224426x.
175. Mitsubishi Chemical Industries, JP-Kokai 58 192 850 A2222 (83 192 850), 1983. *Chem. Abstr.,* **100** (13), 102772d.
176. V. E. Tishchenko, *J. Russ. Phys. Chem. Soc.* **38** (1908) 355, 482.
177. G. M. Villacorta, J. San Filipo, Jr., *J. Org. Chem.* **48** (1983) 1151 – 1154.
178. *Kirk-Othmer,* 3rd ed., **vol. 9,** p. 328.
179. Albemarle Corporation, US 5 672 749, 1997 (W. B. Wautes, R. E. Young, P. R. DeVrou).
180. Kokoku, JP 27-4567, 1952.
181. S. Patai (ed.): *The Chemistry of Acyl Halides,* Interscience, New York 1972.
182. J. Zabicky (ed.): *The Chemistry of Amides,* Interscience Publishers, New York 1970.
183. WO 96/30379, 1996 (A. Parkins).
184. Nitrokemia Ipartelepek, EP 0 418 646 A2, 1990 (J. Paal et al.).
185. Sumitomo, EP 0 568 072 A2, 1993 (Y. Takashima, K. Kumagai, S. Mitsuda).
186. US 4139557, 1976 (M. B. Neduv et al.).
187. Japan Gas-Chemical Co., JP 46 015 082, 1971 (M. Kuraguchi, S. Asano, N. Isogai); CAN 75: 98190.
188. Japan Gas-Chemical Co., JP 46 015 083, 1971 (Y. Kuraishi, S. Asano, T. Oota); CAN 75: 63169.
189. Armstrong World Industries, US 4 956 121, 1989 (T. M. Tymon, A. F. Turbak).
190. Y. Yamazaki, A. Masuda, *Yuki Gosei Kagaku Kyokaishi* **40** (1982) 218 – 226. CAN 96: 199175.
191. Y. N. Pozdnyakovich, E. S. Nazarenko, S. M. Shein, *Khim. Prom-st.* **1** (1985) 5 – 8. Can 102: 115486.
192. D. L. Opdyke, *Food Cosmet. Toxicol.* **13** (1975) Suppl., 901 – 902.
193. C. E. Rumme, K. A. Roberts, *private communication,* Celanese Limited, Chemical Division, Dallas, TX, June 7, 1998.
194. American Conferences of Governmental Industrial Hygienists (ACGIH) (ed): *Threshold Limits Values (TLV), Biological exposure indices (BEI),* Cincinnati, Ohio 1997.
195. H. J. Nicholas, B. E. Thomas, *Brain* **84** (1961) 320 – 328.
196. P. N. Lodal in V. Agreda, J. R. Zoeller (eds.): *Acetic Acid and Its Derivatives,* Dekker, New York 1993 p. 361.
197. I. G. Scott in: Tennesse Easstman Company Technical Report **TEAD-R-DC-706-2254,** Kingport, TN, April 24, 1986.
198. United States Department of Health and Human Services (ed.): *Registry of Toxic Effects of Chemical Substances,* **vol. 1 – 4,** U.S. Government Printing Office, Washington D.C., 1988.
199. A. J. Weiss, *Science* **136** (1962) 151.

Acetic Anhydride and Mixed Fatty Acid Anhydrides

HEIMO HELD, Wacker-Chemie GmbH, Werk-Burghausen, Burghausen, Federal Republic of Germany (Sections 1.1 – 1.3.2, 1.4 – 1.7, and Chaps.2 and 3)

ALFRED RENGSTL, Wacker-Chemie GmbH, Werk-Burghausen, Burghausen, Federal Republic of Germany (Section 1.3.3)

DIETER MAYER, Pharma Forschung Toxikologie, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 4)

1. Acetic Anhydride 1
1.1. Physical Properties 1
1.2. Chemical Properties 2
1.2.1. Acetylation 2
1.2.2. Dehydration 5
1.2.3. Reactions of the α-Protons 5
1.2.4. Reactions of a Single Carbonyl Group 5
1.2.5. Production of Silver Ketenide 5
1.3. Production 6
1.3.1. Ketene Process 6
1.3.1.1. Production of Ketene 6
1.3.1.2. Reaction of Ketene with Acetic Acid 7
1.3.1.3. Pure Anhydride Distillation 8
1.3.1.4. Environmental Problems 8
1.3.2. Oxidation of Acetaldehyde 8
1.3.3. Carbonylation of Methyl Acetate . . 9
1.4. Analysis 11
1.5. Quality Specifications 11
1.6. Storage and Transportation 12
1.7. Uses 12
2. Mixed Fatty Acid Anhydrides . . 12
2.1. Physical Properties 12
2.2. Chemical Properties 13
2.3. Production 13
2.4. Uses 13
3. Economic Aspects 13
4. Toxicology and Occupational Health 13
5. References 14

1. Acetic Anhydride

Acetic anhydride [*108-24-7*], $(CH_3CO)_2O$, was first prepared by C. GERHARDT in 1852 by the reaction of benzoyl chloride with molten potassium acetate. Today it is one of the most important organic intermediates and is used widely in both research and industry.

1.1. Physical Properties

Acetic anhydride, $C_4H_6O_3$, M_r 102.09, *mp* −73.1 °C, *bp* 139.5 °C (at 101.3 kPa), is a colorless liquid with a pungent odor and is strongly lachrymatory. The most important physical data are given below.

Critical pressure	4680 kPa
Critical temperature	296 °C
Vapor pressure	

t, °C	20	40	60	80	100	120
p, kPa	0.4	1.7	5.2	13.3	28.7	53.3

Density	d_{20}^{20} 1.0838; d_{20}^{-195} 1.3290; d_{20}^{-79} 1.277; d_4^{15} (1.0870) up to d_4^{50} (1.0443), see [1]
Refractive index n_D^{20}	1.39038
$n_{322\,nm}^{16}$	1.4174
$n_{667.8\,nm}^{16}$	1.3897, see [2]
^{1}H NMR absorption	see [3]
UV absorption maximum	217 nm
Specific heat capacity (23 – 122 °C)	1.817 J/g
Heat of vaporization	
at 18.5 °C	496.5 J/g
at 137 °C	276.7 J/g
Heat of combustion at constant volume at 25 °C	1804.5 kJ/mol
Thermal conductivity at 25 °C	2.215 mJ cm^{-1}s^{-1}K^{-1} [4]
Electric conductivity at 20 °C	2.3×10^{-8} Ω^{-1}cm^{-1} [5]
Dielectric constant at 20 °C	20.5
Viscosity	
at 15 °C	0.971 Pa s
at 30 °C	0.783 Pa s
(for temperature dependence, see [2], [6])	
Cubic expansion coefficient at 18 °C	1.13×10^{-9} K^{-1}
Adiabatic compressibility constant	4.86×10^{8} cm^2/N
Molecular refraction	22.38 cm^3

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_065

Surface tension

t, °C	15	20	40	60	100	139.5
σ, mN/m	33.4	32.7	30.0	28.1	23.3	18.6

Acetic anhydride is miscible with polar solvents and dissolves in cold alcohol with slow decomposition. The solubility of acetic anhydride in water at 20 °C is 2.6 wt %, with slow decomposition; the solubility of water in acetic anhydride at 15 °C is 10.7 wt %, with gradual decomposition. An overview of the solubilities of several anhydrous organic and inorganic compounds is given in [7].

Data on the kinetics of hydrolysis in water or aqueous mixtures of acids, bases, or organic solvents, with or without additives and at different temperatures, are given in [8]. Rate constants for the solvolysis in alcohols or alcoholic mixtures of organic solvents, with or without additives, are also given.

Melting points of acetic acid – acetic anhydride mixtures are as follows:

% acetic acid	100	80	60	40	20	10
mp, °C	16.7	5.4	−5.8	−19.8	−44.8	−68.1

The vapor-liquid equilibrium curve for mixtures of acetic acid and acetic anhydride is shown in Figure 1; this is of great importance in industry. Further vapor – liquid equilibria, for example, with water, benzene, diketene, propionic acid, pyrimidine, and water + acetic acid are given in [8].

Safety data:

Flash point	52.5 – 53 °C (closed cup, ASTM 56 – 70, DIN 51755)
Ignition temperature	315 °C
Explosive limits in air (20 °C, 101.3 kPa)	

1.2. Chemical Properties

Acetic anhydride undergoes a large variety of chemical reactions and is by far the most researched aliphatic carboxylic acid anhydride. In about half of the several thousand relevant publications and patents over the last 25 years, acetic anhydride was used for the acetylation of OH or NH groups (at least for the primary step). Many special reactions of acetic anhydride are summarized in [9].

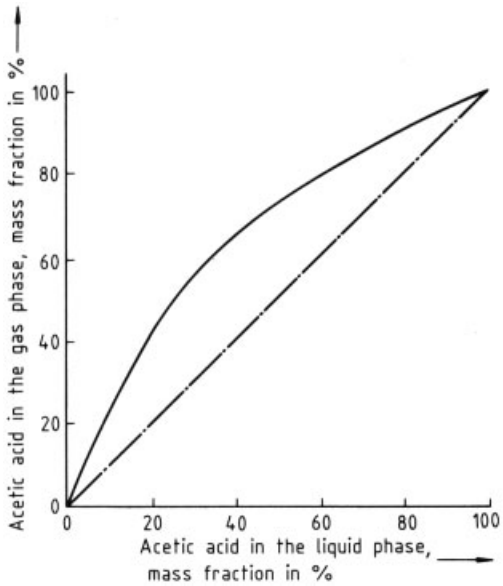

Figure 1. Equilibrium curve: acetic acid – acetic anhydride

1.2.1. Acetylation

O-Acetylation. Acetic anhydride is particularly suitable for the esterification of alcohols, a reaction that is difficult or impossible with acetic acid. Acetic acid is set free in the course of the reaction. Bases and strong acids as well as salts, such as sodium acetate, are suitable as catalysts. Examples of the reaction with hydroxyl groups are the formation of acetyl cellulose, acetylsalicylic acid (Aspirin), and glycerol triacetate (→ Cellulose Esters, → Salicylic Acid, → Glycerol).

Carboxylic Acid reactions are described in Chap. 2 (→ Carboxylic Acids, Aliphatic; → Carboxylic Acids, Aromatic).

Hydrogen peroxide reacts with acetic anhydride to give peracetic acid [*79-21-0*] or diacetyl peroxide [*110-22-5*], depending on the molar ratio of the reactants:

$$(CH_3CO)_2O + 2\,H_2O_2 \rightleftharpoons 2\,CH_3CO_3H$$

$$(CH_3CO)_2O + H_2O_2 \rightleftharpoons CH_3COOH + CH_3CO_3H$$

$$(CH_3CO)_2O + CH_3CO_3H \rightleftharpoons CH_3CO\text{-}OO\text{-}COCH_3 + CH_3COOH$$

N-Acetylation. Acetylation of compounds containing NH groups yields acetamides following the general equation:

$$RR'NH + (CH_3CO)_2O \rightarrow RR'N\text{-}COCH_3 + CH_3COOH$$

where R, R′ = H or alkyl.

Amines. Aliphatic amines usually react without heating. Aniline gives acetanilide (→ Aniline), the acetyl group of which prevents oxidation during subsequent nitration. N-Acetylation generally occurs faster than the acetylation of OH groups. Therefore partial acetylation of compounds with several functional groups is possible. Examples of such reactions are the production of N-acetylamino acids, such as *N*-acetylmethionine-*S*-oxide (**1**) [10] and of N-acetylanthranilic acids (**2**) [11].

1

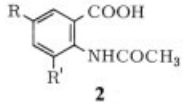

2

Amides and Carbamides. Aliphatic and aromatic carboxylic acid amides as well as carbamides with a free NH group can be acetylated with acetic anhydride. Strong acids, such as sulfuric acid, are most often used as catalysts. The breadth of application of this reaction can be seen from the following products: *N,N,N′,N′*-tetraacetylethylenediamine [*10543-57-4*] [12],

$(CH_3CO)_2NCH_2CH_2N(COCH_3)_2$

and 2,4,6,8-tetraacetylazabicyclo [3.3.1] nonane-3,7-dione [13]:

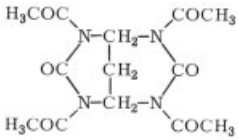

C-Acetylation. Compounds with reactive CH bonds can be acetylated with acetic anhydride, with the aid of a catalyst if necessary. Examples are the production of ethyl α-cyanoacetoacetate [*634-55-9*], $CH_3COCH(CN)COOC_2H_5$, (catalyst: K_2CO_3) [14] and the Friedel-Crafts reaction of acetic anhydride with aromatic hydrocarbons, for example, with benzene to form acetophenone (→ Ketones), as well as with unsaturated hydrocarbons. An interesting example of the latter group is the acetylation of 2,3-dimethylbutadiene tricarbonyl iron in the presence of aluminum chloride [15] to give the following complex:

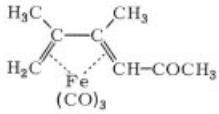

The reaction of ketones with acetic anhydride using boron trifluoride as a catalyst gives β-diketones [16]. Triacetylmethane [*815-68-9*] can be synthesized directly from isopropenyl acetate, acetic anhydride, and aluminum chloride catalyst [17]. In the same way, unsaturated methyl ketones can be produced by acetylation of olefins with acetic anhydride in the presence of zinc chloride as catalyst [18]:

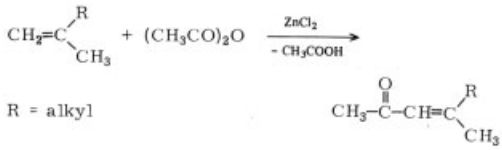

Acetylation of Mineral Acids. The reaction of nitric acid with acetic anhydride gives acetyl nitrate [*591-09-3*], which is used often as a nitrating agent in organic chemistry. For example, it is used in the production of 2-nitrocyclohexanone [*4883-67-4*], a precursor of ε-caprolactam [19]. Acetyl nitrate can be synthesized also from dinitrogen pentoxide and acetic anhydride.

Other strong acids that form mixed anhydrides include sulfuric acid [20], sulfonic acids [21], and hydrochloric acid. The reaction of acetic anhydride with phosphorous acid leads to 1-hydroxyethane-1,1-diphosphonic acid [22]:

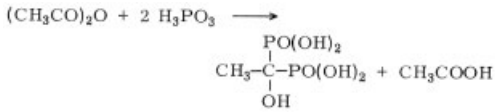

Acetylation of Oxides. Acetic anhydride reacts with antimony trioxide to give antimony triacetate, $Sb(OCOCH_3)_3$ [*5692-86-4*] [23]. The corresponding reaction with chromium trioxide

leads to chromyl acetate solutions [24], which often are used to oxidize olefins and hydrocarbons and which may occasionally explode [25].

Acetylation of Salts. Various salts react with carboxylic acid anhydrides in the same way as the corresponding free acids [26]:

$$\mathrm{Li\text{-}C \equiv CR + (CH_3CO)_2O} \rightarrow \mathrm{CH_3CO\text{-}C \equiv CR + CH_3COOLi}$$

With barium peroxide, diacetyl peroxide [*110-22-5*] is obtained [27]:

$$\mathrm{BaO_2 + (CH_3CO)_2O \longrightarrow (CH_3CO)_2O_2 + BaO}$$

A general method of producing vinyl ketones is the reaction of vinyl magnesium bromides and acetic anhydride [28]:

$$\mathrm{2\ CHR{=}CHMgBr + 2\,(CH_3CO)_2O \xrightarrow{THF} 2\,CHR{=}CHCOCH_3 + MgBr_2 + (CH_3COO)_2Mg}$$

Production of Acetoxy Silanes. Acetic anhydride reacts with silanes according to the equation:

$$\mathrm{-\overset{|}{\underset{|}{Si}}{-}X + (CH_3CO)_2O \longrightarrow -\overset{|}{\underset{|}{Si}}{-}OCOCH_3 + CH_3COX}$$

where X = H, Cl [29]; OR [30]; NR_2 [31].

Addition to Heterocyclic Compounds with Ring Cleavage. These reactions differ only formally from those described above. Examples are the production of ethylene glycol diacetate [*111-55-7*] from ethylene oxide in the presence of strongly acidic [32] or basic [33] catalysts and the production of oxymethylene diacetates from trioxane [34].

Oxidative Addition to Carbon-Carbon Double Bonds. Oxidative addition leads to the corresponding diacetates, as for example in the addition of ethylene, which leads to ethylene glycol diacetate [35]:

$$\mathrm{CH_2{=}CH_2 + (CH_3CO)_2O \xrightarrow[1/2\,O_2]{Re_2O_7} CH_3CO{-}OCH_2CH_2O{-}COCH_3}$$

1,4-Diacetoxy-2-butene [*18621-75-5*] can be obtained from butadiene in a similar manner [36].

Production of Mixed Diacyl Peroxides. Diacyl peroxides can be made by the reaction of oxygen with mixtures of aliphatic aldehydes and acetic anhydride in the presence of sodium acetate [37]:

$$\mathrm{2\,RCHO + (CH_3CO)_2O \xrightarrow{2O_2} 2\,RCO{-}OO{-}COCH_3 + H_2O}$$

Reaction with N-Oxides. Reactions with N-oxides lead to a variety of products, depending on the type of N-oxide. From pyridine-*N*-oxide, 2-acetyloxypyridine is obtained [38]. However, 4-picoline-*N*-oxide yields a mixture of 4-acetyloxymethylenepyridine and 3-acetyloxy-4-methyl- pyridine [39]. In the Polonovski reaction of N-oxides, formaldehyde and unsaturated aldehydes are formed as well as the acid amides [40], as in the following example:

$$\mathrm{(CH_3)_2\overset{\overset{O}{\uparrow}}{N}(CH_2)_n{-}CH{=}CH_2 \xrightarrow{(CH_3CO)_2O} CH_3\overset{\overset{COCH_3}{|}}{N}(CH_2)_n{-}CH{=}CH_2 + CH_2O + CH_3(CH_2)_n{-}CH{=}CH{-}CHO}$$

Reaction with S-Oxides. The reduction of sulfoxides to sulfides with acetic anhydride is known as the Pummerer reaction [41]:

$$\mathrm{CH_3SOR + (CH_3CO)_2O \rightarrow CH_3COOCH_2SR + CH_3COOH}$$

An interesting use of this reaction is the oxidation of primary and secondary alcohols, even when sterically hindered, by mixtures of dimethylsulfoxide (or tetramethylene sulfoxide) and acetic anhydride, to give the corresponding carbonyl compounds [42]:

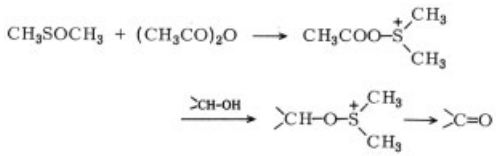

Production of Acylals and Vinyl Acetates. Aldehydes react with acetic anhydride in the presence of acid catalysts to form acylals:

$$RCHO + (CH_3CO)_2O \longrightarrow RCH(OCOCH_3)_2$$

If R has an α-hydrogen atom available, the corresponding vinyl acetate forms by elimination of acetic acid. The temperature required for the reaction can be lowered by adding catalysts [43–45].

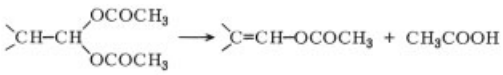

A similar process is particularly suitable for the synthesis of 2,2-dichlorovinyl acetate [*36597-97-4*]. Here, in the absence of an α-proton, cleavage of the acetic acid moiety is facilitated by the presence of zinc [46]:

$$Cl_3CCHO + (CH_3CO)_2O \rightarrow Cl_3CCH(OCOCH_3)_2 \xrightarrow{Zn} Cl_2C{=}CH{-}OCOCH_3 + ClZnOCOCH_3$$

Boron Trifluoride. Acetic anhydride and substituted acetic anhydrides can be converted into acetyl ketones via a boron trifluoride complex that decomposes in warm water [16], [47].

1.2.2. Dehydration

Acetic anhydride is used as a dehydrating agent, for example, in the explosives industry. The most important example of dehydration is in the production of hexogen (1,3,5-trinitrohexahydro-1,3,5-triazine) [*121-82-4*] [16]:

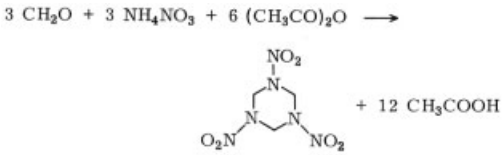

In the production of the nitroester of 1,2,4,5-tetrahydroxy-3,6-dinitrocyclohexane [48], the water-binding property of the anhydride is used again:

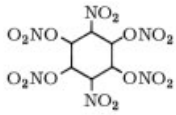

Alkyl cyanides can be made by dehydrating aldoximes with acetic anhydride [49].

$$RCH{=}NOH + (CH_3CO)_2O \longrightarrow RCN + 2\,CH_3COOH$$

Acetic anhydride serves as a dehydrating agent in a large number of cyclization reactions.

1.2.3. Reactions of the α-Protons

The Perkin reaction is used for the production of α,β-unsaturated acids from aromatic aldehydes, such as benzaldehyde, in the presence of potassium acetate or sodium acetate. This reaction is also suitable for synthesizing cinnamalacetic acid [50]:

$$C_6H_5CH{=}CHCHO + (CH_3CO)_2O \xrightarrow[140\,^\circ C]{CH_3COOK} CH_3COOH + C_6H_5CH{=}CHCH{=}CHCOOH$$

Other examples of this type of reaction are the oxidative carboxymethylation with acetic anhydride and oxidizing agents [51] and the conversion of long-chain alkenes to the corresponding carboxylic acid derivatives [52].

1.2.4. Reactions of a Single Carbonyl Group

In some cases, it is possible to make just one of the carbonyl groups of acetic anhydride participate in a reaction. Examples are the reaction with hydrogen cyanide, in the presence of a base, to form (**3**) [53] and in the presence of Grignard reagent (RMgBr) to form (**4**) [54]:

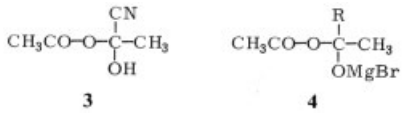

3 **4**

1.2.5. Production of Silver Ketenide

Silver acetate and acetic anhydride react at room temperature in pyridine to form a pyridine complex of silver ketenide. When an excess of anhydride is used, silver ketenide is obtained after

fractional distillation of pyridine and acetic acid [55], [56]:

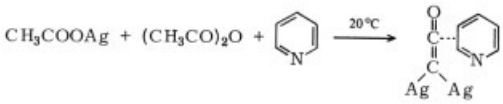

1.3. Production

History. The oldest process for making acetic anhydride is based on the conversion of sodium acetate with an excess of an inorganic chloride, such as thionyl chloride, sulfuryl chloride, or phosphoryl chloride. In this process, half of the sodium acetate is converted to acetyl chloride, which then reacts with the remaining sodium acetate to form acetic anhydride:

$$CH_3COONa + X\text{-}Cl \longrightarrow CH_3COCl + XONa$$

$$CH_3COONa + CH_3COCl \longrightarrow (CH_3CO)_2O + NaCl$$

where $X = SOCl, SO_2Cl, POCl_2$.

A further development, the conversion of acetic acid with phosgene in the presence of aluminum chloride, has the advantage that it allows continuous operation:

$$2\,CH_3COOH + COCl_2 \longrightarrow (CH_3CO)_2O + 2\,HCl + CO_2$$

Two other methods also were used in the past: the cleavage of ethylidene diacetate to form acetaldehyde and acetic anhydride in the presence of acid catalysts, such as zinc chloride, and the reaction of vinyl acetate with acetic acid on palladium(II) contacts to form acetaldehyde and acetic anhydride [16]. Not one of these processes is now of any industrial importance.

Today, acetic anhydride is made mostly by either the ketene process or the oxidation of acetaldehyde. Production by another process, the carbonylation of methyl acetate (Halcon process), was begun in 1983. In Western Europe, 77 % of acetic anhydride is made by the ketene process and 23 % by the oxidation of acetaldehyde. Since production by the Halcon process began at the Tennessee-Eastman plant, 25 % of acetic anhydride in the United States has been made by this process and 75 % by the ketene process.

1.3.1. Ketene Process

The ketene process for the production of acetic anhydride proceeds in two steps: the thermal cleavage of acetic acid to form ketene and the reaction of ketene with acetic acid:

$$CH_3COOH \rightarrow CH_2{=}C{=}O + H_2O \qquad \Delta H = 147\ \text{kJ/mol}$$

$$CH_2{=}C{=}O + CH_3COOH \rightarrow (CH_3CO)_2O \qquad \Delta H = -63\ \text{kJ/mol}$$

1.3.1.1. Production of Ketene

Thermal Cleavage of Acetic Acid. Hot acetic acid vapor is broken into ketene and water at 700 – 750 °C in the presence of traces of phosphoric acid catalyst. The pressure in the reactor is generally reduced so that the ketene can be isolated before it reacts with acetic acid or with water. The cleavage takes place in a multicoil reactor with coils made of highly heat-resistant steel alloys. The alloy Sicromal, containing 25 % Cr, 20 % Ni, and 2 % Si, is particularly suitable. Triethyl phosphate is employed as the catalyst [57].

Figure 2 shows how the vacuum process operates. Acetic acid is fed continuously into the evaporator (a). The acetic acid vapor leaving the evaporator passes over the catalyst evaporator (b), which is supplied continuously with catalyst, and thence into the ketene oven (c). The acetic acid is heated (d) and then cleaved (e). As soon as the cracked gases have left the oven, ammonia is added to prevent the ketene from reacting with the water or with the remaining acetic acid. For the same reason, and also to remove condensable gases, the hot gas is cooled in combined water and brine condensers (f), (g). A 40 – 48 % aqueous solution of acetic acid condenses at about 0 °C and is separated (h) from the cleavage gas (0 – 10 °C). About 4 – 8 % of this acetic acid results from the reaction of the condensed acetic anhydride with water. This process has been developed by the Consortium für Elektrochemische Industrie, the research institute of Wacker-Chemie.

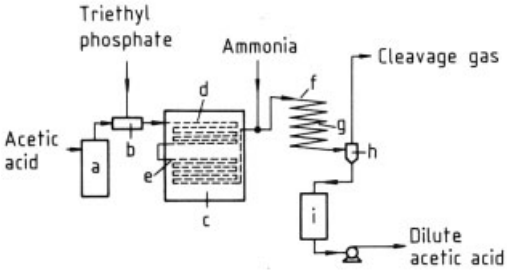

Figure 2. Cleavage of acetic acid by the vacuum process
a) Acetic acid evaporator; b) Catalyst evaporator; c) Ketene oven; d) Convection zone; e) Cleavage zone; f) Water cooler; g) Brine cooler; h) Separator; i) Receiver

The construction of the ketene oven depends strongly on the power required. In small ovens (ca. 50 t/month), separate preheating and cleaving ovens are preferred. The acetic acid is cleaved in a spiral tube which is usually electrically heated. Ovens of medium size (up to ca. 700 t/month), however, are often constructed as single-chamber ovens, generally heated with gas or oil. For capacities of 700 t/month, product loss is to be expected. This is because of the unsatisfactory pressure conditions in the cleavage zone. Ovens with much greater capacities work satisfactorily with three- and four-chamber systems and partial gas flows [58]. It is also possible to operate several ovens with acetic acid supplied from a central evaporator.

Thermal Cleavage of Acetone. This variation of the ketene process (cleavage of acetone into ketene and methane [59]) has no economic significance for the manufacture of acetic anhydride today.

1.3.1.2. Reaction of Ketene with Acetic Acid

Two processes are available for the reaction of ketene with acetic acid. The classical *scrubber process* is being replaced gradually by the *Wacker process* with liquid-ring pumping [60]. A further process [61], in which the reaction occurs at higher temperatures, has not yet gained industrial importance.

Scrubber Process (see Fig. 3). The ketene emerging from the separator (Fig. 2) is absorbed by glacial acetic acid circulating in scrubbers (a) and (b), which are filled with Raschig rings [62]. The circulation is maintained by centrifugal pumps and the heat of reaction removed by coolers (f). Most of the ketene is absorbed in scrubber (a), the rest in scrubber (b). The liquid mixture leaving scrubber (a) contains 85 – 90 % raw anhydride and is collected in (g), cooled and used in scrubber (c) to wash the off-gas from (b). The off-gas from (c) is finally washed with brine-cooled, dilute acetic acid in scrubber (d) in order to remove acid.

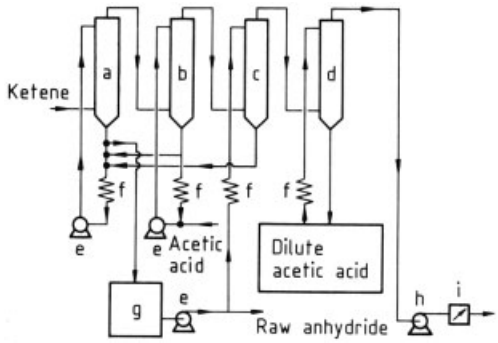

Figure 3. Scrubber process
a), b), c), d) Scrubbers; e) Centrifugal pumps; f) Coolers; g) Collecting vessel; h) Vacuum pump; i) Ring – balance manometer

Wacker Process (Liquid-ring pump process) (see Fig. 4). The ketene emerging from the separator (Fig. 2) is pumped through a Nash Hytor liquid-ring pump (a) at about 13 – 20 kPa. It reacts in the pump, at about 45 – 55°C, with acetic acid. At the same time, the liquid-ring pump provides the vacuum required for the acetic acid cleavage. Reaction and compression heat are removed by passing the raw anhydride through a cooler (c). This serves simultaneously as separator for the off-gas from the cleavage reactor. The off-gas is burnt in the ketene oven after passing the measuring point (d) or is led to a flare system. The raw anhydride (90 %) is collected continuously. Part of the anhydride is returned to the reaction pump, where it is used as the reaction medium. Advantages of this process stem from the impressive simplicity of the apparatus and operation, which lead to high reliability and the possibility of producing raw anhydride of higher purity [60].

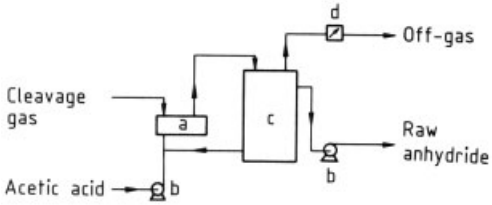

Figure 4. Wacker process (Liquid – ring pump process)
a) Liquid – ring pump; b) Pump; c) Cooler; d) Measuring point

1.3.1.3. Pure Anhydride Distillation

The raw anhydride can be distilled either continuously or discontinuously. In discontinuous distillation, three fractions are obtained:

Forerun (1 %) is about 40 % acetic acid; the rest consists of volatile liquids, particularly methyl acetate and acetone.
Middle run (about 10 %). Depending on the quality of the column, it contains 30 – 60 % acetic anhydride; the rest is acetic acid.
Pure Anhydride. Using discontinuous distillation at normal pressures, the purity of the anhydride cannot exceed 99 % (the remainder is acetic acid) because, at the still temperature required, a certain amount of decomposition occurs. These difficulties can be overcome by operating under reduced pressure.

For continuous distillation, two or three columns are used, depending on the required number of fractions (forerun and middle run are separated). The middle run portion is returned to the reactor as reagent acid. The residue of the distillation contains small amounts of tarry products and is worked up continuously in a Sambay evaporator.

A newer continuous and energy-saving vacuum process works with two evaporators and only one column [63].

The *Wacker process,* described above (see Fig. 4), is also particularly suitable as a workup process. Waste acids, such as those from cellulose acetylation, can be processed after concentration without difficulty. No auxiliary materials are necessary, only energy in the form of gas, oil, or electric current. The process has been adopted by many firms. For the production of 100 kg acetic anhydride, about 122 kg acetic acid is required, taking account of the reconcentrated dilute acetic acid. The yield is over 96 % at about 75 % cleavage.

1.3.1.4. Environmental Problems

No significant environmental problems arise with the processes described above. The off-gas occurring in the production of ketene consists of ca. 45 % carbon monoxide, 15 – 20 % carbon dioxide, ca. 15 % ethylene, ca. 10 % methane, ca. 7 % propylene, and less than 1 % acetic acid and ethane. The remainder is air. The off-gas can be burnt in the ketene oven saving energy (ca. 10 %), or led to a flare system. There are no wastewater problems.

1.3.2. Oxidation of Acetaldehyde

Acetic anhydride can be obtained directly by liquid-phase oxidaton of acetaldehyde. The peracetic acid formed from oxygen and acetaldehyde reacts under suitable conditions with a second molecule of acetaldehyde to form acetic anhydride and water [64]:

$$CH_3CHO + O_2 \rightarrow CH_3CO_3H \xrightarrow{CH_3CHO} (CH_3CO)_2O + H_2O$$

Rapid removal of the reaction water and the use of suitable catalysts are essential in this process. Mixtures of acetic acid and acetic anhydride are always obtained; their ratio can be varied within wide limits by changing the reaction conditions. Generally, the highest possible anhydride yield is sought.

Because of the rapid hydrolysis of acetic anhydride above 60 °C, the process is operated preferably between 40 °C and 60 °C [65]. Suitable catalysts are combinations of metal salts [66]. Particularly important are mixtures of manganese acetate and copper acetate [67], of cobalt acetate and nickel acetate, and of cobalt and copper salts of higher fatty acids [68]. Manganese acetate should hinder the formation of explosive amounts of peracetic acid during the oxidation of acetaldehyde. For increasing the rate of oxidation, the use of pure oxygen at a pressure of several hundred kilopascals instead of air has been proposed [69].

The strongly exothermic reaction requires efficient cooling. For this purpose, the addition of low-boiling solvents has been found to be of assistance. Methyl and ethyl acetates are favored because they form azeotropic mixtures with water (but not with acetic acid or acetic anhydride) and hence allow a rapid, continuous separation of the water formed in the reaction. The ratio of acetic anhydride to acetic acid in the product depends on the ratio of ethyl acetate to acetaldehyde in the initial mixture (Table 1).

Table 1. Formation of acetic anhydride by oxidation of acetaldehyde

Ethyl acetate: acetaldehyde in starting mixture	Acetaldehyde conversion, %	Acetic anhydride yield based on acetaldehyde, %
20 : 80	80	13.5
30 : 70	80	57
60 : 40	80	64
70 : 30	80	68.5

In practice, a 1 : 2 mixture of acetaldehyde and ethyl acetate is oxidized with the addition of 0.05 to 0.1 % cobalt acetate and copper acetate at 40 °C; the ratio of Co:Cu is 1:2. The ratio of acetic anhydride to acetic acid obtained is 56:44, whereas on oxidizing in the absence of ethyl acetate this ratio is only 20:80 [16]. The optimization of other reaction conditions can also lead to an increase in the acetic anhydride-acetic acid ratio. For example, at 55 °C and atmospheric pressure, a ratio of 80:20 was achieved [70]. At a higher temperature (62 – 90 °C, 200 – 300 kPa, acetaldehyde concentration in the final mixture of up to 40 %) a ratio of 75:25 was obtained at high aldehyde conversion [71].

Other suitable low-boiling solvents are methylene chloride, diisopropyl ether, cyclohexanone, or ethylidene diacetate. Nonvolatile esters also can be used as diluents, provided they do not have to be removed from the reaction zone. These include alkyl benzoates and alkyl phthalates [72].

The acetaldehyde oxidation is illustrated in Figure 5 by the process of Usines de Melle [73]. The gas mixture containing oxygen and acetaldehyde is pumped into the reactor (a). The oxidation takes place in the liquid phase and in the presence of catalysts. The reactor effluent is sent through a water-cooled condenser (b) constructed as a separator; non-condensable gases are sent to the packed column (c). Fresh acetaldehyde is introduced at the top of this column. The condensates from both the cooler (b) and the column (c) are distilled to obtain the product. Acetaldehyde is recovered from the branch stream (d) of the non-condensable gas. The other part of the gas flow is supplemented with air and returned to the reactor.

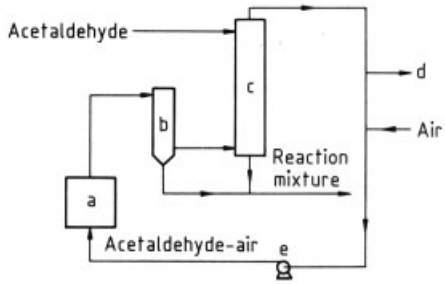

Figure 5. Acetaldehyde oxidation
a) Reactor; b) Condenser; c) Column; d) Branch stream (acetaldehyde recovery from non-condensable gas); e) Pump

Both towers and vessels are suitable as reactors if the heat of reaction can be dissipated. The process of Distillers Co. [69] is shown in Figure 6 as an example. The off-gas contains combustible low-boiling products, such as acetaldehyde, and solvents, such as methyl acetate and ethyl acetate. These can be flared off.

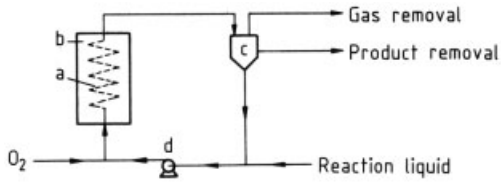

Figure 6. Reactor for the acetaldehyde oxidation (Distillers Co.)
a) Reactor tubing; b) Cooling bath; c) Separator; d) Circulation pump

1.3.3. Carbonylation of Methyl Acetate

The thermal decomposition of acetic acid to form ketene requires a large amount of energy, a disadvantage of the conventional process for the production of acetic anhydride. Moreover, processes based on synthesis gas have been developed that allow the manufacture of products from coal that were hitherto produced from oil. An important example is the acetic acid produc-

tion process developed by Monsanto (→ Acetic Acid).

In 1973 Halcon patented the carbonylation of methyl acetate in the presence of a rhodium catalyst to form acetic anhydride [74]. However, the first plant (Eastman-Kodak Co., in Kingsport, Tenn., USA) using this process was not put into operation before 1983. Because no exact description of the process has so far been available, the following details are based upon information from the patent literature.

Methyl acetate is carbonylated to acetic anhydride in the liquid phase at a temperature of 160 – 190 °C and at a carbon monoxide partial pressure of 2 – 5 MPa:

$$CH_3COOCH_3 + CO \overset{\text{catal.}}{\rightleftharpoons} (CH_3CO)_2O \quad \Delta H = -94.8\ \text{kJ/mol}$$

The starting material, methyl acetate, can be produced by esterification of acetic acid with methanol. However, the methyl acetate obtained as a byproduct of the acetic acid synthesis from methanol and carbon monoxide is used preferably.

Catalysts. Rhodium [74] and nickel compounds [75], activated by CH_3I, HI, LiI, I_2, or other iodides, are particularly appropriate as catalysts. Rhodium catalysts have about a tenfold higher activity than nickel catalysts. The selectivity is higher than 95 % for both catalysts. Their activities and lifetimes are increased by mixing the carbon monoxide used for the synthesis with 2 – 7 % hydrogen. Chromium compounds have been used to shorten the induction phase of the reaction [74]. Kinetic investigations on the rhodium system have shown that the reaction is zero order with respect to the methyl acetate and carbon monoxide concentrations [76]. As possible byproducts, only acetic acid and ethylidene diacetate are mentioned.

The process also can start from dimethyl ether [74]. In this case, dimethyl ether is first carbonylated to methyl acetate, which is then converted to acetic anhydride by using more carbon monoxide in the same reactor:

$$CH_3OCH_3 \overset{CO}{\rightleftharpoons} CH_3COOCH_3 \overset{CO}{\rightleftharpoons} (CH_3CO)_2O$$

Process Description. The carbonylation method is illustrated by the Halcon process, shown schematically in Figure 7.

The methyl acetate is dried with acetic anhydride [77] and is sent to a reactor (b) lined with Hastelloy. Carbon monoxide is compressed (a) to the reaction pressure and then added. The reaction proceeds continuously at 175 °C in the presence of a catalyst consisting of $RhCl_3 \cdot 3\,H_2O$, CH_3I, and LiI. The considerable amount of heat generated by the reaction is removed by heat exchange and is used both for preheating methyl acetate and for the production of low-pressure steam. The unreacted carbon monoxide leaves the top of the reactor, is freed of condensable gases (methyl iodide, methylacetate, acetic anhydride, acetic acid, and ethylidene diacetate) by cooling (f), and is then recirculated. However, part of the circulating gas is separated from the main stream to avoid buildup of inert gases, which may be present in the carbon monoxide. The side stream is washed (g) with a countercurrent of pure acetic anhydride; in this way, the loss of methyl iodide in the offgas can be kept below 0.1 % [78]. This acetic anhydride is combined with the top condensate from the reactor (b), supplemented with fresh catalyst as required, and recirculated. The liquid reaction product leaving the reactor is expanded and subjected to flash distillation (c) [78]. To prevent catalyst decomposition, this distillation is performed in a carbon monoxide-hydrogen atmosphere at about 500 kPa [79]. At the bottom of the evaporator (c) a stream of liquid containing the catalyst is separated and recirculated into the carbonylation reactor. The vapor leaving the top of the evaporator is condensed and passed over adsorbers (d) to remove traces of rhodium and iodine compounds.

The raw anhydride is purified by distillation in three consecutive and continuously operating columns (e). In the first column, methyl iodide and methyl acetate are distilled overhead and recirculated to the carbonylation reactor. In the second column, acetic acid is distilled overhead. The bottom product is distilled in a third column to acetic anhydride of 99 % purity. The bottom product of the third column contains ethylidene diacetate and unidentified high-boiling components. To further reduce the iodide content of the pure anhydride, a solution of potassium acetate in acetic anhydride is added to the top of the column [80].

Other processes. Other processes and catalysts for the carbonylation of methyl acetate have

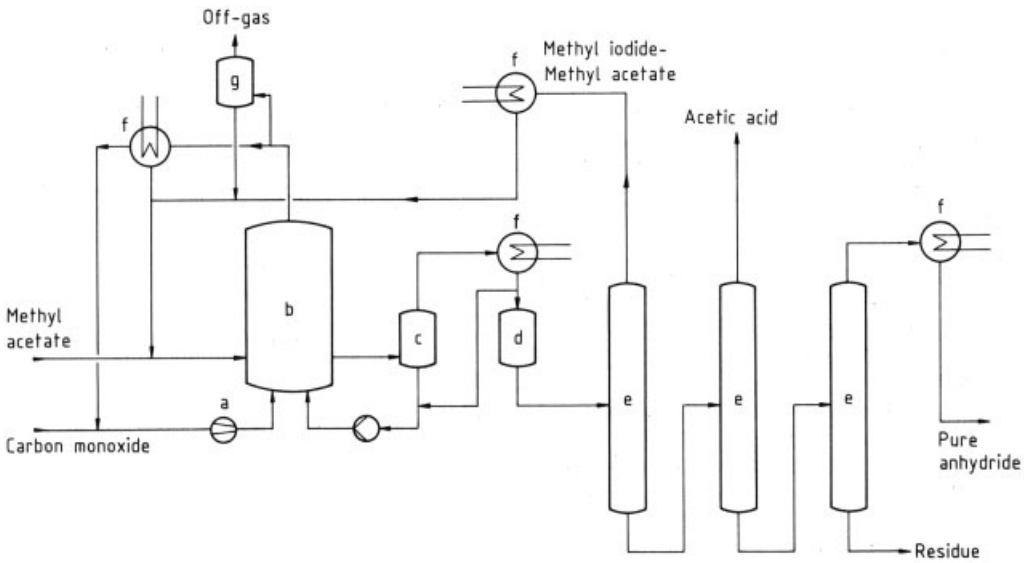

Figure 7. Halcon process for the production of acetic anhydride
a) Compressor; b) Carbonylation reactor; c) Evaporator; d) Adsorber; e) Distillation column; f) Condenser; g) Scrubber

been patented by various companies: Hoechst [81] (rhodium catalyst), Air Products and Chemicals Co.[82] (rhodium catalyst), Mitsubishi Gas and Chemical Co. [83] (nickel catalyst), and Rhône-Poulenc [84] (nickel catalyst). Several patents also describe the carbonylation of esters of higher alcohols and carboxylic acids. This reaction results in the formation of the corresponding mixed anhydrides [74], [83].

Environmental problems do not arise. The off-gas from the acetic anhydride production contains large quantities of carbon monoxide, some inert gases (nitrogen, rare gases, and carbon dioxide), and traces of hydrogen, methane, methyl iodide, and methyl acetate. It can be burned. There are no problems with waste water.

1.4. Analysis

A very reliable method for determining the content of acetic anhydride consists of mixing stoichiometric quantities of the sample and water, then heating carefully to reflux temperature. After completion of the reaction, two drops of concentrated sulfuric acid are added, and the mixture is boiled for 20 min to insure that the last traces of the anhydride have reacted. The anhydride content is calculated from the unused water, which is determined by the Karl Fischer method.

In the *aniline method* [85], the total acid content is first determined. After addition of aniline to a second sample, the aniline number is established with alkali and the acetic anhydride content determined from the difference.

A rapid determination of the anhydride content can be obtained from the refractive index. If the temperature remains constant to within $\pm 0.05\,^{\circ}\mathrm{C}$, the precision obtainable is ± 0.2 %. Gas chromatography also is recommended for purity determination. Analytical methods are discussed in detail in [86].

1.5. Quality Specifications

Acetic anhydride is marketed with more than 95 % purity; the normal product is over 98 % pure, but it is marketed also as over 99 % pure. The color number (APHA) should be below 10 (DIN 53409). The nonvolatile part should not exceed 0.003 %. The product also should contain as few substances as possible that reduce permanganate. According to American Chemical Society specifications, for example, a 2 g sample should not decolorize 0.4 ml of a 0.1 *N* potassium permanganate solution within 5 min. In particular applications, the impurities that

can be oxidized by potassium chromate are also of interest. They may not consume more than 200 ppm of oxygen. The contents of phosphate, sulfate, chloride, aluminum, and iron may not exceed 1 ppm each. Heavy metals should be absent.

1.6. Storage and Transportation

For storage and transportation of pure acetic anhydride, tanks made of aluminum, stainless steel (18 % Cr, 8 % Ni, and 2 % Mo), or polyethylene normally are used, although glass or enamel containers also may be employed. Iron is highly resistant to acetic anhydride, provided moisture is excluded. Therefore it is possible to use iron in the production and workup in certain instances, for example, in pumps and tanks.

Because there are no international arrangements for the storage of dangerous goods, the specifications of individual countries must be observed. Also, acetic anhydride (EG-no. 607–008–00–9) is subject to various industrial working regulations, for example, Appendix I, no. 1.1 of the EEC Guidelines.

Transportation. IMDG-Code, class 8, UN-no. 1715; United Kingdom: Blue Book: Corrosive, IMDG-Code E 8018; United States: DOT Safety Act, Title 46 and Title 49; Cor. M; Europe: RID, ADR, and ADNR: class 8, no. 21 e (from 1985: class 8, no. 32 e), RN 801, 2801, and 6801, respectively. International air transportation: IATA-DGR, class 8, UN-no. 1715, RAR art. no. 9, Cor. M.

1.7. Uses

Acetic anhydride is used chiefly as an acetylating and dehydrating agent; it is used on a large scale for the acetylation of cellulose. Other areas of application for acetic anhydride are:

1) The production of poly(methylacrylimide) hard foam, where acetic anhydride is used for binding the ammonia that is liberated on conversion of two amide groups to an imide group.
2) Acetylated plastic auxiliaries, such as glycerol triacetate, acetyl tributyl citrate, and acetyl ricinolate.
3) Explosives, particularly hexogen production (see Section 1.2.2.).
4) The production of certain types of brake fluids.
5) The production of auxiliaries for drilling fluids.
6) The detergent industry, for the production of cold – bleaching activators such as tetraacetylethylenediamine [12].
7) The dyeing industry, where acetic anhydride is used chiefly in mixtures with nitric acid as a nitrating agent. Here, the solvent and dehydrating properties of acetic anhydride are used.
8) In the preparation of organic intermediates, such as chloroacetylchloride, diacetyl peroxide, higher carboxylic anhydrides, acetates, and the boron trifluoride complex.
9) In the production of pharmaceuticals, such as acetylsalicylic acid, *p*-acetylaminophenol, acetanilide, acetophenacetin, theophyllin, acetylcholine chloride, sulfonamides, a number of hormones and vitamins, and the x-ray contrast agent 2,4,6-triiodo-3,5-diacetylamidobenzoic acid.
10) In the food industry, mainly in the acetylation of animal and plant fats, in order to obtain the desired solubilities; in the production of acetostearins, the edible packing materials; and to clarify plant oils.
11) Flavors and fragrances (production of esters and cumarin).
12) Herbicides such as metolachlor (Dual) and alachlor (Lasso).

2. Mixed Fatty Acid Anhydrides

2.1. Physical Properties

Important mixed anhydrides, that is, anhydrides with two different fatty acid radicals, are compiled in Table 2. The lower anhydrides can be distilled partly undecomposed at reduced pressure. At high temperature they disproportionate fairly rapidly into the symmetrical anhydrides. Only acetoformic anhydride distills (at 127 – 130 °C) at normal pressure, although with partial decomposition to acetic acid and carbon monoxide. The first four mixed anhydrides have odors very similar to pure acetic anhydride; acetic isovaleric anhydride has a fruitlike odor.

Table 2. Mixed fatty acid anhydrides

Anhydride	Formula	*bp* or *mp*, °C	Preferred production method (Section 2.3)
Acetoformic anhydride	$HCOOCOCH_3$	*bp* 32 (at 2.7 kPa)	3
Acetic propionic anhydride	$CH_3COOCOC_2H_5$	*bp* 25 – 27 (at 0.15 kPa) (decomposes at 30 °C)	3
Acetic butyric anhydride	$CH_3COOCO(CH_2)_2CH_3$	–	3
Acetic isobutyric anhydride	$CH_3COOCOCH(CH_3)_2$	–	3
Acetic valeric anhydride	$CH_3COOCO(CH_2)_3CH_3$	(decomposes above 130 °C)	3
Acetic isovaleric anhydride	$CH_3COOCOCH_2CH(CH_3)_2$	–	3
Acetic 2,2-dimethylpropanoic anhydride	$CH_3COOCOC(CH_3)_3$	–	3
Acetic hexanoic anhydride	$CH_3COOCO(CH_2)_4CH_3$	–	3
Acetic octanoic anhydride	$CH_3COOCO(CH_2)_6CH_3$	–	3
Acetic hexadecanoic anhydride	$CH_3COOCO(CH_2)_{14}CH_3$	*mp* 62.5 (decomposes on crystallization from polar solvents)	2,3
Acetic octadecanoic anhydride	$CH_3COOCO(CH_2)_{16}CH_3$	–	3
Butyric tetradecanoic anhydride	$CH_3(CH_2)_2COOCO(CH_2)_{12}CH_3$	*mp* 52.7	2
Hexanoic dodecanoic anhydride	$CH_3(CH_2)_4COOCO(CH_2)_{10}CH_3$	*mp* 42.4	2
Octanoic decanoic anhydride	$CH_3(CH_2)_6COOCO(CH_2)_8CH_3$	*mp* 16	2

2.2. Chemical Properties

Acetoformic anhydride [*2258-42-6*] acts as a formylation agent in acylation reactions. In the other mixed acetic – fatty acid anhydrides, the higher acyl group is more reactive; for example, in the reaction with benzene in the presence of aluminum chloride, compound (**5**) is formed preferably.

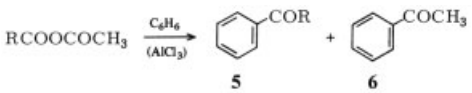

where R = alkyl.

Branching of the higher acyl group at its α position leads to a decrease in the yield of the acylbenzene (**5**) in favor of acetophenone (**6**) [87]. From acetic chloroacetic anhydride, mainly ω-chloroacetophenone is formed [87]. Changes in reaction conditions have little influence on the ratio of reaction products.

2.3. Production

Three processes are generally employed for producing mixed anhydrides:

1) The reaction of acetic anhydride with a higher organic acid [88].
2) Heating an acyl chloride with a fatty acid salt, usually a sodium or potassium salt [89–91].
3) Reacting ketene with a carboxylic acid [92–94].

The latter two processes are more useful than the first. According to the patent literature, mixed anhydrides also can be produced by carbonylating the corresponding esters (Section 1.3.3).

2.4. Uses

Acetoformic anhydride often is used as a formylating agent. Acetic propionic [*13080-96-1*] and acetic butyric [*7165-13-1*] anhydrides are used in the production of acylated cellulose.

3. Economic Aspects

Production data are given in Table 3. Annual capacity data for acetic anhydride are as follows: Western Europe (1978) 357 000 t, (1980) 362 000 t, (1996) 500 000 t; United States (1980) 905 000 t, (1982) 929 000 t, (1996) 1 000 000 t, (2001) 1 160 000 t. No data for the mixed anhydrides are available.

4. Toxicology and Occupational Health

In acute oral studies in rats, an LD_{50} value of 1780 mg/kg was determined for acetic anhy-

Table 3. Production of acetic anhydride (in kt) [95], [96], [97], [98]

	1961	1971	1974	1979	1980	1981	1982	1989	1990	1996	2001
United States	571	686	741			567	481	778	830	1000	1160
Germany*)	32	47	74	91	85	77	76	108	112		
Japan	33	96	115	114	150	145	144	205	1160		

*) 1961 - 1989 Fed. Rep. of Germany

dride. The dermal LD_{50} in rabbits is 4000 mg/kg [99]. Inhalation of 2000 ppm for 4 h causes death in rats [100]. The lowest published lethal concentration (LCLo) for a 4-h exposure of rats is 1000 ppm [101].

In rabbits, skin contact for 24 h causes only mild irritation [99]. These findings are in accordance with observations in humans. Tingling sensations followed by slight erythema 30 min after skin contact have been reported. Pain or vesication has not been observed [102]. In contrast to these observations, acetic anhydride is reported to cause wrinkling, whitening, and peeling, if it is not removed from the skin at once [103]. Acetic anhydride is a severe eye irritant [104]. In rabbits, 250 µg causes severe irritation. Human occupational exposure can cause lacrimation, conjunctivitis, photophobia [105], corneal burns with loss of vision [106], and iritis [107]. It is assumed that acetic anhydride penetrates the corneal epithelium rapidly without hydrolyzing and reaches the iris in concentrations high enough to cause iritis.

*Inhalation*can lead to asthmoid bronchitis with lung edema [108]. Human exposure to concentrations of 800 ppm cannot be tolerated for longer than 3 min because of the burning sensation in the nose and throat. Humans are believed to be generally more sensitive to the irritant effects of inhalation of vapors than animals [109]. *Ingestion* is followed by a burning pain in the stomach, nausea, and vomiting [110]. The acute irritant effects are caused partly by the hydrolysis of acetic anhydride to acetic acid, which itself is an irritant. The rate of hydrolysis depends on the water content of the tissues.

Systemic effects after repeated exposure are unlikely to occur because of the potent warning properties of acetic anhydride. The MAK and TLV are both 5 ppm (20 mg/m^3) [111], [112].

5. References

1. D. T. Lewis, *J. Chem. Soc.* 1940, 33.
2. W. J. Lewis, E. J. Evans, *Phil. Mag.* **13** (1932) 268.
3. L. H. Meyer, A. Saika, H. S. Gutowsky, *J. Am. Chem. Soc.* **75** (1953) 4569; L. H. Allred, E. G. Rochow, *J. Am. Chem. Soc.* **79** (1957) 5361.
4. L. P. Filippov, *Vestnik Moskov. Univ.* **9** (1954) 45; *Chem. Abstr.* **15430** (1955).
5. G. Jander, H. Surawski, *Z. Elektrochem.* **65** (1961) 469.
6. T. V. Malkova, *Zh. Obshch. Khim.* **24** (1954) 1157; Engl. ed. 24 (1954) 1151; K. N. Kovalenko et al., Zh. Obshch. Khim. 26 (1956) 403; Engl. ed. 26 (1959) 427.
7. G. Jaudes, E. Rüsberg, H. Schmidt, *Z. Anorg. Chem.* **255** (1948) 238.
8. *Beilstein,* **4,** 2nd suppl., 387.
9. D. H. Kim, *J. Heterocycl. Chem.* **13** (1976) 179.
10. K. Murihara, *Bull. Chem. Soc. Japan* **37** (1964) 1787.
11. S. S. Parmar, R. C. Arora, *J. Med. Chem.* **13** (1970) 135.
12. BASF DE-OS 2 118 281, 1971; NL 7 204 899, 1972 (C. Palm, G. Matthias).
13. Henkel u. Cie., DE-OS 2 112 557, 1971.
14. J. Hori, H. Midorikawa, *Sci. Pap. Inst. Phys. Chem. Res. (Tokyo)* **56** (1962) 216.
15. A. N. Nesmejanov, K. N. Anisimov, G. K. Magornedov, *Izv. Akad. Nauk SSSR, Ser. Khim.* **4** (1970) 959.
16. *Ullmann,* 3rd ed., **6,** 804.
17. F. Merenyi, N. Nilsson, *Acta Chem. Scand.* **18** (1964) 1368.
18. Houben-Weyl: *Methoden der organischen Chemie,* 4th ed., **vol. VII/2a,** Thieme Verlag, Stuttgart 1983, p. 457.
19. Techn.-Chem., DE-OS 1 940 809, 1969 (D. Sheehan, W. P. Hegarty, A. F. Vellturo, W. A. Gay, D. D. Threlkeld).
20. A. Cassadevall, A. Commeyras, *Bull. Soc. Chim. Fr.* **5** (1970) 1850.
21. A. Cassadevall, A. Commeyras, *Bull. Soc. Chim. Fr.* **5** (1970) 1856.

22. B. Blaser, K.-H. Worms, H. G. Germscheid, K. Wollmann, *Z. Anorg. Allg. Chem.* **381** (1971) 247.
23. G. Gattow, H. Schwank, *Z. Anorg. Allg. Chem.* **382** (1971) 49.
24. J. G. Dawber, *Chem. Ind.* **23** (1964) 973.
25. J. Leleu, *Cah. Notes Doc.* **83** (1976) 281.
26. R. Finding, U. Schmidt, *Angew. Chem.* **82** (1970) 482; *Angew. Chem. Int. Ed. Engl.* **9** (1970) 456.
27. J. D'Ans, J. Mattner, *Chem. Ztg.* **74** (1950) 435.
28. M. S. Newman, A. S. Smith, *J. Am. Chem. Soc.* **67** (1945) 154; *J. Org. Chem.* **13** (1949) 592.
29. H. Kelling, *J. Chem. Soc.* **8** (1968) 391.
30. R. C. Mehrotra, *Pure Appl. Chem.* **13** (1966) 111.
31. J. L. Speier, C. A. Roth, J. W. Ryan, *J. Org. Chem.* **36** (1971) 3120.
32. V. F. Shvets, I. Al-Vakhib, *Kinet. Katal.* **13** (1972) 98.
33. V. F. Shvets, I. Al-Vakhib, *Kinet. Katal.* **16** (1975) 785.
34. J. Tomiska, *Collect. Czech. Chem. Commun.* **28** (1963) 1612.
35. Union Oil, US 3 393 225, 1965 (D. M. Fenton).
36. Toyo Soda Mfg Co., JA 7 319 293, 1969 (T. Ono, T. Yanagihara, H. Okada, T. Koga).
37. J. A. Ol'Dekop, A. N. Sevcenko, J. P. Zjat'Kov, A. P. El'Nickij, *Zh. Obshch, Khim.* **33** (1963) 2771.
38. C. Rüchardt, S. Eichler, O. Krätz, *Tetrahedron Lett.* **4** (1965) 233.
39. H. Iwamura, M. Iwamura, T. Nishida, S. Sato, *J. Am. Chem. Soc.* **92** (1970) 7474.
40. M. Ferles, M. Jankovsky, *Collect. Czech. Chem. Commun.* **36** (1971) 4103.
41. C. R. Johnson, J. C. Sharp, W. G. Phillips, *Tetrahedron Lett.* **52** (1967) 5299.
42. J. D. Albright, L. Goldman, *J. Am. Chem. Soc.* **87** (1965) 4214.
43. R. P. Arganbright, R. J. Evans, *Hydrocarbon Process.* **43** (1964) 159.
44. American Home Products Corp., US 3 663 605, 1970 (R. J. McCaully, G. L. Conklin).
45. Y. Masada, *Kogyo Kagaku Zasshi* **74** (1971) 1149.
46. A. N. Mirskova, E. F. Zorina, A. S. Atavin, *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk* **6** (1971) 72.
47. H. Musso, K. Figge, *Justus Liebigs Ann. Chem.* **668** (1963) 1.
48. ICI, GB 1 107 907, 1965 (A. H. Dinwoodie, G. Fort).
49. J. S. Buck, W. S. Ide in: *Organic Synthesis,* **collective vol. II,** Wiley-Interscience, New York 1943, p. 622.
50. M. Tsuda, H. Tanaka, K. Ikeda, *J. Chem. Soc. Japan, In. Chem. Sect.* **73** (1970) 1888.
51. P. L. Southwick, *Synthesis* **12** (1970) 628.
52. G. T. Nikisin, J. N. Ogibin, J. A. Palanuer, *Izv. Akad. Nauk SSSR* **11** (1967) 2478.
53. B. F. Goodrich Co., DE-AS 1 086 683, 1953 (L. F. Arnold).
54. W. R. Edwards, K. P. Kamman, *J. Org. Chem.* **29** (1964) 913.
55. H. Eck, H. Spes in: J. Falbe (ed.), *Methodicum Chimicum,* **vol. 5,** p. 498.
56. E. T. Blues, D. Bryce-Smith, H. Hirsch, M. J. Simons, *Chem. Commun.* **11** (1970) 699.
57. Consortium für Elektrochemische Industrie, DE 408 715, 1922; DE 417 731, 1924; DE 475 885, 1926; DE 488 573, 1926 (R. Meingast, M. Mugdan); DE 634 438, 1933; DE 687 065, 1933; DE 734 349, 1934; US 2 249 543, 1937 (M. Mugdan, J. Sixt).
58. Eastman Kodak, US 3 403 181, 1966 (E. S. Painter, R. C. Petrey, J. H. Jensen).
59. *Ullmann,* 4th ed., **11,** 80.
60. Wacker-Chemie, DE 1 076 090, 1959 (Th. Altenschöpfer, H. Spes, L. Vornehm).
61. BASF, DE-OS 2 005 970, 1970 (G. Matthias, G. Schulz, C. Palm).
62. Consortium für Elektrochemische Industrie, DE 403 863, 1922 (R. Meingast, M. Mugdan).
63. Wacker-Chemie, DE 2 505 471, 1975 (H. Eck, H. Schwarzbauer, E. Bethe, K. Kaiser, H. Spes).
64. Wacker-Chemie, DE 867 689, 1940 (A. Krug, J. Sixt).
65. Carbide & Carbon Chem. Co., US 2 225 486, 1940 (H. L. Reichart).
66. A. G. f. Stickstoffdünger, DE 699 709, 1934 (J. Lösch, F. Walter, H. Behringer, O. Schlöttig).
67. A. G. f. Stickstoffdünger, DE 708 822, 1934 (J. Lösch, F. Walter, H. Behringer, O. Schlöttig).
68. H. Dreyfus, GB 510 959, 1938.
69. Distillers Co., US 2 514 041, 1946 (A. Elce, H. M. Stanley, K. H. W. Tuerck).
70. D. V. Musenko, G. N. Gvozdovskij, *Zh. Vses. Khim. Ova.* **14** (1969) 263.
71. Hoechst, DE-OS 2 757 222, 1977; EP 0 002 696, 1978; US 4 252 983, 1981 (H. Erpenbach, K. Gehrmann, A. Hauser, K. Karrenbauer, W. Lork).
72. Les Usines de Melle, US 2 658 914, 1950 (L. Rigon). Hoechst, DE 2 757 173, 1977 (H. Erpenbach, K. Gehrmann, A. Hauser, K. Karrenbauer, W. Lork).

73. Les Usines de Melle, DE-AS 1 142 857, 1963 (L. Alheritierc).
74. Halcon International, DE-OS 2 441 502, 1974 (C. Hewlett).
75. Halcon International, US 4 002 678, 1975 (A. N. Naglieri, N. Rizkalla).
76. M. Schrod, G. Luft, *Ind. Eng. Prod. Res. Dev.* **20** (1981) 649.
77. Halcon Research and Development Corp., GB 2 033 385 A, 1980 (C. G. Wan).
78. Halcon Research and Development Corp., US 4 241 219, 1979 (C. G. Wan).
79. Halcon Research and Development Corp., DE-OS 2 940 752, 1979 (R. V. Procelli, V. S. Bhise, A. J. Shapiro).
80. Halcon Research and Development Corp., GB 2 033 901 A, 1980 (P. L. Szecsi).
81. Hoechst, DE-OS 2 836 084, 1978 (H. Erpenbach, K. Gehrmann, H. K. Kübbeler).
82. Air Products and Chemicals, US 4 333 885, 1981 (D. Feitler).
83. Mitsubishi Gas Chemical Co., DE-OS 2 844 371, 1978 (T. Isshiki, Y. Kijima, Y. Miyauchi).
84. Rhône-Poulenc Industries, EP 0 050 084, 1980 (J. Gauthier-Lafaye, R. Perron).
85. A. Menschutkin, B. Wasiljeff, *Z. Anal. Chem.* **60** (1921) 425; T. Ellerington, J. J. Nichols, *Analyst London* **82** (1957) 233.
86. E. F. Joy, A. J. Barnard in: F. D. Snell, C. L. Hilton (eds), *Encyclopedia of Industrial Chemical Analysis,* **vol. 4,** J. Wiley & Sons, New York 1967, p. 102.
87. W. R. Edwards, E. C. Sibille, *J. Org. Chem.* **28** (1963) 674.
88. W. Autenried, G. Thomae, *Ber. Dtsch. Chem. Ges.* **34** (1901) 168; **57** (1924) 423.
89. A. W. Ralston, R. A. Reck, *J. Org. Chem.* **11** (1946) 625.
90. J. B. Polya, T. M. Spotswood, *J. Amer. Chem. Soc.* **71** (1949) 2938.
91. BASF, EP 0 029 176, 1980 (K. Blatt, H. Naarmann).
92. De Bataafsche, GB 389 049, 1932.
93. W. Stevens, A. van Es, *C. R. Acad. Sci. Ser. B* **83** (1964) 863.
94. J. Dickert, A. Krynitsky, *J. Am. Chem. Soc.* **63** (1941) 2511.
95. *Chem. Eng. News* **61** (1983) no. 24, 29.
96. *Chemfacts Japan,* Chemical Data Services, IPC Industrial Press, Sutton, UK 1981.
97. K. Weissermel, H.-J. Arpe *Industrielle Organische Chemie,* Wiley-VCH 1998.
98. *Kirk-Othmer,* 4th ed. online, John Wiley & Sons, NY 2002.
99. *Registry of Toxic Effects of Chemical Substances,* U.S. Department of Health, Education, and Welfare, Public Health Service, Center for Disease Control, NIOSH, Cincinnati, Ohio 1980.
100. H. F. Smyth Jr., C. P. Carpenter, U. C. Pozzani, *A. M. A. Arch. Ind. Hyg. Occup. Med.* **10** (1954) 61.
101. *Toxic and Hazardous Industrial Chemicals Safety Manual,* The International Technical Information Institute, Tokyo 1975, p. 3.
102. H.-J. Oettel, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **183** (1936) 641.
103. E. R. Plunkett, *Handbook of Industrial Toxicology,* Chemical Publ. Co., New York 1968, p. 4.
104. F. A. Patty, *Industrial Hygiene and Toxicology,* 2nd ed., **vol. II,** Interscience, New York 1962, p. 1817.
105. A. Hamilton, H. L. Hardy, *Industrial Toxicology,* 2nd ed., P. B. Hoeber, New York 1949, p. 338.
106. R. S. McLaughlin, *Am. J. Ophthalmol.* **29** (1946) 1355.
107. J. Doull, C. D. Klaassen, M. O. Amdur, *Toxicology. The Basic Science of Poisons,* 2nd ed., Macmillan Publ. Co., New York 1980, p. 284.
108. S. Moeschlin, *Klinik und Therapie der Vergiftungen,* Thieme Verlag, Stuttgart 1980.
109. K. B. Lehmann, J. Wilke, J. Yamada, J. Wiener, *Arch. Hyg.* **67** (1908) 57.
110. N. I. Sax, *Dangerous Properties of Industrial Materials,* 3rd ed., Reinhold Publ. Co., New York 1968, p. 367.
111. Deutsche Forschungsgemeinschaft (ed.), *Maximale Arbeitsplatzkonzentrationen (MAK) 1982,* Verlag Chemie, Weinheim 1982.
112. American Conference of Governmental Industrial Hygienists (ed.), *Threshold Limit Values (TLV) 1982,* ACGIH, Cincinnati, Ohio 1982.

Acetoacetic Acid → Oxocarboxylic Acids

Acetone

Stylianos Sifniades, Allied Signal Inc., Morristown, New Jersey 07962, United States

Alan B. Levy, Allied Signal Inc., Morristown, New Jersey 07962, United States

1.	**Introduction**	1
2.	**Physical Properties**	1
3.	**Chemical Properties**	2
4.	**Production**	3
4.1.	**Cumene Oxidation (Hock Process)**	3
4.2.	**Dehydrogenation of 2-Propanol**	6
4.3.	**Propene Oxidation**	7
4.4.	**Oxidation of 2-Propanol**	8
4.5.	**Oxidation of *p*-Diisopropyl Benzene**	8
4.6.	**Fermentation of Biomass**	8
5.	**Environmental Protection**	8
6.	**Quality Specifications and Analysis**	9
7.	**Storage and Transportation**	10
8.	**Uses**	10
8.1.	**Methyl Methacrylate**	10
8.2.	**Bisphenol A**	11
8.3.	**Aldol Chemicals**	11
8.4.	**Solvent Uses**	11
9.	**Economic Aspects**	12
10.	**Toxicology and Occupational Health**	13
11.	**Derivatives**	13
11.1.	**Acetone Cyanohydrin**	13
11.2.	**Diacetone Alcohol**	15
11.3.	**Miscellaneous Derivatives**	16
12.	**References**	16

1. Introduction

Acetone, 2-propanone, dimethyl ketone, CH_3COCH_3, [*67-64-1*], is the first and most important member of the homologous series of aliphatic ketones. It is a colorless, mobile liquid widely used as a solvent for various polymers. Its largest application, however, is as an intermediate in the synthesis of methyl methacrylate, bisphenol A, diacetone alcohol, and other products.

Acetone was first manufactured by the dry distillation of calcium acetate [*62-54-4*]. Calcium acetate was originally a product of wood distillation, and later was obtained by fermentation of ethanol. Carbohydrate fermentation directly to acetone and butyl and ethyl alcohols displaced these processes in the 1920s. The carbohydrate route, in turn, was replaced in the 1950s and 1960s by the 2-propanol dehydrogenation process and by the oxidation of cumene to phenol [*108-95-2*] plus acetone. Together with direct propene oxidation, these methods account for over 95 % of the acetone produced worldwide.

2. Physical Properties

Acetone has the following physical properties: M_r 58.081; *bp* at 101.3 kPa, 56.2 °C; *mp* − 94.7 °C; relative density, d_4^0 0.81378, d_4^{15} 0.79705, d_4^{20} 0.7908; relative vapor density (air = 1) 2.0025; refractive index n_D^{20} 1.35868; critical temperature 235.0 °C, critical pressure 4.6 MPa (46 bar), critical density 0.278 g/cm^3; cubic expansion coefficient (18 °C) 1.43×10^{-3} K^{-1}; compressibility coefficient (18 °C) 1.286×10^{-6} kPa^{-1} (1.286×10^{-4} bar^{-1}).

Viscosity in mPa · s: 1.53 (− 80 °C), 0.71 (− 40 °C), 0.40 (0 °C), 0.32 (20 °C), 0.27 (40 °C). Surface tension in mN · m^{-1}: 38.1 (− 91.09 °C), 23.9 (15 °C), 23.3 (20 °C), 23.0 (24.8 °C), 22.0 (30 °C), 21.6 (42 °C).

Thermal properties: Specific heat capacity, c_p (20 °C) 2.135 kJ kg^{-1} K^{-1}; heat of fusion (− 95 °C) 98.47 kJ kg^{-1}; heat of vaporization (30 °C) 545.2 kJ kg^{-1}, (0 °C) 588.2 kJ kg^{-1}; molar entropy 0.2001 kJ mol^{-1} K^{-1}; heat of combustion 1804 kJ mol^{-1}; heat of formation (20 °C) 235.3 kJ/mol; thermal conductivity of the liquid 1.976 W m^{-1} K^{-1}.

Vapor Pressure in kPa: 24 (20 °C), 37.3 (30 °C), 56.0 (40 °C), 82.8 (50 °C), 114.8

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_079

(60 °C), 214.8 (80 °C), 372.8 (100 °C), 929.6 (140 °C).

Electrical properties: Electric conductivity (20 °C) $5.5 \times 10^{-8}\ \Omega^{-1}\ \mathrm{cm}^{-1}$; dipole moment (20 °C) 2.69 Debye; dielectric constant of the liquid 21.58 (0 °C), 22.64 (10 °C), 20.70 (25 °C), 19.38 (40 °C); dielectric constant of the vapor 1.0235 (24.8 °C), 1.0277 (29.8 °C).

At ambient temperature acetone is a clear, colorless liquid with a characteristic odor. It is miscible in all proportions with water and polar organic solvents, such as the lower molecular mass alcohols, carboxylic acids, and ethers. It is miscible in limited proportions with nonpolar solvents, such as hydrocarbons. Some azeotropic mixtures are shown in Tables 1 and 2 [3], [4].

Table 1. Acetone binary azeotropes ∗

Second component	Acetone, wt %	*bp* (101.3 kPa), °C
Carbon tetrachloride	88.5	56.08
2-Butylchloride	80	55.75
Hexane	53.5	49.7
Methyl acetate	49	55.65
Diethylamine	38	51.55
Carbon disulfide	33	39.25
tert-Butylchloride	25	49.2
Isoprene	20	30.5
n-Propylchloride	15	45.8
Methanol	14	55.59

∗ Source and further examples [3].

Table 2. Acetone ternary azeotropes ∗

Components (A is acetone)	Composition, wt %	*bp* (101.3 kPa), °C
B water	0.81	38.04
C carbon disulfide	75.21	
B water	0.4	32.5
C isoprene	92.0	
B chloroform	46.7	57.5
C methanol	23.4	
B chloroform	70.2	55.0
C ethanol	6.8	
B methanol	16	51.1
C cyclohexane	40.5	
B methyl acetate	5.6	49.7
C hexane	43.3	

∗ Source and further examples [4].

Acetone dissolves many synthetic resins, e.g., nitrocellulose, acetylcellulose, poly(acrylate esters), and alkyd resins. It also dissolves most natural resins, fats, and oils.

3. Chemical Properties

Pure acetone is essentially inert to air oxidation and to diffuse sunlight under ambient conditions. Its chemical stability diminishes significantly in the presence of water. Acetone may react violently and sometimes explosively, especially in a confined vessel [5]. For example it is particularly sensitive to oxidizing agents, such as nitrosyl chloride [6] [*2696-92-6*], chromium trioxide [7] [*1333-82-0*], and hydrogen peroxide [8] [*7722-84-1*], or organic peroxides [9]. Mixtures of acetone with chloroform [*67-66-3*] may react violently in the presence of alkali [10]. Reaction even may be initiated by surface alkali on new glassware [11]. Acetone has a flash point of −17 °C (closed cup). Flammability limits in air are: lower 2.13 vol %, upper 13 vol %; autoignition temperature 465 °C. The flammability of acetone can be reduced by mixing it with less flammable and/or less volatile solvents [12]. Fires have been started during recovery of acetone from air by adsorption on activated carbon when air flow was too low to effectively remove the heat generated by surface oxidation [13].

Acetone undergoes typical carbonyl reactions with particular ease. Acid- or base-catalyzed self-condensation produces the dimers diacetone alcohol and mesityl oxide and the cyclic trimer isophorone.

Under strongly basic conditions hydrogen cyanide adds to acetone to form 2-cyano-2-propanol (acetone cyanohydrin), an important intermediate in the manufacture of methyl methacrylate and other methacrylate esters (Section 8.1).

In liquid ammonia solution acetone condenses with acetylene [*74-86-2*] in the presence of catalytic amounts of alkali metals to form 2-methyl-3-butyn-2-ol [*115-19-5*], an intermediate in the synthesis of isoprene [14] [*563-46-2*]. Catalytic hydrogenation of acetone yields 2-propanol [*67-63-0*]. Pyrolysis produces methane [*74-82-8*] and ketene [*463-51-4*], a powerful acetylating agent. A more economical source of ketene, however, is the pyrolysis of acetic acid, which produces ketene and water.

Reductive ammonolysis of acetone yields isopropylamine [*75-31-0*]. Condensation with 2 mol phenol in the presence of an acidic catalyst yields bisphenol A (Section 8.2), an important

monomer used in the manufacture of polycarbonate resins.

Perchlorination yields hexachloroacetone [*116-16-5*], which is cleaved into chloroform [*67-66-3*] and sodium trichloroacetate [*650-51-1*] upon treatment with sodium hydroxide.

4. Production

Approximately 83 % of the acetone produced worldwide is manufactured from cumene as a coproduct with phenol. In the United States and Western Europe dehydrogenation of 2-propanol is also important, whereas in Japan catalytic oxidation of propene is used as a second process. Cumene, 2-propanol, and propene together as starting materials account for over 95 % of the acetone produced worldwide. Because propene is used in the manufacture of both cumene and 2-propanol, propene is the ultimate raw material for the production of acetone.

Small amounts of acetone are made by oxidation of *p*-diisopropyl benzene and of *p*-cymene. Coproducts from these reactions are hydroquinone and *p*-cresol, respectively. Acetone is also produced by propene oxidation and as a byproduct of acetic acid manufacture.

Fermentation of cornstarch and molasses to acetone and 1-butanol was important in the past. It is believed to be practiced today to a limited extent in several countries.

4.1. Cumene Oxidation (Hock Process)

(Fig. 1)

Propene [*115-07-1*] is added to benzene [*71-43-2*] to form cumene [*98-82-8*], which is then oxidized by air to cumene hydroperoxide (**1**), and cleaved in the presence of an acid catalyst. Phenol [*108-95-2*] and acetone produced in the process are recovered by distillation.

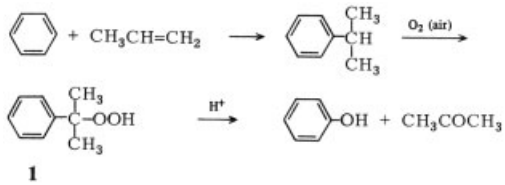

The alkylation of benzene by propene proceeds under typical Friedel – Crafts conditions. In 1996, a number of processes using zeolite catalysts came on-stream. The cumene produced is purified by chemical means and refined by distillation to 99.9 % minimum purity. Oxidation-grade cumene must meet strict quality standards.

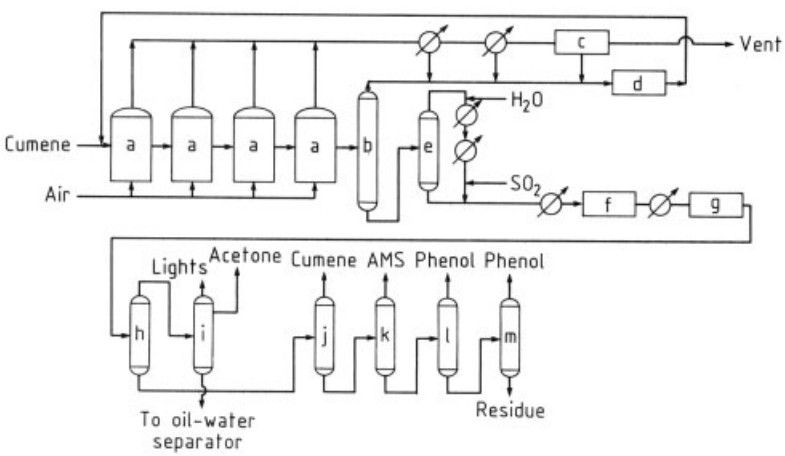

Figure 1. Cumene phenol – acetone process (Allied)
a) Oxidizers; b) Flash column; c) Carbon adsorber; d) Alkaline extraction and wash; e) Cumene hydroperoxide decomposer; f) Dicumyl peroxide decomposer; g) Ion exchange; h) Crude acetone column; i) Acetone-refining column; j) Cumene column; k) α-Methylstyrene column; l) Phenol column; m) Phenol residue topping column
AMS = α-methylstyrene

Table 3. Specifications for oxidation-grade cumene (zeolite process) [15]

Property	ASTM test	Specification
Appearance		Clear, colorless liquid
Color, Pt – Co scale	D1209-79	15 max.
$d_{15.5}^{15.5}$	D891-59	0.864 – 0.867
Acid wash color, W scale	D848-62	2 max.
Sulfur compounds	D853-47	Free from H_2S and SO_2
Copper corrosion	D849-47	No iridescence, gray or black
Distillation range	D950-56	1.0 °C max.
Cumene content		99.93 % min.
Phenolics content		5 ppm max.
Cumene hydroperoxide content		200 ppm max.
Sulfur content		0.1 ppm max.

The newer zeolite-based processes have led to slightly tighter specifications (Table 3) [15].

Cumene Oxidation. The oxidation of cumene is a free-radical chain reaction [16]. The chain initiator is cumene hydroperoxide, the main product of the reaction. The rate of oxygen consumption can be approximated by the following expression:

$$\frac{-\mathrm{d}c_{\mathrm{O_2}}}{\mathrm{d}t} = k_{\mathrm{p}} \cdot c_{\mathrm{RH}} \sqrt{\frac{2k_{\mathrm{i}} c_{\mathrm{ROOH}}}{k_{\mathrm{t}}}}$$

c_{RH} and c_{ROOH} are the concentrations of cumene and cumene hydroperoxide, respectively; k_{i}, k_{p}, and k_{t} are the rate constants for chain initiation, propagation, and termination.

The expression shows that the rate of oxidation is zero in the absence of cumene hydroperoxide. This is not exactly true, because the expression is only an approximation; but the oxidation of cumene does require long induction periods when starting with pure cumene. Consequently, the industrial oxidation always is carried out in a series of continuous reactors; the concentration of cumene hydroperoxide is at least 8 wt % in the first reactor. Because the sum of c_{RH} and c_{ROOH} remains roughly constant during the reaction, the rate of reaction cannot increase indefinitely as c_{ROOH} increases. The maximum rate is achieved at approximately 35 wt % cumene hydroperoxide.

Besides cumene hydroperoxide, both dimethylphenylmethanol and acetophenone are also formed as byproducts during this oxidation. These arise from a secondary chain reaction that proceeds in parallel with the main chain. Byproduct formation is accelerated as the concentration of cumene hydroperoxide increases. For these reasons, most plants operate between 25 and 40 wt % in the last oxidation reactor.

A minor but significant byproduct of the oxidation is dicumyl peroxide [*80-43-3*]. This arises during the termination of the chain reaction. Dicumyl peroxide also contributes to chain initiation [17], but to a much lesser degree than cumene hydroperoxide. Other minor byproducts are formaldehyde and formic acid, which are produced along with acetophenone by methyl group degradation.

The oxygen needed for cumene oxidation is supplied by air. Use of pure oxygen has been suggested [18] but is disfavored by both economic and safety considerations [19]. At low initiation rates, the rate of the reaction is essentially independent of the oxygen concentration at a partial pressure of oxygen over 33 kPa (0.33 bar) [20]. A detailed study of the rate of oxygen uptake in a bubble column as a function of temperature and partial oxygen pressure has been made [21]. The study served as a basis for a mathematical model of the oxidation [22], [23].

Two cumene oxidation processes are used in the United States, which with minor variations are practiced also in the rest of the world [19], [22]. One process was developed by Hercules and licensed by Kellogg (previously BP/Hercules) and GE/Lummus [24]. The other process was developed by Allied and licensed by Allied/UOP [25–28].

In both processes several reactors are employed in series. Fresh and recycled cumene are fed to the first reactor, which may operate at 8 – 12 wt % cumene hydroperoxide. The concentration increases by 4 – 8 wt % in each successive reactor; the last reactor may operate at 25 – 40 wt % cumene hydroperoxide. Fresh air

is pumped in parallel to each reactor and vented at the top after removal of organic vapors.

In the *Hercules process*, the oxidation of cumene is carried out at approximately 620 kPa (6 bar)/90 – 120 °C, in the presence of a sodium carbonate buffer [29]. Under these conditions the residence time in the oxidizer train is 4 – 8 h and the hydroperoxide molar selectivity 90 – 94 %. The spent air is first passed through water cooled and refrigerated condensers in series to remove organic vapors, and is finally vented. The condensate is returned to the oxidizers after treatment [29].

In the *Allied/UOP process* (Fig. 1) the oxidation is carried out at atmospheric pressure. No buffer or promoter is added, but great care is taken to wash all streams recycled to the oxidizer with alkali and water [27]. Temperature is maintained at 80 – 100 °C. Residence time in the oxidizer train is 10 – 20 h and hydroperoxide molar selectivity is 92 – 96 %. Spent air is vented after organic vapors are removed by condensation followed by activated carbon adsorption. The recovered materials are washed with aqueous sodium hydroxide and water, then returned to the oxidizers.

The oxidation of cumene generates approximately 116 kJ of heat per mole of cumene oxidized [30]. Part of this heat is carried to the condensers by organic vapors (this part is larger in the Allied/UOP process because of the lower operating pressure). The rest is removed by heat exchangers.

In both processes cumene hydroperoxide is concentrated to over 80 wt % by evaporation of excess cumene. In the Hercules process the oxidate is washed with water prior to distillation in order to remove the buffer added during oxidation.

Cumene Hydroperoxide Cleavage. Cumene hydroperoxide [*80-15-9*] is cleaved to phenol and acetone in the presence of catalytic amounts of a strong acid. The acid most commonly used is sulfuric acid. Sulfur dioxide is used as catalyst in the Allied/UOP process. Several patents claim the use of solid acids as catalysts for the decomposition [31]. Strongly acidic resins have been used to that effect in the Soviet Union [32]. However, all commercial units use strong mineral acids or SO_2, which generates sulfuric acid in situ, as catalysts.

The cleavage proceeds through an ionic mechanism and releases approximately 252 kJ/mol of cumene hydroperoxide decomposed [30]. The reaction rate accelerates rapidly with increasing temperature. Consequently, decomposition of cumene hydroperoxide commonly is carried out in a continuously stirred reactor in which the steady-state concentration of cumene hydroperoxide is maintained at a low level. The heat released by the reaction can be used to estimate the concentration of hydroperoxide present in the reactor at any time [33].

The molar selectivity of the cleavage to phenol and acetone is higher than 99.5 % at temperatures below 70 °C, but it decreases at higher temperatures as increasing amounts of dimethylphenylmethanol and acetophenone (in addition to those present in the cumene oxidate) are formed (Table 4) [28].

Table 4. Formation of byproducts during cumene hydroperoxide decomposition ∗

Temp. °C	Molar ratio (byproducts/phenol) × 100	
	DMPM equivalents∗∗	Acetophenone
70	0.36	0.06
90	0.61	0.06
110	1.24	0.15
122	2.19	0.25
146	5.04	0.69

∗ Pure cumene hydroperoxide added to phenol – acetone – cumene solution containing initially 0.5 wt % water and 100 ppm sulfuric acid; data from [28].
∗∗ Sum of dimethylphenylmethanol, α-methylstyrene, and their condensation products.

Acetone produced during the cleavage of cumene hydroperoxide can react further. Oxidation by cumene hydroperoxide forms hydroxyacetone [34] to the extent of 0.2 – 0.5 % of acetone present. Self-condensation catalyzed by acid results in diacetone alcohol and mesityl oxide. Conversion of acetone to these condensates is normally below 0.1 % but may increase upon protracted exposure to strong acid. For example, when the cumene hydroperoxide cleavage was carried out with refluxing acetone using a sulfonic acid resin as catalyst, approximately 1.7 % of acetone was transformed to diacetone alcohol and mesityl oxide [32].

Under the conditions of the cumene hydroperoxide cleavage, dimethylphenylmethanol is dehydrated to α-methylstyrene (**2**) and also forms undesirable condensates.

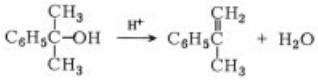

2

Compound (**2**) may be either hydrogenated [35] to cumene and recycled, or recovered and sold.

In the *Hercules process*, the cumene hydroperoxide decomposition is carried out in a constant-flow, stirred tank reactor in the presence of sulfuric acid or another strong mineral acid [19], [28]. The acid is added to the reactor as an acetone solution. The reactor temperature is maintained below 95 °C by refluxing approximately 2.8 kg acetone per kilogram cumene hydroperoxide. The ratio of the quantity of reflux to the quantitiy of hydroperoxide fed is used as a monitor of the cleavage reaction [19].

In the *Allied/UOP process* the cumene hydroperoxide cleavage is carried out at 60 – 80°C in a pressurized, constant-flow, back-mixed reactor. Temperature is controlled by means of heat exchangers in the loop. The catalyst is either sulfuric acid or sulfur dioxide. Up to 5 wt % cumene hydroperoxide remains unreacted. Under these conditions, dimethylphenylmethanol combines with cumene hydroperoxide to form dicumyl peroxide (**3**) [*80-43-3*], which upon subsequent heating to 110 – 140 °C in a short-residence-time plug-flow reactor is cleaved into phenol, acetone, and α-methylstyrene.

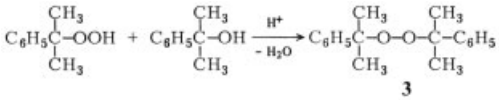

3

This sequence suppresses the formation of condensates by approximately 50 %. Variants of this two-stage process have been patented [36].

Product Separation. In the *Hercules process* the cleavage mixture is neutralized with base and then fed to a separation column. The overheads from this column contain acetone, α-methylstyrene, and cumene; acetone is recovered by distillation, and α-methylstyrene is hydrogenated without prior separation from cumene. This cumene stream is then recycled to the oxidizers. Phenol from the bottoms of the separation column is recovered by distillation.

In the *Allied/UOP process* the cleavage mixture is treated with an ion-exchange resin to remove the acid catalyst and then is distilled. Acetone is removed first in a crude acetone column and purified by distillation with steam in an acetone-refining column. Cumene, α-methylstyrene, and phenol are recovered by sequential distillation of the bottoms from the crude acetone column. Cumene is recycled to the oxidizers after it has been washed with alkali, and α-methylstyrene is marketed.

4.2. Dehydrogenation of 2-Propanol

The hydration of propene [*115-07-1*] gives 2-propanol [*67-63-0*], which is then dehydrogenated to acetone. In the United States a C_3 stream containing 40 – 60 % propene is used for the manufacture of 2-propanol (**4**).

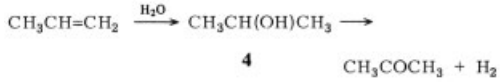

4

The dehydrogenation of (**4**) is endothermic by 66.6 kJ/mol at 327 °C. The equilibrium constant, K_p (bar), obeys the following equation [37]:

$$\log K_p = -2764/T + 1.516 \log T + 1.765$$

The main side reaction is the dehydration of 2-propanol to propene. Other competing reactions are the self-condensation of acetone to diacetone alcohol, which leads to further condensation products.

A large number of catalysts for 2-propanol dehydrogenation have been studied, including copper, zinc, and lead metals, as well as metal oxides, e.g., zinc oxide, copper oxide, chromium-activated copper oxide, manganese oxide, and magnesium oxide. Inert supports, such as pumice, may be used.

Highly active catalysts are the precious metals platinum and ruthenium [39] or 0.25 % platinum on sodium-activated alumina [40]. These catalysts are particularly effective for the dehydrogenation of aqueous 2-propanol, which is obtained by hydration of propene.

All catalysts gradually lose activity because of a buildup of carbon deposits, so the operating temperature is increased as the catalyst ages.

The catalyst is regenerated periodically by burning out the deposits. A good catalyst lasts for several months.

In a typical process, the azeotropic mixture of water and 2-propanol (87.8 wt % 2-propanol) is evaporated (sometimes using steam as carrier) and fed to a catalyst bed in a reactor specially designed for effective heat transfer. Hydrogen, produced downstream, may be mixed with the feed to prevent catalyst fouling. The reactor consists of a multitude of 2.5-mm steel tubes heated by oil, high-pressure steam, hot gases, or molten salts. The reaction produces hydrogen (> 99 % purity) as a valuable byproduct. This is separated by condensing all other components. Acetone is separated by distillation. The process is illustrated in Figure 2. Typical operating conditions are shown in Table 5.

4.3. Propene Oxidation [43], [44]

Direct oxidation of propene (Wacker – Hoechst process) currently is practiced only in Japan. A mixture of acetone (92 % selectivity) and propionaldehyde (2 – 4 % selectivity) is produced.

$$CH_3CH{=}CH_2 + 1/2\ O_2 \xrightarrow{\text{catalyst}} CH_3COCH_3 + CH_3CH_2CHO$$

The process is analogous to the oxidation of ethylene to acetaldehyde by the Wacker process. The catalyst solution typically contains 0.045 M palladium(II) chloride, 1.8 M copper(II) chloride, and acetic acid [45]. The reaction usually is carried out in two alternating stages. In the first stage, air is used to oxidize the metal ions to the +2 oxidation state. In the second, air is removed and propene added. Palladium(II) oxidizes propene, and the resulting palladium(I) is reoxidized by the pool of copper(II). Reaction conditions are 1 – 1.4 MPa (10 – 14 bar) and 110 – 120 °C. Propene conversion is higher than 99 %.

Besides propionaldehyde, chlorinated carbonyl compounds and carbon dioxide also are

Table 5. Gas-phase dehydrogenation of 2-propanol

Company	Catalyst	Temperature, °C	Pressure, kPa	Conversion, %	Selectivity, %	Yield, %	Reference
Standard Oil	ZnO/ZnO_2	400	201 – 304	98.2	90.2	88.6	[38]
Knapsack-Griesheim	$CuO/Cr_2O_3/Na_2O$ pumice	300		89.5	99.0	88.6	[39]
Toyo-Rayon	$CuO/NaF/SiO_2$	300		93.4	100	93.4	[40]
Engelhard Industries	5 % Pt/C	310				92.4	[41]
Usines de Melle	$CuO/Cr_2O_3/SiO_2$	220	151	75	98.2	73.7	[42]

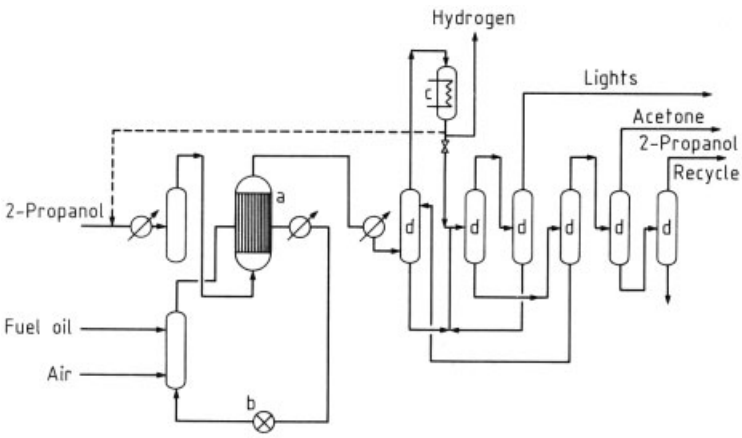

Figure 2. Acetone production via 2-propanol dehydrogenation
a) Reactor; b) Heating loop; c) Refrigeration; d) Distillation columns

formed. Acetone and the byproducts are removed from the catalyst solution by flash evaporation with steam and separated by fractional distillation. Propionaldehyde (*bp* 49 °C) distills in one column and acetone (*bp* 56 °C) distills in the other.

4.4. Oxidation of 2-Propanol [46]

In the absence of catalysts 2-propanol reacts with oxygen via a free-radical reaction to form acetone and hydrogen peroxide.

$$CH_3CH(OH)CH_3 \xrightarrow{O_2} CH_3COCH_3 + H_2O_2$$

Until the mid-1980s the Shell process used hydrogen peroxide for the manufacture of glycerol from propene. The theoretical yield of acetone based on glycerol produced is 1.26 kg/kg. Acetone yields of about 90 % of theoretical were obtained.

4.5. Oxidation of *p*-Diisopropyl Benzene

Acetone is coproduced with hydroquinone [*123-31-9*] from *p*-diisopropylbenzene [*100-18-5*] in a process analogous to the phenol – acetone production from cumene.

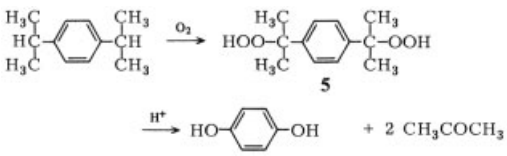

In the Goodyear process [47] *p*-diisopropylbenzene is oxidized by oxygen in the presence of caustic. The *p*-diisopropylbenzene dihydroperoxide (**5**) [*3159-98-6*] formed is crystallized and washed with benzene. It is then dissolved in acetone and cleaved to hydroquinone and acetone in the presence of sulfuric acid. Next the acid is neutralized with ammonia and the ammonium sulfate formed is filtered. Acetone is recovered by distillation from the reaction mixture. Some of this acetone is recycled to the cleavage section while the rest passes through a finishing column for purification to at least 99.5 %. Eastman Chemical and Goodyear Tire & Rubber Company use this process in the United States. Annual US capacity is estimated to be 18 – 20 t/a. Sumitomo Chemical Company and Mitsui Petrochemical Industries of Japan use a similar process to produce *p*-cresol from cymene. Their annual capacity of acetone byproduct is 48 000 t.

4.6. Fermentation of Biomass

The fermentation of cornmeal or molasses by various members of the *Clostridium* genus yields a mixture of 1-butanol, acetone, and ethanol in 2 % overall concentration. The products are recovered by steam distillation and then fractionated.

The process was started during World War II to provide acetone needed for the manufacture of cordite. The last operating plant in the United States (Publicker Industries) closed in 1977.

The mixture of butanol, acetone and ethanol produced has been considered for use as a gasoline substitute in France [48]. Research aimed at increasing the concentration of useful products obtained in the process was carried out in the United States in the early 1980s [49]. The future of the fermentation process is tied to the availability of petrochemical feedstocks. High oil prices during the oil crises of the mid to late 1970s led to renewed interest in the process. Given the low oil prices of the 1990s and the ready availability of feedstocks at reasonable prices, it does not appear that these processes can compete under current conditions.

5. Environmental Protection

Because approximately 70 % of acetone is produced from cumene, a close examination of this process is warranted. Potential pollution sources in a phenol – acetone plant are emissions to the atmosphere and liquid discharge. Atmospheric emissions from the phenol – acetone process in the late 1970s have been estimated [29]. However over the past 20 years, and particularly in the 1990s with the renewal of the Clean Air Act, these emissions have been reduced significantly.

Aqueous streams containing significant amounts of organic substances arise from the

various wash operations and sumps at the plant. Insoluble material is recovered by decantation. Phenol and acetone (0.5 – 3 wt % each) are the most abundant organic compounds remaining in the water after decantation. There are also minor quantities (0.001 – 0.1 wt %) of cumene, α-methylstyrene, dimethylphenylmethanol, acetophenone, formaldehyde, formic acid, and various condensates. Of these compounds, phenol, formaldehyde, and formic acid are listed as hazardous substances in the U.S. Federal Water Pollution Control Act [50], but only phenol is present in sufficient quantities to require removal. Phenol is removed from the aqueous solution by solvent extraction, steam stripping, or adsorption on carbon or resins [51] and subsequently is recovered. The recovered phenol is valuable enough to pay for the capital and operating expenses of phenol abatement. Residual phenol in the water (10 – 500 ppm) is destroyed by biological degradation.

The federal regulatory status of acetone has changed. Acetone was granted VOC-exempt status by EPA on June 16, 1995 [53]. As of August, 1997, forty-four states had promulgated similar state rules. In states that have not yet promulgated state exemptions, acetone may technically still be regulated as a VOC. Acetone is not listed as a hazardous air pollutant (HAP) under section 112(b) of the Clean Air Act (CAA), or as an extremely hazardous substance under EPCRA Section 302. Acetone is also not listed as a priority pollutant under the Clean Water Act. It has been approved under the CAA as a substitute for ozone-depleting substances. Acetone was removed from the Federal Emergency Planning and Community Right-to-Know Act (EPCRA Section 313) list in June of 1995. Acetone is not regulated as a known or suspected carcinogen, and the National Toxicology Program (NTP) has recommended against testing for carcinogenicity because of its low toxicity and absence of any evidence supporting the carcinogenic potential of acetone.

Acetone is listed as a "U" waste under the Resource Conservation and Recovery Act (RCRA) based on its ignitability. "U" wastes are commercial chemicals that must be treated as hazardous wastes when discarded. Because of its RCRA listing it is included in the list of hazardous substances in the Superfund statute (Comprehensive Environmental Response, Compensation, and Liability Act).

6. Quality Specifications and Analysis

Acetone is produced industrially in relatively high purity, the main impurity being water. Table 6 summarizes the quality requirements for commercial 99.5 % acetone. Methods for preparing very high purity acetone from the commercial material are given in reference [54].

Gas chromatography is the most widely used method for the quantitative analysis of acetone. For example, good separation of acetone from other low-boiling organic compounds can be obtained on a 30 m × 0.32 mm Carbowax capillary column. Extensive data on packed column separations are compiled in [55]. Infrared (carbonyl absorption, 1711 cm^{-1}) and ^1H NMR (singlet at ca. 1.05 ppm) spectroscopy may be used for both qualitative and quantitative analysis.

Table 6. Standard specifications for acetone, ASTM D329-90

Property	ASTM test	Specification
Relative density	D268	
20/20 °C		0.7910 – 0.7930
25/25 °C		0.7865 – 0.7885
Color	D1209	≤ 5 on platinum-cobalt scale
Distillation range	D1078	1.0 °C, including 56.1 °C
Nonvolatile matter	D1353	≤ 5 mg/100 mL
Odor	D1296	characteristic, nonresidual
Water	D1364	≤0.5 wt % ∗
Acidity (as acetic acid)	D1613	≤0.002 wt %
Water miscibility	D1722	passes test
Alkalinity (as ammonia)	D1614	≤0.001 wt %
Permanganate time	D1363	≥30 min at 25 °C

∗ This water limit ensures that the material is miscible without turbidity with 19 volumes of 99 % heptane at 20 °C (ASTM D1476).

7. Storage and Transportation

Acetone has a low flash point; therefore, all shipping and storage containers must carry a red, diamond shaped "flammable liquid" label. Strict precautions should be taken to guard against fire hazards whenever acetone is handled. All wiring should be installed as described in Article 500 of the U.S. National Electrical Code or corresponding regulations in other countries. Explosion-proof motors, switches, etc., should be used. Accumulation of static electricity should be prevented by grounding and humidity control. Use of spark-resistant tools is recommended. Small fires may be controlled by use of carbon dioxide or dry chemical extinguishers. "Alcohol"-type foam should be used on larger fires; water spray will reduce the intensity of the flame.

Contact of acetone with oxidants should be avoided because it may lead to explosion [5]. Contamination with chlorinating agents may lead to the formation of toxic chloroketones. Prolonged exposure to direct sunlight may result in the formation of carbon monoxide. Packaging requirements for acetone are described in paragraph 49 CFR 173.242 (bulk), Bulk Packaging with Packaging for Certain Medium Hazard Liquids and Solids, Including Solids with Dual Hazards [56]. Transportation of acetone is covered in paragraph 49 CFR 172.101, Table of Hazardous Materials, of the Department of Transport Regulations [57]. The international transportation codes are IMDG Code D 3102; UN no. 1090; CFR 49, 172.101; RID (ADR, ADNR): Class 3, IATA: flammable liquid. The quantity of acetone in one package may not exceed 5 L in plastic, metal or aluminum, 1 L in glass, or 0.5 L in a glass ampoule in a passenger aircraft. The quantity of acetone in one package may not exceed 60 L in a cargo plane.

8. Uses

The main uses of acetone are as a chemical intermediate and as a solvent. The estimated 1995 acetone consumption by area of application in the USA is shown in Table 7 [59].

Table 7. Estimated 1995 US acetone consumption by area of application [59]

Use	Acetone used, 10^3 t
Acetone cyanohydrin/methacrylates	500
Bisphenol A	203
Aldol chemicals	140 (total)
Methyl isobutyl carbinol	35
Methyl isobutyl ketone	76
Others	22
Solvent use	191
Other uses	90
Total	1124

8.1. Methyl Methacrylate

Acetone is condensed with hydrogen cyanide to form acetone cyanohydrin (**6**) (see Section 11.1), which is next hydrolyzed with sulfuric acid to methacrylamide sulfate (**7**).

Further reaction with methanol yields methyl methacrylate (**8**) [*80-62-6*]. Approximately 0.70 kg of acetone is required per kilogram of methyl methacrylate produced.

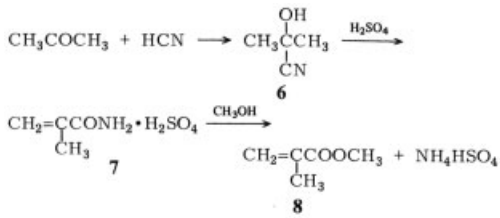

Higher methacrylate esters may be produced either by transesterification of methyl methacrylate or by esterification of methacrylic acid (**9**) [*79-39-0*]; the latter is made by hydrolysis of methacrylamide sulfate:

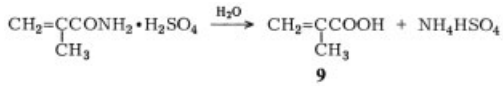

At the end of 1995 there were 22 plants manufacturing MMA in the United States, Western Europe, and Japan. Five basic process routes have been commercialized: The acetone cyanohydrin route; two-stage oxidation of isobutylene to methacrylic acid followed by esterification; two-stage oxidation of *tert*-butyl alcohol to methacrylic acid followed by esterification; hydroformylation of ethylene to propionaldehyde, condensation with formaldehyde

to methacrolein, oxidation, and esterification (BASF); and ammoxidation of *tert*-butyl alcohol to methacrylonitrile, which is hydrolyzed to methacrylamide sulfate and then esterified to MMA (Asahi). One new route has been announced by Mitsubishi Gas Chemicals, which is a recycle version of the acetone cyanohydrin route. A 41×10^3 t/a plant to make MMA and MAA started up in 1997. Worldwide production of MMA in 1996 was about 1682×10^3 t [60].

8.2. Bisphenol A (→ Phenol Derivatives)

Bisphenol A (**10**), 4,4′-isopropylidenediphenol [*80-05-7*] is manufactured by condensation of 2 mol phenol with 1 mol acetone in the presence of an acid catalyst:

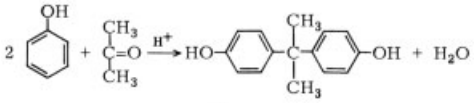

10

Approximately 0.28 kg of acetone is required per kilogram of bisphenol A. In the 1990s, bisphenol A has had the fastest growing demand of the phenol derivatives. Four US. companies produce bisphenol A: Shell (Deer Park, Texas), General Electric (Mount Vernon, Indiana), Dow (Freeport, Texas), and Aristech (Haverill, Ohio). Estimated worldwide usage in 1995 was 1600×10^3 t [61].

8.3. Aldol Chemicals (see Section 11.2, also → Ketones)

These chemicals are produced by condensation of acetone. Two moles of acetone form 1 mol of diacetone alcohol, 4-hydroxy-4-methyl-2-pentanone (**11**) [*123-42-2*]. Subsequent dehydration yields mesityl oxide, 4-methyl-3-penten-2-one (**12**) [*141-79-7*]. Hydrogenation of (**11**) yields 2-methyl-2,4-pentanediol (**13**) [*107-41-5*]. By hydrogenation of (**12**) methyl isobutyl ketone (**14**) [*108-10-1*] is available; further hydrogenation produces 4-methyl-2-pentanol (**15**) [*108-11-2*]. Three moles of acetone are condensed to 1 mol of isophorone, 3,5,5-trimethyl-2-cyclohexen-1-one (**16**) [*78-59-1*].

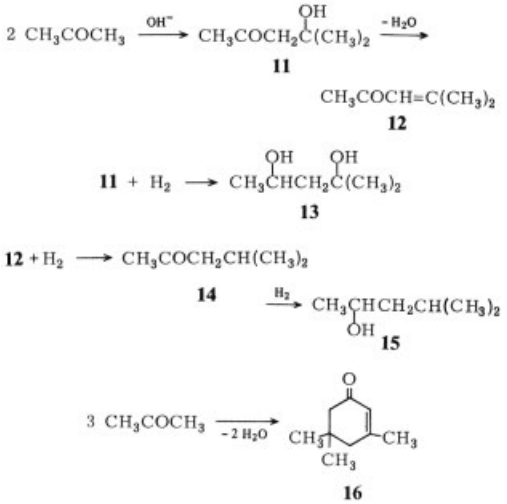

Approximately 1.25 kg of acetone is used per kilogram of methylisobutyl ketone produced. The US 1996 production of ca. 100×10^3 t of MIBK consumed ca 75×10^3t of acetone. The US manufacturers are: Eastman (Kingsport, Tennessee), Shell (Deer Park, Texas), and Union Carbide (Institute, West Virginia) [62]. Methyl isobutyl ketone is used as a solvent for nitrocellulose lacquers, vinyl polymers, and acrylic resins [62].

Diacetone alcohol, mesityl oxide, and isophorone are used mainly as solvents. Their use has diminished in the USA because of their status as photochemically reactive solvents under Rule 66 of Los Angeles County [52]. The primary use of 4-methyl-2-pentanol is for ore flotation, and 2-methyl-2,4-pentanediol is used in hydraulic fluids and printing inks.

8.4. Solvent Uses

Acetone is used as a solvent for paints, varnishes, and lacquers. It is also used as a wash solvent for these materials and as a spinning solvent in the manufacture of cellulose acetate. A small amount of acetone is used as a solvent for acetylene. Approximately 191×10^3 t of acetone was consumed in direct solvent applications. The major solvent market for acetone is in paints and coatings. Consumption of acetone in these applications increased by 9×10^3 t in 1995 because of its delisting as a VOC.

The pharmaceutical industry is also a large consumer of acetone for the manufacture of pharmaceuticals, vitamins, and cosmetics. In 1995 acetone consumption in pharmaceutical and cosmetic applications was $(36-43) \times 10^3$ t.

The removal of acetone from the VOC list has made it more attractive as a solvent, particularly for replacing other chemicals on the VOC list. Following the EPA's August 1995 action, eight states automatically delisted acetone as a volatile organic compound. As of August, 1997, forty-four states had promulgated similar state rules.

9. Economic Aspects

The United States acetone capacity by manufacturer and production process can be found in [63]. World capacity data are given in Table 8 [59]. The United States acetone production for the last two decades is summarized in Table 9 [59]. Worldwide production in 1994 was: USA 1281×10^3 t/a; Western Europe 1200×10^3 t/a; Asia 746×10^3 t/a.

Table 8. World acetone capacity, 1995 [59], [64]

Location	Capacity, 10^3 t
United States	1281
Mexico	22.3
Western Europe	1200
Japan	475
Other Asia	271.4
Others	592
World total	3842

Table 9. United States acetone production, 10^3 t [59]

Year	From 2-propanol	From cumene	Other	Total
1970	379	329	26	734
1975	312	433		745
1980	250	693		943
1985	41	768	3	812
1990	70	972	17	1059
1994	73	1110	20	1203

In 1995 the phenol process accounted for 83 % of all acetone made and 9 % was derived from 2-propanol; only 8 % was produced by all other processes (66×10^3 t/a from propene oxidation in Japan).

The economics of acetone are unusual. The bulk of acetone is made as a coproduct with phenol. Consequently, phenol demand determines to a large extent the availability of acetone. Fortunately acetone serves to some extent the same markets as phenol does. These are mainly the automotive and housing markets. As a result, when economic conditions place a demand on phenol, acetone demand also increases. However, an unusually steep demand for phenol may render acetone an overabundant byproduct [63].

Because the single largest use of acetone is as an intermediate in the manufacture of methacrylates, alternate routes to methacrylates, such as the oxidation of C_4 hydrocarbons, are a potential threat to acetone. When C_4 hydrocarbon stocks are in demand for the manufacture of gasoline additives, they are not likely to be used for acrylate manufacture. However, production of methacrylate by oxidation of C_4 hydrocarbons started in Japan in 1982 [65].

Dehydrogenation of 2-propanol accounted for approximately 80 % of US acetone produced in 1960. As the cumene-based process expanded, the 2-propanol contribution shrank to 52 % in 1970, 27 % in 1980, and 6 % in 1994. If acetone supply does outstrip demand, it is likely that production based on 2-propanol will be further curtailed.

Phenol manufacturing processes that do not coproduce acetone have been developed partly because of fears that there will be a supply/demand imbalance between phenol and acetone. DSM and its licensees produce phenol by the air oxidation of toluene via benzaldehyde and benzoic acid. Subsequently, benzoic acid is decomposed to phenol with a copper catalyst. This route provides phenol, benzaldehyde, and benzoic acid. Mitsui Petrochemical has developed a recycle scheme for converting the acetone byproduct from the cumene hydroperoxide rearrangement back to propylene for feed to the front end of the cumene process. Solutia (formerly Monsanto) and the Boreskov Institute of Catalysis have developed a catalyst capable of oxidizing benzene in high yield with nitrous oxide to give directly phenol. The key to this process is the inexpensive nitrous oxide available as a byproduct from the manufacture of adipic acid. Asahi Chemical has patented a process in which benzene is partly hydrogenated to cyclohexene. The cyclohexene is hydrolyzed to cyclohexanol or oxidized to cyclohexanone; dehydrogenation then gives phenol [66].

10. Toxicology and Occupational Health

Acetone is one of the least toxic industrial solvents [67]. However, exposure to vapor at high concentration should be avoided because it can produce temporary narcosis and cause slight eye irritation. Repeated skin contact with the liquid defats the skin and may cause dermatitis. The liquid is also irritating to the eyes and may cause moderate corneal injury.

The ACGIH [68] has adopted a time-weighted average threshold limit value (TLV-TWA) of 750 ppm, 1.78 g/m^3, and a short-term exposure limit (TLV-STEL) of 1000 ppm, 2.375 g/m^3, for acetone. OSHA regulations [69] set a limit of 2.4 g/m^3. However, the ACGIH has issued a notice of intended change to lower the TLV to 500 ppm (1.19 g/m^3), and the STEL to 750 ppm (1.78 g/m^3) [70]. Exposure limits (TLV-TWA) adopted by the main industrial countries are shown in Table 10 [71]. The odor threshold of acetone is 48 mg/m^3 provided desensitization has not occurred. Animal studies have shown acetone to be relatively nontoxic [67], [71].

Table 10. Time-weighted average threshold limits for acetone [71]

Country	Limit, mg/m^3
Australia	1780
Belgium	1780
Germany	2400
Italy	1000
Japan	470
Netherlands	1780
Former Soviet Union	475
United States	1780

LD_{50} (oral, mouse) 4 – 8 g/kg [67]
LD_{50} (oral, rabbit) 5.3 g/kg [67]
LD_{50} (intraperitoneal, mouse) 1.3 g/kg [67]
LD_{50}(dermal, rabbit) 20 g/kg [71]
Nonteratogenic at 39 or 78 mg per chicken egg [67]
Nonmutagenic in the *Salmonella*/microsome (Ames) test [67]
Nononcogenic on skin of mice, three times a week for 1 year [67]
Moderate corneal injury on rabbit eye [67]
Environmental toxicity: LC_{50} (rainbow trout, 96 h) 5540 mg/L; LC_{50} (bluegill sunfish, 96 h) 8300 mg/L.

Minimum lethal concentration in air (LC_{50}) was 50.1 g/m^3 (21 100 ppm) for rats exposed to acetone vapor for 8 h and 44 g/m^3 (18 500 ppm) for mice exposed for 4 h Human exposure to acetone has been studied [72–77]. Eye and nasal irritation were observed at 1.2 g/m^3. Other effects are similar to those of ethanol, but the anesthetic potency is greater. Prolonged or repeated skin contact may defat the skin and can produce dermatitis. Direct contact of acetone with the eyes can produce corneal injury. Acetone is a solvent of comparatively low acute and chronic toxicity. However it does not have sufficient warning properties to prevent repeated exposures to vapors, which may have adverse effects. There have been no reports that prolonged inhalation of low vapor concentrations result in any serious chronic effects in humans.

Cases of acetone poisoning are rare [67]. In one case, a solvent mixture containing 90 % acetone and 9 % pentane was used to set a cast for a broken leg on a 10-year-old boy [78]. The boy became ill and collapsed 12 h later. After the cast was removed the boy became comatose but recovered completely in 4 days. In another case, a 42-year-old man ingested 200 mL of acetone and became comatose for 12 h [79]. Subsequently, hyperglycemia was diagnosed and attributed to acetone ingestion.

Acetone does not cause neurotoxicity, an occupational disorder caused by exposure to some higher aliphatic ketones and related compounds [80]. Acetone vapor is absorbed with 75 % efficiency by the lungs [76]. The half-life for the elimination of acetone by expired air is approximately 5 h. The metabolism of acetone may proceed through 1,2-propanediol [67], [81], [82].

11. Derivatives

11.1. Acetone Cyanohydrin

Acetone cyanohydrin, 2-hydroxy-2-methylpropanenitrile, $CH_3C(OH)(CN)CH_3$ [*75-86-5*] is an important chemical intermediate for the manufacture of methacrylates (→ Methacrylic Acid and Derivatives). Small amounts of acetone cyanohydrin are used in insecticide manufacture.

Physical Properties. Acetone cyanohydrin is a colorless liquid. The pure compound is practically odorless but usually has an odor of bitter almonds because of traces of hydrogen cyanide. It is very soluble in water and polar solvents and sparingly soluble in hydrocarbons.

M_r 85.11, *mp* −19 °C, relative density d_4^{25} 0.9267, relative vapor density (air = 1) 2.96, refractive index n_D^{25} 1.3980, flash point 73 °C.

Vapor pressure

p, kPa	5.3	3.1	1.3	1.2
t, °C	95	82	74	72

Chemical Properties. Acetone cyanohydrin exhibits the combined characteristics of a nitrile and an alcohol. Under neutral and particularly under alkaline conditions it decomposes to acetone and hydrogen cyanide. The decomposition is inhibited by the addition of small amounts of sulfuric or phosphoric acid; consequently, technical-grade material is stabilized by addition of 0.01 wt % of either acid. Reaction with concentrated sulfuric acid converts acetone cyanohydrin to methacrylamide sulfate; subsequent neutralization with ammonia yields methacrylamide [*79-39-0*]; alcoholysis yields methacrylate esters; alternatively, hydrolysis gives methacrylic acid [*79-41-4*].The yield of methacrylic acid is improved if 3 – 10 % oleum is used instead of 100 % sulfuric acid in the reaction with acetone cyanohydrin [83].

Production. Acetone cyanohydrin is manufactured by the base-catalyzed condensation of acetone with hydrogen cyanide according to the following mechanism [84]:

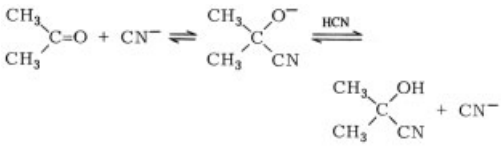

The reaction is reversible but formation of the cyanohydrin is quite favorable; the equilibrium constant is 28 L/mol at 20 – 25 °C [85]. The reaction usually is carried out in the liquid phase. Representative catalysts used industrially are sodium hydroxide [86], potassium hydroxide [87], potassium carbonate [88], and anion-exchange resins. A schematic flowsheet of the Rohm & Haas [89] process is shown in Figure 3. Acetone and liquid hydrogen cyanide are fed continuously to a cooled reactor along with an alkaline catalyst. The catalyst is next neutralized with sulfuric acid and the resulting salt is removed by filtration. The crude product is then distilled in a two-stage process. The overheads from the first column consist mainly of acetone and hydrogen cyanide, which are recycled to the reactor. The second column removes water overhead and leaves 98 % pure acetone cyanohydrin at the bottom. Nitto Chemical claims a two-column distillation system that delivers acetone cyanohydrin of 99.1 % purity [90]. The manufacture of acetone cyanohydrin produces no byproducts other than small amounts of sulfate salts formed during catalyst neutralization. However, the conversion of acetone cyanohydrin to methacrylate in the classical process produces a large amount of ammonium sulfate byproduct, which is usually pyrolyzed to sulfuric acid. An alternative process developed by Mitsubishi Gas Chemical recycles the HCN via formamide. In this process HCN is not directly consumed and no ammonium sulfate is formed.

Uses. By far the largest use of acetone cyanohydrin is as an intermediate in the synthesis of methyl methacrylate [*80-62-6*], methacrylic acid, and higher methacrylate esters. A small amount is converted to methacrylamide. The estimated amount of acetone used in the manufacture of methacrylate esters in 1995 in the United States was 500×10^3 t (Table 7). Based on this estimate, the quantity of acetone cyanohydrin produced as an intermediate was approximately 714×10^3 t.

Certain esters of acetone cyanohydrin, such as 2-chloroethyl-α-cyanoisopropylsulfite [91] and α-cyanoisopropyl-2,6-dichlorobenzoate [92] have strong fungicidal, herbicidal, and insecticidal properties. Nitrilurethanes made from acetone cyanohydrin and substituted phenylisocyanates are intermediates for 4-iminooxazolidin-2-ones, which are plant growth inhibitors [93].

Transportation and Toxicology. Most acetone cyanohydrin is consumed on site for the manufacture of methacrylates. Because of its high toxicity, acetone cyanohydrin is classified

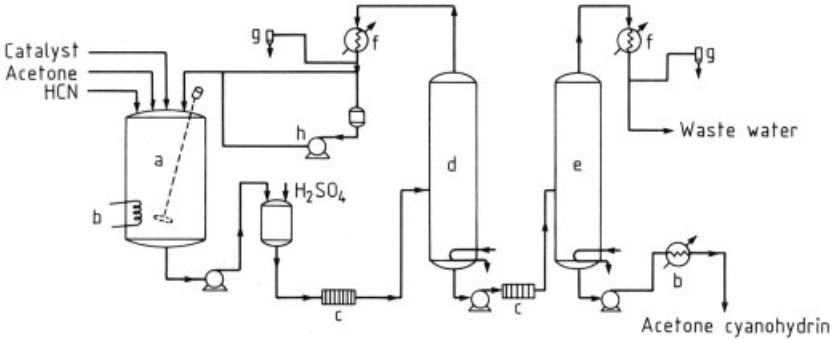

Figure 3. Rohm and Haas acetone cyanohydrin process [89]
a) Reactor; b) Cooling; c) Filter press; d) Concentrator; e) Concentrator; f) Condenser; g) Vacuum jet; h) Pump for acetone and HCN recycle

as a poison B and all shipping containers must carry a "poison inhalation hazard" label [57]. Transport aboard a passenger-carrying or cargo aircraft is forbidden [58]. Acetone cyanohydrin has the following toxicologic properties [94]: LD_{50} (oral, rat) 18.6 mg/kg; LD_{50} (oral, rabbit) 14 mg/kg; LD_{50} (dermal, rabbit) 17 mg/kg; LD_{50} (dermal, guinea pig) 150 mg/kg; aquatic toxicity rating (TLm96) 10 to 1 ppm. Threshold limit values have been established as cyanide: TLV-STEL, 4.7 ppm ceiling limit (5 mg/m^3) [68].

11.2. Diacetone Alcohol

Diacetone alcohol, 4-hydroxy-4-methyl-2-pentanone [*123-42-2*]

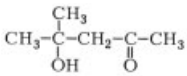

is a dimer of acetone that is used as a solvent and as an intermediate for the manufacture of mesityl oxide, methyl isobutyl ketone, and hexylene glycol.

Physical Properties. Diacetone alcohol is a colorless liquid of mild odor. It is miscible with water and polar solvents and is an excellent solvent for cellulose acetate and various oils and resins. M_r 116.16, *mp* −47 °C, relative density, d_4^{20} 0.9387, refractive index n_D^{20} 1.4235, heat of vaporization at the boiling point at 101.3 kPa 357.1 kJ/kg, specific heat capacity (20 °C) 18.84 kJ kg^{-1} K^{-1}, thermal expansion coefficient (20 °C) 0.00099 K^{-1}, viscosity at 20 °C 2.9 mPa · s, surface tension (20 °C) 31.0 mN/m, dielectric constant (25 °C) 18.2, heat of combustion 3544.5 kJ/mol, flash point 58 °C, lower explosion limit in air 2.6 vol %, auto ignition temperature 624 °C.

Vapor pressure

p, kPa	101.3	1.7	0.108
t, °C	168.1	61.7	20.0

Azeotropic mixture with water: *bp* 98.8 °C/101.3 kPa, 12.7 wt % diacetone alcohol.

Chemical Properties. Diacetone alcohol dehydrates readily in the presence of acids to form mesityl oxide. Catalytic hydrogenation [95] yields hexylene glycol. In the presence of bases, diacetone alcohol reverts to acetone. The reaction is first order in diacetone alcohol and first order in base. The dissociation is accompanied by volume increase; consequently, the reaction is inhibited by pressure [96]. The second-order rate constant is 7.33×10^{-4} L mol^{-1} s^{-1} at 101.3 kPa and decreases to 2.38×10^{-4} L mol^{-1} s^{-1} at 4.05×10^5 kPa. In neutral aqueous solution, dissociation to acetone is very slow at room temperature; it reaches approximately 0.1 % in 1 year.

Diacetone alcohol may be acylated by acetic anhydride under mild conditions to form diacetone alcohol acetate [*1637-25-8*], 4-methyl-4-acetyloxy-2-pentanone, which is claimed as an octane-improving gasoline additive [97]. Con-

densation with urea in the presence of sulfuric acid yields diacetone-monourea, 3,4-dihydro-4,4,6-trimethyl-2(1*H*)-pyrimidone [*4628-47-1*] [98], [99].

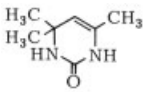

The compound is claimed to improve the egg-laying capacity of hens [98].

Production. Diacetone alcohol is manufactured by self-condensation of acetone in the presence of a basic catalyst. The reaction is exothermic by 14.65 kJ/mol and is easily reversible. The equilibrium concentration of diacetone alcohol is 23.1 wt % at 0 °C and decreases with increasing temperature [100]. Kinetic considerations dictate, however, a higher temperature for the manufacture of diacetone alcohol. An optimum temperature range is 10 – 20 °C.

The self-condensation of acetone is carried out in continuous-flow reactors containing a solid alkaline catalyst, such as barium hydroxide or calcium hydroxide [101], [102]. Anion-exchange resins have been investigated [103], [104], but are not believed to be used commercially. Catalyst performance deteriorates with time, but may last up to 1 year. A patent [105] describes how addition of small amounts of methanol, ethanol, or 2-propanol to the reaction mixture retards catalyst deterioration. The selectivity for diacetone alcohol is 90 – 95 %. Mesityl oxide and higher condensates, such as triacetone alcohol, are the main products. The acetone solution of the crude product is neutralized, e.g., with phosphoric acid [106], prior to concentration under reduced pressure. The recovered acetone is recycled to the condensation reactor and the acidity adjusted by subsequent addition of a base, such as triethylamine. After such treatment, diacetone alcohol of 99.68 % purity was obtained by vacuum distillation [106].

Uses. Diacetone alcohol is an excellent solvent for many natural and synthetic resins. It is used in the coatings industry, especially for hot lacquers, and is also used as a solvent for nitrocellulose, cellulose acetate, and epoxy resins. However, its use has diminished because it is not exempt from restrictions under Rule 66 and related federal regulations [52]. Therefore, US sales have declined more than 50 % since 1978, reaching 10 000 t/a for the past several years and are expected to remain flat. A large portion of diacetone alcohol is used as an intermediate for the manufacture of mesityl oxide, methyl isobutyl ketone, and hexylene glycol.

Transportation and Toxicology. Diacetone alcohol has a flash point of 58 °C, and all transport containers must carry a "flammable liquid" label. Threshold limit values (TLV) for diacetone alcohol vapor at the workplace are: 50 ppm, 240 mg/m^3 (TLV-TWA) [69]; MAK is 50 mL/m^3, 240 mg/m^3. These limits also apply to the other major industrial countries [71]. The toxicologic properties of diacetone alcohol are as follows [107]: LD_{50} (oral, rat) 4 g/kg; LD_{50} (intraperitoneal, mouse) 933 mg/kg; LD_{50} (dermal, rabbit) 13.5 g/kg; aquatic toxicity rating (TLm96) 1000 – 100 ppm.

11.3. Miscellaneous Derivatives

Acetone is used in the production of methyl amyl ketone (MAK, 2-heptanone). MAK is produced by the condensation of *n*-butyraldehyde and acetone. In 1995, US consumption of acetone for MAK was approximately 12×10^3 t, producing ca. 17.5×10^3 t of MAK. Eastman Chemical is the sole US producer. Acetone is also used to produce methyl isoamyl ketone (MIAK) by condensation of acetone with isobutyraldehyde. Eastman Chemical is the sole US producer of MIAK, which is primarily used in lacquers and surface coatings. Other minor uses of acetone include the manufacture of DuPont triazine herbicide Bladex [$(3.5 – 3.6) \times 10^3$ t/a]; Ethoxyquin, a Monsanto antioxidant (2.7×10^3 t/a), as a raw material in the production of hexafluoroacetone, methylbutynol, and pseudoionone; and as an auxiliary blowing agent for the production of flexible polyurethane foam.

12. References

General References

1. *Beilstein* **1** 635, **1 (1)** 335, **1 (2)** 692. **1 (3)** 2696, **1 (4)** 3180. (Acetone); **3** 316, **3 (2)** 224, **3 (3)** 597 (Acetone cyanohydrin).

2. *Ullmann,* 4. Aufl., **7 : 25.**

Specific References
3. *Beilstein* **1** (3) 2707.
4. L. H. Horsley: *Azeotropic Data-III.* Advances in Chemistry Series No. 116, American Chemical Society, Washington, D.C.1973.
5. L. Bretherick: *Handbook of Reactive Chemical Hazards,* CRC Press, Cleveland 1981, p. 362.
6. G. B. Kaufmann, *Chem. Eng. News* **35** (1957) no. 43, 60.
7. R. Delhez, *Chem. Ind.* (London) 1956, 931.
8. H. Seidl, *Angew. Chem. Intern. Ed. Engl.* **3** (1964) 640; *Angew. Chem.* **76** (1964) 716.
9. A. Naponen, *Chem. Eng. News* **55** (1977) no. 8, 5.
10. H. K. King, *Chem. Ind. (London)* 1970, 185.
11. D. H. Grant, *Chem. Ind. (London)* 1970, 919.
12. United States Dept. of the Navy, U.S. Appl. 109 692, 1980 (B. E. Douda, C. F. Parrish, J. E. Short Jr.); *Chem. Abstr.* **93** (1980) 240 992.
13. D. A. Boiston, *Br. Chem. Eng.* **13** (1968) 85.
14. *Hydrocarbon Process.* **44** (1965) no. 11, 231; *Chem. Eng. (N.Y.)* **71** (1964) no. 20, 78.
15. Allied Signal Corp., Specifications for Oxidation Grade Cumene (Partial Listing), 1983.
16. J. A. Howard in: G. H. Williams (ed.): *Advances in Free Radical Chemistry,* **vol. 4,** Academic Press, New York 1972, p. 49.
17. H. C. Bailey, G. W. Godin, *Trans. Faraday Soc.* **52** (1956) 68.
18. A. K. Roby, J. P. Kingsley, *Chem. Tech.* (1996) 41.
19. J. B. Fleming, J. R. Lambrix, J. R. Nixon, *Hydrocarbon Process* **55** (1976) no. 1, 185.
20. H. W. Melville, S. Richards, *J. Chem. Soc.* 1954, 944.
21. K. Hattori, Y. Tanaka, H. Suzuki, T. Ikawa, H. Kubota, *J. Chem. Eng. Jpn.* **3** (1970) 72.
22. C. G. Hagberg, F. X. Krupa, *Chem. React. Eng. Proc. Int. Symp.* **4** (1976) 408.
23. P. Andrigo, A. Caimi, P. Cavalieri d'Oro, A. Fait, L. Roberti, M. Tampieri, V. Tartari, *Chem. Eng. Sci.* **47** (1992) 2511.
24. Hercules Powder Co., US 2 484 841, 1949 (E. J. Lorand).
25. Allied Corp., US 2 613 227, 1950 (G. G. Joris).
26. Allied Corp., US 2 757 209, 1956 (G. G. Joris).
27. Allied Corp., US 3 404 901, 1975 (R. L. Feder, R. Fuhrmann, J. Pisanchyn, S. Elishewitz, T. H. Insinger, C. T. Mathew).
28. Allied Corp., US 4 358 618, 1982 (S. Sifniades, A. A.Tunick, F. W. Koff).
29. J. L. Delaney, T. W. Hughes: *Source Assessment: Manufacture of Acetone and Phenol from Cumene,* Environmental Protection Agency Report No. EPA-600/2-79-019d,1979. Available NTIS PB80-150592.
30. P. R. Pujado, J. R. Salazar, C. V. Berger, *Hydrocarbon Process,* **55** (1976) no. 3, 91.
31. J. F. Knifton, J. R. Sanderson, *Appl. Catal. A* **161** (1997) 199.
32. V. A. Galegov, J. E. Pokrovskaya, V. R. Rakhimov, *Int. Chem. Eng.* **16** (1976) no. 3, 454.
33. Mitsui Petrochemical Industries, Ld. JP 74 46 278, 1974 (T. Akira); *Chem. Abstr.* **84** (1976) 38 444.
34. G. Messina, L. Lorenzoni, O. Cappellazzo, A. Gamba, *Chim. Ind.* **65** (1983) no. 1, 10.
35. J. C. Bonacci, R. M. Heck, R. K. Mahendroo, G. R. Patel, E. D. Allan, *Hydrocarbon Process.* **59** (1980) no. 11, 179.
36. General Electric, US 5 254 75, 1993 (V. M. Zakoshansky).
37. H. J. Kolb, R. L. Burwell, Jr., *J. Am. Chem. Soc.* **67** (1945) 1084.
38. Standard Oil Development Co., US 2 549 844, 1951 (H. O. Mottern).
39. Knapsack-Griesheim AG, GB 804 132, 1958.
40. Toyo Rayon Co., Ltd., JP 68 03 163, 1968 (T. Miyata, M. Sato); *Chem. Abstr.* **70**(1969) 19 573.
41. Engelhard Industries, GB 823 514, 1959; *Chem. Abstr.* **54** (1960) 7562d.
42. Usines de Melle, GB 1 097 819, 1968; *Chem. Abstr.* **68** (1968) 63105.
43. J. Smidt, H. Krekeler, *Hydrocarbon Process Pet Refiner. J.* **42** (1963) no. 7, 149.
44. *Chem. Eng. (N.Y.)* **70** (Sept. 30, 1963) 48.
45. Hoechst, US 3 149 167, 1964 (L. Hornig, E. Paszthory, R. Wimmer).
46. T. Kunugi, T. Matsuura, S. Oguni, *Hydrocarbon Process.* **44** (1965) no. 7, 116.
47. A. H. Olzinger, *Chem. Eng. (N.Y.)* **82** (June 9, 1975) 50.
48. R. Marchal, *Rev Inst Fr Pet* **37** (1982) no. 3, 389.
49. D. I. C. Wang, C. L. Coney, A. L. Demain, R. F. Gomez, A. J. Sinskey, *Degradation of Cellulosic Biomass and Its Subsequent Utilization for the Production of Chemical Feedstocks.* Report 1979, Department of Energy ET/ 20030-1; *Chem. Abstr.* **95** (1981) 153 746.
50. U.S. Code of Federal Regulations 40 117.3, 1981.

51. C. R. Fox, *Hydrocarbon Process.* **57** (1978) no. 11, 269.
52. Rules and Regulations, County of Los Angeles Air Pollution Control District, Los Angeles, CA, Rule 66, amended Aug. 31, 1971; Rule 442, amended March 5, 1982.
53. 40 CFR Part 51 (1998).
54. J. A. Riddick, W. B. Bunger: *Organic Solvents,* Wiley-Interscience, New York 1970, p. 722.
55. G. Zweig, J. Sherma (eds.): *Handbook of Chromatography,* **vol. 1,** CRC Press, Cleveland, Ohio 1972, p. 56.
56. 49 CFR 173.242, 1997.
57. 49 CFR 172.100, 172.101, 1997.
58. IATA Dangerous Goods Regulations 38th ed., Jan. 1997.
59. "Acetone", SRI Consulting, *Chemical Economics Handbook*, April 1996.
60. *Chem. Mark. Rep.* 1997 (March 17), SR4.
61. *Chem. Week,* 1997 (Aug. 27), 68.
62. *Chem. Mark. Rep.* 1996 (Aug. 5), 53.
63. *Chem. Mark. Rep.* 1996 (Jan. 22), 4.
64. Chemical Products Synopsis 1995 (March), A Reporting Service of: Mannsville Chemical Prod. Corp, Adams, NY.
65. *Chem. Mark. Rep.* 1983 (Jan. 17), 3, 17.
66. Asahi Chemical, JP Kokai 02 188 542, 1990, (M. Furuya, H. Nakajima, Y. Fukuoka)
67. W. J. Krasavage, J. L. O'Donoghue, G. D. DiVincenzo: *Patty's Industrial Hygiene and Toxicology,* 4th ed., **vol. 2A,** Wiley-Interscience, New York 1993, p. 149.
68. American Conference of Governmental Industrial Hygienists (ed.): *Threshold Limit Values (TLV),* Cincinnati, Ohio 1995.
69. 29 CFR 1910, 1000, 1997.
70. Acetone (HAZARDTEXT Hazard Management). In: Hall AH @ Rumack BH (Eds.): TOMES System Micromedex, Inc., Englewood, Colorado Edition expires 1/31/98.
71. *Registry of Toxic Effects of Chemical Substances,* **vol. 1,** U.S. Department of Health and Human Services, Washington, D.C. 1985 – 1986, p. 120.
72. K. W. Nelson, J. F. Ege Jr., N. Ross, L. E. Woodman, L. Silverman, *J. Ind. Hyg. Toxicol.* **25** (1943) 282.
73. L. Parmeggiani, C. Sassi, *Med. Lav.* **45** (1954) 431.
74. T. Matsushita, T. Yoshea, A. Yoshimune, T. Inoue, F. Yamata, H. Suzuki, *Jpn. J. Ind. Health* **11** (1969) 477.
75. R. L. Raleigh, W. A. McGee, *J. Occup. Med.* **14** (1972) 607.
76. G. D. DiVincenzo, F. J. Yanno, B. D. Astill, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 329.
77. A. P. Lupulesku, D. J. Birmingham, H. Pinkus, *J. Invest. Dermatol.* **60** (1973) 33.
78. L. C. Harris, R. H. Jackson, *Br. Med. J.* **2** (1952) 1024.
79. S. Gitelson, A. Werczberger, J. B. Herman, *Diabetes* **15** (1966) 810.
80. P. S. Spencer, M. C. Bischoff, H. H. Schaumburg, *Toxicol. Appl. Pharmacol.* **44** (1978) 17.
81. G. A. Mourkides, D. C. Hobbs, R. E. Koeppe, *J. Biol. Chem.* **234** (1959) 27.
82. T. D. Price, D. Rittenberg, *J. Biol. Chem.* **185** (1950) 449.
83. A. A. Michurin, E. A. Sivenkov, E. N. Zilberman, T. I. Tretyakova, *J. Appl. Chem. USSR* **47** (1974) 1383.
84. P. A. S. Smith: *Open Chain Nitrogen Compounds,* **vol. 1,** W. A. Benjamin, New York 1965, p. 217.
85. J. Hine: *Structural Effects on Equilibria in Organic Chemistry,* Wiley & Sons, New York 1975, p. 259.
86. US 2731490, 1953 (G. Barsky).
87. Du Pont, US 2101823, 1954 (H. R. Dittmar).
88. American Cyanamid Co., US 2537814, 1951 (H. S. Davis).
89. M. Salkind, E. H. Riddle, R. W. Keefer, *Ind. Eng. Chem.* **51** (1959) 1232.
90. Nitto Chemical Co. JP-Kokai 7511020, 1975 (K. Nakai, H. Owa, S. Kezuka); *Chem. Abstr.* **84** (1976) 43369.
91. Rohm & Haas Co., US 3052702, 1962 (H. F. Wilson).
92. Tenneco Chemicals, US 3371107, 1968 (J. F. DeGaetano).
93. Etat Français, BE 644178, 1964 (J. Boileau, M. Faidutti, J. P. Konrat, R. Billaz); *Chem. Abstr* **63** (1965) 11567b.
94. [71], vol. 3a, p. 3025.
95. Società Italiana Serie Acetica Sintetica, BE 869056, 1978; *Chem. Abstr.* **90** (1979) 151571.
96. A. Gronlund, B. Andersen, *Acta Chem. Scand.,* Ser. A 33 (1979) 329.
97. Texaco, US 3181 938, 1959 (G. W. Eckert, H. Chafetz).
98. Harvey Research Corp., US 2782197, 1957 (M. T. Harvey).
99. T. Inoi, T. Okamoto, Y. Koizumi, *J. Org. Chem.* **31** (1966) 2700.
100. F. C. Craven, *J. Appl. Chem.* **13** (1963) 71.
101. *Hydrocarbon Process J.* **48** (1969) no. 11, 205.
102. R. A. Garcia, J. V. Sinistera, J. M. Marinas, *React. Kinet. Catal. Lett.* **18** (1981) 33.

103. H. Matyschok, S. Ropuszynski, *Chem. Stosow. Ser. A* **12** (1968) 283; *Chem. Abstr.* **69** (1968) 95877.
104. Z. N. Verkhovskaya, M. Ya. Klimenko, E. M. Zalevskaya, I. N. Bychkova, *Khim. Prom.* **43** (1967), 500; *Chem. Abstr.* **68** (1968) 29204 g.
105. Mitsui Petrochemical Industries, JP-Kokai 80108831, 1980; *Chem. Abstr.* **95** (1981) 97067.
106. J. Przondo, E. Bielous, I. Franek, *Przem. Chem.* **59** (1980), 436; *Chem. Abstr.* **94** (1981) 30132.
107. [71], vol. 1, p. 244.

Acetonedicarboxylic Acid → Oxocarboxylic Acids
Acetonitrile → Nitriles
Acetophenone → Ketones
Acetylacetone → Ketones
Acetyl Chloride → Acetic Acid
Acetylcholine → Neuropharmacology

Acetylene

Peter Pässler, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 1, 3, 4.1, 4.2, 4.4.1, 5.1, 6.2 and 6.3)

Werner Hefner, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 1, 3, 4.1, 4.2, 4.4.1, 5.1, 6.2 and 6.3)

Klaus Buckl, Linde AG, Höllriegelskreuth, Germany (Chap. 2, 4.4.2, 5.1 and 7)

Helmut Meinass, Linde AG, Höllriegelskreuth, Germany (Chap. 5.1 and 5.2)

Andreas Meiswinkel, Linde AG, Linde Engineering Division, Pullach, Germany (Chap. 2, 4.4.2, 5.1, 5.2, 6.1 and 7)

Hans-Jürgen Wernicke, Linde AG, Höllriegelskreuth, Germany (Chap. 6.1)

Günter Ebersberg, Degussa-Hüls AG, Marl, Germany (Chap. 4.3.1, 4.3.2 and 4.3.3)

Richard Müller, Degussa-Hüls AG, Marl, Germany (Chap. 4.3.1 and 4.3.3)

Jürgen Bässler, Uhde GmbH, Dortmund, Germany (Chap. 4.3.4)

Hartmut Behringer, Hoechst Aktiengesellschaft, Werk Knapsack, Germany (Chap. 4.3.4)

Dieter Mayer, Hoechst Aktiengesellschaft, Pharma-Forschung, Toxikologie, Frankfurt, Germany (Chap. 8)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 5
3.1. Industrially Important Reactions 5
3.2. Other Reactions; Derivatives . . . 8
4. Production 9
4.1. Thermodynamic and Kinetic Aspects 9
4.2. Partial Combustion Processes . . 10
4.2.1. BASF Process (Sachsse-Bartholome) 11
4.2.2. Other Partial Combustion Processes 17
4.2.3. Submerged Flame Process 18
4.2.4. Partial Combustion Carbide Process 19
4.3. Electrothermic Processes 20
4.3.1. Production from Gaseous and/or Gasified Hydrocarbons (Hüls Arc Process) 21
4.3.2. Production from Liquid Hydrocarbons (Plasma Arc Process) 25
4.3.3. Production from Coal (Arc Coal Process) 26
4.3.4. Production from Calcium Carbide . 28
4.3.4.1. Wet Generators 29
4.3.4.2. Dry Generators 30
4.3.4.3. Acetylene Purification 32
4.4. Other Cracking Processes 32
4.4.1. Thermal Cracking By Heat Carriers 32
4.4.2. Acetylene as a Byproduct of Steam Cracking 34
5. Safety Precautions, Transportation, and Storage . . . 36
5.1. General Safety Factors and Safety Measures 36
5.2. Acetylene Storage in Cylinders . . 42
6. Uses and Economic Aspects 44
6.1. Use in Metal Processing 44
6.2. Use as Raw Material in Chemical Industry 45
6.3. Competitive Position of Acetylene as Chemical Feedstock 46
7. Propyne 47
8. Toxicology and Occupational Health 50
9. References 51

1. Introduction

Acetylene [*74-86-2*] is the simplest hydrocarbon with a triple bond. In the days before oil gained widespread acceptance as the main feedstock of chemical industry, acetylene was the predominant building block of industrial organic chemistry. The calcium carbide process was the sole route for acetylene production until 1940, when thermal cracking processes using methane and other hydrocarbons were introduced. At first, these processes used an electric arc; then,

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_097.pub2

in the 1950s, partial oxidation and regenerative processes were developed.

However, along with the expansion of the petroleum industry there was a changeover from coal chemistry to petrochemistry, in the 1940s in the United States and in the 1950s in Europe. As a consequence, acetylene lost its competitive position to the much cheaper and more readily available naphtha-derived ethylene and other olefins. This competition between acetylene and ethylene as feedstocks for chemical industry has been much discussed in the 1960s and 1970s [1, 2]. The few hopes, such as BASF's contribution to the submerged flame process, Hoechst's crude oil cracking (HTP), or Hüls' plasma process, have not halted the clear trend toward ethylene as a basic chemical. With the first oil price explosion in 1973, the development of crude cracking processes suffered a setback, and the new processes, such as the Kureha/Union Carbide process, DOW's PCC process (PCC = partial combustion cracking), or the Kureha/Chiyoda/Union Carbide ACR process (ACR = advanced cracking reactor), raise little hope for a comeback of acetylene chemistry. Acetylene production peaked in the United States at 480000 t in the 1960s, and in Germany at 350000 t in the early 1970s [3]. Since then, acetylene production has decreased steadily. In both countries the losses were principally in carbide-derived acetylene; in fact, Germany has produced acetylene for chemical purposes almost exclusively from natural gas and petrochemical sources since 1975.

All acetylene processes, including carbide processes, are high-temperature processes, requiring a large amount of energy. They differ essentially only in the manner in which the necessary energy is generated and transferred. They can be classified into three groups: partial combustion processes, electrothermic processes, and processes using heat carriers. Finally, the use of byproduct acetylene from olefin plants is economically viable in many cases. For each group of acetylene processes several variants have been developed using various feedstocks and techniques. Today, only three processes remain for the commercial production of acetylene: the *calcium carbide route*, in which the carbide is produced electrically, the *arc process*, and the *partial oxidation of natural gas*. Other once popular processes have become uneconomical as the price of naphtha has increased.

Some processes were shelved in the experimental or pilot-plant stage as the importance of acetylene declined. However, other new processes involving the use of coal, sulfur-containing crude oil, or residues as feedstocks for acetylene production are in the pilot-plant stage.

However, the position of acetylene in chemical industry may improve because of the variety of valuable products to which acetylene can be converted with known technology and high yields.

2. Physical Properties

Due to the carbon–carbon triple bond and the high positive energy of formation, acetylene is an unstable, highly reactive unsaturated hydrocarbon. The C – C triple bond and C – H σ bond lengths are 0.1205 and 0.1059 nm, respectively. For the electronic structure of acetylene and a molecular orbital description, see [4]. The acidity of acetylene ($pK_a = 25$) permits the formation of acetylides (see Section 3.2). For comparison the pK_a value of ethylene is 44 and that of acetone 20.

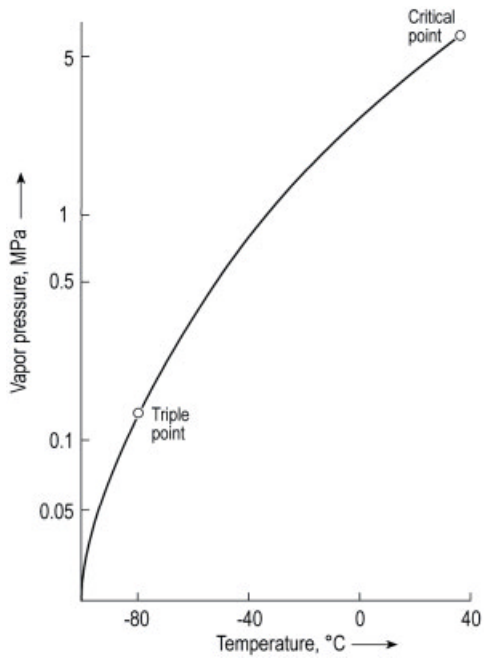

Figure 1. Vapor pressure of acetylene [5, 6]

Under normal conditions acetylene is a colorless, nontoxic but narcotic gas; it is slightly lighter than air. The main physical properties are listed in Table 1. The critical temperature and pressure are 308.32 K and 6.139 MPa. The triple point at 128.3 kPa is 192.4 K. The vapor pressure curve for acetylene is shown in Figure 1. The formation of acetylene is strongly endothermic (ΔH_f = + 227.5 kJ/mol at 298.15 K).

Table 1. Physical properties of acetylene [5 – 8]

Molecular mass	26.0379
Critical temperature	308.32 K (35.17 °C)
Critical pressure	6.139 MPa
Critical volume	0.113 m^3/kmol
Triple point	192.4 K (− 80.75 °C)
Triple point pressure	128.3 kPa
Normal sublimation point and normal boiling point	189.15 K (− 84.0 °C)
Crystal transition point	133.0 K (− 140.15 °C)
Enthalpy of transition	2.54 kJ/mol
Density	760.2 kg/m^3 (131 K) 764.3 kg/m^3 (141 K)
Density (liquid C_2H_2)	465.2 kg/m^3 (273.15 K)
Density (gaseous C_2H_2 at 1 bar)	1.095 kg/m^3 (288.15 K)
Molecular volume (0 °C, 1.013 bar)	22.223 m^3/kmol
Enthalpy of vaporization(calculated)	10.65 kJ/mol (273.15 K)
Enthalpy of sublimation	21.168 kJ/mol (5.55 K)
Enthalpy of formation	227.5 ± 1.0 kJ/mol (298.15 K)
Gibbs free energy of formation	209.2 ± 1.0 kJ/mol (298.15 K)
Entropy of formation	200.8 J mol^{-1} K^{-1} (298.15 K)
Enthalpy of combustion	−1255.6 kJ/mol (298.15 K)
Higher heating value	50 400 kJ/kg
Lower heating value	48 700 kJ/kg
Vapor pressure	2.6633 MPa (273.15 K)
Thermal conductivity	0.0184 W m^{-1} K^{-1} (0 °C, 1.013 bar)
Heat capacity (ideal gas state)	43.990 J mol^{-1} K^{-1} (298.15 K)

Self-decomposition can be initiated when certain pressure limits above atmospheric pressure are exceeded (for details see Section 5.1).

The crystalline structure of solid acetylene changes at − 140.15 °C from a cubic to an orthorhombic phase. The heat of reaction for this phase change is 2.54 kJ/mol [7]; two different values for the enthalpy of fusion are reported in the literature [5, 6]. Figure 2 shows the density of liquid and gaseous acetylene.

Details about flame properties, decomposition, and safety measures are given in Chapter 5.

Solubility coefficients of acetylene in organic solvents are listed in Table 2 [9]. Further solubility data are available as Bunsen absorption coefficients α (20 °C, m^3 (STP) m^{-3} atm^{-1}), as solubilities (g/kg of solvent), and for different pressures (see [10]). The solubilities of acetylene at infinite dilution are shown in Figure 3 for water, methanol, DMF, and *N*-methyl-2-pyrrolidone (NMP) [*872-50-4*]. Figure 4 shows the solubility of acetylene in acetone for various partial pressures and temperatures. While at 20 °C and 1.013 bar 27.9 g (51.0 g) acetylene can be dissolved in 1 kg acetone (DMF) at 20.26 bar 689.0 g (628.0 g) can be dissolved in the same amount of solvent [131]. The heat of solution depends on the concentration of acetylene in the solvent: dissolving 0.5 kg of acetylene in 1 kg of solvent generates 293 kJ for acetone and 335 kJ for DMF. For details on the influence of water, of partial pressure, and deviations from Henry's law, see [9, 10]. Only at temperatures above 25 °C and pressures < ca. 10 bar the solubility of acetylene in acetone follows Henry's law. At higher pressures the solubility increases more rapidly than predicted by Henry's law [132]. The temperature dependence of the solubility of acetylene in DMF at infinite dilution is compared with those of ethylene and ethane in Figure 5 (see also [9, 11] for selectivities).

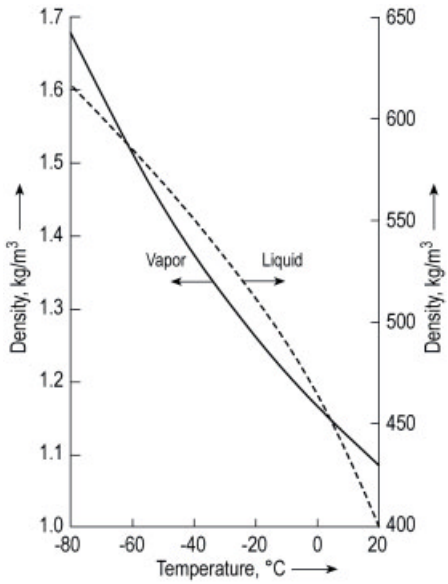

Figure 2. Density of acetylene vapor (at 1.013 bar) and liquid

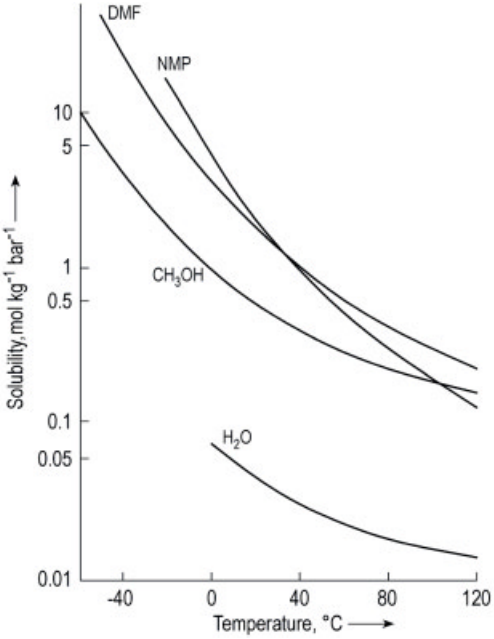

Figure 3. Solubility of acetylene in various solvents at infinite dilution

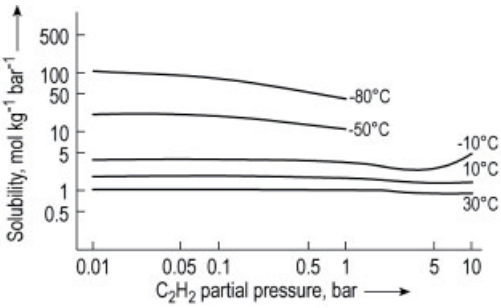

Figure 4. Solubility of acetylene in acetone [9]

The solubility of acetylene in water at 25 °C is 0.042 mol L^{-1} bar^{-1}. Under pressure of acetylene (e.g., > 0.5 MPa at 0 °C) and at temperatures between 268 and 283 K, waxy hydrates of the composition $C_2H_2 \cdot (H_2O)_{\sim 5.8}$ are formed [9, 10]. The hydrates can block equipment; shock waves may initiate self-decomposition.

Liquid oxygen dissolves only traces of acetylene (5.5 ppm at 90 K [12]); the solubilities of ethylene and ethane in oxygen are much higher (factor of 350 and 2280, respectively). The prepurification of the process air in air separation plants with molecular sieves removes acetylene to < 1 ppb provided there is no breakthrough of carbon dioxide. This fact guarantees a safe operation of the downstream equipment [13].

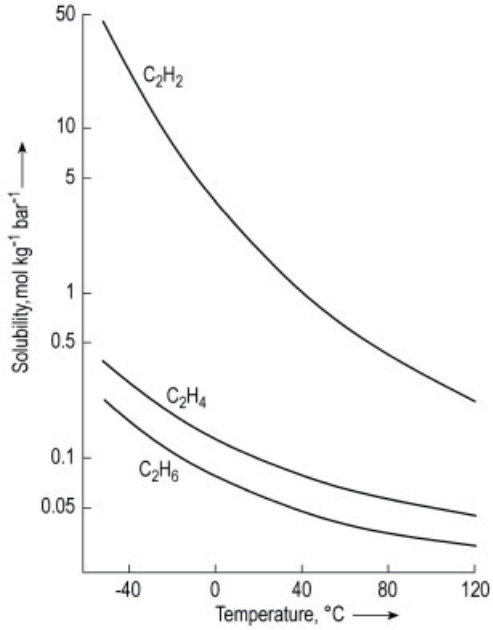

Figure 5. Solubility of C_2 hydrocarbons in DMF at infinite dilution

Typical adsorption isotherms of acetylene are shown in Figure 6 for molecular sieves, activated carbon and silica gel at 25 °C [14]; additional information for activated carbon is summarized in [15].

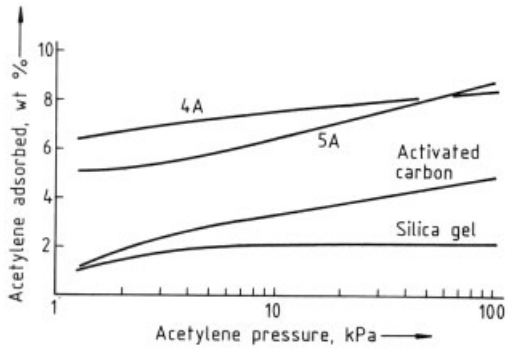

Figure 6. Adsorption isotherms for acetylene on 4A and 5A molecular sieves, activated carbon, and silica gel at 25 °C [14]

Table 2. Solubility coefficients of C_2H_2 in various solvents (in mol kg^{-1} bar^{-1}) [9]

Solvent	C_2H_2 pressure, bar	− 20 °C	25 °C
Methanol	0.98	1.979	0.569
Ethanol	0.98	0.851	0.318
n-Butanol	0.245 – 0.657		0.237
1,2-Dichloroethane	0.4–1.05	0.569	0.218
Carbon tetrachloride	0.98	0.164	0.075
n-Hexane	6.90	0.523	0.264
n-Octane	0.196 – 14.71	0.205	0.146 (0° C)
Benzene	0.98		0.225
Toluene	0.98	0.619	0.214
Xylene (tech.)	0.98	0.528	0.189
4-Methyl-1,3-dioxolan-2-one (propylene carbonate)	0.98	1.137	0.350
Tri-*n*-butylphosphate	0–0.4	2.366	0.614
Methyl acetate	0.98	2.912	0.878
Triethylene glycol	0.98		0.205
Acetone	0.98	4.231	1.069
N-Methyl-2-pyrrolidone	0.98	5.687	1.319
N,*N*-Dimethylformamide	0.98	5.096	1.501
Dimethyl sulfoxide	0.98		1.001
Ammonia	0.98	7.052	2.229

3. Chemical Properties

Because of its strongly unsaturated character and high positive free energy of formation, acetylene reacts readily with many elements and compounds. As a result acetylene is used as raw material for a great variety of substances. Important are addition reactions, hydrogen replacements, polymerization, and cyclization.

Acetylene is more susceptible to nucleophilic attack than, for instance, ethylene. In addition, the polarized C-H bond makes acetylene acidic (pK_a = 25) [16]. Because of this acidity, acetylene is very soluble in basic solvents [17, 18], forming hydrogen bonds with them [19]. Therefore, the vapor pressures of such solutions cannot be described by Raoult's law [20].

The development of the acetylene pressure reactions by W. Reppe (1892 – 1969), BASF Ludwigshafen (Germany) [21 – 23] began modern acetylene chemistry. The most interesting groups of reactions are vinylation, ethynylation, carbonylation, and cyclic and linear polymerization.

3.1. Industrially Important Reactions

Vinylation Reactions and Products [24]. Vinylation is the addition of compounds with a mobile hydrogen atom, such as water, alcohols, thiols, amines, and organic and inorganic acids, to acetylene to form vinyl compounds chiefly used for polymerization.

The two types of vinylation reactions are *heterovinylation* and the less usual *C vinylation.* In the former, the hydrogen atom originates from the heteroatoms O, S, and N, whereas C vinylation occurs when the mobile hydrogen atom is directly bound to a carbon atom. Examples of C vinylation are dimerization and trimerization of acetylene, the synthesis of acrylonitrile from acetylene and hydrogen cyanide, and the addition of acetylene to unsaturated hydrocarbons with activated hydrogen atoms, such as cyclopentadiene, indene, fluorene, and anthracene.

The first industrial vinylation products were acetaldehyde, vinyl chloride, and vinyl acetate. Many other products followed.

Some examples of industrial vinylation processes are given below:

Acetaldehyde [*75-07-0*] (→ Acetaldehyde):

$$HC{\equiv}CH + H_2O \rightarrow CH_3CHO$$

Catalyst: acidic solutions of mercury salts, such as $HgSO_4$ in H_2SO_4 . Liquid-phase reaction at 92 °C.

Vinyl chloride [*75-01-4*] (→ Chlorinated Hydrocarbons):

$$HC{\equiv}CH + HCl \rightarrow CH_2 = CHCl$$

Catalyst: $HgCl_2$ on coal. Gas-phase reaction at 150 – 180 °C.

Vinyl acetate [*108-05-04*] (→ Vinyl Esters):

$$HC{\equiv}CH + CH_3COOH \rightarrow CH_2 = CHOOCCH_3$$

Catalyst: cadmium, zinc, or mercury salts on coal. Gas-phase reaction at 180 – 200 °C.

Vinyl ethers (→ Vinyl Ethers), conjectured reaction steps:

$$ROH + KOH \xrightarrow{-H_2O} ROK \xrightarrow{C_2H_2} RO-CH=CHK$$
$$RO-CH=CHK + ROH \rightarrow RO-CH=CH_2 + ROK$$

where R is an alkyl group. Reaction temperature of 120 – 150 °C; pressure high enough to avoid boiling the alcohol used, e.g., 2 MPa with methanol to produce methyl vinyl ether (acetylene pressure reaction).

Vinyl phenyl ether [*766-94-9*], vinylation with KOH catalyst:

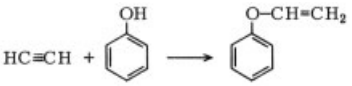

Vinyl sulfides, KOH catalyst:

$$HC{\equiv}CH + RSH \rightarrow CH_2 = CH-S-R$$

Vinyl esters of higher carboxylic acids:

$$HC{\equiv}CH + R-COOH \rightarrow RCOO-CH = CH_2$$

Catalyst: zinc or cadmium salts. Liquid-phase reaction.

Vinyl amines, vinylation with zinc or cadmium compounds as catalyst:

$$R^1R^2NH + HC{\equiv}CH \rightarrow R^1R^2N-CH = CH_2$$

where R^1 and R^2 are alkyl groups.

N-Vinylcarbazole [*1484-13-5*], vinylation of carbazole in a solvent, e.g., *N*-methylpyrrolidone, at 180 °C.

Vinylation of ammonia, complex Co and Ni salts as catalysts, reaction temperature of 95 °C:

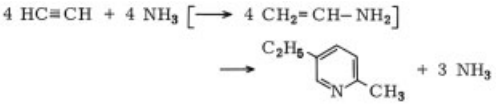

Vinylation of acid amides, potassium salt of the amide as catalyst:

$$HC{\equiv}CH + RCO-NH_2 \rightarrow RCO-NH-CH = CH_2$$

N-Vinyl-2-pyrrolidone [*88-12-0*], vinylation of 2-pyrrolidone with the potassium salt of the pyrrolidone as catalyst.

Acrylonitrile [*107-13-1*], C-vinylation of HCN in aqueous hydrochloric acid with CuCl and NH_4Cl catalyst:

$$HC{\equiv}CH + HCN \rightarrow H_2C = CH-CN$$

Ethynylation Reactions and Products [25]. Ethynylation is the addition of carbonyl compounds to acetylene with the triple bond remaining intact. Reppe found that heavy metal acetylides (see Section 3.2), especially the copper(I) acetylide of composition $Cu_2C_2 \cdot 2\,H_2O \cdot 2\,C_2H_2$, are suitable catalysts for the reaction of aldehydes with acetylene. Alkaline catalysts are more effective than copper acetylide for the ethynylation of ketones. The generalized reaction scheme for ethynylation is:

$$HC{\equiv}CH + RCOR' \rightarrow HC{\equiv}C-C(OH)RR'$$

where R and R′ are alkyl groups or H.

The most important products from ethynylation are propargyl alcohol and butynediol.

Propargyl alcohol, 2-propyn-1-ol [*107-19-7*] (→ Alcohols, Aliphatic):

$$CH{\equiv}CH + HCHO \rightarrow HC{\equiv}CCH_2OH$$

Catalyst: $Cu_2C_2 \cdot 2\,H_2O \cdot 2\,C_2H_2$.

Butynediol, 2-butyne-1,4-diol [*110-65-6*] (→ Butanediols, Butenediol, and Butynediol):

$$HC{\equiv}CH + 2\ HCHO \rightarrow HOCH_2C{\equiv}CCH_2OH$$

Catalyst: $Cu_2C_2 \cdot 2\,H_2O \cdot 2\,C_2H_2$.

Other examples of ethynylation are the reactions of aminoalkanol and secondary amines with acetylene:

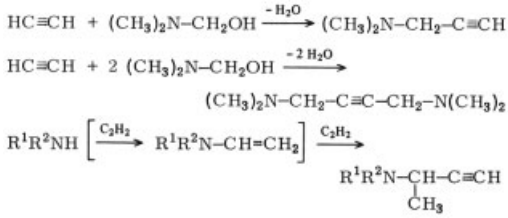

Carbonylation Reactions and Products [26]. Carbonylation is the reaction of acetylene and carbon monoxide with a compound having a mobile hydrogen atom, such as water, alcohols, thiols, or amines. These reactions are catalyzed by metal carbonyls, e.g., nickel carbonyl, $Ni(CO)_4$ [*13463-39-3*]. Instead of metal carbonyls, the halides of metals that can form carbonyls can also be used.

Acrylic acid [*79-10-7*] (→ Acrylic Acid and Derivatives):

$$HC{\equiv}CH + CO + H_2O \rightarrow CH_2 = CH{-}COOH$$

The reaction of acetylene with water or alcohols and carbon monoxide using $Ni(CO)_4$ catalyst was first reported by W. Reppe [26]. If water is replaced by thiols, amines, or carboxylic acids, then thioesters of acrylic acid, acrylic amides, or carboxylic acid anhydrides are obtained.

Ethyl acrylate [*140-88-5*] (→ Acrylic Acid and Derivatives):

Catalyst: nickel salts. Reaction temperature: 30 – 50 °C. The process starts with the stoichiometric reaction (1); afterwards, most of the acrylate is formed by the catalytic reaction (2). The nickel chloride formed in the stoichiometric reaction (1) is recovered and recycled for carbonyl synthesis.

Hydroquinone [*123-31-9*] is formed in a suitable solvent, e.g., dioxane, at 170 °C and 70 MPa [27]. The catalyst is $Fe(CO)_5$:

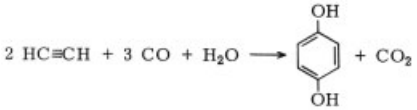

Hydroquinone is formed at 0 – 100 °C and 5 – 35 MPa if a ruthenium carbonyl compound is used as catalyst [28]:

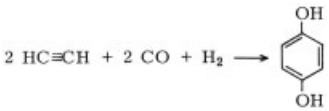

Bifurandiones: The reaction of acetylene and CO in the presence of octacarbonyldicobalt, $(CO)_3Co - (CO)_2 - Co(CO)_3$ [*10210-68-1*], forms a cis–trans mixture of bifurandione. The reaction is carried out under pressure (20 – 100 MPa) at temperatures of about 100 °C [29]:

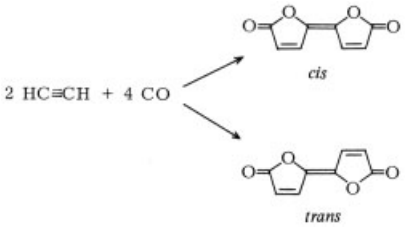

New aspects of such CO insertion reactions have been reported [30].

Cyclization and Polymerization of Acetylene. In the presence of suitable catalysts, acetylene can react with itself to form cyclic and linear polymers.

Cyclization was first observed by Berthelot, who polymerized acetylene to a mixture of aromatic compounds including benzene and naphthalene. In 1940, Reppe synthesized 1,3,5,7-cyclooctatetraene [*629-20-9*] with a 70 % yield at an only slightly elevated pressure:

Reaction temperature of 65 – 115 °C, pressure of 1.5 – 2.5 MPa, $Ni(CN)_2$ catalyst.

The reaction is carried out in anhydrous tetrahydrofuran. The byproducts are mostly benzene (about 15 %), chain oligomers of acetylene of the empirical formulas $C_{10}H_{10}$ and $C_{12}H_{12}$, and a black insoluble mass, called niprene after the nickel catalyst.

If dicarbonylbis(triphenylphosphine)nickel [*13007-90-4*], $Ni(CO)_2[(C_6H_5)_3P]_2$, is used as catalyst, the cyclization products are benzene (88 % yield) and styrene (12 % yield). The reaction is carried out in benzene at 65 – 75 °C and 1.5 MPa [31, 32].

Linear polymerization of acetylene occurs in the presence of a copper (I) salt such as CuCl in hydrochloric acid. Reaction products are vinylacetylene, divinylacetylene, etc. [33]:

$$HC{\equiv}CH + HC{\equiv}CH \rightarrow H_2C = CH{-}C{\equiv}CH$$

A particular polymerization product, known as *cuprene*, is formed when acetylene is heated to 225 °C in contact with copper sponge. Cuprene is chemically inert, corklike in texture, and yellow to dark brown.

Polyacetylene [34, 35] is formed with Ziegler–Natta catalysts, e.g., a mixture of trieth-

ylaluminum, $Al(C_2H_5)_3$, and titanium tetrabutoxide, $Ti(n\text{-}OC_4H_9)_4$, at 10^{-2} to 1 MPa:

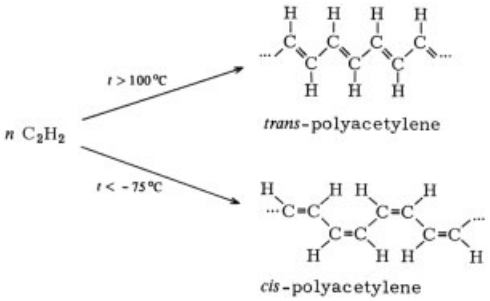

Polymerization can be carried out in an auxiliary inert liquid, such as an aliphatic oil or petroleum ether. The monomer can also be copolymerized in the gas phase.

Polyacetylene is a low-density sponge-like material consisting of fibrils with diameters of 20 – 50 nm. The ratio *cis*- to *trans*-polyacetylene depends on the reaction temperature.

Polyacetylene doped with electron acceptors (I_2, AsF_5), electron donors (Na, K), or protonic dopants ($HClO_4$, H_2SO_4) is highly conductive and has the properties of a one-dimensional metal [35].

3.2. Other Reactions; Derivatives

Metal Acetylides [36]. The hydrogen atoms of the acetylene molecule can be replaced by metal atoms (M) to yield metal acetylides. Alkali and alkaline-earth acetylides can be prepared via the metal amide in anhydrous liquid ammonia:

$$C_2H_2 + MNH_2 \rightarrow MC_2H + NH_3$$

The direct reaction of the acetylene with a molten metal, such as sodium, or with a finely divided metal in an inert solvent, such as xylene, tetrahydrofuran, or dioxane, at a temperature of about 40 °C, is also possible:

$$2\ M + C_2H_2 \rightarrow M_2C_2 + H_2$$

The very explosive copper acetylides, e.g., $Cu_2C_2 \cdot H_2O$, can be obtained by reaction of copper(I) salts with acetylene in liquid ammonia or by reaction of copper(II) salts with acetylene in basic solution in the presence of a reducing agent such as hydroxylamine. Copper acetylides can also form from copper oxides and other copper salts. For this reason copper plumbing should be avoided in acetylene systems.

Silver, gold, and mercury acetylides, which can be prepared in a similar manner, are also explosive.

In sharp contrast to the highly explosive $Cu_2C_2 \cdot H_2O$, the catalyst used for the synthesis of butynediol, $Cu_2C_2 \cdot 2\ H_2O \cdot 2\ C_2H_2$, is not as sensitive to shock or ignition.

Halogenation. The addition of chlorine to acetylene in the presence of $FeCl_3$ yields 1,1,2,2-tetrachloroethane [*79-34-5*], an intermediate in the production of the solvents 1,2-dichloroethylene [*540-59-0*], trichloroethylene [*79-01-6*], and perchloroethylene [*127-18-4*].

Bromine and iodine can also be added to acetylene. The addition of iodine to acetylene stops with formation of 1,2-diiodoethylene.

Hydrogenation. Acetylene can be hydrogenated, partly or completely, in the presence of Pt, Pd, or Ni catalysts, giving ethylene or ethane.

Organic Silicon Compounds [37, 38]. The addition of silanes, such as $HSiCl_3$, can be carried out in the liquid phase using platinum or platinum compounds as catalysts:

$$HC{\equiv}CH + HSiCl_3 \rightarrow CH_2 = CH{-}SiCl_3$$

Oxidation. At ambient temperature acetylene is not attacked by oxygen; however, it can form explosive mixtures with air or oxygen (see Chap. 5). The explosions are initiated by heat or ignition. With oxidizing agents such as ozone or chromic acid, acetylene gives formic acid, carbon dioxide, and other oxidation products. The reaction of acetylene with dilute ozone yields glyoxal.

Hydrates. At temperatures below ca. 15 °C, under pressure, hydrates of the composition $C_2H_2 \cdot 6\ H_2O$ are formed (see Section 2).

Chloroacetylenes [39].

Monochloroacetylene, HC≡CCl, M_r 60.49, *bp* −32 to −30 °C, a gas with nauseating odor that irritates the mucous membranes, is obtained by reaction of 1,2-dichloroethylene with alcoholic NaOH in the presence of $Hg(CN)_2$. It ignites in the presence of traces of oxygen. In air it explodes violently. Chloroacetylene is very poisonous.

Dichloroacetylene, ClC≡CCl, M_r 94.93, *mp* − 66 to − 64.2 °C, a colorless oil of unpleasant odor, explodes in the presence of air or on heating. It is obtained from acetylene in strongly alkaline potassium hypochlorite solution [40] or by reaction of trichloroethylene vapor with caustic alkali.

4. Production

4.1. Thermodynamic and Kinetic Aspects

The production of acetylene from hydrocarbons, e.g.,

$$2\ CH_4 \rightleftharpoons C_2H_2 + 3\ H_2 \quad \Delta H\ (298\ K) = 376.4\ kJ/mol$$

requires very high temperatures and very short reaction times. The main reasons for the extreme conditions are the temperature dependence of the thermodynamic properties (molar enthalpy of formation, ΔH_f, and molar free energy of formation, ΔG_f) of the hydrocarbons; the position of the chemical equilibria under the reaction conditions; and the kinetics of the reaction.

Table 3. Standard molar enthalpies of formation and Gibbs free energy of formation at 298 K

	ΔH_f (kJ/mol)	ΔG_f (kJ/mol)
C (s)	0	0
H_2 (g)	0	0
CH_4 (g)	− 74.81	− 50.82
C_2H_2 (g)	+226.90	+209.30
C_2H_4 (g)	+ 52.30	+ 68.15
C_2H_6 (g)	− 84.64	− 32.90
C_3H_6 (g)	+ 20.43	+ 62.75
C_3H_8 (g)	−103.90	− 23.48
n-C_4H_{10} (g)	−126.11	− 17.10

Thermodynamic data relevant to the hydrocarbon–acetylene system are shown in Table 3 and Figure 7. These data show clearly that at normal temperatures acetylene is highly unstable compared to the other hydrocarbons. However, Figure 7 also shows that the free energy of acetylene decreases as temperature increases, whereas the free energies of the other hydrocarbons increase. Above about 1230 °C, acetylene is more stable than the other hydrocarbons. The temperature at which the acetylene line intersects an other line in Figure 7 is higher the shorter the chain length of the hydrocarbons. Acetylene production from methane requires higher reaction temperatures than production from heavier hydrocarbons.

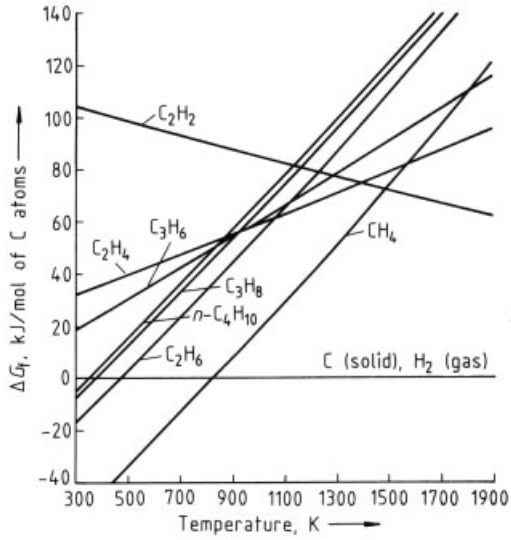

Figure 7. Gibbs free energy of formation per carbon atom of several hydrocarbons as a function of temperature

The equilibrium curve for the methane reaction as a function of temperature (Fig. 8) shows that acetylene formation only becomes apparent above 1000 K (730 °C). Therefore, a very large energy input, applied at high temperature, is required.

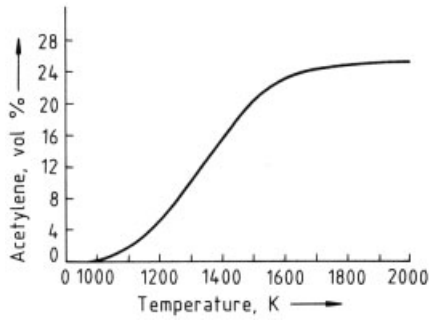

Figure 8. Equilibrium curve for the methane cracking reaction, $2\ CH_4 \rightleftharpoons C_2H_2 + 3\ H_2$

However, even at these high temperatures acetylene is still less stable than its component elements, carbon und hydrogen (see Fig. 7). In

fact, the large difference in free energy between acetylene and its component elements favors the decomposition of acetylene to carbon and hydrogen up to temperatures of about 4200 K.

$$C_2H_2 \rightarrow 2\ C(s) + H_2\ (g) - \Delta G_f\ (298\ K) = -209.3\ kJ/mol$$

Thus cracking and recombination of the hydrocarbons and decomposition of acetylene compete. To achieve reasonable acetylene yields and to avoid the thermodynamically favorable decomposition into the elements, rapid quenching of acetylene produced in the cracking reaction is necessary. In practice, the residence time at high temperature is between 0.1 and 10 ms.

Higher temperatures also increase the rate of conversion of acetylene to byproducts. Again, the residence time must be sufficiently short to prevent this.

In the case of cracking by partial oxidation, the combustion reaction of the hydrocarbon supplies the energy necessary for the production of acetylene from the other part of the hydrocarbon feed:

$$CH_4 + O_2 \rightarrow CO + H_2 + H_2O\ \Delta H\ (298\ K) = -277.53\ kJ/mol$$

$$CO + H_2O \rightarrow CO_2 + H_2\ \Delta H\ (298\ K) = -41.19\ kJ/mol$$

From these reaction enthalpies, the amount of oxygen needed to produce the high reaction temperature can be calculated. Therefore, in addition to the short residence time, the correct methane : oxygen ratio, which also determines the reaction temperature, is essential to obtain good acetylene yields.

4.2. Partial Combustion Processes

In this group of processes, part of the feed is burnt to reach the reaction temperature and supply the heat of reaction. The necessary energy is produced where it is needed. Almost all carbon-containing raw materials can be used as feedstocks: methane, ethane, natural gas liquids (NGL), liquefied petroleum gas (LPG), naphtha, vacuum gas oil, residues, and even coal or coke. Natural gas is especially suitable because it is available in many parts of the world. Only under the conditions of acetylene synthesis can methane be transformed into another hydrocarbon in a single process step, and this is the essential reason for using the thermodynamically unfavorable acetylene synthesis.

The partial combustion processes for light hydrocarbons, from methane to naphtha, all follow similar schemes. The feed and a certain amount of oxygen are preheated separately and introduced into a burner. There they pass through a mixing zone and a burner block into the reaction zone, where they are ignited. On leaving the reaction zone the product mixture is cooled rapidly, either by water or oil. Cooling by water is easier, and more common, but it is thermally less efficient than cooling by oil. Alternatively, the gases can be cooled with light hydrocarbon liquids, which leads to additional acetylene and ethylene formation between 1500 and 800 °C. These processes are usually called *two-step processes*.

Burner design is very important for all partial combustion processes. The residence time of the gas in the reaction zone must be very short, on the order of a few milliseconds, and it should be as uniform as possible for all parts of the gas. Flow velocity within the reaction zone is fixed within narrow limits by the requirements of high yield and the avoidance of preignition, flame separation from the burner block, and coke depositions. A survey of the processes operating according to these principles is given in [7, 41]. Only the BASF process is described here in detail, because it is the most widely used process for the partial combustion of natural gas.

The *submerged flame process*, SFP, was developed by BASF with the aim of producing acetylene from crude oil or its heavy fractions, and thus to be independent of the more expensive refined oil products used in olefin chemistry. One unit of this kind was built in Italy, but it became uneconomic and was shut down after a year of operation [42]. Nevertheless, the process is described in some detail below because of its simple cracking section, because of the simultaneous formation of acetylene and ethylene, and because of its high thermal efficiency and its high degree of carbon conversion (perhaps of even greater importance in the future).

The *partial combustion carbide process*, also developed by BASF, uses coke, oxygen, and lime as feed. It was developed in the 1950s to

reestablish the competitive position of carbide in the face of the new acetylene processes on a petrochemical–natural gas basis. Some attention is given here to the basics of this process, although it has never gone beyond the pilot-plant stage. When petrochemical feedstocks become scarce, this process may have a place in a future coalbased chemistry because it has a higher degree of carbon conversion and a higher thermal efficiency than the electric carbide process.

All these acetylene processes based on partial combustion yield a number of byproducts, such as hydrogen and/or carbon monoxide, which may cause problems if acetylene is the only product desired. Within a complex chemical plant, however, these may be converted to synthesis gas, pure hydrogen, and pure CO and can actually improve the economics of acetylene production.

4.2.1. BASF Process (Sachsse-Bartholome)

The BASF process for the production of acetylene from natural gas has been known since 1950 [43]. Worldwide, some 13 plants used this process in 1983, a total capacity of about 400000 t/a. All use a water quench, except the plant in Ludwigshafen (Germany) operated with an oil quench [44].

The basic idea of partial combustion involves a flame reaction on a premixed feed of hydrocarbon and oxygen. In this way the rate of hydrocarbon conversion is made independent of the gas-mixing rate, which is governed by diffusion. Only then can the residence time in the reaction zone be made much smaller than the average decay time of acetylene. The separate preheating of the reactants to the highest temperature possible before introduction into the burner reduces the consumption of oxygen and the hydrocarbon within the burner. It also causes a higher flame propagation speed and therefore a higher mass flow within the acetylene burner.

The smallest, but most important, part of a partial oxidation acetylene plant is the burner, Figure 9. Its design is nearly identical in the two process variants (i.e., oil and water quench).

At the top of the burner, the preheated reactants, (600 °C in the case of methane) must be mixed (c) so rapidly that there are no domains with a high oxygen concentration. Such domains cause preignition before the reactants are introduced into the reaction zone (g). In fact, the reaction mixture ignites after an induction time depending on the hydrocarbon used as feed and on the preheat temperature, on the order of a few tenths of a second. The maximum preheat temperature is lower for higher hydrocarbons than for methane. Backmixing of the gas between the mixing and the reaction zones is avoided by the diffuser (e), a tube which connects the mixing zone and the burner block (f). Because of its smooth surface and the small opening angle the reaction feed is decelerated gently and backmixing does not occur.

The burner block (f) consists of a water-cooled steel plate with a large number of small channels. The flow velocity through these channels is substantially higher than the flame propagation speed, so that the flame below the burner block cannot backfire into the diffuser. The lower side of the burner block has small openings between the channels through which additional oxygen is fed into the reaction mixture. At these openings small flames form and initiate the flame reaction. The strong turbulence below the burner block stabilizes the flame.

Under unfavorable conditions the flame may appear above the burner block. In this case the oxygen feed must be shut off immediately and replaced by nitrogen. This extinguishes the preignition before it can cause any damage to the equipment. Such preignitions can result from a momentary shift in the oxygen : hydrocarbon ratio or the entrainment of small particles of pyrophoric iron formed from rust in the preheaters.

As mentioned above, the hot gas leaves the reaction chamber within a few milliseconds and passes through sprays of water or oil, which cool the gas almost instantaneously, to about 80 °C in the case of water or 200 – 250 °C in the case of oil. The quench system consists of a set of nozzles that are fed by three annular tubes below the reaction chamber.

The concentrations of the major constituents of the cracked gas depend on the oxygen : hydrocarbon ratio in the feed as shown in Figure 10. As the oxygen supply is increased, the acetylene concentration increases until it passes through a smooth maximum. At the same time there is an increase in the volume of the cracked gas. Thus maximum acetylene production is attained when a little more oxygen is used than the amount

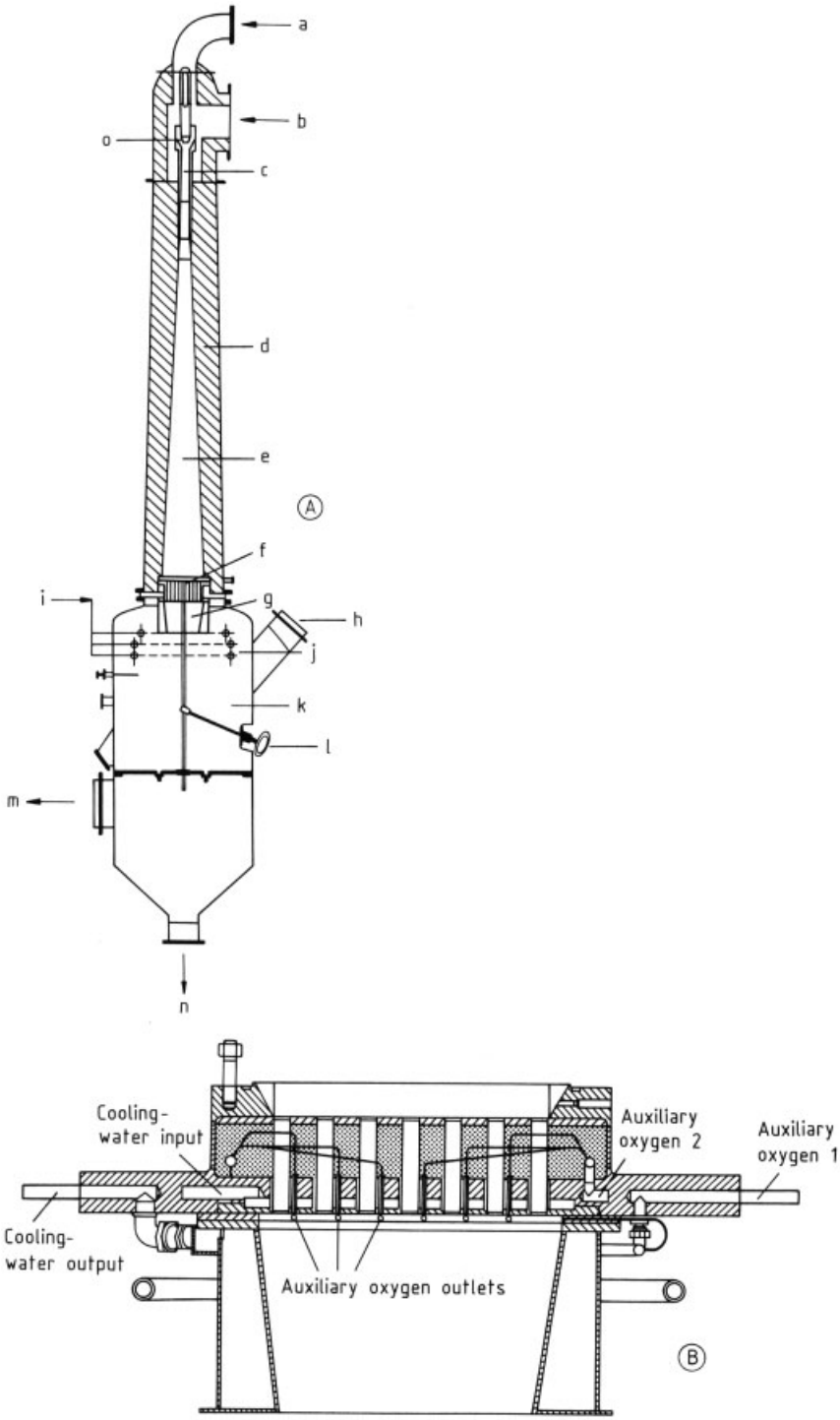

Figure 9. BASF acetylene burner
A) The burner: a) Oxygen: b) Hydrocarbon; c) Mixer; d) Concrete lining; e) Diffuser; f) Burner block; g) Reaction chamber; h) Rupture disk; i) Quench-medium inlet; j) Quench rings; k) Quench chamber; l) Manual scraper; m) Cracked-gas outlet; n) Quench-medium outlet. B) The burner block

required for maximum acetylene concentration in the cracked gas. This is clear from the consumption of natural gas per ton of acetylene produced and the reduction in unconverted methane. When the oxygen : hydrocarbon ratio is too low, the reaction time is insufficient for complete conversion of oxygen, and the cracked gas contains free oxygen. Free oxygen can be tolerated only up to a certain concentration. When the oxygen : hydrocarbon ratio is too high, the increased velocity of flame propagation exceeds the flow velocity in the channels of the burner block, leading to preignitions.

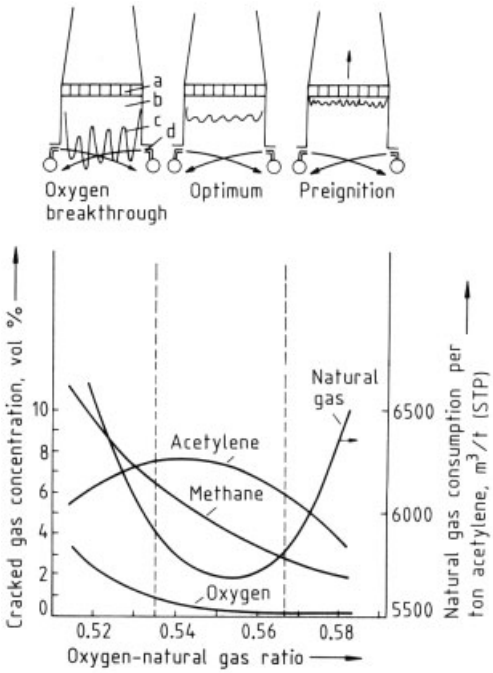

Figure 10. Burner characteristics
a) Burner block; b) Reaction chamber; c) Flame front; d) Quench-medium inlet

Coke deposits in the reaction chamber have to be removed from time to time with a manual or an automatic scraper. Normally, a burner produces 25 t of acetylene per day from natural gas and 30 t per day from liquid feedstocks.

Acetylene Water Quench Process (AWP), Soot Removal (Fig. 11). After quenching with water the cracked gas leaves the burner (b) at 80 – 90 °C. A certain amount of soot is formed in the reaction chamber in spite of the very short reaction time. When natural gas is used as a feedstock, the soot is 50 kg per ton of acetylene, with LPG feedstock it is 250 kg, and with naphtha it is 350 kg. The soot is partly removed from the gas by the quench, then by washing with recirculated water in a cooling column (c), and by passing the gas through an electrofilter (d). After cooling and soot removal, the gas has a pressure slightly above atmospheric, a temperature of about 30 °C, and a soot content of about 1 mg/m^3. The water effluents from the quench system, the cooling column, and the electrofilter carry the washed-out soot. Some gas remains attached to the soot, causing it to float when the soot-containing water flows slowly through basin decanters (e). The upper soot layer, which contains 4 – 8 wt % of carbon, depending on the feedstock, is scraped off the water surface and incinerated.

Acetylene Oil Quench Process (AOP), Soot Removal (Fig. 12). In this process the cracked gas is quenched with oil sprays and leaves the burner at 200 – 250 °C. The oil absorbs the heat from the gas and then passes through waste heat boilers before returning to the quench. The sensible heat of the cracked gas represents more than 15 % of the heating value of the feedstock. The pressure of the generated steam depends on the process configuration and can reach 15 bar (1.5 MPa).

Unlike the water quench process, where the scraped coke deposits sink to the bottom of the quench chamber and are easily removed, in the oil quench the coke deposits do not settle immediately. In order to prevent plugs in the quench nozzles a mill pump (d) is installed immediately underneath the burner column.

The coke and soot content in the quench circuit is kept near 25 % by sending a fraction of the coke-containing oil to externally heated, stirred kettles (coker (e)). In the kettles the volatile matter evaporates very quickly, leading to fluidization of the coke bed. The vapor is returned to the burner column, while the soot is agglomerated. A fine-grained coke is withdrawn from the bottom of the coker.

Because of the cracking losses in the quench a certain amount of quench oil has to be added continuously to the process. This makeup oil is at least 0.15 to 0.3 t per ton of acetylene, depend-

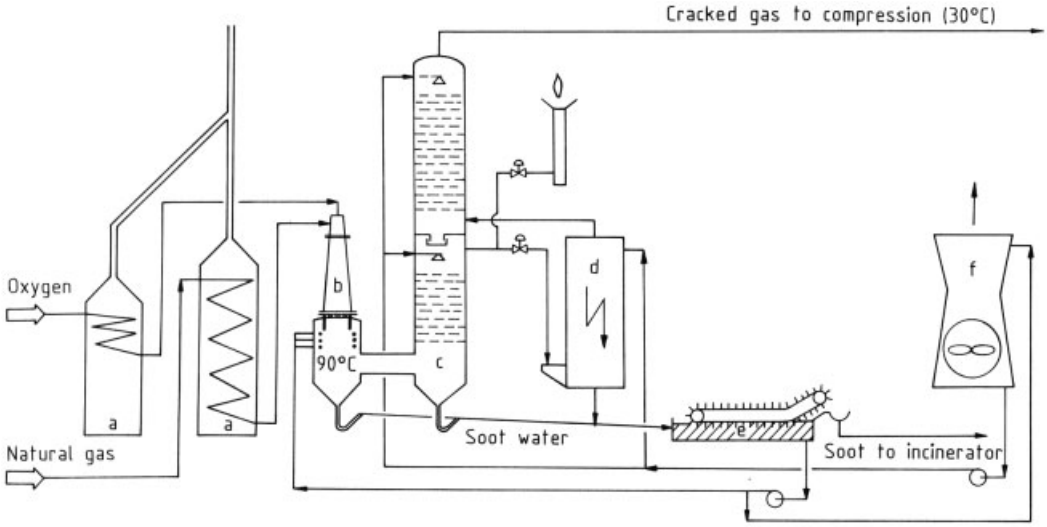

Figure 11. Acetylene water quench process (AWP)
a) Preheaters; b) Acetylene burner; c) Cooling column; d) Electrofilter; e) Soot decanter; f) Cooling tower

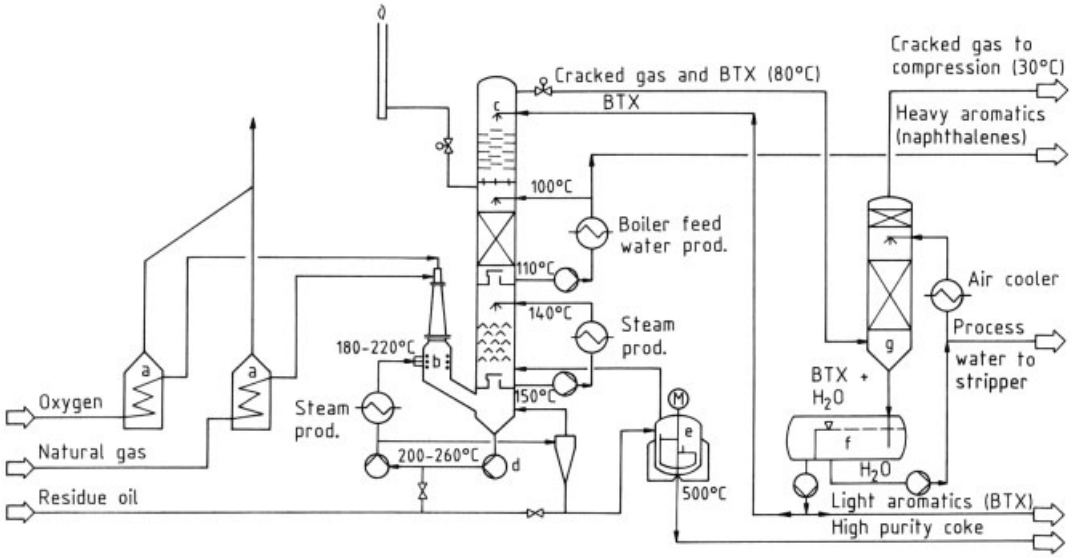

Figure 12. Acetylene oil quench process (AOP)
a) Preheaters; b) Acetylene burner; c) Burner column; d) Mill pump; e) Coker; f) Decanter; g) Final cooler

ing on the stability of the oil used. When residual oil from steam crackers is used, it can be desirable to add up to 1 t of oil per ton of acetylene, because the excess oil is partially converted to light aromatic hydrocarbons.

The cracked gas leaving the quench is cooled in a burner column (c), where there are additional oil circuits for the production of 3-bar steam and for boiler feedwater preheat. At the top of the column a small amount of a low-boiling oil (BTX = benzene, toluene, and xylene) is added to prevent deposit-forming aromatics (mainly naphthalene) from passing downstream into other parts of the plant. The cracked gas, which has to be compressed before separation, is cooled further (g) by water. At this stage most of the BTX condenses and is separated from the water in a large decanter (f).

Table 4 shows the cracked gas compositions for the BASF acetylene oil quench process when natural gas, liquid petroleum gas (LPG), or naphtha is used as feedstock. The water quench process gives very similar compositions. The relative amounts of hydrogen and carbon monoxide formed depend on the hydrogen : carbon ratio of the feedstock used. Even when naphtha is used, almost no ethylene forms. This is because the reaction takes place above 1200 °C where the formation of ethylene is thermodynamically impossible. Only a prequench with additional naphtha or LPG produces additional acetylene and ethylene at intermediate temperatures, as in the case of two-step processes. The higher hydrocarbons require a somewhat lower reaction temperature than methane and have a less endothermic heat of reaction: oxygen consumption per ton of acetylene is lower for the higher hydrocarbons in spite of the lower preheating temperature.

Comparison of Oil Quench and Water Quench Processes. The advantage of the oil quench process is obvious: the heat recovery in the form of steam makes the overall thermal efficiency in relation to primary energy input rather high. If the thermal efficiency for the production of electricity is 33 %, over 70 % of the net heating value of the overall primary energy input is recovered in the form of products and steam. A comparison between the oil quench and water quench (see Table 6) shows that the oil quench requires a net heating value input of 300 – 330 GJ per ton of acetylene, of which 82 GJ (27 – 25 %) is lost, whereas the water quench requires a 288 GJ input, of which 113 GJ (39 %) is lost.

Acetylene Recovery. Liquid acetylene is a dangerous product, even at low temperatures. Separation of the cracked gas by cryogenic processes such as those used in olefin production is clearly ruled out. One exception to this rule is the acetylene recovery unit of the submerged flame process (Section 4.2.3) [45], in which all hydrocarbons except methane are condensed at −165 °C. Otherwise, acetylene is recovered by *selective* absorption into a solvent. This procedure is economical only when the cracked gas is compressed. The upper limit for the pressure is determined by the danger of explosions, and as a rule the partial pressure of acetylene should be kept below 1.4 bar (0.14 MPa).

The solubility of acetylene in the solvents used is between 15 and 35 m^3 (STP) per cubic meter of solvent under process conditions. The dissolved gas is recovered by depressurizing the solvent and by vapor stripping at higher temperatures. All solvents used commercially, *N*-methylpyrrolidone (NMP), methanol, ammonia, and dimethylformamide (DMF), are miscible with water. They are recovered from the gas streams leaving the plant by water scrubbing and distillation.

The kinetics of acetylene formation always lead to the formation of higher homologues of acetylene as byproducts [46], mainly diacetylene, but also methylacetylene, vinylacetylene, and others. These compounds polymerize very easily and must be removed from the cracked gas as soon as possible. Because they are much more soluble in the solvents than acetylene, scrubbing the cracked gas with a small amount of solvent before it enters the acetylene recovery stages is sufficient.

Absorption Section (Fig. 13). Acetylene recovery is illustrated here by the BASF process. *N*-Methylpyrrolidone is used to separate the cracked gas into three streams:

1) Higher homologues of acetylene and aromatics, the most soluble part of the cracked gas. (This is a small stream of gas, which is diluted with crude synthesis gas for safety reasons and is used as fuel.)
2) Product acetylene, less soluble than the higher acetylenes, but much more soluble than the remainder of the gas
3) Crude synthesis gas (off-gas), mainly hydrogen and carbon monoxide

In the prescrubber (b) the cracked gas is brought into contact with a small amount of solvent for removal of nearly all the aromatic compounds and C_4 and higher acetylenes except vinylacetylene. This is done after the compression of the gas if screw compressors are used but before compression if turbo compressors are used because turbo compressors cannot tolerate deposits on their rotors. In the main scrubber (d) the gas is brought into contact with a much larger amount of *N*-methylpyrrolidone (NMP), which dissolves all the acetylene, the remaining homologues, and some carbon dioxide. Crude

Table 4. BASF acetylene oil quench process, cracked gas composition (vol %)

Component *	Raw material (ΔH, kJ/mol)		
	Methane (400)	LPG (325)	Naphtha (230)
H_2	56.5	46.4	42.7
CH_4	5.2	5.0	4.9
C_2H_4	0.3	0.4	0.5
C_2H_2	7.5	8.2	8.8
C_{3+} **	0.5	0.6	0.7
CO	25.8	35.0	37.9
CO_2	3.2	3.4	3.5
O_2	0.2	0.2	0.2
Inerts		balance	

* Dry gas, water, and aromatic compounds condensed out;
** Hydrocarbons with three or more carbon atoms.

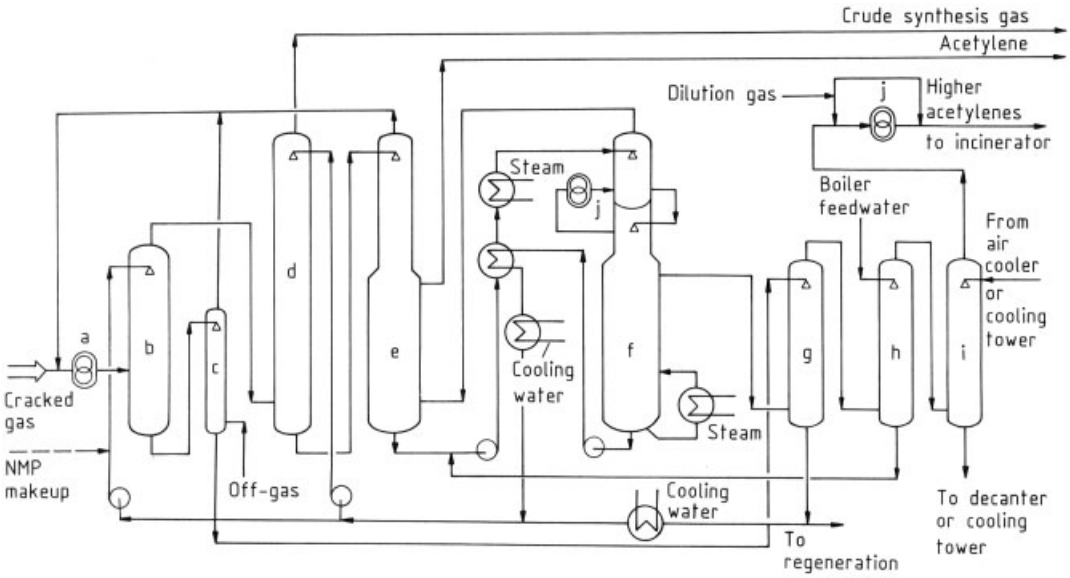

Figure 13. BASF acetylene process — *N*-methylpyrrolidone absorption section
a) Compressor; b) Prescrubber; c) Acetylene stripper; d) Main scrubber; e) Stripper; f) Vacuum column; g) Vacuum stripper; h) Side column; i) Condenser; j) Vacuum pumps

synthesis gas (off-gas) leaves at the top of the column.

The NMP solution is degassed in several steps in which the pressure is reduced and the temperature increased. The stripper (e) operates at pressures and temperatures slightly above ambient. In this tower, the solution is put in contact with a countercurrent gas stream from the subsequent degassing step (f). This leads to the evolution of carbon dioxide, the least soluble of the dissolved gases, at the top of the stripper. The carbon dioxide is recycled to the suction side of the compression and thereby is shifted into the crude synthesis gas. The acetylene product is withdrawn as a side stream from the stripper. The *N*-methylpyrrolidone solution is then completely degassed (f) in two further steps at 110 – 120 °C, first at atmospheric, then at reduced pressure. Vinylacetylene, methylacetylene, and excess process water are withdrawn as bleed streams from the vacuum column (f). The water content of the solvent is controlled by the reboiling rate in the vacuum column. At the bottom of the vacuum column, degassing is completed, and the solvent is cooled and returned to the main scrubber (d).

The small amount of solvent from the prescrubber (b) is stripped with crude synthesis gas for recovery of the dissolved acetylene, the overhead gas being recycled to the suction side of

the compressor. The solvent is then degassed completely in the vacuum stripper (g), a column which also accepts the bleed stream from the vacuum column (f) containing the excess process water together with some higher acetylenes. The overhead vapor of the vacuum stripper contains the higher acetylenes, water, and some NMP vapor. In a side column (h) the NMP is recovered by scrubbing with a small amount of water, which is recycled to the main solvent stream. The gas is cooled (i) by direct contact with water from a cooling circuit to condense most of the water vapor. The higher acetylenes are diluted with crude synthesis gas before they enter and after they leave the vacuum pump (j). The diluted higher acetylenes, which are now at a pressure slightly above atmospheric, can be used as fuel gas, e.g., for soot incineration.

In order to minimize the polymer content of the solvent, about 2 % of the circulating flow is withdrawn continuously from the vacuum stripper circuit and distilled under reduced pressure, leaving the polymers as a practically dry cake for disposal.

The acetylene product from the process as described above has a purity of about 98.4 %, the remainder consisting mainly of propadiene, methylacetylene, and nitrogen. For most applications the purity is increased to 99.7 % by scrubbing with sulfuric acid and sodium hydroxide solutions. Table 5 compares the compositions of crude and purified acetylene. Table 6 compares the consumption and product yields per ton of acetylene for the oil quench process with those for the water quench process.

Table 5. Purity of the acetylene from the BASF process

Component	Crude acetylene, vol %	Purified acetylene, vol %
Acetylene	ca. 98.42	99.70
Propadiene	0.43	0.016
Propyne	0.75	traces
Vinylacetylene	0.05	0
1,3-Butadiene	0.05	0
Pentanes	0.01	0.01
Carbon dioxide	ca. 0.10	0
Nitrogen	ca. 0.30	0.30

4.2.2. Other Partial Combustion Processes

The main features of the BASF process described in detail above are common to all partial oxidation processes. Therefore only the differences between the BASF acetylene burner and burners used in the *Montecatini* and the *SBA processes* [41, 47] are described. These two processes have also attained some importance. The details of the acetylene recovery process depend on the properties of the solvent, but here too the basic principles are the same for all processes.

Montecatini Process . The Montecatini burner [48] has the same main components as the BASF burner: mixing unit, gas distributor, reaction chamber, and quench. The essential difference is the pressure for acetylene synthesis, which can be as high as several bar. This saves compression energy, improves heat recovery from the quench water, which is obtained at 125 °C, and is claimed to make soot removal easier because the cracked gas is scrubbed with water above 100 °C. Although it is well known [7] that acetylene decomposition is accelerated under pressure at high temperatures (> 1000 °C), the acetylene yield is comparable to that obtained at atmospheric pressure because of the short residence time in the reactor. Methanol is used at cryogenic temperatures for acetylene recovery. The main steps of the gas separation are absorption of higher acetylenes and of aromatics, absorption of acetylene, stripping of coabsorbed impurities, and desorption of acetylene.

SBA Process (of the Société Belge de l'Azote). The SBA burner [49] has the same main components as the other processes. However, it has a telescope-like reaction chamber and a device for shifting the quench up and down. Thus it is possible to adjust the length of the reaction zone for optimum residence time at any throughput. The walls of the reaction chamber are sprayed with demineralized water to prevent coke deposits. This eliminates the need to scrape the reaction chamber periodically. Acetylene recovery is carried out with several scrubbing liquids — kerosene, aqueous ammonia, caustic soda, and liquid ammonia, each with its own circuit. After soot is separated from the gas in an electrofilter, higher hydrocarbons are absorbed in kerosene or gas oil. Carbon dioxide is scrubbed in two steps, first with aqueous ammonia and then with caustic soda solution. The acetylene product is absorbed into anhydrous

Table 6. BASF acetylene process, consumption and product yields per ton of acetylene

Consumption and product yields	Oil quench	Water quench
Feed and energy requirements		
Natural gas, 36 000 kJ/m^3(STP) (LHV) *	5833 m^3 = 210 GJ	5694 m^3 = 205 GJ
Oxygen, 0.55 kWh/m^3(STP) **	3400 m^3 = 20.4 GJ	3400 m^3 = 20.4 GJ
Fuel gas	= 12.0 GJ	= 18.0 GJ
Residue oil minimum (surplus)	0.3 (1.0) t = 12.0 (40.0) GJ	
Sulfuric acid	160 kg	160 kg
Sodium hydroxide	5 kg	5 kg
N-Methylpyrrolidone	5 kg	5 kg
Electric energy **	3200 kWh = 34.9 GJ	3100 kWh = 33.8 GJ
Steam, 4 bar	5.0 t = 11.7 GJ	4.5 t = 10.5 GJ
Energy input	301.0 (329.0) GJ	287.7 GJ
Product yields		
Acetylene, 48650 kJ/kg (LHV)	1.0 t = 48.6 GJ	1.0 t = 48.6 GJ
Crude synthesis gas, 12100 kJ/m^3(STP) (LHV)	10600 m^3 = 128.3 GJ	10150 m^3 = 122.8 GJ
Coke (with residue surplus), 35 500 kJ/kg	0.3 (0.46) t = 10.7 (16.3) GJ	–
BTX (with residue surplus), 40 250 kJ/kg (LHV)	0.05 (0.12) t = 2.0 (4.8) GJ	–
Naphthalenes (with residue surplus), 38770 kJ/kg (LHV)	0.0 (0.41) t = – (15.9) GJ	–
Steam (up to 15 bar)	13.0 (14.0) t = 30.3 (32.6) GJ	1.5 t = 3.5 GJ
Energy output	219.9 (246.5) GJ	174.9 GJ
Thermal efficiency	73.0 (74.9) %	60.8 %
Energy losses, absolute per ton acetylene	81.1 (82.5) GJ	112.8 GJ

* If the natural gas contains inerts and higher hydrocarbons, the required input will remain approximately the same on a heating value basis (LHV = low heating value), but the cracked gas analyses and the crude synthesis gas analyses will differ slightly.
** Thermal efficiency of electricity production is assumed to be 33 %.

ammonia and must be scrubbed with water after desorption. All the ammonia–water mixtures are separated in a common distillation unit. This recovery scheme leads to exact separation of the various cracked gas components.

Additional Remarks. The Montecatini and SBA processes can also be operated with two-stage burners. A prequench with light hydrocarbons cools the cracked gas to about 800 °C. After a residence time at this intermediate temperature the gas is cooled down with water. In this way the heat content of the hot gases is used for further cracking of hydrocarbons to yield extra acetylene and olefins. The presence of additional components in the cracked gas requires more process steps in the gas separation units.

4.2.3. Submerged Flame Process

The submerged flame process (SFP) of BASF attracted considerable interest up to 1973 as a partial combustion process for the production of acetylene, ethylene, C_3, and C_4 hydrocarbons, and synthesis gas from feedstock of crude oil and residues, such as Bunker C oil and vacuum residue [44, 45]. Although it was abandoned at the end of 1973, the need to make the most economic use of raw materials has renewed interest in this process [50].

Oxygen compressed to 16 bar (1.6 MPa) feeds a flame that is submerged in the oil. The oil surrounding the flame is partially burnt to obtain the necessary reaction temperature and also acts as the quenching medium. This process differs from the partial oxidation processes using natural gas and lighter hydrocarbons in five main respects:

1) Crude oil can be gasified without the formation of residues, and the process can be operated under certain conditions with heavy fuel oil.
2) All the soot formed is consumed when crude oil feedstock is used, eliminating all the problems associated with the storage, disposal, or utilization of acetylene soot.
3) The heat of reaction is removed by steam generation at 8 bar (0.8 MPa).
4) The process is operated at 9 bar (0.9 MPa) so that the oxygen is the only compressed stream.

The cracked gas is formed at a pressure sufficient for economic separation.

5) The design of the cracking unit is greatly simplified because the reaction feed, fuel, and quenching medium are identical.

The process is described in detail in the literature cited; therefore, only general overviews of the cracking unit (Fig. 14) and the separation unit (Fig. 15) are shown here. The capacities of a submerged flame burner for acetylene and ethylene are 1 t/h and 1.15 t/h, respectively. To produce these quantities, 5000 m^3 (STP) of oxygen and 8 – 10 t of oil are required per hour. The cracked gas shows the following average composition (vol %, the components grouped as streams leaving the separation unit):

Main products	
Acetylene	6.2
Ethylene	6.5
Crude synthesis gas	
Carbon monoxide	42.0
Hydrogen	29.0
Methane	4.0
Inerts	0.6
Other hydrocarbons	
Ethane	0.5
Propane	0.1
Propene	1.2
Propadiene, propyne	0.7
1,3-Butadiene	0.5
Other C_4 and C_{5+} * hydrocarbons	1.5
Remainder	
Carbon dioxide	7.0
Hydrogen sulfide	0.05 – 0.5
Carbon oxide sulfide	0.03 – 0.3

* C_{5+}, five or more carbons

Unlike all other processes the submerged flame process uses low temperatures (–165 °C) to separate the off-gas, consisting of carbon monoxide, hydrogen, and methane, from the C_2 and higher hydrocarbons. On account of the acetylene in the condensed phase, extensive decomposition tests have been carried out. Whereas the cracking unit (Fig. 14) and the amine scrubbing unit have been tested by Soc. Ital. Serie Acetica Sintetica, Milan, on a commercial scale, the remaining purification units (Fig. 15) have not. However, the experience obtained with a pilot plant indicates that major difficulties are not to be expected.

The submerged flame process may become competitive because of its ability to use crude oil and especially residues and because of its low losses on the primary energy input.

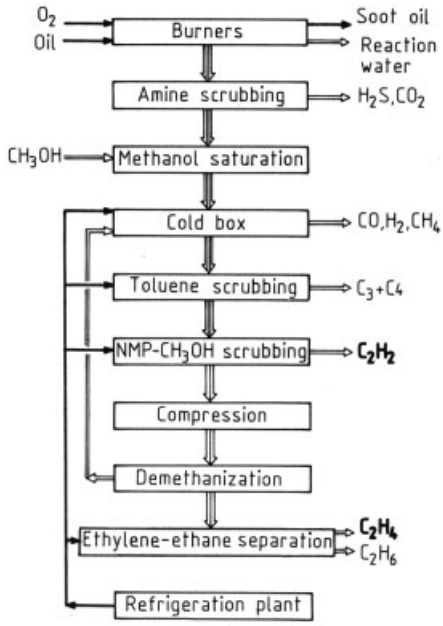

Figure 15. Submerged flame process — purification unit

4.2.4. Partial Combustion Carbide Process

Calcium carbide production from lime and coal requires a large high-temperature heat input (see Section 4.3.4). In the thermal process some of the coal must be burnt to attain the necessary reaction temperature and supply the heat of reaction. The thermal carbide process was developed by BASF [7, 51] from 1950 to 1958 to eliminate the input of electrical energy necessary in the classic carbide process. Starting in 1954, a large pilot plant, with a nominal carbide capacity of 70 t/d, was operated, but in 1958 the more economical petrochemical acetylene production halted further development. Carbide production is just one way of converting coal chemically; other methods include pyrolysis, hydrogenation, and gasification. The question arises as to the conditions under which a thermal carbide process using oxygen can compete with the electric carbide process. The biggest drawback of carbide production in a shaft furnace (Fig. 16) compared to the electric carbide process is the lack of commercial-scale operational experience. Specific disadvantages are greater susceptibility to

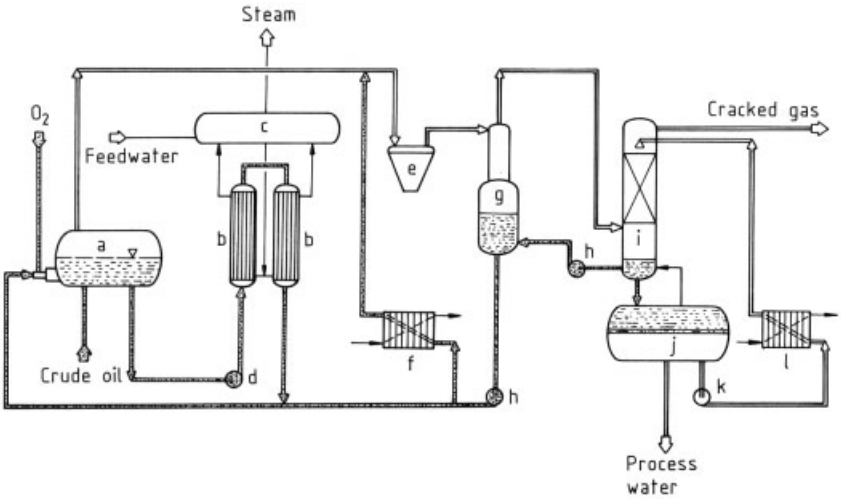

Figure 14. Submerged flame process (SFP) — cracking unit
a) Reactor; b) Oil cooler; c) Steam generator; d) Oil recycle pump; e) Scrubber; f) Naphtha cooler; g) Naphtha separator; h) Naphtha pump; i) Spray cooler; j) Separating vessel; k) Recycle-water pump; l) Recycle-water cooler

disruption because of plugging of the furnace feed, more stringent specifications for the raw materials, more handling of solids, and the large amount of byproduct. There are two main advantages:

1) A thermal efficiency of about 50 % versus about 30 % for the electrothermal process if the thermal efficiency of electricity production is 33 %
2) Carbon monoxide production, which is desirable because carbon monoxide can be converted to synthesis gas by the water-gas shift reaction (→ Gas Production)

Table 7. Partial combustion carbide process, consumption and product yields per ton acetylene

Raw materials	
Coke, dry (88 % C)	5700 kg
Lime (92 % CaO)	3140 kg
Oxygen (98 %) 3560 m^3 (STP)	5090 kg
Total consumption	13 930 kg
Products	
Carbide (80.5 %)	2850 kg ≙ 1000 kg acetylene
Carbon monoxide 7980 m^3 (STP)	9975 kg
(CO 95.5, H_2 2.0, N_2 2.0, CO_2 0.5 vol %)	
Dust	900 kg
Losses	205 kg
Total products	13 930 kg

If the carbon monoxide is converted to synthesis gas and the electrical energy is produced from fossil fuels, production costs are about one third lower for the thermal process than for the electrical process [52] based on the pilot-plant consumption data (Table 7).

4.3. Electrothermic Processes

Because calcium carbide is produced electrothermally, the production of acetylene from this material also is discussed in this group of processes (Section 4.3.4).

Electrothermic processes have the following advantages over partial oxidation:

– The energy requirement for the formation of acetylene can be made independent of the hydrocarbons used as feedstock.
– Hydrocarbon consumption can be reduced by 50 %.
– Provided that electrical energy is available under favorable conditions (nuclear power, hydroelectric power, cheap coal) and/or the availability of hydrocarbons is limited, electrothermic processes are more economical.

In the case of acetylene formation, the electric-arc process offers optimal conditions for the endothermic reaction at high temperatures.

The development of the electric-arc process for cracking light hydrocarbons to acetylene began in 1925 in Germany. The acetylene was to be used as feedstock for butadiene production.

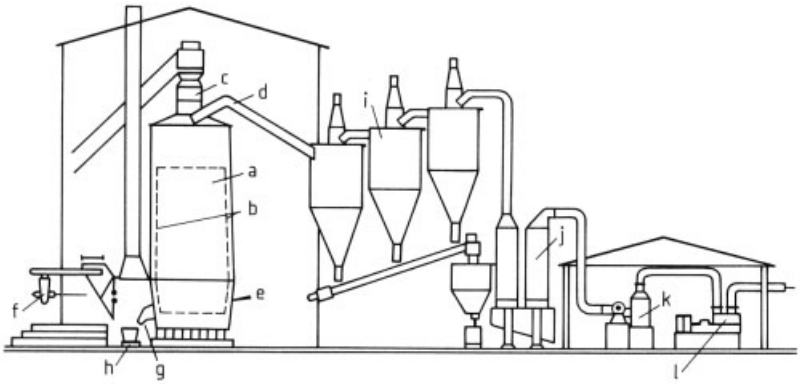

Figure 16. Partial combustion carbide process
a) Carbide furnace; b) Refractory brick lining; c) Charging hopper; d) Gas outlet; e) Oxygen jet; f) Tapping burner; g) Tapping chute; h) Bogey; i) Cyclone; j) Washing column; k) Desintegrator; l) Compressor

In 1940, the first commercial plant was put on stream at Chemische Werke Hüls in Marl, Germany. The Hüls process has since been improved , and the capacity raised, but it retaired the original principles [53].

Feedstock for electric-arc processes may be gaseous or liquid hydrocarbons or even solids such as coal. The design of the arc furnace and the purification section for the cracked products have to be adapted to the different feedstocks. For gaseous or gasified hydrocarbons the classical one-step process is used: the arc burns directly in the gas being cracked. For liquid and solid feeds, a one- or two-step process may be used. In the two-step process hydrogen is first heated in the arc furnace, and then liquid or solid feed is injected into the hydrogen plasma [54]. Figure 17 shows both types of arc furnaces. Because of hydrogen formation during the cracking reaction, the arc burns in a hydrogen atmosphere in both processes. The conductivity and the high rate of ion–electron recombination for hydrogen mean that arcs above a certain length cannot be operated with alternating current at normal frequency and high voltage. All commercial plants therefore run on direct current.

4.3.1. Production from Gaseous and/or Gasified Hydrocarbons (Hüls Arc Process)

The plant for the Hüls arc process includes the arc furnace section itself (Fig. 17 A), which is operated at a pressure of 1.2 bar, and a low and high pressure purification system.

Arc Furnace. A cathode, a vortex chamber, and an anode make up the arc furnace. Cathode and anode are water-jacketed tubes of carbon steel 0.8 m and 1.5 m long, respectively, and with inner diameters of 150 and 100 mm, respectively. The arc burns between cathode and anode with a length of about 1.2 m and with a current of 1200 A. The cathode is connected to the high-voltage side of the rectifier (7.1 kV) and electrically isolated from the other parts of the furnace. Between cathode and anode is the vortex chamber. The gas is injected into it tangentially at a specific velocity to stabilize the arc by creating a vortex. The arc burns in the dead zone, and the striking points of the arc on the electrodes are forced into a rapid rotation so that they only burn for fractions of a millisecond at one point, which gives the electrodes a lifetime up to 1000 h. Temperatures reach 20000 °C in the center of the arc. Because of the tangential flow of the gas, the arc is surrounded by a sharply decreasing coaxial temperature field, and the temperatures at the wall of the electrode are only 600 °C. Thermal losses are therefore limited to less than 10 % of the electrical power input of 8.5 MW.

The residence time of the gas in the arc furnace is a few milliseconds. In this interval, the hydrocarbons are cracked, mainly into acetylene, ethylene, hydrogen, and soot. At the end of the arc furnace, the gases are still at a temperature of about 1800 °C. The high heat content of this gas can be exploited for additional ethylene production by means of a prequench with liquid hydrocarbons. This lowers the temperature to about 1200 °C. Because acetylene

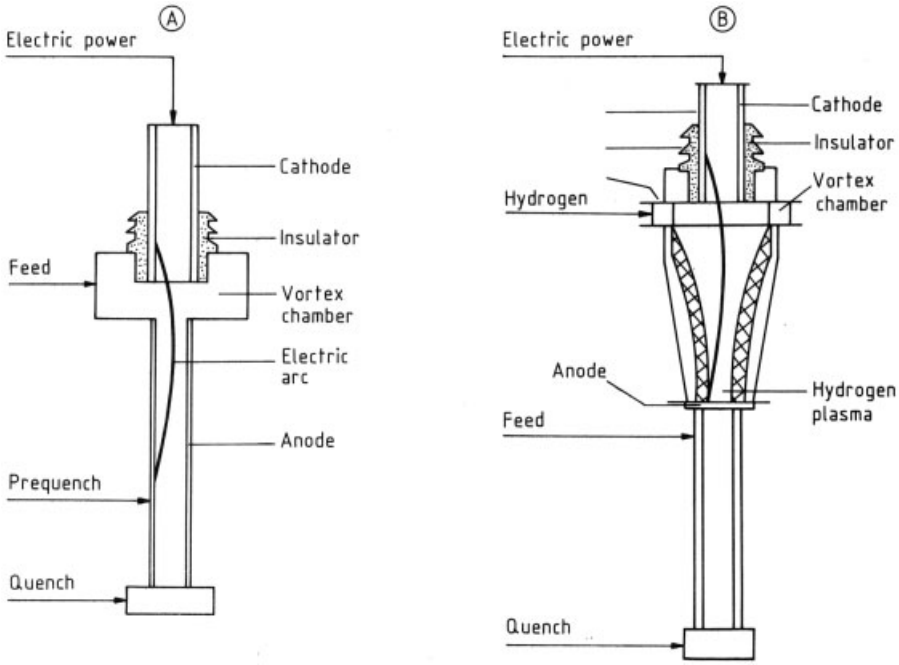

Figure 17. Hüls electric-arc furnaces for gaseous, liquid, and solid feed
A) One-step process; B) Two-step process

rapidly decomposes into soot and hydrogen at these temperatures, the gases must be quenched immediately with water to about 200 °C, i.e., a quench rate of 10^6 °C/s must be achieved.

The *specific energy requirement* (SER) and the acetylene yield depend on the geometry and dimensions of the furnace and electrodes, the velocity distribution of the gas, and the kind of hydrocarbon to be cracked. Once the furnace has been designed, only the hydrocarbons can be varied.

Process Without Prequench. Figure 18 shows acetylene and ethylene yields and the specific energy requirement (SER) of various saturated hydrocarbons under constant conditions without prequench. Methane shows the highest SER and acetylene yield, but the lowest ethylene yield. As the chain length is increased, both acetylene yield and SER decline, corresponding to the declining heat of acetylene formation from the various hydrocarbons.

Normally, pure hydrocarbons are not available. The results obtained from mixtures of hydrocarbons can be expressed as a function of the carbon number, which is the number of moles of carbon atoms bound in hydrocarbons per mole of the gaseous mixture. Figure 19 shows specific amounts of acetylene, ethylene, and hydrogen formed and of hydrocarbon consumed as a function of carbon number. This function enables the Hüls process to be optimized within certain limits, for example, for hydrogen output in relation to acetylene production.

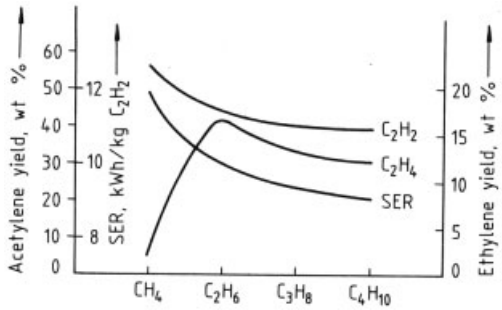

Figure 18. Acetylene yield, ethylene yield, and energy consumption for various hydrocarbons in the Hüls arc process

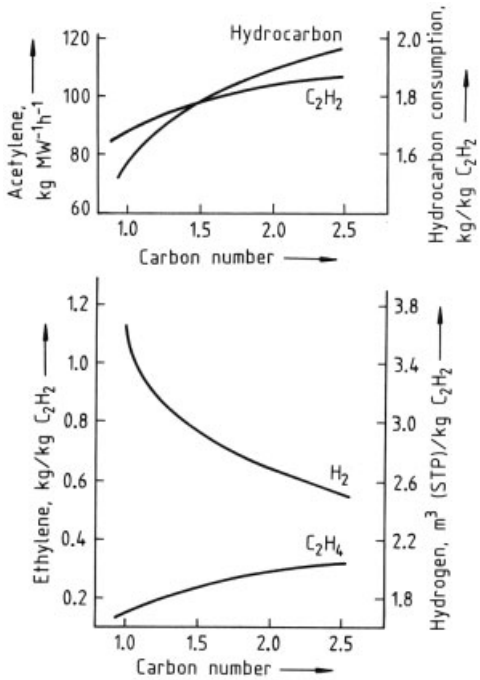

Figure 19. Specific values for acetylene and hydrogen formation

Process with Prequench. Cracking in the prequench section is essentially an ultrasevere steam cracking process. The kind and amount of hydrocarbons used for the prequench can be varied. Figure 20 shows the specific product yield for different prequench rates for feeding methane to the arc furnace and propane to the prequench. Acetylene and hydrogen yields are unaffected, whereas ethylene shows a slight maximum and declines when the temperature is not sufficient at a given residence time. Propene shows a steady increase, and the $C_3 : C_2$ ratio is below 0.25. The relative ethylene yield from various hydrocarbons is as follows: ethane 100, propane 75, *n*-butane 72, isobutane 24, 1-butene 53.

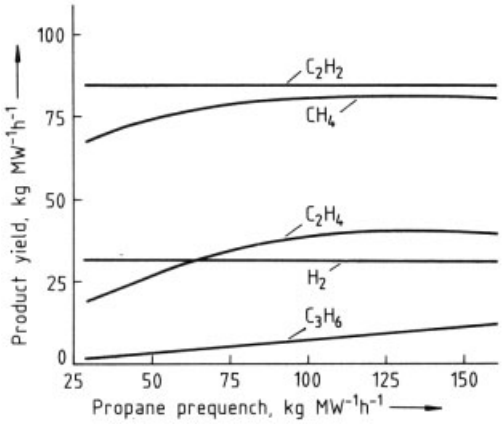

Figure 20. Specific product yield for different prequench rates

Oil Quench. Because the gas temperature of the furnace gas after prequench is on the order of

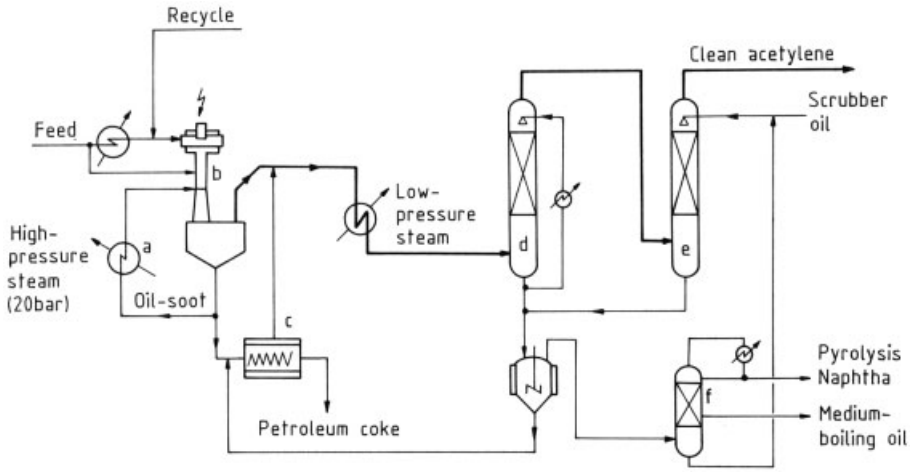

Figure 21. Process with oil quench system
a) Heat recovery; b) Arc furnace; c) Oil recovery; d) Separation of medium-boiling compounds; e) Separation of low-boiling compounds; f) Oil regeneration

1200 °C, an oil quench system has been developed to regain about 80 % of the sensible heat content of the furnace gas as steam by heat exchange. The soot–oil mixture formed can be upgraded to a sulfur- and ash-free high-grade petroleum coke. Figure 21 shows the Hüls system with oil quench.

The Purification System. The process of purification depends on the type of the quench system. In the case of water quenching, 80 % of the carbon black is removed by cyclones as dry carbon black, the remaining 20 % as soot in water-operated spray towers. In a combined oil–water scrubbing system, aromatic compounds are removed and benzene, toluene, and xylene (BTX) are recovered in a distillation process.

Figure 22 shows the principle separation and purification steps for the furnace gas. The gas leaves the first three purification sections with a carbon black content of 3 mg/m^3 and is compressed by four-stage reciprocating compressors to 19 bar (1.9 MPa). The gas is washed in towers with water in a countercurrent flow. At the bottom of the tower, the water is saturated with acetylene, whereas the overhead gas contains less than 0.05 vol % acetylene. The acetylene–water solution is decompressed in four stages. Gas from the first decompression stage returns to the compressor to improve selectivity. The last two stages operate at 0.2 and 0.05 bar (20 and 5 kPa). The gas still contains about 10 vol % of higher acetylenes, which are removed by a cryogenic process. The higher acetylenes are liquefied, diluted with flux oil, stripped, and returned to the arc furnace together with spent hydrocarbon. A more selective solvent such as *N*-methylpyrrolidone or dimethylformamide is preferred to the water wash.

Linde and Hüls have designed an appropriate purification system. Hydrogen and ethylene are separated by well-known technology, such as the cryogenic process or pressure-swing adsorption (see → Adsorption).

Process Data. Hüls operates its plant with a mixture of natural gas, refinery gas, and liquefied petroleum gas. The carbon number varies between one and two. Table 8 shows a typical analysis of feed gas and cracked gas.

Table 8. Typical analysis of feed and cracked gas

	Feed gas, vol %	Cracked gas, vol %
C_2H_2	0.4	15.5
C_3H_4	1.4	0.4
C_4H_2	1.2	0.3
C_4H_4	1.7	0.4
C_2H_4	0.8	6.9
C_3H_6	3.6	1.0
Allene	0.4	0.2
C_4H_8	1.0	0.2
C_4H_6	0.9	0.2
C_5H_6	0.6	0.2
C_6H_6	0.5	0.5
CH_4	64.6	13.8
C_2H_6	7.5	0.4
C_3H_8	3.6	0.3
C_4H_{10}	4.6	1.0
C_5H_{12}	0.5	0.1
H_2	4.5	57.6
CO	0.5	0.6
O_2	0.1	0.0
N_2	1.6	0.4

The Hüls plant has 19 arc furnaces, the number operated depending on the electricity supply. The arc furnaces can be started up and shut down immediately. Two large gas holders provide a storage volume of 350000 m^3 so that the purification section operates on permanent load and there is a dependable supply of products, even if the arc furnace section is operated at higher or lower load.

The plant has an annual capacity of 120000 t acetylene, 50000 t ethylene, 400×10^6 m^3 (STP) hydrogen, 54000 t carbon black and soot, and 9600 t aromatic compounds. The energy consumption is 1.5×10^6 MW h/a.

Specific data for consumption of hydrocarbons and energy and the production of byproducts per ton of acetylene produced are as follows:

Hydrocarbons to the arc furnace	1.8 t
Hydrocarbons for prequench	0.7 t
Energy for the arc furnace	9800 kW h
Energy for gas purification	2500 kW h
Ethylene	0.42 t
Hydrogen	3300 m^3 (STP)
Carbon black and soot	0.45 t
Aromatics	0.08 t
Residue	0.12 t
Heating gas	0.12 t

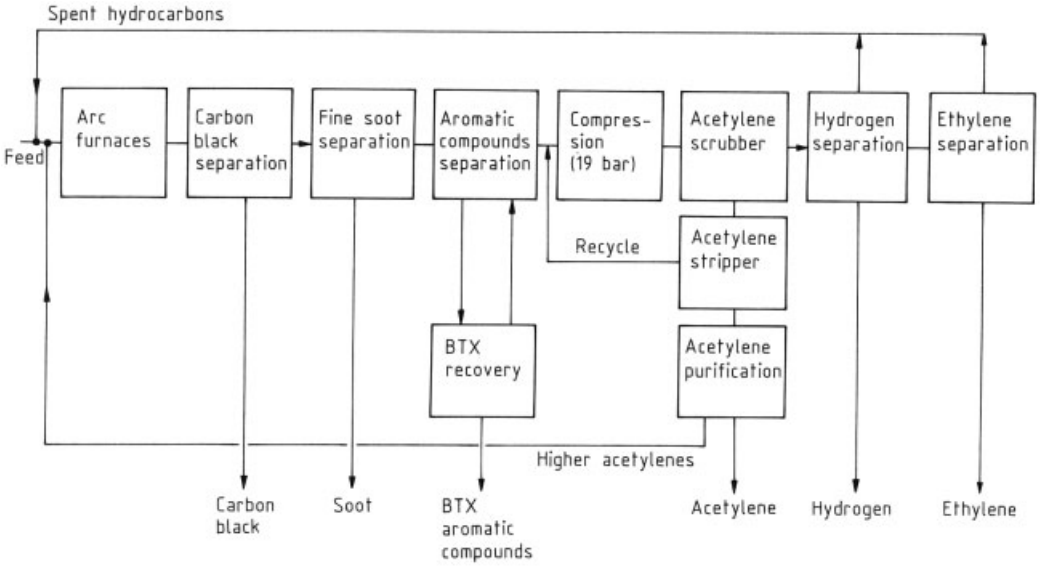

Figure 22. Principal separation and purification steps for the furnace gas of the Hüls arc process

4.3.2. Production from Liquid Hydrocarbons (Plasma Arc Process)

Two different plasma furnaces, each with the appropriate reactor for the cracking of liquid hydrocarbons, were developed by Hoechst and Chemische Werke Hüls in close cooperation. Both units were tested on an industrial scale at a power level of 8 – 10 MW [55]. However, neither process has actually come into use for acetylene production on account of the economics.

The scheme of the plasma generator used by Hüls is shown in Figure 17 B. The unit consists of three parts: the arc furnace, the reactor, and the quench system. The arc burns over a length of 1.6 m at 7 kV d.c. and 1.2 kA, resulting in a power input of 8.5 MW. It is stabilized by hydrogen injected tangentially through the vortex chamber. The thermal efficiency of the furnace is ca. 88 % of the electrical power input. The hydrogen plasma jet passing through the anode nozzle has an energy density of 3.5 kW h/m^3(STP), corresponding to an average temperature of 3500 K. The liquid hydrocarbons (e.g., crude oil) to be cracked are injected into the cylindrical reactor to achieve good mixing with the plasma jet and to avoid the formation of carbonaceous deposits on the wall. Within several milliseconds the hydrocarbons are heated and cracked to acetylene, ethylene, hydrogen, soot, and other byproducts before the mixture is quenched with oil to 300 °C. The acetylene ratio can be adjusted by varying the residence time. By operation of an oil quench with the high-boiling residue of the crude oil, 80 % of the sensible heat content of the cracked gas can be recovered as steam. The soot is taken up by the quench oil and is removed from the system as an oil–soot dispersion having 20 % soot concentration. The unconverted vaporized fractions of the oil are condensed in oil scrubbers at a lower temperature, simultaneously cleaning the gas of the aromatic components and fine soot. These oil fractions are recycled to the reactor and the quench system, respectively.

Tests were carried out with a variety of hydrocarbons from propane to naphtha, but mainly with crude oil and residue oils. The cracking results depend on the chemical nature of the feed. Consumption figures and yields for various feedstocks are given in [55]. For high-boiling petroleum fractions, the acetylene and ethylene yields increase with the content of low-boiling components in the feed (see Fig. 23). Consumption and byproduct yields per ton of acetylene for a Libyan crude oil are summarized below:

Consumption	
Crude oil consumed	3.5 t
Power (d.c.)	10500 kW h
Byproducts	
Ethylene	0.46 t
Hydrogen (99.9 %)	1100 m^3 (STP)
Fuel gas	0.74 t
Soot – oil mixture (20 % soot)	1.2 t

Hoechst used a high-intensity three-phase a.c. arc furnace at 1.4 kV and 4.2 kA, giving a power input of 10 MW [55]. The thermal efficiency was 90 %. Because of the high amperage the graphite electrodes had to be replenished continually. The generator was lined with graphite. Different reactor designs for ethylene: acetylene ratios of 0.5 and 1.0 were developed by varying the mixing intensity of the hydrogen plasma jet with the liquid hydrocarbon. The tests were carried out with naphtha feed (see Table 9). The cracked gas was quenched with residue oil, in a manner similar to that described in the Hüls process.

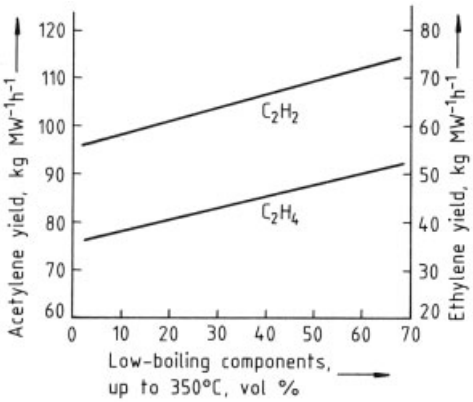

Figure 23. Acetylene and ethylene yields as a function of the low-boiling components

The acetylene concentration in the Hüls process and the Hoechst process was ca. 14 vol % so that in principle the same acetylene separation process can be used as described above for the arc process.

4.3.3. Production from Coal (Arc Coal Process)

Numerous laboratory tests for the conversion of coal to acetylene using the arc or plasma processes have been carried out since the early 1960s [56]. The results can be summarized as follows:

- Acetylene yields up to 30 % can be obtained.
- Because of the rapid heating of the coal in the plasma jet, a higher total gas yield can be achieved than is indicated by the volatiles of the coal measured under standard conditions.
- Hydrogen (instead of argon) plasma gas considerably increases the acetylene yield.

The AVCO arc furnace (Fig. 24) consists of a water-cooled tungsten-tip cathode and a water-cooled anode [57]. The arc is stabilized by a magnetic field surrounding the anode, forcing the anode striking point of the arc to rotate rapidly and so avoiding burnthrough. The dried and finely ground coal is injected by means of a hydrogen gas flow around the cathode. Additional gas without coal is introduced around the cathode and at the anode as a sheath. On passing the arc zone the coal particles are heated up rapidly. The volatiles are released and are cracked to acetylene and byproducts, leaving a residue of fine coke particles covered with soot. After a residence time of some milliseconds the gas–coke mixture is quenched rapidly with water or gases. The use of a prequench system similar to that of the Hüls arc process was also tested. The system pressure can be varied between 0.2 and 1.0 bar (20 and 100 kPa).

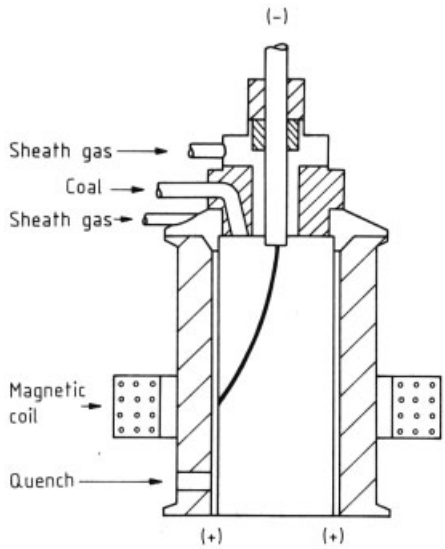

Figure 24. Principal scheme of the AVCO plasma furnace for the pyrolysis of coal

Table 9. Consumption and yield per ton of acetylene for the Hoechst arc process and naphtha feed for different reactor designs

	Low ethylene yield	High ethylene yield
Consumption		
Naphtha	1.92 t	2.50 t
Quench oil	0.53 t	0.63 t
Energy(2-phase a.c.)	9300 kW h	10500 kW h
Byproducts		
Ethylene	0.5 t	0.95 t
Hydrogen	1450 m^3 (STP)	1500 m^3 (STP)
Soot–oilmixture(20 % soot)	0.75 t	1.00 t

Hüls' pilot plant uses the same plasma furnace as for the crude oil cracking, but with 500 kW of power [58]. The dried and ground coal is injected into the plasma jet, and the coal is cracked to acetylene and byproducts in the reactor. The reactor effluent can be prequenched with hydrocarbons for ethylene production or is directly quenched with water or oil. Char and higher boiling components are separated by cyclones and scrubbers, respectively. The problem in the reactor design is to achieve thorough and rapid mixing of the coal with the plasma jet and to avoid forming carbonaceous deposits on the wall. Smaller amounts of deposits can be removed by periodic wash cycles with water. Operation times of 2.5 h by AVCO and 5 h by Hüls have been reported.

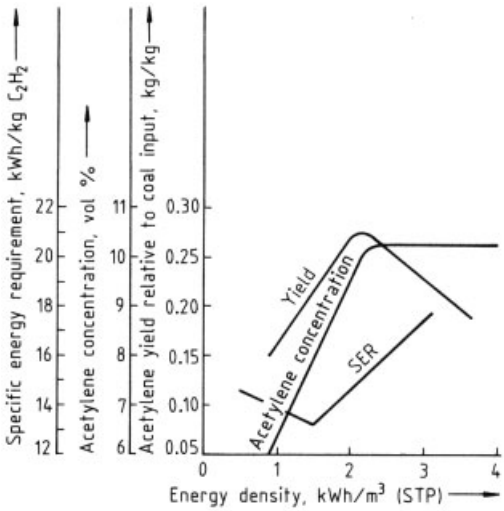

Figure 25. Effect of the energy density of the plasma on the cracking of coal (AVCO)

Experiments published by Hüls and AVCO show that at the optimal residence time the energy density of the plasma jet, the specific power, and the pressure all greatly affect the acetylene yield (Fig. 25 and Fig. 26). Other parameters affecting the yield are the amounts of volatiles in the coal and the particle size. The lowest figures for the specific energy consumption published by AVCO are of the order of 27 – 37 % based on water-free coal.

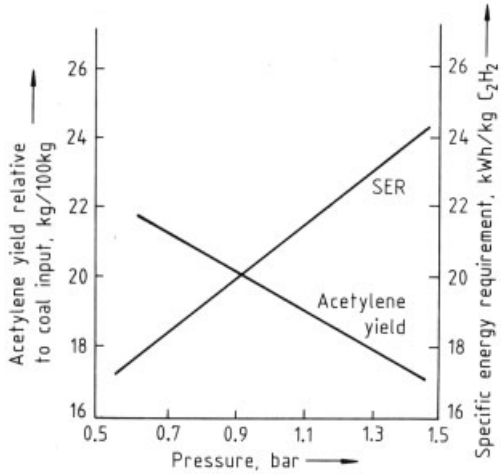

Figure 26. Acetylene yield and specific energy requirement as a function of pressure (Hüls)

In addition to acetylene, the exit gas contains considerable amounts of CO, depending on the oxygen content of the coal. Because nitrogen and sulfur are present in the coal, other byproducts are HCN, CS_2, COS, and mercaptans. The gas separation system is therefore designed accordingly [59]. Depending on the hydrogen content of the coal, the process is either self-sufficient in hydrogen or has a slight surplus. The total gas yield of the coal based on a volatile content in the coal of 33 % is up to 50 %. Thus 50 % of the coal remains as char. Tests with a view to using this char in the rubber industry have been unsuc-

cessful so far. Thus the char can be used only for gasification or as a fuel.

In all the processes under development, the production of ethylene from coal requires several process steps (Fig. 27), resulting in a high capital demand for a production plant. In contrast, the acetylene production from coal arc pyrolysis is straightforward, leading to lower investment costs. Demonstration units on a higher power level are therefore scheduled by both AVCO and Hüls.

4.3.4. Production from Calcium Carbide

At present, the generation of acetylene from calcium carbide (→ Calcium Carbide) is of primary importance for welding and for the production of carbon for batteries. The particular raw-material situation and the use of special processes are two common reasons for continuing to use acetylene generated from carbide in the chemical industry.

The reaction of calcium carbide and water to form acetylene and calcium hydroxide is highly exothermic:

$$CaC_2 + 2\ H_2O \rightarrow C_2H_2 + Ca(OH)_2 \quad \Delta\ H = -129\ \mathrm{kJ/mol}$$

The acetylene generator used for commercial production must therefore be designed to allow dissipation of the heat of reaction. In the event of inadequate heat dissipation, for example, when gasification proceeds with insufficient water, the carbide may become red-hot. Under certain circumstances (including increasing pressure), this may cause the thermodynamically unstable acetylene to decompose into carbon and hydrogen. (For safety precautions see Chap. 5.) Carbide for the production of acetylene is used in the following grain sizes (mm): 2 – 4, 4 – 7, 7 – 15, 15 – 25, 25 – 50, 50 – 80. This classification is virtually identical in most countries: DIN 53922 (Germany); BS 642:1965 (United Kingdom); JIS K 1901 – 1978 (Japan); Federal Specification 0–6–101 b/GEN CHG NOT 3 (United States). In addition, grain size 0 – 3 is used for the dry generation of acetylene.

Pure calcium carbide has a yield number of 372.66. This means that the gasification of 1 kg of carbide yields 372.66 L acetylene at 15 °C and 1013 mbar (101.3 kPa). Commercially available carbide has a yield number of 260 – 300.

A distinction is made between two groups of acetylene generators (with continuous rates of production greater than 10 m^3 acetylene per hour): the wet type and the dry type.

In *wet generators*, the acetylene is converted with a large water excess. In most cases, a lime slurry containing 10 – 20 wt % calcium hydroxide is obtained. The heat of reaction increases the temperature of the generator water and is removed from the reactor with the lime slurry.

In *dry generators*, the water mixed with the carbide is just sufficient for chemical reaction and for dissipating the heat of reaction. The calcium hydroxide is obtained in the form of a dry, easily pourable powder having a residual moisture content of 1 – 6 %. The heat of reaction is dissipated by evaporation of part of the generator water.

Generators are classified according to their working pressure as either low or medium pressure. This classification is governed by the regulations concerning acetylene plants and calcium carbide storage facilities (*Acetylenverordnung*) [60] issued by Deutscher Acetylenausschuß (German acetylene committee) and the associated technical rules for acetylene plants and calcium carbide storage facilities *TRAC* (*Technische Regeln für Acetylenanlagen und Calciumcarbidlager*) [61]. The *Acetylenverordnung* and *TRAC* constitute a comprehensive set of rules for handling acetylene. Recommendations in other countries (e.g., United States) deal only with some aspects, such as safety precautions [62].

Low-pressure generators are designed for a maximum allowable working pressure of 0.2 bar. They must be rated for an internal pressure of at least 1 bar. Lower pressure ratings are possible if proof is given in each particular case that the generator can withstand the expected stress (maximum working pressure, water filling, agitator, etc.; *TRAC 201*).

Medium-pressure generators have a maximum allowable working pressure of 1.5 bar. They must be rated for an internal pressure of 24 bar. A design pressure of 5 bar suffices whenever the generators are equipped with rupture disks of a defined size and specified response pressure (3 – 4.5 bar, *TRAC 201*).

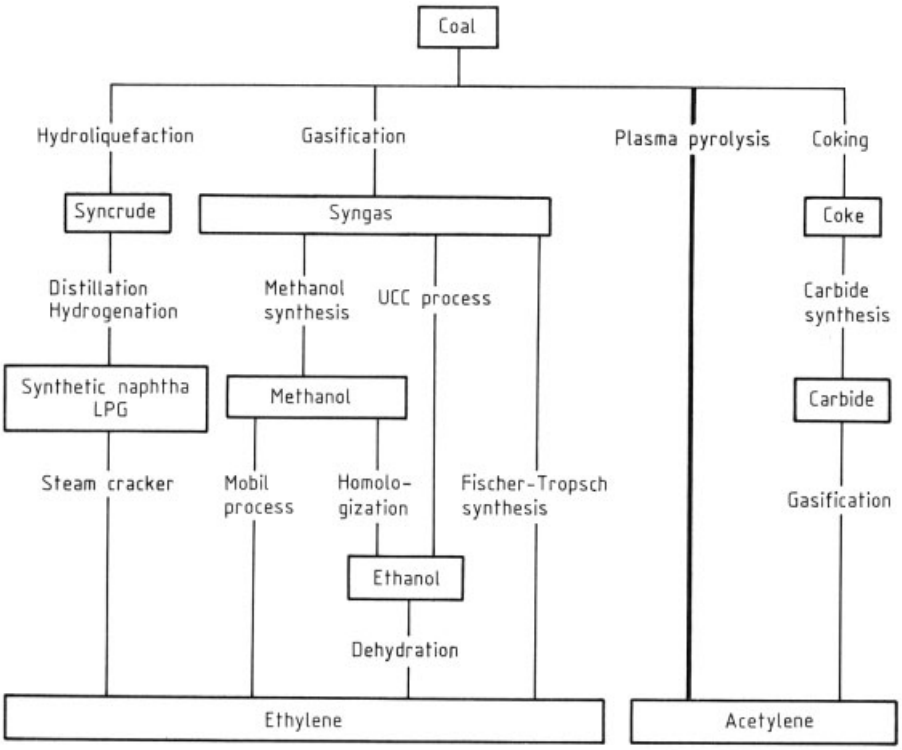

Figure 27. Alternative routes from coal to ethylene and acetylene

4.3.4.1. Wet Generators

Wet generators are used primarily for the production of small amounts of acetylene, e.g., for welding purposes. Wet generators work by one of three different principles [63]:

1) The *carbide-to-water principle,* where the carbide is mixed with a large excess of water at a rate corresponding to the gas withdrawal rate. Most generators work by this principle.
2) The *water-to-carbide principle* (drawer type generators), where water is added at a controlled rate to the carbide, which is held in a replaceable container (drawer).
3) The *contact principle* (basket generators), where the carbide, which is held in a basket, is immersed into the generator water. This type is designed so that the water drifts off the basket as a result of the gas pressure at low gas withdrawal rates and, conversely, returns to the basket when gas withdrawal rates increase.

Medium-Pressure Generators. The Messer Griesheim MF 1009 is a typical carbide-to-water generator (Fig. 28).

The carbide skip (a) is filled with carbide of the 4 – 7 grain size. The skip is connected by gas-tight gates to the hopper (b) and is purged of air by nitrogen or acetylene. The carbide drops into the hopper (b) and is fed continuously by the feeding system (c) to the gasification chamber (d). The gasification chamber contains water up to a level defined by the generator capacity and is equipped with an agitator (e) for whirling the lime slurry. The heat evolved in gasification heats the generator water. For continuous operation the water temperature must not exceed 90 °C; therefore, fresh water is admitted continuously to the gasification chamber. If the defined water level is exceeded, the slurry valve (f), controlled by a float, opens, allowing the excess water and the lime slurry to be discharged from the generator.

The acetylene generated collects above the water and is withdrawn. The feeding system (c) is controlled by the gas pressure, i.e., the rate at

which carbide drops into the gasification chamber varies directly with the rate of gas withdrawal.

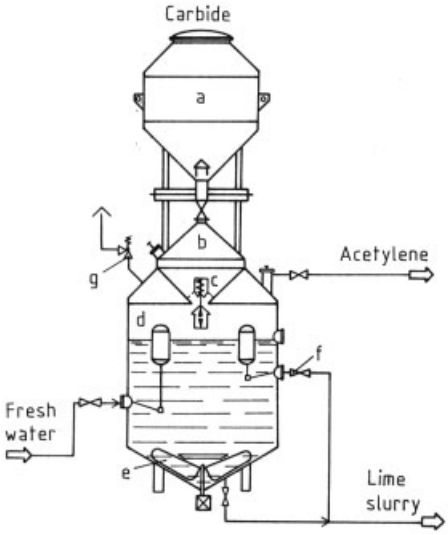

Figure 28. Medium-pressure wet generator
a) Carbide skip; b) Hopper; c) Feeding system; d) Gasification chamber; e) Agitator; f) Slurry valve, g) Safety device

The carbide stock in the hopper (b) is sufficient for about one hour, but the skip (a) can be refilled with carbide and replaced on the hopper so that continuous operation is possible. The wet generator described has a continuous hourly output of 75 m^3 of acetylene. The skip holds 1000 kg of carbide.

Low-Pressure Generators. The working principle of the low-pressure carbide-to-water generators is very similar to that of the medium-pressure carbide-to-water generator described above. In most cases, a downstream acetylene holder, normally of the floating gas bell type, is provided. In contrast to the medium-pressure generator, in which the carbide feed rate is controlled by the acetylene gas pressure, the feed rate in the low-pressure generator is controlled by the position of the bell in the acetylene holder, i.e., by the gas quantity.

Products. The acetylene generated in the wet generator can be used for welding, often without further purification. In certain cases, coke- or gravel-filled purifiers or a wet scrubber are connected downstream from the generator for separating solid or liquid particles. Before it is fed to a synthesis unit, the acetylene must be purified chemically (see Section 4.3.4.3).

The lime slurry formed is fed into pits. Here, the calcium hydroxide settles in the form of a lime dough containing 35 – 75 wt % water (wet lime, carbide lime dough). This dough is used as carbide lime.

4.3.4.2. Dry Generators

Dry generators are mainly used for the production of large quantities of acetylene for chemical synthesis.

Compared to the wet generator, the primary advantage of the dry generator is that the dry calcium hydroxide formed as a byproduct can be used in other processes more easily, more cheaply, and in a more diversified way than the lime slurry obtained in the wet generator [64]. Moreover, lime recycling into the carbide production process is only possible with dry calcium hydroxide.

A high gasification rate and the elimination of the risk of overheating were originally the most important criteria for the design of dry generators. Early designs of dry generators worked by continuous renewal of the reaction surfaces of coarse carbide with the aid of rotating drums, blades, vibrating screens, and similar equipment.

Typical examples are the early generator of Shawinigan Chemicals [65] and the Piesteritz dry generator [66].

Although a number of factors affect the gasification rate of carbide, e.g., density, porosity, and crystalline structure, above all it is the specific surface that affects the carbide gasification rate the most. Hence, dry generators work with finely ground carbide (0 – 3 mm), which gasifies in a fraction of the time needed for coarse carbide. The result is a high space–time yield.

A typical application of this principle is the large-scale *Knapsack dry generator*, which was developed at the Knapsack works of Hoechst. This type of generator is used worldwide and is described in more detail below (Fig. 29).

Carbide of the grain size 0 – 3 falls from the chain conveyor (a) into the subdivided feed bin (b). The chain conveyor is loaded with material from the carbide bin. Because of the recirculating stream of carbide the feed bin (b) is full at

all times. The carbide layer in the feed bin (b) acts as a gas seal between the generator and the carbide conveying system.

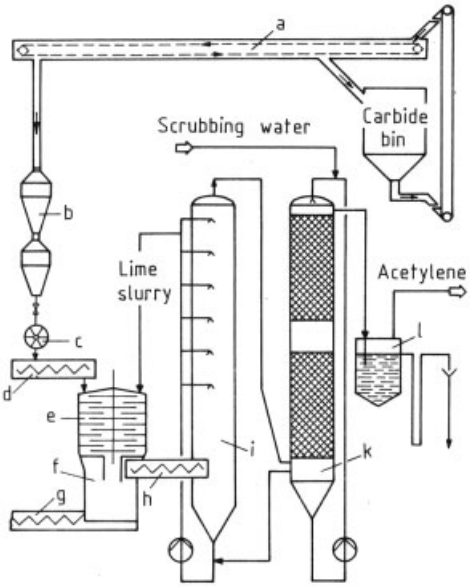

Figure 29. Knapsack dry generator
a) Chain conveyor; b) Feed bin; c) Star wheel; d) Carbide feed screw; e) Generator; f) Lime lockhopper; g) Lime discharge screw; h) Lime scraper; i) First scrubbing tower; k) Second scrubbing tower; l) Dip seal

The carbide is fed to the generator (e) via the star wheel (c) and the carbide feed screw (d). The largest generator of this kind built to date has a diameter of 3.5 m and an overall height of approximately 8.0 m. The generator has up to 13 circular trays. These are so designed as to leave alternate annular gaps on the shell side and at the central agitator shaft. The agitator shaft moves stirrer paddles across the trays.

The carbide first reaches the uppermost tray where the generator water is also admitted. The reaction mixture consisting of carbide, water, and calcium hydroxide is pushed by the stirrer paddles towards the outer edge, drops on to the second tray, returns towards the center, etc. When it reaches the last tray, the carbide has been fully gasified. The calcium hydroxide, which still contains up to 6 % water, drops into the lime lockhopper (f). Here, a lime layer two meters deep serves as the gas closure between generator and lime conveying system. The lime is withdrawn continuously.

The gas leaving the generator through the lime scraper consists of 25 % acetylene and ca. 75 % water vapor. The water vapor is the result of dissipating the major portion of the reaction heat. Depending on the generator load, up to several hundred kilograms of lime hydrate dust are carried along with the acetylene. The lime scraper (h) retains the major portion of this dust and returns it to the generator. The remainder is sent together with the gas into the first scrubbing tower (i). Here, lime slurry is sprayed into the hot acetylene gas (ca. 90 °C) to scrub out the lime dust; part of the water vapor condenses because of the simultaneous cooling. In the second scrubbing tower (k), the acetylene is sprayed with atomized water to cool the gas below 40 °C; additional water vapor condenses here. Any ammonia still present in the gas is also removed.

The acetylene leaves the generator via the dip seal (l). It still contains certain impurities in the form of sulfur and phosphorus compounds.

The Knapsack dry generator is suitable for a carbide throughput of 15 t/h, corresponding to an acetylene quantity of 3750 m^3/h. During this process, about 17.5 t of calcium hydroxide per hour are obtained. The pressure in this low-pressure generator amounts to approximately 1.15 bar (115 kPa).

The dry generator of Shawinigan Chemicals, Montreal [67], also processes finely ground carbide and has a variety of applications. It consists of several superimposed troughs. Carbide and water are fed into the uppermost trough. The reacting mixture, which is constantly kept in motion by blades, flows over a weir onto the trough below, etc. At the uppermost trough, water is admitted at such a rate that carbide-free calcium hydroxide can be withdrawn at the lowermost trough. The generated acetylene is purified in two scrubbing towers and cooled.

The calcium hydroxide formed (carbide lime) has a wide range of applications, e.g., in the building industry (for preparing mortar, cement, etc.), in the chemical industry (for neutralization and for recycling to the carbide furnace), in agriculture (as fertilizer), and for water purification and waste water treatment [64].

4.3.4.3. Acetylene Purification

During the gasification of carbide with water, gaseous compounds become mixed with the acetylene, and these must be removed because they have a harmful effect on the downstream chemical synthesis processes. The impurities are mainly sulfur and phosphorus compounds. They can be removed by one of the following purification processes.

In the first process, *dilute chlorinated water* is used as the oxidizing agent. The chlorine concentration of the water is limited to 1.5 g/L to prevent the formation of unstable chlorine compounds, which present an explosion hazard. The chlorine scrubbing step is followed by a caustic soda scrubber to remove the hydrogen chloride formed during the oxidation process. The disadvantage of this purification process is that considerable quantities of scrubbing water are produced.

The second process uses 98 % *sulfuric acid* as the oxidizing agent [68]. Because very small quantities of sulfuric acid are admitted, it is difficult to dissipate the heats of absorption and reaction. Heating the acetylene results in increased formation and settling of polymerization products in the purification stage. For this reason the gas requires additional cooling in the event that the acetylene contains appreciable quantities of impurities. Moreover, it is recommended that a second scrubbing tower be kept on standby if a high onstream factor is desired (e.g., 91 % $\triangleq$ 8000 h/a operating time).

The sulfuric acid scrubber is followed by a caustic soda scrubber, in which the sulfur dioxide formed during oxidation is removed.

The main advantage of this purification method is that virtually no waste water is obtained. The small amount of polluted, highly concentrated sulfuric acid can be used, for example, in fertilizer plants.

These two purification processes yield the following acetylene purities (by volume):

Acetylene	> 99.5 %
Sulfur, as H_2S	< 10 ppm
Phosphorus, as H_3P	< 10 ppm

As a result of the extremely good sorption properties of the concentrated sulfuric acid, very pure acetylene can be expected.

4.4. Other Cracking Processes

4.4.1. Thermal Cracking By Heat Carriers

Well-known processes using heat carriers, such as the Wulff and Hoechst high-temperature pyrolysis (HTP) processes, are no longer used because they require refined petrochemical feedstocks such as naphtha and liquid petroleum gas. The Wulff process uses refractory material as the heat carrier, whereas the Hoechst HTP process uses hot combustion gases.

Newer processes, which are able to convert crude and heavy distillates into olefins and considerable amounts of acetylene, are still in the pilot-plant stage. These processes include the advanced cracking reactor process developed by Kureha, Chiyoda, and Union Carbide, using high-temperature superheated steam, and Dow's partial combustion cracking process, using hot combustion gases produced from oxygen and fuel oil as the heat carrier.

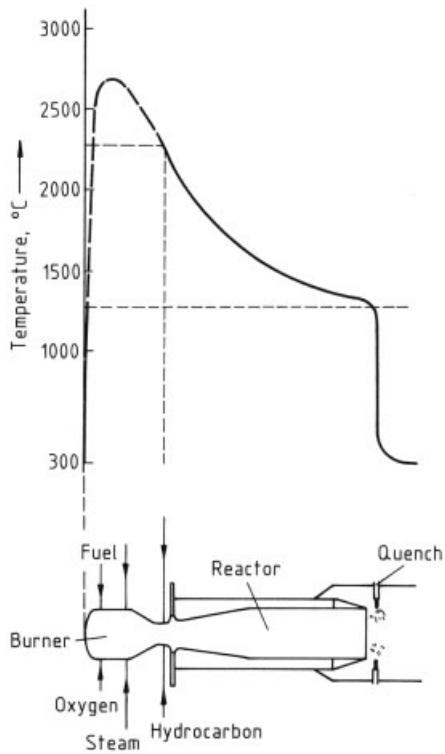

Figure 30. High-temperature pyrolysis (HTP) process

Wulff Process [69 70, p. 58]. This process is based on indirect heat transfer, an approach fundamentally different from the partial-oxidation and electric-arc processes. The hydrocarbon feed is cracked in refractory ovens previously heated by combustion gas. After cracking, the products are quenched outside the reactor. Soot formation is a serious problem because the feed cannot be heated as rapidly as in the partial-oxidation or arc processes. This problem can be diminished by using a feed with a higher hydrogen:carbon ratio. However, methane is not suitable because of the high temperature and high heat of reaction required, resulting in a low conversion rate. Thus the best feed for the Wulff process is ethane or propane.

Hoechst High-Temperature Pyrolysis (HTP) Process (Fig. 30) [1 70, p. 55, 71]. This is a two-stage process. In the first stage, heat is produced in the burner by the combustion of residual cracked gas from the acetylene recovery section (CO, H_2, CH_4) with oxygen. Immediately after combustion, the temperature is about 2700 °C; this is moderated to about 2300 °C by the injection of steam before the reactor is entered. In the second stage, the feedstock naphtha is injected, and the adiabatic cracking reaction takes place. A final temperature of about 1300 °C is reached: This determines the cracked gas composition. By varying the feed rate of naphtha the acetylene–ethylene ratio can be altered from 30 : 70 to 70 : 30. However, thermodynamic and economic considerations show that the optimum ratio is 40 : 60.

After a reaction time of a few milliseconds the cracked gas is quenched to approximately 250 °C by the injection of cracked oil from the process. The oil absorbs heat from the cracked gas and is passed through waste heat boilers, raising the steam pressure. No soot is formed in this process, even when crude oil is used as feed because of the high steam content of the carrier gas.

After the oil crisis of 1973 the process became uneconomical in spite of its high thermal efficiency, and in 1976 it was shut down after 15 years of operation. However, one unit is still running in Czechoslovakia.

Kureha, Chiyoda, Union Carbide Advanced Cracking Reactor (ACR) Process [72]. To avoid dependence on oil refineries or gas processors for the supply of feedstocks, processes for directly cracking crude oil have been developed by various companies for the production of olefins. Some of these processes operate at reaction temperatures intermediate between those of the usual crack processes for olefins and those for acetylene. The ACR process (Fig. 31) uses a multi-port burner to produce a heat carrier

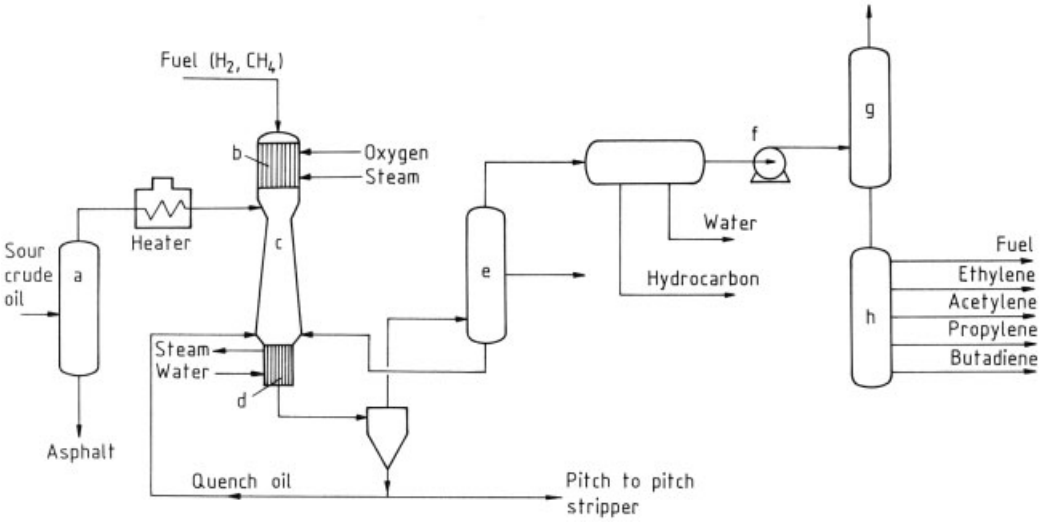

Figure 31. Advanced cracking reactor process (ACR)
a) Crude distillation column; b) Burner; c) Advanced cracking reactor; d) Ozaki quench cooler; e) Oil gasoline fractionator; f) Compressor; g) Acid gas removal column; h) Gas separator

gas of 2000 °C by the combustion of $H_2 - CH_4$ mixtures with oxygen in the presence of steam preheated to 800 °C.

The oil to be cracked is introduced through nozzles into the stream of carrier gas and passes into an advanced cracking reactor, where the reaction takes place adiabatically at 5 bar (0.5 MPa). The initial temperature is 1600 °C; the final temperature at the exit of the reactor is 700 – 900 °C after a residence time of 10 – 30 ms. The cracked gas is quenched by oil in an Ozaki quench cooler (Fig. 32), where steam production up to 120 bar (12 MPa) is possible. This particular boiler design was developed for a high heat transfer rate without coke formation on the exchanger surfaces. Yields reported for Arabian light crude oil are 11.2 wt % hydrogen and methane, 40.7 wt % olefins, and 4.2 wt % acetylene. The acetylene yield is about ten times higher than in usual olefin processes.

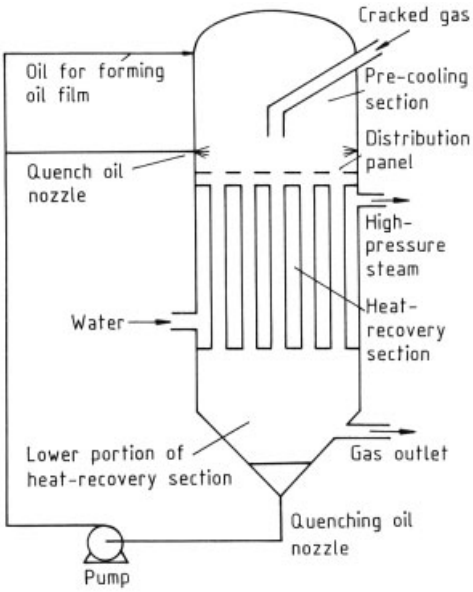

Figure 32. Ozaki quench cooler

Dow Partial Combustion Cracking (PCC) Process [72]. The basic idea of this process is to reduce coking and soot formation considerably when heavy feeds are cracked and when hydrogen is present in the reaction mixture. The PCC process (Fig. 33), which accepts crude oil and heavy residue as feedstock, attains a high partial pressure of hydrogen in the reaction zone by recycling the quench oil (produced in the process) to the burner where it is partially oxidized to yield synthesis gas. Thus there is no need to find a use for the quench oil as in the case of the ACR process. Starting from residual oil boiling above 343 °C, yields are given as 12.4 wt % methane, about 38 wt % alkenes, and 2.5 wt % acetylene. This is seven to eight times more acetylene than that obtained from a steam cracker, but less than the acetylene yield of the ACR process, because of a residence time in the reaction zone which is three to ten times longer.

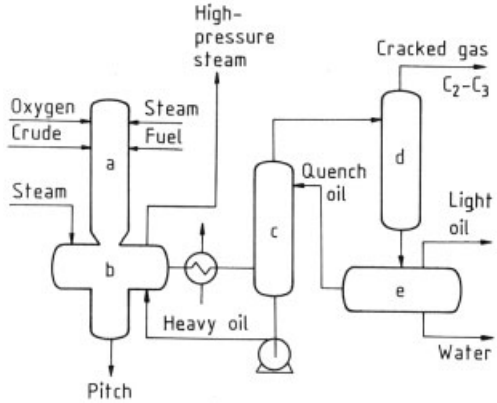

Figure 33. Dow partial combustion cracking process (PCC)
a) Reactor; b) Quench boiler; c) Quench column; d) Stripper; e) Decanter

4.4.2. Acetylene as a Byproduct of Steam Cracking

In a steam cracker saturated hydrocarbons are converted to olefinic products such as ethylene and propene. Besides these desired components, acetylene and many other products are formed in the cracking process (Fig. 36). The concentration of acetylene depends on the type of feed, the residence time, and temperature (cracking severity: expressed as conversion or propene/ethylene ratio P/E). Typical data are given in Table 10. The acetylene concentration in the off-gas from the furnace varies between 0.25 and 1.35 wt %. In certain cases of propane and butane cracking the the raw gas can contain up to 2.1 wt % acetylene with corresponding very low amounts of propyne and propadiene. The corresponding content of acetylene in the C_2 frac-

Table 10. Yields of unsaturated components (wt %) in raw gas from steam cracking

Feedstock	Cracking severity	Acetylene	Propyne	Propadiene
Ethane	65 % convers.	0.4 – 0.50	0.04	0.02
LPG	90 % convers.	0.65 – 1.35	0.63	0.35
Full-range naphtha	P/E: 0.4	0.9 – 1.05	0.81	0.54
Full-range naphtha	P/E: 0.53	0.5 – 0.70	0.68	0.50
Full-range naphtha	P/E: 0.65	0.25 – 0.42	0.46	0.38
Atmospheric gas oil	P/E: 0.55	0.40	0.34	0.29
Hydrocracker residue	P/E: 0.55	0.50	0.36	0.31

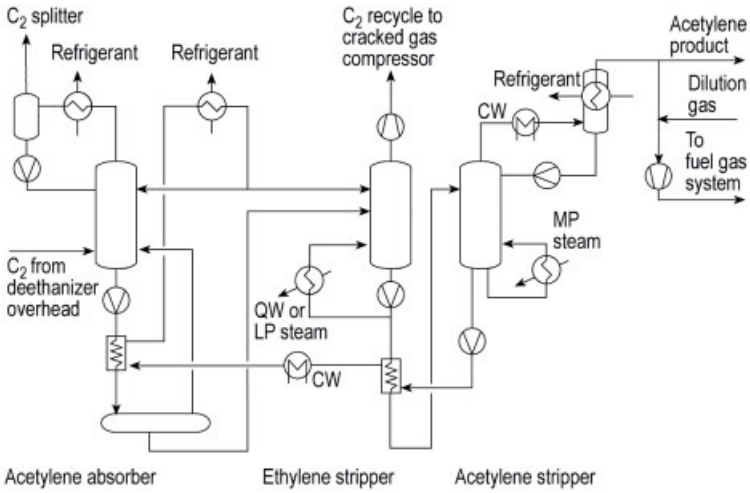

Figure 34. Acetylene recovery process [77]

tion is about 0.4 – 2.5 wt %. An ethylene plant producing 400 000 t/a ethylene produces 4500 – 11 000 t/a acetylene. The acetylene is removed by catalytic selective hydrogenation or solvent extraction. Today the dominating acetylene removal process is selective hydrogenation.

Acetylene Hydrogenation. Most ethylene plants are equipped with a hydrogenation unit. Acetylene is converted selectively to ethylene on a Pd-doped catalyst. Whereas in the past mainly Ni catalysts were used, today Pd catalysts are doped with other metals such as silver to improve selectivity [73]. Typical process conditions are temperatures of about 40 – 120 °C, pressures of 15 – 40 bar and space velocities of 1000 – 120 000 kg L^{-1} h^{-1}. Depending on the type of feed and the plant, there are several process options:

1) Front-end hydrogenation (C_{2-} stream containing H_2, CO, methane, C_2H_2, C_2H_4 and C_2H_6)
2) Raw-gas hydrogenation (hydrogenation before C_2/C_3 separation, stream containing H_2, CO and all hydrocarbons of the raw gas after cracked gas compression).
   This variant of front-end hydrogenation is only used in gas crackers where the content of C_{3+} hydrocarbons is comparable low.
3) Tail-end hydrogenation (pure C_2 stream containing C_2H_2, C_2H_4, and C_2H_6; separate addition of an equimolar amount of hydrogen is necessary)

Typical specifications for acetylene content in the ethylene product are < 2 ppm with a tendency to further reduction to < 1 ppm.

General aspects of the process and the catalyst requirements are reviewed in [73 – 76] (see also → Ethylene, Chap. 5.3.2.2).

Acetylene Recovery. Acetylene is extracted from the C_2 fraction of the steam cracker. The solvent must fulfil the following criteria:

– Melting point lower than the dew point of the feed gas
– High solubility of acetylene at a temperature near the dew point of the C_2 fraction
– High acetylene selectivity ([9, 11])
– High chemical and thermal stability

- No foaming tendency due to traces of hydrocarbons
- Low toxicity
- Low vapor pressure at the operating temperature

After testing many solvents, including DMF, NMP, and acetone, the most suitable solvent for such a process proved to be DMF. The solubility of acetylene as a function of temperature is shown in Figure 5.

The process for the recovery of high-purity acetylene is shown in Figure 34 [77]. The gaseous C_2 mixture, consisting of ethylene, ethane, and acetylene, is fed to the acetylene absorber; the gas stream is contacted with counterflowing lean DMF at a pressure of 0.8 – 3.0 MPa. The process is suitable for the full pressure range prevailing in any of the known ethylene processes. The entire acetylene and some of the ethylene and ethane are dissolved by the solvent. Water and carbon monoxide decrease the solubility of acetylene. Furthermore, water causes hydrolysis of DMF, resulting in additional formation of carbon monoxide. Entrainment of DMF at the top of the column is avoided by a reflux stream. The purified C_2 fraction, containing < 1 ppm of acetylene, is fed to the C_2 splitter. The rich solvent stream is sent to the ethylene stripper, which operates slightly above atmospheric pressure. Ethylene and ethane are stripped off and recycled to the first stage of the cracked gas compression. Any acetylene entrained with the overhead gas is recovered by washing with cold solvent at the top of the stripper. In the acetylene stripper, pure acetylene is isolated from the top of the column. After cooling and heat recovery, the acetylene-free solvent is recycled to the absorber and ethylene stripper. The acetylene product has a purity of > 99.8 % and a DMF content of less than 50 ppm and is available at 10 kPa and ambient temperature.

The material balance and the utilities consumption of an acetylene recovery unit are listed in Tables 11 and 12. At present, more than 112 000 t/a of petrochemical acetylene from twelve olefin plants worldwide is recovered by this technology. With a big drop in the total acetylene market in 2001, the market decreased on average by 6.8 % per year from 1998 to 2003, and only a marginal growth rate is expected until 2009 (max. 2 % per year) [135]. For this reason during the last five years (2006) no new acetylene recovery facilities have been erected. Only some existing plants were expanded. In North America, for example, there are currently only two major companies producing acetylene as byproduct of ethylene production, i.e., Dow Chemical in Taft, LA, and Petroment at Varennes in Canada [136]. Figure 35 shows an industrial plant with a design capacity of 14 400 t/a of high-purity acetylene.

Table 12. Consumption of utilities for an acetylene recovery process operating on the C_2 fraction from a plant producing 400 000 t/a C_2H_4 (Linde) [78]

DMF, kg/h	1.3
Heating steam, t/h	3.9
Cooling water, m^3/h	100
Electrical energy, kW	125
Refrigerant, GJ/h	6.3
Quench water, GJ/h	3.1
Plot area, m × m	15 × 40

A material balance for ethylene plant outputs including acetylene extraction or hydrogenation is shown in Figure 36 [79]. The economic evaluation shows that petrochemical acetylene remains attractive even if the price of ethylene is doubled. It is economical to retrofit acetylene absorption in an existing olefin plant equipped with a catalytic hydrogenation. Isolation of acetylene obtained as an unavoidable byproduct of ethylene production is the economically most attractive route to cover acetylene demand.

A similar process is available for propyne (see Chap. 7)

5. Safety Precautions, Transportation, and Storage

General literature is given in [10, 80, 81].

5.1. General Safety Factors and Safety Measures

Decomposition and Combustion. Acetylene is thermodynamically unstable under normal conditions. Decomposition into carbon and hydrogen can achieve temperatures of about 3100 °C, but due to formation of other products, the temperature reached adiabatically is 2800 – 2900 °C. The decomposition can be initiated by heat of reaction, by contact with a hot body, by an electrostatic spark, by compression

Figure 35. Acetylene recovery plant (name plate capacity: 14 400 t/a of high-purity acetylene)

Table 11. Material balance (mol %) for an acetylene recovery process operating on the C_2 fraction from a plant producing 400 000 t/a C_2H_4 (Linde) [78]

	Gas to absorber	Purified gas	Recycle gas	Product C_2H_2
Methane	trace	trace		
Acetylene	2	1 ppm	4	99.8
Ethylene	82	83.5	85.7	0.2
Ethane	16	16.5	10	trace
C_3	trace	trace	0.3	trace
DMF		1 ppm	trace	trace
Temperature, K	252	249	255	258
Pressure, MPa	2	1.98	0.11	0.12
Flow rate, kmol/h	2186	2126	17.5	52.5

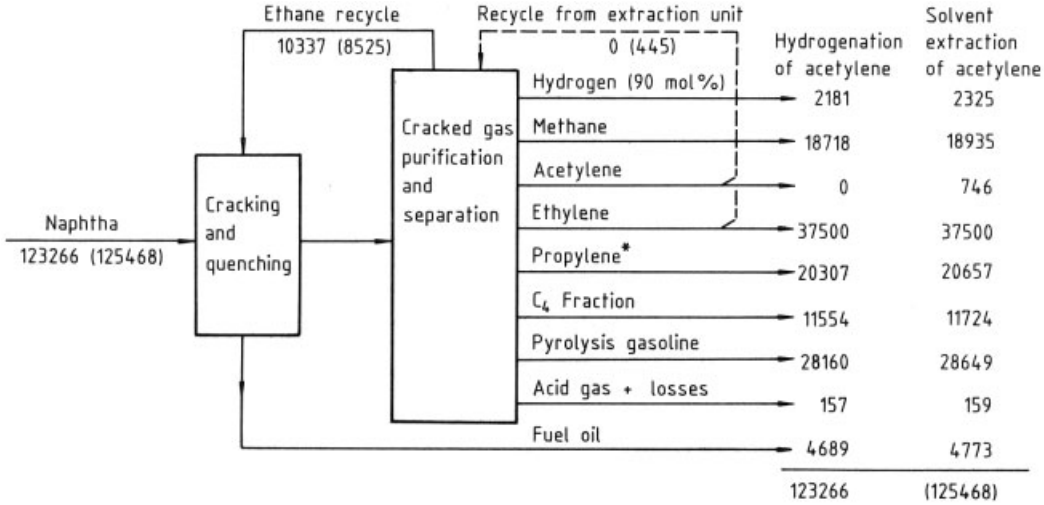

Figure 36. Material balances of a 300 000 t/a ethylene plant equipped with either C_2 hydrogenation or acetylene extraction (all rates in kg/h, the numbers in parentheses are for the solvent extraction process) [79]
* Chemical grade

heating, or by a shock wave. The decomposition induced by heating the wall of the container or pipe is very sensitive to the pressure, the size and shape of the container or the diameter of the pipe, the material of the container, and traces of impurities or other components. Solid particles such as rust, charcoal, alumina, and silica can lower the ignition temperature compared to clean steel pipe.

Decomposition gives rise to different scenarios:

- Working range I (deflagration): a flame produced by decomposition or combustion and propagates at a velocity below the velocity of sound into the unconverted gas (pressure rises simultaneously in front of and behind the flame front)
- Working range III (detonation): the flame propagates at ultrasonic velocity into the unconverted gas (shock wave between low pressure in the unconverted gas and high pressure in the converted gas)
- Working range II (intermediate between I and III): often the propagation velocity of a deflagration is not constant (it increases with increasing density, temperature, and turbulence), and therefore a change from deflagration to detonation is observed

As consequence, design criteria for piping and other components are proposed for the different working ranges and depend on the diameter of the pipe. Limit lines for deflagration and detonation are given in [82] on the basis of the work of Sargent [83]. An extended Sargent diagram is shown in Figure 37.

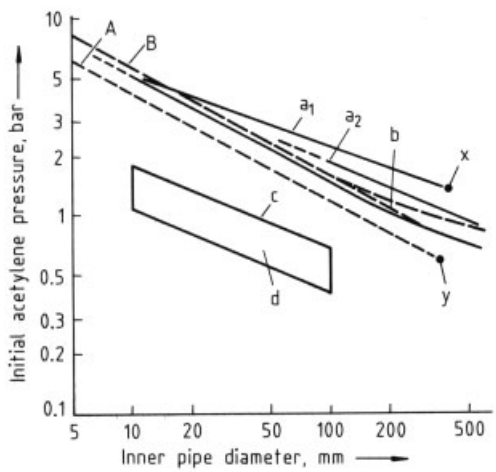

Figure 37. Detonability limits of acetylene [83, 88]
A) Deflagration limit; B) Detonation limit
Detonation limits: a) Thermal ignition in a plain pipe (a_1 melting wire, 20 – 80 J; a_2 detonator cup, ≈2400 J); b) Thermal ignition plus orifice; c) Ignition by chemical reaction in a shock wave; d) Range of possible quasi-detonation depending on ignition energy of shock wave; x) and y) Limiting ignition pressure for thermal ignition with melting Pt wire and with detonator cup, respectively

The limits are influenced by the method of ignition (e.g., melting wire or a detonator cap). Changing the method of ignition from a melting Pt wire (Reppe) to the exploding wire ignitor (ignition energy ca. 70 J) used by BAM resulted in lower stability pressures for acetylene mixtures [84] and pure acetylene [85] (Table 13). A description of the experimental setup can be found in [134].

Mixtures of acetylene with methane have higher stability pressures than those with nitrogen or hydrogen. The phlegmatizing influence of foreign gases increases in the order $H_2 < N_2 < CO_2 < NH_3$, which corresponds to the heat capacities of these gases [134]. For $C_2H_2 - C_2H_4$ mixtures the stability pressure rises with increasing ethylene fraction. However, a significant effect can be observed only at C_2H_2 contents $<$ 50 % [134]. Further information on the effect of additional gases on acetylene decomposition is given in [86]. The dependence of stability pressure on the energy of the ignition source has led to ongoing discussion about its relevance for industrial design and operations [87].

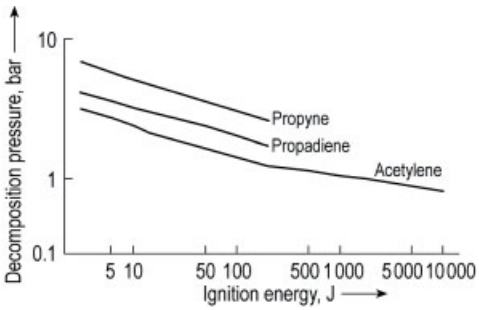

Figure 38. Decomposition pressure versus ignition energy for unsaturated hydrocarbons

Figure 38 shows the decomposition pressure of acetylene, propyne, and propadiene as function of ignition energy. This relationship is the basis for the safe design of processes for the recovery of acetylenic components (see Section 4.4.2 and Chap. 7). Additional investigations have been published on the dependence of deflagration pressure on the flow in pipes [88] and the decomposition of high-pressure acetylene in branched piping [89]. Solid acetylene is not critical with regard to decomposition, provided it is the only material involved [7]. In liquid oxygen, solid acetylene can readily ignite on mechanical impact and react violently [90]. Recommendations for equipment used in gas welding and cutting technology, such as rubber hoses, safety devices, and flame arresters, are given in [91 – 93].

Combustion of Acetylene in Oxygen (Air). The reaction of acetylene and oxygen at 25 °C and 1 bar to form water and CO_2 generates 1255.6 kJ/mol. Temperatures of around 3100 °C can be reached. Figure 39 shows flame temperatures and flame front velocities for mixtures of oxygen with hydrocarbons [94].

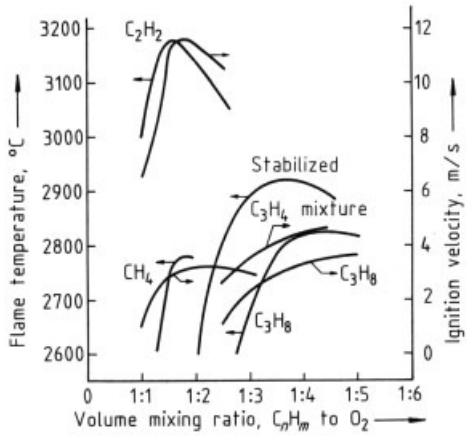

Figure 39. Flame temperatures and ignition velocities of acetylene–oxygen mixtures and mixtures of other hydrocarbons with oxygen

Acetylene allows the highest temperatures and flame front velocities to be attained. The maximum temperature is very sensitive to the mixing ratio, which also determines whether a reducing, neutral, or oxidizing flame exists (Fig. 40).

Fundamental safety data for acetylene–air and acetylene–oxygen mixtures are listed in Table 14. At atmospheric pressure and 25 °C mixtures of 2.4 – 93.0 vol % acetylene in oxygen are explosive; the possibility of self-decomposition at high acetylene must also be taken into account. A large explosion at Acetylene Services Company (ASCO) in the USA in 2005 was caused by accumulation of acetylene in the water-pipe system. Most probably the contact of acetylene with the hot surface of a propane heater caused the explosion [95]. A comparison

Table 13. Stability pressure (bar) of acetylene and acetylene mixtures for two methods of ignition and pressure increase [84, 85]

Mixture	Ignition method		p_{ex}/p_a
	Reppe	BAM	(BAM)
100 % C_2H_2	1.4	0.8	
90 % C_2H_2/10 % N_2	1.8	1.0	8.6
90 % C_2H_2/10 % CH_4	2.1	1.0	7.2
90 % C_2H_2/10 % H_2	1.6	0.9	6.9
50 % C_2H_2/50 % N_2	9.0	3.6	6.3
50 % C_2H_2/50 % CH_4	14.7	12.9	
50 % C_2H_2/50 % H_2	4.7	2.5	5.5

* p_{ex}/p_a: ratio of maximum pressure to pressure before ignition.

of C_2H_2 deflagration and $C_2H_2-O_2$ explosion with TNT explosion is given in [96].

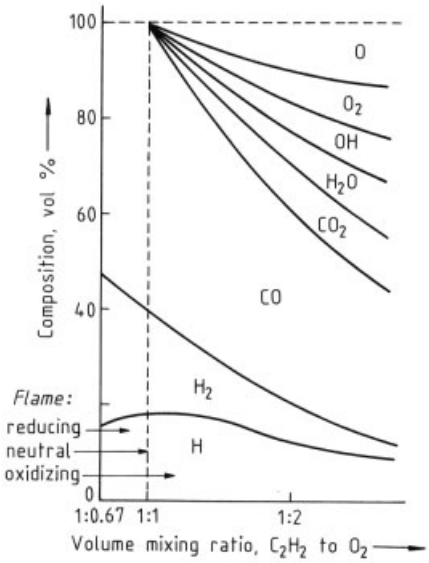

Figure 40. Chemical composition of an oxygen–acetylene flame at its tip versus mixing ratio

Table 14. Fundamental safety data for acetylene–air and acetylene–oxygen mixtures

	Air	Oxygen
Lower flammability limit, vol %	2.5	2.4
Upper flammability limit, vol %	82	93
Flame temperature* , K	2863	3343
Flame front velocity, m/s	1.46	7.6
Increase of pressure (deflagration)	11	50
Detonation velocity, m/s	2300	2900

* Stoichiometric mixture.

Handling of Acetylene. For pure acetylene the prescribed safety instructions, for example, the *Technische Regeln für Acetylenanlagen und Calciumcarbidlager* (Technical regulations for acetylene plants and calcium carbide depots), *TRAC,* [97], have to be strictly followed. The former *Acetylenverordnung* (Decree about acetylene) [53] in Germany has been replaced by the more general *Betriebssicherheitsverordnung* (Decree about safety in plants) [97] as of January 1, 2003. However, it is not possible to formulate general safety instructions for the great variety of chemical processes with acetylene as reaction component under diverse reaction conditions.

Both handling acetylene and experiments with it necessitate critical examination of sources of possible danger. The literature cited can only serve as an aid to decisions on precautions. The safety regulations mentioned above have been determined in experiments with well-defined apparatus dimensions (length, diameter, geometry). For other dimensions they can only serve as an indication of explosive behavior and should not be considered as rigid limits. The development of economical chemical processes involving acetylene at elevated pressures or under other hazardous conditions calls for decomposition tests for the crucial stages where decomposition could occur. This must be done in close cooperation with official testing institutions, such as the Bundesanstalt für Materialprüfung (BAM) in Germany.

In general, the following rules should be observed in handling acetylene:

- Temperature and pressure must be selected so as to avoid liquefaction of acetylene.
- Reactions of acetylene in solvents or with liquid reaction components must be carried out at such acetylene concentrations that explosive decomposition of the acetylene in the liquid phase cannot occur. In many cases this condition is fulfilled at an acetylene loading below 100 m^3 (STP) per cubic meter of solvent. Higher loadings are only permitted if additional precautions are taken, such as filling the volumes containing the liquid with

steel packings. The formation of a separate gas phase has to be avoided.

– The technical rules (such as *TRAC* [97]) are valid for pure gaseous acetylene. If an inert gas, such as nitrogen, is added to the acetylene, higher acetylene partial pressures are permitted.
– In the design of apparatus the partial pressure of acetylene should be selected so that the minimum distance to the decomposition limit is about 20 %. The apparatus should be designed to withstand pressures (1) 12-fold the initial pressure for pure acetylene systems or (2) the initial pressure plus 12-fold the acetylene partial pressure for mixtures and solutions.
– Formation of hydrates (see Chap. 2) under pressure must be avoided because this leads to obstructions in the apparatus and pipelines. The melting point of these compounds is in the range 0 –13 °C; therefore, pressurized acetylene containing water has to be kept above 15 °C.

In addition to the measures for building construction, electrical installations, fire protection, purging, and leak detection, acetylene plants and distribution systems are provided with flame traps and flashback and release valves and locks [98, 99]. Flame traps consist either of tubes immersed into water-filled cylinders (wet trap) or cylinders filled with a packing of high surface area to decelerate the decomposition. A wet arrester, which is used for an acetylene distributing line, is shown in Figure 41. Suitable materials for dry-trap packings are sinter metals, ceramic beads (e.g., Raschig rings), bundles of small tubes, and corrugated metal foils [100, 101].

Tapping points for acetylene distribution units which meet the German *TRAC* rules include a nonreturn valve to avoid intrusion of air from downstream, a sinter-metal flame trap, and a thermo- or pressure-sensitive spring lock. The last closes if a flame is stopped by the trap but still burns outside of the flame trap. Detailed information is given in [98].

Transportation in Pipelines. Acetylene is occasionally transported in pipelines. Figure 42 shows the safety components of an acetylene pipeline between Burghausen and Gendorf, Federal Republic of Germany. The pipeline was operated until 1976 without incident. Its length is 8 km, and the pipes are 300 mm in diameter. Design pressure was 100 bar, although operation pressure was only 2 bar at the inlet and 1.25 bar at the outlet. The pipeline was provided with rupture disks, which open to atmosphere in case of decomposition. Quick closing valves are initiated simultaneously to protect both upstream and downstream equipment. At each end, part of the pipeline is filled with tube bundles to stop propagation of any acetylene decomposition. The flame traps consist of 600-mm-diameter U's filled with Raschig rings.

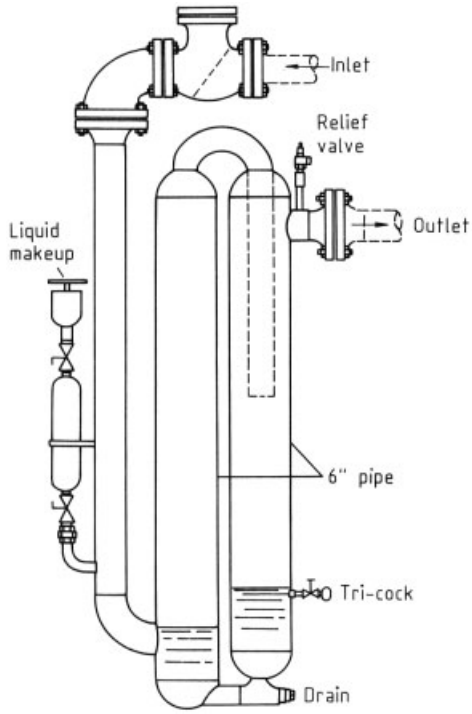

Figure 41. Hydraulic flame trap for acetylene lines (Union Carbide), [102]

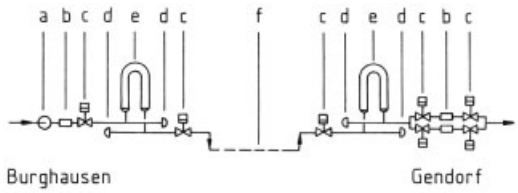

Figure 42. The 8 km acetylene pipeline from the Marathon refinery, Burghausen, to Farbwerke Hoechst, Gendorf, Federal Republic of Germany [103]
a) Compressor; b) Control points; c) Automatic quick-closing valves; d) Rupture disks to atmosphere; e) Flame traps; f) Pipeline

A report [102] is available on an acetylene-decomposition event in a pipeline system, which demonstrates the need for safety measures. Instead of transportation of pure acetylene, pipeline transportation of acetylene solutions in acetone was proposed as safer [103]. In the United States, transportation of acetylene solutions in liquid ammonia was considered for existing ammonia distribution systems [104].

Hazardous Acetylene Traces in Low-Temperature Processes. Acetylene is the most dangerous component in gas mixtures processed in low-temperature plants. In air separators, for example, acetylene can be suspended in liquid oxygen as a solid or as a segregated liquid phase that is quite unstable and tends to uncontrollable and violent decomposition. (The solubility of acetylene in liquid oxygen is low, see Chap. 2) Therefore much attention must be given to checking for and removing acetylene in low-temperature separation plants.

Normally air contains some acetylene, up to 0.3 mL/m^3. In industrial areas, especially in the proximity of petrochemical plants, higher concentrations (up to 1 mL/m^3) can occur. Without any measures, acetylene in the feed air of an air separator would be enriched in the cold section of the unit.

In modern air separators alternating molecular sieve adsorbers are used. The adsorbers obey the following breakthrough sequence:

$$CH_4 - C_2H_6 - \begin{matrix} C_2H_4 \\ CO_2 \\ C_3H_8 \end{matrix} - C_2H_2 - C_3H_6 - C_4 - H_2O$$

To avoid acetylene breakthrough the adsorber is operated for sufficient CO_2 removal and is regenerated when the CO_2 concentration at the adsorber outlet starts to increase. Further details are given in [105 – 109].

Another possible way to remove acetylene and other combustible air contaminants is catalytic oxidation prior to the separation [107, 110]. This method, being expensive, is rarely used.

All air separators are provided with routine analysis systems for acetylene (→ Oxygen). Routine analysis concentrates on the liquid oxygen of the main condenser [111].

The removal of acetylene in cracked gas separation is treated in Section 4.4.2. The processing of other acetylene-containing gases, e.g., coke-oven gas, by low-temperature separation is described in [112].

5.2. Acetylene Storage in Cylinders

Because of its tendency to deflagrate or to detonate, acetylene cannot be compressed and stored in gas cylinders like other gases. At the end of the 19th century attempts to store acetylene under high pressure or liquefied led to fatal detonations in the USA, Paris, and Berlin.

For desensitizing, acetylene stored in gas cylinders is dissolved in a solvent in which the acetylene is very soluble. This solvent is dispersed in a porous solid that completely fills the gas cylinder. As well as giving better solvent distribution the porous material arrests any local acetylene decomposition induced, for instance, by flashback.

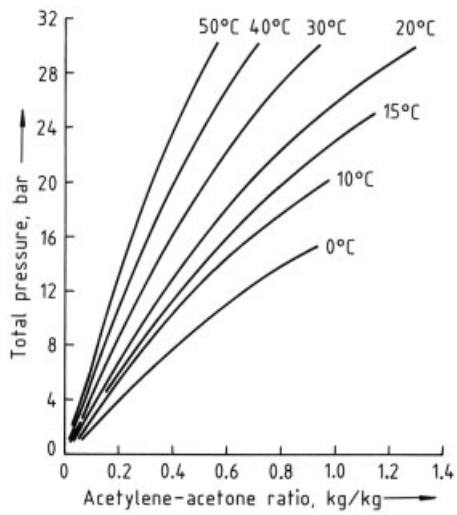

Figure 43. Total pressure of acetylene solution in acetone as a function of acetylene : acetone ratio and temperature [7]

Acetone and dimethylformamide are the preferred solvents for acetylene in cylinders. An advantage of dimethylformamide is its lower vapor pressure, resulting in lower solvent losses during acetylene discharge. A disadvantage is its higher toxicity. The total pressure of an acetone-containing acetylene cylinder depends

on the acetylene : acetone ratio and on temperature as is shown in Figure 43. Deviations from the plotted curves resulting in higher pressures are caused by the porous filling of the gas cylinder, which absorbs acetone, changing the effective acetylene : acetone ratio [113].

Impurities in the acetylene decrease the dissolving capacity of the acetone. Figure 44 shows the effect of moisture on acetylene solubility. As a result, acetylene produced from calcium carbide and water has to be dried.

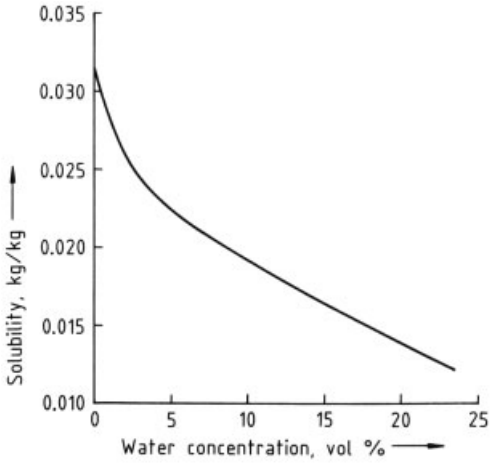

Figure 44. Solubility of acetylene in water-containing acetone at 25 °C and $p_{C_2H_2} = 1$ bar (0.1 MPa) [114]

Calcium carbide-based acetylene contains further impurities that have to be scrubbed out to avoid decreased solubility in acetone. Examples are divinyl sulfide and phosphine: 1 wt % divinyl sulfide in the acetone reduces the acetylene solubility from 35 g/kg to 31 g/kg at 20 °C and an acetylene pressure of 0.1 MPa. Further values, also for phosphine, are given in [114, 115]. The impurities have to be scrubbed out to residual concentrations of 0.5 g of phosphorus and 0.1 g of sulfur per cubic meter of acetylene.

During acetylene production from calcium carbide, disperse calcium hydroxide (0.1 – 1.0 μm) is produced. This contaminates the product gas. The calcium hydroxide present in acetylene filled into acetone-containing gas cylinders catalyzes aldol condensation of the solvent and reduces the solubility for acetylene:

$$2\ (CH_3)_2CO \rightarrow (CH_3)_2COHCH_2COCH_3$$

diacetonealcohol

Therefore, the solids content of calcium carbide-based acetylene for filling acetone-containing gas cylinders must be kept below 0.1 mg/m^3 [114, 116].

The gas cylinders have to be filled with definite amounts of acetylene and solvent. Commercial-grade, seamless gas cylinders which meet specified standards (in Germany, DIN 4664) may be filled with the amounts listed in Table 15.

The amounts are fixed by regulations for handling pressurized gases [118]. Corresponding regulations in the United States have been issued by the Department of Transportation [119].

To determine maximum acetylene filling of gas cylinders, extensive ignition, impact, and heating tests have been worked out [120, 121].

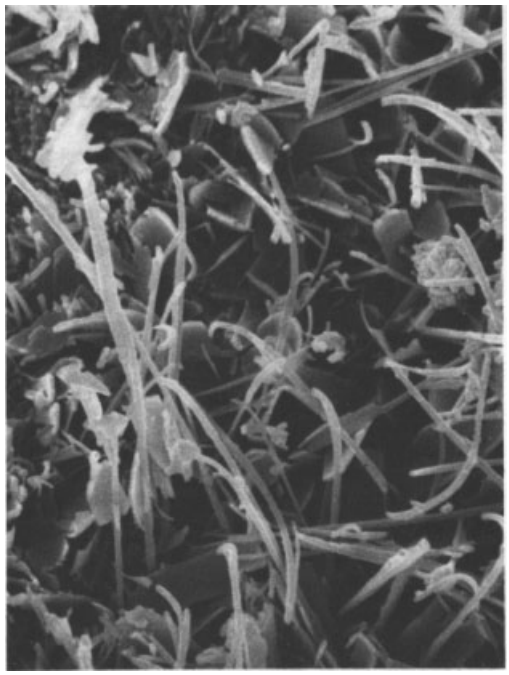

Figure 45. Porous silica material Linde M1 for acetylene cylinders (magnification 1 : 10000)

The porous material in the acetylene gas cylinders must satisfy the following requirements: no interaction with the cylinder material, acetylene, or acetone and suitable mechanical properties, such as sufficient impact resistance. Suitable materials include pumiceous compounds, silica, charcoal, asbestos fiber, and

Table 15. Permitted acetylene and acetone filling of seamless gas cylinders (satisfying German standards and safety rules) [117]

Gas cylinder			Acetone filling, kg			Acetylene filling, kg
Vol. of gas cylinder, L	Outer diameter, mm	Length, mm	Minimum	Maximum		
				General	Exceptional	
3	140	300	0.789	0.8625	0.9375	0.4725
5	140	460	1.315	1.4375	1.5625	0.7875
10	140	850	2.630	2.8750	3.125	1.575
20	204	810	5.260	5.750	6.250	3.150
27	204	1040	7.101	7.7625	8.4375	4.2525
40	204	500	10.520	11.50	12.50	6.30
40	229	1210	10.520	11.50	12.50	6.30

Table 16. Examples of approved silica-based porous materials in German acetylene cylinders [122] (for seamless 40-L cylinders satisfying DIN 4664, Blatt 2)

Material		Approved filling			
Type	Origin [a]	Acetylene [b], kg	Acetone [c], kg	Acetylene : acetone, kg/kg	Maximum pressure [d], bar
Linde M1	Linde, Munich	8.0	12.7	0.63	19
AGA 2	AGA, Hamburg	8.0	12.4	0.645	19
SIAD 2	SIAD, Sabbio	8.0	12.4	0.645	19

[a] Approval only when porous filling is prepared at place of origin;
[b] Maximum;
[c] Desired value;
[d] Gage, at 15 °C.

alkaline carbonates. The porosity of these materials varies between 70 and 80 % [114, 122]. Modern monolithic materials are made preferably from silica, lime, and glass fiber. The mixtures are suspended in water to obtain a pasty material which is filled into the gas cylinders. The material is hardened at about 200 °C and subsequently dried and activated at 350 – 400 °C. A porosity of about 90 % is obtained.

Any porous material to be used for acetylene cylinders has to be examined and approved by competent authorities. The examination includes the determination of the maximum acetylene and solvent filling, the maximum filling pressure, and ignition and impact testing. Table 16 lists three porous materials approved for use in Germany. Figure 45 shows a photograph of "Linde M1" magnified 1 : 10000, clearly revealing the porous structure of such materials.

Methods for examining the materials have been standardized by CPI (Commission Permanente Internat. de l'Acétylène, de la Soudure Autogène et des Industries qui s'y rattachent, Paris) and ISO (International Organization for Standardization) [122].

Discharging acetylene from a gas cylinder leads to acetone losses because the partial pressure of acetone at 15 °C ranges from 0.14 bar at 15 bar total pressure to 0.18 bar at 1 bar total pressure. Solvent loss has to be replaced when an acetylene cylinder is reloaded. Further details concerning acetylene cylinders, porous materials, and valves can be found in [131].

6. Uses and Economic Aspects

6.1. Use in Metal Processing

Acetylene has many applications in the processing of metals and other materials. This is because of the high flame temperature and propagation velocity resulting in high energy densities and rapid heat transfer to the piece being worked. Examinations at Linde showed that for extreme oblique sections the efficiency using acetylene is up to 50 % higher than using propane as fuel gas [133]. The properties of an oxyacetylene flame given here supplement those in Section 5.1.

The temperature profile of an oxyacetylene flame consists of a hotter primary flame and a scattered secondary flame. The highest flame temperature is at the tip of the primary flame (Fig. 46). For material processing the primary flame is the more important.

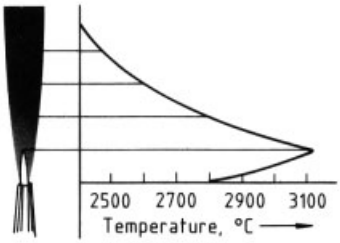

Figure 46. Temperature profile of an oxyacetylene flame

The heating efficiency of the primary flame is the product of the volume-based heat released by the primary flame and the propagation velocity. This is plotted in Figure 47 for the oxyacetylene flame and some other flames. The heat transferred in welding is generated by radiation, convection, and thermal conduction (see Table 17). The heat transfer is promoted by a high temperature gradient between flame and workpiece.

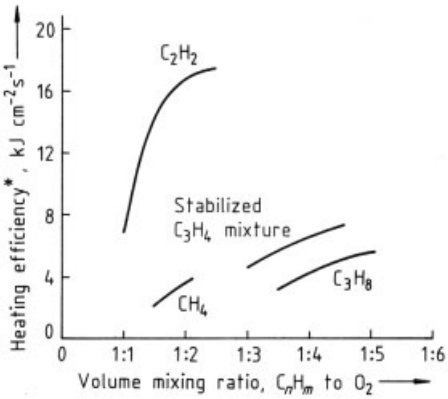

Figure 47. Heating efficiency of acetylene–oxygen mixtures and mixtures of other hydrocarbons with oxygen
* based on area of primary flame cone

Oxidizing, neutral, or reducing (carburizing) flames can be obtained by varying the oxygen : acetylene ratio (Fig. 40). For steel, alumina, and copper welding usually neutral or slightly reducing flames are used, whereas oxidizing flames are preferred for brass welding, cutting, pickling, and surface hardening [124]. Acetylene is burned with oxygen in single torches or in bundles of torches, the chief components of which are the connections for acetylene and oxygen, regulating valves, a mixing chamber (usually of the injection type), a flashback protection element, and a nozzle adapted to the specific applications [125, 126].

Oxyacetylene flames are used in welding, cutting, brazing, soldering, surfacing, flame spraying, heating, hardening, straightening, cleaning, pickling, rust removal, and decarbonizing.

Acetylene–air flames are occasionally used for tin brazing, hot air welding of thermoplastics, glassworking, and paint removal [125], although the convenience and safety of fuels such as propane or butane has displaced acetylene in those applications. Soft and hard soldering, flame hardening, and flame tempering are important applications for the softer acetylene–air flame. For acetylene–air mixtures, self-aspirating Bunsen-type and acetylene–compressed air burners are used.

The different uses of oxyacetylene and acetylene–air flames in metal working, the procedures, and the equipment are comprehensively described in [7] and [127]; other sources of information for oxyacetylene flame properties in welding are [123, 124], and [128].

6.2. Use as Raw Material in Chemical Industry

Because of the diversity of acetylene chemistry (see Section 3.1), acetylene has been used as a starting material for a great variety of industrially important products. These are summarized, together with their applications, in Figure 48.

Between 1960 and 1970, when worldwide acetylene production peaked, most of the products listed in Figure 48 were produced via acetylene. The competition between acetylene and the olefins since the 1970s (see Section 6.3) has resulted in substitution of ethylene and propene for acetylene, especially in the production of acetaldehyde and acrylonitrile. At present, acetylene is used mainly for the production of vinyl chloride, vinyl acetate, and other vinyl esters; acrylic acid; acetylene black; and acetylenic chemicals such as 1,4-butynediol and acetylenic alcohols. For the acetylenic chemicals the acetylene route is either the only commercial production process available or the predominant process. Vinyl chloride, vinyl acetate, and acrylic acid, formerly the main products from acetylene, are pro-

Table 17. Heat transfer in welding [123]

Gas temperature, K	Welding temperature, K 800		1200		1600	
	Q_S	Q_K	Q_S	Q_K	Q_S	Q_K
1000	2.3	13.4				
2000	3.8	55.4	3.4	37.0	2.3	18.5
3000	4.0	83.2	3.8	68.0	3.6	50.4

Q_S, heat transfer by radiation (kJ cm^{-2} h^{-1});
Q_K, heat transfer by convection (kJ $cm^{-2}h^{-1}$) at a gas velocity of 50 m/s

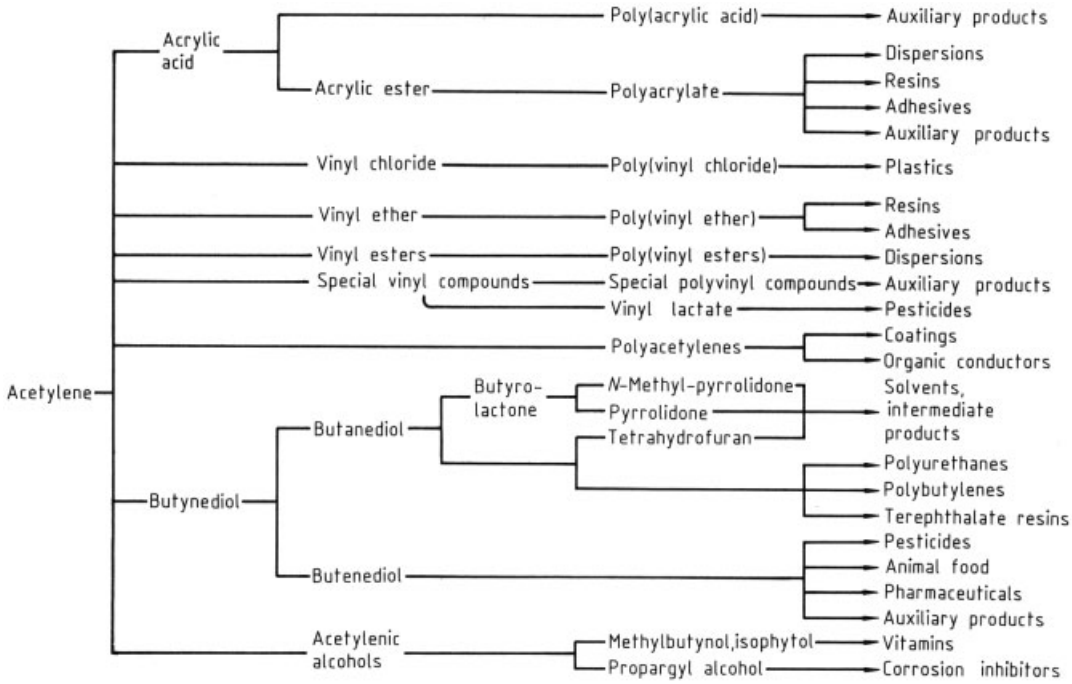

Figure 48. Acetylene as a starting material for industrial products

duced today mainly from ethylene and propene [3].

6.3. Competitive Position of Acetylene as Chemical Feedstock

Today, acetylene plays an important role only in the production of the acetylenic chemicals. The fact that acetylene production has not decreased further seems to indicate that the competition from the olefins is no longer as strong as it was. The main reason for this is that European olefin chemistry depends on refinery products, which have become more expensive than natural gas, the main feedstock for acetylene. Another contributing factor is that acetylene is produced only in old plants, which have low capital costs.

In addition, process improvements, such as an increase in thermal efficiency and optimum use of byproducts by other plants, can make acetylene more competitive. The position of acetylene in chemical industry may be advanced because of the variety of valuable products to which it can be converted in high yields with known technology. Acetylene must compete with ethylene for the production of vinyl chloride and vinyl acetate, and for the production of acrylic acid and its esters it must compete with propene. The Stanford Research Institute [3] has investigated this question in detail. The results, which take into account both capital investment (25 % return) and specific consumption figures for the various processes, show that the prices at which

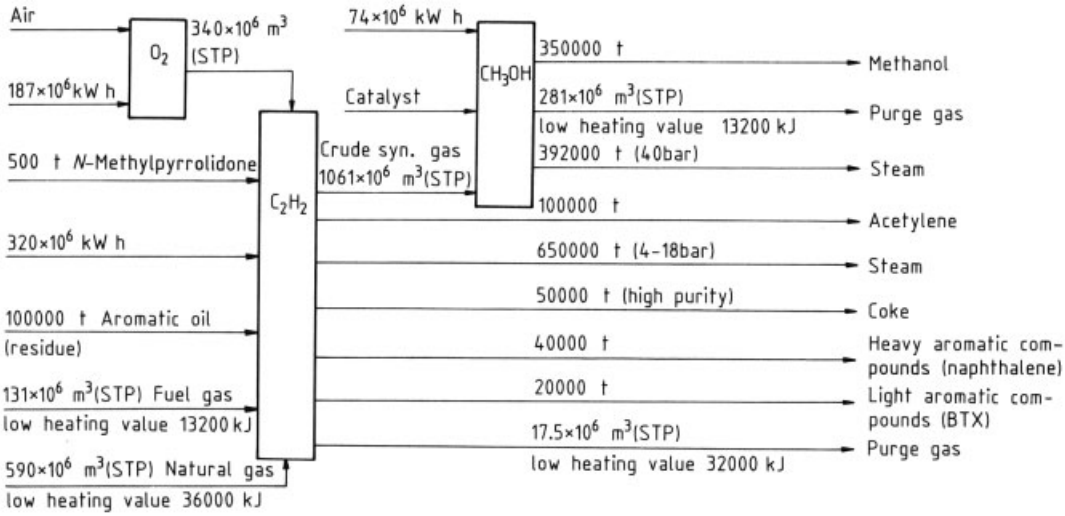

Figure 49. Acetylene–methanol plant

the alternate routes are competitive can be expressed by the following equations:

for vinyl chloride $A = 1.10\ E + 0.42$
for vinyl acetate $A = 1.23\ E + 0.40$
for acrylic acid $A = 1.74\ P + 0.23$

where A is the acetylene price, E is the ethylene price, and P is the propene price, all in \$/kg. For example, if ethylene costs 0.65 \$/kg, acetylene can cost 1.15 \$/kg for the production of vinyl chloride and 1.21 \$/kg for the production of vinyl acetate. If propene costs 0.49 \$/kg, acrylic acid can be profitably produced from acetylene if it costs 1.08 \$/kg or less.

Acetylene prices of 1.08 to 1.21 \$/kg or less can be reached in a new plant only with optimal integration of energy and byproducts within an integrated chemical plant. Figure 49 shows the flow sheet of a 100 000 t/a gas-based partial oxidation acetylene plant in a chemical complex. The required oxygen facility and methanol plant based on acetylene synthesis gas (off-gas) are included. The main products of the complex are acetylene and methanol. The acetylene process is operated with improved quenching technology, allowing a high proportion of energy to be regained in the form of steam. The aromatic residue oil from a steam cracker is converted into high-purity coke and light aromatics.

A production cost estimate based on power, feedstock, and product prices roughly corresponding to market prices in 1982 shows that low production costs for acetylene are possible under the conditions described, in spite of the relatively high natural gas price of 5.5 \$/10^6 BTU (0.021 \$/kW h), which in heating value terms comes very close to that of crude oil (0.022 \$/kW h). The difference in production costs between acetylene from natural gas and ethylene from naphtha (price 333 \$/t, corresponding to 0.028 \$/kW h) is so slight that acetylene from new plants can once again compete with ethylene for certain syntheses, provided that there is a difference between the costs of natural gas and naphtha. This was nearly 0.007 \$/kW h in 1982 in Germany. A similar calculation for the acetylene production by the Hüls arc process based on an ethylene price of 600 \$/t and a hydrogen price 40 % above the heating value results in the same acetylene product value at 0.038 \$/kW h for electrical power. Lower electrical energy costs favor the arc process.

7. Propyne

Propyne [*74-99-7*], methylacetylene, is obtained in cracking processes mostly as a byproduct together with its isomer propadiene [*463-49-0*], allene. Typical concentrations depend on the feedstock and the cracking conditions (Table 10) and vary between 0.3 and 0.8 wt %. The corresponding figures for propadiene are 0.3 – 0.6 wt %.

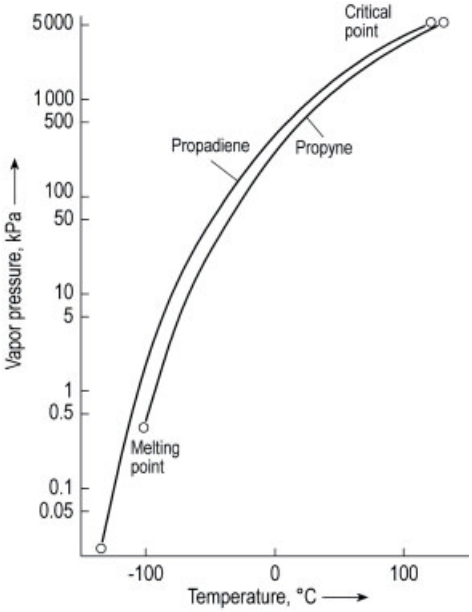

Figure 50. Vapor pressure curves of propyne and propadiene [129]

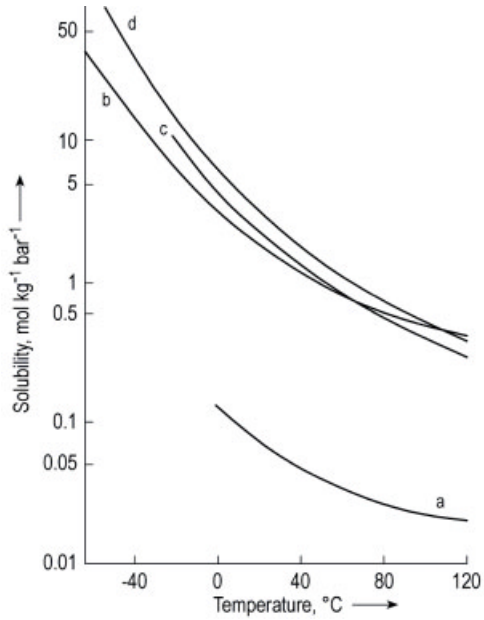

Figure 51. Solubility of propyne in various solvents at infinite dilution
a) Water; b) Methanol; c) *N*-methylpyrrolidone; d) DMF

Pure propyne is a colorless, nontoxic, flammable gas. The important physical properties of propyne are listed in Table 18 [5, 6, 129]. Vapor pressure curves of propyne and propadiene including melting and critical points are shown in Figure 50 [129]. Table 19 summarizes solubility coefficients of propyne for various solvents [9]; further data are available in [10]. The solubility of propyne in various solvents is shown in Figure 51; the solubilities of the C_3 hydrocarbons in DMF are plotted in Figure 52 for infinite dilution.

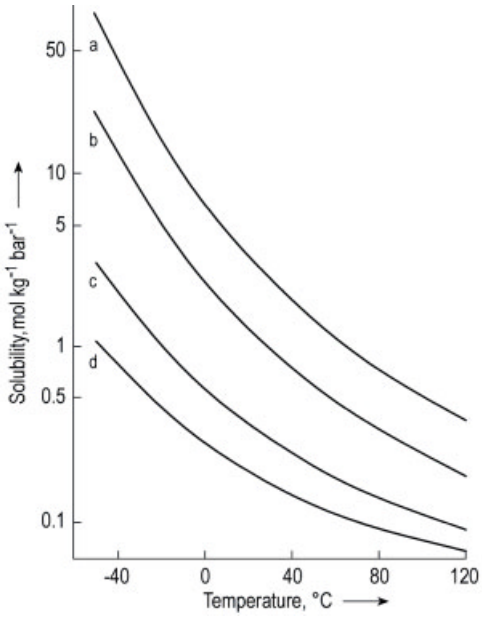

Figure 52. Solubility of the C_3 hydrocarbons in DMF at infinite dilution
a) Propyne; b) Propadiene; c) Propene; d) Propane

The equilibrium between the two isomers of C_3H_4 is reached in the presence of catalysts (for example $Al_2O_3/_3$, SiO_2, activated carbon, and γ-Al_2O_3/Na_2CO_3) [130, 131]. At 270 °C the equilibrium mixture contains 82 % propyne, and at 5 °C, 91.1 % propyne. For calculated data, see [132]. This equilibrium is important for the industrial propyne recovery process.

The decomposition pressure as a function of the ignition energy for propyne (and propadiene) is plotted in Figure 38. The lower and upper flammability limits of propyne in air are about 2.3 and 16.8 vol %. For propadiene only ranges are available: 1.7 – 2.5 and 12 – 17 vol % [133].

Production. Propyne and propadiene can be recovered from cracked gas by solvent extrac-

Table 19. Solubility coefficients of propyne in various solvents (mol kg^{-1} bar^{-1})[9]

Solvent	C_3H_4 pressure, bar	− 20 °C	25 °C
Methanol	0.1	9.099	1.865
1,2-Dichloroethane	0.098 – 0.196	3.276 (0 °C)	1.546
Carbon tetrachloride	0.25	4.732	0.842
n-Octane	0.098	3.412	
Toluene	0.25/1.0	8.644	2.047
Xylene (technical).	0.49/0.98	9.782	1.683
4-Methyl-1,3-dioxolan-2-one	0.49/1.0	8.644	1.183
Triethylene glycol	1.0		0.400 (30 °C)
Acetone	0.6/1.0	35.03	4.186
N-Methyl-2-pyrrolidone	$\leq$ 0.78	6.597 (0 °C)	2.502
DMF	$\leq$ 0.15	12.512	3.003
Water	1.0		0.071
Ammonia	$\leq$ 0.04	20.018	4.436

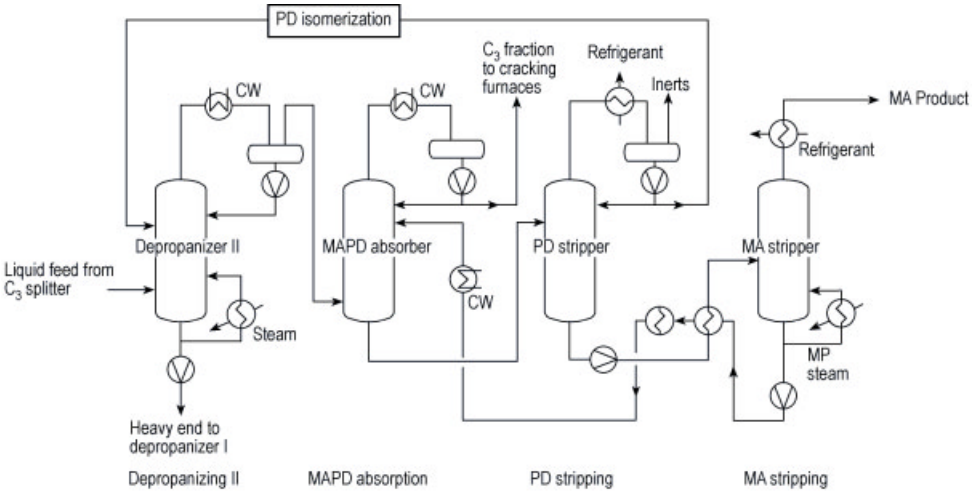

Figure 53. Propyne recovery process (Linde, Shell [77])
MA: methylacetylene; PD: propadiene

tion. The process is outlined in Figure 53 [79, 134] and is similar to the recovery of acetylene (see Chap. 4.4.2). This process was developed by Linde and Shell and is ready for use [70]. Nevertheless, until today (2006) it has not been used in an industrial application. An important step is the catalytic isomerization of propadiene to propyne in the liquid phase on a catalyst [131].

The bottom product from the propene/propane splitter is routed to depropanizer II for removal of traces of C_{4+} hydrocarbons originating from the feed and the isomerization reactor effluent. The overhead is sent to the absorption column. The propane (and some propene) is routed back to the cracking furnace. Propyne (MA) and propadiene (PD) are stripped off in the corresponding columns. The propadiene fraction is converted to propyne in the isomerization reactor and recycled to the feed. Traces of solvent in the pure propyne are removed by cooling. The product can be sent directly to methyl methacrylate (MMA) unit, where MMA is manufactured by reaction of propyne with carbon monoxide and an alcohol in the presence of a Pd-based carboxylation catalyst [131, 135].

Usually propyne and propadiene are undesirable byproducts of the steam cracking process. The crude C_3 fraction can contain up to 6 mol % MAPD, whereas the MAPD content of chemical-grade propene typically has to be $<$ 20 ppm and that of polymerization-grade propene $<$ 2 ppm. To this end MA and PD are removed by selective hydrogenation in the liquid or gas phase. Typical conditions are pressures of 1.5 – 5 MPa and inlet temperatures of 20 – 100 °C. The vapor-phase processes are declining in

importance. Today mainly liquid or composite phase are in use as they have some advantages:

- Operation at lower temperatures (higher selectivity)
- Lower operating costs (no vaporization of the feed and recondensation of the product)
- Regeneration of reactor in situ (removal of polymers)
- Lower frequency of catalyst regeneration
- For high concentration of C_3H_4 in the feed, two reactors are often sufficient; the gas-phase process requires more reactors
- Control of conversion by injection of hydrogen (no excess hydrogen) reduces investment cost
- Control of temperature by use of vaporization minimizes risk of runaway

Table 18. Physical properties of propyne [5, 6, 129]

Molecular mass	40.065
Critical temperature	402.39 K (129.24 °C)
Critical pressure	5.626 MPa
Critical volume	0.1635 m^3/kmol
Melting point	170.45 K (− 102.7 °C)
Dipole moment	2.61×10^{-30} C · m
Density (liquid)	638.92 kg/m^3 (273.15 K)
Normal boiling point	249.97 K (− 23.18 °C)
Enthalpy of vaporization (273.15 K)	20.765 kJ/mol (273.15 K)
Molar volume	0.05962 m^3/kmol (249.91 K)
Enthalpy of formation	185.5 ± 1.0 kJ/mol (298.15 K)
Gibbs free energy of formation	193.8 ± 1.0 kJ/mol (298.15 K)
Entropy of formation	248.4 J mol^{-1} K^{-1} (298.15 K)
Enthalpy of combustion	1938.943 kJ/mol (298.15 K)
Heat capacity (constant pressure)	59.842 J mol^{-1} K^{-1} (273.15 K)
Viscosity (liquid)	1.7500×10^{-4} Pa · s (273.15 K)
Viscosity (gaseous)	8.3300×10^{-6} Pa · s (293.15 K)
Thermal conductivity (liquid) *	0.14560 W m^{-1} K^{-1} (233.45 K)
Thermal conductivity (gas) *	0.014310 W m^{-1} K^{-1} (273.15 K)
Surface tension *	1.47×10^{-2} N/m (273.15 K)

* Predicted or estimated.

As in the case of acetylene hydrogenation, Pd catalysts are used. Several catalyst manufacturers offer special catalysts for C_3 hydrogenation. The selectivity in case of liquid-phase hydrogenation to propene is about 60 – 70 % and depends on the concentration of C_3H_4 in the C_3 feed. For better control of reaction heat, some product is normally recycled to the feed. The cycle time of the catalyst is very sensitive to contaminants (and to byproduct formation in gas-phase processes), especially in the first reactor. A reactivation procedure is necessary when the catalyst loses its activity. In contrast to acetylene hydrogenation, the requirements on MAPD hydrogenation are not that strict. MAPD going to the propane - propylene separation (C_3 splitter) mainly ends up in the bottom product propane which is usually recycled to the cracking furnaces. This means that there is another purification step, which can remove MAPD from the propene product. As propyne and propadiene are different chemical compounds they are hydrogenated with different reaction rates. In contrast, many reports in literature about hydrogenation only consider a simplified model with a combined MAPD fraction and do not discriminate between both isomers.

Stabilized Propyne–Propadiene Mixtures. As a replacement for acetylene, stabilized propyne–propadiene mixtures are available commercially. Trade names are Tetrene or MAPP gas. These mixtures are stabilized by propane, propene, and/or butane and are used for metal cutting, welding, hardening, and brazing. The flame properties are closer to those of propane–propene mixtures. Therefore stabilized C_3H_4 mixtures have not yet won a large-scale market. Further information is available in [136]. In Europe, those mixtures are available only very rarely. In Europe, mixtures with the trade name MAPP gas have a rather low MAPD content actually and mainly consist of propene, propane, and butane. Only in the US the original MAPP mixture has some importance.

8. Toxicology and Occupational Health

Pure acetylene is a simple asphyxiant. When generated from calcium carbide, acetylene is frequently contaminated with arsine, hydrogen sulfide, or phosphine, and exposure to this impure acetylene has often resulted in serious consequences. Commercial acetylene no longer contains these impurities and is therefore less harmful [137].

The lowest published lethal concentration for rats is 9 vol % [138]. Dogs are less sensitive: 80

vol % acetylene in the air is necessary to produce a narcosis accompanied by an increased blood pressure and a decreased pulse frequency (stimulation of vasomotor and vagus centers) [139]. In humans, the inhalation of air containing 10 vol % acetylene has a slight intoxicating effect, marked intoxication occurs at 20 vol %, incoordination at 30 vol %, and unconsciousness within 5 min on exposure to 35 vol %. Inhaling 35 vol. % for 5 – 10 min or 10 vol % for 30 – 60 min is lethal. Symptoms of intoxication are excitement, coma, cyanosis, weak and irregular pulse, and memory failure [140 – 142].

There is no evidence that *repeated exposure* to tolerable levels of acetylene has effects deleterious to health [143]. Inhalation of air with 33 vol % of acetylene by humans led to unconsciousness within 6 min, but when the experiment was repeated within the week the susceptibility to acetylene decreased: 9 min were required on the second exposure and more than 33 min on the third exposure to produce unconsciousness [139].

Acetylene does not irritate the mucous membranes [137]. Neither threshold limit value (TLV) nor a MAK has been established. The standard air concentration allowed for OSHA and NIOSH is 2500 ppm [141].

9. References

1. *Acetylene or Ethylene as Feedstocks for the Chemical Industry*, Conference proceedings Dechema and SCI, Frankfurt/Main, 27 – 29 March 1968, SCI, London 1968.
   E. Schenk: "Acetylen oder Äthylen als Rohstoffe der chemischen Industrie," DECHEMA, Frankfurt/Main 1968.
2. O. Horn, *Erdöl Kohle Erdgas Petrochem.* **26** (1973) no. 3, 129.
3. Yen-Chen Yen, *Acetylene*, Report Nr. 16, Supplement A, Stanford Research Inst., Menlo Park, Calif., Nov. 1981, and Information of the VCI, Frankfurt/Main.
4. D. A. Plattner, Y. Li, K. N. Houk: "Modern Computational and Theoretical Aspects of Acetylene Chemistry", Chapter 1.2. in P. J. Stang, F. Diederich (eds.): *Modern Acetylene Chemistry*, VCH Verlagsgesellschaft, Weinheim, Germany 1995.
5. DIPPR Database, STN International, Design Institute for Physical Property Data c/o DIPPR Project Staff, Pennsylvania State University 167 Fenske Lab, University Park, Pa 16802 U.S.A American Institute of Chemical Engineers, 3345 E. 47th Street, New York, NY 10017, U.S.A., basis February 1998.
6. Beilstein Database, STN International, Beilstein Informationssysteme Carl-Bosch-Haus, Varrentrappstr. 40–42, D-60486 Frankfurt am Main, Germany, basis February 1998.
7. W. Wiechmann, *Amts- und Mitteilungsblatt der Bundesanstalt für Materialforschung und -prüfung (BAM)* (1987) 3, 505.
8. Linde AG: Spezialgasekatalog, Höllriegelskreuth 2006.
9. A. Kruis: "Gleichgewicht der Absorption von Gasen in Flüssigkeiten," in H. Hausen (ed.): *Landolt-Börnstein Zahlenwerte und Funktionen*, **4. Teil, Bestandteil c**, Springer Verlag, Berlin–Heidelberg–New York 1976.
10. S. A. Miller: *Acetylene*, Its Properties, Manufacture and Uses, Ernest Benn, London 1965.
11. F. Rottmayr, H. Reimann, U. Lorber, *Linde Ber. Tech. Wiss.* **30** (1971) 3.
12. H. Schmidt, D. Forney: *Oxygen Technology Survey*, vol. **9**, NASA SP-3090, ASRDI, 1975.
13. E. Lassmann, *Linde Reports on Science and Technology* **57** (1996) 16.
14. UCC isotherm data sheet No. 44, Union Carbide Corp., New York 1980.
15. D. P. Velenuela, A. L. Myers: *Adsorption Equilibrium Data Handbook*, Department of Chem. Engineering, University of Pennsylvania, Philadelphia, Prentice Hall, Englewood Cliffs, NJ 1989.
16. T. F. Rutledge: *Acetylenic Compounds–Preparation and Substitution Reactions*, Reinhold Publ. Co., New York 1968.
17. A. C. Mc Kinnis, *Ind. Eng. Chem.* **27** (1962) 2928.
18. S. A. Miller: *Acetylene*, Its Properties, Manufacture and Uses, Ernest Benn, London 1965.
19. R. C. West, C. S. Kraihanzel, *J. Am. Chem. Soc.* **83** (1961) 765.
20. H. J. Copley, C. E. Holley Jr., *J. Am. Chem. Soc.* **61** (1939) 1599.
21. W. Reppe: *Neue Entwicklungen auf dem Gebiet der Chemie des Acetylens und Kohlenoxids*, Springer Verlag, Berlin-Göttingen-Heidelberg 1949.

22. W. Reppe: *Chemie und Technik der Acetylen-Druckreaktionen*, 2nd ed., Verlag Chemie, Weinheim, Germany 1952.
23. N. von Kutepow in: *Ullmann*, 4th ed., vol. **7**, p. 44.
24. W. Reppe, *Justus Liebigs Ann. Chem.* **601** (1956) 81.
25. W. Reppe et al., *Justus Liebigs Ann. Chem.* **569** (1955) 1.
26. W. Reppe et al., *Justus Liebigs Ann. Chem.* **582** (1953) 1.
27. W. Reppe, N. von Kutepow, A. Magin, *Angew. Chem.* **81** (1969) 717.
28. Lonza, DE 1251329, 1964 (P. Pino, G. Braca, G. Sbrana).
29. BASF, DE-AS 1071077, 1965 (W. Reppe, A. Magin).
Du Pont, DE 1054086, 1955 (J. C. Sauer).
J. C. Sauer, R. D. Cramer, V. A. Engelhardt, T. A. Ford, H. E. Holmquist, B. W. Howk, *J. Am. Chem. Soc.* **81** (1959) 3677.
G. Albanesi, M. Toraglieri, *Chim. Ind. (Milan)* **41** (1959) 189.
30. G. Palyi, G. Varadi, J. T. Horvath, *J. Mol. Catal.* **13** (1981) 61.
31. W. Reppe et al., *Justus Liebigs Ann. Chem.* **560** (1948) 1.
32. G. Schröder: *Cyclooctatetraen*, Verlag Chemie, Weinheim, Germany 1965.
33. J. A. Nieuwland, W. S. Calcott, F. B. Downing, A. S. Carter, *J. Am. Chem. Soc.* **53** (1931) 4197.
34. J. C. W. Chien, *Polym. News* **6** (1979) 52.
D. Bloor, *New Sci.* **93** (1982) no. 1295, 577.
H. Shirakawa, *Kotai Butsuri* **16** (1981) no. 7, 402;
Chem. Abstr. **96** (1982) 172568.
A. J. Heger, A. G. Macdiarmid, *Int. J. Quantum Chem. Quantum Chem. Symp.* **15** (1981) 243.
H. Shirakawa, *Kagaku (Kyoto)* **37** (1982) no. 3, 181;
Chem. Abstr. **96** (1982) 153398.
H. Shirakawa, *Kagaku, Zokan (Kyoto)* 1980, no. 87, 165;
Chem. Abstr. **94** (1981) 139105 x.
35. G. Wegner, *Angew. Chem.* **93** (1981) 352.
36. T. Mole, J. R. Suertes, *Chem. Ind. (London)* 1963, 1727.
37. P. N. Rylander, "Platinum Metal Catalysts in Organosilicon Chemistry," *Engelhard Ind. Tech. Bull.* **10** (1970) no. 4, 130.
38. J. L. Speier, *Adv. Organomet. Chem.* **17** (1979) 407.
39. E. Ott et al., *Ber. Dtsch. Chem. Ges.* **76** (1943) 80 – 91.
40. I. G. Farbenind., DE 495787, 1927.
41. *Chem. Week* **98** (1966) Apr. 16, 90.
42. *Eur. Chem. News* **25** (1974) Feb. 1, 5.
43. H. Sachsse, *Chem. Ing. Tech.* **26** (1954) 245.
E. Bartholome, *Chem Ing. Tech.* **26** (1954) 253.
T. P. Forbath, B. I. Gaffney, *Pet. Refiner* **33** (1954) 160.
44. H. Friz, *Chem. Ing. Tech.* **40** (1968) 999.
45. K. G. Baur, K. Taglieber, *Chem. Ing. Tech.* **47** (1975) 385.
46. M. J. Zundel: *Problèmes de Fabrication d'Acetylène et d'Ethylène à partir d'Hydrocarbures*, Societe de Chimie Industrielle, Paris 1962.
47. *Hydrocarbon Process. Pet. Refiner* **44** (1965) no. 11, 163.
48. Montecatini, GB 1000480, 1962.
49. J. L. Petton et al., *Pet. Refiner* **37** (1958) 180.
50. *Hydrocarbon Process.* **54** (1975) no. 11, 104.
Stanford Research Report no. 109, September 1976. L. Verde, R. Riccardi, S. Moreno, *Hydrocarbon Process.* **57** (1978) no. 1, 159.
51. G. Hamprecht, M. Gettert: "Sauerstoff-Thermisches Calziumcarbid," Festschrift für Carl Wurster, BASF, Ludwigshafen 1960, p. 43.
52. J. Schulze, M. Homann, "Die mögliche Stellung des Acetylens in der zukünftigen Kohlechemie," *Erdöl Kohle Erdgas Petrochem.* **36** (1983) no. 5, 224.
53. H. Gladisch, *Hydrocarbon Process. Pet. Refiner* **41** (1962) 159 – 164.
54. H. Gladisch, *Chem. Ing. Tech* **61** (1969) 204 – 208.
55. K. Gehrmann, H. Schmidt, *Water Air Conserv. Pet. Ind*, 1971, 379.
56. R. L. Bond, W. R. Ladner, G. I. McCommet, *Fuel* **45** (1966) 381 – 395.
R. E. Gannon, V. J. Krukonis, Th. Schoenberg, *Ind. Eng. Chem. Prod. Res. Dev.* **9** (1970) 343 – 347.
R. E. Gannon, S. K. Ubhayakar, *Fuel* **56** (1977) 281.
AVCO Corp., US 179144, 1980 (Ch. Kim).
D. Bittner, H. Baumann, C. Peuckert, J. Klein, H. Jüntgen, *Erdöl Kohle Erdgas Petrochem.* **34** (1981) no. 6, 237 – 242.
R. Müller, C. Peuckert, *Int. Symp. on Plasma Chem.*, 5th, Edinburgh 1981, p. 197 – 202.
57. *Eur. Chem. News* **36** (1981) Feb. 2, 19.
58. R. Müller, *World Hydrogen Energy Conf.*, Conf. Proc., 4th, Pasadena, 1982, p. 885 – 900.
59. GAF, US 219756, 1980 (M. Katz, F. Carluccio).

60. *Verordnung über Acetylenanlagen und Calcium-carbidlager* (Acetylenverordnung–Acet V) vom 27. Februar 1980, Carl Heymanns Verlag, Köln 1980.
61. *TRAC 201 – Acetylenentwickler*, Carl Heymanns Verlag, Köln.
62. Acetylene Transmission for Chemical Synthesis (Recommended Minimum Safe Practices for Piping Systems) International Acetylene Association, New York 1980.
63. C. Hase, W. Reitze: *Fachkunde des Autogenschweißens*, 7th ed., Girardet, Essen 1965.
64. *Carbidkalk*, Hinweise für seine Verwendung, Fachbuchreihe Schweißtechnik, 9th ed., Deutscher Verlag für Schweißtechnik (DVS), Düsseldorf 1968, p. 16.
65. Shawinigan Chemicals, US 1872741, 1931 (R. S. Jane).
66. Bayerische Stickstoff-Werke, DE 714 323, 1938 (R. Wendlandt, R. Neubner).
67. Shawinigan Chemicals, US 1343185, 1942 (A. C. Holm, E. Poirier).
68. Linde, DE 906 005, 1951 (F. Rottmayr). Linde, DE 2 549 399, 1975 (E. Laßmann).
69. *Hydrocarbon Process.* **46** (1967) no. 11, 139. T. Wett, *Oil Gas J.* **70** (1972) 101–110.
70. *Ullmann*, 4th ed., vol. **7.**
71. H. K. Kamptner, *Erdöl Kohle Erdgas Petrochem.* **16** (1963) 547. H. K. Kamptner, W. R. Krause, H. P. Schilken, *Hydrocarbon Process. Pet. Refiner* **45** (1966) no. 4, 187. K. Lissa, *Chem. Anlagen Verfahren* 1970, no. 6, 83.
72. Y. C. Hu, *Hydrocarbon Process.* **61** (1982) no. 11, 109.
73. E. L. Mohundro: Overview on C2 and C3 Selective Hydrogenation in Ethylene Plants, Presentation at the AIChE 15th Ethylene Producers Conference 2003 Spring National Meeting, New Orleans, LA 2003.
74. F. Mey, H. D. Neubauer, R. Schubert, *Petrochemicals and Gas Processing Petroleum Technology quaterly* (1997) Autumn, 119.
75. K. J. Sasaki, *Petrochemicals and Gas Processing Petroleum Technology quaterly* (1997) Autumn, 113.
76. A. Borodzinski, G. C. Bond, *Catal. Rev.* **48** (2006) 91–144.
77. P. Cl. Haehn, Dr. E. Haidegger, Dr. N. Schödel, *Hydrocarbon Engineering* (1997) January/February, 41.
78. Linde München folder C/3.3.e/93
79. D. Sohns, *Linde Reports Science and Technology* **30** (1979) 21.
80. G. Marcks, *Schadenprisma* **3** (1982) 37. D. Lietze, *Amts- und Mitteilungsblatt der Bundesanstalt für Materialprüfung (BAM)* **16** (1986) 1, 23.
81. W. Reppe, *Chem. Ing. Tech.* **22** (1950) 273.
82. CEN/TC 121/SC 7/WG3 German Proposal: *Pipelines in Acetylene Systems,*November 1996, NAS-DIN Deutsches Institut für Normung e.V., D 10772 Berlin or *"Empfehlungen für Acetylenleitungen auf der Grundlage von Arbeitsbereichen",* Industrial Gases Committee IGC Document 9/78/D, F 75880 Paris, CEDEX 18
83. H. B. Sargent, *Chem. Eng.* **64** (1957) 250.
84. Th. Schendler, H.-P. Schulze, *Chem.-Ing.-Tech.* **62** (1990) 1, 41. W. Reppe: *Chemie und Technik der Acetylen-Druck-Reaktionen*, Verlag Chemie, Weinheim, Germany 1951.
85. D. Lietze, H. Pinkofsky, T. Schendler, H.-P. Schulze, *Chem.-Ing.-Tech.* **61** (1989) 9, 736. H. Große-Wortmann, N. Kalkert, H.-G. Schecker, *Chem.-Ing.-Tech.* **53** (1981) 461.
86. A. Williams, D. B. Smith, *Chem. Rev.* **70** (1970) 267. B. A. Ivanov, S. M. Kogarko, *Int. Chem. Eng.* **4** (1964) 4, 670. C. M. Detz, *Combust. Flame* **34** (1979) 187. M. A. Glikin et al., *Sov. Chem. Ind. (Engl. Transl.)* **7** (1975) 1373. A. Baumeier, D. Conrad, S. Dietlen, W. Pezold, T. Schendler und H.-P. Schulze, *Chem.-Ing.-Tech.* **64** (1992) 3, 260.
87. R. Grätz, M. Wagenknecht, *vfdb, Zeitschrift für Forschung, Technik und Management im Brandschutz* (1994) **3**, 103.
88. D. Lietze, *Chem. Ing.-Tech.* **63** (1991) 11, 1148.
89. D. Lietze, *Chem.-Ing.-Tech.* **62** (1990) 3, 238.
90. E. Karwat, *Chem Eng. Progr.* **54** (1958) 10, 96.
91. D. Lietze, *Journal of Hazardous Materials.* **54** (1997) 227.
92. D. Lietze, *J. Loss Prev. Process Ind.* **8** (1995) 6, 319.
93. D. Lietze, *J. Loss. Process Ind.* **8** (1995) 6, 325.
94. A. D. Hewitt: Technology of Oxy-Fuel Gas Processes, Welding and Metal Fabrication, Part 2, 1972 (November) 382. H. Meinass, H. Manhard, J. Schlander, L. Fruhstorfer, *Linde Reports on Science and Technology* **24** (1976) 32. L. Kögel, *Linde Reports on Science and Technology* **32** (1981) 36.

95. U.S. Chemical Safety and Hazard Investigation Board (CSB), Safety Bulletin No. 2005-03-B, January 2006
96. Winnacker-Küchler, 5th ed., vol. 4, p. 899.
97. *Technische Regeln für Acetylenanlagen und Calciumcarbidlager*, TRAC 201 Acetylenentwickler (1973 i.d.F. von 1990) TRAC 202 Acetylenkühler, -trockner und -reiniger (1976 i.d.F. von 1990) TRAC 203 Acetylenverdichter (1974 i.d.F. von 1990) TRAC 204 Acetylenleitungen (1990) TRAC 205 Acetylenspeicher (1973 i.d.F. 1990) TRAC 206 Acetylenflaschenbatterieanlagen (1988 i.d.F. von 1994) TRAC 207 Sicherheitseinrichtungen (1979 i.d.F. von 1999) TRAC 208 Acetyleneinzelflaschenanlagen (1988 i.d.F. von 1999) TRAC 209 Anlagen zur Herstellung und Abfüllung von unter Druck gelöstem Acetylen (Acetylenwerke, Dissousgaswerke) (1982) Deutscher Acetylenausschuß, Bundesanstalt für Arbeitsschutz und Unfallforschung, Dortmund 1969.
98. K. H. Roch, *Amts Mitteilungsbl. Bundesanst. Materialprüf. Berlin* **12** (1982) 283.
99. K. H. Roch, *Schweissen + Schneiden* **25** (1973) 94.
100. D. Lietze, *Amts Mitteilungsbl. Bundesanst. Materialprüf. Berlin* **2** (1972) 9.
101. D. Lietze, *Berufsgenossenschaft* 1976, 435.
102. M. E. Sutherland, M. W. Wegert, *Chem. Eng. Prog.* **69** (1973) no. 4, 48.
103. C. Isting, *Erdöl Kohle Erdgas Petrochem.* **23** (1970) 29.
104. *Chem. Eng. (N.Y.)* **76** (1969) no. 1, 89.
105. E. Karwat, *Chem. Eng. Prog.* **53** (1957) no. 4, 27.
E. Karwat, *Linde Ber. Tech. Wiss.* **13** (1962) 12; "Safety in Air and Ammonia Plants," *Chem Eng. Prog. Tech. Man.* **5** (1963) 43.
G. Klein in: "Luftzerlegungsanlagen", Linde-Arbeitstagung, München 1975.
106. L. W. Coleman: "Safety in Air and Ammonia Plants," *Chem. Eng. Prog. Tech. Man.* **4** (1962) 26.
107. E. Karwat, *Chem. Eng. Prog.* **57** (1961) no. 4, 5.
108. F. G. Kerry, *Chem. Eng. Prog.* **52** (1956) no. 11, 3.
109. J. Reyhing, *Linde Rep. Sci. Technol.* **36** (1983) 14.
110. Engelhard Ind., DE-AS 1283805, 1968 (J. G. Cohn A. J. Haley, Jr.).
111. G. Klein, *Linde Ber. Tech. Wiss.* **17** (1964) 24.
Cryogenics Safety Manual, Part II, British Cryogenics Council, London 1970. H. H. Hofmaier: "Safety in Air and Ammonia plants", Chem Eng. Prog. Tech. Man. 5 (1963) 22.
112. E. Karwat, *Chem. Eng. Prog. Tech. Man.* **2** (1960) A-18.
113. G. Drewes, M. Ermscher, *Chem. Tech. (Leipzig)* **35** (1983) 57.
114. K. H. Möller, C. Stöber, K. Schulze, *Arbeitsschutz* **23** (1972) no. 1, 18.
115. P. Hölemann, R. Hasselmann, *Forschungsberichte des Landes Nordrhein-Westfalen*, no. 765, Westdeutscher Verlag, Köln-Opladen 1959.
116. P. Hölemann, *Forschungsberichte des Landes Nordrhein-Westfalen*, no. 888 and 1151, Westdeutscher Verlag, Köln-Opladen 1960 and 1963.
117. K. H. Möller, *Arbeitsschutz* **22** (1971) no. 1, 6.
118. "Technische Grundsätze für ortsbewegliche Druckgasbehälter," Ziffer 29 and 31, *Arbeitsschutz* **21** (1970) no. 3."Allgem. Verwaltungsvorschrift zu 14, 17 – 19 der Verordnung über ortsbewegliche Behälter und über Füllanlagen für Druckgase," 20. 6. 1968.
119. Department of Transportation, Office of Hazardous Materials Regulations, Code of Federal Regulations 49.
120. K. H. Möller, *Arbeitsschutz* **23** (1972) no. 2, 30.
121. International Organization for Standardization–ISO/TC 58/WG 1.
122. K. H. Möller, *Berufsgenossenschaft* 1972, 375, 422.
123. H. Springmann, *Linde Rep. Sci. Technol.* **34** (1982) 54.
124. E. Zorn, *Mitt. BEFA* **14** (1963) no. 6, 2.
125. *Safety in the Production and Use of Acetylene*, Commission permanente internationale de l'acetylène, Paris 1968.
126. A. D. Hewitt: *Welding and Metal Fabrication*, part 1, Oct. 1972, 347; part 2, Nov. 1972, 382; part 3, Dec. 1972, 416, IPC Science and Tech. Press, Guildford, England.
127. F. Houldcroft: *Welding Processes*, Oxford University Press, London 1975.
J. Ruge: *Handbuch der Schweißtechnik*, Springer Verlag, Berlin 1980.
Bibliographies, American Welding Soc., New York (appears annually).
128. H. Weiler, *Schweißen + Schneiden* **26** (1974) 220.
L. Kögel, *Linde Ber. Tech. Wiss.* **48** (1980) 36.

129. Engineering Sciences Data Unit, "London Vapour Pressures and Critical Points of Liquids", Part 2C: alkadienes and alkynes, Number 86001 (1986).
130. J. F. Cordes, H. Günzler, *Chem. Ber.* **92** (1959) 1055.
J. F. Cordes, H. Günzler, *Z. Naturforschung* **15b** (1960) 682.
C. P. Khulbe, R. S. Mann, *Prep.-Can. Symp. Catal., 5*th (1977) 384.
P. Kos, I. Kiricsi, K. Varga, P. Fejes, *Acta Phys. et. Chem. Szeged* **33** (1987) 109.
F.-D. Zeiseler, G. Zimmermann, *Journal für Prakt. Chemie* **319** (1977) 4, 655.
131. Shell Internationale, EP 0 392 601, 1990 (M. J. Doyle, J. Van Gogh, J. C. Van Ravenswaay Claasen).
132. D. A. Frank-Kamenetzki, V. G. Markovich, *Acta Physicochim, (U.R.S.S.)* **17** (1942) 308.
H. Zeise (ed.): *Thermodynamik*, vol. **III/1**, Verlag Hirzel, Leipzig 1954.
133. T. Redeker, G. Schön, *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe*, 6. Nachtrag Deutscher Eichverlag GmbH, Braunschweig 1990.
134. Linde AG, US 4 701 190, 1986 (P. C. Haehn), EP 0 224 748, 1986 (P. C. Haehn).
135. J. Happel, S. Umemura, Y. Sakakibara, H. Blanck, S. Kunichika, *Ind. Eng. Chem. Process Des. Develop.* **14** (1975) 1, 44.
E. Drent, P. Arnoldy, P. H. M. Budzelaar, *Journal of Organometallic Chemistry*, **455** (1993) 247.
Shell Internationale, EP 0 539 628, 1993 (J. Hengeveld, P. B. de Blank).
136. R. F. Huston, C. A. Barrios, R. A. Holleman, *J. Chem. Eng. Data* **15** (1970) 168.
R. D. Green, *Met. Prog.* **108** (1975) 2, 71.
A. Farwer, *Gas Aktuell* (1982 Nov.) 24.
137. L. T. Fairhall: *Industrial Toxicology*, 2nd ed., The Williams and Wilkins Co., Baltimore 1957, p. 270.
138. The International Technical Information Institute: *Toxic and Hazardous Industrial Chemicals Safety Manual*, Tokyo, 1975, p. 10.
139. T. Sollmann: *A Manual of Pharmacology and its Application to Therapeutics*, 7th ed., W. B. Saunders Co., Philadelphia-London 1949, p. 662.
140. F. Flury, F. Zernik: *Schädliche Gase*, Springer Verlag, Berlin 1931, p. 270.
141. NIOSH, U.S. Dept. of Health and Human Services: Registry of Toxic Effects of Chemical Substances, vol. 1, Washington, DC, 1980, p. 70.
142. E. R. Plunkett: *Handbook of Industrial Toxicology*, Heyden, Barberton, Ohio, 1976, p. 8.
143. F. A. Patty: *Industrial Hygiene and Toxicology*, 2nd ed., Interscience Publ., New York 1962, p. 1205.

Acridine

GERD COLLIN, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany

HARTMUT HÖKE, Weinheim, Federal Republic of Germany

1. Physical Properties 1
2. Chemical Properties 1
3. Production 1
4. Uses 1
5. Toxicology 2
6. References 2

Acridine [*260-94-6*], $C_{13}H_9N$, was discovered in coal-tar anthracene oil by C. GRAEBE and H. CARO in 1870, and identified as dibenzopyridine by C. RIEDEL, A. BERNTHSEN, and F. BENDER in 1883.

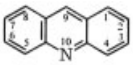

1. Physical Properties

M_r 179.22, *mp* 110 °C, *bp* 345 °C (at 1013 hPa), ϱ at 20 °C 1.1005 g/cm^3. Colorless needles or prisms; volatile in steam; slightly soluble in boiling water, readily soluble in organic solvents such as alcohol, ether, carbon disulfide, and benzene; solutions show blue fluorescence.

2. Chemical Properties

Acridine is a weak tertiary base. With strong acids, it forms crystalline, yellow salts that decompose readily in boiling water. Acridine is stable when heated with concentrated alkali or hydrochloric acid. Nitration with nitric acid yields mainly 2- and 4-nitroacridine and small amounts of dinitroacridines. Reduction gives acridane (9,10-dihydroacridine [*92-81-9*]) preferentially. Oxidation results in the formation of acridinic acid (quinoline-2,3-dicarboxylic acid [*643-38-9*]) or acridone, depending on the oxidizing agent used. Acridine forms quaternary acridinium salts with alkyl or aryl halides and sulfates.

3. Production

Acridine can be separated as sodium acridone sulfonate from the coal-tar anthracene oil boiling between 300 and 360 °C (→ Tar and Pitch) by extraction with dilute sulfuric acid [5] or with aqueous sodium bisulfite. The free base is obtained by decomposition of the salt with caustic soda [6]. Acridine can be synthesized by reduction of acridone or 9-chloroacridine.

Acridone [*578-95-0*], pale yellow needles,

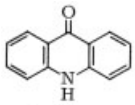

Acridone

mp 354 °C, is obtained by the cyclization of diphenylamine-2-carboxylic acid (phenylanthranilic acid), which is synthesized from aniline and 2-chlorobenzoic acid. *9-Chloroacridine* [*1207-69-8*], almost colorless crystals, *mp* 120 °C, is obtained by reacting phosphoryl chloride with diphenylamine-2-carboxylic acid.

4. Uses

Acridine as such has no commercial significance. In the patent literature it is proposed as an antioxidant, as a polymerization and corrosion inhibitor, as an additive to peroxidic vulcanizing agents for ethylene – propylene rubber, as a thermal stabilizer for polyolefins, and also as an occult blood test reagent, used in paper strips for blood tests. The acridine dyes are made from other starting materials (→ Triarylmethane and Diarylmethane Dyes, Chap. 8.).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_147

5. Toxicology

Animal studies have shown that acridine is slightly toxic: the LD_{50} value is about 2000 mg/kg (rat, oral) [7], [8]. Industrial observations show that exposure to acridine dust or vapor causes strong irritations of the skin and mucous membranes combined with sneezing, with itching or even inflammation of the skin. Some acridine derivatives cause sensitization of the skin, especially on exposure to light. Exposure limits at the workplace (MAK, TLV) are not specified. Resorptive intoxication is unknown.

6. References

General References

1. *Beilstein* **20,** 459; **20 (2),** 171.
2. A. Adrien: *The Acridines,* 2nd ed., St. Martin's Press, New York 1966.
3. R. M. Acheson (ed.): *Acridines. Chemistry of Heterocyclic Compounds,* 2nd ed., **vol. 9,** J. Wiley & Sons, New York 1973.
4. N. Campbell in S. Coffey, M. F. Ansell (eds.): *Rodd's Chemistry of Carbon Compounds,* 2nd ed., **vol. 4 (G),** Elsevier, Amsterdam 1978, pp. 1 – 82.

Specific References

5. Rütgerswerke, DE 688335, 1938 (N. Hviid).
6. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung,* Verlag Glückauf, Essen 1951, p. 191.
7. W. S. Spector: *Handbook of Toxicology,* **vol. 1,** W. B. Saunders Co., Philadelphia – London1956, p. 12.
8. V. B. Kapitul'skii, *Klin. Patog. Profil. Profzabol. Khim. Etiol. Predpr. Tsvetn. Chern. Metall* 1969, no. 2, 179 – 183.

Acridine Dyes → Triarylmethane and Diarylmethane Dyes

Acrolein and Methacrolein

Dietrich Arntz, Degussa Grasse, France
Achim Fischer, Degussa AG, Hanau, Germany
Mathias Höpp, Degussa AG, Frankfurt, Germany
Sylvia Jacobi, Degussa AG, Hanau, Germany
Jörg Sauer, Degussa AG, Marl, Germany
Takashi Ohara, Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
Takahisa Sato, Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
Noboru Shimizu, Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
Helmut Schwind, Degussa AG, Hanau, Germany

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
3. Production 6
3.1. Acrolein by Propene Oxidation 6
3.2. Methacrolein 9
4. Quality and Analysis 10
5. Handling, Storage, and Transportation 10
6. Uses and Production Data 11
7. Toxicology and Ecotoxicology 12
7.1. Toxicology 12
7.2. Ecotoxicology 14
8. References 14

Acrolein is the simplest unsaturated aldehyde. It is an important intermediate in the production processes of a variety of substances. The main use of isolated acrolein is in the production of D,L-methionin, but in far more production processes, acrolein is used without purification. Acrolein is an extremely reactive chemical and, in very low concentrations, acts as a very effective broad-spectrum biocide. Methacrolein is produced commercially as an organic intermediate, which is used as building block for chain expansions or, in limited application, in the synthesis of flavors and fragrances.

1. Introduction

Acrolein [*107-02-8*], propenal, acrylaldehyde, CH_2 =CH-CHO, the simplest unsaturated aldehyde, is a colorless, volatile, toxic, and lacrimatory liquid with a powerful odor.

Acrolein (Latin: acer - acrid, oleum = oil) was discovered 1843 by Redtenbacher in overheated fat [1].

The commercial production of acrolein by heterogeneously catalyzed gas-phase condensation of acetaldehyde and formaldehyde was established by Degussa in 1942. Today, acrolein is produced on a large commercial scale by heterogeneously catalyzed gas-phase oxidation of propene.

Acrolein is an important intermediate for numerous substances (see Chap. 6). The main use of commercial, isolated acrolein is the production of D,L-methionine [2], an essential amino acid used as an animal feed supplement. In the production of acrylic acid, acrolein is not isolated from the gas-phase reaction mixture but is oxidized further on a heterogeneous catalyst. The produced amount of nonisolated acrolein exceeds the amount of isolated acrolein by far. Less important is the production of special chemicals from acrolein.

Several review articles [3 – 9] (see also [92]) and a monograph [10] describe the preparation, reactions, and uses of acrolein.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_149.pub2

Methacrolein [*78-85-3*], 2-methylpropenal, α-methylacrolein, CH_2=C(CH_3)-CHO, is a colorless, volatile, toxic, and lacrimatory liquid with a piercing odor. It is an intermediate in one of the processes for the production of methyl methacrylate (see Section 3.2).

2. Properties

2.1. Physical Properties

The important physical properties of acrolein and methacrolein are compiled in Table 1. The solubility of acrolein in water is limited. It is soluble in many organic solvents, such as alcohols, ethers, and aliphatic or aromatic hydrocarbons. Methacrolein has properties similar to acrolein but is less soluble in water.

Table 1. Physical properties of acrolein and methacrolein

	Acrolein	Methacrolein
M_r	56.06	70.09
bp (101.3 kPa), °C	52.69	68.4
(1.33 kPa), °C	−36	−25
mp, °C	−86.95	−81.0
Relative density, d_{20}^{20}	0.8427	0.8474
Refractive index, n_D^{20}	1.4013	1.4169
Vapor pressure (20 °C), kPa	29.3	16.1
Viscosity (20 °C), mPa · s	0.35	0.49
Solubility (20 °C), g/kg		
in water	260	50
water in	73	36
Critical temperature, °C	233	257
Critical pressure, MPa	5.07	4.36
Critical volume, mL/mol	189	
Heat of vaporization (101.3 kPa), kJ/mol	28.2	29.0
Heat of combustion (25 °C), kJ/mol	1632	2299
Heat of formation (gas, 25 °C), kJ/mol	−74.5	−70.8
Heat of polymerization, kJ/mol	71 – 80	
Specific heat capacity, kJ $mol^{-1}K^{-1}$		
c_p (liquid) (17 – 44 °C)	0.120	
c_p (gas) (27 °C)	0.067	
Flash point, open cup, °C	−18	−15
closed cup, °C	−26	
Flammability limits in air, vol%		
upper	31	
lower	2.8	6.0
Autoignition temperature in air, °C	234	280

2.2. Chemical Properties

Acrolein is an extremely reactive chemical because of its conjugated vinyl and aldehyde groups. It undergoes reactions characteristic of both an unsaturated compound and an aldehyde. The conjugation between the carbon – carbon double bond and the carbonyl group increases the reactivity of both groups, which can react either together or separately [9]. Highly exothermic polymerization can occur spontaneously (see Chap. 5).

Diels – Alder Reaction. Acrolein reacts as both a diene and a dienophile. Thus two molecules of acrolein can form a cyclic dimer, 3,4-dihydro-2*H*-pyran-2-carboxaldehyde [*100-73-2*] [11]:

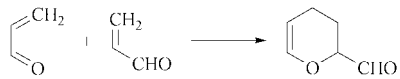

The dimer is formed by the uncatalyzed, thermal reaction of acrolein at approximately 190 °C (1 h for 75 % acrolein conversion) together with polymeric side products. Sufficient stabilization with hydroquinone or complexing compounds, such as polyvalent organic acids [13], is necessary for high yields. This "thermal dimer" is a clear liquid with an unpleasant odor: *bp* 151.3 °C at 101.3 kPa, d_{20}^{20}1.0775. Acid hydrolysis of the dimer yields 2-hydroxyadipaldehyde [*141-31-1*], which can be hydrogenated to form 1,2,6-hexanetriol [*106-69-4*].

Acrolein behaves as a 1,3-diene in reactions with dienophiles in which the electron density of the carbon-carbon double bond is increased by electron-releasing substituents. Vinyl ethers [10, 12, 14, 15] and vinylamines [16] react readily with acrolein to form dihydropyrans. The reaction of methyl vinyl ether and acrolein to form 3,4-dihydro-2-methoxy-2*H*-pyran [*4454-05-1*] is a commercially important example. At a reaction temperature of 160 – 190 °C, reported yields are 80 – 90 % [12].

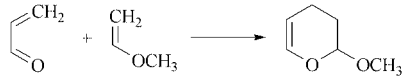

Acid hydrolysis of this product leads to glutaraldehyde [*111-30-8*] (for uses, see page 11).

The electron-deficient vinyl group of acrolein reacts readily with conjugated dienes, such as butadiene or substituted butadienes, forming

derivatives of 3-cyclohexene-1-carbaldehyde [14, 17]:

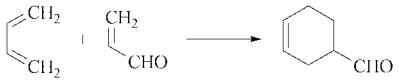

3-Cyclohexene-1-carbaldehyde (1,2,3,6-tetrahydrobenzaldehyde) [*100-50-5*] is formed at 100 – 150 °C in ca. 80 – 90 % yield and is a valuable intermediate for various commercial products. With pentaerythritol a cyclic acetal is formed, which can be used as a stabilizer against ozone degradation in natural or synthetic rubber [18]:

The carbaldehyde also is used for the synthesis of cycloaliphatic epoxides, such as the following:

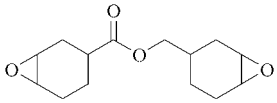

Addition to the Carbon – Carbon Double Bond. The β-carbon atom of acrolein, which is polarized by the carbonyl group, behaves as an electrophile. Therefore nucleophilic reagents, such as alcohols, thiols, water, amines, active methylene compounds, and inorganic and organic acids, add to the carbon – carbon double bond of acrolein in the presence of acidic or basic catalysts. These reactions must be carried out under carefully controlled conditions in order to minimize undesirable side reactions [10].

$CH_2{=}CHCHO + ROH \rightarrow ROCH_2CH_2CHO$	[19, 20]
$+ H_2O \rightarrow HOCH_2CH_2CHO$	
$+ H_2O + H_2 \rightarrow HOCH_2CH_2CH_2OH$	
$+ CH_3COOH \rightarrow CH_3COOCH_2CH_2CHO$	[21]
$+ HCl \rightarrow ClCH_2CH_2CHO$	
$+ HCl + ROH \rightarrow ClCH_2CH_2CH(OR)_2$	
$+ Cl_2 \rightarrow ClCH_2CHClCHO \rightarrow CH_2{=}CCl\text{-}CHO$	
$+ CH_3SH \rightarrow CH_3SCH_2CH_2CHO$	
$+ CH_2(COOR)_2$ * $\rightarrow (ROOC)_2CHCH_2CH_2CHO$	[10, 22]

* or other compounds with an active methylene group

The addition of water under mild acidic conditions gives 3-hydroxypropionaldehyde (HPA) [*2134-29-4*] with high selectivity. Buffer solutions with a pH of 4 – 5 [23, 24] or weak acidic ion-exchange resins [25] are preferentially used as catalysts. Further hydrogenation of the aqueous solutions gives 1,3-propanediol [*504-63-2*] [26, 27]. Direct oxidation of aqueous solutions of 3-hydroxypropionaldehyde at pH 3 with precious metal catalysts produces 3-hydroxypropionic acid [28]. If the oxidation is conducted at above pH 7, the malonate anion is formed in high yield [29].

Acrolein reacts rapidly with hydrogen chloride or hydrogen bromide to form 3-chloropropionaldehyde [*19434-65-2*] [30, 31] or 3-bromopropionaldehyde, but these products easily polymerize, e.g., to trimers and tetramers, in the presence of acids. The preferred procedure for making 3-halopropionaldehyde acetals is therefore simultaneous hydrogen halide addition and acetalization. Acetal yields are about 90 % [31, 32].

Chlorine and bromine add to acrolein in dilute aqueous solution to give 2,3-dihalopropionaldehydes with about 85 % yields; these products can be dehydrohalogenated to form the 2-haloacroleins [33].

These 2-haloacroleins are considered to be potent mutagens [34]. Further halogenation of the 2-haloacroleins provides the 2,2,3-trihalopropionaldehydes, which are valuable intermediates in the synthesis of folic acid [35] (→ Vitamins).

The addition of hydrogen sulfide to two equivalents of acrolein followed by an aldol reaction forms 3-formyl-5,6-dihydrothiopyran [*30058-79-8*] [36].

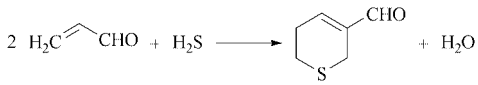

The base-catalyzed addition of methanethiol to form 3-(methylthio)propionaldehyde [*3268-49-3*] is the commercially most important reaction used in the synthesis of the essential amino acid D,L-methionine (→ Amino Acids, [2]).

Reactions of the Aldehyde Group. The selectivity of the acid-catalyzed *acetalization* strongly depends both on the nature of the alco-

hol used and on the catalysts [10]. The most important side reactions yield the corresponding 3-alkoxypropionaldehydes and 3-alkoxypropionaldehyde acetals.

Usually cyclic acetals are formed much more easily, especially with branched diols such as 2-methyl-1,3-propanediol, than the acyclic acetals of lower alcohols [37, 38]. For the preparation of various cyclic acetals, several continuous processes are known [39]. High yields (more than 90 %) of the dimethyl and diethyl acetals can be realized by means of special processes in combination with extraction [40].

Acrolein acetals are valuable intermediates. They have gained interest in the last years as safer forms for the transport of acrolein, especially 2-vinyl-1,3-dioxolane [*3984-22-3*]. The acetals can be cleaved back into acrolein easily in aqueous acidic solutions [41]. These solutions are of particular interest for the use of acrolein as an aquatic herbicide and as a hydrogen sulfide scavenger in oil-field waters.

One of the most interesting cyclic acetals is the spiro diacetal, diallylidene pentaerythritol [*78-19-3*], prepared by condensing pentaerythritol with two moles of acrolein. It can be polymerized with vinyl or other monomers to form spirane resins [42].

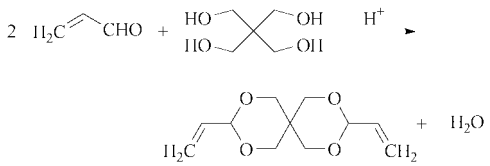

The hydroformylation of acrolein acetals, especially cyclic acetals, leads to the monoacetals of 1,4-butanedial. Subsequent hydrolysis and hydrogenation of this hydroformylation product produces 1,4-butanediol [37, 43].

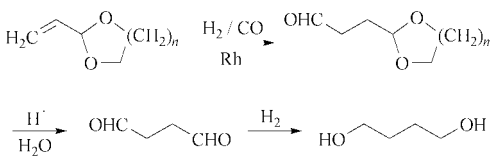

A synthesis of D,L-tryptophan from 1,4-butanedial has been described [44]

Acrolein diacetate [*869-29-4*] 2-propene-1,1-diol, diacetate, is easily prepared in 90 % yield by the acid-catalyzed reaction of acetic anhydride with acrolein [10]:

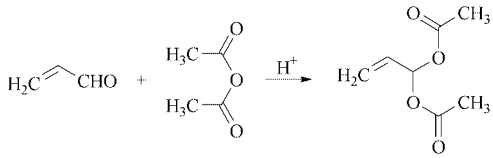

Hydrogen cyanide addition in the presence of mild alkaline catalysts gives acrolein cyanohydrin [*5809-59-6*] in yields of more than 90 % [10, 45, 46]. Acrolein cyanohydrin is a vesicant; it decomposes vigorously in the presence of alkali and is stabilized by traces of acid. Its reaction with acetic anhydride yields acrolein cyanohydrin acetate, which is a valuable intermediate in the synthesis of pharmaceutically and biologically active substances [47 – 49].

Simultaneous Reaction of the Aldehyde and Vinyl Groups. Acrolein has long been recognized as an intermediate in the Skraup synthesis of quinolines from aromatic amines and glycerol [10]. However, little commercial use has been made of acrolein or methacrolein as starting materials for this reaction.

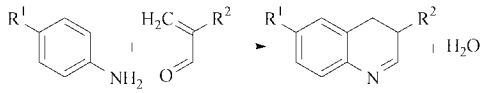

The heterogeneously catalyzed gas-phase condensation of acrolein and ammonia over multicomponent Al_2O_3 or $SiO_2 - Al_2O_3$ catalysts gives 45 % 3-methylpyridine [*108-99-6*] with 20 – 25 % pyridine [*110-86-1*] as a byproduct [10, 50 – 52].

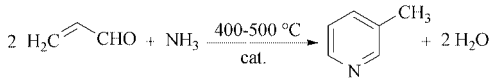

This reaction, usually carried out in a fluidized-bed reactor, has been generalized to a universal synthesis for substituted pyridines [52] (→ Pyridine and Pyridine Derivatives).

Isocyanuric acid reacts with acrolein to form an adduct useful as a cross-linking agent [53].

The bifunctional reactivity of acrolein also has been used in various syntheses of heterocycles [54]. Examples include the reactions with phenols, yielding substituted chromenes or chromanes [55], and the reaction with 2-aminophenols to produce 8-quinolinols [56]. The reaction of acrolein with excess phenol under acidic conditions leads to a mixture of polyphenols with 1,1,3-tris(4-hydroxyphenyl)propane [*4137-11-5*] as the main product [57].

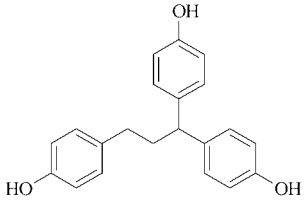

Further reaction with epichlorohydrin leads to glycidyl polyphenols which are useful as epoxy hardeners [58]. Reaction with glycidyl methacrylate followed by polymerization gives heat-resistant materials with high impact resistance [57].

The reaction of acrolein with urea leads to (hexahydro-2-oxo-4-pyrimidinyl urea) [*31036-27-8*], which has been proposed as a fertilizer [10, 59]:

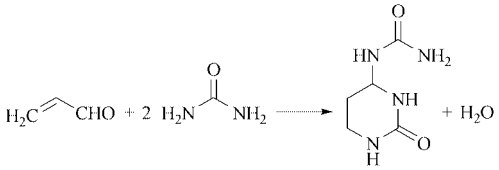

The condensation of acrolein with hydrazine and substituted hydrazines gives pyrazolines in 80 % yield [60, 61]:

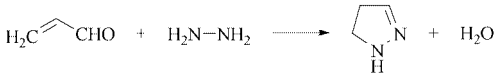

Acrolein reacts with formaldehyde under alkaline conditions to give pentaerythritol [*115-77-5*] [62]. Yields are lower than in the commercial synthesis of pentaerythritol via acetaldehyde and formaldehyde.

Sodium bisulfite reacts with acrolein to form a stable adduct, the disodium salt of 1-hydroxypropane-1,3-disulfonic acid [*35850-94-3*]. This reaction can be conveniently applied for deodorization of spilled acrolein.

Reduction. The selective reduction of acrolein to allyl alcohol [*107-18-6*] in high yields (90 %) is accomplished by hydrogen transfer from secondary alcohols, such as 2-propanol, in a Meerwein – Ponndorf reaction in the liquid phase [10] or from ethanol or 2-propanol in the gas phase over magnesium oxide catalysts [63 – 65] or rare earth oxide catalysts [66] at 300 – 500 °C.

Moreover, the aldehyde group is selectively reduced by metal hydrides, such as sodium borohydride or lithium aluminum hydride [67]. The catalytic hydrogenation to allyl alcohol [63, 68] is either less selective or selective only at very low conversions.

Oxidation. Glycidaldehyde [*765-34-4*] is easily prepared in 80 – 90 % yield by reaction with aqueous hydrogen peroxide at pH 8 – 8.5 [69].

Hydration of the oxirane group gives an aqueous solution of D,L-glyceraldehyde [*56-82-6*] [70].

Acrolein is oxidized commercially with molecular oxygen to form acrylic acid (→ Acrylic Acid and Derivatives).

Polymerization. Acrolein readily polymerizes on heating, on exposure to light, or by the action of various initiators. The properties of the acrolein polymer depend on the polymerization conditions, such as the type of initiator, solvent, and reaction temperature [71].

The functional groups of acrolein (vinyl and aldehyde) can polymerize either separately or together. Radical-initiated polymerization takes place exclusively at the vinyl group, leading to polymers with aldehyde groups **1**, which very

easily yield tetrahydropyran structures **2**. Ionic polymerization gives rise mostly to polymers with vinyl groups **3** along with such polymers as **1** and **4**.

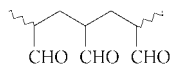

1

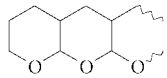

2

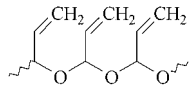

3

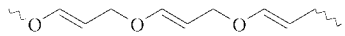

4

Radical polymers are infusible and insoluble in common solvents, whereas the products of anionic initiation are soluble in many organic solvents. In addition to homopolymers, copolymers with most vinylic and acrylic comonomers have been obtained. Furthermore, the carbonyl groups of polyacrolein (**1**) can be modified chemically under mild conditions. Most of the polyacrolein derivatives are soluble in organic solvents or even in water.

These polymers gained interest because of their biocidal activity [72, 73]. The biocidally active component is acrolein, which is slowly released from these polymers [73].

Photochemical (laser) method was used to synthesize ultrafine particles from gaseous mixtures of acrolein and some organosilicon compounds. The polyacrolein particles are in discussion to its applicability to nanotechnology [74].

3. Production

3.1. Acrolein by Propene Oxidation

The first process for acrolein production was commercialized by Degussa in 1942. It was based on the vapor-phase condensation of acetaldehyde and formaldehyde, catalyzed by sodium silicate on silica supports at 300 – 320 °C [75]. This method prevailed until 1959, when Shell began producing acrolein by the vapor-phase oxidation of propene over a cuprous oxide catalyst. The catalyst performance in this process was very poor. In 1957 Standard Oil of Ohio (Sohio) discovered the bismuth molybdate catalyst system, which yielded a fairly good selectivity but still a low propene conversion [76, 77].

$$CH_2 = CH{-}CH_3 + O_2 \rightarrow CH_2 = CH{-}CHO + H_2O$$
$$\Delta H = -340.8 \text{ kJ/mol}$$

The major byproducts of this reaction are acrylic acid and carbon oxides in addition to minor products such as acetaldehyde, acetic acid, formaldehyde, and polyacrolein.

Catalysts for the Oxidation of Propene. The discovery that propene could be oxidized rather selectively to acrolein over copper(I) oxide [78] marked the beginning of the current process of catalytic alkene oxidation to aldehydes over metal oxide catalysts [79]. However, low conversion of propene (20 %) per pass [80], significant recycle of unreacted propene, and low acrolein selectivities were reported for this catalyst. The propene oxidation gathered impetus with the discovery of the bismuth molybdate – bismuth phosphomolybdate system ($Bi_9PMo_{12}O_{52}$ on an SiO_2 support) in 1957 by Sohio [76, 77]. Bismuth -molybdenum oxides also are essential components of commercial catalysts [81 – 84]. The initial step is generally described as coordination of propene on a molybdenum site. Bismuth is involved in the rate-determining hydrogen abstraction from propene. [85]. The mechanism of the main oxidation reaction could be explained by the mechanism of Mars–van Krevelen [86]. The catalysis of propene oxidation to acrolein consists of two reaction cycles: a catalyst reduction cycle (selective product formation), and a catalyst reoxidation cycle (lattice-oxygen regeneration). Only the nucleophilic lattice oxygen (O^{2-}) of the catalyst forms selectively acrolein. In the case the adsorbed electrophilic oxygen reacts with propene, total oxidation to CO_x is favored. To balance the redox properties of the catalyst, additional metal oxides are used.

Modern catalysts are multicomponent metal oxide systems. Knapsack [87] first proposed a three- or four-component catalyst containing Mo-Bi-Fe oxides. Strictly speaking, the multicomponent system consists of different main phases of binary compounds such as α-$CoMoO_4$, β-$CoMoO_4$, $Fe_2(MoO_4)_3$, $Bi_2O_3{\cdot}MoO_3$, $Bi_2O_3{\cdot}2\ MoO_3$ and $Bi_2O_3{\cdot}3\ MoO_3$ [88, 89]. $CoMoO_4$ forms the structural framework of the catalyst covered with iron

molybdate and an overlying layer of bismuth molybdate [90]. It is believed that the active sites of the catalysts are interfaces between $Fe_2(MoO_4)_3$ and β-$CoMoO_4$. Bismuth molybdate is responsible for the selectivity of the catalyst [91].

The optimization of the multicomponent catalyst system [79, 92, 93], to obtain a more selective and also active system with a high space time yield, is still a challenge. The patent application listed in Table 2 of the chemical companies BASF, Nippon Shokubai, Nippon Kayaku, Sabic, or Mitsubishi are some representative examples. The performance of the catalyst could be improved by adding further metal oxides, e.g., of Co, Ni, P, Sb, W, or K and the optimization of the preparation technique. The data of various catalysts that are listed usually refer to short-time results in small, optimized laboratory reactors.

However, acrolein yields depend not only on the chemical compositions of these catalysts, but also on their physical properties, such as shape, porosity, pore-size distribution, and specific surface area, as well as on the reaction conditions and construction of the reactor.

At present the maximum acrolein yield at high propene conversions (up to 98 %) using commercial catalysts is approximately 83 to 90 % with acrylic acid yields of 5 – 10 %. In commercial plants the catalysts are employed at reaction temperatures of 300 – 400 °C, contact times of 1.5 – 3.5 s, and propene concentrations of 5 – 10 vol % of the feed gas at inlet pressures of 150 – 250 kPa. The catalysts have lifetimes of up to ten years, after which they generally have to be replaced. The main reasons for a catalyst change are a reduced economy of the process that is mostly caused by a too low product yield or an increased pressure drop in the reactor.

Production Processes. Production processes are described in [80, 100 – 102]. A simplified flow sheet for the production of acrolein by propene oxidation is shown in Figure 1.

Propene is mixed with air and steam in a molar ratio of approximately 1 : 8 : (1 – 6); steam can be replaced by inert gas, e.g., the off-gas from the absorber. The inlet gas mixture is fed to a multitubular fixed-bed reactor (a) which is cooled by a recirculating molten salt bath.

The reactor is usually operated at 300 – 400 °C with a conversion rate of propene of up to 98 % and inlet pressures of 150 – 250 kPa. The reactor effluent is quenched at the exit to prevent subsequent reactions of acrolein [103]. The reaction gas is then scrubbed with water or water/solvent mixtures in a first column (b) to remove acrylic acid, polymeric compounds, and traces of acetic acid [104]. Byproduct acrylic acid can be recovered from the bottoms [101] and purified; acrylic acid usually forms in 5 to 10 mol % yield based on propene.

The gas is then passed to an absorber (c) where an aqueous solution of acrolein is obtained by absorbing the gas in cold water. Part of the off-gas from the absorber can be used as inert gas for the reactor because it contains only noncondensable components, such as unreacted propene, carbon oxides, oxygen, and nitrogen. The rest is purged as waste gas after it passes through a combustion system.

The aqueous solution of acrolein is sent to a desorption column (d), where it is stripped to give crude acrolein; the bottom stream from this column is cooled and reused as an absorbent. The crude acrolein is distilled to remove low-boiling byproducts, such as acetaldehyde, and heavy ends; acrolein is then obtained as a 96 % pure product that contains only traces of acetaldehyde. Sometimes crude acrolein is used directly. To minimize polymerization, the whole system is stabilized by, e.g., hydroquinone.

New Developments. Selectivity and yield are in the main focus in heterogeneous oxidation catalysis, because the cost of feed materials on the basis of oil escalates. New catalyst materials improve energy and raw materials efficiency and reduce CO_2 formation and emission. Novel approaches to achieve this goal are using oxide nanocatalyst preparation to tune the nature of the active center, oxidant selection to avoid overoxidation, and catalyst arrangement to take advantage of the reaction mechanistic features [105, 106]. An alternative route to acrolein is indicated by numerous publications dealing with the conversion of propane over Mo-V-Te-X-O catalysts [107 – 110]. Actually, new catalysts on the basis of Mo-P-Te-O/SiO_2 and Mo-V-Te/SiO_2 (and MCM 41) show the selective formation of acrolein from propane [111]. However, thermodynamic considerations revealed the limitation of propane as feedstock for the one step synthesis of acrolein [112]. The

Table 2. Industrial catalysts for the selective oxidation of propene to acrolein

Catalyst	T [°C]	X(PE) [%]	Y(AC) [%]	Y(ACA) [%]	Company	Ref.
$Mo_{12}Fe_3Bi_{0.75}\ Co_8O_x$ + Sb, K	350	87	84.5	1.4	-	[89]
$Mo_{12}Fe_2Bi_{1.5}\ Co_{4.4}K_{0,06}O_x$	320	99	89.6	1.6	LG Chem.	[94]
$Mo_{12}Fe_{2,4}Bi_{1,08}\ Co_{9,6}Al_{1,48}V_{0,056}Ag_{0,1764}\ Pd_{0,0019}K_{0,064}O_x$	342	97	95.6	-	Sabic	[95]
$Mo_{12}Fe_{2,94}Bi_{0,8}Co_7Si_{1,52}K_{0,08}O_x$	318	95	87.1	4.1	BASF	[96]
$Mo_{12}Fe_{1,3}Bi_{1,3}Co_6Ni_{2,0}Si_{2,0}K_{0,08}O_x$	310	98.2		92.4*	Nippon Shokubai	[97]
$Mo_{12}Fe_{1,8}Bi_{1,7}Ni_{2,8}Co_{5,2}K_{0,1}O_x$	346	97	84.8	7.4	Nippon Kayaku	[98]
$Mo_{12}Fe_{0,6}Bi_{1,0}Co_{3,3}Ni_{3,3}B_{0,2}K_{0,1}Na_{0,1}\ Si_{24}O_x$	315	98,5	90.1	4.2	Mitsubishi	[99]

* sum AC+ACA

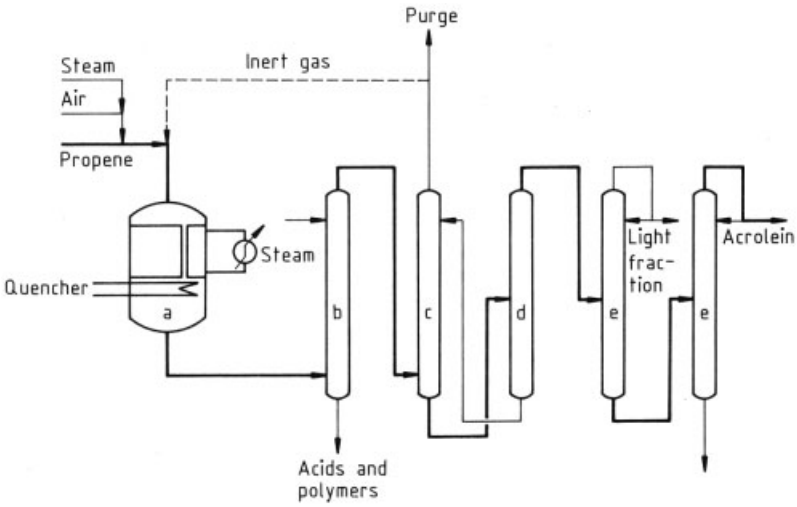

Figure 1. Acrolein production by propene oxidation
a) Oxidation reactor; b) Scrubber; c) Absorber; d) Desorber; e) Fractionators

two stage process based on a first dehydrogenation step of propane to propene followed by a conventional unit for the oxidation of propene to acrolein (PDH-process). The first stage, the dehydrogenation of propane to propene is already commercial available: UOP (Oleflex-Process), ABB Lummus Global (Catofin-Process), Linde (Linde-Process), Snamprogetti/Yarsintez (Fluidized Bed Dehydrogenation-3-Process). In these technical processes, conversions of 30 to 60 % and selectivity to propene of 90 % are obtained. Important side reactions are cracking and hydrogenolysis of propane and propene. Coke deposits on the catalysts necessitate the regeneration of the catalyst. The total amount of propene that is produced by PDH is about 2 % of the total global amount. The oxidative dehydrogenation of propane (ODH process) is still in the development and has not yet led to a commercial process [113 – 116]. However, the direct combination of propane dehydrogenation and selective oxidation of propene to acrolein has not yet been commercialized.

The MTP process (Methanol to Propene) that was developed by Lurgi uses methanol as feedstock for propene. The zeolite-based catalyst for the fixed bed process was developed by Südchemie AG. This technology increases the basis of the raw materials and is the latest example for the increasing relevance of methanol as basis for the chemical production of the future. It is planned to use methanol from a MegaMethanol plant that is sent to an adiabatic DME prereactor where, methanol is converted to DME and water. A methanol, water, DME stream is routed to the MTP® reactor. DME and methanol is

converted by more than 99% with propylene as main product.

3.2. Methacrolein

Methacrolein (MA) was produced by Union Carbide in the 1950s and early 1960s by 2-methylpropene (isobutene) oxidation over a copper(I) oxide catalyst. Today, several companies produce methacrolein commercially as an organic intermediate that is used as building block for chain expansions. Large quantities of methacrolein are produced by several producers [e.g., BASF, Mitsubishi Rayon, Kyodo Monomer (Kuraray, Mitsui), Nihon Methacrylic Monomer (Nippon Shokubai, Sumitomo), Asahi Kasei] as an intermediate in newly developed processes for the production of methyl methacrylate or methacrylic acid. These processes use either C_4 feedstocks [2-methylpropene (isobutene, IB), *tert*-butyl alcohol (TBA)] or the C_2 hydrocarbon ethylene (via hydroformylation) as raw material. However, around 75 % of methyl methacrylate (MMA) worldwide is still produced by the esterification of methacrylamide obtained from acetone cyanohydrin.

Vapor-Phase Oxidation of 2-Methylpropene or *tert*-Butyl Alcohol. Catalysts for the oxidation of 2-methylpropene or *tert*-butyl alcohol are multicomponent metal oxide systems similar to those used to make acrolein. Most of the catalysts for the oxidation of propene have been claimed in the literature to be suitable for the oxidation of 2-methylpropene or *tert*-butyl alcohol. Typical reaction pressure ranges from above 1 bar to about 5 bar, and reaction temperatures range from 350 to 375°C. Higher temperature permits faster reactions but lower selectivity. The reaction times are typically 2 to 4 sec. However, the yields were generally lower than for acrolein. Yields of 80 – 87 % methacrolein and 1 – 5 % methacrylic acid for 90 – 98 % conversions in short-time laboratory experiments have been reported [117]. At high conversion, product gas must not be recycled. For high selectivity, the process generally requires greater dilution of the feed gas with steam (ca. 8 – 10 vol %) or inert gas (ca. 78 – 80 vol %) than is needed for propene oxidation. Impurities, such as butadiene, often oligomerize on the oxidation catalyst causing serious deterioration of its activity. Methacrolein can be separated and purified before it is fed to a second reactor or can be directly led into the second stage to convert the methacrolein to methacrylic acid (MAA). MAA is finally converted in the liquid phase with methanol to obtain methyl methacrylate.

Several plants for producing methyl methacrylate via methacrolein have been constructed in Japan and Korea. In 1983, Nippon Shokubai brought on stream a 15 000 t/a demonstration plant based on the oxidation of 2-methylpropene [118], and later in the same year Mitsubishi Rayon started a 40 000 t/a commercial plant based on the oxidation of *tert*-butyl alcohol [119]. These were the first plants using the new technology. Process conditions are similar to those for propene oxidation [120, 121].

In 2002, Asahi Kasei added 30 000 t/a of methyl methacrylate (MMA) capacity at its 70 000-t/a plant at Kawasaki, Japan. The company uses proprietary technology for the direct oxidative esterification of methacrolein to MMA [122]. In 2002, nearly half of the Japanese methyl methacrylate capacity of 600 000 t/a (Germany 320 000 t/a, total western Europe 743 000 t/a) [123] was produced with processes based on 2-methylpropene or *tert*-butyl alcohol. This corresponds to an estimated methacrolein capacity of 265 000 t/a.

Production from C_2 and C_1 Feedstocks. ASF has developed a process for the production of methacrolein from C_2 and C_1 feedstocks. Ethylene is hydroformylated to propanal, which can be condensed with formaldehyde in a Mannich type reaction to give methacrolein: For each unit of propionaldehyde produced, approximately 0.5 units of ethylene are required, which reflects a 96 to 97 % yield.

$$CH_2 = CH_2 + CO + H_2 \rightarrow CH_3CH_2CHO$$

$$CH_3CH_2CHO + CH_2O \rightarrow CH_2{=}C(CH_3)CHO + H_2O$$

The Mannich condensation can be carried out in aqueous solution of dimethylamine in the presence of acetic acid to from the Mannich base salt. Methacrolein can be distilled from the resulting solution in a yield of 95 %; the aqueous solution can be recycled [124]. Since 1989

BASF has used this process as one reaction step in their production process for methyl methacrylate starting from ethylene (capacity 36 000 t/a).

Alternatively, the crossed aldol condensation of propionaldehyde and formaldehyde (as trioxane) takes place over molecular sieves such as HAMS-1B/ Al_2O_3 at 300°C. The conversion is around 58 % with 98 % selectivity to methacrolein. No amine is required [125].

Dehydrogenation of Isobutyraldehyde. Methacrolein can be prepared by vapor-phase oxidative dehydrogenation of isobutyraldehyde using a catalyst with heteropoly anions containing molybdenum and phosphorus as the main components [126]. Yields are approximately 80 % methacrolein and 7 % methacrylic acid. However, as long as isobutyraldehyde is available only from propene as a byproduct of the oxo synthesis, this route will not be a major process for methacrolein.

A new method of conversion of isobutane to methacrylic acid, which involves contacting the catalyst with isobutane in gas-phase flow reactor in presence of oxygen under reaction conditions was recently reported [127]. However, the synthesis of methacrolein from isobutane is still a challenge.

4. Quality and Analysis

Quality. Typical specifications (wt %) guaranteed by producers of acrolein are as follows:

Purity of acrolein 95 – 97
Water content < ca. 3
Hydroquinone < 0.1 or 0.2
Acetaldehyde < 0.3

Some producers specify the acetaldehyde limit as < 2 % and indicate propene oxide contents of up to 1.5 %.

Analysis. Impurities include water, acetaldehyde, and, depending on the process, small amounts of propionaldehyde, acetone, propene oxide, methanol, and traces of allyl alcohol and ethanol. The contents of acrolein, water, and acetaldehyde can be determined by gas chromatography with a thermal conductivity detector. The Karl Fischer method is not applicable for determining the water in acrolein. Hydroquinone is determined colorimetrically using Millon's reagent to form a yellow complex [10].

5. Handling, Storage, and Transportation

Acrolein is classified as a very toxic, flammable liquid. The liquid vaporizes easily and the vapors are readily flammable in air at between 2.8 and 31.0 vol%. Acrolein vapors are twice as heavy as air. The flash point, − 26 °C, is very low. Preferred fire-fighting agents are foam, powder (not alkaline), spray water, and carbon dioxide. The high toxicity and volatility of acrolein necessitate the use of a respirator in case of fire.

Acrolein polymerizes easily and exothermally; therefore it is stabilized with 0.1 or 0.2 % hydroquinone against radical-initiated polymerization, which can be catalyzed by light, air, heat, or peroxides. Acrolein should be stored and transported in the dark under a blanket of nitrogen at temperatures below 20 °C, and it should be used within three months. The hydroquinone content should be determined (see Chap. 4) and if it is below 0.05 %, hydroquinone must be added to bring it up to 0.1 %.

Highly exothermic ionic polymerization is catalyzed by alkaline compounds, such as caustic soda, ammonia, and amines, or by mineral acids, such as concentrated sulfuric acid. These chemicals initiate polymerization at an explosive rate. Because even a trace amount of contaminant can initiate polymerization, equipment for acrolein handling must be cleaned thoroughly before use. Common inhibitors such as hydroquinone are *absolutely not effective* in preventing these ionic polymerization reactions. Depending upon the amount of ionic contaminants, the temperature rise may be slow enough that the injection of an emergency buffer solution (84 % acetic acid, 8 % hydroquinone, 8 % anhydrous sodium acetate) suffices to control further reaction. Addition of water to stored acrolein must be avoided completely. The acrolein-containing water layer is particularly prone to polymerization. Acrolein vapors polymerize upon condensation.

Because acrolein is toxic, it is not allowed in wastewaters. It is also strongly lacrimatory,

and contact with skin and eyes must be strictly avoided because exposure can result in severe injury. But it can be handled safely under controlled conditions in properly designed equipment [128].

Regulations Governing Transport. Acrolein is classified as a dangerous, flammable, and poisonous substance in various international and regional regulations. For the transportation of inhibited acrolein, the following regulations are mandatory:

International sea transport (IMDG Code): Class 6.1, UN no. 1092, PG I. International air transport (IATA-DGR): forbidden. European road (ADR) and rail (RID) transport: Classes 6.1, 8a). Proper shipping name: Acrolein, inhibited. National regulations: United States (CFR 49): § 172.101 Toxic liquid, flammable. Germany (GGVS): Special permission required for road transport of 1000 kg or more if transported in tanks with a capacity exceeding 3000 L.

For the transportation of inhibited methacrolein the following regulations are mandatory:

International sea transport (IMDG Code): Class 3.2, UN-No. 2396, PG. II. International air transport (IATA-DGR): Class 3, UN no. 2396, PG II. European road (ADR) and rail (RID) transport: Class 3, 17b). Proper shipping name: Methacrylaldehyde, inhibited. National regulations: United States (CFR 49): § 172.101 Flammable liquid, toxic.

6. Uses and Production Data

Methacrolein is an intermediate in two processes for the production of methyl methacrylate: the oxidation of 2-methylpropene or *tert*-butyl alcohol (see Section 3.2) and the Mannich reaction of propionaldehyde with formaldehyde (see Section 3.2). It has also found limited commercial application in the synthesis of flavors and fragrances [129].

Acrolein is a useful chemical intermediate used for the production of numerous chemical products. It is used commercially as a very effective broad-spectrum biocide in very low concentrations of approximately 10 ppm [130, 131]. For example, it is applied to control the growth of aquatic weeds in irrigation waterways [132] or of algae and mollusks in recirculating water systems [131, 133]. Of particular importance is the use of acrolein as a biocide in oil-field brines; it increases the efficiency of oil-field water flooding and is useful in brine disposal operations. Furthermore, it is used in oil-field waters to scavenge malodorous hydrogen sulfide completely [134].

The main field of use includes the production of methionine, methionine hydroxy analog, acrylic acid, 1,3 propanediol, glutaraldehyde, pyridines, flavors and fragrances.

Methionine. Acrolein is mostly used for the production of methionine [*59-51-8*] (→ Amino Acids), which is used as an animal feed supplement [2]. Acrolein is also used to make the methionine hydroxy analogue ($CH_3SCH_2CH_2CH(OH)COOH$) [*583-91-5*], either as an 88 % aqueous solution [135, 136] or the calcium salt [137]. Different values for the bioefficacy of this acid have been published, but it is about 22 – 28 % lower than that of D,L-methionine on an equimolar basis.

Acrylic acid is another commercially important product derived from acrolein (→ Acrylic Acid and Derivatives). It is used to make acrylates.

1,3-Propanediol (→ Propanediols) can be produced by the Degussa/DuPont process from acrolein. Acrolein is hydrolyzed with a acid catalyst to 3-hydroxypropionaldehyde that is hydrogenated to 1,3-propanediol. The process was in operation in Cologne-Wesseling/Germany. Recently, DuPont announced to start the production of bio-propanediol in 2007 by fermentation. Propanediol can be formulated, e.g., into composites, laminates, powder, adhesives and UV-cured coatings, novel aliphatic polyesters, copolyesters, solvents, moldings and anti-freeze [138, 139].

Glutaraldehyde, glutardialdehyde (for preparation from acrolein, see page 2), is supplied industrially in the form of a 25 % or a 50 % aqueous solution; the anhydrous compound is unstable. This dialdehyde is used mainly for leather tanning [140]. Applications also include use as a biocide for industrial water treatment and in oil-field applications [141] and as a dis-

infectant and chemical sterilizer for hospital equipment [142].

Pyridines. Only Daicel is currently using acrolein in the commercial synthesis of 3-methylpyridine (see page 4). Other substituted pyridines can be prepared from acrolein and substituted acroleins [52, 143, 144] (→ Pyridine and Pyridine Derivatives).

Tetrahydrobenzaldehyde (see page 2) is another interesting product, e.g., for the synthesis of pharmaceuticals, fungicides [145], and fragrances [146].

Flavors and Fragrances. A broad variety of compounds, synthesized mainly by Diels – Alder reactions of acrolein or methacrolein (see page 2), are described as flavors and fragrances. Commercially interesting substances are, for example, lyral (**5**) [*31906-04-4*] [147, 148], myrac aldehyde (**6**) [*80450-04-0*] [148], and 5-norbornene-2-carbaldehyde (**7**) [149].

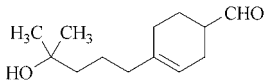

5

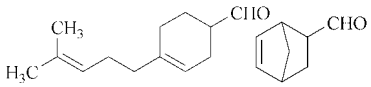

6 **7**

Herbicides. Acrolein cyanohydrin acetate is produced as an intermediate for the production of phosphinotrycin, a nonselective herbicide. Phosphinotrycin is very effective in crop protection together with genetically modified phosphinotrycin-resistant crops [150].

Allyl Alcohol and Glycerol. In a process developed by Shell in 1959, allyl alcohol was synthesized from acrolein by hydrogen transfer from 2-propanol in the vapor phase. The allyl alcohol was then reacted with hydrogen peroxide to form glycerol. However, this process was closed down in 1980.

D,L-Glyceraldehyde (see page 5), which is commercially available as a 40 % aqueous solution, has been proposed as a water-soluble hardener for leather tanning and related applications, as well as for various syntheses.

Acrolein Polymers. Acrolein itself has seldom been used as a monomer for commercial polymerization because of the difficulty in orienting its mode of polymerization and the likelihood of complex cross-linking leading to insoluble products. However, a poly(aldehyde carboxylic acid), produced by oxidative copolymerization of acrolein and acrylic acid [151], is applied industrially as a sequestering agent [152]. Furthermore, many applications of acrolein polymers have been proposed, including those in textile treatment [153], reinforcement of paper [154], and photography [155]. The water-soluble polymeric condensation product of acrolein and formaldehyde in 40 % aqueous solution is used as a biocide, e.g., as an algicide in recirculating cooling waters [156].

Production Data. The production capacities (t/a) of acrolein in the western Europe, USA, and Japan are estimated as follows:

Degussa	150 000
DOW	80 000
Adisseo	80 000
Arkema	70 000
Sumitomo Chem.	35 000
Daicel	10 000

Moreover, all acrylic acid processes involving vapor-phase propene oxidation include the intermediate production and consumption of acrolein.

7. Toxicology and Ecotoxicology

7.1. Toxicology

Acrolein is severely irritating to the skin and the mucous membranes. Its vapor causes strong eye and nasal irritation. Direct contact of liquid acrolein with the eye or skin results in severe burns. Acrolein is very toxic after inhalation exposure and toxic after oral administration and skin contact [157].

Acute toxicity data are:

LD_{50} (rat, oral)	< 11 up to 46 mg/kg
LD_{50} (rabbit, skin, neat acrolein)	562 mg/kg
LD_{50} (rabbit, skin, 20 % aqueous acrolein solution)	335 mg/kg
LC_{50} (rat, inhalation, 4 h)	0.02 mg/L

Even very dilute solutions of acrolein are strongly irritating to skin and mucous membranes [158]. Severe damage to the eye was observed in rabbits with a 1 % acrolein solution in glycol [159]. Due to the high reactivity of the acrolein molecule it is primarily bound locally to the application site. After absorption in the gastrointestinal tract, the main pathway of biotransformation is conjugation with glutathione and subsequent oxidation or reduction of the aldehyde group [158]. Acrolein was reported to be nonsensitizing in the guinea pig maximization test [158].

In subacute to chronic inhalation studies conducted in various species, irritation and inflammation with hyper- and metaplastic changes in the respiratory tract were the primary observed effects. At higher exposure concentrations, additional inflammatory changes in the liver and kidney were observed. After exposure for 13 weeks, (6 h/d, 5 d per week), a concentration of 0.9 mg/m^3 was proven to be the No-Observed-Adverse-Effect Level (NOAEL) in hamsters and rabbits [160], while exposure for 8 to 13 weeks (6 h/d, 5 d per week) at a concentration of 0.9 mg/m^3 caused slight damage to the nasal mucosa and the lung epithelium in some strains of rat [160 – 162]. After continuous exposure of rats (24 h/d, 7 d per week) for 61 d, the NOAEL was 0.15 mg/m^3 [158].

Chronic oral administration of acrolein to rats, mice and dogs resulted in reduced body weight gain, increased mortality and changes in clinicochemical parameters at high dosages. Pathological organ changes, however, were not observed. The NOAEL for rats and dogs was 0.5 mg/kg and 2.0 mg/kg for mice [158].

In vitro acrolein reacts with nucleic acids and inhibits their synthesis. On the basis of the high cytotoxicity as well as the high reactivity of acrolein, there are difficulties in testing the genotoxic potential. From the available studies it may be concluded that acrolein seems to show genotoxic properties predominantely at high, cytotoxic concentrations in in vitro test systems (bacteria, yeasts, mammalian cell cultures), while no mutagenic properties were noted in in vivo studies in mammalian animals [158, 163].

Acrolein is noncarcinogenic after oral administration to rats and mice. After inhalation exposure of hamsters for 52/81 weeks no increase in tumor incidence was observed. In a dermal initiation/promotion study in mice (18 weeks) there was no indication of a possible cocarcinogenic effect of acrolein [158].

Acrolein is classified by IARC in group 3 (inadequate evidence for carcinogenicity in humans, with inadequate evidence in experimental animals and no information in humans) [164].

Acrolein did not show any reproductive or developmental toxicity after oral or inhalative exposure in multigeneration experiments in rats [158, 163]. No developmental toxicity was observed in doses that were not maternally toxic after oral and i.v. administration in rabbits. The overall NOAEL in teratogenicity studies was 2 mg/kg bw per day or higher for developmental effects and 0.75 mg/kg bw per day for parental effects [158, 163].

In humans exposed to acrolein vapors for 5 min, an increased frequency of eyelid closure was observed at 0.69 mg/m^3 and a decrease in respiratory frequency at 1.4 mg/m^3 [165]. The odor threshold was reported to be between 0.05 and 0.8 mg/m^3. At 0.23 mg/m^3, 50 % of the test persons were able to detect the odor [158]. Thus the odor threshold seems to be close to the threshold for eye irritation.

Occupational exposure limits are 0.1 ppm (0.25 mg/m^3) in Germany (8 h TWA and STEL) [166], and 0.1 ppm (0.23 mg/m^3, 8 h TWA), 0.3 ppm (0.7 mg/m^3, STEL) in the United Kingdom [167], and 0.1 ppm (0.23 mg/m^3) ceiling level the United States [168]

Methacrolein is of moderate toxicity after oral, dermal, and inhalative administration to experimental animals.

Acute toxicity data are:

LD_{50} (rat, oral)	140 mg/kg
LD_{50} (rabbit, skin)	111 mg/kg
LC_{50} (rat, inhalation, 4 h)	560 mg/m^3

The effects after inhalative exposure are characterized by irritation of the respiratory tract [169]. A sensory irritation study in mice revealed a 50 % reduction in the respiration rate (RD_{50}) at exposure concentrations of methacrolein vapors of 10.4 ppm following a 30-min exposure period [170].

Methacrolein is corrosive to rabbit skin and severely irritating to rabbit eyes in concentrations of 1 – 5 %. Therefore undiluted

methacrolein is expected to be corrosive to the eye [169].

After subacute (15 d, 6 h/d, 5 d per week) and subchronic (90 d, 6 h/d, 5 d per week) inhalation in rats, the only adverse effect observed was severe irritation of the upper, middle, and lower respiratory tract. The NOAEL was 13 mg/m^3 (14 d study) and 14 mg/m^3 (90 d study) respectively. In the 90 d study, signs of reversibility were observed within the four week recovery period [169].

In humans exposed to methacrolein vapors for 10 min, the eye blinking rate was increased at a concentration of 0.29 ppm, the NOAEL in this study was 0.19 ppm [170]

Methacrolein showed conflicting results in mutagenicity tests in microorganisms [169], which may be related to bacteriotoxic effects. In an in vitro chromosomal aberration assay in Chinese hamster V79 cells, methacrolein caused structural chromosomal aberrations with and without metabolic activation [169].

No embryotoxic or teratogenic effects were observed in maternally nontoxic doses in an inhalation study for developmental toxicity in rats [169].

7.2. Ecotoxicology

Acrolein is of moderate to high toxicity towards aquatic organisms and microorganisms. Toxicity does not seem to increase considerably with exposure time. Ranges of acute toxicity on species of three trophic levels including microorganisms are as follows: In fish LC_{50} (24 – 96 h) values ranged from 0.014 to 2.5 mg/L. In amphibia (tadpoles) an LC_{50} (96 h) of 0.007 mg/L was reported, Daphnia showed EC_{50} (24 – 48 h) values between 0.022 and 0.093 mg/L. In green algae the EC_{50} (1 – 25 h) was 0.69 – 1.8 mg/L. Microorganisms showed IC_{10} (16 – 48 h) values of 0.04 – 1.7 mg/L [158].

Acrolein in a concentration of 1.4 mg/m^3 damaged the leaves of higher plants after an exposure period of 3 h in a fumigation chamber. No toxic effects occurred after 9 h exposure at a concentration of 0.233 mg/m^3 [158].

In a three-generation chronic toxicity study (64 d) in *Daphnia magna*), the MATC (maximal acceptable toxicant concentration) in the second and third generation for reproductive toxic effects ranged from 0.0169 to 0.0336 mg/L [158].

Biodegradation of acrolein only takes place with adapted microorganisms [158]. Based on the low octanol/water partition coefficient (log P_{ow} = − 1.1 – 1.02 [3]) and its high reactivity, bioaccumulation of acrolein is not expected.

Methacrolein. Data on environmental distribution of methacrolein indicate that the majority of the substance will distribute from water bodies to air. At 25 °C, only 0.02 % of the total methacrolein present will remain in the water phase [171]. No data on fish, daphnia, or algae toxicity are available. In a respiration inhibition test with activated sludge, the toxicity threshold concentration of methacrolein was 0.05 mmol/L. Biodegradation, measured as BOD, was about 35 % after 10 d with unadapted microorganisms and > 60 % after 8 d with adapted microorganisms. This indicates ready biodegradability only after acclimatization [172]. Due to its low octanol/water partition coefficient (log P_{ow} = 0.59 [173]) bioaccumulation of methacrolein is not expected.

8. References

1. J. Redtenbacher, *Liebigs Ann. Chem.* **47** (1843) 113–150.
2. Winnacker-Küchler: Chemische Technik, vol. 8, 5th ed., Wiley-VCH, Weinheim 2005, p. 724–727.
3. W. Weigert, H. Haschke, *Chem. Ztg.* **98** (1974) 2, 61.
4. A. Isard, A. Lakodey, F. Weiss, *Chim. Ind. Genie Chim.* **103** (1970) 11, 1341.
5. G. W. Keulks, D. Krenzke, T. Notermann, *Adv. Catal.* **27** (1978) 183–225.
6. N. A. Keiko, M. G. Voronkov, *Russian Chemical Reviews* **62** (1993) 751–767.
7. R. K. Grasselli: "Creating Value from Light Olefins-Production and Conversion", Proceedings of the DGMK-Conference, Hamburg, p. 147–158.
8. W. A. Herrmann in: *Applied Homogeneous Catalysis with Organometallic Compounds*, 2nd ed., vol. 3, Wiley-VCH, Weinheim, Germany 2002, p. 1297–1304.
9. B. Goyau, *Chimica Oggi* **17** (1999) 22–27.
10. C. W. Smith: *Acrolein*, J. Wiley & Sons, New York-London 1962.

11. G. Jenner, H. Abdi-Oskoni, J. Rimmelin, *Bull. Soc. Chim. Fr.* 1977, 341, 983.
12. Distillers, GB 698 736, 1953 (R. G. Hall, B. K. Howe).
Union Carbide, GB 739 128, 1955 (R. I. Hoaglin, R. G. Kelso).
13. Shell Oil, US 3 159 651, 1962 (G. F. Johnson, L. C. Teague).
14. D. P. Schirmann, F. Weiss, G. Bonnard, *Bull. Soc. Chim. Fr.* 1968, 3326.
15. F. Y. Kasumov, S. K. Kyazimov, R. A. Sultanov, *Epoksidnye Monomery Epoksidnye Smoly* 1975, 150;
Chem. Abstr. **85** (1975) 62907 q.
16. G. Opitz, I. Loschmann, *Angew. Chem.* **72** (1960) 523.
G. Opitz, H. Hoffmann, *Liebigs Ann. Chem.* **684** (1965) 79.
17. T. Mukaiyama, S. Aizawa, T. Yamagushi, *Bull. Chem. Soc. Jpn.* **40** (1967) 2641.
L. Givaudan et Cie, EP 12 224, 1978.
S. Watanabe, S. Suga, H. Tsuruta, T. Sato, *J. Appl. Chem. Biotechnol.* **27** (1977) 423.
P. V. Alston, R. M. Ottenbridge, *J. Org. Chem.* **40** (1975) 322.
18. Bayer, EP 73 374, 1981,
EP 67353, 1981 (M. Blazejak, D. Grotkopp, J. Haydn),
DE 2 548 911, 1975 (E. Roos, T. Kempermann, W. Redetzky).
19. Toa Gosei Chem. Ind., JP 7404204, 1974 (H. Ito, H. Inone, K. Kimura, M. Sato).
20. Degussa, US 5 284 979, 1992 (T. Haas, G. Boehme, D. Arntz).
Du Pont, US 3 536 763, 1970 (H. S. Elenterio, T. A. Koch).
Shell, GB 1 185 615, 1969 (E. T. Lutz).
21. Mitsubishi Rayon, JP 7307089, 1973 (Y. Takayami, Y. Nakoyama, M. Asao, Y. Tokumichi).
Mitsubishi Chem. Ind., JP 4043931, 1970;
Chem. Abstr. **81**, 120208 t.;
22. G. V. Krishtal, V. V. Kulganek, V. F. Kucherov, L. A. Yanooskaya, *Synthesis* 1979, 2, 107.
23. Du Pont, US 3 536 763, 1970 (H. S. Elenterio, T. A. Koch).
Shell, GB 1 185 615, 1969 (E. T. Lutz).
24. R. Hall, E. S. Stern, *J. Chem. Soc.* 1950, 490.
25. Degussa, EP 487 903, 1992 (D. Arntz, N. Wiegand).
26. Degussa, EP 572 812, 1993 (D. Arntz, T. Haas, N. Wiegand).
Degussa, EP 535 565, 1993 (D. Arntz, T. Haas, A. Schaefer-Sindlinger).
Ruhrchemie, DE 2 054 601, 1972 (W. Rottig, H. Tummes, B. Cornils).
27. X. Bai, *Hecheng Xianwei Gongye* **25** (2002), 37–40.
28. Degussa, DE-A 19 629 371, 1996 (T. Haas, M. Meier, C. Broßmer, D. Arntz, A. Freund).
29. Degussa, DE-A 19 629 372, 1996 (T. Haas, M. Meier, C. Broßmer, D. Arntz, A. Freund).
30. Degussa, DE 2 115 327, 1972 (H. Wagner, K. Udluft).
31. *Org. Synth.* **2** (1963) 137.
32. Y. Sato, Sh. Sugasawa, *Proc. Jpn. Acad. Ser. B.* **56** (1980) 573.
33. A. Berlande, *Bull. Soc. Chim.* **Fr. 37** (1925) 4, 1385.
34. J. D. Rosen, Y. Segall, J. E. Casida, *Mutation Research* **78** (1980) 113.
35. Takeda Chemical Ind., JP 59 225 155, 1984.
36. Degussa, DE 1 919 504, 1969 (H. Wagner, K. Udluft).
BASF, DE 3 427 404, 1984 (M. Sauerwald, T. Dockner, W. Rohr, G. Reissenweber).
37. Du Pont, US 4 024 159, 1974 (C. C. Cumbo, K. K. Bhatia).
38. Celanese, US 4 003 918, 1975 (O. R. Hughes).
39. Degussa, US 5 688 973, 1995 D. Arntz et al.
Degussa, US 5 216 179, 1991 (M. Hoepp, D. Arntz, S. Bartsch, A. Schaefer-Sindlinger, W. Boeck).
Du Pont, 4 108 869, 1977 (H. B. Copelin).
40. Degussa, DE 3 403 426, 1984 (G. Prescher, J. Andrade, D. Arntz).
41. Betz Laboratories, US 4 851 583, 1988 (E. Bockowski, C. R. McDaniel).
Degussa, US 5 696 052, 1995 (P. Werle, M. Trageser, O. Helmling, H. Jakob).
42. J. I. Mateo, R. Vallve, *Rev. Plast. Mod.* **198** (1972) 24, 921.
Nippon Kayaku JP Kokai 74 101 310, 74 101 311, 1973;
Chem. Abstr. 82, (1974) 72534 t and 111579 h.
E. Takiyama, T. Hanyuda, *Nippon Setchaku Kyokaishi* **13** (1977) 9, 330.
E. Takiyama, *Plast. Age* **21** (1975) 9, 93.
43. Mitsubishi Petrochemical, JP-Kokai 06 305 998, 1993.
Mitsubishi Petrochemical, JP-Kokai 06 107 586, 1992.
44. Degussa, DE 3 043 259, DE 3 043 252, 1980 (A. Kleemann, M. Samson).
45. Nippon Kayaku, JP 7118733, 1971 (G. Yamaguchi, H. Yoshikawa).
Asahi Chemical Industries, JP 7405917, 1974 (C. Shibuya, S. Onchi, S. Hayashi).
46. M. North, *Sci. Synth.* **19** (2004) 235–284.

47. Hoechst, EP 19 227, 1979 (R. Mündnich, M. Finke, W. Rupp, K. Dehmer),
EP 19750, 1979 (M. Finke, R. Mündnich),
DE 3 047 024, 1980 (M. Finke, H. Erpenbach).
48. R. Palm, H. Ohse, H. Cherdron, *Angew. Chem.* **78** (1966) 1093.
49. McIntosh, *Can. J. Chem.* **55** (1977) 4200.
50. H. J. Vebel, K. K. Moll, M. Muehlstaedt, *J. Prakt. Chem.* **312** (1971) 849.
51. S. Doi, I. Hayashi, G. Amano, Y. Hachihama, *Kogyo Kagaku Zasshi.* **63** (1960) 828.
Ube Ind., JP 7622, 1982 (T. Kawato, K. Koga, M. Kuniyoshi).
52. H. Beschke, H. Friedrich, *Chem. Ztg.* **101** (1977) 377.
Degussa, DE 2 639 701, 1976, DE 2 712 694, 1977, and DE 2 639 702, 1976 (H. Beschke, H. Friedrich), DE 2 819 196, 1978 (H. Beschke, A. Kleemann).
53. Monsanto, EP 55 090 and 55 091, 1982 (S. M. Cohen, T. R. Le Blanc).
54. R. D. Larsen, D. Cai, *Sci. Synth,* **15** (2005) 389–549.
55. Hoffmann La Roche, US 4 003 919, 1972 (J. W. Scott, D. R. Parrisch, G. Saucy).
G. Sartori, G. Casiraghi, L. Bolzoni, G. Casnati, *J. Org. Chem.* **44** (1979) 803.
56. Yuki Gosei Kogyo, JP 7312745, 1970 (S. Wagatsuma, I. Yamashita, Y. Kinoshita); *Chem. Abstr.* **79** (1979) 53191 a.
57. BASF, DE-A 3 520 113, 1985 (B. Czauderna, J. Nieberle, R. Peter, D. Nissen).
58. Ciba-Geigy, EP 6 535, 1980 (F. Lohse, F. Gutekunst, R. Schmid, A. Schmitter).
59. Asahi Chemical Industries, JP 57983, 1973 (K. Tanaka, S. Furuhashi, S. Yokoyama).
60. Bayer, EP 48 373, 1980 (U. Heinemann et al.).
Z. Brzozowski, E. Pormanacka-Jankowska, S. Angielski, *Acta Pol. Pharm.* **34** (1977) 3, 279; *Chem. Abstr.* **88** (1978) 120742 g.
61. B. Stanovnik, J. Svete, *Sci. Synth.* **12** (2002) 15–225.
62. Montecatini Edison, DE 1 947 419, 1968 (I. Dahli et al.).
63. Celanese, JP-Kokai 18506, 1978, US 4 072 727, 1978 (T. H. Vanderspurt).
Stamicarbon, US 3 686 333, 1972 (C. J. Duyrerman).
64. Shell Development Co., US 2 767 221, 1953 (A. B. Seaver, H. de V. Finch).
65. A. Shigaki, *Chem. Econ. Eng. Rev.* **2** (1970) 8, 47.
66. Mitsubishi Petrochemicals, EP 582 277, 1993 (Y. Watanabe, M. Kurashige).
67. M. R. Johnson, B. Rickborn, *J. Org. Chem.* **35** (1970) 1041.
68. H. S. Broadbent, G. C. Campbell, W. B. Bartley, J. H. Johnsons, *J. Org. Chem.* **24** (1959) 1847.
Standard Oil (Indiana), US 4 292 452, 1978 (R. J. Lee, D. H. Meyer, D. M. Senneke).
P. N. Rylander, D. R. Stelle, *Tetrahedron Lett.* **20** (1969) 1579.
Celanese, US 4 020 116, 1976,
DE 2 734 811, 1976,
US 4 127 508, 1978,
JP-Kokai 18506, 1978 (T. H. Vanderspurt).
Y. Nagase, H. Hittori, K. Tanabe, *Chem. Lett.* 1983, 1615.
69. G. B. Payne, *J. Am. Chem. Soc.* **80** (1958) 6461,
81 (1959) 4501.
70. Shell Oil, US 2 941 006, 1957 (Ch. R. Greene).
71. R. C. Schulz: "Acrolein Polymers" in *Encyclopedia of Polymer Science and Engineering*, 2nd ed., vol. **1** John Wiley, New York 1985.
72. Biopolymers Ltd., PCT/WO 8 804 671, 1988 (G. J. H. Melrose, C. M. Kleppe, J. W. Langley, J. M. Stewart).
73. Degussa, EP 97 101 161, 1997 (P. Werle, H.-P. Krimmer, M. Trageser, F.-R. Kunz).
74. H. Morita, *Reza Kenkyu* **31** (2003) 103–108.
75. H. Schulz, H. Wagner, *Angew. Chem.* **62** (1950) 105.
76. Standard Oil, US 2 941 007, 1960 (J. L. Callahan, R. W. Foreman, F. Veatch).
77. C. R. Adams, *Chem. Ind. (London)* **52** (1970) 1644.
J. L. Callaghan, R. K. Grasselli, E. C. Milberger, H. A. Strecker, *Ind. Eng. Chem. Prod. Res. Dev.* **9** (1970) 134.
Sohio, US 2 904 580, 1959 (J. D. Idol).
78. Shell Development Co., US 2 451 485, 1948 (G. W. Hearne, M. L. Adams).
79. D. J. Hucknall: *Selective Oxidations of Hydrocarbons*, Academic Press, London-New York 1974.
80. L. F. Hatch, S. Matar, *Hydrocarbon Process.* **57** (1978) 6, 149.
81. T.A. Hanna, *Coord. Chem. Rev.* **24** (2004) 8429–440.
82. R.K. Grasselli, *Top. Catal.* **21** (2002) 79–88.
83. W. Ueda, M.-O. Yoshihiko, *Adv. Catal.* **40** (1994) 233–273.
84. S. B. Abd Hamid et al., *Top. Catal.* **24** (2003) 87–95.
85. R. K. Grasselli in: G. Ertl, H. Knozinger, J. Weitkamp (eds.): *Handbook of Heterogenous Catalysis*, Wiley-VCH, Weinheim 1997, p. 2302.

86. P. Mars, D. W. Krevelen, *Chem. Eng. Sci. (Spec. Suppl.)* **3** (1954) 41.
87. Knapsack, US 3 171 859, 1965 (K. Sennewald, W. Vogt, K. Gehrmann, S. Schafer).
88. D. P. Shashikin, O. V. Udalova, M. D. Shibanova, O. V. Krylov, *Kinet. Catal.* **46** (2005) 535–544.
89. D. P. Shashikin, O. V. Udalova, M. D. Shibanova, O. V. Krylov, *Kinet. Catal.* **46** (2005) 545–549.
90. I. Matsuura, M. W. Wolfs, *J. Catal.* **51** (1979) 174.
91. T. Ono, K. Utsumi, M. Kataoka, Y. Tanaka, F. Noguchi, *Catal. Today 91-92* (2004) 181–184.
92. Surveys on bismuth molybdate catalyst developments: G. W. Keulks: "Selective Oxidation of Propylene", in G. W. Smith (ed.): *Catalysis in Organic Syntheses*, Academic Press, New York 1977.
R. Higgins, P. Hayden, *Catalysis* (London) 1 (1977) 168.
T. Ohara, *Shokubai* **19** (1977) 3, 157; *Chem. Abstr.* **87** (1977) 151636.
I. Matsuura, *Shokubai* **21** (1979) 6, 409.
H. Offermanns, G. Prescher in Houben-Weyl: *Methoden der Organischen Chemie*, vol. **E 3: Aldehyde**, p. 231, Georg Thieme Verlag, Stuttgart 1983.
R. K. Grasselli, J. D. Burrington in W. R. Moser (ed.): *Catalysis of Organic Reactions* (Chem. Ind. Ser., vol. 5), Marcel Dekker, New York 1981.
93. Y. Moro-Oka, W. Ueda, *Adv. Catal.* **40** (1994) 233.
94. LG Chem. Ltd, WO 2005/079980 (J.-H. Kang, W.-H. Lee, M.-H. Kil, S.-H. Moon, B.Y. Jo).
95. Saudi Basic Industries Corporation, WO 2000009260,
US 6143928 (Kh. Karim, Y.S. Bhat, S.I. Zaheer, A.B. Nafisa).
96. BASF AG, EP 1005908 1998, (H. Arnold; K. Harth, H.-P. Neumann, U. Hammon, R. Felder).
97. Nippon Shokubai Kagaku Kogyo Co., Ltd, EP 1055455 2000, (T. Kimura, M. Tanimoto, H. Onodera).
98. Nippon Kayaku Co., Ltd., EP 807465 1997, (K. Wada, A. Iwamoto, Y. Seo, A. Sudo, F. Sakai, K. Shiraishi, H. Miki).
99. Mitsubishi Chem. Corp., US 2005159620 (N. Kanuka, I. Teshigahara).
100. G. E. Schaal, *Hydrocarbon Process.* **52** (1973) 9, 218.
101. W. M. Weigert, *Chem. Eng.* **80** (1973) 15, 68.
102. PCUK, Inf. Chim., Spec. Exp. 97 (1977/78).
103. Degussa, DE 1 910 795, 1969 (H. Hillenbrandt, E. Liebetanz, Th. Lüssling, E. Noll, K. Simon).
104. Degussa, DE 2 263 496, 1972 (E. Noll, H. Schaefer, H. Schmid, W. Weigert).
105. V. C. Corberan, *Catal. Today* **99** (2005) 33–41.
106. J. B. Wagner et al., *J. Catal.* **225** (2004) 78–85.
107. H. Jiang, W. Lu, H. Wan, *Catal. Commun.* **5** (2004) 29–34.
108. C. J. Huang, W. Guo, X. D. Xi, W. Z. Weng, H. L. Wan, *Stud. Surf. Sci. Catal.* **147** (2004) 661–666.
109. R. K. Grasselli, *Catal. Today* **99** (2005) 23–31.
110. D. Vitry, J.-L. Dubois, W. Ueda, *J. Mol. Catal. A: Chemical* **220** (2004), 67–76.
111. X. Yi et al., *Chn. J. Catal.* **24** (2004) 755;
L. Chen, J. Liang, H. Lin, W. Weng, H. Wan, J. Vedrine, *Appl. Catal. A: General* **293** (2005) 49–55.
112. A. Costine, B. K. Hodnett, *Appl. Catal. A: General* **290** (2005) 9–16.
113. M. M. Bettahar, G. Costentin, L. Savary, J. C. Lavalley, *Appl. Catal. A* **145** (1996) 1 – 48.
114. Unverricht, Signe, BASF AG, US 6426433 2002, (O. Machhammer, A. Tenten, H. Jachow, S. Haupt, H. Arnold).
115. Union Carbide Chem & Plastics Technology, US 6166263 2000, (W. G. Etzkorn et al.).
116. Union Carbide Chemicals and Plastics Technology Corporation, US 6187963 2001, (W.G. Etzkorn et al.).
117. N. Song et al., *Journal of Catalysis* **236** (2005), 282–291.
L. Wang, S. Zhang, Z. Li, G. Li, M. Li, *Huagong Xuebao (Chinese Edition)* **55** (2004), 2082–2085.
Mitsubishi Rayon Co., Ltd., WO 2002098827 (S. Watanabe, M. Oh-kita, T. Sato).
Mitsui Toatsu Chemicals, Inc., EP 460932 (T. Honda et al.).
Saudi Basic Ind. Corp., US 2005159621 (J. W. Kaufman et al.).
118. N. Shimizu, *Petrotech (Tokyo)* **6** (1983) 9, 778.
119. *Chem. Econ. Eng. Rev.* **15** (1983) 4, 49.
120. T. Hatsuike, H. Matsuzawa, *Hydrocarbon Process.* **58** (1979) 2, 105.
Y. Oda et al., *Hydrocarbon Process.* **54** (1975) 10, 115.
121. *Process Economics Programm Report* **255** (2004), p. 4–16 to 4–17.
122. *Chemical Week* **164** (2002) no.11, **15**(1).
123. *Process Economics Programm Report* **255** (2004), p. 3–8 to 3–9.

124. BASF, EP 58 927, 1982 (F. Merger, H. J. Foerster), EP 92 097, 1982 (G. Duembgen, G. Fouguet, R. Krabetz, E. Lucas, F. Merger, F. Nees).
M. Tramontini, *Synthesis* 1973, 703.
Celanese, US 2 848 499, 1956 (A. F. Mac Lean, B. G. Frenz).
Ruhrchemie, DE 2 855 504, 1978 (W. Bernhagen et al.).
125. Standard Oil, Ohio, US 4 433 174, 1984 (G. P. Hagen).
126. Mitsubishi Chemical Ind., JP-Kokai 138499, 1977, US 4 146 574.
127. Standard Oil Co, Ohio, WO 2005000463 (R. K. Grasselli, C. G. Lugmair, J. Zysk).
128. Degussa(company publication): *Acrolein,* Frankfurt Manufacturing Chemists Assoc.: *Acrolein,* Chemical Safety Data Sheet, SD-85, 1961. G. Hommel: *Handbuch der gefährlichen Güter,* 3rd ed., vol. 1, Merkblatt 218, Springer-Verlag, Berlin 1980. Kühn-Birett: *Merkblätter gefährliche Arbeitsstoffe,* Verlag moderne Industrie, München 1981.
129. U. Harcher, *Chem. Ztg.* **99** (1975) 54.
Rhône Poulenc, US 3 023 247, 1960, GB 850360, 1960.
International Flavors and Fragrances, US 3 185 629, 1965.
130. Shell, US 2 959 476, 1969 (J. van Overbeek).
131. C. v. Oppel, A. Schiffers, *Energie* **26** (1974) 173.
132. J. L. Brady, C. L. Kissel: "Acrolein in Irrigation Waterways," Proc. Western Aquatic Plant Management Society, 1st Symposium, Denver, Colo. 1982.
133. J. M. Donokue, A. J. Piluso, J. R. Schieber, *Mater. Prot.* **5** (1966) 22.
134. Magna Corp., US 4 215 147, 4 215 148, 1978 (C. L. Kissel, F. F. Caserio).
135. Monsanto, US 4 353 924, 1979 (J. W. Beher, D. L. Mansfield, D. J. Weinkauff).
136. Monsanto, *Chem. Week* **127** (Jul. 16, 1980) 3, 37.
137. Du Pont, US 4 335 257, 1980 (E. W. Cummins, S. I. Gleich, R. M. Vigilant).
138. Polaroid, US 3 932 360, 1974 (L. D. Cerankowski, N. Mattucci, R. C. Baron).
R. C. Baron, R. E. Brooks, K. C. Frisch: *Trimethylene Glycol p-Aminobenzoate*, Can. Ureth. Manufacturers Assoc., April 1978.
139. Chevron Research, US 3 819 521 (M. J. Sims).
140. Union Carbide, US 2 941 859, 1960 (M. L. Fein, E. M. Filachione).
BASF, DE 2 215 948, 1972 (H. Erdmann, F. F. Miller).
141. Union Carbide, US 4 244 876, 1978 (G. H. Warrer, L. F. Theiling, M. G. Freid).
142. Schülke & Mayr, DE 3 032 794, 1980 (W. Münzenmaier et al.).
S. Sankara Subramanian, *Sci. Res. News* **1** (1978) 42.
143. A. Kleemann, *Chem. Ztg.* **191** (1979) 389.
144. D. Spitzner, *Science of Synthesis* **15** (2005) 11–284.
145. BASF, DE 3 121 349 ,1981 (W. Graulich, W. Himmele, Ch. Martin, E.-H. Pommer, H. Siegel).
146. H. Bolens, J. Heydel, *Chem. Ztg.* **97** (1973) 8.
Fritzsche Dodge and Olcott, US 4 287 100, 1978 (K. Kulka, T. Zazulka, J. M. Yurecko).
147. International Flavors & Fragrances, US 4 007 137, 1977 (J. M. Sanders, W. L. Schreiber, J. B. Hall).
148. International Flavors and Fragrances, DE 2 643 062, 1976 (J. M. Sanders et al.).
149. Takasago Perf. Co., DE 2 833 283, 1977 (T. Kobayashi, H. Tsuruta, T. Yoshida).
150. Hoechst, EP 275 957, 1988, (E. Strauch, W. Arnold, R. Alijah, W. Wohlleben, A. Pühler, P. Eckes, G. Donn, E. Uhlmann, F. Hein, F. Wengenmayer).
151. Degussa, DE 1 071 339, 1959 (K. H. Rink).
152. H. Haschke, G. Morlock, P. Kuzel, *Chem. Ztg.* **96** (1972) 199.
153. M. M. Ishanov, U. A. Azizov, M. Nigmankhodzhaeva, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 1013.
154. American Cyanamid, US 3 819 555, 1974 (E. D. Kaufmann).
155. GAF, US 3 595 663, 1971 (S. Emmi), US 3 615 623, 1971 (N. D. Field, D. I. Randall, J. D. Fitzpatrick).
156. Degussa, DE 3 205 487, 1982 (K. H. Rink, W. Merk).
157. Commission Dir. 2001/59/EC, OJ. Eur. Comm. L225 adapting to technical progress for the 28th time Council Directive 67/548/EC, 2001.
158. Beratergremium für umweltrelevante Altstoffe der Gesellschaft deutscher Chemiker: *Acrolein (2-Propenal)*, BUA Stoffbericht Nr. 157, S. Hirzel Wiss. Verlagsges., 1995.
159. T. B. Albin in C. W. Smith (ed.): *Acrolein* Hüthig, Heidelberg 1975.
160. V. J. Feron, A. Kruysse, H. P. Til, H. R. Immel, *Toxicology* **9** (1978) 47 – 57.
161. R. S. Kuztmann, E. A. Popenoe, M. Schmaeler, R. T. Drew, *Toxicology* **34** (1985) 139 – 151.
162. R. S. Kutzmann, R. W. Wehner, S. B. Haber, *Toxicology* **31** (1984) 53 – 65.

163. European Commission, Institute for Health and Consumer Protection, European Chemicals Bureau, Risk Assessment Report Acrylaldehyde, Office for Official Publications of the European Communities, Luxembourg 2001.
164. International Agency For Research on Cancer, World Health Organisation, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, vol. 63, Dry Cleaning, some Chlorinated Solvents and other Industrial Chemicals, WHO, 1995, pp. 337 – 372.
165. A. Weber-Tschopp, T. Fischer, R. Gierer, E. Grandjean, *Z. Arbeitswiss.* **32** (1977) 166 – 171.
166. TRGS 900 Grenzwerte in der Luft am Arbeitsplatz, BArbBl. Nr. 10 (1996) 88, Nr. 4 (1997) 42, Nr. 11 (1997) 27. Ausschuss für Gefahrstoffe - AGS-Geschäftsführung - BAuA - www.baua.de, Ausgabe Januar 2006.
167. Health and Safety Executive (HSE), EH 40/97 Occupational Exposure Limits 1997.
168. ACGIH, American Conference of Governmental Industrial Hygienists, Threshold Limit Values and Biological Exposure Indices, 2006.
169. Berufsgenossenschaft der chemischen Industrie Toxikologische Bewertung Methacrolein, Nr. 108, 1995.
170. S. T. Larsen, G. D. Nielsen, *Toxicol. Lett.* **114** (2000) 197–202.
171. L. T. Iraci, B. M. Baker, G. S. Tyndall, J. J. Orlando, *J. Atmospheric Chem.* **33** (1999) 321–330.
172. V. T. Stack, Jr., *Ind. Eng. Chem.* **49** (1957) 913 – 917.
173. M. D. Barrat, *Toxicol. in Vitro* **10** (1996) 247 – 236.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Acrylic Acid and Derivatives

Standard Article

Takashi Ohara[1], Takahisa Sato[2], Noboru Shimizu[3], Günter Prescher[4], Helmut Schwind[5], Otto Weiberg[6], Klaus Marten[7], Helmut Greim[8]

[1]Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
[2]Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
[3]Nippon Shokubai Kagaku Kogyo Co., Ltd., Osaka, Japan
[4]Degussa AG, Zweigniederlassung Wolfgang, Hanau, Federal Republic of Germany
[5]Degussa AG, Zweigniederlassung Wolfgang, Hanau, Federal Republic of Germany
[6]Degussa AG, Zweigniederlassung Wolfgang, Hanau, Federal Republic of Germany
[7]Sichel-Werke GmbH, Hannover, Federal Republic of Germany
[8]Institut für Toxikologie und Umwelthygiene, TU München, Freising-Weihenstephan, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a01_161.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Acrylic Acid and Esters
1.1. Physical Properties
1.2. Chemical Properties
1.3. Production
1.3.1. Propene Oxidation
1.3.2. Esterification
1.4. Quality Specifications and Analysis
1.5. Storage and Transportation
1.6. Uses
1.7. Some Special Acrylates
1.8. Economic Aspects
1.9. Toxicology and Occupational Health
2. Cyanoacrylates
3. Acrylamide

[Top of Page]

1. Acrylic Acid and Esters

Acrylic acid [79-10-7], 2-propenoic acid, CH_2=CHCOOH, and its esters CH_2=CHCOOR, which are also known as acrylates, are flammable, volatile, mildly toxic, colorless liquids. Hydroquinone or its monomethyl ether is usually added to commercial preparations to inhibit polymerization. Formerly, acrylic acid and acrylates were produced industrially via a variety of routes such as acrylonitrile hydrolysis and the modified Reppe process (see Section Production). However, remarkable progress on the catalytic oxidation of propene to acrylic acid via acrolein has led to almost complete replacement of these earlier processes.

Esters such as methyl, ethyl, *n*-butyl, and 2-ethylhexyl acrylates, as well as acrylic acid, are in worldwide use, primarily for polymers. Other esters, including multifunctional acrylates, are produced for special applications.

Chemically, *acrylamide* (see Chap. Acrylamide) is a derivative of acrylic acid but the amide is produced by hydration of acrylonitrile instead of by amidation of the acid.

1.1. Physical Properties

Acrylic Acid is a clear, colorless liquid, *bp* 141.0 °C (101.3 kPa), *mp* 13.5 °C; it forms crystalline needles in the solid state. Other important physical constants are listed below [1-6]:

M_r	72.06
Refractive index	n_D^{20} 1.4224, n_D^{25} 1.4185
Density	1.060 (10 °C), 1.040 (30 °C), 1.018 (50 °C) g/cm^3
Viscosity at 25 °C	1.149 mPa · s
Critical temperature	380 °C
Critical pressure	5.06 MPa
Heat of vaporization at 101.3 kPa	45.6 kJ/mol
Heat of combustion	1376 kJ/mol
Heat of melting at 13 °C	11.1 kJ/mol
Heat of neutralization	58.2 kJ/mol
Heat of polymerization	77.5 kJ/mol
Dissociation constant at 25 °C	5.5×10^{-5}; $pK_a = 4.26$

Vapor pressure as function of temperature:

t, °C	0	20	40	60	100	120	141
p, kPa	0.31	1.03	2.93	7.2	33.2	63.3	101.3

Acrylic acid is highly miscible with water, alcohols, esters, and many other organic solvents. Figure 1 gives the density of the aqueous solution as a function of water content. Table 1 shows the freezing points of various acetic acid – acrylic acid and water – acrylic acid solutions.

Table 1. Freezing points of acrylic acid mixtures: A with acetic acid, B with water

System A, wt % acetic acid	Freezing point,°C	System B, wt % water	Freezing point, °C [5]
0	13.5	0	13.5
10	7.5	5	5.5
20	0.7	10	1.0
40	–14.1	20	– 5.5
50	–23.5	30	–10.3
50.2	–23.8	37	–12.5
60	–13.4	40	–12.0
80	3.7	60	– 8.0
100	16.6	80	– 4.0
		100	0

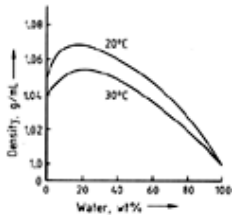

Figure 1. Relationship between density of aqueous acrylic acid solution and water content

Derivatives. Table 2 lists physical properties of representative derivatives other than esters, Table 3 those of five commercial acrylates, and Table 4 those of other acrylates including some diesters.

Table 2. Physical properties of acrylic acid derivatives

	Acrylic anhydride	Acryloyl chloride	Acrylamide
CAS registry number	[2051-76-5]	[814-68-6]	[79-06-1]
Structural formula	$(CH_2 = CHCO)_2O$	$CH_2 = CHCOCl$	$CH_2 = CHCONH_2$
Molecular formula	$C_6H_6O_3$	C_3H_3ClO	C_3H_5NO
M_r	126.11	90.51	71.08
mp, °C			84.5
bp, °C/*p* in kPa	38/0.27	75/101	125/16.6
Density, g/cm^3		1.113 (20 °C)	1.122 (30 °C)
Refractive index, n_D^{20}	1.4487	1.4337	

Table 3. Physical properties of the most important acrylates

Property	Methyl acrylate	Ethyl acrylate	*n*-Butyl acrylate	Isobutyl acrylate	2-Ethylhexyl acrylate
CAS registry number	[96-33-3]	[140-88-5]	[141-32-2]	[106-63-8]	[103-11-7]
Molecular formula	$C_4H_6O_2$	$C_5H_8O_2$	$C_7H_{12}O_2$	$C_7H_{12}O_2$	$C_{11}H_{20}O_2$
M_r	86.09	100.12	128.17	128.17	184.28
mp, °C	–76	–72	–64.6	–61	–90
bp at 101.3 kPa, °C	80.3	99.4	147.4	138	216
Specific heat (l), kJ mol^{-1} K^{-1}	0.48	0.47	0.46	0.46	0.46
Solubility at 25 °C					
in water (g/100 g)	5	1.5	0.2	0.2	0.01
of water in ester (g/100 g)	2.5	1.5	0.7	0.6	0.15
Azeotropes					
with water, *bp*, °C	71	81.1	94.5		
water content, wt %	7.2	15	40		
with methanol, *bp*, °C	62.5	64.5			
methanol content, wt %	54	84.4			
with ethanol, *bp*, °C	73.5	77.5			
ethanol content, wt %	42.4	72.5			
with *n*-butanol, *bp*, °C			119		
n-butanol content, wt %			89		
Heat of vaporization at *bp*, kJ/mol	33.2	34.8	36.5	38.1	47.0
Heat of polymerization, kJ/kg	84.7	77.9	77.3		60.1
Vapor pressure, kPa					
at 0 °C	4.2	1.2	0.14		
at 20 °C	9.3	3.9	0.44		
at 50 °C	35.9	17.3	2.82		0.16
at 100 °C			21.9		2.1
at 150 °C					14.6
Refractive index, n_D^{20}	1.4040	1.4068	1.4190	1.4150	1.4365
Relative density, d_4^{20}	0.9535		0.8998		0.8852
d_{20}^{20}	0.9565	0.9231	0.9015	0.890	0.8869

Viscosity, mPa s					
at 20 °C	0.53	0.69	0.90	0.78	1.7
at 25 °C	0.49	0.55	0.81		1.54
at 40 °C		0.50	0.70		1.2
Autoignition temperature, °C	393	355	267	340	230
Flammability range in air, vol %	2.8 – 25	1.8 – saturated	1.5 – 9.9	1.9 – 8.0	0.6 – 1.8
Flash point					
closed cup, °C	–3	9	41	33	87
open cup, °C	–2	19	47		92

Table 4. Physical properties of acrylic esters and diesters

Ester	CAS registry number	Molecular formula	M_r	*bp*, °C / *p*, kPa	Refractive index, n_D^{20}	Relative density, n_4^{20}
n-Propyl	[925-60-0]	$C_6H_{10}O_2$	114.15	44/5.3	1.4130	0.9078
n-Pentyl	[2998-23-4]	$C_8H_{14}O_2$	142.20	48/0.9	1.4240	0.8920
n-Hexyl	[2499-95-8]	$C_9H_{16}O_2$	156.23	40/0.2	1.4280	0.8882
n-Heptyl	[2499-58-3]	$C_{10}H_{18}O_2$	170.25	57/0.1	1.4311	0.8846
Isopropyl	[689-12-3]	$C_6H_{10}O_2$	114.15	52/14	1.4060	0.8932
sec-Butyl	[2998-08-5]	$C_7H_{12}O_2$	128.17	60/6.7	1.4140	0.8914
tert-Butyl	[1663-39-4]	$C_7H_{12}O_2$	128.17	120/101.3	1.408	0.879
Allyl	[999-55-3]	$C_6H_8O_2$	112.13	47/5.3	1.4320	0.9441
2-Hydroxyethyl	[818-61-1]	$C_5H_8O_3$	116.12	74/0.7	1.4505	1.1038 (25 °C)
2-Hydroxypropyl	[999-61-1]	$C_6H_{10}O_3$	130.14	77/0.7	1.4443	1.5036
Ethylene glycol diester	[2274-11-5]	$C_8H_{10}O_4$	170.17	70/0.1	1.4529	
1,2-Propanediol diester	[25151-33-1]	$C_9H_{12}O_4$	184.19	60/0.04	1.4470	
1.4-Butanediol diester	[31442-13-4]	$C_{10}H_{14}O_4$	198.22	83/0.1	1.4538	

1.2. Chemical Properties

Acrylic acid and its esters undergo reactions characteristic of both unsaturated compounds and aliphatic carboxylic acids or esters. The high reactivity of these compounds stems from the two unsaturated centers situated in a conjugated position. The β carbon atom, polarized by the carbonyl group, behaves as an electrophile; this favors the addition of a large variety of nucleophiles and active hydrogen compounds to the vinyl group. Moreover, the carbon-carbon double bond undergoes radical-initiated addition reactions, Diels–Alder reactions with dienes, and polymerization reactions.

The carboxyl function is subject to the displacement reactions typical of aliphatic acids and esters, such as esterification and transesterification.

Joint reactions of the vinyl and carboxyl functions, especially with bifunctional reagents, often constitute convenient routes to polycyclic and heterocyclic substances.

Acrylic acid and its esters polymerize very easily. The polymerization is catalyzed by heat, light, and peroxides and inhibited by stabilizers, such as the monomethyl ether of hydroquinone or hydroquinone itself. These phenolic inhibitors are effective only in the presence of oxygen. The highly exothermic, spontaneous polymerization of acrylic acid is extremely violent.

In this section are listed typical examples of reactions other than polymerization, which is discussed in Section Uses. Several review articles and monographs [1-7] describe the rich chemistry of acrylates and acrylic acid.

Addition Reactions. Acrylic acid and acrylates combine readily with substances, such as hydrogen, hydrogen halides and hydrogen cyanide, that customarily add to olefins [8]:

$$H_2C = CHCOOR + HX \rightarrow H_2CX - CH_2COOR$$

where R = H, alkyl, or aryl, and X = H, halogen, or CN.

Michael additions of organic substances take place in the presence of basic catalysts, such as tertiary amines, quaternary ammonium salts, and alkali alkoxides:

$$\text{where } X = -C(NO_2)R'R'',\ -CH(COOR')_2,\ -CH\begin{matrix}COOR'\\COR''\end{matrix},\ -CH\begin{matrix}CN\\COOR'\end{matrix},\ \text{or } -CH\begin{matrix}CN\\C_6H_5\end{matrix}$$

and R′, R″ = alkyl or aryl.

Ammonia and amines are sufficiently basic to react without a catalyst: where X = $–NH_2$, –NHR′ [9], [10], –NR′R″ [11], heterocycles [12-14], –NR′COR″, or $–NHNR_2$ [15]. The addition of only one molecule of NH_3 (for addition of two, see below) can be achieved with an aqueous solution of ammonia and ammonium carbonate [16].

The addition of aromatic amines or amides and *tert*-alkyl primary amines is more effectively promoted by acids. Amines may attack both the vinyl and carboxyl functions, but the products of such reactions decompose to give N-substituted amides.

Alcohols [17], phenols, hydrogen sulfide [14], [18], and thiols [19], [20] also add under basic conditions. Hydrogen sulfide in the presence of sulfur and ammonium polysulfide or amine catalysts gives polythiodipropionic acids and esters [19], [21]:

$$CH_2{=}CH{-}COOR \xrightarrow[S,\,H_2S]{(NH_4)_2S_x} S_x(CH_2CH_2COOR)_2$$

where R = H, alkyl, or aryl.

Other examples of HX additions to acrylic acid and acrylates are:

$$\text{where } X = -O_2SR',\ -SO_3Na,\ -OP(OR')_2,\ -\overset{O}{\overset{\|}{C}}{-}R'\ [22],[23],$$

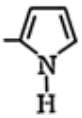

, or

(24)

and R′ = alkyl or aryl [22-24].

Additions of aromatic hydrocarbons are promoted more efficiently by Lewis acids [25].

If further acidic hydrogen atoms are available in the addition product, a second (and third) molecule of acrylic acid or ester adds. This is the case in the reaction of acrylic acid or ester with H_2S, NH_3, RNH_2, and pyrrole.

Other examples of addition reactions are the following:

$$H_2C{=}CH{-}COOR + Hal_2 \longrightarrow \underset{Hal}{CH_2}{-}\underset{Hal}{CH}{-}COOR\ [26]$$

$$+ HOCl \longrightarrow \underset{OH}{CH_2}{-}\underset{Cl}{CH}{-}COOR\ [3]$$

$$+ Hg(OCOCH_3)_2 + CH_3OH \longrightarrow \underset{OCH_3}{CH_2}{-}\underset{HgOCOCH_3}{CH}{-}COOR + CH_3COOH\ [3]$$

$$+ C_6H_5N_2^+Cl^- \longrightarrow \underset{C_6H_5}{CH_2}{-}\underset{Cl}{CH}{-}COOR + N_2\ [3]$$

$$+ CH_2{=}C(CH_3)_2 \longrightarrow CH_2{=}\underset{CH_3}{C}{-}CH_2CH_2CH_2COOR\ [3],\ [27]$$

where R = H, alkyl, or aryl [3, 26, 27].

Heterocyclic substances often can be formed by subsequent reaction of the carboxyl function, especially with bifunctional nucleophiles [23], [28]:

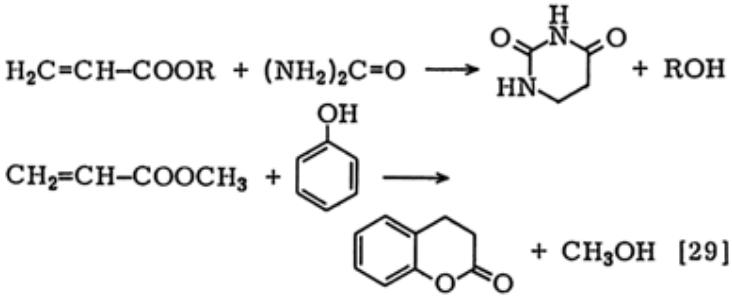

where R = H, alkyl, or aryl.

Substituted ring compounds are formed readily by Diels–Alder reactions [30], [31]:

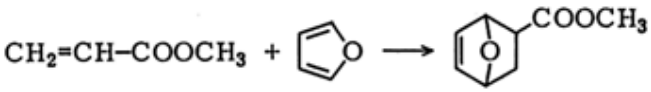

Acrylates also undergo cobalt- or rhodium-catalyzed hydroformylation reactions [32], [33]:

$$2\,CH_2{=}CH{-}COOR + 2\,CO + 2\,H_2 \xrightarrow{\text{catalyst}} CH_2(CHO){-}CH_2COOR + CH_3CH(CHO){-}COOR$$

where R = alkyl or aryl.

At elevated temperature or on longer storage acrylic acid dimerizes:

$$2\ CH_2 = CH - COOH \rightarrow CH_2 = CH - COOCH_2CH_2COOH$$

In the presence of catalysts such as tributylphosphine, acrylates can also dimerize to give 2-methyleneglutarates [34]:

$$2\,CH_2{=}CH{-}COOR \xrightarrow{\text{catal.}} ROOC{-}\overset{\overset{\displaystyle CH_2}{\|}}{C}{-}CH_2CH_2{-}COOR$$

where R = alkyl or aryl.

Reactions of the Carboxyl Group. Acrylic acid is converted readily into its corresponding salts, into acrylic anhydride by reaction with acetic anhydride, or into acryloyl chloride by reaction with benzoyl or thionyl chloride. The esterification of acrylic acid and transesterification of acrylic esters are economically the most important reactions (see Section Esterification).

Some other examples are:

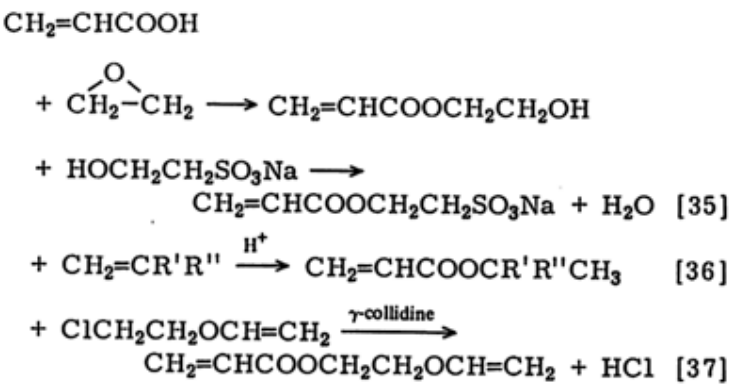

where R′, R″ = alkyl or aryl [35-37].

1.3. Production

Commercial acrylic acid is mostly produced from propene, which is also the raw material for the production of acrolein. In the past, acrylic acid and its esters were produced by various processes some of which are summarized here (see [5], [6], [38], [39]) and are still in use to a small extent.

Processes Based on Acetylene (⟶ Acetylene). The stoichiometric synthesis of acrylic acid and its esters from acetylene proceeds at atmospheric pressure and at 40 °C in the presence of acid and nickel carbonyl:

$$4\ C_2H_2 + 4\ ROH + 2\ HCl + Ni(CO)_4 \rightarrow 4\ CH_2 = CHCOOR + NiCl_2 + H_2$$

where R = H, alkyl, or aryl.

The reaction was discovered by W. REPPE in 1939. Röhm & Haas and Toa Gosei Chemical have used this method as well as the modified, non-stoichiometric Reppe process, but both have been abandoned because of the difficulties in handling the toxic and corrosive nickel carbonyl.

High-Pressure Reppe Process. The process employed by BASF and Badische Corp. proceeds at approximately 14 MPa and 200 °C with a nickel bromide – copper(II) bromide catalyst:

$$C_2H_2 + CO + H_2O \xrightarrow{\text{catal.}} CH_2{=}CHCOOH$$

However, the safety and pollution control problems with nickel carbonyl (formed in the process) and the high cost of acetylene are disadvantages of this process. It has largely been replaced by the direct oxidation of propene although BASF still produces part of its acrylic acid by this process.

Acrylonitrile Hydrolysis (⟶ Acrylonitrile). This method is economically unattractive because of the low yield based on propene and the large quantities of NH_4HSO_4 waste. The process has been abandoned by Ugine Kuhlmann, Mitsubishi Petrochemical, and Mitsubishi Rayon. However, it is still on stream at Asahi Chemical.

Ketene Process [6], [40]. Acetic acid or acetone is pyrolyzed to ketene in this process which has long been abandoned by Celanese and B. F. Goodrich. The many steps and toxicity of β-propiolactone are major disadvantages.

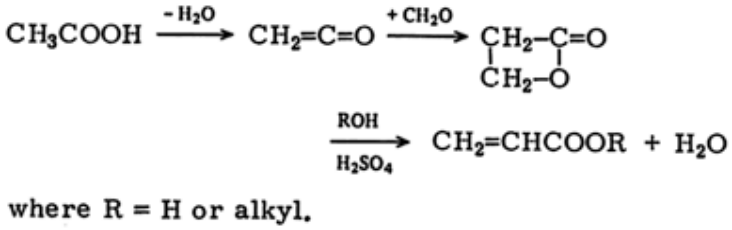

R = H or alkyl.

Ethylene Cyanohydrin Process. Ethylene cyanohydrin is generated by addition of hydrogen cyanide to ethylene oxide. The product then is hydrolyzed to acrylic acid using sulfuric acid. This process was used by Union Carbide and Röhm & Haas, but has been abandoned because of problems in dealing with HCN and the NH_4HSO_4 waste.

1.3.1. *Propene Oxidation*

Propene oxidation involves heterogeneous catalytic oxidation of propene in the vapor phase with air and steam to give acrylic acid. Generally the product leaving the reactor is absorbed in water, extracted with an appropriate solvent, and then distilled to give technical grade glacial acrylic acid.

Oxidation Catalysts. Research on catalysts for propene oxidation to acrylic acid began in the latter half of the 1950 s. The two methods for the heterogeneously catalyzed gas-phase oxidation of propene are single-step and two-step processes:

Single-step process:

$$CH_2{=}CHCH_3 + 3/2\ O_2 \rightarrow CH_2{=}CHCOOH + H_2O \quad \Delta H = -594.9\ \text{kJ/mol}$$

Two-step process:

$$CH_2{=}CHCH_3 + O_2 \rightarrow CH{=}CH_2CHO + H_2O \quad \Delta H = -340.8\ \text{kJ/mol}$$

$$CH_2{=}CHCHO + 1/2\ O_2 \rightarrow CH_2{=}CHCOOH \quad \Delta H = -254.1\ \text{kJ/mol}$$

Many patents have been issued in both cases. The yield in the *single-step process* is at best approximately 50 – 60 % [38], [41-43]. Another drawback is limited lifetime of the catalyst, which is a multicomponent system composed of polyvalent oxides with molybdenum oxide as the main component and tellurium oxide as the promoter. The life of the catalyst is short because of the tendency of tellurium oxide to sublime.

The *two-step reaction* (Fig. 2) requires different reaction conditions and different catalysts to produce optimum conversion and selectivity in each step. Research has focused on this process, in which the oxidation of propene to acrolein and the oxidation of acrolein to acrylic acid employ separate catalysts. The steps are operated at different temperatures to permit high overall efficiency.

Figure 2. Schematic diagram of acrylic acid production (oxidation section)

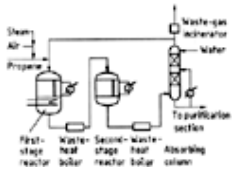

First-stage catalysts are acrolein-selective propene-oxidation catalysts. The total yield of acrolein and acrylic acid is more than 85 % (⟶ Acrolein and Methacrolein).

The early second-stage catalysts [43] for acrolein oxidation to acrylic acid were based mainly on cobalt-molybdenum oxides [44]. They had fairly low activity even at high reaction temperatures and gave yields of less than 70 mol %. Most catalysts are composed of molybdenum and vanadium oxides. In 1959, Distillers first proposed a molybdenum-vanadium catalyst system in which the atomic ratio of molybdenum to vanadium was one-to-one [45]. The maximum yield obtained was 30 % at about 400 °C. Since then further studies have shown that only a relatively small amount of vanadium is required. In addition, other elements and carriers have been shown to increase the activity and yield. They have been used for the preparation of multi-component metal-oxide catalysts that contain one or more of the elements copper, arsenic, uranium, aluminum, tungsten, silver, manganese, germanium, gold, barium, calcium, strontium, boron, tin, cobalt, iron, or nickel in addition to molybdenum and vanadium. Supported on an aluminum sponge, the catalyst described in [46] shows good activity and yield. Table 5 lists patented acrolein oxidation catalysts that have relatively high activities and yields. All of these catalysts are metal oxides.

Table 5. Catalyst for the second step of acrylic acid production

Catalyst composition (support) neglecting oxygen	Reaction temperature, °C	Acrolein conversion, %	One-pass yield of acrylic acid, mol %	References
$Mo_{12}V_{1.9}Al_{1.0}Cu_{2.2}$ (Al sponge)	300	100	97.5	[46]
$Mo_{12}V_3W_{1.2}$ (SiO_2)	240	98.0	87.0	[47]
$Mo_{12}V_3W_{1.2}Mn_3$	255	99.0	93.0	[48]
$Mo_{12}V_2W_2Fe_3$	230	99.0	91.0	[49]
$Mo_{12}V_3W_{1.2}Cu_1Sb_6$	272	99.0	91.0	[50]
$Mo_{12}V_{4.6}Cu_{2.2}W_{2.4}Cr_{0.6}$ (Al_2O_3)	220	100.0	98.0	[51]
$Mo_{12}V_2(Li_2SO_4)_2$	300	99.8	92.4	[52]
$Mo_{12}V_{4.8}Cu_{2.2}W_{2.4}Sr_{0.5}$ (Al_2O_3)	255	100.0	97.5	[53]
$Mo_{12}V_{2.4}Cu_{0.24}$ (SiC)	290	99.5	94.8	[54]
$Mo_{12}V_3W_{1.2}Ce_3$	288	100	96.1	[55]
$Mo_{12}V_{4.7}W_{1.1}Cu_{6.3}$	260	99.0	96.0	[56]

Process Conditions. The conditions in the first step correspond to conditions in the acrolein synthesis (see ⟶ Acrolein and Methacrolein – Production). The catalysts used in the second step require reaction temperatures from 200 to 300 °C and contact times from 1 to 3 s. They give almost 100 % conversion of acrolein and yields of acrylic acid greater than 90 %.

Acid Recovery and Purification. The effluent gas from the second-stage multi-tube reactor in Figure 2 is cooled to about 200 °C and then fed to the absorbing column to be scrubbed with water. Because the effluent gas contains a large amount of steam, acrylic acid usually is obtained as an aqueous solution of 20 to 70 wt % [57]. Alternatively, the acid may be absorbed by an organic solvent such as biphenyl, diphenyl ether, or a carboxylic ester with a boiling point higher than 160 °C [58]. Then the steam in the reaction gas does not condense in the absorbing column, but is discharged with other gases from the column top. This method reduces energy consumption in the subsequent purification step, but it also increases the loss of acrylic acid and solvent from the column top.

After the absorption in water, the acrylic acid is purified by extraction with an organic solvent and then distillation. Various solvents can be used for the extraction. The first group (light solvents) includes those with boiling points lower than acrylic acid, such as ethyl acetate, butyl acetate, ethyl acrylate, and 2-butanone, as well as combinations of these [59]. The second group (heavy solvents) has boiling points higher than acrylic acid (e.g., *tert*-butyl phosphate, isophorone, and aromatic hydrocarbons [60]). Mixtures of these light and heavy solvents form a third group [61], [62].

Figure 3 represents the separation and purification process using a *light extraction solvent*. The aqueous acrylic acid from the absorbing column is introduced into the extraction column (a) countercurrent to an organic solvent. The solvent must have a high distribution coefficient for acrylic acid and low solubility in water, and it must form an azeotrope containing a high

percentage of water. The extract from the top of the extraction column goes to the solvent-separation column (c), where the solvent and water are distilled overhead and the solvent is separated and recycled to the extraction column.

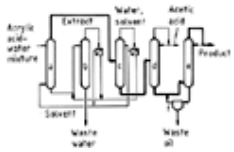

Figure 3. Schematic diagram of acid recovery and purification section

a) Extraction column; b) Raffinate-stripping column; c) Solvent-separation column; d) Light-ends cut column; e) Product column; f) Decomposition evaporator

The bottom stream from the extraction column and the water from the overhead of the solvent-separation column are sent to the raffinate-stripping column (b), where a small amount of solvent is recovered by distillation. The waste water from the raffinate-stripping column is biologically treated or incinerated.

The bottom fraction from the solvent-separation column is fed to the light-ends cut column (d), where acetic acid is distilled off and, if desired, recovered. The crude acrylic acid from the bottom of the light-ends cut column is sent to the product column (e), where acrylic acid of high purity is obtained overhead. The material from the bottom of the product column containing acrylic acid dimer is fed to the evaporator (f), where the dimer is decomposed to the monomer. The evaporator residue, composed of acrylic acid oligomers, polymers, and inhibitors, is withdrawn and burned as waste oil.

Because acrylic acid is readily polymerized, distillation columns are operated with an inhibitor, such as hydroquinone or hydroquinone monomethyl ether, in the presence of oxygen, and at reduced pressure to lower the distillation temperature. The purity of acrylic acid produced by this process usually exceeds 99.5 wt %, and the purified yield is about 98 %.

In a *heavy-solvent extraction process* the solvent is not distilled and therefore the energy consumption is less than in a light-solvent process. However, other problems exist, such as the loss of solvent by decomposition and the inferior quality of the product. In the processes using mixtures of light and heavy solvents, the purification system is complex. The light-solvent process is therefore the most suitable for a commercial plant.

Other purification methods also have been reported. In one of these, the acrylic acid is first oligomerized in aqueous solution in the presence of a catalyst such as sulfuric or phosphoric acid. Next the water is distilled, and finally the residual oligomer is decomposed at 120 to 200 °C to obtain acrylic acid [63]. In another process, the acrylic acid is extracted from the aqueous solution with butyl acrylate or octanol. It is then directly esterified with an alcohol to form an acrylate without isolation of the acrylic acid [64], [65]. These processes have not yet been used commercially, probably because of high energy consumption or problems with the product quality.

1.3.2. *Esterification*

Although acrylic acid can be esterified in the vapor phase [66], [67], the liquid phase esterification is industrially more important. Two types of acid catalyst are used: a strong acid, such as sulfuric acid or *p*-toluenesulfonic acid [68], or a solid acid, such as a cation-exchange resin [69]. Although sulfuric acid is superior to ion-exchange resins, its use causes problems in waste disposal. In general, cation-exchange resins are favored for esterification using such alcohols as methanol and ethanol, whereas sulfuric acid is favored for higher alcohols having slower rates of esterification (e.g., pentanols and octanols). Liquid phase reaction of acrylic acid with ethylene in the presence of sulfuric acid does not seem economically feasible for producing ethyl acrylate [70], because of the large quantities of sulfuric acid that are needed.

Lower Alkyl Acrylates (see Fig. 4). Acrylic acid and a small excess (10 – 30 %) of an alcohol are fed into the fixed-bed reactor (a) which is packed with a cation-exchange resin and operated at a temperature of 60 to 80 °C. The reaction liquid then goes to the ester stripper (b) where the desired ester, water, and unreacted alcohol are removed overhead using part of the bottoms from the light-ends column (e) as reflux. The bottom liquid from b contains unreacted acid and is recycled to the reactor. Part of the recycled liquid is fed into the bottom stripper (c), where high-boiling materials, such as inhibitors, impurities, and polymers, are removed to prevent their accumulation in the reaction system.

The acid-free mixture of ester and alcohol distilled from the ester stripper (b) is fed into the extraction column (d), where the alcohol is extracted with water fed from the top of the column. The raffinate from the top of the column goes into the light-ends cut column (e), where light-ends such as water, acetate, and alcohol are separated overhead.

The extract from the bottom of the extraction column is fed into the alcohol recovery column (f), where the alcohol is recovered for reuse in the reaction. Part of the bottom liquid is reused as extracting water; the rest is taken out as waste, concentrated, and either treated biologically or incinerated.

Crude ester from the bottom of the light-ends column is distilled in the product column (g) to obtain acrylate of high purity. The bottom liquid from the product column is recycled (via the inhibitor tank) to the ester stripper (b) and light-ends cut column (e) to be reused as an inhibitor. However, a part of it is sent to the bottom stripper (c) to recover ester and separate high-boiling materials such as polymers.

Polymerization inhibitors, such as hydroquinone or phenothiazine, are added to each column. The light-ends cut column and the product column are operated at reduced pressure to permit lower distillation temperatures.

Figure 4. Esterification – lower alkyl acrylate process

a) Esterification reactor; b) Ester stripper; c) Bottom stripper; d) Extraction column; e) Light-ends cut column; f) Alcohol-

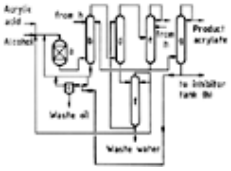

recovery column; g) Product column

This process for making alkyl acrylates is quite economical because only a small excess of alcohol is applied and the inhibitor is reused; this leads to low energy and inhibitor consumptions. The yield reaches 95 % and 97 % based on acrylic acid and on alcohol, respectively. The purity of the product exceeds 99.5 wt %.

Higher Alkyl Acrylates (see Fig. 5). The esterification reaction is preferably carried out batchwise in the presence of an organic solvent as entrainer and sulfuric acid as catalyst. The water formed is separated through the top of the azeotropic-distillation column (b). The reaction conditions are: atmospheric pressure, temperature 85 – 95°C, reaction time 3 – 5 h, molar ratio (alcohol to acid) 1.0 – 1.1.

After completion of the esterification, the reaction liquid is cooled to 60°C and then transferred to tank (c) where the sulfuric acid is neutralized with alkali. The oil and water layers are separated and stored in tanks d and e, respectively. The oil layer is fed into the solvent-recovery column (f) and subsequently into the alcohol-recovery column (g) for distillation. The solvent and alcohol are recovered overhead and reused in the reaction.

The crude ester obtained from the bottom of the alcohol-recovery column is fed into the product column (h) where purified acrylic ester is obtained by distillation. The bottom liquid mainly is recycled to the reactor and the alcohol-recovery column to be reused as supplementary inhibitor. However, part of the liquid is fed into the bottom stripper (i) to recover valuable materials that are resupplied to the product column. High-boiling waste composed of polymers, inhibitors, and other impurities is taken out of the bottom stripper and incinerated.

The water from the water-layer tank (e) is fed into the organic stripper (j) together with the water layer from the top and the bottom of the solvent-recovery column (f). The oil layer obtained from the top of the solvent-recovery column (f) is recycled into the oil-layer tank. The waste obtained from the bottom of the organic stripper (j) is either treated biologically or incinerated after concentration.

As in the lower alkyl acrylate process, hydroquinone, its monomethyl ether, or phenothiazine is added to each column, and the alcohol-recovery and product columns are operated at reduced pressure.

Figure 5. Esterification – higher alkyl acrylate process

a) Esterification reactor; b) Azeotropic-distillation column; c) Neutralization tank; d) Oil-layer tank; e) Water-layer tank; f) Solvent-recovery column; g) Alcohol-recovery column; h) Product column; i) Bottom stripper; j) Organic stripper

The yield reaches 95 % and 96 % based on acrylic acid and on alcohol, respectively. The purity of the product exceeds 99.5 wt %.

1.4. Quality Specifications and Analysis

Production control requires monitoring the propene and oxygen concentrations in the gas phase of the oxidation. These are checked periodically to maintain optimum reaction conditions and avoid entering the range of flammability. Propene is determined by GLC with flame ionization detection, oxygen by a magnetic meter [71].

The purity of acrylic acid and its esters depends on the production method employed. Table 6 shows quality standards for some of these products. The purity of acrylic acid and its esters is commonly determined from the percentage of impurities measured by GLC with a flame ionization detector. Occasionally the purity of acrylic acid is determined by titration with a base.

Table 6. Quality specifications of acrylic acid and esters[a] [6]

		Acrylic acid		Methyl acrylate	Ethyl acrylate	Butyl acrylate	Octyl acrylate
		99 %	80 %, aq.				
Purity [b]	wt %, min	99.0	80.0	99.0	99.0	99.0	99.0
Acid [c]	wt %, max	–	–	0.005	0.005	0.005	0.005
Water	wt %, max	0.20	–	0.05	0.05	0.05	0.05
Color	APHA, max	20	20	20	20	20	20

Inhibitor [d]	ppm	200	200	15 ± 5	15 ± 5	15 ± 5	15 ± 5

[a] Authorized by the Japanese Acrylic Acid and Esters Industrial Association;
[b] determined by GLC;
[c] as acrylic acid;
[d] as hydroquinone monomethyl ether.

In addition to the purity, the polymerization characteristics are important because acrylic acid and acrylates are used chiefly to make polymers. Polymerization characteristics are determined by examining polymerization patterns such as induction period and temperature elevation under fixed polymerization conditions (temperature, concentration, catalyst, etc.) The degree of polymerization of highly purified acrylic acid is greatly decreased in the presence of trace amounts of heavy metals such as copper, or of aldehydes such as acrolein and furfural.

1.5. Storage and Transportation

Acrylic acid and its esters are usually stabilized with inhibitors such as phenothiazine, hydroquinone, or hydroquinone monomethyl ether. Because phenolic inhibitors are only effective in the presence of oxygen, the monomers must be stored under air (usually normal air for acrylic acid and air with reduced oxygen concentration for esters). The safe handling of these products requires the use of proper protective equipment such as rubber gloves and vapor-proof goggles and masks.

Acrylic Acid. Acrylic acid normally contains 50 to 500 ppm of an inhibitor to prevent polymerization. Because of its relatively high corrosiveness, it should be stored in equipment made of or lined with glass, polyethylene, polypropylene, or stainless steel. In addition, it should be kept at 15 to 30 °C and away from direct sunlight. Freezing should be avoided because it tends to localize the inhibitor. If acrylic acid should freeze, however, it should be melted by using a warm water or air bath below 30 °C. Agitation of the acrylic acid during the melting is recommended to avoid any localized heating. Acrylic acid often is used as an 80 % aqueous solution which has a freezing point of –3 to –5 °C.

Acrylic Esters. In general, a lower level of inhibitor is required for acrylic esters than for the acid, although the range is still 50 to 500 ppm. The esters are less corrosive than the acid and thus can be stored in equipment made of or lined with carbon steel or phenolic resin, in addition to glass, polyethylene, and polypropylene. Grades of acrylates containing little or no inhibitor are available. These products should be carefully stored at temperatures of 0 to 10 °C. Methyl and ethyl acrylates have very low flash points and form explosive gas mixtures in air, even at room temperature. Thus, even though oxygen is an effective inhibitor, the oxygen concentration in large storage tanks is usually kept at 6 to 8 vol % to prevent the formation of a flammable mixture.

1.6. Uses

Acrylic Acid. The primary use of acrylic acid is as an intermediate in the production of acrylates. Polymers of the acid and its sodium salts are used increasingly in flocculants and dispersants with the polymeric sodium salts having more industrial importance (⟶ Polyacrylamides and Poly(Acrylic Acids)).

Acrylic Esters. Acrylic esters are used exclusively for the production of polymers (⟶ Polyacrylates). The polymers are used mainly for coatings, paints, adhesives, and binders for leather, paper, and textiles.

About 80 % of the methyl ester produced is used as a copolymer component of acrylic fibers (⟶). The ethyl ester is used for both solvent- and water-based paints, and in textiles as a binder in nonwoven fabrics and flocking. It generally is used in areas where more rigidity is required than can be obtained with the butyl ester. The butyl ester is growing in use, mainly in water-based paints and adhesives. The 2-ethylhexyl ester is used for almost the same purposes as the butyl ester, with a large demand for it in stick-on labels and in the caulking of building materials.

1.7. Some Special Acrylates

Esters with Polyhydric Alcohols. Representative *multifunctional acrylates* are trimethylolpropane triacrylate, pentaerythritol tri- or tetraacrylate, 1,4-butanediol diacrylate, 1,6-hexanediol diacrylate, and poly(ethylene glycol) diacrylate (n = 2 – 14). They usually are produced by direct esterification of acrylic acid with the corresponding polyhydric alcohol in the presence of an entrainer and an acid catalyst, such as sulfuric acid or *p*-toluenesulfonic acid. Because these esters have high boiling points, they cannot be purified by ordinary distillation. Instead, the reaction mixture is neutralized, the entrainer removed, and the product washed with water [3], [72].

The esters are used as cross-linking agents and modifiers in rubber and synthetic resins, in adhesives, and as active diluents in photosensitive resins. They are also applied in the coating and ink industries because they can be cured with ultraviolet light [73], [74] or electron-beam radiation [75]. Proper protection is required when handling these eye and skin irritants.

2-Hydroxyalkyl Acrylates. Two industrially important multifunctional esters are 2-hydroxyethyl acrylate and 2-hydroxypropyl acrylate. These are produced by liquid-phase esterification of acrylic acid with ethylene oxide or propylene oxide in the presence of a Lewis acid catalyst, such as a chromium [76] or ruthenium [77] compound, or the iron salt of an organic acid [78]. Because this reaction readily produces di(alkylene glycol) monoacrylates and alkylene glycol diacrylates as byproducts, a highly efficient catalyst is required. Although vapor-phase catalytic synthesis using magnesium oxide has been proposed [79], the liquid-phase esterification is preferred. These esters are used especially as cross-linking agents in heat-cured paints, adhesives, textile preparations, etc. They are toxic and lacrimatory, cause blistering of the skin, and may give rise to long-term sensitivity. Inhalation of the vapor causes nose, eye, and throat irritation.

Other Derivatives. Halogenated derivatives such as 2-chloroacrylic acid [80], [81], 2,3-dibromopropyl acrylate [82], tetrafluoropropyl acrylate, and octafluoropentyl acrylate, have potential uses as fine chemicals. Dialkylaminoethyl acrylates are produced by transesterifying methyl acrylate with the corresponding amino alcohol.

1.8. Economic Aspects

Western world production capacities of acrylic acid and its esters as of late 1982 were 839500 t/a and 1277500 t/a, respectively (Table 7). Plants based on other processes, such as the ketene and cyanohydrin methods (see Section Production), were shut down during the past decade because of inefficiency.

Table 7. Estimated regional production capacities for acrylic acid and its esters in the western world (t/a)

	Acid		Esters	
	1982	1994*	1982	1993*
United States	430 000	690 000	700 500	587 000
Western Europe	282 000	640 000	380 000	312 000
Japan	117 500	420 000	182 000	206 000
Others	10 000	–	15 000	–
Total	839 500	ca. 2 000 000	1 277 500	ca. 1 200 00

* [118]

In fact it appears that the propene oxidation route will continue to be the most economical process for quite some time. Announced additional capacities of 280000 and 380000 t/a for the acid and esters, respectively, will all be based on propene oxidation.

Consumption of acrylic acid for uses other than as an intermediate in ester production ranges from 5 to 9 % of the total, although demand for and consumption of both acid and esters varies from region to region. Table 8 gives estimated end-use percentages in three regions. Surface coatings provide the largest market for the esters in all three regions.

Table 8. Estimated distribution of end uses of acrylic esters (% of total)

	United States	Western Europe	Japan
Surface coatings	42	35	34
Textiles	23	18	16
Acrylic fibers	6	7	14
Adhesives	5	15	20
Others	24	25	16

1.9. Toxicology and Occupational Health

Acrylic Acid. Acrylic acid is moderately toxic and very corrosive [84]. Ingestion may cause severe gastrointestinal burns. The vapor is an irritant to the eyes and respiratory tract and skin contact may cause burns. Physiological response data are:

LD_{50} 340 mg/kg (rat, oral)
LC_{50} 3600 mg/m^3 (rat, inhalation, 5 L, 4 h)
LD_{50} 280 mg/kg (rabbit, skin)

The TLV on a time weighted average (TWA) is 10 ppm or 30 mg/m^3.

Acrylic Esters. Acrylic esters are of moderately acute toxicity, which decreases with an increase in the number of carbon atoms in the alkyl group (Table 9). Liquid methyl and ethyl acrylates severely irritate the skin and mucous membranes and are corrosive to the eyes, whereas the butyl and 2-ethylhexyl acrylates have less severe effects. Methyl and ethyl acrylate vapors are very lacrimatory, extremely irritating to the respiratory tract, and are corrosive to the eyes, causing corneal injury. The lacrimatory effect of the butyl and 2-ethylhexyl esters is weak, but their vapors may cause dizziness, headache, nausea, and vomiting.

Table 9. Physiological response data and exposure levels of some acrylates

	Methyl acrylate	Ethyl acrylate	Butyl acrylate	2-Ethylhexyl acrylate
LD_{50} (rat, oral), mg/kg	300	1020	3730	5660
LCLo (rat, inhalation, 4 h),	3500	4000	5500	

mg/m^3				
LD_{50} (rabbit, dermal), mg/kg	1243	1950	2000	8480
TLV (TWA)	10 ppm, 35 mg/m^3	5 ppm, ≈ 20 mg/m^3	10 ppm, 55 mg/m^3	
MAK	10 ppm, 35 mg/m^3	25 ppm, 100 mg/m^3		

Methyl and ethyl acrylates can be absorbed through the skin in toxic amounts, and overexposure to the vapor can result in fatal pulmonary edema. However, their noticeable odors and irritating effects reduce the likelihood of significant exposure.

[Top of Page]

2. Cyanoacrylates

Klaus Marten

The monofunctional 2-cyanoacrylates, CH_2 = C(CN)-COOR, have been known for many years. Technical developments, based on the original patents of 1949 [85], led in 1954 to the first viable production process [86], [87]. Since then these compounds have achieved a considerable growth rate on the market. The 2-cyanoacrylates are utilized almost exclusively as adhesives. The methyl, ethyl, butyl, allyl, and methoxyethyl esters are available with different setting characteristics and rheological properties, depending on the requirements of the application.

Physical Properties. Pure 2-cyanoacrylates are clear, colorless liquids at room temperature and have a characteristic odor. Their physical properties are listed in Table 10.

Table 10. Physical properties of industrially important 2-cyanoacrylates

Property	Methyl	Ethyl	Butyl	Allyl	Methoxyethyl
CAS registry number	[137-05-3]	[7085-85-0]	[6606-65-1]	[7324-02-9]	[27816-23-5]
Molecular formula	$C_5H_5NO_2$	$C_6H_7NO_2$	$C_8H_{11}NO_2$	$C_7H_7NO_2$	$C_7H_9NO_3$
M_r	111.10	125.13	153.18	137.14	155.15
bp, °C/*p*, kPa	48 – 49/0.33 – 0.36	54 – 56/0.34 – 0.40	53 – 56/0.27 – 0.33	115/2.53	80 – 82/0.13
Viscosity, mPa s	2.20	1.86	2.08	6.4	–
Heat of polymerization in isobutyronitrile at 25 °C, kJ/mol	57.7	58.2	67.8	–	–
Refractive index, n_D^{25}	1.4406	1.4349	1.4291	1.4426 (20 °C)	–
Density at 20 °C, g/cm^3	1.1044	1.0501	1.0009	1.0578	
Vapor pressure at 25 °C, kPa	<0.27	<0.27	<0.27	–	–

Chemical Properties. The vinyl structure of the 2-cyanoacrylates makes them liable to spontaneous polymerization. This is desirable for their use as reactive adhesives, but causes significant difficulties in synthesizing and especially in purifying them.

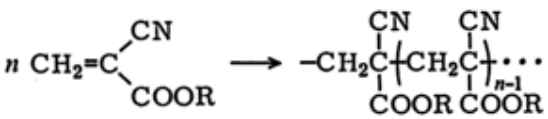

Because the chain propagation reaction can be initiated by either an ionic or a radical mechanism, the rate of polymerization depends on temperature, humidity, light, and the presence of polymerization accelerators, such as peroxides and bases. The 2-cyanoacrylates are, however, adequately stable if stored under cool conditions in the presence of suitable inhibitors. In addition to polymerization, they undergo the other reactions typical of vinyl compounds (e.g., addition).

Production. Many different processes can be used to manufacture 2-cyanoacrylates. The important step in most of the published syntheses is the classical Knoevenagel reaction using formaldehyde and a cyanoacetic ester. In this first step oligomeric cyanoacrylates, water, and other byproducts are formed. The reaction can be catalyzed by bases (e.g., amines) [86], [87], alkali metal tetraborates, metal carbonyls [89], or phase-transfer catalysts [90]. The raw condensation product from

the first reaction is depolymerized thermally (either continuously or discontinuously) [91] and purified by distillation, chromatography on alumina or silica gel [92], crystallization in polyethylene columns [93], or treatment with zinc chloride [94].

Alternatively, 2-cyanoacrylates can be manufactured by the ethoxycarbonylation of cyanoacetylene in the presence of nickel carbonyl [95].

Uses. 2-Cyanoacrylate-based formulations have been a valuable component of modern adhesive technology for many years. Their ability to join the most dissimilar materials quickly, firmly, and durably has insured a wide field of applications. They are used in the electrical and electronics industries as well as many areas of mechanical engineering, such as automobile, ship, and aircraft construction. In addition to these purely technical applications, the 2-cyanoacrylates are used in medicine to close wounds.

Special properties to meet specific requirements are obtained by the use of additives or a manufacturer's proprietary process. Such parameters as hardening time [96-98], viscosity [99], [100], relative strength [101], resistance to hydrolysis [102], [103], heat resistance [104], [105], flexibility of the mature adhesive [106], [107], and shelf life [108-110] commonly are varied (⟶ Adhesives).

Toxicology and Occupational Health. Both the MAK and the TLV exposure limits for 2-cyanoacrylates have been set at 2 ppm. Adequate ventilation of the work place is therefore necessary if these substances are in continual use. Further information can be obtained from the relevant documentation provided by suppliers. Although many years of practical use have not yet led to any significant damage to health, the toxicology of 2-cyanoacrylates requires further investigation [111], [112]

[Top of Page]

3. Acrylamide

Helmut Greim

Properties. Acrylamide [79-06-1], 2-propenamide, $CH_2 = CHCONH_2$, M_r 71.08, *mp* 84.5 °C, usually forms white crystalline platelets. It is very toxic.

Acrylamide is most conveniently handled as an aqueous solution because the dry powder forms a fine, toxic dust. It is soluble in water, alcohols, and acetone but insoluble in benzene and heptane.

Acrylamide and polyacrylamide can be converted into various products by the usual addition reactions of the double bond and especially by reactions of the amide group:

$$\begin{aligned}
RCONH_2 + HCHO &\longrightarrow RCONHCH_2OH \\
+ 2\,HCHO &\longrightarrow RCON(CH_2OH)_2 \\
2\,RCONH_2 + HCHO &\longrightarrow (RCONH)_2CH_2 \\
RCONH_2 + HCHO + NaHSO_3 &\longrightarrow \\
& \qquad RCONH-CH_2SO_3Na + H_2O \\
+ HCHO + R'R''NH &\longrightarrow \\
& \qquad RCONH-CH_2NHR'R'' + H_2O \\
+ NaOCl + 2\,NaOH &\longrightarrow \\
& \qquad RNH_2 + Na_2CO_3 + NaCl + H_2O \\
+ H_2N-NR'R'' &\longrightarrow RCONH-NR'R'' + NH_3
\end{aligned}$$

N-Alkyl derivatives of acrylamide are prepared by the reaction of acryloyl chloride with the corresponding amine, by the dehydrochlorination of 2-chloropropionamide, or by the amination of acrylic acid or esters [113].

Production. Acrylamide was first produced on a large scale, by American Cyanamid, in 1954. It is currently produced from acrylonitrile either by homogeneous sulfuric acid hydration or by heterogeneous catalytic hydration.

Sulfuric Acid Process. One mole of Acrylonitrile (⟶ Acrylonitrile) is added to a solution of 1 mol sulfuric acid and 1 mol of water at 60 °C. The mixture is slowly heated to 80 °C, kept for an hour at this temperature, and then cooled to 40 °C. Ammonium sulfate precipitates on neutralization with ammonia and is filtered out. The acrylamide, which crystallizes on cooling the mother liquor to below 10 °C, is purified by recrystallization from benzene.

Catalytic Hydration Process. Figure 6 shows a flow diagram of this process. A 50 wt % solution of acrylonitrile in water is slurried with a catalyst such as Raney copper and kept at 120 °C for 2.5 h. Conversion of acrylonitrile is greater than 50 % and selectivity to the amide is nearly 100 %. This process conveniently makes an aqueous solution of 30 to 50 wt % acrylamide because of the large amount of water required in the reaction.

Figure 6. Catalytic hydration of acrylonitrile

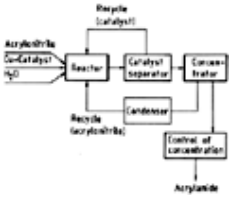

Biological Process. A patented biocatalyst enables the hydration of acrylonitrile at room temperature and pressure in a process developed by Nitto Chemical Industry Co. (Tokyo). The catalyst is an enzyme derived from one of several bacterial genera [114].

Uses. Acrylamide is made mainly into water-soluble polymers and copolymers used in flocculants, papermaking aids, thickening agents, surface coatings, and enhanced oil recovery [113], [115] (⟶ Polyacrylamides and Poly(Acrylic Acids)).

Toxicology and Occupational Health. The health consequences of human exposure to acrylamide have been repeatedly evaluated [118-122]. Most of the information given in the following is derived from these reports in addition to more recent scientific publications.

Effects Assessment. In humans and laboratory animals, single or repeated exposures to acrylamide caused local irritations on contact with the skin and neurological symptoms. The peripheral neuropathy is a result of long-term accumulation and appears after a latent period. Its onset parallels the accumulation of acrylamide bound to proteins in the nervous system and to hemoglobin. In a two year's study in rats exposed to acrylamide via drinking water the lowest observed effect level (LOEL) for peripheral neuropathy was 0.5 mg/kg.

Acrylamide is readily absorbed during oral, inhalation and dermal exposure. In rats the major route of metabolism is direct conjugation with glutathione and subsequent urinary excretion of the conjugate. The compound is also metabolized by cytochrome P-450 2E1 dependent monooxygenase activity forming the epoxide glycidamide [123].

Both, acrylamide and glycidamide interact with proteins and form hemoglobin adducts. Glycidamide interacts with DNA and is considered the genotoxic and carcinogenic intermediate. Mice form more glycidamide than rats and there is an indication that humans form less glycidamide than rats. These species differences in metabolism explain the lower potency to induce tumors in rats than in mice and suggest that humans are less sensitive than rats.

In rats acrylamide induced increased incidences of thyroid adenomas, testicular mesotheliomas, adrenal phaeochromocytomas, fibroadenomas in the mammary gland, and focal hyperplasia in the oral cavity. An increased incidence of astrocytomas was also detected. The lowest effective dose in these studies was 1–2 mg/kg bw/da. (mg per kilogram body weight and day).

In mice skin painting experiments acrylamide initiates the formation of skin tumors. Epidemiological studies have not revealed tumors in humans. Since the carcinogenic potency of acrylamide in animals is weak the epidemiological studies may not be sensitive enough to show an increased tumor incidence in man.

Acrylamide is considered to be genotoxic although contradictory results have been obtained in the various assays. In bacterial systems glycidamide but not acrylamide induced mutations. In mammalian cells in vitro chromosomal aberrations, micronuclei, sister chromatid exchanges, polyploidy and other cytogenetic effects have been observed. In vivo the compound was positive in the bone marrow chromosome aberration test and in the micronucleus assay.

Overall, the long term animal studies, its genotoxicity, and the formation of the genotoxic metabolite glycidamide in humans indicate that acrylamide has the potential to cause cancer in humans. Accordingly, IARC, the EU, and several national committees classified acrylamide as probably carcinogenic to humans.

In reproductive toxicity studies with rats, the no observed adverse effect level (NOAEL) for reduced fertility was 5 mg/kg bw/da and for premature embryonic death 2 mg/kg bw/da.

Exposure. The concern in 2002 is the relatively high exposure of consumers from fried, oven-baked, and deep-fried potato and cereal foods. The Scientific Committee on Food of the European Commission [124] estimates an average daily human intake of about 1 µg/kg considering that differently contaminated food, some of which containing up to 3000 µg acrylamide per kilogram, is consumed. Contamination of certain food had been first detected by the Swedish National Food Agency. They found that especially potato products such as chips and crisps, but also basic staple foods such as bread and crisp bread contained varying high levels of acrylamide. Acrylamide was not detected in boiled foods. These results have been confirmed by analyses of foods in other European countries.

The available information suggests that acrylamide is primarily formed in carbohydrate-rich foods at high temperatures. It is presently being evaluated whether changes in food processing and cooking may decrease formation of acrylamide.

At the workplace exposure occurs during acrylamide and polyacrylamide production, during grouting and during preparation of polyacrylamide gels. Low levels have been detected in the drinking water from the use of polyacrylamide flocculants in water treatment and from the use of polyacrylamides in cosmetics and toiletries [122].

Risk Characterization. The available information indicates that there are concerns regarding neurotoxicity (NOEL 0.5 mg/kg bw/da), impact on the reproduction system (NOEL 2 mg/kg bw/da) and carcinogenicity. Based on the average daily intake by

humans via food of 1 µg/kg bw/da the margin of safety (MOS) between this exposure and the NOEL of neurotoxicity is 500.

Due to its genotoxic potential no dose threshold for the carcinogenic effect of acrylamide can be established. Using a linear extrapolation from the carcinogenicity studies in animals the lifetime cancer risk related to a daily intake for 70 years of 1 µg/kg acrylamide is about 1×10^{-3} (one case of cancer per 1000 individuals).

The TLV-TWA is 0.03 mg/m^3 with a skin notation without a STEL, A3: confirmed animal carcinogen. The TRK (Technische Richtkonzentration) is 0.06 mg/m^3. Classification of acrylamide by IARC: probable carcinogen for humans (2b), MAK: Considered to be carcinogenic to man (Cat. 2).

[Top of Page]

References

1. E. C. Leonard: *Vinyl and Diene Monomer*, part 1, Wiley-Interscience, New York 1970, p. 148.
2. E. Ohmori: *Acrylic Acid and its Polymers*, **vol. I and II**, Shokodo, Tokyo 1973 and 1975.
3. E. H. Riddle: *Monomeric Acrylic Esters*, Reinhold, New York 1954.
4. H. Rauch-Puntigam, T. Völker: *Acryl- und Methacrylverbindungen*, Springer Verlag, Berlin 1967.
5. M. Sittig: *Vinyl Monomers and Polymers*, Noyes Dev. Corp., Park Ridge, N.J., 1966.
6. L. S. Luskin, *High Polym.* **24** (1970) 105.
7. Sh. Suzuki, *Yuki Gosei Kagaku Kyokaishi* **28** (1970) no. 12, 1272. Links
8. Rohm & Haas, US 2647923, 1953 (C. Burton). *Org. Synth. Coll.* **3** (1955), 576.
9. R. Madhav, C. A. Snyder, P. L. Southwick, *J. Heterocycl. Chem.* **17** (1980) no. 6, 1231. V. P. Nekoroshkov, G. L. Kamalov, N. M. Vladyka, *Khim. Promst. Ser. Reakt. Osobo Chist. Veshchestva* 1980, no. 6, 17; *Chem. Abstr.* **95**, 42519 t.
10. K. Thiele, K. Posselt, H. Offermanns, *Arzneim. Forsch.* **28** (1978) no. 11, 2047.
11. Union Carbide, US 4012445, 1972, US 4011223, 1972 (D. C. Priest, M. R. Sander, D. J. Trecker).
12. P. Brun, A. Tenaglia, B. Waegell, *J. Organomet. Chem.* **194** (1980) C 39. Links
13. R. M. Acheson, P. W. Poulter, *J. Chem. Soc.* 1960, no. 2, 2138.
14. A. Etienne, G. Louchambon, P. Givaudean, *C. R. Acad. Sci. Ser. C* **289** (1979) no. 9, 263.
15. A. Le Berre, C. Porte, *Bull. Soc. Chim. Fr.* 1978, 602. B. B. Snyder, R. S. E. Coun, S. Sealfon, *J. Org. Chem.* **44** (1979) 218.
16. Tokyo Fine Chemical, JP 31137, 1980, US 3846489, 1974 (H. Uesugi, T. Takeda).
17. C. E. Rehberg, *J. Am. Chem. Soc.* **68** (1946) 544, Links **69** (1947) 2966, **72** (1950) 2205.
18. BASF, DE 2759162, 1977 (H. Distler, K. Schneider, R. Widder, G. Paulus, H. Schoeppl).
19. Witco Chemical Corp., US 4052440, 1973 (S. Gladstone, S. R. Rao, C. J. Rosshirt).
20. C. D. Hurd, L. L. Gershbein, *J. Am. Chem. Soc.* **69** (1947) 2328. Links
21. Rohm & Haas, US 3769315, 1973 (R. L. Keener, H. Raterink).
22. Soda Sangyo, JP-Kokai 34720, 1978 (Y. Nakatomi, S. Izukawa, T. Ohtsuka); *Chem. Abstr.* **89**, 108175 n.
23. L. Novak, G. Baan, J. Marosfalvi, C. Szantay, *Chem. Ber.* **113** (1980) 2939 – 2949.
24. Inst. of Phys. and Chem. Res., JP-Kokai 57784, 1980, DE 3007592, (K. Koh, H. Yamazaki).
25. Inst. of Phys. and Chem. Res., JP-Kokai 115829, 1980, DE 3007592, (K. Koh, H. Yamazaki). Universal Oil Products, US 3935234, 1976 (P. H. Reichenbacher, T. M. Forsythe, A. K. Sparks, T. Symon).
26. N. D. Lee, D. Faulkner, H. C. Highet, A. R. Philpotts, W. Thain, *J. Appl. Chem.* **3** (1953) 481.
27. C. J. Albisetti, N. G. Fisher, M. J. Hogsed, R. M. Joyce, *J. Am. Chem. Soc.* **78** (1956) 2637, 2640. Links
28. Smithkline Corp., US 4005208, 1975 (P. E. Bender, B. Loev).
29. V. B. Reddy, *Chem. Ind. (London)* 1976, 414.
30. P. De Strong, N. E. Lowmaster, *Synth. Commun.* **13** (1983) no. 7, 537.
31. J. G. Martin, R. Hill, *Chem. Rev.* **61** (1961) 537. Links
32. BASF, BE 585725, 1978.
33. F. Falbe (ed.): *New Syntheses with Carbon Monoxide*, Springer Verlag, Berlin 1980, p. 465.
34. M. Hidai, A. Misono, *Aspects Homogeneous Catal.* **2** (1974) 159.
35. Mitsubishi Rayon, JP-Kokai 30958, 1981 (Y. Nakai, A. Yanagase).
36. Sumitomo Chemical, JP-Kokai 72937, 1982 (S. Shimizu, K. Nakai); *Chem. Abstr.* **97**, 128256 h. Rohm & Haas, JP-Kokai 9219, 1979 (R. J. Piccolini). Sumitomo Chemical, JP-Kokai 65815, 1978 (T. Matsuda, T. Motohashi).
37. Nippon Oil Seal Ind., JP-Kokai 39533, 1980 (H. Idemitsu, M. Kishida, T. Nishikubo, S. Ukai); *Chem. Abstr.* **78**, 158964 n.
38. F. T. Maler, W. Bayer, *Encycl. Chem Process. Des.* **1** (1976) 401.
39. D. J. Hadley, E. M. Evans: *Propylene and its Industrial Derivatives*, J. Wiley & Sons, New York 1973, p. 416 – 497.
40. Celanese Corp., US 2820058, 1958 (O. V. Luke, M. O. Robeson, W. E. Taylor). Celanese Corp., US 3069433, 1962 (K. A. Dunn). B. F. Goodrich Co., US 2 356 459, 1944, US 2361036, 1944 (E. F. King), US 3002017, 1961 (N. Wearsch, A. J. De Paola).
41. B. F. Goodrich Co., US 3392196, 1964. Nippon Shokubai, US 3475488, 1969 (N. Kurata, T. Ohara, K. Oda).

42. S. Sakuyama, T. Ohara, N. Shimizu, K. Kubota, *Chemtech.* **3** (1973) no. 6, 350.
43. D. J. Hucknall: *Selective Oxidations of Hydrocarbons*, Academic Press, London 1974, p. 52.
44. Distillers, JP 20838, 1963, GB 915799, 1963 (D. J. Hadley, R. H. Jenkins).
45. Distillers, GB 903034, 1962 (J. R. Bethell, D. J. Hadley, E. J. Gasson, R. F. Neale).
46. Rikagaku Kenkyusho, JP 26287, 1969, US 3567772, 1971 (M. Yanagita, M. Kitahara).
47. Nippon Kayaku, BE 698273, 1967 (S. Takenaka).
48. Celanese, GB 1267189, 1972, US 3644509, 1972 (G. C. Allen).
49. BASF, FR 2032915, 1970, US 3845120, 1974 (R. Krabetz, H. Engelback).
50. Degussa, JP-Kokai 8360, 1972, DE-OS 2055155, 1972 (J. Hensel).
51. Nippon Shokubai, JP 11371, 1974, DE 2152037 1972 (M. Wada, I. Yanagisawa, M. Ninomiya, T. Ohara).
52. Sumitomo Chemical, JP-Kokai 31923, 1972 (T. Shiraishi, S. Kishiwada, S. Shimizu, S. Honmaru, Y. Nagaoka, K. Jinpo); *Chem. Abstr.* **78**, 30464 g.
53. Nippon Shokubai, JP-Kokai 117419, 1974, DE 2413206 1974 (M. Wada, I. Yanagisawa, N. Ninomiya, T. Ohara).
54. Mitsubishi Petrochem., JP 169, 1974, DE 2164905 1972 (Y. Kadowaki, T. Koshikawa).
55. Sohio, JP-Kokai 83280, 1975, DE 2448804, 1975 (S. R. Dolhyi, E. C. Milberger).
56. JP 28889, 1978, DE 2456100, 1975 (N. Ferlazzo).
57. Knapsack, US 3717675, 1966 (K. Sennewald, A. Hauser, K. Gehrmann, H.-K. Steil, W. Lork). Celanese, US 4156633, 1979, JP-Kokai 59813, 1976 (T. Horlenko *et al.*). Nippon Shokubai, JP-Kokai 95217, 1975 (N. Shimizu); *Chem. Abstr.* **83**, 206785 g.
58. BASF, DE 2241714, 1974, JP 21010, 1981 (D. Gerd *et al.*). Rhône-Poulenc, JP 11896, 1982, US 4219389, 1977, FR 7718136, 1977 (G. Biola, Y. Komorn, G. Schneider).
59. Knapsack, US 3553261, 1971 (K. Sennewald *et al.*). Asahi Kasei, JP 18967, 1971, DE 1 950750 1970 (M. Honda *et al.*).
60. Rhône-Poulenc, JP 11896, 1982, US 4219389, 1980 (G. Biola *et al.*). Knapsack, JP 32843, 1974, DE 1965014, 1971 (K. Sennewald *et al.*).
61. Sumitomo Chemical, GB 1427 223, 1972, US 3968153, 1976 (T. Ohrui *et al.*).
62. BASF, US 3962074, 1976, JP-Kokai 18 412, 1975, DE 2323328, 1974 (W. K. Schropp *et al.*).
63. Rohm & Haas, DE-OS 1168609, 1967, FR 1573704, 1967.
64. Celanese, GB 1182809, 1966, FR 1544368, 1966 (E. G. Schiedling, J. R. Muratorio).
65. Nippon Geon, DE-OS 2035228, 1969 (R. Sato *et al.*).
66. Toa Gosei Chemical, US 3639461, 1972 (H. Ito *et al.*).
67. Asahi Kasei, JP 6131, 1978 (N. Kominami *et al.*).
68. Sumitomo Chemical, US 3875212, 1975 (T. Ohrui *et al.*).
69. Nippon Shokubai, JP 37 404, 1972 (K. Kubota *et al.*), *Chem. Abstr.* **78**, 16746 u.
70. Mitsubishi Rayon, JP 35 883, 1972 (T. Kita *et al.*); *Chem. Abstr.* **78**, 85047 d.
71. L. S. Luskin in F. D. Snell, C. L. Hilton (eds.): *Encyclopedia of Industrial Chemical Analysis*, **vol. 4**, J. Wiley & Sons, New York 1967, p. 181.
72. Sumitomo Chemical, JP-Kokai 70215, 1979 (T. Matsuda, H. Takamatsu); *Chem. Abstr.* **91**, 192843 a.
73. C. B. Rybny, C. A. Defazio, I. K. Shahide, J. C. Trebellas, J. A. Vona, *J. Paint Technol.* **46** (1974) no. 569, 60.
74. S. E. Young, *Prog. Org. Coat.* **4** (1976) 225 – 249. Links
75. C. Bluestein, *Adhes. Age* **25** (1982) no. 12, 19.
76. Nippon Shokubai, JP 300, 1982 (S. Yoshida, M. Daigo, S. Matsumoto, N. Shimizu). Dow, BR Pedido PI 7800250, 1979 (A. E. Gurgiolo).
77. Union Carbide, US 4223160, 1980 (L. G. Hess).
78. Nippon Shokubai, JP-Kokai 161349, 1981 (N. Shimizu, M. Daigo, S. Matsumoto).
79. Nippon Kayaku, JP 61259, 1982 (H. Hayami); *Chem. Abstr.* **87**, 53823 r.
80. Knapsack, JP-Kokai 56917, 1974 (E. Auer, W. Vogt, K. Gehrmann).
81. E. M. Movsumsade, M. G. Mamerov, I. A. Shikhuev, V. B. Papin, *Khim. Promst. (Moscow)* **4** (1980) 204.
82. Sankyo, EP 57058, 1982. (Y. Morisawa, K. Konishi, M. Kataoka).
83. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie*, Wiley-VCH Verlag GmbH, Weinheim, Germany 1998.
84. *Registry of Toxic Effects of Chemical Substances*, U.S. Department of Health, Education, and Welfare, Public Health Service Center for Disease Control, NIOSH, Cincinnati, Ohio 1980. *Patty's Industrial Hygiene and Toxicology*, 3rd ed., **vol 2 A**, Wiley-Interscience, New York 1981, p. 2294.
85. B. F. Goodrich, US 2467927, 1949, US 2467926, 1949 (A. E. Ardis).
86. Eastman Kodak, US 2721858, 1955 (F. B. Joyner, G. F. Hawkins).
87. Eastman Kodak, US 2794788, 1957 (H. W. Coover, Jr. N. Shearer, Jr.).
88. H. S. Holappa *et al.*, *Adhes. Age* **22** (1979) 43.
89. V. Winkovic, DE-OS 2738285, 1979 (Viktor Winkovic).
90. Matsumoto, DE-OS 2944085, 1979 (S. Iwakichi *et al.*).
91. Schering, DE-AS 2027502, 1970 (W. Imöhl, P. Borner).
92. Toa Gosei, JP 7897036, 1978 (K. Kimura, K. Sakabe, A. Motegi, H. Tatemichi); *Chem. Abstr.* **89**, 216475 a.

93. Kores, FR 2143023, 1973 (L. Kamlander).
94. Toa Gosei, JP 78128689, 1978 (K. Kimura, E. Osowa, H. Tatemichi); *Chem. Abstr.* **90**, 104657 a.
95. Matsumoto, JP 7435608, 1974 (S. Iwakichi, E. Kiyoshi); *Chem. Abstr.* **82**, 125805 r.
96. Eastman Kodak, DE-OS 2223026, 1972, US 3 728 375, 1973 (H. W. Coover, J. M. McIntire).
97. Toa Gosei, DE-OS 2 816 836, 1978 (M. Akira *et al.*).
98. Toaka, JP 79123147, 1979 (K. Ohashi, S. Kusayama).
99. Eastman Kodak, US 3654239, 1972 (J. M. McIntire, H. Wicker, Jr.).
100. A. Kawabata, JP-Kokai 79043247, 1979 (A. Kawabata); *Chem. Abstr.* **91**, 40425 c.
101. Eastman Kodak, FR 2010590, 1969.
102. Eastman Kodak, US 33 54 128, 1967 (T. H. Wicker).
103. Toa Gosei, JP 7 776 344, 1977 (K. Ito, A. Yamada, K. Kimura).
104. Schering, DE-AS 2349799, 1973 (B. Brinkman, W. Imöhl).
105. Loctite, DE-AS 2201547, 1971.
106. Loctite, DE-OS 2049 744, 1970 (D. J. O'Sullivan, B. J. Bolger).
107. Toa Gosei, JP 7780336, 1977 (K. Kimura, A. Yamada); *Chem. Abstr.* **87**, 185547 w.
108. Loctite, DE-AS 2042 334, 1970 (D. J. O'Sullivan, B. J. Bolger).
109. Schering, DE-OS 2128985, 1971 (W. Imöhl, P. Borner).
110. Toa Gosei, JP 7313 334, 1969 (I. Kenji, K. Sayoshi, K. Kishichiro, Z. Hiroshi); *Chem. Abstr.80*, 74347 s.
111. R. A. W. Lehmann, G. J. Hayes, F. Leonhard: “Toxicity of alkyl 2-cyanoacrylates, I: Peripheral Nerve,” *Arch. Surg. (Chicago)* **93** (1966) 441 – 446, “II: Bacterial growth,” *Arch. Surg. (Chicago)* **93** (1966) 447 – 450.
112. K. C. Pani, G. Gladieux, R. K. Kulkarni, G. Brandes, F. Leonhard: “The degradation of n-butyl alpha cyanoacrylate tissue adhesive,” *Surgery (St. Louis)* **63** (1968) 481 – 489.
113. Röhm, GB 2 102 426, JP-Kokai 26849, 1983, DE 3130508, 1982 (S. Besecke, G. Schroeder).
114. *Chem. Eng.* **91** (1984) no. 8, 10. *Eur. Chem. News* **42** (1984) no. 1134, 19.
115. Jefferson Chemical Co., US 3 878 247, 1975 (P. H. Moss, R. M. Gipson). Lubrizol Corp., US 3917594, 1975 (D. J. Hoke). Texaco, US 4031138, 1976 (E. C. Nieh, P. H. Moss). BASF, DE-OS 2730094, 1977 (H. Pohlemann, H. Naarmann). Chem. Fabrik Stockhausen, EP 13416, 1979 (B. Gossens, E. Barthell, K. Dahmen, E. Küster).
116. K. Hashimoto, *Sangyo Igaku* **22** (1980) 233. Links
117. P. S. Spencer, H. H. Schaumburg, *EHP Environ. Health Perspect.* **11** (1975) 129 – 133. D. D. McCollister, F. Oyen, V. K. Rowe, *Toxicol. Appl. Pharmacol.* **6** (1964) 172 – 181. Links
118. D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe*, Toxikologisch-Arbeitsmedizinische Begründung von MAK-Werten, Verlag Chemie, Weinheim, Germany 1985, Part 8.
119. IPCS: Environmental Health Criteria 49: Acrylamide, World Health Organization, Geneva 1985.
120. BUA (Beratergremium umweltrelevanter Altstoffe der Gesellschaft Deutscher Chemiker) Acrylamide: *BUA Stoffbericht 103*, S. Hirzel Wissenschaftliche Verlagsgesellschaft, Stuttgart 1993.
121. IARC (International Agency for Research on Cancer): Acrylamide: IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans 60, Lyon 1994, pp. 389–433.
122. EC 2000, Risk Assessment of Acrylamide. European Commission, Joint Research Center, European Chemicals Bureau, Ispra (http://ecb.jrc.it/existing-chemicals/).
123. C. J. Sumner, T. R. Fennell, T. A. Moore, B. Chanas, F. Gonzales, B. I. Ghanayem: “Role of cytochrome P450 2E1 in the metabolism of acrylamide in mice,rdquo; *Chem. Res. Toxicol.* **12** (1999) 1110–1116. Links
124. European Commission: Opinion of the Scientific Committee on Food on new findings regarding the presence of acrylamide in food (expressed on 3 July 2002) SCF/CS/CNTM/CONT/4 Final (http://europa.eu.int/comm/food/fs/sc/scf/index_en.html).

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Acrylonitrile

Patrick W. Langvardt, The Dow Chemical Co., Midland, Michigan 48640, USA

1. Introduction 2
2. Physical Properties 2
3. Chemical Properties 2
4. Production 2
5. Quality Specifications and Chemical Analysis 2
6. Storage and Transportation 2
7. Uses 2
8. Economic Aspects 2
9. Toxicology and Occupational Health . . 2
10. References 2

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_177

Table 5. Acrylonitrile supply and demand in 1981 (thousand tons)∗

Country	Year-end capacity	Production	Imports	Exports	Consumption
Mexico	74	54	20	-	72
United States	1131	906	Neg∗∗	291	598
Brazil	60	57	-	35	24
France	90	56	53	16	94
FR of Germany	370	250	37	26	262
Italy	230	119	97	-	228
United Kingdom	395	225	45	75	195
Spain	75	59	44	-	100
Japan	733	477	71	40	515
South Korea	77	52	76	-	159
Taiwan	132	103	Neg∗∗	9	93

∗∗ Reproduced with permission of the World Petrochemicals Program, SRI International [53].
∗∗ Neg = Negligible.

1. Introduction

In 1893, the French chemist, Ch. Moureau, first prepared acrylonitrile [107-13-1], acrylic acid nitrile, propene nitrile, vinyl cyanide, CH_2=CH-C≡N, M_r 53.06, by dehydrating either acrylamide or ethylene cyanohydrin with phosphorus pentoxide [1]. However, no significant technical or commercial applications were discovered for acrylonitrile until the late 1930s.

Shortly before the Second World War, I. G. Farbenindustrie introduced a synthetic rubber, Buna N, based on a copolymer of butadiene and acrylonitrile. This synthetic rubber was highly resistant to swelling in gasoline, oils, and other nonpolar solvents. At about the same time research began in the United States on similar copolymers termed GR-A, NBR, or nitrile rubber. Projects con-

as end-use products of acrylonitrile in the United States and Japan. Adiponitrile is further converted to hexamethylenediamine (HMDA), used to manufacture nylon 66 (→ Hexamethylenediamine). Acrylamide is used to produce water-soluble polymers or copolymers used for paper manufacturing, waste treatment, mining applications, and enhanced oil recovery.

Nitrile rubbers, the original driving force behind acrylonitrile production, have taken a less significant place as end-use products. They find industrial applications in areas where their oil resistance and low-temperature flexibility are important, such as in the fabrication of seals (O-rings), fuel hoses, and oil well equipment.

In the future, acrylonitrile may be used increasingly for such things as high-nitrile barrier resins

Acylation and Alkylation

Michael Röper, BASF Aktiengesellschaft, Ludwigshafen, Germany

Eugen Gehrer, BASF Aktiengesellschaft, Ludwigshafen, Germany

Thomas Narbeshuber, BASF Aktiengesellschaft, Ludwigshafen, Germany

Wolfgang Siegel, BASF Aktiengesellschaft, Ludwigshafen, Germany

1. Introduction 1
1.1. Catalysts 2
1.1.1. Acidic Catalysts 2
1.1.2. Basic Catalysts 5
1.1.3. Organometallic Catalysts 5
1.2. Alkylating and Acylating Agents . . 5
1.3. Mechanism 6
1.3.1. Alkylation 6
1.3.2. Acylation 7
2. Alkylation and Acylation of Aromatic Compounds 7
2.1. Alkylation 7
2.1.1. Alkylation of Benzene 10
2.1.2. Cycloalkylation 14
2.1.3. Alkylation of Substituted Benzenes . 15
2.1.4. Alkylation of Phenols 16
2.1.5. Alkylation of Aromatic Amines 19
2.1.6. Alkylation of Heteroaromatic Compounds and Related Substrates 19
2.1.7. Miscellaneous Alkylation Reactions . 20
2.2. Acylation 22
2.2.1. Acylation of Benzene and Benzene Derivatives 27
2.2.2. Acylation of Polynuclear Aromatic Compounds 28
2.2.3. Acylation of Heteroaromatic Compounds 29
2.2.4. Acylation of Nonbenzenoid Aromatic Compounds 31
2.2.5. Miscellaneous Acylation Reactions . 31
3. Acylation and Alkylation of Aliphatic Hydrocarbons 33
4. *N*-Alkylation and *N*-Acylation . . . 34
4.1. *N*-Alkylation 34
4.1.1. *N*-Alkylation by Alcohols or Ethers . 34
4.1.2. *N*-Alkylation by Alkyl Halides 35
4.1.3. *N*-Alkylation by Olefins 35
4.1.4. *N*-Alkylation by Carbonyl Compounds (Reductive Alkylation) 35
4.2. *N*-Acylation 36
5. *O*-Alkylation and *O*-Acylation 37
5.1. Synthesis of Esters 38
5.2. Synthesis of Anhydrides 38
5.3. Synthesis of Ethers 39
6. *CH*-Alkylation and *CH*-Acylation of Nonaromatic Compounds 39
7. Synthesis of Metal Alkyls 39
8. References 41

1. Introduction

Alkylation and acylation can be defined as the introduction of an alkyl (R) or an acyl (RCO) group, respectively, into a molecule. These two types of reaction have many features in common. For instance, essentially the same catalysts can be applied, leading to related reaction intermediates. The definition given above covers a wide range of reactions. Although alkylations and acylations often are referred to as acid-catalyzed Friedel – Crafts-type reactions only, alkylation can refer also to the preparation of alkylates in refinery operations. In this article, the major emphasis is on industrial significance; reactions are classified according to substrate type.

Because alkylation and acylation reactions are diverse in nature, there is no universal method of carrying them out. The production of each compound must be considered individually, assessing chemical, engineering, and economic factors.

Industrial Importance. A variety of acylations and alkylations are carried out on a large industrial scale. Some reactions of this type are used predominantly for the synthesis of fine chemicals, such as pharmaceuticals or agricultural chemicals. Some of the processes used

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_185

to produce large-volume chemicals are summarized in Table 1, underlining their commercial value. Further examples are given in the relevant subchapters.

1.1. Catalysts

Although alkylations can be carried out thermally by free-radical pathways, all commercial alkylation and acylation processes are conducted using acidic, basic, or organometallic compounds as promoters. These promoters function by increasing either the nucleophilicity of the substrate (*basic catalysis*) or the electrophilic character of the reagent (*acidic catalysis*).

1.1.1. Acidic Catalysts

The acidic catalysts can be divided into five groups: acidic halides, metal alkyls and alkoxides, proton acids, acidic oxides and sulfides, supported acids and cation-exchange resins.

Acidic Halides. All Lewis acids, including the metal halides known as Friedel – Crafts catalysts, have an electron-deficient center capable of attracting electrons from the basic reagents. Olah has classified the catalytic activity of a variety of halide Lewis acids for the alkylation of benzene or toluene by benzyl chloride [13]. Catalysts of the highest activity give rise to intra- and intermolecular isomerization reactions as well as the desired alkylation. Therefore, depending on the alkylating agent, the substrate, and the reaction conditions, the choice of the catalysts has to be a compromise between activity and product selectivity. The acidic halide catalysts most frequently used are aluminum chloride and aluminum bromide. Other common catalysts are BF_3, BCl_3, BBr_3, $BeCl_2$, $CdCl_2$, $ZnCl_2$, $GaCl_3$, $GaBr_3$, $FeCl_3$, $SbCl_3$, $BiCl_3$, $TiCl_4$, $ZrCl_4$, $SnCl_4$, UCl_4, and $SbCl_5$. In industrial applications, however, aluminum chloride is the catalyst of choice because of its low cost and wide availability.

In their original work, Friedel and Crafts described reactions using anhydrous aluminum chloride. However, it is rather difficult to obtain aluminum chloride or other Lewis acid-type metal halides in an absolutely anhydrous state. Even if this were possible, contamination, for example, moisture during application to a reaction mixture, is extremely difficult to exclude completely. Therefore, under the usual Friedel – Crafts conditions, the presence of such impurities as water, oxygen, hydrogen halides, or organic halides is inevitable. Fortunately, these compounds are beneficial in trace amounts and in many cases they are even essential cocatalysts [14]. These cocatalysts can be proton suppliers, such as hydroxy compounds (alcohols, water) and proton acids (HCl, H_2SO_4, H_3PO_4, RCOOH), leading to very strong conjugated proton acids:

$$HX + AlX_3 \rightleftharpoons H^+AlX_4^-$$

Cation-forming compounds, such as alkyl and acyl halides or oxygen, nitrogen, sulfur, and halide donors, also are effective promoters.

From an environmental point of view, acid halides have the disadvantage to produce waste salts after work-up/neutralization of the product mixture. In an effort to minimize waste formation, reusable aluminum catalysts have been developed. The catalyst forms a liquid clathrate which is insoluble in the Friedel – Crafts reaction products, is separated, e.g., by phase separation, and is recycled to the process [15].

Ionic Liquids. A promising new development is the use of ionic liquids as catalysts and solvents for alkylations and acylations. Chloroaluminate(III) liquids, that form ionic liquids at room temperature, specifically 1-butylpyridinium chloride – aluminum(III) chloride and 1-butyl-3-methylimidazolium chloride – aluminum(III) chloride, allow separation of products from catalysts and provide access to a clean technology, where formation of large amounts of salts as the byproduct is avoided. The design principles for room-temperature ionic liquids, some of their properties, and the rationale for using these neoteric solvents, are discussed in [16] and an indication of the scope of these solvents for future industrial processes is given.

Metal Alkyls and Alkoxides. Because of their Lewis acidity, metal alkyls and alkoxides are active catalysts in Friedel – Crafts reactions. However, the Lewis acidity in these systems,

Table 1. Examples of industrial alkylation processes

Reaction type	Process description		Production, 10^6 t/a ∗
Alkylation of aromatics			
Benzene and ethylene to give ethylbenzene	a)	liquid phase, 85 – 95 °C, 1 bar, catalyst: $AlCl_3$ (BASF, Dow), zeolites (Lummus/Unocal/UOP), MCM-22 zeolite (EBMax, Mobil)	20
	b)	vapor phase, 250 – 450 °C, 10 – 80 bar, catalyst: $Al_2O_3 \cdot SiO_2$ (Koppers), $H_3PO_4 - SiO_2$ (UOP), $BF_3 - Al_2O_3$ (UOP, Alkar process), zeolites ZSM-5 (Mobil/Badger)	
Benzene and propene to give cumene	a)	liquid phase, 35 – 70 °C, 5 – 15 bar, catalyst: solid phosphoric acid (UOP), $AlCl_3$ catalyst (Monsanto/Kellog), H_2SO_4 (Hock process), HF (Hüls)	7.5
	b)	zeolite technologies (Mobil, UOP – QMax, CDTech, Dow/Kellog and Enichem)	
Benzene and propene to give *p*-diisopropylbenzene		liquid phase, 150 – 250 °C, 1 bar, catalyst: $Al_2O_3 \cdot SiO_2$ (Shell), zeolites	0.1
Benzene and $C_{10} - C_{18}$ olefins (or alkyl chlorides) to give alkylbenzenes		liquid phase, 10 – 80 °C, 1 bar, catalyst: HF, H_2SO_4, $AlCl_3$, fluorided silica – alumina, UOP-Detal	0.9
Alkylation of phenols			
Phenol and $C_6 - C_{20}$ olefins to give alkylphenols	a)	liquid phase, 20 – 100 °C, 1 – 20 bar, catalyst: BF_3 , H_2SO_4 , H_3PO_4 – support, cation-exchange resins	
	b)	vapor phase, 300 – 400 °C, 1 – 30 bar, catalyst: BF_3 – support, H_3PO_4 – support, $Al_2O_3 \cdot SiO_2$	0.3
Phenol and methanol to give *o*-cresol		liquid phase, 300 – 360 °C, 40 – 70 bar, catalyst: $Al_2O_3 \cdot SiO_2$, $ZnBr_2$ – HBr	–
Phenol and acetone to give bisphenol A		50 °C, catalyst: H_2SO_4, HCl, cation-exchange resins	1.1
N-Alkylation			
Ammonia and methanol to give methylamines		350 – 400 °C, catalyst: $Al_2O_3 \cdot SiO_2$, H_3PO_4 – support	0.6
O-Alkylation			
Methanol and isobutene to give methyl *tert*-butyl ether		$<$100 °C, catalyst: anion-exchange resins	0.12
Metal alkylation			
Na/Pb alloy and RCl to give PbR_4		70 °C, no catalyst	0.6

∗ estimated for 1992.

particularly for metal alkoxides, is weakened by hyperconjugation effects. Nevertheless, an important industrial application is the alkylation of phenols, in which aluminum phenoxide is used as the catalyst. Again, a conjugated acid is believed to be the active catalyst:

$$C_6H_5OH + Al(OC_6H_5)_3 \rightleftharpoons H^+Al(OC_6H_5)_4^-$$

Aluminum metal can form aluminum phenoxide in phenolic solution.

Rare Earth Metal Trifluoromethansulfonates (Triflates). This type of catalyst can be recycled after use and avoids the stoichiometric formation of metal salts as is the case with conventional homogeneous promoters.

Proton Acids. A variety of proton acids can be used in Friedel – Crafts alkylations, the most important examples being sulfuric acid, anhydrous hydrogen fluoride, trifluoroacetic acid, hydrogen chloride, and phosphoric acids [17]. Combinations of Lewis acids and proton acids, such as $HCl - AlCl_3$, $HCl - GaCl_3$, $HF - BF_3$, $H_2SO_4 - BF_3$, and $HF - SbF_5$, also can be applied and these act as conjugate proton superacids [18], [19]. Similarly, most Lewis acids work under industrial conditions, where traces of moisture or of other proton donors cannot be excluded.

Acidic Oxides and Derivatives. Solid catalysts are progressively substituting liquid acid catalysts in oil refining and synthesis of large

volume chemicals. This development is due to novel environmental legislation as well as economic advantages resulting from superior product qualities of solid catalysts and the possibility to work under noncorrosive conditions (facile construction, handling and disposal). Solid catalysts are generally classified into different groups, i.e., zeolites, mixed oxides, or solid superacids (e.g., sulfated oxides), clays, heteropolyacids or polyoxoanions, supported acids, and cation-exchange resins. The latter two will be discussed in page 4 and page 4

Zeolites (→ Zeolites) represent a class of acidic (or basic) microporous materials with a defined pore structure. The most widely used systems are based on silica and alumina, however, a large variety of other elements have been integrated into the zeolite framework. For a comprehensive review on zeolite types see [20]. Acidity in zeolites is generated by Brønsted and Lewis acid sites. The former are attributed to protons compensating for the negative lattice charge of an oxygen atom in vicinity to an aluminum atom, the latter are believed to result from lattice defect sites and/or extra-framework aluminum species. Zeolites are thermally stable up to 800 °C allowing high-temperature reactions and the thermal regeneration of spent catalysts by burning the carbonaceous residues in oxygen-containing gas streams. Zeolites can be synthesized in a wide variety of pore sizes. Number and strength of Brønsted acid sites as well as crystal sizes may be tuned. The versatility of these materials lead, thus, to a broad field of industrial applications, e.g., fluid catalytic cracking (FCC), alkylation, aromatization, isomerization, and oligomerization [21].

Mixed oxides represent probably the largest group of solid acid catalysts. Various mixtures of (doped) metal oxides have been investigated in acid catalyzed reactions. Excellent reviews were provided by Arata [22] and Tanabe [23–25]. As with zeolites, acidity is generated by the incorporation of heteroatoms (different valence) in a crystal-lattice structure. Some of these materials are quoted superacids due to the presence of acid sites stronger than 100 % H_2SO_4 ($H_0 < -12$) [H_0 = Hammelt acidity function, where $H_0 = pk_a$ (indicator) + log ($c_{\mathrm{indicator}}/c_{\mathrm{H[indicator]+}}$)].

Clays. Natural clays have attracted relatively low attention, however, pillaring those materials with interlayer anions yields porous materials similar to zeolites. Although those materials have been described to be useful for alkylation, especially for larger molecules [26], [27], little is known about their industrial application in alkylation reactions. Generally, the thermal stability and chemical resistance is much smaller than that of zeolites.

Heteropolyacids or polyoxoanion compounds constitute a large category of coordination-type salts and free acids with each member containing a complex and high molecular mass anion. In these anions, two to eighteen hexavalent molybdenum (or tungsten) atoms surround one or more central atoms (heteroatom). Vanadium, niobium, tantalum, and transition metals can replace some of the molybdenum (or tungsten) atoms in the heteropoly structure. Typical examples are $[PMo_{12}O_{40}]^{3-}$, $[SiW_{12}O_{40}]^{4-}$, $[PMo_{10}V_2O_{40}]^{5-}$. An excellent review was given in [28]. As with clays, little is known about the commercialization of these materials in alkylation reactions.

Supported Acids. A variety of fixed-bed processes have been commercialized in which Lewis or Brønsted acids are supported on alumina, silica, or various other supports (e.g., carbon, graphite). A widely used catalyst of this type is "solid phosphoric acid" (phosporic acid adsorbed on kieselguhr), used for the alkylation of benzene with ethylene, butenes, and cyclohexene, but especially with propene. During use, the catalyst loses water which controls the equilibrium between ortho, meta, and poly phosphoric acid and hence catalytic activity. Small amounts (≦0.1 % of feed) of water or of the corresponding alcohol (e.g., isopropyl alcohol for the production of cumene) therefore are added to the reaction mixture in order to regenerate the catalyst and thus maintain activity [29]. Other examples of supported acids are boron trifluoride – alumina and solid superacids, such as SbF_5, TaF_5, and NbF_5, supported on fluorinated alumina or graphite. Graphite-intercalated $AlCl_3$ (i.e., $AlCl_3$ embedded in graphite) also is an effective solid Friedel – Crafts catalyst but loses activity by partial hydrolysis or leaching of the Lewis acid halide from the graphite [30]. Furthermore, Lewis-acidic metal halides supported on layered clays have attracted attention as heterogeneous catalysts.

Cation-Exchange Resins. *Solid Brønsted acids,* such as sulfonated styrene – divinylbenzene cross-linked polymers, are used as an acidic catalyst in a variety of organic reactions. *Cation-exchange resins,* such as Permutit Q, Lewatit, and Dowex 50, are effective catalysts for the alkylation of phenols with olefins, alkyl halides, and alcohols [31]. Catalysts of this type also have been applied in the acylation of thiophene with acetyl chloride or acetic anhydride [32]. Perfluorinated resin sulfonic acids, such as the acid form of DuPont's Nafion, catalyze many alkylations of aromatic hydrocarbons very efficiently [33], [34]. New developments in this field are DuPont's Nafion – silica nanocomposite [35] and Degussa's Deloxan materials [36].

1.1.2. Basic Catalysts

Basic catalysts are preferred for C-alkylations and C-acylations of CH acidic compounds. Here, the base increases the nucleophilicity of the CH acidic compound. The type of catalyst depends upon the substrate to be alkylated or acylated. Often, the compounds are of such low acidity that prior formation of a metal derivative is required. Aliphatic compounds with electron-attracting substituents, for example, CN or CO, usually are depronated by alkoxide, amide, and similar ions:

$$RCH_2CN + NH_2^- \longrightarrow R^{\ominus}CHCN + NH_3$$

$$RCH_2CN + OH^- \longrightarrow R^{\ominus}CHCN + H_2O$$

Commonly used bases include sodium alkoxides, sodium amide, and sodium hydride. Sometimes the powerful triphenylmethylsodium (tritylsodium) or even alkali metals (for example, Na in liquid NH_3 [37] or Na supported on γ-Al_2O_3 [38]) must be used. For example, toluene is alkylated with ethylene or propene to give propylbenzene or isobutylbenzene and cumene is alkylated with ethylene to give (1,1-dimethyl)propylbenzene.

Reactions catalyzed by mild bases were conducted on a wide variety of basic oxides (MgO, ZrO_2, La_2O_3) and basic microporous materials. The side-chain alkylation of toluene with methanol, aldol condensations, Knoevenagel reactions, and Michael additions represent the most common reactions of this type described in literature. A review on basic zeolites was provided in [39].

1.1.3. Organometallic Catalysts [40], [41]

A variety of catalysts based on transition-metal compounds are known in alkylation and acylation, as in the following examples (R,R′ = alkyl, L = $(C_6H_5)_3P$):

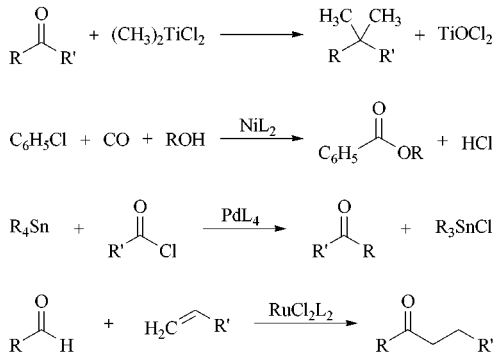

As yet, these catalysts are of no major industrial importance and therefore are not discussed further.

1.2. Alkylating and Acylating Agents

The most widely used alkylating agents are olefins and alkyl halides. The most important olefins are ethylene, propene, dodecene (tetrapropene), and linear olefins with 10 – 20 carbon atoms; all are used in large-scale industrial alkylation processes. Among the alkyl halides, methyl chloride, methyl iodide, ethyl chloride, isopropyl chloride, *tert*-butyl chloride, and benzyl chloride often are used.

Alcohols, ethers, esters of inorganic or organic acids, epoxides, aldehydes and ketones, thiols, and related compounds can be used, but in these cases more than catalytic amounts of the Lewis acid usually are required because the catalyst partly hydrolyzes.

Other powerful alkylating agents are dimethyl sulfate, diethyl sulfate, ethyl *p*-toluenesulfonate, trifluoromethanesulfonic acid, its methyl or ethyl ester, as well as aliphatic diazo compounds and triethyloxonium tetrafluoroborate [*368-39-8*]. Alkylating agents, especially

those of the last group, may present a health hazard.

The acylating agents most frequently used are acyl halides, carboxylic acids, and anhydrides. Less frequently used are ketenes, esters, lactones, etc. Amides [42] and phosgene also can function as acylating agents.

1.3. Mechanism

The majority of industrial alkylations and acylations are acid catalyzed (Friedel – Crafts type). Therefore only these processes are considered here. Both alkylation and acylation can proceed via similar reaction intermediates, involving similar reaction mechanisms.

1.3.1. Alkylation

Friedel – Crafts alkylation is induced by interaction of the acidic catalyst with the alkylating agent, as exemplified below for $AlCl_3$, leading to the formation of a highly polarized complex containing a carbocation:

$$R\text{-}CH{=}CH_2 + HX + AlX_3 \rightleftharpoons [R\text{-}HC^{\oplus}\text{-}CH_3][AlX_4]^-$$

$$RX + AlX_3 \rightleftharpoons [R]^+[AlX_4]^-$$

$$ROH + AlX_3 \rightleftharpoons [R]^+[AlX_3OH]^-$$

R = alkyl

In the case of alkyl halides, carbocations are formed in the complete absence of moisture or any proton acid, whereas in the case of olefins a proton donor is needed. The carbocation then attacks the aromatic substrate, forming the intermediate (**1**), which is transformed into the product by loss of a proton [43]:

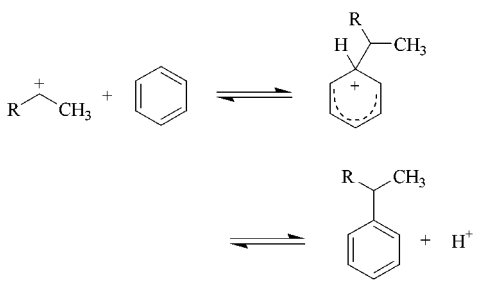

Under the acidic conditions used, both the alkylating agent and the alkylated aromatic compound isomerize. This leads in many cases to a situation in which neither the structure of the alkyl substituent nor the orientation in the nucleus can be predicted.

Friedel – Crafts reactions do not necessarily require free carbenium ions as intermediates; the reaction can proceed also via formation of a polarized complex between the Lewis acid and the alkylating agent followed by attack of the aromatic substrate [44]. However, carbocationic intermediates as postulated in Friedel – Crafts reactions were observed in superacidic media and were studied by spectroscopic methods [45].

Alkylation on Solid Acids. The mechanism of alkylation of aromatics on solid acid catalysts (especially on zeolites) has been reviewed by Venuto [46]. Briefly, alkylation involves an electrophilic addition of a carbocation (formed through protonation of or hydride abstraction from the alkylating agent) to the aromatic ring. However, depending on the basicity of the aromatic ring compared to that of the alkylating agent and depending on temperature either the aromatic ring or the olefin is preferentially adsorbed on the acid sites and thus, forms the carbocation [47]. Although many different heterogeneous catalysts may be used, a major benefit of microporous materials is offered by their shape selectivity. Three different types of shape selective effects have been postulated, i.e., (1) reactant selectivity, (2) transition state selectivity, and (3) product selectivity. Since the pore size of zeolites and the kinetic diameter of the reacting molecules are in the same order of magnitude, bulky isomers/molecules are selectively prevented to (1) enter the zeolite channels, (2) to be produced in the transition state, or (3) to leave the zeolite pores. Commercialized processes using this concept are, e.g., Mobil's Selective Toluene Disproportionation Process (MSTDP) or Mobil's xylene isomerization, both using ZSM-5 zeolites. In the former process, toluene is disproportionated to *p*-xylene and benzene; *m*- and *o*-xylene are produced in minor amounts due to diffusion limitations in the zeolite channels. Mobil/Badger's gas-phase

EB process using a ZSM-5 zeolite and UOP's Q-Max cumene process as well as many other commercial processes (see below) avoid the formation of polysubstituted aromatics effectively due to shape selectivity. A comprehensive review on the industrial application of shape selective catalysis was given by CHEN and GARWOOD [48].

1.3.2. Acylation

The mechanistic features of acylation can be illustrated by the reaction of an acyl halide with an aromatic substrate using $AlCl_3$ as the catalyst. The reaction of the acyl halide with the Lewis acid leads to a more or less polarized complex or even an acylium ion, both of which can act as the active electrophile as shown below [49] (R = alkyl, aryl):

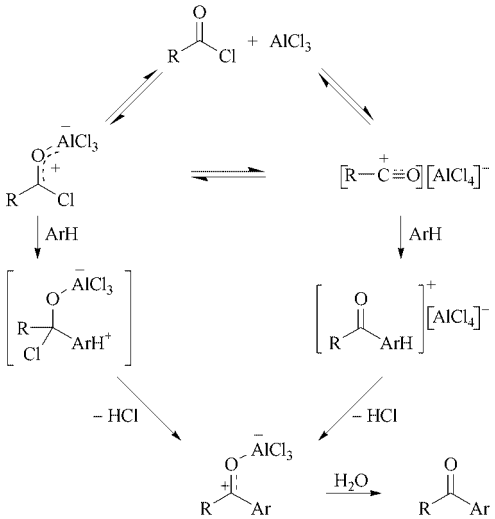

The formation of significant amounts of the free acylium ion can be expected only in highly polar solvents, such as nitrobenzene or nitromethane. The acylium ion mechanism is also favored if steric hindrance occurs either in the acylating compound or in the aromatic substrate [49]. Also, if anhydrides, esters, or the carbonic acids themselves are used as the acylating agents in the presence of proton acids, a high concentration of acylium ions can be expected.

2. Alkylation and Acylation of Aromatic Compounds

The acid-catalyzed conversion of aromatic compounds to their alkylated or acylated derivatives was discovered as early as 1877 by FRIEDEL and CRAFTS. Using anhydrous aluminum trichloride as catalyst, they were able to demonstrate the reactivity of aromatic hydrocarbons toward organic halides and unsaturated compounds, organic acid anhydrides, phosgene, carbon dioxide, and other compounds. Several important industrial processes then emerged, including the synthesis of ethylbenzene, cumene, higher alkylbenzenes, alkylphenols, benzophenone, and anthraquinone. These compounds have found broad application in the production of polymers, detergents, perfumes, pharmaceuticals, dyes, and agricultural chemicals.

2.1. Alkylation

The Friedel – Crafts alkylation of aromatic compounds is an acid-catalyzed electrophilic substitution of an aromatic hydrogen by an alkyl group. A variety of alkylating agents can be used, but typically olefins, alkyl halides, and alcohols are applied. All aromatic substrates accessible to electrophilic substitution can be used, including heteroaromatic compounds and such compounds as ferrocene. These fast, exothermic reactions usually are carried out under mild conditions in the liquid phase, although, for some substrates, vapor-phase processes with more stringent conditions are applied. Using olefins, alkyl halides, and alcohols as alkylating agents, the following overall reactions occur:

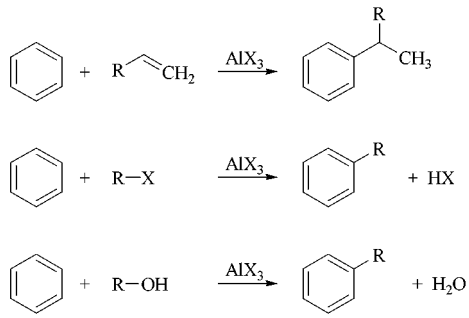

Depending on the acidity of the catalyst, the alkyl moiety is more or less completely isomerized, and the addition follows the Markovnikov rule.

Catalysts for Friedel – Crafts alkylation must be acidic. Depending on reaction conditions, alkylating agents, and aromatic substrates, Lewis as well as Brønsted acids are active. Table 2 lists frequently used catalysts and classifies them on the basis of chemical constitutions.

Although acidic halides, such as $AlCl_3$ or $AlBr_3$, and proton acids, such as HCl, HF, or H_2SO_4, or H_3PO_4 supported on diatomaceous earth are widely used, those catalysts are being progressively substituted by zeolites and inorganic oxides because the latter are environmentally benign, easy to handle and to dispose, thus implying less expensive process design.

All types of bases reduce the acidity of the catalyst and hence its efficiency. This is especially true for organic nitrogen compounds. Heavy resinous materials are formed on the catalyst if oxygen is present even at concentrations as low as 20 ppm [50]. Dienes, such as butadiene in a butene feed, also cause the formation of resinous or tarry materials on the catalyst. Traces of moisture or other proton suppliers in many cases enhance catalytic activity, especially of Lewis acids. Higher amounts of these compounds are detrimental to the catalyst. Therefore, it is not surprising that when alcohols are used, for example, for the alkylation of benzene, 0.5 – 1.0 mol $AlCl_3$ per mole of alcohol is necessary because the water produced reacts with the Lewis acid. In some cases, mild catalyst poisons are used deliberately in order to lower catalyst acidity and to prevent side reactions, such as isomerization. Typical examples are organic nitro compounds, which form stoichiometric complexes with Lewis acids. Compounds such as $AlCl_3 \cdot RNO_2$, $AlBr_3 \cdot RNO_2$, $GaCl_3 \cdot RNO_2$, or SbF_5 catalyze the benzylation of benzene or toluene cleanly, giving moderate yields, without causing the excessive side reactions that occur with the uncomplexed Lewis acids [13]. Similarly, such adducts as $BF_3 \cdot OEt_2$ or $BF_3 \cdot C_6H_5OH$ act as mild Friedel – Crafts catalysts.

Impurities in the catalysts can alter both the yield and the course of the reaction catalyzed by Lewis acids. For example, the presence of ferric chloride in aluminum chloride can decrease product yield markedly, and hence the catalytic activity of Lewis acids depends on the method of their preparation.

Isomerization of the Alkylating Agents. Because of the acidic reaction conditions, the alkylating reagents often undergo rearrangements and are generally transformed into the most highly branched isomer possible. For instance, the reaction of benzene with propyl chloride leads mainly to isopropylbenzene, and that with *n*-butyl chloride to the formation of *sec*-butylbenzene. Under Friedel – Crafts conditions, isomerization of the alkylating agent is initiated by migration of a hydrogen atom and not by direct carbon rearrangements. The driving force is the formation of the most stable carbocation, following the usual trend: primary < secondary < tertiary. For example, the alkylation of benzene with long-chain linear 1-olefins leads not only to the 2-alkylbenzene but in most cases also to a random distribution of the phenyl substituent along the alkyl chain. This effect could result also from the isomerization of the alkylated reaction products, because *n*-butylbenzene is transformed into *sec*-butylbenzene in the presence of a Friedel – Crafts catalyst, such as $AlCl_3$ [51].

Depending on the reaction conditions, alkylation of benzene with tertiary alkyl halides often leads to the formation of secondary alkylbenzenes rather than to the tertiary alkylbenzenes. For example, alkylation of benzene with 2-chloro-2,3-dimethylbutane gives mainly 2,2-dimethyl-3-phenylbutane via isomerization of the primary product *tert*-hexylbenzene:

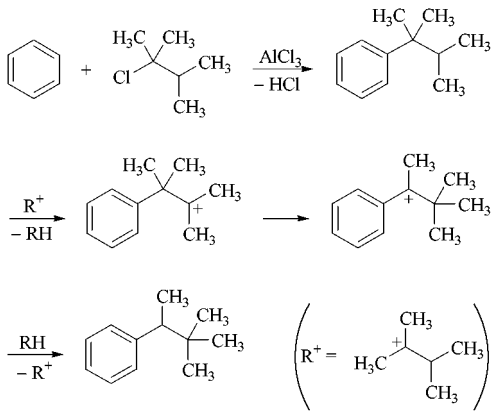

Table 2. Catalysts for the Friedel – Crafts alkylation of aromatic compounds

Group	Examples	Type of acid
Acidic halides	$AlCl_3$, $AlBr_3$, $SnCl_4$, $SbCl_5$, $FeCl_3$	Lewis acids
Metal alkyls and alkoxides	AlR_3, BR_3, ZnR_2, $Al(OPh)_3$	Lewis acids
Proton acids	HCl, HCl – $AlCl_3$, H_2SO_4, HF, H_3PO_4	Brønsted acid
Acidic oxides and derivatives	zeolites, mixed oxides and solid superacids, clays, heteropolyacids	Brønsted/Lewis acid
Supported acids	$H_3PO_4 - SiO_2$, $BF_3 - Al_2O_3$	Brønsted/Lewis acid
Cation-exchange resins	Permutit Q, Amberlite IR 112, Dowex 50, Nafion – silica, Deloxan	Brønsted acid

The primary product, *tert*-hexylbenzene, can be isolated only if less active catalysts, such as ferric chloride, are chosen [52].

Reactivity of Aromatic Substrates. The Friedel – Crafts alkylation of aromatic compounds, as an electrophilic substitution, is influenced by substituents present on the aromatic substrate. For example, electron-donating substituents, such as alkyl, NR_2, or OR (R = alkyl), lead to enhanced reactivity, and the new substituent enters preferentially in the ortho or para position. Therefore the alkylation of benzene with olefins leads to the monosubstituted benzene, which in turn reacts faster with a further olefin molecule than benzene itself does, resulting in a complex product mixture of mono- and polyalkylated benzenes [53]:

$$C_6H_6 + \text{olefin} \xrightarrow{k_1} C_6H_5R$$

$$C_6H_5R + \text{olefin} \xrightarrow{k_2} C_6H_4R_2$$

$$C_6H_4R_2 + \text{olefin} \xrightarrow{k_3} C_6H_3R_3$$

R = isoalkyl; $k_1 < k_2 < k_3$

The degree of polysubstitution, however, is influenced by reaction conditions and is lowered by steric hindrance by bulky alkyl groups, as shown in Table 3.

Benzene can be monoalkylated in good yields by bulky olefins, such as cyclohexene or linear $C_6 - C_{18}$ olefins [54]. In the case of ethylene or propene monoalkylation is made possible by recycling the polyalkylated products and by choosing suitable reaction conditions, as in the industrial synthesis of ethylbenzene and cumene. In the alkylation of monosubstituted benzenes, initially the ortho and para isomers predominate, as Table 4 shows. However, in most cases subsequent isomerization leads to an equilibrium mixture rich in the meta product. Only if the ortho position is sterically hindered is para substitution favored. For example, the reaction of benzene and 4-methylcyclohexene, catalyzed by HF, gives 75 % monoalkyl derivative and 25 % dialkyl derivative, of which 75 % is the para isomer and only 25 % the meta isomer [55]. Similarly, alkylation of cumene by 2-butene over $BF_3 - H_3PO_4$ yields 91 % para product [56], and the dialkylation of benzene with propene over silica – alumina gives preferentially *p*-diisopropylbenzene, which can be used for the synthesis of terephthalic acid and hydroquinone [57]. Substitution of benzene by halides leads to deactivation, as is evident from Table 5.

Table 4. Initial distributions (mass fraction in %) in the alkylation of toluene [53]

Alkylating agent	% Product distribution		
	o	*m*	*p*
C_2H_4	48	30	22
C_3H_6	42	21.5	36.5

Phenols or phenyl ethers react much faster than benzene in alkylation because of the considerably enhanced electron density in the ortho and para positions.

Table 5. $AlCl_3$-catalyzed alkylation with propene in CH_3NO_2 at 25°C [58]

Aromatic compound	k_{rel}	Isomer distribution, mass fraction in %		
		o	*m*	*p*
Benzene	1.00	–	–	–
Fluorobenzene	0.23	41.3	2.1	56.4
Chlorobenzene	0.11	53.9	5.1	41.0
Bromobenzene	0.07	55.4	8.6	36.0

Reactivities as well as product and isomer selectivities in microporous materials largely depend on the type, the number and the strength

Table 3. Relative rate constants for polysubstitution [53]

Reactants	Catalyst	Temp., °C	k_1	k_2	k_3	k_4	k_5	k_6
$C_6H_6 + C_2H_4$	$AlCl_3$	70	1	0.52	0.23	0.08	0.12	0.15
$C_6H_6 + C_3H_6$	$AlCl_3$	70	1	0.85	0.28	0.02	–	–
$C_6H_6 + C_3H_6$	HF	70	1	2.24	0.06	–	–	–
$C_6H_5CH_3 + C_3H_6$	$BF_3 - H_3PO_4$	50	1	0.22	0.06	–	–	–

of the acid sites as well as on the geometry of the zeolite pores (see page 6). Generally, higher temperatures are required for the alkylation process using solid catalysts, exceptions from this rule being only observed with very strong solid acids, i.e., solid superacids. An excellent review on mechanisms and product distributions of organic reactions on zeolites is given in [46].

Dealkylation and Transalkylation. Many of the catalysts used to alkylate aromatic compounds can be used, especially at higher temperatures, to reverse the alkylation reaction [59]. For example, alkylbenzene heated with $AlCl_3$ form mixtures of benzene and polyalkylated benzenes:

$$C_6H_5R \underset{}{\overset{AlCl_3}{\rightleftharpoons}} C_6H_6 + C_6H_5R + C_6H_4R_2 + C_6H_3R_3 + \ldots$$

The tendency of alkylaromatic compounds to disproportionate depends also on the structure of the alkyl substituent. The following order is observed [60]:

$$\textit{tert}\text{-}C_4H_9 > \textit{iso}\text{-}C_3H_7 > \textit{n}\text{-}C_3H_7 > C_2H_5 > CH_3$$

For example, disproportionation and transalkylation of methylbenzenes takes place at 300 – 450 °C in the presence of solid acid catalysts such as $SiO_2 - Al_2O_3$, whereas disproportionation and transalkylation of ethylbenzenes and isopropylbenzenes occur below 100 °C with typical Friedel – Crafts catalysts [53].

This intermolecular alkyl group transfer finds different applications in industrial processes. For example, in the synthesis of ethylbenzene, the undesired dialkylation product, which is formed in considerable amounts, reacts with an excess of fresh benzene:

$$C_6H_4(C_2H_5)_2 + C_6H_6 \rightleftharpoons 2\,C_6H_5(C_2H_5)$$

By this transalkylation process a high conversion of ethylene to ethylbenzene is achieved [61]. The reverse reaction, the disproportionation of monoalkylbenzenes, leads almost exclusively to the meta isomer in the presence of an excess of $BF_3 - HF$ [62]:

$$2\,C_6H_5(C_2H_5) \rightleftharpoons C_6H_6 + C_6H_4(C_2H_5)_2 \text{ (99 \% meta)}$$

This is attributed to the high stability of the complex $[m\text{-}Et_2C_6H_5]^+[BF_4]^-$. A similar type of complex is used commercially by Japan Gas to extract *m*-xylene selectively with $HF - BF_3$ from mixtures of C_8 aromatic compounds [63].

The alkylation – dealkylation equilibria mentioned above also explain product distribution in the alkylation of substituted aromatic compounds. The ortho and para isomers formed initially, which can be isolated after short reaction times (kinetic control), isomerize with prolonged reaction times to the more stable meta isomers (thermodynamic control).

Dealkylation is used commercially to convert toluene to benzene, to isomerize *m*-xylene to the desired ortho and para derivatives, and to disproportionate toluene to benzene and xylenes.

2.1.1. Alkylation of Benzene

There are three major commercial uses of the alkylation of benzene:

1) Alkylation with ethylene to ethylbenzene as an intermediate for styrene production. The production capacity for ethylbenzene was ca. 20×10^6 t/a in 1991, where the USA, Western Europe, and Japan held 5.3, 4.8, and 3.3×10^6 t/a, respectively. More than 95 % of the ethylbenzene produced worldwide is converted to styrene, minor amounts are used as solvents.
2) Alkylation with propene to cumene as an intermediate for phenol production, with acetone as the byproduct. The production capacity for cumene was ca. 7.5×10^6 t/a in

1992, where the USA, Western Europe and Japan held 2, 1.7 and 0.7×10^6 t/a, respectively. The largest fraction of the world's cumene production is converted to phenol in the Hock process, minor quantities are used admixed with other alkylbenzenes as octane enhancers.

3) Alkylation with long-chain olefins (with 10 – 18 carbon atoms) to alkylbenzenes, which are sulfonated to give surface-active compounds. The production capacity for linear alkylbenzenesulfonates (LAB) in 1992 in the USA, Western Europe, and Japan was 180 000, 485 000 and 197 000 t/a, respectively.

Alkylation of benzene with ethylene gives ethylbenzene [*100-41-4*] in a fast, exothermic reaction:

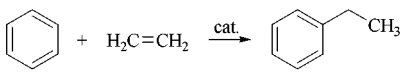

($\Delta H = -113$ kJ/mol)

Considerable amounts of polyethylbenzenes are formed as side products in this process, which can be carried out in the liquid phase or in the vapor phase. *1,4-Diethylbenzene* [*105-05-5*] (→ Hydrocarbons) is of commercial interest because its dehydrogenation product, 1,4-divinylbenzene [*1321-74-0*], is a crosslinking agent for polyolefins and is used widely in the production of ion-exchange resins.

In the *liquid-phase processes,* $AlCl_3$ is the most widely used catalyst, but other catalysts, such as BF_3, $FeCl_3$, $ZrCl_4$, $SnCl_4$, H_3PO_4, or alkaline-earth phosphates, also are used commonly [63]. The reaction is carried out at atmospheric pressure at 85 – 95 °C [64], [65]. For process details, see → Ethylbenzene.

The *vapor-phase* ethylation of benzene is common practice, especially in the United States, and typically is carried out at about 300 °C and 40 – 65 bar over acidic heterogeneous catalysts such as $Al_2O_3 \cdot SiO_2$ (Koppers Co.) or $H_3PO_4 - SiO_2$ (Universal Oil Products). The selectivity to monoalkylation is improved by use of a low ethylene-benzene (0.2 : 1.0) ratio [65], [66].

Universal Oil Products developed the Alkar process, used commercially in several plants since 1966. The catalyst is the noncorrosive boron trifluoride supported on a modified anhydrous alumina, which is also active for de- and transalkylation [67]. An unusually wide range of ethylene concentration (5 – 95 %) in the feed gas can be used, such as that in the off-gas from catalytic cracking, or that from a demethanizer overhead; alternatively, high-purity ethylene can be used. For example, the Alkar process can be adapted readily for use in the refinery. Preferred process conditions are temperatures of about 300 °C and pressures of about 60 bar, leading to quantitative conversion of ethylene [68]. For further details, see → Ethylbenzene.

The isomer distribution (mass fraction in %) obtained by alkylation of benzene with ethylene using the Alkar catalyst is as follows:

Diethylbenzenes:
1,2- (26);
1,3- (49); and
1,4- (25)
Triethylbenzenes:
1,2,3- (3);
1,2,4- (35); and
1,3,5- (62)

A variety of other metal halide Friedel – Crafts-type catalysts and mineral acids have been used on the laboratory scale for benzene ethylation [69], [70]. Of special interest are zeolite catalysts, which have been studied widely for this reaction [71], [72]. A commercial fixed-bed vapor phase process utilizing Mobil's proprietary ZSM-5 zeolites was introduced by Mobil Oil and Badger [73]. The reaction is carried out at 420 – 430 °C and 15 – 20 bar. Pure or diluted ethylene can be used as the feedstock after purification from sulfur, water, and propene. Polyalkylbenzenes can be recycled for transalkylation, resulting in a final yield of ethylbenzene of 98 – 99 mol %. The catalyst is regenerated periodically using oxygen-containing gas streams.

More recently, Raytheon Eng.& Constr. Inc. licensed Mobil/Badger's EBMax Process using Mobil's proprietary MCM-22 zeolite catalyst [74]. This process operates in liquid phase, contrary to Mobil/Badger's ZSM5-zeolite based gas-phase technology [46]. Another fixed-bed, liquid-phase alkylation technology is licensed by Lummus/Unocal/UOP employing a unique

zeolite-based catalyst developed by Unocal. The catalyst is based on a Y-type zeolite [75].

The latest zeolite alkylation technology yields ethylbenzene of 99.5 – 99.9 % purity. Catalyst lifetime is indicated to range between 1 and 3 years. Catalyst regeneration is easily achieved by thermal treatment in oxygen containing gas streams.

Zeolite catalysts offer the advantage of being noncorrosive and environmentally benign compared to liquid Brønsted or Lewis acid catalysts. This renders construction requirements, handling, and disposal much easier, thus, yielding economic benefits.

Amorphous aluminum silicates similar to those utilized in catalytic cracking also have been studied [76]. Although these catalysts give low conversions per pass, they are regenerated easily. Also, polyethylbenzene production is moderate and recycling by transalkylation is possible [76].

Alkylation of benzene with propene leads in an exothermic reaction to cumene [*98-82-8*]:

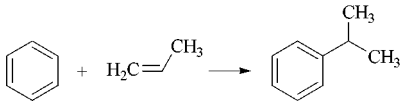

($\Delta H = -113$ kJ/mol)

The first cumene processes were developed between 1939 and 1945 to meet the demand for high-octane aviation gasoline. Today, almost all of the cumene produced as a pure compound is used for the synthesis of phenol (→ Phenol); acetone is also recovered from the process. The remainder is utilized for the production of α-methylstyrene [65], [77].

Both liquid- and vapor-phase processes are used for the commercial synthesis of cumene. The most widely used is the *liquid-phase process* of Universal Oil Products, which operates with a solid phosphoric acid catalyst (SPA, H_3PO_4 supported on diatomaceous earth). The conversion is carried out at 15 – 35 bar and temperatures of 200 – 250 °C. A 5 : 1 molar ratio of benzene to propene is necessary to prevent polyalkylation and to achieve a high conversion (ca. 94 %) of propene. The conversion is lowered to 70 % if a 4 : 1 benzene – propene feed is used. Increase of pressure and temperature enhances conversion rather than yields of cumene [78].

Beside the SPA catalyst, sulfuric acid and $AlCl_3$ are used for the production of cumene. For further information, see [79] and → Hydrocarbons., Chap. 3.7.

Novel, zeolite-based cumene production technologies were developed by Mobil/Raytheon, UOP, CDTech, Dow/Kellogg and Enichem [80–83]. Mobil is using its own MCM-22 zeolite, UOP utilizes a proprietary zeolite in its QMax process. Dow is using a highly dealuminated mordenite zeolite and Enichem employs a BEA-zeolite catalyst [75].

Common to most of the novel processes are two reactors and a distillation unit. In the first reactor the alkylation takes places, while in the second reactor heavy ends (tri- and polyisopropylbenzenes) are transalkylated with fresh benzene. CDTech is using a slightly different reactor concept, i.e., a reactive distillation (→ Reactive Distillation) unit employing a Y-type zeolite. For references see also [84–86]. Mobil's technology is easy to retrofit to existing plants and by mid-1998 six revamps with a total capacity of about 3×10^6 t/a were in operation [81].

Alkylation of cumene with propene, over a pretreated aluminum silicate catalyst and in the liquid phase under pressure at 150 – 250 °C, leads to a mixture of mainly *p*-diisopropylbenzene [*100-18-5*] and *m*-diisopropylbenzene [*99-62-7*] with only traces of the ortho isomer.

p-Diisopropylbenzene (*bp* 210.4 °C at 101.3 kPa) can be separated by distillation from the mixture of *m*-diisopropylbenzene (*bp* 203.2 °C) and *o*-diisopropylbenzene (*bp* 203.8 °C), which is isomerized in a second reactor at 240 – 320 °C over the same aluminum silicate catalyst. In this second reactor, polyalkylbenzenes are transalkylated with benzene to give further *p*-diisopropylbenzene [87]. *m*-Diisopropylbenzene can be produced in a similar manner, with only slight modifications. The diisopropylbenzenes are also important intermediates for the production of hydroquinone (or resorcinol) (→ Hydroquinone).

Alkylation of Benzene with Higher Olefins. Monoalkylbenzenes with 10 – 14 carbon atoms in the alkyl chain, such as dodecylbenzene [*123-01-3*], are important intermediates

for alkylbenzenesulfonates, which are widely used in anionic commercial detergent products (→ Surfactants).

Until about 1965, the most important olefin for this reaction was propene tetramer. However, the resistance of detergents based on propene tetramer to biodegradation led to environmental problems and hence to the introduction of linear alkylbenzenesulfonates. The feedstock for the production of alkylbenzenesulfonates therefore consists of linear olefins with 10 – 14 carbon atoms. The catalysts generally used for this reaction are liquid HF, H_2SO_4, or $AlCl_3$ and the following conditions are suitable [50]:

1) HF: 0 – 10 °C, 4 – 10 mol of benzene to 1 mol alkene, 2 – 3 kg of alkene per kilogram of HF
2) H_2SO_4: 10 – 30 °C, 4 – 5 mol of benzene to 1 mol alkene, 2 volumes of alkene per volume of 96 – 98 % H_2SO_4
3) $AlCl_3$: 55 – 60 °C, 3 – 4 mol of benzene to 1 mol olefin, 25 mol of alkene per mole of $AlCl_3$

The HF process is reported to be the most economical with respect to catalyst consumption. The process is carried out in the liquid phase at atmospheric pressure and 10 °C. Effective stirring and cooling is necessary to remove the heat of the exothermic reaction. This determines the reaction times. The acid strength is maintained at 79 % HF, and a high benzene to olefin ratio is needed to insure complete conversion of olefins and a high selectivity to monoalkylates. The products are separated from the acid in a settler, washed, and distilled at atmospheric pressure to recover benzene and at reduced pressure to isolate the alkylbenzenes [65], [88].

More recently, many solid alkylation catalysts for the production of linear alkyl benzenes have been described in the literature, however, since a slow-down in LAB demand has taken place in the industrialized regions, only UOP (Detal process) has commercialized a new technology using a noncorrosive catalyst to replace HF. UOP has patented a fluorided silica – alumina catalyst [89] which is operated in liquid phase at about 130 °C. The process achieves 92 % linearity (i.e., unbranched alkyl chain, see Table 6) at 100 % conversion and a selectivity of 90 %.

The detergent properties of the alkylbenzenesulfonates are not only influenced by the degree of branching of the alkyl side chain – three to five methyl groups in dodecylbenzene are optimum, whereas the highly branched butene oligomers were not suitable as starting material [64] – but depend also on the position of the phenyl group on the alkyl chain [90]. Irrespective of the position of the olefin double bond, Friedel – Crafts alkylation leads to an equilibrium product mixture with a random distribution of the phenyl over the alkyl chain. The product composition, however, is affected by the nature of the catalyst, as shown in Table 6, where the product distributions obtained from the reaction of benzene with 1-dodecene [*112-41-4*] utilizing $AlCl_3$, HF, and H_2SO_4 as catalyst are listed [54].

Table 6. Isomer distribution of dodecylbenzenes obtained from 1-dodecene and benzene

Phenyldodecane, Mass fraction in %	Catalyst, temperature		
HF, 16 °C	$AlCl_3$, 30 °C	H_2SO_4, 0 °C	
1-	0	0	0
2-	20	32	41
3-	17	22	20
4-	16	16	13
5-	23	15	13
6-	24	15	13

As expected, no terminal phenyldodecanes are observed. With $AlCl_3$ and H_2SO_4 as the catalyst, the 2-phenyl isomer is formed preferentially, and the 3- to 6-phenyl isomers occur in lesser amounts. In contrast, with HF, the 3- to 6-phenyl isomer formation is enhanced. This can be explained by the ability of Lewis acids, such as $AlCl_3$, to isomerize the phenylalkanes formed, whereas the proton acids isomerize only the olefins prior to the final alkylation of the benzene.

Alkylation of benzene with alkyl chlorides also is used commercially for the production of detergent alkylates. The alkyl chloride obtained from the chlorination of paraffins contains about 70 % paraffin. This mixture then reacts with a large excess of benzene in an enameled alkylating tower at 80 °C using $AlCl_3$ as the catalyst. The hydrogen chloride, which is formed in stoichiometric amounts, insures thorough mixing of the reactants and is separated at the top of the reactor. The products are separated from the catalyst in a settler, neutralized, washed, and distilled to recover benzene and paraffin, which are recy-

cled; they are then distilled under reduced pressure to isolate the alkylates [91] (→ Surfactants).

Considerable amounts of indane and tetralin derivatives are formed as byproducts because of the dichloroalkane impurities in the alkyl chloride:

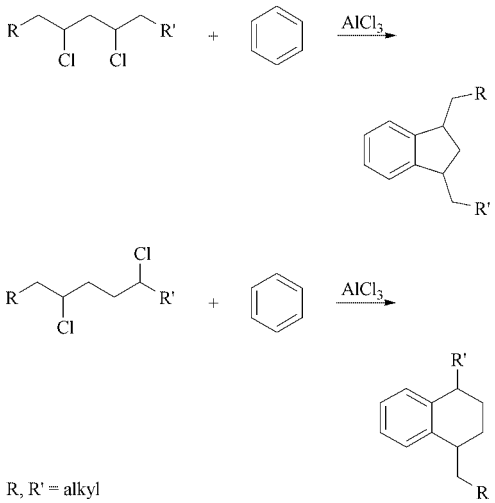

Alkylbenzenes with more than 15 carbon atoms in the side chain lead to alkylbenzenesulfonates that are almost insoluble in water but are extremely soluble in organic media, such as mineral oils. Such solutions in mineral oil are used in the metal-working industry as drilling and cutting fluids or in the textile industry as spinning or batch fluids. Alkylbenzenesulfonates of this type are being used increasingly as flooding agents in the enhanced recovery of petroleum.

Miscellaneous Alkylations of Benzene. Benzene can be alkylated by a variety of other linear, branched, cyclic, aryl- or otherwise substituted olefins, alkyl halides, alcohols, ethers, and compounds such as alkynes and dienes. For a comprehensive review of these reactions, see [92]. Some of these reactions have been commercialized, such as the synthesis from benzene and isobutene of *tert*-butylbenzene [*98-06-6*], which is used for the production of perfumes.

2.1.2. Cycloalkylation

A typical example of *intermolecular* cycloalkylations is the formation of 9,10-dihydroanthracene [*613-31-0*] (together with some diphenylmethane) from dichloromethane and benzene:

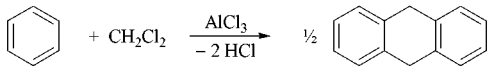

Anthracene derivatives also are formed in this reaction [93].

Intramolecular cycloalkylation is possible for alkylated aromatic compounds carrying a double bond or a functional group, such as a halogen or hydroxy group, in the side chain. For example, 3-phenyl-1-propene [*300-57-2*] is isomerized to give indane [*496-11-7*] in the presence of $AlCl_3$ [94]:

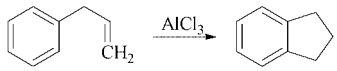

A synthesis for anthraquinone [*84-65-1*], based on cycloalkylation, was developed by BASF:

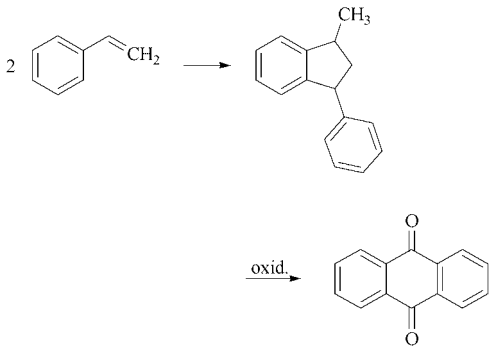

Styrene is dimerized in the liquid phase to 1-methyl-3-phenylindane, utilizing phosphoric acid as catalyst, followed by catalytic vapor-phase oxidation to anthraquinone [95], [96].

Arylation of aromatic compounds can be achieved by reaction with aryl halides, by reaction with diazonium halides, and by dehydrogenating condensation.

Because of their low reactivity, aryl halides can be used only under specific conditions as Friedel–Crafts arylating agents. Their reactivity is F ≫ Cl > Br. Aromatic hydrocarbons can be arylated with fluorobenzene to give the corresponding biphenyls [97]. The decomposition

of aryldiazonium tetrafluoroborates in aromatic solvents gives ring-arylated products as well as fluorobenzene [98].

The Scholl reaction is the dehydrogenating condensation of an aromatic compound in the presence of Friedel–Crafts catalysts. Typical examples are the formation of perylene [*198-55-0*] from naphthalene

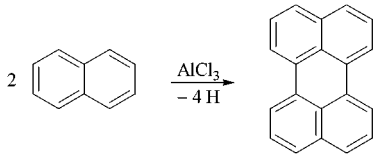

and the formation of 2,2′-dipyridyl [*366-18-7*] from pyridine

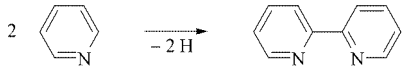

High molecular weight polycondensated aromatic compounds also can be formed [92].

2.1.3. Alkylation of Substituted Benzenes

As pointed out in Section 2.1.1, the alkylation of benzene rarely leads to the monoalkylated product only, but to a complex mixture with di- and polysubstituted compounds. The reactivity as well as the orientation in the second substitution step depends on the electronic and steric properties of the first substituent and on the bulkiness of the alkylating agent.

Some dialkylated benzenes are of considerable commercial importance [65], [73]. For example, diethylbenzenes are useful for improving the octane ratings of motor fuels and are obtained as side products in ethylbenzene production processes. Dialkylbenzenes can be dehydrogenated to give divinylbenzenes that are used for cross-linking

of polyolefins, such as polystyrene in the production of ion-exchange resins. Similarly, ethylmethylbenzene, obtained by ethylation of toluene, is of industrial interest as a starting material for styrenes that are methyl substituted in the aromatic ring. Another example is the production of diisopropylbenzene by reaction of propene with cumene (see page 12).

The alkylation of toluene with propene to give cymene [*25155-15-1*] is carried out commercially, for example, by Sumitomo Chemicals [99]:

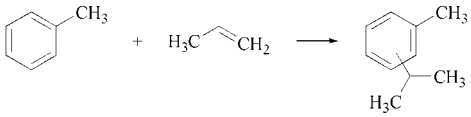

The cymene formed is rich in the meta and para isomers and is oxidized to give acetone and cresols, which consist of ≧99.5 % *m,p*-cresols and ≧60 % *m*-cresol and are used as raw materials for phenolic resins or for plasticizers.

Polymethylated benzenes obtained by disproportionation, for example, of toluene (see Section 2.1.5), or methylation of benzene or toluene [100] can be oxidized to the corresponding acids, which are used in such polymers as polyesters, polyamides, and polyimides.

A vast amount of literature on preparative alkylations of substituted benzenes exists, and this field has been comprehensively reviewed [101]. Some of these reactions are applied for the production of fine chemicals, such as dyes and pharmaceuticals.

Alkylation of higher aromatic compounds, such as biphenyl, indane, and tetralin, or of polynuclear aromatic compounds, such as naphthalene, acenaphthene, anthracene, fluorene, and phenanthrene, also has been reported [101].

The *alkylation of naphthalene* [*91-20-3*] with higher olefins has found commercial application for the production of fluorescent agents used in lubricating oils. The reaction is carried out in the liquid phase using $AlCl_3$ as the catalyst at temperatures up to 180 °C [102]. Another application of naphthalene alkylation is the production of detergents of the "Nekal" type (BASF), which are alkylnaphthalenesulfonates. These compounds were among the first synthetic detergents and are still used in the textile industry as surfactants:

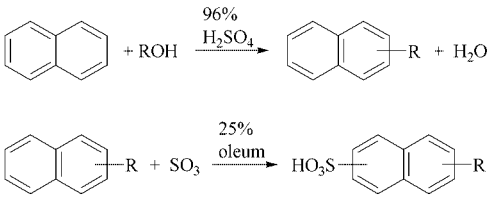

They are prepared in a "one-pot" process by reaction of naphthalene with butanol or hexanol or mixtures of propanol – butanol in the presence of 96 % sulfuric acid at 50 °C. After the alkylation is complete, 25 % oleum is added for sulfonation. The oxidation of 2-isopropylnaphthalene, according to the Hock process, leads to 2-naphthol [*135-19-3*]. 2-Isopropylnaphthalene can be obtained together with small amounts of the 1-isopropyl isomer by reaction of naphthalene with propene using $ZnCl_2$ or aluminum silicates as the catalyst at 150 – 200 °C [103].

2,6-Dialkylnaphthalenes have attracted much interest. They are oxidized to 2,6-naphthalenedicarboxylic acid which is used for the production of the high-performance plastic poly(ethylene-2,6-naphthalene dicarboxylate) (polyethylenenaphthenate, PEN). Various shape-selective catalysts have been described to alkylate naphthalenes, among which mordenite seems to be the most effective [104]. However, another route to 2,6-dimethylnaphthalene was commercialized by Amoco in 1995 at a capacity of about 40 000 t/a. Amoco alkylates *o*-xylene with butadiene over a strongly basic catalyst yielding *o*-tolylpentene, which further undergoes cyclization, dehydrogenation, and isomerization [105].

2.1.4. Alkylation of Phenols

The alkylation of phenols is a very important reaction industrially, and the alkylphenols are used widely in a variety of applications, such as antioxidants, herbicides, insecticides, or polymers. The nucleophilicity of the hydroxyl group is associated with increased electron density in the aromatic ring, which facilitates electrophilic substitution. Therefore alkylation occurs under conditions milder than those needed with aromatic hydrocarbons [101].

Depending on the reaction conditions, catalysts and alkylating agents, predominantly *ortho*- or *para*-monoalkylated, 2,4- or 2,6-dialkylated, and 2,4,6-trialkylated phenols are formed.

Thermodynamically more stable *meta* isomers are formed at higher temperatures by secondary reactions. At low temperatures and low catalyst concentrations, alkyl aryl ethers also are found in significant amounts, whereas at higher temperatures, catalyst concentration, and acid strength, the alkylphenols are formed predominantly [106]. Olefins, alcohols, and alkyl halides are the alkylating agents. For industrial phenol alkylation, three main types of processes have been developed:

1) *Liquid-phase alkylation,* in which the alkylating agent is added to a solution or suspension of the catalyst in the phenol. The catalyst must be separated after the reaction, and the last traces of the catalyst decomposed by neutralization.
2) *Fixed-bed liquid-phase processes,* in which the reactants pass over the catalyst at elevated pressures in tubular reactors, which are cooled externally.
3) *Vapor-phase alkylation,* in which the phenol and the alkylating agent, such as methanol or an olefin, are sent to a fixed-bed reactor operating at 300 – 400 °C. The heat of reaction must be removed efficiently.

The product is recovered in all cases by distillation. The unconverted phenols and alkylating agents are recycled and the various mono-, di-, and trialkylated phenols separated. In order to prevent side reactions, such as isomerization, transalkylation, and disproportionation, during distillation, it is essential to remove all traces of the acidic catalyst.

Catalysts. In the liquid-phase processes, Lewis acid catalysts, such as $AlCl_3$, BF_3, or $Al(C_6H_5O)_3$; proton acids, such as HF, H_2SO_4, H_3PO_4, or *p*-tolylsulfonic acid; or acidic ion-exchange resins, such as Nafion resins or sulfonated polystyrenes, are applied. Catalysts suitable for the vapor-phase processes must be of high thermal stability. Typical examples are montmorillonites ($Al_2O_3 \cdot 4\,SiO_2 \cdot H_2O$), zeolites, such as faujasite, and metal oxides, such as γ-Al_2O_3, MgO, CaO, and $Fe_2O_3 - CrO_3$ [107–109].

Alkylation with alcohols, especially methanol, usually is carried out over metal oxides, preferably MgO. For olefins, the acidic catalysts and γ-Al_2O_3 are used. The heterogeneous catalysts, such as ion-exchange resins or zeolites, gradually are replacing the liquid acids because of easier product separation and lower catalyst consumption [107–110].

Regioselectivity can be controlled using a variety of catalysts. If HF, BF_3, $Al_2O_3 \cdot SiO_2$, or acidic ion-exchange resins are used, the alkylation of phenols with olefins leads predominantly to the *p*-substituted product. In contrast, $Al(C_6H_5O)_3$, which is formed from phenol and metallic aluminum, especially in the presence of traces of mercury salts, leads almost exclusively to the 2-alkylated and 2,6-dialkylated products. The high selectivity for ortho alkylation is explained by a six-membered transition state involving aluminum phenolate and an olefin [111]. Other metal salts of phenol, including those of zinc, magnesium, and calcium, are active. This reaction is not restricted to phenol; 2-naphthol, 2-cresol, or 4-cresol also can be alkylated by ethylene, propene, the butenes, cyclohexene, or other olefins. The highest reactivity is observed for isobutene, the lowest for ethylene [112], [113].

High selectivity to monoalkylated *o*-phenols by reaction of olefins with phenols is achieved with γ-Al_2O_3 as the catalyst. The reaction is carried out at elevated temperatures of 200 – 350 °C in the liquid phase at pressures of 20 – 200 bar. This process is well suited to continuous fixed-bed operation.

Heating alkylphenols, especially in the presence of acidic catalysts, leads to dealkylation, transalkylation, and isomerization [114]. The tendency of the alkylphenols to undergo these reactions depends on the nature of the alkyl substituent and increases in the order primary < secondary < tertiary alkyl; methyl substituents are rather inert. Isomerization leads to an equilibrium where the thermodynamically more stable 3 or 3,5 isomers prevail, and dealkylation occurs more easily in the ortho or para position than in the meta position. The regioselectivity of transalkylations is influenced also by the nature of the catalyst, and, therefore, in the presence of aluminum phenolate the ortho products prevail:

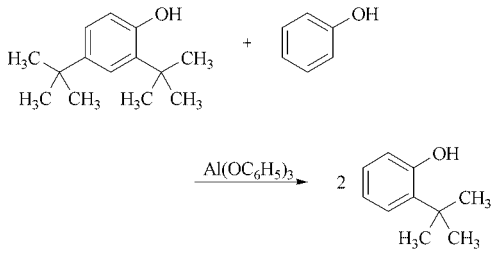

If different alkyl substituents are to be introduced into a phenol, the most sensitive tertiary alkyl is added last.

In industry, alkylphenols often are produced in batch operations, which offer higher flexibility. However, for compounds produced in large amounts, such as cresols, 2,6-xylenol (→ Cresols and Xylenols), 2-*sec*-butylphenol, 4-*tert*-butylphenol, octylphenol, nonylphenol, dodecylphenol, thymol, or 2,6-di-*tert*-butyl-4-methylphenol (→ Phenol Derivatives), continuous processes have been introduced with capacities of up to 25 000 t/a each. A survey of commercially important alkylphenols and their end uses is given in Table 7 [115–117].

Alkylation of phenol with methanol at elevated temperatures over a catalyst yields 2-cresol [*95-48-7*] (60 – 70 % at 50 % phenol conversion and 300 °C) and 2,6-xylenol [*576-26-1*], together with some 4-cresols, 2,4-, and 2,3-xylenols, and a little 3-cresol.

The reaction can be carried out in the liquid phase or in the vapor phase. In the *vapor-phase processes,* the reactants are fed at atmospheric or slightly elevated pressure at 300 – 450 °C into a tubular reactor containing the fixed-bed catalyst [118]. A high selectivity to 2-cresol and 2,6-xylenol is observed with activated Al_2O_3 at 300 °C [119] and with MgO and CeO_2 at 540 and 450 °C, respectively [120]. Use of highly acidic aluminum oxide, aluminum silicates, zeolites, and aluminum phosphate, especially at higher temperatures, leads to transalkylation-isomerization and thus to increased formation of 3-cresol and 4-cresol. In *liquid-phase processes,* catalysts such as Al_2O_3, $Al(OMe)_3$, $Al(OC_6H_5)_3$, or $ZnCl_2$ – HCl are used at elevated pressures [121], [122]. In all cases, the xylenol : cresol ratio increases with the methanol : phenol ratio, with temperature, and with pressure.

2,3,6-Trimethyphenol [*2416-94-6*] is produced by a continuous liquid-phase process from 2,6-xylenol and methanol, using a fixed-bed Al_2O_3 catalyst at 280 – 380 °C in a tubular reactor [123].

Alkylation of Phenols with Olefins. 2-Ethylphenol [*90-00-6*] is formed from ethylene and phenol in the presence of 1 – 2 % $Al(C_6H_5O)_3$ at 320 – 340 °C and 200 bar,

Table 7. Production and end use of alkylphenols [115], [117]

Alkylating agent	Phenol	Alkylphenol	End use
Methanol	phenol	2-, 3-, and 4-cresol	insecticides, herbicides, antioxidants, disinfectants, resins, plasticizers
		2,6-xylenol	polyphenylene oxide resins
		3,5-xylenol	vitamin E
Methanol	3,6-xylenol	2,3,6-trimethylphenol	vitamin E
Ethylene	phenol	2-ethylphenol (2,6-diethylphenol)	photochemicals
Propene	phenol	2-isopropylphenol (2,6-diisopropylphenol)	resins, plasticizers, antioxidants, insecticides, perfumes
Propene	3-cresol	3-methyl-6-isopropylphenol (thymol)	perfumes, disinfectants
	2-cresol	2-methyl-6-isopropylphenol (carvacol)	disinfectants, perfumes
1-Butene	phenol	2-*sec*-butylphenol (2,6-di-*sec*-butylphenol)	insecticides, herbicides, acaricides, resins
Isobutene	phenol	2-*tert*-butylphenol (2,6-di-*tert*-butylphenol)	antioxidants, insecticides
		2,4,6-tri-*tert*-butylphenol	acaricides, perfumes, resins
		4-*tert*-butylphenol	alkylphenol resins
Isobutene	4-cresol	2,6-di-*tert*-butyl-4-methylphenol	antioxidant
2-Methylbutene	phenol	4-*tert*-amylphenol	resins, antioxidants
Diisobutene	phenol	4-*tert*-octylphenol	resins, detergents, antioxidants, emulsifiers, surfactants
Tripropene	phenol	4-isododecylphenol	
Tetrapropene	phenol	4-isododecylphenol	

whereas the alkylation of phenol with propene to 2-isopropylphenol [*88-69-7*] can be carried out at 280 – 300 °C using γ-Al_2O_3 as the catalyst [108], [124]. The propylation of cresols to give thymol [*89-83-8*] or carvacol [*499-75-2*] is achieved in continuous liquid-phase processes using γ-Al_2O_3, aluminum silicates, or zeolites [115]. Under similar conditions, 2-*sec*-butylphenol [*89-72-5*] can be produced from phenol and *n*-butenes [124].

Because of the high reactivity of isobutene, the production of 2-*tert*-butylphenol [*88-18-6*] and 2,6-di-*tert*-butylphenol [*128-39-2*] can be carried out at 100 °C at pressures up to 20 bar in the presence of $Al(C_6H_5O)_3$. 4-*tert*-Butylphenol [*98-54-4*] is prepared at atmospheric pressure at 80 – 140 °C in the presence of H_2SO_4, H_3PO_4, BF_3, or preferably acidic cation-exchange resins [125].

The production of the important antioxidant 2,6-di-*tert*-butyl-4-methylphenol [*128-37-0*] from 4-cresol and isobutene is carried out at 70 °C using H_2SO_4 as the catalyst [126]. The alkylation of phenol with 2-methylbutenes at 100 – 120 °C in the presence of BF_3 or H_3PO_4 yields 4-*tert*-amylphenol [*80-46-6*], in the presence of an excess of 2-methylbutenes also 2,4-di-*tert*-amylphenol [127].

A most important process commercially is the alkylation of phenol with linear or branched olefins containing 6 – 20 carbon atoms to give alkylphenols, which after ethoxylating and sulfating, are used widely as anionic detergents. Diisobutene, tripropene, and tetrapropene are preferred as alkylating agents, and the reaction is catalyzed under mild conditions (50 – 85 °C) by BF_3 in continuous-batch or tubular reactors. Acidic cation-exchange resins, such as macroreticular sulfonated polystyrene, also are excellent catalysts that overcome the problem of catalyst separation. To achieve better temperature control, the reaction may be carried out in two stages, using a catalyst of reduced acidity in the first stage. A process of this type is utilized by Chemische Werke Hüls (now Degussa-Hüls) in Germany to produce nonylphenol [107], [110], [115]. In the first stage of this fixed-bed process, the temperature increases from 80 to 120 °C and in the second stage from 110 to 130 °C. Activated aluminum silicates also are used as catalysts but require higher reaction temperatures and longer reaction times.

A great variety of alkylphenols, haloalkylphenols, and related compounds are produced industrially and are used as oxidation inhibitors, detergents, bactericides, fungicides, insecticides, and disinfectants. This field

has been reviewed extensively (→ Cresols and Xylenols; → Phenol Derivatives) [101], [107], [108], [128].

2.1.5. Alkylation of Aromatic Amines

Because of the basicity of aromatic amines, acidic catalysts are not very suitable for the alkylation of these substrates. For example, the reaction of ethylene with aniline in the presence of the Friedel – Crafts catalyst $AlCl_3$ leads mainly to unidentified resinous products with small amounts of 2-ethylaniline. However, *aluminum anilide,* $Al(C_6H_5NH)_3$, which is formed in situ with the evolution of hydrogen by heating aluminum powder in aniline, is an active catalyst for this reaction. At 300 – 340 °C and 200 bar, 2,6-diethylaniline [*579-66-8*] is formed after 2 – 3 h, whereas at shorter reaction times 2-ethylaniline [*578-54-1*] also is formed [129].

The catalytic activity of aluminum anilide is enhanced by addition of Friedel – Crafts catalysts, such as $AlCl_3$, $SnCl_4$, $TiCl_4$, $SiCl_4$, BF_3, or $ZnCl_2$. Mixtures of alkali metals and $AlCl_3$ in aniline are also active systems, whereas alkali metals alone lead selectively to N-alkylation. The reaction has been extended to a variety of substituted anilines, such as 2-, 3-, and 4-toluidine, 2,6-, 2,4-, and 2,5-xylidine, 2-propylaniline, 2- and 3-chloroaniline, and α-naphthylamine. Diamines, such as *m*-phenylenediamine or 2,4-toluylenediamine, as well as secondary amines, such as diphenylamine, also can be alkylated.

The reactivity of the olefins decreases in the order ethylene > propene > 1-butene > isobutene, which is the reverse of the order observed in phenol alkylation with $Al(OC_6H_5)_3$ as the catalyst. However, in the case of isobutene, Friedel – Crafts catalysts, such as $AlCl_3$, BF_3, or montmorillonite, are active even at 200 – 250 °C, whereas with aluminum anilide only slow reaction occurs. With montmorillonite, 4-*tert*-butylaniline [*769-92-6*] is formed selectively, whereas with BF_3 or $AlCl_3$, mixtures of the 2 and 4 isomers are produced [129], [130].

The alkylation of anilines with methanol is catalyzed by molecular sieves. A complex mixture of 2-toluidine, 2,6- and 2,4-xylidine, and 2,4,6-trimethylaniline is formed. 2,6-Dialkylanilines, for example, 2-ethyl-6-methylaniline [*24549-06-2*], are produced industrially as intermediates for herbicides and dyes [115].

2.1.6. Alkylation of Heteroaromatic Compounds and Related Substrates

A variety of heterocyclic aromatic compounds, such as furanes, thiophenes, and *N*-heterocycles, can be alkylated, although little is known about commercial application of these reactions. Cyclopentadienyl derivatives undergo Friedel – Crafts alkylation in a similar manner.

Furan appears to be much more reactive in Friedel – Crafts alkylations than benzene. However, resinification of furans in the presence of acids occurs quite rapidly, which often makes product separation difficult. Using phosphoric acid on kieselguhr, furan can be alkylated with isobutene in 3 h at 100 °C to give 85 % yield of the alkylates together with small amounts of isooctene:

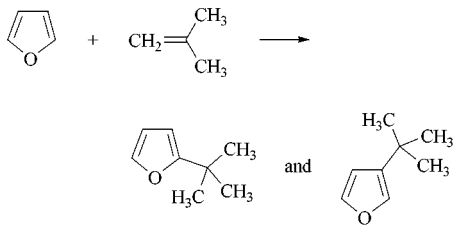

Whereas substitution in the 2 position is favored under mild conditions (50 °C, 1 h), the formation of products substituted in the 3 position is enhanced at longer reaction times (7 h), higher temperatures (150 °C), and greater catalyst concentrations. The formation of 2,5-di-*tert*-butylfuran [*4789-40-6*] also was observed [57], [101], [131].

Thiophene appears to be substantially more reactive in Friedel – Crafts reactions than aromatic hydrocarbons and can be alkylated with olefins, such as propene, 1-butene, 2-methylbutenes, or cyclohexene, and with alcohols, such as isopropyl alcohol and *tert*-butyl alcohol. Acidic catalysts, such as phosphoric acid at 60 °C or alumina silica at 200 °C, were used successfully, whereas H_2SO_4, HF, $BF_3 \cdot OMe_2$,

and $AlCl_3$ led to side reactions. In all cases, the thiophene substituted in the 2 position was observed, as well as higher-boiling material, such as dialkylated products [57], [132]. The removal of thiophene from cracked gasoline by refining with cold concentrated sulfuric acid is probably an alkylation reaction. The olefins present in the gasoline act as alkylating agents to give high-boiling alkylthiophenes, which remain as residue upon distillation [133].

***N*-Heterocycles** can be alkylated in the presence of aluminum amides. Because heterocyclic compounds, such as carbazole and indole, react only slowly with aluminum metal to give the amide, aluminum anilide is used as the catalyst. For example, carbazole reacts with ethylene in the presence of aniline and aluminum powder to give 1-ethylcarbazole [*19275-57-1*]. The reaction takes place in an autoclave at 280 °C and 200 bar.

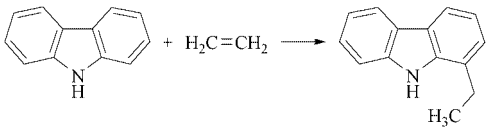

Propene reacts under similar conditions to give 1-isopropylcarbazole [*1484-09-9*], but with a lower yield [130].

Alkylations using acidic catalysts also have been reported. For example, the formation of diisopropylcarbazole from carbazole and propene was observed at 90 – 100 °C in the presence of $AlCl_3$ [134]. The same reaction over an aluminum silicate gave tetra- and pentaisopropylcarbazole at 180 – 200 °C [135], and the ethylation of pyrrole was achieved at 200 – 370 °C over an aluminum silicate [101], [136].

Nonbenzenoid aromatic compounds, such as ferrocenes, fulvenes, azulenes, pseudoazulenes, indolizines, and cyclazines, resemble chemically not only the benzenoid systems but also the olefins. Compared to benzene derivatives, their electrophilic substitution reactions occur under milder conditions. Because these systems are often of low stability, both catalysts and reaction conditions have to be selected carefully.

Ferrocene [*102-54-5*] can be alkylated, in the presence of Lewis acids, such as $AlCl_3$, with a variety of substrates, such as methanol, ethyl, isopropyl, butyl, and benzyl halides, 1,2-dichloroethane, ethylene, propene, and isobutene. The reaction often yields complicated mixtures of homoannular and heteroannular (i.e., both rings are alkylated) products, because of the increased reactivity of the system after the first alkylation step [137], [138]. Alkylation of ferrocene with olefins seems more favorable; that with isobutene to give *tert*-butylferrocene occurs even at room temperature. At 100 – 150 °C, heteroannular tri-*tert*-butylferrocene is obtained in 44 % yield, together with some tetra-*tert*-butylferrocene. Steric hindrance prevents the attachment of more than two *tert*-butyl substituents to the same cyclopentadienyl ring [139]. A more detailed description of ferrocene alkylation and of alkylation reactions of other nonbenzenoid aromatic compounds can be found in [140], [141].

2.1.7. Miscellaneous Alkylation Reactions

The alkylation of aromatic compounds with bi- or polyfunctional reagents leads to a variety of useful mono-, di-, or polynuclear products of commercial interest.

Haloalkylations with bi- or polyfunctional alkylating agents to yield only primary haloalkylation products can be carried out under suitable reaction conditions. Cycloalkylation or linkage of two or more nuclei as shown in the second reaction can be avoided.

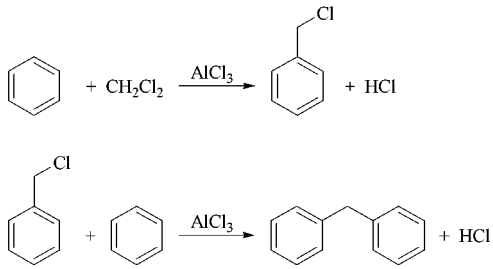

Typical haloalkylating agents are aldehyde – hydrogen halide, haloalkyl ethers, haloalcohols, and similar compounds [92]. Halomethylation, haloethylation, higher haloalkylation, as well as bis- and polyalkylation can be achieved. Zinc chloride is the

catalyst most often used; other catalysts include acidic halides ($ZnCl_2 - AlCl_3$, $SnCl_4$, $AlCl_3$ – ketones, $AlCl_3$ – pyridine) or proton acids (HCl, H_2SO_4, H_3PO_4, *p*-toluenesulfonic acid) [142].

A typical example of this reaction is the chloromethylation with formaldehyde – HCl (Blanc reaction) [143]:

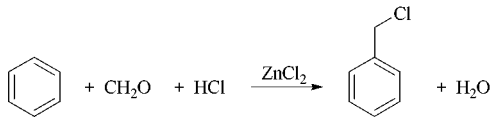

Because this reaction leads to reactive intermediates, it is of considerable synthetic value. However, great care must be exercised because the intermediate chloromethyl ethers are extremely carcinogenic.

Hydroxyalkylation of aromatic compounds with aldehydes and ketones in the presence of acids is industrially very important. The electrophiles attacking the aromatic nucleus are formed from the aldehyde-ketone by reaction with a Lewis or proton acid (R, R′ = H, alkyl, aryl):

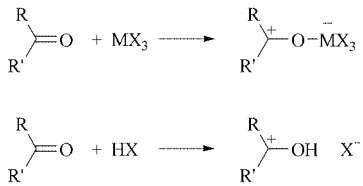

Further reaction proceeds according to the mechanism outlined in Section 1.3.1 and leads to a substituted benzyl alcohol; this, under acidic conditions, reacts with a further molecule of the aromatic compound to yield a substituted diarylmethane. This process is exemplified by the reaction of acetone with phenol, giving 2,2-bis-(4-hydroxyphenyl)propane (bisphenol A) (→ Phenol Derivatives):

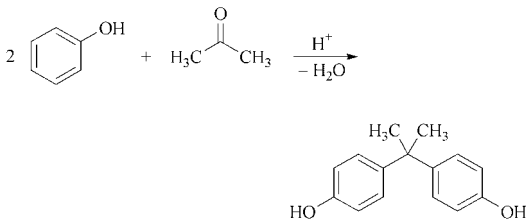

This compound is produced industrially in continuous processes using H_2SO_4 (or, in the Hooker process, dry HCl) as the catalyst and methyl mercaptan as promoter. Almost quantitative yields are obtained at 50 °C [63]. Also, cation-exchange resins are being used increasingly as catalysts, because of the simplified catalyst separation [107]. Bisphenol A is utilized mainly for the production of epoxy- and polycarbonate resins.

Hydroxyalkylations (Table 8) are extremely versatile; numerous alkyl- and aryl-substituted aldehydes and ketones react with a variety of aromatic compounds [144]. For instance, 1,1,1-trichloro-2-bis-(4-chlorophenyl)ethane (DDT) [*50-29-3*], once a widely used insecticide, is formed from chloral and chlorobenzene. The reaction of phenol with formaldehyde leads to hydroxymethylphenols, which condense at elevated temperatures to phenol resins.

If an oxirane is used as the alkylating agent, a β-hydroxyalkyl aromatic compound is obtained. For example, the reaction of ethylene oxide [*75-21-8*] with benzene in the presence of a stoichiometric amount of $AlCl_3$ leads to an aluminum alkoxide, which gives β-phenylethanol [*60-12-8*] on hydrolysis [145].

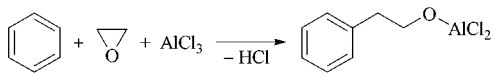

Amidomethylation is possible using the *Tscherniac-Einhorn reaction,* where *N*-hydroxymethylamides or the corresponding imides react with an aromatic compound under mild conditions in sulfuric acid [146].

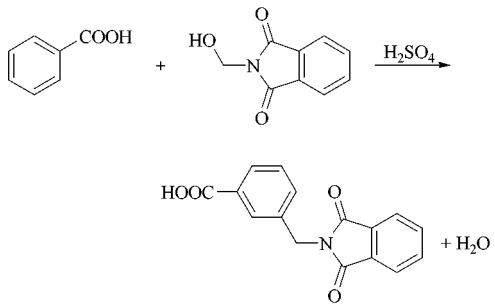

Less reactive aromatic compounds, such as benzoic acids, also can be amidoalkylated; even

Table 8. Alkylation of aromatic compounds by aldehydes and ketones [115]

Alkylating agent	Aromatic compound	Product	End use
H_2CO	C_6H_5OH	$2-HO-C_6H_4-CH_2OH$	salicylaldehyde, salicyl alcohol
H_2CO	$(C_6H_5O)_3B$	$(2-HOCH_2-C_6H_4-O)_3B$	2-cresol
H_2CO	$C_6H_5NH_2$	$(4-H_2N-C_6H_4-)_2CH_2$	methylenediphenyl diisocyanate, resins, adhesives
H_2CO	$2-Cl-C_6H_4NH_2$	$(3-Cl-4-NH_2-C_6H_3)_2CH_2$	polyurethanes
CH_3CHO	$C_6H_5CH_3$	$(4-CH_3-C_6H_4-)_2CHCH_3$	heat-exchanger liquid
CCl_3CHO	C_6H_5OH	$(4-HO-C_6H_4-)_2CHCCl_3$	polymers
HO_2CCHO	C_6H_5OH	$4-HO-C_6H_4-CH(OH)-CO_2H$	4-hydroxybenzaldehyde, 4-hydroxyphenylacetic acid

nitro substituted aromatic compounds can be alkylated using *N*-chloromethylcarboxylic acid amides in the presence of $AlCl_3$ [146].

2.2. Acylation

In Friedel – Crafts acylation, an aromatic ketone is formed by reaction of an aromatic compound with an acylating agent, such as an acyl halide, an acid anhydride, an acid, or an ester, in the presence of an acidic catalyst:

$$C_6H_6 + RCOX \xrightarrow{cat.} C_6H_5COR + HX \quad R = \text{alkyl, aryl}$$

The acylation of aromatic substrates is of considerable industrial interest for making aromatic intermediates used for the production of pharmaceuticals, insecticides, plasticizers, dyes, perfumes, and other commercial products. These speciality products usually are produced on a much smaller scale than alkylation products. Some examples of industrial interest are given in Table 9.

Acylation is extremely versatile, leading to a variety of products that include aromatic aldehydes, alkyl aryl ketones, symmetric and unsymmetric diaryl ketones, and cyclization products [147–149]. Generally, acylating agents react more readily than alkylating agents, and therefore acylations can be carried out usually under quite mild conditions. Electron-donating substituents in the aromatic substrate, such as alkyl, hydroxyl, or alkoxy groups, lead to enhanced reactivity.

On the other hand, aromatic compounds containing electron-withdrawing substituents, such as the nitro, acyl, carboxyl, or nitrile groups, are attacked only with great difficulty, even by the most active acylating agents. The introduction of the first acyl group hinders the entry of a second, and disubstitution in the same ring is therefore rare [147]. Examples for the disubstitution are the diacylation of mesitylene [150] or durene[151] and the diacylation of anisole [*100-66-3*] with a mixture of glacial acetic acid and phosphorus pentoxide [152]:

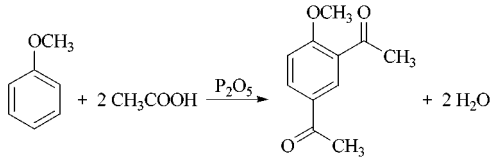

Industrially applied examples of multiple acylation of aromatics can be taken from [153]. As is evident from Table 10, electron-donating substituents facilitate acylation, and the increase of reactivity by polysubstitution is impressive.

Table 10. Relative rate of benzoylation of substituted benzenes [154]

Substrate	Relative rate
Chlorobenzene	0.011
Benzene	1
Toluene	154
m-Xylene	3 910
Mesitylene	125 000
Pentamethylbenzene	139 000

Besides benzene and its derivatives, polynuclear carbocyclic aromatic compounds, such as naphthalene, diphenyl, anthracene, phenanthrene, and pyrene, also can act as substrates in acylation reactions. Electron-rich heterocycles, such as pyrrole, furan, and thiophene, as well as their polycyclic derivatives are especially activated, whereas pyridine and quinoline do not react unless activated by substituents. Non-benzenoid aromatic substrates active in acylation include azulenes, ferrocenes, and indolizines [148].

Table 9. Aromatic ketones by Friedel – Crafts acylation [115]

Acylating agent	Aromatic compound	Product	End use
Acetic anhydride	benzene	acetophenone	perfumes, pharmaceuticals, solvent, plasticizer
Acetic anhydride	toluene	4-methylacetophenone	perfumes
Acetic anhydride	anisole	4-methoxyacetophenone	perfumes
Acetic anhydride	isobutylbenzene	4-isobutylactophenone	pharmaceuticals
Dichloroacetyl chloride	1,2-dichlorobenzene	α,α,2,4-tetrachloroacetophenone	insecticides
Chlorobutyroyl chloride	fluorobenzene	chloropropyl 4-fluorophenyl ketone	pharmaceuticals
Tetrachloromethane	benzene	benzophenone	pharmaceuticals, insecticides, perfumes
Benzoyl chloride	benzene	benzophenone	
Phosgene	*N,N*-dimethylaniline	4,4′-bis-dimethylaminobenzophenone	dyes
Phthalic anhydride	benzene	2-benzoylbenzoic acid	anthraquinone

Although acylation has much in common with Friedel – Crafts alkylation, there are characteristic differences. For example, consumption of the acidic catalyst is usually at least stoichiometric. However, there are examples in the literature where substoichoimetric amounts of Lewis acids are applied. The course of the reaction is more uniform than in Friedel – Crafts alkylation, leading to higher regioselectivity. Also, isomerizations and other side reactions are rare.

Catalysts. Essentially the same catalysts as discussed for the Friedel – Crafts alkylation can be used. In most cases, the catalyst of choice is $AlCl_3$ [*7446-70-0*] [155], [156]. Impurities in $AlCl_3$, such as $FeCl_3$ or traces of water, enhance yields and the reactivity of systems, which otherwise react only slowly [157–159]. In some cases, however, the high activity of $AlCl_3$ leads to undesirable side reactions, for example, the decomposition of reactive heterocycles, such as benzofuran, decomposition of ethers, or rearrangements of alkyl groups [160]. In these cases, the activity of $AlCl_3$ can be moderated by complexing solvents such as nitrobenzene, which form 1 : 1 adducts. BF_3 and $SnCl_4$ are also suitable acylating catalysts for sensitive heterocycles.

If anhydrides, esters, or carboxylic acids are used as the acylating agents, strong proton acids, such as sulfuric acid, perchloric acid, orthophosphoric, and polyphosphoric acid also can be applied. Polyphosphoric acid, originally utilized mainly for intramolecular cyclization reactions, is also being used increasingly in intermolecular acylations.

As already mentioned (Chap. 1), the reaction of an acyl halide with an aromatic substrate, catalyzed by $AlCl_3$, results in a stable complex of the Lewis acid with the aromatic ketone, from which the product is liberated by hydrolysis. Thus, in contrast to Friedel – Crafts alkylations, the catalyst must be usually applied in at least stoichiometric amounts with respect to the acyl halide. Examples for the unusual use of substoichiometric ("truely catalytic") amounts of Lewis acids are ferrous chloride [161–164], zinc or iron oxides [165], cobalt chloride [166], and rare earth metal triflates [167–170].

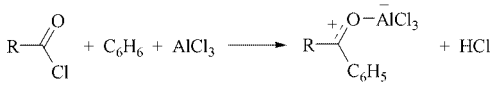

R = alkyl, aryl

At least a twofold amount of catalyst is necessary if carboxylic acids or esters are the acylating agent, and in the case of anhydrides a threefold amount (but 1.5 mol per mole of product):

$$RCO_2H + 2\,AlCl_3 \longrightarrow RCO\text{-}Cl \cdot AlCl_3 + AlOCl + HCl$$

$$RCO_2R' + 2\,AlCl_3 \longrightarrow RCO\text{-}Cl \cdot AlCl_3 + AlOCl + R'Cl$$

$$(RCO)_2O + 3\,AlCl_3 \longrightarrow 2\,RCO\text{-}Cl \cdot AlCl_3 + AlOCl$$

R,R′ = aryl, alkyl)

A catalyst with industrial importance is hydrogen fluoride [171], [172], sometimes in the combination with boron fluoride [173]. This has the advantage of avoiding waste formation due to the possibility to recover the catalyst by distillation. This technology is used commercially for the acylation of isobutylbenzene to 4-isobutylacetophenone, an intermediate for ibuprofen, a large volume anti-inflammatory drug.

Heterogeneous Catalysts. Even if the classical Friedel–Crafts Synthesis with homogeneous catalysts is still the most important process for the preparation of aromatic ketones, it has several disadvantages:

1) The use of stoichiometric or even greater amounts of "catalyst" is often required because of the formation of stable complexes with the products. Lewis acids like $AlCl_3$ are rather reagents then catalysts and contribute to a great extent to the production costs. They must be washed out of the product and cannot be recycled.
2) The reaction system is very corrosive and requires expensive reactors and also an expensive equipment for the purification of the waste gas, which consists not only of HCl, but also of chlorinated hydrocarbons.
3) The catalyst has to be neutralized and disposed together with a large amount of wastewater.

For this reason, much work has been done during the last years to substitute the homogeneous catalysts by heterogeneous catalysts, which do not have to be supplied in stoichiometric amounts and are not corrosive and can be easily separated, without neutralization. They can also be recycled for several times or used in a continuous fixed-bed process.

Heterogeneous acylation catalysts have to be strong Lewis or Brønsted acids. Only few classes of solid acids are strong enough to catalyze this reaction. The most important classes are:

Zeolites
Modified clays
Solid superacids
Heteropolyacids
Proton or Lewis acids on a support
Nafion and Nafion like composites

Zeolites are microporous, crystalline alumosilicates with strong Brønsted or Lewis centers. For the acylation reactions especially medium-pore systems like H-ZSM-5 or large-pore systems like HY zeolite, beta zeolite, or mordenite have successfully been used. Nevertheless only electron-rich aromatic compounds like anisol, alkylbenzenes, furanes, benzofuranes, phenoles or naphthalenes are active enough to react on zeolites in the liquid phase at temperatures below 150 °C.

Zeolites provide not only active acid sites but also the possibility to perform *shape-selective acylations*. The geometry of the micropores can control the regioselectivity of the reaction.

Clays. Clay minerals are sheet silicates with alternating layers of alumina and silica separated by nonbonding sheets. For catalytic purposes these minerals have to be modified to increase the number of acid sites. There are two types of clay catalysts: acid-treated clays and ion-exchanged clays. Acid treatment of clays by H_2SO_4 or HCl removes part of the alumina and also some other naturally occurring cations such as Mg^{2+} or Fe^{3+} from the layers, so that negatively charged alumosilicate sheets are generated, which are electrically neutralized by protons. These Brønsted acid catalysts (like K10) are commercially used for an number of acid-catalyzed reactions. The other type of catalysts are supported reagents, which are prepared by exchanging these protons with Lewis acid cations such as Zn^{2+} or Fe^{3+}. Commercial catalysts of this type, like Envirocat, can be used for the preparation of benzophenones even with less active aromatic compounds.

Solid Superacids. According to the definition of Gillespie superacids are acids, that are stronger than 100 % sulfuric acids. They have a Hammett acidity value H_0 of < -12. A number of solids belong to this class of materials: sulfated oxides, supported Lewis acids, and supported liquid superacids. Only a small number of oxides produces superacid sites on sulfation, including ZrO_2, TiO_2, HfO_2, Fe_2O_3 and SnO_2. The acid sites are generated by treating an amorphous oxihydrate of these elements with H_2SO_4 or $(NH_4)_2SO_4$ and calcining the products at temperatures of 500–650 °C. During the calcination step, the oxides are transformed into the crystalline tetragonal phase, which is covered by a small number of sulfate groups. Instead of H_2SO_4 also H_2MoO_4 or H_2WO_4 can be used to activate the oxide.

Miscellaneous. A number of other classes of catalysts have been descibed for acylation reactions: Heteropolyacids, surface mounted acids (proton or Lewis acids on a support), and Nafion. None of them is in commercial use.

Solvents. The aromatic substrate itself, nonpolar solvents, such as CCl_4 or CS_2, or solvents of medium polarity, such as dichloromethane

or 1,2-dichloroethane, may be used. If $AlCl_3$ is used as the catalyst, heterogeneous systems are formed with these solvents. However, in such solvents as nitromethane or nitrobenzene, homogeneous solutions are obtained, and at the same time the reactivity is reduced because of complex formation. The choice of solvent also can influence the regioselectivity of the acylation [174]:

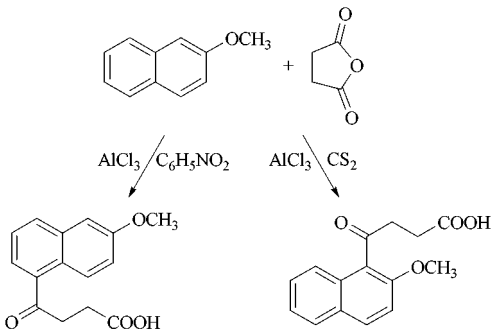

A hazard of this reaction is that the adducts of $AlCl_3$ and nitrobenzene tend to decompose explosively at elevated temperatures.

Acylation using an acyl halide – Lewis acid catalyst system usually is carried out according to one of the following procedures:

1) The acylating agent is added to a cooled solution or suspension of the catalyst in a solvent; the aromatic substrate then is added to the preformed acylating complex (*Perrier method*) [175]
2) All reactants are mixed and cooled, and the catalyst added slowly (*Elbs method*) [176]
3) The aromatic substrate is used as the solvent for the catalyst, and the acylating agent is then added slowly (*Bouveault method*) [177].

The Perrier method often is preferred, because a constant ratio of catalyst to acylating agent is maintained throughout the reaction.

If sulfuric acid is used as the catalyst, a large excess of the acid is needed, to which the reactants are added. In order to avoid ring sulfonation or aldolization reactions, the temperature should be kept below 100 °C.

Acylating Agents. The great versatility of Friedel – Crafts acylation is based on the fact that both the acylating agent and the aromatic substrate can be varied systematically and through a wide range. The reactivity of acylating agents follows roughly the following order:

$$[RCO]^+[BF_4]^- \approx [RCO]^+[ClO_4]^- > RCOO\text{-}SO_3H > RCOX > (RCO)_2O > RCO_2R' > RCONR'_2$$

Whereas the ionic acyl perchlorates and tetrafluoroborates are powerful agents, acid esters and amides are almost inactive in the absence of catalysts. The reactivity of the acyl halides decreases as follows: RCOI > RCOBr > RCOCl > RCOF.

However, the reactivity of acyl halides depends also on R, which may be alkyl or aryl and which can be substituted by halogen as well as by an alkyl, alkoxy, or a nitro group. The aromatic substrate and the type of catalyst also affect the reactivity [178], [179].

Aromatic acyl halides may well contain nitro substituents or be of the pyridine type, as in the synthesis of 3-benzoylpyridine [*5424-19-1*] from benzene and nicotinyl chloride [149]:

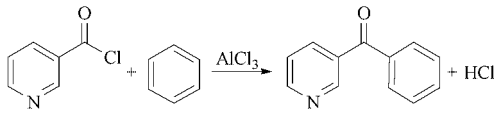

Bifunctional acyl halides also can act as the acylating agent. For example, the reaction of oxalyl chloride with a highly activated aromatic substrate, such as anisole, gives the corresponding benzil derivative [180]:

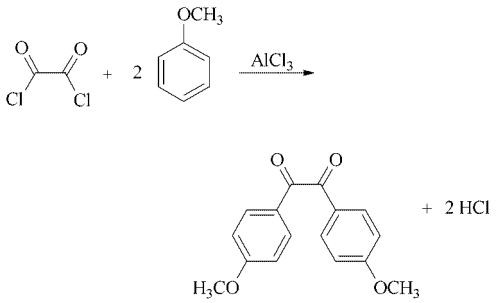

A commercially important reaction of phosgene is the reaction with *N,N*-dimethylaniline, yielding 4,4′-bis(dimethylamino)benzophenone (*Michler's ketone*) [*90-94-8*], which is used in the production of dyes:

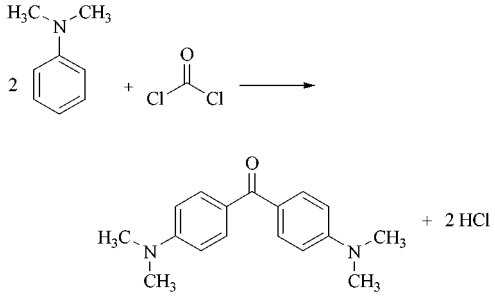

Ketone formation can be avoided if one halogen of the phosgene is replaced by an amido group (R=H, alkyl):

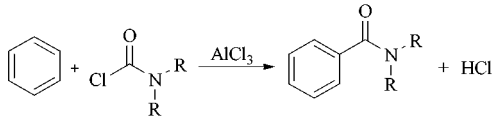

The resulting amide can be hydrolyzed to the acid [181]. This reaction has been considered for an industrial synthesis of *terephthalic acid* [*100-21-0*] [182]:

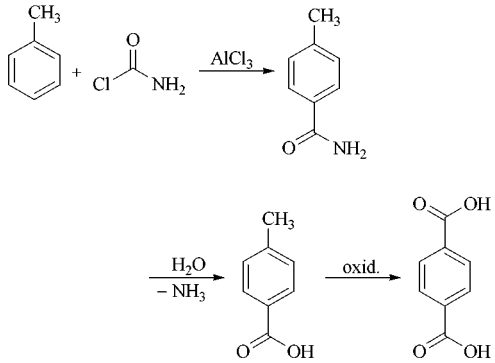

Isocyanates, isothiocyanates, and CO_2 also can be used as acylating agents in the presence of $AlCl_3$ to give aromatic carboxylic acids or their derivatives [92].

If an unsymmetrical bifunctional acylating agent, such as an acid chloride containing an ester group, is used, only the more active acyl group reacts [183]:

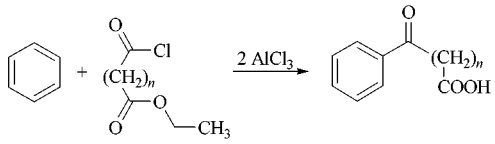

Essentially the same products are obtained with the acid anhydrides, such as succinic, glutaric, or maleic anhydrides, in the presence of 2 mol of $AlCl_3$ per mole of anhydride [148]. By this method, 4-aryl-4-oxobutanoic acids can be obtained, which are intermediates in the Haworth synthesis of polynuclear aromatics [184].

A commercially important synthesis is that of *anthraquinone* [*84-65-1*] from phthalic anhydride and benzene [185] (→ Anthraquinone):

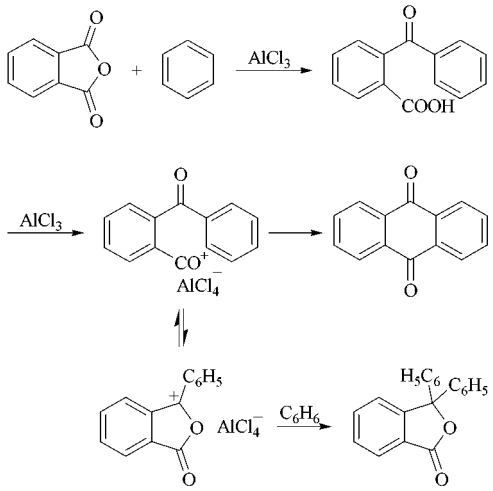

Under strongly acidic conditions, 3,3-diphenylphthalide [*596-29-2*] also can form. The synthesis of anthraquinone by this process is of special importance in the United States.

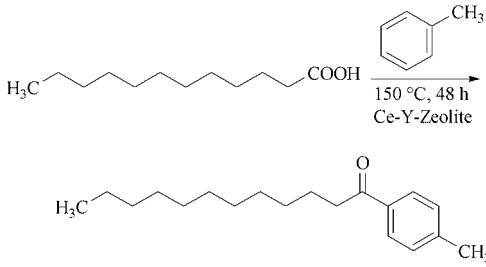

With zeolite catalysts, it is also possible to use nonactivated carboxylic acids. CHICHE found that higher fatty acids react in the liquid phase with toluene on Y-type zeolites to *p*-acyltoluenes with yields up to 96 % [186].

2.2.1. Acylation of Benzene and Benzene Derivatives

Acylation of benzene with acetic anhydride to give acetophenone [*98-86-2*] in 85 % yield was carried out commercially, using $AlCl_3$ as catalyst and keeping the temperature at 30 °C by cooling. Acetophenone is now produced industrially by the oxidation of ethylbenzene or of cumene (→ Ketones, Chap. 7.1.).

Benzene can be acylated with benzoic acid at elevated temperatures in the gas phase to give benzophenone [187]:

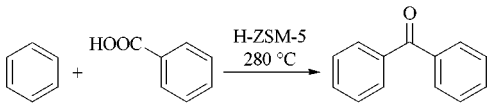

27% yield

Acylation of toluene, for example, using acetyl chloride and aluminum chloride, leads with 97.6 % selectivity to 4-methylacetophenone [*122-00-9*], which is used for the production of perfumes [188].

Acylation of Deactivated Substrates. Benzenoid aromatic compounds with nitro, carboxyl, nitrile, and acyl substituents that hinder electrophilic substitution can be acylated only if an activating second substituent, such as a hydroxy or alkoxy group, is present. Deactivated substrates like chlorobenzene can be acylated to *p*-chlorobenzophenones by use of commercial heterogeneous catalysts like zeolites or sulfated zirconia as solid superacids [189].

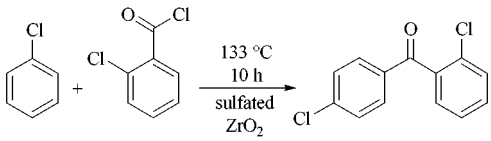

100% yield

Acylation of activated substrates, such as phenol, can be carried out using carboxylic acid – $ZnCl_2$, carboxylic acid – BF_3, or carboxylic acid – polyphosphorus acid as catalysts [190]. Up to 25 % selectivity of ortho substitution is observed for the acylation of phenols [191], and the ortho to para ratio depends strongly on the choice of catalysts and solvents.

Linear alkylphenols can be obtained via acylation followed by reduction. An example is the synthesis of *hexylresorcinol* [*136-77-6*], which is used commercially as a disinfectant [192]:

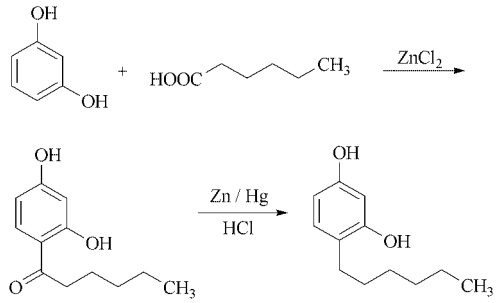

Both phenols and phenyl ethers belong to the most reactive substrates and therefore react under relatively mild conditions. For instance, anisole is acylated conveniently at 40 – 50 °C in the presence of polyphosphoric acid. Phenyl ethers are acylated almost exclusively in the 4 position.

A clean alternative to conventional Friedel – Craft acylation systems, which avoids formation of waste salts and byproducts, uses mixed anhydrides of trifluoroacetic acid which are formed in situ by reaction with carboxylic acids. Low concentrations of H_3PO_4 are used as the catalyst and trifluoroacetic acid is recycled after dehydration [193].

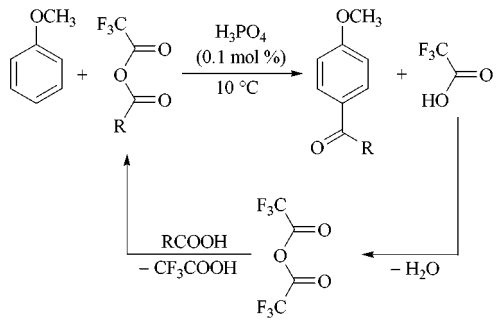

Acylation of Aromatic Amines. After N-acylation, aromatic amines react in a manner similar to alkylaromatic compounds in acylation reactions. Tertiary aromatic amines, such as *N,N*-dimethylaniline, can be acylated in the

4 position using P_2O_5 as catalyst, as is exemplified by the synthesis of Michler's ketone [194], [195].

Cyclization Reactions. The para-directing effect of substituents also is found here. 6-Methoxy-1-tetralone [*1078-19-9*] is obtained in 96 % yield from 4-(3-methoxyphenyl)butyric acid [196], [197]:

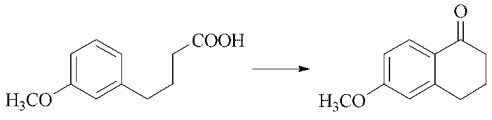

Important examples of intermolecular cyclizations are reactions of phthalic anhydride [*85-44-9*] with benzene and substituted benzenes to give *anthraquinone* [*84-65-1*] and substituted anthraquinones respectively (see bottom):

The reaction of phthalic anhydride with toluene or chlorobenzene leads, in a two-step process, to the 2-substituted anthraquinones, whereas the corresponding reaction with 4-chlorophenol gives 1,4-dihydroxyanthraquinone [*81-64-1*] (*chinizarine*) directly. Oxidation, nitration, chlorination, or amination leads to products that are important intermediates for fine chemicals. The Friedel–Crafts acylation of ethylbenzene with phthalic anhydride yields, in a two-step process, *2-ethylanthraquinone* [*84-51-5*], which is utilized in the synthesis of H_2O_2 [198].

2.2.2. Acylation of Polynuclear Aromatic Compounds

Acylation of naphthalene usually leads to mixtures of 1- and 2-acyl derivatives, the ratio of which depends mainly on the solvent used. Nonpolar solvents, such as CS_2 or CCl_4, favor acylation in the 1 position; polar solvents, such as CH_3NO_2 or $C_6H_5NO_2$, favor the formation of 2-acetylnaphthalenes [199].

Acylation of substituted naphthalenes often gives complex product mixtures. As a rule, acylation of the more highly activated ring is more likely to occur. However, it is possible to control the regioselectivity by the shape selectivity of zeolites. Thus, the acylation of 2-methoxynaphthalene on zeolite HY produces only 1-acetyl-2-methoxynaphthalene while on beta zeolite also 2-acetyl-6-methoxynaphthalene is formed [200].

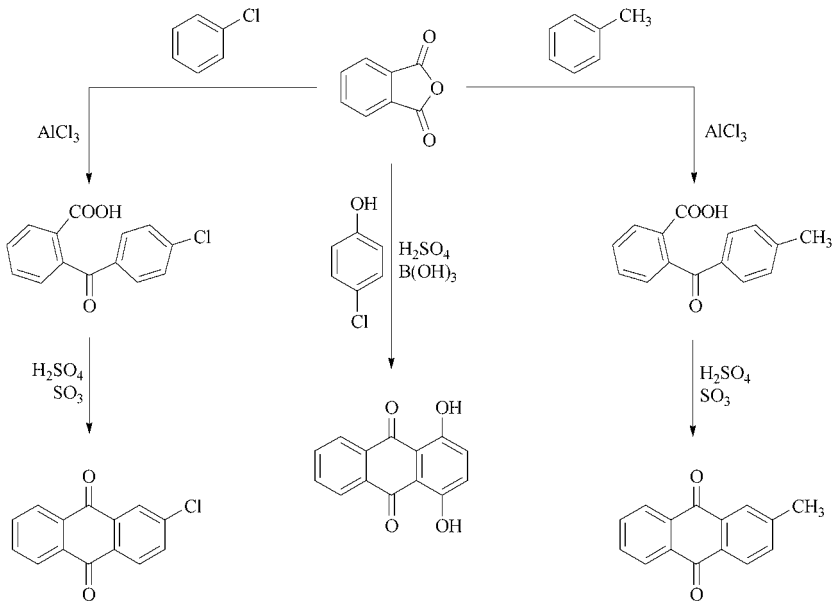

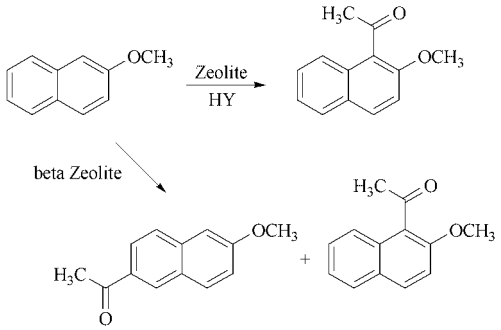

1,5-Diacylation is possible only under more drastic conditions, as in the synthesis of *1,5-dibenzoylnaphthalene* [*83-80-7*]:

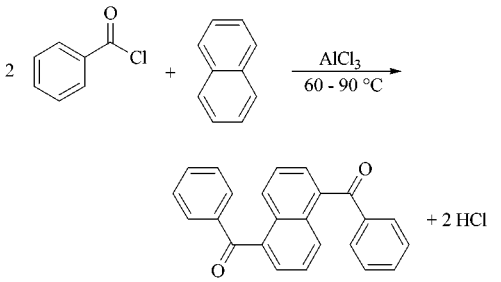

This compound can be fused in sodium aluminum chloride to give *dibenzopyrenequinone* [*128-66-5*], a dye commercialized as Indanthren Goldgelb GK [201].

If naphthalene reacts with diacyl compounds, the 1,8-diacylated products often form [202]:

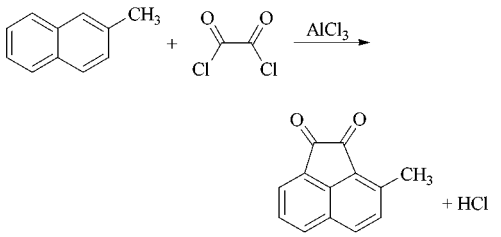

Acylation of Biphenyls. If one ring of a biphenyl is deactivated by an electron-withdrawing substituent, the other ring can be acylated in the 4 position [203]:

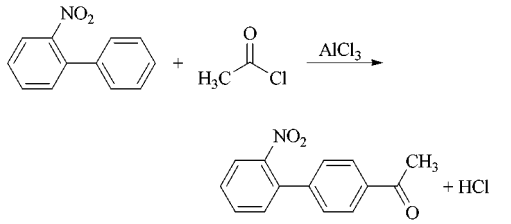

Acylation of Anthracene [*120-12-7*]. In the presence of $AlCl_3$ and under mild conditions in nonpolar solvents (0 °C, benzene), anthracene is acylated at the most reactive 9 position. Isomerization of the ketone – $AlCl_3$ complex is prevented by precipitation. If polar solvents, such as nitrobenzene, are used, a mixture, consisting mainly of 1-acylanthracene and some 2-acylanthracene is formed [204]. Also in this case, 9-acylanthracene is probably the primary product, which is then isomerized [148]:

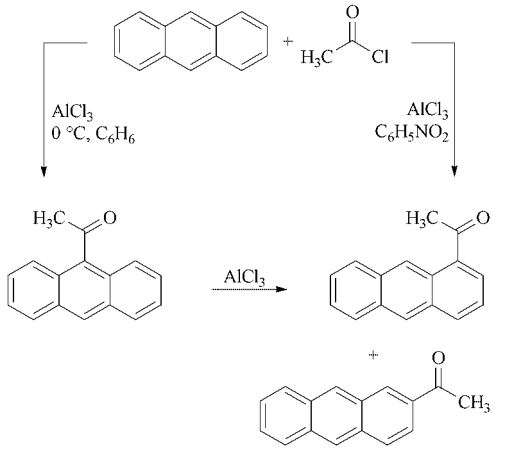

Acylation of Phenanthrene. Similarly, acylation of phenanthrene [*85-01-8*] in nitrobenzene yields mainly the 2- and 3-acyl isomers, whereas use of the less polar solvent dichloroethane gives the 9-acyl isomer as the main product [147], [204].

2.2.3. Acylation of Heteroaromatic Compounds

Because of their high electron density, furan, thiophene, and pyrrole are highly activated in

electrophilic substitution reactions and thus can be acylated under mild conditions.

Acylation of furan leads to high yields of the 2-acyl derivative, especially if anhydrides are used as acylating agents:

The high reactivity of furan [*110-00-9*] allows the use of benzene as solvent. Zinc chloride, boron trifluoride, and phosphoric acid also are suitable catalysts [205]. Acylation also can occur in the 5 position.

Acylation of Thiophene. The reactivity of thiophene [*110-02-1*] can be compared with that of anisole; the compound can be acylated in high yields utilizing acyl halides in the presence of $SnCl_4$ or anhydrides in the presence of phosphoric acid or $ZnCl_2$ as the acylating agent [206].

Acylation of pyrrole [*109-97-7*] may occur even in the absence of catalysts, although $ZnCl_2$, BF_3 , and $AlCl_3$ often are used with anhydrides as acylating agents [207]. Acylation usually occurs in the 2 position but 2,5-diacylation also is possible. 3-Acylation is observed only if the 2 and 5 positions are blocked.

Acylation of Imidazoles. Highly deactivated heterocycles, which can not be acylated in homogenous liquid phase like imidazole or pyrazoles react on Na-Y-zeolites at temperatures of more than 500 °C [208]:

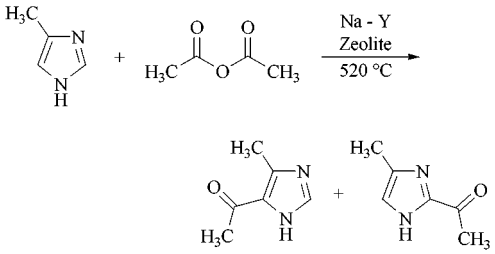

Acylation of Benzofurans. With very reactive heterocycles acylations can be carried out using zeolites in a continuous liquid-phase process, for example the acylation of 2-methylbenzofuran with acetic anhydride [209]:

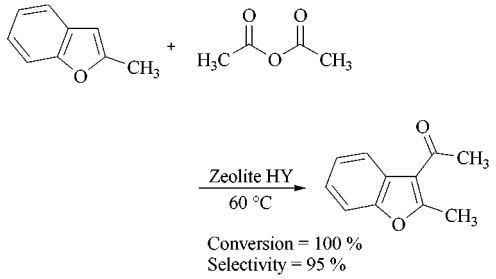

Acylation of Dibenzofuran [*132-64-9*], **Dibenzothiophen** [*132-65-0*], etc. Acylation or diacylation occurs in the 2 or 8 position (3 or 6 position in *carbazole* [*86-74-8*]) [210], [211]:

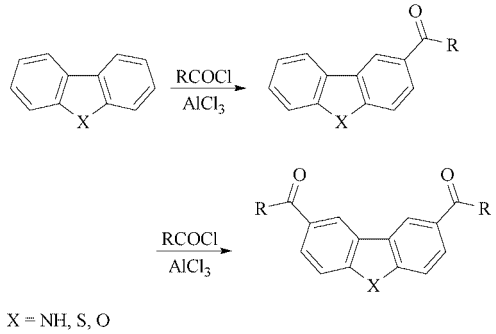

Miscellaneous. Substituted *4* H-*pyrones* have been acylated in excellent yields using trifluoroacetic acid as the catalyst. Even diacylation is possible [212]:

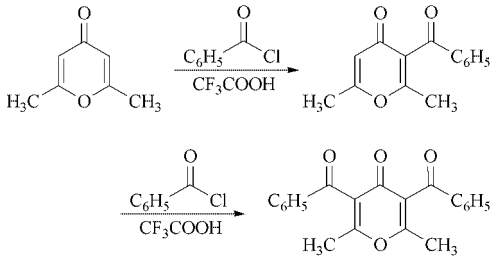

A large variety of further heteroaromatic compounds, such as chroman, xanthenes, 1,3-benzodioxole, and dibenzodioxins, have been acylated; for a review see [147].

2.2.4. Acylation of Nonbenzenoid Aromatic Compounds

Ferrocene [*102-54-5*] can be acylated readily by acyl halides in the presence of aluminum chloride, and its reactivity is comparable to that of phenols [213]. Besides acyl halides, anhydrides may be used as the acylating agent. Monoacylation of ferrocene is possible with BF_3 as catalyst [214]:

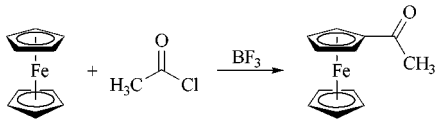

With an excess of acylating agent and of $AlCl_3$, heteroannular diacetylation occurs almost exclusively [215].

Other cyclopentadienyl derivatives also can be acylated and the reactivity of some aromatic substrates has been found to decrease in the following order [215], [216]:

$$C_6H_5OH \approx (C_5H_5)_2Fe > C_6H_5OCH_3 > CH_3C_5H_4Mn(CO)_3 > C_5H_5Mn(CO)_3 > C_6H_6$$

Azulene [*275-51-4*] reacts smoothly with acetic anhydride in the presence of $SnCl_4$ as catalyst; the cyclopentadienyl part is acylated, giving 1-acetylazulene. Under more drastic conditions and in the presence of $AlCl_3$ or $SnCl_4$, 1,3-diacetylazulene is formed using acetyl chloride or acetic anhydride as the acylating agent [217], [218].

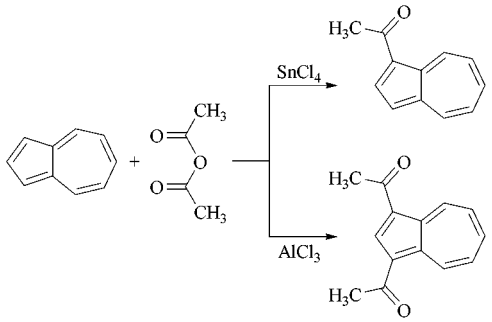

2.2.5. Miscellaneous Acylation Reactions

Among the various reactions leading to aromatic C-acylation products, the formylation reactions, the Houben – Hoesch synthesis, and the Fries rearrangement are of importance.

Formylation of aromatic compounds is simply a special case of acylation and occurs when a formyl halide or formic anhydride reacts with an aromatic compound in the presence of a Friedel – Crafts catalyst, such as $AlCl_3$. However, the scope of this reaction is limited because of the instability of most formylating agents.

Gattermann – Koch Synthesis. A mixture of hydrogen chloride and carbon monoxide in the presence of $AlCl_3$ and copper (I) chloride is used as formylating agent for aromatic compounds [219]:

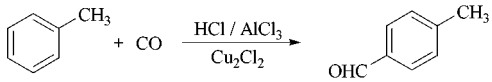

Benzene as well as substituted benzenes or polycyclic aromatics react in the Gattermann – Koch synthesis to give aromatic aldehydes. The *para* isomers are formed preferentially, and aromatic compounds with *meta*-directing substituents do not react in this formylation [220].

A modified Gattermann – Koch synthesis is being considered by Mitsubishi Gas Chemical for the industrial synthesis of *terephthalic acid.* Toluene reacts with CO and $HF-BF_3$ to give *p*-tolualdehyde in 96 % yield, which is then oxidized to terephthalic acid [221], [222].

More recent studies suggest that the acylating agent in these processes is in fact $[CHO]^+$, which is formed by protonation of CO or its complexes. An equimolar CO – HCl mixture as required for this reaction can be prepared as follows:

$$HCOOH + ClSO_3H \longrightarrow H_2SO_4 + CO + HCl$$

$$HCOOH + C_6H_5COCl \longrightarrow C_6H_5COOH + CO + HCl$$

Vilsmeier Aldehyde Synthesis [223]. Aromatic and heterocyclic substrates are formylated by reaction with dialkyl- or arylalkylformamides in the presence of $POCl_3$:

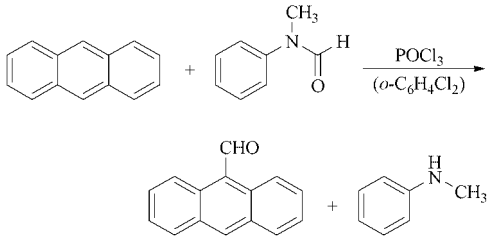

Besides anthracene, pyrene, acenaphthene, and other reactive polynuclear aromatic compounds, phenol ethers, dialkylamines, and *N*-heterocycles, such as pyrrole and indole, can be formylated. The regioselectivity of formylation follows the general rules of Friedel – Crafts reactions and leads usually to the 4-substituted aromatic compounds. Aromatic substrates also can be formylated by use of dichloromethyl methyl ether or trialkyl orthoformate in the presence of $TiCl_4$ or $AlCl_3$, respectively [224].

Houben – Hoesch Synthesis. This synthesis, which is related to the Gattermann reaction, uses nitriles to acylate reactive aromatic substrates, such as phenols, phenyl ethers, pyrroles, thiophenes, and indoles. The reactions are carried out in the presence of HCl and $ZnCl_2$ [225], [226]:

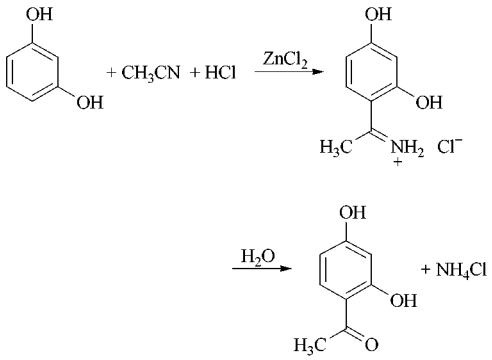

The active electrophile is the imminium ion, formed by addition of HCl to the nitrile and complexed by the Lewis acid. This bulky intermediate leads almost exclusively to acylation in the 4 position and not in the 2 position. Compared to the Friedel – Crafts acylation with acyl halides or anhydrides, the Hoesch synthesis is often more straightforward in the acylation of activated phenols.

Fries Rearrangement. For the synthesis of fine chemicals, the Fries rearrangement of aryl esters, which can be prepared by the acylation of phenols, is a useful method for making phenolic ketones:

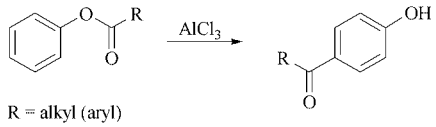

The production of pharmaceuticals, such as adrenalon, synephrin, buphenin, and acebutolol, involves Fries rearrangements. Although the reaction tends to give mixtures of the ortho- and para-acylated phenol, the selective formation of each of the isomers is possible by controlling the reaction conditions [227]:

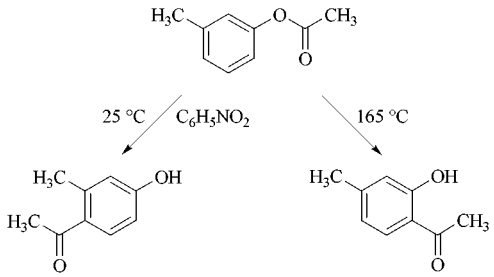

As a rule, para substitution is favored at low temperatures and ortho substitution at high temperatures, but regioselectivity also is affected by solvents and the nature and amount of catalysts.

Numerous substrates that can be substituted both in the phenol part and the acid part are reactive in this rearrangement. Even phenol derivatives containing electron-withdrawing substituents, such as the nitro group, react, as well as naphthol and hydroxy- or dihydroxyphenyl esters [228].

N-Acylanilines also can be rearranged to give 4-acylanilines at elevated temperatures [229]:

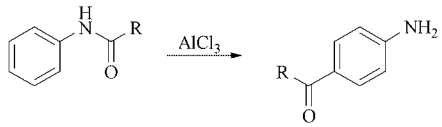

Similarly, the acid part can be varied and all types of alkyl or aryl acids have been used including some diaryl dicarboxyl esters. Sulfonic

acid esters also have been rearranged to give 4-hydroxyphenyl sulfones [230].

Usually, $AlCl_3$ is used as the catalyst, but $SnCl_4$, $TiCl_4$, $FeCl_3$, and $ZnCl_2$ are also active and in some cases more convenient. The rearrangement can be carried out by heating a mixture of the phenolic esters and the catalyst to 80 – 180 °C. Depending on the reactivity of the substrate, the reaction is complete within a few minutes or up to 3 or 4 h. The reaction can be carried out also at lower temperatures in a solvent, such as $C_6H_6NO_2$, $(CH_2Cl)_2$, CS_2, C_6H_5Cl, or petroleum ether. Fairly long reaction times must be used in this case. Another possibility is to start the rearrangement in CS_2, which is then separated by distillation, and to complete the reaction at higher temperatures [231].

The ortho isomer often can be separated from the product mixtures by steam distillation. Volatile products can be fractionally distilled. The para isomer can be separated from nonvolatile product mixtures by crystallization. A more detailed literature survey can be taken from [228, 232, 233].

3. Acylation and Alkylation of Aliphatic Hydrocarbons

Aliphatic hydrocarbons can be both alkylated and acylated, although industrially, alkylation is by far the more important reaction. Because of their nucleophilicity, reactions must be base or acid catalyzed. Base catalysis involves the formation of an organometallic derivative, whereas acidic catalysts lead to carbenium ions as active intermediates.

Acylation and Alkylation of Paraffins. Not very much work has been done on acylation and alkylation of pure paraffins; such reactions are known but never have gained much interest in industry for synthetic purposes [234]. However, the alkylation of toluene with propene to yield isobutylbenzene is carried out commercially at a multithousand t/a scale. Isobutylbenzene [*538-93-2*] is an intermediate for ibuprofen, a widely used anti-inflammatory agent:

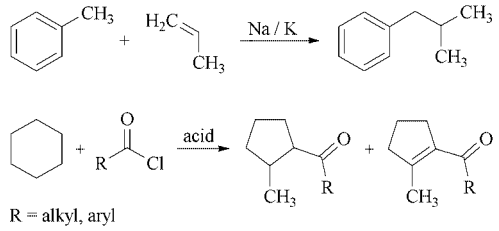

In the presence of very strong acids, such as $HF-TaF_5$, alkanes also can be self-alkylated [235]. The reaction between a paraffin and an olefin [64] to produce a higher paraffin is thermodynamically possible at low temperatures and is exothermic. However, rate and selectivity are low, and side reactions occur. Good results are obtained if highly branched paraffins are introduced. The best known example is the alkylation of isobutane with olefins using acids as the catalyst. This reaction is used on a huge scale by the petroleum industry to make highly branched paraffinic hydrocarbons, which are excellent antiknock components for motor fuels (→ Automotive Fuels, Chap. 4.1.; → Oil Refining, Chap. 3.7.1.). The industrial processes in operation today use large quantities of liquid Brønstedt acids (HF, sulfuric acid). The handling of large quantities of these acids has created environmental and safety concerns. As an alternative, a fixed-bed alkylation process with a supported liquid phase (SLP) catalyst has been developed to the pilot plant stage [236], [237].

Alkylation and Acylation of Olefins. Formally, all oligomerizations and polymerizations of olefins can be considered as alkylation. Table 11 summarizes a few of these reactions. There are some similarities in reaction mechanism between Ziegler – Natta-type catalysts and acylation – alkylation catalysts.

Acylation of olefins may be considered a Friedel – Crafts reaction. The reaction usually occurs with derivatives of carboxylic acids in the presence of an acid catalyst. The Lewis acids, whose activities follow the order $AlCl_3 > BF_3 > SnCl_4 > ZnCl_2$, require anhydrous conditions and temperatures ranging from 0 to 100 °C. At elevated temperatures interfering side reactions can occur.

Table 11. Alkylation of olefins

Olefin	Catalyst	Product
α-Olefin (ethylene, propylene)	Ziegler-Natta ($AlR_3 - TiX_4$)	polymers
Ethylene	Ni chelate	α-olefins
1-Decene	BF_3	lubricants
	$NiX_2 - AlX_nR_m$	dimers (motor fuel, plasticizer, detergents)
Isobutene	H^+	dimers (motor fuel, solvents)
Propene	AlR_3	dimers, oligomers (isoprene, detergents)
Ethylene – butadiene	Rh complexes	1,4-hexadiene

The acylation of olefins with acid chlorides in the presence of aluminum chloride leads to β-haloketones and β-unsaturated ketones [238]:

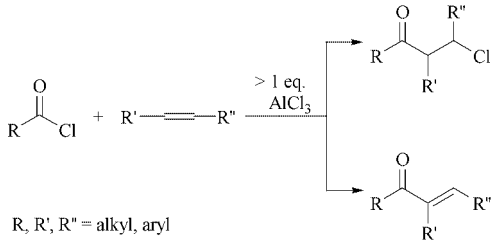

Acylations of olefins have been commercially applied only for the synthesis of fine chemicals, such as the unsaturated ketone (**2**), Vertofix Coeur [*80449-58-7*], (International Flavors & Fragrances, Inc., USA), used in perfume:

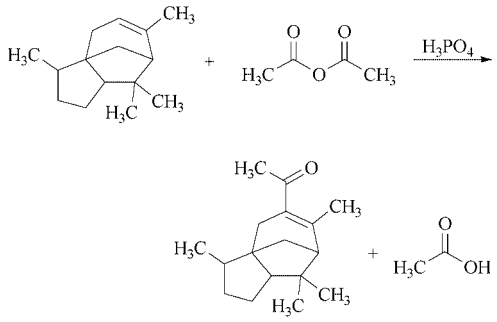

2

4. *N*-Alkylation and *N*-Acylation

An alkyl or acyl group can be attached to the nitrogen atom of an amine in many ways. The particular method chosen depends on the nature of the alkyl or acyl group and the amine used. The products obtained are used mainly as intermediates in the production of detergents, dyes, plasticizers, pharmaceuticals, agricultural chemicals, antioxidants, and vulcanization accelerators.

4.1. *N*-Alkylation

Alcohols, alkyl halides, olefins, and carbonyl compounds react with ammonia or amines to give alkyl derivatives [64]. Further alkylation of a tertiary amine leads to formation of quaternary ammonium compounds, which have found broad application in detergents. For instance, selected amines are alkylated with 1,3-propane sulfonate or epichlorohydrin sulfonate, giving sulfobetaines:

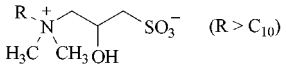

Another derivative typical of this class is cetyltrimethylammonium chloride, made by exhaustive methylation of the primary amine with methyl chloride (→ Surfactants, Chap. 8.1.).

Among the best known "quats" are benzyldimethylalkylammonium chlorides (alkyl, e.g., dodecyl). These are made by methylating the primary amine derived from coconut oil to give the dimethylalkylamine, which is then quaternized by addition of benzyl chloride. The products are marketed as Zephirol, Roccal, Ammonyx, and Triton K 60. These compounds are used widely as disinfectants, antiseptics, and fabric softeners.

4.1.1. *N*-Alkylation by Alcohols or Ethers

Alkylations of amines with alcohols or ethers are of commercial interest. These alkylations are carried out at temperatures up to 300 °C and pressures up to 100 bar. Generally, acidic catalysts, such as HCl, H_2SO_4, BF_3, PCl_3, and H_3PO_4, are used in the liquid phase. In the gas

phase, Al_2O_3, $Al_2O_3 \cdot SiO_2$, or dehydrogenation catalysts are applied. The liquid-phase operation is hampered by severe corrosion problems. The reaction can be carried out in batch or continuous-flow operations. Mixtures of the possible NH-exchange reaction products are obtained. Some control is possible by varying the ratio of reactants and by recycling undesired compounds.

Methylamine, dimethylamine, and trimethylamine [*74-89-5*], [*124-40-3*], [*75-50-3*] are made by the reaction of methanol with ammonia in a continuous-flow system in the presence of a dehydration catalyst, such as aluminum silicate or phosphate, at 350 – 400 °C. All three possible methylamines are formed, but the product distributions can be controlled by reaction conditions (excess NH_3, addition of H_2O) and by recycling the trimethylamine (→ Methylamines).

Higher aliphatic amines, ethyl through pentyl, can be made by processes similar to that described for methylamines. Furthermore, such a reaction can be applied to make ethylenediamine from ethylene glycol.

Aromatic amines can be produced by this route using an acid as catalyst. However, starting from a secondary amine, it is not possible to make monoalkyl derivatives exclusively.

The alkylation of aromatic amines can be used for an efficient synthesis of *N*-methylaniline [*100-61-8*], *N,N*-dimethylaniline [*121-69-7*], *N,N*-diethylaniline [*91-66-7*], *N*-benzyl-*N*-ethyl-*m*-toluidine [*119-94-8*], *m*-chloro-*N*-ethylaniline [*15258-44-3*], *N*-benzylaniline [*103-32-2*], and other aromatic amines, many of which function as dye precursors. For instance, *N,N*-dimethylaniline [*121-69-7*] has been produced on a large scale since 1870 by the alkylation of aniline with methanol. Dimethyl ether can be used instead of methanol, as is known from a BASF process.

4.1.2. *N*-Alkylation by Alkyl Halides

The reaction of an alkyl halide with ammonia or an amine has long been used, particularly in the laboratory, to prepare alkylamines, although industrial processes also are based on this route. For example, pentylamines are made by mixing a solution of ammonia in alcohol with pentyl chlorides. By a similar reaction, ethylenediamine is produced from ammonia and 1,2-dichloroethane. No catalyst is required here.

The reaction of an amine with an alkyl halide is a versatile method for producing a variety of fine chemicals. Examples are the synthesis of sodium *N*-methylglycinate, which is used in the production of compounds (**3**) employed in the textile and cosmetic industries or as corrosion inhibitors, and that of *p-N*-methylaminophenol (**4**), used in the photographic industry.

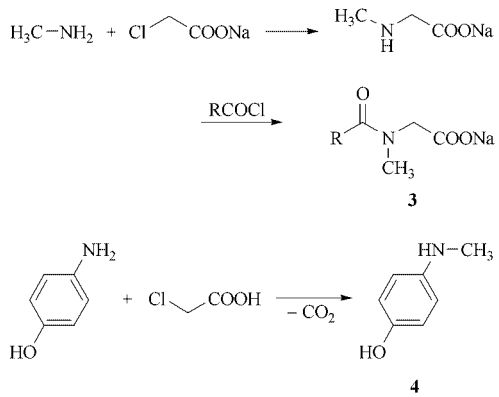

4.1.3. *N*-Alkylation by Olefins

Alkylamines can be prepared by the direct addition of an olefin to ammonia or an amine using alkali metals as catalysts. This method works well with ethylene. For example, mono- and diethylanilines are prepared industrially from ethylene and aniline using alkali metals as catalysts.

Higher olefins also react, but amine yields are lower.

4.1.4. *N*-Alkylation by Carbonyl Compounds (Reductive Alkylation)

Alkyl derivatives can be obtained by the reaction of either ammonia or an amine with a carbonyl compound, such as an aldehyde or a ketone, in the presence of a suitable hydrogenation catalyst and hydrogen [239].

$$RCHO + R'NH_2 + H_2 \longrightarrow RCH_2NHR' + H_2O$$

Reductive alkylation is used commercially, for instance, to introduce branched alkyl groups (R′, R″).

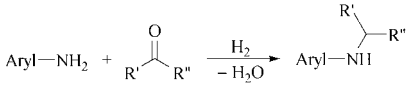

Antioxidants, such as *N,N'*-di-*sec*-butyl-*p*-phenylenediamine [*101-96-2*], are synthesized by this route [240]:

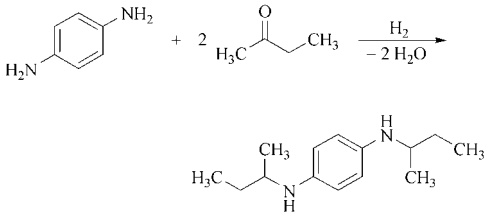

A further example of importance to the dye industry is the synthesis of anilinoethanenitrile (**5**), which is an intermediate in the production of indoxyl (**6**) [*480-93-3*]:

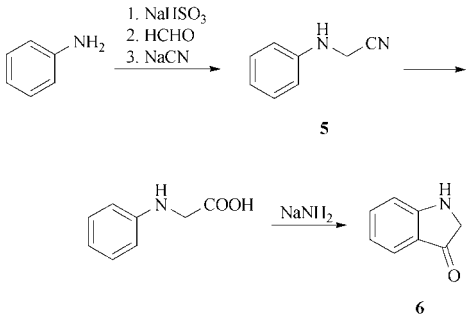

4.2. *N*-Acylation

The acylation of NH groups occurs quite readily. The most common acylating agents are acyl halides, esters, anhydrides, and carboxylic acids.

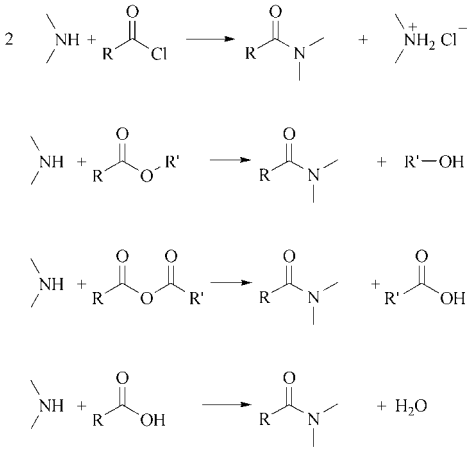

R, R' = alkyl, aryl

Besides amines, many other compounds containing NH groups can be acylated. At high enough temperatures, even diacylation (for example, with primary amines) is possible. However, secondary amines react much more slowly than primary ones.

Acylation at the nitrogen atom often proceeds satisfactorily without a catalyst. This is especially the case for highly nucleophilic NH derivatives with electron-releasing substituents. Common catalytic procedures, if used, involve alkoxides (with esters as alkylating agents), sulfuric or phosphoric acid (with anhydrides or carboxylic acids), and aqueous alkali (with aroyl halides, Schotten – Baumann reaction).

Acylation is frequently used in the laboratory to separate tertiary from primary and secondary amines (Hinsberg reaction). After acylation of primary and secondary amines, the tertiary amines, which cannot be acylated, can be distilled.

Acylating Agents. The use of *carboxylic acids* requires the removal of water from the reaction mixture. This can be done by distillation in the presence of xylene, toluene, or chlorobenzenes. Sometimes catalysts, such as PCl_3, P_2O_5, or phosphites, are used.

Acyl halides often are preferred as acylating agents because they are easily prepared. An inert solvent often is used to precipitate the products. After filtration, the amine salt can be removed from the solid by washing with water. It is noteworthy that in this method half the amine is lost via salt formation. The HCl can be removed by Na_2CO_3, alkali (Schotten – Baumann reaction), or by adding other N bases. Among *anhydrides,* acetic anhydride plays an important role, but for economic reasons, industry prefers to use acids. *Esters* also have found wide application. They are especially suitable if the alcohols formed boil at lower temperatures than the N derivatives, thus removing the alcohol from the equilibrium.

Other acylating agents in use include aldehydes, ketones, ketenes, thiol esters, silicon tetraesters, carbamates, trihaloketones and aldehydes, acyl azides, amides, imides, and hydrazides.

***N*-Acylated products** and intermediates have found many applications, especially in pharmaceuticals, agrochemicals, dyes, and

polymers. N-*Acetylsalicylamide* [*487-48-9*] (**7**) and *p*-ethoxyacetanilide (*phenacetine*) [*62-44-2*] (**8**) are two representative examples of pharmaceuticals synthesized by N-acylation [241].

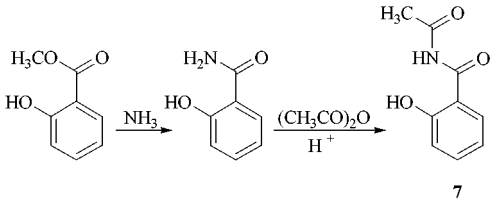

7

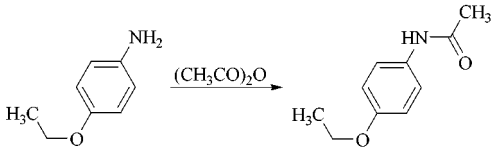

8

Amides of carboxylic acids with such generic names as propanil, monalide, chloranocryl, and allidochlor are used as herbicides [242]. Their synthesis is based on acylation of amines. *N*-Acylcarbamates, which can be synthesized by NH-acylation, are used as insecticides.

Two examples of dye precursors are *N*-(2-carboxyphenyl)glycine (**9**) and so-called "arylide" compounds (**10**):

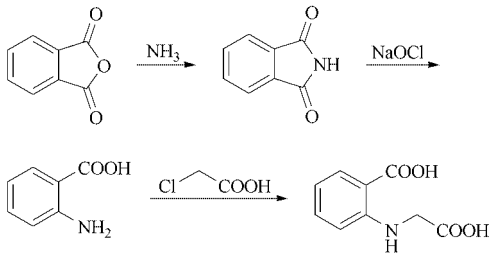

9

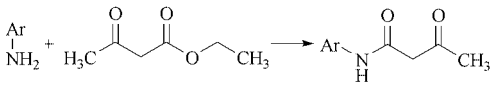

10

N-Acylation to provide a protective group.

The primary NH group in aniline derivatives can be masked by acylating the compound prior to chlorination or nitration. For this purpose, acetanilide, *N*-acetyl-*o*/*p*-toluidine, *N*-acetyl-*o*/*p*-anisidine, *N*-acetyl-*p*-phenetidine, and *N*-acetyl-*p*-aminophenol are prepared as intermediates. The acetyl group can be removed by heating in NaOH. The following reaction is used in the pharmaceutical industry to synthesize *sulfonamides* [241]:

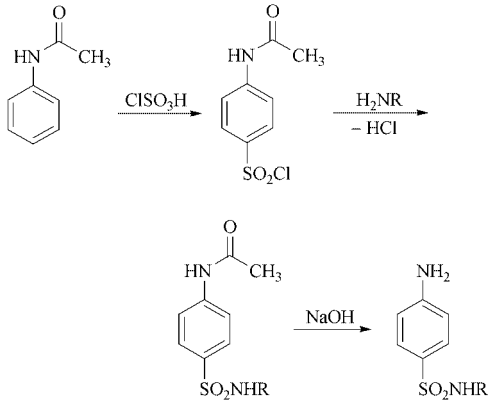

Protection of amine functions is also important in peptide syntheses.

5. *O*-Alkylation and *O*-Acylation

The following types of reactions can be considered:

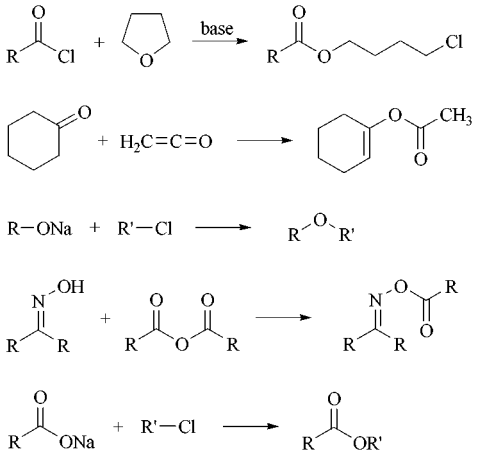

Reactions in which the hydroxyl group is attacked are very wide ranging. However, those alkylations or acylations in which the OH group is attached to a carbon atom are of greatest industrial importance. Table 12 enumerates reactions

and reaction products involving alkylation and acylation of hydroxyl groups attached to carbon.

Table 12. Products of alkylation and acylation of alcohols and carboxylic acids

Alkylation/acylation agents	ROH	RCOOH
RCOX ∗, $(RCO)_2O$	esters	anhydrides
RX ∗	ether	esters
Olefin	ether	esters
∗ X=halide.		

Besides acyl halides, anhydrides, alkyl halides, and olefins, a variety of other alkylation and acylation agents are utilized. Although of minor industrial importance, alkylation with dialkyl sulfates or diazo compounds is a common laboratory method for preparing ethers and esters [243]. The commercial synthesis of *N*-acylcarbamates is based also on an O-alkylation:

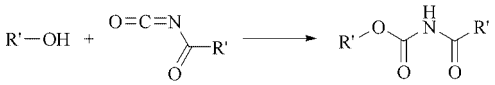

5.1. Synthesis of Esters

Esters can be prepared by three routes: acylation of alcohols, reaction of acids with alkyl or aryl halides, and reaction of acids with olefins (see Table 12). The most readily available acylating agents are selected, together with appropriate catalysts where required. For acyl halides, such bases as pyridine or aqueous alkali are the preferred catalysts; for anhydrides, sodium acetate, sulfuric acid, and BF_3 are preferred. Many fine chemicals, such as pharmaceuticals, are made by this route. The synthesis of *acetylsalicylic acid* (aspirin) using acetic anhydride is an example:

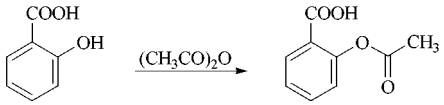

Cellulose Acetate. An ester process of great industrial importance is the synthesis of cellulose acetate for synthetic fibers. Cellulose is added to equal amounts of acetic anhydride and glacial acetic acid, together with a small amount of sulfuric acid as catalyst. The acylation temperature is kept below 30 °C. After 5 – 8 h, water is added to the cellulose triacetate. The hydrolysis is stopped when the required degree of acetylation is reached by running the mixture into a large volume of water followed by centrifuging. For further details, → Cellulose Esters, Chap. 2.1.3.

Direct Esterification Using Olefins or Acetylene. Esters are formed also by alkylation of acids by olefins, but this method often is limited by low catalyst efficiency. Olefins ranging from ethylene to octenes and higher olefins can be used, together with such catalysts as BF_3, $BF_3 \cdot ROR$ (R = alkyl, aryl), or $BF_3 - HF$. Polymerization of reactive olefins often limits this application. The direct esterification of organic acids with olefins has distinct advantages from a processing standpoint over the two-step process in which the olefin is first hydrated to an alcohol with the acid.

Some of these processes are in industrial use. The reaction of α-pinene with carboxylic acids yields bornyl esters that are intermediates in camphor chemistry. Of industrial importance are also the vinyl esters of higher carboxylic acids (fatty acids, Koch acids) which are used as comonomers in polymerizations of vinyl chloride, vinyl acetate, acrylonitrile, styrene, and other compounds. The process is based on the reaction of the acids with acetylene in the liquid phase at 150 – 170 °C at 10 – 15 bar pressure.

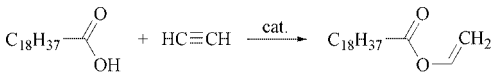

As catalysts, zinc salts or $H_2O - H_2SO_4$ can be used.

5.2. Synthesis of Anhydrides

The acetylation of a carboxylic acid by RCOX yields an anhydride. By this route, unsymmetrical anhydrides can be prepared. Besides acyl halides, amides, esters, and acids, ketenes also can be used. The ketene reaction is of importance for the commercial synthesis of acetic anhydride and a convenient laboratory route to acetic carboxylic anhydrides:

$$RCOOH + CH_2{=}C{=}O \longrightarrow RCOOCOCH_3$$

5.3. Synthesis of Ethers

Ethers can be formed by the alkylation of a metal salt of an alcohol with an alkyl halide (*Williamson synthesis*). The reaction proceeds cleanly with active primary halides and often also with secondary halides. Tertiary alkyl halides are subject to an elimination reaction and are therefore unsuitable for the preparation of tertiary alkyl ethers. The alkyl chlorides are less reactive than the bromides or iodides, but for economic reasons the chlorides are favored for large-scale operations.

Ethers are produced in industry by the reaction of an olefin with an alcohol. This reaction is catalyzed by such acid catalysts as H_2SO_4, ion-exchange resins, phosphoric acids, or zeolites. The reactivity of the olefin decreases with increasing molecular weight: Isobutene reacts more readily than 2-methyl-2-butene. Best results are obtained with olefins containing tertiary carbon atoms because secondary olefins require higher temperatures and catalyst concentrations. Primary alcohols are more suitable than secondary alcohols for this reaction.

Methyl *tert*-butyl ether [*1634-04-4*] (MTBE), an ether of significant commercial interest, is used as an additive in gasoline. It is produced by the reaction of isobutene with methanol, applying an acidic catalyst, such as an anion-exchange resin. Several processes have been developed by Chemische Werke Hüls (now Degussa-Hüls), SNAM Progetti, Atlantic Richfield Co. (Arco), and Sun Oil Co. World capacity is estimated at 12×10^6 t/a for 1992 and 16×10^6 t/a for 2000. The C_4 fraction of naphtha crackers, which is free from butadiene (raffinate I), can be used directly as olefin feed. A liquid-phase system operating at $<100\,°C$ is used (→ Methyl *tert*-Butyl Ether).

Vinyl Ethers. The industrial processes for vinyl ether are based on the reaction of alcohols with acetylene:

$$ROH + HC{\equiv}CH \longrightarrow ROCH{=}CH_2$$

The reaction is carried out in the liquid phase at 120 – 180 °C with alkali hydroxides as catalysts. Low-boiling alcohols require the use of pressures up to 20 bar. Vinyl ethers are used for special polymers, coatings, and adhesives. Of industrial interest are methyl, ethyl, isopropyl, butyl, isobutyl, and 2-ethylhexyl vinyl ethers (→ Vinyl Ethers).

6. *CH*-Alkylation and *CH*-Acylation of Nonaromatic Compounds

Carbon-hydrogen bonds adjacent to electron-attracting groups such as nitro, carbonyl, cyano, sulfone, or phenyl can be C-alkylated or C-acylated [244–246]. The acidity of the C-H bond depends on the nature of the activating group. Procedures that involve the formation and subsequent reaction of anions derived from active CH compounds constitute a very important and synthetically useful class of organic reactions. Examples are the Dieckmann reaction, the Stobbe condensation, the Michael addition, the Claisen condensation, and the Thorpe reaction.

CH-Alkylation and -acylations have been used widely in the laboratory and for the synthesis of many fine chemicals. Generally, a carbanion is formed by hydrogen atom exchange with a base followed by reaction with the alkylation or acylation agent. Bases used are, e.g., sodium alkoxide and sodium amides.

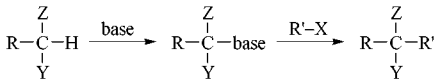

R, R' = alkyl, aryl; Y, Z = electron-withdrawing groups; X = halide

7. Synthesis of Metal Alkyls

The direct alkylation of elements by alkyl halides is frequently the first stage in the synthesis of a variety of organometallic compounds. Especially the main-group metals yield stable derivatives that have found many industrial applications.

In 1900, Grignard discovered that the reaction of magnesium with alkyl halides leads to alkyl – magnesium – halogen compounds, RMgX. Today, Grignard compounds are standard laboratory chemicals.

$$RX + Mg \longrightarrow RMgX$$

Organolithium compounds can be prepared by the direct reaction of alkyl halides with lithium. Large quantities of butyllithium, which is used primarily as an initiator in the polymerization of dienes, is produced by this method.

The scope of organometallic compounds for CH-alkylations is broad and has found many applications, especially in the laboratory. Examples are the following well-known reactions:

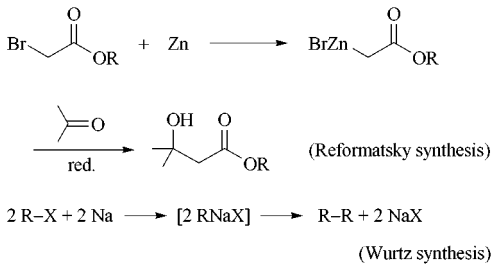

Organometallic compounds also can be prepared by reacting metal hydrides with olefins.

$$AlH_3 + 3\,H_2C{=}CH_2 \longrightarrow AlEt_3$$

Numerous alkylations by transition metals have become known since the 1960s [40] and are gradually becoming accepted in chemical industry, especially for the synthesis of fine chemicals.

Examples of elemental acylation are rather rare and are of no practical importance as yet. However, some interest, especially among fine chemicals, is developing around the use of R–CO–M compounds to control regio- and stereoselectivity in acylations [247].

Aluminum Alkyls [248]. Alkylaluminum derivatives have gained considerable importance since 1953 when ZIEGLER reported their catalytic activity for polymerization and means of their preparation. Their consumption in the USA, Western Europe and Japan was estimated at 30 000 t in 1996.

The method of preparation depends on the particular alkyl derivative but generally involves one of the following alkylation reactions:

$$Al_2Mg_3 + 6\,RCl \longrightarrow 2\,AlR_3 + 3\,MgCl_2$$

$$AlH_3 + 3\,\text{olefins} \longrightarrow Al(\text{alkyl})_3$$

$$2\,Al + 3\,RX \longrightarrow RAlX_2 + R_2AlX$$

The most important derivative is $Al(C_2H_5)_3$ [*97-93-8*], the synthesis of which is based on the reaction of $Al(C_2H_5)_3$ with aluminum, hydrogen, and ethylene. Aluminum alkyls are used for the preparation of Ziegler – Natta-type catalysts, in the polymerization of olefins, and in the alfen and alfol processes for the production of α-olefins and alcohols, respectively.

Aluminum alkyls also are used industrially to synthesize other organometallic compounds, for instance R_4Sn:

$$3\,SnCl_4 + 4\,R_3Al \longrightarrow 3\,R_4Sn + 4\,AlCl_3$$

Organic Silicon Compounds. The direct reaction of silicon with alkyl or aryl halides (*Rochow reaction*) is used for the industrial synthesis of organosilicons:

$$2\,RX + Si \longrightarrow R_2SiX_2$$

The process takes place at high temperature in the presence of a catalyst, such as Cu, Ag, or Zn. Of the silicon derivatives produced, $(CH_3)_2SiCl_2$ [*75-78-5*] dominates the market. It is made by the reaction of silicon with CH_3Cl at 260 – 320 °C, catalyzed by copper. Besides $(CH_3)_2SiCl_2$ (70 – 90 % of the products), all other possible $(CH_3)_xSiCl_y$ compounds are formed in this reaction. The heterogeneous gas-solid reaction can be carried out in a fixed-bed or fluidized-bed reactor. For further details, see [249] and → Silicon Compounds, Organic.

Two other silicon derivatives, $(C_6H_5)_2SiCl_2$ and $C_6H_5SiCl_3$, have some economic importance. Besides the direct synthesis, the addition of Si-H to olefins is used to make silicon derivatives.

Lead Alkyls. Tetraethyllead [*78-00-2*] and tetramethyllead [*75-74-1*] are added to motor fuels to improve the octane number. However, for environmental reasons their use is now limited or banned. In 1977, world production was ca. 700 000 t, ca. 500 000 t in 1980, and 60 000 t in 1998 (USA, Western Europe and Japan). The sole use is in aviation turbine fuels.

Tetraethyllead is made by the reaction of lead – sodium alloy with ethyl chloride.

$$4\,NaPb + 4\,C_2H_5Cl \longrightarrow Pb(C_2H_5)_4 + 3\,Pb + 4\,NaCl$$

A catalyst is not required, but to speed up the reaction, 1 – 2 % acetone is added. Yields of up to 90 % are obtained [250] (→ Lead Compounds). Tetramethyllead is made by a similar route.

8. References

General References

1. F. Asinger: *Chemistry and Technology,* Pergamon Press, Oxford 1968.
2. F. Asinger: *Die Petrochemische Industrie,* **vol. 2,** Akademie Verlag, Berlin (DDR) 1971, p. 1235.
3. J. E. Germain: *Catalytic Conversion of Hydrocarbons,* Academic Press, London 1969, p. 170.
4. *Winnacker-Küchler,* 4th ed., **vol. 5,** p. 219; **vol. 6,** p. 150, p. 267, p. 816.
5. F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart 1975.
6. G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 1 – 4,** J. Wiley & Sons, New York 1964.
7. G. A. Olah (ed.): *Friedel – Crafts Chemistry,* J. Wiley & Sons, New York 1973.
8. G. Stefanidakis, J. E. Gwyn: "Alkylation," in J. J. McKetta, W. A. Cunningham (eds.): *Encycl. Chem. Process. Des.* **2** (1977) 357.
9. C. E. Thomas: *Catalytic Processes and Proven Catalysts,* Academic Press, New York 1970.
10. *Ullmann,* 4th ed., **vol. 7,** p. 101, p. 236; **vol. 14,** p. 672; **vol. 15,** p. 61; **vol. 18,** p. 191; **vol. 22,** p. 455.
11. K. Weissermel, H. J. Arpe: *Industrial Organic Chemistry,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1994.
12. K. Tanabe, M. Misono, Y. Ono, H. Hattori, *Stud. Surf. Sci. Catal.* **51** (1989) 215.
13. G. A. Olah, S. Kobayashi, M. Tashiro, *J. Am. Chem. Soc.* **94** (1972) 7448.

Specific References

14. C. D. Nenitzescu, I. P. Cantuniari, *Chem. Ber.* **66 B** (1933) 1097.
15. Amoco Corp., EP 838 447, 1998 (W. S. Park); *Chem. Abstr.* **128**:296155.
16. K. R. Seddon, *J. Chem. Tech. Biotechnol.* **68** (1997) 351.
17. W. A. Pardee, B. F. Dodge, *Ind. Eng. Chem.* **35** (1943) 273.
18. D. A. McCaulay, A. P. Lien, *J. Am. Chem. Soc.* **74** (1952) 6246.
19. Esso Research and Engineering Co., US 3708553, 1973 (G. A. Olah); *Chem. Abstr.* **78** (1973) 123979 u.
20. R. Szostak: *Handbook of molecular sieves.* Van Nostrand Reinhold, New York, 1992; R. Szostak: *Molecular Sieves – Principles of Synthesis and Identification.* Van Nostrand Reinhold Catalysis Series, Van Nostrand Reinhold, New York, 1989.
21. C. T. O'Conner, E. van Steen, M. E. Dry, *Stud. Surf. Sci. Catal.* **102** (1996) 323.
22. K. Arata, *Adv. Catal.* **37** (1990) 165.
23. K. Tanabe, *Proc. Int. Congr. Catal. 9th, Calgary,* The Chemical Institute of Canada, Ottawa, 1988, vol. 5, p. 85.
24. K. Tanabe, M. Misono, Y. Ono, H. Hattori, *Stud. Surf. Sci. Catal.* **51** (1989) 5.
25. K. Tanabe, Solid Acid and Bases Catalysts, in J. R. Anderson, M. Boudart (eds.): *Catalysis – Science and Technology,* **vol. 2,** Springer, Berlin, 1981, p. 231.
26. J.-R. Butruille, T. J. Pinnavaia, *Clays Control. Environ., Proc. Int. Clay Conf. 10th,* CSIRO Publishing, Melbourne, 1995, p. 92.
27. Contract Chemicals, EP 352 878, 1990 (C. M. Brown, S. J. Barlow, D. J. McQuarrie, J. H. Clark, A. P. Kybett); *Chem. Abstr.* **112**:234978.
28. G. A. Tsigdinos, "Aspects Molybdenum Relat. Chem.", *Top. Curr. Chem.* **76** (1978) 1.
29. W. L. Nelson, *Oil Gas J.* **34** (1965) 104.
30. M. A. M. Boersma in J. L. Burton, R. L. Garten (eds.): *Advanced Materials in Catalysis,* Academic Press, New York 1977, p. 67.
31. B. Loev, J. T. Massengele, *J. Org. Chem.* **22** (1957) 988.
32. Dow Chemical Co., US 2711414, 1955 (T. R. Norton); *Chem. Abstr.* **50** (1956) 5759 d.
33. G. A. Olah, D. Meidar, *Nouv. J. Chim.* **3** (1979) 269.
34. G. A. Olah, D. Meidar, R. Malhotra, J. A. Olah, S. C. Narang, *J. Catal.* **61** (1980) 96.
35. M. A. Harmer, W. E. Farneth, Q. Sun, *J. Am. Chem. Soc.* **118** (1996) 7708.
36. U. Deschler, P. Kleinschmitt, P. Panster, *Angew. Chem.* **18** (1986) 237.
37. M. Mioque, M. D. d'Engenières, *Ann. Chim.* **9** (1964) 123.
38. G. Suzukamo, M. Fukao, T. Hibi, K. Tanaka, K. Chikaishi, *Acid-Base Catal., Proc. Int. Symp.,* Sapporo, VCH, 1988, p. 405.
39. D. Barthomeuf, *Catal. Rev.-Sci. Eng.* **38** (1996) 521.
40. L. G. Volkova, I. Y. Levitin, M. E. Volpin, *Russ. Chem. Rev. (Engl. Transl.)* **44** (1975) 552.
41. B. M. Trost, *Acc. Chem. Res.* **13** (1980) 385.

42. V. I. Minkin, G. N. Dorofeenko, *Russ. Chem. Rev. (Engl. Transl.)* **29** (1960), 599.
43. C. C. Price in R. Adams (ed.): *Organic Reactions,* **vol. 3,** J. Wiley & Sons, New York 1946, p. 1.
44. H. C. Brown, C. R. Smoot, *J. Am. Chem. Soc.* **78** (1956) 6255.
45. G. A. Olah, P. V. R. Schleyer (eds.): *Carbonium Ions,* J. Wiley & Sons, New York 1968 – 1972. D. Bethell, V. Gold: *Carbonium Ions, An Introduction,* Academic Press, New York 1967.
46. P. B. Venuto, *Micropor. Mater.* **2** (1994) 297.
47. G. Mirth, J. A. Lercher, *J. Phys. Chem.* **95** (1991) 3736.
48. N. Y. Chen. W. E. Garwood, *Catal. Rev.-Sci. Eng.* **28** (1986) 185.
49. P. H. Gore, *Chem. Ind.* (London) 1954, 1385.
50. C. E. Thomas: *Catalytic Processes and Proven Catalysts,* Academic Press, New York 1970, p. 87.
51. R. M. Roberts, Y. W. Han, C. H. Schmid, D. A. Davies, *J. Am. Chem. Soc.* **81** (1959) 640. C. D. Nenitzescu, *Experienta* **16** (1960) 332; *Chem. Abstr.* **55** (1961) 3474.
52. L. Schmerling, J. P. West, *J. Am. Chem. Soc.* **76** (1954) 1917.
53. J. E. Germain: *Catalytic Conversion of Hydrocarbons,* Academic Press, London 1969, p. 170.
54. A. C. Olson, *Ind. Eng. Chem.* **52** (1960) 833. V. N. Ipatieff, E. E. Meisinger, H. Pines, *J. Am. Chem. Soc.* **72** (1950) 2772.
55. V. N. Ipatieff, J. E. Germain, H. Pines, *J. Am. Chem. Soc.* **75** (1953) 6056.
56. S. V. Zavgorodnii, L. S. Shvetsova, *Zh. Obshch. Khim.* **28** (1958) 2668; *Chem. Abstr.* **53** (1959) 9104 a.
57. F. Asinger: *Die Petrolchemische Industrie,* **vol. 2,** Akademie Verlag, Berlin 1971, p. 1235.
58. G. A. Olah, S. H. Flood, S. J. Kuhn, M. E. Moffatt, N. A. Overchuck, *J. Am. Chem. Soc.* **86** (1964) 1046. G. A. Olah, S. H. Flood, M. E. Moffat, *J. Am. Chem. Soc.* **86** (1964) 1060, 1065.
59. O. Jacobsen, *Chem. Ber.* **18** (1885) 338.
60. Buu-Hoi, P. Cagniant, *Bull. Soc. Chim.* **9** (1942) 887.
61. A. W. Francis, E. E. Reid, *Ind. Eng. Chem.* **38** (1946) 1194.
62. A. P. Lien, D. A. McCaulay, *J. Am. Chem. Soc.* **75** (1953) 2407.
63. K. Weissermel, H.-J. Arpe: *Industrial Organic Chemistry,* 2nd ed., Verlag Chemie, Weinheim 1978, p. 293.
64. S. H. McAllister in B. T. Brooks, C. E. Boord, S. K. Stewart, Jr., L. Schmerling (eds.): *The Chemistry of Petroleum Hydrocarbons,* **vol. 3,** chap. 37, Reinhold, New York 1955, p. 579.
65. G. Stefanidakis, J. E. Gwyn: "Alkylation," in J. J. McKetta, W. A. Cunningham (eds.): *Encycl. Chem. Process. Des.* **2** (1977) 357.
66. V. N. Ipatieff, L. Schmerling, *Ind. Eng. Chem.* **38** (1946) 400.
67. H. W. Grote, C. F. Gerald, *Chem. Eng. Prog.* **56** (1960) 60.
68. H. W. Grote, *Oil Gas J.* **56** (1958) no. 13, 73.
69. A. V. Topchiev, S. V. Zavgorodnii, V. G. Kryuchkova: *Alkylation with Olefins,* Elsevier, Amsterdam 1964.
70. A. V. Topchiev, S. V. Zavgorodnii, Y. M. Paushkin: *Boron Fluoride and Its Compounds as Catalysts in Organic Chemistry,* Pergamon Press, New York 1959.
71. R. Lacko, M. Chmela, L. Hala, *Ropa Uhlie* **14** (1972) 301; *Chem. Abstr.* **77** (1972) 93252 m.
72. K. M. Minachev, E. S. Mortikov, A. S. Leont'ev, A. A. Masloboev-Shvedov, N. F. Kononov, *Izv. Akad. Nauk. SSSR Ser Khim.* 1971, 258; *Chem. Abstr.* **76** (1972) 99231 e.
73. *Winnacker-Küchler,* 4th ed., **vol. 5,** 1981, p. 219.
74. *Hydrocarbon Process.* March 1997 p. 127.
75. G. Belussi, G. Pazzuconi, C. Perego, G. Girotti, G. Terzoni, *J. Catal.* **157** (1995) 227.
76. A. A. O'Kelly, J. Kellett, J. Plucker, *Ind. Eng. Chem.* **39** (1947) 154.
77. G. T. Austin, *Chem. Eng.* **81** (1974) no. 6, 87; "Chemical Profile – Cumene," *Oil, Paint Drug Rep.* Jan. 1, 1970.
78. P. R. Pujado, J. R. Salazar, C. V. Berger, *Hydrocarbon Process.* March (1976) p. 91.
79. S. H. McAllister, J. Anderson, E. F. Bullard, *Chem. Eng. Prog.* **43** (1947) no. 4, 189.
80. *Europ. Chem. News,* October 21 – 27, 1996 p. 41.
81. *Hydrocarbon Process.* March 1997 p. 120.
82. *Hydrocarbon Process.* March 1995 p. 106.
83. J. D. Shoemaker, E. M. Jones, *Hydrocarbon Process.,* June 1987 p. 57.
84. Eni, EP 340 862, 1989 (F. Cavani, V. Arrigoni, G. Belussi); *Chem. Abstr.* **112**:138717.
85. Union Oil, WO 9 014 160, 1990 (M. West, S. F. Abdo); *Chem. Abstr.* **114**:104771.
86. Chemical Research & Licensing Company, US 4 849 569, 1989 (L. A. Smith, Jr); *Chem. Abstr.* **112**:22750.
87. J. P. Fortuin, M. J. Waale, R. P. van Oosten, *Pet. Refiner* **38** (1959) no. 6, 189.

88. California Research Corp., US 2477382, 1949 (A. H. Lewis, A. C. Ettling); *Chem. Abstr.* **44** (1950) 1137 c.
89. Universal Oil Products, US 5 196 574, 1993; *Chem. Abstr.* **118**:236451.
90. F. N. Baumgartner, *Ind. Eng. Chem.* **46** (1954) 1349.
91. *Ullmann,* 4th ed., **vol. 22,** p. 455.
92. G. A. Olah (ed.): *Friedel – Crafts Chemistry,* J. Wiley & Sons, New York 1973. G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 1 – 4,** J. Wiley & Sons, New York 1964.
93. Y. Lavaux, M. Lombard, *Bull. Soc. Chim. Fr.* **7** (1910) 913.
94. E. Bergmann, *Chem. Rev.* **29** (1941) 529.
95. M. J. Rosen: *Org. Synth.,* **coll. vol. 4,** J. Wiley & Sons, New York 1963, p. 665.
96. BASF, DE 2314695, 1974 (H. Engelbach, R. Krabetz, H. Wistuba, M. J. Sprague); *Chem. Abstr.* **81** (1974) 37431 a.
97. D. V. Nightingale, *Chem. Rev.* **25** (1939) 329. G. A. Olah, W. S. Tolgyesi, R. E. A. Dear, *J. Org. Chem.* **27** (1962) 3441.
98. G. A. Olah, W. S. Tolgyesi, *J. Org. Chem.* **26** (1961) 2053.
99. K. Ito, *Hydrocarbon Process.* **52** (1973) no. 8, 89.
100. N. Naum, P. Onu, N. Bilba, G. Mihaila, V. Ababi, *Rev. Chim. (Bucharest)* **33** (1982) 715; *Chem. Abstr.* **97** (1982) 188828 f.
101. S. H. Patinkin, B. S. Friedman in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 2,** part 1, J. Wiley & Sons, New York 1964, p. 1.
102. J. De Ment, *Pet. Refiner* **22** (1943) no. 8, 83.
103. *Winnacker-Küchler,* 3rd ed., **vol. 4,** p. 191.
104. Y. Sugi, T. Matsuzaki, T. Hanaoka, Y. Kubota, J.-H. Kim, X. Tu, M. Matsumoto, *Stud. Surf. Sci. Catal.* **90** (1993) 397.
105. M. Minamow, "Commercial Facility to Produce Dimethyl-2,6-Naphthalenedicarboxylate Amoco® NDC", press release, Amoco Chemical Company, Feb. 24, 1993; Amoco Corp., US 5 073 670, 1991 (D. L. Sikkenga, J. D. Lamb, I. C. Zaenger, G. S. Williams); *Chem. Abstr.* **116**:128396.
106. Gulf Research and Development Co., US 2655546, 1953 (D. R. Stevens, R. S. Bowman); *Chem. Abstr.* **48** (1954) 10767 g.
107. J. Klein, H. Widdecke, *Chem. Ing. Tech.* **51** (1979) 560.
108. R. Stroh, R. Seydel, W. Hahn in W. Foerst (ed.): *Neuere Methoden der präparativen organischen Chemie,* **vol. 2,** Verlag Chemie, Weinheim 1960, p. 231.
109. G. G. Knapp, T. H. Coffield, J. P. Napolitano, H. D. Orloff, C. J. Worrel: *World Petrol Congr., Proc., 7th.* **vol. 5,** Elsevier, London 1967 (Pub. 1968) p. 403.
110. H. Widdecke, J. Klein, *Chem. Ing. Tech.* **53** (1981) 954.
111. R. Gompper, *Angew. Chem.* **76** (1964) 412; *Angew. Chem. Int. Ed. (Engl.)* **3** (1964) 560.
112. R. Stroh, R. Seydel, W. Hahn, *Angew. Chem.* **69** (1957) 699.
113. A. J. Kolka, J. P. Napolitano, A. H. Filbey, G. G. Ecke, *J. Org. Chem.* **22** (1957) 642.
114. *Houben-Weyl,* **6/1 c,** 1073.
115. *Winnacker-Küchler,* 4th ed., **vol. 6,** p. 150.
116. *Ullmann,* 4th ed., **vol. 18,** p. 191.
117. *Ullmann,* 4th ed., **vol. 15,** p. 61.
118. F. F. Meinhold, J. Bardzig, *Chem. Process. (Chicago)* **31** (Apr. 1968) 32.
119. Coalite and Chemical Products, GB 1064337, 1967 (H. F. Bondy, A. D. Gregory, F. R. Moore, G. E. Munro); *Chem. Abstr.* **68** (1968) 59279 e.
120. General Electric Co., NL 6506830, 1965; *Chem. Abstr.* **65** (1966) 2173 c.
121. M. Inoue, S. Enomoto, *Chem. Pharm. Bull.* **24** (1976) 2199.
122. Union Rheinische Braunkohlen Kraftstoff, DE 1281448, 1968; DE 1297110, 1969 (G. Velling, R. Huettner); *Chem. Abstr.* **70** (1969) 37457 a; **71** (1969) 70293 n.
123. B. E. Leach, *J. Org. Chem.* **43** (1978) 1994.
124. Bayer, DE 1142873, 1963 (W. Hahn); *Chem. Abstr.* **58** (1963) 12468 g.
125. Union Carbide Corp., GB 953929, 1964 (S. Kaufman, R. E. Nicolson); *Chem. Abstr.* **61** (1964) 9435 g.
126. Shell Development Co., GB 557519, 1943 (D. B. Luten Jr., A. De Benedictis); *Chem. Abstr.* **39** (1945) 3304[8].
127. Nalco Chem. Co., US 4092367, 1978 (B. W. Bridwell, C. E. Johnson); *Chem. Abstr.* **89** (1978) 129254 c.
128. A. V. Topchiev, S. V. Zavgorodnii, V. G. Kryuchkova: *Alkylation with Olefins,* Elsevier, Amsterdam 1964, p. 184.
129. R. Stroh, J. Ebersberger, H. Haberland, W. Hahn in W. Foerst (ed.): *Neuere Methoden in der Präparativen Organischen Chemie,* **vol. 2,** Verlag Chemie, Weinheim 1960, p. 155.
130. R. Stroh, J. Ebersberger, H. Haberland, W. Hahn, *Angew. Chem.* **69** (1957) 124.
131. N. I. Shuikin, B. L. Lebedev, V. G. Nikol'skii, O. A. Korytina, A. V. Kessenikh, E. P. Prokof'ev, *Izv. Akad. Nauk. SSSR Ser. Khim.* 1967, 1618; *Chem. Abstr.* **68** (1968) 95607 u.

132. W. M. Kutz, B. B. Corson, *J. Am. Chem. Soc.* **68** (1946) 1477.
133. F. Asinger: *Chemistry and Technology,* Pergamon Press, Oxford 1968.
134. I. G. Farbenindustrie, FR 666718, 1929 (R. Michel).
135. I. G. Farbenindustrie, DE 550494, 1932 (R. Michel).
136. Phillips Petroleum Co., US 2393132, 1946 (C. R. Wagner); *Chem. Abstr.* **40** (1946) 2472[5].
137. S. J. Goldberg, *J. Am. Chem. Soc.* **84** (1962) 3022.
138. K. L. Rinehart Jr., P. A. Kittle, A. F. Ellis, *J. Am. Chem. Soc.* **82** (1960) 2082.
139. A. N. Nesmeyanov, N. S. Kochetkova, *Izv. Akad. Nauk SSSR Ser. Khim.* 1958, 242; *Chem. Abstr.* **52** (1958) 12852 c.
140. K. Hafner, K. L. Moritz in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 4,** J. Wiley & Sons, New York 1965, p. 127.
141. E. G. Perevalova, T. V. Nikitina in E. I. Becker, M. Tsutsui (eds.): *Organometallic Reactions,* **vol. 4,** Wiley-Interscience, New York 1972, p. 177.
142. G. A. Olah, W. S. Tolgyesi in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 2,** part 2, J. Wiley & Sons, New York 1964, p. 659.
143. R. C. Fuson, C. H. McKeever: *Chloromethylation of Aromatic Compounds, Organic Reactions,* J. Wiley & Sons, New York 1942, p. 63.
144. J. E. Hofmann, A. Schriesheim in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 2,** part 2, J. Wiley & Sons, New York 1964, p. 597.
145. I. G. Farbenindustrie, FR 716604, 1931 (H. Hopff).
146. K. Sisido, H. Inada, T. Isida, *Tetrahedron Lett.* 1968, 5267.
147. P. H. Gore in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 3,** part 1, J. Wiley & Sons, New York 1964, p. 1.
148. H. Meyer, G. Aichinger in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart 1975, p. 341.
149. R. C. Fuson: *Reactions of Organic Compounds,* J. Wiley & Sons, New York 1962, p. 44.
150. P. H. Gore, J. A. Hoskins, *J. Chem. Soc. (C)* (1970) 517.
151. A. G. Pinkus, N. Kalyanam, *Org. Prep. Proc.* **10** (1978) 255.
152. G. M. Kosolapoff, *J. Am. Chem. Soc.* **69** (1947) 1651.
153. ICI, GB 2 235 195, 1989 (B. N. Hendy, D. Macdonald Smith, K. H. Patterson); *Chem. Abstr.* **115**:93214.
154. D. P. N. Satchell, R. S. Satchell in S. Patai (ed.): *The Chemistry of the Carbonyl Group,* Wiley-Interscience, New York 1966, p. 233.
155. C. A. Thomas, M. B. Moshier, H. E. Morris, R. W. Moshier: *Anhydrous Aluminum Chloride in Organic Chemistry,* Reinhold Publ. Co., New York 1941.
156. G. Kränzlein: *Aluminiumchlorid in der Organischen Chemie,* Verlag Chemie, Berlin 1939.
157. L. F. Martin, P. Pizzolato, L. S. McWaters, *J. Am. Chem. Soc.* **57** (1935) 2584.
158. W. A. Riddell, C. R. Noller, *J. Am. Chem. Soc.* **52** (1930) 4365.
159. A. Windaus, K. Raichle, *Justus Liebigs Ann. Chem.* **537** (1939) 157.
160. H. Galenkamp, A. C. Faber, *Recl. Trav. Chim. Pays-Bas* **77** (1958) 850.
161. D. E. Pearson, C. A. Buehler, *Synthesis* (1972) 553.
162. BASF, DE 2 204 973, 1972 (G. Eppic, K. Hutmacher); *Chem. Abstr.* **79**:115315.
163. Monsanto, US 2 879 296, 1959 (E. J. Prill); *Chem. Abstr.* **54**:1306[g].
164. Velsicol Chemical Corp., EP 546 133, 1992 (A. E. Rains, T. E. Lea, D. I. Templer); *Chem. Abstr.* **119**:205854.
165. J. O. Morley, *J. Chem. Soc. Perkin Trans II* (1977) 601.
166. J. Iqbal, M. A. Khan, N. K. Nayyar, *Tetrahedron Lett.* **32** (1991) 5179.
167. A. Kawada, S. Mitamura, S. Kobayashi, *Synlett* (1994) 545.
168. S. Kobayashi, *Synlett* (1994) 689.
169. I. Hachiya, M. Moriwaki, S. Kobayashi, *Bull. Chem. Soc. Jpn.* **68** (1995) 2053.
170. I. Hachiya, M. Moriwaki, S. Kobayashi, *Tetrahedron Lett.* **36** (1995) 409.
171. Hoechst Celanese Corp., EP 284 310, 1988 (V. Elango, M. A. Murphy, B. L. Smith, K. G. Davenport, G. N. Mott, G. L. Moss); *Chem. Abstr.* **110**:153916.
172. Hoechst Celanese Corp., EP 479 593, 1991 (D. D. Lindley, T. A. Curtis, T. R. Ryan, E. M. De la Garza, C. B. Hilton, T. M. Kenesson); *Chem. Abstr.* **117**:133852.
173. Elf Atochem, WO 96/28410, 1996 (M. Devic); *Chem. Abstr.* **125**:300625.
174. W. F. Short, H. Stromberg, A. E. Wiles, *J. Chem. Soc.* 1936, 319.

175. G. Perrier, *Chem. Ber.* **33** (1900) 815.
176. K. Elbs, *J. Prakt. Chem.* **33** (1886) 181.
177. L. Bouveault, *Bull. Soc. Chim. Fr.* **17** (1897) 1020.
178. Y. Yamase, *Bull. Chem. Soc. Jpn.* **34** (1961) 480.
179. Y. Yamase, *Bull. Chem. Soc. Jpn.* **34** (1961) 484.
180. T. W. Campbell, *J. Am. Chem. Soc.* **82** (1960) 3126.
181. H. Hopff, *Angew. Chem.* **60** (1948) 245.
182. F. Runge, H. Reinhardt, G. Kühnhanss, *Chem. Tech. (Leipzig)* **8** (1956) 644.
183. S. Grateau, *C. R. Hebd. Seances Acad. Sci.* **191** (1930) 947.
184. R. D. Haworth, *J. Chem. Soc.* 1932, 1125. M. A. Nisbeth, *Chem. Ind. (London)* **47** (1969) 1699.
185. H. Burton, D. A. Munday, *J. Chem. Soc.* 1957, 1727.
186. B. Chiche, A. Finiels, C. Gauthier, P. Geneste, J. Graille, D. Pioch, *J. Mol. Catal.* **42** (1987) 229.
187. Bayer AG, EP 473 023, 1994 (H. Waldmann, M. Hayek, O. Immel, L. Puppe, R. Braden); *Chem. Abstr.* **116**:255321.
188. H. C. Brown, G. Marino, L. M. Stock, *J. Am. Chem. Soc.* **81** (1959) 3310.
189. K. Tanabe, M. Misono, Y. Ono, H. Hattori, *Stud. Surf. Sci. Catal.* **51** (1989) 240.
190. K. Freudenberg, K. Weinges, *Justus Liebigs Ann. Chem.* **590** (1954) 140.
191. N. M. Cullinane, B. F. R. Edwards, *J. Appl. Chem.* **9** (1959) 133.
192. D. Bailey, *Chem. Ind. (London)* 1971, 682.
193. T. P. Smyth, B. W. Corby, *Org. Process Res. Dev.* **1** (1997) 264.
194. H. P. Kaufmann, L. S. Huang, H. Buckmann, *Chem. Ber.* **75** (1942) 1236.
195. V. K. Mehta, M. J. Sacha, S. R. Patel, *J. Indian Chem. Soc.* **33** (1956) 867.
196. W. S. Johnson: *Organic Reactions,* **vol. 2,** J. Wiley & Sons, New York 1944, p. 114.
197. W. E. Bachmann, D. G. Thomas, *J. Am. Chem. Soc.* **64** (1942) 94.
198. *Winnacker-Küchler,* 4th ed., **vol. 6,** p. 267.
199. H. F. Bassilios, S. M. Makar, A. Y. Salem, *Bull. Soc. Chim. Fr.* 1954, 72.
200. G. Harvey, G. Maeder, *Collect. Czech. Chem. Commun.* **57** (1992) 862.
201. *Ullmann,* 4th ed., **vol. 7,** p. 101.
202. P. C. Mitter, H. Mukherjee, *J. Indian Chem. Soc.* **16** (1939) 393.
203. W. S. M. Grieve, D. H. Hey, *J. Chem. Soc.* 1933, 968.
204. P. H. Gore, *J. Org. Chem.* **22** (1957) 135.
205. J. V. Heid, *J. Org. Chem.* **13** (1948) 409.
206. A. I. Kosak, H. D. Harthough, *Org. Synth. Coll.* **3** (1955) 14.
207. W. D. Cooper, *J. Org. Chem.* **23** (1958) 1382.
208. BASF AG, EP 332 075, 1989 (M. Hesse, W. Hoelderich, T. Dockner, H. Koehler); *Chem. Abstr.* **112**:20996.
209. F. Richard, H. Carreyre, G. Perot, *J. Catal.* **159** (1996) 427.
210. N. P. Buu-Hoi, R. Royer, *Recl. Trav. Chim. Pays-Bas* **66** (1947) 533.
211. A. Burger, H. W. Bryant, *J. Org. Chem.* **4** (1939) 119.
212. L. L. Woods, P. A. Dix, *J. Org. Chem.* **24** (1959) 1126.
213. R. B. Woodward, M. Rosenblum, M. C. Whiting, *J. Am. Chem. Soc.* **74** (1952) 3458.
214. C. R. Hauser, J. K. Lindsay, *J. Org. Chem.* **22** (1957) 482.
215. M. Rosenblum, R. B. Woodward, *J. Am. Chem. Soc.* **80** (1958) 5543.
216. J. R. Kozikowski, R. E. Maginn, M. S. Klove, *J. Am. Chem. Soc.* **81** (1959) 2995.
217. A. G. Anderson Jr., J. A. Nelson, *J. Am. Chem. Soc.* **72** (1950) 3824.
218. A. G. Anderson Jr., R. Scotoni Jr., E. J. Cowles, C. G. Fritz, *J. Org. Chem.* **22** (1957) 1193.
219. L. Gattermann, J. A. Koch, *Chem. Ber.* **30** (1897) 1622.
220. L. Bert, *C. R. Hebd. Seances Acad. Sci.* **221** (1945) 77.
221. S. Fujiyama, T. Kasahara, *Hydrocarbon Process.* **57** (1978) no. 11, 147.
222. Mitsubishi Gas Chemical Co., GB 2025956, 1980 (M. Komatsu, T. Ohta, T. Tanaka, R. Oda, Y. Takamizawa); *Chem. Abstr.* **93** (1980) 7864 p.
223. A. Vilsmeier, *Chem. Ztg.* **75** (1951) 133.
224. A. Rieche, H. Gross, E. Hoft, *Chem. Ber.* **95** (1960) 88.
225. P. E. Spoerri, A. S. du Bois, *Org. React. (N.Y.)* **5** (1949) 387.
226. W. Ruske in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 3,** J. Wiley & Sons, New York 1964, p. 383.
227. K. W. Rosenmund, W. Schurr, *Justus Liebigs Ann. Chem.* **460** (1928) 56.
228. J. Kassner, H. Zimmer in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart 1975, p. 432.
229. J. F. J. Dippy, J. H. Wood, *J. Chem. Soc.* 1949, 2719.
230. A. R. Parikh, K. H. Thaker, R. D. Desai, *J. Inst. Chem. (India)* **43 (part 5)** (1971) 183.

231. L. F. Fieser, C. K. Bradsher, *J. Am. Chem. Soc.* **58** (1936) 1738.
232. A. H. Blatt: *Organic Reactions,* **vol. 1,** J. Wiley & Sons, New York 1942, p. 342.
233. A. Gerecs in G. A. Olah (ed.): *Friedel – Crafts and Related Reactions,* **vol. 3,** part 1, J. Wiley & Sons, New York 1964, p. 499.
234. N. I. Shuikin, B. L. Lebedev, *Russ. Chem. Rev. (Engl. Transl.)* **35** (1966) 448.
235. M. Siskin, I. Mayer, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 191.
236. S. I. Hommeltoft, L. Jørgensen, *Erdöl, Erdgas, Kohle* **114** (1998) 248.
237. R. R. Frame, L. O. Stine, H. U. Hammershaimb, B. S. Muldoon, *Erdöl, Erdgas, Kohle* **114** (1998) 385.
238. D. P. N. Satchell, *Q. Rev. Chem. Soc.* **17** (1963) 160.
239. P. N. Rylander: *Catalytic Hydrogenation over Platinum Metals,* Academic Press, New York 1967, p. 291.
240. A. L. J. Beckwith in J. Zabicky (ed.): *The Chemistry of Amides,* Wiley-Interscience, New York 1970, p. 73.
241. H. J. Roth, A. Kleemann: *Arzneistoffsynthese,* Thieme Verlag, Stuttgart 1982.
242. K. H. Büchel: *Pflanzenschutz und Schädlingsbekämpfung,* Thieme Verlag, Stuttgart 1977.
243. H. Zollinger: *Azo and Diazo Chemistry,* Wiley-Interscience, New York 1961, p. 69.
244. H. O. House: *Modern Synthetic Reactions,* 2nd ed., Benjamin, New York 1965, p. 163.
245. T. M. Harris, C. M. Harris, *Org. React. (N.Y.)* **17** (1969) 155.
246. A. C. Cope, H. L. Holmes, H. O. House, *Org. React. (N.Y.)* **9** (1957) 107.
247. J. P. Collman, L. S. Hegedus: *Principles and Application of Organotransition Metal Chemistry,* University Science Books, Mill Valley, Calif., 1980.
248. H. Zeiss: *Organometallic Chemistry,* Reinhold Publ. Co., New York 1960.
249. *Winnacker-Küchler,* 4th ed., **vol. 6,** p. 816.
250. *Winnacker-Küchler,* 4th ed., **vol. 6,** p. 133.

Adhesives

Werner Haller, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chaps. 1 – 3, Sections 9.7 and 9.10)

Hermann Onusseit, Haan, Federal Republic of Germany (Chap. 4, Sections 9.1, 9.2, 9.3, 9.4)

Gerhard Gierenz, Solingen, Federal Republic of Germany (Chap. 5)

Werner Gruber, Korschenbroich, Federal Republic of Germany (Chap. 6, Section 9.9)

Richard D. Rich, Loctite Corporation, Rocky Hill, CT 06067, United States (Chaps. 6 and 8 in part, Section 9.21)

Günter Henke, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 7)

Lothar Thiele, Langenfeld, Federal Republic of Germany (Chap. 8)

Horst Hoffmann, Dorus Klebetechnik GmbH & Co. KG, Bopfingen, Federal Republic of Germany (Section 9.5)

Dieter Dausmann, Münchweiler, Federal Republic of Germany (Section 9.6)

Riza-Nur Özelli, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 9.8)

Udo Windhövel, Monheim, Federal Republic of Germany (Section 9.11)

Hans-Peter Sattler, Kamen, Federal Republic of Germany (Section 9.11 in part, Section 9.12)

Wolfgang Dierichs, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Sections 9.13, and 9.15)

Günter Tauber, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 9.14)

Michael Hirthammer, Henkel Teroson, Heidelberg, Federal Republic of Germany (Section 9.16)

Christoph Matz, Lufthansa Technik AG, Hamburg, Federal Republic of Germany (Section 9.17)

Matthew Holloway, Loctite RD & E, Dublin, Ireland (Section 9.18)

David Melody, Dublin, Ireland (Section 9.19)

Ernst-Ulrich Rust, Bergisch Gladbach, Federal Republic of Germany (Section 9.20)

Ansgar van Halteren, Industrieverband Klebstoffe e.V., Düsseldorf, Federal Republic of Germany (Chap. 10)

1. Introduction 2
2. Adhesion Theories 3
3. Definitions 5
4. Raw Materials for Adhesives 7
5. Classification of Adhesives 12
6. Individual Adhesive Systems 14
6.1. Adhesives That Set Without a Chemical Reaction 14
6.1.1. Solvent-Free Adhesive Systems 14
6.1.2. Adhesive Solutions from Which the Solvents Escape before Bonding . . . 16
6.1.3. Adhesive Solutions from Which the Solvents Evaporate during Bonding . 18
6.1.4. Aqueous Emulsions of Polymeric Compounds 20
6.2. Adhesives Setting by Chemical Reaction 21
6.2.1. Adhesives Setting by Polymerization 21
6.2.2. Adhesives Setting by Polyaddition . . 25
6.2.3. Adhesives Setting by Polycondensation 27
6.2.4. Vulcanizing Adhesives 29
6.2.5. Ultraviolet/Electron Beam (UV/EB) Curing Adhesives 29
6.2.6. Conductive Adhesives 30
7. Bonding Techniques 31
8. Testing of Adhesives 34
9. Applications of Adhesives 36
9.1. Bookbinding 36
9.2. Adhesives for Packaging and Packaging Materials 37
9.3. Nonwovens Hygiene Industry 39
9.4. Hygienic Papers 40
9.5. Gluing of Wood and Wooden Materials 40
9.6. Footwear Adhesives 43

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_221

9.7. Bonding of Plastics 44
9.8. Bonding of Elastomers 46
9.8.1. Adhesion of Rubber 46
9.8.2. Bonding of Rubber 46
9.9. Bonding of Metals 47
9.10. Adhesives for Wallcoverings 49
9.11. Floorcovering Adhesives 49
9.12. Building Construction Adhesives . . 50
9.13. Adhesives for Bonding Textile Fabrics 52
9.14. Flocking Adhesives 52
9.15. Adhesives for Bonding Glass 53
9.16. Adhesives in Automobile Manufacture 53
9.17. Adhesives in Aircraft Construction 55
9.18. Adhesives and Sealants in Electronics 58
9.19. Medical Adhesives 60
9.20. Household Adhesives 62
9.21. Applications of Anaerobic Adhesives. 65
10. Economic Aspects 65
11. References 66

1. Introduction

Adhesives are defined as nonmetallic substances capable of joining materials by surface bonding (adhesion), the bond itself possessing adequate internal strength (cohesion). Adhesive is a generic term and covers other common terms, such as glue, paste, gums, adhesive cement, and bonding agent.

History. Adhesion is among the oldest technologies of mankind [129]. As early as the Stone Age, our ancestors made stone axes and other tools using mineral pitch or wood rosin. Asphalt was used in building the Tower of Babel. In ancient Egypt, it was already customary to veneer furniture with the aid of animal glues. Flour- and casein-based adhesives were applied in ancient Rome. Adhesives of plant origin such as wood rosin were as well-known in China, as was the gum arabic and the caoutchouc ("weeping wood") from the tropical regions of South America and Asia.

In the Middle Ages, the first glue-boiling plants came into being and produced protein glues from animal raw materials (glutin from hides and bones, blood albumin, casein from milk) or starch paste from plants.

With the developments in chemistry in the early 1900s, the technology of adhesives began to boom. Phenolic resins (ca. 1900), melamine resins, and urea resins (ca. 1930), polymer dispersions, epoxy resins (1938), and cyanoacrylates (1957) largely superseded the classic natural adhesives and, with a multitude of bindings agents, form the basis for modern adhesive technology, which is one of the most advanced joining processes.

Composition. An adhesive is composed of basic raw materials, which are called binders [1] and which determine its adhesiveness (adhesion) and its internal strength (cohesion), and of frequently necessary auxiliaries, which establish particular end-use and processing characteristics. The adhesiveness of an adhesive, its internal strength after setting, and its processing characteristics are the fundamental properties that determine its suitability for use in forming adhesive joints. Adhesive joints are the joints formed between substrates and adherents using adhesives.

The binders used for adhesives are primarily high polymers having optimal strength properties. High internal strength (cohesion) is essential if the adhesive in an adhesive joint is to be able to transmit forces from one adherent to the other. Most adhesives contain high molecular mass organic substances as their basic raw materials or reactive organic compounds that are preliminary stages of polymers and that react during the bonding process to form polymers. Inorganic polymers, such as the various types of waterglass, are used only to a very limited extent.

Virtually any standard polycondensate, homopolymer, and copolymer and also polyadducts may be used, provided they can be applied as solutions, dispersions, emulsions, or melts. In addition to these raw materials, auxiliaries such as resins, plasticizers, fillers, thickeners, solvents, antiagers, preservatives, hardeners, or setting retarders, are required, depending on the end use. Their function is inter alia to adjust tack, to improve adhesion, to make flexible, to regulate viscosity, to stabilize, and to influence setting or hardening.

Adhesive Joints. In many cases, adhesion is as effective as other joining techniques, such as riveting, welding, soldering, and screwing; in some cases it complements those techniques, and frequently it affords numerous advantages. On the other hand, certain processing requirements and material properties related to chemical structure can restrict the use of adhesives.

Bonding has the advantage that the joining of adherents can be carried out very rationally, often extremely quickly, and particularly economically [2]. One important feature common to all adhesive joints is the highly uniform distribution of forces over the entire joint area by comparison with rivet and screw joints. The uniform stress levels frequently provide for more optimal utilization of material strength. Accordingly, thinner and hence lighter sections of material may be used. This has resulted in the saving of material and, for example, in the automotive, aircraft, manufacturing, and furniture industries, in totally new structural elements (sandwich elements).

The use of adhesives enables numerous materials, even those differing widely in type, to be joined to one another. This applies above all to materials that cannot be joined to one another by other techniques. Because some adhesives set even in relatively thick layers, any unevenness in the constituent material of the adherents may be smoothed out during bonding, or significant dimensional tolerances may even be bridged by the gap-filling principle. The adhesive layer of an adhesive joint even may have a vibration-dampening effect and, by virtue of its insulating properties, can prevent contact corrosion in adhesive joints involving metals of different normal potential. Another advantage is the fact that many adhesives can be applied at room temperature or, when heat must be used for application, it is sufficient to apply temperatures at which the constituent materials of the adherents are not affected, as can happen in the welding of metals and plastics.

The usability of adhesives can be restricted when stringent requirements are imposed on the thermal stability of an adhesive joint [2]. Like all plastics, adhesives based on organic polymers also show a marked dependence on temperature in their strength properties. The strength properties of an adhesive joint can be kept constant and sufficiently high in only a relatively narrow temperature range. Adhesives cured by thermal cross-linking which are still relatively temperature resistant are generally capable of withstanding temperatures of up to about 150 °C. Using new, but more difficult to apply adhesives based on polyimides [3], it is possible to form adhesive joints which are capable of withstanding temperatures of the order of 250 °C for long periods. On the other hand, some adhesives, when subjected to heavy static stress for long periods, show a tendency even at room temperature to creep, a plastic deformation that can result in separation of the adhesive joint.

2. Adhesion Theories

The atoms or molecules in any solid are held together by so-called cohesion forces. These interatomic and intermolecular forces are conservative in character, i.e., not dependent upon the pretreatment of the material [4, p. 1]. After a break in the solid, the two fragments cannot be put back together in such a way as to restore the solid to its original state. Reactivation of the cohesion forces would be possible only if the original intervals between atoms or molecules could be reestablished. The same may be said also of the joining of two different materials; a mutual convergence over the entire joint area on the radius of interaction of the molecules would result in adhesion of the same order of magnitude as the cohesion forces [5]. In practice such convergence is not possible. There are three reasons for this: first, the surface roughness of the adherents; second, a reduction in the effective surface for boundary layer reactions due to the surface tension of the adhesive; and third, the formation of the so-called weak boundary layer [130], [131]. This is a region within the adhesive film close to the boundary layer where the reactive groups of the molecules are directed to the adherent and consequently do not contribute to the curing process, that is, no cross-linking occurs. Therefore, the weak boundary layer is always a weak link in the adhesive joint. The function of the adhesive in the boundary zone is to compensate for the effects of surface roughness by providing optimal wetting of the adherents.

Over the years, several different theories have been developed as to the mode of action of adhesives, i.e., the mechanism of adhesion. They

extend from the simple mechanical theory via the electrostatic and adsorption theory to the diffusion theory, these being the most important [6]. One feature common to all these theories is that, for some adhesion phenomena, these theories allow qualitatively satisfactory assertions to be made, but in other respects they fail totally and, in some cases, even lead to conflicting results. The adhesion of adhesive to an adherent, on which any joint is based, is clearly not a consistent and isolatable process [7] but a complex addition of various adhesion effects. Because the number of individual processes involved, apart from exceptions, is extremely difficult to estimate, definitive confirmation or rejection of individual theories is hardly possible.

Mechanical Theory of Adhesion. The oldest theory of adhesion is definitely the mechanical theory. It is based on mechanical anchorage of the adhesive in pores and irregularities in the adherent and is discussed primarily in reference to wood and similar porous materials [8].

Electrostatic Theory of Adhesion. According to the electrostatic theory [9], the adhesion forces between adherent and adhesive layer are applied by contact or transfer potentials. These transfer potentials cause the buildup of an electric double layer at the adhesive – adherent boundary and corresponding Coulomb attraction forces between the two components. In principle, the occurrence of transfer potentials is unquestionable, as shown by the electrostatic discharges that can be detected in the destruction of adhesive joints. However, the practical significance of the attraction forces associated with these discharges is still being debated [4, pp. 150 – 153], [10].

Adsorption Theory . As its name indicates, this theory of adhesion draws upon surface forces for explaining the observed phenomena [11]. It regards adhesion as essentially a special property of phase interfaces. The forces that are responsible for adhesion in this process are the so-called secondary valence or van der Waal's forces. These forces have three components, namely Keesom's dipole orienting effect, Debye's induced dipole effect, and London's dispersion effect [12–14]. For these forces to become active, the distances between the molecules of adhesive and adherent must converge toward molecular intervals. This requires complete spreading of the adhesive over the surface of the adherent. Good adhesion can be expected if the adhesive (in liquid form or in a liquid medium) wets the adherent [8]. Accordingly, the adsorption theory may be placed on an entirely thermodynamic basis. Adhesion is thus determined by the ratio between the surface energies of the adhesive and the constituent material of the adherent in the sense that the specific surface energy of the adhesive must be lower than that of the adherent. Accordingly, materials having high surface energy levels, such as metals, and those having medium surface energy levels, such as wood and paper, may be bonded relatively easily. In the case of polymers, bonding becomes increasingly more difficult [15], and finally almost impossible as surface energy decreases (polyolefins and polytetrafluoroethylene). The remedial surface treatment of materials such as these may be interpreted as increasing their surface energy [16].

The adsorption theory also shows that the adhesion forces of two materials are not reciprocal. For example, if a liquid epoxy resin adhesive is allowed to set on the surface of polytetrafluoroethylene or polyethylene, a very weak adhesive joint is formed. If, by contrast, liquid polytetrafluoroethylene or polyethylene is applied to the surface of hardened epoxy resin, strong adhesive joints are obtained [17]. In practice, this theory is not free from contradictions either; above all, it does not answer the question whether the difference in the surface energies between two materials is indicative of the intensity of the adhesion force [18].

Diffusion Theory. The diffusion theory of adhesion is essentially applicable to the bonding of high polymers. According to this principle, adhesion is obtained by the mutual penetration of adhesive and substrate [19], [20]. This mobility is based on the fundamental properties of high polymers: their chainlike structure and resulting mobility, allowing the possibility that the chains possess Brownian molecular movements in a submolecular range. By virtue of their greater mobility, the adhesive molecules normally play a greater part in the diffusion process. However, if the adhesive is in solution, which is generally the case, and if the substrate is slightly soluble in

the solvent, substrate molecules or parts thereof also diffuse into the spread adhesive. A diffusion bond is characterized by the disappearance of a clear boundary between the two phases and by the development of a gradual transition from one phase to the other. The mechanism of adhesion has developed into a three-dimensional process and is no longer confined to one interface.

The interdiffusion of the polymer molecules of adhesive and substrate is dependent upon various parameters, such as pressure, time, temperature, molecule size, and, of course, the reciprocal solubility, as shown by the correlation between the compatibility of the polymers and the quality of the bond [21]. Examples of bonds to which the diffusion theory is applicable include the bonding of PVC-U adherents to PVC in solvents containing tetrahydrofuran and so-called contact bonding, where the diffusion process takes place between two adhesive layers and not between adhesive and substrate [22].

The limits of the diffusion theory of adhesion show up in the adhesion of polymers, for example, to metal or glass. In this case, this theory does not make any useful contribution to the understanding of adhesion.

Other Theories. In addition to these theories, some special cases are discussed in the literature. They include *adhesion through primary valence forces,* for example, in the bonding of metals [8], [23], and so-called *liquid adhesion.* In the latter, a thin film of a liquid of extremely high viscosity produces adhesion through a process in which separation of adherent and substrate results in a flow in the narrow gap which is only possible by overcoming considerable resistance. Liquid adhesion is particularly assumed in pressure-sensitive bonding [16] and in the initial tack of a liquid adhesive.

Summary. None of the theories covers every single aspect of bonding. A promising combination of the theories to bring about an improvement does not exist at present. In addition, the theory of adhesion as a physicochemical phenomenon and physically measurable adhesion are generally still a considerable distance apart, so that the adhesion theories are only reference points. Secondary factors, such as surface roughness, boundary layers with faults (atmosphere, impurities), joint design, type of stress, and aging also influence bond strength to a considerable extent.

3. Definitions

Some of the terms are defined in DIN 16920 [24] and EN 923 [132].

Adherent: a body that is or is intended to be bonded to another body.

Adhesion: state in which two surfaces are held together by interfacial bonds.

Adhesive application: the application and distribution of an adhesive over the surface to be joined.

Adhesive coat: adhesive layer applied to an adherent.

Appearance of separated surfaces: surfaces formed in the event of failure of an adhesive joint; they provide information on the reason for the failure. Distinctions can be drawn visually between:

1) A bond layer that does not cover the entire joint area (uneven adhesive application)
2) Adhesion failure: rupture of the bond layer from the surface of the adherent
3) Cohesion failure: break in the bond layer
4) Insufficient coalescence: separation of two adhesive layers at the interface, for example, in the case of contact adhesives
5) Rupture of the surface layer of an adherent, breakage of the adherent (material failure)

In many cases, several types of failure occur together.

Assembly time: interval between application of adhesive to the adherents and the initiation by heat and/or pressure of the setting process in the assembled joint.

Application weight, spread of adhesive, coverage: quantity of adhesive applied per unit area of a surface. A distinction is drawn between wet application (including solvent or dispersant) and dry application (100 % adhesive). Each

type of adhesive has technically and economically an optimal application weight. In principle, rough and porous substrates (paper, fabrics) require more adhesive than smooth surfaces (glass, metals, films) to obtain whole-surface wetting and bonding. Application weights range from 1 g/m^2 (for example, for bonding films) to several 100 g/m^2, depending on the substrate, the method of application, and the type of adhesive.

Bond layer thickness: the thickness of the set layer of adhesive in millimeters. It may be influenced by the application weight and the fixing pressure. In general, a thin (< 0.001 mm) and uniform bond layer leads to the highest bond strengths.

Bond line: adhesive layer in a bonded joint.

Bond strength: force necessary to bring an adhesive joint to the point of failure in or near the plane of the bond line.

Closed assembly time: interval between assembly of the adhesive joint and the initiation by heat and/or pressure of the setting process in the assembled joint.

Cohesion: state in which the particles of a single substance are held together by primary or secondary valence forces.

Cold flow: deformation of an adhesive layer (or film) at room temperature without external applied load.

Creep: slow inelastic deformation of an adhesive layer under load following initial instantaneous elastic or rapid deformation.

Curing: the setting of an adhesive by chemical reaction (polymerization, polycondensation, polyaddition).

Heat resistance: the maximum temperature that a set adhesive in a joint is capable of withstanding. It depends upon the nature and intensity of the load and must be determined in each individual case.

Initial tack: the ability of a setting adhesive to hold two substrates together immediately after they have been joined. At this stage of the bonding process, strength is determined largely by rheological properties and depends on the cohesion of the unset adhesive. In the case of hot-melt adhesives, the analogous term "hot tack" is used, and this property is determined by the viscosity of the unsolidified melt. The transitions between the use of the terms initial tack and tack are fluent.

Minimum film-forming temperature: the temperature at which an emulsion-based adhesive just no longer dries to form a clear, homogeneous film.

Open assembly time, open time: interval between adhesive application to the adherents and assembly of the adhesive joint. The expression "open time" also is used occasionally for the period of time that elapses in a machine between application of the adhesive and bonding.

Primers: (mixtures of) substances that are applied before the adhesive to one or both of the surfaces to be joined, for improving adhesion and/or durability of the bond.

Setting: process by which an adhesive develops its cohesive strength and hence the physical and chemical properties of its bond.

Setting rate: the increase in strength in the bond line per unit time. It depends upon the setting mechanism and temperature and also may be influenced by the substrate (absorbency), bond layer thickness, etc. The setting rate generally is not constant and decreases toward the end of the setting time.

Setting time: time between joining of the adherents and the setting of the adhesive. It varies from a few seconds to several days, depending on the type of adhesive.

Shear strength: force per unit surface area necessary to bring an adhesive joint to the point of failure by means of shear forces.

Shelf life, storage life: time of storage under stated conditions during which an adhesive can be expected to retain its working properties. Appropriate storage generally is understood to mean that the adhesives are stored in clean, hermetically sealed containers and in cool, dry, well-ventilated rooms. Shelf life is largely dependent upon temperature. Aqueous and in some cases even solvent-containing systems have a tendency for skin formation, separation, and an increase in viscosity, particularly when stored for long periods at relatively high temperatures. They also can be spoiled by microbial attack. Wide fluctuations in temperature (freeze/thaw) lead to coagulation. Shelf lives of less than 12 months generally are indicated on the packing. Manufacturer's instructions should always be followed.

Tack: property of a material that enables it to form a bond immediately on contact with another surface, which can be an adherent or another layer of adhesive. High tack is particularly important in the case of pressure-sensitive adhesives (tapes, labels), for which it is even measured. Tack is generally difficult to measure in setting adhesives.

Thixotropy: decrease in apparent viscosity under shear stress, followed by a gradual recovery when the stress is removed. Thixotropic behavior allows pasty adhesives, for example, to be spread readily and applied in relatively thick layers, in some cases even on vertical surfaces, without running or flowing out of the gap before bonding because the adhesive loses its fluidity immediately after application.

Viscosity: property of a material to increasingly resist deformation with increasing rate of deformation. This property is quantitatively defined as dynamic viscosity or coefficient of viscosity and is often used synonymously with apparent viscosity. The viscosity of adhesives is primarily determined by means of rotational or throughflow viscometers (DIN cup, Ford cup, Zahn cup). Adhesives generally show non-Newtonian behavior. In addition to temperature, any expression of viscosity must also refer to the measuring instrument and measurement parameters (rotating spindle, rate of shear, nozzle diameter).

Working life, pot life: period of time during which a multicomponent adhesive can be used after mixing the components. The pot life of a given system depends primarily on the initial temperature of the mixture and on the quantity used (a high temperature and large quantities result in shorter application times), although it also depends on the mixing conditions (with or without cooling, type of stirrer used), on the ratio of surface area to volume of the mixing vessel, and on the method used for end-point determination. Moisture-curing single-component adhesives also can have a working life, which begins with the contact of atmospheric moisture with the adhesive and which, apart from temperature, is largely dependent upon the existing humidity and upon the degree of moisture uptake during stirring operations.

Standards: comparisons of ASTM, MIL, GB, ISO, and DIN standards can be found in [25] and [26].

Legal Regulations. In view of the formulation of adhesives, particularly the solvent content, their manufacture, packaging, storage, transportation, and application are governed by a large number of regulations, which concern generally goods containing inflammable, explosive, or physiologically harmful constituents. Clean-water and clean-air regulations also must be observed.

4. Raw Materials for Adhesives

The adhesives industry uses many different products of natural and synthetic origin. These are mainly raw materials that are produced in large quantities and hence at reasonable prices. Raw materials that are used solely or primarily for adhesives are relatively rare or, alternatively, represent special types within fairly broad production programs.

Basic raw materials are natural and synthetic polymers or monomers and prepolymers which can form such polymers. The polymers primarily impart the required strength (cohesion) to the adhesive layer. In many cases, however, they have adequate adhesion properties of their own.

The adhesion properties may be improved by addition of resins or special adhesion promoters. Plasticizers and resins improve the flexibility of the polymers, increase the tack of the adhesives, or establish other required product properties, e.g., rheological properties. Rheological and other special properties can be influenced by adding fillers. Polymers are often dissolved in water or an organic solvent to achieve the wetting necessary for good adhesion.

The required service properties frequently can be obtained by using different basic raw materials or different kinds of formulations. Accordingly, the raw materials often cannot be assigned unequivocally to specific types of adhesive. Because almost every polymer and resin and also many other substances in principle may be used for the production of adhesives, it is not possible to provide a complete list of every possible raw material. For this reason, only the most important are listed in the following.

Starch [*9005-25-8*] (→ Starch). Although their significance has decreased, the various types of starch, e.g., mainly potato and wheat starch in Europe, corn starch in North America, and tapioca in Latin America, Africa, and Asia/Pacific region, still represent an important group of raw materials in quantitative terms. Starches are used in native form, in degraded, hot- or cold-soluble form, and as solutions.

In addition, they are used in combination with casein solutions and synthetic resin emulsions. Whereas starch ethers and esters and also aldehyde starches are used to only a limited extent in the adhesives field, the thermal degradation products of starches, the water-soluble dextrins, are still important in quantitative terms as raw materials for adhesives. They are used on their own in aqueous form or as mixtures with synthetic resin emulsions.

Proteins. Whereas vegetable protein is of virtually no significance, milk protein is still important, although increasingly less so, for casein glues. The same applies to collagen from animal skins and bones for glue jellies and glutins.

Cellulose Ethers and Cellulose Esters. Cellulose ethers (→ Cellulose Ethers; → Cellulose Esters), primarily methyl cellulose, are used as aqueous solutions (wallpaper paste) and also as thickeners.

Nitrocellulose [*9004-70-0*] (→ Explosives) is used as a raw material for solvent adhesives (multipurpose adhesives, shoe adhesives), although in recent years it has been increasingly overtaken in importance by other polymers.

Natural Rubber [*9006-04-6*] (→ Rubber, 2. Natural). Pressure-sensitive adhesives, contact adhesives, rubber solutions, and other solvent adhesives are made from natural rubber (generally first-latex crepe). Solutions of regenerated rubber also have some significance. Before dissolution, natural rubber generally has to be degraded by mastication on rolls or in kneaders and solubilized. Soluble types are also available. Natural rubber latices (rubber latex) having solids content of from 35 to 70 % are widely used for special latex adhesives (envelope gumming and the like). Compared with synthetic rubber, natural rubber has better tack but is more sensitive to heat and oxidation, particularly in contact with nonferrous metals.

Synthetic Rubber (→ Rubber, 3. Synthetic). Many different types of synthetic rubber are suitable raw materials for adhesives and sealing compounds. Particularly significant are polychloroprene rubber, styrene – butadiene rubber (SBR), nitrile rubber, and polyisobutylene. Unless these rubbers are available as directly soluble types, they have to be degraded by mastication on rolls or in kneaders and solubilized before dissolution.

Standard contact adhesives and pressure-sensitive adhesives are made from SBR in solution. Styrene – butadiene rubber latices, in some cases containing carboxyl groups, are used for special emulsion-based adhesives. Thermoplastic elastomeric block copolymers of styrene with butadiene or isoprene are of increasing significance in contact adhesives and hot-melt adhesives (→ Thermoplastic Elastomers, Chap. 5.).

Polychloroprene rubbers are the principal raw material for high-quality contact adhesives. Grades with high degrees of crystallinity are used preferably. Polychloroprene latexes are raw materials for aqueous contact adhesives.

Nitrile rubber is used preferably for contact adhesives with improved plasticizer resistance. Polyisobutylene is used in pressure-sensitive ad-

hesives. Butyl rubber is sometimes added to pressure-sensitive and hot-melt adhesives, although it is mainly used in sealing compounds. Epoxy resins and reactive (meth)acrylate adhesives (reactive adhesives) are modified with polychloroprene, butyl, and nitrile rubber. Chlorinated rubber is added in small quantities to contact adhesives and also to rubber-to-metal bonding agents for improving the adhesion properties.

Polyethylene [*9002-88-4*] (→ Polyolefins). Low molecular mass polyethylene, vinyl-modified polyethylene and polyethylene waxes are used for hot-melt adhesives.

Chlorosulfonated polyethylene can used in reactive adhesives in combination with (meth)acrylate monomers and also in rubber-to-metal bonding agents.

Ethylene – vinyl acetate copolymers are the most important polymer raw materials for hot-melt adhesives. They contain 18 – 45 % vinyl acetate and have melt flow indices of from less than 1 to 2500. The quantitatively most important copolymers contain ca. 30 % vinyl acetate and having melt flow indices of 40 – 400. Ethylene – vinyl acetate copolymer emulsions containing ca. 20 % ethylene are used in emulsion-based adhesives.

Polypropylene [*9003-07-0*] (→ Polyolefins) and various copolymers are used in atactic form for standard hot-melt adhesives.

Poly(vinyl esters) [→ Poly(Vinyl Esters)] are one of the most important groups of raw materials for adhesives. Poly(vinyl acetates) are important in solvent adhesives. Synthetic resin emulsions, based on poly(vinyl esters) with a solids content of 50 – 70 % are the principal raw materials for emulsion-based adhesives.

Besides poly(vinyl acetate) emulsions, emulsions of vinyl acetate copolymers with ethylene, (meth)acrylates, vinyl chloride, maleic esters and vinyl laurate are of considerable importance as internally plasticized synthetic resin emulsions.

Poly(vinyl acetate) emulsions for adhesives contain preferably poly(vinyl alcohol) as a protective colloid. Spray-dried poly(vinyl acetate) emulsions, known as redispersion powders, are used in combination with cellulose ethers as binders in floor-leveling compositions and for increasing solids content in emulsion-based adhesives. Poly(vinyl acetates) are used as solid polymers for formulating solvent adhesives.

Poly(vinyl chloride) [*9002-86-2*] homopolymers [→ Poly(Vinyl Chloride)] are processed primarily into plastisol adhesives and special solvent adhesives. Chlorinated PVCs (ca. 65 % Cl) are more readily soluble and are used for certain solvent adhesives.

Copolymers of vinyl chloride are used both as solid resins, for example, in solvent adhesives (vinyl chloride – vinyl acetate copolymers), and as additives in plastisol adhesives, but especially as synthetic resin emulsions for heat-sealing adhesives. Important copolymers are those of vinyl chloride with vinyl acetate, acrylates, or maleates, as well as copolymers of vinylidene dichloride with acrylates or acrylonitrile.

Poly(vinyl alcohol) [*9002-89-5*] (→ Polyvinyl Compounds, Others) is used primarily as a protective colloid for poly(vinyl acetate) emulsions. Relatively small quantities are added to modify emulsion-based adhesives. Aqueous solutions of poly(vinyl alcohol), in some cases combined with fillers and emulsions, are used as paper glues and as labeling adhesives.

Poly(Vinyl Acetals) (→ Polyvinyl Compounds, Others). Poly(vinyl formal) and poly(vinyl butyral) are used in solution as primers, in combination with phenolic resins for specialty reactive adhesives, and as films for the manufacture of laminated glass.

Acrylate and Methacrylate Polymers (→ Polyacrylates; → Polymethacrylates). Poly(ethyl acrylate) and poly(butyl acrylate) solutions and emulsions are important raw materials for pressure-sensitive adhesives. Copolymers of various esters, which give films of tailor-made hardness and which may additionally contain functional groups (carboxyl, amide, amino, methylol, hydroxyl), are used for pressure-sensitive adhesives to improve the adhesion properties or to enable the adhesive layer to be cross-linked to a limited extent.

Methyl acrylates are used as comonomers for increasing the hardness of the adhesive layer.

Acrylate – styrene copolymers are used as emulsions for coating paper.

Solvent adhesives and reactive adhesives are made from homo- and copolymers of methacrylates, generally methyl and ethyl methacrylate and, occasionally, butyl methacrylate. Monomeric (meth)acrylates are also used in reactive adhesive systems (polymerization adhesives). Poly(ethylene glycol) dimethacrylates are the basis of anaerobically curing liquid resins (reactive adhesives). They also are added as adhesion promoters to plastisol adhesives. Acrylate – ethylene copolymers, in some cases with a small content of carboxyl groups, are used instead of ethylene – vinyl acetate copolymers as fusible polymers for special hot-melt adhesives. Salts of polyacrylate and acrylate – acrylic acid copolymers are used as thickeners for aqueous adhesive solutions and emulsion-based adhesives.

α-Cyanoacrylic esters (above all methyl, ethyl, butyl, but also allyl and methoxyethyl esters) are used for one-component reactive adhesives.

Poly(vinyl ethers) [→ Poly(Vinyl Ethers)] are used for specialty pressure-sensitive adhesives.

Polyvinylpyrrolidone [*9003-39-8*] (→ Polyvinyl Compounds, Others) is used in adhesive sticks and in aqueous pressure-sensitive adhesives.

Polystyrene [*9003-53-6*] (→ Polystyrene and Styrene Copolymers) is used as a solution for bonding polystyrene and other plastics.

Polyamides and Polyaminoamides (→ Polyamides) obtained by the condensation of dimerized fatty acids with aliphatic diamines are used in solid form as hot-melt adhesives with particularly pronounced adhesion properties. The low molecular mass condensates, some of which are still liquid and contain free amino groups (amine equivalent ca. 3000 – 12 000), are used widely as hardeners for epoxy resins.

Copolyamides based on C_{12} and C_{14} amino acids or lactams are preferably used as heat-sealing powders for textile applications.

Polyesters (→ Polyesters). Unsaturated polyesters dissolved in styrene or methacrylates are used as reactive adhesives. Hot-melt adhesives can be made from saturated and fusible polyesters based on aliphatic and/or aromatic carboxylic acids, in some cases by using dimerized fatty acids. High molecular mass amorphous polyesters and low molecular mass crystalline polyesters are important raw materials for moisture cross-linking polyurethane adhesives in solution or solvent-free systems.

Low molecular mass polyesters and polycaprolactones containing hydroxyl groups are important raw materials for polyurethane adhesives.

Polyurethanes (→ Polyurethanes), by virtue of their considerable variability, have become very valuable to the adhesives industry over recent decades. They generally show above-average adhesion and cohesion properties. Low molecular mass reaction products of hydroxyl compounds and polyisocyanates, which contain cross-linkable hydroxyl or isocyanate groups, are used for various reactive adhesives. The hydroxyl components are aliphatic and aromatic polyesters, polyethers, castor oil, and even simple polyalcohols.

High-polymer polyurethane rubbers with various crystallization properties are used for solvent adhesives, in some cases in combination with resins and other polymers, or for contact adhesives. In many cases, solutions of this type are also used for heat-reactivatable adhesive coats. By adding polyisocyanates, it is possible to improve both the adhesion properties and the bond strength at elevated temperature.

Polyisocyanates (→ Isocyanates, Organic). Whereas any standard, generally difunctional, polyisocyanate, but preferably toluene diisocyanate (TDI) and methylenediphenyl diisocyanate (MDI), may be used for the production of polyurethane resins, only polyisocyanates with low vapor pressures at room temperature, such as MDI and its homologues, the reaction product of trimethylol propane with TDI, and triphenylmethane-4,4′,4″-triisocyanate, are recommended as hardeners for adhesives for reasons of industrial hygiene.

Nonaromatic isocyanates such as isophorone diisocyanate (IPDI), hexamethylene diisocyanate prepolymers, and dicyclohexylmethane diisocyanate prepolymers have also gained some

importance. Guidelines for the use of isocyanates are given in [27].

Epoxy resins (→ Epoxy Resins) are used primarily in combination with hardeners (polyamines, polyaminoamides, dicyanodiamide, acid anhydrides) for cold-setting and hot-setting reactive adhesives. Liquid epoxy resins of the bisphenol A type with epoxy equivalents of 170 – 300 are used preferably; relatively high molecular mass solid resins are used only occasionally.

In quantitative terms, novolak epoxy resins (epoxy equivalent 120 – 200) and elasticized or otherwise modified bisphenol A (F) resins have relatively little significance.

Epoxy resins are also combined with phenolic resins and poly(vinyl acetate) resins and used as adhesion promoters, for example, in plastisol adhesives.

Phenolic Resins [*9003-35-4*] (→ Phenolic Resins). Curable phenol – formaldehyde resins of the resol type are of considerable importance as hot-setting reactive adhesives for bonding wood and metals. Suitable types are used also in combination with poly(vinyl acetals) and epoxy resins for special, in some cases filmlike, reactive adhesives for bonding metals.

Standard and modified phenolic resins, alkyl phenol resins, etherified phenolic resins, and phenol – terpene resins also are used as primers and resin components in other types of adhesives, for example, in contact cements, pressure-sensitive adhesives, rubber-to-metal bonding agents, and similar adhesive systems.

Resorcinol Resins [*24969-11-7*] (→ Resins, Synthetic). Resorcinol – formaldehyde resins and resorcinol – phenol – formaldehyde resins are used as cold-setting reactive adhesives for waterproof wood bonds and as binders between tire cord and rubber.

Urea Resins [*9011-05-6*] (→ Amino Resins). In quantitative terms, urea – formaldehyde resins are one of the most important groups of raw materials for adhesives. They are used in the form of aqueous solutions or spray-dried powders, generally containing acid hardeners, as reactive adhesives for bonding wood, particularly for the production of plywood and chipboard.

Modified urea resins are also important for other purposes, for example for labeling adhesives and as cross-linking components.

Melamine – formaldehyde resins [*9003-08-1*] (→ Amino Resins), like urea resins, are used primarily for bonding wood. Because they are not particularly stable in aqueous solution, melamine – formaldehyde resins are marketed primarily in powder form.

Nonreactive resins with a variety of compositions are of considerable importance in the formulation of adhesives. In some cases, they are used on their own in solvent adhesives, so-called resin adhesives, but generally in combinations with polymers, in which they perform various functions. Above all, they increase tack, improve adhesion, influence viscosity, fluidity, and sealability, and, in some cases, also act as plasticizers. The most important applications for nonreactive resins are in pressure-sensitive adhesives, contact adhesives, hot-melt adhesives, solvent adhesives, and emulsion-based adhesives. In this context, the term resin covers materials differing very widely in their composition:

Rosin and its derivatives (→ Resins, Natural)
Tall-oil rosins and their derivates (→ Resins, Natural)
Hydrocarbon resins (→ Resins, Synthetic)
Carbamic ester resins
Cyclohexanone resins (→ Resins, Synthetic).

Plasticizers. (→ Plasticizers) are used in adhesive formulations, above all for elastification of the polymer films in emulsion-based adhesives and also for increasing wet tack. They are used, e.g., in homopolymer emulsion-based adhesives. Phthalate plasticizers are the most widely used, diisobutyl phthalate being especially important for emulsion-based adhesives. In addition to phthalate and adipate plasticizers epoxy and phenol sulfonic amide plasticizers are used in PVC plastisols. Besides ester plasticizers, mineral oils and also chlorinated low molecular mass polyethylenes and low molecular mass hydrocarbon resins are used in rubber adhesives.

Other plasticizers, such as benzoates, citric esters, and glycerol triacetate are less important. In hot-melt adhesives, mineral oil and low molecular mass polyolefins are used.

Solvents, primarily for solvent adhesives, but sometimes also added to emulsions in small quantities as temporary plasticizers, are mainly petroleum, toluene, xylene, dichloromethane, trichloroethylene, trichloroethane, ethyl acetate, acetone, 2-butanone, methyl isobutyl ketone, and cyclohexanone. The use of solvents has been reduced significantly by the introduction of solvent-free adhesives (e.g., polyurethanes) in several market segments. Aromatic solvents, such as toluene and xylene, and chlorinated products like trichlorethane and dichloromethane have been nearly eliminated or are used in closed systems.

Fillers are added to adhesives primarily for establishing certain consistencies and occasionally for reducing cost. Preferred fillers are pyrogenic and precipitated silicas, chalks, and light and heavy spar. Fibers and metal powders are used in special cases [28].

Other Additives. Small quantities of standard preservatives, antioxidants, UV/light stabilizers, colored pigments, and defoaming agents are added to many adhesives.

5. Classification of Adhesives

Adhesives can be classified on the basis of chemical composition [29], setting mechanism [30], and adhesion mechanism as pressure-sensitive, adhesion, and diffusion adhesives [31]. Lucke [32] proposed a flexible classification in which each adhesive is characterized and classified according to various characteristic features, such as chemical basis, form of application, application temperature, thermal behavior, and uses. A classification such as this provides for very detailed coverage of the individual features of an adhesive and is a helpful guide for industrial users. However, it is extremely comprehensive and goes beyond the scope of this treatise. For this reason, a classification based on the setting mechanism is given here [30].

1. Setting without Chemical Reaction. The high molecular mass bonding substance is present prior to bonding.

1.1. Application without Volatile Solvents. The adhesive film is obtained by the action of heat.

Hot-melt adhesives, e.g., ethylene – vinyl acetate copolymers, polypropylene, polyamides, polyesters. Uses: paper, fiberboard, plastics, textiles, leather (assembly bonds).
Plastisol Adhesives. PVC or poly(methyl methacrylate) powder + plasticizer + adhesion promoter. Uses: metals (sheet-metal constructions), silicate-containing materials.

1.2. Application of Solutions, Solvent Escapes before Bonding. Adhesive layer is heated before or during bonding.

Heat-Sealing Adhesives and *High-Frequency Welding Auxiliaries,* e.g., vinyl polymers. Uses: paper, plastics, lamination of plastic films to metal foils (packaging, metal-foil lamination).
Contact Cements: elastomers (e.g., polychlorobutadiene) in conjunction with adhesive resins (e.g., alkylphenol or modified phenolic resins); bonding is obtained by contact under pressure between two almost dry adhesive layers. Uses: wood, plastics, rubber, metals.
Pressure-Sensitive Adhesives, e.g., natural and synthetic rubbers in conjunction with modified natural resins and/or synthetic resins; also polyacrylates, poly(vinyl ethers). Uses: nonspecific adhesion to virtually any surfaces, more particularly for the permanently tacky coating of tapes, films, and labels.

1.3. Application of Solutions, Solvent Evaporates during Bonding.

Solutions of polymeric, natural, and/or synthetic substances in organic solvents, e.g., vinyl polymers, natural rubber, synthetic rubbers of low polarity. Uses: paper, plastics, silicate-containing materials, wood.
Solutions of natural and synthetic high molecular mass substances in water, e.g., starch, dextrins, casein, cellulose ethers, water-soluble derivatives of poly(acrylic acid), poly(vinyl alcohol), poly(vinyl pyrrolidone) (adhesive sticks). Uses: paper, fiberboard.
Glutins (glues of animal origin). Uses: wood, paper, fiberboard, moistenable adhesive tapes.

1.4. Emulsions of Water-Insoluble Substances in Water that Escapes during Bonding.

Aqueous emulsions of polymeric compounds, e.g., homopolymers of vinyl acetate and vinyl propionate, vinyl acetate copolymers with ethylene or maleic esters; polyacrylic esters, styrene copolymers. Uses: paper, wood, also bonding to plastics.

2. Setting by Chemical Reaction (Reactive Adhesives). The high molecular mass bonding substance is formed during bonding.

2.1. Formation and Cross-linking of the Adhesive Film by Polymerization.

Two-pack polymerization adhesives: unsaturated polyesters with monomeric vinyl compounds, e.g., styrene or methyl methacrylate. No-mix adhesives and A/B acrylates, applied without premixing. Uses: metals, plastics, silicate-containing materials.
One-pack polymerization adhesives:

a) Cyanoacrylate adhesives setting under the influence of moisture or alkalinity. Uses: metals, plastics, silicate-containing materials, rubber, wood (high-strength spot bonds or small-area bonds).
b) Anaerobic adhesives. Special methacrylate adhesives. Uses: metals (threadlocking and bearing retention).

2.2. Formation and Cross-linking of the Adhesive film by Polyaddition.

Epoxy resin adhesives, polyepoxy compounds in combination with amines, polyamidoamines, or dicarboxylic acid anhydrides. Uses: metals, silicate-containing materials, plastics.
Reactive polyurethane adhesives: polyurethane prepolymers with terminal hydroxyl groups or terminal isocyanate groups. Uses: plastics, metals, silicate-containing materials.
Two-component silicone adhesives: setting induced by catalyst.

2.3. Formation and Cross-linking of the Bond by Polycondensation (elimination of low molecular mass cleavage products during setting).

Polyhydroxymethyl compounds: phenol – formaldehyde resins, also in combination with poly(vinyl formal) resins, nitrile rubber or epoxy resins (as adhesive film), urea – formaldehyde resins, melamine – formaldehyde resins, resorcinol – formaldehyde resins. Uses: wood (plywood manufacture), metals.
Silicone adhesives, one-pack, setting under the influence of atmospheric moisture. Uses: glass, metals, ceramics, plastics; some substrates need a priming coat.
MS polymers, siloxane-alkoxy-modified prepolymers of poly(propylene glycol), one-pack systems. Uses: Many surfaces; no primer necessary.
Polyurethane adhesives, one-pack, setting under the influence of atmospheric moisture. Uses: plastics, metals, silicate-containing materials.
Reactive polyurethane hot-melt adhesives, one-pack, setting under the influence of atmospheric moisture. Uses: bookbinding (high-quality perfect bindings), wood gluing, shoe manufacturing.
Polyimides and polybenzimidazoles. Uses: metals (for high thermal stressing).

2.4. Cross-linking of Polymeric Adhesive Components in the Bond. Rubber-to-metal bonding agents, e.g., halogenated polymers in conjunction with cross-linking agents; adhesion is obtained under vulcanization conditions (elevated temperature and pressure). Uses: natural and synthetic rubber to metals (structural elements for damping vibration).

2.5. Ultraviolet/Electron Beam (UV/EB) Curing Adhesives. Special acrylic acid esters (monomers or oligomers). Uses: for laminating and for pressure-sensitive products.

2.6. Conductive adhesives include electrically and thermally conductive materials. Most fillers that impart electrical conductivity also contribute to thermal conductivity, but thermally conductive adhesives that are electrically insulative are also available.

6. Individual Adhesive Systems

The same classification used in Chapter 5 is used here. Information on the uses of the adhesives discussed here can be found in Chapter 5.

6.1. Adhesives That Set Without a Chemical Reaction

6.1.1. Solvent-Free Adhesive Systems

Hot-melt adhesives are 100 % solids that, in the broadest sense, include all thermoplastic polymers. Polymers that are primarily used as hot-melt adhesives include ethylene – vinyl acetate copolymers (EVA), polyvinyl acetates (PVA), polyethylene (PE), amorphous polypropylene, block copolymers (thermoplastic elastomers), polyamides, and polyesters. The oldest hot-melt adhesive, which has been in use since early times, is sealing wax. In principle, glutins and glue jellies also may be regarded as hot-melt adhesives. However, modern hot-melt adhesives are primarily synthetic products. The simplest hot-melt adhesives are rosin – wax mixtures. But these products have limited strength and thermal stability.

In general, hot-melt adhesives are solid below 80 °C. Ideally, as the temperature is increased beyond this point, the material rapidly melts to a low-viscosity fluid that can be easily applied. Upon cooling, the adhesive sets rapidly. Because these adhesives are thermoplastics, the melting/resolidification process is repeatable. Typical application temperatures of hot-melt adhesives are 150 – 190 °C with melt viscosities in the range of 500 – 3000 mPa s.

Hot-melt adhesives do not undergo a chemical reaction because they simply cool from the melted state to form a solid film at the bond line. These adhesives are available in forms such as pellets, slabs, bars, slugs, and films that allow convenient handling by a variety of application equipment. Because hot-melt adhesives are applied in the molten state, melt viscosity is an important property. The limitations of the application equipment often influence the viscosity range that is selected for a hot-melt adhesive formulation.

Although hot-melt adhesives consist of 100 % solids, they are rarely 100 % polymers in composition. This is due to the limited adhesion of pure thermoplastics like EVA copolymers and their lack of melt properties such as tack and wettability. The components of a hot-melt adhesive can be roughly divided into two categories: polymers and diluents. Typically, diluents are waxes, plasticizers, tackifiers, stabilizers, extenders, and pigments. The functions of the diluents are:

Lowering the viscosity for easy application
Enhancing wettability
Enhancing adhesive strength
Increasing rigidity (extenders) or flexibility (plasticizers)

It is desirable to have a hot-melt adhesive with high strength. But this property is generally accompanied by a high melt viscosity. A low melt viscosity is desirable for application purposes. Therefore hot-melt adhesive formulations must include diluents that decrease the polymer melt viscosity and increase the overall adhesive strength.

Special dispensing equipment must be used to apply hot-melt adhesives. Typical application methods include the use of rollers, screw extruders, and squirting pumps or nozzle applicators.

The polymers are generally of high molecular mass and impart strength and high viscosity to an adhesive. Waxes such as paraffin wax can function as both diluents and antiblocking agents. In addition, waxes promote surface wetting. Similarly, plasticizers such as phthalates, mineral oils, and glycolates provide both surface wetting and adhesive flexibility. Typical tackifiers include rosin, modified rosin, terpenes, modified terpenes, hydrocarbons, and chlorinated hydrocarbons. These materials provide tack and flexibility while promoting surface wetting and adhesion. Stabilizers such as hindered phenols help to maintain adhesive melt viscosity while functioning as antioxidants. Extenders such as talc, clay, and barites lower the cost of hot-melt adhesives and simultaneously help control melt flow.

The following properties of hot-melt adhesives are important:

Melt viscosity is one of the most important properties of a hot-melt adhesive. In general, as the temperature of a polymer increases, its viscosity decreases. Therefore, in a hot-melt adhesive formulation, the melt temperature controls

the viscosity, which greatly influences the extent of surface wetting. The temperature of the melt and the application equipment should be maintained as constant as possible.

The *bond formation temperature* is the minimum temperature below which surface wetting is inadequate. A hot-melt adhesive is applied at a running temperature at which the viscosity is sufficient to wet surfaces properly. Running temperatures that are too high allow more time for surfaces to be wet properly. However, if low penetration is required and if the substrates are porous, this longer time may be detrimental and may result in a bond that contains insufficient adhesive. A running temperature that is too close to the bond-formation temperature may result in a too-rapid solidification of the adhesive bond. As a result, overall properties such as strength may not be optimized.

The *heat stability* of the adhesive is a very important property. Running temperatures that are greater than the adhesive degradation temperature result in charring and decreased overall properties. The addition of stabilizers to the adhesive formulation contributes to the stability of the material by hindering the acceleration of degradation due to the presence of oxygen.

The *tack* of the adhesive indicates the stickiness of the hot-melt as it changes from a liquid to a solid state. This property affects the ability of the adhesive to hold the substrates together.

The *open time* is the lapse of time between the application of the hot-melt to one substrate and the loss of its wetting ability on the second substrate due to solidification. It typically ranges from fractions of seconds to minutes to infinity (in pressure-sensitive hot-melt adhesives).

The *limitations* of hot-melt adhesives are restricted toughness at usable viscosities, low heat resistance, and poor creep resistance. All properties are affected by the polymer and diluents used in the formulation. The application conditions, including the amount of adhesive and the pressure applied to the bond line, also affect end-use properties. In the hot-melt adhesive-bonding operation, a minimum amount of pressure must be applied until the hot-melt becomes solid or sufficient tack develops to hold the substrates in place.

Manufacture. In the manufacturing of hot-melt adhesives, the components are melted in heated stirring vessels and subsequently formed in pelletizing or granulating machines. The melt is allowed to drop onto a cooling belt or to form a film on a cooling belt and subsequently cut. Another possibility is extrusion from an extruder followed by underwater granulation. For manufacturing hot-melt adhesives with higher viscosity and higher filler contents, it is preferable to use heatable kneaders or extruders followed by specialized processing units. Hot-melt sticks (slugs, rods) for application by handguns are produced by injection molding or by cutting hot-melt profiles.

Surface-tacky hot-melt adhesives are marketed in containers or in the form of blocks or cubes in readily removable packs (silicone-coated paper) [33] or in meltable polyethylene bags.

Hot-Melt Chemical Families. EVA copolymers are the cheapest hot-melt adhesives and are used in the greatest quantities. Typical properties of these adhesives include sufficient strength between ca. 30 and 50 °C, but limited upper service temperature of 60 – 80 °C, low hot-melt viscosity, and low creep resistance under load with time. Most EVA-based hot-melt adhesives are used to bond paper, fabrics, wood, and some thermoplastics.

PP polymers based on amorphous polypropylene have better properties than EVA hot-melt adhesives. In particular, higher heat resistance and better adhesion properties can be obtained.

Polyamides and thermoplastic polyesters are classified as high-performance hot-melt adhesives. These adhesives have greater strengths than EVA-based hot-melt adhesives and exhibit adequate bonding from approximately − 40 to 70 °C. Some formulations have upper service temperatures up to 185 °C for applications that do not involve loads. One disadvantage of the higher service temperatures is the higher hot-melt application temperature, which may require special equipment. Both polyamides and polyesters are sensitive to moisture absorption during application. This can result in hot-melt foaming, which leads to voids in the solidified adhesive and decreased strength. Nonetheless, polyamides exhibit good strength and flexibility. Polyesters are even stronger, but are more rigid and are usually highly crystalline. This results in a sharp melting temperature, which is desirable for high-speed bonding. Polyamide and polyester hot-melt adhesives are used to bond

plastics, glass, wood, leather, foam, fabric, rubbers and some metals, such as aluminum, copper, and brass.

Polybutene-based hot-melt adhesives are tough, partially crystalline, and their slow crystallization rates lead to long open times. Copolymers of butene result in softer and more flexible adhesives. In general, polybutene and its copolymers have low temperatures for recrystallization from the melt. This permits stress release in the adhesive bond, which may have been applied to cold surfaces. Polybutene and its olefinic copolymers exhibit good bonding to nonpolar surfaces but poor compatibility with polar substances. These hot-melt adhesives have been used on rubbery substrates and are available as pressure-sensitive adhesives.

Thermoplastic elastomers that are formulated into hot-melt adhesives include polyurethane and block terpolymers, such as styrene – butadiene – styrene; styrene – isoprene – styrene; and styrene – olefin – styrene, in which the olefin component is typically ethylene, propylene, and/or butylene. The saturation in the mid-segment of these terpolymers results in better UV and thermooxidative resistance than that of unsaturated butadiene and isoprene midsegments.

In general, thermoplastic elastomer hot-melt adhesives are not as strong as polyesters. However, they exhibit good flexibility, high elongation, toughness, and vibration resistance. Polyurethanes and block terpolymers are used as hot-melt pressure-sensitive adhesives in tapes and labels and for nonwoven applications.

Pressure-sensitive hot-melt adhesives are surface-tacky and can be used by the application of pressure alone. These adhesives have infinite open times and are used at room temperature. Typically, pressure-sensitive hot-melt adhesives are supplied in form of blocks or cubes in silicone coated packaging units or preferably in polyethylene bags, which after melting form part of the adhesive.

Foamable hot-melt adhesives have been commercially available since 1981. Nitrogen or carbon dioxide is introduced into the hot-melt adhesive and results in a 20 – 70 % increase in adhesive volume. Foaming increases hot-melt spreading and open time. This method is usually used with polyethylene hot-melt adhesives that are applied to selected metal, plastic, paper, porous, and heat-sensitive substrates.

Low-melt adhesives are hot-melt adhesives with an application temperature in the range of 100 – 120 °C instead of 160 – 190 °C. They are used for bonding heat-sensitive materials.

Plastisol adhesives are solvent-free adhesives that require temperatures of 120 – 200 °C for setting. They consist essentially of a dispersion of finely divided poly(vinyl chloride) or polymethacrylate [34] in plasticizers and low molecular mass, heat-reactive substances that act as adhesion promoters, for example, epoxy resins, poly(ethylene glycol) dimethacrylates, phenolic resins; in some cases these are even used in conjunction with corresponding hardeners. Standard heat stabilizers for poly(vinyl chloride) are also added to prevent decomposition of the poly(vinyl chloride) and evolution of HCl during hardening or in the event of subsequent exposure to heat. Depending on the end use, varying amounts of fillers, pigments, or thixotropizing agents may be added to modify the color and fixing behavior. At a certain setting or hardening temperature, the poly(vinyl chloride)/polymethacrylate plasticizer mixture becomes a dispersion gel and also adheres to the substrate through reaction of the adhesion promoter. The various adhesion promoter systems require different activation temperatures, usually in the range of 120 – 200 °C [34].

Production and Storage. Because of their relatively high viscosities, these pasty to highly viscous compositions must be produced in kneader-type stirring vessels capable of applying considerable forces. It is important to ensure that stirring and kneading do not generate excessive temperatures, which can result in gelation of the product. For the same reason, the products must be stored under cool conditions (below 30 °C).

6.1.2. Adhesive Solutions from Which the Solvents Escape before Bonding

This group includes heat-sealing adhesives and high-frequency (HF) sensitive heat-seal coats; these are coating adhesives that are heat-activated during bonding. Contact adhesives and

pressure-sensitive adhesives bond by the same principle but without the action of heat.

Heat-sealing adhesives are applied to the materials to be sealed in the form of solutions, emulsions, or, preferably, in melt or powder form. The solvent-free layer is intended to be nontacky and only to melt and to wet the adherent under the action of heat during sealing. Solidification takes place after cooling. In many cases, sealing is carried out on a coating-to-coating basis.

The choice of raw materials for sealing coatings is very wide and is influenced by the sealing temperature and the sealing behavior of the materials used, by the requirements the joint has to satisfy, and by the adhesion of the sealing layers to the substrates. In many cases, copolymers of vinyl chloride or vinylidene chloride are used in solution, although they are also used in emulsion in conjunction with resins and other polymers. In addition, copolymers of vinyl acetate and polymethacrylates, polyurethanes, and polyesters are also used. Ethylene – vinyl acetate copolymers are used preferably for heat-sealing coatings applied as melts. Heat-sealing powders also are used occasionally; they can be based on fine polyamide or polyethylene powders and also on vinyl acetate – vinyl chloride copolymers.

High-frequency-sensitive heat seal coats are a special type of heat-sealing adhesive [35]. They are also applied in the form of a solution or emulsion to the surface of a substrate and dried. However, welding to the partner substrate is not obtained by the application of heat through the substrate; instead the sealing temperature is generated by high frequency within the adhesive layer itself. Hence, the adhesive layer must have a high dielectric loss factor.

High-frequency sealing auxiliaries generally contain vinyl chloride copolymers in addition to proportions of other polymers, such as polyacrylates, vinyl acetate copolymers, plasticizers, and resins. High-frequency heatable coatings have solids contents of from 30 % (solutions) to 50 % (emulsions). Their activation temperature is in the range of 120 – 150 °C. High-frequency sealing powders consist predominantly of finely divided polyamides.

Contact cements must be applied to both substrate surfaces to be joined. The adhesive layers are joined only when the solvents have almost completely evaporated. A brief, but as high a contact pressure as possible is required for bonding. The strength of the adhesive joint is relatively high immediately after contact bonding and may amount to 50 % of the final strength. In principle, the contact adhesion effect may be obtained with natural rubber solutions, as known, for example, from the repair of bicycle inner tubes with rubber solution. However, the actual contact cements are solutions of synthetic rubbers with resins or solutions of high molecular mass polyurethane elastomers. The rubber components are mainly polychloroprene, nitrile – butadiene, and styrene – butadiene rubbers, and the resins used are primarily phenolic resins, rosins, and hydrocarbon resins. In the case of polychloroprene adhesives, metal oxides are added as stabilizers. In some cases, the rubber starting materials are high molecular mass rubbers that must be masticated on rolls or in internal mixers before dissolution, but mainly directly soluble rubbers are used today. The rubber, resin, and additives are dissolved in the appropriate solvents or solvent mixtures in closed stirring vessels to form the end product. The solid contents of the products varies from 20 to 30 %; the more sprayable types have lower solids contents.

The highest strengths are obtained with a uniform, thin coating of adhesive coupled with a high contact pressure. It is important to allow for adequate evaporation of the solvents and to observe the open time of the contact adhesive.

Contact cements based on nitrile rubber and polyurethane are more resistant to oils and plasticizers than polychloroprene adhesives. To improve adhesion and to increase thermal stability, contact adhesives may be applied together with isocyanate hardeners. Today solvent-free contact adhesives based on aqueous polychloroprene latexes are also available.

Trade Names. In Germany: Pattex, UHU Greenit; in Japan: Cemedine, Bond G, Sunstar; in the USA: Pliobond, Fastbond, Elmer's.

Pressure-sensitive adhesives are permanently tacky substances that adhere spontaneously to the surfaces of most materials with only light pressure [36]. Their adhesiveness is

relatively nonspecific, although the individual products show different adhesion values on different substrates. Pressure-sensitive adhesives are used mostly for coating supports, for example, paper and film tapes, adhesive labels, or self-adhesive decorative sheeting. The adhesives are applied to the supports as solutions, dispersions, or melts. After the adhesive has set, the pressure-sensitive adhesive is present as a thin film on the support and, until it is used, is covered by suitable antiadhesively finished materials, for example silicone paper or siliconized polyethylene film. In some cases, the pressure-sensitive adhesive is also applied to the masking material, and the film of pressure-sensitive adhesive formed is transferred to the actual carrier by the reverse process. The dry application weight varies between 4 and 40 g/m^2, depending on the end use and the roughness of the carrier or substrate to be bonded.

In general, bonds formed with pressure-sensitive adhesives show relatively high resistance to brief and rapid loads. Under permanent load or stress, however, they tend to creep, particularly at elevated temperature. The property spectrum of pressure-sensitive adhesives is characterized by adhesiveness and resistance to heat, aging, and plasticizers.

The raw materials used for pressure-sensitive adhesives are natural and synthetic rubbers in conjunction with modified rosins, phenol – formaldehyde resins, or hydrocarbon resins. In addition to rubber, polyacrylates, polymethacrylates, poly(vinyl ethers), and polyisobutenes also are used frequently, again mostly in combination with resins. Silicone resins are used for special applications.

Pressure-sensitive adhesive dispersions are based primarily on special acrylic ester copolymers, again generally in combination with resins. In addition to suitable resins, two polymer bases are used primarily for hot-melt pressure-sensitive adhesives, namely, ethylene – vinyl – acetate copolymers and styrene – butadiene or styrene – isoprene block copolymers, which also are known as thermoplastic rubbers.

It is said that a pressure-sensitive adhesive always must be composed of a high-polymer base resin, which determines cohesion and specific adhesion, and of tackifying resins (tackifiers). In some systems, the tackifier may be replaced by low molecular mass fractions of the base polymer. To increase cohesion, the base resin in some systems is cross-linked or, in the case of rubber-based pressure-sensitive adhesives, vulcanized after application.

6.1.3. Adhesive Solutions from Which the Solvents Evaporate during Bonding

Solutions of Polymeric Substances and Resins in Organic Solvents. A distinction is be drawn between adhesives that set by evaporation of the solvent and form an adhesive joint by adhesion of the dissolved polymer to the substrate surfaces, and adhesives for plastics, the solvent of which dissolves and swells the plastics surfaces and joins them by means of migration processes. This allows bonding of plastics that are otherwise difficult to join. The polymer solutes in solution adhesives act primarily as thickeners for establishing certain flow properties and retaining the solvent for the duration of the dissolution process. These solid components also may perform a limited gap-filling function.

Adhesion Adhesives. The best known solvent-containing adhesives are the *multipurpose adhesives*. They can not join every type of substrate, but can bond a wide range of substrates with different results. They contain as binders high molecular mass compounds that adhere to many different surfaces, preferably nitrocellulose and poly(vinyl acetate). The solvent mixtures consist of esters and ketones and, in some cases, also contain alcohols. The composition of the solvent is determined by the polymers used, by the required fluidity of the adhesive, and by the required open time or setting time. The solvents may dissolve certain plastics.

Solutions of natural and synthetic rubber, polyurethane rubber, ethylene – vinyl acetate copolymers, polyacrylates, and other polymers with additions of resins and plasticizers are also used. These polymers can be used in solvents or as dispersions in water.

Trade Names.In Germany: UHU Alleskleber, Pritt Alleskleber; in Japan: Saivinol, Cevien-A; in the USA: Elmer's School Glue, Wilhold Glu-Bond.

Solvent Adhesives. In solvent adhesives for bonding plastics, the composition of the solvent

is particularly important with regard to dissolution and swelling of the plastics to be bonded. Synthetic resins from the same class or from a related class are used for thickening.

The PVC adhesives have been used since the 1930s for bonding PVC resins. They are solutions of chlorinated poly(vinyl chloride) in acetone, cyclohexanone, THF, or mixtures thereof. The THF-containing adhesives are solutions of PVC in mixtures of THF, 2-butanone, and cyclohexanone. Organic thickeners are used to obtain pseudoplastic solutions that can be applied in sufficiently thick layers. By virtue of their stronger dissolving and swelling effect, THF adhesives can bridge certain gaps.

Pure solvents, such as aromatic hydrocarbons, halogenated hydrocarbons, solvent mixtures, or even solutions of polystyrene in these solvents are used for bonding polystyrene, ABS, and other styrene copolymers.

Trade Names. Solution adhesives for rigid PVC: Tangit, Dytex (Germany); Hamatide, S-dine (Japan); Bostik, Datey (USA).

Solutions of Natural and Synthetic Polymers in Water.

Aqueous Starch- and Dextrin-Based Adhesives. The starch- and dextrin-based adhesives reached their zenith between 1920 and 1935. Since then, they have been increasingly superseded by synthetic resin-based adhesives, and they are now important only in certain applications. However, they are still used in considerable quantities for applications such as labeling adhesives, paper-sack adhesives, and envelope gums.

The raw materials used for the various types of starch- and dextrin-based adhesives are chiefly potato starch and cornstarch and, to a far lesser extent, tapioca starch and wheat starch. Potato starch is preferred for liquid glues, at least in Europe, where it is widely available.

Native *starch* is not soluble in water. To obtain soluble starches, the starch granule must be gelatinized by the prolonged heating of an aqueous starch suspension above the gelatinization temperature (e.g., 62.5 °C for potato and cornstarch, 67.5 °C for wheat starch) or by treatment with alkali at lower temperature. The gelatinization temperature and also the gelatinization rate can be varied by adding electrolytes.

To obtain certain solution viscosities and structures, the starch molecule must be degraded before and after gelatinization. A starch suspension is exposed before gelatinization to the action of acids, primarily nitric acid, or oxidizing agents, such as hydrogen peroxide, at elevated temperature. The greatest degradation is carried out in the case of dextrination. Because of the different degrees of degradation, the various gelatinization processes, and additives that influence hydration, the adhesives obtained differ considerably in their rheology, adhesive strength, and application properties; they are marketed as ready-to-use adhesives and as cold-water- or warm-water-soluble powdered adhesives. All aqueous starch- and dextrin-based adhesives must be protected against biological degradation by additions of preservatives [37].

Dextrin-based adhesives are available as powdered dextrins, which are dissolved in water before application, and ready-to-use aqueous dextrin adhesives with a dextrin content of 50 – 70 % and viscosities of 3×10^4 to 3×10^7 mPa s.

Dextrins are produced by the thermal and acid-hydrolytic degradation of native starch, generally cornstarch and potato starch. Dextrin-based adhesives are produced by dissolving powdered dextrins or dextrin mixtures of differing degrees of degradation in water. Mixtures of 60 – 70 % dextrin and 30 – 40 % water are heated in glass-lined vessels to 80 – 100°C until a homogeneous solution is formed. Commercial dextrin-based adhesives differ in the type of dextrins or dextrin mixtures used, the solids content, and in the additives present, such as hygroscopic salts ($CaCl_2$, $MgCl_2$), borax, glycerol, sugar or sugar syrup, and similar substances).

The viscosity of dextrin-based adhesives can be reduced by vigorous stirring. However, the viscosity of dextrin-based adhesives generally undergoes a marked increase after production, so that they have to be left to mature. In most cases, dextrin-based adhesives are marketed only after the maturing process is complete.

Mixed Glues. Starches and dextrins also are used in combination with other raw materials in adhesives. For example, the resistance to water of starch-based adhesives can be improved by the addition of melamine – formaldehyde or urea – formaldehyde resins. Mixtures of starch-

based adhesives or dextrin-based adhesives with synthetic resin dispersions are also used.

Casein glues are also still important. Casein is obtained by precipitation from milk protein. The dried powder is mixed with water prior to application and hardens by loss of water at room temperature, with some chemical conversion of protein to more insoluble calcium derivatives. This adhesive is unsuitable for outdoor use but is more resistant to temperature changes and moisture than other water-base adhesives. Casein adhesives tolerate dry heat up to 70 °C and have good resistance to organic solvents. Although they are susceptible to biodegradation, chlorinated phenols can be used to inhibit this behavior. Typical applications include packaging (paper labels for glass), woodworking, and fabrication of interior-grade plywood.

Other Aqueous Adhesive Solutions. Besides starches and dextrins, *cellulose ethers,* primarily methyl cellulose and carboxymethyl cellulose, and in some cases also *starch ethers,* are important as wallpaper pastes and poster glues. The products are generally marketed in the form of powders that are dissolved in water by the user. Aqueous solutions of *poly(vinyl alcohol)* can be used to a limited extent. In most cases, fillers are added to these solutions.

Poly(vinyl pyrrolidone) has acquired significance as a binder in adhesive sticks. In addition to the polymer, these sticks consist primarily of a soap/water gel as builder. After application, the water evaporates, leaving the solid behind as bonding agent [39].

Aqueous Glutins (Animal Glues). The significance of animal glues or glutins has steadily diminished in the last decades. Like gelatin, glutines are obtained by acidic or alkaline hydrolysis of the collagen present in animal bones, hides, etc. (→ Gelatin). The degradation product, which forms colloidal solution in water, is called glutin. Particularly pure glues are known as gelatin glues. In some cases, gelatin waste is also made up into glues.

The glutin solutions that accumulate in the processing of bones and hides are concentrated to a solids content of 30 – 55 %. They solidify on cooling to form firm, cuttable jellies. The slabs or blocks are further reduced in size, dried to a solids content of 85 %, and marketed in various forms. Hide glues are stronger than bone glues and often exceed the strength of wood. These adhesives are primarily used to bond wood, leather, paper, and textiles. Use is limited to interior applications. The setting of glutins is determined primarily by solidification of the gelatin or glutin/water mix on cooling and secondarily by evaporation of the water.

The reversible uptake and release of water by glutins has advantages and disadvantages. The possibility of activating layers of glue jellies by moistening is utilized, for example, in the production of moistenable gum tapes [40].

6.1.4. Aqueous Emulsions of Polymeric Compounds

Emulsion-based adhesives are adhesives that contain aqueous synthetic resin dispersions or emulsions or rubber latices as their principal raw material or active binder [41]. They are milky white, odorless liquids with solids contents of 40 – 60 %. In quantitative terms, emulsion-based adhesives are one of the most important groups of adhesives. They set by the release of water and the formation of a film from the emulsified or dispersed resin or rubber droplets.

The number of resin emulsions and rubber latexes suitable for use as raw materials for adhesives is very large. Emulsions of homopolymers and copolymers of vinyl acetate are the most important. The comonomers for vinyl acetate include maleic esters, acrylates, ethylene and vinyl chloride, unsaturated carboxylic acids, and vinyl esters of relatively long-chain fatty acids, such as vinyl laurate. Apart from these products, polyacrylate homopolymers and copolymers and also styrene copolymers are particularly important in adhesives [42].

Additions of plasticizers, solvents, and resins improve the specific adhesion of the synthetic resin emulsions, and also are capable of lowering the film-forming temperature and increasing the tackiness of the wet adhesive. Poly(vinyl alcohol), cellulose ethers, and other hydrocolloid additives prolong the open time. Poly(vinyl alcohol) also can improve wet tack. Certain applications require fillers, which may be used in such large quantities that pasty mortars and cements are formed.

The elasticity of the set films can be increased by plasticizers and by copolymerization. Redispersion powders, that is, spray-dried synthetic

resin emulsions, are added to synthetic resin emulsions to increase their solids content and their setting rate.

Synthetic resin emulsions containing reactive groups also have appeared on the market. The reactive groups provide for subsequent cross-linking, making it possible, for example, to increase the resistance of the joints to water. Mixed adhesives consisting of synthetic resin emulsions and starch- or dextrin-based adhesives have advantages in certain fields and are generally cheaper.

Latex adhesives are based on natural rubber or on various types of synthetic rubber. Occasionally, mixtures of natural rubber and synthetic resin latices are also used. In general, resins or solvents also are used in the production of latex adhesives.

In the production of latex adhesives, overvigorous stirring can affect the rheological properties of the adhesives, which are often pseudoplastic and differ in their mechanical stability. Characteristic properties of emulsion-based adhesives include viscosity and flow properties, the setting and breaking behavior of the emulsion, the wet tack or green strength of the film, the minimum film-forming temperature, storage behavior at low temperatures (freezing temperature), and specific adhesion to various materials. Films of emulsion-based adhesives are tested only occasionally for tensile strength and elasticity. The redispersibility of the set adhesive films is important with regard to the cleaning of machine components and recycling of paper.

6.2. Adhesives Setting by Chemical Reaction

Before setting, reactive adhesives consist predominantly of reactive low molecular mass monomers and/or oligomers that, during curing, are converted by chemical reactions into high molecular mass, often three-dimensionally cross-linked polymers. Reactive adhesives can be divided into polymerization, polyaddition, and polycondensation adhesives. Vulcanizing reactive adhesives are a special group.

Reactive adhesives are marketed as one-pack, two-pack, and no-mix formulations. Two-pack adhesives must be completely mixed in the correct ratio before use. The mixtures have a limited pot life. One-pack adhesives are easier to handle during application. They are hardened by heat, by the catalytic action of the substrate, or by atmospheric moisture.

Cold-setting reactive adhesives harden within a few hours to several days at room temperature or slightly elevated temperature. An exception are cyanoacrylate adhesives, which cure on several substrates within seconds or minutes, as do UV/EB curable adhesives. Warm-setting adhesives generally require temperatures in the range 80 – 100 °C, and hot-setting adhesives 100 – 250 °C. Adhesive joints of particularly high quality are obtained mainly with warm- and hot-setting systems. In many cases, warm setting leads to higher bond strengths, even in the case of systems that normally harden at room temperature. However, even certain cold-setting adhesives produce outstanding bond strengths under favorable bonding conditions.

To avoid mixing errors in the case of two-pack adhesives, no-mix adhesives are available. In this case, one of the surfaces to be bonded is pretreated with the adhesive resin, the other with the adhesive primer, and the bond is formed instantly when the two surfaces are brought together.

6.2.1. Adhesives Setting by Polymerization

Polymerization adhesives harden through radical or ionic polymerization of the monomers; at the same time, graft polymerization or cross-linking of the dissolved, still polymerizable polymers also may occur.

Two-Component Polymerization Adhesives. The best known types are polymerization adhesives based on solutions of *unsaturated polyesters* in styrene or (meth)acrylates. Peroxides are added as hardeners to the resin component; amines or heavy-metal salts are used as accelerators.

More important and with better properties are the cold-setting, two-component *acrylate adhesives,* which contain methacrylates or acrylates, sometimes mixed with styrene and methacrylic acid as monomer and, in addition, various polymers. The polymers used are primarily synthetic rubbers, such as polychloroprene, styrene – butadiene rubber, butyl rub-

ber, polystyrene, polymethacrylates, and acrylate graft polymers of these polymers. Amines are used as accelerators, and benzoyl peroxide in the form of plasticizer pastes or a powder mixture with fillers is preferred as hardener [43].
Trade Names.In Germany: Agomet, Pattex Stabilit Express; in Japan: Hard-Lock, Sumikitt; in the USA: Versilok.

One-Component Polymerization Adhesives.

Cyanoacrylate Adhesives. Methyl, ethyl, butyl, and methoxyethyl esters of cyanoacrylic acid are used for cyanoacrylate adhesives [44]; soluble polymers and plasticizers are incorporated to regulate viscosity and for elastification. Cyanoacrylate adhesives rapidly polymerize by an ionic reaction mechanism initiated simply by weak bases to form high molecular mass, but largely uncross-linked polymers. In most cases, atmospheric moisture or the film of moisture on the substrate is sufficient to initiate polymerization, because the adhesives are applied in very thin layers. However, this sensitivity to atmospheric moisture means that the adhesives must be stored in tightly sealed form, usually in polyethylene bottles. Cyanoacrylates are used for bonding small items of nearly all substrates. In case of polyethylene or polypropylene special primers are available. Setting is complete with rubbers in seconds, with aluminum in less than one minute.

Cyanoacrylate adhesives are also used in surgery [45]. They enable parts of the body to be joined together and, therefore, must be used carefully.
Trade Names.In Germany: Sicomet, Loctite, Pattex Blitzkleber; in Japan: Aron Alpha, Alteco, Cianobond; in the USA: Superbonder, Permabond, Super Glue, Krazy Glue and Attak.

Anaerobic adhesives Anaerobic adhesives are mixtures of acrylic esters which remain liquid when exposed to air but harden when confined between metal surfaces. These mixtures can be used for a large number of industrial purposes such as locking threaded fasteners, sealing threaded pipe connections, retaining cylindrical machine components, sealing flange joints, bonding structural components, sealing porous metal castings, welds and powdered metal parts and many other applications which are still being found more than forty years after the initial invention [134]. Several reviews have been published which describe anaerobic adhesives and their applications [135–137].

All anaerobic-curing adhesives consist of a monomer, an initiator, one or more accelerators, and stabilizers. The properties of the cured and uncured adhesives can be modified to control viscosity, color, or fluorescence, reduce strength, increase toughness or heat resistance, provide lubrication and reduce settling of solid fillers. In some applications anaerobic sealants cure more rapidly if the surface is treated with chemicals which catalyze their polymerization.

A number of papers discuss the *reaction mechanisms* of anaerobic adhesive cure [138–142]. The polymerization mechanism of anaerobic adhesives is similar to that of other radical initiation systems except for the special way in which the inhibiting effect of oxygen is used to delay the polymerization and in the chemical activation which occurs at the metal surface.

Initiation:

$$\mathrm{I} \longrightarrow \mathrm{I}^{\bullet}$$

$$\mathrm{I}^{\bullet} + \mathrm{M}^{\bullet}$$

Inhibition:

$$\mathrm{M}^{\bullet} + \mathrm{O_2MOO}^{\bullet}$$

Propagation:

$$\mathrm{M} + \mathrm{M}^{\bullet} \longrightarrow \mathrm{M}^{\bullet}$$

Termination:

$$\mathrm{M}^{\bullet} + \mathrm{M}^{\bullet} \longrightarrow \mathrm{M_2}$$

$$\mathrm{M}^{\bullet} + \mathrm{I}^{\bullet} \longrightarrow \mathrm{MI}$$

The reaction rate of oxygen with free radicals is very high, and the resulting peroxy radical is a relatively poor initiator. When the supply of oxygen is used up within a thin bond line the propagation step can provide rapid development of adhesive strength.

An important factor in the initiation of anaerobic adhesive cure is the redox reaction between a hydroperoxide and transition metals with adjacent oxidation states [138].

$$\mathrm{Fe^{2+}} + \mathrm{ROOH} \longrightarrow \mathrm{Fe^{3+}} + \mathrm{RO}^{\bullet} + \mathrm{OH^{-}}$$

$$Fe^{3+} + ROOH \longrightarrow Fe^{2+} + ROO^{\bullet} + H^{+}$$

Other transition metals react similarly, and copper is particularly active.

The use of saccharin and *N,N*-dimethyl-*p*-toluidine results in a substantial acceleration of the initiating reaction. Although each of these components is an accelerator by itself the combination has a strong synergistic effect. It has been suggested that a charge-transfer complex is formed by these materials [139]. It has been suggested that one of the functions of saccharin in anaerobic curing is to dissolve metal ions from the surface that then catalyze the decomposition of CHP [142]. The role of the bonding surfaces and the effect of different types of accelerators account for most of the literature on the cure mechanism of anaerobic adhesives.

The first patents on anaerobic adhesives mentioned only the polyglycol dimethacrylates as *monomers*, mainly tetraethyleneglycol dimethacrylate [143], [144]. The use of acrylic or methacrylic acids to improve adhesion to smooth surfaces was mentioned in [145]. A series of polyurethane polyacrylates were prepared by reaction of toluene diisocyanate, other isocyanates, and isocyanate-terminated oligomers with hydroxyalkyl methacrylates [146]. These monomers could be tailored to provide the strength and toughness for some structural adhesive applications. The incorporation of hard and soft segments into the polyurethane backbones provided significant improvements in toughness, cure-through-gap, and cryogenic strength properties [147].

The use of monomethacrylates in anaerobic formulations was disclosed in a patent assigned to Loctite. Specifically mentioned were hydroxyethyl, hydroxypropyl, cyclohexyl, tetrahydrofurfuryl, dimethylaminoethyl, and glycidyl methacrylates and cyanoethyl acrylate [148].

Methacrylate esters containing residual carboxylic acid groups were prepared by the reaction of hydroxyethylmethacrylate with phthalic anhydride, pyromellitic dianhydride, and benzophenone tetracarboxylic acid dianhydride. The residual acid provided improved adhesion [149], [150]. The reaction product of hydroxyalkylmethacrylates with maleic anhydride also produced monomers with residual acid as well as additional curable unsaturation [151].

Three Bond Company used trimethylolpropane trimethacrylate [152] and ethoxylated bis-phenol A dimethacrylate [153] in anaerobic formulations. These monomers have some advantages in providing improved heat resistance.

At Henkel dicyclopentadienyl methacrylate was used in anaerobic formulations with high strength [154]. Rohm & Haas disclosed the use of dicyclopentenyloxyethyl acrylate and methacrylate in anaerobic formulations [155]. These monomers provide good cure strength on metal parts which have not been degreased and also have lower odor and volatility than the corresponding dicyclopentadienyl esters.

Silicone methacrylates have been formulated into anaerobic adhesives by Dow Corning [156], Toshiba Silicones [157], and Loctite [158], [159].

The most commonly used *initiator* for anaerobic adhesives is cumene hydroperoxide. Many other hydroperoxides have been disclosed such as *tert*-butyl hydroperoxide, *p*-menthane hydroperoxide, diisopropylbenzene hydroperoxide, pinene hydroperoxide and methyl ethyl ketone hydroperoxide [160].

Storage-stable anaerobic formulations can be prepared without hydroperoxide if the methacrylate resin is aerated in the presence of an amide and a tertiary amine [161].

The first *accelerator* used in an anaerobic adhesive was tri-*n*-butylamine [162]. Saccharin was also found to be an effective accelerator [163] and the combination of saccharin and *N,N*-dimethyl-*p*-toluidine was particularly effective if properly stabilized [164].

Many patents have been issued on various accelerators and combinations thereof. 1-Acetyl-2-phenylhydrazine [165], [166], benzenesulfonyl hydrazide [167], dibenzenesulfonamide [168], and other similar compounds have been disclosed.

The use of saccharin is of particular interest, and a number of compounds have been prepared which have a similar chemical structure and reactivity. The reaction of sulfuryldiisocyanate with acetic acid gives a disulfonamide which is an effective accelerator [169]. Similar reactions of *p*-toluenesulfonyl isocyanate and chlorosulfonyl isocyanate can be used to prepare many different compounds which are active accelerators [170], [171]. These methods allow the

preparation of accelerators with improved solubility.

The monomers used in anaerobic adhesives and sealants generally contain at least one free radical *stabilizer* such as hydroquinone or *p*-methoxyphenol. Benzoquinone, naphthoquinone, and similar compounds provide improved shelf stability without retarding the anaerobic cure [172].

The use of a soluble metal-chelating agent such as tetrasodium ethylenediamine tetraacetate effectively stabilizes an anaerobic formulation against small amounts of metal contamination [173]. The wide variety of applications of anaerobic adhesives and sealants is made possible by the modifications which make the viscosity appropriate to the application. An application which requires penetration into close fitting parts should have very low viscosity while a product used with large, loose fitting parts should have a high viscosity. Polymethacrylates, cellulose esters, butadiene – styrene copolymers, acrylonitrile – butadiene – styrene copolymers, poly(vinyl chloride), copolymers of vinyl chloride and vinyl acetate, poly(vinyl acetate), cellulose ethers, polyesters, polyurethanes, and other thermoplastic resins have been used to control the flow characteristics of anaerobic sealants [174].

The flow characteristics of anaerobic formulations can also be controlled by the addition of fumed silica and other solid additives which can impart thixotropic properties [175].

Many anaerobic adhesives and sealants may be required to develop a relatively low strength so that the components can be disassembled for repair or replacement. Many liquid plasticizers have been used for this purpose [176], but the use of a low molecular mass polyester is advantageous [177].

The "toughness" properties of anaerobic adhesives can be enhanced by the addition of a reactive elastomer [178]. The heat resistance of anaerobic adhesives and sealants can be enhanced by the addition of a bis-maleimide [179]. These maleimide additives appear to be relatively unreactive during the initial anaerobic cure. As the adhesive is exposed to high temperatures, the methacrylate backbone degrades and the methacrylates can then copolymerize with the maleimides to form a more heat resistant matrix [180].

The addition of dyes to anaerobic adhesives and sealants assists in identification and inspection of the products. Automated inspection procedures are made possible with dyes that fluoresce under ultraviolet light. Titanium dioxide pigments can make the sealants opaque and more visible.

Solid fillers are added to some anaerobic adhesives and sealants for various purposes. Mica, talc and other mineral fillers can help to provide an "instant seal" capability to anaerobic pipe sealants. The sensitivity of the anaerobic cure system to metal contamination requires that these fillers be chosen very carefully.

Powdered graphite, polytetrafluoroethylene, and polyethylene can function as lubricants in pipe sealants and threadlocking compounds. This lubrication can prevent galling in close-fitting pipe threads [181]. Lubricating additives in thread-locking sealants can provide control of the clamping force exerted by a fastener at a given tightening torque.

The cure speed and adhesion of an anaerobic sealant can be increased by treatment of the surface with a solution of a *primer* or *activator*. Early anaerobic thread lockers were strongly affected by part cleanliness, and degreasing the parts with a chlorinated solvent improved performance dramatically. The condensation product of an aldehyde and a primary or secondary amine, a sulfur-containing radical accelerator, or a compound of an oxidizable transition metal were some of the materials used to activate the anaerobic cure [182–184]. Due to the ozone-depleting potential of chlorinated solvents, acetone, alcohols, or hydrocarbon solvents are generally used at present.

Modified acrylics, also referred to as second-generation acrylics or no-mix adhesives, are composed of a modified acrylic adhesive and a surface activator. Typical modified acrylics are based on cross-linked polymethyl methacrylate grafted to vinyl-terminated nitrile rubber. Carboxyl-terminated rubbers have also been used.

Unlike epoxies, which cure by an ionic polymerization mechanism, modified acrylics cure by free-radical addition. Therefore, careful proportioning of components is not required. In two-component systems, no mixing is required because the adhesive is applied to one substrate,

the activator to the other, and the substrates are joined. Handling strength is rapidly achieved with this fast-curing system.

Modified acrylics have good peel, impact, and tensile lap-shear strengths. High bond strengths are obtained with metals and plastics even if surfaces are oily or improperly cleaned. The cured adhesive exhibits little shrinkage. Resistance to high humidity is good, particularly when bonding plastic substrates. The limitations of modified acrylic adhesives include low high-temperature strength, flammability of the primer, which is an amine (Schiff base) or a peroxide in a solvent, and the odor of the uncured adhesive.

Trade Names. In Germany: Agomet, Multi-Bond; in Japan: Diabond SGA, Cemedine SGA; in the USA: Versilok.

A/B acrylates are a new class of acrylate adhesives. They are two-component adhesives that cure rapidly after application. They avoid using a solvent-based primer as in modified acrylics. The acrylates used are mainly acrylated polyurethane oligomers, compounded with catalyst in one component and accelerator in the other. The two components can be applied 1/1 without mixing as beads side by side. Mixing is performed by coating the materials to be bonded or by a static-mixer applicator. Curing starts immediately after mixing. Handling strength is attained rapidly, and final curing is complete after several hours. An advantage over modified acrylics is the lesser odor due to the use of oligomers instead of monomers.

Trade Names. In Germany: Omnifit A/B; in the USA: Loctite.

6.2.2. Adhesives Setting by Polyaddition

Epoxy Adhesives. Epoxies [47], [48] are among the most widely used structural adhesives. These chemically reactive systems include two-component systems that cure at room temperature or elevated temperature and one-component systems, which usually require heat for curing.

Epoxy thermoset adhesives, available in many formulations, can be used to join most materials. These adhesives have good strength, do not emit volatile substances during curing, and have low shrinkage. However, peel strength and flexibility are low, and epoxies are brittle.

A typical two-component system consists of a resin and a hardener, which are packaged separately. In a one-component system, the resin and hardener are packaged together. Other possible additives include accelerators, reactive diluents, plasticizers, resin modifiers, and fillers.

The most common epoxy resin is based on the diglycidyl ether of bisphenol A. Epoxy resins are sometimes esterified with fatty acids or modified with polyurethane intermediate chains for elastification.

The hardeners used for *hot-setting formulations,* i.e., for hardening temperatures above 100°C, are preferably dicarboxylic acid anhydrides, dicyandiamide, and certain aromatic amines. The hardening temperatures, which are normally in the range 150 – 200°C, can be lowered by adding accelerants [49].

The hardeners for *cold-setting systems* include aliphatic and cycloaliphatic amines and polyamines, adducts of polyamines and epoxy resins, phenol – amine combinations, and condensates of polyamines and dimerized fatty acids (polyaminoamides). Whereas amine hardeners must be used in a stoichiometric ratio to the reactive epoxy groups, polyaminoamides may be overdosed to a certain extent and thus used to elasticize the adhesive resin.

An exothermic reaction occurs during the curing of epoxies. This reaction can be minimized by lowering the temperature of the mixed components in two-component systems, by limiting the batch size, and by using shallow mixing containers.

Epoxies are brittle, especially if cured with an anhydride. Therefore, thermoplastics and rubbery modifiers are often added to decrease the brittleness.

Nylon epoxy, which first became available in 1960, has an approximate maximum service temperature of 180 °C, compared to 80 °C for unmodified epoxies. The main advantage of nylon epoxies is increased flexibility and much higher peel strength, compared to unmodified epoxies. Also this very tough material has excellent tensile lap-shear strength. Fatigue and impact resistance are also good. Limitations include poor peel strength at low temperatures and poor creep resistance. Perhaps the most serious limitation is extremely poor moisture resistance

in both the uncured and cured material. Nylon epoxies are not as durable as elastomer epoxies or thermoplastic-modified epoxies. A typical application of nylon epoxies is in laminates.

Elastomer epoxies generally contain nitrile rubber as the elastomeric component. This system is also referred to as a modified or toughened epoxy. One of the applications of widest use is in films and tapes. Elastomer epoxies cure at low pressures and low temperatures over a short time interval. This is achieved by adding a catalyst to the adhesive formulation. Bond strengths of elastomer epoxies are lower than those of nylon epoxies. However, the major advantage of elastomer epoxies is their sub-zero peel strengths, which do not decrease as fast as those of nylon epoxies. In addition, the moisture resistance of elastomer epoxies is better than that of nylon epoxies but not as good as that of vinyl – phenolics or nitrile – phenolics. Limitations to the use of elastomer epoxies include poor water immersion resistance and poor properties when exposed to marine conditions.

In addition to liquid and pastelike epoxy resin adhesives, solid powdered products are also available commercially. In addition, epoxy resins are used in the form of adhesive films for certain applications. The most important are nylon epoxy resin films consisting of mixtures of relatively high molecular mass polyepoxides and polyamides, and epoxy resin – phenolic resin combinations.

Trade Names. In Germany: Araldite, Metallon, Pattex Kraft-Mix, UHU-plus, Scotch weld, Terokal; in Japan: Sho-bond, KBK, Bond E, Sanyu Resin; in the USA: Scotch Weld, Resiweld, Ross Epoxy Glue, Ten-set Epoxy and Duro E-pox-e Glue.

Reactive Polyurethane Adhesives (for the raw materials → Polyurethanes) date back to the late 1930s and acquired their present economic significance over the past decades [50–52]. They are distinguished by good adhesion to various substrates and, compared with other reactive adhesives, by high elasticity, even at low temperatures. Polyurethane adhesives are marketed both as solvent-free and as solvent-based adhesives.

The generic term polyurethane adhesives covers both adhesives that already contain polyurethane compounds or compounds with a urethane bond and also reactive adhesives in which urethane groups are formed only during the hardening process.

Chemically reactive polyurethanes include both one- and two-component systems. One-component systems are usually based on a polyether polyol treated with a polyisocyanate to give an isocyanate-terminated polymer. A one-component system cures when exposed to moisture at room temperature. One-component polyurethane hot-melt adhesives are also cured by moisture after application. Two-component systems result from the reaction of low molecular mass polyols and isocyanates or from isocyanate-terminated prepolymers with either polyols or polyamines. Two-component systems cure at room and/or elevated temperatures.

One-component heat-curable urethanes are also available. In these formulations, free isocyanate groups are typically blocked by the addition of phenol. The prepolymer is then blended with the polyglycol curing agent and packaged. The mixture is stable until heated to elevated temperatures, whereupon the phenol is released and the isocyanate groups are regenerated. A rapid cure occurs when these groups come into contact with the polyglycol.

Polyurethanes bond to most surfaces. They also have outstanding tensile lap-shear strength at lower temperatures. In fact, polyurethanes have better low-temperature strength than any other adhesive, even epoxy – nylons. Good flexibility, abrasion resistance, and toughness are other advantages of polyurethanes. Limitations include sensitivity to moisture in both uncured and cured adhesives, the toxicity of isocyanates, and poor tensile lap-shear strength at higher temperature. A major advantage, and also the reason for the wide range of applications of polyurethane adhesives, lies in their versatility in regard to raw materials, composition, and curing.

Normally, reactive polyurethane adhesives contain high molecular mass, still reactive polyurethane prepolymers with terminal hydroxyl or terminal isocyanate groups. Prepolymers with terminal hydroxyl groups can be hardened by the addition of polyisocyanates. Depending on the functionality of the prepolymers and polyisocyanate hardeners, curing may involve both cross-linking and linear enlargement of the molecule. High molecular mass polyurethane elastomers contain urethane

groups that can still react with isocyanates and are thus cross-linked (contact cements).

Prepolymers with terminal isocyanate groups can react with and are hardened both with amino or hydroxyl compounds and with water. Hardening with water is preferred for one-pack systems, for which both atmospheric moisture and the film of water on the substrates act as hardeners for the prepolymer. Cross-linking with polyamines is particularly fast compared with curing by polyols.

One-pack polyurethane adhesives can also be manufactured by protecting the isocyanate groups of prepolymers with, for example, phenols or oximes, which are split off by the action of heat during bonding. Protected polyisocyanates can also be added to reactive polyurethane adhesives as latent hardeners. Tertiary amines or organotin compounds can be added to accelerate isocyanate reactions.

Polyurethane adhesives should be stored in tightly sealed containers, because shelf life and adhesive quality can be affected adversely by the uptake of water.

Trade Names. In Germany: Liofol, Macroplast; in Japan: Diabond, Bostik; in the USA: Chemlok.

6.2.3. Adhesives Setting by Polycondensation

Polyhydroxymethyl Compounds. Hydroxymethyl compounds that release water during condensation are used primarily for polycondensation adhesives. Accordingly, the high-quality bonding of impermeable substrates requires relatively high pressures to compensate for the water vapor pressure developed during hot setting. Polycondensation adhesives are used both in pure form and as solutions in water or organic solvents. Hardening can be initiated by the application of heat or by the addition of hardeners.

Polycondensation adhesives include phenolic and resorcinol resins and also urea – and melamine – formaldehyde resins (→ Phenolic Resins; → Amino Resins). Adhesives of particularly high quality are obtained by combining phenolic resins with poly(vinyl formal) resins, nitrile rubber, or epoxy resins. Combinations such as these are often marketed as adhesive film.

Trade Names. In Germany: Redux, Tegofilm, Kaurit, Pressal.

Phenolics or phenol – formaldehyde structural adhesives are chemically reactive systems that cure to form thermosets. In one-component systems, meltable powders (resols) are used as binders for particle board or as alloys (including nitrile – phenolics, vinyl – phenolics, and epoxy – phenolics), which are used in the structural bonding of metals. In two-component systems, the resin and catalyst are mixed and then heated to initiate curing. Both systems cure by a condensation reaction that produces a byproduct.

In general, phenolics are low-cost adhesives with good strength and resistance to biodegradation, hot water, and weathering. Elevated-temperature resistance is also good. Limitations include low impact strength and high shrinkage stresses, which lead to brittleness. Shelf life is limited, the adhesives are dark colored, and they can be corrosive.

Phenolics dominate the wood adhesives market, especially for plywood. The structural adhesive bonding of metals, particularly with phenolic alloys, is another application.

Nitrile – phenolic alloys are composed of nitrile rubber and phenolic with additives. These systems are available in liquid or film form. The advantages include a maximum service temperature of 140°C, low cost, high bond strength, and excellent resistance to water, oil, biodegradation, and salt. Disadvantages include poor to moderately poor low-temperature resistance and high-pressure, high-temperature curing. Typical applications include the bonding of metals, plastics, rubber, wood, glass, and ceramics. This structural adhesive is used in automobile brakeshoes and clutch disks and in aircraft applications.

Vinyl – phenolics are alloys composed of polyvinyl formal (PVF) – phenolics or polyvinyl butyral (PVB) – phenolics. They are available as liquids and films. In general, vinyl – phenolics have a maximum service temperature of 80°C and are equal to phenolic – nitriles in strength. These adhesives are better than epoxies in sandwich structures that require high strength.

PVF – phenolics have good resistance to fatigue, weathering, fungi, salt, humidity, water, and oil. Creep resistance is good in some formulations, but is poor in others above 90°C. These adhesives are used in metal-honeycomb and wood – metal applications.

PVB – phenolics also have good resistance to weathering, fungi, salt, humidity, water, and oil, and creep and fatigue resistance are even better than those of PVF – phenolics. These adhesives are used to bond metal or reinforced plastic facings to paper honeycomb, cork to rubber, and steel to rubber.

Epoxy – phenolic adhesives are also available as liquids or films. They are among the best adhesives for long-term use at 150 – 250 °C. In general, epoxy – phenolics are relatively expensive. They have fairly good shear and tensile strengths over a wide temperature range, but peel strength is poor. Impact resistance is also poor, whereas resistance to weathering, aging, aromatic fuels, glycols, hydrocarbon solvents, and water is good. Typical applications include the bonding of metals, glass, ceramics, phenolics, and honeycomb sandwich structures in aircraft applications.

Silicone Adhesives (for the raw materials → Silicones). Cold-cross-linking one-pack silicone adhesives also harden by a polycondensation mechanism. They are silanols whose OH groups have reacted with acetic acid or amines. Treatment of these compounds with water leads to the elimination of acetic acid or amines and generates the reactive silanols, which cross-link with elimination of water. These products are particularly important as sealing compounds in the building industry.

Usually silicones are available as both one- and two-component systems that cure to thermoset solids. A one-component room-temperature-vulcanizing silicone cures at room temperature upon exposure to atmospheric moisture. Curing is either acidic or nonacidic in the presence of moisture. Adhesives that have an acidic cure have greater unprimed adhesion and a longer shelf life. However, the corrosion of metals by the acid is a potential problem. Thin films of approximately 0.6 mm cure within 90 min., whereas 15 mm films of require ca. 7 d to achieve full cure.

Two-component silicones do not require moisture to cure. During the addition polymerization of silicones, curing is achieved by catalytic action. Pot life, setting time, and cure time are dependent on the catalyst concentration. A system with 5 % catalyst will typically have a 3 h pot life, a 22 h setting time, and a 7 d cure time at room temperature. Increasing the temperature accelerates curing. Addition-polymerized silicones exhibit little shrinkage upon curing and have good high-temperature resistance. In the condensation polymerization of two-component silicones, byproducts are released. These systems are less likely to be inhibited and can be used on a greater variety of substrates. However, reversion of polymerization is a potential problem.

In general, silicone adhesives have good peel strength over a service temperature range of – 60 to 250 °C. Some have limited service to 370°C. Flexibility and impact resistance are good, as are resistance to moisture, hot water, oxidation, and weathering. Typical lap-shear strengths are low in metal-to-metal bonds, with values ranging between 1 and 3.5 MPa. The cost of these adhesives is high.

Silicone adhesives are useful in bonding metals, glass, paper, plastics, and rubbers, including silicone, butyl rubber, and fluoroelastomers.

MS Polymers. A relatively new class of adhesives are the two-component MS polymer adhesives, known as moisture-cross-linkable sealants up to know. Due to their excellent elastic properties they are used for elastic bonding. Chemically they are based on silane-modified prepolymers such as poly(propylene glycol). The end groups of these so-called MS polymers are silanealkoxyl groups which are hydrolyzed by water and then cross-link. In comparison to silicone adhesives, they adhere to all kinds of surfaces without any primer and are paintable in the freshly cured state.

Polyimides and Polybenzimidazoles [53]. The highly temperature resistant polyaromatic adhesive resins of the polyimide and polybenzimidazole type are applied as precondensates in the form of solutions or films and are condensed to completion at 230 – 250°C under a relatively high pressure of 8 – 10 bar. The precondensates remain stable for only a few hours

at room temperature and have to be stored at −18 °C.

These structural adhesives with high-temperature resistance are based on synthetic organics with aromatic (benzene) and/or heterocyclic rings in the main structure. These chemical groups often include imidazoles and substituted imidazoles. These prepolymers have open-ring structures that close upon application of heat. The condensation reaction leads to a highly cross-linked structure. The adhesives are available as liquids and films. High-temperature adhesives, including polyimides and polybenzimidazoles, are expensive materials that are difficult to handle and require long curing times during which volatile substances are released. Polyimides are superior with regard to long-term strength retention at elevated temperatures. At 260 °C in air, polyimides have higher bond strength than epoxy – phenolics. Polybenzimidazoles are stable in air for short-term exposures up to 290 °C. Both polyimides and polybenzimidazoles are moisture-sensitive. Applications of high-temperature adhesives are primarily in the aircraft and aerospace industries for bonding metals.

6.2.4. Vulcanizing Adhesives

Adhesives for joining unvulcanized elastomers to metals or plastics under vulcanization conditions are called rubber-to-metal bonding agents. Mixtures of film-forming components, cross-linking agents, and stabilizers dissolved in organic solvents are used as bonding agents. The solids content varies from about 15 to 25 % according to type. The film-forming components used are chiefly halogenated polymers, whereas the cross-linking agents are, for example, nitroso compounds, isocyanates, and oximes in combination with an oxidizing agent [54]. In many cases, primers are applied additionally to the metal surface to improve resistance to corrosion. Trade Names. In Germany: Chemosil; in Japan: Metalok; in the USA: Chemlok.

6.2.5. Ultraviolet/Electron Beam (UV/EB) Curing Adhesives

Most UV/EB adhesives are based on an addition polymerization curing mechanism. They consist of acrylic esters of various types or combinations of acrylates with aliphatic or aromatic epoxy resins, urethanes, polyesters, or polyethers. Although the epoxy-based systems have higher tensile strengths, their elongations are less than those of the urethane-based systems. In addition, the urethane-based systems have better abrasion resistance. UV/EB adhesives that undergo cationic polymerization are based on epoxies with reactive diluents and cyclic monomers. The main advantage of UV/EB-curable adhesives is rapid curing at room temperature, so they can be used to bond heat-sensitive substrates, such as poly(vinyl chloride). In addition, the rapid cure often eliminates the need to fix parts and greatly increases production rates.

UV/EB-cured adhesives have been used to replace solvent-base adhesives because of the increasing cost of properly recovering and disposing of solvents. Most of these adhesives are single-component materials that require no mixing and produce little waste.

The cross-linked nature of UV/EB-cured adhesives results in good chemical, heat, and abrasion resistance; toughness; dimensional stability; and adhesion to many substrates. Unlike thermal curing, EB or UV curing requires pure raw material grades. The use of dual-curing systems allows opaque substrates to be cured.

The mechanical properties of UV-curable structural adhesives are dependent on polymer molecular mass and cross-linking density. These factors are related to the prepolymer, degree of stiffness or flexibility, and functionality. The overall adhesive strength is affected by:

Adequate UV transmission through the bond line
Adhesive thickness
UV intensity
Postcuring (by heat or exclusion of oxygen, as in anaerobics)

The main components of UV/EB-cured adhesives are reactive oligomers which contribute adhesion, toughness, and flexibility to the overall properties of UV/EB adhesives. Typical reactive oligomers include acrylated epoxy resins and aromatic urethanes.

Monomeric diluents are low molecular mass monofunctional molecules that reduce the viscosity of liquid oligomers. Some, such as

methacrylates, increase the toughness and adhesion. Various types of acrylates are used as reactive monomers.

Cross-linking monomers such as 1,3-butylene glycol dimethacrylate, tripropylene glycol diacrylate, and pentaerythritol tetracrylate are also used in UV/EB adhesive formulations.

Free-radical initiators trigger the cross-linking reaction. In EB-cured adhesives, the electrons act as free-radical initiators for addition polymerization. Therefore, no chemical initiator additives are needed. In UV-cured adhesives, photoinitiators, which release free radicals when exposed to UV radiation, are required to initiate addition polymerization. The most recent UV- and EB-curing systems involve cationic polymerization mechanisms.

Typical UV-curable adhesive applications include the electronics, automotive, medical, optics, and packaging industries, as well as tapes and labels. EB-curable adhesives are used in magnetic tapes and floppy disks, where magnetic particles are bonded to films, as well as in packaging, tapes, and labels.

Laminating adhesives have the largest share of the UV/EB-curable adhesives market. Adhesives are applied and then cured in-line. The lack of solvent in these adhesives eliminates the drying step required when using solvent-based adhesives. The high degree of cross-linking in UV/EB-cured laminates gives high bond strength, good heat resistance, and good chemical resistance. Substrates that are typically laminated include: film to paper, foil, fabric, glass, film, or wood; paper to foil, wood, or paper; wood to wood; and glass to glass or metal.

The pressure-sensitive sector of the UV/EB-curable adhesives market is not nearly as large as that of the laminating-adhesives segment. Because slight changes in exposure time greatly affect peel strength and tack, the formulation of these adhesives is critical. Three approaches are used to make pressure-sensitive adhesives. The first is based on a UV-curable system employing conventional tackifiers blended with UV-reactive moieties. The process must be closely controlled to attain repeatable tack and peel strength. The second approach is based on an EB-curable system that is actually a hot-melt adhesive. EB-cross-linkable thermoplastic rubbers are usually used in these adhesives and result in hot-melt pressure-sensitive adhesives with better heat resistance than conventional elastomeric hot-melt adhesives. The third approach involves the development of inherently tacky oligomers for pressure-sensitive adhesives.

There are several limitations to the use of UV/EB-cured adhesives. EB equipment is expensive. UV equipment is less expensive, but the materials themselves are usually more costly because of the need for photoinitiators. To cure adhesives properly, one substrate must be transparent to UV radiation. However, the necessity of having a transparent substrate has been removed by the introduction of dual-curing adhesives. These adhesives are quickly set by a UV cure and are more fully cured by a second mechanism involving the introduction of heat or moisture or the elimination of oxygen (anaerobics). In EB-curable adhesive, the depth of EB penetration is limited by the density of the material, rather than its opacity.

The cure time of UV adhesives is usually less than 60 s and depends on:

Bond-line joint thickness: As the thickness increases, UV radiation loses its ability to penetrate totally, necessitating a second cure.

Type of substrate: A transparent substrate such as glass with a small gap may take as little as 5 s to cure. Opaque and darker substrates require longer curing times.

Light intensity: The more intense the UV light, the faster the cure.

The curing time of EB adhesives is comparable to that of UV adhesives. In general, EB radiation allows adhesive curing to be achieved at greater depths than is possible with UV radiation. Electrons can pass through substrates that are opaque to UV light. In addition, the area of exposure and the depth of penetration can be controlled by means of the EB conditions.

6.2.6. Conductive Adhesives

Conductive adhesives include both electrically and thermally conductive materials. Most adhesive fillers that result in electrical conductivity in a material also contribute to its thermal conductivity. However, thermally conductive adhesives that are electrically insulative are also available.

The majority of applications involve the use of silver, either in flake or powder form. This

filler is preferred to gold because of its lower cost and its lower volume resistivity (that is, higher electrical conductivity). The best silver-filled epoxies have a volume resistivity of 0.001 Ω cm. With silver-filled adhesives, migration of silver to the surface under conditions of high humidity and direct current can occur. Neither gold- nor silver-coated copper fillers in various adhesives have this migration problem. Epoxies with up to 85 wt % silver filler are available, but adhesives with lower filler loadings generally have better strengths.

Other fillers that provide electrical conductivity include copper and aluminum, although oxide formation on the surfaces of these fillers can occur. This results in lower conductivity because of decreased particle-to-particle contact.

Conductive epoxy resins and polyimides are used as die-attach adhesives. The chloride ion contents of polyimides are inherently lower than those of epoxies. However, these adhesives are more expensive, have lower bond strength, and are more difficult to cure. Epoxies with low contents of ionic impurities, including chloride anions and sodium and potassium cations, have been available since 1981. These materials are fast curing and have excellent high-temperature strength and a glass transition temperature of up to 200 °C.

Oxide fillers are electrically nonconductive but are used to provide thermal conductivity. Alumina, the most commonly used filler, is fairly inexpensive. It can be added in high concentrations to epoxies and silicones without significantly increasing the viscosity of the uncured material. Minimum bond line thicknesses are desirable for thermally conductive adhesives because heat flow is proportional to the ratio of thermal conductivity to bond thickness. Alumina-filled epoxies contain up to 75 wt % filler and have thermal conductivities ranging from 1.38 to 1.73 $W\,m^{-1}\,K^{-1}$.

Other thermally conductive fillers include beryllium oxide, which is expensive and toxic, and boron nitride, which has a loading limitation of ca. 40 wt % in epoxies. Other inorganic oxides, including silica, are also used to provide thermal conductivity.

The chemical families that are used most often to provide electrical and/or thermal conductivity include epoxies, polyurethanes, silicones, and polyimides. Epoxies are the most widely used.

7. Bonding Techniques

Cleaning and Pretreatment of the Surfaces to be Joined. To insure good adhesion of the adhesives, the surfaces to be joined must be cleaned before the adhesive is applied. Layers of grease, oil, and lubricant are removed by means of solvents or aqueous degreasing baths. Loose dirt and oxide layers can be removed by brushing. In many cases, the surfaces to be joined are rubbed down with emery paper or sandblasted to remove firmly anchored layers that would interfere with the bonding process or to enlarge the surface by roughening. Pretreatment processes are used additionally to increase the affinity for bonding or (adsorption) activity of the surface and even to impart an affinity for bonding to certain plastics. In general, acidic pickling baths and often combined acidic and oxidizing pickling baths are used for metals, generating defined oxide layers. Alkaline pickling baths are used relatively infrequently. In plastics that are difficult to bond, the surface is modified by chemical pretreatment processes that increase the surface tension of the plastics and, hence, also their wettability and bondability [55]. The so-called corona treatment, i.e, electrical discharge in the presence of air, is used primarily for pretreating plastics films and, in some cases, also metal foils.

In addition to the cleaning and pretreatment of the surfaces to be joined, primers are applied before the actual adhesive in individual cases, their function being to improve the adhesion of the adhesives and/or favorably to influence the resistance to aging and corrosion of the bond.

Adhesive Application. Water-based adhesives and pigmented solvent-based types should be stirred thoroughly before application to remove possible inhomogeneity. For two-component or multipackage adhesives exact dosage and homogenous mixing of the components is essential. The method of application of the adhesive determines the uniformity of the bond line, the thickness of the adhesive layer, and therefore the quality and economy of the bonding process. The application method itself

depends on the size and nature of the surface, the rheology of the adhesive, the required coating thickness, the number of components to be bonded and the speed of the process.

Manual application of the adhesives with toothed spatulas, blades, casting knives, or brushes is the simplest but least uniform method of application. Pasty adhesives can be applied from cartridges which are squeezed by hand or mechanical devices. This method is also used for moisture-curing polyurethane hot-melt adhesives with heated cartridges.

Low-viscosity adhesives are distributed over large areas by spray guns in the same way as paints (Fig. 1). The spraying process may involve the usual atomization with air or can be airless in the case of solvent-based adhesives.

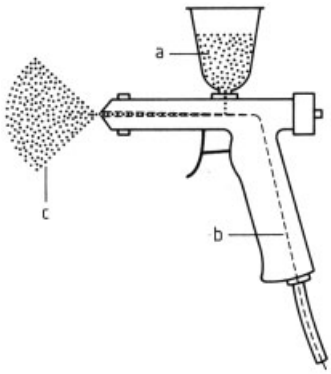

Figure 1. Adhesive application by means of a spray gun
a) Adhesive; b) Compressed air; c) Atomized adhesive

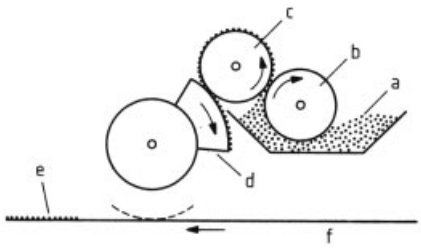

Figure 2. Application by segments
a) Adhesive; b) Dipping roll; c) Transfer roll; d) Rotating segments; e) Applied adhesive; f) Substrate web

Rollers are a versatile application method and range from simple glue rollers fed from a reservoir to highly sophisticated systems with several rolls, suitable for hot-melt, solvent-based, water-based, and solvent-free reactive adhesives (Figs. 2, 3, 4, 5, 6). The coating thickness is regulated by the distance between the rollers, the friction between them (different rotational speeds) and/or by blades. By this means, it is possible to apply from 0.5 g/m^2 up to $>$ 20 g/m^2 in a very precise operation.

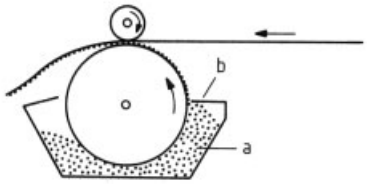

Figure 3. Application by rolls combined with a blade
a) Adhesive; b) Doctor blade

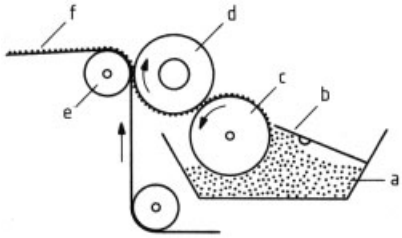

Figure 4. Application by a wheel applicator system
a) Adhesive; b) Blade; c) Dipping roll; d) Shaft with wheel applicator; e) Backing roll; f) Coated substrate web

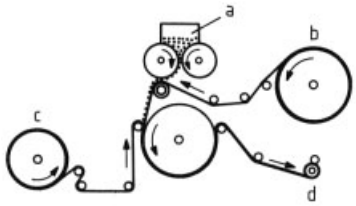

Figure 5. Coat application of hot-melts by rolls (System Pack 400/Kroenert)
a) Hot-melt adhesive; b) Roll I; c) Roll II (cover web); d) Winding up roll

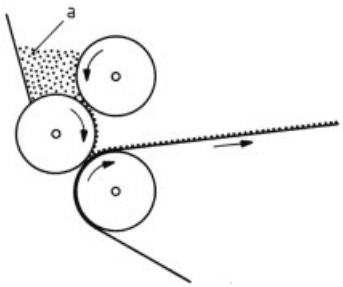

Figure 6. Reserve roll coating system (System Pagendarm)
a) Adhesive

Gravure rollers are used together with doctor blades for a constant coating weight over the width in a narrow range of thickness. Unless the entire surface must be coated with adhesive, slim rollers or so-called wheel applicators are used. If the adhesive is only applied at intervals or with interruptions, segment rollers or wheels, interrupted rollers, toothed wheels, profile rollers, rollers with moving segments, or plungers are used. In addition, blade coating and airbrush coating are used. Hot-melt adhesives must be melted prior to application by rollers. The adhesive is usually transferred from the rolls to the substrates by direct coating in the machine direction. In some cases, reverse-roll coating (Fig. 6) has advantages, especially for low-viscosity adhesives.

Curtain coaters are used for applying adhesives over wide areas by passing the web below the nozzle of the tank (Fig. 7). Highly viscous hot-melt adhesives and thermoplastic polymers are applied from wide-slot nozzles to form a weblike film on the substrate.

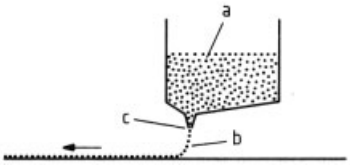

Figure 7. Curtain coater
a) Adhesive; b) Flowing adhesive "curtain"; c) Casting slot

Typically, hot-melt adhesives for folder packaging are liquefied in melting tanks or drum melters and then pumped through heated hoses to the application nozzle. The piston pumps or gear pumps generate sufficiently high pressures to allow contactless application or spraying.

The porous coat system (Nordson) allows the application of discrete, random, and open patterns of hot-melt adhesive to substrates such as films, papers, fabrics, and nonwovens. With the control coat system, hot melt adhesives are applied continuously or intermittently by air-controlled nozzles without contact to the substrate. This technique is used as well for reactive hot-melt adhesives such as moisture-curing polyurethane hot-melt adhesives.

For temperature-sensitive substrates and prolonged open times, gases are incorporated in the melt under pressure. As a result, a hot-melt foam is formed during application [56].

The application units of solvent-free laminating machines consist of three or four rollers with differential speeds to lower the coating weight to 1.5 g/m^2, enabling machine speeds from 250 up to 500 m/min.

Relatively simple handguns (Fig. 8) are used for the workshop or do-it-yourself application of hot-melt adhesives.

Metering and mixing units (Fig. 9) of various constructions (gear pumps, piston pumps, static mixers) are used for the application of two-component adhesives and provide precise dosing and mixing even at elevated temperatures (70 °C).

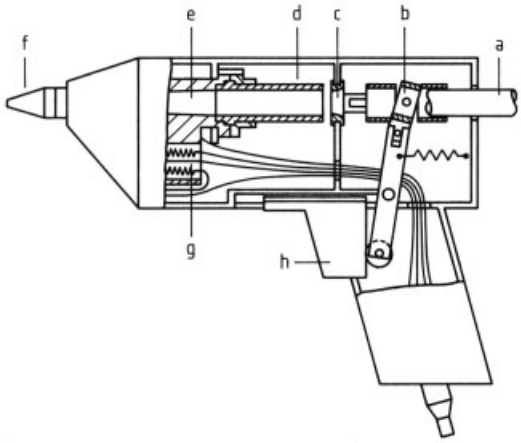

Figure 8. Hand gun for hot-melt (system lang)
a) Adhesive stick; b) Feeding; c) Retaining ring; d) Seal; e) Melting chamber; f) Nozzle; g) Wedge heater or heating wedge; h) Trigger

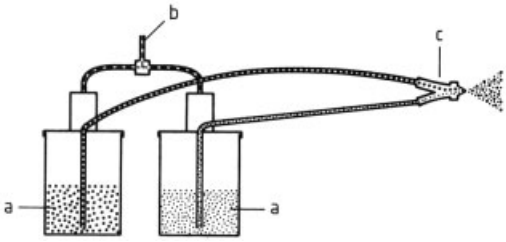

Figure 9. Two component mixing equipment
a) Tanks with metering pumps; b) Compressed air; c) Mixing gun

Anaerobic reactive adhesives and cyanoacrylate adhesives are applied by means of special metering units that dispense small predetermined quantities at the required time intervals.

Joining, Fixing, Drying, and Hardening. After applying the adhesive to one or both surfaces, the substrates have to be fitted together within the open time and fixed until the bond is sufficiently firm. The pressure that has to be applied for bonding is determined only to a minor extent by the adhesive and largely depends on the geometry, nature, and deformability of the substrates and by its recovery forces.

The pressure for joining and fixing to give whole-surface contact can be very high, particularly in the case of large-area bonds, although the adhesive itself does not require such high pressures, apart from contact and hot-setting condensation adhesives. The necessary pressure is usually applied by pressure rollers, laminating rollers, or braces and presses.

The setting time and hence the fixing time can be shortened by heating. Warm-setting adhesives require certain minimum temperatures and times for curing. The energy required may be supplied by recirculating air ovens, drying tunnels in laminating machines and lacquering ovens, by radiant heaters, and by contact heating in heated presses or laminating rollers. Inductive heating of metal-to-metal bonds and high-frequency infrared heating of aqueous adhesives are also used occasionally. Within the last years UV-curing adhesives have been introduced, while electron-beam-curing adhesives are still in the development stage.

Polycondensation adhesives must be subjected during hot hardening to an external pressure that is higher than the vapor pressure of the cleavage products formed during condensation. Accordingly, these adhesives are applied in heated presses or by using pressure bags in heated ovens. The vacuum/pressure process in heated autoclaves is preferably used for uneven, three-dimensional surfaces [57].

A special case of bonding is the heat-sealing of previously applied heat-seal adhesives. Such coatings are converted by heat into a tacky state and bonded by the application of pressure. Thermal activation is carried out by heated sealing bars or plates, hot air, and sometimes by IR radiation.

High-frequency welding is used for special applications, in which cases the necessary heat is generated by the interaction of high-frequency radiation with the dipole moment of the pre-applied heat-seal adhesive itself.

8. Testing of Adhesives

The testing of adhesives comprises two aspects:

1) The adhesive as an individual compound
2) The performance in the actual adhesive joint that is formed between substrates

Generally the following adhesive properties are determined:

Solids content
Rheological properties
Shelf life or storage temperature range
Mechanical properties: tensile strength, modulus, Shore hardness, and elongation to break
Glass transition temperature
Softening point
Flammability and flash point
Health and environmental classification

More important, however, is the performance of the adhesive in an actual joint. Standard test specimens are bonded under standard conditions and subjected to destructive tests. The load values of these tests give information on the adhesive strength between two bonded substrates. Standard destructive test methods are:

Tensile test (EN 26922)
Flatwise tensile shear test (EN 1465)
Peel test (EN 1464/ISO 4578 and EN ISO 28510-1,2)

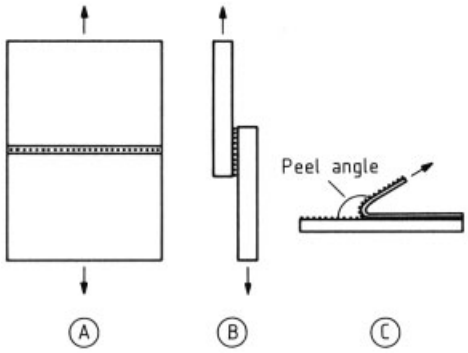

Figure 10. The most important kinds of strain for bonded joints
A) Tensile strain; B) Shear strain; C) Peel strain.

Standard destructive tests are, above all, tensile, shear, and peel tests on specially made, strip-form test specimens (Fig. 10). In general, short-overlap test strips are used for tensile shear tests and flat-bonded test strips for peel tests.

Bending and impact tests occasionally are carried out on these or similar test specimen geometries. Apart from the adhesive itself, the material of the test specimen, surface treatment of the specimen before bonding (sandblasting, pickling, priming), dimensions, overlap surface area, the test speed, the test temperature, and the direction in which the force is applied, all affect the outcome of these tests. Accordingly, the nature and dimensions of the test specimens, the overlap length, the peeling angle, and also the test conditions are laid down in test standards and other guidelines. The test standards [DIN, EN, ASTM, ISO, BS, PSTC (Pressure Sensitive Tape Council)] are designed to give comparable results. Shear strength and tensile strength are expressed in N/mm^2 (strain/joint area), peel strength in N/mm (strain/specimen width). In general, adhesive joints are more resistant to shear and tensile forces than to peel and bending forces.

A distinction can be drawn between:

Static tests to failure under an increasing load
Static long-term strain test below the breaking strain (creep)
Dynamic tests with different loads and frequencies

The static test can be carried out with standard tensile testers that cover the appropriate force range, optionally with a data recorder to obtain a force plot. Dynamic tests require special vibrating devices with constant loads. After the destructive test further information can be gathered from the test specimen. The nature of the break is an important indication of the quality of a bond line. There are three types of failure:

Cohesive failure: break in the bond line
Adhesion failure: break between substrate surface and adhesive (poor adhesion)
Substrate failure, break in the substrate, the bond line is stronger than the substrate (structural bonding)

A combination of the above breakage patterns is possible. This is then indicated by a percentage figure (e.g., 60 % cohesive, 40 % adhesive failure). The above tests are also performed with exposure of the test specimens to extreme environments. Bonds are preferably tested for their resistance to:

Natural weathering
Climate variations (cycling)
Heat aging
Humidity aging
Corrosive atmospheres
Salt spray mist (automotive)
Underwater storage
Submersion in various chemicals such as petroleum, acids, and alkalis

The test conditions are generally much more severe than required for the actual final use of the bonded species. However these exaggerated test conditions deliver valuable information on the durability of the bonded structure in real life conditions. A good understanding has been developed for forecasting the bond line dependability on the basis of the results of aging tests.

The testing of specimens bonded under laboratory conditions is used to compare the performance of adhesives and substrates or the combination thereof, especially in quality-control laboratories. Under real-life conditions laboratory results may not be applicable. Surfaces of substrates, their geometry, curing temperatures, humidity, and many other factors may have an impact on the final bond quality. Therefore it is very important to perform a bond test in the final application even when the laboratory results were satisfactory. In many cases this is the only way to obtain information on the reliability of the bonded construction.

Testing of Anaerobic Adhesives. Testing standards and performance specifications for anaerobic adhesives and sealants have been established by government agencies and industrial organizations in several countries. In the United States there are military specifications for thread lockers, sealants, and retaining compounds. Many of these specifications are being superseded by commercial standards.

Mil-S-22473E, 12 April 1983 "Sealing, Locking and Retaining Compounds: (Single Component)" covers 15 of the earliest "Letter Grade" products. Specifications are set for color, viscosity, locking torque on 3/8-24 steel nuts and bolts and fluid tightness. The effects of immersion in a number of fluids, heat aging at 149 °C and hot strength at 149 °C (or 93 °C for some grades) are also measured. This specification calls for measurement of an "average locking torque" after 90, 180, 270 and 360° of turn.

Mil-S-46163A, 12 July 1983 "Sealing, Lubricating, and Wicking Compounds: Thread-Locking, Anaerobic, Single-Component" covers nine grades of product for sealing (Type I), lubricating (Type II), and wicking (Type III). Specifications are set for color, viscosity, locking torque (break and prevailing torque) on 3/8-16 steel, zinc- and cadmium-plated nuts and bolts, fluid tightness, lubricity, and "wicking" into preassembled fasteners. The immersion, heat aging, and hot strength tests are similar to those specified in Mil-S-22473E.

Mil-R-46082B, 10 June 1983 "Retaining Compounds Single Component, Anaerobic" (Amendment 6, 9 January 1990) covers three types of retaining compounds which are tested with a pin and collar compressive shear specimen. The three types vary primarily in viscosity although there are also some differences in heat resistance and strength. These products are subjected to immersion, heat aging and hot strength tests similar to those described above.

In Britain the Ministry of Defence specification DTD 5628-5633 covers test procedures and performance requirements for a range of products. Five strength bands and four viscosities from penetrating to thixotropic are defined. The torque strengths, including breakloose torque, are tested on M8 nuts and bolts and the shear strength in 12 mm pins and collars [201].

For an improved method of detecting the breakloose torque in a threaded component bonded with an anaerobic thread locker/sealer, see [202].

A British Standard BS 5292 has been prepared which relates to the use of anaerobic sealants on gas appliances.

In Germany standards have been published describing the "Compression Shear Test" (DIN 54452), "Dynamic Viscosity Determination of Anaerobic Adhesives by Rotational Viscometer" (DIN 54453), "Initial Breakaway Test at Bonded Threads" (DIN 54455) and "Torsion Shear Test" (DIN 54455).

In the United States the Industrial Fastener Institute has published standards for "Test Procedure for Locking Ability Performance of Non-metallic Locking Element Type Prevailing Torque Lock Screws" and "Test Procedure for the Locking Ability Performance of Chemical Coated Lock Screws." ASTM has published a "Standard Test Method for Shear Strength of Adhesives Using Pin-and-Collar Specimen" (ASTM D 4562-90, October 1990.)

ISO 10964 (August 15, 1993) "Adhesives – Anaerobic adhesives – Determination of torque strength of anaerobic adhesives on threaded fasteners" describes testing procedures for liquid and preapplied sealants.

9. Applications of Adhesives

It is possible here to select only typical examples for each subheading.

9.1. Bookbinding

Bookbinding with adhesives has played an important role since the beginning of the production of graphical products [185]. Originally, vegetable and animal raw materials were the main ingredients of adhesives, but a steady development of the adhesives took place to fulfill manufacturing requirements. Today, a range of adhesives, tailor-made for particular applications, is available. These applications include perfect binding, case making, casing in, back-lining and gluing-off of thread-stitched book blocks, as well as side-gluing and end-sheet tipping. The most important field is perfect binding.

Perfect Binding of Books. It is impossible to imagine the production of books, brochures and catalogues without perfect binding. New paper qualities such as high-weight coated (HWC) papers, lightweight coated (LWC) papers and recycling papers, high-quality papers such as matt-varnished papers and the increase of the production speed up to 300 books per minute require new bookbinding adhesives. Coated papers can have more than 50 % coat and therefore are difficult to bond. In general, the fiber structure of recycled papers is deteriorated by the recycling process, and these papers must be sized strongly, for instance, to withstand the pulling strain in offset printing, which causes a reduction of absorbency. Therefore, a longer setting time is needed for aqueous adhesives. Strong sizing can also lead to a change of the surface properties, which may cause adhesion problems. Occasional adhesion problems occur because of the often lavish designs of the graphic products.

Varnishing, laminating, embossing, and other refinements have the result that the adhesives can not reach the paper fibers, which are otherwise easy to glue.

Four adhesive types are used for perfect binding:

Glutin glues (animal glues)
Dispersion adhesives
Hot-melt adhesives
Reactive polyurethane hot-melt adhesives

Glutin adhesives are used in form of jellies which are processed at 50 – 70 °C. Due to a sol – gel transition, very fast setting is achieved and a high production speed is reached. However, the adhesion spectrum of glutin glues is normally not very wide.

In contrast, many synthetic polymers possess excellent adhesion characteristics. For example, poly(vinyl acetate) homopolymers combined with external plasticizers show high adhesion, sufficient cohesion, and are easy to use [186]. The improved productivity of the machines has led to the development of fast-drying systems (by high-frequency, IR heating, or gas dryers). The book block can be further processed after cooling for 4 – 5 min.

Hot-melt adhesives are especially useful when high production speeds are required. On rapid cooling, they set fast and achieve the necessary strength quickly. Hot-melt adhesives based on ethylene – vinyl acetate (EVA) copolymers are most commonly used in the graphical industry, although occasionally other thermoplastic polymers are used. Reactive polyurethane (PUR) hot-melt adhesives have proved to be especially useful for high-quality perfect bindings [187]. PUR hot-melt adhesives, which have a very good adhesion characteristics and excellent strength, are applied at 120 – 130 °C. After cooling (physical setting), they react with moisture (chemical setting) to give high-molecular compounds which are practically unmeltable. Besides good adhesion characteristics, they have high chemical resistance (e.g., to the oil of printing inks), high temperature resistance, and good aging resistance. PUR hot-melt adhesives require special premelters and application equipment because of their reactivity. The PUR adhesives combine the advantages of dispersions (e.g., good adhesion, high aging resistance) and of hot-melt adhesives (e.g. high production speeds) in a single system.

Another method of optimizing adhesion and cohesion is to use two different adhesives, one immediately after the other (two-shot application) [188]. The two adhesives can be of the same type (e.g., two dispersion or two hot-melt adhesives), or two different systems can be used (e.g., a dispersion primer followed by a hot-melt or PUR hot-melt adhesive).

Case Making. In case making, when the covering material consists of textiles, plastics film, or paper, as in the backlining of books, hot animal glues are used because they set rapidly by gelatinization and rapidly bond the narrow overlap of the coating material around the board, even under the brief contact pressure normally applied in the machine. In special cases, also emulsion-based adhesives are also used.

Casing In. Depending on the case material and the machine speed, normally emulsion adhesives [poly(vinyl acetate) or copolymer-based] with good flowing properties are used for casing in.

9.2. Adhesives for Packaging and Packaging Materials

Manufacture of Corrugated Board. For this purpose starch suspensions are used almost exclusively. Some of the starch is gelatinized under the heated corrugating rollers which form the corrugation, thus insuring high initial tack. The use of emulsion-based or hot-melt adhesives is confined to special cases. Hitherto, it has been discouraged by the relatively high price of these products, which only occasionally is compensated by other advantages.

Lamination of Paper and Board. Adhesives based on starch, dextrin, glutin and poly(vinyl alcohol), and also emulsion adhesives, mainly poly(vinyl acetate), are used for the lamination of paper and board. Only when high initial tack and a lay-flat effect are required are hot glues based on glutin still in use. Starch-based adhesives modified with silicate, casein-stabilized copolymer dispersions, and rubber latices are used in the lamination of paper to aluminum foils. Modified polyacrylate

solutions and reactive polyurethane adhesives are used mainly for paper-to-plastic lamination.

Adhesive Coating of Paper and Board. Pressure-sensitive coatings, for example, for self-adhesive labels and envelopes include natural rubber–resin combinations in solvents, polyacrylate emulsions, and hot-melt adhesives. Polyacrylate emulsions and hot-melt adhesives are becoming increasingly significant because they are solvent-free.

Heat-sealable coatings are applied from emulsions, solutions, or melts and are activated in different ways: heat-sealable coatings with delayed tack mainly by heat in the absence of heavy pressure, and typical heat-sealing coatings by pressure and heat applied to the bond area as contact or high-frequency heat.

Moisture-reactivated coatings are produced from dextrin glues, dextrin emulsions, mixed glues, and remoistenable hot-melt adhesives.

Bags, Sacks, Pouches, and Boxes. Starch-based adhesives and also emulsions of poly(vinyl acetate) homopolymers and copolymers or mixtures thereof are used in the manufacture of bags, pouches, and carrier bags, depending upon the type of materials (e.g., paper, plastic films, etc.), surface treatment, and the speed of the machine. Machine design factors necessitate the use of pastelike starch-based or emulsion-based adhesives for the bottom gluing of block-bottom or cross-bottom bags.

In making paper sacks, starch-based adhesives normally are used for the longitudinal seam and also for the bottom and bottom patching. For paper sacks with PE free film or inliner, emulsion-based synthetic resin adhesives are used. For the longitudinal seam of the inliner, hot-melt adhesives are used.

The manufacture of folding boxes has long been a domain for free-flowing, low- to medium-viscosity emulsion-based synthetic resin adhesives.

Envelopes. In general, the top-flap gumming of envelopes is made with modified conventional dextrin glues and dextrin–emulsion mixed glues, which dry more quickly and show better ability to stay flat for automatic enveloping. To an increasing extent, modified emulsions and remoistenable hot-melt adhesives are being used. Dextrin glues and synthetic resin emulsions normally are used for bottom-flap and window gluing. Gums for self-seal gumming are based on natural rubber. Self-adhesive top flaps are based on polyacrylate emulsions.

Adhesives for Automatic Packaging Machines. In automatic packaging, package production, filling, and closing are often carried out in a single machine. For this reason, very fast setting adhesives are used.

Considerable recovery forces occur in the closing of prefabricated boxes, particularly in seaming machines, so that hot-melt adhesives often have to be used in such cases. Emulsion-based adhesives are normally used for the dust-tight sealing of cases. Slide boxes, deep-freeze packs, etc., can be line- or spot-sealed with hot-melt adhesives.

Cartons are sealed with emulsion-based or hot-melt adhesives, depending on the length of the pressure belt and the quality of the cardboard. On modern, fast-running machines with up to 120 cases to be closed per minute, hot-melt adhesives are preferably used, due to their fast setting. An interesting combination is the simultaneous use of emulsion-based and hot-melt adhesives. In this case, a spot application of hot-melt adhesive performs the function of the pressure unit or compression section, whereas the actual bonding function is performed by emulsion-based adhesives, generally applied in line form.

Cigarette Manufacture. For gluing cigarette seams, depending on the application system, highly viscous or free-flowing starch glues or emulsion-based adhesives are used. Hot-melt and emulsion-based adhesives are used for cigarette-filter seams, since their viscosity, initial tack, and setting properties are compatible with the special conditions of this manufacturing step, both at low startup speeds and at maximum speeds of 600 m/min. Emulsion-based adhesives are used for bonding the filters to the cigarettes.

Adhesives for Laminated Films. Laminated films are produced by two techniques: lamination and extrusion. In lamination, the low-viscosity adhesive solution is applied by rolls to one of the films in a thickness of a few micrometers and dried. The two films are then combined in the laminator. The adhesives used

are either two-component or one-component, moisture-cross-linking polyurethane adhesives, depending on requirements (type of film, sterilization resistance, sealing seam strength, etc.). Besides the traditional solvent-based systems developments have been made in the direction of solvent-free systems (high-solids products) and of polyurethane emulsion-based systems.

In extrusion, a polymer melt is extruded onto a primary web. An improvement in adhesion or reduction in the extrusion temperature is obtained partly by applying an adhesive as primer to the primary web. Adhesives suitable for this purpose are similar to those used in a lamination. In coextrusion, two polymer melts are combined with one another. If required, a polymer primer is extruded between the two melts.

Bonding of Cellophane. Cellulose hydrate film is used for bag manufacturing or for wrapping operations. Because it readily takes up moisture in its unlacquered state, aqueous emulsion-based adhesives are used primarily for bonding. The addition of glycerol ensures they release their water only slowly so that the film stays flat. Lacquered cellophane is either bonded or sealed. The adhesives do not require any additives to keep the film flat because the film is protected by the lacquer. The adhesives are normally based on poly(vinyl acetate) copolymers or other flexible polymers. Where solvent-based or hot-melt adhesives are used, both types of film may be bonded with the same adhesive, because the absence of water eliminates the problem of keeping the film flat.

Manufacture of Polyethylene Sacks. The one-component and two-component polyurethane adhesives used for the production of laminated films (see above) are used for this purpose, depending on the type of machine.

Labeling of Glass. For labeling glass, dextrin, starch, and casein glues are used. With many machines, the adhesive used is required to show pickup ability, that is, the adhesive applied, for example, to a pallet must be tacky enough that it can pick up a label from the label magazine. Furthermore, the label must not slip after application, and any tendency to curl as a result of moistening must be counteracted. Dextrin glues are well endowed with these properties. In addition, by virtue of their high solubility in water, any glue that remains on the machine is easy to clean off. However, dextrin glues cannot be used where resistance to condensation, water, or ice is required (e.g., in breweries). Therefore, casein glues are mainly used in breweries. Although they are not as concentrated as dextrin glues (ca. 60 %), they develop a high initial tack, even on wet bottles, because they undergo a marked increase in viscosity on contact with the cold bottle. An additional reason for the use of casein adhesives is their wider adhesion range compared to dextrin adhesives. The latter adhere reliably to normal glass bottles, but their adhesion to coated or surface-treated glass bottles is not sufficient, so that casein adhesives have to be used.

Modern automatic high-performance labeling machines are capable of labeling between 500 and 1300 bottles per minute. At lower speeds, starch glues, which do not show such pronounced pick-up properties, are still used because of their lower cost. However, their higher water resistance can cause problems in bottle-washing machines for returnable bottles.

Labeling of Plastic Containers. The glues normally used for glass bottles cannot be used for plastic containers because they do not adhere sufficiently to plastic. Plastic bottles normally consist of polyethylene, polypropylene, poly(ethylene terephthalate), polycarbonate, poly(vinyl chloride), or polystyrene. Can-like containers and tablet tubes are made of polystyrene or polypropylene. Modified copolymer emulsion-based and hot-melt adhesives are used for labeling.

Although there are no major adhesion problems with PVC, PET, and PS, adhesives with stronger adhesive properties must be used for polyethylene and especially polypropylene. However, if the polyethylene surface is treated (e.g., exposed briefly to a gas flame) before labeling, it is possible to use the same adhesives as for PET or PVC.

9.3. Nonwovens Hygiene Industry

In general two types of adhesives are used for the manufacture of baby diapers, feminine-hygiene, and incontinence products. For simple construction purposes, hot-melt adhe-

sives based on amorphous poly(alpha-olefins) (APAOs) are used. High-quality constructions are produced with adhesives based on styrenic block copolymers. For bonding of elastic materials such as polyurethane foam, lycra and natural rubber strands, and hydrophilic and hydrophobic of nonwovens, pressure-sensitive adhesives (PSAs) based on styrenic block copolymers are used. Hot-melt adhesives are the most common types of adhesive. The most widely used synthetic block copolymers are styrene – butadiene – styrene (SBS), styrene – isoprene – styrene (SIS), and styrene – ethylene – butadiene – styrene (SEBS) with different styrene contents.

The adhesives are mainly applied intermittently (up to 650 pieces per minute for baby diapers and up to 1000 pieces per minute for feminine-hygiene products) at line speeds of up to 400 m/min. These requirements place high demands on the performance properties of the adhesives.

Construction Adhesives. APAO-based adhesives are mainly applied by slot nozzles on both the polyethylene backsheet and the nonwovens top sheet. Usual application temperatures are 125 – 140 °C.

The thermoplastic rubber-based adhesives are applied by application methods such as spiral spraying, fiberization, and melt blowing. Application temperatures are usually in the range of 140 – 165 °C.

Elastic Attachment Adhesives. These thermoplastic rubber-based adhesives are usually applied by spraying at 140 – 165 °C. As different elastic materials are used, different adhesives need to be formulated to give the right adhesion performance for each substrate.

Positioning adhesives for feminine-hygiene products are usually thermoplastic rubber-based, pressure-sensitive, and have low viscosities. The tack level can vary from very low to very high depending on the type of product and customer requirements. The positioning adhesive can be applied by both slot and fiberization applicators. The application temperatures depend on the type of substrate and vary from 120 to 170 °C.

9.4. Hygienic Papers

Four processing steps for the production of hygienic papers, (e.g. toilet paper, kitchen rolls) involve the use of adhesives:

Tube Winding. Mainly dextrin adhesives are used and/or plasticized homopolymeric poly(vinyl acetate) emulsions, but glutin glues are also used in special cases.

Pick-up adhesives bond the tissue to the Kraft paper tube. Depending on the machine systems the following adhesives are used: If the adhesive is applied directly to the tube by a slice (machine speed up to 400 m/min), aqueous colloidal solution systems at room temperature are normally used. If the adhesive is applied directly to the tube by a nozzle (machine speed up to 450 m/min), aqueous polyol-based adhesives are often used at room temperature.

When the adhesive is applied by a slice to the application roller, which transfers the glue to the tube (machine speed up to 650 m/min), hot-melt adhesives are used at 60 – 80 °C.

For high speed machines (up to 850 m/min), in which the adhesive is applied directly as a strip by means of a nozzle (temperature of the nozzle 80 °C), aqueous colloidal solution systems are used.

Laminating Adhesives. Aqueous adhesives based on cellulose or poly(vinyl alcohol) for the ply lamination of tissue papers must give clean machinability, good initial tack, and reliable bonding strength of the tissue laminates.

Tail Tie. For this application aqueous adhesives based on cellulose or poly(vinyl alcohol) are used. For the tail tie of tissue paper rolls, products with high wet tack, transparent film, and a low final bond strength are necessary.

9.5. Gluing of Wood and Wooden Materials

Today for the assembly gluing of wood, for gluing veneers, plastic sheets, and films, and for the manufacture of wood-based materials (chipboard, plywood, hardboard, profiles), synthetic

adhesives are used almost exclusively. The traditional use of adhesives based on natural products (glutin and casein glues) is confined to a few special cases only (for example, violin making).

Poly(vinyl acetate) emulsion "white glues" are used widely for the gluing of wood and wood-based materials. They generally contain small quantities of solvents and/or plasticizers to adjust the film-forming temperature and chalk as a pH buffer and filler. They have a solids content of approximately 50 – 65 % and a film-forming temperature of 0 to 15 °C. Special components may be added to increase the setting rate or to prolong the open time. The application weight is 100 – 200 g/m^2, the open time 5 – 25 min, and the clamping time 5 – 45 min for a compressive force of 5 – 20 N/cm^2. The exact figures strongly depend on the glue coating weight and the type of substrate.

Bonds with higher water resistance can be achieved by using special blends or two-component emulsion glues. The second component may be a diisocyanate or particularly an inorganic salt (usually, aluminum nitrate). These hardeners, added in a quantity of at least 5 %, improve the moisture resistance of the joint. However, their addition limits the pot life.

Normal glues develop EN 204/205 Class D2 strengths which are adequate for all indoor applications, for example, for the assembly gluing of furniture and for laminating plastic sheets. Self-cross-linking one- or two-component poly(vinyl acetate) glues, which develop Class D3 and D4 strengths according to the same European standard, may be used for gluing in moist atmospheres (bathrooms) and for exterior assembly gluing (windows, doors).

Furniture Foil Adhesives. These adhesives consist primarily of vinyl acetate – ethylene copolymer emulsions. Rigid PVC and melamine – paper foils, generally 0.2 – 0.4 mm thick, are applied to wood-based materials for decorative purposes. The adhesives must exhibit high tack and high resistance to heat and cold and must counteract the shrinkage of the film (application weight 60 – 100 g/m^2).

Hot-melt adhesives are mainly used for automatic edge-banding applications to bond edge materials to particle board, as well as for profile wrapping with foils, paper, and veneer.

Increasingly, hot-melt adhesives also are being used for the postforming process. Immediately after surface lamination, the overhanging laminate is bonded to the contoured panel edge without an additional adhesive joint.

Hot-melt adhesives are applied on edge-banding machines, profile-wrapping machines, and postforming machines. The adhesive is applied to the substrates with special rollers or slot nozzles. The joint between the cover and the core material is achieved by applying pressure with a roller system while the adhesive sets.

In assembly processes, such as furniture or door manufacture, the hot-melt adhesive is applied in spots or lines by hand or with automatic guns. Often the hot-melt adhesive is exclusively used as an additional assembly support due to its high initial strength, while poly(vinyl acetate) glue is used for achieving the final strength (two-shot process). Hot-melt adhesives for the edge-banding and postforming process are highly viscous products, often filler-free or with a filler content of up to 50 %. For profile wrapping, low-viscosity hot-melt adhesives are used, sometimes with fillers. For assembly bonding, a wide range of assembly hot-melt adhesives varying in viscosity, setting time, and open time is available for different equipment and uses.

As adhesives based on ethylene – vinyl acetate copolymers cover a very wide range of applications, they are used in large quantities. In the case of higher demands on the bond, polyamide- or polyolefin-based hot-melt adhesives are recommended. Highest bonding quality regarding heat and cold resistance, bonding strength, and waterproofness are achieved by using polyurethane-based, moisture-reactive hot-melt adhesives.

EVA hot-melt adhesives have heat resistance up to 85 °C, polyolefin and polyamide hot-melt adhesives up to 130 °C, and reactive polyurethane hot-melt adhesives up to 150 °C. The cold resistance lies between − 20 and − 30 °C; in the case of polyurethane-based hot-melt adhesives, it is even lower than − 30 °C.

Urea formaldehyde (UF) resin glues are the most widely used adhesives in woodworking and are preferably applied by the hot-glue technique. They are marketed as preconden-

sates in the form of 50 – 66 % solutions or as dry powders that must be dissolved in water. Curing aids can be added. In veneering, contact times as short as 30 s are possible for press temperatures of 100 – 120 °C. The most common additive is ammonium chloride. The amount of hardener governs both setting time and pot life. Glue joints in thick boards can be set by high-frequency heating.

For gluing plywood, the UF resins are generally mixed with extenders such as rye flour, followed by the addition of curing aids, the type and amount of which governs the setting characteristics at the processing temperature and pressure. The glues thus prepared are applied to the dry veneers (moisture content 6 – 12 %) by a glue spreader (application weight 120 – 200 g/m^2), and the glue-coated veneers are assembled and then glued in multilayer hot presses or high-speed single-layer presses at 100 – 160 °C and 60 – 200 N/cm^2, depending on the type of wood. For particle board production, unextended UF resins are used preferably. The particles are coated with glue by spraying, the particles of the center layer being coated with less resin than the particles of the two surface layers. Average glue resin content is ca. 10 % (i.e., 6 – 8 % for the center layer and 8 – 12 % for the surface layers). The glue-coated particles are first spread to form a three-layer cake, precompressed, and then compressed in hot presses to form particle board.

UF glues ideally satisfy the requirements for interior use but are unsuitable for exterior application. The requirement for low formaldehyde content must be considered by all producers of particle board. As plywood and veneer glues, with regard to water resistance, UF resins satisfy the requirements of DIN 68705 and EN 314, category IF, and as particle-board binders they meet the requirements of DIN 68763, category V 20.

The glues are cheap, easy to apply, and therefore suitable for the mass production of wood-based materials.

Melamine resin glues are important in the woodworking field when weatherproof glue joints according to DIN 68705, category AW at relatively low press temperatures (under 100 °C) are required. With added extenders moisture-proof glue joints can be obtained that meet the requirements of the same DIN category IF, at press temperatures above 75 °C.

Melamine – Formaldehyde (MF) Resins. The largest field of application for melamine formaldehyde (MF) resins is the manufacture of decorative laminates and the surface finishing of chipboard, hardboard, and profiles. Special papers with weights of 18 – 150 g/m^2 are impregnated with MF resins, dried, and then pressed onto the substrates at 130 – 180 °C. Their colorless transparency, hardness, chemical resistance, and absence of yellowing have made these resins particularly suitable for surface finishing. For technical applications including, e.g., concrete form-work boards, they are also applied by the wet-resin process, in which the resin solution is applied to the surface and cured under heat and pressure.

Phenolic resin glues, for cost reasons generally derived from phenol cresol mixtures, are used for the hot gluing of plywood, the manufacture of chipboard, and as binders for fiberboard. They are marketed in liquid form. Their shelf life is limited to about six months. They set at 130 – 180 °C, generally with addition of hardeners. For setting at room temperature, the addition of a special hardener (usually *p*-toluenesulfonic acid) is essential. The use of phenolic resins requires a relatively low moisture content of the wood (6 – 10 %).

Joints based on phenolic resin glues are of high quality and weatherproof. They are recommended for gluing wood to metal (metal-clad wood).

Resorcinol resin glues set without the assistance of heat under mildly alkaline conditions. Therefore, no damage is caused to the wood by acid hardeners, and there are no restrictions on the moisture content of the wood up to 18 %. The resins can be cured with paraformaldehyde. Wood-to-wood glue joints with resorcinol resins are among the strongest and are weatherproof and resistant to solvents, oils, and fats. For reasons of economy, mixed condensates of resorcinol and phenol are used instead of pure resorcinol.

Contact Adhesives. In countries with a high standard of mechanization, contact adhesives

based on chlorinated rubber for woodwork are used only for on-site repairs and for special purposes, e.g., for gluing plastic sheets or veneers onto rounded substrates. Here, contact adhesives are widely replaced by poly(vinyl acetate)-based wood glues. However, in less developed countries, contact adhesives are still widely used for veneering and laminating plastic sheets to wooden substrates. Their advantage over all other wood glues is the high bonding strength achieved virtually immediately after joining under short-term pressure (several seconds) without the need for sophisticated equipment.

Polyurethane Glues. The search for formaldehyde-free binders resulted in a new generation of polyurethane-based glues for chipboard production. The usage of polyurethane based adhesives for assembly of wooden materials has also increased. These liquid polyurethane adhesives are single-component adhesives that harden by a chemical cross-linking reaction which is initiated by moisture in the material. The water resistance of wooden joint fulfills the requirements of EN 204, D4. Furthermore, the adhesive joint shows high resistance to a variety of organic solvents, weak acids, and alkalis.

Over-all Quality Standards for Wood Glues. There are several national and European standards for the adhesives used in woodworking, of which the following are the most important: EN 204, EN 205, DIN 53255, DIN 68705, and DIN 68761 – 765.

9.6. Footwear Adhesives

One of the most important methods of assembling shoe components is adhesive bonding. In 1949 polychloroprene adhesives replaced the nitrocellulose solutions that had been introduced in 1906. The almost complete versatility of the polychloroprenes allowed leather, textiles, and other natural substances (until then the only shoe materials) to be supplemented by other materials, such as various rubbers and synthetics, or to be replaced altogether, as in the case of toe puffs, counter stiffeners, or heels. The way was thus opened for an astounding development in fashion and shoemaking techniques. Adhesive-based processes, supported by new machines and equipment, led to new bottom-attachment systems, simplified the production of footwear, and drastically reduced the number of production operations. Since 1970 a position of comparable importance has been reached by polyurethane adhesives.

Other adhesives for shoe manufacturing are solvent adhesives based on natural rubber or SBS or SIS block copolymers, dispersions of polyacrylates and poly(vinyl acetate), and a range of latex and hot-melt adhesives.

Closing Room Adhesives. These adhesives are light-colored and nonstaining solutions (or latices) of natural or synthetic rubber; other polymer dispersions are also possible. Wet-stick and pressure-sensitive bondings predominate. Also PSA hot-melt adhesives are of increasing importance for these applications. If higher strengths are demanded of the bonds, for instance, for replacing stitched structural seams, the requirements are met by two-way contact adhesives or by adhesives weldable in the high-frequency process. Ethylene – vinyl acetate hot-melt adhesives are used for applying backings and for sticking-in toe puffs and counter stiffeners, whereas polyamide hot-melt adhesives are used for folding, edge binding, including french binding, etc.

The components to be bonded (linings, interlinings and other reinforcements, and also toe puffs and counter stiffeners) can be precoated with adhesive and then ironed in.

The Lasting Operation. Apart from the bottom attachment, the lasting operation produces the most vital structural bond in the shoe. It involves the fastening of the upper material over the margin of the insole and is subjected to considerable stress; there is normally no supplementary stitching, tack-driving or stapling. A common lasting system is combined pulling-over and cement-lasting with automatic injection of hot-melt adhesive; this is mainly a rapid-setting crystalline polyester or the more flexible polyamide, both supplied in rod or granule form. Lasting by contact bonding with polychloroprene adhesives (or in light shoes also with natural rubber solutions and latices) is limited to particular constructions.

Bonding Soles to the Lasted Uppers. Polychloroprene and polyurethane solvent adhesives, occasionally also dispersions of the same polymers, are used for this purpose. Various pretreatments are used to improve the bond: application of a thin first coat of adhesive, especially for absorbent surfaces; priming, in particular the halogenation of vulcanized rubber and thermoplastic rubber soling; the use of base-coat primers for nylon and other synthetics, as well as UV-curing primers for EVA and other substrates in athletic footwear.

Depending on the required productivity, quality, and performance a wide variety of processes are used in sole bonding. As an example for a contact bond, the single-component adhesive is applied to the sole, usually in advance. The soles can be stored (dust-free) until required in the production process. On the lasted margin of the shoe upper, the adhesive is usually a two-component mix containing isocyanate hardener; bonding must be carried out within the specified open-assembly time.

For bonding with heat activation, the sole is activated by IR irradiation or rapid flash activation: the activation time may be 2 – 6 s, and a surface temperature 55 – 80 °C can be reached. During the process it is the adhesive film that is heated, rather than the sole, which remains cooler. This ensures a more rapid development of bond strength. Soft soles that tend to deform when hot can be bonded at room temperature or with adhesives that have a low activation temperature.

For the various molding-on processes, such as the direct injection molding of thermoplastic rubber or PVC, the vulcanizing-on of rubber, and the low-pressure reaction molding of integral-skin cellular polyurethane, adhesives can seldom be dispensed with: the lasted margin of the upper is coated with a polyurethane adhesive, often containing a pre-mixed latent hardener, or with a self-vulcanizing synthetic rubber solution compounded to match the vulcanization system of the soling compound.

With moisture curing polyurethane hot-melt adhesives a new innovative technology has been introduced in the sole-bonding process. It can be used as a one-way bonding system or as a two-way contact adhesive, applied either to the sole or to the upper. In this process, an adhesive coat must be applied on the other substrate.

Adhesives for Shoe Repairs. These adhesives must be almost universally applicable, because widely differing shoe types and material combinations must be repaired. For this reason, one-component polychloroprene-based adhesives are used almost exclusively, though for PVC and some other plastics materials polyurethane adhesives are quite common.

Testing. In Europe adhesive properties are tested according to European standards (CEN) EN 522, EN 1391, and EN 1392. Specified standard test materials can be obtained from the various research institutes of the European shoe industry.

9.7. Bonding of Plastics

Plastics are bonded to one another or to other materials. They differ widely in their affinity for bonding. Some plastics are easy to bond (e.g., PVC-U, ABS, PS), whereas others can be bonded only after special preparation (e.g., PTFE). On account of the low polarity and surface tension of certain plastics (e.g., polyolefins) there are still no adhesives for pipe joints which are durably resistant to shear strain and pressure. The choice of adhesive is determined by the substrate and the mechanical, thermal, and chemical requirements [69–75].

Poly(Vinyl Chloride). Rigid PVC (PVC-U) and chlorinated PVC (PVC-C) pipes, sheets, and films are used worldwide in the construction industry. They are joined by adhesives consisting of PVC solutions in a mixture of solvents, which operate by the principle of diffusion bonding. The bond is established by diffusion of the adhesive into the surfaces and the temporary dissolution of the PVC (cold swell welding).

For the bonding of uncalibrated pipe joints, adhesives with gap-filling properties are used, in which PVC of a certain molecular mass is dissolved in a solvent consisting predominantly of THF, cyclohexanone, or *N*-methylpyrrolidone. The viscosity range is 2000 – 3500 mPa s. The application time of the adhesive ranges from 1 to 5 min. The adhesives are preferably made thixotropic so that they do not run off the surfaces to be joined after application and to avoid

the damage to the plastic material because of accumulation of adhesive (the excess solvent can cause tensile corrosion). After preliminary cleaning of pipe and fitting with solvent (cleaning and dissolution) and application of the adhesive to both sides, the pipe and the fitting immediately are fitted together without turning. The bond is established by diffusion of the adhesive into the surfaces and temporary dissolution of the PVC (diffusion bonding).

PVC adhesives, usually consisting of solutions of post-chlorinated PVC in methylene chloride, are used in the case of inorganic acids as flow-through media in PVC pipe because of their better chemical resistance. Since PVC adhesives have no gap-filling properties a special application technique is required. This priming method is little used today.

Contact adhesives based on polychlorobutadiene or polyurethane are used for large-area bonds between PVC-U and PVC-C sheets and wood or metal.

Vessels are often lined with thin PVC-U and PVC-C films (1 – 2 mm thick). Contact adhesives are used with usual application techniques. For bonding cooling tower films, special adhesive solutions, which are harmonized in viscosity and thixotropy with the processing techniques (roll or pot application), analogous to PVC pipe adhesives, are used.

Adhesives for bonding plasticized (flexible) PVC (PVC-P) films are adhesive solutions based on THF (80 – 90 %) and PVC-P (10 – 20 %). Adhesives based on nitrile rubber or polyurethane in a plasticizer-resistant formulation for the bonding of PVC-P films made from differently formulated compounds are also suitable.

Adhesives based on polymethacrylates and vinyl copolymers are used for the continuous bonding of PVC-P films to metal sheets (coil coating). In this process, the adhesive is applied by machine to the cleaned sheets, the solvent is removed, and the film subsequently is rolled into the bed of adhesive heated to 140 – 180 °C.

Polystyrene, Styrene – Acrylonitrile Copolymers, and Styrene – Butadiene Copolymers. These plastics may be bonded to themselves or to one another simply by treating the surfaces to be joined with pure solvents, such as dichloromethane, ethyl acetate, or 2-butanone. However, this method of bonding, which also is known as solvent welding, does involve the risk of material corrosion. Therefore, it is advisable to use solutions of polymer raw materials in mixtures of organic solvents with a low risk of material corrosion. Depending on the field of application, especially in combining different materials, contact and nitrile-rubber adhesives containing aliphatic hydrocarbons as solvents also may be used. Pipes of ABS/ASA (ASA: acrylonitrile – styrene – acrylic ester polymer) for high-temperature wastewater run-off systems (max. 90 °C) preferably are joined with solutions of ABS/ASA in acetone or 2-butanone, although they also may be bonded with THF – PVC systems. The solutions are of medium viscosity (max. 2000 mPa s) and have a predominantly thixotropic rheology.

Solutions of ABS in 2-butanone are preferred for use in pressure applications, such as compressed-air conveying systems. As with PVC, the bond is established by diffusion.

Rigid polystyrene foam is used for the manufacture of composite sheets, for example, for containers and prefabricated building boards. The rigid foam is combined with metals, wood-based materials, asbestos, cement, and plastic sheets. Polyurethane adhesives are particularly effective in this field.

Polyolefins. Polyethylene, polypropylene, and polybutene can be bonded only after treatment to increase the surface energy, generally by oxidation, and make the surfaces receptive to bonding. Pretreatment can be carried out with an oxidizing flame, with oxidizing chemicals, or by electrical discharge.

Nowadays, the flaming process is rarely used: in the lining of containers with polyethylene and in the printing of bottles (Kreidel process). Among the chemical pretreatments, immersion in chromosulfuric acid (for a few seconds to a few minutes), followed by careful rinsing with water has proved successful. Thin plastic layers for the production of laminated films for use as packaging materials are best pretreated by corona discharge.

The pretreated polyolefins can be bonded with contact adhesives, epoxy resin adhesives, one- and two-component polyurethane adhesives, and hot-melt adhesives.

Polyamides are bonded with concentrated (ca. 85 %) formic acid. Two-component polyurethane or epoxy resin adhesives are suitable for bonding to metals, for example, for holding bearing bushings in place.

Polymethacrylates can be bonded by treating the surfaces to be joined with a mixture of dichloromethane and dichloroethylene. However, this does involve the risk of material corrosion. In such cases, polymerization adhesives cured by light can be used. Polyurethane, epoxy, and contact adhesives are also suitable.

Fiberglass-Reinforced Plastics. To obtain high-quality pipe connections, the surfaces to be joined must be roughened and the pipes carefully tapered. Polyester adhesives or thermosetting epoxy resin adhesives are best suited for bonding. A pressure of 50 – 100 N/cm^2 is applied to the joined surfaces.

Fluoropolymers. There are no known adhesives for fluoropolymers that give bonds of adequate strength without pretreatment. Simple, low-quality bonds can be obtained with contact adhesives. After chemical pretreatment, for example, with a solution of an alkali metal naphthalenide in THF, polyfluorocarbons can be bonded with high strength by using epoxy and polyurethane adhesives.

Curable Molding Compounds. High-strength bonds involving these plastics are obtained with epoxy resin adhesives. The adherents must be roughened lightly and thoroughly cleaned beforehand. Laminates, including decorative laminates, of phenolic resin- or melamine resin-impregnated papers or fabrics can be bonded with contact adhesives.

9.8. Bonding of Elastomers

In the rubber industry, a fundamental distinction is drawn between "bonding" and "adhesion". Whereas the adhesion of generally vulcanized rubber articles to one another or to metals or plastics is solely a function of the adhesive used, bonding is a process in which vulcanization of the rubber and, at the same time, adhesion between rubber and substrate are obtained by the action of heat and pressure (→ Tires).

9.8.1. Adhesion of Rubber

Contact Adhesives. Nowadays, rubber solutions based on natural rubber or styrene – butadiene rubber, optionally in conjunction with regenerate and depolymerized rubber, are used for large-area bonding where adhesion does not have to meet stringent requirements, as is the case, for example, with the lining of vessels, containers, etc. with rubber.

Today, multipurpose adhesives based on polychloroprene rubber generally are used as adhesives for rubber articles.

Reactive Adhesives. Epoxy and polyurethane adhesives are suitable for bonding polar elastomers, such as nitrile rubber and chloroprene rubber. However, epoxy adhesives form a relatively rigid bond so that they are suitable only for bonding to rigid substrates. In contrast, polyurethane adhesives form elastic bonds and therefore are used for rubber-to-rubber and rubber-to-leather bonding, for example, in the manufacture of profiles and in the shoe industry.

Polyurethane adhesives also are suitable for bonding nonpolar elastomers, for example, natural rubber, styrene – butadiene rubber, or ethylene – propylene terpolymers, after chemical pretreatment of the surface.

Polymerization adhesives based on α-cyanoacrylates (see page 22) are suitable for special bonds involving rubber, because they enable bonds of high strength to be obtained quickly and easily. Medium-viscosity types are preferred for rubber-to-rubber bonds by virtue of the minimal hardening of the joint. In the rubber industry, for example, the adhesives are used for bonding profiles to one another and also for repair work.

Because of the increasingly stringent requirements that rubber articles have had to satisfy over the past 20 years, the bonding process is preferred in the manufacture of rubber – metal composites.

9.8.2. Bonding of Rubber

Rubber-to-Rubber Bonding. In the bonding of unvulcanized rubber mixtures to one another, there is generally no need to use adhesives when the mixtures to be combined with

one another are identical or similar in composition. However, adhesives have to be used for combining different types with one another. Various bonding agents based on halogenated polymers are available commercially for this purpose and also may be used for bonding vulcanized to unvulcanized articles.

Rubber-to-Fabric Bonding. Rubber-to-fabric bonding is of considerable commercial significance in the tire industry and industrial rubber goods such as conveyor belts and rubberized fabrics. In the tire industry, the so-called resorcinol – formaldehyde (RF) latices are still used predominantly for bonding rayon, whereas mainly resorcinol – formaldehyde – vinylpyridine latices are used for bonding nylon and adhesively finished polyester yarns, which, unless specially pretreated, adhere very poorly to rubber.

In industrial rubber goods, which exhibit considerably greater variations in rubber composition, solutions of unvulcanized rubber mixtures are used. They are applied to the fabric as a first coat with the addition of polyisocyanates to improve adhesion. Directly adhering rubber mixtures containing resorcinol – formaldehyde donors and an acidic silica filler are also used for the production of fabric-reinforced rubber materials.

Rubber-to-Metal Bonding. In the Ebonit or hard rubber process [76], the bonding agent is a reaction product of natural or synthetic rubber with 30 – 40 % sulfur, which is thermoplastic in character, adheres firmly to metal, and provides for good adhesion to soft rubber. This process is still in use today in the lining of vessels and, to a certain extent, in the manufacture of large-diameter rolls.

The brass process makes use of the fact that electrodeposited brass of certain composition and crystal structure is capable of directly bonding rubber mixtures, providing their composition is compatible with the adhesion process [77]. Whereas today the brass process is of only minor significance for the production of rubber – metal composites, it is widely used in tire manufacture for bonding the bead wires and for bonding the carcass mixture of steel cord.

Certain polyisocyanate adhesives allow bonding of rubber to metal with far less technical effort than by the brass process. However, polyisocyanates are extremely sensitive to atmospheric moisture and react with compounds containing OH and NH_2 groups, such as antiagers. This reaction leads to a weakening of adhesion.

Nowadays, rubber – metal composites are produced with bonding agents in which the polymer base is usually a halogenated polymer, for example, a mixture of chlorinated and brominated polybutadiene. After the film of bonding agents applied to the metal component has dried, the metal component is bonded to the rubber mixture during the vulcanization process.

Mechanical and/or chemical pretreatment of the metal is essential for corrosion-resistant bonding (for the effect on bonding of the blasting medium used for blasting, see [78]).

The fillers in the rubber have no significant effect on bonding. However, soft rubber mixtures (30 – 40 Shore A) are more difficult to bond than hard rubber mixtures because of their high plasticizer content. The degree of vulcanization is another determining factor as far as the quality of bonding is concerned [79]. Bonding errors generally occur if the mixture is scorched before it comes into contact with the bonding agent surface or if the mixture is undervulcanized or severely overheated [80].

9.9. Bonding of Metals

The adhesives used for bonding metals are preferably reactive adhesives, predominantly epoxies, phenolics, acrylates, polyurethanes, although poly(vinyl chloride) plastisols, MS polymers, and rubber adhesives also are used for elastic bonds [81–85]. The range of bond strengths obtainable extends from high-strength, structural bonds (tensile shear strength ca. 40 N/mm^2) to highly elastic adhesive sealing compounds (tensile shear strength ca. 1 N/mm^2). However, the strength of a bond depends to a large extent upon the nature and direction of the forces acting upon it and on the temperature. Because the intrinsic strength of the cured adhesive resins is not comparable with that of metals, provision must be made for sufficiently large joint areas. Accordingly, the design of adhesive joints in metals must be appropriate for bonding. Resistance to aging and corrosion

depends on the nature of the surface pretreatment and on the type of adhesive used.

Epoxies, polyurethanes, and acrylates represent the principal types of adhesive for structural bonding of plastics and metals. These technologies compete for various applications in the transportation and marine industries because of their strength, good environmental durability and cost effectiveness. Historically, urethanes were chosen for low-temperature applications because their hard segment/soft segment structure typically allows them to retain flexibility in cold environments. Conversely, epoxies were preferred for high-temperature environments because of their strength and high cross-link density. Methacrylates fill the gap between these two technologies. They offer good resistance to high and low temperatures, and they provide the distinct advantages of rapid curing at room temperature, less sensitivity to mixing ratio and higher tolerance of surface contamination. They are proven performers in the transportation and marine industries. Typical applications include bonding fiber-reinforced plastics to metal.

The aeroplane industry was the first to incorporate metal bonding in assembly procedures on a significant scale. The increasing use of high-strength lightweight metals which cannot be satisfactorily welded and the reduction in overall weight through the introduction of light-gauge constructions were other important influencing factors. Light-gauge sheets (skins for fuselage and wings, control surfaces, etc.) could be stiffened more effectively by cementing onto corresponding profiles than by riveting. The problem of reinforced openings (windows, door frames) and of distributing stresses over large cross-sectional areas (rotor blades) was satisfactorily solved by multilayer construction, especially supporting-core and sandwich constructions, the widespread application of which actually was made possible by bonding. In this method, high-shear cores of various structures (e.g., honeycomb), are bonded to thin cover layers or skins of aluminum and other materials. This gives a lightweight, nondenting, torsion-resistant construction suitable for aircraft floors and wings. Suitable adhesives are phenolic – poly(vinyl formal) and epoxy adhesives and also epoxy – nylon adhesive films.

Vehicle manufacture (automobiles, railroad vehicles) is another significant application for metal-to-metal bonding.

In *lightweight metal constructions,* tubes and hollow profiles of any cross section can be designed and bonded as socket joints. Door and window frames are made from steel and aluminum profiles with angles bonded in place (epoxy adhesives), a heat-insulating intermediate layer being applied by bonding or made by casting with an adhesive (polyurethane, epoxy adhesives).

In *bridge building,* adhesive joints (epoxies, polyesters) are used in combination with high-strength, on-torque threaded fasteners (to absorb peel forces) for load-bearing steel constructions.

In the *electrical industry,* it is above all the bonding of sheet packs (dynamo sheets, transformers, motors) and the fixing of small parts (ferrite cores) that are increasing in significance (epoxy resins, cyanoacrylates). Adhesives (epoxy resins) containing conductive additives also are used, for example, for printed circuits.

In *machine construction* and *instrument* and *tool manufacture,* plain bearings, axle bearing guides, bushings, etc., are made by bonding, and punches are cemented into stamping tools with epoxy resins, cyanoacrylates, and acrylate adhesives. Screws, gear wheels, shafts, etc., are secured with anaerobic adhesives.

Adhesive pastes are used as *repair kits,* often in combination with woven glass-filament cloth for reinforcement. In this way, bodywork is repaired, worn surfaces are renewed, and pipes are sealed (epoxies, unsaturated polyester resins).

The sandwich construction method mentioned above involves joining *metals to other materials* by bonding. Further examples include the bonding of brake linings (phenolic adhesives) and compound materials in ski manufacture, where aluminum is bonded to plastics, wood, etc. (phenolic and epoxy adhesives). Highly alloyed steels, beryllium and titanium alloys, and other special metals can be bonded with adhesives (e.g., polyimides, polybenzimidazoles) that have comparable high-temperature resistance.

9.10. Adhesives for Wallcoverings

Wallcoverings, classified by EN 233 – 235 [133], consist of a variety of materials: paper, textiles, glass, plastics (PVC, PE, PS, PU), ceramics, wood, cork, metals, leather. They are attached by hand in the form of sheets (wallpapers), tiles, or profiles to masonry, plaster, concrete, plasterboard and gypsum wallboard, wood, and metals.

For ease in do-it-yourself projects (adjustability, open time, easy application, etc.) and because most substrates to which the coverings are bonded are capable of absorbing moisture, it is preferable to use water-based adhesives in the form of solutions and emulsions. Solvent-containing adhesives are only used on nonporous substrates, for example, metals [86], [87].

The adhesives are based on methyl cellulose, starch ethers, carboxymethyl cellulose, pregelatinized starch, poly(vinyl acetate), polyacrylate, styrene – butadiene latices, polychlorobutadiene, and polyurethanes. The viscosity of the adhesives covers the range from fluid to pasty.

Wallpapers include those with a PVC or metal surface and rough-textured plain wallpapers. They are applied with pastes of pregelatinized starch, starch ethers, carboxymethyl or methyl cellulose in the form of 2 – 10 % solutions. For rough-textured plain wallpapers and heavy wallpapers, moisture resistance and tack are increased by utilizing poly(vinyl acetate) redispersion powders. High solid, emulsion-based adhesives, optionally in conjunction with resin-enhanced pastes (special purpose pastes), are used preferably for heavy vinyl- and metal-faced wallpapers.

Light *textile fabric wallcoverings* are hung with special fabric wallcovering adhesives. Special-purpose pastes containing 10 – 20 % of emulsion-based adhesives also are used.

Heavy wallcoverings of unbacked PVC (1000 g/m^2 or more) can be applied with adhesives based on acrylate or poly(vinyl acetate) copolymer emulsions. The same applies to *metal foils* that are applied underneath wallpaper as a vapor barrier.

Polystyrene and *polyurethane foam* and *mineral wool tiles* are applied to interior and exterior for decoration and insulation purposes (sound and heat). Solventless adhesives based on poly(vinyl acetate), polyacrylate, or styrene – butadiene copolymer emulsions are preferred for this purpose. The viscosity of the adhesives must be adjusted to give the required wet tack.

Ceramic wall tiles are applied either with cement mortar adhesives or with adhesives based on styrene – butadiene copolymer or polyacrylate emulsions. Tile adhesives must be pasty and thixotropic to prevent the tiles from slipping but allow positioning.

Wood panels, chipboard, and similar wooden materials can be fixed with aqueous adhesives based on poly(vinyl acetate) or polyacrylate emulsion. The adhesives must be pasty and have high initial strength. In some cases they are applied from cartridges. Also contact adhesives, special reactive systems (hybrid silicone), and hot-melt adhesives are utilized.

9.11. Floorcovering Adhesives

Flooring adhesives are used for attaching floor coverings to a variety of substrates. The adhesives must develop a strong and durable bond between the floor covering and the substrate. They should not adversely affect coverings, supports and substrates, and after application, should remain odorless and have low emissions. Usually flooring adhesives are applied to the substrate by hand with trowels or sometimes rollers.

Environmental and occupational-health legislation has resulted in a major shift from organic solutions to aqueous emulsions of polymeric compounds [88], and the use of waterborne adhesives is accepted as the state-of-the-art for most floor covering installations. Waterborne flooring adhesives are often based on acrylic copolymers, vinyl acetate – acrylics, vinyl acetate – ethylene – acrylics, or styrene – butadiene rubber latex. They commonly contain natural (e.g., colophony and colophony esters) or synthetic resins (e.g., hydrocarbon resins) and inorganic fillers. In recent years the content of solvent has been reduced to below 5 wt %. The use of high-boiling solvents (glycol ethers, liquid hydrocarbons) has been adopted to some extent for the formulation of solvent-free flooring adhesives [89]. There is now a move to eliminate high-boiling substances and even traces of volatile contaminants from the

adhesives [90]. Several European manufacturers have established a chamber test method to record VOC and to classify low-emission products for flooring [92], [91]. Today solvent-based products are only needed for special applications (e.g., water-sensitive substrates).

Linoleum Adhesives. Linoleum generally is laid with waterborne adhesives based on acrylic emulsions or SBR latex. They contain higher amounts of inert inorganic fillers, and natural and/or synthetic resin is used to accelerate initial tack. Linoleum adhesives have to set rapidly to avoid expansion or shrinkage of the covering. Two-component adhesives consisting of one polymer dispersion and a second cement preparation are available to adhere linoleum to nonabsorbent substrates.

Parquet adhesives can be cold- or hot-spread compositions. The former can be classified as waterborne, solvent-borne, or reactive. The first two types are commonly based on poly(vinyl acetate) emulsions in water or as solutions in organic solvents. Reactive systems are solvent-free two-pack polyurethane adhesives or epoxy resin adhesives. They are preferred for all parquet and wood types which are sensitive to water. All adhesives are required to show high cohesive strength but certain plasticity to absorb movements of the wood through shrinking and swelling in response to changes in atmospheric humidity without failure of the joint. Hot-spread adhesives, which are bituminous or tar-based, are increasingly being replaced by the cold-spread types.

Adhesives for Plastic Floor Covering. Solvent-free adhesives based on acrylic and vinyl acetate – ethylene – acrylic emulsions are used for bonding vinyl, synthetic rubber, and polyolefin flooring. Solvent-containing contact adhesives (rubber- or polychloroprene-based) still are used for some applications (e.g., stairs). When water resistance is required, reactive polyurethane or epoxy resin adhesives are preferred to waterborne emulsions.

In addition to developing a certain minimum strength, the adhesives primarily have to meet demands on the dimensional stability of the covering.

Because of plasticizer migration between covering and adhesive joint, the adhesives for plastic floor coverings should be adjusted carefully in their composition. Plasticizers are commonly used for the production of coverings based on PVC.

Adhesives for Textile Floor Covering. The composition of waterborne adhesives for bonding wall-to-wall carpets is similar to that of linoleum adhesives. High initial tack (wet tack) and a resistance against carpet cleaners is required. Previously used solutions of poly(vinyl ethers) and rosin in mixtures of highly volatile solvents are still used for carpets that are extremely sensitive to water.

Conductive Adhesives. Conductive adhesives are available for bonding conductive floor coverings. The electrical properties are achieved by the addition of carbon fibers, carbon black, or graphite. Conductivity of a floor covering is sometimes required to avoid undesirable static charging.

Tile Adhesives. The thin-set process uses hydraulically hardening thin-set mortars based on cements, emulsion adhesives based on aqueous synthetic resin emulsions, which can be single-component ready to use tile adhesives or two-component adhesives consisting of modified aqueous emulsions and a cement-based compound, and reactive adhesives based on epoxy resins or polyurethanes.

Depending on the intended field of use, cement-based tile adhesives are modified with different amounts of redispersible synthetic resin powder such as co- or terpolymers based on vinyl acetate or acrylates. The recipe must be adjusted to the different types of tiles and natural stones. On account of the various types of tile adhesives corresponding standards may be found, for example in DIN, BSI, AFNOR (Association Française de Normalisation), UEAtc (Union Europeenne pour l'Agrement Technique dans la Construction) or as European Standards prepared by CEN.

9.12. Building Construction Adhesives

The use of adhesives in building construction imposed exacting requirements upon quality and

application techniques [92–98]. In addition to durability and resistance to mechanical stress, temperature and weather resistance are prime requirements. In prefabricated construction, the use of adhesives was an important requirement for efficient assembly techniques.

Adhesives as Additives for Hydraulic Compositions. Synthetic resin dispersions and redispersible synthetic resin powders that are added to cement-containing concrete mixes and mortars must be resistant to alkali and dilutable with water. Redispersible powders are added to cement-based adhesives in factory made mortars, while synthetic resin dispersions are mixed with the powder compound on the building site just before the adhesive is applied. The products that have been specifically developed for this purpose include, in particular, copolymers of vinyl acetate in combination with other esters or with ethylene and terpolymers of vinyl acetate, vinyl chloride, and vinyl laurate and also copolymers based on acrylates, methyl methacrylates and styrene, as well as styrene – butadiene. The additives modify the properties of the concrete. For example, they considerably improve the adherence of new concrete to old concrete. In addition, resistance to dynamic stress is considerably increased. The addition of anionic melamine resins is recommended also for establishing certain rheological properties. Adhesives based on cellulose ethers are added to cement-based systems and gypsum plasters to improve adhesion to the substrate and, by virtue of their hydrophilic character, retain water, thereby preventing mortars and plasters from drying out and improving their strength. Starch and starch ethers are used to modify the adhesion of cement mortars to mineral substrates.

Prefabricated Construction. Sandwich elements are widely used in prefabricated construction. In this case, bonding must meet stringent requirements with regard to stability and durability, especially in the case of exterior wall elements. Deflections attributable to differential temperature or moisture-induced expansion or contraction can amount to as much as 5 % of the length. The adhesive layer must be permeable to water vapor or act as a water barrier, depending on the structure and function of a given element. The adhesives used are two-component reactive resins, based mainly on epoxy resins, but polyurethane resins are used as well, as are elastomer-based contact adhesives that are applied on both sides.

With building components based on synthetic resin concrete (polymer concrete), the cement may be replaced partly or completely by organic binders, such as co- or terpolymeric vinyl acetates, acrylates or epoxy resins. Such systems are described as polymer cement (PC) if the cement as binder has been totally replaced with organic binders, and as polymer cement concrete (PCC) if the cement binder has been partly replaced by organic binders. These systems are mainly used to reface concrete surfaces.

Low-viscosity reactive resins are used to seal cracks in concrete constructions. These epoxy and polyurethane resins show excellent creepability and bond strength to concrete.

Assembly Adhesives. In carcass work, cement mortar is still widely used as a binder for mineral components. The disadvantages of cement mortar include its long curing time, its poor adhesion to old concrete, and its poor resistance to dynamic stress. For these reasons, adhesive mortars have been developed. They are mixtures of liquid, curable synthetic resins, such as epoxy resins or unsaturated polyester resins, blended with mineral additives, generally quartz, and the corresponding hardeners. They establish an excellent bond with hardened concrete, stone, metals, and certain plastics, and are used for applications involving particularly severe stress, for example, for cementing rail supports onto concrete sleepers, for fixing dowels and anchor bolts in concrete, for steel reinforcements, and for the permanent bonding of high-grade concrete elements.

For applications such as swimming pools and water-purification tanks, for which chemical resistance is required, or where resistance to dynamic stress is needed, special epoxy resins are used. Blends of a waterborne synthetic emulsion with a cement compound are suitable for applications with lower requirements. In some cases, e.g. sewage-treatment plants, elastomer-modified bitumen mortars are still used.

Another important application of reactive resins is the production of resin-bonded screeds that, by virtue of their high mechanical and chemical stability, are used, for example, for industrial floors. One particular problem is the

bonding of old concrete to new concrete, where peeling occurs because of the shrinkage of the new concrete layer. By applying intermediate layers of a suitable epoxy resin mortar, it is possible to establish permanent bonds with strengths greater than that of the concrete.

9.13. Adhesives for Bonding Textile Fabrics

Fabric bonding is a relatively new application for adhesives [99–103]. The particular demands that a fabric bond must satisfy make the replacement of conventional thread stitching very difficult:

Adequate bond strength
Resistance to washing and cleaning
No staining or discoloration
No spoiling of texture
Maintain breathability (in the case of spread adhesives)
Rapid setting

Adhesives are used widely in the apparel industry for applying interlinings. The interlinings or nonwovens are precoated with hot-melt adhesives. To maintain breathability, the adhesive is applied spot-wise rather than spread. Suitable adhesives for these and similar purposes are, for example, copolyamides of the 6,6/6,12 type and also boiling-resistant low-pressure polyethylene. 6,36/6,12/6-Terpolyamides based on dimeric acids are used because of their high flexibility, resistance to washing, and favorable application properties. Their softening range is generally from 100 to 150 °C. After cutting to size, the precoated interlining materials are fused using ironing presses (front fixing).

The introduction of other inexpensive thermoplastics into the fabric bonding field has met hitherto with little success. Polyesters, for example, are not sufficiently resistant to dry cleaning. By contrast, polyurethanes show high resistance and favorable bonding properties.

Another area in which adhesives are very widely used in the textile industry is the manufacture of chemically bonded nonwovens (→ Nonwoven Fabrics). The nonwovens are preformed by the dry-layer or wet-layer process and subsequently bonded by spraying or impregnation with adhesives. The binders used generally are products based on acrylates, synthetic rubber, and vinyl polymers. Nitrile and styrene – butadiene rubbers are used for industrial nonwovens because of the need for resistance to oils, fats, and organic solvents.

9.14. Flocking Adhesives

Flocking adhesives are used for bonding short (0.3 – 5 mm) monofilament textile fibers (known as flock or flock material) to a variety of substrates [104]. Compared to mechanical flocking processes, in which the flock is scattered or blown onto a layer of adhesive, electrostatic flocking has the greater importance. In electrostatic flocking, suitable finished fibers are oriented in an high-voltage electrical field and accelerated to such an extent that they penetrate deeply at one end into the open adhesive coating and are firmly anchored therein. The surplus, nonanchored flock is blown or brushed away after drying and setting of the adhesive.

Whereas flocking was formerly carried out for decorative reasons (imitation suede or velvet), technical and economic aspects (noise insulation, friction reduction) also have become significant. Major improvements in flocking adhesives undoubtedly have contributed to this situation. In addition to paperboard, woven fabrics, yarns, films, and wallpapers, plastics, plastic foams, metals, rubber profiles, glass bottles, and, occasionally, even interior walls and car bodies are being flocked to an increasing extent. The adhesives used for flocking primarily are emulsion-based adhesives, generally based on poly(vinyl acetate) or polyacrylates and their copolymers, in some cases with added cross-linking agents. However, solution adhesives (solutions of synthetic resins) and also solventless and solvent-containing reactive adhesives (epoxy and polyurethane adhesives) are also used.

The adhesives must satisfy various selection criteria, of which adhesion, strength, elasticity, and stability of the set adhesive film are the most important. In many cases, colored adhesives also are required. The adhesives generally are applied by spray guns or rolls. They must have sufficiently long open times and spread evenly on the substrate.

Because the adhesive represents the positive pole (earth) in the electrostatic flocking process, through which the charge must be dissipated rapidly, it is required to show a certain conductivity (maximum resistance $10^7\ \Omega$) This is not a problem where water-based adhesives are used, but in the case of solvent-containing adhesives, the necessary conductivity often can be obtained only by means of additives.

9.15. Adhesives for Bonding Glass

The bonding of mineral substrates generally requires high-polarity, low-shrinkage adhesives. These two requirements are satisfied by reactive adhesives [105–111]. In many cases, the bonding of glass additionally requires high transparency and UV stability of the adhesive for optical reasons. The adhesives primarily used are polar polymers such as epoxy resins and polymethacrylates in the form of two-component systems. For large-area compound bonding on an industrial scale, for example, for automobile windshields and bullet-proof glass, glass sheets are combined with laminates of poly(vinyl butyral) and plasticizer. The permeability of transparent glasses to radiation also allows the use of UV-hardening one-component systems based on methacrylates. Elastic glass-to-glass bonds and similar bonds between glass and other materials are obtained with rubber adhesives based on silicone and polysulfide.

The epoxy resins occupy a leading position as glass adhesives. Their hardening at room temperature can be varied over a wide range. Different curing systems are available for this purpose. Thermosetting epoxy resins that harden in a few minutes at 120 – 150 °C are used on industrial scale.

In certain cases, improvements in adhesion can be obtained by additionally priming the glass surface. The primers used are bifunctional products, for example, organosilanes. The adhesives used for bonding utility glass also are required to be resistant to boiling water. Organosilicon epoxy resins have proved to be effective for this purpose, and are used either on their own or in conjunction with conventional epoxy resin adhesives. In addition to their high resistance to water, siloxane – epoxy resin adhesives are of relatively high thermal stability.

9.16. Adhesives in Automobile Manufacture

Introduction. A variety of different adhesives, sealants, and coating compounds with a wide spectrum of properties, from high-strength structural adhesives with energy-absorbing properties (crash resistance) to highly elastic surface-coating compounds, are used in modern automobile manufacture. The boundary between adhesives and sealants is fluid and cannot be defined precisely. It is intended here to discuss only products that are applied for bondings and/or sealings between two parts or in gaps (e.g., in weld seams) which are able to transmit constructive strength, not for surface coating purposes, such as underbody protection.

Although the quantity in which the adhesives thus defined are used per vehicle depends largely on the model, the average value in Europe appears at present to be between about 8 and 12 kg.

For adhesives used in the body shop, strength and elasticity are tested from – 40 °C up to + 90 °C, whereas aging behavior is investigated under temperature load, in humidity, or in aerosols of salt solutions (salt spray test). There has been a steady increase in the use of alternating climate tests either with regular cycles between dry cold and humid heat (VDA) or with a single change from prolonged storage in humid heat to dry cold.

Because automobiles are manufactured in large numbers of up to a few thousand units per day, the adhesives must be formulated in regard to their processability such that they fit into the established production flow and its cycle time. In view of the increasing advance of automation and the use of robots for body assembly, ease of pumping through pipes, often of considerable length, and also appropriate rheology at the application stage are essential requirements. As far as possible, it is preferred to use one-component formulations with wash off resistance that are cured by the heat of the electrocoat, primer, and paint bake ovens. At the trim line, where no bake ovens are available, mainly one-component, moisture-curing or contact adhesives are used. Two-component products will acquire higher significance for bonding of plastics and for assuring rapid through cure. For new

production concepts, the cure speed of trim-shop adhesives will be very important.

Body and Paint Line. On the body line, all adhesives and sealants are applied to the steel surface covered with oil that is used for deep drawing and for corrosion protection. Accordingly, the adhesives and sealants used must be able to absorb the oil. In addition the adhesive must bond to a large number of surface coatings such as zinc, zinc – nickel, electrolytic zinc, fire zinc, prephosphated zinc, and organic coatings.

As hem-flange adhesive on hood and bonnet and on doors, high-strength adhesives are used in order to provide these components with the necessary inherent stability. High impact performance imparts improved crash resistance to the bonded parts and safety for the passenger cell. In addition, the adhesives afford effective protection against corrosion in these particularly sensitive areas (PVC plastisols, acrylics, epoxy resins). For further improving protection against corrosion and to obtain a cleaner appearance, the closed hem-flange seam is sealed with a sealing bead of PVC or acrylic plastisols; some automobile manufacturers are changing their sealing concept (clean paint shop), and these operations are allowed only in the body and/or trim shop. In the hem-flange seam, adhesives have largely replaced the hitherto customary spot welds.

For joining anti-flutter panels to the outer panels of bonnet and trunk, adhesives of relatively low strength and high elasticity are used both at the assembly stage and after electrocoating (PVC plastisols, rubber-based adhesives). These properties are necessary to insure that the bond seams do not "show up" on the outer surface (read-through effect). In this area, adhesive bonding has paved the way for constructions that cannot be obtained by other joining techniques. Rubber-based tapes or compounds are used for joining roof bows to roof panels.

For better protection against corrosion in spot-welded seams, the seams are treated additionally with an elastic adhesive/sealant (spot-weld sealer) based on rubber. Hot-and cold-application systems are in use.

Sealing tapes and/or sealing compounds (rubber-based, vulcanizing with different degrees of expansion or permanently plastic) are used for sealing in the region of fenders, wheel housings, sliding roofs, air louvers, tank supports, and roof linings. Die cut pieces (rubber-based) are used in some areas (e.g., door hinges, mirror mounts). Highly expandable (500 – 1000 %) rubber or EVA-based materials are used for sound absorption in cavities.

To give washing out resistance, pasty adhesives must be solidified by pregelling or precuring before they enter the cleaning process by passing them through a recirculating air oven, IR heaters, or induction heaters. Cleaning is followed by pretreatment (phosphating), electrocoat application, and oven baking at 180 °C. Various weld seams (drip rails, inner seams, outer seams) are later sealed with PVC or acrylic plastisols. The next step is the application of surface primer and final paint with the associated bake ovens.

Trim Line. In the trim shop, the fresh paint surface forms the substrate for a number of bonding operations in which parts prefabricated to a greater or less extent and subsupplied parts are adhesively bonded into the painted body. Because no more bake ovens are involved, cold-curing systems (moisture or two-component systems) are employed.

At the glazing stage (windshield, rear, and side windows), the insertion of a tacky sealing cord (rubber-based) into the elastomeric profile surrounding the glass panel provides for better sealing. In the last 20 years, there has been an steady increase in the use of direct glazing techniques in which a moisture-curing, one-pack polyurethane adhesive joins the glass panel to the painted body. Advantages include very effective sealing, the possibility of larger screens, and lighter body constructions (the glass panel now contributes to body rigidity via the modulus of the adhesive), and more creative freedom (lower air resistance). More recently adhesives for radio antennas integrated into the rear screen have been developed with high electrical resistance. Fast setting materials eliminate any fixing aids and guarantee short drive-away times.

Contact adhesives (polychloroprene, nitrile, or polyurethane rubber) are used for fitting vinyl roofs, roof linings, sound-deadening mats (where they are not already incorporated in the primer-coated body), and carpets. More recent prefabricated roof constructions are designed for bonding with special emulsions.

The waterproofing polyethylene film behind the door panel is sealed with pasty compounds or tapes (rubber-based). Folded lining films and hems on doors and carpets are sometimes secured with hot-melt adhesives (polyamide, polyethylene) because they set relatively quickly.

The mechanical spot fastening of the buffer strips around the body frequently is supported by the "underplacement" of a continuous tape, tacky on both sides (rubber based).

Subsupplied Parts. These parts are generally manufactured by subcontractors and are installed in the painted body in the course of trim line at the automobile assembly plant.

In the interior of the vehicle, prefabricated roofs consist of outer "decorative layers" that are bonded to the layers imparting mechanical rigidity (emulsions, contact adhesives, polyurethane hot-melt adhesives). Contact adhesives, polyurethane adhesives, and high-frequency welding auxiliaries (resin solutions or PVC-based emulsions) are used for attaching textiles or plastic films to the supports made of molded fiberboard or plastics for door inner panels. Emulsions, contact or polyurethane hot-melt adhesives are used for bonding textiles to molded fiberboard or plastics for parcel shelves.

On the exterior of the vehicles, rubber profiles around windows and doors and the trunk lid are "lap" bonded with cyanoacrylate adhesives because continuously extruded profiles cannot be used for sharp corners. To obtain effective sealing and better sliding in the case of sliding windows, these profiles are flocked by means of polyurethane-based flocking adhesives. The lenses of headlamps are joined to the housing by polyamide or polyurethane hot-melt adhesives and epoxy resins. The layers of laminated safety glass are bonded with a film based on poly(vinyl butyral).

Engine, Transmission, Chassis. In engines and transmission systems, components that must transmit torques, such as gear rims, are bonded with anaerobic adhesives (diacrylate based). The same group of products is used for thread locking. Gaskets formed in place in these areas consist of silicones or rubber – phenolic resin.

Rubber-to-metal bonding agents are used for the production of vibration-damping members of rubber and metal, developing their adhesion to the metal during vulcanization of the rubber. The same applies to the production of shaft seals. Clutch and brake linings are applied by means of nitrile rubber – phenolic resin adhesives.

In oil and air filters, the membranes are embedded in adhesives and/or fixed to metal frames (PVC plastisols, polyurethanes).

Future Trends. Short- and medium-term trends, such as the increasing use of plastic (sheet- or injection-molded compounds of glass-fiber-reinforced, unsaturated polyester resins, glass-fiber-reinforced polyurethanes, and polycarbonates) and lightweight alloys such as aluminum and magnesium will require new concepts for adhesives and sealants. New curing mechanisms are being sought to meet future production concepts (modular assembly). In addition the need to separate the multimaterial mix by disbonding processes for recycling and repair are special targets to be met by modern materials. Robots will guarantee further automation, lower baking temperatures will save energy, and the partial introduction of organic-coated steel panels for improved corrosion resistance will require slightly more versatility of the adhesives.

In the long term, materials for improved safety and comfort of vehicles of lightweight construction are required.

9.17. Adhesives in Aircraft Construction

The principle attractiveness of adhesive bonding in comparison to competing joining technologies (e.g., screws, riveting, brazing, welding) for the aircraft industry is mainly due to the following aspects:

High aerodynamic surface quality of bonded components due to smooth and precise contours

Superior properties with respect to stiffness, fatigue, and damage tolerance due to area-spreaded load transfer (no notches) with damping characteristics (vibration, sound)

Nearly universal and simple joining of different materials without any weakening (minimal thermal exposure/stresses, surface damage)

Simple realization of extremely lightweight sandwich and multilayer designs

Gas- and liquid-tight joints (no crevice corrosion)

All these advantages can be utilized for saving weight or increasing life expectancy. Also reductions in production cost may be realized.

Disadvantages are:

Strength properties are comparatively highly influenced by the environment (temperature, moisture, chemicals)
Some processes require relatively high production standards
Nondestructive testing is demanding and of limited significance
Technology-specific destruction mechanisms exist (delamination, bondline corrosion)

However, many of the risks combined with the application of bonding can already be eliminated by means of careful and experienced design.

The application of adhesives in early aircraft manufacturing was not very different from the general evolution of this joining technology. For example, variations in the stability of wooden constructions against moisture due to less suitable adhesives such as casein were experienced. However, impressive parts like the wooden wing of the pre-World War II Fokker F-36 airliner (33 m wing span) were bonded with no in-service problems reported.

The gathering of specific know-how in the aircraft industry started with the transition to metal structures made of aluminum. This step was first made during World War II in the United Kingdom in the light bomber "Hornet". In this aircraft some structural aluminum parts were adhesively bonded with newly developed thermosetting phenolic adhesives instead of the cold-setting acid-hardening urea – formaldehyde type formerly applied with wooden components. This two-component system, a liquid phenolic resin plasticized with poly(vinyl formal) powder, has to be cured at 155 to 175 °C and quickly found its way into civil aircrafts like the DeHavilland "Dove" and the first jet airliner "Comet" [189]. During investigations into the crash series of the latter model, the bonding technology was suspected to be the reason for catastrophic structural failures. The investigations ended with a complete rehabilitation of the bonds, and the adhesive is still in use in its original form or as a one-component precured (B-staged) film adhesive (e.g., skin panels of the Avro Regional Jet).

The most successful application example of this technology is the Fokker F 27/F50 "Friendship", in which about 70 % of the structure (ca. 550 parts) are bonded. The proven durability (> 1000 aircraft in service for up to 30 years) of this design is also due to the superior fatigue properties compared to a riveted construction. Additionally the chromic acid anodizing (CAA) of the aluminum surfaces before bonding resulted in corrosion-resistant joints.

Far less successful were further developments in the 1960s, when epoxy-based adhesives were introduced because of their higher temperature resistance and increased peel strength. Since no difference in mechanical properties were found between anodized and chromic sulfuric acid (CSA) etched aluminum, the advantage of low production costs was exploited. Delaminations of bonds after only two years in service and bondline corrosion events led to the recognition that the etch treatment produced an inconsistent oxide surface. Long-term salt-spray tests revealed the whole to be insufficiently stable to corrosive environments, especially as experienced in the bilge area of aircrafts. Expensive repairs and design modifications were the consequence.

The new standards then established [190], [191] included, besides the reintroduction of anodizing processes, careful edge protection, mandatory use of corrosion-inhibiting primer, and improved epoxy film adhesives with a curing temperature of at least 120 °C; this represents the technology mostly used up to today. Typical design rules require shear stresses of 10 MPa (static) and 1 MPa (dynamic). Any peel stresses should be avoided, but for safety purposes the bonds should bear about 4 N/mm in the floating roller peel test (e.g., EN 2243-2 [192]). In shear stress/shear strain tests (e.g., EN 2243-6 [193]) these adhesives show besides a nearly linear response up to at least 20 MPa and a failure strain of greater than 0.5.

Other lessons learned refer to the design of aluminum sandwich components consisting of a honeycomb core skinned with sheets. This extremely rigid and lightweight design is economically producible only by adhesive bonding, which also achieves superior stability against

sonic fatigue. Durability against corrosion can only be attained if the parts remain completely airtight in service or, in case of acoustic panels around the engine, which are regularly heat dried during service, are well ventilated by perforations with access to all cells. Preventive anodizing and priming with corrosion-inhibiting primers also of the core material has not been applied very often till today due to the higher material costs. Progress in the bonded-sandwich design of the acoustic panels has made a substantial contribution to reduction in noise emissions of jet engines in the last two decades.

Besides the chemical industry, the aircraft industry also pioneered the utilization of titanium alloys and developed adhesive bonding as an option for joining workpieces. Similar to aluminum, the pretreatment process is decisive for successful and durable bonding. Simple etching with fluorine-containing acids resulted in bonds which are highly prone to moisture intrusion by diffusion and adhesive failure. Alkaline etching processes are a better alternative, and most durable bonds are obtained with anodizing procedures [194], [195]. The reliability of adhesively bonded titanium is indicated by the fact that the joints between the titanium fuselage fittings and the carbon fiber reinforced plastic (CFRP) wing skins of the US fighter aircraft F18 "Hornet" are adhesively bonded.

To a certain extent the aircraft industry is returning to its roots by applying more and more fiber-reinforced plastics, starting with G(glass)FRP, going to A(aramid)FRP, and now mostly CFRP, because these materials possess some similarities to wood. For all FRPs the load-spreading properties of adhesively bonded joints is a key issue, and successful adhesive bonding requires careful surface conditioning. Mostly two processes are used: the peel ply technique and mechanical abrasion (sanding).

The peel ply technique integrates economical production of FRP parts with the advantage of obtaining reproducible surface conditions by using an outer woven ply (generally made of polyamide) during cure of the composite part. This ply, which is generally impregnated with an excess of the matrix resin of the composite part during the curing process, is removed (peeled off) just before the subsequent bonding process is started by applying the adhesive to the mating surfaces. A prerequisite for the successful application is that the peel plies are free of release agents and other contaminants.

More versatile are sanding processes. Especially the brushing of the surfaces with Scotchbrite pads and a slurry containing quartz particles and tensides cleans and prepares the substrates reliably. Prepregs can be directly bonded to such surfaces on cured laminates without using additional adhesive.

The workhorse adhesive materials for structural bonding in the aircraft industry are still the modified epoxy film adhesives curing at 120 °C or, for parts exposed to elevated temperatures (> 80 °C), 175 °C. Modification normally means the addition of other polymers that form a secondary phase of microscopic particles in the bondline and act as crack stoppers by distributing the stress of an advancing crack tip in the surrounding volume. For this purpose carboxy-terminated butyl nitrile (CTBN) rubbers are often used, but the addition of thermoplastics (e.g. polyamide, polysulfone) is also common. The epoxy resin itself is often made out of oxirane-terminated bisphenol A monomer and oligomer fractions. Additives are used to improve temperature and fire, smoke, toxicity (FST) properties. Curing agents are generally alkaline (amines, amides), often supported by accelerators, especially in the case of the adhesives curing at 120 °C.

Only for specific tasks with higher thermal stresses are polyimide-based adhesives utilized. Special products can withstand permanent in-service temperatures of 250 °C and more, but the application is economically often less attractive, so that such bonds are found mainly in military and aerospace components.

The most important room temperature (RT) curing two-component adhesives are also based on epoxy chemistry. These versatile materials are not only used for bonding, but also for other purposes like shimming, potting, encapsulation, and even impregnation of fibers and fabrics. Their domain is semistructural and interior joining. Because of early failures, their usage in structural applications (see above) has been restricted, but modern knowledge about failure mechanisms provides a basis for a return to the structural field. More recent developments have created materials with performance properties very similar to those of the tried-and-tested film adhesives. Especially the toughening properties

were improved, and maximum service temperatures generally range from 80 to 175 °C. Hence these materials are today a powerful tool for economical repair and maintenance tasks.

9.18. Adhesives and Sealants in Electronics

Introduction. The use of adhesives in the electronics industry is experiencing a period of rapid growth at the moment. Adhesives are increasingly expected to perform a variety of tasks in electronics assembly, including conducting or insulating electricity, conducting heat, sealing, and protecting, besides the more usual functions of mechanically fixing components and assemblies.

Conductive Adhesives.

Isotropically electrically conductive adhesives (ICAs) are widely used in the electronics industry when high-temperature soldering processes are unsuitable. Typical applications include silicon die attachment, component attachment in multi-chip modules (MCMs), surface mounted printed circuit board repair, and electromagnetic/radio frequency (EMI/RFI) shielding.

In all cases the conductive adhesive consists of high contents of conductive filler in an insulating polymer (adhesive) matrix. The choice of filler and adhesive matrix is dependent on the end use.

Die attachment adhesives are normally epoxy- or polyimide-based and contain 70 – 80 wt % silver flake. Silver flake is the common choice for conductive adhesives as the oxide layer formed on the flake is also conductive.

Polyimides became popular for die-bonding adhesives because they are cleaner (in terms of ionic contaminants) than equivalent epoxy products. The market, however, is dominated by high-purity epoxy adhesives. Polyimides are always applied from solvent solution and require higher curing temperatures than epoxies. They are stable to higher temperatures. Solvent-borne thermoplastics pastes (or cast-film preforms) have been used for some lower reliability die-attachment applications.

Epoxies form the basis of products in the multi-chip module (MCM) bonding and surface mount technology (SMT) repair areas. Two-pack room-temperature products (amine cure) are common in repair applications. Two-pack systems with acid anhydride curing are available as heat-cure products where a more robust final product is required. These two-pack systems are often pre-mixed by the vendor and supplied as a one-pack system requiring low-temperature storage. One-pack latent-heat-cure systems are also available based on dicyandiamide or imidazole adducts.

Electromagnetic/radio-frequency interference shielding materials have to meet much lower demands in terms of overall electrical conductivity (typically 4 – 5 orders of magnitude lower than a silver-flake-filled adhesive). This means that cheaper conductive fillers can be employed, for example, silver-coated copper flake, nickel flake, and carbon black. Typically the adhesive has to form a compliant joint between two mating surfaces, and hence room temperature vulcanizing or heat-cure silicone is often a convenient choice of matrix material.

Anisotropically conductive adhesives conduct in one direction only. They are often referred to as z-axis conductive adhesives. These materials have found widespread use in LCD assembly due to their ability to resolve small gaps between conductor tracks and the fact that solder is not suitable for joining on glass substrates.

The filler loading in anisotropically conductive adhesives is much lower than in conventional isotropic adhesives (typically 5 – 10 vol % as compared to >30 vol % for ICAs). The fillers consist of spherical particles that are often monodisperse; examples include gold-plated polymers, solid gold, silver, and nickel. The filler loading is carefully chosen to insure that the probability of short circuits between adjacent conductors is low but that there is high probability of conduction along the z-axis. The adhesive matrix is usually epoxy-based and the most common product form is B-stagable film. These films are cured for short times at high temperature (20 s at 180 °C) under pressure.

Thermally conductive adhesives are commonly used as an interface between heat sinks and heat sources (e.g., high-power semiconductor devices). Their function is to give a mechanical integrity to the bond between the heat sink

and heat source, but more importantly to eliminate air (which is a thermal insulator) from the interface area.

These adhesives consist of a polymerizable liquid matrix and large volume fractions of electrically insulating thermally conductive filler. Typical matrix materials are epoxies, silicones, urethanes, and acrylates, although solvent-based systems, hot-melt adhesives, and pressure-sensitive adhesive tapes are also available. Aluminum oxide, boron nitride, zinc oxide, and increasingly aluminum nitride are used as fillers for these types of adhesives. The filler loading can be as high as 70 – 80 wt %, and the fillers raise the thermal conductivity of the base matrix from $0.17 - 0.3\,\mathrm{W\,m^{-1}\,K^{-1}}$ up to about $2\,\mathrm{W\,m^{-1}\,K^{-1}}$.

Coating and Protection Products.

Conformal Coatings. A conformal coating is a thin polymeric layer applied to a printed circuit board (PCB) by brushing, dipping, or spraying as the last processing step. The coating provides a barrier which protects the PCB (conductors, solder joints and components) from environmental attack during service life of the product. Not all of the circuit board is coated; areas such as connectors, switches and relays are covered with tapes, rubber caps, or removable masks to prevent the conformal coating impairing their operation. Conformal coatings are available in a number of different formulations; these include polyurethane (solvent and nonsolvent versions), silicones, acrylates, acrylated urethanes, and epoxies.

The choice of coating depends on the chemical and mechanical (abrasion) resistance required, the degree of rework envisaged, and the method of curing. Polyurethane coatings have excellent humidity, chemical, and abrasion resistance. However rework can be difficult. Silicones give the greatest protection during thermal excursions but they are unsuitable where abrasion resistance is required. Epoxies provide excellent chemical and mechanical resistance, but the inherent mechanical stiffness of the coating can be detrimental to stress-sensitive components. Rework is difficult. Acrylated urethanes offer a combination of fast cure and excellent solvent resistance.

Parylene is a completely different class of coating. The *p*-xylylene monomer is stable as a gas at low pressure but polymerizes spontaneously on any surface on which it condenses. The polymer formed has excellent moisture, chemical, mechanical, and electrical properties, but expensive application equipment is required.

Encapsulants encompass a broad range of materials whose primary function is to protect electronic components from detrimental chemical, mechanical, electrical, or thermal environments.

Encapsulants are usually characterized as either potting compounds, glob-tops, or underfills, depending on the end use. Potting compounds are materials that are used to protect final assemblies (relays, component terminals, electrical contact assemblies, etc.). Glob-tops and underfills are specifically used for silicon die protection.

Three different potting compounds are commonly used: silicones (for low-temperature applications), polyurethane (for applications up to 120 °C) and epoxies (for general applications). Potting compounds should provide good adhesion to the assembly materials/substrates, low cure shrinkage (and hence low internal stress), and in large sections, low heat release on curing.

Glob-tops offer protection to silicon die wire-bonded directly onto a substrate (PCB, polyimide flexible circuitry or ceramic). The liquid encapsulant is designed to flow between the wire bonds to form a protective coating over the integrated circuit. In some cases the pitch between wire bonds necessitates the use of low-viscosity liquids, and in these cases a dam of compatible adhesive is first dispensed around the IC to prevent the glob top spreading to other areas of the PCB. Typically, glob tops will have low levels of ionic contaminants and low coefficient of thermal expansion (CTE). The CTE is an important consideration in all encapsulant applications; the primary requirement is that the CTE of the encapsulant closely matches that of the silicon die, substrate, and wire bonds to minimize stress in the joint (particularly under conditions of thermal cycling).

Underfills are a specific class of adhesives designed to protect silicon dies which are soldered active face down onto the PCB. In these flip-chip applications, the underfill material flows beneath the die by capillary action. These materials are generally highly loaded with inorganic fillers to reduce the coefficient of thermal expansion.

Anhydride-cured epoxies give the best combination of properties for most glob-top and underfill applications, in particular their high glass transition temperatures are desirable. Due to the need for single-component materials (for ease of handling and bubble-free packaging) the anhydride and epoxy are premixed. The resultant one-pack system is stored at $-40\,°C$ to give an acceptable product shelf life and a pot life of 8 h (or more) at room temperature. In some (lower reliability) glob-top applications UV curing acrylates or UV cationic epoxies have been used.

Surface Mounting Adhesives. Adhesives are commonly used to hold electronic components in position prior to wave soldering. One-component heat-curing epoxies are the most usual choice of surface mounting adhesive (SMA). These materials are thixotropic pastes based on bisphenol A or F type liquid epoxy resins and latent curing agents. The curing schedule of these materials is typically 3 min at 120 °C, with a six month (or longer) shelf life at 5 °C.

Acrylic SMAs form a smaller class of products. These materials are cured with both UV and heat, the UV curing mechanism insures rapid fixing, while the bulk of the adhesive (under the component) requires a heat-curing step. The shelf life of these products is long (12 months at ambient temperature) but overall product performance is inferior compared to the epoxy analogues.

Various adhesive application methods are available (dispensing, pin transfer and screen/stencil printing) and different adhesive rheologies are required in each case.

Adhesives for Liquid Crystal Display Manufacture. There are three main adhesive applications during the manufacture of a liquid crystal display (LCD): temporary plate fixing, main seal, and end seal. A number of LCD cells are made from single large flat glass plates. In the first process, the mainseal (or LCD gasket) is printed onto the glass to define individual cells, this is usually a solvent-borne epoxy. Once the solvent has been removed, a second glass plate is placed above the first and the mainseal is cured. To hold the two plates in the correct orientation while the mainseal is fully cured a temporary fixing adhesive is used. This is often a UV acrylic material. The final production step involves filling the cell with liquid crystal through a gap left in the mainseal and subsequently sealing the gap. This end-sealing operation is the only occasion when liquid adhesive comes into contact with the liquid crystal and hence the choice of adhesive is vitally important. Specially formulated UV acrylic formulations are most widely used for this application.

9.19. Medical Adhesives

Introduction. The use of adhesives in medical applications has a long history and may date back some 4000 years. Today, adhesives are used in a wide variety of surgical techniques albeit on a limited scale. The type of adhesive used depends on its application. It has to meet certain criteria with regard to its strength, toxicity, degradation pathway, and overall safety and efficacy. This applies equally to pressure-sensitive adhesives used in simple wound dressings to more advanced sealants that are used to plug small holes in blood vessels following cardiovascular surgery. Almost all branches of medicine have experimented with the use of adhesive to replace or augment traditional techniques. This section briefly reviews some of the more commonly used adhesives in medical and dental applications.

Tissue Bonding and Wound Sealing. Skin acts as a barrier to infection and dehydration. When it is breached in the form of a cut it is desirable to have it sealed quickly and effectively. Cyanoacrylates (CAs) have been developed as an alternative to the standard method of suturing (stitching) a wound [196]. CAs have the advantage over sutures that complete tissue closure can be effected, thereby preventing further fluid loss. Normally higher ester forms of CAs such as butyl are used. These result in less tissue inflammation than the methyl and ethyl esters, possibly due to slower degradation. As the polymerization process is exothermic, care must be taken to ensure that the setting speed is not too rapid, otherwise thermal damage to the surrounding tissue may occur. This is circumvented by using very small amounts of the adhesive and by using higher esters. The rate of polymerization

is related to the size of the ester alkyl group. Generally, the larger the alkyl group, the longer the setting time, and hence there is more time for the heat generated to dissipate. Sterilization of CAs can be problematic since it often results in premature polymerization. Sterilization of CAs by dry heat and by gamma irradiation has been described in the literature.

Trade Names.Indermil, Loctite Corporation and Histoacryl, B. Braun, Aesculap, B.B.D. (based on cyanacrylic butyl ester); Dermabond, Closure Medical Corporation (based on cyanacrylic octyl ester).

Fibrin [197] is the body's natural glue or sealant and is produced in response to injury and bleeding. It helps form a clot, thereby sealing damaged blood vessels. Fibrin is formed by the cleavage of fibrinogen (a larger, soluble, blood-based protein) by the enzyme thrombin. This causes a structural modification resulting in the precipitation of fibrin to form a fine mesh which traps escaping blood cells. Fibrin acts well as a sealant and as a hemostatic agent, buts lacks the bond strength of synthetic chemical adhesives such as CAs. The individual components involved in this process can be purified from blood and be used in a variety of medical applications. It is especially useful in cardiovascular surgery.

Although they have quite different properties CAs and fibrin have been used for the same application. For example, both have been used to attach skin grafts to wound sites; to control bleeding and promote hemostasis; in vascular and intestinal anastomoses; and in ophthalmic surgery. Fibrin, being naturally derived, is more biocompatible. However, it is derived from blood and there is concern about transmission of pathogens. Fibrin(ogen) can be sterilized by a combination of either solvent/detergent or heat treatment to inactivate viruses, followed by ultrafiltration.

Trade Names.Tisseel, Immuno AG; Beriplast, Beringwerke.

Adhesives for Protective Dressings and Related Applications. Pressure-sensitive adhesives (PSAs) serve many applications in this area. Acrylates are the preferred elastomeric components of PSAs for general dressings. Acrylic hot-melt PSAs are useful for nonwoven web carriers due to their limited penetration. Specially chosen PSA formulations are used where allergic skin reactions are known or likely. Silicone-based PSAs may also be used in sensitive applications such as the attachment of prostheses and monitoring equipment probes. PSAs are also used to attach drug-delivery systems to the human body. Where the PSA forms a layer between the delivery source and the body tissue, particular care is needed to ensure adhesive function while facilitating drug availability. The design of such products depends largely on the size of the drug molecule.

Orthopedic Adhesives/Bone Cements. Acrylic bone cements [198] are the only group of materials used to anchor long-term implantable devices to the neighboring bone. Though the bone cements have drawbacks, including trauma created by a highly exothermic hardening process, over 90 % of hip and knee repairs function well for 15 years. The bone cements consist of separate powder and liquid components which are mixed carefully prior to application.

Though many products exist, the powder component always contains methacrylate polymer, a polymerization initiator, and a radiopaque medium. The liquid component contains methyl methacrylate with small amounts of accelerator for the hardening process. Hand mixing may be used. However, many proprietary mixing devices exist to give optimal properties and reproducibility.

Trade Names.Simplex, Howmedica; Palacos, Zimmer.

Dental Adhesives. Adhesive technology, cements and composites have added greatly to restorative dental surgery. Restoration of teeth with dental fillings is now often carried out with resin-based composites. These consist essentially of an organic methacrylate functional resin matrix, inorganic fillers, and coupling agents. Successive products have enhanced properties, in terms of the curing process, ease of use, and wear resistance. Light-curing systems are now popular since they allow relatively unlimited working time followed by rapid hardening when exposed to light of the appropriate wavelength. The bonding of crown materials is usually done with similar compositions, though here systems which cure by mixing reactants are necessary since light only reaches peripheral areas.

Glass – ionomer cements have taken a major place in dental treatments as restorative filling materials and also in a range of more adhesive applications due to their ability to bond to both dentine and composite fillers. Acid-etching techniques are well established for the bonding of resins to enamel.
Trade Names.P 100 (Composite Dental Adhesive), 3M; Dyract (Glass Ionomer Cements), Dentsply.

9.20. Household Adhesives

Household adhesives are used for gluing, assembling, or repairing a variety of materials in the household. They include a large number of different adhesive types, usually offered in smaller packages such as tubes, bottles, or tins.

The following list provides a survey of adhesive types frequently used in households:

Paper adhesives in solid (glue sticks), liquid, or cassette form (adhesive rollers)
Multipurpose adhesives
Adhesive tapes and pads
Cyanocrylate adhesives
Hot-melt adhesives
Two-component adhesives
Contact cements
Plastic adhesives
Wood glues
Assembly adhesives

The term do-it-yourself adhesives has also become generally accepted besides the term household adhesives. A clear differentiation between the two ranges of adhesives is not possible. The above-mentioned household adhesives also include DIY adhesives. In addition to the above applications, DIY adhesives are also used for light renovation work in and around the house and are therefore also offered in larger packages or for activities such as wallpaper hanging. They include:

Building chemicals such as tile adhesives, joint mortars, fillers, leveling compounds, and plasters
Wallpaper pastes
Flooring adhesives
Sealants
Polyurethane foams

Glue Sticks (Paper Adhesives, Solid). Glue sticks are used for gluing paper, cardboard, photos, and labels. Most of these sticks are solvent-free and contain a soap gel as builder and natural and/or synthetic polymers as adhesive component. They are also marketed as refillable sticks.

Of late, higher performance glue sticks based, for example, on polyurethane, which can be used to glue other materials such as wood, plastics, and metals have become available.
Trade Names.Pritt Stick (Henkel, Germany); UHU Stic (UHU, Germany); Tesa Klebestift (Beiersdorf, Germany); Fueki Stick (Fueki, Japan); Glue Stic (Avery Division, USA); Glue Stic (Grand & Toy, Canada); Kores Glue Stic (Kores, India); Glue Stick (Scotch 3M, France).

Paper Adhesives, Liquid. Liquid paper adhesives are used in the household for gluing paper and cardboard. These mostly solvent-free adhesives are partly based on natural raw materials such as dextrin, starch, and cellulose derivatives (pastes). They are sold in bottles and tubes but also as pens.
Trade Names.Pritt Pen (Henkel, Germany); UHU klebefix (UHU, Germany).

Adhesive Rollers (Paper Adhesives in Cassette Form). Adhesive rollers are newly developed adhesive systems for gluing paper, cardboard, and photos. The adhesive film is applied to a paper or foil carrier which is rolled up and placed in a cassette, usually with automatic take-up. This adhesive film is then rolled onto the parts to be glued on a solid surface. Adhesive rollers have the advantage over conventional paper adhesives that no drying time needs to be allowed after application of the solvent- and water-free adhesive film. Adhesive rollers are offered in various sizes, sometimes with refill cassettes. They are available for permanent and nonpermanent applications.
Trade Names.Pritt Roller (Henkel, Germany); Uhu multi roller (Uhu, Germany); Tombow (Japan).

Multipurpose adhesives are very versatile and can be used to bond most materials such as cardboard, paper, felt, textiles, wood, metal, glass, and plastics. They are also used for small-scale assembly or as a handicraft adhesive. Depending on the material, the bonding strengths

achieved are different and in most cases on a medium level.

The solvent-free multipurpose adhesives contain transparent to opaque dispersions or solutions based on polyurethane or acrylate. In addition, white dispersions based on poly(vinyl acetate) or acrylate (so-called white glues) are available for all-purpose applications. The solvent-containing types mostly consist of poly(vinyl acetate), vinyl acetate copolymers, or nitrocellulose, dissolved in a solvent mixture of esters, ketones, and alcohols.

Trade Names.Pattex Haushaltskleber transparent (Henkel, Germany); Pritt Alleskleber (Henkel, Germany); Uhu flinke Flasche (Uhu, Germany); Tesa Vielzweckkleber (Beiersdorf, Germany); Elmer's school glue (Borden, USA).

Adhesive Tapes and Pads. Adhesive tapes make use of foils, paper, or fabric as a carrier material which is coated with a permanently sticky film. In the household these tapes, which can be bought in different types and sizes (one-sided or two-sided adhesive tape), are primarily used for packaging, for small-scale assembly, and for masking when doing renovation work.

By contrast, two-sided adhesive pads consist of foamed carrier materials or extruded caoutchouc pastes and are used to fix lightweight objects on to smooth, level surfaces.

Adhesive pads for easily removable bonds that do not require a carrier material and are particularly suited for fixing posters, graphics, and other lightweight objects, even to sensitive substrates such as wallpaper.

Trade Names.Pritt Klebepads (Henkel, Germany); Tesa Klebebänder (Beiersdorf, Germany); Tesa Power Strips (Beiersdorf, Germany); Duck Tapes (Manco, USA); Scotch (3M, USA).

Cyanoacrylate adhesives are used in households for repair work. As these adhesives cure very fast and hardly bridge larger cracks, they are only suitable for small, snug fitting parts. Cyanoacrylate adhesives can be used to bond almost any material, both porous and smooth.

The main ingredient of this adhesive is a cyanoacrylic acid ester (methyl, ethyl, or butyl). Special esters such as methoxyethyl esters are employed for the odorless types. Cyanoacrylate adhesives are mainly available in small aluminum tubes or plastic bottles at different viscosities ranging from highly liquid to gel-like.

Trade Names.Pattex Blitz-Kleber (Henkel, Germany); Super Attak (Loctite, USA); Uhu Sekundenkleber (Uhu, Germany); Krazy Glue (Toagosei, Japan); Aron Alpha (Toagosei, Japan); Cyanolit (3M, USA).

Hot-melt Adhesives. In household applications, hot-melt adhesives are used in cartridges together with the necessary gluing pistols. They are suited for repair and small-scale assembly work, being able to bond almost any material. The special advantage of hot-melt adhesives is that the bonded parts already can be subjected to stress or further processed after a few minutes. A certain drawback is their high processing temperature of ca. 200 °C, which renders the bonding of temperature-sensitive materials such as styrofoam or plasticized PVC more difficult. Lately, so-called low-melt cartridges have come in use which can be processed at a temperature as low as ca. 110 °C. Furthermore, mini-pistols are offered which accommodate smaller cartridges, colored cartridges, or cartridges with a glitter effect.

Those hot-melts that find application in the household and DIY applications are usually based on ethylene – vinyl acetate copolymers (EVA) and contain tackifying resins (e.g. hydrocarbon resins). Today's pistols are equipped almost without exception with a mechanical feed, differ in their heating power, and are partly designed for cordless work.

Trade Names.Pattex Patronen (Henkel, Germany); Uhu Klebepatronen (Uhu, Germany).

Two-component adhesives that are used for household applications differ quite strongly as to their properties. For instance, there are epoxy resin adhesives that are very well suited for bonding glass, china, ceramics and metals. They dry glass-clear, are dishwasher-safe, and therefore exceptionally well suited for repairing dishes or similar household articles. Usually, the two components resin (epoxy) and hardener (amine or mercaptan) are filled in aluminum tubes or plastic double syringes and are mixed before application according to the directions for use. Other two-component adhesives are based on methacrylate and contain a powder hardener (dibenzoyl peroxide in gypsum).

These adhesives bond materials such as stone, plastics, ceramics, and metal; they achieve very high strength and bridge cracks and are therefore well suited for repair and small-scale assembly work.

Further two-component adhesives are available based on polyurethane; here, the resin component contains a polyol (filled or unfilled) and the hardener component a polyisocyanate. These adhesives, too, are suitable for bonding metal, plastics, ceramics, and wood. Special highly filled types can be used both for bonding and as a filler; after curing they can be processed like wood.

Two-component adhesives are used in the household when more difficult materials such as metals, glass, or certain plastics are to be bonded, when high strength is required, or when larger gaps must be bridged.
Trade Names.Pattex Stabilit Express (Henkel, Germany); Pattex Kraft-Mix (Henkel, Germany); Uhu plus (Uhu, Germany); Scotch Weld (3M, USA); Bond E (Japan).

Contact Cements. In the household, contact adhesives are used in many areas for gluing and repairing. They are mainly used when larger and heterogeneous items (e.g., wooden materials with decorative laminates) are to be joined, when difficult parts (e.g., veneer strips on the curves and edges of furniture) are to be fixed, or when flexible materials such as rubber or leather (e.g., shoe repair) must be bonded or repaired. Furthermore, contact adhesives can be used to bond materials such as cork, felt, plastics, soft foams, and metals.

Most contact adhesives contain a solvent and have a formulation that combines a base of synthetic rubber such as polychloroprene or polystyrene – butadiene with reactive phenolic resins and metal oxides. In addition, there are also transparent types based on polyurethane which provide excellent bonding results for soft plastics such as plasticized PVC (used in many household articles). A solvent-free generation of contact adhesives, e.g., based on acrylate, has entered the market. This type can be used to bond solvent-sensitive materials such as polystyrene foam.

Contact adhesives are sold either in cans or tubes with different viscosities (liquid or gel-like). Their application differs from that of other adhesive systems in a few points. The adhesive is applied to both parts to be joined, after which the solvent or water is allowed to evaporate (5 – 40 min, depending on the adhesive). After the evaporation time, bonding is possible within a defined period (open time) by pressing both parts briefly but with the highest possible pressure together.
Trade Names.Pattex Kontaktkleber (Henkel, Germany); Uhu greenit (Uhu, Germany); Elmer's (Borden, USA), Cemendine (Japan).

Plastic adhesives are used in the household or for model-making. They are able to bond nearly all plastics, for example, polystyrene, ABS, plexiglass, polycarbonate, and rigid PVC. Plastic adhesives partly belong to the solvent-welding adhesives. As they contain solvents such as butyl acetate, the surface to be bonded is slightly solvated, and this results in a solvent-welding effect. Plastic adhesives are mostly offered in small bottles, partly equipped with fine-dosing nozzles.
Trade Names.Pattex Plastic (Henkel, Germany); Uhu Plast (Uhu, Germany).

Wood glues are generally ready-for-use, water-based white glues based on poly(vinyl acetate). In the household they are used for assembling or for repairing wooden materials. After applying the glue, the wooden parts to be joined must be cold- or hot-pressed for a certain amount of time. In addition to the conventional wood glues, the market offers special fast-curing and waterproof types, also based on reactive polyurethane.
Trade Names.Ponal Holzleime (Henkel, Germany); Uhu coll (Uhu, Germany); Dorus Holzleim (Dorus, Germany); Elmer's wood glue (Borden, USA), Rakoll (Fuller, USA).

Assembly adhesives are relatively new adhesive systems which also find application in households. Assembly adhesives are able to produce very high bonding strength and high initial tack so that the parts to be joined need not be fixed. They are used whenever the powerful fastening of mostly different materials is required (both small and large areas), i.e., in cases where nailing, screwing and drilling were necessary up to now.

Typical applications include for example the bonding of wooden panels, skirting boards, decorative and insulating panels, plasterboards, tiles, etc. Moreover, it is also possible to bond difficult parts such as mirrors onto absorbent or nonabsorbent substrates, styrofoam sheets on metal, or nonferrous metals.

Assembly adhesives are either available as dispersions or as one-component reaction adhesives in cartridges or tubes.

Trade Names.Pattex Montageklebstoffe (Henkel, Germany); Mastic colle neoprene (Rubson, France); One for all (Bison, The Netherlands); PU-Fixkleber (Compacta, Germany); Sikaflex 221 (Sika, Switzerland); Montagekleber MK (Fischer, Germany); Fix all (Soudal, Benelux); Draft basters (Manco, USA); Multifix (Quilosa, Spain).

9.21. Applications of Anaerobic Adhesives.

Applications of anaerobic adhesives [199] are as follows:

Thread locking: The first applications for anaerobic adhesives were for locking threaded fasteners against vibration. Filling the "inner space" between a nut and bolt with a hard, dense material prevents self-loosening [200].

Thread sealing: The effect of filling the space between threaded parts or the space between inner and outer pipe threads provides a seal which can prevent the leakage of oil and other fluids in machinery as well as prevent corrosion of the threaded parts.

Retaining: Cylindrical press-fits and bearing assemblies can be retained with anaerobic adhesives, allowing accurate alignment and relaxed tolerances. Retaining and sealing of cup plugs and oil seals in castings is a major application.

Impregnation: Powdered metal parts, porous castings, and welds can be sealed against leakage of liquids or gases. This impregnation can also allow such parts to be plated and improves their machinability.

Preapplied Films: Thread lockers and sealants can be coated on threaded parts in the form of a dry-to-the-touch film. These preapplied materials remain inert until assembly releases a quick curing resin or an initiator.

Gasketing: Anaerobic flange sealants can be applied manually or by automated methods such as tracing, stenciling, and screen printing. These products can replace a variety of preformed, precut gaskets, and they can also be used as a gasket dressings.

Structural Bonding: Tough structural bonds can be achieved with some anaerobic adhesives for bonding components such as ferrite magnets, honing stones, identification tags, and decorative inserts.

10. Economic Aspects

More than 1500 adhesives manufacturers worldwide offer a variety of not less than 250 000 products. The worldwide turnover of the adhesives industry is believed to be approximately $\$16 \times 10^9$ corresponding to a total consumption of 6×10^6 t. These figures do not include cement-based products and urea and phenolic resins used a binders for the production of chipboard. Market studies and national industry federation reports are available for the United States, Canada and Mexico, Latin America, western Europe, eastern Europe, Asia/Pacific, Japan, and Australia. The regional distribution of the world adhesives market is given in Table 1.

Table 1. Regional distribution of the world adhesives market

Region	10^6 \$	1000 t
Canada and Mexico	413	187
United States	5799	2277
South America	302	130
Western Europe	3733	1470
Eastern Europe	887	323
Middle East	314	114
Africa	230	86
Asia/Pacific	1882	654
Japan	1925	585
Australia, New Zealand, Oceania	175	65
World total	15 660	5891

Adhesives have a very wide spectrum of applications. They are applied in almost every industrial field. The largest markets are found in the paper, packaging, and board industry (35 %), the building, construction, civil engineering and

craftsmen industry (28 %), and the woodworking industry (15 %). The forecasts for the future growth rates vary. The highest growth rates are expected in the automotive industry and the packaging and the converting sectors.

Considering the product groups, the highest growth rates are expected for hot-melt systems and polymer dispersions and emulsions. A continuing decrease will occur for adhesives based on natural polymers (animal and vegetable), solvent-based adhesives, and systems based on water-soluble polymers.

11. References

1. *Neue polymere Werkstoffe 1969-1974,* Hanser Verlag, München - Wien 1975.
2. M. Michel: *Adhäsion und Klebtechnik,* Hanser Verlag, München 1969.
3. National Materials Advisory Board, National Research Council: Treatise of Adhesion und Adhesives ("Structrual Adhesives with Emphasis on Aerospace Applications"), vol. 4, Marcel Dekker, New York-Basel 1976.
4. J. J. Bikerman: *The Science of Adhesive Joints,* 2nd ed., Academic Press, New York-London 1968.
5. C. Heitler in D. J. Alner (ed.): *Aspects of Adhesion,* **vol. 4,** University of London Press, London-Liverpool-Prescot 1968, p. 73.
6. K. W. Allen in D. J. Alner (ed.): *Aspects of Adhesion,* **vol. 5,** University of London Press, London-Liverpool-Prescot 1969, p. 11.
7. N. J. DeLollis in Ch. V. Cagle (ed.): *Handbook of Adhesive Bonding,* McGraw-Hill, New York 1973, pp. 2/1 – 2/16.
8. W. C. Wake, *Polymer* **19** (1978) no. 3, 291.
9. B. V. Deryagin, N. A. Krotova, V. P. Smilga: *Adhesion of Solids,* Consultants Bureau, New York-London 1978.
10. W. A. Zisman, *Ind. Eng. Chem.* **55** (1963) 18.
11. L. H. Shorpe, H. Schonhorn, *Adv. Chem. Ser.* **43** (1964) 189.
12. R. J. Good in R. L. Patrick (ed.): *Treatise on Adhesion and Adhesives,* **vol. 1,** Marcel Dekker, New York 1967, pp. 9 – 65.
13. D. H. Kaelble: *Physical Chemistry of Adhesion,* Wiley-Interscience, New York 1971.
14. R. H. Ottewill in D. J. Alner (ed.): *Aspects of Adhesion,* **vol. 2,** University of London Press, Liverpool-London-Prescot 1966, p. 9.
15. W. A. Zisman in P. Weiss (ed.): *Adhesion and Cohesion,* Elsevier, Amsterdam-London-New York 1962, p. 176.
16. R. Köhler: "Fortschritte der Klebtechnik in den letzten zehn Jahren," *Chem Ing. Tech.* **42** (1970) no. 9/10, 599.
17. A. Korolew, V. J. Bek, N. A. Grishin, *Vysokomol. Soedin.* **4** (1962).
18. A. H. Walter, *Plaste Kautsch.* **17** (1970) no. 8, 601.
19. S. S. Voyutskii: *Autohesion and Adhesion of High Polymers,* Interscience, New York-London-Sydney 1963.
20. R. M. Vasenin: *Adhesion, Fundamentals and Practice,* Ministry of Technology, MacLaren, London 1969.
21. J. R. Huntsberger in R. L. Patrick (ed.): *Treatise on Adhesion and Adhesives,* **vol. 1,** Marcel Dekker, New York 1967, p. 136 – 137.
22. R. Köhler, *Adhäsion* **3** (1972) 66.
23. W. Brockmann: "Grundlagen und Stand der Metallklebtechnik," *VDI-Taschenbücher T 22,* VDI-Verlag, Düsseldorf 1971, pp. 6 – 9.
24. DIN 16920, June 1981.
25. S. Semerdjiev: *Metal-to-Metal Adhesive Bonding,* Business Books, London 1970.
26. J. Shields: *Adhesives Handbook,* 2nd ed., Newnes-Butterworths, London 1976.
27. Inter. Isocyanate Institute Inc.: *Recommendations for Handling Aromatic Isocyanates,* New York 1976.
28. *Adhes. Age* **25** (Aug. 1982) 30 – 31.
29. I. Skeist: *Handbook of Adhesive* 3rd ed., Van Nostrand Reinhold, New York 1985.
30. R. Köhler, *Kunststoffe* **48** (1958) 441.
31. R. Köhler, *Adhäsion* **16** (1972) 41.
32. H. Lucke: *Kunststoffe und ihre Verklebung,* Verlag Brunke Garrels, Hamburg 1967.
33. D. L. Bateman: *Hot Melt Adhesives,* 3rd ed., Noyes Data Corp., New Jersey, USA, 1978.
34. Teroson, DE 24 54 235, 1974 (Kösters).
35. G. F. Abele: *Hochfrequenz Schweißtechnik,* Zechner-Hurlig Verlag, Speyer 1973, pp. 387 – 391.
36. D. Sates (ed.): *Handbook of Pressure-Sensitive Adhesive Technology,* Van Nostrand Reinhold, New York 1982.
37. J. A. Radley (ed.): *Industrial Uses of Starch and its Derivates,* Applied Science Publ., London 1976.
38. Blattman & Co., CH 256 502, 1949 (Meier).
39. Henkel, DE 1 811 466, 1978 (Muszik, Dierichs).
40. R. Houwink, G. Salomon (eds.): *Adhesion and Adhesives,* 2nd ed., **vol. 1,** Chap. 2, Elsevier, Amsterdam-London-New York 1965.

41. H. Reinhard: "Dispersionen synthetischer Hochpolymerer," part II: Anwendungen, *Chemie, Physik und Technologie der Kunststoffe in Einzeldarstellungen,* vol. 14, Springer Verlag, Berlin-Heidelberg 1969.
42. R. L. Patrick (ed.): *Adhesion and Adhesives,* **vol. 2: Materials,** Marcel Dekker, New York 1969.
43. Lord Corp., DE 2 136 337, 1971 (Duston).
44. L. Skeist (ed.): *Handbook of Adhesives,* Van Nostrand Reinhold, New York 1977.
45. Ch. Gloxhuber, J. Galinke, *Chem. Ing. Tech.* **40** (1968) 485 – 486.
46. V. K. Krieble, US 2 895 950, 1959.
47. H. Lee, K. Neville: *Handbook of Epoxy Resins,* Chap. 21, McGraw Hill Book Comp. (1967)
48. L. Jackson, *Adhes. Age* **6,** (1963) 34.
49. R. J. Arnold, *Mod. Plast.* **41** (1964) 149 – 160.
50. B. A Dombrow: *Polyurethanes,* Van Nostrand Reinhold, New York 1957.
51. R. M. Carvey, *Adhes. Age* **24** (1968) 35 – 37.
52. W. M. Ramsay: *Epoxy and Urethane Adhesives,* Palmerton Publ. Co., New York 1971.
53. H. Lee, D. Stoffey, K. Neville: *New Linear Polymers,* McGraw Hill, New York 1967, p. 207.
54. Henkel, DE 204 1126, 1970 (Klement, Budnowski, Lachmann).
55. K. Katz, C. V. Cagle: *Adhesive Materials,* Forster Publ. Co, Long Beach, USA, 1971.
56. W. H. Cobbs, *Adhes. Age* **22** (Dec. 1979) 31 – 36.
57. W. Brockmann: "Grundlagen u. Stand der Metallklebtechnik," *VDI Taschenbücher T 22,* VDI-Verlag, Düsseldorf 1971.
58. *"Testing of Adhesives",* TAPPI Monogr. Ser. 1974, no. 35.
59. American Society for Testing and Materials, Philadelphia, PA, USA.
60. Deutsches Institut f. Normung, Berlin, s.a. DIN-Normblattverzeichnis (annual) Beuth Verlag, Berlin-Köln.
61. British Standards Institution, London WI, GB.
62. International Standards Organisation Geneva 20, Switzerland.
63. R. L. Patrick (ed.): *Treatise on Adhesion and Adhesives,* **vol. 2: Materials,** Marcel Dekker, New York 1969.
64. F. Kollmann: *Furnier, Lagenhölzer und Tischlerplatten,* Springer Verlag, Berlin 1962.
65. F. Kollmann, E. Kuenzi, A. Stamm: *Principles of Wood Science and Technology,* Springer Verlag, Berlin-New York 1975.
66. H. Baumann: *Leime und Kontaktkleber,* Springer Verlag, Berlin 1967.
67. R. Köhler, L. Skark: "Duroplastische Kunststoffe für die Holzverleimung und Holzveredlung," in *Kunststoffhandbuch*, vol. X, Hanser Verlag, München 1968.
68. C. Schuller: "Handwerkliche und industrielle Verleimungstechnik im Vergleich," *Bau Möbelschreiner* **5** (1981) 33 – 36.
69. "Kunststoffrohre für den Bau", *Kunstst. Bau* **16** (1981) no. 2, 87 – 93.
70. *Kleben von thermoplastischen Kunststoffen,* DVS-Verlag, Düsseldorf 1993.
71. *Klebeanleitung PVC Druckrohre, Verklebung mit THF-Klebestoff,* Kunststoffrohrverein e.V., Bonn 1985.
72. R. Nowack, E. Barth: "Kunststoffrohrleitungen im Bereich der Verfahrenstechnik," *3 R international* **20** (1981) no. 6/7, 337 – 353.
73. H. Fühn: "Kunststoffrohre aus PVC-hart in der Gasverteilung," *3 R international* **17** (1978) no. 10, 611 – 615.
74. G. Poschet, I. Zöhren: "Chemikalienbeständigkeit von Klebverbindungen," *Schweißen + Schneiden* **31** (1979) no. 8, 332 – 335.
75. W. Trautvetter: "Über chloriertes PVC," *Kunstst. Plast.* **13** (1966) no. 2, 54 – 58. K. Binder, J. Fabing: Beständigkeit von PVC/PVC und PVC/Aluminiumverklebungen gegen verschiedenartige äußere Einflüsse, sterreichisches Kunststoffinstitut, Wien. Adhesives for Pipes and Tiltings of Unplasticized PVC, ISO 19E, 20E, 21E, International Standard Organisation. P. Benjamin: "Quality and Quality Control of Unplasticized Polyvinylchloride (uPVC) Pressure Pipes," *Plast. Rubber: Mater. Appl.5* (1980) no. 4, 151 – 160.
76. S. Boström: *Kautschuk-Handbuch,* **vol. 4,** Berliner Union, Stuttgart 1961, p. 120.
77. S. Buchan: *Rubber-to-Metal Bonding,* Crosby Lockwood, London 1959.
78. G. Klement, *Gummi Asbest + Kunstst.* **24** (1971) 430 – 444.
79. DeCrease: *Rubber World* **158** (1968) no. 1, 555 – 557.
80. R. N. Özelli, H. Scheer: *Gummi Asbest + Kunstst.* **27** (1974) 612 – 618.
81. C. V. Cagle: *Adhesive Bonding, Techniques and Applications,* McGraw-Hill, New York 1968.
82. A. Matting: *Metallkleben,* Springer Verlag, Berlin - Heidelberg 1969.

83. T. R. J. Schlieckelmann: *Metallkleben - Konstruktion und Fertigung in der Praxis,* Deutscher Verlag f. Schweißtechnik, Düsseldorf 1972.
84. *Annual Book of ASTM Standards 1971,* part 16, Philadelphia, Pennsylvania.
85. Military Specifications U.S. Army, Navy and Air Force Superintendent of Documents Washington 25, D.C.
86. P. Wahl: *Handbuch Tapezieren-Kleben-Spannen,* Deutsche Verlagsanstalt, Stuttgart 1974
87. Bundesausschuß Farb- und Sachwertschutz (ed.): *Technische Richtlinien für Maler- und Lackierarbeiten,* Nr. 16, Frankfurt 1976.
88. P. Fickeisen, R. Füßl, R. Hummerich, I. Neumann, *Adhäsion* **42** (1998) no. 1/2, 24 – 29.
89. D. Urban, E. Wistuba in D. Distler (ed.): *Wäßrige Polymerdispersionen,* Wiley-VCH, Weinheim 1999, pp. 150 – 154.
90. Rohm & Haas, *Adhäsion* **42** (1998) no. 1/2, 25 – 29.
91. *European Adhesives & Sealants,* **6** (1997) 28.
92. G. Blunk, *KIB Kunstst. Bau* **9** (1968) 17 – 20.
93. G. Tauber, *Seifen Öle Fette Wachse* **92** (1966) 899 – 908.
94. M. Michel, *Tech. Mitt.* **65** (1972) 590 – 594.
95. F. Mang, *Adhäsion* **23** (1979) no. 1, 4 – 9.
96. R. Hinterwaldner, *Holzzentralbl.* **107** (1981) 180 – 181.
97. R. Hinterwaldner, *Adhäsion* **21** (1977) no. 1, 13 – 20.
98. R. Petri, *Boden Wand Decke* 1977, no. 11, 65 – 72.
99. S. Schaaf, *Textilveredlung* **9** (1974) 14 – 25.
100. S. Schaaf, H. Luthi, *Textilveredelung* **11** (1976) 348 – 351.
101. D. V. Kamat, *Text. Chem. Color.* **5** (1973) 25 – 27.
102. E. de Jong, *Adhäsion* **19** (1975) 317 – 324.
103. S. Schaaf, *Am. Dyest. Rep.* **61** (1972) 31 – 40. D. E. Peermann, *Adhes. Age* **24** (1981) no. 12, 23 – 27.
104. U. Maag, *Adhäsion* **20** (1976) 173 – 177, 206 – 209.
105. W. Eichler, *Silikattechnik* **17** (1966) 227 – 228
106. W. Moebes, A. Wende, *Plaste Kautsch.* **9** (1962) 232 – 236
107. W. Dümig, F. Geyer, *Adhäsion* **22** (1978) no. 3, 70 – 74; **22** (1978) no. 4, 107 – 118; **22** (1978) no. 5, 145 – 147; **22** (1978) no. 8, 252 – 261; **22** (1978) no. 9, 288 – 291.
108. R. Becker, *Plaste Kautsch.* **14** (1967) 96 – 98.
109. F. Moser: "Bonding Glass," *Handbook of Adhesives,* Reinhold, New York 1962.
110. *Coating* **2** (1969) 182.
111. W. Brockmann, *Adhäsion* **22** (1978) 107 – 108.
112. J. Schütte: *Der Luftschiffbau Schütte Lanz 1909 – 1925,* R. Oldenbourg, München 1926.
113. R. Houwink, G. Salomon: *Adhesion and Adhesives,* Elsevier, Amsterdam-London-New York 1965.
114. E. Golf: "De Havilland Hornet," *Aircraft Prod.* **8** (1946) no. 4, 211.
115. R. J. Schliekelmann: "Operational Experience with Adhesive Bonded Structures," in: Bonded Joints and Preparation for Bonding, *AGARD-Lecture Ser.* **1979,** no. 102, 1 – 1/1 – 30.
116. R. J. Schliekelmann: *Metallkleben-Konstruktion und Fertigung,* DVS-Fachbuchreihe Schweißtechn., vol. 60, Deutscher Verlag f. Schweißtechnik, Düsseldorf 1972.
117. E. Thrall: "Failures in Adhesively Bonded Structures," in: Bonded Joints and Preparation for Bonding, *AGARD-Lecture Ser.* **1979,** no. 102, 5 – 1/5 – 89.
118. J. E. Mc Carty: "Application of Advanced Structures and Materials to Transport Aircraft Fuselages," *15th Structures, Structural Dynamics and Materials Conference,* AIAA, New York 1974.
119. A. Bethune: "Die Beständigkeit geklebter Aluminiumkonstruktionen," *Adhäsion* **20** (1982) no. 12, 347.
120. W. Brockmann, O.-D. Hennemann, H. Kollek: "Surface Properties and Adhesion in Bonding Aluminium Alloys by Adhesives," *Int. J. Adhesion* **2** (1982) no. 1, 33/40.
121. D. Schulz: "Airbus-Seitenruder in Faserverbundwerkstoffen," *K Kunststoffberater* 1982, no. 4, 18/28.
122. W. E. Quist, G. H. Narayanan, A. L. Wingert: "Aluminium-Lithium Alloys for Aircraft Structure, an Overview," 2. International Aluminium-Lithium Conf., Monterey, 12. 4. – 14. 4. 1983.
123. Skeist Laboratories Inc., Livingston New Jersey.
124. Japan Adhesive Industry Assoc., Tokyo.
125. Frost & Sullivan Inc., New York.
126. Stat. Bundesamt Wiesbaden, FRG.
127. Strategic Analysis, Reading, Pennsylvania, USA.
128. Industrial Aids, London, GB.

129. E. H. Schindel-Bidinelli, W. Gutherz: *Konstruktives Kleben,* VCH Verlagsgesellschaft, Weinheim 1998, pp. 1 – 4. R. Houwink, G. Salomon: *Adhesion and Adhesives,* **vol. 1: Adhesives,** Elsevier, Amsterdam – London – New York 1965, pp. 1, 543 – 546.
130. G. Habenicht: *Kleben,* 3rd ed., Springer, Berlin 1997, p. 295.
131. W. Brockmann, *J. Adh.* **37** (1992) 173 – 179.
132. EN 923, January 1998
133. DIN EN 233, Feb. 1997; DIN 234, Feb. 1997; DIN 235, Oct. 1989.
134. G. S. Haviland: *Machinery Adhesives for Locking, Retaining and Sealing,* Marcel Dekker, New York 1986.
135. C. W. Boeder in S. R. Hartshorn (ed.): *Structural Adhesives Chemistry and Technology,* Chapter 5, Plenum, New York 1986, pp. 217 – 247.
136. J. M. Rooney, B. M. Malofsky in I. Skeist (ed.): *Handbook of Adhesives,* 3rd. ed., Chapter 26, Van Nostrand Reinhold, New York 1990, pp. 451 – 462.
137. R. D. Rich in A. Pizzi, K. Mittal (eds.): *Handbook of Adhesive Technology,* Chapter 29, Marcel Dekker, New York 1994, pp. 467 – 479.
138. D. J. Stamper, *Brit. Polymer J.* **15** (1983) 34 – 39.
139. Y. Okamoto, *J. Adhesion* **32** (1990) 227 – 235.
140. Y. Okamoto, *J. Adhesion* **32** (1990) 237 – 244.
141. T. Okamoto, H. Matsuda, *Nippon Setchaku Kyokaishi* **20** (1984) no. 10, 468 – 471.
142. S. J. Hudak, F. J. Boerio, P. J. Clark, Y. Okamoto, *Surface and Interface Analysis* **15** (1990) 167 – 172.
143. General Electric, US 2 628 178, 1950 (R. E. Burnett, B. W. Nordlander).
144. American Sealants, US 2 895 950, 1957 (V. K. Krieble).
145. Loctite, US 3 300 547, 1964, (J. W. Gorman, B. W. Nordlander).
146. Loctite, US 3 425 988, 1965 (J. W. Gorman, A. S. Toback).
147. Loctite, US 4 018 851, 1975, US 4 295 909, 1980, US 4 309 526, 1980 (L. J. Baccei).
148. Loctite, US 3 435 012, 1965 (B. W. Nordlander).
149. Princeton Chemical Research, US 3 428 614, US 3 451 980, 1966 (A. M. Brownstein).
150. Princeton Chemical Research, US 3 595 969, 1969 (T. H. Shepherd, F. E. Gould).
151. Eschem, US 4 209 604, 1976, US 4 431 787, 1981 (G. P. Werber).
152. Three Bond, JP 44 007 541, 1969 (S. Kiyono, R. Ogawa).
153. Three Bond, US 3 890 273, 1974 (T. Saito).
154. Henkel, US 3 642 750, 1967 (B. Wegemund, J. Galinke).
155. Rohm und Haas, US 4 234 711, 1978 (W. D. Emmons, V. J. Moser).
156. Dow Corning, US 4 035 355, 1975 (R. H. Baney, O. W. Marko).
157. Toshiba Silicone, EP 467 160, 1992 (F. Tetsuo, T. Masahiro, E. Isao).
158. Loctite, US 5 605 999, 1995 (H.-K. Chu, R. D. Rich, I. D. Crossan, R. P. Cross).
159. Loctite, US 5 635 546, 1996 (R. D. Rich, E. Maandi, P. M. Gontarz, H.-K. Chu).
160. Toagosei, US 3 925 322, 1974 (K. Azuma et al.).
161. Borden, US 3 795 641, 1971 (W. A. Lees, D. J. Bennet, J. R. Swire, P. Harding).
162. American Sealants, US 3 041 322, 1959 (V. K. Krieble).
163. American Sealants, US 3 046 262, 1960 (V. K. Krieble).
164. Loctite, US 3 218 305, 1963 (V. K. Krieble).
165. Loctite, US 4 180 640, 1977 (D. P. Melody, D. A. Doherty, J. F. O′Grady, R. D. Rich).
166. Loctite, US 4 287 330, US 4 321 330, 1975 (R. D. Rich).
167. Henkel, US 3 984 385, 1975 (W. Gruber, J. Galinke, J. Keil).
168. Henkel, US 3 985 943, 1974 (W. Gruber, J. Galinke, J. Keil).
169. W. R. Grace, US 4 429 063, 1982 (K. Reich).
170. Loctite, US 4 513 127, 1984 (A. F. Jacobine).
171. Loctite, US 4 622 348, 1984 (A. F. Jacobine, D. M. Glaser).
172. American Sealants, US 3 043 820, 1960 (R. H. Krieble).
173. Loctite, US 4 038 475, 1975 (E. Frauenglass, G. P. Werber).
174. Loctite, US 3 625 875, 1968 (E. Frauenglass, W. E. Cass).
175. Loctite, US 3 547 851, 1968 (E. Frauenglass).
176. Borden, US 3 419 512, 1965 (W. A. Lees, J. R. Swire).
177. Loctite, US 3 794 610, 1971 (A. G. Bachman).
178. Borden, US 4 138 449, 1976 (T. R. Baldwin, D. J. Bennett, W. A. Lees).
179. Loctite, US 3 988 299, 1975 (B. M. Malofsky).
180. L. J. Baccei, B. M. Malofsky, *Polym. Sci. Technol.* **29** (1984) 589 – 601.
181. Loctite, US 4 813 714, 1988 (C. B. Fairey, E. Frauenglass, L. W. Vincent).
182. Loctite, U.S. Patent 3 591 438, 1968 (A. S. Toback, J. T. O'Conner).

183. Loctite, US 3 616 040, 1968 (A. S. Toback).
184. Loctite, US 3 625 930, 1967 (A. S. Toback, W. E. Cass).
185. H. Onusseit, *Adhesives & Sealants Agenda Europe,* Campden Publishing, London 1997, p. 78.
186. H. Onusseit, *Bindereport* (1997) no. 6, p. 373 – 381.
187. H. Onusseit, *European Adhesives & Sealant* (1990) no. 3, p. 26 – 32.
188. H. Onusseit, *Binderreport* (1996) no. 9, p. 465 – 473.
189. P. Allbericci: "Aerospace Applications" in A. J. Kinlock (ed.): *Durability of Structural Adhesives,* Chap. 8, Applied Science Publishers, London 1983.
190. E. W. Thrall, R. W. Shannon (eds.): *Adhesive Bonding of Aluminum Alloys,* Marcel Dekker, New York – Basel 1985.
191. W. Brockmann, O.-D. Hennemann, H. Kollek, C. Matz, "Adhesion in Bonded Aluminum Joints for Aircraft Construction", *Int. J. Adhes. Adhes.* **6** (1986) 115 – 143.
192. EN 2243 – 2,
193. EN 2243 – 6,
194. US 3 959 091, 1971 (Y. Moji, J. A. Marceau).
195. C. W. Matz, "Optimization of the Durability of Structural Titanium Adhesive Joints", *Int. J. Adhes. Adhes.* **8** (1988) 17 – 24.
196. Y. Ika: "Tissue Adhesives," in C. C. Chu, J. A. von Fraunhofer, H. P. Greisler (eds.): *Wound Closure Biomaterials and Devices,* CRC Press, Boca Raton, FL 1996, pp. 317 – 346.
197. D. Sierra, R. Slatz (eds.): *Surgical Adhesives and Sealants,* Technomic Publishing Company, Lancaster, PA 1996.
198. G. Lewis, *J. Biomed Mater. Res. (Appl. Biomater.)* **38** (1997) 155 – 182.
199. Loctite Corporation (ed.): *Loctite Worldwide Design Handbook,* 1996/97 ed., 1995.
200. J. H. Bickford: *An Introduction to the Design and Behavior of Bolted Joints,* 2nd ed., Marcel Dekker, New York 1990, p. 462 – 463.
201. C. L. Brett, *Int. J. Adhes. Adhes.* **2** (1982) no. 1, 19 – 24.
202. E. Maandi, R. D. Rich, *Scientific Computing and Automation* (1994) Aug., 23 – 27.

Adipic Acid

Michael Tuttle Musser, E. I. DuPont de Nemours & Co., Sabine River Laboratory, Orange, Texas 77631, United States

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 2
4.1. Nitric Acid Oxidation of Cyclohexanol 2
4.2. Butadiene-Based Routes 4
4.3. Other Routes 5
5. Byproducts 5
6. Quality Specifications 5
7. Storage and Transportation 6
8. Derivatives 6
8.1. Adiponitrile 6
8.2. Other Derivatives 7
9. Uses 7
10. Economic Aspects 8
11. Toxicology and Occupational Health . 8
12. References 9

1. Introduction

Adipic acid, hexanedioic acid, 1,4-butanedicarboxylic acid, $C_6H_{10}O_4$, M_r 146.14, $HOOCCH_2CH_2CH_2CH_2COOH$ [*124-04-9*], is the most commercially important aliphatic dicarboxylic acid. It appears only sparingly in nature but is manufactured worldwide on a large scale. Its primary application is in the production of nylon 66 polyamide, discovered in the early 1930s by W. H. Carothers of DuPont. Manufacture of nylon 66 polyamide fiber has grown to become one of the dominant processes in the synthetic fiber industry. The historical development of adipic acid was reviewed in 1977 [5].

2. Physical Properties [6]

Adipic acid is isolated as colorless, odorless crystals having an acidic taste. It is very soluble in methanol and ethanol, soluble in water and acetone, and very slightly soluble in cyclohexane and benzene. Adipic acid crystallizes as monoclinic prisms from water, ethyl acetate, or acetone/petroleum ether. Some physical properties of adipic acid follow:

mp, °C	152.1
bp, °C	
at 101.3 kPa	337.5
at 13.3 kPa	265
at 2.67 kPa	222
at 0.67 kPa	191
at 0.133 kPa	159.5
Relative density (170 °C)	1.085
Bulk density, kg/m^3	600 – 700
Solubility, g/100 g water	
at 15 °C	1.42
at 40 °C	4.5
at 60 °C	18.2
at 80 °C	73
at 100 °C	290
Dissociation constants	
k_1	4.6×10^{-5}
k_2	3.6×10^{-6}
Specific heat of liquid (200 °C), kJ kg^{-1} K^{-1}	2.719
Specific heat of vapor (300 °C), kJ kg^{-1} K^{-1}	1.680
Heat of fusion, kJ/kg	115
Heat of vaporization, kJ/kg	549
Heat of solution in water, kJ/kg	
10 – 20 °C	– 214
90 – 100 °C	– 241
Melt viscosity, mPa · s	
at 160 °C	4.54
at 193 °C	2.64

Flammability and explosion data are summarized in the following:

Closed cup flash point	196 °C
Cleveland open cup flash point	210 °C
Autoignition temperature	420 °C
Dust cloud ignition temperature	550 °C
Minimum explosive concentration (dust in air)	0.035 kg/m^3
Minimum cloud ignition energy	600 J
Maximum rate of pressure rise	18.6 MPa/s

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_269

3. Chemical Properties

Adipic acid is stable in air under most conditions, but heating of the molten acid above 230 – 250 °C results in some decarboxylation to give cyclopentanone [*120-92-3*], *bp* 131 °C. The reaction is markedly catalyzed by salts of metals, including iron, calcium [7], and barium [8]. The tendency of adipic acid to form a cyclic anhydride by loss of water is much less pronounced compared to glutaric or succinic acids [9].

Adipic acid readily reacts at one or both carboxylic acid groups to form salts, esters, amides, nitriles, etc. (Chap. 8). The acid is quite stable to most oxidizing agents, as evidenced by its production in nitric acid. However, nitric acid will attack adipic acid autocatalytically above 180 °C, producing carbon dioxide, water, and nitrogen oxides.

4. Production

Early commercial processes for manufacturing adipic acid involved a two-step air oxidation of cyclohexane [*110-82-7*]. Oxidation of cyclohexane to cyclohexanol – cyclohexanone at low conversion was followed by a high-conversion process for air oxidation of the mixture to adipic acid. Currently (2000), however, all large-scale production is via nitric acid oxidation of cyclohexanol [*108-93-0*], cyclohexanone [*108-94-1*], or a mixture of the two [ketone – alcohol (KA) oil].

Differences among commercial processes are mainly in the manufacture of the KA oil. The six carbon atoms of the adipic acid backbone usually come from benzene, which is hydrogenated to cyclohexane, or phenol, which is hydrogenated to cyclohexanol. The cyclohexane is then oxidized with air to KA oil. In the past 20 years, there has been a shift to the lower cost cyclohexane-based process [10]. (For KA production, see → Cyclohexanol and Cyclohexanone).

Since the early 1980s, a great deal of research has been carried out on the synthesis of adipic acid from butadiene and carbon monoxide (Section 4.3). However, no commercial plant based on this technology is currently in operation.

4.1. Nitric Acid Oxidation of Cyclohexanol

Reaction Mechanism. The second step of the conventional process, developed by DuPont in the late 1940s, involves the oxidation of cyclohexanol, cyclohexanone, or a mixture of both with nitric acid [11], [12]. Adipic acid is obtained in greater than 90 % yield. Major byproducts are carbon dioxide, nitrogen oxides, and some lower molecular mass dicarboxylic acids. Some byproducts arising from impurities in the starting KA oil are also present.

The chemical mechanism was discussed originally in 1956 [13] and later in greater detail [14], [15]. The latter reports include kinetic and reactor design considerations. Results of related studies, especially on the later stages of the reaction, were published at about the same time [16–18]. A summary of the findings of these investigations is given in Figure 1.

Cyclohexanol (**1**) is oxidized to cyclohexanone (**2**), accompanied by generation of nitrous acid. The cyclohexanone then reacts by one of three possible pathways leading to the formation of adipic acid (**8**). The major fraction of the reaction occurs via nitrosation to produce 2-nitrosocyclohexanone (**3**), then by further reaction with nitric acid to form the 2-nitro-2-nitrosoketone (**6**). Hydrolytic cleavage of this intermediate gives 6-nitro-6-hydroximinohexanoic acid, also known as nitrolic acid (**9**). This breaks down further to give adipic acid and nitrous oxide, the main unrecovered nitric acid reduction products. Typically 2.0 mol of nitric acid is converted to nitrous oxide for each mole of adipic acid produced.

The second pathway occurs at higher temperature, where nitration predominates. At these elevated temperatures, the pathway via the dinitroketone (**4**) becomes significant.

The third path proposed by the early investigators involves the intermediate formation of the 1,2-diketone (**5**) or its dimer. Conversion of this material to adipic acid in good yield requires the use of a vanadium catalyst. The effect of vanadium on the overall yield suggests a significant contribution by this pathway.

The intermediate nitrosoketone (**3**) can undergo two important side reactions. Multiple nitrosation leads to the intermediate (**10**), which loses carbon dioxide to produce glutaric acid

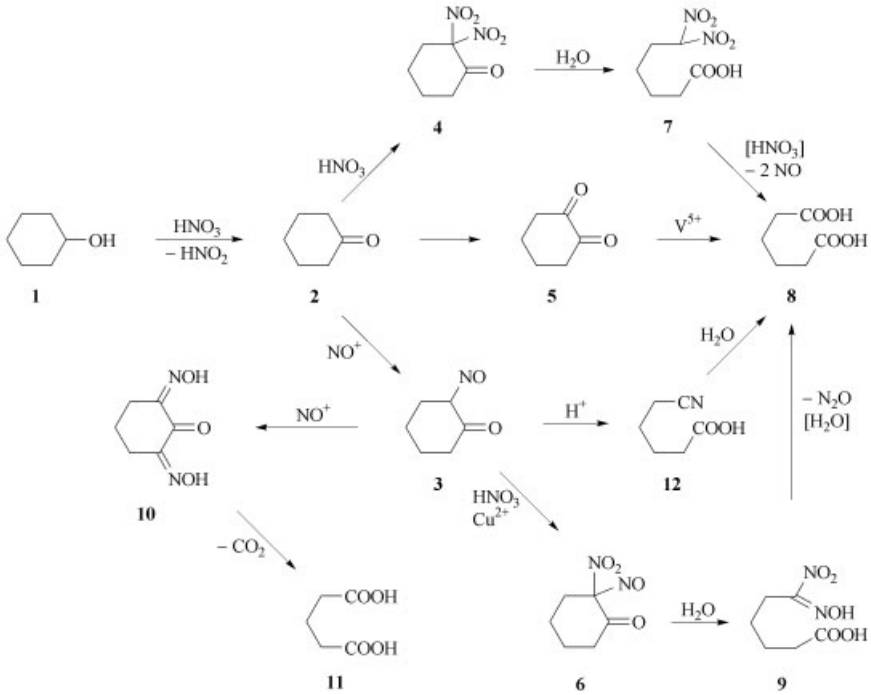

Figure 1. Reaction paths in nitric acid oxidation of cyclohexanol

(**11**) or succinic acid from subsequent reaction with nitric acid. Copper metal is added to the nitric acid to inhibit these reactions. In systems containing a relatively high steady-state concentration of the nitrosoketone (**3**) or the tautomeric oximinoketone, a Beckmann-type rearrangement leads to 5-cyanopentanoic acid (**12**) in minor amounts. This material is slowly hydrolyzed to adipic acid.

Commercial Nitric Acid Oxidation Processes. The basic technology for carrying out the nitric acid oxidation of cyclohexanol – cyclohexanone (KA) remains similar to that described in the early patent literature. Advances have centered on improvement in byproduct removal, catalyst and nitric acid recovery, and suppression of nitrous oxide, a greenhouse gas which was traditionally vented to the atmosphere. Because of the corrosive nature of nitric acid, plants are constructed of stainless steel (type 304L or better), or of titanium in areas of most severe exposure. The block flow diagram in Figure 2 shows a typical layout for a commercial nitric oxidation process [5], [19]. The reaction is carried out in a continuously circulated loop of nitric acid mother liquor (NML) that passes through the entire system, as shown by the bold line.

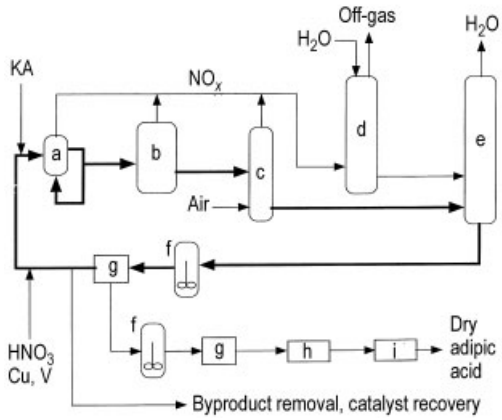

Figure 2. Flow diagram of a process for nitric acid oxidation of cyclohexanone – cyclohexanol
a) Reactor; b) Cleanup reactor; c) NO_x bleacher; d) Nitric acid absorber; e) Concentrator; f) Crystallizer; g) Filter or centrifuge; h) Dryer; i) Cooler

The reactor (a) is essentially a large heat exchanger, controlled at 60 – 80 °C and

0.1 – 0.4 MPa. To this is fed the recycled NML, the KA feed material, the makeup acid containing 50 – 60 % nitric acid and the copper – vanadium catalyst [20], [21]. Residence time in (a) is less than 5 min. In some facilities, the effluent is passed through a second reactor (b) at elevated temperature (110 – 120 °C) [22]. This high-temperature converter (b) can be used to complete the reaction and reduce the amount of impurities which need to be removed through crystallization. The reaction is very exothermic (6280 kJ/kg) and normal heat-exchanger surfaces tend to frost, leading to loss of temperature control. Several different reactor designs have been patented which aid in removing the heat of reaction and minimizing energy usage in the process [23–28]. An excess of recycled NML to KA feed stream of at least 3 : 1 and up to 1000 : 1 is maintained to control the reaction and improve the yield [21].

The product stream is passed through a bleacher (c), in which excess dissolved nitrogen oxides are removed with air and sent to the absorber (d), where they are reabsorbed and recovered as nitric acid. The off-gas from the absorber can be used to initiate the oxidation at lower temperatures by passing it through the KA feed stream before it is fed to the oxidizer [29–31]. Removal of the NO_x from the off-gas by scrubbing with KA has also been described [32]. The water produced in the process is then removed in a concentrating still (e) that is usually operated under vacuum. The concentrated product stream is either recycled to the reactor with diversion of a portion to product recovery or passed to product recovery prior to recycle of the NML filtrate. Crude adipic acid is removed from the NML loop by crystallization (f) followed by subsequent filtration or centrifugation (g) [33–35]. A portion of this effluent stream, which contains high concentrations of glutaric acid, succinic acid, and byproducts, is processed to recover the vanadium and copper catalysts and remove the byproduct acids. Metal recovery is usually accomplished by ion exchange [36]. The crude adipic acid from the first crystallizer (g) is dissolved and recrystallized at least one additional time before proceeding to a dryer (h) and a cooler (i). If the adipic acid is not needed in dry form, the crystals from the centrifuge/filter (g) can be dissolved in water and added to a solution of aqueous 1,6-hexanediamine to make nylon salt.

Other improvements of the conventional process have been described [37], especially in connection with separation and recovery of the dibasic acid byproducts [38–41]. The crude adipic acid is refined to varying degrees, depending upon the end use, but usually is recrystallized from water. Destruction of impurities by refluxing in 60 % nitric acid containing dissolved vanadium has been claimed to produce high-quality product [42].

4.2. Butadiene-Based Routes

In the early 1970s, BASF began an extensive research program on producing dimethyl adipate, a diester that could be hydrolyzed to adipic acid. The process involved carbomethoxylation of butadiene with carbon monoxide and methanol to give methyl 3-pentenoate using a cobalt catalyst and pyridine at high-pressure [43]. The methyl 3-pentenoate was then separated from byproducts by distillation. The second carbomethoxylation to give dimethyl adipate occurs at lower pressure but requires a lower pyridine to cobalt ratio [44]. The hydrolysis of the diester to adipic acid and methanol is a high-yield catalytic process [45]. The overall yield from butadiene appears to be about 70 %. It is believed this process has been demonstrated on a pilot-plant scale, but not yet commercialized.

In the mid-1980s, DuPont also began a major program on a butadiene-based route. In contrast to the BASF diester route, it involved the direct dihydrocarboxylation of butadiene to adipic acid (Fig. 3). The first step [46], which can be catalyzed by palladium, rhodium, or iridium, leads to largely 3-pentenoic acid (**12**). The second step [47], catalyzed by rhodium or iridium gives adipic acid (**13**), 2-methylglutaric acid (**14**), and 2-ethylsuccinic acid (**15**). The advantage of this process was that the 2-methylglutaric and 2-ethylsuccinic acids could be isomerized to adipic acid by the same catalyst system [48]. The catalyst seems to require a halide promoter, such as hydroiodic acid. The solvent for this process is usually a saturated carboxylic acid, such as pentanoic acid, which is a byproduct of the process. Since the late 1980s most major chemical com-

panies have issued numerous patents on variations of these butadiene based routes [49–51].

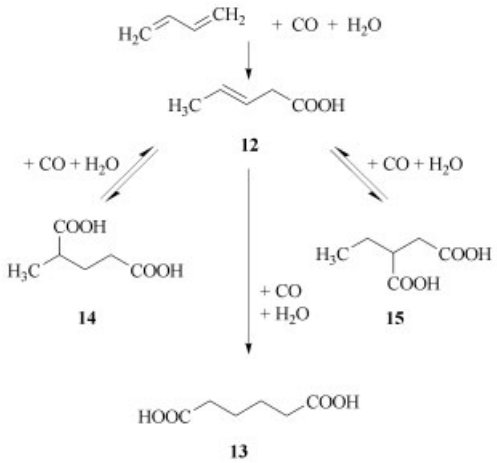

Figure 3. Hydrocarboxylation of butadiene to adipic acid

4.3. Other Routes

In addition to the commercial two-step air/nitric acid oxidation of cyclohexane and the carboxylation/carbomethoxylation of butadiene, several other processes have been investigated. Research at Monsanto on palladium halide catalyzed dicarbonylation of 1,4-disubstituted 2-butenes was reported in early 1984 [52]. This process produces adipic acid from 1,4-dimethoxy-2-butene, carbon monoxide, and palladium chloride at 100 °C after the resulting unsaturated dimethyl ester has been hydrogenated and hydrolyzed.

The one-step oxidation of cyclohexane with nitric acid [53], [54], nitrogen dioxide [55], or air has been described. The one-step all-air oxidation of cyclohexane is economically very attractive and has been heavily researched. Early work was performed by Gulf Research and Development [56–58], Asahi Chemical Industries [59], [60], and others [61]. For example, cyclohexane is oxidized in one step to adipic acid in 70 – 75 % yield, in the presence of a cobalt acetate catalyst in acetic acid as solvent [59]. More recently, there has been renewed interest in this work, and several patents have been issued to Redox Corporation and Bayer. [62–65].

Adipic acid can be produced by ozonolysis of cyclohexene [66] or by addition of a carboxylic acid to cyclohexene and subsequent oxidation of the resulting ester with nitric acid [67]. Formation of adipic acid derivatives by electrolytic coupling of acrylates has also been described [68].

5. Byproducts

The major byproducts of nitric acid oxidation of KA are glutaric acid [*110-94-1*] and succinic acid [*110-15-6*], and minor amounts of pentanoic acid and hexanoic acid are also formed. In commercial operations, the nitric acid reaction medium (NML) contains high concentrations of glutaric and succinic acids, resulting from the recycling of the mother liquor after crystallization of the adipic acid. A portion of this stream is diverted and processed separately to remove the byproduct acids and recover nitric acid and the copper and vanadium catalysts. Early commercial processes discarded these byproduct acids. However, most companies now recover these acids either as a mixture of dibasic acids (DBA) or convert them to dibasic esters (DBE) for a variety of uses.

Following the removal of the copper and vanadium by ion exchange and distillation of the nitric acid in water, methanol can be added to convert the acids to their methyl esters. Then the esters are distilled to give a mixture or the individual esters [69–71].

Sometimes the acids are removed by distillation to produce a mixture of acids and anhydrides, especially glutaric anhydride [*108-55-4*] and succinic acid [*108-30-5*] [72–76]. Separation of the individual acids by crystallization and extraction with organic solvents has been described [77], [78]. Other means of separating the byproduct acids include addition of inorganic salts [79], a $C_1 - C_6$ primary alkylamine [80], or urea [81], and extraction by a ketone solvent [82].

6. Quality Specifications

Commercial adipic acid is one of the purest large-scale manufactured chemicals because of the stringent requirements of its major consumer, the synthetic fibers industry. The U.S.

FDA has approved adipic acid as a food additive. Because essentially all adipic acid manufacturers use a nitric acid oxidation process, impurities are similar. Purity is affected mostly by variations in the synthesis of the KA intermediate and in the extent of adipic acid recrystallization and purification. Some typical specifications for food-grade adipic acid are: color, APHA equivalence (Hazen) 10 max., water 0.2 % max., ash on ignition 10 ppm max., iron 1.0 ppm max., adipic acid content 99.6 % min. [83].

Procedures for analysis of food-grade adipic acid are described in [84]. General methods for water (Karl Fischer), color in methanol solutions (APHA), iron, and other metallic impurities in commercial acid have been summarized [85]. Resin-grade adipic acid frequently has limits for succinic (ca. 50 ppm) and caproic (ca. 30 ppm) acids, and for hydrocarbon oils (ca. 15 ppm). Carboxylic acids can be determined by gas chromatography of their esters or by liquid chromatography of the free acids [86]. Total nitrogen can be determined by chemical reduction and distillation of ammonia from an alkaline solution. Hydrocarbon oil may be determined by IR analysis of a halocarbon extract of a solution of the salt.

7. Storage and Transportation

Adipic acid is conveyed pneumatically or mechanically from the drying equipment to the storage or shipping container. These containers may be aluminum or stainless steel railroad hopper cars, trucks, plastic bags, or drums. Principal hazards in handling adipic acid are the danger of dust explosion and skin or mucus membrane irritation on exposure to the dust. Particle size control and flow characteristics are also important factors due to the tendency of adipic acid that contains excessive fines to cake during storage.

8. Derivatives

8.1. Adiponitrile

The most important derivative of adipic acid is adiponitrile, 1,6-hexanedinitrile, 1,4-dicyanobutane, [*111-69-3*], M_r 108.14, *bp* 298 – 300 °C (at 101.3 kPa), 154 °C (at 1.3 kPa), *fp* 2.4 °C, n_D^{25} 1.4370, d_4^{25} 0.9599, an intermediate in the manufacture of the other major nylon 66 component, 1,6-hexanediamine. The original production process involved conversion of the acid to the dinitrile by liquid- [87] or vapor-phase dehydration [88] of the ammonium salt in the presence of phosphoric acid or a boron – phosphorus catalyst. Although this was the predominate technology used for adiponitrile production in the past, it is no longer used by any major nylon 66 producers.

Other routes which have been used include a process by Celanese, which in the 1960s and 1970s avoided manufacturing adiponitrile by producing 1,6-hexanediamine from ammonolysis of 1,6-hexanediol, which in turn was made by the hydrogenation of adipic acid [89]. This adipic acid based route was shut down around 1980. In 1948, DuPont introduced, and for several years operated, a process based on furfural [90]. From 1951 to 1983, DuPont operated a butadiene chlorination process [91]. The intermediate 1,4-dichloro-2-butene was converted to 3-hexenedinitrile with sodium cyanide and then hydrogenated to adiponitrile. Current adiponitrile manufacture is based on either propylene or direct hydrocyanation of butadiene.

In 1965 Monsanto introduced a process involving the electrolytic coupling of acrylonitrile [92]. This process, or variations of it, is also used in the United Kingdom and Japan. DuPont began the direct hydrocyanation of butadiene in 1972 [93]. Now all DuPont adiponitrile production, including a joint venture with Rhône-Poulenc (now Rhodia) in France, uses this technology. The process consists of a two step hydrocyanation, catalyzed by a nickel(0) phosphite complex and promoted by certain Lewis acids [94–96]. The mixture of isomeric pentenenitriles and methylbutenenitriles produced in the first step is isomerized to predominately 3- and 4-pentenenitrile [97–99]. Subsequent anti-Markovnikov addition of hydrogen cyanide to the pentenenitriles produces adiponitrile.

Other routes that have been revealed include chemical dimerization of acrylonitrile to 3-hexenedinitrile [100–102] and hydrocyanation of butadiene with a copper halide catalyst to yield 3-pentenenitrile [103], followed by disproportionation to dicyanobutenes and

butenes. Finally, another dimerization route to adiponitrile involves the addition of acrylonitrile to 2-methyleneglutaronitrile in the presence of zinc or cobalt complexes and a Lewis base [104]. The dimer is then hydrocyanated to 1,2,4-butanetricarbonitrile followed by dehydrocyanation to 3-hexenedinitrile [105].

8.2. Other Derivatives

Salts. Adipic acid forms alkali metal and ammonium salts that are water-soluble and alkaline earth salts that are only moderately soluble. Their solubilities in 100 g of water are: diammonium salt [*3385-41-9*] 40 g (14 °C), disodium salt [*7486-38-6*] 59 g of hemihydrate (14 °C), dipotassium salt [*19147-16-1*] 65 g (15 °C), calcium salt [*22322-28-7*] 4 g of monohydrate [*18850-78-7*] (13 °C), 1 g of anhydrous salt (100 °C).

The most common salt is poly(1,6-hexanediammonium hexanedioate), produced by interaction of adipic acid with 1,6-hexanediamine. This water-soluble salt, the precursor to nylon 66, is readily shipped or stored prior to the final polyamidation, which occurs with the removal of water. The chemistry of this step has been reviewed [106].

Esters and Polyesters. The esters and polyesters of adipic acid constitute the largest non-polyamide market for adipic acid. Esters made from long-chain alcohols are used as plasticizers and lubricants, while those from short-chain alcohols are used primarily as solvents. Refluxing adipic acid with methanol in the presence of an acid catalyst can produce monomethyl adipate, along with the diester. Electrolysis of the salt of the monoester (Kolbe synthesis) produces dimethyl sebacate, another polyamide precursor. The boiling points of some esters are listed in Table 1. The esters dissolve readily in most organic solvents. While dimethyl adipate is the most commonly used solvent, di-2-ethylhexyl adipate is the most widely used plasticizer. Other simple adipate plasticizers include the *n*-octyl, *n*-decyl, isodecyl, and isooctyl esters. More complex polymeric plasticizers, prepared from glycols, account for a little less than half the adipic acid based plasticizers. Low molecular mass polyester polyols having hydroxyl end groups are used with polyisocyanates to produce polyurethane resins.

Table 1. Boiling points of adipic acid esters

Ester		*p*, kPa	*bp*, °C
Monomethyl	[*627-91-8*]	1.3	158
Dimethyl	[*627-93-0*]	1.7	115
Monoethyl	[*626-86-8*]	0.9	160
Diethyl	[*141-28-6*]	1.7	127
Di-*n*-propyl	[*106-19-4*]	1.5	151
Di-*n*-butyl	[*105-99-7*]	1.3	165
Di-2-ethylhexyl	[*103-23-1*]	0.67	214
Di-*n*-nonyl	[*151-32-6*]	0.67	230
Di-*n*-decyl	[*105-97-5*]	0.67	244

Anhydrides. The usual form of the anhydride produced by dehydrating adipic acid is the linear, polymeric form [*2035-75-8*]. Distillation of the polymeric anhydride is said to produce the monomeric cyclic form, which is very unstable and reverts readily to the linear, polymeric anhydride.

Amide. The diamide, $C_6H_{12}N_2O_2$ [*628-94-4*], *mp* 228 °C, is practically insoluble in cold water. It has been traditionally prepared from the dimethyl ester by treatment with concentrated ammonium hydroxide or by heating the diammonium salt of adipic acid in a stream of ammonia. Other substituted amides can be prepared from amines by the usual synthetic methods.

9. Uses

About 80 % of worldwide adipic acid consumption is used for the manufacture of nylon 66 fibers and resins. This is down from about 87 % in 1981. Table 2 summarizes consumption in three major regions of the world. A small amount of adipic acid is still used captively to produce adiponitrile.

Large amounts are converted to esters for use in plasticizers, lubricants and in a variety of polyurethane resins. The monomeric esters are important plasticizers for poly(vinyl chloride) and other resins, while polymeric esters are used when unusually high plasticizer levels are required. Polyurethane resins employing adipic acid are produced from polyisocyanates and polyester polyols (adipates). These are used

in specialty foams, lacquers, adhesives, surface coatings and spandex fibers for stretch-wear.

Table 2. Adipic acid consumption, 10^3 t/a [107–110]

	United States		Western Europe		Japan	
	1991	1995	1991	1995	1991	1995
Nylon 66 fiber	611	629	267	225	35	34
Nylon 66 resin	115	193	86	110	23	26
Plasticizers	30	38	21	24	13	14
Polyurethane resins	38	39	28	45	18	13
Miscellaneous	17	23	66	88	19	17

Adipic acid is added to gelatins and jams as an acidulant and to other foods as a buffering or neutralizing agent. It is also used to modify the properties of unsaturated polyesters for use in reinforced plastics and alkyd coatings. Polyamide – epichlorohydrin resins employing adipic acid are used to increase the wet strength of paper products. Other miscellaneous applications are in the adhesives, insecticide, tanning and dying, and textile industries. Adipic acid and mixed dibasic acids (DBAs) are being used as buffers in flue gas desulfurization treatment in power plants [111].

10. Economic Aspects

Capacities. Total worldwide annual capacity for adipic acid was 2.5×10^6 t/a in 1999 (Table 3). Although this reflects only a 15 % growth in capacity since 1980, it also reflects the shutting down of several adiponitrile plants which had used adipic acid as a starting material, thus making it available for other uses. The North American capacity was 1.06×10^6 t/a, or 42 % of the total, whereas Western Europe accounted for 35 %, produced mainly by the United Kingdom, France, Germany and Italy. Imports and exports have become significant. In 1995 U.S. exports were 71×10^3 t [112] or 8.5 % of U.S. consumption. This is up from 1.2 % in 1979 [113]. Since 1970, U.S. consumption of adipic acid has increased by 80 %, or about 3 % per year. Growth rates are expected to remain at about 3 % per year through 2000. Regional capacities are shown in Table 3, along with annual capacities for the major producing companies.

Table 3. Worldwide adipic acid capacity as of January 1999 [113]

Region	Capacity, 10^3 t/a	Major producers (capacity, 10^3 t/a)
North America	1058	DuPont (740), Solutia (295), Allied (23)
Western Europe	841	Rhodia (280), BASF (260), DuPont (220), Bayer (55), Radici (60), UCB
Far East	70	Asahi (120), DuPont (115), China (139), Korea (70)
Others	115	

Production. Adipic acid production is dominated by nylon 66 fiber and resin manufacture; as a result, the economic picture for the acid is strongly dominated by the markets for these materials. Less than 15 % of U.S. production is sold on the merchant market, essentially for non-nylon uses. This ratio is higher in Western Europe and Japan. The synthesis of adiponitrile from adipic acid, once significant, is no longer used by any major manufacturers. However, it continues to a very limited degree in some Eastern European countries. The non-nylon uses for adipic acid have grown at about 6 % per year since 1970. Production costs closely parallel raw material prices (cyclohexane and ammonia), which in the late 1990s have fallen with the decline in crude oil prices. The largest growth rate for adipic acid, as well as nylon 66, is in China and the Far East. The projected growth rate in the United States and Western Europe is expected to be slow, so the supply/demand picture should remain relatively constant for the next few years.

11. Toxicology and Occupational Health

Adipic acid is a minor irritant of low oral toxicity. The lowest published lethal dose (LDLo) is 3600 mg/kg (rat, oral), LD_{50} 275 mg/kg (rat or mouse, i.p.), LD_{50} 1900 mg/kg (mouse, oral) [114]. Some delayed body weight increases and changes in certain enzymes and in urea and chloride level in the blood were observed in chronic feeding tests [115]. No teratogenic activity was detected in studies with pregnant mice [116]. In metabolism studies with rats fed ^{14}C-labeled

adipic acid, both unchanged adipic acid and normal metabolic products were detected in the urine [117], [118].

Exposure of the mucous membranes (eyes, respiratory tract) produces irritation; prolonged exposure to the skin can be drying or irritating. In case of spills or leaks, personnel should be protected from inhalation or excessive skin contact. Dusting should be controlled and static sparks should be avoided. Water may be used to flush the area.

Although no TLV or MAK has been established, the airborne exposure should be less than that of an organic nuisance dust: ACGIH (1979) 8-h TWA 10 mg/m^3 (total dust) and 8-h TWA 5 mg/ m^3 (respirable dust) (OSHA TLV is 15 mg/ m^3 for total dust). Toxicity data from representative types of adipic acid derivatives are shown in Table 4.

Table 4. Toxicity data for adipic acid derivatives [119]

Derivative	Oral LD_{50} (rat) mg/kg	Inhalation LC_{50} (rat, 4 h), mg/m^3	Other LD_{50}, mg/kg
Adiponitrile	300	1710	50 (guinea pig, s.c.)
Di-2-ethylhexyl adipate	9110	–	900 (rat, i.v.)
Dimethyl adipate	–	–	1809 (rat, i.p.)
Adipamide	500		
Magnesium adipate	–	–	180 (mouse, i.v.)

12. References

General References

1. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5000A – 608.5001P.
2. M. Sittig: *Dibasic Acids and Anhydrides,* Noyes Development Corp, Park Ridge, NJ, 1966, p. 35 – 50.
3. *Kirk-Othmer,* 4th ed., **1,** 466 – 493.
4. *Ullmann,* 5th ed., **A1,** 269 – 278.

Specific References

5. V. Luedeke in J. McKetta, W. Cunningham (eds.): *Encyclopedia of Chemical Processing and Design,* **vol. 2,** Marcel Dekker Inc., New York 1977, p. 128 – 146.
6. *Ullmann* 5th ed. **A1,** 259.
7. W. Hentzchel, J. Wislicenus, *Liebigs Ann. Chem.* **275** (1983) 312.
8. G. Vavon, A. Apchie, *Bull. Soc. Chim. Fr.* **43** (1928) 667.
9. J. W. Hill, *J. Am. Chem. Soc.* **52** (1930) 4110.
10. W. F Stahl: "Organic Chemicals A-A1," in *Chemical Economics Handbook* , SRI International, Menlo Park, Calif., 1996, 608.5000G
11. DuPont, US 2 557 282, 1951 (C. Hamblett, A. MacAlevy).
12. DuPont, US 2 703 331, 1953 (M. Goldbeck, F. Johnson).
13. H. Godt, J. Quinn, *J. Am. Chem. Soc.* **78** (1956) 1461 – 1464.
14. D. van Asselt, W. van Krevelen, *Recl. Trav. Chim. Pays Bas* **82** (1963) 51 – 56, 429 – 437, 438 – 449.
15. D. van Asselt, W. van Krevelen, *Chem. Eng. Sci.* **18** (1963) 471 – 483.
16. I. Y. Lubyanitskii, R. Minati, M. Furman, *Russ. J. Phys. Chem. (Engl. Transl.)* **32** (1962) 294 – 297.
17. I. Y. Lubyanitskii, *Zh. Obshch. Khim.* **36** (1962) 3431
18. I. Y. Lubyanitskii, *Zh. Prikl. Khim. (Leningrad)* **36** (1963) 819 – 823.
19. *Ullmann,* 5th ed., **A1,** 271.
20. BASF, US 3 564 051, 1971 (E. Haarer, G. Wenner).
21. BASF, GB 1 092 603, 1969 (G. Riegelbauer, A. Wegerich, A. Kuerzinger, E. Haarer).
22. DuPont, US 3 359 308, 1967 (O. Sampson)
23. ICI, US 3 754 024, 1973 (F. Foster, P. Hay).
24. ICI, US 3 950 410, 1976 (J. Lopez-Merono).
25. ICI, US 3 997 601, 1976 (P. Langley).
26. ICI, GB 1 366 082, 1974 (J. Lopez-Merono).
27. ICI, DE 2 435 387, 1973 (P. Langley).
28. F. Hearfield, *Chem. Eng. (London)* 1980 no. 361, 625 – 627.
29. SU 433 784, 1971 (Y. A. Lubyanitskii).
30. El Paso Products Co., US 3 673 245, 1972 (S. Mims).
31. Chem. Werke Hüls, US 3 761 517, 1973 (H. Rohl, W. Eversmann, P. Hegenberg, G. Hellemanns).
32. ICI, GB 1 510 397, 1976 (F. Foster, N. Hutchinson, D. Potter).
33. DuPont, US 2 713 067, 1955 (C. Hamblett).
34. Vickers-Zimmer, US 3 476 804, 1969 (F. Bende, H. Vollinger, K. Pohl).
35. Vickers-Zimmer, US 3 476 805, 1969 (H. Vollinger, K. Pohl, F. Bende).
36. Monsanto, US 3 186 952, 1965 (D. Brubaker, D. Danly).

37. Celanese, US 3 965 164, 1976 (J. Blay).
38. Celanese, US 3 983 208, 1976 (J. Blay).
39. ICI, GB 1 470 169, 1977 (B. Darlow, R. Chase, J. Peters).
40. ICI, GB 1 480 480, 1977 (A. Bowman).
41. Asahi, US 3 673 068, 1972 (M. Seko, A. Yomiyama, T. Miyake, H. Iwashita).
42. BASF, DE 2 624 472, 1977.
43. BASF, US 4 550 195, 1972 (H. W. Schneider, K. Schwirten, R. Kummer, R. Platz).
44. BASF, US 4 310 686, 1982 (H. W. Schneider, R. Kummer, J. Taglieber, F. J. Weiss).
45. BASF, US 4 310 686, 1982 (W. Gerbert, P. Magnussen, V. Schumacher, W. Praetorius, H. Reitz).
46. DuPont, US 5 145 995, 1991 (P. M. Burke).
47. DuPont, US 5 218 144, 1993 (E. M. Atadan).
48. DuPont, US 5 292 944, 1994 (E. M. Atadan, H. S. Bruner).
49. DSM, WO 9 518 089, 1995 (J. G. DeVries, C. B. Hansen).
50. Monsanto, US 4 433 166, 1984 (F. W. Paulik, T. C. Singleton, W. H. Urry).
51. Rhône-Poulenc, US 5 312 974, 1994 (P. Denis, F. Metz, R. Perron).
52. *Chem. Eng. News* **62** (1984) 28 – 29.
53. DuPont, US 3 306 932, 1967 (D. D. Davis).
54. Monsanto, US 3 654 355, 1972 (W. Mueller, C. Campbell, J. Hicks).
55. Kogai Boshi Chosa, JP-Kokai 105 416, 1978 (W. Ando, I. Nakaoka).
56. Gulf R & D, US 3 231 608, 1966.
57. Gulf R & D, US 4 032 569, 1977 (A. Onopchenko, J. Schulz).
58. Gulf R & D, US 4 263 453, 1981 (J. Schulz, A. Onopchenko).
59. K. Tanaka, *Chem. Technol.* **4** (1974), no. 9, 555.
60. Asahi, JP-Kokai 100 022, 1974 (K. Tanaka, S. Handa).
61. Honshu Chem., JP-Kokai 33 891, 1979 (T. Abe, H. Kasamatsu, Y. Ayabe, Y. Isoda).
62. Bayer & Redox, US 5 321 157, 1994 (J. Kollar).
63. Bayer & Redox, US 5 463 119, 1995 (J. Kollar).
64. Bayer, DE 4 427 474, 1996 (C. Casser, M. Hampe, K. Heise, E. Kysela, K. Neumann, H. Ohst, M. Stenger).
65. Bayer, DE 4 428 977, 1996 (C. Casser, K. Heise, K. Neumann, H. Ohst, K. Paetz).
66. Dainippon Ink. & Chem., JP-Kokai 32 245, 1982.
67. Toray Ind., GB 1 402 480, 1975.
68. Monsanto, GB 1 447 772, 1976 (C. Campbell, D. Danly, W. Mueller).
69. El Paso Products Co., US 4 316 775, 1982 (W. Nash).
70. El Paso Products Co., DE 3 043 051, 1982 (N. Cywinski).
71. DuPont, US 3 991 100, 1976 (S. Hochberg).
72. ICI, US 4 191 616, 1980 (B. Baker).
73. Allied Chem., FR 1 347 525, 1963 (J. Benfield, R. Belden).
74. ICI, US 3 511 757, 1970 (W. Costain, B. Terry).
75. BASF, US 3 564 051, 1971 (E. Haarer, G. Wenner).
76. DuPont, CA 707 340, 1965.
77. DuPont, US 3 338 959, 1967 (C. Sciance, L. Scott).
78. Monsanto, US 3 329 712, 1967 (D. Danly, G. Whitesell).
79. Asahi, JP-Kokai 115 314, 1979 (J. Nishikido, A. Tomura, Y. Fukuoka).
80. BASF, DE 3 002 256, 1981 (W. Rebofka, G. Heilen, W. Klink).
81. Asahi, US 4 146 730, 1979 (J. Nishikido).
82. Veba-Chemie, DE 2 309 423, 1974 (H. Heumann, W. Hilt, H. Liebing, M. Schweppe).
83. E. I. DuPont de Nemours Co., *Adipic Acid Product Bulletin E-18722,* Wilmington, Del., 1983, p. 1.
84. National Academy of Sciences: *Food Chemicals Codex,* 2nd ed., National Academy Press, Washington, D.C., 1972, p. 21 – 22.
85. R. Keller in F. Snell, C. Hilton (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 4,** Wiley-Interscience, New York 1967, p. 408 – 423.
86. R. Schwarzenbach, *J. Chromatogr.* **251** (1982) 339 – 358.
87. Rhodiatoce, US 3 299 116, 1967 (R. Romani, M. Ferri).
88. DuPont, US 2 200 734, 1940. Monsanto, US 3 574 700, 1968 (R. Somich).
89. Celanese, FR 1 509 288, 1968 (P. Volpe, W. Humphrey).
90. J. Hardy in H. Simonds, J. Church (eds.): *The Encyclopedia of Basic Materials for Plastics,* Reinhold Publ. Co., New York 1967, p. 293.
91. DuPont, US 2 680 761, 1952 (R. Halliwell). DuPont, US 2 518 608, 1947 (M. Farlow).
92. M. M. Baizer, D. E. Danly, *Chem. Technol.* **10** (1980) no. 10, 161 – 164, 302 – 311.
93. *Eur. Chem News* **23** (1973) no. 2, 17.
94. DuPont, US 3 496 217, 1970 (W. Drinkard, R. Kassal).
95. DuPont, US 3 496 218, 1970 (W. Drinkard).
96. DuPont, US 3 766 237, 1973 (W. Drinkard).

97. DuPont, US 3 526 654, 1970 (G. Hildebrand).
98. DuPont, US 3 536 748, 1970 (W. Drinkard, R. Lindsey).
99. DuPont, US 3 542 847, 1970 (W. Drinkard, R. Lindsey).
100. ICI, US 4 138 428, 1979 (J. Jennings, P. Hogan, L. Kelly).
101. ICI, US 4 316 857, 1982 (A. Gilbert).
102. ICI, US 4 059 542, 1977 (J. Jennings, L. Kelly).
103. ICI, US 4 210 558, 1980 (G. Crooks). Halcon, US 3 954 831, 1976 (O. Onsager).
104. Halcon, US 3 795 694, 1974 (O. Onsager).
105. M. I. Kohan: *Nylon Plastics,* J. Wiley & Sons, New York 1973, p. 14 – 82.
106. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5000N.
107. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5001D.
108. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5001G.
109. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5001L.
110. *Chem. Eng.* **87** (1980) no. 3, 60.
111. W. F. Stahl: "Organic Chemicals A-AI," in *Chemical Economics Handbook* , SRI International, Menlo Park, CA, 1996, 608.5001B.
112. U.S. Dept. of Commerce, Bureau of Census: *U.S. Exports, Schedule B by Commodity,* U.S. Government Printing Office, Washington, D.C., 1980, EM 546.
113. *Chem. Week,* 1999, March 10, 31.
114. J. M. Nielsen (ed.): *Material Safety Data Sheets,* **vol. 1,** General Electric Co., Schenectady, N.Y., 1979. no. 400.
115. M. Krapotkina, *Gig. Tr. Prof. Zabol.* **5** (1981) 46 – 47.
116. Food and Drug Research Labs Inc. (ed.): "Teratogenic Evaluation of FDA-71-50 (Adipic Acid)," *NTIS No. PB221802,* East Orange, N.J., 1972, p. 1.
117. J. Rusoff, *Toxicol. Appl. Pharmacol.* **2** (1960) 316 – 330.
118. D. Guest, G. Katz, B. Astill in G. Clayton, F. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2C,** Wiley-Interscience, New York 1982, p. 4945.
119. R. Tatken, R. Lewis, Sr. (eds.): *Registry of Toxic Effects of Chemical Substances,* **vol. 1,** 1981 – 1982 ed., U.S. Department of Health and Human Resources, U.S. Government Printing Office, Washington, D.C., 1983, p. 286 – 287.

Adiponitrile → Adipic Acid

Alcohols, Aliphatic

Individual keywords: → Alcohols, Polyhydric; → Butanols; → Cyclododecanol, Cyclododecanone, and Laurolactam; → Cyclohexanol and Cyclohexanone; → Ethanol; → 2-Ethylhexanol; → Fatty Alcohols; → Methanol; → Pentanols; → Propanols; for Allyl Alcohol → Allyl Compounds

Jürgen Falbe, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 2 – 4)

Helmut Bahrmann, Ruhrchemie AG, Oberhausen, Federal Republic of Germany (Chap. 2 – 4)

Wolfgang Lipps, Naturwissenschaftlich-technische Akademie Prof. Dr. Grübler, Isny, Federal Republic of Germany (Chap. 1 – 4)

Dieter Mayer, Centre International de Toxicologie, Évreux, France (Chap. 5)

1. Introduction 1
2. Saturated Alcohols 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.3. Production 5
2.3.1. Synthesis from Carbon Monoxide and Hydrogen 5
2.3.2. Oxo Synthesis 5
2.3.3. Hydrogenation of Aldehydes, Carboxylic Acids, and Esters 6
2.3.4. Aldol Condensation of Lower Aldehydes and Hydrogenation of the Alkenals 6
2.3.5. Oxidation of Trialkylaluminum Compounds 7
2.3.6. Oxidation of Saturated Hydrocarbons 7
2.3.7. Hydration of Olefins 8
2.3.8. Homologation of Alcohols 9
2.3.9. Reppe Process 9
2.3.10. Hydrocarboxymethylation 9
2.3.11. Fermentation 9
2.3.12. Guerbet Alcohols 10
2.3.13. Other Processes 10
2.4. Individual Alcohols 10
2.4.1. C_6 Alcohols 10
2.4.2. C_7 Alcohols 12
2.4.3. C_8 Alcohols 13
2.4.4. C_9 Alcohols 15
2.4.5. C_{10} Alcohols 15
2.4.6. Mixtures of Linear C_{12} – C_{18} Alcohols (Detergent Alcohols) 16
2.4.7. C_{13} – C_{18} Isoalcohols 16
2.5. Economic Aspects 17
2.6. Quality Specifications 20
2.7. Storage and Transportation 20
3. Unsaturated Alcohols 21
4. Alkoxides 22
4.1. Properties 22
4.2. Preparation 23
4.3. Uses 24
5. Toxicology 24
6. References 25

1. Introduction

Industrially, the most important alcohols are methanol, ethanol, 1-propanol, 1-butanol, 2-methyl-1-propanol (isobutyl alcohol), the plasticizer alcohols (C_6 – C_{11}), and the fatty alcohols (C_{12} – C_{18}), used for detergents. They are prepared mainly from synthesis gas alone (methanol), from olefins via the oxo synthesis, or by the Ziegler process.

Apart from the applications mentioned above, alcohols are used as solvents and diluents for paints (mainly C_1 – C_6 alcohols) [4], as intermediates in the manufacture of esters and a whole range of organic compounds, as flotation agents, as lubricants, and as fuel or fuel additives, e.g., methanol, ethanol, *tert*-butyl alcohol.

For industrial purposes, isomeric mixtures often are preferred because the pure alcohols are too expensive. Moreover, mixtures of alcohols with differing numbers of carbon atoms can be advantageous for certain purposes. Therefore, the amounts of alcohol mixtures available on the market are similar to the quantities of the pure, individual alcohols.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_279

2. Saturated Alcohols

2.1. Physical Properties

Some physical properties of saturated $C_1 - C_{26}$ alcohols are given in Table 1. Up to C_{10}, the straight-chain alcohols are colorless liquids at room temperature with characteristic odors. The higher alcohols are solid, waxy substances. Because of hydrogen bonding, the boiling points are considerably higher for alcohols, especially lower-molecular-mass ones, than for the corresponding hydrocarbons. With increasing molecular mass, however, the influence of the hydroxyl group becomes less. For example, the degree of association at room temperature is 3.17 for methanol but only 1.12 for decanol.

Methanol, ethanol, and the propanols are completely miscible with water. With increasing molecular mass, the solubility decreases significantly. For example, the solubility of 1-butanol in water is 7.7 %, and that of decanol less than 0.01 % [5]. The data for some azeotropic mixtures with water are compiled in Table 2.

Table 2. Azeotropic mixtures with water [9]

Alcohol	*bp* of alcohol, °C	*bp* of azeotrope, °C	Water in azeotrope, wt %
Ethanol	78.32	78.174	4.0
1-Propanol	97.2	87.7	28.3
2-Propanol	82.4	80.3	12.6
1-Butanol	117.7	92.3	37.0
Isobutyl alcohol	107.9	89.9	33.2
sec-Butyl alcohol	99.4	87.5	27.3
tert-Butyl alcohol	82.5	79.9	11.76
1-Pentanol	138.0	95.8	54.4
2-Pentanol	119.3	91.7	36.5
3-Pentanol	115.4	91.7	36.0
2-Methyl-1-butanol	128.0	93.8	41.5
3-Methyl-1-butanol	132.0	95.15	49.6
2-Methyl-2-butanol	101.8	87.35	27.5
3-Methyl-2-butanol	112.5	91.0	33
1-Hexanol	157.1	97.8	75
1-Heptanol	176.2	98.7	83
1-Octanol	195.15	99.4	90
2-Ethylhexanol	184.7	99.1	80.0

The temperature dependence of vapor pressure, heat of vaporization, specific heat, density, viscosity, surface tension, and thermal conductivity is given in [6] for methanol, ethanol, propanol, and butanol. For nomograms to determine the thermodynamic data of straight-chain alcohols, see [7]. The physical data of alcohols employed as solvents are summarized in [8], [9].

2.2. Chemical Properties

Reactions of alcohols can be characterized by cleavage of the O–H bond or the C–O bond either homolytically or ionically. The chemical properties of greatest industrial importance are as follows.

Oxidation and Dehydrogenation. Under normal conditions alcohols are stable. Oxidation with chemical agents, such as chromic acid or permanganate, leads to a variety of products depending on the nature of the alcohol. Primary alcohols are oxidized first to aldehydes and then to carboxylic acids; secondary alcohols are oxidized to ketones. Catalytic oxidation or dehydrogenation of primary and secondary alcohols on copper, silver, iron, molybdenum, etc., catalysts lead to the formation of aldehydes and ketones. In the Oppenauer oxidation a secondary alcohol is dehydrogenated by an excess of a ketone, e.g., acetone, in the presence of aluminum isopropoxide:

$$R^1-CHOH-R^2 + CH_3-CO-CH_3 \underset{}{\overset{Al(OC_3H_7)_3}{\rightleftharpoons}} R^1-CO-R^2 + CH_3-CHOH-CH_3$$

Reduction. With hydrogen iodide or zinc and hydrochloric acid, alcohols are converted to hydrocarbons. Catalytic hydrogenolysis is especially successful with benzyl alcohols.

Dehydration. Water can be split from alcohols by heating them in the presence of strong acid or by passing them over aluminum oxide, silicic acid, or synthetic zeolites. As a rule, not only those products that result from a β-elimination are formed, but also compounds with an isomerized double bond. The isomerization can be suppressed by the addition of amines. Tertiary alcohols can be dehydrated more easily than secondary or primary alcohols.

Under less severe conditions symmetric ethers are formed from alcohols in an intermolecular reaction:

$$2\ ROH \longrightarrow ROR + H_2O$$

Table 1. Physical properties of saturated alcohols

IUPAC Name		Other common name	Formula	M_r	*bp*, °C	*mp*, °C	n_D^{20}	d_4^{20}
Methanol	[*67-56-1*]	methyl alcohol	CH_3OH	32.04	64.7	− 97.8	1.3285	0.7910
Ethanol	[*64-17-5*]	ethyl alcohol	CH_3CH_2OH	46.07	78.32	−114.5	1.3614	0.7893
1-Propanol	[*71-23-8*]	propyl alcohol	$CH_3CH_2CH_2OH$	60.10	97.2	−126.2	1.3859	0.8035
2-Propanol	[*67-63-0*]	isopropyl alcohol	$CH_3CH(OH)CH_3$	60.10	82.4	− 87.8	1.3771	0.7850
1-Butanol	[*71-36-3*]	butylalcohol	$CH_3(CH_2)_2CH_2OH$	74.12	117.7	− 89.3	1.3991	0.8098
2-Methyl-1-propanol	[*78-83-1*]	isobutyl alcohol	$(CH_3)_2CHCH_2OH$	74.12	107.9	−107.9	1.3959	0.8027
2-Butanol	[*78-92-2*]	*sec*-butyl alcohol	$CH_3CH_2CH(OH)CH_3$	74.12	99.5	−114.7	1.3972	0.8065
2-Methyl-2-propanol	[*75-65-0*]	*tert*-butyl alcohol	$(CH_3)_3COH$	74.12	82.55	25.6	1.3841	0.7867
1-Pentanol	[*71-41-0*]	amyl alcohol	$CH_3(CH_2)_3CH_2OH$	88.15	137.8	− 78.5	1.4100	0.8150
2-Pentanol	[*6032-29-7*]	*sec*-amyl alcohol	$CH_3(CH_2)_2CH(OH)CH_3$	88.15	119.3		1.4053	0.8090
3-Pentanol	[*584-02-1*]	–	$CH_3CH_2CH(OH)CH_2CH_3$	88.15	115.6	− 75	1.4098	0.8218
2-Methyl-1-butanol	[*137-32-6*]	–	$CH_3CH_2CH(CH_3)CH_2OH$	88.15	128	<− 70	1.4098	0.816
3-Methyl-1-butanol	[*123-51-3*]	isopentyl alcohol	$CH_3CH(CH_3)CH_2CH_2OH$	88.15	131.4	−117.2	1.4078	0.812
2-Methyl-2-butanol	[*75-85-4*]	*tert*-amyl alcohol	$CH_3CH_2C(OH)(CH_3)CH_3$	88.15	101.8	− 11.9	1.4052	0.809
3-Methyl-2-butanol	[*598-75-4*]	–	$CH_3CH(CH_3)CH(OH)CH_3$	88.15	112.5		1.4095	0.819
2,2-Dimethyl-1-propanol	[*75-84-3*]	neopentyl alcohol	$CH_3C(CH_3)_2CH_2OH$	88.15	113.4	53.0		0.812
1-Hexanol	[*111-27-3*]	hexyl alcohol	$CH_3(CH_2)_4CH_2OH$	102.18	157.1	− 44.6	1.4178	0.8136
2-Methyl-1-pentanol	[*105-30-6*]	2-methylpentyl alcohol	$CH_3(CH_2)_2CH(CH_3)CH_2OH$	102.18	148		1.4190	0.8254
4-Methyl-1-pentanol	[*626-89-1*]	–	$(CH_3)_2CH(CH_2)_2CH_2OH$	102.18	152.3		1.4134 [a]	0.8110 [a]
4-Methyl-2-pentanol	[*108-11-2*]	methylamyl alcohol	$(CH_3)_2CHCH_2CH(OH)\text{-}CH_3$	102.18	131.8	− 90	1.4113	0.8066
2-Ethyl-1-butanol	[*97-95-0*]	2-ethylbutyl alcohol	$CH_3CH_2CH(C_2H_5)CH_2OH$	102.18	146.5	−114	1.4224	0.8348
1-Heptanol	[*111-70-6*]	heptyl alcohol	$CH_3(CH_2)_5CH_2OH$	116.20	176	−35	1.4233	0.8221
2-Heptanol	[*543-49-7*]	–	$CH_3(CH_2)_4CH(OH)CH_3$	116.20	158–160		1.4213	0.8173
3-Heptanol	[*589-82-2*]	–	$CH_3(CH_2)_3CH(OH)CH_2CH_3$	116.20	156.2	−70	1.4222	0.8210
4-Heptanol	[*589-55-9*]	–	$CH_3(CH_2)_2CH(OH)(CH_2)_2CH_3$	116.20	156	−41.5	1.4199	0.8183
2,4-Dimethyl-3-pentanol	[*600-36-2*]	–	$CH_3\text{-}CH(CH_3)\text{-}CH(OH)\text{-}CH(CH_3)\text{-}CH_3$	116.20	140		1.4246	0.8294

Table 1.(Continued)

IUPAC Name		Other common name	Formula	M_r	*bp*, °C	*mp*, °C	n_D^{20}	d_4^{20}
1-Octanol	[*111-87-5*]	octyl alcohol	$CH_3(CH_2)_6CH_2OH$	130.23	195.15	−16.3	1.4300	0.827
2-Octanol	[*123-96-6*]	capryl alcohol	$CH_3(CH_2)_5CH(OH)CH_3$	130.23	179	−38.6	1.4260	0.8205
2-Ethyl-1-hexanol	[*104-76-7*]	2-ethylhexyl alcohol	$CH_3(CH_2)_3CH(C_2H_5)CH_2OH$	130.23	184.7	−75	1.4315	0.8329
3,5-Dimethyl-1-hexanol	[*13501-73-0*]	–	$(CH_3)_2CHCH_2CH(CH_3)CH_2CH_2OH$	130.23	182.5		1.4250	0.8297
2,2,4-Trimethyl-1-pentanol	[*123-44-4*]	–	$(CH_3)_2CHCH_2C(CH_3)_2CH_2OH$	130.23	168	−70	1.4300	0.839
1-Nonanol	[*143-08-8*]	nonyl alcohol	$CH_3(CH_2)_7CH_2OH$	144.26	213.5	− 5	1.4323	0.8271
5-Nonanol	[*623-93-8*]	–	$CH_3(CH_2)_3CH(OH)(CH_2)_3CH_3$	144.26	193 – 194	−36	1.4299	0.8356
3,5-Dimethyl-4-heptanol	[*19549-79-2*]	–	$C_2H_5CH(CH_3)CH(OH)CH(CH_3)C_2H_5$	144.26	171		1.4330 [b]	0.836 [b]
2,6-Dimethyl-4-heptanol	[*108-82-7*]	diisobutyl carbinol	$CH_3CH(CH_3)CH_2CH(OH)CH_2CH(CH_3)CH_3$	144.26	178	−65	1.4231	0.8121
3,5,5-Trimethyl-1-hexanol	[*3452-97-9*]	–	$CH_3C(CH_3)_2CH_2CH(CH_3)CH_2CH_2OH$	144.26	194	−70	1.4300 [a]	0.8236 [a]
1-Decanol	[*112-30-1*]	decyl alcohol	$CH_3(CH_2)_8CH_2OH$	158.29	232.9	6.4	1.4359	0.8320
1-Undecanol	[*112-42-5*]	undecyl alcohol	$CH_3(CH_2)_9CH_2OH$	172.31	245	14.3	1.4392	0.8298
1-Dodecanol	[*112-53-8*]	lauryl alcohol	$CH_3(CH_2)_{10}CH_2OH$	186.34	259	23.8	1.4428	0.8306 [a]
2,6,8-Trimethyl-4-nonanol	[*123-17-1*]	–	$(CH_3)_2CHCH_2CH(CH_3)CH_2CH(OH)CH_2CH(CH_3)_2$	186.34	225	−60	1.4345	0.8193
1-Tridecanol	[*112-70-9*]	tridecyl alcohol	$CH_3(CH_2)_{11}CH_2OH$	200.33	276	30.6	1.4475	0.8454
1-Tetradecanol	[*112-72-1*]	myristyl alcohol	$CH_3(CH_2)_{12}CH_2OH$	214.39	170 – 173 [d]	38	1.4358 [c]	0.8165
1-Pentadecanol	[*629-76-5*]	pentadecyl alcohol	$CH_3(CH_2)_{13}CH_2OH$	228.42	170 [e]	44		0.8215 [c]
1-Hexadecanol	[*124-29-8*]	cetyl alcohol	$CH_3(CH_2)_{14}CH_2OH$	242.45	177 [e]	49	1.4392 [f]	0.8157 [f]
1-Heptadecanol	[*1454-85-9*]	margaryl alcohol	$CH_3(CH_2)_{15}CH_2OH$	256.48	191 [e]	54		
1-Octadecanol	[*112-92-5*]	stearyl alcohol	$CH_3(CH_2)_{16}CH_2OH$	270.50	210 [g]	57.6 – 58.0		0.8124
1-Nonadecanol	[*1454-84-8*]	–	$CH_3(CH_2)_{17}CH_2OH$	284.53		62		
1-Eicosanol	[*629-96-9*]	eicosanyl alcohol	$CH_3(CH_2)_{18}CH_2OH$	298.56	251 [e]	66		
1-Hexacosanol	[*506-52-5*]	ceryl alcohol	$CH_3(CH_2)_{24}CH_2OH$	382.72	305 [h]	79.5		

[a] 25 °C;
[b] 18 °C;
[c] 50 °C;
[d] 26.7 mbar;
[e] 13.3 mbar;
[f] 60 °C;
[g] 20 mbar;
[h] 26.7 mbar.

The same compounds that accelerate the intramolecular dehydration are suitable as catalysts.

Alcohols as Alkylation Reagents. Alcohols react with ammonia and amines to form *N*-alkyl or *N*,*N*-dialkylamines. Aromatic hydrocarbons are alkylated by alcohols in the presence of Friedel-Crafts catalysts.

Esterification. In the presence of acid catalysts, alcohols react with organic or inorganic acids, acid chlorides, or anhydrides to form esters (→ Esters, Organic).

Addition Reactions. Alcohols add to aldehydes and ketones to form acetals (→ Aldehydes, Aliphatic and Araliphatic). Alkylpolyglycol ethers are obtained with alkylene oxides. The addition to acetylene gives vinyl ether; addition to olefins yields mixed ethers.

For further reactions of alcohols see [10].

2.3. Production

The following processes have been realized on an industrial scale:

1) Synthesis from carbon monoxide and hydrogen (C_1)
2) Oxo synthesis (mostly combined with hydrogenation of the initially formed aldehydes; C_3-C_{20})
3) Hydrogenation of aldehydes, carboxylic acids, or esters
4) Aldol condensation of lower aldehydes and hydrogenation of the alkenals ($C_3 \rightarrow C_6$, $C_4 \rightarrow C_8$, $C_8 \rightarrow C_{16}$)
5) Oxidation of trialkylaluminum compounds (Ziegler process)
6) Oxidation of saturated hydrocarbons
7) Hydration of olefins (C_2-C_4)
8) Homologation of alcohols
9) Hydrocarbonylation by the Reppe process
10) Hydrocarboxymethylation
11) Fermentation processes (C_2-C_5)
12) Guerbet process

The most important industrial processes are the methanol synthesis, with an annual production of ca. 12×10^6 t, and the oxo synthesis, with ca. 4.5×10^6 t. However, the hydration of ethylene and propene to ethanol and 2-propanol, and the oxidation of trialkylaluminum compounds (Alfol process or Ziegler process) also have achieved considerable commercial significance. Fermentation, especially for the production of ethanol, has become important again in certain regions because of the increased price of oil.

2.3.1. Synthesis from Carbon Monoxide and Hydrogen

On a commercial scale only methanol is prepared from synthesis gas (→ Methanol). Synol, isobutylol, oxyl, and similar processes led to a mixture of oxygen-containing compounds with alcohols as the main components [11], [12]. These methods are no longer used in the Western World. Newer developments also produce oxygen-containing compounds [13] or alcohol mixtures [14].

In the IFP (Inst. Français du Pétrole) process for higher alcohols [14], highly activated catalysts are used under low-pressure methanol synthesis conditions. The catalyst contains mixed oxides of copper and cobalt plus at least one other metal (Al, Ce, Cr, Fe, La, Mn, Pr, Nd, Y, or Zn) and at least one Group I or II metal compound. The higher alcohol content can be varied from 20–50 wt % by changing the catalyst composition. A number of byproducts, such as hydrocarbons, esters, and ketones, also are produced.

However, none of these newer processes are used in industry.

Small amounts of higher alcohols are formed as byproducts of the Fischer-Tropsch synthesis [15] (→ Coal Liquefaction).

2.3.2. Oxo Synthesis

For details → Oxo Synthesis.

Alcohols in the range C_3-C_{20} can be prepared by the oxo synthesis, in which olefins react with synthesis gas to form aldehydes, which in turn are hydrogenated.

One particular version of the oxo synthesis is the *Shell process*; the strong hydrogenating activity of the catalyst, $HCo(CO)_3PR_3$, leads to the direct hydrogenation in the oxo reactor of the initially formed aldehyde [16]:

$$R-CH=CH_2 + CO + 2\,H_2 \xrightarrow{\text{cat.}} R-CH_2CH_2CH_2OH$$

The process was first applied commercially on a propene basis in 1963 and used for the manufacture of 1-butanol and 2-ethylhexanol [17]. In 1965 it was developed further to produce detergent alcohols. It is used primarily to make higher alcohols.

Planned or actual expansions of the Shell process capacity in the United States (previously 250000 t/a) by 46000 t/a, of surfactant alcohols in the Federal Republic of Germany by 56000 t/a, and of Mitsubishi's in Japan by 70000 t/a [18], will increase world capacity to some 470000 t/a.

The Shell process has the advantage that olefins with an internal double bond can be hydroformylated, because under the reaction conditions, isomerization of the double bond takes place, and α-olefins are formed. For example, ω-olefin fractions obtained from ethylene by the "SHOP process" of Shell [19] (→ Hydrocarbons, Chap. 2.) can be converted with synthesis gas to alcohols. The alcohol mixtures formed consist of up to 80 % linear compounds and are used in the plasticizer and detergent fields.

2.3.3. Hydrogenation of Aldehydes, Carboxylic Acids, and Esters

Aldehydes can be hydrogenated in the presence of homogeneous or heterogeneous catalysts [20], [21]. Homogeneous systems are advantageous only if sulfur-containing starting materials, which would poison the heterogeneous catalysts, are used or if the hydrogen for the hydrogenation contains carbon monoxide.

Generally, however, heterogeneous catalysts are preferred. These are effective both in the gas phase at temperatures of 90 – 180 °C and pressures of 25 bar and in the liquid phase at 80 – 220 °C and pressures up to 300 bar. The hydrogenation temperature applied in industrial processes represents a compromise between the best possible energy utilization and high catalyst lifetimes.

For continuous processes, catalysts in fixed-bed systems are favored. The aldehyde, either as vapor strongly diluted with excess hydrogen [22] or as liquid together with hydrogen [23], is fed through the high-pressure pipe containing the catalyst bed. To remove the heat generated by the reaction, the hydrogen is circulated through a heat exchanger. The especially important hydrogenation of 2-ethyl-2-hexenal usually is carried out in a single step on a nickel-containing catalyst [24]. Similar one-step processes have been suggested [25]. However, two-step processes are also common, whereby the main hydrogenation takes place in the gas phase (e.g., on copper-containing catalysts), and the second in a liquid phase or trickle-bed process (e.g., on nickel-containing catalysts).

The catalysts usually are supported on aluminum oxide or silica gel. In addition to nickel and copper, also zinc, chromium, and combinations of these metals have been used successfully as catalysts [26].

For the manufacture of fatty alcohols, the corresponding carboxylic acid esters are hydrogenated (for details → Fatty Alcohols). Starting materials are natural fats and oils, which are first transesterified to the methyl esters and then reduced to the alcohols, either with sodium (Bouveault-Blanc reduction) or by catalytic hydrogenation. The reduction with sodium enables the preparation of unsaturated fatty alcohols from the esters of unsaturated fatty acids.

The hydrogenation of fatty acids and fatty acid esters requires more drastic conditions than the aldehyde hydrogenation. The process operates continuously or discontinuously and makes use of copper–chromium oxide catalysts (Adkins catalysts), which are either suspended or placed in a fixed bed. Reaction temperatures of 240 – 300 °C and pressures of 200 – 300 bar are normal. By modifying the catalysts (e.g., by addition of cadmium) unsaturated fatty acid esters can be converted directly to unsaturated fatty alcohols [27].

2.3.4. Aldol Condensation of Lower Aldehydes and Hydrogenation of the Alkenals

In industry, the only source of aldehydes for the aldol condensation is the oxo synthesis [21]. After the isoaldehydes and byproducts are removed, the condensation is catalyzed by acids or bases (→ Aldehydes, Aliphatic and Araliphatic, Chap. 4.2.). Because the reactivity of each aldehyde depends on the chain length and the degree of branching, the reaction conditions are

adapted according to the individual compound. The alkenals, formed from the aldols by elimination of water, are hydrogenated over heterogeneous catalysts. As a rule the same catalysts are used as for the hydrogenation of oxo aldehydes. 2-Ethylhexanol, 2-methylpentanol, and limited amounts of highly branched, isomeric C_{16} and C_{18} alcohols are prepared by this method. The so-called aldox process, in which the aldehyde mixture formed in the oxo synthesis is subjected to aldol condensation in the oxo reactor, has not become established because of the presence of mixed aldols in the final products; however, some mixed aldols of acetaldehyde with higher aldehydes are important.

2.3.5. Oxidation of Trialkylaluminum Compounds

Ethylene can add to triethylaluminum to form a mixture of trialkylaluminum compounds of higher molecular mass [28]. These products can be oxidized with air to the corresponding aluminum alkoxides, which are then hydrolyzed to a mixture of linear primary alcohols with the same number of carbon atoms as the alkyl groups of the trialkylaluminum components [1–3]:

$$-\overset{|}{Al}-C_2H_5 + x\,C_2H_4 \longrightarrow -\overset{|}{Al}-(CH_2CH_2)_x-C_2H_5$$
$$\xrightarrow{O_2} -\overset{|}{Al}-O-(CH_2CH_2)_x-C_2H_5$$
$$\xrightarrow{H_2O} HO-(CH_2CH_2)_x-C_2H_5 + -\overset{|}{Al}-OH$$

On the basis of this reaction, called the Ziegler process, two commercial processes have been developed, one by Conoco, in operation in the United States since 1962 and in the Federal Republic of Germany by Condea Chemie (Conoco and Deutsche Texaco) since 1964; and the other by the Ethyl Corp., in operation since 1965. The principal differences between the two processes are the chain-length distribution and the linearity of the alcohols produced, as well as the technical process characteristics for the generation and control of this distribution.

Alfol Alcohol Process (Conoco Process). The chain-growth reaction is carried out at a temperature as low as possible in order to prevent displacement reactions that lead to the formation of olefins. The chain-length distribution corresponds to a Poisson curve. The resulting alcohols are practically 100 % linear. A broad range of alcohols (C_2-C_{28}) is typical for the Alfol alcohol process. Although the process can be varied to either increase or decrease the chain lengths of the manufactured alcohols, the distribution pattern remains the same.

Process with Controlled Linear Chain Growth. The Ethyl Corp. has successfully developed a process which forms predominantly C_{12} and C_{14} alcohols. The alcohols are up to 95 % linear.

The product distributions obtained in the Conoco and the Ethyl Corp. processes are compared in Table 3.

Table 3. Composition of the alcohol mixtures from the Ziegler process

C no.	Conoco	Ethyl Corp.
6	9.6	1.4
8	16.9	3.2
10	20.7	7.7
12	19.4	34.5
14	15.1	26.3
16	9.8	16.7
18	5.3	8.9
20	3.2	1.3

2.3.6. Oxidation of Saturated Hydrocarbons

Bashkirov Oxidation. The oxidation of aliphatic hydrocarbons with air in the presence of boric acid gives boric acid esters in high yield. These are hydrolyzed in a second step to secondary alcohols in which the hydroxyl groups are distributed statistically along the molecular chain [29–31].

The reaction passes through an intermediate secondary hydroperoxide, for the reaction mechanism see [31], [32]. Normally a mixture of *n*-hydrocarbons with chain lengths between 10 and 16 is used as feedstock. The oxidation is carried out in the liquid phase at 150 – 170 °C in the presence of 4 – 5 wt % metaboric acid. A nitrogen-oxygen mixture (ca. 3.5 % O_2) is used at normal or slightly elevated pressure. In newer plants amines are added as co-catalysts [31]. The resulting metaboric acid esters or boroxines are resistant to oxidation and thermally stable. To obtain economically acceptable selectivities of 80 – 85 % , the level of conversion must be held below 20 %.

Starting material and oxidation byproducts are removed by flash evaporation and are cleaned in alkaline and water scrubbers; the *n*-hydrocarbons are recycled [30]. The metaboric acid esters at the bottom of the flash column are hydrolyzed by the addition of small amounts of water at 80 – 100 °C. Impurities (mostly carbonyl compounds) are removed by alkaline and water washes. After fractional distillation, alcohols are obtained with a purity greater than 98 %. Processing concludes with "hydrofinishing" (hydrogenation over heterogeneous nickel catalysts) to remove colored and odorous substances.

The orthoboric acid in the aqueous solution is converted to metaboric acid by dehydration. Recovery also is possible by crystallization. In more recent process variants the boric acid concentrate is mixed with the *n*-hydrocarbons and the mixture dehydrated, preventing agglomerations and stoppages. The dried slurry is returned to the oxidation reactor.

In the 1950s Bashkirov developed the original German work into a commercial process [28], [33]. The first plant began production in Shebekino/Belgorod, USSR, in 1959 [34]. Plants were operated in the Soviet Union [34] and in Japan [31]. The Union Carbide (UCC) plant in the United States was closed in 1977 [31].

The secondary alcohols produced in this process usually are converted to alkylphenol ethoxides and used as detergents. The costly recirculation of hydrocarbons and boric acid as well as disadvantages in the application of secondary alcohols [35], however, have prevented this process from achieving any great significance in the Western World.

A further important use of boric acid-catalyzed oxidation is the preparation of cyclohexanol and cyclohexanone from cyclohexane [36] (intermediate products in the manufacture of caprolactam, adipic acid, and phenol), and of cyclododecanol (intermediate product in the nylon 12 synthesis) and cyclododecanone from cyclododecane [37]; → Cyclohexanol and Cyclohexanone; → Cyclododecanol, Cyclododecanone, and Laurolactam.

Oxidation with Alkyl Hydroperoxides. Alcohols also can be obtained by hydroxylation of alkanes with alkyl hydroperoxides [38]. Iron porphyrins are especially useful as catalysts.

$$RH + R'OOH \xrightarrow{cat.} ROH + R'OH$$

Alcohols from Fatty Acids Produced by Hydrocarbon Oxidation. The oxidation of hydrocarbons with air in the presence of manganese catalysts leads to a complex mixture of reaction products [39]. The initially formed alcohols are oxidized further to ketones and acids, → Fatty Acids. At present the process is operated mainly in the former Soviet Union. The production of fatty acids in this way has roughly the same order of magnitude as that of fatty acids from native raw materials in the United States [3].

Processing of the reaction mixture is difficult, because not all byproducts can be removed. Therefore, these fatty acids do not reach the same standards of quality as natural fatty acids.

Alcohols are formed from fatty acids by esterification of the fractionated raw acids (mostly in the $C_{10} - C_{15}$ range) with methanol or butanol and subsequent hydrogenation. With linear hydrocarbons as starting materials, linear alcohols can be produced; however, depending on the catalyst, the specific hydrocarbon, and the reaction conditions, the alcohol mixture contains 5 – 15 % branched-chain alcohols. Owing to the impurities present in the fatty acids, the alcohols prepared from them contain odorous substances, so their range of applications is limited.

According to estimates [3], about 10 % of these synthetic fatty acids are processed to alcohols. The annual production in Eastern Block countries is presumed to be at least 50000 t/a.

2.3.7. Hydration of Olefins

A common method for the production of lower alcohols is the hydration of alkenes. In accordance with Markovnikov's rule, secondary and tertiary alcohols are formed (except in the case of ethylene).

$$R-CH=CH_2 + H_2O \xrightarrow{H^+} R-CHOH-CH_3$$

The rate of this reaction is determined by the stability of the intermediate carbenium ion (tertiary > secondary > primary). Therefore the hy-

dration of isobutene proceeds at room temperature in the presence of low H^+ ion concentrations owing to the relative stability of the intermediate tertiary carbenium ion. The hydration of ethylene, in contrast, requires elevated temperatures and pressures [10].

Industrially two variants of the hydration reaction are used. In the so-called *indirect process,* the liquid-phase reaction takes place in two steps. In the first, the olefin reacts with sulfuric acid to form mono- and dialkylsulfates, which, after dilution with water, are hydrolyzed to the alcohol. However, in order to recycle the sulfuric acid, a costly reconcentration is necessary.

In the *direct process,* hydration occurs in the gas phase. Because the reaction is exothermic and is accompanied by a reduction in volume (2 mol of reactant form 1 mol of product), the alcohol formation is favored by high pressure and low temperature. Because the conversion is incomplete, a costly gas recycle is necessary. Phosphoric acid-containing materials, e.g., celite, are efficient catalysts. However, increased use is made of ion exchangers.

The primary use of direct hydration is for the preparation of ethanol from ethylene (→ Ethanol) and of isopropyl alcohol from propene (→ Propanols). It is also important in the production of 2-butanol from a mixture of 1-butene and 2-butene (raffinate II) and of *tert*-butyl alcohol from isobutene.

2.3.8. Homologation of Alcohols

Homologation is the reaction of alcohols with synthesis gas in the presence of complex, multicomponent catalyst systems. Depending on reaction conditions the products are aldehydes or alcohols containing one CH_2 group more than the starting materials [40–43]:

$$RCH_2OH + CO + H_2 \xrightarrow{cat.} RCH_2CHO + H_2O$$
$$RCH_2OH + CO + 2\,H_2 \xrightarrow{cat.} RCH_2CH_2OH + H_2O$$

Although the reaction was originally conceived for the synthesis of ethanol from methanol [44], [45], the scope has been extended to include the production of homologous aldehydes (acetaldehyde from methanol), carboxylic acids (propionic acid from acetic acid), carboxylic acid esters (ethyl acetate from methyl acetate), as well as the synthesis of styrene (via the homologation of benzyl alcohol to 2-phenylethanol with subsequent dehydration) [28], [31].

So far, homologation has not been used industrially because conversion and selectivity, despite considerable advances [28], are still insufficient and because there are problems with recycling the complex homologation catalysts.

2.3.9. Reppe Process

The Reppe hydrocarbonylation of olefins with carbon monoxide and water and using ammonium salts of tetracarbonyldihydrido iron as catalyst leads to alcohols [46] (→ Carbonylation):

$$C_3H_6 + 3\,CO + 2\,H_2O \xrightarrow{cat.} C_4H_9OH + 2\,CO_2$$

As in the oxo synthesis, branched-chain products also are formed (molar ratio of linear to branched-chain alcohols is ca. 9 : 1).

Propene reacts at 90 – 110 °C and 5 – 20 bar to form butanols with yields of 90 %. Approximately 4 % of the propene is hydrogenated to propane. The conversion of higher olefins requires more extreme conditions.

The process cannot compete with hydroformylation. The only plant using it to manufacture butanol from propene (Japan Butanol Co., capacity 30000 t/a) closed several years ago [47].

2.3.10. Hydrocarboxymethylation

Hydrocarboxymethylation is a variant of the Reppe process in which higher olefins react with carbon monoxide and methanol in the presence of a cobalt-pyridine catalyst. The products are esters of carboxylic acids containing one more carbon atom in the parent chain than the olefin feedstock [48]. The esters can be hydrogenated to the alcohols. For economic reasons – at the present time the products can be prepared more cheaply from natural raw materials – the process has not as yet achieved any significance.

2.3.11. Fermentation

Fermentation, probably the oldest process for the manufacture of ethanol (→ Ethanol), is still

practiced on a large scale [49]. The butanol-acetone fermentation of carbohydrate raw materials is no longer of any importance [50]. On a small scale, pentanols are recovered from fusel oils [10], [51].

2.3.12. Guerbet Alcohols

In the Guerbet process saturated primary alcohols are dimerized to α-branched primary alcohols [52]. Normally the reaction is carried out by refluxing the alcohol in the presence of an alkaline condensation agent and a hydrogenation-dehydrogenation catalyst, e.g.:

$$2\,n\text{-}C_6H_{13}OH \xrightarrow[\text{Raney Ni}]{\text{NaOH}} C_6H_{13}\overset{\overset{C_4H_9}{|}}{C}H-CH_2OH + H_2O$$

The water and small amounts of hydrogen produced in the reaction are removed continuously. For the mechanism of the reaction see [53], [54]. If heating is carried out for long periods, trimeric α-branched primary alcohols also are formed.

The yield of dimeric alcohols is about 80 %. It can be increased by recycling the residues and portionwise addition of fresh catalyst [55]. Metallic sodium, as well as a number of other substances have been proposed as condensation agents [10]. For industrial purposes alkali metal hydroxides are preferred.

Alcohols having chains shorter than cetyl alcohol (C_{16}) preferably are dimerized at higher pressure [56]. Using this process, short-chain alcohols can be converted into $C_{10}-C_{20}$ alcohols. Examples include 2-hexyldecanol [57] and 2-octyldodecanol [58], which find application in cosmetics as oily components with favorable solvent properties.

Because alcohols with the typical α-branching [59] are prepared more easily by other methods, e.g., 2-ethyl-1-hexanol by hydroformylation of propene to give butanal and subsequent aldol condensation, the Guerbet reaction has not become established as a large-scale industrial process.

2.3.13. Other Processes

Of commercial interest is the epoxidation of linear α-olefins and subsequent hydrogenating cleavage to plasticizer (C_6-C_{10}) or detergent ($C_{10}-C_{16}$) alcohols [60]. For hydrogenating cleavage a number of catalysts can be applied [61]. The selectivity towards primary alcohols can be influenced by the choice of catalyst, the hydrogenation conditions, and the use of solvents [10], [62].

The hydrolysis of carboxylic acid esters [63] is also of some importance. For a comprehensive review of other processes see [10].

2.4. Individual Alcohols

The following sections describe individual alcohols and special alcohol mixtures, classified according to increasing number of carbon atoms. In each case the specific method of manufacture, the main producers, and trade names, as well as the necessary basic feedstocks are mentioned.

The C_1-C_5 alcohols are described under individual keywords.

2.4.1. C_6 Alcohols

For physical properties see Table 1.

1-Hexanol [*111-27-3*], $CH_3(CH_2)_4CH_2OH$, is prepared according to the Ziegler process (see Section 2.3.5) from ethylene (Alfol 6, Condea; Epal 6, Ethyl Corp.) or is made from natural products derived from coconut or palm oils (Lorol C 6, Henkel). It is used as a solvent, as a basic material for the perfume industry, and for the production of plasticizers (in this case usually as a mixture with higher *n*-alcohols). Nitrates of 1-hexanol are recommended as cetane number improvers. Commercial specifications are given in Table 4.

2-Methyl-1-pentanol [*105-30-6*] is prepared by aldol condensation of propionaldehyde and subsequent hydrogenation of the intermediate 2-methyl-2-pentenal, (see Section 2.3.4). It is used as a solvent.

$CH_3(CH_2)_2CH(CH_3)CH_2OH$

Commercial specifications are given in Table 4. Manufacturers are, e.g., Ruhrchemie and UCC [69], [70].

Table 4. Typical specifications of commercial plasticizer alcohols [9], [64–76]

	Alcohol, wt %	Color, Hazen no. [a]	Density [b], g/cm^3	Boiling range [c], °C	Hydroxyl value [d], mg KOH/g	Acidity [e], mg KOH/g	Moisture [f], wt %	Carbonyl number [g]
1-Hexanol	98	<10	0.819	150–170	540–555	<0.05	<0.5	–
2-Methyl-1-pentanol	≧98	5	0.820	148	–	0.1	<0.3	≈0.5
2-Ethyl-1-butanol	≧98	<10	0.831–0.834	156–150	–	<0.2	<0.2	<0.3
C_6 Alcohol mixture	≧98	10	0.819–0.821	151–159	543	0.003	0.2	0.2
Isooctanol	99	10	0.830–0.834	184–191	428	0.001	0.1	0.15
Isodecanol	99	<10	0.835–0.841	215–225	350	0.05	<0.1	0.3
Isotridecanol	99	10	0.843–0.848	250–266	285	0.05	<0.1	0.4

[a] DIN 53409; ASTM D 1209–69
[b] DIN 51757; ASTM D 1298–67
[c] DIN 53171; ASTM D 1078–70
[d] DIN 53240; ASTM D 1957–63
[e] DIN 53402; ASTM D 1613–66
[f] Karl Fischer reagent
[g] Oximation

4-Methyl-2-pentanol [*108-11-2*] is a byproduct of the synthesis of methylisobutylketone [*108-10-1*] (→ Ketones).

$(CH_3)_2CHCH_2CH(OH)CH_3$

The alcohol is sold by UCC, for example, under the name Methylamylalcohol [69]. It is used as a solvent in the paint industry, as a brake fluid, as a flotation aid [71], and as a fungicide [72].

2-Ethyl-1-butanol [*97-95-0*] is prepared by aldol condensation (→ Aldehydes, Aliphatic and Araliphatic) of 1-butanal and acetaldehyde and subsequent hydrogenation.

$(C_2H_5)_2CHCH_2OH$

It is a neutral, colorless, pleasant-smelling liquid and is employed as a solvent and flow improver for paints and varnishes, as a component in the manufacture of penetrating oils and corrosion inhibitors, as a cleaning agent for printed circuits, as an extracting agent for metal ions, and as the alcohol component in the manufacture of plasticizer phthalates for special uses, of nitric acid esters as cetane number improvers, and of phosphoric acid esters as plasticizer auxiliaries.

Manufacturers are, e.g., Ruhrchemie and UCC. Commercial specifications are given in Table 4.

C_6 Alcohol Mixtures. These are prepared by the hydroformylation of 1-pentenes and subsequent hydrogenation. Esso/Enjay offers, for example, an isomeric mixture under the name Hexanol that consists of 44 % 1-hexanol, 53 % 2-methylpentanol, and 3 % 2-ethylbutanol [64]. The corresponding plasticizer, DHP (dihexylphthalate), has found application only in special cases. Commercial specifications are given in Table 4.

Hexanols have good solvent properties for fats and oils. Because of their lower volatility and higher viscosity, they are superior to pentanols in several applications, e.g., as flow improvers.

C_6–C_{10} Alcohol Mixtures. Condea [65] and the Ethyl Corp. [73] offer mixtures of C_6, C_8, and C_{10} alcohols under the trade names Alfol 610 and Epal 610, respectively. The 100 % linear alcohols are prepared by the Ziegler process (see Section 2.3.5). For the composition and typical data of alcohol mixtures with various C numbers, see Table 5.

The phthalates are suitable as plasticizers for coatings, artificial leather, poly(vinyl chloride) films and sheets, and anticorrosive pastes. The esters of aliphatic acids, also plasticizers, provide cable coverings, which are extremely resistant to cold, films, and artificial leather. They also find application in food wrapping.

Table 5. Composition and typical data of alcohol mixtures with various carbon numbers [65], [67], [73–76]

Manufacturer	Trade name	Chain-length distribution, wt % alcohol	Boiling range, °C	Density (20 °C), g/cm^3	n_D^{20}	Max. water content, wt %	*n*-Alcohol content, wt % (approx.)
Monsanto	Oxo alcohol 7911	C_7 30–34 C_9 35–43 C_{11} 27–31	178–288	0.826–0.832	1.420–1.440	0.01	65–70
	Oxo alcohol 7900	C_7 41–49 C_9 51–59 C_{11} max. 1	n.a.*	0.825–0.831	n.a.	0.01	
	Oxo alcohol 1100	C_{11} approx. 99	243–274	0.831–0.837	1.439	0.01	
ICI	Alphanol 79	$C_7/C_8/C_9$ mixture	177–202	0.835	1.431	0.02	n.a.
Condea (Conoco)	Alfol 610	C_6 20 C_8 35 C_{10} 44	165–235	0.826	n.a.	0.3	100
	Alfol 810	C_8 43 C_{10} 55	195–240	0.827	n.a.	0.2	100
Shell	Dobanol 91	C_9/C_{11} mixture	225–248	0.835	n.a.	0.1	80
	Linevol 79	C_7 44, C_8 36, C_9 20	183–214	0.827–0.833	1.430	0.1	80
	Linevol 911	C_9 20, C_{10} 50, C_{11} 30	228–248	0.833–0.839	1.439	0.1	80
Ethyl Corp.	Epal 610	C_6 17 C_8 36 C_{10} 47	183–242	0.825	1.428	0.08	100
	Epal 810	C_6 1 C_8 45 C_{10} 54	205–240	0.827	1.431	0.02	100
Henkel	Lorol 810	C_6 0–5 C_8 50–66 C_{10} 30–40 C_{12} 3–8	n.a.	approx. 0.82	n.a.	<0.5	100
Ugine Kuhlmann	Linopol 7–11	C_7 33–36 C_9 38–44 C_{11} 21–28	183–245	0.827–0.830	n.a.	<0.1	70
	Acropol 91	C_9 50–53 C_{11} 47–50	215–248	0.830–0.834	n.a.	<0.1	70

* not available.

Mixtures of 1-hexanol and 1-octanol [*111-87-5*] serve as frothing agents (bubbling promoters) in flotation, e.g., of coal [77]. Large amounts are added to aqueous drilling muds to prevent frothing during drilling for oil and gas.

2.4.2. C_7 Alcohols

For physical properties see Table 1.

1-Heptanol [*111-70-6*] in its pure form has very little commercial value. It can be prepared from 1-hexene by oxo synthesis. Isomeric mixtures of various heptanols are of greater significance.

C_7 Alcohol Mixtures. Heptanol mixtures are prepared by the hydroformylation of isohexene, the dimerization product of propene. They have been used by Chisso and Nissan in Japan and also by Exxon for the manufacture of the phthalate plasticizer DIHP (diisoheptyl phthalate) or Jayflex 77, a fast-dissolving plasticizer with good low-temperature properties that is said to be particularly suited for poly(vinyl chloride) flooring and for use in plastisols [78].

C_7–C_{11} Alcohol Mixtures. From a linear $C_6/C_8/C_{10}$ olefin stream of the Ethyl Corp., Monsanto produces the corresponding alcohols with 7, 9, and 11 carbon atoms via oxo synthesis. Depending on the reaction conditions, the content of iso compounds is 30–35 %. In ad-

dition to $C_7/C_9/C_{11}$ mixtures, C_7/C_9 and C_{11} fractions also are available [74]. In the United States, practically the whole production is processed further to phthalates or adipates, used as plasticizers (trade name Santicizer).

Shell also produces mixtures of C_7, C_9, and C_{11} alcohols that are sold under the trade names Dobanal 91 (Europe) and Neodol 91 (United States, Canada). They are prepared by the Shell oxo process from linear olefins with internal double bonds and are ca. 80 % linear; corresponding product distributions are obtained by using α-olefins. The C_7/C_9 and C_9/C_{11} fractions are sold under the trade names Linevol 79 and Linevol 911 [75].

Imperial Chemical Industries (ICI) offers $C_7/C_8/C_9$ alcohol mixtures under the trade name Alphanol 79 [66]. Linopol 7 – 11, a mixture of $C_7/C_9/C_{11}$ primary and α-alkyl-branched alcohols with ca. 70 % *n*-isomer content, is produced by Kuhlmann [76].

Typical data for these mixtures as given in the various companies' data sheets are presented in Table 5. The mixtures are colorless liquids, miscible with most organic solvents but not with water.

Uses. Approximately 70 – 80 % are processed to plasticizers, mostly phthalates, which to a certain extent replace the plasticizer DOP, dioctyl phthalate, di-(2-ethylhexyl)phthalate, made from phthalic anhydride and 2-ethylhexanol (→ Plasticizers). However, DOP still commands the major share of the market.

These esters of dicarboxylic acids are used almost exclusively as plasticizers for PVC. The linear esters exhibit somewhat better low-temperature properties, greater resistance toward oxidation, and lower volatility than DOP. For example, in the United States the plasticizers Santicizer 711 (phthalate) and Santicizer 97 (adipate), manufactured by Monsanto from $C_7/C_9/C_{11}$ alcohol mixtures, have practically displaced DOP in automobile interiors because, owing to the lower volatility, no fogging of the windows takes place.

Further applications of the alcohols depend on their solvent properties. For alcohols with an odd number of carbon atoms, better solvent and wetting properties are claimed than for those with an even number [9], [68].

The alcohol mixtures are used as solvents or solubilizers in the paint and printing ink sector, as components in textile auxiliaries and pesticides, for hormone extraction, and in the surfactant field as foam boosters or antifrothing agents.

2.4.3. C_8 Alcohols

For physical properties see Table 1.

2-Ethyl-1-hexanol [*104-76-7*] is the most important C_8 alcohol (→ 2-Ethylhexanol).

$CH_3(CH_2)_3CH(C_2H_5)CH_2OH$

1-Octanol [*111-87-5*], $CH_3(CH_2)_6CH_2OH$, capryl alcohol. Its esters are widespread in nature; e.g., they occur in grapefruits, oranges, or green tea. 1-Octanol is manufactured by the Alfol process and from natural products. It is sold under the trade names Alfol 8, by Condea, and Lorol C 8, by Henkel. The alcohol is used, for example, in the perfume industry.

2-Octanol [*123-96-6*], $CH_3(CH_2)_5CH(OH)CH_3$, is obtained by the alkaline hydrolysis of castor oil. It is used as a solvent in the paint industry, as a wetting agent in the textile industry, and as a component of brake fluids.

C_8 Alcohol Mixtures (Isooctyl Alcohol). Butenes and propene are codimerized to heptenes which, in turn, are hydroformylated to mixtures of isomeric C_8 aldehydes. These are hydrogenated to yield mixtures of primary alcohols. In addition to 3,4-, 3,5-, and 4,5-dimethyl-1-hexanol and 3- and 5-methyl-1-heptanol, such mixtures contain varying amounts of heptanols and nonanols. They are colorless, slightly viscous, neutral liquids and have a mild fragrance. They are practically immiscible in water (0.06 wt % at 20 °C), but absorb up to 3.5 wt % water at 20 °C. They are miscible with most organic solvents. The mixtures are available under the names isooctanol or isooctyl alcohol from Ruhrchemie, Kuhlmann, ICI, Esso/Enjay, and UCC. The specifications of a typical commercial product are summarized in Table 4.

Isooctyl alcohol is an excellent solvent for many organic compounds, such as fats, oils, and waxes, as well as various rubber formulations

and resins. Isooctyl alcohol is especially suitable for baking enamels because, owing to its low vaporization tendency, cratering and blistering do not occur.

Further applications are as a foam suppressant, an extracting agent for metal salts (Mo, Re, Co) and mineral acids (e.g., phosphoric acid, boric acid from aqueous solutions), an emulsifier and stabilizer in oil emulsions, a modifier for PVC pastes, a polishing agent for acrylic ester polymers, and as a bath additive in galvanization.

Derivatives. As with all higher oxo alcohols, the principal use of isooctyl alcohol is as an esterification component in the preparation of plasticizers, mainly diisooctyl phthalate (DIOP), but also of esters of adipic, sebacic, azelaic, and trimellitic acids. Triisooctyl trimellitate, Jayflex 80-TM, is available from Exxon [78]. As a highly permanent monomeric plasticizer with low viscosity and low specific mass, it competes with polymeric plasticizers.

Various derivatives of isooctyl alcohol, for example, ethers or acetates are suitable as solvents for polymers. If metallic soaps are added, these products can be used as lubricating greases. Some esters are suitable as lubricants or hydraulic fluids.

For the preparation of surfactants, ethoxides or sulfuric acid esters of isooctyl alcohol are recommended. Organophosphorus derivatives, such as metal salts of diisoctyldithiophosphoric acid, are reported to be suitable as additives for lubricants, fuels, and hydraulic fluids.

In agriculture, the isooctyl esters of 2,4-dichloro- and 2,4,5-trichlorophenoxyacetic acid are used as nonvolatile herbicides.

C_8-C_{12} Alcohol Mixtures. Henkel supplies C_8/C_{10} and C_{10}/C_{18} alcohol fractions based on natural fats and oils under the names Lorol 810 and Lorol 818 [67]. The C_8/C_{10} fraction contains 3 – 8 % C_{12} alcohols (Table 5); Lorol 818 contains a wide range of C_8-C_{18} alcohols. Because alcohols from coconut and palm oils are available in only limited quantities, however, by far the greater portion of the alcohols in this range is derived from petrochemical sources. For example, Condea produces Alfol 810 and Alfol 1012 (C_8/C_{10} and C_{10}/C_{12} alcohol mixtures) from ethylene by the Ziegler process [65]. Similar products from the same feedstock are the Epal C_8/C_{10} and C_{10}/C_{12} fractions offered by the Ethyl Corp. [73].

The C_8/C_{10} and C_{10}/C_{12} alcohol mixtures, produced from natural products and by petrochemical methods, are both practically 100 % linear.

Mixtures of 1-octanol and 1-decanol are used by the tobacco industry as growth inhibitors [79].

Derivatives. Esters of aliphatic acids and of trimellitic acid with C_8-C_{10} alcohols are suitable as plasticizers for weather-resistant films and artificial leather as well as for special poly(vinyl chloride) applications, such as thermally stable cable coverings.

Upon addition of 3 – 6 mol ethylene oxide to C_8-C_{10} alcohols, ethoxides are obtained that possess some advantages over longer-chain detergents owing to their excellent solubility in water, improved viscosity, and especially good wettability. They are used as frothing agents in oil drilling, as components in gentle and low-temperature washing liquids, in technical cleansers, and also in the textile industry.

By sulfatation of the ethoxides the corresponding ether sulfates are produced, which possess high stability toward electrolytes and alkaline-earth metal ions. They are used as industrial foaming agents, e.g., for plaster and mortars, and for oil drilling. The sodium salts of alcohol sulfates are used to a slight extent as wetting agents with high stability towards electrolytes.

Primary alkylphosphates are prepared from the alcohols and phosphorus pentoxide. After neutralization with amines or alkali hydroxides, they are used as water-soluble anionic surfactants with high stability toward hydrolysis and as emulsifiers. Other derivatives of these alcohols, especially in the C_8-C_{10} range, include the corresponding amines. They are used for the solvent extraction of heavy metals, such as molybdenum, vanadium, uranium, and thorium, from acidic aqueous solutions. Quaternary ammonium salts with C_8-C_{12} alkyl groups are added to such cleaning agents as bactericides and fungicides.

Ether amines are obtained by cyanoethylation with acrylonitrile and subsequent hydrogenation (Sherex, Henkel). They are used as selective collectors for the flotation of silicates in low-quality iron ores.

2.4.4. C_9 Alcohols

For physical properties see Table 1.

C_9 Alcohol Mixtures. Nonanol mixtures are typical plasticizer alcohol components for the preparation of diisononylphthalate (DINP). Mixtures of C_9 alcohols are prepared by oxo synthesis from dimeric 1-butene and 2-butene (the mixture also is known as raffinate II) or from dimeric isobutene. Because of increased demand for isobutene for the production of methyl *tert*-butyl ether (MTBE), starting materials are now the relatively cheap mixed dimers of *n*-butenes and isobutene (codibutylenes) and dimers of *n*-butene in particular.

Typical commercial olefins of the latter variety are Dimersol from the IFP process and octenes from the UOP process (→ Hydrocarbons, Chap. 2.). Because of the differences in the manufacturing methods, these olefins have considerably different compositions, see Table 6. For specifications of the various alcohols obtained from them, see Table 7. The nonanol from diisobutene is the only one consisting primarily of 3,5,5-trimethyl-1-hexanol; the other alcohols are more or less complex alkyl-branched mixtures. Esso's isononyl alcohol, for example, contains 80 % dimethyl-1-heptanols and 20 % trimethyl-1-hexanol.

Derivatives. Isononyl alcohol is a useful high-boiling solvent for fats, oils, waxes, and resins. It is frequently used as a solvent in the paint industry. In nitrocellulose-based paints, it acts as a latent solvent; the addition of small amounts facilitates the blending with hydrocarbons. In paints based on alkyd resins, shellacs, or urea-formaldehyde resins the viscosity and the flow properties are improved; in baking enamels cratering and blistering are prevented.

As a rule, the mixtures immediately are processed further to plasticizers (for phthalates and esters of other dicarboxylic acids, → Plasticizers).

Isononyl esters of phosphorous, thiophosphorous, and dithiophosphoric acid are employed as additives for high-quality lubricants. The polyacrylates and polymethacrylates of isononyl alcohol are recommended for lowering the pour point temperatures of lubricating oils. They also are used as sealing liquids and hydraulic fluids, e.g., for vacuum and diffusion pumps.

Isononyl alcohol is the first in the series of oxo alcohols used for the industrial synthesis of surfactants. Neutralized isononyl sulfates or phosphates are wetting agents in the textile industry. Sodium diisononyl succinate is suitable as a wetting agent and as an emulsifier for various oils.

2,6-Dimethyl-4-heptanol [*108-82-7*] can be prepared by aldol condensation of acetone and subsequent hydrogenation.

$[(CH_3)_2CHCH_2]_2CHOH$

The alcohol is sold by UCC under the name diisobutylcarbinol. It is used as a reaction medium for the preparation of hydrogen peroxide and as an effective defoaming agent, in this case also in its ester form. The phosphates are extraction solvents for uranium and rare metals.

2.4.5. C_{10} Alcohols

For physical properties see Table 1.

1-Decanol [*112-30-1*], $CH_3(CH_2)_8CH_2OH$, is prepared by a number of producers, e.g., Henkel (Lorol C 10, Lorol C 10−98), by the catalytic reduction at high pressure of coconut oil or coconut fatty acids or esters. Greater, but still not significant quantities are manufactured by the Ziegler process. Companies offering the product include Conoco/Condea and the Ethyl Corp. The perfume industry uses 1-decanol as a raw material for detergents and as a defoaming agent.

C_{10} Alcohol Mixtures. Relatively complex, multi-branched mixtures of primary alcohols are manufactured almost exclusively by the hydroformylation of tripropylene. The main components of the mixture are isomeric trimethylheptanols and 3,5-dimethyloctanol. Isodecanol is sold by Hoechst, Kuhlmann, Esso, UCC, and ICI among others (Table 4).

The C_{10} alcohol mixtures are processed mainly to phthalate plasticizers (DIDP). Because of their low volatility and good low-temperature properties they are gaining steadily in importance for this application. Following

Table 6. Composition of various octenes (in wt %)

Trade name	DIB [a]	Codibutylen [a]	Dimersol [b]	Octene [c]
Butene combination	i + i	i + n	n + n	n + n
Methylheptenes	–	8	52	4
Dimethylhexenes	7	25	40	90
Trimethylpentenes	93	65	2	2

[a] Manufacturer, e.g., Erdölchemie;
[b] IFP/Nissan;
[c] UOP.

Table 7. Specifications of various isononanols

Olefin basis Tradename of the alcohol	DIB Nonanol [a]	Codibutylene	Dimersol Oxocol-900	UOP octene
Boiling range (101.3 kPa), °C	193 – 202	201.7 – 204.4	204.5 – 207.1	201.2 – 205.1
Density d_4^{20}	0.827 – 0.831	0.843	0.835 – 0.837	0.834
Refractive index n_D^{20}		1.4390	1.4365 – 1.4369	1.4365
Color no., Hazen	$\leqq 10$ [a]	–	5	–
Carbonyl no., mg KOH/g	<0.01 [b]	0.1	0.1	0.1
Acid no., mg KOH/g	$\leqq 0.001$ [c]	0.02	0.03	0.02
Water content, %	$\leqq 0.10$	–	–	0.01

[a] Method ISO R 1843;
[b] Method ISO R 1847;
[c] Method ISO R 1846.

esterification with dicarboxylic acids, the C_{10} oxo alcohol mixtures yield synthetic ester oils with outstanding properties [1]. The C_{10} alcohol mixtures are also starting materials for a wide range of derivatives, which are employed as textile auxiliaries, stabilizers, flotation agents, and antifoaming agents, and are used in detergents, pharmaceuticals, cosmetics, and pesticides (esters with 2,4-dichloro- and 2,4,5-trichlorophenoxyacetic acid are herbicides).

2.4.6. Mixtures of Linear C_{12} – C_{18} Alcohols (Detergent Alcohols)

For more details → Fatty Alcohols.

n-Alcohols with carbon chain lengths of 12 to 18, in former times prepared solely from natural products and consequently known as fatty alcohols, are now also manufactured from ethylene by the Ziegler process and by the Shell version of the oxo process (*n*-alcohol content ca. 80 %). These alcohols are important intermediates for a large number of chemical products, but over 95 % of them are used in detergents (→ Surfactants).

To a lesser extent they are used directly as wetting, emulsifying, and foaming agents. For most applications, they are modified by insertion of a hydrophilic group, e.g., by ethoxylation with ethylene oxide to ethoxides, $RO(CH_2CH_2O)_nH$, by sulfatation of the ethoxides to the ether sulfates, e.g., $RO(CH_2CH_2O)_nCH_2CH_2OSO_3Na$, and by direct sulfatation to alcohol sulfates, $ROSO_3Na$.

The products prepared by esterification of the alcohols with acrylic or methacrylic acid and subsequent polymerization are employed mainly in lubricating oils as viscosity index improvers. The C_{12} – C_{15} region is preferred, but C_{16} – C_{18} alcohols also are used.

2.4.7. C_{13} – C_{18} Isoalcohols

Isotridecyl alcohol [*27458-92-0*] is prepared from tetrapropylene by oxo synthesis. The product isolated is actually a complex mixture of primary alcohols with an average of 13 carbon atoms.

Isotridecyl alcohol is a colorless, clear, slightly viscous, neutral liquid with a mild odor. It is manufactured by Hoechst, Kuhlmann, ICI, Esso, and UCC, and consists principally of tetramethyl-1-nonanols. A compilation of the standard specifications is given in Table 4.

Because of its extremely low volatility, the phthalate has special applications as a plasticizer in wire insulation based on PVC. The alcohol also is used as a surfactant raw material, as an antifrothing agent, and as a solvent where low volatility is required, e.g., for preventing cratering and blistering during the drying of paint surfaces.

Isohexadecyl alcohol [*36311-34-9*] is prepared by aldol condensation of isooctylaldehyde and subsequent hydrogenation of the isohexadecenal. A complex mixture of primary alcohols of 2,2-dialkyl-1-ethanols is obtained, wherein the alkyl groups consist of methyl-branched C_6 and C_8 units. The alcohol is manufactured by Esso and Shell.

Isohexadecyl alcohol is used in lubricants, e.g., as the esters of higher fatty acids, dicarboxylic acids, orthophosphoric acid, or unsaturated polybasic acids. Hexadecyl alcohol is also used as a component of detergents and wetting agents, textile aids, softeners, cosmetics, evaporation preventers, and foam breakers. In pure form it is applied as a flotation and extraction agent.

Isooctadecyl alcohol [*27458-93-1*] is a highly branched, primary C_{18} alcohol with 5,7,7-trimethyl-2-(1,3,3′-trimethylbutyl)-1-octanol [*36400-98-3*] as the main component. It is prepared by aldol condensation of isononylaldehyde and subsequent hydrogenation of the unsaturated aldehyde intermediate. It is a highly viscous, neutral, odorless liquid that is practically insoluble in and hydrophobic to water. The alcohol is miscibile with most organic liquids in all proportions.

The alcohol is available from Hoechst. As a highly branched alcohol it is suitable for many applications requiring a low pour point. It is used, therefore, for the manufacture of synthetic lubricants and hydraulic fluids. Because of the low vapor pressure, it is often applied to water surfaces to prevent evaporation (by formation of monomolecular layers).

Isostearyl alcohol is available from Mitsubishi Chem. under the name Diadol 18 G It is a slightly-branched alcohol with a definite structure (2-heptylundecanol) and is prepared either by the aldol condensation of nonanal or by the hydroformylation of a C_{17}-vinylidene olefin. It is used as the waxy component for high-quality cosmetics.

2.5. Economic Aspects

The economic significance of alcohols can be seen clearly from the production figures (in 10^6 t, 1980):

Methanol	12.0
C_6-C_{11} Alcohols	2.0
Ethanol	≈ 4.0
$C_{12}-C_{15}$ Alcohols	0.6
C_3-C_7 Alcohols	2.0
$C_{16}-C_{18}$ Alcohols	0.4

Crude oil and natural gas still predominate as raw materials for their manufacture. However, the oil crisis in the 1970s induced a number of countries to develop coal and natural materials as a substitute. For example, production of ethanol by fermentation or from $C_{16}-C_{18}$ alcohols from fats and oils should increase worldwide.

As a result of the development of new competitive gasification processes, coal is expected to be used increasingly as raw material for the production of synthesis gas from which methanol and C_3-C_{15} alcohols can be made [80]. Table 8 gives a survey of the basic raw materials for higher alcohols. About 90 % of these alcohols are produced from petrochemicals.

About two thirds of the higher alcohols can be classified as plasticizers (C_6-C_{11}) and one third as detergents ($C_{12}-C_{18}$).

Plasticizer Alcohols (C_6-C_{11} and C_{13}-Branched Alcohols). Table 9 shows the 1980 consumption, actual production, and production capacities of plasticizer alcohols in the major Western industrial countries. About 75 % of the available production capacity was being utilized; 85 % of the alcohols produced were consumed in the producer countries, and about 15 % were exported, especially to East Asia. With a near 60 % share of the market, 2-ethylhexanol is the most important plasticizer alcohol. Insufficient production capacity (USA) and an unfavorable raw material supply (Japan) have led to shortages in these countries. Therefore the other C_6-C_{11} alcohols have gained in significance, particularly in the United States. An outline of

Table 8. World capacity of C_6-C_{18} alcohols, 1980 (10^3 t)

	C_6-C_{11} Alcohols	$C_{12}-C_{15}$ Alcohols	$C_{16}-C_{18}$ Alcohols	Total	%
World capacity	2000	615	410	3025	100
Petrochemical basis	1980	510	164	2654	88
Natural sources	20	105	246	371	12

Table 9. Plasticizer alcohols, consumption, production, and capacities in Western Europe, United States, and Japan, 1980 (10^3 t)

	Western Europe	United States	Japan	Total
Consumption				
2-Ethylhexanol	425	193	297	915
Other C_6-C_{11} alcohols	215	259	55	529
	640	452	352	1444
Production				
2-Ethylhexanol	602	167	250	1019
Other C_6-C_{11} alcohols	305	311	57	673
	907	478	307	1692
Capacity				
2-Ethylhexanol	808	215	333	1356
Other C_6-C_{11} alcohols	435	379	83	897
	1243	594	416	2253

the C_6-C_{11} alcohol producers in 1980 with capacity and process type is given in [81].

Table 10 contains a list of the plasticizer alcohols produced in Western Europe, the United States, and Japan in 1980, classified according to process and raw material source. Table 11 shows the relative market shares of various plasticizer alcohols and alcohol mixtures.

With plasticizer alcohols the term "linear" has a different meaning; alcohols from the oxo synthesis are customarily termed linear even if, depending on the process variant, they contain 20–50 % α-alkyl-branched isomers mixed with the *n*-alcohol components. In the strict sense of the word, therefore, only the alcohols from the Ziegler process (5 %) and on the basis of natural raw materials (1 %) can be termed "linear."

Future Aspects. The further development of plasticizer alcohols and the phthalates produced from them is closely connected with the development of PVC, as up to 50 % phthalate is mixed with the polymer. In the Western industrial countries relatively small growth rates are expected.

The following key developments will have an influence on the relative market shares of the different alcohols [1], [82]:

1) The changes in the costs of raw materials. This favors *n*-butene (dimersol) as a feedstock to the detriment of ethylene.
2) The shrinking market for linear phthalates and the trend toward "softer" vinyls favors DIDP.
3) The trend toward plasticizers of higher quality than DOP (fogging problem in automobile furnishings and vinyl roofing) will lead to a shift from 2-ethylhexanol and isooctyl alcohol toward higher alcohols.
4) The trend toward a reduction of the processing temperature in the manufacture of PVC products and toward the use of extra fillers.
5) More stringent environmental regulations during processing (volatility of plasticizers) is promoting the trend toward less volatile plasticizers.

Detergent Alcohols ($C_{12}-C_{18}$). The following figures show the relative worldwide feedstocks for $C_{12}-C_{18}$ alcohols [18]:

	$C_{12}-C_{15}$	$C_{16}-C_{18}$
Petrochemical (%)	83	40
Natural (%)	17	60

Table 10. Production of plasticizer alcohols in the United States, Europe, and Japan, 1980 (10^3 t)

	Process	United States	Western Europe	Japan	Total	%
2-Ethylhexanol	oxo synthesis/aldol cond.	167	602	250	1019	60
Branched C_6-C_{13}	oxo synthesis	186	240	41	467	28
Linear C_7-C_{11}	oxo synthesis	73	25	11	109	6
Linear C_6-C_{10}	Ziegler process	52	30	–	82	5
Linear C_6-C_{10}	from natural sources	–	≈10	5	≈15	1
		478	907	307	1692	100

Table 11. Production of plasticizer alcohols, 1980 (10^3 t), apart from 2-ethylhexanol

	United States	Europe	Japan	Total	%
Hexyl alcohol (non-Ziegler)	7	–	–	7	1
Heptyl alcohol	–	–	28	28	4
Isooctyl alcohol	32	105	–	137	22
Isononyl alcohol	66	50	3	119	19
Isodecyl alcohol	59	75	6	140	22
Linear C_6-C_{11} alcohol	125	25	16	166	26
Tridecyl alcohol	23	10	3	36	6
Total	312	265	56	633	100

Because most of the natural oils and fats (tallow) consist almost entirely of $C_{16}-C_{18}$ fatty acids, which are available in large quantities [18], [83], [84], their significance as raw materials will increase in the years to come, a trend that also will be promoted by the price increase of natural and petrochemical raw materials [80]. On the other hand, the possibilities of using natural raw materials in the $C_{12}-C_{15}$ range are limited because the corresponding lauric oils only make up about 5 – 7 % of world oil and fat production (1980: 3.3×10^6 t) [85] and are mostly used in the food industry. Therefore this range is produced primarily from petrochemical feedstocks.

In 1980, when the total production capacity for surfactant alcohols in the Western World was about 911000 t [86], only 516000 t actually were produced. For producers, process type, and installed capacities in the Western World, see [86]. The relative production figures % for the different processes are:

Oxo process	41 %
Ziegler process	21 %
Natural basis	35 %
Paraffin oxidation	3 %

Consumption figures for the surfactant alcohols and their derivatives are given in Table 12.

Future Aspects. A large growth in the production of surfactant alcohols is expected in the next few years [87] owing to the increasing demand for non-ionic detergents, which exhibit good washing properties even in low-phosphate detergent formulations (environmental regulations). In view of the limited supplies of $C_{12}-C_{15}$ raw materials and the uncertainty in the price development of lauric oils [88], there will be a continued increase of the quantities in this range produced on a petrochemical basis. The natural raw materials otherwise will increase their share in the production of $C_{15}-C_{18}$ alcohols; for corresponding data see [26].

Table 12. Consumption of detergent alcohols, 1980 (10^3 t)

	United States	Western Europe	Japan
Alcohol sulfates	72	31	9
Alcohol ethoxides	63	53	21
Alcohol ether sulfates	64	52	25
Alcohol glyceryl ether			
Sulfonates	6	–	–
Fatty amine oxides	1	–	
Methacrylate derivatives	–	18	
Use as free alcohol	18	21[a]	23[a]
Other derivatives	39		
Total	263	175	78

[a] Figures include use as free alcohol and other derivatives.

2.6. Quality Specifications

The quality control specifications for alcohols are given by DIN [89] or ASTM [90]; analytical procedures are compiled by the Deutsche Gesellschaft für Fettwissenschaft [91] and by Henkel [3]; see also [2].

The most important analytical procedure by far is *gas chromatography*. Chain-length distribution, the proportion of unsaturated, branched, or secondary alcohols, as well as small amounts of esters and hydrocarbons can be determined using suitable columns and conditions.

An important quality criterion for alcohols is the *carbonyl number,* which indicates the degree of impurity caused by carbonyl compounds. It is determined by titration of the hydrochloric acid released by reaction of the carbonyl groups with excess $NH_2OH \cdot HCl$. Its numerical value expresses the mass of CO (mg) in the form of carbonyl groups, that are present in 1 g of product.

The *hydroxyl value* (HV=mg of KOH equivalent to the hydroxyl content of 1 g of alcohol) measures the hydroxyl groups and reflects both the molecular mass and the purity of the sample.

The *saponification number* is the number of mg KOH that is required to neutralize the acids and saponify the esters contained in 1 g of substance. The *iodine value* (g of iodine consumed by 100 g of alcohol) is a measure of the content of unsaturated compounds. The *water content* is determined by the Karl Fischer method [3]. The *color* of the alcohol is usually given in Hazen units (the mass of platinum (mg) present in a solution of K_2PtCl_6 and $CoCl_2 \cdot 6\,H_2O$ in a mass ratio of 1 : 0.8025 in 1000 ml aqueous HCl that has the same color as the alcohol). Hazen numbers of 5 to 10 should not be exceeded.

A number of other tests relating to smell, melting point, boiling range, and color stability also serve as indications of quality and storage stability. One special test is, e.g., the color determination after treatment with concentrated sulfuric acid or after saponification with phthalic acid in the presence of sulfuric acid.

The recognition of the *source* of alcohols and alcohol mixtures (natural, Ziegler, oxo, paraffin oxidation, etc.) is summarized in [92]. A further valuable aid is ^{13}C NMR spectroscopy [93]. Using this technique, it is even possible to identify alcohol fractions from several different feedstock sources [3].

The *heavy-metal content* of alcohols should not exceed 1×10^{-6} g/g because of catalytic effects (reduced storage stability).

How pure an alcohol should be and what impurities it may contain usually depend on the further processing requirements and subsequent application. The specifications usually are agreed upon individually by the manufacturer and customer; therefore, alcohols of differing qualities are obtainable from the same producer.

2.7. Storage and Transportation

Short-chain liquid alcohols can be stored in containers made of noncorroding unalloyed carbon steel or stainless steel. Aluminum containers also are suitable, but not for long-term storage of lower alcohols. For higher alcohols, containers of pure aluminum and frequently of Al-Mg-Mn alloy (DIN 1725/1745) are used. Stainless steel is necessary only for the storage of extremely water-free alcohols. In aluminum vessels, corrosion is to be expected because of the formation of aluminum alkoxides. The storage temperature should be as low as possible.

In order to avoid oxidation, an inert-gas unit should be installed to enable flushing with pure nitrogen, etc. Water content of ca. 0.1 % has a

stabilizing effect; for extremely water-free alcohols the danger of autoxidation is very great.

Liquid alcohols are transported in painted, corrugated iron vessels, in road tankers, or in tank wagons made of normal steel, aluminum, or stainless steel. Solid products, such as the higher alcohols, are sold as flakes contained in multilayer polyethylene-lined paper bags.

Transportation is controlled by a number of general instructions and legal regulations for volatile and flammable substances.

The alcohols are classified on the basis of their specific properties, such as flash point, boiling point, etc.

3. Unsaturated Alcohols

Compared with saturated alcohols, unsaturated ones are of minor importance. In the United States the consumption of allyl alcohol, the most important unsaturated alcohol (→ Allyl Compounds), amounted to 60000 t in 1979. The proportion attributed to the other unsaturated alcohols, which are prepared in processes based either on acetylene or on natural feedstocks, is even smaller.

The unsaturated alcohols are generally colorless, pungent-smelling liquids and even 9-*trans*-octadecenol has a melting point of only 37 °C. The solubility in polar solvents, especially of the lower members, is high.

Some physical properties of unsaturated alcohols are given in Table 13. The accumulation of several functional groups in the molecule leads to a high reactivity. Therefore, they are capable of undergoing a large number of chemical reactions and frequently serve as intermediates in the chemical industry.

Table 13. Physical properties of unsaturated alcohols

Name		Formula	M_r	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}
Allyl alcohol	[*107-18-6*]	CH_2=CH-CH_2OH	58.08	−129	96.95	0.8520	1.4134
Crotyl alcohol, *cis*	[*4088-60-2*]	CH_3-CH=CHOH	72.11	−90.15	123.6	0.8662	1.4342
trans	[*504-61-0*]		72.11		121.2	0.8521	1.4288
Propargyl alcohol	[*107-19-7*]	CH≡C-CH_2OH	56.06	−51 to −48	114–115	0.9478	1.4306
3-Buten-2-ol	[*598-32-3*]	CH_2=CH-CH(OH)-CH_3	72.11	<−100	97.4	0.8413	1.4127
2-Butyn-1-ol	[*764-01-2*]	CH_3-C≡C-CH_2OH	70.09	−2.2	142.7–142.9	0.9373	1.4550
3-Butyn-2-ol	[*2028-63-9*]	CH≡C-CH(OH)-CH_3	70.09		107	0.8858	1.4265
10-Undecen-1-ol	[*112-43-6*]	CH_2=CH-$(CH_2)_8CH_2OH$	170.30	−2	250	0.8495	1.4506
9-Octadecen-1-ol		$CH_3(CH_2)_7CH$=CH$(CH_2)_7CH_2OH$					
cis, oleyl alcohol	[*143-28-2*]		268.49	5.5–7.5	333–335	0.8491	1.4607
trans, elaidyl alcohol	[*506-42-3*]		268.49	35	333	0.8338	1.4522
1-Hexyn-3-ol	[*105-31-7*]	CH≡C-CH(OH)$(CH_2)_2CH_3$	98.08	−80	142	0.882	1.4350
4-Ethyl-1-octyn-3-ol	[*5877-42-9*]	CH≡C-CH(OH)-CH(C_2H_5)-$(CH_2)_3CH_3$	154.14	−45	197.2	0.873	1.4502
2-Methyl-3-butyn-2-ol	[*115-19-5*]	CH≡C-C(OH)(CH_3)-CH_3	84.06	2.6	103.6	0.8672	1.4211
3-Methyl-1-pentyn-3-ol	[*77-75-8*]	CH≡C-C(CH_3)(OH)-CH_2CH_3	98.08	−30.6	121.4	0.8721	1.4318

Crotyl Alcohol [*6617-91-5*], 2-buten-1-ol, $CH_3CH{=}CHCH_2OH$, is made by aldol condensation of acetaldehyde and subsequent hydrogenation. It is of little industrial importance since butanol has been manufactured via the oxo synthesis rather than from crotonaldehyde.

Propargyl Alcohol [*107-19-7*], 2-propyn-1-ol, $HC{\equiv}C\text{-}CH_2OH$, is formed as a byproduct ($\sim$ 5 %) in the butynediol synthesis from acetylene and formaldehyde [94] ($\rightarrow$ Butanediols, Butenediol, and Butynediol):

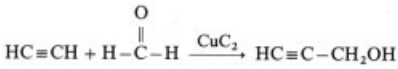

Copper acetylide serves as catalyst. The reaction can be controlled in such a way that either propargyl alcohol (high acetylene pressure, low formaldehyde concentration) or 2-butyne-1,4-diol (excess formaldehyde) is formed preferentially.

Main manufacturers using this process are BASF, GAF, Du Pont, and GAF/Chemische Werke Hüls. Total capacity for butanediol from butynediol has increased since 1975 from 70000 t/a to over 280000 t/a, so that the total capacity of propargyl alcohol is ca. 14000 t/a in 1980.

The alcohol is an excellent rust inhibitor, superior in this respect to allyl alcohol. It retains its rust-inhibiting properties at high temperatures and therefore is used for oil drilling.

Methylbutynol, 2-methyl-3-butyn-2-ol, [*115-19-5*], $HC{\equiv}C\text{-}C(CH_3)_2OH$, is formed by the ethynylation of acetone in base. Since 1977 it has been used as an intermediate for the manufacture of isoprene. Further applications are the preparation of vitamin A and other products and use in metal pickling and plating operations.

Oleyl Alcohol, *cis*-9-octadecen-1-ol, [*143-28-2*], $CH_3(CH_2)_7CH{=}CH(CH_2)_7CH_2OH$, is the most important natural-based unsaturated alcohol. All commercial fatty alcohols derived from natural sources contain oleyl alcohol to a greater or lesser extent. It is prepared by catalytic hydrogenation under pressure of the methyl esters of unsaturated fatty acids; mixed catalysts containing zinc are employed, and the reaction conditions correspond to those for the production of saturated fatty alcohols [95]. The methyl esters of the fatty acids are prepared from beef tallow and olive oil.

Oleyl alcohol is sold by Henkel under the trade name HD-Ocenol. Other manufacturers are New Japan Chemical Co., Ashland Chem. Co., and Kedzierzyn, Poland. Total capacity amounts to ca. 35000 t/a. The alcohol is used principally in the detergent field, where the double bond offers possibilities for applications not covered by the saturated fatty alcohols.

4. Alkoxides

Metal alkoxides can be formally regarded as salts of very weak acids (J. Liebig, 1837).

The alkoxides of alkali metals, magnesium, aluminum, titanium, zirconium, and antimony are commercially the most important ones, e.g., as catalysts in condensation and transesterification reactions, as reducing agents, and as paint additives.

4.1. Properties

The properties of the alkoxides depend on the position of the metal in the periodic table and on the character of the alcohol component. The alkoxides of the highly electropositive metals are solid, nonvolatile, very basic compounds of ionic nature. Polymeric, covalent compounds with a low volatility are formed by the higher multivalent elements; the alkoxides of the lighter transition elements are distillable, monomeric, covalent liquids. The alkali and alkaline-earth metal alkoxides are soluble in organic solvents to only a limited extent owing to their polar nature. Alkali-metal alkoxides crystallize from alcohol solution mostly with 1 – 3 mol alcohol (crystal alcohol), which can be removed by raising the temperature or by azeotropic distillation with benzene or toluene [96]. Heating to above 200 °C generally leads to decomposition [96]. The lower alkoxides of aluminum (with the exception of methoxide) and titanium can be distilled in the vacuum without decomposition. The physical properties of metal alkoxides have been discussed in detail, cf. [96–98].

Alkoxides are very reactive toward compounds having acidic hydrogen atoms, for example, with water:

$$ROM + H_2O \rightleftharpoons ROH + MOH$$

Therefore, alkoxides are stable only in the absence of water. With highly electropositive metals the reaction is reversible so that, e.g., the sodium ethoxide can be obtained from sodium hydroxide and ethanol. With the covalent multivalent alkoxides, however, this reaction is irreversible. The hydrolysis of titanium and aluminum alkoxides has been examined thoroughly [98]. With other alcohols transalcoholization occurs until equilibrium is established:

$$ROM + R'OH \rightleftharpoons R'OM + ROH$$

In this way the alkoxides of higher alcohols can be obtained from lower alkoxides. Reaction with phenol leads to the phenoxides:

$$M(OR)_n + n\,C_6H_5OH \longrightarrow M(OC_6H_5)_n + n\,ROH$$

By reaction with carboxylic acids, the metal salts of the carboxylic acids and the free alcohols are formed:

$$M(OR)_n + n\,R'COOH \longrightarrow M(R'COO)_n + n\,ROH$$

Reaction of trialkyl silanols with alkoxides gives the corresponding trialkylsiloxy metal derivatives [99]:

$$M(OR)_n + n\,R'_3SiOH \longrightarrow M(OSiR'_3)_n + n\,ROH$$

β-Dicarbonyl compounds form chelate complexes with alkoxides.

The alkoxides react with carbonyl compounds in different ways depending on their basicity. With alkali-metal alkoxides enolization or condensation reactions occur. The Tishchenko reaction with aluminum alcoholates leads to esters by disproportionation [100]:

$$2\,RCHO \xrightarrow{Al(OR)_3} RCOOCH_2R$$

In the Meerwein–Ponndorf–Verley reaction, ketones are reduced with alcohols in the presence of aluminum triisopropoxide:

$$R\text{–}CO\text{–}R + R'_2CHOH \rightleftharpoons R_2CHOH + R'\text{–}CO\text{–}R'$$

Transesterification reactions of esters with alkoxides also are known [101]:

$$M(OR)_n + n\,R''COOR' \rightleftharpoons M(OR')_n + n\,R''COOR$$

Other alkoxides catalyze transesterification reactions; e.g., benzoic acid butyl ester and ethanol are obtained from benzoic acid ethyl ester and butanol in the presence of titanium isopropoxide [*546-68-9*] [102]. A detailed description of reactions and applications of alkali and alkaline-earth metal alkoxides has been published [103].

4.2. Preparation

Alkoxides may be obtained by the reaction of alcohols with metals, metal hydroxides, metal halides, or alkoxides of other alcohols; the method depends on the metal and the alcohol component. A summary of the procedures can be found in [96].

Preparation from Metals and Alcohols.

$$M + n\,ROH \longrightarrow M(OR)_n + n/2\,H_2$$

This reaction is the simplest method for the preparation of alkoxides; it is limited to the alkali metals and to magnesium and aluminum. The alkali metals are suspended in an inert solvent [104] or are used as an amalgam [105]. The amalgam is brought into contact with the alcohol in a packed column according to the countercurrent principle. A graphite or Fe–Cr–C alloy contact electrode that cannot be amalgamated is used to speed up the reaction [106]. The longer the alcohol chain, the less vigorous the reaction. Primary alcohols react fastest, tertiary slowest. Magnesium and aluminum must be activated by the addition of catalysts, e.g., mercury chloride, HCl, or iodine [107]. Other catalysts, especially for magnesium, are *p*-toluenesulfonic acid and orthoformic acid esters [108]. By heating the alcohols in a column filled with aluminum turnings, the corresponding aluminum alkoxides can be produced without the aid of a catalyst [109].

Reaction of Alcohols with Metal Hydroxides. The equilibrium reaction of alcohols with metal hydroxides can be driven to completion by azeotropic distillation of the water formed with benzene, toluene, or the alcohol itself [110].

$$n\,ROH + M(OH)_n \rightleftharpoons M(OR)_n + n\,H_2O$$

The method is suitable only for the highly electropositive elements; lithium hydroxide does not react [111].

Reaction of Alcohols with Metal Halides. For the alkoxides of the less electropositive elements the corresponding metal chlorides are reacted with the alcohols:

$$MCl_n + n\,ROH \rightleftharpoons M(OR)_n + n\,HCl$$

The process is primarily suitable for the production of alkoxides of multivalent metals, such as titanium, germanium, and antimony. Mixtures with chloroalkoxides always are obtained. Bases, in particular ammonia, are used to shift the equilibrium toward the formation of pure alkoxides:

$$TiCl_4 + 2\,ROH \rightleftharpoons TiCl_2(OR)_2 + 2\,HCl$$

$$TiCl_2(OR)_2 + 2\,ROH + 2\,NH_3 \rightleftharpoons Ti(OR)_4 + 2\,NH_4Cl$$

Reaction of Metal Alkoxides with Metal Halides and Other Metal Compounds. In particular the alkoxides of multivalent metals can be prepared from their halides (MX_n) by reaction with other alkoxides (preferably of monovalent metals):

$$MX_n + n\,MOR \rightleftharpoons M(OR)_n + n\,MX$$

Nearly all metal alkoxides can be manufactured by this process. Sometimes other metal compounds are employed instead of halides; see [103] for details.

Reaction with Other Alkoxides. The alkoxides of higher or multivalent alcohols also can be prepared by alcohol transfer:

$$n\,R'OH + M(OR)_n \rightleftharpoons (R'O)_nM + n\,ROH$$

The more volatile alcohol distills from the reaction mixture [112]. However, when a large excess of the lower alcohol is used, a lower alkoxide can be obtained from a higher one.

4.3. Uses

Alkoxides of alkali metals are used predominantly as condensation agents in organic synthesis and to introduce the alkoxy group into other compounds. The methoxides and ethoxides of sodium and potassium and the potassium *tert*-butoxide have attained particular significance. For the use of sodium ethoxides, see [103], [113], [114]. Magnesium alkoxides are used as raw materials for driers in paints and varnishes and as components of highly active Ziegler catalysts [115]. Aluminum isopropoxide and aluminum *sec*-butoxide are used commercially as specific reducing agents for various organic compounds (Meerwein-Ponndorf-Verley reaction,see Section 4.1). They also find application in paints to improve resistance to chemicals [116], in textiles to produce impregnating agents, and in cosmetics as components of antiperspirants.

Titanium alkoxides are used in the manufacture of corrosion-proof, high-temperature paints, as esterification catalysts, for the water proofing of textiles, as bonding agents in adhesives, and as catalyst components for the polymerization of olefins.

Besides the solid alkoxides, solutions in the corresponding alcohols are on the market. Alkoxides are very easily hydrolyzed by water, extremely easily in the case of the short-chain alkoxides. Contact with atmospheric moisture must be avoided because spontaneous ignition may occur. Alkoxide fires must be extinguished with dry sand or powder but never with water.

Solid alkoxides are sold mostly in polyethylene bags packed in drums or in corrugated barrels fitted with special lids. Alkoxide solutions, e.g., the 25 % solution of sodium methoxide in methanol, however, also are transported in tankers. During unloading, a dry, inert-gas atmosphere must be provided.

Compared with alcohols, the alkoxides are of minor commercial importance. The 1982 world production is estimated at 10000–12000 t.

5. Toxicology

1-Hexanol. The oral LD_{50} is 4.87 g/kg for rats [117] and 4.0 g/kg for mice [118]. The maximum nonlethal period of time that rats can inhale a saturated atmosphere is 8 h [119].

1-Hexanol is a severe irritant to the rabbit eye [119]. It is moderately irritating to the skin of rabbits after 24-h skin contact [120]. The dermal LD_{50} is 2.53 g/kg, an indication of considerable dermal absorption [119]. 1-Hexanol is not a human skin irritant according to the Epstein test [121]. Also, it does not act as a skin sensitizer [122].

2-Methyl-1-pentanol. The oral LD_{50} for rats is reported to be 1.41 g/kg. For rabbits the dermal LD_{50} is 3.56 g/kg. In the open-patch test, 2-methyl-1-pentanol caused mild skin irritation in rabbits. In humans, inhalation of 50 ppm irritates mucous membranes [123].

4-Methyl-2-pentanol. The oral LD_{50} for rats is 2.59 g/kg [119]. Symptoms of acute intoxication are anesthesia, gastrointestinal irritation, and congestion of the mesenteric blood vessels [119].

In the rabbit eye, 4-methyl-2-pentanol was found to be a mild irritant [119]; after prolonged exposure corneal lesions can occur [124]. The skin irritation is reported to be mild [79], but absorption is considerable. Repeated exposure causes severe drying of the skin due to degreasing [125].

Mice that had inhaled concentrations of 4600 ppm for 1 min showed irritated respiratory passages, drowsiness, and anesthesia; after 10 min, death occurred occasionally, and after 15 min, 8 out of 10 mice had died [125]. For humans, eye irritation is reported after an exposure to 50 ppm for 15 min; more than 50 ppm causes irritation of nose and throat. Exposure to 25 ppm does not cause symptoms to appear [126]; therefore, the TLV and MAK values are set at 25 ppm.

2-Octanol. The oral LD_{50} is 3.2 g/kg for rats [127] and 4.0 g/kg for mice [118]. Rabbit eyes develop a pronounced redness on contact with 2-octanol [118]. In guinea pigs, 2-octanol is a slight skin irritant [128]. Rats were exposed for 4.5 months to a maximum inhalation concentration of 56.3 ppm for 2 h/day, 6 days/week. Hematological changes (anemia) and changes in liver metabolism, kidneys, and myocard were observed [128].

C_9 Alcohol Mixtures. Oral LD_{50} for rats range from 2.98 g/kg [129] to 6.4 g/kg [130]. Percutaneous LD_{50} fall in a similar range [131]. Inhalation of 21.7 mg/L for 4 h causes death in rats [129]. Eye contact causes conjunctivitis and keratitis in rabbits. Inhalation of 33, 99, and 136 ppm by rabbits for 2 months caused degenerations of glial cells in the cerebral cortex and subcortex [132].

C_{10} Alcohol Mixtures (Decyl Alcohols). LD_{50} values differ considerably because of the variable composition of the mixtures. After oral administration to rats the values range from 4.72 g/kg [129] to 25.6 g/kg [130]. Dermal LD_{50} for rabbits are reported to be between 3.56 mL/kg and 18.8 mL/kg [133]. The inhalative LD_{50} for mice is 525 ppm [132].

Eye contact in rabbits causes severe irritative conjunctivitis with corneal injuries [129]. Skin contact in guinea pigs [130] and rabbits [129] for 24 h causes moderate to severe skin irritation.

In a 60-week skin painting study in mice, *n*-decanol showed tumor promoting activities in the presence of dimethylbenzanthracene initiator; however, these concentrations of *n*-decanol alone irritated the skin considerably [134].

6. References

General References

1. E. J. Wickson: "Monohydric Alcohols," *ACS Symp. Ser.* 1981, no. 159 (March 25 – 26, 1980).
2. J. A. Monick: *Alcohols, Their Chemistry, Properties and Manufacture,* Reinhold Publ. Co., New York 1968.
3. *Fettalkohole, Rohstoffe, Verfahren, Verwendung,* 2nd ed., Henkel, Düsseldorf 1982.

Specific References

4. J. Falbe, B. Cornils, *Fortschr. Chem. Forsch.* **11** (1968) 101.
5. L. H. Horsley: *Azeotropic Data,* Am. Chem. Soc., Washington, D.C., 1952.
6. R. W. Gallant, *Hydrocarbon Process.* **45** (1966) no. 10, 171.
7. A. M. P. Tans, *Hydrocarbon Process.* **48** (1969) no. 5, 168.
8. I. Mellan: *Industrial Solvents Handbook,* Noyes Data Corp., Park Ridge 1970, p. 123.
9. *Lösemittel Hoechst, ein Handbuch für Laboratorium und Betrieb,* 5th ed., Hoechst, Frankfurt 1974.
10. B. Cornils in J. Falbe (ed.): *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart 1975.
11. *Ullmann,* 3rd. ed., **vol. 9,** p. 740, **vol. 12,**p. 410.

12. B. Cornils, W. Rottig in J. Falbe (ed.): *Chemierohstoffe aus Kohle,* Thieme Verlag, Stuttgart 1977, p. 323.
13. Union Carbide, US 4 162 262, 1976 (P. C. Ellgen, M. Bhasin). Hoechst, DE 2 925 571, 1979 (H. Hachenberg, F. Wunder, E. Leopold, H.-J. Schmidt).
14. Institut Français du Pétrole, US 4 122 110, 1978 (A. Sugier, E. Freund). *Eur. Chem. News* 1983, no. 12, 25.
15. M. E. Dry, *CHEMTECH* **12** (1982) 744.
16. A. Lundeen, R. Poe, *Encycl. Chem. Process. Des.* 1977, no. 2, 465.
17. R. E. Vincent: "Higher Linear Oxo Alcohols Manufacture," paper presented at ACS Ind. and Eng. Chem. Div. Symposium, Houston, Texas, March 25, 1980.
18. P. Hofmann, *Fette Seifen Anstrichm.* **85** (1983) no. 3, 127.
19. M. Sherwood, *Chem. Ind. (London)* 1982, 994.
20. C. L. Thomas: *Catalytic Processes and Proven Catalysts,* Academic Press, New York-London 1970.
21. J. Falbe: *New Syntheses with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980.
22. Shell, US 3 239 569, 1969 (L. H. Slaugh, P. Hill, R. D. Mullineaux).
23. Chem. Werke Hüls, DE 1 643 856, 1968 (M. Reich).
24. BASF, DE 1 269 605, 1966 (K. Adam, E. Haarer).
25. BASF, DE 1 277 232, 1967 (H. Corr, E. Haarer, H. Hoffmann, S. Winderl). Shell, US 3 278 612, 1966 (C. R. Greene).
26. *Hydrocarbon Process.* **60** (1981) no. 11, 165. National Distillers, US 3 288 866, 1966.
27. H. Bortsch, H. Reinheckel, K. Haage, *Fette Seifen Anstrichm.* **71** (1969) 357, 785, 851.
28. K. Ziegler et al., *Justus Liebigs Ann. Chem.* **629** (1960) 1.
29. *Ullmann,* 4th ed., **vol. 7,** p. 211, **vol. 11,** p. 436.
30. N. J. Stevens, J. R. Livingston, *Chem. Eng. Prog.* **64** (1968) no. 7, 61.
31. *Hydrocarbon Process.* **60** (1981) no. 11, 171; **57** (1978) no. 1, 145.
32. R. Landan, D. Brown, J. L. Russel, *Prepr. Div. Pet. Chem. Am. Chem. Soc.* **15** (1970) 63 – 67.
33. K. Lindner, R. Roland in K. Lindner (ed.): *Tenside, Textilhilfsmittel, Waschrohstoffe,* 2nd ed., **vol. 1,** Wissenschaftliche Verlagsgesellschaft, Stuttgart 1964, p. 344. B. W. Werdelmann, *Fette Seifen Anstrichm.* **76** (1974) 1.
34. I. M. Towbin, D. M. Boljanowskii, *Maslo. Zhir. Promst.* **32** (1966) 29.
35. F. Püschel, *Tenside* **3** (1966) 71.
36. F. Püschel, *Tenside Deterg.* **11** (1971) 147.
37. F. Broich, H. Grasemann, *Erdöl Kohle Erdgas Petrochem.* **18** (1965) 250.
38. D. Mansuy, J. F. Bartoli, J. C. Chottard, M. Lange, *Angew. Chem.* **92** (1980) 938; *Angew. Chem. Int. Ed. Engl.* **19** (1980) 909.
39. D. Osteroth, *Seifen Öle Fette Wachse* **95** (1969) no. 19, 659.
40. H. Bahrmann, W. Lipps, B. Cornils, *Chem. Ztg.* **106** (1982) 249.
41. M. E. Fakey, R. A. Head, *Appl. Catal.* **5** (1983) 3.
42. M. Röper, H. Loevenich, J. Korff, *J. Mol. Catal.* **8** (1982) no. 17, 315.
43. H. Bahrmann, B. Cornils, *Chem. Ztg.* **104** (1980) 39.
44. BASF, DE 843 876, 867 849, 875 346, 1941 bis 1943 (G. Wietzel, A. Scheuermann).
45. J. Wender, M. Orchin, et al., *J. Am. Chem. Soc.* **71** (1949) 4160; **73** (1951) 2656; Science (Washington D.C.) **113** (1951) 206.
46. H. Pichler: *Advances in Catalysis,* **vol. 4,** Academic Press, New York 1952.
47. *CEER Chem. Econ. Eng. Rev.* **9** (1976) 58.
48. J. F. Knifton, *J. Am. Oil. Chem. Soc.* **55** (1978) 496. P. Hofmann, W. H. E. Müller, *Hydrocarbon Process.* **60** (1981) no. 10, 151. P. Hofmann, *Chem. Ztg.* **105** (1981) 311. P. Hofmann, *Fette Seifen Anstrichm.* **85** (1983) no. 3, 126.
49. K. Esser, U. Schmidt, *Process Biochem.* **17** (1982) 46. B. Maiorella, C. R. Wilke, *Adv. Biochem. Eng.* **20** (1981) 43. J. Tremolieres: *Alcohols and Derivatives,* **vol. 2,** Pergamon Press, Oxford-London 1970, p. 81.
50. J. J. H. Hastings, *Econ. Microbiol.* **2** (1978) 31. W. L. Faith, D. B. Keyes, R. L. Clark: *Industrial Chemicals,* J. Wiley & Sons, New York 1965, p. 181.
51. *Ullmann,* 4th ed., **vol. 7,** p. 531.
52. M. Guerbet, *C. R. Hebd. Seances Acad. Sci* **. 128** (1899) no. 51, 1002. S. Veibel, J. I. Nielsen, *Tetrahedron* **23** (1967) 1723. H. Krauch, W. Kunz: *Reaktionen der organischen Chemie,* 4th ed., Hüthig Verlag, Heidelberg 1961, p. 271.
53. E. Supp, *Hydrocarbon Process.* **60** (1981) no. 3, 71.
54. H. Machemer, *Angew. Chem.* **64** (1952) 213.
55. Henkel, DE 2 703 746, 1978 (R. Karl).

56. C. Weizmann, E. Bergmann, M. Salzbacher, *J. Org. Chem.* **15** (1950) 54. R. E. Miller, G. E. Bennett, *Ind. Eng. Chem.* **63** (1961) 33.
57. *Seifen Öle Fette Wachse* **97** (1971) 790.
58. *Technical Service Report: Organische Produkte CPI,* Henkel, Düsseldorf 1980, p. 28.
59. W. Straßberger, *Fette Seifen Anstrichm.* **71** (1969) 215.
60. W. Stein, H. Rutzen, Paper presented at DGF-Tagung in Mainz (1967). Mainz (1967). Henkel Referate **4** (1969) 18.
61. M. S. Newmann, G. Underwood, M. Renoll, *J. Am. Chem. Soc.* **71** (1949) 3362.
62. Henkel, DE 1 139 477, 1963 (W. Stein, H. Rutzen).
63. E. K. Euranto in S. Patai (ed.): *Chemistry of Carboxylic Acid Esters,* Interscience, New York 1969, p. 508.
64. *Technical Service Report on Alcohols,* Esso Chemie, Hamburg 1983.
65. *Technical Service Report on Alfol-Alcohols,* Condea Chemie, Brunsbüttel 1983.
66. *Technical Service Report on Alcohols,* Deutsche ICI, Frankfurt/Main 1983.
67. *Technical Service Report on Fatty Alcohols,* Henkel, Düsseldorf 1983.
68. *Technical Service Reports on Oxoalcohols:* n-Propanol, n- and iso-Butanol, Amylalcohol, Isooctanol, 2-Ethylhexanol, Isononanol, Isodecanol, Isotridecylalcohol, Isooctadecylalcohol, Isohexadecylalcohol, 2-Ethylbutanol, Hoechst, Frankfurt/Main 1974.
69. *Technical Service Report on Alcohols,* Union Carbide Corp., New York 1983.
70. *Oxo Synthesis Products from Ruhrchemie,* 2nd ed., Hoechst, Frankfurt/Main 1971.
71. G. I. Mathier, R. W. Bruce, *Can. Min. J.* **95** (1974) no. 6, 75.
72. Cotton Inc., US 3 778 509, 1973 (H. L. Lewis, E. L. Frick, R. T. Burchill).
73. *Technical Service Report on Alcohols,* Ethyl Internat., Baton Rouge 1976.
74. *Technical Service Report on Oxo Alcohols,* Monsanto (FRG), Düsseldorf 1983.
75. *Shell-Industrie-Chemikalien,* Deutsche Shell Chemie, Frankfurt/Main 1978.
76. *Technical Service Report on Alcohols,* Ugine Kuhlmann, Paris-Cedex 1983.
77. H. Schubert: *Aufbereitung fester mineralischer Rohstoffe,* **vol. 2,** VEB-Deutscher Verlag für Grundstoffindustrie, Leipzig 1967.
78. *Mod. Plast.* **9** (1981) 47.
79. Emery, US 3 438 765, 1969.
80. F. Asinger, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) no. 1, 28.
81. *Chemical Economic Handbook:* "Oxo Chemicals," SRI International, Menlo Park, Calif., 1981.
82. *Chem. Ind. (Düsseldorf)* **34** (1983) 3.
83. M. Buchold, *Chem. Eng.* **90** (Feb. 21, 1983) 42.
84. R. Leysen, *Chem. Ind. (London)* 1982, 428.
85. *Seifen Öle Fette Wachse* **107** (1981) no. 6, 140.
86. *Surfactants and Detergent Raw Materials,* Process Economics Program, SRI International, Menlo Park, Calif., August 1981.
87. D. E. Haupt, P. B. Schwin, *J. Am. Oil Chem. Soc.* **55** (1978) 28. P. L. Layman, *Chem. Eng. News* **60** (1982) no. 1, 13.
88. W. Stein, *Fette Seifen Anstrichm.* **84** (1982) no. 2, 45.
89. *Mineral- und Brennstoffnormen,* DIN-Taschenbuch 32, Beuth-Vertrieb, Berlin-Köln-Frankfurt 1972.
90. *Annual Book ASTM Standards,* part 15, 23, 24, 25, 27, 29 (1978).
91. *Deutsche Einheitsmethoden zur Untersuchung von Fetten, Fettprodukten und verwandten Stoffen,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1950 – 1979.
92. *Ullmann,* 4th ed., **vol. 11,** p. 439.
93. E. Breitmeier, W. Voelter: ^{13}*C-NMR-Spectroscopy,* Verlag Chemie, Weinheim 1974, p. 138.
94. GAF, GB 968 928, 1962 (M. E. Chiddix, O. F. Hecht). BASF, DE 1 284 964, 1967 (D. Neubauer, N. v. Kutepow). Cumberland Chemical Corp., US 3 257 465, 1966 (M. W. Leeds, H. L. Komarowski).
95. Henkel, US 3 729 520, 1973 (H. Rutzen, W. Rittmeister).
96. Houben-Weyl: *Methoden der organischen Chemie, Sauerstoffverbindungen I,* **vol. 6,** part 2, Thieme Verlag, Stuttgart 1963, p. 5 – 70.
97. O. C. Dermer, *Chem. Rev.* **14** (1934) 408.
98. D. C. Bradley, F. A. Cotton, *Prog. Inorg. Chem.* **2** (1960) 303.
99. D. C. Bradley, J. M. Thomas, *J. Chem. Soc.* 1959, 3404.
100. H. Krauch, W. Kunz: *Reaktionen der organischen Chemie,* 3rd ed., Hüthig Verlag, Heidelberg 1966, p. 592.
101. R. C. Mehrota, *J. Am. Chem. Soc.* **76** (1954) 2266.
102. Du Pont, US 2 822 348, 1958 (J. H. Haslam).
103. *Alkali and Alkali Earth Metal Alcoholates-Reactions and Applications,* Dynamit Nobel, Troisdorf-Oberlar.

104. Ethyl Corp., FR 1 070 601, 1952 (T. P. Whaley).
105. Wacker Chemie, DE 928 467, 1953 (A. Gerber, O. Leschhorn, H. Müller, J. Rambausek).
106. Mathieson Alkali Works, US 2 069 403, 1936 (G. L. Cunningham).
107. Armour & Co., GB 840 499, 1960 (S. O. Grimsdick).
108. Dynamit Nobel, BE 741 121, 1968 (M. M. A. Lenz, O. Bleh, E. Termin).
109. Stauffer-Chemical, GB 954 892, 1960 (W. E. Smith, A. R. Anderson).
110. Chemische Werke Hüls, DE 968 903, 1952 (A. Coenen).
111. Dow Chemical, US 2 796 443, 1956 (R. H. Meyer, A. K. Johnson).
112. Continental Oil, GB 874 585, 1960.
113. G. D. Byrkit, E. C. Soule, *Chem. Eng. News* **22** (1944) 1903.
114. M. Sittig: *Sodium, Its Manufacture, Properties and Uses,* Reinhold Publ., New York 1956, p. 260.
115. Hoechst, DE 2 021 831, 1970 (B. Dietrich).
116. F. Schlenker, *Kunststoffe* **47** (1957) 7.
117. E. Bar, F. Griepentrog, *Med. Ernähr.* **8** (1967) 244.
118. G. N. Zaeva, V. J. Fedorova, *Toksikol. Nov. Prom. Khim. Veshchestu.* **5** (1963) 51.
119. H. F. Smyth, C. P. Carpenter, C. S. Weil, *AMA Arch. Ind. Hyg. Occup. Med.* **4** (1951) 119.
120. D. L. Opdyke, *J. Food Cosmet. Toxicol.* **13** (1975) 695.
121. A. M. Kligman, *J. Invest. Dermatol.* **47** (1966) 393.
122. W. L. Epstein, *Report to RIFM,* August 27, 1974.
123. *Registry of Toxic Effects of Chemical Substances,* **vol. II,** U.S. Dept. of Health and Human Services, 1979.1960. Center for Disease Control, NIOSH, Cincinnati, Ohio 45226, Sept. 1980, p. 205.
124. Shell Chemical Corporation, Ind. Hyg. Bull. 1957.
125. W. A. McOmic, H. H. Anderson, *Univ. Calif. Berkeley Publ. Pharmacol.* **2** (1949) 217.
126. L. Silverman, H. F. Schulte, M. W. First, *J. Ind. Hyg. Toxicol.* **28** (1946) 262.
127. D. W. Fassett in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2,** Wiley-Interscience, New York 1963.
128. E. Browning: *Toxicity of Organic Solvents,* Med. Res. Council, Ind. Health Research Report no. 80, London 1937.
129. R. A. Scala, E. G. Burtis, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 493.
130. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2A,** Wiley-Interscience, New York 1981, p. 4622.
131. H. F. Smith, Jr., C. P. Carpenter, C. S. Weil, *J. Ind. Hyg. Toxicol.* **31** (1949) 60.
132. Y. L. Egorov, L. A. Andrianov, *Uch. Zap. Mosk. Nauchno. Issled. Inst. Gig. im. F. F. Erismana* **9** (1961) 47.
133. Y. L. Egorov, *Toksikol. Gig. Prod. Neftekhim. Neftekhim. Proizvod. Vses. Konf. Dokl 1 2nd* 1972, 98.
134. J. Sice, *Toxicol. Appl. Pharmacol.* **9** (1966) 70.

Alcohols, Polyhydric

→ Ethylene Glycol, → Propanediols, → Butanediols, Butenediol, and Butynediol, → Glycerol, and → Sugar Alcohols are separate keywords.

PETER WERLE, Degussa AG, Hanau-Wolfgang, Federal Republic of Germany (Section 2.9, Chaps. 3, 4, 5, 6)

MARCUS MORAWIETZ, Degussa AG, Hanau-Wolfgang, Federal Republic of Germany (Chap. 2, Sections 2.1 – 2.8)

1. General Aspects 1
2. Diols 4
2.1. 1,5-Pentanediol 4
2.2. 1,3-Propanediols 4
2.2.1. 2,2-Dimethyl-1,3-propanediol (Neopentyl Glycol) 4
2.2.2. Hydroxypivalic Acid Neopentyl Glycol Ester 6
2.2.3. 2-Methyl-2-propyl-1,3-propanediol and 2-Butyl-2-ethyl-1,3-propanediol . 7
2.2.4. 2-*sec*-Butyl-2-methyl-1,3-propanediol 7
2.2.5. 1,3-Propanediol 7
2.2.6. 2-Methyl-1,3-propanediol 7
2.3. 1,6-Hexanediol 8
2.4. Hexynediols 8
2.4.1. 3-Hexyne-2,5-diol 8
2.4.2. 2,5-Dimethyl-3-hexyne-2,5-diol . . . 9
2.5. 1,10-Decanediol 9
2.6. 2,2-Bis(4-hydroxycyclohexyl)propane 9
2.7. 1,4-Bis(hydroxymethyl)cyclohexane 9
2.8. 2,2,4-Trimethyl-1,3-pentanediol . . 11
2.9. Vicinal Diols by Hydroxylation of Olefins with Peracids 11
2.9.1. 1,2-Pentanediol 11
2.9.2. Other 1,2-Diols 11
3. Triols 12
3.1. Trimethylolpropane 12
3.2. Trimethylolethane 12
4. Tetrols 13
4.1. Ditrimethylolpropane 13
4.2. Pentaerythritol 13
5. Higher Polyols 15
5.1. Dipentaerythritol 15
5.2. Tripentaerythritol 15
6. Toxicology 15
7. References 16

1. General Aspects

The properties common to all polyhydric alcohols are determined by the hydroxyl groups: hydrogen bonding leads to high boiling points, high viscosity, and solvency for polar substances. Other important characteristics are the structure of the alcohols and the accessibility of the hydroxyl groups (primary, secondary, or tertiary), which influence the chemical behavior and the thermal stability of the hydroxyl functions. Alcohol groups next to a *neo* structure (e.g., pentaerythritol or neopentyl glycol) exhibit higher resistance towards elimination and degradation because of the lack of β-hydrogen atoms. Some of the important physical properties of the alcohols discussed in this article are given in Table 1.

Important reactions of polyhydric alcohols are those with isocyanates to form urethanes, with acids and acid anhydrides to form esters, and with aldehydes or ketones to form acetals or ketals.

Polyhydric alcohols, especially triols and tetrols, are normally regarded as polyols, but in the literature poly(ether alcohol)s such as poly(ethylene glycol)s and/or poly(ester alcohol)s are sometimes also classified as polyols.

Production. Each class of polyhydric alcohol has its own basic production method and raw materials. The most important synthetic strategies are presented in Scheme 1.

The reactions can be classified into the following general categories:

a) Epoxidation of alkenes followed by acid-catalyzed cleavage of the epoxide ring to obtain the 1,2-diol.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_305

Table 1. Physical properties of polyhydric alcohols

Compound	M_r Molecular formula	ϱ, g/cm^3 (°C)	*bp*, °C (kPa)	*mp*, °C	Dynamic viscosity, mPa · s (°C)	Flash point, °C
1,5-Pentanediol	104.15 $C_5H_{12}O_2$	0.992 (20)	240 – 244 (101.3)	ca. – 16	128 (20) 48 (40)	136 [a]
2,2-Dimethyl-1,3-propanediol (Neopentyl glycol)	104.15 $C_5H_{12}O_2$	1.060 (20)	209 – 210 (101.3)	130	174 (25)	103 [b] 151.6 [c]
Hydroxypivalic acid neopentyl glycol ester	204.26 $C_{10}H_{20}O_4$	1.019 (20) 1.000 (85)	292 (decomp., 103.3) 163 (1)	50 – 51	70 (60)	161
2-Methyl-2-propyl-1,3-propanediol	132.20 $C_7H_{16}O_2$	0.913 (60)	230 (102.3) 111 – 113 (0.5)	56 – 58		130
2-Butyl-2-ethyl-1,3-propanediol	160.20 $C_9H_{20}O_2$	0.9266 (50)	132 – 133 (0.7)	40 – 43	8.5 (100)	144
2-*sec*-Butyl-2-methyl-1,3-propanediol	146.22 $C_8H_{18}O_2$		92 – 97 (0.1)	52.5 (decomp.)		
2-Methyl-1,3-propanediol	90.12 $C_4H_{10}O_2$		213 (101.3)	– 91	68 (25)	> 230
1,6-Hexanediol	118.18 $C_6H_{14}O_2$	0.965 (50) 0.927 (104)	252 (101.3)	40 – 42	46.86 (48.6) 6.901 (104)	147 [a]
3-Hexyne-2,5-diol	114.15 $C_6H_{10}O_2$	1.023 (13)	ca. 100 (0.1) 113 – 114 (2)	40 – 70	6 (90)	
2,5-Dimethyl-3-hexyne-2,5-diol	142.20 $C_8H_{14}O_2$			96 – 97		
1,10-Decanediol	174.29 $C_{10}H_{22}O_2$	0.89 (80)	160 (0.65)	71 – 73	15.8 (80)	152 [a]
2,2-Bis(4-hydroxycyclohexyl)-propane [d]	240.37 $C_{15}H_{28}O_2$	0.958 (155)	235 – 245 (2.7)	125 – 163	114 (152)	152 [a]
1,4-Bis(hydroxymethyl)cyclohexane	144.24 $C_8H_{16}O_2$	1.041 (20) 0.994 (100)	286 (100) 163.9 (1.33)	31.5 [e]	218 (70) 40.2 (100)	162
2,2,4-Trimethyl-1,3-pentanediol	146.23 $C_8H_{18}O_2$	0.897	215 – 235 (101.3)	46 – 55	27 (50)	110 [f]
1,2-Pentanediol	104.15 $C_5H_{12}O_2$	0.980 (20)	210 (101.3)			105
2,3-Dimethyl-2,3-butanediol (Pinacol)	118.18 $C_6H_{14}O_2$ $C_4H_{10}O_3$	0.967 (20)	174 (101.3)	42		75
Trimethylolpropane	134.18 $C_6H_{14}O_2$	1.084 (20)	285 (101.3)	58		
Trimethylolethane	120.15 $C_5H_{12}O_2$	1.210 (20)	283 (101.3)	202		
Ditrimethylolpropane	250.34 $C_{12}H_{26}O_5$			112 – 114		
Pentaerythritol	136.15 $C_5H_{12}O_4$	1.396 (20)	276 (4)	260 – 262		240 [a]
Dipentaerythritol	254.28 $C_{10}H_{22}O_7$	1.365 (20)		222		
Tripentaerythritol	372.42 $C_{15}H_{32}O_{10}$	1.300 (20)		248		

[a] Pensky-Marten (closed cup).
[b] BASF.
[c] Eastman.
[d] Isomer distribution: *cis*/*cis* 3 – 3.5 %; *cis*/*trans* 30 – 35 %; *trans*/*trans* 60 – 65 %.
[e] Softening point.
[f] Cleveland (open cup).

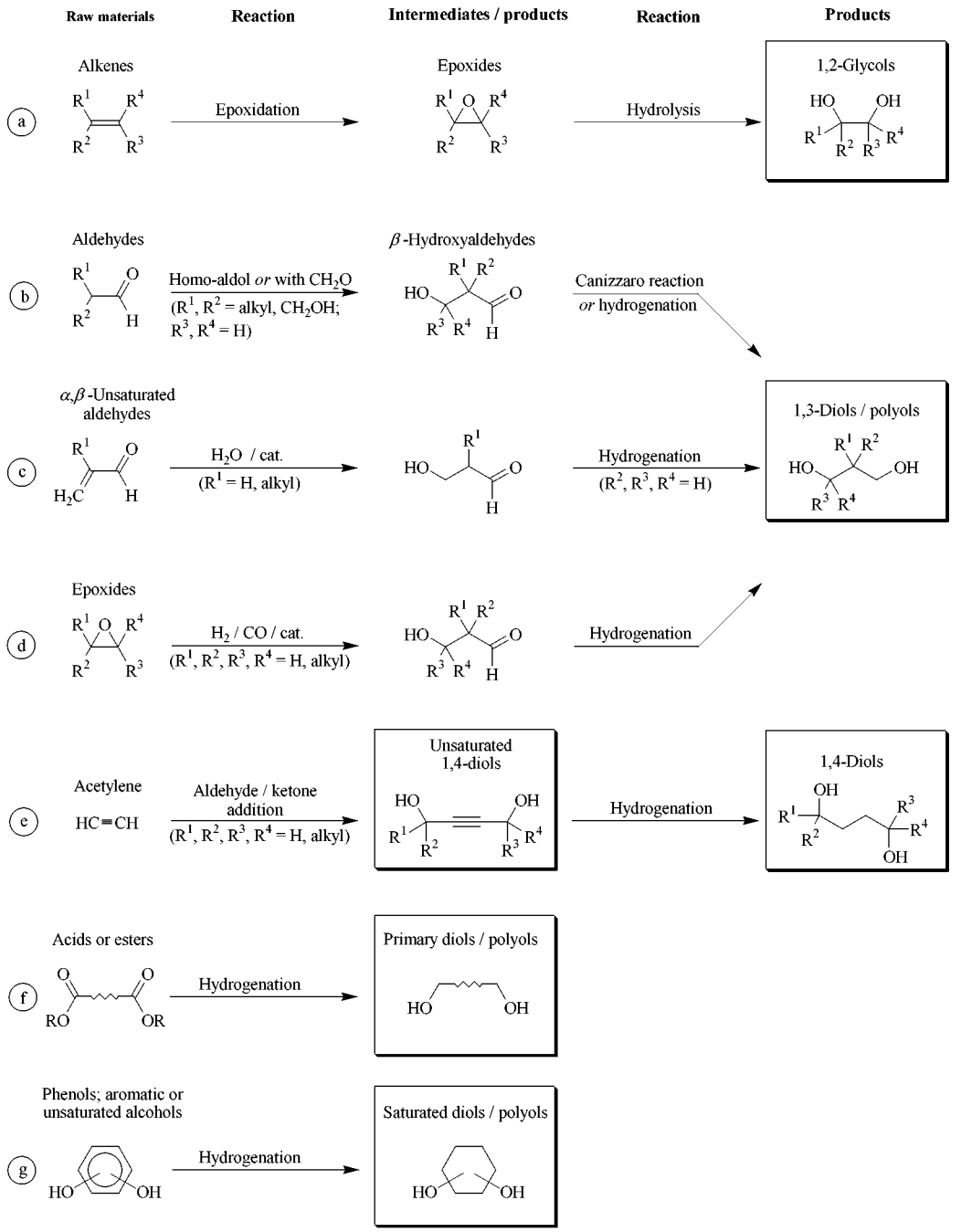

Scheme 1. Synthetic routes to polyhydric alcohols

b) Base-catalyzed aldol addition of formaldehyde to the appropriate higher aldehyde, followed by reduction, either by Cannizzaro reaction with excess formaldehyde and base or by catalytic hydrogenation. Reduction by the Cannizzaro reaction causes the formation of large amounts of formate salt, so that several recrystallization steps are required for the separation of the salts from the products. For the catalytic hydrogenation to be worthwhile, the equilibrium of the aldol addition must lie sufficiently to the right, that is, sufficient β-hydroxyaldehyde must be formed in the first step.
c) Addition of water to α,β-unsaturated aldehydes also results in formation of β-hydroxyaldehydes. The corresponding 1,3-diol can be obtained by subsequent hydrogenation.
d) Hydroformylation of epoxides represents the third method for producing β-hydroxyaldehydes for hydrogenation to 1,3-diols. In contrast to the aldol method (b), the methods based on Michael addition of water (c) and the cleavage of epoxides (a) are limited to the generation of one 1,3-diol unit.
e) Addition of aldehydes or ketones to acetylene to give the alkynediols or the corresponding 1,4-diols by subsequent hydrogenation.
f) Catalytic hydrogenation of carboxylic acids and their esters.
g) Hydrogenation of phenols or phenol derivatives (e.g., bisphenol A).

Uses. The chief use of polyhydric alcohols is as components of saturated and unsaturated polyester resins and alkyd resins. These resins, which are known for their excellent stability, are used in a wide range of varnishes and coatings. The addition of polyisocyanates to polyetherpolyols and polyesterpolyols derived from polyhydric alcohols leads to polyurethanes, which are used for the production of elastomers and various types of foams. Other uses are as synthetic lubricants, plasticizers, dispersants, fibers, and as additives to poly(vinyl chloride). Some of the diols are used in the production of pharmaceuticals and fragrances.

2. Diols

2.1. 1,5-Pentanediol

1,5-Pentanediol [*111-29-5*], pentamethylene glycol, $HOCH_2(CH_2)_3CH_2OH$, is a colorless liquid (see Table 1), soluble in water, alcohol, acetone, and relatively insoluble in aliphatic and aromatic hydrocarbons.

Production. 1,5-Pentanediol usually is produced by catalytic hydrogenation of glutaric acid or of its esters, although dicarboxylic acid mixtures containing glutaric acid also may be used (see Section 2.3). The crude products are purified by distillation. Manufacturers of 1,5-pentanediol are BASF and Ube Ind.; the world capacity is about 500 t/a.

Uses. 1,5-Pentanediol is the starting material for the production of various heterocyclic compounds, such as 1-methylpiperidine, which are used in the synthesis of pharmaceuticals and pesticides. The substance is also used in the production of fragrances and as a component of polyesters and polyurethanes.

2.2. 1,3-Propanediols

2.2.1. 2,2-Dimethyl-1,3-propanediol (Neopentyl Glycol)

Neopentyl glycol [*126-30-7*] has become an industrially important and versatile diol [1], particularly as a building block for polyesters and polyurethanes.

$HOCH_2C(CH_3)_2CH_2OH$

This is because of the wide availability of the starting materials, isobutyraldehyde and formaldehyde, and the extraordinary stability of its derivatives resulting from the quaternary structure and the absence of hydrogen atoms in the β-position.

Physical Properties. Neopentyl glycol (see Table 1), colorless crystals, has a phase transition at 40 – 42 °C. It is hygroscopic at

relative humidities greater than 50 % and sublimes readily, even somewhat below the melting point. Neopentyl glycol is soluble in water, alcohols, and ketones, moderately soluble in hot aromatic solvents, such as benzene and toluene, and relatively insoluble in aliphatic and cycloaliphatic solvents.

Chemical Properties. Pure neopentyl glycol is thermally stable up to the boiling point. However, in the presence of alkali salts or bases, neopentyl glycol decomposes above 140 °C.

In addition to the reactions typical of primary alcohols, such as ester, ether, and carbamate formation, 1,3-diols give six-membered cyclic derivatives with carbonyl compounds, carbonates, phosphites, sulfites, and borates.

Production. The aldol addition of isobutyraldehyde (2-methylpropanal) and formaldehyde gives hydroxypivaldehyde (3-hydroxy-2,2-dimethylpropanal) [*597-31-9*], which is then reduced to neopentyl glycol:

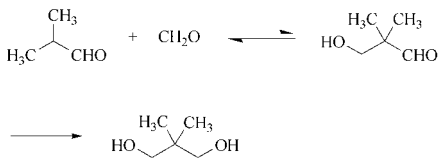

Both aldol addition and subsequent reduction are exothermic. Hydroxypivaldehyde can be reduced either by a crossed Cannizzaro reaction with equimolar amounts of formaldehyde and a base, or by catalytic hydrogenation (see Scheme 1).

Production by the Cannizzaro Reaction. In this process, aldol addition and reduction can be performed either simultaneously [2] or in two stages [3]. However, the individual reactions can be controlled more selectively in the two-stage process. Reduction of hydroxypivaldehyde with formaldehyde/sodium hydroxide solution leads to the coproduction of 0.8 – 0.9 t sodium formate per tonne of neopentyl glycol and to high production costs. The sodium formate has a limited commercial value.

Production by Catalytic Hydrogenation. This process usually is performed in two stages. The type of catalysts used for the aldol addition has a significant effect on the technical requirements of the hydrogenation and purification stages.

If *inorganic bases* such as potassium carbonate [4] or sodium hydroxide [5], [6] are used as catalysts, satisfactory yields (ca. 80 %) of neopentyl glycol can be obtained only with an excess of isobutyraldehyde. Side products formed by aldol addition, Cannizzaro reactions, or Tishchenko reactions are 2,2,4-trimethyl-1,3-pentanediol, neopentyl glycol isobutyrate, and hydroxypivalic acid neopentyl glycol ester (Section 2.2.2). Salts are also formed which interfere with the hydrogenation catalyst or cause decomposition during distillation. Sodium hydroxide always gives some Cannizzaro reaction, which can be partially suppressed by excess isobutyraldehyde and/or water soluble solvents such as aliphatic alcohols [6].

Hydroxypivaldehyde is separated from salts by extraction with dibutylether and hydrogenated on copper chromite catalyst between 175 and 220 °C.

In a variant of this process [7] hydrogenation and hydrogenolysis are performed stepwise at different temperatures (120 – 160 °C and 175 – 190 °C). At the higher temperature, ester byproducts also are hydrogenated to the corresponding alcohols.

Ruhrchemie [8] has avoided the interference of alkali salts, formed during the aldol addition, in the hydrogenation and in the purification stages by diluting the hydroxypivaldehyde with isobutanol, evaporating the mixture, and then hydrogenating the hydroxypivaldehyde in the gas phase at 110 – 150 °C over cobalt, copper, or nickel catalysts.

Tertiary amines are the state-of-the-art catalysts for the aldol addition.[9]. Hydroxypivaldehyde is formed rapidly by the reaction of isobutyraldehyde (up to 10 % excess), formaldehyde, and trialkylamine. The reaction is characterized by the complete conversion of formaldehyde, which is the main difference to the classical processes for trimethylolpropane and pentaerythritol. The excess of isobutyraldehyde is distilled and recycled together with the amine. The reaction runs very selectively and therefore most of the side products which are typical of the production based on the Cannizzaro reaction are negligible. The conditions of the hydrogenation reaction vary from 80 to 200 °C and 35 – 300 bar. Heterogeneous catalysts based on cobalt, cop-

per, or nickel can be used in the fixed-bed reactor [10]. After fractional distillation a very good yield of high-purity neopentyl glycol is obtained.

Quality, Storage, and Transportation. Commercial neopentyl glycol has a purity of about 99 % (determination by GLC). Impurities may include neopentyl glycol monoformate, neofpentyl glycol monoisobutyrate, 2,2,4-trimethyl-1,3-pentanediol, and the cyclic acetal from the reaction of hydroxypivaldehyde and neopentyl glycol.

During storage, higher temperatures and stacking pressures can lead to caking of the lower stack layers. As a hygroscopic material, neopentyl glycol must be stored dry.

Producers. Manufacturers are: BASF, Hoechst Celanese, Degussa-Hüls, Eastman Kodak, Koei, Mitsubishi Gas Chemical, Perstorp, and Polioli. World capacity is estimated at about 220 000 t/a.

Uses. The main use of neopentyl glycol is in saturated polyester resins for coil coatings and internal coating of tins, packing material varnishes, car varnishes, electrical insulation, and wire coatings, as well as in alkyd resins for paint varnishes, household varnishes, and self-drying industrial and ship varnishes. The alkyd resins can be used as solvent-containing, radiation-cured, water-dilutable, and powder varnishes and as high-solids paints. Neopentyl glycol is gaining importance particularly as a diol component in powder coating systems. In addition to the technological advantages of these varnishes, they are highly stable with respect to hydrolysis, to attack by alkali, and to heat; they also are weather-resistant.

Neopentyl glycol imparts the same advantages to unsaturated polyester resins, which are used for gel coats, corrosion resistant containers, and in molding materials for building components, etc. In addition, neopentyl glycol is a component of polyesterpolyols in plasticizers, its esters are synthetic lubricants, it is used as dispersant for titanium dioxide, and as starting material for polyurethanes.

2.2.2. Hydroxypivalic Acid Neopentyl Glycol Ester [11]

Hydroxypivalic acid neopentyl glycol ester, (3-hydroxy-2,2-dimethylpropyl)-3-hydroxy-2,2-dimethylpropionate (HPN) [*1115-20-4*] is the monohydroxypivalate of neopentyl glycol:

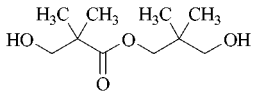

The compound has properties similar to neopentyl glycol but in addition specific characteristics that are responsible for its increasing use. Manufacturers are Eastman Chemical, BASF, Mitsubishi Gas, and Union Carbide. World capacity is estimated at 2000 – 2500 t/a.

Physical Properties. Hydroxypivalic acid neopentyl glycol ester is a colorless crystalline solid with a melting point of 50 °C (see Table 1). It is highly soluble in polar organic solvents; its solubility in water is limited to 27.4 % at 25 °C; conversely, 52.4 % water dissolves in HPN.

Chemical Properties. Pure HPN can be distilled at reduced pressure without decomposition up to about 200 °C. In the presence of alkali metal salt impurities, significant decomposition occurs even at temperatures of about 150 °C. Both hydroxyl groups undergo the usual reactions of primary alcohols. As with neopentyl glycol, the quaternary structure contributes to the excellent stability of HPN derivatives.

Production. Hydroxypivalic acid neopentyl glycol ester is produced by a Tishchenko reaction directly from hydroxypivaldehyde in the presence of a basic catalyst (e.g., aluminum oxide) [12] or classically by thermal treatment [13] (Section 2.2.1). Alternatively, it can be synthesized by esterification of neopentyl glycol with hydroxypivalic acid [14]. High-purity hydroxypivalic acid neopentyl glycol ester is obtained by subsequent distillation [15].

Uses. Hydroxypivalic acid neopentyl glycol ester is a diol modification agent in polyesters, polyurethanes, and plasticizers. As a component of polyester varnish systems, HPN has special

advantages with respect to flow properties, adhesion firmness, flexibility, thermal stability, resistance to light (UV), and impact strength at low temperature. The ester is used as a diol component of powder varnishes, coil coatings (e.g., internal coating of cans), electrophoresis varnishes, and unsaturated polyester resins; it is used also in radiation-cured varnish resins, binders for water lacquers, plasticizers, and polyurethane foams.

2.2.3. 2-Methyl-2-propyl-1,3-propanediol and 2-Butyl-2-ethyl-1,3-propanediol

2-Methyl-2-propyl-1,3-propanediol [*78-26-2*] and 2-butyl-2-ethyl-1,3-propanediol [*115-84-4*] (see Table 1) are produced either by the salt-free procedure based on aldol addition followed by hydrogenation [16] or by the classical combination of hydroxymethylation and Cannizarro-type disproportionation with formaldehyde [17] (see Scheme 1).

The production costs are considerably higher than for neopentyl glycol, since the synthesis of 2-methylpentanal and of 2-ethylhexanal from propanal and butanal requires two more steps: aldol condensation and hydrogenation.

Since ca. 1980 2-butyl-2-ethyl-1,3-propanediol has become a very useful diol for polyester or polyurethane coatings [18]. The world production of 2-butyl-2-ethyl-1,3-propanediol is estimated at about 10 000 t/a; the main producer is Neste.

The best known fine chemicals produced from 2-methyl-2-propyl-1,3-propanediol are the dicarbamates Meprobamate (β-methyl-β-propyltrimethylene dicarbamate [*57-53-4*]), used as a tranquilizer, and Carisoprodol (*N*-isopropyl-β-methyl-β-propyl-trimethylene dicarbamate [*78-44-4*]), used as a muscle relaxant.

2.2.4. 2-*sec*-Butyl-2-methyl-1,3-propanediol

2-*sec*-Butyl-2-methyl-1,3-propanediol (see Table 1) [*813-60-5*] can be made via 2,3-dimethylbutanal by simultaneous aldol addition and Cannizzaro reaction with formaldehyde [25]. So far it has been used only in the production of the tranquilizer Mebutamate (β-*sec*-butyl-β-methyl-trimethylene dicarbamate [*64-55-1*]).

2.2.5. 1,3-Propanediol

1,3-Propanediol [*504-63-2*] is a colorless liquid at room temperature in contrast to 2,2-dialkyl-1,3-diols (see Table 1). 1,3-Propanediol is produced commercially by Degussa starting from acrolein [*107-02-8*].

$$CH_2CHCHO + H_2O \longrightarrow HOHCH_2CH_2CHO$$

$$HOHCH_2CH_2CHO + H_2 \longrightarrow HOHCH_2CH_2CH_2OH$$

The addition of water under mild acidic conditions gives 3-hydroxypropionaldehyde [*2134-29-4*] with high selectivity (see Scheme 1). Preferentially buffer solutions with a pH 4 – 5 [19] or weak acidic ion exchange resins [20] are used as catalysts. Further hydrogenation of this aqueous solutions gives 1,3-propanediol [21]. There is an alternative route via hydroformylation of ethylene oxide and subsequent hydrogenation of the intermediate 3-hydroxypropionaldehyde [22].

A new large market for 1,3-propanediol will be in polyester coatings and in the production of poly(trimethylene terephthalate), a new material for the production of high quality carpet fibers [23].

2.2.6. 2-Methyl-1,3-propanediol

2-Methyl-1,3-propanediol [*2163-42-0*] is a fairly new commercially available diol and is produced by Arco with an estimated capacity of about 7000 t/a [24]. The process is designed for the production of 1,4-butanediol [*110-63-4*] as an alternative to Reppe technology and the acetoxylation of butadiene.

The raw material propylene oxide [*75-56-9*] is isomerized to allyl alcohol [*107-18-6*] in the first step. Hydroformylation of allyl alcohol gives 4-hydroxybutanal (79 %) [*25714-71-0*] and 3-hydroxy-2-methylpropanal (11 %) [*38433-80-6*]. The last step comprises hydrogenation of the *n* and *iso* intermediates to the corresponding diols. 2-Methyl-1,3-propanediol is used in polyester coatings.

2.3. 1,6-Hexanediol

1,6-Hexanediol [*629-11-8*], hexamethylene glycol, $HOCH_2(CH_2)_4CH_2OH$, colorless crystals (see Table 1), is soluble in water and other polar solvents.

Production. 1,6-Hexanediol is produced industrially by the catalytic hydrogenation of adipic acid or of its esters. Mixtures of dicarboxylic acids and hydroxycarboxylic acids with C_6 components formed in other processes (e.g., in cyclohexane oxidation) also can be used.

The acids are hydrogenated continuously at 170–240 °C and at 15.0–30.0 MPa (150–300 bar) on a suitable catalyst either in a trickle-flow (downflow) or a bubble-flow (upflow) fixed-bed reactor. The reactor temperature is controlled by circulating part of the reactor discharge. The hydrogen required for the hydrogenation is fed together with the recycle gas through the recycle gas compressor to the reactor (see Fig. 1).

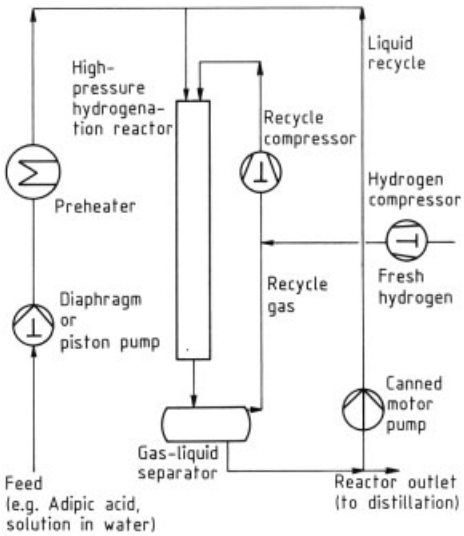

Figure 1. Hydrogenation of adipic acid or of its esters (trickle-flow fixed bed)

Side products of the synthesis are alcohols, ethers, diols, and esters. Pure 1,6-hexanediol is obtained by fractional distillation of the crude reactor discharge.

For the hydrogenation of dicarboxylic acids, catalysts containing cobalt, copper, or manganese are suitable [26]. For the hydrogenation of esters, catalysts such as copper chromite or copper with added zinc and barium are used as "full catalysts" or on inert carriers [27], [28]. Both acids and esters also may be hydrogenated using suspended catalysts.

Quality and Analysis. The assay of the pure product is about 98 %; impurities are various diols and ε-caprolactone as well as traces of water. The color number of the product determined photometrically according to the Pt/Co scale must not exceed 15 APHA. Above 70 °C, 1,6-hexanediol tends to turn yellow.

Storage and Transportation. 1,6-Hexanediol in the form of flakes or as a solidified melt is stored and shipped in barrels. It is transported as a melt in stainless steel containers, tank trucks, or tank cars.

Uses and Capacity. The most important use of 1,6-hexanediol is in the synthesis of polyesters and polyurethanes. In addition it is used in the production of varnishes, adhesives, pharmaceuticals, and textile auxiliaries.

The world capacity of 1,6-hexanediol is ca. 25 000 t/a; manufacturers are BASF and Ube.

2.4. Hexynediols

2.4.1. 3-Hexyne-2,5-diol

3-Hexyne-2,5-diol [*3031-66-1*] is a more or less solid yellow mass of crystals (see Table 1) depending on the relative content of the *meso* form. It is miscible with water and polar solvents.

$CH_3CH(OH)\text{-}C{\equiv}C\text{-}CH(OH)CH_3$

3-Hexyne-2,5-diol is produced by BASF from acetylene and acetaldehyde by the Reppe ethynylation method (→ Acetylene). 3-Hexyne-2,5-diol is used, for example, as a brightener for nickel baths (80 % aqueous solution) or as starting material for production of 2,5-hexanediol [28], but especially in the production of 4-hydroxy-2,5-dimethyl-3-(2*H*)-furanone [*3658-77-3*], trade name Furaneol [29], used as pineapple and strawberry flavor.

2.4.2. 2,5-Dimethyl-3-hexyne-2,5-diol

2,5-Dimethyl-3-hexyne-2,5-diol [*142-30-3*], $(CH_3)_2C(OH)\text{-}C{\equiv}C\text{-}C(OH)(CH_3)_2$, white crystals (see Table 1), is produced by addition of acetone to acetylene. It is an intermediate of minor importance in the production of 2,5-bis(*tert*-butylperoxy)-2,5-dimethyl-3-hexyne, which is used as an initiator in polymerization processes, as well as in the production of moschus fragrances, of chrysanthemumic acid, and of surfactants (surfinols) by oxalkylation (Air Products). Manufacturers are Enichim and Air Products.

2.5. 1,10-Decanediol

1,10-Decanediol [*112-47-0*] (see Table 1) is produced from sebacic acid by a process analogous to that described for 1,6-hexanediol [30] (see Section 2.3). It is an intermediate of minor importance in the production of polyesters and polyurethanes.

$HOCH_2(CH_2)_8CH_2OH$

Manufacturers are Ashland and Degussa-Hüls.

2.6. 2,2-Bis(4-hydroxycyclohexyl) propane

2,2-Bis(4-hydroxycyclohexyl)propane [*80-04-6*] is a crystalline, colorless, and odorless compound (see Table 1).

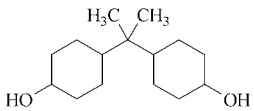

The conformation of the cyclohexane rings gives rise to *cis-cis*, *cis-trans* and *trans-trans* isomers. The industrial product is a mixture of these stereoisomers. The physical properties depend on the relative amounts of the isomers present. The *trans-trans* compound has the lowest energy and is formed from the other isomers at higher temperature. It has the lowest solubility in solvents such as methanol, butanol, and acetone and has the highest melting point of these isomers.

Production. 2,2-Bis(4-hydroxycyclohexyl)propane is produced by hydrogenation of bisphenol A, 2,2-bis(4-hydroxyphenyl)propane [*80-05-7*], an industrially important compound made from acetone and phenol (→ Phenol Derivatives). Bisphenol A is hydrogenated at about 25.0 MPa (250 bar) and at 200 °C on a nickel, cobalt, or ruthenium catalyst [31–33].

Uses. 2,2-Bis(4-hydroxycyclohexyl)propane is used mainly in synthetic materials as a bifunctional component, e.g., as a modifier in the production of saturated and unsaturated polyester resins, and of oil- and fatty-acid-containing alkyd resins. Exchange of hydroxyl for amino gives a diamine which can be used in the production of glass-clear polyamides.

2.7. 1,4-Bis(hydroxymethyl)cyclohexane

1,4-Bis(hydroxymethyl)cyclohexane [*105-08-8*] is commonly known as 1,4-cyclohexanedimethanol (CHDM), also as 1,4-dimethylolcyclohexane, or hexahydro-*p*-xylylene glycol.

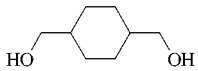

The compound became known in the 1960s mainly for its use in polyester films and fibers highly resistant to hydrolysis.

Physical Properties (see also Table 1). 1,4-Cyclohexanedimethanol is colorless and has a slight odor. The softening point of the *cis* isomer is 43 °C, that of the *trans* isomer 67 °C. *Cis – trans* mixtures can be isomerized at 200 °C in the presence of alkoxides to an equilibrium mixture of 42 % *cis* and 76 % *trans*. The commercial product consists of a mixture of *cis* and *trans* isomers with a typical ratio of 1/3.

1,4-Cyclohexanedimethanol is miscible with water as well as with low molecular mass alcohols; it is very slightly soluble in hydrocarbons and ethers. Its solubility at 20 °C (g in 100 g solvent) in benzene is 1.1, in trichloromethane 5.7, in water 92.0, and in methanol 92.2.

Chemical Properties. CHDM undergoes the characteristic reactions of a diol with two primary hydroxyl functions. In contrast to 1,2- and 1,3-diols reactions of CHDM with aldehydes or ketones do not result in the formation of heterocyclic products, such as 1,3-dioxolanes or 1,3-dioxenanes.

Production. The commercially important process for producing CHDM is the hydrogenation of dimethyl terephthalate (DMT) [*120-61-6*]via dimethyl hexahydroterephthalate (DMHT) [*94-60-0*] in a methanolic solution or in the molten state [34]:

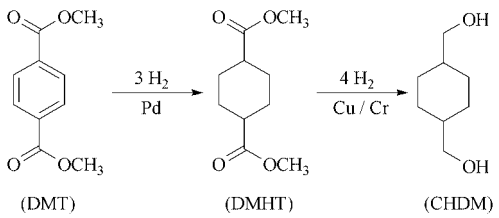

The hydrogenation plant consists of two reactors. In the first reactor, a continuous circulation of product DMHT and of DMT is maintained. The molten DMT is pumped into the inlet of this reactor, which operates at 30 – 48 MPa and 160 – 180 °C with a commercial supported Pd catalyst. Temperature control is achieved by operating at sufficiently high cross-sectional loadings of the mixture of ca. 10 % in product DMHT. This allows the radial dissipation of heat via the reactor walls and largely avoids high temperature peaks. The cooled reactor effluent, which contains only minor amounts of unreacted DMT, is divided: 8 – 10 parts are recycled via a pump and heat exchanger to the reactor inlet. The remaining 1 – 2 parts (the exact amount corresponds to the amount of fresh DMT) are fed continuously into the second reactor for the final Pd-catalyzed hydrogenation.

An advantage of this process is that DMT can be hydrogenated as ca. 10 % dilute feed without handling large amounts of liquid. The yield of DMHT is typically 97 – 98 %, with methyl 4-methyl-4-cyclohexanecarboxylic acid methyl ester [*51181-40-9*] and some 1-hydroxymethyl-4-methylcyclohexane [*34885-03-5*] as main byproducts.

1,4-Bis(hydroxymethyl)cyclohexane (CHDM) is formed in the second stage of the hydrogenation. The crude DMHT of stage one is used in a plant analogous to that of the first stage [40], although any conventional hydrogenation plant can be used since the heat of the ester hydrogenation is comparatively small.

The industrial processes use commercial copper chromite catalysts. The effects of catalyst, residence time, and temperature must be carefully adjusted in order to achieve a *cis*/*trans* ratio of 1/3 to 1/4 in the CHDM product.

A constant isomer ratio is important for further processing of CHDM in polyester formation since the crystal structures of polyesters of *cis*- and *trans*-CHDM differ and thus influence the melting range and density of polyester fibers.

In addition to the side products of the first hydrogenation stage, 4-methyloxymethylhydroxymethylcyclohexane and bis(4-hydroxymethylcyclohexyl) ether can be formed in the second stage of ester hydrogenation. The Eastman process (see above) [34] avoids the formation of high-boiling compounds and thus saves cumbersome separation of side products by fractionation in vacuum. The diol is purified by simply removing methanol and low-boiling compounds. The CHDM of fiber quality obtained by this method is ca. 99 % pure.

Pure CHDM may also be obtained by recycling waste poly(1,4-cyclohexylene dimethylene terephthalate) by first cleaving the polyester in the presence of low molecular mass alcohols and then subjecting the resulting mixture to a hydrogenation analogous to that described above [35]. Estimated world capacity is 55 000 t/a.

Uses. An important application of CHDM is still the production of terephthalate – polyester fibers with lower densities and higher melting points than analogous fibers using ethylene glycol. They are more stable to hydrolysis and have more favorable electrical properties than similar polyesters made from other diols [35].

Unsaturated resins based on CHDM are characterized by very good properties with respect to water absorption, thermal stability, and resistance to environmental and chemical exposure [36]. The use of CHDM in saturated polyester resins leads to powder, waterborne, and solvent-based coatings with excellent hardness and durability [37]. 1,4-Cyclohexanedimethanol is of increasing interest for application in polyurethanes and poly-

carbonates [38] as well as an antifogging agent and as a sensitizing additive in silver bromide emulsions.

2.8. 2,2,4-Trimethyl-1,3-pentanediol

2,24-Trimethyl-1,3-pentanediol [*144-19-4*], TMPD glycol, is a white crystalline solid (see Table 1). It is a versatile diol with a wide range of applications including polyester resins, polyurethane foams, and lubricants.

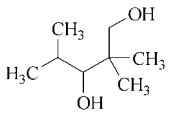

Physical and Chemical Properties. TMPD glycol is only slightly soluble (< 2 %) in water and kerosene, but very soluble (up to 80 %) in alcohols. The solubility in benzene, acetone, and ethers is at most 29 %.

TMPD glycol, with one primary and one secondary hydroxyl function, shows the typical reactivity of a diol. In comparison to other diols, such as CHDM (see Section 2.7) or NPG (see Section 2.2.1), TMPD glycol has an unsymmetrical structure and shielded hydroxyl groups.

Production. TMPD glycol is produced by homo-aldol condensation of isobutyraldehyde [*78-84-2*] in the presence of an alkaline catalyst followed by hydrogenation of the resulting cyclic acetal to give TMPD glycol and isobutanol. Eastman Chemical is the only manufacturer of TMPD glycol.

Uses. TMPD glycol is mainly used as raw material in the production of unsaturated and saturated polyesters (e.g. for waterborne or high-solids coatings) and as an intermediate for polyesters used in the manufacture of polyurethane elastomers and foams [39]. These resins are very stable to hydrolysis and resistant to corrosion, as well as having other desirable features, such as low density, low viscosity, and good compatibility.

2.9. Vicinal Diols by Hydroxylation of Olefins with Peracids

Some vicinal diols are manufactured on a commercial scale by the reaction of an alkene with an organic peracid, such as performic acid, produced in situ, or peracetic acid [41]. An epoxide is formed initially; acid-catalyzed cleavage of the epoxide ring leads to a 1,2-diol or a 2-hydroxyalkyl-ester or a mixture of the two. The 2-hydroxyalkyl ester is hydrolyzed to the 1,2-diol by the subsequent treatment with a base.

2.9.1. 1,2-Pentanediol

1,2-Pentanediol [*5343-92-0*], 1,2-dihydroxypentane, $CH_3(CH_2)_2CH(OH)CH_2OH$, is a colorless and odorless liquid which is miscible with water and polar organic solvents (see also Table 1).

The compound is made by treating a mixture of 1-pentene and formic acid with 40 % hydrogen peroxide at 30 °C [42]. 1,2-Pentanediol has attained commercial importance as a synthetic building block for systemic fungicides [43] and as a moisturizing agent in cosmetics.

2.9.2. Other 1,2-Diols

Pinacol [*76-09-5*], 2,3-dimethyl-2,3-butanediol, $(CH_3)_2C(OH)C(OH)(CH_3)_2$, colorless crystals, is soluble in hot water, alcohol, and ether and slightly soluble in cold water and carbon disulfide. The hexahydrate, *mp* 45 °C, crystallizes as four-sided plates (see also Table 1).

1,2-Hexanediol [*6920-22-5*], (*bp* 108 °C/0.5 hPa; ϱ 0.953 g/cm^3 at 20 °C) is a colorless liquid miscible with water. It is used in printing inks and in cosmetics.

1,2-Octanediol [*1117-86-8*] (*bp* 140 °C/1.6 hPa) is a white waxy solid, slightly soluble in water. Its production is described in [42], [44]. It is used in printing inks and in cosmetics.

3. Triols

3.1. Trimethylolpropane

Properties. Trimethylolpropane [*77-99-6*], 2-ethyl-2-hydroxymethyl-1,3-propanediol, $CH_3CH_2C(CH_2OH)_3$, is a colorless, crystalline, trivalent alcohol (see Table 1). Trimethylolpropane is quite soluble in water and polar organic solvents. The three primary hydroxyl groups undergo the normal OH group reactions.

Production. Trimethylolpropane is made by the base-catalyzed aldol addition of butyraldehyde with formaldehyde followed by Cannizzaro reaction of the intermediate 2,2-bis(hydroxymethyl)butanal with additional formaldehyde and at least a stoichiometric quantity of base (see Scheme 1):

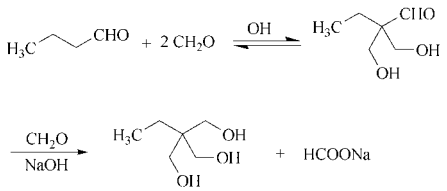

Sodium hydroxide [45], [46] or calcium hydroxide [47], [48] are used as bases in industrial-scale production. The yield in both cases is ca. 90 %. Many attempts have been made to avoid the large quantities of formate byproduct by catalytic hydrogenation of the 2,2-bis(hydroxymethyl)butanal [49], [50]. These processes have been achieved by using tertiary amines as catalyst and transforming the formed ammonium formate into methyl formate [87] or even avoiding the formation of larger amounts of formates [88].

Various procedures have been suggested for separating the trimethylolpropane from the formate. For instance, the concentrated reaction solution can be extracted with an organic solvent, which is then evaporated and the crude trimethylolpropane is purified by vacuum distillation [51], [52], [53]. In another variant [54], the aqueous reaction solution is evaporated until most of the sodium formate crystallizes and is removed by hot filtration. The liquid trimethylolpropane which remains is liberated of residual salts with an ion exchanger and then distilled. The removal of discoloring impurities is described in [55].

Cyclic diethers of the acetal type (1,3-dioxanes, formals) can be converted into TMP and methanol by metal-catalyzed hydrogenation [56].

Uses. Large quantities of trimethylolpropane and its ethoxylated derivatives are used as precursors for urethanes and polyester resins. Another important field of application is in medium-oil and short-oil alkyd resins (→ Alkyd Resins). The resulting lacquers are characterized by excellent resistance to alkali, detergents, and water, combined with outstanding impact resistance and flexibility, as well as excellent clearness and clearness retention.

Reaction products with fatty acids ($C_5 - C_{10}$) are components of synthetic lubricants. A strongly growing market will be the use of these blended polyol esters in chlorine-free, purely fluorocarbon based refrigerant systems. Powder and high-solids coatings represent a strong potential market for TMP, because it contributes to low viscosity. TMP acrylates are used as reactive diluents in UV-cured systems for inks and coatings, and TMP allyl ethers in unsaturated polyesters.

Economic Aspects. World production capacity for trimethylolpropane was estimated at 150 000 t in 1998, of which about 67 000 t was accounted for by Western Europe, 60 000 t by the United States, and 15 000 t by Japan. The main producers are Bayer, Perstorp, and Polioli in Western Europe; Celanese and Perstorp in the United States; and Koei and Mitsubishi Gas in Japan.

3.2. Trimethylolethane

Properties. Trimethylolethane [*77-85-0*], 2-hydroxymethyl-2-methyl-1,3-propanediol, $CH_3C(CH_2OH)_3$, is a colorless, crystalline substance (see Table 1), soluble in water (140 g per 100 g at 25 °C), alcohols, and acetone.

Production. Trimethylolethane is made by aldol condensation of propionaldehyde with formaldehyde, followed by reaction of the intermediate 2,2-bis(hydroxymethyl)propanal with

excess formaldehyde in the presence of sodium hydroxide or lime as basic component (see Scheme 1).

$$H_3CCH_2CHO + 3\,H_2CO + NaOH \longrightarrow H_3CC(CH_2OH)_3 + HCOONa$$

The resulting aqueous solution is freed from excess formaldehyde by distillation under pressure and further concentrated. The formate byproduct can be separated by extracting the trimethylolethane from the residue with organic solvents such as 2-propanol [57] or by almost complete removal of the water and dissolving the trimethylolethane in acetone or methanol.

Uses and Economic Aspects. Trimethylolethane, like trimethylolpropane, is a polyol component of short- and medium-oil alkyd resins, in which it results in a shorter drying time and, because of its shorter alkyl chain, a harder lacquer film compared with trimethylolpropane. However, its economic importance is small compared with that of pentaerythritol or trimethylolpropane. It is produced by Mallinckrodt (USA) and Mitsubishi Gas (Japan) in amounts of several thousand tonnes per annum.

4. Tetrols

4.1. Ditrimethylolpropane

Ditrimethylolpropane [*23235-61-2*] 2,2-[oxybis(methylene)]-bis(2-ethyl)-1,3-propanediol is a colorless and crystalline substance (*mp* 112 – 114 °C), only slightly soluble in water at ambient temperature (2.6 g/100 g at 25 °C).

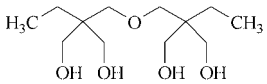

It is a byproduct of TMP production and can be isolated from the residue of the TMP distillation step. Newer investigations show that ditrimethylolpropane can be prepared in high yield and selectivity by the reaction of TMP, formaldehyde, and 2-ethylacrolein [58]. It is used in water-thinnable alkyd resins and as a polyol component in PVC stabilizers based on Ca/Zn alkyl carboxylates. Producers are Perstorp (Sweden) and Mitsubishi Gas (Japan).

4.2. Pentaerythritol

Pentaerythritol [*115-77-5*], 2,2-bis(hydroxymethyl)-1,3-propanediol, $C(CH_2OH)_4$, was discovered by Tollens in 1882 when an aqueous solution of formaldehyde containing some acetaldehyde was allowed to stand with barium hydroxide [59].

Properties. Pentaerythritol is a colorless, crystalline alcohol with four primary hydroxyl groups. It crystallizes in a tetragonal system with. The tetragonal crystals convert into the cubic form at 180 – 190 °C. The compound sublimes on heating in vacuum without dissociation. Important physical data are listed below (see also Table 1):

Heat of combustion (p = const.)	2767 kJ/mol
Heat of formation	931 kJ/mol
Specific heat capacity at 100 °C	255 J/mol
Ignition temperature	390 °C
Solubility in water (per 100 g water)	
at 25 °C	7 g
at 97 °C	77 g

Mixtures of air and pentaerythritol dust are explosive at 490 °C and at dust concentrations of more than 30 g/m^3. The four primary hydroxyl groups undergo the normal reactions of OH groups [60]. Pentaerythritol is oxidized to tris(hydroxymethyl)acetic acid by air in the presence of platinum or palladium [61].

Production. Pentaerythritol is prepared by the reaction of acetaldehyde with formaldehyde in alkaline medium. The aldol addition results in replacement of the three α-hydrogen atoms, leading to trimethylolacetaldehyde (pentaerythrose), which is reduced to pentaerythritol by excess formaldehyde in the presence of base (sodium hydroxide or calcium hydroxide). The yield is approximately 90 %. Catalytic hydrogenation of pentaerythrose, which would avoid the formate byproduct, is not practicable. Higher homologues of pentaerythritol with the general formula

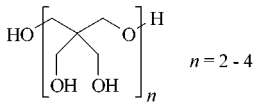

are byproducts of the production of pentaery-

thritol. The product with $n = 2$, known as dipentaerythritol, is the primary contaminant. To minimize its formation it is necessary to increase the ratio of formaldehyde to acetaldehyde beyond the stoichiometric ratio of 4:1 [60]. Additionally, certain amounts of formaldehyde acetals are formed [62]. Both batch and continuous processes for the production of pentaerythritol have been described [63], [64].

To isolate the product, the aqueous reaction solution is first freed of excess formaldehyde and most of the water by distillation under pressure. After further concentration, usually in vacuo, the crude pentaerythritol crystallizes. The mother liquor contains the sodium formate plus residual pentaerythritol. Separation of the formate, recycling of the residual pentaerythritol, and elimination of the wastewater occur in subsequent process steps. The crude pentaerythritol is purified by recrystallization from water after treatment with activated carbon. By using lime, the process runs reversely; first calcium formate separates because of its low water solubility, the mother liquor than contains the dissolved pentaerythrithol. Residual Ca^{2+} or Na^{+} ions can be removed by ion-exchange resins. The wastewater contains a certain amount of dissolved pentaerythritol, which can be recovered [65].

Quality and Analysis. Pentaerythritol is available in various grades which differ in their content of the above-mentioned byproducts including residual ash (alkali metal salts and/or alkaline earth metal salts). Higher ash content often has a detrimental effect on the color of the alkyd resin obtained from the product. Products termed pentaerythritol monograde generally contain more than 97 % pentaerythritol, whereas the technical grade consists of 80 – 95 % pentaerythritol and 5 – 15 % dipentaerythritol.

For *quality control*, the melting point, hydroxyl number, ash content, and the color or iodine number of the melt are determined. Gas chromatographic analysis after silylation with agents such as *N*-methyl-*N*-trimethylsilyltrifluoracetamide [66] gives precise information on the composition of the sample. Analysis by high-pressure liquid chromatography has also been described [67].

Uses. The major application of pentaerythritol (ca. 70 %) is the production of long-oil and medium-oil alkyd resins. The pentaerythritol structure gives the resin outstanding properties both in processing and end use. Resins prepared from pentaerythritol and rosin or tall oil are used for the modification of printing inks and nitrocellulose lacquers and for the preparation of adhesives and core binders. Solvent-free, water-thinnable alkyd resin emulsions can be prepared by using tribasic acids in the esterification process, which gives water-dispersible salts after neutralization.

Pentaerythritol esters of short- to medium-chain carboxylic acids ($C_5 - C_{10}$) are applied as high-temperature synthetic lubricants, lubricant additives, and plasticizers. An emerging market for these esters is as refrigerant lubricants. Because of the ozone-depleting properties of CFCs they are to be replaced by chlorine-free fluorocarbons. However, since this class of compounds differs in polarity the compressor must be lubricated with oils of higher polarity [68], [69] Pentaerythritol tetranitrate is used as a detonator in fuses and, to a limited extent, as a military explosive. The tetranitrate also is used in medicine as a long-term coronary vasodilator.

Finely ground pentaerythritol in combination with metal soaps serves as a nontoxic costabilizer for poly(vinyl chloride). It also is used in flame retardant compositions (intumescent paints). Under the action of heat an insulating foam layer forms, up to 50 times the thickness of the original coating. Pentaerythritol triacrylate is used as an active diluent in radiation-cured paints. Pentaerythritol adducts with ethylene oxide are used as emulsifiers, inorganic phosphoric or phosphoric esters are used as additives in hydraulic fluids, and tetraesters with sterically hindered phenylcarboxylic acids act as a highly effective antioxidant in plastics.

Economic Aspects. In 1998, world production capacity was estimated to be 360 000 t, 71 000 t of which was in the United States, 120 000 t in Europe, 27 000 t in Japan, and 25 000 t in Canada. Smaller producers are located in many other countries. The main manufacturers are Perstorp, Hoechst-Celanese, and Degussa-Hüls. The overall pentaerythritol market is expected to grow at a rate of 1.5 – 2 %/a. The alkyd resin market should be stable, and growing use of fatty acid esters in refrigerant systems can be assumed.

Table 2. Lethal doses of polyhydric alcohols (in mg/kg)

1,4-Butanediol [*25265-75-2*]	[73], [74]
LD_{50} 1500 – 1780 (rat, oral)	
1,2-Butanediol [*584-03-2*]	[73]
LD_{50} 16000 (rat, oral)	
1,3-Butanediol [*107-88-0*]	[75], [76]
LD_{50} 29600 (rat, oral)	
1,5-Pentanediol [*111-29-5*]	[77]
LD_{50} 5890 (rat, oral)	
1,6-Hexanediol [*629-11-8*]	[77]
LD_{50} 3730 (rat, oral)	
2,5-Hexanediol [*2935-44-6*]	[77]
LD_{50} 5000 (rat, oral)	
LD_{50} 16 000 (rabbit, dermal)	
1,2-Hexanediol [*6920-22-5*]	
$LD_{50} > 5000$ (rat, oral)	
1,2-Octanediol [*1117-86-8*]	
LD_{50} 2200 (rat, oral)	
1,3-Propanediol [*504-63-2*]	
LD_{50} 15670 (rat, oral)	
2,2,4-Trimethyl-1,3-pentanediol, (TMPD glycol) [*144-19-4*]	[79]
LD_{50} 2000 (rat, oral),	
LD_{50} 145 (rat, intravenous)	
Neopentyl glycol [*126-30-7*]	
LD_{50} 6400 – 12800 (rat, oral)	[81]
1,4-Bis(hydroxymethyl)cyclohexane (CHDM)[*105-08-8*]	[77], [80]
LD_o 3200 (rat, oral)	
LD_o 1600 (mouse, oral)	
Pentaerythritol [*115-77-5*]	
$LD_o > 5110$ (rat, oral)	[83]
LD_{50} 19 000 (mouse, oral)	[84]
Dipentaerythritol [*124-58-9*]	
$LD_{50} > 2000$ (rat, oral)	
Trimethylolpropane [*77-99-6*]	[82]
LD_{50} 14100 (rat, oral)	

5. Higher Polyols

5.1. Dipentaerythritol

Dipentaerythritol [*126-58-9*], 2,2-[oxybis(methylene)-bis[2-hydroxymethyl]-1,3-propanediol, is a byproduct of pentaerythritol production. Because of its low solubility in water (0.22 g/100 g H_2O at 20 °C and 10.0 g/100 g H_2O at 100 °C) it can be separated from pentaerythritol by fractional crystallization. The yield can be increased by changing the ratio of formaldehyde/acetaldehyde towards stoichiometric. Adding pentaerythritol before starting the reaction also increases the yield. An effective method is to use acrolein instead of acetaldehyde [70], [71].

In the last decade the consumption of dipentaerythritol steadily increased. The main outlets are $C_5 - C_{10}$ ester lubricants for jet engines and refrigerator systems [72], dipentahexaacrylate as cross-linking agent for UV-curing acrylates, for water-resistant intumescent paints, alkyd resins with low viscosity (high solids), and as a nonsubliming polyol component in nontoxic PVC stabilizer systems.

The total installed capacity is ca. 6000 t. Main producers are Hercules, Koei, Perstorp, Polialco, and Degussa-Hüls.

5.2. Tripentaerythritol

Tripentaerythritol [*78-24-0*], 2,2-bis[hydroxy-2,2-bis(hydroxymethyl)propoxy]methyl-1,3-propanediol is only slightly soluble in cold and hot water and can be isolated during the production of dipentaerythritol. It has at present only a very limited economic importance.

6. Toxicology

Table 2 summarizes the acute toxicity data of polyhydric alcohols. According to the lethal doses shown, these compounds are generally little toxic. For comparison, the LD_{50} values of the butanediols also are included in Table 2.

TMPD Glycol, 2,2,4-trimethyl-1,3-pentanediol, is classified as "slightly toxic" [78], [79], see Table 2. In rabbits, moderate eye irritation was observed. Slight to no skin irritation occurred in guinea pigs, and no skin sensitization was found. Repeated skin application studies in humans gave no evidence of irritation, sensitization, photosensitization, or systemic toxic effects. In humans, TMPD glycol is rapidly excreted in the urine, partly unchanged, partly as the glucuronide and sulfate conjugates, and partly as 2,2,4-trimethyl-3-hydroxyvaleric acid.

1,4-Bis(hydroxymethyl)cyclohexane (CHDM) has a low acute toxicity (see Table 2). For metabolism of CHDM in rats, see [80].

Trimethylolpropane is practically nontoxic (see Table 2). Skin irritation was not observed.

Pentaerythritol is practically nontoxic. Health disturbances resulting from handling pentaerythritol have never been observed. High oral doses (50 g) administered to humans resulted only in slightly increased pulse rate [85]. Diarrhea was observed at high doses in animals. Pentaerythritol is readily biodegradable under aerobic conditions according to OECD Guideline No. 301 [86].

1,2-Diols. The oral LD_{50} values in rats for 1,2-hexane-diol, 1,2-octanediol, are included in Table 2. For these 1,2-diols, skin irritation was not observed in rabbits. 1,2-Octanediol caused no eye irritation in rabbits; testing in guinea pigs resulted in a slight sensitization reaction.

7. References

1. B. Cornils, H. Feichtinger, *Chem. Ztg.* **100** (1976) 504 – 514.
2. F. C. Whitmore, A. H. Popkin, H. I. Bernstein, J. P. Wilkins, *J. Am. Chem. Soc.* **63** (1941) 124 – 127. R. W. Shortridge, R. A. Craig, K. W. Greenlee, J. M. Derfer, C. E. Boord, *J. Am. Chem. Soc.* **70** (1948) 946 – 949. BASF, DE 1 057 083, 1957(E. Haarer, K. Ruhl). Riedel de Haen, DE 1 020 614, 1956(H. Müller). Bayer, DE 1 041 027, 1955(H. Danziger, K. Haeseler, G. Schulze). Perstorp, WO 9 601 249, 1996(L. Erlandsaon, S. Johansson). Mitsubishi Gas Chemical, JP 61 091 144, 1986(A. Ninomiya).
3. BASF, DE-AS 1800506, 1968 (F. Merger, W. Fuchs).
4. Eastman Kodak, US 2811562, 1954 (H. J. Hagemeyer, Jr.), US 2895996, 1956 (H. N. Wright, Jr., H. J. Hagemeyer, Jr.), US 3939216, 1974 (R. L. Wright).
5. Eastman Kodak, US 3340312, 1964 (R. B. Duke, M. A. Perry, H. N. Wright). Kuhlmann, FR 1230558, 1959 (F. Le Paire). Ruhrchemie, DE 2045669, 1970 (H. Tummes, G. Schiewe, B. Cornils, W. Pluta, J. Falbe). Hoechst, DE-AS 1768274, 1968 (G. Jacobson, H. Fernholz, D. Freudenberger).
6. Eastman Kodak, DE-OS 1804984, 1968 (H. J. Hagemeyer, Jr., S. H. Johnson).
7. Chemische Werke Hüls, EP 6460, 1978 (M. zur Hausen, M. Kaufhold, E. Lange).
8. Ruhrchemie, DE 2054601 (=US 4094914), 1970 (W. Rottig, H. Tummes, B. Cornils, J. Weber).
9. BASF, DE 1 957 591, 1968(F. Merger, S. Winderl, E. Haarer, W. Fuchs). BASF, DE 1 793 512, 1968(F. Merger, R. Platz, W. Fuchs). Ruhrchemie, DE 3 644 675, 1988(N. Breitkopf, W. Hoefs, H. Kalbfell, F. Thoennessen, P. Lappe, H. Springer). Eastman Kodak, WO 8 903 814, 1989(D. L. Morris). Eastman Kodak, US 4 855 515, 1989(D. L. Morris, B. W. Palmer, T. W. McAninch). Mitsubishi Gas Chemical, EP 343475, 1989(N. Teruyuki, F. Tomiyoshi, K. Seiji, F. Yoshimi). Hoechst, DE 3942792, 1989(G. Daembkes, P. Lappe, H. Springer, F. Thoennessen). Aristech, US 5 146 012, 1991(J. S. Salek, J. Pugach, C. L. Carole, L. A. Cullo). Mitsubishi Gas Chemical, EP 708073, 1994(T. Nimomiya, T. Watanabe, A. Mori, T. Ikebe, A. Iwamoto).
10. Mitsubishi Gas Chemical, EP 484 800, 1990(M. Yoneoka, K. Watabe, G. Matsuda). Mitsubishi Gas Chemical, US 5 395 989, 1990(M. Yoneoka, K. Watabe, G. Matsuda). BASF, WO 9 532 171, 1995(M. Brudermueller, M. Irgang, M. Schmidt-Radde, F. Merger, T. Witzel, D. Kratz, E. Danz, A. Wittwer).
11. J. Arpe, *Chem. Ztg* **. 97** (1973) 53 – 62.
12. Eastman Chemical, US 5 041 621, 1991(D. Mores, G. Luce). BASF, US 5 024 772, 1991(L. Thurman, J. Dowd, K. Fischer).
13. G. K. Finch, *J. Org. Chem.* **25** (1960) 2219 – 2220. Eastman Kodak, DE 1 168 411, 1959(G. Finch).
14. BASF, EP 410 167, 1990(L. R. Thurman, J. P. Dowd, K. J. Fischer).
15. BASF, DE 1 793 512, 1968(F. Merger, R. Pletz, W. Fuchs). BASF, DE 2234110, 1972(F. Merger, G. Dümbgen, W. Fuchs). Eastman Kodak, WO 9 207 815, 1992(G. E. Butler, G. C. Luce, D. L. Morris).
16. Kyowa Yuka, JP 4 069 351, 1990 (S. Mizutani, K. Sato, K. Muto). BASF, 1 927 301, 1969 (F. Meger, R. Pletz, W. Fuchs). Eastman Kodak, US 5 146 004, 1991 (D. L. Morris, W. A. Beavers, W. E. Choate). Neste, WO 9 500 464, 1995 (K. Kulmala, K. Ankner, L. Rintala).
17. US Industrial Chemicals, US 2 413 803, 1947 (W. Tribit). Montrose Chemical, US 2761 881, 1956 (J. Rosin). Chisso, JP 62 129 233, 1985 (K. Tani, K. Saito). Koei, JP 2 062 836, 1988 (K. Doi, T. Moriyama).
18. Herberts, EP 688 840, 1994 (K. Bederke, F. Herrmann, H. Kerber, T. Kutzner, H. M. Schoenrock). Nippon Polyurethanes Kogyo, JP 7 113 005, 1993 (T. Morishima, S. Hama, S. Konishi). BASF, DE 44 01 544, 1994 (H. P. Rink).
19. Degussa, US 5 284 979, 1992 (T. Haas, G. Boehme, D. Arntz). Du Pont, US 3 536 763, 1970 (H. S. Elenterio, T. A. Koch). Shell, GB 1 185 615, 1969 (E. T. Lutz).
20. Degussa, EP 487 903, 1992 (D. Arntz, N. Wiegand).
21. Degussa, EP 572 812, 1993 (D. Arntz, T. Haas, N. Wiegand). Degussa, EP 535 565, 1993 (D. Arntz, T. Haas, A. Schaefer-Sindlinger). Ruhrchemie, DE 2 054 601, 1972 (W. Rottig, H. Tummes, B. Cornils).

22. Shell, WO 9 610 550, 1996 (J. P. Arhancet, T. C. Forschner, J. B. Powell, T. C. Semple, L. H. Slaugh, T. B. Thomason, P. R. Weider). Shell, WO 9745 390, 1997 (J. P. Arhancet, A. N. Matzakos, W. R. Pledger, J. B. Powell, L. H. Slaugh, P. R. Weider). Union Carbide, US 5 449 653, 1995 (J. R. Briggs, J. M. Maher, A. M. Harrison). Hoechst Celanese, US 50 535 562, 1991 (D. T. Kwoliang et al.).
23. Du Pont, WO 97/2354 3, 1997 (J. M. Stouffer, E. N. Blanchard, K. W. Leffew).
24. A. M. Brownstein, *Chemtech* **1991**, 506 – 511. Kuraray, US 4 567 305, 1986 (K. Kikuchi, K. Koga, H. Kojima, M. Matsumoto, S. Miura, M. Tamura, S. Yamashita).
25. Ruhrchemie, DE-AS 2933919, 1979 (W. Bernhagen, J. Weber, H. Bahrmann, H. Springer).
26. Showa Denko, EP 724 908, 1995, (K. Morikawa et al.). BASF, DE 19 500 236, 1995 (R. Dostalek et al.).
27. H. Adkins, *Org. React.* **8** (1954) 1. Eastman Kodak, WO 8 900 886, 1989 (B. L. Gustajson, P. S. Wehner, P. L. Mercer, G. O. Nelson). Ube Industries, DE 4 021 230, 1989 (S. Furusaki, T. Matsuzaki, Y. Yamasaki). Ube Industries, WO 9 510 497, 1995 (S. Furusaki, M. Matsuda, Y. Miyamoto, Y. Shiomi). Bayer, EP 721 928, 1995 (G. Darsow, G. M. Petruck, H. J. Alpers). BASF, WO 9 731 883, 1997 (K. G. Baur, R. Fischer, R. Pinkos, F. Stein, B. Breitscheidel, H. Rust). BASF, WO 9 731 882, 1997 (K. G. Baur, R. Fischer, R. Pinkos, F. Stein, H. Rust, B. Breitscheidel).
28. BASF, EP 295 435, 1987 (H. Mueller, H. Toussaint, J. Schossig).
29. L. Ré, B. Maurer, G. Ohlhoff, *Helv. Chim. Acta* **56** (1973) 1882.
30. Hüls, DE 3 843 956, 1988 (W. Fuhrmann, G. Bub, H. M. Zur). Sagami Chemical Research and Arakawa Chemicals, JP 9 132 541, 1996 (T. Fuchigami, T. Ga, N, Wakassa, K. Tawara).
31. BASF, DE 2 132 547, 1971 (L. Schuster, W. Franzischka, H. Hoffmann, S. Winderl). Institut Français du Pétrole, DE 2 809 995, 1977 (J. Gzillard, C. Lassau). Mitsui Toatsu Chemicals, JP 7 035 300, 1970 (C. Sas, H. Hiari). Honsu Chemical, JP 61 260 034, 1985 (H. Kasamatsu, M. Okada, T. Kimoto). Shinnittetsu Kagaku, JP 6 128 182, 1992 (T. Shimizu, M. Nagano, M. Furumoto). Shin Nippon Rika, JP 6 329 569, 1992 (N. Okajima, M. Nakazawa).
32. Institut Français du Pétrole, DE-OS 2809995, 1977 (J. Gzillard, C. Lassau).
33. Mitsui Toatsu Chemicals, JP 7035300, 1970 (C. Sasa, H. Hirai).
34. Tevco, US 4 301 046, 1981 (M. L. Schlossman). Henkel, DE 2 526 312, 1976 (H. Moeller, C. Gloxhuber, H. J. Thimm). Henkel, DE 2 526 675, 1976 (H. Moeller, H. Schnegelberger, C. Gloxhuber, H. J. Thimm). N. Wilke, *Chem. Ztg.* **95** (1971) 16 – 21. Du Pont, US 3 027 398, 1960 (W. L. Foohey). Eastman Kodak, US 3 334 149, 1964 (G. A. Akin, H. Lewis, T. F. Reid). Eastman Chemical, US 5 414 159, 1993 (P. Appleton, M. A. Wood). Eastman Chemical, US 5 406 004, 1993 (P. H. D. Eastland, J. Scarlett, M. W. M. Tuck, M. A. Wood). Eastman Chemical, US 5395 987, 1993 (C. Rathmell, R. C. Spratt, M. W. M. Tuck). Eastman Chemical, US 5 395 986, 1993 (J. Scarlett, M. A. Wood). Arakawa Chemical, JP 6 228 028, 1993 (H. Tsuji, T. Okazaki, K. Azuma). Towa Kasei, JP 61 292 146, 1992 (M. Magara, Y. Onoda, F. Yamazaki, S. Yoneda, K. Kato).
35. Toyobo, JP 75 142 537, 1975 (T. Mizumoto, H. Kamatani). Toyobo, JP 75 130 738 1976 (T. Mizumoto, H. Kamatani). Eastman Chemical, US 5559 159, 1995 (B. J. Sublett, G. W. Connell).
36. Eastman Kodak, US 3 668 157, 1972 (R. L. Combs, R. T. Bogan). S. Oswitch, *Reinf. Plast.* **17** (1973) 308. E. H. G. Sargent, K. A. Evans, *Plastics* **34** (1969) 721.
37. UCB, DE 2 542 191, 1976 (G. Slickx). Hüls, DE 25 454 800, 1976 (J. Rueter, H. Scholten). Towa Kasei, JP 2 053 881, 1988 (N. Okamoto, Y. Tateno, Y. Ishii, K. Kato).
38. Hüls, DE 2 241 413, 1974 (D. Stoye, W. Andrejewski, A. Draexler). Hüls, EP 790 266, 1996 (E. Wolf). PPG Industries, US 4 859 743, 1988 (R. R. Amborse, J. B. O'Dwyer, B. K. Johnston, D. P. Zielinski, S. Porter, W. H. Tyger). BASF, EP 262 069, 1986 (T. K. Debroy, M. Guagliardo, J. G. Pucknat). Toyo Boseki, JP 6 073 274, 1992 (H. Konishi, K. Fukuda, M. Sakaguchi). Akzo, EP 454 219, 1990 (A. Noomen, J. W. F. L. Seetz, H. Klinkenberg).
39. D. J. Golob, T. A. Odom, R. L. Whitson, *Polym. Mater. Sci. Eng.* **63** (1990) 826 – 832. Shanco Plastics & Chemicals, US 3 979 352, 1976 (J. W. Brady, F. D. Strickland, C. C. Longwith). L. Gott, *J. Coat. Technol.* **48** (1976) 52. Bayer, EP 566 953, 1992 (J. Pedain, H. Mueller, D. Mager, M. Schoenfelder).

Mitsui Toatsu Chemicals, JP 03 195 717, 1989 (M. Sakai, K. Sasaoka, M. Murata, M. Masayuki).
40. Eastman Kodak, US 3334149, 1964, DE 1202269, 1961 (G. A. Akin, H. Lewis, T. F. Reid).
41. W. M. Weigert (ed.): *Wasserstoffperoxid und seine Derivate,* Hüthig Verlag, Heidelberg 1978.
42. Degussa, DE 2937840, 1981 (G. Käbisch, H. Malitius, S. Raupach, R. Trübe, H. Wittmann).
43. Janssen Pharmaceutica N.V., DE 2551560, 1975 (G. van Reet, J. Heeres, L. Wals).
44. Degussa, Henkel, DE 2937831, 1981 (G. Käbisch, H. Malitius, S. Raupach, R. Trübe, H. Wittmann).
45. Celanese, US 3183274, 1955 (M. Robeson).
46. Koei Chemical, JP 21512, 1958 (M. Yamada, H. Terada).
47. Bayer, DE 1041027, 1949 (H. Danzinger, K. Haeseler, G. Schulze).
48. Trojan Powder, US 2468718, 1949 (J. Wyler).
49. J. Jelinek, M. Sýkora, CS 154421, 1973.
50. BASF, DE 2507461, 1975 (F. Merger, S. Winderl, H. Touissant).
51. Celanese, GB 816208 – 9 – 10, 1959.
52. Bayer, DE 1031298, 1958 (K. Bauer, H. Danzinger, G. Schulze).
53. Eastman Kodak, US 3956406, 1974 (B. Pallmer, D. Bondurant).
54. Perstorp, DE 1075102, 1960 (S. Sveninge).
55. Hoechst Celanese, US 5 603 835, 1994(H. Cheung, R. W Laurel, G. C. Seaman).
56. Perstorp, WO 9 701 523, 1996 (B. Wickberg).
57. Celanese, US 2790837, 1954 (M. Robeson).
58. Mitsubishi Gas Chemical, EP 799 815, 1997(T. Ninomiya, T. Watanabe, T. Ikebe, A. Iwamoto).
59. B. Tollens, *Ber. Dtsch. Chem. Ges.* **15** (1882) 1629.
60. E. Berlow, R. H. Barth, J. E. Snow: *The Pentaery-thritols,* Reinhold Publ., New York 1958.
61. Mitsui Toatsu Chemicals, JP 77100415 and JP 77105119, 1977 (T. Kiyoura).
62. P. Werle, G. Nonnenmacher, K. Kruse, *Liebigs Ann. Chem.* 1980, 938.
63. Koei Chemical Co., JP 6818888, 1965 (Y. N. Minato, S. N. Yasuda).
64. Hercules Powder, US 2533737, 1950 (E. Mertz).
65. Virginia Chemicals, US 4277620, 1980 (F. Gupton, H. Ulmer).
66. G. Giesselmann in *Methodicum Chimicum,* **vol. 1,** Thieme Verlag, Stuttgart 1973, p. 254.
67. K. Callmer, *J. Chromatogr.* **115** (1975) 397.
68. R. Nutiu, M. Maties, M. Nutiu, *J. Synth. Lubr.* **7** (1990) 145.
69. Henkel, EP 272 575, 1987(K. H. Schmid, U. Ploog, A. Meffert).
70. CS 197 741, 1983(G. Guba, V. Macho, L. Koudelka, L. Komora).
71. Degussa, DE 19 708 695, 1998, (M. Höpp, D. Arntz, M. Morawietz).
72. DEA, EP 0 499 994, 1992(H.-D. Grasshoff, V. Synek, H. Kolenz).
73. V. K. Rowe et al.: *Patty's Industrial Hygiene and T oxicology,* 3rd ed., **vol. 2C,** Wiley-Intersciene, New York 1982, p. 3874.
74. BASF Aktiengesellschaft (unpublished results 1959 – 1981).
75. L. Fischer et al., *Z. Gesamte Exp. Med.* **115** (1949) 22.
76. A. Loeser, *Pharmazie* **4** (1949) 263.
77. Registry of Toxic Effects of Chemical Substances, NIOSH. U.S. Dept. of Health, Education and Welfare, Public Health Service, Center for Disease Control, Cincinnati, Ohio, 1980.
78. H. C. Hodge, J. H. Sterner, *Am. Ind. Hyg. Assoc. Q.* **10** (1949) 93 – 96.
79. Eastman Kodak, Rochester, N.Y., unpublished results, Health, Safety, and Human Factors Laboratory.
80. G. D. DiVincenzo, D. A. Ziegler, *Toxicol. Appl. Pharmacol.* **52** (1980) no. 1, 10 – 15.
81. NIOSH, Information Profile on Potential Occupational Hazards: Glycols, PB 89-215776 (1982).
82. V. V. Stankevich, *Gig. Sanit.* **32** (1967) 107.
83. Degussa AG, unpublished results.
84. S. J. Plitman, *Gig. Sanit.* **36** (1971) 192 – 196.
85. W. Kutscher, *Z. Physiol. Chem.* **283** (1948) 268 – 275.
86. E. F. King, H. A. Painter, Assessment of Biodegradability of Chemicals in Water by Manometric Respirometry, Final Report, Contract No W/81/217, Commission of the European Communities, Directorate-General, Environment, Consumer protection and Nuclear Safety, 1983.
87. BASF, DE 19 542 036, 1995, (D. Kratz et al.)
88. BASF, DE 19 653 093, 1996, (D. Kratz et al.)

Aldehyde Polymers → Polyoxymethylenes

Aldehydes, Aliphatic and Araliphatic

Separate keywords: → Acetaldehyde, → Acrolein and Methacrolein, → Benzaldehyde, → Butanals, → Chloroacetaldehydes, → Crotonaldehyde and Crotonic Acid, → Formaldehyde, → Glyoxal, → Propanal

Christian Kohlpaintner, Celanese GmbH, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany

Markus Schulte, Celanese GmbH, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany

Jürgen Falbe, Henkel KGaA, Düsseldorf, Federal Republic of Germany

Peter Lappe, Ruhrchemie AG, Oberhausen, Federal Republic of Germany

Jürgen Weber, Ruhrchemie AG, Oberhausen, Federal Republic of Germany

1. Introduction 2
2. Saturated Aldehydes 2
2.1. Physical Properties 2
2.2. Chemical Properties 4
2.3. Production 7
2.4. Individual Saturated C_5 – C_{13} Aldehydes 9
2.4.1. C_5 Aldehydes 9
2.4.2. C_6 Aldehydes 10
2.4.3. C_7 Aldehydes 10
2.4.4. C_8 Aldehydes 11
2.4.5. C_9 Aldehydes 11
2.4.6. C_{10} Aldehydes 11
2.4.7. C_{11} Aldehydes 12
2.4.8. C_{12} Aldehydes 12
2.4.9. C_{13} Aldehydes 12
3. Unsaturated Aldehydes 14
3.1. Physical Properties 14
3.2. Chemical Properties 14
3.3. Production 15
3.4. Individual Unsaturated C_5 – C_{11} Aldehydes 16
3.4.1. C_5 Alkenals 16
3.4.2. C_6 Alkenals 16
3.4.3. C_8 Alkenals 17
3.4.4. C_{10} Alkenals 17
3.4.5. C_{11} Alkenals 17
4. Hydroxyaldehydes 18
4.1. Properties 18
4.2. Production 19
4.3. Individual Hydroxyaldehydes 20
4.3.1. 3-Hydroxypropanal 20
4.3.2. 3-Hydroxybutanal (Acetaldol) 20
4.3.3. 4-Hydroxybutanal 20
4.3.4. 2,2-Dimethyl-3-hydroxypropanal (Hydroxypivaldehyde) 21
4.3.5. 2-Ethyl-3-hydroxyhexanal (Butyraldol) 21
4.3.6. 3,7-Dimethyl-7-hydroxyoctanal (Hydroxycitronellal) 21
5. Araliphatic Aldehydes 22
5.1. Properties 22
5.2. Production 23
5.3. Individual Araliphatic Aldehydes . 23
5.3.1. Phenylacetaldehyde 23
5.3.2. 4-Methylphenylacetaldehyde 24
5.3.3. 2-Phenylpropionaldehyde (Hydratropaldehyde) 24
5.3.4. 3-Phenylpropionaldehyde (Dihydrocinnamaldehyde) 25
5.3.5. 2-(4-Isobutylphenyl)propionaldehyde 25
5.3.6. 3-(4-Isopropylphenyl)-2-methylpropanal (Cyclamenaldehyde) 26
5.3.7. 3-(4-*tert*-Butylphenyl)-2-methylpropionaldehyde (Lilial) 26
5.3.8. Cinnamaldehyde (3-Phenyl-2-propenal) 26
5.3.9. α-Alkylcinnamaldehydes (2-Alkyl-3-phenyl-2-propenals) 27
6. Dialdehydes 27
6.1. Properties 27
6.2. Production 27
6.3. Individual Dialdehydes 28
6.3.1. Succinaldehyde (1,4-Butanedial) . . . 28
6.3.2. Glutardialdehyde (1,5-Pentanedial) . . 29
6.3.3. TCD Dialdehyde 29
7. Acetals 29
7.1. Properties 30
7.2. Production 32
7.3. Uses 32
8. Quality Control 33
9. Analysis 33
10. Storage, Transportation, and Environmental Regulations 34
11. Economic Aspects 34
12. References 35

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_321

1. Introduction

Aldehydes are represented by the general formula RCHO, where R can be hydrogen or an aliphatic, aromatic, or heterocyclic group.

According to IUPAC nomenclature, aldehydes are identified by the ending "al." However, many of them still are called by their common names.

Because of their high reactivity, aliphatic aldehydes are widely used as intermediates in organic synthesis. Sometimes the isolation of a pure aldehyde is very difficult; in such cases, stable derivatives or oligomers are prepared from which the aldehydes can be reisolated.

Aldehydes usually must be kept from contact with air and under certain circumstances must be stabilized during distillation, subsequent storage, and transportation. This applies particularly to unsaturated aldehydes, which have a tendency to polymerize (e.g., on contact with alkali). For commerical purposes aldehydes often are protected by the addition of stabilizers and antioxidants and by a nitrogen atmosphere. When handling aldehydes, care must be taken to prevent either the liquids or their vapors from coming into contact with respiratory organs, eyes, and skin. Gloves and safety glasses are absolutely necessary.

Aldehydes are obtained mainly via the oxo synthesis, by mild oxidation (dehydrogenation) of primary alcohols, and by special olefin oxidation processes. Low concentrations of aldehydes occur naturally in essential oils of various plants. Acetaldehyde is an intermediate product of alcohol fermentation; it is formed by decarboxylation of the intermediate pyruvic acid. Aldehydes also fulfill some important biological functions, e.g., 11-*cis*-retinal in the sight process or as pyridoxal in the transamination of amino acids. Their isolation from natural substances is of commercial significance only in a few cases, e.g., in the production of longer-chain fragrance aldehydes.

2. Saturated Aldehydes

The most important saturated aliphatic aldehydes are formaldehyde, acetaldehyde, propionaldehyde, and butyraldehydes; they are used as starting materials for organic chemicals and polymers (→ Formaldehyde, etc.). Some dialdehydes, e.g., glyoxal, are also commerical products (→ Glyoxal). The higher aliphatic aldehydes are used mostly for the production of alcohols, carboxylic acids, amines, etc. The aldehydes also serve as fragrances in the perfume industry.

2.1. Physical Properties (see Table 1)

Formaldehyde, the simplest saturated aliphatic aldehyde, is a gas at room temperature. The aldehydes up to about C_{12} are liquids; the straight-chain aldehydes from C_{13} and above are solids. Because the hydrogen atom of a formyl group has less tendency to hydrogen bond than the hydrogen atom of a hydroxy group, the boiling points of the aldehydes are considerably lower than those of the corresponding alcohols. The difference in boiling points between aldehydes and alcohols with an intermediate number of carbon atoms is 20 – 40 °C.

Viscosity, density, and refractive index at 20 °C increase with increasing molecular mass. The lower homologues are mobile liquids; aldehydes from heptanal to undecanal have an oily consistency.

There are practically no limits to the miscibility of formaldehyde and acetaldehyde with water. However, with increasing molecular mass the solubility of aldehydes decreases very rapidly. For example, the solubility of hexanal in water is only 0.6 wt % at 20 °C. The aliphatic aldehydes are soluble in alcohols, ethers, and other common organic solvents.

Whereas the lower aldehydes have a pungent smell, the higher homologues in the $C_8 - C_{13}$ range are components of nearly all perfumes and many fragrances. Their odors become weaker with increasing molecular mass.

The characteristic properties of most aliphatic aldehydes are their ease of autoxidation and their tendency to trimerize and/or polymerize. Therefore, aldehydes are protected with an

Table 1. Physical properties of saturated aldehydes

	Name	CAS registry number	Formula	M_r	*mp*, °C	*bp* [a], °C	Density (20 °C), g/cm^3	Refract. index, n_D^{20}	Surface tension (20 °C), mN/m	Viscosity (20 °C), mPa · s	Solubility in H_2O at 20 °C, wt %
C_1	Methanal (formaldehyde)	[*50-00-0*]	HCHO	30.03	−92	-21	0.816 [b]				95
C_2	Ethanal (acetaldehyde)	[*75-07-0*]	CH_3CHO	44.05	−123	20	0.778	1.331		0.2	
C_3	Propanal (propionaldehyde)	[*123-38-6*]	CH_3CH_2CHO	58.08	−81	48	0.797	1.362	21.8	0.4	35
C_4	Butanal (butyraldehyde)	[*123-72-8*]	$CH_3(CH_2)_2CHO$	72.11	−97	75	0.803	1.380	24.6	0.45	7.6
	2-Methylpropanal (isobutyraldehyde)	[*78-84-2*]	$CH_3CH(CH_3)CHO$	72.11	−66	64	0.789	1.373	23.2	0.45	6.7
C_5	Pentanal (valeraldehyde)	[*110-62-3*]	$CH_3(CH_2)_3CHO$	86.13	−91	103	0.808	1.394	27.4	0.54	1.4
	2-Methylbutanal	[*96-17-3*]	$CH_3CH_2CH(CH_3)CHO$	86.13	<−60	92	0.804	1.390		0.55	1.7
	3-Methylbutanal (isovaleraldehyde)	[*590-86-3*]	$CH_3CH(CH_3)CH_2CHO$	86.13	<−70	92,5	0.796	1.388	23.2	0.58	1.4
	2,2-Dimethylpropanal (pivaldehyde)	[*630-19-3*]	$(CH_3)_3CCHO$	86.13	6	74	0.783	1.379		0.67	1.1
C_6	Hexanal (capraldehyde)	[*66-25-1*]	$CH_3(CH_2)_4CHO$	100.16	−56	128	0.814	1.404			0.6
	2-Methylpentanal	[*123-15-9*]	$CH_3(CH_2)_2CH(CH_3)CHO$	100.16	−100	118	0.807	1.400		0.56	0.4
	3-Methylpentanal	[*15877-57-3*]	$CH_3CH_2CH(CH_3)CH_2CHO$	100.16		38–40 (4.3)	0.826	1.407			
	4-Methylpentanal	[*1119-16-0*]	$CH_3CH(CH_3)(CH_2)_2CHO$	100.16							
	2-Ethylbutanal	[*97-96-1*]	$CH_3CH_2CH(C_2H_5)CHO$	100.16	−89	117	0.813	1.402	25.5	0.66	0.5
	2,3-Dimethylbutanal	[*2109-98-0*]	$CH_3CH(CH_3)CH(CH_3)CHO$	100.16		69–72 (21.3)	0.813	1.403			0.6
	3,3-Dimethylbutanal	[*2987-16-8*]	$CH_3C(CH_3)_2CH_2CHO$	100.16							
C_7	Heptanal (enanthaldehyde)	[*111-71-7*]	$CH_3(CH_2)_5CHO$	114.18	−42	153	0.817	1.411			
	2-Methylhexanal	[*925-54-2*]	$CH_3(CH_2)_3CH(CH_3)CHO$	114.18		141 (96.7)		1.4088–1.4100 (25 °C)			
	2,2-Dimethylpentanal	[*14250-88-5*]	$CH_3(CH_2)_2C(CH_3)_2CHO$	114.18		126–127	0.807	1.406			
	2,3-Dimethylpentanal	[*32749-94-3*]	$CH_3CH_2CH(CH_3)CH(CH_3)CHO$	114.18		141	0.845				0.4
C_8	Octanal (caprylaldehyde)	[*124-13-0*]	$CH_3(CH_2)_6CHO$	128.22	−23	175	0.821	1.419			
	2-Methylheptanal	[*16630-91-4*]	$CH_3(CH_2)_4CH(CH_3)CHO$	128.22		51 (1.5) (22 °C)	0.816 (22 °C)	1.414			
	2-Ethylhexanal	[*123-05-7*]	$CH_3(CH_2)_3CH(C_2H_5)CHO$	128.22	<−60	163	0.820	1.416		1.0	<0.02
	2-Propylpentanal	[*18295-59-5*]	$CH_3(CH_2)_2CH(C_3H_7)CHO$	128.22							
	2-Ethyl-3-methylpentanal	[*42347-74-0*]	$CH_3CH_2CH(CH_3)CH(C_2H_5)CHO$	128.22			0.807	1.405			

Table 1. (Continued)

	Name	CAS registry number	Formula	M_r	*mp*, °C	*bp*[a], °C	Density (20 °C), g/cm^3	Refract. index, n_D^{20}	Surface tension (20 °C), mN/m	Viscosity (20 °C), mPa · s	Solubility in H_2O at 20 °C, wt %
C_9	Nonanal (pelargonaldehyde)	[*124-19-6*]	$CH_3(CH_2)_7CHO$	142.24		191	0.825	1.427			
	3,5,5-Trimethylhexanal	[*5435-64-3*]	$CH_3C(CH_3)_2CH_2CH(CH_3)CH_2CHO$	142.24	<−60	173	0.820	1.421	25.5	1.4	<0.01
C_{10}	Decanal (caprinaldehyde)	[*112-31-2*]	$CH_3(CH_2)_8CHO$	156.27	−3	217	0.826	1.428		1.8	
C_{11}	Undecanal	[*112-44-7*]	$CH_3(CH_2)_9CHO$	170.29		120–122 (2.7)		1.433			
	2-Methyldecanal	[*19009-56-4*]	$CH_3(CH_2)_7CH(CH_3)CHO$	170.29	<−60	229	0.830	1.430	31.9	2.5	<0.1
C_{12}	Dodecanal (lauric aldehyde)	[*112-54-9*]	$CH_3(CH_2)_{10}CHO$	184.33	12	254	0.828	1.435		2.8	
	2-Methylundecanal	[*110-41-8*]	$CH_3(CH_2)_8CH(CH_3)CHO$	184.33	−25	246	0.827	1.433	32.4		<0.1
C_{13}	Tridecanal	[*10486-19-8*]	$CH_3(CH_2)_{11}CHO$	198.34	14	126 (3.1)					

[a] at 101.3 kPa unless otherwise specified;
[b] effective density at −19 °C

inert gas atmosphere and minimal amounts of a stabilizer are added if required [1, p. 651].

Because of the differences in boiling points between aldehydes and alcohols, aldehydes prepared by dehydrogenation of the corresponding alcohols generally are separated and purified by distillation. Apart from the lower boiling, short-chain aldehydes, distillation is mostly carried out at reduced pressure [1, p. 654]. Numerous aldehydes form azeotropic mixtures with water (see Table 2) and other substances.

Table 2. Azeotropic mixtures of aldehydes with water [9]

Aldehyde	*bp* of aldehyde, °C	*bp* of azeotrope, °C	wt% H_2O in azeotrope
Propanal	47.9	47.5	2.0
Butanal	74.8	68.0	9.7
2-Methylpropanal	63.3	60.1	9.6
Pentanal	103.3	83.0	19.0
3-Methylbutanal	92.5	77.0	12.0
Hexanal	128.3	91.0	31.3
2-Ethylbutanal	116.7	87.5	23.7
2-Methylpentanal	118.3	88.5	23.0
2-Ethylhexanal	163.6	96.4	51.6

2.2. Chemical Properties

The polarity of the carbonyl group of aldehydes not only facilitates the typical aldehyde reactions – addition of nucleophiles, reduction, and oxidation – but it also makes the α-hydrogen atom acidic. For these reasons, aldehydes can undergo a wide variety of reactions. The major ones are compiled in Table 3 [1, p. 609].

Addition Reactions. Because of the polarity of the carbonyl double bond, aldehydes enter into a wide variety of nucleophilic addition reactions. The simplest case is the intermolecular addition of one molecule of aldehyde to another to form an aldol. These *aldols* (β-hydroxyaldehydes) generally are unstable and react further to form secondary products, i.e., diols, unsaturated aldehydes, or alcohols.

Industrially, crossed aldol condensation of an aldehyde with formaldehyde is used for the production of trimethylolpropane, pentaerythritol, and neopentyl glycol (→ Alcohols, Polyhydric). The aldol condensation of butanal for the production of the plasticizer alcohol 2-ethylhexanol is of major commercial importance (→ 2-Ethylhexanol).

Table 3. General reactions of aliphatic aldehydes

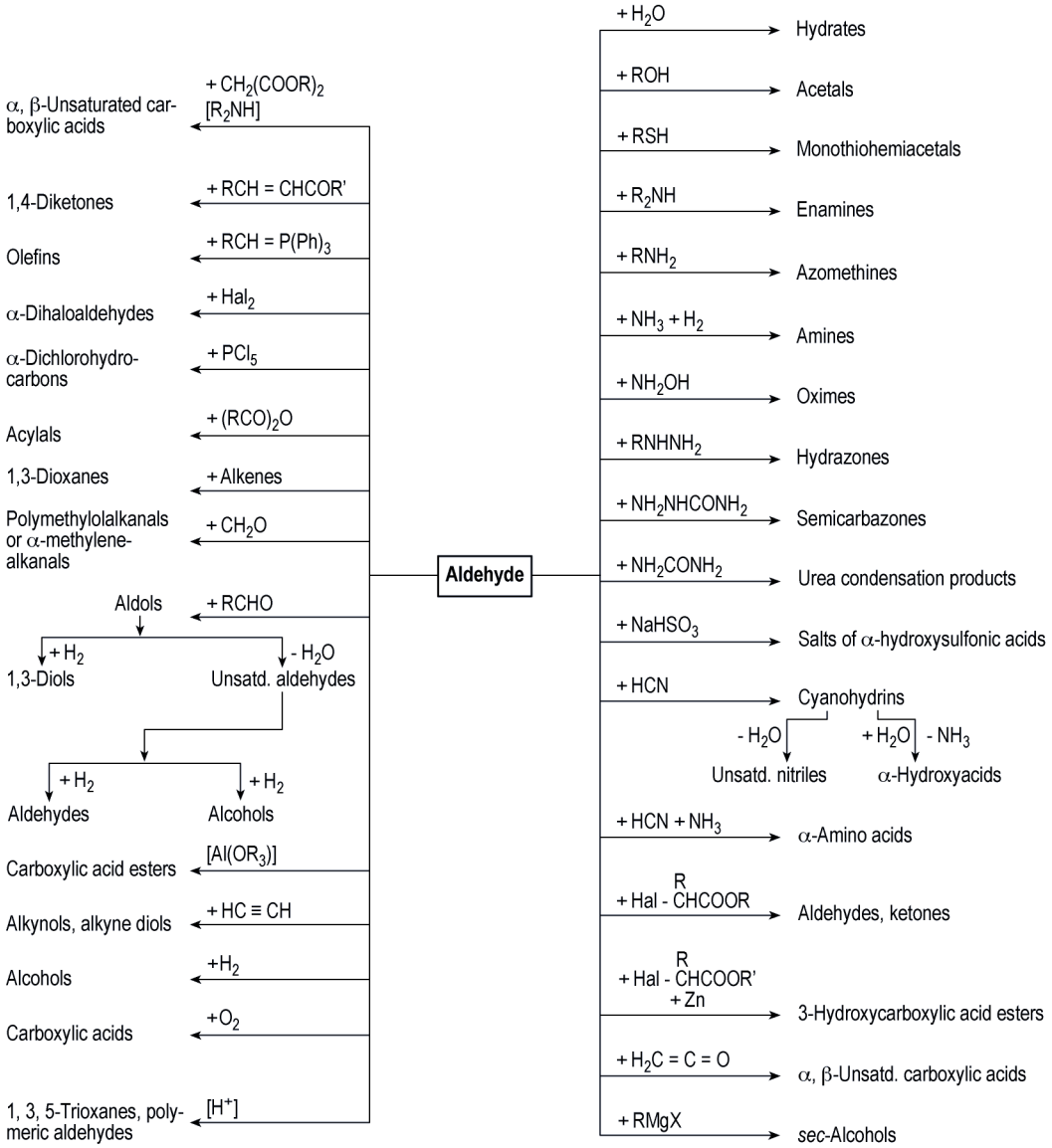

The addition of alcohols and thiols to aldehydes leads to acetals and monothiohemiacetals; these are used for the protection of formyl groups in synthesis because they are resistant to alkali. Reductive amination of aldehydes with ammonia or primary amines in the presence of hydrogen or some other reducing agent gives primary or secondary amines, respectively, e.g.,

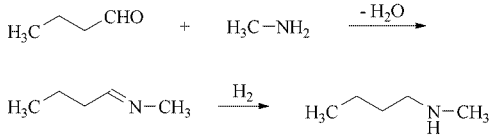

The reaction with secondary amines yields the corresponding enamines, which can be hydrogenated to tertiary amines.

Classically, aldehydes were identified by the preparation of crystalline derivatives, e.g., by reaction with hydrazine, substituted hydrazines (e.g., 2,4-dinitrophenylhydrazine), hydroxylamine, or semicarbazide. However, now other techniques are used (see Chap. 9).

The addition of Grignard compounds leads to secondary alcohols. The Darzens condensation of alkyl chloroacetates in the presence of a strong base yields aldehydes containing one additional methylene group, e.g.,

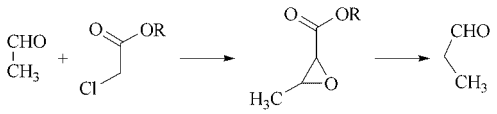

With the alkylhaloacetates in the presence of zinc, 3-hydroxycarboxylates are obtained (Reformatsky reaction).

The addition of sodium hydrogensulfite leads to water-soluble crystalline compounds. This reaction permits the separation of α-methyl-branched aldehydes from the isomeric straight-chain aldehydes [10], [11]. The reaction of hydrogen cyanide with aldehydes is of commercial significance. The highly unstable cyanohydrins can be converted to the unsaturated nitriles by dehydration or to the α-hydroxycarboxylic acids by hydrolysis. If the reaction of hydrogen cyanide with aldehydes is carried out in the presence of ammonia, α-amino acids are obtained.

Polymerization. The formation of cyclic, 2,4,6-trialkyl-1,3,5-trioxane type trimers is catalyzed by various acids.

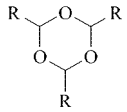

At higher temperature these compounds are unstable and revert to the monomer. With this process pure *n*-alkanals can be isolated from *n*–*iso*-aldehyde mixtures [12].

Mainly oligomers and polymers are formed by the lower aldehydes. Polymeric formaldehyde, referred to as paraformaldehyde, is a mixture of polyoxymethylene glycols $HO(CH_2O)_nH$, where n is 8–100. The oligomers of acetaldehyde, paraldehyde and metaldehyde are 2,4,6-trimethyl-1,3,5-trioxane and the cyclic tetramer, respectively. Polyacetaldehyde is a high-molecular mass polymer with an acetal structure.

Oxidation. A characteristic feature of most aldehydes is their great tendency to autoxidize in a radical chain reaction to the corresponding carboxylic acids. On an industrial scale the oxidation is usually carried out in the liquid phase with oxygen or air. Catalysts often are added to reduce the reaction time and lower the reaction temperature. Salts of transition metals are effective catalysts. For special purposes, hydrogen peroxide, periodic acid, nitric acid, potassium permanganate, chromium trioxide, and peroxo compounds are used as oxidizing agents [1, p. 609]. Another special method is the melting of aldehydes with alkali, e.g., [13].

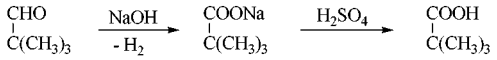

Hydrogenation (Reduction). Catalytic hydrogenation leads to primary alcohols. Both Ni and Cu catalysts have proved to be the most suitable [14]. Normally, the reaction is carried out in the liquid phase on fixed-bed catalysts at 20–200 °C and pressures of up to 30 MPa (→ Alcohols, Aliphatic, Chap. 2.3.3.). Hydrogenation in the gas phase is run continuously. Because of the good heat dissipation this is advantageous, especially with sensitive starting materials.

In addition to catalytic hydrogenation there are many other reduction processes. Complex hydrides such as lithium aluminum hydride and sodium borohydride are used most frequently as reducing agents.

An important reaction characteristic of aldehydes is the Meerwein–Ponndorf–Verley reduction with aluminum alkoxides, in particular with aluminum triisopropoxide. Because this reaction is reversible the carbonyl compound formed is distilled from the alkoxide. The process is recommended for the preparation of alcohols containing halo or nitro groups.

There are numerous other reducing agents for aldehydes; most of these are of interest only for laboratory syntheses. A few

are: potassium – ammonia, trimethyl phosphite, magnesium, aluminum – mercury, trialkylboranes, and aluminum trialkyls.

The reduction of aldehydes to the corresponding saturated hydrocarbons is of no commercial significance. In laboratory syntheses it is achieved with reducing agents such as hydrazine, complex metal hydrides, lithium – ammonia, or hydrogen iodide – phosphorus.

2.3. Production

Although many aldehyde syntheses are known, only a few are used on an industrial scale. Often this is a question of the availability of feedstock.

The most important processes for the preparation of saturated aliphatic aldehydes are:

1) Hydroformylation of olefins (oxo synthesis)
2) Dehydrogenation or oxidation of primary alcohols (mainly for the production of formaldehyde from methanol)
3) Hydration of acetylene for the production of acetaldehyde (→ Acetaldehyde)
4) Oxidation of ethylene to acetaldehyde (→ Acetaldehyde)
5) Oxidation of saturated hydrocarbons (C_3, C_4) for the preparation of lower aldehydes

In addition, some special syntheses for the production of aldehydes required in the perfume industry are of industrial importance.

Oxo Synthesis. The oxo synthesis is the most important process for the production of aldehydes containing at least three carbon atoms (→ Oxo Synthesis). In this process olefins react with synthesis gas (CO, H_2) to form aldehydes with one more carbon atom than the starting material. Pure products are formed only from symmetrical or sterically hindered olefin molecules; otherwise, product mixtures of straight-chain and branched compounds are obtained:

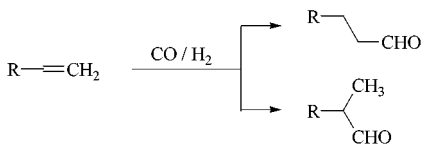

By the selection of suitable catalysts and reaction conditions the *n* : *iso* ratio can be varied over a wide range.

Dehydrogenation/Oxidation of Primary Alcohols. Dehydrogenation, oxidation, and oxidative dehydrogenation proceed according to the following equations:

$$RCH_2OH \rightarrow RCHO + H_2$$
$$\Delta H = +\,84\,\text{kJ/mol for R} = CH_3 \quad (1)$$
$$RCH_2OH + \tfrac{1}{2}\,O_2 \rightarrow RCHO + H_2O$$
$$\Delta H = -\,159\,\text{kJ/mol for R} = H \quad (2)$$
$$RCH_2OH \rightarrow RCHO + H_2\,, H_2 + \tfrac{1}{2}\,O_2 \rightarrow H_2O$$
$$\Delta H = -\,159\,\text{kJ/mol for R} = H \quad (3)$$

Dehydrogenation. The endothermic dehydrogenation reaction of alcohols is carried out at atmospheric pressure and 250 – 400 °C, normally with Cu or Ag catalysts. The catalysts often are activated by the addition of, e.g., Zr, Co, or Cr. The advantage of this process is the simultaneous recovery of hydrogen that can be used without further purification [2, p. 282]. Catalytic dehydrogenation is an equilibrium reaction. Therefore, high temperatures and short residence times are economically advantageous.

The process still has some commercial importance in the preparation of acetaldehyde from ethanol. The gas-phase dehydrogenation is carried out at atmospheric pressure and 270 – 300 °C on a copper catalyst activated with cerium. Twenty-five to 50 % of the ethanol is converted per throughput, with a selectivity of 90 – 95 % acetaldehyde; ethyl acetate, ethylene, crotonaldehyde, and higher alcohols are obtained as byproducts (for further details → Acetaldehyde).

Oxidation. The oxidation according to Equation (2) is carried out with an excess of air or oxygen at 350 – 450 °C over a catalyst containing 18 – 19 wt % Fe_2O_3 and 81 – 82 wt % MoO_3 [15]. This process is used in the production of formaldehyde (→ Formaldehyde).

Oxidative Dehydrogenation. Equation (3) can be separated formally into the endothermic dehydrogenation of the alcohol and the exothermic combustion of the hydrogen formed; the overall reaction therefore can become exothermic. In the industrial process the two reactions take place simultaneously when substoichiometric quantities of oxygen or air are used. In

oxidation and oxidative dehydrogenation the explosion ranges of the alcohol – air mixtures must be considered.

Oxidative dehydrogenation is the most important process for the production of aldehydes from alcohols. Silver catalysts are preferred but copper catalysts are used also.

In formaldehyde production from methanol, 75 – 99 % conversion is achieved on silver crystals (grain size 0.2 – 3 mm), silver nets, or silver on Al_2O_3. Temperatures of 500 – 720 °C and residence times of less than 0.01 s are used (→ Formaldehyde).

In 1979 about 15 % of Western Europe's acetaldehyde production was from ethanol. Silver and copper were the main catalysts employed. Ethanol conversions were 30 – 50 % per throughput, and acetaldehyde selectivity was 85 – 95 %. The reaction temperature ranged between 300 and 600 °C depending on the quantity of air. Ethyl acetate, formic acid, acetic acid, and carbon dioxide were byproducts (for more details → Acetaldehyde).

Fragrance Aldehydes. In addition to the production of formaldehyde and acetaldehyde, dehydrogenation and oxidation also are preferred for the synthesis of fragrance aldehydes. A process developed especially for these products permits the catalytic dehydrogenation of $C_5 - C_{14}$ alcohols in the presence of hydrogen and air; it employs copper or silver catalysts, possibly combined with Zn, Cr, Cr_2O_3 [16]. Another catalyst system based on Cu/MgO, produces octanal with a selectivity of 99 % and an octanol conversion of 58 % at 265 – 330 °C [17]. Other catalyst systems described in the literature are Ag/Na_2O on supports [18], mixtures of MnO, NiO on MgO [19], or CuCl and a nitrogen-containing ligand, such as 2,2′-bipyridyl [20]. For further dehydrogenation catalysts see [2, p. 282].

Oxidation of Hydrocarbons. A process developed by Celanese for the oxidation of C_3 and C_4 alkanes [21] produces a complex reaction mixture requiring costly extraction and distillation steps. In this process propane and propane–butane mixtures are reacted in the gas phase at 425 – 460 °C, 0.7 – 0.8 MPa, and with a conversion of about 20 % (with oxygen deficiency). The reaction proceeds according to a radical mechanism. In addition to acetaldehyde (≈20 %) the reaction mixture consists mainly of formaldehyde (≈15 %), methanol (≈19 %), and organic acids (≈11 %). However, the process is outdated technically. The oxidation of methane or ethane also has no practical significance.

Oxidation of Olefins. The most important industrial process for the production of acetaldehyde is the partial oxidation of ethylene in the aqueous phase in the presence of palladium and copper chlorides (Wacker-Hoechst process, → Acetaldehyde, Chap. 4.3.).

Another oxidation process for the preparation of aldehydes from olefins is the acrolein synthesis from propene (→ Acrolein and Methacrolein, Chap. 3.1.).

Miscellaneous Processes. The process involving addition of water to acetylene for the preparation of acetaldehyde has become more or less insignificant compared with the alternative processes based on ethylene or ethanol (→ Acetaldehyde). In Western Europe the last plants were closed in 1980. Apart from the ready availability of the cheaper ethylene and the good selectivity of its conversion, the use of environmentally detrimental mercury sulfate as catalyst has led to the acetylene process becoming essentially obsolete.

The β-hydroxyaldehydes (aldols) and the resulting α,β-unsaturated aldehydes are of commercial significance as starting materials for the production of saturated aldehydes. Examples are the productions of butanal from crotonaldehyde (→ Butanals) and 2-ethylhexanal from 2-ethyl-2-hexenal (see Section 2.4.4).

Laboratory Processes. Only a few examples of the extraordinarily large number of aldehyde syntheses commonly used on a laboratory scale are mentioned here.

The catalytic reduction of acyl chlorides (Rosenmund reduction) is one of the safest and simplest methods of converting carboxylic acids to aldehydes. Mainly Pd – C or Pd – $BaSO_4$ serve as catalysts, if necessary in the presence of a catalyst poison containing sulfur in an aromatic hydrocarbon solvent, such as benzene or toluene. The yields average between 70 and 90 % [1, p. 418], [22].

Other reducing agents are complex borohydrides or aluminum hydrides, such as

bis(triphenylphosphine)copper(I) tetrahydroborate, $[(C_6H_5)_3P]_2CuBH_4$ [23], and lithium tri-*tert*-butyloxyaluminum hydride [24].

Carboxylic acid esters, amides, hydrazides, and nitriles also can be reduced in many ways to the corresponding aldehydes. Most often complex metal hydrido compounds are used as reducing agents.

The oxidative cleavage of olefins with ozone via the ozonides and subsequent reductive decomposition generally take place under extremely mild reaction conditions. With this process, almost quantitive yields are attained [25].

Apart from the classic processes of oxidative glycol cleavage with lead(IV) acetate (Criegee method) [26] and with periodic acid (Malaprade method) [27], there are a number of other oxidizing agents, such as hydrogen peroxide-ruthenium(III) salts [28], manganese dioxide [29], and *N*-iodosuccinimide [30].

When Grignard compounds react with orthoformate esters, aliphatic or aromatic aldehydes are formed [31]. Allylvinyl ethers, which are easily accessible by the reaction of aldehydes with allyl alcohols, rearrange to 4-alkenals (Claisen rearrangement) when heated; the products subsequently can be hydrogenated to the saturated aldehydes [32].

Comprehensive lists of the various preparative methods are given in [1], [2, p. 203], [5, p. 943], [33].

2.4. Individual Saturated $C_5 - C_{13}$ Aldehydes

The higher aldehydes are hydrogenated mainly to alcohols, which are used as solvents or for the production of plasticizers. In addition, they are intermediates for the production of carboxylic acids, amines, and amino acids. They also are used for the synthesis of agricultural chemicals, pharmaceuticals, disinfectants, dyes, etc. In the perfume industry they are used either as such or for the synthesis of odorants such as α-amylcinnamaldehyde.

2.4.1. C_5 Aldehydes

The most important physical properties of the four isomeric C_5 aldehydes, their CAS registry numbers, and their (low) solubilities in water are compiled in Table 1. They are colorless, flammable liquids with intense, characteristic odors, and they exhibit the typical reactions of aliphatic aldehydes. They form azeotropic mixtures with water and are freely soluble in common organic solvents.

Production. Pentanal and 2-methylbutanal are prepared by the hydroformylation of butene [34]:

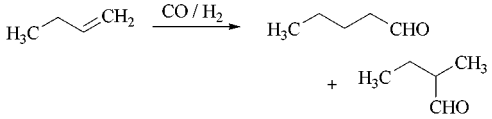

The ratio of the two products depends on the reaction conditions [1, p. 195].

Isomeric mixtures of pentanal and 2-methylbutanal are often commercial products. Also 2-methylbutanal can be prepared by the reaction of butanal with formaldehyde in the presence of secondary amines with subsequent hydrogenation [35]. 3-Methylbutanal (isovaleraldehyde) is prepared by hydroformylation of isobutene. Small quantities (≈5 %) of 2,2-dimethylpropanal are formed also [36].

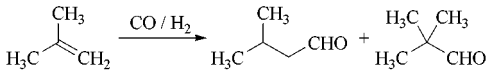

2,2-Dimethylpropanal can be obtained by isomerization of 1,1,2-trimethyloxirane at 80 – 170 °C in the presence of $ZnCl_2$-pumice [37].

Another way to manufacture 3-methylbutanal is by the isomerization of 3-methyl-3-butene-1-ol or of a mixture of 3-methyl-3-butene-1-ol and 3-methyl-2-butene-1-ol, prepared from isobutene and formaldehyde. The isomerization catalyst is CuO – ZnO and the reaction temperature is 200 – 250 °C [38].

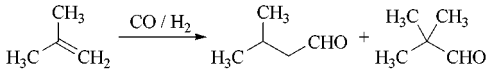

Uses. The C_5 aldehydes are hydrogenated easily to pentanols (amyl alcohols), which are

used chiefly as solvents and for the production of plasticizers, esters, and xanthogenates (→ Pentanols). Zinc diamyldithiophosphate is an additive for lubricating oils.

Further applications of the C_5 aldehydes are in the oxidation to C_5 acids (→ Carboxylic Acids, Aliphatic) and the preparation of amyl amines by aminating hydrogenation. 2-Methylbutanol (from 2-methylbutanal) can be converted to isoprene over dehydration catalysts at 300 – 350 °C [39] and to isoleucine (→ Amino Acids).

3-Methylbutanal is a starting material for 2,3-dimethyl-2-butene, which in turn is converted to 2,3-dimethylbutane-2,3-diol (pinacol) [40] and to methyl *tert*-butyl ketone (pinacolone) [41]. Pinacolone is a valuable starting material for the synthesis of a number of pesticides.

Pharmaceuticals, such as the active substance butizide [42], are synthesized from 3-methylbutanal and the corresponding acid.

The most important manufacturers of C_5 aldehydes in Western Europe are BASF and Ruhrchemie, and in the United States, Union Carbide.

2.4.2. C_6 Aldehydes

The C_6 aldehydes are colorless, mobile liquids. Their most important physical properties and CAS registry numbers are compiled in Table 1. Hexanal occurs naturally, e.g., in lemon and orange oils.

Production. The preparation of the C_6 aldehydes, of which hexanal, 2-methylpentanal, 2-ethylbutanal, and 3-methylpentanal have medium importance, is carried out almost exclusively by hydroformylation or by aldol condensation. For example, hexanal and 3-methylpentanal are prepared by the hydroformylation of 1-pentene [43] and 2-methyl-1-butene [44], respectively. 2-Methylpentanal is obtained by aldol condensation of propanal, and 2-ethylbutanal by the reaction of butanal with acetaldehyde; in both cases the intermediate unsaturated aldehyde must be hydrogenated subsequently.

Uses. The C_6 aldehydes can easily be hydrogenated or oxidized to the corresponding alcohols or carboxylic acids. 2-Methylpentanal and 3-methylpentanal are starting materials for tranquilizers, such as meprobamate and mebutamate [44], [45]. 2-Methylpentanal also is used for the manufacture of various other products, including agrochemicals, perfumes, and catalysts for cross-linking of polyesters.

2.4.3. C_7 Aldehydes

The C_7 aldehydes are colorless, mobile liquids; all form azeotropic mixtures with water. Heptanal (enanthal) occurs naturally in ginger oil. The most important physical data of the C_7 aldehydes and the CAS registry numbers are given in Table 1.

Production. Heptanal, together with undecanoic acid, is obtained by the pyrolytic cleavage of ricinoleic acid esters [46] or by the hydroformylation of 1-hexene in the presence of modified rhodium catalysts. In the latter case, the straight-chain aldehyde is formed almost exclusively [47]. A minor byproduct is 2-methylhexanal, which can be prepared specifically by reaction of hexanal with formaldehyde followed by hydrogenation [35]. 2,3-Dimethylpentanal is prepared by the hydroformylation of 3-methyl-2-pentene [48] or by the amine-catalyzed reaction of 3-methylpentanal with formaldehyde and subsequent hydrogenation [44]:

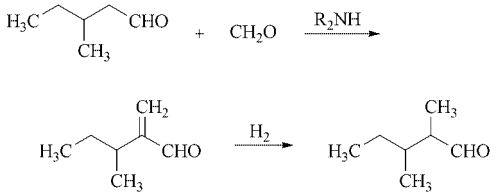

Uses. Heptanal is used principally for the preparation of α-amylcinnamaldehyde, a popular fragrance chemical especially in the field of soap perfumery. It is also a starting material for heptanoic acid, esters of which have applications in lubricants. Manufacturers of the aldehyde and/or the acid are Celanese and Monsanto in the United States and ATO Chimie in Western Europe.

2,3-Dimethylpentanal is required for the synthesis of some pharmaceuticals.

2.4.4. C_8 Aldehydes

The physical data and the CAS registry numbers of the most important C_8 aldehydes are presented in Table 1. The C_8 aldehydes form azeotropic mixtures with water, as do most aliphatic aldehydes, and are miscible with the common organic solvents. Octanal occurs naturally in various citrus oils. Of the C_8 aldehydes, octanal, isooctanal, and 2-ethylhexanal have industrial importance.

Production. Octanal (caprylaldehyde) can be obtained by hydroformylation of 1-heptene [49] or by dehydrogenation of octanol [50]. Isooctanal is used as a mixture of isomers and is prepared by hydroformylation of a heptene mixture, the product of codimerization of propene and butene.

The base-catalyzed aldol condensation of butanal leads to 2-ethyl-2-hexenal, which is converted to 2-ethylhexanal by hydrogenation:

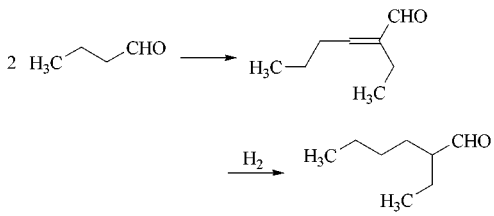

2-Propylpentanal can be prepared either by hydroformylation of 3-heptene [51] or in a multistep synthesis from pentanal [52].

Uses. Octanal is employed in the perfume industry for the preparation of synthetic citrus oils and for the synthesis of α-hexylcinnamaldehyde.

Isooctanal is used in the perfume industry and is also catalytically hydrogenated to isooctanol, which is processed further to plasticizers. The aldol reaction of isooctanal with subsequent hydrogenation leads to isohexadecanol, which can be used for the manufacture of synthetic lubricants and hydraulic fluids. 2-Ethylhexanal is used for the manufacture of 2-ethylhexanol (→ 2-Ethylhexanol) and 2-ethylhexanoic acid (→ Carboxylic Acids, Aliphatic). Aminating hydrogenation leads to 2-ethylhexylamines, which are important intermediates in organic syntheses. The condensation of 2-ethylhexanal with aromatic amines results in products that are used as vulcanizing agents and antioxidants for rubber. 2-Propylpentanoic acid (dipropylacetic acid) and its derivatives are prepared from 2-propylpentanal. The sodium salt is an important antiepileptic.

2.4.5. C_9 Aldehydes

Nonanal (pelargonaldehyde) occurs in several natural oils, such as cinnamon oil, lemon grass oil, citrus oil, and rose oils. For physical data and CAS registry numbers see Table 1. The most important C_9 aldehydes are the isomeric mixtures obtained by hydroformylation of commercial C_8 olefin mixtures.

Production. Nonanal is prepared by catalytic dehydrogenation of nonanol [53], by hydroformylation of 1-octene [54], or by reaction of formic acid and nonanoic acid on a titanium dioxide catalyst. Isononanal, prepared on an industrial scale by hydroformylation of commercial diisobutylene, is an isomeric mixture containing about 95 % 3,5,5-trimethylhexanal [55], [56]:

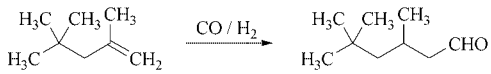

In addition to diisobutylene, other C_8 olefin mixtures are obtained by dimerization of isobutene and butene and are used as feedstock for the production of a number of isononanal mixtures.

Uses. Both nonanal and isononanal are used in the perfume industry. The various isononanals can be converted to isononanol mixtures, which serve as starting materials for the preparation of plasticizers. Upon oxidation isononanoic acid (→ Carboxylic Acids, Aliphatic) is obtained.

Isononanal is an antimicrobial agent [56] and is used for the synthesis of dyestuffs.

2.4.6. C_{10} Aldehydes

Some physical data and the CAS registry numbers are given in Table 1. Decanal is a colorless liquid that is immiscible with water. It has an odor similar to that of orange peel.

Production. Decanal is prepared either by hydroformylation of 1-nonene [57] or by dehydrogenation of 1-decanol on a copper catalyst [17]. 2-Methylnonanal is formed as a byproduct of the hydroformylation of 1-nonene. It can also be prepared by the reaction of nonanal and formaldehyde followed by hydrogenation. Mixtures of *iso*-C_{10} aldehydes are obtained by hydroformylation of commercial C_9 olefins, e.g., tripropylene.

Uses. Besides the use of *iso*-C_{10} aldehyde mixtures as intermediates for isodecanol and isodecanoic acid, C_{10} aldehydes serve primarily as perfumes and fragrance chemicals. The main use of decanal is for citrus tones and for the manufacture of synthetic citrus oils. In addition, the C_{10} aldehydes have value as intermediates in the synthesis of pharmaceuticals and in the polymer and pesticide fields.

2.4.7. C_{11} Aldehydes

Of the C_{11} aldehydes, undecanal and 2-methyldecanal (methyloctylacetaldehyde), possess some commercial value. When pure they are colorless liquids. They have a green, fatty character that blends well with woody and mossy perfume formulations. Undecanal occurs naturally in citrus oils and has a waxy, floral odor. The most important physical data and the CAS registry numbers are presented in Table 1. The aldehydes form azeotropic mixtures with water and are freely soluble in the common organic solvents.

Production. Undecanal is prepared by hydroformylation of decene [57], [58] or by catalytic dehydrogenation of undecanol [17]. Another method is the reduction and subsequent partial hydrogenation of the acid chloride of undecylenic acid, which in turn can be obtained by acidic cleavage of castor oil. 2-Methyldecanal is obtained together with undecanal by the hydroformylation of decene or by the reaction of decanal with formaldehyde and subsequent partial hydrogenation [35], [59]. It is formed also by the Darzens reaction of alkyl chloroacetates with 2-decanone to form glycidic esters, followed by saponification and acidulation [60].

The two isomeric aldehydes can be separated by using the bisulfite adduct [10].

Uses. Undecanal and 2-methyldecanal are used in the perfume industry. Other applications are the syntheses of pharmaceuticals, fungicides, plant growth regulators, bactericides, and disinfectants.

2.4.8. C_{12} Aldehydes

Only dodecanal (lauric aldehyde) occurs naturally in several essential oils, such as rue oil, orange peel oil, and citrus oil. Dodecanal and 2-methylundecanal, because of their pleasant odors, are among the most important toning agents in the aroma industry. The two isomeric aldehydes are colorless liquids, immiscible with water, but freely soluble in most organic solvents. Physical properties and CAS registry numbers are given in Table 1.

Production. Dodecanal is prepared by the catalytic dehydrogenation of dodecanol [61] or by the reduction of the corresponding acid chloride with sodium tetrahydridoborate [62]. 2-Methylundecanal can be obtained by hydroformylation of undecene [58]. Alternatively, the mixture of undecanal and 2-methyldecanal, produced by oxo synthesis, reacts with formaldehyde, followed by partial hydrogenation and distillation [63]. 2-Methylundecanal also is formed from 2-undecanone by the Darzens reaction [60].

2.4.9. C_{13} Aldehydes

The only commercially available C_{13} aldehydes are isomeric mixtures, the composition of which depends primarily on the nature of the olefins that are hydroformylated. Physical properties and CAS registry numbers are given in Table 1.

Production. The C_{12} olefins formed by the oligomerization of butene or propene are converted to the corresponding isomeric tridecanals by hydroformylation. Ziegler α-olefins, olefins from wax cleavage, and olefins produced by metathesis (Shell's SHOP process) also are used as feedstocks.

Table 4. Physical properties of unsaturated aldehydes

	Name	CAS registry number	Formula	M_r	*mp*, °C	*bp* ∗, °C	Density (20 °C), g/cm^3	Refractive index, n_D^{20}	Surface tension (20 °C), mN/m	Viscosity (20 °C), mPa s	Solubility in H_2O at 20 °C, wt%
C_3	Propenal (acrolein)	[*107-02-8*]	CH_2=CHCHO	56.07	−86	53	0.841	1.401			20.6
C_4	*trans*-2-Butenal (crotonaldehyde)	[*123-73-9*]	CH_3CH=CHCHO	70.09	−74	102	0.852	1.437			15.5
	2-Methylpropenal (methacrolein)	[*78-85-3*]	CH_2=C(CH_3)CHO	70.09	−81	68	0.843	1.416		0.5	5.9
C_5	2-Methyl-2-butenal (tiglic aldehyde)	[*497-03-0*]	CH_3CH=C(CH_3)CHO	84.13		117 (73.8)	0.871	1.448			
	2-Pentenal	[*1576-87-0*]	CH_3CH_2CH=CHCHO	84.13		125	0.848	1.443			
C_6	2-Hexenal	[*6728-26-3*]	$CH_3(CH_2)_2CH$=CHCHO	98.15		146	0.842	1.446			
	2-Methyl-2-pentenal	[*623-36-9*]	CH_3CH_2CH=C(CH_3)CHO	98.15	<−60	138	0.854	1.449		0.8	1.4
	2-Isopropylpropenal	[*4417-80-5*]	$CH_3CH(CH_3)CH(CHO)=CH_2$	98.15							
	2-Ethyl-2-butenal	[*19780-25-7*]	CH_3CH=C(C_2H_5)CHO	98.15	<−60	135	0.857	1.440		0.7	
C_8	2-Ethyl-2-hexenal	[*645-62-5*]	$CH_3(CH_2)_2CH$=C(C_2H_5)CHO	126.21	<−70	175	0.852	1.453		1.28	
C_{10}	3,7-Dimethyl-6-octenal (citronellal)	[*106-23-0*]	$(CH_3)_2C=CH(CH_2)_2CH(CH_3)CH_2CHO$	154.26		207–208	0.851	1.448	29.26	1.93	
	3,7-Dimethyl-2,6-octadienal (citral)	[*5392-40-5*]	$(CH_3)_2C=CH(CH_2)_2C(CH_3)$=CHCHO	152.24							
	(2*E*)−3,7-Dimethyl-2,6-octadienal (citral a, geranial)	[*141-27-5*]	$(CH_3)_2C=CH(CH_2)_2C(CH_3)$=CHCHO	152.24		229	0.889	1.490			
	(2*Z*)−3,7-Dimethyl-2,6-octadienal (citral b, neral)	[*106-26-3*]	$(CH_3)_2C=CH(CH_2)_2C(CH_3)$=CHCHO	152.24		120 (2)	0.887	1.487			
C_{11}	10-Undecanal (undecylene aldehyde)	[*112-45-8*]	$CH_2=CH(CH_2)_8CHO$	168.28		103 (0.4)	0.850 (21 °C)	1.446 (21 °C)			

∗ at 101.3 kPa unless otherwise specified in parentheses.

Uses. Tridecanals are used in the perfume and aroma industries. Their catalytic reduction leads to tridecanols, which are of considerable importance as plasticizer and detergent alcohols. Salts, esters, and other derivatives of isotridecanoic acid, prepared by oxidation of the aldehyde, are of value in the paint and plastics industries.

3. Unsaturated Aldehydes

Acrolein is the most important of the unsaturated aldehydes (→ Acrolein and Methacrolein); crotonaldehyde also is manufactured industrially (→ Crotonaldehyde and Crotonic Acid). Besides these, 2-ethyl-2-hexenal (see Section 3.4.3), citral and citronellal (see Section 3.4.4), and undecylene aldehyde (see Section 3.4.5) are important unsaturated aldehydes.

3.1. Physical Properties

The most important physical data for a number of unsaturated aliphatic aldehydes are summarized in Table 4. The α,β-unsaturated aldehydes in the commercially significant C_3-C_{10} range are liquids. Their boiling points are, without exception, higher than those of the corresponding saturated aldehydes. They resemble the saturated aldehydes in that their viscosities, densities, and refractive indices increase with increasing molecular mass. The compounds are only moderately soluble in water; the solubilities decrease with increasing molecular mass. In common organic solvents, solubilities of the unsaturated aldehydes are high. The lower homologues have a pungent smell and have an irritating effect on mucous membranes. However, the higher unsaturated compounds are often used in the preparation of perfumes and aromas.

Most α,β-unsaturated aldehydes autoxidize easily and tend to polymerize, especially in the presence of base. Small amounts of stabilizers, e.g., hydroquinone, are nearly always added for these reasons.

Many α,β-unsaturated aldehydes form azeotropes with water and other substances. Data for some of these are compiled in Table 5.

Table 5. Azeotropic mixtures of unsaturated aldehydes with water [9]

Aldehyde	*bp* of aldehyde, °C	*bp* of azeotrope, °C	wt% H_2O in azeotrope
Acrolein	53	52.4	2.6
Methacrolein	68	63.6	7.7
Crotonaldehyde	102.4	84	24.8
2-Ethyl-2-butenal	135.3	92.7	38
2-Methyl-2-pentenal	138.2	93.5	40
2-Ethyl-2-hexenal	176	97.6	60.9

Like their saturated counterparts, α,β-unsaturated aldehydes can be unequivocally characterized by IR, ^{1}H NMR, ^{13}C NMR, and by UV/visible spectroscopy.

3.2. Chemical Properties

In addition to the typical reactions of saturated aldehydes already described, the chemical properties of the α,β-unsaturated aldehydes are determined by the conjugation of the olefinic double bond with the carbonyl function. This leads to the formation of 1,2 and 1,4 adducts as a result of nucleophilic addition reactions:

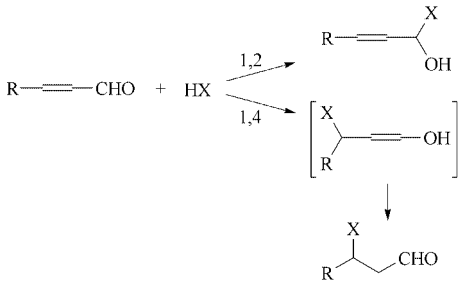

The resulting products play an important role in preparative organic chemistry. A further characteristic reaction of α,β-unsaturated aldehydes is the Diels – Alder conversion, e.g., to alkyl-substituted 2-formyl-2,3-dihydropyranes [64]:

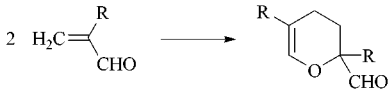

Dienes also can enter into Diels – Alder reactions with unsaturated aldehydes:

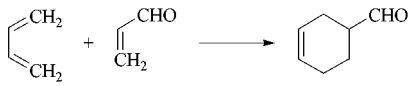

Of special importance for the handling of α,β-unsaturated aldehydes is the previously mentioned tendency toward polymerization, which proceeds in a highly exothermic manner in the presence of base.

Each of the two functionalities present in the molecule, the olefinic double bond and the carbonyl group, can be selectively hydrogenated by a suitable choice of catalyst. For example, the olefinic double bond is hydrogenated selectively in the presence of noble-metal catalysts and the carbonyl group in the presence of modified Pt, Ru, or Os catalysts or by a Meerwein – Ponndorf – Verley reaction with aluminum alkoxides to yield the corresponding saturated aldehydes and α,β-unsaturated alcohols, respectively. The saturated alcohols are obtained by hydrogenation over nickel or copper catalysts.

3.3. Production

Although the lower, commercially important α,β-unsaturated aldehydes, such as acrolein, crotonaldehyde, or 2-ethyl-2-hexenal, are obtained exclusively by synthesis, a number of essential oils serve as raw materials for higher homologues, such as for citral or citronellal.

There are two industrial processes for the manufacture of α,β-unsaturated aldehydes:

1) Oxidation of olefins (preparation of acrolein)
2) Dehydration of aldols obtained by aldol condensation of saturated aldehydes (preparation of crotonaldehyde and 2-ethyl-2-hexenal)

In addition there are some special syntheses for the preparation of aldehydes used in the perfume industry such as dehydrogenation of unsaturated alcohols (preparation of citral from geraniol), and reduction of unsaturated acids (undecylenealdehyde from undecylenic acid).

Oxidation of Olefins. The direct oxidation of olefins is industrially important for the manufacture of acrolein from propene. The conversion proceeds at 300 – 480 °C over variously modified Bi – Mo oxide catalysts. The feedstock is a gaseous mixture of propene, air, and water vapor in a molar ratio of about 1 : 10 : 2. Conversion of up to 98 % and acrolein yields of 78 – 92 % are realized. Byproducts include acetaldehyde, acetic acid, and acrylic acid. Commercial acrolein has a purity of 95 – 97 %. The stabilizer, which is added in all processing steps, is usually hydroquinone (for further details → Acrolein and Methacrolein).

Dehydration of Aldols. The β-hydroxyaldehydes (aldols), see Chapter 4 that are formed as intermediates in the aldol reaction are extremely unstable and decompose with loss of water to form α,β-unsaturated aldehydes. Depending on the reaction conditions, the aldol reaction can lead directly to the unsaturated compound. This process is used principally for the synthesis of crotonaldehyde from acetaldehyde, 2-methyl-2-pentenal from propionaldehyde, and 2-ethyl-2-hexenal from butanal.

The aldol condensation of acetaldehyde with catalytic amounts of dilute sodium hydroxide is carried out at 20 – 25 °C. The reaction mixture is treated with acetic acid to stop the reaction. In the subsequent first distillation, water is cleaved from the acetaldol. The selectivity towards crotonaldehyde reaches values up to 95 %.

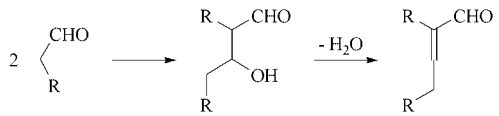

Not only can two identical aldehyde molecules undergo aldol condensation but also two different aldehyde species can do so. In the latter case a mixture of products usually results. By carefully choosing the reactants and reaction conditions, however, it is possible to obtain the desired compound as the main product. For example, acrolein was formerly prepared by the reaction of formaldehyde with acetaldehyde:

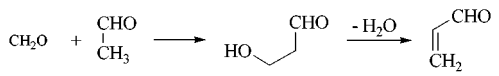

In the same way higher 2-methylenealkanals (2-

alkylacroleins) can be obtained by the reaction of formaldehyde with longer chain aldehydes.

Miscellaneous Processes. For the preparation of some unsaturated aroma aldehydes, the corresponding alcohols are selectively dehydrogenated over copper, copper–zinc, or noble metal catalysts. These processes preferably are carried out under reduced pressure and find application in the manufacture of citral, citronellal, and hydroxycitronellal.

Another method of industrial interest for the synthesis of unsaturated aldehydes is the Claisen rearrangement of allyl vinyl ethers, which are formed as shown in the following reaction scheme [32]:

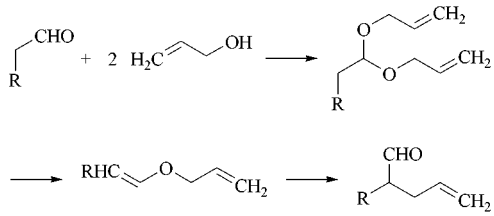

2-Alkenals are obtained by the reaction of unsaturated alkyl halides with the sodium salts of secondary nitrohydrocarbons, e.g., citral in 80 % yield from 1-halo-3,7-dimethylocta-2,6-diene [65]:

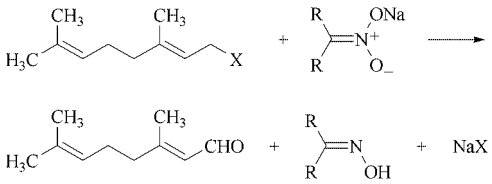

Another possibility is the treatment of acetals with vinyl ethers in the presence of boron trifluoride to give the corresponding β-alkoxyacetals. Under the influence of acids these are converted into α,β-unsaturated aldehydes [66].

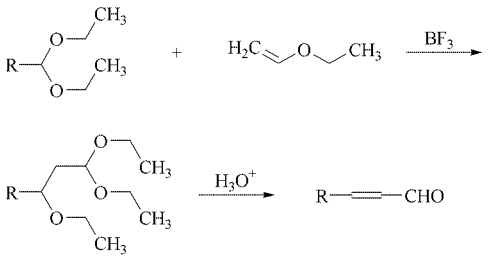

For a review of further synthetic methods, see [1], [2].

3.4. Individual Unsaturated C_5–C_{11} Aldehydes

3.4.1. C_5 Alkenals

Of the C_5 alkenals, only 2-methylcrotonaldehyde (tiglic aldehyde) has any significance.The most important physical data and CAS registry numbers of two C_5 alkenals are presented in Table 4. These compounds undergo most of the reactions characteristic of α,β-unsaturated aldehydes.

Production. Several processes for the preparation of tiglic aldehyde are described in the literature; for example, the condensation of acetaldehyde with propanal in the presence of sodium hydroxide [67]; the reaction of 3,4-epoxy-3–methylbutene with palladium acetylacetonate and triphenylphosphine [68]; the isomerization of 2-ethylacrolein [69]; or a two-step synthesis beginning with the catalytic reaction of isoprene, acetic acid, and oxygen to give 2-methyl-1,4-diacetoxy-2-butene, followed by hydrolysis catalyzed by cation-exchange resins [70].

Uses. Tiglic aldehyde is used mainly in the perfume industry. It is also a starting material for the preparation of terpenes, and the *syn*-oxime is used as a sweetener in food.

3.4.2. C_6 Alkenals

The most important physical data and CAS registry numbers for some unsaturated C_6 aldehydes are given in Table 4. The aldehydes are colorless, highly flammable liquids. A number of essential oils contain *trans*-2-hexenal. When diluted, it has an intense odor resembling that of apples.

Production. *trans*-2-Hexenal (leaf aldehyde) is prepared by reaction of butanal with vinyl ether in the presence of boron trifluoride and subsequent treatment of the reaction mixture with dilute sulfuric acid [71] (see next page).

2-Methyl-2-pentenal is formed by aldol condensation of propanal, 2-ethyl-2-butenal by aldol condensation of butanal and acetaldehyde,

and 2-isopropylacrolein by reaction of 3-methylbutanal with formaldehyde [40].

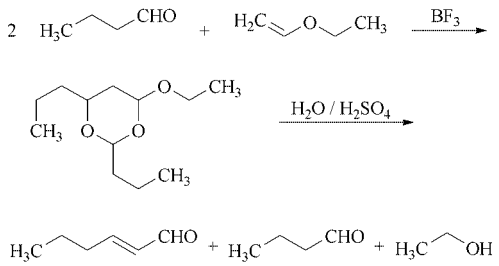

Uses. The perfume and aroma industries use *trans*-2-hexenal as a constituent of floral compositions, 2-methyl-2-pentenal for the manufacture of 2-methylpentanal, and 2-isopropylacrolein as a starting material for the preparation of 2,3-dimethyl-2-butene and pinacol [40]. 2-Ethyl-2-butenal is mainly used for the synthesis of 2-ethylbutyric acid.

3.4.3. C_8 Alkenals

Among the many isomeric C_8 alkenals, *2-ethyl-2-hexenal* is of outstanding economic and technical importance. It is a clear liquid with a characteristic odor and is soluble in practically all organic solvents, although nearly insoluble in water. The most important physical data and the CAS registry number are given in Table 4.

Production. Industrially, 2-ethyl-2-hexenal is prepared by aldol condensation of butanal and subsequent dehydration. The reaction is carried out at 80 – 130 °C and pressures up to 1 MPa in the presence of sodium hydroxide or a basic ion-exchange resin. Many other catalysts are described in the literature [72]. The conversion is over 99 %. The organic phase usually is subjected immediately to further processing to yield either 2-ethylhexanal over a palladium catalyst, or 2-ethyl-1-hexanol over a nickel or copper catalyst [73].

Uses. The main application for 2-ethyl-2-hexenal is the manufacture of 2-ethyl-1-hexanol, the most important plasticizer alcohol with a production capacity (1997) of nearly 10^6 t/a in Western Europe alone and 350 000 t/a in the United States. Condensation products of 2-ethyl-2-hexenal with various aromatic amines have proved to be good vulcanizing agents and antioxidants for rubber.

2-Ethylhexanal, obtainable by selective hydrogenation, is a starting material for the synthesis of the corresponding acid and amines and for the manufacture of pharmaceuticals.

3.4.4. C_{10} Alkenals

The unsaturated C_{10} aldehydes are used preferentially in perfumes and fragrances. Additionally they serve as starting materials for the synthesis of a whole series of terpenoid compounds. The most important representatives are citral a and b (*cis*- and *trans*-3,7-dimethyl-2,6-octadienal) and citronellal (3,7-dimethyl-6-octenal), which occur naturally in many essential oils. They are colorless liquids with odors resembling lemons or balm. Some physical properties and the CAS registry numbers are given in Table 4.

Several industrially established syntheses exist for citral a and b and for citronellal. The isolation from essential oils is now carried out on a large scale only for citronellal. For further details → Flavors and Fragrances.

Uses. Besides the application in the fragrance industry, citral is an intermediate in the vitamin A synthesis and in the preparation of β-ionones [74]. Its acetals, which are relatively stable toward alkali, also are used in the perfume industry. Citronellal is used in small amounts for scenting soaps and detergents. Its principal application is in the preparation of isopulegol, citronellol, and hydroxycitronellal. These compounds also have importance in the fragrance industry.

3.4.5. C_{11} Alkenals

10-Undecenal, undecylene aldehyde, the most important unsaturated C_{11} aldehyde and so far not known to occur naturally, is a colorless liquid with a stifling, flowery odor. For some physical properties and the CAS registry number see Table 4. Undecylene aldehyde is prepared by Rosenmund reduction of undecylenoyl chloride,

which in turn is obtained by cleavage of castor oil. A further possibility is the reaction of undecylenic acid with formic acid over a titanium oxide catalyst. This aldehyde is an important modulating agent in the perfume industry.

4. Hydroxyaldehydes

Of this class of compounds, some β-hydroxyaldehydes in particular are of commercial importance, primarily as intermediates for the preparation of 1,3-diols (see Sections 4.3.1, 4.3.2, and 4.3.4), and α,β-unsaturated aldehydes (see Section 3.3) and their derivatives.

The 2,2-bismethylol compounds of propanal and butanal and the trismethylol compound of acetaldehyde are intermediates in the manufacture of trimethylolethane, trimethylolpropane, and pentaerythritol, respectively (→ Alcohols, Polyhydric).

4.1. Properties

The boiling points of the β-hydroxyaldehydes (aldols) are well above those of the corresponding saturated compounds because of hydrogen bonding. The aldols obtained from lower aldehydes can be vacuum distilled without decomposition only if the salts, present from the condensation reaction, have been completely removed. They decompose upon heating at normal pressure, however [75].

The lower aldols are easily miscible with water and common organic solvents. Some physical data are compiled in Table 6.

For qualitative and quantitative analysis of hydroxyaldehydes, both the hydroxyl and the carbonyl groups can be exploited for the preparation of derivatives.

The chemical behavior of the hydroxyaldehydes is influenced primarily by the ease with which water is cleaved, leading to the formation of α,β-unsaturated aldehydes. This reaction, which is referred to as crotonization, is dependent on the structure of the starting material and on the reaction conditions. It is catalyzed by acids and bases.

Table 6. Physical properties of hydroxyaldehydes

Name	CAS registry number	Formula	M_r	*mp*, °C	*bp* (*p*), °C (kPa)	Density (20 °C), g/cm^3	Refractive index, n_D^{20}
2-Hydroxyethanal (glycol aldehyde)	[*141-46-8*]	$HOCH_2CHO$	60.05	97		1.366 (100 °C)	1.477 (19 °C)
3-Hydroxypropanal	[*2134-29-4*]	$HOCH_2CH_2CHO$	74.05		38 (0.03)		
3-Hydroxybutanal (acetaldol)	[*107-89-1*]	$CH_3CH(OH)CH_2CHO$	88.10		77 (3.0)	1.105	1.4238
4-Hydroxybutanal	[*25714-71-0*]	$HOCH_2(CH_2)_2CHO$	88.10		70 – 72 (1.5)	1.089	1.4450
2,2-Dimethyl-3-hydroxypropanal (hydroxypivaldehyde)	[*597-31-9*]	$HOCH_2C(CH_3)_2CHO$	102.14	ca. 70	ca. 141 (101.3)		
2-Ethyl-3-hydroxyhexanal (butyraldol)	[*496-03-7*]	$CH_3CH_2CH_2CH(OH)CH(C_2H_5)CHO$	144.22		98 – 101 (1.3)	0.9397	1.4426
3,7-Dimethyl-7-hydroxyoctanal (hydroxycitronellal)	[*107-75-5*]	$(CH_3)_2C(OH)(CH_2)_3CH(CH_3)CH_2CHO$	172.27		85 – 87 (0.13)	0.922	1.4488

The catalytic hydrogenation of β-hydroxyaldehydes leads to the formation of 1,3-diols. Important examples are the preparation of propane-1,3-diol from β-hydroxypropanal, 2-ethylhexane-1,3-diol from butyraldol, as well as the previously mentioned preparation of trimethylolethane, trimethylolpropane, and pentaerythritol. The last compound is also formed by reduction with formaldehyde (crossed Cannizzaro reaction).

Monomeric β-hydroxyaldehydes generally are not stable. They react with aldehydes to form so-called aldoxanes (**1**), dimerize to paraldols (**2**), or polymerize.

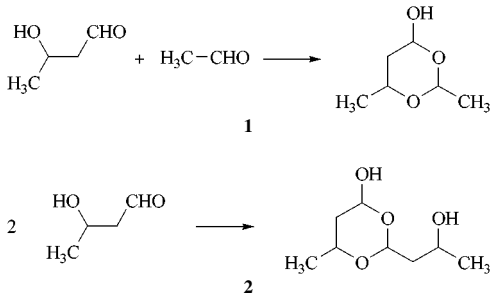

4.2. Production

Aldol Condensation. The aldol condensation, first mentioned in the literature by A. Wurtz in 1872, can be described in the case of a base-catalyzed reaction by the following mechanism [75]:

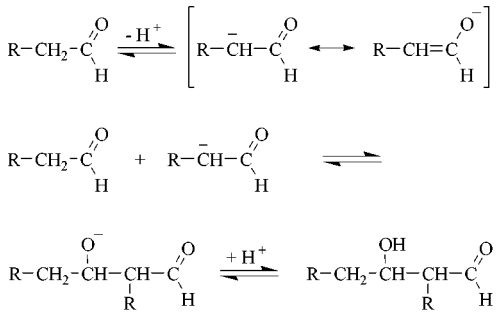

The aldol condensation is a reversible reaction. For the catalytic influence of amines and kinetic studies, cf. [76]; for a detailed description of the aldol reaction, cf. [77].

Aldol condensation, which is only possible for aldehydes with at least one α-hydrogen atom, can be catalyzed not only by bases, as shown above, but also by acids. In general, aldols obtained from aldehydes with more than one α-hydrogen atom can be isolated only at low temperature, because they readily lose water to form α,β-unsaturated aldehydes:

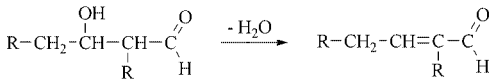

If two different aldehydes with α-hydrogen atoms are subjected to aldol addition, all four possible aldol species generally are formed in varying amounts.

Of more importance for industrial purposes are reactions in which only one reactant possesses α-hydrogen atoms; in this case the second reactant is frequently formaldehyde. For the preparation of mono- or polymethylolalkanals by reaction of formaldehyde with alkanals, cf. [75, p. 89] and Section 4.3.4.

These exothermic reactions are mostly carried out in the liquid phase, and aqueous solutions of sodium hydroxide or alkali carbonates usually serve as catalysts. Other catalyst systems have been described, e.g., zinc- or magnesium-containing zeolites [78], alkali hydroxides in combination with phase-transfer catalysts [79], tertiary amines [80], and basic ion exchange resins [81].

Typical byproducts of aldol condensation, apart from the dimeric aldols, aldoxanes, and α,β-unsaturated aldehydes mentioned earlier, are cyclic acetals, Tishchenko esters, etc. This is especially the case for reactions with formaldehyde, in which a reduction of the aldol by a crossed Cannizzaro reaction can occur [82].

G. Wittig, discovered a modification of the classic aldol reaction that involves the treatment of carbonyl compounds with metalated imines. In this case α,β-unsaturated aldehydes are practically the only products. The value of this procedure has been proved in the field of natural products [83], [84].

Miscellaneous Processes. The addition of water to α,β-unsaturated aldehydes yields β-hydroxyaldehydes. This reaction is used industrially for the preparation of 1,3-propanediol (see

Section 4.3.1). 3-Hydroxypropanal also can be obtained by the hydroformylation of ethylene oxide [85]. 4-Hydroxybutanal is obtained by hydroformylation of allyl alcohol [86]. A review of the hydroformylation of unsaturated alcohols to give hydroxyaldehydes has been published [1, p. 212].

If 2,3-dihydro-1,4-pyran is hydrolized, 5-hydroxypentanal is formed; however, it exists in equilibrium with its cyclic hemiacetal [87]. For example, 2-phenyl-2,3-dihydropyran, obtained by Diels – Alder synthesis from acrolein and styrene, can be cleaved by dilute sulfuric acid to give 5-phenyl-5-hydroxypentanal [88].

4.3. Individual Hydroxyaldehydes

4.3.1. 3-Hydroxypropanal

Some physical properties of 3-hydroxypropanal and its CAS registry number are given in Table 6. This compound is the intermediate in the preparation of 1,3-propanediol. It is extremely unstable and normally is not isolated as a pure substance. It can be distilled in high vacuum and characterized as its 2,4-dinitrophenylhydrazone (*mp* 132.5 – 133 °C).

Production. The addition of water to acrolein is the most important process for the production of 3-hydroxypropanal. A ca. 20 % aqueous solution of acrolein is reacted in the presence of a weakly acidic catalyst or an ion exchange resin [89], [90]. The solution can be either hydrogenated directly to 1,3-propanediol or, more advantageously, first subjected to an extraction with isobutanol (→ Propanediols).

Uses. 3-Hydroxypropanal is used exclusively for the production of 1,3-propanediol.

4.3.2. 3-Hydroxybutanal (Acetaldol)

Some physical properties of 3-hydroxybutanal (acetaldol) and the CAS registry number are given in Table 6. This compound is the lowest molecular mass aldol that can be prepared by aldol condensation. Pure monomeric acetaldol boils at 59 – 60 °C (1300 – 1500 Pa). Higher distillation temperatures lead to crotonaldehyde by dehydration and to acetaldehyde by reversal of the condensation. Polymerization of the aldol during its preparation from acetaldehyde can be prevented by addition of hydroquinone or pyrogallol. When pure, acetaldol polymerizes easily to paraldol; stabilization can be achieved by addition of small amounts of water or acetaldehyde.

Production. In the presence of dilute sodium or potassium hydroxide solutions, acetaldehyde is converted to the aldol over a period of several hours in a water-cooled flow tube at 20 – 25 °C. In order to avoid side and secondary reactions, which would result in appreciable resin formation, the reaction is stopped by adding phosphoric or acetic acid when acetaldehyde conversion has reached 60 %. After the unconverted acetaldehyde is separated, a bottom product is obtained that, apart from small amounts of byproducts such as crotonaldehyde, consists of 72 – 73 wt % acetaldol and ca. 18 wt % water [91].

Uses. Acetaldol is used primarily for the preparation of crotonaldehyde. The dehydration takes place during distillation upon addition of some acetic acid. Acetaldol can also be hydrogenated to 1,3-butanediol. The secondary products, butyraldehyde, butanol, and 1,3-butadiene, which at one time were prepared from acetaldol, are now manufactured almost exclusively by other processes.

Acetaldol is used also for the manufacture of substituted quinaldines which, in turn, are valuable starting materials for dyes and pigments [92].

4.3.3. 4-Hydroxybutanal

Some physical properties of 4-hydroxybutanal and its CAS registry number are given in Table 6. Because no conjugated system can be formed, this aldehyde does not tend to split off water, but forms its cyclic hemiacetal, 2-hydroxytetrahydrofuran. Only 4 % of the free hydroxyaldehyde exists in the equilibrium with the hemiacetal. For chemical characterization the compound is transformed into the 2,4-dinitrophenylhydrazone (*mp* 120 – 122 °C).

Production. Allyl alcohol, obtained from propylene oxide, is converted to 4-hydroxybutanal by rhodium-catalyzed hydroformylation. The formation of 3-hydroxy-2-methylpropanal can be minimized by using a large excess of phosphine ligands [86].

Uses. 4-Hydroxybutanal is used for the synthesis of tetrahydrofuran and 1,4-butanediol. This process was commercialized by Arco in 1990.

4.3.4. 2,2-Dimethyl-3-hydroxypropanal (Hydroxypivaldehyde)

2,2-Dimethyl-3-hydroxypropanal (hydroxypivaldehyde) is characterized by the aldehyde group, the primary alcohol group, and the neopentyl configuration.

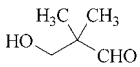

Although the functional groups provide possibilities for a whole range of reactions, the neopentyl structure is responsible for a noticeable stability toward hydrolysis, heat, and light in many derivatives.

Hydroxypivaldehyde is a colorless liquid that can be distilled in vacuum. Within a short time it dimerizes to a substituted 1,3-dioxane; this reaction is reversible [82], [93]. For some physical properties and the CAS registry number see Table 6.

Production. Hydroxypivaldehyde is prepared by the reaction at ca. 50 °C of isobutyraldehyde and a 30 – 37 % aqueous solution of formaldehyde. The catalysts used are normally aqueous solutions of alkali or alkaline-earth hydroxides [94], but the use of alkali carbonates [95], alcoholic alkali hydroxides [96], tertiary amines [80], and basic ion exchangers [81] also has been described. Because the two starting materials are inexpensive and readily available, other processes for the manufacture of hydroxypivaldehyde have been unable to achieve any industrial importance.

Uses. The most important use of hydroxypivaldehyde is the manufacture of neopentyl glycol, which can be prepared either by a crossed Cannizzaro reaction with formaldehyde or by catalytic hydrogenation (→ Alcohols, Polyhydric, Chap. 2.2.1.). Further products with hydroxypivaldehyde as starting material (cf. [82], [93]) are hydroxypivalic acid neopentyl glycol ester, hydroxypivalic acid, and pantolactone (α-hydroxy-β,β-dimethyl-γ-butyrolactone), which is used in the preparation of pharmaceuticals and vitamins [97].

4.3.5. 2-Ethyl-3-hydroxyhexanal (Butyraldol)

Some physical properties and the CAS registry number are given in Table 6.

As is the case with most aldols, 2-ethyl-3-hydroxyhexanal is not isolated as a pure compound but is converted directly under the prevailing reaction conditions to 2-ethyl-2-hexenal. The aldol can be characterized as the 2,4-dinitrophenylhydrazone derivative (*mp* 172 °C).

Production. 2-Ethyl-3-hydroxyhexanal is prepared primarily by aldol condensation of butanal at 30 °C in the presence of an aqueous sodium hydroxide solution and a phase-transfer catalyst; the reaction is stopped by addition of acetic acid [98].

Other catalysts described in the literature include zeolites containing magnesium or zinc [78] and amines, e.g., tri-, tetra-, or hexamethylenediamine [76].

For the preparation of 2-ethyl-2-hexenal without isolation of the aldol see Chapter 3.4.3.

Uses. 2-Ethyl-3-hydroxyhexanal is used for the preparation of 2-ethyl-2-hexenal, 2-ethylhexanol, 2-ethylhexanal, 2-ethylhexanoic acid, and 2-ethyl-1,3-hexanediol.

4.3.6. 3,7-Dimethyl-7-hydroxyoctanal (Hydroxycitronellal)

Some physical properties and the CAS registry number are given in Table 6.

Hydroxycitronellal is a colorless liquid of low viscosity. So far it has not been found in nature. Its odor is reminiscent of linden flowers and lillies of the valley.

Hydroxycitronellal is used in numerous perfumes. The acetals also occasionally find application as perfumes because of their higher stability toward soap. For preparation and further details, → Flavors and Fragrances.

5. Araliphatic Aldehydes

Araliphatic aldehydes are compounds with the general formula

$$Ar \leftarrow CH_2 \rightarrow_x CHO \ (x \geqq 1)$$

where Ar is some aromatic group, usually phenyl.

Although numerous araliphatic aldehydes occur naturally as components of essential oils, they are usually synthesized.

Araliphatic aldehydes resemble their aliphatic counterparts in that they readily undergo autoxidation and polymerization. Therefore, when stored, air should be excluded, and suitable stabilizers should be added.

Many araliphatic aldehydes and secondary products, such as the corresponding acetals, alcohols, and esters, are established compounds in the perfume industry. They are also starting materials for pharmaceuticals, agricultural chemicals, plasticizers, etc.

5.1. Properties

Physical Properties (see also Table 7). The industrially important araliphatic aldehydes are colorless to pale yellow liquids. They can be distilled, preferably under vacuum, without any appreciable decomposition. All have boiling points above 190 °C at atmospheric pressure, and the densities range between 0.95 and 1.05 g/cm^3. With the exception of phenylacetaldehyde which is slightly soluble in water, the higher araliphatic aldehydes are practically insoluble. The compounds are miscible with alcohols, ethers, and other common organic solvents.

Chemical Properties. Apart from their ease of oxidation, the reactivity of the araliphatic aldehydes is determined by the polarity of the carbonyl group, the resulting acidity of the α-hydrogen atoms (if available), and the aromatic

Table 7. Physical properties of araliphatic aldehydes

Name	CAS registry number	Formula	M_r	*mp*, °C	*bp* *, °C	Density (20 °C), g/cm^3	Refractive index, n_D^{20}
2-Phenylacetaldehyde	[*122-78-1*]	$C_6H_5CH_2CHO$	120.16		195	1.027	1.526
4-Methylphenylacetaldehyde	[*104-09-6*]	$CH_3C_6H_4CH_2CHO$	134.18		221 – 222	1.005	1.526
3-Phenylpropanal (dihydrocinnamaldehyde)	[*104-53-0*]	$C_6H_5CH_2CH_2CHO$	134.18		223 (74.5)	1.019	1.527
2-Phenylpropanal (hydratropic aldehyde)	[*93-53-8*]	$C_6H_5CH(CH_3)CHO$	134.18		202 – 205	1.009	1.518
2-(4-Isobutylphenyl)-propanal	[*51407-46-6*]	$(CH_3)_2CHCH_2C_6H_4CH(CH_3)CHO$	190.29				
3-(4-Isopropylphenyl)-2-methyl-propanal (cyclamenaldehyde)	[*103-95-7*]	$(CH_3)_2CHC_6H_4CH_2CH(CH_3)CHO$	190.29		108 – 108.5 (0.3)	0.950	1.507
3-(4-*tert*-Butylphenyl)-2-methylpropanal (lilial)	[*80-54-6*]	$(CH_3)_3C\text{-}C_6H_4\text{-}CH_2CH(CH_3)CHO$	204.31		126 – 127 (0.8)	0.939	1.505
Cinnamaldehyde (3-phenyl-2-propenal)	[*104-55-2*]	$C_6H_5CH{=}CHCHO$	132.17	−8	253	1.050	1.620
α-Amylcinnamaldehyde	[*122-40-7*]	$C_6H_5CH{=}C(C_5H_{11})CHO$	202.3		174 – 175 (2.0)	0.971	1.538
α-Hexylcinnamaldehyde	[*101-86-0*]	$C_6H_5CH{=}C(C_6H_{13})CHO$	216.32		174 – 176 (2.0)	0.950	1.527 (25 °C)
α-Methylcinnamaldehyde	[*101-39-3*]	$C_6H_5CH{=}C(CH_3)CHO$	146.19		150 (10.0)	1.041	1.606 (17 °C)

[a] at 101.3 kPa unless otherwise specified

substituents present in the molecule. These lead to reactions typical of aromatic systems in addition to the reactions of the aldehyde group (see Table 3).

In many cases the aldehyde functionality must be protected before carrying out aromatic electrophilic substitution. However, normally the substituent is already present in the aromatic system before the aldehyde is synthesized.

More important than reactions of the aryl ring are those of the aldehyde group, especially the formation of acetals, which are used in perfumes because of their comparative stability.

In addition to the saturated araliphatic aldehydes, some compounds with olefinic structure, e.g., cinnamaldehyde and α-amylcinnamaldehyde, are important as fragrances. If the aliphatic double bond is conjugated with the carbonyl group, the possibility of 1,4-addition exists (see Section 3.2).

5.2. Production

Most araliphatic aldehydes occur naturally, often widespread, but usually in low concentrations. Therefore, their isolation from natural sources is of relatively little industrial value.

An extraordinarily large number of methods is available for the preparation of araliphatic aldehydes, but only a few are industrially important.

Aldol condensation with subsequent hydrogenation if necessary (Eq. 4):

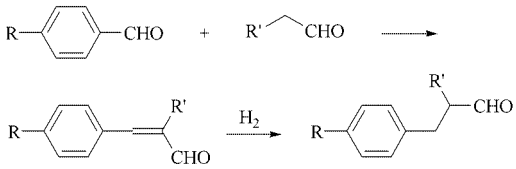

Hydroformylation of styrenes (Eq. 5):

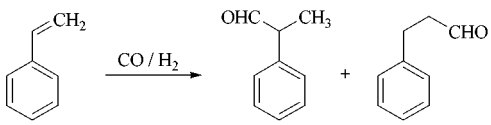

Isomerization of phenylated oxiranes (Eq. 6):

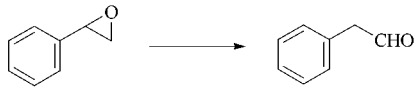

Friedel-Crafts reactions (Eq. 7):

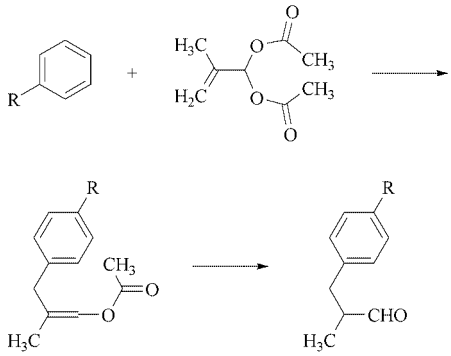

Darzens glycidic ester synthesis (Eq. 8):

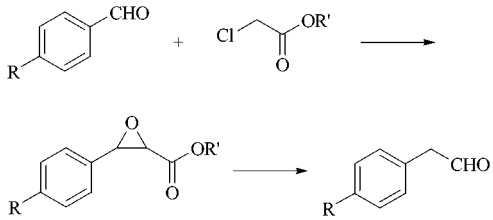

Production of the individual araliphatic aldehydes using these processes is described in the following section.

5.3. Individual Araliphatic Aldehydes

5.3.1. Phenylacetaldehyde

Phenylacetaldehyde, a constituent of numerous essential oils, is a colorless liquid with an odor resembling that of hyacinths and narcissi. Some physical properties and the CAS registry number are given in Table 7.

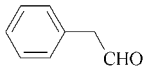

This extremely reactive compound is highly susceptible to oxidation and polymerization reactions. Alcohols are often added, for example, in perfumery, to stabilize the aldehyde by hemiacetal formation. For a comprehensive review, see [99].

Production. The most important method industrially is the isomerization of styrene oxide (Eq. 6). Ion-exchange resins [100], catalysis by chromium trioxide – tungsten trioxide on graphite [101] or silicon dioxide – aluminum trioxide [102], and thermolysis [103] are recommended for the rearrangement.

Phenylacetaldehyde can also be made by the direct oxidation of styrene in the presence of palladium salts and copper(II) chloride in aqueous solutions of glycol ethers [104].

Other possibilities include the catalytic dehydrogenation of 2-phenylethanol on silver or gold catalysts [105], [106], the hydroformylation of benzyl halides in the presence of dicobalt octacarbonyl and sodium carbonate in acetonitrile [107], and the Darzens glycide ester synthesis from benzaldehyde and alkyl chloroacetates (Eq. 8) [108].

Uses. Phenylacetaldehyde is used in the perfume industry to obtain hyacinth and rose nuances. It also imparts these aromas to tea, tobacco, and coffee (→ Flavors and Fragrances). Further applications include the preparation of pharmaceuticals, insecticides, acaricides, disinfectants, and the use as a rate-controlling additive in the polymerization of polyesters with other monomers.

The reaction of phenylacetaldehyde with ammonia and hydrogen cyanide (Strecker reaction) leads to the formation of phenylalanine, an intermediate for the sweetener aspartame, → Sweeteners. Phenylacetic acid can be obtained by oxidation of the aldehyde.

5.3.2. 4-Methylphenylacetaldehyde

Some physical properties and the CAS registry number are given in Table 7.

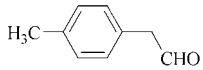

4-Methylphenylacetaldehyde has been identified as a constituent of maize oil. The colorless liquid resembles the corresponding unsubstituted aldehyde in that it is extremely reactive and must be stabilized. The compound has a strong floral odor.

Production. The methods of preparation correspond closely to those of phenylacetaldehyde. Thus, the compound can be prepared by glycidic ester synthesis from 4-methylbenzaldehyde and chloroacetate ester [108], by hydroformylation of 4-methylbenzylhalide with $Co_2(CO)_8Na_2CO_3$ in acetonitrile [107], or by oxidation of 4-methylstyrene [109].

Uses. 4-Methylphenylacetaldehyde is used in perfumes, e.g., in soap, in tobacco fragrances, in the preparation of pesticides, etc.

5.3.3. 2-Phenylpropionaldehyde (Hydratropaldehyde)

For some physical properties and the CAS registry number see Table 7.

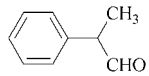

2-Phenylpropionaldehyde (hydratropaldehyde) has long been established in the fragrance industry. It occurs in many natural products and has a strong fruity smell, reminiscent of hyacinths. Surprisingly, 2-phenylpropanal is much less sensitive to oxidation and polymerization than 2-phenylacetaldehyde. For a review on 2-phenylpropanal and the corresponding alcohol, see [110].

Production. In the last few years a series of new syntheses for 2-phenylpropionaldehyde has been described. However, only a few have achieved industrial significance. The most important is the rhodium-catalyzed hydroformylation of styrene.

In this process $[HCo(CO)_4]$ is unsuitable as a catalyst because then styrene is largely hydrogenated [110], [111]. The use of chiral rhodium

complexes leads to the formation of optically active aldehydes when prochiral olefins, such as styrene, are used. Asymmetric hydroformylation, however, has so far found no industrial application because of insufficient purity of the enantiomers [112].

Other synthetic methods are the gas-phase oxidation of α-methylstyrene on Cu, Bi – Mo – P – SiO_2 [113], or Pd salt catalysts [114], the rearrangement of expoxidized α-methylstyrene [100–103], and the dehydrogenation of 2-phenyl-1-propanol on Ag catalysts at reaction temperatures of ca. 600 °C [105], [106].

Uses. 2-Phenylpropanal and its hydrogenation product, 2-phenylpropanol, are used in perfumes. The aldehyde also is a valuable starting material for a series of pharmaceuticals and pesticides, and it serves in the plastics field as a stabilizer, catalyst, and hardener.

5.3.4. 3-Phenylpropionaldehyde (Dihydrocinnamaldehyde)

Some physical properties and the CAS registry number are given in Table 7.

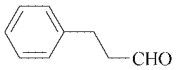

3-Phenylpropionaldehyde resembles most araliphatic aldehydes in that it is only storable to a limited extent. It is a colorless liquid with a pronounced cinnamon aroma. The aldehyde occurs naturally in Ceylon cinnamon oil. For a review, see [110].

Production. The most common preparative process for 3-phenylpropanal is the partial hydrogenation of cinnamaldehyde [115].

The 3-phenylpropanal that is formed during the hydroformylation of styrene can be separated from the isomeric 2-phenylpropanal as the bisulfite adduct [10]. The Rosenmund reduction of dihydrocinnamoyl chloride in the presence of Pd/$BaSO_4$ gives good yields of the aldehyde [116]. The reaction of 2-propen-1-ol and phenylmercury chloride with $CuCl_2$ and $LiPdCl_4$ catalysts in methanol, and the reaction of phenyl bromide and allyl alcohol in the presence of $PdCl_2$ and $NaHCO_3$ in nonpolar solvents also lead to the formation of 3-phenylpropanal [117]. A review of further synthetic methods is given in [110].

Uses. The uses of 3-phenylpropanal correspond in general to those of 2-phenylpropanal.

5.3.5. 2-(4-Isobutylphenyl)propionaldehyde

Some physical properties and the CAS registry number are given in Table 7.

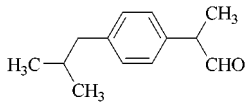

2-(4-Isobutylphenyl)propionaldehyde is a colorless, mobile liquid with a pleasant odor.

Production. Numerous methods are known for the synthesis of 2-(4-isobutylphenyl)propionaldehyde. It can be prepared by reaction of *p*-isobutylacetophenone with methyl chloroacetate using sodium methoxide as catalyst, followed by reaction of the glycidic ester with BF_3 to give 2-hydroxy-3-(4-isobutylphenyl)-3-butenoic acid ester, which is subsequently treated with mineral acid [118]:

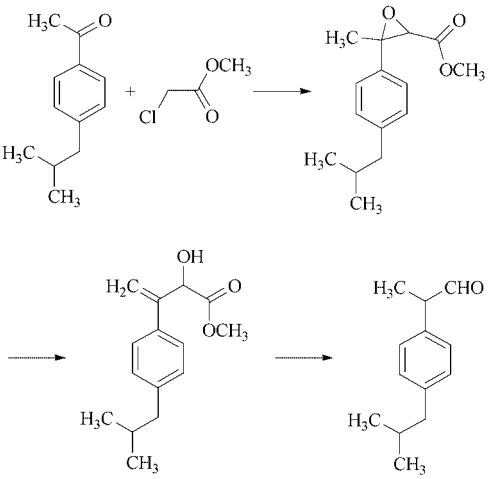

A variant of the process is the hydrolysis of the intermediately formed glycidic ester with al-

kali to give the corresponding salt of glycidic acid, which is then decarboxylated [119].

2-(4-Isobutylphenyl)propanal also is formed by isomerization of 2-(4-isobutylphenyl)-2-methyloxirane on Al_2O_3-SiO_2 [120] or anhydrous $ZnCl_2$ [121], by reaction of 1-(4-isobutylphenyl)-1-chloroethane with dimethylformamide in the presence of Li or Na in tetrahydrofuran [122], and in good yields by Rh- or Co-catalyzed hydroformylation of *p*-isobutylstyrene [123].

Uses. Unlike many araliphatic aldehydes 2-(4-isobutylphenyl)propionaldehyde has no significance in the perfume industry. It is used exclusively for the preparation of the antirheumatic agent Ibuprofen, 2-(4-isobutylphenyl)propionic acid (→ Anti-inflammatory–Antirheumatic Drugs).

5.3.6. 3-(4-Isopropylphenyl)-2-methylpropanal (Cyclamenaldehyde)

Cyclamenaldehyde, 2-methyl-3-(4-isopropylphenyl)propionaldehyde, is one of the fragrance and aroma substances of above average importance, → Flavors and Fragrances. The normal commercial racemate is a colorless to pale yellow liquid. The aldehyde is reviewed in [124].

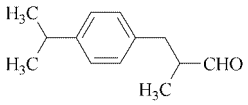

Some physical properties and the CAS registry number are given in Table 7.

In addition to its use in numerous perfume compositions, cyclamenaldehyde is an intermediate for the preparation of fungicides and fungistatic substances [125].

5.3.7. 3-(4-*tert*-Butylphenyl)-2-methylpropionaldehyde (Lilial)

3-(4-*tert*-Butylphenyl)-2-methylpropanal (→ Flavors and Fragrances), a colorless to pale yellow liquid, as yet has not been found in nature. It has a higher stability than the homologous cyclamenaldehyde and therefore is used as scent in soaps. Some physical data and the CAS registry number are summarized in Table 7. A monograph on the compound has been published [126]. More than 1000 t/a of the aldehyde is produced.

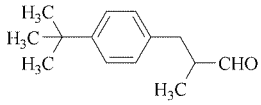

Production. The aldehyde is produced industrially almost solely by aldol condensation of 4-*tert*-butylbenzaldehyde and propionaldehyde to give 4-*tert*-butyl-α-methylcinnamaldehyde, which can be hydrogenated selectively on noble metal catalysts such as Pd, Rh, Pd – Pr_2O_3 on Al_2O_3 [127], [128], or on modified cobalt catalysts [129], [130]. The aldol condensation and the hydrogenation can be carried out in one step in the presence of a hydrogenation catalyst [131].

The Friedel – Crafts reaction of 4-*tert*-butylbenzene with methacrolein or methacrolein diacetate (Eq. 7, in page 23) proceeds in an analogous manner to the preparation of cyclamenaldehyde [132].

Further possibilities are the Rh-catalyzed hydroformylation of 1-(4-*tert*-butylphenyl)-1-methoxypropene and subsequent partial hydrogenation [133], the palladium salt catalyzed reaction of 4-*tert*-butylphenylhalide with methallylalcohol [134], and the dehydrogenation of 3-(4-*tert*-butylphenyl)-2-methylpropanol on silver catalysts [105], [106].

Uses. In addition to its applications in the perfume and aroma industry, lilial is used mainly for the synthesis of substituted 3-(4-*tert*-butylphenyl)-2-methylpropylamines, a new class of substances with fungicidal properties [125]. These compounds are effective against mildew in barley and wheat.

5.3.8. Cinnamaldehyde (3-Phenyl-2-propenal)

Cinnamaldehyde (for more details → Flavors and Fragrances), the simplest unsaturated ar-

aliphatic aldehyde, is produced in considerable quantities. It has outstanding significance as a fragrance. The *trans* isomer is the major product of industrial manufacturing processes and is the predominant naturally occurring one. Some physical properties and the CAS registry number are given in Table 7.

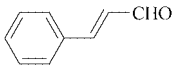

For the *synthesis* only the base-catalyzed condensation of benzaldehyde with acetaldehyde has been adopted on an industrial scale.

Uses. In addition to its application as a fragrance, it is used for the preparation of corrosion inhibitors, as a polymerization inhibitor for conjugated dienes, and for the coating of metals [145].

Partial hydrogenation or oxidation of cinnamaldehyde give 3-phenylpropanal or cinnamic acid, respectively.

5.3.9. α-Alkylcinnamaldehydes (2-Alkyl-3-phenyl-2-propenals)

The most important of the α-alkylcinnamaldehydes are α-amyl-, α-hexyl-, and α-methylcinnamaldehyde (→ Flavors and Fragrances).

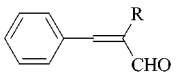

R = CH_3, C_5H_{11}, or C_6H_{13}

Some physical properties and the CAS registry numbers are given in Table 7. The aldehydes are synthesized, analogously to cinnamaldehyde, by aldol condensation. In addition to the considerable use for these compounds in the perfume industry, they are starting materials for their hydrogenated derivatives, and for the preparation of pharmaceuticals and agrochemicals.

6. Dialdehydes

Dialdehydes, compounds bearing two formyl moieties, are of significant commercial interest. Dialdehydes are used as disinfectants, corrosion inhibitors in oil recovery, leather tanning agents, hydrophobizing agents for paper, wallpaper and textiles, biological fixatives, and cross-linkers in thermosetting resins [135]. Glyoxal (ethanedial) is the simplest possible and commercially most important dialdehyde (→ Glyoxal). Because of their bifunctionality dialdehydes are potential intermediates for the synthesis of diols, dicarboxylic acids and diamines, and as monomers for the synthesis of polyesters, polyamides, and polyurethanes. Because of their reactivity, dialdehydes are often converted directly to such more stable derivatives.

6.1. Properties

Physical data of some dialdehydes are compiled in Table 8. Glyoxal has a melting point of 15 °C and boils at 50.4 °C. Propanedial exists as the isomeric 3-hydroxyacrolein. Therefore it is a solid with a melting point of 72 – 74 °C. The higher α,ω-dialdehydes are high boiling liquids. The boiling points of the isomeric branched dialdehydes are lower than those of the linear α,ω-dialdehydes.

Dialdehydes have an intrinsic tendency to undergo polymerization. For long chain α,ω-dialdehydes cyclopolymerization reactions are known [136]. Anhydrous glyoxal polymerizes readily to complex polymeric structures. The monomer can be recovered from this polymer by heating in the presence of phosphorus pentoxide. Similar to formaldehyde, glyoxal reacts with water in a highly exothermic reaction to yield a mixture of linear and cyclic oligomers. In aqueous solution, higher dialdehydes form cyclic hemiacetals.

6.2. Production

Although many syntheses of dialdehydes are described in the patent literature, only a few are of commercial interest, mainly due to lack of selectivity. The most important processes for the preparation of saturated aliphatic and cycloaliphatic dialdehydes are:

– Oxidation of ethylene glycol (→ Glyoxal, Chap. 4.)

Table 8. Physical properties of aliphatic and alicyclic dialdehydes

	Name	CAS registry number	Formula	M_r	*mp*, °C	*bp*∗∗, °C (*p*)	Density (20 °C), g/cm^3	Refractive index n_D^{20}
C_2	Ethanedial (glyoxal)	[*107-22-2*]	OHCCHO	58.04	15	50.4	1.14	1.383
C_3	Propanedial (Malondialdehyde, 3-hydroxyacrolein)	[*542-78-9*]	$OHCCH_2CHO$	72.06	72 – 74			
C_4	Butanedial (succinedialdehyde)	[*638-37-9*]	$OHC(CH_2)_2CHO$	86.09		58 (9)	1.064	1.426
C_5	Pentandial (glutardialdehyde)	[*111-30-8*]	$OHC(CH_2)_3CHO$	100.12	< – 14	188 (decomp.)		
C_6	Hexanedial (adipinaldehyde)	[*1072-21-5*]	$OHC(CH_2)_4CHO$	114.14	– 8	92 – 94 (9)	1.003	1.435
C_{12}	TCD dialdehyde*	[*25896-97-3*]	$OHCC_{10}H_{14}CHO$	192.25	– 33	ca. 288	1.14	

∗ 3(4),8(9)-Bis(formyl)tricyclo[5.2.1.0^{2,6}]decane.
∗∗ at 101.3 kPa unless otherwise specified in parentheses.

– Oxidation of acetaldehyde (→ Glyoxal, Chap. 4.)
– Hydroformylation of dienes
– Oxidative ring opening of cycloalkenes
– Addition of methyl vinyl ether to acrolein (glutardialdehyde)

Hydroformylation of dienes yields different products depending on the structure of the diene and the catalyst applied [1], [3]. Conjugated dienes such as 1,3-butadiene yield saturated monoaldehydes or monoalcohols if unmodified cobalt or rhodium catalysts are used. Initially the single addition of CO/H_2 yields the unsaturated monoaldehyde. The hydroformylation of the unsaturated aldehyde than competes with hydrogenation and isomerization of the double bond. Patent literature describes the bis-hydroformylation of 1,3-butadiene to 1,6-hexanedialdehyde with special modified rhodium catalysts, but commercial realization remains questionable [137].

Nonconjugated dienes tend to isomerize to the thermodynamically more stable α,β-unsaturated aldehydes after single addition of CO/H_2 [3]. Therefore, especially nonconjugated dienes with widely separated ($> C_6$), preferably terminal double bonds are useful for synthesis of dialdehydes by hydroformylation.

Oxidative Ring Opening of Cycloalkenes. In the presence of catalysts, e.g. tungstic acid, cycloalkenes react with hydrogen peroxide via formation of the corresponding epoxides to give linear α,ω-dialdehydes [138].

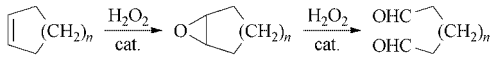

In some cases the addition of boron compounds facilitates the reaction. Oxidation of cycloalkenes in the gas phase has also been described [139].

Miscellaneous Processes. Similar to the formation of monoaldehydes by dehydrogenation, the dehydrogenation of diols yields dialdehydes. Other routes are the ozonization of cycloalkenes and the dimerization of unsaturated aldehydes [3], [140]. Hydroformylation of unsaturated acetals gives the corresponding dialdehydes.

6.3. Individual Dialdehydes

6.3.1. Succinaldehyde (1,4-Butanedial)

Some physical properties of succinaldehyde and its CAS registry number are given in Table 8. To avoid spontaneous polymerization of anhydrous succinaldehyde, it is best stored as a cyclic acetal, e.g., 2,5-dimethoxytetrahydrofuran. Aqueous solutions consist mainly of the cyclic hemiacetal 2,5-dihydroxytetrahydrofuran.

Production. Succinaldehyde is prepared by various routes, e.g., reaction of chlorine with tetrahydrofuran and subsequent acid hydrolysis. Hydroformylation of acrolein acetals (mostly cyclic) yields a mixture of 2- and 3-(1,3-dioxan-2-yl)propanal. Hydrolysis of the latter forms 1,4-butanedial [141].

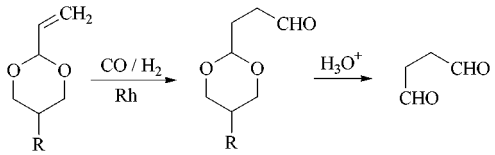

Due to the similar chemical properties, succinaldehyde has been replaced extensively by glutardialdehyde, diminishing its commercial importance significantly.

Uses. Succinaldehyde is used as disinfectant, biological fixative, and cross-linker in thermosetting resins.

6.3.2. Glutardialdehyde (1,5-Pentanedial)

Some physical properties of glutardialdehyde and its CAS registry number are compiled in Table 8. Glutardialdehyde is the most important dialdehyde besides glyoxal. Due to its high reactivity glutardialdehyde is sold as a 25 or 50 wt % aqueous solution or as its bis(sodium bisulfite) adduct.

Production. Glutardialdehyde is produced by gas-phase oxidation of cyclopentene or by Diels – Alder reaction of acrolein with methyl vinyl ether [139], [142]. The latter yields 3,4-dihydro-2-methoxy-2*H*-pyran [*4454-05-1*], acidic hydrolysis of which leads to glutardialdehyde.

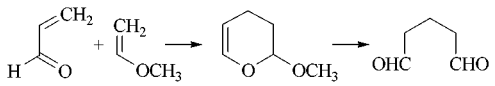

In both cases the dialdehyde is recovered from the reaction mixture by multistage extraction with water.

Uses. Glutardialdehyde is used mainly for leather tanning, as a disinfectant and sterilizer, as a biocide in oil recovery, and in the paper and textile industries to improve wet strength and dimensional stability of fibers.

6.3.3. TCD Dialdehyde

Some physical properties and the CAS registry number of TCD dialdehyde (isomeric mixture), 3(4),8(9)-bis(formyl)tricyclo[5.2.1.0^{2,6}]decane, are given in Table 8. Like other dialdehydes TCD dialdehyde is not very stable and has a significant tendency for polymerization. Distillation under reduced pressure is possible. In practice the dialdehyde is not stored but converted immediately to more stable derivatives.

Production. TCD dialdehyde is produced in up to 95 % yield by hydroformylation of dicyclopentadiene on rhodium catalysts. To achieve bis-hydroformylation, a high pressure of synthesis gas is required (150 – 250 bar) [143].

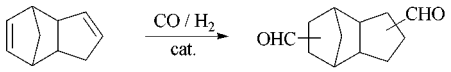

If unmodified cobalt catalysts are used the dialdehyde is mostly hydrogenated to TCD diol, 3(4), 8(9)-bis(hydroxymethyl)tricyclo[5.2.1.0 ^{2,6}]decane, and TCD monoaldehyde monoalcohol. At low synthesis gas pressures, mainly the monoaldehyde is formed.

Uses. TCD dialdehyde is used for the production of the corresponding diamine, dicarboxylic acid, and diol. These compounds are used as monomers for production of poylesters, polyamides, and polyurethanes. Also some applications in the field of flavors and fragrances or dental materials are known [144].

7. Acetals

Acetals can be represented by the general formula

where R can be either hydrogen or some aliphatic, aromatic, or heterocyclic group. In contrast to the hydrates (R′ = H), which normally cannot be isolated, the acetals are generally stable compounds.

Cyclic acetals can be prepared by reaction of aldehydes with diols; 1,2- and 1,3-diols are the most common.

1,3-Dioxolanes

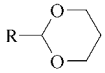

1,3-Dioxanes

For the lower acetal homologues, trivial names are frequently used, e.g., methylal for dimethoxymethane. Because of their stability toward alkali, acetals are used as soap perfume oils. The scents of the acetals are somewhat different and less distinct than those of the corresponding aldehydes.

Acetals are found in wine and other alcoholic drinks. Piperonal, a raw material for aroma compositions, and numerous alkaloids such as piperine, a main alkaloid of black pepper, or berberine, which is found mainly in barberry roots, possess a methylenedioxy group.

7.1. Properties

Physical Properties. The lower acetals are colorless, mobile liquids, with boiling points distinctly higher than those of the corresponding aldehydes. Their solubilities in water decrease rapidly with increasing molecular mass. They have good solubilities in alcohols, ethers, and other common solvents. The most important physical data and the CAS registry numbers of some acetals are presented in Table 9.

As a rule acetals can be purified by distillation. The distillation residue must be handled with extreme care because, as a result of the ether structure of the acetals, it can contain explosive peroxides. For the same reason care should be taken when storing acetals.

Chemical Properties. Acetals are extremely stable toward base, but are hydrolyzed easily by acid. Because of these properties, the acetals provide interesting possibilities for synthesis, especially in carbohydrate chemistry.

In contrast to the aliphatic alkanals, their acetals are extremely resistant to most reducing and oxidizing agents. This is only true to a limited extent for the acetals of aromatic aldehydes, however.

Among the factors that influence the rate of hydrolysis in acids is the structure of the acetal; all the cyclic acetals are more stable than the open-chain species. Difficulties arise during hydrolysis when the aldehyde group is very reactive; secondary reactions, such as aldol condensation, occur.

Acetals can be transacetalized by other carbonyl compounds or alcohols:

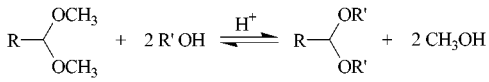

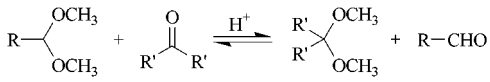

The thermal cleavage of an acetal leads to a vinyl ether and an alcohol:

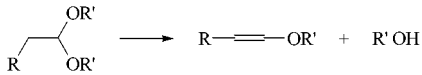

If R′ is an allyl group, Claisen rearrangement to the corresponding unsaturated aldehyde takes place during the thermal cleavage (see page 16).

The cleavage of 4,4-dimethyl-1,3-dioxane, a cyclic acetal, prepared by the Prins reaction of isobutene with formaldehyde (see below), is of industrial importance for the manufacture of isoprene. The dioxane is split at 240 – 400 °C on H_3PO_4/C or $Ca_3(PO_4)_2$ catalysts in the presence of water:

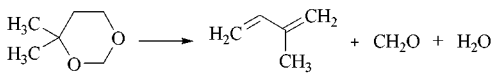

Both BF_3 or $ZnCl_2$ catalyze the reactions of acetals with vinyl ethers to give β-alkoxyacetals, which are converted to the corresponding α,β-unsaturated aldehydes upon treatment with acetic acid:

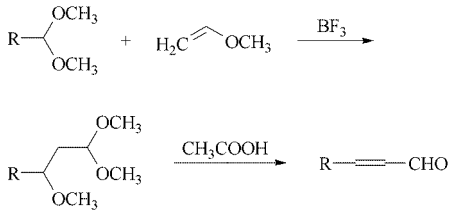

Table 9. Physical properties of some acetals

Name	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C	Density (20 °C), g/m^3	Refractive index, n_D^{23}
Formaldehyde dimethyl acetal (methylal)	[*109-87-5*]	$CH_2(OCH_3)_2$	76.09	− 104.8	45.5	0.8593	1.3513
Formaldehyde diethyl acetal (ethylal)	[*462-95-3*]	$CH_2(OCH_2CH_3)_2$	104.15	− 66.5	89	0.8319	1.3748
Acetaldehyde dimethyl acetal (1,1-dimethoxyethane)	[*534-15-6*]	$CH_3CH(CH_3)_2$	90.12	− 113.2	64.5	0.8502	1.3668
Acetaldehyde diethyl acetal (1,1-diethoxyethane)	[*105-57-7*]	$CH_3CH(OCH_2CH_3)_2$	118.18		103.2	0.8314	1.3834
Propionaldehyde dimethyl acetal (1,1-dimethoxypropane)	[*4744-10-9*]	$CH_3CH_2CH(OCH_3)_2$	101.15		84 – 86	0.8649	
Propionaldehyde diethyl acetal (1,1-diethoxypropane)	[*4744-08-5*]	$CH_3CH_2CH(OCH_2CH_3)_2$	132.21		124	0.8232	1.3924
Butyraldehyde dimethyl acetal (1,1-dimethoxybutane)	[*4461-87-4*]	$CH_3CH_2CH_2CH(OCH_3)_2$	118.18		112 – 113		1.3888
Butyraldehyde diethyl acetal (1,1-diethoxybutane)	[*3658-95-5*]	$CH_3CH_2CH_2CH(OC_2H_5)_2$	146.22		145 – 145.5	0.8320	1.3970
Pentanal diethyl acetal (1,1-diethoxypentane)	[*3658-79-5*]	$CH_3(CH_2)_3CH(OCH_2CH_3)_2$	160.25		162.5 – 163.5	0.8304	1.4021
Hexanal diethyl acetal (1,1-diethoxyhexane)	[*3658-93-3*]	$CH_3(CH_2)_4CH(OCH_2CH_3)_2$	174.28		155 – 158		
1,3-Dioxolane	[*646-06-0*]		74.08	− 95	78	1.0600	1.3974
1,3-Dioxane	[*505-22-6*]		88.12	− 42	105 (100.6)	1.0342	1.4165
2-Methyl-1,3-dioxolane	[*497-26-7*]	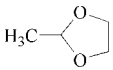	88.12		81 – 82	0.9811	1.4035
2-Methyl-1,3-dioxane	[*626-68-6*]	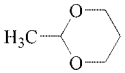	102.13		108.5 – 109.5 (99.3)	0.9701	1.4139

7.2. Production

The primary process for the preparation of acetals is the reaction of an aldehyde with an alcohol. As a rule, a unstable hemiacetal is formed as an intermediate:

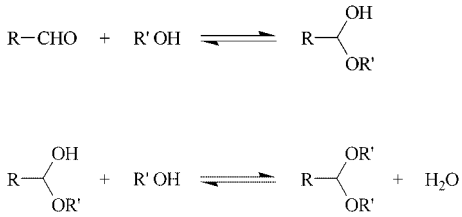

Because the reaction is an equilibrium process, the water must be removed by azeotropic distillation or by the addition of water-adsorbing agents, e.g., molecular sieves, in order to improve the yield. Cyclic acetals are generally obtained in better yield than open-chain acetals. As catalysts, anhydrous sulfuric acid or *p*-toluene sulfonic acid are normally used, although other inorganic acids, oxalic or adipic acid, or ion-exchange resins or molecular sieves also can be employed.

During the acetalization of saturated aliphatic aldehydes that contain an α-hydrogen atom, 1-alkenylethers can be formed, if the reaction temperature is too high:

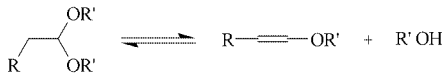

In accordance with a process developed by L. Claisen acetals are formed in good yields from most aldehydes by treating them with orthoformates, mainly the methyl and ethyl esters. Strong acids are normally used as catalysts:

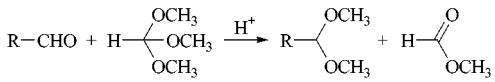

Acetals can also be obtained by reaction of acetylene with alcohols, by addition of alcohol to vinyl ethers, and by treatment of geminal dihalides with alkoxides. Aldehydes and oxiranes form 1,3-dioxolanes, and pyrocatechols and dichloromethane form 1,3-benzodioxoles.

Of industrial importance is the acid-catalyzed reaction of olefins with formaldehyde to give 1,3-dioxanes (Prins reaction):

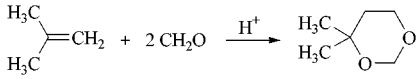

The reaction is carried out at 55 – 75 °C with an excess of formaldehyde.

The reaction of olefins with alcohols under hydroformylation conditions using, for example, phase-transfer catalysts also leads to the formation of acetals [146].

Acetaldehyde dimethyl acetal is obtained at 200 °C from methanol and synthesis gas under catalytic conditions [147]. The methanol conversion is about 60 % with a selectivity of 80 – 85 %. Byproducts are mainly acetaldehyde and ethyl acetate.

Vinyl esters can be converted to acetals by reaction with alcohols in the presence of palladium chloride catalyst:

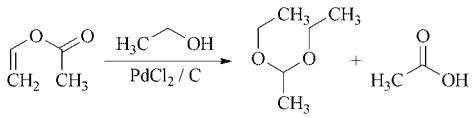

For further methods of synthesis and a review of the reactions of acetals, see [1, p. 615], [148].

7.3. Uses

Acetals frequently are used in preparative chemistry to protect aldehyde functions. The acetals of lower aldehydes, e.g., formaldehyde dimethyl acetal (methylal), serve as solvents for cellulose and cellulose derivatives.

Although acetals have a somewhat weaker fragrance than the corresponding aldehydes, they are used for preparing soap perfume oils, especially because of their stability toward alkali. This applies both for open-chain as well as for cyclic acetals [149].

In polymer chemistry acetals are used for the preparation of epoxy resins [150] and aqueous dispersions [151]. The acetals of polymeric alcohols, such as polyvinyl alcohol, are employed as interlayers in safety glasses. Acetals are also

used for the synthesis of pharmaceuticals, insecticides [152], fertilizers [153], and as constituents of sun lotions [154].

Substituted 1,3-dioxanes are intermediates in the synthesis of conjugated dienes. The IFP (Institut Français du Pétrole) process for the production of isoprene is the most important example [155].

8. Quality Control

The actual method used for the determination of quality primarily depends on the further processing and the special properties of the individual products. In general, high purity of the aldehydes is required.

The composition is analyzed by gas chromatography (cf. Chap. 9). In addition, the following characteristic data usually are determined:

1) Density at 20 °C (DIN 51 757, ASTM D 1298)
2) Refractive index n_D^{20} (DIN 51 423, Bl. 2, ASTM D 1747)
3) Acid value (DIN 53 402, ASTM D 1613)
4) Hazen color index (APHA) (DIN ISO 6271, ASTM D 1209)
5) Boiling range at 101.3 kPa (DIN ISO 51 777, ASTM D 1364)
6) Water content by the Karl Fischer method (methanol-free reagents and solvents)

The following data also can be useful:

1) Carbonyl value (DIN 53 173, ASTM D 2192)
2) Hydroxy value (DIN 53 240, ASTM E 222)
3) Viscosity at 20 °C (DIN 51 757, ASTM D 4052)
4) Iodine number (determination by the Kaufmann method DGF C-V116; for unsaturated aldehydes)

Safety data, such as TLV and MAK values, acute and chronic toxicology data, flash points (DIN 51755, ASTM D 56 or DIN 51758, ASTM D 93), ignition temperatures (DIN 51794, ASTM D 2155), vapor pressures, and solubilities in water, are important under certain circumstances, e.g., for classification and handling of the products during storage and transportation.

9. Analysis

Aldehydes are mainly analyzed by chromatographic and spectroscopic techniques; chemical detection methods also are used.

The most important qualitative and quantitative analytical procedure in industry is *gas chromatography.* The optimum procedure, i.e., filling of the column, carrier gas, detector, etc., depends on the nature of the sample and the individual requirements.

The majority of GC analyses of aldehydes are performed on capillary columns. Packed columns are also still used. Columns with a length of 1 – 3 m and an inner diameter of 2 – 4 mm, and capillary columns (25 – 100 m) with various polar stationary liquids can be used. The most frequently employed support materials are calcinated diatomaceous earths.

The amount of sample that is injected is 0.1 – 0.5 µL for capillary columns with flow separation and 1 – 3 µL for packed columns. In some cases it is advisable to convert the products into derivatives before analysis. Carrier gases which have proven suitable are N_2, He, and H_2. If hydrogen is used, particular care must be taken that both the injection system and the column are free from catalyst deposits; otherwise the aldehydes easily could be hydrogenated. For compounds with particularly low thermal stability, such as the aldols, the sample is added directly into the separating column.

The 2,4-dinitrophenylhydrazones generally are used in *liquid chromatography.* UV detectors serve to identify the constituents of the sample.

For the *chemical identification* of aldehydes, crystalline derivatives of low solubility are used, e.g., 2,4-dinitrophenylhydrazones, semicarbazones, oximes, sodium hydrogen sulfite adducts, and dimedon condensation products. For qualitative detection the reducing properties of the aldehydes (e.g., reaction with Fehling's solution or Tollen's silver nitrate solution) and numerous color reactions can be used. The reducing properties of the aldehydes are not specific, however.

Aldehydes can be detected qualitatively and identified by IR, ^{1}H NMR, and ^{13}C NMR spectroscopy. In IR spectroscopy saturated aldehydes can be characterized by the intensive CO stretching vibration in the range 1740 –

1720 cm^{-1}, and α,β-unsaturated compounds by the absorption at 1705 – 1685 cm^{-1}. The absorption is shifted to lower frequencies for conjugated, unsaturated compounds.

In the ^1H NMR spectrum the formyl H is found at 9.4 – 9.8 ppm in saturated aliphatic aldehydes; the coupling constant is 0 – 3 Hz for the H-C-CHO system. The corresponding values for unsaturated compounds are 9.3 – 10.2 ppm; the coupling constants for the system R–CH=CH–CHO are strongly dependent on whether the compound has E or Z configuration and which substituents are present.

For further details of the analysis of aldehydes see [156].

10. Storage, Transportation, and Environmental Regulations

For storage and transportion, containers of stainless steel normally are used. Vessels lined with polyethylene or other coatings are also suitable. For aldehydes that enter the market as solutions, such as formaldehyde, glyoxal, or glutaraldehyde, aluminum vessels or containers of standard steel should not be used, because the acids formed by autoxidation are corrosive, and the corrosion products can cause discoloration of the aldehyde.

Aldehydes are normally stored under a nitrogen atmosphere. Antioxidants and stabilizers are added to prevent autoxidation.

The formation of cyclic 2,4,6-trialkyl-1,3,5-trioxane trimers is catalyzed by the presence of strong protic and Lewis acids. The trimerization tendency of aldehydes occasionally increases at lower temperatures, so that some compounds cannot be stored over long periods at temperatures under 20 °C without stabilization [1, p. 651].

Some α,β-unsaturated aldehydes, such as acrolein, methacrolein, and crotonaldehyde, can react violently if handled incorrectly in the presence of air or oxygen or some other, in particular, alkaline compound.

Because of their health hazards, saturated and unsaturated aliphatic aldehydes, especially the lower members, should be handled only if special safety precautions are taken. Necessary measures can be found in the regulations on dangerous materials and in the directions on accident prevention, guidelines, leaflets, etc., published by the confederation of trade associations.

For the economically important aldehydes numerous transport regulations and legal directions exist (IMO, IATA, RID, ADR, etc.). The classification of a special aldehyde depends on its specific properties such as flash point, boiling point, solubility in water, toxicity, and ignition temperature.

Because the aldehydes sometimes have a very unpleasant and intense odor, special environmental measures are necessary during manufacture, storage, and dispatch. The aldehyde-containing off-gas from the production normally is drawn off centrally and burned. The wastewaters are treated chemically and biologically.

11. Economic Aspects

The most important process for the manufacturing of aldehydes with more than two carbon atoms is the oxo synthesis. In 1994 total worldwide capacity was 7.2×10^6 t/a with further capacities of $(0.5 – 1.0) \times 10^6$ t/a planned or under construction. Table 10 lists oxo capacities by region for 1984 and 1994.

Table 10. Oxo capacities by region (in 10^6 t/a)

	1984	1994
Western Europe	2.5	2.4
United States	1.4	2.4
Japan	0.5	0.8
Others	0.6	1.6
Total	5.0	7.2

The most important secondary product of butyraldehyde is 2-ethylhexanol, with worldwide capacities of 2.6×10^6 t/a in 1996. The consumption of various oxo chemicals in Western Europe, the United States, and Japan is compiled in Table 11.

For the most important aldehydes used in the perfume industry, few production data are available. Estimated production for the most important fragrance aldehydes are as follows: cinnamaldehyde 1500 t/a, lilial 1500 t/a, hydroxycitronellal 1500 t/a, α-hexylcinnamaldehyde 1000 t/a, α-amylcinnamaldehyde 800 t/a, citral 500 t/a, citronellal 400 t/a, and cyclamenaldehyde 350 t/a.

Table 11. Consumption of oxo chemicals by region (in 10^3 t/a)

	Western Europe		United States		Japan	
	1993	1998	1993	1998	1993	1998
Propionaldehyde	11	12	161	183	1	1
n-Butyraldehyde	1224	1274	1055	1178	572	622
Isobutyraldehyde	133	128	234	263	65	72
Valeraldehydes	11	12	32	35		
$C_6 - C_{13}$ plasticizer alcohols**	430	460	396	430	93	105
Detergent alcohols**	150	174	193	215	47	53
$C_7 - C_9$ oxo fatty acids	11	12	37	44		
Branched-chain oxo acids	15	19	1	1		
Neo acids	50	61	16	19		
Total	2046	2162	2125	2368	778	853

* Excluding 2-ethylhexanol.
** Based on equivalent amounts of aldehyde.

12. References

General References

1. J. Falbe (ed.): "Aldehyde," in *Houben-Weyl: Methoden der Organischen Chemie,* **vol. E 3,** Thieme Verlag, Stuttgart-New York 1983.
2. J. Falbe (ed.): "C-O-Verbindungen," in *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart-New York 1975.
3. J. Falbe (ed.): *New Syntheses with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980.
4. J. Falbe (ed.): *Carbon Monoxide in Organic Synthesis,* Springer Verlag, Berlin-Heidelberg-New York 1970.
5. R. Brettle in: *Comprehensive Organic Chemistry,* **vol. 1,** chap. 5.1, Pergamon Press, Oxford 1979.
6. C. A. Buehler, D. E. Pearson: *Survey of Organic Syntheses,* **vol. 2,** J. Wiley & Sons, New York-London-Sydney-Toronto 1977.
7. S. Patai: *The Chemistry of the Carbonyl Group,* Interscience Publ., London-New York-Sydney 1966.
8. S. R. Sandler, W. Karo: *Organic Functional Group Preparations,* **vol. 12,** Academic Press, New York-London 1968.

Specific References

9. L. H. Horsley: "Azeotropic Data III," *Adv. Chem. Ser.* **116** (1973), ACS, Washington.
10. Ruhrchemie, DE 2459152, 1974 (L. Bexten, H. Noeske, H. Tummes, B. Cornils).
11. BASF, DE 960187, 1952 (W. Hagen).
12. BASF, DE-OS 2218305, 1972 (R. Kummer, H. J. Nienburg, G. Butz).
13. BASF, DE 1022575, 1958 (H. J. Nienburg, H. Böhm).
14. *Katalysatoren Hoechst,* Firmenbroschüre, Hoechst-Ruhrchemie 1979.
15. G. Fagherazi, N. J. Pernicone, *J. Catal.* **16** (1970) 321.
16. H. W. Knol, *Manuf. Chem. Aerosol News* **37** (1966) no. 2, 42 – 45.
17. Stamicarbon, EP 37149, 1980 (T. F. De Graaf, H. J. Delahaye).
18. Moscow Inst. of Fine Chem. Technol., SU 789491, 1978 (M. L. Kaliya et al.).
19. Phillips Petroleum, US 4304943, 1979 (G. Bjornson).
20. C. Jallabert, H. Riviere, *Tetrahedron* **36** (1980) no. 9, 1191 – 1194.
21. Celanese, US 2128908, 1934 (J. E. Bludworth). L. F. Albright, *Chem. Eng.* **74** (1967) no. 17, 165.
22. E. Mosettig, R. Mozingo, *Org. React. (N.Y.)* **4** (1948) 362. K. Harsanyi et al., *J. Med. Chem.* **7** (1964) 623. K. Balenovic et al., *J. Org. Chem.* **21** (1956) 115. Shell Int. Res., EP 5280, 1978.
23. G. W. Fleet, P. J. Harding, *Tetrahedron Lett.* 1979, 975.
24. H. C. Brown, B. C. Subba Rao, *J. Am. Chem. Soc.* **80** (1958) 5377.
25. N. Müller, W. Hoffmann, *Synthesis* 1975, 781. J. J. Pappas et al., *Tetrahedron Lett.* 1966, 4273.
26. R. Criegee, L. Kraft, B. Bank, *Justus Liebigs Ann. Chem.* **507** (1933) 159.
27. L. Malaprade, *C. R. Acad. Sci.* **186** (1928) 382.
28. Mitsui Petrochem. Ind., JP 80102528, 1979 (*Chem. Abstr.* **94** (1981) 46774).
29. I. M. Goldman, *J. Org. Chem.* **34** (1969) 1979. A. J. Fatiadi, *Synthesis* 1976, 65.

30. T. R. Beebe, P. Hii, P. Reinking, *J. Org. Chem.* **46** (1981) 1927.
31. H. Stetter, E. Reske, *Chem. Ber.* **103** (1970) 643. *Vogel's Textbook of Practical Organic Chemistry,* 4th ed., Longman, London-New York 1978, p. 422.
32. G. B. Bennett, *Synthesis* **1977**, 589. S. J. Rhoads, N. R. Raulins, *Org. React.* (*N.Y.*) **22** (1975) 1.Ruhrchemie, DE 2844635, 1978 (J. Weber, W. Bernhagen, H. Springer).
33. C. A. Buehler, D. E. Pearson: *Survey of Organic Syntheses,* **vol. 2,** J. Wiley & Sons, New York-London-Sydney-Toronto 1977, p. 480. S. R. Sandler, W. Karo: *Organic Functional Group Preparations,* **vol. 12,** Academic Press, New York-London 1968, p. 145.
34. Chem. Verwertungs-Ges. Oberhausen, DE 953605, 1956 (G. Schiller). Conoco, US 4183825, 1978 (J. T. Carlock). Exxon Res. & Eng. Co., EP 24088, 1980 (H. L. Mitchell). Davy McKee Oil Chem., EP 16285, 1979 (T. F. Shevels, N. Harris). Agency of Ind. Sci. Technol., DE 2909041, 1978 (I. Ogata, Y. Kawaba, M. Tanaka, T. Hayashi).
35. Ruhrchemie, DE 2855505, 1978 (J. Weber, W. Bernhagen, H. Springer).
36. W. J. Scheidmeir, *Chem. Ztg.* **96** (1972) no. 7, 383. S. S. Kagna, A. G. Trifel, V. J. Gankin et al., GB 1524775, 1977. J. E. Knap, N. R. Cox, W. R. Privette, *Chem. Eng. Prog.* **62** (1966) no. 4, 74 – 78.
37. Yaroslavl. Polytechnic Inst., SU 761450, 1978 (A. V. Irodov et al.).
38. Kuraray Chem. Co., JP 8002632, 1978, JP 7973713, 1977, JP 79109909, 1978 (F. Yamamoto et al.).
39. Erdölchemie, DE 2163396, 1971 (H. Fischer, G. Schnuchel). Int. Synth. Rubber Co., GB 2063297, 1979 (D. G. Timms).
40. Ruhrchemie, DE 2917779, 1979 (J. Weber, W. Bernhagen, H. Springer).
41. Bayer, EP 90246, 1982 (G. Rauleder, H. Waldmann).
42. Ciba-Geigy, GB 861367, 1959.
43. Union Carbide, EP 28378, 1979 (E. Billig, D. L. Bunning). C. U. Pittmann, G. M. Wilemon, *J. Org. Chem.* **46** (1981) no. 9, 1901 – 1905.
44. Ruhrchemie, DE 2933919, 1979 (W. Bernhagen, J. Weber, H. Bahrmann, H. Springer).
45. Carter-Wallace, DE-OS 2939105, 1979. BASF, DE-OS 1957301, 1969 (F. Merger, R. Platz, W. Fuchs).
46. Soc. Organico, FR 952985, 1947 (P. Franck, C. Gregory, M. Genas, O. Kostelitz).
47. Celanese, US 4201714, 1977, US 4201728, 1977 (O. R. Hughes).
48. Ruhrchemie, BE 865728, 1977 (J. Weber, H. Springer).
49. Shell Int. Res., GB 2068377, 1980 (P. Wilhelmus, N. M. van Leeuwen, C. F. Roobeek).
50. Continental Oil, US 4097535, 1977 (K. Yang, K. L. Motz, J. D. Reedy).
51. Ruhrchemie, DE 2844638, 1978 (J. Weber, V. Falk, C. Kniep).
52. Ruhrchemie, DE-OS 2 844 636, 1978 (J. Weber, W. Bernhagen, H. Springer).
53. Kinki University, JP 75130708, 1974 (Y. Matsubara, Y. Fujiwara, C. Hata).
54. Air Products & Chem., US 4230641, 1977 (C. M. Bartish). Union Carbide, US 4148830, 1976 (R. L. Pruett, J. A. Smith).
55. BASF, DE 2604545, 1976 (K. Schwirten, W. Disteldorf, W. Eisfeld, R. Kummer). Ruhrchemie, DE 2737633, 1977 (L. Bexten, B. Cornils, H. D. Hahn, H. Tummes).
56. Henkel, DE 2914187, 1979 (J. Hagen, R. Lehmann, K. Bansemir).
57. Johnson-Matthey, BE 890210, 1980 (M. J. H. Russel, B. A. Murrer).
58. Rhône-Poulenc, FR 2473504, 1979 (J. Jenck).
59. Ruhrchemie, EP 7 609, 1978 (H. Bahrmann, B. Cornils, G. Diekhaus, W. Kascha, J. Weber).
60. M. S. Newman, B. J. Magerlein, *Org. React.* (N.Y.) 5 (1949) 413.
61. Givaudan, US 4154762, 1973 (I. Huang, L. M. Polinski, K. K. Rao).
62. J. H. Babler, B. J. Invergo, *Tetrahedron Lett.* 1981, 11 – 14.
63. Ruhrchemie, DE-OS 2855506, 1980 (H. Springer, J. Weber).
64. Shell, US 2479283, 1949, US 2514688, 1950 (R. R. Whetstone). W. S. Hillman, *J. Am. Chem. Soc.* **71** (1949) 324.
65. J. N. Nazarov et al., *Zh. Obshch. Khim.* **29** (1959) 3965; *J. Gen. Chem. USSR (Engl. Transl.)* **29** (1959) 3925.
66. R. I. Hoaglin, D. H. Hirsh, *J. Am. Chem. Soc.* **71** (1949) 3468. O. Isler et al., *Helv. Chim. Acta* **39** (1956) 249.
67. K. Bernhauer, I. Skudrzyk, *J. Prakt. Chem.* **155** (1940) 310.
68. Y. Nakatani, M. Sugiyama, C. Honbo, *Agric. Biol. Chem.* **39** (1975) 2431.
69. F. G. Fischer, K. Löwenberg, *Justus Liebigs Ann. Chem.* **494** (1932) 273.

70. BASF, DE-OS 2815539, 1978, DE-OS 2847069, EP 4622, 1979 (R. Fischer, H. Weitz).
71. Union Carbide, US 2628257, 1953 (R. I. Hoaglin, D. H. Hirsh). GAF, US 2543312, 1948 (J. W. Copenhaver).
72. Eastman-Kodak, US 4316990, 1980 (D. L. Morris).
73. Ruhrchemie: "Oxo/2-Ethylhexanol Process," *Hydrocarbon Process.*1981, 165. B. Cornils, A. Mullen, *Hydrocarbon Process.* 1980, 93.
74. BASF, EP 44440, 1981 (L. Janitschke, W. Hoffmann).
75. O. Bayer (ed.): "Aldehyde", in *Houben-Weyl: Methoden der organischen Chemie,* **vol. VII/1,** Thieme Verlag, Stuttgart 1954, p. 76. R. P. Bell, *J. Chem. Soc. (London),* **1937,** 1637; **1958,** 1691; **1960,** 2983.
76. L. P. Koshechkina, A. A. Yasnikov, *Ukr. Khim. Zh.* (Russ. Ed.) 40 (1974) no. 9, 948 – 957.
77. A. T. Nielsen, W. J. Houlihan, *Org. React. (N.Y.)* **16** (1968) 1.
78. Esso, US 3729515, 1967 (N. L. Cull, E. M. Gladrow, R. B. Mason, G. P. Hammer).
79. Union Carbide, US 4215076, 1976 (M. L. Deem, K. C. Stueben).
80. Celanese, GB 783458, 1957. BASF, DE-OS 2000699, 1971 (F. Merger, W. Fuchs, S. Winderl, E. Haarer).
81. BASF, DE-AS 1235883, 1965 (R. Platz). Celanese, GB 783458, 1957.
82. B. Cornils, H. Feichtinger, *Chem. Ztg.* **100** (1979) 504 – 514.
83. G. Wittig, H. Reiff, *Angew. Chem.* **80** (1968) 8; *Angew. Chem., Int. Ed. Engl.* **7** (1968) 7. BASF, DE 1199252, 1963 (G. Wittig, H. Pommer, W. Stilz).
84. G. Wittig et al., *Angew. Chem.* **75** (1963) 978.
85. Y. Takegami et al., *Bull. Chem. Soc. Jpn.* **37** (1964) 672. Du Pont, US 3687981, 1972 (F. R. Lawrence, R. H. Sullivan).
86. Shell, GB 2 282 137, 1995 (E. Drent, W. W. Jager). Arco Chemical Technology, US 5 233 093, 1993 (R. Pitchai et al.). Arco Chemical Technology, US 5 166 370, 1993 (F. J. Liotta et al.). Kuraray Co., Daicel Chemical Ind., EP 129 802, 1985 (M. Matsumoto et al.). Daicel Chemical Ind., JP 57 206 629, 1983. *Chem. Ind. (Düsseldorf)* **32** (1980) 275. M. Tamura, S. Kumano, *Chem. Econ. Eng. Ref.* **12** (1980) no. 9, 32.
87. L. E. Schniepp, H. H. Geller, *J. Am. Chem. Soc.* **68** (1946) 1646. G. F. Woods, H. Sanders, *J. Am. Chem. Soc.* **68** (1946) 2111.
88. C. W. Smith et al., *J. Am. Chem. Soc.* **73** (1951) 5274.
89. R. Hall, E. S. Stern, *J. Chem. Soc.* 1950, 490. Du Pont, US 3536763, 1970 (H. S. Eleuterio, T. A. Koch).
90. Ruhrchemie, DE 2054601, 1970 (W. Rottig, H. Tummes, B. Cornils, J. Weber).
91. Chisso Corp., JP 78039407, 1970 (T. Nakamura, M. Takano, K. Fukatsu). Celanese, DE 2204070, 1971 (R. H. Prinz). Melle-Bezons, FR 2058532, 1969.
92. Teijin Chem., JP 79036278, 1977 (T. Harada, O. Magoshi). Teijin Chem., JP 79011139, 1977 (T. Harada, K. Nakagawa).
93. H. J. Arpe, *Chem. Ztg.* **97** (1973) 53. J. F. Jelinek, *Sb. Pr. Vyzk. Chem. Vyuziti Uhli Dehtu Ropy* **14** (1976) 99 – 113, see Chem. Abstr. 85 (1976) 176767. E. Santoro, M. Chiavarini, *J. Chem. Soc. Perkin Trans. 2,* 1978, no. 3, 189 – 192.
94. J. Bathory, O. Repasy, *Petrochemia* **16** (1976) no. 4 – 5, 94 – 97. Ruhrchemie, DE-OS 2045669, 1970 (H. Tummes, G. Schiewe, B. Cornils, W. Pluta, J. Falbe).
95. Eastman-Kodak, US 2811562, 1957 (H. J. Hagemeyer).
96. Hoechst, DE-OS 1768274, 1968 (G. Jacobson, H. Fernholz, D. Freudenberger).
97. BASF, DE 2758883, 1977 (M. Distler, W. Goetze). Soc. Chim. des Charbonnages, BE 843827, 1975 (P. Couderc, S. Hilmoine).
98. Union Carbide, US 4215076, 1976 (M. L. Deem, K. C. Stueben).
99. A. Mueller, *Seifen Öle Fette Wachse* **102** (1976) no. 1, 11 – 12.
100. Toyo Soda, JP 8218643, 1980 (*Chem. Abstr.* **97** (1982) 38670).
101. PCUK, FR 2338920, 1976 (J. C. Volta, J. M. Cognion).
102. Ajinomoto, JP 242278, 1971 (Y. Matsuzawa, T. Yamashita, S. Ninagawa).
103. Cosden Technology, DE-OS 2 501 341, 1974 (J. Watson).
104. Kuraray Chem. Co., JP 8035063, 1978 (Y. Tokito, N. Yoshimura, M. Tamura). J. Vojtko et al., *Zb. Pr. Chemickotechnol. Fak. SVST* 1977, 195 – 200 (Chem. Abstr. 89 (1978) 146285).
105. BASF, EP 4881, 1978 (W. Sauer, W. Fliege, C. Dudeck, N. Petri).
106. Givaudan, FR 2231650, 1973 (I. Huang, L. M. Polinski, K. K. Rao).
107. Sumitomo, EP 34430, 1980 (T. Takano, G. Suzukamo, M. Ishino, K. Ikimi).
108. A. Knorr, A. Weissenborn, E. Laage, US 1899340, 1933.

109. A. Lethbridge et al., *J. Chem. Soc. Perkin Trans. 1* 1973, no. 1, 35 – 38.
110. B. Cornils, R. Payer, *Chem. Ztg.* **98** (1974) 596 – 606.
111. H. Siegel, W. Himmele, *Angew. Chem.* **92** (1980) 182 – 187; *Angew. Chem. Int. Ed. Engl.* **19** (1980) 178. BASF, DE-OS 2132414, 1971 (W. Himmele, H. Siegel, W. Aquila, F. J. Mueller). Ethyl Corp., US 3907847, 1970 (K. A. Keblys).
112. J. M. Brown, *Chem. Ind.* (*London*) **1982**, no. 10, 737. Agency of Ind. Sci. Technol., JP 7757108, 1975 (I. Ogata, M. Tanaka, Y. Ikeda, T. Hayashi). L. Marko in [1], p. 224
113. I. L. Belostotskaya, G. A. Khmeleva, *Termokatal. Metody Pererab. Uglevodorodnogo Syr'ya* 1969, 221 – 224.
114. W. Hafner et al., *Chem. Ber.* **95** (1962) 1575.
115. Heraeus, DE-OS 2613645, 1976 (R. Siepmann, E. Hopf). Engelhard Industries, US 3372199, 1964 (P. N. Rylander, N. Himmelstein). D. V. Sokol'skii et al., *Zh. Org. Khim.* **13** (1977) no. 1, 77 – 80.
116. Seibi Chem. Co., JP 6820 447, 1964 (S. Abe, K. Sato, T. Asami, T. Amakasu, T. Hakura). A. W. Burgstahler, L. O. Weigel, C. G. Shaefer, *Synthesis* 1976, no. 11, 767 – 768.
117. Hercules, US 3658917, 1965 (R. F. Heck). J. B. Melpolder, R. F. Heck, *J. Org. Chem.* **41** (1976) 265. Givaudan, DE-OS 2627112, 1975 (A. J. Chalk, S. A. Magennis).
118. Nisshin Flour Milling, DE 2404158, 1974 (K. Kogure, K. Nakagawa). Kanebo, JP 76101949, 1975 (Y. Izuka, Y. Sawa, T. Kawashima, S. Miura).
119. Boots Pure Drug. Co., FR 1545270, 1966.
120. Ohta Pharmaceutical Co., JP 81154428, 1981. Ota Seiyaku Co., JP 7831637, 1976 (T. Tsukama et al.).
121. Kohjin Co., JP 7746034, 1973, JP 7626835, 1974 (S. Yoshimura, S. Takahashi, M. Ichino, T. Nakamura).
122. Daito Koeki Co., JP 7912336, 1977 (T. Shimazaki, Y. Otsuka, H. Kondo). Daito Koeki Co., JP 7882740, 1976 (Y. Fujimori).
123. Nado Kenkyusho, JP 7797930, 1976 (M. Arakawa). Fujimoto Pharmaceutical Co., JP 7762233, 1975 (M. Fujimoto, S. Nakayama).
124. W. Behrends, L. M. van der Linde, *Perfum. Essent. Oil Rec.* **58** (1967) no. 6, 372 – 378.
125. W. Himmele, E. H. Pommer, *Angew. Chem.* **92** (1980) 176 – 181; *Angew. Chem. Int. Ed. Engl.* **19** (1980) 184. BASF, DE-OS 2727482, 1977 (W. Himmele, E. H. Pommer, N. Goetz, B. Zeeh); DE-OS 2921221, 1979 (W. Himmele, W. Heberle, F. W. Kohlmann, W. Wesenberg). Hoffmann-LaRoche, DE-OS 2752135, 1976 (A. Pfiffner, K. Bohnen); EP 5541, 1978 (K. Bohnen, A. Pfiffner)
126. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **16** (1978) suppl. 1, 659.
127. Universal Oil Prod., FR 1496304, 1966. D. V. Sokol'skii et al., *Zh. Prikl. Khim.* **49** (1976) no. 2, 407 – 411. BASF, DE-OS 2 832 699, 1978 (G. Heilen, A. Nissen, W. Koernig, M. Horner, W. Fliege, G. Boettger). A. M. Pak et al., *Zh. Org. Khim.* **17** (1981) no. 6, 1176 – 1180.
128. Universal Oil Prod., GB 1086447, 1966 (A. Friedman, J. Levy).
129. K. Kogami, J. Kumanotani, *Bull. Chem. Soc. Jpn.* **46** (1973) no. 11, 3562 – 3565.
130. Hasegawa Co., JP 7250095, 1968, JP 7301379, 1968 (J. Kumanotani et al.).
131. BASF, DE-OS 3105446, 1981, EP 58326, 1982 (W. Gramlich, G. Heilen, H. J. Mercker, H. Siegel).
132. Rhône-Poulenc, DE 1145161, 1963 (I. Scriabine). Rhône-Poulenc, US 3023247, 1960. Givaudan, EP 43526, 1981 (R. Valentine, H. A. Brandman).
133. Givaudan, EP 52775, 1980 (Y. Crameri, P. A. Ochsner, P. Schudel).
134. A. J. Chalk, S. A. Magennis, *J. Org. Chem.* **41** (1976) no. 7, 1206.
135. Q. Bone, K. P. Ryan, *Histochem. J.* **4** (1972) 331. Union Carbide, US 2 941 859, 1960 (M. L. Fein, E. M. Filachione). BASF, DE 2 215 948, 1972 (H. Erdmann, F. F. Miller). Union Carbide, US 4 244 876, 1978 (G. H. Warrer, L. F. Theiling, M. G. Freid). Schülke & Mayr, DE 3 032 794, 1980 (W. Münzenmaier et al.). S. S. Subramanian, *Sci. Res. News* **1** (1976) 42.
136. J. Furukawa, J. Nishimura, *J. Polym. Sci. Polym. Symp.* **56** (1976) 437.
137. Union Carbide, WO 9 740 002, 1997 (D. L. Packett et al.). Union Carbide, WO 9 740 001, 1997 (J. B. Briggs et al.). Union Carbide, WO 9 739 998, 1997 (A. S. Guram et al.). DSM, DuPont, WO 9 733 854, 1997 (P. M. Burke et al.). Union Carbide, US 5 312 996, 1995 (D. L. Packett).
138. D. Jingfa, X. Xinhua, C. Haiying, J. Anren, *Tetrahedron* **48** (1992) 3503. Tonen Sekiyu Kagaku, JP 01 190 647, 1990 (H. Inagaki et al.). Toa Nenryo Kogyo, JP 62 029 546, 1987 (H. Furukawa). USSR, NL 7 312 323, 1975 (A. T. Menyailo et al.). Bayer, DE 2 252 719,

1974 (H. Waldmann, W. Schwerdtel, W. Swodenk). Bayer, DE 2 252 674, 1974 (H. Waldmann, W. Schwerdtel, W. Swodenk). Bayer, DE 2 201 455, 1973 (H. Waldmann, W. Schwerdtel, W. Swodenk).
139. Bayer, DE 2 329 586, 1975 (L. Imre, A. Klein, K. Wedemeyer). Bayer, DE 2 201 411, 1973 (L. Irme, K. Wedemeyer).
140. Mitsui Toatsu Chemicals, JP 57 203 024, 1983. Mitsubishi Petrochemical Co., JP 05 310 631, 1994 (J. Ookago, N. Sumya, S. Ichikawa). Mitsubishi Petrochemical Co., EP 537 625, 1993 (Y. Ohgomori, S. Ichikawa, Y. Shuji, T. Yoneyama, N. Sumitani).
141. M. Tamura, S. Kumano, *Chem. Econ. Eng. Rev.* **12** (1980) 32. J. Maeda, R. Yoshida, *Bull Chem. Soc. Jpn.* **41** (1968) 2969. BASF, DE 2 401 553, 1975(R. Kummer).
142. Daicel Chemicals, JP 08 040 968, 1996 (M. Nishioka, M. Saito). Union Carbide, GB 739 128, 1955 (R. I. Hoaglin, R. G. Kelso). Distillers, GB 698 736, 1953 (R. G. Hall, B. K. Howe).
143. Ruhrchemie, DE 1 618 384, 1972(J. Falbe). B. Cornils, R. Payer, H. Tummes, J. Weber, *Eur. Chem. News* **27** (1975) no. 712, 36.
144. ESPE, DE 3 902 417, 1991(O. Gasser, R. Guggenberger, K. Ellrich). ESPE, US 4 131 729, 1979 (W. Schmitt, R. Purrmann, P. Jochum, W. D. Zahler).
145. Philips Gloeilampenfabr., NL 6501841, 1965. M & T Chemicals, GB 1168225, 1967.
146. Ethyl Corp., US 4209643, 1978 (K. H. Shin).
147. Union Rhein. Braunkohlen Kraftstoff, BE 882655, 1979 (J. Korff, M. Fremery, J. Zimmermann). Union Rhein. Braunkohlen Kraftstoff, BE 890964, 1980. Mitsubishi Gas Chem. Ind., GB 2079746, 1980.
148. O. Bayer (ed.): *"Aldehyde",* see [75] p. 413. H. Meerwein (ed.): "Acetale," in Houben-Weyl: Methoden der organischen Chemie, vol. VI/3, Thieme Verlag, Stuttgart 1965, p. 203. F. A. J. Meskens, *Synthesis* 1981, 501 – 522. A. J. Showler, P. A. Darley, *Chem. Rev.* **67** (1967) 427.
149. Henkel, EP 39029, 1980 (H. Upadek, K. Bruns). Henkel, DE 2648109, 1976 (K. Bruns et al.). Firmenich, NL 7415594, 1973 (G. Ohloff, W. K. Giersch). Naarden Int., NL 7305488, 1973.
150. Shell Int. Res., BE 816134, 1973.
151. BASF, DE 2509739, 1975 (P. Engler et al.).
152. BASF, DE 2636 278, 1976 (H. Zinke-Allmang, W. Scheidmeir).
153. Akademie Wissenschaften der DDR, DD 142044, 1979 (J. Freiberg, H. Seeboth, R. Goersch, H. Bergner).
154. Henkel, DE 2533048, 1975 (H. Moeller, J. Conrad, C. Gloxhuber, H. J. Thimm).
155. Inst. Français du Pétrole, US 2997480, 1959.
156. L. F. Tietze: "Nachweismethoden und quantitative Bestimmung von Aldehyden," in [1], p. 673.E. Pretsch, T. Clerc, J. Seibl, W. Simon: *Tabellen zur Strukturaufklärung organischer Verbindungen,* Springer Verlag, Heidelberg 1976. D. H. Williams, I. Fleming: *Spektroskopische Methoden in der organischen Chemie,* Thieme Verlag, Stuttgart 1971. E. Heuser: "Analytik der Carbonylgruppen," in Houben-Weyl: Methoden der organischen Chemie, vol. 2, Thieme Verlag, Stuttgart 1953, p. 434 – 467. W. Büchler, A. Becker, A. Walter in: *Methodicum Chimicum,* **vol. 1/1,** Thieme Verlag, Stuttgart 1973, p. 216.

Aldols → Aldehydes, Aliphatic and Araliphatic
Alfol Process → Fatty Alcohols
Algin → Polysaccharides
Alkali and Chlorine Products → Chlorine
Alkali and Chlorine Products → Sodium Hydroxide

Alkaloids

Many of the alkaloids described under this keyword possess pharmaceutical importance. For more details on their use in chemotherapy, see the corresponding keywords.

Osamu Hoshino, Science University of Tokyo, Tokyo, Japan

Masatoshi Murakata, Science University of Tokyo, Tokyo, Japan

Miyuki Ishizaki, Science University of Tokyo, Tokyo, Japan

Tetsuji Kametani, Hoshi University, Tokyo, Japan

Toshio Honda, Hoshi University, Tokyo, Japan

Keiichiro Fukumoto, Tohoku University, Sendai, Japan

Masataka Ihara, Tohoku University, Sendai, Japan

1. Introduction 2
2. Phenylalkylamines 3
2.1. Simple Aromatic Amines 3
2.2. Ephedra Alkaloids 3
3. Pyrrolidine, Piperidine, and Pyridine Alkaloids 4
3.1. Pyrrolidine Alkaloids 4
3.2. Piperidine Alkaloids 5
3.2.1. Hemlock Alkaloids 5
3.2.2. Areca Alkaloids 5
3.2.3. Pomegranate Alkaloids 6
3.2.4. Lobelia Alkaloids 6
3.2.5. Histrionicotoxins 6
3.3. Pyridine Alkaloids 6
3.3.1. Tobacco Alkaloids 6
3.3.2. Theonelladin Alkaloids 7
4. Tropane Alkaloids 7
4.1. Chemical and Physical Properties 7
4.2. Preparation 8
4.3. Pharmacological Properties . . 9
5. Pyrrolizidine Alkaloids 9
6. Indolizidine Alkaloids 10
6.1. Ipomoea Alkaloids 10
6.2. Thylophora Alkaloids 11
6.3. Elaeocarpus Alkaloids 11
6.4. Dendrobatid Alkaloids 11
6.5. Miscellaneous 11
7. Quinolizidine Alkaloids 12
7.1. Lupinine and Related Alkaloids 12
7.2. Lythraceae Alkaloids 13
8. Quinoline, Quinazoline, and Acridone Alkaloids 14
8.1. Quinoline Alkaloids 14
8.2. Quinazoline Alkaloids 14
8.3. Acridone Alkaloids 15
9. Isoquinoline Alkaloids 15
9.1. Simple Isoquinoline Alkaloids 16
9.2. Phenylisoquinoline Alkaloids . 16
9.3. 1-Benzylisoquinoline Alkaloids 17
9.4. 2-Benzylisoquinoline Alkaloids 18
9.5. Pavine and Isopavine Alkaloids 18
9.6. Bisbenzylisoquinoline Alkaloids 19
9.7. Cularine and Related Alkaloids 21
9.8. Proaporphine Alkaloids 21
9.9. Aporphine Alkaloids 21
9.10. Protoberberine Alkaloids 22
9.11. Protopine Alkaloids 24
9.12. Phthalide Alkaloids 25
9.13. Ochotensine and Related Alkaloids 25
9.14. Rheadan Alkaloids 26
9.15. Morphine and Related Alkaloids 26
9.16. Hasubanan Alkaloids 28
9.17. Emetine and Related Alkaloids 28
9.18. Erythrina Alkaloids 29
9.19. Cephalotaxine Alkaloids 30
9.20. Protostephanine Alkaloids . . . 30
9.21. Benzophenanthridine Alkaloids 30
9.22. Dibenzopyrrocoline Alkaloids . 31

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_353

9.23. Phenethylisoquinoline Alkaloids . . . 31
9.24. Miscellaneous Isoquinoline Alkaloids . . . 32
10. Amaryllidaceae Alkaloids . . . 33
10.1. Lycorine Type . . . 33
10.2. Crinane Type . . . 33
10.3. Nariclasine Type . . . 34
10.4. Galanthamine Type . . . 34
10.5. Montanine Type . . . 35
10.6. Miscellaneous . . . 35
11. Indole Alkaloids . . . 36
11.1. Non-Tryptamine Type . . . 36
11.2. Non-Isoprenoid Type . . . 36
11.3. Isoprenoid Type . . . 37
11.3.1. Mould Metabolites . . . 37
11.3.2. Ergot Alkaloids . . . 38
11.3.3. Monoterpenoid Alkaloids . . . 39
11.3.3.1. Corynanthe Group . . . 39
11.3.3.2. Heteroyohimbine . . . 40
11.3.3.3. Yohimbine . . . 40
11.3.3.4. Sarpagine . . . 42
11.3.3.5. Ajmaline . . . 42
11.3.3.6. Picraline . . . 42
11.3.3.7. Vobasine . . . 43
11.3.3.8. Pleiocarpamine . . . 43
11.3.3.9. Strychnine . . . 43
11.3.3.10. Akuammicine . . . 44
11.3.3.11. Ellipticine . . . 44
11.3.3.12. Aspidospermine . . . 45
11.3.3.13. Aspidofractine . . . 45
11.3.3.14. Eburnamine . . . 46
11.3.3.15. Catharanthine . . . 46
11.3.3.16. Ibogamine . . . 46
11.3.3.17. Cleavamine . . . 47
11.3.3.18. Andranginine . . . 47
11.3.4. Bis-Indole Alkaloids . . . 47
11.3.5. Biogenetically Related Quinoline Alkaloids . . . 48
12. Lycopodium Alkaloids . . . 50
13. Terpene Alkaloids . . . 50
13.1. Monoterpene Alkaloids . . . 50
13.2. Sesquiterpene Alkaloids . . . 51
13.3. Diterpene Alkaloids . . . 52
14. Steroidal Alkaloids . . . 53
14.1. Alkaloids of the Apocynaceae . . . 53
14.2. Buxus Alkaloids . . . 53
14.3. Alkaloids of the Asclepiadaceae 53
14.4. Solanum Alkaloids . . . 54
14.5. Veratrum and Fritillaria Alkaloids . . . 54
14.6. Phyllobates Alkaloids . . . 54
15. Miscellaneous Alkaloids . . . 55
15.1. Muscarine Alkaloids . . . 55
15.2. Imidazole Alkaloids . . . 55
15.3. Polyamine Alkaloids . . . 55
15.4. Securinine Alkaloids . . . 56
15.5. Tetrodotoxin . . . 56
15.6. Aaptamine and Related Alkaloids . . . 56
15.7. Eudistomidins . . . 57
15.8. Manzamines . . . 58
15.9. Panuamine and Haliclonadiamine . . . 58
16. References . . . 59

1. Introduction

The typical alkaloid is an alkaline compound having one or more heterocyclic nitrogen atoms. It is found in nature, usually in plants. Some of the alkaloids show pronounced physiological action, and several alkaloids, e.g., reserpine, quinine, and morphine, have been used in medicine. It is no longer customary to include the purine and pyrimidine bases in the alkaloid group.

The alkaloids usually occur in the form of salts with organic acids (oxalic, acetic, malic, meconic, or succinic acid). Therefore these compounds most often have been isolated as free bases by treatment of the source plants with an alkali solution. Traditionally, the alkaloids have been purified by recrystallization of the free base or a salt, but chromatographic methods are also used frequently. For detection of alkaloids the Mayer reagent (potassium mercuric iodide), Wagner reagent (iodine-potassium iodide), Dragendorff reagent (potassium bismuth iodide), Sonnenschein reagent (phosphomolybdic acid), and Scheibler reagent (phosphotungstic acid) are used. These are called alkaloid reagents.

Structure determination has been carried out by the usual chemical and physicochemical methods. The frequently used chemical methods are the following: zinc distillation, dehydrogenation on palladium or selenium, and alkali fusion to determine the basic skeleton and Hofmann degradation, Emde degradation, and von Braun degradation to elucidate the nitrogen part of the molecule. In addition, controlled oxidation and reduction processes have been used to determine some functional groups. Today,

physicochemical techniques are the best way to elucidate the structure of an alkaloid. The following methods should be available to an alkaloid laboratory: ultraviolet and infrared spectroscopy, proton and carbon-13 nuclear magnetic resonance (NMR), optical rotatory dispersion (ORD), circular dichroism (CD), mass spectrometry, and X-ray analysis. In the near future, computer analysis systems involving all these physicochemical methods are expected to increase the effectiveness of structural determination.

Biosynthesis. Alkaloids are biosynthesized from amino acids via amines or aldehydes. This mechanism was established by BARTON, BATTERSBY, and SCOTT with tritium, carbon-13, and carbon-14 tracer experiments. The key reactions in biogenesis are Mannich-type condensation, aldol condensation, methylation with methionine, and oxidative coupling. In some cases, biogenetic theories have been fruitful guides in further structure determination (emetine and strychnine) and total synthesis (morphine). In this article, alkaloids are classified structurally, in the standard way, and one or two typical compounds in each group are briefly discussed in respect to chemistry and application.

2. Phenylalkylamines

Benzylamines and phenethylamines from natural sources are included in this chapter. Although ephedra alkaloids are also phenethylamine derivatives, they are discussed separately because of their characteristic pharmacological activities.

2.1. Simple Aromatic Amines

The *trans*-configuration of capsaicine (**1**), isolated from *Capsicum* species, was established by stereospecific synthesis [1]. Phenethylamines, such as dopamine (**2**) and tyramine (**3**), are important precursors of isoquinoline alkaloids and important neuroamines. Acanthoidine (**4**) and acanthoine (**5**) are found in *Carduus acanthoides.* The structure of the cactus alkaloid mescaline (**6**) was confirmed by synthesis from 3,4,5-trimethoxybenzoic acid [2].

Physical Properties.

Mescaline (**6**) [*54-04-6*], $C_{11}H_{17}NO_3$, M_r 211: *mp* 35 – 36 °C, *bp* 180 °C (at 1.6 kPa). Hydrochloride: *mp* 181 °C.

Pharmacological Properties. Mescaline (**6**) depresses the central nervous system of the frog and also mammals. The best-known action of this and the related cactus alkaloids is the production of color hallucinations. Acanthoidine (**4**) and acanthoine (**5**) exhibit some hypertensive activity.

Therapeutic Use: Experimental psychotomimetic.

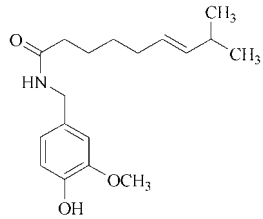

1 Capsaicine

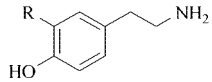

2 R = OH Dopamine
3 R = H Tyramine

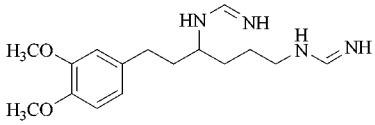

4 Acanthoidine

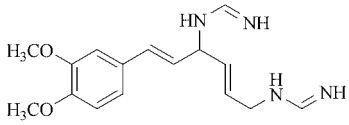

5 Acanthoine

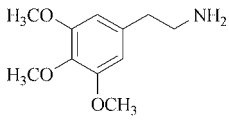

6 Mescaline

2.2. Ephedra Alkaloids

Plants of genus *Ephedra* contain optically active alkaloids. Some are useful in medicine. The two major alkaloids, (−)-ephedrine (**7**) and (+)-pseudoephedrine (**8**), are diastereoisomers.

Physical Properties.

(−)-*Ephedrine* (**7**) [*299-42-3*], $C_{10}H_{15}NO$, M_r 165: *mp* 38.1 °C. Hydrochloride: *mp* 217 – 218 °C or 220 – 221 °C, $[\alpha]_D$ − 34 ° (wa-

ter). (±)-Form: *mp* 79 °C. Hydrochloride: *mp* 187 – 188 °C.

(+)-*Pseudoephedrine* (**8**) [*90-82-4*], $C_{10}H_{15}NO$, M_r 175: *mp* 119 °C, $[\alpha]_D^{20}$ +51 ° (ethanol). Hydrochloride: *mp* 181 – 182 °C, $[\alpha]_D^{20}$ +62 ° (water). (±)-Form: *mp* 118 °C.

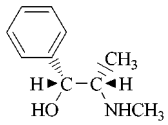

7 Ephedrine

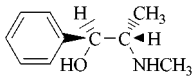

8 Pseudoephedrine

Preparation. (−)-Ephedrine can be obtained from the Ma Huang drug (*Ephedra "vulgaris", E. sinica* Stapf, *E. equisetina* Bunge) and other types of *Ephedra* [3]. The powdered drug is alkalized and extracted with benzene. The solution obtained is extracted with dilute acid, and the ephedrine crystallized after concentration of the aqueous phase.

Synthetic ephedrine continues to replace natural ephedrine.

Pharmacological Properties. Ephedrine (**7**) was originally employed as a mydriatic until the similar pharmacological action of adrenaline was observed. The mydriatic action of (−)-ephedrine is greater than that of (+)-ephedrine.

Therapeutic Use: Adrenergic (bronchodilator) [3].

Trade Names for (−)-*Ephedrine:* L-Sedrin (Lilly), Lexofedrin (Lexington). *Hydrochloride:* Darophedrin (Daro), Eggophedrin (Eggochemia), Ephedrin [Nagai] (Dainippon), Ensan Ephedrin (Meiji, Taisho), Ephedrin (Knoll), Ephetonin (Merck). *Sulfate:* Ectasule-Minus (Fleming), Ephedsol (Lannett), Iso-Phedrizem (Zemmer), Isofedrol (Blue Line). *Resinate:* Ephedrate (Upjohn), Ephedrol (Lilly), Ephedronal (Blue Line).

Trade Names for (+)-*Pseudoephedrine Hydrochloride:* Afrinol (Schering Corp.), Besan (Tutag), Eltor (Dow Chemical), First Sign (Williams), Novafed 120 (Dow Chemical), Sinufed (Hauck), Sudaped (Wellcome), Symptom 2 (Park Davis), Tussaphed (Knoll). *Sulfate:* Drixoral (Schering Corp.).

3. Pyrrolidine, Piperidine, and Pyridine Alkaloids

3.1. Pyrrolidine Alkaloids

Hygrine (**9**, racemic mixture) is a liquid tertiary aminoketone. The freshly isolated compound has a small negative optical rotation, but the free base racemizes rapidly. Cuscohygrine (**10**), isolated from the Peruvian *coca* shrub together with hygrine, is also oxidized to *N*-methylproline with chromic acid. The structure of shihunine (**11**), found in the Chinese plant *Dendrobium loliohense,* was determined by conversion into 1-methyl-2-phenylpyrrolidine [4].

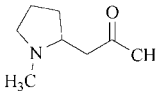

9 Hygrine

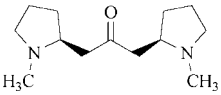

10 Cuscohygrine

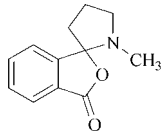

11 Shihunine

α-Kainic acid, isolated from the algae *Digenea simplex* [5] and *Centrocerus clavulatum* [6], has been shown to possess composition and configuration **14** on the basis of chemical [7] and X-ray evidence [8]. Kainic acid was enantioselectively synthesized from (*S*)-glutamic acid. Ene reaction of the 1,6-diene **12** produced the pyrrolidine **13**, which was converted into (−)-α-kainic acid [9].

(−)-Domoic acid (**19**) was found in the red algae *Chondria armata* [10]. The structure first proposed was revised by X-ray analysis and the total synthesis [11] of the optically active compound **19**. Diels-Alder reaction between the enone **15**, derived from (*S*)-glutamic acid, and diene **16** gave the bicyclic compound **17**, which was transformed into (−)-domoic acid (**19**) via the aldehyde **18**.

Physical Properties.

Hygrine (**9**) [*496-49-1*], $C_8H_{15}NO$, M_r 141: *bp* 76.5 °C (at 1.5 kPa), *bp* 81 °C (at 1.9 kPa). Picrate: *mp* 149 – 151 °C.

Cuscohygrine (**10**) [*454-14-8*], $C_{13}H_{24}N_2O$, M_r 224: *bp* 169 – 170 °C (at 3.1 kPa), *bp* 118 – 125 °C (at 0.3 kPa). Picrate: *mp* 216 – 217 °C (decomp.).

α-Kainic acid (**14**) [*487-79-6*], $C_{10}H_{15}NO_4$, M_r 213: *mp* 251 °C (decomp.), $[\alpha]_D^{24}$ – 14.8 ° (water).

Pharmacological Properties. Both kainic acid (**14**) and domoic acid (**19**) show potent neuronal excitatory activity.

α-Kainic acid is used therapeutically as an ascaricide.

Trade Names for α-Kainic Acid: Digenin (Takeda).

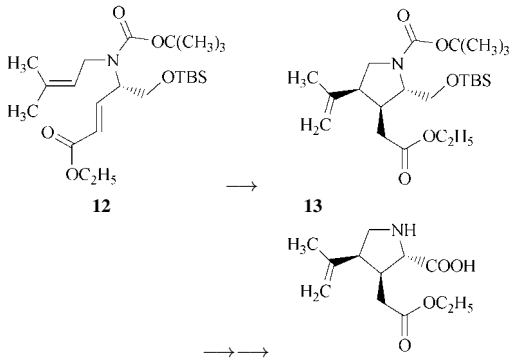

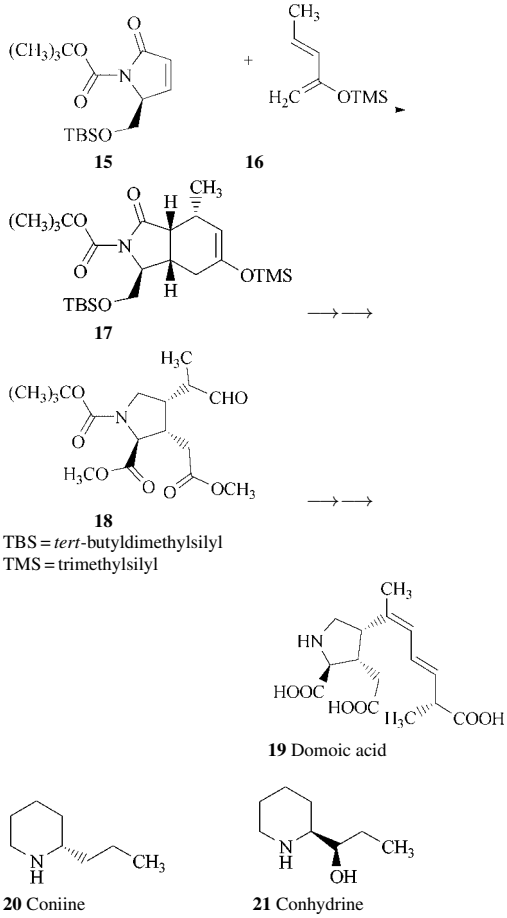

3.2. Piperidine Alkaloids

Some monocyclic piperidine alkaloids are classified according to the natural source into hemlock, areca, pomegranate, sedum, and lobelia alkaloids. Histrionicotoxins, the principal active ingredients of neotropical arrow poisons, have a spiro fused piperidine.

3.2.1. Hemlock Alkaloids

One of major components of hemlock, *Conium maculatum,* is coniine (**20**). The relative configuration of the other hemlock alkaloid, conhydrine (**21**), was assigned to the erythro form.

Physical Properties.

Coniine (**20**) [*458-88-8*], $C_8H_{17}N$, M_r 127: *bp* 166 °C, $[\alpha]_D^{19}$ + 16 ° (water).

Conhydrine (**21**) [*495-20-5*], $C_8H_{17}NO$, M_r 143: *mp* 121 °C, *bp* 226 °C, $[\alpha]_D$ + 10 ° (water).

Pharmacological Properties. Coniine (**20**) is highly toxic, inducing nausea and vomiting at an early stage and paralysis of the motor nerve endings followed by depression of the central nervous system.

3.2.2. Areca Alkaloids

Arecoline (**22**) [*63-75-2*], $C_8H_{13}NO_2$, M_r 155: *bp* 209 °C, *bp* 92 – 93 °C (at 0.9 kPa), is isolated from seeds of the betel nut palm *Areca catechu.* The compound is a powerful stimulant for the central nervous system and is used therapeutically as a cholinergic and anthelmintic.

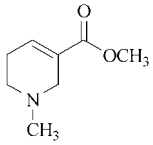

22 Arecoline

3.2.3. Pomegranate Alkaloids

The root bark of the pomegranate, *Punica granatum,* contains several alkaloids. Pelletierine (**23**) was synthesized by acylation of α-picolyllithium with acetic anhydride followed by hydrogenation. Pelletierine racemizes on treatment with bases and for this reason is usually isolated as the racemate.

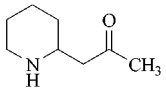

23 Pelletierine

Pelletierine (**23**) [*2858-66-4*], $C_8H_{15}NO$, M_r 141. Sulfate: *mp* 135 – 138 °C, $[\alpha]_D^{25}$ – 29.5 ° (water).

Pharmacological Properties. Pelletierine has been employed as an anthelmintic: It is highly toxic to tapeworms.

3.2.4. Lobelia Alkaloids

*Lobelia inflata,*commonly known as Indian tobacco, is an emetic herb rich in alkaloids. It is found in the Great Smoky Mountains. The main alkaloid, lobeline (**24**), was synthesized via a Mannich condensation of glutaraldehyde, methylamine, and benzoylacetic acid at pH 4.

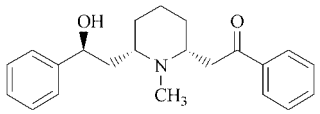

24 Lobeline

Lobeline (**24**) [*90-69-7*], $C_{22}H_{27}NO_2$, M_r 337: *mp* 130 – 131 °C, $[\alpha]_D^{15}$ – 43 ° (ethanol). Hydrochloride: *mp* 178 – 180 °C, $[\alpha]_D^{20}$ – 43 ° (water).

Therapeutical Use: Respiratory stimulant.

3.2.5. Histrionicotoxins

Histrionicotoxin (**25**), isolated from *Dendrobates histrionicus,* contains the unique 1-azaspiro[5.5]undecane ring system. The structure and the stereochemistry at the C-2, C-6, C-7, and C-8 positions were elucidated by X-ray diffraction analysis of the hydrochloride [12].

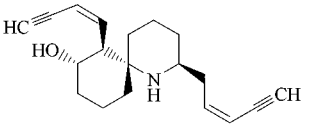

25 Histrionicotoxin 283A

Pharmacological Properties. Histrionicotoxin (**25**) is a highly active venom and a mucosal-tissue irritant for mammals and reptiles. Both histrionicotoxin and its perhydroderivative selectively bind to the acetylcholine receptor and interrupt transsynaptic transmission of neuromuscular impulses [13]. Both compounds block postsynaptic membrane depolarization but do not interfere with acetylcholine binding. These toxins may prevent membrane depolarization by reversible bonding to the receptor ion channel or "ion conductance modulator" [13].

3.3. Pyridine Alkaloids

3.3.1. Tobacco Alkaloids

Tobacco alkaloids, include nicotine (**26**), nicotyrine (**27**), and anabasine (**28**). Nicotine is the principal alkaloid from tobacco plants (genus *Nicotiana*). Chromic acid oxidation gives nicotinic acid (**29**), whereas distillation from lime affords pyrrole and methylamine. Nicotyrine (**27**) is the corresponding tetradehydro compound, whereas anabasine (**28**) is an isomer of nicotine.

Physical Properties.

Nicotine (**26**) [*54-11-5*], $C_{10}H_{14}N_2$, M_r 162: *bp* 123 – 125 °C (at 2.3 kPa), $[\alpha]_D^{20}$ – 169 °. Hydrochloride: $[\alpha]_D^{20}$ + 104 ° (water).

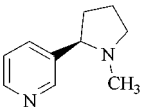

26 Nicotine

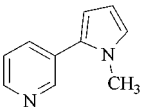

27 Nicotyrine

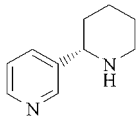

28 Anabasine

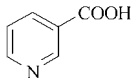

29 Nicotinic acid

Nicotyrine (**27**) [*487-19-4*], $C_{10}H_{10}N_2$, M_r 158: *bp* 280–281 °C (at 99 kPa), *bp* 150–151 °C (at 2.1 kPa). Tartrate: *mp* 105–106 °C.

Anabasine (**28**) [*494-52-0*], $C_{10}H_{14}N_2$, M_r 162: *bp* 270–272 °C, *bp* 145–147 °C (at 1.9 kPa). Hydrochloride: $[\alpha]_D$ +16.5 ° (water).

Pharmacological Properties. LD_{50} in mice: 0.3 mg/kg i.v., 230 mg/kg orally for nicotine.

Nicotine (**26**) possesses the property of inducing an initial and transient stimulation, followed by depression, and finally paralysis of the autonomic ganglion.

Therapeutic Use: Ectoparasiticide. Nicotine has been used as an anthelmintic. Anabasine is used as an insecticide.

3.3.2. Theonelladin Alkaloids

Theonelladins A–D (**30**)–(**33**) were isolated from the Okinawan marine sponge *Theonella awinhoei*. Their structures were determined spectroscopically [14]. Theonelladin C (**32**) was isolated as a mixture with traces of anallogues having branched alkyl chains. These alkaloids exhibit potent antineoplastic activity against marine lymphomas L1210 cells (IC_{50} 4.7 µg/mL for A, 1.0 µg/mL for B, 3.6 µg/mL for C, 1.6 µg/mL for D) and human epidermoid carcinoma KB cells (IC_{50} 10 µg/mL for A, 3.6 µg/mL for B, 10 µg/mL for C, and 5.2 µg/mL for D) in vitro [14]. They also have powerful Ca^{2+}-releasing activity (twenty times more potent than caffein).

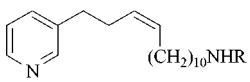

30 R = H Theonelladin A
31 R = CH_3 Theonelladin B

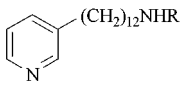

32 R = H Theonelladin C
33 R = CH_3 Theonelladin D

4. Tropane Alkaloids

About 50 alkaloids are known that have a tropane skeleton (**34**). All tropane alkaloids possess a hydroxy or acyloxy group at the C-3 position. They are found in the plants of Solanaceae, Convolvulaceae, Erythroxylaceae, and Rhizophoraceae. The most important source plants are *Atropa belladonna, Erythroxylum coca, Datura metel, D. stramonium, Scopola carniolica,* and *Hyoscyamus niger*.

34

4.1. Chemical and Physical Properties

Structure. The structures of important tropane alkaloids were determined by degradation methods about 100 years ago. The stereochemistries were worked out about 50 years ago. Both types of configurational isomers are known for the relative position of the nitrogen bridge and the substituent at C-3. For example, (−)-hyoscyamine (**35**) is *trans*, whereas tropacocaine (**37**) is *cis*.

The *cis* conformation was determined by acyl migration: Heating *N*-acetylnorpseudotropine hydrochloride (**38**) at 160 °C gave quantitatively *O*-acetylnorpseudotropine hydrochloride (**40**) via the cyclic intermediate **39**. On the other hand, no migration occurs with the α-hydroxyl compounds.

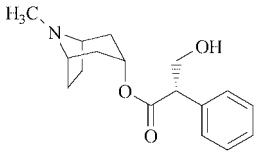

35 (−)-Hyoscyamine
(±)-form: Atropine

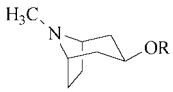

36 R = H Pseudotropine
37 R = COC_6H_5 Tropacocaine

The acyl part of (−)-hyoscyamine (**35**) is (−)-tropic acid, which easily undergoes racemization. Acidic or basic treatment of (−)-hyoscyamine affords the racemate, atropine.

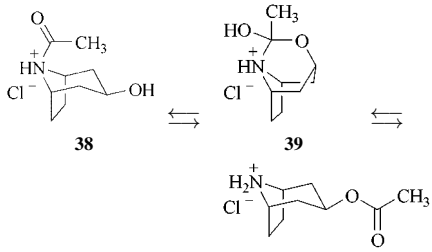

38 **39** **40**

Cocaine (**42**) belongs to the β-series and carries a methoxycarbonyl group at the C-2 position. The absolute configuration was determined by the correlation with L-glutamic acid. Ecgonine (**41**) is the corresponding hydroxy acid.

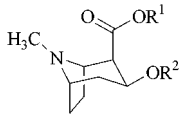

41 $R^1 = R^2 = H$ Ecgonine**42** $R^1 = CH_3$; $R^2 = COC_6H_5$ Cocaine

The tropane alkaloid (−)-hyoscine (**43**) has an oxygen group on the pyrrolidine ring. Its racemate is the important alkaloid scopolamine. This alkaloid was converted by either acidic or alkaline hydrolysis into oscine (**45**) via scopine (**44**). The tetrahydrofuran ring formation was interpreted as a backside nucleophilic attack of the hydroxy oxygen upon either bridgehead of the epoxide ring, i.e., the C-6 or the C-7 position.

Physical Properties.

Atropine (Hyoscyamine) (**35**) [*51-55-8*], $C_{17}H_{23}NO_3$, M_r 289. (±)-Form: *mp* 114–116 °C. (−)-Form: *mp* 108.5 °C, $[\alpha]_D^{20}$ −21.0° (ethanol). Hydrobromide of the (−)-form: *mp* 152 °C.

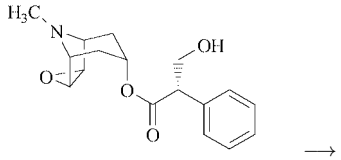

43 (−)-Hyoscine
(±)-form: Scopolamine

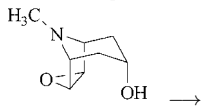

44 Scopine

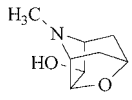

45 Oscine

Tropine (**47**) [*120-29-6*], $C_8H_{15}NO$, M_r 141: *mp* 63 °C, *bp* 233 °C.

Pseudotropine (**36**) [*135-97-7*], $C_8H_{15}NO$, M_r 141: *mp* 109 °C, *bp* 241 °C. Hydrochloride: *mp* 282 °C (decomp.).

Cocaine (**42**) [*50-36-2*], $C_{17}H_{21}NO_4$, M_r 303: *mp* 98 °C, *bp* 187–188 °C (at 0.01 kPa), $[\alpha]_D^{18}$ −35° (50 % ethanol), $[\alpha]_D^{20}$ −16° (chloroform). Hydrochloride [*53-21-4*]: *mp* 195 °C, $[\alpha]_D$ −72° (water). Nitrate [*5913-62-2*]: *mp* 58–63 °C. Sulfate [*5913-65-5*].

Ecgonine (**41**) [*481-37-8*], $C_9H_{15}NO_3$, M_r 185: *mp* 198 °C, $[\alpha]_D^{15}$ −45° (water). Hydrochloride: *mp* 246 °C: $[\alpha]_D^{15}$ −59° (water).

Hyoscine (**43**) (±)-form: Scopolamine [*51-34-3*], $C_{17}H_{21}NO_4$, M_r 303: $[\alpha]_D^{20}$ −28° (water). Hydrobromide [*114-49-8*]: *mp* 195 °C, $[\alpha]_D^{25}$ −24 to −26° (water).

4.2. Preparation

Synthesis. An effective synthesis of tropane alkaloids was developed by Robinson and Schöpf. Tropinone (**46**) was obtained in one step by the reaction of succinic aldehyde, methylamine, and acetonedicarboxylic acid [15]. Catalytic hydrogenation of **46** gave tropine (**47**). Tropine and its derivatives can be prepared commercially by this method. If this reaction is carried out with monomethyl acetonedicarboxylate, the ecgonine derivatives can be synthesized.

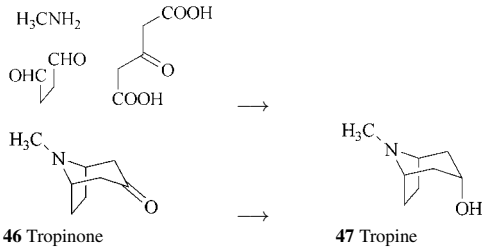

46 Tropinone **47** Tropine

Commercial Production. The procedure to extract tropane alkaloids from Solanaceae plants, e.g., the deadly nightshade, is as follows: The dried and finely divided materials are extracted with ethanol. Evaporation of the extract gives a residue, which is further extracted with water and then with 0.5 % sulfuric acid. The aqueous extracts are washed with ether or chloroform and then made basic by addition of ammonia under cooling. The alkaloids are extracted with ether and then with chloroform. Evaporation of the solvents gives the alkaloidal fractions, which are then converted to salts and purified by recrystallization.

Cocaine (**42**) is produced industrially from natural sources. The alkaloid is extracted from coca leaves with dilute sulfuric acid and hydrolyzed to ecgonine (**41**). The purified base is esterified with methanol and then acylated with benzoyl chloride to cocaine.

4.3. Pharmacological Properties

Atropine first stimulates but eventually depresses the central nervous system, causing hallucinations, delirium, and convulsions, and finally coma. Scopolamine has anticholinergic activity. Cocaine (**42**) is a potent anaesthetic, mydriatic, and sedative.

Therapeutic Use.

Atropine is used as an anticholinergic. It is employed principally in medicine as a mydriatic, to dilate the pupil of the eye. It also is used when paralysis of parasympathetic nervous activity is desired, e.g., to suppress branchial or intestinal spasms.

Cocaine is used as a topical anesthetic and a central nervous system stimulant.

Ecgonine is a topical anesthetic, and *scopolamine* is a anticholinergic. The last has been used also as a sedative, a preanesthetic, and to control motion sickness.

Trade Names.

Atropine: Atropinol (Dr. Winzer). *Sulfate:* Atropen (Survival Tech.), Atropine Dispersa (Dispersa), Atropisol (Cooper Vision), Atrop (Sigma), Eyesule (Duncan Flockhart), Isopto-Atropin (Alcon), Liotropina (Sifi), Ocu-Atro (Orion), Oftan-atropin (Star), Skiatropin (Chauvin-Blache), Atratan (Mallinckrodt), Tropintran (Kwizda), Ryusan Atropine (Santen, Iwaki, Torii, Takeda, Tanabe), Opulets Atropine (Alcon), AtropinPOS (Ursapharm), Atropin sulfat (Hameln), Atropin Drobena (Drobena), Atropin Sulfuric Thilo (Thilo), Atropin Sufuric (Cascan), Minims Atropine (S. N. P.). *Methionitrate:* Eumycrin (Winthrop).

Cocaine: Psicaine (Merck), Delcaine (Roquies), Ensan Cocaine (Sankyo, Takeda, Shionogi, Dainippon).

Scopolamine Borate: Boro-Scopol (Winzer). *Hydrobromide:* Isopto Hyoscine (Alcon), Scopolamine Dispersa (Dispersa), Scops (Spret-Mauchant), Triptone (Commerce Drug), Minims Hyoscine (S. N. P.), Hisco (Kyorin). *Butyl bromide:* Buscopan (Boehringer Ingelheim).

5. Pyrrolizidine Alkaloids

Occurrence and Preparation. A number of pyrrolizidine alkaloids are isolated from *Senecio* plants, the largest genus of the Compositae. Generally hydrolysis of the alkaloids gives necine bases and necic acids. Thus retronecine (**48**) was obtained by the hydrolysis of echimidine (**56**), monocrotaline (**57**), crobarbatine (**58**), and indicine *N*-oxide (**59**). Heliotridine (**49**) is formed from heliotrine (**55**). Alkaloids **48** and **49** are stereoisomers at the C-7 position. All four possible stereoisomers, hastanecine (**50**) [16], dihydroxyheliotridane (**51**) [16], platynecine (**52**), and turneforcidine (**53**), are obtained by degradation of parent alkaloids or as natural products. The structures were established by careful spectroscopical analysis. They can be synthesized by many methods [16–23].

Trachelanthamidine (**54**) [18] was synthesized based on the presumed biosynthetic pathway [19]: The dialdehyde **60** was allowed to stand for several days at pH 7 to give **61**. Sodium borohydride reduction gave racemic trachelanthamidine. Asymmetric synthesis was also attempted [20].

(±)-Retronecine (**48**) and (±)-turneforcidine (**53**) were synthesized via sulfenocycloamination ring closure. Regio- and stereoselective [3.3] sigmatropic rearrangement between the 3-pyrrolidinone **62** and the *cis*-butenediol derivative **63** gave the ketone **64**, which was then transformed into the amine **65**. Treatment of the amine **65** hydrochloride with phenylsulfenyl chloride and base-induced ring closure produced the pyrrolizidine (**66**), which was converted into (±)-retronecine (**48**) and (±)-turneforcidine (**53**) in few steps [23].

Alexine (**67**) [*116174-63-1*], $C_8H_{15}NO_4$, M_r 189: *mp* 162 – 163 °C, $[\alpha]_D^{20}$ +40 ° (water), was isolated from the pods of *Alexa leiopetala* (Leguminoceae) and its structure was elucidated on the basis of spectroscopic evidence (MS, ^{1}H and ^{13}C NMR). A total synthesis of **67** starting from D-glucose has been reported [25].

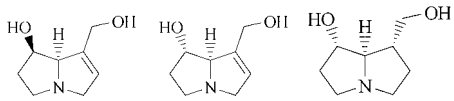

48 Retronecine **49** Heliotridine **50** Hastanecine

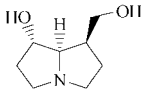

51 Dihydroxyheliotridane

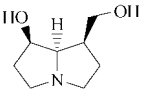

52 Platynecine

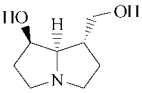

53 Turneforcidine

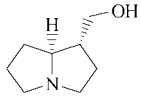

54 Trachelanthamidine

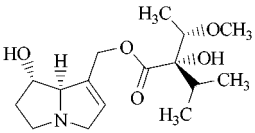

55 Heliotrine

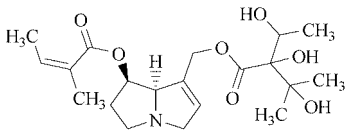

56 Echimidine

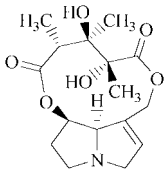

57 Monocrotaline

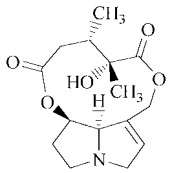

58 Crobarbatine

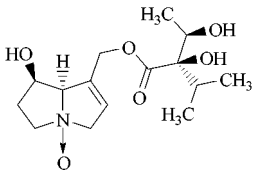

59 Indicine *N*-oxide

Pharmacological Properties. The pyrrolizidine alkaloids exhibit selective toxic action on the liver. This toxicity is associated with the metabolic conversion of the pyrrolizidine nucleus to a pyrrole [24]. *Indicine N-oxide* has proved to be highly active against transplanted neoplasms, Walker-256 carcinosarcoma, melanoma B-16, leukaemia L-1210, leukaemia P-388, and leukaemia P-1534 in mice and rats, inhibiting up to 91 percent of transplanted growth. *Alexine* is weak inhibitor of glucosidase activity [26].

6. Indolizidine Alkaloids

6.1. Ipomoea Alkaloids

Ipalbine (**68**), which is the *O*-β-D-glucoside of ipalbidine (**69**), occurs in *Ipomoea alba.* It is dextrorotatory, $[\alpha]_D^{23}$ +32.5° (ethanol). The structure was confirmed by the total synthesis from 2-methoxypyrroline and methyl acetoacetate [27]. The seeds of *Ipomoea muricata*, a plant grown in Senegal, affords ipomine (**70**) and ipalbidine (**69**). Acid hydrolysis of ipomine gives ipalbidine, D-glucose, and *p*-coumaric acid, whereas enzymatic hydrolysis with emulsin yields ipalbidine (**69**) [28].

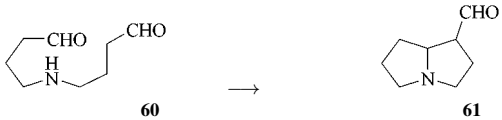

60 → **61**

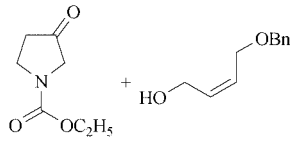

62 + 63 →

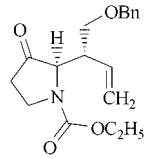

64 →

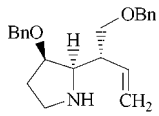

65 Bn = $CH_2C_6H_5$

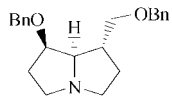

→ **66**

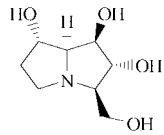

67 Alexine

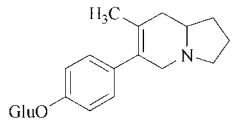

68 Ipalbine
Glu = glucose

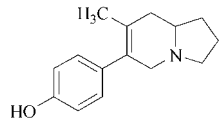

69 Ipalbidine

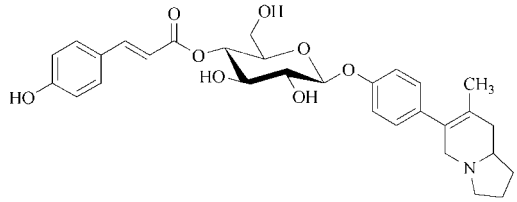

70 Ipomine

6.2. Thylophora Alkaloids

Septicine (**71**) and thylophorine (**72**), both isolated from *Ficus septica,* have been synthesized by a number of groups [29–35]. Both are synthesized through 1,3-dipolar cycloaddition [35].

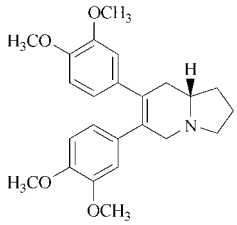

71 Septicine

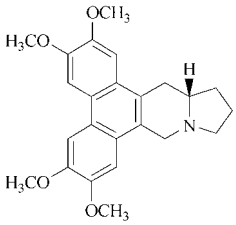

72 Tylophorine

Pharmacological Properties. Thylophorine (**72**) and its positional isomers showed antitumor activity.

6.3. Elaeocarpus Alkaloids

The structure of elaeocarpine (**73**) was determined by X-ray analysis of its hydrobromide [36]. Racemates of elaeokanine A (**74**) and elaeokanine C (**75**) have been synthesized by a nitrone 1,3-dipolar cycloaddition [37], a versatile method for synthesis of elaeokanine alkaloids [38].

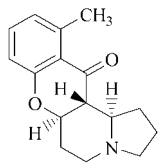

73 Elaeocarpine

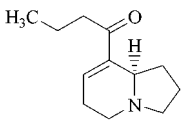

74 Elaeokanine A

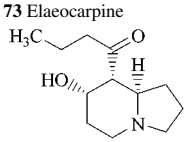

75 Elaeokanine C

6.4. Dendrobatid Alkaloids

Pumiliotoxins A (**76**) and B (**77**) and gephyrotoxin (**78**) were isolated from skin extracts of frogs belonging to the Dendrobatid family [12]. Pumiliotoxins A and B have in common the unusual 6-alkylideneindolizidine (1-azabicyclo[4.3.0]nonane) ring system. They differ only in the side chain. Overman synthesized the dendrobatid toxin 251 D (**79**), which had been isolated from *Dendrobates tricolor*, via iminium-ion vinylsilane cyclization [39].

Structure of gephyrotoxin (**78**) was established by X-ray crystallographic analysis of its *p*-bromobenzoate [40]. Total synthesis of gephyrotoxin was accomplished by Kishi and Hart [41–43]. The stereoselectivity of the Kishi synthesis was controlled through hydrogenation reactions. On the other hand, Hart controlled the stereochemistry on cyclization of carbinolamide with formic acid.

Pharmacological Properties. Pumiliotoxin B (**77**) shows powerful cardiotonic and myotonic activity, which is believed to involve calcium-dependent mechanisms [44].

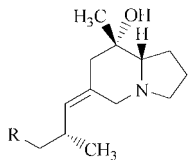

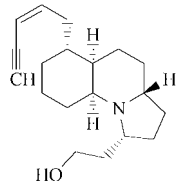

78 Gephyrotoxin

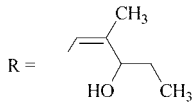

76 Pumiliotoxin A

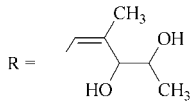

77 Pumiliotoxin B

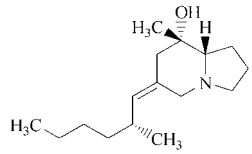

79 Toxin 251 D

6.5. Miscellaneous

Certain legume forages used for cattle feed have been known to produce excessive salivation when infested by *Rhizoctonia leguminicola.* This was shown to be caused by the presence of *slaframine* (**80**) [45]. The structure and absolute stereochemistry were established by Rine-

HART [46]. Slaframine has been synthesized by three groups [47–49]. Dieckmann cyclization was used to construct the indolizidine skeleton in the first two syntheses [47], [48], whereas the key step of the other synthesis was the intramolecular Diels-Alder reaction [49].

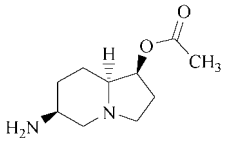

80 Slaframine

Slaframine itself is not bioactive but is oxidized in vivo to a compound that can interact with muscarinic acetylcholine receptors, producing the observed symptoms [50].

Castanospermine (**81**) [*79831-76-8*], $C_8H_{15}NO_4$, M_r 175: *mp* 212 – 215 °C (ethanol), $[\alpha]_D^{25}$ + 79.7° (water), was isolated from the toxic seeds of the Australian legume *Castanospermum australe* and its structure was determined by spectroscopic considerations [51]. The first total synthesis of (+)-castanospermine starting from D-glucose was reported by GANEM [52]. Since castanospermine shows inhibitory activity against tumors and HIV, extensive synthetic studies have been performed. Synthesis of **81** and related alkaloids is reviewed in [53]. More recently **81** has been prepared from methyl α-D-glucopyranoside [54].

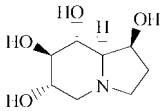

81 Castanospermine

7. Quinolizidine Alkaloids

The quinolizidine ring system is a common structural unit in many alkaloids. Lupinine and related alkaloids and the Lythraceae alkaloids are discussed in this section.

7.1. Lupinine and Related Alkaloids

Occurrence and Synthesis. Bicyclic, tricyclic, and tetracyclic compounds are found in a large family of the *Leguminosae.* Representative compounds are lupinine (**82**), cytisine (**84**), and sparteine (**85**). Matrine (**86**) represents another type of tetracyclic lupine alkaloids.

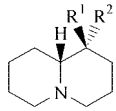

82 R^1 = H, R^2 = CH_2OH Lupinine
83 R^1 = CH_2OH, R^2 = H Epilupinine

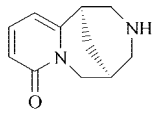

84 Cytisine

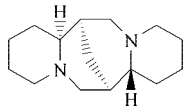

85 Sparteine

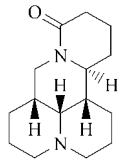

86 Matrine

Lupinine (**82**), isolated from yellow lupine seed, is the simplest member of the family. A synthesis along lines of the presumed biosynthesis was carried out by VAN TAMELEN [55].

Lupinine (**82**) is the less stable isomer. Its axial hydroxymethyl group is hydrogen-bonded to the nitrogen atom.

Sparteine (**85**) has four asymmetric centers, two of them interdependent because of the bridge. There should be three pairs of enantiomers, and in fact three different diastereoisomers have been isolated from the plants *Cytisus ratisbonensis, C. scoparius, Lupinus barbiger, L. laxus, L. luteum, L. niger, Retama sphaerocarpa, Sarothamnus scoparius, Spartium junceum,* and *S. scoparium.* 8-Oxo-α-isosparteine (**89**) was synthesized very simply by condensation of the racemic or *meso* forms of anaferine (**87**) with formaldehyde. The intermediate products are the spiranoid ketals **88**, which were converted on treatment with acetic anhydride into **89** [56].

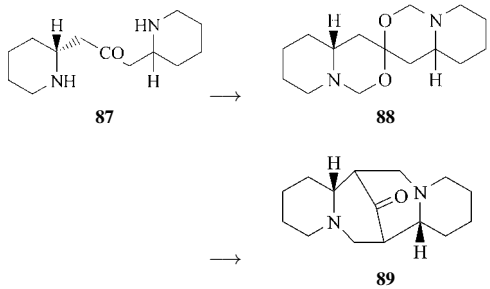

87 → **88** → **89**

Matrine (**86**) occurs in many species of *Sophora.* Most of the structural features were recognized through degradation experiments. The straightforward synthesis starts from the Schiff base **90** of the ethyl ester of β-alanine and ethyl β-ketopimelate. Hydrogenation of **90** affords **91**, which is converted by Dieckmann condensation and hydrolysis to **92**. Monocyanoethylation by the enamine method gives **93**, and hydrogenation provides ($\pm$)-matrine (**86**).

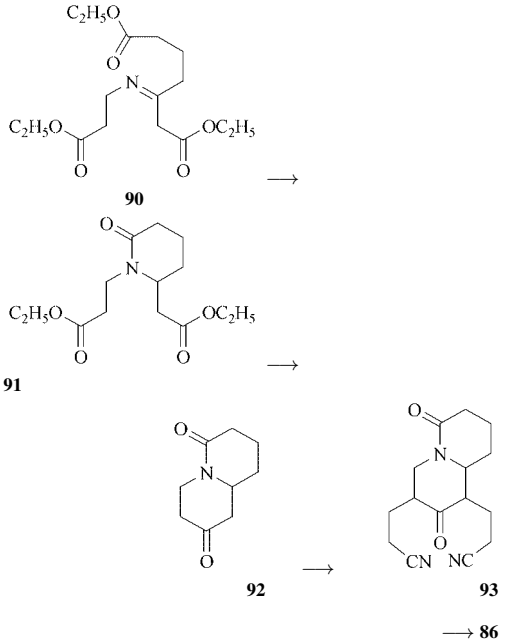

Physical Properties.

Lupinine (**82**) [*486-70-4*], $C_{10}H_{19}NO$, M_r 169: *mp* 68.5 – 69.2 °C, *bp* 160 – 164 °C (at 0.5 kPa), *bp* 269 – 270 °C (at 100.6 kPa), $[\alpha]_D^{28}$ $-21°$ (ethanol). Hydrochloride: *mp* 208 – 213 °C, $[\alpha]_D$ $-14°$ (water).

Cytisine (**84**) [*485-35-8*], $C_{11}H_{14}N_2O$, M_r 190: *mp* 152 – 153 °C, *bp* 218 °C (at 0.3 kPa), $[\alpha]_D^{17}$ $-120°$ (water).

Matrine (**86**) [*519-02-8*], $C_{15}H_{24}N_2O$, M_r 248: *mp* 87 °C, $[\alpha]_D^{20}$ $+38°$ (ethanol). Picrate: *mp* 167 – 169 °C.

Sparteine (**85**) [*90-39-1*], $C_{15}H_{26}N_2$, M_r 234: *bp* 173 °C (at 1.1 kPa), $[\alpha]_D^{21}$ $-16.4°$ (ethanol).

Pharmacological Properties. Sparteine (**85**) has little effect on the central nervous system but paralyzes the motor nerve terminations and the sympathetic ganglia. It depresses circulation.

Therapeutic Use of Sparteine: Oxytocic.

Trade Names for Sparteine: Spartopan (Paillusseau). *Sulfate:* Actospar (Sandoz), Depasan (Giulini), Spartocin (Ayerst, Benton), Tocosamine sulfate (Trent). *Theophyllinate:* Spartofillina (Amelix, Terapeutico M. R.), Uterospan (Nihonzoki).

7.2. Lythraceae Alkaloids

Occurrence and Synthesis. Macrocyclic quinolizidine alkaloids, isolated from members of the *Lythraceae,* can be classified into three structural types. These are represented by the biphenyllactone cryogenine (**94**), the diphenyl ether lactone vertaline (**95**) [57–59], and the carbocyclic biphenyl lythrumine (**97**). Several total syntheses of lactonic lythraceae alkaloids have been reported, most using Mannich reactions of pelletierine with substituted benzaldehydes, biaryls, or diaryl ethers to assemble the quinolizidine moiety.

Vertaline (**95**) and decaline (**96**) were synthesized via [3 + 2] cycloaddition [59].

Pharmacological Properties. Some lythraceae alkaloids increase blood glucose level and lower mean blood pressure; they show sedative activity in dogs and guinea pigs.

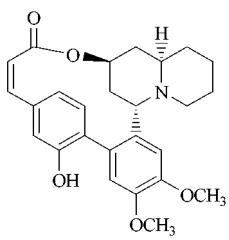

94 Cryogenine

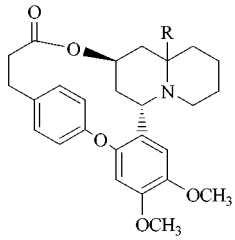

95 R = α-H Vertaline
96 R = β-H Decaline

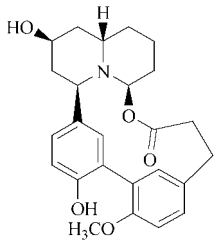

97 Lythrumine

8. Quinoline, Quinazoline, and Acridone Alkaloids

8.1. Quinoline Alkaloids

Occurrence and Synthesis. Most quinoline alkaloids are derived from anthranilic acid. Dictamnine (**98**) is a representative of furoquinoline alkaloids. Ribalinidine (**99**), which has the dihydropyranoquinoline structure, is isolated as the L-form. Bucharamine (**100**), spectabiline (**103**), and balfourodine (**104**) contain the dihydrofuroquinoline ring system. Ravenine (**101**) and ravenoline (**102**) possess 1,1-dimethylallyl and 1,2-dimethylallyl moieties, respectively.

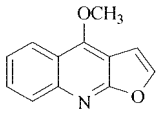

98 Dictamnine

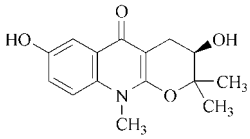

99 Ribalinidine

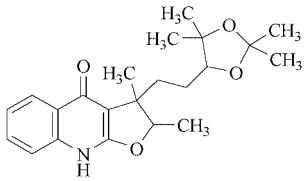

100 Bucharamine

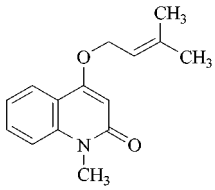

101 Ravenine

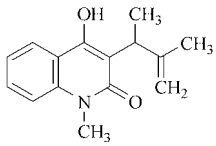

102 Ravenoline

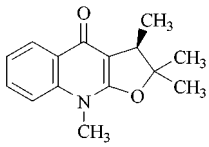

103 Spectabiline

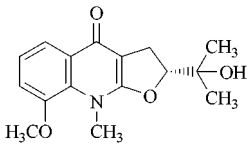

104 Balfourodine

Pumiliotoxin C (**105**) was originally isolated from the skin of the neotropical frogs *Dendrobates pumilio* and *D. auratus* [40], [60]. This alkaloid has been synthesized by several groups [61], [62]. YAMAMOTO developed a practical synthesis. The enone **106**, prepared via the Stobbe condensation of 2-methylcyclohexanone, was converted to the *cis*-fused hexahydroindanone **107** by stereoselective hydrogenation and elaborated via alkylaluminum-promoted Beckmann rearrangement of the oxime tosylate **108** to (±)-pumiliotoxin C (**105**) [63].

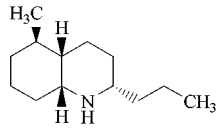

105 Pumiliotoxin C

↑

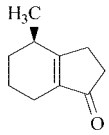

106

→

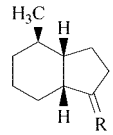

107 R = O
108 R = oxime tosylate

Physical Properties. *Dictamnine* (**98**) [*484-29-7*], $C_{12}H_9NO_2$, M_r 199: *mp* 133 °C. Hydrochloride: *mp* 170 °C (decomp.). Picrate: *mp* 163 °C.

8.2. Quinazoline Alkaloids

Occurrence and Synthesis. A number of quinazoline alkaloids are isolated from the *Rutaceae* plants. Anthranilic acid is thought to be the precursor in the biogenesis. Vasicine (**109**), isolated from *Adhatoda vasica* and *Pegunum harmala,* can be oxidized with $KMnO_4$ to give 4-quinazolone. Arborine (**110**), glycosminine (**111**), glomerine (**112**), crysogine (**113**), rutecarpine (**114**), euxylophoricines A (**115**) and C (**116**), and evodiamine (**117**) were synthesized in a one-pot reaction by iminoketene cycloaddition [64], [65]. Treatment of the sulfinamide anhydrides **120** and **121**, derived from anthranilic acid (**118**) or *N*-methylanthranilic acid (**119**), with imines or primary or secondary amides gives the quinazolone derivatives **123**. The iminoketenes **122** were postulated as the intermediates.

Physical Properties.

Vasicine (**109**) [*6159-55-3*], $C_{11}H_{12}N_2O$, M_r 188. (±)-Form: *mp* 210 °C (from ethanol). (−)-Form: *mp* 212 °C, $[\alpha]_D^{14}$ −254° (*c* = 2.4 in chloroform) or −62° (*c* = 2.4 in ethanol). Hydrochloride: *mp* 208 °C.

Rutecarpine (**114**) [*84-26-4*], $C_{18}H_{13}N_3O$, M_r 287: *mp* 259.5 – 260 °C.

Evodiamine (**117**) [*518-17-2*], $C_{19}H_{17}N_3O$, M_r 303: *mp* 278 °C, $[\alpha]_D^{15}$ +352° (acetone), $[\alpha]_D$ +440° (chloroform).

Pharmacological Properties. Vasicine (**109**) is an active bronchodilator. On injection rutecarpine (**114**) increases arterial pressure.

8.3. Acridone Alkaloids

Acridone alkaloids are also derived biogenetically from anthranilic acid. It was reported that certain bases of this class possess antitumor properties. Acronycine (**125**) was synthesized from the acridone derivative **124** [66].

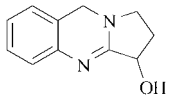

109 Vasicine

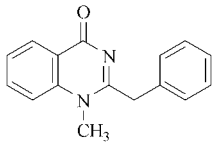

110 Arborine

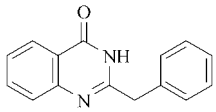

111 Glycosminine

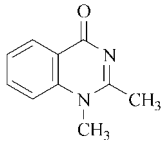

112 Glomerine

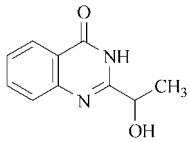

113 Crysogine

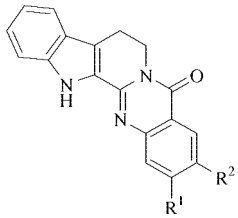

114 $R^1 = R^2 = H$
Rutecarpine
115 $R^1 = R^2 = OCH_3$
Euxylophoricine A
116 $R^1 + R^2 = OCH_2O$
Euxylophoricine C

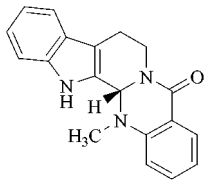

117 Evodiamine

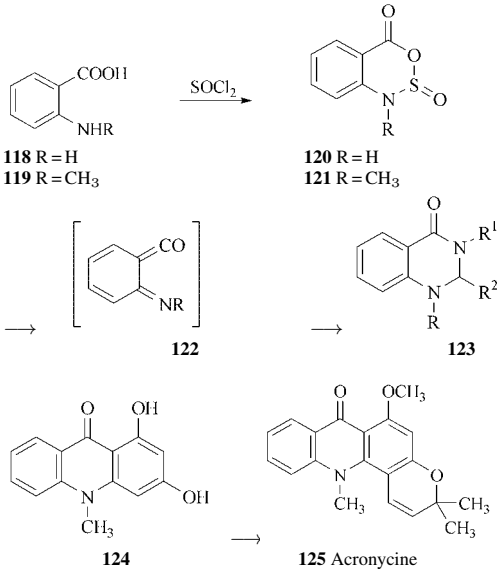

9. Isoquinoline Alkaloids

The isoquinoline alkaloids comprise a large number of naturally occurring compounds of varying complexity [67–69]. They are classified as follows:

1) simple isoquinoline alkaloids
2) phenylisoquinoline alkaloids
3) 1-benzylisoquinoline alkaloids
4) 2-benzylisoquinoline alkaloids
5) pavine and isopavine alkaloids
6) bisbenzylisoquinoline alkaloids
7) cularine and related alkaloids
8) proaporphine alkaloids
9) aporphine alkaloids
10) protoberberine alkaloids
11) protopine alkaloids
12) phthalide alkaloids
13) ochotensine and related alkaloids
14) rheadan alkaloids
15) morphine and related alkaloids
16) hasubanan alkaloids
17) emetine and related alkaloids
18) erythrina alkaloids
19) cephalotaxine alkaloids
20) protostephanine alkaloids
21) benzophenanthridine alkaloids
22) dibenzopyrrocoline alkaloids
23) phenethylisoquinoline alkaloids

Furthermore, some isoquinoline derivatives have been found in bacteria. The 1-

benzylisoquinoline alkaloids, derived biogenetically from tyrosine or phenylalanine, are transformed to a number of the above isoquinoline alkaloids by phenol oxidation [70] or modification in the presence of some plant enzymes [71], [72].

9.1. Simple Isoquinoline Alkaloids

Occurrence and Synthesis. The simple isoquinoline alkaloids, such as salsoline (**126**), anhalamine (**127**), corypalline (**128**), thalifoline (**129**), and pilocereine (**130**), are tetrahydroisoquinolines substituted in the C-1 position by methyl, hydrogen, oxygen, or isobutyl groups. The stereochemistry of the optically active alkaloids is determined by conversion to an amino acid of known configuration. Optical rotatory dispersion (ORD) and circular dichroism (CD) spectra are useful for determination of absolute configuration. The isoquinoline moiety is usually obtained by the Bischler-Napieralski reaction [73], the Pictet-Spengler reaction [74], which mimics physiological reaction conditions, or the Pomeranz-Fritsch reaction [75] and its modification [76]. All of these classical methods for forming the isoquinoline ring have been employed for the synthesis of the simple isoquinoline alkaloids and some of the alkylisoquinoline alkaloids.

Pilocereine (**130**), isolated from *Lophocereus schottii,* is the only example of a phenolic oxidation product in simple isoquinoline alkaloids. Tracer experiments showed that the phenolic oxidation of lophocerine (**131**, racemic mixture) by enzymes in the plant furnishes pilocereine (**130**) [77], [78]. This coupling process was reproduced in the laboratory by potassium ferricyanide oxidation of **131** at pH 6 [79].

Physical Properties.

Salsoline (**126**) [*89-31-6*], $C_{11}H_{15}NO_2$, M_r 193: *mp* 221 °C, $[\alpha]_D^{20}$ +34.5° (0.1 M hydrochloric acid). Hydrochloride: *mp* 174 – 175 °C, $[\alpha]_D$ +31.0° (methanol).

Anhalamine (**127**) [*643-60-7*], $C_{11}H_{15}NO_3$, M_r 209: *mp* 189 – 191 °C. Hydrochloride: *mp* 258 °C.

Pilocereine (**130**) [*2552-47-8*], $C_{45}H_{65}N_3O_6$, M_r 744: *mp* 176.5 – 177 °C. Methyl ether: *mp* 103 – 105 °C, solidifies and remelts at 153 – 155 °C.

Pharmacological Properties. Salsoline (**126**) shows an antihypertensive activity.

Therapeutic Use of Salsoline: Antihypertensive.

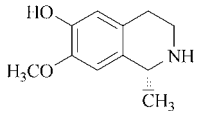

126 (+)-Salsoline

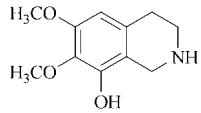

127 Anhalamine

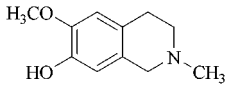

128 Corypalline

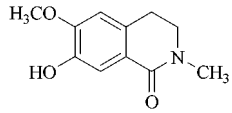

129 Thalifoline

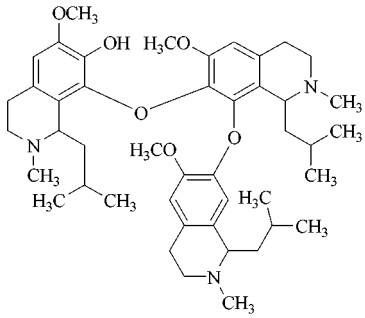

130 Pilocereine

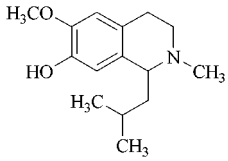

131 Lophocerine

9.2. Phenylisoquinoline Alkaloids

The unique structure of rufescine (**132**), isolated from *Abuta imene* and *A. rulesceus,* was determined by spectroscopic methods and confirmed by synthesis [80].

The structure of the optically active cryptostyline I (**133**) was derived from NMR, UV, and mass spectra. Final confirmation for the structure, as well as the determination of the absolute configuration, was obtained by X-ray analysis [81]. Cryptostyline II (**134**) and III (**135**) were synthesized via the Bischler-Napieralski reaction [81].

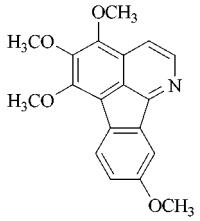

132 Rufescine

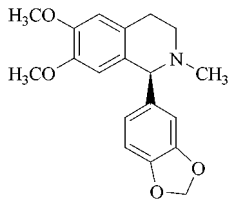

133 Cryptostyline I

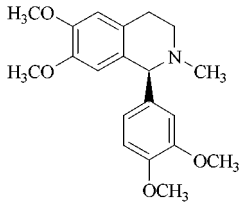

134 Cryptostyline II

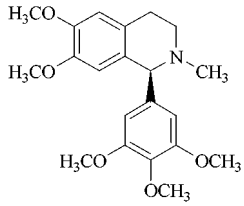

135 Cryptostyline III

9.3. 1-Benzylisoquinoline Alkaloids

Structure, Occurrence, and Synthesis. The structures of the simple 1-benzylisoquinolines and the biogenetically related alkaloids were established by classical methods involving Hofmann degradation and oxidation; the phenolic functions were usually protected by O-ethylation.

The absolute configuration of this class of alkaloids was initially established by Corrodi and Hardegger [82], who ozonized (−)-*N*-norlaudanosoline (**137**) to give the known *N*-carboxyethyl-l-aspartic acid. The stereochemistry of related alkaloids was then determined by comparison with **137** or its derivatives.

The structures of the newer benzylisoquinolines have been determined by physical methods, including mass spectra [83], ^{1}H NMR, ^{13}C NMR, and UV. Further, the chirality of the alkaloids was established by optical rotatory dispersion [84], [85] and circular dichroism [86], where the *S*-series of 1-benzylisoquinolines show positive Cotton effects in the 280 – 240 nm region.

Papaverine (**136**), norlaudanosoline (**137**), coclaurine (**138**), reticuline (**139**), and laudanosine (**140**) are 6,7-dioxygenated isoquinoline derivatives, whereas petaline (**141**) possesses the 7,8-dioxygenated isoquinoline structure. Aobamine (**142**), isolated from *Corydalis ochotensis var. raddeana* [87], has a formyl group and must be derived from protoberberine alkaloids.

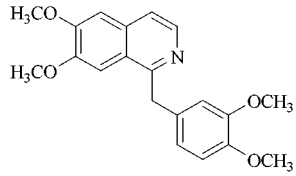

136 Papaverine

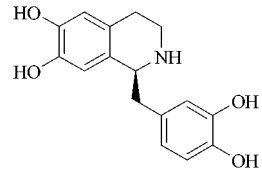

137 Norlaudanosoline

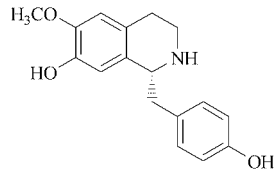

138 (+)-Coclaurine

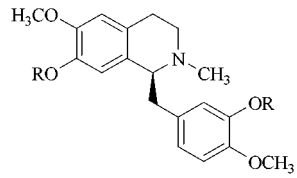

139 R = H (−)-Reticuline
140 R = CH_3 (+)-Laudanosine

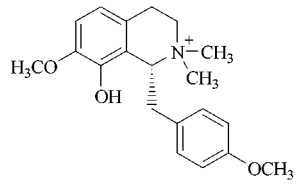

141 Petaline

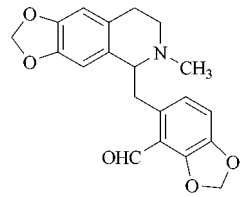

142 Aobamine

The simple benzylisoquinoline alkaloids have mostly been synthesized via the Bischler-Napieralski reaction. Although the phenolic hydroxyl group is usually protected as the benzyl

ether during synthesis, the phenolic function in amides need not be protected in the Bischler-Napieralski reaction [88].

Simple 1-benzylisoquinolines have also been prepared from isoquinolines devoid of a substituent in the 1-position by application of the Grignard reaction [89], the Reissert reaction [90], or the Stevens rearrangement [91], [92].

Papaverine (**136**) was synthesized via the Bischler-Napieralski reaction as follows: the Schotten-Baumann reaction between the phenethylamine **143** and the acyl chloride **144** gave the amide **145**. Cyclization with phosphorous pentoxide afforded the 3,4-dihydroisoquinoline **146**. Mild dehydrogenation of **146** gave papaverine [93].

The Pictet-Gams modification was also used for the synthesis of papaverine [94].

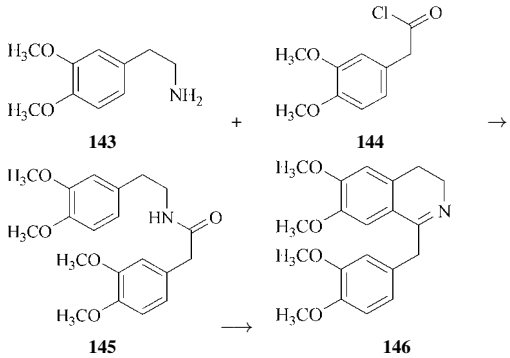

Reticuline (**139**) is an important precursor of protoberberine, morphine, aporphine, spirobenzylisoquinoline, and rheadan alkaloids [95].

Physical Properties.

Papaverine (**136**) [*58-74-2*], $C_{20}H_{21}NO_4$, M_r 339: *mp* 147 °C. Hydrochloride [*61-25-6*]: *mp* 220 – 225 °C.

Coclaurine (**138**) [*486-39-5*], $C_{17}H_{19}NO_3$, M_r 285: *mp* 220 – 221 °C. Hydrochloride: *mp* 263 – 264 °C.

Reticuline (**139**) [*485-19-8*], $C_{19}H_{23}NO_4$, M_r 329. (±)-Form: *mp* 146 °C. Picrate. (±)-Form: *mp* 188 – 190 °C. Perchlorate. (+)-Form: *mp* 203 – 204 °C, $[\alpha]_D^{18}$ + 88.3 ° (ethanol).

Laudanosine (**140**) [*2688-77-9*], $C_{21}H_{27}NO_4$, M_r 357. (±-Form): *mp* 114 – 115.5 °C. (+)-Form: *mp* 89 °C, $[\alpha]_D^{16}$ + 106 ° (97 % ethanol), $[\alpha]_D^{16}$ + 130 ° (chloroform).

Pharmacological Properties. Small doses of papaverine (**136**) produce a light sleep that does not become deeper as the dose is increased because reflex irritability and some stimulating action is brought about. There is a tendency for the heart to slow, an effect more pronounced than with morphine or codeine.

Therapeutic Use of Papaverine: Smooth muscle relaxant and cerebral vasodilator.

Trade Names for Papaverine Chloride: Artegodan (Artesan), Dipav (Lemmon), Pameion (Simes), Panergon (Mack), Pavabid (Marion), Pavakey (Key), Paveron (Luar, Karlspharma), Therapav (Coopar, Berlex), Vasal (Tutag), Ensan Papaverin (Iwaki, Nakakita, Torii). *Papaverine Sulfate:* Copavin (Lilly), Maspaver (Juste).

9.4. 2-Benzylisoquinoline Alkaloids

The structure of sendaverine (**147**), a 2-benzylisoquinoline isolated from *Corydalis aurea* and present in the alkaloid mixture corpaverine [96], [97], was largely determined by IR, UV, NMR, and mass spectra [98]. The related alkaloid corgoine (**148**) was obtained from *Corydalis gortshakovii* [99].

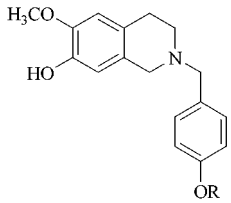

147 R = CH_3 Sendaverine **148** R = H Corgoine

Sendaverine (**147**) was synthesized by utilizing the Bischler-Napieralski or the Pictet-Spengler reaction [98]. Sendaverine (**147**) was also synthesized via isochroman-3-ones prepared by thermolysis of benzocyclobutenes [100].

9.5. Pavine and Isopavine Alkaloids

The pavine alkaloids, represented by argemonine (**149**), are characterized by a cyclooctane ring system. The structure of **149** was determined from spectral data and confirmed by comparison with the synthetic sample obtained from

papaverine by reduction with tin and hydrochloric acid [101]. The absolute configuration was established by correlation with L-aspartic acid as well as by optical rotatory dispersion [102].

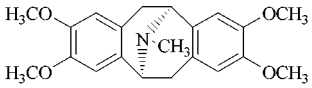

149 Argemonine

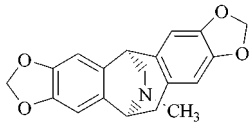

150 Reframidine

The isopavine alkaloid reframidine (**150**) was obtained via the aziridinium salt [103]. The syntheses of several isopavine alkaloids have been reported by DYKE [104] and UMEZAWA [105].

9.6. Bisbenzylisoquinoline Alkaloids

Structure and Synthesis. Bisbenzylisoquinoline alkaloids are classified by the bonding position of the ether. Dauricine (**151**) is a tail-tail mono ether, liensinine (**152**) is a head-tail mono ether, and oxyacanthine (**153**) is a head-head and tail-tail bis ether. Berbamine (**154**) is a different type of head-head and tail-tail bis ether. Trilobine (**155**) has a head-head and tail-tail tri ether structure, and tubocurarine (**156**) is a head-tail bis ether. Thalicarpine (**157**) and pakistanamine (**158**) have aporphine and proaporphine moieties, respectively.

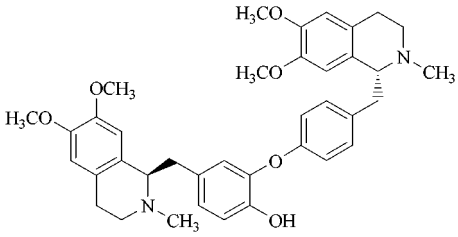

151 Dauricine

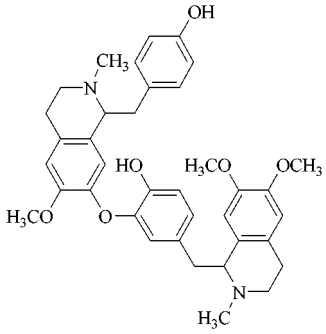

152 Liensinine

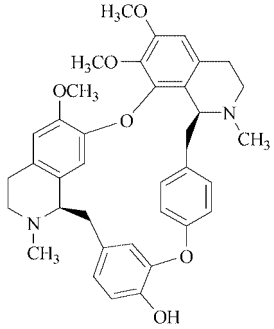

153 Oxyacanthine

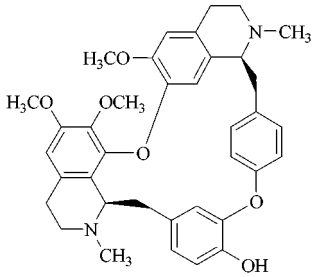

154 Berbamine

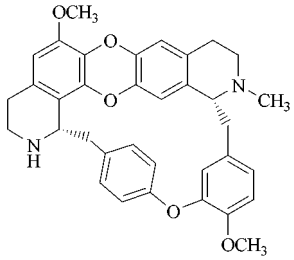

155 (+)-Trilobine

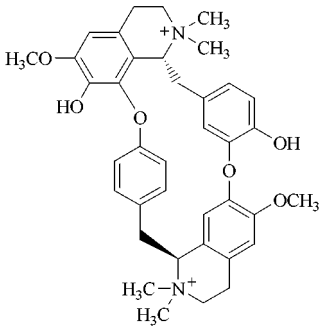

156 Tubocurarine

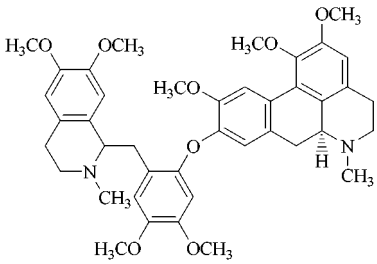

157 Thalicarpine

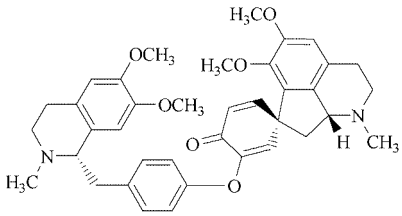

158 (+)-Pakistanamine

Classical degradative methods were used to determine the structures of the bisbenzylisoquinoline alkaloides. In addition, the fragmentation pattern obtained by mass spectroscopy has proved useful in "typing" the bisbenzylisoquinoline alkaloids, while their NMR spectra have been employed not only to locate the position of the methoxy group but also to narrow the stereochemical possibilities. NMR examination has shown that the chemical shifts of the aromatic methoxy group (due to the proximity of the ring system) can aid in identifying the type of oxygen bridge. The absolute configuration of the bisbenzylisoquinolines was readily determined from the chirality of products obtained by cleavage with sodium in liquid ammonia or by measurement of the optical rotatory dispersion (ORD) [106], [107].

Synthesis of bisbenzylisoquinoline alkaloids is well reviewed by SHAMMA and GEOGRIEV [108].

Bisbenzylisoquinoline alkaloids have been prepared from simple 1-benzylisoquinoline precursors by classical methods involving a combination of the Bischler-Napieralski and Ullmann reaction [109–112] or by the oxidative coupling of appropriately substituted phenolic substrates [113]. Presently the most promising avenue to the bisbenzylisoquinolines appears to be via an improved Ullmann diaryl ether synthesis which uses pentafluorophenyl copper in dry pyridine [114].

Physical Properties.

Dauricine (**151**) [*524-17-4*], $C_{38}H_{44}N_2O_6$, M_r 624: *mp* 115 °C, $[\alpha]_D^{11}$ − 139 ° (methanol). Dimethiodide: *mp* 204 °C, $[\alpha]_D^{26}$ − 114 ° (water).

Oxyacanthine (**153**) [*548-40-3*], $C_{37}H_{40}N_2O_6$, M_r 608: *mp* 216 – 217 °C, $[\alpha]_D^{20}$ + 131.5 ° (chloroform). Dihydrochloride: *mp* 270 – 271 °C, $[\alpha]_D$ + 185.5 ° (water).

Berbamine (**154**) [*478-61-5*], $C_{37}H_{40}N_2O_6$, M_r 608: *mp* 197 – 210 °C, $[\alpha]_D^{20}$ + 114.6 ° (chloroform). Hydrochloride: $[\alpha]_D^{20}$ + 63.2 ° (water). Dihydrochloride: *mp* 270 °C (decomp.).

Trilobine (**155**) [*6138-73-4*], $C_{35}H_{34}N_2O_5$, M_r 562: *mp* 237 °C, $[\alpha]_D$ + 307 ° (chloroform). *N*-methyl derivative: *mp* 215 °C, $[\alpha]_D$ + 317 ° (chloroform).

Tubocurarine chloride (**156**) [*57-94-3*], $C_{37}H_{42}N_2O_6Cl_2$, M_r 681. (+)-Form: *mp* 270 °C (decomp.), $[\alpha]_D^{22}$ + 190 ° (water), $[\alpha]_D^{23}$ + 219 ° (methanol). (−)-Form: *mp* 268 °C, $[\alpha]_D^{20}$ − 258 ° (water).

Thalicarpine (**157**) [*5373-42-2*], $C_{41}H_{48}N_2O_8$, M_r 696: *mp* 160 – 161 °C, $[\alpha]_D^{25}$ + 133 ° (methanol), $[\alpha]_D^{25}$ + 89 ° (chloroform).

Pharmacological Properties. KUPCHAN studied the structural requirements for tumor-inhibitory activity by bisbenzylisoquinoline alkaloids and related compounds. Only one linkage of the isoquinoline unit appears to be necessary, and activity is seemingly unaffected by the configuration of the asymmetric centers or whether the nitrogen is secondary or tertiary.

However, the presence of two methylimino groups destroys activity.

Therapeutic Use of Tubocurarine Chloride: The (+)-form is used as a skeletal muscle relaxant [115] and as a diagnostic aid (myasthenia gravis).

9.7. Cularine and Related Alkaloids

The cularine-type alkaloids are characterized by an oxepine ring and a 7,8-dioxygenated isoquinoline moiety. The gross structure of cularine (**159**) was assigned by systematic Hofmann degradation and reductive cleavage of the biphenyl ether linkage [116], and confirmed by X-ray crystallography [117], [118]. The structures of the related alkaloids were established by comparing the mass spectra with that of cularine (**159**). A new cularine alkaloid norsarcocapnine (**160**) was isolated from *Ceratocapnos heterocapra* and the structure was elucidated on the basis of spectroscopic evidence [119]. The structure of the complex alkaloid cancentrine (**161**) was established by X-ray analysis of a degradation product [120].

The cularine-type alkaloids are biosynthesized in plants by phenol oxidation of the 8-oxygenated benzylisoquinoline. Laboratory demonstration of this biosynthesis was attempted [121], [122]. Cularine and related alkaloids were prepared by the Pomeranz-Fritsch and related reactions [123] and by the Ullmann reaction [124]. A synthesis of (+)-cularine (**159**) was achieved via oxylium ions and diastereoselective reductive methylation [125].

9.8. Proaporphine Alkaloids

The representative proaporphine alkaloids are glaziovine (**162**), pronuciferine (**163**), orientalinone (**164**), and litsericine (**165**).

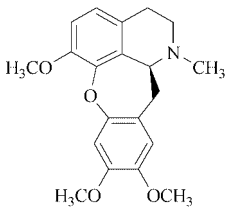

159 Cularine

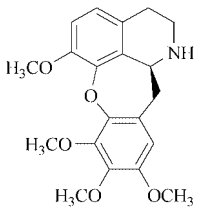

160 (+)-Norsarcocapnine

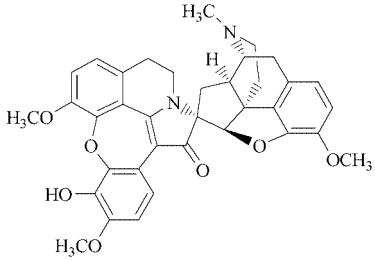

161 Cancentrine

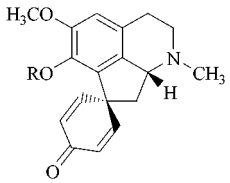

162 R = H (+)-Glaziovine
163 R = CH_3 (+)-Pronuciferine

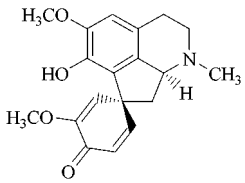

164 (−)-Orientalinone

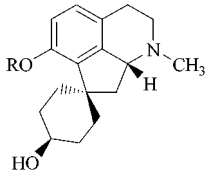

165 Litsericine

The proaporphine alkaloids possess a tetracyclic nucleus incorporating a cross-conjugated cyclohexadienone or related system. Even before they were isolated from nature, their existence was suggested by Barton [126] as biogenetic precursors converted by phenolic oxidation to the aporphines. Indeed, this mechanism has been used to determine the structures of the proaporphines.

The proaporphine alkaloids have been prepared by a variety of classical and biomimetic methods [127–131].

9.9. Aporphine Alkaloids

Structure and Synthesis. Isoboldine (**166**), glaucine (**167**), corytuberine (**168**), isothebaine (**169**), anonaine (**170**), and bulbocapnine (**171**) are typical aporphine alkaloids.

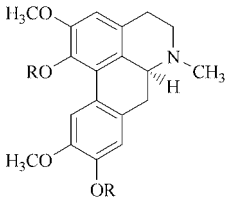

166 R = H Isoboldine
167 R = CH_3 Glaucine

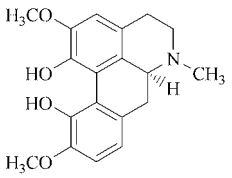

168 Corytuberine

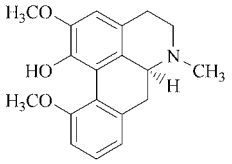

169 Isothebaine

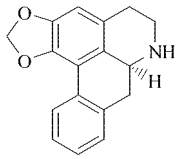

170 Anonaine

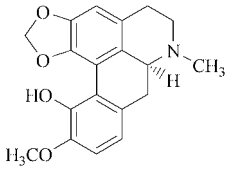

171 Bulbocapnine

The structures of the aporphine alkaloids are characterized by a 9,10-dihydrophenanthrene system. Both classical methods and UV, NMR, and mass spectra have been effectively applied. The absolute configurations have been determined by relating the benzylisoquinoline laudanosine of known chirality to (−)-glaucine (**167**) [132] and by optical rotatory dispersion (ORD) [133], [134]. The marked ORD abnormalities are probably due to steric distortion of the biphenyl ring system [135].

The aporphine alkaloids have been synthesized by the Pschorr reaction [136], either from the accessible 1-(2-aminobenzyl)isoquinolines [137–139] or from the 8-aminoisoquinolines, which are difficult to synthesize [140]. Decomposition of the diazonium intermediates by photolysis instead of the usual thermolysis improved the yields [141]. The synthesis of certain aporphines by photolytic cyclization of 1-benzylisoquinolines has been reported [143], [144]. In addition, aporphines have been synthesized in a biogenetic-type sequence by phenolic oxidation of 1-benzylisoquinoline precursors [145–152]. Developments in the synthesis of aporphine alkaloids are reviewed in [153].

Physical Properties.

Corytuberine (**168**) [*517-56-6*], $C_{19}H_{21}NO_4$, M_r 327: *mp* 240 °C (decomp.), $[\alpha]_D^{20}$ +283° (ethanol). Hydrochloride: *mp* 250 °C, $[\alpha]_D^{20}$ +168° (water).

Isothebaine (**169**) [*568-21-8*], $C_{19}H_{21}NO_3$, M_r 311: *mp* 203–204 °C, $[\alpha]_D^{18}$ +285° (ethanol).

Pharmacological Properties. Glaucine (**167**) brings about slight narcosis in animals, but the narcosis is interrupted by epileptiform convulsions. Glaucine also is a depressant for the heart and blood vessels and can damage striated muscle. Corytuberine (**168**) does not induce narcosis in frogs but does induce increased reflex irritability. In warm-blooded animals it causes tonic convulsions and stimulates the secretion of saliva and tears. It also slows the pulse rate by vagus action, increasing the blood pressure during convulsions.

Bulbocapnine (**171**) phosphate is sometimes used in the treatment of paralysis agitans and St. Vitus' dance.

9.10. Protoberberine Alkaloids

Structure and Synthesis. Coreximine (**172**) and xylopinine (**174**) are 2,3,10,11-tetraoxygenated protoberberines, whereas scoulerine (**173**) and canadine (**175**) are 2,3,9,10-tetraoxygenated. Kikemanine (**176**) and capaurimine (**177**), isolated from *Corydalis pallida var. tenuis* [154], possess a methoxy group at C-9 and a hydroxyl group at C-10. Oritentalidine (**178**), isolated from *Papaver* species [155], [156], has a fused *m*-dioxin system on ring D. Coryenchirine (**179**) carries a methyl group at the C-8 position. Berberine (**180**) and palmatine (**181**) are the dehydro quaternary bases. The structure that was initially proposed for capaurimine [157], was revised to **177** on the basis of spectroscopic [158] and synthetic studies [159] confirmed by X-ray analysis [160].

NMR [161], UV spectroscopy, and mass spectrometry have also been used to elucidate the structure of protoberberines. The chirality was established by correlation with (−)-*N*-norlaudanosine [82], and the conformation was determined by IR spectroscopy (Bohlmann bands in the region 2850–2750 cm^{-1}) [162], ^1H NMR [163], ^{13}C NMR [164], optical rotatory dispersion [165], and relative rates of quaternization [166].

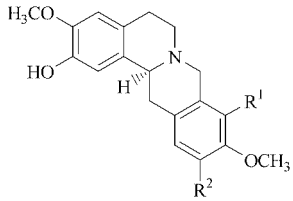

172 R^1 = H, R^2 = OH
Coreximine
173 R^1 = OH, R^2 = H
Scoulerine

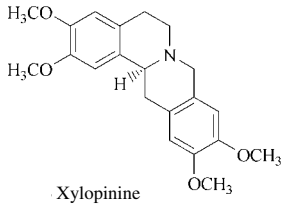

Xylopinine

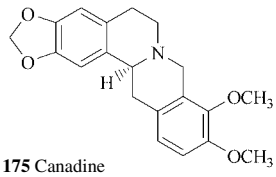

175 Canadine

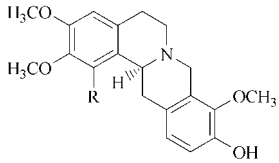

176 R = H, Kikemanine
177 = OH, Capaurimine

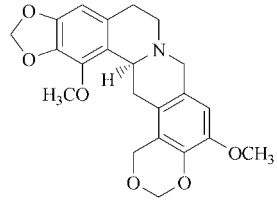

178 Orientalidine

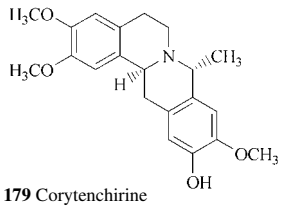

179 Corytenchirine

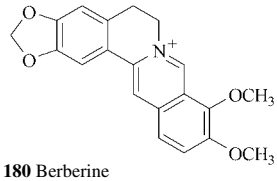

180 Berberine

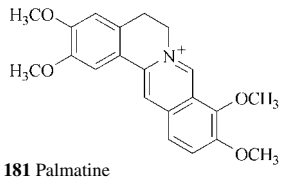

181 Palmatine

Tetrahydroprotoberberines have been synthesized for the most part from benzyltetrahydroisoquinolines and formaldehyde by the Pictet-Spengler reaction. Under the usual conditions, involving an acid catalyst, benzyl moieties containing methoxy groups cyclize mainly to 10,11-substituted products, whereas those with hydroxyl functions yield a mixture of 9,10- and 10,11-disubstituted tetrahydroprotoberberines. To favor formation of the 9,10-isomers, pH has been carefully controlled [167–170] or the alternate site of cyclization blocked by a bromine substituent [171]. For example, the phenolic benzylisoquinoline **139** and formaldehyde cyclized at low pH to the 10,11-disubstituted protoberberine, racemic coreximine (**172**), whereas buffer control at pH 6.3 afforded racemic scoulerine (**173**) as the main product. The latter was also obtained when the bromosubstituted precursor **182** was cyclized under the usual mineral acid conditions and the resulting bromo intermediate **183** reductively dehalogenated.

The isomeric 10,11-disubstituted tetrahydroprotoberberines have also been synthesized by the modified Pomeranz-Fritsch reaction [172] and photocyclization of 2-acetyl-1-benzylideneisoquinolines [173]. KAMETANI and his coworkers [174] converted the 1-benzocyclobutenyl-3,4-dihydroisoquinoline (**185**) by thermolysis into the protoberberine (**186**), which was reduced to give xylopinine (**174**).

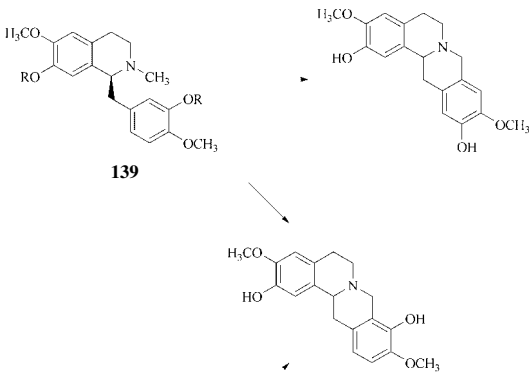

139

173

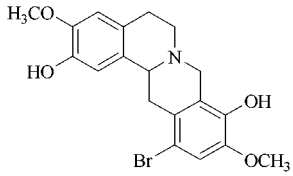

183

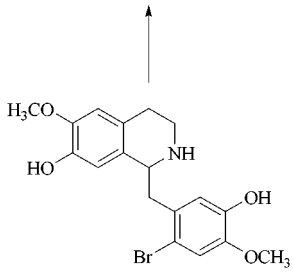

182

Physical Properties.

Canadine (**175**) [*5096-57-1*], $C_{20}H_{21}NO_4$, M_r 339. (±)-Form: *mp* 132 °C, $[\alpha]_D^{15}$ +299° (chloroform), *mp* 172 °C.

Berberine (**180**) [*2086-83-1*] $C_{20}H_{18}NO_4$, M_r 336. Hydroxide: *mp* 145 °C.

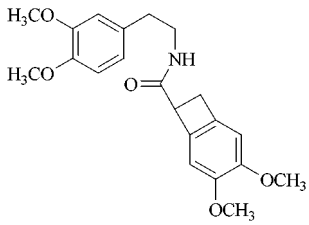

184

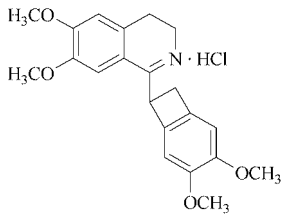

185

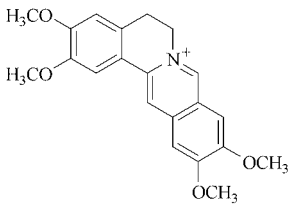

186

Palmatine (**181**) [*3486-67-7*], $C_{21}H_{22}NO_4$, M_r 369. Iodide: *mp* 239 °C (decomp.). Nitrate: *mp* 239 °C (decomp.).

Pharmacological Properties. Berberine (**180**) is moderately toxic to larger animals, causing cardiac damage, dyspnoea, lowered blood pressure, and paresis in rabbits. After oral administration to humans large amounts appear in the urine. It has some trypanocidal action and has been used as an adjunct to quinine in the treatment of malaria.

Therapeutic Use of Canadine: Sedative, muscle relaxant.

Therapeutic Use of Berberine: Bitter stomachic, antibacterial, antimalarial, and antipyretic.

Trade Names for Berberine Chloride: Pheroberine (Kanebo). *Sulfate:* Sutopunine (Kyorin).

9.11. Protopine Alkaloids

Structure and Synthesis. The protopine alkaloids such as protopine (**187**) and cryptopine (**188**) contain a ten-membered ring system which undergoes a transannular reaction between the basic nitrogen and the carbonyl function to form protoberberines. This transformation and classical degradation have been used to elucidate the structure of these alkaloids.

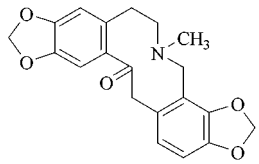

187 Protopine

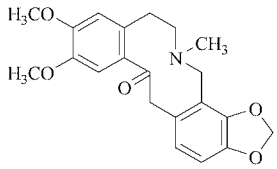

188 Cryptopine

Two general methods have been utilized for the synthesis of protopines, both based on the transformation of protoberberines [175], [176].

Physical Properties.

Protopine (**187**) [*130-86-9*], $C_{20}H_{19}NO_5$, M_r 353: *mp* 208 °C. Methiodide: *mp* 217 °C (decomp.).

Cryptopine (**188**) [*482-74-6*], $C_{21}H_{23}NO_5$, M_r 369: *mp* 220 – 221 °C.

Pharmacological Properties. The alkaloids have been shown to possess an inhibiting action on isolated frog heart, muscle, or nerve and a stimulating action on guinea-pig intestine.

9.12. Phthalide Alkaloids

Structure and Synthesis. Hydrastine (**189**) and narcotine (**190**) are representative phthalide alkaloids. Aobamidine (**191**) is a seco form.

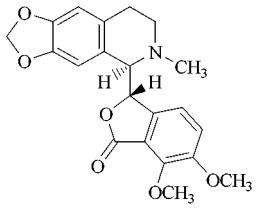

189 Hydrastine

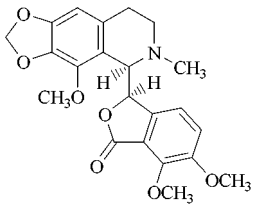

190 Narcotine

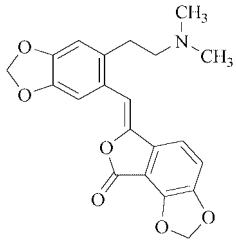

191 Aobamidine

The structures of the phthalides, exemplified by (−)-β-hydrastine (**189**), are characterized by a δ-lactone joined at the C-1 position of the tetrahydroisoquinoline ring. This structure was determined by mild oxidative cleavage and conversion into known protoberberine alkaloids [177]. The absolute configurations have been assigned by chemical and physical methods [178] and confirmed by X-ray analysis [179].

The phthalideisoquinolines **190** have been classically synthesized as a mixture of two diastereoisomers [180], [181].

Physical Properties.

Hydrastine (**189**) [*118-08-1*], $C_{21}H_{21}NO_6$, M_r 383: *mp* 132 °C, $[\alpha]_D^{20}$ − 50 ° (ethanol), $[\alpha]_D$ − 85 ° (acetone). Hydrochloride: *mp* 116 °C, $[\alpha]_D^{17}$ + 127 ° (dilute hydrochloric acid).

Narcotine (Noscapine) (**190**) [*128-62-1*], $C_{22}H_{23}NO_7$, M_r 413: *mp* 176 °C. Camphorsulfonate: *mp* 188 – 191 °C, $[\alpha]_D^{33}$ + 32.7 ° (water).

Pharmacological Properties. Hydrastine (**189**) causes strychnine-like convulsions in the frog. When large doses are administered, sometimes the convulsions are followed by paralysis. It has been employed on occasion as an internal styptic in uterine haemorrhage. Narcotine (**190**) resembles thebaine in its action, being a reflex stimulant rather than a narcotic.

Therapeutic Use of Hydrastine Hydrochloride: uterine hemostatic, antiseptic.

Therapeutic Use of Narcotine: Antitussive.

Trade Names for Narcotine: Tuscapine (Fisons), Narcotine (Taisho, Hishiyama, Fusso), Noscapine (Nakakita). *Hydrochloride:* Capval (Dreluso), Codipect (Kwizda), Lyobex (Lappe), Narotussin (Bidogici Italia, Ebewe), Nectaclon (Merck), Narcotine M (Isei). *Resinate:* Capval (Dreluso).

9.13. Ochotensine and Related Alkaloids

Ochotensine (**192**), fumaricine (**193**), raddeanine (**194**), and raddeanone (**195**) contain the 1,1-spirobenzylisoquinoline structure. The structure of ochotensine was elucidated by a combination of Hofmann and Emde degradations and NMR, and confirmed by X-ray analysis of ochotensine methiodide [182].

Ochotensine (**192**) has been synthesized by the Pictet-Spengler reaction [183], [184]. SHAMMA and his co-workers converted the protoberberine-type compound into the ochotensine derivative [185], while KAMETANI and his associates synthesized it by thermolysis of benzocyclobutene [186].

9.14. Rheadan Alkaloids

Structure, Occurrence, and Synthesis. Rhoeadine (**196**), alpinigenine (**197**), and oreogenine (**198**), found in *Papaver* species, are representative rheadan alkaloids. Stepinonine (**199**), isolated from *Stephania japonica* [187], is a dimeric alkaloid having a seven-membered nitrogen-containing ring.

The structure of the rheadan alkaloids was elucidated by a combination of Hofmann degradation and spectroscopic methods and confirmed by X-ray analysis [188].

The first synthesis of rhoeadine (**196**) was accomplished via a spirobenzylisoquinoline derivative by IRIE [189]. Several other synthetic methods [190–192] leading to rheadan alkaloids have been developed.

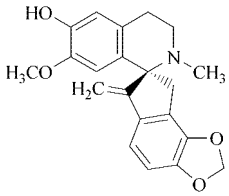

192 Ochotensine

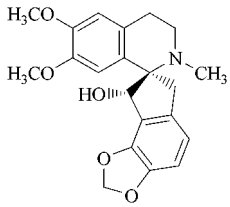

193 Fumaricine

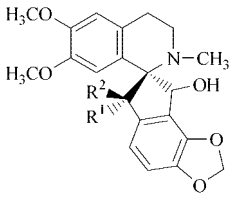

194 R^1 = OH, R^2 = H Raddeanine
195 $R^1 + R^2$ = O Raddeanone

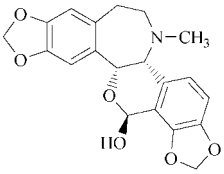

196 Rhoeadine

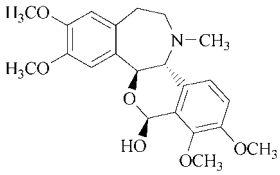

197 Alpinigenine

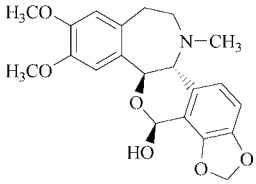

198 Oreogenine

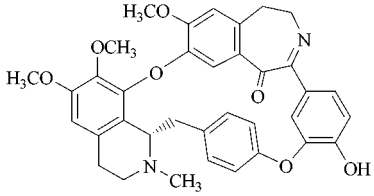

199 Stepinonine

Physical Properties.
Rhoeadine (**196**) [*2718-25-4*], $C_{21}H_{21}NO_6$, M_r 383: *mp* 252 – 254 °C, $[\alpha]_D^{23}$ +235 ° (chloroform).

9.15. Morphine and Related Alkaloids

Structure and Synthesis. Representative morphine alkaloids are morphine (**200**), codeine (**201**), thebaine (**202**), salutaridine (**203**), pallidine (**204**), flavinantine (**205**), and sinomenine (**206**). The structures of these alkaloids were elucidated by a combination of Hofmann degradation and IR, UV, NMR, and mass spectroscopies. Absolute configuration was determined by optical rotatory dispersion [193], [194].

In contrast to the protoberberine and related alkaloids, which originate from (+)-reticuline, morphine is derived from the antipodal benzylisoquinoline (−)-reticuline. Reticuline undergoes an alternate coupling mode involving bond formation between the benzyl moiety and the C-4 a position to give the morphinandienone alkaloids flavinantine (**205**) [195] and salutaridine (**203**). The latter is the biogenetic precursor of morphine [196].

Total synthesis of morphine (**200**) was achieved first by GATES [197] and then by GINSBURG [198]. It has also been prepared by BARTON and his associates [199] with biogenetic-type sequences. KAMETANI and his associates reported the synthesis of the key intermediate salutaridine (**203**) by the Pschorr reaction of the aminoisoquinoline **207** [200] as well as by photolysis of 2′-bromoreticuline (**208**, racemic mixture) [201].

Physical Properties.
Morphine (**200**) [*57-27-2*], $C_{17}H_{19}NO_3$, M_r 285: *mp* 254 °C (decomp.). Monohydrate: *mp* 254 – 256 °C (decomp.), $[\alpha]_D^{25}$ − 132 ° (methanol). Hydrobromide [*630-81-9*]: $[\alpha]_D^{15}$ − 100.4 ° (water). Hydrochloride [*52-26-6*]:

mp 200 °C (decomp.), $[\alpha]_D^{25}$ −113.5° (water). Methyl bromide [*125-23-5*]: *mp* 265–266 °C.

(**201**) [*76-57-3*], $C_{18}H_{21}NO_3$, M_r 299: *mp* 154–156 °C, $[\alpha]_D^{15}$ −136° (ethanol). Hydrobromide [*125-25-7*]: *mp* 190–192 °C, $[\alpha]_D^{22}$ −96.6° (water). Hydrochloride [*1422-07-7*]: *mp* 280 °C, $[\alpha]_D^{22}$ −108° (water). Methyl bromide [*125-27-9*]: *mp* 260 °C. Phosphate [*52-28-8*]. Sulfate [*1420-53-7*].

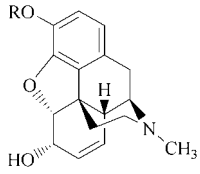

200 R = H Morphine
201 R = CH_3 Codeine

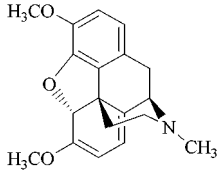

202 Thebaine

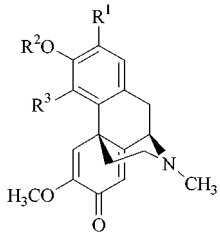

203 R^1 = H, R^2 = CH_3, R^3 = OH Salutardine
204 R^1 = OH, R^2 = CH_3, R^3 = H Pallidine
205 R^1 = OCH_3, R^2 = H, R^3 = H Flavinantine

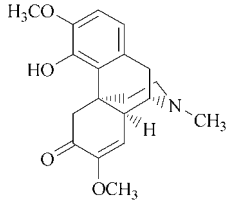

206 Sinomenine

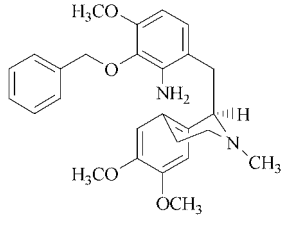

207

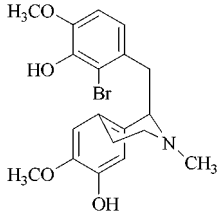

208

Codeine

Thebaine (**202**) [*115-37-7*], $C_{19}H_{21}NO_3$, M_r 311: *mp* 193 °C, $[\alpha]_D^{15}$ −219° (ethanol), $[\alpha]_D^{23}$ −230° (chloroform).

Salutaridine (**203**) [*1936-18-1*], $C_{19}H_{21}NO_4$, M_r 327: *mp* 197–198 °C, $[\alpha]_D^{12}$ +114° (methanol).

Sinomenine (**206**) [*115-53-7*], $C_{19}H_{23}NO_4$, M_r 329: *mp* 161 °C, $[\alpha]_D^{26}$ −71° (methanol). Hydrochloride: *mp* 231 °C (decomp.), $[\alpha]_D^{17}$ −82° (water).

Preparation. One method for the production of morphine and codeine starts from finely powdered poppy plants [202]. The plants are extracted with quicklime in a suitable solvent. After separating the extract and washing the residue, the filtered solution is acidified, freed from the solvent in vacuo, and the residue extracted at a pH of ca. 9 with amyl alcohol or another suitable organic solvent. Morphine and codeine are obtained from this extract with dilute acid.

New variations are described in [203].

Pharmacological Properties. *Morphine* (**200**) exerts both marked depressing and marked stimulating effects on the central nervous system. The depression affects the brain, particularly the sensation of pain and respiration. The cerebral motor functions are less affected. Pharmacologically, *codeine* (**201**) resembles morphine but is less toxic, and its depressant action is less marked and prolonged. Its stimulating action involves not only the spinal cord but also the lower parts of the brain.

Sinomenine (**206**) produces increased reflex excitability and respiratory paralysis in mice, frogs, and rabbits.

Therapeutic Use: Morphine is used as a narcotic analgesic; codeine is used as a narcotic analgesic and antitussive.

Trade Names for Morphine: Duromorph (L. A. B.), Nepenthe (Evans). *Sulfate:* Morphine Sulfate (Elkins-Sinn, Philips Roxane, Wyeth), MST Continus (Napp). *Hydrochloride:* Morphin-Amphiden "NBK" (Merck, Knoll, Boehringer), Morphin Thilo (Thilo), Ensan Morphine (Sankyo, Shionogi, Dainippon, Takeda, Tanabe). *Tartrate:* Cyclimorph (Calmic).

Trade Names for Codeine: Codeopect (Nyegaard), Codicept (Sanol). *Phosphate:* Codeinfos (Union Quimico-Farmaceutica), Tricodein (Solac, Zyma), Codyl (Boehringer Ingelh.), Codeinum Phosphoricum (Merck, Knoll, Cas-

can, Takeda, Dainippon, Isei, Tanabe, Sankyo, Shionogi). *Resinate:* Codipertussin (Taeschner).

9.16. Hasubanan Alkaloids

The structures of hasubanonine (**209**) and cepharamine (**210**) have been derived by the use of physical methods on the degradation products [204]. The structure of the related, novel halogen-containing alkaloid acutuminine (**211**) was proposed on the basis of spectra and confirmed by X-ray analysis [205].

(±)-Cepharamine (**210**) and (±)-hasubanonine (**209**) have been synthesized via Robinson annelation [206], [207]. Photolytic and phenolic oxidative methods have been used to synthesize hasubanan alkaloids [208], [209].

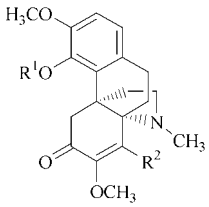

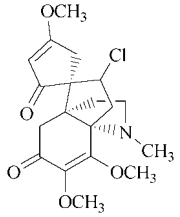

209 $R^1 = CH_3$, $R^2 = OCH_3$ Hasubanonine
210 $R^1 = H$, $R^2 = H$ Cepharamine

211 Acutuminine

Hasubanonine (**209**) [*1805-85-2*], $C_{21}H_{27}NO_5$, M_r 373: *mp* 116 °C.

9.17. Emetine and Related Alkaloids

Structure and Synthesis. Protoemetine (**212**), emetine (**213**), and tubulosine (**214**) contain the benzoquinolizidine skeleton. The elucidation of the structure of emetine (**213**) involved a variety of degradative and physicochemical methods as well as biogenetic considerations. The gross structure was proposed by Robinson [210] on the speculation that emetine is biogenetically derived from a protoberberine-type precursor via Woodward fission. The relative and the absolute configurations were determined mainly by Battersby [211] and van Tamelen [212]. Ipecoside (**215**) is a biogenetic precursor of the ipecacuanha alkaloids [213].

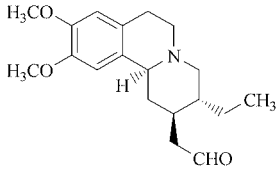

212 Protoemetine

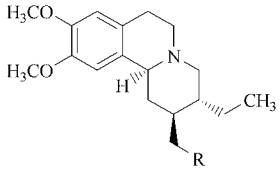

R =

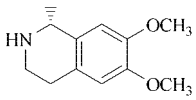

213 Emetine

R =

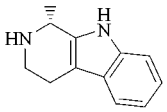

214 Tubulosine

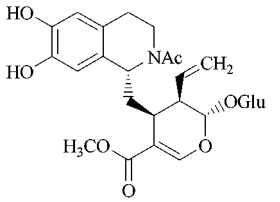

Ac = acetyl
Glu = glucose
215 Ipecoside

Emetine has been synthesized by many groups [214–220]. Openshaw [215] obtained the (−)-benzoquinolizidine **218** by the reaction of the 3,4-dihydroisoquinoline **216** and the Mannich base **217** followed by treatment with (−)-camphor-10-sulfonic acid. The ketone **218** was transformed into emetine (**213**) via the unsaturated ester **219**. Furthermore, properly substituted benzo[a]quinolizidine **222** was synthesized in a one-pot reaction from the 1-methyl-3,4-dihydroisoquinoline **220** and dimethyl 3-methoxyallylidenemalonate (**221**) in methanol. (±)-Emetine (**213**) and (±)-tubulosine (**214**) were stereoselectively synthesized from the acetal **222** [218–220].

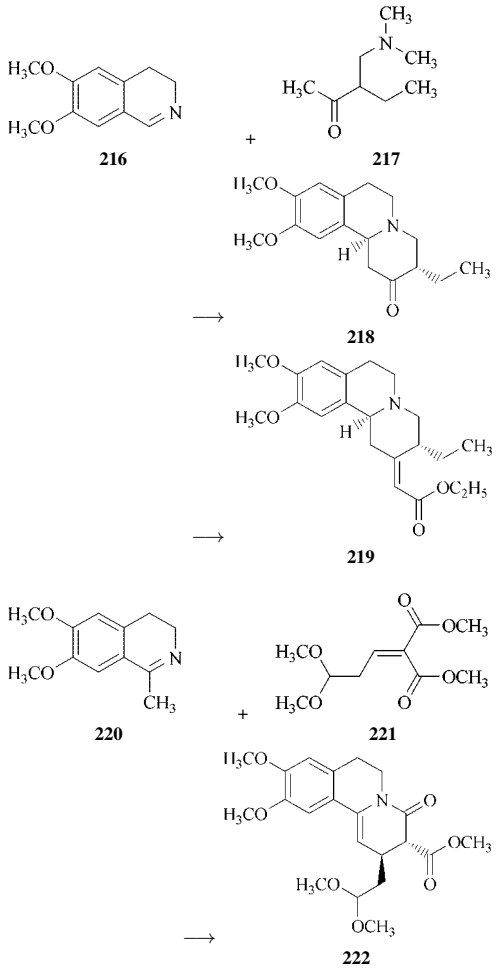

Physical Properties. *Emetine* (**213**) [*483-18-1*], $C_{29}H_{40}N_2O_4$, M_r 480: *mp* 74 °C, $[\alpha]_D^{20}$ −50° (chloroform). Hydrochloride [*316-42-7*]: *mp* 235–255 °C, $[\alpha]_D$ +11° (water).

Preparation. Emetine is produced on an industrial scale by both synthesis and isolation from roots of Ipecacuanha. The powdered drug is alkalized with ammonia and then extracted with ether. The base is neutralized with aqueous sulfuric acid, and then the solution is alkalized with sodium hydroxide and potassium hydroxide and extracted again with ether, whereby the emetine goes into the ether phase. After evaporation, the residue of the ether phase is dissolved in methanol and mixed with ethanolic hydrobromic acid to crystallize emetine hydrobromide. From 120 to 200 kg of *Radix ipecacuanha* are required to produce one kilogram of raw emetine hydrobromide.

Pharmacological Properties and Therapeutic Use. Emetine (**213**) is employed as an antiamebic and is sometimes employed as an emetic. In small doses it is used as an expectorant although more commonly ipecacuanha itself or one of its galenical preparations is employed.

Trade Names for the Hydrochloride: Emetine (City Chem. Corp., Inland Alkaloid Co., Lilly, Upjohn), Emetinum Hydrochloricum (Riedel).

9.18. Erythrina Alkaloids

Structure and Synthesis. The erythrina alkaloids are characterized by a tetracyclic spiro amine system containing either two conjugated double bonds or a single unsaturated bond. One of the rings is either aromatic or a δ-lactone. The three major types are exemplified by erysotrine (**223**), erythratine (**224**), and α-erythroidine (**225**). The elucidation of the structures of these alkaloids, summarized in an excellent review by Mondon [221], involved the efforts of many investigators and utilized almost all available physicochemical methods.

The total synthesis of the erythrina alkaloid erysotrine (**223**) was achieved by Mondon and Nestler [222]. A total synthesis of (+)-erysotrine (**223**) is reported in [223].

On the basis of biogenetically patterned sequences [224] racemic dihydroerysodine (**226**) and erythratidine (**227**) were synthesized.

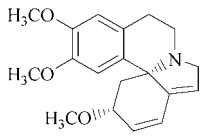

223 Erysotrine

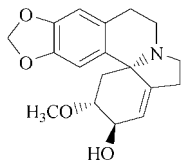

224 Erythratine

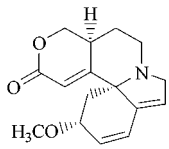

225 α-Erythroidine

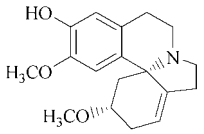

226

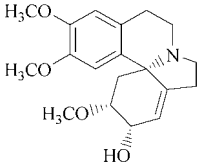

227

Physical Properties.

α-Erythroidine (**225**) [*466-80-8*], $C_{16}H_{19}NO_3$, M_r 273: *mp* 58 – 60 °C, $[\alpha]_D^{27}$ + 136 ° (water). Hydrochloride: *mp* 226 – 228 °C, $[\alpha]_D^{32}$ + 118 ° (water).

9.19. Cephalotaxine Alkaloids

Cephalotaxine (**228**), isolated form *Cephalotaxus harringtonia var. drupacae,* possesses a characteristic 1-azaspiro[4.4]nonene structure. Harringtonine (**229**) is an ester derivative. Cephalotaxine (**228**) was synthesized by WEINREB [225], SEMMELHACK [226], [227], and IKEDA [228].

Pharmacological Property. Harringtonine (**229**) shows a potential antitumor activity [229].

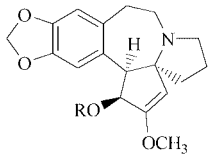

228 R = H Cephalotaxine

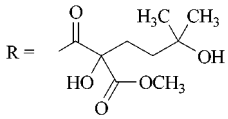

229
Harringtonine

9.20. Protostephanine Alkaloids

Protostephanine (**230**), isolated from *Stephania japonica,* contains a dibenz[*d,f*]azocine system. The nine-membered heterocyclic protostephanine has been prepared by classical methods [230] and a biosynthetic pathway [231].

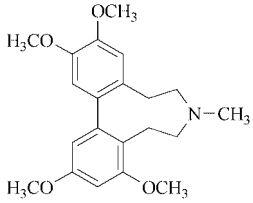

230 Protostephanine

9.21. Benzophenanthridine Alkaloids

Structure, Origin, and Synthesis. Representative benzophenanthridine alkaloids are nitidine (**231**), sanguinarine (**232**), chelidonine (**233**), and corynoline (**234**). All were isolated from *Papaveraceae* plants.

Benzophenanthridine alkaloids have been synthesized by the general procedure developedby BAILEY and co-workers [232] as well as by methods which used the Pschorr reaction [233], thermolysis [234], and photolysis [235].

The structure of corynoline (**234**), isolated from *Corydalis incisa,* was confirmed by an X-ray analysis of the *p*-bromobenzoate [236]. Corynoline was synthesized by two different methods: the photocyclization of enamides [237] and the condensation of a piperonylidenemethylamine with a substituted homophthalic anhydride [238].

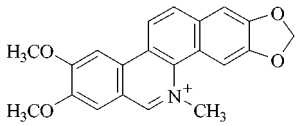

231 Nitidine

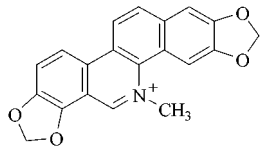

232 Sanguinarine

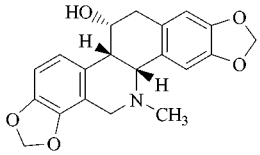

233 Chelidonine

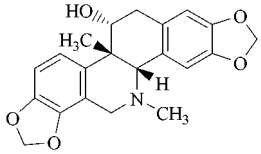

234 Corynoline

Physical Properties.
Chelidonine (**233**) [*476-32-4*], $C_{20}H_{19}NO_5$, M_r 353: *mp* 135 – 136 °C, $[\alpha]_D^{22}$ +115° (ethanol), $[\alpha]_D^{20}$ +117° (chloroform).
Sanguinarine (**232**) [*2447-54-3*], $C_{20}H_{14}NO_4$, M_r 332: *mp* 278 – 280 °C (from water), *mp* 210 – 211 °C (from chloroform + ethanol).

Pharmacological Property. Sanguinarine (**232**) produces glaucoma in animals.

9.22. Dibenzopyrrocoline Alkaloids

The dibenzopyrrocoline alkaloids, isolated from Australian *Lauraceae* plants, have a five-membered ring system containing a shared quaternary nitrogen. This is exemplified by cryptowoline (**235**). This representative alkaloid was synthesized by oxidative coupling of laudanosine before it was isolated from nature [239].

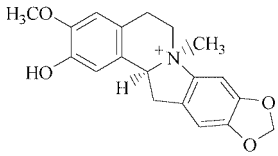

235 Cryptowoline

9.23. Phenethylisoquinoline Alkaloids

Structure and Synthesis. Various types of phenethylisoquinoline alkaloids have been found in nature. For example, autamnaline (**236**), melanthioidine (**237**), kreysiginone (**238**), multifloramine (**239**), androcymbine (**240**), and schelhammerine (**241**) are the homologues of the corresponding 1-benzylisoquinoline, bisbenzylisoquinoline, proaporphine, aporphine, morphine, and erythrina alkaloids, respectively. Colchicine (**242**) is also included in this class for biogenetic reason.

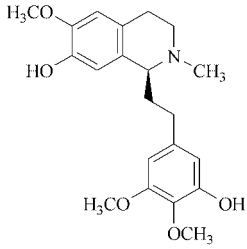

236 Autumnaline

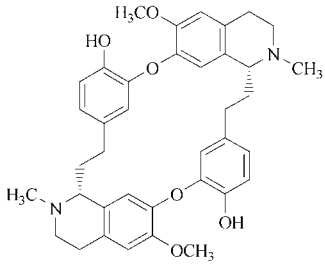

237 Melanthiodine

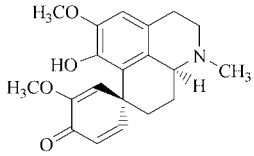

238 Kreysiginone

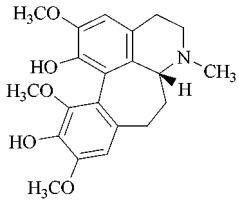

239 (−)-Multifloramine

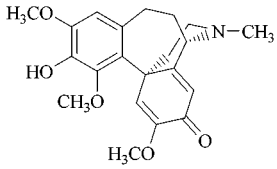

240 Androcymbine

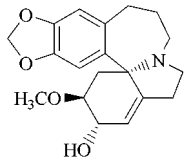

241 Schelhammerine

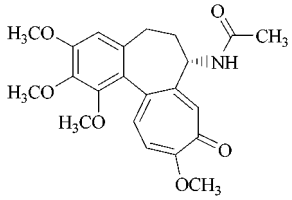

242 Colchicine

Although the biosynthesis of all the phenethylisoquinoline alkaloids has not yet been fully elucidated, the probable pathways have been indicated by tracer experiments for homomorphinandienones and homoaporphines. Strong evidence was provided for the route by which the homomorphinandienone alkaloid androcymbine (**240**), itself derived from the phenethylisoquinoline **236**, is transformed into colchicine (**242**) [240].

Bisphenethylisoquinoline melanthioidine (**237**) has been synthesized by a double Ullmann reaction [241], [242]. In addition, a biogenetic route to **237** via phenol oxidation has been studied [243].

Phenol oxidation has also been used in the synthesis of the homoaporphines [244–246]. Androcymbine (**240**) was obtained by photolytic dehydrobromination [247] and the photo-Pschorr reaction [248]. Colchicine (**242**) has been synthesized by a number of methods [249–253].

Physical Properties.

Colchicine (**242**) [*64-86-8*], $C_{22}H_{25}NO_6$, M_r 399: *mp* 142 – 150 °C, *mp* 157 °C, $[\alpha]_D^{17}$ – 121 ° (chloroform).

Preparation. The seeds of the meadow saffron (*Colchicum autumnale* L.) are extracted with boiling alcohol to obtain colchicine. Because of its solubility, it can be separated from accompanying substances: the extract is evaporated and the residue mixed with water. The undissolved resins and fats are filtered off and the filtrate extracted with chloroform. After evaporation of the chloroform solution the procedure is repeated with the residue. When the purity is sufficient, the colchicine crystallizes from the concentrated chloroform solutions as a chloroform complex [254].

Pharmacological Properties. Colchicine (**242**) has no direct action upon the heart but is a capillary poison, among other things. Large doses cause paralysis ascending the central nervous system and accompanied by vasomotor and respiratory paralysis.

Therapeutic Use: Gout suppressant, treatment of familial Mediterranean fever.

Trade Names: Colchicine (Lilly, Danbury Pharm., Purepac Pharm., Shionogi, Tokyo Tanabe).

9.24. Miscellaneous Isoquinoline Alkaloids

Structure and Synthesis. The structure of mimosamycin (**243**), isolated from *Streptomyces lavendulae,* was determined by X-ray analysis [255]. Mimosamycin (**243**) has been synthesized via Pomeranz-Fritsch reaction [256]. On the other hand, the structure of saframycin B (**244**) also isolated from *S. lavendulae,* was elucidated by comparison of spectra of saframycin B and its congener saframycin C. The structure of the congener had been determined previously by X-ray crystallography [257]. The first synthesis of (±)-saframycin B was carried out by FUKUYAMA and SACHLEBEN [258].

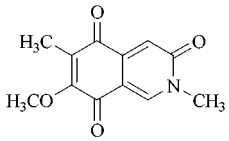

243 Mimosamycin

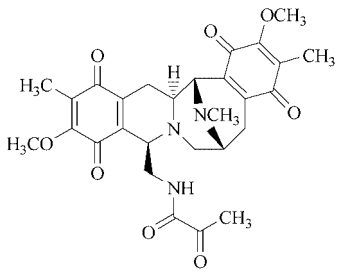

244 Saframycin B

Pharmacological Properties. Mimosamycin (**243**) and saframycins show antibiotic activity against gram-positive bacteria. Saframycins also exhibit antitumor activities.

10. Amaryllidaceae Alkaloids

A large number of alkaloids isolated from *Amaryllidaceae* species are classified by their biogenetic origin from norbelladine (**245**) into five groups (Sections 10.1, 10.2, 10.3, 10.4, 10.5) [259], [260]. The tracer experiment by BARTON and COHEN showed that **245** was derived from tyrosine and phenylalanine, and that lycorine (**250**) and haemanthamine (**259**) originate from **246** via o,p′- and p,p′-oxidative coupling, respectively. Galanthamine (**265**) is formed from **247** by p,o′ coupling [126], [261–264].

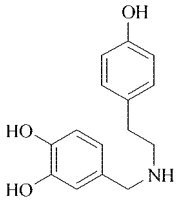

245 Norbelladine

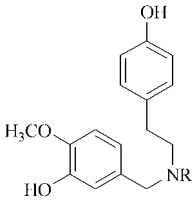

246 R = H
247 R = CH_3

10.1. Lycorine Type

Structure. The structure, including the absolute stereochemistry, of lycorine has been worked out on the basis of spectra, chemical degradations, and finally X-ray analysis, to be **250** [265]. The synthesis of the lycorine skeleton **249** of correct stereochemistry (B and D are *trans*) was achieved elegantly by STORK in 1979 [266], employing an intramolecular Diels-Alder reaction of the amide **248**. Many papers concerned with the synthesis of this group of alkaloids have appeared [267–269], [270–272].

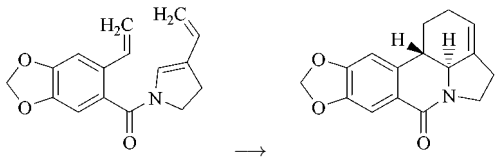

248 → **249**

Pulviine (**251**) possesses a structure similar to that of lycorine, and it could be converted into lycorenine (**252**) in several steps. Reduction with lithium aluminum hydride, tosylation of the benzylic OH group, cyclization, and demethylation transforms **252** into **251**, thus establishing the stereochemistry of both alkaloids. Since pretazettine (**253**) also has a *trans* B – D ring juncture, it was classified in this group. Pretazettine (**253**) is converted to tazettine on heating in water at 70 °C.

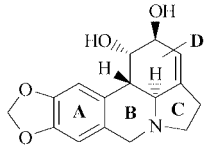

250 (−)-Lycorine

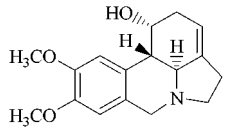

251 (−)-Pulviine

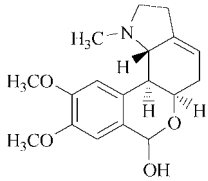

252 (+)-Lycorenine

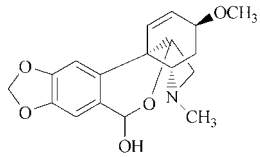

253 (+)-Pretazettine

Physical Properties.
Lycorine (**250**) $C_{16}H_{17}NO_4$, M_r 287: *mp* 275 – 280 °C decomp. (from ethanol), $[\alpha]_D^{16}$ − 129 ° (98 % ethanol).

Pharmacological Properties. The alkaloids in this group are relatively nontoxic to mammals. Lycorine given orally or subcutaneously in large doses to the dog or cat causes vomiting and diarrhea. Blood pressure does not seem to be affected.

10.2. Crinane Type

The structure of crinine (**254**) was elucidated by Hofmann degradation of oxocrinine (**255**) to give the dienone **256**. Reduction then provided the tetrahydro derivative **257** [273]. The stereochemistry of maritidine (**258**), a tertiary base containing two methoxy groups and an allyl alcohol function, was established [274] on the basis of optical rotatory dispersion and circular dichroism curves. The absolute configuration of crinane alkaloids was finally solved by X-ray analysis of a haemanthamine (**259**) derivative [275].

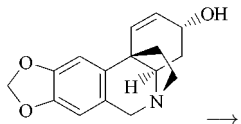

254 (−)-Crinine →

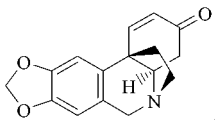

255 →

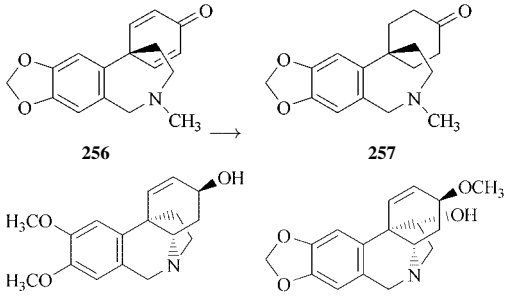

256 **257**

258 (+)-Maritidine **259** (+)-Haemanthamine

10.3. Nariclasine Type

Structure and Synthesis. Lycoricidine (**260**) and lycoricidinol (**261**) possess potent biological activity and were isolated from the bulbs of *Lycoris radiata* Hebb. Their structures, excluding stereochemistry, were elucidated by spectroscopic and chemical evidence [275]. Lycoricidine (**260**) may be derived biogenetically from **254** by phenolic coupling and subsequent degradation of **262**. Narciclasine (**261**), which is identical with lycoricidinol, was also isolated from *Narcissus incomparabilis* and all *Narcissus* sp. The structure was determined by spectral data [276] and X-ray analysis [277]. Later, pancratistatin (**264**), which is more highly functionalized with hydroxyl groups, was found in the roots and bulbs of *Pancratium littorale* [278].

The first total synthesis of (±)-pancratistatin (**264**) was achieved by DANISHEFSKY [279]. The syntheses of optically active lycoricidine (**260**) [280], narciclasine (**261**) [281], pancratistatin (**264**), [282], [283] and 7-deoxypancratistatin (**263**) [284], [285] have been reported.

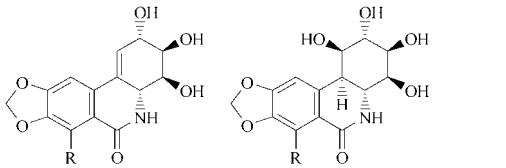

260 R = H Lycoricidine
261 R = OH Narciclasine (Lycoricidinol)
263 R = H 7-Deoxypancratistatin
264 R = OH Pancratistatin

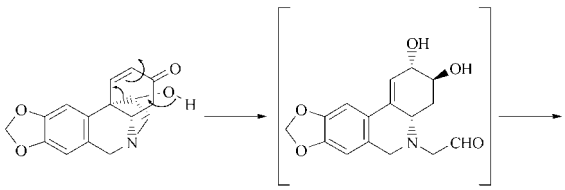

262 **260**

Physical Properties.
Lycoricidine (**260**) [*19622-83-4*], $C_{14}H_{13}NO_6$, M_r 291: *mp* 214.5 – 215.5 °C.
Narciclasine (**261**) [*29477-83-6*], $C_{14}H_{13}NO_7$, M_r 307: *mp* 232 – 234 °C (decomp.), $[\alpha]_D$ + 145 ° (c = 1.5 in ethanol).
Pancratistatin (**264**) [*96203-70-2*], $C_{14}H_{15}NO_8$, M_r 325: *mp* 322 – 324 °C (from DMF – methanol – diethyl ether), $[\alpha]_D$ + 48 ° (c = 1.0 in dimethyl sufoxide).

Pharmacological Properties. Lycorocidine (**260**) acts as a plantgrowth regulator and inhibits division of Ehrlich ascites cells. Narciclasine (**261**) is a strong tumor inhibitor [LD_{50} 5 mg/kg (s.c. mice)]. In addition, pancratistatin (**264**) has remarkable inhihitory activity against P-388 leukaemia cells (ED_{50} 0.01 µg/mL) and M 5076 sarcoma (53 – 84 % life extension at 0.38 – 3.0 mg/kg).

10.4. Galanthamine Type

Structure and Synthesis. The structure of galanthamine (**265**) was elucidated by the reaction with concentrated hydrobromic acid to yield the catechol. The dimethyl ether of this catechol was then subjected to Emde degradation, and the oxidation of the resulting phenethylamine (**267**) provided galanthamic acid (**268**) and the monocarboxylic acid **269** [286]. Moreover, manganese dioxide oxidation of galanthamine afforded narwedine (**266**) [287]. The structure of **265** was confirmed by X-ray analysis, including its absolute configuration [288]. The first synthesis of galanthamine via a biogenetically patterned pathway by phenolic coupling of belladine (**247**) was reported by BARTON [289]. Since then, many attempts to synthesizes galanthamine have been reported, because of its pharmacological activity [290–292].

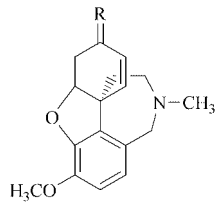

265 R = H β-OH (−)-Galanthamine
266 R = O Narwedine

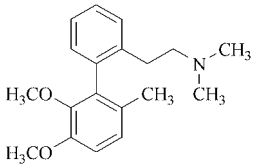

267

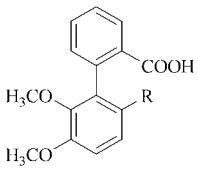

268 R = CO_2H **269** R = CH_3

Physical Properties.

Galanthamine (**265**) [*357-70-0*], $C_{17}H_{21}NO_3$, M_r 287: *mp* 126 – 127 °C (from benzene), $[\alpha]_D^{20}$ – 188.8 ° (*c*= 1.378 g/ 100 ml in ethanol).

Narwedine (**266**) [*510-77-0*], $C_{17}H_{19}NO_3$, M_r 285, *mp* 189 – 192 °C, $[\alpha]_D$ +405 ° (*c* = 1.0 in chloroform).

Pharmacological Properties. Galanthamine inhibits acetylcholinesterase activity and is sold in the USSR. LD_{50} in mice: 8.0 mg/kg i.v., 18.7 mg/kg orally.

Therapeutic Use: Cholinesterase inhibitor.

Trade Names: Nivalin (Chimimport, Bulgaria), Nivalin (Werfft Chemie, Austria), Galanthamin (Diamant).

10.5. Montanine Type

Structure and Synthesis. The montanine-type alkaloids montanine (**271**), pancracine (**272**), and coccinine (**273**) have a unique structure with a 5,11-methanomorphanthridine skeleton. The new montanine-type alkaloid montabuphine (**274**) with a 5,11-methanomorphanthridine skeleton was isolated from the bulbs of *Boophane flava*, and its structure was determined on the basis of spectroscopic considerations [293]. The total syntheses of **271** – **273** were reported by Hoshino [294], Overman [295], and Weinreb [296].

Physical Properties.

Montanine (**271**) [*642-52-4*], $C_{17}H_{19}NO_4$, M_r 301: oil, monohydrate, *mp* 88 – 89 °C (from water), $[\alpha]_D^{26}$ – 97.6 ° (*c* = 0.57 in chloroform); acetone solvate, *mp* 57 – 60 °C (acetone); chloroform solvate, *mp* 59 – 65 °C (chloroform). Weak hypotensive and convulsive agent (dog); LD_{50} 42 mg/kg (i.v. dog).

Pancracine (**272**) [*21416-14-8*], $C_{16}H_{17}NO_4$, M_r 287: *mp* 272 – 273 °C (mehtanol), $[\alpha]_D^{25}$ – 74 ° (*c* = 0.02 in methanol).

Coccinine (**273**) [*132242-50-3*], $C_{17}H_{19}NO_4$, M_r 301: *mp* 162 – 163 °C, $[\alpha]_D^{27}$ – 189 ° (*c* = 1.9 in ethanol). Convulsive action at high doses; LD_{50} 17.5 mg/kg (i.v. dog).

Montabuphine (**274**) [*168611-77-6*], $C_{17}H_{19}NO_4$, M_r 301: *mp* 162 – 164 °C, $[\alpha]_D^{22}$ + 157 ° (*c* = 0.106 in ethanol).

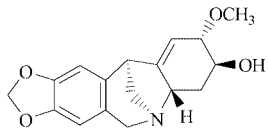

271 Montanine

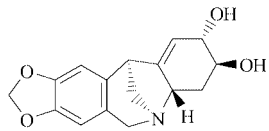

272 Pancracine

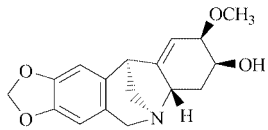

273 Coccinine

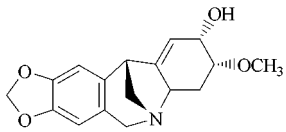

274 Montabuphine

10.6. Miscellaneous

The structures of mesembrine (**275**) and joubertiamine (**276**) were evident from their spectral properties. The absolute configuration of **275** was confirmed by X-ray analysis [297]. Biosynthetic studies for the alkaloids of this family were carried out by tracer work. There are a number

of reports concerning the synthesis of this family of alkaloids [298], [299]. The first synthesis of **275** was reported by SHAMMA in 1965.

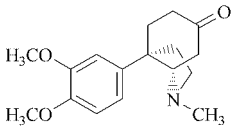

275 Mesembrine

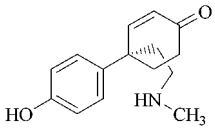

276 Joubertiamine

Physical Properties.

Mesembrine (**275**) $C_{17}H_{23}NO_3$, M_r 289: *bp* 186–190 °C (at 0.04 kPa), $[\alpha]_D^{20}$ −55.4° (methanol).

Mesembrine shows local anaesthetic and mydriatic activities. Asymmetric syntheses of **275** have been performed [300–302].

Cherylline, isolated from several *Crinum* species [303], was assigned structure **277**, which possesses a *p*-hydroxyphenyl group at the C-4 position on the basis of spectroscopic and chemical considerations. Biogenetically, cherylline might arise from **246** via the quinone methide **278** by cyclization. An alternate pathway involves mantanine-type intermediate **279** arising from **246** by phenolic oxidative coupling and rearrangement of the 5,10*b*-ethanophenanthridine skeleton. Acid-catalyzed cyclization of **280** led to cherylline via the quinone methide **278** [304].

11. Indole Alkaloids

Indole alkaloids constitute the largest group of alkaloids. They are of interest on account of their structures – often extremely complex – and their surprising physiological activities. Most contain a tryptamine unit as a readily distinguishable feature or a modified structure, and tryptamine is a precursor in the biosynthesis of many of these alkaloids.

11.1. Non-Tryptamine Type

Mukonine (**281**) was isolated from the stem bark of *Murraya koenigii* Spreng. It is the methyl ester of mukonic acid (1-methoxycarbazole-3-carboxylic acid) [305]. Mupamine (**282**), isolated from *Clausena anisata* (Willd.) Oliv., was characterized as a pyranocarbazole alkaloid on the basis of its spectra [306].

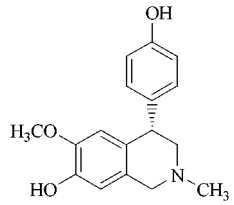

277 (−)-Cherylline

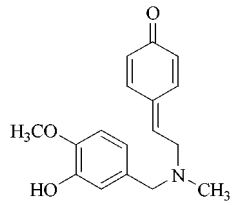

278

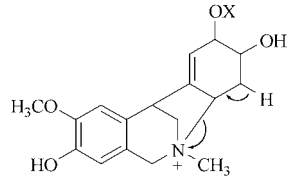

279

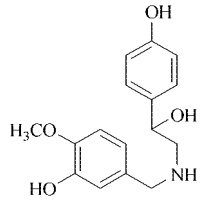

280

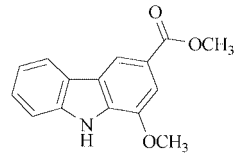

281 Mukoine

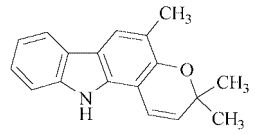

282 Mupamine

11.2. Non-Isoprenoid Type

Tryptamine (**283**) and serotonin (**284**) represent the basic skeleton and are biosynthetic precursors for this group of alkaloids. Harmine (**285**) is a typical β-carboline alkaloid. Although the structure relationship between physostigmine (**286**) and tryptophan appears obscure, it is also included in this group. Also calycanthine (**287**) does not possess a tryptamine structure as a distinguishable feature; however, it is included in an indole alkaloid because most of its congeners are tryptamine derivatives.

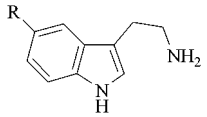

283 R = H Tryptamine
284 R = OH Serotonin

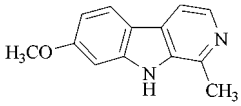

285 Harmine

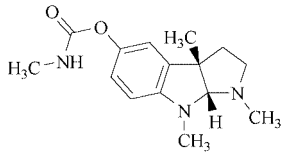

286 Physostigmine (Eserine)

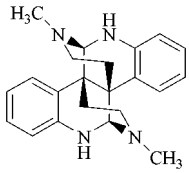

287 Calycanthine

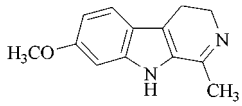

288 Harmaline

Physical Properties.

Harmine (**285**) [*442-51-3*], $C_{13}H_{12}N_2O$, M_r 212: *mp* 261 °C decomp. (from methanol).

Physostigmine (**286**) [*57-47-6*], $C_{15}H_{21}N_3O_2$, M_r 275: *mp* 105 – 106 °C (from ether or benzene), $[\alpha]_D^{17}$ − 76 ° (chloroform), $[\alpha]_D^{25}$ − 120 ° (benzene).

Harmaline (**288**) [*304-21-2*], $C_{13}H_{14}N_2O$, M_r 214: *mp* 229 – 231 °C (from ethanol).

Pharmacological Properties. Harmine is less toxic than harmaline. Harmaline brings about tremors and convulsions in moderate doses. Lethal doses in mammals cause respiratory paralysis and decreased body temperature. The contraction of most kinds of smooth muscle is diminished. An exception is uterine muscle. A large number of doses reduce blood pressure by weakening the cardiac muscle [307]. Physostigmine is known to be a cholinesterase inhibitor. It stimulates the central nervous system and causes respiratory paralysis. Large doses reduce blood pressure, cause a convulsion of skeletal muscles, and act as a cardiac depressant. Its sulfate and salicylate are used medically. The effects of calycanthine resemble those of strychnine: calycanthine stimulates the spinal cord but depresses the heart. Its hydrochloride reduces blood pressure and cardiac contraction in anaesthetized cats. LD_{50} i.v. in mice: 38 mg/kg for harmine. LD_{50} i.p. in mice: 0.64 mg/kg for eserine salicylate. LD_{50} s.c. in rats: 120 mg/kg for harmaline.

11.3. Isoprenoid Type

11.3.1. Mould Metabolites

Brevianamide E (**289**) [308], isolated from culture medium of *Penicillium brevicompactum,* is a toxic mould metabolite. Structure **289** was proposed mainly on the basis of spectroscopic evidence and plausible biogenetic argument. Its absolute configuration was confirmed by the total synthesis from L-tryptophan [309]. Echinuline (**290**) and aurechinuline (**291**) [310] are also toxic fungi metabolites. They are isolated from several *Aspergillus* species. Characteristic features of these alkaloids are isopentenyl groups on the aromatic ring and a diketopiperazine ring.

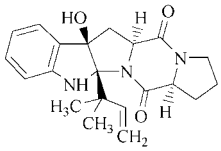

289 Brevianamide E

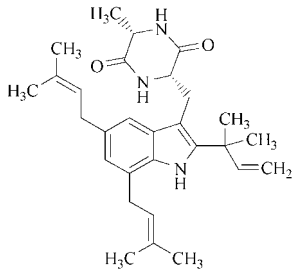

290 Echinuline

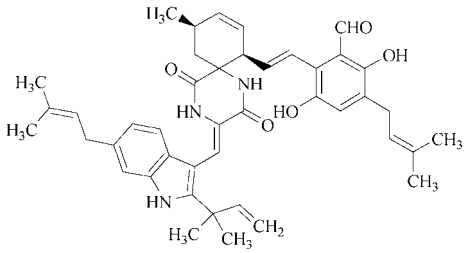

291 Aurechinuline

11.3.2. Ergot Alkaloids

Structure, Occurrence, and Synthesis. Lysergic acid (**292**) and isolysergic acid (stereoisomers in respect to the carboxyl group at the 8-position) are the main cleavage components of ergotamine-type alkaloids. Agroclavine (**293**) [311] was isolated from the mycelium of *Claviceps purpurea*. Clavine-type alkaloids, which are tricyclic ergot alkaloids, were found in several *Claviceps* species. Chanoclavine I (**294**) [312] and chanoclavine II (**295**) [313] were shown to be stereoisomers in respect to substituents on ring C. Ergotamine (**296**) [314] and ergosine (**297**) [315] contain a peptide unit in which the amino acid components are different. Paliclavine (**298**) [313] is an allyl alcohol rearrangement product of chanoclavine. Chemical correlation of these alkaloids (**292**–**295**), e.g. catalytic reduction over palladium, allowed elucidation of the structures of these natural products.

Many reports concerning synthesis have appeared.

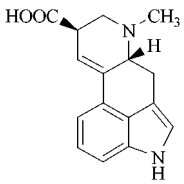

292 Lysergic Acid

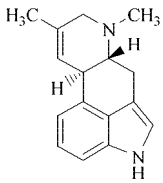

293 Agroclavine

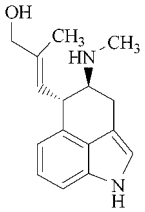

294 Chanoclavine I

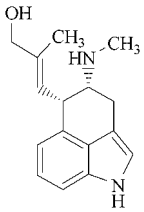

295 Chanoclavine II

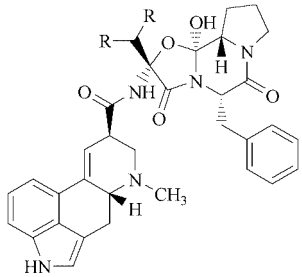

296 R = H Ergotamine
297 R = CH_3 Ergocristine

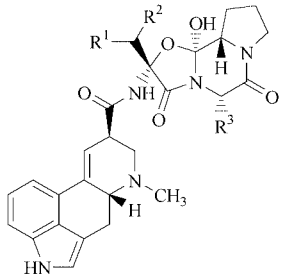

298 R^1 = H, R^2 = H, R^3 = $CH_2CH(CH_3)_2$
Ergosine
299 R^1 = CH_3, R^2 = CH_3, R^3 = $CH_2CH(CH_3)_2$
Ergocryptine
300 R^1 = CH_3, R^2 = CH_3, R^3 = $CH(CH_3)_2$
Ergocornine

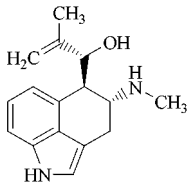

301 Paliclavine

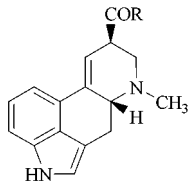

302 R = $NHCH(CH_3)CH_2OH$
Ergobasine
303
R = $HNCH(CH_2OH)(CH_2CH_3)$
N-[1-(Hydroxymethyl)propyl]-lysergamide

Physical Properties.

Lysergic acid (**292**) [*82-58-6*], $C_{16}H_{16}N_2O_2$, M_r 268: *mp* 238 °C (from water), $[\alpha]_D^{20}$ +32° (pyridine).

Isolysergic acid, $C_{16}H_{16}N_2O_2$, M_r 268: *mp* 218 °C.

Ergocristine (**296**) [*511-08-0*], $C_{35}H_{39}N_5O_5$, M_r 609: prisms, *mp* 160–175 °C decomp. (from acetone), $[\alpha]_D^{20}$ −280° (*c* = 1.0 g/100 ml in chloroform).

Ergocryptine (**299**) [*511-09-1*], $C_{32}H_{41}N_5O_5$, M_r 575: prisms, *mp* 210–212 °C (from methanol), $[\alpha]_D^{20}$ −191° (*c* = 1.0 g/100 ml in chloroform).

Ergocornine (**300**) [*564-36-3*], $C_{31}H_{39}N_5O_5$, M_r 561: *mp* 182–184 °C decomp. (from methanol), $[\alpha]_D^{20}$ −186° (*c* = 1.0 g/100 ml in chloroform).

9,10-Dihydroergotamine [*511-12-6*], $C_{33}H_{37}N_5O_5$, M_r 583: prisms, *mp* 239 °C (from aqueous 90 % acetone), $[\alpha]_D^{20}$ −64° (pyridine).

9,10-Dihydroergocristine [*17479-19-5*], $C_{33}H_{41}N_5O_5$, M_r 611: hexagonal plates, *mp* 180 °C (from acetone), $[\alpha]_D^{20}$ −56° (pyridine).

9,10-Dihydroergocryptine [*17479-14-0*], $C_{32}H_{43}N_5O_5$, M_r 577: plates or polyhedrons, *mp* 235 °C (from ethanol), $[\alpha]_D^{20}$ −41° (pyridine).

9,10-Dihydroergocornine [*17479-17-3*], $C_{31}H_{41}N_5O_5$, M_r 563: plates, *mp* 187 °C (from ethanol), $[\alpha]_D^{20}$ −48° (pyridine).

N-[1-(Hydroxymethyl)propyl]-lysergamide (Methylergometrine, Methylergonovine) (**303**) $C_{20}H_{25}N_3O_2$, M_r 339: *mp* 172 °C (from benzene), $[\alpha]_D^{20}$ −45° ($c = 0.4$ in pyridine).

N-[1-(Hydroxymethyl)propyl]-1-methyllysergamide (Methysergide), $C_{21}H_{27}N_3O_2$, M_r 353: *mp* 194–196 °C, $[\alpha]_D^{20}$ −45° ($c = 0.5$ in pyridine).

N-Cyclopentyllysergamide, $C_{21}H_{25}N_3O$, M_r 335: *mp* 121–122 °C, $[\alpha]_D^{20}$ −27.9° (pyridine).

Pharmacological Properties. Lysergic diethylamide is a known antagonist to serotonin, and sometimes it causes psychological trouble. Ergolines may be potential inhibitors of prolactin and mammary tumours. Ergotamine stimulates the smooth muscles of only the uterus and blood vessels. Therefore it increases blood pressure by constricting blood vessels. It does not affect other smooth muscles. The tartrate is usually used for relief of migraine. The pharmacology of ergot alkaloids has been reviewed.

Therapeutical Use. Ergocristine: Uterotonic, vasodilator.

Ergocryptine: vasodilator, antihypertensive agent.

Ergocornine: vasodilator, antihypertensive agent.

9,10-Dihydroergotamine: against migraine.

9,10-Dihydroergocristine: antihypertensive agent, vasodilator, against migraine.

9,10-Dihydroergocornine: antihypertensive agent, against migraine.

N-[1-(Hydroxymethyl)propyl]-1-lysergamide: uterotonic.

N-[1-(Hydroxymethyl)propyl]-1-methyllysergamide: antiserotonin activity (migraine, carcinoid syndrome).

N-Cyclopentyllysergamide: Uterotonic.

Trade Names. 9,10-Dihydroergotamine: Dihydergot (Sandoz). *Tartrate:* Eldoral (Dumex). *Methylsulfonate:* Dihydergot (Sandoz), Detms (Rentschler), Dihytamin (Arzneimittelwerk Dresden, Temmler Pharma), Dihydroergotamine (Lafarge), Ergont (Desitin), Ikaran (Sinbio), Séglor (Millot), Tonopres (Boehringer Ingelheim), Dihydergot (Sankyo), Restal (Tokyo Tanabe), Dihygotamine (Kaken).

9,10-Dihydroergocornine: Together with 9,10-dihydroergocristine and 9,10-dihydroergocryptine in a mixed preparation: Hydergin (Sandoz), Hydergin (Sankyo).

N-[1-(Hydroxymethyl)propyl]lysergamide: Methylergobasin, Methylergobrevin (Arzneimittelwerk Dresden), Methylergometrin (Sandoz), Methylergonovin (Burroughs-Wellcome). *Methylergometrin maleinate:* Methergin (Sandoz). *Malate:* Levospan (Isei), Metenarin (Teikoku Zoki), Ryeonovin (Morishita), Spametrin-S (Yamanouchi), Takimetrin-M (Kanebo).

N-[1-(Hydroxymethyl)propyl]methyllysergamide: Deseril (Sandoz), Sandert (Sandoz).

N-Cyclopentyllysergamide: Cepentyl.

11.3.3. Monoterpenoid Alkaloids

Feeding experiments on *Vinca rosea* showed that the non-tryptamine unit of monoterpene alkaloids was derived from two molecules of mevalonic acid. Further investigation suggested that secologanin, derived from geraniol, was a key precursor of these alkaloids. The non-tryptamine part can be of the (**A**) Corynanthe-Strychnos type, (**B**) Aspidosperma type, or (**C**) Iboga type.

11.3.3.1. Corynanthe Group

Vincoside (**304**) and strictosidine (**305**) are stereoisomers at the 3-position. They are synthesized from secologanin by condensation with tryptamine [317]. Vincoside (**304**) was readily converted into vincoside lactam (**308**) by treatment with aqueous sodium carbonate at room temperature, whereas strictosidine (**305**) was transformed into nacycline (**310**) on treatment with methanolic hydrochloric acid [318]. Zenk indicated that terpenoid indole alkaloids of 3β configuration also arise biosynthetically from strictosidine [319]. The biosynthetic conversion proceeds with loss of hydrogen at the 3-position to form the 3α-*Corynanthe* alkaloids, whereas the hydrogen is retained in the 3β series. Corynantheine (**306**) is a typical indole alkaloid bearing an indoloquinolizidine skeleton. Three other

possible diastereomers have been found in nature: hirsuteine, corynantheidine, and spiociliatine. Antirhine [320] was isolated from *Antirrhea putaminosa,* and the structure was assigned as **307**. Strictosidine was correlated chemically with corynantheine and antirhine. The cinchona alkaloid was shown to be derived from corynantheal (**311**) biosynthetically.

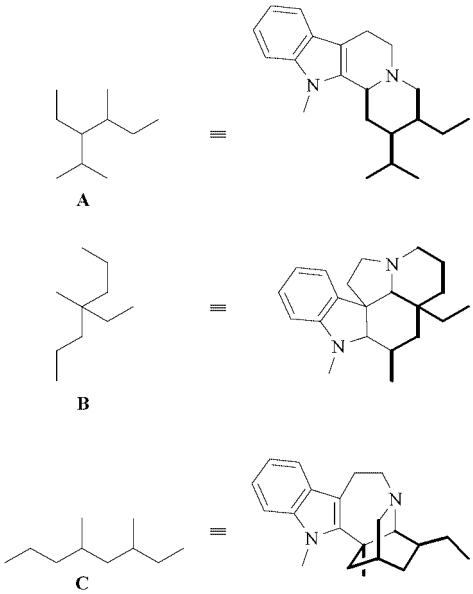

11.3.3.2. Heteroyohimbine

Ajmalicine (**312**) is an early stage of alkaloid in the biosynthesis of indole alkaloids from tryptamine and secologanin. It is converted to the oxindole alkaloid mitraphylline (**313**) in *Mitragyna parvifolia* [321].

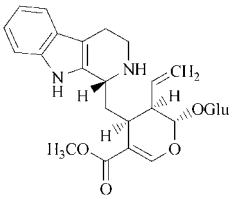

304 Vincoside
Glu = glucose

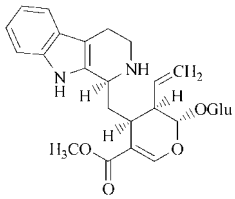

305 Strictosidine

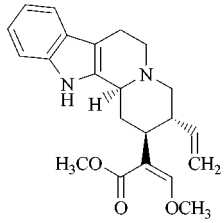

306 Corynantheine

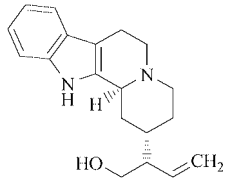

307 Antirhine

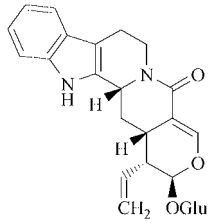

308 Vincoside lactam

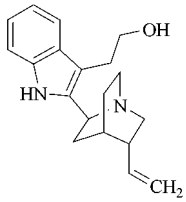

309 Cinchonamine

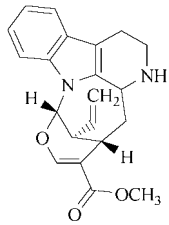

310 Nacycline

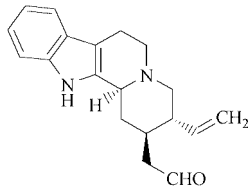

311 Corynantheal

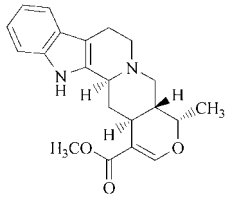

312 Ajmalicine

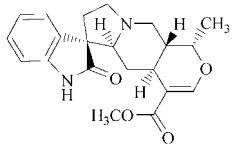

313 Mitraphylline

Ajmalicine (**312**) [*483-04-5*], (δ-Yohimbine, Raubasine, Vinceine, Vincain), $C_{21}H_{24}N_2O_3$, M_r 352: colorless prisms (from methanol), *mp* 257 °C (decomp.), $[\alpha]_D^{20}$ − 60 ° (chloroform).

Therapeutic Use: Against central and peripheral ischemia. Antihypertensive reagent and tranquilizer.

Trade Names: Hydrosarpan (Servier), Circolene (Inverni Della Beffa), Sarpan (Farge), Raubaserp (Ethica), Loparol (Boehringer Mannheim).

11.3.3.3. Yohimbine

Occurrence and Synthesis. Yohimbine (**314**) was found in *Corynanthe* and related trees.

Its epimers are known as alloyohimbine (**315**) and α-yohimbine (**316**). Synthesis of yohimbine (**314**) was reported by VAN TAMELEN [322]. Reserpine is a major constituent of *Rauwolfia serpentina*. Alkali hydrolysis of **318** afforded reserpic acid (**317**). WOODWARD [323] reported the stereoselective total synthesis of reserpine (**318**).

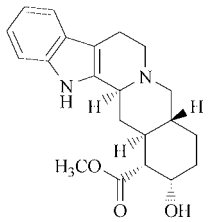

314 Yohimbine

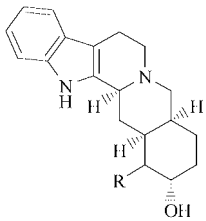

315 R = α-CO_2CH_3 Alloyohimbine
316 R = β-CO_2CH_3 α-Yohimbine

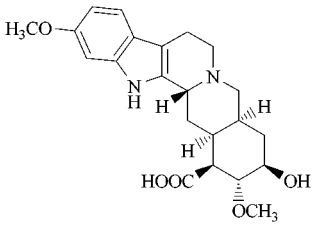

317 Reserpic acid

Physical Properties.

Yohimbine (Quebrachine) (**314**) [*46-48-5*], $C_{21}H_{26}N_2O_3$, M_r 354: colorless needles (from dilute ethanol) *mp* 234 – 235 °C (decomp.), $[\alpha]_D^{20}$ +62.2 ° (ethanol).

Reserpine (**318**) [*50-55-5*], $C_{33}H_{40}N_2O_9$, M_r 608: colorless crystals (from acetone), *mp* 262 – 266 °C (decomp.), $[\alpha]_D^{23}$ − 118 ° (chloroform).

Rescinnamine (reserpinine) (**319**) [*24815-24-5*] $C_{35}H_{42}N_2O_9$, M_r 634: colorless needles (from benzene), *mp* 238 – 239 °C, $[\alpha]_D^{24}$ − 97 ° (chloroform).

Syrosingopine (**321**) [*84-36-6*], $C_{35}H_{42}N_2O_{11}$, M_r 666: *mp* 175 – 179 °C (from acetone).

Methoserpine (methoserpidine, 10-methoxy-11-demethoxyreserpine) (**320**), $C_{33}H_{40}N_2O_9$, M_r 608: colorless crystals (from methanol), *mp* 173 °C, $[\alpha]_D^{20}$ − 142 ° (chloroform).

Deserpidine (**322**) [*131-01-1*], $C_{32}H_{38}N_2O_8$, M_r 578: crystallizes from methanol in three crystal types: α-type, *mp* 228 – 232 °C; β-type, *mp* 230 – 232 °C; γ-type *mp* 138 °C, resolidification at 175 °C. $[\alpha]_D^{20}$ − 163 ° (c = 0.5 in pyridine).

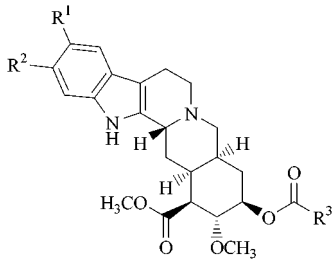

318 R^1 = H, R^2 = OCH_3, Reserpine

R^3 =

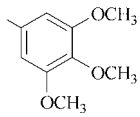

319 R^1 = H, R^2 = OCH_3, Rescinnamine

R^3 =

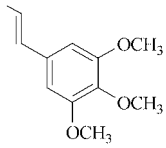

320 R^1 = OCH_3, R^2 = H, Methoserpidine

R^3 =

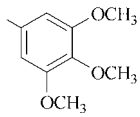

321 R^1 = H, R^2 = OCH_3, Syrosingopine

R^3 =

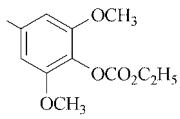

322 R^1 = R^2 = H, Deserpidine

R^3 =

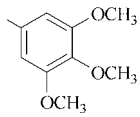

Preparation. The isolation of yohimbine from *Pausinystalia Johimbe* (K. Schum.) Pierre (*Corynanthe yohimbe* K. Schum.) is described by F. CHEMNITIUS [324].

The following procedure is suitable for isolating reserpine from *Rauwolfia serpentina* [325]:

Finely powdered roots (750 g) of *Rauwolfia serpentina* Benth. are moistened with water (800 ml) and left to stand for three hours to ensure even moistening and swelling. The drug is then extracted for 20 h with 3 – 4 l of benzene in a

Soxhlet extractor. The benzene extract is shaken three times with 200 ml of 10 % phosphoric acid, and the phosphoric acid solution adjusted to pH 6.5 – 7.0 with hydrogencarbonate. Then this solution is shaken five times with 200 ml of benzene, the benzene extract dried with sodium sulfate, and the solution evaporated in vacuo. A pale foamy residue (4.4 – 5.0 g) remains. Reserpine can be isolated from this concentrate by liquid chromatography. It is eluded, first with benzene, and then with benzene/acetone (2 : 1), from a column of neutral aluminum oxide. The alkaloid is recrystallized from hot acetone and melts at 262 – 263 °C with decomposition.

Pharmacological Properties. Yohimbine is a local anaesthetic, reduces blood pressure, and increases depth and frequency of respiratory movements. It is known to be an adrenergic blocking agent. It has been used as an aphrodisiac. The diacetyl derivative is also an effective vasodilator. Reserpine restricts local arteries and reduces preservation of catecholamine and serotonin. Reserpine is a tranquilizer and is used to prevent aortic rupture in turkeys. It is an antihypertensive agent.

Therapeutic Use. Rescinamine: tranquilizer, antihypertensive agent, sympathicolytic.

Syrosingopine: antihypertensive agent.

Methoserpine: antihypertensive agent.

Deserpidine: tranquilizer.

Trade Names. Yohimbine: Yohimbin "Buchler", Yohimbine "Houde" (Lirca), Yohimbin-Spiegel (Kalichemie).

Reserpine: Eskaserp (Smith Kline & French), Rau-Sed (Squibb), Réserpine Chantereau, Reserpin (Leo), Réserpine Nativelle, Rivasin (Giuliani), Sandril (Lilly), Serpate (Vale), Reserpin (Hameln), Abicol (Boots), Seominal (Winthrop), Sedaraupin (Boehringer Mannheim), Serpasil (Ciba), Spansule (Smith Kline & French), Reserpoid (Upjohn), Rauvilid (Pharmacia), Apoplon (Daiichi), Resermin (Takeda), Toriserpin (Torii), Resera (Sanwa).

Rescinamine: Moderil (Pfizer), Rozex (Teisan), Apolon (Toyama), Ruconax (Sato), Paresinan (Wakamoto), Anaprel (Servier), Recinate (Ohta).

Methoserpine: Decaserpyl (Roussel).

Deserpidine: Harmonyl (Abbott), Deserpidine (Ono).

11.3.3.4. Sarpagine

Sarpagine was isolated from *Rauwolfia serpentina*. Its stereochemistry was assigned as **323** on the basis of spectroscopic evidence.

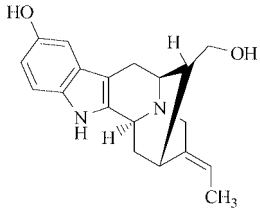

323 Sarpagine

11.3.3.5. Ajmaline

Ajmaline was found in the root of *Rauwolfia serpentina* to the extent of about 0.5 %. Its structure was determined by WOODWARD and ROBINSON [326]. The first total synthesis of ajmaline (**324**) was reported by MASAMUNE [327].

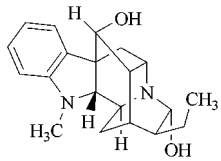

324 Ajmaline

Ajmaline (**324**) [*4360-12-7*], $C_{20}H_{26}N_2O_2$, M_r 326: prisms, adduct with 1 mol of methanol (from methanol), *mp* 158 – 160 °C, $[\alpha]_D^{18}$ + 131 ° (c = 0.4 g/100 ml in chloroform).

Pharmacological Properties. Ajmaline shows an anti-ephedrine activity and is used as an anti-arrhythmic drug, antihypertensive, and tranquilizer.

Therapeutic Use: antiarrhythmic agent.

Trade Names: Gilurytmal (Giuliani), Cardiorythmine 5 (Servier), Ritmos (Inverni Della Beffa), Raugallin (Chibe), Ajmal (Kanto Ishi), Ajmalin (Mochida), Ajsal (Teisan).

11.3.3.6. Picraline

Picraline (**325**) and related alkaloids were found in several *Alstonia* species. Their structures were

determined on the basis of spectroscopic evidence. Picralinal (**326**) has possible hypotensive and anticancer properties.

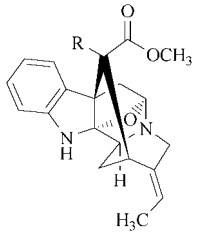

325 R = CH_2OAc Picraline **326** R = CHO Picralinal Ac = $COCH_3$

11.3.3.7. Vobasine

These alkaloids have a 2-acylindole skeleton in their structure. Dregamine (**328**) is a dihydro compound of vobasine (**327**). The conversion of the *N*-oxide of **327** and **328** into the ervatamine-type alkaloids was achieved in vitro. Therefore, the synthesis of dregamine reported by KUTNEY constitutes a formal synthesis of 20-*epi*-ervatamine (**329**) [328].

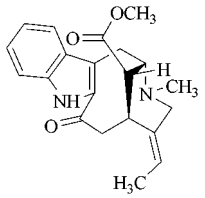

327 Vobasine

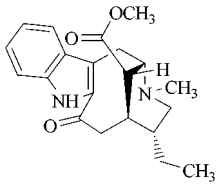

328 Dregamine

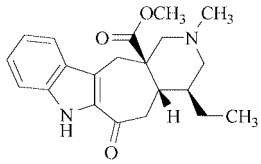

329 Ervatamine

11.3.3.8. Pleiocarpamine

Pleiocarpamine (**330**), isolated from *Pleiocarpamutica* Benth., is a C-20 indole alkaloid, which is converted into the isomeric epipleiocarpamine on treatment with strong bases. Subsequent reduction with lithium aluminum hydride and *N*-methylation with methyl iodide afforded C-mavacurine (**331**).

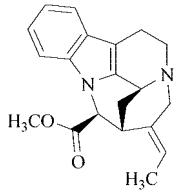

330 Pleiocarpamine

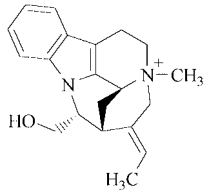

331 C-Mavacurine

11.3.3.9. Strychnine

Structure, Occurence, and Synthesis. Strychnine (**332**) is the most important of the *Strychnos* alkaloids. Its structure was initially determined by ROBINSON and WOODWARD and confirmed by X-ray analysis [329]. Total synthesis of strychnine was accomplished by WOODWARD, who employed a catechol cleavage reaction (Woodward fission) [330]. Strychnine was converted into caracurine VII (**334**), also isolated from *Strychnos toxifera,* by WIELAND and GUMLICH [331]. Investigation of biosynthesis of strychnine (**332**) revealed that geissoschizine and the Wieland-Gumlich aldehyde (**334**) were key precursors. The *N*-acetyl derivative of the Wieland-Gumlich aldehyde is diaboline. Aspidospermatidine (**336**) and condylocarpine (**337**) were also placed in this group because of their structural relationship to strychnine.

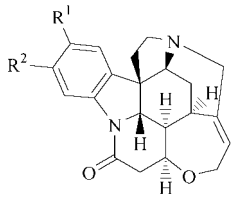

332 R^1 = H, R^2 = H Strychnine

333 $R^1 = OCH_3$, $R^2 = OCH_3$ Brucine

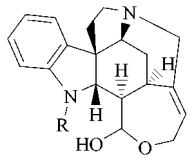

334 R = H, Wieland-Gumlich aldehyde (Caracurine VII)

335 R = Ac Diaboline

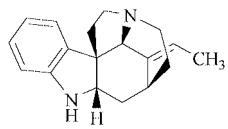

336 Aspidospermatidine

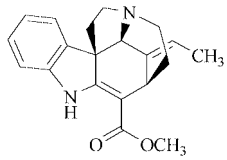

337 Condylocarpine

Physical Properties.

Strychnine (**332**) [*57-24-9*], $C_{21}H_{22}N_2O_2$, M_r 334: colorless rhombohedrons (from ethanol), *mp* 268–290 °C (depends on heating rate), $[\alpha]_D$ − 145 ° (chloroform).

Brucine (**333**) [*357-57-3*], $C_{23}H_{26}N_2O_4$, M_r 394: colorless prisms (from ethanol/water), *mp* 178 °C, $[\alpha]_D$ − 127 ° (chloroform).

Wieland-Gumlich aldehyde (**334**) [*466-85-3*] $C_{19}H_{22}N_2O_2$, M_r 310: *mp* 213–214 °C decomp. (from acetone/methanol), $[\alpha]_D^{22}$ − 133.8 ° (c = 0.52 in methanol).

Diaboline (**335**) [*509-40-0*], $C_{21}H_{24}N_2O_3$, M_r 352: *mp* 187 °C (from ether), $[\alpha]_D^{22}$ + 37.8 ° (c = 1.72 in chloroform).

Preparation. Only strychnine is isolated on an industrial scale, and only strychnine and products obtainable from it are of practical importance.

Vomiting nuts (*Nux vomica*) are used as the starting material for the isolation of strychnine. The raw nuts contain 2–3 % alkaloids, of which approximately half is strychnine. The alkaloids are set free with calcium hydroxide, extracted continuously with benzene or toluene, and transferred from the organic phase to an aqueous sulfuric acid phase. Precipitation with soda solution gives a raw alkaloid mixture, from which pure strychnine is obtained by crystallization with 50 % ethanol and recrystallization. Brucine can be isolated from the mother liquors; colubrine and vomicine, from the final mother liquors.

Pharmacological Properties. Strychnine is a highly toxic alkaloid and affects the spinal cord to produce excessive reflex irritability. Lethal doses lead to the death of mammals by respiratory failure. In large doses, strychnine paralyzes the central nervous system, leading instantaneously to the death of mammals. It is used in medicine as a tonic and stimulant; however, its major use is as a vermin killer. MLD (minimum lethal dose) orally in rats: 5 mg/kg for strychnine.

Therapeutic Use. Strychnine: analeptic agent.

Brucine: analeptic agent.

Strychnic acid: depot analeptic agent.

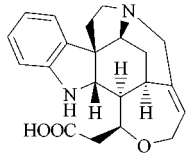

338 Strychnic acid

Trade Names. Strychnine: Strychnine (Mallinckrodt, MSD), Strychnin. nitric. (Dr. Buchler), Strychnin Nitrate (Torii, Kyoritsu).

Strychnine dimethylarsinate: Strychnium Kadoylicum (Boehringer Ingelheim). *Strychnine Sulfate* (Lilly), *Strychnine Phosphate* (Smith Klein Fujisawa).

Brucine: Genostrychnin (Amido), Invocan (Pharm. Fabr. Hameln), Movellan (Asta).

Strychnic acid : Movellan (Asta).

11.3.3.10. Akuammicine

Akuammicine (**339**) is a minor constituent of *Picralima klaineana* Pierre and is chemically derived from strychnine [332].

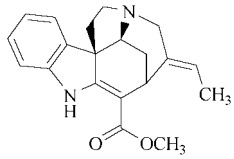

339 Akuammicine

Physical Properties.

Akuammicine (**339**) [*639-43-0*], $C_{20}H_{22}N_2O_2$, M_r 322: *mp* 182 °C (from dilute ethanol), $[\alpha]_D^{16}$ − 745 ° (c = 0.994 in ethanol).

11.3.3.11. Ellipticine

Structure, Occurrence, and Synthesis. Ellipticine (**340**) [333] and olivacine (**341**) [334] lack the normal tryptamine unit in their fundamental structure; however, they are placed with the Aspidosperma alkaloids as they were found in *Aspidosperma* species. Interest in the synthesis of these alkaloids stems from their marked antitumor activity. The first synthesis of ellipticine (**340**) was achieved by WOODWARD [335]. Olivacine (**341**) was synthesized by several groups [336]. Uleine (**342**), apparicine (**343**), and vallesamine (**344**) also do not bear

a tryptamine unit as a distinguishable feature in their fundamental structure, but they are normally grouped under the *Aspidosperma* family since they occur in *Aspidosperma* species. Synthesis of uleine (**342**) was achieved stereoselectively by BÜCHI [337], whereas KAMETANI [338] published the efficient synthesis of several members of the dasycarpidone-uleine series.

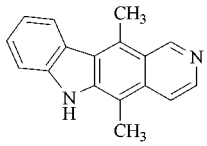

340 Ellipticine

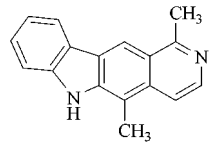

341 Olivacine

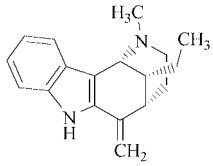

342 Uleine

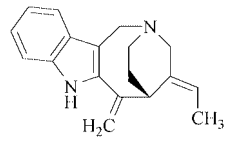

343 Apparicine

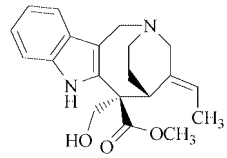

344 Vallesamine

Physical Properties.

Ellipticine (**340**) [*519-23-3*], $C_{17}H_{14}N_2$, M_r 246: *mp* 311 – 315 °C decomp. (from ethyl acetate).

Olivacine (**341**) [*484-49-1*], $C_{17}H_{14}N_2$, M_r 246: *mp* 317 – 325 °C (from methanol).

Pharmacological Properties. Ellipticine and olivacine exhibited antitumor activity in appropriate biological evaluation in various tumor systems. LD_{50} i.v. in mice: 19.5 – 22.4 mg/kg for ellipticine.

11.3.3.12. Aspidospermine

Aspidosperma alkaloids are derived biosynthetically from tryptophan and secologanin via geissoschizine, stemmadenine, and dehydrosecodine. Many syntheses and synthetic approaches were attempted along this pathway, which involves an intramolecular Diels-Alder reaction of dehydrosecodine. STORK's first synthesis of aspidospermine (**345**) utilized an enamine annelation reaction and Fischer indole cyclization of a phenylhydrazone derivative [339]. The synthesis of vindoline (**348**), a major component of *Vinca rosea,* was accomplished by BÜCHI [340] and KUTNEY [341], and this alkaloid is an important compound because it is the precursor of dimeric indole alkaloids possessing antitumor activity. Treatment of vincadifformine, obtained from tabersonine (**347**) by partial reduction with formic acid and formamide followed by reduction, afforded vincadine and quebrachamine (**346**).

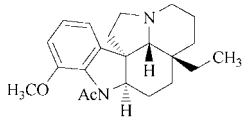

345 Aspidospermine
Ac = $COCH_3$

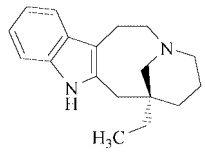

346 Quebrachamine

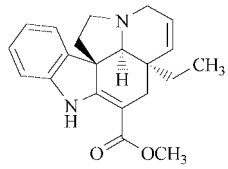

347 Tabersonine

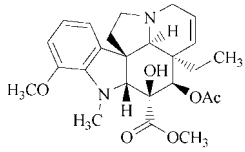

348 Vindoline

Physical Properties.

Aspidospermine (**345**) [*466-49-9*], $C_{22}H_{30}N_2O_2$, M_r 354: *mp* 208 °C (from petroleum ether), $[\alpha]_D^{15}$ – 100.2 ° (ethanol), $[\alpha]_D$ – 92 ° (chloroform).

Quebrachamine (**346**) [*4850-21-9*], $C_{19}H_{26}N_2$, M_r 282: *mp* 145 – 147 °C, $[\alpha]_D^{20}$ – 109 to – 110 ° (acetone).

Vindoline (**348**) [*2182-14-1*], $C_{25}H_{32}N_2O_6$, M_r 456: *mp* 164 – 165 °C for needles (from acetone + petroleum ether), *mp* 174 – 175 °C for prisms, $[\alpha]_D^{20}$ – 18 ° (chloroform).

11.3.3.13. Aspidofractine

The hexacyclic alkaloids are considered to be derived biogenetically from aspidospermine-type pentacylic alkaloids. In fact, minovincine (**353**) was chemically converted into aspidofractinine (**352**) by treatment with hydrochloric acid and subsequent reduction [377]. The degradation of

kopsine (**351**) to aspidofractinine (**352**) indicated its absolute configuration, [343].

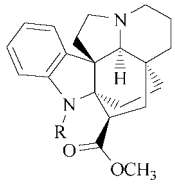

349 R = CHO Aspidofractine
350 R = CO_2CH_3 Pleiocarpine

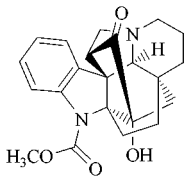

351 Kopsine

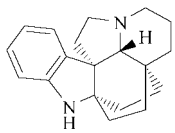

352 Aspidofractinine

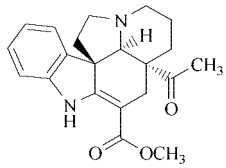

353 Minovincine

Kopsine (**351**) [*559-48-8*], $C_{22}H_{24}N_2O_4$, M_r 380: *mp* 217–218 °C decomp. (from ethanol), $[\alpha]_D^{27}$ −14.3° ± 1° (*c* = 2.0 in chloroform).

11.3.3.14. Eburnamine

Eburnamine-type alkaloids are the rearrangement products of aspidospermine-type compounds. 1,2-Dehydroaspidospermidine *N*-oxide (**358**) was converted into eburnamine (**354**) by treatment with triphenylphosphine in aqueous acetic acid [344]. Moreover, vincamine (**355**) was formed from tabersonine in several steps. (−)-Eburnamonine (**356**) was also prepared from vincamine (**355**). Schizophylline (**357**) is also grouped under *Aspidosperma* family on the basis of its natural source; however, its biosynthetic pathway has not yet been determined.

Physical Properties.
Vincamine (**355**) [*1617-90-9*], $C_{21}H_{26}N_2O_3$, M_r 354: *mp* 232–233 °C (from acetone or methanol), $[\alpha]_D^{23}$ +41° (pyridine).
Eburnamonine (**356**) [*474-00-0*], $C_{19}H_{22}N_2O$, M_r 294: *mp* 174 °C, $[\alpha]_D^{25}$ +89° (chloroform).

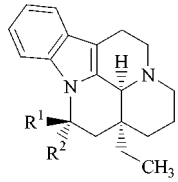

354 R^1 = H, R^2 = OH Eburnamine
355 R^1 = OH, R^2 = CO_2CH_3 Vincamine
356 $R^1 + R^2$ = O Eburnamonine

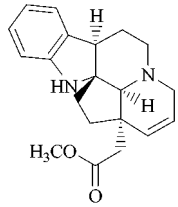

357 Schizophylline

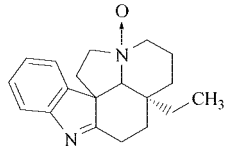

358

11.3.3.15. Catharanthine

Catharanthine (**359**) is a minor constituent of *Vinca rosea* [345] and is derived from dehydrosecodine by an intramolecular cyclization. This alkaloid is important as the upper half of the dimeric indole alkaloids vinblastine and vincristine.

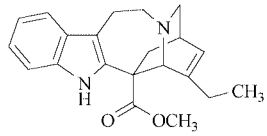

359 (+)-Catharanthine

Physical Properties.
Catharanthine (**359**) [*2468-21-5*], $C_{21}H_{24}N_2O_2$, M_r 336: *mp* 126–128 °C (from methanol), $[\alpha]_D^{27}$ +29.8° (chloroform).

11.3.3.16. Ibogamine

Ibogamine (**360**) and ibogaine (**361**) were isolated from *Tabernanthe iboga,* and their structures were determined by X-ray analysis [346] and total synthesis [347].

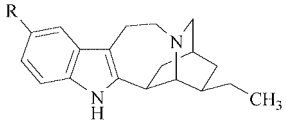

360 R = H (−)-Ibogamine **361** R = OCH_3 (−)-Ibogaine

It is reported that ibogaine inhibits the action of serum cholinesterase.

Physical Properties.
Ibogaine (**361**) [*83-74-9*], $C_{20}H_{26}N_2O$, M_r 310: *mp* 152–153 °C (from ethanol), $[\alpha]_D^{20}$ −53 ° (95 % ethanol).

11.3.3.17. Cleavamine

Cleavamine (**362**) was derived from catharanthine (**359**) by treatment with acid. Its stereochemistry was confirmed by X-ray analysis [348]. Cleavamine was chemically related to ibogamine (**360**). Oxidation of 16α-carbomethoxycleavamine with mercuric acetate afforded two products, one of which was identical with the dehydration product of pandoline (**364**) except for its optical rotation. Dihydrocleavamine (**363**) was also converted into pseudovincadifformine (**368**) by oxidation with mercuric acetate.

Two new indole alkaloids, ibophyllidine (**366**) and iboxyphylline (**367**), were obtained from *Tabernanthe iboga.* Their structures were assigned as **366** and **367** on the basis of spectroscopic evidence. These two alkaloids were suggested to be derived from pandoline (**364**) biosynthetically. The conversion of pandoline (**364**) into velbanamine (**369**) was achieved by catalytic hydrogenation in trifluoroacetic acid followed by acid catalyzed decarboxylation.

11.3.3.18. Andranginine

Andranginine (**370**), isolated from *Craspidospermum verticillatum*, was synthesized from 19-iodotabersonine (**371**) via the olefin **372** by dehydrohalogenation followed by heating in methanol, to confirm its biosynthetic pathway.

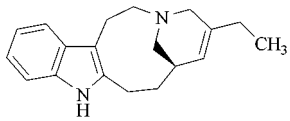

362 (+)-Cleavamine

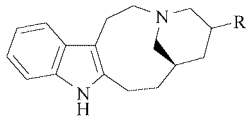

363 R = α-Et
20α-Dihydrocleavamine
R = β-Et
20β-Dihydrocleavamine

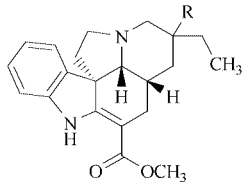

364 R = α-OH
Pandoline
365 R = β-OH
Isopandoline

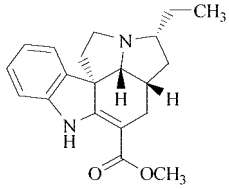

366 Ibophyllidine

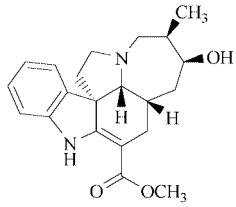

367 Iboxyphylline

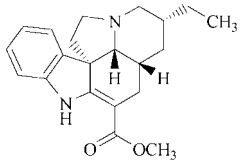

368 Pseudovincadifformine

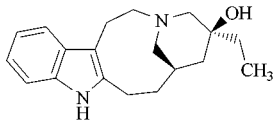

369 (+)-Velbanamine

11.3.4. Bis-Indole Alkaloids

Strucure and Synthesis. Vinblastine (**373**) is an important alkaloid because of its biological activity. Vinblastine seems to be derived biosynthetically from vindoline and catharanthine on the basis of tracer experiments with a cell-free enzyme system [349]. Synthesis of vinblastine was accomplished by POTIER, who employed a modified Polonovski reaction to control the stereochemistry at the 18-position [350]: Catharanthine *N*-oxide was treated with trifluoroacetic anhydride in the presence of vindoline to give dehydrovinblastine (**375**). Catalytic reduction afforded deoxyvinblastine. Repeated Polonovski reaction followed by introduction of a hydroxyl group completed the synthesis of vinblastine. Vincristine (**374**) was synthesized from vinblastine (**373**) by selective oxidation of the *N*-methyl

function of vindoline with the Jones reagent at low temperature. Villalstonine (**376**) is obtained from several *Alstonia* species.

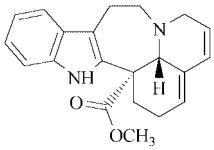

370 Andranginine

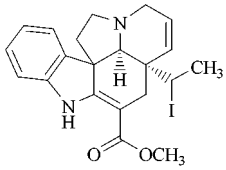

371

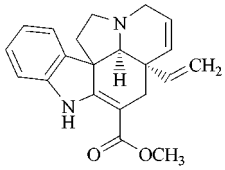

372

Physical Properties.

Vinblastine (Vincaleukoblastine) (**373**) [*865-21-4*], $C_{46}H_{58}N_4O_9$, M_r 831: *mp* 211 – 216 °C decomp. (from methanol), $[\alpha]_D^{26}$ + 42 ° (chloroform).

Vincristine (22-oxovincaleukoblastine, Leurocristine) (**374**) [*57-22-7*], $C_{46}H_{56}N_4O_{10}$, M_r 844: *mp* 218 – 220 °C decomp. (from methanol), $[\alpha]_D^{26}$ + 17.0 ° (chloroform).

Preparation. Vinblastine and Vincristine occur in *Catharanthus roseus* (L). The plant contains only very small amounts of vinblastine and vincristine, and their isolation is correspondingly difficult [351]. First the alkaloids present are separated by extraction into alkaloid tartrate soluble in benzene and alkaloid tartrates insoluble in benzene. Vinblastine and vincristine belong to the first group. They then are separated by chromatography on aluminum oxide deactivated with acetic acid.

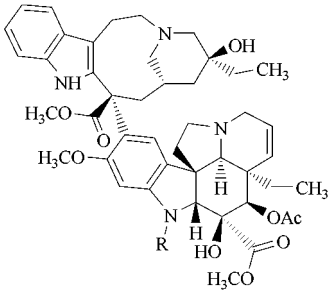

Ac = $COCH_3$
373 R = CH_3 Vinblastine
374 R = CHO Vincristine

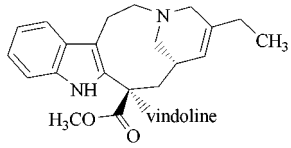

375 Dehydrovinblastine

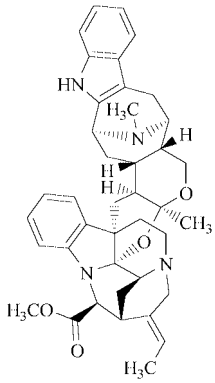

376 Villalstonine

Pharmacological Properties. Vinblastine (**373**) exhibits an antitumor activity and is also active against certain leukaemias. Vincristine (**374**) is active against leukaemias and causes either thrombocytosis with no effect upon the leucocytes or thrombocytopenia accompanied by peucopenia in rats. LD_{50} i.v. in mice: 17 mg/kg for vinblastine. LD_{50} i.p. in mice: 5.2 mg/kg for vincristine.

Therapeutic Use. Vinblastine: Hodgkin's disease and other lymphomas, chorioepithelial carcinoma.

Vincristine: childhood leukaemia.

Trade Names. Vinblastine: Velban, Velbe (Lilly), Exal (Shionogi), Vinblastin (Kyorin).

Vincristine: Oncovin (Lilly), Kyocristine (Kyorin), Vincosid (Leo).

11.3.5. Biogenetically Related Quinoline Alkaloids

Structure and Synthesis. These alkaloids are biogenetically derived from appropriate indole alkaloids. Chemical correlation between cinchona alkaloids and indole alkaloids, of the corynanthe group, were carried out based on the consideration of their biosynthetic pathway. KAMETANI reported the transformation of an indole ring into a quinoline ring involving a photooxygenation reaction [352]. The

use of this reaction also led to the biomimetic synthesis of camptothecine (**377**). Streptonigrin was produced by *Streptomyces flocculus,* and its structure was determined as **381** [353]. Total synthesis of streptonigrin was reported by Weinreb and Kende [354]. The structure of meloscine (**382**), characterized by a unique framework, was confirmed by X-ray analysis [355].

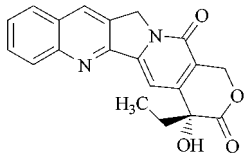

377 Camptothecine

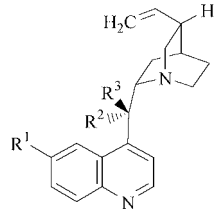

378 $R^1 = OCH_3$, $R^2 = H$, $R^3 = OH$ (−)-Quinine
379 $R^1 = H$, $R^3 = H$, $R^2 = OH$ (+)-Cinchonine
380 $R^1 = OCH_3$, $R^2 = OH$, $R^3 = H$ (+)-Quinidine

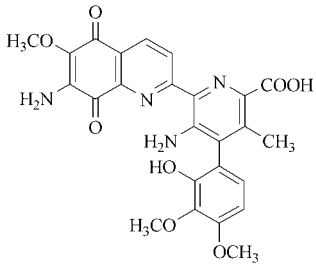

381 Streptonigrin

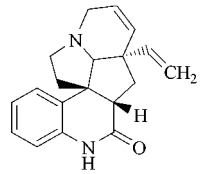

382 (+)-Meloscine

Physical Properties.

Camptothecine (**377**) [*7689-03-4*], $C_{20}H_{16}N_2O_4$, M_r 348: *mp* 264 – 267 °C decomp. (from methanol + acetonitrile), $[\alpha]_D^{25}$ + 31.3 ° (chloroform – methanol).

Quinine (**378**) [*130-95-0*], $C_{20}H_{24}N_2O_2$, M_r 324: *mp* 174.4 – 175.5 °C, $[\alpha]_D$ − 158 ° (ethanol). $C_{20}H_{24}N_2O_2 \cdot H_2O$ (from water): *mp* 57 °C.

Cinchonine (**379**) [*118-10-5*], $C_{19}H_{22}N_2O$, M_r 294: *mp* about 265 °C (from ethanol or ether), begins to sublime at 220 °C, $[\alpha]_D$ + 229 ° (ethanol).

Quinidine (**380**) [*56-54-2*], $C_{20}H_{24}N_2O_2$, M_r 324: *mp* 173.5 °C (from ether, ethanol, or water), $[\alpha]_D^{15}$ + 334.1 ° (in 0.1 N sulfuric acid).

Streptonigrin (**381**) [*3930-19-6*] $C_{25}H_{22}N_4O_8$, M_r 506: *mp* 275 °C decomp. (acetone or dioxane).

Dihydroquinidine, $C_{20}H_{26}N_2O_2$, M_r 326: *mp* 169.5 °C (from ether or ethanol), $[\alpha]_D^{15}$ − 299 ° (c = 0.025 M in 0.1 N sulfuric acid).

Preparation of Quinine. The finely powdered *cinchona* peel is treated with dissolved lime and 5 % sodium hydroxide solution and then extracted with aromatic hydrocarbons at 60 °C. The raw alkaloid is removed from the organic solvent by shaking with dilute sulfuric acid. The concentrated sulfuric acid-raw alkaloid solution is heated to its boiling point, and 10 % sodium hydroxide solution added until the solution is almost neutralized. Depending on the pH, the mono- or bisulfate crystallizes on cooling. Quinine is purified by recrystallization of either the mono- or bisulfate.

Pharmacological Properties. Quinine and cinchonine exhibit activity against *Plasmodium* species and are used as antimalarials. Quinine and its salts show bitter stomachic, analgesic, and antipyretic activities, whereas quinidine shows an antiarrhythmic activity. Camptothecin is suggested to have antitumor and antileukaemia activity. It prolonged life as much as 100 % when used in treatment of leukaemia L 1210 in mice. It was subjected to phase II clinical studies with gastrointestinal cancer; however, the study had to be dropped because of high toxicity. 10-Hydroxycamptothecin is expected to be more promising in treatment of certain tumors. Streptonigrin also exhibits antineoplasmic activity, but it causes severe and prolonged bone-marrow depression.

Therapeutic Use. Quinidine and Dihydroquinidine: antiarrhythmic agent.

Trade Names. Quinine-carboxy-3-salicylate: Hivernine (Vaillant), *Ethyl carbonate:* Chininum aethylcarbonicum (Buchler). *Quinine*

sulfate: Coco-Quinin (Lilly), Quine (Rowell). *Quinine hydrochloride:* Kinurea Terrial "H" (Dr. Dumesny). *Dihydrochloride:* Chininum dihydrochloricum (Buchler).

Quinidine Gluconate: Duraquin (Parke Davis). *Quinidine phenyl-S-ethyl-5-barbiturate:* Natisedine (Nativelle). *Quinidine-polygalacturonate:* Galactoquin (Mundipharma), Naticardina (Chinoin). *Quinidine sulfate:* Cin-Quin (Rowell), Optochinidin (Boehringer Mannheim), Quinidex LA (Robins), Novoquinidin (Novopharm), Systodin (Buchler). *Quinidine hydrogensulfate:* Kiditard (Delandale), Kinidin Durules (Astra), Optochinidin retard (Boehringer Mannheim), Quinidurule (Searle-Astra), Quinidex Extentabs (Robins).

Dihydroquinidine: Hydroquinidine "Houde", Serecor (Houde).

12. Lycopodium Alkaloids

Structure, Occurrence, and Synthesis. Since the structure determination of annotinine by WIESNER in 1957 [356], the chemistry of lycopodium alkaloids has progressed rapidly. The typical alkaloid found in the *Lycopodium* family is lycopodine. Its structure was assigned as **383** by MANSKE [357]. Several alkaloids of this group have been chemically related. For example, lycodine (**386**) was synthesized from lycopodine in four steps [358]. The structures of annotine (**384**) and serratinine (**385**) were confirmed by X-ray analysis [359]. Selagine [360], isolated from *Lycopodium selago,* was assigned as **387** on the basis of spectra and chemical degradation. Total synthesis of the *Lycopodium* alkaloids lycopodine, lycodine, and lycodoline was accomplished by HEATHCOCK [361].

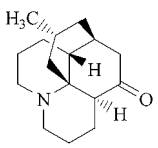

383 Lycopodine

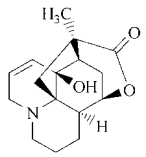

384 Annotine

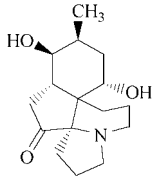

385 Serratinine

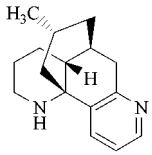

386 Lycodine

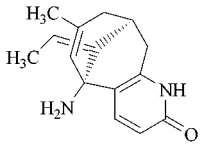

387 Selagine

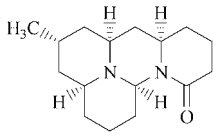

388 Cernuine

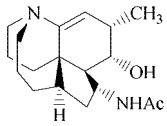

389 Serratinidine
Ac = $COCH_3$

Physical Properties.

Lycopodine (**383**) [*466-61-5*], $C_{16}H_{25}NO$, M_r 247: *mp* 114 – 115 °C, $[\alpha]_D$ – 24 ° (ethanol).

Lycodine (**386**) [*20316-18-1*], $C_{16}H_{22}N_2$, M_r 242: *mp* 118 – 119 °C, $[\alpha]_D$ – 10 ° (ethanol).

Selagine (**387**) [*116-28-9*], $C_{15}H_{18}N_2O$, M_r 242: *mp* 224 – 226 °C, $[\alpha]_D^{25}$ – 99 ° (methanol).

Pharmacological Properties. The *Lycopodium* alkaloids are moderately toxic. They seem to show a variety of activities including pressor activity, stimulation and contraction of the uterus, and paralysis. However, none of these have been useful in medicine.

13. Terpene Alkaloids

13.1. Monoterpene Alkaloids

Structure. The monoterpene family of alkaloids arises in the condensation of iridoids or secoiridoids with ammonia. Gentianine (**390**) is well known as the monoterpene pyridine alkaloid and is the most widely distributed. Synthesis of gentianine (**390**) was achieved by GOVINDACHARI [362]. Tecomine, isolated from *Tecoma* species, was assigned as **391** by JONES [363]. A synthesis of tecomine (**391**) was achieved by HANAOKA [364].

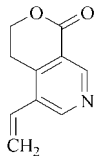

390 Gentianine

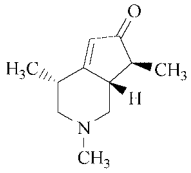

391 Tecomine (Tecomanine)

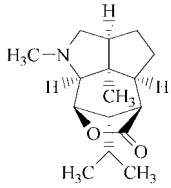

392 Dendrobine

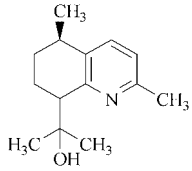

393 Fabianine

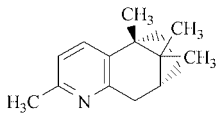

394 (−)-Patchoulipyridine

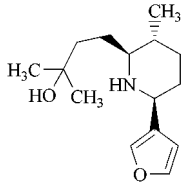

395 (−)-Nupharamine

Physical Properties.

Gentianine (**390**) [*439-89-4*], $C_{10}H_9NO_2$, M_r 175: *mp* 82 – 83 °C (from ether).

Tecomanine (**391**) [*6878-83-7*], $C_{11}H_{17}NO$, M_r 179: *bp* 125 °C (at 0.01 kPa), $[\alpha]_D^{24}$ − 175 ° (chloroform).

Pharmacological Properties. Gentianine has a low toxicity and stimulates the central nervous system. In large doses it can paralyze. Gentianine also may exert anti-inflammatory, hypotensive, and muscular-relaxant activities. The leaves of various *Tecoma* species are used in Mexico by natives for the control of diabetes. Tecomine citrate shows hypoglycemic activity.

13.2. Sesquiterpene Alkaloids

Structure and Occurrence. Dendrobine (**392**) was isolated from genus *Dendrobium.* Its structure was deduced to be **392** on the basis of spectroscopic evidence and chemical correlations [365]. The biosynthesis of dendrobine is consistant with its derivation from mevalonate via farnesol, germacrane, and cadalane skeletons. Synthesis of **392** was first accomplished by INUBUSHI [366]. Furthermore, formal total synthesis of (−)-dendrobine was reported by SHIBASAKI [367]. Fabianine (**393**) and patchoulipyridine (**394**) were derived biosynthetically from the sesquiterpenes. (−)-Patchoulipyridine (**394**) was made from patchouli alcohol [368], [369]. (−)-Nupharamine (**395**) and other related bases are distributed in *Nuphar* species. The synthesis of (±)-fabianine (**393**) and (−)-patchoulipyridine (**394**) has been reported [370], [371].

Physical Properties.

Dendrobine (**392**) [*2115-91-5*], $C_{16}H_{25}NO_2$, M_r 263: *fp* 134 °C (sublime), $[\alpha]_D^{16}$ − 51.5 ° (ethanol).

Fabianine (**393**) [*6871-51-8*], $C_{14}H_{21}NO$, M_r 219: *bp* 74 °C (at 6.7 Pa), $[\alpha]_D < 1$ °.

Patchoulipyridine (**394**) [*6517-97-1*], $C_{15}H_{21}N$, M_r 215: *bp* ca. 80 °C (at 13.3 Pa), *mp* 24 – 26.5 °C (white waxy needles), $[\alpha]_D^{28}$ − 31.3 ° (ethanol).

(−)-*Nupharamine* (**395**) [*17812-38-3*], $C_{15}H_{25}NO_2$, M_r 251: *bp* 130 – 134 °C, $[\alpha]_D^{22}$ − 35.04 ° (chloroform).

Pharmacology. Dendrobine reduces blood pressure and depresses respiration. In large doses it also diminishes cardiac activity and brings about moderate hyperglycemia. It shows weak analgesic and antipyretic activities.

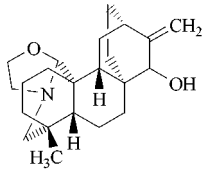

396 Atisine

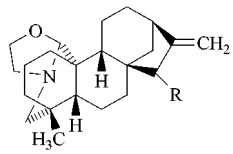

397 R = α-OH Veatchine
398 R = β-OH Garryfoline

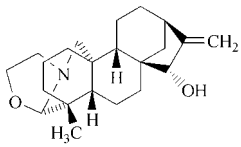

399 Garryine

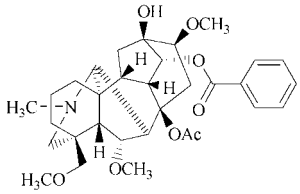

400 Delphinine

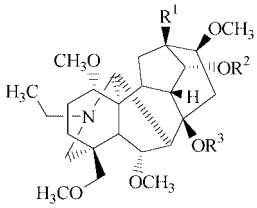

401 $R^1 = H$, $R^2 = H$, $R^3 = H$ Chasmine
402 $R^1 = OH$, $R^2 = COC_6H_5$, $R^3 = COCH_3$ Aconitine

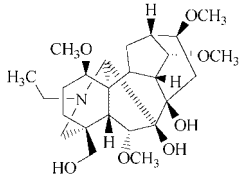

403 (+)-Lycoctonine

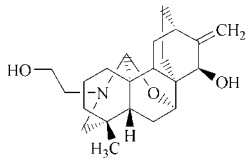

404 Ajaconine

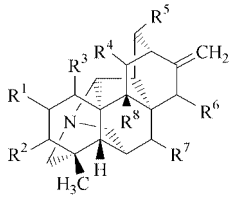

405 $R^1 = R^4 = \beta$-OH, $R^3 = OH$, $R^2 = R^3 = R^6 = R^7 = R^8 = H$ Hetisine
406 $R^1 = R^6 = R^7 = \beta$-OH, $R^5 = OH$, $R^2 = OCOC_6H_5$, $R^3 = R^4 = R^8 = H$ Ignavine
407 $R^1 = \alpha$-OH, $R^2 = R^4 = R^5 = R^7 = H$, $R^3 = R^6 = R^8 = \beta$-OH Hypognavinol
408 $R^1 = R^2 = R^3 = R^5 = R^7 = R^8 = H$, $R^4 = R^6 = OH$ Kobusine

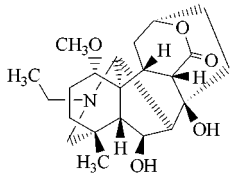

409 Heteratisine

13.3. Diterpene Alkaloids

Structure, Occurrence, and Synthesis. Diterpene alkaloids, isolated from plants of the *Delphinium* and *Aconitum* genera (Ranunculaceae) and *Inula royleana* (Compositae), are of interest because of their physiological activities and their unique carbon framework. These alkaloids are possibly derived biogenetically from tetracyclic or pentacyclic diterpenes in which the nitrogen atom of β-aminoethanol, methylamine, or ethylamine forms a substituted piperidine ring.

Atisine (**396**) is the major constituent of several *Aconitum* species. The conversion of the atisane skeleton to the aconitine (**402**) skeleton was achieved [372] along the commonly accepted biogenetic route for aconitine-type alkaloids. Veatchine (**397**) and garryine (**399**) have a kaurene skeleton, and they are the positional isomers of an aminoethanol unit. Synthesis of these last three alkaloids was accomplished by NAGATA [373], whereas the efficient synthesis of the key intermediate for these alkaloids was achieved by KAMETANI [374]. Delphinine (**400**), chasmanine (**401**), and aconitine (**402**) are grouped into the aconitine type, which has no oxygen functional group or substituent at the C-7 position, whereas the lycoctonine type bears oxygen functional groups at the C-6 and C-7 positions. The structure of the alkaloids (**404**) – (**408**) resembles that of atisine (**396**), and these are placed in the atisane group. The alkaloids in the heteratisine (**409**) group always possess a lactone moiety. New alkaloids have been found in *Delphinium* species [375], [376].

Physical Properties.

Atisine (**396**) [*466-43-3*], $C_{22}H_{33}NO_2$, M_r 343: *mp* 57 – 60 °C. Hydrochloride: *mp* 311 – 312 °C decomp. (from dilute ethanol), $[\alpha]_D^{25}$ + 28 ° (ethanol).

Veatchine (**397**) [*76-53-9*], $C_{22}H_{33}NO_2$, M_r 343: *mp* 119 – 120 °C (from dilute acetone), $[\alpha]_D^{27.5}$ – 69.01 ° (ethanol).

Garryine (**399**) [*561-51-3*], $C_{22}H_{33}NO_2$, M_r 343: *mp* 74 – 82 °C (from dilute acetone), $[\alpha]_D^{27.5}$ – 84.23 ° (ethanol).

Delphinine (**400**) [*561-07-9*], $C_{33}H_{45}NO_9$, M_r 599: *mp* 197.5 – 199 °C (from ethanol), $[\alpha]_D^{20}$ + 25 ° (ethanol).

Aconitine (**402**) [*302-27-2*], $C_{34}H_{47}NO_{11}$, M_r 645: mp 204 °C, $[\alpha]_D$ + 17.3 ° (chloroform).

Lycoctonine (**403**) [*26000-17-9*], $C_{25}H_{41}NO_7$, M_r 467: *mp* 143 °C, $[\alpha]_D^{20}$ + 53 ° (ethanol).

Ajaconine (**404**) [*545-61-9*], $C_{22}H_{33}NO_3$, M_r 359: *mp* 172 °C (from dilute ethanol), $[\alpha]_D^{18}$ – 119 ° (ethanol).

Pharmacological Properties. Aconitine, delphinine, and related alkaloids irritate the skin and cause death due to respiratory failure and cardiac and vasomotor damage. Atisine is relatively nontoxic in comparison to aconitine.

14. Steroidal Alkaloids

14.1. Alkaloids of the Apocynaceae

Conessine was found in several *Holarrhena* species. Its structure was deduced to be **410** by chemical conversion to 3β-acetoxy-20α-aminopregn-5-ene [377]. Irehdiamine A (**411**), was isolated from *Funtumia elastica* [378]. The structure of conanine (**412**) was confirmed by X-ray analysis of its *N*-oxide derivative [379]. A formal total synthesis of (+)-connecine (**410**) was reported by MEYERS [380].

Conessine (**410**) $C_{24}H_{40}N_2$, M_r 356: *mp* 127 – 128.5 °C (from acetone), $[\alpha]_D^{20}$ +25.3° (ethanol).

Although conessine produces local anaesthesia, it causes local necrosis on subcutaneous injection. This alkaloid inhibits the growth of *Mycobacterium tuberculosis* in vitro. Conessine is used mainly as a remedy for amoebic dysentery.

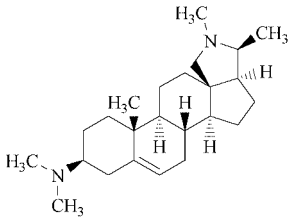

410 (+)-Conessine

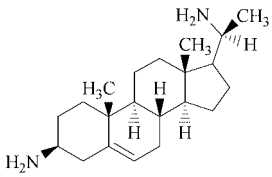

411 Irehdiamine A

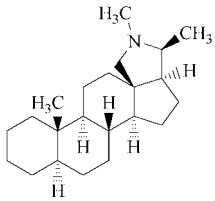

412 Conanine

14.2. Buxus Alkaloids

The buxus family of alkaloids bears a cyclopropane ring on the B-ring or a seven-membered B-ring. Cyclobuxine D (**413**), isolated from *Buxus* species, was chemically correlated to cycloeucalenol (**414**) [381]. The structure of buxozine C (**415**) was determined in spectroscopic studies [382].

Cyclobuxine D (**413**) $C_{25}H_{42}N_2O$, M_r 386: *mp* 245 – 247 °C (decomp.), $[\alpha]_D^{23}$ +98° (chloroform).

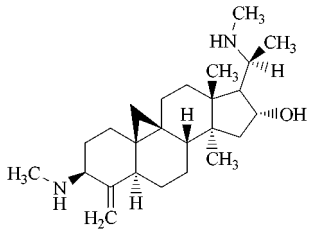

413 Cyclobuxine D

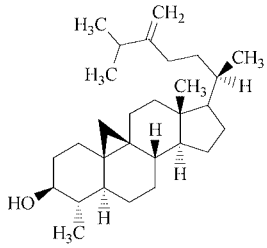

414 Cycloeucalenol

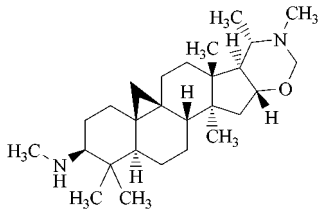

415 Buxozine C

14.3. Alkaloids of the Asclepiadaceae

Stephanthraniline A (**416**), isolated from *Stephanotis japonica,* has an *N*-methylanthraniloyl ester moiety [383]. Other steroidal alkaloids are found in *Stephanotis* species [384]. Its

base-catalyzed hydrolysis afforded the known pregnane derivative sarcostin.

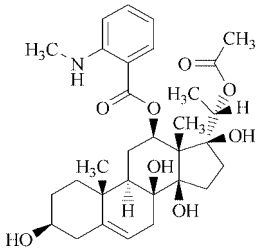

416 Stephanthraniline A

Stephanthraniline A (**416**) [*65429-40-5*] $C_{31}H_{43}O_8$, M_r 557: *mp* 170–173 °C (from acetone–hexane), $[\alpha]_D$ +17.9 ° (chloroform).

14.4. Solanum Alkaloids

Solasodine (**417**) is distributed in *Solanum* species. It is becoming an important starting material for the synthesis of steroids in the pharmaceutical industry [385]. Some of the *Solanum* alkaloids exhibit antibiotic effects on certain fungi and bacteria.

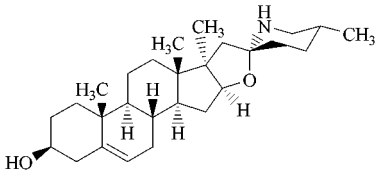

417 Solasodine

Solasodine (**417**) [*126-17-0*], $C_{27}H_{43}NO_2$, M_r 413: *mp* 200–202 °C (from methanol), $[\alpha]_D^{25}$ −98 ° (methanol), $[\alpha]_D$ −113 ° (chloroform).

Solasodine was shown to possess cardiotonic and antiphlogistic effects on mice and some effects on vascular permeability and the central nervous system.

14.5. Veratrum and Fritillaria Alkaloids

Jervine (**418**) and verticine (**419**) have C-nor-D homosteroidal linkage and were synthesized from hecogenin [386], [387].

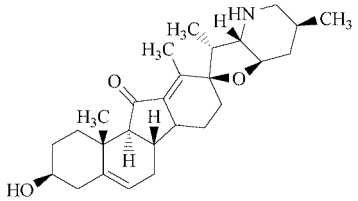

418 (−)-Jervine

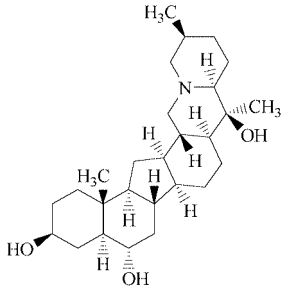

419 Verticine

Jervine (**418**) [*469-59-0*], $C_{27}H_{39}NO_3$, M_r 425: *mp* 243.5–244.5 °C /378.(from dilute methanol), $[\alpha]_D^{20}$ −150 ° (ethanol), $[\alpha]_D^{20}$ −167.6 ° (chloroform).

Verticine (**419**) [*23496-41-5*], $C_{27}H_{45}NO_3$, M_r 431: *mp* 223–224 °C (from ethanol), $[\alpha]_D^{16}$ −19.4 ° (ethanol), $[\alpha]_D^{17}$ −20 ° (chloroform).

Jervine was suggested to cause antiaccelerator cardiac action. In large doses it produces vasodilatation and a fall in blood pressure.

14.6. Phyllobates Alkaloids

Batrachotoxin (**420**) and batrachotixin A (**421**) were isolated from the skin of neotropical poison-dart frogs, genus *Phyllobates* [388], and their structures were established by spectroscopic and chemical studies. The former alkaloid is the most potent cardiotoxin known. The absolute configuration of batrachotoxinin A was determined by X-ray analysis [389]. The LD_{50} value of batrachotoxin is 2 µg/kg (s.c. in mice), while that of batrachotoxinin A is 1000 µg/kg (s.c. in mice). Toxic effect is caused by selective and irreversible increase in permeability of membranes to Na^+ ions. Several reports on synthetic studies have appeared [390–392].

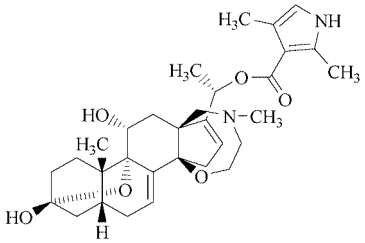

420 Batrachotoxin

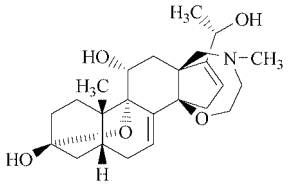

421 Batrachotoxinin A

Physical Properties

Batrachotixin (**420**) [*23509-16-2*], $C_{31}H_{42}N_2O_6$, M_r 538: $[\alpha]_D^{24}$ $-5°$ to $-10°$; $[\alpha]_{300}^{24}$ $-260°$ ($c = 0.23$ in methanol).

15. Miscellaneous Alkaloids

15.1. Muscarine Alkaloids

Muscarine (**422**) is found in the fly mushroom, *Amanita muscaria,* and is a hallucinogenic choline analogue. Its structure was confirmed by X-ray analysis [393]. Muscarine is a parasympathomimetic drug. L-Arabinose was converted into muscarine by EUGSTER [394]. Total synthesis of (+)-muscarine was accomplished by ADAMS [395].

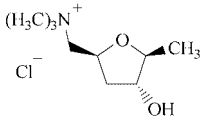

422 (+)-Muscarine

Muscarine (**422**) [*300-54-9*], $C_9H_{20}ClNO_2$ M_r 210: *mp* 180 – 181 °C (from ethanol), $[\alpha]_D^{25}$ +8.1° (ethanol). LD_{50} i.v. in mice: 0.23 mg/kg.

15.2. Imidazole Alkaloids

Anantine (**423**) is derived from cinnamic acid and histamine, which arises from histidine by decarboxylation. Histamine (**425**) occurs widely in nature as a result of putrefactive processes. The flavoprotein diamine oxidase converts histamine to the aldehyde and ammonia. Its phosphate is used as a diagnostic indicator for gastric secretion or pheochromocytoma. Anantine and related alkaloids were also reported [396]. Furthermore, the synthesis of (±)-anantine and (±)-isoanantine was achieved by NINOMIYA [397].

(S)-(E)-(−)-Anantine (**423**) [*50656-82-1*], $C_{15}H_{15}N_3O$, M_r 253: *fp* 204 °C, $[\alpha]_D$ $-549°$.

(S)-(E)-(−)-Isoanantine (**424**), $C_{15}H_{15}N_3O$, M_r 253: *fp* 190 °C (from acetone), $[\alpha]_D$ $-347°$.

Histamine (**425**) [*51-45-6*], $C_5H_9N_3$, M_r 111: *mp* 83 – 84 °C (from chloroform), *bp* 209 – 210 °C (at 2.4 kPa). LD_{50} i.p. in mice: 2020 mg/kg.

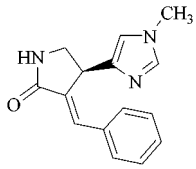

423 Anantine

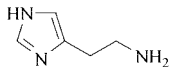

425 Histamine

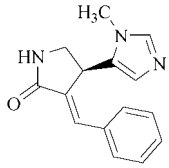

424 Isonatine

15.3. Polyamine Alkaloids

The family of polyamine alkaloids exhibits a variety of biological effects: antiviral, antitumor, and blood-pressure depressing activities. Maytansine is isolated from *Maytenus ovatus* Loes. Its structure was assigned as **426** by KUPCHAN [398]. Maytansine shows cytotoxic action, and it is expected to be an anticancer agent. The synthesis of (−)-maytansine and (−)-maytansinol was reported by COREY [399] and ISOBE [400]. The chemistry of spermine and spermidine alkaloids such as codonocarpine (**427**) and lunarine (**429**) is reviewed in [401]. Also isocodonocarpine (**428**) was isolated from the root bark of *Capparis decidua* [402]. The synthesis of condocarpine and (±)-lunarine was achieved by NAGAO [403], [404].

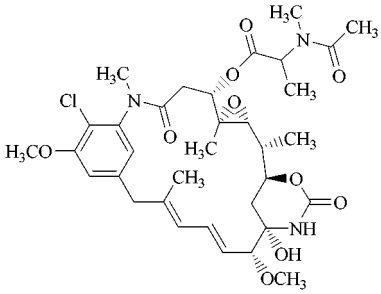

426 Maytansine

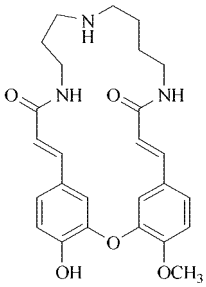

427 Codonocarpine

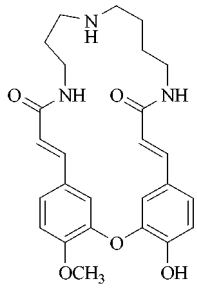

428 Iscadonocarpine

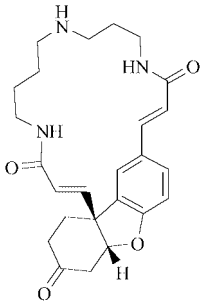

429 (+)-Lunarine

Maytansine (**426**) [*35846-53-8*], $C_{34}H_{46}ClN_3O_{10}$, M_r 692: *mp* 171 – 172 °C, $[\alpha]_D^{26}$ – 145 ° (chloroform).

(E,E)-codonocarpine (**427**) [*33442-48-7*], $C_{26}H_{31}N_3O_5$, M_r 465: *mp* 187 °C.

(+)-*Lunarine* (**429**) [*24185-51-1*], $C_{25}H_{31}N_3O_4$, M_r 437: *fp* 238 – 240 °C, $[\alpha]_D$ + 344 ° (ethanol).

15.4. Securinine Alkaloids

Securinine (**430**) is isolated from *Securinega suffructicosa* Rehd. Its structure was confirmed by X-ray analysis [405]. A report on stereoisomers of securine has appeared [406]. A synthesis was reported by Horii [407]. (–)-Securinine and its derivatives were synthesized by Wanner [408].

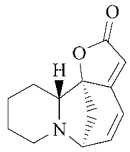

430 Securinine

Securinine (**430**) [*5610-40-2*], $C_{13}H_{15}NO_2$, M_r 217: *mp* 142 – 143 °C (from ethanol), $[\alpha]_D^{20}$ – 1042 ° (ethanol).

15.5. Tetrodotoxin

Tetrodotoxin (**431**) is identical with tarichatoxin, the poison of the California salamander *Taricha torosa* [409]. Tetrodotoxin is obtained from the ovaries of suitable fish after precipitating the protein by chromatography on active charcoal. Usually, 8 – 9 g of pure tetrodotoxin are obtained from 1000 g of roe. The structure of the compound was established by chemical, spectroscopic, and X-ray investigations. *rac*-Tetrodotoxin was synthesized [410]. The synthesis of optically active tetrodotoxine is reviewed in [411]. Tetrodotoxin is extremely poisonous, having a LD_{50} value with mice of 8 – 10 µg/kg. Consumption of poisonous fish causes paralysis of the respiratory center and respiratory muscles, leading to paresthesia, motor paralysis, and after 6 – 8 h death. Tetrodotoxin acts on the cell membranes such that the flow of Na^+ ions into the cells is hindered.

Tetrodotoxin (**431**) [*4368-28-9*], $C_{11}H_{17}N_3O_8$, M_r 319: *mp* > 200 °C, $[\alpha]_D^{25}$ – 8.64 ° (dilute acetic cid).

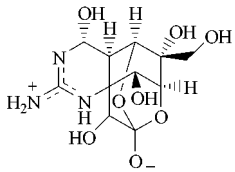

431 Tetrodotoxin

15.6. Aaptamine and Related Alkaloids

Aaptamine (**432**), demethylaaptamine (**433**), and demethyloxyaaptamine (**434**) were isolated from the marine sponge *Aaptos aap-*

tos [412]. The structures, with a new heterocylic system, were determined from spectroscopic evidence and by chemical degradation, and the biological activity of the alkaloids was evaluated. Aaptamine possesses potent α-adrenorecepter blocking activity on vascular smooth muscle. Demethyloxyaaptamine is cytotoxic against Hela cells (ED_{50} 0.87 µg/mL) and has potent antimicrobial activities against gram-positive and -negative bacteria such as *Staphylococcus aureus* [minimum inhibitory concentration (MIC) 3.13 mg/mL], *Bacillus subtilis* (MIC 6.25 mg/mL) and *protecus vulgaris* (MIC 12.5 mg/mL). The biological activity of **432** is weaker (about half) than that of **434**. Several reports concerned with total syntheses of aaptamine have appeared [413–421]. A synthesis of demethyloxyaaptamine was also reported [414].

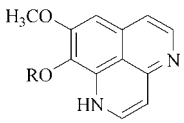

432 R = CH_3 Aaptamine
433 R = H Demethylaaptamine

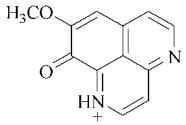

434 Demethyloxyaaptamine

Physical Properties.

Aaptamine (**432**) [*85547-22-4*], $C_{13}H_{12}N_2O_2$, M_r 228: *mp* 110 – 113 °C (hydrochloride).

Demethylaaptamine (**433**) [*88839-98-9*], $C_{12}H_{10}N_2O_2$, M_r 214: *mp* 248 – 251 °C (hydrochloride).

Demethyloxyaaptamine (**434**) [*88839-99-0*], $C_{12}H_8N_2O_2$, M_r 212: *mp* 210 – 212 °C (hydrochloride).

15.7. Eudistomidins

The alkaloids possessing a β-carboline skeleton eudistomidin A – F (**435** – **440**) were isolated from the Okinawan tunicate *Eudistoma glaucus* [422]. The structures were determined on the basis of spectroscopic evidence. The stereochemistry of eudistomidin F was not established.

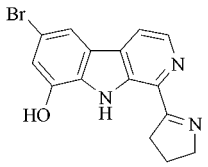

435 Eudistomidin A

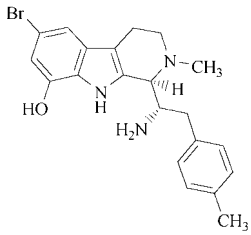

436 Eudistomidin B

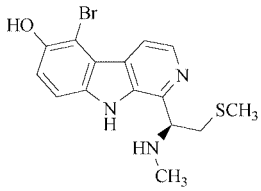

437 Eudistomidin C

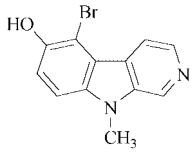

438 Eudistomidin D

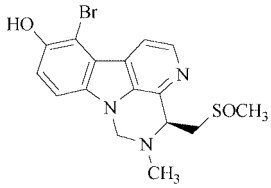

439 Eudistomidin E

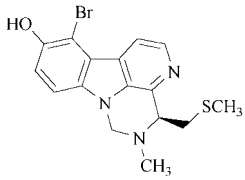

440 Eudistomidin F

Physical Properties.

Eudistomidin A (**435**) [*102673-53-0*], $C_{15}H_{12}BrN_3O$, M_r 330: *mp* 225 – 230 °C .

Eudistomidin B (**436**) [*125443-66-5*], $C_{21}H_{24}BrN_3O$, M_r 398: *mp* 81 – 83 °C, $[\alpha]_D^{22}$ – 54 °C (methanol), $[\alpha]_D$ – 76.4 °C (chloroform).

Eudistomidin C (**437**) [*125422-17-5*], $C_{15}H_{16}BrN_3OS$, M_r 366: *mp* 120 – 122 °C, $[\alpha]_D$ + 15.6 °C (methanol).

Eudistomidin D (**438**) [*125422-19-7*], $C_{12}H_9BrN_2O$, M_r 277: *mp* 180 °C.

Eudistomidin E (**439**) [*136094-35-4*], $C_{16}H_{16}BrN_3O_2S$, M_r 394: amorphous solid, $[\alpha]_D$ +19 °C (methanol)).

Eudistomidin F (**440**) [*136094-36-5*], $C_{16}H_{16}BrN_3OS$, M_r 378: amorphous solid, $[\alpha]_D$ +16 °C (methanol)).

Pharmacological Properties. Eudistomidin A is the first inhibitor of calmodulin-activated brain phosphodiesterase (2×10^{-5} mol/L) of marine origin. Eudistomidins B–D exhibit potent cytotoxic activity against murine leukemia L1210 (IC_{50} 3.4 µg/mL for B, 0.36 µg/mL for C, and 2.4 µg/mL for D) and L1210 (IC_{50} 3.1 µg/mL for B, 0.42 µg/mL for C, and 18 µg/mL for D). Eudistomidin B also activates rabbit heart muscle actomyosin ATPase (93 % at 3×10^{-5} mol/L). Eudistomidin C exhibits cytotoxic and calmodulin antagonistic activity ($IC_{50} = 3 \times 10^{-5}$ mol/L). Eudistomidin D exhibits cytotoxic activity and induces Ca^{2+} release from the sarcoplasmic reticulum about ten times more potently than caffein.

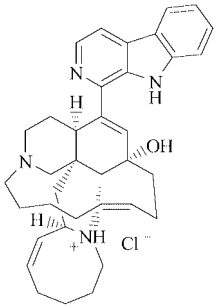

441 Manzamine A

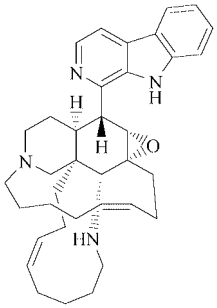

442 Manzamine B

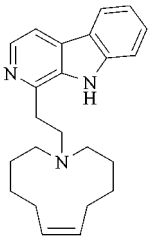

443 Manzamine C

15.8. Manzamines

Manzamines A–C (**441**)–(**443**) were isolated from the Okinawan marine sponges *Haliclona* sp. and *Pellina* sp. The structures were determined on the basis of spectroscopic evidence and X-ray analysis [423]. Manzamine A exhibits antitumor activity against P-388 leukemia cells (IC_{50} 0.07 µg/mL). IC_{50} 6 µg/mL for mansamine B, 3 µg/mL for manzamine C.

Physical Properties.

Manzamine A (**441**) [*104196-68-1*], $C_{36}H_{44}N_4O$, M_r 548: *mp* >240 °C, $[\alpha]_D$ +50 °C (chloroform).

Manzamine B (**442**) [*112663-92-0*], $C_{36}H_{44}N_4O$, M_r 550: *mp* 198–203 °C, $[\alpha]_D$ +89 °C (chloroform).

Manzamine C (**443**) [*112693-24-0*], $C_{23}H_{29}N_3$, M_r 547: *mp* 77–82 °C.

15.9. Panuamine and Haliclonadiamine

Papuamine (**444**) and haliclonadiamine (**445**) were isolated from a marine sponge *Haliclona* sp. [424], [425]. The structures were elucidated by spectroscopic studies. The synthesis of (+)-papuamine (**444**) and (−)-haliclonadiamine (**445**) were reported by Bartrett [426] and Heathcock [427]. Papuamine shows antifungal activity, and the alkaloids inhibit the growth of *Candida albicans, Bacillus substilis,* and *Staphylococcus aureus* (1.0–5.0 µg/mL).

Physical Properties

Papuamine (**444**) [*112455-84-2*], $C_{25}H_{40}N_2$, M_r 368: *mp* 167.5–169 °C, $[\alpha]_D$ −150° (methanol).

Haliclonadiamine (**445**) [*117065-24-4*], $C_{25}H_{40}N_2$, M_r 368: oil, $[\alpha]_D$ −18.2° (methanol).

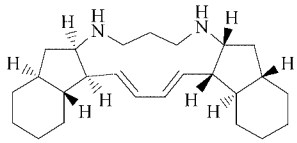

444 Papuamine

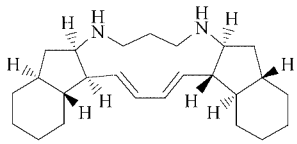

445 Haliclonadiamine

16. References

1. E. K. Nelson, L. E. Cawson, *J. Am. Chem. Soc.* **45** (1923) 2179. L. Crombie, S. H. Dandegaonker, K. B. Simpson, *J. Chem. Soc.* 1955, 1025.
2. E. Späth, *Monatsh. Chem.* **40** (1919) 129.
3. F. E. Rymill, C. A. McDonald, *Q. J. Pharm. Pharmacol.* **10** (1937) 463.
4. Y. Inubushi, Y. Tsuda, T. Kunita, S. Matsumoto, *Chem. Pharm. Bull.* **12** (1964) 749.
5. S. Murakami, T. Takemoto, Z. Shimizu, *Yakugaku Zasshi* **73** (1953) 1026.
6. G. Impellizzeri, S. Mangiafico, G. Oriente, M. Piatelli, S. Sciuto, E. Fattorusso, S. Magno, C. Santacroce, D. Sica, *Phytochemistry* **14** (1975) 1549.
7. S. Murakami, T. Takemoto, Z. Tei, K. Daigo, *Yakugaku Zasshi* **75** (1975) 866, 869.
8. H. Watase, Y. Tomiie, I. Nitta, *Bull. Chem. Soc. Jpn.* **31** (1958) 714.
9. W. Oppolzer, K. Thirring, *J. Am. Chem. Soc.* **104** (1982) 4978.
10. T. Takemoto, K. Daigo, Y. Kondo, K. Kondo, *Yakugaku Zasshi* **86** (1966) 874.
11. Y. Ohfune, M. Tomita, *J. Am. Chem. Soc.* **104** (1982) 3511.
12. B. Witkop Experientia 27 (1971) 1121. J. W. Daley, I. L. Karle, W. Meyers, T. Tokuyama, J. A. Walters, B. Witkop, *Proc. Natl. Acad. Sci. USA.* **68** (1971) 1870. T. Tokuyama, K. Uenoyama, G. Brown, J. W. Daly, B. Witkop, *Helv. Chim. Acta* **57** (1974) 2597. I. L. Karle, *J. Am. Chem. Soc.* **95** (1973) 4036. J. W. Daly, T. Tokuyama, T. Fujiwara, R. J. Highet, I. L. Karle, *J. Am. Chem. Soc.* **102** (1980) 830.
13. J. Elliot, M. A. Raftery, *Biochemistry* **18** (1979) 1968.
14. J. Kobayashi, T. Maruyama, Y. Ohizumi, *Tetrahedron Lett.* **30** (1989) 4833.
15. R. Robinson, *J. Chem. Soc.* **111** (1917) 762. C. Schöpf, G. Lehmann, *Justus Liebigs Ann. Chem.* **518** (1935) 1.
16. S. Danishefsky, R. McKee, R. K. Singh, *J. Am. Chem. Soc.* **99** (1977) 7711.
17. J. J. Tufariello, G. E. Lee, *J. Am. Chem. Soc.* **102** (1980) 373.
18. G. A. Kraus, K. Neuenschwander, *Tetrahedron Lett.* **21** (1980) 3841.
19. N. J. Leonard, S. W. Blum, *J. Am. Chem. Soc.* **82** (1960) 503.
20. S. Takano, N. Ogawa, K. Ogasawara, *Heterocycles* **16** (1981) 915.
21. F. L. Warren, *Fortschr. Chem. Org. Naturst.* **12** (1955) 198; **24** (1966) 329. R. V. Stevens in J. ApSimon (ed.): *The Total Synthesis of Natural Products,* **vol. 3,** J. Wiley & Sons, New York 1977, p. 439.
22. T. Furuya, M. Hikichi, *Yuki Gosei Kagaku Kyokaishi* **35** (1977) 653.
23. T. Ohsawa, M. Ihara, K. Fukumoto, T. Kametani, *Heterocycles* **19** (1982) 2075.
24. A. R. Mattocks, *Nature (London)* **232** (1971) 475.
25. G. W. J. Fleet, M. Haraldsson, R. J. Nash, L. E. Fellows, *Tetrahedron Lett.* **29** (1988) 5441.
26. R. J. Nash et al., *Tetrahedron Lett.* **29** (1988) 2487.
27. A. E. Wick, P. A. Bartlett, D. Dolphin, *Helv. Chim. Acta* **54** (1971) 513.
28. A. M. Dawidar, F. Winternitz, S. R. Johns, *Tetrahedron* **33** (1977) 1733.
29. T. R. Govindachari, M. V. Lakshmikantham, S. Rajadurai, *Tetrahedron* **14** (1961) 284.
30. R. B. Herbert, C. J. Moody, *J. Chem. Soc. D* 1970, 121.
31. B. Chauncy, E. Gellert, *Aust. J. Chem.* **23** (1970) 2503.
32. A. J. Liepa, R. E. Summons, *J. Chem. Soc. Chem. Commun.* 1977, 826.
33. S. M. Weinreb, N. A. Khatri, J. Shringarpure, *J. Am. Chem. Soc.* **101** (1979) 5073.
34. N. A. Khatri, H. F. Schmitthenner, J. Shringarpure, S. M. Weinreb, *J. Am. Chem. Soc.* **103** (1981) 6387.
35. H. Iida, M. Tanaka, C. Kibayashi, *J. Chem. Soc. Chem. Commun.* 1983, 271.
36. S. R. Johns, J. A. Lamberton, A. A. Sioumis, R. I. Willing, *Aust. J. Chem.* **22** (1969) 775.
37. J. J. Tufariello, S. A. Ali, *Tetrahedron Lett.* 1979, 4445.
38. H. Otomasu, N. Takatsu, T. Honda, T. Kametani, *Heterocycles* **19** (1982) 511.
39. L. E. Overman, K. L. Bell, *J. Am. Chem. Soc.* **103** (1981) 1851.
40. J. W. Daly, B. Witkop, T. Tokuyama, T. Nishikawa, I. L. Karle, *Helv. Chim. Acta* **60** (1977) 1128.
41. R. Fujimoto, Y. Kishi, J. F. Blount, *J. Am. Chem. Soc.* **102** (1980) 7154.
42. R. Fujimoto, Y. Kishi, *Tetrahedron Lett.* 1981, 4197.
43. D. J. Hart, K. Kanai, *J. Am. Chem. Soc.* **105** (1983) 1255.
44. M. A. Maleque, E. X. Albuquerque, J. E. Warnick, J. W. Daly, Y. Nimitkitpaisan, *Fed. Proc. Fed. Am. Soc. Exp. Biol* **38** (1979) 1399.
45. D. P. Rainey, E. B. Smalley, M. H. Crump, F. M. Strong, *Nature (London)* **205** (1965) 203.

46. R. A. Gardiner, K. L. Rinehart, J. J. Snyder, H. P. Broquist, *J. Am. Chem. Soc.* **90** (1968) 5639.
47. D. Cartwright, R. A. Gardiner, K. L. Rinehart, *J. Am. Chem. Soc.* **92** (1970) 7615.
48. W. J. Gensler, M. W. Hu, *J. Org. Chem.* **38** (1973) 3848.
49. R. A. Gobao, M. L. Bremmer, S. M. Weinreb, *J. Am. Chem. Soc.* **104** (1982) 7065.
50. F. P. Guengerich, S. D. Aust, *Mol. Pharmacol.* **13** (1977) 185.
51. L. D. Hohenschutz et al., *Phytochemistry* **20** (1981) 811.
52. R. C. Bernotas, B. Ganem, *Tetrahedron Lett.* **25** (1984) 165. H. Hamana, N. Ikota, B. Ganem, *J. Org. Chem.* **52** (1987) 5492.
53. K. Burgess, I. Henderson, *Tetrahedron* **48** (1992) 4045, and references cited therein.
54. H. Zhao, D. R. Mootoo, *J. Org. Chem.* **61** (1996) 6762, and references cited therein.
55. E. E. van Tamelen, R. L. Foltz, *J. Am. Chem. Soc.* **91** (1969) 7372.
56. C. Schöpf, G. Benz, F. Dürsch, W. Burkhardt, T. Rokohl, *Justus Liebigs Ann. Chem.* **755** (1972) 86.
57. M. Hanaoka, N. Ogawa, Y. Arata, *Chem. Pharm. Bull.* **22** (1974) 973; **24** (1976) 1045.
58. D. J. Hart, K. Kanai, *J. Org. Chem.* **47** (1982) 1555.
59. K. Shishido, K. Tanaka, K. Fukumoto, T. Kametani, *Tetrahedron Lett.* **24** (1983) 2783.
60. J. W. Daley, T. Tokuyama, G. Habermehl, I. L. Karle, B. Witkop, *Justus Liebigs Ann. Chem.* **729** (1969) 198.
61. Y. Inubushi, T. Ibuka, *Heterocycles* **8** (1977) 633.
62. G. Habermehl, H. Andres, K. Miyahara, B. Witkop, J. W. Daley, *Justus Liebigs Ann. Chem.* 1976, 1577.
63. K. Maruoka, T. Miyazaki, M. Ando, Y. Matsumura, S. Sakane, K. Hattori, H. Yamamoto, *J. Am. Chem. Soc.* **105** (1983) 2831.
64. T. Kametani, T. Higa, C. V. Loc, M. Ihara, M. Koizumi, K. Fukumoto, *J. Am. Chem. Soc.* **98** (1976) 6186; **99** (1977) 2306.
65. T. Kametani, C. V. Loc, T. Higa, M. Ihara, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1977, 2347.
66. J. Hlubucek, E. Ritchie, W. C. Taylor, *Chem. Ind.* (London) 1969, 1809.
67. T. Kametani: *The Chemistry of the Isoquinoline Alkaloids,* Hirokawa Publ. Co. Inc., Tokyo, Elsevier Publ. Co., Amsterdam 1968.
68. T. Kametani: *The Chemistry of the Isoquinoline Alkaloids,* **vol. 2,** The Sendai Institute of Heterocyclic Chemistry, Sendai, Japan 1974.
69. T. Kametani in J. ApSimon (ed.): *The Total Synthesis of Natural Products,* **vol. 3,** J. Wiley & Sons, New York 1977.
70. W. I. Taylor, A. R. Battersby: *Oxidative Coupling of Phenols,* Marcel Dekker, New York 1967.
71. K. Mothes, H. R. Schütte: *Biosynthese der Alkaloide,* VEB Deutscher Verlag der Wissenschaften, Berlin 1969.
72. T. Kametani, K. Fukumoto, M. Ihara: "Substrate Behavior" in E. E. van Tamelen (ed.): *Bioorganic Chemistry,* **vol. II,** Academic Press, New York 1978.
73. W. M. Whaley, T. R. Govindachari, *Org. React. (NY)* **6** (1951) 74.
74. W. M. Whaley, T. R. Govindachari, *Org. React. (NY)* **6** (1951) 151.
75. W. J. Gensler, *Org. React. (NY)* **6** (1951) 191.
76. J. M. Bobbitt, D. N. Roy, A. Marchand, C. W. Allen, *J. Org. Chem.* **32** (1967) 2225.
77. D. G. Odonovan, H. Horan, *J. Chem. Soc. C* 1969, 1737.
78. D. G. Odonovan, E. Barry, H. Horan, *J. Chem. Soc. C* 1971, 2398.
79. J. M. Bobbitt, R. Evermann, M. Schubert, *Tetrahedron Lett.* 1963, 575.
80. M. P. Cava, K. T. Buck, A. I. daRocha, *J. Am. Chem. Soc.* **94** (1972) 5931.
81. A. Brossi, S. Teitel, *Helv. Chim. Acta* **54** (1971) 1564.
82. H. Corrodi, E. Hardegger, *Helv. Chim. Acta* **39** (1956) 889.
83. H. Budzikiewicz, C. Djerassi, D. H. Williams: *Structure Elucidation of Natural Products by Mass Spectrometry,* **vol. 1: Alkaloids,** Holden-Day Inc., San Franciso 1964.
84. J. C. Craig, M. Martin-Smith, S. K. Roy, J. B. Stenlake, *Tetrahedron* **22** (1966) 1335.
85. T. Kametani, M. Ihara, *J. Chem. Soc. C* 1968, 1305.
86. N. S. Bhacca, J. C. Craig, R. H. F. Manske, S. K. Roy, M. Shamma, W. A. Slusarchyk, *Tetrahedron* **22** (1966) 1469.
87. T. Kametani, M. Takemura, M. Ihara, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1977, 390.
88. S. Teitel, A. Brossi, *J. Heterocycl. Chem.* **5** (1968) 825.
89. G. Grethe, M. Uskokovic, A. Brossi, *J. Org. Chem.* **33** (1968) 2500.
90. H. W. Gibson, F. D. Popp, A. C. Noble, *J. Heterocycl. Chem.* **3** (1966) 99.

91. G. Grethe, H. L. Lee, M. R. Uskokovic, A. Brossi, *Helv. Chim. Acta* **53** (1970) 874.
92. T. Kametani, T. Kobari, K. Fukumoto, M. Fujiwara, *J. Chem. Soc. C* 1971, 1796.
93. E. Späth, A. Burger, *Ber. Dtsch. Chem. Ges.* **60** (1927) 704.
94. A. Pictet, A. Gams, *Ber. Dtsch. Chem. Ges.* **43** (1910) 2384.
95. T. Kametani, M. Ihara, *Heterocycles* **13** (1979) 497.
96. T. Kametani, K. Ohkubo, I. Noguchi, R. H. F. Manske, *Tetrahedron Lett.* 1965, 3345. T. Kametani, K. Ohkubo, I. Noguchi, *J. Chem. Soc. C* 1966, 715.
97. T. Kametani, K. Ohkubo, R. H. F. Manske, *Tetrahedron Lett.* 1966, 985.
98. T. Kametani, K. Ohkubo, *Chem. Pharm. Bull.* **15** (1967) 608.
99. M. U. Ibragimova, M. S. Yunusov, S. Y. Yunusov, *Khim. Prir. Soedin.* 1970, no. 6, 638.
100. T. Kametani, Y. Enomoto, K. Takahashi, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1979, 2836.
101. A. R. Battersby, R. Binks, *J. Chem. Soc.* 1955, 2888.
102. R. P. K. Chan, J. C. Craig, R. H. F. Manske, T. O. Soine, *Tetrahedron* **23** (1967) 4209.
103. T. Kametani, S. Hirata, K. Ogasawara, *J. Chem. Soc. Perkin Trans. 1* 1973, 1466.
104. S. F. Dyke, A. C. Ellis, *Tetrahedron* **27** (1971) 3803.
105. B. Umezawa, O. Hoshino, *Heterocycles* **3** (1975) 1005.
106. M. Curcumilli-Rodostamo in R. H. F. Manske (ed.): *The Alkaloids,* **vol. XIII,** Academic Press, New York 1971.
107. M. P. Cava, K. T. Buck, K. L. Stuart in R. H. F. Manske (ed.): *The Alkaloids,* **vol. XVI,** Academic Press, New York 1977, p. 249.
108. M. Shamma, V. S. Geogriev in R. H. F. Manske (ed.): *The Alkaloids,* **vol. XVI,** Academic Press, New York 1977, p. 319.
109. T. Kametani, H. Iida, K. Sakurai, *J. Chem. Soc. C* 1969, 500.
110. M. Tomita, K. Fujitani, Y. Aoyagi, Y. Kajita, *Chem. Pharm. Bull.* **16** (1968) 217.
111. E. Fujita, A. Sumi, *Chem. Pharm. Bull.* **18** (1970) 2591.
112. Y. Inubushi, Y. Masaki, S. Matsumoto, F. Takami, *J. Chem. Soc. C* 1969, 1547.
113. J. M. Bobbitt, R. C. Hallcher, *J. Chem. Soc. D* 1971, 543.
114. M. P. Cava, A. Afzali, *J. Org. Chem.* **40** (1975) 1553.
115. S. M. Kupchan, H. W. Altland, *J. Med. Chem.* **16** (1973) 913.
116. R. H. F. Manske, *J. Am. Chem. Soc.* **72** (1950) 55.
117. J. Kunitomo, K. Morimoto, K. Yamamoto, Y. Yoshikawa, K. Azuma, K. Fujitani, *Chem. Pharm. Bull.* **19** (1971) 2197.
118. H. Shimanouchi, Y. Sasada, T. Honda, T. Kametani, *J. Chem. Soc. Perkin Trans. 2* 1973, 1226.
119. R. Suau, M. Vaslpuesta, M. Silva, *Phytochemistry* **28** (1989) 3511.
120. G. R. Clark, R. H. F. Manske, G. J. Palenik, R. Rodrigo, D. B. Maclean, L. Baczynskyj, D. E. F. Gracey, J. K. Saunders, *J. Am. Chem. Soc.* **92** (1970) 4998.
121. A. H. Jackson, G. W. Stewart, *J. Chem. Soc. D* 1971, 149.
122. T. Kametani, K. Fukumoto, M. Fujihara, *J. Chem. Soc. D* 1971, 352.
123. T. Kametani, K. Fukumoto, *J. Chem. Soc.* 1963, 4289. T. Kametani, S. Shibuya, S. Seino, K. Fukumoto, *J. Chem. Soc.* 1964, 4146.
124. S. Ishiwata, T. Fujii, N. Miyaji, Y. Satoh, K. Itakura, *Chem. Pharm. Bull.* **18** (1970) 1850.
125. J. D. F. de Sousa, J. A. R. Rodrigues, R. A. Abranovitch, *J. Am. Chem. Soc.* **116** (1994) 9745.
126. D. H. R. Barton, T. Cohen: Festschrift Arthur Stoll, Birkhäuser, Basel 1957, p. 117.
127. K. Bernauer, *Helv. Chim. Acta* **51** (1968) 1119.
128. T. Kametani, H. Yagi, *J. Chem. Soc.* **2967,** 2182.
129. T. Kametani, T. Sugahara, H. Sugi, S. Shibuya, K. Fukumoto, *J. Chem. Soc. D* 1971, 724. T. Kametani, H. Sugi, S. Shibuya, K. Fukumoto, *Chem. Ind. (London)* 1971, 818.
130. Z. Horii, Y. Nakashita, C. Iwata, *Tetrahedron Lett.* 1971, 1167.
131. S. Ishiwata, K. Itakura, K. Misawa, *Chem. Pharm. Bull.* **18** (1970) 1219. S. Ishiwata, K. Itakura, *Chem. Pharm. Bull.* **18** (1970) 1841.
132. F. Faltis, E. Alder, *Arch. Pharm. (Weinheim, Ger.)* **284** (1951) 281.
133. C. Djerassi, K. Mislow, M. Shamma, *Experientia* **18** (1962) 53.
134. J. C. Craig, S. K. Roy, *Tetrahedron* **21** (1965) 395.
135. A. H. Jackson, J. A. Martin, *J. Chem. Soc. C* 1966, 2181.
136. D. F. DeTar, *Org. React. (NY)* **9** (1957) 409.
137. T. Kametani, I. Noguchi, *J. Chem. Soc. C* 1967, 1440.

138. M. Shamma, W. A. Slusarchyk, *Tetrahedron Lett.* 1965, 1509.
139. M. S. Gibson, G. W. Prenton, J. M. Walthew, *J. Chem. Soc. C* 1970, 2234.
140. S. Ishiwata, K. Itakura, *Chem. Pharm. Bull.* **17** (1969) 2261.
141. T. Kametani, M. Koizumi, K. Shishido, K. Fukumoto, *J. Chem. Soc. C* 1971, 1923. T. Kametani, H. Sugi, S. Shibuya, K. Fukumoto, *Chem. Pharm. Bull.* **19** (1971) 1513.
142. M. P. Cava, M. J. Mitchell, S. C. Havlicek, A. Lindert, R. J. Spangler, *J. Org. Chem.* **35** (1970) 175.
143. S. M. Kupchan, R. M. Kanojia, *Tetrahedron Lett.* 1966, 5353. S. M. Kupchan, J. L. Moniot, R. M. Konoja, J. B. O'Brien, *J. Org. Chem.* **36** (1971) 2413.
144. T. Kametani, S. Shibuya, H. Sugi, O. Kusama, K. Fukumoto, *J. Chem. Soc. C* 1971, 2446.
145. T. Kametani, A. Kozuka, K. Fukumoto, *J. Chem. Soc. C* 1971, 1021.
146. A. R. Battersby, T. J. Brocksom, R. Ramage, *J. Chem. Soc. D* 1969, 464.
147. M. S. Gibson, J. M. Walter, *Chem. Ind. (London)* **965,** 185.
148. S. V. Kessar, S. Batra, S. S. Gandhi, *Indian J. Chem.* **8** (1970) 468.
149. T. Kametani, S. Shibuya, K. Kigasawa, M. Hiiragi, O. Kusama, *J. Chem. Soc. C* 1971, 2712.
150. T. Kametani, K. Fukumoto, T. Nakano, *J. Heterocycl. Chem.* **9** (1972) 1363.
151. T. Kametani, M. Ihara, T. Takemura, Y. Satoh, H. Terasawa, Y. Ohta, K. Fukumoto, K. Takahashi, *J. Am. Chem. Soc.* **99** (1977) 3805.
152. T. Kametani, M. Ihara, *J. Chem. Soc. Perkin Trans. 1* 1980, 629.
153. O. Hoshino in Attu-ur-Rahman (ed.): *Studies in Natural Products Chemistry,* **vol. XVI,** Stereoselective Synthesis (Part J), Elsevier, Amsterdam 1995 p. 503, and references cited therein.
154. T. Kametani, M. Ihara, T. Honda, *J. Chem. Soc. C* 1970, 1060.
155. S. Pfeifer, D. Thomos, *Pharmazie* **22** (1967) 701.
156. T. Kametani, A. Ujiie, M. Ihara, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1975, 1822.
157. R. H. F. Manske, *J. Am. Chem. Soc.* **69** (1947) 1800.
158. T. Kametani, M. Ihara, T. Honda, *J. Chem. Soc. C* 1969, 2342.
159. T. Kametani, T. Honda, M. Ihara, *J. Chem. Soc. C* 1971, 2396.
160. T. Kametani, M. Ihara, T. Honda, H. Shimanouchi, Y. Sasada, *J. Chem. Soc. C* 1971, 2541.
161. H. Kaneko, S. Naruto, *Yakugaku Zasshi* **91** (1971) 101.
162. F. Bohlmann, *Chem. Ber.* **91** (1958) 2157.
163. M. Uskoković, H. Bruderer, C. Von Planta, T. Williams, A. Brossi, *J. Am. Chem. Soc.* **86** (1964) 3364.
164. T. Kametani, K. Fukumoto, M. Ihara, A. Ujiie, K. Koizumi, *J. Org. Chem.* **40** (1975) 3280.
165. G. Snatzke, J. Hrbek, Jr., L. Hruban, A. Horeau, F. Šantavý, *Tetrahedron* **26** (1970) 5013.
166. M. Shamma, C. D. Jones, J. A. Weiss, *Tetrahedron* **25** (1969) 4347.
167. A. R. Battersby, R. Southgate, J. Staunton, M. Hirst, *J. Chem. Soc. C* 1966, 1052.
168. T. Kametani, K. Fukumoto, T. Terui, K. Yamaki, E. Taguchi, *J. Chem. Soc. C* 1971, 2709.
169. T. Kametani, T. Honda, M. Ihara, *J. Chem. Soc. D* 1970, 1253.
170. T. Kametani, A. Ujiie, M. Ihara, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1976, 1218.
171. T. Kametani, M. Ihara, *J. Chem. Soc. C* 1967, 530;1968, 1305.
172. D. W. Brown, S. F. Dyke, G. Hardy, M. Sainsbury, *Tetrahedron Lett.* 1968, 5177.
173. G. R. Lenz, N. C. Yang, *Chem. Commun.* 1967, 1136.
174. T. Kametani, K. Ogasawara, T. Takahashi, *J. Chem. Soc. D* 1972, 675; *Tetrahedron* **29** (1973) 73.
175. R. D. Haworth, W. H. Perkin, Jr., *J. Chem. Soc.* 1926, 1769.
176. K. W. Bentley, A. W. Murray, *J. Chem. Soc.* 1963, 2497.
177. A. R. Pinder, *Chem. Rev.* **64** (1964) 551.
178. R. H. F. Manske, H. L. Holmes: *The Alkaloids,* **vol. IV,** Academic Press, New York 1954.
179. S. Teitel, S. O'Brien, A. Brossi, *J. Org. Chem.* **37** (1972) 1879.
180. E. Hope, R. Robinson, *J. Chem. Soc.* **105** (1914) 2085.
181. R. D. Haworth, A. R. Pinder, R. Robinson, *Nature (London)* **165** (1950) 529.
182. S. McLean, M.-S. Lin, A. C. MacDonald, J. Trotter, *Tetrahedron Lett.* 1966, 185.
183. H. Irie, T. Kishimoto, S. Ueyo, *J. Chem. Soc. C* 1968, 3051.
184. . S. McLean, M.-S. Lin, J. Whelan, *Tetrahedron Lett.* 1968, 2425.
185. M. Shamma, C. D. Jones, *J. Am. Chem. Soc.* **91** (1969) 4009; **92** (1970) 4943.

186. T. Kametani, T. Takahashi, K. Ogasawara, *J. Chem. Soc. Perkin Trans. 1* 1973, 1464.
187. T. Ibuka, T. Konoshima, Y. Inubushi, *Tetrahedron Lett.* 1972, 4001.
188. C. S. Huber, *Acta Crystallogr. Sect. B* **28** (1972) 982.
189. H. Irie, S. Tani, H. Yamane, *J. Chem. Soc. Perkin Trans. 1* 1972, 2986.
190. W. Klötzer, S. Teitel, J. F. Blount, A. Brossi, *J. Am. Chem. Soc.* **93** (1971) 4321.
191. K. Orito, R. H. F. Manske, R. Rodrigo, *J. Am. Chem. Soc.* **96** (1974) 1944.
192. M. Shamma, L. Töke, *J. Chem. Soc. Chem. Commun.* 1973, 740.
193. S. Döpke, J. Flentje, P. W. Jeffs, *Tetrahedron* **24** (1968) 4459.
194. K. K. W. Bentley in S. W. Pelletier (ed.): *Chemistry of the Alkaloids,* Chap. 5, Van Nostrand Reinhold Co., New York 1970.
195. K. L. Stuart, V. Teetz, B. Franck, *J. Chem. Soc. D* 1969, 333.
196. D. H. R. Barton, A. J. Kirby, G. W. Kirby, *J. Chem. Soc. C* 1968, 929.
197. M. Gates, G. Tschudi, *J. Am. Chem. Soc.* **78** (1956) 1380.
198. D. Elad, D. Ginsburg, *J. Chem. Soc.* 1954, 3052.
199. D. H. R. Barton, D. S. Bhakuni, R. James, G. W. Kirby, *J. Chem. Soc. C* 1967, 128.
200. T. Kametani, M. Ihara, K. Fukumoto, H. Yagi, *J. Chem. Soc. C* 1969, 2030.
201. T. Kametani, K. Fukumoto, S. Shibuya, H. Nemoto, T. Nakano, T. Sugahara, T. Takahashi, Y. Aizawa, M. Toriyama, *J. Chem. Soc. Perkin Trans. 1* 1972, 1435. T. Kametani, K. Fukumoto, *J. Heterocycl. Chem.* **8** (1971) 341.
202. Hoffmann-La Roche, CH 180189, 1934.
203. *Chem. Eng. News* 1982(Apr. 19) 41. J. Trojanek, F. Kafka, J. Vit, Z. Cekan, CS 110847, 1964.
204. M. Tomita, T. Ibuka, Y. Inubushi, Y. Watanabe, M. Matsui, *Chem. Pharm. Bull.* **13** (1965) 538. S. M. Kupchan, M. I. Suffness, D. N. J. White, A. T. McPhail, G. A. Sim, *J. Org. Chem.* **33** (1968) 4529.
205. M. Tomita, Y. Okamoto, T. Kikuchi, K. Osaki, M. Nishikawa, K. Kamiya, Y. Sasaki, K. Matoba, K. Goto, *Chem. Pharm. Bull.* **19** (1971) 770.
206. Y. Inubushi, T. Ibuka, M. Kitano, *Chem. Pharm. Bull.* **19** (1971) 1820.
207. T. Ibuka, K. Tanaka, Y. Inubushi, *Tetrahedron Lett.* 1970, 4811.
208. T. Kametani, H. Nemoto, T. Kobari, K. Shishido, K. Fukumoto, *Chem. Ind. (London)* 1972, 538.
209. T. Kametani, T. Kobari, K. Fukumoto, *J. Chem. Soc. Chem. Commun.* 1972, 228.
210. R. Robinson, *Nature (London)* **162** (1948) 524.
211. A. R. Battersby, R. Binds, T. P. Edwards, *J. Chem. Soc.* 1960, 3474.
212. E. E. van Tamelen, P. E. Aldrich, J. B. Hester, Jr., *J. Am. Chem. Soc.* **81** (1959) 507, 6214.
213. A. R. Battersby, R. J. Parry, *J. Chem. Soc. D* 1971, 901.
214. E. E. van Tamelen, C. Placeway, G. P. Schiemenz, I. G. Wright, *J. Am. Chem. Soc.* **91** (1969) 7359.
215. H. T. Openshaw, N. Whittaker, *J. Chem. Soc. C* 1963, 1461.
216. C. Szántay, L. Töke, P. Kolonits, *J. Org. Chem.* **31** (1966) 1447.
217. S. Takano, M. Sasaki, H. Kanno, K. Shishido, K. Ogasawara, *Heterocycles* **7** (1977) 143.
218. T. Kametani, Y. Suzuki, H. Terasawa, M. Ihara, *J. Chem. Soc. Perkin Trans. 1* 1979, 1211.
219. T. Kametani, S. A. Surgenor, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1981, 920.
220. T. Kametani, Y. Suzuki, M. Ihara, *Can. J. Chem.* **57** (1979) 1679.
221. A. Mondon in S. W. Pelletier (ed.): *Chemistry of the Alkaloids,* Van Nostrand Reinhold Co., New York 1970, p. 173.
222. A. Mondon, H. J. Nestler, *Angew. Chem.* **76** (1964) 651.
223. Y. Tsuda, S. Hosoi, N. Katagiri, C. Kaneko, T. Sano, *Chem. Pharm. Bull.* **41** (1993) 2087.
224. D. H. R. Barton, R. B. Boar, D. A. Widdowson, *J. Chem. Soc. C* 1970, 1208.
225. J. Auerbach, S. M. Weinreb, *J. Am. Chem. Soc.* **94** (1972) 7122; **97** (1975) 2503.
226. M. F. Semmelhack, B. P. Chong, L. D. Jones, *J. Am. Chem. Soc.* **94** (1972) 8629. M. F. Semmelhack, B. P. Chong, R. D. Stauffer, T. D. Rogerson, A. Chong, L. D. Jones, *J. Am. Chem. Soc.* **97** (1975) 2507.
227. S. M. Weinreb, M. F. Semmelhack, *Acc. Chem. Res.* **8** (1975) 158.
228. M. Ikeda, M. Okano, K. Kosaka, M. Kido, H. Ishibashi, *Chem. Pharm. Bull.* **41** (1993) 276.
229. R. G. Powell, D. Weisleder, C. R. Smith, Jr., *J. Pharm. Sci.* **61** (1972) 1227.
230. B. Pecherer, A. Brossi, *J. Org. Chem.* **32** (1967) 1053.
231. A. R. Battersby, A. K. Bhatnagar, P. Hackett, C. W. Thornber, J. Staunton, *Chem. Commun.* 1968, 1214.

232. A. S. Bailey, R. Robinson, *J. Chem. Soc.* 1950, 1375. A. S. Bailey, C. R. Worthing, *J. Chem. Soc.* 1956, 4535.
233. S. F. Dyke, M. Sainsbury, B. J. Moon, *Tetrahedron* **24** (1968) 1467.
234. W. Oppolzer, K. Keller, *J. Am. Chem. Soc.* **93** (1971) 3836.
235. M. Onda, K. Yonezawa, K. Abe, *Chem. Pharm. Bull.* **19** (1971) 31.
236. T. Kametani, T. Honda, M. Ihara, H. Shimanouchi, Y. Sasada, *Tetrahedron Lett.* 1972, 3729.
237. I. Ninomiya, O. Yamamoto, T. Naito, *J. Chem. Soc. Perkin Trans. 1* 1980, 212.
238. M. Cushman, A. Abbaspour, Y. P. Gupta, *J. Am. Chem. Soc.* **105** (1983) 2873.
239. R. Robinson, S. Sugasawa, *J. Chem. Soc.* 1932, 789.
240. A. R. Battersby, R. B. Herbert, E. McDonald, R. Ramage, J. H. Clements, *Chem. Commun.* 1966, 603. A. C. Barker, A. R. Battersby, E. McDonald, R. Ramage, J. H. Clements, *Chem. Commun.* 1967, 390.
241. A. R. Battersby, R. B. Herbert, L. Mo, F. SÎantavy, *J. Chem. Soc. C* 1967, 1739.
242. T. Kametani, S. Takano, K. Haga, *Chem. Pharm. Bull.* **16** (1968) 663.
243. T. Kametani, S. Takano, T. Kobari, *J. Chem. Soc. C* 1969, 9.
244. T. Kametani, M. Koizumi, in R. H. F. Manske (ed.): *The Alkaloids,* **vol. XIV,** Academic Press, New York 1973, p. 265.
245. T. Kametani, F. Satoh, H. Yagi, K. Fukumoto, *J. Org. Chem.* **33** (1968) 690.
246. A. R. Battersby, R. B. Bradbury, R. B. Herbert, M. H. G. Munro, R. Ramage, *Chem. Commun.* 1967, 450.
247. T. Kametani, M. Koizumi, *J. Chem. Soc. C* 1971, 3976.
248. T. Kametani, M. Koizumi, K. Fukumoto, *J. Org. Chem.* **36** (1971) 3729.
249. J. Schreiber, W. Leimgruber, M. Pesaro, P. Schudel, T. Threlfall, A. Eschenmoser, *Helv. Chim. Acta* **44** (1961) 540.
250. E. E. van Tamelen, T. A. Spencer, D. S. Allen, R. L. Orvis, *J. Am. Chem. Soc.* **81** (1959) 6341; *Tetrahedron* **14** (1961) 8.
251. R. B. Woodward, *Harvey Lect.* **59** (1963) 31.
252. S. Sunagawa, J. Nakamura, K. Nakazawa, *Chem. Pharm. Bull* **. 9** (1961) 81; **10** (1962) 281.
253. A. I. Scott, F. McCapra, R. L. Buchanan, A. C. Day, D. W. Young, *Tetrahedron* **21** (1965) 3605. K. Kotani, F. Miyazaki, S. Tobinaga, *J. Chem. Soc. Chem. Commun.* 1974, 300.
254. J. W. Cook, J. D. Loudon in R. H. F. Manske, H. L. Holmes (eds.): *The Alkaloids.* **vol. II,** Academic Press, New York 1977, p. 261 – 329.
255. T. Hata, H. Fukumi, S. Sato, K. Aiba, C. Tamura, *Acta Crystallogr. Sect. B* **34** (1978) 2899.
256. H. Fukumi, H. Kurihara, H. Mishima, *Chem. Pharm. Bull.* **26** (1978) 2175.
257. T. Arai, K. Takahashi, A. Kubo, S. Nakahara, S. Sato, K. Aiba, C. Tamura, *Tetrahedron Lett.* 1979, 2355.
258. T. Fukuyama, R. A. Sachleben, *J. Am. Chem. Soc.* **104** (1982) 4957.
259. S F. Martin in A. Brossi (ed.): *The Alkaloids,* **vol. XXX,** Academic Press, New York 1987, p. 251, and references cited therein.
260. O. Hoshino in G. A. Cordell (ed.): *The Alkaloids,* **vol. LI,** Academic Press, New York 1998, p. 323, and references cited therein.
261. A. R. Battersby, R. Binks, S. W. Breuer, H. M. Rales, W. C. Wildman, R. J. Highet, *J. Chem. Soc.* 1964, 1595.
262. A. R. Battersby, H. M. Fales, W. C. Wildman, *J. Am. Chem. Soc.* **83** (1961) 4095.
263. D. H. R. Barton, G. W. Kirby, J. B. Traylor, G. M. Thomas, *J. Chem. Soc.* 1963, 4545.
264. H. M. Fales, W. C. Wildman, *J. Am. Chem. Soc.* **86** (1964) 294. W. C. Wildman, N. E. Himer, *J. Am. Chem. Soc.* **89** (1967) 5265.
265. M. Shiro, T. Sato, H. Koyama, *Chem. Ind. (London)* 1966, 1229.
266. G. Stork, D. J. Morgans, *J. Am. Chem. Soc.* **101** (1979) 7110.
267. S. F. Martin, C.-Y. Tu, M. Kimura, S. H. Simonsen, *J. Org. Chem.* **47** (1982) 3763.
268. T. Weller, D. Seebach, *Tetrahedron Lett.* 1982, 935.
269. P. W. Jeffs, N. A. Cortese, J. Wolfram, *J. Org. Chem.* **47** (1982) 3881.
270. B. Umesawa et al., *Tetrahedron,* **40** (1984) 1783. O. Hoshino et al., *Chem. Lett.* 1991, 1365. O. Hoshino et al., *J. Chem. Soc. Perkin Trans. l,* 1996, 571.
271. B. K. Boeckman, Jr., S. W. Goldstein, M. A. Walters, *J. Am. Chem. Soc.* **110** (1988) 8250.
272. A. G. Schultz, M. A. Holoboski, M. S. Smyth, *J. Am. Chem. Soc.* **115** (1993) 7905; **118** (1996) 6210.
273. W. C. Wildman, *J. Am. Chem. Soc.* **80** (1958) 2567.
274. G. G. DeAngelis, W. C. Wildman, *Tetrahedron Lett.* 1969, 729.
275. J. Clardy, F. M. Hauser, D. Dahm, R. A. Jacobson, W. C. Wildman, *J. Am. Chem. Soc.* **92** (1970) 6337.

276. T. Okamoto, Y. Torii, Y. Isogai, *Chem. Pharm. Bull.* **16** (1968) 1860.
277. F. Piozzi, C. Fuganti, R. Mondelli, G. Geriotti, *Tetrahedron* **24** (1968) 1119. F. Piozzi, M. L. Marino, C. Fuganti, A. Di Martino, *Phytochemistry* **8** (1969) 1745.
278. A. Immirzi, C. Fuganti, *J. Chem. Soc. Chem. Commun.* 1972, 240.
279. G. R. Pettit, V. Gaddamidi, G. M. Cragg, D. L. Herald, Y. Sagawa, *J. Chem. Soc. Chem. Commun.* 1984, 1693. G. R. Pettit et al., *J. Nat. Prod.* **49** (1986) 995.
280. S. Danishefsky, J. Y. Lee, *J. Am. Chem. Soc.* **111** (1989) 4829.
281. N. Chida, M. Ohtsuka, S. Ogawa, *Tetrahedron Lett.* **32** (1991) 4441. N. Chida, M. Ohtsuka, S. Ogawa, *J. Org. Chem.* **58** (1993) 4525.
282. T. Hudlicky, H. F. Olivo, *J. Am. Chem. Soc.* **114** (1992) 9694.
283. B. M. Trost, S. R. Puliey, *J. Am. Chem. Soc.* **117** (1995) 10143.
284. X. Tian, R. Maruya, K. Koenigsberger, T. Hudlicky, *Synlett* 1995, 1125. X. Tian, T. Hudlicky, K. Koenigsberger, *J. Am. Chem. Soc.* **117** (1995) 3643. T. Hudlicky, X. Tian, K. Koenigsberger, *J. Am. Chem. Soc.* **118** (1996) 10752.
285. G. E. Keck, S. F. McHardy, J. A. Murry, *J. Am. Chem. Soc.* **117** (1995) 7289.
286. S. Kobayashi, T. Shingu, S. Uyeo, *Chem. Ind. (London)* 1956, 177. S. Kobayashi, S. Uyeo, *J. Chem. Soc.* 1957, 638.
287. S. Uyeo, S. Kobayashi, *Chem. Pharm. Bull.* **1** (1953) 139.
288. D. J. Williams, D. Rogers, *Proc. Chem. Soc. London* 1964, 357.
289. D. H. R. Barton, G. W. Kirby, *J. Chem. Soc.* 1962, 806.
290. T. Kametanai et al., *J. Chem. Soc. C* 1971, 1043. T. Kametani et al., *J. Org. Chem.* **36** (1971) 1295. T. Kametani, M. S. Premila, K. Fukumoto, *Heterocycles* **4** (1976) 111.
291. K. Shimizu, K. Tomioka, S. Yamada, K. Koga, *Heterocycles* **8** (1977) 277.
292. J. Szewczyk, A. H. Lewin, F. I. Carrol, *J. Heterocycl. Chem.* **25** (1988) 1809. J. Szewczyk, J. W. Wilson, A. H. Lewin, F. I. Lewin, *J. Heterocycl. Chem.* **32** (1995) 1195.
293. F. Viladomat, J. Bastida, C. Codina, W. E. Campbell, S. Mathee, *Phytochemistry* **40** (1995) 307.
294. O. Hoshino, M. Ishizaki, *Chem. Lett.* 1990, 1817. M. Ishizaki, O. Hoshino, Y. Iitaka, *Tetrahedron Lett.* **32** (1991) 7082. M. Ishizaki, O. Hoshino, Y. Iitaka, *J. Org. Chem.* **57** (1992) 7285. M. Ishizaki, K. Kurihara, E. Tanazawa, O. Hoshino, *J. Org. Chem.* **58** (1993) 3877.
295. L. E. Overman, J. Shim, *J. Org. Chem.* **56** (1991) 5005; **58** (1993) 4662.
296. J. Jin, S. M. Weinreb, *J. Am. Chem. Soc.* **119** (1997) 5773.
297. P. Coggon, D. S. Farrier, P. W. Jeffs, A. T. McPhail, *J. Chem. Soc. B* 1970, 1267.
298. S. Takano, Y. Imamura, K. Ogasawara, *Tetrahedron Lett.* 1981, 4479.
299. G. E. Keck, R. R. Webb II, *J. Org. Chem.* **47** (1982) 1302 and references cited therein.
300. S. Takano, Y. Imamura, K. Ogasawara, *Tetrahedron Lett.* **22** (1981) 4479. S. Takano, K. Samizu, K. Ogasawara, *Chem. Lett.* 1990, 1239.
301. H. Kosugi, Y. Miura, H. Hanna, H. Uda, *Tetrahedron: Asymmetry* **4** (1993) 1409.
302. H. Nemoto, T. Tanabe, K. Fukumoto, *Tetrahedron Lett.* **35** (1994) 6499. H. Nemoto, T. Tanabe, K. Fukumoto, *J. Org. Chem.* **60** (1995) 6785.
303. A. Brossi, G. Grethe, S. Teitel, W. C. Wildman, D. T. Bailey, *J. Org. Chem.* **35** (1970) 1100.
304. M. A. Schwartz, S. W. Scott, *J. Org. Chem.* **36** (1971) 1827.
305. D. P. Chakraborty, P. Bhattacharyya, S. Roy, S. P. Bhattacharyya, A. K. Biswas, *Phytochemistry* **17** (1978) 834.
306. I. Mester, J. Reisch, *Justus Liebigs Ann. Chem.* 1977, 1725.
307. Raymond-Hamet, *C. R. Hebd. Seances Acad. Sci.* **221** (1945) 387.
308. A. J. Birch, J. J. Wright, *Tetrahedron* **26** (1970) 2329.
309. T. Kametani, N. Kanaya, M. Ihara, *J. Am. Chem. Soc.* **102** (1980) 3974.
310. S. Inoue, N. Takamatsu, Y. Kishi, *Yakugaku Zasshi* **97** (1977) 553. S. Inoue, N. Takamatsu, K. Hashizume, Y. Kishi, *Yakugaku Zasshi* **97** (1977) 582.
311. M. Abe, *Nippon Nogei Kagaku Kaishi* **27** (1953) 18.
312. W. Acklin, T. Fehr, D. Arigoni, *Chem. Commun.* 1966, 799.
313. D. Stauffacher, H. Tscherter, *Helv. Chim. Acta* **47** (1964) 2186.
314. A. Hofmann, A. J. Frey, H. Ott, *Experientia* **17** (1961) 206.
315. A. Stoll, A. Hofmann, T. Petrzilka, *Helv. Chim. Acta* **34** (1951) 1544.
316. H. Tscherter, H. Hauth, *Helv. Chim. Acta* **57** (1974) 113.
317. A. R. Battersby, A. R. Burnett, P. G. Parsons, *J. Chem. Soc. C* 1969, 1193.

318. R. T. Brown, C. L. Chapple, G. Lashford, *J. Chem. Soc. Chem. Commun.* 1975, 295.
319. M. Rueffer, N. Nagakura, M. H. Zenk, *Tetrahedron Lett.* 1978, 1593.
320. S. R. Johns, J. A. Lamberton, J. L. Occolowitz, *Aust. J. Chem.* **20** (1967) 1463.
321. E. J. Shellard, P. J. Houghton, *Planta Med.* **21** (1972) 16.
322. E. E. van Tamelen, M. Shamma, A. W. Burgstahler, J. Wolinsky, R. Tamm, P. E. Aldrich, *J. Am. Chem. Soc.* **80** (1958) 5006.
323. R. B. Woodward, F. E. Bader, H. Bickel, A. J. Frey, R. W. Kierstead, *J. Am. Chem. Soc.* **78** (1956) 2023, 2657; *Tetrahedron* **2** (1958) 1.
324. F. Chemnitius, *Chem. Ztg.* **51** (1927) 169.
325. C. H. Boehringer Sohn, DE 967469, 1957; GB 772122, 1957.
326. R. B. Woodward, *Angew. Chem.* **68** (1956) 13. F. A. L. Anet, D. Chakravarti, R. Robinson, E. Sclitter, *J. Chem. Soc.* 1954, 1242.
327. S. Masamune, S. K. Ang, C. Egli, N. Nakatsuka, S. K. Sarkar, Y. Yasunari, *J. Am. Chem. Soc.* **89** (1967) 2506.
328. J. P. Kutney, G. R. Eigendorf, H. Matsue, A. Murai, K. Tanaka, W. L. Sung, K. Wada, B. R. Worth, *J. Am. Chem. Soc.* **100** (1978) 938.
329. R. N. Chakravarti, R. Robinson, *Nature (London)* **160** (1947) 18. R. B. Woodward, *Nature (London)* **162** (1948) 155 A. F. Peerdemann, *Acta Crystallogr.* **9** (1956) 824.
330. R. B. Woodward, *Tetrahedron* **19** (1963) 247.
331. H. Wieland, W. Gumlich, *Justus Liebigs Ann. Chem.* **494** (1932) 191.
332. C. Weissmann, H. Schmid, P. Karrer, *Helv. Chim. Acta* **44** (1961) 1877.
333. S. Goodwin, A. F. Smith, E. C. Horning, *J. Am. Chem. Soc.* **81** (1959) 1903.
334. J. Schmutz, F. Junziker, *Pharm. Acta Helv.* **33** (1958) 341.
335. R. B. Woodward, G. A. Iacobucci, F. A. Hochstein, *J. Am. Chem. Soc.* **81** (1959) 4434.
336. T. Kametani, T. Suzuki, Y. Ichikawa, K. Fukumoto, *J. Chem. Soc. Perkin Trans. 1* 1975, 2102, and references cited therein.
337. G. Büchi, S. J. Gould, F. Näf, *J. Am. Chem. Soc.* **93** (1971) 2492.
338. T. Kametani, T. Suzuki, *J. Org. Chem.* **36** (1971) 1291.
339. G. Stork, J. E. Dolfini, *J. Am. Chem. Soc.* **85** (1963) 2872.
340. M. Ando, G. Büchi, T. Ohnuma, *J. Am. Chem. Soc.* **97** (1975) 6880.
341. J. P. Kutney, U. Bunzli-Trepp, K. K. Chan, J. P. de Souza, Y. Fujise, T. Honda, J. Katsube, F. K. Klein, A. Leutwiler, S. Morehead, M. Rohr, B. R. Worth, *J. Am. Chem. Soc.* **100** (1978) 4220.
342. H. K. Schnoes, K. Biemann, *J. Am. Chem. Soc.* **86** (1964) 5693.
343. A. Guggisberg, A. A. Gorman, B. W. Bycroft, H. Schmidt, *Helv. Chim. Acta* **52** (1969) 76, and references cited therein.
344. G. Hugel, B. Gourdier, J. Levy, J. Le Men, *Tetrahedron Lett.* 1974, 1597.
345. N. Neuss, M. Gorman, *Tetrahedron Lett.* 1961, 206.
346. G. Arai, J. Coppola, G. A. Jeffrey, *Acta Crystallogr.* **13** (1960) 553.
347. G. Büchi, D. L. Coffen, K. Kocsis, P. E. Sonnet, F. E. Ziegler, *J. Am. Chem. Soc.* **88** (1966) 3099.
348. N. Camerman, J. Trotter, *Acta Crystallogr.* **17** (1964) 384.
349. J. P. Kutney, L. S. L. Choi, T. Honda, N. G. Lewis, T. Sato, K. L. Stuart, B. R. Worth, *Helv. Chim. Acta* **65** (1982) 2088.
350. P. Mangeney, R. Zo Andriamialisoa, N. Langlois, Y. Langlois, P. Potier, *J. Am. Chem. Soc.* **101** (1979) 2243.
351. G. H. Svoboda, I. S. Johnson, M. Gorman, N. Neuss, *J. Pharm. Sci.* **51** (1962) 707 . Symposium Pagass, Lloydia 27 (1964) 275 – 485.
352. T. Kametani, T. Ohsawa, M. Ihara, *J. Chem. Soc. Perkin Trans. 1* 1981, 1563.
353. Y.-Y. Chiu, W. N. Lipscomb, *J. Am. Chem. Soc.* **97** (1975) 2525.
354. S. M. Weinreb, F. Z. Basha, S. Hibino, N. A. Khatri, D. Kim, W. E. Pye, T.-T. Wu, *J. Am. Chem. Soc.* **104** (1982) 536. A. K. Kende, D. P. Lorah, R. J. Boatman, *J. Am. Chem. Soc.* **103** (1981) 1271.
355. W. E. Oberhänsli, *Helv. Chim. Acta* **52** (1969) 1905.
356. K. Wiesner, W. A. Ayer, L. R. Fowler, Z. Valenta, *Chem. Ind. (London)* 1957, 564.
357. R. H. F. Manske, L. Morion, *J. Am. Chem. Soc.* **69** (1947) 2126.
358. F. A. L. Anet, M. V. Rao, *Tetrahedron Lett.* 1960, 9.
359. M. Przybylska, F. R. Ahmed, *Acta Crystallogr.* **11** (1958) 718. K. Nishio, T. Fujiwara, K. Tomita, H. Ishii, Y. Inubushi, T. Harayama, *Tetrahedron Lett.* 1969, 861.
360. Z. Valenta, H. Yoshimura, E. F. Rogers, M. Ternbah, K. Wiesner, *Tetrahedron Lett.* 1960, 26. H. Yoshimura, Z. Valenta, K. Wiesner, *Tetrahedron Lett.* 1960, 14.

361. C. H. Heathcock, E. F. Kleinman, E. S. Binkley, *J. Am. Chem. Soc.* **104** (1982) 1054.
362. T. R. Govindachari, K. Nagarajan, S. Rajappa, *J. Chem. Soc.* 1957, 2725.
363. E. M. Dickinson, G. Jones, *Tetrahedron* **25** (1969) 1523.
364. T. Imanishi, M. Hanaoka, *Yuki Gosei Kagaku Kyokaishi* **41** (1983) 403.
365. Y. Inubushi, Y. Sasaki, Y. Tsuda, J. Nakano, *Tetrahedron Lett.* 1965, 1519.
366. Y. Inubushi, T. Kikuchi, T. Ibuka, K. Tanaka, J. Saji, K. Tokane, *J. Chem. Soc. Chem. Commun.* 1972, 1252.
367. N. Uesaka et al., *J. Org. Chem.* **59** (1994) 5633, and references cited therein.
368. G. Büchi, I. M. Goldman, D. W. Mayo, *J. Am. Chem. Soc.* **88** (1966) 3109.
369. J. Koyama, T. Okatani, T. Ogura, K. Tagahara, H. Irie, *Chem. Pharm. Bull.* **39** (1991) 481.
370. T. Sugita, J. Koyama, K. Tagahara, Y. Suzuta, *Heterocycles* **24** (1986) 29.
371. T. Honda, F. Ishikawa, S. Yamane, *J. Chem. Soc. Chem. Commun.* 1994, 499, and references cited therein.
372. J. P. Johnston, K. H. Overton, *J. Chem. Soc. Perkin Trans. 1* 1972, 1490.
373. W. Nagata, T. Sugasawa, M. Narisada, T. Wakabayashi, Y. Hayase, *J. Am. Chem. Soc.* **89** (1967) 1483.
374. T. Kametani, Y. Kato, T. Honda, K. Fukumoto, *J. Am. Chem. Soc.* **98** (1976) 8185.
375. M. Benn, J. F. Richardson, *Heterocycles* **24** (1986) 1605.
376. S. W. Pelletier, S. A. Ross, J. T. Etse, *Heterocycles* **27** (1988) 2467.
377. E. J. Corey, W. R. Hertler, *J. Am. Chem. Soc.* **80** (1958) 2903.
378. T. Ho, K. Huu, Goutarel, *Bull. Soc. Chim. Fr.* 1963, 594.
379. M. Cesario, J. Guilhem, *Chryst. Struct. Commun.* **6** (1977) 247.
380. M. E. Kopach, A. H. Fray, A. I. Meyers, *J. Am. Chem. Soc.* **118** (1996) 9876, and references cited therein.
381. K. S. Brown, S. M. Kupchan, *J. Am. Chem. Soc.* **86** (1964) 4424.
382. Z. Voticky, L. Dolejs, O. Bauerova, V. Paulik, *Collect. Czech. Chem. Commun.* **42** (1977) 2549.
383. S. Terada, K. Hayashi, H. Mitsuhashi, *Chem. Pharm. Bull.* **25** (1977) 2802.
384. K. Yoshikawa, N. Okada, Y. Kann, S. Arihara, *Chem. Pharm. Bull.* **44** (1996) 1790.
385. J. Kloosterman, *Chem. N.Z.* **43** (1979) 146; *Chem. Abstr.* **92** (1980) 111209.
386. T. Masamune, M. Tagasugi, A. Murai, K. Kobayashi, *J. Am. Chem. Soc.* **89** (1967) 4521.
387. J. P. Kutney, C. C. Fortes, T. Honda, Y. Murakami, A. Preston, Y. Ueda, *J. Am. Chem. Soc.* **99** (1977) 964.
388. F. Märki, B. Witkop, *Experimentia* **19** (1963) 329. J. W. Daly, B. Witkop, P. Bommer, K. Biemann, *J. Am. Chem. Soc.* **87** (1965) 124. T. Tokuyama, J. W. Aly, B. Witkop, *J. Am. Chem. Soc.* **91** (1969) 3931.
389. R. D. Gilardi, *Acta Crystallogr.* **B26** (1970) 440.
390. J. F. W. Keana, R. R. Schumaker, *J. Org. Chem.* **41** (1976) 3840.
391. P. Magnus, T. Leapheart, C. Walker, *J. Chem. Soc. Chem. Commun.* 1985, 1185.
392. P. Hudson, G. Pairudeau, P. J. Parsons, A. W. Jahans, M. G. B. Drew, *Tetrahedron Lett.* **34** (1993) 7295.
393. F. Jellinek, *Acta Crystallogr.* **10** (1957) 277.
394. C. H. Eugster, *Adv. Org. Chem.* **II** (1960), 427.
395. J. Adams, M.-A. Poupart, L. Grenier, *Tetrahedron Lett.* **30** (1989) 1753, and references cited therein.
396. L. Tchissambou, M. Benechie, F. Khuong-Huu, *Tetrahedron* **38** (1982) 2687, and references cited therein.
397. T. Naito, Y. Honda, O. Miyata, I. Ninomiya, *Chem. Pharm. Bull.* **41** (1993) 217.
398. S. M. Kupchan, Y. Komoda, W. A. Court, G. J. Thomas, R. M. Smith, A. Karim, C. J. Gilmore, R. C. Haltiwanger, R. F. Bryan, *J. Am. Chem. Soc.* **94** (1972) 1354.
399. E. J. Corey, L. O. Weigel, A. R. Chamberlin, H. Cho, D. H. Hua, *J. Am. Chem. Soc.* **102** (1980) 6613.
400. M. Kitumara, M. Isobe, Y. Ichikawa, T. Goto, *J. Am. Chem. Soc.* **106** (1984) 3252, and references cited therein.
401. H. Matsuyama, *Yuki Gosei Kagaku Kyokaishi* **39** (1981) 1151.
402. V. U. Ahmad, N. Ismail, A. R. Amber, *Phytochemistry* **28** (1989) 2493.
403. Y. Nagao, E. Fujita, *Heterocycles* **17 (special issue)** (1982) 537.
404. Y. Nagao, S. Takao, T. Miyasaka, E. Fujita, *J. Chem. Soc. Chem. Commun.* 1981, 286.
405. S. Imado, M. Shiro, Z. Horii, *Chem. Ind. (London)* 1964, 1691.
406. N. H. Lajis, O. B. Guan, M. V. Sargent, B. W. Skelton, A. H. White, *Aust. J. Chem.* **45** (1992) 1893, and references cited therein.
407. Z. Horii, M. Hanaoka, Y. Yamawaki, Y. Tamura, S. Saito, N. Shigematsu, K. Kotera, H. Yoshikawa, Y. Sato, H. Nakai, N. Sugimoto, *Tetrahedron* **23** (1967) 1165.

408. K. Th. Wanner, *Eur. J. Pharm. Sci.* **2** (1994) 62.
409. R. B. Woodward, *Pure Appl. Chem.* **9** (1964) 49.
410. Y. Kishi, T. Fukuyama, M. Aratani, F. Nakatsubo, T. Goto, S. Inoue, H. Tanino, S. Sugiura, H. Kakoi, *J. Am. Chem. Soc.* **94** (1972) 9219.
411. M. Isobe, T. Nishikawa, N. Fukami, T. Goto, *Pure Appl. Chem.* **59** (1987) 399.
412. H. Nakamura, J. Kobayashi, Y. Ohizumi, *Tetrahedron Lett.* **23** (1982) 5555.
413. T. R. Kelly, M. P. Maguire, *Tetrahedron* **41** (1985) 3033.
414. T. Sakamoto, N. Miura, Y. Kondo, H. Yamanaka, *Chem. Pharm. Bull.* **34** (1986) 2760.
415. H. Nakamura, J. Kobayashi, Y. Ohizumi, *J. Chem. Soc. Perkin Trans 1* 1987, 173.
416. A. Bassoli, G. Maddinelli, B. Rindstone, S. Tollai, F. Chioccara, *J. Chem. Soc. Chem Commun.* 1987, 150.
417. C. J. Pelletier, M. P. Cava, *J. Org. Chem.* **52** (1987) 616.
418. R. G. Andrew, R. A. Raphael, *Tetrahedron* **43** (1987) 4803.
419. S. Hibino, E. Sugino, T. Choshi, K. Sato, *J. Chem. Soc. Perkin Trans 1* 1988, 2429.
420. P. Balczewski, M. K. J. Mallon, J. D. Street, J. A. Joule, *Tetrahedron Lett.* **31** (1990) 569.
421. P. Molina, A. Vidal, I. Barquero, *Synthesis* 1996, 1199.
422. J. Kobayashi, H. Nakumara, Y. Ohizumi, Y. Hirata, *Tetrahedron Lett.* **27** (1986) 1191. J. Kobayashi et al., *J. Org. Chem.* **55** (1990) 3666. O. Murata et al., *Tetrahedron Lett.* **32** (1991) 3539.
423. R. Sakai, H. Higa, *J. Am. Chem. Soc.* **108** (1986) 6404. R. Sakai, S. Kohomoto, T. Higa, C. W. Jeffort, G. Bernardinelli, *Tetrahedron Lett.* **28** (1987) 5493.
424. B. J. Baker, P. J. Scheuer, J. N. Shoolery, *J. Am. Chem. Soc.* **110** (1988) 965.
425. E. Fahy et al., *Tetrahedron Lett.* **29** (1988) 3427.
426. A. G. M. Bartrett, M. L. Boys, T. L. Boehm, *J. Org. Chem.* **61** (1996) 685.
427. T. S. McDermott, A. A. Mortlock, C. H. Heathcock, *J. Org. Chem.* **61** (1996) 700.

Alkoxides → Alcohols, Aliphatic

Alkyd Resins

FRANK N. JONES, Coatings Research Institute, Eastern Michigan University, Ypsilanti, MI 48197, USA

1. **Introduction** 1
2. **Production** 3
2.1. **Raw Materials** 3
2.2. **Polymerization Theory** 5
2.3. **Formulation and Process Types; Laboratory Processing** 6
2.4. **Manufacture** 8
3. **Film Formation by Coatings Based on Alkyd Resins** 8
4. **Modified Alkyds** 11
5. **Low-Solvent Coatings** 13
6. **Markets and End Uses** 15
7. **References** 16

1. Introduction

Alkyd resins are, literally, polyesters. They are made by condensation polymerization of three types of monomers:

1) Polyols
2) Polybasic acids
3) Fatty acids or triglyceride oils

Because of their distinctive properties, alkyds often are classified separately from other polyesters, as they are in this encyclopedia.

The largest use of alkyds is for surface coatings (paints, enamels, lacquers, and varnishes) in which the resins function as binders, forming a tough, continuous film that adheres to the object coated. Alkyds are the most versatile of coating binders, and they are used extensively in all major categories of coatings: architectural, industrial, and special purpose (see Chap. 6). Despite gradual replacement of alkyds with other types of resins, a substantial fraction of all organic coatings applied worldwide use alkyds as a primary binder or as an additive.

Alkyds are also important ink binders. Other uses include caulks, adhesives, and plasticizers.

Relative to most commercial polymers, alkyd resins often have low glass transition temperatures T_g. Consequently, they are viscous, tacky materials that are difficult to handle. Most often, the handling problems are overcome by dissolving the resins in organic solvents to give solutions that can be poured and pumped.

Because of their low T_g, most alkyds are useful as coating binders only after they are chemically cured to harden the film after the liquid coating has been applied. Curing is accomplished by cross-linking chemical reactions. Cross-linking may result from chemical reaction of the coating film with oxygen in the air or from chemical reaction with a second resin or substance. The second resin, called a cross-linker, is blended with the alkyd at the time of coating manufacture or application. Certain alkyds are added to acrylic and nitrocellulose lacquers, where their role is to plasticize the brittle thermoplastic resins. Alkyds used in this way do not cross-link.

History. Paints containing vegetable oils have been used for at least 600 years. A major goal in the early development of alkyd resins was to combine vegetable oil technology with polyester technology to make improved materials. Probably the first polyester resin was made by BERZELIUS [18] in 1847. It was an intractable product of glycerol and tartaric acid. In 1901 W. SMITH [19] reported a brittle glycerol – phthalic anhydride polymer, and by 1910 workers at General Electric were attempting to develop more tractable materials for electrical end uses. Work was probably hampered by lack of an accepted theory of polymerization, but several patents were obtained.

The General Electric program eventually led to preparation in 1921 of glycerol – phthalic anhydride – fatty acid polymers. By 1926 these polymers were being used commercially as adhesives, and research on their use in coatings was underway [20]. The first account of this work was published in 1929 [21]. A U.S. patent was issued in 1933 [22] but was declared invalid in

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_409

1935. Technical progress presumably was aided by emergence of polymerization theory in the late 1920s [23], [24].

An important early development was the monoglyceride process for making alkyds directly from vegetable oils, eliminating the expense of preparing and purifying glycerine and fatty acids. In the 1930s, linseed oil was the most common binder for architectural paints, but alkyds were beginning to replace it. Alkyds offered faster drying, less yellowing, and superior protective properties. By 1940 U.S. consumption of alkyds exceeded 4×10^4 t/a.

By the early 1950s, alkyds had become the preeminent binder for coatings. U.S. consumption had reached 2×10^5 t/a and was rising. However, water-based architectural paints were being introduced. These paints generally used polyvinyl or polyacrylic emulsion resins as binders and contained little or no alkyd. During the 1950s, 1960s, and 1970s, water-based paints gradually captured most of the architectural coatings market from alkyd- and oil-based paints. The newer paints were popular because they could be cleaned up with water and were economical to produce because of low petroleum prices. Despite declining use in architectural paints, total use of alkyds continued to increase during most of this period because of rapidly growing use in a huge variety of industrial and special purpose coatings and in inks. Alkyd consumption in the USA reached a plateau at about 3×10^5 t/a in 1960–1980 and then began declining at a rate of about 2 %/a, reaching about 2.3×10^5 t/a in 1996 [25]. Worldwide consumption of alkyds is probably declining at a slower rate and perhaps holding steady.

Alkyd technology is being adapted to meet contemporary needs. Concern about air quality led to development of alkyds for use in high-solids and water-reducible coatings. Rising petrochemical prices have increased interest in all polymers made partly from renewable resources. Besides, alkyds are entrenched in innumerable specialized end uses, so wholesale replacement would be difficult.

Definitions and Classifications. Terminology relating to alkyd resins can be confusing because it has never been uniform and because it has changed considerably over the years. The ASTM defines an alkyd as "a synthetic resin made from polyhydric alcohols and polybasic acids; generally modified with resins, fatty oils, or fatty acids" [26]. The DIN definition is similar [27].

For this article, slightly narrower and more exact definitions have been adopted. They are conveniently represented as follows:

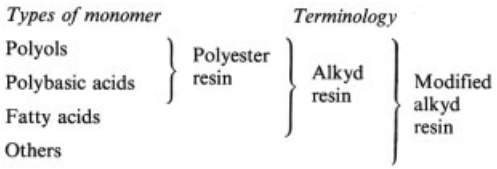

In this article the word "alkyd" is used as both a noun and an adjective. Therefore, "alkyd" and "alkyd resin" are synonymous. This usage is widespread, although some authors feel that only the latter usage is correct.

The scope of this article includes alkyds and modified alkyds. Polyesters are treated under separate keywords (see → Polyester Resins, → Polyesters).

A brief retracing of the history of resin terminology may help clarify some of the terms that may be encountered. "Alkyd" is a respelling of "alcid", a word apparently coined at General Electric for the reaction product of polyfunctional alcohols and organic acids. Therefore, in the 1920s "alkyd" meant what is now meant by "polyester". The resins now called alkyds were called "oil-modified alkyds" during the 1930s. For a time, almost all commercial alkyds contained oil, so the phrase "oil-modified" was redundant and fell out of use. As polyester resins gained popularity in coatings, they began to be called "oil-free alkyds", a term that is now becoming obsolete. In the older literature, the term "glyptal" is found. It is synonymous with "alkyd". It was originally a General Electric trade name.

Today, the term "modified alkyd" usually refers to a resin in which a fourth type of monomer (a modifier) is incorporated in the polymer molecules. This definition is used in this article.

Because alkyds are a large and diverse family of resins, several *subclassifications* are often used. These groupings describe the features that most strongly affect resin properties and end

uses, such as the type and amount of fatty acid and the identity of other components.

Fatty Acid Type. There are two broad groupings:

1) *Drying alkyds* contain enough unsaturated fatty acids to make curing by oxygen possible.
2) *Nondrying alkyds* contain lower levels of unsaturated fatty acids and are not polymerized appreciably with oxygen.

To be more specific, the particular oil used in preparing the resins is often stated.

Fatty Acid Amount. A classification called "oil length" arbitrarily groups alkyds by the amount of oil or fatty acid present. For alkyds made directly from oils, the oil length is given by Equation (1).

$$\text{Oil length} = (\text{mass of oil used to make alkyd resin} \times 100)(\text{mass of reactants-mass of water evolved in process})^{-1} \quad (1)$$

For alkyds made from fatty acids, the formula is amended to correct for the extra mass of water lost (usually about 4 %) during esterification of the fatty acids, i.e., the factor of 100 in Equation (1) is replaced by 104.

Many authors classify alkyds by "oil length", a term apparently derived from terminology in varnish production in which oils were blended with other ingredients such as rosin or phenolic resins. Oil-length classifications are approximate, mainly because the ranges tabulated in Table 1 are standardized in Europe but not, apparently, in the rest of the world. Different authors assign somewhat different dividing points between short and medium, medium and long, etc. Despite being inexact, oil length classifications are used widely. The ranges used in Europe and the USA are listed in Table 1 [28].

Table 1. Percentage of oil in alkyd classes

Oil length	Europe (DIN 55 945)	Common U.S. usage
Short oil alkyds	<40	<45
Medium oil alkyds	40–60	45–55
Long oil alkyds	60–70	55–70
Very long oil alkyds ("alkyd oils")	>70	>70

Other authors, however, give different ranges for the USA; in one case they are the same as STOYE gives for Europe [29]. In view of the ambiguities involved, it is safest to work with the actual oil length numbers, not the verbal ranges.

Other Constituents. When the polyol and polybasic acid are not specified in describing an alkyd, it probably contains glycerol and phthalic anhydride, the most common ingredients. Other ingredients often are mentioned in describing a resin, e.g., "an isophthalic alkyd".

Modified alkyds usually, but not always, are classified according to the generic type of modifer, e.g., "a styrenated alkyd" or "a silicone-modified alkyd".

2. Production

2.1. Raw Materials

Well over one hundred raw materials are used in commercial alkyd manufacture. Each resin contains a minimum of three monomers and often five to seven. Only the most common raw materials are listed here.

Polyols (→ Alcohols, Polyhydric, → Ethylene Glycol, → Glycerol). Alkyds produced by processes that start with triglyceride oils (see Section 2.3) always contain at least some glycerol. Other polyols often are used, leading to alkyds of mixed polyol content.

Processes starting with fatty acids (see Section 2.3) allow the use of any polyol or polyol mixture to achieve the desired spectrum of properties. Glycerol [*56-81-5*] can be used, but less expensive synthetic polyols are a common choice. 1,1,1-Trimethylolpropane [*77-99-6*] is a popular synthetic alternative to glycerol. Pentaerythritol [*115-77-5*], with four hydroxyl groups per molecule, is popular because it builds highly branched polymers and because it is relatively inexpensive. Mixtures of pentaerythritol and ethylene glycol [*107-21-1*] may be used as a low-cost substitute for glycerol.

Polybasic Acids and Their Anhydrides. Phthalic anhydride [*85-44-9*] (→ Phthalic Acid and Derivatives) is used more than any other polybasic acid for reasons of economy and easy processing. Isophthalic acid [*121-91-5*] is often used where superior weatherability and corrosion resistance are required. Cycloaliphatic

diacids, such as 1,4-cyclohexanedicarboxylic acid [*1076-97-7*], are also said to give better weatherability than phthalic acid.Trimellitic anhydride [*552-30-7*] (→ Carboxylic Acids, Aromatic) is used mainly in water-reducible alkyds. Many alkyds contain a small percentage of maleic anhydride [*108-31-6*] (→ Maleic and Fumaric Acids), which is said to reduce resin color and may enhance oxidative cross-linking.

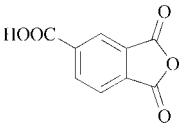

1,2-Polybasic acids are never used as such. Instead, their internal cyclic anhydrides are used. The anhydrides form nominally the same polymer composition as the acids, but they are easier to handle, easier to polymerize, more stable, and less costly. In large-scale operations phthalic anhydride and maleic anhydride are kept hot and are handled in liquid form, an economy not feasible with the high-melting acids.

Triglyceride Oils and Fatty Acids (→ Fats and Fatty Oils, → Fatty Acids, → Drying Oils and Related Products). *Triglycerides* containing unsaturated and mixed fatty acids can remain liquid at temperatures well below 0 °C. For example, linseed oil contains a mixture of C_{18} fatty acids with zero, one, two, and three double bonds per C_{18} chain; it does not freeze at the very cold temperatures to which flax seeds may be exposed. A typical molecule might have the structure:

$$\begin{array}{lll} H_2C-O-\overset{\overset{\displaystyle O}{\|}}{C}(CH_2)_7(CH{=}CHCH_2)_2(CH_2)_3CH_3 & & \text{(linoleic)} \\ \quad | & & \\ HC-O-\overset{\overset{\displaystyle O}{\|}}{C}(CH_2)_7(CH{=}CHCH_2)_3CH_3 & & \text{(linolenic)} \\ \quad | & & \\ H_2C-O-\overset{\overset{\displaystyle O}{\|}}{C}(CH_2)_7CH{=}CH(CH_2)_7CH_3 & & \text{(oleic)} \end{array}$$

The total number of double bonds in a fatty acid, the Wijs iodine number, is measured by titration with iodine [30]. This number, however, correlates poorly with the reactivity of a triglyceride oil toward oxygen, a critical factor for coatings based on drying alkyds. Fatty acid residues containing only one double bond are almost unreactive. Those containing more than one double bond separated by a single methylene group, e.g., –CH=CHCH$_2$CH=CH--, react readily with oxygen and impart drying capablities to the oil or alkyd. Linoleic and linolenic acids have two and three such double bonds, respectively. Conjugated double bonds, e.g., –CH=CHCH=CH--, within the fatty acid are even more reactive.

Nondrying alkyds, for which reactivity with oxygen is unwanted, are made from triglyceride oils that contain a high proportion of saturated fatty acids, such as coconut oil. They can also be made from synthetic fatty acids.

While the fatty acid makeup of each type of oil is somewhat characteristic, the exact composition of the fatty acid mixture may vary widely with soil, climate, strain of plant, extraction process, and other factors. For example, the iodine number of soybean oil may vary from 103 to 151 [30]. Commercial oils are reliably within a much narrower range, for example 128 – 135 for soybean oil [31].

Usually the double bonds of the fatty acids are in the relatively unreactive *cis*, nonconjugated configuration. Certain oils, such as tung, oiticica, and dehydrated castor, contain conjugated and/or *trans* double bonds and are used to make alkyds for fast air-drying and low-bake coatings.

Some of the oils frequently used in alkyds are listed in Table 2 along with their iodine numbers and an indication of the degree of unsaturation of the fatty acids present. The numbers are averages of published data and, of course, are approximate. A wide variety of other oils is used, including tung oil, whale oil, and fish oils from various sources.

The *fatty acids* used in alkyd production are usually made by hydrolyzing (saponifying or "splitting") triglyceride oils to obtain fatty acid mixtures. Linseed oil fatty acid and soybean oil fatty acid [*68308-53-2*] are used widely. The fatty acid composition of the oil is not substantially changed by the hydrolysis process. Some fatty acid producers offer fractionated products that are enriched in particular acids.

An important fatty acid mixture not derived from a seed oil is tall oil fatty acid [*61790-12-3*] (→ Tall Oil). It is a byproduct of kraft pulping. The name derives from the Swedish word for pine. The composition and performance are fairly similar to those of soybean oil fatty acid. The presence of about 10 % of fast-reacting con-

Table 2. Makeup of oils used in alkyds

Oil		Iodine no.	Typical fatty acid composition, % ∗				
			0	1	2	3	4–6
Linseed	[*8001-26-1*]	190	11	18	16	55	–
Manhaden	[*8002-50-4*]	180	37	30	2	1	30
Safflower	[*8001-23-8*]	145	10	12	78	tr.	–
Sunflower	[*8001-21-6*]	135	10	25	65	<1	–
Dehydrated castor	[*64147-40-6*]	135	<1	7	92 ∗∗	–	–
Soybean	[*8001-22-7*]	132	15	25	52	8	–
Coconut	[*8001-31-8*]	8	91	7	2	–	–

∗ Indicates percentage of fatty acid chains containing zero, one, two, etc., double bonds per fatty acid chain
∗∗ Contains a high proportion of conjugated double bonds

jugated double bonds compensates for its lower iodine number of about 125. Tall oil fatty acid always contains some rosin as an impurity; it is sold in grades specifying the maximum amount of rosin. The 1.5 % and 4 % grades are used widely in alkyds. Tall oil produced in Europe has a somewhat different composition than tall oil produced in North America; resin production formulas developed with one type are likely to fail if the other type is substituted.

Synthetic fatty acids are used in alkyds to a modest extent.

Modifiers (see Chap.4).

Solvents. Immediately after production almost all alkyds are thinned in organic solvent to make them pumpable and pourable. Generally, these solvents are not removed at the time of coating manufacture, and they become part of the coating formula. Medium and long oil types usually are thinned with mineral spirits. Short oil alkyds require more polar solvents, often blends of aromatic hydrocarbon solvents. Alkyds to be used in high-solids coatings are sometimes thinned with ester or ketone solvents, which are efficient viscosity reducers. Water-reducible alkyds are thinned with water-miscible solvents such as 2-butoxyethanol or 1-propoxy-2-propanol.

Many alkyd resin and paint formulas were changed during the 1980s and 1990s to avoid solvents classified in the USA as "hazardous air pollutants" (HAPs). Similar changes were made in other countries.

2.2. Polymerization Theory

When a difunctional acid or anhydride (AA) and a difunctional alcohol (BB) are heated together, a stepwise reaction occurs that builds up linear chain molecules:

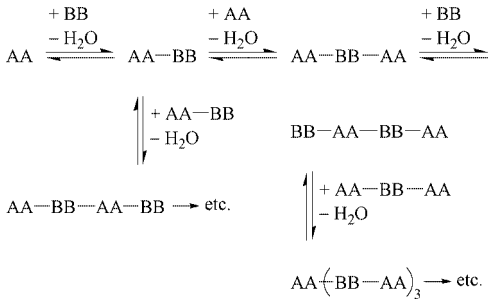

Compositions containing short chains (3 to about 20 monomer units per chain) are called "oligomers". Those containing long chains are "polymers". Under ideal conditions polymer chains containing thousands of monomer units can be made. Conditions necessary for formation of very high molecular mass chains include the following:

1) The mixture must be heated long enough to complete the reaction.
2) Equimolar quantities of AA and BB must be present.
3) No alternative reactions may be possible.
4) AA and BB must contain no monofunctional reactive impurities.

The average molecular mass can be deliberately held at lower levels by altering any of the above ideal conditions, for example:

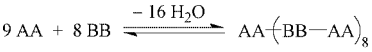

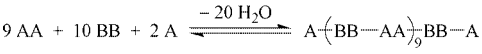

where A represents a monofunctional acid. The actual products of such reactions are mixtures of various chain lengths.

If a trifunctional monomer BBB is added, chain branching occurs:

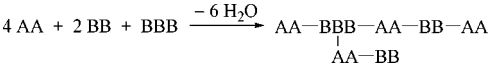

If substantial amounts of BBB are present, continued polymerization leads to formation of a three-dimensional network, or gel:

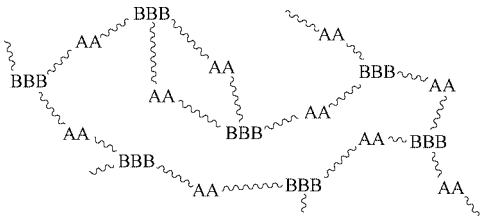

One of the goals in alkyd resin synthesis is to prepare structures that are branched but are not quite gelled. A fairly close approach to gelation is needed to form a polymer which readily forms cross-linked networks after application as a coating. However, gelation in the reactor causes a total loss of materials and a difficult clean-up job.

Much research has been directed toward understanding polymerization and developing general principles to help the alkyd formulator achieve the desired degree of polymerization and to optimize properties. Early theories of polymerization were developed by CAROTHERS [23], [29] and by KIENLE et al. [20], [22], [24], [34], [35]. CAROTHERS hypothesized a uniform, stepwise polymerization process. In his model, gelation occurs as the polymer molecular mass approaches infinity. The condition for gelation would be

$$p_g = \frac{2}{F_{av}}$$

where

p_g is the extent of reaction at the gel point and

F_{av} is the average functionality of the monomer mixture.

According to this equation pure difunctional monomers can never gel, but a mixture of glycerol and phthalic anhydride ($F_{av} = 2.5$) would gel at $p_g = 0.8$; that is, it would gel when 80 % of the functional groups have reacted.

In practice gelation usually occurs at lower p_g than predicted by the Carothers theory. To explain this fact FLORY et al. [32], [33] proposed a probabilistic theory: gelation occurs when a relatively small proportion of the monomer molecules first becomes bonded into a three-dimensional network. FLORY's theory is satisfactory for predicting polymerizations of the chemically straightforward monomers used to make polyesters. However, alkyd formulators must deal with monomers that are far from straightforward. Complications include

- Variable starting materials
- Competing side reactions
- Reaction reversibility
- Differential reactivity of the functional groups
- Differential solubility effects
- Nonuniform process conditions

Gradual progress has been made toward a general and reliable theory of alkyd formation by several authors; see, e.g., [6], [38–44]. A completely precise theory is unlikely, mainly because alkyd polymerizations probably are not homogeneous during the late stages of reaction. Most alkyds contain small, soft gel particles as well as soluble material [44], and these gel particles are in some cases essential to the attainment of optimum properties [45], [46]. Despite this difficulty, certain alkyd polymerizations can be accurately modeled with the aid of computers and advanced statistical techniques [47]. Such theoretical methods offer considerable help and insight to the alkyd formulator. They can reduce reliance on empirical trial and error, but they will not completely eliminate it for the foreseeable future.

2.3. Formulation and Process Types; Laboratory Processing

Alkyds are made in batch processes, of which several types have evolved.

Fusion Process Using Fatty Acids. The fusion process starting with fatty acids is the oldest and most straightforward. A representative formulation is given in Table 3 [48].

Table 3. Composition of long oil alkyd

	Equivalents COOH	OH	g	% in alkyd
Soybean fatty acids	1.246	–	349.0	60.0
Pentaerythritol	–	3.545	125.5	21.5
Phthalic anhydride	2.000	–	148.0	25.4
Totals	3.246	3.545	622.5	106.9
H_2O			−40.4	−6.9
Theoretical yield of resin			582.1	100.0

Like nearly all alkyd formulations, it contains an excess of hydroxyl groups, 9.2 % in this case. This excess is essential to assure that the rate of viscosity increase will be controllable as the polymerization process approaches completion. To visualize the consequences of not using an excess of OH groups, suppose that exactly 3.246 equivalents of pentaerythritol have been used. The F_{av} would have been 2.07, and CAROTHERS' theory predicts an exponential increase in viscosity as the reaction approaches completion, ending in gelation at $p_g = 0.96$. As FLORY's theory and practical experience demonstrate, gelation would actually occur at lower p_g. Therefore in the late stages of reaction, viscosity would be increasing rapidly and exponentially. However, the 9.2 % excess of OH groups assures that the number average chain length will remain modest even when nearly all of the carboxyl groups have reacted, limiting viscosity at the desired level. Chain lengths of individual molecules in the resin will vary from the oligomer to the high-polymer range.

Viscosity could be controlled by using an excess of acid, but inferior coatings would result because of moisture sensitivity imparted by the unconverted COOH groups in the film. Unconverted OH groups impart less moisture sensitivity, and they play a useful role in certain cross-linking reactions (see Chap. 3).

To prepare the above resin on a *laboratory scale* the ingredients are placed in a round-bottom reaction flask, heated at about 240 °C, and stirred mechanically. Inert gas (N_2 or CO_2) is passed through the reaction mass to facilitate water removal; the water is condensed and measured to estimate the extent of reaction. As the reaction approaches completion (5 – 10 h), samples are removed and the acid value and/or viscosity are measured. When the desired degree of reaction has been reached, the reaction mass is cooled slightly and is diluted with about an equal volume of mineral spirits.

Fusion Process Using Triglyceride Oils. The fusion process starting from triglyceride oils is often called the *alcoholysis* or the *monoglyceride process*. It is popular because the oil provides an economical source of all of the fatty acid and part of the polyol. It is a two-stage, batch process:

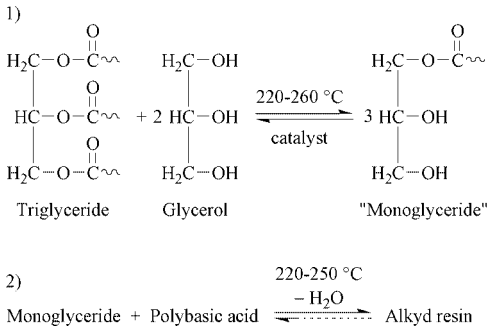

The first, or alcoholysis, stage is an ester interchange reaction. Equilibrium is reached within an hour or two at sufficiently elevated temperature when the reaction is catalyzed by litharge (PbO) or by certain Li or Ca salts. Although the product is called a monoglyceride, it is actually a mixture of isomers of mono-, di-, and triglycerides and of glycerol. Polyols other than glycerol often are used and yield even more complex mixtures. In the second stage polybasic acids or anhydrides are polymerized with the monoglyceride mixture to form the resin. Other ingredients often are added during the second stage.

Procedures and equipment for laboratory preparation are generally similar to those used for the fatty acid process.

Both fusion processes suffer from several practical difficulties. To assure reasonable process rates the water of reaction must be removed by "sparging" the reaction mass with a stream

of inert gas. Without sparging, process rates are slow because the esterification reaction is highly reversible. With sparging, however, the gas stream sweeps vapors of phthalic anhydride and polyol out of the reaction mass. These materials may be lost altogether or they may condense in the upper parts of the apparatus from which they may fall back into the reactor, causing dark or nonuniform product.

Solvent Processes. To overcome these problems solvent processes are used widely, both with fatty acid and triglyceride starting materials. In these processes a few percent of unreactive, water-immiscible solvent is added to the formulation. Commercial xylene, also called xylol (mixed dimethylbenzenes and ethylbenzene, *bp* 138 – 140 °C), is used commonly. Typically a stirred alkyd polymerization mass containing 4 % xylene will boil vigorously at about 250 °C. The xylene vapors form an azeotrope with water and carry it out of the reactor. Usually the vapor stream is passed through a "partial condenser" operated at about 100 °C; this allows water vapor to escape, but it condenses most of the xylene and almost all of the monomers and returns them to the reactor. Whether or not a partial condenser is used, the vapors are condensed. Water and xylene separate into two layers, and the xylene is returned to the reactor.

Laboratory equipment for solvent processing has been described in detail [49]. Losses of xylene and of monomers can be minimized with proper operation, and uniform products can be produced.

2.4. Manufacture

A variety of equipment for alkyd production has been described [5], [8], [50]. Equipment for *solvent processing* will be described here because the trend is in that direction. However, fusion processes, which require simpler equipment, often produce satisfactory products. The fusion process is preferable for isophthalic alkyds.

Figure 1 shows a versatile configuration designed for the solvent process. The reactor typically would be constructed of no. 316 stainless steel. Capacity of commercial-scale reactors ranges from 2 to 35 m^3. A turbine agitator driven by a powerful electric motor is necessary. Heating and cooling generally are accomplished by a heat-transfer fluid flowing either through a jacket around the reactor (as shown in Figure 1) or through coils within the reactor. High-pressure steam and gas-flame heating have been used but are less satisfactory and more hazardous.

Whereas the reactor and overhead shown in Figure 1 are the heart of an alkyd plant, other equipment is needed for efficient, large-scale production. Such equipment includes the following:

1) Raw material storage and metering equipment (phthalic anhydride often handled in hot, molten form)
2) A manhole in the reactor top for adding solids and to provide access for cleaning
3) A heater for heat-transfer fluid
4) A thin tank about 2.5 times as large as the reactor, equipped with a condenser and agitator
5) Product storage tanks and drumming equipment
6) Environmental control equipment, including incinerators for gaseous and liquid waste
7) Filters
8) Safety and fire protection equipment, including rupture disks, blowout vents, and sprinklers

Automated alkyd reactors have been installed in Europe, North America, and Asia [51]. Improved safety, quality, and productivity are claimed. However, several factors have discouraged retrofitting of reactors to automate them. Although the necessary microprocessors are relatively inexpensive, automation requires costly feedback systems and servomotor driven valves. Detailed process studies are necessary to program the microprocessor to deal with all contingencies.

3. Film Formation by Coatings Based on Alkyd Resins

There are two mechanisms by which a liquid coating can be transformed into a tough, protective film after it has been applied:

1) Lacquer drying by simple solvent evaporation

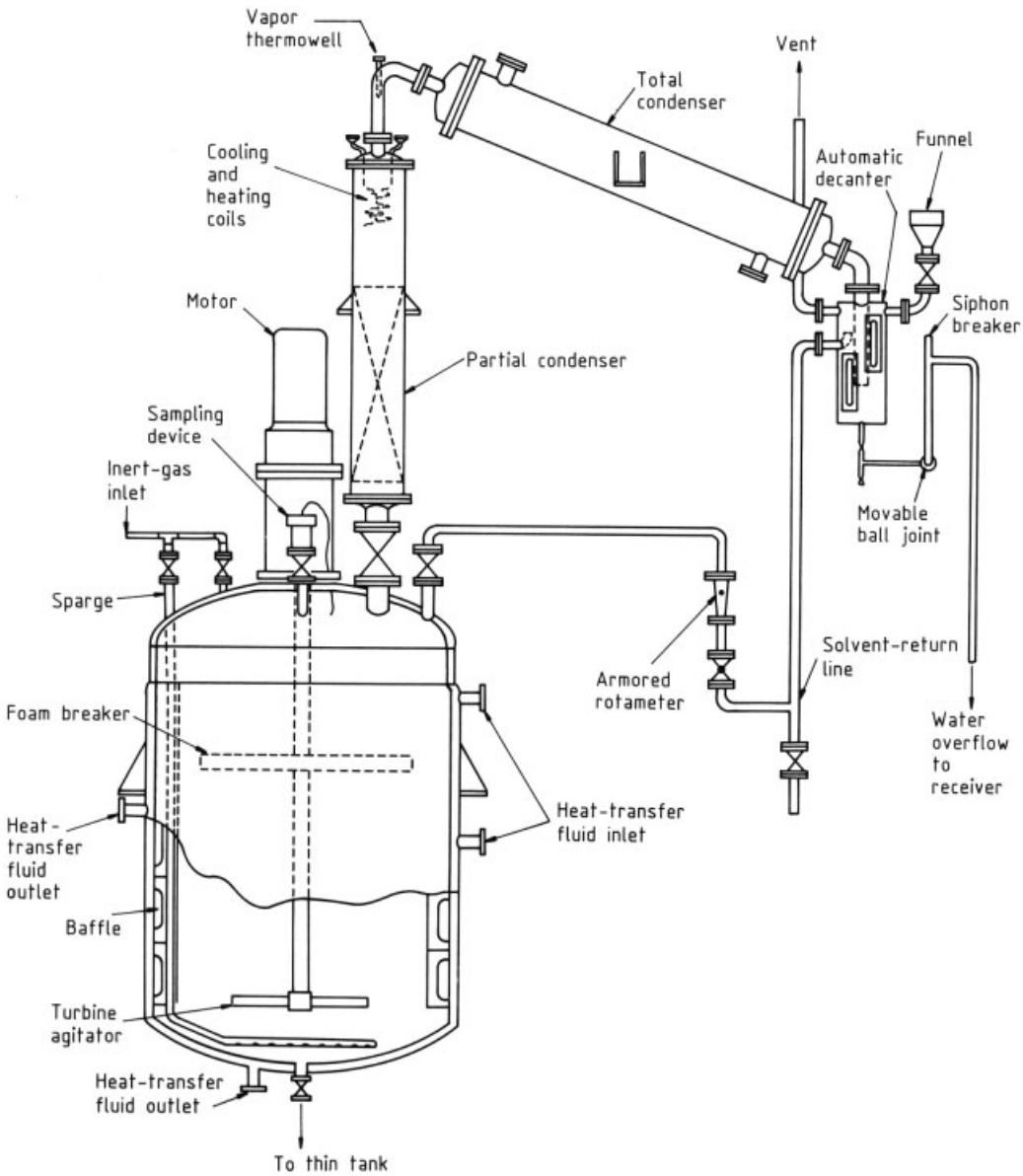

Figure 1. Alkyd solvent-processing equipment with partial condenser (adapted from [30])

2) Chemical cross-linking by reaction of the polymeric binders to form a cross-linked polymer network

Coatings based on alkyd resins often depend on both mechanisms to some degree. Chemical cross-linking plays a vital role in converting a soft, tacky alkyd into a hard, tough film.

Drying Alkyds. The majority of alkyd coatings are designed for film formation at ambient or slightly elevated temperatures. The chemical mechanism of curing involves autoxidation of the unsaturated fatty acid residues in the resin. This chemically complex process has been studied extensively [52–58], but it is not yet completely understood. Most of the studies have involved model compounds such as monoesters of

fatty acids rather than the more complex alkyd resins.

Autoxidation of alkyds is a chain reaction that proceeds by a free-radical mechanism. A key step is the formation of a hydroperoxide on an allylic methylene group:

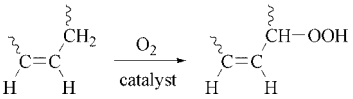

The hydroperoxide is capable of several further reactions, some of which lead to formation of stable, covalent bonds with sites on other resin molecules. This bonding builds up a three-dimensional polymeric network, and a tough film is formed.

Small amounts of catalysts (driers) that promote autoxidation and cross-linking are added to drying-type alkyd coatings. Driers are generally mixtures of oil-soluble salts of Co, Mn, Zr, Ca or other metals [59]: soluble Pb salts are also effective but have been banned in many countries.

Fatty acid chains containing only one double bond are autoxidized slowly. Those containing allylic methylene groups activated by two double bonds react over 100 times faster. Reaction sequences probably involve rearrangement and isomerization of the double bonds, for example:

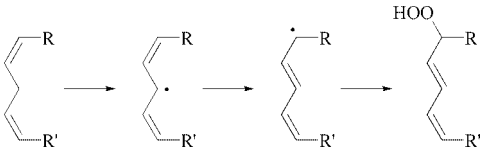

Studies on model compound [56–58] confirm the mechanism shown above for the early stages of cross-linking, but reveal that subsequent film-forming reactions are extremely complex and beyond the scope of this article.

Useful drying rates are achieved only if appreciable concentrations of polyunsaturated fatty acid residues are present. Soybean, safflower, sunflower, and tall oil alkyds are about the least unsaturated that dry well. These oils have high linoleic acid (two double bonds) content but relatively low levels of fatty acids with three or more double bonds. They are preferred for decorative coatings where resistance to yellowing is desired and moderate cross-linking is adequate.

Higher levels of cross-linking and therefore better protective properties are achieved by use of fatty acids or oils with substantial contents of triply unsaturated fatty acid residues (linseed or tung) or of those containing multiple unsaturated fatty acid residues (fish oils). Blends of fatty acids or oils are used to reduce cost and sometimes to minimize premature cross-linking during alkyd manufacture.

Drying rate is often an important consideration. Oils containing two conjugated double bonds dry faster than those containing similar levels of nonconjugated double bonds. Thus, chemically modified oils and fatty acids with two conjugated double bonds (dehydrated castor oil and conjugated tall oil fatty acids) find considerable use. Fastest of all are oils that contain fatty acids with three conjugated double bonds, such as tung oil.

Rates of autoxidation are useful above about 0 °C and sharply increase with increasing temperature. A significant advantage of drying oil alkyd-based paints over latices is that they can form satisfactory films when applied at 0 °C or even a bit below.

Drawbacks of all alkyd paints paints include a tendency of the films to turn yellow and become brittle with age. Their low permeability to water can also lead to blistering when painted on wood. Use of high-linoleic, low-linolenic oils such as safflower and sunflower can minimize these problems, but can not eliminate them. Short oil alkyds are used in applications where rapid partial drying is needed. Although they have less cross-linking capability than longer oil types, they quickly develop some degree of hardness by lacquer drying as the solvent evaporates. Quick drying is desirable in many manufacturing situations, as it permits gentle handling within 15 – 30 min after coating. Film properties improve for several days or weeks as the slower cross-linking process occurs.

Nondrying Alkyds. Alkyds are used extensively in industrial coatings that are designed to be cured at elevated temperatures (baking or stoving enamels). Such enamels are usually blends of alkyds with aminoplasts, e.g., alkylated melamine – formaldehyde or urea – formaldehyde resins. Cross-linking occurs by

acid-catalyzed condensation of the OH groups of the alkyd with the aminoplast:

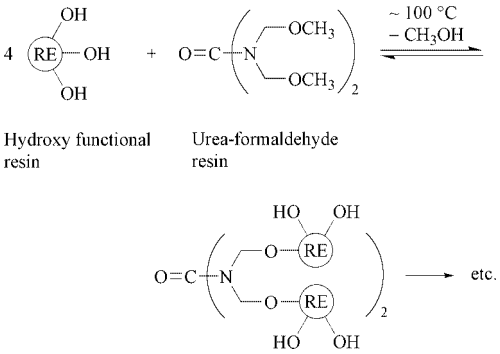

Other chemical cross-linkers can be blended with alkyds. For example, alkyd – polyisocyanate blends are used as binders for two-package, ambient temperature curing polyurethane coatings.

Alkyds for baking enamels do not need to be autoxidizable and often are made from oils containing only saturated fatty acids, coconut, for example. Soybean and tall oil alkyds often are used in alkyd – amino enamels; such enamels cure by a mixture of condensation and autoxidative cross-linking.

Polyvalent metal alkoxides, such as $Al(OC_4H_9)_3$, can serve as trifunctional cross-linkers. They react with hydroxyl groups on the alkyd resin to form $Al(O–alkyd)_3$ structures within the film. This expedient can enhance hardness and through-cure of the paints.

Nondrying alkyds are used as plasticizers in lacquers. Their role is to soften high molecular mass thermoplastic resins that otherwise would be too hard and brittle for coatings use. Examples are nitrocellulose – alkyd blends for furniture finishes and acrylic – alkyd blends for automotive refinish lacquers. Note that in Europe the term "lacquers" is frequently applied to all types of organic coatings, while in North America the term tends to be limited to coatings made from thermoplastic, non-cross-linking binders.

4. Modified Alkyds

Roughly one-fourth of commercial alkyd usage is of modified alkyds, resins in which modifiers are at least partly interpolymerized with the usual alkyd monomers. The more important types are described in this section.

Chain-stopped alkyds contain aromatic monobasic acids (see Section 2.2), such as benzoic acid [*65-85-0*] or 4-(1,1-dimethylethyl)benzoic acid (*p-tert*-butylbenzoic acid) [*98-73-7*]. These modifiers improve lacquer-dry characteristics and harden films. However, because they do not participate in covalent cross-linking, the films tend toward brittleness.

Inclusion of monobasic acids during polymerization reduces F_{av}; to compensate, highly functional polyols such as pentaerythritol are used. Because pentaerythritol and benzoic acid are relatively inexpensive, low-cost resins with reasonably good properties are attainable. Uses are diverse, a major one being traffic-control striping. Chain-stopped alkyds are useful in high-solids coatings (see Chapter 5).

For premium-quality chain-stopped resins, 4-*tert*-butylbenzoic acid affords an excellent combination of fast dry and film plasticity. However, use of this modifier is limited by concerns about its potential for long-term health effects.

Styrenated alkyds are made in a two-stage process [54], [60]. First a drying oil alkyd is produced using a formulation designed to limit molecular mass. The resin is then thinned, usually with xylene, and styrene [*100-42-5*] is added to the solution. Addition of a free-radical initiator starts the second stage of the process, the addition polymerization of the styrene. This step is effected at higher temperature (about 140 °C) than most addition polymerizations, conditions that favor grafting of polystyrene segments onto the alkyd. The product is a mixture of graft copolymer, ungrafted alkyd, and homopolystyrene [61]. Apparently sufficient grafting occurs to "compatibilize" the mixture.

Styrene content ranges up to 60 %. The homopolystyrene ($T_g \approx 100$ °C) segments impart hardness, chemical resistance, and very fast lacquer dry characteristics. However, high levels of styrene reduce weatherability and toughness.

Acrylated alkyds are made by a similar process using acrylic or methacrylic monomers in place of styrene [62]. These monomers are more costly than styrene, but their presence improves resistance to sunlight. For example, an alkyd modified with methyl methacrylate [*80-62-6*] (homopolymer $T_g \approx 105$ °C) would be simi-

lar in most properties to a styrenated alkyd but would be more weatherable. Alkyds modified with lower T_g acrylic or methacrylic monomers can be made readily. They are less important because they offer few advantages over unmodified alkyds.

Other olefinic monomers, such as acrylonitrile [*107-13-1*], α-methylstyrene [*98-83-9*], and mixed 3- and 4-vinyltoluene (VT), are used commercially to modify alkyds. The latter mixture affords the commercially important "VT alkyds", which lacquer dry nearly as fast as styrenated alkyds but have better film properties. Uses include fast-dry industrial enamels and aerosol paints.

Silicone alkyds are prepared by interpolymerizing hydroxyfunctional organosilicone oligomers with alkyd resins [63]. Methyl/phenyl-substituted silicones are used. The proportion of silicone ranges from 15 to 55 %. These premium quality resins have superior resistance to heat and weather. Drying oil silicone alkyds are used in heat-resistant coatings such as stove enamels and in exterior applications in which prolonged retention of properties and appearance saves substantial cost by reducing the frequency of repainting, e.g., petroleum storage tanks. Blends of nondrying silicone alkyds with melamine resins have been used in coil-applied coatings for exterior siding.

Thixotropic alkyds are modified at high temperatures with an oil-miscible polyamide resin, for example, a polymer [*51178-84-8*] of octadecadienoic acid dimer [*6144-28-1*] and 1,2-diaminoethane [*107-15-3*]. Proportions and reaction conditions must be carefully controlled [64]. Paints containing these resins are viscous, but their viscosity drops under mechanical shear and then recovers when the shear is removed. It is unclear from published information whether their behavior is simple shear thinning (near instantaneous recovery when shear is removed) or genuine thixotropy (time-delayed recovery). Regardless of methanism, the paints resist running and sagging when brushed or sprayed in thick coats. Applications include bridge and marine paints.

Uralkyds are usually air-drying compositions in which part of the polybasic acid is replaced by a polyfunctional isocyanate, most commonly toluene diisocyanate [*26471-62-5*] (→ Isocyanates, Organic) to form a polyester – polyurethane interpolymer. The resins are usually made in a two-stage process, the reactive toluene diisocyanate being added in the second stage at relatively low temperature.

An excess of OH groups is used so that all of the isocyanate reacts to form urethane linkages. Cross-linking is accomplished by the usual autoxidative mechanism. Presence of the urethane linkages imparts outstanding toughness, water resistance, and abrasion resistance; thus, urethane alkyds are used in premium-quality varnishes, for example, for gymnasium floors. A drawback is that resins containing toluene diisocyanate are brownish and yield yellowish varnishes that darken with age. This problem can be almost eliminated by use of the more costly aliphatic polyisocyanates in place of toluene diisocyanate.

A second class of uralkyds is prepared with a deficiency of OH groups so that unconverted isocyanate groups are present in the resin. After application these resins cross-link by reaction of atmospheric moisture with the isocyanate groups. They form exceptionally tough films. However, the coatings are difficult to apply and may cause skin irritation.

Epoxy esters do not fit the strict definition of alkyd resins, but they are closely related. These resins usually are formed by condensing an unsaturated fatty acid with an epoxy resin of low to intermediate molecular mass (→ Epoxy Resins). Epoxy ester coatings combine part of the toughness, adhesion, and corrosion resistance of epoxies with the easy application and cure of alkyds. They also share the main shortcomings of epoxies: surface chalking when exposed to sunlight and discoloration with age. Epoxy esters are used where their adhesion and corrosion resistance are needed and where surface weatherability is unimportant, for example, in auto body primers and in industrial maintenance paints.

Phenolic modified alkyds are used in applications in which hardness and moisture resistance are important. High-solids formulations have been developed [65].

Rosin and rosin esters (→ Resins, Natural) are used as modifiers mainly to reduce cost. Coating properties are similar to those of chain-stopped alkyds, but films yellow with age and have inferior weatherability. These modifiers are used in wall primers.

5. Low-Solvent Coatings

In most coating processes, the solvent in the coating evaporates into the atmosphere. Formerly the amounts were thought harmless, but by the 1960s it was recognized that the solvents used in alkyd and other coatings contribute substantially to air pollution. In this context, the solvents are called the "volatile organic content" (VOC) of the coating. In air, organic solvents react in a complex photochemical process to produce ozone, one of the most troublesome constituents of photochemical smog.

In the late 1960s and early 1970s governments in many industrial areas adopted "technology forcing" regulations designed to compel coatings users to find ways to reduce VOC emissions. These regulations stimulated research on alkyds during the 1970s and 1980s. Two major approaches were used (→ Paints and Coatings): replacing most of the solvent with water (water-borne coatings), or sharply increasing the proportion of coatings solids to solvent (high-solids coatings). It is estimated that 10 % of alkyd usage is now of the water-borne and high-solids types. The most common approach to water-borne alkyds is to make special alkyds (see below), dissolve them in water-miscible organic solvents, and then add water. These are called "water-reducible" resins. Additional approaches have been introduced more recently, such as solvent-free dispersions of alkyd resins in water. All approaches require substantial changes in conventional alkyd resin technology.

Water Reducible Resins. To make alkyd resins suitable for use in water-reducible paints, it is necessary to chemically change them so they are miscible with blends of water and limited amounts of organic solvent. This is accomplished by attaching hydrophilic sites to the polymer molecules. Numerous procedures for doing so have been published [66–68].The most common method is to place acidic carboxy groups on the molecule and then to "salt" the resin with a base, usually ammonia or an amine. A representative method uses a two-stage process involving trimellitic anhydride [69] (mass parts in parenthesis):

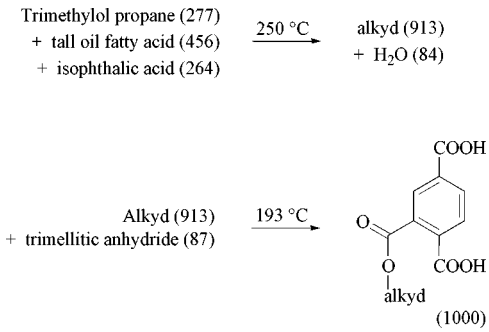

Reduced temperature is used in the second stage to encourage selective reaction of the anhydride group. However, there are several competing side reactions: at 193 °C the anhydride reaction is not completely selective, and acid – ester interchange reactions occur to some degree. For this reason, the process requires careful control to assure consistent product.

If enough carboxyl groups are present to give an acid value of about 45 mg of KOH per gram of resin or higher, the resin forms a transparent mixture when salted and dissolved in 20:80 solvent:water medium. Such resins are sometimes called "water-soluble" alkyds, but their abnormal solution rheology, light scattering, and other evidence indicate that they are seldom, if ever, true solutions. The terms "water-reducible", "water-thinnable", or "water-dilutable" are more descriptive.

Similar resins having acid numbers below about 45 form hazy or milky dispersions when salted and thinned with water. They are often termed "water-dispersible" as well as "water-reducible" alkyds. Such resins are gaining commercial importance because lowering the acid number seems to improve corrosion resistance of the coatings.

Most types of modified alkyds can be made in water-reducible form. For example, water-reducible acrylated [70], silicone-modified [71], and urethane-modified [72] alkyds have been developed.

A disadvantage of this approach is that the ester bond that connects trimellitic anhydride residues to the alkyd resin is more vulnerable to hydrolysis than the ester bonds in most polyesters. For this reason, stability of water-borne paints containing such resins is limited to 3 – 6 months. If a substantial fraction of the crit-

ical ester bond is hydrolyzed, the paint becomes unstable and unusable.

Numerous other approaches to formulating alkyds in aqueous media have been studied, often in efforts to improve package stability. Among the more promising are to copolymerize hydrophilic poly(ethylene oxide) segments into the resin [73], to emulsify hydrophobic alkyds with conventional surfactants [74], to prepare acrylated alkyds in which the acrylic polymer segments contain –COOH or other hydrophylic groups, and to prepare transparent microemulsions [75]. These approaches can be combined. The emulsification approach has been widely studied in recent years [76], [77], and is the basis for commercial products. Careful selection of emulsifiers, pigments, and additives is critical to assure stability and acceptable application characteristics and to avoid adverse interactions that deactivate driers [77].

High-Solids Coatings. There seems to be no standard definition of the term "high-solids coatings". The level of 70 vol % solids has been suggested as a minimum for high solids, but such a definition is arbitrary and does not conform to general usage. In this article, the term is used in a vague but useful way: a high-solids coating is one that is substantially more concentrated than similar coatings that were in general use in the 1960s. For example, an appliance finish that is sprayed at 62 vol % solids is called a high-solids coating because earlier generations were sprayed at 34 %. To emphasize the relative aspect of the definition, some authors prefer the term "higher-solids coatings".

The essential technical problem in formulating high-solids coatings is to find ways to achieve low viscosity without using very much volatile organic solvent. Many formulation and application variables can be manipulated in pursuit of this goal, but the most important is resin viscosity [78]. Therefore, the challenge to the alkyd resin chemist is to reduce the resin viscosity in solution without serious adverse effect on coating properties. To accomplish this difficult task it is necessary to reduce the molecular mass of the resin while increasing its functionality for cross-linking. Especially important is reducing the proportion of soluble high molecular mass molecules in the polymer mixture. These molecules contribute disproportionately to the viscosity [79]; unfortunately, their removal may harm film properties [46]. These difficulties have, to a degree, been overcome, and high-solids alkyd coatings are gaining commercial acceptance.

In air-dry, high-solids alkyds, the most difficult problem is to attain commercially acceptable drying rates without excessive yellowing. Resins that autoxidize very rapidly can be made by use of conjugated or highly unsaturated fatty acids. However, coatings made with such resins tend to turn yellow quickly and to cure too quickly at the surface of the coating film. Surface cure creates an oxygen barrier film that unacceptably slows cure in the interior of the coating. Solutions proposed to solve this problem include use of reactive diluents [80–82], use of auxiliary cross-linkers [83], and use of special autoxidation catalysts [84]. In each case, specially tailored alkyd resins are required.

Alkyds also are used in high-solids baking enamels, where they usually are cross-linked with amino resins, and in two-package alkyd – isocyanate enamels. In both of these applications, alkyds compete with polyester and acrylic high-solids resins.

Ryer presents an overview of new alkyd chemistry and gives several specific formulas [85]. For example, a chain-stopped alkyd suitable for coatings for agricultural and heavy equipment is composed of 72.8 parts (by weight) of soya fatty acid, 44.9 parts of pentaerithritol, 40.0 parts of phthalic anhydride, and 24.4 parts of benzoic acid. Given that the water of reaction is 13.1 parts, one can use Equation (1) to calculate the oil length of this alkyd as 44.8.

Compliance Solvents. In the USA, three solvents, acetone, methyl acetate, and 4-chlorobenzotrifluoride, have been exempted from regulations limiting VOC emissions because of their very low photochemical reactivity. Exemption of one or two more solvents, including *tert*-butyl acetate, is expected. These regulatory changes are very favorable for the future of alkyds, since resourceful formulators can use them to help satisfy air-pollution regulations that limit VOC. Two chlorinated organic solvents, CH_3CCl_3 and CH_2Cl_2, were previously exempted [87] but are little used because of concerns about their safety, corrosiveness, and greenhouse effects.

Latex Modifiers. A few percent of alkyd resin is often added to water-borne architectural paints, the principal binder of which is a polyvinyl or polyacrylic emulsion resin. The alkyd improves adhesion to chalky or oily surfaces. Addition of much larger proportions of alkyd, up to 50 %, is feasible [74], but incentives to commercialize this technology have been lacking because latices have been less expensive than alkyd resins most of the time since 1970.

6. Markets and End Uses

U.S. Tariff Commission states that alkyd resin production in the USA was about $(3-3.5) \times 10^5$ t/a between 1960 and 1980 and then began a slow decline, reaching about 2.3×10^5 t/a by 1996 [25]. The decline is thought to be continuing at a rate of about 2 %/a. Worldwide production figures are not readily available. If the proportion of alkyd used in the rest of the world is the same as in the USA, gobal production would be roughly three times as great as the U.S. figures. The proportion appears to be higher in some major markets, and this suggests that global production may be on the order of 10^6 t/a.

A very high percentage of alkyd production is used in coatings (→ Paints and Coatings). In this market alkyds compete with many other generic types of resins. While they are gradually being displaced by other types of resins, alkyds continue to compete effectively in many end uses. The continuing popularity of alkyds seems to be based on three factors:

1) *Versatility.* Whereas alkyds cannot match the outstanding features of some competing resin types (hardness and weatherability of acrylics, corrosion resistance and adhesion of epoxies, and toughness and abrasion resistance of polyurethanes), they can be formulated to provide a satisfactory balance of properties for many purposes.
2) *Ease of Use.* Well formulated alkyd coatings tend to be easier to apply by spray, brush, or roller than other generic types. They seem less susceptible to popping, sagging, cratering, and other appearance defects. They wet oily or dirty surfaces relatively well. Alkyds are package stable and do not have to be formulated in two-package systems. Air-dry formulations dry quickly, and baking enamels cure at low temperatures. An important but often overlooked factor in the continuing popularity of alkyds is that many existing finishing facilities and procedures have been designed for alkyds, and it has proved difficult to substitute other types of coatings without expensive capital investments.
3) *Economy.* During the 1960s and early 1970s petrochemical monomers, such as acrylics and vinyls, were less costly than the mixture of petrochemical and biomass ingredients used in alkyds. By 1983 the situation had reversed; soybean oil and tall oil fatty acids were selling at less than half the price of basic acrylic monomers. The cost advantage of alkyds is diluted by the petrochemical content of alkyd resins, which averages roughly 50 %, and the need, in most cases, to use petrochemical-based solvents. Since 1983, the relative costs of alkyd resins and 100 % petrochemical-based resins have fluctuated roughly in proportion to the prices of petroleum and of vegetable oils. Soybean oil currently sell for about 0.55 $/kg in the USA, while key vinyl and acrylic monomers (including styrene) sell for 1.5 to 3.5 times as much.

A summary of where alkyds fit in each of the three major coatings market segments follows.

Trade sales, or *architectural coatings*, comprise roughly 52 % of the physical volume and 42 % of the value of coatings sold in the USA. It is estimated that 55 % of the total alkyd usage in paints is in this sector [17][25]. Water-borne formulations based on polyacrylic and polyvinyl emulsion resins have captured over three-fourths of this market in the USA. Their chief advantages are easier cleanup, and better exterior durability. Alkyd- and alkyd – vegetable oil-based paints remain popular with many coating contractors because they can be applied faster and at lower temperatures; emulsion resin paints do not form films properly below about 5 – 8 °C, but alkyds can be used at 0 °C or even lower.

Alkyd-based paints retain a major share of certain trade sales markets, for example, stains, varnishes, high-gloss enamels, and primers. In these market segments, the hard, tough, cross-

linked films of alkyds offer advantages over latices, and their superior adhesion is important.

As noted above, a few percent of alkyd is often added to acrylic emulsion paint formulations to improve adhesion.

Industrial coatings are, by definition, coatings applied at the factory to manufactured goods. They account for roughly 31 % of the volume and 36 % of the value of coatings sold in the USA About 21 % of the alkyds used in paints go into these markets [17], [25].

Alkyd – melamine automotive topcoats have been almost completely replaced by more weatherable acrylic – melamine – formaldehyde and acrylic – polyurethane topcoats. Alkyds still find uses in auto and truck production, for example, in coatings for under-hood parts.

Factory wood finishing is a large user of coatings. Alkyds are used extensively both in alkyd – aminoplast thermosetting enamels and as plasticizers for nitrocellulose lacquers.

Other large-scale manufacturing operations that use alkyds include the following:

1) Metal containers – exterior can finishes
2) Coil coatings – flexible enamels for light fixtures, siding, Venetian blinds, etc.
3) Metal office furniture – hard alkyd – aminoplast enamels
4) Farm implements – fast air-drying enamels

Alkyd coatings are popular in smaller scale factory finishing operations, often because of their ease of application. Examples are tools, fixtures, drums, toys, and shelving. These diverse operations often are called the *general industrial* coatings market; it is a substantial consumer of alkyds.

Special purpose coatings include all organic coatings that do not fit into the above categories. It is a diverse market that is often characterized by sophisticated, specialized coatings. It accounts for the remaining 17 % of the volume and 22 % of the value of U.S. coatings. About 24 % of U.S. alkyd production goes to this segment [17], [25]. Alkyds and modified alkyds share this market with polyacrylic, epoxy, and polyurethane resins. Alkyds are used wherever they are satisfactory because in most cases they are the least expensive and easiest to apply. There are many examples:

1) Highway maintenance – alkyds, thixotropic alkyds, and epoxy esters for bridges, etc.
2) Traffic-control striping – usually very fast drying chain-stopped or styrenated alkyds blended with chlorinated rubber resin; a large market
3) Auto and truck refinish – epoxy ester primers and chain-stopped alkyd topcoats; alkyds losing share to acrylics
4) Marine – alkyds used mainly above the water line; fish oil alkyds popular for rust penetration, silicone alkyds for gloss retention
5) Industrial maintenance – linseed and silicone alkyds for refineries, tanks, factories, etc.

Noncoatings uses. The ink industry is a significant user of alkyd binders, mainly of the longoil isophthalic type. Other uses include caulks, adhesives, sealants, and linoleum. At one time alkyds were used as binders in fiberglass-reinforced plastics, but they have been supplanted almost completely by unsaturated polyesters.

7. References

General References

1. R. H. Kienle: "Alkyd Resins, Development of and Contribution to Polymer Theory", *Ind. Eng. Chem.* **41** (1949) 726 – 729.
2. *The Chemistry and Processing of Alkyd Resins,* Monsanto Co., St. Louis, Mo., 1952.
3. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y., 1953.
4. A. R. H. Tawn: "Some Problems in the Fundamental Study of Alkyd Resins", *J. Oil Colour Chem. Assoc.* **39** (1956) 223 – 252.
5. C. R. Martens: *Alkyd Resins,* Reinhold Publ. Corp., New York 1961.
6. T. C. Patton: *Alkyd Resin Technology: Formulating Techniques and Allied Calculations,* Wiley-Interscience, New York 1962.
7. R. P. A. Sims, W. H. Hoffman, W. J. Steward in W. O. Lundberg (ed): *Autooxidation and Antioxidants,* **vol. II,** J. Wiley & Sons, New York 1962, pp. 683 – 694.
8. J. B. Blegen, W. R. Fuller: *Alkyd Resins,* Unit Five of Federation Series on Coatings Technology, Federation of Societies for Paint Technology, Philadelphia 1967.
9. R. A. Myers, J. S. Long (eds.): *Treatise on Coatings,* **vol. 1,** pt. 1, Marcel Dekker, New York 1967.

10. F. W. Billmeyer, Jr.: *Textbook of Polymer Science,* 2nd ed., J. Wiley & Sons, New York 1971.
11. M. W. Formo in D. Swern (ed.): *Bailey's Industrial Oil and Fat Products,* 4th ed., **vol. I,** J. Wiley & Sons, New York 1979, pp. 687 – 816.
12. T. C. Patton: *Paint Flow and Pigment Dispersion,* 2nd ed., J. Wiley & Sons, New York 1979.
13. *Annual Book of ASTM Standards, Paints, Related Coatings and Aromatics,* **vol. 6.** 01, 6.02, and 6.03, ASTM, Philadelphia (revision issued annually).
14. K. Holmberg, *High Solids Alkyd Resins,* M. Dekker, New York, 1987.
15. A. R. Marrion (ed.): *The Chemistry and Physics of Coatings,* Royal Society of London, London, 1994
16. D. Stoye, W. Freitag, *Resins for Coatings,* Hanser, Munich, 1996, pp. 60 – 80.
17. Z. W. Wicks, Jr., F. N. Jones, S. P. Pappas, *Organic Coatings Science and Technology,* 2nd Ed., Wiley, New York, 1999.

Specific References

18. J. Berzelius, *Rapp. Annu. Inst. Geol. Hong.* **26** (1847).
19. W. Smith, *J. Soc. Chem. Ind. London* **20** (1901) 1075 – 1076.
20. R. H. Kienle, *Ind. Eng. Chem.* **41** (1949) 726 – 729.
21. General Electric, US 1893873, 1933 (R. H. Kienle).
22. R. H. Kienle, C. S. Ferguson, *Ind. Eng. Chem.* **21** (1929) 349 – 352.
23. W. H. Carothers, *J. Am. Chem. Soc.* **51** (1929) 2548 – 2559.
24. R. H. Kienle, A. G. Hovey, *J. Am. Chem. Soc.* **51** (1929) 509 – 519; **52** (1930) 3636 – 3645.
25. K. McNally et al., *The Coatings Agenda America 1999/2000,* Campden, London, 1999, pp. 83 – 84.
26. *Annual Book of ASTM Standards, Paints, Related Coatings and Aromatics,* **vol. 6.01** § D 16 – 82 a ASTM, Philadelphia 1983, p. 13.
27. DIN 55945 (1968).
28. D. Stoye, W. Freitag, *Resins for Coatings,* Hanser, Munich, 1996, pp. 60 – 80.
29. *Organic Coatings Science and Technology,* 2nd Ed., Wiley, New York, 1999, p. 268.
30. N. O. V. Sonntag in D. Swern (ed.): *Bailey's Industrial Oil and Fat Products,* 4th ed., **vol. I,** J. Wiley & Sons, New York 1979, pp. 289 – 477.
31. *Annual Book of ASTM Standards, Paints, Related Coatings and Aromatics, D,* **vol. 6.03,** ASTM, Philadelphia 1959, pp. 282 – 285
32. C. Williams, *Inform* **11** (2000) 580 – 588.
33. W. H. Carothers, *Trans. Faraday Soc.* **32** (1936) 38.
34. R. H. Kienle, P. A. Vander Muelen, F. E. Petke, *J. Am. Chem. Soc.* **61** (1939) 2258, 2268 – 2271.
35. R. H. Kienle, F. E. Petke, *J. Am. Chem. Soc.* **62** (1940) 1053 – 1056; **63** (1941) 481– 484.
36. P. J. Flory, *J. Am. Chem. Soc.* **63** (1941) 3083, 3091, 3096 – 3100.
37. P. J. Flory, *Chem. Rev.* **39** (1946) 137 – 197.
38. W. H. Stockmayer, *J. Chem. Phys.* **11** (1943) 45; **12** (1944) 125.
39. W. H. Stockmayer, *J. Polym. Sci.* **9** (1952) 69; **11** (1953) 424.
40. L. C. Case, *J. Polym. Sci.* **26** (1957) 333 – 350.
41. M. Jonason, *J. Appl. Polym. Sci.* **4** (1960) 129 – 140.
42. H. S. Lilley, *J. Appl. Polym. Sci.* **5** (1961) no. 16, S16 – S18.
43. E. G. Babolek, *Am. Paint J.* **42** (15 Sept. 1958).
44. E. G. Babolek, E. R. Moore, S. S. Levy, C. C. Lee, *J. Appl. Polym. Sci.* **8** (1964) 625 – 657.
45. H. Hata, J. Kumoanotani, Y. Nishizawa, H. Tomita, *XIVth FATIPEC Congr.* 1978, 358 – 364.
46. J. Kuomanotani, H. Hata, H. Masuda, *Proc. VIIIth Intl. Conf. Org. Coatings Sci. Technol.,* Athens 1982, 123 – 144.
47. C. M. Bruneau, *XVth FATIPEC Congr.,* Amsterdam 1980, vol. I, 123 – 144.
48. *Alkyd Report No. 1.2,* Hercules, Wilmington, Del.
49. *Alkyd Report No. 1.3,* Hercules, Wilmington, Del.
50. A. G. Hovey, *Ind. Eng. Chem.* **41** (1949) 730 – 737.
51. W. S. Armstrong, B. F. Coe, *Chem. Eng. Prog.* **79** (1983) 56 – 61.
52. H. Wexler, *Chem. Rev.* **64** (1964) 591 – 611.
53. J. H. Hartshorn, *J. Coat. Technol.* **54** (1982) 53 – 61.
54. J. Scheiber, *Farbe Lack* **63** (1957) 443.
55. N. A. Falla, *J. Coat. Technol.* **64** (1992) no. 815, 55 – 60.
56. J. C. Hubert, R. A. M. Venderbosch, W. J. Muizebelt, R. P. Klaasen, K. H. Zabel, *J. Coat. Technol.* **69** (1997) no. 869, 59 – 64.
57. W. J. Muizebelt, J. J. Donkerbroek, M. W. F. Nielen, J. B. Hussern, M E. F. Biemond, R. P. Klaasen, K. H. Zabel, *J. Coat. Technol.* **70** (1998) no. 876, 83 – 93.

58. W. J. Muizebelt, J. C. Hubert, M. W. F. Nielen, R. P. Klaasen, K. H. Zabel, *Prog. Org. Coat.* **40** (2000) 121 – 130.
59. R. W. Hein, *J. Coat. Technol.* **71** (1999) no. 898, 21 – 25.
60. Lewis Berger & Sons, DE 975352, 1949.
61. E. F. Redknap, *J. Oil Colour Chem. Assoc.* **43** (1960) 260.
62. Rohm & Haas, DE 912752, 1951; DE 1002 381, 1953.
63. Dow Corning, Midland, Mich., Bulletin 03-003.
64. T. F. Washburn Co., US 2633649, 1952.
65. J. S. Fry, C. N. Meriam, G. J. Misko, *Am. Chem. Soc. Div. Org. Coat. Appl. Polym. Sci. Preprints* **47** (1982) 540 – 548.
66. R. R. Engelhardt, *Double Liaison Chim. Peint.* **27** (1980) no. 296, 207 – 213; *Pigm. Resin Technol.* 1979 (Aug.) 5 – 15.
67. F. Sheme, S. Belote, L. Gott, *Polym. Paint Colour J.* 1975, 784 – 790.
68. Bayer, DE 2837552, 1979; US 4321 170, 1982 (K. Zabrocki et al.).
69. Amoco Chemicals Corp., Tech. Bull. TMA-1086 (1981).
70. E. Levine, E. J. Kuzma, *J. Coat. Technol.* **51** (1979) no. 657, 35 – 48.
71. K. Fey, W. Finzel, *Am. Paint Coat. J.* 1982 (June 28) 41 – 45.
72. J. C. Laout, J. F. May, C. Nicaud, *XVth FATIPEC Congr.,* Amsterdam 1980, vol. II, 227-246.
73. Ashland Oil & Refining Co., US 3442835, 1969 (G. M. Curtice, D. D. Taft).
74. B. G. Bufkin, J. R. Grawe, *J. Coat. Technol.* **50** (1978) no. 647, 65 – 96.
75. Perstorp AB, Report 1978 (S. E. Friberg, E. G. Gillberg-LaForce, K.-H. Falklin).
76. B. Bergenstahl, A. Hofland, G. Ostberg, A Larsson in C. Salamone (ed.): *Polymeric Materials Encyclopedia,* CRC Press, Boca Raton, FL, 1996, pp. 154 – 160.
77. P. K. Wissenborn, A. Motiejauskaite, *Prog. Org. Coat.,* **40** (2000) 253 – 266.
78. L. W. Hill, Z. W. Wicks, Jr., *Prog. Org. Coat.* **10** (1982) no. 1, 55 – 89.
79. K. Holmberg, J.-A. Johansson, *Proc. VIIIth Int. Conf. Org. Coatings Sci. Technol.,* Athens 1982, 255 – 275.
80. D. B. Larson, W. D. Emmons, *Proc. VIIIth Int. Conf. Org. Coatings Sci. Technol.,* Athens 1982, 193 – 216.
81. S. Enomoto et al., *J. Appl. Polym. Sci.* **22** (1978) 252 – 265.
82. E. Levine, E. J. Kuzma, M. T. Nowak, *Mod. Paint Coat.* **66** (1976) no. 8, 23 – 30.
83. D. J. Love, *J. Coat. Technol.* **53** (1981) no. 680, 55 – 58.
84. V. Verkholantsev, *Eur. Coat. J.* (2000) no. 1/2, 76 – 83.
85. D. Ryer, *Paint and Coatings Industry* **14** (1998) no. 1, 76 – 83.
86. A. Heitkamp, J. Hamre, *J. Water Borne Coat* **.** **6** (1983) no. 2, 20 – 28.
87. F. N. Jones, C. P. Alexander, J. M. Larson, *Mod. Paint Coat.* **71** (1981) no. 8, 54 – 56.

Alkyl Amines → Amines, Aliphatic
Alkylation → Acylation and Alkylation
Alkyl Phenols → Phenol Derivatives
Allyl Alcohol → Allyl Compounds

Allyl Compounds

Ludger Krähling, Deutsche Solvay-Werke GmbH, Rheinberg, Federal Republic of Germany (Chaps. 1, 6)

Jürgen Krey, Deutsche Solvay-Werke GmbH, Rheinberg, Federal Republic of Germany (Chaps. 1, 6)

Gerald Jakobson, Deutsche Solvay-Werke GmbH, Rheinberg, Federal Republic of Germany (Chaps. 1, 6)

Johann Grolig, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 2 – 5)

Leopold Miksche, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 6)

1. Allyl Chloride 1
1.1. Physical Properties 1
1.2. Chemical Properties 2
1.3. Production 3
1.3.1. Chlorination of Propene 3
1.3.2. Other Production Processes 5
1.4. Handling, Environmental Protection, Storage, and Transportation . 6
1.5. Quality and Analysis 7
1.6. Uses 7
1.7. Economic Aspects 7
2. Allyl Alcohol 7
2.1. Physical Properties 8
2.2. Chemical Properties 8
2.3. Production 9
2.3.1. Hydrolysis of Allyl Chloride 9
2.3.2. Isomerization of Propene Oxide 9
2.3.3. Hydrolysis of Allyl Acetate 10
2.3.4. Hydrogenation of Acrolein 10
2.4. Quality and Analysis 11
2.5. Uses 11
2.6. Methallyl Alcohol 11
3. Allyl Esters 11
3.1. Properties 11
3.2. Production 12
3.2.1. Oxidation of Olefins 12
3.2.2. Esterification 13
3.2.3. Transesterification 13
3.2.4. Other Production Methods 13
3.3. Uses 14
3.3.1. Polymer Production 14
3.3.2. Other Uses 15
4. Allyl Ethers 15
4.1. Properties and Uses 15
4.2. Production 16
5. Allylamines 16
5.1. Properties and Uses 16
5.2. Production 17
6. Toxicology and Occupational Health 17
7. References 19

1. Allyl Chloride

Allyl chloride [*107-05-1*], the only chloropropene of industrial importance, was first produced in 1857 by A. Cahours and A. W. Hofmann by reacting phosphorus chloride with allyl alcohol. The name allyl is derived from the latin *allium,* meaning garlic. Inhalation of even small amounts of allyl chloride produces, after a short time, the characteristic odor of garlic on the breath.

At the end of the 1930s, IG Farbenindustrie and the Shell Development Co. developed the high-temperature chlorination of propene, permitting large-scale production of allyl chloride with good yields. A significant part of the development was done by the Shell Chemical Co. when erecting a commercial plant in 1945. Dow, Solvay, and Asahi-Kashima developed their own processes.

1.1. Physical Properties

Allyl chloride, 3-chloropropene, CH_2=CH-CH_2Cl, is a colorless, mobile liquid with a penetrating, pungent odor, M_r 76.53, *mp* −134 °C, *bp* (101.3 kPa) 44.4 °C, n_D^{20} 1.416, n_D^{25} 1.413 [1–5]. Temperature-dependent physical data are given in Table 1.

Expansion coefficient (0 – 30 °C)	0.0041 K^{-1}
Heat of combustion (CH_2=CH-CH_2Cl (g), HCl (g), CO_2 (g), H_2O (g))	1846 kJ/mol
Latent heat of vaporization at *bp*	26 kJ/mol
Critical temperature, t_{crit}	240 °C
Critical pressure, p_{crit}	4.79 MPa
Ratio of specific heat capacities c_p/c_v (at 14 °C)	1.137

Vapor pressure equation

$$\log p = 19.1403 - 2098.0/T - 4.2114 \cdot \log T;$$

$(p \,\mathrm{in\,kPa},\ T \,\mathrm{in\,K})$ [3].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_425

Table 1. Temperature dependence of the physical properties of allyl chloride

	Temperature, t, °C						
	10	15	20	25	30	40	50
Density, ϱ_4, g/cm^3	0.950	0.944	0.938	0.931	0.925	0.911	0.898
Specific heat capacity (of liquid), c_p, kJ kg^{-1} K^{-1} at 101.3 kPa	1.633		1.666		1.700	1.733	1.771
Viscosity, η, μPa · s	368	347	336	315	307	282	
Surface tension, mN/m		28.9	23.1		21.8		
Solubility in 15 % hydrochloric acid, wt%				0.120			0.176

Allyl chloride is miscible with most solvents in general use (e.g., octane, toluene, acetone); the solubility (mass fraction in %) of allyl chloride in water at 20 °C is 0.36, that of water in allyl chloride, 0.08. Azeotropic data for allyl chloride are presented in Table 2.

Table 2. Azeotropes of allyl chloride

Component	*bp* at 101.3 kPa, °C	Allyl chloride, mass fraction, %
Water	43	97.8
Methanol	40	90
Ethanol	44	95
2-Propanol	45	98
Formic acid	45	92.5

Flash point (closed cup)	− 27 °C
Autoignition point	392 °C
Flammability limits in air	3.28 and 11.15 vol%

1.2. Chemical Properties

Allyl chloride is a very reactive compound undergoing the usual addition and polymerization reactions at the double bond. Also, because the chlorine atom can be exchanged readily with other groups, allyl chloride is a suitable starting material for the synthesis of a wide range of allyl derivatives.

Reactions of the Double Bond. The reaction of allyl chloride with oxygen in the liquid phase at ca. 120 °C and in the presence of metal acetates or hydrogen peroxide yields glycerol monochlorohydrin. Halogens add readily to the double bond yielding the corresponding trihalogeno compounds. The reaction with hypochlorous acid, yielding 2,3- and 1,3-glycerol dichlorohydrins (which are then dehydrochlorinated with alkali to give epichlorohydrin), is of great industrial importance. Allyl chloride reacts with hydrogen halides to form 1,2-dihalogeno compounds. In the presence of peroxides, the reaction with hydrobromic acid yields 1-bromo-3chloropropane (Kharasch effect), but in highly concentrated hydrogen peroxide solution, 1,2-dibromo-3-chloropropane is formed [6]. Addition reactions of silanes [7], boranes [8], carboranes [9], and phosphorus trichloride [10], as well as cycloadditions of allyl cations with alkenes [11], are known. Allyl chloride polymerizes with sulfur dioxide to form polysulfones [12].

Reactions of the Chlorine Atom. Formerly, the most important reaction industrially was hydrolysis to allyl alcohol, in which small amounts of diallyl ether formed as byproduct. However, today allyl alcohol is produced increasingly by the isomerization of propene oxide (see Section 2.3.2). The chlorine atom is replaced easily by iodide, cyanide, isothiocyanate, sulfide, polysulfides (giving rubber-like condensation products [13]), and alkyl thiols [14]. The salts of carboxylic acids yield allyl esters (e.g., diallyl phthalate), which are easily polymerized to allyl resins or are copolymerized with other monomers. Sodium allyl sulfonate is obtained on reaction with sodium sulfite [15]. The reaction of allyl chloride with ammonia yields a mixture of mono-, di-, and triallylamines [16], [17], and the reaction with primary and secondary amines makes the corresponding alkylallylamines [18]. Organic polycarbonates are formed from a mixture of tertiary amines, sodium carbonate, alkylene glycol, carbon dioxide, and allyl chloride [19]. Phase-transfer-catalyzed carbonylation (e.g., with $Ni(CO)_4$

or $(Me_4N)_2[Ni_6(CO)_{12}]$) in the presence of sodium hydroxide leads to the sodium salt of vinylacetic acid [20], [21]. In the presence of alkalis, allyl chloride reacts with polyols to form the corresponding allyl ethers. Further reactions are described in [22].

1.3. Production

1.3.1. Chlorination of Propene

Allyl chloride is produced on a large scale by the high-temperature (300 – 600 °C) chlorination of propene:

$$CH_2{=}CH\text{-}CH_3 + Cl_2 \rightarrow CH_2{=}CH\text{-}CH_2Cl + HCl \qquad \Delta H^0_{298} = -113\ \text{kJ/mol}$$

At these temperatures, chlorination proceeds by a free-radical chain mechanism, whereby the hydrogen atom in the allyl position is substituted preferentially by the chlorine, giving allyl chloride [23].

Byproducts. Below 200 °C propene reacts with chlorine mainly by addition to the double bond to give 1,2-dichloropropane; above 300 °C, this reaction is suppressed and the formation of allyl chloride predominates so that 1,2-dichloropropane is only a byproduct. The compounds *cis*- and *trans*-1,3-dichloropropene arise from a secondary reaction of allyl chloride, in which a further hydrogen atom is substituted by chlorine. Small amounts of other chlorination products are formed also:

$$CH_3{-}CH{=}CH_2 + Cl_2 \rightarrow CH_3{-}CHCl{-}CH_2Cl \qquad \Delta H^0_{298} = -184\ \text{kJ/mol}$$

$$ClCH_2{-}CH{=}CH_2 + Cl_2 \rightarrow CH_2Cl\text{-}CH{=}CHCl + HCl \qquad \Delta H^0_{298} = -101\ \text{kJ/mol}$$

$$CH_3{-}CH{=}CH_2 + Cl_2 \rightarrow CH_3{-}C(Cl){=}CH_2 + HCl \qquad \Delta H^0_{298} = -121\ \text{kJ/mol}$$

$$CH_3{-}CH{=}CH_2 + Cl_2 \rightarrow CH_3{-}CH{=}CHCl + HCl$$

The most important variables in the industrial chlorination process of propene to allyl chloride are the temperature and the ratio of propene to chlorine, whereas pressure and residence time have only a slight effect on the allyl chloride yield [24]. Because the dominant reaction below 200 °C is addition to form 1,2-dichloropropane, the mixing temperature of propene and chlorine must be kept above 250 – 300 °C. The best yields of allyl chloride in industrial reactors are obtained at a maximum reaction temperature of 500 – 510 °C. If the reactor temperature is increased further, spontaneous pyrolysis occurs, with the formation of soot and high-boiling tars. Under laboratory conditions and temperatures of ca. 600 °C, benzene is formed and the yield of allyl chloride decreased [24], [25].

The *maximum reaction temperature* can be influenced by the temperatures of propene and chlorine entering the reactor and by the ratio of propene to chlorine. In general, only the propene is preheated. If the chlorine is also preheated, expensive construction materials must be used to avoid the danger of a "chlorine fire." For a given propene: chlorine ratio, the preheating temperature of the propene is fixed; for example, for $C_3H_6 : Cl_2 = 3$ the temperature is set at ca. 300 °C and for $C_3H_6 : Cl_2 = 5$, at ca. 400 °C [24].

Propene : Chlorine Ratio. The formation of byproducts decreases with increasing propene excess [26]. On the other hand, the cost of processing the propene rises simultaneously. The optimum reaction conditions are therefore strongly influenced by economic considerations, including the demand for the dichloride byproducts as nematocides [27]. The byproducts can be employed also as starting material for the production of C_1- or C_2-type solvents.

The *pressure* in the reactor has little influence on the yield or the distribution of products and is determined only by the pressure drop in the propene circulation system of the plant.

The *residence time* has only a small effect on the yield of allyl chloride. At high temperatures (300 – 600 °C), the chlorine has completely reacted after 1 – 3 s [24]. Too long a residence time, however, leads to thermal decomposition of the allyl chloride [25].

Reactor Construction. Industrial-scale reactors mainly operate adiabatically, even though higher yields would be expected when operating isothermally. Because the reaction is rapid and exothermic, the amount of heat lost through the reactor wall is not significant.

The technically simplest and oldest reactor type is the tube reactor, which sometimes is equipped with facilities for gas distribution and soot removal [25]. Many other reactor designs are known [26], [28–35], but all reactors are designed to achieve the mixing of the two reactants as rapidly and as thoroughly as possible in order to reduce the secondary reaction to form 1,3-dichloropropene. Under the good mixing conditions obtained in a highly turbulent flow, propene and chlorine are fed into the reaction zone at velocities up to 300 m/s. The importance of optimum mixing conditions is demonstrated by a cyclone reactor, in which, at a molar ratio of 3 : 1, similar yields are obtained as at a ratio of 5 : 1 with other designs [36]. Many reactors are equipped with a cooling jacket.

If the chlorine is distributed among several reactors arranged in a cascade, a chlorine conversion of up to 86 % can be achieved [37], [38]. With this arrangement, higher preheating temperatures are possible because of a higher propene : chlorine ratio in the first reactor.

In another design of a cascade of reactors only the first reactor is charged with preheated propene and chlorine. Liquid propene and gaseous chlorine (molar ratio of 1 : 1) are fed into the other reactors. The advantages of this procedure are the cooling of the reaction gas and a quasiconstant ratio of the reactants over all stages. At an overall molar ratio of 3.2 : 1, the same yield is achieved as at a ratio of 7 : 1 in a single reactor [39].

Feed Preparation. The *purity* of the propene and the chlorine is important. Organic impurities in the propene cause the formation of byproducts and loss of chlorine. Especially propane leads to the formation of chlorine derivatives (1-chloropropane, 2-chloropropane) which are difficult to separate. Polymer-grade propene with a purity of 99.5 % often is employed as the feedstock. The use of revaporized chlorine insures adequate purity. Small amounts of inert gases are thereby introduced into the process, the venting of which means propene losses. Both raw materials must contain as little water as possible. This is a particularly important condition for the choice of materials for the various stages of the plant (see below).

A process similar to that employed by Shell, for example [24], [40], is shown in Figure 1. Liquid propene is vaporized, then preheated to 350 – 400 °C (b), and fed, together with gaseous chlorine, into the reactor (c) via a mixing jet. The chlorine reacts completely, thereby increasing the temperature to 500 – 510 °C (under optimum conditions). Even under the best conditions, small amounts of carbon are formed; these catalyze the chlorination. A protective film of vitreous carbon deposits on the reactor walls. This material, which also contains highly chlorinated materials and tar, must be cleaned from the reactor walls at intervals of 4 – 8 weeks. Two parallel reactor chains often are in use so that partial production can be maintained during the cleaning. Another possibility is alternate operation of the two chains.

The gas stream leaving the chlorination reactor is precooled (d) and led to a prefractionator (e), the overhead temperature of which is maintained at ca. – 40 °C by feeding liquid propene. This effectively separates all chlorinated hydrocarbons; the bottom product is free of propene and hydrogen chloride.

The gaseous mixture drawn off overhead is separated by absorption with water (i) into aqueous hydrogen chloride of commercial quality and propene. Then the propene is washed with caustic soda in a scrubber (k) to remove traces of hydrogen chloride. After compression to 1.2 MPa (12 bar), it is liquefied in a condenser (l, q). Water is separated (m) and liquid propene is dried by adsorption (n) and returned to the storage tank (a).

The bottom product of the prefractionator contains 80 % allyl chloride, 3 % 2-chloro-1-propene [*557-98-2*] and other low boilers, 16 % dichlorides (mainly 1,2-dichloropropane [*78-87-5*] and *cis*- and *trans*-1,3-dichloro-1-propene [*542-75-6*]), and 1 % 1,2,3-trichloropropane [*96-18-4*] and other heavy boilers [40]. These four fractions are separated by distillation (f, g, h).

The basic principle of most production processes is as shown in Figure 1, although it has been modified for the synthesis and the fractionation processes [41–44]. For example, cooler (d)

in Figure 1 can be replaced by quench cooling [44], whereby the hot reaction gases are cooled by the evaporation of propene or of condensed reaction products.

Construction Materials. The choice of materials in the allyl chloride synthesis (propene circulation) depends on the individual process operations, temperature, and pressure. In principle, normal carbon steel is resistant where the water content of the product streams is low enough. If the chlorine stream is not preheated, carbon steel also can be utilized in the reactor area. However, materials resistant to chlorine at high temperatures (chromium – nickel steels, nickel, cupronickel alloys) often are employed in the reactor region. Graphite and PTFE coatings are favored for the absorption of the hydrogen chloride, whereas rubberized steel can be employed for the caustic soda scrubber. Carbon steel is suitable for the entire chlorinated hydrocarbon fractionation plant. However, in places when large amounts of water are present, expensive materials, such as nickel and cupronickel, are necessary [45]. Dry allyl chloride can be stored in steel vessels without any danger of corrosion.

1.3.2. Other Production Processes

Several other processes have been suggested for the production of allyl chloride, but none of these has been operated commercially.

Catalytic Chlorination of Propene. This process uses tellurium-containing catalysts [46]; allyl chloride yields of up to 82 % are obtained. The main organic byproduct is 17 % 2-chloropropane (isopropyl chloride). Unreacted propene and the hydrogen chloride produced in the reaction can be converted to allyl chloride in a second reaction step by oxychlorination.

The dehydrochlorination of 1,2-dichloropropane [47–49] produces only a 55 % yield of allyl chloride plus a large amount of monochloropropenes. The process is of no commercial interest even though the starting material is dichloropropane, produced in large quantities during the production of propene oxide.

Oxychlorination. Oxichlorination was developed for the production of allyl chloride in order to utilize cheaply available hydrogen chloride instead of chlorine as the feedstock [50–62]. Palladium, vanadium, tellurium, copper, lithium, and manganese, as well as their chlorides and oxides and mixtures thereof, have been suggested as catalyst systems.

Most of these oxychlorination processes start from propene [50–60], although two of them start from propane [61], [62]. The following process was tested in a pilot plant by Hoechst [58].

Propene, hydrogen chloride, oxygen, and 2-chloropropane (either produced in a subsidiary reactor or introduced from outside) react in a main fluid-bed reactor at 200 – 260 °C and 0.1 MPa (1 bar) gauge pressure ($\Delta H = -218$ kJ/mol). The carrier catalyst contains tellurium, vanadium pentoxide, phosphoric acid, and a nitrogen compound as promoter. A part of the catalyst flow is treated in a side stream with air and nitric acid to remove coke and maintain constant reactivity. In a subsidiary reactor, 2-chloropropane is produced from propene, hydrogen chloride, and ferric chloride solution. The unreacted propene, 2-chloropropane, and hydrogen chloride are separated and recycled. The yield of allyl chloride is 88 – 94 % based upon propene. The purity of the propene is not particularly critical.

Some of the suggested oxychlorination processes have disadvantages. For example, the activities of the catalysts deteriorate quickly as a consequence of the volatility of the metal salts employed and large volumes must be passed through the reactor because the conversion per pass is low. Also, it is difficult to remove the highly dilute allyl chloride from the reaction mixture without excessive loss of propene by reaction with oxygen. A suggestion [60] attempts to avoid these disadvantages by using manganese dioxide as both catalyst and carrier for oxygen. The following reactions take place in the main reactor:

$$MnO_2 + 4\,HCl \longrightarrow MnCl_2 + Cl_2 + 2\,H_2O$$

$$C_3H_6 + Cl_2 \longrightarrow C_3H_5Cl + HCl$$

The catalyst is then reoxidized and activated with oxygen:

$$MnCl_2 + O_2 \rightarrow MnO_2 + Cl_2$$

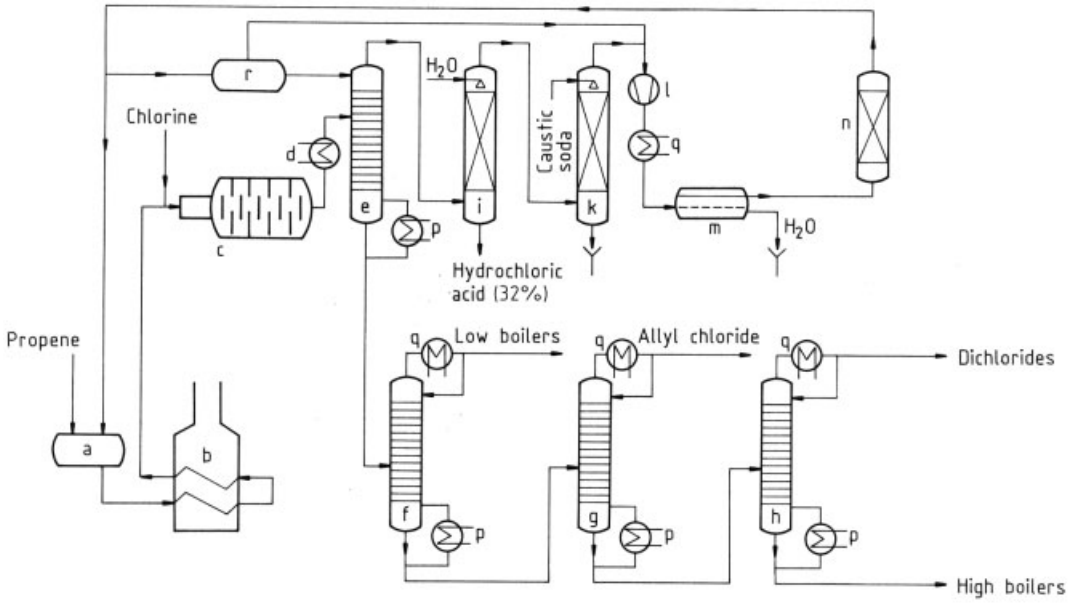

Figure 1. Allyl chloride production by high-temperature chlorination of propene
a) Storage vessel for liquid propene; b) Evaporator and superheater for propene; c) Reactor; d) Cooler; e) Prefractionator; f) Light-ends column; g) Allyl chloride purification column; h) Dichloropropene column; i) Hydrogen chloride absorber; k) Gas washer; l) Compressor; m) Decanter for removing water; n) Propene dryer; p) Evaporator; q) Condenser; r) Cold propene storage vessel

The process achieves an allyl chloride yield of between 71 and 81 % based on the consumption of propene.

1.4. Handling, Environmental Protection, Storage, and Transportation

Handling and Environmental Protection. Allyl chloride is a highly reactive, highly toxic, easily ignitable substance. Therefore, very stringent standards exist in many countries for atmospheric emission. Allyl chloride should be handled in closed systems in order to fulfill these requirements. Gases containing allyl chloride or byproducts of the production process must be purified by condensation in cold traps, by absorption, by adsorption, or in special cases by combustion before they are vented [63]. The compensation technique has proved of value during transfer from one vessel to another.

Any contamination of the soil, of underground water supplies, watercourses, or of wastewater must be avoided. Any danger of such contamination must be reported immediately to the responsible authorities.

Allyl chloride-containing wastes can be disposed of without difficulty in special incinerators, such as those used for the disposal of solvents, where the combustion gases are treated in absorbers or scrubbers to remove the hydrogen chloride formed.

Because of the high volatility and low flash point of allyl chloride, plants for its production and processing must comply with the relevant standards for fire and explosion protection.

Combustion of chlorinated hydrocarbons produces hydrogen chloride, so that fire fighters should be equipped with suitable protective suits and portable breathing apparatus. Combustion in the absence of adequate air supplies can lead to the formation of carbon monoxide.

Allyl chloride is highly reactive. Therefore contact with other substances can lead to vigorous, exothermic, and even explosive reactions. This applies particularly to alkali and alkaline-earth metals, but also to aluminum and zinc and to strong oxidizing agents, such as concentrated sulfuric acid. The anhydrous halides (e.g., chlo-

rides) of the metals mentioned above also react vigorously with allyl chloride [5]. In plants producing or processing allyl chloride, careful consideration must be given to the safety aspects of these exothermic reactions [64]. In 1996, allyl chloride has been classified by the EC as "dangerous for the environment" and "very toxic to aquatic organisms".

Storage. Pure, dry allyl chloride (water content $\leqq$ 200 mg/kg) does not corrode iron and can be stored in ferrous containers for months at ambient temperatures without significant changes in quality. Lined vessels should be used if traces of iron chlorides are objectionable. Unwanted coloration can be prevented by the addition of a suitable stabilizer (e.g., propene oxide).

Transportation of the highly toxic and highly flammable allyl chloride is governed by many regulations [65]. International marine transportation is governed by the IMDG Code, D 3108, class 3.1, UN no. 1100; transportation in Europe is governed by RID, ADR, and ADNR class 6.1, no. 4 a, Rn 601, 2601, and 6601, respectively. European Economic Community: Yellow Book 78/79, EG-no. 602–029–00-X; Great Britain: Blue Book: Flammable Liquid, IMDG Code E 3023; United States: CFR 49: 172 – 189, Flammable Liquid (DOT Regulations) [5].

The most important regulations state that:

1) Vessels may be filled only to a maximum of 93 % of their volumes or the RID (Suppl. X and XI) must be complied with.
2) Road transport is allowed (in Germany a maximum of 3 t per tank container).
3) No pipe ducts or pipe connections are allowed below the surface of the liquid. Filling and emptying connections must be tightly closed and should be additionally secured by means of sealed caps. The tanks must be tested regularly to insure that they withstand total pressures of 0.6 MPa (6 bar).
4) Welded drums constructed of steel or stainless steel must be used for transport in drums. Frequent changes in regulations are recorded. Each transport must be examined individually with regard to the countries concerned.

1.5. Quality and Analysis

Commercial allyl chloride is at least 97.5 wt% pure and contains 1-chloropropene, 1-chloropropane, and 1,5-hexadiene as impurities. In addition, it contains a stabilizer, such as propene oxide, to scavenge any hydrogen chloride produced hydrolytically on long-term storage. Allyl chloride containing traces of iron has a pinkish tinge. The quantitative analysis of allyl chloride is performed exclusively by gas chromatography.

1.6. Uses

Allyl chloride [22], [66] is an important intermediate in the petrochemical industry, used chiefly for the production of epichlorohydrin, which in turn is used in the production of epoxy resins. Allyl chloride is a starting material in the synthesis of various chemicals, such as glycerol, many esters [of which those of phthalic, phosphoric, and carboxylic acids are the most important (Section 3.2.4)] and some allyl ethers (Section 4.2) and allylamines (Section 5.2).

Further compounds made from allyl chloride are: *allyl isothiocyanate* (synthetic mustard oil), used in plant protection agents and pharmaceutical preparations; *allyl sulfonate,* used as an electroplating-bath additive [67] and in the production of carbon fibers [68]; *allylsilane,* used for the production of additives for the rubber industry; and *cyclopropane,* an anesthetic.

1.7. Economic Aspects

More than 90 % (ca. 800 000 t worldwide in 1997) of the allyl chloride produced is used for the production of epichlorohydrin. Less than 55 000 t worldwide went into other applications.

2. Allyl Alcohol

Allyl alcohol was first prepared in 1856 by A. Cahours and A. W. Hofmann by saponification of allyl iodide. For general literature on allyl alcohol, see [69–74].

2.1. Physical Properties

Allyl alcohol [*107-18-6*], 2-propen-1-ol, CH_2=CH-CH_2OH, C_3H_6O, M_r 58.08, *mp* − 129 °C, *bp* 96.9 °C (101.3 kPa), is a colorless, mobile liquid with an irritating odor.

Relative density,	d_4^{20} 0.8520;	d_4^{25} 0.8476;	d_{20}^{20} 0.8535
Refractive index,	n_D^{20} 1.4133;	n_D^{25} 1.4111;	n_D^{30} 1.4090
Critical temperature, t_{crit}			271.9 °C
Heat of vaporization at 101.3 kPa			39.98 kJ/mol
Specific heat capacity of vapor, c_p (g), (20 °C and 101.3 kPa)			2.428 kJ kg^{-1}K^{-1}
Specific heat capacity of liquid c_p (l), (20.5 – 95.5 °C)			2.784 kJ kg^{-1}K^{-1}
Heat of combustion at constant *p*			1853.8 kJ/mol

Vapor pressure vs. Temperature

t, °*C*	20	40	60	80
p, kPa	2.3	7.7	21.6	52.6

Viscosity	
at 15 °C	1.486 mPa · s
at 30 °C	1.072 mPa · s
Surface tension	
at 20 °C	25.68 mN/m
at 60 °C	22.11 mN/m
Dipole moment	1.63 D
Dielectric constant	
at 16.2 °C, λ = 60 cm	20.3
Ignition limits in air	
at 100 °C, 101.3 kPa	2.5 – 18.0 vol %
Flash point (closed cup)	22.2 °C

Allyl alcohol is miscible with water and organic solvents in all proportions at 20 °C. Table 3 contains the most important azeotropic data for binary and ternary allyl alcohol azeotropes.

2.2. Chemical Properties

The allylic hydroxyl group and the olefinic double bond can undergo numerous reactions, such as oxidation, reduction, hydrogenation, condensation (formation of ethers and esters), and addition. At room temperature, allyl alcohol is a stable liquid. When heated to temperatures above about 100 °C, allyl alcohol forms water-soluble polymers (polyallyl alcohols) that react with alkenoic acids to give drying oils [75].

Allyl alcohol also can be grafted to polyimides [76] or copolymerized with styrene in the presence of oxygen [77]. Condensation of allyl alcohol with methyl glucoside polyethers, followed by bromination and addition of isocyanates, yields flame-resistant polyurethane foams [78].

Hydroxylation. The reaction of allyl alcohol with hydrogen peroxide in the presence of catalytic amounts of tungstic acid yields glycerol with excellent selectivity (→ Glycerol). This commercially important process was developed by Shell [79]. The hydroxylation of allyl alcohol can be carried out also by organic hydroperoxides, such as ethylbenzene hydroperoxide, in the presence of a vanadium catalyst [80]. Organic hydroperoxides oxidize allyl alcohol selectively to glycidol (oxiranylmethanol) [*556-52-5*] [81]. Glycidol is also obtained by the reaction of allyl alcohol with peracids [82].

Allyl alcohol can be *dehydrogenated* with air in the gas phase over palladium, silver, or copper catalysts to give acrolein [83], [84]. The *oxidation* of allyl alcohol in the liquid phase (150 – 200 °C, Pd catalyst) yields acrylic acid and acrolein with 83 % combined selectivity [85].

Catalytic *hydrogenation* of allyl alcohol gives 1-propanol [*71-23-8*] [86]. *Chlorination* in aqueous solution yields mono- and dichlorohydrins of glycerol, which can be hydrolyzed to glycerol [87]. *Bromination* proceeds very selectively in aqueous solution, buffered by calcium chloride [88]; 2,3-dibromo-1-propanol [*96-13-9*], used in flame-resistant materials [89], is formed.

Allyl alcohol reacts with allyl chloride to give diallyl ether [*557-40-4*] (*bp* 94 °C) and with methallyl chloride to give allyl methallyl ether [*14289-96-4*]. Carbon tetrachloride adds to allyl alcohol by radical initiation, leading to 2,4,4,4-tetrachloro-1-butanol [*3290-70-8*], which is a useful intermediate in flame-retardant technology [90].

Hydroformylation of allyl alcohol in the presence of cobalt carbonyl yields 4-hydroxybutyraldehyde [*25714-71-0*] [91] in 30 % yield. With rhodium-complex catalysts, the 4-hydroxybutyraldehyde yield is improved to 80 % [92].

The aldehyde then can be hydrogenated to 1,4-butanediol (→ Butanediols, Butenediols, and Butynediols), a useful monomer for the production of polyurethanes [93]. When the hydroformylation of allyl alcohol is carried out in the vapor phase over rhodium triphenylphos-

Table 3. Azeotropes of allyl alcohol

Component		*bp*, °C	Allyl alcohol content, wt %	
Binary azeotropes				
Water		88.89	72.3	
Benzene		76.75	17.3	
Diallyl ether		89.8	30.0	
Allyl chloride		79.9	18	
Trichloroethylene		80.95	16	
Component I	Component II	*bp*, °C	Content, wt %	
			Allyl alcohol	Component II
Ternary azeotropes				
Water	benzene	68.2	9.1	83.6
Water	diallyl ether	77.8	8.7	78.9

phine complexes on porous carriers, tetrahydro-2-furanol [*5371-52-8*] is obtained in 94 % yield [94]. Carbonylation of allyl alcohol in acetic acid solution with a palladium chloride catalyst gives 3-butenoic acid [95].

2.3. Production

The raw material in all commercial processes is propene. However, although allyl alcohol can be made via a variety of intermediates, only the processes proceeding via allyl chloride and propylene oxide are of major commercial importance.

2.3.1. Hydrolysis of Allyl Chloride

Allyl alcohol is produced commercially by Shell and Dow from allyl chloride [24], [26]. Allyl chloride is hydrolyzed by a 5 – 10 % sodium hydroxide solution at 150 °C and 1.3 – 1.4 MPa, yielding 85 – 95 % allyl alcohol. Byproducts are diallyl ether (5 – 10 %), chloropropenes, propionaldehyde, and high-boiling material. In order to obtain commercially satisfactory yields of allyl alcohol, it is necessary to mix the allyl chloride thoroughly with the aqueous alkaline solution, maintaining a constant pH value. The allyl chloride has to be converted nearly quantitatively, because its high corrosiveness does not allow economical recovery. The reaction therefore is carried out in a recycle reactor; the most suitable reactor material is nickel.

A commercial plant for the production of allyl alcohol by continuous allyl chloride hydrolysis [24] is shown in Figure 2. Water is removed from the raw allyl alcohol by azeotropic distillation with diallyl ether as entrainer.

For some reactions, the raw allyl alcohol – water azeotrope (72 % allyl alcohol) may be used without further purification.

2.3.2. Isomerization of Propene Oxide

The catalytic rearrangement of propene oxide is of increasing commercial interest and has partially replaced the hydrolysis of allyl chloride. The most commonly used catalyst is lithium phosphate.

Vapor-Phase Process. Propene oxide vapor passes at 250 – 350 °C over a lithium phosphate catalyst containing up to 30 % inert material. The catalyst is arranged in a fixed bed. Conversions are about 70 – 75 %; allyl alcohol selectivity is 97 %. Space-time yields of 0.5 kg allyl alcohol per liter of catalyst per hour are obtained. Olin Mathieson uses a lithium phosphate catalyst which contains 1 % alkali hydroxide and which has to be regenerated after 40 h by washing with acetone [96]. The catalyst developed by Chemische Werke Hüls [97], containing 73.6 % lithium phosphate and 17.5 % silica, needs no regeneration. Even after 1200 h, an allyl alcohol selectivity of 97.3 % is maintained. Lithium arsenate also has been proposed as an isomerization catalyst but has no essential advantage over lithium phosphate [98]. An older, chromium oxide-catalyzed process, developed

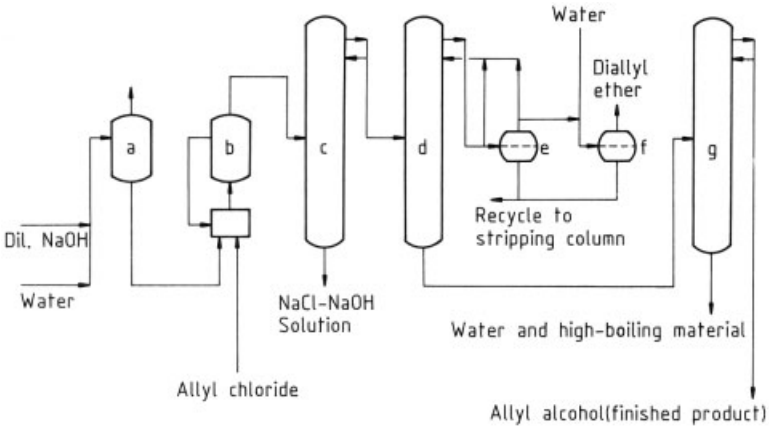

Figure 2. Allyl alcohol production
a) Preheater; b) Hydrolysis reactor with recycle system; c) Stripping column; d) Dehydration column; e) Separator; f) Diallyl ether washing tank; g) Allyl alcohol purification column

by Wyandotte Chemical Corp. [99], gives much lower conversions and selectivities.

Liquid-Phase Process. The Progil process [100] is carried out with a finely ground lithium phosphate catalyst suspended in high-boiling solvents, such as terphenyl or alkyl benzenes. Propene oxide is bubbled through the suspension at 280 °C. Conversion is 60 %, allyl alcohol selectivity 92 %, and the space-time yield 1 kg allyl alcohol per liter per hour. Catalyst lifetimes of 500 – 1000 h are achieved. Jefferson Chemical Co. patented a similar process using a mixture of biphenyl and diphenyl ether as solvent [101].

2.3.3. Hydrolysis of Allyl Acetate

Propene, acetic acid, and oxygen react in the gas phase over palladium catalysts to form allyl acetate, which in turn is hydrolyzed to allyl alcohol:

$$C_3H_6 + CH_3COOH + 1/2\,O_2 \xrightarrow{Pd} CH_2{=}CHCH_2OCOCH_3 + H_2O$$

$$CH_2{=}CHCH_2OCOCH_3 + H_2O \rightarrow CH_2{=}CHCH_2OH + CH_3COOH$$

$$C_3H_6 + 1/2\,O_2 \rightarrow CH_2{=}CHCH_2OH$$

The acetic acid is recovered and reused in the oxidation of propene. Therefore, only propene and oxygen are required as raw materials for the preparation of allyl alcohol.

This process needs no chlorine and therefore may be of commercial interest in the future. Allyl acetate can be hydrolyzed by heating at about 230 °C and 3 MPa, according to a process developed by Hoechst [102], [103]. In the Bayer process hydrolysis is carried out catalytically over an acid cation exchanger (sulfonated polystyrene) at 100 °C [104], [105]. The overall yield of allyl alcohol, relative to the propene feed, is 90 %.

2.3.4. Hydrogenation of Acrolein

Allyl alcohol also can be obtained by catalytic hydrogenation of acrolein in the vapor phase. Use of cadmium – zinc catalysts gives yields up to 70 % [106]. The hydrogenation also can be carried out over silver – cadmium alloys on inert carriers, such as alumina or silica, also giving yields up to 70 % [107].

Ethanol or isopropyl alcohol can reduce acrolein to allyl alcohol in the presence of a mixture of magnesium oxide and zinc oxide [108], [109]. The reaction takes place in the vapor phase at 400 °C; allyl alcohol yields of 80 % have been reported [108].

2.4. Quality and Analysis

Commercial allyl alcohol has the following specifications: content of allyl alcohol 98.0 wt % (minimum), content of water 0.3 wt % (maximum), boiling range 95 – 98 °C. The purity of allyl alcohol is determined by GC. In the absence of saturated alcohols, allyl alcohol can be determined also by acetylation with *N*-acetylpyridinium chloride and back titration of the unconverted *N*-acetylpyridinium chloride. Water is determined by Karl Fischer titration.

2.5. Uses

Allyl alcohol is an intermediate in the production of polymerizable allyl ethers and esters, especially diallyl phthalate. The bulk of the allyl alcohol produced commercially is consumed in the production of glycerol. Polymeric allyl alcohol reacts with unsaturated fatty acids to give drying oils [75]. Sulfur dioxide and allyl alcohol yield polymeric allylsulfonic acids, which have been proposed as intermediates for plasticizers and textile auxiliaries [110]. Allyl alcohol can be copolymerized with other monomers [76], [77] and then used as an intermediate in the production of flame-resistant materials [78], [89], [90] or as a nematocide, fungicide, or preservative.

2.6. Methallyl Alcohol

Methallyl alcohol [*513-42-8*], 2-methyl-2-propen-1-ol, $CH_2{=}C(CH_3)\text{-}CH_2OH$, C_4H_8O, M_r 72.11, *bp* 114.5 °C (101.3 kPa), d_4^{20} 0.8515, n_D^{20} 1.4255, is only partially miscible with water (19.4 wt % methallyl alcohol in water, 33.8 wt % water in methallyl alcohol, at 20 °C). The methallyl alcohol-water azeotrope (*bp* 92 °C) contains 59.8 wt % methallyl alcohol.

Methallyl alcohol is prepared by hydrolysis of methallyl chloride in a 10 wt % aqueous sodium hydroxide solution [111], [112]. Pure methallyl alcohol can be obtained by methanolysis of methallyl acetate in the presence of catalytic amounts of bases [113]. Methallyl alcohol is used in the preparation of polymerizable esters.

3. Allyl Esters

3.1. Properties

Physical Properties. The lower allyl esters, such as allyl acetate, methallyl acetate, and allyl acrylate, are colorless, mobile liquids of low viscosity, with pungent odors, which cause violent irritation of the mucous membranes. The higher allyl esters, such as diallyl phthalate, have high viscosities and boiling points and are nearly odorless. The physical properties of the commercially most important allyl esters are summarized in Table 4.

Chemical Properties. Industrially, the most important property of allyl esters is their ability to polymerize in the presence of oxygen or peroxides. The allyl esters readily undergo hydrolysis and transesterification. Allyl acetate can be oxidized by oxygen and acetic acid over a palladium catalyst to 2-propene-1,1-diol diacetate [*869-29-4*] [114].

Addition of chlorine in the presence of catalytic amounts of transition-metal chlorides yields 2,3-dichloro-1-propanol acetate [*589-96-8*] [115]. Acetic acid can add to allyl acetate in the vapor phase (phosphoric acid catalyst, 200 °C) [116] or in the liquid phase with cation exchangers as catalysts [117], yielding 1,2-propanediol diacetate [*623-84-7*].

Substitution. Hydrogen chloride converts allyl acetate to allyl chloride; the reaction takes place at 100 °C in the liquid phase. A combination of copper(I) chloride and iron(III) chloride which is soluble in allyl acetate, is used as a catalytic system [103], [118], giving conversions of allyl acetate up to 94 %, and an allyl chloride selectivity of 96 %. Over copper catalysts, hydrocyanic acid converts allyl acetate to allyl cyanide (3-butenenitrile) [*109-75-1*] [119–121].

Hydroformylation of allyl acetate creates a mixture of acetoxybutyraldehydes [91], [122], [123], the main product being 4-acetoxybutyraldehyde [*6564-95-0*] in yields of about 70 %. This aldehyde is an intermediate in a proposed production method for 1,4-butanediol and butyrolactone.

Table 4. Physical data of important allyl esters

	CAS Registry No.	Mol. Formula, M_r	*bp*, °C/*p*, kPa	d_4^{20}	n_D^{20}
Allyl acetate (acetic acid 2-propenyl ester)	[*591-87-7*]	$C_5H_8O_2$ 100.12	103.5 – 104.5/101.3	0.9277	1.4050
Methallyl acetate (acetic acid 2-methyl-2-propenyl ester)	[*820-71-3*]	$C_6H_{10}O_2$ 114.14	124/101.3	0.9239 [a]	1.4050
Allyl acrylate (2-propenoic acid 2-propenyl ester)	[*999-55-3*]	$C_6H_8O_2$ 112.13	122/101.3	0.9441	1.4320
Allyl methacrylate (2-methyl-2-propenoic acid 2-propenyl ester)	[*96-05-9*]	$C_7H_{10}O_2$ 126.16	67/6.7	0.9335	1.4358
Diallyl maleate (*cis*-butenedioic acid di-2-propenyl ester)	[*999-21-3*]	$C_{10}H_{12}O_4$ 196.20	114 – 116/0.67 109 – 110/0.4	1.0773	1.4699
Diallyl fumarate (*trans*-butenedioic acid di-2-propenyl ester)	[*2807-54-7*]	$C_{10}H_{12}O_4$ 196.20	142/3.7 93/0.4	1.0768	1.4675
Diallyl succinate (butanedioic acid di-2-propenyl ester)	[*925-16-6*]	$C_{10}H_{14}O_4$ 198.22	105/0.4	1.0510	1.4517
Diallyl adipate (hexanedioic acid di-2-propenyl ester)	[*2998-04-1*]	$C_{12}H_{18}O_4$ 226.27	141 – 142/1.1	1.0235	1.4540
Diallyl sebacate (decanedioic acid di-2-propenyl ester)	[*3137-00-6*]	$C_{16}H_{26}O_4$ 282.38	163/0.4	0.976	1.4551
Diallyl phthalate (1,2-benzene dicarboxylic acid di-2-propenyl ester)	[*131-17-9*]	$C_{14}H_{14}O_4$ 246.26	161/0.55 150/0.15	1.120	1.5190
Diallyl isophthalate (1,3-benzene dicarboxylic acid di-2-propenyl ester)	[*1087-21-4*]	$C_{14}H_{14}O_4$ 246.26	181/0.55	1.124	1.521 [c]
Diethylene glycol bis(allyl carbonate) (2,5,8,10-tetraoxa-9-oxotridec-12-enoic acid 2-propenyl ester)	[*142-22-3*]	$C_{12}H_{18}O_7$ 274.27	162/0.25	1.143	1.4503
Triallyl cyanurate (2,4,6-tris (2-propenyloxy)-1,3,5-triazine)	[*101-37-1*]	$C_{12}H_{15}O_3N_3$ 249.27	162/0.25	1.1133 [b]	1.540 [c]
Triallyl phosphate (phosphoric acid tri-2-propenyl ester)	[*1623-19-4*]	$C_9H_{15}O_4P$ 218.19	108 – 110/1.0 93 – 94/0.14	1.0815	1.4500

[a] d_{20}^{20}
[b] at 30 °C
[c] at 25 °C

Allyl acetate can undergo a metathesis reaction leading to 2-butene-1,4-diol diacetate (*cis*- and *trans*-isomers) [*18621-75-5*] and ethylene. With a catalyst system consisting of rhenium heptoxide on alumina promoted by tetramethyltin [*594-27-4*], allyl acetate conversions of 17 % are obtained. The selectivity for 2-butene-1,4-diol diacetate is 96 % [124].

3.2. Production

3.2.1. Oxidation of Olefins

Vapor-phase oxidation of olefins in the presence of lower alkanoic acids and over palla-

dium catalysts yields the corresponding alkenyl esters.

The olefins are attacked in the allylic position. For example, propene is oxidized in the presence of acetic, propionic, or butyric acid to give allyl acetate, allyl propionate, or allyl butyrate, respectively. In a similar way, the methallyl esters are obtained by oxidation of isobutene.

Allyl Acetate. Propene is oxidized with oxygen in the presence of acetic acid in a tubular reactor [125–127]. The solid catalyst is located in a plurality of tubes and contains metallic palladium, alkali acetate, and promoters, such as compounds of iron or bismuth [125], [126]. The reaction takes place in the vapor phase at 50 – 250 °C and elevated pressure. Because of the high exothermicity of the reaction ($\Delta H = -1890$ kJ/mol) and to avoid explosion hazards, propene and acetic acid are converted only partially. The Bayer process [125], [128] yields allyl acetate with selectivities higher than 90 mol%, carbon dioxide being the sole byproduct.

Methallyl acetate is produced from isobutene, acetic acid, and oxygen by a similar process [129].

Liquid-Phase Oxidation. Olefins can be oxidized also in the liquid phase with a catalyst system consisting of palladium chloride and copper(II) chloride (redox system) [130]. This reaction is not selective; major byproducts are 1-propene-1-ol acetate, 1-propene-2-ol acetate, propionaldehyde, and acetone.

3.2.2. Esterification

A general method for the preparation of allyl esters is the reaction of allyl alcohol with the free acids, acid anhydrides, or acid chlorides. Catalysts frequently used are aromatic sulfonic acids (*p*-toluenesulfonic acid and naphthalene-2-sulfonic acid) [128]. Strong mineral acids are of less value because they can decompose the allyl alcohol. Polymerizable acids are esterified in the presence of phenolic inhibitors. The following allyl esters are prepared by this method: *diallyl phthalate*, *diallyl isophthalate*, *diallyl maleate*, *diallyl fumarate*, *triallyl cyanurate* (from cyanuric chloride and allyl alcohol), and *diethylene glycol bis*(*allyl carbonate*). The last of these compounds is obtained from allyl alcohol and the corresponding bischloroformate (prepared from diethylene glycol and phosgene).

3.2.3. Transesterification

Higher allyl esters can be prepared by transesterification of the corresponding methyl esters with allyl alcohol; sodium methanolate is the preferred catalyst [131], [132]. *Diallyl phthalate* and *diallyl sebacate* can be prepared in this way. The transesterification of methyl or ethyl esters with allyl acetate has been proposed [133], [134]. This method would be of industrial interest if allyl acetate were commercially available as a primary product, for example, by oxidation of propene [133]. Suitable transesterification catalysts are alkoxides or alkoxide complexes of elements of the first to third main groups of the periodic table or alkoxides of titanium and zirconium [134]. The ester is mixed with twice as much allyl acetate and catalytic amounts of the alkoxides and heated to reflux temperature. Methyl acetate (*bp* 57 °C) and ethyl acetate (*bp* 77 °C) have the lowest boiling points and are removed from the equilibrium by distillation. The desired allyl ester can be isolated in excellent yields and high purity.

3.2.4. Other Production Methods

The reaction of allyl chloride with the alkali salts of mono- or dicarboxylic acids to give the corresponding allyl esters is well known. These salts also may be formed during the reaction. In this way, *diallyl phthalate* is obtained from the reaction of allyl chloride and phthalic anhydride in the presence of sodium hydroxide, sodium carbonate, tertiary amines, or quaternary ammonium salts [135].

Allyl acrylate is prepared by the pyrolysis of allyl lactate, *allyl methacrylate* in a similar manner from allyl 2-hydroxyisobutyrate [136]. *Substituted tetrahydrophthalic acid allyl esters* are obtained by Diels-Alder reaction of the corresponding substituted butadienes and diallyl maleate [137]. The best method for the preparation of *triallyl phosphate* is the condensation of phosphorus trichloride with allyl alcohol to give *triallyl phosphite* [*102-84-1*] followed by oxidation with air [138]. *Diethylene glycol bis*(*allyl*

carbonate) can be prepared in 87 % yield by heating allyl chloride, diethylene glycol, and carbon dioxide in the presence of sodium carbonate and triethylamine in an autoclave [139].

3.3. Uses

3.3.1. Polymer Production

Polymer production is the most important application of allyl esters. They are especially suitable as components in copolymers. Some allyl esters are used for cross-linking or curing of polyolefins (graft copolymerization). Optimum cross-linking is obtained by adding specific peroxides or by high-energy irradiation. The cross-linked polymers and copolymers of allyl esters give thermoset articles of excellent heat resistance. Other applications of these polymers are in the production of casting sheets, molding material, electric and optical devices of high resistance, and flame-retardant materials.

Diethylene glycol bis(allyl carbonate) usually is referred to as *CR 39 monomer*. It is bulk polymerized by the addition of relatively large (compared to similar polymerizations) amounts of peroxides. The commercially preferred peroxide initiator is isopropyl percarbonate [*105-64-6*]. The homopolymer (*CR 39 polymer*) [*25656-90-0*] is suitable for the production of cast sheets, lenses, and other articles that have excellent resistance to scratching, impact, and heat as well as outstanding optical properties [140–142]. This polymer is resistant to common solvents with the exception of oxidizing acids. It is produced in the United States by PPG Ind. [140]. Even better qualities can be obtained by reacting CR 39 monomer with other monomers, such as methyl methacrylate [*80-62-6*] or triallyl isocyanurate [*1025-15-6*]. These copolymers are used for the production of light-weight lenses with very hard surfaces and high refractive indices [143].

Diallyl phthalate is an important monomer for the production of thermosetting molding compounds, which must have good dimensional stability and electrical properties, and be resistant to heat and solvents. Diallyl phthalate can be polymerized or copolymerized. The preferred technique is first to prepare a prepolymer or precopolymer in solution. This usually is done by dissolving the diallyl phthalate monomer in 2-propanol, adding 50 % hydrogen peroxide at about 105 °C, and precipitating the prepolymer from the cooled, viscous solution with excess 2-propanol. Precopolymers can be prepared by adding comonomers, such as triallyl cyanurate, acrylates, vinyl compounds, acrylonitrile, styrene, or diallyl isophthalate. The prepolymers and precopolymers are bulk polymerized by mixing the prepolymer or precopolymer with large amounts of free-radical initiators (benzoyl peroxide, *tert*-butyl perbenzoate) and molding at temperatures up to 200 °C and 40 – 45 bar for about 15 min. Because of their outstanding mechanical and electrical properties, copolymers containing diallyl phthalate are suitable for specialty coating and for embedding, especially in the production of electronic devices. For example, the moisture-sensitive epoxy compounds now used in light-emitting diode (LED) displays can be replaced by stable diallyl phthalate epoxy encapsulating resins [144], [145].

By adding inorganic materials to diallyl phthalate prepolymer compositions, reinforced thermosetting molding compounds can be obtained. Glass cloth or paper can be impregnated with a solution of prepolymer, monomer, and peroxide initiator. After removal of the solvent, the glass cloth or paper is cured to give the desired film-protected material, which is used for decoration, stain-resistant overlays for household articles, and furniture.

Diallyl isophthalate (DAIP) has the advantage over the ortho isomer that it polymerizes faster and gives polymers of better heat resistance. The DAIP prepolymer is less stable than the diallyl phthalate prepolymer. Cured moldings from DAIP monomer-prepolymer compositions can be better processed than cured diallyl phthalate moldings because of their greater fluidity. Compositions containing DAIP prepolymers can be used for the production of hard, translucent, abrasive-resistant, and laser-trimmable coatings [146].

Diallyl esters of aliphatic dicarboxylic acids, e.g., diallyl succinate, adipate, and sebacate, are preferred for the preparation of soluble, thermoplastic precopolymers that sometimes

contain vinyl monomers. These precopolymers can also be used for graft copolymerization with preformed polymers; heat- and solvent-resistant adhesives, plastics, and coatings are produced by this technique.

Diallyl maleate and diallyl fumarate have additional activated double bonds and copolymerize readily with styrenes, vinyl ethers, and acrylates. The less reactive allyl groups remain intact and can subsequently be cross-linked by heating with peroxides or treating with high-energy radiation. As in the case of the saturated diallyl esters, soluble precopolymers are obtained by heating the appropriate mixture of the monomers in the presence of peroxides [147]. Preferential uses are castings and moldings, plastisols, and adhesives.

The allyl acrylic monomer of most importance is allyl methacrylate; allyl acrylate is of secondary importance. Both are used in small amounts for the preparation of soluble, fusible copolymers that can be cross-linked in a subsequent reaction to form insoluble, heat-resistant plastics and coatings. Homopolymers also can be produced by radical-initiated polymerization but are too expensive and are of poorer quality than copolymers. About 10 % of allyl methacrylate and allyl acrylate is copolymerized with other acrylic esters. Preformed vinyl polymers can be cross-linked and processed to form graft polymers. A copolymer of 10 % allyl methacrylate and 90 % vinyl chloride is a flame-retardant additive for polypropylene in amounts of up to 30 % relative to the polyolefin [148]. Allyl acrylate is used in the production of dental plastics, optical lenses, reinforced plastic adhesives, coatings, rubbers, and components in printing processes. It can also be used to modify the properties of cotton [149], [150] or nylon 6 [151].

Triallyl cyanurate is used preferentially as a cross-linking agent in copolymers. On heating it may polymerize violently and then isomerize to the more stable triallyl isocyanurate [*1025-15-6*] with the allyl groups attached to the nitrogen atoms [152], [153].

At 30 °C viscous solutions of prepolymers form slowly. Triallyl cyanurate is used for the production of heat- and solvent-resistant coatings and moldings, reinforced plastics, and adhesives. Addition of 5 – 10 % triallyl cyanurate to polyester – styrene or methyl methacrylate yields cast sheets of improved mechanical and thermal stability. Short-term heating of triallyl cyanurate with polymers in the presence of peroxides at 180 °C gives cross-linking; valuable copolymers are thus obtained from poly(vinyl chloride) elastomers [154] and fluoroelastomers [155]. Ethylene polymers and copolymers also may be cross-linked under similar reaction conditions [156]. High-impact plastics have been obtained by grafting butyl acrylate-triallyl cyanurate copolymer with a styrene – acrylonitrile mixture [157]. Other examples of peroxide-initiated curing with triallyl cyanurate (2 – 5 %) at 150 – 160 °C are polyurethanes [158], nylons [159], cellulose [160], polyoxyethylene [161], vinyl-substituted polysiloxanes [162], and acrylate copolymers [163]. Crosslinking of polycarbonate with triallyl cyanurate by UV irradiation in the presence of polythiols gives scratch-resistant coatings [164].

Triallyl phosphate can explode on heating to about 130 °C and polymerizes rapidly on exposure to air. A possible commercial application of this reactive monomer is the cross-linking with polyolefins in the presence of peroxides and, if necessary, in the presence of foaming agents [165].

3.3.2. Other Uses

Some allyl esters find application as plasticizers, textile auxiliaries, and insecticides. Allyl cinnamate [*1866-31-5*] is a component of perfumes. 2,3-Dibromopropyldiallyl phosphate [*33528-41-5*] has been proposed as a flame retardant in poly(ethylene terephthalate) fibers [166]. Allyl esters also are suitable as diluents for the less reactive peroxides and as dispersion media for pigments and fillers.

4. Allyl Ethers

4.1. Properties and Uses

Simple allyl ethers, such as diallyl ether or alkyl allyl ethers, have only minor commercial importance. Valuable allyl ethers are polyol allyl ethers and allyl ethers that contain an epoxy group.

Allyl glycidyl ether [*106-92-3*], (allyloxymethyl)oxirane, M_r 114.14, *bp* 87.5 – 88 °C (10.9 kPa), d_4^{20} 0.9678, n_D^{20} 1.4345, is a toxic liquid used as an additive for epoxy resins and as a comonomer in polyglycols and polyolefins. Crosslinking is achieved by irradiation or by thermally induced peroxide decomposition. Copolymerization of allyl glycidyl ether with acrylamide or *N,N′*-methylenebis(acrylamide) yields water-soluble polymers containing epoxy groups. These can be condensed with pharmacologically active (controlled-release) compounds to form pharmacologically active polymers [167].

Ethylene glycol diallyl ether [*7529-27-3*], 1,2-bis(allyloxy)ethane, $C_8H_{14}O_2$, M_r 142.20, *bp* 35 – 37 °C (0.13 kPa), d_4^{20} 0.8940, n_D^{20} 1.4340, is polymerized oxidatively with cobalt acetate and air in methanol solution soaked on the surface of filter paper. The resulting polymeric product confers a high degree of hydrophilicity to the filter paper [168].

Partial allyl ethers of trimethylolpropane and pentaerythritol are used to increase the drying rate of alkyd drying-oil coatings. By condensing a mixture of various pentaerythritol allyl ethers with polyesters and ethylene glycol monobutyl ether [*111-76-2*] and then curing with cobalt octanoate, a film of high tensile hardness and flexibility can be obtained [169].

Tetraallyl pentaerythritol ether [*1471-18-7*], $C_{17}H_{28}O_4$, 296.41, *bp* 124 – 125 °C (0.13 kPa), d_4^{20} 0.9497, n_D^{20} 1.4595, can be added to lubricating oil to improve its viscosity index (i.e., to reduce temperature effects) [170].

Allyl Ethers of Carbohydrates. Sucrose polyethers copolymerize with acrylic acid to yield branched, water-soluble polymers. Allyl ethers of starches have been proposed as air-drying protective coatings and as varnishes but have found only limited commercial interest because of their instability and water sensitivity.

4.2. Production

Allyl ethers of low molecular mass, such as diallyl ether and alkyl allyl ethers, are prepared by heating allyl alcohol or mixtures of allyl alcohol with monoalkanols in the presence of mineral acids. The more valuable polymerizable allyl ethers of polyols are prepared by reaction of allyl chloride with the polyols in the presence of sodium hydroxide [171], [172] or by treating an aqueous solution of the polyol with allyl chloride in a two-phase system to which a phase-transfer catalyst has been added. Using tetrabutylammonium bromide [*1643-19-2*] as the phase-transfer catalyst, up to 99.7 % tetraallyl pentaerythritol ether is obtained from the reactants [170].

5. Allylamines

5.1. Properties and Uses

Allylamine [*107-11-9*], $CH_2{=}CHCH_2NH_2$, C_3H_7N, M_r 57.10, *bp* 54.5 °C, d_4^{20} 0.7621, is a liquid with an ammoniacal odor. It can be converted to *N*-allylamides or -imides by reaction with esters, acid anhydrides, or acid chlorides. Allylamine is oxidized at 400 °C on a bismuth molybdate catalyst; acrylonitrile and propionitrile are the main products (85 % combined yield) [173]. Plasma polymerization of allylamine gives polymers that can be used as reverseosmosis membranes. A possible future application of these membranes is in closed environmental systems, such as manned spacecraft [174], [175]. Allylamine is useful as a corrosion inhibitor when pickling steel in acid [176]. The rhodium-catalyzed reaction of allylamine with carbon monoxide at 120 °C and 10 MPa yields 67 % γ-butyrolactam [*616-45-5*] [177].

Diallylamine [*124-02-7*], $C_6H_{11}N$, M_r 97.16, *bp* 111 – 112 °C, can be cyclized to give substituted pyridines [178]. It is an intermediate in the production of pharmaceuticals and resins.

Triallylamine [*102-70-5*], $C_9H_{15}N$, M_r 137.23, *bp* 150 – 151 °C, has been proposed as a catalyst for the production of polyesters [179] and as an initiator for the polymerization of butadiene [180].

Methallylamine [*2878-14-0*], $CH_2{=}C(CH_3)$-CH_2NH_2, C_4H_9N, M_r 71.12, *bp* 78.8 °C, d_4^{20} 0.782, is dehydrogenated to methacrylonitrile

[*126-98-7*] over a silver catalyst. Copolymerization of methallylamine with acrylonitrile increases the affinity of polyacrylonitrile fibers for dyes [181].

5.2. Production

Allyl chloride reacts with aqueous ammonia at about 100 °C to give a mixture of the hydrochlorides of allyl-, diallyl-, and triallylamine, from which the pure amines are obtained by distillation [182]. Methallylamine is produced in a similar manner [183].

Table 5. Hepatotoxic effect ∗ of some allyl and propyl compounds in rats [190]

	Dosage, mg/kg	Degree of liver necrosis
Allyl alcohol	25	1.5
Allyl formate	40	2
Allyl acetate	45	2.5
Allyl butyrate	85	1.5
Allyl caproate	75	2.5
Allyl heptanoate	165	3
1-Propanol	2160	0
Propyl formate	1330	0
Propyl acetate	3120	0
Propyl butyrate	5000	0

∗ Degree of macroscopically visible liver necrosis on day 5 after oral application of 1/3 of the LD_{50} for four consecutive days. Degree 1: just clearly visible; degree 3: severe necrosis.

Pure allylamine can be prepared by hydrolysis of allyl isothiocyanate [184], thermal cleavage of allyl dithiocarbamate [185], hydrolysis of diallyl cyanamide [186], or ammonolysis of triallylamine [187]. Allylamine may be produced also by electroreduction of acrylonitrile on lead cathodes [188]. Polyallylamines are readily obtainable by catalytic hydrogenation of poly(acrylonitrile) over a Raney nickel catalyst [189].

6. Toxicology and Occupational Health

Allyl compounds are stronger irritants to the skin and mucous membranes and are more toxic than the corresponding alkyl compounds (Table 5). Many allyl compounds are absorbed easily through the skin and mucous membranes. After absorption, aliphatic allyl compounds can cause severe liver damage, in contrast to comparable alkyl compounds (Table 5). Other target organs are mainly the central nervous system, the kidneys, and the hematopoietic system.

Allyl Chloride [*107-05-1*]**.**

Acute and Subacute Toxicity. LD_{50} = 460 mg/kg (rat, oral) [191]; LD_{50} = 3.7 mg/kg (rabbit, percutaneous) [5]; LC_{50} = 11 mg/L (rat, inhalation, 2 h) [191]. The inhalation of 3 ppm allyl chloride during 7 h/d on 5 days a week was tolerated by a group of rats, guinea pigs, and rabbits for 180 days without irreversible damage occurring. An analogous test using 8 ppm over a period of 35 days led to damage of the liver and kidneys [192]. Further experiments demonstrate a neurotoxic effect of allyl chloride, in particular to the peripheral nerves of cats and rabbits [191], [193], [194].

Carcinogenicity, Mutagenicity, Embryotoxicity. So far, little information on the *carcinogenic activity* of allyl chloride is available [195–199]. A carcinogenic effect is suspected, although a definite statement cannot be made at present [195]. The *mutagenicity* of allyl chloride has been confirmed in various tests [200–203]. A toxic effect on the development of rat embryos, as well as increased *embryo mortality,* was established in rats that had inhaled air containing allyl chloride at a concentration of 3.1 mg/m^3. At 0.29 mg/m^3 no negative effects occurred [204].

General Characteristic Effects. Allyl chloride causes strong irritation of the skin and the mucous membranes. It is absorbed by inhalation, by ingestion, and through the skin. Direct contact and exposure to air – allyl chloride mixtures lead to strong irritation of the eyes, mucous membranes, and nasopharyngeal cavity. Acute symptoms are numbness, unconciousness, and – particularly upon repeated and lengthy exposure to higher concentrations – pulmonary edema. The heart and circulatory system, liver, and kidneys are endangered. At longer exposure times and higher concentrations, fatal poisoning is possible. After prolonged exposure, chronic damage of the liver and kidneys occurs [5], [205], [206].

Occupational Health. Because of its toxicity, allyl chloride is classified as a poisonous material. The threshold limit value (TLV) and the MAK are 1 ppm (8 h time weighted aver-

age) [207–209]. The short-term exposure limit is twice as high: 2 ppm [205], [208]. Because of its suspected carcinogenic potential, allyl chloride is classified in group III B by the MAK commission [207]. In the former USSR, the maximum allowed workroom concentration is set at 0.1 ppm [210].

The particularly pungent odor of allyl chloride cannot serve as an adequate warning signal. The odor perception thresholds, at which 50 and 100 % of exposed persons perceive allyl chloride vapors in the air, are 3 to 6 and 25 ppm, respectively [5]. These values are considerably above the permissible working concentrations.

So far, few investigations in occupational medicine on the effects of allyl chloride exist [193], [205], [211].

Allyl Alcohol [*107-18-6*]**.** With respect to its irritating and toxic effects, allyl alcohol is the best investigated aliphatic allyl compound. The odor perception threshold is approximately 0.8 ppm [212], and 6 – 12 ppm causes irritation of the nose. At a concentration of 25 ppm, severe irritation of the eye occurs, with lacrimation, photophobia, blurred vision, and retrobulbar pain [212]. Corneal necrosis results in temporary blindness in persons exposed to higher concentrations [213]. Sensitive persons may react with slight irritation of the eyes at concentrations of 2 – 5 ppm [214], [215]. In animal experiments daily inhalation of concentrations between 2 and 7 ppm (7 h/d) has been tolerated without irritation in dogs, rabbits, guinea pigs, and rats during periods of up to 6 months [192].

The threshold limit value (TLV) and the MAK value are 2 ppm [208], [209]. The short-term exposure limit is twice as high, namely 4 ppm.

A single oral application of allyl alcohol (0.02 mL/kg) caused liver necrosis in rats and the death of 7 out of 12 animals [216]. The oral and cutaneous LD_{50} for rabbits is 50 – 80 mg/kg, and for rats 64 mg/kg [217], [218]. A single dermal application of allyl alcohol to dogs (approximately 0.2 mL/kg) caused lethal gastrointestinal hemorrhages within a few hours [217]. In rats hepatotoxic effects with liver cell necrosis were observed after a single application of 62.5 µL/kg by gavage [219] and after a single intraperitoneal injection of 0.5 mL/kg [220]. On the other hand, the daily application of 50 ppm of allyl alcohol (4.8 – 6.2 mg/kg) with the drinking water was tolerated by rats for 15 weeks without effect [221]. In cases of skin contamination, the danger of systemic intoxication brought on by the high absorption rate of allyl alcohol through the skin (one drop of allyl alcohol equals approximately 50 mg) exists in addition to local irritation [218].

Allyl Esters. The toxicological effects of allyl esters are quite similar to those of allyl alcohol. Local irritation and damage to the central nervous system and the liver caused by absorption are the main effects. For *allyl acetate* [*591-87-7*], the dermal LD_{50} in rabbits is 1000 mg/kg and the LC_{50} in rats is 1000 ppm after 1 h of inhalation. Comparable values are reported for *diallyl maleate* [*999-21-3*] [212]. Table 5 shows the hepatotoxicity of simple allyl esters, compared with some propyl esters. The hepatotoxicity of *diallyl phthalate* [*131-17-9*] is probably in the same range as that of others in the table, but it causes only slight irritation of the skin or mucous membranes [222]. *Allyl methacrylate* [*96-05-9*] has an oral LD_{50} in rats of 430 mg/kg and the dermal LD_{50} in rabbits is 500 mg/kg. Allyl methacrylate is easily absorbed through the skin. The inhalational LCLo (lowest lethal concentration) in rats has been found to be 500 ppm [223].

Allyl Ethers.

Allyl glycidyl ether [*106-92-3*] has an LD_{50} of 390 mg/kg (mouse, oral) and an LD_{50} of 1600 mg/kg (rat, oral). In rabbits, the dermal LD_{50} is 2550 mg/kg. Inhalational toxicity has been estimated: LC_{50} = 270 ppm (mice, 4 h) and 670 ppm (rats, 8 h) [224], [225].

Allyl glycidyl ether causes only slight irritation of the skin but is strongly irritating and corrosive to the eye. Following inhalation, the irritation of the respiratory tract can lead to pulmonary edema or to secondary bronchopneumonia [224], [225]. After absorption, systemic intoxication with disorders of the central nervous system and morphologic damage to the liver, kidneys, and spleen can occur [224], [225]. In humans, dermal sensitization has been observed besides the irritating effects [224], [225]. The threshold limit (TLV) and the MAK values are 10 ppm or 45 mg/m^3 [208], [209].

Diallyl ether [*557-40-4*] smells like horseradish and severely irritates the mucous membranes; the LD_{50} is approximately 270 mg/kg (rat, oral) [226]. Diallyl ether is absorbed easily through the skin; the LD_{50} is 540 mg/kg (rabbits, dermal, estimate) [226].

Ethylene glycol diallyl ether [*7529-27-3*] has an oral LD_{50} in rats of 1020 mg/kg. Exposure to it for 24 h causes severe irritation of the skin (500 mg) and eyes (250 µg) of rabbits [227].

Allyl ethyl ether [*557-31-3*] and *allyl vinyl ether* [*3917-15-5*] are irritants to the skin, mucous membranes, and eyes [228]. For allyl vinyl ether, the oral LD_{50} in rats is 450 mg/kg [228].

Allylamines. Allylamines also are irritating to the skin and mucous membranes and are corrosive to the eye. In general, allylamines are absorbed easily through the skin. The dermal LD_{50} of *allylamine* [*107-11-9*] in the rabbit is 35 mg/kg [229]. The oral LD_{50} has been determined to be 106 mg/kg (rat) and 57 mg/kg (mouse) [229], [230]. For rats, the inhalational LC_{50} is 286 ppm (4 h). In humans, the odor perception threshold of allylamine is approximately 2.5 ppm, whereas the TCLo (lowest toxic concentration) has been found to be 5 ppm (5 min). Above 2.5 ppm, irritating effects on the respiratory tract can occur; 14 ppm is intolerable [231].

In contrast to other allyl compounds, allylamine, diallylamine [*124-02-7*], and triallylamine [*102-70-5*] can cause myocardial degeneration as well as damage to the kidneys and liver. This has been observed in rats and rabbits [230]. Allylamines cause a severe primary fibrosis of the myocardium [232], [233]. The oral application of allylamine in the drinking water over 81 – 104 days caused a dose-dependent myocardial degeneration in rats [233].

7. References

1. *Ullmann,* 4th ed., **9,** 466.
2. Solvay Note CMP 8049 V-2, 20 (internal document), 1976.
3. *Kirk-Othmer,* 3rd ed., **5,** 764.
4. *Beilstein* **3** (1), 699.
5. Shell Chemical Co.: Allyl Chloride Toxicity and Safety Bulletin 10/76, SC 195 – 76, 1976.
6. Chemische Fabrik Kalk, DE 2913277, 1980 (H. Jenker, R. Strang).
7. M. Capka, CS 187167, 1979 (M. Capka).
8. M. J. Hawthorne, J. A. Dupont, *J. Am. Chem. Soc.* **80** (1958) 5830.
9. J. Plesek, Z. Plzak, J. Stuchlik, T. Hermanek, *Collect. Czech. Chem. Commun.* **46** (1981) no. 8, 1748.
10. L. Z. Soborowskii, Ju. M. Zinovev, L. I. Muler, *Zh. Obshch. Khim.* **29** (1959) 3907. *Beilstein* **1** (4), 739.
11. H. Klein, H. Mayr, *Angew. Chem.* **93** (1981) 1069; *Angew. Chem. Int. Ed. Engl.* **20** (1981) 1027.
12. E. M. Fettes, F. O. Davis, N. E. Gaylord (ed.): *Polyethers* part III, Interscience Publishers, New York 1962, pp. 225 – 270.
13. H. Jacobi, W. Flemming, US 2259470, 1941.
14. M. K. Gadzhiev, SU 810684, 1979.
15. Kaustik Sterlitamak Industrial Enterprises USSR, SU 859357, 1981 (E. N. Denisov, Yu. D. Morozov).
16. A. M. Mezheritskii, SU 578301, 1976 (A. M. Mezheritskii, M. M. Krivenko, N. N. Vdovenko).
17. A. M. Mezheritskii, SU 654609, 1979 (A. M. Mezheritskii, M. M. Krivenko, N. N. Vdovenko).
18. Akad. Wissenschaft DDR, DD 136497, 1978 (D. Ballschuh, R. Ohme, J. Rusche).
19. Tokuyama Soda Corp., JP 8105442, 1981; JP 8105443, 1981 (S. Koyanagi, N. Hasegawa, T. Shimizu, S. Katsushima, I. Kaneho).
20. M. Foa, L. Cassar, *Gazz. Chim. Ital.* **109** (1979) no. 12, 619.
21. *Ullmann,* 4th ed., **9,** 163.
22. C. E. Schildknecht: *Allyl Compounds and their Polymers, High Polymers,* **vol. XXVIII,** Wiley-Interscience, New York-London-Sydney-Toronto 1973.
23. G. W. Hearne, T. W. Evans, H. L. Jale, M. C. Hoff, *J. Am. Chem. Soc.* **75** (1953) 1392. Shell Oil Co., US 2939879, 1957 (A. De Benedictis).
24. A. W. Fairbairn, H. A. Cheney, A. J. Cherniavsky, *Chem. Eng. Prog.* **43** (1947) 280 – 290.
25. L. M. Porter, *J. Am. Chem. Soc.* **78** (1956) 5571.
26. E. C. Williams, *Trans. Am. Inst. Chem. Eng.* **37** (1941) 157 – 207.
27. L. W. Chubb, *Ind. Chem.* **30** (1954) 491 – 496.
28. Shell Development Co., US 2643272, 1947; US 2763699, 1956 (C. P. Van Dijk, F. J. F. van der Plas); GB 765764, 1955; GB 790166, 1956.

29. Union Carbide Corp., US 3054831, 1958.
30. Columbia-Southern Chemical Corp., GB 901680, 1962.
31. Nissan Chemical Ind., BE 657 267, 1964.
32. Asahi Denka Kogyo Co., JP 42/7366, 1966; JP 73/32087, 1968; JP 73/26732, 1969.
33. Asahi Glass Co., JP 73/30250, 1968.
34. Asahi Electrochem. Co., JP 72/13006, 1967; JP 72/ 13007, 1967; JP 74/39961, 1969; JP 75/16332:4, 1969.
35. BASF, DE 1114184, 1959.
36. P. Klucovsky, J. Dykyj, *Acta Chim. Acad. Sci. Hung.* **36** (1963) nos. 1 – 4, 145 – 155.
37. Shell Development Co., GB 761831, 1954.
38. G. A. Oshin, SU 706392, 1979.
39. Solvay Cie., DE 1960063, 1969.
40. D. L. Jabroff, J. Anderson, *World Pet. Congr. Proc. III, Sect. V* (1951) 22 – 30.
41. Halcon International, DE 1200286, 1965; DE 1210806, 1966.
42. L. Prochazka, CS 138599, 1968.
43. Dow Chemical Corp., DE 2540336, 1975.
44. Halcon International, DE 1215691, 1966.
45. *DECHEMA-Werkstofftabellen A-Z DWT 53,* Dechema-Verlag, Frankfurt 1953.
46. Hoechst, DE 1960158, 1969.
47. Shell Development Co., US 2207193, 1940 (H. P. A. Groll).
48. BASF, DE 926246, 1952.
49. Bayer, DE 1210801, 1966.
50. Hoechst, DE 1224302, 1964; DE 1230780, 1964; DE 1237554, 1965; DE 1237555, 1965; DE 1243670, 1965; DE 1274112, 1966; DE 1282637, 1967; DE 1283828, 1966; DE 1793132, 1968.
51. Hoechst, DE 1222913, 1964; DE 1224301, 1964; DE 1227006, 1964; DE 1230781, 1965.
52. Monsanto Chem. Co., US 2966525, 1959; GB 935088, 1960 (D. E. Steen).
53. ICI, DE 1234705, 1963.
54. Toyo Soda, GB 1016094, 1963.
55. Deutsche Texaco, DE 1300930, 1967.
56. El Paso Products Co., DE 1280847, 1966.
57. Showa Denko Kabushiki Kaisha, DE 1817281, 1968.
58. Hoechst, *Oil Gas J.* **68** (1970) 58.
59. Shell Chemical Co., DE 2114302, 1971 (W. Rootsaert, R. Van Helden, F. Wattimina).
60. Institut Neftechimitscheskich Processov Imeni Akademika Ju. G. Mamedalieva Akademii Nauk Azerbaidschanskoj SSR, Baku, SU 2585604, 1978.
61. El Paso Products Co., DE 1293744, 1966.
62. Lummus Co., DE 2426640, 1974 (H. D. Schindler, M. Sze, H. Riegel).
63. C. A. Peterson, J. A. Key, F. D. Hobbs, J. W. Blackburn, H. S. Basdekis et al.: *Organic Chemical Manufacturing,* **vol. 10,** Selected Processes, Report 3, Glycerin and its Intermediates, EPA Report 450/3-80-028 e (1980).
64. N. Piccinini, U. Anatra, G. Malandrino, D. Barone, S. Donato: "Safety analysis for an allyl chloride," *Plant/Oper. Progr.* **1** (1982) 1,69.
65. G. Hommel: *Handbuch der gefährlichen Güter* 1, Springer Verlag, Berlin-Heidelberg-New York 1980.
66. F. Asinger: *Die petrolchemische Industrie,* **vol. II,** Akademie-Verlag, Berlin 1971, pp. 946 – 950 and pp. 953 – 960.
67. Xantia National Corp., FR 2472161, 1981.
68. Toray Industries Corp., JP 8137152, 1981; JP 8137153, 1981.
69. *Beilstein,* **1** 436; **1** (1), 224; **1** (2), 474; **1** (3), 1873.
70. F. Andreas, K. Gröbe: *Propylenchemie,* Akademie-Verlag, Berlin 1969.
71. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* Verlag Chemie, Weinheim-New York 1976.
72. H. Warson: "Allyl alcohol and derivatives," *Polym. Paint Colour J.*1976, 698 – 699.
73. R. K. Grasselli, J. D. Burrington, *Adv. Catal.* **30** (1981) 133 – 163.
74. C. E. Schildknecht, *Encycl. Chem. Process. Des. 1977,* **2,** 460.
75. American Cyanamid Co., GB 573757, 1945.
76. Plastics Engineering Co., US 4168360, 1978 (G. F. D'Alelio).
77. Osaka Soda Co., JP 7987788, 1977 (A. Suzui).
78. F. H. Otey, C. A. Wilham, C. R. Russell, *Ind. Eng. Chem. Prod. Res. Dev.* **17** (1978) no. 2, 162 – 164.
79. T. P. Forbeth, B. J. Gaffney, *Pet. Refiner* **34** (1955) no. 12, 160.
80. Sumitomo Chemical Co., JP 8277636, 1980.
81. Martinez de la Cuesta, E. Costa Novella, P. J. E. Rus Martinez, J. Torregrosa Anton, J. Augado Alonso. *An. Quim.* **76** (1980) no. 3, 380 – 386.
82. K. Yamagishi, O. Kageyama, *Hydrocarbon Process.* **55** (1976) no. 11, 139 – 144.
83. Shell Development Co., US 2042220, 1934 (H. P. A. Groll, H. W. de Jong).
84. Bataafsche Petroleum Mij., FR 788921, 1935.
85. National Distillers and Chemical Corp., US 4051151, 1975 (J. H. Murib).
86. D. V. Sokol'skii, O. A. Tyurenkova, V. A. Dashevskii, E. I. Seliverstova, *Kinet. Katal.* **7** (1966) no. 6, 1032.

87. Olin Mathieson Chemical Corp., US 3037059, 1959 (D. W. Kaiser).
88. Chemische Fabrik Kalk, DE-OS 2526653, 1975 (H. Jenkner, O. Koenigstein).
89. Du Pont, US 3283013, 1962 (R. W. Rimmer).
90. Phillips Petroleum Co., US 3471579, 1967 (D. H. Kubicek).
91. H. Adkins, G. Krsek, *J. Am. Chem. Soc.* **71** (1949) 3051.
92. C. U. Pittman Jr., W. D. Honnick, *J. Org. Chem.* **45** (1980) no. 11, 2132.
93. M. Tamura, S. Kumano, *CEER Chem. Econ. Eng. Rev.* **12** (1980) no. 9, 32 – 35.
94. Stamicarbon B.V., EP-A 38609, 1980 (N. A. de Munck, J. J. F. Scholten).
95. Chevron Research Co., US 4189608, 1976 (V. P. Kurkov).
96. Olin Mathieson Chemical Corp., US 2986585, 1959 (W. I. Denton).
97. Chemische Werke Hüls, DE-AS 1271082, 1964 (W. Knepper, G. Hoeckele).
98. Shell Oil Co., US 3209037, 1961 (R. W. Fourie, R. L. Maycock, G. H. Riesser).
99. Wyandotte Chemical Corp., US 2479632, 1945 (L. H. Lundsted, E. C. Jacobs).
100. Progil, DE-AS 1197077, 1961 (E. Charles, M. E. Degeorges, A. Thizy).
101. Jefferson Chemical Co., US 3238264, 1961 (R. L. Rowten).
102. Hoechst, DE-OS 1949537, 1969 (G. Roscher, H. Schmitz).
103. H. J. Schmidt, G. Roscher, *Compend. Dtsch. Ges. Mineralölwiss. Kohlechem.* **75 – 76** (1975) 318 – 322.
104. Bayer, DE-OS 1933538, 1969 (B. Engelhard, J. Grolig, M. Martin, K.-H. Reissinger, G. Scharfe, et al.).
105. Bayer, DE-OS 2009742, 1970 (G. Scharfe, W. Schwerdtel, W. Swodenk, B. Engelhard, J. Grolig et al.).
106. Degussa, DE 888691, 1952 (H. Brendlein).
107. Celanese Corp., US 4127508, 1976 (T. H. Vanderspurt).
108. Bataafsche Petroleum Mij., GB 619014, 1949 (S. A. Ballard, H. de V. Finch, E. A. Peterson).
109. K. Yamagishi, *CEER Chem. Econ. Eng. Rev.* **6** (1974) no. 7, 40.
110. IG-Farbenindustrie, FR 894673, 1943.
111. Shell Development Co., US 2072015, 1932 (A. W. Tamele, H. P. A. Groll).
112. Shell Development Co., US 2072016, 1932 (A. W. Tamele, H. P. A. Groll).
113. Bayer, DE-OS 1939142, 1969 (W. Swodenk, G. Scharfe, J. Grolig).
114. Bayer, DE-AS 1904236, 1969 (W. Swodenk, G. Scharfe, J. Grolig).
115. Bayer, DE-OS 2121251, 1971 (J. Grolig, G. Scharfe, W. Swodenk).
116. Hoechst, DE-OS 2219915, 1972 (H. Fernholz, D. Freudenberger).
117. Bayer, DE-OS 2019428, 1970 (J. Grolig, G. Scharfe, W. Swodenk).
118. Hoechst, DE-OS 1768242, 1968 (H. Fernholz, H. Wendt).
119. Hoechst, DE-OS 2102263, 1971 (H. Krekeler, H. Fernholz, D. Freudenberger, H.-J. Schmidt, F. Wunder).
120. Hoechst, DE-OS 2124755, 1971 (H. Fernholz, D. Freudenberger).
121. Bayer, DE-OS 1960380, 1969 (P. Kurtz).
122. General Electric Co., DE-OS 2425653, 1974 (W. E. Smith).
123. Bayer, DE-AS 2430082, 1974 (C. Rasp, G. Scharfe, J. Grolig).
124. J. C. Mol, E. F. G. Woerlee, *J. Chem. Soc. Chem. Commun.* 1979, 330.
125. Bayer, FR 1346219, 1962 (H. Holzrichter, W. Krönig, B. Frenz).
126. Bayer, DE-OS 1901289, 1969 (W. Krönig, G. Scharfe).
127. Hoechst, DE-OS 1903954, 1969 (H. Fernholz, F. Wunder, H.-J. Schmidt).
128. D. Swern, E. F. Jordan, *J. Am. Chem. Soc.* **70** (1948) 7.
129. Bayer, DE-OS 1933537, 1969 (G. Scharfe, J. Grolig, W. Swodenk, M. Martin).
130. Consortium, FR 1370867, 1963.
131. Chem. Werke Witten, DE-AS 1211625, 1964.
132. Du Pont, US 2218439, 1937 (H. S. Rothrock).
133. Bayer, DE-OS 1933536, 1969 (W. Swodenk, G. Scharfe, J. Grolig).
134. ICI, DE-OS 1568810, 1966 (D. K. V. Steel).
135. FMC Corp., US 3250801, 1962 (H. Stange, W. B. Tuemmler).
136. C. H. Fischer, C. Rehberg, L. Smith, *J. Am. Chem. Soc.* **65** (1943) 763 and 1003.
137. Shell Development Co., US 2445627, 1944 (R. C. Morris, R. M. Horowitz).
138. Hooker Chemical Corp., US 3136804, 1964 (J. J. Hodan, C. F. Baranaukas).
139. Tokuyama Soda Co., JP 8102937, 1979.
140. PPG, US 2379218, 1945 (W. R. Dial, C. Gould).
141. PPG, US 4139578, 1975 (G. L. Baughman, H. C. Stevens).
142. Deutsche Spezialglas, DE-OS 2938098, 1979 (H. Fricke, H. Schillert).
143. Hoya Lens K. K., JP 8223901, 1980.
144. A. M. Usmani, *J. Elastomers Plast.* **13** (1981) no. 3, 170 – 176.

145. A. M. Usmani, I. O. Salyer, *J. Mater. Sci.* **16** (1981) no. 4, 915 – 926.
146. North American Philips Corp., US 4350784, 1981 (K. E. Baum).
147. Montecatini, IT 630731, 1961.
148. Stauffer Chemical Co., US 4022849, 1971 (Jung Il Jin, P. Kraft).
149. E. El-Alfy, M. I. Khalil, A. Hebeisch, *J. Polym. Sci. Polym. Chem. Ed.* **19** (1981) no. 12, 3137 – 3143.
150. M. I. Khalil, F. I. Abdel-Hay, A. Hebeish, *Angew. Makromol. Chem.* 1982, 103 and 143.
151. F. I. Abdel-Hay, M. I. Khalil, A. Hebeish, *J. Appl. Polym. Sci.* **27** (1982) no. 4, 1249 – 1258.
152. J. K. Gillham, C. C. Mentzer, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **13** (1972) no. 1, 247.
153. H. Boos, K. R. Hauschildt, *Angew. Makromol. Chem.* **25** (1972) 75.
154. Ethyl Corp., US 3539488, 1970 (O. E. Klopfer, D. Hornbaker).
155. Raychem., GB 1255493, 1971 (R. J. Penneck).
156. G. Kerrutt, *Kautsch. Gummi Kunstst.* **24** (1971) 384.
157. Hitachi Chem., JP 75124986, 1975 (F. Shoji et al.).
158. E. N. Sotnikova et al., *Kozh. Oburvna Promst.* **13** (1971) no. 7, 55; *Chem. Abstr.* **75,** 141737.
159. Yunichik, JP 7038735, 1970 (Y. Mori et al.).
160. K. Nitzl, A. Dietl, DE-OS 1953121, 1971.
161. General Electric Co., DE-OS 1815306, 1969 (F. F. Holub, M. M. Safford).
162. Wacker-Chemie, FR 1546377, 1968.
163. American Cyanamid Co., DE-OS 2119149, 1971 (N. P. Ermidis).
164. General Electric Co., US 4199648, 1978 (H. L. Curry, W. L. Hall).
165. K. Rauer, K. R. Adlassnig, J. Groepper, H. Hofmann, *Res. Discl.* 1981, no. 212, 436.
166. Hoechst, DE-OS 1903954, 1969 (H. Fernholz, F. Wunder, H.-J. Schmidt).
167. J. Pitha, S. Zawadzki, B. A. Hughes, *Macromol. Chem.* **183** (1982) no. 4, 781.
168. Monsanto Co., US 4289864, 1980 (D. N. van Eenam).
169. Esso Research and Eng. Co., FR 2071240, 1969.
170. Elf France, BE 885670, 1979 (R. Leger, R. Nouguier, J. C. Fayard, P. Maldonado).
171. Ciba-Geigy, EP-A 46731, 1980 (F. Lohse, C. E. Monnier).
172. Bayer, DE-OS 2437789, 1974 (H. Haupt).
173. J. D. Burrington, C. T. Kartisek, R. K. Grasselli, *J. Catal.* **75** (1982) no. 2, 225 – 232.
174. D. Peric, A. T. Bell, M. Shen, *J. Appl. Polym. Sci.* **21** (1977) no. 10, 2661 – 2673.
175. P. V. Hinman, A. T. Bell, M. Shen, *J. Appl. Polym. Sci.* **23** (1979) no. 12, 3651 – 3656.
176. M. I. V. S. Florencio, *Rev. Port. Quim.* **22** (1980) no. 1 – 2, 48; *Chem. Abstr.* **96,** 55956.
177. Texaco Dev. Co., DE-OS 2 750 250, 1976 (J. F. Knifton).
178. C. Dauphin, L. David, B. Jamilloux, A. Kergomard, H. Veschambre, *Tetrahedron* **28** (1972) no. 4, 1055.
179. M. F. Sorokin, E. L. Gershanova, SU 328124, 1969.
180. Phillips Petroleum Co., US 3652456, 1969 (F. E. Naylor).
181. A. Ageev, A. I. Ezvielev, E. S. Roskin, L. D. Mazo, SU 323409, 1966.
182. Shell Development Co., US 2216548, 1938.
183. M. Tamele, C. J. Ott, K. E. Marple, G. Hearne, *Ind. Eng. Chem.* **33** (1941), 115 – 120.
184. *Org. Synth. Coll.* **2** (1943) 24.
185. BASF, DE 845516, 1940 (W. Stade, W. Flemming).
186. *Org. Synth. Coll.* **1** (1941) 201.
187. Monsanto Co., US 4091019, 1976 (R. A. Keppel, J. S. McConaghy).
188. Y. D. Smirnov, L. I. Saltikova, A. P. Tomilov, *Zh. Prikl. Khim. (Leningrad)* **43** (1970) no. 7, 1620.
189. Shell Development Co., US 2456428, 1944 (J. H. Parker).
190. J. M. Taylor et al., *Toxicol. Appl. Pharmacol.* **6** (1964) 378.
191. L. Boquin, D. Shuwei, Y. Airu, X. Yinlin, G. Taibao, C. Tao, *Ecotoxicol. Environ. Saf.* **6** (1982) 19 – 27.
192. T. Torkelson, M. A. Wolf, F. Oyen, V. K. Rowe, *Am. Ind. Hyg. Assoc. J.* **20** (1959) 217.
193. H. Fengsheng, S. Dingguo, G. Yupu, L. Boquin, *Chin. Med. J. (Peking, Engl. Ed.)* **93** (1980) no. 3, 177.
194. F. He, J. M. Jacobs, F. Scaravilli, *Acta Neuropathol.* **55** (1981) no. 2, 125.
195. B. L. van Duuren: "Ethylene Dichloride, a Potential Health Risk," *Bambury Report,* vol. 5, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1980.
196. B. L. van Duuren, B. M. Goldschmidt, G. Loewengart, A. C. Smith, S. Melchionne, I. Seldman, D. Roth, *JNCI J. Natl. Cancer Inst.* **63** (1979) no. 6, 1433.
197. J. C. Theis, M. B. Shimkin, L. K. Poirier, *Cancer Res.* **39** (1979) 391.
198. E. K. Weisburger, *EHP Environ. Health Perspect.* **21** (1977) 7.

199. Carcinog. Tech. Rep. S – U.S., NCI-CG-TR-73, National Cancer Institute, USA 1978.
200. L. Fishbein, *Sci. Total Environ.* **11** (1979) no. 2, 111.
201. E. Eder, T. Neudecker, D. Lutz, D. Henschler, *Chem. Biol. Interact.* **38** (1982) 303.
202. E. C. McCoy, L. Burrows, H. S. Rosenkranz, *Mutat. Res.* **57** (1978) 11.
203. M. Bignami, G. Conti, L. Conti, R. Crebelli, F. Misuraca, *Chem. Biol. Interact.* **30** (1980) 9.
204. S. M. Alizade, F. G. Guseinov, L. I. Denisko, *Azerb. Med. Zh.* **59** (1982) no. 2, 66.
205. U.S. Department of Health, Education, and Welfare, HEW-Publication No. (NIOSH) 76-204.
206. R. Kühn, K. Birett, *Merkblätter gefährliche Arbeitsstoffe,* 15th suppl., 12/81 A 47, Ecomed Verlagsgesellschaft, Landsberg 1981.
207. Berufsgenossenschaft der chemischen Industrie: *Technische Regeln für gefährliche Arbeitsstoffe, MAK-Werte 1983,* UVV der BG-Chemie, Anlage 4, Jedermann Verlag, Heidelberg 1983.
208. American Conference of Governmental Industrial Hygienists (ACGIH, ed.): *Threshold Limit Values 1982,* Cincinnati, Ohio 1982.
209. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentrationen* (MAK) 1982, Verlag Chemie, Weinheim 1982.
210. Deutsche Solvay, Note CMP-8049, V-1, 47 (internal document), 1976.
211. M. Häusler, R. Lenich, *Arch. Toxicol.* **23** (1968) 209.
212. M. K. Dunlap et al., *AMA Arch. Ind. Health* **18** (1958) 303.
213. H. F. Smyth, *Am. Ind. Hyg. Assoc. J.* **17** (1956) 129.
214. Shell Chemical Corp., *Ind. Hyg. Bull. SC* 1957, 57 – 78.
215. C. P. McCord, *JAMA J. Am. Med. Assoc.* **98** (1932) 2289.
216. O. Strubelt et al., *Arch. Toxicol.* **22** (1967) 236.
217. H. Oettel, *Hippokrates* **40** (1969) 285.
218. American Conference of Governmental Industrial Hygienists, *Am. Ind. Hyg. Assoc. J.* **23** (1962) 419.
219. P. G. Pagella, D. Faini, C. Turba, *Arzneim. Forsch.* **31** (1981) 1448.
220. H. M. Maling, B. Highman, M. Williams, W. Saul, W. Butler, Jr., B. B. Brodie, *Toxicol. Appl. Pharmacol.* **27** (1974) 380.
221. F. M. B. Carpanini, I. F. Gaunt, J. Hardy, S. D. Gangoli, K. R. Butterworth, *Toxicology* **9** (1978) 29.
222. F. A. Patty: *Industrial Hygiene and Toxicology,* **vol. 2,** Wiley-Interscience, New York 1963, p. 1907.
223. *Am. Ind. Hyg. Assoc. J.* **30** (1969) 470.
224. C. H. Hine, J. K. Kodama, J. S. Wellington, M. K. Dunlap, H. H. Anderson, *AMA Arch. Ind. Health* **14** (1956) 250.
225. C. H. Hine, V. K. Rowe: *Patty's Industrial Hygiene and Toxicology,* 2nd ed., **vol. 2,** Wiley-Interscience, New York 1963, p. 1598.
226. H. F. Smyth et al., *J. Ind. Hyg. Toxicol.* **31** (1949) 60.
227. J. V. Marhold, in *Sborník výsledků toxikologickeho vyšetření latex. pripravku,* Institut pro výchovu vedoucích pracovníků Chemickeho průmyslu, Praha, Prague 1972, p. 38.
228. C. J. Kirwin, E. E. Sandmeyer in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A,** Wiley-Interscience, New York 1981, p. 2515.
229. C. H. Hine et al., *Arch. Environ. Health* **1** (1960) 343.
230. R. J. Guzman, *Arch. Environ. Health.* **2** (1961) 62.
231. R. R. Beard, J. T. Noe in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 B,** Wiley-Interscience, New York 1981, p. 3157.
232. P. J. Boor, E. S. Reynolds, *Am. J. Pathol.* **86** (1977) 49 a.
233. P. J. Boor, M. T. Moslem, E. S. Reynolds, *Toxicol. Appl. Pharmacol.* **50** (1979) 581.

Alumina → Aluminum Oxide

Aluminothermic Processes

In this keyword, only the general aspects of aluminothermic processes are discussed: The technical production of the individual metals and alloys is described in the corresponding keywords, e.g.: ferroboron → Boron and Boron Alloys; chromium and ferrochromium → Chromium and Chromium Alloys; ferromolybdenum → Molybdenum and Molybdenum Compounds; ferroniobium → Niobium and Niobium Compounds; ferrotitanium → Titanium and Titanium Alloys; ferrotungsten → Tungsten, Tungsten Alloys, and Tungsten Compounds; ferrovanadium → Vanadium and Vanadium Compounds

FRED W. HALL, formerly London & Scandinavian Metallurgical Co. Ltd., Rotherham, United Kingdom

1. Introduction 1
2. Basic Metallurgy 2
2.1. Basic Thermochemistry 2
2.2. Fundamental Reactions 4
2.3. Auxiliary Reactions 4
2.4. Fluxes 5
2.5. Formation of Intermetallic Compounds 5
2.6. Thermochemistry in Practice 6
3. Commercial Production 7
3.1. Pot Design and Construction 8
3.2. Charge and Charge Preparation . . . 9
3.3. Feeding 9
3.4. Further Processing 10
3.5. Safety Precautions 10
4. Other Uses of Aluminothermic Processes 10
5. Possible Further Developments 10
6. References 11

1. Introduction

The usual method of reduction by carbon is inapplicable to the reduction of refractory oxides and ores. The products commonly covered under this title are usually required to be low in carbon and they are avid carbide formers. Also, no conventional refractories would stand up to the long reduction times at the high temperatures required. For such products, advantage is taken of the high temperatures achieved and the rapidity of metallothermic reactions using reactive metals, notably Al.

Early examples of such reactions are BERZELIUS' reduction of K_2TaF_7 by Na and WÖHLER's reduction of $AlCl_3$ by Na, in 1825 and 1828, respectively. However, the father of aluminothermic processes was undoubtedly GOLDSCHMIDT who, in 1898, described the production of low-carbon, high-melting-point metals, without extraneous heat, by feeding an exothermic mix into an already ignited first portion [1].

The term "aluminothermic processes" can cover a wide field. For the present purposes it is taken, as is generally understood, to be the manufacture, by Al reduction of refractory oxides or ores, of metals and alloys mainly used in the steel and superalloy industries. Because some important alloys are produced aluminosilicothermically side by side with those produced aluminothermically, they are included in the survey.

The main alloys and metals under discussion are Cr metal [*7440-47-3*], FeB [*11108-67-1*], FeMo [*11121-95-2*], FeNb [*11108-69-3*], FeTi [*12719-90-3*], FeV [*12604-58-9*] and FeW [*12604-57-8*]. There are many minor variations, such as CrC, where superior purity or the absence of Fe is required for alloys used in superalloy production, and FeNb from Nb_2O_5 instead of ore, where extra purity is required, again for superalloys.

Aluminothermic Mn [*7439-96-5*], formerly of importance, has been superseded by electrolytic Mn, and aluminothermic production of FeTi has largely given way to induction furnace melting of Fe and Ti scraps. The latter method also now substitutes for the former, dangerous, highly exothermic production of 55 – 60 % TiAl and 55 – 60 % ZrAl. When possible, combined electroaluminothermic methods can be used.

Because most basic reactions require extra heat this is provided by Al and a vigorous oxidant. Aluminum is an expensive fuel. At an Al

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_447

powder price of 1400 $/t and a small-scale electricity cost of 4.9 C/kW h the heating effect of Al combustion costs about three times that of electricity.

Although a minor proportion of production, some FeB, FeNb, FeTi, and FeV is produced in the arc furnace. In view of the small tonnages involved, such productions usually occur only where an arc furnace exists for other purposes.

Apart from the necessity of metallothermic reduction advantages of the process as compared to conventional reduction processes are:

- Very rapid reaction with less heat loss by radiation and convection
- Much less gas volume
- Easy accomodation of diverse productions
- Small plant investment (important, considering the small production).

A disadvantage of metallothermic production is that no refining of the metal is possible. It is important to leave the slag in situ to allow metal drops to settle into the regulus, so that one is faced with a large depth of slag quickly crusting over. Alternatively, if the slag is tapped, requiring subsequent ore dressing processes to recover included metal, the metal regulus quickly becomes sluggish. Therefore, the composition of the mix must be regulated very carefully to produce the optimum combination of oxygen and reactant metal levels in the product, and all ingredients must be low in harmful impurities. Obvious ones are C, S, and P. Other impurities in the various ores used for different products are As, Sn, Pb, Sb, and Cu. Nitrogen is not usually a problem because the metal regulus formed is protected by the slag layer. Chromium metal, used in superalloys and electric heating elements, is an extreme case, where, apart from low impurities generally, certain impurities, such as Ga, Pb, and B, must be kept down to a few parts per million.

The worldwide – excluding the former USSR and China – production of metal and alloys by aluminothermic processes is shown in Table 1.

Table 1. Production of metal and alloys by aluminothermic processes

	t/a	Expressed as
Cr	4550	Cr
FeB	1350	18 % B
FeMo ∗	40500	65 % Mo
FeNb	16300	65 % Nb
FeV	25000	80 % V
FeW	4000	80 % W

∗ The FeMo represents about one third of the total Mo used in the steel industry; the remaining two thirds are used as MoO_3.

2. Basic Metallurgy

A metallothermic reduction of a metal compound is possible when the reductant metal has a greater affinity for the nonmetal element of the compound than the desired metal. In various branches of metallurgy the nonmetal may be a halide, sulfur, or oxygen. The reductant metal is usually Al, Si, Mg, Ca, or Na. For instance, Ti metal is produced by reduction of halides with Na or Mg and V metal by reduction of V_2O_5 with Ca in a bomb reaction (to prevent vaporization of the Ca). Interesting examples are the reduction of NiO and MoO_3 by Fe in the arc furnace. Aluminothermic processes, as previously defined, depend upon the high affinity of Al for oxygen as compared to that of many other metals, an example being:

$$Cr_2O_3 + 2\,Al \rightarrow 2\,Cr + Al_2O_3$$

2.1. Basic Thermochemistry

Aluminum has a high affinity for oxygen, and the ease of reduction of a metal oxide depends on the difference in oxygen affinity of Al and the metal. The affinity for oxygen is indicated by its heat of formation, ΔH, expressed for comparison purposes as ΔH per mol/oxygen (kJ/mol O). Table 2 gives a selection of ΔH values at the standard temperature of 298 K.

Table 2. Heats of formation of oxides [2]

Oxide	$-\Delta H_{298}$, J/mol oxide	$-\Delta H_{298}$, kJ/mol O
CaO	634.7	634.7
MgO	601.7	601.7
Al_2O_3	1678.5	559.4
ZrO_2	1101.5	550.6
TiO_2	945.4	472.7
SiO_2	911.0	455.5
B_2O_3	1272.8	424.1
Na_2O	415.3	415.3
Ta_2O_5	2047.3	409.5
Nb_2O_5	1900.8	380.2
Cr_2O_3	1130.4	376.8
V_2O_5	1551.6	310.2
WO_3	843.2	280.9
Fe_3O_4	1117.5	279.3
Fe_2O_3	821.9	273.8
FeO	264.6	264.6
MoO_3	745.7	248.7

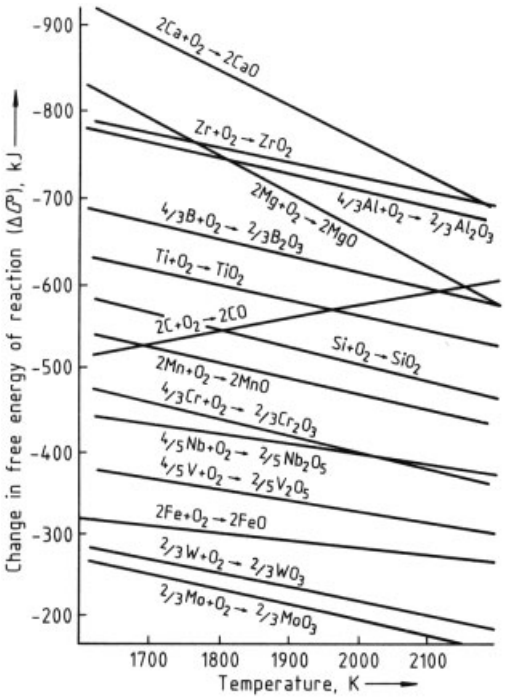

Figure 1. Change in free energy of oxide formation reactions [3]

A quantitative measure for the stability of an oxide is the Gibbs free energy of formation, ΔG, which is related to the heat of formation by the formula:

$$\Delta G_T^0 = \Delta H - T\Delta S$$

where S is the entropy and is always zero or positive because it corresponds with the increasing atomic disorder of the system with temperature. Figure 1 shows the free energy of formation of the oxides of interest at high temperatures. All oxides become less stable as temperature increases. The more negative the free energy of formation at a given temperature the greater the tendency for that metal to reduce an oxide of less negative free energy.

The free energies of reaction of oxides with Al and Si are given in Figures 2 and 3.

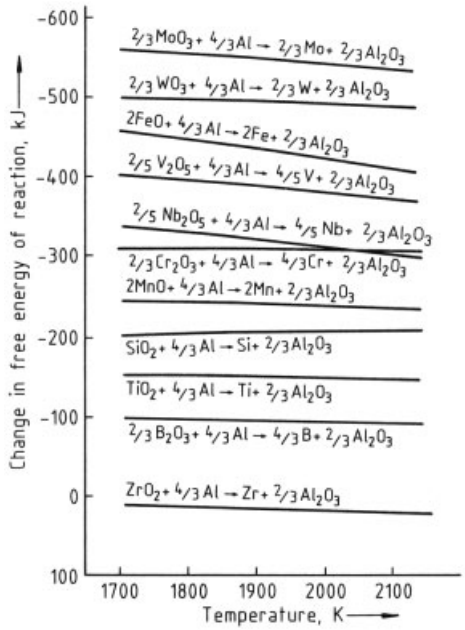

Figure 2. Change in free energy of reaction in the reduction of oxides by aluminum

Although the free energy of reaction is a quantitative measure, not all data necessary for such complex reactions as commercial metallothermic reactions are readily available and it is usual for the practicing metallurgist to rely on heats of formation for calculation of heats of reaction.

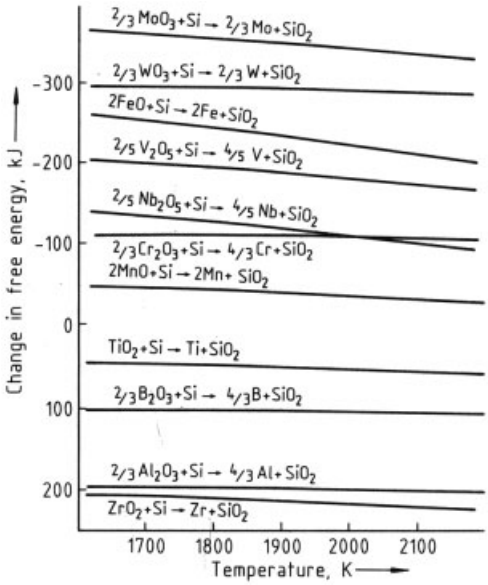

Figure 3. Change in free energy of reaction in the reduction of oxides by silicon

2.2. Fundamental Reactions

Although Ca and Mg may be the most suitable reactants, they are not used in the reductions under review. Their basic costs are high compared to Al and Si and 16 parts by weight O require 40 parts Ca, 24 parts Mg, 18 parts Al, or 14 parts Si. Furthermore, the boiling points of Ca and Mg are 1 484 °C and 1 090 °C, respectively. The heats of reaction of the oxides of interest with Al and Si are given in Table 3.

Table 3. Heats of reaction per mole reductant

	$-\Delta H_{298}$, kJ
3/4 TiO_2 + Al → 3/4 Ti + 1/2 Al_2O_3	130.2
3/4 SiO_2 + Al → 3/4 Si + 1/2 Al_2O_3	156.0
1/2 B_2O_3 + Al → B + 1/2 Al_2O_3	202.6
3/10 Nb_2O_5 + Al → 3/5 Nb + 1/2 Al_2O_3	268.8
1/2 Cr_2O_3 + Al → Cr + 1/2 Al_2O_3	273.8
3/10 V_2O_5 + Al → 3/5 V + 1/2 Al_2O_3	373.9
1/1 Wo_3 + Al → 1/2 W + 1/2 Al_2O_3	417.4
3/8 Fe_3O_4 + Al → 9/8 Fe + 1/2 Al_2O_3	420.2
1/2 Fe_2O_3 + Al → Fe + 1/2 Al_2O_3	428.3
3/2 FeO + Al → 3/2 Fe + 1/2 Al_2O_3	442.1
1/2 MoO_3 + Al → 1/2Mo + 1/2 Al_2O_3	466.4
2/3 WO_3 + Si → 2/3 W + SiO_2	348.8
1/2 Fe_3O_4 + Si → 2/3 Fe + SiO_2	352.1
2/3 Fe_2O_3 + Si → 4/3 Fe + SiO_2	363.0
2 FeO + Si → 2 Fe + SiO_2	381.8
2/3 MoO_3 + Si → 2/3 Mo + SiO_2	414.1

A guide for a successful aluminothermic reaction, permitting adequate separation of metal and slag, is a $-\Delta H_{298}$ higher than 300 kJ/mol [4]. This is confirmed in practice: additional heat is required for Cr production, whereas the reaction producing 80 % FeV is so vigorous that it requires cooling. This guide does not apply to silicothermic reactions.

2.3. Auxiliary Reactions

Many reactions require additional heat to produce fully liquid metal and slag. Except for producing pure Cr, the first step is to make an addition of iron oxide plus reactant, giving a higher average heat of reaction. This also has the advantage of producing significantly lower melting points (Table 4).

Table 4. Melting points of some metals (in °C) and of alloys of these metal with iron

V	1 902	80 % FeV	1 770
B	2 180	16 % FeB	1 550
Nb	2 468	65 % FeNb	1 560
Mo	2 620	65 % FeMo	1 900
W	3 410	80 % FeW	2 500

In silicothermic reactions, a proportion of Al is usually used to increase the exothermicity particularly, because the products, e.g., FeW and FeMo, have such high melting points.

Because the alloy specifications have become traditional, although not necessarily ideal, there is a limit to the amount of iron oxide and Al to be added, and, indeed, it may be more economical to increase the heat content of the mixture furhter, by using highly oxidic compounds plus Al. The reactions and the heats of reaction (in kJ/mol Al) for the additives most commonly used are:

$$K_2Cr_2O_7 + 4\,Al \rightarrow 2\,Cr + 2\,Al_2O_3 + K_2O$$
$$\Delta H_{298} = -\,422.0\,\text{kJ/mol Al}$$

$$3\,BaO_2 + 2\,Al \rightarrow 3\,BaO \cdot Al_2O_3$$
$$\Delta H_{298} = -\,812.0\,\text{kJ/mol Al}$$

$$NaClO_3 + 2\,Al \rightarrow Al_2O_3 + NaCl$$
$$\Delta H_{298} = -\,869.2\,\text{kJ/mol Al}$$

$$3\,KClO_4 + 8\,Al \rightarrow 4\,Al_2O_3 + 3\,KCl$$
$$\Delta H_{298} = -\,839.0\,\text{kJ/mol Al}$$

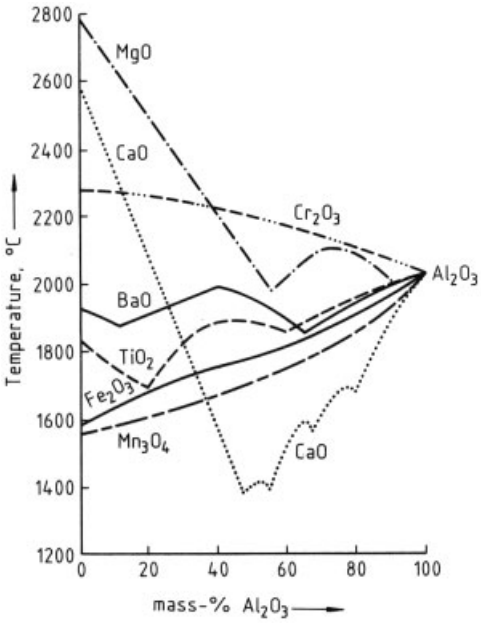

Figure 4. Melting point curves of binary alumina slags

In some cases additional heat was applied formerly extraneously by preheating the charge. This has been given up because of practical difficulties of heating a mixture of low heat conductivity and the danger of premature ignition, with loss of the charge and damage to the preheating equipment. The cheap nitrates are not used now because products low in nitrogen are required and the metals in question are avid nitride formers. BaO_2 is useful because it not only acts as an oxidant, but is a strong base and assists the progress of the reaction by combining with the Al_2O_3 produced. However, its high molecular mass and its one available oxygen atom make it a very expensive oxidant, and its use is confined to highly exothermic mixes, e.g., for TiAl and ZrAl, where it helps to smooth the reaction. The common oxidant is $NaClO_3$, but $KClO_4$ is to be preferred because it is more stable and $NaClO_3$ is hygroscopic. In Cr mixes, $K_2Cr_2O_7$ is used and is preferred to CrO_3 on health grounds as well as expense.

2.4. Fluxes

Both Al_2O_3 and SiO_2 have high melting points, 2050 °C and 1722 °C, respectively. The aim in metallothermic reactions is to produce a fluid slag, so as to permit good separation of slag and metal. In the case of aluminosilico reactions, residual iron oxide and Al_2O_3 are already good fluxes, and lime also is used because the strongly basic CaO, apart from fluxing, lowers the activity of Al_2O_3 and SiO_2.

In aluminothermic reactions residual oxides may play a part, the compound $Al_2O_3 \cdot TiO_2$ having a melting point of 1590 °C; Na_2O and K_2O, not completely vaporized, also assist. The usual flux is CaO which has a pronounced effect on the melting point of Al_2O_3. It is not used in Cr manufacture as it spoils the slag for use in refractories and calcium chromite is a stable compound.

Figure 4 shows the effect on melting point of additions of several oxides to Al_2O_3.

The energy required to fuse the flux is partially compensated by the heats of formation of the compound oxide (Table 5).

Table 5. Heats of formation of compound oxides

Compound	$-\Delta H_{298}$, kJ/mol
$Al_2O_3 \cdot SiO_2$	192.6
$2Al_2O_3 \cdot B_2O_3$	69.5
$BaO \cdot Al_2O_3$	100.5
$CaO \cdot Al_2O_3$	15.1
$CaO \cdot SiO_2$	89.2

2.5. Formation of Intermetallic Compounds

The heat of reaction of the mixture can be increased by the formation of intermetallic compounds. Table 6 gives some examples. However, this is restricted in that most of the higher heats of formation belong to unwanted compounds because the usual aim is to keep residual elements, such as Al and Si, to a minimum. Of the alloys considered it is only in the case of FeTi and to a lesser extent in that of FeB that aluminides have some significance.

Table 6. Heats of formation of intermetallic compounds

Compound	ΔH_{298}, kJ/mol
FeTi	40.6
FeB	29.3
$FeAl_3$	112.2
FeSi	77.0
TiAl	80.8
CaSi	150.7
AlB	67.0

All reactions are equilibria. Commercial alloys must be low in oxygen and most alloys can be produced at 1 % Al or Si without appreciable oxygen content. This is not the case with FeTi, where the equilibrium cannot be driven further than Al levels of 4 – 6 %.

2.6. Thermochemistry in Practice

A commercial metallothermic reaction must provide sufficient heat to drive the reaction as far to the right side of the equation as possible, plus that heat required to bring the products to a temperature sufficiently high to permit separation of metal and slag and to provide for heat losses. The amount of total heat must be the minimum required, because heat is expensive and because excess heat supplied to an exothermic reaction moves the equilibrium in the undesired direction.

Heats of formation have not always been determined at the high operating temperatures. Working on heats of formation at 298 K requires knowing the specific heats, latent heats, and, in charges employing sodium salts and chlorates, the heats of vaporization. Slags can be complex so that little idea can be had of their heats of formation. The amount of heat losses cannot be forecast and varies with the size of charge. Losses can be established and compensated for only by trial and error.

Therefore, available information is used as much as possible, accompanied by intelligent extrapolations of such information. Final adjustments to the charge are a matter of trial and error.

A very simple example is:

$$Cr_2O_3 + 2\,Al \rightarrow 2\,Cr + Al_2O_3 \quad \Delta H_{298} = -548\,kJ$$

Heat required for products at 2000 °C:

Al_2O_3	$H_{2273} - H_{298}$	356 kJ
2 Cr	$H_{2273} - H_{298}$	184 kJ
amounts to:		540 kJ

According to this, there is a heat balance, so there should be a satisfactory result. However, it is found in practice that 125 kJ extra heat are required, provided by Al plus $K_2Cr_2O_7$ and/or $KClO_4$. This 125 kJ (for a certain size of charge) is a measure of the heat losses. More would be required for small charges and less for bigger charges. It is interesting that this conforms with the guide.

Chemical composition of ores is some guide to improving the heat balance. Generally the higher the state of the oxide in the ore, the higher the heat of reaction with Al or Si. A classical example was that of Mn ore where the reaction $3\,MnO + 2\,Al \rightarrow 3\,Mn + Al_2O_3$ did not supply sufficient heat. A preoxidation to Mn_3O_4 produced a satisfactory result. Preoxidation also is found beneficial to some ilmenites.

There are many other factors affecting the fine control of an exothermic reaction, and it is not always economic to indulge in laboratory control of all of them. They mainly affect the speed of the reaction, which should be an optimum to minimize heat losses but not to create excess heat with its adverse effect on the equilibrium.

Granulometry. If the charge is to be reacted in bulk then the ideal would be a charge of concrete-type granulometry, giving maximum contact and minimum air inclusion. However, to provide effective heat control, charges usually are fed. There should be an optimum particle size of the ingredients, therefore, coarse enough to minimize dust losses but fine enough to give an adequate speed of reaction. Aluminum powder can be controlled, because suppliers will supply within reasonable size ranges. In some cases, such as the highly exothermic FeV reaction, Al pellets and turnings can be used and they are used also to some extent in the very competitive production of the cheap alloy FeTi. These are crude raw materials and different parcels require slight adjustments to the mix. Ores are a different matter. These must usually be used as they arise; the coarser they are in any one production, the more the amount of heat required. Where ores are extremely fine, leading to high dust losses, it may be economical to pelletize.

Mineralogic Structure of Ores. Complex ores, such as pyrochlore, can vary widely in structure, and different parcels require adjustments to the mix.

Bulk Density. This is related to granulometry, but even in Cr_2O_3, which would be expected to be uniform from parcel to parcel and where grain sizes are of a few microns, variations in

the supplier's final calcination temperature can affect the bulk density greatly and hence the recovery.

Mass of Charge. The percentage of heat losses varies inversely with the size of charge. These losses cannot be measured, but when the mass of a charge is increased, less heat per unit charge is required; this decrease can be obtained by an intelligent estimate finalized by trial and error.

Rate of Feeding. This is very much an art and efficient feeding can minimize some of the above mentioned difficulties. The general aim is to maintain a cover of unreacted mix on the surface of the liquid slag, so as to minimize radiation losses but not to feed too fast and produce too hot a reaction.

It can be seen from the above that whereas thermochemistry plays a vital role in providing the basic knowledge for any metallothermic production, there are other "fine tuning" adjustments that are a matter of experience.

3. Commercial Production

There are two methods of promoting the reaction. One is to fill the reaction vessel with the charge and ignite it via an easily ignitable mixture, such as BaO_2 or Na_2O_2 plus aluminum; the reaction then proceeds from top to bottom of the charge. This is a wasteful method because only approximately one third of the capacity of the reaction vessel is used. The mix calculation must be very accurate because no control of the speed of reaction can be exercised. Because the charge is always covered by liquid slag, dust losses should be low, but there can be considerable loss by splashing owing to the violence of the reaction. The short reaction time should mean low heat losses, but this advantage is counteracted by the intense radiation from an always liquid surface. This method is seldom used, but there are some reactions that are so exothermic that they are dangerous and practically impossible to feed, such as those for the minor products 55 – 60 % TiAl and ZrAl. This "firing down" method is used sometimes also for FeW production, where the extra heat supplied to achieve the high melting point requires a highly exothermic reaction near the limit of feedability.

The second method, almost universal, is to use a charge slightly more exothermic than that for "firing down," to ignite a small portion in the reaction vessel, and to feed the remaining charge. By doing so, the full capacity of the vessel is used so that vessel preparation costs per unit production are approx. one third those of the "firing-down" method. Another great advantage is that the speed of the reaction can be controlled by varying the feeding rate so that it gets neither too hot nor too cold. As mentioned previously, this is quite an art, but it is surprising how soon a skilled worker becomes used to visual control of the temperature of the reaction.

There are also two types of reaction vessel. One is a mobile steel pot, lined with refractory. The other is a sand bed into which cylindrical holes are dug and lined with firebrick. This latter method usually is confined to aluminosilico reactions, such as those for FeW and FeMo.

Pot reactions are carried out in firing chambers, where the under pressure, as compared to that outside, is kept to the minimum required for complete exhaust of fume consistent with the charge feeder's need to observe the reaction but to keep dust losses to a minimum. *Sand bed reactions* are usually under a movable hood connected to the main flue. Whereas tall chimneys were once the solution to disposal of fume, modern conceptions of pollution control demand efficient dust collection. Owing to the volatility of MoO_3, this production demands its separate dust collector, the recovery from which makes the difference between profit and loss on the operation. For other productions, because of the small tonnage of individual products and because the fume is largely Al_2O_3, CaO, NaOH, etc., with low contents of valuable oxide, a communal collector is used. Collectors are of many types, but the most efficient is the reverse-jet filter. Because of the presence of CaO, NaOH, etc., the filter, when not in use, must always be kept above 100 °C to prevent caking of the fume on the bags.

Figure 5 shows the arrangement, in principle, for feeding, reaction in a chamber, and fume extraction. After firing, the reaction pots must be left to cool to allow final settling out of metal droplets and eventual solidification. Depending on the size of the reaction, solidification can take

a long time, but 24 h, which often fits in with the overall routine of the production, is usually sufficient. The pot is then stripped, leaving a block of metal with adherent block of slag. The whole is then carefully quenched in a water tank, which causes the slag block to separate and embrittles the metal to some extent, facilitating subsequent crushing.

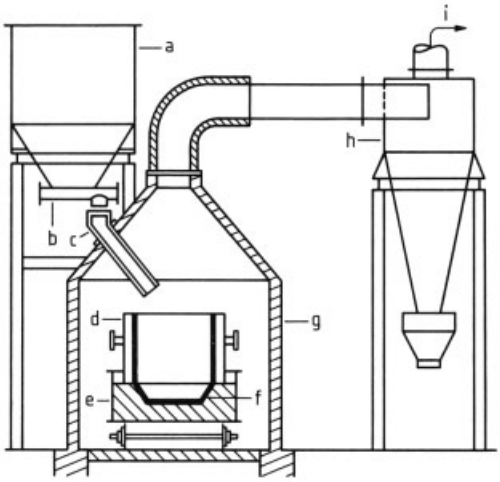

Figure 5. Reaction chamber with feeding hopper, reaction vessel, and dust collector
a) Feeding hopper; b) Feed-regulating valve; c) Feeding chute; d) Separable refractory-lined steel plate cylinder; e) Refractory-lined base; f) Optional magnesite inner lining; g) Firebrick-lined chamber; h) Filter; i) Offtake to chimney

The sides of the metal block are surrounded by an adherent mixture of metal and slag created by penetration into the refractory lining. This easily breaks away and is consumed in the next reaction. The top surface, containing pockets of very adherent slag, requires removal by automatic hammers and, in the last resort, by hand chipping hammers. The bottom of the block usually requires less attention. Final cleaning of the whole block is by shot blasting.

The block is coarsely broken up by a mobile hammer of the pile driver type and crushed to desired size in jaw crushers and/or rotary crushers with intermediate screenings. There is immense variety in customer's size requirements, ranging from, say, 15-cm lumps for bulk users down to graded powders for welding rod manufacturers. An increasing demand is for bags of 6 mm and down alloy, each containing a fixed weight of the desired alloying element.

3.1. Pot Design and Construction

Although there is theoretically no limit to the size of the charge, as compared to steel production, the weights are small for the following reasons:

Costs. Overall, raw materials are expensive. It is the aim of a producer to try to keep pace evenly with sales, holding modest stocks of raw materials and finished product. The total annual production of any one producer is small and diverse, with, in some cases, several specifications for one product.
Handling Facilities. The total weight of a refractory-lined pot full of metal and slag is about three times that of the metal produced. Crane size is limited by cost and the small building required for carrying out the reactions, and for total operation lift trucks are convenient.
Preliminary Metal Breaking. The small tonnages involved preclude a very expensive, large, preliminary breaking plant. Some alloys, particularly FeMo, are very tough and metal block thickness is usually limited to 30 – 35 cm.
Metal Specification. Some residual specifications are very tight, notably that for Al in Cr metal. A mishap in the weighing of the charge can produce an off-grade block that may be in stock for a long time.
Refractory Failure. Although rare, there can be an occasional failure of the refractory, producing breakout of metal and slag at high economic loss.

Therefore, metal block weights usually vary between 1000 and 2500 kg and, depending on the density of metal and slag, two pot sizes suffice. A small pot of ca. 200 kg metal capacity is useful for intermediate size investigations on new raw materials. A typical pot covering 1000 – 1750 kg metal would have internal dimensions of ca. 100 cm diameter and 125 cm height and one for 2500 kg would have the same height but be of 120 cm diameter in order to avoid too thick a metal block.

The outer construction consists of a plate steel cylinder attached to a heavy steel base by quick-acting lugs. The base may be on wheels or adapted for lift truck operation. The lining

is of moistened, crushed slag from chromium or ferrovanadium production, because this is a reasonably pure corundum. To ease cleaning of the metal block, magnesite tiles may be used in the metal zone, particularly for the hotter reactions. After preparation of the refractory base, either from bricks or ramming of crushed slag, the sides are rammed around a tapered, withdrawable template. Ramming is by hand, by pneumatic or electric hammer, or by jolting on a foundy jolt molding machine. The lining must then be dried for several hours to remove moisture. It is taken up to a dull red heat.

Basic sand bed operation is to construct a well in the sand bed from unmortared firebricks which takes the whole molten charge. After sufficient time is allowed for metal to settle out, a channel is made through the sand at the slag-metal interface level and the slap tapped off. A variation is to use a similar well in the sand bed sufficient to accomodate only the metal, topped by a firebrick-lined steel shell to accommodate the slag, with ample sand seal at the junction. After the reaction, the slag is tapped at the junction and the shell used several times.

3.2. Charge and Charge Preparation

Ideally, there would be an optimum size for raw materials, fine enough for a successful reaction but as coarse as possible to minimize dust losses. Fortunately, most oxides and ores are of reasonable size in these respects. Wolfram ores may have to be ground sometimes and it is known that very fine ores, such as pyrochlore, have been pelletized. Chromium oxide, at a grain size of a few µm, can be expected to give high dust losses but the grains have a strong tendency to self-adherence so that the particles are much larger. Even so, there is a slightly noticeable difference in recovery when different oxides are used, this varying directly with the final calcining of the oxide during its manufacture. The calcination increases grain size and density.

Auxiliary oxidants are of satisfactory grain size with the proviso that $NaClO_3$ tends to agglomerate on storage and must be sieved down to max. 1 mm before use.

The minimum size of Al is limited by dust losses and the risk of dust explosions. It should not be below ca. BSS 240 = 64 µm. The size used in exothermic reactions then depends on the reaction. For chromium it is on the order of average 0.25 mm, for FeB and FeNb, 0.45 mm, and the Al powder can be partially substituted by cheaper foil powder. In the highly exothermic FeV reaction, where the V_2O_5 is in the form of flakes or lumps, only a little coarse powder is used to assist the start of the reaction, the bulk being pellets, chopped wire, and/or pure turnings. Despite the fine size of ilmenite and rutile, the same applies to FeTi production. Owing to the relative cheapness of the ores, Al is the most expensive ingredient, and cheap aluminum is used at the expense of Ti recovery.

In aluminosilico reductions only powders are used, in the case of FeMo because the ore is fine, in the case of FeW to obtain maximum speed of reaction so as to achieve the high temperature required. The reductants are Al powder and 75 % FeSi, on the order of 0.25 mm average.

The amount of flux used is kept to a minimum in aluminothermic charges. Where sodium salts are used as auxiliary oxidants, residual Na_2O already lowers the melting point of the slag. A calcium salt is the usual flux because it has such a significant effect on the melting point of Al_2O_3 (see Fig. 4). Instead of CaO, CaF_2 often is used as its low melting point of 1418 °C assists in promoting a smooth reaction. The amount used is often 1 – 2 % of the total charge, never more than 5 %.

Significant quantities of CaO – CaF_2 are used in aluminosilico thermic reactions because it is important for their success to tie the SiO_2 produced. As can be seen from Table 4, CaO has a strong affinity for SiO_2.

After the components have been weighed out accurately, they are mixed completely in a gently acting mixer, such as a drum with lifters, V blender, or rotating cube. Duration of mixing can be determined only by practice and can be determined initially by analyzing samples taken at intervals.

The charge is then dropped into a feeding hopper.

3.3. Feeding

When a charge has been prepared accurately to produce the desired result, success depends on the efficiency of feeding. The usual method is to

control the exit valve of the feeding hopper according to the visual observation of the speed of the reaction, making slight adjustments to compensate for momentary coldness or hotness. Additional finer control is often applied by means of a handrake in the feeding chute. A fast reaction, such as that for FeV, often requires cooling, and a second small hopper may be used to feed a portion of crushed FeV slag. A FeW reaction is fed as fast as possible.

3.4. Further Processing

Pot stripping, block cleaning, and crushing require no further description than described in the introduction to Chapter 3, except to emphasize the importance of avoiding cross contamination. This is particularly the case with FeB. A small amount of B is very beneficial to a steel designed for it but an unknown amount can be disastrous.

3.5. Safety Precautions

It must be realized that exothermic production, inadequately supervised, can be a very dangerous operation. Aluminum powder and oxidants are themselves fire hazards, and risk is magnified, when they are in contact, maybe to explosive proportions.

Aluminum powder and oxidants must be stored separately and the latter kept clear of carbonaceous material. Only sufficient for a single charge should be brought to the mixing department, and, during weighing, the oxide or ore should be sandwiched between the Al powder and the auxiliary oxidant to avoid direct, concentrated contact.

Mixing plants should be grounded to avoid buildup of static electricity. During firing, overfeeding can cause violent eruptions. Lumpy $NaClO_3$ can cause explosions. In case of an inadvertent lining failure, a large enough well should be provided to accept all the charge.

The main causes of disastrous explosions are Al dust explosions, either primary, because of careless handling, or secondary, over an overfed reaction. It also has been known for the metal and slag, apparently solid, to be quenched while still liquid internally, creating a disastrous explosion.

Employees must be protected from inhalation of dust and physical contact with charge components. This is particularly the case with hexavalent compounds of Cr, which can cause the well-known chrome ulcers.

Housekeeping must be exceptional. Fine dust accumulated on roof beams has been known to combust spontaneously.

4. Other Uses of Aluminothermic Processes

A significant use of an aluminothermic process is the welding of railway and tramway line joints, where the reaction takes place in a small crucible, and the steel produced is tapped onto the joint. In a similar manner, heavy cracked steel components, such as rolls, are welded and worn parts are built up and remachined. Developments over the last few years are the repair of ingot mold stools, where the steel stream has worn a cavity, and of cracked ingot mold lugs.

The use of exothermic reactions in incendiary bombs, where the intense $Fe_2O_3 - Al$ reaction ignites the magnesium casing, is well known. Other military uses are for flares, which in the case of sea rescue give off colored smokes.

Milder reactions, damped down by insulating materials, are used in feeder head compounds, which are spread over the surface of liquid steel ingots in order to conserve the heat and minimize piping. For steel castings, similar compositions are formed into sleeves inserted in the sand mold to insure liquidity over the whole of the pouring time. Where large amounts of alloying addition are made to steel in the ladle, the cooling effect of the addition may be counteracted by the use of exothermic briquettes composed of alloy, Al, and oxidant. Minor uses are for jointing of Al and Cu cables by a $CuO - Al$ reaction, and as slow reactions for hand warmers and food cans for explorers.

5. Possible Further Developments

Wherever alloys with low content are required, aluminothermic processes are necessary, and they have been brought to such an efficiency that little further improvement can be foreseen.

Improved refractories may permit further transfer of some reactions to the arc furnace, saving Al plus auxiliary oxidant, but at present, those which can be envisaged possibly to stand up to the longer exposure are very expensive and have to compete with the self-produced refractory from an exothermic reaction.

Plasma processes are a possibility, the very high temperatures encouraging the oxide–carbon reaction, but up to now there has been no success in driving the reaction to low C levels. Here again, refractories are a problem.

A relaxation of maximum impurity levels would be of mutual economic benefit to alloy maker and steel maker. Stringent specifications for superalloys are obvious, but when the small proportion of alloy used in many steels and the wealth of purification methods used in bulk steel production are considered, then some relaxation in impurity levels should be possible, enabling use of cheaper, more impure ores. Higher Al and/or Si levels would give better recovery and, where present specifications are very low, lower oxygen levels.

6. References

1. H. Goldschmidt, *Justus Liebigs Ann. Chem.* **301** (1898) 19 – 28.
2. O. Kubaschewski, C. B. Alcock: *Metallurgical Thermochemistry,* 5th ed., Pergamon Press, London-New York 1979.
3. N. P. Elyutin et al: *Production of Ferroalloys,* 2nd ed., Israel Programme for Scientific Translations, S. Monon, Jerusalem 1961.
4. W. Dautzenberg, in: Dürrer & Frank, (G. Volkert, K.-D. Frank eds.): *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin-Göttingen-Heidelberg-New York 1972.
5. E. M. Levin et al.: *Phase Diagrams for Ceramists,* American Ceramic Society Inc., Columbus, 1964, Supplement 1969.
6. W. G. Moffatt: *Handbook of Binary Phase Diagrams,* General Electric Corp., Schenectady 1973 (plus updates).

Aluminum

William B. Frank, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chaps. 1 – 9)
Warren E. Haupin, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chaps. 1 – 9)
Robert K. Dawless, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chap. 5)
Douglas A. Granger, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chap. 6)
Maurice W. Wei, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chap. 7)
Kenneth J. Calhoun, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chap. 8)
Thomas B. Bonney, Aluminum Company of America, Alcoa Center, Pa. 15069, United States (Chap. 9)

1. Introduction 1
2. Properties of Pure Aluminum 2
2.1. Mechanical Properties 2
2.2. Physical Properties 3
3. Chemical Properties 4
3.1. Oxidation of Aluminum 4
3.2. Reactions with Aqueous Solutions . 5
3.3. Reactions at High Temperatures . . 5
3.4. Corrosion 5
3.5. Gases and Aluminum 5
4. Production 6
4.1. History of the Electrolytic Reduction of Alumina 6
4.2. Raw Materials 6
4.3. Hall–Héroult Cell for Aluminum Production 8
4.3.1. Electrolyte 9
4.3.2. Electrode Reactions 10
4.3.3. Current Efficiency 12
4.3.4. Cell Voltage 12
4.3.5. Heat Balance 13
4.3.6. Fluid Dynamics 13
4.4. Thermodynamic Considerations . . 14
4.5. Alternate Processes 14
5. Producing High-Purity Aluminum . 15
6. Aluminum Casting: Remelt Ingot . 16
7. Environmental Protection 17
7.1. Air Emission 17
7.2. Wastewater Discharge 18
7.3. Solid Waste 19
7.4. Regulatory Requirements 19
8. Economic Aspects 19
9. Toxicology and Occupational Health 20
10. References 21

1. Introduction

Aluminum [*7429-90-5*] is the most abundant metallic element in the earth's crust. It is normally found combined with other elements, and occurs rarely, if at all, in its pure state. However, elemental aluminum particles have been discovered in lunar soil [13]. Aluminum appears in a wide variety of minerals combined with oxygen, silicon, the alkali and alkaline-earth metals, and fluorine, and as hydroxides, sulfates, and phosphates. Aluminum has become the predominant nonferrous metal in use, yet it is one of the newest of the common metals. Aluminous minerals are quite stable; large amounts of energy and high temperatures are required to reduce these compounds to metal. This is the reason often given to explain why the metal has been isolated and produced commercially only in relatively recent times.

In the late 18th and early 19th centuries many famous scientists worked at isolating aluminum, including Lavoisier, Dalton, Berzelius, Davy, and Oersted. In 1825 Hans Christian Oersted discovered a method to prepare anhydrous aluminum chloride, a compound that was to play an important role in the production of the metal. Oersted is sometimes credited with the discovery of aluminum but more often that honor is accorded to Friedrich Wöhler. His isolation of aluminum in 1827 was based on the reaction in a porcelain crucible of anhydrous aluminum chloride with potassium.

In 1854, H. Sainte-Claire Deville improved the method of preparing aluminum by substituting sodium for potassium as the reductant for aluminum chloride. He was the first to

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_459

record that fused cryolite serves as a solvent for aluminum oxide. DEVILLE emerges as the outstanding figure in aluminum technology in the period after its discovery and before commercialization of the Hall–Héroult process. His development of the chemical reduction process raised aluminum from a laboratory curiosity to a useful metal.

The chemical production process was further refined throughout the second half of the 19th century and the aluminum industry grew significantly during this period. Chemical production of aluminum by reduction of cryolite or aluminum fluoride with sodium or potassium was also under development in the late 19th century. The metal exhibited at the Paris Exposition of 1889 was produced at Alliance Aluminium of London, probably by sodiothermic reduction of cryolite. Reduction methods were used into the 1890s, when electrolyte production became dominant.

In the late 19th century many unsuccessful attempts were made to electrolyze aluminum from aqueous solutions. Aluminum was first prepared electrolytically by a method discovered independently in 1854 by DEVILLE in France and BUNSEN in Germany. A sodium chloride-aluminum chloride fusion was electrolyzed in a porcelain crucible using carbon electrodes. Production based on this electrolysis of aluminum chloride was begun in 1975 but operation was suspended in 1982. Development of the process is continuing (see Section 4.5).

For a discussion of the Hall–Héroult electrolytic process, see Section 4.1.

2. Properties of Pure Aluminum

Certain physical and chemical properties of aluminum depend primarily on purity. There is no generally accepted nomenclature for the degrees of purity of aluminum. The following classification is suggested:

% Aluminum	Designation
<99.5	alloy or scrap
99.5 – 99.9	commercial purity
99.9 – 99.99	high purity
>99.99	super purity

The aluminum produced by the Hall–Héroult process is of commercial purity. Iron and silicon are the predominant impurities. Aluminum of super purity, produced by the three-layer electrolytic process, first became available in 1920 (see Chap. 5). The properties of this grade of metal have been studied extensively. Methods for preparing aluminum of even higher purity include zone refining, fractional crystallization, and preparation from aluminum alkyls. Electrical resistivity at low temperatures is employed as a measure of purity for high purity and super purity aluminum. Newer methods for analysis of trace impurities, including activation analysis, have improved the sensitivity and scope of analyses for very pure materials. Aluminum purer than 99.9999 % has been prepared and characterized [14].

Many applications of aluminum and its alloys are based upon its inherent properties of low density, high electrical and thermal conductivities, high reflectivity, and great resistance to corrosion. Pure aluminum is soft and lacks strength, but it can be alloyed with many other elements to increase strength and impart a number of useful properties. Alloys of aluminum are light, strong, and readily formable by many metal-working processes; they can be cast, joined, or machined easily and accept a wide variety of finishes. The properties and processing of alloys are discussed under the keyword → Aluminum Alloys.

2.1. Mechanical Properties

The mechanical properties of aluminum and its alloys are discussed under → Aluminum Alloys. Some mechanical properties of aluminum of several purities are given in Table 1 [15].

Table 1. Mechanical properties of pure aluminum at room temperature

Purity	Tensile strength		Elongation in 50 mm
%	MPa	ksi *	%
99.99	45	6.5	50
99.8	60	8.7	45
99.6	70	10.2	43

* kips (1000 pounds) per square inch

The data are from different sources and caution should be exercised in any direct comparison. Difficulties occur because of problems in anal-

ysis, temper of the specimen, and test methods. However, even small amounts of impurities present in commercial aluminum raise the tensile strength and hardness over that of the purest aluminum in the table.

2.2. Physical Properties

The physical properties of pure aluminum are summarized in Table 2 [15].

Table 2. Physical properties of pure aluminum

Property	Value
mp	933.5 K
bp	2767 K
Heat of fusion	397 J/g
Heat of vaporization	1.08×10^{-4} J/g
Heat capacity	$0.90\,\mathrm{J\,g^{-1}K^{-1}}$
Density (solid)	
theoretical density based on lattice spacing	$2699\,\mathrm{kg/m^3}$
polycrystalline material	$2697-2699\,\mathrm{kg/m^3}$
Density (liquid)	$2357\,\mathrm{kg/m^3}$ at 973 K $2304\,\mathrm{kg/m^3}$ at 1173 K
Thermal neutron cross section	$(2.32\pm0.03)\times10^{-25}\,\mathrm{cm^2}$ (0.232±0.003 barn)
Lattice constant (length of unit cube)	4.0496×10^{-10} m at 298 K
Coefficient of expansion	23×10^{-6}/K at 293 K
Thermal conductivity	$2.37\,\mathrm{W\,cm^{-1}K^{-1}}$ at 298 K
Volume resistivity	$2.655\times10^{-8}\,\Omega\,\mathrm{m}$
Magnetic susceptibility	$16\times10^{-3}\,\mathrm{mm^{-3}\,mol^{-1}}$ at 298 K
Surface tension	8.68×10^{-3} N/cm at the *mp*
Viscosity	$0.0012\,\mathrm{Pa\cdot s}$ at the *mp*

Atomic Structure, Nuclear Properties [15]. Aluminum has an atomic number of 13 and atomic mass of 26.98154 based on ^{12}C. The only abundant isotope, ^{27}Al, is stable and consists of 14 neutrons and 13 protons. Except for a single isotope, ^{26}Al, which has a half-life of 10^6 years, all isotopes have half-lifes of less than 8 s and are of negligible abundance. The naturally occurring isotope has a low cross section for thermal neutrons of $2.3\times10^{-25}\,\mathrm{cm^2}$, increasing in an irregular manner to $6\times10^{-25}\,\mathrm{cm^2}$ at 700 – 800 MeV. The nuclear magnetic moment is $1.84\times10^{-26}\,\mathrm{A\cdot m^2}$.

Crystal Structure. Aluminum crystallizes in a face-centered cubic lattice that is stable from 4 K to the melting point. The coordination number is 12 with four atoms to the unit cell. The edge length of the unit lattice cube for pure aluminum is 4.049596×10^{-10} m at 298 K. Hence, the atomic diameter of aluminum is 2.86×10^{-10} m and its atomic volume $9.999\times10^{-6}\,\mathrm{m^3/mol}$. The lattice parameter is affected only slightly by impurities.

Thermal Expansion. Values for the coefficient of thermal expansion are given in Table 3 [15]. The coefficient of thermal expansion is probably isotropic, meaning it is the same in all directions.

Table 3. Thermal expansion coefficient of pure aluminum

Temperature, K	α ($\times10^{-6}$/K)	Temperature, K	α ($\times10^{-6}$/K)
25	0.5	400	24.9
50	3.5	500	26.5
75	8.1	600	28.2
100	12.0	700	30.4
150	17.1	800	33.5
200	20.2	900	37.3
293	23		

Thermal Conductivity. Above 100 K the thermal conductivity of well-annealed 99.99 % aluminum is relatively insensitive to the impurity level. Below 100 K thermal conductivity becomes highly sensitive to the level of impurities. Values for thermal conductivity are given in Table 4 [15].

Table 4. Thermal conductivity of aluminum

Temperature, K	Thermal conductivity, $\mathrm{W\,cm^{-1}K^{-1}}$	Temperature, K	Thermal conductivity, $\mathrm{W\,cm^{-1}K^{-1}}$
Solid			
0	0	300	2.37
1	41.1	400	2.40
2	81.8	500	2.36
3	121	600	2.31
4	157	700	2.25
5	188	800	2.18
9	239	900	2.10
10	235	933.52	2.08
20	117		
30	49.5	Liquid	
40	24.0	933.52	0.907
50	13.5	1000	0.930
100	3.02	1100	0.964
150	2.48	1200	0.994
200	2.37	1300	1.02
250	2.35	1400	1.05
273.2	2.36	1500	1.07
298.2	2.37		

Electrical Resistivity. The application of aluminum as an electrical conductor depends upon the low electrical resistivity of unalloyed aluminum. The electrical conductivity of pure aluminum at room temperature is 64.94 % of that specified for copper in the International Annealed Copper Standard (IACS). At temperatures below 50 K, the electrical resistivity of aluminum is less than that of copper and silver of very high purity. Aluminum becomes superconducting below 1.2 K.

The resistivity of aluminum below 100 K is highly sensitive to purity. The *residual resistivity ratio* (RRR), or the ratio of electrical resistivity at room temperature to that at 4.2 K (boiling point of helium), is sometimes used as a measure of purity. Resistivity ratios of more than 3×10^4 have been reported for 99.999 % aluminum.

Values for the electrical resistivity of aluminum are given in tabular form (Table 5) and schematically (Fig. 1).

Table 5. Electrical resistivity of pure aluminum

Symbol	Temperature, K	Resistivity, $\mu\Omega\cdot$cm	Sample
△	1.65	9.25×10^{-5}	99.9998 % aluminum, single crystal, large diameter (≈10 mm)
	4.22	9.25×10^{-5}	
	14	2.61×10^{-4}	
	20.4	7.33×10^{-4}	
	58	8.70×10^{-2}	
	63.5	0.117	
	77.4	0.210	
	90.31	0.351	
	111.6	0.614	
•	273	2.50	99.9 % aluminum, 0.05 % silicon, 127-µm diameter wire
	373	3.62	
	473	4.78	
	573	6.00	
	673	7.29	
	773	8.63	
	873	10.10	
	923	10.90	
○	933	10.95 (s)	99.99 % aluminum
	933	24.2 (l)	
	1173	27.75	
	1273	29.2	
	1473	32.15	
- - -	4.2	5.7×10^{-3}	99.999 % aluminum annealed at 150 °C (300 °F) for 4 h
	77	0.22	
	4.2	2.3×10^{-3}	
	77	0.22	99.9999 % aluminum annealed at 150 °C (300 °F) for 4 h

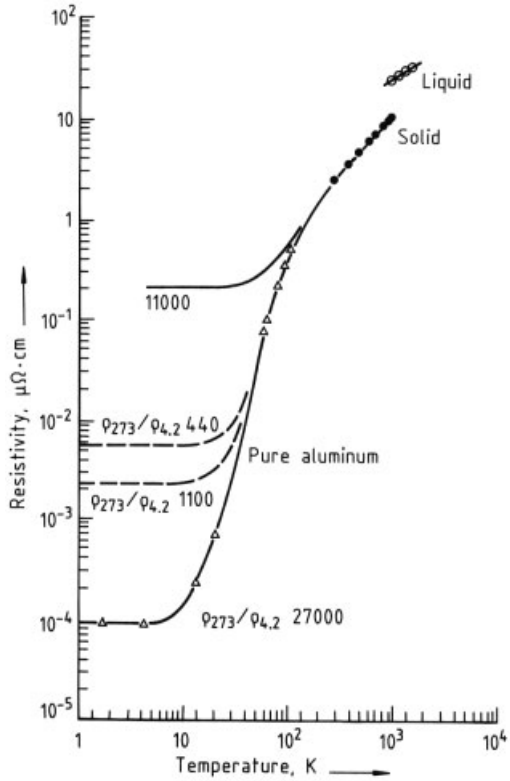

Figure 1. Electrical resistivity of aluminum [15]

3. Chemical Properties

The 13 electrons in the aluminum atom are distributed: $1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^1$. With few exceptions the valence of aluminum in chemical compounds is 3 +. At elevated temperatures aluminum is monovalent in gaseous molecules, such as AlCl, AlF, and Al_2O. Such species disproportionate at lower temperatures into the normal trivalent condensed compounds and elemental aluminum.

For reactions of aluminum not covered in this chapter, see → Aluminum Compounds, Inorganic and → Aluminum Compounds, Organic.

3.1. Oxidation of Aluminum

Although aluminum is one of the most reactive of the common commercial metals, it is remarkably stable in many oxidizing environments. It

owes its stability to the continuous film of aluminum oxide that rapidly grows on a nascent aluminum surface exposed to oxygen, water, or other oxidants. The molecular volume of the oxide is about 1.3 times greater than that of the aluminum consumed in the oxidation reaction. The surface layer, therefore, is under compressive stress and rapidly heals when damaged. In dry oxygen, the surface layer attains a limiting thickness that is a function of temperature. At room temperature this thickness is about 2.5 – 3.0 nm.

Film thickness is also a function of the amount of water vapor present. At room temperature and 100 % relative humidity, about twice as much oxide is formed as in dry oxygen. In both cases, however, the same rate laws apply. Two film layers generally form in wet environments; the continuous oxide layer closest to the metal surface changes to a hydroxylated film at the solid-gas interface. At higher temperatures and on aluminum alloys, especially those containing magnesium and copper, more complex film structures develop, and oxide growth no longer can be described by simple time laws [16].

3.2. Reactions with Aqueous Solutions

Large pieces of aluminum of purity greater than 99.95 % resist attack by most acids but dissolve in aqua regia. Therefore, aluminum is used to store nitric acid, concentrated sulfuric acid, organic acids, and many other reagents. However, its oxide film dissolves in alkaline solutions; corrosion is rapid, producing soluble alkali-metal aluminate and hydrogen:

$$2\,Al + 2\,OH^- + 6\,H_2O \rightarrow 2\,[Al(OH)_4]^- + 3\,H_2(g)$$

Aluminum is amphoteric and can react with mineral acids to form soluble salts and evolve hydrogen:

$$2\,Al + 6\,H_3O^+ + 6\,H_2O \rightarrow 2\,[Al(H_2O)_6]^{3+} + 3\,H_2(g)$$

Molten aluminum can react explosively with water [17], [18]. The mechanism of the reaction is not well understood. Molten aluminum should not be allowed in contact with moist tools or containers.

3.3. Reactions at High Temperatures

Molten aluminum reduces many compounds containing oxygen (→ Aluminothermic Processes). These reactions are used in the manufacture of certain metals and alloys by the thermite reaction, which is the reaction of a metal oxide with aluminum to produce aluminum oxide (Al_2O_3) and the free metal.

Aluminum reduces silicates, particularly glasses. The reaction can start well below the melting point of aluminum.

3.4. Corrosion

Atmospheric corrosion is generally electrochemical in nature, depending on current flow between anodic and cathodic areas. Attack is often nonuniform and limited to specific areas of the surface. The corrosion resistance of some alloy sheet can be improved by cladding the sheet with a thin layer of unalloyed aluminum; however, the aluminum must be anodic to the base metal. Aluminum performs poorly in direct contact with more noble metals in the presence of an electrolyte; galvanic attack of aluminum may take place near the contact area.

3.5. Gases and Aluminum

Hydrogen [*1333-74-0*] is appreciably soluble in both solid and molten aluminum. Other gases reported to be present occur when nonmetallics in the metal react with the environment. Molten aluminum reacts readily with carbon monoxide [*630-08-0*] and carbon dioxide [*124-38-9*]; it also reacts with water vapor in the atmosphere, adsorbed water, water present in hydrated oxide films on scrap, and water adsorbed on or combined in refractories. Solid aluminum reacts with moisture in the atmosphere of a furnace to form oxides and hydrogen. Hydrogen is considerably less soluble in solid than in molten aluminum. The gas diffuses out of the metal at a rate determined by the temperature. However, when the reaction of aluminum with water vapor leads to a high activity of hydrogen at the solid-gas interface, the concentration gradient can be reversed. As a result, the rate of outgassing is decreased

and a higher level of hydrogen is retained in the metal. In solid aluminum, hydrogen in excess of the solution limit can precipitate as H_2 at grain boundaries, thus lowering the hydrogen ion concentration and in turn further facilitating the diffusion of hydrogen ions from the interface into the metal.

4. Production

The only method now used industrially to produce primary aluminum is the Hall–Héroult process. Production of aluminum before its development was discussed previously in this article. Alternate means of producing aluminum are treated in Section 4.5.

4.1. History of the Electrolytic Reduction of Alumina

The technological elements of the process – electrolysis of fused salts to produce metals (including aluminum), the use of cryolite as a flux to dissolve alumina, and the use of carbon electrodes – had been exploited for some time prior to 1886. The workable electrolytic process was discovered independently, and almost simultaneously, in early 1886 by CHARLES MARTIN HALL in Oberlin, Ohio, and PAUL L. T. HÉROULT in Gentilly, France. Both these young scientists were familiar with the work of SAINTE-CLAIRE DEVILLE.

In less than three years, the invention had been implemented industrially in North America and in Europe. In November 1888 aluminum was first produced commercially by the electrolytic reduction of alumina by HALL and others in a company that later was to become the Aluminum Company of America. At about the same time, HÉROULT was associated with a company (later to be known as Alusuisse) that operated aluminum electrowinning cells at Neuhausen, Switzerland.

KARL J. BAYER, an Austrian chemist, was issued a patent, DE 43977, in July 1887 for an improved method of producing alumina from bauxite. Bauxite [*1318-16-7*], discovered by P. BERTHIER in 1821, is named for Les Baux, the village in the south of France near which it was first found. With the development of a process to produce pure aluminum oxide from this abundant ore, the technology was then complete to spur rapid growth of the aluminum industry in Europe and North America in the last decade of the 19th century. The increase in the production of primary aluminum during the past hundred years is shown in Figure 2 [19].

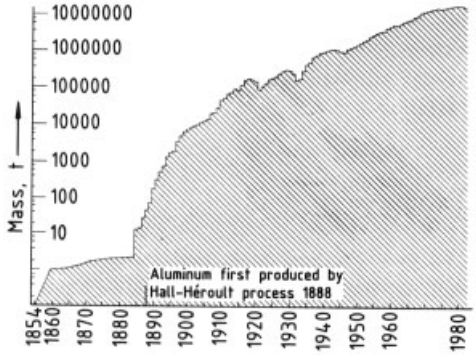

Figure 2. Annual world production of primary aluminum 1854 – 1982 [7]

4.2. Raw Materials

Carbon. In the industrial electrowinning (separation by electrolysis) of aluminum, part of the energy for reducing alumina is supplied as electricity and part comes from consumption of the carbon *anode.* Carbon is also used as the cathode lining. Because 0.4 – 0.5 kg of anode is consumed for each kilogram of aluminum produced, this represents the major carbon requirement. Because the ash from the carbon will contaminate either the aluminum produced or the electrolyte, high-purity carbon is desirable. Certain impurities, such as vanadium, are particularly harmful in that they catalyze air burning of the carbon. Other impurities, such as phosphorus, accumulate in the electrolyte and undergo cyclic redox reactions (partial reduction followed by reoxidation), consuming electric current without producing product. The coke residue from petroleum refining is quite pure and, therefore, has been the major source of carbon for anodes. The structure of petroleum coke (→ Petroleum Coke) varies depending on the nature of the petroleum feedstocks used at the refinery, the refinery flowstream, and the coking conditions used. This coke produced at about

500 °C requires calcining at about 1200 °C to remove volatile constituents and increase its density before it is blended into the anode mix. After calcination, the coke is ground and mixed with crushed spent anodes and sufficient coal-tar pitch to allow molding into anode blocks by pressing or by vibrating. They are baked at 1000 – 1200 °C, causing the pitch to carbonize, forming strong carbon blocks. These blocks are made with one or more sockets into each of which is fastened a steel stub by pouring cast iron around it. These stubs both conduct electric current into the anode and support the anodes in the cell. The cost of prebaked carbon anodes in the United States was about $ 0.40/kg ($ 0.15/pound) in 1983.

Anthracite has been the major constituent in the cell *cathode* blocks, although graphite and metallurgical coke have been used to some extent. The anthracite is calcined at 1200 °C or higher, crushed and sized, mixed with coal-tar pitch, molded into blocks, and baked. These blocks, mortared together with a carbonaceous seam mix, form the pot lining, which is the container for both the aluminum and the electrolyte. High purity is not as important for the cathode blocks because leaching of impurities is very slow. Consumption of cathode carbon amounts to 0.02 – 0.04 kg of carbon per kilogram of aluminum produced. The life of a pot (typically 2 – 6 years) generally is terminated by failure of the carbon pot lining.

Aluminum Oxide. Depending on its purity and losses in handling, 1.90 – 1.95 kg of alumina are consumed in producing 1 kg of aluminum. The cost of alumina at United States smelting facilities was $ 0.20 – 0.26/kg ($ 0.09 – 0.12/pound) in 1983. The preparation of metallurgical alumina and its required physical and chemical properties are discussed under → Aluminum Oxide.

Electrolyte Materials. The electrolyte for electrowinning aluminum is basically a solution of aluminum oxide in cryolite [*15096-52-3*]. The presence of cryolite is essential for dissolution of alumina. Cryolite usually comprises more than 75 % of the electrolyte which typically also contains calcium fluoride (4 – 8 %), excess aluminum fluoride (5 – 15 %), alumina (1 – 6 %), and sometimes lithium fluoride [*7789-24-4*] (0 – 5 %) and magnesium fluoride [*7783-40-6*] (0 – 5 %). These additives lower operating temperature and increase current efficiency.

The mineral cryolite is the double fluoride of sodium and aluminum and has a stoichiometry very near the formula Na_3AlF_6 and a melting point of about 1010 °C. It has been found in substantial quantities only in Greenland, and was mined extensively there in the early 20th century but now is essentially exhausted. Synthetic cryolite can be produced by reacting hydrofluoric acid with an alkaline sodium aluminate solution:

$$6\,HF + 2\,NaOH + NaAlO_2 \rightarrow Na_3AlF_6 + 4\,H_2O$$

Cryolite also can be recovered from used pot linings. The lining is crushed and treated with dilute sodium hydroxide solution to dissolve fluorides. After being filtered, the solution is neutralized with carbon dioxide to precipitate the cryolite.

Cryolite is produced directly in reduction cells by reaction of the soda impurity in the feed alumina with added aluminum fluoride (→ Fluorine Compounds, Inorganic):

$$3\,Na_2O + 4\,AlF_3 \rightarrow 2\,Na_3AlF_6 + Al_2O_3 \qquad (1)$$

Electrolyte generated by the above reaction must be tapped from the cells periodically. In modern smelters with dry scrubbing equipment for fume treatment and cell lives greater than 3 years, cryolite is a byproduct rather than a raw material in producing aluminum.

Synthetic cryolite could be purchased in the United States for $ 500 – 600/t in 1983.

Aluminum Fluoride. Aluminum fluoride, AlF_3 [*7784-18-1*], may comprise as much as 15 wt% of the electrolyte in excess of the amount represented by the cryolite composition. Aluminum fluoride is consumed during normal operation by three major mechanisms. First, losses of aluminum fluoride by vaporization are appreciable; the most volatile species present in the electrolyte is sodium tetrafluoroaluminate [*13821-15-3*], $NaAlF_4$, having a partial pressure of 200 – 600 Pa over the operating melt, depending on composition and temperature (→ Fluorine Compounds, Inorganic). Second, aluminum fluoride is depleted by hydrolysis:

$$2\,AlF_3 + 3\,H_2O \rightarrow Al_2O_3 + 6\,HF$$

And finally, aluminum fluoride is consumed by reaction with the soda present in feed alumina (Eq. 1).

Fume capture and scrubbing efficiencies have improved aluminum smelters. Fluoride previously lost by vaporization of $NaAlF_4$ and hydrolysis of bath is now almost completely recycled to the cells. Nevertheless, aluminum fluoride consumption amounts to 0.02 – 0.04 kg AlF_3 per kilogram of aluminum product. In 1983 technical anhydrous aluminum fluoride could be purchased in the United States for about $ 350/t.

4.3. Hall–Héroult Cell for Aluminum Production

All commercial production of aluminum today is done in Hall–Héroult cells. Employing prebaked carbon anodes or self-baking Söderberg anodes. The Hall–Héroult cell with *prebaked anodes* is shown in Figure 3. Essentially pure alumina is fed into the previously discussed cryolite base electrolyte. Electric current deposits aluminum into a pool of molten aluminum held under the electrolyte in the carbon lined cavity of the cell. Oxygen from the alumina deposits electrolytically onto the carbon anode dipping into the electrolyte and reacts with (burns) the anode. Cells typically range from 9 to 12 m long, 3 to 4 m wide, and 1 to 1.2 m high. Thermal insulation surrounds the carbon lining of the cell to control heat losses. Although carbon is the material known to withstand best the combined corrosive action of molten fluorides and molten aluminum, even carbon would have a very limited life in contact with the electrolyte at the sides of the cell were it not protected by a layer of frozen electrolyte. The thermal insulation is adjusted carefully to maintain a protective coating on the walls but not on the bottom, which must remain substantially bare for electrical contact. Steel collector bars in the carbon cathode conduct electric current from the cell. These bars are inserted into holes that have been sized carefully so that thermal expansion forms a tight electrical contact, or cemented in place with a carbonaceous cement containing metal particles, or bonded in place with cast iron.

The electrical resistivity of prebaked anodes ranges from 0.005 to 0.006 Ω cm. Current density at the anode face is 0.6 – 1.3 A/cm^2.

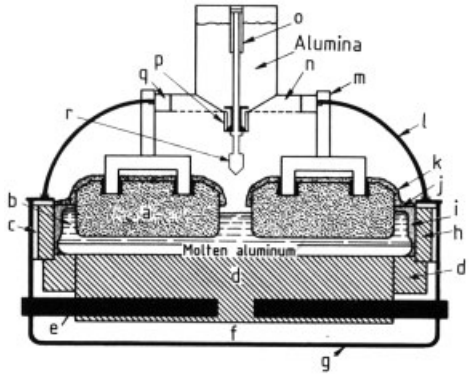

Figure 3. Hall–Héroult cell with prebaked anodes
a) Carbon anode; b) Electrolyte; c) Insulation; d) Carbon lining; e) Current collector bar; f) Thermal insulation; g) Steel shell; h) Carbon block; i) Ledge; j) Crust; k) Alumina cover; l) Removable covers; m) Anode rods; n) Fume collection; o) Air cylinder; p) Feeder; q) Current supply; r) Crust breaker

The *Söderberg anode* (Fig. 4) uses a premixed "paste" of petroleum coke and coal-tar pitch. This mixture is added at the top of a rectangular steel casing that is typically 6 – 8 m long, 2 m wide, and 1 m high. Heat from the electrolyte and heat from the electric current passing through the anode bakes the carbonaceous mix as it progresses through the casing.

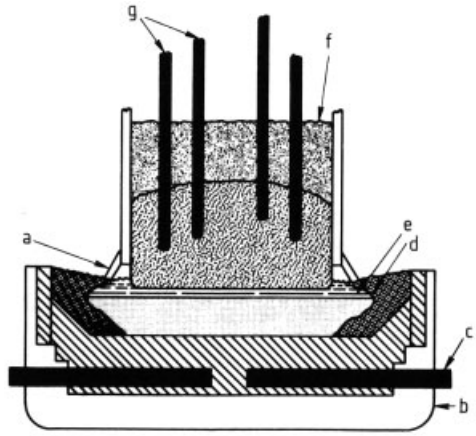

Figure 4. Aluminum electrolyzing cell with Söderberg anode
a) Manifold gas; b) Steel shell; c) Current collector bars; d) Frozen ledge; e) Molten electrolyte; f) Coke and tar paste; g) Current supplying pins

The baked portion extends past the casing and into the molten electrolyte. Baked mix replaces anode being consumed at the bottom surface. Electric current enters the anode through either

vertical or sloping steel spikes. These spikes are pulled and reset to a higher level as they approach the lower surface. Söderberg anodes have an electrical resistivity about 30 % higher than that of prebaked anodes. Because of the resulting lower power efficiency and the greater difficulty in collecting and disposing of baking fumes, Söderberg anodes are being replaced by prebaked anodes, even though the former save the capital, labor, and energy required to manufacture the latter.

4.3.1. Electrolyte

Pure cryolite melts at about 1010 °C. Alumina and other additives lower the melting point, allowing operation at 940 – 980 °C. The cryolite-aluminum fluoride-alumina system (Fig. 5) [20] has binary eutectic points at 961 °C and 694 °C and a ternary one at 684 °C. Calcium fluoride and lithium fluoride further reduce the liquidus temperature (Figs. 6 [21] and 7 [22]). *Calcium fluoride* is seldom added intentionally. Because of a small amount of calcium oxide impurity in the alumina, it attains a steady-state concentration of 3 – 8 % in the melt. At this level calcium is codeposited into the aluminum and emitted in the off-gas at a rate equal to its introduction. Magnesium fluoride accumulates to 0.1 – 0.3 %, in the electrolyte by the same mechanism as calcium fluoride. Some operators add up to 5 % MgF_2 because it expels carbon dust from the electrolyte by decreasing the electrolyte's ability to wet carbon.

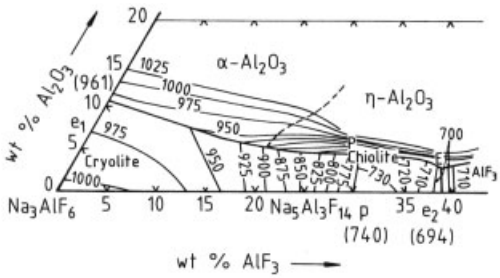

Figure 5. The Na_3AlF_6–AlF_3–Al_2O_3 system [20]

Calcium fluoride, in addition to lowering the liquidus temperature, decreases the vapor pressure and solubility of reduced species in the electrolyte for better current efficiency. Detrimentally, it lowers alumina solubility and electrical conductivity and increases density, viscosity, and surface tension of the electrolyte. *Lithium fluoride,* in addition to lowering the melting point, decreases the vapor pressure, density, reduced species solubility, and viscosity; it also increases electrical conductivity. The only negative effect appears to be lowered alumina solubility. Its high cost, however, requires that its benefits be weighed against the price. *Aluminum fluoride* decreases solubility of reduced species and lowers surface tension, viscosity, and density. It has the undesirable effects of decreasing alumina solubility and electrical conductivity and increasing vapor pressure. Aluminum fluoride

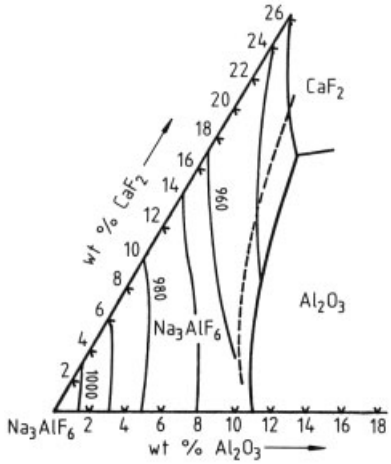

Figure 6. The Na_3AlF_6–Al_2O_3–CaF_2 system [21]

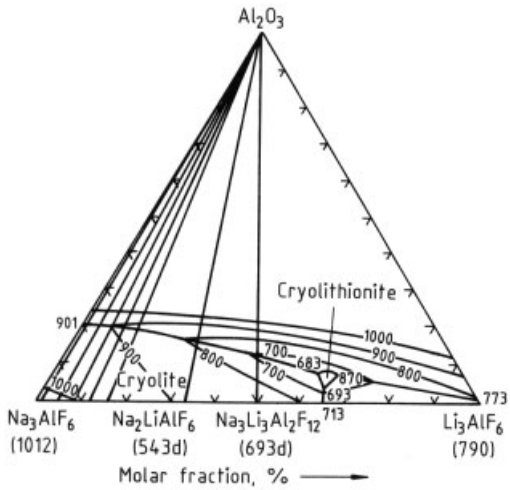

Figure 7. The Na_3AlF_6–Li_3AlF_6–Al_2O_3 system [22]

acts as a Lewis acid with sodium fluoride acting as a Lewis base. Neutrality has been defined arbitrarily as a molar ratio of sodium fluoride to aluminum fluoride of 3 : 1. Control of electrolyte acidity or the NaF : AlF_3 molar ratio, referred to as the *cryolite ratio* (R_c), is of importance to cell operation. Lithium fluoride is a slightly weaker Lewis base than sodium fluoride. Magnesium fluoride and calcium fluoride are weak Lewis acids.

Ionic Structure of the Melt. There is general agreement that molten cryolite is completely ionized to sodium ions and hexafluoroaluminate ions:

$$Na_3AlF_6 \rightarrow 3\,Na^+ + [AlF_6]^{3-}$$

Also it is well established that the hexafluoroaluminate ion dissociates further:

$$[AlF_6]^{3-} \rightarrow [AlF_4]^- + 2\,F^-$$

At $R_c = 3$, $[AlF_6]^{3-}$ is about 30 % dissociated [23]. Raman spectroscopy has showed that the dissociation increases with decreasing cryolite ratio to complete dissociation at $R_c = 1$ [24].

The nature of the species formed when alumina is added is not so well established. Combination of the results of cryoscopic measurements, Raman spectrographic data, and equilibrium studies and vapor pressure measurements (reviewed in [23]) leads to the conclusion that $[Al_2O_2F_4]^{2-}$ and $[Al_2OF_6]^{2-}$ are the two major oxygen-containing ions in the melt. Possible reactions for their formation are:

$$2\,Al_2O_3 + 2\,[AlF_6]^{3-} \rightarrow 3\,[Al_2O_2F_4]^{2-}$$

$$Al_2O_3 + 4\,[AlF_6]^{3-} \rightarrow 3\,[Al_2OF_6]^{2-} + 6\,F^-$$

4.3.2. Electrode Reactions

Cathode Reaction. Even though Na^+ is the principal current carrier, it does not discharge at the cathode. The reversible electromotive force for the formation of liquid aluminum is about 0.24 V lower than that for the formation of sodium gas at 101.3 kPa (1 atm) for the range of compositions and temperatures used industrially. Bowman [25], using cyclic voltammetry, stationary electrode polarography, differential pulse polarography, and chronopotentiometry, found evidence only for a reversible three-electron transfer process. There was no evidence for a chemical reaction, either preceding or following the electron-transfer process. This ruled out discharge of sodium at low activity followed by a chemical reaction to form aluminum, and also eliminated dissociation of the $[AlF_6]^{3-}$ or $[AlF_4]^-$ to form Al^{3+} and F^- ions in the double layer preceding charge transfer. Rolin, et al. [26] believed this latter process does take place. If so, dissociation must be too rapid to be detected by Bowman's techniques.

The cathode overvoltage can be represented by [27]:

$$\eta_{cc} = [RT\,(1.375 - 0.125\,R_c)\,/\,(1.5\,F)]\,\ln\,(i/0.25) \quad (2)$$

where R is the gas constant, T is the temperature (K), R_c is the mole ratio NaF : AlF_3, F is the Faraday constant, and i is the electrode's current density (in A/cm^2). Although this relationship mathematically looks like activation overvoltage, actually the cathode overvoltage is caused by an increase in the NaF : AlF_3 ratio at the aluminum surface [28]. Sodium ions carry the current while complex aluminum anions discharge. This requires a diffusional flux of $[AlF_6]^{3-}$ and $[AlF_4]^-$ ions to the cathode interface with a similar diffusional flux of Na^+ and F^- ions away from the interface. Using Fick's first law, the resulting concentration gradients were calculated and good agreement was found between the measured overvoltages and the electromotive forces between the two aluminum half-cells, one containing electrolyte of the bulk composition and the other, electrolyte corresponding to the calculated interfacial composition.

Anode Reactions. The primary anode reaction can be written:

$$C + 2\,O^{2-} \rightarrow CO_2 + 4\,e^-$$

However, O^{2-} ion is not present in the bulk electrolyte; instead, oxygen is present as structurally large complexes, i.e., $[Al_2OF_6]^{2-}$ and $[Al_2O_2F_4]^{2-}$. Thermodynamically, oxygen depositing onto carbon at 950 – 1000 °C should equilibrate to about 99 % CO and 1 % CO_2.

However, based on either net carbon consumption or the volume of gas produced, the primary anode product is essentially all CO_2. The high anodic overvoltage implies that reaction kinetics cause this surprising displacement from thermodynamic equilibrium. Rotating disk [29] and impedance [30] measurements indicate that there is a small diffusional overvoltage, probably caused by reaction within the pores of the electrode. Using the general treatment for heterogeneous reaction control [31], overvoltage data can be expressed by the relationship:

$$\eta_{SA} = (vRT/pnF)\ \ln\left(i/i^0\right) \tag{3}$$

where v is the number of executions of rate-controlling steps to produce one overall step, p is the reaction order, n is the number of electrons transferred in one overall step, and i^0 is the reaction limiting current density. The reaction order, $p = 0.57$, was found from measurements of overvoltage versus current density [32]. In industrial practice, the reaction order ranges from 0.4 to 0.6, varying with carbon reactivity and porosity. The reaction-limiting current density, i^0, goes from 0.0039 to 0.0085 A/cm^2 as alumina concentration varies from 2 to 8 wt %. The reaction order also has been determined [32] from the rate of change of the limiting current with reactive species concentration and a similar value obtained:

$$p = \frac{\mathrm{d}\ln i^0}{\mathrm{d}\ln c_{Al_2O_3}} = 0.56 \tag{4}$$

Measurements [33] of the ordinary combustion of graphite have shown that when oxygen reacts both in pores and on the surface, a chemical reaction of approximately half-order results. Linear sweep voltammograms showed voltage peaks with increasing current density corresponding to discharge of CO_2, COF_2, and CF_4 [34].

The following anode reaction mechanism is consistent with these observations. Oxyfluoride ions dissociate within the double layer to oxygen ions:

$$[Al_2O_2F_4]^{2-} \rightarrow O^{2-} + Al_2OF_4$$

$$Al_2OF_4 + [Al_2OF_6]^{2-} \rightleftharpoons [Al_2O_2F_4]^{2-} + 2\,AlF_3$$

Oxygen ions discharge upon the carbon surface (surf) forming an activated complex, CO:

$$C\,(surf) + O^{2-} \rightarrow C{*}O^{-}\,(surf) + e^{-}$$

$$C{*}O^{-}\,(surf) \rightarrow C{*}O\,(surf) + e^{-}$$

The activated complex converts very slowly to CO (g) through an adsorbed (ads) intermediate:

$$C{*}O(surf) \rightarrow CO\,(ads)\ \text{slow}$$

$$CO(ads) \rightarrow CO\,(g)\ \text{slow}$$

Carbon burning in pure oxygen at 940 – 970 °C has a combustion rate equivalent to between 0.1 and 0.2 mA/cm^2. One would expect this reaction to proceed at a similarly slow rate. As available surface sites become covered with C∗O, oxygen is deposed at higher energy (overvoltage) onto a carbon site already bonded to oxygen, producing unstable $[C{*}O_2]^{-}$, which breaks carbon to carbon bonds almost immediately as it is formed:

$$C{*}O\,(surf) + O^{2-} \rightarrow [C{*}O_2]^{-}\,(surf) + e^{-}$$

$$[C{*}O_2]^{-}\,(surf) \rightarrow CO_2\,(ads) + e^{-}$$

$$CO_2\,(ads) \rightarrow CO_2\,(g)$$

The adsorbed CO_2 quickly desorbs as CO_2 (g). This mechanism explains both the high anodic overvoltage and the primary production of CO_2 instead of the thermodynamically favored CO.

When there is insufficient alumina in the electrolyte, the cell experiences a phenomenon called *anode effect.* Bubbles grow larger and larger on the anode until the electrolyte no longer wets the anode. With a constant potential the current falls to a low value; but with the constant current source used industrially, the cell potential rises to $\geqq 30$ V. Current then penetrates the gas film by a multitude of small electric arcs or sparks. The gas produced at the anode changes from CO_2 to CO, with significant quantities (3 – 25 %) of CF_4 and minor amounts of C_2F_6 generated. Fluorocarbon compounds most likely deposit on the surface of the anode and trigger the anode effect. As alumina is depleted, anode overvoltage increases. At about 1.2 V anode overpotential, sufficient thermodynamic activity of fluorine is produced to cause fluorine to bond to the carbon. Even though these low-surface-energy carbon-fluorine compounds decompose on the surface to CF_4 and C_2F_6 at cell temperature, their rate of formation can exceed the rate of thermal decomposition, producing high coverage. Once the cell is on anode effect,

restoring the alumina concentration is not sufficient to return it to normal operation. The gas film must be broken by splashing aluminum, by interrupting the current momentarily, or by lowering the anodes to expose new areas not contaminated with fluorine.

4.3.3. Current Efficiency

According to Faraday's law, 1 kA h of electric current should produce 0.3356 kg of aluminum, but only 85 – 95 % of this amount is obtained. The principal loss mechanism is recombination of anodic and cathodic products. Reduced species go into solution in the electrolyte at the aluminum-electrolyte interface (Fig. 8). Disagreement exists over whether the dissolved species is metallic sodium, monovalent aluminum, subvalent sodium, colloidal aluminum, or a combination of these species. Equation (4) produces a thermodynamic activity of sodium in the melt, whereas Equation (5) produces an activity of aluminum monofluoride.

$$\mathrm{Al(l)} + 3\,\mathrm{NaF(soln)} \rightleftharpoons 3\,\mathrm{Na(soln)} + \mathrm{AlF_3(soln)} \tag{4}$$

$$2\,\mathrm{Al(l)} + \mathrm{AlF_3(soln)} \rightleftharpoons 3\,\mathrm{AlF(soln)} \tag{5}$$

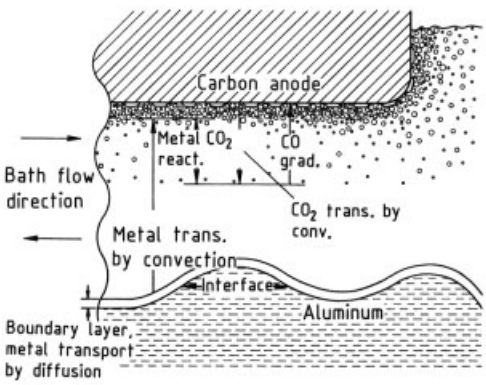

Figure 8. Current efficiency loss by metal reoxidation

Metal going into solution must first diffuse through the metal-electrolyte boundary layer (Fig. 8). The metal is then transported by convection to the vicinity of the anode. Here it reacts with carbon dioxide. Chemical reactions appear to be fast compared to mass transport. The rate-controlling step was previously assumed to be diffusion of dissolved metal through the boundary layer at the metal-electrolyte interface. However, ref. [35] indicates mixed control with diffusion at both the aluminum-electrolyte interface and the bubble-electrolyte interface being important.

There are several mechanisms accounting for additional minor losses in current efficiency. New cell linings absorb sodium, with Equation (4) maintaining an equilibrium activity of sodium. Fortunately, the lining saturates early in the cell's life but until this occurs, current efficiency is low. When a metal dissolves in a molten salt it usually imparts electronic conductivity to the melt, thereby lowering current efficiency. Some investigators have found a small electronic conductance for cryolite-base melts but others have not. Such elements as phosphorus and vanadium, which can be reduced partially at the cathode and then reoxidized at the anode, lower efficiency.

4.3.4. Cell Voltage

The voltage of a Hall–Héroult cell is made up of a number of components:

$$E_{\mathrm{cell}} = E^0 + \eta_{\mathrm{CA}} + \eta_{\mathrm{CC}} + \eta_{\mathrm{SA}} + \eta_{\mathrm{SC}} - I\,(R_{\mathrm{A}} + R_{\mathrm{B}} + R_{\mathrm{C}} + R_{\mathrm{X}})$$

where E^0 is the thermodynamic equilibrium voltage described under Thermodynamic Considerations (Section 4.4); η_{CA} is the concentration overpotential at the anode; η_{SA} is the surface overpotential at the anode described by Equation (3); η_{CC} is the concentration overpotential at the cathode described by Equation (2); η_{SC} is the surface overpotential at the cathode, generally negligible; I is the total cell current; R_{A} is the electrical resistance of the anodes or anode; R_{B} is the effective resistance of the bath, allowing for fanning out of current as it flows from anode to cathode and the increased bath resistivity caused by gas bubbles (for details, see [27]); R_{C} is the cathode resistance; and R_{X} is the resistance external to the cell but included in calculating power consumption. The anode concentration overpotential can be calculated by:

$$\eta_{\mathrm{CA}} = (RT/2F)\,\ln\,[i_{\mathrm{c}}/\,(i_{\mathrm{c}} - i)] \tag{5}$$

The quantity i_{c} is the concentration-limited current density and it varies between 1.0±

0.5 A/cm^2 at 1 wt % Al_2O_3 and 15 A/cm^2 at 8 wt % Al_2O_3. The specific power consumption in kilowatt-hours per kilogram of aluminum can be calculated:

$$\text{kW h/kg} = 298.06\, E_{\text{cell}} / \%\, CE$$

where % CE is the percentage current efficiency or Faraday efficiency and E_{cell} is the cell voltage.

4.3.5. Heat Balance

Heat balance considerations are important in designing and operating a cell at the optimum temperature. Calculation is complicated by simultaneous heat flow and heat generation in conductors that carry electric current into and out of the cell. More than 50 % of the heat loss may occur through the anodes and top crust. The sidewalls must be protected from erosion by a ledge of frozen electrolyte maintained by extracting the exact amount of heat required for the desired ledge thickness. These calculations have been described [36], [37]. The complex geometry and the interaction between heat flow and electrical flow generally require a computer using either a finite element or finite difference technique for heat balances. Figure 9 [38] shows a typical heat loss distribution for a center-fed cell and Figure 10 [38], the typical temperature distribution.

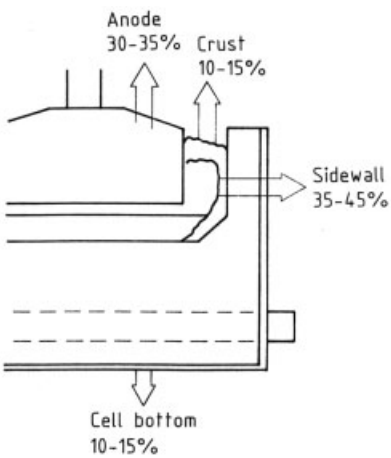

Figure 9. Cell heat loss distribution [38]

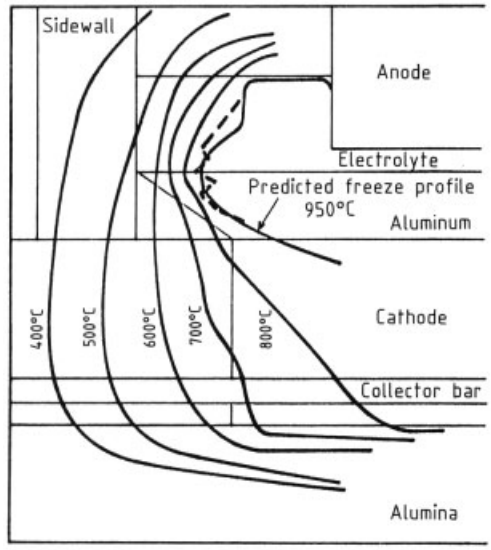

Figure 10. Cell lining isotherms [38]
Dashed line: measured profile

4.3.6. Fluid Dynamics

The large electric currents used in modern cells generate relatively strong magnetic fields within the cell. These magnetic fields interact with electric currents and exert Lorentz forces, producing movement of liquid conductors. Three types of magnetic disturbances have been observed: a vertical or static displacement of the metal pad, a circulating flow within the metal pad with velocities as high as 20 cm/s, and finally a wave motion in the metal pad. The last is probably the most harmful because waves may lead to electrical short circuiting. Calculation of the magnetic fields and electric current flow patterns is complicated. Computer programs have been designed to make these calculations and describe the fluid-dynamic consequences. These calculations have been refined to the point where cells of over 2.5×10^5 A have been designed and operated with high efficiency. Natural convection caused by temperature gradients or composition differences in the electrolyte are insignificant compared with the movement induced by magnetic forces and gas bubbles. Gas bubbles produce significant stirring in the electrolyte and are the dominant force in bath movement, whereas electromagnetic forces predominate in metal pad movement.

4.4. Thermodynamic Considerations

The thermochemistry of constituents of the electrolyte and of the electrolyte-aluminum system is treated in [23]. Thermodynamic data used in the following section are from the *JANAF* (Joint Army-Navy-Air Force) *Thermochemical Tables* [39] and the 1978 supplement to the tables [40].

For pure α-alumina, reduction to aluminum can be represented:

$$1/2\ \alpha\text{-}Al_2O_3(s) \rightarrow Al(l) + 3/4\ O_2(g)$$

At 960 °C (1233 K), $\Delta G^0_{1233} = +642.12$ kJ and $\Delta H^0_{1233} = +845.62$ kJ.

In the industrial production of aluminum, a portion of the energy is supplied by the combustion of carbon anodes to carbon dioxide. The overall cell reaction can be represented:

$$1/2\ \alpha\text{-}Al_2O_3(s) + 3/4\ C(s) \rightarrow Al(l) + 3/4\ CO_2(g) \quad (6)$$

At 960 °C, $\Delta H^0_{1233} = +549.3$ kJ.

Additional energy is required to heat alumina and carbon from room temperature to operating temperature. The alumina fed to the cell usually is not pure α-alumina. For γ-alumina, the overall process can be represented:

$$1/2\ \gamma\text{-}Al_2O_3(s, 298) + 3/4\ C(s, 298) \rightarrow Al(l, 1233) + 3/4\ CO_2(g, 1233)$$

The enthalpy requirement for transformation of the reactants at 25 °C (298 K) to products at 960 °C (1233 K) is:

$$\Delta H_{r(35)} = \Delta H^0_{r1233} + 1/2\,(H^0_{1233} - H^0_{298})\ \gamma\text{-}Al_2O_3 + 3/4\,(H^0_{1233} - H^0_{298})\ C(s) = 606.4\ \text{kJ}$$

This value corresponds to a theoretical energy requirement of 6.25 kW h per kilogram of aluminum produced.

The reversible cell potential (decomposition voltage) for Reaction (6) can be calculated from the relation:

$$\Delta G^0_{1233} = -nFE^0_{\text{sat},\,1233} \quad (7)$$

where ΔG^0_{1233} is the Gibbs free energy change for Equation (6), 3.450×10^5 J; n is the number of electrons per unit cell reaction, i.e., 3; F is the Faraday constant, 96.487 kJ V^{-1} mol^{-1}; and $E^0_{\text{sat},1233}$ is the standard electrode potential, in volts, at 960 °C for the reaction with all reactants and products at unit activity. Substituting the appropriate values into Equation (7) and solving for the cell potential:

$$E^0_{\text{sat},\,1233} = -1.192V$$

The above decomposition potential applies to electrolyte saturated with alumina at 950 °C. The decomposition potential from melts with alumina activity less than unity can be calculated [27] from the equation:

$$E^0 = -E^0_{\text{sat}} + (RT/6F)\ \ln a_{Al_2O_3}$$

The activity of alumina in cryolite-base melts can be approximated [27] from the equation:

$$a_{Al_2O_3} = \left(c_{Al_2O_3}/c^*_{Al_2O_3}\right)^{2.77}$$

where $c_{Al_2O_3}$ is the concentration of alumina in the electrolyte, in wt %, and $c*_{Al_2O_3}$ is the saturation concentration of alumina, in wt %.

4.5. Alternate Processes

Although the Hall–Héroult process has gained industrial dominance, it has several inherent disadvantages. The most serious are the large capital investment required and the high consumption of costly electrical power. There are also the costs of the Bayer alumina refining plant and of the carbon anode plant. Many of the aluminum-producing countries must import alumina or bauxite. The supply of petroleum coke is limited. These deficiencies have spurred research to find alternate processes.

The earliest commercial process for producing aluminum was sodiothermic reduction of aluminum halides. Reduction of aluminum chloride by manganese has been investigated [41]. However, neither can compete with the Hall–Héroult process. Many attempts have been made at direct carbothermic reduction of alumina but these have resulted in very low yields, owing to the formation of solid aluminum carbide, aluminum suboxide vapor, and aluminum vapor that reacts with carbon monoxide as the temperature drops on the way from the furnace. Yields as high as 67 % can be obtained by staging the reactions to produce aluminum carbide at

1930 – 2030 °C, which then reacts with alumina to produce aluminum and carbon monoxide at 2030 – 2130 °C [42].

Better yields result from adding to the furnace a metal (or a metal oxide that is subsequently reduced to a metal), such as iron, silicon, or copper, to alloy with the aluminum and lower its vapor pressure. Of course it is then necessary to extract the aluminum from the alloy. In principle this can be accomplished by electrolytic refining, by fractional crystallization, or by monohalide distillation. In the latter process the aluminum extraction takes place at 1000 – 1400 °C:

$$AlCl_3 + 2\,Al(alloy) \rightarrow 3\,AlCl(g)$$

The AlCl gas is transported to a cooler zone, 600 – 800 °C, where pure aluminum is formed:

$$3\,AlCl(g) \rightarrow 2\,Al(pure) + AlCl_3$$

These combined processes to date have proved noncompetitive. Selective solution of aluminum from the alloy by using a volatile metal, such as mercury, lead, bismuth, cadmium, magnesium, or zinc, has been investigated. Following extraction, the volatile metal is distilled, leaving pure aluminum. If $FeAl_3$, $TiAl_3$, or Al_4C_3 is formed, neither electrolysis nor volatile metal extraction will remove the aluminum from the compound.

The aluminum vapor pressure can be lowered also by alloying the aluminum with aluminum carbide. Over 40 wt % aluminum carbide is soluble in aluminum at 2200 °C [43]. This observation led to a process in which an alloy of aluminum and aluminum carbide was produced at 2400 °C. When this alloy was tapped from the furnace and allowed to cool slowly, the aluminum carbide crystallized into an open lattice, the interstices of which were filled with pure aluminum. Pure aluminum was then removed either by leaching with molten chloride fluxes or by vacuum distillation. The aluminum carbide residue was recycled to the arc furnace. Alternately, the aluminum carbide could be distilled destructively above 2200 °C to produce aluminum and a residue of pure graphite.

In a joint project with the U.S. Department of Energy concluded in 1983, Alcoa investigated producing aluminum-silicon alloy carbothermally in a blast furnace. The required high temperature was obtained by using a pure oxygen blast. Using a low-pressure blast furnace to produce an aluminum-silicon alloy was found to be infeasible because of severe bridging and interruption of the burden movement caused by total reflux of Al, Al_2O, and SiO vapors [44], [45]. Addition of iron improved operation but required an improved technique for extracting the aluminum from the dilute aluminum alloy. An alloy having higher aluminum content could be obtained if a significant portion of the process energy was supplied by electrical power [46]. Japanese researchers are continuing to explore the aluminum blast furnace concept aiming for a low-silicon, high-iron alloy [47]. This improves the blast furnace efficiency but complicates extraction of aluminum from the alloy. They propose to overcome this difficulty by extracting the aluminum with lead [48].

In 1976 Alcoa described a smelting process wherein aluminum chloride, dissolved in a molten sodium chloride-lithium chloride electrolyte, was electrolyzed in a bipolar electrode cell to produce aluminum and chlorine. The chlorine was recycled to a fluid-bed chemical reactor, where it reacted with alumina, pyrolytically coated with carbon from fuel oil, to produce aluminum chloride, carbon dioxide, and carbon monoxide. This reaction was highly exothermic. The aluminum chloride was desublimed to separate it from the gas and recycled to the cell to produce more aluminum and chlorine. This process required 30 % less electrical power than their best Hall–Héroult cells. Because the entire process is a closed system, it is also environmentally more attractive than the Hall–Héroult process. Problems with the chemical plant and low demand for aluminum caused the pilot plant to be shut down temporarily in 1982. Research on the plant was continuing in 1984.

5. Producing High-Purity Aluminum

The Hall–Héroult process has produced aluminum with a purity as great as 99.95 %. Other techniques, such as electrolytic refining or fractional crystallization, are required to produce purer metal. Alcoa developed a three-layer electrolytic aluminum-refining cell in the 1920s. Copper was added to the bottom layer of impure aluminum to increase its density above that

of the fused salt electrolyte. The electrolyte contained 25 – 30 % sodium fluoride, 30 – 38 % aluminum fluoride, and 30 – 38 % barium fluoride to increase the electrolyte's density so that pure aluminum would float to the top. Electric current entered the anode alloy through carbon or graphite blocks at the bottom of the cell and left through graphite electrodes dipping into the high-purity aluminum on top. Aluminum was transported preferentially. Metals more difficult to oxidize than aluminum remained in the anode or alloy layer, and metals more difficult to reduce than aluminum remained in the electrolyte. A ledge of frozen electrolyte on the carbon sidewalls prevented electrical short circuiting between the floating pure aluminum layer and the carbon wall. The cell operated at 950 – 1000 °C and produced 99.98 % pure metal.

In 1932 the Pechiney Company in France operated a cell using a fluoride-chloride electrolyte and nonconducting magnesia brick sidewalls. The electrolyte consisted of 17 % sodium fluoride, 23 % aluminum fluoride, and 60 % barium chloride. The cell operated at 750 °C and produced 99.995 % pure aluminum. An all-fluoride electrolyte (18 % sodium fluoride, 48 % aluminum fluoride, 16 % calcium fluoride, 18 % barium fluoride) also can be used in magnesia-lined cells. Modern electrolytic refining cells (Fig. 11) add a segregation sump [49]. The sump, normally operated at least 30 °C cooler than the aluminum alloy anode, serves as a charging port for aluminum and permits impurity removal. As impurities concentrate in the bottom layer, saturation is reached and crystals form in the cooler sump area, where they can be removed. This procedure greatly extends the operating life of the cell.

Fractional crystallization also is used commercially to refine aluminum. As aluminum freezes most impurities preferentially concentrate in either the liquid or the solid. A molten zone is moved along an aluminum bar in zone refining. Impurities that lower the melting point of aluminum concentrate at one end of the bar, whereas impurities that raise the melting point concentrate at the other end. After several passes the ends are removed, leaving a middle portion of high-purity aluminum.

Super-purity aluminum is produced by another fractional crystallization process [50]. In the first step, peritectic elements are removed

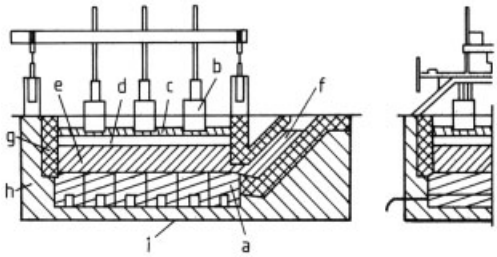

Figure 11. Longitudinal and cross sections of 15000-ampere aluminum refining cell
a) Anode (carbon blocks); b) Cathode (graphite electrodes); c) Refined aluminum; d) Electrolyte; e) Alloy serving as anode; f) Segregation sump; g) Magnesia brick; h) Refractory brick; i) Iron shell

from molten aluminum by reacting them with boron to form borides, which are precipitated and settled before the metal is transferred to a second furnace. Here the surface of molten aluminum is cooled with forced air, causing metal to crystallize. When the furnace is nearly filled with crystals, the remaining impure aluminum is drained. Then the crystals are remelted to yield very pure metal. Aluminum of purity higher than 99.995 % has been produced from 99.9 % pure smelting-grade aluminum by this process.

6. Aluminum Casting: Remelt Ingot

Metal tapped from the Hall–Héroult cell may be treated in one of several ways. First, it may be transported in the molten state to a holding furnace, where, along with solid scrap, it is alloyed, fluxed, filtered, and cast into a form suitable for fabrication by rolling, forging, etc., or for remelting by a foundry. It may be transported further (over distances up to 200 km) on trucks, equipped with well-insulated containers to prevent freezing, to a foundry where it is alloyed and used to make shaped castings. Both these methods of utilizing the molten metal are favored because energy required for remelting is conserved.

Alternatively, where ingot casting facilities or foundries are not nearby, metal is cast using cast iron molds into "primary" ingots for remelting. Generally, these unalloyed remelt ingots come as stackable 22-, 340-, or 680-kg ingots. Casting machines, such as the one illustrated in Figure 12, are often fully automated, with production rates of up to 20000 kg/h. An important con-

sideration is metal transfer from the ladle to the mold, which must be as quiescent as possible in order to minimize oxide contamination. Unalloyed ingot is sold in various purity grades, such as P 0506 or P 1520, depending on the maximum Si and Fe levels; e.g., P 0506 is aluminum with 0.05 % Si and 0.06 % Fe maximum.

Figure 12. Casting machine for the production of open-mold remelt ingot (courtesy Gautschi Electrofours SA)

The larger sizes of unalloyed remelt ingot may be cast on semicontinuous DC (direct chill) casting equipment (→ Aluminum Alloys, Chap. 3.2.) with a "T" cross section for ease of handling and stacking. This ingot usually is cut into lengths weighing approximately 340 kg.

Primary metal producers also make remelt ingot, which is prealloyed for the foundry industry. This is known as foundry ingot and generally is alloyed so that its composition lies within the limits of the alloy desired in the final shaped casting (→ Aluminum Alloys). Foundry ingot may be cast into open molds or continuously cast, cut to length, and stacked (Fig. 13). The most common masses for this ingot, which is remelted in furnaces smaller than those used by wrought alloy producers, are approximately 14 kg and 22 kg. The general practice, where applicable, is to premodify and grain refine this foundry ingot. Also the molten metal is filtered and fluxed in a manner similar to that employed for wrought alloy ingot in order to insure optimum metal quality for the foundry user.

Figure 13. Different types of aluminum castings
Continuously cast "T" section bar (left); Open-mold foundry ingot (right)

7. Environmental Protection

In the primary production of aluminum there are three principal areas of environmental concern: air emissions, wastewater discharges, and solid waste. In the early years of the aluminum industry, most of the effort was directed toward controlling fluoride emission from the cells. The technology for fluoride control is now well established throughout the industry. Other significant air emission sources exist in the anode production process.

Wastewater effluent usually is not a major problem, except in those plants where wet processes are still used to scrub the air emissions from the cells and anode baking furnaces.

Cathode linings in aluminum cells are removed when lining failure occurs and large amounts of this material, called spent pot lining, must be discarded. Historically, spent pot lining has been stockpiled or used as landfill. However, the aluminum industry still needs to find disposal methods that also can provide for the recovery of raw materials.

7.1. Air Emission

Cell Room Emission. Atmospheric emissions produced by an aluminum cell consist of particulate and gaseous fluoride (predominately HF), alumina, carbon dust, sulfur dioxide, carbon dioxide, and carbon monoxide. Typical values for total fluorides generated by the cell related to aluminum production range from 20 to 30 kg/t. Approximately the same amounts of alumina and carbon dusts also are produced. Sulfur dioxide varies according to the sulfur

content of the coke used to prepare the anodes, whereas carbon monoxide emissions generally are less than 30×10^{-6} parts by volume.

Control of Cell Emissions. To control cell emissions, each cell is equipped with a closely fitted hood to collect the evolved gases, which are then ducted to pollution-control equipment. After air dilution the cell gas temperature ranges from 85 to 125 °C, depending on the ambient temperature. Both wet and dry systems are employed to treat cell gas. Some of the older smelters use a combination of electrostatic precipitators for solids removal following by spray towers for removing hydrogen fluoride gas and some sulfur dioxide. Wet precipitators also are quite effective for collecting both gaseous and particulate fluorides; some sulfur dioxide also may be collected by wet precipitators.

Because of inherent disadvantages in wet systems, not the least of which is potential water pollution, new smelters employ *dry scrubbers* for controlling cell emission. A dry scrubber uses either a fluidized bed or direct injection of metal-grade alumina into the gas stream to remove the hydrogen fluoride gas by sorption. The larger the specific surface area of the alumina, the greater is its sorptive capacity for hydrogen fluoride.

The alumina usually is recovered from the gas stream by a fabric filter, which also removes the other entrained particulate species. Dry scrubber efficiencies are generally greater than 98 % for both the gaseous fluoride and the particulate matter. Dry scrubbers are considered to be state of the art control technology. The aluminum fluoride that is recycled to the cells represents a significant material recovery value.

Emissions from Anode Preparation. In the production of anodes, coke dust and pitch fumes are generated at the mixers and at the anode forming equipment. These emissions must be exhausted from the work area for worker protection.

Pitch fumes, coke dusts, and fluorides are the chief pollutants in the waste gases from the anode baking furnaces. In most countries, limits are established for aluminum fluorides (0.05 kg per ton of aluminum) and particulate (dust) emissions, or opacity limits are set in lieu of particulate limitations.

Control Systems for Anode Baking Furnaces. The control systems for the baking furnace emissionsare similar to those described for the cell, with both wet and dry systems being utilized. Fluidized alumina beds and alumina injection systems are used to control gaseous fluorides and hydrocarbons from most new furnaces. A major difference, however, is the need to cool the baking furnace gas in order to condense the hydrocarbons so that they can be extracted by the dusts on the bag filter.

7.2. Wastewater Discharge

Fume Collection Systems. Wastewater generation has been reduced drastically with the development of dry systems for controlling emissions. Some older facilities still rely on wet collection devices, which result in a wastewater stream contaminated with acid, fluoride, and organic material. Conventional treatment involves addition of lime and prolonged retention times (12 h to several days) for neutralization and removal of the calcium fluoride precipitate and a significant portion of the dissolved organic compounds, e.g., benz[*a*]anthracene, benzofluoranthene, benzo[*a*]pyrene, chrysene, and phenanthrene. At least 98 % of these can be removed by biologic treatment in a facultative lagoon that employs aerobic as well as anaerobic bacteria. Partial decomposition of the organic material occurs in the anaerobic bottom layer, whereas the aerobic bacteria present in the upper levels complete the degradation process. Other organic removal mechanisms in the facultative lagoon include skimming, adsorption, and sedimentation.

Sodium hydroxide and alum are used in place of lime for fluoride removal at some plants. This treatment creates a fluoride sludge that, after drying, can be used as a cryolite replacement in the aluminum cell.

Pot Wash Water. After failure of a cathode lining, most of the metal and bath are tapped from the cell and materials remaining within the shell are removed. This "pot digging" operation can be done either "wet" or "dry". In "wet digging," water is applied to the pot to shock thermally and to fracture the bath, metal, and lining material that remain in the pot.

The best approach to handling pot wash water is to minimize the amount of wastewater generated through good management practices during the soaking operation. Excess water is collected in a sump and used as makeup to the next pot brought in for relining. Under no circumstances is this water discharged to a receiving stream because it may contain cyanide (10 – 200 mg/L total CN^-) and fluoride (100 – 600 mg/L).

Plant Runoff. Although aluminum smelting is a dry process, some water is required for services, such as boiler makeup, various equipment cooling needs, and area wash water. A major source of wastewater at a smelting plant is stormwater runoff. Although intermittent, the contaminant loading is significant if efforts are not made to keep courtyards relatively free of dust and to protect spent pot linings from the elements. To minimize the environmental impact of storm water and the other miscellaneous flows, many primary aluminum plants direct their storm sewers through an equalization/holding basin before discharge. This basin is designed to remove suspended solids and floating matter (oil and grease) and also to equalize the fluoride concentration.

7.3. Solid Waste

The material removed from the pot after liner failure is referred to as spent pot lining and represents the largest volume of solid waste associated with the smelting process. Over 3×10^5 t of spent pot linings were estimated to have been generated in 1983 by the worldwide aluminum industry.

The Environmental Protection Agency in the United States initially classified spent pot lining as a hazardous waste because of its cyanide content. Pot lining subsequently was removed from the hazardous waste list pending further studies.

Present disposal/management practices include "temporary" storage in curbed, roofed, protected areas to minimize leachate generation, and burial in landfills lined with impervious materials to prevent contamination of the surrounding environment. These procedures are costly because of the large volumes involved. Because of the carbon and fluoride content of the solid waste, research is directed toward recovering or utilizing these valuable components. Potential applications for spent pot lining include recovery and reuse of the fluoride and carbon constituents, use as a flux material in the steel industry, use as a supplemental fuel by cement manufacturers, and use as fuel in fluidized-bed boilers.

7.4. Regulatory Requirements

Regulatory limits throughout the world have been summarized in an International Primary Aluminium Institute publication [51]. In the United States new carbon bake furnace emissions are limited to 20 % opacity and the total fluoride as a function of aluminum produced is set at 0.05 kg/t. New potroom emissions are limited to 10 % opacity and 0.95 kg/t total fluoride. The effluent limitation guidelines are presently being revised with emphasis on control of toxic pollutants.

8. Economic Aspects

The geographic distribution of primary aluminum production has undergone important changes in the past decade or so. Table 6 shows production figures in 1972 and 1982 for several major countries and the shares of the Western World they represent.

Table 6. Primary aluminum production for selected countries and share of Western World total ($\times 10^3$ t/a)

	1972		1982	
	Prod.	%	Prod.	%
United States	3740	41	3270	31
Canada	920	10	1070	10
Japan	1009	11	350	3
Norway	558	6	640	6
West Germany	444	5	710	7
France	394	4	380	4
Western World	9206	100	10500	100

The decline in production in the United States and Japan is significant. The change that is taking place is largely a result of the rapid rise in energy costs that began with the oil crises of 1973 and 1979 – 1980. The cost of power has become the critical element in the economic viability of primary aluminum facilities. Table 7 shows the estimated percentage of total costs for major cost

components in the production of aluminum for pre- and post-oil embargo periods.

Table 7. Cost breakdown for primary aluminum production for an average smelter in the Western World (%)

	Pre-1973	Post-1980
Alumina	35	30
Power	15	30
Labor	20	15
Miscellaneous	30	25
Total	100	100

Prior to the first oil embargo in 1973, the range of variable costs to produce aluminum was narrow among all smelters. Since then, the range of costs has remained small except for power. The cost of oil- and gas-based electricity escalated to the extent that some smelters were paying 20 times the rate per kilowatt hour of the lowest cost smelters. This is the reason that the Japanese industry has reduced its primary capacity from its peak of 1.6×10^6 t/a to a 1983 level of 0.7×10^6 t/a. Some United States capacity, particularly natural-gas powered smelters in the southwest, has been permanently closed, and further shutdowns are likely.

Largely because of escalating power costs, geographic dispersion of primary capacity is occurring in the Western World. From the end of 1972 to the end of 1982, capacity more than doubled in Africa, more than quadrupled in Latin America, and decreased by 40 % in East Asia. Regional capacities, from data supplied by the International Primary Aluminium Institute for 1972 and estimated for 1982, are given in Table 8.

Table 8. Year-end primary aluminum production capacity ($\times10^3$ t/a)

	1972	1982
Africa	254	615
North America	5430	6135
Latin America	228	1065
Eastern Asia	1323	790
Southern Asia	354	843
Western Europe	3056	4200
Oceania	345	700
Total	10990	14348

Much of the new capacity built in the past decade is nonintegrated; that is, the facility is not tied directly to bauxite or alumina production and no fabricating operations are associated with the smelting facility. This helped promote the trading of aluminum on the London Metal Exchange, and on the aluminum contract on the New York Commodities Exchange.

The creation of a terminal market has accelerated the development of unalloyed aluminum as a commodity product, with the price volatility that commodity status entails.

9. Toxicology and Occupational Health

The bulk of the epidemiological literature confirms that, in humans, aluminum and its compounds exhibit a very low order of toxicity [52–55]. Whether taken into the body by ingestion or inhalation, very little aluminum finds its way into body tissues. The aluminum content of the adult human body is 50 – 150 mg [56], [57]. Because of its ubiquitous nature, aluminum also is found in many foods that make up our normal diets. The amount ingested ranges from as little as 1 to >100 mg/d [58]. The use of aluminum cooking utensils and food packaging does not contribute significantly to the amount that is ingested normally [59], [60]. Regardless of the route of intake, most of the aluminum is eliminated through the gastrointestinal tract.

For many years aluminum, in the form of $Al(OH)_3$, was used widely at high dosages as an antacid, with no apparent ill effects. Although aluminum is known to have neurotoxic properties [61–64], it does not exhibit these properties in normal human beings because it is eliminated readily from the body. On the other hand, for those who have impaired renal function, it can be neurotoxic. There are strong implications that aluminum is a major factor in dialysis encephalopathy, an ultimately fatal condition that sometimes occurs in those with renal failure who undergo dialysis. The major symptoms are speech difficulties, dementia, and seizures. Traces of aluminum in the water used for the preparation of dialysis fluid can accumulate in the patient because the impaired kidneys are unable to eliminate it.

Those with renal failure, whose aluminum serum levels are elevated as a result of dialysis or the oral administration of aluminum compounds, are also at special risk of developing osteomalacia, a softening of the bone that re-

sults in pain and an increased incidence of bone fractures.

Although aluminum has been implicated as being associated with Alzheimer's disease, a condition of premature senility [65], [66], the evidence to date suggests that it is not causal. Aluminum's role, if it plays a part at all, appears to be secondary to some unknown primary agent [67], [68].

With a few exceptions, most of the literature dealing with *inhalation exposures* of humans and experimental animals to aluminum and its compounds shows these materials to be relatively innocuous. However, there are reports contending that aluminum and aluminum oxide dusts are fibrogenic. A critical examination of the original references reveals that the evidence for this contention is inconsistent and mitigated by compounding variables.

Cases of pulmonary fibrosis attributed to aluminum powder exposures as reported in the literature [69–74] are equivocal. In most of these instances, environmental conditions were poor, dust exposures were extremely high, materials other than aluminum were known or suspected to be present, smoking and work histories of those affected were not given, and the types of fibrosis seen were not always consistent. Of these cases, 47 in all, one was reported from Japan and the remainder from Europe. No case has been reported from North America, even though similar production operations were conducted there over the same period of time.

A study of 125 employees involved in a powder stamping operation in the United States failed to show any evidence of fibrosis [75]. Their exposure times ranged from 6 to 23 years (mean average of 12). Furthermore, a therapeutic administration of heavy doses of aluminum powder, by inhalation, to silicotics over a period of 30 – 50 weeks showed no detrimental effects to the patients [75]. A report in 1973 indicated that the failure of inhaled metallic aluminum powder to cause pulmonary fibrosis in experimental animals "parallels the clinical experience in the United States but it leaves unexplained the occurrence of fatal pulmonary fibrosis in England from the inhalation of pyro aluminum powder" [76]. Worldwide, no case has been reported since 1962.

In 1947 cases of pulmonary fibrosis were attributed to aluminum oxide fume exposure [77]. Subsequent investigation demonstrated the presence of silica in these fumes [78]. Silica generally is accepted today as the principal cause.

The American Conference of Governmental Industrial Hygienists continues to recommend a 10 mg/m^3 TLV for aluminum metal and alumina dust, the same as for nuisance or biologically inert materials [79]. They also have established a lower limit applicable to aluminum pyro powders and aluminum welding fumes, 5 mg/m^3. This concentration is consistent with limits for nuisance materials with particle sizes in the respirable range. Health regulatory agencies in Australia, the Federal Republic of Germany, the Netherlands, Sweden, and many other countries consider aluminum oxide dusts to be nuisance particulates [80]. No MAK values have been set.

10. References

General References

1. *World Aluminum Abstracts,* American Society for Metals, Metals Park, Ohio. (This abstract bulletin is useful for keeping abreast of technical and economic developments in aluminum production. All published literature, government reports, theses, and patent literature are covered in this monthly journal).
2. R. Beck, H. Lax, W. Prast, J. Scott in J. Keefe (ed.): *Aluminum: Profile of the Industry,* McGraw-Hill, New York 1982.
3. M. H. Govett, J. Larsen (eds.): *The World Aluminium Industry,* **vol. 1,** Australian Mineral Economics Pty. Ltd., Sydney, Australia, 1981.
4. L. F. Mondolfo: *Aluminum Alloys: Structure and Properties,* Butterworths, Boston 1976.
5. J. E. Hatch (ed.): *"Aluminum, Its Properties and Physical Metallurgy,"* American Society for Metals, Metals Park, Ohio, 1984.
6. *Aluminum Statistical Review,* Annual Publication of the Aluminum Association, 818 Connecticut Avenue, NW, Washington, D.C. 20006.
7. M. Rolin: *Le Procédé Héroult-Etude de l'Electrolyse,* Institut National des Sciences Appliquees de Lyon, 1977; English translation-N. Phillips: The Electrochemistry of the Hall–Héroult Process, American Society for Metals, 1980.
8. M. Rolin: *Le Procede Hall–Héroult; l'Aspect Industriel,* Institut National des Sciences Appliquees de Lyon, 1979.

9. M. Rolin: *L'Electrolyse de l'Aluminium, (Le Procédé Hall–Heroult)* . Institut National des Sciences Appliquées de Lyon, Villeurbanne, France 1981.
10. K. Grjotheim, C. Krohn, M. Malinovsky, K. Matiasovsky, J. Thonstad: *Aluminium Electrolysis. Fundamentals of the Hall–Héroult Process,* Aluminium Verlag GmbH, Düsseldorf 1982.
11. K. Grjotheim, B. J. Welch: *Aluminium Smelter Technology,* Aluminium Verlag GmbH, Düsseldorf 1980.
12. W. E. Haupin, W. B. Frank: "Electrometallurgy of Aluminum," in J. O'M. Bockris, B. E. Conway, E. Yeager, R. E. White (eds.): Electrochemical Processing,*Comprehensive Treatise of Electrochemistry,* **vol. 2,** Plenum Publishing, New York and London, 1981, p. 301 – 325.

Specific References

13. N. A. Ashikhmina, A. I. Bogotikov, A. V. Gorshkov, A. V. Mokhov et al., *Dokl. Akad. Nauk. SSSR.* **746** (1979) no. 4, 958 – 961.
14. G. Ibe: "High-purity Aluminum for Electrotechnical Applications," *ETZ Elektrotech. Z* **101** (1980) no. 6, 366 – 368.
15. J. E. Hatch (ed.): *Aluminum Its Properties and Physical Metallurgy,* American Society for Metals, Metals Park, Ohio 1984, Chap. 1.
16. K. Wefers: " Properties and Characterization of Surface Oxides on Aluminum Alloys," *Aluminium,* **57** (1981) 722 – 726
17. A. W. Lemmon, Jr.: "Explosions of Molten Aluminum and Water," *Light Met.* (N.Y.) 1980 817 – 836.
18. P. D. Hess, R. E. Miller, W. E. Wahnsiedler, C. N. Cochran: "Molten Aluminum/Water Explosions," *Light Met. (N.Y.)* 1980 837 – 857.
19. K. J. Brondyke: "The Aluminum Industry in 1982 and Outlook for the 80's," *J. Metals* **54** (1983) no. 4, 63 – 67.
20. P. A. Foster, Jr.: "Phase Diagram of a Portion of the System Na_3AlF_6-AlF_3-Al_2O_3", *J. Am. Ceram. Soc.* **58** (1975) no. 7 – 8, 288 – 291.
21. A. Fenerty, E. A. Hollingshead: "Liquidus Curves for Aluminum Cell Electrolyte III. Systems Cryolite and Cryolite-Alumina with Aluminum Fluoride and Calcium Fluoride,"*J. Electrochem. Soc.* **107** (1960) no. 12, 993 – 997.
22. D. F. Craig, R. T. Cassidy, J. J. Brown, Jr.: "A Review of Phase Equilibria in the Na_3AlF_6-LiF- CaF_2-AlF_3-Al_2O_3 System," *Applications of Phase Diagrams in Metallurgy and Ceramics,* vol. 1, National Bureau of Standards SP-496, Washington, D.C., March 1978, p. 272 – 345.
23. K. Grjotheim, C. Krohn, M. Malinovsky, K. Matiasovsky, J. Thonstad: *Aluminium Electrolysis – Fundamentals of Hall–Héroult Process,* Aluminium Verlag GmbH, Düsseldorf 1982, p. 81 – 93, 119 – 126.
24. B. Gilbert, G. Mamantov, G. M. Begun, *Inorg. Nucl. Chem. Lett.* **10** (1974) 1123.
25. K. A. Bowman, Ph. D. Dissertation, University of Tennessee, Knoxville, March 1977.
26. J. J. DelCampo, M. P. Millet, M. Rolin, *Electrochim. Acta* **26** (1981) no. 1, 59 – 67.
27. W. E. Haupin, W. B. Frank: "Electrometallurgy of Aluminum," *Comprehensive Treatise of Electrochemistry,* vol. 2, Plenum Publishing, New York 1981, p. 308.
28. J. Thonstad, S. Rolseth, *Proceedings at 3rd ICSOBA Conference,* Nice, France September 1973, p. 657.
29. J. P. Millet, M. Rolin: "Electrolysis of Alumina in Molten Cryolite, Study of the Anodic Discharge Mechanism with Rotating Disk Electrode," *Fourth International Congress for the Study of Bauxites, Alumina, and Aluminum,* vol. 3, Athens, Greece October 9 – 12 1978, p. 265 – 286.
30. J. Thonstad, *Electrochim. Acta,* **15** (1970) 1569.
31. K. Vetter: *Electrochemical Kinetics,* Academic Press, New York 1967.
32. J. Thonstad, *Electrochim. Acta* **15** (1970) 1581.
33. G. Blyholder, J. Eyring, *J. Phys. Chem.* **61** (1957) 682.
34. A. J. Calandra, C. E. Castellano, C. M. Ferro, *Electrochim. Acta* **24** (1979) 425.
35. S. Rolseth, J. Thonstad, *Light Met. (N.Y.)* 1981, 289 – 301; *Proc. AIME Annual Meeting,* Chicago, Ill., Feb. 1981.
36. W. E. Haupin, *J. Met.* **42** (1971) no. 7, 41.
37. J. Thonstad, S. Rolseth, *Light Met.* 1983, 415 – 435; *Proc. AIME Annual Meeting,* Atlanta, Ga., March 1983.
38. N. Jarrett, W. B. Frank, R. Keller: "Advances in the Smelting of Aluminum," *Metallurgical Treatises,* The Metallurgical Society of AIME, Warrendale, Pa. 1981, p. 137 – 158.
39. *JANAF Thermochemical Tables,* 2nd ed., National Standards Reference Data Service, National Bureau of Standards, Washington, D.C., June 1971.

40. *JANAF Thermochemical Tables,* 1978 supplement; *J. Phys. Chem. Ref. Data* **7** (1978) no. 3, 793.
41. C. Toth, US 3615359 and 3615360, 1971.
42. C. N. Cochran: "Calculated Equilibrium for Carbothermic Reduction of Alumina," *Electrochem. Soc.,* Toronto May 1975.
43. P. T. Stroup, *Trans. Metall. Soc. AIME,* **203** (1964) 356.
44. M. J. Bruno: "Overview of Alcoa Direct Reduction Process Technology," *Light Met.*1984 1571 – 1590; *Proc. AIME Annual Meeting,* Los Angeles, Calif., Feb. 27 – Mar. 1, 1984.
45. R. L. Troup: "Feasibility of an Aluminum-Silicon Blast Furnace," *Light Met.*1984, 1591 – 1612; *Proc AIME Annual Meeting,* Los Angeles, Calif., Feb. 27 – Mar. 1, 1984.
46. D. T. Stevenson: "Feasibility of an Aluminum-Silicon Blast-Arc Process," *Light Met.*1984, 1613 – 1631; *Proc. AIME Annual Meeting,* Los Angeles, Calif., Feb. 27 – Mar. 1, 1984.
47. M. Dokiya, M. Fujishige, T. Kameyama, et al.: "Aluminum Blast Furnace," *Light Met.*1983 651 – 672; *Proc. AIME Annual Meeting,* Atlanta, Ga., Mar. 6 – 10, 1983.
48. A. Kozawa: "A New Nonelectric Aluminum Production Process, The Kuwahara Process," *New Materials and New Processes,* vol. 2, 1983, p. 374 – 378; Seminar, New Japanese Technologies and Materials, Case Center for Electrochemical Sciences, Cleveland, Ohio, October 25, 26, 1982.
49. H. Ginsberg, *Aluminium* **23** (1941) 131.
50. S. C. Jacobs, US 3303019 1967.
51. International Primary Aluminum Institute, *Environmental Legislation Survey,* London, U.K. 1981.
52. I. R. Campbell, J. S. Cass, J. Cholak, R. A. Kehoe: "Aluminum in the Environment of Man," *Arch. Ind. Hlth.* **15** (1957) 359.
53. J. R. J. Sorenson, I. R. Campbell, L. B. Tepper, R. D. Lingg: "Aluminum in the Environment and Human Health," *Environ. Health Perspect.* **8** (1974) 3.
54. G. L. Krueger, T. K. Morris, R. R. Suskind, E. M. Widner: "The Health Effects of Aluminum Compounds in Mammals," *Crit. Rev Toxicol.* **13** (1984) 1 – 24.
55. Hygienic Guide "Aluminum and Aluminum Oxide," *Amer. Ind. Hyg. Ass.* revised (1978).
56. W. Huepke, "Die Spurenstoffe," *Muench. Med. Wochenschr.* **92** (1950) 351.
57. *Report of the Task Group on Reference Man,* International Commission on Radiation Protection, ICRP Publication No. 23, Pergamon Press, Oxford 1975.
58. J. R. J. Sorenson, I. R. Campbell, L. B. Tepper, R. D. Lingg: "Aluminum in the Environment and Human Health," *Environ. Health Perspect.* **8** (1974) 3.
59. A. J. Lehman, W. I. Patterson: "F and DA Acceptance Criteria: Basic Considerations in Determinations Safety of Chemicals Used in Food-Packaging Materials," *Mod. Packag.* **28** (1955) 115, 172.
60. *Safety of Cooking Utensils,* FDA Fact Sheet, U.S. Food and Drug Administration, Washington, D.C., July 1971.
61. L. Kopeloff, S. Barrera, N. Kopeloff: "Recurrent Convulsive Seizures in Animals Produced by Immunologic and Chemical Means," *Am. J. Psych.* **98** (1942) 881 – 902.
62. H. M. Wisniewski, R. D. Terry, C. Pena, E. Streicher, et al.: "Experimental Production of Neurofibrillary Degeneration," *J. Neuropath. Exp. Neurol.* **24,** (1965) 139.
63. I. Klatzo, H. M. Wisniewski, E. Streicher: "Experimental Production of Neurofibrillary Degeneration: 1. Light Microscopic Observations," *J. Neuropath. Exp. Neurol.* **24** (1965) 187 – 199.
64. D. R. Crapper, A. J. Dalton: "Alterations in Short-Term Retention, Conditioned Avoidance Response Acquisition and Motivation Following Aluminum Induced Neurofibrillary Degeneration," *Physiol. and Behav.* **10** (1973) 925 – 933.
65. D. R. Crapper, S. S. Krishnan, S. Quittkat: "Aluminum, Neurofibrillary Degeneration and Alzheimer's Disease," *Brain* **99** (1976) 67 – 80.
66. G. A. Trapp, G. D. Miner, R. L. Zimmerman, A. R. Mastri, et al.: "Aluminum Levels in Brain in Alzheimer's Disease," *Biol. Psych.* **13** (1978) 709 – 718.
67. W. R. Markesbery, W. D. Ehmann, T. I. M. Hossain, M. Alauddin, et al.: "Instrumental Neutron Activation Analysis of Brain Aluminum in Alzheimer Disease and Aging," *Ann. Neurol.* **10** (1981) 511 – 516.
68. G. P. Cooper, G. L. Krueger, E. M. Widner: *Neurotoxicity of Aluminum,* University of Cincinnati, Cincinnati 1981, p. 27.
69. G. Goralewski: "Die Aluminiumlunge – eine neue Gewerbeerkrankung," *Ztschr. Ges. inn. Med.* **2** (1947) 665 – 673.
70. M. Ueda, Y. Mizoi, Z. Maki, R. Maeda, et al., *J. Med. Sci.* **4** (1958) 91.

71. J. W. Jordan: "Pulmonary Fibrosis in a Worker Using an Aluminum Powder," *Br. J. Ind. Med.* **18** (1961) 21.
72. J. Mitchell, G. B. Manning, M. Molyneux, R. E. Lane: "Pulmonary Fibrosis in Workers Exposed to Finely Powdered Aluminum," *Br. J. Ind. Med.* **18** (1961) 10.
73. N. P. Edling, *Acta Radiol.* **56** (1961) 170.
74. A. I. G. McLaughlin, G. Kazantzis, E. King, D. Teare, et al.: "Pulmonary Fibrosis and Encephalopathy Associated with the Inhalation of Aluminum Dust," *Br. J. Ind. Med.* **19** (1962) 253 – 263.
75. D. W. Crombie, J. L. Blaisdell, G. MacPherson, *Can. Med. Assoc. J.* **50** (1944) 318.
76. P. Gross, R. A. Harley, Jr., R. T. P. deTreville: "Pulmonary Reaction to Metallic Aluminum Powders," *Arch. Environ. Health* **26** (1973) 227.
77. C. Shaver, A. R. Riddell: "Lung Changes Associated with the Manufacture of Alumina Abrasives," *J. Ind. Hyg.* **29** (1947) 145.
78. C. M. Jephcott, *J. Occup. Med.* **5** (1948) 701.
79. American Conference of Governmental Industrial Hygienists (ACGIH) (ed.): *Threshold Limit Values* (TLV) 1983 – 1984, Cincinnati, Ohio, 1983, p. 10, 35, 52.
80. *The Measurement of Employee Exposures in Aluminium Reduction Plants,* International Primary Aluminium Institute, London 1982, p. 47.

Aluminum Alloys

J. Paul Lyle, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, United States

Douglas A. Granger, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, United States

Robert E. Sanders, Aluminum Company of America, Alcoa Technical Center, Pittsburgh, Pennsylvania 15219-1859, United States

1. Introduction 1
2. Metallurgy of Aluminum Alloys . . 2
2.1. Composition 2
2.2. Solidification and Cast Structure . . 4
2.2.1. Grain Size and Dendrite Arm Spacing 4
2.2.2. Segregation 5
2.2.3. Porosity 6
2.2.4. Nonmetallic Inclusions 7
2.3. Thermal and Mechanical Processes 7
2.3.1. Preheating 7
2.3.2. Non-Heat-Treatable Alloy Processes . 8
2.3.3. Heat-Treatable Alloy Processes 9
3. Production Methods for Aluminum Alloys 11
3.1. Melting and Metal Treatment 12
3.1.1. Impurities in the Molten Metal 12
3.1.2. Alloying 13
3.1.3. Recycling and Skim Treatment 13
3.1.4. Molten Metal Treatment 13
3.2. Casting Processes for Ingot 14
3.3. Shaped Casting 15
3.3.1. Die Casting 16
3.3.2. Permanent Mold Casting 17
3.3.3. Sand Casting 17
3.4. Mill Products 17
3.4.1. Plate, Sheet, and Foil 17
3.4.2. Extrusions 17
3.4.3. Forgings 18
3.4.4. Other Mill Products 18
3.5. Powder and Paste 18
4. Commercial Alloys, Products, and Uses 18
4.1. Aluminum Association Designations for Alloys and Tempers 18
4.2. Alloys for Castings 20
4.3. Alloys for Wrought Products 22
4.4. Comparison of Designation Systems 29
4.5. Alloys for Powder Metallurgy Products 31
5. Manufacture of Articles from Aluminum Products 31
5.1. Shaping 31
5.2. Joining 32
5.2.1. Mechanical Joining 32
5.2.2. Metallurgical Joining 32
5.2.3. Adhesive Bonding 33
5.3. Cleaning and Finishing 33
5.3.1. Mechanical Cleaning and Finishing . 34
5.3.2. Chemical Cleaning and Finishing . . . 35
5.3.3. Chemical Etching 35
5.3.4. Chemical and Electrolytic Brightening 35
5.3.5. Chemical Conversion Coating 35
5.3.6. Anodizing 36
5.3.7. Metal Coatings 38
5.3.8. Paints and Organic Films 38
5.3.9. Porcelain Enamels 40
6. Corrosion Resistance of Aluminum 40
6.1. Pitting Corrosion 41
6.2. Electrode Potentials 41
6.3. Crevice Corrosion 43
6.4. Intergranular Corrosion and Stress-Corrosion Cracking 43
6.5. Exfoliation Corrosion 43
6.6. Ratings for Resistance to Corrosion 44
6.7. Corrosion Prevention 44
7. Major Markets 44
8. References 45

1. Introduction

Most primary aluminum is used for metallurgical purposes: aluminum-based alloy castings and wrought products, alloying additions to metals, coatings on metals to prevent corrosion and oxidation, and reducing agents to produce metals from their oxides and fluorides. Of these metallurgical uses, aluminum-based alloy castings and wrought products make up about 85 – 90 % of aluminum shipments in the United States.

A comprehensive summary of the properties, production, fabrication, and use of aluminum alloys was published in 1967 [1–3]. Summaries of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_481

properties and physical metallurgy are given in [4], [5]. Commercial aluminum alloy products are described in [6], [8]. Commercial alloys of other metals that contain aluminum are given in [7], [8, pp. 237 ff.], [9]. Heat treatment [10], surface treatment [11], and joining [12] of aluminum alloys have been described.

2. Metallurgy of Aluminum Alloys

The properties of aluminum alloys depend on the metallurgical structure, which in turn depends on composition, solidification processes, and post-solidification thermal and deformation processing [1], [4], [5].

2.1. Composition

Many alloying elements are added to aluminum to enhance its properties (Table 1) [13], [14]. The major elements increasing strength are copper (see Fig. 21), magnesium (Fig. 1), manganese (Fig. 2), silicon (Fig. 3), and zinc (Fig. 4). All these have maximum solid solubility limits in excess of 1.5 % (Table 2) [1, pp. 362 – 3]. Elements with lower solubilities, e.g., chromium (Fig. 5), zirconium, and titanium, are added to certain alloys to aid in controlling aspects of metallurgical structure, e.g., as-cast grain size and the nature of recrystallization, and thus to affect the characteristics of the alloys. Of the elements commonly added, copper, magnesium, silicon, and zinc have high diffusion coefficients; other elements diffuse only slowly [1, p. 118].

Table 1. Elements in commercial aluminum alloys: maximum specified, wt %

Element	Wrought alloys	Casting alloys
Ag	–	1.0
Be	–	0.3
Bi	0.7	–
Co	1.9	0.3
Cr	0.40	0.6
Cu	6.8	11.0
Fe	2.0	2.0
Mg	5.6	10.6
Mn	1.8	0.8
Ni	2.3	3.0
O	0.5	–
Pb	0.7	0.25
Sb	–	0.3
Si	13.5	23.0
Sn	22.0	7.0
Ti	0.2	0.35
V	0.15	0.2
Zn	8.7	8.0
Zr	0.3	0.3

Commercial alloys are usually multicomponent systems, and the phase diagrams are frequently complex. One important precipitation-hardening system, Al – Mg – Si, is shown in Figure 6 (page after next) in the form of contour maps representing the solvus, solidus, and liquidus surfaces.

Table 2. Solid solubility of elements in aluminum

	Cr	Cu	Fe	Mg	Mn	Si	Zn *	Zr	Quasi-Mg_2Si	Quasi-$MgZn_2$
Eutectic or peritectic temperature, °C	661 **	548	655	450	659	577	382	661 **	595	475
Solubility at eutectic temperature, wt %	0.77 **	5.65	0.052	14.9	1.82	1.65	82.3	0.28 **	1.85	16.9
Solubility, wt % at										
650 °C	0.71	0.50	0.049	0.6	1.67	0.12	2.4	0.25	0.19	
600 °C	0.47	2.97	0.025	3.6	1.03	1.00	14.6	0.15	1.37	
550 °C	0.27	5.55	0.013	7.0	0.65	1.30	27.4	0.08	1.47	
500 °C	0.15	4.05	0.006	10.6	0.35	0.80	40.7	0.05	1.08	
450 °C	0.10	2.55		14.9		0.48	64.4		0.78	14.8
400 °C	0.07	1.50		11.5		0.29	81.3		0.52	10.7
350 °C		0.85		8.7		0.17	81.5			7.2
300 °C		0.45		6.3		0.10	79.0			4.5
250 °C		0.20		4.5		0.07	22.4			2.64
200 °C				2.9		0.05	12.4			1.34

* Miscibility gap between 31.6 % and 77.7 % Zn at 275 °C, closing at 60 % Zn at 353 °C
** Peritectic instead of eutectic

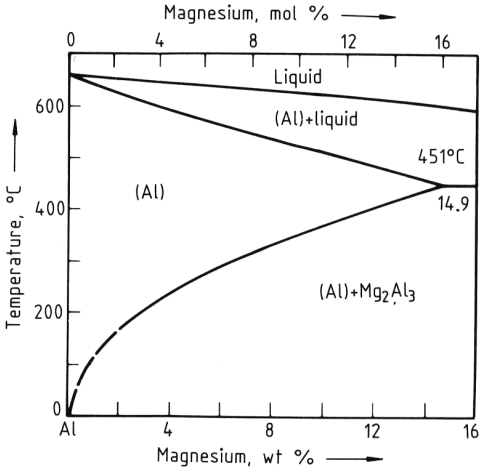

Figure 1. Aluminum – magnesium phase diagram (up to 16 wt % Mg)

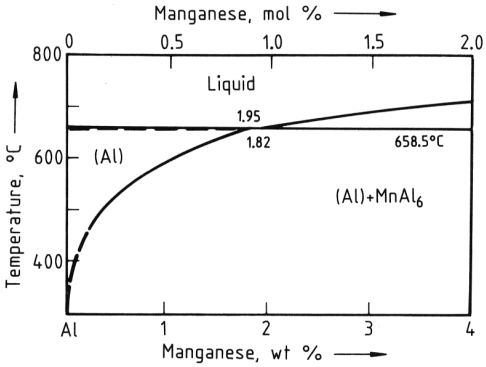

Figure 2. Aluminum – manganese phase diagram (up to 4 wt % Mn)

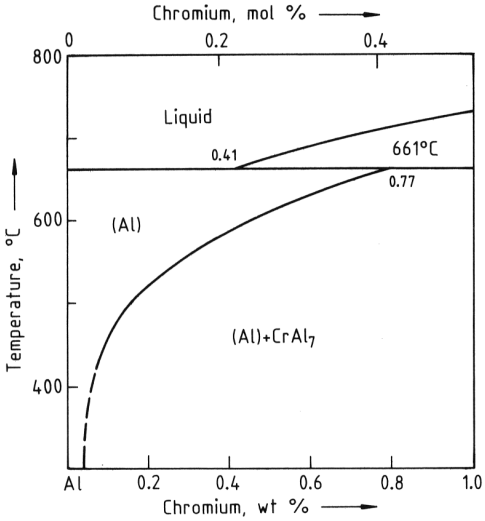

Figure 5. Aluminum – chromium phase diagram (up to 1 wt % Cr)

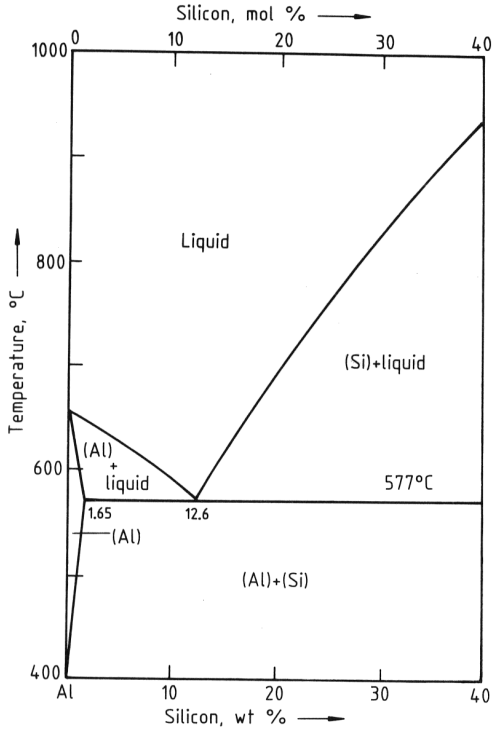

Figure 3. Aluminum – silicon phase diagram (up to 40 wt % Si)

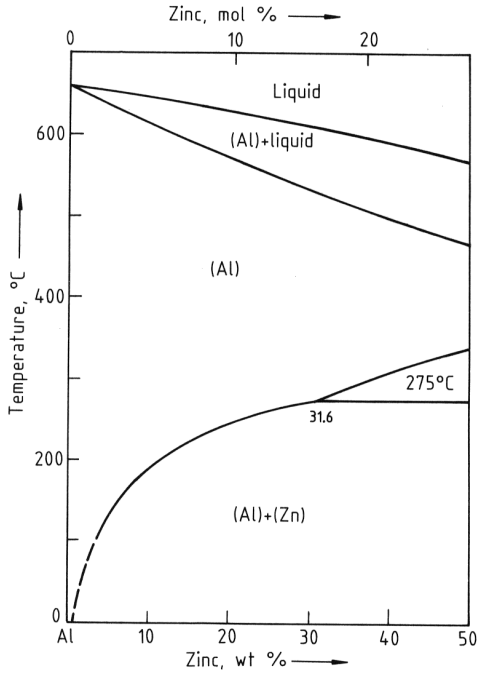

Figure 4. Aluminum – zinc phase diagram (up to 50 wt % Zn)

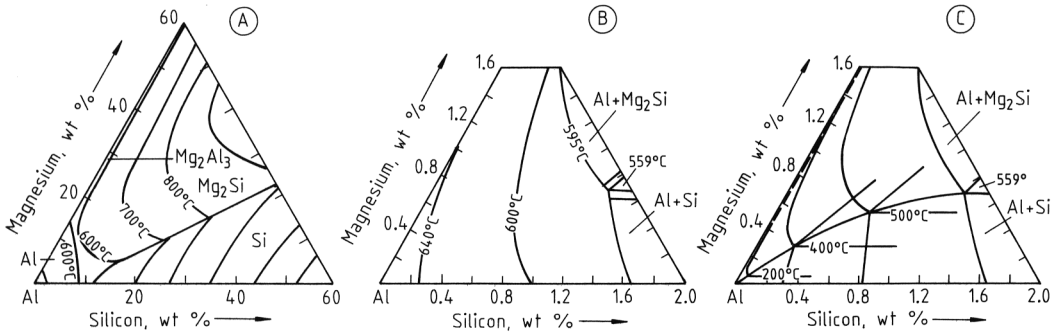

Figure 6. Al – Mg – Si phase diagram (aluminum-rich portion)
A) Liquidus; B) Solidus; C) Solvus

The composition of a specific alloy is often written as, e.g., Al – 5.1 Mg – 0.79 Mn – 0.11 Cr (Aluminum Association Number 5456), meaning an alloy consisting of 5.1 wt % Mg, 0.79 wt % Mn, 0.11 wt % Cr, and the rest aluminum and the usual impurities.

2.2. Solidification and Cast Structure

2.2.1. Grain Size and Dendrite Arm Spacing

The transformation from the liquid to the solid state in aluminum alloys is assisted by heterogeneous nucleation. Cast grain size is controlled by adding potent grain-refining nuclei to the melt to minimize the amount of undercooling and maximize the number of nucleation sites. Once nucleation has occurred, growth continues under commercial casting conditions with a dendritic structure, illustrated in Figure 7. An important characteristic of a dendrite is the dendrite arm spacing (DAS) or cell size. The DAS is related to the cooling rate during freezing, as shown in Figure 8 [15]. This relationship is described by the equation

$$d = kQ^{-n}$$

where d is the DAS in µm, k is a constant with a value of 35, Q is the cooling rate in K/s, and n is an exponent with a value of 0.33.

In wrought alloy ingot 500 mm thick, the DAS varies from about 15 µm close to the surface to 80 µm at the center. Roll-cast strip (6 mm thick) has a DAS ranging from 5 to 12 µm. In contrast, the grain size of grain-refined direct chill cast ingot ranges from 50 to 300 µm, depends mainly on nucleation and growth conditions, and is not strongly affected by cooling rate.

Figure 7. Typical dendritic growth forms: photomicrograph of electropolished as-cast 2124 alloy taken under polarized light

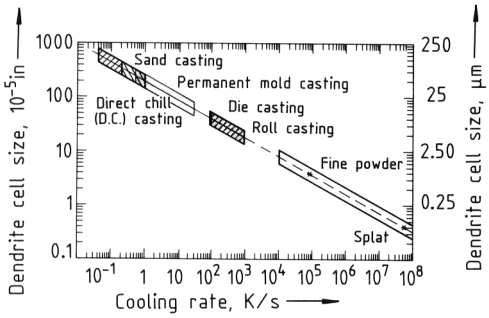

Figure 8. Dendrite cell size (DAS) as a function of cooling rate

2.2.2. Segregation

In most aluminum alloys the solute (or alloying element) is less soluble in the solid phase than in the liquid phase (see Fig. 9), and addition of solute lowers the freezing point. As a consequence, when an aluminum alloy solidifies, solute is rejected at the solid – liquid interface, resulting in (1) solute buildup in the liquid ahead of the freezing front and (2) breakdown of the planar solidification interface. The former gives rise to segregation in the cast structure, commonly called coring, and is directly responsible for the dendritic growth form. Segregation occurs on both microscale and macroscale. *Microsegregation* is that which forms on the scale of the dendrite arm spacing, whereas *macrosegregation* describes solute differences over distances of several centimeters.

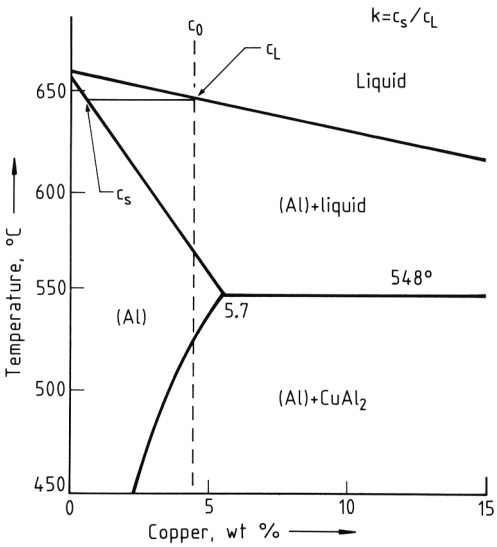

Figure 9. Al – Cu phase diagram (up to 15 wt % Cu)

Microsegregation is best described in terms of the solute distribution between two dendrite arms. A typical solute distribution profile in a direct chill cast ingot is illustrated in Figure 10. (The copper at different locations within the volume element was determined by electron-probe microanalysis.) If the solidifying metal between two dendrite arms is considered a closed volume element with no mass flow of liquid in or out (a reasonable first approximation), then the redistribution that takes place can be described adequately by the Scheil equation:

$$c_s = c_0 k\,(1-)\,f_s^{k-1}$$

where c_s is the composition of the solid at fraction solid f_s, c_0 is the original alloy composition, and k is the equilibrium partition ratio (see Fig. 9) [16].

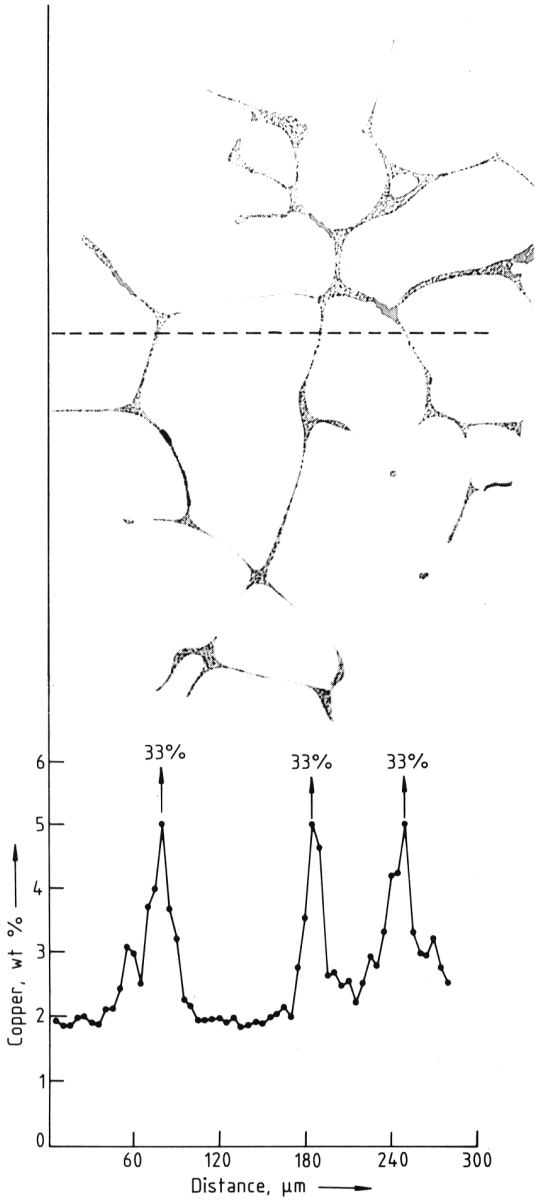

Figure 10. Copper distribution between dendrite arms in direct chill (D.C.) cast 2124 alloy
– – – – – Analysis path

Macrosegregation arises because of the flow of segregated liquid from one area of a solidifying casting to another. A typical solute distribution across a 600-mm-thick commercial direct chill ingot of 2024 alloy is illustrated in Figure 11. The driving forces for macrosegregation are (1) shrinkage contraction that accompanies the transformation of liquid to solid, (2) density differences in the liquid phase, and (3) convective stirring that arises from the temperature gradient across the crater of a solidifying ingot.

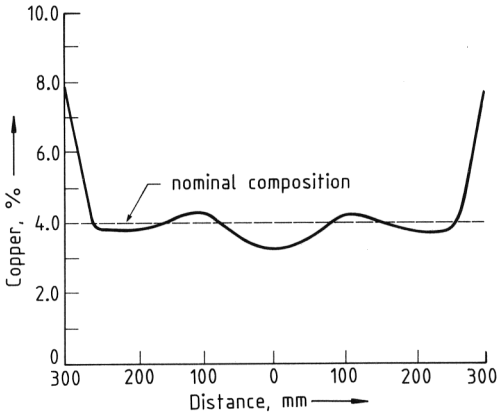

Figure 11. Variation in copper concentration across a 600-mm-thick direct chill (D.C.) cast ingot of Al – Cu – Mg alloy

Macrosegregation has been described with some success by analogy to fluid flow through a porous medium, i.e., with D'Arcy's law

$$v = -\frac{k}{\eta\varphi_l}\left(\nabla P + \varrho_l g\right)$$

where v is the mean interdendritic flow velocity, k is a constant for the permeability of the medium, P is pressure, g is the acceleration due to gravity, η is viscosity, φ_l is the volume fraction of liquid, and ϱ_l is the liquid density [16].

2.2.3. Porosity

Porosity in aluminum ingot and castings arises from two principal sources: (1) shrinkage contraction and (2) the rejection of dissolved gas on freezing. The characteristic feature of *shrinkage porosity,* which arises from inadequate feeding of the solidifying casting with molten metal, is its angularity. In contrast, *gas porosity* is more or less spherical. All commercial aluminum castings contain some gas above the solid-state equilibrium level, and all shrinkage porosity contains some gas.

Hydrogen is the only gas that is significantly soluble in aluminum. At the melting point of aluminum, hydrogen is 20 times more soluble in the liquid than in the solid (Fig. 12) [17]. Even when the hydrogen level is reduced below the equilibrium concentration in molten aluminum by fluxing, the low solubility of the gas in the solid causes hydrogen to go into the remaining liquid. When the internal gas pressure exceeds atmospheric pressure, a bubble of gas can form. For a given gas content, the amount of porosity is increased by increased alloy freezing range and decreased cooling rate (Fig. 13) [18].

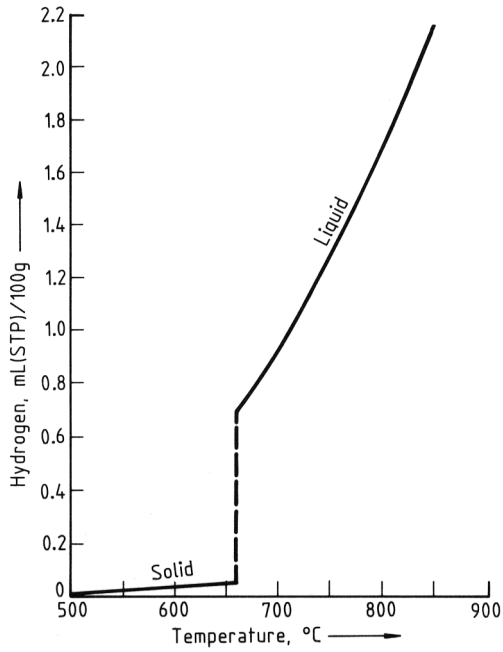

Figure 12. Solubility of hydrogen at 0.1 MPa (1 bar) in aluminum

There is no measurable porosity, regardless of the conditions, if the hydrogen content is below about 0.10 mL per 100 g. At these low hydrogen contents the internal pressure in the liquid cannot exceed atmospheric pressure. Instead, the gas in excess of the solid-state equilibrium level at the freezing point coprecipitates with vacancies behind the solid – liquid freezing front. The pores are extremely small (1 µm) and contain hydrogen at pressures of about 20 bar (2 MPa).

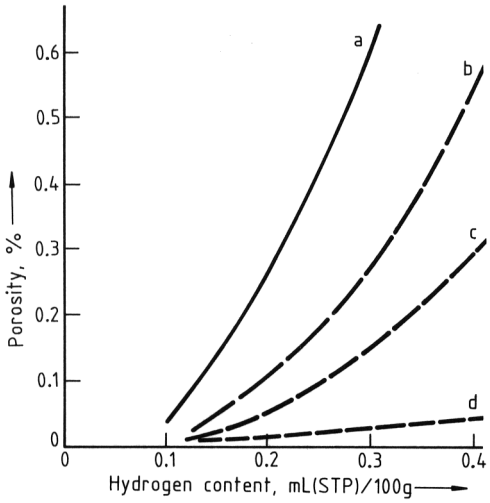

Figure 13. Relationship between porosity and hydrogen content in semicontinuously direct chill (D.C.) cast ingot a) Thick ingot; b) Thin ingot; c) 99.2 % Al; d) 99.99 % Al Both thick (a) and thin (b) ingots are long freezing range alloy 7xxx.

This porosity is not readily discernible in the cast structure, but the pores expand considerably on heating. The terms primary and secondary porosity are used to distinguish macroscopic and microscopic forms of hydrogen porosity [19].

2.2.4. Nonmetallic Inclusions

There are three major types of solid inclusions in aluminum alloys: (1) oxides, (2) borides, and (3) carbides.

Molten aluminum reacts readily with moisture in the atmosphere. The resulting aluminum *oxide* film is generally quite coherent and inhibits further growth. But whenever the molten metal surface is disturbed, e.g., during the mixing in of alloying elements or tapping from the furnace to a transfer trough, a fresh oxide layer is formed immediately on the newly exposed surface. Therefore, oxides found in cast structures are usually filmlike. An exception is spinel, $MgAl_2O_4$, which grows on Al – Mg alloy melts when the first-formed MgO film crystallizes and "breakaway" oxidation ensues. Spinel is particulate.

Borides originate from the presence of the transition elements Zr, V, and Ti along with boron in the Hall – Héroult cell. These impurities, which come from the alumina feed and anode material, are somewhat soluble at 950 °C, but transition metal borides precipitate as the metal cools to furnace temperature (750 °C). Another source of boride inclusions, but usually of much smaller size (0.5 to 3 µm), is the Al – Ti – B master alloy used for grain refining.

Carbides, generally in the form of Al_4C_3, also arise from the cell metal. At 950 °C carbon has a solubility of 50 ppm in molten aluminum. As the metal cools, the solubility drops to 1 ppm, with the resultant precipitation of carbide particulate.

2.3. Thermal and Mechanical Processes

2.3.1. Preheating

Ingot to be fabricated to wrought products is usually preheated or homogenized [4, Chap. 5]. This treatment, often near the melting temperature, lasts a few hours to a few days. Preheating alleviates microsegregation (coring and soluble grain boundary segregation) by diffusion and makes the ingot more amenable to hot deformation. Preheating followed by slow cooling is used to bring certain elements, e.g., manganese, into solid solution and then precipitate them into a more desirable form, usually very fine dispersoids, for control of recrystallization, formability, toughness, finishing, etc. During homogenization, secondary porosity increases in size, but the porosity will eventually disappear, collapsing under the force of surface tension, if the conditions in the furnace are favorable for the loss of hydrogen from the ingot.

During solidification, insoluble intermetallics, such as $FeAl_3$, precipitate with an angular geometry. During homogenizing heat treatments this type of precipitate becomes less angular, and the fabricability of the ingot is enhanced.

There is sufficient residual stress in a direct chill cast ingot to cause recrystallization and grain growth in alloys containing no insoluble particles, as shown in Figure 14.

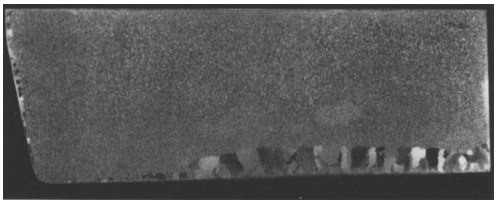

Figure 14. Abnormal grain growth in a direct chill (D.C.) cast 5252 alloy ingot homogenized at 565 °C for 4 h (0.09X) Aqua regia etch

2.3.2. Non-Heat-Treatable Alloy Processes

The non-heat-treatable alloys are strengthened by dispersion, solid solution, and work hardening [4, Chap. 4]. The common alloy systems of this type are unalloyed aluminum containing controlled levels of impurities (mostly Fe and Si), Al – Mn, and Al – Mg. Strength increases with increasing content of the alloying component (Fig. 15). Magnesium is more soluble in aluminum than Fe and Si or Mn, and Al – Mg alloys are capable of reaching the highest strengths. Strength increases also when the dislocation density is increased by deformation: the increase is determined by the composition and the amount and temperature of deformation. Figure 16 shows for three alloys, how the yield strengths are increased and the elongations decreased by rolling at room temperature.

The strain induced by deformation can be reduced by heating; the rate of reduction is determined by alloy, amount of strain, temperature, and time. Figure 17 shows the effects of heating time and temperature on the yield strength and elongation of alloy 1100 given about 75 % cold reduction (H18 temper). The effect of lower strains (H12 and H14) at 204 °C is illustrated in Figure 18. The initial drop in strength is caused by recovery processes involving changes in dislocation structure. Higher temperatures and longer times increase the degree of recrystallization and thus the ductility, and decrease strength, until full annealing is achieved. Heating to achieve properties between strain hardened and fully annealed is known as partial annealing.

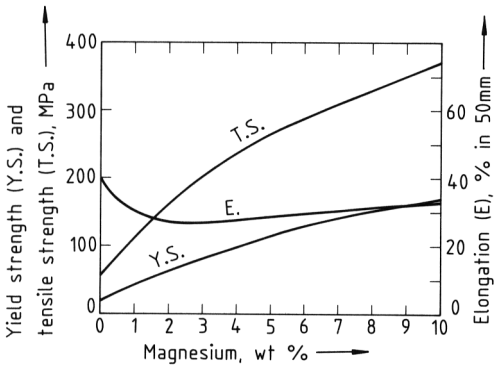

Figure 15. Effect of magnesium content on tensile properties

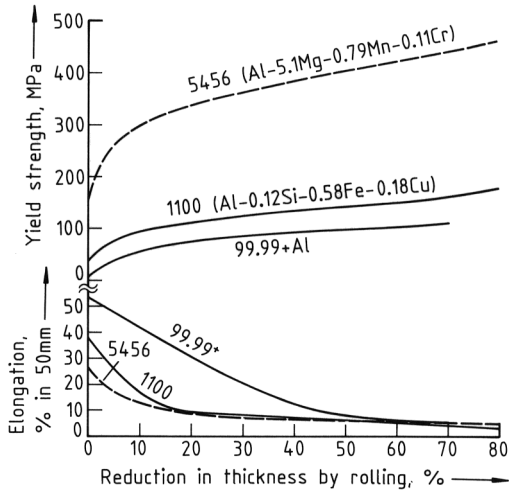

Figure 16. Effect of cold rolling on yield strength and elongation of three alloys

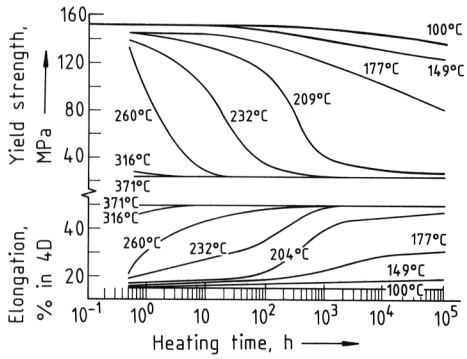

Figure 17. Effect of heating time and temperature on yield strength and elongation of 1100-H18
The gage length over which the elongation is measured is four times the diameter of the specimen (4D)

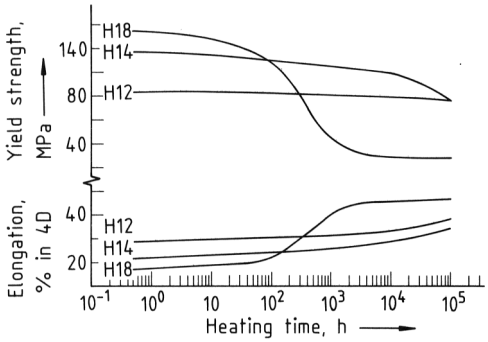

Figure 18. Effect of strain (H12, H14, H18) and heating time at 204 °C on yield strength and elongation of 1100 alloy

Strain-hardened Al – Mg alloys recover at room temperature, the rate of recovery increasing with Mg content and amount of original strain. This long-term recovery can be prevented by stabilizing at moderate temperatures. Figure 19 illustrates the effects of the recovery process and stabilization on an alloy cold worked on four strain levels.

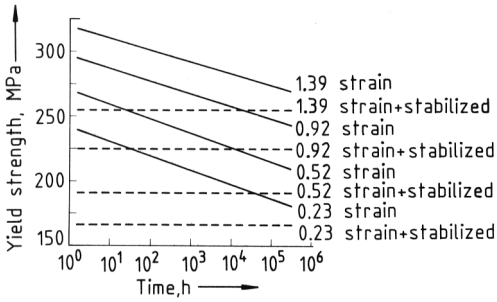

Figure 19. Recovery at room temperature of strain-hardened and stabilized (8 h, 143 °C) Al – 2.43 Mg – 0.25 Cr alloy sheet

2.3.3. Heat-Treatable Alloy Processes

Heat-treatable aluminum alloys depend on precipitation from supersaturated solid solution to develop high strength [4, Chap. 5], [10, pp. 675 – 718]. The basic requirements for alloying elements are high solid solubility at elevated temperatures, low solubility at room and moderate temperatures, and precipitates that strain the lattice. The typical precipitation sequence is

supersaturated solid solution → Guinier – Preston zones → metastable phases → stable phases

Examples of the sequence are given in Table 3 for several systems [20–22]. The major commercial heat-treatable alloy systems are Al – Cu, Al – Cu – Mg, Al – Cu – Si – Mg, Al – Si – Mg – Cu, Al – Mg – Si – Cu, Al – Mg – Si, Al – Zn – Mg, and Al – Zn – Mg – Cu. Alloys containing lithium, which has a solid solubility limit of 4 wt % at 600 °C, have been the subject of extensive research and have become commercial in limited applications.

Table 3. Aluminum alloy precipitation sequences

Al – Cu	S.S.$\alpha \rightarrow$GP[1]$\rightarrow$GP[2] or $\theta'' \rightarrow \theta' \rightarrow \theta(CuAl_2)$
Al – Si	S.S.$\alpha \rightarrow$Si
Al – Mg	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow \beta' \rightarrow \beta(Mg_3Al_2)$
Al – Cu – Mg	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow$S$' \rightarrow$S(Al_2CuMg)
Al – Mg – Si	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow \beta' \rightarrow \beta(Mg_2Si)$
	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow \eta' \rightarrow \eta(MgZn_2)$
Al – Zn – Mg – Cu	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow \eta' \rightarrow$T$(Mg_3Zn_3Al_2)$
	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow$T$' \rightarrow$T$(Mg_3Zn_3Al_2)$
Al – Li	S.S.$\alpha \rightarrow \delta'(Al_3Li) \rightarrow \delta(AlLi)$
Al – Li – Cu	S.S.$\alpha \rightarrow$G.P. zones$\rightarrow \theta'' \rightarrow \theta' \rightarrow \theta(Al_2Cu) \rightarrow T_1(Al_2CuLi)$
Al – Li – Mg	S.S.$\alpha \rightarrow \delta'(Al_3Li) \rightarrow Al_2LiMg$

As the precipitation sequence proceeds, lattice strain and strength increase, whereas ductility decreases and then increases. The strength can increase at room temperature (natural aging) or at moderate temperatures (artificial aging or precipitation heat treatment). Aging past peak strength is known as overaging and for practical purposes occurs only at elevated temperatures.

The kinetics of the precipitation sequence and the magnitude of the effects on properties are determined by composition, temperature, and strain. Some effects of composition on *natural aging* are illustrated in Figure 20. The Al – Cu – Mg alloy (2024) effectively reaches peak strength after about 100 h. The Al – Mg – Si

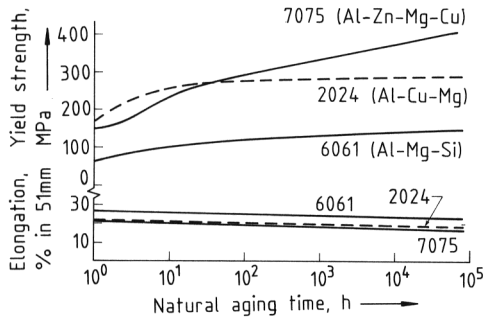

Figure 20. Effect of natural aging time on strength and ductility of some aluminum alloys

alloy (6061) also hardens rapidly for the first 100 h, but strength continues to rise more rapidly than for the Al – Cu – Mg alloy for at least ten years. The strength of the Al – Zn – Mg – Cu alloy (7075) rises even more rapidly than that of the Al – Mg – Si alloy, with no sign of leveling. Ductilities, as indicated by elongation tests, fall slowly as strengths rise.

The classic system for studying precipitation hardening and *artificial aging* is Al – Cu. The equilibrium phase diagram (Fig. 21) shows the maximum solid solubility of 5.65 wt % at 548 °C and a solubility of less than 0.2 % at 250 °C. The aging sequence of an Al – 4 Cu alloy is illustrated in Figure 22 [23]. (Al – 4 Cu is an alloy consisting of 4 wt % Cu and the rest aluminum.) After solution heat treating above 500 °C and quenching to retain a supersaturated solid solution, artificial aging proceeds at moderately elevated temperatures as shown, with initially rising and then falling hardness. The changes in hardness are associated with the appearance of G – P (Guinier – Preston) zones and their transformation to θ' and θ.

Artificial aging accelerates the precipitation sequence; higher strengths are obtained within practical time frames. (Compare the curves in Figure 23 with the curve for 2024 in Figure 20.) The decrease in strength from *overaging* is associated with changes in metallurgical structure, resulting in more stable properties and dimensions and increased toughness and resistance to corrosion, in particular to exfoliation and stress-corrosion cracking in the Al – Zn – Mg – Cu alloys. These effects are illustrated in Figures 24 and 25 [24].

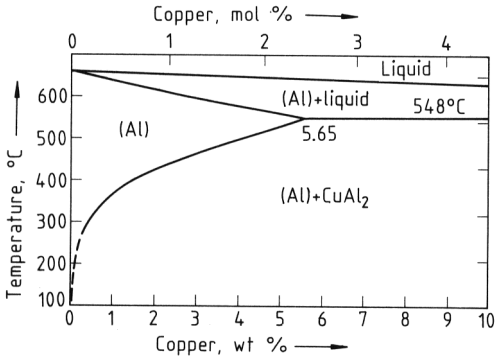

Figure 21. Aluminum – copper phase diagram (up to 10 wt % Cu)

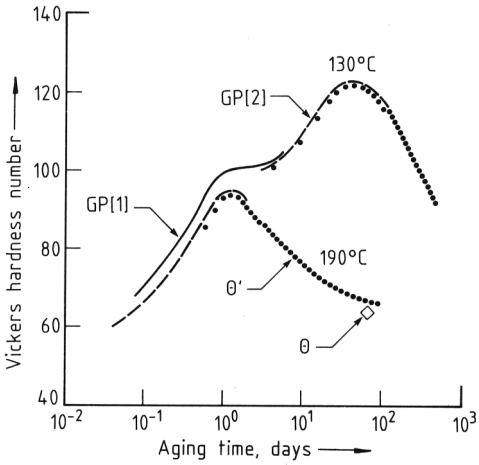

Figure 22. Effect of aging time and temperature on hardness and structure of Al – 4 Cu alloy

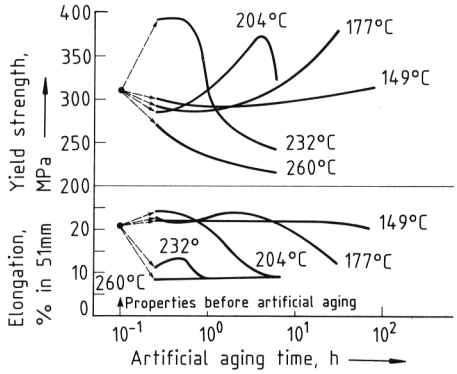

Figure 23. Effect of aging temperature and time on strength and ductility of an Al – Cu – Mg alloy (2024) not cold worked between quenching and aging

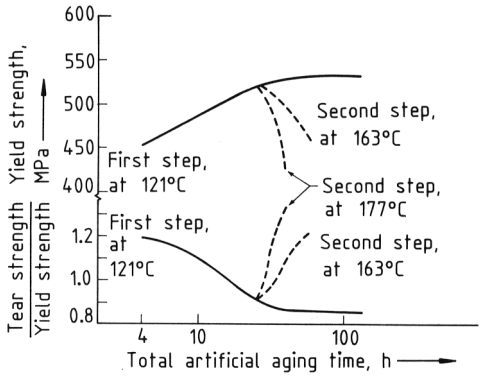

Figure 24. Effect of overaging on strength and toughness of an Al – Zn – Mg – Cu – Cr alloy (7075)

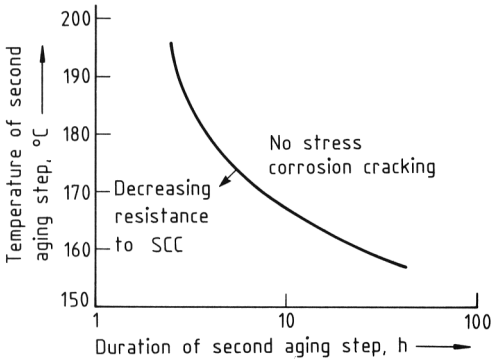

Figure 25. Effect of overaging time and temperature on stress-corrosion cracking of Al – Zn – Mg – Cu – Cr alloys containing 4.5 – 7.5 % Zn, 1.5 – 3.8 % Mg, 0.75 – 2.5 % Cu, and 0.05 – 0.5 % Cr

The precipitation sequence is accelerated also by *strain* introduced after quenching. The effects of strain on natural aging and on artificial aging of an Al – Cu – Mg alloy are shown in Figures 26 and 27, respectively. The introduction of strain after artificial aging also increases strength and reduces ductility (Fig. 28).

The preceding discussion is essentially in terms of strength and ductility. However, much of the manipulation of metallurgical structure by composition, solidification rate, thermal practices, and deformation of commercial products is done to obtain an optimum combination of properties and characteristics, including strength, ductility, toughness, formability, corrosion resistance, finishing characteristics, fatigue resistance, and machinability.

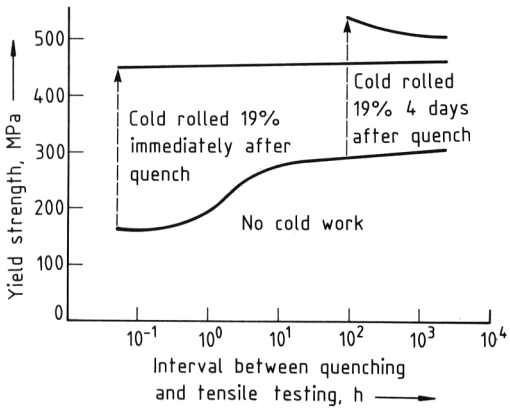

Figure 26. Effect of cold work after quenching and natural aging on yield strength of Al – Cu – Mg alloy (2024)

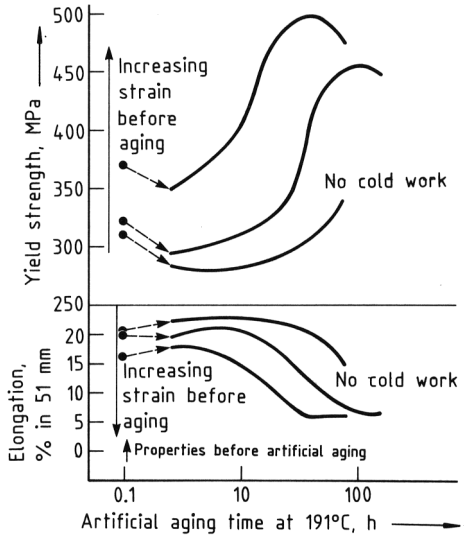

Figure 27. Effect of strain between quenching and artificial aging of Al – Cu – Mg alloy (2024)

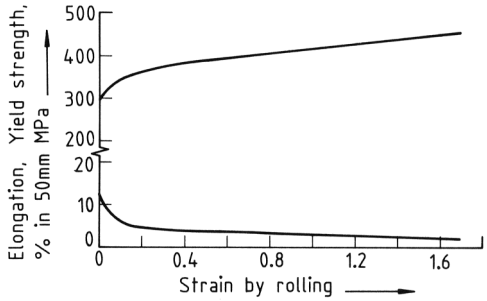

Figure 28. Effect of strain after artificial aging on strength and ductility of an Al – Mg – Si alloy (6061)

3. Production Methods for Aluminum Alloys

The flow of aluminum from reduction plants to end product, as well as the important scrap loops, is shown schematically in Figure 29. Producers of mill products and castings use both primary and secondary metal to produce the intermediate products that manufacturers need to produce final products.

The production of mill products and castings begins with the melting and casting operation. Castings are made in final or near-final shapes. For mill products, ingots are cast with rectangular or circular cross section and subsequently

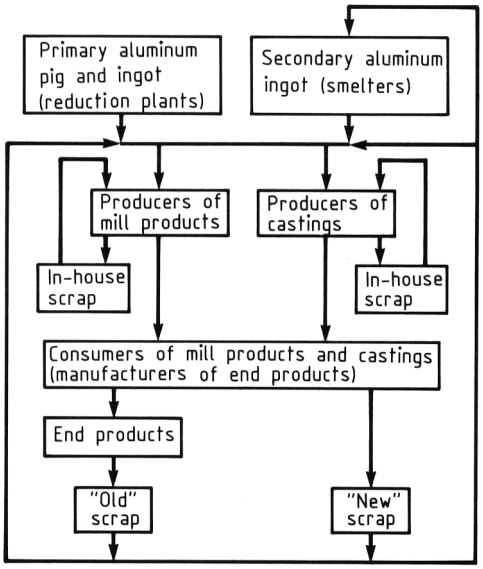

Figure 29. Flow of aluminum alloy products

worked down to the desired shapes. Alternate methods for producing thin sheet and foil consist of continuously casting and rolling thin wide strip. In similar processes for producing rod, bar, and wire, the alloy is continuously cast into a small bar, rolled through grooved rolls, and drawn through dies. Other processes begin with atomized powders, which subsequently are compacted and then deformed.

3.1. Melting and Metal Treatment

Aluminum and aluminum alloys are melted before conversion into castings, ingot for subsequent fabrication or remelting, particulate for chemical applications and alloying with other metals, and powders for pigments and powder metallurgy. The components of the melt may include molten metal directly from the reduction cell, primary ingot, scrap aluminum, remelted scrap ingot, and alloying elements [3, Chap. 1].

3.1.1. Impurities in the Molten Metal

The principal impurities in aluminum from the Hall – Héroult cell are listed in Table 4. The ranges are shown as a guide and are not absolute values.

Table 4. Impurities in aluminum from the Hall-Héroult cell

Impurity	Usual concentration range, %
Fe	0.06 – 0.35
Si	0.05 – 0.15
H	ca. 3×10^{-5} *
Na	0.002 – 0.004
Ca	0.002 – 0.004
Li	0.002 – 0.004
Ti, Zr	0.005 – 0.020
V, Cr	0.005 – 0.020
Al_2O_3	variable
Al_2MgO_4 (spinel)	variable
Al_4C_3	variable
TiB_2, etc.	variable

* 0.20 – 0.60 mL per 100 g Al

Hydrogen is the only gas significantly soluble in molten aluminum. It arises in the reaction of molten metal with moisture in the atmosphere, refractories, or reaction products. The standard free energy (ΔG°) for the reaction

$$3\,H_2O\,(g) + 2\,Al\,(l) \longrightarrow Al_2O_3 + 3\,H_2\,(g)$$

is high:

$$\Delta G^\circ = -979100 - 719\,T\log T + 413\,T \quad (\Delta G^\circ \text{ in J/mol})$$

The molecular hydrogen dissociates and enters the metal interstitially:

$$H_2\,(g) \rightleftharpoons 2\,H\,(soln)$$

At equilibrium:

$$K = \frac{a_H^2}{a_{H_2}} \approx \frac{c^2}{P}$$

where K is the equilibrium constant, a_H and a_{H_2} are activities, c is the concentration of dissolved gas, and P is the gas pressure. The concentration of dissolved gas is given then by

$$c = k\sqrt{P} \quad \text{where} \quad k = \sqrt{K}$$

commonly known as Sievert's law. Reaction of molten aluminum with moisture is not only the source of hydrogen but also the principal source of oxide contamination. The alkali metals and alkaline earth metals (Li, Na, and Ca) come mainly from the electrolyte in the Hall – Héroult cell and from refractories. Borides originate from impurities in the carbon anodes and alumina, whereas carbon enters into solution by dissolution of the anodes.

3.1.2. Alloying

Alloying is normally carried out in the temperature range 700 – 730 °C. The lower-melting additions, such as Zn, Mg, Cu, and Si, are made in elemental form, whereas the higher-melting point additions are made as master alloys or briquettes (Table 5).

Table 5. Aluminum master alloys and briquettes for alloying

Composition	Principal constituents
Al – 10 % Cr	$CrAl_7$
Al – 10 % Mn	$MnAl_4$ and $MnAl_6$
Al – 25 % Fe	Al_3Fe
Al – 6 % Zr	Al_3Zr
Al – 3 or 4 % B	AlB_2 and AlB_{12}, respectively
75 % Mn – Al	Al and Mn powder
75 % Cr – Al	Al and Cr powder

3.1.3. Recycling and Skim Treatment

Great emphasis is placed on the recovery of aluminum from scrap. The principal motivation is energy savings; about 20 times as much energy is needed to produce aluminum from ore as to produce the equivalent amount from scrap. Recycling of beverage cans is increasing in the United States as primary producers, beverage companies, secondary metal dealers, and civic organizations continue to open collection centers. The growth of recycling of aluminum cans in the United States is shown in Table 6 [25, p. 30].

Table 6. Aluminum can reclamation, 1986 – 1996

Year	Aluminum collected, 10^6 kg	No. of aluminum cans collected, 10^9	Percentage of aluminum cans collected
1986	550	33.3	48.7
1987	596	36.6	50.5
1988	672	42.5	54.6
1989	753	49.4	60.8
1990	863	55.0	63.6
1991	879	56.8	62.4
1992	956	62.7	67.9
1993	900	59.5	63.1
1994	959	64.7	65.4
1995	900	62.7	62.2
1996	879	62.8	63.5

The primary problems encountered in the recovery of aluminum from scrap are *melt loss* and *contamination.* Prevention of large losses when melting oily scrap, lightweight cans, foil, etc., requires special techniques, including melting in molten salt (NaCl or NaCl + KCl) and ablative melting, i.e., submerging the scrap below the surface of a molten aluminum charge. The principal contaminants encountered in recycled metal are silicon, iron, and titanium (from paint pigment on beverage containers). Because there are no simple methods of removing these elements from molten aluminum, scrap must be presorted before it is melted. Many foundry alloys can contain high levels of these elements, but for wrought alloys the only way to recycle scrap is to dilute the contaminants with pure metal.

The *skim* that forms on molten aluminum may be 70 – 95 % unoxidized aluminum metal, depending on the type of charge, furnace, holding temperature, alloy, and fluxing treatment. A variety of methods have been devised to recover metal from skim. Most depend on a combination of mechanical crushing, screening, and mixing in a molten salt bath to separate metal from the oxidized fraction. The salt bath is normally an eutectic mixture of NaCl and KCl with 5 – 10 % cryolite.

3.1.4. Molten Metal Treatment

Most uses of aluminum alloys demand high quality with respect to contaminants. For example, sheet for containers, plate, sheet and forgings for aircraft structures, memory disks for computers, and articles with a decorative finish, such as automotive trim, all require low levels of nonmetallic impurities. Hydrogen is particularly deleterious because it causes gas porosity in the cast ingot, possibly leading to defects such as blisters in sheet and extrusions and reduced mechanical properties in forgings and thick plate.

Even under the most favorable melting conditions, some degassing treatment is usually required prior to pouring high-quality ingot or castings. Treatment may be either batch in the furnace or continuous during transfer of the melt from the furnace to the casting unit. In the batch method the melt is purged with Ar, N_2, or Cl_2 or a mixture of Cl_2 with either of the inert gases. However, because furnace fluxing is inefficient and environmentally damaging, it is being replaced by continuous processes, which often

combine gas removal with filtration. A number of these systems are described in [26]. Figure 30 illustrates a typical bed filtration system (Alcoa 528) combined with a gas purge for hydrogen and alkali metal reduction (Alcoa 622). Figure 31 shows a successful method of filtration with a porous ceramic plate [27].

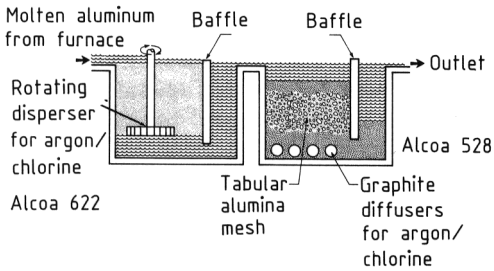

Figure 30. Molten metal fluxing and filtering system with Alcoa 622 and 528 metal treatment units

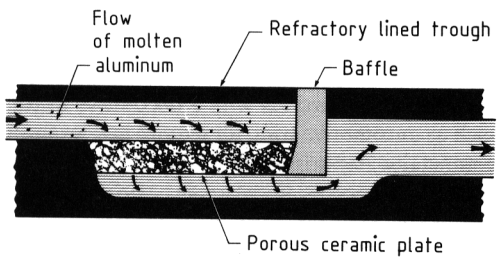

Figure 31. Filtration with a porous ceramic plate

Before the metal is cast into ingot, *grain refiner* is usually added to control the grain structure of the ingot. Reducing the as-cast grain size has a number of benefits, such as reducing the sensitivity of the metal to hot cracking during solidification, enhancing the hot workability of the ingot, and producing a wrought structure that gives a uniform appearance on etching and anodizing. Grain refinement is carried out most effectively by the addition of titanium and boron in the form of an Al – Ti – B master alloy. The master alloy may be added to the furnace but preferably is added to the molten metal stream between the furnace and the casting unit. Typical titanium additions range from 0.003 % to 0.015 %.

3.2. Casting Processes for Ingot

Direct chill (D.C.) *casting* is the primary method of ingot production in the aluminum industry [28]. Large ingots with either a rectangular or circular cross section are cast for fabrication by rolling to sheet or extruding to shaped profiles. Some circular cross section ingot is used as forging stock and starting material for rod and wire production. Figure 32 is a schematic illustrating the D.C. casting process developed around 1940, and Figure 33 shows a later modification in which level transfer is used to avoid turbulence with the associated oxide generation and hydrogen pickup. The level-pour method produces high-quality ingot for forged products of highest internal integrity. Direct chill ingot is usually cast in multiples, and the drop is terminated when the ingot length reaches the depth of the casting pit. A *continuous casting* process, in which the ingot is cast horizontally, as depicted in Figure 34, is widely used to produce a limited number of alloys when large amounts of ingot of one size are required.

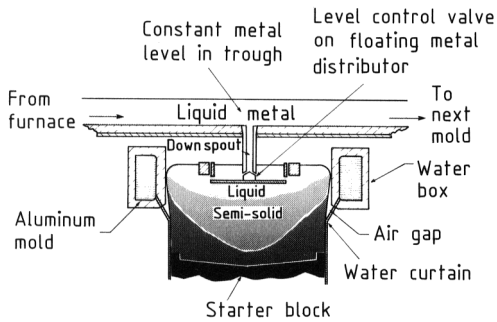

Figure 32. Vertical direct chill (D.C.) casting

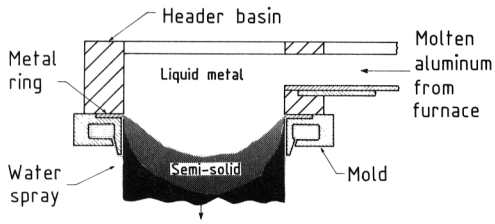

Figure 33. Level-pour mold assembly

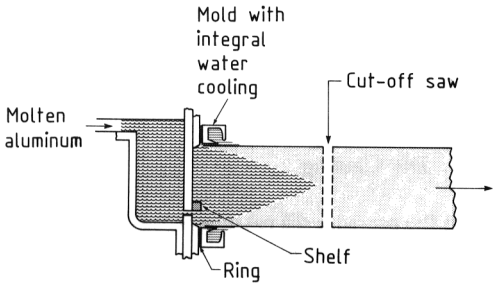

Figure 34. Horizontal direct chill (D.C.) casting

A development in semicontinuous thick ingot casting is *electromagnetic casting* (EMC). In this method the traditional mold is replaced by an electromagnetic field that contains the column of molten metal until it enters the direct water cooling system. This produces a smooth-surface ingot free from the subsurface segregation of the older methods (Fig. 35) [29]. In some applications the EMC ingot can be fabricated without any prior surface machining or scalping. The major capital investment for scalping equipment and the melt loss, which occurs when scalper chips are remelted, are thus avoided.

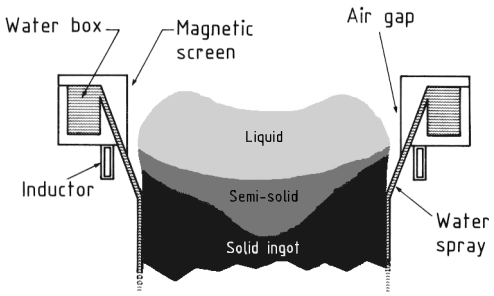

Figure 35. Electromagnetic casting (EMC) process

Although thick-sectioned ingot is used for the bulk of aluminum alloy production, there is an increasing amount of continuous cast strip 6 to 18 mm thick. Productivity is lower, but casting at high solidification rates, closer to final shape, allows elimination of many energy-consuming steps, such as scalping, homogenizing, and hot rolling. All the major thin-section casting methods use the moving mold concept. Two of the most popular are the twin-roll Hunter Engineering process (Fig. 36) and the Hazelett twin-belt method (Fig. 37). Another method of casting close to the final shape is the Properzi process. A bar with a trapezoid cross section of about 32 cm^2 is continuously cast and rolled to a 9-mm diameter rod. This method provides stock for conductor wire production. High energy costs will accelerate the trend from thick ingot casting to the less energy-intense, thin-section, continuous casting.

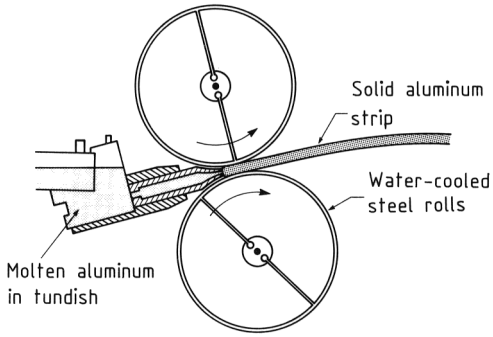

Figure 36. Twin-roll Hunter caster

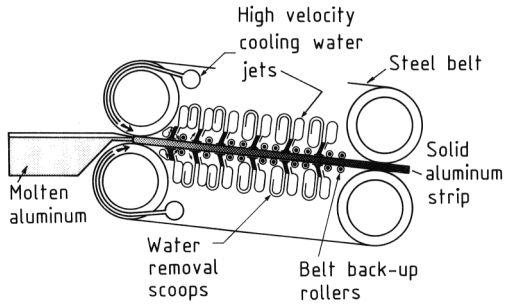

Figure 37. Hazelett twin-belt caster

3.3. Shaped Casting

Aluminum alloys are suitable for all common shaped casting methods: pressure die, permanent mold, sand casting, plaster, and investment [3, Chap. 2], [8, pp. 140 – 79]. The first three methods are the most commonly used. Pressure die casting (also called die casting) is the most widely used. Production by die casting, in terms of kilograms of metal cast, is more than twice that by permanent mold and sand casting combined. Typically, die castings are small in size, permanent mold castings are intermediate, and sand castings are largest. Section thickness usually increases in the same order. Die castings have the best surfaces, sand castings the worst.

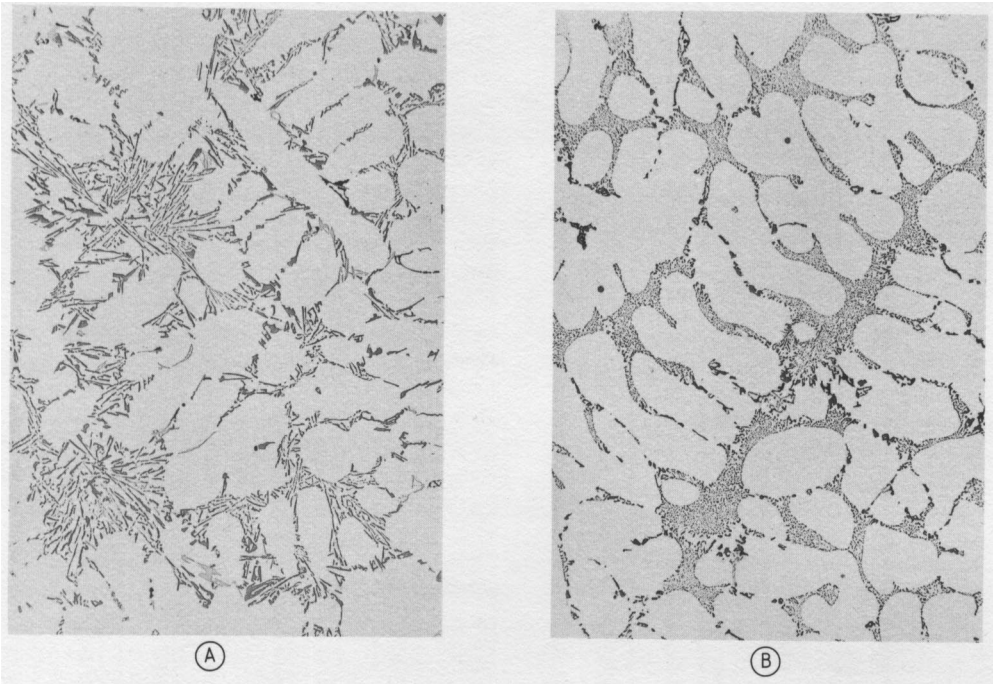

Figure 38. Effect of sodium modification on microstructure of sand-cast A357.0 alloy as polished (120x)
A) Not modified; B) Modified

By far the most common casting alloys are the *hypoeutectic Al – Si* type (Si in the range 4 – 12 %). Die castings freeze more rapidly than permanent mold castings, which freeze more rapidly than sand castings. In order to maximize the mechanical properties in permanent mold and sand castings, modifiers are added to the hypoeutectic Al – Si alloys. The most common modifiers are Na and Sr, small additions (0.005 – 0.1 %) having the effect of refining the usually platelike silicon in the eutectic and giving it a rodlike form. These structural changes, illustrated in Figure 38 A and B, help to increase both strength and ductility of the alloys. Antimony (0.05 – 0.1 %) may also be used to refine the eutectic structure; however, it does not modify the alloy and the benefits are fully achieved only after heat treatment. The *hypereutectic Al – Si* alloys, used primarily in die casting, are modified with phosphorus, because aluminum phosphide is a potent nucleus for silicon.

3.3.1. Die Casting

In the pressure die casting process the molten metal is injected at high velocity into a reusable metal mold. The combination of the thin sections and high pressure insures a rapid freezing rate (Fig. 8), but the accompanying contamination by gas entrapped as the metal enters the mold detracts from the potential advantages of the fine metallurgical structure. In addition, gas contamination prevents heat treatment to develop precipitation strengthening. However, the true potential of die casting will be realized only by adopting techniques such as evacuating the mold before injecting the molten metal to minimize gas pickup.

Because the die is made of steel, with no give, the alloys used must be resistant to hot cracking and reasonably fluid to fill the thin sections typical of die castings. In general, the alloys contain at least 0.6 % Fe, to reduce "soldering," i.e., sticking of the cast part to the die as it is being ejected.

3.3.2. Permanent Mold Casting

Like die casting, permanent mold casting (also called gravity die casting) employs a reusable metal mold. Because the mold is gravity fed, the pouring rate is relatively slow, but the metal mold produces rapid freezing. Generally, permanent mold castings are larger, with thicker sections, than die castings. They have good mechanical properties and are usually sound, provided the alloys exhibit good fluidity and resistance to hot cracking. Mechanical properties can be improved further by heat treatment.

3.3.3. Sand Casting

Sand casting is the least expensive process, but the mold, which is formed by ramming sand mixed with bonding agent around a pattern, is not reusable. Like permanent mold casting, the sand mold is gravity fed. The principal disadvantages are slow cooling rates (Fig. 8), poor surface finish, and the minimum wall thickness of 4 mm. For the same alloy the slow cooling rate results in lower mechanical properties than with permanent mold casting. Nevertheless, sand casting is widely used, particularly if the casting is large or if the number of items to be produced does not justify the expense of a metal mold. An advantage of sand casting is that quite complex shapes can be made with alloys that are not particularly resistant to hot cracking because sand, unlike metal, can give when the alloy contracts on freezing.

3.4. Mill Products

Mill products or wrought products are produced by working ingot into the desired shape. The most common mill product is sheet, but significant quantities of plate, foil, extrusions, drawn tube, forgings, impacts, rolled rod, bar, shapes, and wire are produced [5, Chap. IX].

3.4.1. Plate, Sheet, and Foil

Plate, sheet, and foil are collectively referred to as *flat rolled products.* By convention, plate is thicker than 6.30 mm, sheet is between 0.15 mm and 6.30 mm, and foil is less than 0.15 mm thick. Flat rolled products are produced in rolling mills by passing the metal between two cylindrical rolls where it is subjected to high compressive forces to decrease thickness and increase length. When flat rolled products are made from ingot, initial passes are always at elevated temperatures, but final passes may be hot or cold, depending on desired thickness and temper. When continuously cast strip is the starting material, initial passes may be either hot or cold. Intermediate heating may be used to promote homogeneity and to remove prior strain. After the final rolling pass, heat-treatable alloys are annealed or solution heat treated and aged to produce the desired temper: some alloy-tempers require strain after quenching or aging. Nonheat-treatable alloys are annealed, partially annealed, or stabilized to produce the desired temper.

The size of ingot used for making flat rolled products varies widely in thickness, width, and length, depending on such factors as alloy, product size, and equipment size. Commercial ingot varies from 200 × 1000 mm to 600 × 1900 mm in cross section and is up to 9 m in length. The maximum size of flat rolled products depends on alloy, temper, and thickness. Maximum size of flat plate may vary from 1300 to 5300 mm wide and 23 – 25.4 m long. The maximum weight per piece commercially available is about 10 000 kg. Maximum size of flat sheet is 900 – 3350 mm wide and 4.6 – 19.8 m long. Sheet is also supplied in coils up to 2500 mm wide and 12 000 kg in weight. Foil is supplied in coils up to 1900 mm wide.

3.4.2. Extrusions

Extrusions are produced in extrusion presses by forcing an ingot or billet to flow from a container through a die to produce an elongated shape. The diversity of shapes and sizes is great, ranging from the simplest round rod to cross sections of great complexity. Table 7 summarizes the shape and size capabilities of the aluminum extrusion industry. Extrusions may be given thermal and mechanical treatments to produce the desired temper, or they may be given further mechanical deformation by drawing to tube or wire. Extruded stock can also be fabricated further by forging and drawing.

Table 7. Capabilities of the extrusion industry

Alloy	almost all commercial wrought alloys
Section thickness	1 – 300 mm
Diameter of circumscribing circle (smallest circle that completely encloses the shape)	6 – 300 mm
Weight/length	0.07 – 300 kg/m
Straight length	33.5 m max.
Weight per piece	1200 t max.
Tubing and pipe diameter	600 – 840 mm

The capability of extrusion presses is described in terms of the maximum force that the press can exert on the ingot or billet. Capacity ranges from a few hundred to 13000 t, but most presses fall in the range 1500 – 2500 t [3, Chap. 3].

3.4.3. Forgings

Forging is metalworking performed in hammers and presses. The compressive force is applied either locally – the different parts of the forging are progressively worked on open dies (hand forgings) – or over the entire surface of the forging in closed dies (die forgings). Forging is done both hot and cold. Starting stock is ingot or partly fabricated forms such as extruded bar, rolled bar, and plate. Forging hammers range in capacity from 0.2 to 25 t and make parts weighing up to 100 kg. Mechanical presses range in capacity from 300 to 10000 t and make parts weighing up to 25 kg with plan areas up to 0.3 m^2. Hydraulic presses range in capacity from 500 to 75000 t and make parts weighing from a few kg to 1500 kg with plan areas up to 3 m^2 [3, Chap. 5].

3.4.4. Other Mill Products

Tubing and pipe are produced as extrusions that may be reduced further, in both size and wall thickness, by subsequent drawing through dies. An alternate method is forming sheet into a long hollow shape or a spiral and butt welding the edges to make a tube or pipe.

Rolled rod, bar, and some shapes are produced in rolling mills. For this purpose the rolls have a series of circumferential grooves to impart the desired shape.

Wire may be round, square, hexagonal, or octagonal in cross section. By convention it is less than 10 mm in diameter or between flats. It is produced by drawing rod or bar through a series of dies. Cable is made by stranding wire. ACSR (aluminum conductor, steel reinforced) is made by stranding aluminum wires around steel wire.

Impacts are parts made in a confining die from a metal slug by a rapid, single-stroke application of force through a punch. The metal flows back along the punch or through an opening in the punch or die.

3.5. Powder and Paste

Most aluminum powder is made by atomizing molten aluminum with compressed air; the particle size is controlled by the atomizing conditions. The particles are very rough but approximately equiaxial. Atomized powders are used in a number of chemical and metallurgical procedures.

Pigments are usually produced by ball milling atomized powder in the presence of lubricants to produce thin flakelike shapes having length and breadth 10 to 100 times greater than thickness. After milling, the pigment may be dried at moderately elevated temperature at reduced pressure, or solvent may be added to make a paste.

4. Commercial Alloys, Products, and Uses

A number of systems are used throughout the world to designate aluminum alloys and tempers. In the discussion that follows, the Aluminum Association designations are used for casting and wrought alloys.

4.1. Aluminum Association Designations for Alloys and Tempers

The Aluminum Association designation system for *casting alloys* is shown in Table 8 [6, pp. 8 – 9]. In the 1xx.x group the second and third digits indicate the minimum aluminum percentage above 99.00 %. In the 2xx.x through 9xx.x groups, the second and third digits have no compositional significance but serve only to identify different alloys within the group. The digit to

the right of the decimal point indicates product form: 0 for castings; 1 and 2 for ingot, 2 indicating a narrower composition range within the broader limits of the xxx.1 composition. A modification of the original alloy or impurity limits is indicated by a letter prefix. (I, O, Q, and X may not be used as the prefix.)

Table 8. Casting alloy designation system

Major alloying element	Designation
≧ 99.00 % Al or purer	1xx.x
Cu	2xx.x
Si + Cu and/or Mg	3xx.x
Si	4xx.x
Mg	5xx.x
Zn	7xx.x
Sn	8xx.x

The Aluminum Association system for *wrought alloys* is shown in Table 9 [6, pp. 7 – 8]. In the 1xxx group the last two digits indicate the minimum aluminum percentage above 99.00 %. For example, the alloy 1060 is ≧ 99.60 % Al. Only one alloy, 1100, has a specified alloying addition, 0.05 – 0.20 % Cu. In the 2xxx through 7xxx groups the last two digits have no composition significance but serve only to identify individual alloys in the series. The second digit indicates a modification of the original alloy. The prefix X indicates an experimental wrought alloy.

Table 9. Wrought alloy designation system

Major alloying element	Designation
≧99.00 % Al	1xxx
Cu	2xxx
Mn	3xxx
Si	4xxx
Mg	5xxx
Mg + Si	6xxx
Zn	7xxx
Other	8xxx

Temper designations are based on the sequence of basic treatments used to produce the temper (Table 10) [6, pp. 9 – 12]. Some basic treatments are applicable to both wrought and casting alloys. Variations of basic tempers for *non-heat-treatable* alloys are listed in Table 11. The third digit, when used, indicates a variation of a two-digit temper. It is used when the temper or mechanical properties, or both, differ from but are close to those for the two-digit H temper designation, or when some other product characteristic is affected significantly. Variations of the basic tempers are shown for *heat-treatable* alloys in Table 12. Additional digits, the first not zero, may be added to designation T1 through T10 to indicate a variation in treatment that significantly alters the characteristics of the product.

Table 10. Basic temper designations

Designation	Description
F	*as fabricated.* Applies to the products of shaping processes in which thermal conditions or strain hardening is not specially controlled. For wrought products there are no mechanical property limits.
O	*annealed.* Applies to wrought products annealed to the lowest strength temper, and to cast products annealed to improve ductility and dimensional stability. The O may be followed by a digit other than zero.
H	*strain hardened (wrought products only).* Applies to products strengthened by strain hardening, with or without supplementary thermal treatment to reduce strength. The H is always followed by two or more digits.
W	*solution heat treated.* An unstable temper, applies only to alloys aging spontaneously at room temperature after solution heat treatment. The designation is specific only when the period of natural aging is indicated, e.g., W 1/2 hr.
T	*thermally treated, stable tempers other than F, O, or H.* Applies to products thermally treated, with or without supplementary strain hardening, to produce stable tempers. The T is always followed by one or more digits.

Table 11. Non-heat-treatable alloy-tempers

Designation	Description
H 1	*strain hardened only.* Applies to products strain hardened to the desired strength without supplementary thermal treatment. The number following this designation indicates the degree of strain hardening.
H 2	*strain hardened and partially annealed.* Applies to products strain hardened more than the desired final amount and then reduced in strength to the desired level by partial annealing. The number following this designation indicates the degree of strain hardening remaining after partial annealing.
H 3	*strain hardened and stabilized.* Applies to products strain hardened and stabilized by low-temperature thermal treatment or heating during fabrication. This designation is applicable only to those alloys which, unless stabilized, gradually age soften at room temperature. The number following this designation indicates the degree of strain hardening remaining after stabilization.

Table 12. Heat-treatable alloy-tempers

Designation	Description
T 1 ∗	cooled from an elevated-temperature shaping process and naturally aged to a substantially stable condition.
T 2 ∗	cooled from an elevated-temperature shaping process, cold worked, and naturally aged to a substantially stable condition.
T 3	solution heat treated, cold worked, and naturally aged to a substantially stable condition.
T 4	solution heat treated and naturally aged to a substantially stable condition.
T 5 ∗	cooled from an elevated-temperature shaping process and artificially aged.
T 6	solution heat treated and artificially aged.
T 7	solution heat treated and overaged/stabilized. Applies to wrought products artificially aged after solution heat treatment and carried beyond maximum strength to control characteristics other than mechanical properties.
T 8	solution heat treated, cold worked, and artificially aged.
T 9	solution heat treated, artificially aged, and cold worked.
T 10 ∗	cooled from an elevated-temperature shaping process, cold worked, and artificially aged.

∗ Press quenched temper designations.

4.2. Alloys for Castings

Over 230 compositions for casting alloys have been registered with The Aluminum Association [13]. Nominal compositions of the most common alloys are shown in Table 13. Also shown are a few alloys, e.g., 201.0, 390.0, and 852.0, which are used in small volume but illustrate the range of commercial compositions. Some properties that are affected by compositions but not by temper are given in Table 13.

Casting alloys are ranked by suitability for casting, i.e., by their foundry characteristics, in Tables 14 and 15. The rankings are relative to other alloys in each casting category; i.e., die, permanent mold, or sand. Resistance to hot cracking is the ability to withstand contraction stresses on cooling. Die filling capability and fluidity indicate the ability of the molten metal to flow into all parts of the mold. Die castings are ranked on their tendency to solder to the die. Pressure tightness is an indication of the soundness of the cast part and the freedom from leakage of fluid under pressure. Solidification shrinkage reflects the alloy not being subject to shrinkage porosity.

Typical tensile properties of some die casting alloys are shown in Table 14 [30, ASTM B85].

Table 13. Composition and properties of casting alloys

Alloy	Product[a]	Nominal composition, wt %					Properties[b]			
		Cu	Mg	Mn	Si	Others	*m.r.*, °C	$\varrho^{25\,c}$, g/cm^3	$\bar{\alpha}$, 10^{-6}/K	*E*, GPa
201.0	S	4.6	0.35	0.35		0.70 Ag, 0.25 Ti	535 – 650	2.77	19.3	71.0
355.0	S or P	1.2	0.50		5.0		545 – 620	2.71	22.4	70.3
C355.0		1.2	0.50		5.0		545 – 620	2.71	22.4	70.3
356.0	S or P		0.32		7.0		555 – 615	2.69	21.5	72.4
A356.0	S or P		0.35		7.0		555 – 615	2.69	21.5	72.4
A357.0	S or P		0.60		7.0	0.15 Ti, 0.005 Be	555 – 615	2.68	21.6	71.7
360.0	D		0.50		9.5		555 – 595	2.63	21.0	71.0
380.0	D	3.5			8.5		540 – 595	2.74	21.0	71.1
390.0	D	4.5	0.60		17.0		505 – 650	2.73	18.0	81.2
413.0	D				12.0		575 – 582	2.66	20.4	
B 443.0	S or P				5.2		575 – 630	2.69	22.0	71.0
513.0	P		4.0			1.8 Zn	580 – 640	2.65	24.0	
518.0	D		8.0				535 – 620	2.57	24.1	
520.0	S		10.0				450 – 605	2.57	25.0	66.0
713.0	S or P	0.70	0.35			7.5 Zn	595 – 640	2.81	24.0	
852.0	S or P	2.0	0.75			6.25 Sn, 1.2 Ni	205 – 635	2.88	23.3	

[a] D for die casting, P for permanent mold casting, and S for sand casting;
[b] Melting range *m.r.*, density ϱ^{25}, average coefficient of thermal expansion $\bar{\alpha}$ over the range 20 – 100 °C, and modulus of elasticity *E* in tension;
[c] Actual density slightly lower because of porosity

Table 14. Characteristics and tensile properties of die casting alloys

Alloy	Foundry characteristics				Temper	Tensile properties *			Product characteristics						
	Resistance to hot cracking	Die filling capability	Soldering to die	Pressure tightness		T S, MPa	Y S, MPa	El, %	Corrosion resistance	Machining	Polishing	Electro-plating	Anodizing appear-ance	Chemical oxide coating	Strength at elevated tempera-ture
360.0	1	3	2	2	F	315	170	2.5	2	3	3	2	3	3	1
380.0	2	2	1	2	F	320	160	3.0	4	3	3	1	3	4	3
390.0	4	1	2	4	F	280	245	1.0	3	5	5	3	5	5	3
					T5	295	260	1.0	3	5	5	3	5	5	3
413.0	1	1	1	1	F	300	140	2.5	2	4	5	3	5	3	3
518.0	5	5	5	5	F	310	190	6.0	1	1	1	5	1	1	4

* Tensile strength (T S), yield strength (Y S), and elongation in 50 mm (El)

Table 15. Characteristics of permanent mold and sand casting alloys

Alloy	Foundry characteristics					Product characteristics								
	Resist-ance to hot cracking	Fluidity	Solidification shrinkage	Pressure tightness	Normally heattreated	Corrosion resistance	Machining	Polishing	Electro-plating	Anodizing appearance	Chemical oxide coating	Strength at elevated tempera-ture	Suitability for welding	Suitability for brazing
Permanent mold castings														
355.0, C355.0	12	2	1	yes		3	3	3	2	4	2	2	2	no
356.0, A356.0	12	1	1	yes		2	3	3	1	4	2	3	2	no
A357.0	12	1	1	yes		2	3	3	1	4	2	3	2	no
B 443.0	11	2	1	no		2	5	4	2	4	2	4	1	limited
513.0	45	4	5	no		1	1	1	4	1	1	3	5	no
852.0	55	5	5	aged only		3	1	1	5	4	5	*	5	no
Sand castings														
201.0	43	4	3	yes		4	1	1	1	2	2	1	4	no
C355.0	11	1	1	yes		3	3	3	1	4	2	2	2	no
A356.0	11	1	1	yes		2	4	5	2	4	2	3	2	no
A357.0	11	–	1	yes		2	3	4	–	4	–	2	1	–
B 443.0	11	1	1	no		2	5	5	2	5	2	4	1	limited
520.0	24	5	5	yes		1	1	1	4	1	1	*	5	no
713.0	54	4	3	aged only		2	1	1	2	2	3	5	4	yes
852.0	55	5	5	aged only		3	1	1	5	4	5	*	5	no

* Not suitable for elevated temperature

Table 16. Minimum tensile properties of permanent mold and sand casting alloys

Alloy	Temper	Permanent mold			Sand casting		
		Tensile strength, MPa	Yield strength, MPa	Elongation in 50 mm, %	Tensile strength, MPa	Yield strength, MPa	Elongation in 50 mm, %
201.0	T6	–	–	–	414	345	5.0
	T7	–	–	–	414	345	3.0
355.0	T6	255	–	1.5	221	138	2.0
	T51	186	–	–	172	124	–
	T71	234	186	–	207	152	–
	T62	290	–	–	–	–	–
	T7	248	–	–	–	–	–
C355.0	T61	276	207	3.0	–	–	–
	T6	–	–	–	248	172	2.5
356.0	F	145	–	3.0	131	–	2.0
	T6	228	152	3.0	207	138	3.0
	T7	–	–	–	214	–	–
	T51	172	–	–	159	110	–
	T71	172	–	3.0	172	124	3.0
A356.0	T61	255	179	5.0	–	–	–
	T6	–	–	–	234	166	3.5
A357.0	T61	310	248	3.0	–	–	–
B 443.0	F	145	41	2.5	117	41	3.0
513.0	F	152	83	2.5	–	–	–
520.0	T4	–	–	–	290	152	12.0
713.0	T5	221	152	4.0	207	152	3.0
852.0	T5	186	–	3.0	166	124	–

Table 16 lists tensile properties of some permanent mold and sand castings, minimum guaranteed values for separately cast test bars [30, ASTM B26, ASTM B108].

Castings are also ranked by product characteristics (Tables 14 and 15). Resistance to corrosion is based on general corrosion in a standard salt spray test. Machining is a composite rank based on ease of cutting, chip characteristics, quality of finish, and tool life. Polishing is based on ease of polishing and quality of finish. Electroplating rank indicates the ability to take and hold an electroplated coating. Anodizing rank is based on lightness of color, brightness, and uniformity of the clear anodized coating applied in sulfuric acid electrolyte. The chemical oxide coating is ranked on the combined resistance of coating and base alloy to corrosion. The elevated temperature strength is based on tensile and yield strength up to 260 °C. Die castings are not generally solution heat treated, but some permanent mold and sand castings are. Castings may be given a low-temperature aging or stabilization treatment without a separate solution heat treatment.

Some uses of aluminum alloy castings are listed in Table 17 (see next page).

4.3. Alloys for Wrought Products

Almost 400 domestic and international chemical compositions have been registered with the Aluminum Association for wrought alloys [14], and more than 1000 alloy-temper-product combinations have been described [6], [8].

Nominal compositions are listed in Table 18. Typical properties that depend on composition but not on temper are illustrated in Table 19. The melting range is for a thoroughly homogenized structure. The modulus of elasticity is an average of tension and compression. The modulus is about 2 % higher in compression than in tension.

Some characteristics and properties for selected alloy-tempers are given in Table 20. Corrosion ratings A through E are relative ratings, in order of decreasing merit, based on exposures to sodium chloride solution by intermittent spraying or immersion. Alloys with A and B ratings can be used in industrial and seacoast atmospheres without protection. Alloys with C, D, or E ratings generally should be protected, at least on faying surfaces. The general corrosion ratings of thick sections of 2011-T3, 2024-T3, and 7075-T6 would be lower than given in

Table 17. Typical uses of aluminum casting alloys

Alloy	Typical applications
201.0	applications requiring highest strength and moderate elongations (T6): structural castings, aerospace parts, truck and trailer castings. Applications requiring elevated-temperature strength (T6, T7): gasoline engine cylinder heads and pistons, rocker arms, connecting rods; turbine and supercharger impellers; missile fins. Applications requiring high strength and energy absorption capacity: gear housings, aircraft landing gear, ordnance
355.0	applications requiring good castability, high strength, and pressure tightness: pump bodies, liquid-cooled cylinder heads, crankcases, aircraft fittings, impellers
C355.0	stronger and more ductile than 355.0: aircraft, missile, and other structural parts
356.0	intricate castings requiring good strength and ductility: transmission cases, truck axle housings, truck wheels, cylinder blocks, cylinder heads, fan blades, marine hardware
A356.0	stronger and more ductile than 356.0: aircraft and missile parts, auto transmission cases
A357.0	aircraft and missile parts requiring weldability, strength, and toughness
360.0	applications requiring corrosion resistance better than 380.0
380.0	most widely used die-casting alloy: lawnmower housings, gear cases, cylinder heads for air-cooled engines, parts for auto and electrical industries
390.0	auto cylinder blocks, four-cycle air-cooled engines, air compressors, parts requiring abrasion resistance, low thermal expansion, or elevated-temperature strength
413.0	thin-walled intricate designs requiring excellent castability
B 443.0	general-purpose alloy: cooking utensils, waffle irons, architectural and marine fittings
513.0	castings requiring anodic finishes and corrosion resistance: marine fittings, auto trim, pipe couplings
518.0	escalator parts, conveyor components, aircraft and marine hardware
520.0	strong and ductile structural uses: aircraft fittings, trucks and bus frame components, brackets
713.0	general-purpose alloy: large castings requiring strength without heat treatment
852.0	bearings and bushings for heavy loads

Table 20. The corrosion ratings of 5182 and 5456 may be lowered by exposure to elevated temperature for long periods.

Stress-corrosion cracking ratings are based on service experience and laboratory tests of specimens exposed to the 3.5 % sodium chloride alternate immersion test.

A) No known instance of failure in service or laboratory tests
B) No known instance of failure in service, limited failures in laboratory tests on short transverse specimens
C) Service failures with sustained tension stress acting in short transverse direction relative to grain structure, limited failures in laboratory tests of long transverse specimens
D) Limited service failures with sustained longitudinal or long transverse stress

Ratings A through D for workability and A through E for machinability are relative ratings in order of decreasing merit. Ratings A through D for weldability, brazeability, and solderability also are relative ratings. For example:

A) Generally weldable by all commercial procedures and methods
B) Weldable with special techniques or for specific applications that justify preliminary trials or testing to develop welding procedure and weld performance
C) Limited weldability because of crack sensitivity or loss in resistance to corrosion and mechanical properties
D) No commonly used welding methods have been developed

A wide variety of alloy-temper-product combinations are available as indicated by the partial listing in Table 21. Some typical uses for wrought aluminum alloy products are in Table 18.

Table 18. Nominal compositions and uses of wrought aluminum alloys

Alloy	Percentage of alloying elements (balance is Al + normal impurities)								Uses
	Si	Cu	Mn	Mg	Cr	Ni	Zn	Others	
1060			≧ 99.60 % Al						chemical equipment, railroad tank cars
1100		0.12	≧ 99.00 % Al						sheet metal work, spun holloware, fin stock
1199			≧ 99.99 % Al						electrolytic capacitor foil, chemical equipment, railroad tank cars
1350			≧ 99.50 % Al						electrical conductors
2011		5.5						0.40 Pb, 0.40 Bi	screw machine products
2014	0.8	4.4	0.8	0.5					truck frames, aircraft structures
2024		4.4	0.6	1.5					truck wheels, screw-machine products, aircraft structures
2124		4.4	0.6	1.5					military supersonic aircraft
2195		4.0		0.4				1.0 Li, 0.11 Zr	space applications, cryogenic tankage
2219		6.3	0.30					0.06 Ti, 0.10 V, s0.18 Zr	structural uses at high temperatures (to 315 °C), high-strength weldments
2618	0.18	2.3		1.6		1.0		1.1 Fe, 0.07 Ti	aircraft engines
3003		0.12	1.2						cooking utensils, chemical equipment, pressure vessels, sheet metal work, builders' hardware, storage tanks
3104		0.15	1.0	1.2					can bodies, sheet metal work, storage tanks
3105			0.6	0.5					residential siding, mobile homes, rain-carrying goods, sheet metal work
4032	12.2	0.9		1.0		0.9			engine pistons
4043	5.2								welding electrode
5005				0.8					appliances, utensils, architectural, electrical conductors
5050				1.4					builders' hardware, refrigerator trim, coiled tubes
5052				2.5	0.25				sheet metal work, hydraulic tube, appliances
5056			0.12	5.0	0.12				cable sheathing, rivets for magnesium, screen wire, zippers
5083			0.7	4.4	0.15				unfired welded pressure vessels, marine, cryogenics, drilling rigs, transportation equipment, missile components, ballistic armor
5086			0.45	4.0	0.15				unfired welded pressure vessels, marine, cryogenics, drilling rigs, transportation equipment, missile components
5154				3.5	0.25				welded structures, storage tanks, pressure vessels, saltwater service
5182			0.35	4.5					can ends, automobile body sheet
5252				2.5					automotive and appliance trim
5254				3.5	0.25				hydrogen peroxide and chemical storage vessels
5356			0.12	5.0	0.12			0.13 Ti	welding electrode
5454			0.8	2.7	0.12				welded structures, pressure vessels, marine service, truck panels
5456			0.8	5.1	0.12				high-strength welded structures, storage tanks, pressure vessels, marine applications
5657				0.8					anodized auto and appliance trim
6005	0.8			0.50					heavy-duty structures requiring good corrosion resistance, truck and marine, railroad cars, furniture, pipelines
6111	0.9	0.7	0.3	0.9					automobile body panels
6022	1.3		0.08	0.6					automobile body panels

Table 18. (Continued)

Alloy	Percentage of alloying elements (balance is Al + normal impurities)								Uses
	Si	Cu	Mn	Mg	Cr	Ni	Zn	Others	
6061	0.6	0.28		1.0	0.20				heavy-duty structures requiring good corrosion resistance, railroad cars, furniture, pipelines, truck and auto wheels, tooling plate
6063	0.40			0.7					pipe railing, furniture, architectural extrusions
6066	1.4	1.0	0.8	1.1					forgings and extrusions for welded structures
6070	1.4	0.28	0.7	0.8					heavy-duty welded structures, pipelines
6101	0.50			0.6					high-strength bus conductors
6151	0.9			0.6	0.25				moderate-strength, intricate forgings for machine and auto parts
6201	0.7			0.8					high-strength electric conductor wire
6262	0.6	0.28		1.0	0.09			0.6 Pb, 0.6 Bi	screw machine products
6351	1.0		0.6	0.6					heavy-duty structures requiring good corrosion resistance, truck and tractor extrusions
6463	0.40			0.7					extruded architectural and trim sections requiring bright finishing
7005			0.45	1.4	0.13		4.5	0.04 Ti, 0.14 Zr	heavy-duty structures requiring good corrosion resistance, trucks, trailers, dump bodies, sports equipment
7049		1.5		2.5	0.16		7.6		aircraft and other structures
7050		2.3		2.25			6.2	0.12 Zr	aircraft and other structures from plate, forgings, or extrusion
7072							1.0		finstock, cladding alloy
7075		1.6		2.5	0.23		5.6		aircraft and other structures
7175		1.6		2.5	0.23		5.6		aircraft and other structures, forgings
7178		2.0		2.8	0.23		6.8		aircraft and other structures
7475		1.5		2.3	0.22		5.7		aircraft and other structures requiring high fracture toughness

Table 19. Properties ∗ of wrought aluminum alloys

Alloy	*m.r.,* °C	ϱ^{25}, g/cm^3	$\bar{\alpha}_1$, 10^{-6}/K	*E,* GPa
1100	640 – 655	2.71	23.6	69
2011	540 – 640	2.83	22.9	70
2024	500 – 635	2.78	23.2	73
2036	555 – 650	2.75	23.4	71
2219	545 – 645	2.84	22.3	73
3003	640 – 655	2.73	23.2	69
3004	630 – 655	2.72	23.9	69
4032	530 – 570	2.68	19.4	79
4043	575 – 630	2.68	22.0	–
5052	605 – 650	2.68	23.8	70
5182	575 – 640	2.65	24.1	71
5456	570 – 640	2.66	23.9	71
6010	585 – 650	2.70	23.2	69
6061	580 – 650	2.70	23.6	69
6066	560 – 645	2.72	23.2	69
7005	–	2.78	–	72
7050	488 – 635	2.82	23.0	72
7075	532 – 635	2.80	23.6	72
7090	550 – 635	2.85	23.8	75

∗ Melting range *m.r.,* density ϱ^{25}, average coefficient of thermal expansion $\bar{\alpha}_1$ over the range 20 – 100 °C, and average modulus of elasticity *E*

Table 20. Typical tensile properties, electrical resistivity, and product characteristics of some wrought aluminum alloys

Alloy-temper	Tensile properties				Electrical resistivity, $10^{-9}\ \Omega$ m	Corrosion resistance		Workability (cold)	Machina-bility	Weldability			Brazeability	Solderability
	Tensile strength, MPa	Yield strength, MPa	Elongation, % *			General	SCC			Gas	Arc	Resistance spot & seam		
			1.6 mm thick	12.5 mm dia.										
1100−0	90	34	35	–	29	A	A	A	E	A	A	B	A	A
1100-H18	165	150	5	–	30	A	A	C	D	A	A	A	A	A
2011-T3	380	295	–	15	44	D	D	C	A	D	D	D	D	C
2011-T8	405	310	–	15	38	D	B	D	A	D	D	D	D	C
2024−0	185	76	20	20	34	–	–	–	D	D	D	D	D	C
2024-T3	485	345	18	–	57	D	C	C	B	C	B	B	D	C
2024-T861	517	490	5	–	45	D	B	D	B	D	C	B	D	C
2036-T4	340	195	24	–	42	C	–	B	C	–	B	B	D	–
2219−0	170	75	18	–	39	–	–	–	–	D	A	B	D	–
2219-T87	475	395	10	–	57	D	B	D	B	A	A	A	D	–
3003−0	110	42	30	40	34	A	A	A	E	A	A	B	A	A
3003-H18	200	185	4	10	43	A	A	C	D	A	A	A	A	A
3004−0	180	69	20	25	41	A	A	A	D	B	A	B	B	B
3004-H38	285	250	5	6	41	A	A	C	C	B	A	A	B	B
4032-T6	380	315	–	9	49	C	B	–	B	D	B	C	D	–
4043−0	145	69	22	–	41	B	A	–	C	–	–	–	–	–
4043-H18	285	270	1	–	41	B	A	–	C	–	–	–	–	–
5052−0	195	90	25	27	49	A	A	A	D	A	A	B	C	D
5052-H38	290	255	7	–	49	A	A	C	C	A	A	A	C	D
5182−0	275	140	25	–	–	A	A **	A	D	C	A	B	D	D
5182-H19	420	395	4	–	–	A	A **	D	B	C	A	A	D	D
5456−0	310	160	–	22	59	A **	B **	B	D	C	A	B	D	–
5456-H321	350	255	–	14	59	A **	B **	C	D	C	A	A	D	–
6010-T4	290	170	24	–	42	A	A	B	C	A	A	A	A	B
6061−0	125	55	25	30	37	B	A	A	D	A	A	B	A	B
6061-T451	240	145	22	25	43	B	B	B	C	A	A	A	A	B
6061-T651	310	275	12	17	40	B	A	C	C	A	A	A	A	B
6066−0	150	83	–	18	43	C	A	B	D	D	B	B	D	–
6066-T6	395	360	–	12	47	C	B	C	B	D	B	B	D	–
7005-T53	393	345	15	–	45	B	B	C	A	B	B	B	B	B
7050-T7451	510	455	–	11	43	C	B	D	B	D	C	B	D	D
7050-T7651	530	475	–	11	44	C	B	D	B	D	C	B	D	D
7075-T6	570	505	11	11	52	C	C	D	B	D	C	B	D	D
7075-T73	505	435	13	–	43	C	B	D	B	D	C	B	D	D

* Elongation in 50 mm, %, for specimens 1.6 mm thick or 12.5 mm in diameter;
** May be different for material held at elevated temperature for an extended period

Table 21. Examples of commercially available alloy-temper-product combinations

Alloy	Sheet	Plate	Tube		Pipe	Structural shapes	Extruded wire, rod, bar, and shapes	Rolled or cold finished			Rivets	Forgings and forging stock	Foil	Fin stock	Can sheet
			Drawn	Extruded				Rod	Bar	Wire					
1100	0 H12 H14 H16 H18	0 H12 H14 H112	0 H12 H14 H16 H18 H113	0 H112			0 H112	0 H112 H14 F	0 H112 F	0 H112 H12 H14 H16 H18	0 H14	H112 F	0 H19	0 H14 H18 H19 H25 H111 H113 H211	
2024	0 T3 T361 T4 T72 T81 T861	0 T351 T361 T851 T861	0 T3	0 T3510 T3511 T3 T81 T8510 T8511			0 T3 T3510 T3511 T81 T8510 T8511	0 H13 T4 T351 T6 T851	0 T351 T4 T6 T851	0 H13 T36 T4 T6	T4				
3003	0 H12 H14 H16 H18	0 H12 H14 H112	0 H12 H14 H16 H18 H25 H113	0 H112	H18 H112		0 H112	0 H112 F H14	0 H112 F	0 H112 H12 H14 H16 H18	0 H14	H112 F	0 H19	0 H14 H18 H19 H25 H111 H113 H211	
3004	0 H32 H34 H36 H38	0 H32 H34 H112	0 H34 H36 H38	0											H19
5052	0 H32 H34	0 H32 H34	0 H32 H34					0 F H32	0 F	0 H32 H34	0 H32		0 H19		

Table 21. (Continued)

Alloy	Sheet	Plate	Tube		Pipe	Structural shapes	Extruded wire, rod, bar, and shapes	Rolled or cold finished			Rivets	Forgings and forging stock	Foil	Fin stock	Can sheet
			Drawn	Extruded				Rod	Bar	Wire					
	H36	H112	H36							H36					
	H38		H38							H38					
5182	0														H19
5456	0	0										H112			
	H116	H112										F			
	H321	H116													
		H321													
6061	0	0	0	0	T6	T6	0	0	0	0	T6	F			
	T4	T451	T4	T1			T1	H13	T4	H13		T6			
	T6	T651	T6	T4			T4	T4	T451	T4		T652			
				T4510			T4510	T451	T6	T6					
				T4511			T4511	T6	T651	T89					
				T51			T51	T651		T93					
				T6			T6			T913					
				T6510			T6510			T94					
				T6511			T6511								
6066			0	0			0					F			
			T4	T4			T4					T6			
			T6	T4510			T4510								
				T4511			T4511								
				T6			T6								
				T6510			T6510								
				T6511			T6511								
7075	0	0	0	0			0	0	0	0	T6	F			
	T6	T651	T6	T6			T6	H13	T6	H13	T73	T6			
	T73	T7351	T73	T6510			T6510	T6	T651	T6		T652			
	T76	T7651		T6511			T6511	T651	T73	T73		T73			
				T73			T73	T73	T7351			T7352			
				T73510			T73510	T7351							
				T73511			T73511								
							T76								
							T76510								
							T76511								

Table 22. Designations of approximately equivalent casting alloys ∗ [31]

Aluminum Association	Australia	Canada	Federal Republic of Germany	Italy	Japan	United Kingdom
355	AP 309; H49−13	.SC 51N		3600; G-AS5CG; G-AlSi5CuMg	AC4D	L78; LM16
C355		.SC 51P				
356	AP 601 ∗∗; H49−18	.SG 70N	3.2364	3599; G-AS7GM; G-AlSi7MgMn	AC4C	LM25
A356	BP 601; CP 601; H49−18	.SG 70 P	3.2374; G-AlSi7Mg; GK-AlSi7Mg			5028; L99
360	AP 605; BP 605; H49−23			3051; 5074; GD-AS9GF; GD-AlSi9MgFe		
380	AP 313	.SC 84P	G-AlSi8Cu3; GD-AlSi8Cu3; GK-AlSi8Cu3			
A380	BP 313; H49−16	.SC 84N		3601; G-AS8,5C; G-AlSi8,5Cu		LM24
413	AP 401 ∗∗	.S 12P	GD-AlSi12			LM20
518	AP 503; H49−19	.G 8		5080; GD-AG7,5F; GD-AlMg7,5Fe		
520	AP 505; H49−8	.G 10		3056; G-AG10; G-AlMg10	AC78	L53; LM10

∗ The Australian and Canadian designations refer to the Aluminum Association .2 alloys, whereas the German, Japanese, and British refer to the .0 alloys. The Italian designations generally refer to both .0 and .2 alloys. For example, AP 309 is equivalent to 355.2 and L78 is equivalent to 355.0;
∗∗ AS for .1 alloys, e.g., AS 601 for 356.1

4.4. Comparison of Designation Systems

The Aluminum Association's wrought alloy designation system is in common use worldwide. International wrought alloy registrations are made through The Aluminum Association's International Accord. Table 22 (see above) compares the designations of approximately equivalent casting alloys, and Table 23 (see next page) does the same for wrought alloys. Much more extensive lists are given in [31]. Inspection of Table 23 shows that the designations of wrought alloys are similar, or the same, in the United States, Australia, Canada, Italy, and Japan. The systems in Germany, Austria, the Netherlands, and Switzerland are similar. National organizations issuing specifications and designations are listed in Table 24 for selected countries [31, Appendix F].

Table 24. Organizations that issue specifications

Country	Organization, city
Australia	Standards Association of Australia, North Sydney
Austria	Österreichisches Normungsinstitut, Wien
Canada	Canadian Standards Association, Etobicoke, Ontario
China, People's Republic of	China Association for Standardization, Peking
France	Association Française de Normalisation, Paris
Germany	Deutsches Institut für Normung, Berlin
International	International Organization for Standardization, Genève
Italy	Ente Nazionale Italiano di Unificazione, Milano
Japan	Japanese Industrial Standards Committee, Tokyo
Netherlands	Nederlands Normalisatie-instituut, Rijswijk
Spain	Instituto Nacional de Racionalizacion del Trabajo, Madrid
Switzerland	Verein Schweizerischer Maschinen Industrieller, Zürich
Russia	Russian Committee for Standards, Moskva
United Kingdom	British Standards Institution, London
United States	The Aluminum Association, Washington

Table 23. Designations of approximately equivalent wrought alloys [32]

Aluminum Association	Australia	Austria	Canada	France	Germany	Italy	Japan	Netherlands	Spain	Switzerland	United Kingdom
1100	A1100		1100	A45			1100		L-3002 UNE 38−119		
1199	A1199			A99							1
2011	A2011		2011	A-U5PbBi	AlCuBiPb	2011	2011		L-3192 UNE 38−322	AlCu6BiPb	FC1
2014	B2014		2014	A-U4SG	AlCuSiMn	2014	2014	AlCu4Mg	L-3130 UNE 38−313	AlCu4SiMn	H15
2024	A2024	AlCuMg2	2024	A-U4G1	AlCuMg2	2024	2024	AlCu4Mg1	L-3140 UNE 38−314	AlCu4Mg1,5	2L97 L98 L109 ∗ L110 ∗
2618				A-U2GN			2N01		L-3171 UNE 38−320		H16
3003	A3003		3003	A-M1	AlMnCu	3003	3003 3203	AlMn1	L-3810 UNE 38−381	AlMn	
3004	A3004			A-M1G	AlMn1Mg1	3004			L-3820 UNE 38−382		
3105					AlMn0,5Mg0,5						N31
4043	B4043		4043	A-5S							N21
5005	A5005			A-G0,6		5005	5005	AlMg1	L-3350 UNE 38−335	AlMg1	N41
5050				A-G1,5					L-3380 UNE 38−338	AlMg1,5	
5052	A5052		5052	A-G2,5C	AlMg2,5	5052	5052		L-3360 UNE 38−336		
5056	A5056	AlMg5	5056				5056				
5083	B5083		5083	A-G4,5MC	AlMg4,5Mn	5083	5083	AlMg4,5Mn	L-3321 UNE 38−340	AlMg4,5Mn	N8
5086	A5086		5086	A-G4MC	AlMg4Mn				L-3322 UNE 38−341	AlMg4	
5356	B5356		5356	A-5G							
5454	A5454		5454	A-G2,5MC	AlMg2,7Mn	5454			L-3391 UNE 38−345	AlMg2,7Mn	N51
6061	A6061		6061	A-GSUC		6061	6061		L-3420 UNE 38−342		H20
6063	B6063	AlMgSi0,5	6063				6063	AlMgSi	L-3441 UNE 38−337		H9
6101	B6101	E-AlMgSi					6101		L-3431 UNE 38−343		
7072	D7072								L-3721 UNE 38−372		
7075	A7075	AlZnMgCu1,5	7075	A-Z5GU	AlZnMgCu1,5	7075	7075		L-3710 UNE 38−371	AlZn6MgCu1,5	2L88 ∗ 2L95, 2L96

∗ Core metal of clad alloys

4.5. Alloys for Powder Metallurgy Products

Powder metallurgy is used to achieve unique combinations of properties not obtainable by other techniques. Two categories of powder-metallurgy products are made: high-performance wrought products, and sintered parts.

Sintered parts are made from mixtures of unalloyed powders or mixtures of unalloyed and alloyed powders. Lubricant to prevent aluminum adhering to compacting tools may be mixed with the powder or applied to the tools. The powders are usually compacted cold into the desired shape in automatic powder metallurgy presses. The part then is sintered in an inert atmosphere at a temperature that causes limited melting to promote interparticle bonding as well as alloying of the components. The part may be used as sintered and naturally aged (T1), or it may be heat treated and aged to increase strength (T6). Sintered parts retain some porosity, which is controlled by the compacting and sintering conditions. Table 25 summarizes composition and properties of powder-metallurgy alloys [34].

Table 25. Composition and properties of aluminum powder-metallurgy alloys

Alloy	Nominal composition, wt %						Temper	Density, g/cm^3	Tensile strength, MPa	Yield strength, MPa	Elongation, % in 25 mm	Comments
	Si	Cu	Mg	Zn	Other	Lubrication						
Alcoa 601AB	0.6	0.25	1.0			1.5	T1-sintered	2.55	145	95	6	good strength, ductility, corrosion resistance
							T6	2.55	240	230	2	
Alcoa 201AB	0.8	4.4	0.5			1.5	T1-sintered	2.64	210	180	3	high strength
							T6	2.64	330	330	2	
Alcoa 202AB		4.0				1.5	T1-sintered	2.56	160	75	10	good ductility, suitable for cold- formed parts
							T6	2.56	230	145	7	
Alcoa 602AB	0.4		0.6			1.5	T1-sintered	2.55	120	60	9	good electrical conductivity, ductility, finishing characteristics
							T6	2.55	180	170	2	
Alcan 67		0.5					T1-sintered	2.52	105	55	12	good electrical conducitivty, ductility
Alcan 76		1.6	2.5	5.6	0.20 Cr		Sintered	2.51	205	150	3	high strength
							T6	2.51	310	275	2	
Alcan 91					Tribaloy 26.3		T1-sintered	3.05	95		2	wear resistance
							T6	3.05	105		1	
Alcan 4090	1.0	1.0					T1-loose sintered	1.35	0.7 (breaking strength)			high-porosity parts for controlling contamination, pressure, sound, catalytic reactions

5. Manufacture of Articles from Aluminum Products

Aluminum alloy mill products and castings are the starting materials for the manufacture of a great many articles, ranging from agricultural equipment to zippers. The value added by subsequent operations varies with the article, ranging from a small fraction to many times the cost of the mill product or casting. The major manufacturing operations are shaping, joining, and finishing.

5.1. Shaping

The shapes of aluminum products are changed by deformation and metal removal. *Deformation* involves stretching, bending, upsetting, and extruding and changes the metallurgical structure. Little or no metal is lost. On the other hand, *metal removal* usually is accompanied by substantial loss of metal, but there is no change in

change in metallurgical structure. Removal includes cutting, grinding, and chemical dissolution [3, Chaps. 6 – 8],[8, pp. 180 – 90].

The choice of method is determined by many factors, including alloy-temper, the form of the mill product, and the form and characteristics of the final product. Table 20 classifies a number of alloy-tempers according to workability; information on additional alloy-tempers is in [3, Chaps. 6 – 8], [8, pp. 46 – 9]. The machinability ratings in Tables 14 and 20 are based on the nature of the machine chips (small, or long and stringy), surface finish, and tool wear. These ratings can serve only as a preliminary guide because the type and shape of machining tools, cutting speeds, lubrication, and other factors also must be considered [3, Chap. 8], [8, pp. 187 – 90].

5.2. Joining

Aluminum products are joined by many commercial methods: mechanical, metallurgical, and adhesive [3, Chaps. 11 – 15], [12], [8, pp. 191 – 203].

5.2.1. Mechanical Joining

Mechanical joining is used to join aluminum articles to other aluminum articles and to other materials. Aluminum alloy threaded fasteners, rivets, bolts, and nails are often used. Also used often are lock seams, crimping, and other integral joints that do not use separate fasteners. Generally, the condition of the aluminum surface does not have an effect on the initial strength of the joint, but surface preparation and post-joining protection have major effects on leak tightness and durability.

The most ubiquitous examples of mechanical joints are probably in two-piece beverage cans. Cross sections through two parts of an end are shown in Figure 39. The end is joined to the can body with a double-lock seam (Fig. 39 A). The tab is joined to the end by an integral rivet (Fig. 39 B). The rivet is first formed as a small bubble in the can end, and the tab is placed on the end so that the bubble goes through a hole in the tab. Finally, the bubble is flattened to fasten the tab to the end.

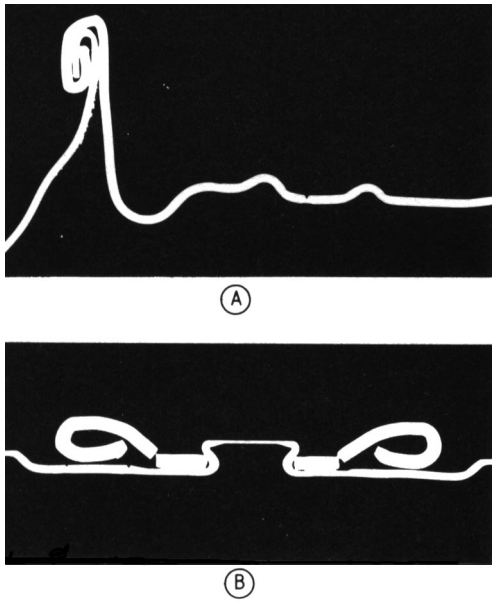

Figure 39. Mechanical joints in beverage can (≈ 2.5X)
A) Double seam joining end to body; B) Integral rivet joining tab to end

5.2.2. Metallurgical Joining

The metallurgical joining processes are welding [12, pp. 373 – 99, 534 – 44, 609 – 768], brazing [12, pp. 1022 – 32], and soldering [3, pp. 525 – 48]. These methods all involve establishing metal-to-metal bonds by melting. Welding is achieved also by solid-state processes involving deformation to break up oxide films and insure contact between faying surfaces.

Aluminum alloys vary widely in their suitability for joining by metallurgical techniques. Some alloy-tempers are categorized in Tables 15 and 20; information on other alloy-tempers is in [8, pp. 46 – 49, 144]. The rankings are based on how easily defect-free joints are made and how closely the joints approach the properties, corrosion resistance, and finishing characteristics of the parent metal. Aluminum alloy specialty products made for joining applications include filler wire for welding (Table 26) and brazing (Table 27) and brazing sheet (Table 28). Surface pretreatment and protection of the surface during metallurgical joining are important factors affecting the quality of the finished joint.

Table 26. Filler alloys for inert gas welding of aluminum alloys

Alloy	Nominal composition, wt%						
	Si	Cu	Mn	Mg	Cr	Ti	Other
1100		0.13					
2319		6.2	0.30			0.15	0.10 V 0.18 Zr
4043	5.3						
4047	12.0						
4145	10.0	4.0					
5183			0.75	4.8	0.15		
5356			0.13	5.0	0.13	0.13	
5554			0.75	2.7	0.13	0.13	
5556			0.75	5.1	0.13	0.13	
5654				3.5	0.25	0.10	
242.0		4.0		1.6			2.0 Ni
295.0	1.1	4.5					
355.0	5.0	1.25		0.50			
356.0	7.0			0.30			

Table 27. Brazing filler alloys

AWS∗	Nominal composition, wt%						Temperature, °C	
classification	Si	Cu	Mg	Cr	Ti	Bi	Solidus	Liquidus
BAlSi-2							575	615
	7.5							
BAlSi-3	10.0		0.15				520	585
		4.0						
BAlSi-4	12.0						575	580
BAlSi-5	10.0				0.20		575	590
BAlSi-6 ∗∗			2.5				560	605
	7.5							
BAlSi-7 ∗∗	10.0		1.5				560	595
BAlSi-8 ∗∗	12.0		1.5				560	580
BAlSi-9 ∗∗	12.0		0.3				560	580
BAlSi-10 ∗∗	11.0		2.5				560	580
BAlSi-11 ∗∗	10.0		1.5			0.11	560	595

∗ American Welding Society;
∗∗ Solidus and liquidus temperatures in a vacuum differ from those in the table

Table 28. Aluminum brazing sheet

Brazing sheet no.	Sides clad	Core alloy	Cladding alloy
7	1	3003	4004
8	2	3003	4004
11	1	3003	4343
12	2	3003	4343
13	1	6951	4004
14	2	6951	4004
21	1	6951	4343
22	2	6951	4343
23	1	6951	4045
24	2	6951	4045

Nominal composition, wt %

Alloy	Si	Cu	Mn	Mg
3003		0.12	0.12	
4004	9.8			1.5
4045	10.0			
4343	7.5			
6951	0.35	0.28		0.6

5.2.3. Adhesive Bonding

Adhesive bonding is widely used to join aluminum to itself and to other materials. Applications range from aircraft structures to packaging. Suitability for adhesive bonding is not usually a factor in alloy selection; the alloy is selected on its other characteristics and properties. Surface preparation of the adherends strongly affects initial strength and durability. Several pretreatment methods are in use. Adhesive bonding is sometimes combined with other joining methods such as spot welding (weldbonding) and mechanical joining [35].

5.3. Cleaning and Finishing

An outstanding characteristic of aluminum alloys is the ability to accept decorative and protective finishes. Extensive research and development has resulted in many different types of finishes and many processes to produce finishes [3, Chaps. 16 – 22], [11, pp. 571 – 610]. Because cleaning and pretreating the surface prior to applying the final coating are so important, the manufacture of a commercial article usually involves a sequence of cleaning and finishing steps. Generally, the procedures do one or more of the following:

1) Remove dirt, lubricant, oxide, and other debris left by prior operations
2) Remove some of the base metal
3) Deform the surface of the base metal
4) Convert aluminum to aluminum oxide
5) Coat with other materials

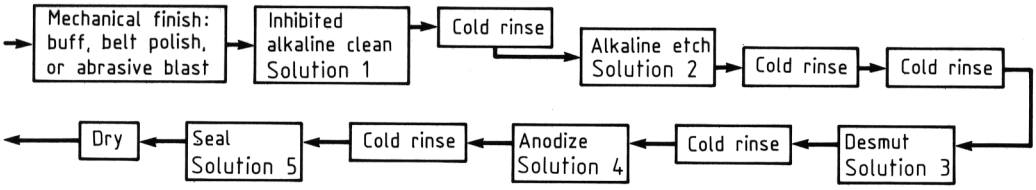

Figure 40. Operations sequence in sulfuric acid anodizing of architectural parts

Solution no.	Type of solution	Composition	Operating temperature, °C	Cycle time, min
1	alkaline cleaning	alkali, inhibited	60 – 70	2 – 4
2	alkaline etching	NaOH, 5 wt %	50 – 71	2 – 20
3	desmutting	HNO_3, 25 – 35 vol %	room	2
4	anodizing	H_2SO_4, 15 wt %	21 – 25	5 – 60
5	sealing	water, pH 5.5 – 6.5	100	5 – 20

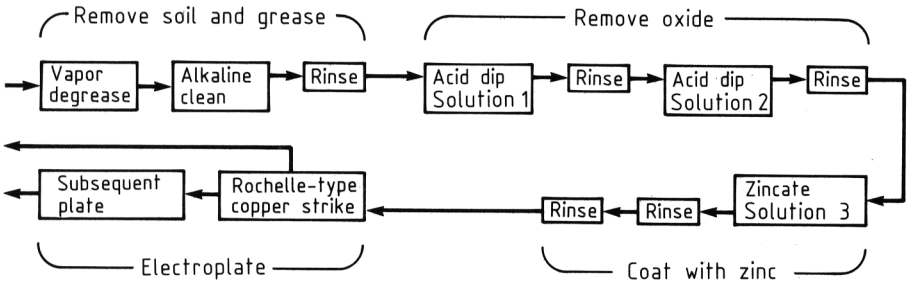

Figure 41. Preplating surface preparation sequence suitable for alloys 1100, 3003, 3004, 2011, 2017, 2024, 5052, 6061, 208, 295, 319, and 355

Solution	Type of solution	Composition	Concentration	Operating temperature, °C	Cycle time, s
1	acid dip	H_2SO_4	15 vol %	≥ 85	120 – 300
2	acid dip	HNO_3	50 vol %	room	15
3	zincating	NaOH	525 g/L	16 – 27	30 – 60
		ZnO	100 g/L		

The multistep nature of finishing processes is illustrated in Figure 40 by showing a sequence for anodizing architectural parts [11, p. 590]. One sequence for preparing alloys for electroplating is shown in Figure 41 [11, p. 604]. Although many methods are used, they can be divided into nine categories:

1) mechanical
2) chemical cleaning
3) etching
4) brightening
5) chemical conversion coating
6) anodizing
7) metal coatings (immersion and electroless coating, electroplating)
8) painting and organic coatings
9) enameling

5.3.1. Mechanical Cleaning and Finishing

Mechanical methods may be used to remove asperities, smoothing the surface, or to produce a uniformly rough surface. The method can remove metal (e.g., grinding, polishing, buffing, blasting) or deform the surface (e.g., embossing, hammering, shot peening). Surface appearance

ranges from mirrorlike to hammered, depending on the method and alloy.

5.3.2. Chemical Cleaning and Finishing

Chemical cleaning and finishing treatments are used to:

1) Remove oxides, lubricants, and other matter left on the surfaces from prior operations or storage
2) Smooth or roughen surfaces to impart specular or mat appearances
3) Develop decorative or protective coatings
4) Deposit metallic films

The surfaces that result from these treatments are usually a pretreatment for other finishes, but they are sometimes the final surface of the article.

The least aggressive cleaners are the *organic solvents.* They are used to remove oils and greases but do not remove oxides or metal. *Alkaline cleaners* range from mild to aggressive. They are commonly used to dissolve or remove dirt and can be formulated to etch. Etching is controlled by inhibitors, such as silicates. *Acid cleaners* are used to remove oxide films, welding and brazing fluxes, and smut from other cleaning operations. Smut is a term applied to the dark deposit formed when alloys containing high concentrations of silicon, copper, iron, and manganese are etched in alkaline solutions.

5.3.3. Chemical Etching

Chemical etching removes metal to produce mat finishes and to reduce or remove surface scratches, extrusion die lines, and other small defects. Etching promotes surface uniformity, thus simplifying subsequent finishing treatments. *Alkaline etching* works well with most aluminum alloys, provided etching is followed by desmutting. However, *acid etching* is preferred for 99.99+ aluminum and alloys containing large amounts of magnesium or silicon.

The most common alkaline etching solution is NaOH plus a sequestering agent to retard the formation of hydrated alumina. Acid etching solutions are usually nitric – hydrofluoric and sulfuric – chromic acids. The high-silicon alloys are also etched in hydrofluoric acid solutions.

5.3.4. Chemical and Electrolytic Brightening

Chemical brightening, chemical polishing, and bright dipping are synonymous. The rough surface is smoothed and brightened by dissolving the high spots. Acid oxidizing solutions are used most commonly, for example, phosphoric – nitric and phosphoric – sulfuric acids. The solutions can be tailored to specific alloys.

Electrolytic brightening, or electropolishing, also is used to produce smooth bright surfaces. The aluminum is the anode. A variety of baths are used, including fluoroboric acid, sodium carbonate – trisodium phosphate, and sulfuric – phosphoric – chromic acid. These brightening processes are followed by desmutting, anodizing, and sealing.

The surfaces produced range from specular to diffuse. Generally the quality of finish is highest for super pure aluminum (99.99 %) containing up to 2 % Mg and high-purity aluminum (99.7 to 99.85 %). However, these finishing methods are also applied to other alloys: e.g., 5457, 5357, 6463, 6063, 5052, 1100, 5005, 3003, and 6061, the quality of finish decreasing generally in this order.

5.3.5. Chemical Conversion Coating

Chemical conversion coatings are surface films formed by reactions that convert the surface of the aluminum into one of the components of the coating. Because the coatings are integral with the surface, they have excellent adherence. Although conversion coatings are thinner and less protective than anodic coatings, they give adequate corrosion and abrasion protection for many applications and are excellent substrates for lacquers, paints, plastic films, adhesives, and lubricants.

Many solutions have been used, but they generally contain agents to do at least two things: (1) dissolve the original aluminum oxide coating and some of the underlying aluminum and (2) form insoluble aluminum compounds (oxides, phosphates, or chromates) on the surface. Inhibitors are sometimes added to control the rate at which aluminum dissolves.

5.3.6. Anodizing

Anodizing converts the aluminum surface to aluminum oxide. The alloy and the conditions for forming the oxide can be chosen to produce a wide variety of useful characteristics. Uses include:

1) Decorative colors and textures
2) Substrates for paints, metal plating, adhesives, and photographic and lithographic emulsions
3) Corrosion resistance
4) Wear resistance
5) Electrical insulation
6) Quality control (inspection of parts for cracks and metallurgical nonuniformity)

Two principal types of anodized coatings are recognized: barrier and porous (Fig. 42) [36], [37]. The mechanism for the formation of the barrier coating is ionic: aluminum ions from the base metal combine with the oxygen of water in the electrolyte. The limiting thickness of the barrier layer is determined by the distance that the ions can penetrate at the applied potential, about 140 nm/V. If the oxide is appreciably soluble in the electrolyte, current continues to flow, and a relatively thick, porous oxide layer can develop.

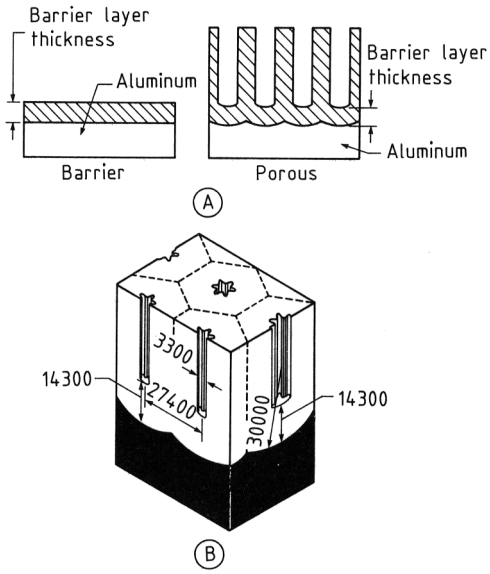

Figure 42. Structure of anodic coatings
A) Two principal types; B) Structure of 120-V phosphoric acid coating with dimensions of pore, cell, cell wall, and radius of curvature shown in nm

The limiting barrier layer thickness is then less than 140 nm/V. The porous oxide cells are predominately hexagonal, and each cell contains a pore. The dimensions of the cells and pores depend primarily on the potential, and secondarily on the electrolyte. The porous anodic coatings are essentially amorphous and can be sealed in boiling water. Anodizing at voltages above 100 V can form some eta alumina (η-Al_2O_3).

Some conventional anodizing processes are shown in Table 29 (next page). The most common electrolytes for conventional thick porous anodized coatings are *sulfuric acid* solutions. These coatings have a wide range of characteristics; the conventional coatings are decorative or protective and range in thickness from 4 to 30 µm. *Chromic acid* anodized coatings are thinner and less abrasion resistant than the sulfuric acid coatings, but they are more flexible and more resistant to deformation and heating without crazing. *Oxalic acid* anodized coatings are somewhat harder than the sulfuric acid coatings.

Table 30 (page after next) lists a number of processes that produce thicker and harder anodized coatings than the conventional processes.

Alloy Composition. The anodizing characteristics of aluminum are affected significantly by impurities and alloying elements. Alloys containing copper and silicon require higher voltage than unalloyed aluminum to produce a given thickness, but magnesium lowers the voltage. Appearance, continuity, abrasion resistance, dielectric strength, composition, and structure of the anodic layer are all affected by alloy composition.

Coloring. Pure aluminum (99.95 %) produces the most transparent coating, dullness generally increasing with the impurities and alloy content. Anodic coatings are colored by certain alloying elements (integral colors). Aluminum – silicon alloys such as 4043 provide durable gray coatings widely used as exterior architectural finishes. Chromium and copper in solid solution yield light gold finishes, and manganese imparts tan to brown shades.

Porous anodic coatings can be impregnated with organic and inorganic dyes and pigments during anodizing by incorporating the dyes and pigments in the electrolyte or during a post-anodizing treatment.

Table 29. Conventional anodizing processes, [11, p. 587]

Process bath	Amount, wt %	Temperature, °C	Duration, min	Voltage, V	Current density, A/dm^2	Film thickness, µm	Appearance	Remarks
Sulfuric acid								
Sulfuric acid	10	18	15 – 30	14 – 18	1 – 2	5 – 17	colorless, transparent	hard, unsuitable for coloring
Water	90							
Alumilite								
Sulfuric acid	15	21	10 – 60	12 – 16	1.3	4 – 23	colorless, transparent	good corrosion protection
Water	85							
Oxydal								
Sulfuric acid	20	18	30	12 – 16	1 – 2	15 – 20	colorless, transparent	good corrosion protection, suitable for variegated and golden coloring
Water	80							
Anodal and Anoxal								
Sulfuric acid	20	18	50	12 – 16	1 – 2	20 – 30	colorless, transparent	for coloring to dark tones, bronze, and black
Water	80							
Commercial chromic acid process								
Chromic acid	5 – 10	40	30 – 60	0 ∗	0.5 – 1.0	4 – 7	gray to iridescent	good chemical resistance, poor abrasion resistance, suitable for parts with narrow cavities, electrolyte not detrimental
Water	95 – 90							
Eloxal GX								
Oxalic acid	2 – 10	20 – 80	30 – 80	20 – 80	0.5 – 30	5 – 60	colorless to dark brown	hard films, abrasion resistant, coloring depends on alloy
Water	98 – 90							
Ematal								
Oxalic acid	1.2	50 – 70	30 – 40	120	3	12 – 17	gray, opaque, enamel-like	hard, dense film, extreme abrasion resistance
Titanium potassium oxalate	1							
Citric acid	0.025							
Boric acid	0.2							
Water	98							

∗ The voltage is increased from zero to a limit determined by the amperage

Table 30. Processes for hard coating, [11, p. 592]

Process bath	Amount, wt%	Temperature, °C	Duration, min	Voltage, V	Current density, A/dm^2	Film thickness, µm	Appearance	Remarks
Martin Hard Coat (MHC)								
Sulfuric acid	15	−4 to 0	80	20–75	2.7	50	light to dark gray or bronze	very hard, wear resistant
Water	85							
Alumilite 225 and 226								
Sulfuric acid	12	10	20 or 40	10–75	28	25–50	light to dark gray or bronze	very hard, wear resistant, allows higher operating temperature than MHC
Oxalic acid	1							
Water	87							
Kalcolor								
Sulfosalicylic acid	7–15	18–24	–		1.5–4	15–35	light yellow to brown to black	self-coloring process, the light-fast colors depend on alloy chosen
Sulfuric acid	0.3–4							
Water	81–93							
Lasser								
Oxalic acid	0.75	1–7	≦20	*	*	700	colorless	hard, thick coatings produced with special cooling process
Water	99.25							

* The voltage is a rising ramp from 50 to 500 V; the amperage is determined by the voltage

Sealing. Many anodic coatings are porous and subsequently are sealed to improve resistance to staining, resistance to corrosion, and durability of the colors. Sealing is accomplished by immersion in hot water. The mechanism is generally thought to involve conversion of the essentially amorphous alumina surface of the anodic coating to crystalline alpha alumina hydrate, increasing the volume and therefore closing the pores.

5.3.7. Metal Coatings

Certain metals can be deposited on the surface of aluminum articles by reaction (immersion and electroless processes) or electrolysis (electroplating). The former are usually functional, but the latter can be functional or decorative.

Immersion coating involves the reduction of cations by aluminum, whereas *electroless coating* involves the reduction of cations by reducing agents in the solution. Immersion coatings of zinc and tin plus copper are used primarily as a base for electroplating other metals. Tin has been used on aluminum aircraft engine pistons. Nickel is the most common electroless coating on aluminum, but cobalt, copper, gold, palladium, etc., can also be deposited by this method.

Metals are *electroplated* on aluminum for appearance; increased wear-, abrasion-, or erosion-resistance; lower electrical resistance; better solderability; and improved friction characteristics. Table 31 shows a few uses of electroplated products.

Electroplating of aluminum articles is made more complicated by the nature of the aluminum oxide surface and the reactivity of aluminum when the oxide is removed. Additional complications result from reactions between aluminum and the electroplating solutions, and galvanic reactions between aluminum and the plated metal. The metallurgical structure of aluminum alloys consists of solid solutions, constituents, dispersoids, and precipitates, each having a different reactivity. All of these forms must be considered to obtain useful electroplates.

5.3.8. Paints and Organic Films

Aluminum surfaces that have been properly pretreated are excellent substrates for paints and

Table 31. Uses of electroplated aluminum products, [11, p. 602]

Product	Form	Preplating treatment	Electroplating system	Thickness, µm	Purpose
Automotive					
Bearings	sheet	none	Pb + Sn + Cu	6 + 32	prevent seizing
Bumper guards	castings	buff and zincate	Cu + Ni + Cr	2.5 + 51 + 0.8	appearance, corrosion resistance
Tire molds	castings	none	hard Cr	51	wear and corrosion resistance
Aircraft					
Hydraulic parts, landing gears, small engine pistons	forgings	machine and zincate	Cu flash + Cu + hard Cr	2.5 + 25 + 76	wear resistance
Electrical and electronics					
Busbars, switchgears	extrusions	zincate	Cu flash + Cu + Ag	8 + 5	nonoxidized surface, solderability, corrosion resistance
Intermediate-frequency housings	die castings	zincate	Cu flash + Cu + Ag + Au	13 + 13 + 0.6	surface conductivity, solderability, corrosion resistance
Microwave fittings	die castings	zincate	Cu flash + Cu + Ag + Rh	0.25 + 13 + 0.5	smooth, nonoxidized interior, corrosion-resistant exterior
General Hardware					
Screws, nuts, bolts	castings	buff and zincate	Cd (on threads)	13, 0.5 on threads	corrosion resistance
Spray guns and compressors	die castings	buff and zincate	hard Cr	51	appearance
Window and door hardware	die castings	barrel burnish and zincate	brass	8	appearance, low cost
Household					
Coffee maker	sheet	buff and zincate	Cr	5	appearance, cleanness, resistance to food
Refrigerator handles, salad makers, cream dispensers	die castings	buff and zincate	Cu + Ni + Cr	2.5 + 13 + 0.8	contamination appearance, cleanness, resistance to food contamination
Personal products					
Compacts, fountain pens	sheet	buff and zincate	Cu flash + brass	5	appearance, low cost
Jewelry	sheet	buff and zincate	brass + Au	8 + 0.25	appearance, low cost

organic films. Optimum *pretreatment* is multistep and may include combinations of mechanical cleaning, chemical cleaning, conversion coating, anodizing, and priming. Pretreatment ranges in effectiveness from solvent cleaning, marginal even for interior service, to anodizing, which can give corrosion protection in severe environments. Conversion coatings are excellent bases for organic coatings and laminating adhesives, and they provide some corrosion protection.

Primer coatings are formulated specifically to promote the adherence and performance of finish coats. *Plastic coatings,* e.g., vinyl, epoxy, butyrate, polyethylene, and nylon, can be applied in thicknesses of 0.1 – 0.5 mm by melting fine particles of resin on the surface to form a continuous film. *Plastic films,* e.g., poly(vinyl chloride) and poly(vinyl fluoride), are bonded adhesively to aluminum.

5.3.9. Porcelain Enamels

Porcelain enamel is a glass film fused onto the surface of the metal. The components of the enamel, a mixture of glass and inorganic pigment particles, or frit, are applied as a water suspension. The part and the mixture are heated to dry and fuse the frit. The frit must melt below 550 °C, and its coefficient of thermal expansion must be close to that of the aluminum substrate.

Compositions of frits are complex, as illustrated by three examples in Table 32 [11, p. 511]. The lead-based enamels have high gloss and good weather resistance. The phosphate base has a lower melting temperature but is not as resistant to water. The barium enamels melt at higher temperature and have good resistance to chemicals.

Aluminum alloys for porcelain enameling are limited by their melting range, chemical compatibility with the glass, and strength after firing. The commonly used alloys are 1100, 3003, 6061, 443.0, and 356.0. Surface preparation consists of (1) removing dirt, grease, and oil; (2) removing thickened, irregular oxide films; and (3) applying a conditioning film.

Table 32. Compositions of frits

Constituent	Composition, wt%		
	Lead enamel	Barium enamel	Phosphate enamel
PbO	14 – 45	–	–
SiO_2	30 – 40	25	–
Na_2O	14 – 20	20	20
K_2O	7 – 12	25	–
Li_2O	2 – 4	–	4
B_2O_3	1 – 2	15	8
Al_2O_3	–	3	23
BaO	2 – 6	12	–
P_2O_5	2 – 4	–	40
F	–	–	5
TiO_2	15 – 20	*	*

* 7 to 9 wt % added to frit during mill preparation of the enamel slip

6. Corrosion Resistance of Aluminum

Corrosion literature deals with the theory and the many observations of the role of metallurgical structure, environment, stress, and time [4, Chap. 7], [8, pp. 204 – 36]. Most of the effort has been pragmatic, and the information consists of systematic observations on the performance of a variety of alloys, tempers, and protective systems in natural and artificial environments.

The oxide that forms on aluminum is bonded strongly to the metal and reforms very rapidly after damage in most environments. Consequently the corrosion resistance of aluminum is typically excellent, even though pure aluminum is more electropositive than all other structural metals except magnesium and probably beryllium. The stability region of the oxide is shown in a Pourbaix diagram of pH vs. potential (Fig. 43) [38]. Aluminum is protected by its oxide film, i.e., it is passive, in the pH range 4 – 8.5. The exact range depends on the form of oxide, the rate at which the oxide dissolves in the electrolyte, and the temperature. Outside the passive range, in both acid and alkaline solutions, aluminum can corrode. However, in some cases, corrosion does not occur even outside the passive range if the oxide is not soluble in the electrolyte or if the oxide is maintained by a strongly oxidizing solution.

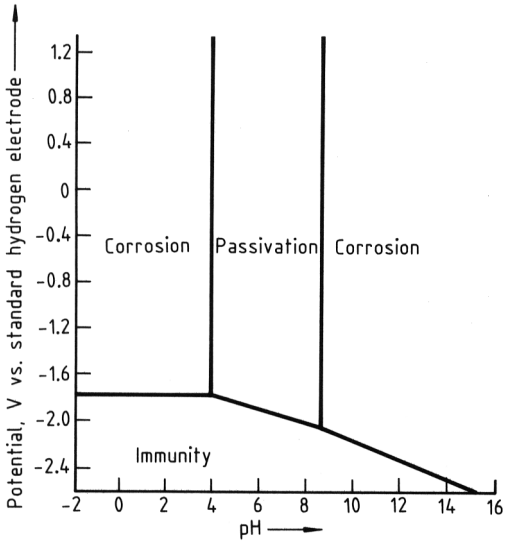

Figure 43. Pourbaix diagram for aluminum with an oxide film of hydrargillite ($Al_2O_3 \cdot 3H_2O$) at 25 °C

6.1. Pitting Corrosion

Corrosion of aluminum in the passive region is usually by pitting in the presence of halides and oxygen. In this kind of environment, aluminum is easily polarized to its pitting potential; i.e., it develops an electrode potential that exceeds the pitting potential. In deaerated solutions this is not possible. In general, the pitting potentials of aluminum in solutions of most other salts are significantly more cathodic than they are in halides, and the pitting potential is not reached.

6.2. Electrode Potentials

Because many types of corrosion proceed by electrochemical mechanisms, the electrode potentials (or solution potentials) of the phases in aluminum alloys are important for understanding corrosion. Electrode potentials are the potentials of the phases versus a reference electrode in a standard solution. Work by Brown and Mears led to wide acceptance of the 0.1 N calomel electrode, an aqueous solution containing 53 g NaCl + 3 g H_2O_2 per liter, and 25 °C for the study of aluminum alloys [1, p. 209]. The potentials for this solution are particularly useful because the behavior is similar to the behavior observed in marine and other saline environments. The potential values given here are referred to the 0.1 N calomel electrode, although the saturated calomel electrode is now preferred as the reference electrode. The saturated calomel electrode gives potentials that are 0.08 V more cathodic.

Figure 44 shows the effects of the major alloying elements on the solution potential of solid-solution alloys. Table 33 gives the solution potentials of some casting alloys, Table 34 some wrought alloys, and Table 35 some second-phase constituents in aluminum alloys.

Table 33. Electrode potentials of casting alloys

Alloy	Temper	Type *	Potential, V **	Nominal composition, wt%		
				Cu	Mg	Si
208.0	F	S	−0.77	4.0		3.0
295.0	T4	S, P	−0.70	4.5		1.1
	T6	S, P	−0.71			
	T62	S, P	−0.73			
296.0	T4	S, P	−0.71	4.5		2.5
308.0	F	P	−0.75	4.5		4.5
319.0	F	S	−0.81	3.5		6.0
	F	P	−0.76			
355.0	T4	S, P	−0.78	1.2	0.50	5.0
	T6	S, P	−0.79			
356.0	T6	S, P	−0.82		0.32	7.0
443.0	F	S	−0.83			5.2
	F	P	−0.82			
514.0	F	S	−0.87		4.0	
520.0	T4	S, P	−0.89		10.0	

* S for sand cast and P for permanent mold;
** Versus 0.1 N calomel electrode

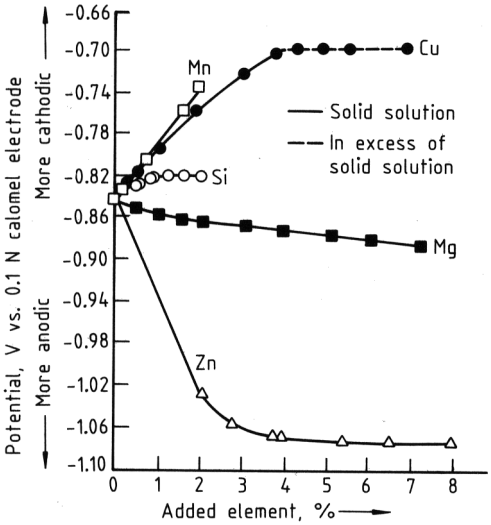

Figure 44. Effects of alloying elements on the electrode potential of aluminum

Table 34. Electrode potentials of wrought alloys

Alloy [a]	Temper	Potential, V [b]
1060	all [e]	−0.84
1100	all [e]	−0.83
2024	T3	−0.69 [c]
	T4	−0.69 [c]
	T6	−0.81
	T8	−0.82
3003	all [e]	−0.83
3004	all [e]	−0.84
5050	all [e]	−0.84
5456	all [e]	−0.87
6061	T4	−0.80
	T6	−0.83
7005	T6	−0.94
7072	all [e]	−0.96
7075	T6	−0.83 [d]
	T73	−0.84 [d]
	T76	−0.84 [d]

[a] Nominal compositions shown in Table 18;
[b] Versus 0.1 N calomel electrode;
[c] Varies ±0.01 V with quenching rate;
[d] Varies ±0.02 V with quenching rate
[e] Some value for all tempers.

Table 35. Electrode potentials of second-phase constituents in aluminum alloys

Phase	Potential, V ∗
Si	−0.26
Al_3Ni	−0.52
Al_3Fe	−0.56
Al_2Cu	−0.73
Al_6Mn	−0.85
$CuMgAl_2$	−1.00
$MgZn_2$	−1.05
Al_8Mg_5	−1.24

∗ Versus 0.1 N calomel electrode

The type and amount of corrosion are related to the electrode potentials of the phases in aluminum alloys. The second phases usually have potentials that are significantly different from the aluminum matrix. The matrix can be either pure aluminum or solid solutions of other elements and aluminum. When the second phase particles are anodic to the matrix, the particles can be dissolved from the surface of the metal. When the second phase is cathodic, second phase particles can accelerate corrosion of the matrix adjacent to the particles. Adjacent areas can also have different amounts of an element in solid solution, thus creating a significant anodic – cathodic relationship. This is common in some tempers of heat-treatable alloys. These metallurgical structure factors are often involved in localized, intergranular, and exfoliation corrosion and in stress-corrosion cracking.

Differences in electrode potentials between two aluminum alloys or between an aluminum alloy and another metal can also lead to galvanic corrosion. Table 36 gives electrode potentials for some other metals. A related situation occurs when a cathodic metal precipitates on the aluminum surface during corrosion; the cathodic metal (e.g., Cu, Hg, Pb) may enter the system from the environment or the alloy. This is sometimes called deposition corrosion.

Table 36. Solution potentials of some nonaluminum base metals

Metal	Potential, V ∗
Magnesium	−1.73
Zinc	−1.10
Cadmium	−0.82
Mild carbon steel	−0.58
Lead	−0.55
Tin	−0.49
Copper	−0.20
Stainless steel (series 300, type 430)	−0.09
Nickel	−0.07
Chromium	−0.40 to +0.18

∗ Versus 0.1 N calomel electrode

However, potential differences can be used to advantage. Alclad products consist of a strong core alloy clad with alloy anodic to the core, usually by 0.08 – 0. 10 V Corrosion will penetrate the cladding but not enter the core; instead, corrosion proceeds laterally as the cladding is consumed. Table 37 lists alloy pairs used in Alclad products.

Table 37. Combinations of alloys in Alclad products

Core alloy ∗	Cladding alloy					
	Alloy no.	Nominal composition, wt%				
		Si	Mn	Mg	Zn	Cr
2014	6003	0.7		1.2		
	6053	0.7		1.2		0.25
2024	1230		≧99.30 %			
2219	7072				1.0	
3003	7072				1.0	
3004	7072				1.0	
	7013		1.3		1.8	
6061	7072				1.0	
7075	7072				1.0	
	7008			1.0	5.0	0.18
	7011		0.20	1.3	4.8	0.12
7178	7072				1.0	

∗ Nominal compositions shown in Table 18

6.3. Crevice Corrosion

A concentration cell may be set up if there is an electrolyte between two aluminum faying surfaces, or between an aluminum surface and a nonaluminum surface, and one region of the electrolyte is oxygen-depleted relative to the region that is adjacent to the atmosphere. This may occur between the faying surfaces in a joint or in a stack or coil of aluminum sheet. The oxygen-depleted area is anodic and corrosion may occur there.

6.4. Intergranular Corrosion and Stress-Corrosion Cracking

In some environments and under some metallurgical conditions, aluminum alloys are susceptible to *intergranular attack.* If the alloy-temper is not susceptible to stress-corrosion cracking, then intergranular corrosion tends to decrease in time, and its effect on strength for a given depth of attack is about the same as pitting corrosion. For *stress-corrosion cracking* (SCC) to proceed, four conditions must be met: (1) a metallurgical structure susceptible to SCC, (2) an environment that can cause SCC, (3) a sustained tensile stress, and (4) time. Only the alloys containing large amounts of soluble elements, e.g., copper, magnesium, silicon, and zinc, are susceptible. For most of the commercial alloys, SCC-resistance tempers have been developed, and effective protection systems have also been developed.

The electrochemical theory of SCC developed about 1940 dealt with SCC on the basis of continuous areas of anodic material in or along the grain boundaries [39]. These anodic areas are precipitates of an anodic phase in the grain boundary or anodic solid-solution zones adjacent to the grain boundary. In the Al – Cu alloys the solute-depleted zones are most anodic. In the Al – Mg alloys the precipitate Al_8Mg_5 is the most anodic. The theory is now seen to be incomplete or inadequate and work continues, especially on the possibility of hydrogen playing a significant role [40]. A most promising elastic-plastic fracture mechanics model was developed. It deals with stress-corrosion cracking damage and the strength and toughness of the material [41].

6.5. Exfoliation Corrosion

Exfoliation corrosion, also known as lamellar layer or stratified corrosion, proceeds along paths approximately parallel to the surface. The corrosion product is voluminous, causing delamination and swelling. It is usually associated with a metallurgical structure consisting of long, flat grains and probably results from aligned intergranular or subgrain-boundary precipitates or from layers of slightly different composition. Metallurgical practices to prevent or minimize these susceptible structures have been developed.

6.6. Ratings for Resistance to Corrosion

From the vast amount of detailed information available, ratings of resistance to ordinary corrosion and to stress-corrosion cracking have been compiled for a number of alloy-tempers (Tables 14, 15, and 20, also see [8]). These ratings are useful, but the selection of the optimum alloy-temper-product for a particular application is many faceted and requires expertise.

The rating system for resistance to stress-corrosion cracking has been expanded to acknowledge that susceptibility depends not only on alloy-temper but also on product form and the direction of sustained stress with respect to the metallurgical grain structure. Generally, susceptibility is lowest when the stress is longitudinal, i.e., parallel to the major direction of metal flow. Susceptibility can be higher when stress is normal to the major flow and parallel to the minor flow (long transverse). Susceptibility is highest when the stress direction is normal to both major and minor flows (short transverse) [8, pp. 213 – 14].

6.7. Corrosion Prevention

Methods of preventing corrosion are well developed [4, Chap. 7]. The alloy-temper should be selected for suitability in the service environment. Corrosion risks should be minimized by proper design of the part and the equipment in which it serves. Suitable protective finishes, e.g., conversion coatings, anodized film, organic paints, and films, should be chosen. In some cases cathodic protection can be used.

7. Major Markets [25], [42]

By far the greatest part of aluminum shipments is in the form of aluminum alloys, and production of primary aluminum and consumption of aluminum are the best indicators of the importance of aluminum alloys.

World production of primary aluminum is summarized in Table 38, and per capita consumption of aluminum in selected countries is summarized in Table 39 (see page after next).

The major markets for aluminum products are shown in terms of use by major industries in the United States in Table 40 (see page after next). To eliminate duplication, the figures exclude intra-industry shipments for further fabrication.

Table 38. World primary aluminum production in 10^3 t, 1986 – 1996

Country	1986	1990	1994	1996
Total World [a]	15,412	19,295	19,206	20,607
Africa	554	605	590	1,027
Cameroon	84	93	89	80
Egypt	175	179	188	190
Ghana	125	174	141	137
South Africa	170	159	172	620
North America	4392	5615	5554	5859
Canada	1355	1567	2255	2282
United States	3037	4048	3299	3577
Latin America	1389	1787	1978	2018
Argentina	148	166	176	185
Brazil	757	931	1185	1190
Mexico	37	68	–	11
Suriname	29	32	32	32
Venezuela	418	590	585	600
Asia	1488	2013	3266	3611
Bahrain	178	213	447	450
China, P.R. [a]	410	850	1450	1780
India	257	433	472	518
Indonesia	219	186	222	225

Table 38. (Continued)

Country	1986	1990	1994	1996
Asia (cont.)				
Iran	40	60	116	118
Japan	140	34	17	17
Korea, Peop. Rep.	10	–	–	–
Korea, Republic of	19	2	–	–
Tajikistan [b]	–	–	235	198
Turkey	60	61	60	60
United Arab Emirates	155	174	247	245
European Union	2412	2365	2081	2183
Austria [c]	–	–	–	–
France	322	326	384	365
Germany [d]	826	740	505	577
Greece	124	150	144	130
Italy	243	232	176	185
Netherlands	266	270	219	226
Spain	355	353	338	362
Sweden [c]	–	–	84	98
United Kingdom	276	294	231	240
Other Europe	4059	5420	4149	4316
Austria [c]	93	89	–	–
Azerbaijan [b]	–	–	15	10
Bosnia-Herzeg. [e]	–	–	10	10
Croatia [e]	–	–	26	30
Czechoslavakia [a f g]	33	70	–	–
Hungary	74	75	31	25
Iceland	76	87	99	104
Norway	726	845	858	874
Poland [h]	48	46	50	52
Romania	269	168	119	140
Russia [b]	–	–	2670	2800
Serbia-Montenegro [e]	–	–	7	36
Slovakia [g]	–	–	60	50
Slovenia [e]	–	–	80	70
Sweden [c]	78	96	–	–
Switzerland	80	72	24	27
Ukraine [b]	–	–	100	88
U.S.S.R. [a]	2300	3523	–	–
Yugoslavia	282	349	–	–
Oceania	1118	1490	1588	1656
Australia	882	1230	1317	1372
New Zealand	236	260	271	284

[a] Estimated by the U.S. Geological Survey. [b] Reported as part of the Soviet Union until 1992. [c] Joined European Union in 1994.
[d] German Democratic Republic & Federal Republic of Germany combined in 1990.
[e] Reported as part of Yugoslavia until 1992.
[f] All production capacity formerly reported as Czechoslavakia is now reported as Slovakia.
[g] Includes secondary unalloyed ingot.
[h] Includes primary alloyed ingot.

8. References

General References

1. K. R. Van Horn (ed.): *Aluminum, vol. I: Properties, Physical Metallurgy, and Phase Diagrams,* American Society for Metals, Metals Park, Ohio, 1967.
2. K. R. Van Horn (ed.): *Aluminum, vol. II: Design and Application,* ASM, 1967.
3. K. R. Van Horn (ed.): *Aluminum, vol. III: Fabrication and Finishing,* ASM, 1967.
4. J. E. Hatch (ed.): *Aluminum: Properties and Physical Metallurgy,* ASM, 1984.
5. D. Altenpohl: *Aluminum Viewed From Within,* Aluminium-Verlag, Düsseldorf 1982.
6. *Aluminum Standards and Data – 1982,* The Aluminum Association, Washington, D.C.
7. Properties and Selection: "Irons and Steels," *Metals Handbook,* 9th ed., **vol. 1,** ASM, 1978.
8. Properties and Selection: "NonFerrous Alloys and Pure Metals," *Metals Handbook,* 9th ed., **vol. 2,** ASM, 1979.

Table 39. Per capita aluminum consumption of selected countries, 1986 – 1996 (in kg/a) ∗

Country	1986	1990	1994	1996
Argentina	2.1	1.5	3.1	3.2
Australia	17.3	18.6	20.0	18.6
Austria	14.5	21.5	21.6	22.5
Belgium	10.6	12.4	8.3	na
Brazil	3.1	2.1	3.0	3.5
Cameroon	3.1	1.9	1.1	1.6
Canada	18.7	21.8	25.9	24.3
Denmark	9.6	17.7	29.8	19.5
Finland	8.4	10.4	12.4	11.6
France	14.4	17.4	17.5	17.6
Germany	25.5	29.7	25.0	24.0
Greece	1.5	5.8	na	na
Hungary	17.0	15.2	na	na
Iceland	8.2	5.4	12.0	17.3
India	0.4	0.5	0.5	na
Ireland	5.0	6.1	na	20.3
Italy	14.6	19.4	21.0	18.6
Japan	20.3	31.4	31.8	31.6
Mexico	1.3	2.5	2.7	2.6
Netherlands	15.0	16.1	25.0	25.7
New Zealand	10.2	10.2	16.7	14.7
Norway	18.5	19.0	21.1	20.9
Panama	0.8	2.4	2.9	5.1
Philippines	0.1	0.2	na	na
Portugal	5.0	6.3	na	na
Russia	na	na	2.7	2.7
El Salvador	0.2	0.3	0.5	0.1
Saudi Arabia	5.1	na	na	na
Singapore	11.5	19.1	na	na
South Africa	4.2	11.9	3.4	4.5
Spain	5.7	7.4	8.7	na
Sweden	20.5	24.8	21.0	26.5
Switzerland	20.4	22.6	21.0	16.5
Turkey	1.6	4.3	3.5	na
United Kingdom	11.1	14.0	14.0	18.0
United States	26.4	27.0	31.2	31.3
Venezuela	6.6	8.3	5.7	5.4

∗ na: Not available.

Table 40.U.S. average annual net shipments by major market

Major Market	1995		1996		% change,	% annual growth,
	%	10^6 kg	%	10^6 kg	1995 – 1996	1986 – 1996
Building &construction	12.7	1218	13.7	1310	7.5	– 1.0
Transportation	27.3	2613	27.6	2640	1.0	7.2
Consumer durables	6.5	622	6.6	634	2.0	1.4
Electrical	6.9	658	7.0	666	1.2	0.5
Machinery &equipment	6.0	571	5.9	568	– 0.6	4.0
Containers &packaging	24.1	2312	22.8	2180	– 5.7	1.4
Other	2.9	280	2.9	280		0.8
Domestic, total	86.3	8275	86.5	8278		2.5
Exports	13.7	1310	13.5	1290	– 1.5	7.6
Total Shipments	100.0	9585	100.,0	9568	– 0.2	3.1

9. Properties and Selection: "Stainless Steels, Tool Materials and Special-Purpose Metals," *Metals Handbook,* 9th ed., **vol. 3,** ASM, 1980.
10. Heat Treating, *Metals Handbook,* 9th ed., **vol. 4,** ASM, 1981.
11. "Surface Cleaning, Finishing, and Coating," *Metals Handbook,* 9th ed., **vol. 5,** ASM, 1982.
12. "Welding, Brazing, and Soldering," *Metals Handbook,* 9th ed., **vol. 6,** ASM, 1983.

Specific References

13. *Registration Record of Aluminum Association Alloy Designations and Chemical Composition Limits for Aluminum Alloys in the Form of Castings and Ingot,* The Aluminum Association, Washington, D.C., revised Aug. 1982.
14. *Registration Record of Aluminum Association Designations and Chemical Composition Limits for Wrought Aluminum and Wrought Aluminum Alloys,* The Aluminum Association, Washington, D.C., revised Jan. 1984.
15. R. E. Spear, G. R. Gardner, *Trans. Am. Foundrymen's Soc.* **71** (1963) 209 – 15.
16. M. C. Flemings: *Solidification Processing,* McGraw-Hill, New York 1974.
17. C. E. Ransley, H. Neufeld, *J. Inst. Met.* **74** (1947/48) 599 – 620.
18. D. E. J. Talbot, *Int. Metall. Rev.* **20** (1975) 166 – 84.
19. D. E. J. Talbot, D. A. Granger, *J. Inst. Met.* **92** (1963/64) no. 9, 290 – 97.
20. G. W. Lorimer in K. C. Russell, H. I. Aaronson (ed.): *Precipitation Processes in Solids,* The Metallurgical Society of AIME, Warrendale, PA 1978, pp. 87 – 119.
21. E. A. Starke, Jr., T. H. Sanders, Jr.: *Aluminum-Lithium Alloys II,* The Metallurgical Society of AIME, Warrendale, PA 1984.
22. L. F. Mondolpho: *Aluminum Alloys: Structure and Properties,* Butterworths, London – Boston 1976.
23. J. M. Silcock, T. J. Heal, H. K. Hardy, *J. Inst. Met.* **82** (1953/54) 239 – 48.
24. Aluminum Company of America, US 3198676, 1965 (D. O. Sprowls, J. A. Nock, Jr.).
25. *Aluminum Statistical Review for 1982,* The Aluminum Association, Washington, D.C.
26. E. F. Emley, *Met. Technol. (London)* **3** (1976) 118 – 27.
27. J. E. Dore, J. C. Yarwood, *Light Met. (N.Y.)* **2** (1977) 171 – 89.
28. E. F. Emley, *Int. Met. Rev.* **21** (1976) 75 – 115.
29. H. A. Meier, G. B. Leconte, A. M. Odok, *Light Met. (N.Y.)* **2** (1977) 223 – 33.
30. *1983 Annual Book of ASTM Standards-Section 2. Nonferrous Metal Products, vol. 02.02: Die-Cast Metals; Aluminum and Magnesium Alloys,* American Society for Testing and Materials, Philadelphia, PA 1983.
31. W. F. Kehler: *Handbook of International Alloy Compositions and Designations, vol. III,* Aluminum Metals and Ceramics Information Center, Battelle Columbus Lab., Columbus, Ohio, 1980.
32. *Registration Record of International Alloy Designations and Chemical Composition Limits for Wrought Aluminum and Wrought Aluminum Alloys,* The Aluminum Association, Washington, D.C., revised 1 June 1980.
33. R. M. Hart: *Wrought P/M Alloys 7090 and 7091,* Aluminum Company of America, Alcoa Center, PA 1981.
34. *Met. Prog.* **124** (1983) no. 1, 45.
35. R. Knoll: *Materials and Processes – Continuing Innovations* [Proc. Conf.], Anaheim, CA, 12 – 14 April 1983, Society for the Advancement of Material and Process Engineering, Azusa, Calif., 1983, pp. 133 – 48.
36. M. S. Hunter, P. E. Fowle, *J. Electrochem. Soc.* **101** (1954) 481 – 85.
37. F. Keller, M. S. Hunter, D. L. Robinson, *J. Electrochem. Soc.* **100** (1953) 411 – 19.
38. M. Pourbaix et al.: *Atlas of Electrochemical Equilibria in Aqueous Solutions,* Pergamon Press, Oxford, NY 1966.
39. D. O. Sprowls, R. H. Brown, in R. W. Staehle, A. J. Forty, D. Van Rooyen (eds.): *Fundamental Aspects of Stress-Corrosion Cracking,* National Association of Corrosion Engineers, Houston, TX 1969.
40. V. A. Marichev, *Werkst. Korros.* **34** (1983) no. 6, 300 – 308.
41. D. O. Sprowls, R. J. Bucci, B. M. Ponchel, R. L. Brazill, P. E. Bretz: *A Study of Environmental Characterization of Conventional and Advanced Aluminum Alloys for Selection and Design,* Aluminum Company of America, Alcoa Lab., Alcoa Center, PA, for NASA, Contract NAS1-16424, dated 1984 May 31.
42. *Aluminum Statistical Review – Historical Supplement,* The Aluminum Association, Washington, D.C. 1982.

Aluminum Compounds, Inorganic

Otto Helmboldt, Giulini Chemie GmbH, Ludwigshafen, Germany (Chap. 1 and 2)

L. Keith Hudson, Aluminum Company of America, Alcoa Center, Pennsylvania, USA (Chap. 2.1)

Chanakya Misra, Aluminum Company of America, Alcoa Center, Pennsylvania, USA (Chap. 2.1)

Karl Wefers, Aluminum Company of America, Alcoa Center, Pennsylvania, USA (Chap. 2.1)

Wolfgang Heck, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 3)

Hans Stark, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 4.1 and 5)

Max Danner, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Germany (Chap. 4.2, 4.3 and 5)

Norbert Rösch, Clariant Produkte (Deutschland) GmbH, Gersthofen, Germany

Related Articles → Aluminum Oxide is a separate keyword. Aluminum fluoride and cryolite → Fluorine Compounds, Organic

1. Aluminum Sulfate and Alums 1
1.1. Aluminum Sulfate 1
1.1.1. Properties 1
1.1.2. Production 2
1.1.3. Uses 4
1.2. Alums 5
1.2.1. Potassium Aluminum Sulfate 6
1.2.2. Ammonium Aluminum Sulfate 7
1.2.3. Sodium Aluminum Sulfate 8
2. Aluminates 8
2.1. Sodium Aluminate 8
2.2. Barium Aluminates 9
3. Aluminum Alkoxides 10
4. Aluminum Chloride 10
4.1. Anhydrous Aluminum Chloride . . 10
4.1.1. Properties 10
4.1.2. Production 11
4.1.3. Quality Specifications and Analysis . 11
4.1.4. Handling, Storage, and Transportation 12
4.1.5. Uses 12
4.2. Aluminum Chloride Hexahydrate . 13
4.3. Basic Aluminum Chlorides 13
5. Toxicology 15
6. References 15

1. Aluminum Sulfate and Alums

1.1. Aluminum Sulfate

Aluminum sulfate is the second most important industrial compound of aluminum, after aluminum oxide. Aluminum sulfate was first used in Paris in 1844 to replace potassium alum. Today, it has taken over almost all areas of application that potassium alum originally had.

1.1.1. Properties

Aluminum sulfate is almost insoluble in anhydrous alcohol but readily soluble in water; aqueous solutions are acidic. Literature data on solubility and the structure of the precipitate in water differ markedly and should be used with caution (see Table 1). Previously, $Al_2(SO_4)_3 \cdot 18\ H_2O$ was thought to crystallize from aqueous solution under normal conditions (20 °C, 1 bar). This is now doubted, because such factors as hydrolysis, oversaturation, shifts in equilibrium, and, particularly, poor crystal formation make definitive characterization difficult [5]. The common form of aluminum sulfate, generally considered to be $Al_2(SO_4)_3 \cdot 18\ H_2O$ [*7784-31-8*], occurs in nature as alunogen (hair salt) and can be crystallized from hydrochloric acid solution as microscopically small, white needles. However, some researchers ascribe a hydrate water content of 17 mol to this form of aluminum sulfate [6]. A hydrate containing 27 mol of water can be prepared readily and in high purity [1]. Other well-defined aluminum sulfates contain 16, 10, and 6 mol of water. A total of 39 basic and 3 acidic aluminum sulfates, as well as 13 different hydrates of the neutral salt, are described in the literature [7]. The existence of aluminum

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_527.pub2

sulfates with 14, 13, 12, 9, 7, 4, 2, and 1 mol of water can be concluded from the vapor-pressure curves and the dehydration curves of the $Al_2(SO_4)_3 - H_2O$ and $Al_2(SO_4)_3 - Al(OH)_3 - H_2O$ systems [8, 9].

In industrial practice, the hydrate water content is unimportant because "crystalline" aluminum sulfate is a ground, microcrystalline solid with variable water content, which is obtained from a melt. The desired Al_2O_3 content is adjusted within certain limits by heating. At a temperature above 340 °C, anhydrous aluminum sulfate [*10043-01-3*] is formed, a white powder, ϱ 2.71 g/cm^3, that decomposes above 770 °C to aluminum oxide [9].

1.1.2. Production

In Germany and most European countries, aluminum sulfate is produced on a large scale only from aluminum hydroxide and sulfuric acid by the Giulini process. In this process, aluminum sulfate is obtained relatively easily with high purity. Production by the action of sulfuric acid on aluminum-containing ores (clays and bauxite with high silicon and low iron content, e.g., bauxite with $SiO_2 > 5$ %, Fe_2O_3 ca. 1 %) is still important in some countries.

Giulini Process [10]**.** The Giulini method for producing various grades of aluminum sulfate, containing between 8 and 23 % Al_2O_3, is shown schematically in Figure 1.

A pressure-resistant, stirred vessel (a) is filled with aluminum hydroxide (moist or dry). The calculated quantity of warm sulfuric acid is added from a preheater, and the mixture is stirred. In calculating the sulfuric acid concentration required to obtain an SO_3 content of ca. 1 % below stoichiometric, all process steps involving the introduction or the removal of water must be accounted for. Generally, acid of density 1.6 g/cm^3 is used. The reaction starts after 60 – 300 s and is complete after 10 – 12 min. The heat of reaction causes the temperature to rise to ca. 170 °C, while the pressure rises to 5 – 6 bar. The mixture must not be stirred for more than 1 h because otherwise the aluminum sulfate can hydrolyze to give insoluble basic aluminum sulfate and strongly acidic sulfate melt. An autoclave unit allowing batches of 2.5 t can produce ca. 50 t of aluminum sulfate in a 10-h shift.

The melt is led into a copper container where it is concentrated by flash evaporation (b). From the evaporator, the melt is sucked into a well-isolated vacuum tank (c), which is evacuated to the vapor pressure of the aluminum sulfate melt. This vacuum cooling avoids incrustation of the heat exchanger surfaces.

The melt falls from the vacuum container into the mixer, where it is seeded at 85 °C with 1 – 2 % aluminum sulfate powder. The pulp-like product reaches the "crystallization belt," a smooth, heat-resistant, trough-form rubber conveyor belt (e), and crystallizes there in ca. 30 min. Because of the high heat of crystallization, the material has a temperature of ca. 90 °C and cannot be broken to fine size in one step (f). It passes over air-cooled conveyor belts until it has cooled to 40 °C, after which it is ground (h) and sieved (i). The goods are filled (j) into paper or jute sacks or transported loose in silo cars. Aluminum sulfate with 17.2 % Al_2O_3 produced in this way contains only 0.01 % insoluble material; therefore, digestion of the aluminum hydroxide is almost complete. For transportation as a solution the Al_2O_3 content is adjusted to ca. 8 % to avoid crystallization during transport.

Production from Bauxite. Finely ground bauxite (for example, 60 % Al_2O_3, 1.5 % Fe_2O_3, 1.6 % TiO_2, 3.0 % SiO_2, 32 % H_2O) also can be used as starting material. In this case, 3 mol H_2SO_4 are charged onto 1 mol Al_2O_3. The Fe_2O_3 component in bauxite is disregarded because the Al_2O_3 is only 97 – 98 % digested and therefore sufficient sulfuric acid is available for production of aluminum sulfate containing 17.5 % Al_2O_3. The aluminum sulfate obtained by digestion of bauxite contains ca. 0.5 % Fe_2O_3 and ca. 2.2 % insoluble residue.

Production from Less Pure Starting Materials. Acid digestion of predominately silicon-rich raw materials gives a solution of aluminum sulfate. The purity depends on the process and starting materials. Iron, which strongly interferes, is precipitated with calcium hexacyanoferrate (II) as Berlin blue (iron hexacyanoferrate (II)), with calcium sulfide as iron sulfide, or by hydrolysis as basic iron sulfate. For details, see [11]. The clear solution is de-

Table 1. Solubility of aluminum sulfate as a function of temperature (grams of anhydrous salt per 100 g water)

t, °C	0	10	20	30	40	50	60	70	80	90	100	Reference
	31.3	33.5	36.15	40.36	45.73	52.13	59.10	66.23	73.14	80.83	89.11	[1]
	31.2	33.5	36.4	40.4	46.1	52.2	59.2	66.1	73.0	80.0	89.0	[2]
	27.5	27.6		28.0	28.8	29.9	31.0	32.8	36.6	38.74	46.85	[3]
	23.9	25.0	26.9		31.5		36.4		41.7		47.0	[4]

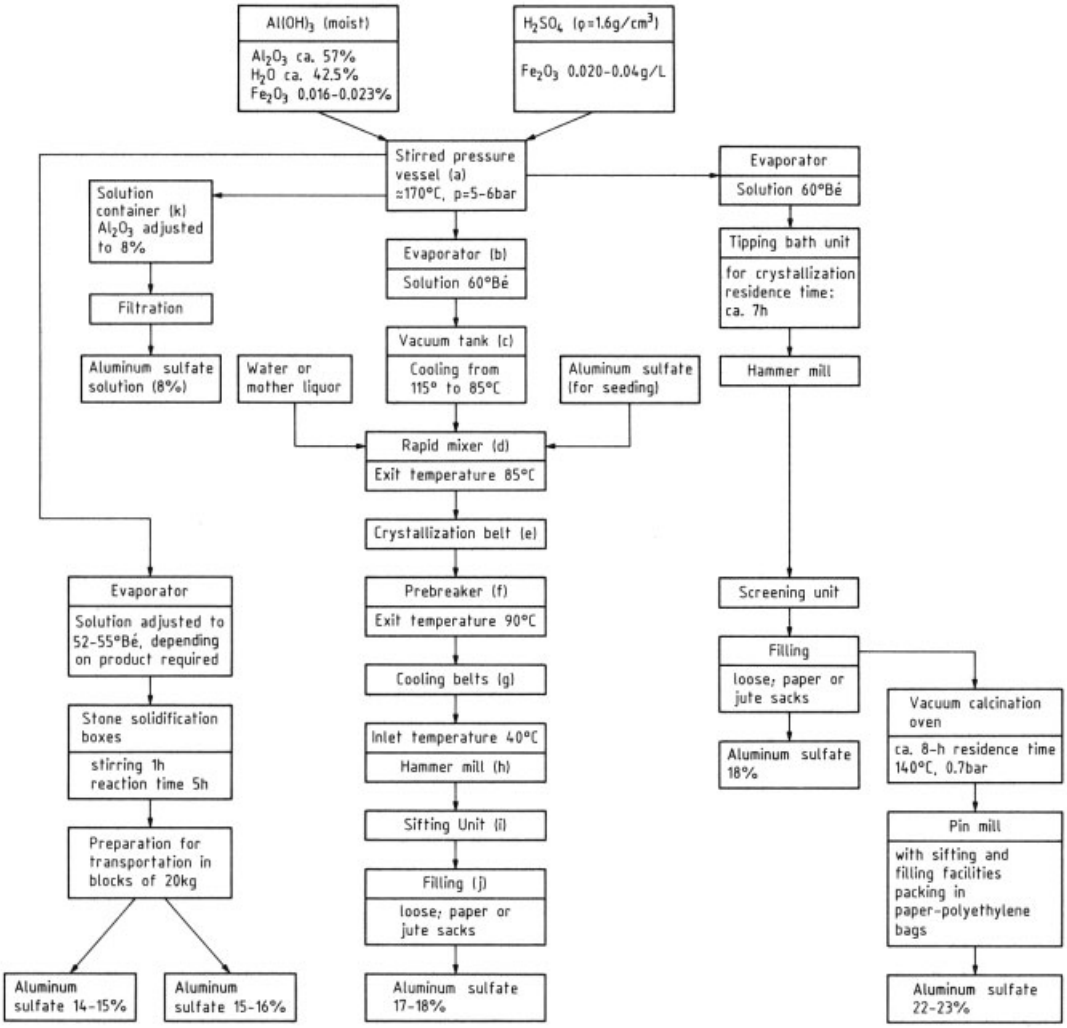

Figure 1. Production of aluminum sulfate by the Giulini process [10]

canted and sold as a liquid or concentrated to 61.5 °Bé, allowed to solidify, and milled. The final product, which contains ca. 0.5 % Fe_2O_3 and 0.1 % insoluble material, is the *technical grade.* In addition, there is an *iron-free grade* having an Fe_2O_3 content of ca. 0.005 %.

The *Kretzschmar* process is used to produce very pure, iron-free aluminum sulfate by digestion of clay (for example, 40 – 43 % Al_2O_3, 53 – 56 % SiO_2, 2 – 4 % Fe_2O_3) with sulfuric acid. The greater part of the impurities is removed and crystals are separated from the solution by stirring. The formation of colloids is avoided by using vacuum apparatus. Pure, large crystals of $Al_2(SO_4)_3 \cdot 18\,H_2O$ (15.3 % Al_2O_3) can be separated easily from the impure mother liquor by

centrifuging [12]. The residue from the digestion process (SiO_2) can be converted with lime to calcium hydrosilicate, which increases the hardness of and also plasticizes lime mortar.

In a process developed by the U.S. Bureau of Mines, alcohol is used to reduce the viscosity of the oversaturated aluminum sulfate mother liquor [13].

Olin Mathieson Chemical Corp. developed an economical process for producing high-quality aluminum sulfate from the clay or waste shale of coal mines. Large crystals (1.5 – 3 mm) with iron contents of less than 0.03 % are produced in a patented crystallizer [14]. A process for producing aluminum sulfate from alum-containing ores is given in [15].

Aluminum sulfate is produced from waste "red mud" (from the aluminum oxide industry) by suspending the mud in water and passing sulfur dioxide through the suspension until pH 2 is reached. After filtration and removal of the sulfur dioxide in vacuo to give a pH of 4.5 – 5.0, $Al(OH)SO_3$ and $SiO_2 \cdot n\,H_2O$ precipitate. The precipitate is filtered and treated with sulfuric acid, whereby aluminum sulfate dissolves. [16].

Commercial Grades. Whereas $Al_2(SO_4)_3 \cdot 18\,H_2O$ theoretically contains 15.3 % Al_2O_3, commercial grades of aluminum sulfate contain 14 – 15 %, 15 – 16 %, 17 – 18 %, 18 %, or 22 – 23 % Al_2O_3 (Table 2). Generally, the aluminum sulfate containing 17 – 18 % Al_2O_3 (water-soluble aluminum content, calculated as Al_2O_3) is used most frequently. The calculated hydrate water content of this aluminum sulfate is 13 mol. The commercial grade containing 17 – 18 % Al_2O_3 is delivered also with various special qualities, such as low arsenic or low iron content.

The *basicity* of the product is defined as the excess in Al_2O_3 over the stoichiometric SO_3 : Al_2O_3 ratio. The 17 – 18 % Al_2O_3 grade has a basicity of 0.1 – 1 %, i.e., the Al_2O_3 content exceeds the stoichiometric amount by 0.1 – 1 %. Occasionally, the term "basicity" is defined as the ratio % SO_3 to % Al_2O_3. For commercial aluminum sulfate, this ratio is ca. 2.30, whereas the theoretical value is 2.35. This weakly basic aluminum sulfate has the advantage that it is less hygroscopic and therefore hardly affects either the metal parts of the apparatus during use or paper sacks during transport. Also, free acid damages cellulose fibers in the paper industry.

1.1.3. Uses

About two thirds of the total aluminum sulfate production is used for treating water. About one half of the total production goes into the paper industry (paper sizing, pH adjustment, wastewater purification) [17, p. 246].

Paper Industry. Aluminum sulfate is used for precipitating and fixing sizing agents, wet-strength agents, and basic dyes; for improving retention; for dispersing resin particles that stick together and block sieves; as the starting material of a high-quality slip for coating glossy paper (satin white); and for the production of lake dyes and wallpaper. For an overview of uses in newspaper factories, see [18]; for a detailed discussion on the importance of aluminum sulfate in paper production, see [5]. Aluminum sulfate with more than 0.2 % Fe_2O_3 gives paper a yellow tinge and cannot be used for good-quality white paper.

Water Purification. Aluminum sulfate is an important flocculating agent for purifying water [19]. The mechanism is as follows: Positively charged aluminum ions, which are hydrated or hydroxylated with water via numerous intermediate stages, neutralize the negative charge on the colloidal material in the water. As a result of mutual adsorption, the material flocculates, sediments at various rates, and finally settles as sludge after 15 – 150 min. To neutralize the hydrogen atoms formed on hydrolysis of aluminum sulfate, some carbonate hardness is consumed:

$$Al_2(SO_4)_3 + 6\,HCO_3^- \rightarrow 2\,Al(OH)_3 + 6\,CO_2 + 3\,SO_4^{2-}$$

For 50 mg/L of the 17 – 18 % Al_2O_3 grade 0.25 mmol/L of carbonate hardness (based on alkaline-earth ions) is consumed. Generally, 5 – 50 g aluminum sulfate is sufficient to purify 1 m^3 of water. An advantage of using aluminum sulfate is the lack of postflocculation, provided the correct amount is added. Disadvantages are the formation of free carbon dioxide, the increase of non-carbonate hardness, such as calcium sulfate, and the pH remaining between 5 and 7. If no natural bicarbonate hardness is

Table 2. Commercial grades of aluminum sulfate

Grade	Al_2O_3, wt %	Fe_2O_3, wt %	Insoluble material, wt %	Basicity *, %
8 % (solution)	8	0.004	–	1
14 – 15 %	14 – 15	0.006 – 0.008	0.04	0.1 – 1.0
15 – 16 %	15.1 – 16.0	0.006 – 0.008	0.04	0.1 – 1.0
17 – 18 %	ca. 17.2	< 0.01	0.03	0.1 – 1.0
18 %	ca. 18.0	< 0.01	< 0.03	0.8 – 1.5
22 – 23 %	ca. 22.8	< 0.02	ca. 0.03	1.0 – 1.8

* Defined in text

present, an alkaline substance that also has a dispersing effect, such as lime or sodium aluminate, must be added. Aluminum sulfate, used in combination with calcined leached clay, is suitable for binding mercury ions, for example, in sea water [20].

Other Uses. Aluminum sulfate with a very low iron content (below 0.01 %) is used as a mordant in dyeing; a higher iron content is unacceptable because this leads to color changes. Further uses are for pickling of seeds, deodorizing of mineral oils, tawing, and producing aluminum hydroxide gel employed, for example, as a filler for synthetic rubber. Aluminum sulfate has some importance also as a catalyst support. Finally, it is the starting material for almost all other aluminum compounds.

Production and Capacity Data. Aluminum sulfate production in the USA was 1 168 000 t in 1980 and 1 075 000 t in 1982 [21]; that in Japan was 757 000 t in 1980 [22]. The main U.S. producer of aluminum sulfate is Allied Corp. The main Japanese producers are Nikkei Kako Co. and Scintoma Aluminum Smelting Co.

1.2. Alums

Historical Aspects. *Alumen*, from which the word alum is derived, was known to the ancient Greeks and Romans as an astringent and also as a mordant in dyeing wool. Alum was employed in processing skins, for embalming animal and human corpses, and for fireproofing wood. However, what was known at first as alum was not a defined substance but referred to both alum-containing minerals and mixtures of alum with iron vitriol. Paracelsus was the first to distinguish true alum from iron vitriol. In the years 1776 – 1798, Chaptal and Vauquelin established that alum was a double salt of potassium and aluminum sulfates and that the potassium ions could be replaced by ammonium ions. The raw material of the ancient alum industry was alum stone (alunite) or alum shale. Alum stone contains all components of the alum in the correct ratios and is processed by roasting and leaching. Roasting, aging for several years, and leaching of alum shale gave an aluminum solution from which the alum was precipitated with alkali. The alum industry played an important role during the whole of the Middle Ages.

The industrial production of the alums ceased to have particular importance when it became possible to produce aluminum sulfate economically in high purity, because only the aluminum content is of practical importance. Most of the alum production methods now have only limited, partly only historical, interest. Clay and other alkali-containing silicates, particularly bauxite, were used as raw materials and were digested using alkali or acid. Today, alums are produced only from aluminum hydroxide, which is obtained by the Bayer process from bauxite (→ Aluminum Oxide).

General Properties. Alums are crystalline double salts of the general formula (cation 1)$^+$ (cation 2)$^{3+}$ (anion^{2-})$_2$ · 12 H_2O. The most useful alums are those with trivalent aluminum cations and sulfate anions, $M^+Al^{3+}(SO_4^{2-})_2 \cdot 12\ H_2O$. The individual components of the alum can be replaced, retaining both the regular crystalline form and the hydrate water. Alums with the following trivalent metals are known: iron, chromium, cobalt, manganese, titanium, vanadium, gallium, indium, scandium, rhodium, and iridium. The monovalent component can be an alkali metal, or it can be ammonium, alkylammonium, arylammonium, or thallium. The existence of a lithium alum is uncer-

tain. Known combinations include potassium–aluminum, cesium–aluminum, potassium–chromium, and hydroxyammonium–aluminum. A selenate alum also is known: $KAl(SeO_4)_2 \cdot 12\ H_2O$.

All alums crystallize in strongly refracting octahedra or cubes [1]. They have an astringent taste and rot-proofing, protein-precipitating properties.

In *aqueous solution*, alums show all the chemical properties that their components show separately. Physical properties, such as color, electrical conductivity, and freezing-point lowering, are the sum of the properties of the components, provided the solutions are very dilute. At higher concentrations, complexes, such as $[Al(SO_4)_2 \cdot (H_2O)_2]^-$, are formed. The solubilities of the alums in water decrease from sodium to cesium (see Table 3). On heating, the alums lose their water of crystallization either partially or completely. However, part of the water is taken up again at normal temperature and humidity.

1.2.1. Potassium Aluminum Sulfate

Potassium aluminum sulfate [*7784-24-9*], known as potassium alum, $KAl(SO_4)_2 \cdot 12\ H_2O$, ϱ 1.75 g/cm^3, M_r 474.4, and ammonium aluminum sulfate are the aluminum compounds known longest. Potassium alum occurs in nature as efflorescence on alum shale and in volcanic areas on trachyte and lava as feather alum.

Potassium alum crystallizes in large, colorless, transparent octahedra, which melt at 92.5 °C in their own water of crystallization. The hardness of the crystals on the Mohs scale is 2. The octahedra absorb long-wavelength IR radiation almost completely, but are transparent to visible light. Certain substances (hydroxides, carbonates, borates, carbamides, also metals and organic dyes) promote the formation of basic aluminum sulfate by binding free sulfuric acid in the mother liquor. Under these conditions, the cubic form is favored (cubic or Roman alum).

Potassium alum is stable in air of normal humidity. Dehydration does not begin below 30 °C, but at 65 °C nine moles of water are lost. Literature data on the dehydration are contradictory. Investigations on thermal decomposition are given in [23]. According to these results, K_2SO_4, γ-Al_2O_3, and 3 $K_2SO_4 \cdot Al_2(SO_4)_3$ form at 780 °C, whereas K_2SO_4, α-Al_2O_3, and $K_2O \cdot 12\ Al_2O_3$ are produced at 1400 °C. When heated above its melting point, potassium alum dehydrates, forming calcined alum (alum ustum), KAl $(SO_4)_2$ [*10043-67-1*]. At red heat, SO_3 is released.

Potassium alum is soluble in dilute acid but almost insoluble in anhydrous alcohol, acetone, and methyl acetate. Because of the marked increase in solubility in water with temperature, potassium alum can be purified more easily than other aluminum salts by recrystallization and, in particular, can be freed from iron sulfate. Mixed crystals form readily with ammonium sulfate.

Basic potassium aluminum sulfate, $K[Al(OH)_2]_3(SO_4)_2 \cdot 3/2\ H_2O$, occurs in nature as löwigite. The compound is made synthetically as a rather insoluble, amorphous powder by heating aluminum sulfate, water, and an excess of potassium sulfate or by heating potassium alum with water in the ratio 1 : 4 to 200 °C [24]. Another basic potassium aluminum sulfate, $K[Al_3(OH)_6(SO_4)_2]$ [*1302-91-6*], containing less water, is the alum stone (alunite) occurring in nature.

Production. Aluminum hydroxide (wet hydrate; Al_2O_3 ca. 57 %, Fe_2O_3 0.016 – 0.023 %, H_2O 42.5 %) and sulfuric acid ($\varrho \approx 1.6$ g/cm^3; Fe_2O_3 0.020 – 0.040 g/L) react in a stirred, corrosion-resistant pressure boiler at 5 – 6 bar to form aluminum sulfate as shown in Figure 1. The aluminum sulfate melt is then led, with release of pressure, into a copper container. Here, a stoichiometric quantity of potassium sulfate (as chloride-free as possible) is added. The solution is heated to ca. 100 °C for 2 – 3 h and adjusted to 40 – 44 °Bé with mother liquor. After filtration from insoluble material, the alum melt is left in crystallization boxes for 10 days, after which the alum is removed in blocks. The product obtained in this way contains < 0.001 % Fe_2O_3, 0.001 – 0.004 % chloride, and < 0.01 % insoluble material.

For the production of *alum crystals*, the alum melt is brought into a stirred crystallization bath, where it is cooled to ca. 40 °C by air bubbles. The pulp is then separated from the mother liquor by centrifuging and washed. The mother liquor is collected and returned to the copper container. The potassium alum is dried at 50 –

Table 3. Solubility of some aluminum alums as a function of temperature (grams of anhydrous salt per 100 g water)

t, °C	0	10	20	30	40	50	60	70	80	90	100
Sodium	56.2	60.5	61.5	66.7							
Potassium	3.0	4.0	5.9	7.9	11.7	17	25	40	71	109	154
Ammonia	2.6	4.5	6.6	9.0	12.4	15.9	21.1	26.9	35.2	50.3	70.8
Rubidium	0.71	1.09	1.40		3.1	4.98			21.6		
Cesium	0.2	0.3	0.4		0.8	1.24			5.3		22.8
Thallium	3.05	4.4	6.2		12.6		26.0				

60 °C. The salt, which forms lustrous crystals, is sieved and packed in paper bags that are lined with polyethylene.

A process for obtaining potassium alum from alum-containing ores is given in [15], and the production of basic potassium and sodium alums from synthetic alum is described in [25].

Uses. As already mentioned, the industrial importance of potassium alum has declined considerably. However, it is still employed in tawing skins, as a mordant in dyeing, and as a coagulating agent for latex. Because of its astringent and protein-precipitating properties, potassium alum is used in the pharmaceutical and cosmetics industries. The use of potassium alum as a styptic pencil, because of its blood-stanching property, is very widespread and popular. The most important application today is in the gypsum industry, which employs potassium alum as hardening agent and setting accelerator for the production of marble cement and alabaster plaster. However, for purifying water and sizing paper, potassium alum has been replaced completely by aluminum sulfate. In the paper industry, aluminum sulfate is designated traditionally, although incorrectly, as an alum.

1.2.2. Ammonium Aluminum Sulfate

Ammonium aluminum sulfate [*7784-26-1*], also called ammonium alum, $NH_4Al(SO_4)_2 \cdot 12\ H_2O$, M_r 453.33, ϱ 1.64 g/cm^3, *mp* 93.5 °C, occurs in nature as shermigite. Crystals doped with other alums show birefringence. The solubility of ammonium alum in water is similar to that of potassium alum, with which it forms a continuous series of mixed crystals. Ammonium alum is slightly soluble in dilute acids and glycerol but insoluble in absolute alcohol. In aqueous solution, ammonium alum is neutral.

Data on loss of water on heating are in disagreement. Gel'Perin and Chebotkevich [26] have reported that water of crystallization is released in three stages: first to give the hydrate with 21 mol water, then to that with 3 mol water, and finally to the anhydrous product, so that it is in fact more correct to formulate ammonium alum with 24 mol water. Above 193 °C, decomposition with the release of ammonia begins. On glowing above 1000 °C, sulfur trioxide is lost, leaving an aluminum oxide residue.

Production. Ammonium alum is made mainly by dissolving aluminum hydroxide in sulfuric acid and adding ammonium sulfate but sometimes by the reaction of ammonia gas with aluminum sulfate and sulfuric acid. The procedure is analogous to the potassium alum process. To obtain a quality particularly low in iron oxide or other metal oxides (< 0.0001 % Fe_2O_3; normal quality ca. 0.001 % Fe_2O_3), as is required for production of synthetic gems, very pure starting materials must be used. Ammonium alum occurs as an intermediate in the "aloton" process, which operates in a sulfuric acid medium. This process had some importance prior to 1945 in Germany and the United States for aluminum hydroxide and aluminum oxide production (→ Aluminum Oxide)

Uses. In Europe, ammonium alum is not used in large quantities. Applications include those in dressing furs in tanning, in the production of very fine aluminum oxide particles for polishing metallographic surfaces, and, in some countries outside Europe, as a disinfectant. In the United States, ammonium alum is important as an additive in baking powder; ca. 500 t/a are produced for this purpose. More recently, ammonium alum has gained considerable importance as starting material for the production of the finely powdered, loose aluminum oxide of high purity that is required for synthesizing corundum gems, such as rubies and sapphires. This quality

of aluminum oxide is made by heating ammonium alum (or ammonium aluminum selenate alum) to 1000 °C.

1.2.3. Sodium Aluminum Sulfate

Sodium aluminum sulfate [*7784-28-3*], sodium alum, $NaAl(SO_4)_2 \cdot 12\ H_2O$, melts at 61 °C in its water of crystallization. In nature, it occurs as the mineral mendozite. Again, the data on thermal dehydration disagree. Sodium alum is insoluble in absolute alcohol but is much more soluble in water than all other alums. Alum powder (very fine crystalline form) is not available commercially. These two facts make it difficult to obtain sodium alum free from iron. Because of this and also because of its strong tendency to age, sodium alum has never gained the same importance as the other alums. In Europe, its use has been abandoned. In the United States, however, sodium alum is still utilized in relatively large quantities (ca. 3000 t/a) in baking powder.

Production. In the United States, sodium alum is produced by adding a clear solution of sodium sulfate to aluminum sulfate. After dilution to 30 °Bé and subsequent heating, a sludge of potassium sulfate, sodium silicate, and caustic soda is added to improve the purity of the product. The mixture is pumped into a stirred vessel, where it is mixed for several hours. During this stage, the ratio of aluminum sulfate to sodium sulfate is adjusted to the stoichiometric amount. Afterwards, the melt is pumped into an evaporator and concentrated to such an extent that it solidifies to a hard cake on pouring into a cooling tank. This sodium alum cake is then heated and finally ground to the desired size (99 % through a sieve of 100 mesh) [17, p. 250].

2. Aluminates

Only the aluminates of barium and sodium have importance in industry. Aluminates occur also in cement and in spinels.

2.1. Sodium Aluminate

Sodium aluminate [*1302-42-7*] is an important commercial inorganic chemical. It functions as an effective source of aluminum hydroxide for many industrial and technical applications. Commercial grades of sodium aluminate are available in solid and liquid forms. Pure sodium aluminate (anhydrous) is a white crystalline solid having a formula variously given as $NaAlO_2$, $Na_2O \cdot Al_2O_3$, or $Na_2Al_2O_4$. Commercial grades, however, always contain more than the stoichiometric amount of Na_2O, the excess being on the order of 1.05 to 1.50 times the formula requirement. Hydrated forms of sodium aluminate are crystallized from concentrated solutions.

Sodium aluminate has no defined melting point; it softens above 1700 °C when sodium begins to evaporate slowly, leaving aluminum oxide.

Production. The chief commercial process for the manufacture of sodium aluminate is the dissolution of aluminum hydroxides in sodium hydroxide solution. Aluminum trihydroxide (gibbsite) from the Bayer process (→ Aluminum Oxide) can be dissolved in 10 – 30 % aqueous NaOH solution at a temperature near the boiling point.

The use of more concentrated NaOH solutions leads to a semisolid product. The process is carried out in steam-heated vessels of nickel or steel, and the aluminum hydroxide is boiled with ca. 50 % aqueous sodium hydroxide until a pulp forms. After this is poured into a tank and cooled, a solid mass containing about 70 % $NaAlO_2$ is formed. After being crushed, this product is dehydrated in a rotary oven heated either directly or indirectly by burning hydrogen. The resulting product contains ca. 90 % $NaAlO_2$ and 1 % water, together with 1 % free NaOH. The solubility of the salt produced in this way depends strongly on how much excess sodium hydroxide is used.

Alternatively, bauxite can be used directly as the alumina source. Bauxites containing gibbsite are extracted at 150 °C and 5 bar, whereas boehmite containing bauxite requires higher temperatures (230 °C) and pressures. The sodium aluminate solution obtained from the digestion process is separated from any impurities and then concentrated to the commercial liquid grade by evaporation. The solid product is obtained by drying the liquid.

The sinter method also has been used to produce sodium aluminate. By sintering sodium carbonate directly with Bayer aluminum trihydroxide in rotary sintering kilns at 1000 °C, an essentially anhydrous product can be obtained. When sintering bauxite, it is essential to leach the sinter mass with water and separate the impurities.

Uses. The major use of sodium aluminate is for water treatment, including both potable and industrial waters. Sodium aluminate is used as an adjunct to water softening systems, as a coagulant aid to improve flocculation, and for removing dissolved silica. Sodium aluminate dissolves in water to give a solution that has a pH value of 8. The aluminum hydroxide that precipitates from this solution has excellent flocculation properties and coagulates other impurities present in the water. Depending on the impurities, the conditions of precipitation can be improved by adding aluminum sulfate [27].

In *construction technology*, sodium aluminate is employed to accelerate the solidification of concrete, mainly when working during frost, under water, and in humid soil (see Table 4). Compared to calcium chloride, which acts in a similar manner, sodium aluminate has the advantage of greater setting acceleration at low water-to-cement ratios [28]. Because it does not attack the reinforcing metals, it can be used for work with steel-reinforced concrete. Too rapid setting can cause cracks in the concrete, affecting the final strength of the building. By adding substances such as oxoacids [29], sugar, naphthene [30], K_2CO_3, or Na_2SO_4 [31] the final strength increases without affecting the setting rate significantly. Sodium aluminate increases the resistance of mortar to water, alkali, and acid. For the production of expanded concrete, sodium aluminate serves to activate nitrogen-releasing substances [32]. To stabilize foam formation in the production of light refractory bricks, up to 5 % (relative to the total dry mass) sodium aluminate is added.

Increasingly, sodium aluminate is being used for the production of *synthetic zeolites*, used as catalyst supports or as catalysts [33 – 36] and as adsorbents. Sodium aluminate is one of the principal sources of *alumina* of the preparation of alumina adsorbents and catalysts.

Table 4. Acceleration of setting by sodium aluminate in a mixture of 6 parts sand, 2 parts concrete, and 1 part water

% $NaAlO_2$ in concrete	Setting started	Setting completed
0	4 h 30 min	6 h
1	30 min	1 h 40 min
1.5	15 min	30 min
2	10 min	15 min
10	instantly	

In the *paper industry*, sodium aluminate increases the opacity [37], retention of fibers and filling materials [38], and the paper strength [39]. Also, it stabilizes the pH value of the water circulation [40] and increases the dispersion stability of titanium dioxide [41].

Another application of sodium aluminate is as a *pickle* to protect metallic surfaces (copper [42], aluminum [43], among others). This effect is based on the formation of a thin, very firmly adhering layer of aluminum hydroxide or, after heat treatment, of aluminum oxide on the surface, which does not affect the metallic luster. In the *enamel industry*, sodium aluminate is added to enamel mixtures to achieve low-melting covers [44].

Further applications of sodium aluminate are in lithography for the production of printing inks [45] and print forms [46], in the detergent and varnish industries, in the production of water-insoluble floor waxes [47], as a dispersion agent for production of high-purity asbestos [48], as an additive to drilling fluids [49], and in the textile industry as a mordant in dyeing and in printing cloth. Many other uses of sodium aluminate have been reported. These include inhibition of glass etching by alkaline solutions; protection of steel surface during galvanizing; improving dyeing, antipiling, and antistatic properties of polyester synthetic fibers; as an additive to foundry sand molds and cores; and as a binder in the ceramics industry.

2.2. Barium Aluminates

The industrially important barium aluminates are $BaO \cdot 6\ Al_2O_3$ [*12254-17-0*], *mp* 1915 °C; $BaO \cdot Al_2O_3$ [*12004-04-5*], *mp* 1815 °C; and $3\ BaO \cdot Al_2O_3$ [*12004-05-6*], *mp* 1425 °C [50]. The first two crystallize hexagonally.

Barium aluminates are produced by melting bauxite with coal and barite. Leaching the melt gives a solution of barium aluminate, from

which the salt is obtained by evaporation. Very pure barium aluminates can be produced by sintering mixtures of aluminum oxide with barium carbonate. All barium aluminates, including such hydrated compounds as $BaO \cdot Al_2O_3 \cdot 4\,H_2O$, $2\,BaO \cdot Al_2O_3 \cdot 5\,H_2O$, $BaO \cdot Al_2O_3 \cdot 7\,H_2O$, and $7\,BaO \cdot 6\,Al_2O_3 \cdot 36\,H_2O$, hydrolyze in water, forming hydrargillite (gibbsite), $Al(OH)_3$, and relatively soluble barium hydroxide.

Barium aluminate is used for purifying water because the Ba^{2+} ion precipitates both the sulfates and the carbonates, whereas the aluminate ions form insoluble calcium aluminate. A further use of barium aluminates is as a special cement, for example, as binding agent for the production of high-temperature refractories, and in the production of radiation shields [51].

3. Aluminum Alkoxides

Aluminum alkoxides (alcoholates) are solid or liquid compounds of covalent character. They are readily soluble in hydrocarbons, but sparingly soluble in alcohols. Hydrolysis with water occurs readily, giving aluminum hydroxide and the corresponding alcohol. For a general review of aluminum alkoxides, see [52]. Industrially, only the isopropoxide (isopropylate) and *sec*-butoxide (*sec*-butylate) are important. These compounds are used to adjust the viscosity of varnishes, to impregnate textiles, as intermediates in the production of pharmaceuticals, and as antitranspirants in cosmetics. In the industrial production of ketones and aldehydes, aluminum alkoxides are employed as reducing agents (Meerwein-Ponndorf reaction).

Aluminum isopropoxide [*555-31-7*], $Al(OCH(CH_3)_2)_3$, is a white solid, M_r 204.25, *mp* 118 °C, *bp* 125 – 130 °C at 533 Pa, d_4^{20} 1.0346, flash point 26 °C. It is usually produced by direct reaction of aluminum and isopropyl alcohol in the presence of mercury(II) chloride catalyst. In a later version of the process [53], the alcohol was heated under reflux in a column filled with aluminum chips; no catalyst was required and nearly quantitative conversion was obtained. The alkoxide can be purified by distillation. In an alternative production method, excess isopropyl alcohol is added to a solution of aluminum chloride in benzene; the hydrogen chloride formed is removed by introducing dry ammonia into the reactor and filtering off the ammonium chloride that precipitates.

Aluminum *sec*-butoxide [*2269-22-9*], $Al(OC_4H_9)_3$, is a colorless liquid, M_r 246.3, *bp* 180 °C at 533 Pa, d_4^{20} 0.9671, flash point 26 °C, and is produced in a manner similar to aluminum isopropoxide.

4. Aluminum Chloride

4.1. Anhydrous Aluminum Chloride

Oerstedt first prepared anhydrous aluminum chloride [*7446-70-0*] in 1825 by the reaction of chlorine gas with a mixture of alumina and carbon. This compound has acquired great significance in organic chemistry as a catalyst, particularly for Friedel-Crafts syntheses and allied reactions for the production of alkylated aromatics, dyestuffs, pharmaceuticals, and perfumery chemicals. For general literature, see [1] and [54].

4.1.1. Properties

Physical Properties. In the solid and gas phases at temperatures up to 400 °C, aluminum chloride is present as the dimer, Al_2Cl_6. Dissociation of the dimer progresses with rising temperature and is quantitative above 800 °C. Solid aluminum chloride crystallizes to form a monoclinic layer lattice. Pure aluminum chloride is white, but the commercial product usually has a yellowish or grayish tinge because of small amounts of iron chloride or aluminum impurities. Physical properties of anhydrous aluminum chloride [55]:

M_r	133.34
ϱ at 25 °C	2.44 g/cm^3
Sublimation temperature at 101.3 kPa	181.2 °C
Triple point, at 233 kPa	192.5 °C
Heat of formation at 25 °C	–705.63 ± 0.84 kJ/mol
Heat of sublimation of dimer at 25 °C	115.73 ± 2.30 kJ/mol
Heat of fusion	35.35 ± 0.84 kJ/mol
Heat of solution at 20 °C	–325.1 kJ/mol

Chemical Properties. Anhydrous aluminum chloride reacts extremely violently with water, evolving hydrogen chloride. The hexahydrate [*7784-13-6*], $AlCl_3 \cdot 6\,H_2O$, M_r 241.44, is formed in this reaction. In aqueous solution, aluminum chloride is partially hydrolyzed to hydrochloric acid and aluminum oxychloride, AlClO [*13596-11-7*]. For this reason, anhydrous aluminum chloride cannot be obtained by concentrating the solution and calcining the hydrate. When aluminum chloride is heated with γ-alumina, aluminum oxychloride, AlClO, forms, but this reaction goes in the reverse direction at temperatures above 700 °C:

$$AlCl_3 + Al_2O_3 \rightleftharpoons 3\,AlClO$$

If aluminum chloride vapor is passed at 1000 °C under reduced pressure over molten aluminum, volatile aluminum monochloride [*13595-81-8*], AlCl, is formed, but this decomposes immediately into the elements in cooler zones of the reactor. This method has been adopted for purifying aluminum.

Reaction between aluminum chloride and other metal halides, such as $CaCl_2$, $CrCl_3$, and $FeCl_3$, gives mixed halides. Eutectic melts with other metal chlorides are of industrial significance, for example, that with sodium chloride is used as a solvent in chlorination reactions.

Anhydrous aluminum chloride dissolves readily in polar organic solvents. As a Lewis acid it forms addition compounds with numerous electron donors, such as hydrogen chloride, hydrogen sulfide, sulfur dioxide, sulfur tetrachloride, phosphorus trichloride, ethers, esters, amines, and alcohols.

4.1.2. Production

The starting materials are either aluminum or pure aluminum oxide. Bauxite now has no economic significance as a raw material because of the iron chloride always present in the product.

Chlorination of Aluminum [56]. Today, most anhydrous aluminum chloride is made by chlorinating aluminum.

Chlorine is passed through molten aluminum in ceramic-lined, tube-shaped reactors. The reaction is highly exothermic:

$$2\,Al\,(s) + 3\,Cl_2\,(g) \rightarrow Al_2Cl_6\,(s) \quad \Delta H^\circ = -1411\,kJ/mol$$

The temperature in the reactor is maintained at 670 – 850 °C by controlling the admission rates of chlorine and aluminum and by cooling the reactor walls with water. The aluminum usually is replenished in the form of lumps. The difficulty of controlling the large heat of reaction can be overcome also by dividing the process into a number of small units.

The aluminum chloride vapor leaving the reactors is passed through ceramic-lined tubes into large, air-cooled iron chambers. Solid aluminum chloride is withdrawn from the condenser walls at regular intervals, ground (insuring exclusion of moisture), and classified by sieving. Chlorine in the off-gas is removed by conventional methods, such as absorption in caustic soda solution.

Chlorination of Pure Aluminum Oxide. Chlorination of alumina is advantageous over the formerly widely used bauxite process (less reactor corrosion and higher-purity product), and at the same time it avoids the high raw material costs involved in metal chlorination [57].

Carbon monoxide and chlorine are partially converted to phosgene over an activated charcoal catalyst. The gaseous reaction mixture enters a brick-lined, fluidized-bed reactor, where it reacts with finely divided γ-alumina to yield aluminum chloride. The reaction is exothermic enough (ca. 300 kJ/mol, based on $AlCl_3$) to maintain the temperature at 500 – 600 °C without external heat. Consequently, the process permits large units, and the low reaction temperature insures that the brick lining has a long life.

The aluminum chloride vapor is filtered through a bed of coarse pumice chips, condensed, and further processed as described above. The off-gas contains chlorine, phosgene, and large amounts of carbon dioxide. The chlorine is removed by scrubbing, and the phosgene is hydrolyzed with water. Extremely pure aluminum chloride is obtained by resublimation from molten sodium aluminum chloride.

4.1.3. Quality Specifications and Analysis

Anhydrous aluminum chloride is ground and marketed as powder or granules. Aluminum is

usually determined by atomic spectrometry, the chloride by argentometry. The assay is 98 – 99 % $AlCl_3$. The main impurity is iron (0.05 wt % max. or 0.01 wt % for resublimed product). Sampling and analysis of anhydrous aluminum chloride must be carried out in an atmosphere of dry air or nitrogen.

4.1.4. Handling, Storage, and Transportation

Because of its corrosive and irritant action, anhydrous aluminum chloride is classified as a dangerous substance. National and international regulations must be observed, such as 67/548/EEC, the key European Economic Community directive dealing with the classification, packaging, and labeling of dangerous substances.

Handling. Goggles, gloves, and protective clothing must be worn when handling aluminum chloride. A fume cupboard or a respirator with filter type B/St against acid gases should be used. Because hydrogen chloride is evolved when aluminum chloride is exposed to water, any spills must be taken up dry, and only small, residual amounts can be washed away with plenty of water. Sodium bicarbonate or slaked lime should be used for neutralization.

Storage and Transportation. Aluminum chloride is dispatched in vented steel drums or in tank trucks or rail tankcars that can be emptied pneumatically with dry air or nitrogen (dew point below − 40 °C). Because the product tends to cake, it should not be stored for more than six months. International marine transportation is governed by the IMDG code, class 8, UN no. 1726. RID, ADR, ADNR: Class 8, no. 11b, Rn 801, 2801, 6801, respectively. EEC: Yellow Book 78/79, EG no. 013−003−00−7. UK: Blue Book, Corrosives & IMDG code E 8031. USA: DOT regulations, corrosive solid, CFR 49, 172.101.

4.1.5. Uses

Anhydrous aluminum chloride is an important *Friedel-Crafts catalyst* in the chemical and petrochemical industries [58, 59]. A principal application is the alkylation of benzene by alkyl halides to form alkylbenzenes that are consumed in the production of synthetic detergents, such as alkylbenzenesulfonates. Aluminum chloride also catalyzes the liquid-phase ethylation of benzene with ethylene to yield ethylbenzene, most of which is used in the production of styrene.

Ethyl chloride is produced mainly by the reaction of hydrochloric acid and ethylene in the presence of aluminum chloride; the consumption has significantly decreased because of the declining demand for tetraethyl lead as an antiknock additive.

In the dyestuffs industry, aluminum chloride is widely used as catalyst in the production of anthraquinone and its derivatives and of pigments, such as phthalocyanine green. A further application of aluminum chloride is as a finish for titanium dioxide pigments, which are in this way protected from oxidation by an oxide layer.

Until 1960, aluminum chloride was used extensively in petroleum refining for cracking and isomerization, but it has now been replaced by zeolite catalysts. It is still employed in reforming hydrocarbons, as a polymerization catalyst in the production of hydrocarbon resins, and for the production of butyl rubber.

It is a catalyst also in the production of compounds, such as aromatic aldehydes, ketones, and 2-phenyl-ethanol, used in fragrances. Other applications for anhydrous aluminum chloride include the production of aluminum borohydride, lithium aluminum hydride, as well as compounds of phosphorus and sulfur.

Most of these reactions with aluminum chloride produce a solution that is often used as a flocculating agent in the treatment of waste water.

Aluminum chloride also is an intermediate in the production of aluminum by the Alcoa smelting process [60] (→ Aluminum).

Data on the consumption of anhydrous aluminum chloride in the USA are given in Table 5.

Table 5. U.S. anhydrous aluminum chloride consumption in 1993 (not including the amounts consumed in the production of aluminum) [61]

Alkylate detergents	2200 t
Ethylbenzene	4000 t
Hydrocarbon resins	3900 t
Pharmaceuticals	4000 t
Titanium dioxide	2200 t
Miscellaneous	4900 t
Total	21200 t

4.2. Aluminum Chloride Hexahydrate

Properties. The white hydrate $AlCl_3 \cdot 6\ H_2O$ [*7784-13-6*], M_r 241.43, crystallizes hexagonally and dissolves exothermally in water (133 g hexahydrate per 100 g water at 20 °C) [62]. The solubility increases only slightly with temperature. The aqueous solutions are strongly acidic because the salt hydrolyzes. Hydrogen chloride is evolved on heating concentrated solutions. The structure of aqueous solutions of $AlCl_3 \cdot 6\ H_2O$ is discussed in [63].

Production and Uses. Aqueous solutions of $AlCl_3 \cdot 6\ H_2O$ can be obtained by dissolving aluminum hydroxide in hydrochloric acid: the hexahydrate crystallizes when hydrogen chloride passes into cold (about 20 °C) saturated solution. In this way, impurities, particularly iron (III) chloride, can be removed [62]. Aluminum chloride hexahydrate has only minor importance as such; for example, it is used for hardening photographic layers. However, it is of industrial importance as an intermediate in the production of aluminum oxide [64, 65].

4.3. Basic Aluminum Chlorides

Properties. Basic aluminum chlorides [*1327-41-9*], aluminum hydroxychlorides, aluminum chloride hydroxides, poly(aluminum hydroxychlorides), have the general formula $Al_2(OH)_{6-n}Cl_n \cdot x\ H_2O$. The individual compounds are best defined either by the molar ratio of aluminum to chloride ($2/n$) or by their basicity, defined as $(1 - n/6) \cdot 100\,\%$. Below a minimum water content, which depends on n, the compounds are unstable; they decompose to aluminum hydroxide and aluminum oxide on heating to above 80 °C, releasing water and hydrogen chloride. All basic aluminum chlorides with $n = 1 - 5$ can be isolated, and all are white, in some cases crystalline substances, partially soluble in the lower alcohols [66, 67], but readily soluble in water. For example, 170 g of the compound $Al_2(OH)_5Cl \cdot 2.5\ H_2O$ dissolves in 100 g water at 20 °C. However, the viscosity of the solution prevents further amounts from dissolving. The *electrical conductivities* of the aqueous solutions, shown in Figure 2, are characteristic of the basic aluminum chlorides.

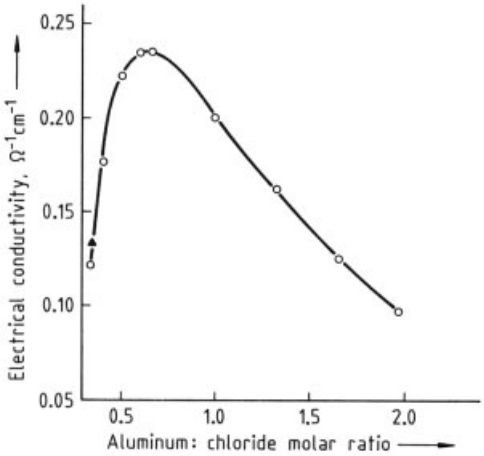

Figure 2. Electrical conductivity of freshly prepared basic aluminum chloride solutions: ○) from Locron L (ca. 50 % aqueous solution of $Al_2(OH)_5Cl \cdot 2.5\ H_2O$) and hydrochloric acid at constant aluminum oxide concentration (11.5 %) and 20 °C; ▲) prepared directly from $AlCl_3 \cdot 6\ H_2O$

Aqueous solutions of basic aluminum chlorides are acidic because the compounds hydrolyze (see Table 6). The stability of the solutions depends on the basicity: compounds with aluminum to chloride molar ratios of ca. 0.5 : 1 to about 1.6 : 1, corresponding to basicities of ca. 30 – 75 %, decompose to give insoluble basic aluminum chloride at a rate that depends on the temperature and concentration of the solution [68]. On the other hand, compounds of basicity greater than 75 % are very stable in aqueous solution. For example, compounds with the approximate composition $Al_2(OH)_5Cl \cdot 2 - 3\ H_2O$ [*12042-91-0*] are so stable at concentrations > 50 % that no decomposition occurs, even on boiling for days.

Table 6. pH values of aqueous solutions of $Al_2(OH)_5Cl \cdot 2.5\ H_2O$ at 20 °C (mass fractions, w, in %)

w	pH
60	3.4
30	4.0
15	4.2
10	4.3
5	4.4
3	4.4

In contrast to the aluminum chlorides of low basicity (below 65 %), the compounds of high basicity (above 65 %) do not crystallize from their aqueous solutions, but form

glassy masses. When solutions in the basicity range 30 – 65 % are evaporated, the compound $Al_2(OH)_{3.7}Cl_{2.3} \cdot 6.05\ H_2O$ (basicity 62 %) always crystallizes [69]. On addition of alkali to basic aluminum chlorides, aluminum hydroxide precipitates. Hydrochloric acid converts all basic aluminum chlorides to the hexahydrate:

$$Al_2(OH)_5Cl + 5\ HCl + 7\ H_2O \rightarrow 2\ AlCl_3 \cdot 6\ H_2O$$

Chemical Structure. Basic aluminum chlorides have been known for a long time [1, p. 205], but only recently have investigations given significant insight into their chemical structures [70 – 73]. Basic aluminum chlorides are mixtures of complex compounds of various degrees of polymerization. Spectroscopic and kinetic investigations of the basic aluminum chlorides and of their solutions [70 – 73] lead to the conclusion that the complex ion $[Al_{13}O_4(OH)_{24}(H_2O)_{12}]^{7+}$ is present and is in equilibrium with its polymeric forms. Variations observed in the properties of these aqueous solutions such as viscosity and pH are caused by polymerization and depolymerization [74].

Analysis. Basic aluminum chlorides are so stable that the aluminum content can be determined only after decomposition of the complex ion [75, 76]. The usual procedure is to mix the solution containing aluminum with 60 % sulfuric acid and 0.1 M disodium ethylenediaminetetraacetate, followed by ca. 20 % sodium hydroxide. The pH is adjusted to 4.7 – 4.9 at 60 – 70 °C. After the solution is cooled to about 25 °C and buffer solution (acetic acid, ammonium acetate), acetone, and indicator (dithizone in ethanol) are added, the solution is backtitrated with 0.1 M zinc sulfate. One milliliter of 0.1 M disodium ethylenediaminetetraacetate is equivalent to 2.698 mg of aluminum.

Production. Basic aluminum chlorides are made by dissolving either aluminum hydroxide or metallic aluminum in hydrochloric acid. Aged aluminum hydroxide leads only to compounds of basicity up to 65 % [68]. To obtain aluminum chlorides of high basicity, freshly precipitated aluminum hydroxide [77] must be used. However, the preferred method is to dissolve aluminum in hydrochloric acid either thermally or electrochemically:

$$2\ Al + HCl + 5\ H_2O \rightarrow Al_2(OH)_5Cl + 3\ H_2$$

This method is exemplified by the following process [78]:

Aluminum electrodes are set up at distances of 40 mm in a corrosion-resistant electrolysis cell. Hydrochloric acid (ϱ 1.04 g/cm^3) is poured into the cell at a rate such that the reaction temperature remains at ca. 80 °C. After all the hydrochloric acid has been added, the voltage between the first and last electrodes is set so that an average voltage of 0.6 V per electrode pair is obtained at a current density of ca. 170 A/m^2. The formation of explosive gas mixtures must be avoided throughout the reaction by diluting the hydrogen with air or nitrogen to below the explosion limit. After about 70 h, the density of the electrolyte increases to about 1.2 – 1.3 g/cm^3, and the pH value to about 3.5. At the same time the cell voltage falls. The electrolyte then consists of a solution containing aluminum and chloride in the molar ratio 2 : 1, corresponding to a basicity of 83 %. The corresponding bromides as well as iodides are obtained by the same method. The solid halides are obtained from the solutions by careful evaporation or by spray drying [79]. Recently another method may be of interest: Controlled thermic decomposition of $AlCl_3 \times 6\ H_2O$ [86].

Uses and Quality. The most important basic aluminum chloride with regard to consumed quantities are the polyaluminum chlorides (PAC) with basicities in the range of 35 – 60 %. PAC is widely used as a flocculating agent for purifying water. In water treatment PAC can be used for purifying surface water, sewage, and wastewater from industry and municipalities and for water treatment in swimming pools. A second important application for PAC is as fixing sizing agent in the paper industry. Because of the change from acidic to neutral or alkaline sizing, large quantities of aluminum sulfate had to be substituted by PAC. Various qualities are available with aluminum contents between 7.5 and 18.0 %, calculated as aluminum oxide.

Another important basic aluminium chloride has the composition $Al_2(OH)_5Cl \cdot 2 – 3\ H_2O$ and is known in the literature as aluminum chlorohydrate. Very stringent purity specifications apply to this substance. The maximum levels permitted are Fe 100 ppm, SO_4 500 ppm,

Pb 20 ppm, As 3 ppm. These are required for one of its major applications, namely, that in the cosmetics industry as the effective component of antiperspirants [80, 81]; trade names are Chlorhydrol (Reheis, USA) and Locron (Clariant). Other important areas of application for this aluminum chloride are in the production of catalysts and highly temperature-resistant fibers based on Al_2O_3, as a hydrophobic agent for impregnation of cotton textiles, for tanning leather, as a retention agent in paper production, as a binding agent for fire-resistant ceramic products as a hardening agent for rapid fixing baths in the photographic industry and as a flucculant for purifying swimming-pool water.

5. Toxicology

Aluminum sulfate is classified as nontoxic; an LD_{50} value of 6207 mg/kg (mouse, oral) is reported [82]. *Alum* solutions are known for their astringent (tissue-contracting) effects. A TLV of 2 mg/m^3 was established for water-soluble aluminum salts [83].

Anhydrous aluminum chloride has a corrosive and irritant effect on the skin, respiratory tract, and eyes. The toxic effect is caused by the evolution of hydrochloric acid when the product is exposed to moist air. The resulting respiratory difficulties vary from mild irritation to coughing. The MAK value (1995) and TLV (1983) for hydrochloric acid are both 7 mg/m^3 [83, 84].

The extensive use of the *basic aluminum chloride* $Al_2(OH)_5Cl \cdot 2-3\,H_2O$ in the cosmetic industry worldwide over more than 30 years has shown that this compound is not irritating to the skin in the concentrations used commercially. Although intermittent application of 150 mg $Al_2(OH)_5Cl \cdot 2-3\,H_2O$ over a period of 3 days was found to cause mild skin irritation in humans, a dose of only 7.5 mg of $AlCl_3 \cdot 6\,H_2O$ caused the same degree of irritation [85].

6. References

1. *Gmelin*, System No. 35 "Aluminium," part B.
2. N. A. Lange: *Handbook of Chemistry*, 10th ed., McGraw-Hill, New York 1961.
3. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Metal Organic Compounds*, 4th ed., **vol. 1**, Van Nostrand, Princeton, N.J., 1958.
4. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker*, 3rd. ed., vol. 1, Springer Verlag, Berlin 1967.
5. J. Kaltenbach, *Papier (Darmstadt)* **17** (1963) 14.
6. N. O. Smith, P. N. Walsh, *J. Am. Chem. Soc.* **76** (1954) 2054.
7. J. L. Henry, G. B. King, *J. Am. Chem. Soc.* **72** (1950) 1282.
8. P. Barret, R. Thiard, *C. R. Hebd. Seances Acad. Sci.* **260** (1965) 2823.
9. H. A. Papazian, P. J. Pizzolato, R. R. Orrell, *Thermochim. Acta* **4** (1972) 97.
10. Giulini, DE 1125894, 1960 (J. Rüter, H. Cherdron);
    DE 1146042, 1960 (J. Rüter, F. Fäßle).
11. *Ullmann*, 3rd ed., **vol. 3**, p. 425.
12. H. Kretschmar, *Aluminium (Düsseldorf)* **39** (1963) 624;
    US 2951743, 1957.
13. E. A. Gee, W. K. Cunningham, R. A. Heindl, *Ind. Eng. Chem.* **39** (1947) 1178.
14. *Chem. Eng. (N.Y.)* **68** (1961) 36.
15. G. B. Skakhtakhinskii, A. N. Khalilov, *Mater. Konf. Molodykh Uch. Inst. Neorg. Fiz. Khim. Akad. Nauk Az. SSR* 1968, 34;
    Chem. Abstr. **74** (1971) 101 183.
16. Jui-Hsiung Tsai, US 3574537, 1967.
17. *Kirk-Othmer*, 3rd ed., **vol. 2**.
18. R. G. Lebel, *Pulp Pap. Mag. Can.* **71** (1970) 23.
19. K. E. Oehler: "Technologie des Wassers," in *Handbuch der Lebensmittelchemie,* vol. VIII, part 1: Wasser und Luft, Springer Verlag, Berlin 1969.
20. Boliden, SE 341381, 1971.
21. *Chem. Eng. News* **61** (1983) 32.
22. Chemfacts Japan, Chemical Data Series, IPC Industrial Press, Sutton, Surrey UK 1981.
23. M. M. Kazakov et al., *Uzb. Khim. Zh.* **16** (1972) no. 2, 10;
    Chem. Abstr. **77** (1972) 50847.
24. A. Mitscherlich, *J. Prakt. Chem.* **83** (1861) 477.
25. G. L. Akhmetora et al., *Tr. Khim. Metall. Inst. Akad. Nauk Kaz. SSR* **15** (1970) 110.
26. N. I. Gel'Perin, G. V. Chebotkevich, *Zh. Prikl. Khim. (Leningrad)* **44** (1971) 420.
27. I. Kh. Kovarskaya, *Nauch. Issled. Inst. Akad. Kommunal Khoz.* **3** (1967) 77 – 87.
28. M. Venuat, *Mater. Constr. Trav. Publics* 1967, no. 626 – 672, 394 – 446.
    Ind. Liants Hydrauliques, Paris.

29. Progil, US 3433657, 1964; FR 6941443, 1969.
30. P. P. Stupaschenko, E. P. Kholoshin, *Izv. Vyssh. Uchebn. Zaved. Stroit. Arkhit.* **11** (1968) no. 7, 105 – 108.
31. Promstroiniiprockt Inst. Rostov, SU 229889, 1967.
32. Winkler, Kaspar & Co., ZA 6801274, 1968.
33. FMC Corp., GB 1136688, 1967.
34. Esso Research & Eng. Co., US 3380915, 1965.
35. Eastman Kodak, US 3285863, 1964.
36. Office National Industriel del'Azote, FR 1433595, 1965.
37. E. Hechler, *Wochenbl. Papierfabr.* **96** (1968) no. 23/24, 868 – 872.
38. *Abstr. Bull. Inst. Pap. Chem.* **37** (1967) no. 7, 415 – 416.
39. Nalco Chemical Co., US 3264174, 1962.
40. W. Vogel: *Einsatz von Natriumaluminat im Wasserkreislauf von Papierfabriken*, PTS-WAF-Seminar, München, 4 – 6 Nov. 1980.
41. Cabot Corp., US 3212911, 1961.
42. Imperial Metal Ind. (Kynoch), FR 1507547, 1967.
43. Compagnie Générale du Duralumin et du Cuivre, FR 1470389, 1966.
44. T. Chmiel et al., PL 55018, 1966.
45. Eastman Kodak, FR 1516507, 1968.
46. Aluminum Co. of America, US 3290151, 1963; CS 126439, 1963.
47. American Cyanamid, US 3328328, 1964.
48. Union Carbide, US 3297516, 1963.
49. VEB Cerital-Werk Mieste, NL 6503849, 1965.
50. G. Purt: "Binary System $BaO \cdot Al_2O_3$," *Radex Rundsch.* **4** (1960) 201.
51. A. Branski: "Refractory Barium Aluminate Cements," *Tonind. Ztg. Keram. Rundsch.* **85** (1961) no. 6, 125 – 129.
52. D. C. Bradley: "Metal Alkoxides," in F. A. Cotton (ed.): *Prog. Inorg. Chem.* **2** (1960) 303.
53. Anderson Chemical Corp., US 2965663, 1960 (W. E. Smith, A. R. Anderson).
54. J. Andrioly, J. Enezian, *Ind. Chim. (Paris)* **55** (1968) 203 – 213.
55. *JANEF Thermochemical Tables*, NSRDS-NBS 37, 2nd ed., National Bureau of Standards, Gaithersburg, Md., 1971.
56. R. L. de Beauchamp: "Preparation of Anhydrous Aluminum Chloride," *Inf. Circ. U.S. Bur. Mines*, no. 8412, 1969.
57. J. Hille, W. Dürrwächter, *Angew. Chem.* **72** (1960) 850 – 855.
58. C. A. Thomas: *Anhydrous Aluminum Chloride in Organic Chemistry*, ACS Monograph Series, Reinhold Publ. Co., New York 1941.
59. G. A. Olah: *Friedel-Crafts and Related Reactions*, Wiley-Interscience, New York 1963.
60. H. J. Gardener, K. Grjotheim, B. J. Welch: *Alumina Production Until 2000*, ICSOBA-Hungary, Hungalu, Budapest 1981.
61. *Chemical Economics Handbook*, SRI International, Oct., 1994.
62. R. R. Brown, G. E. Daut, R. v. Mrazek, N. A. Gokcen, *Rep. Invest. U.S. Bur. Mines*, no. 8379, 1979, p. 1.
63. J. F. McIntyre, R. T. Foley, B. F. Brown, *Inorg. Chem.* **21** (1982) 1167.
64. K. B. Bengtson, 108th Annual Meeting, American Institute of Mining, Metallurgical and Petroleum Engineers, New Orleans 1979; *Light Met. (N.Y.)* 1979, vol. 1, p. 217.
65. J. Cohen, J. Mercier: Proceedings, 105th Annual Meeting, American Institute of Mining Engineering, Las Vegas 1976; *Light Met. (N.Y.)* 1976, vol. 2, p. 3.
66. Hoechst, DE 2408751, 1974 (M. Danner, K. Zeisberger).
67. Hoechst, DE 2704850, 1977 (M. Danner, K. Zeisberger).
68. Hoechst, DE 2309610, 1973 (M. Danner, M. Krieg).
69. Hoechst, DE 2907671, 1979 (M. Danner, R. Schlosser).
70. S. Schönherr, H.-P. Frey, *Z. Anorg. Allg. Chem.* **452** (1979) 167.
71. S. Schönherr, H. Görz, W. Gessner, M. Winzer, D. Müller, *Z. Anorg. Allg. Chem.* **476** (1981) 195.
72. F. v. Lampe, D. Müller, W. Gessner, A.-R. Grimmer, G. Scheler, *Z. Anorg. Allg. Chem.* **489** (1982) 16.
73. S. Schönherr, H. Görz, R. Bertram, D. Müller, W. Gessner, *Z. Anorg. Allg. Chem.* **502** (1983) 113.
74. B. Clauss, *Angew. Makromol. Chemie* **217** (1994) 139.
75. E. Wänninen, A. Ringbom, *Anal. Chim. Acta* **12** (1955) 308.
76. M. Danner, R. Schlosser, E. Hambsch (Hoechst), unpublished results, 1981.
77. C. F. Asche & Co., DE 1041933, 1957 (K. Langer).
78. Hoechst, DE 1174751, 1961 (J. König, H.-K. Platzer).
79. Hoechst, DE 2263333, 1972 (M. Danner, M. Krieg, K. Matschke).

80. J. J. Fitzgerald, A. H. Rosenberg: "Antiperspirants and Deodorants"in K. Laden (ed.): *Cosmetics Science and Technology Series*, vol. 20, 2nd ed., Marcel Dekker, New York 1999, p. 83.
81. Cosmetic, Toiletry and Fragrance Association (CTFA), Washington, D.C., 1971.
82. *Br. J. Ind. Med.* **23** (1966) 305.
83. American Conference of Governmental Industrial Hygienists (ACGIH) (ed.): *Threshold Limit Values 1982*, Cincinnati, Ohio, 1982.
84. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentrationen (MAK) 1995*, Verlag Chemie, Weinheim, Germany 1995.
85. V. A. Drill, P. Lazar (eds.): *Cutaneous Toxicity*, Academic Press, New York 1977, p. 127.
86. A.I.R. Lippewerk, DE 4202 937, 1995 (G. Kudermann, et al.).

Aluminum Compounds, Organic

Michael J. Krause, Albemarle Corporation, Baton Rouge, LA 70820, United States

Frank Orlandi, Albemarle Corporation, Baton Rouge, LA 70820, United States

Alfred T. Saurage, Albemarle Corporation, Baton Rouge, LA 70820, United States

Joseph R. Zietz Jr., Albemarle Corporation, Baton Rouge, LA 70820, United States

1. Introduction 1
2. Alkylaluminums and Derivatives . . 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.2.1. Reactions with Olefins 3
2.2.2. Reactions with Oxygen 4
2.2.3. Reactions with Metal Compounds . . 5
2.2.4. Reactions with Proton-Donating Materials 5
2.2.5. Alkylaluminums in Organic Synthesis 5
2.3. Production of Trialkylaluminums and Alkylaluminum Chlorides . . . 6
2.3.1. Hydroalumination to Produce Trialkylaluminums 6
2.3.2. Reaction of Aluminum With Alkyl Halides to Make Alkylaluminum Sesquichlorides 7
2.3.3. Reduction of Alkylaluminum Sesquichlorides to Make Trialkylaluminums 8
2.3.4. Reaction of Acids With Trialkylaluminums to Produce Alkylaluminum Chlorides 9
2.3.5. Reproportionation Reactions 9
2.3.6. Olefin Elimination and Displacement Reactions 9
2.3.7. Other Organoaluminum Conversions 10
2.4. Uses of Alkylaluminum Compounds 10
2.4.1. Stoichiometric Applications 10
2.4.2. Catalytic Applications 11
2.5. Quality Specifications 11
3. Aluminoxanes 12
3.1. Production 12
3.2. Physical and Chemical Properties . 13
3.3. Uses 13
3.3.1. Mechanism of Polymerization Using Aluminoxanes 14
3.3.2. Production of Stereoregular Polymers using Aluminoxanes 14
3.3.3. Polymers Produced by Aluminoxanes 16
3.4. Analysis and Quality Specifications 16
4. Handling, Storage, and Transportation 17
5. Waste Disposal and Environmental Protection 18
6. Economic Aspects 19
7. Toxicology and Occupational Health 19
8. References 19

1. Introduction

Compounds containing an aluminum – carbon bond are some of the most important, highest volume, highest value organometallic compounds in production. Products ultimately made using these compounds include alcohols, olefins, polyethylene, polypropylene, heat stabilizers, pharmaceuticals, agricultural chemicals, specialty polyolefins, ceramics, and electronic materials. These products are valued in the tens of billions of dollars per year.

Trialkylaluminums and triarylaluminums, R_3Al, alkylaluminum chlorides, R_nAlCl_{3-n} ($n = 1$, 1.5, 2), alkylaluminum hydrides, R_2AlH, and aluminoxanes, $(RAlO)_n$, all have major industrial uses. In addition, various aluminate complexes, $M(I)R_nAlX_{4-n}$, $M(II)[R_nAlX_{4-n}]_2$ – where M represents a metal in group 1 or 2 of the periodic table – have been made.

The organic groups, R, can also possess double or triple bonds or can contain heteroatoms. The R group can bear inorganic nucleophilic substitutents. In addition, various organic groups can bonded to the aluminum atom through oxygen, sulfur, nitrogen, and other atoms (see general references [1–8] for the complete scope of organoaluminum materials).

Organoaluminum chemistry came of age with the studies by Ziegler and asso-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_543

ciates, commencing in the early 1950s at the Max Planck Institut für Kohlenforschung, Mülheim/Ruhr, Germany. Of great technical importance was the discovery of the very economic *direct synthesis* of trialkylaluminum compounds, particularly triethylaluminum and triisobutylaluminum, from metallic aluminum, hydrogen, and the olefins ethylene and isobutylene. This process, requiring intermediate formation of aluminum hydride bonds followed by hydroalumination of the olefin, regiospecific in the case of isobutylene, made inexpensive trialkylaluminum materials available.

Other important discoveries within this group included transition-metal coordination catalysts for olefin polymerization; the olefin elimination reaction, which is the reverse of hydroalumination of an olefin; carboalumination of olefins, especially ethylene chain growth; and the controlled air oxidation of aluminum alkyls to alkoxides (for contributions of the ZIEGLER group in organoaluminum chemistry, see [4], [9], [10] → Ziegler Processes).

Other important technical developments, spurred by the availability of economic organoaluminum raw materials, included such diverse applications as alkylation of metal salts, the syntheses of olefin and olefin oligomers, and the use of these compounds as precursors to electronic and ceramic materials.

The report by KAMINSKY and SINN in 1976 that zirconocene – methylaluminoxane (MAO) catalyst systems were highly active in ethylene polymerization has led to a major expansion of research in polyolefin chemistry [11] (→ Metallocenes; → Polyolefins, Chap. 1.3.2.3.). These MAO containing catalyst systems produce a broad slate of conventional polyolefin polymers, sometimes with vastly improved properties, as well as new, exciting polyolefin materials.

2. Alkylaluminums and Derivatives

2.1. Physical Properties

Organoaluminum materials vary in consistency from colorless, mobile liquids through low-melting waxes to crystalline solids. They generally are soluble in aliphatic or aromatic hydrocarbon solvents.

Physical properties of a group of representative alkylaluminum compounds are listed in Table 1. Normal trialkylaluminum compounds exist largely as associated dimers at 25 °C (structure (**1**)). Trialkyls branched on the β carbon, such as triisobutylaluminum, are monomeric. The dialkylaluminum hydrides and amides are trimeric (structure (**2**)). Alkylaluminum chlorides, bromides and iodides are usually dimeric (structure (**3**)). Fluorides are more highly associated. Alkylaluminum alkoxides and mercaptides generally form trimers or higher molecular aggregates.

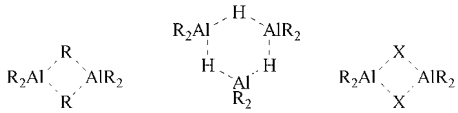

(**1**) (**2**) (**3**)

Alkyl bridges are considerably weaker than hydride, halide, or other heteroatom bridges. Dimeric trialkylaluminum compounds are in equilibrium with their monomers [12], [13]. The increased dissociation of the dimers with increasing temperature must be considered in reaction rate computations because the monomer generally is the reactive species. The carbon-aluminum bonds in trialkylaluminum compounds are dynamic, rapidly exchanging intermolecularly even under ordinary conditions, and the exchange rate increases with temperature.

2.2. Chemical Properties

The C-Al bonds of organoaluminum materials undergo a wide variety of chemical reactions. For industrial applications, reactions with olefins, metal compounds, carbonyl compounds, oxygen, and proton-donating materials are important in many of the high volume stoichiometric uses of alkylaluminums (see also [1–3], [5–8]). The Lewis acidic properties and reducing characteristics of alkylaluminums are important in high volume catalytic olefin polymerization and oligomerization applications [14]. In addition, alkylaluminums are used to perform a variety of synthetic organic transformations [3–8] and in the clean thermal decomposition to aluminum metal as a basis for metalloorganic chemical vapor deposition (MOCVD) technique

Table 1. Physical properties of alkylaluminum compounds

Compound	M_r	Density, g/mL (°C)	*mp*, °C	*bp*, °C (*p*, kPa)
Trimethylaluminum, $(CH_3)_3Al$ [*75-24-1*]	72.1	0.7478 (25)	15.3	127 (101.3)
Triethylaluminum, $(C_2H_5)_3Al$ [*97-93-8*]	114.2	0.8324 (25)	−46	185 (101.3)
Tri-*n*-propylaluminum, $(n\text{-}C_3H_7)_3Al$ [*102-67-0*]	156.2	0.8207 (25)	<−60	56 (2.7×10^{-2})
Tri-*n*-butylaluminum, $(n\text{-}C_4H_9)_3Al$ [*1116-70-7*]	198.3	0.8235 (25)	<−60	110 (0.4)
Triisobutylaluminum, $(i\text{-}C_4H_9)_3Al$ [*100-99-2*]	198.3	0.7822 (25)	1.0 (1.3×10^{-5})	58–60
Tri-*n*-hexylaluminum $(n\text{-}C_6H_{13})_3Al$ [*1116-73-0*]	282.5	0.8374 (25)	−87	
Tri-*n*-octylaluminum, $(n\text{-}C_8H_{17})_3Al$ [*1070-00-4*]	366.6	0.8336 (25)	<−40	
Diethylaluminum hydride, $(C_2H_5)_2AlH$ [*871-27-2*]	86.1	0.7938 (25)	−59	49–51 (1.3×10^{-4})
Diisobutylaluminum hydride, $(i\text{-}C_4H_9)_2AlH$ [*1191-15-7*]	142.2	0.7989 (25)	−80	89–105 (1.3×10^{-5})
Sodium aluminum diethyl dihydride, $NaAl(C_2H_5)_2H_2$ [*17836-88-3*]	110.1	0.87 * (20)		
Dimethylaluminum chloride, $(CH_3)_2AlCl$ [*1184-58-3*]	92.5	0.9946 (25)	−21	125.7 (101.3)
Diethylaluminum fluoride, $(C_2H_5)_2AlF$ [*367-44-2*]	104.1	0.992 (24)		90–91 (14)
Diethylaluminum chloride, $(C_2H_5)_2AlCl$ [*96-10-6*]	120.6	0.9709 (25)	−74	125 (6.7)
Diethylaluminum bromide, $(C_2H_5)_2AlBr$ [*760-19-0*]	165.0			147–148 (6.7)
Diethylaluminum iodide, $(C_2H_5)_2AlI$ [*2040-00-8*]	212.0	1.5772 (25)	−43	105 (6.7)
Diisobutylaluminum chloride, $(i\text{-}C_4H_9)_2AlCl$ [*1779-25-5*]	176.7	0.9118 (25)	−40	108 (0.13)
Methylaluminum sesquichloride, $(CH_3)_3Al_2Cl_3$ [*12542-85-7*]	205.4	1.1626 (25)	23	143.7 (101.3)
Ethylaluminum sesquichloride, $(C_2H_5)_3Al_2Cl_3$ [*12075-68-2*]	247.5	1.0956 (25)	−21	209.6 (101.3)

in electronic or ceramic applications [15]. The latter are, however, much smaller volume applications.

Alkylaluminums form complexes with nonprotic Lewis base solvents such as ethers or tertiary amines. They react with protic or halogenated solvents such as water, alcohols, ketones, esters, or many chlorocarbons.

2.2.1. Reactions with Olefins

The *hydroalumination reaction* and the reverse process, olefin elimination, are important in economic syntheses of organoaluminum materials [16]:

Table 1. (Continued)

Compound	M_r	Density, g/mL (°C)	*mp*, °C	*bp*, °C (*p*, kPa)
Ethylaluminum sesquibromide, $(C_2H_5)_3Al_2Br_3$ [*12327-99-0*]	380.9			113 (0.53)
Methylaluminum dichloride, CH_3AlCl_2 [*917-65-7*]	112.9		73.6	159 (101.3)
Ethylaluminum dichloride, $C_2H_5AlCl_2$ [*563-43-9*]	127.0	1.227 (35)	31	115 (6.7)
Isobutylaluminum dichloride, *i*-$C_4H_9AlCl_2$ [*1888-87-5*]	155.0	1.1247 (25)	−30	112 (1.3)
Diethylaluminum ethoxide, $(C_2H_5)_2AlOC_2H_5$ [*1586-92-1*]	130.2	0.851 (20)	2	108 – 109 (1.3)
Diethylaluminum thioethoxide, $(C_2H_5)_2AlSC_2H_5$ [*17002-76-5*]	146.2			84 – 85 (6.7×10^{-3})
Diethylaluminum cyanide, $(C_2H_5)_2AlCN$ [*5804-85-3*]	111.1			162 (2.7×10^{-3})
Diethylaluminum dimethylamide, $(C_2H_5)_2AlN(CH_3)_2$ [*13154-82-0*]	129.2			67 – 69 (1.6×10^{-2})

∗ 25 % solution in toluene.

$$R'_2Al-H + CH_2=CHR \rightarrow R'_2AlCH_2CH_2R \quad (1)$$

The rate of addition of the Al – H bond to olefins is greatest for ethylene, and decreases with 1-alkenes, 2-alkyl-1-alkenes, and 1,2-dialkylethylenes. Stability to thermal elimination of the alkene takes the same order.

Triethyl- and higher tri-*n*-alkylaluminum compounds readily add to ethylene [*74-85-1*] or 1-alkenes in a *carboalumination reaction* [17]:

$$R''_2Al(CH_2CH_2R) + CH_2=CHR' \rightarrow R''_2Al(CH_2CHR'CH_2CH_2R) \quad (2)$$

where R, R′ = H or alkyl. When R′ is an alkyl group, the aluminum atom adds to the methylene, as opposed to the methine carbon, in a ratio of ca. 9 : 1. With ethylene, this reaction can repeat, resulting in a *chain growth process.*

$$R''_2Al\,(CH_2CH_3) + n\,CH_2=CH_2 \rightarrow R''_2Al[(CH_2CH_2)_nCH_2CH_3] \quad (3)$$

Starting with triethylaluminum and ethylene, the distribution of alkyl groups in the product follows the Poisson statistical equation. The shape and therefore, the carbon number peak of the curve formed by plotting carbon number vs. weight percentage is a function of the amount of ethylene added. This process results in a primarily linear trialkylaluminum mixture.

Alkylaluminum compounds with branches in the β position add to another olefin molecule only with difficulty. But, these branched compounds readily eliminates a branched 2-alkyl-1-alkene, generally called a *vinylidene olefin:*

$$R''_2Al(CH_2CHR'CH_2CH_2R) \rightarrow CH_2=CR'CH_2CH_2R + R''_2AlH \quad (4)$$

where R = H or alkyl; R′ = alkyl

2.2.2. Reactions with Oxygen

The carbon-aluminum bonds in organoaluminum compounds react readily with oxygen [*7782-44-7*] by an exothermic radical chain mechanism that, under controlled conditions, can give high yields of aluminum alkoxide but low yields of aryloxide with alkyl and aryl groups, respectively [1–4], [18]. With a trialkylaluminum material, the overall reaction is:

$$2\,R_3Al + 3\,O_2 \rightarrow 2\,Al(OR)_3 \quad (5)$$

The molar heat of reaction ($\Delta H = -1071$ kJ/mol at 25 °C [19]) is sufficiently great that, exposure of concentrated organoaluminum materials to air can produce an uncontrolled reaction with oxygen. The heat is so concentrated that the kindling temperature of the material is exceeded and spontaneous flaming occurs.

Establishing conditions for effective control of the oxidation reaction was an important contribution in development of the large volume industrial process for linear primary alcohols utilizing organoaluminum chemistry.

2.2.3. Reactions with Metal Compounds

Several types of reactions of organoaluminum compounds with metal salts can be distinguished.

Salts of more electronegative main-group metals or metalloids, particularly those in groups 12 through 15, frequently undergo *alkylation* by organoaluminum materials [20]. An example, with a group 12 metal salt, is the reaction of triethylaluminum with zinc chloride [*7646-85-7*] to form diethylzinc [*557-20-0*], which can be separated by distillation from coproduct diethylaluminum chloride.

$$2\,Al(C_2H_5)_3 + ZnCl_2 \rightarrow Zn(C_2H_5)_2 + Al(C_2H_5)_2Cl \quad (6)$$

Another industrially important example is the production of tin(IV) alkyls from stannic chloride [*7646-78-8*] [21]:

$$4\,R_3Al + 3\,SnCl_4 \rightarrow 3\,SnR_4 + 4\,AlCl_3 \quad (7)$$

$R = n\text{-}C_6H_{13}$ or $n\text{-}C_8H_{17}$

Saltlike complexes are formed with the more electropositive elements, especially group 1 and 2 metals [22]. For example, when triethylaluminum is heated with sodium fluoride [*7681-49-4*] at 100 – 120 °C, a stable complex results.

$$(C_2H_5)_3Al + NaF \rightarrow Na[Al(C_2H_5)_3F] \quad (8)$$

Complex formation is also possible with quaternary ammonium, phosphonium, or sulfonium type salts. These complexes are used as ionic liquids. Liquid clathrate formation is an interesting property of these saltlike complexes and has potential applications in coal liquefaction [23].

Reductions often occur when alkylaluminum compounds are allowed to react with transition-metal compounds. Depending on reactants, stoichiometry, and conditions, either completely reduced, finely divided, very active metals result, or reduced coordination complexes are formed [10], [14]. The latter, which include Ziegler – Natta and other coordination catalysts, show remarkable catalytic effects in oligomerizations, polymerizations, and other reactions of olefinic materials.

2.2.4. Reactions with Proton-Donating Materials

Most protonolysis reactions of alkylaluminum compounds are very exothermic, but under controlled conditions they can be used to prepare such materials as alkylaluminum alkoxides [24], [25], cyanides [26], [27] mercaptides [25], [28], halides, or dialkylamides [25]:

$$R_3Al + HZ \rightarrow R_2AlZ + RH \quad (9)$$

The active hydrogen reagent (HZ) may be mineral acid, water, an alcohol or a mercaptan, a phenol, carboxylic acid, ammonia or amine, or even mono- or unsubstituted acetylene.

When conditions are not controlled, thermal degradation also can take place, forming a mixture containing olefins and hydrogen as well as the alkane in the case of alkylaluminum materials. With water or other reagents that form aluminum-oxygen bonds, the reaction is highly exothermic. If water is added to an organoaluminum compound that is burning in air, therefore, a very violent reaction can ensue.

Protonolysis is also an important technique for preparing organoaluminum compounds for analysis.

2.2.5. Alkylaluminums in Organic Synthesis

The reactions of alkylaluminum reagents with organic functional groups has been the subject of much research [1–8]. Two further reactions of alklyaluminums have reasonable industrial significance.

First, numerous examples have been reported of *alkylation of the carbonyl carbon* in acyl halides, ketones, acid anhydrides, ketenes, esters, lactones, and aldehydes using alkylaluminum compounds. Alcohols are formed after hydrolysis [3], [4], [5].

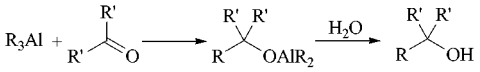

(10)

Diisobutylaluminum hydride or other complex organoaluminum hydrides are useful reagents for reduction of the aldehyde or ketone carbonyl function to an alcohol [1–5], [7], [8].

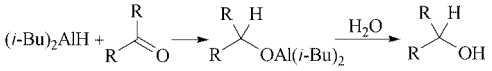

(11)

Second, *alkynes* can be stereoselectively hydroaluminated. After hydrolysis, the products formed are exclusively cis-alkenes [6].

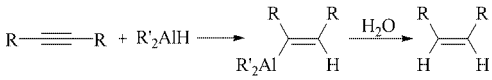

(12)

2.3. Production of Trialkylaluminums and Alkylaluminum Chlorides

Trialkylaluminums are available industrially via the reactions of aluminum [*7429-90-5*] with hydrogen plus alkenes, or by reduction of alkylaluminum halides. *Alkylaluminum halides* are produced by reaction of hydrogen halides with trialkylaluminums, or by direct reaction of aluminum with an alkyl halide. *Alkylaluminum hydrides* and other products are produced by several conversion methods.

Often the method of synthesis has a great influence on the quality of an organoaluminum material in terms of the presence of trace components. In addition, the hydroalumination method is inherently safer to practice industrially than the reaction of aluminum metal with an alkyl chloride followed by reduction.

2.3.1. Hydroalumination to Produce Trialkylaluminums

Under the activating influence of an alkylaluminum compound, again in the absence of air or moisture, finely divided aluminum metal reacts directly with hydrogen [*1333-74-0*] to form an Al – H bond. A 1-alkene or 2-alkyl-1-alkene can then add by hydroalumination to give the primary trialkylaluminum product [16].

With ethylene and *n*-1-alkenes, a two-step approach can be used.

$$4\,R_3Al + 2\,Al + 3\,H_2 \rightarrow 6\,R_2AlH \qquad (13)$$

$$R = C_2H_5,\ i\text{-}C_4H_9,\ n\text{-}C_4H_9,\ n\text{-}C_6H_{13},\ n\text{-}C_8H_{17},$$
$$110 - 140\,^\circ C,\ 5 - 20\,MPa$$

$$6\,R_2AlH + 6\,(1\text{-alkene}) \rightarrow 6\,R_3Al$$
$$80 - 100\,^\circ C \qquad (14)$$

In the first step, Equation 13, the Al – H bond is formed and stabilized as the dialkylaluminum hydride. A relatively high temperature is required to achieve an acceptable reaction rate. Air-blown aluminum powder is suitable for the process. Formation of Al – H bonds can be accelerated by milling the aluminum powder in the presence of the trialkylaluminum product [29], or by alloying certain transition metals, such as titanium or zirconium, with the aluminum powder [30]. Among other catalysts reported for the reaction are sodium alcoholates [31], sodium metal [32], and sodium hydride [32].

In the second step, Equation 14, the olefin adds rapidly at a decreased temperature where unwanted carboalumination of the olefin is negligible.

Triethylaluminum (TEA), tri-*n*-butylaluminum (TNBA), triisobutylaluminum (TIBA), tri-*n*-hexylaluminum (TNHA), and tri-*n*-octylaluminum (TNOA) can be produced by the two-step approach, which has been referred to as the three-for-two process. Both steps can also be carried out simultaneously: in this case, the feed ratio of hydrogen to olefin must be controlled carefully [33–37]. Both batch and continuous triethylaluminum processes have been described in patents [33–35].

Straight-chain alkylaluminums made by the hydroalumination process generally contains some residual Al – H groups. They also contain higher-chain groups that form through the competing ethylene chain – growth reaction. When higher trialkylaluminum products are made from 1-alkenes, the process-related impurities include, along with Al – H bonds, some β-branched alkyl groups, which can be referred to as 1-alkene dimer groups. Triisobutylaluminum contains no long-chain impurities because isobutene [*115-11-7*] does not undergo chain growth.

The synthesis of lower hydride straight-chain alkyls is accomplished by increasing the 1-alkene concentration during production. The materials, however, contain higher levels of longer-chain alkyl byproducts.

Crude triethylaluminum can be purified by fractional distillation. Distillation of the higher trialkylaluminum products is difficult because of their relatively low vapor pressures. Furthermore, temperatures required for distillation of TNBA, TIBA, TNHA, and TNOA are usually high enough to result in thermal elimination of 1-alkene and the concomitant introduction of Al – H groups. These more thermally sensitive alkyls are purified by either by filtration or by wiped film evaporation.

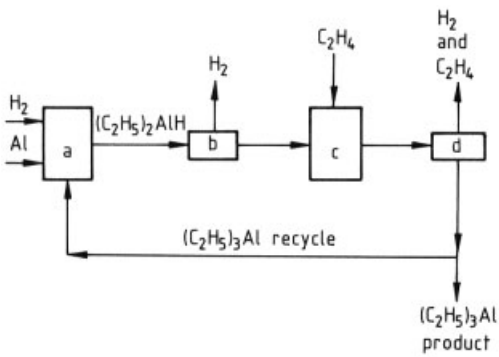

Figure 1. Continuous two-step process for triethylaluminum
a) Hydriding reactor; b) Flash drum; c) Ethylation reactor; d) Flash drum.

Process Example. A flow scheme for a continuous two-step industrial production of triethylaluminum is represented in Figure 1 [36]. The hydriding reactor (a) is of the vertical cylinder type, 3.8 m^3 in volume. In a typical example, a slurry is prepared containing 7 wt % aluminum powder (0.1 wt % zirconium catalyst content) and 45 wt % recycled triethylaluminum in a paraffin hydrocarbon fraction, boiling range 177 – 260 °C. The slurry is pumped into the reactor at a rate of 1.5 m^3/h, at a reactor temperature of 132 °C, while hydrogen is introduced at the bottom to maintain a partial pressure of 10 MPa. Diethylaluminum hydride is formed according to Equation 4, and is present at a concentration of 83 wt % in the stream flowing from the reactor.

Excess hydrogen is removed in flash drum (b), wherein pressure is decreased to 3 MPa and temperature is lowered to prevent thermal reversal of the reaction.

In the ethylation reactor (c) sufficient ethylene is added to react completely with the Al – H bonds (Eq. 5). Temperature is held at around 85 °C or somewhat higher. The addition rate increases with temperature, but carboalumination of ethylene also can occur to form *n*-butyl groups, especially at higher temperatures. Hydrogen and ethylene partial pressure is kept low in (c) to minimize hydrogenation of ethylene to ethane and to reduce excess carboalumination products.

Excess ethylene and hydrogen are removed in flash drum (d). A portion of the triethylaluminum product solution is recycled to reactor (a). The remainder may be used as such, or, if the solvent is a hydrocarbon of higher boiling range, pure triethylaluminum may be separated from it by distillation.

2.3.2. Reaction of Aluminum With Alkyl Halides to Make Alkylaluminum Sesquichlorides

Methyl, ethyl, and other alkyl or aralkyl halides that are not dehydrohalogenated readily, can react with aluminum metal in an exothermic process to form organoaluminum sesquihalides in high yields [38], [39]. An important example is the reaction of methyl chloride [*74-87-3*] with aluminum [*7429-90-5*].

$$3\,CH_3Cl + 2\,Al \rightarrow (CH_3)_3Al_2Cl_3 \qquad (15)$$

The reaction is carried out with aluminum in the form of turnings, shavings, granules, or powder. Oxygen and moisture must be rigorously

excluded. The reaction can be initiated with a small amount of mercury or iodine. It also can be started by treating the aluminum with an alkylaluminum halide. A portion of the residue from a previous reaction is effective.

The products are equilibrium mixtures of the codimer ($R_2AlX \cdot RAlX_2$) and homodimers [$(R_2AlX)_2$ and $(RAlX_2)_2$], in which the two aluminum atoms of each component are halogen-bridged as shown by compound (**3**), Section 2.1.

In the reaction of methyl chloride [*74-87-3*] with aluminum, small amounts of ethane are formed by an apparent Wurtz-like reaction. With ethyl chloride, some butane is produced; also, dehydrochlorination of ethyl chloride can occur to form byproduct ethylene and excess Al – Cl bonds in the product. The ethylene may form oligomers in the mixture by a cationic mechanism. The byproduct reactions are favored by higher temperatures and therefore can be minimized by careful temperature control. When byproduct reactions take place to a significant extent, the excess Al – Cl content in the $R_3Al_2Cl_3$ product can be decreased by addition of the calculated amount of R_3Al. Overall, however, it is critical to control reaction conditions as slight excursions can result in catastrophic events.

Products can be recovered by settling and decantation, filtration, or distillation. Large quantities of methyl- and ethylaluminum sesquichlorides are produced industrially by this method. They are used as alkylating agents, catalyst components, and for making other methyl- or ethylaluminum chlorides.

Process Example. Production of ethylaluminum sesquichloride by reaction of ethyl chloride with aluminum is illustrated by this patent example [39]. A $9.5 \times 10^{-2}\,m^3$ closed steel reactor, fitted with a reflux condenser and equipped for vacuum distillation of product, was first purged with nitrogen. After addition of 13.6 kg of aluminum powder and a catalyst mixture composed of 3.6 kg diethylaluminum chloride and 410 g iodine [*7553-56-2*], the reactor contents were stirred and heated slowly to 130 °C. The temperature was maintained at 120 – 150 °C during addition of 50 kg ethyl chloride over a period of 3 h.

The product can be flash distilled under vacuum, or it can be drained from the reactor and clarified by settling or by filtration of black residual metal solids. Iodine can be omitted from the procedure, in which case the induction period may be increased.

2.3.3. Reduction of Alkylaluminum Sesquichlorides to Make Trialkylaluminums

The alkylaluminum sesquihalide products from aluminum reactions with alkyl halides can be converted to dialkylaluminum halide or trialkylaluminum materials by treatment with *active metals*, such as sodium [*7440-23-5*] or magnesium [*7439-95-4*] [38], [39]. For example, diethylaluminum chloride or triethylaluminum can be produced from ethylaluminum sesquichloride by sodium reduction.

$$2\,(C_2H_5)_3Al_2Cl_3 + 3\,Na \rightarrow 3\,(C_2H_5)_2AlCl + Al + 3\,NaCl \quad (16)$$

$$3\,(C_2H_5)_2AlCl + 3\,Na \rightarrow 2\,(C_2H_5)_3Al + Al + 3\,NaCl \quad (17)$$

A magnesium-aluminum alloy can provide the reduction function concurrently with reaction between aluminum and the alkyl halide.

$$4\,C_2H_5Cl + Al_2Mg \rightarrow 2\,(C_2H_5)_2AlCl + MgCl_2 \quad (18)$$

The sesquichloride reduction process is currently the most economical route available for production of trimethylaluminum.

Trialkylaluminum products made by this process generally contain trace levels of residual chloride but do not contain Al – H or other low-level components found in the aluminum-hydrogen-olefin processes (Section 2.3.1).

The components of ethylaluminum sesquichloride cannot be separated economically by fractional distillation because of unfavorable vapor pressure relationships. However, diethylaluminum chloride can be distilled from the crystalline sodium chloride complex that forms with ethylaluminum dichloride [22].

$$(C_2H_5)_3Al_2Cl_3 + NaCl \rightarrow (C_2H_5)_2AlCl + Na[Al(C_2H_5)Cl_3] \quad (19)$$

Process Example. Diethylaluminum chloride has been made industrially according to Equation 16. In a patent example [39], ethylaluminum sesquichloride (26.5 kg) added to a nitrogen-purged $9.5 \times 10^{-2}\,m^3$ reactor, was heated to 175 °C. Then, while the mixture was stirred vigorously, 1.1 kg sodium was added over a 30 min period; the mixture was further heated at 155 – 190 °C for 60 min. The diethylaluminum chloride product distilled from the reactor at 100 – 161 °C (1.3 – 6.1 kPa).

In this example, an excess of ethylaluminum sesquichloride was employed to facilitate draining the voluminous byproduct salt and aluminum solids from the reactor. In an alternate approach, a heavy hydrocarbon oil, added prior to reaction, may be employed to remove the solids in slurry form, permitting the use of a stoichiometric ratio of ethylaluminum sesquichloride and sodium reactants.

2.3.4. Reaction of Acids With Trialkylaluminums to Produce Alkylaluminum Chlorides

Diethylaluminum chloride (DEAC), ethylaluminum sesquichloride (EASC), and ethylaluminum dichloride (EADC) can all be prepared by the reaction of triethylaluminum with HCl:

$$(C_2H_5)_3Al + n\,HCl \rightarrow (C_2H_5)_{(3-n)}AlCl_n + n\,C_2H_6 \qquad n = 1, 1.5, 2 \quad (20)$$

The products or their derivatives are used as components of catalysts for the production of polyolefins and some elastomers.

2.3.5. Reproportionation Reactions

When alkylaluminum halide or trialkylaluminum compounds and aluminum trihalides are combined, reproportionation reactions occur rapidly under ambient blending conditions, releasing small amounts of energy [38], [39]. These reactions are of industrial importance in making, e.g., diethylaluminum chloride or ethylaluminum dichloride. Either can be formed, depending upon reactants and stoichiometry chosen.

$$2\,(C_2H_5)_3Al + AlCl_3 \rightarrow 3\,(C_2H_5)_2AlCl \quad (21)$$

$$(C_2H_5)_3Al + 2\,AlCl_3 \rightarrow 3\,C_2H_5AlCl_2 \quad (22)$$

$$(C_2H_5)_3Al_2Cl_3 + (C_2H_5)_3Al \rightarrow 3\,(C_2H_5)_2AlCl \quad (23)$$

$$(C_2H_5)_2AlCl + AlCl_3 \rightarrow 2\,(C_2H_5)AlCl_2 \quad (24)$$

This method is particularly useful for making alkylaluminum halides that cannot be produced via the reactions of aluminum with alkyl halide, e.g., diisobutylaluminum chloride and isobutylaluminum dichloride. These are preferred over the corresponding ethyl homologues as co-catalysts in certain olefin polymerization systems.

2.3.6. Olefin Elimination and Displacement Reactions

2-Alkyl-1-alkene olefins are readily eliminated from β-branched trialkylaluminum compounds. In the most common of these compounds, TIBA, a substantial level of Al – H bonds are present at equilibrium. Diisobutylaluminum hydride (DIBAH) is commercially produced by slow thermolysis of TIBA with concurrent removal of the formed isobutene:

$$(i\text{-}C_4H_9)_3Al \rightarrow (i\text{-}C_4H_9)_2AlH + i\text{-}C_4H_8 \quad (25)$$

The greater stability of unbranched trialkylaluminum compounds relative to branched trialkylaluminum compounds in TIBA or DIBAH forms the basis for a general synthesis of triethyl- and higher linear trialkylaluminum materials from triisobutylaluminum or diisobutylaluminum hydride [40].

$$(i\text{-}C_4H_9)_3Al + 3\,RCH{=}CH_2 \rightarrow (RCH_2CH_2)_3Al + 3\,i\text{-}C_4H_8 \quad (26)$$

When the displacing olefin is a 1-alkene, the competing carboalumination reaction of 1-alkene with product trialkylaluminum can be minimized by careful temperature control and, particularly in a batch system, by maintaining a relatively high Al – H concentration. The latter condition is achieved by continuous removal of isobutene from the system while the displacing 1-alkene is fed at a limited rate.

The triisobutylaluminum displacement process can be used for production of triethylaluminum, owing to greater stability of ethylaluminum over isobutylaluminum groups. The resulting triethylaluminum product contains some residual isobutyl groups, in addition to the Al – H and chain – growth alkyl groups inherent in the ethylene hydroalumination process. When using 1-alkenes in the displacement process, small amounts of dimer alkyl groups are formed by carboalumination.

Triethyl-, tri-*n*-propyl-, or tri-*n*-butylaluminum products made by the displacement process can be purified by distillation. Displacement products of the higher alkylaluminums of good quality can be obtained by beginning with triisobutylaluminum or diisobutylaluminum hydride of good purity, carrying out the reaction with a slight excess of 1-alkene, and then carefully removing excess 1-alkene at a minimum temperature under vacuum.

Several trialkylaluminum products, e.g., TEA, TNHA, and TNOA, are made industrially by this process. A product mixture, referred to as "isoprenylaluminum", is produced by partial displacement of isobutene from triisobutylaluminum by isoprene [41], [42].

2.3.7. Other Organoaluminum Conversions

Preparation of a typical *organoaluminum fluoride,* diethylaluminum fluoride, has been reported via the reaction of diethylaluminum chloride with sodium fluoride [43].

$$(C_2H_5)_2AlCl + NaF \rightarrow (C_2H_5)_2AlF + NaCl \quad (27)$$

Diethylaluminum iodide can be prepared by concurrent reaction of iodine [*7553-56-2*] and aluminum with triethylaluminum [44], [45], [46].

$$4\,(C_2H_5)_3Al + 3\,I_2 + 2\,Al \rightarrow 6\,(C_2H_5)_2AlI \quad (28)$$

2.4. Uses of Alkylaluminum Compounds

Industrial uses for organoaluminum materials include a number of stoichiometric and catalytic applications that consume hundreds of thousands of tonnes each year.

2.4.1. Stoichiometric Applications

Undoubtedly the stoichiometric use that consumes the most alkylaluminum material is the production of linear primary alcohols (→ Alcohols, Aliphatic, → Ziegler Processes). Triethylaluminum is the single most important organoaluminum material on a volume basis.

A long-chain trialkylaluminum intermediate is first made by chain growth of ethylene on TEA or by the displacement of the alkyl group by a higher-chain α-olefin. The long-chain alkylaluminum compound is oxidized with oxygen, and the resulting aluminum trialkoxide is hydrolyzed to produce the primary alcohol.

The products are even carbon number, primary alcohols, that are large-volume intermediates to poly(vinyl chloride) plasticizer esters ($C_6 - C_{10}$) and numerous surfactant derivatives ($C_{10} - C_{18}$).

The second largest volume use of alkylaluminums is to make linear α-olefins. These are produced first by chain growth of ethylene on triethylaluminum to produce a long-chain alkylaluminum compound. This long-chain alkylaluminum compound is reacted with another olefin such as ethylene which displaces the long-chain alkyl as the α-olefin and regenerates the triethylaluminum. Product olefins are separated from the TEA, which is recycled.

Linear α-olefins are highly versatile intermediates to a number of important products [46]. The main uses are in the production of copolymers, surfactants, and lubricants. Linear low density polyethylene is a copolymer of ethylene with either C_4, C_6, or C_8; C_{18} is used as a comonomer with either maleic anhydride [*108-31-6*] or 1-vinyl-2-pyrrolidinone [*88-12-0*]. Surfactants are made from C_8, C_{10}, and C_{12} alcohols as such or from the derivatized alcohol sulfonates, ethoxylates, sulfonated ethoxylates, and alkylated phenols. Poly-α-olefins, the major basestock in synthetic hydrocarbon lubricants, are based on C_{10} oligomers; derivatives of C_{20} and higher fractions are used as oil pour-point depressants, engine oil detergents, and other engine oil additives or as paper sizing agents. In addition, C_{20} and higher fractions are used as waxes and in investment casting.

Another stoichiometric application of alkylaluminum compounds is in the alkylation of metals or metalloid salts to produce organometallic products. These products include derivatives of zinc, boron, silicon, lead, mercury, tin, and phosphorus. The largest volume of materials produced this way are tetraalkyltin(IV) compounds. These compounds are derivatized to alkyltin carboxylates or to alkyltin mercaptides which are used in such applications as poly(vinyl chloride) heat stabilizers, microbiocides, and catalysts [47].

Uses of organoaluminum materials in organic syntheses, particularly in the production of specialty chemical products, e.g., pharmaceuticals, fragrances, and certain pesticides, have received much attention [3–6]. Alkylaluminums, particularly high purity trimethylaluminum, are also used as precursors to high purity alumina and aluminum containing electronic materials and ceramics.

2.4.2. Catalytic Applications

The longest known industrial use of organoaluminum materials is in Ziegler – Natta catalysts for olefin and diene polymerizations [14]. In this use alkylaluminums reduce a transition metal compound, typically of titanium or vanadium, to a catalytically active oxidation state. They then alkylate the catalyst to form the metal – alkyl bond that is the site of monomer insertion and polymerization. They also serve as poison scavengers within the polymerization reactor. Additionally, they can become a part of the active coordination catalyst complex.

Industrially important polymers that utilize Ziegler – Natta or other coordination catalysts in their production include high-density polyethylene, isotactic polypropene, both plastic and elastomeric ethylene-propylene copolymers, ethylen – propene – diene (EPDM) terpolymers, isotactic poly-1-butene, cis-polydiene elastomers, high-density polyethylene, and the linear low- and medium-density polyethylenes that contain low levels of butene, hexene or octene comonomers.

Important products from reactions using other coordination catalysts are, for example, cyclic dimers or trimers from butadiene and linear α-olefin oligomers from ethylene; C_6 and C_8 materials from propene and linear butene dimerization, used in gasoline blending and C_9 oxoalcohol production, respectively; and synthetic lubricants, hydraulic fluids, and waxes from oligomerizing ethylene and 1-alkenes. The catalysts also are used for the ring-opening polymerization of cyclic olefins, e.g., in conversion of cyclopentene to polypentenamer elastomer, or the olefin redistribution reaction, as in the conversion of propene to its equilibrium mixture with ethylene and 2-butene.

Effective homogeneous hydrogenation catalysts can be made by reduction of transition-metal salts with alkylaluminum compounds.

2.5. Quality Specifications

For some applications of alkylaluminum materials, e.g., for syntheses of olefins from trialkylaluminum compounds, the major quality criterion is simply total R_3Al content. The processes are not affected significantly by trace impurities. For other applications, however, such as in the use of R_3Al or R_2AlX in preparing certain coordination catalysts for olefin polymerizations, even low levels of hydride or alkoxide can affect catalyst activity. In addition, keeping the halogen : aluminum atomic ratio within a narrow set of limits can be important.

When diethylaluminum chloride is used as an activator for some Ti(III)-based polypropylene catalysts, the presence of hydrogen as hydride at levels in excess of 0.05 mol % of the ethyl groups may decrease stereoselectivity. The presence of oxygen as ethoxide at O : Al atomic ratios below 0.1 may increase activity for some systems, whereas above 0.1 the activity usually is decreased.

The Cl : Al atom ratio must be controlled between 0.97 and 1.03, either near the lower limit or the higher limit, depending on other catalyst components and on various process factors.

During the analysis of organoaluminum materials, extreme care must be taken, because exposure to traces of air can lead to oxidation or hydrolysis. Also, thermal degradation may result during the hydrolysis that precedes gas chromatography [24].

Halogen and *total aluminum* content in organoaluminum materials are determined by advanced methods on hydrolysis solutions that

are based on state-of-the-art titration equipment and computerized data enhancement. Standard deviations of the advanced methods are significantly lower than conventional methods. *Active aluminum,* i.e., the aluminum in trialkylaluminum and dialkylaluminum hydride form, can be determined by careful protonolysis with ammonia or amines [48], [49].

Dialkylaluminum hydrides have been, in the past, assayed by spectrophotometric measurement of complexes with an azomethine compound, such as isoquinoline [49]. The isoquinoline analysis has been replaced with more accurate and precise analytical methods based on gas chromatography and NMR. Hydrolysis of organoaluminum compounds followed by capillary gas chromatography of the gaseous phase using a megabore molecular sieve column is effective for determining hydride and alkyl content of organoaluminum compounds. The liquid organic hydrolysis products are determined by standard capillary gas chromatography techniques. Another approach to measurement of hydride and alkyl group content in organoaluminum materials is ^{1}H-NMR spectroscopy. The ^{1}H-NMR analysis avoids the lengthy and labor intensive gas evolution analyses and directly measures the hydride and alkyl groups. No calibration is needed. Further, the limit of detection is very low and the precision is high. In a single NMR analysis the hydride, identity of alkoxide impurities, and alkyl group content can be determined. Best results are obtained using a 300 MHz or higher field NMR instrument [50].

Oxygen at trace levels in the form Al – O – Al is generally not measured directly in organoaluminum materials. Alkoxide oxygen is determined readily, however, even at low concentrations directly by NMR.

3. Aluminoxanes

3.1. Production

Aluminoxanes are produced by the partial hydrolysis of alkylaluminum compounds with water [5], [51]:

$$n\,R_3Al + n\,H_2O \longrightarrow \left(\begin{array}{c} R \\ | \\ Al-O \end{array}\right)_n + 2n\ RH \qquad (29)$$

Methylaluminoxane (MAO) [*120144-90-3*], is by far the industrially most significant aluminoxane. The direct hydrolysis of trimethylaluminum (TMA) with water, however, is extremely exothermic and difficult to control. Thus, small scale, continuous processes, rather than large scale batch processes, are used for the direct hydrolysis of TMA to MAO [52].

MAO can also be produced by reaction of TMA with organic acids, ketones, or carbon dioxide [51], [53]. The synthesis of MAO from TMA and CO_2 actually has substantial literature precedent [5], [51], [54]. Carbonation of an organoaluminum compound with carbon dioxide forms an aluminum carboxylate:

$$R_3Al + CO_2 \rightarrow R_2Al\text{-}OCOR \qquad (30)$$

With an excess of carbon dioxide, a trialkylaluminum compound gives $R_2AlOCOR$ in good yield, but conversion of the second alkyl group to carboxylate proceeds with difficulty. Conditions have not been found for reaction of the third group. Reaction of excess trialkylaluminum with carbon dioxide produces what was originally reported as a low molecular mass alkylaluminoxane dimer and aluminum carboxylate.

$$R_2AlOCOR + 2\,R_3Al \rightarrow (R_2Al)_2O + R_2AlOCR_3 \qquad (31)$$

Isobutylaluminoxane. In contrast, the direct hydrolysis of triisobutylaluminum (TIBA) to produce isobutylaluminoxane (IBAO) proceeds less violently and can be performed in large batch scale operations [55], [56].

Mixed alkylaluminoxanes are aluminoxanes that typically contain 60 to 85 mol % methyl groups while the remaining alkyl groups are of a higher carbon number. A methyl – isobutyl mixed alkyl aluminoxane can be produced by diluting IBAO with TMA and then hydrolyzing the excess alkyl groups [55], [56]. Mixed alkyl aluminoxanes containing other alkyl groups can also be produced by hydrolyzing mixtures of TMA with another alkylaluminum [57].

3.2. Physical and Chemical Properties

Solvent free MAO is a glassy solid at room temperature. Other aluminoxanes such as IBAO and mixed alkylaluminoxanes are viscous oils at room temperature. Most aluminoxanes, however, are used as solutions in hydrocarbons. IBAO and mixed alkylaluminoxanes are available in toluene or saturated hydrocarbon solutions. Unmodified MAO is sold only in toluene because it is only sparingly soluble in saturated hydrocarbons.

Solid state molecular structures have been determined for many different *tert*-butylaluminoxane oligomers; two representatives are shown below [58]:

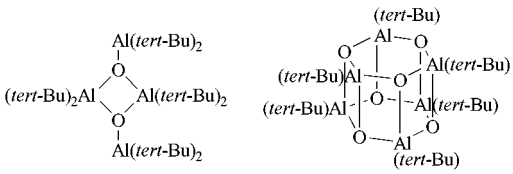

The structure of most aluminoxanes, however, is not well defined. Aluminoxanes are mixtures of oligomers of $n = 2$ to 20. Many structures have been proposed for aluminoxanes including rings, cages, and linear molecules [51]. It is likely that multiple equilibria exist in solution which interconvert the oligomer sizes and structures, particularly for MAO.

All commercially available MAO products contain residual TMA associated with the MAO. The level of TMA varies among manufacturers [59].

Although MAO can be produced as a clear, colorless solution in toluene, undisturbed solutions will precipitate hard gels over time as shown in Figure 2. Eventually, a 30 wt % solution of MAO will form a single gelatinous mass. Although the mass can be shaken or stirred to permit the material bulk to flow, the MAO solution will remain opaque with the presence of small, hard gels that cannot be redissolved into hydrocarbon solution. Figure 3 shows that refrigeration of MAO can significantly retard gel formation. IBAO and mixed alkyl aluminoxanes are much more stable and less likely to form gels than MAO.

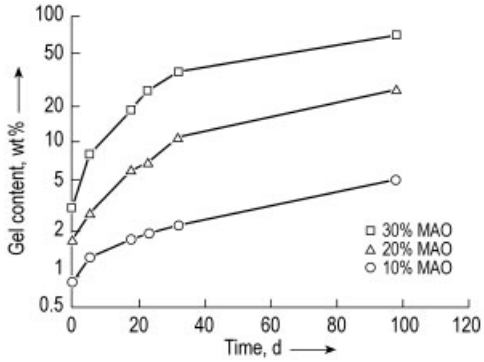

Figure 2. Formation of gels in MAO – toluene solutions over time at 25 °C

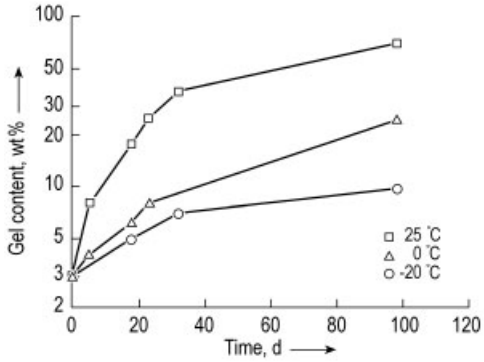

Figure 3. Formation of gels in 30 wt % MAO in toluene solutions at different temperatures

Like other organoaluminum compounds, aluminoxanes are extremely reactive toward oxygen and water and toward oxygenated and halogenated compounds.

3.3. Uses

Aluminoxanes are used industrially exclusively as cocatalysts in the polymerization of olefins (→ Metallocenes; → Polyolefins, Chap. 1.3.2.3.). Cyclopentadiene derivatives of zirconium, titanium, and hafnium are the primary catalysts used industrially [60]. But organometallic compounds based on nickel are also industrially important [61], and catalysts based on cobalt and iron have been identified [62].

Activities of 1 to 10 t of polyethylene per mole of transition metal are observed in these catalyst systems. Different molar alu-

minum : transition metal atomic ratios of one thousand to several hundred to one hundred or even less are used, depending on the catalyst system. A multitude of zirconocene and hafnocene derivatives, when used with MAO, are all active for the polymerization of olefins. No obvious correlation between catalyst activity and catalyst structure, however, has been shown [60].

3.3.1. Mechanism of Polymerization Using Aluminoxanes (→ Metallocenes; → Polyolefins, Chap. 1.3.2.3.)

The role of MAO in metallocene catalyst polymerization is fourfold. First, it alkylates the transition metal. Second, it extracts an alkyl group from the transition metal to generate an active cationic polymerization catalyst. Third, it acts as a bulky, inert anion to the highly reactive transition metal cation. Fourth, it scavenges impurities from the polymerization reactor [5].

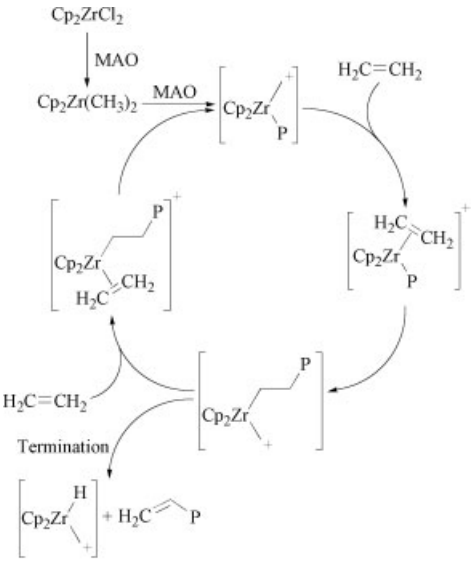

Figure 4. Mechanism of olefin polymerization using zirconocene with MAO catalysts

The mechanism of olefin polymerization using zirconocene/MAO catalysts has been well studied [63–67]. A representative mechanism is shown in Figure 4. When using bis(cyclopentadienyl) derivatives of zirconocene dichloride as the catalyst precursor, MAO first alkylates the zirconocene center to form a zirconocene dimethyl complex. The MAO further reacts with the zirconocene dimethyl compound, formally extracting a methyl anion, to produce a 14-valence-electron zirconocene monomethyl cation that has a vacant coordination site and a metal – alkyl bond.

The incoming olefin first coordinates to the vacant site on the metal, after which the alkyl group bound to the zirconocene center migrates to the β position on the olefin. This regenerates a zirconocene monoalkyl cation, which continues to insert olefins into the growing polymer chain. By this mechanism, the active polymerization site moves back and forth from one side of the molecule to the other.

The polymer chain is terminated by β-hydride elimination leaving the active catalyst site, a transition metal monohydride cation, which can begin the polymerization process anew.

After generating the active zirconocene methyl cation, the altered MAO acts as a surprisingly inert anion to the polymerization. Most common "inert" anions such as BF_4^-, BPh_4^-, and PF_6^- poison the zirconocene monoalkyl cation catalyst. Unique fluorinated borates such as $B(C_6F_5)_4^-$ [68], [69] and $MeB(C_6F_5)_3^-$ [64], [70] also do not react with the active catalyst and can be used industrially in these systems. Since the metallocene/borane systems are well defined molecular compounds, they have permitted the study and elucidation of the polymerization mechanism of this system to a degree never before seen with other catalysts.

On an industrial scale, MAO also acts as a scavenger of low level catalyst poisons such as oxygen, water, amines, sulfides, etc. The borane systems do not have the same poison scavenging ability that aluminoxanes do, and thus require much more strenuously clean reactor conditions.

3.3.2. Production of Stereoregular Polymers using Aluminoxanes

The so-called "windshield wiper" mechanism described in Figure 4 facilitates the production of stereoregular polymers, as shown in Figure 5. For example, consider the family of linked bis-indenyl zirconocene *ansa*-metallocene complexes shown below:

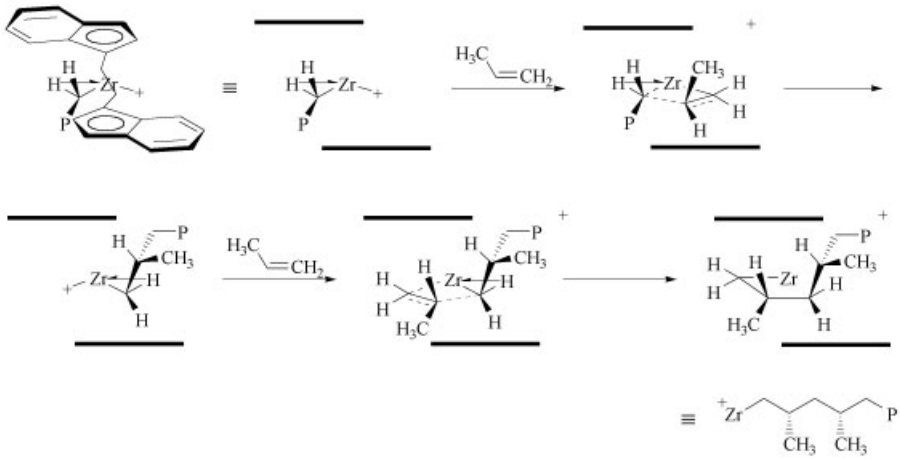

Figure 5. Synthesis of stereoregular polymers using *ansa*-metallocenes with MAO. (The bridged bis-indenyl ligand is depicted by two heavy lines for clarity.) The resting state of the catalyst is a zirconocene alkyl cation. Steric interactions direct the growing polymer chain to the open area away from the bulky indenyl ligand. This stereochemistry is further locked into place by an agostic interaction between an α-hydrogen on the polymer chain and the zirconium atom. Propene inserts, forming a metallocycle intermediate, with the propene methyl group trans to the polymer chain to minimize steric interaction. Migratory addition of the polymer chain switches the positions of the open coordination site and the growing polymer chain. Again, the intermediate is stabilized by an α-agostic interaction, and the polymer chain is directed away from the bulky indene ligand. Another propene coordinates to the vacant coordination site, the process repeats, and isotactic polypropylene is formed.

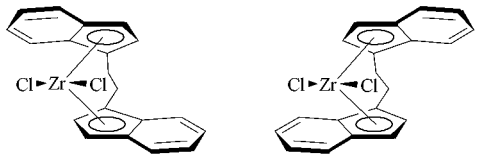

rac-isomers

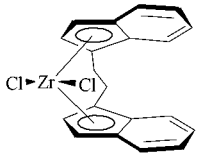

meso-isomer

The salient structural feature of these metallocene compounds, first studied by Brintzinger et. al., is that the two cyclopentadiene rings are linked together by a bridging group, which locks the relative stereochemistry of the two rings in place [71]. The *ansa*-zirconocene system exists as three different isomers: a racemic pair of enantiomers, called *rac*-, and a *meso*- diastereomer. When used with MAO cocatalyst, the *rac*- isomers polymerize α-olefins; the incoming monomer inserts into the growing polymer chain such that the pendant alkyl groups all wind up on the same side of the polymer molecule, producing isotactic stereochemistry. The *meso*- isomer produces atactic polypropylene under the same conditions.

In contrast, the cyclopentadiene – fluorenyl linked derivative shown below forces the pendant alkyl groups to alternate from one side of the polymer molecule to the other, creating syndiotactic polymer [63].

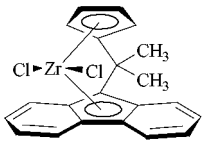

In all these cases, the minimization of steric interaction between the polymer, ligands, and incoming monomer drives the ultimate configuration of the polymer. An α-agostic interaction of the polymer with the transition metal center also helps control polymer stereochemistry [72].

3.3.3. Polymers Produced by Aluminoxanes

High molecular mass polyolefins are commonly produced industrially using either chromium on silica catalysts (→ Polyolefins, Chap. 1.3.2.1.) or Ziegler – Natta catalysts based on titanium chloride or vanadium catalysts (→ Polyolefins, Chap. 1.3.2.2.). These traditional catalysts have several chemically different active sites, each of which produce polymer of slightly different composition. As a result, these catalysts produce a bulk polymer that has a moderately broad to very broad molecular mass distribution (MMD). When producing ethylene – α-olefin copolymers, the different sites also have different affinities for the comonomer (α-olefin). Thus, copolymers made with Ziegler – Natta or chromium on silica catalysts also have a broad composition distribution. Typically, the comonomer is preferentially incorporated into the lower molecular mass molecules of the bulk polymer when using these traditional catalysts.

In contrast, transition metal – MAO catalysts are molecular catalysts and produce polymers of much narrower MMD and composition distribution. As a result, this family of catalysts is often referred to as "single site" catalysts, regardless if the transition metal compound is a zirconocene derivative, another metallocene, a "half-sandwich" complex such as the Dow "constrained geometry" catalyst [61], or a nickel catalyst [95]. (The label "single site", however, is a misnomer. Only solution catalysts that utilize high MAO aluminum to transition metal ratios are truly single site and produce polymer with a single Flory distribution. Most metallocene – MAO catalysts, especially heterogeneous catalysts supported on a carrier such as silica, produce polymer that has multiple Flory distributions [74].)

The first high volume industrial use of metallocene – MAO catalyst has been in the production of polyethylene, particularly linear low density polyethylene (LLDPE). The narrower MWD and composition distribution ultimately translate to a polymer that has higher impact strength, lower extractable components, better clarity, and better heat sealing properties than polymer made from traditional catalysts [75].

Likewise, ethylene – α-olefin copolymers can be made with much higher comonomer levels from metallocene catalysts than from Ziegler – Natta catalysts. These polymers have application in the plastomers and elastomers markets [76].

Isotactic polypropylene can be made industrially using MAO with derivatives of the *ansa*-zirconocene compound, *rac*-$EtInd_2ZrCl_2$ described above [77]. Derivatives of the MAO – $[Me_2Si(Cp)(Flu)ZrCl_2]$ (Flu = fluorenyl, η^5-$C_5H(C_4H_4)_2$) catalyst system produce syndiotactic polypropylene [78]. Isotactic polypropylenes produced by metallocene – MAO catalysts tend to have narrower MMD and higher isotactic index than those polymers produced by traditional catalysts.

Synthetic rubbers, which are terpolymers of ethylene, propylene, and either 1,4-butadiene or 2-ethylidenenorbornene (EPDMs) (→ Metallocenes, Chap. 2.; → Rubber, 3. Synthetic, Chap. 3.2.2.) can also be produced by zirconocene – MAO catalysts [79].

$CpTiCl_3$ or its derivatives and MAO catalyst systems are used commercially to produce syndiotactic polystyrene [80].

It has been reported that zirconocene derivatives or nickel catalysts can be used with an aluminoxane cocatalyst to produce linear α-olefins [81], [82]. Neither of these systems, however, has yet been commercialized.

Metallocene – aluminoxane catalysts are used industrially in high-pressure autoclave, solution, slurry, and gas-phase polymerization processes. These catalyst systems are used as both in homogeneous solutions and supported on a carrier such as silica.

3.4. Analysis and Quality Specifications

Alkylaluminum compounds have traditionally been analyzed to determine aluminum content, the molar ratio of alkyl groups to aluminum, and the identity of the alkyl groups and/or halides bonded to aluminum. The ratio and identity of alkyl groups is determined by hydrolysis of the alkylaluminum compound and gas chromatography analysis of the hydrolysis gases. MAO has also been analyzed and characterized using these techniques: aluminum content by titration and methyl group to aluminum ratio by digestion methods.

NMR has been used to detect and determine the presence of other species such as aluminum hydrides and other low level impurities in alkylaluminums. NMR of MAO reveals that there is always some residual TMA present in MAO. Indeed, NMR can be used to determine the absolute amounts of TMA and MAO in MAO product, as well as the methyl to aluminum ratio [59]. For this, NMR is superior to titration techniques which measure "acidic aluminum" and tend to overestimate the amount of TMA and consequently underestimate the amount of MAO.

MAO is a performance product: its quality is best measured in how well it performs in a given catalyst system. Specifications based on percent aluminum in solution, percent TMA, percent MAO and methyl to aluminum ratio are used and are useful to measure how consistent the manufacture of MAO is. MAO is still not well understood enough, however, to directly measure its critical features that relate to product quality.

4. Handling, Storage, and Transportation

In view of the reactivity of the industrially useful organoaluminum materials with air and water, much effort has been expended in developing special shipping, handling, and storage procedures (e.g., see merchant brochures in reference [83], [84]).

Organoaluminum materials generally are protected with an inert nitrogen atmosphere. Low moisture ($< 3.5\,mg/m^3$) and oxygen (< 0.001 vol %) contents are recommended. Activity toward air of the organoaluminum materials is decreased by dilution with inert solvents (e.g., aliphatic or aromatic hydrocarbons).

Chlorinated solvents should be avoided with organoaluminum compounds. Alkylaluminums react with polyhalogenated organic solvents, such as chloroform or carbon tetrachloride, to form unstable intermediates, especially at high temperature [85].

For shipping purposes, organoaluminum materials are classed as spontaneously combustible liquids or, in the case of some dilute solutions, as water-reactive, flammable liquids. This classification depends on the performance of the blend in certain tests. For an example of a test, see [86].

Portable steel shipping tanks range in size from small metal cylinders for laboratory and pilot-plant studies to ones with $7.5\,m^3$ capacity (Fig. 6). In addition, shipments are made in tank trailers and rail tank cars in the United States and in some other countries.

International shipments of organoaluminum products in portable tanks are made according to the United Nations Transport of Dangerous Goods Code, wherein they are listed in Table 12.2 [87]. The products appear in the International Maritime Organization IMO-IMDG code portable tank tables under their UN numbers [88]. Alkylaluminums and alkylaluminum halides are identified, respectively, by UN 3051 and UN 3052. Non-pyrophoric alkylaluminum blends and MAO are identified by UN 3207.

In Europe, international agreements cover transport of organoaluminum materials by rail (RID) [89] and by road (ADR) [90], under Class 4.2 and 4.3, Appendix X in both cases. In the United States, they are shipped according to U.S. Department of Transportation (DOT) regulations 49 CFR 173.124 (c) and 173.125 for organometallic compound solutions, water-reactive, flammable and 173.124 (b) (1) and 173.125 for spontaneously combustible liquids [91].

Handling and storage facilities are designed with safe flushing and cleaning in mind, and can be protected from air by a continuous nitrogen flow through the equipment and out through an oil seal drum, under a positive pressure with respect to the outside atmosphere [83], [84], [92]. Fire walls are provided for operator safety and equipment protection. Concrete pads, sloping to a fire-safe area, are recommended under tanks in the plant.

Should lines and tanks in organoaluminum service require cleaning, extreme caution is urged during the cleaning process. A buildup of solids may result from accidental thermal, oxidative, or hydrolytic decompositions. The equipment is cleaned by a combination of inert solvent flushing, steam, water, and an aqueous caustic solution (< 25 wt %). A nitrogen purge must be maintained on the equipment during each stage of the cleaning process to safely vent hydrogen and volatile hydrocarbons that may form.

Spills of organoaluminum materials or their hydrocarbon solutions present special hazards.

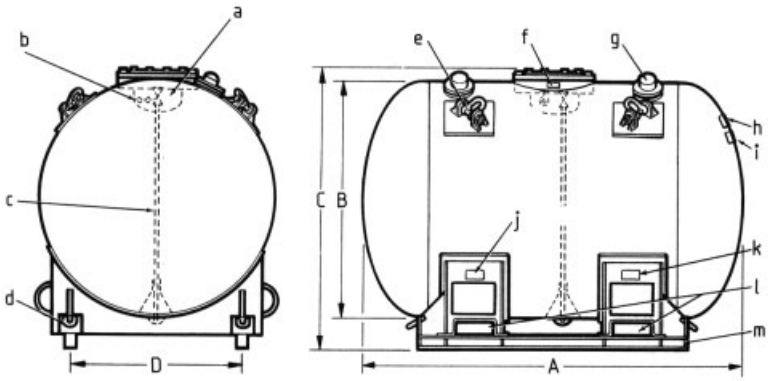

Figure 6. Portable shipping tank used by Albemarle Corporation
Dimensions, (in cm) A 304.5; B 190.0; C 224.0; D 140.0

Tare	2880 kg
Design pressure	12.31 kg/cm^2
Relief valve set pressure	12.31 kg/cm^2
Cubage	13.0 m^3

a) Recessed dome; b) 3/4″ Vapor space connection for nitrogen and vent; c) 1 1/2″ Dip pipe for liquid removal; d) Tie down and towing ring; e) Lifting lugs (4); f) Lloyd's stamp; g) Relief valves (2); h) ASME nameplate; i) DOT nameplate; j) Owner's nameplate; k) Dutch Railroad nameplate; l) Fork lift openings; m) Skids for handling

There is, of course, a danger of spontaneous flaming. Even in the absence of fire, hydrogen, hydrocarbon vapors, and hydrogen halides may be released.

For small, contained spills, absorption in dry powder extinguishing agents or blanketing with vermiculite may be used. Once absorbed or blanketed, the materials may be deactivated by slow aging in air, or by controlled burning.

Dry powder fire extinguishers are effective in the control of small to moderate sized fires resulting from a spill of undiluted or solutions of organoaluminum compounds [93]. Carbon dioxide is somewhat effective on solution fires, but less effective on fires of undiluted organoaluminum compounds. For larger fires, in isolated regions, controlled burning may be the best approach. Application of a water fog has been found effective in alkylaluminum fire control in plant process areas. Water fog systems have been installed in some alkylaluminum handling and reaction systems [83].

The amount and type of personal protective equipment recommended depend on the magnitude of potential exposure [83], [84]. Fire retardant clothing, with long sleeves, face shield, and nonporous gloves that can be removed quickly is suggested when working in the laboratory with small quantities. For larger potential exposure, full suits made of aluminized, fire-resistant material are worn, along with goggles and a hood for head protection, heat-resistant gloves, and combination spats and shoe covers or heavy leather or rubber boots to protect legs and feet.

5. Waste Disposal and Environmental Protection

Some organoaluminum compounds are extremely sensitive to oxygen or water and may spontaneously burst into flames in their presence.

Small amounts of residues from laboratory-scale organoaluminum syntheses or reactions generally can be deactivated to air or moisture by dilution with a hydrocarbon solvent under an inert atmosphere, followed by protonolysis with an alcohol, such as isopropyl alcohol. Finally, water can be added carefully until no further reaction is observed. Larger volumes can be destroyed by incineration. Disposal of the ash from incineration, consisting mainly of metal oxides, should be handled in a manner that is consistent

with all local, state, and federal regulations and environmental laws.

Hydrolysis, under acidic or basic conditions, is an effective method for the destruction of organoaluminum compounds. The alkyl waste should be diluted with a high flash point, low vapor pressure diluent to a metal concentration of less than 5 wt % prior to hydrolysis [83], [84].

Organoaluminum materials produce no unexpected environmental effects. Once in contact with the atmosphere or soil, either they degrade rapidly to hydrocarbons and inorganic products, or, on flaming, they burn to form the expected combustion products.

6. Economic Aspects

Merchant producers of organoaluminum products for industrial use include Albemarle Corp., with plants in Houston, Tex. and Orangeburg, S. C., USA, and Feluy, Belgium; Akzo Nobel, Deer Park, TX, USA and Rotterdam, The Netherlands; Nippon Aluminum Alkyls Ltd., Osaka, Japan; Tosoh Akzo Chemical Co., Yamaguchi, Japan; and Witco AG, Bergkamen, Germany and Mapleton, IL, USA. In addition there is considerable "captive use" production by a number of companies, for making polyolefin catalysts and for intermediates in the synthesis of organic and organometallic products.

Capacity, production, and total consumption figures for organoaluminum materials are not published.

7. Toxicology and Occupational Health

The effects of organoaluminum compounds on the skin can range from a slight irritation to a severe burn, depending on concentration. These burns generally are treated in the same manner as thermal burns; in all cases, the affected areas should be flushed immediately with water.

Breathing fumes produced by organoaluminum materials in contact with air or combustion products from the burning materials should be avoided. Effects of the combustion products on humans are essentially the same as effects of smoke from the burning of organic substances. In addition, "fume fever" may result from prolonged breathing of the aluminum-containing particulate matter. The threshold limit values (TLVs) in air of "Aluminum, alkyls (NOC)" are described by the ACGIH and the NIOSH. A time-weighted average (TWA) for a normal 8-h workday of 2 mg/m^3 of airborne particulate matter, expressed as aluminum, is specified [94], [95].

Acknowledgments

The authors would like to thank Drs. Tom Hanlon, Won Park, Steven Herron, and Ray Dawson for helpful suggestions and comments.

8. References

1. H. Lehmkuhl, K. Ziegler, H.-G. Gellert, "Methoden zur Herstellung und Umwandlung von organischen Aluminium-Verbindungen," in G. Bähr, P. Burba, H. Lehmkuhl, K. Ziegler (eds.): *Houben-Weyl, Methoden der organischen Chemie,* 4th ed., vol. 13, part 4, Georg Thieme Verlag, Stuttgart 1970, p. 1 – 34.
2. T. Mole, E. A. Jeffery: *Organoaluminum Compounds,* Elsevier, Amsterdam 1972.
3. J. R. Zietz, Jr., G. C. Robinson, K. L. Lindsay: "Compounds of Aluminum in Organic Synthesis," in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 7,** Pergamon Press, Oxford 1982, p. 365 – 464.
4. J. J. Eisch: "Aluminum," in G. Wilkinson, P. G. A. Stou, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 1,** Chapter 6, Pergamon Press, Oxford 1982, p. 555 – 682. J. R. Zietz, Jr., G. C. Robinson, K. L. Lindsay in G. Wilkinson, P. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 7,** Chapter 46, Pergamon Press, Oxford 1982, p. 365 – 464.
5. J. J. Eisch: "Aluminum," in E. W. Abel, F. G. A. Stone, G. Wilkinson, A. McKillop (eds.): *Comprehensive Organometallic Chemistry II,* **vol. 11,** Elsevier Science Ltd., Oxford 1995, p. 431 – 502.
6. John J. Eisch: "Hydroalumination of C=C and C≡C," in Barry M. Trost and Ian Fleming (eds.): *Comprehensive Organic Synthesis,* **vol. 8,** Oxford Pergamon Press, New York 1991, p. 733 – 761.

7. P. A. Chaloner: "Preparation and use of organoaluminum compounds in organic synthesis," in F. R. Hartley (ed.): *The chemistry of the metal-carbon bond,* **vol. 4,** John Wiley and Sons, Chichester 1987, p. 411 – 472.
8. *The Use of Aluminum Alkyls in Organic Synthesis*, Ethyl Corp. Brochure, Baton Rouge 1969, Supplement 1969 – 1972, 1972 – 1978.
9. K. Ziegler, *Angew. Chem.* **76** (1964) 545.
10. K. Ziegler, "Organo-Aluminum Compounds," in H. Zeiss (ed.): *Organometallic Chemistry*, Reinhold Publ. Co., New York 1960, Chapter 5, p. 194 – 269.
11. A. Andresen et al., *Angew. Chem. Int. Ed. Engl.* **15** (1976) 630 – 632.
12. M. B. Smith, a. *J. Phys. Chem.* **71** (1967) 364 – 370; b. *J. Organomet. Chem.* **70** (1974) 13 – 33.
13. M. B. Smith, *J. Organomet. Chem.* **70** (1974) 13 – 33.
14. J. Boor: *Ziegler-Natta Catalysts and Polymerizations,* Academic Press, New York 1979.
15. R. M. Laine (ed.): *Inorganic and Organometallic Polymers with Special Properties,* Kluwer, Dordrecht 1992, p. 435.
16. K. Ziegler, H.-G. Gellert, H. Lehmkuhl, W. Pfohl, et al., *Justus Liebigs Ann. Chem.* **629** (1960) 1 – 13.
17. K. Ziegler, H.-G. Gellert, K. Zosel, E. Holzkamp, et al., *Justus Liebigs Ann. Chem.* **629** (1960) 121 – 166.
18. K. Ziegler, F. Krupp, K. Zosel, *Justus Liebigs Ann. Chem.* **629** (1960) 241 – 250.
19. M. B. Smith, Ethyl Corp. unpublished results, April, 1963.
20. J. P. Collman, L. S. Hegedus, J. R. Norton, R. G. Finke: *Principles and Applications of Organotransition Metal Chemistry,* University Science Books, Mill Valley, CA 1987, p. 94.
21. *Kirk-Othmer,* 4th ed., **vol. 24,** 1994, pp. 122 – 161.
22. K. Ziegler, R. Köster, H. Lehmkuhl, K. Reinert, *Justus Liebigs Ann. Chem.* **629** (1960) 33 – 49.
23. J. L. Atwood: "Anionic and Cationic Organoaluminum Compounds," in G. H. Robinson (ed.): *Coordination Chemistry of Aluminum,* VCH Publishers, Inc., New York 1993, pp. 197 – 232.
24. T. R. Crompton, *Anal. Chem.* **39** (1967) 1464 – 1465.
25. N. Davidson, H. C. Brown, *J. Am. Chem. Soc.* **64** (1942) 316 – 324.
26. W. Nagata, M. Yoshioka, *Tetrahedron Lett.* 1966, 1913 – 1918.
27. W. Nagata, M. Yoshioka, *Org. Synth.* **52** (1972) 90 – 95.
28. T. Hirabayashi, H. Imaeda, K. Itoh, S. Sakai, et al., *J. Organomet. Chem.* **19** (1969) 299 – 308.
29. Schering AG, DE 1158517, 1961 (W. Marx); *Chem. Abstr.* **60** (1964) 5550 c.
30. Continental Oil Co., GB 932539, 1963 (F. J. Radd, W. W. Woods); *Chem. Abstr.* **60** (1964) 548 g.
31. Sumitomo Chemical Co., DE 1216304, 1966 (K. Toyoshima, H. Ryu, A. Matsui, E. Ichiki); *Chem. Abstr.* **65** (1966) 3906 f.
32. Goodrich-Gulf Chemicals, US 3100786, 1963 (H. B. Fernald); *Chem. Abstr.* **55** (1961) 8291 i.
33. Ethyl Corp., US 2886581, 1959 (H. E. Redman);
34. Sumitomo Chemical Co., US 3 016 396, 1962 (T. Irie, T. Kagiya, M. Asada).
35. Schering AG, US 3 960 912, 1976 (K. H. Muller, H. J. Hubert).
36. Continental Oil Co.; ZA 635722, 1963 (P. A. Lobo, D. C. Coldiron); *Chem. Abstr.* **63** (1965) 13314 h.
37. Continental Oil Co.; US 3373179, 1968 (A. L. Lewis); *Chem. Abstr.* **69** (1968) 44011 h.
38. A. von Grosse, J. M. Mavity, *J. Org. Chem.* **5** (1940) 106 – 121.
39. Shell Development Co., CA 582180, 582181, 585587, 1959 (R. W. King, D. J. Movsovic).
40. K. Ziegler, H. Martin, F. Krupp, *Justus Liebigs Ann. Chem.* **629** (1960) 14 – 19. K. Ziegler, W.-R. Kroll, W. Larbig, O.-W. Steudel, *Justus Liebigs Ann. Chem.* **629** (1960) 53 – 89.
41. Du Pont, GB 926699, 1963; *Chem. Abstr.* **59** (1963) 4059 e. *Chem. Abstr.* **78** (1973) 30511 v.
42. Schering AG, FR 2 072 306, 1971.
43. K. Ziegler, R. Köster, *Justus Liebigs Ann. Chem.* **608** (1957) 1 – 7.
44. Montecatini, FR 1321239, 1962; *Chem. Abstr.* **59** (1963) 11556 h.
45. Schering AG, DE 1 122 925, 1965.
46. G. R. Lappin, J. D. Sauer (eds.): *Alpha Olefins Applications Handbook,* Marcel Dekker, Inc., New York 1989.
47. *Kirk-Othmer,* 4th ed., **vol. 12,** pp. 1071 – 1091.
48. K. Ziegler, H.-G. Gellert, *Justus Liebigs Ann. Chem.* **629** (1960) 20 – 22.
49. W. P. Neumann, *Justus Liebigs Ann. Chem.* **629** (1960) 23 – 33. E. Bonitz, *Chem. Ber.* **88** (1955) 742 – 747.
50. L. S. Simeral, *Today's Chemist At Work* **2** (1993) Nov/Dec, 52 – 57.

51. S. Pasynkiewicz, *Polyhedron* **9** (1990) 429 – 453.
52. Albemarle Corporation, US 5 599 964, 1997 (J. K. Roberg, R. E. Farritor, E. A. Burt).
53. Akzo Nobel N.V., WO 97/23288, 1997 (Gregory M. Smith et al.).
54. K. Ziegler, F. Krupp, K. Weyer, W. Larbig, *Justus Liebigs Ann. Chem.* **629** (1960) 251 – 256.
55. Texas Alkyls, Inc., US 4 960 878, 1990 (Clark C. Crapo, Dennis B. Malpass).
56. Texas Alkyls, Inc., US 5 041 584, 1991 (Clark C. Crapo, Dennis B. Malpass).
57. Ethyl Corporation, US 5 157 008, 1992 (S. A. Sangokoya, M. S. Howie, A. L. Dunaway).
58. M. R. Mason, J. M. Smith, S. G. Bott, A. R. Barron, *J. Am. Chem. Soc.* **115** (1993) 4971 – 4984.
59. D. W. Imhoff, L. S. Simeral, S. A. Sangokoya, J. H. Peel, *Organometallics* **17** (1998) 1941 – 1945.
60. Exxon Chemical Patents, Inc., US 5 324 800, (H. C. Welborn, Jr., J. A. Ewen).
61. G. M. Benedikt, B. L. Goodall, L. H. McIntosh III, L. F. Rhodes, L. M. Wojcinski II: "Homo- and Co-polymers derived from Mulitcyclic Olefin Monomers: The Quest for Higher T_g Materials," in G. M. Benedikt, B. L. Goodall (eds.): *Metallocene-Catalyzed Polymers,* Plastics Design Library, Norwich, NY 1998 pp. 57 – 62.
62. Michael Freemantle, *Chem. Eng. News* (1998) April 13, 11 – 12.
63. J. A. Ewen, *J. Amer. Chem. Soc.* **106** (1984) 6355 – 6364.
64. X. Yang, C. L. Stern, T. J. Marks, *Angew. Chem. Int. Ed. Engl.* **31** (1992) 1375 – 1377.
65. C. Sishta, R. M. Hathorn, T. J. Marks, *J. Am. Chem. Soc.* **114** (1992) 1112 – 1114.
66. H. Sinn, W. Kaminsky, *Adv. Organomet. Chem.* **18** (1980) 99 – 149.
67. R. F. Jordan, C. S. Bajgur, R. Willett, B. Scott, *J. Am. Chem. Soc.* **108** (1986) 7410 – 7411.
68. Exxon Chemical Patents, Inc., US 5 198 401, 1993 (Howard W. Turner, G. G. Hlatky, R. R. Eckman).
69. Exxon Chemical Patents, Inc., US 5 599 761, 1997 (Howard W. Turner).
70. X. Yang, C. Stern, T. J. Marks, *J. Am. Chem. Soc.* **113** (1991) 3623 – 3625.
71. W. Kaminsky, K. Kulper, H. H. Brintzinger, F. R. W. P. Wild, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 507 – 508.
72. W. E. Piers, K. Kulper, J. E. Bercaw, *J. Am. Chem. Soc.* **112** (1990) 9406 – 9407.
73. The Dow Chemical Company, US 5 064 802, 1991 (J. C. Stevens, D. R. Neithamer).
74. Y. V. Kissin, *J. Polym. Sci.: Part A: Polym. Chem.* **33** (1995) 227 – 237.
75. *Plastics Technology,* 1991, November, p. 15 – 19.
76. K. Swogger, *Rubber & Plastics News,* 1994, Feb. 28, p. 36 – 38.
77. H. H. Brintzinger, D. Fischer, R. Mulhaupt, B. Rieger, R. M. Waymouth, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 1143 – 1170.
78. *Plastics Technology,* 1993, April, p. 28 – 32.
79. Uniroyal Chemical Company, WO 9 738 019, 1997 (D. J. Gillis, R. Karpeles).
80. *Plastics Technology,* 1993, Feb., p. 26 – 30.
81. J. S. Rogers, G. C. Bazan, C. K. Sperry, *J. Am. Chem. Soc.* **119** (1997) 9305 – 9306.
82. C. M. Killian, L. K. Johnson, M. Brookhart, *Organometallics* **16** (1997) 2005 – 2009.
83. Albemarle Corp., "Handling Procedures for Aluminum Alkyl Compounds and Methylaluminoxane (MAO)."
84. Akzo Nobel, "Metal Alkyls". Witco, "Properties and Handling, Aluminum Alkyls."
85. W. H. Thomas, *Ind. Eng. Chem. Prod. Res. Dev.* **21** (1982) 120 – 122.
86. W. L. Mudry, D. C. Burleson, D. B. Malpass, S. C. Watson, *J. Fire & Flammability,* **6** (1975) 478 – 487.
87. U. N. Committee of Experts on the Transport of Dangerous Goods (ed.): *Transport of Dangerous Goods,* Table 12.2, United Nations Publication, New York 1993, p. 345.
88. IMO (ed.): *IMCO,* revised ed., London 1996, p. 008701.
89. RID, 7th ed., Appendix 10, Her Majesty's Stationery Office, London 1978, p. 365 ff.
90. ADR, United Nations, London 1978, Marginal 2430 and Appendix Blb, Marginal 212400 ff.
91. DOT: CFR 49 (Transportation), U.S. Government Printing Office, Washington, D.C., 1997, paragraphs 173.124(b) (1), 173.124 (c), 173.125.
92. L. J. Governale, W. B. Ruhlin, R. F. Silvus, *Chem. Eng. Prog.* **61** (1965) 105 – 110.
93. W. B. Heck, Jr., R. L. Johnson, *Ind. Eng. Chem.* **54** (1962) 35 – 38. E. A. Zeratsky, in C. H. Vervalin (ed.): Fire Protection Manual,Gulf Publishing, Houston 1973, p. 376 – 377.
94. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 4th ed., **vol. 2 C,** J. Wiley & Sons, New York 1994, p. 1886, 1899 – 1900.
95. American Conference of Governmental Industrial Hygienists (ed.): Guide to Occupational Exposure Values-1998, Cincinnati, OH 1998, p. 4.

Aluminum Oxide

L. Keith Hudson, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, USA

Chanakya Misra, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, United States

Anthony J. Perrotta, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, USA

Karl Wefers, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, USA

F. S. Williams, Aluminum Company of America, Alcoa Center, Pennsylvania 15069, USA

1. General Aspects 1
1.1. Aluminum Hydroxides 2
1.2. Aluminum Oxide Hydroxides . . . 3
1.3. Aluminum Oxide, Corundum . . . 4
1.4. The $Al_2O_3 - H_2O$ System 4
1.5. Thermal Decomposition of Aluminum Hydroxides 5
1.6. Aluminates and Related Compounds 7
2. Bauxite, the Principal Raw Material 7
2.1. Definition and Geology 7
2.2. Composition and Properties 7
2.3. Genesis of Bauxites 9
2.4. Major Bauxite Deposits 10
2.5. Economic Aspects 11
3. Bayer Process 11
3.1. History and Procedure 11
3.1.1. Bauxite Preparation 12
3.1.2. Digestion 13
3.1.3. Equipment 15
3.1.4. Residue Separation 17
3.1.5. Precipitation 20
3.1.6. Impurities 23
3.1.7. Calcination 25
3.1.8. Evaporation 26
3.1.9. Residue Disposal 27
3.1.10. Energy in the Process 28
3.2. Economic Aspects 29
4. Other Processes for Alumina Production 29
4.1. Raw Materials 29
4.2. Alkaline Processes 30
4.3. Acid Processes 31
5. Metallurgical Alumina 32
5.1. Alumina Properties Required for Smelting 32
5.2. Typical Specifications for Metallurgical Alumina 33
6. Industrial Alumina Chemicals . . 33
6.1. Aluminum Hydroxides 33
6.2. Adsorbent and Catalytic Aluminas 34
6.2.1. Preparation of Activated Aluminas . 34
6.2.2. Adsorbent Applications 36
6.2.3. Catalytic Applications 37
7. Ceramic Uses of Alumina 38
7.1. Calcined Alumina 39
7.2. Fused Alumina 39
8. Toxicology and Industrial Hygiene 39
9. References 39

1. General Aspects

Almost 4000 years ago Egyptians and Babylonians used aluminum compounds in various chemicals and medicines. Herodotus mentioned alum in the fifth century B.C. and Pliny referred to "alumen," now known as alum, as a mordant to fix dyes to textiles around 80 A.D. In 1754 Marggraf showed that a distinct compound existed in both alum and clays. In 1761 the French chemist Guyton de Morveau proposed the name "alumine" for the base in alum, identified in 1787 by Antoine Lavosier as the oxide of a then-undiscovered element. By the 1700s the earthy base alumina was recognized as the potential source of a metallic element.

Greville (1798) described a mineral from India that had the composition Al_2O_3 and named it corundum [*1302-74-5*], derived from the native name of this stone. Haüy (1801) called a mineral diaspore [*14457-84-2*] (from the Greek "diaspora" meaning dispersion) because it decrepitated on heating. Its composition, $Al_2O_3 \cdot H_2O$, was determined by Vaquelin in 1802. Gibbsite [*14762-49-3*], named after the American mineralogist G. Gibbs, was found

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a01_557

by DEWEY in 1820; TORREY (1822) showed this mineral to have the composition $Al_2O_3 \cdot 3\,H_2O$. The name hydrargillite [*14762-49-3*] was given to a similar mineral found later in the Ural Mountains. Using the newly developed technique of X-ray diffraction, BÖHM and NICLASSEN [13] identified a crystalline aluminum oxide hydroxide, $Al_2O_3 \cdot H_2O$, later named böhmite (boehmite). BÖHM discovered a second type of trihydroxide, $Al_2O_3 \cdot 3\,H_2O$, a year later. FRICKE [14] suggested the name bayerite [*20257-20-9*] for this compound, believing it to be the product of the Bayer process, which he later identified as gibbsite. Only few occurrences of natural bayerite have been reported [15]. VAN NORDSTRAND et al. [16] reported a third form of trihydroxide, which was later named nordstrandite [*13840-05-6*] in his honor.

1.1. Aluminum Hydroxides

A general classification of the various modifications of aluminum hydroxides is shown in Figure 1. The best defined *crystalline forms* are the three trihydroxides, $Al(OH)_3$: gibbsite, bayerite, and nordstrandite, and two modifications of aluminum oxide hydroxide, AlO(OH): boehmite and diaspore. Besides these well-defined crystalline phases, several other forms have been described in the literature [2], [3]. However, there is controversy as to whether they are truly new phases or simply forms with distorted lattices containing adsorbed or interlamellar water and impurities.

Identification of the different hydroxides and oxides is best carried out by X-ray diffraction methods [3]. Mineralogical and structural data are listed in Tables 1 and 2.

Gelatinous hydroxides may consist of predominantly X-ray indifferent aluminum hydroxide or pseudoboehmite. The X-ray diffraction pattern of the latter shows broad bands that coincide with strong reflections of the well-crystallized oxide hydroxide boehmite.

The aluminum hydroxides found abundantly in nature are gibbsite, boehmite, and diaspore. *Gibbsite* and *bayerite* have similar structures. Their lattices are built of layers of anion octahedra in which aluminum occupies two thirds of the octahedral interstices. In the gibbsite structure, the layers are somewhat displaced relative to one another in the direction of the *a* axis. The hexagonal symmetry of this lattice type (brucite type) is lowered to monoclinic. Triclinic symmetry was found in larger gibbsite single crystals from the Ural Mountains [17].

In bayerite the layers are arranged in approximately hexagonal close packing. Because of shorter distances between the layers, the density is higher than in the case of gibbsite. The crystal class and space group of bayerite have not yet been established clearly.

The individual layers of hydroxyl ion octahedra in both the gibbsite and the bayerite structures are linked to one another only through weak hydrogen bonds. Bayerite does not form large single crystals. The most commonly observed growth forms are spindle- or hourglass-shaped somatoids. The long axis of these somatoids stands normal to the basal plane; i.e., the somatoids consist of stacks of $Al(OH)_3$ layers (Fig. 2). The effect of alkali ions on the structures of $Al(OH)_3$ types was investigated by several workers [3], [18], [19]. Intercalation of Li^+ ions transforms gibbsite to the hydrotalcite structure [20].

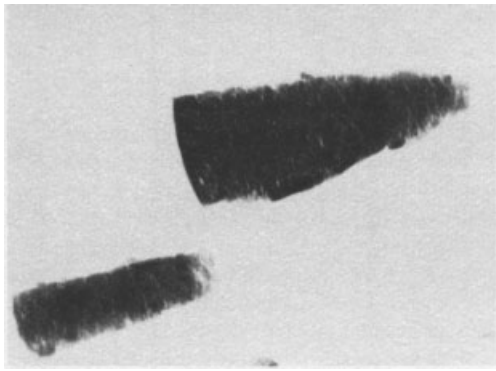

Figure 2. Somatoids of bayerite

Gibbsite crystals of appreciable size are not uncommon. Clear pseudohexagonal platelets about 1 mm in diameter are known from Arö in Norway. Prismatic crystals 0.5 – 1 mm in length are occasionally produced in the Bayer process.

Nordstrandite, the third form of $Al(OH)_3$, was described by VAN NORDSTRAND [16] and others [21]. The structures of nordstrandite and bayerite were investigated [22] and compared with those of monoclinic and triclinic gibbsite, which had been determined previously [17].

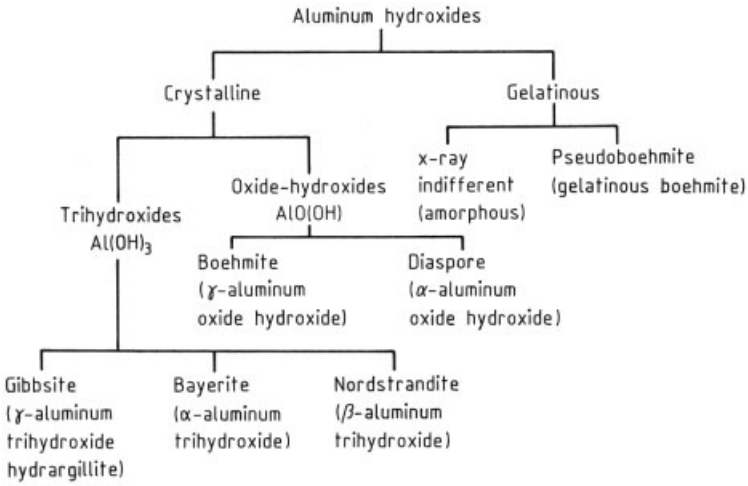

Figure 1. Classification of aluminum hydroxides
(All figures courtesy of Aluminum Company of America)

Table 1. Mineralogic properties of oxides and hydroxides [3]

Phase	Refractive index n_D^{20}				Cleavage	Brittleness	Mohs hardness	Luster
	α	β	γ	Average				
Gibbsite	1.568	1.568	1.587	–	(001) perfect	tough	2.5 to 3.5	pearly vitreous
Bayerite	–	–	–	1.583	–	–	–	–
Boehmite	1.649	1.659	1.665	–	(010)	–	3.5 to 4	–
Diaspore	1.702	1.722	1.750	–	(010) perfect	brittle	6.5 to 7	brilliant pearly
	ε	ω		Average				
Corundum	1.7604	1.7686		–	none	tough when compact	9	pearly adamantine

The lattice of nordstrandite is built of the same, electrically neutral $Al(OH)_3$ octahedral layers that form the structural elements of gibbsite and bayerite [14]. The lattice period amounts to 1.911 nm in the direction normal to the layer. This corresponds to the sum of identical layer distances of bayerite plus gibbsite. The identical nordstrandite structure consists of alternating double layers, in which the OH octahedra are arranged once in the packing sequence of bayerite, and then in that of gibbsite. Material containing continuous transitions from bayerite through nordstrandite to gibbsite has been prepared through the proper selection of precipitation conditions [18].

1.2. Aluminum Oxide Hydroxides

Pseudoboehmite is formed during aging of X-ray indifferent hydroxide gels as a precursor of trihydroxide. The reflections of pseudoboehmite are broadened not only because of the very small particle size, but also because of variable distances of the AlO(OH) double chains, which form the structural element of pseudoboehmite as well as of well-crystallized boehmite. The lattice spacing in the direction of the *c* axis increases by 0.117 nm for each mole of excess water [22].

Boehmite consists of O, OH double layers in which the anions are in cubic packing.

Table 2. Structural properties of oxides and hydroxides [3]

Phase	Formula	Crystal system	Space group	Molecules per unit cell	Unit axis length, $\times 10^{-1}$ nm			Angle	Density, g/cm^3
					a	*b*	*c*		
Gibbsite	$Al(OH)_3$	monoclinic	C_{2h}^5	4	8.68	5.07	9.72	94°34′	2.42
Gibbsite	$Al(OH)_3$	triclinic	–	16	17.33	10.08	9.73	94°10′ 92°08′ 90°0′	–
Bayerite	$Al(OH)_3$	monoclinic	C_{2h}^5	2	5.06	8.67	4.71	90°16′	2.53
Nordstrandite	$Al(OH)_3$	triclinic	C_i^1	4	8.75	5.07	10.24	109°20′ 97°40′ 88°20′	–
Boehmite	AlO(OH)	orthorhombic	D_{2h}^{17}	2	2.868	12.227	3.700	–	3.01
Diaspore	AlO(OH)	orthorhombic	D_{2h}^{16}	2	4.396	9.426	2.844	–	3.44
Corundum	Al_2O_3	hexagonal (rhomb.)	D_{3d}^6	2	4.758	–	12.991	–	3.98

The aluminum ions are octahedrically coordinated. These layers are composed of chains of $[AlO(OH)]_2$ extending in the direction of the *a* axis [23]. The double layers are linked by hydrogen bonds between hydroxyl ions in adjacent planes. Boehmite crystals exhibit perfect cleavage parallel to the (010) plane.

In the *diaspore* structure the oxygen ions are nearly equivalent, each being joined to another oxygen center through a hydrogen ion. The anions are hexagonally close packed [24]. The position of the hydrogen ion has been established by neutron diffraction [25]. The O – H – O distance is 0.265 nm. By infrared studies, the bond energy for the hydrogen bridges in diaspore was determined to be 28.7 kJ/mol, compared with 20.1 kJ/mol for boehmite [26].

1.3. Aluminum Oxide, Corundum

The hexagonally closest packed α-Al_2O_3 modification is the only stable oxide in the $Al_2O_3 - H_2O$ system. Corundum is a common mineral in igneous and metamorphic rocks. Red and blue varieties of gem quality are called ruby and sapphire, respectively. The lattice of corundum is composed of hexagonally closest packed oxygen ions forming layers parallel to the (0001) plane. Only two-thirds of the octahedral interstices are occupied by aluminum ions. The structure may be described roughly as consisting of alternating layers of Al and O ions. The corundum structure was determined in the early 1920s [27]; numerous workers later confirmed and refined these data [3]. Properties of corundum are listed in Tables 1 and 2.

1.4. The $Al_2O_3 - H_2O$ System

Under the equilibrium vapor pressure of water, crystalline $Al(OH)_3$ converts to AlO(OH) at about 375 K. The conversion temperature appears to be the same for all three forms of $Al(OH)_3$. At temperatures lower than 575 K, boehmite is the prevailing AlO(OH) modification, unless diaspore seed is present. Spontaneous nucleation of diaspore requires temperatures in excess of 575 K and pressures higher than 20 MPa. In the older literature, therefore, diaspore was considered the high-temperature form of AlO(OH). The first reaction diagram of the phase transitions in the $Al_2O_3 - H_2O$ system was published in 1943 [28]. These workers determined the gibbsite $\rightarrow$ boehmite conversion temperature to be 428 K. Boehmite transformed to diaspore above 550 K; diaspore converted to corundum, α-Al_2O_3, at 725 K. Similar results were reported in 1951 [29].

The system was reinvestigated in 1959 [30] and in 1965 [31]. A phase diagram based on these data is shown in Figure 3. Diaspore is

the stable modification of AlO(OH); boehmite is considered metastable, although it is kinetically favored at lower temperatures and pressures. This is because the nucleation energy is lower for boehmite than for the considerably more dense diaspore. Nucleation is additionally facilitated by the possibility that boehmite can grow epitaxially on $Al(OH)_3$. In the $Al_2O_3 - Fe_2O_3 - H_2O$ system, the presence of the isostructural goethite, α-FeO(OH), lowers the nucleation energy for diaspore so that this AlO(OH) modification crystallizes at temperatures near 373 K [32]. This observation explains the occurrence of diaspore in clays and bauxite deposits that have never been subjected to high temperatures or pressures.

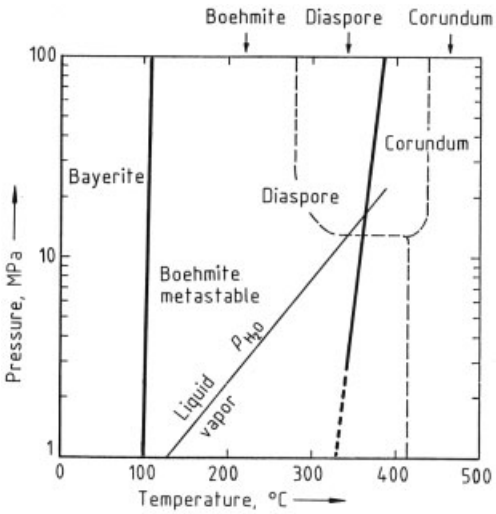

Figure 3. The $Al_2O_3 - H_2O$ system
Dashed lines [21], solid lines [22], [28]

Nomenclature. Although there is fairly good agreement in the more recent literature on phase fields and structures of the crystalline phases in the $Al_2O_3 - H_2O$ system, the nomenclature is still rather unsystematic.

Bayerite, gibbsite (hydrargillite), and nordstrandite are trihydroxides of aluminum, and not oxide hydrates. The designation "aluminum oxide monohydrate" for boehmite and diaspore is also incorrect. Both are true oxide hydroxides. Molecular water has been determined only in poorly crystallized, nonstoichiometric pseudoboehmite.

The designation of the modifications of aluminum hydroxides and oxides lacks uniformity just as much as does the nomenclature of the compounds. According to the general usage in crystallography, the most densely packed structures are designated as α-modifications [3]. Bayerite, diaspore, and corundum fall within this class. The compounds with cubic packing sequence, gibbsite and boehmite, have been designated by the symbol γ. Nordstrandite can be classified as β-$Al(OH)_3$ when regarding this compound not as an intergrowth of bayerite and gibbsite, but as an independent modification.

1.5. Thermal Decomposition of Aluminum Hydroxides

When aluminum hydroxides or oxide hydroxides are heated in air at atmospheric pressure, they undergo a series of compositional and structural changes before ultimately being converted to α-Al_2O_3. These thermal transformations are topotactic. Despite a loss of 34 or 15 % of mass for the trihydroxides or oxide hydroxides, respectively, the habit of the primary crystals and crystal aggregates changes very little. This leads to considerable internal porosity, which may increase the specific surface area of the material to several hundred m^2/g. Structural forms develop that, although not thermodynamically stable, are well reproducible and characteristic for a given temperature range and starting material. These transition aluminas have been the subject of numerous investigations because of their surface activity, sorptive capacity, and usefulness in heterogeneous catalysis. The literature in this field of physical chemistry has been reviewed up to 1987 [3].

The simplest transformation is that of diaspore to corundum. As the structures of these two compounds are very similar, the nucleation of α-Al_2O_3 requires only minor rearrangement of the oxygen lattice after the hydrogen bonds are broken. A temperature below 860 – 870 K is sufficient for complete conversion. The newly formed corundum grows epitaxially on the decomposing diaspore, with the (0001) plane of Al_2O_3 parallel to the (010) plane of AlO(OH) [33]. Transformation to corundum (α-Al_2O_3) proceeds through an intermediate α'-Al_2O_3 phase [34].

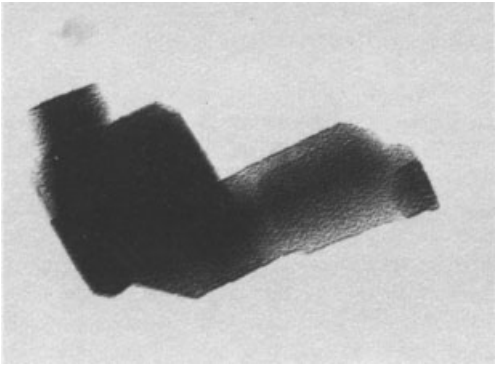

Figure 4. Gibbsite heated to 573 K

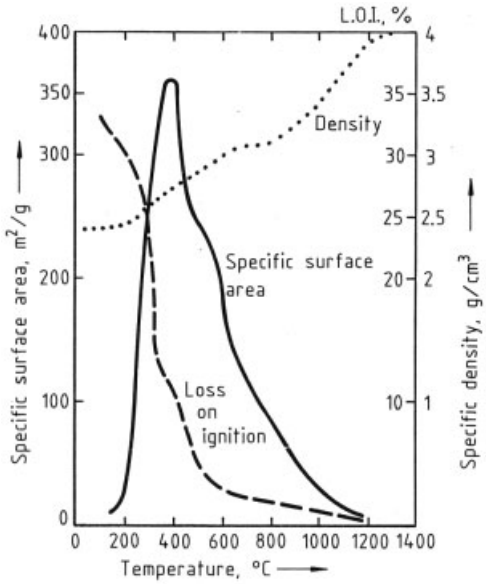

Figure 5. Specific surface area, loss on ignition (LOI), and density of heated $Al(OH)_3$

The thermal transformation, at ambient pressure, of boehmite and the trihydroxides to α-Al_2O_3 requires considerably more structural rearrangements and is generally not completed until the temperature reaches at least 1375 – 1400 K. The first step in the reaction sequence is the diffusion of protons to adjacent OH groups and the subsequent formation of water [35], [36]. This process begins at a temperature near 475 K. If this water cannot diffuse rapidly out of larger trihydroxide particles, hydrothermal conditions may develop locally, resulting in the formation of γ-AlO(OH). With increasing loss of water, a large internal porosity develops. The lattice voids left by the escaping water are not readily healed because of the slow diffusion in this low temperature range. The voids are oriented parallel and perpendicular to the basal plane of the trihydroxide crystals (Fig. 4).

The highest surface area and lowest crystalline order of the solid (not counting newly formed boehmite) is obtained at a temperature around 675 K. With increasing temperature the surface area decreases, while the density of the solid shows progressively higher values (Fig. 5). This trend is the result of progressive reordering and consolidation of the solid.

During the thermally driven consolidation and reordering, the solid goes through structural stages that are influenced by the nature of the starting material as well as by heating rates, furnace atmosphere, and impurities. The general reaction paths are illustrated in Figure 6, which shows the various intermediate transition forms that have been identified during the reordering process.

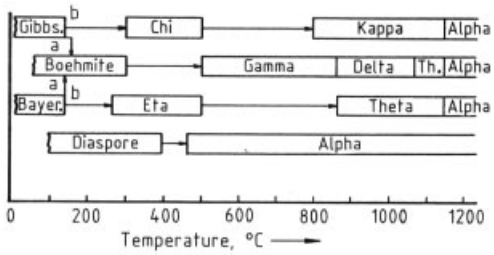

Figure 6. Decomposition sequence of aluminum hydroxide

Transition oxides formed at lower temperatures are mostly two-dimensional, short-range ordered domains within the texture of the decomposed hydroxides. Extensive three-dimensional ordering begins at about 1050 K. Until completely converted to corundum, the solid retains considerable amounts of OH^- ions. Most likely protons are retained to maintain electroneutrality in areas deficient of cations. Therefore, the presence of protons may retard the reordering of the cation sublattice. The high surface area ($> 75\,m^2/g$) of γ-Al_2O_3 has been shown to provide thermodynamic stability [37]. Addition of fluorine to the furnace atmosphere removes protons. As a result, rapid transition to α-Al_2O_3 occurs at temperatures as low as

1150 K. Markedly tabular corundum crystals form, possibly because the preceding transition alumina is mostly two-dimensionally ordered [38].

Transition forms other than those shown in Figure 6 can be obtained by hydrothermal treatment [3]. The structures of various transition forms have been investigated [17], [22], [39].

1.6. Aluminates and Related Compounds

Sodium oxide forms several compounds with aluminum oxide. These so-called β-aluminas represent a group of aluminates having the same, or very similar, structures but variable chemical composition. Their molar ratios ($Na_2O : Al_2O_3$) can vary between 1 : 1 and 1 : 11. The 1 : 1 sodium aluminate, $NaAlO_2$ [*1302-42-7*], exists in at least two allotropic modifications. Orthorhombic β-$NaAlO_2$ is stable below 750 K; the higher – temperature γ modification is tetragonal. For preparation and properties of technical sodium aluminate, → Aluminum Compounds, Inorganic, Chap. 2.1.

The 1 : 11 β-Al_2O_3 crystallizes from melts containing aluminum oxide and sodium oxide or other sodium compounds. A 1 : 5 β-alumina has been prepared by heating α-Al_2O_3 with $NaAlO_2$ or Na_2CO_3 at about 1325 K.

Several other β-aluminas containing CaO, BaO, or SrO in 1 : 6 ratio also have been reported [3].

β-Alumina is of interest to its use as a solid electrolyte in sodium – sulfur secondary batteries (→ Batteries).

2. Bauxite, the Principal Raw Material

2.1. Definition and Geology

The term bauxite [*1318-16-7*] is used for sedimentary rocks that contain economically recoverable quantities of the aluminum minerals gibbsite, boehmite, and diaspore. The name derives from the description by BERTHIER, in 1821, of a sediment that occurred near the village of Les Baux in the Provence, France. Originally assumed to be a dihydrate of aluminum oxide, bauxite was later recognized as being composed of aluminum hydroxide, iron oxide and hydroxide, titanium dioxide, and aluminosilicate minerals [2].

Early in this century, major bauxite deposits were found in various parts of the Tertiary and Cretaceous limestone formations of the European Alps; also in several locations on the North American continent, e.g., in Arkansas, Alabama, and Georgia. Since the 1920s, extensive deposits have been discovered in the tropical and subtropical climate belts.

The oldest known bauxites developed in the Precambrian; the youngest are of recent origin. Deposits may occur as extensive, flat bodies blanketing areas of many square miles; they may form irregularly shaped fillings of dolinas in old karst surfaces, or lenses several hundred meters in diameter and tens of meters thick. Allochtonous, i.e., displaced bauxites, are also common; erosion of primary deposits and redeposition of the detritus in valleys or along mountain slopes has frequently led to mineable accumulations of ore.

Many of the geologically younger bauxites are covered only by thin layers of soil; others are buried by coastal or alluvial sediments. Older deposits, especially in the Balkans or the Ural Mountains, often are overlain by carbonate rocks hundreds of meters thick. Regardless of their geologic age, all bauxites were formed during continental periods [40].

2.2. Composition and Properties

The chemical composition of bauxites of various origin is given in Table 3 [2]. Aluminum oxide, iron oxide, and titanium and silicon dioxides are the major chemical components of all bauxites. Alkali and alkaline earth compounds are rarely found.

Gibbsite, γ-$Al(OH)_3$, is the predominant aluminum mineral in the geologically young bauxites of the tropical climate belt. Mesozoic and older bauxite contain mostly boehmite (γ-AlO(OH)) or diaspore (α-AlO(OH)). Since their formation, many of the older bauxites have been buried under considerable layers of younger sediments and often were subjected to tectonic stress. Boehmite and diaspore formation appears

Table 3. Principal chemical constituents of various bauxites, wt %

Country and location	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	Loss on ignition
Australia					
Darling Range	37	26.5	16.4	1.1	19.3
Weipa	58	4.5	6.9	2.5	26.8
Brazil					
Trombetas	52	5.1	13.9	1.2	28.1
France					
Southern districts	57	4.6	22.6	2.9	15.1
Guyana					
Mackenzie	59	4.9	2.9	2.4	30.4
Guinea					
Friguia	49	6.1	14.2	1.6	28.1
Boke	56	1.5	7.9	3.7	30.1
Hungary					
Halimba	52	6.6	23.5	2.9	18.1
India					
Orissa	46	2.7	22.4	1.1	24.2
Indonesia					
Bintan	53.5	3.9	12.1	1.6	29.2
Jamaica					
Clarendon	47.8	2.6	17.6	2.3	27.3
Surinam					
Onverdacht	59	4.3	3.1	2.5	30.9
Moengo	54	4.2	10.4	2.8	28.9
United States					
Arkansas	51	11.2	6.6	2.2	28.4
Former Soviet Union					
Severouralsk	54	6.2	14.8	2.4	15.7
Former Yugoslavia					
Mostar	52	3.9	21.2	2.7	16.2

to be related to an increasing degree of metamorphism. However, both minerals also occur in young deposits, although in minor quantities. The chemical environment obviously plays as important a role in the formation of boehmite and diaspore as do pressure and stress [40].

Dissolution of gibbsite requires the mildest conditions in the Bayer process (see Section 3.1.2). Higher temperatures and alkali concentrations are necessary for the digestion of boehmite and diaspore. Technically, both oxide hydroxides can be processed without difficulties. The abundance of high-grade gibbsitic bauxites, however, has made boehmite- and diaspore-rich ores economically less attractive.

Goethite, α-FeO(OH) [*1310-14-1*], and hematite, α-Fe_2O_3 [*1309-37-1*], are the most prevalent iron minerals in bauxites. They are practically inert under the conditions of the Bayer process. In both materials, some of the iron may be replaced isomorphically by aluminum ions. This amount of aluminum is included in the chemical analysis of the bauxite but is normally not extracted in the digest. Magnetite (Fe_3O_4) is found in some European bauxites; pyrite (FeS_2) and siderite ($FeCO_3$) also may occur. Decomposition of pyrite may lead to high sulfur levels in the process solutions.

Anatase, TiO_2 [*1317-70-0*], is the titanium mineral found most frequently in bauxites. Rutile, TiO_2 [*1317-80-2*], occurs in some European

deposits; $FeTiO_3$ is also present in minor quantities, especially in titanium-rich bauxites. Titanium dioxide minerals are attacked under only the most severe conditions of the Bayer process.

Silicon dioxide may occur as quartz [*14808-60-7*], as in the bauxites of the Darling Range in Western Australia. Most commonly, however, SiO_2 is associated with the clay minerals kaolinite, halloysite, or montmorillonite. These aluminosilicates react with sodium aluminate solutions to form insoluble sodium aluminum silicates during digest, causing loss of sodium hydroxide and extractable alumina. The amount of so-called reactive silica is one of the major factors determining quality and price of the ore.

Minor constituents, such as chromium, vanadium, zinc, and gallium, have little effect on the Bayer process or on the quality of the final product. Some tend to accumulate in the recirculated process solutions (e.g., gallium and vanadium) and must be removed periodically by appropriate treatments.

The physical properties of bauxites, i.e., texture, hardness, and density, can vary widely. Geologically old diaspore bauxites, especially those high in iron oxide, are very hard and can reach densities of 3.6 g/cm^3. Young tropical deposits, in contrast, may have an earthy, soft texture and a density around 2.0 – 2.5 g/cm^3. Allochtonous bauxite often consists of hard nodules embedded in a soft, usually clayey matrix. Porous textures also occur. The color of bauxites is largely determined by the type and particle size of the prevalent iron mineral. Highly dispersed goethite tends to be yellow to orange, whereas dark brown tones usually are associated with coarser hematite. Colors can vary greatly within a single ore body.

Hardness, texture, and the amount of overburden determine the methods applied for bauxite mining. Deposits in Greece, former Yugoslavia, Hungary, and the former Soviet Union require deep mining to depths of several hundred meters, often complicated by the difficulties of controlling water levels in the porous limestone formations. Tropical bauxites frequently are located so close to the surface that they can be recovered with normal earth-moving equipment.

2.3. Genesis of Bauxites

Many of the geologically young bauxite deposits are located in the savannah region, which extends north and south of the tropical rain forest belt. The climate of this region is characterized by a high mean annual temperature and abundant precipitation during the rainy season. Deposits occur on gently sloping hills or on peneplains. The stratigraphic evidence shows that these bauxites have formed in situ. Parent rocks may be coarse-grained, igneous rocks such as syenite, phonolite, basalt, or gabbro. However, large deposits also developed on kaolinitic sandstones, on phyllites, and on schists. A layer of kaolinitic clay is frequently found between the ore body and the parent rock.

Bauxite probably forms during long periods of low geologic activity when the combination of high temperature, abundant precipitation, and good vertical drainage favors intensive chemical weathering. The sequence of leaching begins with removal of alkali followed by the removal of alkaline earths. Oxides of iron, aluminum, titanium, and silicon are mobilized and reprecipitated as hydroxides and oxides. Aluminum hydroxide and silica form kaolinite, $Al_4(OH)_8Si_4O_{10}$. This sequence first leads to the formation of tropical soils (laterites). Bauxitization follows when the climatic, chemical, and topographic conditions prevail long enough to allow the removal of silicon dioxide as well.

The bauxite deposits of the Mediterranean region, the Caribbean Islands, and many other locations that are associated with tertiary and older limestone formations (karst bauxites) were formed by a similar weathering process. Parent materials were lateritic soils and clays transported into the karst region and deposited in depressions. During extended terrestrial periods, high mean temperatures, copious precipitation, and good vertical draining through the porous limestone bedrock facilitated a thorough desilification. In the older literature, the parent material for karst bauxitization was reported to be clayey residue left after weathering of substantial layers of carbonate rocks [41]. Researchers have shown conclusively that igneous rocks were the source of this material [42–44].

Comprehensive reviews on the geology, mineralogy, and genesis of bauxites have been published [40], [45].

Table 4. World bauxite reserves and production, 10^3 t [46]

Country	Mine production		Reserves
	1996	1997	
United States	n.a.	n.a.	40 000
Australia	43 100	43 500	7 900 000
Brazil	9700	9700	2 900 000
China	6 200	7000	2 000 000
Guinea	14 000	14 000	900 000
Guyana	2000	2000	900 000
India	5100	5500	200 000
Jamaica	11 829	12 000	2 000 000
Russia	3300	3300	200 000
Surinam	4000	4000	600 000
Venezuela	5600	5600	350 000
Other countries	8928	8900	4 400 000
World total (rounded)	114 000	115 000	28 000 000

2.4. Major Bauxite Deposits

Until the 1950s, the European aluminum industry was supplied from the karst bauxite deposits of France, Hungary, former Yugoslavia, and Greece. United States sources (Arkansas) and ore from Surinam provided the raw material for North American production. Since then, the picture has changed dramatically. The four largest bauxite producers of 1990, namely, Australia, Guinea, Jamaica, and Brazil (Table 4), hardly would have been mentioned in 1950. Today, they provide more than half of the world's total bauxite output.

Africa. The Savannah region covers an area of the African continent that has experienced low geologic activity for a long time. In this belt, which stretches from the Ivory Coast to Madagascar, very large bauxite deposits were found. The major production is currently concentrated in Guinea and Ghana. Cameroon, Sierra Leone, Mali, and the Congo region, among other areas, have substantial reserves.

Australia. Major deposits are located in the Darling Range in Western Australia, on the Gove Peninsula in the Northern Territory, and on the Cape York Peninsula in Queensland. Bauxite also occurs in New South Wales, in Victoria, and on the island of Tasmania. The Australian deposits developed between the Eocene and Pliocene epochs on substrates ranging from Precambrian sandstones to Tertiary basalts. Gibbsite is the predominant aluminum mineral, although some boehmite occurs in all but the Western Australian deposits.

South America. On the outer slopes of the old Guyana Shield, many economically important deposits have been found. They are located in the Amazon Basin of Brazil, in Colombia, Venezuela, Surinam, Guyana, and French Guyana. Surinam and Guyana have been producing for more than 60 years, whereas the Brazilian Amazon deposits have been mined only recently. Bauxite also is produced in the Pocos de Caldas area in Southern Brazil, state of Minas Gerais. The South American bauxites generally are geologically young, gibbsitic ores.

Caribbean. Jamaica and the Dominican Republic have major reserves of karst bauxites that occur on Tertiary limestones under generally very thin overburden. Although gibbsite is the main aluminum mineral, some boehmite also is present. Jamaica has been one of the world's leading bauxite producers for the past 20 years.

North America. The only economically important deposit is located in Arkansas, where gibbsitic bauxite developed on nepheline syenite during the Eocene. Less than 5×10^7 t of bauxite remain, and the grade is declining.

Europe. Except for a few commercially insignificant occurrences, all European bauxites are of the karst type. The oldest deposits (Devonian/Mississippian) are those of the Tikhvin area in the former Soviet Union; the youngest are the Eocene bauxites of former Yugoslavia. Most European deposits developed during the Lower and Upper Cretaceous, e.g., the diaspore and boehmite bauxites of France, Greece, and Romania, and the gibbsite and boehmite bauxites found in Hungary, former Yugoslavia, and

Italy. Today, all European mines combined contribute only about 15 % of world production.

Asia. Major deposits of gibbsite bauxites occur in India; the island of Kalimantan (Indonesia) has large potential reserves. Numerous bauxite deposits, most of them containing geologically old, diaspore-rich ore, were found in China. Substantial deposits also are located in Western Siberia.

2.5. Economic Aspects

The proved reserves of bauxite shown in Table 4 are sufficient to supply the world aluminum industry for a few centuries. Total resources are estimated by the U.S. Geological Survey at 55 to 75×10^9 t. Because of the worldwide distribution of significant ore deposits, a disruption of bauxite supply for political reasons appears highly unlikely. In 1974 several major bauxite-producing nations formed the International Bauxite Association (IBA) with the intent of increasing control over the exploitation of their bauxite deposits. Although levies were increased substantially, competition from countries not associated with the IBA helped maintain a reasonable price structure.

Dissolution of gibbsite requires the mildest conditions in the Bayer process (see Section 3.1.2). Higher temperatures and alkali concentrations are necessary for the digestion of boehmite and diaspore. Technically, both oxide hydroxides can be processed without difficulty. The abundance of high-grade gibbsitic bauxites, however, has made boehmite- and diaspore-rich ores economically less attractive.

Economic and political considerations favor refining of bauxite near the deposit and shipment of either alumina or aluminum ingot. Brazil, Surinam, and Australia have smelters, although their refining capacity by far exceeds the demands of domestic metal production. Jamaica and Guinea refine a substantial portion of their own bauxite production.

About 25 % of all bauxite mined is used for producing refractories, abrasives, catalysts, adsorbents, and other industrial chemicals.

3. Bayer Process

3.1. History and Procedure

In 1855 the French mining engineer Louis le Chatelier obtained alumina from bauxite by sintering with sodium carbonate at 1200 °C and leaching the sodium aluminate with water. Aluminum hydroxide was then precipitated from the sodium aluminate solution by carbon dioxide. Early use of aluminum hydroxide was chiefly as a mordant in the textile dyeing industry. The demand for pure alumina increased rapidly when it became the raw material for aluminum production upon development of the Hall – Héroult cell. The Austrian chemist Karl Josef Bayer received German patent 43977 in August, 1888 for a new, improved process for production of aluminum hydroxide from bauxite, and the process became known as the Bayer process in his honor. Bayer initially worked on his process at the Tentelev chemical plant near St. Petersburg in Russia. Between 1888 and 1900 he supervised the construction of Bayer process plants in Germany, England, France, Italy, and the United States. The Merrimac Chemical Company in

Table 5. World alumina production, 10^3 t Source: Roskill's Metals Databook 1997

Country	Production	
	1996	1997
Australia	13 008	13 349
Brazil	2 147	2 800
Canada	1 064	1 100
China	2 080	n.a.
France	425	430
Germany	994	1 000
Greece	630	630
Guinea	630	622
Hungary	353	n.a.
India	1 650	1 660
Ireland	1 186	1 300
Italy	857	880
Jamaica	3 030	3 200
Japan	743	750
Spain	1 095	1 100
Surinam	1 579	1 600
Turkey	172	170
United Kingdom	110	110
United States	4 530	4 780
Venezuela	1 641	1 700
Yugoslavia	150	300
World total	43 600	45 000

Massachusetts first used the Bayer process in the United States, and the Alcoa plant in East St. Louis, Illinois, was constructed in 1901. Since then, plants have been built in more than 25 countries, and the present world alumina capacity is over 40×10^6 t/a (Table 5).

The important features exploited in the Bayer process are that boehmite, gibbsite, and diaspore can be dissolved in NaOH solutions under moderate hydrothermal conditions; the solubility of Al_2O_3 in NaOH is temperature dependent; most other components of the bauxite are quite inert in the process; and the silica that does dissolve subsequently forms a nearly insoluble compound. These features permit formation of a sodium aluminate solution, physical separation of the impurities, and precipitation of pure $Al(OH)_3$ from the cooled solution.

The main features of the Bayer process have remained unchanged for the last 100 years, although the scale of operations has been enlarged considerably due to chemical engineering developments. Figure 7 shows the flow sheet of the process as it is now practiced.

Each operation in the process is carried out in a variety of ways. The process begins with preparation of the bauxite by blending for uniform composition followed by grinding. In most plants the bauxite is ground while suspended in a portion of the process solution. This slurry is mixed with the balance of the heated NaOH solution, then treated in a digester vessel at well above atmospheric pressure. The digest reaction is:

$$Al(OH)_3 + Na^+ + OH^- \longrightarrow Na^+ + Al(OH)_4^-$$

Additional reactions convert impurities such as SiO_2, P_2O_5, and CO_2 to relatively insoluble compounds. The slurry leaves the digester at a temperature above its atmospheric boiling point and is cooled by flashing off steam as the pressure is reduced in several stages. The flashed steam is used to heat the slurry and the solution going to the digester. The bauxite residue solids are separated from the sodium aluminate solution in two steps so that the coarse fraction is processed separately from the fine. Both residue fractions are washed and discarded. The solution, being free of solids, is cooled and seeded with fine crystals of $Al(OH)_3$; this causes the $Al(OH)_4^-$ ions to decompose to $Al(OH)_3$, thereby reversing the reaction that previously had taken place in the digester. Again, the heat removed in cooling the solution is used to heat a colder stream in the process. After the precipitation reaction has proceeded to the point that about half of the Al_2O_3 in the solution has been removed, the mixture of solids and solution is sent to classifiers. The fine $Al(OH)_3$ particles are returned to the process to serve as seed. The coarse particles are washed and calcined to Al_2O_3. Excess solution introduced in washing the product and the residue must be removed by evaporation. In some cases the solution is treated to remove both organic and inorganic impurities before the solution is recycled through the plant.

3.1.1. Bauxite Preparation

The bauxite entering the refinery must be uniform and sufficiently fine that extraction of the Al_2O_3 and the other operations are successful. The chemical composition of bauxite varies. Uniformity is improved by blending material mined from several pits and, if necessary, by adding bauxite from storage piles. Ground bauxite cam be stpred as a slurry in surge tanks before it is pumped to digestion. These agitated tanks are operated so that plant feed is uniformly blended for several hours. Sometimes bauxites are dried to improve handling or washed to remove clay.

Hard bauxite is reduced to particles finer than 2 cm in roll or cone crushers and hammermills. Before it enters the process it is ground further to less than 0.15 cm.

Previously, fine grinding was done in dry mills operating in closed circuit with vibrating screens. Such operation required very dry bauxite to avoid blinding the screens. This resulted in a dusty working environment. In most modern plants, the bauxite is mixed with a portion of the process solution and is ground as a slurry. Rod mills and ball mills are used most frequently. The ground slurry may be passed over screens or through cyclones, with the fine particles progressing and the coarse ones being returned to the mills.

In all-wet grinding the bauxite feed and the flow of solution are controlled to keep the solids content of the slurry between 45 and 55 %. The power consumed by the mill drive is an indicator

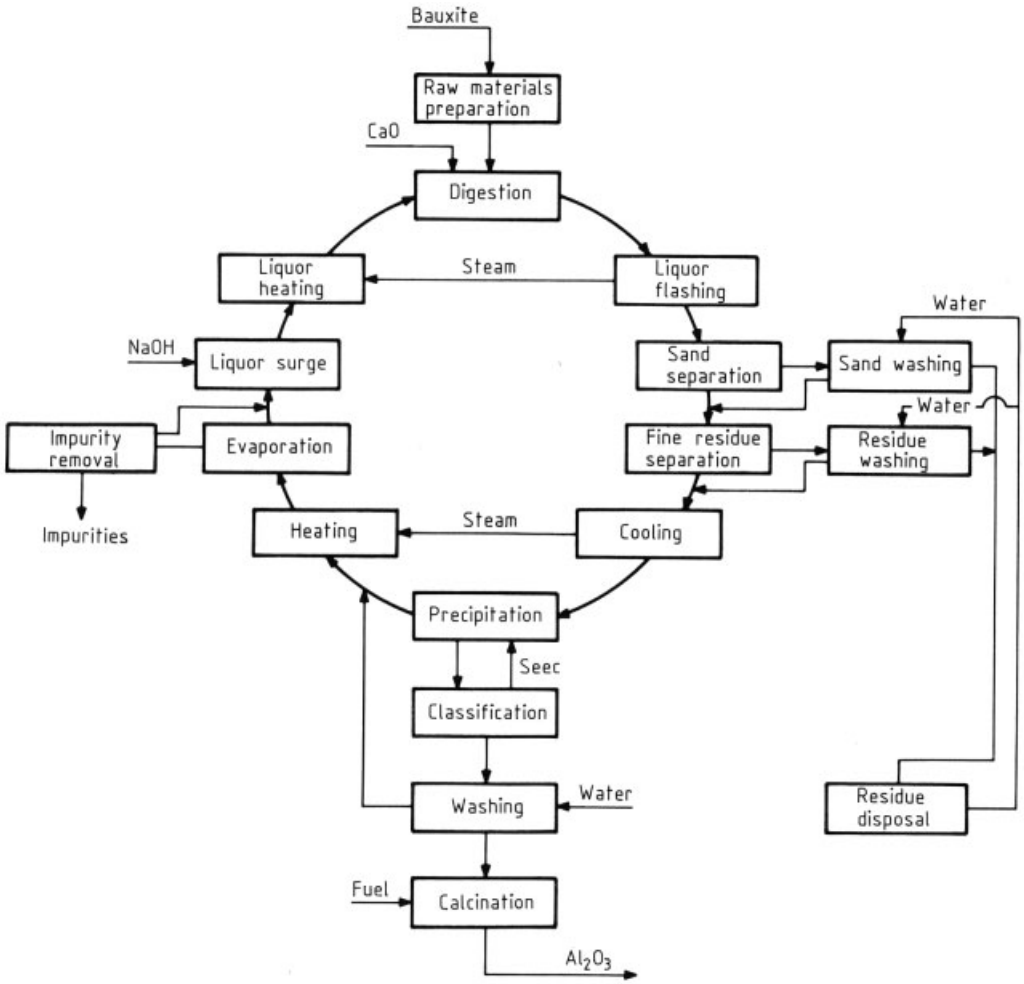

Figure 7. Bayer process flow sheet

of the amount of grinding being done and may be used to control the feed to the mill. On a longer time scale the particle size of the product can be used to make changes in the grinding operation.

3.1.2. Digestion

In digestion, all of the Al_2O_3 in the bauxite must be extracted. A solution is produced that contains the maximum Al_2O_3 concentration that can remain stable through the rest of the process. This must be accomplished while using a minimum amount of energy.

The *conditions for digestion* can and do vary widely. The first consideration is whether the available Al_2O_3 is present as gibbsite, boehmite, diaspore, or a mixture of these minerals. The dissolution rates of the three are quite different. Generally, if the bauxite contains mixed phases, the digestion conditions will be chosen on the basis of the least soluble compound. Any $Al(OH)_3$ or AlOOH left undissolved can act as seed in the clarification step, causing precipitation of $Al(OH)_3$ while the residue is still in contact with the solution. In the *sweetening process,* boehmitic bauxite is digested under relatively mild conditions, producing an intermediate Al_2O_3 concentration. In a separate vessel, gibbsitic bauxite is added to the flow from the first digest to raise the Al_2O_3 concentration to the desired level. Another approach is

to digest only the gibbsite in a mixed bauxite. The residue is separated from the solution and redigested under more severe conditions to recover the boehmite. This method reduces the flow to the high-temperature digest and so requires lower capital and operating costs.

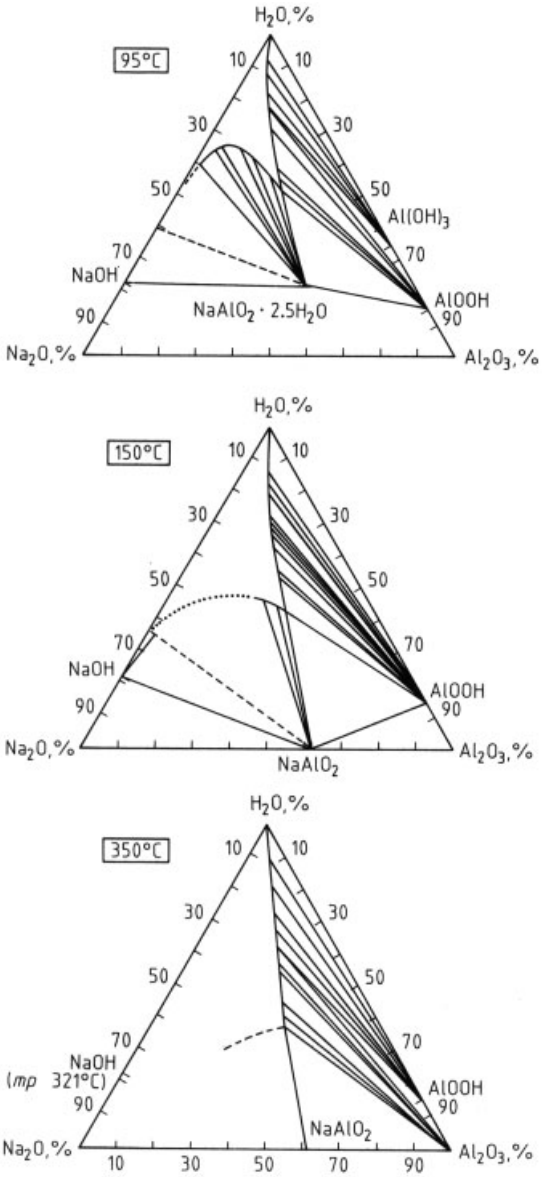

Figure 8. Phase diagram Na_2O–Al_2O_3–H_2O

The solubility data (Fig. 8) show that the Al_2O_3 concentration in the process solution can be increased by increasing the temperature, the NaOH concentration, or both [1–4]. As a result, operating conditions in plants vary widely. Higher digest temperatures result in higher pressures, making the equipment more expensive. There also is the need for more heat-exchange equipment, which further increases capital costs. High concentrations, on the other hand, permit increased production from a given flow rate and, hence, from a given plant installation. Precipitation is thought to occur better at lower concentrations, but use of low precipitation concentrations while digesting at high concentrations requires dilution and additional operating costs for subsequent evaporation. Choosing digester conditions involves balancing these physical factors with local economics, together with the designer's experience. This has resulted in the spectrum of operating conditions given in Table 6.

The conditions listed in the first line of Table 6 are those for digesting bauxite at the atmospheric boiling point. Quite high alkali concentrations are required and the evaporation requirement is an extraordinary 5.3 t of water for a tonne of Al_2O_3. Most plants digesting gibbsitic bauxite use the conditions on the second line of Table 6. Boehmitic bauxite is digested using one of two general sets of conditions. The first is European practice, in which higher concentrations and dilution prior to precipitation are preferred to higher digest temperatures. When American companies began processing boehmitic bauxite from the Caribbean area, most chose the same concentrations used for gibbsitic bauxite; therefore, a higher temperature was required.

The second important reaction in digestion is *desilication.* In the equation below, kaolinite is used to illustrate the reaction of siliceous minerals with the process solution:

$$Al_2O_3 \cdot 2\,SiO_2 \cdot 2\,H_2O + 6\,NaOH$$
$$\rightarrow 2\,NaAlO_2 + 2\,Na_2SiO_3 + 5\,H_2O$$

The soluble products react to form a series of precipitates with zeolite structure, having a composition of approximately $Na_8Al_6Si_6O_{24}(OH)_2$.

Depending on temperature and concentration, the ratio of $Al_2O_3 : \times SiO_2$ in the zeolite (tectosilicate) structure can vary. For each Al^{3+} replacing Si^{4+} in the lattice, one Na^+ is taken up to maintain charge neutrality. Anions, such as SO_4^{2-} or CO_3^{2-} may substitute for OH^- in the structure. The formation of the zeolitic desilica-

Table 6. Commercial digestion conditions

Bauxite type	Temperature, K	c_{NaOH}, g/L	Final $c_{Al_2O_3}$, g/L
gibbsitic	380	260	165
	415	105 – 145	90 – 130
boehmitic	470	150 – 250	120 – 160
	510	105 – 145	90 – 130
diaspore *	535	150 – 250	100 – 150

* CaO is added to digests to accelerate dissolution of diaspore.

tion product (DSP) therefore leads to costly loss of sodium hydroxide. This reaction, however, is necessary for lowering the level of dissolved SiO_2 to less than 0.6 g/L, the maximum acceptable concentration.

The rate-determining step in the desilication reaction is the nucleation and crystallization of the desilication product. Therefore, the presence of seed particles is important; without seeds, solutions containing 0.75 g/L SiO_2 will not react in 40 min at 415 K. This has led to the paradoxical situation that some bauxites contain too little reactive SiO_2 for good desilication. The slurry blending and storage operation discussed earlier can be an important part of the desilication process. If the storage temperature is above 355 K, about 80 % of the reaction takes place in 8 h. More important, seed is formed so the desilication reaction in the digest is not delayed. Very low SiO_2 concentrations can be achieved if an excess of CaO is charged. At high temperatures, and with the lime additions, a less soluble desilication compound (cancrinite) is formed. European plants use longer holding times during or after digestion to facilitate desilication.

Some CaO is added to the digest even when extreme desilication is not required. The calcium reacts with the carbonate and phosphate compounds as follows:

$$CaO + H_2O \rightarrow Ca(OH)_2$$

$$Ca(OH)_2 + Na_2CO_3 \rightarrow CaCO_3 + 2\,NaOH \text{ (causticization)}$$

$$5\,Ca(OH)_2 + 3\,Na_3PO_4 \rightarrow Ca_5(PO_4)_3OH + 9\,NaOH$$

The last reaction controls the level of phosphates in the process solution; this impurity can affect clarification adversely. Causticization in the digest was more important when prices were such that the sodium required by the process was supplied more cheaply as Na_2CO_3. Its current importance is in control of the CO_3^{2-} concentration, which, at high levels, can affect precipitation. In modern practice, the causticization reaction is carried out on dilute process solutions outside the digester. Maintaining low concentrations and having much of the NaOH combined as $NaAlO_2$ are favorable to more complete reaction of CO_3^{2-} with Ca^{2+}.

3.1.3. Equipment

The equipment for digestion includes the reactor vessel, heat-exchange equipment, and pumps. The first reactors were horizontal vessels with crude agitators mounted on axial shafts. These were filled with a slurry of bauxite, lime, and process solution. They were closed and heated individually by injecting high-pressure steam. At the end of the designated holding time, a discharge valve was opened and a slurry was forced into another vessel at atmospheric pressure. Only a small portion of the steam flashing from the slurry as the pressure was reduced could be recovered. Continuous operation was introduced in the 1930s in which heat could be exchanged between the hot stream leaving the digester and the cooler, incoming stream. This greatly reduces the energy requirements.

Reactors. Digester vessels, whether batch or continuous, provide intimate contact between the bauxite and the solution for a period long enough to complete the extraction and desilication processes. The equipment designed for this purpose is quite diverse. Agitated horizontal vessels have been replaced by vertical units to avoid the mechanical problems with sealing

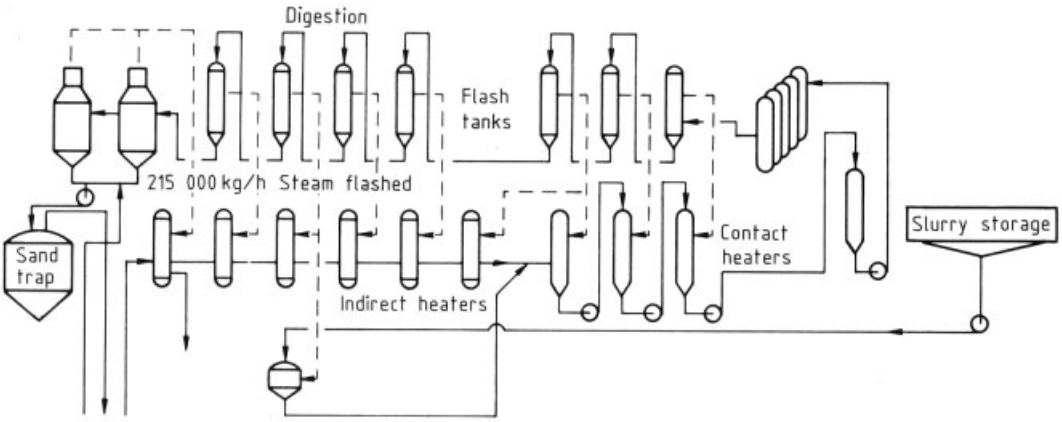

Figure 9. High-temperature digestion and heat recovery

the agitator shaft. Both designs have used a series of vessels to minimize short circuiting of the flow. European practice includes use of a series of vertical agitated vessels, each operating at a higher temperature than the preceding one. These units are heated by flashed steam in internal steam coils. A German digester design uses concentric pipes; the liquid being heated flows through the inner pipe while the hot slurry from the digester is returned through the outer pipe. This tube digester has been modified by placing several tubes for the cold slurry in parallel inside a larger outer tube. This reduces the length of the unit. The variety of digester design is illustrated in Figures 9, 10. The flow through single-digester units may be as high as 12 m^3/min.

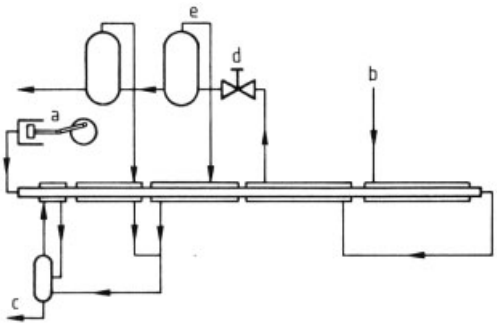

Figure 10. Tube digester
a) Piston pump; b) Heat; c) Condensate; d) Flow control valve; e) Steam

The usual *construction material* for digesters is mild (low-carbon) steel, despite data showing that, at the temperatures and NaOH concentrations involved, there may be stress corrosion cracking resulting from alkali embrittlement. The presence of the $[Al(OH)_4]^-$ ions reduces the activity of OH^- in the solution. Finely divided Fe_2O_3 in the bauxite quickly saturates the solution with Fe^{3+}, suppressing corrosion of the metal. Some digestion equipment has been nickel plated for safety.

Flashing. In most plants the hot slurry from the digester is flashed in a series of pressure vessels until it reaches the atmospheric boiling point. The steam generated in flashing is used in heat exchangers to heat the flow of liquor and bauxite coming to the digester. Flashing also removes water from the process stream, thereby accomplishing a significant portion of the evaporation needed. Normally, heaters are tubular, but coils of tubing also are used. As few as three stages of flashing may be needed when the digestion temperature is 418 K, and this may be increased to ten stages for a unit operated at 515 K. Figure 9 is a schematic diagram of a high-temperature digestion unit.

Heat Exchangers. Operating practice has been divided as to whether the bauxite slurry should be heated separately or whether it should be mixed with the rest of the solution and then heated. These two modes of operation have been designated the single- and the double-stream processes. Those favoring separate heating of the bauxite slurry, the double-stream option, feared rapid deposition of desilication product on the heat-transfer surfaces. Experience has shown that, particularly if the bauxite slurry is

stored for blending, those fears are unfounded. The rate of fouling is no higher than for heating clean liquor. In both options there is slow deposition of desilication product on the heat-exchanger surface because the desilication reaction does not reach equilibrium in the digester and continues at a slow rate wherever the solution temperature is raised. The encrustation must be removed periodically by washing with dilute HCl or H_2SO_4, or by mechanical means to maintain a high heat-transfer rate.

In some plants using the double-stream system, the bauxite slurry is heated in contact heaters. These heaters are designed much like barometric condensers. The liquid is forced to fall through the steam in the heater in droplets or in thin films. The large surface area exposed results in high heat-transfer rates, and because there is no metallic surface, there is no fouling. The penalty is higher dilution.

The heat needed to raise the temperature above that achieved in the flashing system can be introduced by direct injection of steam. More often the final heating to the digest temperature is done in a separate heat exchanger, with the energy coming from steam, hot oil, or molten salts.

Pumps. *Centrifugal pumps* are used in most digester systems. Those used for pumping slurries are built of wear-resistant alloys. A great deal of maintenance is required to keep the shaft seals operating well. In some plants, concern about pump wear has led to the use of *diaphragm pumps.* In such pumps a check valve admits the slurry to a chamber on one side of a flexible diaphragm. When that chamber is full, oil at high pressure is forced into a chamber on the other side of the diaphragm. This moves the slurry through another check valve into the digester. Therefore, no rapidly moving parts are in contact with the slurry. In a Japanese modification, called the *Hydrohoist,* the diaphragm is replaced by a sphere that is free to move the length of a cylinder. Slurry is kept in one end of the cylinder and the drive fluid, which is clean process solution, is in the other end. Because of this choice of drive fluid, leakage past the sphere is of no consequence and tight clearances are not required.

Measurement and Control. Operating variables that must be controlled are temperature, solution concentration, and degree of desilication. The holding time is fixed by the design of the digester vessels. As noncondensable gases are formed by the reaction of the solution with organic materials in the digest, it is necessary to vent these gases so that the digester remains full. Venting can be controlled by sensing the liquid level with a float or a radioactive sensor. Temperature measurement and control are relatively simple because only the temperature from the final heater is critical. Concentration control comes from blending the bauxite and careful proportioning of the slurry and clean solution flows. Originally, feedforward control was used in which the composition of the incoming materials was used to set the rates at which they were added. In modern practice feedback control is based on the composition of the exiting solution. Chemical analyses of the solution can be used, but the time lag involved makes this control imprecise. Nearly real-time control has been achieved by using the fact that the conductivity of the solution at a given temperature is a linear function of the mass ratio of Al_2O_3 and NaOH. This is the variable of interest. In modern plants the output of the sensor goes to a computer, which in turn controls the amount of bauxite slurry pumped. Such control has reduced the variance in concentrations to about a third of the values experienced earlier. This permits operation closer to the maximum safe concentration with less risk of premature precipitation of Al_2O_3 in clarification.

3.1.4. Residue Separation

The next step in the process is the separation of bauxite residue solids from the solution. A wide variety of equipment and procedures is used in this operation. The methods chosen depend on the quantity and properties of the residue.

The particle size distribution of the residue solids is usually bimodal. The coarse fraction over 100 µm in diameter is termed *sand*, whereas the rest of the solids are finer than 10 µm. The sand fraction may range from 5 to nearly 50 % of the total. The lowest amounts are in the Caribbean bauxites and the highest are in the residue from Western Australian bauxite. In the plant the sand fraction is separated from the process stream in liquid – solids cyclones or in more

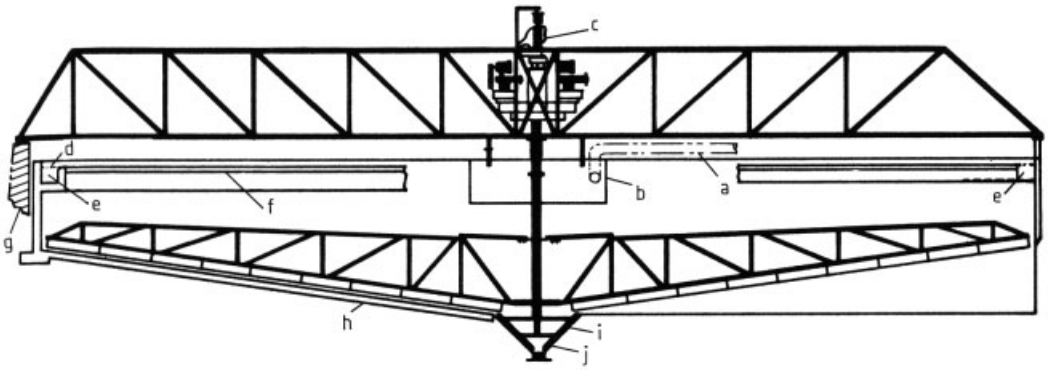

Figure 11. Residue thickener
a) Feedpipe; b) Feedwell; c) Motorized lifting device; d) Liquid level; e) Overflow launder; f) Weir; g) Off-tank support; h) Tank; i) Cone scraper; j) Discharge cone

primitive settling devices. A variety of equipment, including Dorr, Hardinge, and Aikens classifiers, have been used to wash the sand free of process solution before it is discarded. In all of these units the sand moves countercurrently to the wash water so that a maximum of washing is done with a minimum of water. The wash solution is added to the balance of the process flow and proceeds to further clarification steps for removal of the fine solids.

In most plants the fine residue fraction is settled in raking *thickeners* of the type illustrated in Figure 11. These vessels may exceed 49 m in diameter. Some older plants used multideck thickeners. Difficulties in keeping the multideck units in balanced operation have led to almost exclusive use of single-deck units in new construction.

The desanded slurry is fed at the center of the thickener and clarified solution overflows at the perimeter. As the solution flows radially across the thickener, both the horizontal and vertical velocities become very low. The solids, having a higher specific gravity than the solution, sink to the bottom of the thickener. A settling rate of 1.5 m/h is sufficient for commercial operation. The solution overflowing the thickener usually contains less than 0.3 g/L solids, whereas the underflow ranges from 15 to 35 wt % solids. The rotating mechanism has plows mounted at an angle to the arm. These slowly move the settled pulp across the bottom of the thickener to a discharge port. A few units have been designed with the discharge at the perimeter rather than at the center. The objective was to avoid the need to elevate the units to provide access to a center discharge.

The fine solids behave as a relatively stable colloidal suspension, so they settle slowly if not treated further. The addition of *flocculants* improves the clarity of the thickener overflow, the settling rate of the solids, and the solids content of the underflow pulp. Flocculants act by binding the very fine particles into flocs that may be several millimeters in diameter. The ratio of mass to drag forces is increased so the flocs settle more rapidly.

Starch from grains or roots was the first flocculant; the dosage varied from 0.5 to 3 kg per tonne of bauxite. The amount required increases as the surface of the residue increases and is affected by the mineral composition of the residue. Starch has the advantages of being inexpensive and ubiquitous. It does react with the NaOH in the solution to add organic compounds to the solution. In the 1950s water-dispersible polymers were proposed as flocculants. The first compounds were effective when added at 10 % of the starch charge, but they were expensive so they offered no economic advantage. More recently, flocculants based on acrylate – acrylamide copolymers have begun to replace starch. The functional groups of these anionic flocculants can be modified to suit specific operating conditions. Different materials may be used in the thickeners and in the washing operation. In some cases a small amount of starch is added to the synthetic flocculant to improve overflow clarity. These synthetic materials

now offer an economic advantage because they have been made less expensive and more effective.

With the residue concentrated in the thickener underflow, the next task is to *wash* it free of process solution so that it can be discarded. This is usually done in countercurrent decantation systems using washing thickeners that are similar in design to those used for settling. The washing operation is done in as many as seven stages with the solids moving countercurrent to the wash stream. This is illustrated in Figure 12. An equation can be written for each stage expressing a material balance between soluble materials and dilution. Computer models of the washing system allow accurate prediction of performance.

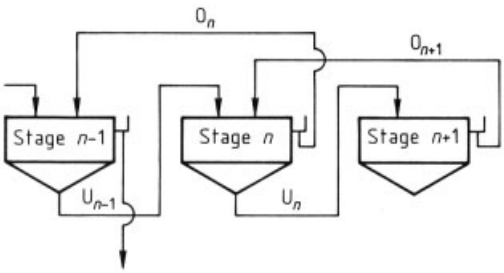

Figure 12. Flows at stage *n* in countercurrent decantation (O = overflow; U = underflow)

In *residue washing,* the objective is to minimize both the value of the soluble salts lost and the cost of evaporating the added wash water. The solubles in the discarded pulp can be reduced by increasing the solids content of the underflow. This reduces the amount of solubles carried from one stage to the next so that dilution is more effective and can be accomplished by using more flocculant or by increasing the holding time in the thickener. There is some risk in the latter change. If some undissolved gibbsite or boehmite remains in the residue, it may induce precipitation with a resulting loss of Al_2O3.

In some plants vacuum drum-type washing filters are used to replace some or all of the washing thickener stages. The solids content of the residue may be increased to 50 – 60 wt % by filtration. Therefore, only about 33 % as much solution is needed compared to plants using thickeners and this results in a great increase in the effectiveness of washing. Calculations for countercurrent washing show that two stages discharging at 55 % solids are more effective than six stages operating at 20 % solids.

Another approach only recently introduced into plant practice is called the deep thickener. This equipment was developed in the British coal industry to concentrate the slimes from coal washing. In the deep thickener, increased quantities of flocculant are used, and a column of residue pulp up to 10 m deep is maintained. The hydraulic pressure of the deep column appears to densify the pulp so that underflow solids approaching the concentration of filter cake are obtained. The overflow from these units contains more solids than the flow from standard thickeners, but the units are finding use in concentrating residue for washing and for disposal. The simplicity of the equipment adds to its appeal.

The final stage of clarification is *polish filtration* of the solution, sometimes called clear pressing. Most of the residue solids have been removed from the process solution in the previous steps, but product purity must be protected by removing the few particles of solids remaining. Filter presses are used most widely.

Another approach to polish filtration is use of a stationary filter in which the medium is uniform, fine sand a meter deep. Gravity provides the force to move the solution through the sand bed. During operation, the fine solids are mixed with the sand, from which they are elutriated at the end of the cycle. Such filters can be automated and are especially effective with residues that are difficult to process.

Close attention is required to keep the solids content of the filtrate down to 0.5 mg/L to protect product purity. Light transmission of the filtrate is used as an index of clarity. Purity originally was determined visually, but nephelometers are used now. If evidence of solids is found, the filtrate from each filter is scanned to locate the source. Some solids may pass through the filter cloth before a layer of solids forms to serve as a barrier, so it is common practice to recycle the first portion of filtrate.

Control in clarification has two major objectives; the solids must be removed and the washed residue must be prepared for discard, using a minimum of wash water. The first objective is achieved by measuring the cloudiness of the filtrate. High solids in the underflow are dependent in part on the retention time in the thickener, so measurements must be made of the pulp level in

the unit. Automatic determination of the interface between clarified liquid and pulp has been done by ultrasonic and by optical scanning. The density of the underflow pulp is monitored by a device that measures absorption of radiation or by gravimetric measurement. If the pulp is too dilute the rate of withdrawal is decreased. This control can be overridden if the pulp level in the unit rises too high so that there is concern that the solids in the filter feed may become excessive. The flows of solids and wash water also must be measured and proportioned for efficient operation.

3.1.5. Precipitation

The filtered solution has a temperature of 375 K and it must be cooled to 335 – 345 K before precipitation. This *cooling* is usually done in a flashing system analogous to that used in recovering heat in digestion. Because the temperatures are below the atmospheric boiling point, the flash vessels and the heat exchangers must operate under vacuum. Most of the heat removed by flashing is transferred to the colder solution returning from precipitation. Plate-type liquid – liquid heat exchangers have also been used. The solution can be cooled further midway through the precipitation operation. This allows higher recovery of Al_2O_3 without using conditions that lead to excessive nucleation of product.

Recovery of the $Al(OH)_3$ from the process solution is known as precipitation, decomposition, or Ausrührung. The reaction is the reverse of the digester reaction given earlier. The cooling done after digestion and filtration has moved the solution into an area of the solubility diagram known as the metastable region. The concentration and temperature are such that the solution is supersaturated with $Al(OH)_3$ but is not saturated enough for spontaneous crystallization. This is illustrated by Figure 13. At the digest temperature, T_1, the Al_2O_3 concentration is increased to point P on the diagram. Cooling to temperature T_2 results in crossing the solubility curve into the metastable region. Addition of $Al(OH)_3$ seed at this point causes precipitation and the concentration of Al_2O_3 approaches the solubility curve at T_2. The additional cooling recently adopted moves the solution conditions further to the left in the diagram above a lower point on the solubility curve.

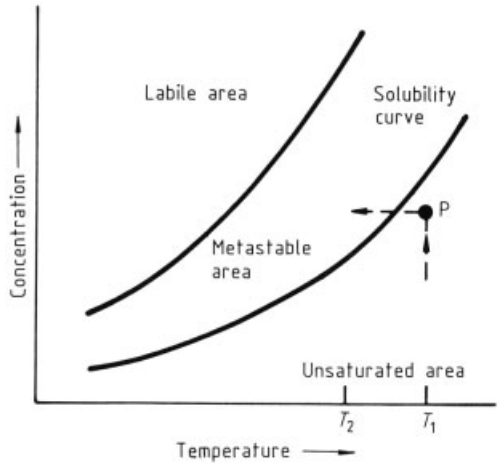

Figure 13. Generalized solubility curve

The *kinetics* of the precipitation reaction are represented by the equation:

$$\frac{-\mathrm{d}c}{\mathrm{d}t} = k\exp\left(-E/RT\right) A\left(c-c_\infty\right)^2$$

where:

c = Al_2O_3 concentration, g/L
k = constant
E = activation energy, about 59000 J/mol
R = 8.31441 J mol^{-1} K^{-1}
T = temperature, K
A = seed area, m^2/L
c_∞ = concentration at equilibrium

The reaction is second order, i.e., the rate is affected by the square of the difference between the actual Al_2O_3 concentration and the equilibrium concentration. The operating temperature affects the equilibrium concentration and the reaction rate. Because seed area is important, the seed charge must be as high as can be maintained while meeting other operating objectives.

The first objective in precipitation is to produce $Al(OH)_3$ that, when calcined, meets the specifications for metallurgical alumina. These specifications are discussed in Section 5.2. The second objective is to obtain a high yield from each volume of solution. This increases the plant capacity and reduces the amount of energy spent in heating and pumping the solution. At the same time, the processes serving to create new $Al(OH)_3$ particles must be controlled so that the

number of seed particles created equals the number of particles leaving the system as product. This requires balancing nucleation, agglomeration, growth, and particle breakage, so a combination of science and art has developed.

About 2×10^{10} nuclei must be formed in each liter of solution to balance the process. The rate of *nucleation* is a strong function of both the degree of supersaturation and of temperature. Very few nuclei are formed if the temperature is above 350 K. A change in the operating temperature of only 5 K will change a system from a net producer to a net consumer of nuclei. Temperature changes of 2 – 3 K are the most frequently used control measures.

The nuclei *grow* by accumulation of $Al(OH)_3$ on their surface if they become large enough to be viable seeds. The rate of growth can be as high as 9 µm/h, but it is generally much lower. The growth rate is independent of the particle diameter. Particles also increase in size by *agglomeration.* By a process that is not well understood, the nuclei form clusters in the first few hours of the cycle in batch precipitation. When the supersaturation is low, this agglomeration does not occur. Some speculate that when the rate of precipitation is high, the $Al(OH)_3$ is not well crystallized when it is first deposited on the seed. This later may serve as the bond for agglomeration. At low supersaturation the deposited material is better ordered. The action of mechanical equipment in precipitation can *break* fragments off crystals or break up agglomerates, creating secondary nuclei, but this is not a major factor in most plants.

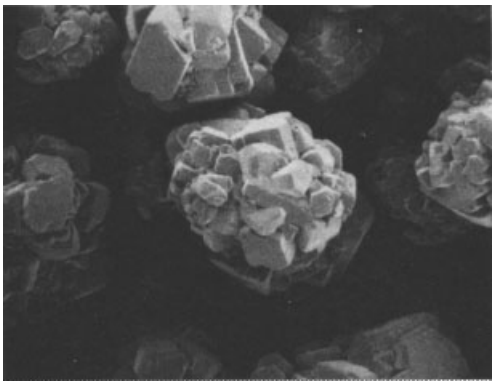

Figure 14. Technical aluminum hydroxide (gibbsite)

Figure 14 is an electron micrograph of some $Al(OH)_3$ produced in a plant operation. The particles are obviously the result of agglomeration. Those agglomerates having fewer particles can be shown by thin-section micrographs to have grown radially from an agglomerate of relatively few nuclei.

In Figure 15, the Al_2O_3 concentrations in experimental precipitations is shown as a function of time and temperature. The slopes of these curves indicate the rates at which precipitation is progressing. All curves show a high rate at the start and appear to be approaching a final value asymptotically. The final concentrations were well above the saturation level, but in these as in plant operation the rate became unusually low, so the process was terminated.

Precipitators are vertical cylinders; the height, which may be 30 m, is usually 2.5 – 3.0 times the diameter. The seeded slurry is circulated to maintain intimate contact between solids and solution. Early precipitators circulated the material by means of a central air lift pipe. Air introduced at the bottom of the pipe reduced the apparent density of the slurry within the pipe. This caused slurry to flow into the bottom of the pipe to establish circulation. Modern precipitators are circulated mechanically by impellers up to 3 m in diameter operating in a draft tube. In these units the flow is downward through the tube, so the tank bottoms are nearly flat to reverse the flow and cause upward motion of the slurry. Tanks with air lifts have conical bottoms to direct the flow to the central air lift.

Originally, precipitators were filled individually in *batch operation.* This method had several disadvantages, chief of which was the need for many operators because the operating cycle of a single precipitator required at least 15 separate operations. Control was difficult and batch operation left equipment out of service part of the time.

Nearly all plants built in the last 30 years have used the *continuous system.* Up to 13 tanks are placed in series so that the flow of seed and solution moves by gravity through channels connecting the tank tops. Continuous flows are much easier to measure and to automate. The number of flows is reduced, so control is better and the labor requirement is reduced. There are some operating problems. One is to avoid passage of the slurry across the tank tops, thereby reducing the retention time. Movement of solids from tank to tank is also a problem. An ideal system

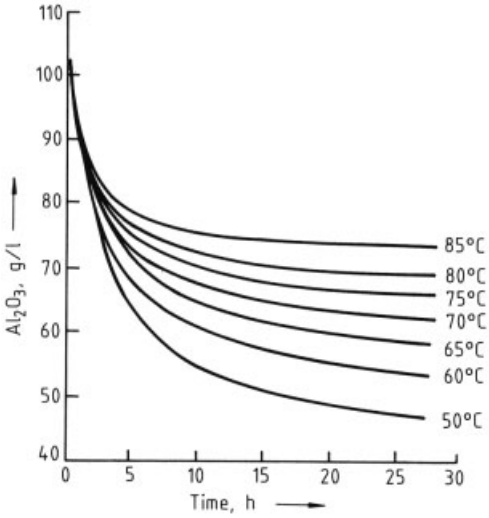

Figure 15. Effect of temperature on precipitation

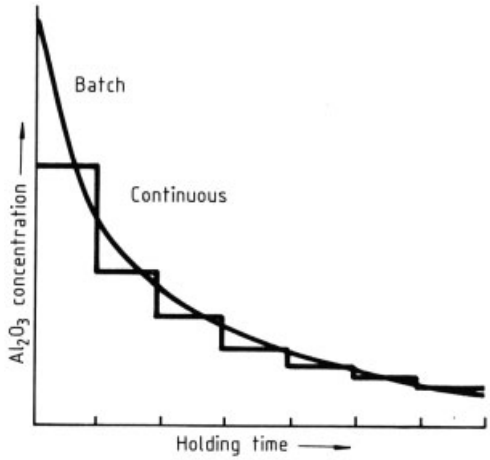

Figure 16. Al_2O_3 concentration profiles in batch and continuous precipitation

would retain fine particles in the tanks and move only the coarse material to the discharge. This is not possible, so a compromise method of operation must be found. Precipitation takes place at constant conditions in each tank instead of following a continuous curve, as in Figure 16. This tends to decrease production per tank, but the loss is more than recovered because the tanks are always in use.

Several patterns have been used to supply the *seed for precipitation.* The earliest was to neutralize a portion of the solution. An $Al(OH)_3$ gel was precipitated that changed into fine crystallites on aging. This very fine material was grown through at least one cycle to provide an active seed for product precipitation. The more common approach is to use fine particles, produced in the operation, as seed. The slurry leaving precipitation is classified into a coarse product fraction and one or more fine fractions. The fines are returned to the process to grow to product size. Classification is done by controlling the flow rate through tanks so that the fines are elutriated. Liquid – solid cyclones are coming into use because they provide more accurate separations. The seeding and classification systems have become increasingly complex. In modern plants at least three fractions are separated. In some plants the agglomerates formed from fine seed must pass through the system once more before becoming product. This reflects the belief that the agglomerates must be made tougher by deposition of additional $Al(OH)_3$ so that they do not disintegrate in calcination and subsequent handling.

In a modification [47], a relatively small charge of fine seed is added to the first tank in a series. This promotes agglomeration. Then, additional coarser seed is added to the second or third tank to provide the large seed area needed for high production. As indicated previously, cooling may be done between units in a series to increase the supersaturation in the final tanks, thereby increasing the production rate. A further change in modern plants is to filter the seed slurry being recycled. This decreases the amount of spent solution returning to the system and increases both the Al_2O_3 concentration in the first tank of the series and the overall production.

The control variables available to the precipitation operator are the temperatures in the system; the flow per unit, which translates into holding time; the amount of seed; and the particle size of the seed. Some of these variables are difficult to control, so the system usually cycles slowly from coarser than desired to finer. There is enough inventory in the process so that minor cycles can be tolerated without affecting plant output.

3.1.6. Impurities

The impurities in the product Al_2O_3 are affected by all of the foregoing operations, including selection of the bauxite. There are at least three classes of impurities. The first is residue solids that are not removed in clarification. The large surface area of $Al(OH)_3$ in precipitation effectively sweeps up these solids into the product. The preventive measure is increased vigilance in filter maintenance and operating. The second class is materials that are in true solution, such as sodium and gallium. The third group includes materials that are in solute form at the temperature of digestion but precipitate later in the process. Their solubilities may be only milligrams per liter and return of the solution to equilibrium at lower temperature may be slow. If precipitation is not completed before filtration, the remaining impurities appear in the product. There are also other impurities that do not fall neatly into these categories.

Sodium is the main component of the process solution and is also the largest contaminant of the product. The Na_2O content of alumina can be decreased by increasing the seed charge and by increasing the precipitation temperatures. The real variable that these operating changes help control is the rate of crystal growth. Higher temperatures, greater seed area, and lower Al_2O_3 concentrations all slow growth and give impurities more time to diffuse from the surface. By proper control of these variables an order of magnitude decrease in the Na_2O content of a commercial product has been achieved. Some sodium is lost in the Hall – Héroult cells during smelting by absorption into the cell lining and by conversion to sodium metal, so complete absence of sodium in metallurgical alumina is not necessary. The target is to balance the input to the cells with the losses. This requires Al_2O_2 containing about 0.4 – 0.5 wt % Na_2O.

Gallium is a ubiquitous component of aluminous ores. Its chemistry is similar to that of aluminum, so it accumulates in Bayer process solutions until an equilibrium is reached at about 0.2 g/L. The gallium content of Al_2O_3 is a linear function of the gallium concentration in the solution. Therefore, the amount leaving in the product equals the input. Because gallium is in true solution, the same changes that lower the Na_2O content of the product lower the gallium content. Only in the production of high-purity metal has there been concern about the gallium content of alumina. About 10 t/a of gallium is recovered from Bayer process solutions. This is the best source for the production of gallium (→ Gallium and Gallium Compounds).

Silicon is a component of many aluminum alloys, yet the specification for Al_2O_3 is less than 0.02 % SiO_2. This value is only 25 % of that specified 50 years ago. The SiO_2 is in solution, but it appears to be added to the product through a slow continuation of the desilication reaction (Section 3.1.2). There is also evidence that if the rate of Al_2O_3 deposition on seed is not high, the seed surface becomes at least partially covered with desilication product, thereby becoming less active as a seed for $Al(OH)_3$ precipitation and more active toward SiO_2. This impurity is controlled by driving the desilication reaction nearer to completion. Lowering the SiO_2 concentration of the solution below 0.1 g/L gives a product containing less than 0.003 % SiO_2.

Potassium is undesirable in the Al_2O_3 because it may destroy the graphite in Hall – Héroult cells by intercalation, i.e., it diffuses between the layers of the graphite structure, thus expanding its volume. Although it is soluble in Bayer solutions, there has not been a recorded instance of K_2O concentrations becoming high enough to affect the Al_2O_3 quality.

Iron(III)oxide (Fe_2O_3) as an impurity in Al_2O_3 has been the subject of much investigation. It is soluble up to ≈ 0.1 g/L in the solutions at digest temperatures (400 – 500 K), but only to ≈ 0.001 g/L at 333 K [48]. This reconciles the observations that Fe_2O_3 behaved at times as if it were in solution and as a colloid at others. There are several iron minerals that can serve as sources for the impurity. Hematite (α-Fe_2O_3) and goethite (α-FeO(OH)) are present in most bauxites. Goethite is slowly converted to hematite in high-temperature digests, particularly if CaO is present. This conversion improves clarification because hematite settles better than goethite. The mineral siderite ($FeCO_3$) present in some bauxites not only is a source of iron but also reacts with NaOH to increase the amount of causticization needed. Pyrite (FeS_2) is almost insoluble in Bayer solutions; its adverse effects occur when the sulfur is oxidized to sulfate in the presence of air and water. The sulfate will

react with NaOH to form Na_2SO_4 which has to be removed periodically.

Paradoxically, there is less difficulty with Fe_2O_3 as an impurity in processing bauxites containing a large amount of iron oxides. This is probably because the iron oxides serve as seed and hasten hydrolysis of the $NaFeO_2$ formed in the digest. Running the solution over iron oxide particles before filtration has been effective in controlling this impurity. Another process for lowering the iron content of Al_2O_3 is termed step precipitation. The filtered Bayer solution is lightly seeded with $Al(OH)_3$ for a short time to precipitate up to 3 g/L Al_2O_3. The seed and the small amount of precipitation effectively sweep the colloidal iron hydroxide particles from the solution, leaving an Fe_2O_3 concentration of less than 0.005 g/L.

Calcium also is a common impurity. The presence of increasing amounts of Na_2CO_3 in the Bayer solution seems to increase the CaO in the product. The calcium content of alumina, as with sodium, needs to be controlled such that the Hall – Héroult cell bath can come to an equilibrium level acceptable for cell performance. Calcium can be controlled by step precipitation or by lowering the carbonate content of the solution.

Lithium is also acceptable in Hall – Héroult cells. It can be removed by step precipitation.

Other metallic impurities that appear in many bauxites in small amounts are chromium, copper, manganese, titanium, vanadium, and zinc. Generally, these impurities are not so concentrated that removal processes are required. Sulfide salts have been added to precipitate copper and zinc. The chromium is a problem only if it is oxidized; the organic salts in the solution are usually a sufficient reductant. The addition of manganese compounds as oxidants for organic material has been reported.

Anions, such as carbonate, chloride, fluoride, and sulfate, are known to interfere with $Al(OH)_3$ precipitation. The carbonate ion is controlled by addition of lime, although processes have been reported in which Na_2CO_3 is removed by evaporative crystallization or by cooling the solution to 260 K. The sulfate and chloride ions usually are controlled naturally, as their salts are incorporated in the desilication product in sufficient quantities to maintain an equilibrium. There are commercial operations in which the sulfate salt is removed by evaporation and crystallization. The least soluble salt is schairerite ($Na_2SO_4 \cdot NaF$), and when the fluoride ion is depleted, burkeite ($2Na_2SO_4 \cdot Na_2CO_3$) appears. The fluoride ion is almost never present in quantities large enough to be harmful.

Many bauxites contain *phosphate minerals*. Apatite ($3Ca_2P_2O_8 \cdot CaF_2$) does not react in the process; indeed, this compound or its hydroxy counterpart is formed when soluble phosphates react with calcium in the digest. Aluminum phosphates, such as wavellite ($4AlPO_4 \cdot Al(OH)_3 \cdot 9H_2O$) do react, forming Na_3PO_4 as the soluble phosphate. The phosphate ion interferes with flocculation of bauxite residues containing goethite, probably by competing for attachment sites [49]. The concentration of phosphates never is high enough to interfere with Al_2O_3 purity. Addition of CaO to the digest is the control measure.

Organic matter in bauxite, whether it be roots and twigs from plants or humic acids from decayed matter, reacts in the digest to form a wide variety of organic compounds. Some operation process streams have exceeded 15 g/L organic carbon in solution. There is evidence that larger molecules are oxidized in the process all the way to Na_2CO_3 and sodium oxalate ($Na_2C_2O_4$). In addition some, but not all, of the organic compounds may interfere with precipitation [4].

The effect of sodium oxalate on processing has not been quantitatively determined, yet many plants incorporate equipment for crystallization of the oxalate salt from the solution. Others wash this salt from the $Al(OH)_3$ seed and recover it from the wash water. The evidence is overwhelming that operations are improved by the oxalate removal techniques. Early removal processes included evaporation to high concentrations followed by centrifugal separation of the gelatinous salt mass. In other plants, a portion of the solution was mixed with bauxite or $Al(OH)_3$ and heated to 1280 K. This destroyed the organic compounds and formed $NaAlO_2$, which could be recovered in the process. A German process adds magnesium salts to the digest to remove the deleterious organic compounds. Still others are investigating methods of oxidizing the organic material in the solution [50].

3.1.7. Calcination

The final operation in production of alumina is calcination. The temperature of the $Al(OH)_3$ is raised above 1380 K resulting in the reaction:

$$2\,Al(OH)_3 \rightarrow Al_2O_3 + 3\,H_2O$$

As discussed in Section 2.2, this reaction can take several pathways and several transition forms may appear. The end of all pathways is α-alumina. In older European practice, a major portion of the alumina was converted to the α-phase, sometimes by the addition of a fluoride salt to lower the temperature of transformation. In American practice, calcination always has been less severe; so, normally, $\leqq 20$ % of the alumina is in the α-phase.

Before calcination, it is necessary to *wash* the process solution from the coarse $Al(OH)_3$. This is done countercurrently, using storage tanks and filters. As in residue washing, increasing the number of washing stages and the concentration of solids leaving each stage improves the effectiveness of washing. Vacuum filters of several designs have been used for the final separation. The early Oliver and Dorr drum filters have been replaced by horizontal rotary filters because better washing can be achieved. The quality of the wash water is of some concern because such impurities as calcium and magnesium can be adsorbed on the surface of the $Al(OH)_3$.

Earlier, calcination was done in reverberatory furnaces. These had neither the capacity nor the thermal efficiency required, so they were soon replaced by *rotary kilns.* These kilns are cylinders that may be 3.5 m in diameter and over 80 m long. They are mounted on bearings and rotate about an axis inclined at a small angle to the horizontal. Damp $Al(OH)_3$ from the filters enters the upper end and slowly tumbles toward the lower end, traveling against a stream of hot gas formed by combustion of natural gas or oil at the discharge end. Early kilns were less than 40 m long, but longer units have better thermal efficiency and higher capacity.

Product coolers were added to kilns so that part of the heat in the incandescent product could be transferred to the combustion air. Early coolers followed designs developed in the cement industry and were themselves rotary units much like, although smaller than, the kilns. Improved coolers fluidize the product Al_2O_3 around heat-exchange surfaces, often in several stages. In such coolers, the heat can be recovered in combustion air or in wash water and the product can be cooled readily to 425 K.

The progressive adoption of large-capacity *stationary calciners* has reduced fuel consumption for alumina calcination to ca. 3.1 GJ per tonne of alumina, which compares to ca. 4.5 GJ per tonne for earlier rotary kilns. Several stationary calciner designs with capacities of > 2500 t/d alumina are now commercially available. In stationary calciners, a combination of fluid beds and stages in which the solids are transported in suspension is used. Figure 17 is a schematic diagram of a commercial unit capable of calcining 1500 t/d.

The $Al(OH)_3$ is washed and dewatered on the filter (a) and is conveyed into the flash dryer (b), where it encounters hot gas from cyclone (g). The gas evaporates the free water from the particle surfaces. Because the temperature exceeds 480 K, a portion of the water of hydration also is driven from the crystals. Cyclone (c) separates the solids from the gas. The gas, cooled almost to its dew point, leaves the unit through the electrostatic precipitator (j), where dust is recovered. Vessel (d) serves as a surge buffer to smooth the flow of solids through the unit. The solids are separated in cyclone (g) to enter the vessel (f) where combustion of fuel is taking place. Vessel (e) allows a holding time at the calcination temperature so that the water content of the product is low. A small amount of fuel can be burned in this vessel if a higher temperature is desired. In this design the flow of solids is counter to the flow of the gas.

The solids are cooled, also in counterflow, by the combustion air in unit (h). Final cooling is in a fluidized bed unit (i). In the fluidized unit a small amount of air is introduced through a membrane at the bottom. This lifts but does not entrain the solids, so they behave much like a fluid. Heat is transferred at a high rate to air or water flowing through tubes submerged in the bed. The product Al_2O_3 can be cooled below 400 K with all of the recovered energy being used in the process.

The great advantage of the stationary units is that the heat required for calcination is 3100 MJ/t. The capital costs are lower than for rotary kilns of equal capacity, and the mainte-

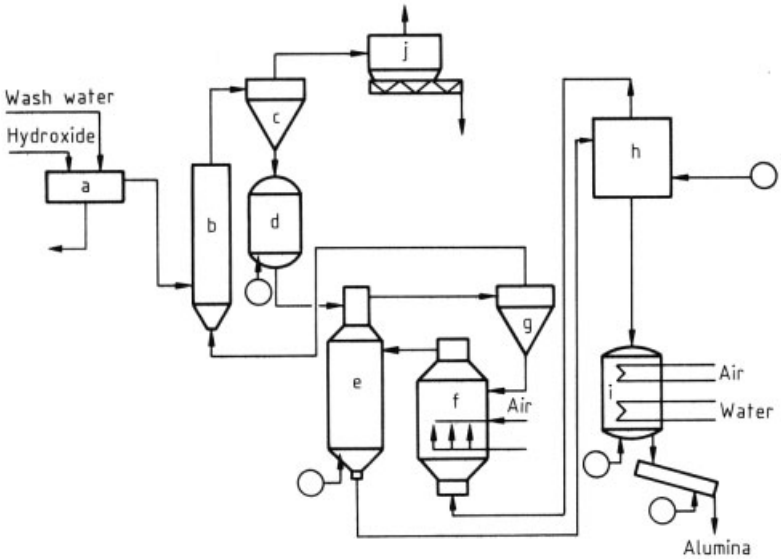

Figure 17. Alcoa flash calciner
a) Filter; b) Flash dryer; c) Cyclone; d) Fluidized-bed dryer; e) Holding vessel; f) Furnace; g) Cyclone; h) Multistage cyclone cooler; i) Fluidized-bed cooler; j) Electrostatic precipitator

nance costs are lower primarily because the refractories are not subjected to rotational stresses. Because the capacity per unit is high and automation is quite easy, the labor requirements are low. The stationary units are being installed in retrofitting older plants as well as in new construction.

Alumina calcined in stationary kilns has a lower α-alumina content and a higher surface area than alumina calcined at the same temperature in a rotary kiln. This reflects the shorter residence time and the absence of a high-temperature flame in the static units. Photomicrographs show that the rapid temperature rise creates vapor pressure within the particles that tends to fracture large crystals.

The characteristics of the metallurgical alumina are controlled mostly by the time and temperature of calcination. The important parameter is the reaction between the $Al(OH)_3$ feed and the flow of fuel. The fuel must be free of impurities that can contaminate the alumina. Sulfur and vanadium in fuel oil are limited by concern for product purity. The amount of combustion air must be sufficient to burn the fuel completely without the use of a great excess that increases the gas flow and thereby reduces the thermal efficiency. Factors affecting fuel economy and alumina quality in modern stationary calciners are reviewed in [51].

3.1.8. Evaporation

As is shown on the flow sheet for the Bayer process (Fig. 7), the solution continuously cycles through the plant. Consequently, any wash water used must be evaporated so that the solution volume can be controlled. About 10 % of the flow is evaporated in the two cooling areas by being flashed into steam. In high-temperature digesters, the amount of flash evaporation is even larger. At least one plant processing high-grade bauxite was designed without any additional evaporation capacity. In most cases, the economics require installation and operation of evaporators. The objective is to minimize the summed costs of evaporation and the value of the soluble salts lost by incomplete washing. The minimum is usually attained with a net dilution of the residue of 1.5 – 2.0 kg/kg.

Table 7 lists the input and exit streams that compromise the dilution balance in an operating plant. Efficiency demands that all of these flows be monitored.

Table 7. Dilution balance

Inputs	Losses
residue wash	evaporation
sand wash	–
$Al(OH)_3$ wash	heat interchange flash
free moisture in bauxite	$Al(OH)_3$ to calcination
water in gibbsite and boehmite	free water with $Al(OH)_3$
injected steam	water with residue
purge water	vapor from solution surfaces
sodium hydroxide	
cleanup water (maintenance)	
uncontrolled dilutions	

Several evaporator designs are used in Bayer plants, but nearly all designs are used in *multiple-effect configurations.* In such units each stage operates at a lower pressure than the preceding one. Therefore, the vapor evaporated from the solution in the first stage is at a temperature high enough to heat the solution in the second stage, causing it to boil. This continues to the final stage, where the vapor is condensed. The condenser operates under vacuum so a jet pump is necessary to remove noncondensable gases from it. The more times the latent heat is used, i.e., the more stages that are present, the more water can be evaporated by the fuel steam used in the first effect. With Bayer solutions the maximum number of effects is six. This is because the boiling point of the solution is elevated 5 – 8 K by the dissolved solids. This reduces the steam temperature in each effect; with many effects there is little temperature difference available to cause evaporation.

An unconventional design called *continuous regenerative evaporators* has been used in some plants. These units are similar to the heating and flashing equipment used in digestion in that the feed is heated through up to ten stages without evaporation; the hot solution is then flashed through an equal number of stages and the steam is used for heating the feed. These units are not as efficient as most multiple-effect evaporators. The design choice is based on economics. Special evaporator designs may be used when it is necessary to crystallize organic or inorganic salts from the solution. In such designs pumps are used to increase the thermally induced flow through the tubes in the heaters. The heat-transfer surfaces in evaporators must be cleaned periodically to remove encrustations of desilication product and soluble salts.

Evaporators are controlled primarily by changing the amount of steam used. The flow of feed solution is regulated so that crystallization of soluble salts is either induced or avoided, depending on the operating mode. Temperature readings at each effect are compared to design values to indicate operating difficulties.

3.1.9. Residue Disposal

The most important environmental problem in the Bayer process is disposal of the bauxite residue. The solution left with the residue after an economical amount of washing is still very alkaline and cannot be allowed to contaminate ground water. Furthermore, the desilication product in the residue which is in contact with water has the capacity to exchange sodium for hydrogen. An aqueous slurry of the residue that had been washed with 1000 times its mass of distilled water still reached a pH value of 10.5 on standing. The undrained fine residue, even after years of consolidation, does not have enough strength to support buildings or equipment. These properties make disposal a difficult problem.

Previously, disposal was to a marine environment where the alkalinity was diluted by large quantities of water. This method has been used in the seas off Europe and Japan and in a river in the United States. Studies by environmentalists have indicated little damage to flora or fauna by the residue in a disposal area [52]. Today, however, environmental concern is so great that any new refinery is unlikely to be permitted to use marine disposal.

Early inland refineries simply dammed a convenient valley or built retaining dikes on flat land to form residue disposal areas. In some cases, the sandy portion of the residue was used to build the dikes. This method can be effective and cheap if care is taken to protect the surroundings by proper sealing techniques. The compacted residue has a lower permeability than many clays; yet, there have been isolated leaks into aquifers from such impoundments. Improved designs have been developed.

In Germany, retaining dikes are built with a portion of the structure designed to be porous. The dilute solution draining from the residue is channeled by the porous sections into water-

treatment facilities before being discarded. In some cases, the dike material exchanged ions with the solution so that little treatment was needed. The draining allowed consolidation of the residue so that it could support equipment. An American firm used a drained lake in which the bottom of the disposal area was covered with sand in which draining pipes were laid. The residue was pumped to these areas in the traditional manner as a dilute slurry. Most of the conveying water was decanted, while some percolated through the deposit to the drains. This design has two advantages: the residue consolidates better, so the storage capacity of an area is increased, and the hydraulic head on the bottom is reduced to zero, so that leakage is unlikely. Lakes have been built using the sand drains on top of seals of clay and plastic film for additional environmental safety.

Another modern disposal method, called *dry stacking*, takes advantage of the thixotropic nature of the residue. By some method, usually vacuum filtration, the residue is concentrated to 35 – 50 wt % solids. The slurry is agitated to reduce its viscosity by as much as two orders of magnitude and pumped to a disposal area. There it flows in lava-like fashion over the surface, establishing a slope away from the discharge point. In the absence of shear, the viscosity of the slurry increases and flow stops. Water does not separate from the slurry and the slope causes rain to run off rapidly, so the surface usually is losing water to the atmosphere. The surface becomes deeply fissured, further assisting drying. In about 90 days the residue may dry to 75 wt % solids, far drier than in any of the other disposal methods. In this state, it can support heavy earth-moving equipment and can be recovered or used for increasing the height of retaining dikes. This method maximizes the storage capacity of a given area and seems to pose the minimum threat to the environment. It does require construction of a permanent lake area for water storage and for cooling water. If any use is to be made of the residue, recovery is relatively easy.

Many investigations have been directed toward finding a commercial use for bauxite residue. The high iron content of some residues suggested production of pig iron. The quality of the iron was poor and the amount of slag formed exceeded the original amount of residue. Similarly, chemical processes have been developed for recovery of Al_2O_3, Na_2O, and TiO_2 from residue. Although all are technically possible, none has been feasible economically. Small quantities of residue have been used in making Portland cement, and smaller quantities have been used as a mold wash, as an insulating material, and, after reaction with H_2SO_4, as a water treatment. No chemical use is likely to consume a significant part of the residue [1], [5].

The residue is claylike and can be used in ceramic materials. The sodium causes formation of glasses at 1450 K, giving a vitreous bond. Bricks have been made commercially, but economic factors and other shortcomings of the brick have eliminated this use. Sintering the residue into aggregate for concrete may be economical where natural aggregate is not available.

In Texas, agronomists have shown that the residue can be used to neutralize acidic soil, replacing limestone. The cost of preparing the material for application and the logistics argue against extensive use. The conclusion is that large-scale use of bauxite residue is unlikely, so efforts should be directed toward reclamation of disposal areas.

3.1.10. Energy in the Process

Approximately 16 MJ are required to produce a kilogram of Al_2O_3. The worldwide range is 7.4 – 32.6 MJ/kg [48]. Variations in the quality of bauxite, plant design, and the size of the plants are the reasons for the large differences. Even with existing plants, large improvements in energy usage can be made. The Aluminum Association (USA) reports that the energy used per unit of alumina production in 1980 was 68 % of the energy used in 1972 [54]. The energy used in mining is less than 5 % of the total and is difficult to change. The energy for transporting bauxite is double that for mining it, because the average tonne of bauxite used in the United States is transported 4700 km. The situation worldwide is different because much of the bauxite is refined close to the mines.

Within the refinery, more than half of the energy is used for pumping and heating the solution and for evaporation. The previous discussion (Section 3.1.3) has shown that plant design can affect heat recovery and minimize en-

ergy use. Emphasis is being placed on increasing the amount of Al_2O_3 recovered from less than 50 g/L to over 65 g/L; some claim yields as high as 100 g/L. Because energy is more closely related to the flow of solution than to yield, the energy savings are large. The static calciners are nearly as efficient as they can be made because the temperatures of the gas and of the product leaving the units are very low. The change from rotary kilns has made up a large portion of the savings. There are still opportunities to improve operating practice and design to reduce energy consumption [55].

3.2. Economic Aspects

The largest cost elements in alumina production are raw materials, energy, and capital-related costs for the production equipment. Labor, operating supplies, and miscellaneous costs are much smaller than these three.

Bauxite is the most important raw material. Its cost includes those for mining, transport, levies, and taxes. In addition, the relative quality of the bauxite ore influences the expenditures for necessary reagents. Mining is relatively inexpensive because most deposits are covered with only a shallow overburden and can be mined with efficient equipment. Since the formation of the International Bauxite Association (IBA), levies and taxes have become a large part of the cost of bauxite. These costs have become relatively stable because most are related to the selling price of aluminum. The availability of bauxite from outside the IBA countries has allowed the effect of supply and demand to influence this cost. Where levies are not imposed, some system of taxation provides income to the producing nation. Transportation costs vary from very small to as much as half the delivered cost of the bauxite. Some bauxite travels less than 5 km by truck or conveyor belt to the refinery. Other bauxite may travel more than halfway around the world. The characteristics of bauxite most important in influencing its value are the available alumina and the reactive silica contents. The latter quantity has a great effect on the amount of sodium hydroxide and lime required in processing. Bonuses are given for high available alumina values, and penalties are charged for excessive silica.

In page 28, values were given showing the fourfold range in energy used to produce alumina. The most important variables are the plant design and the quality of the bauxite being processed. The most energy-efficient bauxites are gibbsitic, with very low residue content.

The capital costs also show wide variation. In 1990 U.S. dollars new installations may cost from $ 700 to $ 900 per annual tonne of capacity. The variables are design factors, location, capacity, and the properties of the bauxite to be processed. Much of the world capacity was constructed before construction costs were inflated and so has lower capital costs. The effect of capacity is also large, and plants producing 2×10^6 t/a have costs well below those of plants producing less than 10^5 t/a.

Alumina is traded on the world market as a commodity. Not all smelters have alumina capacity dedicated to them; they buy alumina on contract. Long-term contracts are common. In other instances, alumina is bought when needed at spot market prices. Since 1980, alumina has sold on the spot market for less than $ 150 to over $280 per tonne, reflecting a change in the relationship between supply and demand. Economics do not always control production because many corporations own refineries to supply their smelters. In other instances, national governments own a significant portion of a refinery and for political reasons may choose to operate in a noncompetitive situation. Costs per tonne of alumina can, under unfavorable conditions, exceed $ 200, although efficient plants may produce at roughly half this figure.

4. Other Processes for Alumina Production

4.1. Raw Materials

Many investigations have sought processes to replace the Bayer process by using raw materials other than bauxite [2]. Clay, primarily kaolinite, has been most considered because it can contain up to 39 % Al_2O_3 (→ Clays). Other materials, including anorthosite, nepheline, coal wastes, and fly ash, have been candidate raw materials. Yet, less than 2 % of the world supply of alumina is not made by the Bayer process. This happens

only where use of a domestic raw material and production of a desirable byproduct change the economic picture. Alumina from ores other than bauxite normally costs 1.5 – 2.5 times that from the Bayer process.

Because aluminum is amphoteric, both acid and alkaline processes have been developed. Most of the processes in each class follow the same general flow sheets.

4.2. Alkaline Processes

Sodium compounds, such as Na_2CO_3, react at 1280 K with the Al_2O_3 in aluminous ores to produce water-soluble $NaAlO_2$. This compound can be leached from the sinter with water and the solution treated to remove impurities; the purified solution is neutralized with CO_2 to recover $Al(OH)_3$. The last step regenerates the Na_2CO_3 for recycle. Difficulty arises because the aluminous ores contain SiO_2, which also reacts in the sinter to form soluble Na_2SiO_3. In processing, the desilication reaction discussed in Section 3.1.2 takes place, so the net recovery of Al_2O_3 is small or zero. If 2 mol of CaO, usually as limestone, is charged for every mole of SiO_2 in the raw material, insoluble calcium silicate compounds are formed. Under proper leaching conditions the $NaAlO_2$ can be recovered from such sinters with only slight loss.

The process has been used commercially in two American plants to recover most of the Na_2O and Al_2O_3 lost in the desilication product formed while treating high-silica native bauxite. A schematic flow sheet is given in Figure 18. The same process was investigated by the U.S. Bureau of Mines in a large pilot plant using anorthosite as the raw feed [51]. [Anorthosite is a mixture of the minerals anorthite ($CaAl_2SiO_8$) and albite ($NaAlSi_3O_8$).] Unless the composition, sintering, and leaching were closely controlled, the residue formed a gel in leaching and became very difficult to filter.

In a variation of the sintering approach, *calcium* replaces the sodium so that calcium aluminate ($3CaO \cdot Al_2O_3$) is formed as well as calcium silicate ($CaSiO_3$). The calcium aluminate reacts with Na_2CO_3 in an aqueous leach to form $NaAlO_2$.

$$3\,CaO \cdot Al_2O_3 + 3\,Na_2CO_3 + 2\,H_2O \rightarrow 2\,NaAlO_2 + 4\,NaOH + 3\,CaCO_3$$

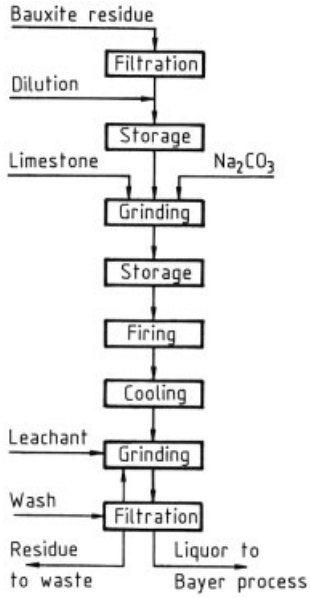

Figure 18. Lime/soda sinter process flow sheet

In carbonation, $Al(OH)_3$ is formed and the Na_2CO_3 solution is regenerated. Nepheline, $(Na,K)AlSiO_4$, is treated by this process in the former Soviet Union. The calcium silicate residue is processed to make about 10 t of cement per tonne of alumina. The two products make the process viable here.

In the Pedersen process, sintering was replaced with a reducing fusion so that ferric ions in the ore were reduced to metal. Iron and the calcium slag were separated by decantation. During cooling, the $CaSiO_3$ passes through a crystalline phase change and the resulting stresses reduce the slag to powder [1], [2].

The sinter processes suffer economically for several reasons. The energy for sintering and for evaporation of the leach solution must be added to the energy required from operations analogous to those in Bayer processing. The capital investment is increased by the need for sintering equipment, and although limestone is inexpensive, the quantity required is so large that the expense is considerable.

4.3. Acid Processes

All of the acid processes follow the general flow sheet given in Figure 19. Usually the clay is prepared by grinding and roasting to about 1000 K. The roasting changes the kaolinite to meta-kaolin, from which the aluminum can be dissolved as the acid salt. The roasted clay is leached in an acid solution, usually at the atmospheric boiling point. Some investigators have chosen to leach at elevated temperature for processing advantages even though corrosion problems become more severe [1]. The siliceous residue is separated using sedimentation and filtration as in the Bayer process. Small amounts of iron salts remain in the clarified solution. These salts are removed by extraction with an organic compound that forms a complex with the iron but not with the aluminum salts. The organic solution is decanted from the aqueous phase and is treated to separate the iron salt and to regenerate the organic extractant.

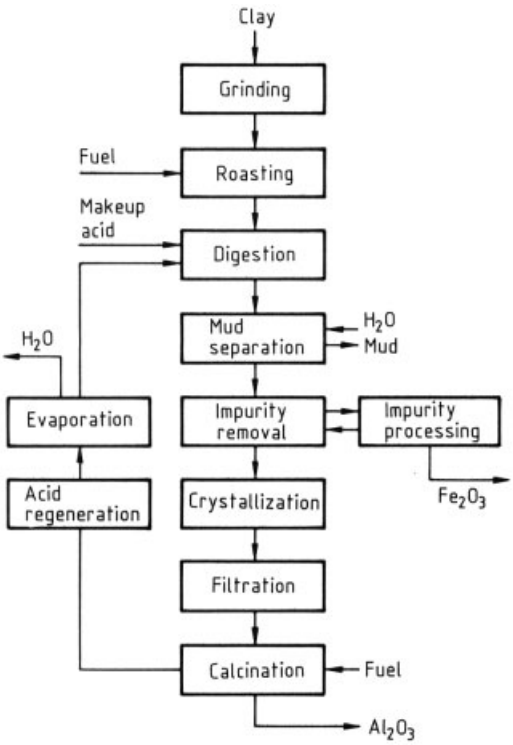

Figure 19. Generalized acid process flow sheet

A hydrated aluminum salt is recovered from the aqueous solution, usually by evaporation and cooling to cause crystallization. In the *hydrogen chloride process,* advantage is taken of the low solubility of $AlCl_3$ in HCl. The HCl gas is absorbed in the solution both to regenerate the acid for leaching and to precipitate $AlCl_3 \cdot 6H_2O$. In all acid processes the hydrated salt is decomposed to Al_2O_3 by heating to 1300 K. Both water and the acid radical are driven off, and the acid is absorbed for recycling.

The U.S. Bureau of Mines sponsored a cooperative effort with several aluminum companies to investigate acid processes for recovery of alumina from clay [56]. For several reasons the HCl process was preferred: the reagent is inexpensive, processing conditions are not severe, and the salt can be recovered without evaporation. The HCl is not decomposed in calcination, and the water content of the acid salt is half that of other acid salts. Despite these features the process is not competitive economically with the Bayer process. A major fault is the amount of energy used to roast the clay, to evaporate the solution, and to decompose the acid salt. The investment is increased because of the corrosive solutions. A process disadvantage is that the Al_2O_3 is physically different from Bayer alumina, so smelting practice has to be changed for this aluminum source.

Pechiney and Alcan jointly operated a pilot plant using the H^+ *process,* a modification of the HCl process in which H_2SO_4 is added to the digest [52]. The combination of acids eliminated the need to roast the clay. Other processes use H_2SO_4, H_2SO_3, HNO_3, or NH_4HSO_4 as reagents to attack the clay, but no commercial use has been made of any of them.

Clay is a major impurity of coal, so the ash from coal may be considered an overroasted clay. The acid processes, with minor modifications, will extract alumina from ash [57]. Anorthosite that is high in anorthite reacts with boiling HCl; this approach has been investigated in Norway and Canada [58]. So far, all of these methods have served only to enrich the technical literature.

Two aluminous minerals that do not contain silica have attracted attention and, because of their composition, require different technology. Alunite, $K_2SO_4 \cdot Al_2(SO_4)_3 \cdot 2Al_2O_3 \cdot 6H_2O$ [*12588-67-9*], has been used commercially in

Russia. Both H_2SO_4 and K_2SO_4 are useful byproducts of alumina production by this method. The alunite is heated to drive off the hydrated water; additional heating under reducing conditions decomposes the aluminum sulfate without adversely affecting alumina recovery. The solid residue from the roast is a mixture of K_2SO_4, Al_2O_3, and gangue. The K_2SO_4 is dissolved in water and recovered. Alumina is extracted from the gangue in a modification of the Bayer process. In other approaches using alunite, the solutions are kept acidic so the alumina is recovered as a salt [59].

Dawsonite, $Na_2O \cdot Al_2O_3 \cdot 2\,CO_2 \cdot 2\,H_2O$ [*12011-76-6*], is found in some oil shales. The temperatures used to recover the oil from the shale decompose the dawsonite to $NaAlO_2$, which is soluble in dilute NaOH. The $Al(OH)_3$ and Na_2CO_3 can be recovered from the solution by carbonation. The economics are such that recovery of the oil must be competitive before mining and byproduct recovery can be considered.

5. Metallurgical Alumina

Aluminum production is the principal application for alumina; more than 92 % of world alumina production is used for this purpose. The property requirements for commercial metallurgical alumina are therefore of considerable importance to the alumina industry. Specifications have responded to changes in energy costs, aluminum cell design, cell gas scrubbing techniques, environmental regulations, working conditions in smelters, and the technology for alumina calcination. In general, there has been a shift away from relatively small-particle-size, highly calcined, "floury" alumina to a coarse, free flowing, dust-free, less calcined, "sandy" alumina of narrower particle sizing and higher chemical purity (→ Aluminum).

5.1. Alumina Properties Required for Smelting

Five developments have had a significant impact on alumina properties.

Cell Design. New, high current efficiency, prebaked anode cells with automatic center-feed systems require that the alumina be consistent, with trouble-free handling properties to insure proper conveying and volumetric metering from the feeder. These criteria are best met by a free-flowing, moderate- to low-calcined alumina with relatively coarse and narrow particle size distribution.

Cell Gas Dry Scrubbing. The use of dry scrubber systems employing cell-feed alumina as adsorbent for fluorides in effluent gas from the cells dictates other requirements for metallurgical alumina:

1) High adsorption capacity for hydrogen fluoride. This property is closely related to the specific surface area of the alumina, which is higher for the lower calcined aluminas.
2) Attrition resistance.
3) Free flowability.
4) Higher chemical purity to compensate for capture in the dry scrubber of impurities which are recycled to the cell.

Pot Room Working Conditions. Use of low-calcined, high surface area alumina as a cover for the cell bath reduces fluoride evolution within the pot room. Working conditions in the smelter are degraded by dust caused by fine particles of alumina. Reduction of the fines fraction ($< 44\,\mu m$) in the alumina and high attrition resistance are important for reducing dust.

Use of Stationary Calciners and Pneumatic Handling Systems for Alumina. The replacement of rotary alumina kilns by the energy-efficient stationary calciners has resulted in producing a different type of calcined alumina for smelter feed. This difference is reflected in the interrelationship between degree of calcination (measured by the mass loss on ignition), specific surface area, and the α-Al_2O_3 content [60].

The strength of the alumina particles has become of concern not only because of relatively higher breakdown in the stationary calciners, but also because of attrition occurring in pneumatic unloading and conveying equipment and in fluid-bed dry-scrubbing systems. The generation of fine particles in such equipment, apart from causing unacceptable dusting conditions in

the smelters, often results in troublesome segregation problems in alumina storage bins and bunkers.

Electrolyte Composition and Temperature. The trend to operate cells at lower temperatures with electrolyte having a lower bath ratio ($NaF : AlF_3$) has decreased the solubility and rate of dissolution of alumina in the electrolyte. The rate of dissolution is greater for aluminas having higher surface areas and low content of α-Al_2O_3. Both properties can be achieved by a low degree of calcination.

5.2. Typical Specifications for Metallurgical Alumina

The considerations discussed in Section 5.1 have contributed to the evolution of the general specifications used in production and international trading (Table 8). These values are only representative and considerable variation exists in actual practice depending on price, availability, smelting practices, and many other factors. Consistency of quality is a major consideration.

Table 8. Typical properties of metallurgical alumina

Physical property	
Particle size distribution, wt %	
+100 mesh (Tyler)	< 5
+325 (44 µm)	> 92
− 325	< 8
Bulk density, kg/L	
loose	0.95 – 1.00
packed	1.05 – 1.10
Specific surface area, m^2/g	50 – 80
Moisture (to 573 K), wt %	< 1.0
Loss on ignition (573 – 1473 K), wt %	< 1.0
Attrition index (modified Forsythe – Hertwig method)	increase in < 44 µm particles 4 – 15 wt %
α-Al_2O_3 content (by optical or X-ray method), %	< 20
Chemical analysis	wt %
Fe_2O_3	< 0.015
SiO_2	< 0.015
TiO_2	< 0.004
CaO	< 0.040
Na_2O	< 0.400

6. Industrial Alumina Chemicals

Alumina, in various forms, is one of the inorganic chemicals produced in greatest volume today. Although production of aluminum metal consumes ca. 90 % of all alumina, an increasing amount is being applied in the chemical industry for fillers, adsorbents, catalysts, ceramics, abrasives, and refractories. With the development and growth of applications and markets for alumina chemicals, all the major alumina producers have, over the years, converted a part of their capacity to produce various alumina chemicals. In fact, some of the older, smaller alumina refining plants have been totally converted to alumina chemicals production in order to be economically viable. Chemical uses account for nearly 8 % of the world production.

6.1. Aluminum Hydroxides

Aluminum hydroxides constitute a versatile group of industrial chemicals. Important uses requiring large quantities are as fillers in plastic and polymer systems and for the production of aluminum chemicals. A moderate amount is used for the production of alumina-based adsorbents and catalysts.

Aluminum hydroxide meets most of the requirements for an effective filler: white or near-white color; large volume production base, resulting in price and supply stability; consistency of physical and chemical properties; a wide range of particle size distributions; chemical inertness; and nontoxicity. However, its increasing popularity as a plastics filler is strongly related to its fire-retardant and smoke-suppressant properties, which justify the somewhat higher price compared with calcium carbonate and other mineral fillers. Aluminum hydroxide acts as a fire retardant by adsorbing heat through endothermic dehydration and dilution of pyrolytically produced combustion gas by the released steam. Dehydration and water release processes become significant at temperatures above 500 K. The smoke-inhibiting activity of aluminum hydroxide filler has been attributed to promotion of solid-phase charring in place of soot formation. Although offering these desirable features,

aluminum hydroxide has certain disadvantages that impose some limitations on its uses as a filler. Like other nonreinforcing mineral fillers, it generally lowers strength. Because it undergoes thermal decomposition, it is not suitable for processing above 500 K. These factors are responsible for the larger use of aluminum hydroxide in latex carpet backings and in glass-reinforced polyesters, where processing temperatures below about 480 K generally are used. These are also chemically cross-linked or fiber-reinforced systems, in which loss of strength caused by a nonreinforcing filler may not be very important. Aluminum hydroxide filler has been used less extensively in thermoplastics, e.g., poly(vinyl chloride) and polyethylene, and elastomeric materials. For use as a filler, the crystalline aluminum hydroxide from the Bayer process is dried and ground to particles $\leqq 10\,\mu m$ size. Special grades with increased whiteness and a variety of both particle size ranges and chemical purity also are available commercially.

Fine, precipitated aluminum hydroxide having a uniform particle size ($\approx 1\,\mu m$) is used in paper making as a filler pigment and as a coating. As a filler, it disperses rapidly with low sedimentation. Improved printing properties are reported for the hydroxide-filled paper. The application of fine, platy aluminum hydroxide as a paper coating is well established in the paper industry. It gives a coating of high brightness, opacity, and gloss (→ Paper and Pulp).

Technical aluminum hydroxide obtained from the Bayer process is 99.5 % pure. It dissolves readily in strong acids and bases. For these reasons, aluminum hydroxide is the preferred raw material for the production of a large number of aluminum compounds. These include pure, iron-free aluminum sulfate (used in the paper industry and for water purification), aluminum fluoride, synthetic zeolites, and sodium aluminate. Aluminum hydroxide also is used in the glass industry and in cosmetic and pharmaceutical preparations. An important cosmetic use of aluminum hydroxide is in toothpaste. The mildly abrasive hydroxide cleans and polishes teeth.

The price of aluminum hydroxides ranged from \$0.20 to \$0.60 per kg in 1990, the price range reflecting the cost of additional processing of the usual Bayer process product to suit application requirements. These include grinding, higher purity, classification, surface treatment, etc. Pharmaceutical-grade gel hydroxides were at the top of the price range.

6.2. Adsorbent and Catalytic Aluminas

Activated aluminas represent another group of technically important alumina chemicals. Principal uses are as drying agents, adsorbents, catalysts, and catalyst carriers. These products are obtained by thermal dehydration of different aluminum hydroxides in the 520 – 1070 K temperature range (see Section 1.5).

6.2.1. Preparation of Activated Aluminas

Bayer aluminum hydroxide is the chief source of commercial activated alumina products. Powder forms of activated alumina are produced by heating the hydroxide directly at 575 – 1825 K in ovens or in rotary or fluidized-bed calciners. The products have surface areas of 200 – 350 m^2/g and losses on ignition of 3 – 12 wt % (at 575 – 1473 K). Such products are used as decolorizing agents for organic chemicals and as starting material for the production of aluminum fluoride (→ Fluorine Compounds, Inorganic). Other uses include chromatographic and catalytic applications in organic chemistry.

Granular activated alumina produced from Bayer plant crust is one of the oldest commercial forms of this product and still is used widely. A flow sheet of the production process is shown in Figure 20. The activated material is a hard, nondusting product. Table 9 lists some properties of a commercial product of this type (Alcoa F-1). A similar granular product has been produced by compacting Bayer hydroxide by mechanical pressure (Martinswerk GmbH, Federal Republic of Germany). The process utilizes a roll-type compactor. The product from the compactor is broken up and sieved to the required size fractions and activated at 675 – 875 K in a rotary calciner.

Table 9. Typical properties of Alcoa activated aluminas

	F-1	H-151	S-100
Al_2O_3, %	92	90	95
Na_2O, %	0.58	1.6	0.35
SiO_2, %	0.12	2.0	0.03
Fe_2O_3, %	0.06	0.03	0.05
Loss on ignition, %	7	6	5
Loose bulk density:			
g/cm^3	0.83	0.82	0.80
lb/ft^3	52	51	50
Packed bulk density:			
g/cm^3	0.85	0.85	0.75
lb/ft^3	55	53	47
Pore characteristics and surface area of typical Alcoa activated aluminas:			
Helium (true) density, g/cm^3	3.25	3.40	3.15
Mercury (particle) density, g/cm^3	1.42	1.38	1.24
Micro pore volume (pores < 3.5 nm), cm^3/g	0.017	0.023	0.012
Macro pore volume (pores > 3.5 nm), cm^3/g	0.023	0.020	0.037
Total pore volume, cm^3/g	0.40	0.43	0.49
Total porosity, %	56.3	59.4	60.6
Pore diameter at 50 % total pore volume, nm	17.7	3.5	4.7
Primary pore size range, nm	0 – 10000	0 – 40	0 – 500
BET surface area, m^2/g	250	360	260
Pore diameter, nm			
$10^6 - 10^7$	0.0031	0.0003	0.0000
Pore volume, cm^3/g			
$10^5 - 10^6$	0.0045	0.0001	0.0001
$10^4 - 10^5$	0.0059	0.0001	0.0004
$10^3 - 10^4$	0.0029	0.0004	0.0033
$10^2 - 10^3$	0.0045	0.0093	0.0047
$35 - 10^2$	0.0021	0.0099	0.0289
2 – 35	0.0165	0.0232	0.0115
	0.0395	0.0433	0.0489

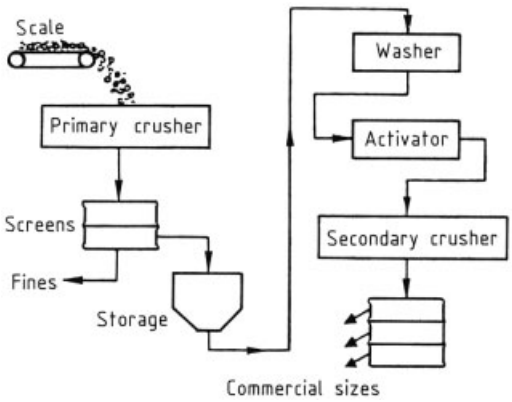

Figure 20. Flow sheet for production of activated alumina from Bayer plant crust

Fast dehydration of Bayer aluminum hydroxide, either by vacuum or by exposure to high-temperature gas (1050 – 1300 K) for a few seconds, has been used for the production of ball-shaped, activated alumina having properties superior to the granular product. This process results in the formation of nearly amorphous ϱ-Al_2O_3. The product is finely ground and using water as binder, formed into spherical agglomerates in a rotating pan agglomerator. Rehydration of ϱ-Al_2O_3 with water leads to crystallization of bayerite, causing the agglomerates to harden. Reactivation of the hard balls at 675 – 775 K produces the activated product (5 – 20 mm in diameter) having a surface area of 320 – 380 m^2/g.

Alumina gels also have been used for the manufacture of activated aluminas. These gels are produced by neutralization of aluminum sulfate or ammonium alum by NH_4OH, or from sodium aluminate by neutralization with acids, CO_2, $NaHCO_3$, and $Al_2(SO_4)_3$. The gelatinous aluminum hydroxide precipitate is filtered and thoroughly washed and dried. The dried prod-

uct is activated, milled, and agglomerated to a spherical product. Other forming processes, such as extrusion, pelletizing, and tableting, also can be used. The product is finally activated at 675 – 875 K to a loss on ignition value of about 6 %. Although gels of various textures can be prepared, the usual industrial adsorbent products have very small pores (less than 4 nm in diameter) and surface areas in the range of 300 – 400 m^2/g. A flow sheet of a manufacturing process is shown in Figure 21. Table 9 reports data on a commercial, gel-based product (Alcoa H-151).

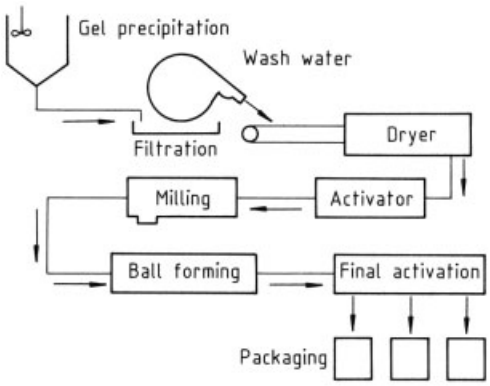

Figure 21. Production of gel-based activated alumina

Aluminum oxide hydroxide (boehmite), a by-product from the Ziegler process for linear alcohol production, is another source of activated alumina. The high purity of this material favors its use in catalytic applications. The fine-particle boehmite is normally extruded to various shapes. The material is claimed to serve as its own binder when peptized with glacial acetic acid. The extrudate is cut to the required size, dried, and activated at 775 – 875 K. The surface area of the activated product is 185 – 250 m^2/g. Commercial bayerite also has been used to produce activated alumina; the product is preferred in some catalytic applications.

6.2.2. Adsorbent Applications

A major application of activated alumina is in the field of adsorption, where its high surface area, pore structure, strength, and chemical inertness favor its use. The alumina performs important technical functions, such as gas and liquid drying, water purification, and selective adsorption in the petroleum industry.

Gases that have been dried successfully by alumina desiccants are:

acetylene	cracked gas	hydrogen	oxygen
air	ethane	hydrogen chloride	propane
ammonia	ethylene	hydrogen sulfide	propene
argon	freon	methane	sulfur dioxide
carbon dioxide	furnace gas	natural gas	
chlorine	helium	nitrogen	

In many applications an alumina desiccant can dry gas to a lower dew point than any other commercially available desiccant. The static water adsorption capacities of two typical activated aluminas (Alcoa F-1 and H-151) in contact with air at different relative humidity conditions are shown in Figure 22.

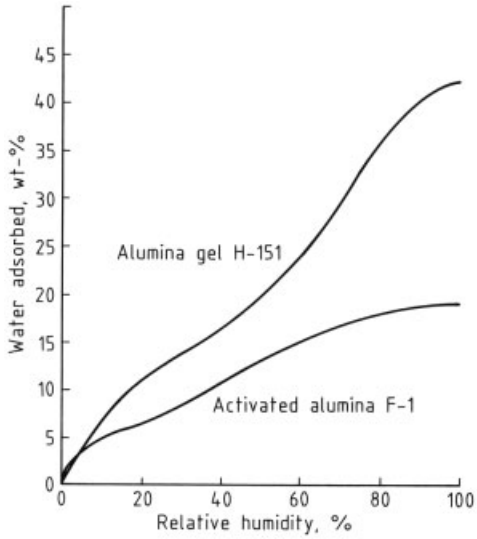

Figure 22. Static water adsorption capacity of typical commercial activated aluminas at 25 °C (298 K)

Moist gas usually is dried by passing it through a column (or tower) packed with the adsorbent. Adsorption of water on activated alumina is strongly exothermic, releasing between 45 and 55 kJ of heat per mole of water adsorbed. This factor must be considered in the design of the drying tower. Both granular and spherical forms of activated alumina can be used in desiccant beds. Sizes from 2 to 20 mm are available commercially. The granular products have lower

surface areas and their application is based on low cost per kilogram, low water load, and low stream pressure. Spherical products, produced by either the gel or the fast dehydration processes, have larger surface areas, a narrower pore structure, and a high adsorption capacity, and they are relatively more expensive than the granular variety. The spherical kind usually is specified for high-pressure, high-moisture removal duties. The alumina desiccant is regenerated by passing a current of hot, dry gas (475 K) through the bed, usually countercurrent to the main gas flow.

Liquids that can be dried with activated alumina include aromatic hydrocarbons, higher molecular mass alkanes, gasoline, kerosene, cyclohexane, power system coolants, lubricants, and many halogenated hydrocarbons. Liquids that are highly adsorbed on alumina (e.g., ethanol and methanol), react or polymerize in contact with activated alumina. Those containing components that tend to deposit on the alumina surface cannot be dried by activated alumina. Regeneration schemes for liquid dehydration units are varied and depend on the liquid being dried. In some applications the liquid being dried is vaporized, heated, and passed through the desiccant bed to adsorb water. Hot, dry gases also are used for regeneration.

An evolving application of activated alumina is in water purification. Several important contaminants have been removed from water successfully and economically in pilot plants as well as in large-scale treatment plants. These include reduction of fluoride concentration in drinking water and in some industrial effluents, and removal of color and odor from effluent water from dye works and paper mills. Removal of phosphate and arsenic also has been investigated.

Activated alumina can be used to separate one or more components from a gas or liquid stream by taking advantage of differences in adsorption or desorption kinetics. For example, a short cycle process has been used to recover heavy hydrocarbons from a stream of lighter hydrocarbons. Often a regeneration scheme can be devised that permits cyclic use of the alumina. In other instances, such as removal of catalyst in polyethylene and hydrogen peroxide production, it is more economical not to recycle the alumina.

6.2.3. Catalytic Applications

Alumina is used in many industrial catalytic processes, both as the catalyst and as a support for catalytically active components. In many instances, the alumina support contributes to catalytic activity and so assumes an essential role in the catalyst system. Other catalytic uses of alumina take advantage of its strength, heat resistance, and inertness.

Although the common adsorbent aluminas derived from Bayer hydroxide have many catalytic applications, high-purity materials (e.g., boehmite from the Ziegler process for linear alcohol production) are sometimes preferred. Because of the lower cost of Bayer process-based products, such techniques as water and acid washing of the activated product have been employed to reduce the alkali (Na_2O) content of the Bayer material. The sodium oxide is known to have a negative influence in many catalytic processes. Catalyst-forming methods include tableting, pelletizing, compacting, ball forming (agglomeration), and extrusion. Such factors as purity, surface area, pore volume, and size distribution, and rate of deactivation influence catalyst performance and selectivity. In addition, crushing strength and resistance to attrition of the catalyst pellets are important considerations in practical operation of catalytic reactors. Although these physical aspects of alumina catalysts are well characterized, the surface structure and chemistry responsible for catalytic activity still remains unclear. Many investigators have attributed catalytic activity to the intrinsic acidity of the surface of activated alumina. First, the combination of two neighboring OH^- groups to form water during the dehydration process leaves behind an exposed Al^{3+} ion, which, because of its electron deficiency, behaves as a Lewis acid site. In addition, hydroxyl groups are retained during thermal decomposition of aluminum hydroxide (see Section 1.5). These OH^- ions on the surface may act as proton donors (Brønsted acids). These Lewis and Brønsted acid sites have been looked upon as the active catalytic centers. Further "defect" structures are formed with increasing degree of dehydration [56]. Of the types of defects created during dehydration, those assumed to have the greatest catalytic importance are the triplet va-

cancies. They provide unusual exposure of the aluminum ions in the underlying layer and constitute strong acid sites.

Catalytic applications of alumina are extensive. Important industrial processes employing alumina as a catalyst by itself include alcohol dehydration and the Claus process for sulfur recovery (→ Sulfur). Dehydration of alcohols over activated alumina is one of the oldest catalytic processes. Typical reaction conditions for olefin production are 575 – 675 K and atmospheric pressure. Lower temperatures favor the formation of ethers. The most suitable aluminas for alcohol dehydration catalysis are those that have large surface areas (150 – 200 m^2/g) and possess good thermal and hydrothermal stability. Coke formation occurs over a period of several hundred operating hours and the catalyst must be regenerated by burning off the carbon with hot air at 773 – 873 K.

The largest present-day catalytic application of activated alumina itself is in the Claus process, which is used to recover sulfur from hydrogen sulfide (H_2S).

$$2\,H_2S + SO_2 \rightarrow 2\,H_2O + 3\,S \quad \Delta H = -147\,\text{kJ}$$

The reaction is carried out catalytically in two or more conversion stages using alumina catalysts. Reaction temperature in the first stage is around 625 K. At this temperature, the conversion of H_2S is only about 65 %. Subsequent lower temperature catalytic stages are used to further reduce the H_2S concentration. Spherical, high-strength activated alumina catalysts are used in the Claus converters. Service life as high as 5 years has been reported. Deactivation of the alumina catalyst occurs by sulfation, thermal aging, and carbon and/or sulfur deposition. Regeneration of Claus catalyst involves removal of sulfur and burning off of carbon deposits.

Alumina-supported catalysts are used extensively in the petroleum and chemical industries. In general, the petroleum industry catalysts have high surface area and high porosity; the support is mostly activated alumina. On the other hand, many typical chemical process catalysts (e.g., ammonia synthesis, steam reforming) are characterized by lower surface area ($< 20\,m^2/g$) and are nonporous or have very large diameter pores. The carrier in this case is inert and consists mostly of calcined or sintered alumina products.

Two different methods have been used commonly for preparation of alumina-supported catalysts: impregnation and coprecipitation. The alumina support used in the *impregnation process* generally has been formed into its final shape (extrudates, tablets) prior to the impregnation step. Impregnation with a salt solution of the active species is then carried out, followed by drying and thermal decomposition of the salt. In the *coprecipitation procedure,* hydroxides of aluminum together with the active component are precipitated from a salt solution by neutralization with ammonia or alkali. The washed precipitate is dried, powdered, and processed (e.g., by extrusion) to the desired shape and finally activated by thermal dehydration. The coprecipitation method is used when the active species must be present in high concentrations or when more uniform distribution of the active component is desired.

The technical and patent literature contains innumerable examples of the use of alumina as a catalyst support. Some important examples are the catalytic dehydration of *n*-butane to butadiene (used in synthetic rubber), using a chromia-impregnated alumina catalyst; molybdenum – alumina catalysts, used in hydrorefining operations (e.g., desulfurization) in petroleum refining; and pelleted catalysts containing platinum, palladium, and rhodium on an alumina base, used in automobile exhaust catalytic converters.

United States production of adsorbent-grade activated aluminas, both granular and spherical, amounted to nearly 250 000 t in 1990. Price of the cheaper, granular product was quoted around \$ 0.40 – 0.50 per kg. Price of the spherical product ranged from \$ 0.55 to \$0.65 per kg. Total catalytic applications of alumina in the United States were estimated to be around 400 000 t in 1990. Price of preformed alumina for catalytic applications ranges from \$ 0.60 to \$ 4.00 per kg, depending on source and purity. Alumina-based Claus catalyst was priced between \$ 3.00 and \$ 3.50 per kg.

7. Ceramic Uses of Alumina

Alumina is used extensively as a ceramic material (→ Ceramics, Advanced Structural Products). Products range from relatively low-calcined grades of polishing aluminas to the

extremely hard, fused alumina and synthetically produced sapphire. The characteristics that make alumina valuable in ceramic applications are high melting point (2325 K), hardness (9 on the Mohs scale), strength, dimensional stability, chemical inertness, and electrical insulating ability. These, together with availability in large quantities at moderate prices, have led to extensive and varied uses of alumina as a ceramic material.

7.1. Calcined Alumina

Ceramic aluminas are generally produced by calcining Bayer aluminum hydroxide at temperatures high enough for the formation of α-Al_2O_3. By control of calcination time and temperature and by the addition of mineralizers, such as fluorine and boron, the crystallite size in the calcined product can be varied from 0.2 to 100 µm. These calcined aluminas can be categorized broadly according to their sodium content. There are two general types: those having about 0.5 % Na_2O and low-soda grades with a content < 0.1 %.

Reactive alumina is a material manufactured by dry grinding calcined alumina to particle sizes smaller than 1 µm. The large surface area associated with very fine particles and the high packing densities obtainable considerably lower the temperatures required for sintering.

Tabular aluminas are manufactured by grinding, shaping, and sintering calcined alumina. The thermal treatment at 1900 – 2150 K causes the oxide to recrystallize into large, tabular crystals of 0.2 – 0.3 mm.

7.2. Fused Alumina

For ceramic applications and for the production of abrasives, fused aluminas are manufactured by melting a suitable raw material in an electric arc furnace (→ Abrasives). Calcined alumina from the Bayer process is used as a starting material for the highest quality fused alumina. Bauxites with varying levels of iron oxide, silicates, and titanium minerals are melted to produce the brown or less pure black qualities.

8. Toxicology and Industrial Hygiene

For a discussion concerning the toxicology and industrial hygiene of aluminum oxides, → Aluminum, Chap. 9.

9. References

General References

1. H. Ginsberg, F. W. Wrigge: *Tonerde und Aluminium, Teil 1, Die Tonerde,* W. De Gruyter, Berlin 1964.
2. H. Ginsberg, K. Wefers: *Aluminium und Magnesium, vol. 15, Die Metallischen Rohstoffe,* Enke Verlag, Stuttgart 1971.
3. K. Wefers, C. Misra: *Oxides and Hydroxides of Aluminum,* Alcoa Technical Paper no. 19, revised, Pittsburgh 1987.
4. T. G. Parson: *The Chemical Background of the Aluminum Industry; Lectures, Monographs, and Reports,* no. 3, The Royal Institute of Chemistry, London 1955.
5. P. Barrand, R. Gadeau: *l'Aluminium,* part 1, Editions Eyrolles, Paris 1964.
6. A. N. Adamson, E. J. Bloore, A. R. Carr: *Basic Principles of Bayer Process Design, Extractive Metallurgy of Aluminum,* **vol. 1,** Interscience Publishers, New York 1963.
7. A. N. Adamson: "Aluminum Production Principles and Practice," *The Chemical Eng.* (London), June 1970, 156 – 171.
8. S. I. Kuznetsov, V. A. Derevyankin: *The Physical Chemistry of Alumina Production by the Bayer Process,* Moscow 1964.
9. W. H. Gitzen: *Alumina as a Ceramic Material,* The Amer. Cer. Society, Columbus, Ohio 1970.
10. C. Misra: *Industrial Alumina Chemicals, ACS Monograph 184,* The Amer. Chem. Society, Washington, DC 1986.
11. L. Jacob, Jr. (ed): *Bauxite,* The Amer. Chem. Society, Westerville, Ohio 1990.
12. L. D. Hart (ed): *Alumina Chemicals Science and Technology Handbook,* Society Of Mining Engineers (AIME), New York 1984.

Specific References

13. J. Böhm, H. Niclassen, *Z. Anorg. Allg. Chem.* **132** (1923) 1.
14. R. Fricke, *Z. Anorg. Allg. Chem.* **175** (1928) 249; **179** (1929) 287.
15. T. G. Gedeon, *Acta Geol. Acad. Sci. Hung.* **4** (1956) 95.

16. R. A. Van Nordstrand, W. P. Hettinger, C. Keith, *Nature (London)* **177** (1956) 713.
17. H. Saalfeld, *Neues Jahrb. Mineral. Abh.* **95** (1960) 1.
18. H. Ginsberg, W. Hüttig, H. Stiehl, *Z. Anorg. Allg. Chem.* **309** (1961) 233; **318** (1962) 238.
19. K. Wefers, *Naturwissenschaften* **49** (1962) 204.
20. K. R. Poppelmeier, S. J. Hwu, *Inorg. Chem.* **26** (1987) 3297 – 3306.
21. U. Hauschild, *Z. Anorg. Allg. Chemie* **324** (1963) 15.
22. B. C. Lippens, Ph. D. Dissertation, University Delft (1961).
23. H. D. Megaw, *Z. Kristallogr.* **87** (1934) 185.
24. F. J. Ewing, *J. Chem. Phys.* **3** (1935) 203.
25. W. R. Busing, H. A. Levy, *Acta Crystallogr.* **11** (1958) 798.
26. O. Glemser, E. Hartert, *Z. Anorg. Allg. Chem.* **283** (1956) 111.
27. W. H. Bragg, *J. Chem. Soc.* **121** (1922) 2766.
28. A. W. Laubengayer, R. S. Weiss, *J. Am. Chem. Soc.* **65** (1943) 247.
29. G. Ervin, E. F. Osborn, *J. Geol.* **59** (1951) 381.
30. G. C. Kennedy, *Am. J. Sci.* **257** (1959) 563.
31. A. Neuhaus, H. Heide, *Ber. Dtsch. Keram. Ges.* **42** (1965) 167.
32. K. Wefers, *Erzmetall* **20** (1967) 13, 71.
33. H. Schwiersch, *Chem. Erde* **8** (1933) 252.
34. A. H. Carim et al., *J. Am. Ceram. Soc.,* **80** (1997) 2677 – 2680.
35. W. Feitknecht, A. Wittenbach, W. Buser: *4th Symposium on Reactivity of Solids,* Elsevier, Amsterdam 1961, p. 234.
36. F. Freund, *Ber. Dtsch. Keram. Ges.* **42** (1965) 23; **44** (1967) 141.
37. J. M. McHale et al., *Science* **277** (1977) 788 – 791.
38. K. Wefers, *Erzmetall* **17** (1964) 583.
39. H. C. Stumpf, A. S. Russell, J. W. Newsome, C. M. Tucker, *Ind. Eng. Chem.* **42** (1950) 1398.
40. I. Valeton: "Bauxites," *Dev. Soil Sci.* **1**, (1972).
41. J. G. deWeisse: *Les Bauxites de l'Europe Centrale,* Univ. Lausanne 1948.
42. V. Z. Zans, *Geonotes* **5** (1959) 123.
43. M. E. Roch: "Les bauxites du Midi de la France," *Rev. Gen. Sci. Pures Appl. Bull. Assoc. Fr. Av. Sci.* **66** (1958) no. 5/6, 151 – 156.
44. H. Erhart: "La Genèse des Sols," *Evolution des Sciences,* Masson, Paris 1956.
45. G. Bardossy: *Bibliographie des Travaux Concernant les Bauxites,* ICSOBA, Paris 1966.
46. U.S. Geological Survey, Mineral Commodity Summaries, January 1998.
47. Nippon Light Metal Col., DE 2 807 245, 1977 (M. Kanehara, A. Kainuma, M. Fujiike).
48. P. Basu, "Reactions of Iron Minerals in Sodium Aluminate Solutions," *Light Met. (New York)* **1983,** 83 – 98.
49. D. K. Grubbs, S. C. Libby, J. K. Rodenburg, K. Wefers, *J. Geol. Soc. Jam.* **4** (1980) 176 – 186.
50. K. Yamada, T. Harato, H. Kato, *Light Met. (N.Y.)* 1981, 117 – 128.
51. C. Misra, *Travaux ICSOBA* **24** (197) 194 – 204.
52. *UNEP/UNIDO Workshop Envir. Asp. Alumina Prod.,* Paris 1981.
53. K. Bielfeldt, K. Kämpf, G. Winkhaus, *Erzmetall* **29** (1976) 120 – 125.
54. The Aluminum Association, Inc., N.Y., Report no. 4, Sept. 15, 1981.
55. G. Lang, K. Solymar, J. Steiner, *Light Met. (N.Y.)* 1981, 201 – 214.
56. H. W. St. Clair, D. A. Elkins, W. M. Mahan, R. C. Merritt, M. R. Howcroft, M. Hayashi, Bulletin 577, U.S. Bureau of Mines, 1959.
57. K. B. Bergston, *Light Met. (N.Y.)* 1979, 217 – 282.
58. N. Gjelsvik, *Light Met. (N.Y.)* 1980, 133 – 148.
59. V. S. Sazlin, A. K. Zapolskii, *Tsvetn. Metall. UDK 669.712.2* (1968) 64.
60. E. Barrilon, *Erzmetall* **31** (1978) 519 – 522.
61. H. Knözinger, P. Ratnasamy, *Catal. Rev. Sci. Eng.* **17** (1978) 31 – 70.

Amalgams → Mercury, Mercury Alloys, and Mercury Compounds
Americium → Transuranium Elements
Amidosulfuric Acid → Sulfamic Acid
Amine Oxides → Surfactants

Amines, Aliphatic

Ethanolamines and Propanolamines and Methylamines are separate keywords; quaternary ammonium compounds, fatty amine oxides → Surfactants

Karsten Eller, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 – 9)

Erhard Henkes, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 – 9)

Roland Rossbacher, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 10)

Hartmut Höke, Weinheim, Federal Republic of Germany (Chap. 10)

1. Introduction . . . 1
2. General Chemical Properties . . . 2
2.1. Salt Formation . . . 2
2.2. Conversion to Carboxamides . . . 2
2.3. Conversion to Sulfonamides . . . 2
2.4. Reaction with Carbonyl Compounds . . . 3
2.5. Reaction with Carbon Dioxide and Carbon Disulfide . . . 3
2.6. Reaction with Epoxides . . . 3
2.7. Alkylation . . . 3
2.8. Formation of Isocyanates and Ureas . . . 3
2.9. Reaction with Acrylonitrile . . . 4
2.10. Formation of Isonitriles . . . 4
2.11. Oxidation . . . 4
2.12. Dealkylation . . . 4
3. General Production Methods . . . 5
3.1. Production from Alcohols . . . 5
3.2. Production from Carbonyl Compounds . . . 6
3.3. Production from Nitriles . . . 6
3.4. Production from Alkyl Halides . . . 7
3.5. Production from Nitro Compounds 7
3.6. Production from Olefins . . . 7
3.7. Other Processes . . . 8
4. Lower Alkylamines . . . 8
4.1. Physical Properties . . . 8
4.2. Storage and Transportation . . . 9
4.3. Quality Specifications and Analysis 9
4.4. Production . . . 9
4.5. Uses . . . 12
4.6. Economic Aspects . . . 14
5. Cycloalkylamines . . . 14
6. Cyclic Amines . . . 19
7. Fatty Amines . . . 25
7.1. Properties . . . 25
7.2. Production . . . 27
7.3. Analysis and Quality Control . . . 29
7.4. Uses . . . 30
7.5. Economic Aspects . . . 31
8. Diamines and Polyamines . . . 32
8.1. Diamines . . . 32
8.1.1. 1,2-Diaminoethane (Ethylenediamine) . . . 32
8.1.2. Diaminopropanes . . . 34
8.1.3. Higher Diamines . . . 36
8.2. Oligoamines and Polyamines . . . 37
8.2.1. Physical Properties . . . 37
8.2.2. Chemical Properties . . . 39
8.2.3. Production, Analysis, and Uses . . . 39
9. Chiral Amines . . . 40
10. Toxicology and Occupational Health . . . 40
10.1. General Aspects . . . 40
10.2. Toxicology of Specific Amines . . . 42
10.2.1. Alkylamines, Cyclic Amines, and Polyamines . . . 42
10.2.2. Fatty Amines . . . 46
11. References . . . 46

Based on the corresponding article from the fifth edition, written by Gerd Heilen, Hans Jochen Mercker, Dieter Frank, Richard A. Reck, and Rudolf Jäckh

1. Introduction

Primary, secondary, and tertiary amines are distinguished on the basis of the number of hydrogen atoms in ammonia that have been replaced by organic groups. Substitution at the nitrogen atom by a fourth substituent gives quaternary ammonium compounds. The lower aliphatic amines are those with up to six carbon atoms per alkyl chain. Long-chain amines, with more than eight atoms per carbon chain, are generally known as fatty amines and are discussed in Chapter 7.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_001

In 1849, WURTZ prepared methylamines and ethylamines by hydrolysis of the corresponding alkyl isocyanates, trialkyl cyanurates, and alkylureas [1]. However, HOFMANN was the first to use the terms "primary", "secondary", and "tertiary" and to carry out fundamental work on the synthesis, properties, and structure of amines [2].

In general, naturally occurring amines are relatively complex compounds (e.g., alkaloids, vitamins, and amino acids); however, some lower alkylamines and diamines, such as tetramethylenediamine (putrescine) and pentamethylenediamine (cadaverine), are known as degradation or decomposition products of proteins.

With a worldwide production (2000) of several 100 000 t/a, the aliphatic amines are among the most important organic intermediates in the chemical industry. The range of uses of these compounds is correspondingly wide. Major uses include those in the production of agrochemicals (in particular, herbicides), dyes, drugs, surfactants, and plastics, as auxiliaries for the rubber, textile, and paper industries, and as anticorrosion agents and process chemicals for gas scrubbing.

2. General Chemical Properties

The chemistry of the aliphatic amines is determined by the electron lone pair on the nitrogen atom and by the tendency of the hydrogen atoms bonded to nitrogen to be replaced by other substituents.

2.1. Salt Formation

Because they carry alkyl substituents, the aliphatic amines are stronger bases than ammonia (see Table 1); with acids, they form salts that are very soluble in water but insoluble in organic solvents. This property, combined with the difference in solubility between amine and salt, makes amines good acid acceptors and solvents for gas scrubbing and for certain extraction processes (e.g., in the synthesis of semisynthetic penicillins).

2.2. Conversion to Carboxamides

Amines react with carboxylic acids and their esters, chlorides, and anhydrides to give the corresponding substituted carboxamides:

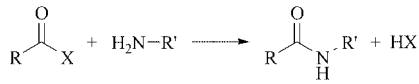

X = OH, OR, Cl

Very good yields are obtained, particularly with carboxylic acid chlorides, for which the reaction is highly exothermic. With carboxylic acids, the reaction often stops at the initially formed ammonium salt. This type of reaction is utilized industrially, for example, in the synthesis of various herbicides with an acid amide structure.

2.3. Conversion to Sulfonamides

The reaction with benzenesulfonyl chloride is utilized for distinguishing among primary, secondary, and tertiary amines (Hinsberg test), as well as for their preparative separation. Whereas primary amines form alkali-soluble *N*-alkyl benzenesulfonamides, secondary amines give alkali-insoluble *N,N*-dialkyl benzenesulfonamides, and tertiary amines do not react under these conditions. Limitations can arise with long-chain primary amines, which are alkali-insoluble despite their acidic hydrogen, and with tertiary amines, which may undergo quaternization with the sulfonyl chloride.

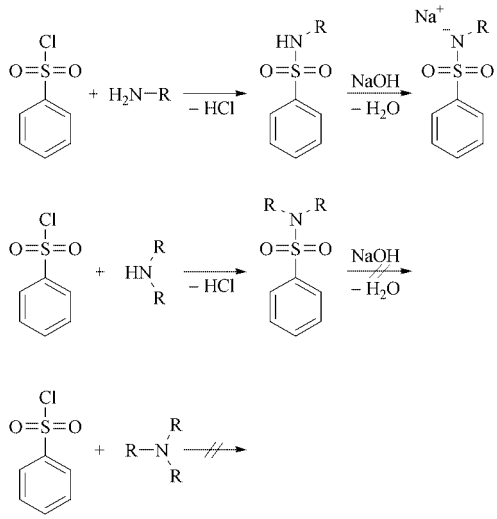

2.4. Reaction with Carbonyl Compounds

Depending on the reaction conditions and the compound employed, carbonyl compounds react with amines to form imines (Schiff bases) (**2**) or enamines (**3**); these products can be hydrogenated to give more highly alkylated amines. This reaction is an important method for synthesizing higher amines. The intermediate hemiaminals (**1**) are usually not isolable. Aldehydes generally react faster than ketones. Secondary amines can only form enamines, and tertiary amines neither of the two products.

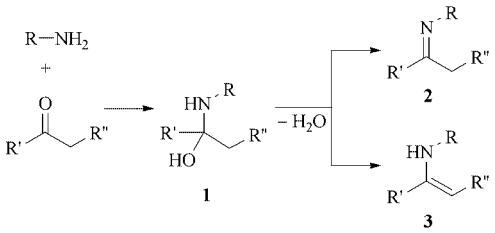

An amine, an aldehyde, and a compound with an activated hydrogen atom, such as a ketone, can react to form a Mannich base.

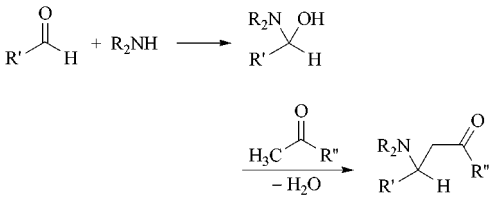

In this case, the condensation product between the aldehyde and the amine is attacked by the C – H acidic compound, in this case the ketone.

2.5. Reaction with Carbon Dioxide and Carbon Disulfide

The carbamic acid or dithiocarbamic acid formed in this reaction is unstable but can be isolated as a salt or an ester.

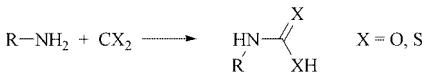

Dithiocarbamates obtained from various amines play a key role as vulcanization accelerators in the rubber industry.

2.6. Reaction with Epoxides

Primary amines react with epoxides to give a mixture of mono- and dioxyalkylated derivatives, whereas secondary amines give only monooxyalkylated compounds, and tertiary amines form quaternary ammonium compounds.

This oxyalkylation reaction is one of the most important industrial reactions of aliphatic amines and is utilized for the preparation of flocculants, surface coating resins, drug intermediates, and products for gas scrubbing (→ Ethanolamines and Propanolamines).

2.7. Alkylation

The reaction of amines with alkyl halides and dialkyl sulfates to give, ultimately, quaternary ammonium compounds is utilized in preparative pharmaceutical chemistry and for the preparation of anticorrosion agents and biocides.

The reaction of alkyl halides with ammonia is not very useful for the preparation of primary or secondary amines since these are stronger bases and preferentially attack the halides. Tertiary amines can be prepared, though. On a laboratory scale, primary amines can be obtained by using tetramethylenehexamine, and secondary amines with cyanamide.

2.8. Formation of Isocyanates and Ureas

The reaction of phosgene with primary amines leads first to the corresponding carbonyl chloride; subsequent cleavage of hydrogen chloride

gives the alkyl isocyanate. Excess amine reacts with the isocyanate with formation of ureas:

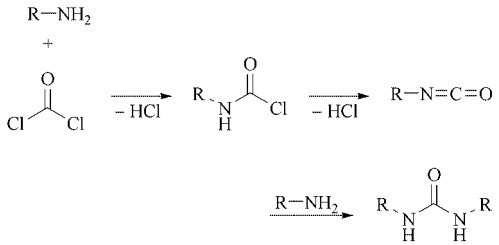

This reaction is important in the preparation of various herbicides with urea, carbamate, or thiocarbamate structures and, particularly in the case of polyfunctional amines, in polyurethane chemistry. The reaction of secondary amines with phosgene proceeds via an analogous intermediate chloride to form *N*,*N*′-tetraalkylureas.

2.9. Reaction with Acrylonitrile

The addition of a primary or secondary amine to acrylonitrile to form an aminopropionitrile is utilized industrially on a large scale for the preparation of higher diamines and polyamines, because the nitrile can be readily hydrogenated to the amine.

$$RNH_2 + CH_2{=}CHCN \longrightarrow RNHCH_2CH_2CN$$
$$[\longrightarrow RNH(CH_2)_3NH_2]$$

2.10. Formation of Isonitriles

Primary amines react with trichloromethane under basic conditions to form isonitriles:

$$RNH_2 + CHCl_3 + 3\,KOH \longrightarrow RNC + 3\,KCl + 3\,H_2O$$

Because of the strong unpleasant odor of the isonitriles, the reaction can be used as a test for primary amines. For the lower amines, the reaction can be used synthetically as well. An alternative route uses the *N*-alkylformamides to synthesize isonitriles.

2.11. Oxidation

In contrast to their salts, the free amines are sensitive to oxidation, giving various products, depending on the oxidizing agent and the type of amine. Tertiary amines are oxidized by hydrogen peroxide to amine oxides, whereas the corresponding compounds formed from primary and secondary amines undergo further reaction to give the corresponding hydroxylamines or aldoximes:

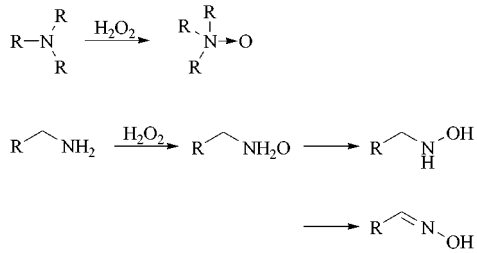

Oxidation with nitrous acid can be used to distinguish among primary, secondary, and tertiary amines. Primary amines undergo diazotization followed by loss of gaseous nitrogen and afford alcohols by reaction of the intermediate carbenium ion with water. Secondary amines react to give yellow *N*-nitrosoamines. Normally, tertiary amines do not react.

$$RCH_2NH_2 + HNO_2 \longrightarrow RCH_2OH + H_2O + N_2$$

$$(RCH_2)_2NH + HNO_2 \longrightarrow (RCH_2)_2N{-}N{=}O + H_2O$$

The formation of nitrosamines can be a serious problem, particulary with secondary amines, since they are highly mutagenic. During normal use they are not formed, but upon contact with nitrites or nitrous oxides traces can be produced that give rise to problems during handling. Air contamination with secondary amines is thus considered quite problematic.

2.12. Dealkylation

Tertiary amines can be dealkylated by converting them to basic quaternary ammonium salts, which decompose upon heating. If only methyl groups are present, they are eliminated as methanol or dimethyl ether; all higher alkyl groups furnish alkenes, e.g.,

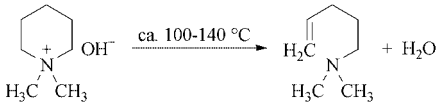

This reaction, known as the Hofmann degradation, is seldom utilized synthetically, but used to be important for elucidating the structure of unknown amines.

3. General Production Methods

3.1. Production from Alcohols

The reaction of the appropriate alcohol with ammonia over a suitable catalyst is now the most common process for the preparation of lower alkylamines. In this reaction, the product is always a mixture of primary, secondary, and tertiary amines because the primary amine formed initially can react further with one or two molecules of alcohol. While the initial conversion of the alcohol to the primary amine is close to thermoneutral, the formation of secondary and tertiary amines is exothermic and thus thermodynamically favored. Also, the primary amine is more nucleophilic than ammonia and hence has a higher reactivity. The product distribution can be controlled to a certain extent by means of the reaction conditions (temperature, excess of ammonia, residence time). Because the mixture of products usually obtained does not correspond to market requirements, amines that are not marketable can be recycled. Thus, the yield of the desired amine can be increased to above 90 %.

Previously, this reaction was carried out with pure dehydration catalysts (e.g., alumina, silica, titania, thorium oxide, tungsten oxide, chromium oxides, phosphates or various mixed oxides such as silica – alumina, clays, or zeolites [3]) and temperatures up to 500 °C; however, use of these catalysts is now restricted to the production of methylamines. For the conversion of an alcohol containing two or more carbon atoms, catalysts possessing hydrogenating and dehydrogenating properties have become important. The catalysts used are mainly based on nickel, cobalt, copper or iron, and to a lesser extent platinum or palladium. Noble metal catalysts have a tendency to cause C – C or C – N bond cleavage and can hence give lower selectivities. Promoters such as Ag, Zn, In, Mn, Mo, and alkali metals are added to improve catalyst performance. Mostly, the active components are supported on a carrier such as Al_2O_3, SiO_2, or ZrO_2. Zeolites have been suggested as carriers to improve the selectivity towards primary amines [4]. In this process, the alcohol, ammonia, and hydrogen are passed continuously over the catalyst in a fixed-bed reactor. The reaction takes place at about 0.5 – 25 MPa and about 100 – 250 °C, depending on the catalyst and on whether the liquid-phase or the gas-phase process is used.

Mechanistic details of the process have been reviewed [5]. The first and rate-determining step is the dehydrogenation of the alcohol to the carbonyl compound. Addition of ammonia is followed by loss of water to yield an imine or an enamine, which is hydrogenated to the final amine. Side reactions are amine disproportionation, especially at higher reaction temperatures, and to a lesser degree aldol condensation of the intermediate aldehydes, formation of Schiff base from these aldehydes and the amine product, and nitrile formation, which occurs at high temperatures and low hydrogen pressures.

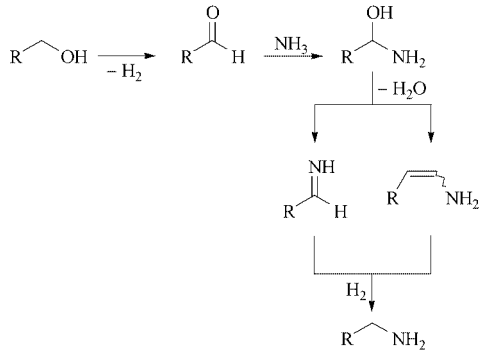

A two- to eightfold excess of ammonia is used to shift the equilibrium towards the primary amines. Although hydrogen is not required as a direct reactant, in its absence imines, enamines, and even nitriles are formed. Hydrogen also maintains the activity of the catalyst by removing carbonaceous deposits and metal carbides or nitrides and prevents disproportionation of the amine products [6]. The various versions of the process have been reviewed in detail [7]. Several surface-bound intermediates were identified

by Fourier transform IR spectroscopy [8]. While the amination of diols and polyols in principle follows the same rules as for the monools, the occurrence of numerous side reactions, for instance, cyclizations, is a complicating factor [9].

The same method is also useful for alkylating primary or secondary amines instead of ammonia. When an amine and an alcohol having different aliphatic substituents react, the product is a mixed aliphatic amine. For example, *N*,*N*-dimethylethylamine can be obtained from dimethylamine and ethanol. To avoid transalkylation at the nitrogen atom in reactions of this type, copper catalysts are recommended [10]. Copper chromite is a particularly effective catalyst for long-chain tertiary amines such as dimethyldodecylamine [11].

3.2. Production from Carbonyl Compounds

Aldehydes and ketones are used in preference to alcohols for amine synthesis if this is more economical. Generally, this is the case only for the lower aldehydes (obtained from oxo synthesis) and acetone obtained as a byproduct in the production of phenol (Hock synthesis).

The reaction between a carbonyl compound and ammonia or an amine occurs in two steps: the imine or Schiff base formed initially is hydrogenated with hydrogen in a second step to give the amine. The usual procedure is similar to that for the conversion of alcohols; the reaction mixture, comprising the carbonyl compound, ammonia and hydrogen, is passed over a fixed-bed catalyst. In some cases, it can be advantageous to carry out the reaction in two stages. The carbonyl compound and ammonia or the amine react first, the water of reaction is removed, and only then is the hydrogenation carried out [12].

The essential difference between the amination of an aldehyde or ketone and that of an alcohol is that in the former hydrogen is a reactant and is consumed in a stoichiometric amount. The substantially higher heat of reaction of this process (e.g., 60.4 kJ/mol for acetone [13] versus 7.1 kJ/mol for isopropyl alcohol) necessitates a fundamentally different reactor design. In general, the process is carried out in the vapor phase at 100 – 160 °C and atmospheric or slightly above atmospheric pressure. High-pressure processes can also be employed if the heat of reaction can be removed by appropriate measures, such as a high recycle rate or with a multitubular reactor. Since the reaction temperatures are lower than for the reaction of alcohols, in principle, higher selectivities can be achieved.

Generally, the same catalysts as for the hydrogenative amination of alcohols can be used; control of product distribution by adjusting the amount of excess ammonia and workup by distillation are also much the same. Comprehensive reviews of industrial and preparative possibilities for synthesis of amines from carbonyl compounds are available [7], [14].

It is also possible to use other reducing agents instead of hydrogen, but these are of little industrial importance. This is often due to the difficulty of handling and the cost of the reagents (e.g., selenophenol or $NaBH_4$). The Leuckart – Wallach reaction, in which ammonium formate is used to aminate the carbonyl compound with formation of CO_2 as byproduct, has found some use.

3.3. Production from Nitriles

Frequently, nitriles can be obtained more economically than the corresponding alcohols or carbonyl compounds. They are synthesized by ammoxidation, the addition of an amine or an alcohol to acrylonitrile, or from the corresponding carboxylic acid. Nitriles are converted into the corresponding amines, preferably by catalytic hydrogenation:

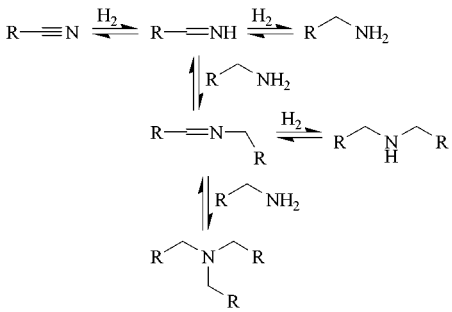

Noble metal (palladium, platinum, rhodium), nickel, or cobalt catalysts are generally em-

ployed, but iron catalysts have been used to an increasing extent [15], largely for economic reasons. Whereas the supported noble metals permit the use of relatively mild reaction conditions (20 – 100 °C, 0.1 – 0.5 MPa), pressures as high as 25 MPa and temperatures of up to 180 °C are required with nickel or cobalt catalysts. When di- or polynitriles are hydrogenated to di- or polyamines, a base such as NaOH is often added to the catalyst [16]. There are various versions of the process, but those most often used are the batch process with a suspended catalyst and the continuous process with a fixed-bed catalyst.

Side reactions resulting in the formation of secondary amines always occur, but these can be suppressed by the addition of ammonia, sodium hydroxide solution, or an acid [7]. Alternatively, process conditions can be adjusted to make the formation of secondary or tertiary amines the principal reaction. This alternative is of interest, for example, in the synthesis of di- and triethylamines from acetonitrile [17] or dibutylamine from butyronitrile [18].

The hydrogenation of acrylonitrile is a special case. Whereas saturated amines are normally the only products of this reaction, the use of copper chromite as a catalyst instead of noble metals, nickel, or cobalt leads to allylamine and diallylamine [19].

Many important nitrile hydrogenations to form the corresponding amines are based on previous addition reactions of acrylonitrile. Alcohols add to acrylonitrile only in the presence of a basic catalyst such as NaOH, KOH, or quaternary ammonium hydroxides. Since amines are themselves sufficiently basic, addition occurs readily. Sometimes higher temperatures are needed for cyanoethylation, and stabilizers have been described to avoid polymerization of acrylonitrile [20].

3.4. Production from Alkyl Halides

The reaction of an alkyl halide with ammonia or an amine gives an alkylammonium halide, from which the amine can be liberated with a caustic alkali solution. Although a standard method of preparative chemistry, this route is industrially important only for the preparation of ethylenediamine, the homologous polyamines, and a few special amines such as allylamine and some small-volume pharmaceuticals. The reasons for this restricted use are the lack of cheap starting materials, corrosion and product-quality problems that arise in the processing of halides, and the need to dispose of the salt that is formed.

3.5. Production from Nitro Compounds

Whereas the reduction of nitro compounds is one of the most important methods for the synthesis of aromatic amines, this route has not become important for the production of aliphatic amines. Because the availability of appropriate nitroalkanes is restricted, this method is used only in a few special cases, e.g., the production of 2-amino-1-butanol, the precursor for the antituberculotic ethambutol.

The reduction proceeds via the nitroso or hydroxylamine intermediate, in some cases even at room temperature, and gives the amine in yields of 90 % or more. Platinum, palladium, rhodium, nickel, or copper is used as the hydrogenation catalyst [7]. The reaction is highly exothermic (ca. 500 kJ/mol per nitro group), and therefore heat must be removed rapidly to control the process. Nitroparaffins are more difficult to hydrogenate than their aromatic counterparts [21].

Functionalization of alkenes with NO or NO_2 has been described; the reaction follows a radical path and affords a mixture of products, among them nitroalkenes. Reduction of the latter also leads to amines, but has found no industrial application so far. Likewise, primary amines were obtained from paraffins by nitration with HNO_3 or liquid NO_2 and hydrogenation of the resulting nitroparaffin/ketone mixture in the presence of ammonia [22]. The preparation of 1,2-diamines from nitroolefins and *O*-alkylhydroxylamines has been described [23].

3.6. Production from Olefins

Amines having a tertiary alkyl group adjacent to the nitrogen atom (e.g., *tert*-butylamine) are difficult to obtain by conventional synthetic methods. These compounds can be prepared readily by addition of hydrogen cyanide to an alkene, e.g., 2-methylpropene, in an acidic medium:

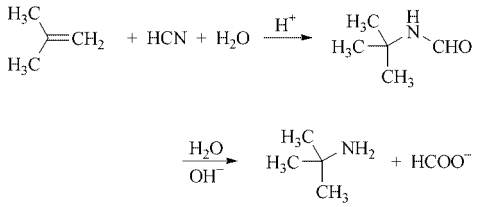

This process, known as the Ritter reaction [24], is carried out at 30 – 60 °C, and the resulting formamide intermediate is hydrolyzed when the reaction mixture is heated to ca. 100 °C. To liberate the amine, the acidic mixture is neutralized to give sodium formate and the salt of the original acid, i.e., sodium sulfate. Because hydrocyanic acid is difficult to handle and causes corrosion problems, one carbon atom is lost as formate, and significant amounts of salts (ca. 3.3 kg/kg *tert*-butylamine) must be disposed of in deep wells, this process is restricted to amines that cannot be obtained by one of the methods described in Sections 3.1, 3.2, 3.3, 3.4, and 3.5. Besides *tert*-butylamine and *tert*-octylamine, these are only a few amines of pharmaceutical interest.

Although advantageous thermodynamically, the direct addition of ammonia or an amine to an olefinic double bond is restricted to a few cases in which either an activated amine or a compound with an activated double bond is employed, for example, the addition of an amine to acrylonitrile. It is however possible to achieve the addition of ammonia to unactivated alkenes under pressure and by using catalysts. The most effective catalysts are acidic zeolites. The first successful demonstration was the formation of ethylamines from ethylene and ammonia with H-mordenite [25]. Although the conversion for the direct amination of alkenes is low, being limited by kinetic factors and the thermodynamic equilibrium, the selectivity for monoalkylamines can be excellent given the proper choice of zeolite and alkene. Small-pore zeolites are most suitable for ethylamines [26], while medium-pore zeolites can be used to prepare *tert*-butylamine [27]. Certain large-pore zeolites have been found to offer improvements over medium-pore zeolites [28–30].

Homogeneously catalyzed amination of alkenes is restricted to some special cases [31], [32]. Although not very selective, hydroformylation of alkenes in the presence of ammonia or amines [33] might find applications in future. So far it is hampered by the fact that in addition to the amines arising from the *n*-aldehyde and isoaldehyde, significant amounts of alcohols are formed besides dimerization products [34].

3.7. Other Processes

The hydrogenation of aromatic amines to cycloalkylamines is industrially restricted to the production of cyclohexylamines (see Chap. 5). A process for synthesizing amines from carbon monoxide, hydrogen, and ammonia or amines over an activated iron catalyst by a method of the Fischer – Tropsch-type was described [35], but this is not used on an industrial scale. All other synthetic routes described in the literature, for instance reduction of amides [36], are so far essentially laboratory methods that are suitable for the preparation of small quantities and are carried out on an industrial scale only in isolated cases. For details see [3].

4. Lower Alkylamines

4.1. Physical Properties

Ethylamine is gaseous at room temperature; diethylamine, triethylamine, and higher amines up to about twelve carbon atoms per alkyl chain are liquid, and long-chain amines containing still higher alkyl groups are solid.

The short-chain amines are readily soluble in water, alcohol, ether, and conventional organic solvents. Amines containing more than five carbon atoms are only partially or very sparingly soluble in water. The solubility in water generally decreases with increasing temperature, and in some cases the miscibility gap vanishes at lower temperatures. For example, triethylamine is completely miscible with water below about 18 °C but is only partially miscible with water above this temperature. Therefore, industrial phase separations (amine/water) are often carried out at elevated temperatures.

All alkylamines, but particularly the lower members, which have high vapor pressures, have a characteristic ammonia odor. The odor decreases with increasing substitution.

Physical properties of importance for characterizing and handling the alkylamines are summarized in Table 1. For further data, reference may be made to the relevant tables and the producers' data sheets.

4.2. Storage and Transportation

The aliphatic amines are usually stored in carbon steel or stainless steel containers; relatively small amounts can be kept in glass or ceramic vessels. Nonferrous metals such as copper, aluminum, zinc, lead, and their alloys (e.g., brass) are not resistant to amines. Aqueous amine solutions are basic and slowly etch glass. Ethylamine must be stored under pressure because it boils below room temperature. Isopropylamine (*bp* 32.4 °C) is usually stored in cooled and isolated tanks or in pressurized containers. To facilitate storage and transport, these amines are also produced and distributed in the form of aqueous solutions (e.g., 40 % or 70 %). The aliphatic amines have a virtually unlimited shelf life, but they should be stored under nitrogen to avoid contact with carbon dioxide (resulting in formation of carbamates) and atmospheric moisture.

All of the amines discussed here are inflammable. Some of them have very low flash points (see Table 1), and they form explosive mixtures with air. Hence, when these compounds are stored and transported, the relevant statutory regulations for inflammable liquids or pressurized gases must be observed. Because these vary greatly from country to country, they cannot be discussed in detail here.

To prevent odor nuisance (odor thresholds are below 0.1 ppm in some cases) and impermissible emissions, special precautions must be taken during storage and, in particular, during transfer from one container to another. Where the process cannot be carried out in a closed system and relatively large amounts are handled, absorption plants for washing off-gas must be provided. Scrubbing with water is frequently adequate, but better treatment of off-gas is achieved by scrubbing with an acid. For smaller plants, adsorption on active carbon filters is advisable under certain circumstances. This method has the additional advantage that the adsorbed amines can be recovered.

The toxicity (Chap. 10) of these amines also makes it necessary to take certain precautions with regard to occupational safety. These include, in particular, avoiding inhalation of amine vapors by using respiratory equipment and reliable prevention of skin and eye contact with fluid amines by wearing protective clothing.

4.3. Quality Specifications and Analysis

In the case of individual amines, the quality specifications are very high. Particularly for amines with a wide range of uses (e.g., ethylamine), specifications require a purity of above 99.7 %, although usually 98.0 – 99.5 % purity is considered adequate. Some end uses impose additional maximum values on the content of water, ammonia, and unsaturated amines or, in the case of tertiary amines, on the content of primary and secondary amines.

The amine content is determined by acid titration of the basic nitrogen, and the purity is determined preferably by gas chromatography. Also of importance is the determination of the water content by the Karl Fischer method, titration of the imine content by oximation, and the determination of physical properties.

4.4. Production

With few exceptions, aliphatic amines of industrial importance are produced by any one of the methods described in Chapter 3. The main routes start from the corresponding hydroxyl or carbonyl compounds. The choice of process depends mainly on the following criteria:

1) Price of the starting materials
2) Availability of the starting materials
3) Conversion of any existing plants to allow for different pressures in the two processes
4) Availability of hydrogen (hydrogen consumption and pressure differ in the two processes)
5) Fiscal reasons (taxes or subsidies on particular materials, e.g., ethanol)

Because these factors vary greatly from country to country, and in some cases even from region to region, and are changing continually, the

Table 1. Physical properties of commercial alkylamines

Compound	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}	Flash point, °C	pK_b (25 °C)
Ethylamine	[*75-04-7*]	$C_2H_5NH_2$	45.09	−80.6	16.6	0.6829	1.3663	−52	3.25
Diethylamine	[*109-89-7*]	$(C_2H_5)_2NH$	73.14	−50	56.3	0.7056	1.3864	−23	2.88
Triethylamine	[*121-44-8*]	$(C_2H_5)_3N$	101.19	−115	89.3	0.7275	1.4010	−11	3.24
Propylamine	[*107-10-8*]	$C_3H_7NH_2$	59.11	−83	47.8	0.7173	1.3870	−30	3.41
Dipropylamine	[*142-84-7*]	$(C_3H_7)_2NH$	101.19	−63	109.2	0.7400	1.4050	7	3.09
Tripropylamine	[*102-69-2*]	$(C_3H_7)_3N$	143.26	−93.5	156	0.7558	1.4171	36	3.35
Isopropylamine	[*75-31-0*]	$(CH_3)_2CHNH_2$	59.11	−95.2	32.4	0.6886	1.3742	−37	3.37
Diisopropylamine	[*108-18-9*]	$[(CH_3)_2CH]_2NH$	101.19	−61	84	0.7169	1.3924	−17	3.43 (20 °C)
Butylamine	[*109-73-9*]	$C_4H_9NH_2$	73.14	−49.1	77.8	0.7414	1.4031	−8	3.39
Dibutylamine	[*111-92-2*]	$(C_4H_9)_2NH$	129.25	−62	159	0.7670	1.4177	39	3.04
Tributylamine	[*102-82-9*]	$(C_4H_9)_3N$	185.36	−70	213	0.7771	1.4291	70	3.11
Isobutylamine	[*78-81-9*]	$(CH_3)_2CHCH_2NH_2$	73.14	−85.5	68	0.7360	1.3988	−16	3.59
Diisobutylamine	[*110-96-3*]	$[(CH_3)_2CHCH_2]_2NH$	129.25	−77	139	0.7450	1.4090	25.5	3.18
Triisobutylamine	[*1116-40-1*]	$[(CH_3)_2CHCH_2]_3N$	185.36	−21.8	191.5	0.7684	1.4252		3.68
1-Methylpropylamine (D, L)	[*13952-84-6*]	$C_2H_5CH(CH_3)NH_2$	73.14	−10.4	63.5	0.7246	1.3932	−20	3.43
Bis(1-methyl)propylamine (D, L)	[*626-23-3*]	$[C_2H_5CH(CH_3)]_2NH$	129.25	−70	135	0.7534	1.4111	21	
1,1-Dimethylethylamine	[*75-64-9*]	$(CH_3)_3CNH_2$	73.14	−67.5	44.4	0.6958	1.3784	−9	3.39
Pentylamine	[*110-58-7*]	$C_5H_{11}NH_2$	87.16	−55	104.4	0.7547	1.4118	7	3.36
Dipentylamine	[*2050-92-2*]	$(C_5H_{11})_2NH$	157.30	−32	202	0.7771	1.4272	71	2.82
Tripentylamine	[*621-77-2*]	$(C_5H_{11})_3N$	227.44		240–5	0.7907	1.4366	88	
2-Pentylamine	[*625-30-9*]	$C_3H_7CH(CH_3)NH_2$	87.16		91.5	0.7384	1.4027		
3-Pentylamine	[*616-24-0*]	$(C_2H_5)_2CHNH_2$	87.16		91	0.7479	1.4063		
2-Methylbutylamine	[*96-15-1*]	$C_2H_5CH(CH_3)CH_2NH_2$	87.16		95.4	0.7490	1.4114	3	
3-Methylbutylamine	[*107-85-7*]	$(CH_3)_2CH(CH_2)_2NH_2$	87.16	−60	95	0.7505	1.4083	4	3.98
Bis(3-methylbutyl)amine	[*544-00-3*]	$[(CH_3)_2CH(CH_2)_2]_2NH$	157.30	−44	188	0.7669	1.4235		
Tris(3-methylbutyl)amine	[*645-41-0*]	$[(CH_3)_2CH(CH_2)_2]_3N$	227.44		235	0.7848	1.4331	82	
Hexylamine	[*111-26-2*]	$C_6H_{13}NH_2$	101.19	−19	130	0.7660	1.4180	34	3.44

Table 1. (Continued)

Compound	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}	Flash point, °C	pK_b (25 °C)
Octylamine	[*111-86-4*]	$C_8H_{17}NH_2$	129.25	0	179.6	0.7826	1.4294	58	3.35
2-Ethylhexylamine	[*104-75-6*]	$C_4H_9CH(C_2H_5)CH_2NH_2$	129.25	<−70	169.2	0.7894	1.4313	53	
Decylamine	[*2016-57-1*]	$C_{10}H_{21}NH_2$	157.30	17	220.5	0.7936	1.4369		3.36
N-Methylbutylamine	[*110-68-9*]	$C_4H_9NHCH_3$	87.16	−75	91	0.7341	1.4010	0	6
N-Ethylbutylamine	[*13360-63-9*]	$C_4H_9NHC_2H_5$	101.19	−70	108–9	0.7398	1.4040	9	
N,N-Dimethylethylamine	[*598-56-1*]	$C_2H_5N(CH_3)_2$	73.14	<−70	36	0.6754	1.3720	−45	
N,N-Dimethylisopropylamine	[*996-35-0*]	$(CH_3)_2CHN(CH_3)_2$	87.16	−136	66	0.7151	1.3905	−27	
N-Ethyldiisopropylamine	[*7087-68-5*]	$[(CH_3)_2CH]_2NC_2H_5$	129.25	<−50	127	0.751	1.4140	10	
Allylamine	[*107-11-9*]	$CH_2{=}CH\text{-}CH_2\text{-}NH_2$	57.09	−88	53.2	0.7621	1.4205	−20	4.28
Diallylamine	[*124-02-7*]	$(CH_2{=}CH\text{-}CH_2)_2NH$	97.16	−88	111	0.7874	1.4387	21	4.87
N,N-Dimethyl-2-chloroethylamine	[*107-99-3*]	$ClCH_2CH_2N(CH_3)_2$	107.58		110		1.4286		
N,N-Diethyl-2-chloroethylamine	[*100-35-6*]	$ClCH_2CH_2N(C_2H_5)_2$	135.64		146–7	0.9166	1.4358		
3-Methoxypropylamine	[*5332-73-0*]	$CH_3O(CH_2)_3NH_2$	89.14		116–9	0.8727	1.4191	48	

reasons for the choice of a particular process depend strongly on individual circumstances. Currently (2000) the commercial use of the alcohol route is prevalent.

The aliphatic amines of large-scale industrial importance (mainly C_2-C_4) can be produced in a multipurpose plant that can be used for the production of various amines under similar reaction conditions, with only small differences in workup.

The reaction of an alkyl halide with ammonia or an amine typically results in formation of sodium halide as an undesired byproduct. This synthetic route is used for ethylenediamine and special amines such as allylamines (Chapter 3.4). Ethylenediamine can also be produced from ethanolamine and ammonia or directly from ethylene oxide.

Other production processes have become important only for a few amines. Examples are the production of *tert*-butylamine from isobutene and hydrogen cyanide or, since 1987, by direct amination [37] (Section 3.6).

Another industrial route to alkylamines is the addition of amines to acrylonitrile followed by hydrogenation. Examples are 1,3-diaminopropane and dipropylenetriamine, produced by combination of ammonia and acrylonitrile and subsequent hydrogenation, and the hydrogenation of 3-alkoxypropionitriles, which are produced by addition of alcohols to acrylonitrile to give 3-alkoxypropylamines (Section 3.3).

When the conversion of alcohol to amine is performed in the liquid phase under high pressure (typically 20 MPa) the plant can consist of a synthesis unit with a cylindrical reactor, heat exchangers, high- and medium-pressure (3 MPa) separators and a distillation unit. The distillation towers separate ammonia/low-boiling alkylamine, water, and high-boiling residues and remove the monoalkylamine from di- and trialkylamine. Ammonia/low-boiling alkylamine is generally used in excess and recycled to the synthesis reactor. Water is typically isolated by phase separation and fed to a wastewater treatment system. If the reaction mixture does not meet market requirements, the excess amines can be returned to synthesis stage since the mono-, di-, and trialkylamines are in thermodynamic equilibrium (see Section 3.1).

Reactions in the gas phase are typically carried out in low-pressure reactor systems which

require only one gas – liquid separator (3 MPa) and can be set up with cheaper equipment. Figure 1 shows a simplified flow diagram of the catalytic gas-phase reaction of ethanol with ammonia. In this case all three ethylamines are separated in a specially designed distillation unit and marketed.

To construct a multipurpose plant, it must first be decided whether the synthesis is to be run at low to medium (0.1 – 3 MPa) or high pressure (5 – 30 MPa). The design of the distillation unit depends on the range of amines to be produced. For the whole range of lower alkylamines (starting with ethylamines) a complicated system of five to six distillation towers is needed. With three distillation towers, a plant for a restricted portfolio of alkylamines can be run. The first pressure tower (typically 1.5 – 1.8 MPa) separates ammonia/low-boiling amine, the second water, and the third the desired alkylamine. The excess amines are reycled to the synthesis unit. For instance, a monoalkylamine can be produced by feeding pure alcohol and ammonia or by feeding recycled di- and trialkylamines, fresh alcohol, and ammonia. Instead of ammonia low-boiling amines such as methylamine or dimethylamine can be used for the synthesis of mixed amines. An example is the production of *N,N*-dimethylethylamine from ethanol and dimethylamine.

4.5. Uses

Ethylamines. Of the amines discussed here, with a worldwide market volume (2000) of about 80 000 t/a, the ethylamines are the most important lower aliphatic alkylamines. For tax reasons, in Japan ethylamines are produced from acetaldehyde, while in other countries ethanol is mostly used. The general production processes are similar; differences lie in the catalyst used, the distillation conditions, and the recycle streams (e.g., [39]). Production of di- and triethylamine from acetonitrile is described in [17].

The principal use of *monoethylamine* is in the production of herbicides of the triazine type by reaction with cyanuric chloride. Important commercial products are atrazine (Novartis), ametryne (Novartis), cyanazine (ACC, DuPont), simazine (Novartis), terbutryn (Novartis), desmetryn (Novartis), terbumeton (Novartis), secbumeton (Novartis), simetryn (Novartis) and terbutylazine. Lately, the demand for atrazine has decreased due to environmental concerns, and it has been banned in certain western European countries. A triazine derivative has also been suggested as stabilizer for chlorine in swimming pools [40]. Ethylamine, via *N*-ethylcyclohexylamine, also ends up in the herbicide cycloate (see below). The fungicide cymoxanil (DuPont) is also made from ethylamine. The ethylamine demand for rubber applications, for instance, toluene ethylsulfonamide, is declining.

A large percentage of the total *diethylamine* production is used for *N,N*-diethylaminoethanol, which is formed by reaction with ethylene oxide and finds application as a corrosion inhibitor and in coatings. Diethylamine is also used for the production of vulcanization accelerators. Reaction with carbon disulfide gives a dithiocarbamate that can be oxidized to a thiuram disulfide. The most important commercial products are sodium and zinc diethyldithiocarbamate and tetraethylthiuram disulfide. Diethylamine is used for the production of certain agrochemicals, such as rice herbicides [e.g., Napropamide (Zeneca) and Thiobencarb (Kumiai)] or insecticides [e.g., Phosphamidon (Novartis, DuPont) and Pirimiphos-methyl (Zeneca)], pharmaceuticals, radical scavengers (diethylhydroxylamine), and insect repellents (*N,N*-diethyl-3-methylbenzamide, Autan).

Of the three ethylamines, *triethylamine* has the widest range of uses and is also the most expensive product owing to the unfavorable distribution of monoethylamine, diethylamine, and triethylamine in production. Much of the triethylamine is used as an organic acid acceptor in synthesis or as a salt former in precipitation and purification operations. Important examples of these uses are in the synthesis of semisynthetic penicillins and cephalosporins, and as solubilizers for herbicides in combination with 2,4-dichlorophenoxyacetic acid or 2,4,5-trichlorophenoxyacetic acid. Triethylammonium is the cationic component of the herbicide Garlon 3A (Dow Elanco). Other important fields of use include polyurethane catalysts (for example, the hardening of core sands in cold-box casting), curing catalyst for phenol – formaldehyde resins, anticorrosion

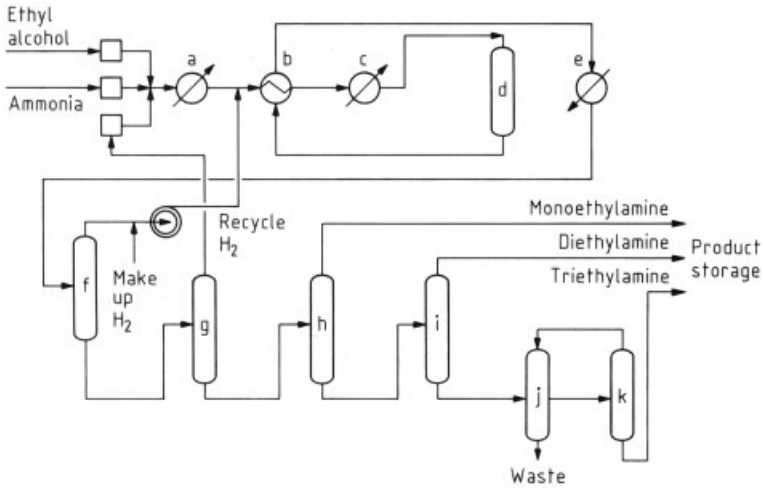

Figure 1. Continuous process for the production of ethylamines [38]
a) Vaporizer; b) Heat exchanger; c) Superheater; d) Catalytic converter; e) Product cooler; f) Gas separator; g) Ammonia column; h) Monoethylamine column; i) Diethylamine column; j) Decanter; k) Triethylamine column

agents, paper, textile and photographic auxiliaries, and anodic electrocoating.

Propylamines. The commercial demand for *n*-propylamines and isopropylamines is together about the same as that for the ethylamines. *Monoisopropylamine* is the most important product and is used particularly in the production of herbicides such as bentazon (BASF), Roundup (Monsanto), imazapyr (ACC) and the triazines ametryne (Novartis), atrazine (Novartis), desmetryn (Novartis), prometryn (Novartis), pramitol (Novartis), dipropetryn (Novartis), and propazine (Novartis, Makteshim-Agan and Drexel). Other applications encompass the nematicide fenamiphos (Bayer), the fungicide iprodione (Rhône-Poulenc), insecticides, pharmaceuticals, and surfactants.

Dipropylamine is required principally as a starting material for the synthesis of herbicides, e.g., trifluralin (Dow), oryzalin (Dow), EPTC (Zeneca), and vernolate (Zeneca). The other propylamines are of minor importance compared to monoisopropylamine and dipropylamine. *Diisopropylamine* is mostly used for agrochemicals such as the herbicides diallate and triallate (Monsanto), but also for rubber accelerators such as DIBS (*N,N*-diisopropyl-2-benzothiazylsulfenamide). *n-Propylamine* is used for agrochemicals and pharmaceuticals. *Tripropylamine* is an intermediate for pharmaceuticals and corrosion inhibitors.

Butylamines. *Diisobutylamine* is the principal starting material for the herbicide Butylate (Zeneca). *tert-Butylamine* is primarily employed for the synthesis of vulcanization accelerators, for instance, TBBS (*N-tert*-butyl-2-benzothiazylsulfenamide) and TBSI (*N-tert*-butyl-2-benzothiazylsulfenimide). *tert*-Butylamine is also used for the production of herbicides, i.e., terbacil (DuPont) and the triazines terbutylazine, terbumeton (Novartis), and terbutryn (Novartis) and insecticides (e.g., diafenthiuron from Novartis). A derivative has also been proposed as stabilizer for cosmetics [41]. The synthesis of vulcanization accelerators of the sulfenamide type is an important outlet for *dibutylamine* (e.g., sodium dibutylthiocarbamate, zinc dibutyldithiocarbamate or tetrabutylthiuram disulfide); the demand for this amine for other purposes (e.g., corrosion protection, flotation agents, cutting oils, plastics, and intermediates for insecticides (e.g., carbosulfan) and pharmaceuticals) is less important.

Monobutylamine is an intermediate for the production of plasticizers (e.g., the rubber accelerator dibutylthiourea), agrochemicals (e.g., DuPont's fungicide benomyl), pharmaceuticals (e.g., the antidiabetic tolbutamide), emulsifying agents, dyes, and UV absorbers.

Tributylamine is also used commercially (a few 100 t/a) as a polymerization catalyst and as an acid acceptor. *sec-Butylamine* is consumed in small amounts for herbicides such as bromacil (DuPont) or secbumeton (Novartis).

Other Amines. Of the C_{4+} monoamines, only few are of commercial importance. *Octylamine* is required, for example, for the production of biocides, as a solvent for pesticides, and as precursor for *N*-octylpyrrolidone – which is used for the cleaning of printing machines – whereas *2-ethylhexylamine* is employed for the production of oil additives and for insecticides. *Di-2-ethylhexylamine* is used for oil additives, dyes, and pharmaceutical applications. *Tri-2-ethylhexylamine* (*bp* 337.9 °C) is also commercially available and mainly used in coatings. *Tridecylamine* (mixture of isomers) is used for the production of the fungicide tridemorph (BASF). A mixture of the *pentylamines* finds use in drilling applications. *N,N-Dimethylethylamine* is an important epoxy and urethane catalyst, for instance, in sand-core applications. It is made from ethanol and dimethylamine. *N,N-Dimethylisopropylamine*, usually obtained from acetone and dimethylamine under hydrogenating conditions, is also used in foundry applications. *N-Ethyldiisopropylamine* (Hünig's base) finds use as a good acid acceptor. *Mono-*, *di-*, and *trihexylamine* are also commercially available, but represent very minor products. *Isoamylamine* (3-methyl-1-butaneamine) is an intermediate for pharmaceuticals and agrochemicals.

Some 10 000 t/a of tertiary alkyl primary amines are manufactured by Rohm & Haas by means of the Ritter reaction. These primene amines range from *tert*-octylamine to C_{18} – C_{22}.

Of the unsaturated amines, *diallylamine* has the most significant market (ca. 1000 t/a); it is used for the production of *N,N*-diallyldichloracetamide. Copolymers of dimethyldiallylammonium chloride with acrylates or acrylamides are used for water treatment. *N-Ethylmethallylamine* is employed for herbicide production (Dow's sonalan).

Although *chloroalkylamines* (mainly chloroethyl- and chloropropyl-*N,N*-dialkylamines) are not important in terms of the amounts employed, they are important starting materials for the production of a number of pharmaceuticals (principally neuroleptics).

Although not a lower alkylamine, *polyisobutenamine* (PIBA) is produced with the same technology by amination of the hydroformylation product of polyisobutene (PIB) with a molecular mass of about 1000 [42]. About 10 000 t/a PIBA is used as an detergent in finished additive packages for gasoline to improve the mechanical functioning of the engine, lower fuel consumption, and give improved emission characteristics.

4.6. Economic Aspects

Worldwide production capacity for aliphatic amines including fatty amines but excluding methylamines is estimated to be ca. 500 000 t/a, of which about 40 % is attributable to the United States and about 30 % to western Europe. Because of the large number of producers and the varying capacities of multipurpose plants, exact figures for annual worldwide production are not available. Apart from those who produce mainly for their own requirements, the most important producers of lower aliphatic amines are Air Products, Hoechst Celanese, and Elf Atochem in the United States; BASF, Elf Atochem, Hoechst Celanese, and ICI in Western Europe; Daicel in Japan; and Nordeste/Oxiteno in Brazil.

Manufacturing costs vary significantly depending on the availability of the starting materials and the type of process, e.g., continuous dedicated plant vs. batch process in a multipurpose plant.

5. Cycloalkylamines

Cyclopentylamine [*1003-03-8*], aminocyclopentane, $C_5H_{11}N$, M_r 85.15, is a colorless liquid, *mp* −85 °C, *bp* 106 – 108 °C (at 101.3 kPa), d_4^{20} 0.8689, n_D^{20} 1.4482, flash point 17 °C, with an ammoniacal odor.

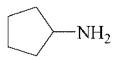

Production. Cyclopentylamine is produced from cyclopentanone and ammonia over nickel

catalysts at 20 MPa and 150 – 200 °C in the presence of circulating hydrogen over a fixed-bed catalyst.

Uses. Cyclopentylamine is used in small amounts as an intermediate for the potato and rice fungizide pencycuron (Bayer, Nihon Tokushu, Noyaku Seizo).

Cyclohexylamine [*108-91-8*], aminocyclohexane, cyclohexanamine, hexahydroaniline, hexahydrobenzenamine, aminohexahydrobenzene, $C_6H_{13}N$, M_r 99.18, is a colorless liquid, *mp* − 17.8 °C, *bp* 134.5 °C (at 101.3 kPa), d_4^{25} 0.8647, n_D^{20} 1.4592.

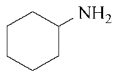

Viscosity

t, °C	0	20	50	75
η, Pa · s	3.73	2.10	1.14	0.77

Vapor pressure

t, °C	20	60	100	120
p, kPa	1.43	8.66	35.99	65.31

Specific heat capacity

t, °C	20	70	145
c, J g^{-1}K^{-1}	2.366	2.583	2.910

Heat of vaporization	399.86 J/g
Flash point (closed cups)	26.5 °C
Ignition temperature (class T3)	265 °C
Ignition range in air	1.6 – 9.4 vol %

Cyclohexylamine is infinitely miscible with water and conventional organic solvents. With water it forms an azeotrope that contains 44.2 % cyclohexylamine and boils at 96.4 °C.

Chemical Properties. Chemically cyclohexylamine has much in common with the acyclic aliphatic amines [3], [43]. Cyclohexylamine reacts with chlorine to form *N,N*-dichlorocyclohexylamine [44]. For conversion to azomethines and their chemistry, see [45]. *N*-Cyclohexylidenecyclohexylamine reacts with chloramine to give 1-cyclohexyl-3,3-pentamethylenediaziridine, which can be hydrolyzed to give cyclohexylhydrazine [46]. Cyclohexylamine and formaldehyde together react with peracetic acid to give 2-cyclohexyloxaziridine [47]. In addition to using alkyl halides, alkyl sulfates, or alkyl phosphates, cyclohexylamine can be alkylated with an alcohol in the presence of a catalyst, such as aluminum oxide, copper, nickel, cobalt, or platinum (reviewed in [48], [49]), or by the Leuckart – Wallach method [50].

Production. Cyclohexylamine can be produced by catalytic hydrogenation of aniline under pressure. High yields are obtained with nickel or cobalt catalysts treated with basic oxides [51]. Raney cobalt treated with calcium oxide and sodium carbonate gives a cyclohexylamine yield of more than 96 % at 6 MPa and 230 °C [52]. Noble metals also are employed as catalysts [53].

The alternative synthesis from cyclohexanol and ammonia over supported cobalt catalysts [54] is also an important method; calcium silicoaluminates are also suitable catalysts [55]. In this process, cyclohexanol reacts at 20 MPa and 220 °C with at least 3 mol of ammonia in the presence of circulating hydrogen over a fixed-bed catalyst. Using the same continuous procedure, cyclohexanone can be hydrogenated with ammonia under amination conditions [56]; nickel or cobalt catalysts are used for this process, which is carried out at 0.1 – 20 MPa. Cyclohexylamine can also be obtained by reacting phenol with hydrogen and ammonia over a rhodium catalyst [57].

Quality. The purity of the commercial product is above 99.5 % and is determined by gas chromatography; contaminants are ammonia and water.

Storage and Transportation. Cyclohexylamine can be stored and shipped in iron tanks. Nonferrous metals, particularly copper-containing materials, are attacked and are therefore unsuitable. The amine discolors on contact with air and therefore must be kept under nitrogen.

Uses. Cyclohexylamine is used primarily as corrosion inhibitor and vulcanization accelerator. Alone or mixed with other compounds, it has an anticorrosive action [58], for example, when used as an additive in heating oil or in the operation of steam boilers. It is also employed as a salt-forming and amide-forming component in many applications, for example, in the above mentioned vulcanization accelerators (CBS, *N*-cyclohexyl-2-benzothiazylsulfenamide [59]), plasticizers [salts with dodecanethiol and mer-

captobenzothiazole (CMBT)] [60]], emulsifiers (emulsification of active ingredients in water with cyclohexylammonium alkylbenzenesulfonate [61]) and coagulators (salt with acetic acid). Salts of $C_{10}-C_{14}$ fatty acids prevent foam formation in mineral oils [62]. Cyclohexylamine functions as a hardener for epoxy resins and as a catalyst for polyurethanes. Sodium cyclohexylsulfamate and calcium cyclohexylsulfamate (cyclamates) are important artificial sweeteners. In polyamide polymerizations, cyclohexylamine is employed as chain terminator to control the molecular mass. Other applications of cyclohexylamine are for instance in herbicide production. An example is DuPont's Hexazinone [3-cyclohexyl-6-(dimethylamino)-1-methyl-1,3,5-triazine-2,4-(1*H*,3*H*)-dione]. The total market for cyclohexylamine is around 20 000 t/a. Manufacturers include Air Products, BASF, Bayer, and Hoechst Celanese.

***N*-Methylcyclohexylamine** [*100-60-7*], $C_7H_{15}N$, M_r 113.2, is a colorless liquid, *mp* −9 °C, *bp* 150 °C (at 101.3 kPa), d_4^{25} 0.8533, n_D^{20} 1.4551, η 5 mPa · s (at 25 °C).

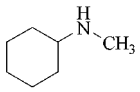

Vapor pressure

t, °C	20	40	60	80	100	120	140
p, kPa	0.48	1.72	4.68	11.48	24.13	46.26	81.31

Flash point	36.1 °C
Ignition temperature (class T3)	255 °C
Explosion limits in air	2.2 and 10.5 vol % (corresponding to 103 and 494 g/m^3)
Specific heat capacity	2.14 J g^{-1} K^{-1} (at 20 °C) 2.47 J g^{-1} K^{-1} (at 140 °C)

N-Methylcyclohexylamine is infinitely miscible with water, methanol, acetone, toluene, and cyclohexane.

Production. *N*-Methylcyclohexylamine can be prepared by a procedure similar to that used for cyclohexylamine, i.e., by hydrogenation of methylaniline over a supported nickel catalyst [63] or from cyclohexanone and methylamine under hydrogenation conditions [64]. Cyclohexylamine reacts with methanol over copper, zinc, or copper–calcium catalysts [65].

Uses. *N*-Methylcyclohexylamine is used as a component of vulcanization accelerators [66], but the market size is relatively small.

***N,N*-Dimethylcyclohexylamine** [*98-94-2*], cyclohexyldimethylamine, $C_8H_{17}N$, M_r 127.23, is a colorless liquid, *mp* <–50 °C, *bp* 159 °C (at 101.3 kPa), d_4^{25} 0.8467, n_D^{20} 1.4519, η 3 mPa · s (at 25 °C).

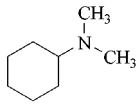

Vapor pressure

t, °C	20	40	60	80	100	120	140
p, kPa	0.29	1.05	3.21	8.29	18.13	35.99	63.72

Flash point	38.1 °C
Ignition temperature (class T4)	200 °C
Explosion limits in air	0.79 and 7.0 vol % (corresponding to 41.8 and 370 g/m^3)
Heat of vaporization	285.6 J/g
Specific heat capacity	1.88 J g^{-1}K^{-1} (at 20 °C) 2.51 J g^{-1}K^{-1} (at 140 °C)

The compound is sparingly soluble (1 %) in water, but the solubility of water in dimethylcyclohexylamine is ca. 20 %. *N,N*-Dimethylcyclohexylamine is miscible with conventional organic solvents.

Production. *N,N*-Dimethylcyclohexylamine is produced by hydrogenation of dimethylaniline at 180 °C and 6 MPa [67] or by amination of cyclohexanone under hydrogenation conditions in the presence of dimethylamine [68]. In principle, it could also be obtained from phenol by reaction with dimethylamine, but this route is commercially less attractive.

Uses. *N,N*-Dimethylcyclohexylamine is a catalyst for the production of polyurethane, particularly for making foams, and for this purpose it is used as the free base [69] or as the salt of an organic acid [70]. The curing temperature of baking finishes comprising polyurethane-forming substances can be reduced by 50–80 °C by adding weakly acidic derivatives of the amine [71]. Like pyridine, dimethylcyclohexylamine catalyzes certain reactions and is slightly more efficient than pyridine in the preparation of acid chlorides with thionyl

chloride [72]. It can be used as corrosion inhibitor and as an antioxidant in fuel oils.

***N*-Ethylcyclohexylamine** [*5459-93-8*], $C_8H_{17}N$, M_r 127.2, is a colorless liquid, *mp* −43 °C, *bp* 165 °C (at 101.3 kPa), d_4^{20} 0.846, n_D^{20} 1.4525, η 1.39 mPa · s (at 20 °C).

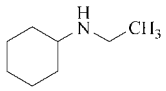

Vapor pressure

t, °C	20	40	80	120
p, kPa	0.24	0.84	6.13	27.59

Flash point	46 °C
Ignition temperature (class T3)	245 °C
Explosion limit in air	1.2 and 7.6 vol %

N-Ethylcyclohexylamine is only slightly soluble (2.3 %) in water, but the solubility of water in ethylcyclohexylamine is 25 %. The amine is miscible with conventional organic solvents.

Ethylcyclohexylamine is produced by hydrogenation of ethylaniline at 180 °C and 6 MPa [67] or by amination of cyclohexanone under hydrogenation conditions in the presence of ethylamine [68]. The amine is a starting material for the preplant sugar beet, grass, and spinach herbicide cycloate (Zeneca's Ro-Neet), which is obtained by converting *N*-ethylcyclohexylamine into *S*-ethyl-*N*-ethyl-*N*-cyclohexylthiocarbamate. A minor use of *N*-ethylcyclohexylamine is for vulcanization accelerators.

Dicyclohexylamine [*101-83-7*], dodecahydrodiphenylamine, *N,N*-dicyclohexylamine, *N*-cyclohexylcyclohexanamine, perhydrodiphenylamine, $C_{12}H_{23}N$, M_r 181.32, is a colorless liquid, *mp* −0.1 °C, *bp* 256 °C (at 101.3 kPa), *bp* 133 °C (at 2.6 kPa), d_4^{25} 0.9104, n_D^{20} 1.4852.

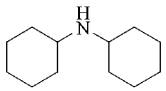

Flash point	105 °C
Ignition temperature (class T3)	240 °C
Explosion limits in air	0.83 and 4.6 vol % (corresponding to 62 and 350 g/m^3)
Specific heat capacity	1.88 J g^{-1} K^{-1} (at 20 °C) 2.47 J g^{-1} K^{-1} (at 140 °C)

Dicyclohexylamine is sparingly soluble in water (ca. 0.16 % at 28 °C) but is readily soluble in the conventional solvents.

Production. Dicyclohexylamine can be formed as a byproduct in the production of cyclohexylamine from aniline or cyclohexanol/cyclohexanone [73]. When desired as the end product, dicyclohexylamine is obtained by amination of cyclohexanone under hydrogenation conditions [74] or, in higher yield, from cyclohexanone and cyclohexylamine over Pd/C at a hydrogen pressure of 0.4 MPa [75]. Dicyclohexylamine can also be obtained by amination of phenol under hydrogenation conditions, or by the reaction of phenol with aniline over Pd/C at 0.5 MPa under hydrogenation conditions [76].

Uses. Dicyclohexylamine and its salts have good anticorrosion properties in the vapor phase [77]. Paper treated with the nitrite is used as anticorrosion paper. The nitrite also imparts these properties when incorporated into alkyd resins [78]. Like the other cyclohexylamines, dicyclohexylamine is useful as a component of vulcanization accelerators (for instance DCBS, *N,N*-dicyclohexyl-2-benzothiazylsulfenamide, or DCMBT, the salt with 2-mercaptobenzothiazole) and of pesticides. It is also consumed as a processing chemical for antibiotics and as fuel oil additive.

***N*-Methyldicyclohexylamine** [*7560-83-0*], *N,N*-dicyclohexylmethylamine, *N*-cyclohexyl-*N*-methyl-cyclohexanamine, $C_{13}H_{25}N$, M_r 195.35, is a liquid, *bp* 275 °C (at 101.3 kPa), d_4^{25} 0.9207, n_D^{25} 1.4881, η 10.2 mPa · s (at 25 °C), flash point 66 °C.

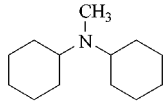

Vapor pressure

t, °C	63	90	150	190	248
p, kPa	0.133	0.599	2.66	13.33	66.65

N-Methyldicyclohexylamine is sparingly soluble in water and THF but readily soluble in acetone, methanol, toluene, cyclohexane, and chloroform. It is prepared by methylating dicyclohexylamine with dimethyl sulfate or by the Leuckart – Wallach reaction [79] and has uses similar to those of the other cyclohexylamines, e.g., as a urethane foam catalyst.

4,4′-Diaminodicyclohexylmethane [*1761-71-3*], methylene bis(4,4′-cyclohexylamine), bis(4-aminocyclohexyl)methane, 4,4′-methylenebiscyclohexanamine, $C_{13}H_{26}N_2$, M_r 210.36, *mp* 60 – 62 °C (cis,cis), 64 – 65 °C (trans,trans), *bp* 330 – 331 °C (at 101.3 kPa), d_4^{20} 0.96, flash point 153.5 °C, is a water insoluble solid.

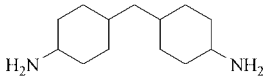

The large-volume product 4,4′-diaminodicyclohexylmethane is manufactured by several companies by the catalytic hydrogenation of methylene dianiline. The isomer distribution in the yield is controlled, since the main product from the amine is the isocyanate, which is used in urethane coatings. Other applications include fibers, pigments, films, and epoxy curing agents.

2,2-Bis(4-aminocyclohexyl)propane [*3377-24-0*], isopropylidenedi(cyclohexylamine), $C_{15}H_{30}N_2$, M_r 238.42, *mp* 51 – 53 °C, *bp* 155 – 160 °C (at 0.65 kPa), d_4^{20} 0.99.

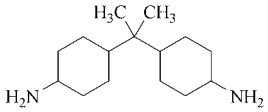

2,2-Bis(4-aminocyclohexyl)propane can be obtained by amination of the corresponding bishydroxy compound in the liquid phase under pressure. Product yields of ca. 90 % are obtained over hydrogenation catalysts containing palladium [57], ruthenium [80] or cobalt, manganese, and phosphoric acid [81]. Alternatively, hydrogenation of the condensation product of aniline and acetone, bisaniline A, furnishes the same compound. The industrial product is a mixture of three stereoisomers. Concentration of the trans,trans isomer by treating the isomer mixture with phosphoric acid under reduced pressure at 230 °C or by recrystallization from branched octanes [82] is important because polymerization of the pure isomer gives polyamides of better quality.

Cyclooctylamine [*5452-37-9*], $C_8H_{17}N$, M_r 127.23, is a liquid, *mp* – 48 °C, *bp* 80 °C (at 1.3 kPa), *mp* of hydrochloride 244 – 245 °C, *mp* of picrate 193 – 194 °C, d_4^{20} 0.9280, n_D^{20} 1.4804, flash point 62 °C.

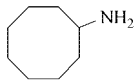

The amine can be produced by the Ritter reaction of cyclooctanol with hydrocyanic acid [83], by reduction of cyclooctanone oxime with sodium and an alcohol [84], by Ritter reaction of cyclooctene with hydrocyanic acid [85], or by catalytic hydrogenation of cyclooctanone under amination conditions [86]. Cyclooctylamine is used as a component of urea herbicides, e.g., cycluron (BASF). Its chemical behavior is analogous to that of other primary amines.

Cyclododecylamine [*1502-03-0*], $C_{12}H_{25}N$, M_r 183.34, *mp* 31 – 32 °C, *bp* 95 – 96 °C (at 0.25 kPa), $n_D^{32.5}$ 1.4849; *mp* of hydrochloride 268 – 270 °C, *mp* of picrate 232 – 234 °C, flash point 121 °C.

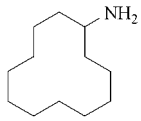

Cyclododecylamine is produced by reduction of cyclododecanone oxime with sodium and an alcohol [87], by catalytic reduction of nitrocyclododecane with hydrogen [88], by Ritter reaction of cyclododecene with hydrocyanic acid [89], or by hydrogenation of cyclododecanone under amination conditions [90]. The reaction of cyclododecylamine with propylene oxide leads to a product that can undergo cyclization to give

the fungicide Dodemorph (BASF). The compound behaves chemically as a primary amine.

6. Cyclic Amines

Pyrrolidine [*123-75-1*], tetrahydropyrrole, C_4H_9N, M_r 71.12, is a colorless liquid, *mp* −60 °C, *bp* 87–88 °C (at 101.3 kPa), d_4^{20} 0.8576, n_D^{20} 1.4428, η 1.1 mPa · s (at 20 °C).

Flash point	3 °C
Ignition temperature (class T2)	345 °C
Explosion limits in air	1.6 and 10.6 vol % (corresponding to 47 and 314 g/m^3)
Specific heat capacity	2.20 J g^{-1} K^{-1} (at 21 °C) 2.25 J g^{-1} K^{-1} (at 65 °C) 2.27 J g^{-1} K^{-1} (at 77 °C)
Heat of vaporization	529 J/g (at 0.1 MPa)

Pyrrolidine is infinitely miscible with water and conventional organic solvents such as methanol, acetone, ether, and chloroform. It acts as a lachrymator.

Chemically, pyrrolidine behaves like a secondary amine in every respect. For example, it undergoes Leuckart–Wallach and Mannich reactions and is readily converted into an enamine. In the presence of a catalyst, such as platinum at 360 °C or rhodium at 650 °C [91], pyrrole is formed. In the presence of a copper catalyst, *N*-methylpyrrolidone is converted into *N*-methylpyrrolidine.

Production. Pyrrolidine can be produced from butanediol and ammonia, e.g., over an aluminum thorium oxide catalyst at 300 °C [92] or over a nickel catalyst at 200 °C and 20 MPa under hydrogenation conditions. It can also be produced from THF and ammonia over aluminum oxide at 275–375 °C [93].

Uses. Pyrrolidine is used virtually exclusively for the synthesis of pharmaceuticals, including antibiotics. In addition, it is used for producing vulcanization accelerators.

***N*-Methylpyrrolidine** [*120-94-5*], 1-methylpyrrolidine, $C_5H_{11}N$, M_r 85.15, colorless liquid; *mp* −90 °C, *bp* 80–81 °C (at 101.3 kPa), d_4^{20} 0.7992, n_D^{20} 1.4236, flash point −21 °C.

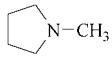

Piperidine [*110-89-4*], hexahydropyridine, pentamethyleneimine, azacyclohexane, cyclopentimine, cypentil, hexazane, $C_5H_{11}N$, M_r 85.15, derives its name from the alkaloid piperine, the pepper flavoring, in which piperidine is present as piperic acid piperidide.

Piperidine, a colorless liquid, *mp* −10.5 °C, *bp* 106.4 °C (at 101.3 kPa), with a sharp ammonia odor, is hygroscopic and fumes in air; d_4^{20} 0.8613 (the temperature dependence of the density is given in [94]), n_D^{20} 1.4532.

Vapor pressure

t, °C	13.0	25.0	40.0	51.0	67.8	78.8	87.2
p, kPa	2.0	4.0	8.0	13.3	26.7	40.0	53.3

The vapor pressure curve is given in [95]; flash point 3 °C, dissociation constant 1.6×10^{-3} (at 25 °C). Piperidine is infinitely soluble in water, lower alcohols and ketones, ethers, aliphatic and aromatic hydrocarbons, ethyl acetate, and DMF. It forms an azeotrope with water:

p, kPa	100.8	53.3	26.7
bp, °C	93.7	76.9	61.4
Piperidine, mass fraction in %	63.7	68.1	76.0

The preparation of enamines from piperidine takes place particularly smoothly [96], [97], [98]. Unlike open-chain amines, cyclic amines react with aldehydes to give aminals as intermediates [99]. Only pyrrolidine is comparably versatile.

Production. Pyridine obtained from tar distillation is contaminated by sulfur-containing compounds and therefore can be hydrogenated economically only in a two-stage process. The crude pyridine is first hydroraffined over a sulfidic metal catalyst at 280–310 °C and then hydrogenated quantitatively to piperidine at

120 – 160 °C over a Ni – Al_2O_3 catalyst [100]. Sulfur-free pyridine can be hydrogenated over a ruthenium [101] or cobalt catalyst.

The preparation of piperidine by aminolysis of 1,5-pentanediol under hydrogenation conditions at 225 – 260 °C and 20 MPa over a cobalt catalyst has been described [102], but can also be affected with nickel catalysts. A similar preparation from tetrahydrofurfuryl alcohol also has been described [103]. When piperidine is prepared by hydrogenating glutarates [104], glutaric acid [105], or glutaraldehyde [106], the reaction is carried out in the presence of a large excess of ammonia.

Uses. The secondary amine piperidine is highly reactive and is therefore frequently employed as an intermediate for pharmaceuticals [107] and for plant protection agents. It is also used as a vulcanization accelerator in rubber manufacture and as an oil or fuel additive [108].

Piperidine and, in many cases, piperidine acetate are useful catalysts for condensation reactions, e.g., the Knoevenagel reaction [109], aldol condensation [110], and the condensation of a nitroparaffin with an aldehyde [111]. However, for the last of these reactions, diethylamine is the preferred catalyst. The use of piperidine is particularly advisable where the reactants or products are unstable in the presence of stronger bases.

***N*-Methylpiperidine** [*626-67-5*], $C_6H_{13}N$, M_r 99.18, colorless liquid; *mp* – 18 °C, *bp* 106 – 107 °C (at 101.3 kPa), d_4^{20} 0.8157, n_D^{20} 1.4378, flash point 3 °C.

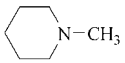

Uses. *N*-Methylpiperidine is used for the production of a growth regulator for cotton plants.

***N*-Ethylpiperidine** [*766-09-6*], $C_7H_{15}N$, M_r 113.20, colorless liquid; *bp* 131 °C (at 101.3 kPa), d_4^{20} 0.8237, n_D^{20} 1.4440, flash point 19 °C.

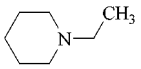

Uses. *N*-Ethylpiperidine has found use as a catalyst in polycarbonate manufacture.

Hexamethyleneimine [*111-49-9*], HMI, azacycloheptane, hexahydroazepine, homopiperidine, perhydroazepine, $C_6H_{13}N$, M_r 99.18, is a liquid, *mp* – 37 °C, *bp* 139 °C (at 101.3 kPa), d_4^{20} 0.8799, n_D^{20} 1.4658, flash point 30 °C, ignition temperature 255 °C.

Hexamethyleneimine and water form a 1/1 azeotrope that boils at 95 – 95.5 °C. The hexamethyleneimine – water solubility curve is given in [112]. The chemical behavior of hexamethyleneimine is determined entirely by the secondary amino group.

Production. Hexamethyleneimine is produced in 84 % yield by heating hexamethylenediamine at 350 °C in a stream of hydrogen. The catalyst, ammonium vanadate on activated alumina, is prereduced with hydrogen at 500 °C [113]. Residues from the industrial distillation of hexamethylenediamine can be converted into hexamethyleneimine over aluminum silicate or aluminum oxide in a stream of nitrogen [114]. For the preparation from caprolactam, see [115]. Hexamethyleneimine can be prepared by dimerizing acrolein, reducing the product to 2-hydroxymethyltetrahydropyran, expanding the ring to give oxepane, followed by treatment with ammonia over aluminum oxide at 350 °C [116].

Uses. The most important use is the conversion of hexamethyleneimine into *S*-ethylhexahydro-1*H*-azepine-1-carbothioate, the selective rice herbicide Molinate (C_2H_5S–CO–NC_6H_{12}, Zeneca). Minor amounts of hexamethyleneimine are used for the production of rubber vulcanization accelerators, resins, pharmaceuticals, and textile auxiliaries; numerous patents describe the anticorrosive action of hexamethyleneimine and its derivatives.

Morpholine [*110-91-8*], tetrahydro-2*H*-1,4-oxazine, diethylene oximide, diethylenimide oxide, 1-oxa-4-azacyclohexane, C_4H_9NO, M_r 87.12, is a colorless hygroscopic liquid; *mp*

$-5\,°C$, *bp* 128.2 °C (at 101.3 kPa), d_4^{20} 1.007, n_D^{20} 1.4542, η 2.3 mPa · s (at 20 °C).

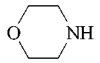

Vapor pressure

t, °C	10	20	40	60	80	100	120
p, kPa	0.57	1.1	3.2	8.3	20.5	40.9	81.8

Critical pressure	5.302 MPa
Critical temperature	344 °C
Flash point	43 °C
Ignition temperature	275 °C
Explosion limits in air	1.8 vol % and 15.2 vol %
Specific heat capacity	1.2 J g^{-1}K^{-1} (at 25 °C) 2.15 J g^{-1}K^{-1} (at 100 °C)
Heat of vaporization	505 J/g (at 25 °C) 425 J/g (at 128 °C)

Morpholine is infinitely miscible with water and conventional solvents such as methanol, acetone, benzene, and glycol. The mixture consisting of 88 % morpholine and 12 % water has a freezing point below $-50\,°C$ and a viscosity of 7.45 mPa · s at 20 °C.

Chemically, morpholine behaves as a secondary amine in every respect. It forms complexes with many cations. In addition to the Leuckart – Wallach and Mannich reactions, the conversion to enamines has become important [96], [97]. Morpholine, pyrrolidine, and piperidine are preferred to the open-chain amines for these reactions because they react faster with the carbonyl component.

Production. Morpholine can be obtained from diethanolamine by cyclization with sulfuric acid or via diethanolamine hydrochloride [117] or from di(2-chloroethyl) ether and ammonia [118]. Industrially more important is the amination of diglycol under hydrogen pressure and over a cobalt or nickel catalyst [42], [119]. Other catalysts are described in [120]. An important byproduct frequently obtained is 2-(2-aminoethoxy)ethanol, which is the result of simple amination without subsequent cyclization; after separation of morpholine, this compound can be recycled to the synthesis reactor.

Quality and Storage. When produced from diglycol, commercial morpholine is more than 99 % pure; the main impurities are *N*-ethylmorpholine and ethylenediamine. Morpholine can be stored for an unlimited time in iron containers if it is protected from atmospheric moisture and carbon dioxide. Copper, zinc, and their alloys are unstable towards morpholine.

Uses. Morpholine is an intermediate for a large number of pharmaceuticals, crop protection agents, dyes, and optical brighteners. An important use is the conversion to vulcanization accelerators or sulfur donors, which are based on 2-mercaptobenzothiazole, dithiocarbamic acid, thiuram polysulfides, or sulfenamides, e.g., MBS [2-(4-morpholinylmercapto)benzothiazole], DTDM (4,4′-dithiodimorpholine), and MBSS (4-morpholinyl-2-benzothiazyl disulfide). Salts of morpholine with long-chain fatty acids, such as oleic or stearic acid, have waxlike properties and are used as emulsifiers.

Because water and morpholine have similar volatilities and the latter is an anticorrosion agent, it is useful in steam cycles [121], aqueous hydraulic liquids [122], and similar systems. Up to 550 °C, no substantial loss occurs as a result of decomposition.

An aqueous solution of morpholine (2 M) can be used for removing CO_2, H_2S, or HCN from gases [123]. In the extraction of aromatic compounds from hydrocarbon mixtures, morpholine is highly selective [124].

Morpholine Derivatives. *N*-Methylmorpholine and *N*-ethylmorpholine are used as catalysts for the production of polyurethane foams. *N*-Formylmorpholine is used as a selective solvent for the extraction of high-purity aromatic compounds in the Formex process (Snamprogetti).

N-Methylmorpholine [*109-02-4*], 4-methyl-1-oxa-4-azacyclohexane, $C_5H_{11}NO$, M_r 101.15, *mp* $-65\,°C$, *bp* 114 °C (at 101.3 kPa), d_4^{20} 0.919, n_D^{20} 1.4347, η 0.90 mPa · s (at 20 °C), flash point 13 °C.

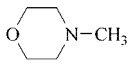

Methylmorpholine can be obtained from diethylene glycol and methylamine or by reacting morpholine with formaldehyde to give *N*-hydroxymethylmorpholine and reducing the product with excess formaldehyde [125], formic acid [126], or hydrogen over cobalt or manganese catalysts under pressure [127]. Another important reaction is the alkylation of morpho-

line with methanol at a total pressure above 3.5 MPa and a hydrogen partial pressure above 1 MPa [128].

Alkylmorpholines are also obtained by cyclization of bis(2-chloroethyl) ether with an amine [129] or of diethylene glycol with an amine in the presence of hydrogen and a hydrogenation catalyst [130].

N-Ethylmorpholine [*100-74-3*], $C_6H_{13}NO$, M_r 115.18, *mp* −63 °C, *bp* 134–137.5 °C (at 101.3 kPa), d_4^{20} 0.913, n_D^{20} 1.4415, η 1.05 mPa · s (at 20 °C), flash point 27 °C, is miscible with water.

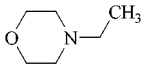

N-Formylmorpholine [*4394-85-8*], 4-morpholinecarbaldehyde, $C_5H_9NO_2$, M_r 115.13, *mp* 20–21 °C (anhydrous), 9 °C (5 % water), –3 °C (10 % water), *bp* 244 °C (anhydrous), 172 °C (2 % water), 151 °C (5 % water), 132 °C (8 % water) (all at 101.3 kPa), d_4^{20} 1.1268, n_D^{20} 1.4835, flash point 124 °C, is hygroscopic.

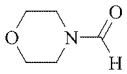

N-(2-Hydroxyethyl)morpholine [*622-40-2*], 4-(2-hydroxyethyl)morpholine, 4-morpholinoethanol, $C_6H_{13}NO_2$, M_r 131.17, *mp* −1 °C, *bp* 227 °C (at 100.9 kPa), d_4^{20} 1.0733, n_D^{20} 1.4760, flash point 99 °C, is used as an intermediate for pharmaceuticals.

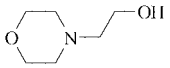

N-(2-Aminoethyl)morpholine [*2038-03-1*], 2-morpholinoethylamine, 4-(2-aminoethyl)-morpholine, $C_6H_{14}N_2O$, M_r 130.19, *mp* 24 °C, *bp* 205 °C (at 101.3 kPa), d_4^{20} 1.001, n_D^{20} 1.4755, flash point 175 °C, is used as an intermediate for pharmaceuticals.

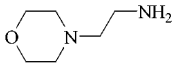

Piperazine [*110-85-0*], 1,4-diethylenediamine, hexahydropyrazine, dispermine, hexahydro-1,4-diazine, $C_4H_{10}N_2$, M_r 86.14, forms strongly hygroscopic, colorless lamellar crystals with a typical amine odor, *mp* 109.6 °C, *bp* 148 °C (at 101.3 kPa), *mp* of the hexahydrate 42.0 °C.

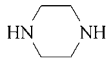

Vapor pressure

t, °C	110.8	121.7	125	140.6
p, kPa	30.3	44.6	49.6	79.7

Flash point	98 °C
Ignition temperature	340 °C
Explosion limits in air	3.93 kPa and 12.0 kPa (corresponding to 139 g/m^3 and 430 g/m^3)

These are the saturated vapor pressures above piperazine at 64.0 °C and 90.5 °C.

Specific heat capacity	2.94 J g^{-1} K^{-1} (at 130 °C) 2.99 J g^{-1} K^{-1} (at 140 °C)
Heat of vaporization	47.3 J/g

Piperazine is readily soluble in water, methanol, and ethanol but only slightly soluble in diethyl ether, benzene, and heptane. Piperazine has the chemical properties of a secondary amine. Vapor-phase dehydrogenation over copper chromite or palladium gives pyrazine.

Production. Piperazine is obtained as a byproduct in the production of ethylenediamine, in particular when ethanolamine is used as starting material [131]. If ethanolamine reacts with ammonia at 150–220 °C and 10–25 MPa, piperazine can be distilled from the reaction mixture, which also contains unreacted ethanolamine, ethylenediamine, diethylenetriamine, aminoethylethanolamine, and polyamines.

Piperazine is obtained with a purity of 99 %, and can be converted to flakes, even after conversion to the hexahydrate. Frequently, it is transported as an approximately 65 % aqueous solution, which has a low melting point of about 45 °C and hence corresponds roughly to an eutectic mixture. The purity is determined in most cases by gas chromatography. If it can be assumed that organic impurities are absent, the purity can be determined titrimetrically (HCl or $HClO_4$) in aqueous solution.

Storage. Since piperazine is corrosive, the flakes are stored in barrels lined with a polyethylene sack. To avoid yellowing, the barrels should be air tight and not exposed to direct sunlight.

The aqueous solution is stored at 50 – 60 °C in insulated iron tanks that can be heated.

Uses. Piperazine, in the form of the hexahydrate or as salts, is used as an anthelmintic in human and veterinary medicine. Furthermore, the piperazine ring is a constituent of a large number of other pharmaceuticals. It is also an intermediate for dyes, corrosion inhibitors, rubber vulcanization accelerators, insecticides, and surfactants.

The production of polyamides from piperazine and aliphatic dicarboxylic acids has long been known [132], but has remained unimportant. The polymers so formed have a heat resistance superior to that of conventional polyamides but require the use of stabilizers [133].

The use of 1,4-bis(1-aziridinylcarbonyl)-piperazine or 1,4-bis(chloroacetyl)piperazine for curing gelatine for photographic purposes was proposed [134].

***N*-Methylpiperazine** [*109-01-3*], 1-methylpiperazine, $C_5H_{12}N_2$, M_r 100.16, colorless hygroscopic liquid, *mp* – 6 °C, *bp* 138 °C (at 101.3 kPa), d_4^{20} 0.9021, n_D^{20} 1.4658, flash point 42 °C.

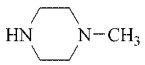

***N*-Ethylpiperazine** [*5308-25-8*], 1-ethylpiperazine, $C_6H_{14}N_2$, M_r 114.19, colorless liquid, *mp* – 60 °C, *bp* 157 °C (at 101.3 kPa), d_4^{20} 0.8989, n_D^{20} 1.4690, flash point 47 °C.

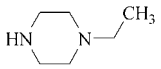

***N,N′*-Dimethylpiperazine** [*106-58-1*], 1,4-dimethylpiperazine, $C_6H_{14}N_2$, M_r 114.19, colorless liquid, *mp* – 1 °C, *bp* 131 – 132 °C (at 101 kPa), d_4^{20} 0.8509, n_D^{20} 1.4463, flash point 18 °C, can be used as catalyst for polyurethane foams and as an intermediate for cationic surfactants. It is obtained from piperazine and methanol or formaldehyde under aminating conditions.

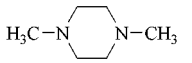

***N*-(2-Aminoethyl)piperazine** [*140-31-8*], AEP, 1-(2-aminoethyl)piperazine, 2-piperazin-1-ylethylamine, $C_6H_{15}N_3$, M_r 129.21, colorless liquid, *mp* – 19 °C, *bp* 218 – 222 °C (at 101.3 kPa), d_4^{20} 0.9842, n_D^{20} 1.5000, flash point 93 °C, η 15 mPa · s (at 20 °C) is corrosive, toxic, and hygroscopic. AEP is a byproduct of ethylenediamine production and is used as a hardener for epoxy resins, as an urethane catalyst, and also acts as a corrosion inhibitor [135]. It is also an intermediate for dyes, surfactants and various pharmaceuticals.

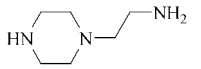

***N*-(2-Hydroxyethyl)piperazine** [*103-76-4*], HEP, 1-(2-hydroxyethyl)piperazine, 1-piperazineethanol, $C_6H_{14}N_2O$, M_r 130.19, *mp* – 10 °C, *bp* 246 °C (at 101.3 kPa), d_4^{20} 1.0595, n_D^{20} 1.5058, flash point 135 °C. HEP is obtained during ethyleneamine production and used for corrosion inhibitors, pharmaceuticals, surfactants, and synthetic fibers.

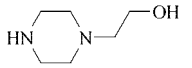

Triethylenediamine [*280-57-9*], TEDA, 1,4-diazabicyclo[2.2.2]octane, Dabco, $C_6H_{12}N_2$, M_r 112.17, is a highly symmetrical molecule with a cage structure. The colorless extremely hygroscopic crystals melt at 159.8 °C; *bp* 174 °C, pK_{a1} 2.95, pK_{a2} 8.60, flash point 50 °C.

Triethylenediamine reacts virtually quantitatively with bromine to give a 1/1 adduct. With alkyl halides it forms quaternary salts, even in nonpolar solvents. Apart from its highly nucle-

ophilic nature, triethylenediamine exhibits catalytic activity in base-catalyzed reactions.

Production. Triethylenediamine can be produced from ethylenediamine or ethanolamine [136], diethanolamine [137], or diethylenetriamine [138] with a variety of different catalysts. All processes give only moderate yields. For example, in the presence of a catalyst consisting of 86 % silicon dioxide and 12 % aluminum oxide at 360 °C and atmospheric pressure, diethylenetriamine gives about equal amounts of piperazine, alkylpiperazines, pyrazines, and ethylenediamine and homologues in addition to the bicyclic compound. Fractional distillation, subsequent recrystallization from acetone, and washing with petroleum ether give 95 % pure triethylenediamine; sublimation gives a 98 % pure product. Processes starting from piperazine derivatives (*N*-hydroxyethylpiperazine, *N,N′*-bishydroxyethylpiperazine or *N*-aminoethylpiperazine) are described in [137], [139]. Dealuminated or untreated ZSM-5 zeolites are claimed to produce TEDA from various precursors, including ethanolamine and piperazine [140]. Compared to the amorphous catalysts, the zeolite catalysts have the advantage of shape selectivity.

Uses. The only important industrial use of triethylenediamine is as a hardener for polyurethane foams [141]. However, more effective aza- and diazabicyclooctane derivatives intended for the same purpose have been the subject of patent applications [142].

Hexamethylenetetramine [*100-97-0*], HMTA, hexamine, methenamine, 1,3,5,7-tetraazaadamantane, 1,3,5,7-tetraazatricyclo[3.3.1.1(3,7)]decane, $C_6H_{12}N_4$, M_r 140.19, is a highly symmetrical molecule with an adamantane structure. Trade names: Urotropin, Formin, Aminoform. It crystallizes from ethanol as colorless, hygroscopic rhombododecahedra.

At reduced pressure (2 kPa), hexamethylenetetramine sublimes at 230 – 270 °C, virtually without decomposition, $mp > 270$ °C (decomp.), ϱ 1.3394 g/cm^3 (at 22 °C), enthalpy of formation 124.1 ± 0.75 kJ/mol. For further thermodynamic data, see [143]. For the IR spectrum, see [144]. The vapor pressure between 20 and 280 °C can be calculated from the formula $\log(p/\text{kPa}) = -524.8/(T/\text{K}) + 1.334$ [145]. Flash point 250 °C.

At 12 °C, 81.3 g of hexamethylenetetramine dissolves in 100 g of water; the solubility decreases slightly with increasing temperature. The dissociation constant is 1.4×10^{-9} (temperature not given). The solubility of hexamethylenetetramine in 100 g of various solvents at 20 °C is as follows: 13.4 g in chloroform, 7.25 g in methanol, 2.89 g in absolute ethanol, 0.65 g in acetone, 0.23 g in benzene, 0.14 g in xylene, 0.06 g in ether, and near zero in petroleum ether.

Hexamethylenetetramine gives a monobasic reaction, but with strong acids it is dibasic. The equilibrium $C_6H_{12}N_4 + 6\,H_2O \rightleftharpoons 6\,CH_2O + 4\,NH_3$ in aqueous solution permits the use of hexamethylenetetramine as a formaldehyde or ammonia donor. Thermolysis leads to hydrocyanic acid in a yield of 73 % at 800 °C, and 92 % at 1200 °C [146]. Nitration gives 1,2,3-trinitro-1,2,3-triazine, which is an important explosive (hexogen, cyclonite, RDX). The intact tricyclic structure can be alkylated and arylated at the nitrogen atom and halogenated at the carbon atom. It undergoes a large number of addition reactions: adduct formation with phosgene, disulfur dichloride, and many inorganic salts.

Production. Formaldehyde and ammonia can be converted into hexamethylenetetramine in the gas phase, in aqueous solution, or in a suspension in an inert solvent, although the most important of these procedures is that carried out in water. The usual process gives a pure product in yields near 90 % and includes an economical drying step.

If aqueous solutions of ammonia (27 %) and formaldehyde (30 %) react continuously at up to 95 °C in a V4A stainless steel tubular reactor, it is necessary to dewater a 14 % solution of hexamethylenetetramine [147]. If formaldehyde and ammonia are reacted in the vapor phase, the product is precipitated with cold water [148] or is discharged with an aqueous solution of hexamethylenetetramine [149].

The heat of reaction in aqueous solution is 230 kJ/mol, and in the vapor phase 745 kJ/mol. Therefore, the presence of a minimum amount of water is desirable. In this method [150], ammonia gas and formaldehyde gas are passed

continuously into the boiling reaction solution. The boiling point is brought to 50 – 70 °C. The heat of hydration and heat of reaction vaporize the water of condensation. The hexamethylenetetramine is removed in crystalline form or as a concentrated solution.

In another method, ammonia gas and formaldehyde solution react at pH 7.5 – 8 and 50 – 90 °C, and the product is granulated by spray drying in a hot inert gas stream (220 – 250 °C), or by spraying it onto hexamethylenetetramine particles in a fluidized bed in an air stream at 80 – 115 °C [151].

Quality. As a rule, the industrial product is 99 % pure. The compound is purified by stepwise crystallization (see [152]). The addition of paraffin, benzoic acid, acid amides, or diatomaceous earth gives free-flowing products [153].

Uses. Hexamethylenetetramine can be used as an ammonia or formaldehyde donor, for example, in the production of phenol resins, urea – formaldehyde resins and in fuel tablets. Further uses are the production of the high explosives hexogen and octogen, vulcanization accelerators, and anticorrosion agents. Proposed uses relate to the activation of chlorites as bleaches [154] and the preservation of foodstuffs [155], paints, finishes, and latexes [156], and as antibacterial agent for the urinary tract.

7. Fatty Amines

Straight-chain primary, secondary, and tertiary amines with chain lengths between 8 and 24 carbon atoms are commonly known as fatty amines. This group of compounds also includes derivatives such as the *N*-alkyl-1,3-propanediamines. Of commercial importance are fatty amine mixtures, such as coco amine, tallow amine, hydrogenated tallow amine, oleylamine, and soya amine, which are derived from naturally occurring fatty acids. More recently, alkylamines of similar structure were produced from synthetic feedstocks such as olefins or paraffins, and these also are called fatty amines (β-amines, branched chain amines, aminoalkanols). Only partly covered are fatty amine ethoxylates, fatty alkyl quaternary ammonium salts, alkyl betaines and fatty amine oxides, although all of these are fatty amine derivatives.

7.1. Properties

Physical Properties. Table 2 lists typical compositions of fatty amine mixtures. The properties of these can be varied to some extent by blending chain lengths with the desired physical characteristics. Fatty acids are also available in fractionated form and thus narrower ranges of fatty amines may be produced.

Solubilities of selected fatty amines in polar and nonpolar organic solvents are given in Table 3. Only the short-chain fatty amines with 8 – 10 C atoms per alkyl chain are of limited solubility in water.

Table 3. Solubility (g per 100 mL at 30 °C) of fatty amines in organic solvents [158]

	2-Propanol	Hexane
Primary amines		
Dodecylamine	∞	∞
Tetradecylamine	458	216
Hexadecylamine	169	64.8
Octadecylamine	86	27.9
Secondary amines		
Dioctylamine	∞	∞
Didodecylamine	55	27.5
Dioctadecylamine	1.2 (50 °C)	2.1 (40 °C)
Tertiary amines		
Trioctylamine	∞	∞
Tridodecylamine	23.9	∞
Trioctadecylamine		36.4

Vapor pressures at ambient temperature are generally very low and decrease from 26.5 Pa for dodecylamine to 2.7×10^{-2} Pa (0.27×10^{-3} mbar) for eicosylamine. Some physical properties of pure compounds and commercial mixtures are given in Tables 4 and 5.

The oscillation of *melting points* between the series of odd-chain and even-chain homologues is shown in Figure 2. This effect disappears in derivatives of primary fatty amines.

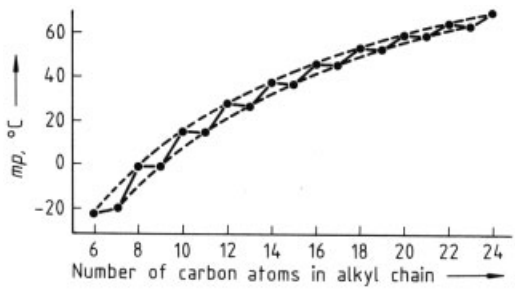

Figure 2. Melting points of primary amines

Table 2. Mixture composition of primary fatty amines (mass fraction in %) [157]

	Number of carbon atoms in alkyl chain													
	Saturated alkyl									Unsaturated alkyl ∗				
	6	8	10	12	14	15	16	17	18	14′	16′	18′	18″	Other
Amine														
Coco	0.5	8	7	50	18		8		1.5			6	1	
Hydrogenated-tallow				1	4	0.5	30	1.5	60	0.5	0.5	2		
Tallow				1	3	0.5	29	1	23	1	3	37	1.5	
Oleyl				0.5	3.5	0.5	4	1	5	1.5	5	76	3	
Soya				0.5	1		16		15		1	49.5	13	4

∗A single prime indicates one double bond; a double prime, two double bonds.

Table 4. Physical properties of fatty amines [159–162]

	Formula	M_r	*mp*, °C	*bp* ∗, °C	Hydrochloride decomp. temp., °C	Acetate *mp*, °C
Pure primary amines						
Octylamine	$C_8H_{17}NH_2$	129.25	−0.1	180.0	198	
Decylamine	$C_{10}H_{21}NH_2$	157.30	15.9	221.8	194	
Dodecylamine	$C_{12}H_{25}NH_2$	185.35	28.2	259.1	179 – 181	56.0 – 56.3
Tetradecylamine	$C_{14}H_{29}NH_2$	213.41	38.0	292.3	150 – 167	66.0 – 66.8
Hexadecylamine	$C_{16}H_{33}NH_2$	241.46	46.0	321.9	148 – 166	74.0 – 74.2
Octadecylamine	$C_{18}H_{37}NH_2$	269.52	53.1	348.5	158 – 161	79.8 – 80.0
Eicosylamine	$C_{20}H_{41}NH_2$	297.57	59.5	372.5	151 – 154	85.0 – 85.2
Docosylamine	$C_{22}H_{45}NH_2$	325.63	65.3	394.6	150 – 153	88.8 – 89.0
Commercial mixtures of primary amines						
Coco amine			12 – 17	130 – 227		ca. 50
Oleylamine				200 – 210		
Tallow amine			32 – 40	200 – 230		ca. 55
Hydrogenated-tallow amine			48 – 56			ca. 60
Soya amine			27 – 30			

∗ At 101.3 kPa for pure compounds, at 3.6 kPa for mixtures

Table 5. Physical properties of secondary and tertiary fatty amines

	Formula	M_r	*mp*, °C		*bp*, °C (*p*,Pa)
Secondary amines			(α)	(β)	
Dioctylamine	$(C_3H_{17})_2NH$	241.46	14.8	26.5	115 (133)
Didecylamine	$(C_{10}H_{21})_2NH$	297.57	33.6	41.4	152 (133)
Didodecylamine	$(C_{12}H_{25})_2NH$	353.68	47.2	51.0	183 (133)
Ditetradecylamine	$(C_{14}H_{29})_2NH$	409.79	58.1	58.8	209 (133)
Dihexadecylamine	$(C_{16}H_{33})_2NH$	465.90	67.0		231 (133)
Dioctadecylamine	$(C_{18}H_{37})_2NH$	522.01	74.7		250 (133)
Dicoco amine			40 – 47		
Dihydrogenated-tallow amine			60 – 65		
Tertiary amines					
N-Methyldioctadecylamine	$(C_{18}H_{37})_2NCH_3$	536.03	40		252 – 259 (6.7)
N,N-Dimethyldodecylamine	$C_{12}H_{25}N(CH_3)_2$	213.41	−20.5		135 (1333)
N,N-Dimethyloctadecylamine	$C_{18}H_{37}N(CH_3)_2$	287.49	22.5		202 (1333)

Almost all amines derived from natural sources have an even number of carbon atoms. Diamines and 2-(alkylimino)diethanols [*N,N*-bis(2-hydroxyethyl)alkylamines] are low-melting pasty solids. Unsaturation decreases the melting points of the amines relative to the fully saturated compounds. Fatty amines and their derivatives are cationic surfactants (→ Surfactants) and show strong substantivity to the negatively charged surface of most solids, e.g., cotton, minerals, and metals, and thus change the physical properties of the substrates [163], [164].

Chemical Properties. Fatty amines, like short-chain amines, are stronger bases than am-

monia. Their ionization constants [165] do not change significantly with chain length (see Table 6), and their reactions are similar to those known from the lower homologues.

Table 6. pK_b Values of amines in water at 25 °C *

R	RNH_2	R_2NH	R_3N
Methyl	3.38	3.29	4.24
Ethyl	3.37	3.02	
Octyl	3.35	3.00	
Octadecyl	3.40	3.00	

* The pK_b of NH_3 in water at 25 °C is 4.79.

The following are industrially important reactions:

1) *Salt formation* with inorganic and organic acids. The hydrochlorides are slightly more soluble and the acetates considerably more soluble in water than the free amines. Acetates and salts of higher fatty acids are readily soluble in organic solvents [158]. The melting points of the salts are higher than those of the free amines.
2) *Conversion* into *N*-alkyl carboxamides with carboxylic acids or their derivatives.
3) Reaction with *carbon dioxide* to give carbamates, which are formed as contaminants in amines exposed to air.
4) *Alkylation* with excess of strong nucleophiles, such as methyl chloride, benzyl chloride, or dimethyl sulfate, leading eventually to quaternary ammonium compounds. This reaction is normally carried out in aqueous alcohol as the solvent and in the presence of alkali as the proton acceptor [166], [167].
5) *Oxyalkylation*with ethylene oxide, a more selective reaction than alkylation. The uncatalyzed reaction with primary amines at 150 – 200 °C yields almost exclusively 2-(alkylimino)diethanols (*N,N*-bis(2-hydroxyethyl) fatty amines). Base catalysis even at temperatures as low as 100 °C leads to bis(polyoxyethyl) amines [168]. Under mild conditions (80 °C, 350 Pa) and in the presence of water or acids, secondary and tertiary amines react readily with ethylene oxide to form the corresponding bis- or mono(2-hydroxyethyl) quaternary ammonium salts [169–171]. Long-chain epoxides undergo a similar reaction [172], but products from ethylene and propylene oxide are the only important commercial products.
6) Reaction of fatty amines with *halogenated carboxylic acids* or with *lactones* to give amphoteric compounds (betaines).
7) *Michael addition* to activated double bonds. For example, the important 1,3-propanediamines are synthesized by the reaction of a primary fatty amine with acrylonitrile followed by hydrogenation of the intermediate 2-(alkylimino)ethanenitrile [173].
8) *Mannich reactions* with formaldehyde and nucleophiles, such as aldehydes, nitroparaffins, or phenols, to form polyfunctional derivatives [174–176].
9) *Reaction with phosgene* to form isocyanates and ureas.
10) *Oxidation* by hydrogen peroxide, peroxoacids, or ozone to give amine oxides. Industrially, oxidation of tertiary amines to amine oxides by hydrogen peroxide is very impotant since the products find use as surfactants.

7.2. Production

Fatty amines can be produced from natural fats and oils or from synthetic raw materials [177].

Production from Fatty Nitriles. The most important source of fatty amines is still fatty nitriles, which are themselves formed from fatty carboxylic acids and ammonia over optional dehydrating catalysts (Al_2O_3, SiO_2, ZnO, bauxite, or salts of Ti, Fe, Mn, or Co) in batch liquid-phase reactors or continuous liquid- and vapor-phase reactors at 160 – 430 °C and 0.03 – 0.7 MPa. The principal raw materials are coconut oil, palm kernel oil, and tallow.

Depending on the process, the nitriles are purified by distillation or used as such. The fatty acid feedstocks may be fractionated beforehand as well. Hydrogenation of nitriles can lead to all three types of amine, depending on reaction conditions. Hydrogenation may be batch or continuous and is sensitive to catalyst poisons from the natural feedstock.

Primary amines are formed by hydrogenation of fatty nitriles, typically at 80 – 140 °C and 1 – 4 MPa, over nickel or cobalt catalysts. Raney cobalt or copper chromite catalysts are

used to obtain *unsaturated amines* [178–180]. To avoid formation of secondary amines, ammonia is commonly used as a suppressant; a solvent may also be used.

Secondary Amines. If ammonia is continuously or periodically vented from the reactor and temperatures of 150 – 210 °C and pressures of 5 – 20 MPa are maintained, saturated and unsaturated secondary amines can be obtained in yields of > 90 % [181], [182]. Catalysts mainly used are nickel, cobalt, and copper chromite.

Tertiary Amines. Symmetrical trialkylamines are obtained similarly to the secondary amines, but usually mixtures of primary, secondary, and tertiary amines result. Tertiary amines may also be produced from nitriles via the imine RCH=NH and Schiff base RCH=N–CH_2R by using supported nickel catalysts at 230 °C and 0.7 MPa hydrogen partial pressure [183].

More important are the *N*-alkyl-*N*-methylalkylamines and the *N,N*-dimethylalkylamines, which can be synthesized by the Leuckart reaction of primary or secondary amines [184], [185]:

$$R_2NH + CH_2O + HCOOH \longrightarrow R_2NCH_3 + CO_2 + H_2O$$

$$RNH_2 + 2\,CH_2O + 2\,HCOOH \longrightarrow RN(CH_3)_2 + 2\,CO_2 + 2\,H_2O$$

Industrially, the reduction of the formaldehyde may be achieved with hydrogen and a nickel catalyst, e.g.:

$$RNH_2 + 2\,CH_2O + 2\,H_2 \longrightarrow RN(CH_3)_2 + 2\,H_2O$$

N,N-dimethylalkylamines can also be produced from fatty nitriles and dimethylamine [186]. Both types of methyl-substituted tertiary amines are accessible also from primary alcohols [187], [188], and this route is used almost exclusively.

Efforts to use methanol or hexamethylenetetramine to methylate the nitriles with different catalysts to obtain *N,N*-dimethylalkylamines or *N*-methyldialkylamines are descibed in [189].

Production from Alcohols or Carbonyl Compounds. Fatty amines are produced by reductive alkylation of ammonia or substituted amines with primary linear alcohols ($C_8 - C_{18}$) at 90 – 190 °C under low pressure (atmospheric up to 0.7 MPa) in the presence of a Raney nickel catalyst. If water is continuously removed from the reactor, only secondary and tertiary amines are obtained [190–192]. Similarly, aldehydes can be converted into primary amines by reductive amination at 110 – 120 °C and 1.5 MPa with rare earth-promoted cobalt catalysts [193]. Oxo aldehydes react with methylamine to form imines, which are then hydrogenated to give *N*-methylalkylamines at 115 °C and 3 MPa with a Raney nickel catalyst [194]. Similarly, both alcohols and aldehydes can be converted into *N,N*-dimethyl tertiary amines with dimethylamine. Alcohols are converted into amines at 230 °C at atmospheric pressure with copper chromite catalysts, whereas fatty aldehydes require noble metal, copper chelate, or copper carboxylate catalysts [195–198].

The widely used fatty alkyl dimethylamines are nowadays almost exclusively produced from long-chain natural, oxo, or Ziegler fatty alcohols ($C_{12} - C_{18}$). Conversion is achieved by reacting dimethylamine with the fatty alcohol at elevated temperatures and pressures with copper, copper chromite, or nickel catalysts. Alternatively, the fatty alcohol can initially be converted to the primary fatty amine by using ammonia and a nickel catalyst and subsequently treated with formaldehyde as described above. Some other two-stage processes starting from fatty alcohols and with sodium alkyl sulfates or chlorides as intermediates have been described, but direct conversion with dimethylamine is the most attractive.

Production from Olefins. α-Olefins can be converted to 1-bromoalkanes by radical addition (anti-Markovnikov) of hydrogen bromide [199]. Reaction with ammonia furnishes primary fatty amines, and using dimethylamine results in tertiary fatty amines. A higher amount of branched chain isomers is likely for the α-olefin route as compared with the alcohol route.

Under hydrocarbonylation conditions, α-olefins and dimethylamine form *N,N*-dimethylalkylamines directly if highly selective noble metal catalysts are used [200]. Use of the less specific tributylphosphinecobalt catalysts leads to mixtures of amines and alcohols [201].

In a *Ritter reaction*, α-olefins react with hydrogen cyanide or acetonitrile in concentrated sulfuric acid or aqueous hydrogen fluoride, via the intermediate 2-acylaminoalkane, to

give branched alkyl primary amines (β-amines). The reaction is not specific, and positional isomers are also formed [202], [203].

Aluminum alkoxides, as formed in the Ziegler process, can be used to convert secondary amines into tertiary amines [204].

Production from Alkane Derivatives. *Nitroalkanes* with a statistical distribution of nitro groups can be hydrogenated over Pd/C at 190 °C and 3.9 MPa to form secondary alkyl primary amines in yields of up to 93 % [205]. Nickel catalysts can also be used. Selectively *chlorinated alkanes* react with dimethylamine to yield up to 65 % terminal tertiary amine [206].

Production from Other Starting Materials. Glycerides and methyl esters of fatty acids can be converted into fatty nitriles at 220 – 300 °C provided ammonia is vented rapidly. Yields of nitrile and glycerine in excess of 80 % were reported with zinc carboxylate or aryl sulfate catalysts [207]. Under hydrogenation conditions over Zn – Al catalysts at high pressures (10 – 30 MPa) and temperatures (200 – 350 °C), coconut glycerides can be converted into primary coco amine. Glycerine, however, decomposes completely under these conditions [208].

Primary fatty amides can be converted in high yields into odd-chain primary amines by Hofmann degradation of the chloroamide intermediate [209]:

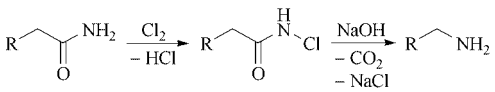

These fatty amines have an uneven number of carbon atoms and are hence distinct from the more common primary amines from the nitrile route.

***N*-Alkyl-1,3-propanediamines.** These so called fatty diamines are obtained by acrylonitrile addition to fatty amines followed by hydrogenation in the presence of ammonia:

$$RNH_2 + CH_2{=}CHCN \longrightarrow RNH(CH_2)_2CN$$

$$RNH(CH_2)_2CN + 2\,H_2 \longrightarrow RNH(CH_2)_3NH_2$$

Acrylonitrile addition may be catalyzed by the presence of methanol and is conducted at temperatures between 30 and 90 °C. Hydrogenation is performed under the same conditions as for the fatty nitrile hydrogenations above. Polyamines may also be obtained similarly.

7.3. Analysis and Quality Control

Because of their relatively high basicity, fatty amines can be titrated easily with acids. Most of the amines are insoluble in water; therefore, organic solvents, such as isopropyl alcohol, are commonly used. End points are determined colorimetrically (bromophenol blue) or potentiometrically. In the second case, perchloric acid and glacial acetic acid are preferred as titrant and solvent, respectively.

Industrial mixtures of amines normally contain up to 90 % of the main compound. These mixtures are analyzed by stepwise titrations with acid. The total basicity is determined first. Primary amines then are removed with salicylaldehyde, and primary and secondary amines can be removed with phenylisothiocyanate or acetic acid, so that the remaining free amine can be titrated with acid. Unsaturated amines are determined by means of the iodine number. Because amine groups also consume iodine, the original Wijs method must be slightly modified by using acetic acid as the solvent. The resulting amine salt absorbs the halide at a much slower rate than the double bonds.

Quaternary ammonium compounds usually are determined by solvent partition titration with anionic surfactants or sodium tetraphenylboron [210]. Iodine numbers of quaternary ammonium compounds are obtained in sodium lauryl sulfate containing chloroform to prevent the free iodine from being retained by the nonaqueous phase. A summary of approved wet chemical test methods (AOCS, ASTM) is given in [211].

Chain-length distribution and hydrocarbon content of fatty amines are determined by gas chromatography [212–216]; nitrile and amide impurities are determined by IR spectroscopy [217], [218].

More recently, NMR spectroscopy has been used for determination of quaternary ammonium compounds in the presence of free amines [219] and also for establishing the relative amounts of

primary, secondary, and tertiary amines in mixtures [220].

Fatty amines can be stored in carbon steel containers; aluminum, copper or their alloys are to be avoided. Quaternary amines are more corrosive and thus require lined drums or stainless steel. With water hydrates are formed by primary and secondary amines and carbamates with CO_2.

7.4. Uses

Fabric Softeners. By far the largest use of fatty amines and in particular their quaternary derivatives used to be in fabric softening. Prior to the early 1950s, fabric softeners were not needed because most detergents were based on tallow-derived soap formulations and some of the active detergent remained on the textile and resulted in a softer dried fabric. However, the advent of synthetic detergents created the market for fabric softeners in home care and institutional laundries.

Most of these softeners are used in *rinse cycles* where a 3 – 8 % formulated aqueous dispersion is added leaving 0.1 – 0.2 % active quaternary ammonium compound on the surface of the clothes. Since the 1980s, concentrates have been developed, which contain up to 50 % active ingredients. In another usage, a quaternary fabric softener is applied to a substrate, such as paper, to woven or nonwoven cloth, or to a foam and then added to the *drying cycle*. The formulation contains a quaternary ammonium compound and a transfer agent that allows the active ingredient to be transferred to the cloth during the drying cycle. The prevention of static buildup is actually more important than the softening action in this application. A third area of application are *wash cycle* softeners, in which quaternaries are combined with other detergents to achieve both cleaning and softening.

The fatty quaternary ammonium compounds used were formerly mainly dimethyl dihydrogenated tallow ammonium chloride, but this compound was replaced in many countries by other surfactants because of its lack of biodegradability and potential aquatic toxicity. Nowadays the most widely used quaternaries are ester quats, which are not derived from fatty amines but from fatty acids, alcoholamines (e.g., triethanolamine), and methylating agents.

Amine ethoxylates, mainly derived from primary tallow, hydrogenated tallow and coconut amines, are also used in large amounts for textiles.

Flotation. One of the first commercial uses of primary fatty amines was in the beneficiation of potash [221]. In this process, fatty amine salts adhere to the surface of potassium chloride, allowing it to be removed by froth flotation and leaving the sodium chloride in the tailings. The technique spread to other mineral beneficiation processes [222], the most important of which are phosphate rock, precious metals, lead, zinc, copper, molybdenum, tin, fluorspar, feldspar, magnesite, limestone, mica, manganese, silica, and coal. In general, any negatively charged surface can be removed from a system with an amine salt. Ore flotation accounts for the largest direct use of primary fatty amines. Other fatty amines used for mining are quaternary componds and some diamines, but these are less important.

Corrosion Inhibition. Large quantities of amines are used in corrosion inhibition. The cationic amino group adheres to the surface of metals and prevents attack by corrosive liquids and gases. Primary amines and *N*-alkyl-1,3-propanediamines are used most frequently for this purpose. The amines may be neutralized with oleic, naphthenic, and acetic acids to adjust their solubility. A typical oil-soluble inhibitor formulation contains *N*-tallow-1,3-propanediamine dioleate.

Asphalt Emulsification. The cationic nature of fatty amines makes them useful as emulsifiers for asphalt. Primary amines, diamines, polyamines, alkoxylated amines, and fatty quaternary ammonium compounds can be used in this application. The formulation used depends on the properties of the emulsion desired. The cationic emulsion can be rapid, medium, or slow set. In all cases a cationic emulsifier is used at a low concentration (0.2 – 2.0 wt % of the emulsion). When the cationic emulsion is mixed with the negatively charged aggregate, for instance limestone, the emulsion breaks, thereby setting the road-covering formulation. Fatty amines, because of their positive charge, also increase the

adherence of the asphalt to the aggregate. For di- and polyamines, asphalt emulsification is by far the largest application and they are also the most important surfactants for this purpose.

Other Applications. Significant amounts of primary amines are used as fertilizer anticaking agents. Frequently, fatty amine salts are also added to other mineral products to prevent caking of the powdered mineral.

There are many other applications for fatty amines in *agriculture* as emulsifiers, adjuvants, and intermediates in pesticide production. Much of their usefulness relies on the controllable lipid solubility, which aids penetration or absorption of many active pesticides. Among the emulsifiers used are ethoxylated amines and quaternary ammonium compounds. Some of the surfactants also can act as foaming agents or foam stabilizers [223].

Neutralization of various herbicidal acids with fatty amines enhances their activity because the corresponding salts are more soluble in oil and less volatile than the free amines [224–227]. In some cases, the *N*,*N*-dimethylalkylamines were found to control the growth of suckers or unwanted leaves, thereby producing a better grade of product, for example, in tobacco [228]. The reaction of dodecylamine and cyanamide leads to dodecylguanidine (Dodine from ACC, Agrimont, Chemolimpex, and Drexel), which is an excellent fungicide for fruit trees.

Many fatty alkylamine derivatives are used as additives in plastic formulations. For example, bis(2-hydroxyethyl)alkylamines are added to polypropylene or polyethylene as internal *antistatic agents*, either as 100 % amines or as a 75 % solid masterbatch to prevent static charge buildup at low humidities.

Fatty amines, diamines, and ethoxylated amines also can be used as *pigment-grinding aids* and as *dispersants* for pigments in paints, coatings, and magnetic tape. The most widely used compound is *N*-tallow-1,3-propanediammonium dioleate. The amine compound coats the pigment particle, making it more readily and evenly dispersible in the formulation.

The *organo-clay* market also is a very large user of quaternary fatty amines. Most of this market is satisfied with methyl or benzyl quaternary ammonium chlorides produced from mono- or dihydrogenated tallow amine. In either case, the cations of the quaternary ammonium chloride are exchanged with those of the clay, mostly bentonite, to produce clay platelets separated with a hydrophobic compound. These organo-clay products are used as thickening agents, mainly in drilling muds and oil-based coatings, but also in the paint industry.

Some primary amines, amine salts, di- and polyamines, e.g., lauryl and tallow amines, posess biocidal properties; monoalkyl quaternary amines with a $C_{10}-C_{18}$ chain act as germicides and antiseptics and therefore have a wide range of biocidal applications, for example, in households, swimming pools, and oil fields.

Primary amines are also important lubricants since their polar nitrogen end attaches to metals while the nonpolar chains form an oily film above the surface. They are also used as petroleum additives. Secondary amines have few direct uses and are mainly used as intermediates for ethoxylation or quaternization. Tertiary amines with three long chains are of minor commercial significance, but can be used, for instance, as deicers for motor fuels. The fatty alkyldimethylamines are used almost exclusively as intermediates for alkyl betaines, amine oxides, and quaternary ammonium salts.

7.5. Economic Aspects

The economics of fatty amines production generally depend on the worldwide fat and oil market process. Most of the fatty amine products are made from tallow and coconut oils; others are made from oils such as soybean, palm, and cottonseed. Minor amounts of tall oil acids are used. When pure chain lengths are required, these can be obtained by fractionating the acid mixture found in the initial triglyceride. Pure fractions are much more expensive than mixtures. Primary and secondary amines have a much larger market than tertiary amines.

The nitrile route, which is mainly used for the production of primary and secondary amines, accounted for 300 000 t/a world capacity in 1996 for all alkylamines and their derivatives and has decreased since the early 1990s from about 370 000 t/a due to declining demand for the poorly biodegradable dihydrogenated tallow dimethyl ammonium salts for fabric softeners. The

fatty alcohol and α-olefin route, which is mainly used for the production of alkyldimethylamines and their derivatives, is less important but is expected to surpass the nitrile route in the long run. In Europe, the largest producers employing the nitrile route are Akzo Nobel, Clariant, Ceca, and Olefina, in North America Witco and Akzo Nobel, in South America Akzo Nobel, Clariant, and Quimi Kao, in Asia Kao and Lion-Akzo.

The most important producers of *N,N*-dimethyl fatty amines are Procter & Gamble, Albemarle, and Lonza in the United States, Clariant, Albright & Wilson, and Kao in Europe, and Kao and Lion-Akzo in Asia. Worldwide demand was estimated at about 100 000 t/a in 1998. In 1988 it was only ca. 75 000 t/a. North America accounts for 59 % of the installed capacity, Europe for 28 %, and Asia for 8 %.

Because most fatty amines offered commercially contain mixed alkyl chain lengths, trade names are common; a comprehensive list can be found in [229]. In Europe, about 20 000 t were sold in 1996 as the underivatized primary, secondary or tertiary amines.

8. Diamines and Polyamines

8.1. Diamines

8.1.1. 1,2-Diaminoethane (Ethylenediamine)

(For triethylenediamine, see page 23)

Physical Properties. 1,2-diaminoethane [*107-15-3*], EDA, $C_2H_8N_2$, M_r 60.10, is a colorless liquid, *mp* 10.9 °C, *bp* 117 °C (at 101.3 kPa), d_4^{20} 0.8978, n_D^{20} 1.4571.

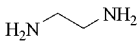

Viscosity

t, °C	25	60	100
η, mPa · s	1.35	0.7	0.4

Vapor pressure

t, °C	20	40	60	80	100
p, kPa	1.21	3.87	11.13	26.53	55.99

Flash point	39.4 °C	
Ignition temperature (class T2)	390 °C	
Explosion limits in air	2.7 and 16.6 vol % (corresponding to 67.4 and 415 g/m^3)	
Specific heat capacity	3.41 J g^{-1} K^{-1}	(at 100 °C)
	3.45 J g^{-1} K^{-1}	(at 110 °C)
	3.50 J g^{-1} K^{-1}	(at 120 °C)
Heat of vaporization	636.4 J/g (at 0.1 MPa)	
	660 J/g (at 40 kPa)	

Vapor pressures, thermodynamic data, densities, and viscosities have been shown in graphical form [230]. Diaminoethane is infinitely miscible with water, methanol, diethyl ether, acetone, and benzene but is only slightly soluble in lower hydrocarbons. It reacts, however, with carbon tetrachloride or methylene chloride. It forms binary azeotropes with water, several alcohols, or toluene [231].

Chemical Properties. As a primary diamine, diaminoethane forms stable mono- and diacid salts. The free base can be liberated from these salts by inorganic alkali solutions. The mononitrate and dinitrate are highly explosive in the dry state. The reaction of diaminoethane with chloroacetic acid or with formaldehyde and hydrogen cyanide or an alkali-metal cyanide to give ethylenediaminetetraacetic acid or its salts is particularly important (→ Ethylenediaminetetraacetic Acid and Related Chelating Agents). Diaminoethane forms complex salts with salts of copper, manganese, cobalt, and other transition elements.

If diaminoethane is heated with a fatty acid under conditions favoring dehydration, the product is a mono- or diamide, which can be converted into an imidazoline by cyclization. At high temperatures and pressures and in the presence of a catalyst, ammonia is eliminated from diaminoethane, and the resulting product undergoes cyclization to give piperazine. Diaminoethane reacts with aldehydes and ketones to give Schiff bases, which in turn can be hydrogenated to give secondary amines. The Schiff base can also cyclize to afford imidazolidines. Condensation with 1,2-diketones results in the elimination of water to give the corresponding 2,3-dihydropyrazines, which can be readily dehydrogenated to pyrazines [232]. Phosgene reacts with diaminoethane to give the hydrochloride of *N,N′*-ethyleneurea (2-imidazolidone) [233]. One mole of diaminoethane reacts with

two moles of carbon disulfide in the presence of a basic catalyst to give ethylene bisdithiocarbamate [234]. With only one mole of carbon disulfide, ethylenethiourea is formed. Similarly, ethyleneurea is obtained from carbon dioxide. With two moles of acrylonitrile, diaminoethane undergoes a smooth cyanoethylation reaction [235].

Production. 1,2-Diaminoethane is mainly produced by treating ethylene dichloride (EDC) with aqueous or liquid ammonia at about 100 °C in the liquid phase [236]. This so-called EDC process has been modified frequently [237]; the reactant ratio, product recycle, pH, reactor geometry, temperature, and pressure control the product mix. Byproducts include the higher oligomers diethylenetriamine (DETA), triethylenetetramine (TETA), and tetraethylenenpentamine (TEPA). An unavoidable coproduct of the EDC process is the amine hydrochloride, which must be neutralized with caustic soda, lime, or other bases to form, e.g., sodium or calcium chloride. Ethylenediamine is either extracted or distilled from the aqueous stream after neutralization. If deemed necessary, the higher amines can be recycled to optimize EDA production and vice versa. About two-thirds of the installed ethylenediamine capacity still relies on the EDC process.

A second process is the reaction of monoethanolamine with ammonia and hydrogen over a nickel or cobalt catalyst at 20 MPa and 150 – 230 °C [238]. The yield of 1,2-diaminoethane is 74 %, based on an ethanolamine conversion of 93 %. Byproducts are mainly DETA and piperazine, but also aminoethylpiperazine (AEP) and hydroxyethylpiperazine (HEP). The monoethanolamine process has cost advantages over the ethylene dichloride process, circumvents chloride disposal, and prevents the formation of chlorinated hydrocarbons. Lower amines can be recycled to increase the amount of higher ethanolamines; cyclic products are much more abundant than in the EDC process. Product distribution is dependent on reactant ratio, temperature, pressure, and conversion. Nickel – rhenium catalysts have often been employed [239], but many other nickel catalysts with different promoters have also been suggested, for instance, Ni – Re – B supported on alumina [240] or Ni – Co – Cu – Re – B on silica [241].

A third important and very cost effective production process is the reaction of ethylene oxide with ammonia [242], as developed by Berol (now Akzo) in Sweden. Among the byproducts of this process are DETA, piperazine and di- and triethanolamine. Union Carbide has recently brought a similar process into operation. The production of diaminoethane by the reaction of formaldehyde with hydrocyanic acid, ammonia, and hydrogen [243], of chloroacetyl chloride with ammonia [244], and by hydrogenation of aminoacetonitrile [245] are of lesser importance.

Quality, Analysis, and Storage. The purity of diaminoethane, determined by titration and gas chromatography, is usually above 99 %. For storage, containers made of stainless steel and aluminum are preferred to avoid deterioration of the color number when the product is stored for long periods. Because it is corrosive, EDA should not be stored in containers made of copper, copper alloys, or uncoated iron. Frequently containers made of iron coated with plastics, such as Lupolen (BASF) or phenolic resins, are used for shipping and storing diaminoethane. Tin-lined drums can also be employed. Because the amine is hygroscopic and reacts with carbon dioxide in the air, tanks must be flushed with nitrogen before use. This also avoids discoloration during storage.

Uses. Diaminoethane is mainly used in detergents, resins, crop protection agents, paper chemicals, lubricants, and pharmaceuticals. The chelating agent ethylenediaminetetraacetic acid (EDTA) and its salts are important products, in particular in the United States. They are made by reaction of EDA with formaldehyde and hydrogen cyanide or sodium cyanide together with an excess of sodium hydroxide. Tetraacetylethylenediamine (TAED), which is obtained from EDA, is used in the detergent industry as a cold-bleach activator. It has become the fasted-growing EDA product in the last years. Hydroxyethylethylenediamine is also converted into chelating agents. Several nitrogen-containing surfactants are produced from fatty acids and EDA or aminoethylethanolamine, available from EDA and ethylene oxide. Nonreactive polyamide resins produced from EDA and diacids also consume large amounts of EDA. They find use as adhesives and for printing inks.

The zinc and manganese salts of ethylenebisdithiocarbamic acid, obtained from EDA, are used as fungicides (maneb, mancozeb, zineb or metiram) and have a significant market. Other fungicides based on EDA are known, e.g., those of the imidazoline type.

The EDA-derivative *N,N*-ethylene bis(stearamide) (EBS) has lubricating properties and is used as a mold-release agent, for the processing of thermoplastic resins, in paper coating, wire production, etc. EDA is also used for gasoline and lubricant additives and for surfactant production.

Diaminoethane – polyester condensates that are hydroxymethylated with formaldehyde can be used as plasticizers for phenol – formaldehyde resins [246]. The use of condensates of diaminoethane, epoxides, and urea as nitrogen-containing polyol components for the production of polyurethane foams is proposed in [247]. The addition of EDA to viscose fiber spinning baths is said to improve the tensile strength of the fiber [248]. In the wet spinning of polyurethane fibers, diaminoethane is a rapid-action curing agent [249]. The incorporation of diaminoethane into diisocyanate – polyester prepolymers results in useful polymers for the production of elastic polyurethane fibers [250]. Ethyleneurea derivatives prepared from diaminoethane are used for textile finishing. Diaminoethane is employed as a stabilizer for rubber latex. In the mineral oil industry, diaminoethane can act as a stabilizer for halogen-containing high-pressure lubricating oils [251] and, in the form of Schiff bases with ketones, is also used as a metal deactivator [252].

Economic Aspects. Ethylenediamine and its homologues are large-scale chemical products. The global demand for ethyleneamines in 1998 was estimated at 250 000 t [253]; capacities are around 340 000 t/a. The European market is ca. 100 000 t/a with an annual growth rate of 3 – 4 % [254]; the U.S. market has a similar size, and the Japanese is ca. 15 000 t/a. Major producers are Union Carbide, Dow, Akzo Nobel, Tosoh, BASF, and Bayer. In the United States and Japan, the ethylenediamine market is smaller than that of the higher ethyleneamines, and vice versa in Europe.

8.1.2. Diaminopropanes

In the past few years, the demand for diaminopropanes and in particular for *N*-substituted 1,3-diaminopropanes has increased sharply in a large variety of fields.

1,2-Diaminopropane [*78-90-0*], propylene-1,2-diamine, $C_3H_{10}N_2$, M_r 74.13, *mp* – 36.6 °C, *bp* 120.9 °C (at 101.3 kPa), d_4^{20} 0.872, n_D^{20} 1.446, η 1.6 mPa · s (at 20 °C), flash point 33 °C, is a colorless liquid that is very hygroscopic and miscible with water and most organic solvents. It forms azeotropes with butanol, 2-methylpropanol, and toluene and has a strong ammoniacal odor.

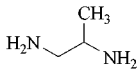

Because 1,2-dichloropropane does not react smoothly with ammonia, 1,2-diaminopropane is usually prepared by aminating a mixture of 2-amino-1-propanol and 1-amino-2-propanol, which is obtained from propylene oxide and ammonia [255]. The process differs from corresponding processes for the production of diaminoethane in that the temperature in the catalytic amination is higher (190 – 230 °C). Producers of 1,2-diaminopropane include Bayer and BASF.

1,2-Diaminopropane is used as an intermediate for crop-protection agents, e.g., Basfungin (BASF). Analogous to the ethylenediamine derivatives, the zinc dithiocarbamate propineb (Bayer) acts as a fungicide. Especially in the United States, derivatives of the amine are used as fuel and lubricant additives.

1,3-Diaminopropane [*109-76-2*], propylene-1,3-diamine, trimethylenediamine, propanediamine, $C_3H_{10}N_2$, M_r 74.13, *mp* – 11.8 °C, *bp* 136 °C (at 101.3 kPa), d_4^{20} 0.8834, n_D^{20} 1.459, η 2.0 mPa · s (at 20 °C), flash point 48 °C. 1,3-Diaminopropane is a colorless liquid that is miscible with water and most organic solvents.

1,3-Diaminopropane is produced in a two-stage continuous process, under a pressure of

10 – 20 MPa. In a first reactor, acrylonitrile reacts with excess ammonia at 70 – 100 °C to give 2-aminopropionitrile and bis(cyanoethyl)amine. The mixture of aminonitriles is then hydrogenated in a downstream reactor over a fixed-bed catalyst (cobalt or nickel) at 60 – 120 °C to give 1,3-diaminopropane and bis(aminopropyl)amine [256]. The yield of 1,3-diaminopropane can be increased by adding a polar solvent and water or by using a large excess of ammonia. Most of the 1,3-diaminopropane produced is converted into the textile-finishing agent 1,3-dihydroxymethylhexahydropyrimid-2-one. 1,3-Diaminopropane is also utilized for the production of a chelating agent used in the photo industry and for anticorrosion agents.

***N*-Substituted Diaminopropanes.** *N*-Substituted 1,3-diaminopropanes can be prepared from acrylonitrile and an amine, also with such a two-stage procedure. It is usually more efficient to produce the substituted aminopropionitrile batchwise and then hydrogenate it continuously in the presence of ammonia [257]. In some cases, hydrogenation also can be carried out batchwise over Raney nickel or Raney cobalt. A primary amine or ammonia reacts with acrylonitrile in a molar ratio of about 1/2 to give (after subsequent hydrogenation) alkylbis(aminopropyl)amine, $RN[(CH_2)_3NH_2]_2$, as the principal product, whereas until the molar ratio exceeds 1/1, the principal product is a 1-amino-3-alkylaminopropane, $RNH(CH_2)_3NH_2$. Dialkylamines give 1-amino-3-dialkylaminopropanes.

Major producers of diamines, oligoamines, and polyamines of acrylonitrile include Huntsman and BASF.

The following liquid, water-miscible *N*-substituted diaminopropanes are industrially important. Their chief use is as components of epoxy resins:

1-Amino-3-methylaminopropane [*6291-84-5*], *N*-methyl-1,3-propanediamine, methylaminopropylamine, $C_4H_{12}N_2$, M_r 88.15, *bp* 140 – 141 °C (at 101.3 kPa), d_4^{20} 0.852, n_D^{20} 1.4479, flash point 35 °C, is a byproduct of the production of *N*-methyldipropylenetriamine and results from addition of methylamine to acrylonitrile followed by hydrogenation.

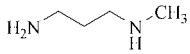

1-Amino-3-dimethylaminopropane [*109-55-7*], 3-dimethylamino-1-propylamine, DMAPA, *N,N*-dimethyl-1,3-diaminopropane, $C_5H_{14}N_2$, M_r 102.18, *mp* < – 60 °C, *bp* 133 °C (at 101.3 kPa), d_4^{20} 0.8112, n_D^{20} 1.4350, η 1.6 mPa · s (at 20 °C), flash point 32 °C, ignition temperature 305 °C (class T2), is a colorless liquid and a lachrymator.

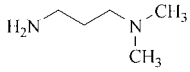

It is produced by addition of dimethylamine to acrylonitrile and subsequent hydrogenation in the presence of ammonia. DMAPA is used primarily for the synthesis of betaines which are employed as personal care products and for dish washing. Of minor importance is the consumption for the production of flocculants and ion exchangers.

1-Amino-3-diethylaminopropane [*104-78-9*], *N,N*-diethyl-1,3-propanediamine, *N,N*-diethyl-1,3-diaminopropane, *N,N*-diethylaminopropylamine, $C_6H_{18}N_2$, M_r 130.2, *mp* – 60 °C, *bp* 169 °C (at 101.3 kPa), d_4^{20} 0.829, n_D^{20} 1.4424, η 1.4 mPa · s (at 20 °C), flash point 53 °C is made by addition of diethylamine to acrylonitrile and subsequent hydrogenation. The market for this product is relatively small.

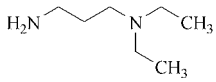

1-Amino-3-cyclohexylaminopropane [*3312-60-5*], *N*-cyclohexylpropylene-1,3-diamine, *N*-cyclohexyl-1,3-propanediamine, $C_9H_{20}N_2$, M_r 156.27, *mp* – 16 °C, *bp* 113 °C (at 1.3 kPa), d_4^{20} 0.9154, n_D^{20} 1.4810, η 9.4 mPa · s (at 20 °C), flash point 103 °C, ignition temperature 205 °C (class T3) is made by addition of cyclohexylamine to acrylonitrile and subsequent hydrogenation.

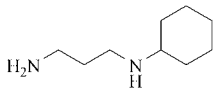

N,N,N',N'-Tetramethyl-1,3-propanediamine [*110-95-2*], $C_7H_{18}N_2$, M_r 130.23, *mp* < –70 °C, *bp* 145 – 146 °C (at 1.3 kPa), d_4^{19} 0.7837, n_D^{20} 1.4234, flash point 31 °C can be made by methylation of 1,3-diaminopropane with formaldehyde or from dimethylaminopropionitrile and dimethylamine and is an intermediate for corrosion inhibitors, textile auxiliaries, flotation agents, and emulgators. It is also employed as a catalyst for polyurethanes and as a hardener for epoxy resins.

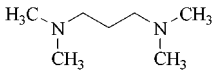

8.1.3. Higher Diamines

1,3-Diaminobutane [*590-88-5*], $C_4H_{12}N_2$, M_r 88.15, *bp* 138 – 141 °C (at 101.3 kPa), d_4^{20} 0.8536, n_D^{20} 1.4449, can be produced from crotonaldehyde by amination under hydrogenation conditions [258].

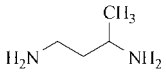

Diamines with four to five carbon atoms are important intermediates for pharmaceuticals. However, the production of these diamines is restricted by the fact that the starting materials are difficult to obtain and that there is a strong tendency for cyclization during preparation.

1,4-Diaminobutane [*110-60-1*], tetramethylenediamine, putrescine, $C_4H_{12}N_2$, M_r 88.15, *mp* 27 – 28 °C, *bp* 158 – 159 °C (at 101.3 kPa), d_4^{25} 0.8704, n_D^{20} 1.4569, flash point 63 °C, occurs as colorless crystals that are soluble in water and conventional organic solvents.

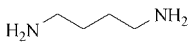

1,4-Diaminobutane can be produced from 1,4-dichloro-2-butene or 1,4-dihalobutane (obtainable from the reaction of butadiene or tetrahydrofuran with a halogen), preferably via the phthalimido compounds. Alternatively, 1,4-diaminobutane can be prepared by hydrogenating succinodinitrile [259].

1,5-Diaminopentane [*462-94-2*], pentamethylenediamine, cadaverine, $C_5H_{14}N_2$, M_r 102.18, *mp* 9 °C, *bp* 178 – 180 °C (at 101.3 kPa), d_4^{25} 0.867, n_D^{20} 1.4582, flash point 62 °C, can be prepared by methods similar to those described for 1,4-diaminobutane. The starting materials are 1,5-dichloropentane (obtainable from tetrahydropyran), glutarodinitrile, or glutaraldehyde (e.g., Relugan, BASF). The compound is hygroscopic and has a very bad odor.

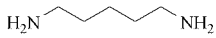

1,8-Diaminooctane [*373-44-4*], octamethylenediamine, $C_8H_{20}N_2$, M_r 144.26, *mp* 52 – 53 °C, *bp* 225 – 226 °C (at 101.3 kPa), flash point 165 °C is hygroscopic. It is used for the production of fungicides and is obtained by hydrogenation of suberonitrile (1,6-dicyanohexane).

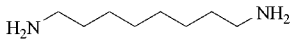

1,10-Diaminodecane [*646-25-3*], decamethylenediamine, $H_2N(CH_2)_{10}NH_2$, $C_{10}H_{24}N_2$, M_r 172.31, *mp* 61 – 62 °C, *bp* 140 °C (at 1.6 kPa).

1,12-Diaminododecane [*2783-17-7*], $H_2N(CH_2)_{12}NH_2$, $C_{12}H_{28}N_2$, M_r 200.37, *mp* 66 – 67 °C, *bp* 304 °C (at 101.3 kPa), flash point 155 °C can be used as a spacer in affinity chromatography.

Diamines containing more than six carbon atoms can be obtained readily by hydrogenating the corresponding dinitriles over a Raney nickel catalyst in a batch procedure under pressure at up to 125 °C. A five- to tenfold excess of liquid ammonia is used, with or without methanol solvent [260].

Higher diamines can be prepared continuously, as described for the production of hexamethylenediamine. For methods of preparing specific higher diamines, see [3].

2-(Diethylamino)ethylamine [*100-36-7*], *N,N*-diethylethylenediamine, 1-amino-2-(*N,N*-diethylamino)ethane, *N,N*-diethyl-1,2-ethanediamine, diethylaminoethylamine,

$C_6H_{16}N_2$, M_r 116.21, *mp* < −70 °C, *bp* 146 °C (at 101.3 kPa), d_4^{20} 0.8222, n_D^{20} 1.4360, flash point 30 °C.

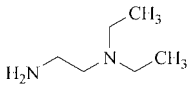

1-Diethylamino-4-aminopentane [*140-80-7*], *N,N*-diethyl-1,4-pentanediamine, novoldiamine, $C_9H_{22}N_2$, M_r 158.29, *bp* 191 °C (at 101.3 kPa), d_4^{20} 0.821, n_D^{20} 1.4442, flash point 68 °C, is a colorless liquid that is soluble in water and the usual organic solvents.

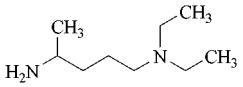

The diamine is produced by a conventional method, for example, by aminating 1-diethylaminopentan-4-one under hydrogenation conditions, by hydrogenating the oxime, or by aminating 1-diethylaminopentan-4-ol. The ketone is obtained from 2-chloroethyldiethylamine by synthesis of the acetoacetate, and 1-diethylaminopentan-4-ol is obtained by Mannich condensation of diethylamine, formaldehyde, and 1-butyn-3-ol followed by hydrogenation of the product, or is prepared from 1-methyltetrahydrofuran [261], [262]. Producers of 1-diethylamino-4-aminopentane are BASF and some Chinese producers. Derivatives of 1-diethylamino-4-aminopentane are particularly important as antimalarial agents [262] (e.g., chloroquine from Bayer).

3-Aminomethyl-3,5,5-trimethylcyclohexylamine [*2855-13-2*], isophorone diamine, IPDA, (5-amino-1,3,3-trimethyl-1-cyclohexylmethyl)amine, 1-amino-3-aminomethyl-3,5,5-trimethylcyclohexane, $C_{10}H_{22}N_2$, M_r 170.30, *mp* 10 °C, *bp* 247 °C (at 101.3 kPa), d_4^{20} 0.922, n_D^{20} 1.4880, flash point 112 °C.

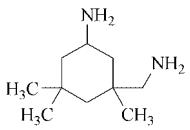

This specialty diamine is produced by Degussa-Hüls and BASF. It is made from isophorone nitrile, ammonia, and hydrogen, which leads to simultaneous imination and hydrogenation of the resulting imine and nitrile groups [263]. Catalysts employed are mainly cobalt or ruthenium. Isophorone diamine is employed as a hardener for epoxy resins and as an intermediate for isophorone diisocyanate; its market is considerable. Very important for the applications is the *cis*/*trans* isomer ratio, which should be as high as possible, since the trans isomer reacts more sluggishly.

***N,N,N′,N′*-Tetramethyl-1,6-hexanediamine** [*111-18-2*], *N,N,N′,N′*-tetramethylhexamethylenediamine, $C_{10}H_{24}N_2$, M_r 172.31, *bp* 209–210 °C (at 101.3 kPa), d_4^{20} 0.7981, n_D^{20} 1.4363, flash point 73 °C, can be made from adipodinitrile and dimethylamine or from 1,6-hexanediamine and formaldehyde and is used for instance as a catalyst for polyurethanes and hardener for epoxy resins.

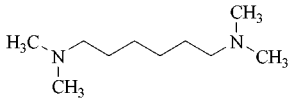

8.2. Oligoamines and Polyamines

8.2.1. Physical Properties

Diethylenetriamine [*111-40-0*], DETA, bis(2-aminoethyl)amine, 2,2′-diaminodiethylamine, 3-azapentane-1,5-diamine, $C_4H_{13}N_3$, M_r 103.17, is a colorless liquid, *mp* −39 °C, *bp* 207 °C (at 101.3 kPa), d_4^{20} 0.9542, n_D^{20} 1.4859, η 7 mPa · s (at 20 °C).

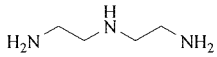

Vapor pressure

t, °C	20	89	174
p, kPa	0.049	1.333	40.0

Flash point	102 °C
Ignition temperature	330 °C (class T2)
Specific heat capacity	3.44 J g^{-1} K^{-1} (at 190 °C) 3.48 J g^{-1} K^{-1} (at 200 °C) 3.53 J g^{-1} K^{-1} (at 210 °C)
Heat of vaporization	456.4 J/g (at 0.1 MPa) 485.7 J/g (at 40 kPa)
Dissociation constant	7.07×10^{-5}
Surface tension	43.8 mN/m (at 25 °C) 40.8 mN/m (at 50 °C)
Coefficient of expansion	9.1×10^{-4} (at 55 °C).

The amine is infinitely miscible with water, methanol, acetone, ether, and benzene, but insoluble in heptane. Because of its reactivity, CCl_4 cannot be used as solvent. DETA is hygroscopic and has a strong ammoniacal odor.

Triethylenetetramine [*112-24-3*], TETA, *N,N′*-bis(2-aminoethyl)ethylenediamine, 1,8-diamino-3,6-diazaoctane, 1,4,7,10-tetraazadecane, $C_6H_{18}N_4$, M_r 146.24, is a colorless to yellowish oily viscous liquid, *mp* − 35 °C, *bp* 277 °C (at 101.3 kPa), d_4^{20} 0.9818, n_D^{20} 1.4986, η 26.7 mPa · s (at 20 °C).

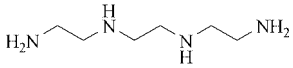

Vapor pressure

t, °C	20	144	240
p, kPa	0.001	1.33	40.0

Flash point	135 °C
Heat of vaporization	374.7 J/g (at 0.1 MPa) 397.8 J/g (at 40 kPa)
Dissocation constant	6.7×10^{-5}

TETA is hygroscopic, corrosive, and has a strong ammoniacal odor. With water a crystalline hydrate is formed. Like DETA, it is completely miscible with water and many polar organic solvents, but less so with lipids; with CCl_4 a violent reaction occurs. Its four pK_a values are 3.32, 6.67, 9.20, and 9.92. Technical-grade TETA is sometimes available as a distillation cut that also contains branched isomers and cyclic compounds.

Tetraethylenepentamine [*112-57-2*], TEPA, 1,4,7,10,13-pentaazatridecane, $C_8H_{23}N_5$, M_r 189.30, is a slightly yellowish viscous liquid, *mp* − 40 °C, *bp* 340 °C (at 101.3 kPa), d_4^{20} 0.9994, n_D^{20} 1.505, η 96.2 mPa · s (at 20 °C). Technical-grade TEPA may be available as a distillation cut that also contains branched isomers and cyclic compounds.

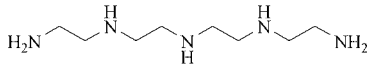

Vapor pressure

t °C	20	195	300
p, kPa	0.0013	13.5	40

Dissociation constant	6.5×10^{-5}
Flash point	163 °C

Tetraethylenepentamine is infinitely miscible with water, methanol, acetone, benzene, and diethyl ether but is insoluble in heptane and reacts violently with CCl_4.

Pentaethylenehexamine [*4067-16-7*], PEHA, $C_{10}H_{28}N_6$, M_r 232.37, *mp* − 26 °C, *bp* 136 – 144 °C (at 0.02 kPa), d_4^{20} 1.0020, flash point 186 °C, commercially is of minor importance. Technical-grade PEHA may be available as a distillation cut that also contains branched isomers and cyclic compounds.

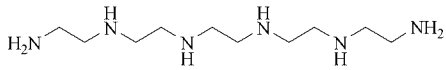

Dipropylenetriamine [*56-18-8*], 1-amino-3-(3-aminopropyl)aminopropane, bis(aminopropyl)amine, $C_6H_{17}N_3$, M_r 131.22, *mp* − 15.1 °C, *bp* 238 °C (at 101.3 kPa), d_4^{20} 0.9386, n_D^{20} 1.4846, flash point 118 °C, vapor pressure < 0.001 kPa at 20 °C, dissociation constant 3.96×10^{-4}, η 9.6 mPa · s (at 20 °C), is miscible with water, methanol, and conventional organic solvents.

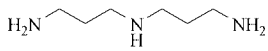

***N*-Methyldipropylenetriamine** [*105-83-9*], 1-amino-3-*N*-methyl-*N*-(3-aminopropyl)-aminopropane, $C_7H_{19}N_3$, M_r 145.25, *mp* − 32 °C, *bp* 234 °C (at 101.3 kPa), d_4^{20} 0.9040, n_D^{20} 1.4732, η 7.4 mPa · s (at 20 °C), flash point 107 °C, ignition temperature 250 °C (class T3),

is miscible with water and conventional organic solvents.

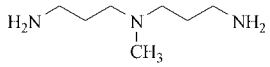

3-(2-Aminoethyl)-aminopropylamine [*13531-52-7*], *N*-(2-aminoethyl)-1,3-diaminopropane, $C_5H_{15}N_3$, M_r 117.19, *mp* −10 °C, *bp* 80 °C (at 0.3 kPa), d_{20}^{20} 0.9401, n_D^{20} 1.4805, flash point 98 °C.

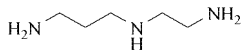

***N,N′*-Bis(3-aminopropyl)-1,2-diaminoethane** [*79554-59-9*], *N,N′*-bis(3-aminopropyl)ethylenediamine, $C_8H_{22}N_4$, M_r 174.3, *mp* −1.5 °C, *bp* 170 °C (at 0.3 kPa).

$H_2N(CH_2)_3NHCH_2CH_2NH(CH_2)_3NH_2$

8.2.2. Chemical Properties

Because they possess terminal primary amino groups, condensates of 1,2-diaminoethane react very much like 1,2-diaminoethane itself. The higher condensates of 1,3-diaminopropane have similar properties. Dibasic acids react with oligoamines and polyamines to give polyamides [264]. A number of urea compounds can be obtained from oligoamines, e.g., diethylenetriamine, or from polyamines [265]. These amines react with aliphatic dihalides to give water-soluble cationic products [266] and with epichlorohydrin to give anionic polymers [267].

In a manner similar to diaminoethane, diethylenetriamine reacts with formaldehyde and sodium cyanide to give diethylenetriaminepentaacetic acid, which also has complex-forming properties. Diethylenetriamine undergoes also a smooth cyanoethylation reaction [235]. The oligoamines react directly with formaldehyde to give permethylated products. At elevated temperatures, diethylenetriamine can be cyclized to piperazine; if the temperature is about 400 °C, the cyclization reaction gives pyrazine directly [268].

8.2.3. Production, Analysis, and Uses

Diethylenetriamine is obtained as a byproduct in the synthesis of 1,2-diaminoethane from dichloroethane (ethylene dichloride, EDC), ethylene oxide, or ethanolamine and ammonia. Triethylenetetramine, tetraethylenepentamine, and pentaethylenehexamine are other byproducts from the EDC process for 1,2-diaminoethane. More selective production methods have been described as well, in particular for DETA, mostly starting with EDA [269].

A number of polyamines are obtained by cyanoethylation of diaminoethane or a diaminopropane with acrylonitrile, followed by hydrogenation of the product [133]. For instance, dipropylenetriamine is produced as byproduct by hydrogenation of the addition product of ammonia and acrylonitrile [256]; it is used as curing agent for epoxy resins. *N*-Methyldipropylenetriamine is obtained from methylamine and acrylonitrile.

The regulations governing storage and transportation of the oligoamines and polyamines are the same as those for 1,2-diaminoethane. For storage of DETA, TETA, and TEPA, plastic-lined, stainless steel or aluminum containers are best suited. Carbon steel leads to iron contamination. A nitrogen atmosphere without carbon dioxide is necessary for DETA. TETA and TEPA are sometimes stored under air, but nitrogen is superior. Purity is determined preferably by titration, and analysis is performed by fractional distillation. The oligoamines, e.g., diethylenetriamine, can be also analyzed by gas chromatography, but this method is unsuccessful for the higher polyamines, owing to their low vapor pressures.

Many uses of the higher acyclic ethyleneamines are similar to those of EDA itself, the most important being lubricant oil additives, polyamide resins, epoxy curing agents, surfactants, and oil field chemicals [231]. Polyisobutenylsuccinimides derived from TETA, TEPA, or PEHA are important motor oil additives. Polyamide resins produced from DETA or TETA with adipic acid are treated with epichlorohydrin and used as wet-strength additives in paper manufacture. Polyamine-modified urea – formaldehyde resins are also used as wet-strength additives for cellulose fibers. DETA and

TETA or their ethoxylation products are used as curing agents for epoxy resins. Polyamine amides possessing terminal primary amino groups, which are prepared from fatty acids and polyamines, serve the same purpose.

Amidoamines obtained from the reaction of a fatty acid and a polyamine, in particular DMAPA and DETA, have similar uses to the more expensive fatty amines, e.g., as asphalt emulsifiers, flotation agents, and corrosion inhibitors. If the solid amidoamines are further heated to 200 – 240 °C, cyclization to imidazoline derivatives is induced. Since these are liquids they posses handling advantages over the amidoamines; the uses are similar, and the polyamines mostly used for production are DETA and TETA. Polyamine derivatives are also used for textile finishing. The most important DETA applications are chelating agents (diethylenetrinitrilopentaacetic acid) and wet-strength paper additives. The latter application is increasing since the paper industry is replacing melamine – formaldehyde resins with polyamide – epichlorohydrin resins which operate under neutral pH conditions.

9. Chiral Amines

With the trend in modern pharmaceutical and agricultural chemistry towards enantiomerically pure compounds, amines, being important intermediates in both areas, came into demand in their chiral forms. Some can be taken from the chiral pool and others made from it, but a growing proportion of the market for these specialities is being synthesized. Although enantiomeric syntheses or resolutions are sometimes available [270], they tend to use very expensive reagents, so that enzyme-catalyzed processes offer a significant economic advantage. Enzymatic resolution of a racemic mixture affords a pure enantiomer, albeit with loss of the other enantiomer. Both enantiomers are obtained in pure form by enantioselective acylation, after which the (*R*)-amide and the unchanged (*S*)-amine are separated by simple distillation [271], [272]. Formation of diastereomeric salts by neutralizing a racemic amine with an optically active acid and subsequent fractional crystallization can also be used for optical resolution [273].

The chiral amines can be directly incorporated in pharmaceuticals or used as chiral auxilaries in synthesis. For instance, (*S*)-1-phenylethylamine has been used for the optical resolution of the analgesic Naproxen [274].

10. Toxicology and Occupational Health

10.1. General Aspects

The aliphatic amines are highly irritating or corrosive. In their nonprotonated form they are lipid soluble and therefore penetrate rapidly into the lower layers of the exposed tissue. Even in low concentration the vapors cause swelling and damage in the mucous membranes of the eye and the respiratory tract.

Acute Toxicity and Local Effects. Generally, the LD_{50} (rat, oral or dermal) is on the order of 100 – 1.000 mg/kg [275]. The symptoms are caused mainly by high penetrability, corrosion, alkalosis, and, in some cases, by functional interference with nerve-impulse conduction. The LD_{50} values of the hydrochlorides are usually higher because they do not cause corrosion and alkalosis: 1600 – 3200 mg/kg (rat, oral) [275]. Acute toxic values, local effects on exposed skin and eyes, and occupational exposure limits in Germany [276] and the United States [277] are compiled in Table 7.

Irritation and corrosion of the skin and mucous membranes were observed frequently in humans [275], [316], also on exposure to the vapor [317]. Accidental inhalation of large amounts can cause headache, nausea, excitation, and short convulsions [275]. Some aliphatic amines, especially di- and polyamines, can cause sensitization of the skin and the respiratory tract (allergies) [275], [318].

β-Chloroalkylamines have an alkylating potential in addition to their corrosiveness. Therefore, they are suspected carcinogens.

Systemic Toxicity. In neutral media the amines are almost completely protonated to the ammonium compounds. Therefore, interaction with the cation channels of the nerve-cell membranes is possible. This may result in a change of permeability for physiologically important

Table 7. Survey of acute toxicological data and occupational exposure limits of aliphatic amines

Substance	CAS no.	Acute toxicological data						Occup. exposure limits [d] [276], [277]	
		LD_{50} [a] mg/kg	Source	LC_{50} (inhal.) [b] g/m^3	Source	Local irritation [c]	Source	MAK (1998), mL/m^3	TLV (1997), mL/m^3
Ethylamine	[*75-04-7*]	400 – 800 (o, rt)	[278], [279], [280]	12.6 (4 h, 20 – 22 °C, rt)	[281]	c	[282]	10	10
Diethylamine	[*109-89-7*]	450 – 650 (o, rt, ms) 580 – 820 (d, rbt)	[282], [283] [287], [288]	12.1 (4 h, 20 – 22 °C, rt)	[283], [284]	c	[283], [285], [286]	5	5
Triethylamine	[*204-44-8*]	460 – 590 (o, rt, ms)	[283], [288], [289], [282], [290]	10.9 (4 h)	[291]	c	[292], [293]	1	1
Dipropylamine	[*142-84-7*]	200 – 690 (o, rt) 925 (d, rbt)	[294], [282]	ca. 4.2 (4 h)	[294]	c	[295]	n.e.	n.e.
Isopropylamine	[*75-31-0*]	120 – 820 (o, rt)	[283], [296], [281]	> 9.8 to < 19.6 (4 h)	[283], [296]	c	[283], [297]	5	5
Cyclohexylamine	[*108-91-8*]	710 (o, rt) 320 (d, rbt)	[298]	> 16 (4 h)	[299]	c	[275]	10	10
4,4'-Diaminodicyclohexylmethane	[*1761-71-3*]	300 – 1000 (o, rt, ms, rbt)	[281], [300], [301]		-	c	[281], [300], [302]	n.e.	n.e.
Piperazine	[*110-85-0*]	ca. 2200 (o, rt)	[281]	no mortality (8 h, satur.)	[281]	c	[281]	n.e.	5 mg/m^3 (as HCl)
1-(2-Aminoethyl)piperazine	[*140-31-8*]	1500 – 2100 (o, rt) ca. 870 (d, rbt)	[281], [294]	no mortality (8 h, satur.)	[281]	c	[281], [294]	n.e.	n.e.
Pyrrolidine	[*123-75-1*]	300 – 450 (o, rt)	[303]	> 4.5 – 11.7 (4 h, rt) 5.9 (o, rt); 1.3 (2 h, ms)	[281] [306]	c	[304], [305]	n.e.	n.e.
Ethylenediamine	[*107-15-3*]	ca. 700 – 1400 (o, rt)	[275], [281]	> 5 (6 h)	[281]	c	[275], [281]	10	10
3-Aminopropyl dimethylamine	[*109-55-7*]	930 – 1800 (o, rt)	[307]	> 4.3 (4 h) (irritating)	[281]	c	[281]	n.e.	n.e.
Diethylenetriamine	[*111-40-0*]	≥ 1080 (o, rt) ca. 700 – 1000 (d, rbt)	[275], [308], [281] [309]	no mortality (8 h, satur.)	[281], [309]	c	[281]	n.e.	1
Triethylenetetramine	[*112-24-3*]	ca. 2800 – 4940 (o, rt) 550 – 820 (d, rbt)	[298], [310]	no mortality (8 h, satur.)	[310]	c	[310]	n.e.	n.e.
Hexamethylene imine	[*111-49-9*]	ca. 350 (o, rt)	[281]	-	-	c	[281]	n.e.	n.e.
Hexamethylene tetramine	[*100-97-0*]	9200 (o, rt) 1850 (o, ms)	[311]	-	-	no, weak	[312]	n.e.	n.e.
Morpholine	[*110-91-8*]	1050 – 1600 (o, rt) ca. 500 – 900 (o, ms)	[313], [314] [315]	-	-	c	[313], [314] [315]	20	20

[a] rt = rat, ms = mouse, rbt = rabbit, o = oral, d = dermal. [b] satur. = saturated atmosphere, rt = room temperature. [c] c = corrosive. [d] n.e. = not established

cations and thus in inhibition of impulse conduction along the nerve fibers. The essential structural element of most local anesthetics is a tertiary amino group. If significant amounts of these substances are present in the blood, effects adverse to the impulse conduction of the heart and especially to the inhibitory neurons of the CNS result. The clinical symptoms range from excitation to clonic convulsions. Consequently, excitation of the CNS with increased blood pressure and short spasms were observed along with irritation and injury to the respiratory tract on accidental inhalation of large amounts of aliphatic amines.

Long-chain mono- and dialkylamines are potential histamine liberators and may produce reddening and edema of the skin and mucous membranes, itching, decreased blood pressure, tachycardia, and bronchioconstriction [275], [319]; the maximum effect occurs with C_{10} monoamines and C_{14} diamines. Certain aliphatic di- and polyamines, such as putrescine [*110-60-1*] ($H_2N(CH_2)_4NH_2$), spermidine [*124-20-9*] ($H_2N(CH_2)_4NH(CH_2)_3NH_2$), and spermine [*71-44-3*] ($H_2N(CH_2)_3NH(CH_2)_4NH(CH_2)_3$-$NH_2$), are found characteristically in proliferating cells and are formed under the influence of hormones, natural growth stimuli, specific tumor promoters, and other factors. Their physiological action is apparently involved in cell division and tissue proliferation [320].

Metabolism. In many cases, little is known about the metabolism of alkylamines. Generally, once absorbed, alkylamines do not accumulate and are rapidly released from the organism, mostly into the urine, either unchanged (e.g., diethylamine, cyclohexylamine) or metabolically converted. Reactions catalyzed by the cytochrome P450 system include the oxidative deamination of primary alkylamines and *N*-dealkylation of secondary and tertiary amines [321]. This enzyme system competes with an unspecific flavin-containing monooxygenase that produces *N*-oxides from tertiary amines and hydroxylamines from secondary amines. Furthermore, *N*-methyltransferases may alkylate mono-, di-, and trialkylamines [321]. In addition, by monoamine oxidase, primary, secondary, and tertiary amines can be oxidatively dealkylated stepwise to aldehydes (or carboxylic acids), ammonia, and hydrogen peroxide [321]. The oxidation rate increases with increasing chain length; methylamine is not transformed by monoamine oxidase [322], [323]. Short-chain mono- and diamines are also oxidized by diamine oxidase (histaminase); those most readily oxidized are 1,4-butanediamine and the $C_5 - C_6$ monoamines.

Secondary amines (including those with cyclic structures), and to some extent tertiary amines, can be chemically transformed into carcinogenic *N*-nitrosodialkylamines in the presence of nitrites or nitrates under reducing conditions.

10.2. Toxicology of Specific Amines

10.2.1. Alkylamines, Cyclic Amines, and Polyamines

Ethylamine [*75-04-7*], alkaline liquid, in neat form gaseous at room temperature, is readily absorbed through the skin and respiratory tract [282], but easily excreted in the urine after partial metabolism [324]. Exposure to the vapor may transiently impair vision in humans ("blue haze") [325]. The odor threshold in air is ca. $0.027 - 0.27$ mL/m^3 (≤ 0.5 mg/m^3) [326], [327]. In rabbits, inhalation of 50 mL/m^3 (94 mg/m^3) for 6 weeks resulted in local and systemic organ lesions (corneal erosions and edemas, pneumonia, peribronchitis, focal muscular degeneration of the heart) [285]. Transient physiological effects were seen after prolonged exposure of rats to vapors ($0.06 - 3.69$ mg/m^3, 90 d) with no effects at 0.015 mg/m^3. At the highest concentration, morphological changes in the lung and cerebral neurons apparently were reversible [326].

Diethylamine [*109-89-7*], alkaline liquid, once absorbed, is rapidly excreted unchanged in the urine [328]. Exposure to low vapor concentrations are reported to transiently impair vision in humans (reduction of sensitivity to light): no-effect threshold $0.022 - 0.028$ mL/m^3 (≤ 0.08 mg/m^3) [326], [329]. The odor threshold in air is ca. $0.02 - 0.3$ mL/m^3 (≤ 0.9 mg/m^3) [329], [330]. Transient physiological effects in rats after 90-d exposure to vapors were reported to already be visible at $0.37 - 4.19$ mg/m^3

with no effects at 0.05 mg/m^3, but with morphological, apparently reversible changes in the lung and cerebral neurons at the highest concentration [326]. On the other hand, 50 mL/m^3 (150 mg/m^3) was the lowest concentration to show mild, but significant effects in the lung and liver of rabbits on prolonged inhalation [285]. There were no visible signs of toxicity following exposure of rats to 25 mL/m^3 (75 mg/m^3) for 120 d, but a moderate, bronchiolar hyperplasia of lymphoid cells [286]. It was assumed that diethylamine may exert a certain neurotoxic effect, which, however, was not confirmed by others [286].

Under experimental conditions, no sensitizing potential was observed [331]. Diethylamine itself showed no mutagenic potential. Although the carcinogenic *N*-nitrosodiethylamine is formed in the presence of nitrite, combined administration in the feed and drinking water gave no evidence of increased tumor incidence in a lifelong study with rats [332].

Triethylamine *204-44-8*, oily, alkaline liquid, once absorbed, is rapidly excreted unchanged or as its *N*-oxide in the urine [333], [334]. Exposure to the vapor may transiently impair vision in humans ("blue haze"): concentrations without visual effect are in the range of 1.4 – 2.7 mL/m^3 [293], [335]. Its exciting effect on the CNS is more pronounced than that exerted by the other lower ethyl homologues; effects on parasympathetically controlled musculature and inhibition of nerve pulse transmission in sypmpathetic ganglia can be explained by inhibition of cerebral monoamine oxidase [292], [335]; however, triethylamine does not inhibit neuromuscular impulse transmission [292]. Under experimental conditions and in humans, no sensitizing potential was observed [335]. The odor threshold in air is reported from < 0.09 to 1.4 mL/m^3 (≥ 0.33 mg/m^3) [326], [335].

Dipropylamine [*142-84-7*], alkaline liquid, may cause headache, nausea, faintness, and lung edema on inhalation. The odor threshold in air is 0.02 – 0.1 mL/m^3 (≤ 0.4 mg/m^3) [295].

Isopropylamine [*75-31-0*], alkaline liquid: Signs of respiratory irritation may occur at about 24 mg/m^3 and above (≥ 10 mL/m^3) in humans [297]. In rats, a NOAEL was defined as 100 mg/m^3 for prolonged inhalation (1 month) [336]. The substance was void of any teratogenic effects after exposure of pregnant rats to up to 1000 mg/m^3 [337]. An odor threshold in air is reported to be between 5 – 10 mL/m^3 (≤ 24 mg/m^3) [297].

Cyclohexylamine [*108-91-8*], alkaline liquid, is readily absorbed by the oral and respiratory route, but only slowly through the skin, and is rapidly excreted in the urine almost unchanged [338]. It has a strong fishlike odor (odor threshold ca. 2.5 mL/m^3 = 10.4 mg/m^3) [299], [338]. First signs of a sympathomimetic effect in humans (elevation of the arterial blood pressure) are reported at a daily dose of 5 mg/kg [338].

Cyclohexylamine has been studied intensively because it is a metabolite of the sweetener saccharin. Intake of either up to 300 mg/kg of the HCl salt daily in the feed of rats for two years and lifelong [339–341] or of up to about 750 mg/kg daily in the feed of mice (0.3 and 0.5 % HCl salt) for ≥ 80 weeks [342], [343] did not lead to higher incidence of tumors. Accelerated basal metabolic rate and reduced spermatogenesis were observed in rats after 90 d [344] at daily doses of 2000 and 6000 ppm (HCl salt), but no such symptoms were evident at 600 ppm (ca. 20 mg amine per kilogram) daily intake. In other long-term feeding studies, no convincing evidence of toxicity was apparent at doses of 5 to ca. 20 mg kg^{-1} d^{-1} [299].

In a multigeneration study (two years, 150 mg/kg daily of cyclohexylammonium chloride), the fertility of rats was not affected adversely in spite of slight atrophy of the testicles [341]. Based on these and other results [345], [346], it is concluded that no mutagenicity, carcinogenicity, or teratogenicity is suspected for cyclohexylamine [347], [338].

4,4′-Diaminodicyclohexylmethane [*1761-71-3*], alkaline solid, was shown to have skin-sensitizing potential [302], [348]. On resorption, predominant clinical signs of intoxication were hyperexcitability, excessive salivation, and convulsions in animals; repeated oral dosing (50 – 100 mg kg^{-1} d^{-1} in rats and mice for 10 and 16 d, respectively, 50 mg/kg in dogs for up to 82 weeks) resulted in gastrointestinal hemorrhagic inflammations, nephritis, and fatty changes to the liver [281], [300], [348]. No

changes in methemoglobin or in other hematological parameters were induced in cats on repeated doses of 50 and 100 mg kg^{-1} d^{-1}, given twice per week over 2 months [281], [301]. Several in vitro [281], [349] and in vivo mutagenicity studies [281], [301] gave no evidence of a mutagenic or clastogenic potential.

Piperazine [*110-85-0*], alkaline solid, is almost completely absorbed from the intestinal tract on oral ingestion. Acute intoxications occasionally follow therapeutic treatment of animals with piperazine salts (as anthelmintics), and symptoms include adverse neurologic events such as ataxia, muscle tremor, muscular weakness, and vomiting [350]. In humans, headache, nausea, coordination disorders, and exanthems have been observed occasionally during repeated intake of piperazine salts in daily doses of 30 – 75 mg/kg [351]. Apparently, sensitizations of the skin and also the respiratory tract can occur; these are also caused by the dihydrochloride [352–354]. Based on several observations, especially in workers in Germany, piperazine was included in the list of allergens which are also suspected of inducing respiratory sensitization, but no MAK value was established [355], [356].

Although experimental data suggest that piperazine itself is unlikely to be mutagenic and carcinogenic, there is strong evidence that the combination with nitrite produces malignant tumors in animals: for reviews, see [357], [358].

1-(2-Aminoethyl)piperazine [*140-31-8*], alkaline liquid: Sensitization was observed in animal experiments [359].

Pyrrolidine [*123-75-1*], alkaline liquid: As with other amines, sympathomimetic action is observed and results in increased blood pressure and a tendency to convulsions [304]. In an inhalation study on male rats [360], exposure to 2.6 mg/m^3 for 6 months caused increasing excitation of the nervous system, inhibition of diuresis and spermatogenesis without impairment of spermatozoon mobility and fertility; however, morphological alterations were evident from an increase in the number of tubuli seminiferi with desquamation of the seminiferous epithelium. Experimental data suggest that pyrrolidine itself is unlikely to be mutagenic [303] and carcinogenic: in limited long-term studies in rats and mice in the presence of nitrite in drinking water to favor nitrosation, there was no evidence of a carcinogenic effect [361], [362].

Ethylenediamine [*107-15-3*], alkaline liquid: Dermatitis related to primary irritation and also to sensitization occurs frequently [275]. Sensitization of the mucous membranes of the respiratory tract with asthmatic symptoms has also been described [363], [364]. After oral administration in rats and mice at up to 1000 mg kg^{-1} d^{-1} of the dihydrochloride salt, no teratogenic effects were produced [365–367], and there was no evidence for an impairment of reproduction in a two-generation study [365]: 250 mg kg^{-1} d^{-1} dihydrochloride salt (ca. 112 mg kg^{-1} d^{-1} base) can be considered the NOAEL [368]. Several in vitro tests gave no evidence for a mutagenic potential [369–371].

3-Dimethylamino-1-propylamine [*109-55-7*], alkaline liquid: There are indications of contact and possibly respiratory allergy in relation to occupational exposure [307]. Less than 0.9 ppm of 3-dimethylaminopropylamine in the workplace air may adversely affect respiratory function [372]. The skin-sensitizing potential was verified in an animal study. In an oral 28-d study with rats (gavage, 7 times per week.), clinical symptoms including pulmonary and cardiac failure were noted at 250 mg kg^{-1} d^{-1}, but no treatment-related effects at a level of 50 mg kg^{-1} d^{-1} [307].

Diethylenetriamine [*111-40-0*], alkaline liquid: sensitization of the skin and possibly respiratory tract occurs frequently [275], [309]. Experience from occupational medicine suggests that the latter has to be considered the main adverse effect that may occur in humans [309]. Following oral administration, diethylenetriamine is readily absorbed and rapidly excreted, mainly via the urine and feces [309]. In rats, no histopathological changes were observed after 90-d feeding up to the highest dietary level of 15 g/kg (ca. 1200 mg kg^{-1} d^{-1}), [373] while in cats and rabbits, repeated oral doses of ca. 200 mg kg^{-1} d^{-1} and above as base or HCl salt resulted in severe gastritis, lung edema or pneumonia, and kidney lesions [281], [309].

Triethylenetetramine [*112-24-3*], alkaline liquid: may induce lung edema on inhalation of its vapors. The compound is also an effective skin sensitizer [275]. Following repeated dosing (ca. 50 mg each, 17 – 55 times) onto the skin of pregnant and nonpregnant guinea pigs, significant amounts were absorbed by the strongly irritated skin, leading to toxic effects in the kidneys, liver, brain, and placenta and causing abortion [310], [374], [375]. Triethylenetetramine proved to be a mutagen in vitro [376] but not in vivo [310].

Hexamethyleneimine, hexahydro-1*H*-azepine [*111-49-9*], for acute toxicity data: see Table 7.

Hexamethylenetetramine [*100-97-0*], HMTA, urotropin, is rapidly absorbed from the gastrointestinal tract, and the majority is excreted unchanged in the urine [311].

Irritation and Sensitization. In animals, HMTA exhibited a skin-sensitizing potential [377]. Several reports in industry [378–380] suggest an association between HMTA exposure and the increased occurrence of irritative, allergic events among workers: this was often encountered with hot processing in the rubber and foundry industries. Sensitized persons frequently showed positive responses on patch testing, although the causative agent remained obscure [379], [381]. They can suffer from asthmatic or asthmalike attacks on inhaling HMTA [382].

Systemic Toxicity. Experience from therapeutic use in the past, especially for the prevention of recurrent urinary tract infections in human patients, support the low toxicity on prolonged oral ingestion of up to 4 g/d per individual. Some 1 – 7 % of the patients experienced untoward events such as gastrointestinal intolerance (nausea, vomiting, abdominal cramps) and unspecific skin reactions (pruritus, urticaria, erythematous eruptions) [383]. Chronic animal studies with HMTA administered in the diet or drinking water did not reveal any evidence of a carcinogenic potential and account for a high chronic NOAEL of more than 1000 mg kg^{-1} d^{-1} [384–386].

Mutagenicity. The majority of studies failed to reveal any in vitro or in vivo genotoxic activity. At cytotoxic or sublethal doses ($\leq$ 10 g/kg) [311], a weak mutagenic effect was found in male mice [387], but there was no positive response at up to 1500 mg/kg [388], while a cytogenetic assay after single oral and repeated doses of one-third of the LD_{50} in mice was clearly negative [389]. Furthermore, the compound is described as mutagenic for larvae of *Drosophila* (dose not specified) [382].

Teratogenicity and Reproduction. In several multigeneration studies on rats, no treatment-related abnormalities or developmental defects in the offspring or reproductive defects in the parental generation were seen following high oral doses of up to more than 1500 mg kg^{-1} d^{-1} [311], [386]. In a study with female beagles (600 and 1250 ppm of HMTA in the feed (15 – 31 mg kg^{-1} d^{-1}), from the 4th to 56th day after mating), slight embryotoxic effects were found at higher doses but no malformations [390].

Morpholine [*110-91-8*], alkaline liquid, is well absorbed by the oral, dermal, and inhalation route, but rapidly excreted, mainly unchanged, in the urine. Experimental evidence suggests that morpholine is unlikely to induce skin sensitization [314], [391].

Common signs of toxicity following repeated dosing are local irritation and inflammations of the stomach, respiratory tract, and eyes, as well as systemic effects primarily on the liver and kidneys. In rats, exposure to 250 mL/m^3 (890 mg kg^{-1} d^{-1}, 6 h/d, 5 d/week, 90 d) [392] and to up to 150 mL/m^3 (543 mg/m^3, 6 h/d, 5 d/week, 104 weeks) [393], [313] produced focal erosions and squamous-cell metaplasia of the nasal cavities and turbinates and ocular irritation, but no hematological or organ effects; at 90 mg/m^3 (25 mL/m^3, subchronic) and 36 mg/m^3 (10 mL/m^3, chronic), no treatment-related effects at all were identified. These data may be taken as NOAELs, although one earlier Russian publication [394] claimed that some adverse effects were conspicuous on the spleen and in the red and white blood counts in rats and guinea pigs after four-month inhalation of 70 mg/m^3 and less (see also [313]).

After oral application, no toxic signs were found in mice receiving 70 – 140 mg kg^{-1} d^{-1} base in drinking water as morpholine oleate for 90 days [396].

In a subacute feeding test (four weeks, rats), a daily dose of 323 mg/kg led to increased weight of the suprarenal gland accompanied by retarded development of body weight; daily doses of 27 and 93 mg/kg were without effect [303], [313].

In isolated cases, morpholine showed a weak mutagenic acticity; however, most of the experimental in vitro and in vivo data confirmed that purified morpholine did not give any evidence of a mutagenic potential [281], [313], [314], [395]. Furthermore, longterm studies (in rats and hamsters with feed containing 0.1 % base [396], in mice with drinking water containing up to 0.24 % base [397], and inhalation of up to 530 mg/m^3 in rats [398]) failed to reveal an inherent cancerogenic potential of morpholine [313], [314]. However, there is a possibility of nitrosation in the presence of traces of nitrous acid or nitrogen oxides [303], [399]: *N*-nitrosomorpholine is a potent carcinogen [313], [395].

10.2.2. Fatty Amines

Skin and Eye Irritation. Fatty amines are local irritants of eyes, skin, and mucous membranes. Both the primary amines and the propanediamines are generally regarded as severely irritating and possibly corrosive. Even dilute solutions (5 %) of fatty amines may cause irreversible damage to the eye; cornea inflammation caused by 1 % solutions heals without permanent effects. Concentrated solutions of salts of fatty amines and quaternary ammonium compounds are almost neutral but they have an irritating potential similar to the free bases. Ethoxylation of fatty amines reduces the irritating effects.

A small number of workers became sensitized to selected fatty amines. Symptoms included skin rash, dermatitis, eye swelling, and a sensation of the skin, described as "crawling". Appropriate skin and eye protection must be worn when handling fatty amines.

Inhalation. At ambient temperatures only the short-chain amines, especially if branched, have sufficient vapor pressure to present any hazard by inhalation. The LC_{100} of 2-ethylhexylamine is 1.3 mg/L (rat, 4 h) [400], whereas exposure to air saturated with *N,N*-di(tridecyl)tridecanamine for eight hours resulted in no deaths [401].

Acute and Chronic Oral Toxicity. None of the fatty amines are highly toxic by ingestion. The more corrosive ones, such as coco-1,3-diaminopropane, have oral LD_{50} values as low as 147 mg/kg, whereas compounds of higher molecular weight, such as di(hydrogenated-tallow) dimethylammonium chloride, are relatively nontoxic, with oral LD_{50} values of 7000 mg/kg and higher [402].

Chronic feeding studies have shown that rats fed with 500 ppm of 1-octadecylamine in the diet for two years showed no observable adverse health effects, whereas dogs fed the same level for twelve months showed mainly nonpathologic irritant effects on the mucosa of the gastrointestinal tract [403].

11. References

1. A. Wurtz, *C. R. Hebd. Seances Acad. Sci.* **28** (1849) 223.
2. E. Fischer, *Ber. Dtsch. Chem. Ges. Sonderheft* 1902, 197.
3. *Houben-Weyl*, vol. **9/1.**
4. Air Products, EP 311 900, 1987 (M. Deeba).
5. A. Baiker, J. Kijenski, *Catal. Rev. Sci. Eng.* **27** (1985) 653.
6. A. Baiker in M. Guisnet et al. (ed.): *Heterogeneous Catalysis and Fine Chemicals,* Elsevier, Amsterdam 1988, 283.
7. J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive,* Thieme Verlag, Stuttgart 1978.
8. A. Baiker, *Chem. Ind. (Dekker)* **53** (1994) 91.
9. A. Fischer, T. Mallat, A. Baiker, *Catal. Today* **37** (1997) 167.
10. BASF, DE 2 838 184, 1980 (W. Schroeder, H. Toussaint, W. Franzischka).
11. A. Baiker, W. Richarz, *Ind. Eng. Chem., Prod. Res. Dev.* **16** (1977) 261.
12. Ruhrchemie, DE-OS 2 048 750, 1970 (H. Goethel, B. Cornils, H. Feichtinger, H. Tummes, J. Falbe).
13. J. Pašek, P. Richter, *Chem. Prum.* **19** (1969) 63.
14. W. S. Emerson in R. Adams: *Organic Reactions,* **vol. 4,** J. Wiley & Sons, New York 1948, p. 174.
15. ICI, DE 2 429 293, 1975 (T. Dewdney, D. Dowden, B. Hawkins, W. Morris).
16. Henkel, WO 98/19991, 1996 (A. F. Elsasser).

17. Erdölchemie, DE 2 839 134, 1978 (K. Mainusch, R. Schorsch, B. Schleppinghoff), DE 2 839 135, 1978 (K. Mainusch, B. Schleppinghoff, K. Munster).
18. Air Products, DE 19 602 679, 1996 (G. A. Vedage, J. N. Armor).
19. C. Barnett, *Ind. Eng. Chem. Prod. Res. Dev.* **8** (1969) 145.
20. LCE Partnership, US 5 196 589, 1993 (A. J. O'Lenick, J. M. Clumpner).
21. Engelhard Industries, GB 925 458, 1960.
22. Gulf Research & Development Co., US 3 366 686, 1964 (R. W. Rosenthal, R. Seekircher); Texaco, US 3 739 027, 1971 (W. C. Gates, Jr.).
23. Mitsui Petrochemical Ind., JP 10 095 753, 1996.
24. J. J. Ritter, J. Kalish, *J. Am. Chem. Soc.* **70** (1948) 4084.
25. Air Products, US 4 307 250, 1981 (J. Peterson, H. Fales), US 4 375 002, 1981 (J. Peterson, H. Fales).
26. Air Products, EP 77 016, 1982 (M. Deeba, W. J. Ambs).
27. BASF, DE 3 326 579 1983 (V. Taglieber, W. Hölderich, R. Kummer, W. D. Mross, G. Saladin). BASF, DE 3 327 000 1983 (V. Taglieber, W. Hölderich, R. Kummer, W. D. Mross, G. Saladin).
28. BASF, DE 19 524 241, 1995 (U. Dingerdissen, R. Kummer, P. Stops, J. Herrmann, H.-J. Lützel, K. Eller).
29. BASF, DE 19 530 177, 1995 (U. Dingerdissen, R. Kummer, P. Stops, U. Müller, J. Herrmann, K. Eller).
30. BASF, DE 19 545 875, 1995 (K. Eller, R. Kummer, P. Stops). BASF, DE 19 545 876, 1995 (K. Eller, R. Kummer, P. Stops).
31. D. M. Roundhill, *Catal. Today* **37** (1997) 155.
32. D. Steinborn, R. Taube, *Z. Chemie* **26** (1986) 349.
33. T. Rische, P. Eilbracht, *Synthesis* **11** (1997) 1331.
34. J. F. Knifton, J. J. Lin, D. A. Storm, S. F. Wong, *Catal. Today* **18** (1993) 355.
35. H. Kölbel, I. Abdulahad, M. Raleh, *Erdöl Kohle Erdgas Petrochem.* **28** (1975) no. 8, 385.
36. Stepan Co., WO 98/3 262, 1996 (M. J. Nepras, R. J. Bernhardt, C. J. Sporer).
37. *Europachemie* **1 – 2** (1994) 6.
38. *Hydrocarbon Process.* **60** (Nov. 1981) 146.
39. Daicel, EP 841 320, 1996 (Y. Tsuji).
40. FMC Corp., EP 857 694, 1997 (J. F. Kramer, F. O'Brien).
41. 3V Inc., US 5 759 525, 1996 (A. Malpede, G. Raspanti).
42. BASF, EP 695 572, 1994 (L. Franz et al.).
43. S. Patai: *The Chemistry of the Amino Group,* Interscience, London-New York-Sydney 1968.
44. L. K. Jackson, G. N. R. Smart, G. F. Wright, *J. Am. Chem. Soc.* **69** (1947) 1539.
45. S. Patai: *The Chemistry of the Carbon Nitrogen Double Bond,* Interscience, London 1970.
46. E. Schmitz, *Angew. Chem.* **73** (1961) 23; *Houben-Weyl,* **vol. 10/2** (1967) 71.
47. H. Krimm, *Chem. Ber.* **91** (1957) 1057; *Houben-Weyl,* **vol. 10/4** (1968) 449.
48. L. Spialter, J. A. Pappalardo: *The Acyclic Aliphatic Tertiary Amines,* MacMillan Publ. Co., New York 1965, p. 29.
49. V. A. Nekrasova, N. I. Shuikin, *Russ. Chem. Rev. (Engl. Transl.)* **34** (1965) 843.
50. Monsanto, GB 716 649, 1952.
51. *Houben-Weyl,* **vol. 11/1** (1957) 687; ICI, GB 969 542, 1964 (J. D. Seddon); K. Smekal, G. Esser, DD 27 348, 1964. BASF, DE 1 921 467, 1969 (H. Corr, W. Friedrichsen).
52. Mitsui Chem. Ind., JP 3 180, 1968 (T. Sasa, H. Hirai).
53. Monsanto, US 271 869, 1981 (H. L. Merten, G. R. Wilder); US 3 196 179, 1962 (R. M. Robinson); BASF, EP 813 906, 1996 (T. Rühl et al.).
54. BASF, DE 1 543 377, 1966 (K. Adam, E. Haarer); Inst. Cercetari, RO 63 243, 1975 (E. Dutkay).
55. Halcon Internat., NL 6 401 010, 1964.
56. Inst. Français du Pétrole, GB 1 050 589, 1965 (B. Choffe, R. Stern); Quaker Oats, FR 1 446 554, 1965 (A. P. Dunlop, D. G. Manly); Mitsubishi Chem. Ind., JP 4 332, 1968 (R. Matsura, T. Otaki); Toa Gosei Chem. Ind., JP 19 897, 1970 (T. Kawaguchi, T. Matsubara); JP 19 898, 1970 (T. Kawaguchi, T. Matsubara).
57. BASF, DE 3 045 719, 1980 (N. Goetz, P. Jacobs, L. Hupfer, H. Trussaint, W. Reiss).
58. Dearborn Chem., US 2 956 889, 1956 (W. L. Denman); *Chem. Eng. (N.Y.)* **62** (1955) no. 4, 140.
59. Monsanto, US 2 191 657, 1937 (M. W. Harman).
60. US Rubber, GB 898 319, 1962.
61. Monsanto, US 2 731 338, 1951 (E. A. Fike, H. L. Morrill).
62. Esso, GB 793 737, 1958 (J. Scott, R. J. Westland).
63. BASF, DE 1 921 467, 1970 (W. Corr, W. Friedrichsen).
64. W. S. Emerson, *Org. React. (N.Y.)* **4** (1948) 218.

65. Shell Development, US 2 580 284, 1949 (T. J. Deahl, F. H. Stross).
66. Goodyear, US 3 086 018, 1957 (A. F. Hardman); Monsanto, US 2 930 777, 1955 (H. M. Leeper).
67. DuPont, US 3 376 341, 1964 (C. R. Bauer). BASF, DE 851 189, 1944 (O. Stichnoth).
68. W. S. Emerson, *Org. React. (N.Y.)* **4** (1948) 174. BASF, DE 1 543 354, 1966 (H. Corr, E. Haarer).
69. Bayer, DE 826 641, 1949 (E. Windemuth).
70. Bayer, DE 962 113, 1953 (F. Brochhagen, A. Höchtlen).
71. Bayer, DE 946 173, 1953 (W. Bunge, K. H. Mielke).
72. J. F. Norris, A. E. Bearse, *J. Am. Chem. Soc.* **62** (1940) 953.
73. BASF, DE 805 518, 1949 (O. Stichnoth). Ugine Kuhlmann, FR 1 530 477, 1967.
74. Witco Chem., US 3 551 486, 1968 (J. M. Solomon).
75. Abbott Lab., FR 1 333 693, 1963 (R. M. Robinson).
76. Dow, US 2 571 016, 1949 (J. Dankert). Abbott Lab., FR 1 34 3 391, 1963 (R. M. Robinson); US 3 351 661, 1967 (F. H. Van Munster).
77. I. Garz, K. Schwabe, *Werkst. Korros.* **14** (1965) 842. Nagynyomasu Kiserleti Intezet, HU 16 386, 1976 (J. Gemes).
78. R. Johannsen, *Farbe Lack* **70** (1964) 189.
79. S. Hünig, M. Kiessel, *J. Prakt. Chem.* **277** (1958) 224.
80. DuPont, DE 1 468 779, 1964 (L. D. Brake).
81. BASF, DE 1 272 919, 1966 (P. Raff, H. G. Peine).
82. BASF, DE 2 036 502, 1970; DE 2 038 023, 1970 (K. Adam, H. Hoffmann).
83. R. Jacquier, H. Christol, *Bull. Soc. Chim. Fr.* 1954, 556.
84. L. Ruzicka, M. W. Goldberg, M. Hürbin, *Helv. Chim. Acta* **16** (1933) 1339.
85. L. Schuster, *Tetrahedron Lett.* **1963**, 2001. BASF, DE 1 178 846, 1964 (L. Schuster).
86. BASF, FR 1 468 354, 1966 (K. Adam, E. Haarer).
87. V. Prelog, M. Fausy El-Neweihy, O. Häfliger, *Helv. Chim. Acta* **33** (1950) 365.
88. H. Meister, *Justus Liebigs Ann. Chem.* **679** (1964) 83.
89. L. Schuster, *Tetrahedron Lett.* **1963,** 2001. BASF, NL 6 408 389, 1965.
90. BASF, FR 1 528 400, 1967.
91. J. M. Patterson, P. Drenchko, *J. Org. Chem.* **24** (1959) 878.
92. I. G. Farbenind., DE 701 825, 1938 (W. Reppe).
93. DuPont, US 2 525 584, 1948 (C. A. Bordner).
94. J. Timmermans, *J. Chim. Phys. Phys. Chim. Biol.* **34** (1937) 693.
95. S. A. G. Osborn, D. R. Douslin, *J. Chem. Eng. Data* **13** (1968) 543.
96. A. G. Cook: *Enamines,* Marcel Dekker, New York 1969.
97. I. Szmuszkovicz, *Adv. Org. Chem.* **4** (1963) 1.
98. K. Blàha, O. Cesvinka, *Adv. Heterocycl. Chem.* **6** (1966) 147.
99. G. Laban, R. Bayer, *Z. Chem.* **8** (1968) 165.
100. K. Smeykal, K. K. Moll, *Chem. Tech. (Leipzig)* **19** (1967) 92.
101. A. A. Ponomarev, A. S. Ĉegolja, *Khim. Geterotsikl. Soedin.* 1966, 239.
102. Celanese, DE 1 802 122, 1968 (D. R. Larkin).
103. Quaker Oats Co., US 3 163 652, 1962 (D. G. Manly). Inst. of Petrochem. Synth., SU 2 478 887, 1977 (A. N. Bashkirov).
104. ICI, GB 971 187, 1962 (C. Gardner, G. A. Silverstone).
105. Rhône-Poulenc, FR 1 475 961, 1966 (G. Chichery, P. Perras).
106. ICI, NL 6 605 495, 1966.
107. Ciba, US 2 826 583, 1954 (K. Hoffmann, J. Heer).
108. Phillips Petroleum, US 2 771 737, 1951 (C. R. Scott, A. L. Ayers).
109. G. Jones, *Org. React. (N.Y.)* **15** (1967) 204.
110. A. T. Nielsen, W. J. Honlihan, *Org. React. (N.Y.)* **16** (1968) 1.
111. G. B. Bachmann, M. T. Atwood, *J. Am. Chem. Soc.* **78** (1956) 484.
112. E. N. Zilbermann, E. D. Skorikowa, *Zh. Obshch. Khim.* **23** (1953) 1659.
113. Bayer, DE 738 448, 1941 (H. Raab).
114. BASF, DE 1 138 780, 1960 (A. Palm, H. Stanger).
115. Artemev, GB 1 570 647, 1980 (D. M. Popov, E. V. Genkina). BASF, DE 2 837 290, 1978 (W. Schröder, W. Franzischka). Mitsubishi, JP 82 183 771, 1982.
116. Monsanto, US 3 635 952, 1969 (D. A. Tyssee).
117. Soc. Prod. Chim. Marles Kuhlmann, FR 1 583 461, 1969.
118. *Houben-Weyl,* **vol. 11/1** (1957) 45, 129.
119. Wyandotte, DE 1 049 864, 1957 (W. K. Langdon, E. Jaul).
120. Goodyear, DE 3 002 342, 1979 (K. J. Frech). BASF, DE 3 125 662, 1981 (W. Schröder, W. Lengsfeld, G. Heilen). Jefferson, US 3 155 657, 1961 (W. C. Bedoit).

121. M. F. Obrecht, *Effluent Water Treat. J.* **4** (1964) 279.
122. Chem. Werke Hüls, BE 642 387, 1964.
123. H. Koppers GmbH, US 3 555 782, 1968 (H. Deringer).
124. E. Cinelli, P. L. Girotti, R. Tesei, *Hydrocarbon Process. Pet. Refiner* **42** (1963) no. 8, 141.
125. Wyandotte, US 3 167 551, 1961 (E. A. Weipert). Agency Ind. Sci. Techn., JP 80 43 008, 1980 (Y. Sugi).
126. Jefferson, US 3 210 349, 1961 (N. B. Godfrey).
127. BASF, DE 1 793 380, 1968 (E. Fürst, H. Toussaint).
128. Jefferson, US 3 087 928, 1961 (N. B. Godfrey).
129. Pennsalt Chem., US 3 155 656, 1960 (R. H. Goshorn, R. Ferren).
130. Jefferson, US 3 151 113, 1962 (P. S. Advani, G. P. Speranza). ICI, GB 1 106 084, 1964 (L. L. Dufty).
131. BASF, FR 1 575 059, 1968. Union Carbide, US 3 112 318, 1957 (R. C. Lemmon, R. C. Myerly).
132. DuPont, US 2 130 948, 1937 (W. H. Carothers).
133. M. M. Epstein, C. W. Hamilton, *Mod. Plast.* **37** (1960) 142.
134. Eastman Kodak, US 2 950 197, 1960 (C. F. H. Allen, E. R. Webster). AGFA, DE 1 130 283, 1962 (W. Himmelmann, H. Ulrich).
135. J. R. Lindsay Smith, A. U. Smart, M. V. Twigg, *J. Chem. Soc., Perkin Trans. 2* **1992**, 939.
136. Air Products and Chemicals, US 4 362 886, 1981 (M. E. Ford, T. A. Johnson).
137. Air Products and Chemicals, US 3 166 558, 1961 (R. L. Mascioli).
138. Houdry Process Corp., US 2 937 176, 1956 (E. C. Herrick).
139. Air Products and Chemicals, EP 69 322, 1981 (J. E. Wells). Chem. Werke Hüls, DE 1 445 578, 1963 (W. Thomas, H. Steinbrink, M. zur Hansen). Bayer, DE 2 846 813, 1978 (L. Imre, W. Horstmann, H. G. Leopold).
140. Air Products and Chemicals, US 5 741 906, 1996 (J. N. Armor, J. G. Santiesteban, H. Li). Air Products and Chemicals, EP 842 935, 1996 (J. N. Armor, J. G. Santiesteban, H. Li).
141. Houdry Process Corp., DE 1 120 691, 1958 (M. Orchin).
142. National Starch and Chem. Corp., DE 2 040 607, 1970 (H. H. Stockmann).
143. S. S. Chang, E. F. Westrum, *J. Phys. Chem.* **64** (1960) 1547.
144. A. Chentin, I. P. Mathieu, *J. Chem. Phys.* **53** (1956) 106.
145. G. Klipping, I. N. Stranski, *Z. Anorg. Allg. Chem.* **297** (1958) 23.
146. VEB Gärungschemie Dessau, DE 1 076 104, 1959 (H. Plischke, W. Zettl). H. Plischke, *Z. Chem.* **1** (1961) 217.
147. Stachowski, PL 106 934, 1980. Celanese, US 2 640 826, 1949 (A. F. Maclean, A. L. Stautzenberger).
148. Bayer, DE 977 337, 1953 (R. Ludwig, F. Halle).
149. Montecatini, BE 601 096, 1960. Soc. Belge de l'Azote et des Prod. Chim., GB 962 925, 1962 (R. Englebert, A. Genkenne, A. Lefebre).
150. F. Meissner, E. Schwiedessen, D. F. Othmer, *Ind. Eng. Chem.* **46** (1954) 724.
151. Tenneco, US 3 538 199, 1968 (S. Weiss, D. X. Klein); BE 743 974, 1969 (S. Weiss).
152. Phillips Petroleum, US 3 547 597, 1967 (G. E. Hayes).
153. Degussa, DE 1 104 967, 1961 (H. J. Mann). BASF, US 2 912 435, 1957 (H. Scholz).
154. Degussa, DE 1 085 491, 1959 (A. Lehn).
155. Bayer, FR 895 645, 1962. Food Machinery and Chem. Corp., US 2 833 656, 1953 (A. F. Kalmar, H. F. Fitzpatrick).
156. Dow, GB 960 732, 1964; US 3 268 399, 1964 (R. R. Langner); US 3 524 854, 1968 (S. J. Kuhn).
157. "Armeen Aliphatic Amines", *Armak Product Data Bulletin* 72-8 (1972).
158. *Kirk-Othmer,* 2nd ed., **vol. 2,** pp. 130, 132.
159. Melting points were calculated using references [160][161] , and internal Armak data converted into a polynomial. Vapor pressures were correlated in a similar way using the Antoine equation given in [162]
160. R. R. Dreisbach: "Physical Properties of Chemical Compounds, III", *Adv. Chem.* 1961, no. 29. 1.
161. *Selected Values of Properties of Chemical Compounds,* Thermodynamics Research Center, Department of Chemistry, Texas A & M University, June 30, 1971.
162. J. A. Dean: *Lange's Handbook of Chemistry,* 11th ed., McGraw-Hill, New York 1973, p. 10 – 31.
163. G. Tauber, A. May, *Tenside Deterg.* **19** (1982) no. 3, 151 – 156.
164. A. Weiss, *Tenside Deterg.* **19** (1982) no 3, 157.
165. C. W. Hoerr, M. R. McCorkle, A. W. Ralston, *J. Am. Chem. Soc.* **65** (1943) 328, 329.
166. Kao Soap Co., JP Kokai 7 616 603, 1976 (U. Nishimoto, K. Okabe, M. Takaku).
167. Akzona Inc., US 4 237 064, 1980 (R. A. Reck).

168. H. L. Sanders, J. B. Braunwarth, R. B. McConnell, R. A. Swenson, *J. Am. Oil Chem. Soc.* **46** (1969) no. 3, 167 – 170.
169. Henkel DE 2 117 427, 1972 (G. Demmering).
170. Armour, GB 1 223 730, 1967 (R. A. Reck).
171. Akzona Inc., US 4 368 127, 1983 (J. M. Richmond).
172. Union Carbide, US 3 636 114, 1972 (E. Tobler, D. J. Foster).
173. Armour, US 3 222 402, 1965 (M. C. Cooperman).
174. R. S. Egly, Nitroparaffin Symposium, Commercial Solvents Corp., Chicago, Jan. 25, 1956.
175. Akzona Inc., US 4 334 887, 1982 (D. Frank, L. D. Metcalfe).
176. Akzona Inc., US 4 357 181, 1982 (D. Frank, L. D. Metcalfe).
177. S. Billenstein, G. Blaschke, *J. Am. Oil Chem. Soc.* **61** (1984) 353.
178. P. Jowett, GB 1 475 689, 1973.
179. W. R. Grace & Co., GB 1 388 053, 1975 (R. A. Diffenbach).
180. Chemcell, US 3 574 754, 1971 (G. Specken).
181. Hoechst, DE 1 941 290, 1971 (H. Mueller, W. Kuehn, G. Mueller-Schiedmayer, J. Strauss).
182. Hoechst, DE 1 280 243, 1968 (H. Oberrauch, W. Froehlich, H. Lewinsky); *Chem. Abstr.* **70** (1969) 11 108.
183. Archer-Daniels-Midland Co., US 3 264 254, 1966 (E. J. Sawyer).
184. Armour, GB 860 922, 1961 (S. H. Shapiro, F. Pilch).
185. Lion Fat and Oil, US 4 248 801, 1981 (S. Tomidokoro, M. Sato, D. Saika).
186. Hoechst, US 3 444 205, 1969 (W. Froehlich, H. Oberrauch, K. Fischer).
187. Kao Soap Co., JP Kokai 81 152 441, 1981.
188. BASF, DE 2 639 648, 1978 (H. Hoffmann, H. Mueller, H. Toussaint, A. Wittwer).
189. L. Barrault, Y. Pouilloux, *Catal. Today* **37** (1997) 137.
190. Archer-Daniels-Midland Co., US 3 223 734, 1965 (H. T. Fallstad, A. E. Rheineck).
191. Gulf Research & Development Co., US 2 953 601, 1960 (A. C. Whitaker).
192. BASF, DE 2 645 712, 1978 (H. Hoffmann, H. Mueller, H. Toussaint, A. Wittwer); US 4 207 263, 1980.
193. Japan Catalytic Chem. Ind., JP Kokai 7 424 905, 1974 (H. Koike, N. Kurata, Y. Okuda).
194. Imperial Chem. Ind., GB 2 048 866, 1979 (C. B. Hanson, G. E. B. Smith).
195. Gulf Research and Development Co., US 4 251 465, 1981 (H. E. Swift, R. A. Innes, P. Adams).
196. UOP, DE 3 014 455, 1981.
197. Kao Soap Co., US 4 254 060, 1980 (H. Kimura, K. Matsutani, S. Tsutsumi).
198. Kao Soap Co., US 4 210 605, 1980 (F. Hoshino, H. Kimura, K. Matsutani).
199. Gulf Research and Development Co., US 3 742 060, 1973 (R. W. Lagally, G. D. Johann).
200. UOP, US 4 250 115, 1981 (T. Imai).
201. Shell Oil Co., US 3 234 283, 1966 (H. V. Finch, R. E. Meeker).
202. M. R. McCorkle, P. L. Dubrow, B. E. Marsh, *J. Am. Oil Chem. Soc.* **45** (1968) 10A.
203. Armour, US 3 338 967, 1966; US 3 530 153, 1966 (R. H. Potts, E. J. Miller, A. Mais).
204. Ethyl Corp., GB 1 115 238, 1968.
205. Texaco Development Corp., US 3 739 027, 1971.
206. Continental Oil Co., US 3 371 118, 1968 (A. J. Lundeen, K. L. Motz).
207. Hoechst, US 4 234 509, 1980; DE 2 813 294, 1978; DE 2 737 607, 1979 (S. Billenstein, B. Kukla, H. Stuehler).
208. Henkel, DE 1 288 595, 1969 (H. Rutzen); US 3 579 585, 1971 (H. Rutzen).
209. SNIA Viscosa, US 4 198 348, 1980 (C. A. Pauri).
210. C. Jungermann: *Cationic Surfactants,* Marcel Dekker, New York 1970, Chap. 13.
211. L. D. Metcalfe, *J. Am. Oil Chem. Soc.* **56** (1979) 786A.
212. K. G. van Senden, *Recl. Trav. Chim. Pays Bas* **84** (1965) 1459.
213. DGF-Einheitsmethoden M-V-6 (57).
214. G. Grossi, *J. Gas Chromatogr.* **3** (1965) no. 5, 170.
215. L. D. Metcalfe, A. A. Schmitz, *J. Gas Chromatogr.* **2** (1964) no. 1, 15.
216. L. D. Metcalfe, R. J. Martin, *Anal. Chem.* **44** (1972) 403.
217. R. B. Stage, J. B. Stanley, P. B. Moseley, *J. Am. Oil Chem. Soc.* **49** (1972) no. 2, 87.
218. G. F. Okunev, L. G. Morozova, E. M. Shmeleva, *Zavod. Org. Sint.* **2** (1972) 173; *Chem. Abstr.* **83** (1975) 178 196 v
219. A. E. Merbach, *Mitt. Geb. Lebensmittelunters. Hyg.* **66** (1975) 176.
220. F. Mozayeni, *Appl. Spectrosc.* **33** (1979) 520.
221. DuPont, US 2 088 325, 1937 (J. E. Kirby).
222. DuPont, US 2 132 902, 1938 (S. Lenhar).
223. L. L. Johnsen, *Weeds* **13** (1965) 123 – 130.
224. Armour, US 3 246 015, 1966 (H. L. Lindaberry, W. W. Abramitis).

225. Akzona, US 3 506 433, 1970 (W. W. Abramitis, R. A. Reck).
226. Sherwin-Williams, US 3 166 469, 1965 (H. A. Pass, M. S. H. Nurse, B. J. Watt).
227. Armour, US 3 135 656, 1964 (W. W. Abramitis, R. A. Reck).
228. Armour, US 3 223 517, 1965 (W. W. Abramitis, R. A. Reck).
229. *Kirk-Othmer,* 4th ed., **2,** 405.
230. R. W. Gallant, *Hydrocarbon Process.* **48** (1969) 143.
231. *Kirk-Othmer*, 4th ed., **8**, p. 74.
232. I. Flament, M. Stoll, *Helv. Chim. Acta* **50** (1967) 1754.
233. N. Puschin, R. Mitič, *Justus Liebigs Ann. Chem.* **532** (1937) 300.
234. Rohm & Haas, US 2 326 643, 1941 (W. F. Hester).
235. H. Bruson, *Org. React. (N.Y.)* **5** (1949) 79.
236. IG Farbenind., *BIOS Final Rep.* no. 1154, no. 5; Toyo Soda, JP 79 26288, 1979 (S. Otsuki).
237. Simon Carves, GB 1 147 984, 1965 (J. G. Blears, P. Simpson). Nippon Soda, JP 6325, 1965.
238. Dow, US 2 861 995, 1956 (G. F. Mac Kenzie). BASF, DE 1 172 268, 1962 (S. Winderl, E. Haarer). Union Carbide, US 3 112 318, 1957 (R. C. Lemon, R. C. Myerly).
239. Union Carbide, US 5 750 790, 1995 (S. W. King).
240. Union Carbide, EP 737 514, 1995 (S. W. King). Union Carbide, JP 08 2997 98, 1995 (S. W. King).
241. Dow, WO 96/38 226, 1995 (D. Chang, F. A. Sherrod).
242. *Eur. Chem. News* **20** (1971) no. 495, 16.
243. Rohm & Haas, JP 1909, 1964.
244. Prod. Chim. Pechiney-St. Gobain, US 3 281 471, 1963 (P. Chassaing, J. Coillard).
245. DuPont, US 2 429 876, 1944 (W. F. Gresham).
246. Monsanto, FR 1 426 289, 1964 (J. R. Le Blanc).
247. Bayer, FR 1 411 269, 1964 (H. J. Hennig, G. Braun).
248. Phrix, DE 1 185 331, 1959 (F. Kaiser, P. Weber).
249. Thiocol Chem., GB 1 038 355, 1964.
250. Polythane, BE 633 253, 1963 (A. D. Schneider, T. V. Peters).
251. Texas Corp., US 2 696 473, 1951 (S. J. Sokol).
252. Shell Dev., US 2 285 878, 1942 (E. R. White).
253. *Chemical Market Reporter*, Feb. 16, 1998.
254. *Eur. Chem. News*, Feb. 9 – 15, 1998.
255. BASF, FR 1 347 648, 1963 (S. Winderl, E. Haarer).
256. BASF, DE 1 034 184, 1957 (H. Scholz, H. Günthert).
257. *Houben-Weyl,* **vol. 11/1** (1957) 272, 565.
258. *Houben-Weyl,* **vol. 11/1** (1957) 606.
259. *Houben-Weyl,* **vol. 11/1** (1957) 92, 95, 558. Nitto Chemical, JP 79 40524, 1979 (S. Tomita).
260. *Houben-Weyl,* **vol. 11/1** (1957) 558, 567.
261. A. Wingler, *Angew. Chem.* **61** (1949) 52.
262. *Houben-Weyl,* **vol. 11/1** (1957) 61, 72, 131, 498. E. Buse, *Naturw. Rdsch.* **50** (1997) 253.
263. Degussa, DE 4 343 890, 1993 (T. Haas, D. Arntz, D. Most); Degussa, DE 4 343 891, 1993 (T. Haas, D. Arntz, D. Most); Degussa, DE 195 40 191, 1995 (T. Haas, R. Burmeister, D. Arntz, K.-L. Weber, M. Berweiler); BASF, EP 449 089, 1990 (F. Merger, C.-U. Priester, T. Witzel, G. Koppenhoefer, W. Harder); BASF, DE 19 507 398, 1995 (G. Voit, T. Witzel, B. Breitscheidel, W. Harder, H. Luyken, A. Paul, K.-H. Roß, P. Wahl); BASF, EP 757 980, 1995 (T. Witzel, G. Voit).
264. Cyanamid, US 2 371 104, 1943 (R. H. Kienle).
265. Cyanamid, US 2 554 475, 1947 (T. J. Suen).
266. Talbott Dev. Assoc., US 2 694 633, 1950 (D. K. Pattiloch).
267. Cyanamid, US 2 469 683, 1945 (J. R. Dudley).
268. Cyanamid, US 2 414 552, 1942 (H. F. Pfann).
269. US 5 410 086, 1991 (L. M. Burgess).
270. X. Verdaguer, U. E. W. Lange, S. L. Buchwald, *Angew. Chem.* **110** (1998) 1174; *Angew. Chem. Int. Ed.* **37** (1998) 1103; A. Tungler, K. Foder, *Catal. Today* **37** (1998) 191.
271. *Chem. Market Reporter*, Dec. 8, 1997, p. 14.
272. Bayer, DE 19 621 686, 1996 (U. Stelzer).
273. Ajinomoto, EP 845 455, 1994 (T. Hijiya, T. Yonekawa, T. Takemoto).
274. Abermarle Corp., WO 97 47 572, 1996 (R. E. Young, H. V. Phan, T. Manimarant, R. C. Zumstein).
275. R. R. Beard, J. T. Noe: "Aliphatic and Alicyclic Amines," in C. D. Clayton, F. Clayton (eds.): *Patty's Industrial Hygiene & Toxicology,* vol. II B, Wiley Interscience, New York 1982.
276. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentrationen (MAK),* Wiley-VCHs, Weinheim 1998.
277. Am. Conf. of Governmental Industrial Hygienists (ed.): *Threshold Limit Values (TLV),* Cincinnati, Ohio, 1997.
278. H. F. Smyth et al., *Arch. Ind. Hyg.* **10** (1954) 61 – 68.
279. G. Dura et al., *Arch. Toxicol., Suppl.* **8** (1985) 481 – 487.

280. D. Fasset in C. D. Clayton, F. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* **vol. II B,** Wiley Interscience, New York 1981, p. 3148.
281. BASF Aktiengesellschaft, unpublished results.
282. W. L. Sutton in F. A. Patty (ed.): *Industrial Hygiene and Toxicology,* **vol. II,** Interscience Publ., New York 1962, pp. 2037 – 2044.
283. H. F. Smyth et al., *Arch. Ind. Hyg.* **4** (1951) 119 – 122.
284. BIBRA (The British Industrial Biological Assoc.), Toxicity Profile "Diethylamine and its Hydrochloride", Carshalton/Surrey 1989.
285. H. Brieger, W. A. Hodes, *Arch. Ind. Hyg.* **3** (1951) 287.
286. D. W. Lynch et al., *Fund. Appl. Toxic.* **6** (1986) 559 – 565.
287. E. Anderson, B. Irvholm, *Arbete och Haelsa* **23** (1994) 1 – 8.
288. G. Z. Kagan, *Gig. i Sanit.* **30** (1965) 351.
289. P. Schmidt et al., *Z. Ges. Hyg.* **20** (1974) 393.
290. Union Carbide Corp., TSCATS, OTS 0515610 (July 07, 1979)
291. C. H. Hine et al., *Arch. Environ. Health* **1** (1960) 65 – 74.
292. "Triethylamin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1983.
293. "Triethylamine", *Dang. Prop. Mat. Rep.* **14** (1994) 2 –27.
294. H. F. Smyth et al., *Am. Ind. Hyg. Assoc. J.* **23** (1962) 95 – 107.
295. "Dipropylamine", *Dang. Prop. Mat. Rep.* **7** (1987) 54 – 58.
296. Monsanto Agric. Comp., TSCATS, OTS 0542011 (July 17, 1985)
297. I.N.R.S., *Cahiers de Notes Documentaires* **88** (1977) 379 – 382.
298. H. Smyth, C. Carpenter, C. Weil, U. Pozzani, J. Striegel, J. Nycum, *Am. Ind. Hyg. Assoc. J.* **30** (1969) 470.
299. BIBRA (The British Industrial Biological Assoc.), Toxicity Profile "Cyclohexylamine and its Common Salts" , Carshalton/Surrey 1991.
300. DuPont, TSCATS, OTS 0215026 (1982), OTS 0571265 (1992), and OTS 0556352 (1992)
301. "Dicykan", in BG Chemie (ed.): *Toxikologische Bewertung* **15,** Heidelberg 1986.
302. DuPont, TSCATS, OTS 0571204, and OTS 0571506 (1992)
303. C. F. Reinhardt, M. S. Britelly in G. D. Clayton, F. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* **vol. II A,** Wiley-Interscience, New York 1981, p. 2671 – 2822.
304. H. J. Trochimowicz et al. in: G. D. Clayton, F. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* **Vol. II E,** Wiley-Interscience, New York, 1994, p. 3285.
305. "Pyrrolidin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), Verlag Chemie, Weinheim 1977.
306. G. N. Zaeva, N. N. Ordynskaya, L. J. Duninina, N. J. Osipenko, V. N. Ivanov, *Gig. Tr. Prof. Zabol.* **2** (1974) 29 – 32.
307. GDCh-Beratergremium (BUA): "N,N-Dimethylpropandiamin", *BUA-Report 197,* S. Hirzel, Stuttgart 1996.
308. F. S. Malette, E. V. Haan, *Arch. Ind. Hyg. Occup. Med.* **5** (1952) 311.
309. BG Chemie (ed.): "Diethylentriamin", in: *Toxikologische Bewertung* **182,** Heidelberg 1992.
310. BG Chemie (ed.): "Triethylentetramin", in: *Toxikologische Bewertung* **181,** Heidelberg 1991.
311. "Hexamethylentetramin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1991.
312. Degussa AG, Report No. 84-0046-DKT and 84-0047-DKT, 1984 (unpubl.).
313. Final report on the Safety Assessment of Morpholine, *J. Am. Coll. Toxicol.* **8** (1989) 707 – 748.
314. BUA (GDCh Advisory Committee on Existing Chemicals of Environmental Relevance), Morpholin, BUA-Report **56** (Dec. 1990), VCH Verlagsgesellschaft, Weinheim.
315. UNEP, Environmental Health Criteria for Morpholine **179,** WHO, Genf (1995).
316. L. B. Bourne, F. J. M. Miller, L. B. Alberman, *Br. J. Ind. Med.* **16** (1959) 81.
317. C. P. Carpenter, H. F. Smyth, jr., *Am. J. Ophthalmol.* **29** (1946) 1363 – 1372.
318. R. E. Eckardt, *EHP Environ. Health Perspect.* **17** (1976) 103 – 106.
319. T. Sollman: *A Manual of Pharmacology,* 8th ed., Saunders, Philadelphia 1957.

320. D. R. Morris, L. J. Marton: *Polyamines in Biology and Medicine,* Marcel Dekker, New York-Basel 1981.
321. I. G. Sipes, A. J. Gandolfi in M. O. Amdur et al. (eds.): *Casarett and Doull's Toxicology* 4th ed., Pergamon Press, New York 1991, pp. 88 – 126.
322. H. Blaschko, *Pharmacol. Rev.* **4** (1952) 415.
323. M. L. Simenhoff: "Metabolism of Aliphatic Amines," *Kindney Int. Suppl.* **3** (1975) 314 – 317.
324. "Ethylamin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikolog.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1984.
325. A. J. Amor, *Manufact. Chem. Manufact. Perfumer* **10** (1949) 540.
326. P. C. Tkachev, *Gig. i Sanit.* **36** (1971) 334 – 339.
327. T. M. Hellmann, F. H. Small, *J. Air Pollut. Contr. Assoc.* **24** (1974) 979.
328. "Diethylamin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1984.
329. N. R. Kosiborod, *Gig. i Sanit.* **30** (1968) 31.
330. T. M. Hellmann, F. H. Small, *Chem. Engin. Progr.* **69** (1973) 75.
331. "Diethylamin", in H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), Suppl. VCH Verlagsgesellschaft, Weinheim 1996.
332. H. Druckrey et al., *Arzneimittel-Forsch.* **13** (1963) 320.
333. B. Akessen, S. Skerfving, *Br. J. Ind. Med.* **45** (1988) 262 – 268.
334. B. Akessen et al., *Toxicol. Appl. Pharmacol.* **100** (1989) 529 – 538.
335. "Triethylamin", in H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1996.
336. B. R. Dudek et al., *Toxicologist* **11** (1991) 88.
337. Monsanto Comp., TSCATS, OTS 0522377 (Feb. 08, 1990)
338. "Cyclohexylamin", in D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), Verlag Chemie, Weinheim 1979.
339. D. Schmähl, *Arzneim. Forsch.* **23** (1973) 1466 – 1470.
340. J. F. Gaunt, J. Hardy, P. Grasso, S. D. Gangolli, K. R. Butterworth, *Fed. Cosmet. Toxicol.* **14** (1976) 255 – 267.
341. D. L. Oser, S. Carson, G. E. Cox, E. E. Vogin, S. S. Sternberg, *Toxicology* **6** (1976) 47 – 65.
342. J. Hardy, J. F. Gaunt, J. Hooson, R. J. Hendy, K. R. Butterworth, *Fed. Cosmet. Toxicol.* **14** (1976) 269 – 276.
343. R. Kroes et al., *Toxicology* **8** (1977) 285.
344. J. F. Gaunt, M. Sharrat, P. Grasso, A. B. Landsdown, S. D. Gangolli, *Food Cosmet. Toxicol.* **12** (1974) 609 – 624.
345. M. S. Legator, K. A. Palmer, S. Green, K. W. Peterson, *Science* **165** (1969) 1139 – 1140.
346. J. H. Turner, D. H. H. Hutchinson, *Mutat. Res.* **26** (1974) 407.
347. P. Cooper, *Food Cosmet. Toxicol.* **15** (1977) 69 – 70.
348. G. L. Kennedy, *J. Appl. Toxicol.* **11** (1991) 367 – 371.
349. E. Zeiger, *Environ. Mutagen.* **9** (1987) 1 – 110.
350. R. A. Lovell, *Vet. Clin. North Amer. Small Animal Practice* **20** (1990) 453 – 468.
351. J. R. Berger, M. Globus, E. Melamed, *Arch. Neurol.* **36** (1979) 180 – 181.
352. S. Fregert, *Manual of Contact Dermatitis,* Scandinavian University Books, Munksgaard 1974.
353. J. Pepys, C. A. C. Pickering, H. W. G. London, *Clinical Allergy* **2** (1972) 189 – 196.
354. L. Hagmar, T. Bellander, B. Bergbö, B. G. Simonsson, *JOM J. Occup. Med.* **24** (1982) 193 – 197.
355. Deutsche Forschungsgemeinschaft (ed.): *MAK- und BAT-Werte-Liste,* Wiley-VCH, Weinheim 1998.
356. "Piperazin", in H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim Suppl. 1996.
357. "Piperazine" in Scientific Basic for Swedish Occupational Standards VI, P. Lundberg (ed.): *Arbete och Haelsa* **32** (1985) 22 – 41.
358. Dutch Expert Committee for Occupational Standards: Health-based recommended exposure limit for piperazine, Min. of Social Affairs and Employment, RA 7/91 (Jan. 1992).

359. A. Thorgeirson, *Acta Derm. Venereol.* **58** (1978) 332 – 336.
360. G. N. Zaeva et al., *Gig. Tr. Prof. Zabol.* **2** (1974) 29 – 32.
361. H. Garcia, W. Lijinski, *Z. Krebsforsch.* **79** (1973) 141 – 144.
362. A. M. Pearson et al., *Meat Sci.* **7** (1982) 259 – 268.
363. C. U. Dernehl, *Ind. Med. Surg.* **20** (1951) 541.
364. S. Lam, M. Chang-Yeung, *Am. Rev. Resp. Dis.* **121** (1980) 151 – 155.
365. R. S. H. Yang et al., *Fundam. Appl. Toxicol.* **4** (1984) 539.
366. L. R. De Pass et al., *Fundam. Appl. Toxicol.* **9** (1987) 687.
367. B. D. Hardin et al., *Teratog. Carcinog. Mutag.* **7** (1987) 29.
368. "1,2-Diaminoethan", in H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikol.-arbeitsmed. Begründung v. MAK-Werten,* Deutsche Forschungsgemeinschaft (DFG), VCH Verlagsgesellschaft, Weinheim 1990, Suppl. Pregnancy.
369. R. S. Slesinski et al., *Mut. Res.* **124** (1983) 299 – 314.
370. S. Zimmering et al., *Environ. Mutagen.* **7** (1985) 87 – 100.
371. U.S. EPA: Health and Environmental Effects Document for Ethylene Diamine, Washington, EPA/600/8-89/004.
372. R. E. Brubaker, H. J. Muranko, D. B. Smith, G. J. Beck, G. Scovel, *J. Occup. Med.* **21** (1979) 688 – 690.
373. Union Carbide Corp.: unpublished reports, 1988.
374. W. Dobryszycka, J. Kulpa, A. Woyton, J. Woyton, J. Szacki, A. Dziboa, *Arch. Immunol. Ther. Exp.* **33** (1975) 867 – 870.
375. H. Szacki, J. Woyton, A. Dzioba, J. Rabczynski, A. Woyton, *Arch. Immunol. Ther. Exp.* **22** (1974) 123 – 128.
376. Union Carbide Corp.: unpublished reports, 1981 and 1987.
377. Degussa AG, Report No. 85-0057-DKT, 1985 (unpubl.).
378. H. J. Cronin, *JAMA* **83** (1924) 250 – 251.
379. L. L. Schwartz et al. in: *Occupational Diseases of the Skin,* 3rd ed., Lea & Febinger, Philadelphia (1957) 546 – 573.
380. H. H. Gelfand, *J. Allergy* **34** (1963) 374 – 381.
381. M. Kiec-Swierczynska, *Medycyna Pracy* **45** (1994) 393 – 398.
382. I. A. Rapoport, *C. R. Acad. Sci. USSR* **54** (1946) 65 – 67.
383. R. Gleckmann et al., *Am. J. Hosp. Pharm.* **36** (1979) 1509 – 1512.
384. W. Lijinsky, H. W. Taylor, *Food Cosmet. Toxicol.* **15** (1977) 269 – 274.
385. G. Della Porta et al., *Food Cosmet. Toxicol.* **6** (1968) 707 – 715.
386. H. Netvig, J. Anderson, W. Wulff-Rasmussen, *Food Cosmet. Toxicol.* **9** (1971) 491 – 500.
387. G. Röhrborn, F. Vogel, *Dtsch. Med. Wochenschr.* **92** (1967) 2315 – 2321.
388. K. H. Baldermann, T. M. Roehrborn, *Humangenetik* **4** (1967) 112 – 126.
389. M. Vujosevic et al., *Acta. Vet. (Beograd)* **36** (1986) 91 – 94.
390. H. Hurni, H. Ohder, *Food Cosmet. Toxicol.* **11** (1973) 459 – 462.
391. X. Wang, R. R. Suskind, *Contact Derm.* **19** (1988) 11 – 15.
392. C. C. Conaway et al., *Fundam. Appl. Toxicol.* **4** (1984) 465 – 472.
393. R. D. Harbison et al., *Fundam. Appl. Toxicol.* **12** (1989) 491 – 507.
394. N. V. Migukima, *Toksikol. Nov. Prom. Khim. Veshchestv* **13** (1973) 92 – 100.
395. P. N. Magee, J. M. Barnes, *Adv. Canc. Res.* **10** (1967) 163 – 246.
396. P. M. Newberne, R. C. Shank, *Fd. Comet. Toxicol.* **11** (1973) 819 – 825.
397. M.-A. Shibata et al., *Fd. Chem. Toxicol.* **25** (1987) 569 – 574.
398. R. D. Harbison et al., *Fundam. Appl. Toxicol.* **12** (1989) 491 – 507.
399. K. Brunnemann, S. Hecht, D. Hoffmann, *J. Toxicol. Clin. Toxicol.* **19** (1982/83) 661 – 688.
400. H. F. Smyth, C. P. Carpenter, C. S. Weil, *J. Ind. Hyg. Toxicol.* **31** (1949) 60 – 63.
401. H. F. Smyth, C. P. Carpenter, C. S. Weil, *Toxicol. Appl. Pharmacol.* **28** (1974) 313 – 319.
402. *Armak Product Data Bulletin* 81-6 (1981), 80-8 (1980), and 73-6 (1973), Armak Co., Chicago.
403. W. B. Deichmann, J. L. Radomski, W. E. MacDonald, R. L. Kascht, R. L. Erdmann, *AMA Arch. Ind. Health* **18** (1958) 483 – 487.
404. M.-A. Shibata et al., *J. Toxicol. Environ. Health* **22** (1987) 187 – 194.

Amines, Aromatic

Amines, Aliphatic; Aminophenols; Aniline; Benzidine and Benzidine Derivatives; Naphthalene Derivatives; Phenylenediamines and Toluenediamines; Toluidines and Xylidines are separate keywords

PETER F. VOGT, Mobay Chemical Corp., Pittsburgh, Pennsylvania 15205, United States(Chaps 2 – 4)

JOHN J. GERULIS, Mobay Chemical Corp., Pittsburgh, Pennsylvania 15205, United States(Chap. 5)

1. Introduction 1
2. Physical and Chemical Properties 3
2.1. Physical Properties 3
2.2. Chemical Properties 3
2.2.1. Acetylation and Similar Reactions . 3
2.2.2. *N*-Alkylation and Condensations . . 4
2.2.3. *C*-Alkylation 4
2.2.4. Cyclization 4
2.2.5. Diazotization 5
2.2.6. Halogenation 6
2.2.7. Nitration 6
2.2.8. Oxidation 6
2.2.9. Sulfonation 6
2.2.10. Reduction 7
2.2.11. Other Reactions 7
3. Production 7
3.1. Reduction 7
3.1.1. Chemical Reduction 7
3.1.2. Catalytic Hydrogenation 9
3.2. Nucleophilic Substitution 12
3.2.1. Exchange of Halide 13
3.2.1.1. Activated Halides 13
3.2.1.2. Unactivated Halides 13
3.2.2. Exchange of Hydroxyl and Ether Groups 14
3.2.3. Exchange of Sulfo Groups 15
3.3. Other Processes 15
4. Economic Aspects 15
5. Toxicology, Occupational Health, and Environmental Protection . . 15
5.1. Toxicity 15
5.2. Exposure Limits 16
5.3. Protective Measures 17
5.4. Environmental Monitoring 18
6. References 18

1. Introduction

Definitions. *Aromatic amines* are organic nitrogen compounds that may be considered derivatives of ammonia (NH_3) with at least one of the hydrogen atoms replaced by an aryl group. The nitrogen must be bound directly to the aromatic ring and so be able to interact with the aromatic π-electron system. The amine can be primary, secondary, or tertiary depending on whether one, two, or three of the protons are replaced by alkyl or aryl groups. The simplest aromatic amine is derived from benzene and is called aniline or benzenamine ($C_6H_5NH_2$). The amines from toluene are referred to as toluidines and from naphthalene as naphthylamines. Pyridine, pyrrole, and others in which nitrogen forms part of the ring are treated not as aromatic amines but as heterocyclic compounds.

Nomenclature. Aromatic amines most often are named by adding the suffix "amine" to the name of the radical (or radicals) replacing one or more of the hydrogen atoms in ammonia, except when some trivial name exists. According to IUPAC nomenclature, the amino (NH_2) or modified amino group (NHR, NRR′) is considered a substituting group of the aromatic hydrocarbon. Therefore, these compounds are named as derivatives of benzene, toluene, naphthalene, and others, e.g., *N,N*-dimethylaminobenzene, $C_6H_5N(CH_3)_2$.

Aniline is generally used as the parent structure for its derivatives unless a carbon atom is attached to the benzene ring or unless a function of higher seniority than amino is present in the molecule. Important derivatives of aniline are named as such, e.g., *N*-methylaniline, *p*-nitroaniline, chloroaniline, whereas the sulfonic acids are named as derivatives

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_037

of benzene (aminobenzenesulfonic acid) or by their trivial names (metanilic acid for *m*-aminobenzenesulfonic acid). If two of the hydrogen atoms of ammonia are replaced by aryl groups, the compounds are generally described as diarylamines, with diphenylamine (DPA) as the parent compound. Aromatic amines with two amino groups on the same benzene ring are named phenylenedi-amines. *Chemical Abstracts* called these compounds benzenediamines in 1972.

Historical Development. In 1826 O. UNVERDORBEN first obtained aniline by the dry distillation of indigo and called it "krystallin". In 1834 F. RUNGE found it in coal tar; in 1841 C. F. FRITZSCHE prepared the same oily liquid by heating indigo with potash and gave it the name "aniline," from "anil" (the Portuguese word for indigo), which was identical to the "benzidam" of N. N. ZININ (1841). The structure was proved by A. W. VON HOFMANN in 1843, who showed that it was obtained from the reduction of nitrobenzene. This is still the most common method of production. Diphenylamine, the simplest diarylamine, was first prepared by VON HOFMANN in 1863 by destructive distillation of triphenylmethane dyes.

Major Uses and Importance. Analgesics and Antipyretics; Antioxidants; Azo Dyes; Chemotherapeutics; Fungicides, Agricultural; Insect Control; Optical Brighteners; Ozone; Photography; Pigments, Organic; Polyurethanes; Rubber 4., Chemicals and Additives; Weed Control are separate keywords.

Aromatic amines are widely used as dye intermediates, especially for azo dyes, pigments, and optical brighteners; as intermediates for photographic chemicals, pharmaceuticals, and agricultural chemicals; in polymers via isocyanates for polyurethanes; and as antioxidants. They have been listed as corrosion inhibitors of mild steel in the pickling process [1].

The oldest use of aromatic amines is as intermediates for *dyes*. Substituted anilines and especially naphthylamines have been precursors of azo dyes since the middle of the 1800s. Even with the replacement of the once predominant anthraquinone dyes with more and more azo dyes derived from aromatic amines, the previous strong demand of aniline in the dye industry has decreased markedly in the United States to less than 4 % of its former magnitude. Once over 700 dyes listed in the Colour Index (C. I.) [2] were prepared from aniline and its derivatives; now only a few are produced in commercial quantities.

The production and sales volume for the aniline derivatives as a group is generally growing only at the average rate of expansion of the chemical market as a whole. United States production of many of them has ceased, and the supply is shifting to Japan, Korea, China, and other East Asian countries along with textile and dye production.

p-Nitroaniline [*100-01-6*], with a United States production of 6200 t in 1981 [3] and an import volume of 630 t in 1983 [4], is probably the largest in this group. *p*-Nitroaniline is used mainly in the production of dyes, antioxidants, and pharmaceuticals.

Anisidines, especially *o*-anisidine [*90-04-0*], with a limited United States production and imports of 570 t in 1983 [4], is an important intermediate for pigments and azo dyes. Japan reported a production of only 320 t in 1981 [5], down from a high of 997 t in 1974. *Chloroanilines* are used mostly for agricultural products. *4,4′-Diamino-2,2′-stilbenedisulfonic acid* [*81-11-8*] has maintained its strength as the largest volume raw material for optical brighteners (United States use 5000 t in 1984).

By far the largest share of the *aniline* consumption (67 % of total 1984 consumption) is used for the manufacture of isocyanates, primarily for 4,4′-methylenebis(phenylisocyanate), (MDI) to make polyurethanes. About 20 % of the aniline consumption in the United States and in Western Europe versus 27 % in Japan is used in the rubber industry for the manufacture of antioxidants; vulcanization accelerators, such as 2-mercaptobenzothiazoles (MBT) [*149-30-4*]; diphenylguanidines; for condensation products of aniline with aldehydes and ketones, e.g., 2,2,4-trimethyl-1,2-dihydroquinoline (TMDHQ) [*147-47-7*] and its polymer [*26780-96-1*]. Many herbicides, fungicides, insecticides, and animal repellents are made from aniline or its derivatives. Between 1960 and 1980, United States aniline consumption grew at an average rate of 8.8 % annually and then slowed down to an average rate of 3.1 % in the period 1980 – 1984.From 1980 to 1993

aniline consumption grew from 0.29 to 0.49 Mio t in th United States, in Western Europe from 0.30 to 0.59 Mio t, and in Japan from 0.07 to 0.16 Mio t [6].

Analgesics and sulfonamides are examples of important pharmaceuticals derived from aromatic amines. Many sulfa drugs start from acetanilide (antifebrin) [*103-84-4*], which itself has been used as an antipyretic and analgesic agent.

Diarylamines readily donate their amine hydrogen atom to terminate various free-radical reactions. For this reason they are widely used as antioxidants for various polymers and elastomers, and as stabilizers for nitrocellulose [7]. Nearly all commercial synthetic rubbers are protected with alkylated diphenylamines. Because they darken on oxidation, diphenylamines cannot be used in light colored rubber and polymer compositions. In this case *N,N′*-di-β-naphthyl-*p*-phenylenediamines are used instead. Some diarylamines retard the effect of intense gamma or neutron radiation [8]. Diarylamines have been used to stabilize sulfur trioxide, are used as corrosion inhibitors in roofing asphalt, and are essential in promoting cross-linking in poly(vinyl chloride) [9] and in polyethylene [10]. Diarylamines containing other functional groups, such as amino or hydroxy, are used as dye intermediates [11], antiozonants, in color photography, and for hair dyeing and have been patented for use in the image formation process in photographic films [12], [13].

2. Physical and Chemical Properties

2.1. Physical Properties

All aromatic amines are either solids or liquids at room temperature. The ones without other functional groups are colorless, high-boiling materials having a characteristic odor. Most oxidize on exposure to air and darken. They are only slightly soluble in water.

In contrast to the aliphatic amines, which are stronger bases than ammonia, the aromatic amines are only weakly basic with basicity constants $K_B \leqq 10^{-10}$ in water. Their onium salts, therefore, are ionized in water and act as Brønsted acids.

The effect of the phenyl groups becomes even more dominant in diphenylamine ($K_B = 10^{-14}$), which can form salts only with strong acids that are completely ionized in water, whereas triphenylamine rarely forms salts and has no basic properties.

The physical properties of the most important aromatic amines are given under → Aniline.

2.2. Chemical Properties

The chemical reactions of aromatic amines are similar to those of aliphatic amines. These reactions are influenced greatly by the basicity of the amine nitrogen. Therefore, such electron-withdrawing substituents as nitro and cyano reduce the base strength of the amine nitrogen when substituted in the ortho or para positions of the benzene ring(s). Aromatic amines undergo a variety of chemical reactions. The ones of greatest industrial significance are described in this section.

2.2.1. Acetylation and Similar Reactions

Primary and secondary aromatic amines react with carboxylic acids, acid chlorides, anhydrides, and esters to form a wide variety of commercially important amides. For example, *p*-acetanisidine, *N*-(4-methoxyphenyl)acetamide [*51-66-1*], is obtained in nearly theoretical yield, especially if the water formed is removed (e.g., by distillation).

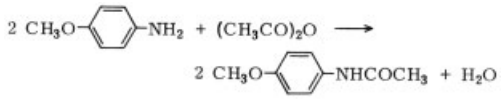

Acid chlorides, in general, give the best yields of pure product. Polymeric aromatic amides have been made from *m*- and *p*-phenylenediamines by reaction with iso- or terephthaloyl chloride. These so-called aramid fibers are marketed by Du Pont (United States) under the trade names Nomex, poly(1,3-phenyleneisophthalamide) [*24938-60-1*], and as Kevlar, poly(1,4-phenyleneterephthalamide) [*24938-64-5*] [14], [15] (→ Fibers, 4. Synthetic Organic).

The reaction with phosgene gives isocyanates or ureas depending on the reaction conditions.

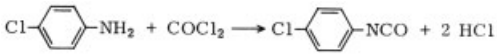

Commercially important isocyanates are obtained from toluenediamine, methylenedianiline, and other aromatic primary diamines. The monoisocyanates are used in such agricultural products as herbicides, insecticides, bacteriostats, fungicides, and pesticides. The major use of the diisocyanates is as intermediates for polyurethanes.

Reaction with dimethylformamide in the presence of sodium methoxide yields *N*-arylformamides in 45 – 88 % yield [16], [17]. With aromatic aldehydes, Schiff bases (azomethines, Ar - CH = N - Ar′) are formed, and with chloral and hydroxylamine, oximinoacetanilides result in 80 – 91 % yield [18]. Dithiocarbamic acid ammonium salts are formed with CS_2 and NH_3 in toluene. These salts give aryl isothiocyanates (70 – 90 % yield) when treated with phosgene [19].

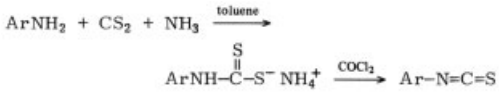

Instead of toluene, water also can be used as the solvent, followed by treatment with heavy metal salts, e.g., $Pb(NO_3)_2$, $CuSO_4$, $FeCl_3$ [20]. Using CS_2 alone in ethanol or acetone, aromatic amines react to yield *N,N′*-diarylthioureas.

With cyanogen chloride, *N,N*-diarylguanidines are formed in 90 % yield.

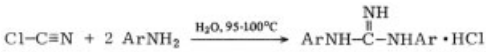

2.2.2. *N*-Alkylation and Condensations

Alkylation of aromatic amines to the corresponding *N*-alkyl- and *N,N*-dialkylamines is generally of importance only for aniline (→ Aniline), [21–24].

Substituted mono- or diethanolamines can be produced in high yields from ethylene or propylene oxide and an aromatic amine.

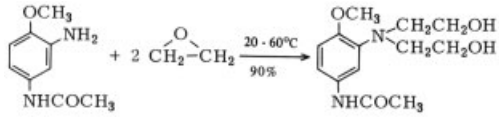

The resulting products, as such, or after further acetylation of the hydroxy groups, are important coupling compounds for azo dyes.

Aldehydes, ketones, acetals, and orthoformates give a variety of condensation products depending on the reaction conditions. For example the reaction of aromatic amines with ketones gives Schiff bases that can be hydrogenated to the alkyl-substituted amines. In the case of phenylenediamine derivatives, important antioxidants and antiozonants for rubber are produced, such as *N*-(1,3-dimethylbutyl)-*N′*-*p*-phenylenediamine [*793-24-8*], which is sold under various *trade names:* Vulkanox 4020 (Mobay, Bayer), Flexone 7L (Uniroyal), Santoflex 13 (Monsanto), UOP 588 (UOP), Wingstay 300 (Goodyear).

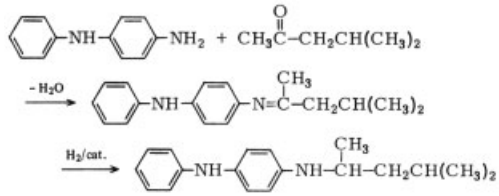

2.2.3. *C*-Alkylation

Under Friedel – Crafts conditions, the aromatic ring can be alkylated with alkenes to form the corresponding alkylanilines at temperatures of 200 – 280 °C and 20 MPa (200 bar). The alkylation takes place in the positions ortho to the amino group and if one is occupied only monoalkylation results [25–29].

These compounds are used as chain extenders for polyurethanes (e.g., 3,5-diethyl-2,4-diaminotoluene), and for herbicides (e.g., 2,6-diethylaniline and 2-methyl-6-ethylaniline).

2.2.4. Cyclization

Aromatic amines undergo numerous cyclization reactions to give a variety of *N*-heterocyclic

compounds. Quinoline derivatives are obtained with aldehydes and ketones. Aniline condenses with acetone in the presence of hydrochloric acid, *p*-toluenesulfonic acid, acidic clays, or acidic ion-exchange resins to give 2,2,4-trimethyl-1,2-dihydroquinoline (TMDHQ) [*147-47-7*], which on oligomerization is used as an antioxidant in the rubber industry.

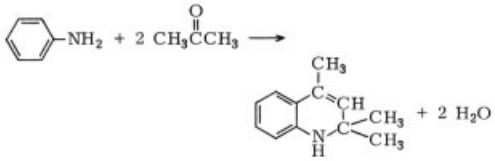

With aldehydes and styrenes, 1,2,3,4-tetrahydroquinolines result [30], and with glycerol or acrolein, quinoline [*91-22-5*] is formed in the Skraup synthesis [31]. Many of the substituted quinolines are intermediates for antimalarial agents. In the Combes quinoline synthesis [32], aniline and 1,3-diketones are condensed to make 2,4-disubstituted quinolines. With hydroxyketones or with haloketones, aromatic amines yield substituted indoles [33]. Aromatic amines and 2-aminophenols condense with CS_2 to form 2-mercapto-1,3-benzothiazole. With sulfur monochloride (S_2Cl_2) 1,2,3-benzodithiazole is formed and can be hydrolyzed to the corresponding 2-aminothiophenol in yields of up to 93 % [34].

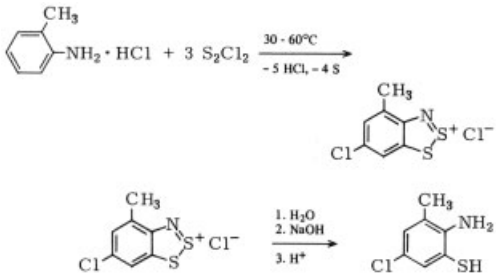

Benzofuroxane [*480-96-6*] results from the treatment of *o*-nitroaniline with sodium hypochlorite [35].

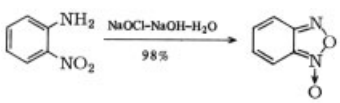

Diphenylamines form carbazoles when heated with iodine as the catalyst, whereas heating diphenylamine with sulfur in the presence of various catalysts results in phenothiazine [*92-84-2*].

2.2.5. Diazotization

Primary aromatic amines react with nitrous acid to form diazonium salts.

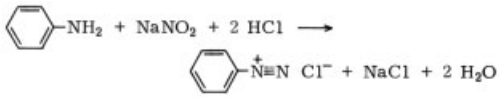

This reaction has considerable general utility and forms the basis for azo dye production by coupling of the diazonium salt with aromatic amines and phenols (→ Azo Dyes). In the Sandmeyer reaction the diazo group is replaced by chlorine from a CuCl solution. The reaction also is used for the introduction of Br, CN, I, and sometimes also aryl and alkyl groups. Secondary aromatic amines react with nitrous acid to form *N*-nitroso compounds in the same way as do aliphatic amines.

Tertiary aromatic amines do not give *N*-nitroso compounds but undergo *C*-nitrozation.

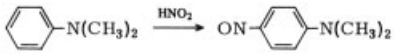

Diaryldiazenes are obtained by the oxidation of aromatic amines with $KMnO_4$ in aqueous KOH [36] or by the reaction of aromatic amines with aromatic nitrosamines [37].

$$Ar-NO + H_2N-Ar' \xrightarrow[70-95\%]{CH_3COOH} Ar-N=N-Ar' + H_2O$$

o-Phenylenediamines react with nitrite at neutral to slightly acidic pH to undergo ring closure to the corresponding benzotriazoles [38], [39]; these are important corrosion inhibitors for Cu-containing alloys.

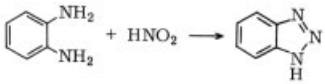

2.2.6. Halogenation

The chlorination of aniline salts in aqueous solutions is so fast that only the trichloroaniline can be isolated [40]. Chlorine in the presence of hydrogen chloride in an anhydrous solvent yields 2,4,6-trichloroaniline [*634-93-5*] [41]. In aqueous sulfuric acid at 110 °C aniline gives chloranil, 2,3,5,6-tetrachloro-2,5-cyclohexadiene-1,4-dione [*118-75-2*], in 91 % yield [42].

2.2.7. Nitration

Generally, nitration of aromatic amines with nitric acid results in concomitant oxidation of the amino function. If it is protected, predominantly the para- and ortho-nitro isomers are formed. Nitration of *p*-acetanisidine [*51-66-1*] in sulfuric acid yields 2-nitro-4-acetaminoanisole [*50651-39-3*] as the major product. This compound can be hydrolyzed subsequently to the free amine.

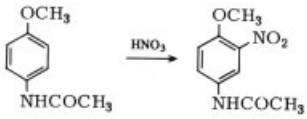

The nitration of *p*-acetaminotoluene [*103-89-9*], however, leads predominantly to substitution ortho to the acetamino group.

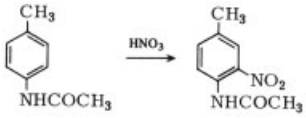

2.2.8. Oxidation

Aromatic amines are very susceptible to oxidation and most turn brown or black if exposed to air for extended periods. The course of oxidation depends on the nature of the oxidizing agent. With hydrogen peroxide and percarboxylic acids (e.g., peracetic acid), oxidation to the corresponding *N*-phenylhydroxylamine or to nitroso or nitrobenzene may occur, depending on temperature and the amount of oxidizing agent present [43]. 2,6-Dichloroaniline [*608-31-1*] is oxidized to 2,6-dichloronitrobenzene by trifluoroperacetic acid [44].

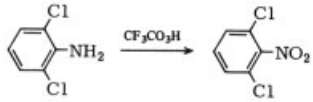

Oxidation with MnO_2 or $Na_2Cr_2O_7$ in sulfuric acid leads to quinones; e.g., 1-naphthylamine is oxidized to 1,4-naphthoquinone. Oxidation of aniline with dichromate or chlorate in the presence of copper or vanadium yields aniline black.

2.2.9. Sulfonation

Reaction of aromatic amines with sulfuric acid, oleum, or sulfur trioxide leads to the introduction of the sulfonic acid group mainly in the positions ortho or para to the amino group. 2-Methoxy-5-methylaniline [*120-71-8*] is sulfonated with fuming sulfuric acid to 4-amino-5-methoxy-2-methylbenzenesulfonic acid [*6471-78-9*] (the intermediate to make the dye FD & C Red 40 [*25956-17-6*] [2]).

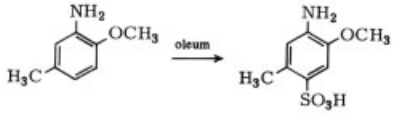

1-Naphthylamine [*134-32-7*] is sulfonated with oleum at room temperature to give 1-amino-4- [*84-86-6*] and 5-naphthalenesulfonic acid [*84-89-9*].

Many aromatic amines also sulfonate via the *baking process* in which the amine sulfate is heated to 220 °C under vacuum to yield either the *para*- or *ortho*-sulfonic acid. The process is very selective and gives high yields of the *p*-sulfonic acid, e.g., aniline leads to sulfanilic acid (4-aminobenzenesulfonic acid [*121-57-3*]). If the para position is occupied, sulfonation takes place ortho to the amino group; e.g., *p*-toluidine [*106-49-0*] gives 2-amino-5-methylbenzenesulfonic acid [*88-44-8*] [45]. 1-Naphthylamine yields only 1-amino-4-naphthalenesulfonic acid [*84-86-6*], also known as naphthionic acid [11]. This compound is used as a dye intermediate for Congo Red, Fast Red A, Azo Rubin, and others.

2.2.10. Reduction

Hydrogenation of aromatic amines results in the corresponding cycloaliphatic amines. This process is of greatest importance with amines that do not contain other easily reducible functional groups. Such catalysts as Raney nickel, cobalt – alumina, and more exotic ones are used [46]. Hydrogenation of aniline using an alumina–silicate catalyst [47] leads to 80 % cyclohexyl-amine [*108-91-8*] and 15 % dicylohexylamine [*101-83-7*]; bis(4-amino-3-methylphenyl)methane is reduced in 89 % yield on an alumina catalyst [48]. Hydrogenation of aniline over a Raney nickel catalyst gives 95 % dicyclohexylamine [49]. Good results have been obtained with rhodium catalysts. 4,4′-Methylenedianiline (MDA) [*101-77-9*] has been hydrogenated on a 5 % rhodium – aluminum catalyst to give bis(4-aminocyclohexyl)methane in good yield [48].

2.2.11. Other Reactions

In a reversal of the Bucherer reaction, naphthylamines can be hydrolyzed to the corresponding naphthol in yields approaching 100 %. Sulfonic acid groups on the ring accelerate the reaction [50], [51].

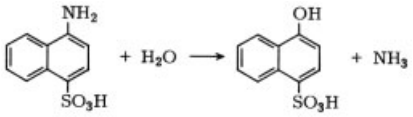

3. Production

Generally, three types of reactions are used for the production of aromatic amines:

1) *Reductions:* using such metals as Fe, Zn, Sn, and Al or their salts; sulfur compounds; electrochemical methods; and catalytic hydrogenation
2) *Nucleophilic substitutions:* replacement of such substituents as halogen, hydroxyl, alkoxy, and sulfo groups
3) *Rearrangements and degradations:* including the benzidine and Beckmann rearrangements, and the Schmidt and Hofmann degradations.

Only the first two reaction types are of general importance. Chemical rearrangements and degradations seldom lead to clean reaction products in high yield.

3.1. Reduction

All aromatic carbon – nitrogen compounds with nitrogen oxidation numbers $\geqq -2$ can be reduced to aromatic amines (for a summary see [52]). Only the reduction of nitro compounds has gained widespread industrial acceptance, be-cause the starting materials can be prepared with a high degree of specificity for a wide variety of compounds. Other aromatic nitrogen compounds cannot be produced economically enough to be used as starting materials for aromatic amines.

3.1.1. Chemical Reduction

In the chemical reduction of aromatic nitro compounds to aromatic amines, the hydrogen that becomes attached to the nitrogen normally comes from the solvent, in many cases water, or from the added acid. The most important reductants are metals, such as iron, tin, and zinc; phosphorus, sulfides, sulfites, and sulfur dioxide are used also. Because oxidation of the reducing agent may result in a useless waste product, which has to be removed in an environmentally safe way, chemical reduction has lost its importance in favor of catalytic reduction for the large-scale production of products.

Reduction with Iron. The reduction with iron in dilute acid was described by Béchamp in 1854 and the process still carries his name. Iron is still the most important metal for the reduction of nitro compounds to aromatic amines, and generally, all nitro compounds can be reduced to the corresponding aromatic amines with iron in water and acid. Side products are expected only if the molecule contains other substituents that can be reduced easily (e.g., nitroso, azo, hydrazine, sulfoxide, and other nitro groups) or can be saponified [52]. Carbon – carbon multiple bonds are not attacked. Bayer (Germany) and Mobay (United States) both use this iron

reduction for the production of iron oxides with aniline as the resulting side product. For a further discussion, → Aniline, Section 2.3, and [52–55].

Another example of industrial importance is the reduction of 4,4′-dinitro-2,2′-stilbenedisulfonic acid to 4,4′-diamino-2,2′-stilbenedisulfonic acid [*81-11-8*], which is still done exclusively using iron in neutral medium. Also, many small-volume dye intermediates are reduced with iron.

Reduction with Other Metals. Aromatic nitro groups also can be reduced with zinc, tin, and aluminum, or their salts in acidic, neutral, or alkaline medium [52].

These reducing agents are generally very mild and do not interfere with, e.g.,-OH, -OR, -COOH, -CO-Ar, halogen, or -CN groups. However, except for laboratory applications they are not of any industrial significance. Instead, catalytic hydrogenation is used because of waste water regulations.

Reductions with Sulfide, Hydrogensulfite, and Sulfur Dioxide. Compounds with several nitro groups can be reduced to nitro amines if a stoichiometric amount of *sulfide* is used [56], [57]. In this case, sodium or ammonium sulfide is the reagent of choice. *m*-Nitroaniline [*99-09-2*] can be obtained in 87 % yield from 1,3-dinitrobenzene [56]:

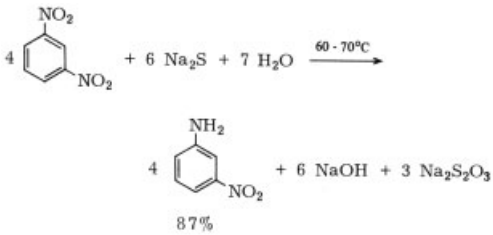

The addition of sulfur reduces the required amount of sodium sulfide. 4-Nitrotoluene is reduced nearly quantitatively to 4-aminobenzaldehyde [*556-18-3*] with sodium sulfide and sulfur [58], [59].

Azo groups stay unchanged if the reduction with sodium sulfide is carried out at 40 – 70 °C, if the reaction times are kept short, and if sulfide is not in excess. However, halogen substituents are replaced easily by -SH [56], [57]. When excess reducing agent is not a problem, ammonium hydrogensulfide in the presence of ammonia can be used [57], [60].

The reduction of aromatic nitro, nitroso, or azo compounds with *sulfite* or *hydrogensulfite* to the corresponding aromatic amines is known as the Piria reaction (Piria, 1851). In aqueous or alcoholic medium, mainly *N*- and *C*-sulfonic acids result [61]. On treatment of these with mineral acid, a mixture of aromatic amines and aminobenzenesulfonic acids is formed.

The quinone oxime intermediate facilitates substitution of hydrogensulfite onto the aromatic ring in the reduction of *ortho*- and *para*-nitrosophenols or naphthols:

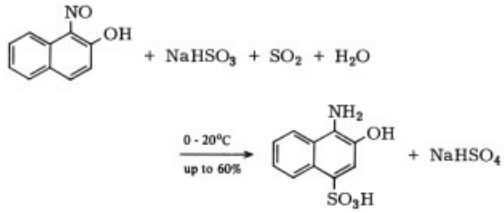

If the hydroxyl group is in the para position, the sulfo group will enter in the 2 – position. Substituents in the 3 – position (e.g., Cl or COOH) are replaced by H [62].

Only after methods were found to reduce the amount of *C*-sulfonation did reduction with SO_2 gain industrial importance. This reduction is best carried out in a closed system in strongly acidic media (e.g., 15 – 40 % aqueous sulfuric acid) at temperatures of 80 – 180 °C with iodine, hydrogen iodide, or iodine salts as catalysts. Glass-lined reactors are the best [63].

Sodium dithionite reduces the aromatic nitro and nitroso group nearly quantitatively [64]:

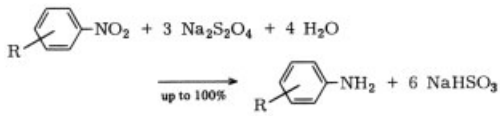

When possible, reductions with sulfide, sulfite, and other sulfur compounds have been replaced by catalytic hydrogenations, resulting in cleaner reaction mixtures, less corrosion (resulting when SO_2 is used), and fewer environmental problems.

Electrochemical Reduction of Nitro Compounds. Electrochemical reduction is generally

a special case of the chemical reduction. An inorganic compound is used as the reducing agent; the oxidized form is then reduced at the cathode and can react again.

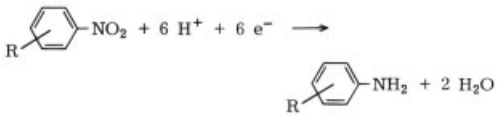

The cathode is made of Pb, Sn, Ni, or Cu, and 15 – 20 % hydrochloric acid is used on the cathode side of a semi-permeable membrane. On the anode side 30 % sulfuric acid is employed [65].

Although electrochemical reductions have been known for more than 100 years, they have remained mostly curiosities and have not been used commercially to any great extent in the past. However, interest is growing as better cells and membranes become available. Some products have been produced electrochemically, especially in India; e.g., *p*-aminobenzoic acid [*150-13-0*], *p*-aminophenol [*123-30-8*], and many others have reached pilot-plant scale or beyond [66], [67].

3.1.2. Catalytic Hydrogenation

Reduction with Hydrazine. The reduction of aromatic nitro compounds with hydrazine represents, in most cases, a special variation of catalytic hydrogenation, where hydrazine is the source of the hydrogen. The mechanism of decomposition of hydrazine on precious-metal catalysts differs depending on the pH value, with the amount of hydrogen generated per mole of hydrazine increasing with higher pH [68]. At weakly alkaline or neutral conditions, 1 mol H_2 is generated:

$$3\,N_2H_4 \xrightarrow{\text{Pt, Pd, or Ni}} 2\,NH_3 + 2\,N_2 + 3\,H_2$$

If barium hydroxide or calcium carbonate is added, 2 mol of H_2 are formed:

$$N_2H_4 \xrightarrow{\text{Pd or Pt}} N_2 + 2\,H_2$$

The reduction with hydrazine sometimes occurs without a catalyst but in most cases is done in the presence of a hydrogenation catalyst, such as Raney nickel, Pd on carbon, or Pd on $CaCO_3$ [69], [70]. Ideally 1 mol of hydrazine gives four reduction equivalents depending on the pH and catalyst, but in practice fewer are formed, and an excess of hydrazine must be used. Carbon – carbon double bonds and carbonyl groups are not attacked [71]. Depending on the reaction conditions, the intermediates of the reduction can be isolated.

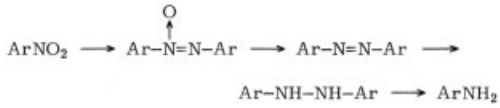

In the reduction using hydrazine without a catalyst, carbonyl-containing compounds are reduced to the corresponding hydroxy compounds via the hydrazone intermediate [72]. The reduction of aromatic nitro compounds with hydrazine has no advantage over catalytic reduction with much cheaper hydrogen, except for laboratory work and very small product volumes, where it allows use of normal pressure. Examples of catalytic reduction with hydrazine hydrate on Pd / C and on Raney nickel can be found in [70], [73–75].

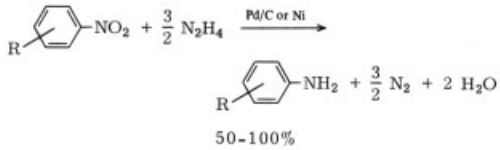

Cyclohexene also has been used on a laboratory scale as a hydrogen source for the catalytic hydrogenation of polynitro compounds [76].

Reduction with Hydrogen. Primary aromatic mono- or polyamines generally are produced by the catalytic hydrogenation of the corresponding nitro compound, either in the vapor or in the liquid phase, with or without a solvent [46], [77–79]. All large-scale products, such as aniline (→ Aniline), or *o*- and *p*-toluidine (→ Toluidines), 2,4- and 2,6-diaminotoluene (→ Phenylenediamines and Toluenediamines), and 1-naphthylamine (→ Naphthalene Derivatives), are produced in this way. For the reaction meachanism see [80], [81].

Catalytic reduction of nitro compounds is very exothermic. In the hydrogenation of nitrobenzene in the liquid phase, 553.5 kJ/mol are released, whereas hydrogenation in the vapor

phase at 200 °C releases 493.2 kJ/mol. Unless this heat is dissipated properly, explosions can result, especially if thermal decomposition of the nitro compound occurs or condensation reactions are initiated as with chloro – nitro compounds [52]. To reduce these hazards, the concentration of nitro compound, the amount and partial pressure of hydrogen, the temperature, and the activity of the catalyst are controlled.

Vapor-Phase Hydrogenation. Industrial use of vapor-phase hydrogenation is limited by the boiling point and the thermal stability of the nitro compounds. Most aniline produced in the United States is made by vapor-phase hydrogenation of nitrobenzene over copper – silica catalyst in $\geqq$ 99 % yield, whereas Pd – alumina is used by Bayer in Europe and Brazil.

Liquid-Phase Hydrogenation. It is preferable to hydrogenate most aromatic nitro compounds in the liquid phase. In this case the pressure and temperature can be changed independently. The temperature is limited by the hydrogenation reaction of the aromatic ring which occurs above 170 – 200 °C. Normally the reduction is carried out at 100 – 170 °C. Some compounds (e.g., 1-bromo-4-nitrobenzene) must be hydrogenated at lower temperatures (20 – 70 °C) to avoid cleavage of sensitive groups. Pressures of 1 – 15 MPa (10 – 150 bar) are used industrially. For the hydrogenation of sensitive nitro aromatic compounds lower pressures are advisable (0.1 – 5 MPa, 1 – 50 bar) [82–84]. Generally the pressure only influences the speed of the reaction by increasing the phase transfer and saturation of the catalyst with hydrogen. The reaction time depends on many parameters, such as hydrogen pressure, concentration, temperature, activity and concentration of the catalyst, and mixing. Often an induction period is observed, which is independent of the activity of the catalyst [85], [86]. Generally the reaction time is from a few minutes to several hours. Considerably longer times are necessary if these parameters are not optimized.

Because of the exothermic nature of the reaction, many safety precautions must be observed, especially for the industrial hydrogenation of aromatic polynitro compounds in the liquid phase without solvents [87–89]. The exothermic reaction is controlled by the continuous addition of small quantities of the nitro compound, thus keeping its concentration below 2 % [90]. Deionized water can be added to remove the heat of reaction by continuous evaporation. The water also affects the activity of the catalyst; e.g., the more water present in the batch hydrogenation of dinitrotoluene without solvent, the lower the activity of the Pd on carbon catalyst [91], [92]. If the amine shows good water solubility, water can be used as the solvent. Water also can be used in cases where the nitro compound forms water-soluble salts with alkali, such as with nitrocarbonic or sulfonic acids. Sometimes 30 – 40 % solutions of the amine in water can be produced directly in the hydrogenation reactor without an additional concentration step [52], [93].

Solvents. Methanol and 2-propanol are preferred; also dioxane, tetrahydrofuran, and *N*-methylpyrrolidone have been used; e.g., *o*-nitrophenol is hydrogenated in good yields in methanol with Pt on carbon. The concentration of the nitro compound can vary substantially. According to [94] a 33 wt% solution of dinitrotoluene in methanol is fed into the hydrogenation reactor, whereas Olin recommends a 15 – 35 wt% solution [95]. In the hydrogenation with a water-immiscible solvent, such as toluene, the water must be removed, as in solvent-free hydrogenation, in order to maintain the activity of the catalyst. Very polar nitro compounds, e.g., bis(4-amino-3-nitrophenyl)sulfone, are hydrogenated in liquid ammonia [96].

Hydrogenation Catalysts. For vapor-phase hydrogenation only metals or metal derivatives on supports are used in fixed beds or fluidized beds, whereas metals with large surfaces (Raney nickel) or metals and metal derivatives on supports with large surfaces (carbon, alumina) are used for the liquid-phase hydrogenation. In practice only Raney nickel, Raney nickel – iron, Raney cobalt, and Raney copper are used as pure-metal catalysts because of their relatively low cost. Precious-metal catalysts, such as Pt and Pd, generally are used at concentrations of 0.5 – 5 wt% on support material with large surfaces, such as charcoal, silica, aluminum oxide, or alkaline-earth carbonates. The powdered form of the catalyst is used in slurries and the pellet form in fixed beds [97]. Because the support material may activate or deactivate the catalyst [81], it is important to find the optimum conditions that not only allow the catalyst to

be used repeatedly and to be regenerated but also are economical for large-scale (> 500 t/a) products [88], [98]. Decomposition and oxidation products generally deactivate catalysts by blocking the active surfaces [81]. Even in small quantities, compounds containing sulfur, arsenic, or antimony are catalyst poisons. However, with increased oxidation state (e.g., $As^{3-} \rightarrow As^{5+}$ or $S^{2-} \rightarrow S^{4+}$) these compounds have less effect as catalyst poisons. With noble metal catalysts, such ions as halide, sodium, and magnesium, cobalt metal, and CO_2 reduce the activity. Nitro aromatic compounds containing sulfur are better reduced with molybdenum or tungsten sulfide [99], or with nickel catalysts that contain zinc or calcium carbonate [102]. In the hydrogenation with Pd on carbon, copper salts are catalyst poisons, whereas in the hydrogenation of dinitrotoluene, both nitrophenol and nitrocresol are strong catalyst poisons and decomposition activators [46], [101].

Hydrogenation of Halide-Substituted Nitro Compounds. Whereas most aromatic nitro compounds are hydrogenated at 100 – 170 °C, lower temperatures are used with halogen-substituted nitro compounds to avoid halide – hydrogen exchange. The catalysts have reduced activity, e.g., sulfide-poisoned metal catalysts. 1-Chloro 2 – nitrobenzene can be reduced at 80 °C and 1.0 – 5.0 MPa (10 – 50 bar) with a 5 wt% Pt on carbon catalyst, which is treated at 22 °C with dilute sulfuric acid, then with H_2 followed by H_2S [102]. The catalyst can be recycled up to 30 times without any further purification. The reaction time is 30 min per batch and only 0.1 % chlorine is cleaved. Therefore, a product of high purity and stability can be obtained in nearly quantitative yield. Also 3-chloroaniline [*108-42-9*], 4-chloroaniline [*106-47-8*], 2,5-dichloroaniline [*95-82-9*], 3,4-dichloroaniline [*95-76-1*], 2,6-dichloroaniline [*608-31-1*], 6-chloro-2-aminotoluene [*87-60-5*], and 4-chloro-2-aminotoluene [*95-79-4*] can be prepared in high yield and purity by this method.

A similar sulfide-poisoned catalyst has been described [103], in which 5 % platinum on carbon catalyst was treated with 1 % aqueous sulfuric acid, hydrogen, and sodium sulfide solution. 2,5-Dichloronitrobenzene was hydrogenated with this catalyst at 90 – 100 °C, 1.0 – 3.0 MPa (10 – 30 bar), 40 min per batch. The catalyst was recycled at least 30 times without loss of activity or selectivity while chlorine cleavage was kept $\leqq$ 0.1 %; yield of 2,5-dichloroaniline [*95-82-9*] was 97 %. Also a 1 % Pt on carbon catalyst was treated with dimethylsulfoxide in water before it was used in the hydrogenation of chloronitro aromatic compounds in 98.4 % yield [104]. Additional treatment with hydrazine hydrate increased the activity of the catalyst with yields of 99.6 % for 2-chloroaniline [*95-51-2*]. Modification of Pt on carbon with lead, bismuth, or silver salts for the hydrogenation of chloronitrobenzene at 100 °C gave chloroaniline in 99 % yield, whereas untreated catalyst at 77 °C resulted in 100 % hydrolysis [105]. 3-Bromonitrobenzene has been hydrogenated using Pt on barium or strontium carbonate at 20 °C, 0.1 MPa (1 bar) [106]. The catalytic activity of Raney nickel can be modified also. For the hydrogenation of 3-chloronitrobenzene the addition of 0.1 – 20 % of thiocyanate to the catalyst has proved beneficial [84].

Equipment and Construction Materials. Hydrogenation in the vapor phase is carried out only continuously, but in the liquid phase it can be done in batch or continuous operation de-pending on volume. Continuous liquid-phase hydrogenation is used exclusively for a relatively few large-scale products, such as dinitrotoluene, nitrotoluene, and 1-nitronaphthalene [92]. The vast majority of aromatic amines have small annual volumes (< 500 t) and are produced by batch hydrogenation with catalyst slurries. This allows fast changeover from one product to another, which is not possible when fixed-bed catalysts are used. The intensive mixing of the three phases (hydrogen gas, nitro compound solution, and solid catalyst) is of great importance. The transport of the reactants toward the active centers of the catalyst, and of the products away from them, must be rapid for good reaction.

Although batch catalytic hydrogenations traditionally are done in stirred, steel or stainless steel autoclaves (Fig. 1), improvements in performance are being realized increasingly through loop reactor technology. Production units employing the loop reactor principle (Fig. 2) have been available for the last two decades and have been applied successfully to the catalytic hydrogenation of aromatic nitro compounds by Buss (Basel, Switzerland) [107–110]. The major advantages reported over stirred autoclaves are increased heat and mass trans-

fers and improved reaction selectivity. In continuous hydrogenation, generally shorter batch cycle times and higher product yields result. In addition catalyst usage is often lower.

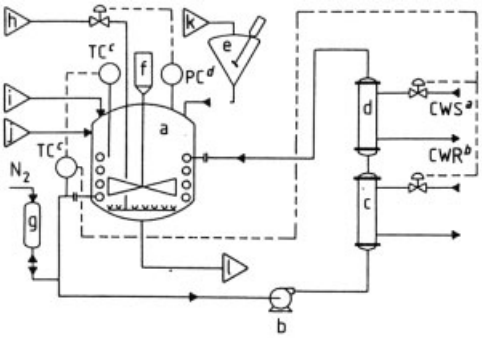

Figure 1. Conventional stirred autoclave system [107–109]
a) Reactor with internal coil; b) Recirculation pump; c) Heater; d) Cooler; e) Catalyst preparation vessel; f) Agitator; g) Expansion vessel; h) Hydrogen; i) Reactant; j) Solvent; k) Catalyst; l) Product; [a] Cold water supply; [b] Cold water return; [c] Temperature control; [d] Pressure control

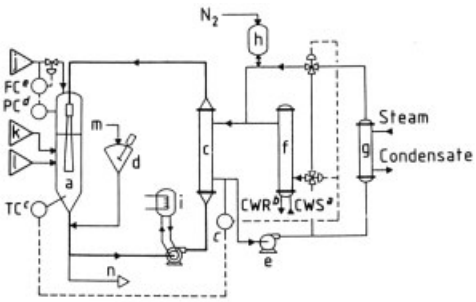

Figure 2. Loop reactor system [107–109]
a) Reactor with mixing nozzle; b) Primary loop recirculation pump; c) Primary loop heat exchanger; d) Catalyst preparation vessel; e) Secondary loop recirculation pump; f) Secondary loop cooler; g) Secondary loop heater; h) Secondary loop expansion vessel; i) Seal fluid cooler; j) Hydrogen; k) Reactant; l) Solvent; m) Catalyst; n) Product; [a] Cold water supply; [b] Cold water return; [c] Temperature control; [d] Pressure control; [e] Flow control

The basic operating principle of a loop reactor is shown in Figure 3 [107–110]. The batch charge and slurry-type catalyst are pumped continuously through an external heat exchanger to a mixing nozzle in the top head of the autoclave. The nozzle acts as an eductor, and the action of the slurry passing through it creates intimate gas – liquid – solid mixing. The chemical reaction primarily takes place in a well-defined zone inside the nozzle prior to gas – liquid disengagement at the nozzle exit. The autoclave acts as a holding vessel. A loop reactor allows for simple and reliable transfer from pilot plant to production, because the scaleup is based only on the nozzle configuration and pump-around rate. In the case of a stirred autoclave, scaleup is more difficult because of the ratio between a small pilot vessel and a large production vessel (e.g., 1 : 1500 to 1 : 12500). Temperature and concentration gradients as well as longer reaction times in the production vessel often result in undesired side reactions and a decreased product yield. Loop reactors minimize these problems by limiting the reaction to a highly agitated, well-defined zone, with a considerably decreased cycle time.

3.2. Nucleophilic Substitution

Practically all non-carbon substituents on the aromatic ring can be exchanged for amino groups. However, only a few of these substitution reactions are of industrial importance, including:

1) Exchange of halide
2) Exchange of hydroxyl and ether groups
3) Exchange of sulfo groups

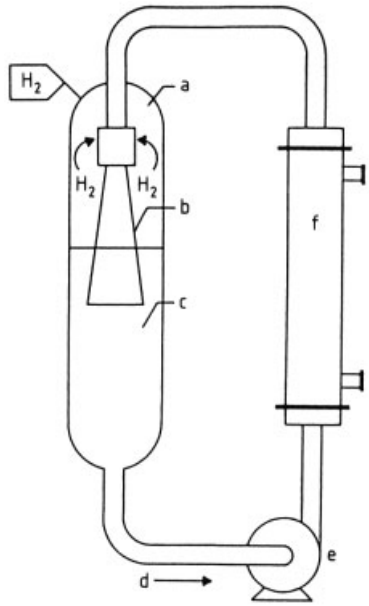

Figure 3. Operating principle of a loop reactor
a) Autoclave; b) Mixing and reaction zone; c) Reactant, solvent, catalyst suspension; d) Flow; e) Primary loop recirculation pump; f) Primary loop heat exchanger

3.2.1. Exchange of Halide

Exchange of halide is universally useful for the production of amines with mono-, poly-, or heterocyclic substituents as well as for diaryl-amines [111–113], and arylpolyamines [114], [115]. Many of the products are used as dye or pigment intermediates. The substitution of a halide on the aromatic ring with ammonia or with amines occurs via two different mechanisms, depending on whether the halide is activated or unactivated.

3.2.1.1. Activated Halides

Aromatic halides containing ortho and/or para electron-withdrawing substituents, e.g., $-NO_2$ or $-CN$, undergo nucleophilic aromatic substitution [116], [117]. Both *o*- and *p*-chloronitrobenzene react with ammonia at 170 °C, whereas 2,4-dinitrochlorobenzene requires only 70 °C and no catalyst [118], [119].

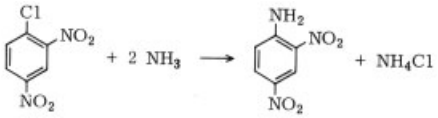

The more basic the amine (unless it is sterically hindered), the faster the reaction with the haloaromatic compound. In the reaction of methylamine and ammonia with a solution or suspension of 2,4-dinitrochlorobenzene at higher temperature, good yields of *N*-methyl-2,4-dinitroaniline are obtained, rather than 2,4-dinitroaniline [120].

Reaction Conditions. *4-Nitroaniline* [*100-01-6*] is produced in an agitated titanium reactor in 99.3 % yield by reacting 4-chloronitrobenzene [*100-00-5*] with a tenfold excess of aqueous ammonia at 175 °C and 4.2 MPa (42 bar) for 10 h [121].

4-(4-Methylanilino)—3-nitrobenzenesulfonamide

is prepared in 90.1 % yield by heating 4-chloro-3-nitrobenzenesulfonamide and *p*-toluidine [*106-49-0*] for 5 h at 130 °C [111]. 4-Nitrodiphenylamine [*836-30-6*] (precursor for antiozonants for rubber) is made by the condensation of 4-nitrochlorobenzene with aniline in chlorobenzene. *N*-Alkylated *p*-phenylenediamine derivatives, however, cannot be produced in aqueous medium.

3.2.1.2. Unactivated Halides

Unactivated halides undergo elimination – addition reaction through a benzyne intermediate. This requires higher temperatures than nucleophilic aromatic substitution and results in isomeric mixtures if other substituents are present. The industrial usefulness of this reaction is therefore limited.

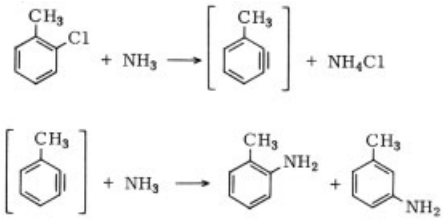

Chlorobenzene and ammonia react at a reasonable rate only above 200 °C [122], generally with the addition of a catalyst, such as copper(I) and copper(II) salts or their amine complexes [115], [123], [124]. Also metal amides (e.g., $NaNH_2$, $LiNEt_2$) are used. The energy needed to exchange the halide is the reverse of the order for activated halides: $F \gg Cl \gg Br > I$

Reaction Conditions. The reaction is generally carried out in aqueous suspension or in solution at elevated temperature and pressure. Because the reaction of arylhalides with amines is exothermic, it is especially important in the batch process to watch the temperature while heating to reaction conditions. This normally poses no problem in the continuous reaction. The hydrogen halide released during the reaction is neutralized either by excess ammonia or by the addition of bases, such as sodium carbonate or calcium oxide [113]. With excess amine the reaction generally proceeds with nearly quantitative yield. With an insufficient excess of strongly basic amines in aqueous solution, side reactions such as hydrolysis or the formation of diarylamines, can occur. Cyano groups are hydrolyzed to amides in aqueous solution, but this reaction can be avoided by using an inert solvent. Carboxylic acids are converted easily to the corresponding amide by reaction in nonaqueous media.

Material of Construction. The presence of halides and amines at elevated temperature and pressure can cause severe corrosion; therefore, the selection of materials and equipment becomes difficult. Autoclaves of normal steel can be used for laboratory work, but considerable reduction in the wall thickness must be expected (several millimeters per year). In addition, stress corrosion can lead to much faster deterioration of the reactor. Low-alloy steels are not useful for production units, especially with continuous working reactors. Stainless steels (chromium- and nickel-containing austenitic steels of the 316 types) were used with success for reactions in aqueous media. Because even these materials become pitted, other more corrosion-resistant alloys, such as Hastelloy and Inconel, have come into use. In addition, zirconium, titanium, and tantaclad (explosion clad tantalum on steel) have been used as reactor materials [121].

3.2.2. Exchange of Hydroxyl and Ether Groups

The displacement of an aromatic hydroxyl group by an amine is not only possible with phenols and cresols but also with quinoline and isoquinoline derivatives. In addition ammonia can replace alkoxy groups that are activated by nitro, halo, or cyano in the ortho or para positions. For example, 2,4-dinitroanisole [*119-27-7*] reacts with ammonia at 50 – 200 °C to form 2,4-dinitroaniline [*97-02-9*] in 50 – 90 % yield [125]. Interest in the reaction of phenols with amines is often a function of the cost of the phenol raw material. The principal exchange processes are:

1) *Reaction in aqueous medium without a catalyst.* 3-Nitrosalicylic acid reacts with aqueous ammonia at increased temperature and pressure to yield 2-amino-3-nitrobenzoic acid [126].
2) *Reaction in the presence of an acid catalyst.* *m*-Aminophenols are produced in the reaction of resorcinol [*108-46-3*] and aqueous ammonium hydroxide or alkylamines with boric acid as the catalyst [127]. Under anhydrous conditions with the same catalyst, 3-(β- and γ-hydroxyalkylamino)phenols are obtained from resorcinol and ethanolamine or 3-amino-1 – propanol [128]. A similar process uses an acid activated alumina – silica catalyst [129].
3) *In a multiphase system*, nitrophenols and ammonia react to yield nitroanilines, preferably using a solvent in which the reaction product has good solubility [130]. Many of the new phase-transfer catalysts are being applied to these reactions.
4) *p-Nitrosoaniline* and *p*-nitroso-*N*-phenylaniline [*156-10-5*] have been produced in liquid ammonia in the presence of either an ammonium salt or a tertiary amine, respectively, and an organic solvent [131].
5) *p-Nitrosophenols* are reacted with alkylamines via the *p*-nitrosophenylethers to give the corresponding *N*-substituted *p*-nitrosoanilines. The water that is eliminated in this process must be removed to drive the reaction to completion [132].
6) *The Halcon process* is the industrially important vapor-phase amination of phenol [133], [134]. Phenols or naphthols react in the gas phase with ammonia at 250 °C on metallic or nonmetallic oxides, e.g., MgO, B_2O_3, Al_2O_3, SiO_2, TiO_2, or mixtures of these. The Halcon process has been used by U.S. Steel Chemicals since 1982 in its plant at Haverhil, Ohio for the production of aniline from phenol (aniline capacity 100000 t/a) [135].
7) *The Bucherer reaction* is the reaction of 1- or 2-naphthols with aqueous amine or ammonium solutions using hydrogen sulfite as the catalyst. It is nearly exclusively used for the naphthalene series [116]. The use of hydrogensulfite reduces the reaction temperature by 50 – 150 °C and results in high yield (85 – 95 %) and purity. The reaction mechanism has been elucidated [136].

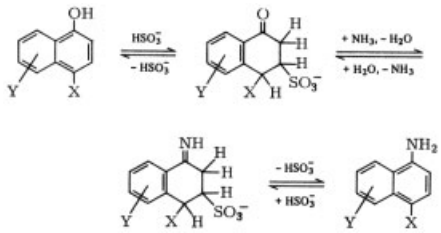

3.2.3. Exchange of Sulfo Groups

Ammonia can replace the sulfo groups in benzene, naphthalene, and anthraquinone derivatives. However, only in the anthraquinone series is this exchange of industrial importance (→ Anthraquinone Dyes and Intermediates). Until recently 1-aminoanthraquinones were made by reaction of 1-sulfoanthraquinones with amines. To destroy the released hydrogensulfite, stoichiometric amounts of an oxidizing agent, e.g., sodium 3-nitrobenzenesulfonate, were added. Increasingly 1-aminoanthraquinone [*82-45-1*] is made by reduction of the corresponding nitroanthraquinone. This method avoids the mercury-catalyzed sulfonation to the 1-sulfoanthraquinone. 1-Cyclohexylaminoanthraquinone-5-sulfonic acid is obtained from anthraquinone-1,5-disulfonic acid and cyclohexyl-amine in aqueous medium under pressure at 100 – 200 °C [138]. The hydrogensulfite is oxidized with sodium perchlorate or 3-nitrobenzenesulfonic acid.

3.3. Other Processes

Only a few of all other possible methods for the production of aromatic amines described in the literature should be mentioned here. 4,4′-Methylenedianiline is produced on a large scale by condensation of aniline with formaldehyde in aqueous or aqueous – methanolic solution. Diphenylamine is made by vapor-phase condensation of aniline over alumina or titanium catalyst at 450 – 500 °C [139] or in the liquid phase at 175 – 450 °C [140]. The rearrangement of arylhydroxylamines in dilute sulfuric acid leads to *o*- and *p*-aminophenols [141]. 3,3′-Dichlorobenzidine [*91-94-1*], an important pigment intermediate, is produced by Bofors (plant in Muskegan, MI) using the benzidine rearrangement starting from *o*-chloronitrobenzene.

4. Economic Aspects

The economically most important aromatic amines are discussed in separate keywords. For the others, data are generally unavailable.

5. Toxicology, Occupational Health, and Environmental Protection

5.1. Toxicity

The toxicity of aromatic amines has long been of concern to the chemical industry because of the serious nature of both acute and chronic effects of exposure to aromatic amines. Despite this, the toxicology of many of the individual compounds has not been studied thoroughly. Therefore, a discussion of aromatic amine toxicity in general is more appropriate than a discussion of individual compounds.

The primary hazards associated with aromatic amine exposure are methemoglobinemia and carcinogenesis. Both of these effects are attributed to activation products of metabolic *N*-oxidation [142]. Acute exposure, from either inhalation or skin contact, results in anoxia because of an impairment of the oxygen transport system of the blood. In aromatic amine exposure the heme iron of hemoglobin is oxidized from Fe(II) to Fe(III). The resulting product, methemoglobin, cannot combine with oxygen or carbon monoxide. The development of *methemoglobinemia* after exposure is often insidious; i.e., the onset of symptoms may be delayed for hours. Usually the first symptom is headache; anoxia, including cyanosis of lips, nose, and earlobes, develops when the methemoglobin blood level reaches 15 %. At levels of 40 %, weakness and dizziness are experienced, and at $>$ 70 %, coma and death may occur.

The extent to which various aromatic amines can cause methemoglobin formation varies widely. However, the arylhydroxylamines are recognized as the strongest methemoglobin formers in this group. The reason for their potency is that they are involved in a cyclic reaction. Phenylhydroxylamine reacts with hemoglobin to form methemoglobin and nitrosobenzene [143]. Nitrosobenzene is then reduced back to phenylhydroxylamine, which then is available to oxidize more hemoglobin. A small amount of nitrosobenzene appears to be reduced all the way to its amine, which eventually stops the reaction. As much as 50 mmol of methemoglobin may be produced by 1 mmol of hydroxylamine [144].

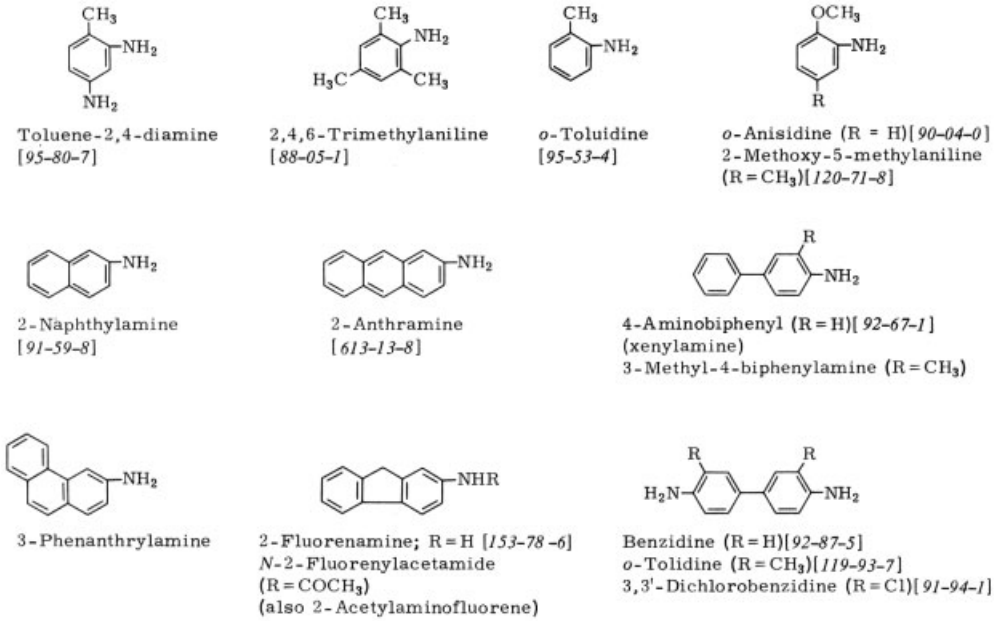

Figure 4. Examples of carcinogenic amines

The *carcinogenic potential* of aromatic amines is of greatest concern. A number of them are known or are suspected of being carcinogens. Indeed it has been reported that the epidemiology of aromatic amine carcinogenesis is essentially the epidemiology of human cancer of industrial origin [145]. The first suspicion of aromatic amine carcinogenesis was raised in 1895 when cases of bladder cancer in Swiss dye workers were reported [146]. Since then a number of epidemiologic investigations and animal studies have identified which aromatic amines are carcinogenic [147–149]. Because most aromatic amines have not been examined thoroughly for carcinogenic potential, researchers are attempting to develop theoretical models for determining which ones are carcinogenic [150–152]. Structural requirements, such as the type of aromatic ring system, the position of the amine group on the ring, and the presence and position of other substituents, are believed to be important to carcinogenic potential. Chemical carcinogens exert their effect by causing DNA damage, which leads to tumor formation. The metabolic reaction products, rather than the aromatic amines themselves, are believed to damage the DNA. These reaction products are electrophilic and are covalently bonded to the nucleic acid bases of DNA. The following reaction sequence has been proposed, where $ArNH_2$ is the aromatic amine [153].

$$ArNH_2 \xrightarrow{oxid.} ArNHOH \xrightarrow[pH<7]{} ArN\overset{+}{O}H_2 \rightarrow Ar\overset{+}{N}H \rightarrow$$

covalent bonding to nucleophilic sites on critical macromolecules

This reaction sequence implies that structural factors in the aromatic amine favoring the cation formation (e.g., electron-donating substituents ortho and para to the amino group) enhance carcinogenicity. Figure 4 is a sampling of some carcinogenic amines [154].

In addition to the effects cited in Table 1 [144], aromatic amines can cause a wide variety of other symptoms, such as eye and skin irritation. By drawing correlations from animal studies, exposure to some of these compounds may lead to liver and kidney damage.

5.2. Exposure Limits

Various countries have established their own exposure limits for workers [156–160]. As an illustration, Table 2 lists those aromatic amines

Table 1. Effects of the chronic administration of primary amines on the blood, liver, and bladder of dogs

	Methemoglobin formation	Toxicity and/or carcinogenesis (liver)	Bladder carcinogenesis
4-Aminobiphenyl [*92-67-1*]	+++	-	+++
2-Naphthylamine [*91-59-8*]	+	-	++
1-Naphthylamine [*134-32-7*]	-	-	-
Benzidine [*92-87-5*]	-	-	+
Aniline [*62-53-3*]	++	-	-
Dichlorobenzidine [*91-94-1*]	-	+	++
Methylenedianiline [*101-77-9*]	-	+++	+- b ∗
4,4′-Methylene-bis(2-chloroaniline) [*101-14-4*]	-	+	++
4,4′-Methylene-bis(2-methylaniline) [*838-88-0*]	-	+++	-
Bis aniline A [*2479-47-2*]	-	+-	+
4-Aminobiphenylamine [*101-54-2*]	-	++	+- b ∗

∗ Indicates only preneoplastic changes, no tumors.

Table 2. ACGIH aromatic amines workplace exposure limit (mg/m^3) [161]

Substance	TWA[a]	STEL[b]	MAK[c]
4-Aminobiphenyl (skin) [*92-67-1*]	A/b [d]	A/b [d]	III A 1 [f]
Aniline & homologues (skin) [*62-53-3*]	10	20	A 2 [e]
Anisidine (*o-, p*-isomers) (skin) [*29191-52-4*]	0.5	-	0.5
ANTU (1-Naphthylthiourea) [*86-88-4*]	0.3	0.9	0.3
Benzidine (skin) [*92-87-5*]	A/b [d]	A/b [d]	III A 1 [f]
3,3′-Dichlorobenzidine (skin) [*91-94-1*]	A 2 [e]	A 2 [e]	III A 2[g]
N,N-Dimethylaniline (skin) [*121-69-7*]	25	50	25
Diphenylamine [*122-39-4*]	10	20	-
N-Isopropylaniline (skin) [*643-28-7*]	10	20	-
N-Methylaniline (skin) [*100-61-8*]	2	5	9
4,4′-Methylene-bis(2-chloroaniline) (skin) [*101-14-4*]	0.22, A 2 [e]	-	III A 2[g]
4,4′-Methylenedianiline (skin) [*101-77-9*]	0.8	4	-
2-Naphthylamine [*91-59-8*]	A/b [d]	A/b [d]	III A 1 [f]
p-Nitroaniline (skin) [*100-01-6*]	3	-	6
N-Phenyl-2-naphthylamine [*135-88-6*]	A 2	-	A 2 [e]
p-Phenylenediamine [*106-50-3*]	0.1 (skin)	-	0.1
o-Tolidine [*119-93-7*]	A 2 [e]	-	A 2 [e]
o-Toluidine (skin) [*95-53-4*]	9, A 2 [e]	-	-
Triphenylamine [*603-34-9*]	5	-	-
m-Xylene-α,α'-diamine [*1477-55-0*]	C [h] 0.1	-	-
Xylidene (all isomers except 2,4-xylidene) [*1300-73-8*]	10 (skin)	-	25

[a] Time-weighted average (8-h workday and a 40-h week);
[b] Short-term exposure limit (15-min TWA exposure);
[c] Maximum concentration values in the workplace;
[d] A recognized carcinogen. No assigned TLV. No exposure permitted by any route;
[e] A suspect carcinogen for man. Exposures should be controlled carefully;
[f] Compounds capable of inducing malignant tumors as shown by experience with humans;
[g] Unmistakably carcinogenic in animal experimentation only;
[h] Ceiling limit; TLV should not be exceeded even instantaneously, avoid skin contact, exposures are likely by skin absorption.

with exposure limits established in the United States (ACGIH) [161] and the Federal Republic of Germany (MAK) [162]. These recommendations are in addition to the U.S. Occupational Safety and Health Administration's Subpart Z exposure limits [163].

5.3. Protective Measures

Exposure to aromatic amines occurs primarily through two routes: skin contact and inhalation. Because of the potential health hazards, strict control measures are necessary to prevent exposure. From an industrial hygiene standpoint,

exposure is best prevented through the use of engineering controls. Whenever possible, processes containing aromatic amines should be enclosed. In a closed process exposure can occur only when equipment fails or when process streams are sampled. If a process is not enclosed, ventilation (both general diluted and local exhaust) should be used to prevent or reduce exposure. Ventilation specifications for a number of industrial processes are available [164]. When ventilation controls are not sufficient to lower exposure, respirators should be used until effective controls are installed. When respirators are used, a complete respiratory protection program should be instituted that includes regular training, maintenance, inspection, cleaning, and evaluation. Where skin contact may occur, workers should wear impervious clothing, including gloves and boots. Contaminated clothing should be stored in closed containers until discarded or cleaned. Safety showers should be available in the immediate area where exposure can occur so that the body can be washed quickly.

5.4. Environmental Monitoring

Routine monitoring of the airborne aromatic amine concentration should be performed to evaluate worker exposure. Measurements should be taken on a single 8-h sample or two 4-h samples to determine the 8-h TWA exposure. In addition, short-term samples should be taken to determine peak exposure when workers are doing specific tasks with a potential for exposure. Air samples should be taken in the worker's breathing zone. Several methods are available for airborne aromatic amine monitoring [165, 166] [167].

6. References

1. M. N. Desai, *Trans. Saest* **16** (1981) no. 2, 77 – 87. *Chem. Abstr.* **95** (1981) 105208 p.
2. *Colour Index,* 3rd ed., Society of Dyers and Colourists, Bradford, England, and American Association of Textile Chemists and Colorists, North Carolina, United States, 1971.
3. United States International Trade Commission, *Publication 1588,* U.S. Government Printing Office, Washington, D.C., 1984.
4. United States International Trade Commission, *Publication 1548,* "Import of Benzenoid Chemicals and Products 1983," Graphics Branch, Washington, D.C., July 1984, pp. 10, 23.
5. Japan Dyestuff Industry Assn., *Japan Chem.* 1982 (April 15) 8.
6. K. Weissermel, H.-J. Arpe, *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998.
7. W. L. Semon in Davis, Bake (ed.): *The Chemistry and Technology of Rubber,* Reinhold Publ. Co., New York 1937, p. 414.
8. R. G. Bauman, J. W. Born, *J. Appl. Polym. Sci.* **1,** (1959) 351.
9. G. S. Petrov, V. F. Prosvirkina, *Zh. Prikl. Khim. (Moscow)* **30,** (1956) 1660. *Chem. Abstr.* **52** (1958) 5028 f.
10. G. Oster, US 3014799, 1961.
11. H. E. Fierz-David, L. Blangey: *Fundamental Processes of Dye Chemistry,* 5th ed., Interscience, New York 1949, pp. 99, 140, 179, 306.
12. Horizon Inc., US 3082086, 1963 (R. H. Sprague).
13. Horizon Inc., US 3056673, 1962 (E. Wainer).
14. Du Pont, US 3673143, 1971 (T. I. Bair, P. W. Morgan).
15. R. E. Wilfong, J. Zimmerman, *Rubber World* **166** (1972) 40, 47.
16. A. Galat, G. Elion, *J. Am. Chem. Soc.* **65** (1943) 1566.
17. G. R. Petit, E. G. Thomas, *J. Org. Chem.* **24** (1959) 895.
18. C. S. Marvel, G. S. Hiers, *Org. Synth. Coll.* **1** (1941) 327.
19. K. H. Slotta, H. Dressler, *Ber. Dtsch. chem. Ges.* **63** (1930) 889, 894. *Houben-Weyl* **9,** 871.
20. F. B. Dains, R. Q. Brewster, C. P. Olander, *Org. Synth. Coll.* **1** (1941) 447.
21. Shell Development Co., US 2580284, 1951 (T. J. Deahl, F. H. Stross, M. D. Taylor).
22. Du Pont, US 3649693, 1975 (D. R. Coulson).
23. Biller, Michealis & Co., US 2991311, 1961 (M. Thoma).
24. A. K. Bhattacharyya, D. K. Naudi, *Ind. Eng. Chem. Prod. Res. Dev.* **14** (1975) no. 3, 162.
25. R. Stroh, J. Ebersberger, H. Haberland, W. Hahn, *Angew. Chem.* **69** (1957) 124.
26. Ethyl Corp., US 4219502, 1980 (K. G. Ihrmann, M. Brandt).
27. Ethyl Corp., US 4128582, 1978 (L. J. Governale, J. C. Wollensak).

28. Ethyl Corp., US 3923892, 1974 (O. H. Klopfer).
29. Ethyl Corp., US 3843698, 1974 (J. H. Dunn).
30. K. D. Hesse, *Justus Liebigs Ann. Chem.* **741** (1970) 117.
31. C. W. Smith: *Acrolein,* J. Wiley & Sons, New York 1962, p. 105.
32. R. C. Elberfield: *Heterocyclic Compounds,* **vol. 4,** J. Wiley & Sons, New York 1952, pp. 6 – 80.
33. H. J. Roth, P. Lepke, *Arch. Pharm.* (Weinheim, Germ.) 305 (1972) 159.
34. *Houben-Weyl* **9,** 40.
35. F. B. Mallory, *Org. Synth. Coll.* **4** (1963) 74.
36. *Houben-Weyl* **4 (1 b),** 662.
37. H. D. Anspon, *Org. Synth. Coll.* **3** (1955) 711. *Houben Weyl* **10 (3),** 332.
38. Sherwin-Williams Co., US 4363914, 1982 (J. W. Long, L. Vacek).
39. Bayer, US 4424360, 1984 (F. Hagedorn, W. Evertz).
40. *Houben-Weyl* **5 (3),** 705.
41. Ethyl Corp., US 2675406, 1954 (H. D. Orloff, J. F. Napolitano).
42. U. S. Rubber Co., GB 588642, 1947.
43. *Houben-Weyl* **10** (1), 1054.
44. A. S. Pagano, W. D. Emmond, *Org. Synth. Coll.* **5** (1973) 367.
45. C. F. H. Allen, J. A. van Allen, *Org. Synth. Coll.* **3** (1955) 824.
46. R. J. Peterson: *Hydrogenation Catalysts,* Noyes Data Corp., Park Ridge, N.J., 1977, pp. 139 – 144, 254 – 267.
47. BASF, US 3799983, 1974 (H. Corr et al.).
48. Upjohn Co., US 3856862, 1974 (T. H. Chung et al.).
49. BASF, DE 805518, 1951 (O. Stichnoth).
50. N. L. Drake, *Org. React.* **1,** (1942) 105.
51. W. W. Hartmann, J. R. Byers, J. B. Dickey, *Org. Synth. Coll.* **2** (1948) 451.
52. *Houben-Weyl* **11 (1),** 341, 394.
53. G. W. Gray et al., *J. Chem. Soc. B.* 1952, 1959.
54. D. C. Owsley, J. J. Bloomfield, *Synthesis* 1977 118.
55. S. A. Cogswell: *Chemical Economics Handbook,* SRI International, Menlo Park, Calif., 1984, p. 611.5032 K.
56. *Houben-Weyl* **11 (1)** 409, 474.
57. H. K. Porter, *Org. React.* **20** (1973) 455.
58. *Houben-Weyl* **7 (1),** 156.
59. Monsanto Chemical Co., US 2745614, 1957 (O. DeGarmo, E. J. McMullen).
60. G. G. Henderson, M. M. J. Sutherland, *J. Chem. Soc. Trans.* **97** (1910) 1616.
61. J. F. Bunnett, R. E. Zahler, *Chem. Rev.* **49** (1951) 398.
62. *Houben-Weyl* **9,** 519.
63. Bayer, US 3417090, 1965 (H. Pelster, C. König, R. Pütter); FR 1531889, 1967 (H. Pelster, C. König, R. Pütter).
64. *Houben-Weyl* **11 (1),** 437.
65. *Houben-Weyl* **11 (1),** 472.
66. R. Jansson, *Chem. Eng. News* **62** (1984) no. 47, 43 – 57.
67. M. M. Baizer, H. Lund: *Organic Electrochemistry,* 2nd ed., Marcel Dekker, New York 1983, pp. 295 – 313.
68. T. Kaufmann et al. in W. Foerst (ed.): *Neure Methoden der präparativen organischen Chemie,* **vol. 4,** Verlag Chemie, Weinheim 1966, p. 62.
69. D. Balcom, A. Fürst, *J. Am. Chem. Soc.* **75** (1953) 4334.
70. A. Fürst et al., *Chem. Rev.* **65** (1965) 51 – 69.
71. T. L. Fletcher et al., *J. Org. Chem.* **25** (1960) 996.
72. E. J. Colter, S. S. Wang, *J. Org. Chem.* **27** (1962) 1517.
73. Filiala Cluj a Academiei R. S. Romania, DE-OS 2007656, 1971 (M. Ionesca, I. Mester).
74. V. V. Korshak et al., SU 218898, 1965. *Ullmann,* 4th ed., 7 : 398.
75. P. M. Bavin, *Org. Synth. Coll.* **5** (1973) 30.
76. R. A. Raphael, E. C. Taylor, H. Wynberger, *Advan. Org. Chem.* **2** (1960) 350.
77. American Cyanamid Co., US 2891094, 1951 (O. C. Kaskalitis et al.).
78. Monsanto, US 3944615, 1976 (A. F. M. Iqbal).
79. Du Pont, DE-AS 1187243, 1965 (J. R. Kosak).
80. S. Tsutsumi, H. Terade, *Kogyo Kagaku Zasshi* **54** (1951) 527.
81. L. F. Albright et al., *Chem. Eng. (New York)* **74** (1967) 197 – 202, 251 – 256.
82. Engelhard-Minerals & Chem. Corp., DE 2042368, 1971 (S. G. Hindin, D. L. Blair, D. R. Steele).
83. Du Pont, US 3499034, 1970 (R. A. Gonzales).
84. Bayer, DE-OS 1643379, 1971 (W. Böhm, A. Wissner).
85. P. N. Rylander: *Catalytic Hydrogenation over Platinum Metals,* Academic Press, New York 1967, p. 169.
86. B. R. James: *Homogeneous Hydrogenation,* J. Wiley & Sons, New York-London 1973, p. 180.
87. Du Pont, US 3328465, 1967 (L. Spiegler).
88. Du Pont, US 3127356, 1964 (J. M. Hamilton, L. Spiegler).
89. Olin Corp., US 3935264, 1976 (S. K. Bhutani).

90. ICI, US 3154584, 1964 (C. Gardner, R. G. New).
91. Allied Chem. Corp., US 2976320, 1961 (L. O. Winstrom, R. D. Samdahl, P. S. Perch).
92. Olin Mathieson Chem. Corp., US 3356729, 1967 (W. I. Denton, P. D. Hammond).
93. Du Pont, US 2619503, 1952 (R. G. Benner, A. C. Stevenson).
94. Mobay Chem. Corp., US 3517063, 1970; BE 645406, 1964 (B. R. Nason).
95. Olin Mathieson Chem. Corp., BE 661047, 1965 (J. J. Cimerol, W. M. Clark, W. I. Denton).
96. Monomers & Des. Inst., SU 233660, 1968 (A. V. Ivanov et al.).
97. *Houben-Weyl* **4 (2),** 137.
98. Allied Chem. & Dye Corp., US 2671763, 1954 (L. O. Winstrom, W. B. Harris).
99. BASF, DE-AS 1233408, 1967 (K. Ehrmann).
100. Office Official Industrial L'azote, BE 646607, 1964.
101. M. Sittig: *Amines, Nitrites and Isocyanates, Processes and Products,* Noyes Dev. Corp., Park Ridge, N.J., 1969.
102. Farbwerke Hoechst AG, US 3761425, 1973 (K. Baessler, K. Mayer).
103. Farbwerke Hoechst AG, US 3929891, 1975; US 3803054, 1974 (K. Habig, K. Baessler).
104. Degussa, US 3941717, 1976 (G. Volheim et al.).
105. Engelhard Minerals & Chem. Corp., DE 2042368, 1970; US 3666813, 1972 (S. G. Hindlin, D. L. Bair).
106. Engelhard Industries Inc., FR 1440419, 1965.
107. G. M. Leuteritz, *Chemie-Anlagen & Verfahren* **3** (March 1971) 49 – 54.
108. G. M. Leuteritz, *Process Eng.* **Dec. 1973** 62, 63.
109. G. M. Leuteritz et al., *Hydrocarbon Process.* **June 1976** 99, 100.
110. R. J. Malone, *Chem. Eng. Prog.* **53** (1980) 53 – 59.
111. M. Day, A. T. Peters Jr., *J. Soc. Dyers Colour.* **83** (1967) no. 4, 137 – 143.
112. S. K. Bhattacharyya et al., *Indian J. Technol.* **5** (1967) no. 7, 232 – 233.
113. Universal Oil, FR 1471866, 1967 (M. Earl, A. Nielsen).
114. Phillips Petroleum Co., US 3484487, 1969 (J. S. Dix).
115. Du Pont, DE-OS 1695590, 1971 (J. Pikl).
116. *Houben-Weyl* **11 (1),** 26.
117. J. F. Bunnet, R. F. Zahler, *Chem. Rev.* **49** (1951) 273.
118. F. B. Wells, C. F. H. Allen, *Org. Synth. Coll.* **2** (1948) 221.
119. F. Pietra, *Q. Rev. Chem. Soc.* **23** (1969) 504 – 521.
120. Etat Francais, FR 1332501, 1963 (J. Issoire).
121. Monsanto, DE-OS 1768518, 1968 (T. A. Garabedian).
122. Mobil Oil Co., US 3231616, 1966 (D. J. Jones).
123. Kanto Elektrochem. Ind. Co., JP-Applic. 671864, 1967; 6726289, 1967.
124. J. F. Bunnet, *Chem. Rev.* **49** (1951) 395.
125. *Houben-Weyl* **11 (1),** 189.
126. General Aniline, GB 1115535, 1966 (C. H. Chang). GAF Corp., US 3468941, 1969 (C. H. Chang).
127. Sterling Drug Inc., US 3102913, 1963 (R. E. Werner).
128. BASF, BE 703695, 1967.
129. Monsanto, US 3170956, 1965 (J. F. Olin).
130. Sherwin-Williams Co., NL 7017339, 1970.
131. Hercules, US 3338966, 1967 (J. C. Snyder); US 3340302, 1967 (H. L. Young).
132. Hercules Powder Co., GB 926897, 1961.
133. Halcon International Inc., FR 1 405 816, 1965 (M. R. S. Barker); BE 740758, 1969 (J. L. Russell); US 3578714, 1971 (J. L. Russell); BE 744936, 1970 (M. Becker, R. Fernandez).
134. Mitsui Petrochemical Ind. Ltd., JP 7123052, 1968.
135. I. McKechnie, F. Bayer, J. Drennan, *Chem. Eng. (N.Y.)* **87** (Dec. 29, 1980) 26 – 27.
136. A. Rieche, H. Seeboth, *Justus Liebigs Ann. Chem.* **638** (1960) 43 – 110.
137. N. L. Drake, *Org. React.* **1** (1942) 110.
138. Bayer, DE 1222051, 1965 (H. J. Schulz, H. S. Bien).
139. Amer. Cyanamid Co., US 2943111, 1960 (G. L. Wiesner); US 3118944, 1964 (G. I. Addis).
140. B. F. Goodrich Co., US 3071619, 1959 (H. J. Kehe, R. T. Johnson et al.).
141. E. D. Hughes, C. K. Ingold, *Quart. Rev.* **6** (1952) 45.
142. H. G. Neumann, E. W. Wieland, *Ind. Med. Surg.* **42** (Nov-Dec 1973) no. 11, 15 – 19.
143. M. Kiese, *Pharmacol. Rev.* **18** (1966) 1091 – 1161.
144. J. L. Radomski, *Annu. Rev. Pharmacol. Toxicol.* **19** (1979) 129 – 157.
145. H. G. Parkes in C. E. Scarle (ed.): *Chemical Carcinogens,* American Chemical Society Monograph No. 173, American Chemical Society, Washington, D.C., 1976, pp. 462 – 480.
146. L. Rehn, *Arch. Klin. Chir.* **50** (1895) 588 – 600.

147. *Second Annual Report on Carcinogens,* U.S. Dept. of Health and Human Services, U.S. Govt. Printing Office, December 1981.
148. International Agency for Research on Cancer (IARC): *Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man (4),* Lyon 1974.
149. International Agency for Research on Cancer (IARC): *Monographs on the Evaluation of Carcino-genic Risk of Chemicals to Man* (16), Lyon 1978.
150. S. S. Thorgeirsson, H. A. J. Schut, P. J. Wirth, E. Dybing (ed.): *Mutagenicity and Carcinogenicity of Aromatic Amines, Molecular Basis of Environmental Toxicity,* Ann Arbor Science Publishers, Inc., Ann Arbor, Mich., 1980, pp. 275 – 292.
151. R. A. Thuraisingham, S. H. M. Nilar, *J. Theor. Biol.* **86** (1981) 557 – 580.
152. K. Yuta, P. C. Jurs, *J. Med. Chem.* **24** (1981) 241 – 251.
153. E. C. Miller, *Can. Res.* **38** (1978) 1479.
154. J. H. Weisburger, G. M. Williams (ed.): *Casarett and Doull's Toxicology, The Basic Science of Poisons,* 2nd ed., Macmillan Publ. Co., New York 1980, p. 93.
155. G. W. Thorn et al. (ed.): *Harrison's Principles of International Medicine,* 8th ed., McGraw-Hill, New York 1977, p. 1710 – 1713.
156. Deutsche Forschungsgemeinschaft (ed.): *Report No. 11 of the Commission for Investigation of Health Hazards of Chemical Compounds in the Work Area,* Bonn, Federal Republic of Germany, 1979.
157. Ontario Ministry of Labor, *Exposure Criteria for Potentially Harmful Agents and Substances in Workplaces,* Toronto, Ontario, 1981.
158. Z. B. Smelyanskiy, I. P. Ulanova, *Gig. Tr. Prof Zabol* **3** (1959) no. 5, 7 – 15.
159. *Recommendation for Permissible Concentration, etc. (Japan), Sangyo Igaku* **13** (1971) 475 – 484.
160. *Permissible Levels of Occupational Exposure to Airborne Toxic Substances,* WHO Technical Report Series No. 415, WHO, France 1968.
161. American Conference of Governmental Industrial Hygienists (ACGIH) (ed.): Threshold Limit Values for Chemical Carcinogens in Work Air (TLV), Cincinnati, Ohio, 1982.
162. Deutsche Forschungsgemeinschaft (DFG) (ed.): *Maximum Concentrations at the Workplace (MAK),* Verlag Chemie, Weinheim 1982.
163. Occupational Safety and Health Administration, *OSHA Safety and Health Standards,* 29 CFR 1910, U.S. Government Printing Office, Washington, D.C., Nov. 7, 1978.
164. American Conference of Governmental Industrial Hygienists (ACGIH) (ed.): *Industrial Ventilation-A Manual of Recommended Practice,* 17th ed., Cincinnati, Ohio, 1982.
165. M. A. Pinches, R. F. Walker, *Ann. Occup. Hyg.* **23** (1980) 335 – 352.
166. E. E. Campbell, G. O. Wood, R. G. Anderson, *IARC Sci. Publ.* **40** (1981) 109 – 118.
167. D. N. Meddle, *Analyst (London)* **106** (1981) Issue 1261, 1088 – 1095.

Amino Acids

Karlheinz Drauz, Degussa AG, Hanau-Wolfgang, Germany
Ian Grayson, Degussa, Middlesbrough, United Kingdom
Axel Kleemann, Hanau, Germany
Hans-Peter Krimmer, Degussa AG, Hanau-Wolfgang, Germany
Wolfgang Leuchtenberger, Degussa AG, Düsseldorf, Germany
Christoph Weckbecker, Degussa AG, Hanau-Wolfgang, Germany

1. Introduction and History 1
2. Properties 2
2.1. Physical Properties and Structure 2
2.2. Chemical Properties 4
2.3. Important Amino Acids 7
2.3.1. Proteinogenic Amino Acids 7
2.3.2. Other Important Amino Acids 9
3. Industrial Production of Amino Acids 13
3.1. General Methods 14
3.2. Production of Specific Amino Acids 15
3.2.1. L-Alanine 15
3.2.2. L-Arginine 15
3.2.3. L-Aspartic Acid and Asparagine . . . 16
3.2.4. L-Cystine and L-Cysteine 17
3.2.5. L-Glutamic Acid 17
3.2.6. L-Glutamine 18
3.2.7. L-Histidine 19
3.2.8. L-Hydroxyproline 19
3.2.9. L-Isoleucine 19
3.2.10. L-Leucine 19
3.2.11. L-Lysine 20
3.2.12. D,L-Methionine and L-Methionine . 21
3.2.13. L-Phenylalanine 21
3.2.14. L-Proline 23
3.2.15. L-Serine 23
3.2.16. L-Threonine 23
3.2.17. L-Tryptophan 24
3.2.18. L-Tyrosine 24
3.2.19. L-Valine 25
4. Biochemical and Physiological Significance 25
5. Uses 26
5.1. Human Nutrition 26
5.1.1. Supplementation 28
5.1.2. Flavorings, Taste Enhancers, and Sweeteners 30
5.1.3. Other Uses in Foodstuff Technology 31
5.2. Animal Nutrition 32
5.3. Pharmaceuticals 35
5.3.1. Nutritive Agents 36
5.3.2. Therapeutic Agents 36
5.4. Cosmetics 40
5.5. Agrochemicals 41
5.5.1. Herbicides 41
5.5.2. Fungicides 43
5.5.3. Insecticides 45
5.5.4. Plant Growth Regulators 45
5.6. Industrial Uses 45
6. Chemical Analysis 46
7. Economic Significance 48
8. Toxicology 49
9. References 50

1. Introduction and History

The proteins, although they occur in an almost infinite variety, are composed of a relatively small number of basic building blocks, all α-amino acids. In addition, the amino acids fulfill certain regulatory functions in the metabolism and are required for the biosynthesis of other functional structures. This review is limited, for the most part, to the protein-forming amino acids, because they are by far the most widely distributed in nature and are of considerable economic interest.

The ca. 20 different α-amino acids found in proteins are simple organic compounds, in which an amino group and a side chain (R) are attached alpha to the carboxyl function. The R group may be aliphatic, aromatic, or heterocyclic and may possess further functionality.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_057.pub2

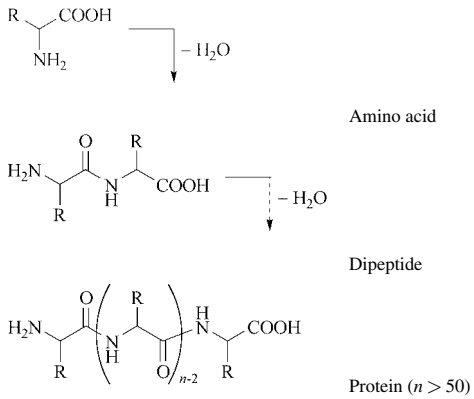

At present over 200 naturally occurring α-amino acids are known [1–7]. Table 1 shows the α-amino acids found in proteins, where they occur exclusively as the L-enantiomers. D-Amino acids have been found only in the cell walls of some bacteria, in peptide antibiotics, and in the cell pools of some plants [10, 14, 15]. The L-amino acids are commonly abbreviated with three-letter or one-letter codes, as shown in Table 1 [4]. In addition, the abbreviation Glx or Z is used to refer to either glutamic acid or glutamine, and the abbreviation Asx or B is used to refer to either aspartic acid or asparagine. More recently

History. The history of amino acid chemistry began in 1806, when two French investigators, Vauquelin and Robiquet, isolated asparagine from asparagus juice. It was not until 1925 that Schryver and Burton isolated threonine from oat protein, the last of the ca. 20 protein-forming amino acids to be discovered. Strecker synthesized alanine in 1850 from acetaldehyde, ammonia, and hydrogen cyanide. Escher established the hypothesis of essential amino acids. Emil Fischer discovered that the amino acids were building blocks of the proteins. Abderhalden synthesized threonine from acrylic acid derivatives and methanol. Rose et al. recognized threonine as the last of the eight essential amino acids. D,L-Methionine was produced industrially in Germany in 1948, and in 1956 L-glutamic acid was produced by fermentation in Japan.

Origin of Amino Acids. The first amino acids were probably produced on the earth more than 3×10^9 years ago via "prebiotic synthesis" in the primordial atmosphere. The concept of prebiotic synthesis is based on laboratory experiments in which glycine, alanine, aspartic acid, glutamic acid, and other compounds were produced by the action of an electrical discharge on a simulated primordial atmosphere consisting of methane, hydrogen, water, and ammonia [16]. Since then, traces of amino acids have been detected in moon rocks, meteorites, and interstellar space.

2. Properties

2.1. Physical Properties and Structure

α-Amino acids are nonvolatile, white, crystalline compounds with no defined melting points. They are relatively stable on heating, generally decomposing at 250–300 °C. Both the low volatility and the thermal stability result from the low-energy dipolar structure (zwitterion, inner salt, betaine), which the amino acids assume in the solid state.

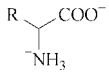

Evidence for this structure is provided by infrared and Raman spectra in which the bands typical of -NH_2 and -COOH moieties are absent. Equilibrium in solution also lies almost exclusively on the side of the dipolar form; therefore, amino acids are insoluble in nonpolar solvents and usually not very soluble in polar ones. The only amino acids that exhibit any appreciable solubility in alcohol are proline and hydroxyproline. Solubility in water depends on the pH: the minimum is at the isoelectric point.

This solubility minimum at the isoelectric point is quite useful for purifying and recrystallizing amino acids. The analytical technique for separating amino acid mixtures by electrophoresis is based on the fact that a specific amino acid does not migrate in an electric field at its isoelectric point, p*I*, a physical constant for each amino acid.

The structures and physical properties of the most important α-amino acids are given in Section 2.3.1.

Stereochemistry. With the exception of glycine, the simplest amino acid (R = H), all natural α-amino acids are chiral compounds occurring in two enantiomeric (mirror-image) forms.

Table 1. Common amino acids of proteins

Trivial name	IUPAC abbreviation	One-Letter Code	Protein	Content, g/100 g	
				[8]	[10]
L-Alanine	Ala	A	silk fibroin	25	29.7
L-Arginine	Arg	R	salmin	87	
			edostin	17	
			wool	10	
			gelatin		8.3
			rat liver histone		15.9
L-Asparagine	Asn	N			
L-Aspartic acid	Asp	D	edestin	12	
			hemoglobin	9–10	
			barley globulin		10.3
L-Cysteine	Cys	C	wool keratin		11.9
			human hair keratin		14.4
			feather keratin		8.2
L-Cystine	Cys-Cys, $(Cys)_2$, Cyss	C			
L-Glutamic acid	Glu	E	gliadin	47	
			zein	31	
			wheat gliadin		39.2
			maize zein		22.9
L-Glutamine	Gln	Q			
Glycine	Gly	G	gelatin	26	
			silk fibroin	44	
L-Histidine	His	H	hemoglobin	7	
L-Hydroxyproline	Hyp	–	gelatin	15	
L-Isoleucine	Ile	I	edestin	21	
			hemoglobin	29	
			serum proteins	20	
			oat globulin		4.3
			beef serum albumin		2.6
L-Leucine	Leu	L	edestin	21	
			hemoglobin	29	
			serum proteins	20	
			maize zeins		19
L-Lysine	Lys	K	serum albumin	13	
			serum globulin	6	
			horse myoglobin		15.5
L-Methionine	Met	M	egg albumin	5	
			casein	3	4.1
			β-lactoglobulin		3.2
L-Phenylalanine	Phe	F	zein	8	
			egg albumin	5	7.7
			serum albumin		7.8
L-Proline	Pro	P	gelatin	17	16.3
			gliadin	13	
			salmin		6.9
			casein		10.6
L-Serine	Ser	S	silk fibroin	13	16.2
			trypsinogen		16.7
			pepsin		12.2
L-Threonine	Thr	T	casein	4	
			human hair keratin		8.5
			avidin		10.5
L-Tryptophan	Trp	W	fibrin	3	
			egg lysozyme		10.6
L-Tyrosine	Tyr	Y	fibrin	6	
			silk fibroin	13	
			papain		14.7
L-Valine	Val	V	casein	8	
			beef sinew		17.4
			beef aorta		17.6

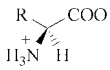

L-Amino acid

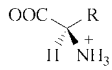

D-Amino acid

The prefixes L and D express the absolute configuration at the α-carbon atom by means of the formal stereochemical relationship to L- or D-glyceraldehyde, the reference substance introduced by Emil Fischer in 1891. In addition to the spacial representations shown above, the Fischer projections are also universally recognized and used:

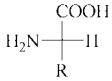

L-Amino acid

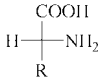

D-Amino acid

Polarimetric determination of the specific rotation $[\alpha]^t_D$ can be used to differentiate between the two enantiomers and to check their optical purity. The molecular rotation $[M]^t_D$ is less common:

$$[M]^t_D = \frac{M_r}{100} \cdot [\alpha]^t_D$$

M_r molecular mass; t temperature; D 589.3 nm (wavelength of the sodium D line). Other methods for investigating the structure of amino acid enantiomers include the Cotton effect (change in molecular rotation as a function of the wavelength of plane-polarized light), as reflected in optical rotational dispersion (reversal of the direction of the molecular rotation at the wavelength of the absorption maximum), and circular dichroism (differing absorption for left- and right-handed circularly polarized light). L-Amino acids exhibit a positive carbonyl Cotton effect, D-amino acids a negative one.

Isoleucine, threonine, and hydroxyproline each contain two chiral carbon atoms; therefore, they appear in four stereoisomeric forms. Cystine, which likewise contains two chiral carbons, has only three stereoisomers: L-, D-, and *meso*-cystine, the meso form having a plane of symmetry. According to the Cahn–Ingold–Prelog rule (R, S rule) [17], all proteinogenic L-α-amino acids, with the exception of L-cysteine and L-cystine, are S; the D-α-amino acids are R. According to this system, L-threonine, for example, is termed (2S, 3R)-threonine. However, the L- and D-descriptors are still in widespread use for the simple L-α-amino acids.

Absorption Spectra. Aliphatic amino acids exhibit no absorption in the UV region above 220 nm, with the exception of cystine (240 nm). The aromatic amino acids, phenylalanine, tyrosine, and tryptophan, absorb between 250 and 300 nm [1, vol. 2]. The exact position of the maximum and the molar extinction coefficient ϵ are affected by the pH of the aqueous solution.

Two infrared bands are especially characteristic of amino acids: 1560–1600 cm^{-1} ($-COO^-$) and ca. 3070 cm^{-1} $\left(-NH_3^+\right)$. These bands are also evidence of the bipolar nature of the amino acids.

2.2. Chemical Properties

Acidity and Basicity. The chemical properties of the α-amino acids are primarily the properties of the amino and carboxyl groups. Amino acids react with strong acids as proton acceptors (bases) and with strong bases as proton donors (acids). In acidic medium they are present predominantly as cations; in basic medium they are present predominantly as anions. Figure 1 shows the titration curves of glycine, glutamic acid, and lysine. The pK_1 and pK_2 values correspond to the inflection points of the titration curve, the pH value where the concentrations of the zwitterion form and the cationic or anionic form are equal. The pK_1, pK_2, and pI values can be calculated from the titration curves with the Henderson–Hasselbalch equation:

$$pK_1 = pH - \log \frac{\left[R-CH(NH_3^+)-COO^-\right]}{\left[R-CH(NH_3^+)-COOH\right]}$$

$$pK_2 = pH - \log \frac{\left[R-CH(NH_2)-COO^-\right]}{\left[R-CH(NH_3^+)-COO^-\right]}$$

$$pI = \frac{pK_1 + pK_2}{2}$$

Amino acids with additional basic or acidic groups (arginine, lysine, glutamic acid, cysteine)

exhibit additional pK values. Amino acids act as buffers in the region of their pK values.

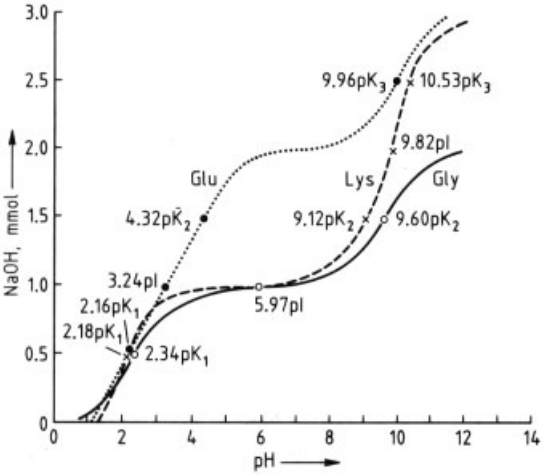

Figure 1. Titration curves of glycine, glutamic acid, and lysine [9]

The pK_1 values show the amino acids to be considerably stronger acids than acetic acid. However, because of intramolecular protonation of the amine moiety by the carboxyl group, aqueous solutions of amino acids are only weakly acidic. The pH values of aqueous monoaminomonocarboxylic acids lie between 5.5 and 6.0. Solutions of the acidic amino acids aspartic acid and glutamic acid have pH values of ca. 2. The weakly basic amino acid histidine has a pH of 7.5 in aqueous solution; the more strongly basic amino acids lysine and arginine have pH values of ca. 11–12. These differences in acidities and basicities are utilized in the separation of amino acid mixtures by ion-exchange chromatography and electrophoresis.

Reactions. Because of their bifunctional and sometimes trifunctional character, the α-amino acids are capable of taking part in a variety of chemical reactions. Comprehensive treatments of these may be found in monographs and reviews [1, 9–12, 18].

The introduction of groups protecting the amino, carboxyl, and side-chain functions is especially important for peptide synthesis [13]. Additionally, α-amino acids play a prominent role as intermediates in the synthesis of heterocycles [19]. The chiral α-amino acids and their derivatives are inexpensive and, for the most part, readily available synthons for numerous natural products and pharmaceuticals. In many cases optically active amino acids can also be used to induce chirality during the course of a synthetic process [20].

α-Amino acids form *chelate-like complexes* with heavy metal ions. The best known are the dark blue, easily crystallized copper chelates:

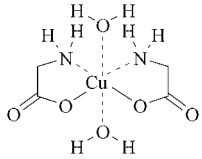

Bis(glycinato) copper(II) hydrate

This complex formation can be used to protect both the α-amino and the carboxyl group during synthesis of ϵ-N-acetyl derivatives or carbamates of lysine or other side-chain derivatives. In these copper or cobalt complexes the α-carbon atom is activated sufficiently to react with aldehydes. A well-known example is the alkaline condensation of the glycine–copper complex with acetaldehyde, resulting in threonine. Especially important is the reaction of α-amino acids with ninhydrin to form a blue-violet dye, the basis of a sensitive optical method for detecting amino acids (see Chap. 6).

Free amino acids react with *nitrous acid* to yield α-hydroxycarboxylic acids, with retention of configuration. Volumetric measurement of the nitrogen gas set free is the basis of Van Slyke's method for the quantitative analysis of amino acids. Reaction of amino acid esters with nitrous acid gives the acid-labile diazocarboxylic acid esters. Treatment of N-alkyl- or N-arylamino acids with nitrous acid yields the N-nitroso derivatives, which can be dehydrated to sydnones in the presence of acetic anhydride:

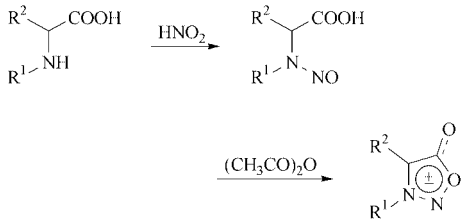

α-Amino acids have found use as *organocatalysts* for asymmetric organic reactions, for example asymmetric versions of the aldol or the Mannich reactions [21]. The amino acids most often employed are L- or D-proline, their derivatives, or short-chain peptides based on proline.

Oxidizing agents attack the amino group, converting the amino acid into an iminocarboxylic acid. These are unstable and either hydrolyze to α-oxocarboxylic acids or decompose, after decarboxylation, into ammonia and the aldehyde containing one carbon atom less. Diketones, triketones, *N*-bromosuccinimide, or silver oxide may serve as the oxidizing agent. The best known example is the reaction with ninhydrin. Oxidative deamination also can be carried out enzymatically with D- or L-amino acid oxidases. This also proceeds via the α-iminocarboxylic acids, which subsequently are hydrolyzed to α-oxo acids:

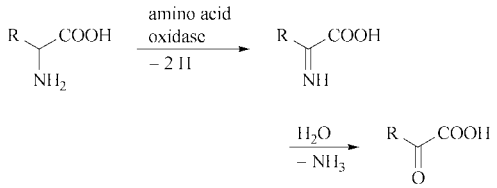

This enantioselective enzymatic oxidative deamination is the basis of analytical methods for the determination of amino acid enantiomers.

The *N-acylation* of α-amino acids with acyl chlorides or anhydrides under Schotten-Baumann conditions produces *N*-acyl-α-amino acids, which have numerous uses. For example, the *N*-acetyl derivatives of D,L-amino acids (e.g., alanine, valine, methionine, phenylalanine, tryptophan) are intermediates in the production of L-amino acids by enzymatic resolution using aminoacylases (see Section 3.1). Amino acids acylated with naturally occurring fatty acid residues are used industrially as biodegradable surfactants.

If free amino acids are heated above 200 °C, especially in the presence of soda lime or metal ions, they readily *decarboxylate* to form amines. The enzymatic decarboxylation of amino acids gives biogenic amines. Some of these amines are physiologically active neurotransmitters (histamine, tyramine, dopamine, serotonin).

The *esters* are usually prepared by direct esterification, e.g., reaction of the α-amino acid with anhydrous alcohol in the presence of anhydrous hydrogen chloride. The initial product is the hydrochloride of the ester, which is liberated by addition of base. On standing or warming, the free esters of α-amino acids can eliminate alcohol to form 2,5-diketopiperazines:

Amino acid esters are useful intermediates for peptide synthesis because the carboxyl function is protected. The esters may be converted to amino alcohols by treatment with a strong reducing agent, such as lithium aluminum hydride.

Several cyclic derivatives of the α-amino acids are of industrial importance. The *hydantoins* or imidazolidine-2,4-diones, which have been used as intermediates in the synthesis of α-amino acids, are most conveniently prepared by treatment of aldehyde cyanohydrins with ammonium carbonate or urea. They also may be obtained by reacting amino acids with cyanates or isocyanates, the reaction proceeding via the ureidocarboxylic acid (hydantoic acid derivative):

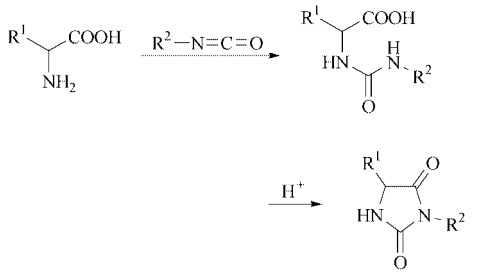

Hydantoin derivative

The thiohydantoins are obtained by the reaction with isothiocyanates.

Dehydration of *N*-acylamino acids with acetic anhydride or carbodiimide yields *1,3-oxazolin-5-ones* (azlactones):

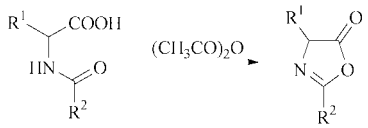

Racemization occurs readily during the reaction. The azlactones are intermediates in the synthesis of amino acids.

The *N-carboxylic acid anhydrides* (Leuchs anhydrides, 1,3-oxazolidine-2,5-diones) are often used in peptide synthesis, especially to prepare poly-α-amino acids. These reactive amino acid derivatives are obtained by treating the amino acid with phosgene in the presence of tertiary amines

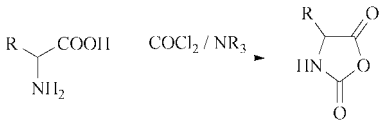

or by elimination of benzyl chloride from *N*-benzyloxycarbonylamino acid chlorides.

2.3. Important Amino Acids

2.3.1. Proteinogenic Amino Acids

L-Alanine [*56-41-7*], 2-aminopropionic acid, $C_3H_7NO_2$, M_r 89.09, *mp* 314 °C (decomp.), $[\alpha]_D^{25}$ + 14.47° (c = 10.03 in 6 M HCl), solubility 16.51 (25 °C) g/100 g H_2O, p*I* 6.01, dissociation constants: pK_1 2.34, pK_2 9.69.

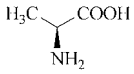

L-Arginine [*74-79-3*], 2-amino-5-guanidinovaleric acid, 2-amino-5[(aminoiminomethyl)amino]pentanoic acid, $C_6H_{14}N_4O_2$, M_r 174.20, *mp* 244 °C (decomp.), $[\alpha]_D^{25}$ + 27.58° (c = 2 in 6 M HCl), solubility 14.87 (20 °C) g/100 g H_2O, p*I* 10.76, dissociation constants: pK_1 2.01, pK_2 9.04 (α-NH_2), pK_3 12.48 (guanidyl). Hydrochloride: [*1119-34-2*], $C_6H_{15}ClN_4O_2$, M_r 210.66, *mp* 220 °C (decomp.), $[\alpha]_D^{25}$ + 11.7° (c = 5 in H_2O), solubility 75.1 (20 °C) g/100 g H_2O.

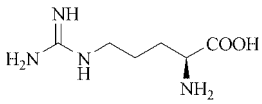

L-Asparagine [*70-47-3*], 2-aminosuccinamic acid, 2,4-diamino-4-oxobutanoic acid, $C_4H_8N_2O_3$, M_r 132.13, *mp* 236 °C (decomp.), $[\alpha]_D^{20}$ + 32.6° (c = 1 in 0.1 M HCl), solubility 3.11 (28 °C) g/100 g H_2O, p*I* 5.41, dissociation constants: pK_1 2.02, pK_2 8.8.

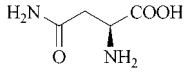

L-Aspartic acid [*56-84-8*], aminosuccinic acid, 2-amino-1,3-butanedioic acid, $C_4H_7NO_4$, M_r 133.10, *mp* 270 °C (decomp., sealed tube), $[\alpha]_D^{25}$ + 25.4° (c = 2 in 5 M HCl), solubility 0.5 (25 °C) g/100 g H_2O, p*I* 2.98, dissociation constants: pK_1 2.1, pK_2 3.86 (β-COOH), pK_3 9.82.

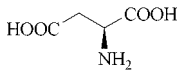

L-Cysteine [*52-90-4*], 2-amino-3-mercaptopropionic acid, 3-mercaptoalanine, $C_3H_7NO_2S$, M_r 121.16, *mp* 240 °C (decomp.), $[\alpha]_D^{25}$ + 9.7° (c = 8 in 1 M HCl), solubility 28 (25 °C) g/100 mL solution, 16 (20 °C) g/100 g H_2O, p*I* 5.02, dissociation constants: pK_1 1.71, pK_2 8.27 (–SH), pK_3 10.78. Hydrochloride monohydrate: [*7048-04-6*], $C_3H_{10}ClNO_3S$, M_r 175.64, *mp* (anhydr.) 178 °C (decomp.), $[\alpha]_D^{25}$ + 6.53° (calculated as cysteine) (c = 2 in 5 M HCl), solubility >100 (20 °C) g/100 g H_2O.

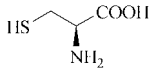

L-Cystine [*56-89-3*], 2,2′-diamino-3, 3′-dithiobis(propionic acid), 3,3′-dithiobis(2-aminopropanoic acid), $C_6H_{12}N_2O_4S_2$, M_r 240.30, *mp* 260°C (decomp.), $[\alpha]_D^{25}$ −212° (c = 1 in 1 M HCl), solubility 0.011 (25 °C) g/100 g H_2O, p*I* 5.02, dissociation constants: pK_1 1.04, pK_2 2.05 (-COOH), pK_3 8.0 (–NH_2), pK_4 10.25 (–NH_2).

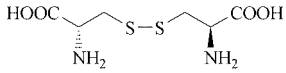

L-Glutamic acid [*56-86-0*], 2-aminoglutaric acid, 2-amino-1,4-pentanedioic acid, $C_5H_9NO_4$, M_r 147.13, *mp* 224–225 °C (decomp.), $[\alpha]_D^{25}$ +31.5° (c = 2 in 5 M HCl, 20 °C), solubility = 0.843 (25 °C) g/100 g H_2O, p*I* 3.08, dissociation constants: pK_1 2.1, pK_2 4.07, pK_3 9.47.

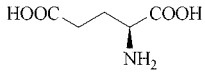

L-Glutamine [*56-85-9*], 2-amino-glutaramic acid, 2,5-diamino-5-oxopentanoic acid, $C_5H_{10}N_2O_3$, M_r 146.15, *mp* 185-6 °C (decomp.), $[\alpha]_D^{25}$ + 31.8° (*c* = 2 in 1 M HCl), solubility 3.6 (19 °C) g/100 g H_2O, p*I* 5.65, dissociation constants: pK_1 2.17, pK_2 9.13.

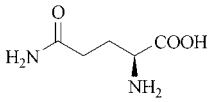

Glycine [*56-40-6*], aminoacetic acid, $C_2H_5NO_2$, M_r 75.07, *mp* 262, 292 °C (decomp.), solubility 24.99 (25 °C) g/100 g H_2O, p*I* 6.06, dissociation constants: pK_1 2.35, pK_2 9.13.

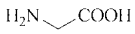

L-Histidine [*71-00-1*], α-amino-1*H*-imidazole-4-propionic acid, 1*H*-imidazole-4-alanine, $C_6H_9N_3O_2$, M_r 155.16, *mp* 277, 287 °C (decomp.), $[\alpha]_D^{25}$ + 13.0° (*c* = 1 in 6 M HCl), solubility 4.29 (25 °C) g/100 g H_2O, p*I* 7.64, dissociation constants: pK_1 1.77, pK_2 6.1 (imidazolyl), pK_3 9.18. Hydrochloride monohydrate: [*5934-29-2*], $C_6H_{12}ClN_3O_3$, M_r 209.63, *mp* 259 °C (decomp.), solubility 16.99 (20 °C) g/100 g H_2O.

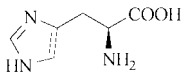

L-Hydroxyproline [*51-35-4*], *trans*-4-hydroxy-2-pyrrolidinecarboxylic acid, $C_5H_9NO_3$, M_r 131.13, *mp* 274 °C (decomp.), $[\alpha]_D^{22}$ −75.2° (*c* = 2 in H_2O), solubility 36.11 (25 °C) g/100 g H_2O, p*I* 5.82, dissociation constants: pK_1 1.92, pK_2 9.73.

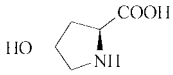

L-Isoleucine [*73-32-5*], 2-amino-3-methylvaleric acid, 2-amino-3-methylpentanoic acid, $C_6H_{13}NO_2$, M_r 131.18, *mp* 285-6 °C (decomp.), $[\alpha]_D^{25}$ + 40.6° (*c* = 2 in 6 M HCl), solubility 4.117 (25 °C) g/100 g H_2O, p*I* 6.02, dissociation constants: pK_1 2.36, pK_2 9.68.

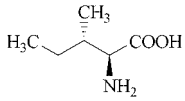

L-Leucine [*61-90-5*], 2-amino-4-methylvaleric acid, 2-amino-4-methylpentanoic acid, 2-aminoisocaproic acid, $C_6H_{13}NO_2$, M_r 131.18, *mp* 293-5, 314-5 °C (decomp.), $[\alpha]_D^{25}$ −10.9° (*c* = 0.52 in H_2O), solubility 2.19 (25 °C) g/100 g H_2O, p*I* 5.98, dissociation constants: pK_1 2.36, pK_2 9.6.

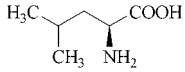

L-Lysine [*56-87-1*], 2,6-diaminohexanoic acid, 2,6-diaminocaproic acid, $C_6H_{14}N_2O_2$, M_r 146.19, *mp* 225-5 °C (decomp.), $[\alpha]_D^{25}$ + 25.9° (*c* = 2 in 5 M HCl), solubility >100 (25 °C) g/100 g H_2O, p*I* 9.47, dissociation constants: pK_1 2.18, pK_2 8.95 (α-NH_2), pK_3 10.53. Hydrochloride: [*657-27-2*], $C_6H_{15}ClN_2O_2$, M_r 182.65, *mp* 253-6 °C (decomp.), solubility 72.5 (25 °C) g/100 mL solution.

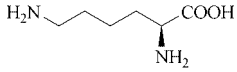

L-Methionine [*63-68-3*], 2-amino-4-(methylthio)butyric acid, 2-amino-4-(methylthio)butanoic acid, $C_5H_{11}NO_2S$, M_r 149.21, *mp* 283 °C (decomp.), $[\alpha]_D^{28}$ + 23.4° (*c* = 5 in 3 M HCl), solubility 5.37 (20 °C) g/100 g H_2O, p*I* 5.74, dissociation constants: pK_1 2.28, pK_2 9.21.

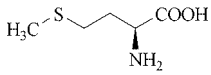

L-Phenylalanine [*63-91-2*], 2-amino-3-phenylpropionic acid, α-aminobenzenepropanoic acid, $C_9H_{11}NO_2$, M_r 165.19, *mp* 283-4 °C (decomp.), $[\alpha]_D^{20}$ −35.1° (*c* = 2 in H_2O), solubility 2.965 (25°C) g/100 g H_2O, p*I* 5.48, dissociation constants: pK_1 1.83, pK_2 9.13.

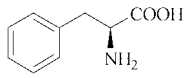

L-Proline [*147-85-3*], 2-pyrrolidinecarboxylic acid, 2-carboxypyrrolidine, $C_5H_9NO_2$,

M_r 115.13, *mp* 220–222 °C (decomp.), $[\alpha]_D^{25}$ −85.0° ($c = 1$ in H_2O), solubility 162.3 (25 °C) g/100 g H_2O, p*I* 6.3, dissociation constants: pK_1 2.0, pK_2 10.6.

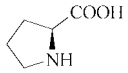

L-Serine [*56-45-1*], 2-amino-3-hydroxypropionic acid, 2-amino-3-hydroxypropanoic acid, $C_3H_7NO_3$, M_r 105.09, *mp* 228 °C (decomp.), $[\alpha]_D^{26}$ −6.8° ($c = 10$ in H_2O), solubility 35.97 (20 °C) g/100 g H_2O, p*I* 5.68, dissociation constants: pK_1 2.21, pK_2 9.15.

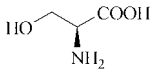

L-Threonine [*72-19-5*], 2-amino-3-hydroxybutyric acid, 2-amino-3-hydroxybutanoic acid, $C_4H_9NO_3$, M_r 119.12, *mp* 253 °C (decomp.), $[\alpha]_D^{26}$ −28.6° ($c = 2$ in H_2O), solubility 9.03 (20 °C) g/100 g H_2O, p*I* 6.16, dissociation constants: pK_1 2.71, pK_2 9.62.

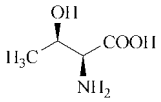

L-Tryptophan [*73-22-3*], 2-amino-3-(3′-indolyl)propionic acid, α-amino-1*H*-indole-3-propanoic acid, $C_{11}H_{12}N_2O_2$, M_r 204.23, *mp* 290–292 °C, 281 °C (decomp.), $[\alpha]_D^{25}$ −32.15° ($c = 1$ in H_2O), solubility 1.14 (25 °C) g/100 g H_2O, p*I* 5.88, dissociation constants: pK_1 2.38, pK_2 9.39.

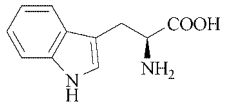

L-Tyrosine [*60-18-4*], 2-amino-3-(4-hydroxyphenyl)propionic acid, [3-(4-hydroxyphenyl)]alanine, 2-amino-3-(*p*-hydroxyphenyl)propionic acid, α-amino-4-hydroxybenzenepropanoic acid, $C_9H_{11}NO_3$, M_r 181.19, *mp* 342-4 (sealed tube), 297–298 °C (decomp.), $[\alpha]_D^{25}$ −7.27° ($c = 4$ in 6 M HCl), solubility 0.045 (25 °C) g/100 g H_2O, p*I* 5.63, dissociation constants: pK_1 2.2, pK_2 9.11, pK_3 10.07 (–OH).

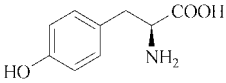

L-Valine [*72-18-4*], 2-amino-3-methylbutyric acid, 2-aminoisovaleric acid, $C_5H_{11}NO_2$, M_r 117.15, *mp* 315 °C (decomp.), $[\alpha]_D^{20}$ +26.7° ($c = 3.4$ in 6 M HCl), solubility 8.85 (25 °C) g/100 g H_2O, p*I* 5.96, dissociation constants: pK_1 2.32, pK_2 9.62.

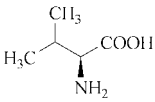

2.3.2. Other Important Amino Acids

β-Alanine [*107-95-9*], 3-aminopropionic acid, $C_3H_7NO_2$, M_r 89.09, dissociation constant: pK_1 3.6, occurrence: apple, constituent of pantothenic acid, carnosine, anserine.

D-Alanine [*338-69-2*], 2-aminopropionic acid, D-Ala, $C_3H_7NO_2$, M_r 89.09, application: LHRH-antagonists, e.g., Abarelix (antineoplastic) [22].

D,l-Alanine [*302-72-7*], 2-aminopropionic acid, $C_3H_7NO_2$, M_r 89.09, *mp* 295 °C (decomp.), solubility 16.72 (25 °C) g/100 g H_2O, p*I* 6.11, dissociation constants: pK_1 2.35, pK_2 9.87.

α-Aminoisobutyric acid [*62-57-7*], 2-amino-2-methylpropionic acid, Aib, $C_4H_9NO_2$, M_r 103.12, application: growth hormone secretagogues, e.g., MK-0677 [23].

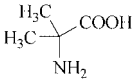

L-α-Aminobutyric acid [*1492-24-6*], 2-aminobutyric acid, L-Abu, $C_4H_9NO_2$, M_r 103.12, application: Levetiracetam (anticonvulsant) [24].

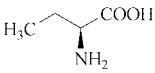

γ-Aminobutyric acid (GABA) [*56-12-2*], 4-aminobutyric acid, $C_4H_9NO_2$, M_r 103.12, occurrence: citrus fruits, sugar beet, brain.

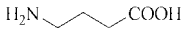

D,L-Aspartic acid [*617-45-8*], aminosuccinic acid, 2-amino-1,3-butanedioic acid, $C_4H_7NO_4$, M_r 133.10, *mp* 275 °C (decomp., sealed tube), solubility 0.775 (25 °C) g/100 g H_2O, p*I* 2.98, dissociation constants: pK_1 2.1, pK_2 3.86 (β-COOH), pK_3 9.82.

L-Azetidine-2-carboxylic acid [*2133-34-8*], $C_4H_7NO_2$, M_r 101.10, application: Ximelagatran (thrombin inhibitor) [25].

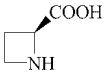

Betaine [*107-43-7*], carboxymethyl-trimethyl ammonium betaine, $C_5H_{11}NO_2$, M_r 117.15, occurrence: sugar beet.

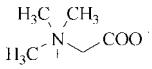

L-Carnitine [*541-15-1*], (3-carboxy-2-hydroxypropyl)trimethyl ammonium betaine, $C_7H_{15}NO_3$, M_r 161.2, occurrence: Lys metabolite, muscle.

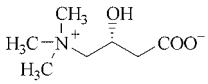

L-Citrulline [*372-75-8*], 2-amino-5-ureidopentanoic acid, $C_6H_{13}N_3O_3$, M_r 175.19, *mp* 234–237 °C , 222 °C (decomp.), $[\alpha]_D^{25}$ + 24.2° (c = 2 in 5 M HCl), solubility 10.3 (20 °C) g/100 g H_2O, p*I* 5.92, dissociation constants: pK_1 2.43, pK_2 9.41; occurrence: urea cycle, watermelon.

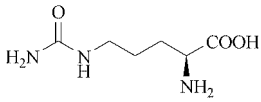

D-Citrulline [*13594-51-9*], 2-amino-5-ureidopentanoic acid, D-Cit, $C_6H_{13}N_3O_3$, M_r 175.19, application: Cetrorelix (antineoplastic) [26].

Creatine [*57-00-1*], *N*-(aminoiminomethyl)-*N*-methylglycine, *N*-methylguanetic acid, $C_4H_9N_3O_2$, M_r 131.14, occurrence: muscle of vertebrates.

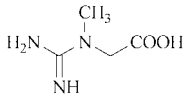

D-Cyclohexylalanine [*58717-02-5*], 2-amino-3-cyclohexylpropionic acid, D-Cha, $C_9H_{17}NO_2$, M_r 171.25, application: (thrombin inhibitors) [27].

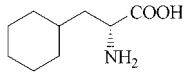

(3*S*,4a*S*,8a*S*)-Decahydroisoquinolinecarboxylic acid [*115238-58-9*], decahydroisoquinoline-3-carboxylic acid, $C_{10}H_{17}NO_2$, M_r 183.26, application: Nelfinavir (antiviral) [28].

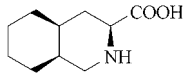

L-2,3-Diaminopropionic acid [*4033-39-0*], L-Dap, $C_3H_8N_2O_2$, M_r 104.11, application: Imidapril (antihypertensive) [29].

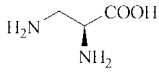

l-3,4-Dihydroxyphenylalanine (DOPA) [*59-92-7*], 2-amino-3-(3,4-dihydroxyphenyl)propionic acid, $C_9H_{11}NO_4$, M_r 197.17, occurrence: Tyr metabolite, faba bean.

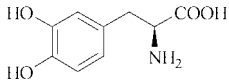

D-Glutamic acid [*6893-26-1*], 2-aminoglutaric acid, 2-amino-1,4-pentanedioic acid, D-Glu, $C_5H_9NO_4$, M_r 147.12, application: Spiroglumide (CCK-B-antagonist) [30].

L-Homocysteine [*6027-13-0*], 2-amino-4-mercaptobutyric acid, $C_4H_9NO_2S$, M_r 135.18, occurrence: Met metabolite, mushrooms, application: Omapatrilat (ACE inhibitor) [31].

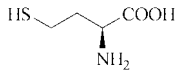

D-*p*-Hydroxyphenylglycine [*22818-40-2*], amino-(4-hydroxyphenyl)acetic acid, D-Phg(OH), $C_8H_9NO_3$, M_r 167.15, application: Amoxicillin (antibiotic) [32].

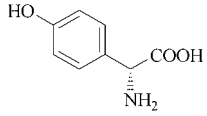

L-5-Hydroxytryptophan [*4350-09-8*], 2-amino-3-(5-hydroxy-1*H*-indol-3-yl)propionic acid, $C_{11}H_{12}N_2O_3$, M_r 220.22, occurrence: serotonin precursor.

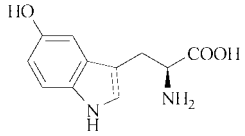

D,L-Isoleucine [*443-79-8*], 2-amino-3-methylvaleric acid, 2-amino-3-methylpentanoic acid, $C_6H_{13}NO_2$, M_r 131.18, *mp* 292 °C (decomp.), solubility 2.011 (25 °C) g/100 g H_2O, p*I* 6.04, dissociation constants: pK_1 2.32, pK_2 9.76.

D,L-Leucine [*328-39-2*], 2-amino-4-methylvaleric acid, 2-amino-4-methylpentanoic acid, 2-aminoisocaproic acid, $C_6H_{13}NO_2$, M_r 131.18, *mp* 293-5, 332 °C (decomp.), solubility 1.00 (25 °C) g/100 g H_2O, p*I* 6.04, dissociation constants: pK_1 2.33, pK_2 9.74.

L-*tert*-Leucine [*20859-02-3*], 2-amino-3,3-dimethylbutyric acid, L-Tle, $C_6H_{13}NO_2$, M_r 131.18, application: div. pharmaceuticals [33].

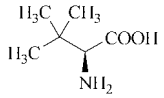

D,L-Lysine hydrochloride [*70-53-1*], $C_6H_{15}ClN_2O_2$, M_r 182.65, *mp* 264 °C (decomp.), solubility 35.98 (20 °C) g/100 g H_2O.

D,L-Methionine [*59-51-8*], 2-amino-4-(methylthio)butyric acid, 2-amino-4-(methylthio)butanoic acid, $C_5H_{11}NO_2S$, M_r 149.21, *mp* 281 °C (decomp.), solubility 3.35 (25 °C) g/100 g H_2O, p*I* 5.74, dissociation constants: pK_1 2.28, pK_2 9.21.

***S*-Methyl-L-cysteine** [*1187-84-4*],2-amino-3-methylthiopropionic acid L-Cys(Me), $C_4H_9NO_2S$, M_r 135.18, application: Kynostatin-272 (antiviral) [34].

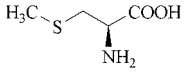

L-*S*-Methylmethionine (vitamin U) [*4727-40-6*], 3-amino-3-carboxy-propyl)-dimethyl sulfonium, $C_6H_{13}NO_2S$, M_r 163.24, occurrence: cabbage, asparagus.

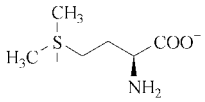

D-3-(2′-Naphthyl)-alanine [*76985-09-6*], 2-amino-3-naphthalen-2-ylpropionic acid, D-Nal, $C_{13}H_{13}NO_2$, M_r 215.25, application: LHRH-antagonists, e.g., Abarelix (antineoplastic) [35].

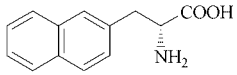

D-Ornithine [*70-26-8*], 2,5-diaminovaleric acid, $C_5H_{12}N_2O_2$, M_r 132.16, *mp* 140 °C (decomp.), $[\alpha]_D^{25}$ + 16.5° (c = 4.6 in H_2O), solubility, p*I* 9.7, dissociation constants: pK_1 1.94, pK_2 8.65 (a-NH), pK_3 10.76; occurrence: urea cycle, shark liver, application: Atosiban (tocolytic) [36]. Hydrochloride: [*3184-13-2*], $C_5H_{13}ClN_2O_2$, M_r 168.6, *mp* 215 °C (decomp.), $[\alpha]_D^{25}$ + 28.3° (calculated as ornithine) (c = 2 in 5 M HCl), solubility 54.36 (20 °C) g/100 g H_2O.

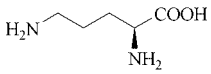

D-Penicillamine [*52-67-5*], 2-amino-3-mercapto-3-methylbutyric acid, $C_5H_{11}NO_2S$, M_r 149.21, occurrence: hydrolysis product of penicillin.

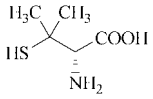

D-Phenylalanine [*673-06-3*], 2-amino-3-phenylpropionic acid, α-aminobenzenepropanoic acid, D-Phe, $C_9H_{11}NO_2$, M_r 165.19, application: Nateglinide (antidiabetic) [38].

D,L-Phenylalanine [*150-30-1*], 2-amino-3-phenylpropionic acid, α-aminobenzenepropanoic acid, $C_9H_{11}NO_2$, M_r 165.19, *mp* 284–288 °C, 320 °C (decomp.), solubility 1.29 (25 °C) g/100 g H_2O, p*I* 5.91, dissociation constants: pK_1 2.58, pK_2 9.24

D-*p*-Cl-Phenylalanine [*14091-08-8*], 2-amino-3-(4-chlorophenyl)propionic acid, D-Phe(Cl), $C_9ClH_{10}NO_2$, M_r 199.63, application: LHRH antagonists, e.g., Abarelix (antineoplastic) [39].

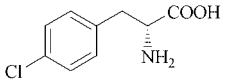

L-*p*-NO_2-Phenylalanine [*949-99-5*], 2-amino-3-(4-nitrophenyl)propionic acid, L-Phe(NO_2), $C_9H_{10}N_2O_4$, M_r 210.17, application: Zolmitriptan (antimigraine) [40].

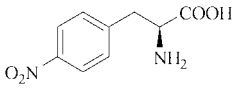

***S*-Phenyl-L-cysteine** [*34317-61-8*], 2-amino-3-phenylthiopropionic acid, L-Cys(Ph), $C_9H_{11}NO_2S$, M_r 197.61, application: Nelfinavir (antiviral) [41].

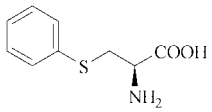

D-Phenylglycine [*875-74-1*], 2-amino-phenyl-acetic acid, D-Phg, $C_8H_9NO_2$, M_r 151.16, application: Ampicillin (antibiotic) [42].

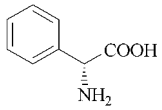

L-Pipecolic acid [*3105-95-1*], piperidine-2-carboxylic acid, L-Pec, $C_6H_{11}NO_2$, M_r 129.16, occurrence: legumes, metabolite of Lys, application: Ropivacaine (local anesthetic) [43].

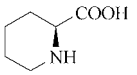

L-Piperazinecarboxylic acid [*147650-70-2*], piperazine-2-carboxylic acid, $C_5H_{10}N_2O_2$, M_r 130.15, application: Indinavir (antiviral) [44].

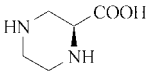

D-Piperidine-3-carboxylic acid [*25137-00-2*], nipecotic acid, $C_6H_{11}NO_2$, M_r 129.7, application: Tiagabine (anticonvulsant) [45].

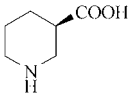

D,l-Proline [*609-36-9*], 2-pyrrolidinecarboxylic acid, 2-carboxypyrrolidine, $C_5H_9NO_2$, M_r 115.13, *mp* 205 °C (decomp.), p*I* 6.3, dissociation constants: pK_1 2.0, pK_2 10.6.

D-Proline [*344-25-2*], 2-pyrrolidinecarboxylic acid, 2-carboxypyrrolidine, D-Pro, $C_5H_9NO_2$, M_r 115.13, application: Eletriptan (antimigraine) [46].

D-3-(3′-Pyridyl)-alanine [*70702-47-5*], 2-amino-3-pyridin-3-ylpropionic acid, D-Pal, $C_8H_{10}N_2O_2$, M_r 166.18, application: LHRH antagonist, e.g. Abarelix (antineoplastic) [47].

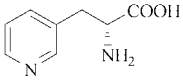

L-Pyrrolysine [*448235-52-7*], N^6-[(3*R*)-1,5-didehydro-3-methyl-D-prolyl]- L-lysine, $C_{12}H_{21}N_3O_3$, M_r 255.31, the 22nd natural amino acid to be discovered, occurrence: archaea as part of methane-producing enzymes [37].

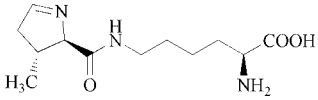

L-Saccharopine [*997-68-2*], 2-(5′-amino-5′-carboxypentylamino)pentanedioic acid, $C_{11}H_{20}N_2O_6$, M_r 276.27, occurrence: baker′s and brewer′s yeast.

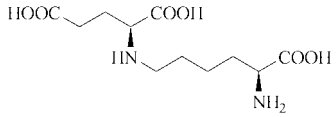

L-Selenocysteine [*10236-58-5*], 3-selenyl-L-alanine, Sec, U, $C_3H_7NO_2Se$, M_r 168.05, the 21st natural amino acid to be discovered [48]. Often found in the form of its dimer, *R*,*R*-selenocystine [*29621-88-3*], $C_6H_{12}N_2O_4Se_2$, M_r 334.09.

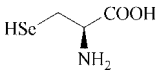

L-Selenomethionine [*3211-76-5*], (*S*)-2-amino-4-(methylseleno)butanoic acid, (*S*)-2-amino-4-(methylseleno)butanoic acid, $C_5H_{11}NO_2Se$, M_r 196.11, occurrence: plant tissue, application: dietary supplement [49].

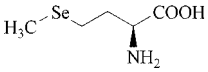

D,l-Serine [*302-84-1*], 2-amino-3-hydroxypropionic acid, 2-amino-3-hydroxypropanoic acid, $C_3H_7NO_3$, M_r 105.09, *mp* 246 °C (decomp., sealed tube), solubility 5.023 (25 °C) g/100 g H_2O, p*I* 5.68, dissociation constants: pK_1 2.21, pK_2 9.15.

D-Serine [*312-84-5*], 2-amino-3-hydroxypropionic acid, 2-amino-3-hydroxypropanoic acid, D-Ser, $C_3H_7NO_3$, M_r 105.09, application: D-cycloserine (cognition enhancer) [50].

L-Thiazolidine-4-carboxylic acid [*34292-47-7*], L-Tia, $C_4H_7NO_2S$, M_r 133.16, application: Kynostatin-272 (antiviral) [34].

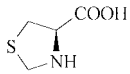

D,L-Threonine [*80-68-2*], 2-amino-3-hydroxybutyric acid, 2-amino-3-hydroxybutanoic acid, $C_4H_9NO_3$, M_r 119.12, *mp* 234-5 °C (decomp.), solubility 20.5 (25 °C) g/100 g H_2O, p*I* 6.16, dissociation constants: pK_1 2.71, pK_2 9.62.

L-Thyroxine [*51-48-9*], 2-amino-3-[4-(4-hydroxy-3,5-diiodophenoxy)-3,5-diiodophenyl]propionic acid, $C_{15}H_{11}NO_4$, M_r 269.18, occurrence: thyroid gland.

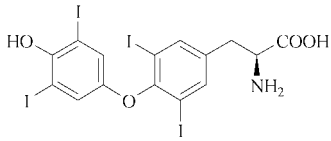

D,L-Tryptophan [*54-12-6*], 2-amino-3-(3′-indolyl)propionic acid, α-amino-1*H*-indole-3-propanoic acid, $C_{11}H_{12}N_2O_2$, M_r 204.23, *mp* 285 °C (decomp.), solubility 0.25 (30 °C) g/100 g H_2O, p*I* 5.88, dissociation constants: pK_1 2.38, pK_2 9.39.

D,L-Tyrosine [*556-03-6*], 2-amino-3-(4-hydroxyphenyl)propionic acid, [3-(4-hydroxyphenyl)]alanine, 2-amino-3-(*p*-hydroxyphenyl)propionic acid, α-amino-4-hydroxybenzenepropanoic acid, $C_9H_{11}NO_3$, M_r 181.19, *mp* 340 °C, 318 °C (decomp.), solubility 0.351 (25 °C) g/100 g H_2O, p*I* 5.63, dissociation constants: pK_1 2.2, pK_2 9.11, pK_3 10.07 (–OH).

D,L-Valine [*516-06-3*], 2-amino-3-methylbutyric acid, 2-aminoisovaleric acid, $C_5H_{11}NO_2$, M_r 117.15, *mp* 298 °C (decomp., sealed tube), solubility 7.09 (25 °C) g/100 g H_2O, p*I* 6.0, dissociation constants: pK_1 2.29, pK_2 9.72.

D-Valine [*640-68-6*], 2-amino-3-methylbutyric acid, D-Val, $C_5H_{10}NO_2$, M_r 116.14, application: Fluvalinate (insecticide) [51].

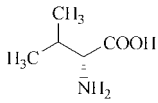

3. Industrial Production of Amino Acids

Amino acids can be classified as natural, including the proteinogenic (occurring in proteins) amino acids, or as nonnatural. Originally, only

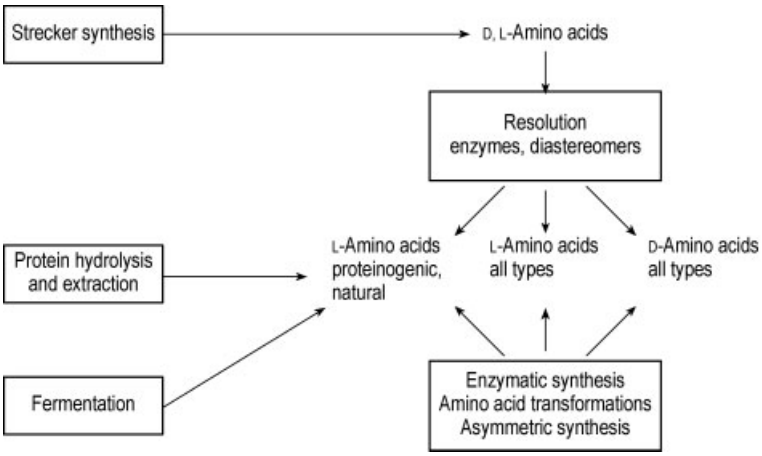

Figure 2. Routes for production of amino acids

20 amino acids were believed to be genetically coded, but the recent discoveries of selenocysteine and pyrrolysine have demonstrated that additional amino acids could be coded by subverting a stop codon. The following sections will focus on production methods for the principal proteinogenic amino acids, which are of major importance in nutrition and as feed additives.

3.1. General Methods

Four basic processes (see Fig. 2) are suitable for the production of amino acids: chemical synthesis (including asymmetric synthesis), extraction, fermentation, and enzymatic routes. The *classical chemical synthesis* is applied to produce either the achiral amino acid glycine or racemic amino acids, for instance D,L-methionine. If L-amino acids are to be manufactured by this route, the chemical synthesis has to be followed by a resolution step. In some cases L-amino acids are directly produced from prochiral precursors by means of enzymes as chiral catalysts. For L-cysteine, transformation of another amino acid (cystine) is an important method. Catalytic asymmetric syntheses have been developed for many L-amino acids and their D-antipodes, but few have been applied on an industrial scale. The *extraction process* has the advantage of offering access to nearly all the proteinogenic L-amino acids by isolation from protein hydrolysates. The starting materials are protein-rich products, such as keratin, feathers, blood meal, or technical gelatin (see Fig. 3). This process is now generally being replaced by other methods, because of customer preference for materials not of animal origin. Some amino acids, such as L-tyrosine, can be economically isolated from plant residues, for example sugar beet molasses.

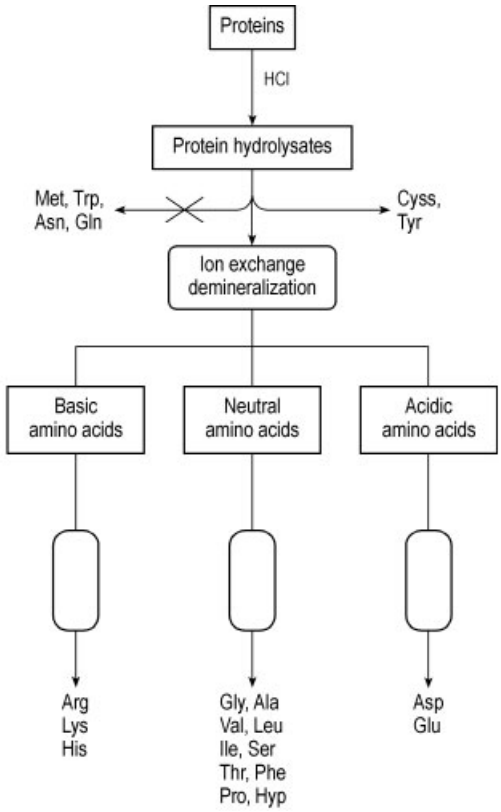

Figure 3. Extraction of amino acids

Table 2. Production of amino acids, methods, and production volume

Type	Amino acid	Preferred production method
I	L-Alanine	enzymatic catalysis, fermentation
I	L-Asparagine	extraction, chemical synthesis
I	L-Glutamine	fermentation, extraction
I	L-Histidine	fermentation, extraction
I	L-Hydroxyproline	fermentation, extraction
I	L-Isoleucine	fermentation, extraction
I	L-Leucine	fermentation, extraction
I	L-Methionine	enzymatic resolution
I	L-Proline	fermentation, extraction
I	L-Serine	fermentation, extraction
I	L-Tyrosine	extraction
I	L-Valine	enzymatic catalysis, fermentation
II	L-Arginine	fermentation, extraction
II	L-Cysteine	fermentation, enzymatic catalysis (via reduction of L-cystine)
II	L-Tryptophan	fermentation
III	Glycine	chemical synthesis
III	L-Aspartic acid	enzymatic catalysis
III	L-Phenylalanine	fermentation
III	L-Threonine	fermentation
IV	D,L-Methionine	chemical synthesis and enzymatic resolution
IV	L-Glutamic acid	fermentation
IV	L-Lysine	fermentation
Type I	100–1000	t/a
Type II	1000–8000	t/a
Type III	8000–100 000	t/a
Type IV	100 000–800 000	t/a

The use of overproducing microbial strains (→ Biotechnology; → Enzymes, Chap. 6.5) in *fermentation processes* with sucrose or glucose as carbon source is currently the most economic production method for bulk amino acids such as monosodium L-glutamate, L-lysine hydrochloride, and L-threonine. In the last ten years, this has become the method of choice for production of other amino acids, such as L-phenylalanine, L-tryptophan, and L-cysteine. Classical breeding and mutagenesis as well as recombinant DNA techniques are generally employed to modify the biosynthetic pathway and force the bacteria to overproduce the corresponding metabolites. Industrially used mutants are usually characterized by genetic markers introduced by methods of deregulation, e.g., auxotrophic mutants, in order to release key enzymes from strict regulation by metabolites (feedback inhibition and repression). A different way to obtain overproducing strains consists of screening of regulatory mutants, in that important enzymes are resistant to toxic compounds, i.e., analogues or derivatives of amino acids. In such cases high amino acid concentrations do not inhibit the enzymatic key in the pathway of biosynthesis and thus result in high production rates of the desired amino acid. The fourth method, *enzymatic catalysis*, uses whole cells or active cell components (enzymes) as such or, if possible, as immobilized biocatalysts in continuously operated reactors. The competitiveness of enzymatic processes depends on the availability and price of substrate, the activity and stability of the involved enzyme(s) and on the simplicity of product recovery. For some general reviews on amino acid synthesis, emphasizing the chemical and biocatalytic methods, see [52–56].

In Table 2 the amino acids and their preferred production methods are listed, classified by the size of production volume. Table 3 specifies the amino acids produced by direct fermentation of carbohydrates.

3.2. Production of Specific Amino Acids

3.2.1. L-Alanine

L-Alanine is industrially produced from L-aspartic acid by means of immobilized *Pseudomonas dacunhae* cells in a pressurized bioreactor [69]. In direct fermentation microorganisms usually accumulate D,L-alanine because of alanine racemase also present. With a D-cycloserine resistant mutant selected from *Brevibacterium lactofermentum*, it is possible to obtain 46 g/L D-alanine with an enantiomeric excess (e.e.) of 95 % [58]. An alanine racemase-deficient mutant of *Arthrobacter oxydans* was reported, that produces 75 g/L L-alanine from glucose with a yield of 52 % and 95 % e.e. [57]. A small amount of L-alanine is still isolated from protein hydrolysates.

3.2.2. l-Arginine

L-Arginine is today produced mainly by fermentation. Suitable strains for fermentation are deregulated mutants derived from *Corynebacterium glutamicum* and *Bacillus subtilis* accumulating 25–35 g/L L-arginine from glucose [70]. Very potent strains of *Serratia marcescens*

Table 3. Amino acids produced by direct fermentation from carbohydrates

Amino acid	Type of mutant	Amino acid yield, g/L	Reference
L-Alanine (95 % e.e.)	*Athrobacter oxydans*	75	Hashimoto and Katsumata (1994) [57]
D-Alanine (95 % e.e.)	*Brevibacterium lactofermentum*	46	Yahata et al. (1993) [58]
L-Arginine	*Serratia marcescens* AU^r-1	100	Chibata et al. (1983) [59]
L-Glutamine	*Brevibacterium flavum* AJ 3409	57	Yoshihara et al. (1992) [60]
L-Histidine	*Serratia marcescens*	40	Sugiura et al. (1987) [61]
L-Isoleucine	*Escherichia coli* H-8461	30	Kino et al. (1993) [62]
L-Leucine	*Brevibacterium flavum* AJ 3686	19.5	Yoshihara et al. (1992) [63]
L-Lysine	*Corynebacterium glutamicum*	>120	Oh et al. (1990) [64]
L-Methionine	*Pseudomonas putida* VKPM V-4167	3.5	University Odessa (1992) [65]
L-Phenylalanine	*Escherichia coli* MWPWJ304/pMW16	51	Ikeda et al. (2003) [56]
L-Proline	*Brevibacterium flavum* AP113	97.5	Kocarian et al. (1986) [66]
L-Serine	*Methylobacterium* sp. MN43	65	Ikeda et al. (2003) [56]
L-Threonine	*Escherichia coli* BKIIM B-3996	85	Debabov et al. (1990) [67]
L-Tryptophan	*Corynebacterium glutamicum* KY9218/pKW9901	50	Ikeda et al. (1994) [68]
L-Valine	*Corynebacterium glutamicum VR 3*	99	Ikeda et al. (2003) [56]

derived from mutants obtained by transduction, having feedback-insensitive and derepressive enzymes of arginine biosynthesis and 6-azauracil resistance, are able to produce 60–100 g/L L-arginine [59]. A small amount of L-arginine is isolated from protein hydrolysates.

3.2.3. l-Aspartic Acid and Asparagine

L-Aspartic acid is industrially manufactured by an enzymatic process in which aspartase (L-aspartate ammonia lyase, EC 4.3.1.1) catalyzes the addition of ammonia to fumaric acid [71]. Advantages of the enzymatic production method are higher product concentration and productivity and the formation of fewer byproducts. Thus L-aspartic acid can be easily separated from the reaction mixture by crystallization.

A process involving continuous production of L-aspartic acid by means of carrier-fixed aspartase isolated from *Escherichia coli* was first commercialized in Japan [72]. An economically attractive process for L-aspartic acid uses resting or dried cells with a high aspartase content, e.g., a 4.5 % suspension of aspartase-containing *Brevibacterium flavum* cells could be recycled in seven repeated batches and concentrations up to 166 g/L L-aspartate could be achieved [73]. In 1973, an immobilized cell system based on *E. coli* cells entrapped in polyacrylamide gel lattice was introduced for large-scale production [74]. This example represents the first industrial application of immobilized microbial cells in a fixed-bed reactor. Further improvements, immobilization of the cells in κ-carrageenan, provided remarkably increased operational stability, resulting in biocatalyst half-lives of almost two years [75]. A column packed with the κ-carrageenan-immobilized system allows a theoretical productivity of 140 $g\,L^{-1}\,h^{-1}$ L-aspartate.

As L-aspartic acid gained importance as intermediate for the manufacture of the dipeptide sweetener aspartame (methyl ester of L-aspartyl-L-phenylalanine) improved processes were developed. For continuous production of L-aspartic acid *Escherichia coli* strains immobilized with polyurethane [76] or with polyethylenimine and glass fiber support [77] and κ-carrageenan-immobilized *Pseudomonas putida* [78] have been reported and protected, respectively. Less successful was the search for strains that are overproducing L-aspartate by fermentation of sugar. A pyruvate kinase-deficient mutant of *Brevibacterium flavum* accumulates up to 22.6 g/L L-aspartic acid in glucose-containing medium [79], not enough to be competitive with the enzymatic processes. On the other hand, genetic recombination techniques have helped to improve aspartase-containing strains. An *asp*A gene bearing plasmid (pBR322:*asp*A-*par*) was able to elevate aspartase formation in *Escherichia coli* K 12 about 30-fold [80].

L-Asparagine can be isolated as a byproduct from the production of potato starch. A simple synthesis of L-asparagine starts from L-aspartic acid which is esterified to the β-methyl ester followed by treatment with ammonia [81].

3.2.4. L-Cystine and L-Cysteine

L-Cysteine used to be produced almost exclusively by hydrolysis of hair or other keratins. The amino acid isolated was L-cystine, which was reduced electrolytically to L-cysteine. L-Cysteine has also been prepared from β-chloro-D,L-alanine and sodium sulfide with cysteine desulfhydrase, an enzyme obtained from, e.g., *Citrobacterium freundii* [82].

Today, however, the main processes for cysteine production are biological. A direct fermentation process has been developed for the manufacture of L-cystine, using a modified E. coli bacterium [83]. The technology has been extended to prepare other modified L-cysteine analogues [84]. An enzymatic process for L-cysteine has been successfully developed using microorganisms capable to hydrolyze 2-amino-Δ^2-thiazoline 4-carboxylic acid (ATC) which is readily available from methyl α-chloroacrylate and thiourea. A mutant of *Pseudomonas thiazolinophilum* converts D,L-ATC to L-cysteine in 95 % molar yield at product concentrations higher than 30 g/L [85].

3.2.5. L-Glutamic Acid

In 1957, a soil bacterium was discovered [86] which was able to excrete considerable amounts of L-glutamate. Fermentation processes using strains of this bacterium, later called *Corynebacterium glutamicum*, have been successfully commercialized not only for L-glutamic acid but also for the production of other economically important amino acids [87, 88]. Numerous coryneform microorganisms have been isolated and found to be able to overproduce L-glutamic acid and other amino acids. Examples of these microorganisms are *Brevibacterium flavum*, *Brevibacterium lactofermentum*, and *Microbacterium ammoniaphilum*. Because of minor differences in the character of those bacteria [89] which are all gram-positive, nonspore-forming, nonmotile and all require biotin for growth, the name of genus *Corynebacterium* was suggested for these coryneform bacteria [90].

For industrial production of L-glutamic acid, molasses (sucrose), starch hydrolysates (glucose) and ammonium sulfate are generally used as carbon and nitrogen sources, respectively [91]. Key factors in controlling the fermentation are the presence of biotin in optimal concentration — to optimize cell growth and the excretion of L-glutamate — and sufficient supply of oxygen to reduce the accumulation of byproducts, such as lactic and succinic acid. In biotin-rich fermentation media the addition of penicillin or cephalosporin C favors the overproduction of L-glutamic acid due to effects on the cell membrane. The supplementation of fatty acids also results in an increased permeability of the cells thus enhancing glutamate excretion. In the past the mechanism of glutamate excretion was simply explained as a "leakage" or "overflow" phenomenon [92]. In the beginning of the 1990s it was reported that a specific carrier system exists [93, 94], which is responsible for active glutamate transport in *Corynebacterium glutamicum.* Generally the intracellular accumulation of glutamate does not reach levels sufficient for feedback control in glutamate overproducers due to rapid excretion of glutamate. However, the regulatory mechanisms of L-glutamic acid biosynthesis have been studied intensively to obtain mutants with increased productivity. Two enzymes have been shown to play key roles in the biosynthesis of L-glutamic acid [95].

1) Phosphoenolpyruvate carboxylase (PEPC) catalyzes carboxylation of phosphoenolpyruvate to yield oxaloacetate; it is inhibited by L-aspartic acid and repressed by both L-aspartic and L-glutamic acids.
2) α-Ketoglutarate dehydrogenase (KDH) converts α-ketoglutarate to succinyl-CoA. In L-glutamate overproducing strains KDH limits further oxidation of α-ketoglutarate to carbon dioxide and succinate, thus favoring formation of L-glutamic acid.

In L-glutamate overproducing strains the K_m value of KDH for α-ketoglutarate was nearly two magnitudes lower than that of L-glutamic acid dehydrogenase (GDH) which catalyzes the last step, the reductive amination of α-ketoglutarate to L-glutamate. Consequently, V_{max} of GDH was proven to be about 150 times higher than that of KDH. The *Corynebacterium glutamicum gdh* gene has been isolated and characterized [96]. A strain of *Microbacterium ammoniaphilum* cultured under biotin-deficient conditions produced 58 % of L-glutamic acid

formed from glucose via phosphoenolpyruvate, citrate, and of α-ketoglutarate and the other 42 % via the tricarboxylic acid (TCA) or the glyoxylate cycle [97]. In large-scale production the formation of trehalose very often reduces the product yield. Trehalose consists of two α-1,1 bound glucose molecules and is excreted by the bacteria as a material with protects the cell against high osmotic pressure (osmoprotectant). A process was recently developed and successfully industrialized in which trehalose formation is controlled and decreased by culturing the overproducing mutant in media containing invert sugar from molasses [98, 99].

Today wild type isolates as well as mutants developed by classical breeding or even strains constructed by modern techniques, using cell fusion or recombinant DNA methods [100], are available for industrial L-glutamate production. In addition a new type of fermentation process was reported which uses a strain that overproduces L-glutamic acid and L-lysine simultaneously. The cultivation of an auxotrophic regulatory mutant of *Brevibacterium lactofermentum* in a medium, supplemented with polyoxyethylenesorbitan monopalmitate as surface-active agent, resulted in the accumulation of 162 g L-amino acids per liter (105 g L-lysine · HCl + 57 g L-glutamic acid). This corresponds to a production rate of 3.8 g (L-lysine · HCl + L-glutamic acid) per liter per hour [101].

The success of an economic production basically depends on the experience in fermentation technology, in up- and downstream processing and on the skill of employees in research, development, and production. A multistep inoculation procedure followed by the fed-batch mode (→ Biotechnology, Chap. 6.2) in the main fermentation up to 500 m^3 scale is still the preferred technology for the production of L-glutamic acid.

Critical operations are the steam-forced sterilization of the fermenter and batchwise or continuous sterilization of the culture medium to prevent contamination by foreign microbes. During fermentation the temperature, pH, dissolved oxygen, and the sugar consumption have to be controlled as important process parameters. When the fermentation is completed after 40–60 h, the production strain may have accumulated more than 150 g/L L-glutamic acid. After deactivation of the broth, the recovery of the product is started by biomass separation using centrifuges or ultrafiltration units, concentration of the centrifugate or filtrate, followed by crystallization, filtration and drying. A scheme of production steps is given in Figure 4.

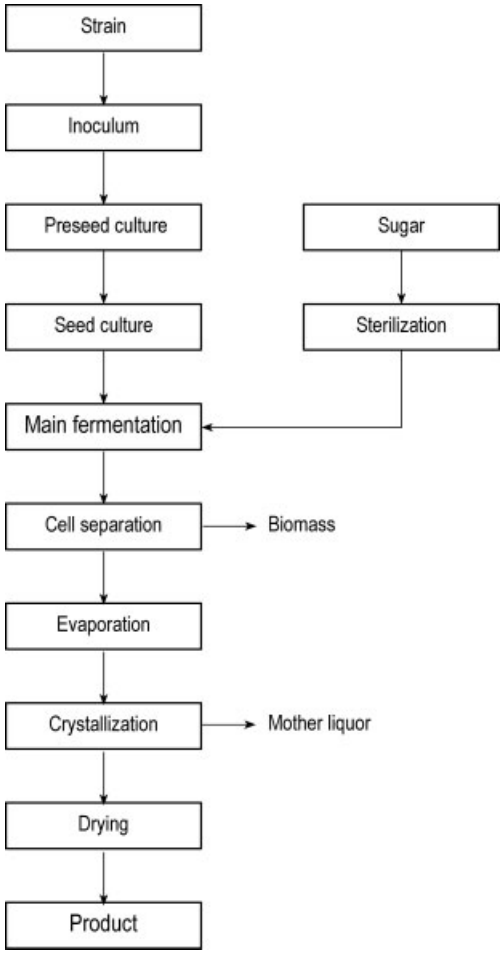

Figure 4. Flow diagram of fermentation and downstream processing of L-glutamic acid

3.2.6. L-Glutamine

Microbial L-glutamine producers were selected from wild type glutamate-producing coryneform bacteria. A sulfaguanidine resistant mutant of *Brevibacterium flavum* accumulates 41 g/L L-glutamine in 48 h from 10 % glucose [102]. A yield of 44 % was achieved by the mutant *Brevibacterium flavum* AJ3409 [60].

3.2.7. L-Histidine

Efficient L-histidine fermentation can be performed with strains of *Corynebacterium glutamicum* and *Serratia marcescens*. A mutant of *Corynebacterium glutamicum* having resistance to 8-azaguanine, 1,2,4-triazole-3-alanine, 6-mercaptoguanine, 6-methylpurine, 5-methyltryptophan and 2-thiouracil, produces 15 g/L of L-histidine. Only after introduction of these resistance markers by mutagenesis is the mutant capable of releasing this amount of L-histidine into the nutrient solution. A strain of *Serratia marcescens* having both characters, feedback-insensitive and derepressed histidine enzymes combined with transductional techniques and 6-methylpurine resistance, accumulates 23 g/L L-histidine [103]. By amplification of the genes *his*G, *his*D, *his*B, *his*C in *Serratia marcescens* the final concentration of L-histidine could be elevated from 28 g/L to 40 g/L [61].

L-Histidine can also be produced by isolation from blood-meal hydrolysates [104].

3.2.8. L-Hydroxyproline

L-Hydroxyproline (*trans*-4-hydroxy-L-proline, 2*S*,4*R*-hydroxyproline) is today mainly obtained by enzymatic hydroxylation of L-proline with genetically modified *E. coli* [105, 106]. The proline 4-hydroxylase specifically produces the 4-*trans* isomer. This method has almost completely replaced the isolation from hydrolysates of gelatin or other collagens. An unusual feature of this amino acid is that it is not incorporated into collagen during biosynthesis at the ribosome, but is formed from L-proline by a post-translational hydroxylation reaction.

3.2.9. L-Isoleucine

An advantageous fermentation method for production of L-isoleucine is the use of chemically synthesized substrates that only require a few steps to be converted to L-isoleucine. Among these the natural precursors 2-ketobutyrate [107] or D,L-2-hydroxybutyrate [108] have been used for the production with *Corynebacterium glutamicum*. A leucine requiring mutant of *Corynebacterium glutamicum*, with increased D-lactate utilization and consuming D,L-2-hydroxybutyrate, accumulates 13.4 g/L L-isoleucine. However, exploitation of this process is hampered by formation of by-products [109]. Sugar based L-isoleucine processes have been developed with strains of *Corynebacterium glutamicum* [110], *Serratia marcescens* [111], and *Escherichia coli* [112–114]. The mutant *Escherichia coli* H-8285, being resistant to thiaisoleucine, arginine hydroxamate, and D,L-ethionine accumulates 26 g/L L-isoleucine in 45 h in a fed-batch process [115]. Introduction of resistance to 6-dimethylaminopurine in strain H-8285 resulted in a mutant *Escherichia coli* H-8461 that increased L-isoleucine accumulation to 30.2 g/L [62]. The biosynthesis of L-isoleucine has been investigated in detail on the level of involved genes [63], thus recombinant strains are being constructed with high productivity and selectivity. An appropriate balance of homoserine dehydrogenase and threonine dehydratase activities in the construct *Corynebacterium glutamicum* DR17/pECM3::*ilv*A(V323A) with feedback-resistant aspartate kinase create a specific productivity of 0.052 g L-isoleucine per gram dry biomass per hour [116]. The recombinant strain *Escherichia coli* AJ13100 produces L-isoleucine from glucose with high selectivity in 30 % yield [117].

3.2.10. L-Leucine

L-Leucine can be manufactured by fermentation, precursor fermentation, or, less commonly today, by isolation from protein hydrolysates. Using a fed-batch culture of *Corynebacterium glutamicum* ATCC 13032 32 g/L 2-ketoisocaproate can be converted to 24 g/L L-leucine by the transaminase B reaction [118]. With the method of direct fermentation L-leucine can be produced by either α-aminobutyric acid-resistant mutants of *Serratia marcescens* or by 2-thiazole-alanine-resistant coryneform strains [119]. *Brevibacterium flavum* AJ3686 accumulates 19.5 g/L L-leucine (15 % yield) [120].

3.2.11. L-Lysine

The most potent microorganisms to overproduce L-lysine are mutants derived from *Corynebacterium glutamicum,* a gram-positive bacterium first introduced as an L-glutamate producing microbe; the wild strains themselves are not able to excrete L-lysine. Mainly auxotrophic and regulatory mutants of this bacterium have been developed by classical breeding methods and mutagenesis. The following techniques have been applied to channel the metabolic pathway in the biosynthesis of amino acids:

Screening for auxotrophic mutants, in order to release key enzymes, for instance aspartate kinase, from strict regulation by metabolites (feedback inhibition).

Screening for regulatory mutants, in which aspartate kinase is resistant to toxic analogues of L-lysine, such as *S*-(2-aminoethyl)-L-cysteine (AEC) or *O*-(2-aminoethyl)-L-serine. In such strains high lysine concentrations do not inhibit the enzymatic key step, the formation of 4-aspartylphosphate from L-aspartase catalyzed by aspartate kinase.

Screening for mutants having amino acid auxotrophy combined with deregulation.

Screening for regulatory mutants having additional enzyme defects (reduced pyruvate kinase) and low levels of other enzymes (citrate synthase).

As a modern technique, cell fusion with the method of protoplast fusion has been successively applied for breeding of industrial microorganisms [121, 122]. This technique allows the combination of positive characteristics of different strains such as high selectivity and high productivity. In fermentation with media of inhibitory osmotic stress (i.e., high substrate concentration results in high osmotic pressure on the cell that inhibits the performance of the mutant) the sugar consumption rate and L-lysine production rate of some mutants can be stimulated by the addition of glycine [123]. Another attractive approach is the development of thermophilic strains. A mutant of *Corynebacterium thermoaminogenes* was patented which is capable of growing at a temperature $> 40\,^{\circ}C$ and accumulating L-lysine in the culture medium [124].

Meanwhile, regulation of lysine excretion in overproducing strains is known in detail [125, 126]. Thus more attention should be paid to process development in the fields of molecular biology, biochemistry, and physiology and to finding new approaches for developing improved overproducing mutants [127]. The most specific and well-directed methods for strain development are offered by recombinant DNA techniques.

In principle all genes encoding the relevant enzymes in L-lysine biosynthesis have been isolated, characterized, and amplified in coryneform bacteria to enhance L-lysine formation [128]. As an example, amplification of *dap*A gene that codes for dihydrodipicolinate synthase in *Corynebacterium glutamicum* resulted in 35 % higher overproduction of L-lysine compared to the parent strain [129]. Another option for strain improvement is the transformation of *dap*A gene together with a *lys*C gene, coding for aspartate kinase with decreased feed back inhibition in *Corynebacterium glutamicum* [130].

In fed-batch culture and under appropriate conditions the favorable mutants for lysine production are able to reach final concentration of about 120 g/L L-lysine, calculated as hydrochloride [64]. Fermentation processes are performed in big tanks up to 500 m^3 size. An optimized feeding strategy practiced with a computer aided process control system may enable high conversion yield and productivity in large scale fed-batch cultivation [131]. Apart from the specific fermentation know how, inoculation, sterilization, and feeding strategy the recovery process and the quality of the product both can be decisive factors to guarantee competitiveness. The conventional route of lysine downstream processing is characterized by:

Removal of the bacterial cells from fermentation broth by separation or ultrafiltration

Absorbing and then collecting lysine in an ion exchange step

Crystallizing or spray drying of lysine as L-lysine hydrochloride

An alternative process consists of biomass separation, concentration of the fermentation solution, and filtration of precipitated salts. The liquid product contains up to 50 % L-lysine base, that is stable enough to be marketed [132]. In the 1990's, a new concept for lysine production was introduced. Here the lysine containing fermentation broth is immediately evaporated, spray-

dried, and granulated to yield a feed-grade product, which contains lysine sulfate. Its lysine content correspond to that of a material which contains to at least 60 % of L-lysine hydrochloride [133, 134]. The process avoids any waste products usually present in the conventional L-lysine hydrochloride manufacture.

3.2.12. D,L-Methionine and L-Methionine

L-Methionine and its antipode D-methionine are of equal nutritive value, thus the racemate can directly be used as feed additive.

The most economic way for production of D,L-methionine is the chemical process based on acrolein, methyl mercaptan, hydrogen cyanide, and ammonium carbonate (see Fig. 5). β-Methylthiopropionaldehyde, formed by addition of methyl mercaptan to acrolein, is the intermediate that reacts with hydrogen cyanide to give α-hydroxy-γ-methylthiobutyronitrile. Treatment with ammonium carbonate leads to 5-(β-methylthioethyl)hydantoin that is saponified by potassium carbonate giving D,L-methionine in up to 95 % yield, calculated on acrolein [135].

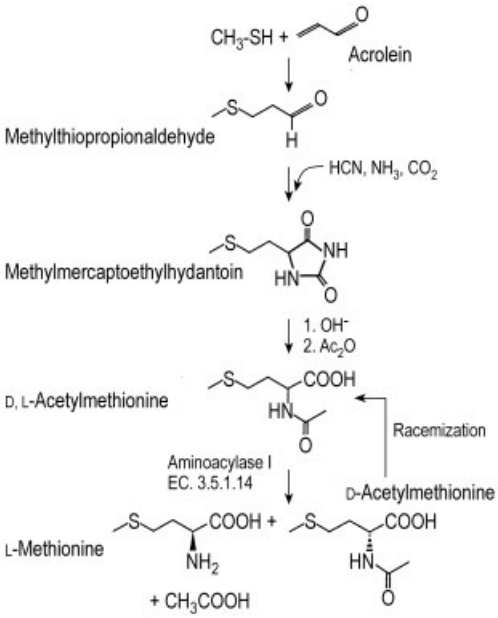

Figure 5. Degussa process for production of L-methionine

The production method of choice for L-methionine is still the enzymatic resolution of racemic *N*-acetyl-methionine using acylase from *Aspergillus oryzae*. The production is carried out in a continuously operated fixed-bed or enzyme membrane reactor [136].

Alternatively, L-methionine may be produced by microbial conversion of the corresponding 5-substituted hydantoin. With growing cells of *Pseudomonas* sp. strain NS671, D,L-5-(2-methylthioethyl)hydantoin was converted to L-methionine; a final concentration of 34 g/L and a molar yield of 93 % have been obtained [137].

Biosynthesis of L-methionine and its regulation in bacteria is well known. Although some promising concepts, for example utilization of sulfate, sulfite, or thiosulfate as sulfur sources for microbes have been suggested [138], it was not possible so far to develop strains that are able to excrete remarkable amounts of L-methionine into the culture medium.

3.2.13. L-Phenylalanine

Several chemical, biocatalytic, and fermentation methods for producing L-phenylalanine have been developed, some of which are of industrial significance (see Fig. 6).

In previous large-scale production processes for L-phenylalanine two enzymatic methods were applied:

1) Resolution of *N*-acetyl-D,L-phenylalanine by carrier-fixed microbial acylase: This process provided pharmaceutical-grade L-phenylalanine, but suffered from the disadvantage that the D-enantiomer had to be racemized and recycled.
2) Stereoselective and enantioselective addition of ammonia to *trans*-cinnamic acid, catalyzed by L-phenylalanine ammonia lyase (PAL, EC 4.3.1.5): PAL-containing *Rhodotorula rubra* was used in an industrial process [139] to supply L-phenylalanine for the first production campaign of the sweetener aspartame. When continuously operated in an immobilized whole cell reactor, the bioconversion reached concentration up to 50 g/L L-phenylalanine at a conversion of about 83 % [140]. Other processes started from phenylpyruvate with L-aspartic acid as amine donor using immobilized cells of *Escherichia*

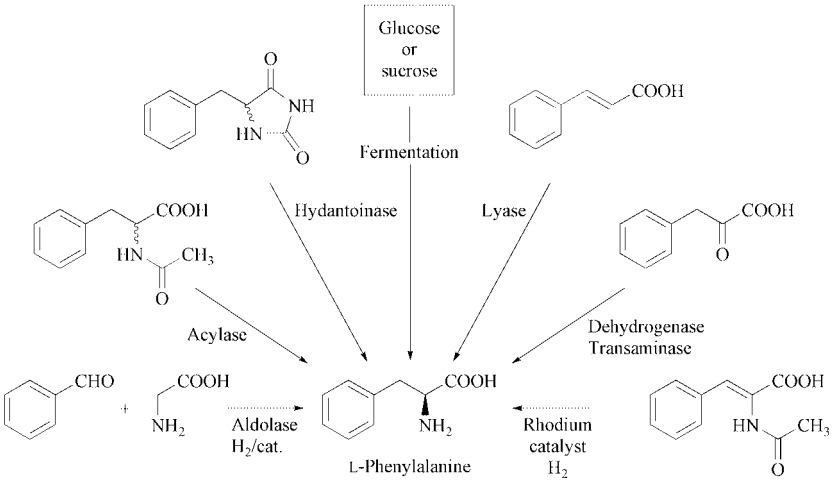

Figure 6. Asymmetric syntheses for production of L-phenylalanine

coli [141] or from α-acetamidocinnamic acid and immobilized cells of a *Corynebacterium equi* strain [142]. In both cases L-phenylalanine concentrations up to 30 g/L and more (molar yields as high as at least 98 %) were reached.

However, fermentation processes based on glucose-consuming L-phenylalanine overproducing mutants of *E. coli* and coryneform strains turned out to be more economical. The biosynthetic pathway for aromatic amino acids in bacteria is strictly regulated [143]. L-Phenylalanine is formed in ten enzymatic steps starting from erythrose-4-phosphate and phosphoenolpyruvate. The biosynthesis is governed by the first key enzyme 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (DAHPS) which is inhibited by both L-phenylalanine and L-tyrosine (by acting on the enzyme itself) and repressed by L-tyrosine (by acting on the according gene of the DNA). The other important enzyme is prephenate dehydratase (PDT) also inhibited by L-phenylalanine, but stimulated by L-tyrosine. To overcome these regulatory mechanisms either auxotrophs of *Corynebacterium glutamicum* have been constructed or L-phenylalanine analogues, e.g., 4-aminophenylalanine and 4-fluorophenylalanine, have been applied. The latter variant leads to resistant mutants of *Brevibacterium flavum* or *lactofermentum* [144]. These auxotrophic and regulatory mutants are able to produce more than 20 g/L of L-phenylalanine in a medium containing 13 % glucose. Similar results can be obtained by tyrosine auxotrophic regulatory mutants of *E. coli* [145]. With recombinant DNA techniques it was possible to improve overproducing strains of coryneform bacteria as well as of *E. coli.* Amplification of a deregulated DAHPS gene was achieved in a phenylalanine producer of *Brevibacterium lactofermentum* [146]. For optimal production of L-phenylalanine in fed-batch cultivation the critical specific glucose uptake rate has to be controlled. The specific feed rate during fermentation has to be adjusted below a critical limit, since otherwise the *E. coli* producer will be forced to excrete acetate [147]. A suitable profile of the specific glucose feed rate prevents acetate formation and leads to improved L-phenylalanine production with a final concentration up to 46 g/L and a corresponding yield of 18 %. L-Phenylalanine is recovered from the fermentation broth either by two-step crystallization or by an ion-exchange resin process. The preferred cell separation technique is ultrafiltration; and the filtrates may be treated with activated carbon for further purification. Instead of ion-exchange resins nonpolar, highly porous synthetic adsorbents are recommended to remove impurities [148, 149]. An alternative process in which a cell separator is integrated in the fermentation pa rt, thus allowing cell recycling, was suggested for L-phenylalanine production and may lead to prospective developments [150].

3.2.14. L-Proline

L-Proline is still produced to a small extent by isolation from protein hydrolysates, but today direct fermentation using analogue-resistant mutants of coryneform bacteria or *Serratia marcescens* is an economic alternative production method [151]. An isoleucine auxotrophic mutant of *Brevibacterium flavum* having resistance to sulfaguanidine and D,L-3,4-dehydroproline (DP) is able to accumulate 40 g/L L-proline. *Brevibacterium flavum* AP113 is claimed to produce 97.5 g/L L-proline; this mutant is characterized by isoleucine auxotrophy, resistance to DP, and osmotic pressure and incapable to degrade L-proline [66]. A proline oxidase-less strain of *Serratia marcescens*, having resistance to DP, thiazoline-4-carboxylate and azetidine-2-carboxylate, overproduces 58.5 g/L L-proline into the culture medium [152]. By amplification of the genes *pro*A and *pro*B in this type of regulatory mutant, a construct was obtained which yields 75 g/L L-proline [153].

3.2.15. L-Serine

L-Serine is obtained by microbial/ enzymatic conversion of glycine using immobilized resting cells or crude cell extracts. *Hyphomicrobium* strains possess the serine pathway and are able to produce L-serine from methanol and glycine. Methanol is oxidized by methanol dehydrogenase to formaldehyde which in turn is converted in an aldol-like reaction with glycine to L-serine. The reaction is catalyzed by serine hydroxymethyltransferase (SHMT) [154]. *Hyphomicrobium* sp. NCIB10099 was found to produce 45 g/L L-serine from 100 g/L glycine and 88 g/L methanol in three days [155]. In an enzyme bioreactor with a feedback control system a crude extract from *Klebsiella aerogenes* containing SHMT has been used to synthesize L-serine from glycine and formaldehyde in the presence of tetrahydrofolic acid and pyridoxal phosphate. In this bioreactor a serine concentration of 450 g/L with an 88 % molar conversion of glycine at a volumetric productivity of $8.9\,g\,L^{-1}\,h^{-1}$ could be achieved under optimized conditions [156]. With whole cells of *Escherichia coli* MT-10350 L-serine is produced by treatment of an oxygenated aqueous glycine solution (485 g/L) with aqueous formaldehyde for 35 h at 50 °C in a molar yield of 89 % based on glycine [157]. Extraction of protein hydrolysates is used to a smaller extent.

3.2.16. L-Threonine

Up to the end of the 1980s, L-threonine was mainly used for medical purposes, in amino acid infusion solutions and nutrients. It was manufactured by extraction of protein hydrolysates or by fermentation using mutants of coryneform bacteria in amounts of several hundred tons per year worldwide. The production strains were developed by classical breeding. They were auxotrophic and resistant to threonine analogues such as α-amino-β-hydroxyvalerate (AHV), and reached product concentrations up to 20 g/L. These strains possessed deregulated L-threonine pathways with feedback inhibition-insensitive aspartate kinase and homoserine dehydrogenase [158, 159]. In the 1990s strain developments, using both conventional methods and recombinant DNA techniques, have been very successful. Potent classically selected mutants suggested for industrial production are the species *Brevibacterium flavum*, *Providentia rettgeri*, *Serratia marcescens*, and *Escherichia coli.* However, in the competition between the favorable candidates, strains of *Escherichia coli* proved to be superior to other bacteria. Although the pathway of L-threonine biosynthesis in *Escherichia coli* is much more regulated than that in *Corynebacterium glutamicum*, new *Escherichia coli* strains with excellent yields and productivity in threonine formation could be constructed by genetic engineering. L-Threonine is successfully marketed as feed additive with a worldwide demand of more than 10 000 t/a. Production strains are based on *Escherichia coli* K-12 constructs harboring plasmids containing the *thr* operon that consists of the genes *thr*A, *thr*B, and *thr*C [160]. Further improvements resulted in strains capable to accumulate more than 80 g/L in about 30 h with a conversion yield of more than 40 % [161]. The strain stability could be further improved, for example by integrating the threonine operon into the chromosome [162]. The recovery of feed-grade L-threonine is rather simple. After fermentation is completed, cell mass is removed by centrifugation or ultrafiltration, the

filtrate is concentrated, depigmentated and L-threonine isolated by crystallization [163].

3.2.17. L-Tryptophan

L-Tryptophan is one of the limiting essential amino acids required in the diet of pig and poultry. A mature and growing market for L-tryptophan as feed additive is in development,based on improved microbiological processes, and production has increased to several thousand tons per year. The most attractive production processes for tryptophan are based on microorganisms used as enzyme sources or as overproducers:

Enzymatic production from various precursors
Fermentative production from precursors
Direct fermentative production from carbohydrates by auxotrophic and analogue resistant regulatory mutants

L-tryptophan is synthesized from indole, pyruvate, and ammonia by the enzyme tryptophanase [164] or from indole and L-serine/D,L-serine by tryptophan synthase [165, 166]. Although production in enzyme bioreactors is quite efficient and concentrations of L-tryptophan up to 200 g/L could be achieved by condensation of indole and L-serine [167], these process variants were not economic due to the high costs of the starting materials. The microbial conversion of biosynthetic intermediates such as indole or anthranilic acid to L-tryptophan has also been considered as alternative for production. Whereas indole consuming mutants of *Corynebacterium glutamicum* produced about 10 g/L L-tryptophan [168], strains of *Bacillus subtilis* and *Bacillus amyloliquefaciens* reached final concentrations > 40 g/L L-tryptophan with anthranilic acid as carbon source [168, 169]. The process with anthranilic acid as precursor has been commercialized in Japan. However, the manufacturer using genetically modified strains derived from *Bacillus amyloliquefaciens* IAM 1521 was forced to stop L-tryptophan production. L-Tryptophan produced by this process was stigmatized because of side products found in the product causing a new severe disease termed eosinophilia-myalgia syndrome (EMS) [170]. One of the problematic impurities, "Peak E", was identified as 1,1′-ethylidene-*bis*-(L-tryptophan), a product formed by condensation of one molecule acetaldehyde with two molecules of tryptophan [171].

In other processes, i. e., direct fermentation using overproducing mutants and carbohydrates as carbon sources, formation of such impurities does not occur. In the 1990s striking progress has been made in the development of auxotrophic and deregulated mutants of *Brevibacterium flavum*, *Corynebacterium glutamicum*, and *Bacillus subtilis*. The biosynthesis of L-tryptophan and its regulation have been reviewed in detail for the different species [172–174]. The precise knowledge about the structure of the *trp* operon in *Escherichia coli* comprising the *trp* promoter and the genes *trp*E, *trp*D, *trp*C, *trp*B, and *trp*A which are coding for the enzymes anthranilate synthase (AS), phosphoribosyl anthranilate transferase (PRT), indole glycerol phosphate synthase (IGP) and tryptophan synthase (TS), respectively, was the benefit for further strain improvements.

Thus recombinant DNA techniques have been used to increase the capability of overproduction especially in strains of *Corynebacterium glutamicum* and *Escherichia coli*. One concept was realized successfully by amplification of *trp* operon genes together with *ser*A which codes for phosphoglycerate dehydrogenase. This key enzyme in L-serine biosynthesis should provide enough L-serine in the last step of L-tryptophan formation. Production strains are able to accumulate 30–50 g/L L-tryptophan with yields higher than 20 % based on carbohydrate. Isolation of L-tryptophan is still a major issue, because the amino acid is sensitive to oxygen and heat [167]. However, at a typical production concentration of 50 g/L, more than half of the L-tryptophan product crystallizes from the medium [68, 175].

3.2.18. L-Tyrosine

L-Tyrosine was, until recently, produced exclusively from protein hydrolysates. Its low solubility in water enables a quite simple isolation of the amino acid. Enzymatic catalysis using a b-tyrosinase (tyrosine-phenol lyase) from *Erwinia herbicola* gives tyrosine by a three-component synthesis from phenol, pyruvate, and ammonia.

This process has been commercialized for the enzymatic synthesis of the drug L-DOPA, where phenol is replaced by catechol [54]. L-Tyrosine is formed in nature as part of the shikimic acid pathway, as are L-phenylalanine and L-tryptophan. Recombinant DNA technology has been used to develop strains of *Corynebacterium glutamicum* that are specific for L-tyrosine, and production yields of 20–25 g/L have been obtained [175].

3.2.19. L-Valine

L-Valine is produced industrially in pharmaceutical quality by enzymatic resolution of *N*-acetyl-D,L-valine. Using direct fermentation the branched-chained amino L-valine can be produced by either α-aminobutyric acid-resistant mutants of *Serratia marcescens* or by 2-thiazolealanine-resistant coryneform strains [176]. *Brevibacterium lactofermentum* AJ12341 produces 39 g/L L-valine (28 % yield) [177].

4. Biochemical and Physiological Significance

The biosynthesis of amino acids begins with atmospheric nitrogen, which is reduced to ammonia by bacteria and plants. Ammonia is used by plants, by bacteria, and, to a limited extent, by ruminants as a raw material for amino acids. Amino acids, in turn, serve as starting materials for the synthesis of proteins and a variety of other nitrogen-containing compounds, such as the purine and pyrimidine bases in nucleic acids. Bacterial degradation leads, once again, to ammonia and nitrogen [178].

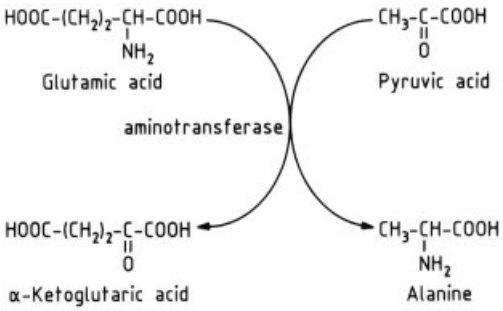

Figure 7. The transamination of amino acids

α-Ketoglutaric acid plays the central role in the assimilation of ammonia. Its transamination product, glutamic acid, in turn, can provide its amino group for the synthesis of other amino acids, e.g., alanine (Fig. 7).

Humans, animals, and some bacteria are incapable of synthesizing all the necessary amino acids in their own intermediary metabolism; i.e., they are heterotrophs and are therefore dependent on the biosynthetic capability of plants. Proteins that are consumed as foods by humans and animals are hydrolyzed to amino acids by the digestive enzymes. The amino acids are resorbed in the upper part of the intestine and enter the liver by way of the portal vein. The liver is the central organ for metabolism and homoeostasis of the plasma amino-acid level. The body's various requirements are met from the pool of free amino acids (Fig. 8), ca. 50 g in adult humans.

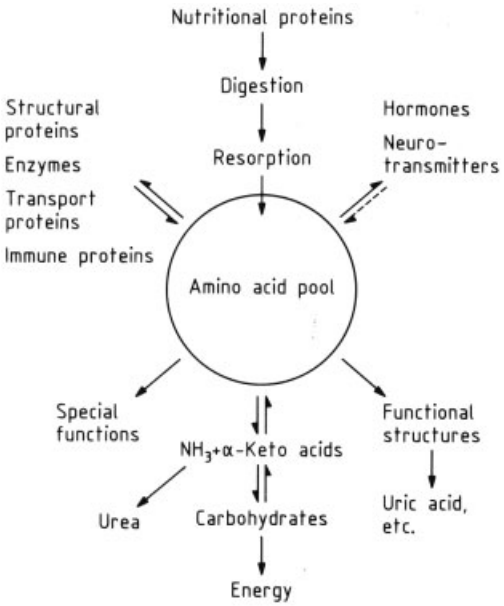

Figure 8. The amino acid pool and functions of amino acids in the intermediary metabolism

The lion's share of the amino acids ($\approx$ 300 g/d for adults) is required for synthesis of proteins [179]: structural proteins, enzymes, transport proteins, and immune proteins. Additionally, amino acids are required for the synthesis of oligopeptides and polypeptides that fulfill regulatory functions in the body, i.e., hormones.

Some amino acids or their metabolites are directly active as hormones or facilitate the transmission of nerve impulses (neurotransmitters), e.g., serotonin. Furthermore, there are amino acids that serve special functions, such as methionine, which is a methyl group donor. Finally, a series of amino acids serves as precursors for the biosynthesis of other structures. For example, glycine is used in the construction of the porphyrin skeleton. Amino acids are metabolized to produce energy in the case of a protein-deficient or a protein-excess diet.

The free amino acids in the amino acid pool undergo numerous transformations (Fig. 9), involving transamination and oxidative deamination, by which other amino acids can be synthesized. The α-keto acids are intermediates and in addition allow amino acids entrance into the carbohydrate (through pyruvate) and fatty acid (through acetylcoenzyme A) metabolisms. A distinction is therefore drawn between glucogenic and ketogenic amino acids.

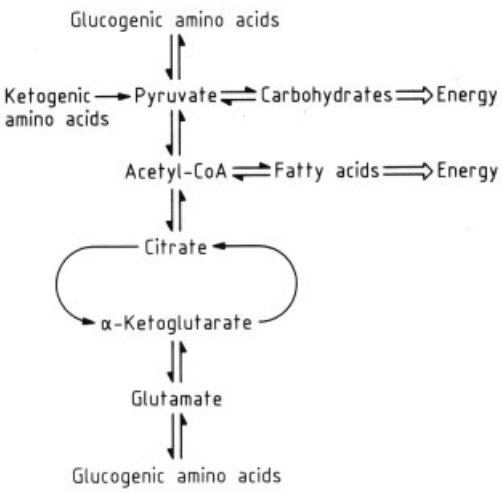

Figure 9. Intermediary metabolism of amino acids (simplified)

D-Amino acids occur in the cell pool of plants and gram-positive bacteria and as building blocks in peptide antibiotics and bacterial cell walls [10, 14]. They do not occur in human or animal metabolism; proteins are made exclusively of L-amino acids. The traces of D-amino acids detected in metabolically inert protein (teeth, eye lenses) are believed to originate from racemization. Orally ingested D-amino acids are resorbed from the intestinal lumen more slowly than the L-form. The D-enantiomers cannot be utilized or can be utilized only to a slight extent as essential amino acids [180]. The one important exception is D-methionine. Animals and adult humans convert D-methionine into L-methionine by transamination. The α-keto acid of methionine is an intermediate. Otherwise, D-amino acids are degraded with the help of D-amino acid oxidases [181] to be used as an energy source [182].

The major end products of amino acid metabolism are urea, uric acid, ammonium salts, creatinine, and allantoin. The loss of nitrogen via these metabolites stabilizes at about 22 g protein per day after a few days on a protein-free diet.

Inborn disorders in amino acid metabolism [183] can lead to marked alterations in the excretion profile. These disorders usually take the form of an enzyme or transport deficiency [178, 184]. The most common example is phenylketonuria, a disruption of the normal metabolic pathway from phenylalanine to tyrosine caused by severe limitation in the activity of the phenylalanine hydroxylase [185].

Humans and animals are not capable of producing all the required L-amino acids in their intermediary metabolism. Therefore, they are dependent on an external source of these *essential* amino acids (Table 4). In situations of increased requirements (rapid growth, stress, trauma), histidine and arginine also become essential for humans. Cysteine and tyrosine may be essential for infants during their first few weeks, because their intermediary metabolism does not yet function well enough to produce these from methionine and phenylalanine in sufficient quantities.

5. Uses

The uses of amino acids have been treated in review articles [19, 186–189].

5.1. Human Nutrition

In addition to their nutritive value, amino acids are important flavor precursors and taste enhancers. In foods for humans, the flavor uses of amino acids represent the dominant factor in

Table 4. Essential (+) and semiessential (±) amino acids

	Baby	Adult	Rat	Chicken	Hen	Cat	Salmon
L-Arginine	±	±	±	+	±	+	+
L-Cysteine	+ (?)						
Glycine				+			
L-Histidine	+	±	+	+	±	±	+
L-Isoleucine	+	+	+	+	+	+	+
L-Leucine	+	+	+	+	+	+	+
L-Lysine	+	+	+	+	+	+	+
L-Methionine	+	+	+	+	+	+	+
L-Phenylalanine	+	+	+	+	+	+	+
L-Threonine	+	+	+	+	+	+	+
L-Tryptophan	+	+	+	+	+	+	+
L-Tyrosine	+ (?)						
L-Valine	+	+	+	+	+	+	+

Table 5. Average amino acid content of some foodstuffs (mg/100 g) [a] [190]

Food	Ile	Leu	Lys	Cyss	Met	Phe	Tyr	Thr	Trp	Val	Arg	His	Protein, %	Moisture content, %
Maize, grain	350	1190	254	147	182	464	363	342	67 [M]	461	398	258	9.5	12.0
Rice, husked	300	648	299	84	183	406	275	307	98 [M]	433	650	197	7.5	13.0
Wheat, whole grain	426	871	374	332	196	589	391	382	142 [M]	577	602	299	12.2	12.0
Wheat, flour, 70–80 % extr. rate	435	840	248	304	174	581	277	321	128 [M]	493	422	248	10.9	12.0
Potato (*Solanum tuberosum*)	76	121	96	12	26	80	55	75	33 [M]	93	100	30	2.0	78.0
Bean (*Phaseolus vulgaris*)	927	1685	1593	188	234	1154	559	878	223 [M]	1016	1257	627	22.1	11.0
Soybean, milk	171	278	195	57	50	175	133	128	48 [M]	165	253	84	3.2	92.0
Soy protein, isolate [191]	4147	7119	5777	1008	1092	4644	3458	3211	1080	4210	6767	2378	75.7	4.7
Lettuce, leaves (*Lactuca sativa*)	50	83	50	24	–	67	35	54	10 [M]	71	59	21	1.3	94.8
Tomato (*Solanum lycopersicum*)	20	30	32	7	7	20	14	25	9 [M]	24	24	17	1.1	93.8
Apple (*Malus silvestris*)	13	23	22	5	3	10	6	14	3	15	10	7	0.4	84.0
Orange (*Citrus sinensis*)	23	22	43	10	12	30	17	12	6	31	52	12	0.8	87.4
Beef, veal, edible flesh	852	1435	1573	226	478	778	637	812	198 [M]	886	1118	603	17.7	61.0
Fish, fresh, all types	900	1445	1713	220	539	737	689	861	211 [M]	1150	1066	665	18.8	74.1
Milk, cows, untreated	162	328	268	28	86	185	163	153	48 [M]	199	113	92	3.5	87.3
Milk, human	48	104	81	16	19	41	39	53	20 [M]	54	46	30	1.2	87.6
Cheese, all types	956	1864	1559	76	530	950	973	725	217 [M]	1393	651	556	18.0	51.0
Egg, whole	778	1091	863	301	416	709	515	634	184 [M]	847	754	301	12.4	74.0

[a] Chemical determination.
[M] Microbiological determination.

Table 6. Essential amino acid requirements of humans

Amino acid	Suggested patterns of requirements [194], g/100 g protein			Adult requirement, mg kg^{-1} d^{-1}			
	Infant child	School-age	Adult 1973	FAO/WHO 1980 (144)	NAS/NCR	Rose[a]	Hegsted[b]
His	1.4	–	–	–	–	–	–
Ile	3.5	3.7	1.8	10	12	10	10
Leu	8.0	5.6	2.5	14	16	11	13
Lys	5.2	7.5	2.2	12	12	9	10
Met + Cys[c]	2.9	3.4	2.4	13	10	14	13
Phe + Tyr[d]	6.3	3.4	2.5	14	16	14	13
Thr	4.4	4.4	1.3	7	8	6	7
Trp	0.85	0.46	0.65	3.5	3	3	3
Val	4.7	4.1	1.8	10	14	14	11
Total	37.3	32.6	15.2	83.5	91	81	80

[a] For men.
[b] For women.
[c] Cys can partly cover the total *S*-amino acid requirement.
[d] Tyr can partly cover the total aromatic amino acid requirement.

total market value. In animal nutrition, amino acids are used almost exclusively for their nutritive value.

Addition of small amounts of amino acids to improve the nutritive value of proteins is known as supplementation. Both supplementation and the combination of proteins with complementary amino acids are used to increase the biologic value of proteins. Usually the supply of at least one of the essential amino acids lies below the requirement. This, the limiting amino acid, determines what percentage of the protein (or, more precisely, its amino acids) can be used to meet the body's amino acid requirements. In most cases, methionine is the first limiting amino acid. Sometimes it is lysine; now and then it is both together.

The contents of essential amino acids found in several animal and vegetable foodstuffs are compiled in Table 5. Considerable variations may be present in the amino acid contents of a given foodstuff.

The published requirements for the individual essential amino acids differ. The values (Table 6) usually contain safety factors and therefore are higher than the minimum requirement. Requirement values were first determined by Rose [192]; those published by Hegsted [193] are considered the most reliable at present. The amino acid requirement pattern suggested by the FAO/WHO [194] is considered optimal for the greatest part of the population.

The "average safe level of daily protein intake for men and women," based on these amino acid requirement figures, is given as 0.55 g/kg body weight. The acute daily protein requirement, however, varies between 0.5 and 2.5 g/kg with age and constitution [195]. The Deutsche Gesellschaft für Ernährung (DGE) recommends a daily consumption of 0.9 g/kg body weight [196, 197]. The Committee on Dietary Allowances, Food and Nutrition Board of the National Academy of Sciences (NAS), USA, cites 0.8 g/kg as a desirable level of daily protein consumption [198, 199].

5.1.1. Supplementation

In general, animal protein contains the essential amino acids in larger quantities and in a more favorable ratio than vegetable protein, which is often deficient in essential amino acids. Lysine is the first limiting amino acid in wheat, rye, barley, oats, maize, and millet, whereas methionine is the first limiting amino acid in meat, milk, soybeans, and other beans. The second limiting amino acids are usually threonine (wheat, rice) and tryptophan (maize, rice, casein). The limiting amino acids for several foodstuffs are listed in Table 7.

Improving the biologic value of vegetable protein in human nutrition is practiced for economic and dietary reasons. Combining different protein types is not always practical. Often a complementary protein is unavailable, too expensive, or not of acceptable taste. In these cases supplementation with amino acids is the simplest method of increasing the biologic value of

proteins. There is a monumental volume of literature on the subject of amino acid supplementation [200–207].

The biologic value is an important criterion for the evaluation of proteins or amino acid mixtures. It can be determined experimentally [208, 209]. In principle, all methods measure the ability of the nutritional protein to replace body protein. Table 8 lists the biologic values of nutritional proteins as determined by the minimum requirement [195]. Whole egg protein is the reference in this scale.

Table 7. Limiting amino acids in foodstuffs

Proteins	First limiting amino acid	Second limiting amino acid(s)
Peanut	Thr	Lys and Met
Fish	Met	Lys
Casein	Met	Trp
Torula yeast	Met	
Sesame	Lys	
Skim milk	Met	
Beans	Met	
Sunflower seed	Lys	Thr
Soy protein	Met	Lys
Wheat	Lys	Thr
Rice	Lys	Thr and Trp
Rye	Lys	Thr and Trp
Gelatin	Trp	
Maize	Lys	Trp and Thr

Another scale of evaluation is the protein efficiency ratio (PER). This is the daily weight gain of young animals, usually rats, under standard feeding conditions (see Table 8). In an improved method, the animals are fed diets of various protein levels, and the protein efficiency is then determined using regressional analysis.

The net protein retention (NPR) method and the net protein utilization (NPU) method are more accurate than the PER method because they consider the protein (NPR) or total nitrogen (NPU) requirement for maintenance. The "chemical score," in which the availability of the amino acids is not considered, is suitable for a gross estimation of the biologic protein quality. In this method the amino acid content is determined analytically and compared with the amino acid pattern of a reference protein, e.g., the FAO/WHO provisional scoring pattern (Table 9). This method provides an immediate picture of the size of amino acid gaps and the sequence of limiting amino acids.

$$\text{Score} = \frac{\text{Content of amino acid in test protein}}{\text{Content of amino acid in reference protein}} \times 100$$

Table 9. FAO/WHO provisional scoring pattern 1973 [194]

Amino acid	g/100 g protein
L-Isoleucine	4.0
L-Leucine	7.0
L-Lysine	5.5
L-Methionine + L-cystine *	3.5
L-Phenylalanine + L-tyrosine **	6.0
L-Threonine	4.0
L-Tryptophan	1.0
L-Valine	5.0
Total	36.0

* Cys can partly cover the total *S*-amino acid requirement.
** Tyr can partly cover the total aromatic amino acid requirement.

As can be seen in Table 10, addition of ca. 0.1–0.5 % of the limiting amino acid to such basic foodstuffs as wheat, rice, maize, and soy-

Table 8. Protein quality of food and minimum requirements (human)

Food	Biologic value $g\ kg^{-1}\ d^{-1}$ [195]	Minimum requirement *, ratio (rat) [195]	Protein efficiency [190]
Whole egg	100	35	3.92
Beef	92	39	2.30
Cow's milk	88	40	3.09
Potato	86 [210]	41 [210] **	3.0 [211]
Fish			3.55
Casein			2.50
Soybean	84	42	2.32
Rice	81	44	2.18
Rye flour	76	46	
Maize	72	49	1.18
Beans	72	49	1.48
Wheat flour	56	63	0.6

* Of the protein part.
** Calculated.

Table 10. Increase of the protein efficiency ratio (PER) by supplementation with amino acids [190, 203]

Food (protein content)	L-Lys·HCl, %	L-Thr, %	D,L-Trp, %	D,L-Met, %	PER *
Wheat flour (10 %)					0.65
	0.2				1.56
	0.4				1.63
	0.4	0.15			2.67
Rice (7.8 %)					1.50
	0.2	0.1			2.61
Maize (8.75 %)					1.41
	0.4		0.07		2.33
Soybean milk (10 %)					2.12
				0.3	3.01
Extruded soy protein (10%)					1.99
				0.23	2.62

* Reference protein: casein, PER=2.50.

beans raises the protein efficiency in rat growth tests impressively.

Clinical studies [212–214] and field trials have solidified the evidence of the benefits to human nutrition of amino acid supplementation [215–218]. However, the general supplementation of basic foodstuffs, such as bread and rice, is not yet practiced extensively. In dietary nutrition, however, supplementation already plays an important role (Table 11).

Some infants exhibit lactose or cow's milk protein incompatibility. The formulas marketed for this condition often are based on isolated soybean protein and are supplemented with L-methionine to increase the biologic value. The advantageous effects of L-methionine supplementation on the physical development of infants has been demonstrated in a series of clinical studies [219, 220]. Additionally, the food for pregnant and nursing women, seniors, overweight persons, and athletes can also be supplemented. Extruded soy protein, which is used in large quantities as a meat extender and vegetarian meat substitute, can be supplemented with *N*-acetyl-L-methionine [221].

5.1.2. Flavorings, Taste Enhancers, and Sweeteners

Free amino acids occur in almost all protein-based foods. In some foods their concentration is several percent. Foodstuffs having a relatively high concentration of free amino acids include fruit juices [222], cheese [223, 224], beer [225], and seafood [226]. Approximately 85 % of the free amino acids in orange juice is proline, arginine, asparagine, γ-aminobutyric acid, aspartic acid, serine, and alanine [227].

Amino acids are relatively tasteless. Nonetheless, they contribute to the flavor of foods. They have characteristic synergistic flavor-enhancing and flavor-modifying properties, and they are precursors of natural aromas [228–230]. Amino acids and protein hydrolysates are therefore useful additives in the food industry. The sodium salt of L-glutamic acid (MSG) exhibits a particularly pronounced flavor-enhancing effect, leading to the introduction of a fifth human taste concept of *umami* [231], and has been recognized as a flavoring factor for seaweed, sake, miso, and soy sauce since 1908. The substance is used in concentrations of 0.1–0.4 % as an additive for spices, soups, sauces, meat, and fish, usually in combination with nucleotides [230]. In 2000, the existence of a specific receptor for umami was confirmed [232].

L-Cysteine especially enhances the aroma of onion [233] and is therefore used to rearomatize dried onions. Glycine, which has a refreshing, sweetish flavor, occurs abundantly in mussels and prawns. It is considered to be an important flavor component of these products. When used as an additive for vinegar, pickles, and mayonnaise, it attenuates the sour taste and lends a note of sweetness to their aroma. D,L-Alanine is used for the same purpose in the Far East [234]. Glycine is used to mask the aftertaste of the sweetener saccharin [235, 236].

The L- and D-amino acids usually exhibit pronounced flavor differences. Many L-enantiomers taste weakly bitter, whereas their optical antipodes, the D-amino acids, taste sweet [237–

Table 11. Amino acids in dietetic products

Protein/protein hydrolysate	Supplemented amino acid	Use	Indication
Cow's milk, casein, whey protein	L-Cyss and/or L-Lys · HCL	infant nutrition	adapted nutrition
Soy protein	L-Met	infant nutrition	lactose incompatibility and milk protein allergy
Casein/yeast	L-Lys · HCl	meal supplement	protein malnutrition, in place of conventional nutrition

239] (Table 12). For the most part, dipeptides and oligopeptides have bitter flavors. One of the few exceptions is the methyl ester of the dipeptide L-aspartyl-L-phenylalanine (Aspartame) [240, 241], which is 150–200 times sweeter than sucrose (→ Sweeteners).

Table 12. Tastes of L- and D-amino acids* [239]

	L-Amino acid	D-Amino acid
Alanine	sweet (12–18)	sweet (12–18)
Arginine	bitter	neutral
Asparagine	neutral	sweet (3–6)
Aspartic acid	acidic/neutral	acidic/neutral
Cysteine	sulfurous	sulfurous
Glutamine	neutral	sweet (8–12)
Glutamic acid	acidic/"glutamate-like"	acidic/neutral
Glycine	sweet (25–35)	sweet (25–35)
Histidine	bitter	sweet (2–4)
Isoleucine	bitter	sweet (8–12)
Leucine	bitter	sweet (2–5)
Lysine	sweet/bitter	sweet
Methionine	sulfurous	sweet/sulfurous (4–7)
Phenylalanine	bitter	sweet (1–3)
Proline	sweet/bitter (25–40)	neutral
Serine	sweet (25–35)	sweet (30–40)
Threonine	sweet (35–45)	sweet (40–50)
Tryptophan	bitter	sweet (0.2–0.4)
Tyrosine	bitter	sweet (1–3)
Valine	bitter	sweet (10–14)

* Threshold values for sweet taste in parenthesis, µmol/mL; the threshold value for sucrose is 10–12 µmol/mL.

The free amino acids are used widely in foodstuff technology as precursors for aromas and brown food colors [242]. The flavors are formed during foodstuff production, e.g., during the ripening of cheese [243, 244], the fermentation of alcoholic beverages [245, 246], or the leavening of dough [247, 248], or foodstuff cooking, e.g., frying, roasting, boiling, by the Maillard reaction between amino acids and reducing sugars (nonenzymatic browning) [228, 249, 250]. The Strecker degradation of amino acids plays a central role in this process. A broad spectrum of aroma-intensive, volatile compounds forms [251, 252]. The most important classes are aliphatic carbonyl compounds and heterocycles, such as furans, pyrones, pyrroles, pyrrolidines, pyridines, imidazoles, pyrazines, quinoxalines, thiophenes, thiolanes, trithianes, thiazoles, and oxazoles.

It is often possible to assign certain aromas to specific amino acids [252]. For example, the sulfur-containing amino acid cysteine is primarily responsible for the formation of meat flavor. Proline seems to be important for the aroma of bread crust. Phenylalanine, as well as the branched-chain amino acids leucine and valine, is important for the characteristic flavor of chocolate. Valine and leucine are also involved in the aroma of roasted nuts. Methionine plays a key role in the aroma of French fries. The flavors of such products as precooked foods, snack articles, and spices may be improved by addition of the proper Maillard aromas. One variation is adding the precursors of the Maillard aromas, i.e., amino acid plus sugar, to the foodstuff and allowing the fragrance to form in situ. Some aroma profiles that can be prepared from amino acids are compiled in Table 13.

Table 13. Amino acids for Maillard flavors *

Meat, poultry [253]	Cys, Cyss, Gly, Glu, Ala, Met, His, Ser, Asp, Pro
Bread, cracker, biscuit [247, 254]	Pro, Lys, Arg, Val, His, Leu, Glu, Phe, Asp, Gly, Gln
Chocolate, cocoa [255]	Leu, Phe, Val, Glu, Ala
Honey [256–258]	Phe
Cream, butter [259]	Pro, Lys, Ala, Gly, His
Nut, peanut [260]	Leu, Val, Ile, Pro, Glu, Gln, His, Phe, Asp, Asn
Potato [261]	Met
Tobacco [257, 262]	Asn, Arg, GABA, Gln, Ala, Gly, Orn, Glu, Asp, Leu, Val, Thr, Pro, Tyr, Phe

* Key amino acids underlined.

5.1.3. Other Uses in Foodstuff Technology

Amino acids are used in the foodstuff industry for purposes other than supplementation and flavoring. L-Cysteine, for example, is used by the baked goods and pasta industry as a flour additive [263, 264]. As a reducing agent, it relaxes

wheat gluten proteins (by cleavage of the disulfide linkages), homogenizes the dough, accelerates dough development, and improves the structure of the baked product, while allowing shorter kneading times.

Because they are capable of forming complexes with metals, amino acids act as antioxidants for fats and fat-containing foodstuffs [265]. This effect is strengthened by primary antioxidants, such as α-tocopherol. Melanoidines, which are formed during the Maillard reaction, are stronger antioxidants than the amino acids themselves [266]. Maillard products are also reported to be preservatives [267]. Glycine apparently exhibits a special preservative effect [268].

There has been a recent upsurge in the sale of sports drinks and "energy bars", foodstuffs containing supplements with selected amino acid. These drinks were originally designed for performance athletes, but are now in widespread use, and the market is growing rapidly [269]. Amino acids used in functional beverages and foods include creatine, which is intended to help build muscle; the branched-chain amino acids L-isoleucine, L-leucine, and L-valine, which are important components of cell membrane proteins and assist synthesis of body protein and reduce central fatigue; and L-arginine, which is a regulatory agent for the circulation (production of NO). These amino acids have a higher rate of uptake in the body than protein, and improve muscle strength and recovery after exercise [270].

5.2. Animal Nutrition

The use of amino acids for the nutrition of monogastric animals is based on the same foundation as the supplementation of human foodstuffs and the clinical experience with humans. In practice, the enrichment of animal feeds and formulated feeds with amino acids, especially methionine and lysine, represents far greater quantities than does human nutrition. By supplementing feeds or formulated feeds with the first limiting amino acid an obvious cost reduction can be achieved, while maintaining the quality of the ration [271].

Of the ca. 20 amino acids found in feed protein, about one half are essential for monogastric animals (see Table 4). Most natural feeds are relatively poor in methionine and lysine (Table 14). The requirement of our livestock, however, is comparatively high [272].

When formulating a feed mix for a given animal type, the manufacturer has two choices for meeting the requirement of a particular amino acid. He may use either an excess of feed protein that contains large amounts of this amino acid or a minimum of natural protein and supplement it with synthetic amino acid. Because methionine, lysine, and threonine are commercially available and inexpensive, they are often used in formulated feed. L-Tryptophan, which in many cases is the third or fourth limiting amino acid, is becoming more popular as an ingredient for feed supplements, particularly for pigs.

Amino Acids Content of Feedstuffs. Effective supplementation requires an exact knowledge of the natural amino acid content of both the individual feedstuffs and the formulated feed mix: the desired rates of supplementation must always be capable for being measured analytically. Ion-exchange and high-pressure liquid chromatography are reliable and proven methods for this.

The amino acid contents of individual feedstuffs are published internationally in a large series of tabulations (see, e.g., Table 14). However, such data must be current and reliable.

Amino Acid Requirements of Livestock. Determining amino acid requirement of animals requires difficult, time-consuming experiments. The values derived from these experiments are not constants valid for all times but vary depending on environmental (heat stress, disease), genetic (sex, breed), and dietary factors (protein level, energy level, feed intake). There are essentially three methods for determining these requirements: carcass or milk analysis, synthetic rations, and semisynthetic rations.

For the first method, the amino acid content of a carcass is taken as a first-order approximation to the amino acid requirement of an animal. The same method can be applied for young suckling mammals by analyzing the amino acid content of the milk.

In the second method, the test animal is fed a synthetic mixture of all amino acids, along with an otherwise balanced ration (control group). The requirement for a single amino acid is determined by reducing its content in the diet to zero

Table 14. Amino acid composition of feedstuffs, wt % [272, 273]

Feedstuff	Dry matter	Crude protein	Met	Met+Cys	Lys	Thr	Trp	Arg	Leu	Ile	Val	His
Alfalfa	88	17.7	0.27	0.50	0.86	0.76	0.25	0.81	0.72	1.27	0.91	0.36
Barley	88	10.6	0.18	0.42	0.38	0.36	0.12	0.53	0.37	0.73	0.52	0.24
Beans, field	88	25.4	0.20	0.52	1.63	0.90	0.22	2.29	1.03	1.89	1.15	0.67
Blood meal	91	88.8	1.03	2.17	7.96	3.85	1.42	3.94	1.17	11.39	7.69	5.65
Corn	88	8.5	0.18	0.37	0.25	0.31	0.06	0.40	0.29	1.05	0.41	0.26
Corn gluten feed	88	19.0	0.32	0.71	0.58	0.68	0.11	0.85	0.59	1.69	0.89	0.59
Corn gluten meal	88	60.6	1.43	2.52	1.02	2.08	0.31	1.93	2.48	10.19	2.79	1.28
Feather meal	91	81.1	0.61	4.74	2.08	3.82	0.54	5.62	3.86	6.79	5.88	0.93
Fish meal	91	62.9	1.77	2.34	4.81	2.64	0.66	3.66	2.57	4.54	3.03	1.78
Meat and bone meal	91	49.1	0.68	1.18	2.51	1.59	0.28	3.45	1.34	2.98	2.04	0.91
Meat meal	91	48.8	0.68	1.24	2.44	1.63	0.30	3.42	1.40	2.99	2.13	0.93
Oat	88	12.6	0.22	0.57	0.53	0.44	0.14	0.87	0.48	0.92	0.66	0.31
Rapeseed meal,	88	34.8	0.70	1.59	1.95	1.53	0.45	2.15	1.37	2.47	1.77	0.97
Rice	88	7.3	0.20	0.37	0.26	0.26	0.09	0.60	0.29	0.59	0.40	0.18
Sesame meal	88	41.1	1.15	1.97	1.01	1.44	0.54	4.86	1.47	2.74	1.85	0.98
Sorghum	88	9.3	0.17	0.34	0.22	0.31	0.10	0.38	0.37	1.21	0.46	0.23
Soybean meal, 44 % CP*	88	44.0	0.64	1.31	2.75	1.76	0.57	3.28	2.01	3.44	2.09	1.21
Soybean meal, 48 % CP	88	47.6	0.69	1.41	2.98	1.89	0.61	3.52	2.16	3.71	2.23	1.31
Sunflower meal	88	33.5	0.77	1.36	1.19	1.25	0.40	2.75	1.37	2.15	1.66	0.88
Tapioca	88	3.3	0.04	0.09	0.12	0.11	0.04	0.18	0.11	0.19	0.14	0.08
Triticale	88	11.6	0.21	0.49	0.42	0.39	0.12	0.61	0.42	0.79	0.55	0.29
Wheat	88	13.3	0.21	0.50	0.38	0.38	0.15	0.64	0.44	0.87	0.56	0.32
Wheat bran	88	15.7	0.25	0.58	0.64	0.52	0.22	1.07	0.49	0.98	0.72	0.44
Wheat gluten feed	88	14.4	0.22	0.52	0.46	0.46	0.19	0.86	0.45	0.89	0.67	0.39
Wheat gluten meal	88	74.3	1.17	2.79	1.24	1.89	0.68	2.59	2.65	5.20	2.88	1.54

* Crude Protein

and from there on supplementing it stepwise up to an amount where the animal performs as well as the control group.

The third and most commonly used method is that a basal diet consisting of typically used feeds is formulated to be deficient in one amino acid but adequate in all other nutrients. To this basal diet graded levels of the amino acid in question are added until the performance of the animal approaches a maximum. The response in growth to stepwise increasing of amino acid supplementation follows the law of diminishing returns and can be described best by an exponential function.

Amino acid recommendations and requirements vary between species and for each species with age. There are several physiological reasons for that:

The various proteins deposited in the animal's body differ considerably in their amino acid composition. In poultry, for example, the proportion of lysine is much higher in muscle protein than in feather protein. In contrast, methionine and cystine are required in a higher percentage for the formation of feather protein and for maintaining metabolic functions than for muscle protein.

Not only the amino acid composition of the different proteins varies but also the quantity of each protein deposited in the animal changes with age. The daily accretion of muscle protein increases up to a certain age depending on the species and declines thereafter. The daily protein requirement for metabolic functions rises continually as the animal grows. Another reason for amino acid recommendations changing with age is the fact that young animals have a high potential for growth but at the same time a relatively small feed intake capacity which requires a high nutrient density in the diet.

Table 15. Amino acid recommendations, wt %, for poultry [273] and pigs [274, 275]

Species	Metabolizable energy, MJ/kg	Crude protein	Met	Met+ Cys	Lys	Thr	Trp
Broiler							
Starter	12.7	21	0.57	0.92	1.27	0.80	0.20
Grower	13.2	18	0.44	0.76	1.00	0.65	0.19
Finisher	13.2	17	0.41	0.74	0.95	0.63	0.11
Laying hen (105 g feed intake per day)	12.0	16	0.38	0.71	0.80	0.52	0.15
Turkey (weeks of age)							
0–4	11.7	28	0.66	1.15	1.80	1.06	0.31
5–8	12.1	25	0.59	1.04	1.60	0.95	0.27
9–12	12.6	22	0.53	0.92	1.40	0.84	0.24
13–16	13.0	19	0.46	0.80	1.20	0.73	0.20
> 16	13.4	16	0.43	0.75	1.10	0.67	0.19
Pig (kg live weight)							
10–19	10.4	18	0.48	0.87	1.40	0.94	0.25
20–30	10.2	17	0.40	0.74	1.15	0.79	0.23
40–70	9.8	14.5	0.35	0.64	0.95	0.68	0.19
70–105	9.6	13.5	0.30	0.55	0.82	0.59	0.16
Sow							
Lactating	9.8	16.5	0.36	0.65	1.00	0.70	0.20
Pregnant	8.9	12.5	0.23	0.42	0.70	0.46	0.14

The amino acid recommendations for domestic monogastric animals are listed in Table 15.

Economics of Amino Acid Supplementation. The purpose of modern, formulated feed mixes is to meet all nutritional requirements of the animal at a minimum cost, and amino acids and proteins are usually among the most expensive components of a feed mix. The performance of a feed mix is measured primarily by feed utilization and weight gain. Both feed utilization and weight gain, as well as other factors such as laying performance or feather or hair growth, are directly dependent on a sufficient supply of amino acids. Methionine and lysine play the major roles. Methionine in the form of D,L-methionine and lysine in the form of L-lysine · HCl are used in place of or as a supplement to natural methionine and lysine sources [276].

Linear programming is the method of choice for simultaneously optimizing a ration and minimizing cost [277]. This method allows simultaneous consideration of all demands that are made on the ration. A prerequisite is exact data on the nutrient content of all available feedstuffs and additives as well as their prices and availabilities. Additionally, the restrictions, i.e., the dietary requirements that the ration has to fulfill, must be known. Table 16 shows two examples of feed mixes formulated by the linear programming method commonly used.

Protein and Amino Acid Digestibilities of Feed Ingredients. The nutritive value of protein for monogastric animals is determined not only by the amino acid composition of the diet but also by digestibility of the individual amino acids, in particular the amino acids likely to be limiting. Over the last decades, considerable research has been carried out that demonstrates large differences in amino acid digestibilities between feeds.

Amino acid digestibilities can be determined according to the ileal or fecal analysis method. The fecal analysis method measures the amount of each amino acid consumed and excreted in feces. The amino acid digestibilities determined according to the ileal analysis method are calculated based on the intake and amount of each amino acid passing at the end of the distal ileum. The ileal analysis method should be considered as an improvement on the fecal analysis method, since the protein or amino acid absorbed in the large intestine make little or no contribution to the protein status of the animal. The ileal analysis method is also very sensitive to differences in amino acid digestibilities, as these result from processing conditions or from inherent differences between samples of the same feedstuff. Digestibilities measured over the entire digestive tract may be altered by resident bacteria. The bacteria may break down some of the amino acids, convert them to other amino acids, or even produce new amino acids. Measuring ami-

Table 16. Two examples of formulated feeds

Broiler feed composition, wt %		Pig fattening feed composition, wt %	
Yellow corn	28.00	Feed grain (barley, wheat, corn)	35.00
Wheat	20.00	Soybean meal (44 % crude protein)	19.00
Soybean meal (48 % crude protein)	30.00	Tapioca meal	20.00
Tapioca meal	10.00	Corn gluten feed	15.00
Fat	7.00	Meat and bone meal (45 % crude protein)	3.00
Meat and bone meal (45 % crude protein)	3.00	Fat	3.00
Mineral premix	1.25	Beet molasses	2.00
Vitamin-trace element premix	0.50	Mineral premix	2.43
D,L-Methionine	0.25	Vitamin-trace element premix	0.50
	100.00	D,L-Methionine	0.07
			100.00

no acids excreted in the feces will not reflect unabsorbed amino acid, but rather unabsorbed amino acids after possible alteration by the bacteria.

Amino acid digestibility values from the literature, determined with the ileal analysis method (Table 17) show large differences in amino acid digestibilities between feeds and among different samples of the same feed.

The differences in amino acid digestibility of feeds can be attributed to various factors. These include, e.g., fiber levels, heat damage during processing and the presence of antinutritional factors that interfere with nutrient digestion and utilization.

Studies have clearly shown improvements in diet formulation practices, when diets are formulated on the basis of ileal digestible rather than on supply of limiting amino acids. This holds especially true when a good quality protein supplement is replaced by protein supplement(s) of lower quality combined with supplementary amino acids. In consequence, amino acid requirements should be expressed as ileal digestible, rather than as total amino acid requirements.

Amino Acids as a Measure to Reduce the N-Output from Livestock Production. Animal production accounts for a significant portion of nitrogen containing compounds released into the environment. In areas with intensive livestock production this might result in environmental problems, especially if N-requirements for crop fertilization and N-output from livestock production get out of balance. In this case nitrogen containing compounds are released into the surface and ground water where they accumulate.

The total amount of nitrogen produced is dependent on both the average number of animals per unit available on and the efficiency of conversion of feed protein into body protein. This efficiency is often impaired due to diets that are not balanced for the specific amino acid requirements of the animal fed. The portion of protein that is fed without supplying an adequate mix of amino acids is thus excreted without being utilized by the animals. Meanwhile the animal's requirements for amino acids are well known which allows a reduction in the total protein content of diets as long as the diet is supplemented with amino acids in accordance with the animal's specific requirements. As a result, feed protein utilization is maximized and also water intake as a means to excrete excess nitrogen via the kidney is reduced.

In consequence the N-output from livestock production can be reduced down to 65 % when supplemental amino acids are used together with a reduction of the protein content of the diet.

5.3. Pharmaceuticals

The pharmaceutical industry requires amino acids at a rate between 2000 and 3000 t/a worldwide. More than half of this is used for infusion solutions. During the last few years the potential of amino acids and their derivatives as active ingredients for pharmaceuticals has been recognized clearly, and considerable growth can be predicted.

Table 17. Range of ileal digestibilities (%) of lysine, methionine, and threonine in different feeds for pigs modified from [274]

Amino acids	Lysine	Methionine	Threonine
Cereal grains			
Barley	65–79	72–88	64–76
Wheat	62–81	79–92	51–78
Corn	71–82	88–92	74–79
Protein supplements			
Soybean meal (44 % CP*)	85–89	77–90	73–81
Canola meal	74–76	81–87	66–67
Meat and bone meal	58–67	72–79	53–62
Cottonseed meal	53–70	65–82	55–69

* Crude Protein

5.3.1. Nutritive Agents

Infusion Solutions. Parenteral nutrition with L-amino acid infusion solutions is a well-established component of clinical nutrition therapy. A standard infusion solution contains the eight classical essential amino acids, the semiessential amino acids L-arginine and L-histidine, and several nonessential amino acids, generally glycine, L-alanine, L-proline, L-serine, and L-glutamic acid.

Also available are special infusion solutions tailored to the requirements of particular groups, such as newborn infants, seniors, or patients with an extreme negative nitrogen balance. Solutions rich in the branched-chained amino acids leucine, isoleucine, and valine and poor in methionine and aromatic amino acids are available for liver-disease patients. Solutions containing only essential amino acids are available for kidney patients. Enzymatic protein hydrolysates, which were used as infusion solutions until the 1980s have disappeared almost completely from the market. They were not available in the optimal composition, and there were often compatibility problems. Only pure, crystalline L-amino acids are used in modern infusion solutions. The solutions (up to 10 %), which also contain electrolytes in addition to amino acids, are sterile and pyrogen-free.

The simultaneous administration of carbohydrates is necessary for optimal utilization of the amino acids. Glucose is normally a separate infusion. Some commercially available amino acid infusion solutions contain an energy source in the form of sugar alcohols (sorbitol, xylitol), which do not enter into a Maillard reaction with the amino acids.

Normally, parenteral nutrition is only practiced over a limited time. In principle, however, total parenteral nutrition over many years is possible. In such a case, all essential nutrients (unsaturated fatty acids, vitamins, and trace elements) must be provided.

Elemental Diets. Enteral nutrition is also a means of providing the essential nutrients [278]. Elemental diets, which were developed originally for the astronauts [389], contain chemically defined nutritive components. In addition to free amino acids the mixtures generally contain carbohydrates, fats, minerals, and vitamins in a combination adapted to the requirements. In many cases, elemental diets are used as an alternative and supplement to parenteral nutrition. They have high nutritional value and are totally resorbable. They are largely independent of the digestive function of the pancreas and reduce the intestinal bacteria flora. Amino acid elemental diets generally are used in cases of anatomic, functional, or enzymatic defects [279].

Formula diets based on peptides are gaining ground as an alternative to elemental diets based on L-amino acids. According to [280], short-chained peptides are resorbed rapidly via a peptide transport system in the gut, therefore in a process that is independent of amino acid transport. Compositions of nitrogen-free amino acid analogues (keto acids and hydroxy acids) have come into use for the special case of kidney insufficiency (chronic renal failure).

Elemental diets or formula diets are administered orally or via a nasogastric tube directly into the gastrointestinal tract.

5.3.2. Therapeutic Agents

Many therapeutic agents are derivatives of natural or nonnatural amino acids. Examples are benserazide, captopril, and dextrothyroxine. They are described under keywords

such as → Spasmolytics, → Antihypertensives, → Antihypotensives, or → Thyrotherapeutic Agents. Only therapeutically useful amino acids and simple derivatives are treated here.

Amino Acids and Salts. The amino acids and their simple salts that are important therapeutic agents are compiled in Table 18. The proprietary names listed represent only a selection.

***N*-Acetylcysteine** [*616-91-1*], $C_5H_9NO_3S$, M_r 163.2, *mp* 109–110 °C, $[\alpha]_D^{20}$ + 5° (*c* = 3, H_2O), is a mucolytic and secretolytic agent. It is also indicated for the treatment of paracetamol poisoning.

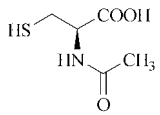

N-Acetylcysteine is prepared by reaction of cysteine hydrochloride monohydrate with acetic anhydride in the presence of sodium acetate [281, 282].

Trade names: Exomouc (Bouchara), Fluimucil (Zambon) and numerous others [283].

Carbocisteine (carbocysteine) [*638-23-3*], *S*-carboxymethyl-L-cysteine, $C_5H_9NO_4S$, M_r 179.2, *mp* 204–207 °C (decomp.), $[\alpha]_D^{20}$ −34.0 to −36.0° (*c* = 10, H_2O), is used to treat disorders of the respiratory tract associated with excessive mucus.

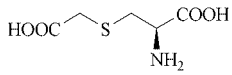

Synthesis involves *S*-alkylation of L-cysteine with chloroacetic acid in the presence of sodium hydroxide [284, 285].

Trade names: Mucodyne (Sanofi-Aventis) and numerous others [283].

Fudosteine [*13189-99-5*], *S*-(3-hydroxypropyl)- L-cysteine, $C_6H_{13}NO_3S$, M_r 179.24, *mp* 198–202 °C, $[\alpha]_D^{20}$ −22° (*c* = 1, H_2O), was launched by Mitsubishi Pharma in 2001 for the treatment of bronchial hypersecretion.

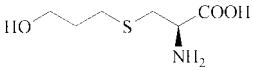

Fudosteine is synthesized by condensation of L-cysteine with 3-bromopropyl alcohol in aqueous NaOH, or by condensation of L-cysteine with allyl alcohol by means of aqueous potassium persulfate [286].

Trade names: Cleanal, Spelear (Mitsubishi Pharma).

Levodopa [*59-92-7*], (*S*)−3-(3,4-dihydroxyphenyl)alanine, $C_9H_{11}NO_4$, M_r 197, *mp* 285.5 °C (decomp.), $[\alpha]_D^{20}$ −12.15° (*c* = 4 in 1 M HCl).

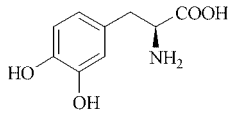

Levodopa is widely used for treatment of Parkinson's disease, most often in combination with peripheral decarboxylase inhibitors such as benserazide and carbidopa. For manufacture and trade names → Parkinsonism Treatment.

Trade names: Larodopa (Roche) and numerous others [283].

Mecysteine hydrochloride [*18598-63-5*], cysteine methyl ester hydrochloride, methyl L-2-amino-3-mercaptopropionate hydrochloride, $C_4H_{10}ClNO_2S$, M_r 171.66, *mp* 140–141 °C, is used in the treatment of disorders of the respiratory tract associated with excessive mucus. It is prepared by esterification of L-cysteine hydrochloride monohydrate with methanol in the presence of hydrogen chloride [287].

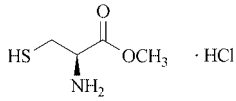

Trade names: Acthiol (Joullié, France), Actiol (Lirca, Italy), Visclair (Sinclair, UK), Aslos-C (Nissin, J), Epectan (Seiko, Japan), Moltanine (Tohok.-Tokyo Tanabe, Japan), Radcol (Nippon Universal, Japan), Sekinin (Tokyo Hosei, Japan).

Methiosulfonium chloride (iodide) [*1115-84-0*], L-methylmethionine sulfonium chloride, vitamin U, $C_6H_{14}ClNO_2S$, M_r 199.7; *mp* 134 °C (decomp.). Iodide [*3493-11-6*], $C_6H_{14}INO_2S$, M_r 291.1.

Table 18. Amino acids and their salts as drugs

Compound		Formula	M_r	Medical use	Trade names
L-Alanine	[*56-41-7*]	$C_3H_7NO_2$	89.09	parenteral nutrition, stimulant of glucagon secretion	
L-Arginine	[*74-79-3*]	$C_6H_{14}N_4O_2$	174.20	parenteral nutrition, stimulation of pituitary release of growth hormone and prolactin, stimulation of pancreatic release of glucagon and insulin	
L-aspartate	[*7675-83-4*]	$C_{10}H_{21}N_5O_6$	307.31	treatment of hyperammonemia	Argihepar, Laevil, Potentiator, Sargenor, Sorbenor
L-glutamate	[*4320-30-3*]	$C_{11}H_{23}N_5O_6$	321.34	treatment of hyperammonemia	Modumate
hydrochloride	[*1119-34-2*]	$C_6H_{15}ClN_4O_2$	210.66	treatment of hyperammonemia, electrolyte concentrate for i.v. solutions	Argivene, R-Gene, Polilevo
L-pyroglutamate		$C_{11}H_{21}N_5O_5$	303.27		
2-oxoglutarate	[*16856-18-1*]	$C_{11}H_{20}N_4O_7$	320.3	treatment of hepatic disorders	Leberam, Anetil, Argiceto, Eucol
L-Asparagine	[*70-47-3*]	$C_4H_8N_2O_3$	132.13	parenteral nutrition	
monohydrate	[*5794-13-8*]	$C_4H_{10}N_2O_4$	150.13		
D,L-Aspartic acid	[*617-45-8*]	$C_4H_7NO_4$	133.10		
magnesium, tetrahydrate	[*52101-01-6*]	$C_8H_{20}N_2O_{12}Mg$	360.54	cardiac agent, management of fatigue, mineral supplement	Magnesium Verla,
potassium, semihydrate	[*923-09-1*] (anhydrous)	$C_4H_6NO_4K \cdot 1/2\,H_2O$	180.20	cardiac agent, management of fatigue, mineral supplement	Trommcardin, Trophicard-Köhler
sodium, monohydrate		$C_4H_8NO_5Na$	173.10		
L-Aspartic Acid	[*56-84-8*]	$C_4H_7NO_4$	133.10	parenteral nutrition	
ferrous, tetrahydrate		$C_8H_2ON_2O_{12}Fe$	392.1	treatment of iron deficiency	Sideryl, Spartocine
magnesium, dihydrate	[*2068-80-6*]	$C_8H_{16}N_2O_{10}Mg$	324.52	management of fatigue and heart conditions	
potassium, semihydrate	[*1115-63-5*] (anhydrous)	$C_4H_6NO_4K \cdot 1/2\,H_2O$	180.20	management of fatigue and heart conditions	Magnesiocard, Corroverlan
sodium, monohydrate	[*3792-50-5*] (anhydrous)	$C_4H_8NO_5Na$	173.10	parenteral nutrition	
Betaine	[*107-43-7*]	$C_5H_{11}NO_2$	117.15		
citrate	[*17671-50-0*]	$C_{11}H_{19}NO_9$	309.27	lipotropic	Flacar, Panstabil
hydrochloride	[*590-46-5*]	$C_5H_{12}ClNO_2$	153.61	lipotropic, gastric acidifier	Aciventral, Acidol, Pesimuriat
monohydrate	[*17146-86-0*]	$C_5H_{13}NO_3$	135.15	lipotropic	Hepaderichol
L-Citrulline	[*372-75-8*]	$C_6H_{13}N_3O_3$	175.19	treatment of hepatic disorders	Polilevo
L-Cysteine	[*52-90-4*]	$C_3H_7NO_2S$	121.16	treatment of damaged skin, topically in ophthalmology, detoxicant	Reducdyn, Hepa-Loges, Irradian, Cicatrex, Felacomp
hydrochloride	[*52-89-1*]	$C_3H_8ClNO_2S$	157.62		Cheihepar, Choldestal
hydrochloride monohydrate	[*7048-04-6*]	$C_3H_{10}ClNO_3S$	175.64		
L-Cystine	[*56-89-3*]	$C_6H_{12}N_2O_4S_2$	240.30	parenteral nutrition, lipotropic agent, treatment of hair and nail damage	Cystin "Brunner", Gerontamin, Pantovipar, Priorin
L-Glutamic acid	[*56-86-0*]	$C_5H_9NO_4$	147.13	parenteral nutrition, dietary supplement, treatment of hyperammonemia	Aciglut, Glutamin-Verla
calcium, dihydrate	[*5996-22-5*]	$C_{10}H_{20}N_2O_{10}Ca$	368.2		Vivacalcium
hydrochloride	[*138-15-8*]	$C_5H_{10}ClNO_4$	183.54	gastric acidifier	Bioprotein-Holzinger, Pansan, Pepsalara
monosodium, monohydrate	[*142-47-2*] (anhydrous)	$C_5H_{10}NO_5Na$	187.13	flavoring and seasoning of food	
magnesium, tetrahydrate	[*19238-50-7*]	$C_{10}H_{24}N_2O_{12}Mg$	388.62	tonics, mineral supplements	Magnesium Verla, Glutergen
potassium, monohydrate	[*19473-49-5*] (anhydrous)	$C_5H_{10}NO_5K$	203.24	tonics, mineral supplements	

Table 18. (Continued)

Compound		Formula	M_r	Medical use	Trade names
magnesium, hydrobromide		$C_{10}H_{17}BrN_2O_8Mg$	397.5	tranquilizer	Psicosoma, Psychoverlan
L-Glutamine	[*56-85-9*]	$C_5H_{10}N_2O_3$	146.15	parenteral nutrition, treatment of mental disorders and alcoholism	
Glycine	[*56-40-6*]	$C_2H_5NO_2$	75.07	parenteral nutrition, antacid in conjunction with calcium carbonate	Cicatrex, Felacomp, Gastripan-K, Mirogastrin
aluminum, hydrate	[*13682-92-3*]	$C_2H_6NO_4Al$ $(+xH_2O)$	135.1	antacid	Acidrine, Al-Glycin
L-Histidine	[*71-00-1*]	$C_6H_9N_3O_2$	155.16	parenteral nutrition, essential amino acid for infants	
acetate		$C_8H_{13}N_3O_4$	215.21		
hydrochloride monohydrate	[*5934-29-2*]	$C_6H_{12}ClN_3O_3$	209.63		Anti-rheuma, Rollkur-Ankermann, Laristine, Plexamine
L-Isoleucine	[*73-32-5*]	$C_6H_{13}NO_2$	131.18	parenteral nutrition, treatment of hepatic coma, dietary supplement	
L-Leucine	[*61-90-5*]	$C_6H_{13}NO_2$	131.18	parenteral nutrition, treatment of hepatic coma, dietary supplement	numerous combinations
D,L-Lysine	[*70-54-2*]	$C_6H_{14}N_2O_2$	146.19	formation of salts with acidic drugs (to enhance solubility)	
monohydrochloride	[*70-53-1*]	$C_6H_{15}ClN_2O_2$	182.65	geriatric	Jestrosemin
acetylsalicylate	[*34220-70-7*]	$C_{15}H_{22}N_2O_6$	326.34	soluble form of acetylsalicylic acid (aspirin) for injection	Aspisol, Delgesic
L-Lysine	[*56-87-1*]	$C_6H_{14}N_2O_2$	146.19	parenteral nutrition, dietary supplement, prophylaxis of	numerous combinations
acetate	[*57282-49-2*]	$C_8H_{18}N_2O_4$	206.24	herpes simplex infection (?)	
L-aspartate	[*27348-32-9*]	$C_{10}H_{11}N_3O_6$	279.30	dietary supplement	
L-glutamate	[*5408-52-6*]	$C_{11}H_{23}N_3O_6$	293.32	dietary supplement	
L-malate	[*71555-10-7*]	$C_{10}H_{20}N_2O_7$	280.28	dietary supplement	
monohydrate	[*39665-12-8*]	$C_6H_{16}N_2O_3$	164.21	dietary supplement	
monohydrochloride	[*657-27-2*]	$C_6H_{15}ClN_2O_2$	182.65	treatment of hypochloremia, alkaloses	Aktivanad, Athensa, Omnival, Vivioptal
D,L-Methionine	[*59-51-8*]	$C_5H_{11}NO_2S$	149.21	lipotropic and choleretic agent	numerous combinations
L-Methionine	[*63-68-3*]	$C_5H_{11}NO_2S$	149.21	parenteral nutrition, dietary supplement, lipotropic agent, treatment of paracetamol poisoning	
L-Ornithine					
acetate	[*60259-81-6*]	$C_7H_{16}N_2O_4$	192.22	parenteral nutrition	
L-aspartate	[*3230-94-2*]	$C_9H_{19}N_3O_6$	265.27	treatment of hepatic disorders	Hepa-Merz
monohydrochloride	[*3184-13-2*]	$C_5H_{13}ClN_2O_2$	168.62	parenteral nutrition	Ornitaine, Polilevo
2-oxoglutarate	[*5191-97-9*]	$C_{10}H_{18}N_2O_7$	278.14	treatment of hepatic disorders (hyperammonemia)	Ornicetil
D-Phenylalanine	[*673-06-3*]	$C_9H_{11}NO_2$	165.19	antidepressant	
L-Phenylalanine	[*63-91-2*]	$C_9H_{11}NO_2$	165.19	parenteral nutrition	
L-Proline	[*147-85-3*]	$C_5H_9NO_2$	115.13	parenteral nutrition, dietary supplement, starting material for captopril and enalapril	
L-Pyroglutamic acid	[*98-79-3*]	$C_5H_7NO_3$	129.07	formation of salts with basic drugs	
D,L-Serine	[*302-84-1*]	$C_3H_7NO_3$	105.09	starting material for benserazide	Aktiferrin
L-Serine	[*56-45-1*]	$C_3H_7NO_3$	105.09	parenteral nutrition, dietary supplement	Sulfolitruw
L-Threonine	[*72-19-5*]	$C_4H_9NO_3$	119.12	parenteral nutrition, dietary supplement	
L-Tryptophan	[*73-22-3*]	$C_{11}H_{12}N_2O_2$	204.23	parenteral nutrition, antidepressant, sleep inducer, dietary supplement	Optimax, Kalma, Pacitron
L-Tyrosine	[*60-18-4*]	$C_9H_{11}NO_3$	181.19	parenteral nutrition, dietary supplement	
L-Valine	[*72-18-4*]	$C_5H_{11}NO_2$	117.15	parenteral nutrition, dietary supplement, treatment of hepatic coma	

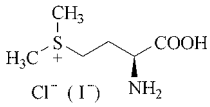

Methiosulfonium chloride is used for its protective effect on the liver and gastrointestinal mucosa, whereas the iodide finds use for rheumatic disorders. The compounds are made by heating L-methionine with methyl chloride or methyl iodide [288].

Trade names: Chloride: Ardesyl (Beytout, France), withdrawn from the market; Cabagin (Kowa, Japan). Iodide: Lobarthrose (Opodex, France), withdrawn from the market.

Oxitriptan [*4350-09-8*], (*S*)-5-hydroxytryptophan, $C_{11}H_{12}N_2O_3$, M_r 220, *mp* 273 °C (decomp.), $[\alpha]_D^{22}$ −32.5° (*c* = 1 in H_2O), $[\alpha]_D^{22}$ + 16.0° (*c* = 1 in 4 M HCl).

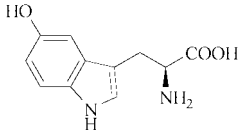

This intermediate in mammalian biosynthesis of serotonin is used as an antidepressant. It is produced either by total synthesis (analogous to L-tryptophan via the 5-benzyloxy derivative) [289–291] or by fermentation with *Chromobacterium violaceum* [292].

Trade names: Levothym (Promonta Lundbeck, Germany), Lévotonine (Panpharma, France), Oxyfan (Coli, Italy), Tript-OH (Sigma-Tai, Italy).

D-Penicillamine [*52-67-5*], D-3-mercaptovaline, D-β,β-dimethylcysteine, $C_5H_{11}NO_2S$, M_r 149.21, *mp* 198.5 °C, $[\alpha]_D^{25}$ −63° (*c* = 0.1, pyridine). Hydrochloride [*2219-30-9*], $C_5H_{12}ClNO_2S$, M_r 185.7, *mp* 177.5 °C (decomp.) $[\alpha]_D^{25}$ −63° (*c* = 1 in 1 M NaOH).

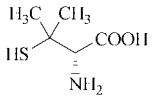

D-Penicillamine is a chelating agent that aids the elimination of toxic metal ions, e.g., copper in Wilson's disease. It is used, as an alternative to gold preparations, in the treatment of severe rheumatoid arthritis. It is useful in treating cystinuria because it reacts with cystine to form cysteine-penicillamine disulfide, which is much more soluble than cystine.

D-Penicillamine is produced by hydrolysis of benzylpenicillin (→ Antibiotics, Chap. 4) via its Hg(II) complex [293] or by total synthesis. In the synthesis, isobutyraldehyde, sulfur, and ammonia are condensed to 5,5-dimethyl-2-isopropyl-Δ^3-thiazoline, which, on reaction with hydrogen cyanide, gives 5,5-dimethyl-2-isopropyl-thiazolidine-4-carbonitrile. Hydrolysis with boiling hydrochloric acid yields D,L-penicillamine hydrochloride. Cyclization with acetone and formylation leads to D,L-3-formyl-2,2,5,5-tetramethylthiazolidine-4-carboxylic acid, which can be resolved with (−)-phenylpropanolamine via the diastereomeric salts. Hydrolysis with hydrochloric acid leads to D-penicillamine hydrochloride [294].

Trade names: Cuprimine (Merck Sharp & Dohme, USA), Depen (Carter-Wallace, USA), Distamine (Dista, UK), Metalcaptase (Heyl, Germany), Trolovol (ASTA Medica AWD, Germany), Trisorcin (Merdele, Germany), Trolovol (Bayer, France), Pendramine (ASTA Medica, UK), Pemine (Lilly, Italy).

5.4. Cosmetics

Amino acids are a major component of the "natural moisturizing factor" (NMF) that protects the skin surface from dryness, brittleness, and a deleterious environment [295]. The epidermis of the skin contains about 15 % water, which, in the presence of amino acids, principally serine, citrulline, and alanine, forms a stable water-in-oil emulsion with the skin lipids in the form of a thin layer. The amino acids simultaneously stabilize the pH of the skin (the acidic layer), and can be absorbed into the skin and hair. Because of these properties, amino acids [296, 297], protein hydrolysates [298], and proteins [296] are widely utilized in skin and hair cosmetics, e.g., in mild skin creams, skin cleansing lotions, and hair shampoos. It has been claimed that 60 % of hair strength is accomplished by three amino acids, i.e., histidine, lysine, and tyrosine, and supplementing these in hair shampoos can reduce loss of hair strength and improve shine [299]. Methionine is absorbed into the scalp and absorbed by the hair fibers as cysteine, which

is responsible for the cross-linking of keratin [300]. A report published in 2004 claims that amino acids comprise the fastest growing area of new ingredients in the moisturizers and humectants market [301]. Amino acids manufactured by fermentation rather than extracted from animal protein have been increasingly promoted as being more "natural", and this has helped to increase the market for these products.

The sodium salts of the reaction products of fatty acids with amino acids, such as glutamic acid [302–305], or short-chain peptides (protein hydrolysates) [306, 307] are surfactants. They are effective, skin-compatible cleaners and emulsifiers, which are used in shampoos, shower gels, baby baths, medicinal skin cleansers, cold-wave preparations, etc. [303, 308, 309]. Arginine has been used as the cation in soaps to improve the foam volume, and the free amino acid has been used in conditioners to improve sensory properties [310]. The sulfur-containing amino acids exhibit a special normalizing effect on skin metabolism, e.g., in cases of excess skin lipid production (seborrhea), dandruff, or acne. Substances utilized for this purpose include derivatives of cysteine (e.g., *S*-carboxymethylcysteine), homocysteine (2-amino-4-mercaptobutyric acid), and methionine [311]. Amino acids are also used in hair lotions, where they are reported to have a nutritive effect [312]. Cysteine, which acts as a reducing agent, is gaining importance, especially in Japan, as a substitute for thioglycolic acid in permanent wave preparations that are less damaging to hair [313]. Use of aluminum, tin, and zirconium complexes of amino acids, especially glycine, as deodorants [314] and antiperspirants [314, 315] has been reported.

The use of peptides as mimics of human protein fragments in cosmetics has been developed since 2000. One such product is Matrixyl (Sederma), a pentapeptide (Palmitoyl-Lys-Thr-Thr-Lys-Ser), which is a mimic of a precollagen fragment [316]. Matrixyl is applied in antiwrinkle creams. Another example is the dipeptide *N*-acetyl-Tyr-Arg-cetyl ester, registered as Calmosensine (Sederma), which stimulates endorphin precursor release in skin cells and acts as a mild analgesic [317].

5.5. Agrochemicals

An increasing number of pesticides [318] are derivatives of natural or nonnatural amino acids. Important amino acids like Glyphosate or Glufosinate are major products in the agrochemical market. The synthesis of the active ingredient may start with an amino acid, but very often the amino acid moiety is formed during the chemical synthesis.

5.5.1. Herbicides (→ Weed Control)

All pesticide segments (i.e., herbicides, fungicides, insecticides) contain compounds with amino acid substructures, but the major application of pesticidal amino acids is in the herbicide area [319].

Glyphosate [*1071-83-6*] [320], Roundup, $C_3H_8NO_5P$, M_r 169.1, *mp* 200 °C, is the isopropylamine salt of a phosphorus containing glycine derivative with a hybrid structure [321].

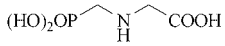

Its mode of action is unique; it inhibits different enzyme systems that leads to the blocking of different amino acid biosyntheses. The substance is used as a nonselective contact herbicide to control deep rooted weeds. The sales [322] in 1997 grew to about $\$2180 \times 10^6$ and reaching this figure it is the worlds top selling pesticide. The growth rates were double digits in the 1990s (17 %) and for 2002 an annual production of 180 000 t was estimated. Additional driving impact on the sales of glyphosate has been achieved by the introduction of glyphosate resistant crops (1996: soy beans, 1997: canola and cotton, 1998: maize) and the consequent distribution of genetically engineered seeds in combination with this herbicide.

The classical synthesis starts with iminodiacetic acid, phosphorous trichloride and formaldehyde, but also sequences using glycine and dimethyl phosphate have been applied. There are no confirmed reports of resistance. The expiration of the original patent and the advent of generic producers guarantee an increase of production of the substance becoming a bulk chemical.

Sulfosate [*81591-81-3*] [323], Touchdown, $C_6H_{16}NO_5PS$, M_r 245.2, is the trimethylsulfonium salt of Glyphosate also applied as a nonselective herbicide that was introduced in 1989. As in the case of Glyphosate, sales are growing well.

Glufosinate [324], Basta, Phosphinothricine [*53369-07-6*], $C_5H_{12}NO_4P$, M_r 181.1, *mp* 215 °C (for the ammonium salt). This phosphorous analogue of glutamic acid was introduced in 1984 as a nonselective herbicide in the speciality market, but meanwhile also resistant crops have been developed, which are distributed in combination with the herbicide (Liberty Link). In 1997 the sales were at $170 × 10^6.

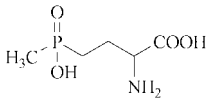

The contact herbicide inhibits the plant specific enzyme glutamine synthetase. The chemical synthesis requires acrolein, hydrocyanic acid and methyl phosphinous ester, resulting in a racemic product [325]. As the L-enantiomer is 40 times more active than the D-compound, many asymmetric synthetic routes are under investigation.

Bialofos [*35597-43-4*] [326], Bilanafos, $C_{11}H_{22}N_3O_6P$, M_r 323.3, *mp* 160 °C (decomp.), is a natural occurring tripeptide of the sequence L-phosphinothricine-L-alanine-L-alanine isolated from *Streptomyces viridochromogenes, Streptomyces hygroscopicus* and others. The compound was launched as a nonselective contact herbicide in 1984. It has the same mode of action as Glufosinate, but has much lower sales.

Herbicides Based on 2-Methylvaline. In 1985, the first member of a new and very important class of selective herbicides based on 2-methylvaline was introduced. These imidazolinone herbicides are acetolactate synthase (ALS) inhibitors [327] meaning that the synthesis of branched amino acids is blocked [328].

The synthesis of the racemic active ingredients starts from methyl isopropyl ketone which is converted to 2-methylvaline and finally cyclized to an imidazolinone with an *ortho*-carboxy substituted aromatic or heteroaromatic ring system. The sales of the imidazolinone herbicides grew with 6.8 % per annum from 1992 to 1997. The sales for Imazethapyr, the star product against mono- and dicotyl weeds in soya were about $540 × 10^6 in 1997.

Imazaquin [*81335-37-7*], 2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-3-quinoline carboxylic acid, Scepter, $C_{17}H_{17}N_3O_3$, M_r 311.3, *mp* 219–224 °C, was launched in 1984.

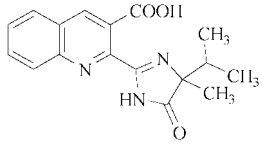

Imazapyr [*81334-34-1*], 2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-nicotinic acid, Arsenal, $C_{13}H_{15}N_3O_3$, M_r 261.3, *mp* 169–173 °C, was launched in 1985.

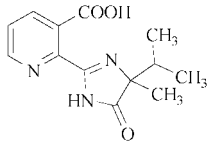

Imazamethabenz [*81405-85-8*], 2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-5-methylbenzoic acid methyl ester, Assert, $C_{15}H_{18}N_2O_3$, M_r 274.3, *mp* 113–153 °C (of methyl ester), was launched in 1986.

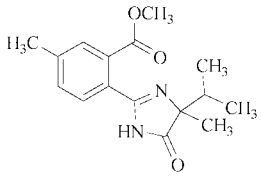

Imazethapyr [*81335-77-5*], 5-ethyl-2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)nicotinic acid, Pursuit, $C_{15}H_{19}N_3O_3$, M_r 289.3, *mp* 169–173 °C, was launched in 1987.

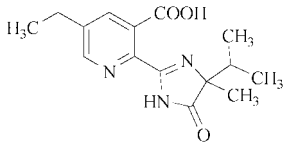

Imazapic [*104098-48-8*], 5-methyl-2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-nicotinic acid, Cadre, $C_{14}H_{17}N_3O_3$,

M_r 275.3, *mp* 204–206 °C, was launched in 1997.

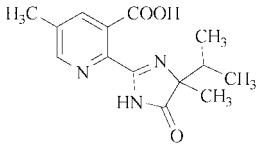

Imazamox [*114311-32-9*], 2-(4-isopropyl-4-methyl-5-oxo-4,5-dihydro-1*H*-imidazol-2-yl)-5-methoxymethylnicotinic acid, Raptor, $C_{15}H_{19}N_3O_4$, M_r 305.3, *mp*, was launched in 1997.

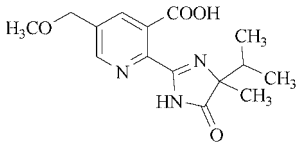

Herbicides derived from Acylphenyl Amino Acids. Acylphenyl amino acids are a group of herbicides with decreasing importance. The development of enantiomerically pure compounds are new concepts for improving the substance performance.

Benzoylprop [*22212-56-2*] [329], Suffix, $C_{16}H_{13}C_{12}NO_3$ (for ethyl ester), has an alanine subunit, which is formed by reacting 3,4-dichloroaniline with racemic ethyl 2-chloropropionate and benzoyl chloride. This herbicide is mainly used in wheat. The racemic active ingredient and the analogue with an 3-chloro-2-fluoroaniline aromatic are more and more replaced by the D-alanine derivative (racemic switch).

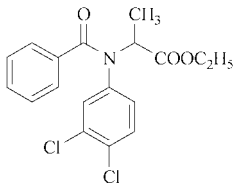

Flamprop-M [*63782-90-1*] [330], Mataven L, Suffix BW, $C_{19}H_{19}ClFNO_3$, M_r 363.8, *mp* 72.5–74.5 °C (data for isopropyl ester), is synthesized from 3-chloro-4-fluoroaniline, benzoyl chloride and methyl *S*-2-chloro propionate. This compound is used as selective grass herbicide in wheat and barley.

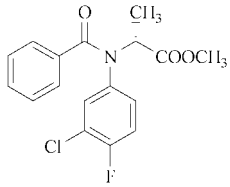

Amino Acids in Protox Herbicides. Inhibitors of the protoporphyrinogene oxidase [331], also known as Protox, are important herbicides blocking the biosynthesis of chlorophyll in plants. Phenyl substituted heterocycles have a strong impact on that enzyme [332]. These herbicides can contain amino acids in two ways. On the one side, amino acids like proline, 4-oxaproline [333] or pipecolinic acid [334] could be part of the heterocycle forming a hydantoin, on the other side amino acids like alanine can form the substituent at 5-position of the aromatic system. The preemergence activity and selectivity is increased using the R-enantiomer instead of the racemic alanine.

Herbicides Based on Glycine. Two herbicides based on a glycine residue are the chloroacetanilide herbicide Diethatyl [*38725-95-0*], Antor, $C_{14}H_{18}ClNO_3$ M_r 283.75, produced by Hercules Agrochemicals [318]

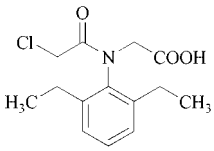

and the triazine herbicide Eglinazine [*68228-19-3*], $C_7H_{10}ClN_5O_2$, M_r 231.64, introduced by Nitrokémia Ipartelepek [318]

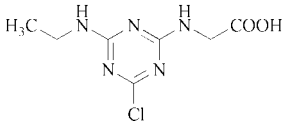

5.5.2. Fungicides (→ Fungicides, Agricultural)

Valinamide Fungicides. A lot of acylvaline anilides are claimed to be active against a broad spectrum of fungi, but especially the control of *Plasmopara* and *Phytophthora* [335] is pointed out. The literature indicates [336] that the *S*-compound is the more active enantiomer. This

group of compounds may be divided into valine derivatives and nonvaline derivatives. The valine is usually acylated to form an urethane or a maleimide. The carboxylic group is condensed with a substituted aniline, an alkyl benzylamine or an alkyl homobenzylamine [337, 338].

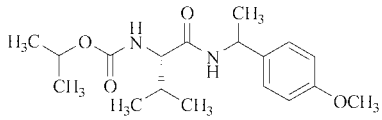

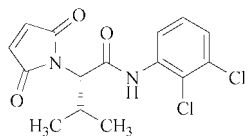

Two valinamide fungicides have been introduced into the market.

Iprovalicarb [*140923-17-7*], Melody, $C_{18}H_{28}N_2O_3$, M_r 320.43, has been introduced by Bayer CropScience for the prevention of downy mildew and *Phomopsis viticola* in vines [339]. The valine moiety is in the L-form, but the stereochemistry of the second asymmetric center is not defined.

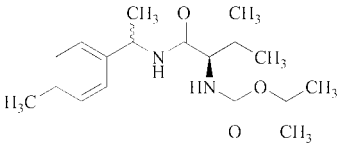

Benthiavalicarb [*413615-35-7*], $C_{15}H_{18}FN_3O_3S$, M_r 339.38, has been developed by Kumiai Chemicals Industry for treatment of tomatoes and potato late blight [340].

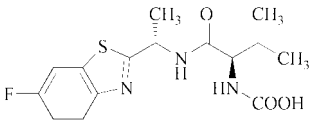

Anilide Fungicides. A number of anilide fungicides with a market volume of approx. $610 × 10^6 in 1997 have been established since the 1990s. Three amino acid derivatives are sold as fungicides with systemic activity.

Metalaxyl [*57837-19-1*] [341], Ridomil, $C_{15}H_{21}NO_4$, M_r 279.3, *mp* 71.8–72.3 °C (data for isopropyl ester). This most important fungicide of its class is synthesized by condensation of 2,6-dimethylaniline with methoxyacetyl chloride and subsequent coupling with racemic methyl 2-chloropropionate.

This fungicide is mainly used for the seed treatment business, in wine and in vegetables. The compound is sold as a stand-alone product or in a variety of mixtures with other fungicides. It was introduced in the market in 1978. The (*R*)-isomer, based on D-alanine, is known as Metalaxyl-M [*70630-17-0*] or Ridomil Gold, and was introduced by Syngenta in the USA in 1996.

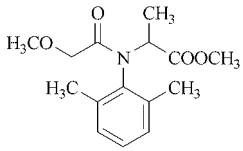

Furalaxyl [*57646-30-7*], Fongarid, $C_{17}H_{19}NO_4$, M_r 301.3, *mp* 70 and 84 °C (dimorphic) (data for isopropyl ester). Furalaxyl is a condensation product of furan-2-carboxylic acid and 2,6-dimethylaniline which is subsequently coupled with methyl 2-chloropropionate. The compound was introduced in 1984.

Benalaxyl [*71626-11-4*] [342], Galben, $C_{20}H_{23}NO_3$, M_r 325.4, *mp* 78–80 °C. In the commercial synthesis again 2,6-dimethylaniline is coupled with phenylacetyl chloride and subsequently methyl 2-bromopropionate. The single enantiomer product Benalaxyl-M [*98243-83-5*], based on D-alanine, has also been registered.

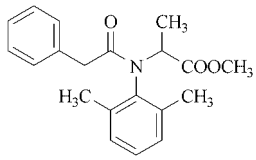

Hydantoins. Some amino acids being marketed as fungicides are sold as their hydantoins. One established product is iprodione.

Iprodione [*36734-19-7*], Rovral, $C_{13}H_{13}Cl_2N_3O_3$, M_r 330.2, *mp* 134 °C. The synthesis uses 3,5-dichloroaniline which is coupled with glycine and phosgene. The urea substructure is formed by reaction with isopropylamine and phosgene. Many fungicide mixtures with Iprodione are on the market.

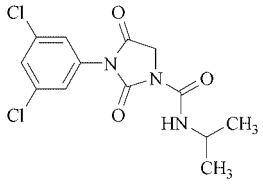

Perfuazoate [*101903-30-4*], Healthied, $C_{18}H_{23}N_3O_4$, M_r 345.4. This rice fungicide has been introduced by Showa Denko for the treatment of rice seedlings in Japan [318]. It is based on a racemic homoalanine pentenyl ester backbone.

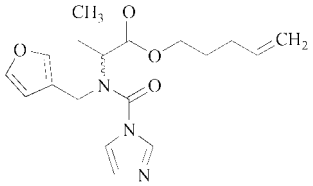

5.5.3. Insecticides (→ Insect Control)

Only a few amino acid based insecticides are distributed in the market.

tau-Fluvalinate [*102851-06-9*] [343], Klartan, Mavrik, $C_{26}H_{22}ClF_3N_2O_3$, M_r 502.9, *bp* 164 °C (9.33 Pa) is the most important compound. This substance, preferably used in cotton fields, but also in corn, rape tomatoes and vegetables, is a pyrethroid with contact and stomach action. The market of that class had a volume of $1530 × 10^6 in 1997. The chemical synthesis starts with 3,4-dichlorotrifluoromethylbenzene and D-valine. The carboxylic acid is esterified with the cyanohydrin of 3-phenoxybenzaldehyde. The compound is marketed with the defined stereochemistry at the valine moiety, but the cyanohydrin subunit is a *R/S* mixture.

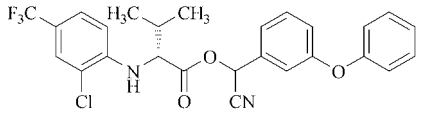

Imiprothrin, Pralle [*72693-72-5*], $C_{17}H_{22}N_2O_4$, M_r 318.4, has been introduced by Sumitomo as a household insecticide against cockroaches and other crawling insects, and is a pyrethroid based on glycine hydantoin. It is sold as a mixture of *cis*- and *trans*-isomers [318]. The long synthetic pathway leads to the vinyl substituted cyclopropane carboxylate, which forms an ester with 3-hydroxymethyl hydantoin. The resulting substance is *N*-alkylated with 3-bromopropyne.

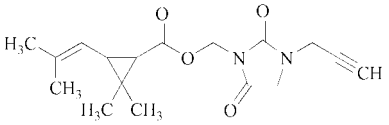

5.5.4. Plant Growth Regulators (→ Plant Growth Regulators)

Two compounds with amino acid structure are highlighted because of their direct chemical relationship to Glyphosate.

Glyphosine [*2439-99-8*] [344], Polaris, $C_4H_{11}NO_8P_2$ was introduced in 1973 and is preferably used in sugar beets. The synthesis starting with glycine uses two equivalents of formaldehyde and of phosphorous trichloride under oxidative conditions.

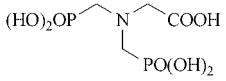

A second derivative of Glyphosate is a cyclic phosphaoxazole being formed from Glyphosate and formaldehyde. This compound is claimed to control the growth of some grasses without discoloration or phytotoxicity.

Aviglycine [*49669-74-1*], aminoethoxyvinylglycine, ReTain, $C_6H_{12}N_2O_3$, M_r 160.2, is produced by *Streptomyces* spp. and was isolated and purified by Abbott. It is marketed by Valent Biosciences for control of fruit ripening as it inhibits the production of endogenous ethylene [318].

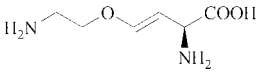

5.6. Industrial Uses

Polyfunctional compounds, such as amino acids and their derivatives, have been applied in a

wide variety of industrial areas. This is particularly because of their physicochemical properties, such as high thermal stability, low volatility, amphoterism, buffering capacity, and the ability to form complexes. However, properties of amino acids such as environmental acceptability and low toxicity, are becoming increasingly important. A few of the recorded uses have been listed below, and include acylamino acid monomers for epoxy resins [345]; amino acid dispersing agents for pigments in coloring polyester fibers [346]; *N*-acylamino acid dispersants for polyurethanes in water [347]; amino acid setting retarders for cement [348]; zinc salts of *N*-acyl-derivatives of basic amino acids and *N*-acylamino acids for the thermal stabilization of PVC [349]; polyglutamic acid esters and polyaspartic acid esters coatings for natural and synthetic leather [350]; amino acid hardening agents for methacrylate resins [351]; *N*-acylamino acid, amino acid ester, and amino acid amide gel-forming agents for oils [352]; basic amino acid vulcanization accelerators for natural rubber [353]; amino acid [354] and *N*-acylamino acid [355] corrosion inhibitors for metals; amino acids to stabilize the latent image of photographic emulsions [356]; and amino acid brighteners in galvanic baths [357, 358].

6. Chemical Analysis

Amino acids do not have defined melting points but decompose over a broad range between 250 and 300 °C. Therefore, they must be transformed into derivatives before melting points are useful for identification. Phenylisothiocyanate is used to yield the phenylthiohydantoin amino acid (PTH amino acid) [8], or 2,4-dinitrofluorobenzene (Sanger's reagent) is used to yield the dinitrophenyl amino acid [8].

Spectroscopic methods for the identification of amino acids include infrared [359], Raman, ^{1}H-NMR, and ^{13}C-NMR spectroscopy [360] of free amino acids or PTH derivatives and mass spectrometry of PTH derivatives [361]. Ultraviolet spectroscopy is important only for aromatic amino acids. The different methods for qualitative and quantitative analysis of amino acids, particularly of mixtures from protein hydrolysis and biological fluids, have been reviewed [362]. Separation methods are described in [363].

Various *staining methods* may be used for the qualitative identification of α-amino acids [364]. Some of these dye-forming reactions are suitable for quantitative analysis within the validity range of the Lambert–Beer law. By far the most important is the reaction with ninhydrin, which yields a red-violet to blue-violet dye ($\lambda_{max} = 570$ nm).

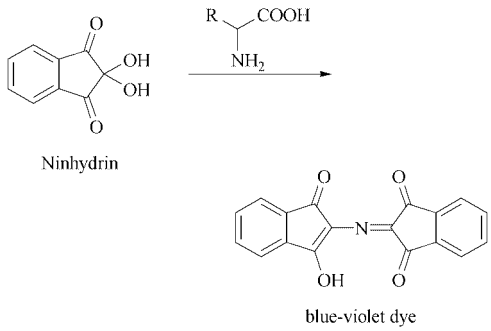

The imino acids proline and hydroxyproline form a structurally different dye, with an absorption maximum at 440 nm.

Fluorescent reagents also have been used successfully for quantitative analysis. The amino acid is converted into a strongly fluorescent derivative, which increases the sensitivity by orders of magnitude. Typical fluorescent reagents are *o*-phthalaldehyde–2-mercaptoethanol [365], 1-dimethylaminonaphthalene-5-sulfonyl chloride (dansyl chloride) [366], and 4-phenylspiro [furan-2(3*H*)-1′-phthalan]-3,3′-dione (fluorescamine) [367].

The *separation* of amino acid mixtures is possible by electrophoresis or chromatography. The latter is especially useful, techniques including paper, thin layer, ion-exchange, high-pressure liquid, and gas chromatography. The different techniques have been reviewed and compared in [368]. Paper chromatography is generally carried out in two dimensions. A number of eluents are available for this purpose. Quite often the amino acid is converted to its dinitrophenyl derivative.

More common than paper chromatography is *thin layer chromatography* (TLC) of either the free amino acids [369] or their PTH derivatives [370]. Silica gel, aluminum oxide, cellulose powder, or polyacrylamide may be used as carrier, but silica is preferred. An indicator reagent, most often ninhydrin, is used for detection [371]. For detection of very small quantities the amino

acids separated on an TLC plate can be converted to the fluorescamine derivatives [372]. The detection limit is 10 pmol. The time required for the analysis of dinitrophenyl amino acids can be reduced by using high-pressure thin layer chromatography [373]. TLC is often specified as an analytical test to show the levels of foreign amino acids in a sample, particularly when the amino acid is from natural sources.

Ion-exchange chromatography [374, 375] on organic resins (Dowex, Amberlite, etc.) has proven to be the most exact and reliable method for the separation and quantitative analysis of amino acids. Before the automatic analysis technique was introduced [376], complete analysis of an amino acid mixture required 24 h. Today 2 h is the rule. Sodium citrate or lithium citrate buffer solutions are the eluents. The eluate is reacted with ninhydrin [374, 377] or *o*-phthalaldehyde [378]. With ninhydrin, 1–20 nmol amino acid can be measured with an accuracy of $\pm$ 1–5 %. The detection limit with *o*-phthalaldehyde is in the picomole range. Ion-pair chromatography on a porous graphitic carbon stationary phase has also been applied. Detection is by evaporative light scattering, or by electrospray mass spectrometry [379].

Ion-exchange chromatography is the method of choice for analyzing amino acids in feeds, foodstuffs, and biologic fluids [380]. In general, analysis is preceded by a hydrolysis, which degrades the proteins and peptides to their component amino acids. However, acidic hydrolysis can destroy Cys, Met, and Trp residues, and special techniques may be necessary to analyze for these amino acids. Figure 10 shows a sample aminogram of broiler feed.

Utilization of *high-pressure liquid chromatography* (HPLC) allows a further reduction in analysis time. Originally, the amino acids were converted to derivatives, e.g., to dansyl amino acids [381], PTH amino acids [382], or dinitrophenyl derivatives [383], before analysis. Reversed phases are the preferred stationary phases. Ninhydrin, *o*-phthalaldehyde–mercaptoethanol [384], or fluorescamine [385] are the usual reagents for detection. An HPLC method with direct UV detection has also been described [386] for analysis of infusion solutions. Modern automated amino acid analyzers are now available, that use quaternary gradient elution to separate all the common amino acids efficiently without prior derivatization. These machines generally incorporate postcolumn visualization with ninhydrin and photometric quantification. Today this method rivals ion-exchange chromatography as the preferred analytical method.

Gas chromatographic methods [387] are useful for the analysis of complex amino acid mixtures. However, the amino acids must be converted into volatile derivatives, e.g., PTH amino acids [388], methyl esters of *N*-trifluoroacetylamino acids [389], *n*-butyl esters of *N*-trifluoroacetylamino acids [390], *N*-trimethylsilylamino acids [391], or *N,O*-bis-trimethylsilylamino acids [392]. Gas chromatography coupled with electron-capture negative-ionization mass spectrometry (GC–ECNI–MS) has been used to determine amino acids at the femtomole scale, using ^{2}H or ^{13}C labeled amino acids as reference standards [393].

Electrophoresis [394], which employs the differing rates of migration in an electric field, is becoming more important [395]. The technique is already well-established for protein analysis, and has been coupled with electrospray ionization mass spectrometry (CE–ESI–MS) to determine underivatized amino acids at micromolar level [396]. Capillary isotachophoresis is a new high-resolution electrophoresis technique for amino acids.

The chromatographic separation of amino acid enantiomers is the subject of intensive investigation. Separation is currently possible by gas chromatography [397] and high-pressure liquid chromatography [398], using optically active phases or chiral solvents [399]. A wide range of chiral stationary phases is now available for analytical or preparative separations. Another method is to use a chiral precolumn derivatizing agent. Marfey's reagent (1-fluoro-2,4-dinitrophenyl-5-L-alanine amide) has been successfully applied to separate complex mixtures of amino acids into their separate enantiomers, and is effective at the nanomolar scale [400, 401]. Marfey's reagent forms diastereomeric salt pairs with the amino acids, that are particularly well separated by HPLC.

The *microbiological analysis* of amino acids is based on the fact that several L-amino acids are essential for certain bacteria strains. The growth of the bacteria cultures under standard conditions can be quantitatively evaluated (acidime-

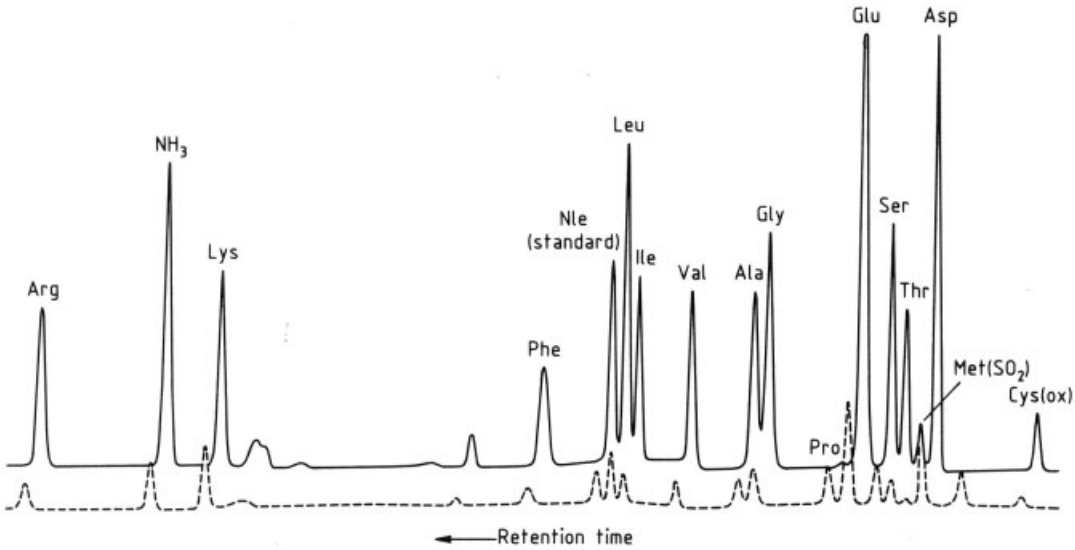

Figure 10. Amino acid chromatogram of a broiler feed
Internal standard: norleucine. Solid curve: UV detection at $\lambda = 570$ nm. Dotted curve: UV detection at $\lambda = 440$ nm for Pro (and Hyp).

try or turbidimetry) and related to the amino acid concentration. Lactic acid bacteria [402] (Lactobacteriaceae) can be used to analyze 19 L-amino acids. Typical test microbes include *Leuconostoc mesenteroides* (ATCC 8042), *Lactobacillus arabinosus 17–5* (ATCC 8014), and *Streptococcus faecalis R* (ATCC 9790, 8043). The conventional microbiological methods are quite complicated but can be simplified by automation [403].

A series of L- or D-amino acids can be analyzed by *enzymatic methods.* L-Amino acids, or the enantiomeric purity of D-amino acids, can be determined with bacteria decarboxylases by measurement of CO_2 formed. D-Amino acids, or the enantiomeric purity of L-amino acids, can be determined with kidney D-amino acid oxidases by measurement of the O_2 consumption. Enzymes that react only with a single amino acid allow determination of that amino acid, e.g., arginine using liver arginase. An improved enzymatic method consists of the use of enzyme electrodes that contain the enzymes [404], or microorganisms having a special enzyme in a fixed form. Enzyme electrodes, however, are relatively unstable [405].

The *quantitative determination of crystallized amino acids* is carried out by acidimetric titration in nonaqueous medium [406, 407]. Glacial acetic acid is a suitable solvent. Formic acid may be added to improve solubility. The titrant is perchloric acid. Formol titration by the method of Sörensen can be used for aqueous solutions but is less accurate.

Standards of purity for individual L- and D,L-amino acids that are used in drugs or as food additives are published in pharmacopeias [406, 408] and food codices [409].

7. Economic Significance

The 2005 world market for amino acids is estimated at more than 3.3×10^6 t/a (Table 19). The "big three" amino acids, sodium L-glutamate, D,L-methionine, and L-lysine · HCl, account for approximately 95 % of the volume (Table 20). The other amino acids play only a small role. The dominant amino acid, sodium glutamate, is used almost exclusively as a taste enhancer. D,L-Methionine and L-lysine · HCl are used almost exclusively to improve the nutritive value of animal feeds. L-Threonine is also used mainly in animal feeds. L-Aspartic acid and L-phenylalanine are used principally for the manufacture of the intense sweetener aspartame. The other amino acids have diversified applications. With the exception of glycine, they are more expensive than the big three amino acids. In terms of

volume, feed additives used about 1 500 000 t of amino acids, pharmaceuticals 17 000 t and sweeteners around 17 000 t (2005) [410]. Table 21 shows the market value for amino acids in 2004 broken down by field of use [52]. The total value of the global amino acids market in 2004 was estimated at \$ 4.5×10^9 [52].

Table 19. Amino acid production, 2005

Amino acid	Quantity, t/a [410–413]
L-Alanine	1500
L-Arginine *	3000
L-Asparagine	200
L-Aspartic acid	15 000
L-Cysteine *	7000
L-Glutamic acid *	1 690 000
L-Glutamine	2000
Glycine	16 000
L-Histidine	300
L-Hydroxyproline	100
L-Isoleucine	500
L-Leucine	800
L-Lysine *	850 000
D,L-Methionine	600 000
L-Methionine	400
L-Phenylalanine	14 000
L-Proline	800
L-Serine	300
L-Threonine	85 000
L-Tryptophan	2000
L-Tyrosine	150
L-Valine	1100
D-Phenylglycine + D-*p*-hydroxyphenylglycine	9000
Others	5000
Total	3 300 000

* Free amino acid and salts.
* Free amino acid and derivatives.

Table 20. Percentage of individual amino acids as a part of the total market, 2005

Amino acid	Quantity, %
L-Glutamic acid (Na)	51
D,L-Methionine	18
L-Lysine (HCl)	26
Other amino acids	5

Table 21. Market value by field of use, 2004

Use	Value, %
Animal nutrition	56
Human nutrition	32
Specialty *	12

* Pharmaceuticals, cosmetics, agrochemicals, and industrial uses.

The main manufacturers of amino acids are located in Japan (e.g., Ajinomoto, Kyowa Hakko, Tanabe) and in Europe (Degussa, Wacker, Amino, Flamma). Recently there has been an increasing trend towards production in China, with established manufacturers setting up plants there.

8. Toxicology

Excess amino acids are rapidly disposed of by increased metabolic degradation. Should the amino acid dose be suddenly increased, e.g., by extremely high protein consumption, within about two days the liver adaptively increases the levels of amino acid-catabolizing enzymes — transaminases, enzymes of the urea cycle, cystathionase, tryptophan pyrrolase, etc. The excess amino acids are to a large extent used to provide energy. The nitrogen is eliminated as urea. A smaller portion is used in protein synthesis, mainly liver protein and plasma albumin.

When too little protein or no protein is consumed or when the component amino acids are imbalanced, alteration of the ribosome profile occurs in the liver, and ribonucleic acids are catabolized. The manifestation of chronic protein deficiency is known as marasmus (slight deficiency) and kwashiorkor (extreme deficiency). Protein deficiency is usually coupled with calorie deficiency (protein–calorie malnutrition). This manifests itself on the biochemical level as a negative nitrogen balance, indicating a reduction in protein inventory. Initially, the labile enzyme and plasma proteins are consumed, the greatest losses occurring first in the liver, then in the musculature. Brain, heart, and kidneys suffer only minimal protein loss. Symptoms of protein synthesis disorders include disturbances in wound healing and bone growth, lowered resistance to infection and stress, loss of fertility and appetite, and anorexia. The urinary excretion of 3-methylhistidine is a common indicator of the catabolism of muscle protein.

The total absence of an essential amino acid in the diet is more serious than protein deficiency. In this case the proteins or amino acids in the diet are totally worthless because protein synthesis can occur only by degradation of body protein. A general interruption of the protein synthesis results. This manifests itself rapidly as a drop in enzyme activity and an impoverishment of the plasma proteins. Noticeable symptoms are loss of appetite and weight, alteration of cornea and lens, anatomic organ alterations, and an increased rate of mortality. In addition

there appear specific deficiency symptoms characteristic of the missing amino acid or acids.

The metabolic disturbances brought about by gross divergences from the optimal amino acid pattern have three different causes [414, 415]: imbalance, antagonism, and toxicity.

Amino acid imbalance manifests as an appearance of deficiency symptoms for the first limiting amino acid when the other amino acids are consumed in great excess. The symptoms of imbalance are eliminated by administration of the first limiting amino acid in sufficient quantities. The main symptom is a severe reduction of food or feed assimilation and depression of growth. The depression of growth in rats has been investigated intensively by adding individual L- and D,L-amino acids to basal diets of various protein levels [416, 417].

*Amino acid antagonism*is caused by competition for common transport systems. An example is antagonism of the branched-chain amino acids isoleucine, leucine, and valine. The symptoms are reversible. In a study with young rats, addition of 5 % L-leucine to a low-protein diet (9 % casein) reduced the plasma levels of isoleucine and valine, depressed growth, and reduced feed consumption [418]. These effects were eliminated after a latent period of three days by small doses of L-isoleucine (0.16 %) and L-valine (0.15 %).

Amino acid toxicity occurs when very large quantities of one or more amino acids are consumed and is characterized by total failure of the adaptive mechanisms. The toxic level has been studied by adding increasing quantities of individual amino acids to a protein basal diet [415]. The toxicity of individual amino acids depends on the total protein consumption. Imbalance, antagonism, and toxicity are less pronounced when overall protein consumption is sufficient but become more severe at lower levels of protein consumption. The consumption of toxic amounts of amino acids increases their concentration in the plasma and brain. Because of the blood–brain barrier, however, the increase in the brain is not as great [419]. Failure of the adaptation mechanisms during consumption of large excess of an amino acid can lead to accumulation of the amino acids or certain metabolites in the organism, leading directly or indirectly (e.g., by influencing hormone secretion) to anatomic or functional damage.

The toxicity of amino acids has been reviewed [415, 418]. The acute toxicities of most L-amino acids and some derivatives have been determined [420–422]. The toxicology of D-amino acids is discussed in review articles [415] and other publications [421, 422]. There is no evidence to date that the D-enantiomers of the α-amino acids found in proteins exhibit specific toxic effects. Their LD_{50} values are generally higher than those of the L-amino acids.

9. References

General References

1. J. P. Greenstein, M. Winitz: *Chemistry of the Amino Acids*, vol. **1, 2, and 3**, J. Wiley & Sons, New York–London 1961.
2. A. Meister: *Biochemistry of the Amino Acids*, 2nd ed., vol. **1**, Academic Press, New York 1965.
3. I. Wagner, H. Musso: "Neue natürliche Aminosäuren" *Angew. Chem.* **95** (1983), 827–920;
   Angew. Chem. Int. Ed. Eng. **22** (1983) 816 (review of the literature since 1956).
4. IUPAC and IUB Joint Commission on Biochemical Nomenclature, *Pure Appl. Chem.*, **56** (1984) 595.
5. G. C. Barrett (ed.): *Amino Acid Derivatives*, Oxford University Press, Oxford 1999.
6. *Houben-Weyl*, **E16d**, 1992.
7. S. E. Wolkenberg, R. M. Garbaccio, *Science of Synthesis* **20a** (2006) 385.
8. Th. Wieland et al.: "Methoden zur Herstellung und Umwandlung von Aminosäuren und Derivaten," *Houben-Weyl*, 11/2.
9. H. D. Jakubke, H. Jeschkeit: *Aminosäuren, Peptide, Proteine*, Verlag Chemie, Weinheim, Germany 1982.
10. B. Weinstein (ed.): *Chemistry and Biochemistry of Amino Acids, Peptides and Proteins*, vol. **1–5**, Marcel Dekker, New York 1971 –1978.
11. D. Barton, W. D. Ollis: *Comprehensive Organic Chemistry*, vol. **2, Chap. 9.6, p. 815; vol. 5** Pergamon Press, New York 1979.
12. G. Krügers: "Aminocarbonsäuren" in: *Methodicum Chimicum*, vol. **6**, p. 611.
13. E. Wünsch: "Synthese von Peptiden," *Houben-Weyl*, **15/1 and 15/2.**

Specific References

14. S. K. Bhattacharyya, A. B. Banerjee, *Folia Microbiol. (Prague)* **19** (1974) 43.

15. T. Robinson, *Life Sci.* **19** (1976) 1097.
16. R. E. Dickerson, *Spektrum der Wissenschaften* 1979, no. 9, 98.
17. R. S. Cahn, C. Ingold, V. Prelog, *Angew. Chem.* **78** (1966) 413; *Angew. Chem. Int. Ed. Engl.* **5** (1966) 385, 511.
18. K. Lübke, E. Schröder, G. Kloss: *Chemie und Biochemie der Aminosäuren, Peptide und Proteine*, vol. **I and II**, Thieme Verlag, Stuttgart 1975.
19. A. Kleemann, *Chem.* Ztg. 106 (1982) 151–167.
20. K. H. Drauz, A. Kleemann, J. Martens, *Angew. Chem.* **94** (1982) 590; *Angew. Chem. Int. Ed. Engl.* **21** (1982) 584.
21. A. Berkessel, H.w Gröger: *Asymmetric Organocatalysis*, Wiley-VCH, Weinheim 2005.
22. *Drugs of the Future* **23** (1998) 1057.
23. R. P. Nargund, A. A. Patchett, M. A. Bach, M. G. Murphy, R. G. Smith, *J. Med. Chem.* **41** (1998) 3103.
24. *Drugs of the Future* **19** (1994) 111.
25. H. J. Federsel, *Chirality* **15** (2003) 128.
26. A. Müller, E. Busker, J. Engel, B. Kutscher, M. Bernd, A. V. Schally, *Int. J. Peptide Prot. Res.* **43** (1994) 264.
27. Europ. Pat. Appl. EP 0 530 167 A 1, 1993 (B. Atrash, D. M. Jones, M. Szelke).
28. *Drugs of the Future* **22** (1997) 371.
29. *Drugs of the Future* **17** (1992) 551.
30. *Drugs of the Future* **23** (1998) 751.
31. J. A. Robl et al., *J. Med. Chem.* **40** (1997) 1570.
32. S. Budavari (ed.): *The Merck Index*, 12th ed., Merck & Co., Inc., Whitehouse Station, NJ 1996, p. 97.
33. A. S. Bommarius, M. Schwarm, K. Stingl, M. Kottenhahn, K. Huthmacher, K. Drauz, *Tetrahedron: Asymmetry* **6** (1995) 2851.
34. *Drugs of the Future* **21** (1996) 1022.
35. *Drugs of the Future* **23** (1998) 1057.
36. *Drugs of the Future* **19** (1994) 985.
37. J. F. Atkins, R. Gesteland, *Science* **296** (2002) 1409.
38. *Drugs of the Future* **18** (1993) 503.
39. *Drugs of the Future* **23** (1998) 1057.
40. *Drugs of the Future* **22** (1997) 260.
41. *Drugs of the Future* **22** (1997) 371.
42. S. Budavari (ed.): *The Merck Index*, 12th ed., Merck & Co., Inc., Whitehouse Station, NJ 1996, p. 99.
43. *Drugs of the Future* **14** (1989) 767.
44. *Drugs of the Future* **21** (1996) 600.
45. *Drugs of the Future* **18** (1993) 1129.
46. *Drugs of the Future* **22** (1997) 221.
47. *Drugs of the Future* **23** (1998) 1057.
48. V. N. Gladyshev, D. L. Hatfield, *J. Biomed. Sci.* **6** (1999) 151.
49. G. N. Schrauzer, *J. Nutrition* **130** (2000) 1653.
50. *Drugs of the Future* **19** (1994) 988.
51. R. Meister (ed.): *Farm Chemical Handbook 1990*, Meister Publishing Co., Willoughby, OH 1990, p. C136.
52. W. Leuchtenberger, K. Huthmacher, K. Drauz, *Appl. Microbiol. Biotech.* **69** (2005) 1.
53. T. Hermann, *J. Biotechnol.* **104** (2003) 155.
54. M. Breuer et al., *Angew. Chem. Int. Ed.* **43** (2004) 788.
55. J. Jones: *Amino Acid and Peptide Synthesis*, Oxford University Press, Oxford 2002.
56. M. Ikeda, in T. Scheper, R. Faurie, J. Thommel (eds.): *Advances in Biochemical Engineering/Biotechnology* **79** (2003) 1.
57. S. Hashimoto, R. Katsumata, L-Alanine production by alanine racemase-deficient mutant of Arthrobacter oxydans in: *Proc. Ann. Meet. (Agric. Chem. Soc. Jpn.)* (1994) 341.
58. S. Yahata, H. Tsutsui, K. Yamada, T. Konehara: "Fermentative production of D-alanine:" in: Proc. Ann. Meet. Agric. Chem. Soc. Jpn., (1993), 92.
59. Tanabe Seiyaku, JP-Kokai 58-9692, 1983 (I. Chibata, M. Kisumi, T. Tagaki).
60. Ajinomoto, US 5 164 307, 1992 (Y. Yoshihara, Y. Kawahara, Y. Yamada).
61. M. Sugiura, S. Suzuki, M. Kisumi, *Agric. Biol. Chem.* **51** (1987) 371–377.
62. Kyowa Hakko, EP-A 557 996, 1993 (K. Kino, K. Okamoto, Y. Takeda, Y. Kuratsu).
63. C. Keilhauer, L. Eggeling, H. Sahm, *J. Bacteriol.* **175** (1993) 5595–5603.
64. Cheil Sugar, FR 2 645 172, 1990 (J. W. Oh, S. J. Kim, Y. J. Cho, N. H. Park, L. H. Lee).
65. University Odessa, SU 1 730 152, 1992 (I. I. Brown et al.).
66. Sci. Res. Technol Inst. Amino Acids, USSR, DE-OS 3 612 077, 1986 (Š. M. Kocarian et al.).
67. All Union Sci.-Res. Inst. Genetics and Microorg. Gen. + Selec Univ. Odessa, WO 90/04636, 1990 (V. G. Debabov et al.).
68. M. Ikeda, K. Nakanishi, K. Kino, R. Katsumata, *Biosci. Biotech. Biochem.* **58** (1994) 674–678.
69. M. Furui, K. Yamashita, *J. Ferment, Technol.* **61** (1983) 587–591.
70. H. Yoshida: "Arginine, Citrulline, and Ornithine," in K. Aida et al. (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 131–143.

71. I. Chibata, T. Tosa, T. Sato: "Aspartic Acid," in H. W. Blanch, S. Drew, D. I. C. Wang (eds.): *Comprehensive Biotechnology*, vol. **3** , Pergamon Press, Oxford 1985.
72. Y. Yokote et al., *J. Solid-Phase Biochem.* **3** (1978) 247–261.
73. M. Terasawa, H. Yukawa, Y. Takayama, *Process Biochem.* **20** (1985) 124–128.
74. T. Sato et al., *Biotechnol. Bioeng.* **17** (1975) 1779–1804.
75. I. Chibata, T. Tetsuya, T. Sato, *Appl. Biochem. Biotechnol.* **13** (1986) 231–240.
76. M. C. Fusee, W. E. Swann, G. J. Calton, *Appl. Environ. Microbiol.* **42** (1981) 672–676.
77. Genex, EP-A 197 784, 1986 (W. E. Swann, A. C. Nolf).
78. J. Michelet, A. Deschamps, J. M. Lebeault, *3rd Eur. Congr. Biotechnol.*, vol. **2**, Verlag Chemie, Weinheim, Germany 1984, pp. 133–138.
79. M. Mori, I. Shiio, *Agric. Biol. Chem.* **48** (1984) 1189–1197.
80. N. Nishimura, S. Komatsubara, M. Kisumi, *Appl. Environ. Microbiol.* **53** (1987) 2800–2803.
81. Maggi, DE 2 449 711, 1979 (P. Hirsbrunne, R. Bertholet).
82. Mitsubishi Chemical Ind., US 3974031, 1976 (H. Yamada, K. Gumagai, H. Ohkishi).
83. Consortium für Elektrochemische Industrie GmbH, WO 01/27307, 2001 (T. Maier, C. Winterhalter).
84. T. H. P Maier, *Nat. Biotechnol*. **21** (2003) 422.
85. K. Sano, K. Mitsugi, *Agric. Biol. Chem.* **42** (1978) 2315–2321.
86. S. Kinoshita, S. Ukada, M. Shimono, *J. Gen. Appl. Microbiol.* **3** (1957) 139–205.
87. S. Kinoshita, K. Tanaka: "Glutamic acid," in K. Yamada, et al. (eds.): *The Microbial Production of Amino Acids*, John Wiley & Sons, New York 1972, pp. 263–324.
88. L. Eggeling: *Handbook of Corynebacterium Glutamicum*, CRC Press, Boca Raton, FL 2005.
89. B. J. Eikmans, M. Kircher, D. J. Reinscheid, *FEMS Microbiol. Lett.* **82** (1991) 203–208.
90. W. Liebl, M. Ehrmann, W. Ludwig, K. H. Schleifer, *Int. J. Syst. Bacteriol.* **41** (1991) 225–235.
91. Y. Hirose, H. Enei, H. Shibai: "L-Glutamic acid fermentation," in H. W. Blanch, S. Drew, D. I. C. Wang (eds.): *Comprehensive Biotechnology*, vol. **3**, Pergamon Press Ltd., Oxford 1985.
92. M. Kikuchi, Y. Nakao: "Glutamic acid," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 101–116.
93. R. Krämer, *BioEngineering* **9** (1993) 51–61.
94. R. Krämer, *FEMS Microbiol. Rev.* **13** (1994) 75–93.
95. I. Shiio, K. Ujigawa, *Agric. Biol. Chem.* **42** (1980) 1897–1904.
96. E. R. Börmann, B. J. Eikmanns, H. Sahm, *Molecular Microbiol.* **6** (1992) 317–326.
97. T. E. Walker et al., *J. Biol. Chem.* **257** (1982) 1189–1195.
98. H. Yoishii et al., *Nippon Nogei Kagaku Kaishi* **67** (1993) 949–954.
99. H. Yoishii, M. Yoshimura, S. Nakamori, S, Inoue, *Nippon Nogei Kagaku Kaishi* **67** (1993) 955–960.
100. Ajinomoto, JP 56 148 295, 1981 (T. Tsuchida, K. Miwa, S. Nakamori, H. Mimose).
101. M. Shirasuchi et al., *Biosci. Biotechnol. Biochem.* **59** (1995) 83–86.
102. T. Tsuchida et al., *Agric. Biol. Chem.* **51** (1987) 2089–2094.
103. K. Araki: "Histidine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 247–256.
104. M. B. Vickery, *J. Biol. Chem.* **143** (1942) 77.
105. Kyowa Hakko, EP-A 759 472, 1996 (A. Ozaki, H. Mori, T. Shibasaki).
106. J. Ogawa, S. Shimizu, *Curr. Opin. Biotechnol*. **13** (2002) 367.
107. I. Eggeling, C. Cordes, L. Eggeling, H. Sahm, *Appl. Microbiol. Biotechnol.* **25** (1987) 346–351.
108. E. Scheer, L. Eggeling, H. Sahm, *Appl. Microbiol. Biotechnol.* **28** (1988) 474–477.
109. C. Wilhelm et al., *Appl. Microbiol. Biotechnol.* **31** (1989) 458–462.
110. Ajinomoto, Patent 5 164 307, 1992 (Y. Yoshihara, Y. Kawahara, Y. Yamada).
111. S. Komatsubara, M. Kisumi, I. Chibata, *J. Gen. Microbiol.* **119** (1980) 51–61.
112. Ajinomoto, EP-A 519 113, 1992 (V. A. Livshits et al.).
113. Kyowa Hakko, EP-A 595 163, 1994 (T. Nakano, T. Azuma, Y. Kuratsu).
114. Forsch. Z. Jülich, DE 44 00 926, 1995 (B. Möckel, L. Eggeling, H. Sahm).
115. Kyowa Hakko, US 5 362 637, 1994 (K. Kino, Y. Kuratsu).
116. S. Morbach, H. Sahm, L. Eggeling, *Appl. Environ. Microbiol.* **61** (1995) 4315–4320.

117. Ajinomoto, EP-A 685 555, 1995 (K. Hashiguchi, H. Kishino, N. Tsujimoto, H. Matsui).
118. U. Groeger, H. Sahm, *Appl. Microbiol. Biotechnol.* **25** (1987) 352–236.
119. S. Komatsubara, M. Kisumi: "Histidine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amio Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 233–246.
120. Ajinomoto, US 5 164 307, 1992 (Y. Yoshihara, Y. Kawahara, Y. Yamada).
121. M. Karasawa, O. Tosaka, S. Ikeda, H. Yoshi, *Agric. Biol. Chem.* **50** (1986) 339–346.
122. Kyowa Hakko, US 4 623 623, 1986 (T. Nakanashi et al.).
123. Kawahara et al., *Appl. Microbiol. Biotechnol.* **34** (1990) 87–90.
124. Ajinomoto, US 5 250 423, 1993 (Y. Murakami, H. Miwa, S. Nakamori).
125. A. Erdmann, B. Weil, R. Krämer, *Appl. Microbiol. Biotechnol.* **42** (1994) 604–610.
126. A. Erdmann, B. Weil, R. Krämer, *Biotechnol. Lett.* **17** (1995) 927–932.
127. R. Krämer, *J. Biotechnol.* **45** (1996) 1–21.
128. M. S. N. Jetten, A. J. Sinskey, *Crit. Rev. in Biol.* **15** (1995) 73–103.
129. Kyowa Hakko, EP-A 197 335, 1986 (R. Katsumata, T. Mitzukami, T. Oka).
130. Forsch. Z. Jülich, EP-A 435 132, 1991 (J. Cremer, L. Eggeling, H. Sahm).
131. Y.-C. Liu, W.-T. Wu, J.-H. Tsao, *Bioprocess. Eng.* **9** (1993) 135–139.
132. Eurolysine, EP-A 534 865, 1993 (P. Lucq, C. Domont).
133. Rhone-Polenc, EP 122 163, 1984 (N. Rouy).
134. Degussa, EP 533 039, 1993 (W. Binder et al.).
135. Degussa, DE 2 421 167, 1974 (T. Lüßling, K. Müller, G. Schreyer, F. Theissen).
136. W. Leuchtenberger, U. Plöcker: "Amino acids and hydroxycarboxylic acids," in W. Gerhartz (ed.): *Enzymes in Industry*, Production and Applications, VCH, Weinheim, Germany 1990.
137. T. Ishikawa et al., *Biosci. Biotech. Biochem.* **57** (1993) 982–986.
138. Genencor, WO 93/17112, 1993 (J. C. Lievense).
139. Genex, US 4 598 047, 1986 (J. C. McGuire).
140. C. T. Evans, C. Coma, W. Petreson, M. Misawa, *Biotechnol. Bioeng.* **30** (1987) 1067–1072.
141. G. J. Calton et al., *Biotechnology* **4** (1986) 317–320.
142. C. T. Evans et al., *Biotechnology* **5** (1987) 818–923.
143. A. J. Pittard, *Amer. Soc. Microbiol. Conf. Proc.* (1987) 368–394.
144. I. Shiio: "Tryptophan, phenylalanine, and tyrosine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 188–206.
145. S. O. Hwang et al., *Appl. Microbiol. Biotechnol.* **22** (1985) 108–113.
146. H. Ito et al., *Agric. Biol. Chem.* **54** (1990) 707–713.
147. B. K. Konstantinov, N. Nishio, T. Yoshida: "Glucose feeding strategy accounting for the decreasing oxidative capacity of recombinant Escherichia coli in fed-batch cultivation for phenylalanine production," *J. Ferment. Bioeng.* **70** (1990) 253.
148. Ajinomoto, US 4 584 400, 1986 (M. Ootani, S. Sano, I. Kusumotu).
149. A. M. Vasconcellos, A. L. Neto, D. M. Grassiano, C. P. De Oliveira, *Biotechnol. Bioeng.* **33** (1989) 1324–1329.
150. Ajinomoto, US 5 362 635, 1994 (T. Hirose et al.).
151. F. Yoshinaga: "Proline," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 117–120.
152. M. Sugiura, T. Takagi, M. Kisumi, Abst. 31st Symp. Amino Acid and Nucleic Acid (Japan), 1982, p. 10.
153. Y. Imai, T. Takagi, M. Sugiura, M. Kisumi, Abst. Ann. Meet Agr. Chem. Soc. (Japan), 1984, p. 100.
154. Y. Izumi et al., *Appl. Microbiol. Biotechnol.* **39** (1993) 427–432.
155. T. Yoshida, T. Mitsunaga, Y. Izumi, *J. Fermentation Bioeng.* **75** (1993) 405–408.
156. H.-Y. Hsiao, T. Wie, *Biotechnol. Bioeng.* **28** (1986) 1510–1518.
157. Mitsui Toatsu, US 5 382 517, 1995 (D. Ura, T. Hashimukai, T. Matsumoto, N. Fukuhara).
158. K. Shimura: "Threonine," in H. W. Blanch, S. Drew, D. I. C. Wang (eds.): *Comprehensive Biotechnology*, vol. **3**, Pergamon Press, Oxford 1985.
159. S. Nakamori: "Threonine and Homoserine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Progress in Industrial Microbiology*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 173–182.

160. Genetics Ind., US 4 278 765, 1981 (V. G. Debabov et al.).
161. Ajinomoto, EP-A 593 792, 1994 (V. G. Debabov et al.).
162. Eurolysine, FR 2 627 508, 1989 (F. Richaud et al.).
163. Ajinomoto, FR 2 588 016, 1987 (M. Ootani, T. Kitahara, K. Akashi).
164. A. Yokota, T. Takao, *Agric. Biol. Chem.* **48** (1984) 2663–2668.
165. Mitsui Toatsu, EP-A 341 674, 1989 (K. Ishiwata et al.).
166. Mitsui Toatsu, EP-A 438 591, 1991 (S. Ogawa, S. Iguichi, S. Morita, H. Kuwamoto).
167. B. K. Hamilton et al., *Trends Biotechnol.* **3** (1985) 64–68.
168. J. Plachy, S. Ulbert, *Acta Biotechnol.* **10** (1990) 517–522.
169. Showa Denko, JP 62 186 786, 1987 (E. Takinishi, H. Takamatsu, K. Sakimoto, Y. Yajima).
170. A. N. Mayeno, G. J. Gleich, *TIBTECH* **12** (1994) 346–352.
171. K. Sakimoto, Y. Torigoe, *Curr. Prospects Med. Drug Saf.* (1994) 295–311.
172. K. Nakayama: "Tryptophan," in M. Moo-Young (ed.): *Comprehensive Biotechnology*, vol. **3**, Pergamon Press, Oxford 1985, 621–631.
173. I. Shiio: "Tryptophan, Phenylalanine, and Tyrosine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds.): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 188–206.
174. T. K. Maiti, S. P. Chatterje, *Hindustan Antibiotics Bulletin* **33** (1991) 26–61.
175. M. Ikeda, *Appl. Microbiol. Biotech.* **69** (2006) 615.
176. S. Komatsubara, M. Kisumi: "Isoleucine, Valine, and Leucine," in K. Aida, I. Chibata, K. Nakayama, K. Takinami, H. Yamada (eds): *Biotechnology of Amino Acid Production*, Kodansha, Tokyo/Elsevier, Amsterdam 1986, pp. 131–143.
177. Ajinomoto, EP 287 123, 1993 (N. Katsurada, H. Uchibori, T. Tsuchida).
178. D. A. Bender: *Amino Acid Metabolism*, J. Wiley & Sons, London–New York–Sydney–Toronto 1975, p. 2.
179. H. N. Munro, *Drug Intell. Clin. Pharm.* **6** (1972) 216.
180. M. L. Sunde, *Poult. Sci.* **51** (1972) 44.
181. R. F. Barker, D. A. Hopkinson, *Ann. Hum. Genet.* **41** (1977) 27.
182. L. D. Stegink: *Clinical Nutrition Update: Amino Acids*, Am. Med. Assoc. Publ., Chicago 1977, pp. 198–205.
183. K. Schreier, *Ärztl. Fortbildung* **19** (1971) 107.
184. U. Porath, K. Schreier, *Med. Ernähr.* **11** (1970) 229.
185. H. Bickel, S. Kaiser-Grubel, *Dtsch. Med. Wochenschr.* **96** (1971) 1415.
186. Y. Izumi, I. Chibata, T. Itoh, *Angew. Chem.* **90** (1978) 187;
Angew. Chem. Int. Ed. Eng. **17** (1978) 176.
187. M. S. Sadovnikova, V. M. Belikov, *Russ. Chem. Rev. (Engl. Transl.)* **47** (1978) 199.
188. K. Drauz, A. Kleemann, J. Martens, *Angew. Chem.* **94** (1982) 590;
Angew. Chem. Int. Ed. Engl. **21** (1982) 584.
189. Winnacker-Küchler, 5th ed., vol. **8**, 715.
190. FAO: *Nutritional Studies No. 24, Amino Acid Content of Foods and Biological Data on Proteins*, Interprint (Malta), Rome 1970.
191. Degussa, unpublished
192. W. C. Rose, *Nutr. Abstr. Rev.* **27** (1957) 631.
193. D. M. Hegsted, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **22** (1963) 1424.
194. Report of a Joint FAO/WHO Ad Hoc Expert Committee: *Energy and Protein Requirements*, Rome 1973.
195. E. Kofrányi, *Ernähr. Umsch.* **23** (1976) 205.
196. Dtsch. Ges. für Ernährung: *Empfehlungen für die Nährstoffzufuhr*, 3rd ed., Umschau-Verlag, Frankfurt 1975.
197. Dtsch. Ges. Für Ernährung: *Referenzwerte für die Nährstoffzufuhr*, Umschau Buchverlag, Frankfurt 2000.
198. *Recommended Dietary Allowances*, 9th ed., Nat. Acad. of Sci., Washington, D.C., 1980.
199. J. J. Otten, J. P. Hellwig, L. D. Meyers (eds.): *Dietary Reference Intakes: the Essential Guide to Nutrient Requirements*, National Academies Press, Washington DC 2006.
200. G. K. Parman, *J. Agric. Food Chem.* **16** (1968) 169.
201. H. H. Ottenheym, P. J. Jenneskens, *J. Agric. Food Chem.* **18** (1970) 1010.
202. N. S. Scrimshaw, A. M. Altschul: *Amino Acid Fortification of Protein Foods*, MIT Press, Cambridge, MA 1971.
203. E. E. Howe, G. R. Jansen, E. W. Gilfilian, *Am. J. Clin. Nutr.* **16** (1965) 315.
204. A. M. Altschul, *Nature (London)* **248** (1974) 643.
205. J. Kato, N. Muramatsu, *J. Am. Oil Chem. Soc.* **48** (1971) 415.

206. J. Mauron, *Z. Ernährungswiss. Suppl.* **23** (1979) 10.
207. D. M. Hegsted, *Am. J. Clin. Nutr.* **21** (1968) 688.
208. L. D. Satterlee, H. F. Marshall, J. M. Tennyson, *J. Am. Oil Chem. Soc.* **56** (1979) 103.
209. H. W. Staub, *Food Technol. (Chicago)* **32** (1978) 57.
210. F. Jekat, *Fette Seifen Anstrichm.* **79** (1977) 273.
211. J. C. Somogyi in: *Die Bedeutung der Eiweiße in unserer Ernährung*, Issue 48 a, Schriftenreihe der Schweizer Vereinigung für Ernährung, 1982, p. 3.
212. G. G. Graham et al., *Am. J. Clin. Nutr.* **22** (1969) 1459.
213. G. R. Jansen, C. F. Hutchison, M. E. Zanetti, *Food Technol. (Chicago)* **20** (1966) 323.
214. R. Bressani et al., *J. Nutr.* **79** (1963) 333.
215. J. L. Iwan, *Cereal Sci. Today* **13** (1968) 202.
216. D. Rosenfield, F. J. Stare, *Mod. Gov.* **11** (1970) 47.
217. S. N. Gershoff et al., *Am. J. Clin. Nutr.* **30** (1977) 1185.
218. H. N. Parthasarathy et al., *Can. J. Biochem.* **42** (1964) 385.
219. S. J. Fomon et al., *Am. J. Clin. Nutr.* **32** (1979) 2460.
220. A. L. Jung, S. L. Carr, *Clin. Pediatr. (Philadelphia)* **16** (1977) 982.
221. Procter & Gamble, US 3878305, 1975 (R. A. Damico, R. W. Boggs).
222. S. Wallrauch, *Flüss. Obst* **44** (1977) 386.
223. H. D. Pruss, I. P. G. Wirotama, K. H. Ney, *Fette Seifen Anstrichm.* **77** (1975) 153.
224. C. Ambrosino et al., *Minerva Pediatr.* **18** (1966) 759.
225. C. A. Masschelein, J. Van de Meerssche, *Tech. Q. Master Brew. Assoc. Am.* **13** (1976) 240.
226. T. Take, H. Otsuka, *Chem. Abstr.* **70** (1969) 46270 m.
227. J. Koch, *Flüss. Obst* **46** (1979) 212.
228. G. Baumann, K. Gierschner, *Dtsch. Lebensm. Rundsch.* **70** (1974) 273.
229. A. Askar, H. J. Bielig, *Alimenta* **15** (1976) 3.
230. W. Hashida, *Food Trade Rev.* **44** (1974) 21.
231. *Food Technol. (Chicago)* **34** (1980) 49.
232. B. Lindemann, *Nat. Neurosci.* **3** (2000) 99.
233. S. Schwimmer, D. G. Guadagni, *J. Food Sci.* **32** (1967) 405.
234. Riken Kagaku, JP 7249707, 1972 (H. Watanabe et al.); *Chem. Abstr.* **79** (1973) 114219 q.
235. Pillsbury Comp., US 3510310, 1970.
236. C. Colburn, *Am. Soft Drink J.* **126** (1971) 16.
237. R. S. Shallenberger, T. E. Acree, C. Y. Lee, *Nature (London)* **221** (1969) 556.
238. J. Solms, L. Vuataz, R. H. Egli, *Experientia* **21** (1965) 693.
239. H. Wieser, H. Jugel, H.-D. Belitz, *Z. Lebensm. Unters. Forsch.* **164** (1977) 277.
240. R. H. Mazur, J. M. Schlatter, A. H. Goldkamp, *J. Am. Chem. Soc.* **91** (1969) 2684
241. L. A. Pavlova et al., *Russ. Chem. Rev. (Engl. Transl.)* **50** (1981) 316.
242. W. J. Herz, R. S. Shallenberger, *Food Res.* **25** (1960) 491.
243. Lever Brothers, US 3922365, 1975 (K. H. Ney et al.).
244. H. Tanaka, Y. Obata, *Agric. Biol. Chem.* **33** (1969) 147.
245. M. Giaccio, L. Surricchio, *Quad. Merceol.* **16** (1977) 151.
246. H. Valaize, G. Dupont, *Ind. Agric. Aliment.* **68** (1951) 245.
247. E. L. Wick, M. DeFigueiredo, D. H. Wallace, *Cereal Chem.* **41** (1964) 300.
248. A. A. M. El-Dash, Dissertation, Kansas State University 1969 (Univ. Microfilms Inc., Ann Arbor, Michigan, No. 69-21123).
249. W. Baltes, *Ernähr. Umsch.* **20** (1973) 35.
250. M. Angrick, D. Rewicki, *Chem. Unserer Zeit* **14** (1980) 149.
251. H. E. Nurstein, *Food Chem.* **6** (1980) 263.
252. T. A. Rohan, *Food Technol. (Chicago)* **24** (1970) 29.
253. Maggi, DE 2246032, 1973 (R. J. Gasser).
Z. Mielniczuk et al., *Acta Aliment. Pol.* **2** (1976) 213.
Chas. Pfizer & Co., US 3365306, 1968 (M. A. Perret).
Y.-P. C. Hsieh et al., *J. Sci. Food Agric.* **31** (1980) 943.
R. A. Wilson, *J. Agric. Food Chem.* **23** (1975) 1032.
P. van de Rovaart, J.-J. Wuhrmann, US 3930044, 1975.
R. Schroetter, G. Woelm, *Nahrung* **24** (1980) 175.
254. H. Kisaki, *Chem. Abstr.* **69** (1968) 9824 d.
Ajinomoto, FR 2005896 (1969).
G. Rubenthaler, Y. Pomeranz, K. F. Kinney, *Cereal Chem.* **40** (1963) 658. I. R. Hunter, M. K. Mayo, US 3425840, 1969. Y. H. Liau, C. C. Lee, *Cereal Chem.* **47** (1970) 404. Y.-Y. Linko, J. A. Johnson, B. S. Miller, *Cereal Chem.* **39** (1962) 468. G. L. Bertram, *Cereal Chem.* **30** (1953) 126. Research Corp., US 3268555, 1966 (L. Wiseblatt).
Hoffmann-La Roche, US 3547659, 1970 (W. Cort, L. Neck). M. Rothe, *Nahrung* **24** (1980)

185. A. A. El-Dash, A. A. Johnson, *Cereal Chem.* **47** (1970) 247. L. Wiseblatt, H. F. Zoumut, *Cereal Chem.* **40** (1963) 162. M. Rothe, *Ernährungsforschung* **5** (1960) 131.
255. G. Ziegleder, D. Sandmeier, *Dtsch. Lebensm. Rundsch.* **78** (1982) 315. W. Mohr, E. Landschreiber, Th. Severin, *Fette Seifen Anstrichm.* **78** (1976) 88. S. Turos, US 4346121, 1982.
256. E. Cremer, M. Riedmann, *Monatsh. Chem.* **96** (1965) 364.
257. Yuki Gosei Kogyo, US 3478015, 1969 (I. Onishi, A. Nishi, T. Kakizawa).
258. Fuji Oil Co., GB 1357511, 1974; GB 1488282, 1977.
259. Naarden Int., NL 7712745, 1979.
260. J. A. Newell, M. E. Mason, R. S. Matlock, *J. Agric. Food Chem.* **15** (1967) 767.
261. S.-C. Lee, B. R. Reddy, S. S. Chang, *J. Food Sci.* **38** (1974) 788. Research Corp., US 3814818, 1974 (S. S. Chang, B. R. Reddy). T. Y. Fan, M. H. Yueh, *J. Food Sci.* **45** (1980) 748. P. T. Arroyo, D. A. Lillard, *J. Food Sci.* **35** (1970) 769.
262. Yuki Gosei Kogyo, DE 1593733, 1972 (I. Onishi et al.). Japan Monopoly, Tanabe Seiyaku, US 3722516, 1973 (K. Suwa et al.). Philip Morris, US 4306577, 1981 (D. L. S. Wu, J. W. Swain). Reynolds Tobacco, US 3996941, 1976 (C. W. Miller, J. P. Dickerson, C. E. Rix).
263. R. G. Henika, N. E. Rodgers, *Cereal Chem.* **42** (1965) 397.
264. J. M Bruemmer, W. Seibel, H. Stephan, *Getreide Mehl Brot* **34** (1980) 173. Patent Technology, US 3803326, 1974. J. Geittner, *Gordian* **79** (1979) 202.
265. R. Marcuse, *Fette Seifen Anstrichm.* **63** (1961) 940.
266. H. Iwainsky, C. Franzke, *Dtsch. Lebensm. Rundsch.* **52** (1956) 129. H. Lingnert, C. E. Eriksson, *J. Food Process. Preserv.* **4** (1980) 161.
267. N. Watanabe et al., JP 7314042, 1973; *Chem. Abstr.* **79** (1973) 145052 j.
268. A. G. Castellani, *Appl. Microbiol.* **1** (1953) 195. Nisshin Flour Milling, JP 7319945, 1973(G. Ogawa, K. Taguchi); *Chem. Abstr.* **81** (1974) 76689 z. Nippon Kayaku, JP-Kokai 81109580, 1981; *Chem. Abstr.* **95** (1981) 202313 b.
269. *US Functional Beverages Market*, Frost & Sullivan, London, 2004.
270. Ajinomoto Co. Inc.: Amino Acid Link News, No. 14 (May 2006). Available at http://www.ajinomoto.com/aminoacid/link/E/pdf/Vol.14E.pdf.
271. M. Pack: *Amino Acids in Animal Nutrition*, Coral Sanivet, Bucharest 2002.
272. M. Kirchgeßner: *Tierernährung*, 11th ed., DLG-Verlag, Frankfurt 2004.
273. *Nutrient Requirements of Poultry*, 9th revised ed., National Academies Press, Washington DC 1994.
274. *Nutrient Requirements of Swine*, 10th revised ed., National Academies Press, Washington DC 1998. P. L. M. Berende, H.-L. Bertram, *Z. Tierphysiol. Tierernähr. Futtermittelk.* **49** (1983) 30. H.-L. Bertram, P. L. M. Berende, *Kraftfutter* (1983) 46.
275. Degussa Feed Additives: available at http://www.aminoacidsandmore.com.
276. M. L. Scott, M. C. Nesheim, R. J. Young: *Nutrition of the Chicken*, 3rd ed., M. L. Scott, Ithaka, New York, 1982.
277. W. Prinz, A. Becker, *Arch. Geflügelk.* **29** (1965) no. 2, 135.
278. R. Chernoff, *J. Am. Diet. Assoc.* **79** (1981) 426. M. R. Polk, *Am. Pharm.* **NS 22** (1982) 25.
279. W. F. Caspary, *Dtsch. Ärztebl.* 1978 (2. Febr.) 243. H. Kasper, *Aktuel. Ernährung* **1** (1978) 22.
280. D. M. Matthews, S. A. Adibi, *Gastroenterology* **71** (1976) 151. M. T. Lis, R. F. Crampton, D. M. Matthews, *Br. J. Nutr.* **27** (1972) 159.
281. Mead Johnson, US 3091569, 1963; US 3184505, 1965.
282. H. A. Smith, G. Gorin, *J. Org. Chem.* **26** (1961) 820.
283. A. Kleemann, J. Engel: *Pharmaceutical Substances*, 4th ed., Thieme Verlag, Stuttgart 2001.
284. Rech. et Propagande Scientif., FR 1288907, 1962.
285. Degussa, US 4129593, 1978.
286. *Drugs of the Future* **23** (1998) 374.
287. M. Bergmann, G. Michalis, *Ber. Dtsch. Chem. Ges.* **63** (1930) 987.
288. Degussa, DE 1239697, 1963.
289. May & Baker, GB 845 034, 1957.
290. B. Witkop et al., *J. Am. Chem. Soc.* **76** (1954) 5579.
291. A. J. Morris et al., *J. Org. Chem.* **22** (1957) 306.
292. E. A. Bell et al., *Nature (London)* **210** (1966) 529.

293. Distillers, GB 854339, 1957.
Squibb, US 3281461, 1966.
Heyl & Co., DE 2114329, 1971;
DE 2413185, 1974.
294. Degussa, DE 1795299, 1968;
DE 1795297, 1968;
DE 2032952, 1970;
DE 2123232, 1971;
DE 2156601, 1971;
DE 2335990, 1973;
DE 2138122, 1971;
DE 2258411, 1972;
DE 2304055, 1973.
295. K. Schrader, *Am. Cosmet. Perfum.* **87** (1972) 49.
S. Tatsumi, *Am. Cosmet. Perfum.* **87** (1972) 61.
O. K. Jacobi, *Am. Cosmet. Perfum.* **87** (1972) 35.
G. Hopf, J. König, G. Padberg, *Kosmetologie* **4** (1971) 132.
A. Szakall, *Arch. Klin. Exp. Dermatol.* **201** (1955) 331.
K. Laden, R. Spitzer, *J. Soc. Cosmet. Chem.* **18** (1967) 351.
H. W. Spier, G. Pascher, *Klin. Wochenschr.* **31** (1953) 997.
296. Y. Kumano et al., *J. Soc. Cosmet. Chem.* **28** (1977) 285.
297. L'Oreál, DE 2807607, 1978 (J.-C. Ser et al.).
Unilever, DE 2337342, 1973 (G. F. Johnston et al.).
Orlane Paris, DE 2524297, 1975 (A. Meybeck, H. Noel).
Shiseido, US 4035513, 1977 (Y. Kumano).
298. E. S. Cooperman, *Am. Cosmet. Perfum.* **87** (1972) 65.
Colgate Palmolive, GB 1573529, 1980.
299. Proctor and Gamble: Press Release, 23 August 2004; available at http://www.uk.pg.com.
300. C. R. Robbins: *Chemical and Physical Behaviour of Human Hair*, Springer Verlag, Berlin 1994.
301. *European Markets for Active Ingredients in Personal Care*, Frost & Sullivan, London, 2004.
302. M. Takehara et al., *J. Am. Oil Chem. Soc.* **50** (1973) 227;
51 (1974) 419.
303. Ajinomoto, Kawaken Fine Chemicals, US 4273684, 1981 (T. Nagashima et al.)
304. Ajinomoto, GB 1483500, 1977.
305. Kawaken Fine Chemicals. JP-Kokai 75117806, 1975 (K. Nakazawa);
Chem. Abstr. **84** (1976) 8858 r.
306. G. Schuster, H. Modde, E. Scheld, *Seifen Öle Fette Wachse* **38** (1965) 477.
G. Schuster, H. Modde, *Parfüm + Kosmet.* **45** (1964) 337.
307. Estee Lauder, US 4005210, 1977 (J. Gubernick).
308. American Cyanamid, US 3988438, 1976.
Ajinomoto, DE 2010303, 1970 (R. Yoshida et al.).
309. R. Yoshida, M. Takehara, *Chem. Abstr.* **84** (1976) 6843 h.
310. M. Koyama, *J. Dispersion Sci. Tech.* **26** (2005) 785.
311. Dominion Pharmacal., US 4176197, 1979 (B. N. Olson).
L'Oréal, US 4002634, 1977 (G. Kalopissis, C. Bouillon).
L'Oréal, GB 1397623, 1975 (G. Kalopissis, G. Manoussos).
L'Oréal, DE 1492071, 1965 (G. Kalopissis).
312. Mare Corp., US 4201235, 1980 (V. G. Ciavatta).
Unilever, EP 8171, 1981 (G. P. Mathur et al.).
313. Kyowa Hakko Kogyo, US 4139610, 1979 (Y. Miyazaki et al.)
Hans Schwarzkopf, DE 958501, 1957 (J. Saphir, E. Kramer).
K. Yoneda et al., DE 2951923, 1979.
314. Schuylkill Chem., GB 1516890, 1978.
315. Procter & Gamble, EP 47650, 1982.
Unilever, GB 1597498, 1981 (K. Gosling, M. R. Hyde).
316. Sederma, US 6974799, 2003 (K. Lintner).
317. Sederma, WO 98/07744, 1998 (D. Greff).
318. C. Tomlin (ed.): *The Pesticide Manual*, 13th ed., British Crop Protection Council, Farnham 2003.
319. B. Hock, C. Fedtke, R. R. Schmidt: *Herbizide*, 1st ed., Georg Thieme Verlag Stuttgart 1995.
320. *Proc. North. Cent. Weed Control Conf.* **26** (1971) 64.
321. P. Knuuttila, H. Knuuttila, *Acta Chem. Scand.* **33** (1979) 623.
322. *Agrochemical Service*, Wood Mackenzie Consultants Limited, Edinburgh 1996.
323. Zeneca, US 4 315 765.
324. *Z. Pflanzenkr. Pflanzenschutz* (1981) Sonderheft IX, 431.
325. Hoechst, DE-OS 2 717 440, 1976.
326. K. Tachibana et al., *Abstr. 5th Int. Congr. Pestic. Chem.,* IVa Abstract 19.
327. D. W. Ladner, *Pestic. Sci.* **29** (1990) 341–356.
328. M. J. Muhitch, D. L. Shaner, M. A. Stidham, *Plant Physiol.* **83** (1987) 451–456.
329. T. Chapman et al., *Symp. New. Herbic. 3rd ed.* (1969) 40.

330. R. M. Scott et al., *Proc. Br. Crop. Prot. Conf.–Weeds* **2** (1976) 723.
331. S. O. Duke et al., *Pestic. Sci.* **40** (1994) 265–277.
332. N. Mito, R. Sato, M. Miyakado, H. Oshio, S. Tanaka, *Pestic. Biochem. Physiol.* **40** (1991) 128–135.
333. Degussa, WO 94/03458, 1994 (M. Schäfer, H. Baier, K. Drauz, H.-P. Krimmer, S. Landmann).
334. Sumitomo, JP 62 158 280.
335. BASF, DE 409 432, 1994.
336. BASF, DE 407 023, 1994.
337. American Cyanamid Co., EP 101 098, 1993.
338. Bayer, DE 030 062, 1990.
339. S. Sauza et al., *Phytoma–la défense des vegetaux* **558** (2003) 51.
340. Y. Miyake et al., *Nippon Noyaku Gakkaishi* **30** (2005) 390.
341. P. A. Urech, *Proc. Br. Crop. Prot. Conf.–Pests Dis.* **2** (1977) 623.
342. Garavaglia et al.: *Atti Simp. Chim. Antiparassitari, 3rd, Piacenza* 1981.
343. C. A. Henrick et al., *Pestic. Sci.* **11** (1980) 224.
344. C. A. Porter, L. E. Ahlrichs, *Hawaii Sugar Tech. Rep. 1971* **30** (1972) 71.
345. Ciba Geigy, EP 18948, 1980.
346. Emori Shoji, JP 78097022, 1978.
347. Dainichi Seika Kogy, JP 78079990, 1978.
348. L. Mueller, *Zem. Kalk Gips* **27** (1974) 69.
349. Ajinomoto, DE 2533136, 1975 (R. Yoshida et al.).
350. Kyowa Fermentation, GB 1400741, 1975.
Honny Chemicals, GB 1402758, 1975 (Y. Nakagoshi).
Kyowa Hakko Kogyo, DE 2229488, 1972 (Y. Fujimoto et al.).
Toyo Cloth, US 3676206, 1972 (K. Nishitani et al.).
351. Sanyo Trading Co., DE 2716758, 1977 (M. Onizawa, S. Ohmiya).
352. Ajinomoto, JP 81035179, 1981.
Ajinomoto, JP-Kokai 7522801, 1975 (T. Saito et al.);
Chem. Abstr. **84** (1976) 33494 b.
353. Sanyo Trading Co., DE 2602988, 1976;
US 4069213, 1978;
DE 2604053, 1976 (M. Onizawa);
DE 2658693, 1976 (M. Onizawa, S. Ohmiya).
354. R. M. Saleh, A. M. Shams El Din, *Corros. Sci.* **12** (1972) 688.
355. Ajinomoto, JP-Kokai 7426145, 1974 (Y. Kita et al.);
Chem. Abstr. **81** (1974) 53195 w.
356. Ilford, GB 1378354, 1974 (A. D. Ezekiel).
Ilford, DE 2316632, 1973 (R. Jefferson).
357. Fr. Blasberg & Co., DE 2050870, 1973 (W. Immel, W. Adams).
Yokozawa Chemical Ind., JP 7410572, 1974 (K. Aoya et al.);
Chem. Abstr. **81** (1974) 57631 h.
Sony Corp., DE 2325109, 1973 (S. Fueki et al.).
358. A. Steponavicius, R. Visomirskis, *Electrodepos. Surface Treat. Lausanne* **1** (1972) 37.
359. F. S. Parker, D. M. Kirschenbaum, *Spectrochim. Acta* **16** (1960) 910.
M. Tsuboi, T. Takenishi, A. Nakamura, *Spectrochim. Acta* **19** (1963) 271.
J. F. Pearson, M. A. Slifkin, *Spectrochim. Acta Part A* **28** (1972) 2403.
360. C. S. Tsai et al., *Can. J. Biochem.* **53** (1975) no. 9, 1005.
361. H. Hagenmeyer et al., *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **25 B** (1970), 681.
362. C. Cooper, W. Packer, K. Williams: *Amino Acid Analysis Protocols*, Methods in Molecular Biology Series, **159**, Humana Press, Totowa, NJ 2000.
363. W. S. Hancock: *Handbook for the Separation of Amino Acids, Peptides and Proteins*, vol. **1**, CRC Press, Boca Raton, FL, 1984.
364. E. Scoffone, A. Fontana, *Mol. Biol. Biochem. Biophys.* **8** (1970) 185.
365. M. Roth, *Anal. Chem.* **43** (1971) 880.
J. R. Benson, P. E. Hare, *Proc. Natl. Acad. Sci. USA* **72** (1975) 619.
366. E. Bayer et al., *Anal. Chem.* **48** (1976) 1106.
367. S. Udenfried et al., *Science (Washington, D.C.)* **178** (1972) 871.
368. I. Molnár-Perl: *Quantitation of Amino Acids and Amines by Chromatography*, Elsevier, Oxford, 2005.
369. C. Haworth, R. W. A. Oliver, *J. Chromatogr.* **64** (1972) 305.
E. Stahl: *Dünnschichtchromatographie*, Springer Verlag, Berlin 1967, p. 701.
A. R. Fahmy et al., *Helv. Chim. Acta* **44** (1959) 245.
370. M. Kubota et al., *Anal. Biochem.* **64** (1975) no. 2, 494.
K. D. Kulbe, *Anal. Biochem.* **44** (1970) 548.
P. A. Laursen, *Biochem. Biophys. Res. Commun.* **37** (1969) 663.
371. A. Wolf, *Prax. Naturwiss. Chem.* **23** (1974) no. 3, 74.
372. H. Nakamura, J. J. Pisano, *J. Chromatogr.* **121** (1976) 33.

373. K. Macek, Z. Deyl, M. Smrž, *J. Chromatogr.* **193** (1980) 421.
374. P. B. Hamilton, *Anal. Chem.* **35** (1963) 2055.
375. S. Blackburn: *Amino Acid Determination*, Marcel Dekker, New York 1978.
376. D. H. Spackman, W. H. Stein, S. Moore, *Anal. Chem.* **30** (1958) 1190.
377. S. Moore, W. H. Stein, *J. Biol. Chem.* **211** (1954) 907.
378. H.-M. Lee et al., *Anal. Biochem.* **96** (1979) 298.
379. P. Chaimbault et al., *J. Chromatogr. A*, **870** (2000) 245.
380. J. Weiss: *Handbook of Ion Chromatography*, Wiley-VCH, Weinheim 2004.
381. E. Bayer et al., *Anal. Chem.* **48** (1976) 1106.
J. M. Wilkinson, *J. Chromatogr. Sci.* **16** (1978) 547.
K.-T. Hsu, B. L. Currie, *J. Chromatogr.* **166** (1978) 555.
T. Jamabe, N. Takei, H. Nakamura, *J. Chromatogr.* **104** (1975) 359.
A. Khayat, P. K. Redenz, L. A. Gorman, *Food Technol. (Chicago)* **36** (1982) 46.
382. A. P. Graffeo, A. Haag, B. L. Karger, *Anal. Lett.* **6** (1973) no. 6, 505.
J. K. De Vries, R. Frank, C. Birr, *FEBS Lett.* **55** (1975) no. 1, 65.
P. Frankhauser et al., *Helv. Chim. Acta* **57** (1974) 271.
C. C. Zimmermann, E. Appella, J. J. Pisano, *Anal. Biochem.* **77** (1977) 569.
G. Frank, W. Strubert, *Chromatographia* **6** (1973) no. 12, 522.
383. H. Beyer, U. Schenk, *J. Chromatogr.* **89** (1969) 483.
384. T. A. Kan, W. F. Shipe, *J. Food Sci.* **47** (1981) 338.
H. Umagat, P. Kucera, L. F. Wen, *J. Chromatogr.* **239** (1982) 463.
385. W. Voelter, K. Zech, *J. Chromatogr.* **112** (1975) 643.
386. R. Schuster, *Anal. Chem.* **52** (1980) 617.
387. A. Darbre, *Biochem. Soc. Trans.* **2** (1974) 70.
B. M. Nair, *J. Agric. Food Chem.* **25** (1977) 614.
388. J. J. Pisano, T. J. Bronzert, H. B. Brewer, *Anal. Biochem.* **45** (1972) 43.
389. A. Darbre, A. Islam, *Biochem. J.* **106** (1968) 923.
390. C. W. Gehrke, R. W. Zumwalt, K. Kuo, *J. Agric. Food Chem.* **19** (1971) 605.
391. K. Ruhlmann, W. Giesecke, *Angew. Chem.* **73** (1961) 113.
392. D. L. Stalling, C. W. Gehrke, R. W. Zumwalt, *Biochem. Biophys. Res. Commun.* **31** (1968) 4.
393. M. W. Duncan, A. Poljak, *Anal. Chem.* **70** (1998) 890.
394. W. Grassmann, K. Hannig, *Houben-Weyl*, **1/1**, 708.
395. V Poinsot et al., *Electrophoresis* **27** (2006) 176.
396. T. Soga, D. N. Heiger, *Anal. Chem.* **72** (2000) 1236.
397. I. Abe, S. Musha, *J. Chromatogr.* **200** (1980) 195.
G. J. Nicholson, H. Frank, E. Bayer, *HRC CC J. High Resolut. Chromatogr. Chromatogr. Commun.* **2** (1979) 411.
W. A. König, G. J. Nicholson, *Anal. Chem.* **47** (1975) 951.
398. W. Lindner, *Chimia* **35** (1981) 294.
V. A. Davankov et al., *Chromatographia* **13** (1980) no. 11, 677.
399. P. E. Hare, E. Gil-Av, *Science (Washington, D.C.)* **204** (1979) 1226.
400. C. B'Hymer, M. Montes-Bayon, J. A. Caruso, *J. Sep. Sci.* **26** (2003) 7.
401. R. Bhushan, H. Bruckner, *Amino Acids* **27** (2004) 231.
402. L. M. Henderson, E. E. Snell, *J. Biol. Chem.* **172** (1947) 15.
I. Grote, *Mühle Mischfuttertech.* **116** (1979) 465.
403. H. Itoh, T. Morimoto, I. Chibata, *Anal. Biochem.* **60** (1974) 573.
404. Ch. Calvot et al., *FEBS Lett.* **59** (1975) no. 2, 258.
S. J. Updike, G. P. Hicks, *Nature (London)* **214** (1967) 986.
405. Ajinomoto, FR 2421380, 1979.
406. *European Pharmacopoeia*, 5th ed., European Directorate for the Quality of Medicine, Strasbourg 2005.
407. W. Seaman, E. Allen, *Anal. Chem.* **23** (1951) no. 4, 592.
H. P. Deppeler, G. Witthans, *Fresenius Z. Anal. Chem.* **305** (1981) 273.
408. *United States Pharmacopeia*, 29th ed., US Pharmcopeial Convention, Inc., Rockville, MD 2006.
409. *Food Chemicals Codex*, 5th ed., National Academies Press, Washington, D.C. 2003.
410. Ajinomoto: Amino Acids Business, 2005 (available at http://www.ajinomoto.com).
411. E. O. Camara Greiner, R. Gubler, K. Yokose: *Market Research Report: Amino Acids*, SRI International, Menlo Park, CA, 2003.
412. *European Amino Acids Market*, Frost & Sullivan, London, 2006.

413. K. Brown: *Amino Acids: Highlighting Synthesis Applications*, Business Communications Co. Inc., Norwalk, CT 2005.
414. H. N. Munro, *Adv. Exp. Med. Biol.* **105** (1978) 119.
415. A. E. Harper, N. J. Benevenga, R. M. Wohlhueter, *Physiol. Rev.* **50** (1970) 428.
416. R. G. Daniel, H. A. Waisman, *Growth* **32** (1968) 255.
417. H. E. Sauberlich, *J. Nutr.* **75** (1961) 61.
418. K. Lang: *Biochemie der Ernährung*, 4th ed., Steinkopff, Darmstadt 1979.
419. Y. Peng et al., *J. Nutr.* **103** (1973) 608.
420. Degussa, unpublished, 1973 and 1983. R. J. Breglia, C. O. Ward, C. I. Jarowski, *J. Pharm. Sci.* **62** (1973) 49. P. Gullino et al., *Arch. Biochem. Biophys.* **58** (1955) 253. O. Strubelt, C.-P. Siegers, A. Schütt, *Arch. Toxicol.* **33** (1974) 55. W. Braun, *Strahlentherapie* **108** (1959) 262. H. Gutbrod et al., *Acta Hepatol.* **5** (1957) no. 1/2, 1. I. Petersone et al., *Eksp. Klin. Farmakoter.* **3** (1972) 5. D. G. Gallo, A. L. Sheffner, DE 2018599, 1971. Transbronchin, Homburg Pharma, Frankfurt 1971. L. Bonanomi, A. Gazzaniga, *Therapiewoche* **30** (1980) 1926. W. F. Riker, H. Gold, *J. Am. Pharm. Assoc.* **31** (1942) 306.
421. G. Maffii, G. Schott, M. G. Serralunga, *Res. Prog. Org. Biol. Med. Chem.* **2** (1970) 262.
422. Y. Kawaguchi et al., *Iyakuhin Kenkyu* **11** (1980) 635. Kaken Chemical, JP 52083940, 1976 (S. Suzuki). P. Gullino et al., *Arch. Biochem. Biophys.* **64** (1956) 319. Degussa, unpublished, 1981 and 1982. E.-J. Kirnberger et al., *Arzneim. Forsch.* **8** (1958) 72.

Aminophenols

Stephen C. Mitchell, Imperail College, London, United Kingdom

Rosemary H. Waring, University of Birmingham, Birmingham, United Kingdom

1. Aminophenols 1
1.1. Physical Properties 1
1.2. Chemical Properties 2
1.3. Production 4
1.3.1. Reduction 4
1.3.2. Substitution 5
1.3.3. Purification 6
1.4. Storage and Handling 6
1.5. Analysis 6
1.6. Environmental Aspects 6
1.7. Uses 7
1.8. Toxicology 7
2. Aminophenol Derivatives 7
2.1. Derivatives of 2-Aminophenol 8
2.2. Derivatives of 3-Aminophenol 12
2.3. Derivatives of 4-Aminophenol 13
3. References 15

1. Aminophenols

Aminophenols and their derivatives are of steadily increasing commercial importance, both in their own right and as intermediates in the chemical and dye industries. They are amphoteric and can behave either as weak acids or weak bases, but the basic character usually predominates. 2-Aminophenol (**1**) and 4-aminophenol (**2**) are oxidized easily. This is the basis for their main applications, for example, as photographic developers. In contrast, 3-aminophenol (**3**) is fairly stable in air and is not oxidized rapidly.

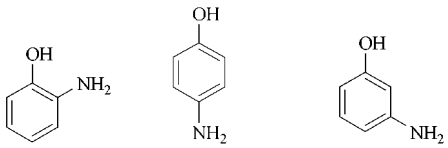

1 **2** **3**

1.1. Physical Properties

The simple aminophenols exist in three isomeric forms depending on the relative positions of the hydroxyl and amino groups around the benzene ring. At room temperature they are solid crystalline compounds. The commercial-grade compounds are usually impure because of contamination with oxidation products and may take on a yellow-brown or pink-purple hue, the 2- and 4-aminophenols being more susceptible to this phenomenon than the 3-isomer. The solubilities of these compounds in common solvents and their spectral characteristics are given in Tables 1 and 2, respectively [3–12].

Table 1. Solubility ∗ of aminophenols in common solvents

	2-Aminophenol	3-Aminophenol	4-Aminophenol
Acetone	3	3	2
Acetonitrile	3	3	2
Benzene	1	1	0
Chloroform	1	1	0
Diethyl ether	2	2	1
Dimethylsulfoxide	3	3	3
Ethanol	3	3	1
Ethyl acetate	3	3	2
Toluene	1	1	1
Water			
Cold	1	2	1
Hot	2	3	2

∗ 0, insoluble; 1, slightly soluble; 2, soluble; 3, very soluble

2-Aminophenol (**1**) [*95-55-6*], 2-hydroxyaniline, 2-amino-1-hydroxybenzene, C.I. Oxidation Base 17, C.I. 76520, C_6H_7NO, M_r 109.13, forms white orthorhombic bipyramidal needles when crystallized from water or benzene, *mp* 174 °C. The crystals have eight molecules to the elementary cell and a density of 1.328 g/cm^3 (1.29 also quoted) [13–16]. At reduced pressure (1.47 kPa), 2-aminophenol sublimes rapidly at 153 °C without decomposition [17]. Acid – base dissociation constants are pK_1 4.72 (water at 21 °C), 4.66 (1 vol % ethyl

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_099

Table 2. Spectral characteristics of the aminophenol isomers

	2-Aminophenol	3-Aminophenol	4-Aminophenol
IR[a], cm^{-1}	3380, 3300, 1600, 1510, 1470, 1270, 900, 740	3370, 3310, 1600, 1470, 1390, 1260, 1180, 910	3050 – 2580, 1500, 1470, 1240, 970, 830, 750
UV[b], nm	233, 285 (methanol) 229, 281 (water) 235, 288 (cyclohex.)	287 (methanol) 270 (0.1 M HCl) 234, 284 (cyclohex.)	234, 301 (methanol) 229, 294 (water) 235, 304 (cyclohex.)
NMR[c], ppm	6.9 – 7.5, 8.6 (TFA)	4.7, 6.0, 6.1, 6.8, 8.8 (DMSO)	7.1, 7.4, 8.7 (TFA)

[a] Only IR absorption bands reported as very strong are included (accuracy: $\pm 10\,cm^{-1}$) [4], [5].
[b] UV spectra: cyclohex. = cyclohexane containing 2 vol % diethyl ether [6–8].
[c] Proton chemical shift spectra over the range of 0 – 15 ppm ($\pm$ 0.1 ppm); TFA, trifluoroacetic acid; DMSO, dimethylsulfoxide. When complex spectra caused by second-order effects or overlapping resonances were encountered the range was recorded [10], [11].

alcohol in water at 25 °C); pK_2 9.66 (water at 15 °C), 9.71 (water at 22 °C) [18], [19]. Salts: hydrochloride [*51-19-4*], needles, *mp* 207 °C; formate, *mp* 120 °C; oxalate, *mp* 167.5 °C (decomp.); acetate, *mp* 150 °C.

3-Aminophenol (**3**) [*591-27-5*], 3-hydroxyaniline, 3-amino-1-hydroxybenzene, C_6H_7NO, M_r 109.13, forms white prisms when crystallized from water or toluene, *mp* 122 – 123 °C. The orthorhombic crystals have a tetramolecular unit and a density of 1.195 g/cm^3 (1.206 and 1.269 also quoted) [14–16]. 3-Aminophenol boils at 164 °C at 1.47 kPa with slight decomposition [17]. Acid – base dissociation constants are: pK_1 4.17 (water at 21 °C), 4.31 (1 vol % aqueous ethyl alcohol at 25 °C); pK_2 9.87 (water at 22 °C) [18]. Salts: hydrochloride [*51-81-0*], prisms, *mp* 229 °C; hydrobromide, prisms, *mp* 224 °C; hydroiodide, prisms, *mp* 209 °C; sulfate [*66671-80-5*], plates or needles, *mp* 152 °C; oxalate, *mp* 275 °C.

4-Aminophenol (**2**) [*123-30-8*], 4-hydroxyaniline, 4-amino-1-hydroxybenzene, C_6H_7NO, M_r 109.13, forms white plates when crystallized from water, *mp* 189 – 190 °C (decomp.). The crystals exist in two forms. The α form (from alcohol, water, or ethyl acetate) is the more stable and has an orthorhombic pyramidal structure containing four molecules per unit cell. It has a density of 1.290 g/cm^3 (1.305 also quoted). The less stable β form (from acetone) exists as acicular crystals that turn into the α form on standing; they are orthorhombic bipyramidal or pyramidal and have a hexamolecular unit [14], [15], [20]. At reduced pressure (40 Pa) 4-aminophenol sublimes at 110 °C with slight decomposition. Boiling points quoted at various pressures are as follows: 284 °C (101.3 kPa), 174 °C (1.47 kPa), 167 °C (1.07 kPa), 150 °C (0.4 kPa), 130.2 °C (0.04 kPa), decomposition usually occurring. Acid – base dissociation constants are: pK_1 5.50 (water at 21 °C), 4.86 (water at 30 °C), 5.48 (1 vol % aqueous ethyl alcohol at 25 °C); pK_2 10.30 (water at 22 °C), 10.60 (water at 30 °C) [18], [21]. Salts: hydrochloride [*51-78-5*], prisms, *mp* 306 °C (decomp.); hydrosulfate [*15658-52-3*], needles, *mp* 272 °C; oxalate, *mp* 183 °C; acetate [*13871-68-6*], *mp* 183 °C; chloroacetate, needles, *mp* 148 °C; trichloroacetate, needles, *mp* 166 °C.

1.2. Chemical Properties

The chemical properties and reactions of the aminophenols are to be found in detail in many standard chemical texts and only a summary is given here [22].

The acidity of the phenols is depressed by the presence of an amino group on the benzene ring; this phenomenon is most pronounced with 4-aminophenol. They also behave as weak bases, giving salts with both mineral and organic acids. The aminophenols are true ampholytes, with no zwitterion structure; hence they exist either as neutral molecules (**4**) or as ammonium cations (**5**) or phenolate anions (**6**), depending on the pH value of the solution.

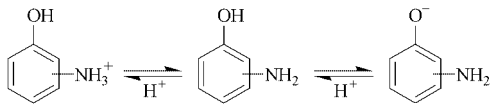

5 **4** **6**

However, deviations from theoretical curves observed during acid – base titrations have led to postulation of the existence of half-salt complex cations B_2^+, formed by the association of an ammonium cation, B^+, with a neutral molecule, B. This association phenomenon is most apparent with 4-aminophenol but is also displayed by the other isomers [21].

The aminophenols are chemically reactive, undergoing reactions involving both the aromatic amino group and the phenolic hydroxyl moiety, as well as substitution on the benzene ring. Oxidation leads to the formation of highly colored polymeric quinoid structures. 2-Aminophenol undergoes a variety of cyclization reactions.

Alkylation. All of the possible mono-, di-, and trimethylated aminophenols are known. *N*-Monoalkylation occurs when the aminophenol is heated with the appropriate alkyl halide or with an alcohol and Raney nickel; equal or even better results can be achieved by using aldehydes or ketones in place of the alcohol. Specific alkylation of the hydroxyl group to form methoxyanilines (anisidines) or ethoxyanilines (phenetidines) is difficult because of the reactivity of the amino group; mixed alkylated products usually are obtained. 3-Methoxyaniline may be prepared by methylation of 3-aminophenol under alkaline conditions, but it is more usual to protect the amino group and to methylate 3-acetylaminophenol, followed by hydrolysis. The other anisidines and phenetidines are prepared indirectly by reduction of the nitro analog:

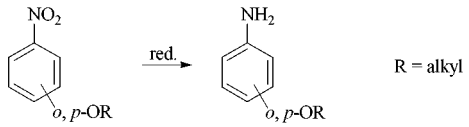

Acylation. Acylation of the aminophenol (using acetic anhydride in alkali or pyridine, acetyl chloride and pyridine in toluene, or ketene in ethanol) usually leads to *N*-acylated products. If an excess of reagent is used, however, especially with 2-aminophenol, *O,N*-diacylated products are formed. Aminophenyl carboxylates (*O*-acylated aminophenols) normally are prepared by the reduction of the corresponding nitrophenyl carboxylates, which is of particular importance with the 4-aminophenol derivatives. A migration of the acyl group from the *O* to the *N* position is known to occur for both 2- and 4-aminophenol acylated products. 2-Aminophenyl ethyl carbonate (**7**) rearranges slowly in dilute acid to ethyl 2-hydroxyphenyl carbamate (**8**); the corresponding 4 derivative does not undergo this particular reaction [23].

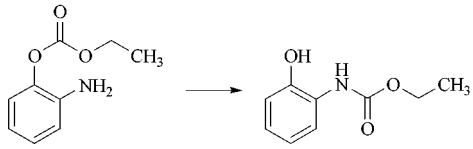

7 **8**

Diazonium Salt Formation. The aromatic amino group of aminophenols can be converted to the diazonium salt with sodium nitrite in aqueous acid, although difficulties may be encountered when the aminophenol is oxidized easily or of low solubility. Crystalline diazonium salts have been isolated by using the hydrochloride or sulfate of the appropriate aminophenol under anhydrous conditions. Such diazo derivatives find extensive use in the dye [24], [25].

Cyclization Reactions. Owing to the close proximity of the amino and hydroxyl groups in 2-aminophenol, this isomer is particularly susceptible to cyclization and condensation reactions. Oxidation by iron(III) chloride, enzymes, or light or autoxidation on silica thin-layer plates gives 2-aminophenoxazin-3-one (**9**). Further oxidation with iron(III) cyanide or heating with potassium hydroxide in ethanol gives a five-ring structure, triphenoxdioxazine (benzoxazinophenoxazine) (**10**). 2-Aminophenol and its derivatives are useful starting materials for the synthesis of phenoxazones, phenoxazines, benzoxazoles, and thiobenzoxazoles. Most of these condensation reactions involve heating at 200 – 300 °C with a suitable catalyst [22].

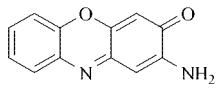

9

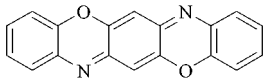

10

Condensation Reactions. Substituted diphenylamines or diphenyl ethers are obtained from aminophenols and suitable reactants by elimination of ammonia or hydrogen chloride:

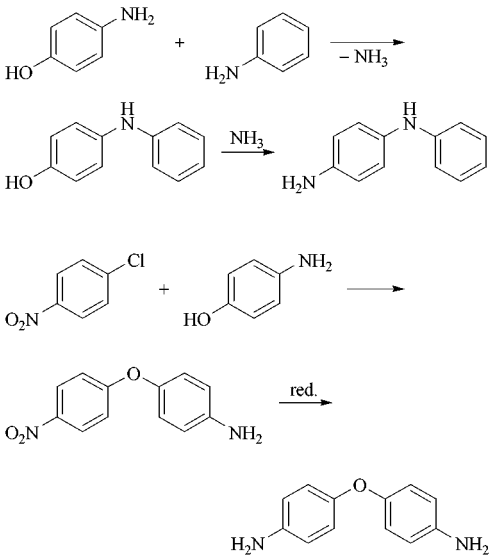

Reactions of the Benzene Ring. Both the amino and hydroxyl groups are electron-donating moieties, and many substituted derivatives are known. The controlled interaction of aminophenols with chlorine or bromine in glacial acetic acid can give a variety of mono-, di-, tri-, or tetra-halogenated products. The use of concentrated sulfuric acid or oleum, with or without heat, gives aromatic sulfonic acids. The sulfonic acid group enters the *ortho* or *para* position relative to the hydroxyl group. Further treatment with oleum leads to the formation of disulfonated compounds. The carboxylation of *m*-aminophenol leads to the formation of *p*-aminosalicylic acid.

1.3. Production

Aminophenols are made either by reduction of nitrophenols or by substitution. Reduction is accomplished with iron or with hydrogen in the presence of a catalyst. The last mentioned is the method of choice today for the production of 2- and 4-aminophenol.

1.3.1. Reduction

Iron Reduction [26]. The reduction of nitrophenols with iron turnings takes place in weakly acidic solution or suspension. Before the iron – iron oxide sludge is separated from the solution [27] the product aminophenol must be made water soluble by adding sodium hydroxide. The resulting sodium aminophenolate is very susceptible to oxidation in aqueous solution; various methods are recommended for its purification (see Section 1.3.3). Subsequent to this, the aminophenols are precipitated from acidic solution by neutralization with base and in the absence of air; reducing agents usually must be added during this procedure.

When 2-nitrophenol is reduced with iron, red insoluble color lakes are formed as byproducts that decrease the yield. Therefore, the iron reduction of 2-nitrophenol is of minor industrial importance today.

Catalytic Reduction [28–30]. Catalytic reduction usually takes place in solution, emulsion, or suspension in autoclaves or pressurized vessels; after the catalyst is added, the vessel is pressurized with hydrogen. Water and methanol are the preferred solvents; in water the addition of alkali hydroxide [31–34], alkali carbonate [31], [34], or acid [34], [35] has been recommended. Nickel [36] – preferably Raney nickel [32], [33], [37], [38] – or supported precious metals [31], [39], such as platinum or palladium [34], [40] on activated carbon, or the oxides of these metals [35], [41] are used as catalysts. The catalyst life can be extended, the catalyst consumption decreased, and the product quality enhanced by adding organic solvents that are not miscible with water [32], [34]. The preferred hydrogen pressure is 2 MPa; the hydrogenation can also be performed at atmospheric pressure or at higher pressure up to 6 MPa. The reaction temperature does not exceed 100 – 110 °C.

Reduction of Nitrobenzene. When reducing nitrobenzene in acidic medium [42], the intermediate phenylhydroxylamine [43] rearranges to 2- and 4-aminophenol before it is further reduced to aniline. The main product of this reaction is 4-aminophenol; byproducts are 2-aminophenol, aniline, and 4,4′-diaminodiphenyl ether:

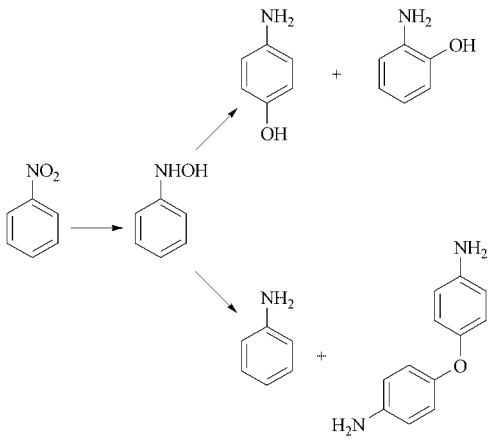

In the past, metals in dilute sulfuric acid were used as reducing agents [44]. Today, the reducing agent is hydrogen in the presence of precious-metal catalysts [43], [45], [46], e.g., palladium or platinum. Other catalysts have been suggested: molybdenum and platinum sulfide [47], [48], and a platinum – ruthenium mixed catalyst [49]. Either the catalysts are used as their oxides, or they are supported on activated carbon.

Dilute aqueous mineral acid is used as reaction medium, for example, dilute sulfuric acid; acidic salts also can be added to the reduction medium [47].

In a two-step process, nitrobenzene first is selectively reduced to phenylhydroxylamine with hydrogen in the presence of Raney copper and in an organic solvent, such as 2-propanol [43]. The product rearranges to 4-aminophenol after addition of dilute sulfuric acid [50].

The addition of *wetting agents* increases the aminophenol yield [51]; these agents must be water soluble and stable in the presence of sulfuric acid. Quaternary ammonium salts that contain at least one alkyl group with at least ten carbon atoms are suitable, e.g., dodecyltrimethylammonium chloride. The reaction usually is performed below 100 °C, either at atmospheric or at higher pressure. Hydrogen is added during the reaction as consumed. The addition of inert organic solvents further increases the yield of 4-aminophenol and the product quality [52].

In another variant, only 88 % of the nitrobenzene is reduced [53]; after that, the reaction mixture consists of two phases with the precious-metal catalyst (palladium on activated carbon) remaining in the unreacted nitrobenzene phase. Therefore, phase separation is sufficient as workup, and the nitrobenzene phase can be recycled directly to the next batch. The aqueous sulfuric acid phase contains 4-aminophenol and byproduct aniline. After neutralization, the aniline is stripped, and 4-aminophenol is obtained by crystallization after the aqueous phase is purified with activated carbon.

Electrolytic reduction also is possible [44], [54–57]; this method causes less concern over pollution than metal – acid reduction systems, but it has not yet found industrial application. Electrolysis of nitrobenzene or phenylhydroxylamine in the presence of sulfuric acid or of azoxybenzene in acid solution yields specifically 4-aminophenol [58].

1.3.2. Substitution

Substitution of various groups by amino or hydroxyl groups is industrially unimportant for the production of 2- and 4-aminophenol [59], but this type of reaction is used for the synthesis of 2- or 4-aminophenol derivatives.

However, 3-aminophenol cannot be obtained easily by reduction. It is made mainly by the reaction of 3-aminobenzenesulfonic acid with sodium hydroxide or by the reaction of resorcinol with ammonia. Substitution of the sulfonic acid group in 3-aminobenzenesulfonic acid is accomplished by caustic soda fusion (5 – 6 h; 240 – 245 °C). The product is purified by vacuum distillation [60].

Alternatively, resorcinol reacts with ammonia [61], for example, in the presence of diammonium phosphate and arsenic pentoxide [62] or ammonium sulfite to form 3-aminophenol. The compound also may be made by hydrolysis of 3-aminoaniline [63], [64].

1.3.3. Purification

Generally, aminophenols can be purified by sublimation at reduced pressure and higher temperature. 3-Aminophenol may be purified by vacuum distillation; to obtain a colorless product sulfur dioxide is added during distillation [65] or the distillate is collected under a blanket of an unreactive liquid of lower density, such as water [66].

Another method for purifying aminophenols is the treatment of their aqueous solutions with activated carbon [67], [68]. During this treatment, sodium sulfite, sodium dithionite, or disodium ethylenediaminotetraacetate [68] is added to increase the quality and stability of the products and to chelate heavy-metal ions that would catalyze oxidation. Addition of sodium dithionite, hydrazine [68], or sodium hydrosulfite [69] also is recommended during precipitation or crystallization of aminophenols.

Contaminants, which are usually present in the 4-aminophenol made by catalytic reduction, can be reduced or even removed completely by a variety of procedures: treatment with 2-propanol [70]; with aliphatic, cycloaliphatic or aromatic ketones [71]; with aromatic amines [72]; with toluene or low molecular mass alkyl acetates [73]; with phosphoric acid, hydroxyacetic acid, hydroxypropionic acid, or citric acid [74]; or by extraction with methylene chloride, chloroform [75], or nitrobenzene [76].

1.4. Storage and Handling

Under atmospheric conditions, 3-aminophenol is the most stable of the three isomers. Both 2- and 4-aminophenol are unstable; they darken on exposure to air and light and should be stored in brown glass containers, preferably in an atmosphere of nitrogen. The use of activated iron oxide in a separate cellophane bag inside the storage container inhibits the discoloration of 4-aminophenol [77]. The salts, especially the hydrochlorides, are more resistant to oxidation and should be used where possible.

1.5. Analysis

Aminophenols have been detected in waste water by investigating UV absorptions at 220, 254, and 275 nm [78] and also in parts per thousand by using potentiometric titrations in a two-phase chloroform–water medium [79] or potentio-argentometric titrations of liberated bromide after reaction with bromine [80]. More specifically, 2-aminophenol can be identified using an iron(II) sulfate–hydrogen peroxide reagent [81] or dimerized in acidic solution to 2-hydroxyisophenoxazin-3-one, as intensively colored dye [82]. 4-Aminophenol can be analyzed by voltammetry [83], [84]. A colorimetric method using 3-cyano-*N*-methoxypyridinium perchlorate as reagent detects 4-aminophenol in the presence of *N*-acetyl-4-aminophenol [85]. 3-Aminophenol has been analyzed colorimetrically by oxidation in base and subsequent extraction of a violet quinoneimide dye [86].

Chromatographic methods for the separation and identification of aminophenols also have been described. Thin layer chromatography provides a rapid and convenient method of separating the isomers from many derivatives, and subsequent spraying with a variety of chromogenic reagents gives additional information [87], [88]. Several gas–liquid [89] and high-pressure liquid [90–92] chromatographic separation techniques have also been reported, and the use of selected-ion mass spectrometry greatly increases detection sensitivity [93], [94].

1.6. Environmental Aspects

As the aminophenols are oxidized easily they tend to remove oxygen from solutions. Hence, if they are released from industrial waste waters into streams and rivers they will deplete the capacity of these environments to sustain aquatic life. The adsorption of aminophenols onto metal ferrocyanides and activated carbons from bituminous coal has met with limited success but may provide a possible means of removal [95], [96]. Biological degradation of 3-aminophenol has been carried out in aeration tanks with adapted microflora from active sludge; the process took 4 months [97]. An enzymatic method for removing aminophenols from waste waters has been described using horseradish peroxidase to cross-link and precipitate the compounds as insoluble polymers [98]. Degradative metabolism by bacteria [99–101] and the im-

mobilization of microbial cells onto Ca-alginate beads [102] may open up a new approach.

1.7. Uses

Both 2- and 4-aminophenols are strong reducing agents and are employed as photographic developers under the trade names of Atomal and Ortol (2-aminophenol); Activol, Azol, Certinal, Citol, Paranol, Rodinal, Unal, and Ursol P (4-aminophenol); they may be used alone or in combination with hydroquinone. The oxalate salt of 4-aminophenol is marketed under the name of Kodelon.

The aminophenols are versatile intermediates and are employed in the synthesis of virtually every class of stain and dye. In addition, 2-aminophenol (Oxidation Base 17; C.I. 76520) is specifically used for shading leather, fur, and hair from grays and browns to yellowish brown. 3-Aminophenol has found application as a hair colorant and as a coupler molecule in hair dyes [103], [104]. 4-Aminophenol is used as an intermediate in the synthesis of pharmaceuticals, as a wood stain imparting a roselike color to timber [105], and as a dyeing agent for fur and feathers.

As a result of the close proximity of the amino and hydroxyl groups on the benzene ring and their ease of condensation with suitable reagents, 2-aminophenol is a principal intermediate in the synthesis of such heterocyclic systems as oxyquinolines, phenoxamines, and benzoxazoles. The last-named compounds have been used as inflammation inhibitors [106]. 3-Aminophenol has found use as a stabilizer of chlorine-containing thermoplastics [107], although its major use is as an intermediate in the production of 4-amino-2-hydroxybenzoic acid, a tuberculostat. Similarly, nitrogen-substituted 4-aminophenols have long been known as antipyretics and analgesics, and the production of these derivatives is a major use of 4-aminophenol.

1.8. Toxicology

In general, aminophenols are irritants. Their toxic hazard rating is slight to moderate, but repeated contamination may cause general itching, skin sensitization, dermatitis, and allergic reactions [108], [109]. Immunogenic conjugates are spontaneously produced on exposure to 2- and 4-aminophenol [110]. Methemoglobin formation with subsequent cyanosis is another possible complication. Inhalation of 4-aminophenol may precipitate this event and also cause bronchial asthma, [111]. 3-Aminophenol is less hazardous than the others. The sulfonated derivatives are less irritating than the unsulfonated compounds.

2-Aminophenol is neuroactive, inducing spike discharges when instilled into the cerebroventricle of the rat [112]. 4-Aminophenol is nephrotoxic and strongly inhibits proximal tubular function [113–115]. Respiration, oxidative phosphorylation, and ATPase activity are inhibited in rat kidney mitochondria [116]. 2-Aminophenol and 4-aminophenol, but not 3-aminophenol, are teratogenic in the hamster [117]. 4-Aminophenol is known to inhibit DNA synthesis and alter DNA structure in human lymphoblasts [118]. The aminophenols also have been shown to be genotoxic, as evidenced by the induction of sister chromatid exchanges and micronucleus tests [119–122].

Obviously, care should be taken in handling these compounds; prolonged exposure should be avoided. Contaminated clothing should be removed immediately and the affected area washed thoroughly with running water for at least 10 min.

2. Aminophenol Derivatives

The derivatives of the aminophenols have important uses in both the photographic and the pharmaceutical industries. They are also extensively applied as precursors and intermediates in the synthesis of more complicated molecules, especially those used in the staining and dye industry. All of the major classes of dyes have representatives that incorporate substituted aminophenols. The varying degrees of ease with which the aminophenols can be diazotized and coupled has led to their use in the manufacture of azo compounds. The sulfonated aminophenols also are employed in this context; these compounds, however, are treated elsewhere (→ Benzenesulfonic Acids and Their Derivatives). Details concerning which aminophenol derivatives and sulfonated compounds

are produced commercially as dye intermediates have been reviewed and are given in the publications listed in reference [123]. The more commonly encountered simple derivatives of the aminophenols can be found in the review articles cited in the references and also in standard organic chemistry texts [124]. A few examples, which have specific uses or are manufactured in large quantities, have been selected for detailed discussion.

2.1. Derivatives of 2-Aminophenol

2-(*N*-Methylamino)phenol (**11**) [*611-24-5*], C_7H_9NO, M_r 123.15, *mp* 99 °C, forms colorless plates; the aqueous hydrochloric acid solution turns reddish brown when $FeCl_3$ solution is added.

Compound **11** is synthesized by reaction of 2-aminophenol with phosgene, alkylation of the intermediate benzoxazolone, and cleavage of *N*-methylbenzoxazolone with hydrochloric acid [125]. Alternatively, 2-*N*-acetyl-*N*-methylaminophenol is heated at 140 °C in the presence of hydrochloric acid [126]:

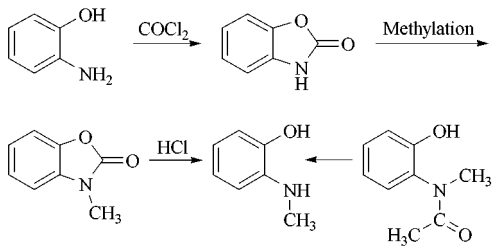

2-Methoxyaniline (**12**) [*90-04-0*], *o*-anisidine, C_7H_9NO, M_r 123.15, *mp* 6.22 °C, *bp* 224 °C (102.1 kPa), *bp* 105 – 106 °C (1.85 kPa), d_4^{20} 1.0923, $d_4^{98.7}$ 1.0263, n_D^{20} 1.5738, dynamic viscosity η 2.211 mPa · s (55 °C), η 1.051 mPa · s (98.7 °C), dipole moment 1.62 D (vapor), is slightly soluble in water (1.38 wt % at 25 °C).

The compound is produced from 2-methoxynitrobenzene by reduction with iron [127] or by hydrogenation in the presence of precious-metal catalysts [128] (platinum or palladium on activated carbon; solvent: toluene) or of Raney nickel [129] (solvent: methanol or 2-propanol).

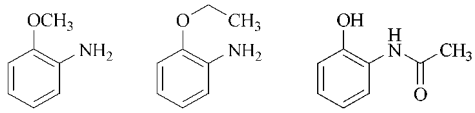

12 **13** **14**

2-Ethoxyaniline (**13**) [*94-70-2*], *o*-phenetidine, $C_8H_{11}NO$, M_r 137.18, d_{20}^{20} 1.0513, *bp* 232.5 °C (101.3 kPa), is produced from 2-ethoxynitrobenzene by reduction with iron or by catalytic hydrogenation in the presence of precious-metal catalysts or Raney nickel [130].

2-Acetamidophenol (**14**) [*614-80-2*], 2-hydroxyacetanilide, $C_8H_9NO_2$, M_r 151.16, forms colorless crystals, *mp* 209 °C, that are quite soluble in ethanol and hot water. Production is by acetylation of 2-aminophenol with acetic anhydride [131].

2-Amino-5-nitrophenol (**15**) [*121-88-0*], $C_6H_6N_2O_3$, M_r 154.13, forms orange needles, *mp* 207 – 208 °C, that are sparingly soluble in water. Production is by nitration of benzoxazolone (**16**) and separation from the 2-amino-4-nitrophenol isomer after treatment with hydrochloric acid [132]. It induces Cu(II)-dependent DNA cleavage, presumably via the generation of active oxygen species [133].

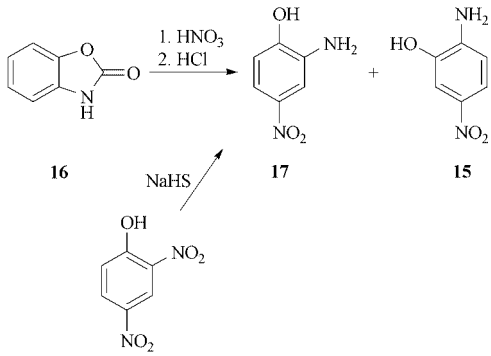

2-Amino-4-nitrophenol (**17**) [*99-57-0*], $C_6H_6N_2O_3$, M_r 154.13, 2-hydroxy-5-nitroaniline, forms orange prisms from water that are hydrated with one molecule of water of crystallization, *mp* 80 – 90 °C. They can be

dehydrated over sulfuric acid to the anhydrous form, *mp* 143 – 145 °C, soluble in alcohol, ether, acetic acid, and warm benzene and slightly soluble in water.

Compound **17** is produced commercially by the partial reduction of 2,4-dinitrophenol; this may be achieved electrolytically with vanadium [134] or chemically with polysulfide, sodium hydrosulfide, or hydrazine and copper [135–137]. Alternatively, 2-acetamidophenol or 2-methylbenzoxazole may be nitrated in sulfuric acid to yield a mixture of 4- and 5-nitro derivatives that are then separated and hydrolyzed with sodium hydroxide [138].

The major use of this compound is in the production of mordant and acid dyes. 2-Amino-4-nitrophenol also has found limited use as an antioxidant and light stabilizer in butyl rubbers and as a catalyst in the manufacture of hexadiene.

2-Amino-4-nitrophenol has been shown to be a skin irritant and continuous exposure should be avoided. Toxicologic studies have shown it to be nonaccumulative [139] but suggest it is carcinogenic [133].

2-Amino-4,6-dinitrophenol (**18**) [*96-91-3*], $C_6H_5N_3O_5$, M_r 199.13, 4,6-dinitro-2-aminophenol, picramic acid, forms dark red needles from alcohol and prisms from chloroform, *mp* 169 – 170 °C. The compound flashes at 210 °C in contact with an open flame, ignites rapidly, and burns relatively fast. 2-Amino-4,6-dinitrophenol is soluble in glacial acetic acid, water, alcohol, benzene, and aniline and sparingly soluble in ether and chloroform.

The compound can be prepared from 2,4,6-trinitrophenol (picric acid) by reduction with sodium hydrosulfide [140], with ammonia – hydrogen sulfide followed by acetic acid neutralization of the ammonium salt [141], with ethanolic hydrazine and copper [137], or electrolytically with vanadium sulfate in alcoholic sulfuric acid [134]. Heating 4,6-dinitro-2-benzamidophenol in concentrated HCl at 140 °C also yields picramic acid [142]. It is a major intermediate in the manufacture of colorants, especially mordant dyes. It has also been used as an indicator dye in titrations (yellow with acid, red with alkali) and as a reagent for albumin determination. The induction of sister chromatid exchange and micronuclei formation by this compound suggest potential health hazards [143], [144].

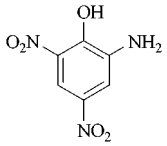

18

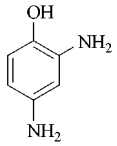

19

2,4-Diaminophenol (**19**) [*95-86-3*], 4-hydroxy-*m*-phenylenediamine, $C_6H_8NO_2$, M_r 124.14, forms leaflets that darken on exposure to air and melt at 78 – 80 °C with decomposition. The compound is soluble in acid, alkali, alcohol, and acetone and sparingly soluble in chloroform, ether, and ligroin. 2,4-Diaminophenol usually is sold as the sulfate [*72556-58-2*] (Diamol) or dihydrochloride salt [*137-09-7*] (Acrol, Amidol).

2,4-Diaminophenol can be prepared from 2,4-dinitrophenol by catalytic hydrogenation or, less conveniently, by metal reduction in acid solution (Béchamp method) [145–150]. Alternatively, electrolytic reduction and subsequent hydroxylation of 1,3-dinitrobenzene or 3-nitroaniline in sulfuric acid can be undertaken [151–153].

The dihydrochloride salt is used as a photographic developer. It also is applied as an intermediate in the manufacture of fur dyes, in hair dyeing, as a reagent in testing for ammonia and formaldehyde, and as an oxygen scavenger in water to prevent boiler corrosion [154].

2-Amino-4-chlorophenol [*95-85-2*], C_6H_6ClNO, M_r 143.58, forms colorless plates, *mp* 139 – 140 °C, that are extremely sensitive to oxidation when wet. For its production 2,5-dichloronitrobenzene is converted to 4-chloro-2-nitrophenol by reaction with sodium hydroxide; this product is then reduced either with iron [155], with hydrazine, or with hydrogen [156] in the presence of Raney nickel or platinum catalysts.

2-Amino-4-methylphenol [*95-84-1*], C_7H_9NO, M_r 123.15, forms crystals, *mp* 135 °C (decomp.), that are almost insoluble in cold water. The compound is quite soluble in ethanol, diethyl ether, and chloroform but is almost in-

soluble in benzene. 2-Amino-4-methylphenol is obtained by reduction of 2-nitro-4-methylphenol either with hydrazine or catalytically [157].

2-Methoxy-5-methylaniline (**20**) [*120-71-8*], 2-amino-4-methylanisole, 3-amino-4-methoxytoluene, $C_8H_{11}NO$, M_r 137.18, forms needles or plates, *mp* 52 °C, *bp* 235 °C (101.3 kPa), that can be steam distilled. The compound is quite soluble in ethanol, diethyl ether, and benzene, but not very soluble in cold water. Production is by alkylation of 4-methyl-2-nitrophenol and subsequent reduction with iron [158].

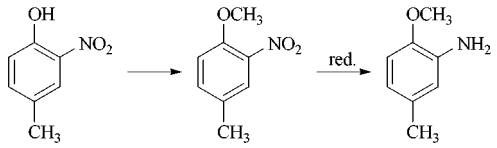

20

2-Methoxyacetanilide (**21**) [*93-26-5*], $C_9H_{11}NO_2$, M_r 165.19, crystallizes as needles, *mp* 86.5 °C, *bp* 303 – 305 °C (101.3 kPa). The compound is quite soluble in acetic acid and hot water. Production is by acetylation of 2-methoxyaniline with acetic anhydride.

2-Methoxy-5-chloroaniline (**22**) [*2401-24-3*], C_7H_8ClNO, M_r 157.61, *mp* 84 °C, is quite soluble in ethanol and can be steam distilled. Production is by reduction of the corresponding nitro compound with iron [159].

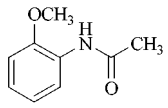

21

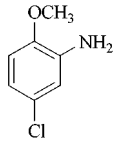

22

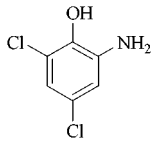

23

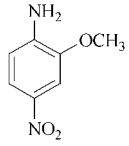

24

2-Amino-4,6-dichlorophenol (**23**) [*527-62-8*], $C_6H_5Cl_2NO$, M_r 178.03, is sparingly soluble in water and ethanol and decomposes when heated above 109 °C. Industrial production is by reduction of the corresponding nitro compound with iron [160] or with hydrazine.

2-Methoxy-4-nitroaniline (**24**) [*97-52-9*], $C_7H_8N_2O_3$, M_r 168.15, *mp* 140 – 142 °C, is prepared by acetylation and subsequent nitration of 2-methoxyaniline to form intermediate **33** , which is then deacetylated to **24**.

2-Methoxy-5-nitroaniline (**25**) [*99-59-2*], $C_7H_8N_2O_3$, M_r 168.15, *mp* 118.5 – 119 °C, is prepared either by partial reduction of 2,4-dinitroanisole (**26**) with ammonium or sodium sulfide [161] or by nitration of 2-methoxyaniline (**27**) [162].

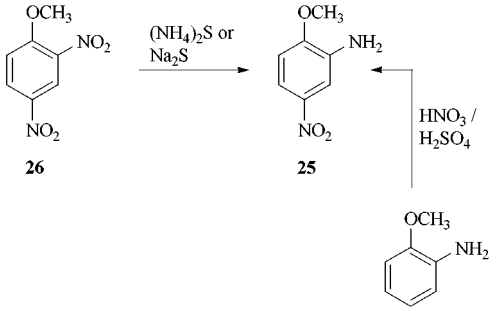

26 **25** **27**

2-Amino-4-chloro-5-nitrophenol (**28**) [*6358-07-2*], $C_6H_5ClN_2O_3$, M_r 188.58, forms yellow needles that darken when heated above 200 °C and decompose above 225 °C. The compound is scarcely soluble in hot water but quite soluble in ethanol. Industrial production is by reaction of 2-amino-4-chlorophenol with phosgene, nitration of the product 5-chlorobenzoxazolone (**29**) with nitric acid to form 5-chloro-4-nitrobenzoxazolone (**30**), and subsequent hydrolysis with sodium hydroxide [163].

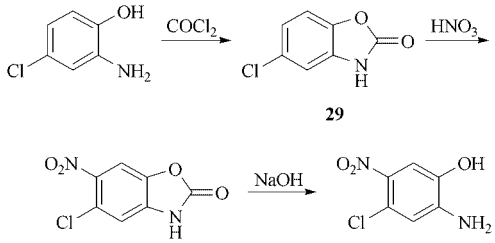

2-Amino-6-chloro-4-nitrophenol (**31**) [*6358-09-4*], $C_6H_5ClN_2O_3$, M_r 188.58, crystallizes from water in yellow needles that contain 1 mol of water. Dehydration occurs when the needles are heated above 100 °C; the crystals turn scarlet and finally melt at 160 °C. The compound is sparingly soluble in water and quite soluble in ethanol and diethyl ether. Industrial production is by partial reduction of 6-chloro-2,4-dinitrophenol with sodium hydrosulfide or ammonium sulfide [164].

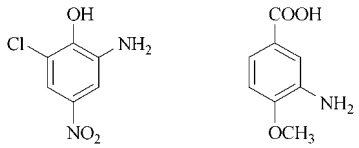

31 32

3-Amino-4-methoxybenzoic acid (**32**) [*2840-26-8*], $C_8H_9NO_3$, M_r 167.16, *mp* 204 – 205 °C, forms needles or prisms that are sparingly soluble in cold water and also in boiling water (0.12 wt %). The crystals are sparingly soluble in diethyl ether and quite soluble in hot ethanol. Production is by reduction of the corresponding nitro compound.

4-Amino-2-methoxy-acetanilide (**34**) [*5329-15-7*], 4-acetamido-3-methoxyaniline, $C_9H_{12}N_2O_2$, M_r 180.2, *mp* 120 – 121 °C, is quite soluble in ethanol, ethyl acetate, acetone, and chloroform. The solubility in diethyl ether is low. Production is by acetylation of 2-methoxyaniline and subsequent nitration with nitric acid [165]. Nitro compound **33** is then reduced with iron in dilute acetic acid [166].

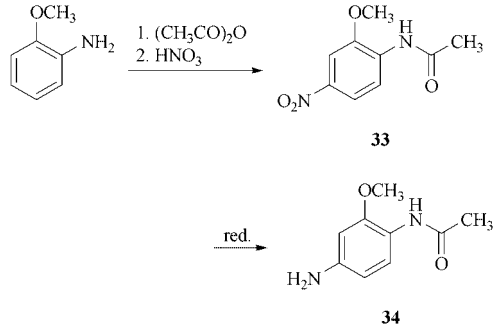

2-Methylbenzoxazole (**35**) [*95-21-6*], C_8H_7NO, M_r 133.16, *mp* 8.5 – 10 °C, *bp* 178 °C, n_D^{20} 1.5497, is made by cyclization of 2-acetamidophenol. Its chief use is for the production of 2-amino-5-nitrophenol (**15**).

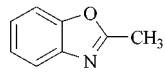

35

Orthocaine [*536-25-4*], 3-amino-4-hydroxybenzoic acid methyl ester, methyl 3-amino-4-hydroxybenzoate, Orthoform New, $C_8H_9NO_3$, M_r 167.16, *mp* 143 °C, is almost insoluble in cold, but soluble in hot water, and very soluble in alcohol, ether, and dilute alkali. For preparation, see [167], [168].

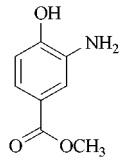

The compound is used as a topical anesthetic.

Acetarsone [*97-44-9*], 3-acetamido-4-hydroxyphenylarsonic acid, acetarsol, stovarsol, $C_8H_{10}AsNO_5$, M_r 275.08, *mp* 240 – 250 °C (decomp.), is practically insoluble in alcohols or dilute acids, slightly soluble in water, and freely soluble in dilute aqueous alkali. For its preparation, see [169], [170].

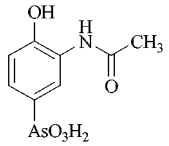

Salts of acetarsone are used in the treatment of intestinal amoebiasis, trade name Acetarsol, of vaginal trichomoniasis, trade name S.V.C. (May and Baker), and as a constituent of a mouthwash, trade name Pyrex (Bengué). Because of acetarsone's toxicity, safer drugs have been developed; lethal dose LD_{50} 150 mg/kg (rabbit, oral).

2.2. Derivatives of 3-Aminophenol

3-(*N,N*-Dimethylamino)phenol (36) [*99-07-0*], 3-hydroxydimethylaniline, $C_8H_{11}NO$, M_r 137.18, forms white needles, *mp* 87 °C. Boiling points quoted at various pressures are 265 – 268 °C (101.3 kPa), 206 °C (13.3 kPa), 194 °C (6.7 kPa), 153 °C (0.7 kPa). The compound is soluble in alkali, mineral acid, ethanol, ether, acetone, and benzene and practically insoluble in water.

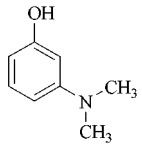

36

3-(*N,N*-Dimethylamino)phenol can be prepared by heating resorcinol with an aqueous solution of dimethylamine and its hydrochloride at 200 °C under pressure for 12 h [171], [172]. Alternatively, the treatment of dimethylaniline with oleum at 55 – 60 °C, followed by fusion with sodium hydroxide at 270 – 300 °C, gives 3-(*N,N*-dimethylamino)phenol [173]. In addition, 3-aminophenol may be methylated with dimethyl sulfate under neutral conditions [174] or its hydrochloride salt heated with methanol at 170 °C under pressure for 8 h to give the desired product [175]. The compound is used primarily as an intermediate in the production of basic (Red 3 and Red 11) and mordant (Red 77) dyes.

3-(*N*-Methylamino)phenol (37) [*14703-69-6*], C_7H_9NO, M_r 123.15, *bp* 170 °C (1.6 kPa), is easily soluble in ethyl acetate, ethanol, diethyl ether, and benzene; it is sparingly soluble in cold water, better in hot water. Industrial synthesis is by heating 3-(*N*-methylamino)benzenesulfonic acid with sodium hydroxide at 200 – 220 °C [176] or by the reaction of resorcinol with methylamine in the presence of aqueous phosphoric acid at 200 °C [177].

3-(*N,N*-Diethylamino)phenol (38) [*91-68-9*], $C_{10}H_{15}NO$, M_r 165.23, forms rhombic bipyramidal crystals, *mp* 78 °C, *bp* 276 – 280 °C (101.3 kPa), *bp* 209 – 211 °C (1.6 kPa). Industrial synthesis is analogous to the previously described synthesis of 3-(*N,N*-dimethylamino)phenol: from resorcinol and diethylamine, by reaction of 3-(*N,N*-diethylamino)benzenesulfonic acid with sodium hydroxide, or by alkylation of 3-aminophenol hydrochloride with ethanol.

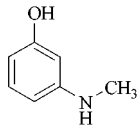

37

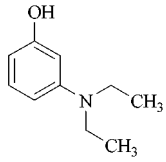

38

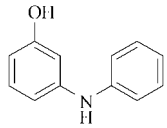

39

3-(*N*-Phenylamino)phenol (39) [*101-18-8*], $C_{12}H_{11}NO$, M_r 185.22, *mp* 81.5 – 82 °C, *bp* 340 °C (101.3 kPa), is slightly soluble in ethanol, diethyl ether, acetone, benzene, and water. The compound is made by heating resorcinol and aniline at 200 °C in the presence of aqueous phosphoric acid [62] or calcium chloride [178]. In another process, 3-aminophenol is heated with aniline hydrochloride at 210 – 215 °C [179].

4-Amino-2-hydroxybenzoic acid (40) [*65-49-6*], 4-aminosalicylic acid (PAS), $C_7H_7NO_3$, M_r 153.13, white crystals from alcohol, melts at 150 – 151 °C with effervescence

and darkens on exposure to light and air. A reddish-brown crystalline powder is obtained on recrystallization from ethanol – ether. 4-Amino-2-hydroxybenzoic acid is soluble in dilute solutions of nitric acid and sodium hydroxide, alcohol, and acetone; slightly soluble in water and ether; and virtually insoluble in benzene, chloroform, or carbon tetrachloride. It is unstable in aqueous solution and decarboxylates to form 3-aminophenol. Because of the instability of the free acid it is usually prepared as the hydrochloride salt, *mp* 224 °C (decomp.), dissociation constant pK_a 3.25.

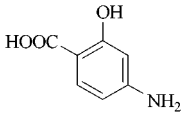

40

4-Amino-2-hydroxybenzoic acid is manufactured by carboxylation of 3-aminophenol under pressure with ammonium carbonate at 110 °C [180] or with potassium bicarbonate and carbon dioxide at 85 – 90 °C [181] and subsequent acidification.

The major use of this compound is as a bacteriostatic agent against tubercle bacilli. The compound also is used as an adjunct to streptomycin and isoniazid. The free acid and its sodium, potassium, and calcium salts are marketed under many trade names. Up to 10 – 15 g of the sodium salt may be administered each day, although prolonged use may give rise to toxic symptoms. Oral LD_{50} of the free acid in mice is 4 g/kg.

2.3. Derivatives of 4-Aminophenol

4-(*N*-Methylamino)phenol (41) [*150-75-4*], 4-hydroxy-*N*-methylaniline, C_7H_9NO, M_r 123.15, forms needles from benzene, *mp* 87 °C, *bp* 168 – 169 °C (2 kPa); it is slightly soluble in alcohol and insoluble in ether.

Industrial synthesis involves the decarboxylation of *N*-(4-hydroxyphenyl)glycine (**47**) at elevated temperature in such solvents as chlorobenzene – cyclohexanone [182]. It also can be prepared by the methylation of 4-aminophenol [183] or from methylamine by heating with 4-chlorophenol and copper sulfate at 135 °C in aqueous solution [184] or with hydroquinone at 200 – 250 °C in alcoholic solution [185], [186].

Its chief use is as a component in photographic developers. Because the free compound is unstable in air and light, it usually is marketed as the sulfate salt [*55-55-0*], Metol, *mp* 260 °C (decomp.). It also finds application as an intermediate for fur and hair dyes and, under certain circumstances, as a corrosion inhibitor for steel. Prolonged exposure to 4-(*N*-methylamino)phenol has been associated with the development of dermatitis and allergies.

4-(*N,N*-Dimethylamino)phenol (42) [*619-60-3*], 4-hydroxydimethylaniline, $C_8H_{11}NO$, M_r 137.18, forms large rhombic crystals from ether – hexane or ether – ligroin that melt at 75 – 76 °C, *bp* 101 – 103 °C (66.7 Pa). The compound forms a salt with sulfuric acid, *mp* 208 – 210 °C [187], [188].

Methylation of 4-aminophenol with a methyl halide under pressure produces 4-(*N,N*dimethylamino)phenol. The major competing product, 4-hydroxyphenyltrimethyl ammonium halide (or the corresponding base), also yields 4-(*N,N*-dimethylamino)phenol on distillation. Alternatively, it can be synthesized by dealkylation of 4-methoxydimethylaniline with hydroiodic acid at reflux temperature for 10 h [189] or by the photodecomposition of 4-dimethylaminobenzenediazonium tetrafluoroborate [187].

The compound is an intermediate in several synthetic reactions and recently has found extensive use in experimental toxicity studies in animals. It has been shown to cause methemoglobinemia; its metabolism in humans has been discussed in [190–195].

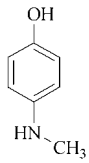

41

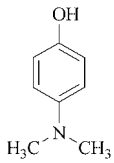

42

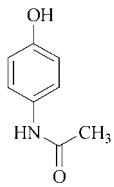

43

4-Hydroxyacetanilide (**43**) [*103-90-2*], 4-acetamidophenol, acetaminophen, paracetamol, forms large white monoclinic prisms from water that melt at 169 – 171 °C. The compound is odorless and has a bitter taste. 4-Hydroxyacetanilide is insoluble in petroleum ether, pentane, and benzene; slightly soluble in ether and cold water; and soluble in hot water, alcohols, dimethylformamide, 1,2-dichloroethane, acetone, and ethyl acetate. The dissociation constant, pK_a, is 9.5 (25 °C).

Production is by the acetylation of 4-aminophenol. This can be achieved with acetic acid and acetic anhydride at 80 °C [196], [197], with acetic anhydride in pyridine at 100 °C [198], with acetyl chloride and pyridine in toluene at 60 °C [199], or by the action of ketene in alcoholic suspension [200]. 4-Hydroxyacetanilide also may be synthesized directly from 4-nitrophenol. The available reduction – acetylation systems include tin with acetic acid [201], hydrogenation over Pd – C in acetic anhydride [202], and hydrogenation over platinum in acetic acid [203]. Other routes include rearrangement of 4-hydroxyacetophenone hydrazone with sodium nitrite in sulfuric acid [204] and the electrolytic hydroxylation of acetanilide [205].

4-Hydroxyacetanilide is used as an intermediate in the manufacture of azo dyes and photographic chemicals. The compound possesses antipyretic and analgesic properties and is used widely in this context. The oral LD_{50} in rats is 3.7 g/kg.

4-Methoxyaniline [*104-94-9*], C_7H_9NO, has M_r 123.55, *mp* 58.5 °C, *bp* 123.2 – 123.5 °C (1.9 kPa), dipole moment 1.8 D (benzene), d_4^{55} 1.092, n_D^{67} 1.5559, viscosity η 3.215 mPa · s (55 °C). For its industrial production 4-nitroanisole is reduced either with sodium sulfide [206] or with hydrogen in the presence of precious-metal catalysts or Raney nickel.

4-Ethoxyaniline [*156-43-4*], $C_8H_{11}NO$, M_r 137.18, *mp* 2.4 °C, *bp* 249.9 °C (101.3 kPa), $d_4^{14.9}$ 1.0652, $n^{15.9}{}_{\alpha}$ 1.5572, is a colorless compound at room temperature when pure; in the light or in the presence of air, the liquid turns brown. 4-Ethoxyaniline is slightly soluble in water.

Industrial production of 4-ethoxyaniline is by reduction of 4-nitrophenetole with iron [207] or with hydrogen under pressure in the presence of a catalyst. Alternatively, 4-chlorophenetole is heated with aqueous ammonia in the presence of copper(I) oxide at 225 °C; the yield of 4-ethoxyaniline is 85 % [208].

4-Ethoxyacetanilide (**44**) [*62-44-2*], phenacetin, *p*-acetophenetidine, 4-acetamidophenetole, $C_{10}H_{13}NO_2$, M_r 179.21, is a white crystalline powder, *mp* 134 – 135 °C. The compound is odorless and has a slightly bitter taste; it is sparingly soluble in cold water and more soluble in hot water, alcohol, ether, and chloroform. At relative humidities between 15 and 90 % the equilibrium moisture content is about 2 % (25 °C).

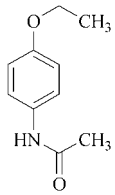

44

The main production route to 4-ethoxyacetanilide is by catalytic reduction of 4-nitrophenetole with hydrogen and subsequent acetylation using acetic anhydride. The compound also can be synthesized by ethylating 4-nitrophenol with ethyl sulfate in alkali, reducing the nitro group to an amino group with iron in acid, and then acetylating by boiling with glacial acetic acid [209]. Alternatively, 4-aminophenol may be ethylated with ethyl iodide in alcoholic alkali; the resulting 4-phenetidine is then acetylated. The acetylation also may be carried out first, followed by the ethylation.

4-Ethoxyacetanilide possesses both antipyretic and analgesic properties but is of little value for the relief of severe pain. Its use for prolonged periods should be avoided because one of its minor metabolites (2-hydroxyphenetidine) is nephrotoxic and may be involved in the formation of methemoglobinemia. The oral LD_{50} in rats is 1.65 g/kg [210].

5-Aminosalicylic acid (**45**) [*89-57-6*], 5-amino-2-hydroxybenzoic acid, $C_7H_7NO_3$, M_r 153.13, forms colorless crystals that turn

brown on heating above 250 – 260 °C and melt at 283 °C (decomp.). The crystals are insoluble in ethanol and cold water and sparingly soluble in hot water.

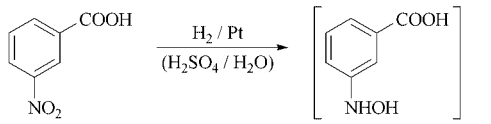

46

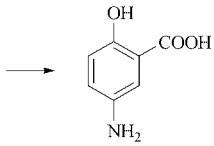

45

The compound is made by reduction of 3-nitrobenzoic acid at 115 – 145 °C and 3.5 MPa in dilute sulfuric acid and in the presence of platinum catalyst. The intermediate hydroxylamine derivative (**46**) rearranges to **45** under these conditions [211]. Alternatively, 5-nitrosalicyclic acid is reduced, for example, with zinc in hydrochloric acid [212].

***N*-(4-Hydroxyphenyl)glycine (47)**

[*122-87-2*], 4-hydroxyphenylaminoacetic acid, 4-oxyanilinoacetic acid, photoglycine, $C_8H_9NO_3$, M_r 167.16, forms aggregate spheres or shiny leaflets from water; it turns brown at 200 °C, begins to melt at 220 °C, and melts completely with decomposition at 245 – 247 °C. The compound is soluble in alkali and mineral acid and sparingly soluble in water, glacial acetic acid, ethyl acetate, alcohol, ether, acetone, chloroform, and benzene.

N-(4-Hydroxyphenyl)glycine can be prepared from 4-aminophenol and chloroacetic acid [213], [214] or by alkaline hydrolysis of the corresponding nitrile with subsequent elimination of ammonia [215].

N-(4-Hydroxyphenyl)glycine is used as a photographic developer under the trade names of Glycine, Iconyl, and Monazol. It also is applied as a photoresist in the dye industry and serves as an intermediate in the production of 4-(*N*-methylamino)phenol (Metol) by liberation of CO_2. *N*-(4-Hydroxyphenyl)glycine is used in analytical chemistry for the determination of iron, phosphorus, and silicon and as an acid indicator in bacteriology. Prolonged use of this compound may result in kidney damage [216].

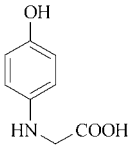

47

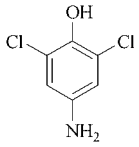

48

4-Amino-2,6-dichlorophenol (48)

[*5930-28-9*], $C_6H_5Cl_2NO$, M_r 178.03, can be sublimed and melts at 167 °C. The compound is soluble in ethanol, diethyl ether, acetone, and acetic acid; it is sparingly soluble in benzene and almost insoluble in water.

3. References

General References

1. K.-F. Wedemeyer, in "Phenole,"*Houben-Weyl* **Teil 1 and 2** (1976).
2. *Beilstein,* **13,** 354 – 549; **13 (2),** 164 – 308.

Specific References

3. E. G. Smith: *The Wiswesser Line-Formula Chemical Notation,* McGraw-Hill, New York 1968.
4. G. Varsanyi: *Assignment for Vibrational Spectra of Seven Hundred Benzene Derivatives,* Akademiai Kiado, Budapest – Adam Hilger, London 1974, pp. 136, 206, 253, 473, 476, 478.
5. C. J Pouchert (ed.): *The Aldrich Library of Infrared Spectra,* 3rd ed., **vol. 3,** Aldrich Chemical Co., Milwaukee, Wisc. 1981, pp. 718 B, 720 D, 725 D.
6. W. F. Forbes, I. R. Leckie, *Can. J. Chem.* **36** (1958) 1371 – 1380.
7. J. C. Dearden, W. F. Forbes, *Can. J. Chem.* **37** (1959) 1294 – 1304.
8. *Sadtler Catalog of Ultraviolet Spectra,* The Sadtler Research Co., Philadelphia, Pa., SAD no. 236, 1894, 3509.
9. *Aldermaston Eight-peak Index of Mass Spectra,* Imperial Chemical Industries, Organics Division, Manchester, & Mass Spectrometry Data Centre, Atomic Weapons Research Establishment (AWRE), Aldermaston, U.K., p. 46.

10. C. J. Pouchert, J. R. Campbell (eds): *The Aldrich Library of NMR Spectra,* **vol. 5.**, Aldrich Chemical Company, Milwaukee, Wisc., 1974 – 75, pp. 45 C, 51 B, 57 A.
11. *Sadtler Catalog of NMR Spectra,* The Sadtler Research Co., Philadelphia, Pa., SAD no. 717, 1176, 10220.
12. J. G. Grasselli (ed.): *CRC Atlas of Spectral Data and Physical Constants for Organic Compounds,* CRC Press, Cleveland, Ohio, 1973, p. B 754.
13. S. Ashfaquzzaman, A. K. Pant: *Acta Crystallogr., Sect. B.* **B 35** (1979) no. 6, 1394 – 1399; *Chem. Abstr.* **91** (1979) 66707.
14. W. A. Caspari, *Phil. Mag.* **4,** 1927, no. 7, 1276 – 1285.
15. J. D. H. Donnay, H. M. Ondik in: *Crystal Data; Determinative Tables,* 3rd ed., U.S. Dept. of Commerce National Bureau of Standards & Joint Committee on Powder Diffraction Standards, Washington, D.C., 1972, pp. 0 – 23, 0 – 73, 0 – 97, 0 – 106.
16. F. H. Allen et al.: "Crystal Engineering and Correspondence Between Molecular and Crystal Structures. Are 2- and 3-aminophenols anomalous?", *J. Am. Chem Soc.* **119** (1997) 3477 – 3480.
17. N. V. Sidgwick, R. K. Callow, *J. Chem. Soc. (Trans.)* **125** (1924) 522 – 527.
18. R. Kuhn, A. Wassermann, *Helv. Chim. Acta* **11** (1928) 1 – 30.
19. V. H. Veeley, *J. Chem. Soc.* **93** (1908) 2122 – 2144.
20. R. W. G. Wyckoff, in: *Crystal Structures,* 2nd ed., **vol. 6,** part 1, Interscience, New York 1969, pp. 186 – 190.
21. G. Chuchani, J. A. Hernández, J. Zabicky, *Nature* **207** (1965) 1385 – 1386.
22. S. Coffey (ed.): *Rodd's Chemistry of Carbon Compounds,* 2nd ed., **vol. 3 A,** Elsevier Publishing Co., Amsterdam 1971, pp. 352 – 363.
23. J. H. Ransom, *Am. Chem. J.* **23** (1900) 1 – 50; *J. Am. Chem. Soc.* **22** (1900) 89 – 91.
24. K. H. Saunders: *The Aromatic Diazo Compounds and Their Technical Application,* Arnold, London 1949, pp. 17 – 20.
25. H. Zollinger: *Azo and Diazo Chemistry,* Interscience, New York 1961, pp. 51 – 53 and 250 – 265.
26. *BIOS Report* **986** (1946) 45.
27. CS 159564, 1975 (D. Kulda, J. Fuka, J. Ott, Z. Misar).
28. CPC International, US 3535382, 1970 (B. Brown, F. Schilling).
29. Mallinckrodt Inc., EP 41837, 1980 (W. R. Clingan, E. L. Derrenbacker, T. J. Dunn).
30. Mallinckrodt Inc., US 4264529, 1975 (T. J. Dunn).
31. Universal Oil, FR 1354430, 1964.
32. BASF, DE 1244196, 1967.
33. Du Pont, US 2183019, 1939.
34. Seiko Chem. Ind. Co. Ltd., JP 75135042, 1975.
35. Abbott Lab., US 3079435, 1963.
36. Sumitomo, JP 30294/68, 1968.
37. Monsanto, US 2035292, 1933.
38. P. Sidheswaran, V. Krishnan, A. N. Bhat: "Catalytic Reduction of p-Nitrophenol to p-Aminophenol,." *Indian J. Chem. Sect. A.* **36A** (1997) 149 – 152.
39. Sumitomo, JP 2701/69, 1969.
40. Abbott Lab., FR 1307415, 1962.
41. Du Pont, US 2947781, 1960.
42. Constructors John Brown Ltd., DE-OS 2026039, 1970. Continental Oil Co., US 3338806, 1967. Miles Laboratories, Inc., DE 1066589, 1959. Miles Lab. Inc., GB 856436, 1960. Ionics Incorp., US 3103473, 1963. Warner-Lambert Pharmaceutical Co., US 2998450, 1958. National Carbon Co., US 2427433, 1947.
43. Du Pont, FR 1392098, 1965.
44. Hoechst, DE 96853, 1897. Belvedere Chem. Ltd., US 2132454, 1935. Eastman Kodak Co., US 2446519, 1948, US 2525515, 1950.
45. Soc. des Usines Chimiques Rhône-Poulenc, FR 1559841, 1969. Frontier Chem. Co., US 3265735, 1966. ICI, GB 856366, 1960. Prochim S. A., FR 1338899, 1963. Du Pont, GB 713622, 1954.
46. D. Lui, C. Chen: "Synthesis of p-Aminophenol by Catalytic Hydrogenation of Nitrobenzene over Pt/Activated Carbon Catalyst," *Cuihua Xueboa* **17** (1996) 544 – 546.
47. Koppers Co., US 3953509, 1976.
48. Du Pont, US 198249, 1938.
49. International Nickel Ltd., NL 6814316, 1969.
50. GB 802619, 1958 (R. Debus, P. Leprince).
51. CPC Internat. Inc., US 3535382, 1970.
52. Engelhard Minerals & Chem. Corp., DE-OS 2118334, 1971. Seiko Chem. Co. Ltd., JP Kokai 75142525, 1975.
53. R. G. Benner, DE-OS 1643255, 1971.
54. M. Noel, P. N. Anantharama, H. V. K. Udupa, *J. Appl. Electrochem.* **12** (1982) no. 3, 291 – 298.
55. E. Theodoridou, D. Jannakondakis, *Z. Naturforsch., B* **36 B** (1981) no. 7, 840 – 845.

56. K. S. Udupa, G. S. Subramanian, H. V. K. Udupa, *Trans. Soc. Adv. Electrochem. Sci. Technol.* **7** (1972) no. 2, 49 – 50.
57. Q. Gao, S. Chen: "Electrolytic Reduction of p-Nitrophenol," *Huaxue Shijie* **35** (1994) 432 – 434.
58. Continental Oil, US 3338806, 1967 (W. Harwood).
59. Trimb. Ltd., AU 205525, 1957. Akt.-Ges. f. Anilinfabr., DE 205415, 1908.
60. *BIOS Report 986,* British Intelligence Objectives Subcommittees, Eng. 1946, p. 45 – 46, 412.
61. Hoechst, DE-OS 2140786, 1973; Leonhardt & Co., DE 49060, 1889.
62. Eastman Kodak Co., US 2376112, 1945.
63. Universal Oil Products, FR 1354430, 1964 (J. Levy).
64. Mitsui Toatsu Chem. Inc., JP Kokai 8120553, 1981.
65. Toaka Dyestuffs MFG. Co. Ltd., JP Kokai 7611, 722, 1976.
66. BASF, DE 1104970, 1961.
67. Bayer, DE-OS 1902418, 1970.
68. ICI, GB 1038005, 1966 (Z. Grzymalski, F. Mirek).
69. PL 46829, 1963.
70. ICI, GB 1038078, 1966.
71. Soc. des Usines Chimiques Rhône-Poulenc, FR 1564882, 1969.
72. Howard Hall & Co., DE-OS 2050943, 1971.
73. Howard Hall & Co., DE-OS 2050927, 1971.
74. Howard Hall & Co., DE-OS 2054282, 1971.
75. MacFarlan Smith Ltd., DE-OS 2103548, 1971.
76. CPC Internat. Inc., US 3876703, 1975.
77. Mitsui Toatsu Chem. Inc., JP Kokai 8081843, 1980.
78. K. Urano, K. Kawamoto, K. Hayashi, *Suishitsu Odaku Kenkyu* **4** (1981) 43 – 50.
79. R. A. Hux, Su Puon, F. F. Cantwell, *Anal. Chem.* **52** (1980) 2388 – 2392.
80. F. A. N. El-Dien, M. A. Zayed, G. G. Mohamed: "Microdetermination of Aminophenols and Cresols via Their Reactions with bromine," *Egypt. J. Chem.* **39** (1996) 115 – 128.
81. C. S. P. Sastry, K. V. S. S. Murty, *Natl. Acad. Sci. Lett. (India)* **5** (1982) 15 – 17.
82. N. H. Cnubben, B. Blaauboer, S. Juyn, J. Vervoort, I. M. Rietjens: "A Spectrophotometric Assay for the Detection of 2-Aminophenols in Biological Samples," *Anal. Biochem.* **220** (1994) 165 – 171.
83. D. J. Miner, J. R. Rice, R. M. Riygin, P. T. Kissinger, *Anal. Chem.* **53** (1981) 2258 – 2263.
84. O. Niwa, Y. Xu, H. B. Halsall, W. R. Heinemann: "Small-volume Voltammetric Detection of 4-Aminophenol with Interdigitated Array Electrodes and its Application to Electrochemical Enzyme Immunoassay," *Anal. Chem.* **65** (1993) 1559 – 1563.
85. M. A. Korany, D. Heber, J. Schnekenburger, *Talanta* **29** (1982) 332 – 334.
86. A. Mazzeo-Farina, M. A. Ionio, A. Laurenzi, *Ann. Chim. (Rome)* **71** (1981) 103 – 109.
87. L. Reio, *J. Chromatogr.* **1** (1958) 338 – 373.
88. S. C. Mitchell, R. H. Waring, *J. Chromatogr.* **151** (1978) 249 – 251.
89. A. Ya. Yakobi, L. A. Oberemok, *Khim. Promst. Ser. Metody Anal. Kontrolya Kach. Prod. Khim. Promst.* **2** (1981) 46 – 47.
90. N. T. Bernabei, V. Ferioli, G. Gamberini, R. Cameroni, *Farmaco Ed. Prat.* **36** (1981) 249 – 255.
91. A. Brega, P. Prandini, C. Amaglio, E. Pafumi: "Determination of Phenol, m-, o- and p-Cresol, p-Aminophenol and p-Nitrophenol in Urine by High-performance Liquid Chromatography," *J. Chromatogr.* **535** (1990) 311 – 316.
92. L. A. Stevenson, *IARC Sci. Publ.* **40** (1981) 219 – 228.
93. N. Tanada et al.: "Identification of Human Hair Stained with Oxidation Hair Dyes by Gas Chromatographic – Mass Spectrometric Analysis," *Forensic Sci. Int.* **52** (1991) 5 – 11.
94. N. Tanada et al.: "Demonstration of Oxidation Dyes on Human Hair," *Forensic Sci. Int.* **64** (1994) 1 – 8.
95. B. B. Tewara, A. Kamaluddin: "Interaction of o-Aminophenol and o-Nitrophenol with Copper, Zinc, Molybdenum, and Chromium Ferrocyanides," *J. Colloid Interface Sci.* **193** (1997) 167 – 171.
96. C. Moreno-Castilla, J. Rivera-Ultrilla, M. V. Lopez-Ramon, F. Carrasco-Marin: "Adsorption of Some Substituted Phenols on Activated Carbons from a Bituminous Coal," *Carbon* **33** (1995) 845 – 851.
97. L. N. Zayidullina, E. A. Korneeva, A. S. Lukyanova, *Vodosnabzh. Sanit. Tekh.* **8** (1981) 12 – 13.
98. A. M. Klibanov, B. N. Alberti, E. D. Morris, L. M. Felshin, *J. Appl. Biochem.* **2** (1980) 414 – 421.
99. U. Lendemann, J. C. Spain: "2-Aminophenol 1,6-Dioxygenase: a Novel Aromatic Ring Cleavage Enzyme Purified from Pseudomonas pseudoalcaligenes JS45," *J. bacteriol.* **178** (1996) 6227 – 6232.

100. S. Takenaka et al.: "Novel Genes Encoding 2-Aminophenol 1,6-Dioxygenase from Pseudomonas species AP-3 Growing on 2-Aminophenol and Catalytic Properties of the Purified Enzyme," *J. Biol. Chem.* **272** (1997) 14727 – 14732.
101. S. Takenaka, S. Murakami, R. Shinke, K. Aoki: "Metabolism of 2-Aminophenol by Pseudosomas sp. AP-3: Modified meta-Cleavage Pathway," *Arch. Microbiol.* **170** (1998) 132 – 137.
102. J. Beunink, H. J. Rehm: "Coupled Reductive and Oxidative Degredation of 4-Chloro-2-Nitrophenol by a Co-immobilized Mixed Culture System," *Appl. Microbiol. Biotech.* **34** (1990) 108 – 115.
103. Alberto Culver Co., US 4297098, 1980 (G. F. Dasher, T. J. Schamper).
104. Wella AG, GB 2085483, 1981 (E. Konrad, H. Mager).
105. Matsushita Electric Works Ltd., JP Kokai 8120162, 1981.
106. M. Terashima, M. Ishi, Y. Kanaoka, *Synthesis* **6** (1982) 484 – 485.
107. Ciba-Geigy AG, EP 48222, 1982 (H. O. Wirth, J. Buessing, H. H. Friedrich).
108. *Patty,* 3rd ed., **vol. 2 A,** 2443 – 2445.
109. B. Santucci, A. Cristaudo, C. Cannistraci, A. Amantea, M. Picardo: "Hypertrophic Allergic Contact Dermatitis from Hair Dye," *Contact Dermatitis* **31** (1994) 169 – 171.
110. M. Cirstea, G. Suhaciu, M. Cirje, G. Petec, A. Vacariu, *Rev. Roum. Morphol. Embryol. Physiol. Physiol.* **17** (1980) 91 – 96.
111. M. Katsumata: "An industrial Toxicological Study on Workers who Synthesize Aminophenol and Anisidine from Chloronitrobenzene," *Nippon Ika Daigaku Zasshi* **61** (1994) 590 – 601.
112. Y. Nishijima: "Effects of Tryptophan Metabolites on Brain Function. Electrocorticographical study," *Okayama Igakkai Zasshi* **104** (1992) 471 – 482.
113. J. D. Tange, B. D. Ross, J. G. G. Ledingham, *Clin. Sci. Mol. Med.* **53** (1977) 485 – 492.
114. H. Song, C. A. Lang, T. S. Chen: "The Role of Glutathione in p-Aminophenol-induced Nephrotoxicity in the Mouse," *Drug Chem. Toxicol.* **22** (1999) 529 – 544.
115. G. Kanbak, M. Inal, C. Baycu: "The Role of Free Radicals in p-Aminophenol-induced Nephrotoxicity: does Reduced Glutathione have a Protective Effect?" *Clin. Chim. Acta* **252** (1996) 61 – 71.
116. C. A. Crowe, I. C. Calder, N. P. Madsen et al., *Xenobiotica* **7** (1977) 345 – 356.
117. J. V. Rutkowski, V. H. Ferm, *Toxicol. Appl. Pharmacol.* **63** (1982) 264 – 269.
118. R. Yoshida et al.: "Mutagenicity of p-Aminophenol in E. coli WP2uvrA/pKM101 and its Relevance to Oxidative DNA Damage," *Mutat. Res.* **415** (1998) 139 – 150.
119. N. K. Hayward, M. F. Lavin, P. W. Craswell, *Biochem. Pharmacol.* **31** (1982) 1425 – 1429.
120. G. Kirchner, U. Bayer, *Hum. Toxicol.* **1** (1982) 387 – 392.
121. J. B. Majeska, H. E. Holden: "Genotoxic Effects of p-Aminophenol in Chinese Hamster Ovary and Mouse Lymphoma Cells: Results of a Multiple Endpoint Test," *Environ. Mol. Mutagen.* **26** (1995) 163 – 170.
122. S. M. Sicardi, J. L. Martiarena, M. T. Iglesias: "Mutagenic and Analgesic Activities of Aniline Derivatives," *J. Pharmaceut. Sci.* **80** (1991) 761 – 764.
123. *Colour Index,* 3rd ed., The Society of Dyers and Colorists, Bradford, Yorkshire, England, vol. 4, 1971, p. 4001 – 4689, 4691 – 4863; vol. 6, 1975, p. 6391 – 6404, 6407 – 6410.
124. *Beilstein,* **13,**354 – 549; **13 (2),** 164 – 308.
125. J. Ransom, *Am. Chem. J.* **23** (1900) 34.
126. F. Lees, F. Shedden, *J. Chem. Soc.* **83** (1903) 756.
127. J. Schwyzer: *Die Fabrikation pharmazeutischer und chemisch-technischer Produkte,* Springer Verlag, Berlin 1931, p. 203.
128. General Aniline & Film Corp., FR 1321689, 1963.
129. E. Profft, *Dtsch. Chem. Z.* **2** (1950) 194.
130. J. R. Reasenberg, E. Lieber, G. B. L. Smith, *J. Am. Chem. Soc.* **61** (1939) 384.
131. R. Meldola, G. H. Woolcott, E. Wray, *J. Chem. Soc.* **69** (1896) 1321.
132. R. Nodzu et al., *Yakugaku Zasshi* **79** (1959) 1378.
133. F. Chen et al.: "Metal-mediated Oxidative DNA Damage Induced by Nitro-2-aminophenols," *Cancer Lett.* **126** (1998) 67 – 74.
134. H. Hofer, F. Jakob, *Ber. Dtsch. Chem. Ges.* **41** (1908) 3187 – 3199.
135. *P.B. Report 12272* U.S. Department Commission Office Technology Service, Washington, D.C., 1945, p. 39 – 46.
136. W. W. Hartmann, H. L. Silloway, *Org. Synth.* **25** (1945) 5 – 7.
137. S. Kubota, K. Nara, S. Onishi, *Yakugaku Zasshi* **76** (1956) 801.

138. *P. B. Report 74197* U.S. Department Commission Office Technology Service, Washington, D.C., 1947, p. 695 – 697.
139. M. P. Slyusar, *Nauchn. Tr. Ukr. Nauchno Issled. Inst. Gig. Tr. Profzabol.* **27** (1958) 103; *Chem. Abstr.* **54** (1960) 13435 h.
140. *P. B. Report 85172* U.S. Department Commission Office Technology Service, Washington, D.C., 1948, p. 242.
141. G. Egerer, *J. Biol. Chem.* **35** (1918) 565.
142. H. Hubner, *Justus Liebigs Ann. Chem.* **210** (1881) 328 – 396.
143. Z. C. Heng, J. Nath, X. Liu, T. M. Ong: "Induction of Chromosomal Aberrations by 2,4-Dichloro-6-aminophenol in Cultured V79 cells," *Teratogen. Carcinogen. Mutagen.* **16** (1996) 81 – 87.
144. Z. C. Heng, T. Ong, J. Nath: "In Vitro Studies on the Genotoxicity of 2,4-Dichloro-6-Nitrophenol Ammonium (DCNPA) and its Major Metabolite," *Mutat. Res.* **368** (1996) 149 – 155.
145. W. E. Bradt, *J. Phys. Chem.* **34** (1930) 2711 – 2718.
146. E. A. Braude, R. P. Linstead, K. R. H. Wooldridge, *J. Chem. Soc.* (1954) 3586 – 3595.
147. T. Nelson, H. C. S. Wood, A. G. Wylie, *J. Chem. Soc.* 1962, part I, 371 – 372.
148. L. Gattermann, *Ber. Dtsch. Chem. Ges.* **26** (1893) 1844 – 1856.
149. W. Hemilian, *Ber. Dtsch. Chem. Ges.* **8** (1875) 768.
150. H. Pomeranz, DE 269542, 1914.
151. Friedrich Bayer & Co., DE 75260, 1894.
152. K. Udupa, G. Subramanian, H. Udupa, *Bull. Acad. Pol. Sci. Ser. Sci. Chim.* **9** (1961) no. 6, 419.
153. Eastman Kodak Co., US 2525515, 1950 (F. R. Bean).
154. Betz Laboratories Inc., US 4279767, 1980 (J. A. Muccitelli).
155. H. E. Fierz-David, L. Blangey: *Grundlegende Operationen der Farbenchemie,* Springer Verlag, Wien 1943, p. 105.
156. A. M. Popow, *Anilinokras. Promst.* **3** (1933) 391; *Chem. Zentralbl.* 1935, I, 226.
157. W. J. Close, B. D. Tiffany, M. H. Spielmann, *J. Am. Chem. Soc.* **71** (1949) 1265.
158. *BIOS Final Report* **986** M. O. de Vries, *Recl. Trav. Chim. Pays-Bas* **28** (1909) 288.
159. Akt.-Ges. f. Anilinf., DE 137956, 1902.
160. *P. B. Report 74051,* U.S. Department Commission Office Techn. Service, Washington, D.C., 1946, p. 281.
161. A. Cahours, *Justus Liebigs Ann. Chem.* **74** (1850) 301. C. Niemann, J. F. Mead, A. A. Benson, *J. Am. Chem. Soc.* **63** (1941) 609.
162. Du Pont, US 1998794, 1932.
163. Bayer, DE 186655, 1905.
164. *P. B. Report 74197* U.S. Department Commission Office Techn. Service, Washington, D.C., 1947 610. A. Faust, H. Müller, Justus Liebigs Ann. Chem. 173 (1874) 315.
165. Fabr. de Prod. Chimiques de Thann, DE 98637, 1898. Amer. Cyanamid, US 2459002, 1944.
166. IG Farbenind., DE 552267, 1932.
167. A. Einhorn, B. Pfyl, *Justus Liebigs Ann. Chem.* **311** (1900) 34 – 66.
168. K. Auwers, H. Rohrig, *Ber. Dtsch. Chem. Ges.* **30** (1897) 988 – 998.
169. G. W. Raiziss, J. L. Gavron, *J. Am. Chem. Soc.* **43** (1921) 582 – 585.
170. G. W. Raiziss, B. C. Fisher, *J. Am. Chem. Soc.* **48** (1926) 1323 – 1327.
171. A. Leonhardt & Co. DE 49060, 1889.
172. Badische Anilin- und Soda-Fabrik, DE 121683, 1901.
173. Ges. F. Chem. Ind., DE 44792, 1888.
174. M. L. Crossley, P. F. Dreisbach, C. M. Hoffmann, R. P. Parker, *J. Am. Chem. Soc.* **74** (1952) 573 – 578.
175. Badische Anilin- und Soda-Fabrik, DE 44002, 1888.
176. BASF, DE 48151, 1889.
177. Eastman Kodak Co., US 2376112, 1942.
178. A. Calm, *Ber. Dtsch. Chem. Ges.* **16** (1883) no. 2, 787.
179. BASF, DE 46869, 1888.
180. J. T. Sheehan, *J. Am. Chem. Soc.* **70** (1948) 1665 – 1666.
181. H. Erlenmeyer, B. Prijs, E. Sorkin, E. Suter, *Helv. Chim. Acta* **31** (1948) 988 – 992.
182. Industrial Dyestuff Co., US 1993253, 1934, US 2101749, US 2101750, 1937.
183. K. Fricker, DE 449047, 1927.
184. Akt.-Ges. f. Anilin-Fabr., DE 205414, 1908.
185. E. Merck AG, DE 260234, 1913.
186. R. N. Harger, *J. Am. Chem. Soc.* **41** (1919) 270 – 276.
187. J. DeJonge, R. Dijkstra, *Recl. Trav. Chim. Pays-Bas* **68** (1949) 426 – 429.
188. H. Wieland, *Ber. Dtsch. Chem. Ges.* **43** (1910) 712 – 728.
189. F. G. Bordwell, P. J. Boutan, *J. Am. Chem. Soc.* **78** (1956) 87 – 91.
190. J. E. Bright, T. C. Marrs, *Arch. Toxicol.* **50** (1982) 57 – 64.

191. T. C. Marrs, J. E. Bright, D. W. Swanston, *Arch. Toxicol.* **50** (1982) 89 – 92.
192. R. Elbers, S. Soboll, H. G. Kampffmeyer, *Biochem. Pharmacol.* **29** (1980) 1747 – 1753.
193. P. Jansco, L. Szinicz, P. Eyer, *Arch. Toxicol.* **47** (1981) 39 – 45.
194. P. Eyer, H. G. Kampffmeyer, *Biochem. Pharmacol.* **27** (1978) 2223 – 2228.
195. P. Eyer, H. Gaber, *Biochem. Pharmacol.* **27** (1978) 2215 – 2221.
196. H. Fierz-David, W. Kuster, *Helv. Chim. Acta* **22** (1939) 82 – 112.
197. Warner-Lambert Pharmaceutical Co., US 2998450, 1961.
198. A. L. LeRosen, E. D. Smith, *J. Am. Chem. Soc.* **70** (1948) 2705 – 2709.
199. V. R. Olsen, H. B. Feldman, *J. Am. Chem. Soc.* **59** (1937) 2003 – 2005.
200. F. Stern, DE 453577, 1927 (M. Bergmann).
201. H. N. Morse *Ber. Dtsch. Chem. Ges.* **11** (1878) 232 – 233.
202. Nepera Chemical Co., US 3341587, 1967 (B. Duesel, G. Wilbert).
203. J. H. Burckhalter, F. H. Tendick, E. M. Jones, P. A. Jones, W. F. Holcomb, A. L. Rawlins, *J. Am. Chem. Soc.* **70** (1948) 1363 – 1373.
204. D. E. Pearson, K. Carter, C. M. Greer, *J. Am. Chem. Soc.* **75** (1953) 5905 – 5908.
205. C. F. Boehringer & Soehne GmbH, DD 1259344, 1968 (H. Staudinger, V. Ullrich).
206. *BIOS Final Report* 986.
207. J. Schwyzer: *Die Fabrikation pharmazeutischer und chemisch-technischer Produkte,* Springer Verlag, Berlin 1931, p. 204.
208. Dow, US 1932653, 1928.
209. Monsanto Chemicals Co., US 2887513, 1959 (C. M. Eaker, J. R. Campbell).
210. P. K. Smith in: *Acetophenetidin,* a Monograph, Interscience, New York 1958.
211. Du Pont, US 2198249, 1938.
212. H. Weil, M. Traun, S. Marcel, *Ber. Dtsch. Chem. Ges.* **55** (1922) 2664.
213. H. Vater, *J. Prakt. Chem.* **29** (NF) (1884) 286 – 299.
214. R. Medola, H. S. Foster, R. Brightman, *J. Chem. Soc.* **111** (1917) 551 – 553.
215. L. Galatis, *Helv. Chim. Acta* **4** (1921) 574 – 579.
216. B.-N. Li, T. Kemeny, J. Sos, *Acta Med. Acad. Sci. Hung.* **19** (1963) suppl. 19, 111.

Aminoplastics → Amino Resins

Amino Resins

Hans Diem, BASF Aktiengesellschaft, Ludwigshafen/Rh., Federal Republic of Germany

Günther Matthias, BASF Aktiengesellschaft, Ludwigshafen/Rh., Federal Republic of Germany

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Raw Materials 3
5. Production 5
5.1. Principle Aspects 5
5.1.1. Hydroxymethylation 6
5.1.2. Condensation 6
5.1.3. Principal Components of an Amino Resin Solution 7
5.2. Production Processes 7
5.2.1. Batchwise Production 7
5.2.2. Continuous Production 10
5.2.3. Production of Special Products and Foams 11
6. Environmental Protection 12
7. Types of Resins and Their Properties 13
7.1. Urea Resins (UF Resins) 13
7.1.1. Pure Urea – Formaldehyde Resins . . 13
7.1.2. Mixed Condensates 14
7.1.3. Condensation of Urea with Other Aldehydes 14
7.1.4. Resins from Urea Derivatives 14
7.1.5. Modified Urea Resins 15
7.2. Melamine Resins 17
7.2.1. Unmodified Melamine Resins 17
7.2.2. Modified Melamine Resins 17
7.3. Urethane Resins 18
7.4. Cyanamide and Dicyanodiamide Resins 18
7.5. Sulfonamide Resins 18
7.6. Aniline Resins 19
8. Analysis 19
9. Storage and Shipment 20
10. Uses 21
11. Economic Aspects 26
12. References 27

1. Introduction

Amino resins are condensates formed when carbonyl compounds react with compounds containing amino, imino, or amide groups, liberating water. Before they have hardened, these products consist principally of oligomers, which are also called prepolymers.

A very large number of carbonyl- and nitrogen-containing compounds yield condensates of relatively low molecular weight. The type of interaction is determined by the functional groups. Among such compounds, formaldehyde, urea, and melamine have become the most important, especially because of their relatively low prices. These products are obtained by large-scale production in modern industrial plants. The amino resins offer technical advantages in many applications.

History. The history of amino resins began with the synthesis of the two main raw materials, urea and formaldehyde; the former was prepared by F. Wöhler in 1824, and the latter by A. Butlerov in 1859. However, the first investigations of urea – formaldehyde (UF) resins were not carried out until about 1880. The first reaction products were reported by B. Tollens (1884) and C. Goldschmidt (1887), and the basic chemistry of the amino resins was established about 1908 [14].

F. Pollack and K. Ripper (1921) synthesized a clear transparent material, which they called "Pollopas." This product never became commercially important, but it did give rise to a number of experiments. The first commercial products, manufactured by E. C. Rossiter of British Cyanides Co. in 1924 and 1925, led to molding materials still in use today.

The melamine resins were introduced in Germany by Henkel in 1935 [15]. Their properties are similar to those of the urea – formaldehyde resins, but the resins are of higher quality. For example, they are chemically stable after the resin has hardened and become insoluble.

An ever-increasing number of mixtures of urea – formaldehyde and melamine – formaldehyde resins are produced for the most diverse

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_115

uses. Lower costs and the better performance of the products made from these mixtures account for their popularity.

Amino resins find their major use as adhesives. They are used in the production of plywood, waferboard, strandboard, and particle board. Other types are employed for the production of laminated beams and parquet flooring, and for furniture assembly. Amino resins are also used to strengthen paper, to produce molding compounds and the flexible backing of carpets, as leather auxiliaries and soil conditioners, and for protective surface coatings.

2. Physical Properties

Amino resins are available commercially as concentrated solutions or solids, e.g. powders. The solutions generally contain water as a solvent, but in the surface-coating resins the solvents are alcohols. The solutions range from colorless and clear to milky and cloudy and are tacky to the touch. Their viscosities are substantially higher than that of water, from 20 to 70000 mPa · s – in general from 200 to 2000 mPa · s – measured at 20 °C. The amino resins can be odorless or, depending on the content of free formaldehyde, have a more or less sharp odor. The densities of the solutions at the commercially available concentrations of 55 – 80 % solids are from 1.23 to 1.31 g/cm^3. An important property of the resins is the amount of water that can be added without precipitating solids. At relatively low temperatures, i.e., below about 20 °C, pronounced cloudiness and an unusually large increase in viscosity are observed in many cases.

Some resin solutions exhibit the phenomenon of pseudoplasticity (thixotropy), which may or may not be desirable. Thixotropy means that the viscosity decreases as the shearing force increases [16], [17].

In powder form amino resins are white. In the spray-dried, commercially available form they have a particle size generally of 15 – 70 µm, but always < 200 µm. This particle size is a result of the spray-drying process [18]. The bulk density of the commercially available products is 0.5 – 0.8 kg/L. The powder resins dissolve in water to give milky liquids. Flourlike additives can be mixed with some resins, which then assume the color of the additive to a greater or lesser extent.

3. Chemical Properties

Even when they are prepared from only two components, the amino resins are mixtures of various substances. For example, in the case of urea – formaldehyde resins the major components of the mixture are various hydroxymethyl compounds and oligomers composed of different combinations of basic units and substituted basic units. These oligomers are of different sizes. As an approximation, the basic unit can be described as methylene – urea (**1**), which can carry various substituents at the nitrogen atoms:

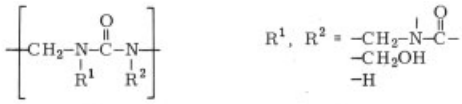

1

Hence, it is not possible to attribute specific chemical properties to these substances, as in the case of a pure compound.

However, some chemical properties are common to all or most amino resins.

The chemical properties given below are important with regard to processing the resins. The equations given for urea – formaldehyde also apply in principle to other reactants employed for these classes of products.

– Formaldehyde is slowly split off from hydroxymethyl compounds and undergoes addition reactions with other amino groups in accordance with the equilibrium equation [19], [20]:

$$H_2N\text{–}CO\text{–}NH_2 + CH_2O \rightleftharpoons H_2N\text{–}CO\text{–}NH\text{–}CH_2OH$$

– Hydroxymethyl compounds undergo slow condensation in the absence of heat or react with urea to give methylene compounds, such as methylene – diurea or methylene – oligourea:

$$H_2N\text{–}CO\text{–}NH\text{–}CH_2OH + H_2N\text{–}CO\text{–}NH_2 \longrightarrow H_2N\text{–}CO\text{–}NH\text{–}CH_2\text{–}NH\text{–}CO\text{–}NH_2 + H_2O$$

$$2\,H_2N\text{–}CO\text{–}NH\text{–}CH_2OH \longrightarrow H_2N\text{–}CO\text{–}NH\text{–}CH_2\text{–}NH\text{–}CO\text{–}NH_2 + CH_2O + H_2O$$

These reactions are responsible for the gradual increase in viscosity that amino resins undergo during storage [21].

– It is assumed that bis-methylene – urea ethers (**2**) are formed [13, p. 324], [20], [22], but amino resin solutions contain only small amounts of these ethers. Under certain production conditions cyclic ethers, so-called urones (**3**), also occur.

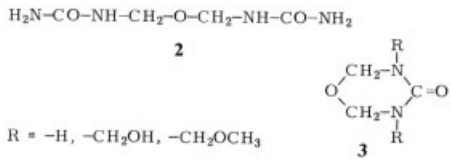

– Amino resin solutions generally are rendered slightly alkaline, i.e., pH 8 – 9, in order to stabilize them, but the pH slowly decreases as formic acid is formed.

These reactions, especially condensation with the formation of methylene bridges, are accelerated by acid catalysts and elevated temperatures. Hence, in processing amino resins, acids or acid-forming substances are important as hardeners that accelerate the condensation. Alkali-catalyzed condensation [13, pp. 331 – 332], [23] is of no practical importance.

The hydroxymethyl groups can be converted with alcohols to the corresponding ethers. This reaction is utilized in surface-coating resins and to a lesser extent in glues for particle board and in textile and paper auxiliaries.

The free formaldehyde, which is present in concentrations of from 0.1 % to about 2 %, is of practical importance in many cases because it reacts with ammonium salts of acids to give hexamethylenetetramine and the free acid, which acts as a hardener.

Resin powders also condense over time to give relatively high molecular mass products, but in this case the reaction is slower by a factor of at least five.

4. Raw Materials

The principal carbonyl component is formaldehyde, but furfural and glyoxal are employed also. Only formaldehyde is important in industrial production. Other aldehydes are employed, but these play a very minor role.

The most important amine-containing components are urea and melamine, but thiourea, guanidine, cyanamide, and dicyanodiamide, and to a lesser extent acid amides, guanamines, and aniline, are employed also. Further components include the derivatives of these compounds, e.g., substituted ureas.

Relevant information on raw materials used to produce amino resins can be found in [1, pp. 3 – 85].

Products that are added in relatively small amounts to effect slight but often important changes in the properties are called modifiers. These include alcohols, amines, aminoalcohols, and a very large variety of other organic and inorganic substances [1, pp. 7 – 122].

Finally, it should also be mentioned that inert inorganic salts sometimes are added during the production or processing of glue resins [8, p. 34], [24], [25].

Formaldehyde, CH_2O, M_r 30.03, [*50-00-0*] (production and properties → Formaldehyde). In the production of amino resins, formaldehyde is used either in the form of solid paraformaldehyde, in aqueous solution, or as a urea – formaldehyde precondensate containing from 4 to 5 mol of formaldehyde per mole of urea. Paraformaldehyde is a mixture of hydrated polyoxymethylene molecules, with a degree of polymerization of 6 – 100 units. The melting point ranges from 120 to 170 °C, depending on the molecular mass.

$$HO-[CH_2O]_n-H \quad n = 6-100$$

The most frequently used commercial form is a 30 – 55 % aqueous solution of formaldehyde. These solutions are clear, caustic, and inflammable, have a sharp odor, and are infinitely miscible with water. Formaldehyde is present in aqueous solutions predominantly as the hydrate (HO–CH_2–OH), which is in equilibrium with its oligomers:

$$n\,CH_2(OH)_2 \rightleftharpoons HO(CH_2O)_nH + (n-1)\,H_2O$$

Depending on the method of production, the solutions are contaminated with a few hundredths of a percent of formic acid and from 0.5 to 2 % of

methanol. The content of formic acid and methanol increases during storage, as a result of the Cannizzaro reaction:

$$2\,CH_2O + H_2O \longrightarrow HCOOH + CH_3OH$$

Moreover, on long-term storage at low temperatures, paraformaldehyde separates as a hazy suspension or a precipitate. This process is retarded by the addition of stabilizers, e.g., up to 12 % of methanol or triazine derivatives.

Depending on the concentration and the methanol content, the flash point is about 60 – 85 °C. The regulations governing flammable liquids do not apply to aqueous formaldehyde solutions.

Formaldehyde solutions can be stored and transported in containers made of stainless steel, aluminum, polyester resin, or appropriately lined iron.

Finally, aqueous solutions of urea – formaldehyde precondensates also are employed. Commercially available solutions contain from 70 to 80 % of the active ingredients, calculated as urea and formaldehyde. The molar ratio of formaldehyde to urea ranges from 4 : 1 to 5 : 1, and the mass ratio is therefore 2 : 1 to 2.5 : 1. These solutions are substantially more resistant to polymer precipitation than are pure formaldehyde solutions. In the urea – formaldehyde precondensates, the major part of the formaldehyde is present in the form of urea – hydroxymethyl compounds, a smaller part as formaldehyde or its oligomers in the hydrated form, and a very small part as condensates. In this last case, the urea molecules are linked mainly via methylene bridges.

Biologic Effect. The maximum allowable concentration (MAK value) recommended by the Senate Committee of the Deutsche Forschungsgemeinschaft (German Research Organization) in the Federal Republic of Germany is 1 ppm. The corresponding values for the United States of America (given by OSHA) and Japan are 3 ppm and 2 ppm, respectively. At present, regulations vary from country to country (→ Formaldehyde, Chap. 10.).

Other Aldehydes. Other aldehydes play a minor role. *Glyoxal* [*107-22-2*] is available commercially in the form of 30 – 40 % aqueous solutions. The purity is 97 – 99 %. Depending on the production method, the solutions contain 0.1 – 0.2 % acid constituents. The color number should have a maximum value of 15 (APHA). Glyoxal is hydrated in the aqueous solution.

Urea [*57-13-6*], carbamide, diamide of carbonic acid, H_2NCONH_2, M_r 60.06 (→ Urea).

Urea is available commercially as a white crystalline product or as prills. For amino resin production, a technical-grade product is usually adequate.

It should be noted that fertilizer urea frequently is contaminated with ammonium salts, and these may act as hardeners and cause the glue mixture to solidify even in the production stage. Furthermore, the prilled urea can be mixed with sulfur or formaldehyde to keep it free flowing. Sulfur-coated urea is not recommended for resin production. The formaldehyde forms insoluble condensates on the surface of the prills, and these condensates can lead to a slight cloudiness when the prills are dissolved.

The dissolution of urea in water is an endothermic process. The enthalpy of solution (ΔH) is + 257.5 J/g, measured at 10 °C in a 20-fold amount of water.

In the production of amino resins, urea is employed in solid form, in solution, or as a urea – formaldehyde precondensate.

Thiourea [*62-56-6*], thiocarbamide, NH_2CSNH_2, M_r 76.12 (→ Thiourea).

Thiourea is a relatively expensive raw material and is therefore not very important industrially. Apart from its use in special products, it serves only as a modifier, for example, in impregnating resins. Thiourea improves the resin properties, e.g., hardness, surface gloss, and water resistance.

The melting point of thiourea is 180 °C. Its solubility in water is 9 % at 100 °C, substantially lower than that of urea.

Substituted Ureas. Cyclic ethylene ureas and propylene ureas (2-oxoimidazolidine and 2-oxohexahydropyrimidine) are employed for the preparation of textile auxiliaries. Because of their high price and their limited ability to form condensates, they are not used in large amounts for any other purposes.

Melamine [*108-78-1*], 2,4,6-triamino-1,3,5-triazine, cyanurotriamide, M_r 126.13 (→ Melamine and Guanamines). Melamine (**4**) is a white powder that sublimes on heating. The particle diameter depends on the production process and varies from 5 to 300 µm. The rate of solubility in formaldehyde solutions depends on the particle size and is an important parameter given in the information leaflets of melamine producers. The color number of a 50 % melamine – formaldehyde solution may not exceed 20 (APHA). This is important with respect to raw materials for surface coatings and to impregnating resins.

Melamine derivatives, e.g., benzoguanamine (**5**), are employed as modifiers.

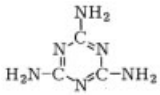

4

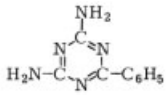

5

Modifiers. A very large number of additives from all areas of organic and inorganic chemistry have been used to modify amino resins; most of these additives are no longer important [1, pp. 107 – 122]. Aliphatic alcohols, sugars, polysaccharides, and caprolactam are additives for special purposes and are still used today.

5. Production

5.1. Principle Aspects

Amino resin production consists of two stages: *hydroxymethylation* and *condensation*. Depending on the reaction conditions, hydroxymethylation is accompanied to a greater or lesser extent by condensation. The following equations illustrate the process in a simple case, for example, for formaldehyde plus urea:
Hydroxymethylation:

$$NH_2CONH_2 + 2\,CH_2O \rightleftharpoons HOCH_2\text{–}NHCONH\text{–}CH_2OH$$

Condensation:

$$HOCH_2\text{–}NHCONH\text{–}CH_2OH + NH_2CONH_2 \longrightarrow HOCH_2\text{–}NHCONH\text{–}CH_2\text{–}NHCONH_2 + H_2O$$

Further condensation leads to longer chains of the basic unit

$$\text{–}\underset{|}{N}\text{–}CO\text{–}\underset{|}{N}\text{–}CH_2\text{–}.$$

Hydroxymethylation is carried out in a slightly alkaline or slightly acidic medium, whereas the condensation requires a more strongly acidic solution. The process for the production of the resin is continued until the resulting product is an oligomeric mixture that is still soluble or fusible. It is protected against further condensation by rendering the medium alkaline. Condensation of the resin by means of acid catalysis, i.e., cross-linking to give a substantially infusible product, is not carried out until the resin is put to use. This process is called *curing*. However, free hydroxymethyl groups are present even in this state.

The monomolecular and hydroxymethyl compounds with low degrees of polymerization and up to six urea – formaldehyde units are virtually all known and can be prepared in pure form [1, p. 20]. Aqueous solutions of the oligomers with low degrees of polymerization do not have a sufficiently long shelf life. Only when the concentrations of hydroxymethyl and NH_2 groups are reduced by partial condensation or when the products are present in solid form do they possess an industrially useful shelf life, i.e., several weeks or more.

Condensations carried out in an excess of an alcohol, such as methanol or butanol, with the addition of acid and removal of water as necessary, produce curable etherified amino resins. These are soluble in nonaqueous solvents and can be mixed with alkyd resins, epoxy resins, etc. They are used as starting materials in the production of surface coatings.

Resin glues based on urea, melamine, or both have been subjected to partial etherification to stabilize them against further condensation [26–28].

Preventing the resins from solidifying during the production process and selectively achieving certain properties – such as long shelf life, low formaldehyde content, strength, and resistance to swelling of particle board after curing – are the basic difficulties encountered. Frequently, novel properties are required on short notice by the market.

5.1.1. Hydroxymethylation

Hydroxymethylation can be carried out using either alkaline or acid catalysis, but the former is faster. The reaction mechanisms in acidic and alkaline media differ. In acids, condensation takes place at the same time. Hence, hydroxymethylation is carried out industrially in an alkaline or slightly acidic solution.

In hydroxymethylation, a semiaminal bond is formed:

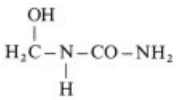

Monohydroxymethylurea

This is cleaved readily, and hydroxymethylation reactions are reversible and establish an equilibrium [32]. Although mono-, di-, and trihydroxymethylurea are formed and isolated easily, tetrahydroxymethylurea is virtually nonexistent in aqueous solution. The complete hydroxymethylation of all amino groups to form hexahydroxymethylmelamine can be carried out readily. The hydroxymethylation of urea is slightly exothermic:

$$H_2N\text{-}CO\text{-}NH_2 + CH_2O \rightleftharpoons H_2N\text{-}CO\text{-}NH\text{-}CH_2OH$$

$$\Delta H = -23\ \text{kJ/mol}$$

The ΔH data given in the literature fluctuate about this value by ± 5 kJ/mol. The enthalpy of hydroxymethylation of melamine, relative to the amount of formaldehyde, is -15 kJ/mol [29], [30].

Reaction rates have been measured frequently, but they are sensitive to many parameters [19], [29].

Industrial processes produce variable mixtures of oligomers with mixed repeat units, so the overall reaction rate is also variable and cannot be defined. Moreover, the important resin properties have not been correlated with the individual reaction rates. In general, the hydroxymethylation rate increases with increased concentration, temperature, and pH. Substituents also have an effect; adjacent electron-withdrawing groups reduce the reaction rate, as do bulky substituents.

5.1.2. Condensation

The condensation is catalyzed by acids. Catalysis by strong alkalis has been described also, but the resulting amino resins are not important industrially [23]. The condensation can take place by the following routes:

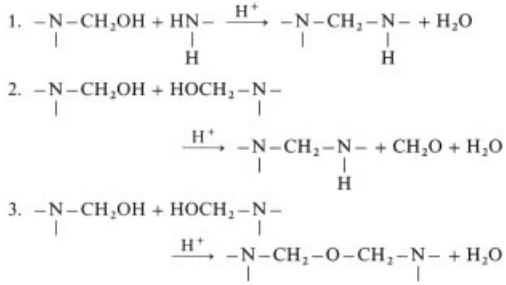

Although the presence of methylene compounds in amino resins can be established unambiguously by means of ^{13}C NMR, bis(ureamethylene) ethers can be detected only by an indirect procedure [31]. Carbon-13 NMR spectroscopy permits the elucidation of many of the structures but does not distinguish between the **C** atoms shown in boldface in (**6**) and (**7**), although their third nearest neighbors are different.

This problem also occurs in the case of melamine resins [33].

The methylene bonds between two urea molecules are very stable and can be hydrolyzed only by strong acids at elevated temperatures (see Chap. 8). The condensation reaction is slightly exothermic; the enthalpy is -28 kJ/mol [29]. The stability of the methylene-ether bond is between that of the hydroxymethyl bond and that of the methylene bond.

It is impossible to give a single reaction rate for the condensation process because there are many different starting materials (hydroxymethyl compounds) and many end products (methylene compounds).

The type of solvent also has an effect on the condensation rate [34]. The dependence of the reactivity of hydroxymethyl compounds on the chemical composition of the amino compounds can be interpreted on the basis of several qualitative rules [35].

5.1.3. Principal Components of an Amino Resin Solution

It is possible to identify the most important structural moieties of the oligomers in amino resin solutions by means of ^{1}H and ^{13}C NMR spectroscopy [22], [31], [36].

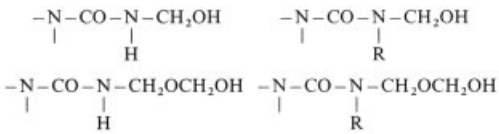

The hydrates of the formaldehyde oligomers $HO(CH_2O)_nH$, where n is about 5, can be detected along with the following hydroxymethyl structures:

Ethers, such as hydroxymethyl methyl ether, also can be identified in the spectrum by distinguishing the C atoms of the parent alcohol.

Finally, the structures (**8**) and (**9**) and ring compounds, including cyclic ethers of the urone type, also appear.

Asymmetrical **8** Symmetrical **9**

In a typical particle board glue, the ^{13}C NMR spectrum might show, for example, the following distribution of the carbon atoms of the original formaldehyde (mole fractions):

formaldehyde 0.03
hydroxymethyl compounds 0.72
methylene compounds 0.15
cyclic compounds 0.10

5.2. Production Processes

Amino resins usually are prepared in 5 – 20 m^3 batches. However, continuous processes are employed also, particularly for the production of resin glues. The patent literature describes processes where a particular stage is operated continuously while other stages are operated batchwise.

Because the reaction rate and the composition of the end product are influenced by a large number of parameters, the reaction must be controlled carefully. Important reaction parameters are:

1) purity or composition of the starting materials
2) molar ratio in each of the reaction stages
3) type and amount of modifier
4) reactant concentrations
5) the pH value at each stage of the reaction
6) temperature at each stage of the reaction
7) type and concentration of catalyst
8) amount of buffer salt
9) reaction time in each stage

The large number of reaction parameters and resultant adaptations of the process lead to a very extensive body of relevant literature, particularly patent literature. Only the important processes are described here, but a comprehensive review can be found in [1, pp. 101 – 105].

5.2.1. Batchwise Production

The batchwise procedure is the most widely used method for the industrial production of amino resins, but it has the disadvantage of a relatively small production capacity. Nevertheless, this process permits an extensive variety of products and frequent changes of product.

The reactions are carried out in stirred kettles, in two or more stages, at 70 – 100 °C. In the first stage, which is carried out in a slightly acidic to alkaline solution, the principal reaction is hydroxymethylation. In subsequent stages, the condensation is carried out, water is split off, and higher molecular weight products of increasing viscosity are obtained. The duration of the condensation depends on the properties desired in the product. In general, the products are concentrated to 60 – 70 % aqueous solutions by evaporation. The starting materials, urea or melamine, are employed in powder form. Urea also can be used in solution. Formaldehyde is employed as a 37 – 55 % aqueous solution. In a few special cases, formaldehyde is used in the solid form, paraformaldehyde. Finally, it is possible to employ urea – formaldehyde precondensates in concentrations of not more than 80 % by mass of the active ingredients.

Figure 1 shows a diagram of a plant for the batchwise production of aqueous solutions.

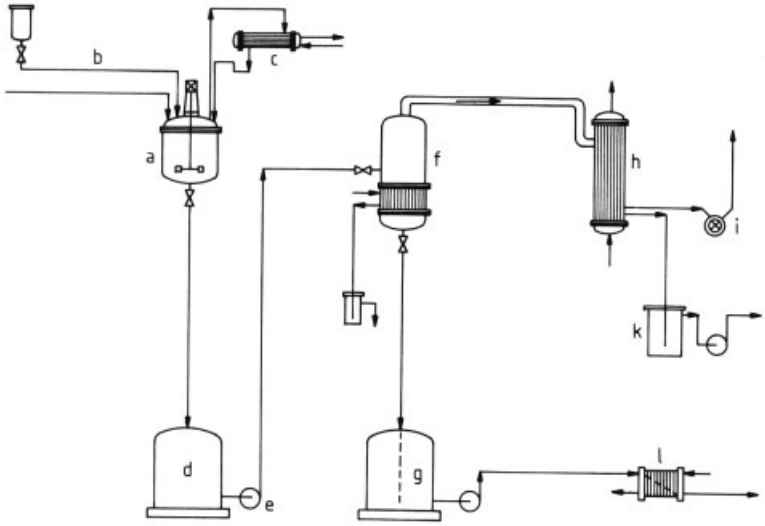

Figure 1. Plant for the batchwise production of aqueous amino resin solutions
a) Stirred kettle; b) Raw material feed; c) Reflux condenser; d) Temporary container; e) Pump; f) Vaporizer; g) Container for finished product; h) Condenser; i) Vacuum pump; k) Vessel for condensed vapors; l) Cooler

Aqueous Solutions. The reactor consists of a kettle equipped with heating and cooling devices, a stirrer, inlets for feeding in the raw materials and for measuring devices, a valve at the bottom for emptying the kettle, and a manhole for cleaning.

The material used is mainly stainless steel, material No. St. 1.4541 or St. 1.4571. Internal coils or external, welded-on, semicircular pipes provide for heating with steam or cooling with cold water. Stirring is carried out by means of a powerful unit consisting of disk or anchor stirrers. Nozzles are welded on in order to meter in the raw materials and to permit measurement of the temperature and pH.

When the condensation is complete, the product is evaporated. This is carried out in the stirred kettle, or more economically in a single- or multistage tubular vaporizer, with the pressure reduced to protect the resin. The quality of the resin is then compared to its specifications and it is discharged into a storage container.

The product tanks and the pipelines can be of aluminum or steel because the medium is slightly alkaline. It is advantageous to provide lined iron tanks.

A typical, simple resin glue for particle board production is made as follows [37], [38]: 60 parts (by mass) of solid urea are introduced into 158 parts of a 40 % formaldehyde solution that has been rendered neutral. The mixture is kept at 80 °C for 10 min, during which time hydroxymethylation takes place. The pH is then brought to 4 – 5 by adding a 10 % formic acid solution. The condensation causes the temperature to rise to 97 – 100 °C within a few minutes. This temperature is maintained for about 30 min or until a mixture of 1 part of reaction solution and 5 parts of water gives a white precipitate at 15 °C (sharp endpoint). The reaction mixture is then brought to pH 7 with a 25 % sodium hydroxide solution. This terminates the condensation. A further 16 parts (by mass) of urea are then added, so that the molar ratio of formaldehyde to urea in the end product is 1.65. The resulting solution has a solids content of 51 %. The resin solution is evaporated under reduced pressure to the commercial concentration of 66 %.

This 66 % solution of amino resin forms a milky solution and has a viscosity of about 500 mPa · s (at 20 °C) and a density of 1.290 g/cm^3.

A typical resin for the impregnation of paper webs used in producing decorative laminates can be made as follows [12, p. 258]:

126 parts (by mass) of melamine are introduced into a stirred solution of 120 parts of a 40 % formaldehyde solution and 70 parts of

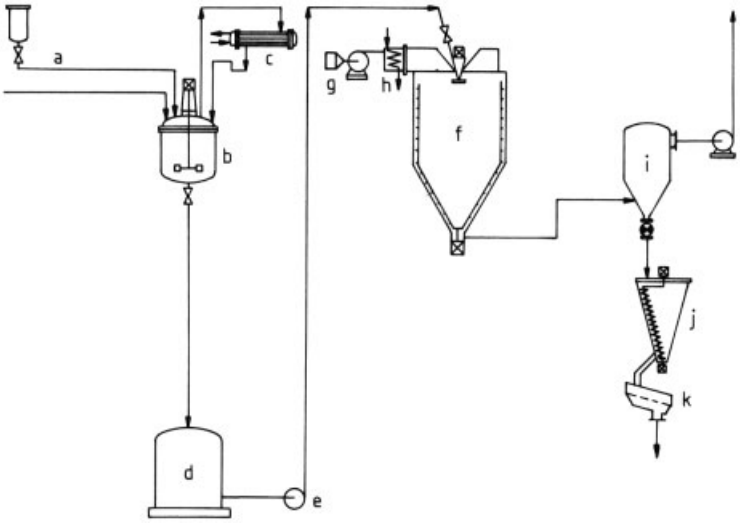

Figure 2. Plant for the production of amino resin powders by spray drying [18, p. 501]
a) Starting material feed; b) Stirred kettle; c) Reflux condenser; d) Temporary container; e) Pump; f) Spray drier; g) Blower; h) Air heater; i) Filter; j) Mixing bin; k) Vibrating sieve

water at room temperature. The reaction mixture is brought to pH 9 by adding aqueous sodium hydroxide solution and is then heated very rapidly, i.e., in 20 – 30 min for an industrial-scale mixture, to a temperature of 100 °C. After the melamine has dissolved, a process that takes place exothermally, sodium hydroxide solution is added continuously to keep the pH value of the reaction mixture at 8.5 – 8.8 throughout the condensation. Condensation is carried out, with gentle refluxing and stirring, until the reaction mixture is compatible with water when mixed in the ratio 1 : 1.5. This is the case when a sample of the reaction mixture exhibits slight cloudiness when water, at 1.5 times the amount of solution, is added at 20 °C. The reaction solution is then cooled rapidly and brought to pH 9.5 – 10 at room temperature by the addition of sodium hydroxide solution. The resultant impregnating resin solution has a solids content of about 55 % and is usually clear.

Amino Resin Powders. In the production of amino resin powders, an aqueous resin solution is first produced (Fig. 2) and then fed into a spray drier, where it is atomized by a spray disk or nozzle. The resultant droplets are dried in the stream of hot gas produced by heating air indirectly in a heat exchanger or by mixing hot exhaust gas with air. The powder is collected in a tower and in cyclone separators or filters. It passes via a vibrating sieve into a mixing bin and then pours into sacks or drums.

Etherified Products Containing Solvent. Products of this type, which are used as starting materials for making surface coatings, can be produced in a reactor like that described in the production of aqueous solutions of amino resins (Fig. 3). The reactor must be equipped with a sufficiently large heat exchanger to vaporize the mixture of water and excess alcohol, e.g., butanol. Downstream from the heat exchanger is a water separator. The separated aqueous phase still contains from 1 to 2 % formaldehyde and 20 % alcohol and therefore must be worked up. At the appropriate alcohol content, this solvent phase is recycled to the production process.

A butylated melamine – formaldehyde resin for surface coatings can be produced in the following way [11, p. 69]:

60 parts (by mass) of melamine, 60 parts (by mass) of formaldehyde (calculated at 100 %) in the form of a 40 % aqueous solution, and 220 parts (by mass) of butanol are refluxed for 30 min in a stirred kettle equipped for heating. Distillation is then carried out until water no longer evaporates. Then a further 20 – 25 parts (by mass) of anhydrous butanol are distilled.

The mixture is cooled, the resin content is determined, and the concentration is then brought to 50 % with butanol.

5.2.2. Continuous Production

The continuous industrial production of amino resins was begun in order to meet rising demand. The disadvantages of continuous production are that the amount produced per unit time in a particular plant can vary only within relatively narrow limits because prolonged residence under otherwise constant conditions increases the proportion of molecules with high degrees of condensation. Changing product is not simple either. If the plant is not cleaned out beforehand, a product having a composition between the old product and the new will be formed for a time. For this reason, the continuous production process is employed where the variety of products is restricted to a relatively narrow range. On the other hand, continuous procedures give a very uniform product quality.

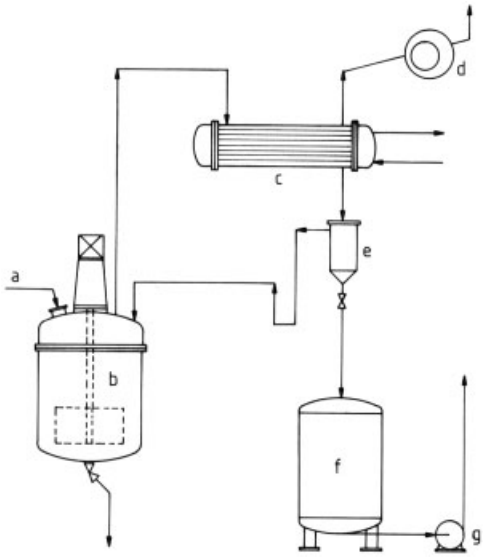

Figure 3. Plant for the batchwise production of etherified amino resins
a) Raw material feed; b) Stirred kettle; c) Condenser; d) Vacuum pump; e) Separating vessel (water separator); f) Product vessel; g) Pump

The patent literature describes a large number of continuous processes. They differ only in process technology for the product flow. In general, the differences include variations in the temperature, pH, concentration, or modifiers. The process and product flow otherwise remain unchanged. In many cases the publications only describe laboratory- or pilot-scale plants that have not been scaled up for industrial use. Of the continuous processes that have been described, only those with basic differences in type are discussed here.

Aqueous Solutions. The apparatuses employed for the continuous production of aqueous amino resin solutions are tube reactors, stirred kettle cascades, or combinations of the two. The danger of incrustation by highly condensed, insoluble products is greater in the continuous process than in the batch process. In the latter, the residues are dissolved by fresh resin solutions; this is not the case in the continuous process because very little backmixing occurs.

A continuous two-reactor process was described by the Girdler Corp., United States of America, in 1943 [39]. The reactants were mixed thoroughly in the first reactor and heated to about 150 °C under pressure. In the second stage, the solution under pressure was maintained at this temperature and condensation was carried out. The pressure of the reactor was then released, with the result that volatile constituents, especially water, escaped from the resin and the solution cooled.

According to a 1949 patent granted to Sherwood Paints, UK, the reaction is carried out in a packed column [40] and water vapor is expelled by means of a countercurrent gas stream.

A process of Allied Chemical Corp., United States of America (1951), operates in a similar manner and employs a tubular coil as the reactor [41]. The resin solution is subsequently evaporated. Another process of Allied Chemical Corp. is carried out under pressure, using injected steam (1952) [42].

According to an application by Spumalit-Anstalt, Liechtenstein, the starting materials are passed over a solid catalyst (polycarbamide resin) under a pressure of 400 kPa at 140 – 150 °C to form the resin [43].

A process of Rütgerswerke, Germany, from 1955 also uses a fixed-bed catalyst, an ion exchanger [44]. The resins obtained are used as starting materials for making molding compounds.

A Swedish application by Skanska Attifabriken AB from 1955 describes a process in which starting materials are heated separately and then sprayed into a chamber to undergo partial condensation [45]; the reaction is completed in another chamber. The advantage claimed is that none of the final products participates in the initial stages of the condensation.

An application by Du Pont, United States of America, in 1956, describes a continuous process for producing etherified formaldehyde – urea resins [46]. The stepwise reaction takes place in a cascade of two to four stirred kettles. The free alcohols (e.g., butanol, isobutanol, and others) are distilled in the last stage, and some of them are recycled.

In a Meissner process from 1966, the hydroxymethylation stage is carried out continuously and the subsequent condensation stage is operated batchwise [47]. The continuously flowing product stream from the first stage passes through a stirred kettle, which is one of a large number of kettles, so that the product flow need not be interrupted.

Other continuous processes have been worked out by Rheinpreussen AG [48] and the St. Regis Paper Co. [49], but these processes predominantly relate to phenoplasts.

A BASF process (1971) is shown diagrammatically in Figure 4 [50]. The reaction is carried out in a series of three or more stirred kettles. At the beginning of the cascade predominantly hydroxymethylation takes place. Condensation occurs later in the cascade. The reaction is terminated in the last section by the addition of sodium hydroxide solution. The resin solution passes into a temporary container and is subsequently evaporated under reduced pressure in a single- or multi-stage vaporizer. The vapors, containing about 2 % formaldehyde, are worked up and the effluent is discharged for treatment. The end product is cooled and pumped into a tank.

Other cascade processes have been described in applications of SIR, Milano, Italy, (1972) [51] and Montedison (1980) [52].

Finally, mention should be made of the process in which formaldehyde is absorbed directly in urea solutions to produce highly concentrated, formaldehyde-rich solutions that are used as starting materials for glue resins. Examples of conventional commercial products are Kaurit-Konzentrat 244 fl. or 255 fl. (BASF), Formurea 80 (Montedison), and Saruform 80 (Chemische Werke Saar Lothringen).

Stamicarbon [53] and American Cyanamid [54] have described the continuous production of concentrated, formaldehyde-rich melamine – formaldehyde solutions. These too can be used as starting materials for amino resins.

Resins in Powdered Form. Resins in powdered form also can be produced by a continuous procedure if the spray solutions, which in general are produced batchwise, are fed continuously from a temporary container into a spray tower [18].

5.2.3. Production of Special Products and Foams

In addition to impregnating resins, special products include foamed resins and resins for paper auxiliaries, label adhesives, concrete liquefiers, leather auxiliaries, carpet coatings, microcapsules, etc. Because of the relatively small individual amounts and the large variety of products, these resins are produced virtually always in batches. The principle of the production is the same as that for the manufacture of resin glues. The proprietary know-how relates to modifiers and small but important variations in the manner in which the process is carried out.

The foamed resins (→ Foamed Plastics) are an interesting type of amino resin. They are used to produce solid foams, which have a wide variety of applications. The productions of the foamed resin solution and the end product, i.e., the foam, are discussed together.

In 1927, E. C. Bayer discovered that porous, heat-insulating materials are obtained if a foam or a foam-forming substance is added to a glue solution and the mixture is foamed and allowed to harden. In 1933, a process for the production of a foamed urea formaldehyde resin was developed by I. G. Farbenindustrie. In this process, a foaming agent and air were used to produce a primary foam that itself contained the acid, e.g., phosphoric acid, for curing the amino resin. Curing begins after a few seconds and is complete after 12 – 24 h at room temperature. After drying, the foamed urea formaldehyde resins have

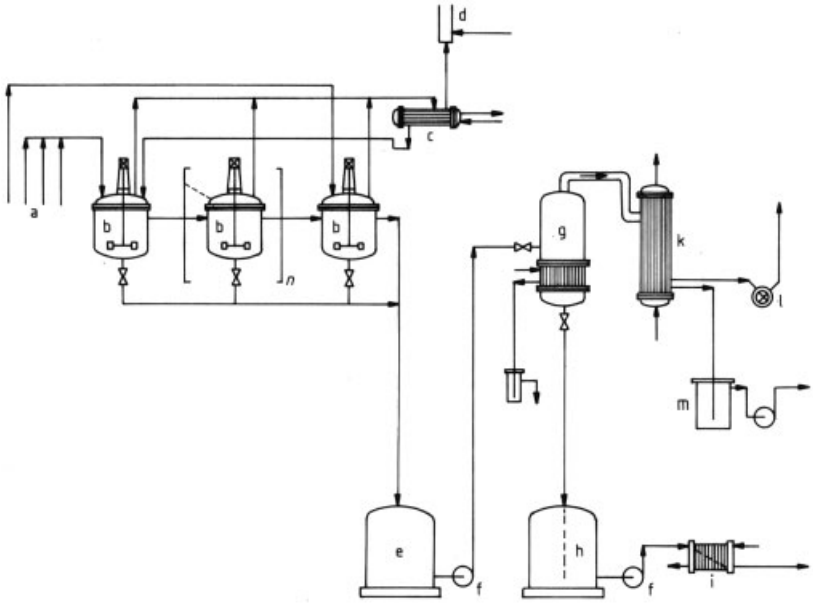

Figure 4. Plant for the continuous production of aqueous amino resin solutions [50]
a) Starting material feeds; b) Stirred kettle; c) Reflux condenser; d) Flare; e) Temporary container; f) Pump; g) Vaporizer; h) Product container; i) Cooler; k) Condenser; l) Vacuum pump; m) Container for condensed vapors

bulk densities of 5 – 70 kg/m^3. More than 60 % of the foam is in the form of open cells.

Whereas UF foams previously could be prepared only in factories of fixed location, W. Bauer in 1953 described a transportable apparatus that permitted on-site production of the foam by means of compressed air [55]. A number of other apparatuses subsequently were developed. The state of the art is reported in [56] with respect to the design of foam-producing apparatus and a novel high-performance apparatus of this type is described. The development of foam-producing apparatuses led to much wider applications: soil conditioning in agriculture, heat insulation, sound insulation, filling of mining cavities, carrier material for cleaning fluids, and covering for landfills.

6. Environmental Protection

All necessary measures possible, according to the state of the particular art, must be taken during the production and processing of amino resins to avoid air and water pollution. Regulations constantly become more stringent.

Regulations or recommendations concerning concentration of formaldehyde in rooms are:

Denmark 0.15 mg/m^3
Finland 0.15 mg/m^3
Italy 0.12 mg/m^3
Netherland 0.12 mg/m^3
Sweden 0.48 mg/m^3
Switzerland 0.24 mg/m^3
Germany 0.10 mg/m^3

Effluents [57]. The condensed vapors or effluents from amino resin production may not be introduced into water without biologic treatment. It should be further ensured that these effluents cannot penetrate into ground water.

Waste waters or partially hardened residues inevitably result when machines and containers are cleaned. The amount of wash water or solvent should be kept as small as possible. If the waste water contains predominantly urea or formaldehyde, it can be reused in the production process. Otherwise, it is eliminated, for example, by incineration.

The biologic treatment of industrial effluents from amino resin production is carried out pri-

marily because of the formaldehyde, which is toxic. The usual effluents contain formaldehyde in concentrations that can be degraded readily by bacteria, which are always present. In general, a critical factor in the biologic treatment of any effluent is the concentration in which the various toxic components are present.

Air. The treatment of waste gas by washing with water, catalytic combustion, or thermal combustion is used in industry. The waste gases should be very concentrated to minimize treatment costs.

Solid Wastes. Amino resins can be solidified by hardening them with ammonium chloride or phosphoric acid, or they can be absorbed in sawdust or sand. In this form, they may be deposited in a landfill in accordance with the relevant provisos and legal requirements.

7. Types of Resins and Their Properties

In addition to the simple types of resin that are produced in the largest amounts, i.e., the urea – formaldehyde and melamine – formaldehyde resins, other types of resin with different properties have been developed. These are modifications, i.e., versions of a resin obtained by use of relatively small amounts of additive or by changing the aldehyde or amino component. However, even where the composition is fixed the resin type can be varied by changing the reaction procedure. This leads to different distributions of oligomers or to different substitutions of the repeat units. The distribution of hydroxymethyl groups is the determining feature.

7.1. Urea Resins (UF Resins)

All the urea – formaldehyde condensates [*9011-05-6*] containing the repeat unit **1** (see Chap. 3), from monomolecular product to high molecular weight ones, are useful substances. In the past few years, a number of discoveries have been made concerning the composition of UF resins. Systematic investigations into the process of condensation and the structural composition had long been impossible because of the lack of suitable analytical methods. Industrial development took place largely on an empirical basis. However, using a combination of analytical methods, e.g., ^{1}H NMR spectroscopy, ^{13}C NMR spectroscopy, C, H, N, O elementary analysis, determination of hydroxymethyl groups, and determination of total formaldehyde, it is possible in principle to determine the structural units and their concentrations [22]. However, the relationship between the technical properties of the amino resins or the materials produced from these resins and the distribution of the structural units is still unclear in most cases, so development is still largely empirical.

7.1.1. Pure Urea – Formaldehyde Resins

Important macroscopic parameters are the molar ratio of formaldehyde to urea, the viscosity, and the gelling time. Urea – formaldehyde condensates having a molar formaldehyde : urea ratio of 1 usually give insoluble polymethylene – ureas that do not possess any adhesive properties. If, however, the molar ratio is reduced from a relatively high value gradually or stepwise in the course of the reaction, the resultant resin possesses adhesive properties. These properties improve as the molar ratio increases, but the undesirable release of formaldehyde increases simultaneously.

The highly viscous resin solutions obtainable by prolonged industrial condensation of urea and formaldehyde are low molecular mass products. They contain about six to eight urea units, as can be determined by gel chromatographic [21] and cryoscopic measurements and by end-group determinations. This also applies to the completely insoluble powdered or glassy condensates [58]. During storage, the concentration of methylene groups increases at the expense of the hydroxymethyl groups. The viscosity also increases, initially slowly and virtually linearly, and then, after a critical time, rather steeply. Some or all of the hydroxymethyl groups generally can be blocked by etherification, and the shelf life of the amino resin is thus extended [27], [59], [60]. In principle the same reactions take place when the resin is hardened by the user, but they are substantially more rapid.

However, a residual amount of hydroxymethyl groups remains after hardening [21]. Because they are susceptible to hydrolysis, even the cured urea resins are more water sensitive than the melamine or phenol resins.

The gelling time [11, p. 174] or reactivity of the resin is an important parameter for the user. This time can be determined relatively precisely, and is the time in which the resin changes from a fluid to a virtually immobile solid following the addition of a certain amount of hardener at a particular temperature. The reactivity increases with increasing molar ratio, content of hydroxymethyl groups, and content of free formaldehyde.

7.1.2. Mixed Condensates

Industrial amino resins contain mixtures of various condensates and cocondensates. Amino resin solutions are always polymeric mixtures having various degrees of condensation and differing distributions of functional groups. Resin-forming monomers employed as textile auxiliaries constitute an exception, but these in themselves are not resins. The term mixed condensates is rather loosely applied to those resins which are formed in the condensation of urea and formaldehyde when the reaction is carried out using substances that also form resins with formaldehyde. These are, for example, melamine, phenol, organic acid amides, sulfonamides, and carbamates. Therefore, in the presence of melamine or phenol, it is possible to prepare urea resins with improved strength and swelling properties for the manufacture of particle board, i.e., weather-resistant particle board. Modification with melamine gives resin glues that can be used to produce good-quality particle board even when the molar ratio of formaldehyde to urea is low.

Mixed resins with other resin formers, e.g., acrylic acid, acrylates [61], [62], acrylamide [1, pp. 109] and their derivatives, are also sometimes advantageous. These are also mixtures of polymers, of cocondensates, or of homologous polymers and cocondensates.

7.1.3. Condensation of Urea with Other Aldehydes

Condensation of urea with isobutyraldehyde, acetaldehyde, or crotonaldehyde is carried out industrially to obtain sparingly soluble nitrogen fertilizers, i.e., controlled release fertilizers, rather than resins [63], [64]. The condensation of urea or melamine with glyoxal gives products that can be employed as textile auxiliaries. Products obtained by the subsequent condensation of UF resins with glyoxal (e.g., DMDHEU [*1854-26-8*]) also are used for textile finishing [65], [66].

7.1.4. Resins from Urea Derivatives

Condensation with urea derivatives or substitution products are well known, but these reactions are of industrial importance only for making textile and paper auxiliaries. Reactions with aldehydes can take place depending on the hydrogen atoms still present at the nitrogen, but the ability to undergo condensation decreases sharply with the size and number of substituents.

The reaction of a symmetrically disubstituted urea derivative, e.g., *N,N′*-dimethylurea or dihydroxymethylurea dimethyl ether, with formaldehyde in an acidic medium gives cyclic urea derivatives of the urone type [67, p. 9]:

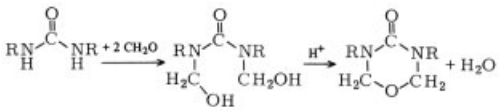

Dihydroxymethylurone dimethyl ether, bis(methoxymethyl)urone [*7388-44-5*] is used industrially to provide cellulose fabrics with a wrinkle-resistant finish. It has middling reactivity and stability to hydrolysis [68, pp. 164 – 174]. The urones do react with cellulose but form only minor amounts of self-cross-linked resins, such as the urea – formaldehyde compounds.

When S is substituted for O in urea, thiourea is obtained. The latter undergoes all of the principal condensation reactions of urea. In a weakly alkaline medium, *N*-hydroxymethyl compounds and *S*-hydroxymethyl compounds are formed. The tautomeric structure NH_2–C(SH) = NH may be assumed initially for the

latter. The thiourea – formaldehyde condensates are more water resistant than the corresponding urea condensates. Some of the thiourea condensates are assumed to be cyclic compounds [69].

7.1.5. Modified Urea Resins

There is very extensive literature on the possible methods of modifying amino resins [1, pp. 95] and only the most important modifications are discussed here.

Reactions with Alcohols. The hydroxymethyl groups can be etherified with alcohols in an acidic medium [13, pp. 334].

$$-\underset{\underset{H}{|}}{N}-CH_2-OH + ROH \overset{H^+}{\rightleftharpoons} -\underset{\underset{H}{|}}{N}-CH_2OR + H_2O$$

The equilibrium constants for the etherification of monohydroxymethylurea and dihydroxymethylurea with methanol and ethanol have been measured [59]. The enthalpy of reaction, estimated from the temperature dependence of the equilibrium constant, is $\Delta H = -6$ to -8 kJ/mol. The equilibrium constants (K) differ depending on the substitution:

	$K_{30\,°C,\,R=CH_3}$
$H_2N-CO-NH-CH_2OH + ROH \rightleftharpoons H_2N-CO-NH-CH_2OR + H_2O$	0.6
$HOCH_2-HN-CO-NH-CH_2OH + ROH \rightleftharpoons ROCH_2-HN-CO-NH-CH_2OH + H_2O$	7.6
$ROCH_2-HN-CO-NH-CH_2OH + ROH \rightleftharpoons ROCH_2-HN-CO-NH-CH_2OR + H_2O$	2.0

$$K = \frac{[\text{Ether}] \cdot [H_2O]}{[\text{ROH}] \cdot [\text{Hydroxymethyl}]}$$

As the number of C atoms in the alcohol increases, the equilibrium is shifted and the reaction rate reduced so that the etherification becomes more difficult. The ethers of the low-boiling alcohols can be transetherified with higher boiling alcohols by heating the mixture in weakly acidic solution and distilling the more volatile alcohols [13, p. 336].

Frequently only some of the hydroxymethyl groups of resin glues and impregnating resins undergo etherification with monohydric or polyhydric alcohols of less than five carbon atoms. As a result, some of the hydroxymethyl groups are protected against condensation, and the resins have a longer shelf life [1, p. 115], [70], [71]. In the starting materials for surface coatings, 20 – 100 % of the hydroxymethyl groups undergo etherification, depending on the type. The amino resins etherified with butyl groups are the most important for wood coatings, e.g., parquet coating. Where complete etherification takes place, the reaction is carried out using an entraining agent – the alcohol itself, toluene, or xylene – and the water is distilled azeotropically.

Ethers of one- to three-carbon alcohols with low molecular mass hydroxymethyl compounds of urea, ethylene urea (imidazolidinone [*28906-87-8*]), propylene urea (1,3-bis(hydroxymethyl)tetrahydro-2-(1*H*)-pyrimidinone [*65405-39-2*]), and melamine have achieved some importance in the field of textile auxiliaries [3, p. 168], [59], [68]. These compounds are precondensates having relatively low molecular masses, about 600.

Urea resins etherified with butyl groups can be employed as starting materials for making surface coatings. However, when baked alone, they give brittle and water-sensitive films. Therefore they cannot be used alone as binders. In combination with nitrocellulose, or with alkyd, epoxy, or acrylate resins, they give excellent baked finishes.

Reactions with Ammonia and Amines. Amines are very reactive toward hydroxymethyl compounds obtained from urea and formaldehyde. Ammonia reacts with the hydroxymethyl group to form an amino or imino group, eliminating water:

$$\text{-NH-CH}_2\text{OH} + NH_3 \longrightarrow \text{-NH-CH}_2\text{NH}_2 + H_2O$$

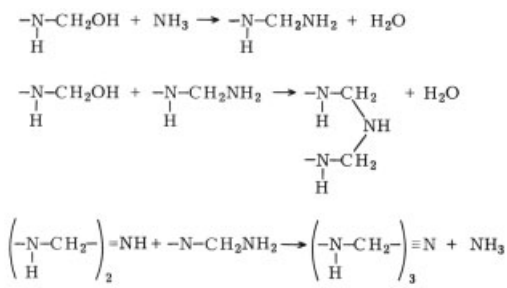

The reaction of ammonia with formaldehyde is exothermic ($\Delta H = -57$ kJ/mol) [1, p. 49] and gives hexamethylenetetramine. This reaction has been used industrially to trap formaldehyde

from waste gases, for example, from particle board production [72–74]. The reaction of urea and aldehydes with primary amines gives cyclic compounds (triazines) [75] of the type **10**, where R′ can be hydroxymethyl and its methyl and ethyl ethers.

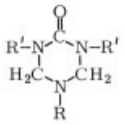

10

Ammonia, primary or secondary amines, or polyamines containing two or more primary amino groups are used for the production of amino-modified resins [3, p. 206]. They are more stable to water than the unmodified resins produced by a similar procedure.

As shown in the equation above, ammonia and amines can react with the hydroxymethyl groups in urea, in a weakly acidic or alkaline medium, to give *N*-aminomethyl groups. In strongly acidic solution, these eliminate the amine substituents to form condensates. This condensation is similar to that starting with hydroxymethyl groups. Even very highly condensed urea resins are able to bind large amounts of ammonia and amines because a substantial part of the formaldehyde is still present in the form of hydroxymethyl groups. This reaction is utilized industrially, for example, to deodorize resins or particle board [72].

Condensation of the resin molecule with an amine produces a resin that has a cationic character; this increases the resin's affinity for cellulose fibers when it is used as an auxiliary for textiles or paper. These basic urea resins are therefore mainly important because they improve the wet tensile strength of paper.

Reaction with Sulfites. Just as formaldehyde hydrate reacts with a sulfite, so the hydroxymethyl groups of the amino resins react with a hydrogen sulfite or a sulfite to give an aminomethanesulfonate [1, p. 120]:

$$\text{R-NH-CH}_2\text{OH} + [\text{HSO}_3]^- \longrightarrow [\text{R-NH-CH}_2\text{SO}_3]^- + \text{H}_2\text{O}$$

This imparts an anionic character to the resins [3, p. 206]. The pH of the solution controls its sulfite content.

$$[\text{SO}_3\text{H}]^- \rightleftharpoons [\text{SO}_3]^{2-} + \text{H}^+$$

To produce resins of this type, the reaction of, e.g., urea and formaldehyde is begun in a neutral or alkaline medium in order to achieve hydroxymethylation. When the desired degree of conversion has been achieved, condensation is carried out using sulfite or hydrogen sulfite in a slightly acid solution [76]. However, the resin can react with a hydrogen sulfite because many hydroxymethyl groups are still present. Water-insoluble resins can even be rendered soluble by posttreatment with a hydrogen sulfite. Although, like unmodified resins, the amino resins produced using sulfites exhibit an increase in viscosity on aging, they remain water soluble because of their hydrophilic ionic groups.

The anionic resins are of particular industrial importance in the field of water-soluble or water-dispersible adhesives. Other uses include textile auxiliaries [77], paper auxiliaries [78], leather auxiliaries, and deodorants [79] for UF resins [80].

In addition to modification with sulfite, the relevant patents mention many other modifiers, such as aldehydes and triazines. The affinity of the anionic resins for cellulose (which carries a slight negative charge) is lower than that of the cationic resins; therefore, the fibers must be posttreated with aluminum sulfate to fix the resins.

The modification of urea–formaldehyde resin glues with ligninsulfonic acid should be mentioned also [81], [82]. The products are dark resin solutions that have been used from time to time in the woodworking industry. The mechanism involved in the incorporation of the nonuniform lignin is still unclear [1, p. 120]. Lignin contains phenolic groups. Incorporation of the lignin components imparts relatively high water resistance, among other things. The lignins are of different origins and compositions, so that the production process must be adapted empirically to the different lignin varieties. This fact and the dark coloration have prevented more extensive use of lignin wastes from wood pulp production.

7.2. Melamine Resins

7.2.1. Unmodified Melamine Resins

Melamine [*9003-08-1*], like urea, can react with formaldehyde [9, p. 65]. The addition of formaldehyde at the amino group is carried out at pH 9 – 10 and takes place very rapidly. One mole of melamine combines with 6 mol of formaldehyde to form hexahydroxymethylmelamine. Melamine in itself is only slightly soluble even in hot water [9, p. 60]. However, it dissolves readily in aqueous formaldehyde solutions, forming hydroxymethylmelamines. A small part of the melamine remains physically dissolved. When heated, hydroxymethylmelamines undergo cross-linking [9, pp. 66] and form resins that become more insoluble the longer they are heated. Melamine condenses with formaldehyde in acid solution to form hydroxymethyl compounds. This is followed by a more rapid polycondensation, yielding cross-linked products, than in the case of urea – formaldehyde [13, p. 359]. Melamine – formaldehyde resins can be hardened thermally in the absence of a hardener or they can be hardened in the absence of heat, using a very strongly acidic hardener. The hardening processes, as in the case of the urea resins, are not entirely clear. Crosslinking takes place principally via methylene bridges, but it also has been suggested that methylene-ether bridges play a role. The melamine formaldehyde condensates possess good thermal stability, optical properties, and water resistance [13, p. 350].

7.2.2. Modified Melamine Resins

Reaction with Alcohols. Melamine hydroxymethyl compounds can be etherified with alcohols in the same manner as hydroxymethylurea compounds [13, p. 361]. The most important use of the products is as surface-coating resins, but they also are used as textile auxiliaries and even as melamine – urea – formaldehyde resins and impregnating resins. As in the case of the urea resins, etherification is carried out at relatively high temperatures with acid catalysis. Where etherification is carried out to a substantial degree, it is important to remove the water or excess alcohol from the weakly acidic solution, because cross-linking otherwise competes with etherification. The mono- to hexaalkyl ethers of the C_1 to C_4 alcohols can be prepared readily. Good conversions are obtained for higher alcohols only by means of transetherification [83].

Methyl ethers of melamine hydroxymethyl compounds are readily soluble in water even at room temperature and crystallize readily [3, p. 361]. The aqueous solutions of the methyl ethers prepared from higher hydroxymethyl compounds, i.e., those with a higher molar ratio of hydroxymethyl groups, become cloudy on heating and clarify again when cooled. Most of the methyl ethers are soluble not only in water but also in benzene and similar solvents. The products etherified with higher alcohols, e.g., butanol or benzyl alcohol, are water insoluble but are readily soluble in alcohols and some hydrocarbons. In addition to monoalcohols, polyhydric alcohols, e.g., ethylene glycol, glycerol, 3-hydroxymethyl-1,5-pentanediol, pentaerythritol, can be employed as etherification agents. This etherification must be carried out in neutral solution because the competing cross-linking reaction otherwise takes place more rapidly and forms insoluble products.

Like the corresponding urea resins, the etherified melamine – formaldehyde resins used as starting materials for surface coatings are condensates having low molecular masses, 600 – 1400 [84]. These are presumably triazine structures linked via methylene bridges and contain from two to four melamine groups. The ether groups are distributed randomly over the triazine rings.

Reaction with Amines. Amines and aminocarboxylic acids react readily with the hydroxymethyl group of hydroxymethylmelamines in the same way as with urea resins. Modification with amines or amino acids improves the hydrophilic character of the melamine resins. Condensation with amino acids gives alkali-soluble melamine resins that are used to produce impregnating or molding materials [85]. Basic products, which form salts with acids, are obtained by modification with aminoalcohols, e.g., aminoethanol [86], and are used as textile auxiliaries. These products are not of major industrial importance.

Reaction with Hydrogen Sulfite. Melamine resins, like urea resins, react with hydrogen sulfites. Industrially, the cheapest product, i.e., sodium sulfite or sodium hydrogen sulfite, is used, and the reaction is carried out in an acidic or basic medium. The products are readily water-soluble melamine resins that can be precipitated with acids. Hydrogen sulfite-modified melamine – formaldehyde resins are used industrially as concrete liquefiers in the building materials industry. In addition to sulfite, other modifiers described in the patent literature are dithionite and formaldehyde sodium sulfoxylate ($NaHSO_2 \cdot CH_2O$) [87], [88].

Other Modifiers. The modification of hydroxymethylmelamine resins with polyalcohols, sugars, toluenesulfonamides, lactams, nitriles [89–91], acrylates, and polyacrylonitrile [1, pp. 109], [92] is of great importance in the production of impregnating resins. Such modifications improve the large number of laminating properties required and stabilize the resin solution against clouding and excessively rapid increase in viscosity.

Mixed condensates and cocondensates of melamine with other compounds – e.g., phenol, urea, cyanamide, and dicyanodiamide – that react with formaldehyde have also been described in a large number of publications. For example, phenol-modified melamine – urea – formaldehyde resins are used as resin glues for weather-resistant particle board [93], [94]. These resins, too, can be modified in a large variety of ways.

7.3. Urethane Resins

Condensation of a urethane (carbamate) of the formula $H_2N–COOR$ with an aldehyde gives a distinct compound or a resin, depending on the production conditions [8, p. 220], [13, p. 371]. In an alkaline medium, carbamates react with formaldehyde to give hydroxymethyl compounds that split up into the starting materials at higher temperatures. In an acidic medium, they cross-link to form methylene bridges and produce resins that are used to manufacture molding materials. A patent has been applied for regarding the use of urethane resins derived from long-chain aliphatic alcohols as plasticizers [3, p. 220]. Urethane – urea – formaldehyde mixed resins have been described also. The reaction of cyanuric acid with polyhydric alcohols and formaldehyde gives resins that can be used as paper and textile auxiliaries and as starting materials for surface coatings.

In another procedure, urea is converted to the appropriate carbamate by heating in the presence of an alcohol and eliminating ammonia. The subsequent reaction with formaldehyde and cross-linking are carried out as described above [13, p. 374]. The resultant highly viscous solutions can be used as label adhesives and for the adhesive bonding of cardboard boxes.

The properties of urethane resins vary depending on the components and method of production. They range from oily to brittle depending on the degree of condensation. Lower aliphatic and cycloaliphatic alcohols yield hard resins, whereas higher alcohols give waxlike resins.

7.4. Cyanamide and Dicyanodiamide Resins

Cyanamide and dicyanodiamide react with formaldehyde both in acid and in alkaline media [8, p. 199]. Cyanamide reacts with an equimolar amount of formaldehyde at pH 6 – 7 to give a white precipitate, which loses water to form a crystalline compound. In general, cyanamide and dicyanodiamide are used as modifiers [1, pp. 118], [13, pp. 382]. Cyanamide – formaldehyde and dicyanodiamide – formaldehyde resins have been used as mixed resins for starting materials in making molding compositions, as dyeing and tanning auxiliaries, for textile finishing, and for preserving urea – formaldehyde foams.

7.5. Sulfonamide Resins

The most commonly used basic component for this type of resins is *p*-toluenesulfonamide:

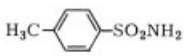

With formaldehyde, *p*-toluenesulfonamide forms resins that are fusible but not curable, whereas disulfonamides or amides possessing

other groups capable of condensation give insoluble products that are infusible [95]. The special properties of *p*-toluenesulfonamide resins include compatibility with cellulose surface coatings, pale natural color (in contrast to that of phenol resins), and low tendency to yellow.

Only the *p*-toluenesulfonamide – formaldehyde condensates have become important industrially, and these only as cocondensates or modifiers. Only the combinations with urea – melamine resins, phenol resins, and phthalate resins are of importance [96]. They also are added to melamine – formaldehyde impregnating resins as plasticizers in amounts of 3 – 8 % [1, pp. 109].

7.6. Aniline Resins

Aniline resins include the condensates of formaldehyde with aromatic amines, such as aniline [*25214-70-4*], toluidines, xylidines, naphthylamines, and phenylenediamines [13, pp. 192].

In alkaline solution, the first stage, i.e., hydroxymethylation, leads to a simple condensate:

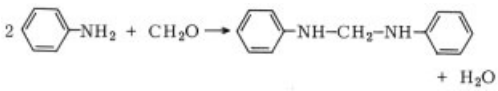

Condensation to give a cyclic product has been reported also.

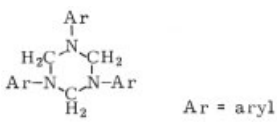

The reaction with formaldehyde in an acid medium gives resins in which methylene groups link the aromatic rings directly, as in the case of the phenoplasts (→ Phenolic Resins). The para position is activated by salt formation at the amino group, so formaldehyde can add at this position to form *p*-aminobenzyl alcohol.

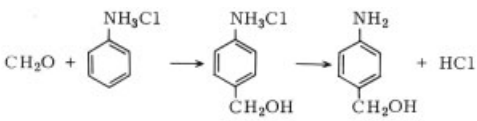

Polymer chains are formed and water is eliminated in this process. In the presence of an excess of formaldehyde, three-dimensional networks also form.

The amino groups have a directing function, so tertiary aromatic amines also can form resins. Unlike the phenol resins, aniline resins do not harden further. The aniline resins are therefore thermoplasts in contrast to the resins discussed above, which are duroplasts. The aniline resins are not very important industrially. They have been used in small amounts in the production of molding materials [1, p. 120] and ion-exchange resins and, because of their high resistance to tracking currents, they have been mixed with epoxy resins and used for wire enamels and insulating coatings [97].

8. Analysis

Analysis, insofar as it is used for product control during production, storage, and use, amounts to testing various types of resins for characteristic parameters or selected specifications. Complete elucidation of the chemical structure, including the sizes and amounts of the oligomers, would entail an unacceptably great expense.

Characterization of the products usually includes the following determinations [11, pp. 171], [98–100]:

density in g/cm^3 (in the case of foams, in kg/m^3)
solids content in % by mass
viscosity in mPa · s, usually at 20 °C
free formaldehyde
total formaldehyde
free urea
total urea
melamine
methoxy content
ash content (incombustible modifiers)

The urea – formaldehyde or amino groups – formaldehyde molar ratio is calculated from the analytical data.

Further tests may be employed if special properties are required of the resins. The development of novel products and applications increases the need for analytical effort.

Solids Content. The determination of the solids content in resin solutions is a determination of

the dry matter content. The value is useful only for comparison. It is determined by evaporating the volatile components at 100 – 130 °C over a fixed period of time. This procedure removes not only the water present as a solvent and the free formaldehyde, but also the water and formaldehyde formed during the drying process in the condensation. Therefore the dry matter content is by no means to be equated with the active ingredient content. More exact data are obtained by the determination of water by the Karl Fischer method [101], by freeze drying, or, in organic solvents, by the Aufhauser method. In industry, however, the water content is estimated by the drying method detailed above because this method is simple and quick.

Formaldehyde [102–104]. The free formaldehyde can be determined directly and the total formaldehyde can be determined after the product has been hydrolyzed completely. The hydroxymethyl content can be determined by hydrolysis under relatively mild conditions. The formaldehyde bonded in the form of methylene groups is calculated by difference.

Urea. The total urea can be determined by hydrolysis and the Kjeldahl procedure. The free urea is degraded to ammonia by the enzyme urease [105]. If no other form of nitrogen is present in the resin, the urea content can be calculated from the nitrogen content, which is determined by combustion analysis.

Melamine in amino resins is determined by hydrolysis of the product under mild conditions followed by titration with perchloric acid. Resins having a low degree of condensation can be titrated directly using a computerized automatic apparatus.

The *molar ratio* of urea to formaldehyde can be calculated from the total content of formaldehyde. This ratio provides information on the formaldehyde elimination to be expected.

Methanol is generally present in the form of methoxy groups bonded to hydroxymethyl groups via an ether bond; it is determined by the Zeisel's procedure.

Other components are important, particularly where the resins are used as paper or textile auxiliaries or as starting materials for making surface coatings. Some of these are maleic acid, glyoxal, phenol, and amines and sulfites. The methods of determination are discussed in textbooks on analysis or, for example, in specific articles. These substances are not determined during routine product control.

For research and development purposes, paper chromatography, thin-layer chromatography [106], column chromatography [107], gel permeation chromatography, ^{13}C NMR spectroscopy [22], [31], [36], ^{15}N NMR spectroscopy [119], ^{1}H NMR spectroscopy (using dimethylsulfoxide as the solvent) [31], [108], and infrared spectroscopy are used.

9. Storage and Shipment

Aqueous or alcoholic solutions of amino resins are stored in steel tanks with or without internal epoxide coatings, in stainless steel tanks, or in containers made of fiberglass-reinforced polyester resin. Temperatures above 20 – 25 °C must be avoided because they bring on rapid condensation, which increases the viscosity and soon renders the resin unprocessible. Rotary pumps or – in the case of highly viscous liquids – positive-displacement pumps, e.g., gear pumps, are employed.

Powdered amino resins are packaged in airtight, plastic-lined paper sacks that are sealed against moisture. Certain types of paper bags can withstand tropical climates. Under high pressure, the materials may solidify. Spilled product is taken to a waste dump.

Liquid resin loads of up to about 250 kg are dispatched in small polyethylene barrels, whereas greater loads, up to about 1 t, are dispatched in polyethylene containers encased in sheet metal. Larger amounts are transported in containers, railroad tank cars, tank trucks, or marine tankers. Amino resins must be disposed of if they are damaged or become unstable during transport in tanks.

The relevant regulations governing the disposal of amino resins differ from country to country and even within some countries. A common feature of all the regulations, however, is that the resin must be hardened by the addition of acid curing agents and must be delivered to a licensed landfill. Where leakage occurs during transport, it is advisable to absorb the resin solution with any appropriate agent that is available (e.g., sand or sawdust) and then to deliver the resulting material for incineration. In the Federal Republic of Germany, amino resins are not

included among substances regarded as endangering water.

10. Uses

The properties of amino resins have been adapted by industry to a very large number of applications. New uses are being developed constantly and any list is bound to be incomplete.

Uses include the following: adhesives; impregnating resins; molding materials; starting materials for making surface coatings; auxiliaries for paper, textiles, leather, and flotation; strengtheners for building materials; concrete liquefiers; binders for glass fibers and foundry sand casting; fire lighters; emery papers; flame retardant coatings; flameproofed combustible items; foamed resins for many purposes; grinding wheels; ion-exchange resins; sewage flocculants; and microcapsule production.

The largest amount of amino resins is consumed by the woodworking industry. These resins, along with the phenol resins, have largely replaced the natural glues based on proteins (glutin, casein, soybean, and blood albumin glues). Impregnating resins are used as resin glues in the particle board, plywood, and furniture industries. Impregnating resins also are used to impregnate papers for decorative laminates and for coating wood particle board.

Resin Glues for the Woodworking Industry. The amino resins used as glues in the woodworking industry are available commercially predominantly in the form of 55 – 70 % aqueous solutions or as solids, for example, powders. The powder resins entail lower transport costs but are more expensive because of the costs involved in drying them and must be dissolved by the user. Whereas the powder resins have shelf lives of 1 year or longer if stored correctly, the liquid products, depending on the type, have shelf lives of only 2 – 6 months. During this period the viscosity remains within limits that permit the resin to be pumped with a conventional pump. The amino resins are more or less stable intermediates; condensation is completed by the addition of a curing agent and/or by heat treatment. A curing agent, e.g., ammonium chloride, is virtually always required for the urea resins, whereas melamine resins can be cured at 120 °C. Compared to the melamine or phenol resin glues, the particle board made from urea resin glue has lower water resistance and is not resistant to boiling. If these properties are required, melamine resin glues or mixed condensates of formaldehyde, urea, melamine, and/or phenol are employed.

In addition to water resistance, the tendency to release formaldehyde is very important. If a low level of formaldehyde release is desired, the formaldehyde content of the resin glue must be correspondingly low. In some countries the glue must receive the approval of the authorities.

Over the past few years, MDF (medium-density fiberboard) resins have become increasingly important. The MDF resins are pure or melamine-modified urea – formaldehyde resins that have a low viscosity, about 200 mPa · s. Other modifiers may be used also. Medium-density fiberboard is used in the production of furniture with structured surfaces [1, pp. 179], [109, pp. 266 – 275].

From an economic point of view, the urea resins are the most important amino resin glues because they are inexpensive and possess technical properties adequate for most purposes.

The main use is in the production of wood particle board. These resins also play an important role as glues for plywood and blockboard and for furniture construction. The gluing of load-bearing wood components and shutter panels and the production of shaped articles are further examples of the large variety of uses of amino resin wood glues.

Impregnating Resins. Urea – formaldehyde or melamine – formaldehyde impregnating resins are used to treat papers for decorative purposes. The resins themselves are brittle products and it may be necessary to modify them with a plasticizer. Because of their good properties, i.e., hardness, abrasion resistance, scratch resistance, and stability to light, the amino resins are particularly suitable for this purpose. The impregnated papers are used either to produce decorative laminates or for coating wood particle board. Papers impregnated with melamine resin are more heat stable than those impregnated with urea resin [110].

The paper webs impregnated with the amino resin are dried in a furnace at 70 – 200 °C and

are stored under air-conditioning until they are used.

Table 1. Trade names of amino resins used as resin glues and impregnating resins

Supplier	Urea resins	Melamine resins
Nobel Hoechst Chim. France	Nobelamine	Nobelamine
Société Chimique des Charbonnages S.A. CdF-Chimie France	Pressamine	Pressamine
BP Chemicals UK	Cellobond	–
Montedison Italy	Xylocolla	–
Società Italiana Resine SIR Italy	Sirit	–
Aika Japan	Aika-Uri	–
Dainippon Ink Japan	Plyamine	–
Mitsui Toatsu Chemical Japan	Uramin	–
Nippon Carbide Japan	–	Nika Resin
Nippon Chemical Japan	–	Suntop
Nippon Kasei Japan	Suiso-Bond	–
Shooei Chemical Japan	S-Resin	–
Sumitomo Bakelite Japan	Igeta	–
Dyno Industrier Norway	Dynomin	Dynomin
Ciba-Geigy Switzerland	Aerolite	Melolam
Amer. Cyanamid USA	UFR Urac	Cymel Melurac
Borden USA	Cascomite	–
Monsanto USA	– –	Scripset Resimene
Reichhold USA	Floramine	Diaron
Rohm and Haas USA	Uformite	Uformite
BASF FRG	Kaurit	Kauramin
Cassella (Hoechst) FRG	–	Madurit
CSL, Chemische Werke Saar Lothringen FRG	SAR	–
Henkel FRG	–	Pressal
Resart FRG	–	Resart
Südwestchemie FRG	Supraplast	Supraplast

Various types of amino resin-impregnated papers are required for decorative laminates. The most important are papers that impart a decorative pattern to the board which itself consists of several layers of phenol resin-impregnated kraft paper.

Ground films and decorative films are likewise amino resin-impregnated papers and are used in the same way as wood veneers in the furniture industry. These films generally are produced from urea resins and to a certain extent also from melamine – urea resin mixtures.

Molding Materials. Molding materials consist of a filler, e.g., cellulose or sawdust, and a binder, which can be a urea resin or a melamine resin. The molding materials obtained from melamine resins are superior to those obtained from urea resins because of their better hardening properties, higher mechanical strength, and greater resistance to tracking currents and moisture [3, pp. 68 – 100].

The molding materials are employed for objects in daily use; e.g., white electrical plugs, switches, covers, telephone accessories, and buttons are made from urea resin. Melamine resin is used for camping, hospital, and canteen tableware, and for junction boxes and insulating com ponents. Mixed resins that combine some of the properties of the individual resins are also used.

The most important trade names of amino resins used as resin glues and impregnating resins are given in Table 1.

Raw Materials for Surface Coatings. Pure amino resins are not very useful as surface coatings. The majority of the commercially available amino resins used as starting materials for surface coatings are etherified with butanol or isobutanol and dissolved therein [3, pp. 124 – 148]. Special brands are also modified with methanol, ethanol, or propanols.

Because the surface coatings obtained using amino resins are too brittle, these resins cannot be used alone as binders. However, they are very useful in combination with other binders. Urea resins and melamine resins are the most important cross-linking components for alkyd resins.

Physically drying finishes are obtained by adding amino resins to nitrocellulose finishes. The addition of these resins improves the hardness, body, and stability to light. The principal field of use is in furniture finishing.

Table 2. Trade names of amino resins used as surface coating resins

Supplier	Urea resins	Melamine resins
Vianova Austria	Viamin	Viamin
UCB Belgium	–	Synkamine
Syntova France	Syntamine	Syntamine
British Industrial Plastics (BIP Chemicals) UK	–	Beetle
British Resin Prod. UK	Epok	Epok
British Paints UK	–	Syntex
ICI UK	Paralac	–
Resia Italy	–	Resmelin
Società Italiana Resine SIR Italy	Siramin	Siramin
Hitachi Kasei Japan	Melan	Melan
Mitsui Toatsu Chemical Japan	Uvan	Uvan
Sanwa Chemical Japan	–	Nikalac
Sumitomo Chemical Japan	Sumimal	Sumimal
Scado-Archer-Daniels Netherlands	Scadonur	Scadomex
Synres Nederland Netherlands	Synresin	Synresin
Synthese Netherlands	Setamine	Setamine
Dyno Industrier Norway	Dynomin U	Dynomin M
Perstorp Norway	Peramin	–
Svenska Olieslageri Sweden	Soamin	Soamin
Ciba-Geigy Switzerland	Cibamin H	Cibamin M
Plüss Staufer Switzerland	–	Plusamid
Allied Chemicals USA	Plaskon	Plaskon
Amer. Cyanamid USA	–	Plaskon
Monsanto USA	–	Resimene
Reichhold USA	Plyamin	Super Beckamine
Rohm and Haas USA	Uformite	Uformite
Texaco USA	–	Pantoxyl
BASF FRG	Plastopal	Luwipal
Hendricks & Sommer FRG	Heso-amin	Heso-amin
Hoechst FRG	–	Maprenal
Jäger FRG	Jägamin	Jägamin
Stickstoffwerk Piesteritz FRG	Piadurol	Meladurol

Acid-curable finishes are obtained by adding alkyd resins or saturated polyester resins. These finishes can be hardened by adding an acid, e.g., hydrochloric acid, *p*-toluenesulfonic acid, or a phosphoric acid derivative, at an elevated temperature. The resulting coatings are very hard and resistant to scratches, solvents, and the effects of light [3, p. 130]. Urea resins are more important than melamine resins, particularly in furniture finishing and parquet sealing.

Coatings that can be cross-linked by heat treatment alone are obtained by combining amino resins with alkyd resins, saturated polyester resins, heat-curable acrylate resins, or epoxy resins. Because of their good mechanical properties, their gloss, and their resistance, these baked finishes are extremely important in industrial metal finishing, for example, for automobile bodywork, household appliances, metal furniture, and containers [111, p. 177].

Because they are etherified and essentially free of water, the amino resins used in the surface-coating industry have shelf lives of from 6 months to several years at room temperature. They are dispatched in plain or galvanized sheet metal containers. Only stainless steel or aluminum tanks are used for transport by tanker trucks and railroad tank cars.

The most important trade names of urea and melamine resins used as starting materials for surface coatings are summarized in Table 2.

Paper Auxiliaries. In the paper industry, amino resins are used to increase the dry tensile strength and especially the wet tensile strength of the product [3, pp. 205 – 209], [112], [113]. In contrast to the procedure followed in the textile industry (see Textile Auxiliaries below), the resin components are added to the paper stock before condensation. They are adsorbed onto the fiber and then undergo condensation in an acid medium. The maximum strength of the paper is reached only after a maturing period of 2 – 4 weeks. Based on the fiber, 1 – 5 wt% of resin is sufficient to achieve good wet strength. The amino resins also are used for the surface finishing of papers. They impart water resistance to starch-containing paper-coating compositions.

The urea resins, which constitute the most important group, are generally hydroxymethyl compounds that have been modified by the addition of amines to give them a cationic character

or by the addition of sulfites to provide an anionic character. After modification, these compounds are subjected to a condensation reaction in an acid medium. The resultant ionic resins are water soluble. The ionic character also enables them to be adsorbed onto anionic paper pulp, but anionic resins must be used with aluminum sulfate [112, p. 43].

Melamine resins also are important to the wet strength of paper. In the absence of other modifiers, they react with acids to form strongly acid cationic colloids that are stabilized by an electric charge and consist of partially condensed hydroxymethyl melamines. Melamine resins modified with aminocaproic acid are also suitable substitutes, either alone or in combination with other polymers, for some or all of the colophony soaps used in the internal sizing of paper [112, p. 42].

A further field of use is the production of moisture-resistant starch pastes for the adhesive bonding of corrugated board [114].

Trade names for aminoplasts used as paper auxiliaries are summarized in Table 3.

Table 3. Amino resins used as paper auxiliaries

Supplier	Resin
Rousselot France	Resin
British Industrial Plastics (BIP Chemicals) UK	Beetle Resin
Mitsui Toatsu Chemical Japan	Uramin
Sumitomo Chemical Japan	Sumires Resin
AKZO Netherlands	Etadurin
Chemanord Norway	Casco-Resin
Amer. Cyanamid USA	Parez Paramel Regular Resin
Hercules Powder USA	Kymene
Monsanto USA	Scriptite Resloom
Rohm and Haas USA	Uformite
BASF FRG	Urecoll
Hoechst FRG	Madurit

Textile Auxiliaries. The textile auxiliaries based on derivatives of urea, melamine, and formaldehyde [3, pp. 166–181] are predominantly monomeric, and therefore should not be regarded as resins. They contain only unavoidable, small, random amounts of higher condensates. The molecules react with the primary hydroxyl groups of the cellulose on contact with the cellulose fibers. The relatively weak hydrogen bridges of the cellulose molecules are thereby supplemented by additional, stronger covalent bonds. Self-condensation takes place on the surface of synthetic fibers, which do not react with the "amino resin" monomers. This type of auxiliary is discussed here, as are others that stiffen the finish of synthetic fabrics.

The pure urea–formaldehyde adducts, mono- and dihydroxymethylurea, are water soluble and crystallize readily. In crystalline form, they possess the largest share of the market for easy-care and wrinkle-resistant finishes for cellulose textiles and mixtures of cellulose fibers with synthetics and wool. The pastes also consist principally of these basic substances, but some of them are etherified with lower alcohols and exhibit a minor amount of condensation. They are therefore used to provide wrinkle-resistant and wash-and-wear finishes as well as stiffened finishes on synthetic fabrics. The fabrics treated in this manner possess improved wash resistance. Hydroxymethyl compounds of urea that exhibit a high degree of condensation and possess a high concentration of ether groups are used to produce highly elastic stiffening effects and good wash resistance in cellulose fabrics. Because of the relatively low concentration of "amino resin" monomers (about 1 mol of resin per kg of cellulose), cross-linking between the cellulose groups is strongly favored over self-cross-linking. A disadvantage of this class of products is the relatively low stability of the acid treatment liquors in the bath, where condensation can take place easily between the hydroxymethylurea molecules.

An important group is constituted by hydroxymethyl compounds of cyclic urea derivatives; examples are dihydroxymethylethylene urea (**11**), dihydroxymethylpropylene urea (**12**), and dihydroxymethylurone (**13**).

Acyclic compounds, such as various alkyl carbamates, are also typical finishing agents. In contrast to the pure urea–formaldehyde compounds, these exhibit little tendency to form self-cross-linked resins and react predominantly with

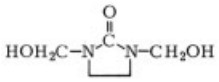

11

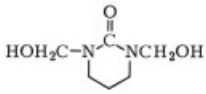

12

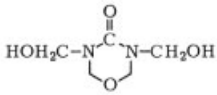

13

cellulose to cross-link the fibers. The other advantages, over and above the mechanical properties, are resistance to chlorinated water, lightfastness, and shade constancy. These are closely related to the chemical structure and the resistance of the C–N and N–O bonds to hydrolysis.

Table 4. Amino resins used as textile auxiliaries

Supplier	Resin
PCUK France	Diamonine
Dai Nippon Ink and Chemicals Japan	Beckamin
Mitsui Toatsu Chemical Japan	Uramin
Sumitomo Chemical Japan	Sumitex Resin
Sandoz Switzerland	Finish
Monsanto USA	Resloom
Sun Chemical USA	Permafresh
BASF FRG	Fixapret Kaurit
Ciba-Geigy/Pfersee FRG	Knittex Lyofix
Henkel FRG	Stabitex
Hoechst FRG	Arkofix
Hoechst FRG	Cassurit
Thor-Chemie FRG	Quecodur

Hydroxymethyl melamine derivatives and their ethers are of some importance for providing high-grade finishes, i.e., embossed finishes, Schreiner finishes, chintz finishes, and plisse finishes, but they have largely been replaced by the reactant types. The finishes obtained using hydroxymethylmelamines are more wash resistant than those made with hydroxymethylureas. Hydroxymethylmelamine compounds are preferred for embossed finishes (Everglaze) where the fabric can be damaged mechanically.

The most important trade names of the textile auxiliaries based on amino resins are summarized in Table 4.

Table 5. Trade names of amino resins used as leather auxiliaries

Supplier	Resin
Hodgson UK	Neosyn
BASF FRG	Relugan
Bayer FRG	Retingan
Henkel FRG	Drasil
Chemtan USA	Chemtan

Leather Auxiliaries. Resin retannage of chrome-tanned leather is carried out using hydroxymethyl compounds of urea, thiourea, melamine, dicyanodiamide, and ethers of these [3, p. 211], [80]. Aqueous solutions of the resins should penetrate the hide before condensation begins. The acid condensation catalyst (pH 3.9 – 4.5) should not reduce the period of activity of the tanning liquor or damage the hide. As in the case of the paper auxiliaries, the resins are rendered cationic using amines or anionic by means of a bisulfite. The anionic resins based on melamine or dicyanodiamide make leather very strong and supple and leave the color unchanged. The hydroxymethyl groups react with the active groups of the collagen fiber, which has a high content of glycine, proline, and oxyproline.

The trade names of a few leather auxiliaries are given in Table 5.

Other Uses of Amino Resins. The number of possible uses is increasing constantly. The uses listed at the beginning of this section are summarized with the appropriate references in Table 6.

The importance of many uses, e.g., flameproofing agents, has declined greatly. Some applications, e.g., casting resins, have proved unsatisfactory in practice, whereas others are only at the development stage and their industrial suitability cannot yet be assessed.

Table 6. Uses for amino resins

Application	Resin and use	References
Binders for glass fiber webs	Reinforcing glass fibers in webs by means of urea resins	[9, pp. 201, 214]
Foams	Soil conditioning, covering landfills, heat insulation, lining tunnels in mines (→ Foamed Plastics)	[115], [1, pp. 196], [3, p. 161]
Foundry sand binders	Binders for sand-casting in the foundry industry; casting molds and cores; modification with furfuryl alcohol	[3, pp. 122]
Concrete liquefiers and auxiliaries for building materials	Sulfite-modified melamine resins	[87], [88]
Flame-proofing agents	Flame-proofing wood by means of amino resins containing phosphorus, boron, bromine, etc.	[1, p. 206], [3, p. 215]
Ion-exchange resins	Melamine-formaldehyde condensates modified with amines Used for recovering silver from photographic effluents	[3, p. 225]
Coatings for abrasive papers	Urea – formaldehyde – furfuryl alcohol condensates for imparting water-resistance	[3, p. 226]
Wood preservatives	Melamine – formaldehyde condensates impart good aging resistance; no shrinkage; no seasoning cracks	[3, p. 227]
Microcapsules	Urea – formaldehyde resins	[1, p. 289], [116]
Firelighters	Emulsion of urea – formaldehyde resin with a liquid fuel and hardeners is processed to give small moldings. Lighting oil-fired furnaces, charcoal grills, etc.	[117]
Sewage flocculants	Urea – formaldehyde resins cause flocculation	[118]

Table 7. Production of amino resins in the OECD countries

Year	Production (10^6 tons)
1969	1.95
1970	2.03
1971	2.20
1972	2.55
1973	3.13
1974	2.79
1975	2.47
1976	2.79
1977	2.94
1978	3.09
1979	3.33
1980	3.14
1981	3.01
1982	2.69
1983	2.66

Comments: The figures for 1974 to 1976 and for 1979 exclude Portugal. The figures include the phenol resins produced in the Federal Republic of Germany. However, this qualification does not greatly change the above figures for world production of amino resins.

11. Economic Aspects

Amino resins have become increasingly important since World War II. The OECD statistics "The chemical industry" and the production statistics of the Federal Statistics Office of the Federal Republic of Germany refer the production for the OECD countries shown in Table 7.

The figures for Europe, USA and Japan are given in Table 8.

The production of particle board, which consumes the largest amount of amino resins in the Federal Republic of Germany, reached a high level by the late 1970s but diminished thereafter.

Table 8. Production of amino resins for Europe, USA and Japan

	Production in 10^3 tons					
	1978	1979	1980	1981	1982	1983
Europe	1790	1890	1810	1690	1560	1420
USA	600	711	604	610	524	613
Japan	706	734	716	629	606	632

12. References

General References

1. B. Meyer: *Urea-Formaldehyde Resins,* Addison-Wesley, London 1979.
2. H. Baumann: *Leime und Kontaktkleber,* Springer Verlag, Berlin – Göttingen – Heidelberg – New York 1967.
3. A. Bachmann, T. Bertz: *Aminoplaste,* VEB-Verlag f. Grundstoffindustrie, Leipzig 1970.
4. R. Houwink, A. J. Staverman: *Chemie und Technologische Kunststoffe,* 4th ed., **vol. I: pp.** 102 – 103, **vol. II/1:** pp. 623 – 639, **vol. II/2:** pp. 681 – 685, Akadem. Verlagsgesellschaft, Leipzig 1962.
5. H. Petersen: *Grundzüge der Aminoplastchemie,* W. Pansegrau Verlag, Berlin 1968 (Reprint from Kunstoffjahrbuch, 10th Issue).
6. E. Plath, L. Plath: *Taschenbuch der Kitte und Klebstoffe,* Wissenschaftl. Verlags-GmbH, Stuttgart 1963.
7. J. F. Blais: *Amino Resins,* Reinhold Publ. Co., New York 1959.
8. P. Talet: *Aminoplastes,* Dunod, Paris 1951.
9. C. P. Vale: *Aminoplastics,* Cleave-Hume Press Ltd., New York 1950.
10. R. Vieweg, E. Becker: "Duroplaste", in *Kunststoff-Handbuch,* vol. X, Hanser Verlag, München 1968, pp. 134.
11. G. F. D'Alelio: "Aminoplasts," in *Kunstoff-Praktikum,* Hanser Verlag, München 1952, pp. 50 – 69.
12. D. Braun, H. Cherdron, W. Kern: *Praktikum der makromolekularen organischen Chemie,* 2nd ed., Hüthig Verlag, Heidelberg 1971.
13. E. Müller (ed.) in Houben-Weyl, *Methoden der organischen Chemie,* **vol. XIV/12: Makromolekulare Stoffe,** part 2, Thieme Verlag, Stuttgart 1963, pp. 319 – 400.

Specific References

14. E. Einhorn, A. Hamburger, *Ber. Dtsch. Chem. Ges.* **41** (1908) 24.
15. Henkel, DE 647303, 1935 (W. Hentrich, R. Köhler).
16. H. Umstätter: *Einführung in die Viskosimetrie und Rheometrie,* Springer Verlag, Berlin – Göttingen – Heidelberg 1952, pp. 5, 17.
17. A. Eucken: *Lehrbuch der chemischen Physik,* 2nd ed., **vol. II,** Akad. Verlagsges., Leipzig 1944, pp. 1119 – 1120.
18. K. Masters: *Spray Drying,* Leonhard Hill Books, London 1972, pp. 489.
19. A. F. Price, A. R. Cooper, A. S. Meskin: "Urea-Formaldehyde Reaction System, an Experimental Investigation," *J. Appl. Polym. Sci.* **25** (1980) 2597 – 2611.
20. G. Zigeuner, *Kunstoffe* **41** (1951) no. 7, 221 – 224.
21. D. Braun, F. Bayersdorf, *Angew. Makromol. Chem.* **89** (1980) 183 – 200.
22. V. Horn, G. Benndorf, K.-P. Rädler: " Strukturgruppenanalyse von Harnstoff-Formaldehyd-Leimen", *Plaste Kautsch.* **25** (1978) no. 10, 570 – 575.
23. K. Kumlin, R. Simonson: "Urea-Formaldehyde Resins," *Angew. Makromol. Chem.* **93** (1981) 27 – 42.
24. Teukros, DE 2745951, 1976.
25. Borden Inc., GB 1319110, 1970.
26. Hitachi Chem. K. K., JP 56024415, 1979.
27. Evidenzbüro Österreichischer Zuckerfabriken, DE 2928003, 1979 (G. Greber, H. Andres, W. Pichler).
28. BASF, EP 52211, 1980 (W. Weiss, H. Petersen, H. Etling, V. M. Duda, L. Lelgemann, W. Pfalzgraf).
29. A. Šebenik, U. Osredkar, M. Žigon, I. Vizovišek: "Study of the reaction between Urea and Formaldehyde by DSC and ^{13}C NMR Spectroscopy," *Angew. Makromol. Chem.* **102** (1982) 81 – 85.
30. T. J. Perelpekova, F. G. Igranova, V. D. Moiseev: "Calorimetric study of the methylolation of urea by formaldehyde," *Khim. Promst. Ser. Proizvod. Pererab. Plastmass Sint. Smol.* 1981 no. 5, 12 – 14; *Chem. Abstr.* **96** (1982) 35762j.
31. B. Tomita, S. Hatono: "Urea-Formaldehyde Resins. III. Constitutional Characterisation by ^{13}C Fourier Transform NMR Spectroscopy," *J. Polym. Sci.* **16** (1978) 2509 – 2525.
32. S. H. Nam, J. I. Ha: "Study of melamine formaldehyde resins. 1. Equilibrium constants in methylolation," *Hwahak Kwa Hwahak Kongop* **19** (1976) 180 – 182, 194; *Chem. Abstr.* **86** (1977) 44 266 w.
33. W. J. Blank, *J. Coat. Technol.* **51** (1979) 62.
34. H. Petersen, *Chem. Ztg.* **95** (1971) 695.
35. H. Petersen, *Textilveredlung* **5** (1970) 570 – 588.
36. M. Chiavarini, N. Del Fanti, R. Bigatto: "Compositive Characterization of Urea-Formaldehyde Adhesives by NMR Spectroscopy," *Angew. Makromol. Chem.* **46** (1975) no. 695, 151 – 162.

37. British Ind. Plastics, GB 524350, 1939.
38. ICI, DE 2726617, 1977 (D. A. Hubbard).
39. Girdler Corp., US 2456192, 1943 (H. G. Houlton).
40. Sherwood Paints, DE 863417, 1949.
41. Allied Chem. and Dye Corp., US 2658054, 1951 (G. A. Coleman, R. B. Greene, J. H. Merriam, S. P. Miller, R. F. Shannon).
42. Allied Chem. and Dye Corp., US 2729616, 1952 (M. H. Bigelow, H. N. Spurlock).
43. Spumalit-Anstalt, FR 1104018, 1954.
44. Rütgerswerke, DE 1029150, 1955 (H. Sauer, A. Tusch).
45. Skanska Attifabriken Aktiebolag, GB 800201, 1951.
46. Du Pont, US 2849421, 1956 (T. D. Weldin).
47. J. Meissner, DE 1570998, 1965 (L. Brucker-Voigt).
48. Rheinpreussen AG für Bergbau und Chemie, BE 711431, 1967.
49. St. Regis Paper Comp., US 2688606, 1951 (G. P Schmitt, C. Werberig).
50. BASF, DE 2109754, 1971 (F. Brunnmüller, H. Schatz, J. Mayer, O. Grabowsky).
51. Società Italiana Resine, DE 2263125, 1972 (S. Vargiu, G. S. Sesto, G. Mazzolani, U. Nistri).
52. Montedison, EP 38180, 1980 (D. Cannaloni, N. Conti).
53. Stamicarbon, BE 635085, 1963.
54. American Cyanamid Corp., FR 955810, 1947 (K. L. Lynch, A. J. Grossman).
55. W. Bauer, DE 1043628, 1953 (W. Bauer).
56. L. Unterstenhöfer, *Kunststoffe* **57** (1967) no. 11, 850.
57. *Lehr- und Handbuch der Abwassertechnik,* Verlag W. Ernst & Sohn, Berlin – München 1969.
58. H. Staudinger, K. Wagner, *Makromol. Chem.* **12** (1954) 168, 194.
59. H. Sobue, K. Murakami, C. Y. Tae, *Kobunshi Kagaku* **12** (1955) 240 – 257, *Chem. Abstr.* **51** (1957) 734 a.
60. BASF, EP 52211, 1980 (W. Weiss, H. Petersen, H. Etling, U. M. Duda, L. Lelgemann, W. Pfalzgraf).
61. H. Scheibler, *Z. Angew. Chem.* **53** (1940) 303.
62. Th. Goldschmidt, DE 2734628, 1977 (A. Laqua, U. Holtschmidt, E. Schamberg, D. Hellwig).
63. I. G. Farbenind., DE 733496, 1938 (K. Ott, K. Hamann).
64. Mitsubishi Kasei Kogyo, DE 1146080, 1951 (M. Hamamoto, S. Yasuhiro).
65. BASF, DE 889152, 1952 (B. von Reibnitz).
66. Phrix Werke, US 2876062, 1953 (E. Torke, J. König).
67. H. Petersen, *Text. Rundsch.* **16** (1961) no. 9/10, 2 – 11.
68. H. Petersen, *Text. Res. J.* **38** (1968) 156 – 176.
69. H. Petersen, *Text. Res. J.* **40** (1970) 335 – 344.
70. BASF, EP 62900, 1982 (C. Dudeck, E. Weber, H. Diem, O. Wittmann).
71. BASF, EP 53762, 1982 (C. Dudeck, E. Weber, H. Diem, J. Mayer, O. Wittmann, G. Lehmann).
72. Verkor N. V., Belgien, DE-OS 2903254, 1979.
73. Fraunhofer-Gesellschaft, DE 2829021, 1978 (E. Roffael, L. Mehlhorn).
74. Aktiebolaget Statens, EP 27583, 1980 (A. W. Westling).
75. M. Paquin, *Angew. Chem.* **60** (1948) 267 – 271.
76. BASF, DE 889225, 1949 (H. Scheuermann, J. Lenz).
77. Böhme Fettchemie, US 2978359, 1954 (H. Wedell).
78. The Resinous Products and Chem. Comp., US 2407599, 1946 (R. W. Auten, J. L. Rainey).
79. West Point Manufacturing Comp., US 2870041, 1958 (H. M. Waddle, J. F. Cotton, R. E. Hudson).
80. Nopco Chem. Comp., US 3063781, 1959 (C. A. Fetscher, S. L. Lipowski).
81. E. Roffael: "Fortschritte in der Verwendung von Sulfitablaugen als Binde- und Zusatzmittel bei der Herstellung von Holzspanplatten", *Adhäsion* (1979) no. 11, 334 – 336; (1979) no. 12, 368 – 370.
82. Champion Int. Corp., US 3994850, 1976 (W. N. Willegger).
83. A. Gams, G. Widmer, W. Fisch, *Helv. Chim. Acta* **24** (1941) 302 E – 319 E.
84. D. Braun, W. Pandjojo: "Gelchromatographische Untersuchungen von ungehärteten Kondensaten aus Melamin und Formaldehyd", *Angew. Makromol. Chem.* **80** (1979) 195 – 205.
85. BASF, DE 888169, 1949 (H. Scheuermann).
86. I. G. Farbenind., FR 880189, 1942.
87. BASF, DE 2424379, 1974 (U. Hampel, R. Petri, H. Petersen, J. Lenz, G. Matthias, W. Reuther).
88. Süddeutsche Kalkstickstoffwerke, DE 1671017, 1966 (H. Aignesberger, H. Michaud).
89. Cassella, DE 3104420, 1981 (H. Hönel, W. Michel, S. Piltsch, K. Schlüter, A. Wolf).
90. Société Chimique des Charbonnages, FR 80/09 193, 1980.

91. BASF, DE 1595368, 1966 (H. Etling, H. Henkel, F. Meyer).
92. Torbjörn Reitberger, SE 8000 – 663, 1980.
93. BASF, DE 2020481, 1970 (J. Mayer, C. SchmidtHellerau).
94. BASF, DE 2951957, 1979 (J. Mayer, C. SchmidtHellerau, C. Dudeck, H. Diem, H. Schatz).
95. L. McMaster, *J. Am. Chem. Soc.* **56** (1954) 204 – 206.
96. The Resinous Products and Chem. Comp., US 2160196, 1937 (H. A. Bruson, J. W. Eastes).
97. Dr. Beck and Co., DE 888294, 1949 (H. Beck, A. Wiegand).
98. H. Schindlbauer, B. Holzhöfer, J. Schuster, *Kunststoffe* **69** (1973) no. 3, 158 – 162.
99. W. G. K. Taylor: "Aminoplastics", *Rep. on the Progress of Appl. Chem.,* vol. XLIX, Society of Chemical Industry, London 1964, pp. 572 – 581.
100. G. Widmer: "Quantitative Bestimmung von Harnstoff und Melamin in Gemischen von Harnstoff- und Melamin-Formaldehyd-Kondensationsprodukten", *Kunststoffe* **46** (1956) no. 8, 359 – 362.
101. DIN 51777:
102. F. Käsbauer, D. Merkel, O. Wittmann: "Bestimmung von freiem Formaldehyd und N-Methylol-Formaldehyd in Harnstoff-Formaldehyd-Kondensaten", *Fresenius Z. Anal. Chem.* **281** (1976) 17 – 21.
103. H. Petersen, L. Klug, H. Hahn, W. Huber, *Farbe + Lack* **87** (1981) no. 8, 647 – 652.
104. G. Groh, H. Petersen, L. Klug, *Farbe + Lack* **87** (1981) no. 9, 744 – 748.
105. AOAC-Methods (Association of Official Analytical Chemists), 13th ed., Washington 1980, p. 17.
106. D. Braun, P. Günther: "Methoden zur Verfolgung der Harnstoff-Formaldehyd-Harz-Synthese", *Kunststoffe* **72** (1982) no. 12, 785 – 790.
107. K. Kumlin, R. Simonson: "Urea-Formaldehyde Resins, 1. Separation of Low Molecular Weight Components in Urea-Formaldehyde Resins by means of Liquid Chromatography," *Angew. Makromol. Chem.* **68** (1978) no. 1025, 175 – 184.
108. M. Chiavarini, R. Bigatto, N. Conti: "Synthesis of Urea-Formaldehyde Resins: NMR Studies on Reaction Mechanisms," *Angew. Makromol. Chem.* **70** (1978) no. 1020, 49 – 58.
109. H. J. Deppe, K. Ernst: *Taschenbuch der Spanplattentechnik,* 2nd ed., DRW-Verlag, Leinfelden 1982.
110. T. Götze, P. Dörries, *Plastverarbeiter* **33** (1982)no. 9, 1118 – 1122.
111. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* **vol. IV,** Heenemann, Berlin 1976.
112. F. Wultsch: *Hilfsmittel und ihre Anwendung in der Papiererzeugung,* G. Staib Verlag, Biberach, 1966.
113. "Wet Strength Paper and Paperboard," *TAPPI Monogr. Ser.* **29** (1965) 1 – 32.
114. K. Holmberg: "Alkali Catalysed Curing of Starch-Melamine Resin Systems," *Polym. Bull. (Berlin)* **6** (1982) 553 – 558.
115. H. Baumann: *Plastoponik, Schaumstoffe in der Agrarwirtschaft,* Hüthig Verlag, Heidelberg 1967.
116. Fuji Photo Film Co., DE 2905632, 1979 (Y. Hoshi, M. Matzukana).
117. BASF, DE 1794131, 1968 (F. Brunmüller, K. Neumann, J. Lenz).
118. Société Française Hoechst, EP 5390, 1979 (C. Trouve, J. Cabestany).
119. J. R. Ebdon, P. E. Heaton, W. T. S. O'Rourke, J. Parkin, *Polymer* **25** (1984) 821 – 825.

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Ammonia

Standard Article

Max Appl[1]

[1]Dannstadt-Schauernheim, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a02_143.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Occurrence and History
2. Physical Properties
3. Chemical Properties
4. Production
4.1. Historical Development
4.2. Thermodynamic Data
4.3. Ammonia Synthesis Reaction
4.3.1. General Aspects
4.3.2. Catalyst Surface and Reaction Mechanism
4.3.3. Kinetics
4.4. Catalysts
4.4.1. Classical Iron Catalysts
4.4.1.1. Composition
4.4.1.2. Particle Size and Shape
4.4.1.3. Catalyst-Precursor Manufacture
4.4.1.4. Catalyst Reduction
4.4.1.5. Catalyst Poisons
4.4.2. Other Catalysts
4.4.2.1. General Aspects
4.4.2.2. Metals with Catalytic Potential
4.4.2.3. Commercial Ruthenium Catalysts
4.5. Process Steps of Ammonia Production
4.5.1. Synthesis Gas Production
4.5.1.1. Feedstock Pretreatment and Raw Gas Production
4.5.1.2. Carbon Monoxide Shift Conversion
4.5.1.3. Gas Purification
4.5.2. Compression
4.5.3. Ammonia Synthesis
4.5.3.1. Synthesis Loop Configurations
4.5.3.2. Formation of Ammonia in the Converter
4.5.3.3. Waste-Heat Utilization and Cooling
4.5.3.4. Ammonia Recovery from the Ammonia Synthesis Loop
4.5.3.5. Inert-Gas and Purge-Gas Management
4.5.3.6. Influence of Pressure and Other Variables of the Synthesis Loop
4.5.3.7. Example of an Industrial Synthesis Loop
4.6. Complete Ammonia Production Plants
4.6.1. Steam Reforming Ammonia Plants
4.6.1.1. The Basic Concept of Single-Train Plants
4.6.1.2. Further Developments
4.6.1.3. Minimum Energy Requirement for Steam Reforming Processes
4.6.1.4. Commercial Steam Reforming Ammonia Processes
4.6.1.4.1. Advanced Conventional Processes
4.6.1.4.2. Processes with Reduced Primary Reformer Firing
4.6.1.4.3. Processes Without a Fired Primary Reformer (Exchanger Reformer)
4.6.1.4.4. Processes Without a Secondary Reformer (Nitrogen from Air Separation)
4.6.2. Ammonia Plants based on Partial Oxidation
4.6.2.1. Ammonia Plants Based on Heavy Hydrocarbons
4.6.2.2. Ammonia Plants Using Coal as Feedstock
4.6.3. Waste-Heat Boilers
4.6.4. Single-Train Capacity – Mega Ammonia Plants
4.7. Modernization of Older Plants (Revamping)
4.8. Material Considerations for Equipment Fabrication
5. Storage and Shipping
5.1. Storage
5.1.1. Pressure Storage
5.1.2. Low-Temperature Storage
5.1.3. Underground Storage

5.1.4. Storage of Aqueous Ammonia
5.2. Transportation
6. Quality Specifications and Analysis
7. Environmental, Safety, and Health Aspects
7.1. Environmental Aspects of Ammonia Production and Handling
7.2. Safety Features
7.3. Health Aspects and Toxicity of Ammonia
8. Uses
9. Economic Aspects
10. Future Perspectives

[Top of Page]

1. Occurrence and History

Occurrence. Ammonia, NH_3, occurs in nature almost exclusively in the form of ammonium salts. Natural formation of ammonia is primarily by decomposition of nitrogen-containing organic materials or through volcanic activity. Ammonium chloride can deposit at the edges of smoldering, exposed coal beds, as observed in Persia before 900 A. D. Similar deposits can be found at volcanoes, for example, Vesuvius or Etna, in Italy. Ammonia and its oxidation products, which combine to form ammonium nitrate and nitrite, are produced from nitrogen and water vapor through electrical discharges in the atmosphere. These ammonium salts, as well as those arising from industrial and automotive exhausts, supply significant quantities of the nitrogen needed by growing plants when eventually deposited on the earth's surface. Ammonia and its salts are also byproducts of commercial processing (gasification, coking) of fossil fuels such as coal, lignite, and peat (see Fig. 1).

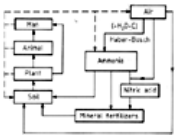

Figure 1. The nitrogen cycle

History. The name ammonia is derived from “sal ammoniacum” (Oasis Ammon in Egypt, today Siwa). Sal ammoniac was known to the ancient Egyptians; the Arabs were aware of ammonium carbonate. Free ammonia was prepared for the first time in 1774 by J. B. PRIESTLEY. In 1784, C. L. BERTHOLLET recognized that ammonia was composed of the elements nitrogen and hydrogen. W. HENRY, in 1809, determined the volumetric ratio of the elements as 1 : 3, corresponding to the chemical formula NH_3.

Following the discovery of the nature and value of mineral fertilization by LIEBIG in 1840, nitrogen compounds were used in increasing quantities as an ingredient of mineral fertilizers. At the end of the last century ammonia was recovered in coke oven plants and gas works as a byproduct of the destructive distillation of coal. The produced ammonium sulfate was used as fertilizer. Another nitrogen fertilizer was calcium cyanamide which was manufactured by the Frank – Caro process from 1898 onwards. Since both sources of nitrogen were limited in quantity they did not suffice for fertilization. Therefore, it was necessary to use saltpeter from natural deposits in Chile. In 1913, the first Haber – Bosch plant went on stream, representing the first commercial synthesis of ammonia from the elements. Subsequently, other ammonia plants were started up, and synthetic ammonia and/or its derivatives soon replaced Chile saltpeter. Today ammonia is a commodity product of the chemical industry. World production capacity in 2005 exceeded 168 × 10^6 t ammonia.

[Top of Page]

2. Physical Properties

Molecular Properties. Corresponding to its nuclear charge number, the nitrogen atom possesses seven shell electrons. One electron pair is in the ground state 1 *s* (K shell), and five electrons are distributed over the four orbitals with the principal quantum number 2 (L shell). Of these, one electron pair occupies the 2 *s* level and three unpaired electrons, respectively, a half of the remaining three levels, 2 p_x, 2 p_y, 2 p_z. The unpaired electrons can enter into electron-pair bonds with the 1*s* electron of three hydrogen atoms. Thus, the three half-occupied orbitals of the L shell become about fully occupied (formation of an octet of the neon type in accordance with the octet theory of Lewis – Langmuir).

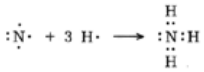

The nitrogen atom is at the apex of a pyramid above the plane of the three hydrogen atoms, which are arranged in an equilateral triangle. The H-N-H bond angle is about 107° [7]. Although covalent, the nitrogen – hydrogen bonds have a polar contribution because of the stronger electronegativity of nitrogen relative to hydrogen. Because of polarization of the bonds and the unsymmetrical molecular arrangement, the ammonia molecule has a considerable dipole moment, 1.5 D [8].

As the ammonia molecule possesses the same electron configuration as water (isosterism) and similar bond angles (water vapor bond angle 105°, dipole moment 1.84 D), ammonia and water behave similarly in many reactions. Ammonia and water are diamagnetic. The dielectric constant of liquid ammonia is about 15 and greater than that of most condensed gases; therefore, liquid ammonia has a considerable ability to dissolve many substances. The ammonia molecule, with its free electron pair, can combine with a proton.

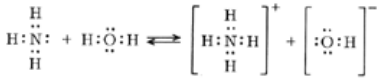

In the resulting ammonium ion, the nitrogen atom is situated in the middle of a tetrahedron whose corners are occupied by hydrogen atoms. The four hydrogen atoms are equivalent in their behavior. According to LINUS PAULING, the positive excess charge is distributed about equally over all five atoms.

Compilations of Physical Data. The results of comprehensive investigations of the physical properties of ammonia have been published in [9] and [12]. Both papers provide numerous equations for physical properties derived from published data, the laws of thermodynamics, and statistical evaluation. These equations are supplemented by lists and tables of thermodynamic quantities and an extensive collection of literature references.

Moreover, data on physical properties and the complex systems important in synthesis may be found in [13-16], and of course, in the well-known tabulations *Landolt-Börnstein* [17] and *Handbuch der Kältetechnik* [18], among others. The most important physical data are compiled in Table 1.

Table 1. Properties of ammonia

M_r	17.0312

Molecular volume (at 0 °C, 101.3 kPa)	22.08 L/mol
Specific gas constant *R*	0.48818 kPa m^3 kg^{-1} K^{-1}
Liquid density (at 0 °C, 101.3 kPa)	0.6386 g/cm^3
Gas density (at 0 °C, 101.3 kPa)	0.7714 g/L
Liquid density (at –33.43 °C, 101.3 kPa)	0.682 g/cm^3
Gas density (at –33.43 °C, 101.3 kPa)	0.888 g/L
Critical pressure	11.28 MPa
Critical temperature	132.4 °C
Critical density	0.235 g/cm^3
Critical volume	4.225 cm^3/g
Critical compressibility	0.242
Critical thermal conductivity	0.522 kJ K^{-1} h^{-1} m^{-1}
Critical viscosity	23.90×10^{-3} mPa · s
Melting point (triple point)	–77.71 °C
Heat of fusion (at 101.3 kPa)	332.3 kJ/kg
Vapor pressure (triple point)	6.077 kPa
Boiling point (at 101.3 kPa)	–33.43 °C
Heat of vaporization (at 101.3 kPa)	1370 kJ/kg
Standard enthalpy of formation (gas at 25 °C)	–46.22 kJ/mol
Standard entropy (gas at 25 °C, 101.3 kPa)	192.731 J mol^{-1} K^{-1}
Free enthalpy of formation (gas at 25 °C, 101.3 kPa)	–16.391 kJ/mol
Net heating value, *LHV*	18.577 kJ/g
Gross heating value, *HHV*	22.543 kJ/g
Electrical conductivity (at –35 °C), very pure	1×10^{-11} Ω^{-1} cm^{-1}
Commercial	3×10^{-5} Ω^{-1} cm^{-1}
Ignition temperature acc. to DIN 51794	651 °C
Explosive limits	
NH_3 – O_2 mixture	
(at 20 °C, 101.3 kPa)	15 – 79 vol % NH_3
NH_3 –air mixture	
(at 0 °C, 101.3 kPa)	16 – 27 vol % NH_3
(at 100 °C, 101.3 kPa)	15.5 – 28 vol % NH_3

p – V – T Data. The *p – V – T* data in Table 2 are calculated from the equations in [12] and further data in [9], [13-18]. Measured *p – V – T* values for liquid ammonia in the pressure range from 7 to 180 MPa (70 – 1800 bar) and at temperatures from –20 °C to 40 °C may be found in [19]. Detailed information on compressibility gained from experimental data can be found in [24-29]. An equation of state for liquid ammonia is given in [20]; for a more simple form, see [16]. Further data may be found in references [21-23]. Properties of liquid ammonia from –50 to 65 °C and from saturation pressure up to 370 bar are reported in [30].

Table 2. Specific volume of ammonia in L/kg

***p*, MPa**	**Temperature, °C**											
	–33	**–20**	**–10**	**0**	**10**	**20**	**50**	**100**	**150**	**200**	**250**	**300**
0.1	1172.001	1235.492	1284.303	1333.094	1381.870	1430.634	1576.882	1820.589	2064.339	2308.163	2552.061	2796.020
0.5	1.467	1.503	1.533	1.566	257.613	269.524	303.259	356.206	407.353	457.602	507.299	556.610
1	1.466	1.503	1.533	1.565	1.600	1.638	144.981	173.883	200.728	226.667	252.117	277.272
2	1.465	1.501	1.531	1.563	1.598	1.636	64.713	82.476	97.322	111.121	124.391	137.352
3	1.464	1.500	1.530	1.562	1.596	1.634	1.772	51.740	62.785	72.589	81.818	90.721
4	1.463	1.499	1.528	1.560	1.594	1.631	1.768	36.088	45.448	53.305	60.532	67.414
5	1.462	1.498	1.527	1.558	1.592	1.629	1.764	26.361	34.981	41.713	47.754	53.432
6	1.461	1.497	1.526	1.557	1.590	1.627	1.760	19.379	27.944	33.968	39.229	44.109
7	1.460	1.495	1.524	1.555	1.589	1.625	1.756	2.171	22.859	28.422	33.137	37.449
8	1.459	1.494	1.523	1.554	1.587	1.623	1.752	2.151	18.986	24.251	28.565	32.455
10	1.458	1.492	1.520	1.551	1.583	1.618	1.745	2.115	13.375	18.385	22.163	25.467
15	1.453	1.486	1.514	1.543	1.574	1.608	1.727	2.048	4.321	10.478	13.632	16.176
20	1.448	1.481	1.508	1.536	1.566	1.599	1.712	1.999	2.780	6.534	9.408	11.574
25	1.443	1.475	1.501	1.529	1.558	1.590	1.697	1.959	2.516	4.474	6.957	8.863
30	1.438	1.470	1.496	1.522	1.551	1.581	1.684	1.926	2.376	3.540	5.443	7.113
35	1.434	1.465	1.490	1.516	1.544	1.573	1.672	1.898	2.282	3.094	4.494	5.922

40	1.429	1.460	1.484	1.510	1.537	1.565	1.660	1.873	2.211	2.839	3.893	5.088
45	1.424	1.454	1.479	1.504	1.530	1.557	1.650	1.850	2.155	2.670	3.497	4.491
50	1.419	1.449	1.473	1.498	1.523	1.550	1.639	1.830	2.108	2.547	3.221	4.054
55	1.414	1.444	1.468	1.492	1.517	1.543	1.630	1.811	2.068	2.453	3.020	3.727
60	1.409	1.439	1.462	1.486	1.511	1.536	1.620	1.794	2.033	2.377	2.865	3.476

Caloric Data. Enthalpy and entropy of solid ammonia are given in [13]. Enthalpy and entropy may be calculated by the equations in [12]. Further data are given in [9-11] and [13-18]. An enthalpy log *p* diagram can be found in [18].

Specific Heat. In the range of – 45 °C to 50 °C the specific heat (in kJ kg^{-1} K^{-1}) of liquid ammonia can be calculated using the equation:

$$c_p = -3.787 + 0.0949\,T - 0.3734 \times 10^{-3}\,T^2 + 0.5064 \times 10^{-6}\,T^3$$

where $T = \vartheta + 273.15$.

Assuming ammonia to be an ideal gas in the range 300 – 2000 K the following equation represents the specific heat:

$$c_p = 1.4780 + 2.09307 \times 10^{-3}\,T - 2.0019 \times 10^{-7}\,T^2 - 8.07923 \times 10^{-11}\,T^3$$

The values of the specific heat at constant pressure in Table 3 have been calculated according to [12] and have been extrapolated to 500 °C using data from [9]. Further data can be found in [13-17], [31] and [32].

Table 3. Specific heat of ammonia at constant pressure in kJ kg^{-1} K^{-1}

***p*, MPa**	**Temperature, °C**													
	–33	**–20**	**–10**	**0**	**10**	**20**	**50**	**100**	**150**	**200**	**250**	**300**	**400**	**500**
0.1	2.324	2.236	2.197	2.175	2.163	2.159	2.174	2.240	2.328	2.422	2.518	2.612	2.797	2.966
0.5	4.468	4.528	4.572	4.618	2.673	2.553	2.383	2.339	2.385	2.459	2.542	2.628	2.815	2.981
1	4.466	4.526	4.569	4.615	4.668	4.736	2.707	2.470	2.456	2.504	2.574	2.653	2.838	2.998
2	4.462	4.521	4.563	4.607	4.659	4.725	3.785	2.787	2.611	2.597	2.636	2.696	2.886	3.028
3	4.458	4.516	4.557	4.600	4.651	4.714	5.039	3.216	2.788	2.697	2.701	2.742	2.934	3.060
4	4.454	4.511	4.552	4.594	4.642	4.703	5.014	3.832	2.994	2.805	2.769	2.788	2.981	3.091
5	4.450	4.506	4.546	4.587	4.634	4.693	4.991	4.826	3.236	2.923	2.840	2.837	3.028	3.121
6	4.446	4.501	4.541	4.580	4.626	4.683	4.969	6.854	3.526	3.050	2.916	2.887	3.067	3.168
7	4.442	4.497	4.535	4.574	4.618	4.673	4.947	6.776	3.882	3.190	2.994	2.939	3.069	3.203
8	4.439	4.492	4.530	4.567	4.610	4.664	4.927	6.551	4.330	3.342	3.077	2.992	3.155	3.238
10	4.432	4.484	4.520	4.555	4.595	4.645	4.889	6.215	5.724	3.695	3.255	3.103	3.242	3.304
15	4.415	4.463	4.496	4.527	4.561	4.603	4.806	5.701	20.410	4.965	3.776	3.407	3.470	3.463
20	4.400	4.444	4.474	4.501	4.530	4.566	4.736	5.396	8.827	6.899	4.399	3.739	3.644	3.607
25	4.387	4.427	4.454	4.477	4.502	4.532	4.676	5.187	6.826	8.074	5.046	4.080	3.812	3.734
30	4.376	4.412	4.435	4.456	4.477	4.502	4.624	5.032	6.048	7.426	5.543	4.397	3.980	3.861
35	4.367	4.399	4.419	4.436	4.454	4.475	4.578	4.912	5.614	6.605	5.741	4.652	4.104	3.967
40	4.361	4.387	4.404	4.418	4.432	4.450	4.538	4.814	5.330	6.033	5.693	4.818	4.229	4.074
45	4.358	4.377	4.391	4.402	4.413	4.427	4.502	4.734	5.128	5.643	5.539	4.894	4.310	4.159
50	4.359	4.369	4.379	4.387	4.395	4.407	4.469	4,665	4.975	5.365	5.364	4.903	4.392	4.244
55	4.367	4.363	4.369	4.374	4.379	4.388	4,440	4.607	4.855	5.157	5.202	4.873	4.442	4.301
60	4.385	4.360	4.361	4.362	4.365	4.370	4.413	4.556	4.757	4.996	5.060	4.824	4.492	4.359

Properties of Saturated Ammonia Liquid and Vapor. Table 4 is derived from [12]. Further data may be found in [9], [15], [17], and [18]. For calculations from reduced data, among others, see [33], [34].

Table 4. Properties of saturated ammonia liquid and vapor

***t*, °C**	***p*, kPa**	**Specific volume**		**Enthalpy**		**Heat of**	**Entropy**	
		Liquid, L/kg	**Vapor, L/kg**	**Liquid, kJ/kg**	**Vapor, kJ/kg**	**vaporization, kJ/kg**	**Liquid, kJ kg^{-1} K^{-1}**	**Vapor, kJ kg^{-1} K^{-1}**
–40	71.72	1.4490	1551.6	180.53	1568.7	1388.1	0.8479	6.8017
–35	93.14	1.4620	1215.4	202.80	1576.5	1373.1	0.9423	6.7105
–30	119.49	1.4754	962.9	225.19	1584.1	1358.9	1.0352	6.6241
–25	151.54	1.4892	770.95	247.69	1591.5	1343.8	1.1266	6.5419
–20	190.16	1.5035	623.31	270.31	1598.6	1328.3	1.2166	6.4637
–15	236.24	1.5184	508.49	293.05	1605.4	1312.4	1.3053	6.3890
–10	290.77	1.5337	418.26	315.91	1611.9	1296.0	1.3927	6.3175
– 5	354.77	1.5496	346.68	338.87	1618.0	1279.1	1.4787	6.2488
0	429.35	1.5660	289.39	361.96	1623.6	1261.7	1.5636	6.1826
5	515.65	1.5831	243.16	385.17	1628.9	1243.8	1.6473	6.1188
10	614.86	1.6009	205.55	408.51	1633.7	1225.2	1.7299	6.0571
15	728.24	1.6194	174.74	432.01	1638.1	1206.1	1.8116	5.9972
20	857.08	1.6387	149.31	455.67	1642.0	1186.3	1.8922	5.9390
25	1002.7	1.6590	128.19	479.52	1645.3	1165.8	1.9721	5.8822
30	1166.6	1.6801	110.54	503.57	1648.1	1144.5	2.0512	5.8267
35	1350.0	1.7023	95.699	527.86	1650.3	1122.4	2.1295	5.7722
40	1554.6	1.7257	83.150	522.40	1651.9	1099.5	2.2075	5.7185
45	1781.7	1.7505	72.484	577.22	1652.8	1075.6	2.2849	5.6656

50	2033.0	1.7766	63.373	602.36	1653.0	1050.6	2.3619	5.6131

Vapor – Liquid Equilibria for the Ammonia, Hydrogen, Nitrogen, Argon, Methane System (Fig. 2). Because of the great importance of absorption processes in synthesis loop engineering (see Section Inert-Gas and Purge-Gas Management), these binary and multicomponent systems have come under experimental and theoretical reexamination. The new efforts are characterized above all by consideration of thermodynamic relationships in combination with equations of state.

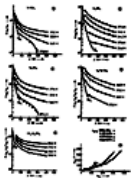

Figure 2. Equilibrium solubilities (*K* factors) for argon (A), hydrogen (B), nitrogen (C), hydrogen : nitrogen 3 : 1 (D), and methane : nitrogen in liquid ammonia (E). F shows the dependence of the methane partial pressure ($p \cdot yCH_4$) on the concentration of methane dissolved in the liquid, xCH_4, for three different gas mixtures

A – C and E according to [48], F according to [44], D was developed from data reported in [46]

In [35] various publications are compared with one another and equilibrium methods for the ranges 273 – 323 K and 4.9 – 49 MPa (49 to 490 bar) are reported. Investigations of binary systems are reported in [36-44] and equations and compilations in [45], [46], and [399], among others. The basis for the saturation concentration (Table 5) and solubilities in liquid ammonia (Table 6) are unpublished BASF calculations produced with the help of the Lee – Kesler equation of state [47] and publications [37-46]. The ammonia content of the gas may be calculated with the formulas from [40].

Table 5. Ammonia concentration at saturation in 1 : 3 mixtures of nitrogen and hydrogen, in vol %

t, °C	*p*, MPa							
	5	10	15	20	30	40	50	100
–30	2.9	1.7	1.34	1.15	0.95	0.9	0.87	0.8
–20	4.6	2.70	2.05	1.58	1.5	1.38	1.3	1.2
–10	6.8	4.06	3.10	2.63	2.22	2.03	1.94	1.75
0	10.0	5.80	4.50	3.80	3.20	2.90	2.75	2.5
10	15.0	8.25	6.40	5.36	4.50	4.05	3.84	3.55
20	19.6	11.4	8.6	7.35	6.05	5.55	5.25	4.9
30	20.0	15.2	11.5	9.80	8.20	7.50	7.10	6.70
40	26.5	20.0	15.0	14.0	10.75	9.80	9.40	9.0

Table 6. Solubility of hydrogen and nitrogen in liquid ammonia for 1 : 3 mixtures of nitrogen and hydrogen in cm^3 gas per gram NH_3 (STP)

t, °C		Total pressure, MPa							
		5	10	15	20	30	40	50	100
–30	N_2	0.632	1.161	1.637	2.036	2.743	3.291	3.754	5.317
	H_2	1.291	2.598	3.896	5.157	7.554	8.625	12.380	22.287
–20	N_2	0.864	1.438	2.036	2.568	3.452	4.195	4.195	8.852
	H_2	1.581	3.167	4.759	6.248	9.294	12.518	15.219	33.945
–10	N_2	0.922	1.730	2.488	3.129	4.258	5.207	6.027	8.853
	H_2	1.845	3.869	5.756	7.623	11.444	14.954	18.349	33.945
0	N_2	1.081	2.088	2.994	3.811	5.269	6.457	7.532	11.218
	H_2	2.215	4.560	6.956	9.163	13.740	18.220	22.461	41.863
10	N_2	1.226	2.461	3.583	4.564	6.342	7.800	9.190	14.085
	H_2	2.585	5.490	8.360	11.178	16.590	22.056	27.300	51.783
20	N_2	1.372	2.874	4.414	5.415	7.654	8.303	11.285	17.733
	H_2	2.969	6.423	9.970	13.338	20.142	26.731	33.310	63.530
30	N_2	1.505	3.315	4.954	6.418	9.169	11.619	13.968	22.528
	H_2	3.300	7.557	11.719	15.711	24.272	32.343	40.396	79.345
40	N_2	1.584	3.717	5.767	7.647	10.967	13.954	16.852	28.527
	H_2	3.631	8.761	13.745	18.916	28.028	39.127	49.451	97.253

A mutual interaction of solubilities exists in all multicomponent systems. The interaction with methane is pronounced. This interdependence is treated in detail in [44] and [49-51].

Surface Tension. The following equation, from [52], represents the surface tension of ammonia (σ in Nm^{-1}) over the range –50 to 50 °C.

$$\sigma = 26.55 \times 10^{-3} - 2.3 \times 10^{-4}\vartheta$$

Experimental data are given in [53]. The surface tension of aqueous ammonia solutions over the ranges 0.1 – 1.2 MPa (1 – 12 bar) and 20 – 100 °C can be found in [54].

Dynamic Viscosity. A survey on dynamic viscosity data for the ranges 30 – 250 °C and 0.1 – 80 MPa (1 – 800 bar) appears in [15]; further data occur in [55] and [56]. Experimental viscosity data and correlating equations for the range 448 – 598 K and for pressures up to 12.1 MPa (121 bar) can be found in [57]. Generalized equations for viscosity calculations on refrigerants are presented in [58].

The following, from [52], is a formula for interpolation along the vapor pressure curve for liquid ammonia (viscosity η in N s m^{-2}):

$$\eta = 1.949 \times 10^{-4} - 1.72 \times 10^{-6}\vartheta + 7 \times 10^{-10}\vartheta^2$$

An interpolation formula for ammonia vapor for the temperature range –20 to 150 °C at 0.098 MPa is given in [52]:

$$\eta = 9.83 \times 10^{-6} + 2.75 \times 10^{-8}\vartheta + 2.8 \times 10^{-11}\vartheta^2$$

Reference [59] reports viscosity measurements for the hydrogen – ammonia system; [60] for nitrogen – ammonia, oxygen – ammonia, methane – ammonia, and ethylene – ammonia; and [61] for argon – ammonia.

Thermal Conductivity. In [62], experimental data and calculation methods for ammonia liquid and vapor for the ranges 0.1 – 49 MPa (1 – 490 bar) and 20 – 177 °C are reported. The anomaly in the thermal conductivity at the critical point is discussed in full detail. A nomogram for these data appears in [63]. A third-degree equation for the range 358 – 925 K can be found in [64]. For the liquid phase the following formula may be used to interpolate in the range from –10 to 100 °C [52] (thermal conductivity λ in W m^{-1} K^{-1}):

$$\lambda = 0.528 - 1.669 \times 10^{-2}\vartheta - 6.2 \times 10^{-6}\vartheta^2$$

Also from [52] is the formula for ammonia vapor at 101.3 kPa and –20 to 150 °C:

$$\lambda = 0.0217 + 1.17 \times 10^{-4}\vartheta + 1.87 \times 10^{-2}\vartheta^2$$

The thermal conductivities of gas mixtures of ammonia with argon, neon, hydrogen, and methane are reported in [60] and [65-67].

Solubility in Water. Tables 7 and 8 and Figures 3 and 4 show miscellaneous information. General physicochemical and chemical engineering handbooks can be consulted for further data.

Table 7. Table for determining the percentage ammonia content of aqueous solutions from the density at 15 °C

Density, g/cm^3	Ammonia content	Density, g/cm^3	Ammonia content
1.000	0.00	0.935	17.12
0.995	1.14	0.930	18.64
0.990	2.31	0.925	20.18
0.985	3.55	0.920	21.75
0.980	4.80	0.915	23.35
0.975	6.05	0.910	24.99
0.970	7.31	0.905	26.64
0.965	8.59	0.900	28.33
0.960	9.91	0.895	30.03
0.955	11.32	0.890	31.73
0.950	12.74	0.885	33.67
0.945	14.17	0.880	35.60
0.940	15.63		

Table 8. Heat of solution of ammonia in water, in kJ/mol NH_3

t, °C	Mixture ratio (moles of water per mole ammonia)						
	1	2.33	4	9	19	49	99
0	30.69	34.25	35.17	35.80	36.22	36.47	36.09
20	27.38	32.87	33.66	34.50	34.67	34.83	34.92
40	24.53	31.99	32.91	33.62	33.87	33.95	34

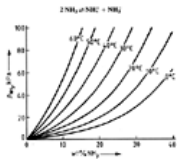

Figure 3. Ammonia partial pressure in aqueous ammonia solutions

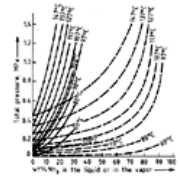

Figure 4. Total vapor pressure and composition of the liquid (——) and vapor (– – –) in the system NH_3 –H_2O

Isotherms

Reference [68] gives *p* – *V* – *T* values for ammonia containing up to 0.5 wt % water at pressures in the range 113 – 221 kPa.

Solubility in Liquid Ammonia. Liquid ammonia is a good solvent for many salts as well as for some metals and nonmetals. Because of its dipole moment, the ammonia molecule interacts with these ions to form solvates in a manner analogous to the water molecule in aqueous solutions. Solutions in liquid ammonia show significant electrical conductance. Pure ammonia, like water, itself has a conductivity that, although limited, is based on dissociation according to:

$$2\,NH_3 \rightleftharpoons NH_4^+ + NH_2^-$$

Inorganic salts such as nitrates, nitrites, iodides, cyanides, thiocyanides, and acetates are very soluble in liquid ammonia. The hydroxides, fluorides, and salts with di- and trivalent anions, such as oxides and sulfides, in general are insoluble. Ammonium salts are especially soluble.

A compilation of the solubilities of organic compounds in liquid ammonia shows notable solubility of saccharoses [76].

Liquid ammonia in combination with water as antisolvent and methylamine as prosolvent is also an excellent extraction medium for all types of petroleum fractions [81].

[Top of Page]

3. Chemical Properties

The ionic complex NH_4^+ results from addition of a proton to the ammonia molecule. This ammonium ion is similar to the alkali metal ions in its alkaline and salt-forming attributes.

The free radical NH_4 has been isolated only in the form of its amalgam (DAVY, 1811) which precipitates the more noble metals, such as copper, cadmium, and zinc, from their solutions. Solutions of ammonium in liquid ammonia are blue and usually exhibit properties similar to liquid ammonia solutions of sodium and potassium.

Aqueous solutions of ammonia react as weak bases because hydroxyl ions result from the addition of a proton or H_3O^+ from the water to the ammonia molecule. However, the equilibrium is very strongly shifted to the side of free ammonia, as may be seen from the equilibrium constant at 18 °C:

$$\frac{[NH_4^+][OH^-]}{[NH_3]} = K[H_2O] = K'_{NH_3} = 1.75 \times 10^{-5}$$

A 1 N solution of ammonium hydroxide has a pH of 11.77 – at 18 °C.

Ammonium salts result from reaction with acids in aqueous solution or in the gas phase (⟶ Ammonium Compounds).

Ammonia is oxidized by *oxygen* or air, depending on the reaction conditions, to NO, NO_2, or N_2O or to nitrogen and water. On alkaline surfaces, such as quicklime or soda lime, ammonia – air mixtures are oxidized readily at 350 °C to nitrite and further to nitrate. The reaction rate can be increased by adding traces of nickel and cobalt oxides to the catalyst surfaces [69]. Gaseous ammonia reacts very violently or even explosively with nitrogen oxides to form nitrogen, water, ammonium nitrate or nitrite. The reaction with N_2O requires ignition. The controlled catalytic oxidation of ammonia to produce nitric acid or pure NO for hydroxylamine production is described in ⟶ Nitric Acid, Nitrous Acid, and Nitrogen Oxides – Industrial Production; ⟶ Hydroxylamine – Catalytic Hydrogenation of Nitric Oxide.

At higher temperatures, especially in the presence of surface-active materials, ammonia forms *hydrogen cyanide* with carbon monoxide, methane, or charcoal. The catalytic oxidation of methane in the presence of ammonia is employed for the industrial production of hydrogen cyanide. With CO_2, SO_3, and P_2O_5, NH_3 forms amides of carbonic acid (⟶ Urea), sulfuric acid (⟶ Sulfuric Acid and Sulfur Trioxide), and phosphoric acid (⟶ Phosphoric Acid and Phosphates).

Metals may replace one or all of the hydrogen atoms of ammonia. The *amide*, for example, sodium amide, $NaNH_2$, is obtained with evolution of hydrogen by passing ammonia vapor over metallic sodium. Sodium amide reacts with N_2O to form highly explosive sodium *azide*. An ammonium azide, NH_4N_3, may be prepared also. The *imide*, for example, lithium imide, Li_2NH, results from replacing a further hydrogen atom by metals. Finally, there are the *nitrides*, such as lithium or magnesium nitride, Li_3N or Mg_3N_2. The nitrides of reactive metals, such as lithium, magnesium, calcium, aluminum, and titanium, are formed directly from the elements at elevated temperature. In many cases, it is better to obtain nitrides through the action of ammonia on the metals or metal compounds at higher temperature. A nitride also may result from thermal decomposition of an amide.

Sulfur, phosphorus, and halogens also can take the place of the hydrogen atoms in ammonia. At relatively low temperatures, halogens react with ammonia to form *nitrogen – halogen compounds* and *ammonium halides*. At higher temperatures the products are nitrogen and hydrogen halides. With sodium hypochlorite, ammonia forms *hydrazine hydrate* (Raschig synthesis) with chloramine, H_2NCl, as an intermediate.

Ammonia undergoes numerous industrially important reactions with organic compounds. Reaction with alkyl halides or alcohols serves to manufacture *amines* or *imines*. For example, methanol forms mono- through trimethylamine; dichloromethane yields ethylene imine in the presence of calcium oxide. With organic acid halides, the reaction leads to the *acid amide*, with elimination of hydrogen chloride. Likewise, acid amides result from acylating ammonia with esters, acid anhydrides, or the acids themselves (above 100 °C); for example, formamide from methyl formate. Catalytic hydrogenation converts the acid amides to *primary amines*. Adding ammonia to aldehydes and ketones and splitting off water leads through unstable, intermediate amino compounds to three different forms of stable end products: for example, on reaction with formaldehyde, *hexamethylenetetramine* (urotropine) is obtained; with acetaldehyde, *ammono acetaldehyde*, which can again be dissociated hydrolytically; and in the case of benzaldehyde, *hydrobenzamide*. With ethylene and propylene oxides, aqueous ammonia reacts to form *ethanol*- or *propanolamine*. By reaction of olefins with ammonia, especially employing corona electrical discharges, *alkylamines* can be obtained, e.g., up to 10 % yield of ethylamine, or 15 % of 1-aminopropane. Otherwise, these compounds may be prepared by decomposing the acid amide or by reacting ammonia with alkyl halides in multistage processes [70]. The catalytic gas-phase oxidation of olefins in the presence of ammonia on vanadium- or molybdenum-containing catalysts, so-called *ammonoxidation*, results in economic yields of the commercially important *acid nitriles*, for example, acrylonitrile from propylene [71]. Ammonoxidation of *o*-xylene yields phthalodinitrile in a single reaction step.

Reactions in Liquid Ammonia. Reactions in liquid ammonia [72] sometimes proceed differently from those in water because of the solubility difference of many salts between water and ammonia. For example, the reaction

$$2\,AgNO_3 + BaCl_2 \underset{NH_3}{\overset{H_2O}{\rightleftharpoons}} 2\,AgCl + Ba(NO_3)_2$$

leads in water to precipitation of silver chloride, and in ammonia, to precipitation of barium chloride. In liquid ammonia, ammonium salts have an “acid” character. For example, alkali and alkaline-earth metals, cerium, lanthanum, manganese, cobalt, nickel, and iron and its alloys are soluble in solutions of ammonium bromide, iodide, cyanide, and thiocyanide, with formation of the corresponding metal amine salt and evolution of hydrogen:

$$2\,NH_4^+ + Mg \rightarrow Mg^{2+} + 2\,NH_3 + H_2$$

Correspondingly, metal amides in liquid ammonia have a “basic” character. The reaction of ammonium salt with metal amide in liquid ammonia is analogous to the neutralization of acid and base in water. The heats of neutralization in ammonia are even larger than in water. The process of hydrolysis corresponds to ammonolysis in ammonia. This results in ammonobasic compounds, for example, in the infusible precipitate $HgNH_2Cl$ from $HgCl_2$.

Liquid ammonia is a good solvent for white phosphorus. Rhombic sulfur is dissolved by liquid ammonia at –11.5 °C. Once dissolved, at lower temperature a blue solution results in which partial ammonolysis of sulfur to sulfur nitride and ammonium sulfides or polysulfides takes place.

Liquid ammonia's ability to dissolve alkali and alkaline-earth metals has been well known for a long time. In concentrated solutions, the metals largely remain in the metallic state. The magnetic properties and the electrical conductivity, which is comparable to that of mercury, confirm this. In the more dilute blue solutions, the metals are completely dissociated to positive metal ions and solvated electrons [73]. The ammoniacal solutions allow preparation of many compounds otherwise unobtainable or obtainable only with difficulty. Examples are the sodium selenides, Na_2Se through Na_2Se_6, tellurides up to Na_2Te_4, antimonides, arsenides, polystannides, and polyplumbides. Introducing oxygen into the solution of alkali metals forms peroxides, Na_2O_2, K_2O_2, as well as dioxides, KO_2, RbO_2, and CsO_2. White explosive salts of the hypothetical acetylenediols, such as KO–C≡C–OK, result with carbon monoxide [74]; with acetylene, NaC≡CH; and with phosphine, KPH_2. Reactions of ammoniacal metal solutions with halogen-containing organic compounds can be used for the synthesis of higher hydrocarbons, amines, and free radicals. For example, reacting triphenylchloromethane with dissolved sodium produces the triphenylmethyl radical $Ph_3C^{\bullet}$.

A survey of organic reactions in liquid ammonia appears in [75].

[Top of Page]

4. Production

Ammonia is the second largest synthetic chemical product; more than 90 % of world consumption is manufactured from the elements nitrogen and hydrogen in a catalytic process originally developed by FRITZ HABER and CARL BOSCH using a promoted iron catalyst discovered by ALWIN MITTASCH. Since the early days there has been no fundamental change in this process. Even today the synthesis section of practically every ammonia plant has the same basic configuration as the first plants. A hydrogen – nitrogen mixture reacts over the iron catalyst (today's formulation differs little from the original) at elevated temperature in the range of 400 – 500 °C (originally up to 600 °C) and pressures above 100 bar with recycle of the unconverted part of the synthesis gas and separation of the ammonia product under high pressure. End of the 1990s a ruthenium-based catalyst was introduced which allows slightly lower synthesis pressure and is used in a few world-size plants (KAAP process of Kellogg – Braun – Root [1392-1394].

BOSCH was already well aware that the production of a pure hydrogen – nitrogen mixture is the largest single contributor to the total production cost of ammonia [93]. So, in contrast to the synthesis reaction, dramatic changes happened over the years in the technology of synthesis-gas generation, and technical ammonia processes differ today mainly with respect to synthesis-gas preparation and purification. The elements nitrogen and hydrogen are abundantly available in the form of air and water, from which they can be separated by physical methods and/or chemical reactions using almost exclusively fossil energy. The predominant fossil fuels are natural gas, liquefied petroleum gas (LPG), naphtha, and higher petroleum fractions; coal or coke is used only under special economic and geographical conditions (China, India, South Africa). Recovery of ammonia as byproduct of other production processes, e.g., coke ovens, is no longer of great importance.

Of course, some of the hydrogen comes also from the hydrocarbons themselves (methane has the highest content), and the carbon acts as a reducing agent for water and in some processes may also facilitate the separation of oxygen from nitrogen by formation of carbon dioxide which can be removed by various operations.

4.1. Historical Development

The catalytic synthesis of ammonia from its elements is one of the greatest achievements of industrial chemistry. This process not only solved a fundamental problem in securing our food supply by production of fertilizers but also opened a new phase of industrial chemistry by laying the foundations for subsequent high-pressure processes like methanol synthesis, oxo synthesis, Fischer – Tropsch Process, coal liquefaction, and Reppe reactions. Continuous ammonia production with high space yields on large scale combined with the ammonia oxidation process for nitric acid, which was developed immediately after ammonia synthesis, enabled the chemical industry for the first time to compete successfully against a cheap natural bulk product, namely, sodium nitrate imported from Chile.

The driving force in the search for methods of nitrogen fixation, of course, was to produce fertilizers. In principle there are three ways of breaking the bond of the nitrogen molecule and fixing the element in a compound:

1. To combine the atmospheric elements nitrogen and oxygen directly to form nitric oxides
2. To combine nitrogen and hydrogen to give ammonia
3. To use compounds capable of fixing nitrogen in their structure under certain reaction conditions.

A vast amount of research in all three directions led to commercial processes for each of them: the electric arc process, the cyanamid process, and ammonia synthesis, which finally displaced the other two and rendered them obsolete.

The availability of cheap hydrolelectric power in Norway and the United States stimulated the development of the *electric arc process*. Air was passed through an electric arc which raised its temperature to 3000 °C, where nitrogen and oxygen combine to give nitric oxide. In 1904 CHRISTIAN BIRKELAND performed successful experiments and, together with SAM EYDE, an industrial process was developed and a commercial plant was built, which by 1908 was already producing 7000 t of fixed nitrogen. Working in parallel, SCHOENHERR at BASF developed a different electric arc furnace in 1905. The Norwegians and BASF combined forces in 1912 to build a new commercial plant in Norway. However, since at this time pilot-plant operation of ammonia synthesis was already successful, BASF withdrew from this joint venture soon after. Nevertheless, the Norwegian plants operated throughout World War I and had total production of 28 000 t/a of fixed nitrogen with a power consumption of 210 000 kW [33]. The specific energy consumption was tremendous: 60 000 kW per tonne of fixed nitrogen. Had this electricity been generated from fossil fuels this figure would correspond to about 600 GJ per tonne nitrogen, which is about 17 times the consumption of an advanced steam-reforming ammonia plant in 1996.

The cyanamide process, [33], [77-79], developed by FRANK and CARO in 1898, was commercially established by 1910. Calcium carbide, formed from coke and lime in a carbide furnace (⟶ Calcium Carbide), reacts with nitrogen to give calcium cyanamide, which can be decomposed with water to yield ammonia. The process was energetically very inefficient, consuming 190 GJ per tonne of ammonia. Some other routes via barium cyanide produced from barytes, coke and nitrogen, or using the formation of titanium nitride were investigated in Ludwigshafen by BOSCH and MITTASCH but did not appear promising. In 1934, 11 % of world nitrogen production (about 2×10^6 t/a) [80] was still based on the cyanamide process, and some plants even continued to operate after World War II.

After BERTHOLLET proved in 1784 that ammonia consists of nitrogen and hydrogen and was also able to establish the approximate ratio between these elements, many experiments in the 1800s were aimed at its *direct synthesis*, but remained unsuccessful [82-84]. One of the reasons for the lack of success was the limited knowledge of thermodynamics and the incomplete understanding of the law of mass action and chemical equilibrium. It was the new science of physical chemistry, which developed rapidly in the late 1800s, that enabled chemists to investigate ammonia formation more systematically.

Around 1900 FRITZ HABER began to investigate the ammonia equilibrium [85] at atmospheric pressure and found minimal ammonia concentrations at around 1000 °C (0.012 %). Apart from HABER, OSTWALD and NERNST were also closely involved in the ammonia synthesis problem, but a series of mistakes and misunderstandings occurred during the research. For example, OSTWALD withdrew a patent application for an iron ammonia synthesis catalyst because of an erroneous experiment, while NERNST concluded that commercial ammonia synthesis was not feasible in view of the low conversion he found when he first measured the equilibrium at 50 – 70 bar [257], [86].

After a controversy with NERNST, HABER repeated his measurements at atmospheric pressure and subsequently at higher pressures [87-90], overcoming his colleague's preoccupation with the unfavorable equilibrium concentrations. HABER concluded that much higher pressures had to be employed and that, perhaps more importantly, a recycle process was necessary.

The amount of ammonia formed in a single pass of the synthesis gas over the catalyst is much too small to be of interest for an economic production. HABER therefore recycled the unconverted synthesis gas. After separating the ammonia by condensation under synthesis pressure and supplementing with fresh synthesis gas to make-up for the portion converted to ammonia, the gas was recirculated by means of a circulation compressor to the catalyst-containing reactor. This process, described in the patent DRP 235 421 (1908), became the basis for the industrial manufacture of ammonia and since then the same principle has found widespread application in numerous high-pressure processes. HABER also anticipated the preheating of the synthesis gas to reaction temperature (at that time 600 °C) by heat exchange with the hot exhaust gas from the reactor, the temperature of which would be raised by the exothermic ammonia formation reaction sufficiently (about 18 °C temperature rise for a 1 % increase of the ammonia concentration in converted synthesis gas).

In 1908 HABER approached BASF (Badische Anilin & Soda Fabrik at that time) to seek support for his work and to discuss the possibilities for the realization of an industrial process. His successful demonstration in April 1909 of a small laboratory scale ammonia plant having all the features described above finally convinced the BASF representatives, and the company's board decided to pursue the technical development of the process with all available resources. In an unprecedented effort, CARL BOSCH, together with a team of dedicated and highly skilled co-workers, succeeded in developing a commercial process in less than five years [91-98]. The first plant started production at Oppau in September 1913 and had a daily capacity of 30 t of ammonia. Expansions increased the capacity to about 250 t/d in 1916/17 and a second plant with a capacity of 36 000 t/a went on stream in 1917 in Leuna. Further stepwise expansions, finally reaching 240 000 t/a, already decided in 1916, came in full production only after World War I [77]. After World War I ammonia plants were built in England, France, Italy, and many other countries based on a BASF license or own process developments, with modified process parameters, but using the same catalyst.

Up to the end of World War II, plant capacities were expanded by installing parallel lines of 70 – 120-t/d units, and synthesis-gas generation continued to be based on coal until the 1950s. With growing availability of cheap petrochemical feedstocks and novel cost-saving gasification processes (steam reforming and partial oxidation) a new age dawned in the ammonia industry. The development started in the USA, where steam reforming of natural gas was used for synthesis gas production. This process was originally developed by BASF and greatly improved by ICI, who extended it to naphtha feedstocks. Before natural gas became available in large quantities in Europe, partial oxidation of heavy oil fractions was used in several plants. The next revolution in the ammonia industry was the advent of the single-train steam reforming ammonia plants, pioneered by M. W. KELLOGG and others. The design philosophy was to use a single-train for large capacities (no parallel lines) and to be as far as possible energetically self-sufficient (no energy import) by having a high degree of energy integration (process steps in surplus supplying those in deficit). Only through this innovative plant concept with its drastic reduction in feedstock consumption and investment costs, could the enormous increase in world capacity in the following years became possible. Increasing competition and rising feedstock prices in the 1970s and 1980s forced industry and engineering companies to improve the processes further.

The LCA process of ICI (ICI LCA Process.) and the KRES/KAAP process, which is the first process since 1913 to use a non-iron synthesis catalyst, are recent advances that make a radical breakaway from established technology.

A short survey of the history of ammonia process technology can be found in [95], [96], [99], [100], [406], [1259], [1318], [1409].

4.2. Thermodynamic Data

Ammonia synthesis proceeds according to the following reaction:

$$0.5\,N_2 + 1.5\,H_2 \rightleftharpoons NH_3 \quad \Delta H_{298} = -46.22\,kJ/mol$$

To fix a kilogram of nitrogen in ammonia requires reacting 2.4 m^3 (STP) of hydrogen and 0.8 m^3 (STP) of nitrogen. About 3.27 MJ of heat is evolved. Table 9 is a compilation of the most important thermodynamic data for the reaction at atmospheric pressure.

Table 9. Thermodynamic data for the reaction 0.5 N_2+1.5 H_2 ⇌ NH_3 at atmospheric pressure

	t, K							
	300	400	500	600	700	800	900	1000
ΔH, kJ	–46.35	– 48.48	–50.58	–52.04	–53.26	–54.28	–55.06	–55.68
ΔS, J K^{-1}	–99.35	–105.63	–110.03	–112.71	–114.55	–115.89	–116.77	–117.48
$\Delta G/T$, J K^{-1}	–55.22	–15.66	8.88	26.25	38.48	48.02	55.56	61.81
c_p, NH_3, J mol^{-1} K^{-1}	35.52	38.80	41.97	45.04	47.98	50.80	53.49	56.03
c_p, H_2, J mol^{-1} K^{-1}	28.868	29.199	29.278	29.341	29.454	29.634	29.89	30.216
c_p, N_2, J mol^{-1} K^{-1}	29.144	29.270	29.601	30.132	30.777	31.451	32.117	32.724

Chemical Equilibrium. The reaction equilibrium has been investigated experimentally and theoretically many times. Values for the equilibrium constant are available for pressures up to 350 MPa (3500 bar).

GILLESPIE and BEATTIE [101] (see also [14]) were by far the most successful experimentally in establishing a firm basis for an analytical expression of the equilibrium constant in the range of industrial interest. The values in Tables 10 and 11 were calculated using their equation. A detailed description, with literature data and many tables, appears in [14]. A description of the equilibrium using the REDLICH – KWONG equation of state is given in [400].

Table 10. Ammonia content (in mol %) in equilibrium synthesis gas; N_2 : H_2=1 : 3

t, °C	p_{abs}, MPa										
	5	10	20	30	40	50	60	70	80	90	100
300	39.38	52.79	66.43	74.20	79.49	83.38	86.37	88.72	90.61	92.14	93.39
310	36.21	49.63	63.63	71.75	77.35	81.51	84.73	87.29	89.35	91.03	92.42
320	33.19	46.51	60.79	69.23	75.12	79.53	82.98	85.74	87.98	89.83	91.35
330	30.33	43.45	57.92	66.64	72.79	77.46	81.13	84.09	86.52	88.52	90.20
340	27.64	40.48	55.04	63.99	70.39	75.29	79.18	82.34	84.95	87.12	88.94
350	25.12	37.60	52.17	61.31	67.93	73.04	77.14	80.49	83.28	85.62	87.59
360	22.79	34.84	49.33	58.61	65.41	70.72	75.01	78.55	81.52	84.02	86.15
370	20.64	32.21	46.53	55.89	62.85	68.33	72.80	76.52	79.66	82.33	84.61
380	18.67	29.71	43.79	53.19	60.26	65.89	70.53	74.42	77.72	80.54	82.97
390	16.87	27.36	41.12	50.50	57.66	63.41	68.19	72.23	75.69	78.67	81.25
400	15.23	25.15	38.53	47.86	55.06	60.91	65.81	69.99	73.59	76.71	79.43
410	13.74	23.08	36.04	45.26	52.47	58.39	63.40	67.69	71.42	74.68	77.54
420	12.40	21.16	33.65	42.72	49.91	55.87	60.96	65.36	69.20	72.58	75.57
430	11.19	19.38	31.37	40.26	47.39	53.37	58.50	62.99	66.93	70.43	73.53
440	10.10	17.74	29.20	37.87	44.92	50.88	56.05	60.60	64.63	68.22	71.43
450	9.12	16.23	27.15	35.57	42.50	48.43	53.61	58.20	62.29	65.97	69.28
460	8.24	14.84	25.21	33.36	40.16	46.03	51.19	55.80	59.95	63.69	67.08
470	7.46	13.57	23.39	31.26	37.89	43.67	48.81	53.42	57.60	61.39	64.85
480	6.75	12.41	21.69	29.55	35.71	41.38	46.46	51.06	55.25	59.09	62.60
490	6.12	11.36	20.10	27.34	33.61	39.16	44.17	48.74	52.92	56.78	60.33
500	5.56	10.39	18.61	25.54	31.60	37.02	41.94	46.46	50.62	54.48	58.06
510	5.05	9.52	17.24	23.84	29.68	34.95	39.77	44.22	48.36	52.20	55.80
520	4.59	8.72	15.96	22.24	27.86	32.97	37.68	42.05	46.13	49.96	53.55
530	4.19	8.00	14.77	20.74	26.13	31.07	35.65	39.94	43.96	47.75	51.32
540	3.82	7.34	13.68	19.34	24.49	29.26	33.71	37.89	41.84	45.58	49.13
550	3.49	6.74	12.67	18.02	22.95	27.54	31.85	35.92	39.79	43.47	46.97
560	3.20	6.20	11.74	16.80	21.49	25.90	30.06	34.02	37.80	41.41	44.86
570	2.93	5.70	10.88	15.65	20.13	24.35	28.37	32.20	35.88	39.41	42.81
580	2.69	5.26	10.09	14.59	18.84	22.88	26.75	30.46	34.04	37.48	40.81
590	2.47	4.85	9.36	13.60	17.64	21.50	25.22	28.80	32.26	35.62	38.87
600	2.28	4.48	8.69	12.69	16.52	20.20	23.76	27.22	30.57	33.83	37.00

Table 11. Equilibrium ammonia content (in mol %) in the presence of inert gases (2.5 mol % Ar; 7.5 mol % CH_4); N_2 : H_2=1 : 3

t, °C	p_{abs}, MPa										
	5	10	20	30	40	50	60	70	80	90	100
300	32.11	43.10	54.27	60.62	64.95	68.12	70.56	72.49	74.03	75.28	76.31

310	29.52	40.51	51.98	58.63	63.20	66.59	69.22	71.31	72.99	74.37	75.51
320	27.05	37.96	49.66	56.57	61.38	64.98	67.79	70.04	71.87	73.38	74.63
330	24.71	35.46	47.32	54.45	59.48	63.29	66.28	68.70	70.67	72.31	73.68
340	22.51	33.03	44.96	52.30	57.52	61.52	64.69	67.27	69.39	71.16	72.65
350	20.46	30.68	42.62	50.11	55.51	59.68	63.02	65.76	68.03	69.93	71.54
360	18.56	28.42	40.30	47.90	53.46	57.79	61.29	64.18	66.59	68.63	70.36
370	16.80	26.27	38.01	45.69	51.37	55.85	59.49	62.52	65.08	67.24	69.10
380	15.19	24.33	35.77	43.48	49.26	53.86	57.64	60.81	63.49	65.79	67.77
390	13.72	22.31	33.59	41.29	47.14	51.85	55.74	59.03	61.84	64.27	66.36
400	12.39	20.50	31.48	39.13	45.02	49.81	53.81	57.21	60.14	62.68	64.89
410	11.18	18.81	29.44	37.01	42.91	47.76	51.84	55.34	58.38	61.03	63.35
420	10.08	17.24	27.49	34.93	40.83	45.71	49.86	53.44	56.57	59.32	61.75
430	9.10	15.79	25.62	32.92	38.77	43.67	47.86	51.52	54.73	57.57	60.09
440	8.21	14.45	23.85	30.97	36.75	41.64	45.87	49.58	52.86	55.78	58.39
450	7.41	13.22	22.17	29.09	34.78	39.65	43.88	47.63	50.97	53.96	56.65
460	6.70	12.09	20.59	27.29	32.87	37.68	41.92	45.68	49.06	52.11	54.87
470	6.06	11.05	19.10	25.56	31.02	35.77	39.97	43.75	47.16	50.25	53.06
480	5.49	10.11	17.71	23.92	29.23	33.90	38.06	41.83	45.25	48.38	51.24
490	4.97	9.24	16.41	22.36	27.51	32.08	36.20	39.94	43.36	46.51	49.40
500	4.51	8.46	15.19	20.89	25.87	30.33	34.37	38.08	41.49	44.64	47.56
510	4.10	7.74	14.07	19.50	24.30	28.64	32.61	36.26	39.65	42.79	45.72
520	3.73	7.09	13.02	18.19	22.81	27.02	30.89	34.49	37.84	40.97	43.90
530	3.40	6.50	12.05	16.96	21.39	25.46	29.24	32.76	36.07	39.17	42.09
540	3.10	5.97	11.16	15.81	20.05	23.98	27.65	31.09	34.34	37.41	40.31
550	2.83	5.48	10.33	14.73	18.79	22.57	26.12	29.48	32.66	35.68	38.56
560	2.59	5.04	9.57	13.73	17.60	21.23	24.67	27.93	31.04	34.01	36.84
570	2.38	4.64	8.87	12.79	16.48	19.96	23.28	26.44	29.47	32.38	35.17
580	2.18	4.27	8.22	11.92	15.43	18.76	21.95	25.01	27.96	30.80	33.54
590	2.01	3.94	7.63	11.12	14.44	17.63	20.69	23.65	26.51	29.28	31.96
600	1.85	3.64	7.08	10.37	13.52	16.56	19.50	22.35	25.12	27.81	30.43

Heat of Reaction. HABER investigated the heat of reaction at atmospheric pressure [102]. Numerous authors have estimated the pressure dependence under various assumptions. Today, most people use the Gillespie – Beattie equation [103]. This equation was used in calculating the values in Table 12. For further data see, for example, [14]. Reference [104] contains test results for the range 120 – 200 MPa (1200 – 2000 bar) and 450 – 525 °C. Additional literature can be found in [105]

Table 12. Heat of reaction ΔH (in kJ) for the reaction 0.5 N_2+1.5 H_2 $\rightleftharpoons$ NH_3, from [101]

t, °C	p_{abs}, MPa										
	0 (ideal)	10	20	30	40	50	60	70	80	90	100
0	44.39	55.20	66.02	76.84	87.66	98.47	109.29	120.11	130.92	141.74	152.56
25	44.92	53.48	62.03	70.59	79.15	87.70	96.26	104.82	113.38	121.93	130.49
50	45.44	52.37	59.30	66.23	73.16	80.09	87.02	93.95	100.88	107.80	114.73
75	45.96	51.69	57.41	63.14	68.86	74.58	80.31	86.03	91.76	97.48	103.20
100	46.47	51.28	56.10	60.91	65.72	70.53	75.34	80.16	84.97	89.78	94.59
125	46.97	51.08	55.19	59.29	63.40	67.51	71.62	75.72	79.83	83.94	88.04
150	47.46	51.02	54.57	58.12	61.68	65.23	68.78	72.34	75.89	79.44	83.00
175	47.95	51.06	54.17	57.28	60.39	63.50	66.61	69.72	72.83	75.94	79.05
200	48.42	51.17	53.92	56.68	59.43	62.18	64.93	67.69	70.44	73.19	75.94
225	48.88	51.34	53.80	56.26	58.71	61.17	63.63	66.09	68.55	71.01	73.47
250	49.33	51.54	53.76	55.97	58.19	60.40	62.62	64.84	67.05	69.27	71.48
275	49.76	51.77	53.79	55.80	57.81	59.82	61.84	63.85	65.86	67.87	69.88
300	50.18	52.02	53.86	55.70	57.54	59.38	61.22	63.06	64.90	66.74	68.59
325	50.59	52.28	53.98	55.67	57.37	59.06	60.75	62.45	64.14	65.83	67.53
350	50.99	52.55	54.12	55.69	57.26	58.82	60.39	61.96	63.53	65.10	66.66
375	51.36	52.82	54.28	55.74	57.20	58.66	60.12	61.58	63.04	64.49	65.95
400	51.73	53.09	54.46	55.82	57.18	58.55	59.91	61.28	62.64	64.00	65.37
425	52.07	53.35	54.64	55.92	57.20	58.48	59.76	61.04	62.32	63.60	64.88
450	52.40	53.61	54.82	56.03	57.23	58.44	59.65	60.86	62.06	63.27	64.48
475	52.71	53.86	55.00	56.14	57.29	58.43	59.57	60.72	61.86	63.00	64.14
500	53.01	54.09	55.18	56.26	57.35	58.43	59.52	60.60	61.69	62.78	63.86
525	53.28	54.31	55.35	56.38	57.42	58.45	59.48	60.52	61.55	62.58	63.62
550	53.53	54.52	55.51	56.50	57.48	58.47	59.46	60.45	61.44	62.42	63.41
575	53.76	54.71	55.66	56.60	57.55	58.50	59.44	60.39	61.33	62.28	63.23
600	53.98	54.88	55.79	56.70	57.61	58.52	59.43	60.34	61.24	62.15	63.06

Physical Properties of the Reactants. Various authors have measured the *p* – *V* – *T* behavior of nitrogen, hydrogen, and 3 : 1 hydrogen – nitrogen mixtures. Reference [14] contains a survey. For the specific heat, thermal conductivity, and viscosity of the reactants, see [106], [107]. It is to be noted that heats of mixing must be considered for synthesis gas. This applies especially for high pressures and low temperatures. Reference [108] gives viscosities of hydrogen – nitrogen and hydrogen – ammonia mixtures.

4.3. Ammonia Synthesis Reaction

4.3.1. *General Aspects*

Usually, a system having an exothermic heat of reaction under operating conditions should react spontaneously. However, to form ammonia from hydrogen and nitrogen molecules, significant energy input is required for the nitrogen molecule to achieve the activated state. This is because of its high dissociation energy of 941 kJ/mol, which is considerably higher than that of hydrogen. According to estimates [109], initiation of ammonia synthesis homogeneously in the gas phase

requires an activation energy of 230 – 420 kJ/mol. For purely thermal energy supply with a favorable collision yield, this activation barrier requires temperatures well above 800 – 1200 K to achieve measurable reaction rates. However, at such high temperatures and industrially reasonable pressures, the theoretically achievable ammonia yield is extremely small because of the unfavorable position of the thermodynamic equilibrium. In fact, all older attempts to combine molecular nitrogen purely thermally with atomic or molecular hydrogen failed. On the other hand, both ammonia and hydrazine result from reacting atomic nitrogen with atomic hydrogen (see for example [109, p. 186]).

At pressures above 200 MPa (2000 bar), the synthesis of ammonia proceeds even in the absence of specific catalysts. At such extreme pressures the vessel walls appear to catalyze the formation of ammonia.

In the homogeneous phase under thermodynamically favorable temperature conditions, the formation of ammonia may be forced by employing other forms of energy, such as electrical energy or ionizing radiation. The principal difficulty with these so-called plasma processes, which also impedes their economic use, is that the energy supplied is useful only in part for ammonia formation. A greater part is transformed in primary collision and exothermic secondary processes into undesirable heat or unusable incidental radiation.

In the catalytic combination of nitrogen and hydrogen, the molecules lose their translational degrees of freedom by fixation on the catalyst surface. This drastically reduces the required energy of activation, for example, to 103 kJ/mol on iron [110]. The reaction may then proceed in the temperature range 250 – 400 °C. In 1972, it was discovered that electron donor – acceptor (EDA) complexes permit making ammonia with measurable reaction rate at room temperature.

This section concentrates mainly on the ammonia synthesis reaction over iron catalysts and refers only briefly to reactions with non-iron catalysts. Iron catalysts which are generally used until today in commercial production units are composed in unreduced form of iron oxides (mainly magnetite) and a few percent of Al, Ca, and K; other elements such as Mg and Si may also be present in small amounts. Activation is usually accomplished in situ by reduction with synthesis gas. Prereduced catalysts are also commercially available.

Numerous investigations have been performed to elucidate the mechanism of catalytic reaction of nitrogen and hydrogen to form ammonia. References [111-114] give reviews of the older and some of the newer literature. Since the 1980s a large variety of surface science techniques involving Auger electron spectroscopy, X-ray photoelectron spectroscopy, work-function measurements, temperature-programmed adsorption and desorption, scanning tunnelling microscopy, and others have been developed [117], [118]. Many of these methods are based on interaction of slow electrons, ions, or neutral particles and exhibit high sensitivity to surface structures. With these powerful tools the kinetics of nitrogen and hydrogen adsorption and desorption could be investigated, and it was also possible to identify adsorbed intermediates. The results of these experiments allow the mechanism of ammonia synthesis in the pressure range of industrial interest to be elucidated [115, 116, 119].

As with every catalytic gas-phase reaction, the course of ammonia synthesis by the Haber – Bosch process can be divided into the following steps:

1. Transport of the reactants by diffusion and convection out of the bulk gas stream, through a laminar boundary layer, to the outer surface of the catalyst particles, and further through the pore system to the inner surface (pore walls)
2. Adsorption of the reactants (and catalyst poisons) on the inner surface
3. Reaction of the adsorbed species, if need be with participation by hydrogen from the gas phase, to form activated intermediate compounds
4. Desorption of the ammonia formed into the gas phase
5. Transport of the ammonia through the pore system and the laminar boundary layer into the bulk gas stream

Only the portion of the sequence that occurs on the catalyst surface is significant for the intrinsic catalytic reaction. Of special importance is the adsorption of nitrogen. This assumption is decisive in representing the synthesis reaction kinetics. The transport processes occurring in the pores of the catalyst in accordance with the classical laws of diffusion are of importance in industrial synthesis (see also Sections Particle Size and Shape and Kinetics).

4.3.2. *Catalyst Surface and Reaction Mechanism*

Earlier studies [142, 144-150, 157], [272] had already suggested that on iron catalysts nitrogen adsorption and dissociation can be regarded as the rate-determining step for the intrinsic reaction. This has now been fully confirmed by microkinetic simulations based on the results of surface science studies [115, 129, 151]. In single-crystal experiments it was found that the activation energy for dissociative nitrogen adsorption is not constant; it increases as the surface becomes increasingly covered with adsorbed nitrogen species. Under reaction conditions it can attain values of 63 – 84 kJ/mol [119, 141, 144, 145] which are typical for ammonia synthesis. In the case of nitrogen the first step seems to be an adsorption in molecular state followed by chemisorption in the atomic state. From experimental results it was concluded that for hydrogen a direct transition from the gaseous H_2 molecule into the chemisorbed H_{ad} is most likely, and evidence for the stepwise hydrogenation of surface nitrogen atomic species was found. In IR spectroscopic investigations, NH and NH_2 species were identified [152], and secondary ion emissions detecting NH^+ also confirmed the presence of NH species on the surface [180]. It was further shown that at lower temperatures nitrogen becomes adsorbed only in the molecular state, but subsequently dissociates when the temperature is raised. Isotopic experiments with $^{30}N_2$ and $^{29}N_2$ showed that the surface species resulting from low-temperature adsorption was molecular, whereas that from high-temperature adsorption was atomic [182]. Ammonia synthesis is highly sensitive to the orientation of the different crystal planes of iron in the catalyst [174-178]. Measurements on defined single crystal surfaces of pure iron performed under ultrahigh vacuum [153-155] clearly showed that Fe(111) is the most active surface. This was also demonstrated by the rate of ammonia formation on five different crystallographic planes for unpromoted iron at 20 bar, as shown in Figure 5.

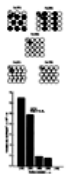

Figure 5. Ammonia formation rate on different iron surfaces [123], [151]

Activation energy, reactivity, adhesion coefficient (the probability that a nitrogen molecule striking the surface will be adsorbed dissociatively) and work function show a clear dependence on surface orientation [123], [173]:

Ammonia synthesis activity:
(111) > (211) > (100) > (210) > (110)
Activation energy of nitrogen adsorption:
(111) > (100) > (110)
Work function:
(210) > (111) > (211) > (100) > (110)
Surface roughness:
(210) > (111) > (211) > (100) > (110)

Ultrahigh vacuum (UHV) experiments with single crystals show that the activation energy of the nitrogen adsorption at zero coverage increases from about zero for Fe (111) to 21 kJ/mol for Fe(110) [115], [144], [170-172]. These values increase significantly for higher coverage [171], [172]. Adsorption and desorption under higher pressure on finely dispersed catalyst indicate that the reaction is highly activated under these conditions with high coverage, about 100 kJ/mol [142], [143].

According to [170], [171] adsorption on the planes Fe(111) and Fe(110) is associated with a regrouping of the surface atoms.

A possible explanation for the high activity of faces (111) and (211) is that these are the only surfaces which expose C_7 sites (iron with seven nearest neighbors) to the reactant gases. There are theoretical arguments [160] that highly coordinated surface atoms should show increased catalytic activity due to low-energy charge fluctuations in the *d*-bands of these highly coordinated atoms. This argument might probably be the key for the special role of C_7 sites. Other reasons discussed are based on charge transfer and interaction of iron *d*-bands with antibonding 2 π^* orbitals of nitrogen [156].

Promotion with potassium of single iron crystals enhances the sticking probability for nitrogen dissociation much more on the Fe(100) and (110) than on the Fe(111)

(factors 280, ca. 1000 and 8, respectively) to the effect that the differences in surface orientation disappear [1316]. A similar effect was not found for the ammonia synthesis at 20 bar and catalyst temperature of 400 °C: only a two-fold increase of the ammonia formation rate was measured for Fe(111) and Fe(100), and the face (110) was found to be inactive with and without potassium [123]. Other experiments [158] show that even the least active face Fe(110) becomes as active for the synthesis as Fe(111) after addition of alumina with subsequent annealing with oxygen and water vapor. A proposed mechanism for these findings — backed by X-ray photoelectron spectroscopy, temperature programmed desorption and electron microscopy — assumes that first alumina forms an iron aluminate $FeAl_2O_4$ on the surface. This new surface then may serve as a template on which iron grows with (111) and (211) orientation upon exposure to the synthesis-gas mixture in the reaction [159].

Based on these experimental results a reaction scheme for the ammonia synthesis may be formulated comprising the following sequence of individual steps [115]:

$$H_2 + * \rightleftharpoons 2\,H_{ad} \quad (1)$$

$$N_2 + * \rightleftharpoons N_{2,ad} \quad (2)$$

$$N_{2,ad} \rightleftharpoons 2\,N_{ad} \quad (3)$$

$$N_{ad} + H_{ad} \rightleftharpoons NH_{ad} \quad (4)$$

$$NH_{ad} + H_{ad} \rightleftharpoons NH_{2,ad} \quad (5)$$

$$NH_{2,ad} + H_{ad} \rightleftharpoons NH_{3,ad} \quad (6)$$

$$NH_{3,ad} \rightleftharpoons NH_3 + * \quad (7)$$

The progress of the reaction may be described in the form of an energy profile, as shown in Figure 6. Industrial ammonia synthesis in the homogeneous gas phase is not feasible because of the high dissociation energies for the initial steps. The reaction over a catalyst avoids this problem since the energy gain associated with the surface atom bonds overcompensates these dissociation energies and the first steps have actually become exothermic.

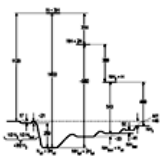

Figure 6. Schematic energy profile of the progress of ammonia synthesis on Fe (energies in kJ/mol) [115]

Dissociative nitrogen adsorption remains nevertheless the rate-determinating step, not so much on account of its activation barrier but rather because of the very unfavorable pre-exponential factor in its rate constant. The subsequent hydrogenation steps are energetically uphill, but the energy differences involved can easily be overcome at the temperatures (ca. 700 K) used in industrial ammonia synthesis. It is, however, quite apparent that the rate-controlling step would switch from dissociative nitrogen adsorption to hydrogenation of adsorbed atomic nitrogen species if the temperature were lowered sufficiently because of these differences in activation energy.

Some critics [161] of the above energy diagram question the low net activation barrier from the gas phase. The arguments are based on an analysis of activation energies from early measurements of the nitrogen adsorption kinetics on singly (Al_2O_3) promoted catalysts and on the results of experiments with supersonic molecular beams [162], [163].

An attempt to explain these differences is given in [164]. In more recent investigations of the adsorbed nitrogen species [166-169] a second molecularly adsorbed species was detected. This so-called α state was interpreted as a bridge-bonded species with electron donation from the surface to the antibonding π levels of N_2, whereas the γ state is regarded as a terminally bound species. Thus the following picture for the nitrogen adsorption emerges:

$$N_2 \rightarrow S^* - N_2(\gamma\,\text{state}) \rightarrow S^* - N_2 - S^*(\alpha\,\text{state}) \rightarrow 2\,S^* - N(\beta\,\text{state})$$

where S* denotes a surface atom.

For industrial catalysts made by careful reduction of magnetite fused with nonreducible oxide promoters the important role of the (111) face seems to be confirmed [122]. However, the question whether the active industrial catalyst exposes mostly (111) faces remains unresolved. If not, further improvements of the catalyst are at least theoretically possible [165]. This was indeed the case with the new AmoMax-10 catalyst [1321], [1322] of Südchemie. This catalyst in its oxidic form is based on wustite instead of magnetite and has a significantly better activity at lower temperature than magnetite. Wustite is a nonstoichiometric iron oxide($Fe_{1-x}O$) with a cubic crystal structure, x ranging from 0.03 to 0.15. With still (111) faces present in the reduced state the catalyst has a higher specific surface area and an improved pore structure.

A critical evaluation of the present knowledge of the mechanism of the synthesis reaction was made by SCHLÖGL [172].

Other reaction mechanisms have been debated for reaction temperatures below 330 °C [126], [177], [179-190]. These propose participation of diatomic nitrogen, or of adsorption complexes containing diatomic nitrogen, in the rate-determining step (see 111, 112 for further literature).

NIELSEN et al. investigated ammonia synthesis on a commercial Topsøe catalyst, KM IR, over a wide temperature range. They found evidence that a different reaction mechanism predominates below and above 330 °C [191]. Also, at low temperatures, chemisorbed hydrogen blocks the catalyst surface [192]. The latter finding is in agreement with the observations of ERTL's group [141].

Reaction Mechanism on Non-Iron Catalysts. Non-iron systems which exhibit some potential to catalyze ammonia synthesis can be divided into the following groups [195]:

- Platinum group metals such as Ru, Os Ir, Pt (no nitrides)
- Mn, Co, Ni, Tc, Rh and their alloys (no nitride formation under synthesis conditions)
- Mn, Mo, V (present as nitrides under the reaction conditions)

In the non-iron systems the rate-determining step is also dissociative adsorption of nitrogen and the catalyst effectivity seems to be primarily dictated by the activation energy of the dissociation reaction [195]. This is somewhat surprising in view of the marked differences in the heats of adsorption of nitrogen and the adsorption activation energy. This even holds for tungsten, which has no significant activation energy and a high adsorption enthalpy for nitrogen, so that hydrogenation of adsorbed atomic nitrogen could be expected to be the rate-determining step. The factor common with the iron catalyst is the structure sensitivity.

The only system which seems to be promising for industrial application is ruthenium promoted with rubidium on graphite as carrier (see Section Commercial Ruthenium Catalysts). Further information on structure, activity and reaction mechanism of non-iron catalysts is given in [112], [195-197], [202]. Specific references: vanadium [193], uranium [194], molybdenum [198-200], tungsten [201].

4.3.3. *Kinetics*

Knowledge of the reaction kinetics is important for designing industrial ammonia synthesis reactors, for determining the optimal operating conditions, and for computer control of ammonia plants. This means predicting the technical dependence on operating variables of the rate of formation of ammonia in an integral catalyst volume element of a converter.

High pressure promotes a high rate of ammonia formation; high ammonia concentration in the synthesis gas (recycle gas) restricts it (Fig. 7). High temperatures accelerate ammonia formation but imply a lower value of the equilibrium ammonia concentration and so a lower driving force. Therefore, the rate of formation at first

increases with rising temperature but then goes through a maximum as the system approaches thermodynamic equilibrium (Fig. 8). A similar situation exists for the dependence of the reaction rate on the ratio of the hydrogen and nitrogen partial pressures; with lower temperature, the maximum rate shifts to a lower hydrogen – nitrogen ratio (Fig. 9). Figure 9 presents data obtained using a commercial iron catalyst, Topsøe KMIR. The data show a sharp drop in reaction rate with declining temperature at $H_2/N_2 = 3:1$ ratio in contrast to a $H_2/N_2 = 1:1$ ratio. This may be attributed to a hindering effect of adsorbed hydrogen at low temperature [192].

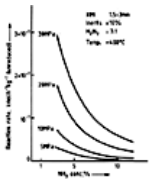

Figure 7. Reaction rate for NH_3 synthesis. Dependence on the ammonia concentration at various pressures.

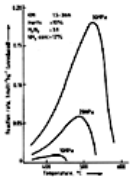

Figure 8. Reaction rate for NH_3 synthesis. Dependence on the temperature at various pressures.

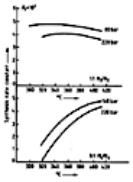

Figure 9. Ammonia synthesis rate constant dependence on hydrogen – nitrogen ratio

Equations for describing ammonia synthesis under industrial operating conditions must represent the influence of the temperature, the pressure, the gas composition, and the equilibrium composition. Moreover, they must also take into consideration the dependence of the ammonia formation rate on the concentration of catalyst poisons and the influence of mass-transfer resistances, which are significant in industrial ammonia synthesis.

Since the beginning of commercial ammonia synthesis, a large number of different kinetic equations have been suggested, emanating in each case from a proposed reaction mechanism or from empirical evaluations. A critical review of the data and equations published up to 1959 appears in [203]. A discussion of kinetics proposed up to 1970, insofar as they have been based on measurements in the operating range of commercial interest, can be found in [126]. An evaluation of present knowledge is given in [115], [127-129].

Contradictory data on the kinetics of ammonia synthesis, especially in the earlier literature, in some circumstances may reflect a lack of attention to the influence of impurities in the gas. If oxygen compounds are present in the synthesis gas, reversible poisoning of the adsorbing areas, in accordance with an equilibrium depending on the temperature and the water vapor – hydrogen partial pressure ratio, must be taken into account when developing rate equations (see also Section Catalyst Reduction).

Experimental Measurements of Reaction Kinetics. The reaction expressions discussed in the following model the intrinsic reaction on the catalyst surface, free of mass-transfer restrictions. Experimental measurements, usually made with very fine particles, are described by theoretically deduced formulas, the validity of which is tested experimentally by their possibility for extrapolation to other reaction conditions. Commonly the isothermal integral reactor is used with catalyst crushed to a size of 0.5 – 1.5 mm to avoid pore diffusion restriction and heat-transfer resistance in the catalyst particles. To exclude maldistribution effects and back mixing, a high ratio of bed length to bed diameter is chosen. Sometimes the catalyst is also diluted with inert material. In some investigations, differential reactors were used. To exclude any poisoning by the synthesis gas, very pure reactants were prepared by decomposition of anhydrous ammonia [127].

Classical Expressions for Ammonia Synthesis Kinetics. The first expression useful for engineering purposes was the Temkin – Pyzhev Equation (1) proposed in 1940 [204], [205]. It is based on the assumption that dissociative adsorption is the rate-determining step, that hydrogen and ammonia have no significant influence on nitrogen adsorption, and that the kinetics of nitrogen adsorption and desorption can be described adequately by Elovich-type adsorption on an energetically inhomogeneous surface. For many years this kinetic expression was the basic design equation for ammonia converters. Values for the factors α between 0.5 [207-209] and 0.75 [14], [210], [211] were used. A problem with this equation was that the α values (reaction order) were dependent on temperature, and the rate constants on pressure [191], [210-214]. More serious (not so much for industrial purposes, where the converter feed has always a certain ammonia content) was the fact that for zero ammonia content, as in some laboratory measurements, the equation gives an infinitely high reaction rate. To avoid this, a simpler expression (Eq. 2) was often used [212], [213].

$$(1)\quad v = k_1 p_{N_2}\left(\frac{p_{H_2}^3}{p_{NH_3}^2}\right)^{\alpha} - k_{-1}\left(\frac{p_{NH_3}^2}{p_{H_2}^3}\right)^{1-\alpha}$$

$\alpha = 0.5-0.75$

$$(2)\quad v = k' p_{H_2}^{\alpha} p_{N_2}^{1-\alpha}$$

Near equilibrium (fugacities instead of partial pressures in Eq. 1)

Higher pressure, some ammonia in the synthesis gas

Far from equilibrium

$$(3)\quad v = \frac{k_{-1}^0\left(a_{N_2}K_a^2 - \frac{a_{NH_3}^2}{a_{H_2}^3}\right)}{\left(1 + K_3\frac{a_{NH_3}}{a_{H_2}^w}\right)^{2\alpha}}$$

$w = 1.5;\ \alpha = 0.75$

$$(4)\quad v = \frac{k p_{N_2}^{1-\alpha}\left(1 - \frac{1}{K}\frac{p_{NH_3}^2}{p_{N_2}p_{H_2}^3}\right)}{\left(\frac{1}{p_{H_2}} + \frac{1}{K}\frac{p_{NH_3}^2}{p_{N_2}p_{H_2}^3}\right)^{\alpha}\left(1 + \frac{1}{p_{H_2}}\right)^{1-\alpha}}$$

An important modification was made by TEMKIN [206] who incorporated hydrogen addition to the adsorbed nitrogen as a second rate-determining step (Eq. 4). ICI demonstrated that this equation gives a better fit with experimental data [218]. It was also shown later that the original Equation (1) is a simplified form of a more general model which can be derived from the concept of energetically homogeneous (Langmuir – Hinshelwood adsorption isotherm) as well as for heterogeneous surfaces (Elovich-type isotherm). The applicability of a particular equation resulting from this concept also depends on the state of reduction of the catalyst [214, 215] and the type of promoter [216]. Equation (3), used by NIELSEN et al. [219], is a combination of these model equations, developed by OZAKI et al. [220], that uses fugacities instead of partial pressures. A similar equation is found in [217]. Additionally, a number of modified equations were proposed and tested with existing experimental data and industrial plant results [221-224]. Near the thermodynamical equilibrium, Equation (3) transforms into Equation (1) [191], [126].

Surface science based ammonia kinetics [115], [116], [127], [128] are presently still viewed as an academic exercise rather than as a practical tool for engineering. The large amount of available data on nitrogen and hydrogen adsorption from ultrahigh-vacuum studies on clean iron surfaces, acquired with all the modern spectroscopic techniques, has prompted some research groups, such as BOWKER et al. [226], [227] and STOLTZE and NØRSKOW [228], [229], to attempt the generation of a kinetic expression for ammonia synthesis from a detailed microscopic model of the reaction mechanism consisting of a number of discrete steps at molecular and atomic level. Potential energy diagrams for the various intermediate steps and species, were set up and Arrhenius expressions for each single step with known or estimated values for all pre-factors and activation energies were formulated. The best results have been achieved so far by calculating the overall rate from the rate of dissociation of the adsorbed nitrogen and equilibrium constants for all other reaction steps. The adsorption – desorption equilibria were treated with approximation of

competitive Langmuir-type adsorption and by evaluation of the partition functions for the gaseous and adsorbed species. The data from single-crystal experiments for potassium-promoted Fe(111) surface were used for the rate of the dissociative nitrogen adsorption. Comparison of the calculated ammonia yields with those determined experimentally on a commercial Topsøe KM1 catalyst showed an agreement within a factor better than 2. Figure 10 demonstrates these encouraging results. These conclusions have been confirmed by calculations results of another independent group [129], [231].

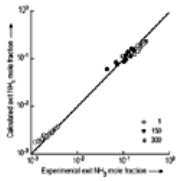

Figure 10. Comparison of ammonia concentrations calculated from surface science kinetics with experimentally measured values [229], [230]

To compete with the empirical models (Temkin and improved expressions) for the best fit to experimental data cannot be the prime objective of the microkinetic approaches. Rather, they are means of checking whether our knowledge and understanding of the elementary steps correspond to the reality of catalysts under industrial synthesis conditions.

Transport Phenomena. For practical application, the above kinetic equations have to be modified to make allowance for mass and energy transfer since the reaction rates actually observed in a commercial converter are lower. One aspect is interparticle mass transfer and heat transfer through the stagnant film which surrounds the catalyst particles. The high velocity of the gas passing through the converter creates sufficient turbulence to keep the film thickness rather small in relation to the catalyst grain size. For this reason the largest concentration gradient (with respect to the concentration in the bulk gas stream) is within the catalyst particles. Since the thermal conductivity of the iron catalysts is much higher than that of the synthesis gas, the major temperature difference is in the external gas film, while the catalyst particles themselves operate under approximately isothermal conditions. As can be seen from Table 13 the differences in temperature and ammonia concentration between the bulk gas stream and the external catalyst surface are small. It also appears that the effects are oppositely directed and will partly compensate each other. So it can be concluded that their combined influence on the reaction rate is negligible compared to inaccuracies of the experimental data for the intrinsic catalyst activity [127].

Table 13. Mass and heat transfer effects at the external surface of catalyst particle

Position in catalyst bed, vol% from inlet	NH_3 concentration in bulk gas, mol%	NH_3 concentration at catalyst surface, mol%	Temperature in bulk gas, °C	Temperature at catalyst surface, °C
0	2.500	2.592	400.0	401.4
20	4.500	4.288	428.1	429.5
40	5.960	6.045	455.1	456.5
60	7.000	7.778	481.9	483.2
80	9.300	9.361	505.5	506.6
100	10.500	10.536	522.7	523.3

For the particle sizes used in industrial reactors (≤ 1.5 mm), intra particle transport of the reactants and ammonia to and from the active inner catalyst surface may be slower than the intrinsic reaction rate and therefore cannot be neglected. The overall reaction can in this way be considerably limited by ammonia diffusion through the pores within the catalysts [225]. The ratio of the actual reaction rate to the intrinsic reaction rate (absence of mass transport restriction) has been termed as pore effectiveness factor *E*. This is often used as a correction factor for the rate equation constants in the engineering design of ammonia converters.

For pore diffusion resistances in reactions having moderate heat evolution, the following phenomena characteristically hold true in industrial ammonia synthesis [232]: in the temperature range in which transport limitation is operative, the apparent energy of activation falls to about half its value at low temperatures; the apparent activation energy and reaction order, as well as the ammonia production per unit volume of catalyst, decrease with increasing catalyst particle size [225][233-235]. For example at the gas inlet to a TVA converter, the effective rate of formation of ammonia on 5.7-mm particles is only about a quarter of the rate measured on very much smaller grains (Fig. 11) [126].

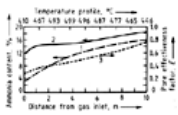

Figure 11. Ammonia content in the bulk stream (1) and in the catalyst pores (at r = 0.5 R) (2) and pore effectiveness factor, E (3)

21.4 MPa; 12 % inerts; SV = 15000 h^{-1}; particle diameter, 2 R = 5.7 mm

Mathematical models for calculating these effectiveness factors involve simultaneous differential equations, which on account of the complex kinetics of ammonia synthesis cannot be solved analytically. Exact numerical integration procedures, as adopted by various research groups [126], [238-240], are rather troublesome and time consuming even for a fast computer. A simplification [241] can be used which can be integrated analytically when the ammonia kinetics are approximated by a pseudo-first-order reaction [234-236], according to the equation:

$$r = k_v\left(c_{NH_3,\text{equilibrium}} - c_{NH_3}\right)$$

For this case, the pore effectiveness factor *E* is a function of the so-called Thiele modulus *m* [237]:

$$E = \frac{3}{m}\left[\frac{1}{\text{tnh}\ m} - \frac{1}{m}\right]$$

The Thiele modulus *m* is defined by

$$m = \frac{d_{eff}}{2}\sqrt{\frac{k_v}{D_{eff}}}$$

where

d_{eff} = effective particle diameter

D_{eff} = effective diffusion coefficient of ammonia in the catalyst particle

k_v = reaction rate constant referred to a unit of particle volume

tnh = hyperbolic tangent

The practical application of kinetic equations to the mathematical calculation of ammonia synthesis converters is described in [218], [238], [242-248].

4.4. Catalysts

The ammonia synthesis catalyst may be viewed as the heart of an ammonia plant. For a given operating pressure and desired production, it determines the operating temperature range, recycle gas flow, and refrigeration requirement. As a result, it directly fixes vessel and exchanger design in the synthesis loop. It also indirectly influences the make-up gas purity requirement, and so the operating pressure, and capital cost, and energy consumption for synthesis gas production and purification. Although the proportionate cost of catalysts compared to the total cost of a modern ammonia synthesis plant is negligible, the economics of the total

process are determined considerably by the performance of the ammonia catalyst [249].

Industrial catalysts for ammonia synthesis must satisfy the following requirements:

1. High *catalyst activity* at the lowest possible reaction temperatures in order to take advantage of the favorable thermodynamic equilibrium situation at low temperatures. Average commercial catalysts yield about 25 vol % ammonia when operating at 40 MPa (400 bar) and 480 °C catalyst end temperature, which corresponds to a 535 °C equilibrium temperature. With catalysts that would function at a reaction temperature about 100 K lower, a yield of 45 vol % ammonia can be expected with the same approach to equilibrium, or the pressure may be reduced to 15 MPa (compare Tables 9 and 10).
2. The highest possible *insensitivity to oxygen- and chlorine-containing catalyst poisons*, which may be present in even the very effectively purified synthesis gas of a modern process (see Sections Catalyst Poisons, and Gas Purification). In assessing the newly developed catalyst systems recommended for operation at very low temperatures (see Section Other Catalysts), it must be kept in mind that the effect of poisons, for example, oxygen compounds, may become more severe as temperature declines (see Fig. 23).
3. *Long life*, which is determined essentially by resistance to thermal degradation and to irreversible poisoning (see Section Catalyst Reduction). In older high-pressure plants (60 – 100 MPa), catalyst life was a big issue; because the catalysts in these plants showed a markedly reduced life owing to the severe operating conditions, the necessary downtime for removing, replacing, and reducing the catalyst had a considerable effect on the ammonia manufacturing cost. In modern single-train ammonia plants, conventional iron catalysts achieve service lifetimes up to 14 years.
4. *Mechanical Strength*. Insufficient pressure and abrasion resistance may lead to an excessive increase in converter pressure drop, and so to a premature plant shutdown. For example, mechanical disintegration during operation along with oxygen sensitivity thwarted the industrial application of uranium carbide catalysts [250].
5. Because of the high and increasing world demand for ammonia, a *reliable primary raw material source*. For example, osmium, which was planned as the first industrial catalyst, is so scarce that, in 1910, as a precautionary measure for this option, BASF had secured almost the total world supply [250].

The ammonia synthesis catalyst problem has been more intensively studied than the catalysis of any other industrial reaction. At BASF A. MITTASCH et al. started a tremendous program, in which up to 1911 more than 2500 different formulations were testet in more than 6500 runs. The experiments were finally brought to an end in 1922 after a total of 22 000 tests. They tested almost all elements of the periodic table for their suitability as ammonia catalysts [82], [251]. From these experiments came a series of technical findings, for example, concerning the relationships between catalytic effectiveness and the strength of the nitrogen bond and reducibility, or relating to the mechanism of opposing activation or inactivation in doubly promoted systems.

In principle, metals or metal alloys are suitable as ammonia catalysts, above all those from the transition-metal group [252] (Table 14). Metals or metal compounds for which the chemisorption energy of nitrogen is neither too high nor too low show the greatest effectiveness (Figs. 12, 13), [253], [254], so the magnetite-based catalyst proved suitable for industrial use.

Table 14. Effectiveness of various elements as catalysts, promoters, or catalyst poisons

	Catalysts	Promoters	Poisons
I		Li, Na, K, Rb, Cs	
II		Be, Mg, Ca, Ba, Sr	Cd, Zn
III	Ce and rare earths	Al, Y, La, Ce and rare earths	B, Tl
IV	(Ti), (Zr)	Si, Ti, Zr, Th	Sn, Pb, C
V	(V)	Nb, Ta	P, As, Bi
VI	(Cr), Mo, W, U	Cr, Mo, W, U	O, S, Se, Te
VII	(Mn), Re		F, Cl, Br, I
VIII	Fe, Ni, Co, Ru, (Rh), Os, (Ir)		

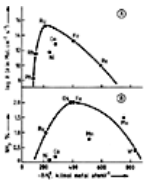

Figure 12. The rate constants of ammonia decomposition (A) on and the ammonia synthesis capacities (B) of metals as a function of $-\Delta H_0^0$. (Mol. denotes molecule, mol denotes mole)

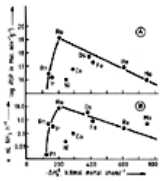

Figure 13. Catalytic activity of carbon-supported metals promoted by metallic potassium as a function of $-\Delta H_0^0$.

A) The rate of isotopic equilibration of N_2 at 623 K, 20 kPa of pN_2 (Mol. denotes molecule, mol denotes mole);

B) The rate of ammonia synthesis at 523 K, 80 kPa of total pressure

4.4.1. *Classical Iron Catalysts*

From the early days of ammonia production to the present, the only catalysts that have been used have been iron catalysts promoted with nonreducible oxides. Recently, a ruthenium-based catalyst promoted with rubidium has found industrial application. The basic composition of iron catalysts is still very similar to that of the first catalyst developed by BASF.

The catalytic activity of iron was already known well before the advent of industrial ammonia synthesis. RAMSAY and YOUNG used metallic iron for decomposing ammonia. PERMAN [255], as well as HABER and OØRDT [256], conducted the first catalytic synthesis experiments with iron at atmospheric pressure. NERNST [257] used elevated pressures of 5 – 7 MPa. Pure iron showed noticeable initial activity which, however, could be maintained for longer operating periods only with extremely pure synthesis gas.

The ammonia synthesis catalyst problem could be considered solved when the catalytic effectiveness of iron in conversion and its onstream life were successfully and substantially improved by adding reduction-resistant metal oxides [251] (Table 15). The iron catalysts promoted with aluminum and potassium oxides proved to be most serviceable [258]. Later, calcium was added as the third activator. Development work in the United States from 1922 can be found in [259].

Table 15. Effect of various elements or their oxides on the activity of iron catalysts in ammonia synthesis

a) positive:
Al, Ba, Be, Ca, Ce, Cr, Er, K, La, Li, Mg, Mn, Mo, Na, Nb, Nd, Rb, Sm, Sr, Ta, Th, Ti, U, V, W, Y, Zr

b) negative:
As, B, Bi, Br, C, Cd, Cl, F, J, P, Pb, S, Sb, Sn, Te, Tl, Zn

c) doubtful:
Au, Co, Cu, Hg, Ir, Ni, Os, Pd, Pt, Si

Modern catalysts additionally contain other promoters that were present in the older catalysts only as natural impurities from the raw materials. Onstream life and performance were enhanced considerably by optimizing the component ratios (Section Composition), conditions of preparation (Section Catalyst-Precursor Manufacture), and catalyst particle size and form (Section Particle Size and Shape). The high-purity gas of modern processes and the trend to lower synthesis pressures especially favor the development of more active and easily reducible types of catalysts, at some sacrifice in temperature stability and resistance to poisons. To some extent even today, ammonia plant operating conditions and types of converters (Section Formation of Ammonia in the Converter) can differ greatly one from another. Thus, individual catalyst manufacturers now offer several catalyst types in various particle size distributions, in oxidic and prereduced states.

4.4.1.1. *Composition*

Table 16 gives a composition survey of commercial ammonia catalysts in the years 1964 – 1966. The principal component of oxidic catalysts is more or less stoichiometric magnetite, Fe_3O_4, which transforms after reduction into the catalytically active form of α-iron.

Table 16. Examples of commercial ammonia catalysts from the years 1964 – 1966. Values for the composition in weight %. The numbers at the beginning of the lines are keyed to the catalyst origin.

Origin, type	Fe total	FeO	Fe_2O_3	Al_2O_3	MgO	SiO_2	CaO	K_2O	Other	Particle size, mm	Bulk density, kg/L
1	68.6	36.07	57.85	3.30	0.09	0.75	2.13	1.13	–	2 – 4	2.37
2 normal	60.0	32.91	60.18	2.90	0.37	0.35	2.80	0.54	–	4 – 10	2.94
2	68.2	31.30	62.53	2.90	0.30	0.35	1.65	0.97	–	6 – 10	2.80
2 HT	66.9	32.47	59.18	2.95	1.55	0.40	2.95	0.50	–	6 – 10	2.80
2 prereduced	88.1			3.70	0.43	0.45	3.60	0.70	–	6 – 10	2.30
3	71.3	39.22	58.2	1.80	0.18	0.27	1.43	0.89	–	2 – 4	2.86
4	66.3	22.27	49.0	0.59	4.47	0.77	0.65	0.50	0.7 Cr_2O_3		
4 prereduced	90.6			0.10	6.08	1.23	0.10	0.86	1.05 Cr_2O_3		
5	71.5	33.0	65.5	2.96	1.55	–	0.20	0.01	–		
6 (1964)	69.5	23.85		3.15	0.26	0.40	1.85	1.10	–	3 – 9	2.71
6 (1966)	66.9			2.73	0.29	0.43	1.84	1.15	–	5 – 10	2.73
7 prereduced	90.4			3.12	1.00	0.46	0.25	0.58	0.4 MnO	5/5	2.55
8	68.4	35.35		3.16	0.56	0.50	3.54	0.58	–	2 – 4	2.61
9	70.0	32.14		3.17	0.28	0.10	2.40	0.32	–	2 – 4	2.81
10	70.8	33.62		1.58	0.28	1.14	0.67	1.57	–	2 – 4	2.66
11 normal (1964)	66.7	35.95	56.97	3.27	0.67	0.55	3.00	0.65	–		
11 (1966)	68.2	38.70	54.60	2.42	0.35	0.64	2.85	0.58	–		
11 normal (1964)				3.0	0.3	0.5	2.0	1.0	–		
11 (1966)	69.5	38.20	56.70	2.34	0.35	0.57	1.85	0.57	–		
11 HT	66.3	38.22	52.38	2.94	3.56	0.30	2.66	0.63	–		
11 prereduced	84.9			3.62	0.43	0.94	4.70	0.68	–	5 – 11	2.11
12				3.9	4.0	0.8	2.3	1.8	–		
13		23.15		2.9	0.1	0.42	3.12	0.52	–		
14				4		0.7	3	1			

The degree of oxidation of industrial catalysts has a considerable influence on their catalytic properties. MITTASCH in 1909 established that catalysts manufactured by reducing a magnetite phase were superior to those prepared from other oxides. For industrial catalysts, the highest ammonia yields are observed with an Fe(II) – Fe (III) ratio of 0.5 – 0.6, about the degree of oxidation of stoichiometrically composed magnetite [260-262] (Fig. 14). To obtain optimal catalyst composition, careful control of the manufacturing process, especially the melting conditions, which determine the oxygen content, is necessary.

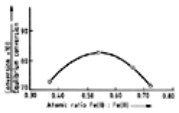

Figure 14. Dependence of the ammonia yield on the degree of oxidation of the iron in the unreduced catalyst

In general, the catalysts contain varying quantities of the oxides of aluminum, potassium, calcium, magnesium, and silicon as promoters. Patents recommend adding sodium [263], beryllium [264], vanadium [265], uranium [307], or platinum [266]. Reference [267] describes cesium-containing catalysts. Catalysts patented by Lummus [268] and Ammonia Casale [269] contain cerium as additional promoter. ICI [270] has developed a cobalt-containing catalyst, as has Grande Paroisse [271].

Nature of the Surface of Commercial Iron Catalysts. Freshly reduced commercial iron catalysts that contain aluminum, potassium, and calcium oxides as basic promoters consist of approximately 30-nm primary crystallites; the spaces between them form an interconnecting system of pores. Besides a maximum at a pore radius of about 10 nm that originates on reduction of the Fe_3O_4 (magnetite) phase of the nonporous oxidic catalyst, the pore distribution curve (Fig. 15) generally shows a peak at 25 – 50 nm that is formed on reduction of the wustite phase [126], [130]. The pore volume measures 0.09 – 0.1 cm^3/g; with an apparent density of 4.8 – 4.9 g/cm^3, accordingly, the pores represent 44 – 46 % of the volume of a catalyst granule [14]. The surface of the walls of the pores, the so-called inner surface, amounts to about 15 m^2/g.

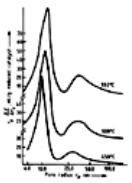

Figure 15. Pore size distribution of a commercial catalyst after reduction at various temperatures [130]

The novel AmoMax catalyst [1321], [1322] of Südchemie is iron-based but uses wustite instead of magnetite—which previously was considered undispensible for the production of industrial ammonia catalysts—has an improved pore structure and higher specific surface area.

The composition of the outermost atomic layers of the pore walls deviates considerably from the overall average concentrations. Auger electron spectroscopic (AES) measurements on an industrial catalyst (BASF S 6-10) have shown that a significant enrichment of the promoters into the surface results using the unreduced as well

as the reduced catalyst [119] (see Table 17). The free iron surface of the reduced BASF catalyst [119] and Topsøe catalyst KM-I [131] comprises only a fraction of the total surface, as could be deduced from the results of prior investigations [126], [132-138].

Table 17. Composition by volume of an industrial ammonia catalyst in comparison to the surface composition before and after reduction (an approximately 10^{-4} cm^2 size typical surface). Numerical values in atomic % [119]

	Fe	K	Al	Ca	O
Volume composition	40.5	0.35	2.0	1.7	53.2
Surface composition before reduction	8.6	36.1	10.7	4.7	40.0
Surface composition after reduction	11.0	27.0	17.0	4.0	41.0

The aluminum oxide promoter exists partly in the form of larger crystallites and, moreover, is relatively homogeneously distributed over the iron area of the surface, although with low concentration [119], [131]. After reduction, about 1 wt % of the alumina also remains statistically distributed in the form of $FeAl_2O_4$ molecular groups built into the α-iron lattice of the reduced catalyst [139], [140]. According to [119] the potassium, in the form of a K + O adsorbed layer, covers about 20 – 50 % of the iron surface. According to [130], [131], a correlation exists between the distribution of the potassium and that of aluminum and/or silicon. Calcium oxide segregates, essentially at the grain boundaries, into separate regions, probably as a mixture of the silicate and ferrite [121]. Auger spectroscopic investigations on reduced BASF and Topsøe catalysts reveal large local differences in composition [119], [130]. Large, apparently homogeneous regions that have originated from reduction of Fe_3O_4 crystallites alternate with nonhomogeneous regions that are formed by the reduction of FeO crystals or consist of amorphous phases [130].

Extensive studies in the last decade have provided a more refined picture of the morphology of the active catalyst (reduced state) and its precursor (oxidic state). A review is given in [124]. With methods such as scanning transmission electron microscopy (TEM) and electron microdiffraction a textural hierarchy has been modeled. Macroscopic particles in the reduced catalyst are confined by fracture lines running through a system of blocks consisting of stacks of slabs in parallel orientation. This structure is already preformed in the preparation of the catalyst precursor, and in the reduction process a further subdivision of the slabs into even smaller platelets might occur. This texture is stabilized by structural promoters, which act as spacers and “glue,” separating neighboring platelets and thus providing voids for the interconnection of the pore system. There is also evidence that the basal plane of many platelets has the Fe(111) orientation [122], [124].

Influence of the Promoters. Promoters can be arranged in different groups according to the specific action of the metal oxides:

Structural stabilizers, such as Al_2O_3, produce a high inner surface during reduction and stabilize it under thermal stress by restraining iron crystallite growth [158], [262], [273]. The ability of the various metal oxides to create a high specific surface decreases in the following order [274]:

$$Al_2O_3 > TiO_2 > Cr_2O_3 > MgO > MnO = CaO > SiO_2 > BeO$$

So-called *electronic promoters*, such as the alkali oxides, enhance the specific activity (based on a unit surface) of iron – alumina catalysts. However, they reduce the inner surface or lower the temperature stability and the resistance to oxygen-containing catalyst poisons [275], [276]. In the alkali-metal series, the promoter effect increases with increasing atomic radius, and the destructive effect with decreasing atomic radius [277]. In striving to improve the activity or stability of iron catalysts, a multitude of structural and electronic promoters has been investigated, among them rare-earth oxides [278], [279], such as Sm_2O_3 [280], Ho_2O_3, Dy_2O_3, and Er_2O_3 [281].

Promoter oxides that are reduced to the metal during the activation process and form an alloy with the iron (see also Section Other Catalysts) are a special group. Among those in use industrially, cobalt is of special interest [282], [283].

The effect of a given promoter depends on concentration and on the type of promoter combination and the operating conditions, especially the reaction temperature and the synthesis gas purity [265], [275], [277], [284-287].

A graphic picture is conveyed in [277] of how the activity of a quadruply promoted (4 % Al_2O_3 , 1 % K_2O, 1 % CaO, 1 % SiO_2) catalyst changes with varying promoter concentration and operating conditions (Fig. 16). Under normal operating conditions [14 – 45 MPa, 380 – 550 °C, 10 000 – 20 000 m^3 m^{-3} h^{-1} (STP)], the optimal activity corresponds to a composition of 2.5 – 3.5 % CaO, 2.3 – 5.0 % Al_2O_3, 0.8 – 1.2 % K_2O, and 0 – 1.2 % SiO_2 (Fig. 16 B). Raising or lowering the concentration of a particular oxide causes a reduction in activity. Changes in the potassium and aluminum oxide concentrations have an especially strong influence. Adding magnesium oxide decreases the catalyst performance. If before the test the catalyst is overheated at 700 °C for 72 h (Fig. 16 C) or poisoned with 2000 cm^3 water vapor per cubic meter of gas at 550 °C (Fig. 16 D), the optimum composition shifts to higher Al_2O_3 and SiO_2 and lower K_2O and CaO concentrations. Magnesium oxide addition now shows a favorable effect.

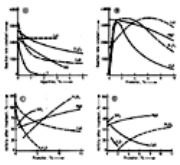

Figure 16. Dependence of the catalyst activity on various factors

A) Concentration of impurities; B) Concentration of promoters; C) Overheating to 700 °C with increasing promoter concentrations; D) Poisoning with water at increasing promoter concentrations

For ammonia plants operating at pressures up to 35 MPa (350 bar), catalyst end temperatures of 520 – 530 °C maximum, and with highly purified synthesis gas, the preferred catalysts contain 2.5 – 4.0 % Al_2O_3 , 0.5 – 1.2 % K_2O, 2.0 – 3.5 % CaO, 0 – 1.0 % MgO, and a natural content of about 0.2 – 0.5 % SiO_2 [262]. Less active but more poison- and temperature-resistant catalysts containing up to 3.6 % magnesium oxide were recommended for older plants, for example, classical Casale plants, which operated at up to 80 MPa (800 bar) pressure and in which catalyst end temperature reached 650 °C (see for example [14]). An industrial catalyst for operating temperatures up to 550 °C is stabilized against deterioration by 2 – 5 % V_2O_5 besides 3.5 – 4.0 % Al_2O_3, 2.0 – 2.5 % CaO, and 0.7 % K_2O [265]. For higher operating temperatures, still higher V_2O_5 contents are recommended. Silicon dioxide additions shift the optimum potassium oxide concentration to higher values [275]. For example, the Bulgarian catalyst K-31 contains 3.9 % Al_2O_3, 2.3 % CaO, 0.4 % MgO, 0.8 % SiO_2, and 1.8 % K_2O. An older Norsk Hydro catalyst, besides the usual additives, contained 1.14 % SiO_2 and 1.57 % K_2O. The ICI catalyst with composition 5.2 % CoO, 1.9 % CaO, 0.8 % K_2O, 2.5 % Al_2O_3 , 0.2 % MgO, 0.5 % SiO_2, remainder Fe_3O_4, is substantially more active than the conventional cobalt-free catalysts [270], [288]. Reducing the synthesis pressure and/or the synthesis temperature should enable application of the Lummus [268], [289] and Ammonia Casale [269] cerium-containing catalysts.

The effect of the promoters on the rate of reduction and the temperature required for reducing the iron oxide phase is also significant in industrial practice. The structural promoters, such as Al_2O_3, lower the rate of reduction [124], [290], [291]. Magnesium oxide-activated iron, said to be thermally stable up to 650 °C, needs a higher reduction temperature than aluminum oxide-promoted iron catalyst [292]. Greater differences in reducibility have also been observed in commercial catalysts with similar chemical composition [293] and in connection with particular oxide phases, such as $CaFe_3O_5$ and FeO [293-295]. Early work [296] referred to the rough parallels between the reducibility and the thermal stability of catalysts. All published experience appears to demonstrate that it is not possible to combine in a catalyst high thermal stability with easy reducibility and high activity at low temperatures. Hence it may be advantageous to use a combination of active and thermally-resistant catalysts in the same converter.

Mechanism of the Promoter Effect. The action of the so-called structural promoters (stabilizers), such as Al_2O_3, is closely associated with their solubilities in the iron oxide matrix of the unreduced catalyst or with the capability of the regular crystallizing magnetite to form solid solutions with iron – aluminum spinels [14], [297-299].

The solid solutions of Fe_3O_4 and the spinel $FeAl_2O_4$ have a miscibility gap below 850 °C [300]; at 500 °C, the solubility limit in the magnetite mixed crystal is a maximum of 7.5 % $FeAl_2O_4$, i.e., 3 % Al_2O_3, referred to iron. Higher alumina contents lead to separation into two phases, whereby only the portion dissolved in the magnetite phase appears to be responsible for the specific promoter action of alumina [132], [133], [301]. According to [302], there exists a close connection with the mechanism of the reduction which consists schematically of the following partial steps [303]:

1. Phase-boundary reaction of hydrogen with oxygen ions of the magnetite lattice, with formation of water vapor and release of electrons
2. Formation of metallic iron nuclei by combination the electrons with Fe^{2+} ions
3. Diffusion of Fe^{2+} ions and electrons to the nuclei and growth of the nuclei to iron crystals of various size

A more detailed discussion of the reduction mechanism is given in Section Catalyst Reduction.

In the presence of dissolved aluminum ions, at not too high a temperature, the diffusion rate of the iron ions in the magnetite lattice is low. Hence, nucleus formation proceeds rapidly relative to crystal growth. Therefore, small iron crystallites, about 30 nm, form with correspondingly large specific surfaces. The aluminum probably remains partly in the iron crystallite in the form of very small $FeAl_2O_4$ areas statistically distributed over the lattice [139], [140], [302], [305], where an $FeAl_2O_4$ molecule occupies seven α Fe lattice positions [112].

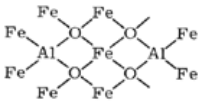

According to this concept, the stabilizer function of alumina reduces to paracrystalline lattice defects; an analogous effect is to be expected with Cr_2O_3, Sc_2O_3 , etc. [139], [305]. Another theory is based on the observation that during reduction part of the alumina precipitates with other promoters into the surface of the iron crystallite in a molecularly dispersed distribution [135], [306] or in small islands [131]. This “patchy” monolayer of alumina acts like a “spacer” between iron atoms of neighboring crystallites and prevents sintering by means of a “skin” effect [140]. (See also [112], [121], [130].)

Insofar as small crystals of nonreducible oxides dispersed on the internal interfaces of the basic structural units (platelets) will stabilize the active catalyst surface Fe (111), the paracrystallinity hypothesis will probably hold true. But the assumption that this will happen on a molecular level on each basic structural unit is not true. The unique texture and anisotropy of the ammonia catalyst is a thermodynamically metastable state. Impurity stabilization (structural promotion) kinetically prevents the transformation of platelet iron into isotropic crystals by Ostwald ripening [122]. Thus the primary function of alumina is to prevent sintering by acting as a spacer, and in part it may also contribute to stabilizing the Fe(111) faces [125], [165].

Calcium oxide, which also acts as a structural promoter [130] has a limited solubility in magnetite. It tends to stretch the magnetite lattice [112]. In the main, on cooling the magnetite melt, it separates at the grain boundaries as $CaFe_3O_5$ (at very rapid cooling rates) [293] and, in the presence of SiO_2, forms poorly reducible intermediate layers of calcium ferrite and silicates [130], [294]. In the reduced catalysts, it segregates between the iron crystallites [120], [137] and so possibly prevents sintering together at high operating temperatures. One may also presume that by partial neutralization of the “acid” components by calcium, more potassium is made available for activating the iron [218].

Potassium is likewise scarcely soluble in magnetite because of its ionic size [112]. In the unreduced catalysts, separate potassium- and iron-rich regions were found [119]. The appearance of a $K_2Fe_{22}O_{34}$ phase besides an unidentified phase, however, has been proved [308]. According to [309], during reduction, the emerging K_2O migrates to the iron crystallite surface. While doing this it reacts with the more or less homogeneously distributed aluminum (silicon) compounds. In this way, it is distributed over the iron phase. In the reduced catalysts, the potassium exists as a K + O adsorption layer that covers about 20 – 50 % of the iron surface [119]. According to [136], [310], potassium associates partly with the alumina in the surface, partly as KOH with iron [1282]. It was found that the enhancement of the catalysts' specific activity by potassium oxide is accompanied by a decrease in the electron work function [310-312]. The promoting effect of potassium seems to be based on two factors which probably act simultaneously. One mechanism is the lowering of the activation energy for the dissociative adsorption of nitrogen [111], [119]. The explanation is based on an electrostatic model [115], [156], [313]. As indicated by the strong decrease in the work function upon potassium adsorption, there is a considerable electronic charge transfer to the substrate, which creates a $M^{\delta-} – K^{\delta+}$ dipole. A nitrogen molecule adsorbed near such a site will experience a more pronounced back-bonding effect from the metal to its antibonding π orbitals. This will increase the bond strength to the metal and further weaken the N – N bond, as can be seen from a further reduction of the N – N stretching frequency [314]. The other effect consists of lowering the adsorption energy of ammonia, which avoids hindering of nitrogen adsorption by blocking (poisoning) of the catalyst surface by adsorbed ammonia molecules [109, p. 234], [119]; hence, potassium oxide ought to improve the catalyst performance less at low than at higher operating pressures [210 p. 316], [315].

The negative effect of K_2O concentrations higher than about 0.58 % has not been explained unequivocally [158]. With increasing potassium concentration, this manifests itself by the increasing size of the average iron crystallite or the decreasing specific surface in the reduced catalyst [275], [308]. Since potassium oxide prevents the formation of solid solutions between alumina and magnetite to a certain extent [316], the recrystallization-promoting effect of higher K_2O concentrations may be attributed to a lowering of the portion of alumina dissolved in the magnetite phase [308]. Remarks in [309] and [317] reveal another possible interpretation: the K_2O located at the phase boundary surface and not bound to acid or amphoteric oxides may be converted by water vapor concentrations over 10^{-2} ppm in the synthesis gas into potassium hydroxide or by hydrogen into potassium and potassium hydroxide, which would exist in molten form at operating conditions [309], [1282].

An extensive review on promoters can be found in [116].

4.4.1.2. *Particle Size and Shape*

The choice of particle size and shape of commercial ammonia catalysts is determined mainly by two factors:

1. Catalyst performance
2. Pressure drop

From the standpoint of space – time yield, it is desirable to use the finest possible particle, which, practically speaking, is about 1 – 2 mm (Fig. 17); however, with decreasing particle size, the pressure drop and the risk of destructive fluidization of the catalyst increase (Fig. 18).

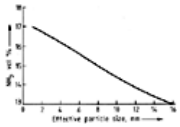

Figure 17. Influence of the particle size on the ammonia production (BASF catalyst).

25 MPa; SV = 12 000 $m^3 m^{-3} h^{-1}$ (STP); gas composition (vol %): N_2 20.59, H_2 60.06, NH_3 2.55, Ar 5, CH_4 10

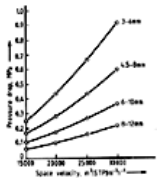

Figure 18. Pressure drop in the catalyst bed for various catalyst particle size ranges, from [14].

Depth of the catalyst bed, 7 m; reaction pressure 27.1 MPa; reaction temperature 450 °C

For processes operating at pressures of 25 – 45 MPa (250 – 450 bar) and at space velocities of 8000 – 20 000 $m^3 m^{-3} h^{-1}$ (STP) a grain size of 6 – 10 mm is

preferred. Larger granulations, for example, 8 – 15 mm or 14 – 20 mm, are used only in plants where the lowest possible pressure drop is essential because of very high gas velocities. In catalyst zones in which the ammonia formation rate is so high that the allowable temperature limits are exceeded, it may be advantageous to use coarse particles for suppressing the reaction. Radial-flow converters and the horizontal crossflow Kellog converter (Fig. 52), which operate at comparatively low gas velocities [318], allow the use of small granulations (1.5 – 3 or 2 – 4 mm) with optimal use of the converter volume. Fluidized-bed processes, which were explored especially in the Soviet Union, have so far been unsuccessful [319], [320].

Two effects cause the low production capacity of coarse-grained catalyst: first, large grain size retards transport of the ammonia from the particle interior into the bulk gas stream, because this proceeds only by slow diffusion through the pore system. Slow ammonia diffusion inhibits the rate of reaction. At the high reaction rate typical for the converter inlet layer, only a surface layer of the catalyst grains, about 1 – 2 mm thick, participates in the reaction.

The second effect is a consequence of the fact that a single catalyst grain in the oxidic state is reduced from the outside to the interior of the particle [321]: the water vapor produced in the grain interior by reduction comes into contact with already reduced catalyst on its way to the particle outer surface; this induces a severe recrystallization [14]. The effect is very significant. As an example, if the particle size increases from about 1 to 8 mm, the inner surface decreases from 11 – 16 m^2/g to 3 – 8 m^2/g [326].

To allow for the influence of various particle shapes and size distributions within a defined sieve fraction, in lay-out calculations it is customary to employ an effective particle diameter, d_{eff}, as nominal size. The diameter d_{eff} is defined as the ratio of equivalent diameter Λ and a form factor Ψ. Λ is equal to the diameter of a sphere with a volume equal to the (average) volume of the particles, and Ψ is the average ratio of the particle surface to the surface of a sphere of equal volume.

Table 18 shows the relation between the catalyst size classification, the equivalent particle diameter, and the percentage saving in catalyst or converter volume relative to the 6 – 10 mm standard size [288].

Table 18. Effect of catalyst size on catalyst volume

Catalyst size classification, mm	Approximate equivalent particle diameter, mm	Relative catalyst volume, %
6 – 10	7.5	100
4.5 – 8	5.5 – 6.5	92 – 95
3 – 6	4.5 – 4.7	88 – 90
1.5 – 3	2.0 – 2.2	80 – 82
1 – 1.5	1.2 – 1.3	77 – 79

An irregular grain shape, for example with a shape factor of 1.5, has a more favorable effective activity for the individual particle and for radial intermixing of mass and heat in an industrial converter [327] than a more cubic or spherical shape, with a shape factor close to one. According to a patent by Chemie Linz AG [328], the catalyst particle ought to be 2 – 20 times as long as it is broad, preferably 5 – 10 mm long and 1 – 2 mm thick (broad). On the other hand, regular shapes have the advantages of greater abrasion resistance and lower pressure drop (see Fig. 19) [401].

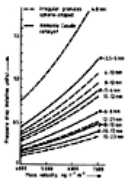

Figure 19. Comparison of the Ammonia Casale spherical catalyst and irregularly shaped catalyst [288]

The advantages of regular catalyst shapes and the need to compensate for the above-described negative effects of larger grain size by a system of macropores in the oxidic and reduced catalyst stimulated various attempts to manufacture shaped, macroporous catalysts. Various manufacturing techniques have been proposed [329-338]. As an example, magnetite is melted with the additives at high temperature (> 1600 °C) and the melt is cooled, broken, and ground to powder. After water is added and, if required, a binding agent, such as bentonite [337], or a promoter salt, such as cerium nitrate [269], the powder is pelletized. The pellets subsequently are dried and sintered in an inert atmosphere at higher temperatures (about 1350 °C).

The application of macroporous catalysts ought to be especially useful for very low synthesis pressures and in plants in which large catalyst particles must be used for reasons of low pressure drop. For example, the performance of the macroporous Topsøe catalyst KMG 6 mm at 5 MPa synthesis pressure is said to be at least equivalent to KM I 1.5 – 3 mm and distinctly superior to KM I 6 – 10 mm [339].

4.4.1.3. *Catalyst-Precursor Manufacture*

The term “ammonia catalyst” commonly refers to the oxidic form consisting of magnetite and oxidic promoters. In fact this is only the catalyst precursor which is transformed into the active catalyst composed of α-iron and promoters by reduction with synthesis gas, usually in situ. The reduction step is very important for catalyst performance.

The first effective catalysts were made by the oxygen-melt process. The purest possible iron (e.g., Swedish charcoal iron), together with the additives, was burned to Fe_3O_4 in a stream of oxygen. This process was largely replaced by melt processes in which natural or, less frequently, synthetic magnetite, together with the activators, was melted electrically or in electric arc furnaces [261], [340]. The cooled melt is ground to the proper granulation and reduced with hydrogen – nitrogen mixtures.

A process developed by Farbenfabriken (formerly Friedrich Bayer) had only local significance. Complex iron cyanides were decomposed thermally in the presence of hydrogen. The hydrodecomposition proceeds via the carbide and nitride, finally to α iron. This is activated in the well-known manner by reduction-resistant metal oxides. The so-called Mont Cenis process employed such catalysts [341]. Repeatedly described as a means of manufacture, although only in the scientific literature, is coprecipitation of the catalyst components, for example, from aqueous solutions of the metal salts, with subsequent calcining and reduction [342-345]. With magnesium oxide as support, very small (under 10 nm) iron particles are obtained with high specific iron surfaces [345], similar to those obtained by exchange of magnesium ions by iron ions in the surfaces of magnesium hydroxy carbonate crystals [345], [346]. Granulation and sintering techniques have been used for the preparation of shaped macroporous iron catalysts, which, however, have not gained industrial importance.

Impregnating the pore surface of prereduced passivated catalysts is a possibility for incorporating promoters into iron ammonia catalysts. A United Kingdom patent [347], by way of example, claims catalysts manufactured by impregnating the reduced catalysts with cerium salts. Improving the performance and life of industrial catalysts by radioactivity and X-rays [348], treatment with ultrasound [349], [350], mechanical treatment [351], or high-frequency, alternating-field heating [352] has been attempted also.

The superiority of the catalyst manufacturing processes that use a molten iron oxide stage is mainly due to the fact that above 1000 °C in air, magnetite, Fe_3O_4, is the thermodynamically stable oxide phase of iron [82], [353]. Magnetite leads to especially efficient catalysts, and its electrical conductivity allows the use of economical electrical melting processes.

In 1996 the prices of commercial ammonia catalysts were about 2 $/lb (3.58 €/kg) for oxidic and about 5.5 $/lb (10.23 €/kg) for prereduced. Therefore, they are among the least expensive catalysts.

Operating conditions in the individual manufacturing steps — proportioning and mixing the raw materials, melting, cooling, crushing or if necessary grinding and preforming, and reduction — influence the quality of the finished catalyst (Fig. 20) [249].

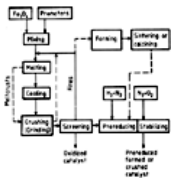

Figure 20. Ammonia catalyst manufacture

The raw materials — usually, natural magnetite, lime, potash, and alumina — must, as far as possible, be free of catalyst poisons (see Section Catalyst Poisons). Many ores have too high a content of free or bound silica, which can be lowered with magnetic separators [354]. Melting is accomplished in electrical resistance or induction furnaces (arc furnaces in the past) operating at 1600 – 2000 °C. The walls of these furnaces should consist of a weakly basic tamping material indifferent to the melt, such as magnesium oxide, which is also an activating component of the catalyst [355]. A homogeneous distribution of the promoters in the magnetite melt and a degree of oxidation at or under that of stoichiometric magnetite should be obtained. This is said to be promoted by initially overheating to temperatures up to 3500 °C [356] (Fig. 21). It is claimed to be advantageous to bring the promoters into the melt as common chemical compounds that are isomorphous with magnetite [357].

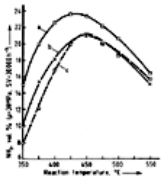

Figure 21. Effect of the melting temperature and rate of cooling of the melt on the activity of ammonia catalysts

a) Melt overheated to 3500 °C (rapid cooling); b) Melt temperature 1800 – 2000 °C (rapid cooling); c) Melt temperature 1800 – 2000 °C (slow cooling)

Induction furnaces are optimal for the melting operation. Their good temperature control permits accurate adjustment of the degree of oxidation. Since the melt is held in constant turbulent motion by the magnetic field produced in the primary coil, it is well mixed, even for short melt times. In comparison to the most frequently used resistance furnaces, plant cost and power consumption are higher.

Another important factor in catalyst manufacture is the melt cooling rate, which is affected by the design and dimensions of the ingot molds. Quenching or fast cooling in thin sheets leads to a less abrasion-resistant, sharp-edged chip after crushing. Very slow cooling results in a more cubic chip, but with inferior catalyst quality (Fig. 21). In practice, slow cooling is avoided.

With falling temperature, both the solubility of the activator oxides (see Mechanism of the Promoter Effect.) in magnetite and the rate of adjustment to the new phase equilibrium decline. Therefore, by rapid cooling of the melt the activator oxide distribution can be frozen in a condition corresponding to that of a higher temperature [358]. According to [293], there may exist relationships between the melt cooling rate, the appearance of certain phases, and the reducibility of the catalysts.

4.4.1.4. *Catalyst Reduction*

The reduction of oxidic catalyst is generally effected with synthesis gas. The magnetite is converted into a highly porous, high surface area, highly catalytically active form of α-iron. The promoters, with the exception of cobalt, are not reduced [14].

To ensure maximum effectiveness of the catalyst, a defined reduction procedure must be followed. Above all, it is important to hold the partial pressure of the resulting water vapor as low as possible and to insure that the water vapor does not come into contact with regions that have already been reduced. High temperature and high water vapor partial pressure markedly accelerate premature catalyst aging by recrystallization. Therefore, the reduction should be carried out at high gas velocities [about 5000 – 15 000 $m^3\ m^{-3}\ h^{-1}$ (STP)], at the lowest temperatures sufficient for complete reduction, and at not too high pressures (7 – 12 MPa in low-pressure, 25 – 30 MPa in high-pressure plants) to hold the exothermic formation of ammonia under better control during the reduction. When the reduction of the oxidic catalyst is carried out in the production plants, long reduction times are needed at low temperatures and low pressures with a consequential loss of production.

In practice, the reduction temperature is raised stepwise by using the exothermic heat of ammonia formation. The progress of the reduction is controlled according to the catalyst temperature and the water concentration by means of the synthesis gas flow. As a rough guideline, the water content of the gas effluent from the catalyst should not exceed 2 – 3 g/m^3 (STP). Under these conditions, depending on its size and operating pressure, a synthesis converter with a fresh load of oxidic catalyst attains its full production capacity in 4 – 10 d.

The minimum temperatures necessary for reduction are somewhat different for the various catalyst types. Catalysts conventionally employed in medium-pressure plants may be reduced from about 340 to 390 °C, although a slow induction period starts somewhat lower. Generally, temperatures above 440 °C are required to complete the reduction.

The reducibility of industrial catalysts is dependent on both the combination of promoters and the degree of oxidation. The FeO (wustite) phase is reduced faster and at lower temperatures than the Fe_3O_4 (magnetite) phase [293]. According to [293], the rather considerable differences in the reduction rates of commercial catalysts with similar compositions may be attributed to differences in manufacturing methods or operating conditions. Commonly, the manufacturers hold these in strict secrecy.

Some older publications deal with the influence of catalyst granulation on the optimum reducing conditions [360]. Directions for reduction in multibed converters that combine fine- and coarse-grained catalyst appear in [361]. The influence of the hydrogen/nitrogen ratio during reduction on the catalyst performance after reduction is discussed in [362].

The influence of the reduction conditions (gas flow rate, temperature in the range 300 – 600 °C, nitrogen content in the hydrogen in the range 0 – 100 %) on the production capacity of an ammonia catalyst has been investigated [368].

The gas/solid reaction between magnetite and hydrogen has been studied in great detail by rate measurements, microscopy, and X-ray diffraction [363-366]; a summary is given in [124]. On the atomic scale the reaction is controlled by two processes:

1. Metallic iron is formed from wustite by direct chemical reaction controlled in the initial phase by the reaction rate (activation energy ca. 65 kJ/mol) and in the final stage by diffusion processes involving hydrogen and water on the reaction site:

   $$FeO + H_2 \rightarrow Fe + H_2O$$

2. The chemical reaction creates an iron(II) ion concentration gradient in the solid. This gradient leads to a rapid diffusion of iron(II) ions from magnetite through wustite to the chemical reaction interface, where they are reduced and precipitated as iron nuclei. This is made possible by the structural defects of the wustite. The precipitation of further wustite nuclei on the magnetite/wustite reaction interface seems to be effected by ion/electron diffusion processes rather than by direct contact of magnetite with hydrogen [367]:

   $$O^{2-} + H_2 \rightarrow H_2O + 2\,e^-$$

   $$Fe_3O_4 + Fe^{2+} + 2\,e^- \rightarrow 4\,FeO$$

The topotactic reduction process leads to a core and shell structure which is visible under the optical microscope and is shown schematically in Figure 22 (a). Figure 22 (b) shows the wustite/magnetite interface as observed by electron microscopy [124]. This concept was developed for single-crystal magnetite grains, but the shell and core model is also valid for commercial polycrystalline catalysts [369], [124]. Information on the reduction kinetics, including industrial catalysts, can be found in [122], [321], [322-325]. Newer findings seem to question the shell and core model with the topotactic reaction interface [172]. On account of the simplifications involved this model should be regarded as a formalistic approach to describe the reaction kinetics rather than a mechanistic approximation. A more detailed atomistic picture leads to the assumption that there is no homogeneous topotactic reaction interface. According to this concept the active catalyst should not grow as a shell around a precursor core but as a core within an oxidic matrix.

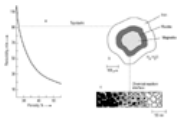

Figure 22. Mechanism of catalyst reduction [117]

a) Reducibility of catalyst under standard conditions as a function of its porosity; b) Core and shell structure of catalyst; c) Reaction interface

Prereduced, stabilized catalyst types, introduced on the market some years ago, have gained a considerable market share. Prereduced catalysts have the full pore structure of active catalysts, although the pore surface has been oxidized to a depth of a few atomic layers to make these catalysts nonpyrophoric.

Reactivating such catalysts usually takes only 30 – 40 h. Ammonia formation begins at substantially lower temperatures, so that altogether the downtime of a production unit is reduced markedly. This and further advantages, such as reducing the risk of damaging the catalyst during activation by too high a local concentration of water, the quantitatively lower yield of aqueous ammonia solution (which can be added to the production), as well as the roughly 20 % lower bulk weight (which reduces the design loadings and costs of the internals of large converter units), make using prereduced catalysts increasingly attractive, especially in single-train plants, in spite of higher prices. The somewhat inferior mechanical strength is a disadvantage that requires special care when a charge is being loaded into the converter. When removed from the hydrogen – nitrogen atmosphere at low to moderate pressure in the producing furnace, the prereduced catalysts are pyrophoric and must be stabilized before transport and installation by passivating the surface. Usually, the method recommended by BURNETT is used. The reduced charge is treated with nitrogen containing 100 – 1000 ppm oxygen at 50 – 70 °C (maximum 95 °C) and a pressure of 0.1 – 0.2 MPa (1 – 2 bar) and up [359]. The reducible oxygen content of the prereduced catalyst ranges between 2 and 7 %. Part of this is only loosely bound and is removed in reactivation even below 200 – 300 °C.

Detailed data on the manufacturing steps most important to the catalyst performance, reduction, prereduction, passivating, and reactivation, appear in [14], including a discussion of the most important literature in this field.

4.4.1.5. *Catalyst Poisons*

The activity of an ammonia synthesis catalyst may be lowered by certain substances, commonly referred to as poisons. These substances can be minor gaseous constituents of the synthesis gas or solids introduced into the catalysts during the manufacturing procedure, derived from impurities in the natural magnetite from which the catalyst is made. These latter should not play a major role with catalysts from manufacturers of repute and are not discussed in detail in this section because of the proprietary nature of the production processes. General measures to avoid this sort of contamination include selecting a rather pure magnetite, the application of pretreatment processes, and the use of high-purity promoters. The melting process itself may also contribute to minimizing the content of some minor impurities. For gaseous poisons in the synthesis gas, a distinction can be made between permanent poisons that cause irreversible damage to the catalyst and temporary poisons which lower the activity while present in the synthesis gas. In contrast to temporary poisons, permanent poisons can be detected by chemical analysis. Oxygen-containing compounds such as H_2O, CO, CO_2, and O_2 are the most common temporary poisons encountered in ammonia synthesis.

Oxygen compounds have a reversible effect on iron catalysts at not too high temperatures. That is, the activity of a damaged catalyst may be practically completely restored by reduction with clean synthesis gas. Equivalent concentrations of oxygen compounds, for example, 100 ppm of O_2 or CO_2 and 200 ppm CO or H_2O, lead to the same degree of poisoning, presumably because as soon as they enter the catalyst bed they rapidly and completely transform into H_2O [14]. The damage depends approximately linearly on the quantity of adsorbed water taken up by the catalyst [370], which is proportional to $\sqrt{pH_2O}/\sqrt{pH_2}$ [371]. Corresponding to the adsorption equilibrium, the degree of poisoning therefore rises with growing partial pressure ratio, pH_2O/pH_2 and falls with increasing temperature.

Under the assumption of a displacement equilibrium in accordance with

$$(N_2)_{ads} + H_2O_{gas} + 2\,H_{2\,gas} \rightleftharpoons (O)_{ads} + 2\,NH_3$$

I. A. SMIRNOV et al. set up a rate equation for ammonia synthesis [372], [373] that takes the effect of water vapor into consideration over a wide range of temperature and pressure:

$$v = \frac{k_1 p_{N_2} - k_{2p^2_{NH_3}}/p^3_{H_2}}{\left(p^2_{NH_3}/p^2_{H_2} + C_{p_{H_2O}/p_{H_2}}\right)^{1/2}}$$

and found the values of *C* listed in Table 19.

Table 19. *C* values for the ammonia synthesis rate equation

Catalyst	Temperature, °C				
	400	425	450	475	500
Fe + Al_2O_3	0.63	0.39	0.24	0.17	0.12
Fe + Al_2O_3 + K_2O	0.74	0.46	0.28	0.20	0.14

A more recent investigation [374], [375], [376] proposed multiplying the rate equation by a correction factor 1 – Θ, where $\Theta = a + bT + cT \ln X_{H_2O}$ and X_{H_2O} is the molar fraction of H_2O. Some authors assume a different route for the formation of adsorbed atomic oxygen [377], [378]:

$$H_2O(g) + 3* \rightleftharpoons 2\,H* + O*$$

where * denotes a surface site.

The equivalence of H_2O, CO, CO_2 and O_2 [379] with respect to their poisoning effect has been confirmed.

The experimentally determined effect of water vapor concentrations up to about 30 ppm on the activity of a commercial catalyst (BASF S 6–10) at 30 MPa is evident from Figure 23.

Figure 23. Reversible effect of increasing water vapor concentrations in the synthesis gas on the activity of industrial ammonia catalysts

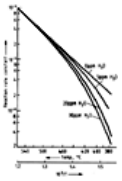

With continuing exposure oxygen compounds also cause irreversible damage to the catalyst activity that is causally linked with growth of the iron primary crystallite [376], [380]. This is probably one of the main causes of the decline in converter performance over the course of the catalyst operating life. This damage depends on the water-vapor partial pressure and is especially serious, in contrast to reversible poisoning, at high temperatures. In a pilot plant, Österreichische Stickstoffwerke (ÖSW), Linz, established that at 30 MPa and a water-vapor content of 250 ppm, for example, the production declined by about 15 % per month. For a carbon monoxide content of 5 ppm, they determined about a 4 – 5 % decrease in activity per year. The influence of the operating temperature is evident from the data. The performance of a catalyst operated at 570 °C is about 35 – 40 % under that of a catalyst charge operated at 520 °C. That is, the higher the temperature, the greater is the harmful effect of oxygen compounds [381]. Corresponding to the data, they established that above all, temperatures above 500 – 520 °C must be avoided in order to achieve a converter charge operating period of more than 2 – 3 years. If the temperature is kept at this level, generally recommended for modern plants, then even at CO concentrations of 20 ppm in the recycle loop gas, no serious deterioration of the catalyst performance is observed after a month of operation. With the highly purified synthesis gas of modern synthesis processes (with final gas purification by methanation the carbon monoxide level is lowered to below 5 ppm), operating periods of up to 14 years can be achieved for a converter charge without significant loss of activity (Fig. 24).

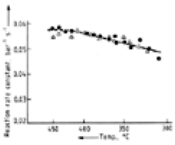

Figure 24. Activity of BASF ammonia catalyst S 6–10 in a commercial converter after 14 years of operation

• fresh catalyst; △ after 14 years operation

As already mentioned in Section Composition, the concentration and combination of promoters affect the degree of irreversible damage. This must be considered in the choice of catalyst for a particular plant.

Sulfur, Phosphorus, and Arsenic Compounds. Sulfur, occasionally present in synthesis gases from coal or heavy fuel oil, is more tightly bound on iron catalysts than oxygen. For example, catalysts partially poisoned with hydrogen sulfide cannot be regenerated under the conditions of industrial ammonia synthesis. Compounds of phosphorus and arsenic are poisons but are not generally present in industrial syngas. There are indications that these permanent poisons exert the most detrimental effect when present as hydrogen compounds and are less harmful in higher oxidation states [382].

With regard to the sulfur bound on the catalyst surface, differences exist between the various types of ammonia catalysts, especially between those containing or free of alkali and alkaline earths. Pure iron and catalysts activated simply with alumina chemisorb S_2N_2 or thiophene when treated with concentrations of H_2S that lie below the equilibrium for the FeS bond, a maximum of 0.5 mg of sulfur per m^2 of inner surface or free iron surface; this corresponds to monomolecular coverage [383], [384]. The monolayer is also preserved on reduction with hydrogen at 620 °C, whereas FeS formed by treatment above 300 °C with high H_2S concentrations is reducible as far as the monolayer. For total poisoning, 0.16 – 0.25 mg S/m^2 is sufficient. Like oxygen, sulfur promotes recrystallization of the primary iron particle.

Under similar poisoning conditions, alkali- and alkaline earth-containing industrial catalysts adsorb more H_2S. In spite of this, however, in terms of activity, they are more stable toward the action of sulfur and are partially regenerable [384]. In a catalyst bed, most of the sulfur already has been taken up in the gas inlet layer. A catalyst sulfur content of several 100 ppm suffices to impair its activity [14].

In industrial plants, sulfur may reach the ammonia converter in various forms. In some plants, traces of H_2S and COS may not be removed in the upstream purification steps and so may enter the converter with the make-up gas. However, the sulfur contained in the compressor oil constitutes the main danger. On cracking the oil to lower molecular mass hydrocarbons, sulfur is freed as H_2S. It is therefore very important to use an oil with low sulfur content in ammonia plants, especially those still using reciprocating compressors. If after mixing with the recycle gas the make-up gas first runs through an ammonia condensation stage in which the H_2S and also to a certain extent COS are very effectively washed out by condensing ammonia, a sulfur content of the oil of 0.1 – 0.2 wt % ought to be sufficiently low. Otherwise, a value under 0.1 % is recommended [385]. In modern plants designed with centrifugal compressors and in which the sulfur content of the synthesis gas is extremely low because of very effective purification (about 0.5 – 1 µg S/m^3 (STP)), sulfur poisoning is of lesser importance than carbon monoxide and chlorine poisoning.

Chlorine compounds. The permanent poisoning effect of chlorine compounds is two orders of magnitude worse than that of oxygen compounds. Concentrations of about 0.1 ppm are viewed as the uppermost allowable limit in order not to affect adversely the life of ammonia catalysts [385]. The deactivation effect is based at least in part on the formation of alkali chlorides that are volatile at the upper synthesis temperatures.

Further information on catalyst poisoning is given in [375], [376].

4.4.2. *Other Catalysts*

4.4.2.1. *General Aspects*

For a long time efforts to improve the efficiency of industrial ammonia production concentrated on synthesis gas production, and major progress was achieved over the years. In ammonia synthesis itself considerable progress was made in converter design and recovery of the reaction energy at high temperature, but there has been no substantial improvement in the catalyst since the 1920s. The standard commercial iron catalyst still requires relatively high pressures (usually in excess of 130 bar), high temperatures (400 – 500 °C) and large reactor volumes (more than 60 m^3 for a capacity of 1500 t/d) to achieve good economics, although in a few cases a pressure as low as 80 bar has been used. From the vast amount of experimental and theoretical studies of the iron catalyst one can conclude that there is only limited potential for further improvement. Substantial energy savings would require lowering the synthesis pressure considerably, down to the synthesis gas production level, say. However, to compensate for the less favorable equilibrium situation much lower operating temperatures would be necessary, because otherwise too low an ammonia concentration would result, and additional energy would be needed for recovery, thus cancelling the energy saving from synthesis gas compression. To reach this goal a synthesis catalyst with a volumetric efficiency some two orders of magnitude greater than magnetite would be necessary. Process studies show an energy saving potential of about 1 GJ per tonne NH_3 [1279].

In the search for an alternative catalyst, most metals have been tested, either as primary components or as promoters. Much of this work was performed in the early, pioneering studies in the BASF laboratories [250], [251]. Most of the studies in the following years concentrated on the magnetite system in the sense of more fundamental and general catalytic research. Rising energy costs since the mid-1960s have given a new incentive to the search for other catalyst systems with improved performance. The first development which found commercial application was a cobalt-modified magnetite catalyst introduced in 1984 by ICI. With similar kinetic characteristics its volumetric activity is about twice that of the standard iron catalyst. The only other catalyst system which exhibits a promising potential for industrial application is based on ruthenium [195]. These new efforts to find improved catalysts could use methods and knowledge of modern surface science as developed on the example of the magnetite catalyst. Structural sensitivity and nature of the nitrogen adsorption and dissociation steps could serve as guidelines [196].

For the overall performance of potential catalysts in practical application additional factors, such as number of active sites, physical form, and porosity must also be taken into account. The classical commercial iron catalyst is an unsupported catalyst. First of all iron is a cheap material and secondly by the incorporation of alumina a surface area similar to that attained in highly dispersed supported catalysts can be obtained. Of course, for an expensive material such as the platinum group metals, the use of a support material is the only viable option. The properties of the supported catalyst will be influenced by several factors [195]

– Adequate surface of the carrier to achieve a reasonable metal loading.

- Dispersion stability by using a more active support with strong interaction between the support phase and the metal precursor. Too strong an interaction may cause difficulties in the metal reducing.
- Promoter localization with respect to metal sites and support sites.
- Gas transport effects will be governed by pore size, pore distribution, and tortuosity.
- Anion retention capability, which plays a role in the catalyst preparation by impregnation of a support with metal impregnation of the support.

4.4.2.2. *Metals with Catalytic Potential*

Materials that show significant ammonia synthesis activity can be divided into three categories according to their ability to form nitrides:

1. Platinum group metals: no stable nitrides (Ru, Os, Ir, Pt)
2. Metals forming nitrides unstable under reaction conditions (Mn, Fe, Co, Ni, Tc, Re)
3. Metals likely to be present as nitrides under synthesis conditions (groups 3 – 6 of the periodic table)

Of the platinum group metals only ruthenium and osmium show an activity superior to iron, though only in presence of alkali metal promoters, as may be seen from Table 20[386].

Table 20. Ammonia synthesis activity of metals supported on carbon with potassium metal promotion (mL NH_3/ mL catalyst, 573 K, 128.313 kPa, H:N = 3:1)

	Fe 0.72	Co 0.4	Ni 0.04	
Mo 0.6	Ru 22.2	Rh 0.52	Pd 0	
Re 0.36	Os 5.6	Ir 0.68	Pt 0.008	

Although osmium was the first active catalyst used by HABER in 1909 [387] in his laboratory-scale unit to demonstrate the technical viability of the high-pressure recycle process, this metal never became an industrial catalyst because of its limited availability and its dangerous properties. The interest therefore shifted to ruthenium. The most active ruthenium catalysts use a graphite support and alkali metal promotion, preferentially with rubidium or cesium [394], [398]. It is unlikely that the promoter is in metallic state as its high vapor pressure would probably lead to substantial losses under synthesis conditions. It is assumed that a charge transfer complex $M^{+}\cdots C^{-}$ is formed between the metal and the graphite. A major advantage of graphite is its ability to stabilize high loadings of alkali metals. A special limitation until recently was the chemical reactivity of carbon supports. In a typical ammonia synthesis environment with high hydrogen partial pressure, the catalyst may also catalyze the methanization of carbon, which would lead to destruction of the support. Indeed, this phenomenon has been observed [388], but it can be avoided by careful heat treatment of the support above 1500 °C [389]. With this modified carbon material lifetimes of at least six years are expected [386].

The group of metals forming low-stability or unstable nitrides includes Mn, Fe, Co, Ni, Tc, and Re. As in the case of iron a clear structural sensitivity was found for rhenium but the role of promoters remains the subject of discussion. There are also indications of structure sensitivity for cobalt and nickel. It was attempted to improve the activity of the classical magnetite catalyst by alloying with nickel or cobalt. The only commercial catalyst is a cobalt containing magnetite [390].

Of the group of metals forming stable nitrides, only molybdenum is of some interest. Under synthesis conditions it is present as a nitride with some ammonia formation activity and structural sensitivity [391]. Molybdenum also seems to exhibit activity in biological nitrogen fixation [392] and is synthetically active at ambient conditions in the air-sensitive Glemser compounds [392], [393].

The results of the intensive research in this field over the last decades demonstrate that, irrespective of the catalyst, the rate-determining step in the ammonia synthesis reaction is the dissociation of the nitrogen molecule and the catalyst effectivity is determined in the first instance by the activation energy of the dissociation reaction. The other common factor for the ammonia catalysis is the structure sensitivity of molecular nitrogen adsorption. Only if both conditions are favorable, and other factors such as hydrogen and ammonia inhibition do not play a major role, can a sufficient overall reaction rate be expected. The available data show that these conditions are fulfilled only for a limited number of metals: iron, ruthenium, and osmium. On account of its very strong ammonia inhibition rhenium is not an option. Extensive literature on non-iron catalysts is given in [195], [196]; kinetic investigations are reported in [196].

4.4.2.3. *Commercial Ruthenium Catalysts*

Since the early days of industrial ammonia synthesis only minor improvements have been achieved for the magnetite system: optimization of manufacturing procedures, promoter concentrations, and particle size to give somewhat higher activity and longer service life.

A notable development for the magnetite system was the introduction of cobalt as an additional component by ICI in 1984 [395], [396]. The cobalt-enhanced catalyst formula was first used in an ammonia plant in Canada using ICI Catalco's AMV process (later also in other AMV license plants) and is also successfully applied in ICI's LCA plants in Severnside. Commercially successful is a new type of iron-based catalyst, the AmonMax-10 of Südchemie which uses wustite instead of magnetite. This catalyst is already used in seven ammonia plants with a total capacity of 5600 t/d [1320-1322], [1410]. Production of synthetic wustite is described in [1324], its application for ammonia synthesis catalyst in [1323]. For a literature report on promoted iron catalysts see [1325].

In 1979 BP disclosed to M.W. Kellogg a new catalyst composed of ruthenium on a graphite support [395], [397]. In October 1990, after a ten-year test program, Kellogg started the commercialization of the Kellogg Advanced Ammonia Process (KAAP) using this catalyst [397], which is claimed to be 10 – 20 times as active as the traditional iron catalyst: The KAAP catalyst. According to the patent [390] the new catalyst is prepared by subliming ruthenium-carbonyl $[Ru_3(CO)_{12}]$ onto a carbon-containing support which is impregnated with rubidium nitrate. The catalyst has a considerably higher surface area than the conventional catalyst and, according to the patent example, it should contain 5 wt % Ru and 10 wt % Rb. Besides having a substantially higher volumetric activity, the promoted ruthenium catalyst works best at a lower than stoichoimetric H/N ratio of the feed gas as shown in Figure 25. It is also less susceptible to self-inhibition by NH_3 (Figure 26) and has excellent low-pressure activity.

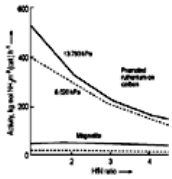

Figure 25. Effect of H/N ratio on activity of Ru and Fe_3O_4 catalysts [195]

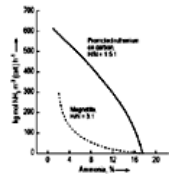

Figure 26. Ammonia inhibition of Ru and Fe_3O_4 catalysts [195]

Three ammonia plants in Trinidad, each with a nameplate capacity of 1850 t/d, use the KAAP ruthenium catalyst [1317]. The first layer of the four-bed reactor [1318], [406] is loaded with conventional magnetite catalyst, the others with the ruthenium – graphite catalyst. For the special properties of ruthenium, which have to be considered in handling these catalysts see [406]. Topsøe has developed a ruthenum catalyst which uses boron nitride instead of graphite as support. Boron nitride, occasionally called “white graphite”, because of its structure similar to graphite, is completely stable towards hydrogenation under all conditions relevant to technical ammonia synthesis. A Ba-Ru/BN catalyst proved completely stable in a 5000 h test at 100 bar and 550 °C using a 3:1 H_2 – N_2 mixture [1326]. There is no industrial

application so far. A discussion on ruthenium as ammonia synthesis catalyst with an extensive literature review is found in [1389].

The potential for ruthenium to displace in the long run iron in new plants will depend on whether the benefits of its use are sufficient to compensate the higher costs. Ruthenium prices increased by a factor of four in 2000 to 170 $/oz. due to new applications in the electronic industry. Ruthenium price development is occasionaly published in FINDS, a Stokes Engineering Publication.

In common with the iron catalyst ruthenium will also be poisoned by oxygen compounds. Even with some further potential improvements it seems unlikely to reach an activity level which is sufficiently high at low temperature to allow operation of the ammonia synthesis loop at the pressure level of the syngas generation. An overview over ammonia synthesis catalyst development history and newer research is given in [1326].

4.5. Process Steps of Ammonia Production

The term “ammonia synthesis” is increasingly used when referring to the total ammonia production process. Synthesis conditions are no longer viewed in isolation. Of course, they are an important consideration in the total process but can be determined properly only in relation to the total plant integration (see Section Complete Ammonia Production Plants). A literature summary of ammonia production is contained in [402] and [403]; the United States patent literature in the field from 1972 to 1980 is covered in [304]. More modern and comprehensive reviews of ammonia production technology can be found in [404-406], [1318], [1132]. The journal *Nitrogen*, now named *Nitrogen + Syngas*, published by British Sulphur, presents an update of the state of the art from time to time.

The complete process of industrial ammonia production may be subdivided into the following sections:

A. Synthesis gas production
   1. Feedstock pretreatment and gas generation
   2. Carbon monoxide conversion
   3. Gas purification
B. Compression
C. Synthesis and purge gas management

The most fundamental changes over the years have occurred in synthesis gas production and gas compression. In the synthesis section itself, some progress has been made in converter design and optimization of heat recovery.

Synthesis gas preparation, carbon monoxide conversion, and gas purification for ammonia synthesis are treated in detail in the article ⟶ Gas Production. Therefore, only a short abstract is presented here.

4.5.1. *Synthesis Gas Production*

The goal is preparing a pure mixture of nitrogen and hydrogen in the stoichiometric ratio of 1 : 3. The raw materials are water, air, and a carbon-containing reducing medium, that, for its part, may contain hydrogen (natural gas, CH_4 ; naphtha, ≈ CH_2; petroleum, ≈ CH) and nitrogen; for example, natural gas from the Slochteren field in the Netherlands contains 14 % nitrogen.

Usually only the carbon-containing materials and hydrogen from other sources are regarded as raw materials in the narrow sense because of the abundance of air, which provides all of the nitrogen, and water, which generally supplies most of the hydrogen. The term feedstock is often applied to the total consumption of fossil fuel, although strictly speaking a distinction should be made between gasification feed and fuel for energy generation.

Certain raw materials for synthesis gas production that were once of primary importance currently are used only under special economic and geographical circumstances (e.g., China, where 66 % of production is based on coal). These include solid fuels, coke oven gas, and hydrogen produced by electrolysis. Reference [409] covers coke oven gas as a feedstock for ammonia synthesis and references [410-413] describe producing hydrogen by water electrolysis for ammonia production.

Table 21 provides an overview of the raw material sources (apart from water and air) for world ammonia capacity.

Table 21. Feedstock distribution of world ammonia production capacity

	1962		1972		1983		1998	
	10^3 t N	%	10^3 t N	%	10^3 t N	%	10^3 t N	%
Coke oven gas and coal	2800	18	4600	9	7200	8	16500	14
Natural gas	7800	50	32100	63	66850	74	94300	77
Naphtha	2050	13	10700	21	9050	10	7300	6
Other petroleum products	2950	19	3600	7	7200	8	4400	3
Total	**15600**	**100**	**51000**	**100**	**90300**	**100**	**122500**	**100**

Table 21 indicates that new ammonia plants are based almost exclusively on natural gas and naphtha. This trend is also expected to continue in the near future. Naturally, the regional distribution is diverse. In North America, for example, natural gas dominates, with 95 % of capacity. In the EU, 86 % of capacity is based on natural gas and 8 % on naphtha [502].

The capital cost and the specific energy requirement (i.e., feed and fuel, and so the manufacturing cost) largely depend on the raw material employed [414], [415]. Table 22 shows the relative capital cost and the relative energy requirement for a plant with a capacity of 1800 t/d ammonia. For the natural gas based plant the current best value of 28 GJ per tonne NH_3 is used. If water electrolysis (4.5 kWh/Nm^3 H_2) is used together with an air separation unit for the nitrogen supply the energy requirement amounts to 34 GJ per tonne NH_3 when the electric energy is valuated just with the caloric equivalent, which would be only justified when electric power is generated from water power. In case of electricty generation from fossil energy with 40 % efficiency the consumption figure is 85 GJ per tonne NH3, which is 300 % of the consumption of a modern steam reforming plant [1364]. The investment could according to a rather rough estimate be about three times the investment for a steam reforming plant with natural gas.

Table 22. Relative ammonia plant investment and relative energy requirement for 1800 t/d NH_3

	Natural gas	Naphtha	Fuel oil	Coal
Relative investment	1.0	1.15	1.5	2.5
Relative specific energy requirement (based on lower heating values)	1.0	1.1	1.3	1.6

4.5.1.1. *Feedstock Pretreatment and Raw Gas Production*

The chemical reaction of hydrocarbons with water, oxygen, air, or any combination of these is generally referred to as gasification. It yields a gas mixture made up of CO and H_2 in various proportions along with carbon dioxide and, where air is used, some nitrogen. Any carbon containing feedstock will undergo a reaction according to Equation (5) or (6) or both simultaneously.

$$[CH_x] + H_2O \rightleftharpoons CO + H_2 + x/2\,H_2 \Delta H > 0 \quad (5)$$

$$[CH_x] + 1/2\,O_2 \rightleftharpoons CO + x/2\,H_2 \quad \Delta H < 0 \tag{6}$$

Light hydrocarbons ranging from natural gas (methane) to naphtha (max. C_{11}) undergo reaction with steam over a catalyst according to Equation (5) which is usually called steam reforming. Corresponding to Equation (6), commonly known as partial oxidation, all carbon-containing feedstocks can be processed in a noncatalytic reaction with oxygen (together with a minor amount of steam for process reasons, which gives rise to a simultaneous reaction according to Eq. 5). An additional equilibrium reaction involved in any gasification process is the water gas shift reaction (7).

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad \Delta H^0_{298} = -41\ \mathrm{kJ/mol} \tag{7}$$

Although reaction in the right hand direction is favored by lower temperatures it is responsible for the initial carbon dioxide content of the raw synthesis gas. To maximize the hydrogen yield, this reaction is carried out in a separate step over a different catalyst at a lower temperature than the preceding gasification step (Section Carbon Monoxide Shift Conversion).

From Equation (5) it can be seen that in the steam reforming variant the proportion of hydrogen supplied by the feedstock itself increases with its hydrogen content. It attains the theoretical maximum of 66 % with methane. The hydrogen – oxygen bond energy in water is higher than the hydrogen – carbon bond energy in the hydrocarbon. The positive enthalpy per mole of hydrogen therefore decreases as the proportion of hydrogen contributed from the feedstock itself increases. Natural gas consists predominantly of methane and is therefore the most hydrogen-rich and energetically the best raw material for the steam-reforming route. In the partial-oxidation route less hydrogen is produced in the primary gasification step and the raw synthesis gas has a rather high CO content.

The raw gas composition is thus strongly influenced by the feedstock and the technology applied. But for the different feedstocks there are some constraints on the applicability of the various gas generation processes. The catalytic steam reforming technology can only be applied to light hydrocarbon feedstock (up to naphtha) but not for heavy hydrocarbons such as fuel oil or vacuum residue. These raw materials contain a substantial amount of sulfur and also minor quantities of heavy metals, which would poison the sensitive reforming catalyst. In addition cracking reactions will occur on the catalyst, depositing carbon, which not only blocks the catalyst pores but also restricts interparticle flow. Thus for heavy feedstocks the only choice is noncatalytic partial oxidation, which, however, is capable of processing any type of hydrocarbon feedstock. The various commercial coal gasification processes may also be classified as partial oxidations.

Steam Reforming Processes. (see also ⟶ Gas Production – Steam Reforming of Natural Gas and Other Hydrocarbons). As the nickel-containing catalysts are sensitive to poisons, any sulfur compounds present in the hydrocarbon feedstock have to be removed by hydrodesulfurization, generally with a combination of cobalt – molybdenum and zinc oxide catalysts [560], [561], [639].

Adsorption on activated carbon is an alternative when the feed is natural gas with a rather low sulfur content.

$$R - SH + H_2 \rightarrow RH + H_2S \tag{8}$$

$$H_2S + ZnO \rightarrow ZnS + H_2O \tag{9}$$

The general overall reaction can be formulated as

$$C_nH_{(2^n+2)} + n\,H_2O \rightleftharpoons nCO + (2n+1)H_2 \tag{10}$$

or more specifically for methane, usually the major constituent of natural gas, as:

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2 \quad \Delta H = +206\ \mathrm{kJ/mol} \tag{11}$$

Simultaneously to this equilibrium the water gas shift reaction (Eq. 7) proceeds.

To introduce nitrogen to achieve the required stoichiometric hydrogen/nitrogen ratio for ammonia synthesis, the reforming reaction is split into two sections. In the first section, the primary reformer, the reaction proceeds in indirectly heated tubes filled with nickel-containing reforming catalyst and is controlled to achieve a partial conversion only [in conventional plants 65 % based on methane feed, leaving around 14 mol % methane (dry basis) in the effluent gas]. In the following secondary reformer — a refractory-lined vessel filled with nickel catalyst — the gas is mixed with a controlled amount of air introduced through a nozzle (burner). By combustion of a quantity of the gas the temperature is raised sufficiently (to about 1200 °C) that the endothermic reforming reaction is completed with the gas adiabatically passing the catalyst layer. In this way the outlet temperature is lowered to around 1000 °C, and a residual methane content of 0.5 % or lower (dry basis) is attained in conventional plants [404, 406]. Nitrogen already present in the natural gas tends to cause a reduced specific air ratio in the secondary reformer and a reduced secondary reformer temperature rise. Therefore, to maintain the same methane leak, the primary reformer exit temperature must be increased.

The primary reformer consists of a multitude of reformer tubes loaded with the nickel catalyst (15 – 25 % NiO on α-aluminum oxide, calcium aluminate, or magnesium aluminum spinel support) [1318] in a furnace box in which the heat needed for the reaction is transferred to the tubes by radiation. The heat is generated in burners, generally gas-fired, in the furnace box.

A special consideration is the lifetime of the expensive reformer tubes, made of highly alloyed chromium-nickel steel by centrifugal casting, because under the severe reaction conditions the material exhibits creep which finally leads to rupture [1327], [1390]. The time to rupture for a specific material depends on the tube-wall temperature and on the internal pressure. This limits the reforming pressure, which to save energy in synthesis-gas compression should be as high as possible. As the reforming reaction is endothermic and proceeds with volume increase the negative effect of a pressure increase (lower conversion) has to be compensated by a higher reaction temperature and hence higher wall temperatures, but this is limited by the material. Another possibility to compensate is a higher steam surplus (steam/carbon ratio), but this is economically unfavorable. The furnace box usually accommodates 200 – 400 tubes (depending on plant capacity), 10 – 13 m long, with an inner diameter of 75 – 140 mm and a wall thickness of 11 – 18 mm. The standard material for a long time was HK 40 (20 Ni/25 Cr) but for replacements and new plants, HP modified (32 – 35 Ni/23 – 27 Cr stabilized with about 1.5 % Nb) is being increasingly used on account of its superior high-temperature properties [425]. With this latter tube material a reforming pressure of 40 bar is possible at outer tube-wall temperatures of around 900 °C . A further improvement are *microalloys* that additionally contain Ti and Zr [406], [1318], [1373], [1376].

According to the disposition of the burners primary reformers can be classified as top-fired, side-fired, terraced-wall, or, less common, bottom-fired reformers [406], [1318]. The steam/carbon ratio used in modern commercial primary reformers for natural gas is between 2.8 and 3.5, and markedly higher for naphtha, for which an alkalized nickel catalyst has to be used to prevent carbon deposition, which causes catalyst deactivation and local overheating of tubes (hot bands and hot spots). A special requirement for the catalyst is a low content of silica, which could be volatile under the hydrothermal conditions and deposit downstream of the secondary reformer at lower temperature, causing fouling of the waste-heat recovery surfaces.

When higher hydrocarbons (naphtha) are used, a so-called “rich gas stage” can be utilized upstream of the primary reformer. In this, the higher hydrocarbon – steam mixture is transformed at relatively low temperature (400 – 500 °C) and steam – carbon ratios less than two into a methane-rich gas. This is then converted in the primary reformer under normal reforming conditions. For this rich-gas stage, three processes are commercially available: the British Gas Process [426], the BASF – Lurgi Process [427-429], and the Japan Gasoline Process [430].

This process has recently become the subject of renewed interest for increasing the capacity of existing natural-gas-based ammonia plants in which the primary reformer has been identified as bottle-neck. Under the name pre-reforming, it is installed up-stream of an existing tubular reformer [431-438], [406], [1318]. The natural gas enters such a pre-reformer with a temperature of 530 °C instead of being fed to the primary reformer tubes with the same temperature. A temperature drop of about 60 – 70 °C occurs in the catalyst bed due to the overall endothermic reaction. Medium-grade heat is used to reheat the exit gas to the correct primary reformer entrance temperature. This compensation heat may be derived from a variety of sources, including flue gas, process gas, or gas turbine exhaust.

Another possibility for reducing the load on the primary reformer is to transfer part of the conversion duty to the secondary reformer with application of an super-stoichiometric amount of air. This requires removal of the surplus of nitrogen from the synthesis gas either ahead of the synthesis loop (by cryogenic methods or pressure swing adsorption) or by purge gas and hydrogen recovery [457], [458]. In the extreme case the whole reforming reaction could be performed without a tubular reformer by autothermal catalytic reforming in a design similar to a secondary reformer. In this case it would be necessary to use oxygen or oxygen-enriched air instead of air [404], [446], [459], [460], [1318], [1413, 1414]. Unlike to a secondary reformer, which is fed with partially reformed gas, having a substantial concentration of hydrogen, the autothermal reformer (ATR) is fed directly with the hydrocarbon feedstock. Because of the higher heat of reaction in the internal combustion (temperature >2000 C), the flow conditions, heat release characteristics and the risk of soot formation are very different from the situation of a normal secondary reformer. Therefore special considerations in the design of burner and reactor are necessary. See also ⟶ Gas Production.

A recent development which avoids a fired primary reformer is the exchanger reformer, which with some simplification may be viewed as tubular heat exchanger with

the catalyst inside the tubes, which are heated by the hot secondary reformer effluent flowing on the shell-side. In some designs the tubes may be open at the lower end, in which case the gas flow on the shell side consists of a mixture of the off-gases from the secondary reformer and from the reformer tubes. Commercially operating designs are the GHR of ICI (Figure 27) [406], [461-464] and the KRES of M. W. Kellogg [466-470], [1318-1320], [1398] (Fig. 28). Similar concepts are offered by other licensors and contractors e.g., Braun & Root, or Topsøe [406].

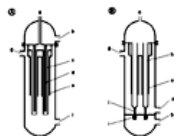

Figure 27. ICI gas heated reformer

A) GHR; B) AGHR

a) Tubeside inlet; b) Tubeside outlet; c) Scabbard tube; d) Bayonet tube; e) Sheath tube; f) Shellside inlet; g) Shellside outlet; h) Catalyst tube; i) Seal; j) Tail pipe; k) Catalyst; l) Refractory lining

ICI has come out with a modified design, the AGHR, with “A” standing for advanced. As shown in Figure 27 B, the bayonet tubes are replaced by normal tubes attached to a bottom tubesheet using a special seal to allow some expansion. In this way the delicate double tubesheet of the GHR is avoided. The seal which prevents leakage of methane-rich gas to the secondary reformer effluent flowing on the shell side has a unique design which is subject to patent applications of ICI. The AGHR will allow a single-line concept for worldscale plants whereas with the GHR several parallel units for large plants would be necessary [406].

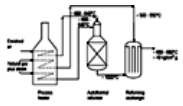

Figure 28. Kellogg reforming exchanger System (KRES) [406]

Because of the smaller size compared to a conventional fired reformer considerable investment savings can be achieved. To close the heat balance between the exchanger reformer and secondary (autothermal) reformer, the latter has to take on a higher reforming duty, which may be achieved by using an over-stoichiometric amount of air or oxygen-enriched air. In some configurations the exchanger reformer is partially by-passed, part of the feed being fed directly into the autothermal reformer.

Haldor Topsøe has developed an exchanger reformer called HTER which may be used in various process configurations as described in [1328]. The design may be also used as a convective reformer, called HTCR with a dedicated burner.

A concept developed by Uhde goes a step further in this direction: exchanger reforming and subsequent noncatalytic partial oxidation, which provides the reaction heat, are accommodated in a single vessel. This combined autothermal reformer (CAR) design, shown in Figure 29 was operated in a demonstration unit producing 13 000 m^3/h of synthesis gas [472-475].

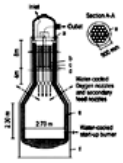

Figure 29. Uhde combined autothermal reformer (CAR)

a) Sandwich type tubesheet; b) Enveloping tube; c) Reformer tubes; d) Tubesheet; e) Refractory lining; f) Water jacket

A unique steam reforming process [439-441] has been developed in Japan to the pilot-plant stage. It reportedly can operate without upstream desulfurization and should be able to gasify naphtha, crude oil, and atmospheric or vacuum residues.

For literature on steam reforming see [404-408], [420], [442-448]; reaction kinetics and thermodynamics [1318]; for steam reforming of naphtha, [451-453]; for steam reforming catalysts, [454], [455], [406], [456], [1318], [1331] for reformer design, [406], [449, 450], [1318]. An oververview of the steam reforming technology, its historical development, present state of the art, newer developments and future perspectives together with quotation of newest literature is given in [1327]. Detailed discussion on exchanger reforming and autothermal reforming is found in [1328].

Partial Oxidation. Hydrocarbons or coal will react with an amount of oxygen insufficient for total combustion to CO_2 according to:

$$C_nH_m + n/2\,O_2 \rightarrow n\,CO + m/2\,H_2 \tag{12}$$

$$C + 1/2\,O_2 \rightarrow CO \quad \Delta H = -123\,kJ/mol \tag{13}$$

In practical operation some steam must always be added, the quantity depending on feedstock and process configuration, so that the following reactions proceed in parallel:

$$C_nH_m + n\,H_2O \rightarrow n\,CO + (n + m/2)H_2 \tag{14}$$

$$C + H_2O \rightarrow CO + H_2 \quad \Delta H = +119\,kJ/mol \tag{15}$$

As the overall reaction is exothermic no external heat supply is necessary. Since in some processes with coal feedstock (e.g., the Lurgi Process) the reaction according to Equation (15) may proceed to a considerable extent, they are more often referred to as coal gasification rather than as partial oxidation, but this is just a matter of definition.

Certain processes have achieved particular significance. The Texaco and Shell processes can handle fuel oil, vacuum residues, and coal. The Lurgi process handles lump coal, operates at 2.5 MPa, the Koppers – Totzek process operates on coal dust at atmospheric pressure.

Partial Oxidation of Hydrocarbons. The two dominant processes are the Texaco Syngas Generation Process (TSGP) and the Shell Gasification Process (SGP). In both the reaction is performed in an empty pressure vessel lined with alumina. The reactants (oil and oxygen, along with a small amount of steam) are introduced through a nozzle at the top of the generator vessel. The nozzle consists of concentric pipes so that the reactants are fed separately and react only after mixing at the burner tip or in the space below. The temperature in the generator is between 1200 and 1400 °C. Owing to insufficient mixing with oxygen, about 2 % of the hydrocarbon feed is transformed into soot, which is removed from the gas by water scrubbing. Depending on the process configuration the gas is either cooled by quenching or in a waste-heat boiler. The TSGP and SGP are rather similar, they mainly differ in the nozzle design, soot removal and recirculation, and process gas cooling. The reaction pressures may be as high as 80 bar; there are no mechanical or material limitations to raising the gasification pressure further, but with respect to the overall ammonia process this might be beyond the energy optimum because of the increasing energy demand for nitrogen and oxygen compression. Maximum raw gas generation capacity of a single generator corresponds to about 1000 t/d ammonia. World-scale ammonia plants based on partial oxidation exist in Germany, India, China, and other countries [465]. Further information is found for the TSGP in [421-424], [486], [487, 489-491], for the SGP in [471], [477-485], [492]; see also ⟶ Gas Production – Partial Oxidation of Hydrocarbons.

Partial Oxidation of Coal. So far the Koppers – Totzek, Texaco and Lurgi gasifiers, and probably the Winkler process in some smaller installations, have been used in ammonia plants, but the successful demonstration of the Shell process in other applications makes it a potential candidate for ammonia production, too. Additional processes in different stages of technical development are the HTW and the Dow process. Information on the status and the development in the gasification of coal can be obtained from [493-496], [498-501], [503]. A special case is China, where a number of small plants with a capacity of 30 – 80 t/d still produce their synthesis gas with the today largely outdated water-gas process which operates at atmospheric pressure. Coke or anthracite is reacted intermittently with air and steam in a fixed bed. The heat produced by the exothermic reaction of coal and air in the “blow phase” is stored in the fixed bed and provides the heat needed for reaction of coal with steam in the “run phase” [99], [497].

The Koppers – Totzek Process [504], [506-509], [510] used in several ammonia plants in China, India, and South Africa, operates practically at atmospheric pressure. Dry coal dust is fed to the two (sometimes four) burners of the gasifier. Oxygen, together with a small amount of steam, is introduced immediately at the head of the burners, and the mixture enters the reaction zone with high velocity. The residence time is less than 1 s and the temperature is 1500 – 1600 °C. The gas leaving the top of the reactor vessel is cooled in a waste-heat boiler, followed by a water scrubber to remove carbon and ash traces from the raw gas and to effect further cooling. Liquid slag is withdrawn from the reactor bottom. The process can handle bituminous coal and lignite. To overcome the disadvantage of atmospheric pressure operation, a version was developed capable of operating at 25 – 30 bar. This Prenflow (pressurized entrained flow) process [511], [1132] is being tested in a 48 t/d pilot plant.

The Texaco Coal Gasification Process [495], [518-525] is rather similar to the Texaco partial oxidation process for heavy hydrocarbons. An aqueous slurry containing 60 – 70 % coal is fed by reciprocating pumps to the generator at a pressure of 20 – 40 bar. Waste-heat boiler, quencher, and carbon scrubber are especially adapted to deal with the ash and slag introduced with the bituminous coal feed.

In its coal gasification process [495], [512-517]. Shell has completely departed from the concept of its process for partial oxidation of heavy hydrocarbons. With reversed flow pattern in the gasifier (from bottom to top), dry coal dust is introduced via lock hoppers into the reactor vessel operating at 20 – 40 bar.

The Lurgi Dry Gasifier [495], [526-530], [532], [533] performs the reaction in a moving bed, usually operating at 25 – 30 bar. Crushed coal with a particle size of 4 – 40 mm enters the top of the gasifier through a lock hopper and is evenly distributed over the cross-section of the coal bed surface by a distribution disk equipped with scraper arms. Ash with only 1 % of residual carbon is removed at the bottom of the gasifier by a revolving grid with slots through which steam and oxygen are introduced. The temperature in the lower section of the bed is around 1000 °C, and at the top where the raw gas exits about 600 °C. As a result of this lower temperature the raw gas has an increased content of impurities such as tars, phenols, and some higher hydrocarbons. In addition the methane content is relatively high (up to 10 – 15 %), so that purification and conditioning of the raw gas is a rather elaborate task. The process actually can use any sort of coal and can handle ash contents higher than 30 %. The British Gas/Lurgi Slagging Gasifier [495], [528], [531] operates without a grate with withdrawal of the liquid slag.

The classic Winkler gasifier, which operates at atmospheric pressure, is today still in use in some smaller plants (e.g., in China). In a further development (HTW process) by Rheinbraun it has been tested in a demonstration plant with lignite at 10 bar [534], [535].

The Advanced Coal Gasification Process (ACGP) [537], a new concept developed by AECI Engineering, Kynoch Ltd and Babcock Wilson, is an entrained gasification process at atmospheric pressure. Pulverized coal and oxygen are injected near the base of the gasifier through eight burners using a proprietary feed technique which employs a special pump. The gasifier is a slender column with a square cross-section. The combustion area is lined with heat resistant ceramic material. The upper part is made of finned tubes which are welded together and are circulated with water by thermosyphon action. So far no commercial installation exists but it is claimed that one unit can produce the synthesis gas for 750 t/d ammonia.

A block diagram showing steam reforming and partial oxidation together with the further steps needed to transform the raw gas into pure make-up gas for the synthesis is shown in Figure 30.

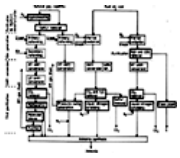

Figure 30. Alternative process steps for generation and purification of synthesis gas

4.5.1.2. *Carbon Monoxide Shift Conversion*

As ammonia synthesis needs only nitrogen and hydrogen, all carbon oxides must be removed from the raw synthesis gas of the gasification process. Depending on feedstock and process technology, this gas contains 10 – 50 % carbon monoxide and also varying amounts of carbon dioxide. In the water gas shift reaction, traditionally known as carbon monoxide shift conversion (Eq. 7), the carbon monoxide serves as reducing agent for water to yield hydrogen and carbon dioxide. In this way not only is the carbon monoxide converted to readily removable carbon dioxide but also additional hydrogen is produced:

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad \Delta H = -41.2\,\mathrm{kJ/mol} \tag{7}$$

As no volume change is associated with this reaction, it is practically independent of pressure, but as an exothermic process, it is favored by lower temperatures, which shift the equilibrium to the right-hand side. Even with a low excess of steam in the gas, the equilibrium concentrations of CO are low; for example, 0.2 vol % at 220 °C and 0.12 vol % at 200 °C for a steam/gas ratio of 0.4.

To keep the temperature low the heat of reaction must be removed in appropriate way and to achieve a sufficient reaction rate effective catalysts have to be applied. The process is therefore performed in steps, with intermediate heat removal between the individual catalyst beds in which the reaction runs adiabatically. Quasi-isothermal reactors have been developed in which cooling tubes run though the catalyst layers. As the process configuration and catalysts are to some extent different for steam reforming and partial oxidation, they are treated separately here.

Shift Conversion in Steam Reforming Plants. In the traditional plant concept, the gas from the secondary reformer, cooled by recovering the waste-heat for raising and superheating steam, enters the high-temperature shift (HTS) reactor loaded with an iron – chromium catalyst at 320 – 350 °C. After a temperature increase of around 50 – 70 °C (depending on initial CO concentration) and with a residual CO content of around 3 % the gas is then cooled to 200 – 210 °C for the low temperature shift (LTS), which is carried out on a copper – zinc – alumina catalyst in a downstream reaction vessel and achieves a carbon monoxide concentration of 0.1 – 0.3 vol %.

In the unreduced state the HTS catalyst is iron(III) oxide (Fe_2O_3) containing additionally 5 – 10 % chromic oxide (Cr_2O_3). During operation, it is reduced more or less stoichiometrically to the composition of magnetite (Fe_3O_4). This catalyst is active in the temperature rage of 300 – 500 °C. Steam surplus is not only necessary for thermodynamic reasons but also to suppress undesirable side reactions. Decreasing the steam surplus lowers the oxygen to carbon ratio in the HTS to such an extent that the atmosphere can reduce magnetite partially to metallic iron. In addition the Boudouard reaction can occur under these conditions. The resulting carbon is deposited within the catalyst particles causing their disintegration, and iron carbides will be formed, which are effective Fischer – Tropsch catalysts that lead to the formation of some methane and higher hydrocarbons [406], [544], [545]. Newly introduced HTS catalysts with additional copper promotion suppress this side reaction [546] and are therefore less sensitive to lower steam-to-gas ratios. The classical HTS iron catalyst is resistant against sulfur compounds, but this is of greater importance in partial oxidation processes and less for the practically sulfur-free steam reforming gas.

In some ammonia process schemes operating without a secondary reformer and applying pressure swing adsorption (PSA) for further purification (KTI PARC) only a HTS is used.

The LTS catalyst, supplied in pellets like the HTS catalyst, consists of 40 – 55 % copper oxide, 20 – 30 % zinc oxide, the balance being alumina. The catalyst properties are influenced far more by the formulation and manufacturing procedure [547] than by its chemical composition. The copper oxide is reduced in situ with hydrogen and a carrier gas (usually nitrogen) to form fine copper crystallites on which the activity depends. Sulfur, usually present as H_2S, has to be below 0.1 ppm, but even with such low concentrations, the catalyst is slowly poisoned. The ZnO adsorbs the sulfur and it finally transforms into bulk ZnS. When the ZnO is exhausted in a given layer of the catalyst, the H_2S causes deactivation of the copper by sintering. The poisoning process moves through the catalyst as a relatively sharp front and can be seen in the change of the catalyst temperature profile over time [548], [549]. The LTS catalyst is protected by a guard bed, formerly loaded with ZnO, but nowadays usually with LTS catalyst [558]. Changing the guard bed more frequently prolongs the service life of the main LTS catalyst bed. Traces of chlorine compounds [557], which may be introduced with the natural gas or more often with the process air to the secondary reformer, may also deactivate the LTS catalyst by accelerating the sintering of the copper particles. Unlike sulfur poisoning, chlorine is more diffusely distributed over the whole catalyst bed by migration as volatile zinc and copper chlorides.

As chemical composition and formulation of the LTS catalyst are very similar to methanol production catalysts, small quantities of methanol are formed and are found in the process condensate after cooling the LTS effluent. In a consecutive reaction, amines (mainly methylamine) are formed from the methanol and traces of ammonia originating from the secondary reformer and the HTS. These pollutants are removed from the process condensate by steam stripping and ion exchange. Byproduct formation is higher with fresh catalyst and declines with operating time. New catalyst types with increased activity and higher selectivity have reduced the problem. The tendency for methanol formation increases with decreasing steam/gas ratio [552].

A relatively new process concept is the intermediate temperature shift (ITS) [550] which performs the shift conversion in a single step. The catalyst is based on a copper – zinc – alumina formulation and optimized for operating in a wider temperature range than the standard LTS catalyst. The reaction heat can be removed by use of a tube-cooled reactor raising steam or heating water for gas saturation to supply process steam in the reforming section (Linde LAC, ICI Catalco LCA). In a new plant using the spiral-wound Linde reactor [551], a methane slip of only 0.7 mol % (dry basis) is achieved. Further purification is performed by PSA. Generally the shift conversion reactors have an axial gas flow pattern, but recently radial gas flow configurations have been chosen in some instances.

Additional literature on shift conversion can be found in [406-408], [418], [419], [536], [538-543], [404], [1318].

Shift Conversion in Partial Oxidation Plants. The raw synthesis gases from partial oxidation of heavy hydrocarbons and coal differ mainly in two aspects from that produced from light hydrocarbons by steam reforming. First, depending on the feedstock composition, the gas may contain a rather high amount of sulfur compounds (mainly H_2S with smaller quantities of COS); second, the CO content is much higher, in some cases in excess of 50 %. The sulfur compounds (see Section Gas Purification) can be removed ahead of the shift conversion to give a sulfur-free gas suitable for the classical iron HTS catalyst. In another process variant the sulfur compounds are removed after shift conversion, which thus has to deal with a high-sulfur gas. As the standard iron catalyst can tolerate only a limited amount of sulfur compounds, the so-called dirty shift catalyst is used in this case. This cobalt – molybdenum – alumina catalyst [417], [418], [554-556], [638] is present under reaction conditions in sulfidized form and requires for its performance a sulfur content in the gas in excess of 1 g S/m^3. Reaction temperatures are between 230 and 500 °C.

Irrespective of the catalyst type used, the high initial carbon monoxide concentration means that the reaction must generally be performed in steps, with intermediate cooling. It has been reported that the CO content can be reduced from 50 to 0.8 % in a single step in a large hydrogen plant by using a quasi-isothermal reactor (e.g., the Linde spiral-wound reactor).

4.5.1.3. *Gas Purification*

In further purification, carbon dioxide, residual carbon monoxide, and sulfur compounds (only present in the synthesis gas from partial oxidation) have to be removed as they are not only a useless ballast but above all poisons for the ammonia synthesis catalyst.

The total sulfur contained in the coal and hydrocarbon feedstock is converted in gasification to H_2S and a smaller amount of COS, which are removed as described below. In contrast steam reforming requires removal of sulfur from the natural gas and light hydrocarbon feedstocks upstream of gasification to avoid poisoning of the sensitive reforming catalysts. This is usually performed by hydrodesulfurization and adsorption of the H_2S by ZnO. As this is an essential part of the steam reforming process it is treated in Section Feedstock Pretreatment and Raw Gas Production.

The classical method for CO_2 removal is to scrub the CO_2 containing synthesis gas under pressure with a solvent capable of dissolving carbon dioxide in sufficient quantity and at sufficient rate, usually in countercurrent in a column equipped with trays or packings. The CO_2-laden solvent is flashed, often in steps, to around atmospheric pressure, and the spent scrubbing liquid is subsequently heated and regenerated in a stripping column before being recycled to the pressurized absorption column. In the early days of ammonia production water, often river water, served as solvent in a once-through process without regeneration and recycling.

Today a variety of solvents are used and they can be categorized as physical or chemical solvents. In the physical solvents the carbon dioxide dissolves without forming a chemical compound, and this allows recovery simply by flashing. In the chemical solvents the carbon dioxide becomes fixed as a chemical compound, which needs heat for its decomposition. At low carbon dioxide partial pressures, the chemical solvents absorb substantially more carbon dioxide than the physical solvents; at higher partial pressures the physical solvents (according to Henry's Law, the loading is approximately proportional to the CO_2 partial pressure) have a higher loading capacity than the chemical solvents, for which the solubility approaches a saturation value. Figure 31 shows the loading characteristic for various solvents.

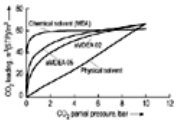

Figure 31. CO_2 loading characteristics of various solvents

As both sour gases, CO_2 and H_2S, have good solubility in the applied solvents, special process configurations are required for partial oxidation gases to recover separately a pure CO_2 fraction and an H_2S-rich fraction suitable for sulfur disposal.

Very suitable for the partial pressure range of CO_2 in steam reforming plants (4 – 7 bar) are chemical solvents based on aqueous solutions of potassium carbonate or alkanolamines containing additional activators to enhance mass transfer and, in some cases, inhibitors to limit or prevent corrosion processes. Primary and secondary amines, for example, monoethanolamine (MEA) and diethanolamine (DEA) exhibit a high mass-transfer rate for carbon dioxide but have a high energy demand for regeneration. For this reason tertiary amines are commonly used today, for example, methyldiethanolamine together with an activator. Triethanolamine does not achieve the required final CO_2 concentration, and in the few cases where it was used it was followed by additional scrubbing with monoethanolamine (MEA).

The potassium carbonate processes from the various licensors differ with respect to the activator, corrosion inhibitor and to some extent in the process configuration.

Over the years considerable progress has been achieved in improving the efficiency of the carbon dioxide removal systems. The first generation of single-train steam reforming ammonia plants used MEA and consumed about 5.8 GJ per tonne NH_3, which was about 14 % of the total energy consumption. Table 23 demonstrates the progress made in energy consumption.

Table 23. Heat requirements for regeneration in CO_2 removal systems [406], [603]

Process	Heat requirement, kJ/mol CO_2
MEA (without Amine Guard)	209
MEA (with Amine Guard III)	116
Benfield hot potash (1-stage)[a]	88
BASF aMDEA process (1-stage)[a]	73
Benfield LoHeat	28 – 35
BASF aMDEA process (2-stage)[b]	28 – 30

[a] Single-stage regeneration.
[b] Two-stage regeneration.

The first progress was made by addition of corrosion inhibitors (e.g., Amine Guard introduced by Union Carbide) [603], [634-637], which allow a higher loading of the solvent. The energy consumption of the modern systems, as shown in the last two lines of Table 23 depends largely on process configuration and required final purity which may range from 50 to 1000 ppm in the purified gas. There is also a trade-off between energy consumption and investment costs.

The hot potash systems, for example, the Benfield process licensed by UOP, differ in the type of activator used to increase the reaction rate between the CO_2 and the solvent. The activators enhance mass transfer and thus influence not only the regeneration energy demand (circulation rate of the solvent) but also the equipment dimensions. The following activators are used: in Benfield [567-570] and Carsol [571], [572] process ethanolamines; in the Giammarco Vetrocoke process glycine (originally arsenic oxide) [575-578]; in the Catacarb system [565], [566] amine borate; in the Exxon Flexsorb HP process [573], [574] a sterically hindered amine; in the Carbosolvan process [580], [584] sulfosolvan B. New activators named ACT-1 (UOP) [585], [586] and LRS-10 [587], [588] (British Gas) have been introduced. All hot potash systems need corrosion inhibitors the concentration of which must carefully to be monitored.

A typical example of an efficient modern hot potash system is the Benfield LoHeat [589] system of UOP. The high thermal efficiency of this two-stage adsorption process using lean and semi-lean solvent is achieved by recompression of the flash steam with an injector or a mechanical vapor compressor. UOP has also developed a number of other process configurations [573], [584], [590], [591], [1280].

BASF's aMDEA [592-594], [595-598] process uses an aqueous solution of the tertiary amine methyldiethanolamine (MDEA) with a special amine activator. No corrosion inhibitors are necessary, and unlike MEA no solvent degradation is observed, so recovery installations are not required. Owing to the low vapor pressure of MDEA and the activator, there are no losses of the active solvent components. The carbon dioxide binds much less strongly to MDEA than to MEA, and the solvent character is more like a hybrid between a strong chemical and a purely physical solvent. On account of the relative weak binding forces, a substantial amount of carbon dioxide can be recovered simply by flashing to low pressure, and only a small amount has to be recovered by stripping. The process is very versatile: increasing the activator concentration shifts the character of the solvent more to the chemical side and vice versa, and process configurations include two- and single-stage designs, the latter allowing old MEA units to be revamped [599-601] simply by swapping the solvent without changing the process equipment.

Physical solvent based processes can also be applied in steam reforming ammonia plants. The Selexol process (UOP) [565], [605-613] uses polyethylene glycol dimethyl ether, the Sepasolv MPE process (BASF) [565], [593]polyethylene glycol methyl isopropyl ether, and the Fluor Solvent Process [614-617] polypropylene carbonate. These solvents are stable, noncorrosive, nontoxic, and not very volatile, but they have a rather high capacity for adsorbing water. For this reason the raw gas has to be dry, which is usually achieved by chilling and operating at low temperatures. The Selexol Process for example operates at 5 °C. The carbon dioxide is recovered simply by flashing; no heat is necessary. Before the solvent is recycled it has to be vacuum treated and/or air stripped to remove traces of carbon dioxide. When comparing the energy consumption of physical and chemical solvent based processes for a given application not only the heat for solvent regeneration but also the mechanical energy for solvent circulation has to be taken into account.

Raw synthesis gas produced by partial oxidation has a carbon dioxide partial pressure between 10 and 30 bar, depending mainly on feedstock and gasification pressure. In the lower half of this partial pressure range chemical solvents based on tertiary amines (e.g., MDEA) might be suitable, but at higher values physical solvents become increasingly preferable. Examples of physical solvent processes are Selexol Process, Sepasolv MPE process, Lurgi Purisol Process (*N*-methylpyrrolidone) [581-583], Sulfinol process (tetrahydrothiophene-1,1-dioxide + diisopropanol) [579], [604], and, probably the most important in this connection, the Rectisol process (Linde and Lurgi) [565], [618-622], [1396], with methanol as solvent operating at –15 to – 40 °C.

The presence of sulfur compounds, mainly H_2S along with a minor quantity of COS, introduces some complications for CO_2 removal from partial oxidation gases. Since the sour gases are both soluble in the solvents and a separate recovery in the regeneration stage is only partially possible (only pure CO_2 can be obtained along with a CO_2 fraction more or less rich in H_2S) two principal plant flowsheets are possible. In one, a first scrubbing stage, which removes the sulfur compounds, is positioned upstream of the shift conversion. The second stage, removing pure CO_2, is located downstream of the shift reactors. This concept reduces the effort required to receive a highly concentrated H_2S fraction suitable for further processing in a Claus plant or a sulfuric acid plant. In the other arrangement, the sour gases are removed after shift conversion, which in this case has to operate over dirty shift catalyst (see Section Carbon Monoxide Shift Conversion) with the consequence that the H_2S is diluted by a large amount of CO_2. In this case, too, methanol regeneration in the Rectisol process can be tailored to achieve a H_2S-rich fraction for downstream processing and pure CO_2, a task difficult to fulfil with the other physical solvents mentioned.

Extensive surveys and additional literature of the CO_2 removal processes are found in [564], [602], [603], [623], [624-627], [628-633], [1332], [1333], [1396] a theoretical analysis is given in [562].

Final Purification. After bulk removal of the carbon oxides has been accomplished by shift reaction and CO_2 removal, the typical synthesis gas still contains 0.2 – 0.5 vol % CO and 0.005 – 0.2 vol % CO_2. These compounds and any water present have to be removed down to a very low ppm level, as all oxygen-containing substances are poisons for the ammonia synthesis catalyst [640].

Methanation. Methanation is the simplest method to reduce the concentrations of the carbon oxides well below 10 ppm and is widely used in steam reforming plants. It is actually the reverse reaction of steam reforming of methane:

$$CO + 3\,H_2 \rightleftharpoons CH_4 + H_2O \quad \Delta H = -206\,\mathrm{kJ/mol} \tag{16}$$

and

$$CO_2 + 4\,H_2 \rightleftharpoons CH_4 + 2\,H_2O \quad \Delta H = -165\,\mathrm{kJ/mol} \tag{17}$$

The advantages of simplicity and low cost more than outweigh the disadvantages of hydrogen consumption and production of additional inerts in the make-up gas to the synthesis loop.

The reaction is carried out over a supported nickel catalyst at a pressure of 25 – 35 bar and a temperature of 250 – 350 °C. The required catalyst volume is relatively small. If a breakthrough of carbon monoxide from the low-temperature shift or carbon dioxide from the absorption system occurs, the intensely exothermic methanation reaction can reach temperatures exceeding 500 °C very quickly [641]. For example, 1 % CO_2 breakthrough leads to an adiabatic temperature rise of 60 °C. Controls should be installed and other security measures taken to avoid these high temperatures because the catalyst may be damaged or the maximum allowable operating temperature of the pressure vessel wall exceeded.

Methanation as final purification for the raw gas from partial oxidation was proposed by Topsøe [642]. In this case the shift conversion is carried out in two stages with a special sulfur-tolerant shift catalyst followed by removal of hydrogen sulfide and carbon dioxide in an acid gas removal unit. Because of the potential danger of a sulfur break-through causing poisoning, the normal copper – zinc – alumina catalyst is usually not applied, which is surprising as the same risk exists in partial oxidation based methanol plants for the similarly composed methanol catalyst.

Selectoxo Process. The Selectoxo process (Engelhard) reduces the hydrogen consumption of the methanation system, as well as the inert gas content of the purified synthesis gas fed to the synthesis loop. After low-temperature shift conversion, the cooled raw gas is mixed with the stoichiometric quantity of air or oxygen needed to convert the carbon monoxide to carbon dioxide. The mixture is then passed through a precious-metal catalyst at 40 – 135 °C to accomplish this selective oxidation [643-646]. The carbon dioxide formed by the Selectoxo reaction adds only slightly to the load on the downstream carbon dioxide absorption system.

Methanolation [647], [648] has been proposed for partially replacing methanation. It converts the residual carbon oxides to methanol, preferably at higher pressure in an intermediate stage of synthesis gas compression. Methanol is removed from the gas by water scrubbing. The methanol may be recycled to the steam reformer feed or recovered as product. As full conversion of the carbon oxides is not achieved, a clean up methanation unit must follow the methanolation section.

Dryers. It is energetically advantageous to add the purified synthesis gas at a point in the synthesis loop where it can flow directly to the synthesis converter (see Section Synthesis Loop Configurations). For this reason water and traces of carbon dioxide must be removed from the make-up gas downstream of methanation. This is accomplished by passing the make-up gas through molecular sieve adsorbers.

Cryogenic methods are usually used for final purification of partial oxidation gases, but may be also incorporated in steam reforming plants. A prominent example is the Braun Purifier process [457], [649-655].

The purifier is a cryogenic unit placed downstream of the methanator and its duty is to remove the nitrogen surplus introduced by the excess of air used in the secondary reformer of the Braun ammonia process (Processes with Reduced Primary Reformer Firing). Additionally the inert level in the synthesis loop is reduced through this unit because methane is completely and argon is partially removed from the make-up gas. Another advantage of the process is that it separates the front-end and the synthesis loop, permitting the H/N ratio in the synthesis loop to be set independent of the secondary reformer. The purifier is a relatively simple unit composed of feed/effluent exchanger, a rectifier column with an integrated condenser and turbo-expander. At –185 °C methane and argon are washed out. The cooling energy is supplied by expansion of the raw gas over the turbo-expander (pressure loss about 2 bar) and expanding the removed waste gas to the pressure level of the reformer fuel.

Liquid Nitrogen Wash [656-658]. Normally, for the partial oxidation processes, only a high-temperature shift conversion is used. This results in a carbon monoxide content of the gas after shift conversion in the range 3 – 5 vol %. Copper liquor scrubbing for carbon monoxide removal, commonly employed in early plants, has become obsolete and is now operated in only a few installations. Not only does it have a high energy demand, but it is also environmentally undesirable because of copper-containing wastewater. Liquid nitrogen wash delivers a gas to the synthesis loop that is free of all impurities, including inert gases and is also the means for adding some or all of the nitrogen required for synthesis.

The nitrogen is liquefied in a refrigeration cycle by compression, cooling, and expansion. It flows to the top of a wash column, where it countercurrently contacts precooled synthesis gas from which most of the methane and hydrocarbons have been condensed. All of the cold equipment is installed in an insulated "cold box." The wash column temperature is about –190 °C. Liquid nitrogen wash systems are in operation at pressures up to 8 MPa corresponding to the highest gasification pressures. Careful surveillance of the inlet gases is required. Water and carbon dioxide in the inlet gas will freeze, causing operating difficulties. Traces of nitric oxide (NO) may react with olefinic hydrocarbons, causing explosions [659], [660].

Normally, an air separation plant is installed in conjunction with liquid nitrogen wash for economy in operation. In modern plants, the air separation and the nitrogen wash frequently are closely integrated with one another so that economies can be realized in the refrigeration system.

Pressure Swing Adsorption. This process can be used to replace the LT shift conversion, carbon dioxide removal, methanation, and even the secondary reformer as well [406], [627], [661-664]. It uses molecular sieves as adsorbents in a series of vessels operated in a staggered cyclic mode changing between an adsorption phase and various stages of regeneration. The regeneration of the loaded adsorbent is achieved by stepwise depressurization and by using the gas from this operation to flush other adsorbers at a different pressure level in the regeneration cycle. The hydrogen recovery may be as high as 90 % depending on the number of adsorbers in one line, which may be as high as 10. Very high purity can be obtained, with about 50 ppm argon and less than 10 ppm of other impurities.

The process scheme for the ammonia plant may consist of production of pure hydrogen followed by separate addition of pure nitrogen from an air separation unit [665-667]. In a special version the nitrogen can be added in the PSA unit itself to increase hydrogen recovery [668-671]. In some processes it may also remove the excess nitrogen introduced with the process air fed to the secondary reformer, e.g. the LCA Process of ICI (ICI LCA Process.). Since this technology has proven its reliability in rather large hydrogen plants for refineries it is now also used for world-scale ammonia plants , e.g., the Linde LAC process (Linde LAC Process.).

4.5.2. *Compression*

Up to the mid-1960s reciprocating compressors were used to compress the synthesis gas to the level of the synthesis loop, which was around 300 bar in the majority of the plants at that time. Higher pressures were used in a few installations, for example, in Claude and Casale units. Prior to about 1950 gas generation processes and shift conversion operated at essentially atmospheric pressure. The gas was first compressed to the level of the CO_2 removal section (usually 25 bar) and afterwards to around 300 bar for final purification (at that time usually copper liquor scrubbing) and synthesis. Reciprocating compressors with as many as seven stages in linear arrangement with intermediate cooling were used, whereby the CO_2 removal section was usually installed between the 3rd and 4th stages. Machines with a suction volume up to 15 000 m^3 (STP) for the first stage were not uncommon. Huge flywheels were designed as the rotors of synchronous motors (ca. 125 rpm) with two crankshafts on both sides connected over crossheads with the piston rod for the horizontally arranged stages. In some instances gas engines were used as drivers.

The rapid technical progress in the hydrocarbon based technologies of steam reforming and partial oxidation made it possible to generate the synthesis gas at a pressure level sufficient for the CO_2 removal operation. As gasification proceeds with a considerable volume increase and feedstocks such as natural gas are usually already available under pressure at battery limits, considerable savings in compression energy are achieved in this way.

Along with the introduction of pressure gasification, horizontally balanced compressors in which the cylinders are in parallel configuration on both sides of a common crankshaft became the preferred design. In these machines a good dynamic balance can readily be achieved, higher speeds are possible and also the use of asynchronous motors is possible. The low height of the arrangement has less severe requirements for foundations, allows simpler piping connections and facilitates maintenance. When gas engine drivers (two-stroke type) were used instead of electric motors, some designs applied a common crankshaft for the piston rods of the gas machine cylinders and compressor cylinders. In a very few cases steam turbines with special speed reduction gears have been used. In smaller plants, the various compression services, e.g., natural gas, process air, and synthesis gas compression, were apportioned among the crankshaft throws in such a manner that a single compressor can perform all compression duties [672]. Further information on reciprocating compressors is given in [673], [674], [677], [678].

One of the most important features of the energy integrated single-stream steam reforming ammonia plant pioneered by M. W. Kellogg in 1963 was the use of centrifugal compressors for the compression duties of synthesis gas and recycle, process air, and refrigeration. From this time onwards application of centrifugal compressors became standard practice in most ammonia plants irrespective of the synthesis gas generation technology. The fundamental advantage of these machines are low investment (single machines even for very large capacities) and maintenance cost, less frequent shutdowns for preventive maintenance, and high reliability (low failure rate) [958]. In most cases the centrifugal compressors in ammonia plants are directly driven by steam turbines. This avoids the losses associated with generation and transmission of electric power. For this reason the overall efficiency of a plant with steam-driven centrifugal compressors is superior, although the centrifugal compressors are inherently less efficient than reciprocating units. A further advantage is that centrifugal compressors require only a fraction of the space needed for reciprocating compressors.

Manufacturing capabilities limit the minimum possible passage width (today about 2.8 mm) at the outer circumference of a centrifugal compressor impeller and this imposes a limit on the minimum effective gas volume leaving the last impeller. Unless the total volumetric gas flow has a reasonable relationship to the passage width of the last impeller and the pressure ratio, excessive pressure losses would occur within the passage and in the diffusers between the impellers, rendering the machine extremely ineffective. The first single-train ammonia plants with a capacity of 550 – 600 t/d had to lower the synthesis pressure to 145 – 150 bar to meet the required minimum gas flow condition. Today, with improved manufacturing techniques, the minimum gas flow from the last wheel is 350 m^3 for synthesis gas with a molecular mass of about 9 and an efficiency of around 75 %. This corresponds to a capacity of 400 t/d at 145 bar. As newer synthesis catalysts allow a pressure of 80 bar in the synthesis loop (ICI's LCA Process, Kellogg's KAAP) a centrifugal compressor could be used down to 220 t/d. Of course, for today's world-scale capacities of 1200 – 2000 t/d these technical limitations have no influence on the synthesis pressure, which even for plants with 1800 t/d is between 155 and 190 bar [404], [406], [675], [676], [679].

The tensile strength of the steels normally used to manufacture the compressor impellers allow a maximum wheel tip speed of about 330 m/s, which limits the pressure increase attainable by each impeller. A pressure increase, for example, from 25 to 200 bar would require 18 – 20 impellers. However, a compressor shaft must have sufficient rigidity to avoid excessive vibration, and this limits the possible length such that a compressor shaft cannot accommodate more than eight or nine impellers. It is therefore necessary to arrange several compressor casings in series, with compression ratios from 1.8 to 3.2.

To overcome the pressure drop (5 – 20 bar) in the synthesis loop re-compression of the recycle gas is required. In practically all modern ammonia plants, the shaft of the final casing also bears the impeller for the compression of the recycle gas. Depending on synthesis configuration, mixing of make-up gas and recycle can be performed inside the casing or outside (three or four-nozzle arrangement; Fig. 34).

In older plants which used a reciprocating compressor for the recycle, a recycle cylinder was often mounted together with the other cylinders on the reciprocating frame. Sometimes special rotary compressors, so-called mole pumps were also used, with the unique feature that compressor and electric driver were completely enclosed in a common high-pressure shell. In old Casale plants, the make-up gas was introduced into the high pressure recycle loop and acted as the driving fluid of an injector which compressed the recycle gas.

Today modern plant concepts for world-scale capacity plants tend to limit the number of compressor casings to two. Figure 32 shows an example of the synthesis gas compressor of a large ammonia plant.

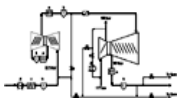

Figure 32. Centrifugal compressor for make-up and recycle gas compression of an ammonia plant (courtesy of Uhde)

a) Air cooler; b) Separator; c) Silencer; d) Water cooler

Geared or metal diaphragm couplings are used to connect the shafts of the individual casings (two in Figure 32). These flexible couplings prevent possible compressor damage resulting from slight misalignment and shaft displacement.

Sealing of the rotating shaft against the atmosphere is an important and demanding task. The high pressures and the high rotational speeds involved do not allow mechanical contact shaft seals. Usually, liquid-film shaft seals with cylindrical bushings (floating rings) are applied [680]. In this concept, an oil film between the shaft and a floating ring, capable of rotation, provides the actual sealing. The floating ring is usually sealed to the compressor casing by O-rings. Seal oil flows between both halves of the floating ring. Part of the oil returns to the reservoir, while the remainder flows against the gas pressure into a small chamber from which, together with a small quantity of gas, it is withdrawn through a reduction valve. Figure 33 is a schematic diagram of a liquid film shaft seal.

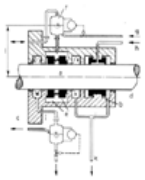

Figure 33. Liquid film shaft seal with cylindrical bushing for a high-pressure centrifugal compressor

a) Shaft; b) Bearing; c) Pressure side; d) Ambient side; e) Floating seal rings; f) Seal oil surge vessel; g) Seal oil; h) Lube oil; i) Drain to gas – oil separator; k) Drain to lube oil tank; l) Elevation for seal oil head

The seal oil pressure in the floating ring cavity must always be slightly higher than the gas pressure within the casing which is provided by static height difference of the oil level in the elevated oil buffer vessel. In this way normally no oil should enter into the synthesis gas. As the labyrinth section on the pressure side is connected with the suction side through an equalizing line, it is necessary to seal the compressor shaft towards the atmosphere only at suction pressure level.

Often, seal oil supply is combined with the lubricating oil system, with oil reservoir, filters and (in part) pumps in common.

The high rotational speeds and the relatively large masses of the compressor rotors place high demands on the performance of the bearings. This is especially true for the thrust bearings, which must withstand high thrust forces. The minimum clearances necessary at the labyrinths and the impellers allow practically zero wear. Any wear resulting in friction could lead instantly to severe damage. For these reasons, in addition to measuring bearing temperatures, the axial position of the rotor and the radial vibrational deflections are continuously monitored by sensors. An increased vibrational amplitude is often an early indication of mechanical faults, such as rotor imbalance, bearing damage, occurrence of friction, or misalignment.

An interesting concept is dry (oil-less) gas seals for centrifugal compressors. Development dates back to 1969 and the first commercial application was in a natural-gas compressor in 1973. Since then they have been widely used in off-shore service [681-683]. Only recently have they gained acceptance in the ammonia industry, where several compressors for synthesis gas and refrigeration duty equipped with dry seals have been successfully placed in service. Nitrogen is used as an inert fluid for the seal, which is achieved at the radial interface of rotating and stationary rings. During operation the seal is not completely tight; some of the seal gas flows back to the suction side to be re-compressed, and a small amount from the suction side may go to the atmospheric side and is sent to the flare on account of its content of combustibles. During stops, when the shaft is not rotating, the seal ring is pressed tight against the seat by means of a spring and the differential gas pressure. Dry gas seals in combination with oil-lubricated bearings (dry/wet) have the advantage that a much smaller oil system is required and that there is no contact between oil and gas, which eliminates an emission source.

A new development, already in commercial service, but so far not used in ammonia plants, is magnetic bearings [681], [684]. Magnetic bearings promise a wider temperature range, are less prone to wear, are less prone to developing vibrations due to imbalance and require less maintenance. A combination of magnetic bearings and dry seals (dry/dry) could totally eliminate the oil system.

Integrated geared centrifugal compressors developed by DEMAG and GHH are a new development which might become of interest for plants of smaller capacity operating at lower pressure. The driver (e.g., a steam turbine) drives a common gear to which the individual compression stages are connected. Each stage has a single impeller which runs with very high speed, for example, 25 000 rpm or higher. Compressors with three or four stages are in operation, for example, a methanol synthesis compressor for 75 000 m^3/h and a pressure of 75 bar. It is likely that this concept can be extended to a final pressure of 120 bar with ammonia synthesis gas.

Compressor control is achieved basically by controlling the rotational speed of the driver, in modern plants often with the help of a distributed control system (DCS). If the volumetric flows through the machine at start-up or during reduced load operation deviate too far from the design values, it is necessary to re-circulate gas through individual stages or through the whole machine. Otherwise the compressor can enter a state of pulsating flow, called surge, which could cause damage. Anti-surge control (minimum by-pass control, kickback control) is designed to prevent this condition, as well as to minimize the incidence and degree of uneconomical re-circulation. A point of discussion is sometimes the minimum load at which the compressor can run without kickback, which means without loss of efficiency. Usually a load of 75 % is possible, and with special impeller design, the value may be lowered to 70 % but with a slight sacrifice of full-load efficiency.

Other compression duties in the plants, such as process air in steam reforming plants and air, and oxygen, and nitrogen compression in partial oxidation plants, are also performed by centrifugal compressors. Also for the ammonia compression in the refrigeration section centrifugal compressors are normally in service. In some cases screw compressors have been used for this duty on account of their good efficiency and load flexibility, which is of interest in plants where the ammonia product is split between direct users at the site and cold storage in changing ratios.

Criteria for compressor selection and economic comparisons are discussed in [685-692, 704]. Additional information is given in [693-703].

In modern plants the synthesis gas compressors, including recycle, are almost exclusively driven by a steam turbines. These are generally extraction turbines with a condensing section. Steam is extracted at suitable pressure levels (e.g. 45 – 55 bar) to provide, for example, the process steam in steam reforming plants, and for other drivers, e.g., air compressor, ammonia compressor, boiler feed water pumps, and blowers.

As failures and breakdowns of these large rotary machines could lead to long and expensive repairs and to a corresponding loss of production it is advisable to keep the essential spare parts, for example, spare rotors, in stock. In older steam turbines, sometimes blade failures occurred, but this is no longer a problem due to improved blade design and shroud bands, which are standard today.

Gas turbines have also been used as drivers for compressors in ammonia plants. The exhaust may be used for steam production, for preheating duties, or as combustion air in the primary reformer [705-709].

4.5.3. *Ammonia Synthesis*

Under the conditions practical for an industrial process ammonia formation from hydrogen and nitrogen

$$N_2 + 3\,H_2 \rightleftharpoons 2\,NH_3 \quad \Delta H = -92.44\ \mathrm{kJ/mol}$$

is limited by the unfavorable position of the thermodynamic equilibrium, so that only partial conversion of the synthesis gas (25 – 35 %) can be attained on its passage through the catalyst. Ammonia is separated from the unreacted gas by condensation, which requires relatively low temperatures for reasonable efficiency. The unconverted gas is supplemented with fresh synthesis gas and recycled to the converter. The concentration of the inert gases (methane and argon) in the synthesis loop is controlled by withdrawing a small continuous purge gas stream. These basic features together with the properties of the synthesis catalyst and mechanical restrictions govern the design of the ammonia synthesis converter and the layout of the synthesis loop. Evaluation criteria are energy consumption, investment and reliability.

4.5.3.1. *Synthesis Loop Configurations*

A number of different configurations are possible for the synthesis loop. They can be classified according to the location of ammonia condensation and the point at which the make-up gas is introduced. Figure 34 shows the principal possibilities.

If the make-up gas is absolutely free of catalyst poisons, such as water and carbon dioxide (for example, after molecular sieve dehydration or liquid nitrogen wash), it can be fed directly to the synthesis converter (Fig. 34 A). After the gas leaves the synthesis converter, ammonia is condensed by cooling and the recycle gas is referred to the recycle compressor. This represents the most favorable arrangement from a minimum energy point of view. It results in the lowest ammonia content at the entrance to the converter and the highest ammonia concentration for condensation.

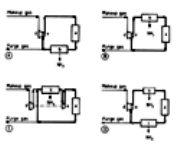

Figure 34. Schematic flow diagrams of typical ammonia synthesis loops

A) Synthesis loop for pure and dry make-up gas; B) Product recovery after recycle compression; C) Product recovery before recycle compression (four-nozzle compressor design); D) Two stages of product condensation

a) Ammonia converter with heat exchangers; b) Ammonia recovery by chilling and condensation; c) Ammonia recovery by condensation at ambient temperature; d) Synthesis gas compressor; e) Recycle compressor

When the make-up gas contains water or carbon dioxide, advantage is taken of the fact that these materials are absorbed completely by condensing ammonia. This requires that the condensation stage be located partially or wholly between the make-up gas supply point and the converter. This arrangement has the disadvantage that the ammonia concentration for condensation is reduced by dilution with the make-up gas. Also, at equal condensing temperature, a higher ammonia concentration exists at the inlet to the converter. Figure 34 B shows the simplest such configuration. An additional drawback of this arrangement is that all the ammonia produced must be compressed with the recycle gas in the recycle compressor.

The scheme shown in Figure 34 C, the frequently used “four-nozzle compressor”, avoids this waste of energy. With this arrangement, recycle compression follows directly after condensing and separating the ammonia. In this configuration, it is possible to cool the recycle gas using cooling water or air immediately before admixing the make-up gas (i.e., before diluting the recycle gas) and thereby to reduce the energy expenditure for refrigerated cooling.

Splitting the cooling step for ammonia condensation also offers advantages when the recycle gas is compressed together with the make-up gas. This applies especially at synthesis pressures above about 25 MPa (250 bar). At these pressures, a greater portion of the ammonia formed can be liquefied by cooling with cooling water or air (see Fig. 34D).

When ammonia-containing recycle gas and carbon dioxide containing make-up gas mix together under certain conditions of concentration and temperature, precipitation of solid ammonium carbamate can result.

In recent years, also as a retrofit in existing plants, molecular sieve drying of make-up gas has increasingly been applied in order to realize the energy-saving arrangement of the synthesis loop corresponding to Figure 34 A.

4.5.3.2. *Formation of Ammonia in the Converter*

The central part of the synthesis system is the converter, in which the conversion of synthesis gas to ammonia takes place. Converter performance is determined by the reaction rate, which depends on the operating variables (cf. Section Kinetics). The effect of these parameters is discussed briefly in the following.

With increasing pressure, ammonia formation increases (Fig. 35). This results not only from the more favorable equilibrium situation for the reaction, but also from the effect on the reaction rate itself. In industrial practice, there are plants that operate at about 8 MPa (80 bar), but there are also those that operate at more than 40 MPa (400 bar). Today, plants are built mainly for synthesis pressures of 150 – 250 bar. Typical operating parameters for modern synthesis loops with different pressures are listed in Table 24.

Table 24. Typical operating parameters for modern synthesis loops at 140 and 220 bar (1000 t/d NH_3) [404, p. 226]

Parameters	Inlet pressure, bar	
	140	220
Inlet flow, Nm^3/h	500 000	407 000
Inlet NH_3 conc., mol %	4.1	3.8
Outlet NH_3 conc., mol %	17.1	19.9
Inlet inert conc., mol %	8.0	12.0
NH_3 separator temperature, °C	–5	–5
Relative catalyst volume	1	0.6

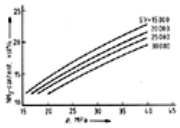

Figure 35. Performance for a four-bed quench converter as a function of operating pressure with space velocity (per hour) as parameter; 10 % inerts in the inlet synthesis gas

Converter performance decreases with increasing inert gas content (Fig. 36). The usual range is 0 – 15 vol %. For a secondary loop based on purge gas, it can be 30 % or more (see Section Inert-Gas and Purge-Gas Management).

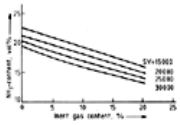

Figure 36. Performance of a converter as a function of inlet inert gas (CH_4 and Ar) content with space velocity (per hour) as parameter, inlet NH_3 content is 3.5 %; 30 MPa pressure; catalyst particle size is 6 – 10 mm

Converter performance also diminishes (Fig. 37) with increasing oxygen content of the synthesis gas. Today, a level of 10 ppm in the make-up gas, corresponding to about 3 ppm in the converter inlet gas, is usually, not exceeded (cf. Section Catalyst Poisons).

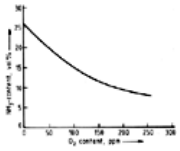

Figure 37. Performance of a converter as a function of oxygen content (all oxygen-containing impurities) in the inlet synthesis gas

In contrast to the above-mentioned variables, the dependence of the converter performance on the H_2/N_2 ratio shows a true maximum (Fig. 38). The optimum conversion at high space velocity [SV = m^3 (STP) gas $h^{-1} \cdot m^{-3}$ catalyst] lies close to an H_2/N_2 ratio of 2 and approaches 3 at low space velocities. The reason is that equilibrium plays a greater role at low space velocities and has a maximum at a ratio of 3, except for small corrections [14] with regard to the behavior of real gases. Usually, the ratio is adjusted to 3, because in most plants, conversions near equilibrium are attained.

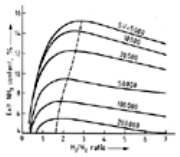

Figure 38. Ammonia conversion as a function of hydrogen/nitrogen ratios in the inlet synthesis gas with space velocity (per hour) as parameter; 9.7 MPa pressure [1136]

In practice, space velocities vary from about 12 000 h^{-1} at about 15 MPa (150 bar) to about 35 000 h^{-1} at about 80 MPa (800 bar). Usually, with increasing space

velocity, the ammonia concentration in the effluent synthesis gas from a given converter does indeed go down (Figs. 35, 36, 38). However, the operating point normally chosen means that the increase in gas flow rate more than compensates for the reduced ammonia concentration. Thus, a still higher ammonia production rate is achieved. Plant operation often takes advantage of these phenomena. For example, this characteristic can be used to maintain ammonia production rate when the synthesis catalyst ages and its activity declines. Increasing converter flow rate and declining synthesis catalyst activity can reach a point, even with careful control, where the reaction "blows out" and production ceases. This occurs when the heat of reaction is no longer sufficient to provide the temperatures necessary for operation of the feed – effluent heat exchanger. The heat exchanger then fails to heat the cold converter feed gas to the required reaction "ignition" temperature. Achieving maximum ammonia production requires operating in the neighborhood of this "blow out" point, in turn requiring very careful control [710-712]. If the synthesis converter is to be operated in this region, then it is advisable to oversize the converter feed – effluent heat exchanger system to attain a higher degree of control stability.

Converter Design. Design of ammonia synthesis reactors is not just the calculation of the required catalyst volume; other parameters have to be considered, too, and for some of them optimum values have to be found. This raises the question of the definition of optimum. In the early days with more strict material and fabrication-related limitations the converters were usually designed for minimum high-pressure volume, and this meant maximum use of the catalyst. Today the objective is to optimize the heat recovery (at the highest possible level) and to minimize the investment for the total synthesis loop.

The design of an ammonia converter is a demanding engineering and chemical engineering task. To calculate the parameters for the design, including dimensions and number of catalyst beds, temperature profiles, gas compositions, and pressure drop, a suitable mathematical model is required.

Two differential equations describe mathematically the steady-state behavior of the reactor section of a converter. The first models the concentration – position relationship for transformation of the reactants to products, i.e., the reaction kinetic equation (cf. Section Kinetics). The second handles the temperature – position behavior of the reacting synthesis gas, the catalyst, and the vessel internals. The form of the latter is characteristic of the type of converter. The temperature profile depends not only on the rate of reaction heat evolution but also on the method and nature of the system for removing heat from the catalyst bed or beds. Additional equations describe the behavior of the separate feed – effluent heat exchanger system [713-717]. General information on converter calculations are given in [127], [716-718]. Computer programs and applications can be found in [719-722]. For a discussion of modeling of different converter types, see [713-717], [723], [724]. Models for multibed quench converters are described in [723-730]. Tubular reactors are treated in [732-739]. The individual effects of the operational parameters are evaluated in [14], [130], [731].

The reaction temperature profile is of particular importance because the reaction rate responds vigorously to temperature changes. Figure 39 plots lines of constant reaction rate illustrating its dependence on temperature and ammonia concentration in the reacting synthesis gas. The line for zero reaction rate corresponds to the temperature – concentration dependence of the chemical equilibrium. From Figure 39 it is apparent that there is a definite temperature at which the rate of reaction reaches a maximum for any given ammonia concentration. Curve (a) represents the temperature – concentration locus of maximum reaction rates. To maintain maximum reaction rate, the temperature must decrease as ammonia concentration increases.

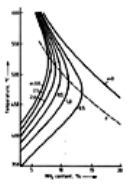

Figure 39. Nitrogen reaction rate v in m^3 NH_3 / (m^3 catalyst · s) as a function of temperature and ammonia concentration at 20 MPa pressure and 11 vol % inerts in the inlet synthesis gas

a) Locus of temperatures resulting in maximum reaction rate at a given ammonia concentration

If the objective in design or operation were optimizing catalyst utilization, then Figure 39 shows that the converter temperature – composition profile should follow curve (a), which corresponds to maximum reaction rate at all points. It is also obvious that in reality this "ideal" temperature – concentration profile cannot be achieved. For example, a synthesis gas with about 3 % ammonia concentration entering the converter cannot be heated to the "ideal" temperature by heat exchange because the very high temperature required does not exist in the converter system. To reach the "ideal" temperature, the first portion of the catalyst must initially operate adiabatically. Consideration of the service life of the catalyst requires that this maximum initial temperature not exceed that recommended by the manufacturer, usually 530 °C (cf. Section Catalyst Reduction). Following this initial adiabatic temperature rise, it is possible to minimize the required catalyst volume by cooling the reacting synthesis gas such that, as ammonia formation progresses, the temperature follows curve (a). In the days when converters were designed to operate at very high pressures and temperatures and before the advent of improved construction materials, the converter design represented a real limitation on plant capacity. To maximize converter output and plant capacity to achieve the most favorable overall manufacturing cost, it was necessary to optimize catalyst utilization. A converter temperature – concentration profile was often compared to the "ideal" for optimum usage of high-pressure vessel and catalyst volumes [731], [740].

Commercial Ammonia Converters.
Principal Reactor Configurations. Commercial converters can be classified into two main groups:

1. Internally cooled with cooling tubes running through the catalyst bed or with catalyst inside the tubes and the cooling medium on the shell side. The cooling medium is mostly the reactor feed gas, which can flow counter- or cocurrently to the gas flow in the synthesis catalyst volume (*tube-cooled converters*).
2. The catalyst volume is divided into several beds in which the reaction proceeds adiabatically. Between the individual catalyst beds heat is removed by injection of colder synthesis gas (*quench converters*) or by indirect cooling with synthesis gas or via boiler feed water heating or raising steam (*indirectly cooled multibed converters*).

The gas flow can have an axial, cross-flow or radial flow pattern. The different cooling methods can be combined in the same converter.

The severe conditions of high pressure, high temperature, and high hydrogen partial pressures place strict requirements on the construction materials and design for both groups. For example, almost all converters consist of an outer pressure vessel containing a separate inner vessel housing the catalyst and other internals, such as gas distributors and heat exchangers. Relatively cool converter feed gas flows through the annular space between the outer pressure shell and the internal "basket". This shields the outer shell from the high-temperature "basket", permitting use of comparatively low-alloy chromium – molybdenum steels for its construction. Often, part of the converter feed – effluent heat-exchange system surface is placed within the converter pressure shell. By this means, the nozzle penetration through the pressure shell for the converter effluent gas is also maintained at relatively low temperature. Today, this latter feature is not always necessary; the state of the art in converter construction materials now allows design of exit nozzles for the maximum anticipated temperatures, i.e., up to about 530 °C. References [745] and [746] review some literature on ammonia converters.

Tube-Cooled Converters. To remove the heat evolved in the synthesis reaction, converters have been designed in which cooling tubes run through the catalyst bed. With these tubes, the heat is transferred to the converter feed gas to heat it to the reaction ignition temperature or to an external cooling medium. The known designs for such converters are suited only for small production capacities and therefore currently of limited interest. When designed to utilize the heat of reaction for heating the converter feed gas, such converters have the further disadvantage that temperature control is sluggish and temperature oscillations dampen out very slowly, if at all, a phenomenon called "hunting".

Typical of such converters is the countercurrent design of the Tennessee Valley Authority (TVA) [710], [737], [740], [749-751], [757-759]. Part of the feed gas to this unit enters the converter at the top and flows down through the annular space between the pressure shell and the basket. The main gas flow enters the converter at the bottom and joins the shell cooling gas. The mixture is heated to about 200 °C in an internal exchanger located beneath the catalyst bed. The gas then enters the cooling tubes that run through the catalyst bed (Fig. 40 A). There it absorbs the heat released in the reaction and reaches the required reaction ignition temperature of about 400 °C. The reaction begins almost adiabatically in the catalyst bed. As the reacting gas temperature rises, the temperature difference between the reacting gas and the cooling tubes increases, resulting in increasing heat removal (Fig. 40 B). As the reacting gas reaches the bottom of the catalyst bed, the rate of reaction begins to decrease sufficiently, because of high ammonia concentration, that cooling predominates and the temperature of the reacting gas begins to fall. Figure 40 C shows the converter temperature – concentration profile. The reaction temperature at first somewhat exceeds that for maximum reaction rate but eventually falls below curve (a). Cold converter feed gas can be admitted at a point within the internal heat exchanger to control both the total system and reaction temperature profiles through bypassing. Casale has employed a similar converter design concept [760].

Figure 40. Countercurrent tube-cooled converter (TVA converter) [740]

A) Converter layout; a) Catalyst; b) Heat exchanger; c) Cooling tubes; d) Main gas inlet; e) Vessel wall cooling gas inlet; f) Temperature control gas (cold-shot) inlet; g) Gas exit

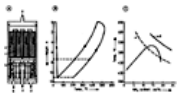

B) Gas temperature profile through the converter; C) Ammonia concentration versus temperature (cf. Fig. 39)

Other tube-cooled converters with countercurrent flow are the Mont Cenis reactor [741-744], the original Haber – Bosch reactors [741], the Claude converter [559], [563], and the older Fauser design [559]. These converters were all used in relatively small plants and are now obsolete.

An interesting rebirth of the countercurrent principle is the new ICI tube cooled converter used in the LCA process(Fig. 41) [1404].

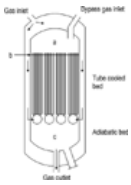

Figure 41. ICI tube-cooled ammonia converter

a) Top of catalyst bed; b) Cooling tubes; c) Catalyst;

The Nitrogen Engineering Corporation (NEC) converter applies cocurrent flow by means of bayonet tubes [740], [749], [751-758]. This design places maximum heat-exchange temperature difference — therefore, maximum cooling performance — at the catalyst bed inlet, the point where maximum reaction rate — therefore maximum rate of heat evolution — is taking place. The intent is to obtain closer approach to curve (a) (cf. Fig. 42 C). Figure 42 A shows the general arrangement of the converter, catalyst bed, and heat exchanger; Figure 42 B, the temperature profile of the system. Chemico, a derivative of NEC, continued to apply such converters with only slight changes [760].

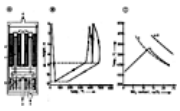

Figure 42. Cocurrent-flow tube-cooled converter [740]

A) Converter; a) Catalyst; b) Heat exchanger; c) Cooling tubes; d) Gas inlet; e) Temperature control gas (cold-shot) inlet; f) Gas exit

B) Gas temperature profile through the converter;

C) Ammonia concentration versus temperature (cf. Fig. 39)

Ammonia Casale developed a new converter concept which uses no tubes but arranged cooling plates. The plates are radially immersed in the axial – radial catalyst bed to remove the reaction heat while it is formed. Figure 43 [1415] shows the temperature profile of this pseudo-isothermal reactor. The profile follows the line of the maximum reaction rate to obtain the highest possible conversion per pass from a given catalyst volume. The use of plates allows a design without tubesheets and simplifies catalyst loading and unloading and the use of small size high active catalyst. The new Casale IAC ammonia converter is suited for very high capacities up to 4500 t/d for the the so-called mega-ammonia plants presently in discussion. The first commercial use of a converter of this type is operating successfully in a 1050 stpd (953 mtpd) plant in Trinidad [1336]. In a recent Casale company brochure the plate arrangement is shown taking a methanol converter as an example (Fig. 44).

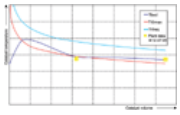

Figure 43. Expected thermal profile of catalyst along the converter and comparison with plant data [1415]

Figure 44. Internals of a Casale isothermal converter (courtesy of Ammonia Casale)

Today, the great majority of synthesis converters are designed with catalyst distributed in several beds, within one or more reactor vessels. In each bed, the synthesis gas reacts adiabatically, and direct or indirect cooling is provided between the catalyst beds for cooling the reacting mixture from a temperature above to a value below the “ideal” (curve (a), Fig. 45).

Multibed Reactors with Direct Cooling (Quench Converters). In quench converters cooling is effected by injection of cooler, unconverted synthesis gas (cold shot) between the catalyst beds. The catalyst beds may be separated by grids designed as mixing devices for main gas flow and cold shot, or be just defined by the location of cold gas injection tubes as for example in the ICI lozenge converter.

In this type of converter only a fraction of the recycle gas enters the first catalyst layer at about 400 °C. The catalyst volume of the bed is chosen so that the gas leaves it at ca. 500 °C (catalyst suppliers specify a maximum catalyst temperature of 530 °C). Before it enters the next catalyst bed, the gas is “quenched” by injection of cooler (125 – 200 °C) recycle gas. The same is done in subsequent beds. In this way the reaction profile describes a zig-zag path around the maximum reaction rate line. A schematic drawing of a quench converter together with its temperature/location and temperature/ammonia concentration profiles is presented in Figure 45.

Figure 45. Multibed converter with quench cooling [1137]

A) Converter; a) Catalyst; b) Heat exchanger; c) Quench gas inlets; d) Gas inlet; e) Gas exit

B) Gas temperature profile through the converter;

C) Ammonia concentration versus temperature (cf. Fig. 39)

A disadvantage is that not all of the recycle gas passes over the entire catalyst volume so that considerable ammonia formation occurs at higher ammonia concentration and therefore at lower reaction rate. Therefore a higher catalyst volume is needed compared to an indirectly cooled multibed converter. However, no extra space is required for interbed heat exchangers, so that the total volume of the high-pressure vessel will remain about the same as that for the indirectly cooled variant [775].

As the quench concept was well suited to large capacity converters it had a triumphant success in the early generation of large single-train ammonia plants constructed in the 1960s and 1970s. Mechanical simplicity and very good temperature control contributed to the widespread acceptance. For example, M. W. Kellogg alone has installed more than 100 of its quench converters. Though being increasingly replaced by the indirect-cooling concept by revamp or substitution they are still extensively used. Descriptions of earlier designs of Kellogg, BASF, and Uhde can be found in [746-748], [757].

The most important example is the M. W. Kellogg three- or four-bed converter [746], [776-778] (Figure 46). In this design, the catalyst “basket” is not easily removable from the pressure vessel. The catalyst can be changed by draining it at the bottom of the converter through “downcomers” that connect all catalyst beds with one

another. The converter feed – effluent exchanger, attached to the top head, is designed for disassembly.

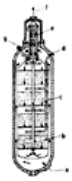

Figure 46. Kellogg four-bed vertical quench converter

a) Gas inlet; b) Catalyst bed; c) Basket; d) Quench; e) Interchanger; f) Gas outlet; g) Bypass

Other designs were used by BASF [746], Casale [746], Chemico, Grand Paroisse [745], ICI [746], Uhde [746], and others.

An interesting variant in this group is the ICI lozenge converter [288], [741], [746], [760], [776]. This design (Fig. 47) has a single catalyst bed that is divided into several zones (usually four) by quench gas distributors, through which colder recycle gas is injected evenly across the whole cross section of the catalyst bed. For this reason it is justifiable to classify this converter as a multibed type. The distributors consist of banks of transverse sparge pipes which deliver gas at regular intervals along their length. The spargers are in a void space within horizontal mesh-covered structures, whose cross section is lozenge shaped so that the catalyst particles can flow freely past them during loading and unloading. A special version of this reactor concept is the opposed flow design [779], [780], suggested for very large capacities. In this configuration the converted gas is collected and withdrawn from the middle of the catalyst bed, with down-flow in the upper half and up-flow in the lower half of the catalyst bed. The uninterrupted catalyst bed is maintained in the opposed-flow converter. A design similar to that of ICI with direct injection of the quench gas into the single catalyst bed has been proposed by Chemico [781].

Figure 47. ICI lozenge converter

a) Quench gas distributors; b) Heat exchanger; c) Catalyst discharge nozzle; d) Tube for thermocouples

Converters with axial flow face a general problem: with increasing capacity the depth of the catalyst beds must be increased because for technical and economic reasons it is not possible to increase the bed diameter, and thus the pressure vessel diameter, above a certain limit. To compensate for the increasing pressure drop conventional axial flow converters have to use relatively large catalyst particles, which have the disadvantage of lower activity compared to smaller particles mainly on account of diffusion restriction. *Radial gas flow* in the converter avoids this dilemma, and with this concept it is possible to design converters for very large capacities without excessive diameters and with low pressure drop, even with small catalyst particle size. The advantages of radial flow are discussed in [783-786]. Radial flow has been also applied in tube-cooled converters [787]. The first radial flow converter introduced commercially and then widely used was the Topsøe S 100 converter [741], [746], [776], [789-794], which has two catalyst layers with a catalyst particle size of 1.5 – 3 mm. Figure 48 shows a schematic of the converter [404].

Figure 48. Haldor Topsøe S100 converter

a) Outer internal lid; b) Inner internal lid; c) First catalyst chamber; d) Inner annular space; e) Perforated center tube; f) Catalyst support plate 1; g) Second catalyst chamber; h) Transfer tube; i) Outer annular space; j) Catalyst support plate 2; k) Heat exchanger; l) Refractory fiber; m) Pressure shell; n) Refractory cement

The major part of the gas enters the vessel at the top and flows down as shell cooling gas. It then passes through the feed – effluent heat exchanger and flows upwards through a central pipe to the first catalyst bed, which is traversed from the inside to the outside. After the effluent from the first bed has been quenched with cooler recycle gas, it enters the second bed and passes through it in the inward direction. The cold gas enters through the bottom of the vessel and is mixed with the inlet gas to the first bed for temperature control.

Radial flow quench converters have also been used by Chemoproject [796], Österreichische Stickstoffwerke [797], and Lummus [798].

Axial – radial flow pattern was introduced by Ammonia Casale. Converters with strictly radial gas flow require mechanical sealing of the top of each catalyst bed and dead catalyst volume with little or no flow to avoid bypassing of the catalyst. In the Casale concept there is no need for a dead catalyst zone as the annular catalyst bed is left open at the top to permit a portion of the gas to flow axially through the catalyst. The remainder of the gas flows radially through the bulk of the catalyst bed. As shown in Figure 49 this is achieved by leaving the upper part of the catalyst cartridge at outlet side unperforated so that gas entering from the top is forced to undergo partially axial flow.

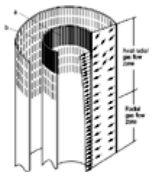

Figure 49. Ammonia Casale axial – radial flow pattern

a) Unperforated wall; b) Perforated wall

Crossflow was chosen as a different approach by M. W. Kellogg in their horizontal quench converter to obtain low pressure drop even with small catalyst particles [745], [770], [790], [791], [800], [801, 1271]. The catalyst beds are arranged side by side in a removable cartridge which can be removed for catalyst loading and unloading through a full-bore closure of the horizontal pressure shell. As the cartridge is equipped with wheels it can be moved in and out on tracks, thus needing no crane. The gas flows vertically from the top to the bottom. The temperature difference between the top and the bottom requires special design measures to prevent uneven circumferential warming of the pressure shell and to avoid bending.

Multibed Converters with Indirect Cooling. In converters of this type cooling between the individual beds is effected by indirect heat exchange with a cooling medium, which may be cooler synthesis gas and/or boiler feed water heating and steam raising. The heat exchanger may be installed together with the catalyst beds inside one pressure shell, but an attractive alternative, especially for large capacities, is to accommodate the individual catalyst beds in separate vessels and use separate heat exchangers. This approach is especially suitable when using the reaction heat for raising high-pressure steam. The indirect cooling principle is applied today in most large new ammonia plants, and also in revamps an increasing number of quench converters are modified to the indirect cooling mode. Figure 50 shows a schematic of the principle together with temperature/location and temperature/ammonia concentration profile.

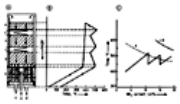

Figure 50. Multibed converter with indirect cooling

A) Converter; a) Catalyst; b) Heat exchanger; c) Cooling section; d) Gas inlet; e) Temperature control gas (cold-shot) inlet; f) Gas exit

B) Gas temperature profile through the converter;

C) Ammonia concentration versus temperature (cf. Fig. 39)

Converters with indirect cooling have been known since the early days of ammonia production, for example, the Fauser – Montecatini reactor [741], [745], [746], [757], [803-805]. In this converter, tube coils between catalyst beds transfer the reaction heat to a closed hot water cycle under pressure, operating by natural draft. The hot water releases the absorbed heat in an external steam boiler generating about 0.8 t of steam per tonne of ammonia at about 45 bar (ca. 250 °C).

In the well-known Uhde – Chemie Linz converter with three catalyst beds — described in various versions [731], [745], [746], [761-764], [776] — the indirect cooling is provided by converter feed gas. Feed gas enters at the top, passes down the annulus between basket and shell to cool the pressure wall, flows through the shell side of the lower feed – effluent heat exchanger, and then via the center pipe and interbed exchangers to the top of the first catalyst bed. The gas passes downwards through the catalyst beds and the tube side of the interbed exchangers and the lower heat exchanger to leave the reactor vessel. For trimming purposes a quench is foreseen.

Further development of the radial flow concept used in the quench converter Topsøe Series 100 has led to the successful launch of the Topsøe Series 200 converter [552], [765], [766], [783], [807-811], designed for indirect cooling. Two versions are shown in Figure 51, with and without a lower internal heat exchanger. A “cold shot” ahead of the first catalyst bed is installed for temperature adjustment.

Figure 51. Topsøe Series 200 converter

a) Pressure shell; b) Interbed heat exchanger; c) 1st Catalyst bed; d) Annulus around catalyst bed; e) 2nd Catalyst bed; f) Lower heat exchanger; g) Cold bypass; h) Cold bypass pipe

In the converter without a lower exchanger the feed gas enters at the bottom and flows as pressure wall cooling gas to the top of the converter. After passing the centrally installed interbed exchanger on the tube side, the gas is mixed with cold gas for temperature adjustment and passes through the first catalyst bed radially from the outside to the inside. The exit gas flows through the shell side of the interbed exchanger before it enters the second bed, which is crossed in the same direction as the first one. The Topsøe Series 300 converter contains three catalyst beds and two central interbed exchangers in a single pressure shell.

Casale [813-818] has also successfully commercialized converters based on the axial – radial flow concept with indirect cooling.

Kellogg has re-engineered its horizontal cross-flow quench converter for indirect cooling [745], [770], [771], [827], [828]. (Fig. 52) As in the quench version the pressure shell has a full-bore closure to remove the catalyst cartridge for loading and unloading. The reactor contains two catalyst beds, with the second one split into two parallel sections. Reactor feed gas passing between cartridge and shell is used to keep the pressure wall cool, and an inlet – outlet heat exchanger is located between first and second bed. A cold shot is installed for adjusting the inlet temperature of the first catalyst bed. A horizontal design with indirect cooling is also proposed in [829].

Figure 52. Kellogg horizontal intercooled ammonia converter

a) Inlet; b) Interbed heat exchanger; c) Bypass; d) Bed 1; e) Bed 2; f) Outlet

In the C. F Braun ammonia synthesis concept [772], [773], [819-822], separate vessels were used for the two catalyst beds in the original version. The feed – effluent heat exchanger was located between the first and second reactor vessels, and the second catalyst vessel was followed by a waste-heat boiler for high pressure steam. This basic concept was already introduced in the heyday of the quench converters and prior to the first energy crisis. In a later version three catalyst vessels [822-826] with boilers after the second and the third were used. The gas flow is axial and the mechanical design rather simple. Contrary to most converter designs, in which the pressure shell is kept below 250 °C by means of insulation or by flushing with cooler gas, the pressure vessel wall of the C. F. Braun reactors is at 400 °C, which is possible with modern steels in compliance with the Nelson diagram (see Section Material Considerations for Equipment Fabrication). Figure 53 shows the two-bed version of the C. F. Braun converter.

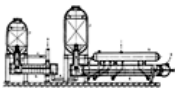

Figure 53. C. F. Braun converters with interbed heat exchanger and waste-heat boiler

a) Gas inlet; b) Feed – first bed effluent exchanger; c) Bypass for temperature control; d) First bed; e) Line to second bed; f) Second bed; g) waste-heat boiler (Borsig); h) Steam drum; i) Steam outlet; k) Gas outlet

To withstand the high outlet temperature level (450 – 500 °C) needed for the high-pressure boilers, a special design is employed to keep the outlet nozzle cool. C. F. Braun has devised a good solution by direct coupling the boilers and the exchanger to the converter.

Uhde has used three catalyst beds for capacities up to 2300 t/d (Fig. 54). The first two are accommodated in a single pressure vessel together with an inlet – outlet exchanger. Then a waste-heat boiler generating high-pressure steam cools the gas before it enters the second vessel containing the third bed, which discharges through a second external high-pressure boiler. The gas flow is in radial mode in all three catalyst beds.

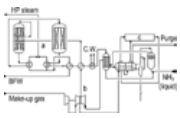

Figure 54. Uhdes standard synloop configuration

a) Ammonia converter; b) Syngas compressor; c) Refrigeration

A new development is the Uhde dual pressure process [1343], [1351-1355], suited for very large capacities. The syngas with 110 bar from the low-pressure casing of the compressor is fed to a three-bed, intercooled, once-through converter which can produce a third of the total ammonia yield. From the effluent about 85 % of the ammonia produced is removed from the gas before it is compressed in the high presure casing to 210 bar to enter the well-known Uhde two-vessel-three-bed combination as described above. Figure 55 shows the Uhde dual pressure concept. The first plant based on this concept with a capacity of 3300 t/d will come on stream 2006 in Al-Jubail, Saudi Arabia (Safco IV). This is so far world-wide the largest capacity in a single synthesis (see also Single-train Capacity Limitations – Mega-Ammonia Plants).

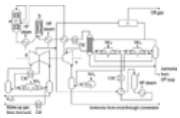

Figure 55. Uhde dual pressure ammonia synthesis concept

a) First ammonia converter; b) Second ammnia converter; c) Purge gas recovery unit; d) LP casing; e) HP casing; f) Once-through ammonia converter

Topsøe has now introduced a hot-wall converter with only one catalyst bed and no internal heat-exchange equipment, similar to C.F. Braun, but with radial flow. Three of these converters can be combined with an external heat exchanger and two high-pressure boilers to give an arrangement as described for Braun. This Topsøe Series 50 converter, can also be combined with the Series 200 reactor and two external high-pressure boilers to attain the configuration as described for Uhde (Topsøe Series 250).

Kellogg has developed for its ruthenium catalyst based KAAP ammonia process [406], [469] a special converter design. Four radial flow beds are accommodated in a single pressure shell with intermediate heat exchangers after the first, second and third bed. The first bed is loaded with conventional iron catalyst, the following ones with the new ruthenium catalyst. Figure 56 is a simplified sketch of the synthesis loop of the KAAP Figure 57 shows the internals of the KAAP converter [1375].

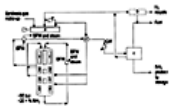

Figure 56. Kellogg advanced ammonia process (KAAP)

a) Compressor; b) Purge gas recovery unit; c) Refrigeration

Figure 57. KAAP four bed ammonia converter

a) Bed no. 1 magnetite catalyst; b) Bed no. 2 KAAP catalyst; c) Bed no. 3 KAAP catalyst; d) Bed no. 4 KAAP catalyst

For revamps Kellogg has also proposed a two-bed version completely loaded with ruthenium catalyst to be placed downstream of a conventional converter [470].

Optimizing the Temperature Profile and Bed Dimensions in Multibed Converters. With closer approach to equilibrium, the volume of catalyst required in the individual beds increases, requiring greater total catalyst volume and appropriate changes in the catalyst bed dimensions. The literature has frequently treated the problem of optimizing both the distribution of catalyst volume between a given number of beds and the bed inlet temperatures [713], [715], [740], [795, 799, 802, 806]. Reference [812] examines optimizing the temperature profile in a given four-bed quench converter. For additional literature on this important issue in converter layout and design, see also Converter Design..

4.5.3.3. *Waste-Heat Utilization and Cooling*

The reaction heat of ammonia synthesis is 46.22 kJ/mol at STP, which corresponds to 2.72 GJ per tonne NH_3. Utilization of this heat at the highest possible temperature for generating high-pressure steam is a major contribution to the high total energy efficiency of modern ammonia plants [830]. Early converters, operating at about 300 bar, equipped with a lower heat exchanger for raising the inlet temperature for the (first) catalyst bed to the ignition temperature (ca. 400 °C) , received the converter feed at about ambient temperature and therefore had outlet temperatures of ca. 250 °C (ca. 15.5 °C per mol % of ammonia formed).

Initially there was practically no heat recovery, and nearly the total heat content of the gas down to ambient temperature and thus the reaction heat was transferred to cooling water. Subsequently plants were modified to use this heat to some extent, but the low temperature level allowed only boiler feed water heating and generation of low-pressure steam (ca. 20 bar). In some instances, water circulation was installed to use this heat in other plants or production steps. As for any type of converter the outlet temperature rises with increasing inlet temperature (ΔT is determined by the degree of conversion), in further developments an additional heat exchanger for converter feed versus the converted gas was installed, downstream of the above-mentioned heat recovery. In this way the temperature level at which heat could be recovered was increased, ultimately to the point where the inlet temperature to waste-heat recovery is equal to the outlet temperature of the last catalyst bed. In practice this corresponds to moving the lower heat exchanger (which in multibed converters exchanges feed to the first catalyst bed against effluent from the last bed) partially or completely to a position outside of the converter and downstream of the waste-heat recovery installation. In this way the waste-heat downstream of the synthesis converter in modern plants is available in the temperature range around 480 to 290 °C. The steam pressure, formerly 100 bar, has now been raised to 125 bar, which means that the gas can be cooled in the boiler to ca. 350 °C ; the remaining recoverable heat is used for boiler feed water heating.

The trend followed in newer plants is to increase conversion per pass with the result of higher ammonia outlet concentrations and lower outlet temperatures from the last bed. However, as optimum energy efficiency of the whole ammonia plant requires maximum high-pressure steam generation, part of the heat must be recovered before the reaction is completed in the reactor system. This can be accomplished [807], [808], [830], [831] by using three catalyst beds in separate pressure vessels with boilers after the second and the third vessel and an inlet – outlet heat exchanger for the first catalyst bed.

Advanced ammonia concepts produce as much as 1.5 t of high-pressure steam per tonne of ammonia, which correspond roughly to 90 % of the standard reaction enthalpy. Figure 58 is a temperature – enthalpy diagram for a converter system corresponding to the original C. F. Braun arrangement. High-pressure steam (113 bar, 320 °C) is generated after the second catalyst bed.

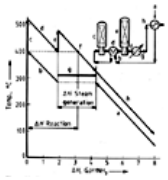

Figure 58. Temperature versus enthalpy diagram for a two-bed system with steam generation

a) Heating in main feed effluent exchanger; b) Further heating in feed – first bed effluent exchanger; c) Temperature rise in first bed; d) Cooling in feed – first bed effluent exchanger; e) Temperature rise in second bed; f) Cooling by steam generation; g) Temperature level of waste-heat boiler; h) Cooling in main feed effluent exchanger

Appropriate designs of waste-heat boilers are described in Waste-Heat Boilers for High-Pressure Steam Generation.

4.5.3.4. *Ammonia Recovery from the Ammonia Synthesis Loop*

In all commercial plants ammonia is recovered from the synthesis loop by cooling the synthesis gas to condense the ammonia under synthesis pressure. The liquid ammonia product is separated from the gas, which is recycled. Arrangement and location of the ammonia separator(s), recirculation compression, addition of make-up gas and extraction of purge gas are discussed in Section Synthesis Loop Configurations; see also Figure 34).

In older high-pressure synthesis loops (> 450 bar) cooling by water and or air is sufficient to obtain the required low residual ammonia concentration in the gas. In modern plants, which operate at moderate pressures, the cooling has to be supplemented by refrigeration, for which a mechanical ammonia refrigeration cycle, with one or more temperature levels is generally used. Refrigeration down to – 25 °C is used, which with inclusion of the necessary temperature difference in the chiller requires ammonia evaporation at about atmospheric pressure. The amount of ammonia vaporized (and consequently reliquefied by compression and water or air cooling) can be substantial. At a loop pressure in the range 100 to 150 bar the quantity of reliquefied ammonia may be twice the ammonia product flow.

The liquid ammonia of the high-pressure separator is flashed to about 20 bar, whereby the majority of the dissolved gases are released in the let-down vessel. This gas is normally used as a fuel, preferably after removal of ammonia vapor to avoid NO_x formation in the combustion furnace.

The ammonia from the let-down vessel may be sent directly to downstream users or flashed further to atmospheric pressure for storage in a cold tank. All ammonia vapors removed from flash gases and from purge gas by water scrubbing can be recovered as pure liquid product by distillation if there is no direct use for the aqueous ammonia.

Absorption refrigeration with aqueous ammonia instead of a mechanical refrigeration system [832], [839], [845], [847], [853], [855] did not find widespread application.

Although ammonia condensation was already used in HABER's lab-scale ammonia plant and in early pilot plants of BOSCH, the first commercial units of BASF used ammonia absorption in water to remove the product from the cool synthesis loop gas, because various technical problems were encountered with refrigeration at that time. It was only in 1926 that ammonia condensation was introduced in the Haber – Bosch plants. In the early 1920s LUIGI CASALE successfully used condensation for his first plant. Water cooling was sufficient on account of the very high synthesis pressure.

Recovery of ammonia by water scrubbing offers the advantage of achieving a very low residual ammonia content, but the drawback is that the whole recycle gas has to be dried afterwards and in addition distillation of aqueous ammonia is necessary to yield liquid ammonia. Nevertheless the scrubbing route was again proposed for a synthesis loop to be operated under extremely low pressure (around 40 bar) [862]. Snam Progetti [288], [866], [867] has proposed removing the ammonia from the loop gas at ambient temperature down to 0.5 mol % by absorption in dilute aqueous ammonia.

The extent to which the ammonia concentration in the gas can deviate from that expected for ideal behavior can be seen from Tables 4 and 5. For example, at 30 bar the ammonia vapor pressure according to Table 4 is 1.167 bar, corresponding to 5.84 mol % at the total pressure of 20 bar. In contrast, Table 5 gives the ammonia concentration in a 1/3 nitrogen/hydrogen mixture at 20 bar total pressure as 9.8 mol %.

4.5.3.5. *Inert-Gas and Purge-Gas Management*

Apart from nitrogen and hydrogen, the fresh make-up gas supplied to the synthesis loop usually contains small quantities of inert gases. These include methane (from gas generation), argon (from the process air), and helium (from the natural gas). Because they are inert , they tend to concentrate in the synthesis loop and must be removed to maintain the loop material balance. A portion of the inert gases dissolves in the liquid produced in the ammonia separator. Table 6 gives a rough approximation of the extent to which these inerts are dissolved. Figure 2 gives vapor – liquid equilibrium ratios for use in making precise calculations of the dissolved quantities of inerts. If the synthesis gas pressure is high, for example, 300 bar), and the inert gas concentration in the synthesis loop make-up gas low enough, for example, under 0.2 vol % [868], then dissolution in the product ammonia suffices to remove the inerts from the synthesis loop.

With a higher inert gas content in the make-up gas this method is not applicable, because the required partial pressure of the inert gas at equilibrium in the loop gas would become so high that a synthesis under moderate pressure would be virtually impossible. So, in addition to removal as dissolved gases (flash gas), inerts have also to be removed from the gas phase by withdrawing a small purge-gas stream from the loop. At the same time expensive synthesis gas is also lost from the loop, which lowers the ammonia yield. Therefore, determining the appropriate inert gas concentration requires a precise economic calculation.

A high inert gas level has various drawbacks. It decreases the specific converter performance by reducing the hydrogen and nitrogen partial pressures. The gas recycle flow is increased by the amount of inert gas. Piping and equipment must correspondingly be increased in size, and the associated power consumption for recycle gas increases. Moreover, there is an unfavorable effect on condensation of the ammonia product. Because of the dilution, less ammonia can be condensed from the recycle synthesis gas by less expensive air or water cooling or higher temperature level refrigeration.

There are several possibilities for reducing the losses associated with the purge gas. The most capital-intensive method consists of feeding the purge gas to a second synthesis loop operating at a slightly lower pressure [869], [870]. As this loop can operate at a very high inert level (40 % or more), only a very small final purge stream is necessary and no recompression is needed. Up to 75 % of the hydrogen from the first-loop purge stream can be recovered. This system has the advantage that nitrogen is also recovered, but it is too expensive for use in modern plants and revamps. For this reason other methods have been developed.

Hydrogen Recovery by Cryogenic Units [848]. Ammonia is first removed from the purge gas by cooling or in a water wash operating at 7.5 MPa (75 bar). Molecular sieve adsorbers then eliminate moisture and traces of ammonia (Fig. 59). The dry, ammonia-free purge gas from the adsorbers next enters the cold box. Heat exchange with cold product hydrogen fraction and with gas rejected to fuel cools the purge gas to a temperature of about –188 °C (85 K). Partial condensation liquefies methane and argon as well as some of the nitrogen and helium. These are removed in a separator, leaving a hydrogen-rich gas.

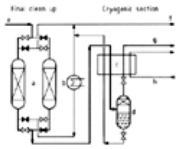

Figure 59. Simplified flow diagram of cryogenic hydrogen recovery unit

a) Molecular sieve adsorbers; b) Heater; c) Plate-fin exchanger; d) Separator; e) Ammonia-free purge gas; f) Fuel gas to reformer burners; g) Hydrogen product to syngas compressor; h) NH_3 refrigerant

The liquid flows through a control valve, reducing its pressure, and into a brazed aluminum (plate-fin or core-type) heat exchanger. The hydrogen-rich gas also flows into the same exchanger (in separate passages) where the vaporizing liquid and the hydrogen are warmed by cooling the entering purge gas. Liquid ammonia from the ammonia plant may be used to provide additional refrigeration, especially at plant startup.

The warm hydrogen-rich gas flows back to the suction side of the second stage of the synthesis gas compressor (6.5 – 7 MPa). About 90 – 95 % of the hydrogen and 30 % of the nitrogen in the purge gas can be recovered.

The remaining gas, with a high concentration of inerts, serves as fuel for the primary reformer. After heating in a preheater, a portion serves to regenerate the molecular sieves and then likewise flows to reformer fuel.

Cryogenic hydrogen recovery units are supplied by firms such as Costain Engineering (formerly Petrocarbon Development) [782], [849-851], [924], [936-941], Linde [852], and L'Air Liquide, among others. Reference [788] reports on the changes in operating conditions of an ammonia plant resulting from the operation of a cryogenic hydrogen recovery unit.

Hydrogen Recovery by Membrane Separation. The Monsanto Prism membrane separator system uses selective gas permeation through membranes to separate gases. This principle has been applied to separating hydrogen from other gases [942-945] (⟶ Membranes and Membrane Separation Processes). The membranes are hollow fibers with diameters of about 0.5 mm. The fiber is a composite membrane consisting of an asymmetric polymer substrate and a polymer coating. The design of a single separator module (length, 3 – 6 m; diameter, 0.1 – 0.2 m) resembles a shell and tube heat exchanger. A bundle with many thousands of hollow fibers is sealed at one end and embedded in a tubesheet at the other. The entire bundle is encased in a vertical shell (Fig. 60).

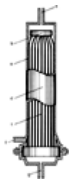

Figure 60. Gas separator module

a) Nonpermeate gas outlet; b) Fiber bundle plug; c) Hollow fiber; d) Separator (length, diameter, and number of separators determined by ammonia process); e) Coated carbon steel shell; f) Feed stream of mixed gases; g) Permeate gas outlet

The purge gas is water scrubbed at 135 – 145 bar, reducing the ammonia concentration to less than 200 ppm. The scrubbed purge gas is heated to 35 °C and sent directly to the Prism separators. Trace concentrations of ammonia and water vapor in the gas stream pose no problem to the membrane. Therefore, a dryer system is not required.

The gas stream enters the separator on the shell side, i.e., the outside of the hollow fibers. Hydrogen permeates through the wall of the fibers. Hydrogen-rich permeate gas flows down the bore of the fiber and through the tubesheet and is delivered at the bottom of the separator. The remaining (nonpermeating) gases, nitrogen, methane, and argon, are concentrated on the shell side, recovered through the top and pass to the next separator module. Several separators operate in series. The rate of permeation decreases across a bank of separators as the hydrogen partial pressure differential across the membrane approaches zero. Therefore, a second bank of separators with lower pressure on the tube side is used to increase the hydrogen recovery. Of the recovered hydrogen, 40 – 70 % leaves the first bank of separators at 7 MPa (70 bar) and is returned to the second-stage suction of the syngas compressor. The second bank permeate hydrogen is recovered at 2.5 – 2.8 MPa (25 – 28 bar) and returned to the first-stage suction of the syngas compressor. Overall hydrogen recovery is 90 – 95 %. The remaining nonpermeate gas stream normally flows to primary reformer fuel.

The main advantages of the Prism separator system are simplicity, ease of operation, and low maintenance. Reference [946] compares membrane and cryogenic separation units for a large ammonia plant.

For further literature on gas separation by membranes, see [854], [856], [857], [947-949], [950].

Membrane technology is also offered by other licensors; an example is the Polysep Membrane System of UOP [858]. In addition to the systems based on hollow fibers, membrane modules have been developed in which the membrane is in the form of a sheet wrapped around a perforated center tube using spacers to separate the layers. The raw gas flows in axial direction in the high pressure spacer and the permeate is withdrawn in the low pressure spacer. Linde, for example, offers such a module under the name Serarex [859], [860].

Hydrogen Recovery by Pressure Swing Adsorption. Pressure swing adsorption on zeolite molecular sieves (PSA) (see Section Gas Purification) may be also used for hydrogen recovery from purge gas [663]. The process, originally developed by Union Carbide under the name HYSIV, is now marketed as Polybed PSA by

UOP [858], [951-953]. PSA technology is also offered by Linde and other companies. If required, the process also offers the possibility to supply pure hydrogen from the purge gas for other uses. PSA units usually operate at adsorption pressures of 20 – 30 bar and achieve recovery rates higher than 82 % for hydrogen and 20 % for nitrogen. Carbon-based adsorbents for pressure swing adsorption have also been investigated [951], [952] and a process developed by Bergbau-Forschung [953] is offered by Costain.

Hydrogen Recovery Using Mixed Metal Hydrides. A proprietary hydrogen separation process utilizing the reversible and selective adsorption capability of mixed metal hydrides has been proposed. The hydride, such as $LaNi_5$, FeTi, or Mg_2Cu, is in the form of ballasted pellets. The ballast material serves as a heat sink to store the heat of adsorption. Subsequently, this is used to supply the heat of desorption. The ballast also is the binder for the pellets, preventing attrition. Each type of metal hydride is susceptible to certain contaminants. Therefore, selection of the metal hydride must be based on the analysis of the gas to be treated. No ammonia removal step is required upstream of the unit. The system yields 99 mol % hydrogen product at a recovery efficiency of 90 – 93 % [954], [860], [861]. No commercial installations are known to be in operation in ammonia production processes.

Argon Recovery from Ammonia Purge Gas. The waste gas from hydrogen recovery plants is more highly enriched in argon than the purge gas. If potential markets for argon exist, then it may be possible to supplement the hydrogen recovery plant with one for recovering argon. Cryogenic argon recovery from ammonia purge gas is discussed in [955-957]. Typical argon recoveries are in excess of 90 %, with a purity of 99.999 %.

4.5.3.6. *Influence of Pressure and Other Variables of the Synthesis Loop*

The influences of individual parameters can be summarized as given in [846, p. 231]:

- Pressure: increasing pressure will increase conversion due to higher reaction rate and more favorable ammonia equilibrium.
- Inlet temperature: there are two opposed effects as increasing temperature enhances reaction rate but decreases the adiabatic equilibrium concentration.
- Space velocity: increasing the space velocity normally lowers the outlet ammonia concentration, but increases total ammonia production.
- Inert level: increasing the inert level lowers the reaction rate for kinetic and thermodynamic reasons (Section Kinetics).
- Hydrogen/nitrogen ratio: a true maximum reaction rate for a certain H/N ratio; at lower temperatures the maximum lies at lower H/N ratios (Section Kinetics). Position of the maximum also depends on the space velocity values (Section Formation of Ammonia in the Converter).
- Recycle rate: at constant pressure and production rate, the consequence of higher recycle rate is a lower ammonia concentration. In this case the difference to the equilibrium concentration and thus the reaction rate increases with the result that less catalyst is required. However, the temperature level for waste-heat recovery decreases and with lower temperature differences, larger heat exchanger surface areas become necessary and the cross sections of piping and equipment have to be enlarged on account of the higher gas flow.
- Separator temperature: together with pressure and location of make-up gas addition, the temperature of the ammonia separator determines the ammonia concentration at the converter inlet. A lower temperature means lower ammonia concentration, which translates into either a lower catalyst volume or a higher conversion.
- Catalyst particle size (Section Particle Size and Shape): smaller catalyst particles give higher conversion because of lower diffusion restrictions (higher pore efficiency)

The question of the best synthesis pressure is rather a difficult one and the answer is extremely dependent on optimization parameters such as feedstock price, required return on investment, and site requirements. In principle it is possible to calculate the minimum amount of mechanical work needed in the synthesis loop. If plots of kilowatt hours versus synthesis pressure for make-up gas compression, recycle, and refrigeration are superimposed the result will be a minimum, for which a value of 155 bar is reported in [863], [864]. The result is strongly dependent on assumed boundary conditions, and other studies came to higher values in the range 180 – 220 bar. However, such diagrams should be interpreted with care. First, this type of plot is strongly influenced by the catalyst activity, and thus by any factors that affect it, such as grain size and minimum possible temperature for the first bed, and especially the equilibrium temperature of the last bed [865]. Second, this approach considers only the mechanical energy and ignores completely the recovered energy of reaction, its energy level, and its impact on the energy balance of the complete ammonia plant. Third, the temperature attainable at a given site by air- or water cooling will affect the refrigeration duty. And fourth, the type of front end can profoundly alter the result. The front-end pressure determines the suction pressure of the synthesis gas machine. In a plant based on partial oxidation with an operating pressure of 80 bar, less than half as much energy is needed to compress the make-up gas to 180 bar, for example, as in a steam reforming plant with a suction pressure of only 25 bar.

Thus the problem of choosing the best synthesis pressure is complex because not only does the entire energy balance have to be examined, but also the mechanical design and the associated investment costs. For an actual project, the costs for energy (i.e., the feedstock) have to be weighed against investment. With the exception of some older processes which used extremely high pressures (Casale, Claude) a pressure of around 300 bar was common for the old multistream plants operated with reciprocating compressors. The centrifugal compressor is one of the most important features of single-train ammonia plants. The first of these plants, with a capacity of 600 t/d, were built in the mid-1960s and the maximum attainable synthesis pressure was restricted to ca. 150 bar by technical limitations of the centrifugal compressor, which needed a minimum gas flow (see also Section Compression). Of course, at current plant sizes of 1200 – 2000 t/d this constraint is no longer of importance. These plants usually operate in the pressure range of 170 – 190 bar and it is a confirmation of the argumentation presented above that the different pressures are not reflected in the overall energy consumption of the complete ammonia plant. Analysis of the effects of various parameters on the energy consumption of the synthesis loop are also reported in [807], [808]. At constant equilibrium temperature for the effluent from the last bed (constant ammonia concentration) energy consumption of the loop was found to be almost independent of pressure in the range of 80 – 220 bar [865].

4.5.3.7. *Example of an Industrial Synthesis Loop*

Figure 61 is an example of modern ammonia synthesis loop (Krupp – Uhde) with two converter vessels and three indirectly cooled catalyst beds producing 1500 t/d NH_3 at 188 bar.

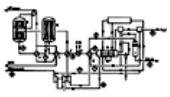

Figure 61. Example of a synthesis loop for 1500 t/d NH_3 (Krupp – Uhde)

a) Converter with two radial beds and internal heat exchange; b) Converter with one radial bed; c) Waste-heat recovery; d) Heat exchanger; e) Water cooling; f) Heat exchanger; g) No. 1 separator; h) Refrigerated cooling; i) No. 2 separator; j) NH_3 recovery; k) H_2 recovery

No.	Flow rate, kmol/h	*T*, °C	*p*, MPa	Gas composition, mol %					Production of NH_3 liquid, t/h
				N_2	H_2	CH_4	Ar	NH_3	
1	28344	295	18.80	23.53	61.20	8.20	2.89	4.18	
2	25368	469	18.48	20.43	50.79	9.15	3.23	16.40	
3	25368	401	18.38	20.43	50.79	9.15	3.23	16.40	
4	24672	336	18.26	19.60	47.99	9.43	3.31	19.67	
5	21613	21	18.02	22.33	54.71	10.67	3.78	8.51	
6	28344	0	17.88	23.54	61.20	8.20	2.88	4.18	
7	3060	21	18.02	0.30	0.55	0.50	0.09	98.55	51.36
8	3678	20	2.50	0.02	0.03	0.12	0.01	99.82	62.51
9	320	38	0.25	46.55	21.72	24.34	7.21	0.18	0.01
10	7667	35	3.23	25.89	72.73	1.07	0.31	0.00	
11	374	38	3.25	5.22	90.20	2.86	1.72	0.00	
12	8041	35	18.20	24.93	73.54	1.16	0.37	0.00	

The gas enters the converter (a) at 295 °C and is subsequently heated in the internal heat exchanger to 390 °C before it enters the first catalyst layer. The outlet gas from the first layer then passes through the aforementioned heat exchanger and enters the second bed, after which the gas leaves the converter with 469 °C and

passes a waste-heat boiler generating 125 bar steam. The inlet gas of the second vessel, which accommodates the third catalyst bed, has a temperature of 401 °C and the outlet enters a further waste-heat boiler generating 125 bar steam.

4.6. Complete Ammonia Production Plants

The previous sections mainly considered the individual process steps involved in the production of ammonia and the progress made in recent years. The way in which these process components are combined with respect to mass and energy flow has a major influence on efficiency and reliability. Apart from the feedstock, many of the differences between various commercial ammonia processes lie in the way in which the process elements are integrated. Formerly the term ammonia technology referred mostly to “ammonia synthesis technology” (catalyst, converters, and synthesis loop), whereas today it is interpreted as the complete series of industrial operations leading from the primary feedstock to the final product ammonia.

The major determinant for process configuration is the type of feedstock, which largely governs the mode of gas generation and purification. The other important factor is the plant capacity, which, together with consumption and costs of feedstock and energy, is decisive for the production economics. An important development was the concept of single-train plants, first introduced for steam reforming based production by M.W. Kellogg in 1963 with a capacity of 600 t/d. Before then maximum capacities had mostly been about 400 t/d, with several parallel trains in the synthesis gas preparation stage and the synthesis loop. Today world-scale plants have capacities of 1200 – 2000 t/d. The lowest capital cost and energy consumption result when steam reforming of natural gas is used. In addition site requirements can influence the layout considerably. In contrast to a stand-alone plant, ammonia production at a developed industrial site may import and/or export steam and power, which affects the total energy consumption.

With the exception of the Koppers – Totzek coal gasification process, which operates at near atmospheric pressure, all modern gasification processes operate at elevated pressure. Steam reforming of light hydrocarbons at 30 – 40 bar and partial oxidation of heavy hydrocarbons at 40 – 90 bar are generally used.

4.6.1. *Steam Reforming Ammonia Plants*

4.6.1.1. *The Basic Concept of Single-Train Plants*

The innovative single-train concept, introduced in 1963 by Kellogg, was a technical and an economical breakthrough and triggered a tremendous increase in world ammonia capacity. No parallel lines, even for high capacity, and a highly efficient use of energy, with process steps in surplus supplying those in deficit, were the main features. Figure 62 shows a block flow diagram and the gas temperature profile for a steam reforming ammonia plant [1377].

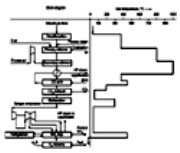

Figure 62. Block diagram and gas temperature profile for a steam reforming ammonia plant

High level surplus energy is available from the flue gas of the reformer and the process gas streams of various sections, while heat is needed, for example, for the process steam for the reforming reaction and in the solvent regenerator of the carbon dioxide removal system. Because considerable mechanical energy is needed to drive compressors, pumps, and blowers, it was most appropriate to use steam turbine drives, since sufficient steam could be generated from the waste-heat. As the level was high enough to raise high-pressure steam (100 bar) it was possible to use the process steam first to generate mechanical energy in the synthesis gas compressor turbine before extracting it at the pressure level of the primary reformer. Table 25 lists all significant energy sources and sinks within the process.

Table 25. Main energy sources and sinks in the steam reforming ammonia process

Process section	Origin	Contribution
Reforming	primary reforming duty	demand
	flue gas	surplus
	process gas	surplus
Shift conversion	heat of reaction	surplus
CO_2 removal	heat of solvent regeneration	demand
Methanation	heat of reaction	surplus
Synthesis	heat of reaction	surplus
Machinery	drivers	demand
Unavoidable loss	stack and general	demand
Balance	auxiliary boiler or import	deficit
	export	surplus

The earlier plants operated at deficit, and needed an auxiliary boiler, which was integrated in the flue gas duct. This situation was partially caused by inadequate waste-heat recovery and low efficiency in some energy consumers. Typically, the furnace flue gas was discharged in the stack at rather high temperature because there was no air preheating and too much of the reaction heat in the synthesis loop was rejected to the cooling media (water or air). In addition, efficiency of the mechanical drivers was low and the heat demand for regenerating the solvent from the CO_2 removal unit (at that time aqueous MEA) was high. Maximum use was made of direct steam turbine drive, not only for the major machines such as synthesis gas, process air, and refrigeration, but even for relatively small pumps and blowers. The outcome was a rather complex steam system. Even after substitution of the smaller turbines by electric motors, the steam system in the modern plant is still a complex system as shown in Figure 64. Ammonia plant steam systems are described in [404], [406], [1318], [961-964].

The first generation of the single-train steam reforming plants is discussed in [420], [959], [960], [979], [980], [988], [995], and the required catalysts are reviewed in [539], [553], [999]. A survey of the development of the steam reforming concept through 1972 can be found in [969]. Other references which cover the development of the steam reforming before the introduction of the single-train concept (1940 to 1960) can be found in [404 p. 276].

The new plant concept had a triumphant success story. By 1969, 30 new Kellogg large single-train plants with capacities of 1000 t/d or more were in operation, and other contractors were offering similar concepts.

The decrease in energy consumption compared to the older technology was dramatic, and with the low gas prices at that time it is understandable that greater emphasis was placed on low investment cost, although there was a considerable potential for further reducing the energy consumption.

With the advent of the single-train steam reforming plants, it became standard for licensors and engineering contractors to express the total net energy consumption per tonne of ammonia in terms of the lower heating value of the feedstock used. The total net energy consumption is the difference between all energy imports (mainly feedstock) and all energy exports (mostly steam and/or electric power) expressed as lower heating value of the consumed feedstock, whereby electric power is converted with an efficiency of 25 – 30 % and steam is accounted for with its caloric value.

4.6.1.2. *Further Developments*

The significant changes in energy prices from 1973 onwards were a strong challenge to process licensors, engineering contractors and plant owners to obtain better energy efficiency. The overall energy consumption was reduced from around 45 GJ per tonne NH_3 for the first large single-train units to less than 29 GJ per tonne NH_3 (Table 26).

Table 26. Development of the net energy consumption of natural gas based steam reforming ammonia plants (real plant data) in GJ per tonne NH_3

Year	1966	1973	1977	1980	1991
Plant	A	B	C	D	E
Feed	23.90	23.32	23.48	23.36	22.65
Fuel, reformer	13.00	9.21	7.35	5.56	5.90
Fuel, auxiliary boiler	2.60	5.02	3.06	1.17	
Export					0.55
Total	39.50	37.55	33.89	30.18	28.00

Energy saving modifications are described in [288], [867], [970-978], [1000], [1001], [1005], [1018], [1025], [1030], [1254]. For catalyst improvements see [1395]. Some of the most important improvements compared to the first generation of plants are discussed below.

Feedstock Purification. In feedstock purification, mainly desulfurization, adsorption on active carbon was replaced by catalytic hydrogenation over cobalt – molybdenum or nickel – molybdenum catalyst, followed by absorption of the H_2S on ZnO pellets with formation of ZnS. By itself this measure has no direct influence on the energy consumption but is a prerequisite for other energy saving measures, especially in reforming and shift conversion.

Reforming Section. In the reforming section energy savings were achieved by several, often interrelated, measures, of which the most important are the following:

- Reduction of the flue gas stack temperature to reduce heat losses to the atmosphere [552]
- Avoiding excessive heat loss by better insulation of the reformer furnace
- Introduction of combustion air preheating [1018]
- Preheating the reformer fuel
- Increased preheat temperatures for feed, process steam and process air
- Increased operating pressure (made possible by using improved alloys for the reformer tubes and improved catalysts)
- Lowering of the steam to carbon ratio [1032]
- Shifting some reformer duty from primary to secondary reformer with the use of excess air [457], [458] or oxygen-enriched air [1276] in the secondary reformer, including the possibility of partially bypassing the primary reformer [1033], [1036]
- Installing a prereformer or rich-gas step is another possibility to reduce primary reformer duty and stack temperature of the flue gas [431-438], especially in LPG- and naphtha-based plants

A more recent development that breaks away from the usual plant configuration is to replace the traditional fired primary reformer with an exchanger reformer which uses the heat of the effluent of the secondary reformer [461-464], [466-469], [1035], [1036]. Also other applications have been reported in which flue gas [1035] heat from the fired reformer is used to perform a part of the reforming.

Shift Conversion. Improved LT shift catalysts can operate at lower temperatures to achieve a very low residual CO content and low byproduct formation. A new generation of HT shift catalysts largely avoids hydrocarbon formation by Fischer – Tropsch reaction at low vapor partial pressure, thus allowing lower steam to carbon ratio in the reforming section (see Section Carbon Monoxide Shift Conversion).

Carbon Dioxide Removal Section. In the carbon dioxide removal section the first generation of single-train plants often used MEA with a rather high demand of low-grade heat for solvent regeneration. With additives such as UCAR Amine Guard [603], [637], solvent circulation could be reduced, saving heat and mechanical energy. Much greater reduction of energy consumption was achieved with new solvents and processes, for example BASF aMDEA or Benfield LoHeat. Other hot potash systems (Giammarco Vetrocoke, Catacarb) and physical solvents (Selexol) were introduced (Section Gas Purification).

Final make-up gas purification was improved by removing the water and carbon dioxide traces to a very low level by using molecular sieves. Some concepts included cryogenic processes with the benefit of additional removal of methane and argon.

Ammonia Synthesis Section. In the ammonia synthesis section conversion was increased by improved converter designs (see Section Formation of Ammonia in the Converter), larger catalyst volumes, and to some extent with improved catalysts. The main advances in converter design were the use of indirect cooling instead of quenching, which allowed the recovery of reaction heat high pressure steam. Radial or cross-flow pattern for the synthesis gas instead of axial flow was introduced. All modern plants include installations for hydrogen recovery (cryogenic, membrane, or PSA technology; see Section Inert-Gas and Purge-Gas Management).

Machinery. Developments in compressor and turbine manufacturing have led to higher efficiencies.

- Gas turbine drive for a compressor and/or an electric generator combined with the use of the hot exhaust as combustion air for the primary reformer or raising steam (combined cycle) [705-709], [1039].
- Employing electric motors instead of condensation turbines [1040].
- Application of liquid and gas expansion turbines can recover mechanical work (e.g., let-down of the CO_2-laden solvent, liquid ammonia, purge and fuel gas).

Steam system and waste-heat recovery were improved by the following measures: increased pressure and superheating temperature of high-pressure steam; providing a part of the process steam by natural gas feed saturation [458], [603], [965-967]; inclusion of a steam superheater downstream of the secondary reformer [964], [968].

Process Control and Process Optimization. Progress in instrumentation and computer technology has led to increased use of advanced control systems and computerized plant optimization. Advanced control systems [981-987], [1133-1135] allow operating parameters to be kept constant in spite of variations in external factors such as feedstock composition or ambient or cooling water temperatures. These systems may be operated in open loop fashion (set values changed manually by the operator) or in closed fashion (set points automatically adjusted to optimum values by using a computer model with input of operational and economic data). Also plant simulation [989-992] is possible by using extensive computer models of complete plants. These models can simulate in real time the dynamic response to changes in operating parameters, plant upsets, etc. Such systems are used for off-line optimization studies and for operator training [993], [994].

The above list, by no means complete, is also a survey of options for plant revamps (Section Modernization of Older Plants (Revamping)). Quite a number of options can be found in papers presented at the "AIChE Annual Symposium Ammonia Plants and Related Facilities" giving practical experience and presenting case stories [1350].

Many of these elements are strongly interrelated with each other and may affect different sections of the plant concept. It is thus a demanding engineering task to arrive at an optimum plant concept, which can only defined by the conditions set by the feedstock price, the site influences, and the economic premises of the customer. An evaluation of the individual merits of the described measures in terms of investment and operational cost in a generalized form is not possible and can be done only from case to case in real project studies.

To illustrate the forgoing discussion of the concept of the single-train steam reforming plant, Figure 63 presents a modern low-energy ammonia plant with flow sheet and process streams (UHDE process).

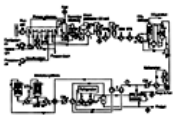

Figure 63. Modern integrated single-train ammonia plant based on steam reforming of natural gas (Uhde process)

a) Sulfur removal; b) Primary reformer; c) Steam superheater; d) Secondary reformer; e) waste-heat boiler; f) Convection section; g) Forced draft fan; h) Induced draft fan; i) Stack; k) HT and LT shift converters; l) Methanator; m) CO_2 removal solvent boiler; n) Process condensate separator; o) CO_2 absorber; p) Synthesis gas compressor; q) Process air compressor; r) Ammonia converter; s) High-pressure ammonia separator; t) Ammonia and hydrogen recovery from purge and flash gas

	1	2	3	4	5	6	7	8	9
CH_4, mol %	91.24	91.24	14.13	0.60	0.53	0.65	1.16	24.34	0.12
C_nH_m, mol %	5.80	5.80							
CO_2, mol %	1.94	1.94	10.11	7.38	18.14	0.01			
CO, mol %			9.91	13.53	0.33	0.40			
Ar, mol %				0.28	0.25	0.30	0.37	7.21	0.01
H_2, mol %			65.52	54.57	59.85	73.08	73.54	21.72	0.03
N_2, mol %	1.02	1.02	0.33	23.64	20.90	25.56	24.93	46.55	0.02
NH_3, mol %								0.18	99.82
Drygas, kmol/h	1713.7	534.43	5296.4	8414.2	9520.7	7764.0	8041.4	319.9	3676.6
H_2O, kmol/h			3520.6	4086.1	2979.6	22.8	13.3	0.2	0.9
Total, kg/h	30213	9422	121183	199555	199555	70084	71004	6496	62626
p, MPa	5.00	0.25	3.95	3.90	3.61	3.43	3.23	0.25	2.50
T, °C	25	25	808	976	229	50	35	38	20

Figure 64 shows a simplified diagram of the steam system. Even in such an advanced plant the quantity of steam generated from waste-heat is as much as 3.4 times the weight of ammonia produced.

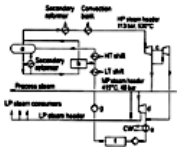

Figure 64. Steam system of a modern steam reforming

a) Steam drum, 125 bar; b) NH_3 loop; c) Turbine for syngas compressor; d) Turbine for process air compressor and alternator; e) Surface condenser; f) Condensate treatment; g) BFW pump

4.6.1.3. *Minimum Energy Requirement for Steam Reforming Processes*

The energy saving measures described in Section Further Developments have considerably reduced the demand side (e.g., CO_2 removal, higher reforming pressure, lower steam to carbon ratio, etc.). On the supply side, the available energy has been increased by greater heat recovery. The combined effects on both sides have pushed the energy balance into surplus. Because there is no longer an auxiliary boiler which can be turned down to bring the energy situation into perfect balance, the overall savings usually could not be translated into further actual reduction of the gross energy input to the plant (mainly natural gas). In some cases this situation can be used advantageously. If the possibility exists to export a substantial amount of steam, it can be economically favorable (depending on feedstock price and value assigned to the steam) to deliberately increase the steam export by using additional fuel, because the net energy consumption of the plant is simultaneously reduced (Table 27).

Table 27. Increase of plant efficiency by steam export (GJ per tonne NH_3)

	Plant		Difference
	A	B	
Natural gas	27.1	32.6	+ 5.5
Electric power	1.1	1.1	
Steam export		– 6.4	– 6.4
Total energy	28.2	27.3	– 0.9

A reduction in gross energy demand, that is, a lower natural gas input to the plant, can only be achieved by reducing fuel consumption, because the actual feedstock requirement is determined by the stoichiometry. So the only way is to decrease the firing in the primary reformer, which means the extent of reaction there is reduced. This can be done by shifting some of the reforming duty to the secondary reformer with surplus air or oxygen-enriched air, although this makes an additional step for the removal of surplus nitrogen necessary. A more radical step in this direction is total elimination of the fired primary reformer by using exchanger reformers like the ICI GHR and the Kellogg KRES.

Based on pure methane, it is possible to formulate a stoichiometric equation for ammonia production by steam reforming:

$$CH_4 + \underbrace{0.3035\ O_2 + 1.131\ N_2}_{1.4345\ \text{air}} + 1.393\ H_2O \longrightarrow CO_2 + 2.262\ NH_3$$

$$\Delta H = -86\ \text{kJ/mol and}\ \Delta F = -101\ \text{kJ/mol at 25 °C}$$

From a mere thermodynamic point of view, in an ideal engine or fuel cell, heat and power should be obtainable from this reaction. Since real processes show a high degree of irreversibility, a considerable amount of energy is necessary to produce ammonia from methane, air and water. The stoichometric quantity of methane derived from the above reaction is 583 m^3 per tonne of ammonia, corresponding to 20.9 GJ per tonne NH_3 (LHV), which with some justification could be taken as minimum value. If full recovery of the reaction heat is assumed, then the minimum would be the lower heating value of ammonia, which is 18.6 GJ per tonne NH_3. Table 28 compares the specific energy requirement for ammonia production by steam reforming with the theoretical minimum.

Table 28. Specific energy requirement for ammonia production compared to the theoretical minimum

	GJ per tonne NH_3 (LHV)	% theory
Classical Haber – Bosch (coke)	80 – 90	(338 – 431)
Reforming, 0.5 – 10 bar (1953 – 55)	47 – 53	225 – 254
Reforming, 30 – 35 bar (1965 – 75)	33 – 45	139 – 215
Low energy concepts (1975 – 84)	29 – 33	139 – 158
Modern concepts (since 1991)	< 28	134
Stoichiometric CH_4 demand	20.9	= 100

Comparison of energy consumption figures without precise knowledge of design and evaluation criteria can be misleading. First of all the state of the ammonia product should be noted. Relative to the delivery of liquid ammonia to battery limits at ambient temperature, the production of 3 bar ammonia vapor at the same temperature would save 0.6 GJ per tonne NH_3, while delivery as liquid at – 33 °C would need an extra 0.3 GJ per tonne NH_3. The temperature of the available cooling medium has a considerable influence. Increasing the cooling water temperature from 20 to 30 °C increases energy consumption by 0.7 GJ per tonne NH_3 . A detailed energy balance, listing all imports and exports, together with the caloric conversion factors used for steam and power is needed for a fair comparison of plants. The beneficial effect of energy export to the net energy consumption is discussed above. Gas composition is also of some importance. Nitrogen content is marginally beneficial: 10 mol % nitrogen leads to a saving of about 0.1 GJ per tonne NH_3, whereas a content of 10 mol % carbon dioxide would add 0.2 GJ per tonne NH_3 to the total consumption value [996 p. 263].

Energy requirements and energy saving possibilities are also discussed in [288], [867], [1000], [1001], [1005], [1018], [1025], [1030], [1254].

The energy consumption figures discussed so far represent a thermodynamic analysis based on the first law of thermodynamics. The combination of the first and second laws of thermodynamics leads to the concept of ideal work, also called exergy. This concept can also be used to evaluate the efficiency of ammonia plants. Excellent studies using this approach are presented in [997 p. 258], [998]. Table 29 [997] compares the two methods. The analysis in Table 29 was based on pure methane, cooling water at 30 °C (both with required pressure at battery limits), steam/carbon ratio 2.5, synthesis at 140 bar in an indirectly cooled radial converter.

Table 29. Energy analysis of a low energy ammonia plant (GJ per tonne NH_3)

	HHV	LHV	Exergy
Input			
Natural gas consumption			
Reformer feed	24.66	22.27	23.28
Reformer fuel	7.49	6.78	7.08
Auxiliary boiler fuel	0.34	0.29	0.33
Total consumption	32.49	29.34	30.69
Losses			
Reforming	0,38	0,38	4,94
Steam generation	0.33	0.33	2.39
Shift, CO_2 removal, methanation	1.30	1.30	0.67
Synthesis	1.70	1.70	1.55
Turbines and compressors	6.50	6.50	0.54
Others (including stack)	1.30	0.68	0.46
Total losses	11.51	10.89	10.55
NH_3 product	20.98	17,12	20.14
Efficiency, %	64,60	58,40	65,60

Almost 70 % of the exergy loss in the process occurs in the reforming section and in steam generation. From conventional first law analysis it can be seen that almost all of the losses are transferred to the cooling water. As the analysis assumes water in liquid state, the LHV analyses in Table 29 is not completely balanced. For a perfect balance the heat of evaporation of water (as a fictive heating value) would have to be included.

4.6.1.4. *Commercial Steam Reforming Ammonia Processes*

Especially with ammonia processes based on steam reforming technology it has become a habit to differentiate between processes from various licensors and engineering contractors. This is not so much the case for partial oxidation plants (see Section Ammonia Plants based on Partial Oxidation). Strong competition together with increased plant size and the associated financial commitment has reduced the number of licensors and engineering contractors to a few companies capable of offering world-scale plants, often on a lump-sum turnkey basis. In some cases these companies sub-license their processes and special engineering know-how to competent engineering companies possessing no knowledge of their own in the ammonia field. There are also several smaller companies with specific and sometimes proprietary know-how which specialize in revamps of existing plants or small plant concepts.

In the following, each of the commercially most important processes is discussed in some detail and a shorter description of economically less important processes is given. The process configuration offered and finally constructed by a given contractor may vary considerably from case to case, depending on economic and site conditions and the clients' wishes. Thus plants from the same contractor and vintage often differ considerably. It is possible to categorize steam reforming plants according to their configuration in the reforming section:

1. Advanced conventional processes with high duty primary reforming and stoichiometric process air in the secondary reformer
2. Processes with reduced primary reformer firing and surplus process air
3. Processes without a fired primary reformer (exchanger reformer)
4. Processes without a secondary reformer using nitrogen from an air separation plant

In principle the amount of flue gas emitted should be related to the extent of fired primary reforming, but generalizations are questionable, because sometimes the plant layout, as dictated by site requirements, may considerably change the picture for the specific flue gas value.

4.6.1.4.1. Advanced Conventional Processes

Kellogg Low-Energy Ammonia Process [1042], [1043], [1254-1259], [1277]. The Kellogg process is along traditional lines, operating with steam/carbon ratio of about 3.3 and stoichiometric amount of process air and low methane slip from the secondary reformer. The synthesis pressure depends on plant size and is between 140 and 180 bar. Temperatures of the mixed feed entering the primary reformer and of the process air entering the secondary reformer are raised to the maximum extent possible with today's metallurgy. This allows reformer firing to be reduced and, conversely, the reforming pressure to be increased to some extent to save compression costs. An important contribution comes from Kellogg's proprietary cross-flow horizontal converter, which operates with catalyst of small particle size, low inlet ammonia concentration, and high conversion. Low-energy carbon removal systems (Benfield LoHeat, aMDEA, Selexol) contribute to the energy optimization.

When possibilities to export steam or power are limited, part of the secondary reformer waste-heat is used, in addition to steam generation, for steam superheating, a feature in common with other modern concepts. Proprietary items in addition to the horizontal converter are the traditional Kellogg reformer, transfer line and secondary reformer arrangement, waste-heat boiler, and unitized chiller in the refrigeration section.

According to Kellogg 27.9 GJ per tonne NH_3 can be achieved in a natural gas based plant with minimum energy export, but with export of larger quantities of steam this value could probably be brought down to about 27 GJ per tonne NH_3. Figure 65 shows a simplified flowsheet of the process [1259] with Selexol CO_2 removal systems (other options are, e.g., Benfield or BASF aMDEA).

Figure 65. M.W. Kellogg's low energy process

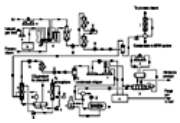

a) Feed gas compressor; b) Desulfurization; d) Primary reformer; e) Air compressor; f) Secondary reformer; g) Heat recovery; h) High temperature shift converter; i) Low temperature shift converter; j) Condensate stripper; k) CO_2 absorber; l) CO_2 flash drum; m) Recycle compressor; n) Semi-lean Pump; o) Stripper (other options are, e.g., Benfield or BASF aMDEA); p) Stripper air blower; q) CO_2 lean pump; r) Methanator feed preheater; s) Methanator; t) Synthesis gas compressor; u) Dryer; v) Purge gas H_2 recovery; w) Ammonia converter; x) Start-up heater; y) Refrigeration exchanger; z) Refrigeration compressor

Haldor Topsøe Process. In addition to technology supply, Haldor Topsøe also produces the full catalyst range needed in ammonia plants. The energy consumption of a basically classic plant configuration has been reduced considerably by applying systematic analysis and processes engineering. Descriptions and operational experience are given in [552], [433], [1018], [1049-1060].

Topsøe offers two process versions. The first operates at steam/carbon ratio of 3.3 and with rather high residual methane content from the secondary reformer. Shift conversion is conventional, the Benfield or Vetrokoke process is used for carbon dioxide removal, and the synthesis pressure depends on plant size ranging between 140 and 220 bar when the proprietary Topsøe two-bed radial converter S 200 is used. A simplified flowsheet is presented in Figure 66.

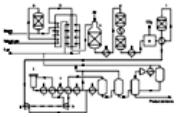

Figure 66. Haldor Topsøe's low energy process

a) Desulfurization; b) Primary reformer; c) Secondary reformer; d) Shift conversion; e) CO_2 removal; f) Methanation; g) Main compressor; h) Recycle compressor; i) Heat recovery; j) Converter

An actual plant has reported a consumption of 29.2 GJ per tonne NH_3 [1058].

The second version has a S/C ratio of 2.5 and shift conversion with medium- and low-temperature catalysts, both copper-based. For CO_2 removal Selexol or aMDEA is chosen. The synthesis is performed at 140 bar with a Topsøe two-bed S 200 radial converter, followed by a single-bed radial S 50 converter (S 250 configuration). After the converters, high-pressure steam is generated. An additional proprietary item is the side-fired reformer.

For this most energy-efficient concept a figure of 27.9 GJ per tonne NH_3 is claimed [1057].

Uhde Process. Uhde, in the ammonia business since 1928, markets a low-energy ammonia plant with classic process sequence and catalysts [1061-1068], [1343]. High plant reliability at competitive overall costs was a major objective. A process flow diagram together with the main process stream is presented in Figure 63.

Key features are the high reforming pressure (up to 43 bar) to save compression energy, use of Uhde's proprietary reformer design [1062] with rigid connection of the reformer tubes to the outlet header, also well proven in many installations for hydrogen and methanol service. Steam to carbon ratio is around 3 and methane slip from the secondary reformer is about 0.6 mol % (dry basis). The temperature of the mixed feed was raised to 580 °C and that of the process air to 600 °C. Shift conversion and methanation have a standard configuration, and for CO_2 removal BASF's aMDEA process (1340 kJ/Nm^3 CO_2) is preferred. Synthesis is performed at about 180 bar in Uhde's proprietary converter concept with two catalyst beds in the first pressure vessel and the third catalyst bed in the second vessel.

After each converter vessel high pressure steam (125 – 130 bar, up to 1.5 t per tonne NH_3) is generated (Uhde also offers its own boiler design in cooperation with an experienced boiler maker). Heat exchange between inlet and outlet of the first bed is performed in the first vessel, and gas flow in all beds is radial. When only a minimum of energy export (steam or power) is possible, the process heat from the secondary reformer outlet is partly used to raise high-pressure steam, and partly for superheating high-pressure steam. Refrigeration uses screw compressors with high operational flexibility and efficiency. Instead of the standard synloop Uhde offers the Dual Pressure Process for very large capacities (Fig. 55) (see Section Formation of Ammonia in the Converter). In this variant a once-through converter at lower pressure (100 bar), which produces about one third of the capacity, is followed by Uhde's standard loop (Fig. 54) at around 200 bar [1343], [1351-1355].

Achieved net energy consumption is about 28 GJ per tonne NH_3 and Uhde's engineers expect values of below 27 GJ per tonne NH_3 when a gas turbine and large steam export is included [1066].

LEAD Process (Humphreys & Glasgow, now Jacobs) [867], [1044]. The LEAD process is a highly optimized conventional approach with synthesis at 125 bar and two converter vessels, the first of which contains two catalyst beds with axial-flow quenching, while the second has a third bed with small particle size catalyst and radial flow. A consumption of 29.3 GJ per tonne NH_3 is claimed.

Exxon Chemical Process. The Exxon Chemical process [1070], [1071], was specifically designed for the company's own site in Canada and so far not built for third parties. It uses a proprietary bottom-fired primary reformer furnace and a proprietary hot potash carbon dioxide removal system with a sterically hindered amine activator. Synthesis loop and converter are licensed by Haldor Topsøe A/S. Synthesis is carried out at 140 bar in a Topsøe S-200 converter and total energy consumption is reported to be 29 GJ per tonne NH_3.

Fluor Process. The Fluor process [288], [845], [867], [1041] uses the proprietary propylene carbonate based CO_2 removal system with adsorption refrigeration using low level heat downstream of the low-temperature shift. Methanation and CO_2 removal are placed between the compression stages and thus operate at higher pressure. With a value of 32 GJ per tonne NH_3 [845] this is not really a low-energy concept.

Lummus Process. For the Lummus Process schemes [288], [867], [1045], [1047], [1048] a consumption of 29.6 [867] to 33.5 GJ per tonne NH_3 [1046] is quoted. In the synthesis section either an axial flow quench converter or a radial flow converter with indirect cooling is used. CO_2 removal is performed with a physical solvent, and there are no special features compared to other conventional process configurations.

Integrating the ammonia and urea process into a single train was proposed by Snam Progetti to reduce investment and operating costs [1097].

4.6.1.4.2. Processes with Reduced Primary Reformer Firing

Braun Purifier Process [457], [651], [654], [655], [867], [1069], [1072], [1073], [1261], [1262-1268]. Characteristic of the low-energy Braun purifier process (Fig. 67) is the reduced primary reformer duty, which is achieved by shifting a proportion of the reforming reaction to the secondary reformer by using about 150 % of the stoichiometric air flow. The excess nitrogen introduced in this way is removed after the methanation step in a cryogenic unit known as a purifier [1263], which also removes the methane and part of the argon. The result is a purer synthesis gas compared to conventional processes, and only minimal purge from the loop is required. A typical flow diagram of this process is shown in Figure 67.

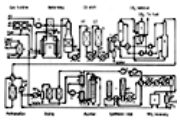

Figure 67. The Braun purifier ammonia process

a) Sulfur removal; b) Primary reformer; c) Convection section; d) Secondary reformer; e) waste-heat boiler; f) Process air compressor; g) Gas turbine; h) High- and low-temperature shift converters; i) CO_2 removal solvent reboiler; k) CO_2 absorber; l) CO_2 desorber; m) CO_2 stripper; n) Methanator; o) Driers; p) Purifier heat exchanger; q) Expansion turbine; r) Purifier column; s) Synthesis gas compressor; t) Synthesis converters; u) waste-heat boiler; v) High-pressure ammonia separator; w) Ammonia letdown vessel; x) Ammonia recovery from purge gas

Synthesis is carried out in the proprietary Braun adiabatic hot-wall converter vessels (Fig. 53). Each catalyst bed (of which three are now used in newer plants [1262]) is accommodated in a separate vessel with an inlet – outlet heat exchanger after the first and high-pressure steam boilers after the following. The smaller furnace produces less flue gas and consequently less waste-heat, which makes it easier to design a balanced plant with no energy export. The lower reforming temperature allows a reduction of the steam/carbon ratio to about 2.75 without adverse effects on the HT shift, because of the less reductive character of the raw gas on account of its higher CO_2 content. In energy balanced plants, the use of waste-heat in the secondary reformer effluent is split between steam raising and steam superheating.

The concept shows great flexibility [1264] for design options. It is possible, for example, to aim for minimal natural gas consumption, even at the cost of importing some electric power. On the other hand, it is possible to improve the overall efficiency further by exporting a greater amount of energy. In this latter case it is advantageous to incorporate a gas turbine to drive the process air compressor. The hot exhaust (about 500 °C) of the turbine contains 16 – 17 mol % of oxygen and can serve as preheated combustion air of the primary reformer. In addition it is possible to include an electric generator to cover the plant demand and export the surplus. The C. F. Braun process can attain 28 GJ per tonne NH_3 in a balanced plant, but with steam export and realization of the available improvement possibilities a value of 27 GJ per tonne NH_3 seems feasible.

ICI AMV Process. The ICI AMV process [458], [1060], [1074-1086], [1404] also operates with reduced primary reforming (steam/carbon ratio 2.8) and a surplus of process air in the secondary reformer, which has a methane leakage of around 1 %. The nitrogen surplus is allowed to enter the synthesis loop, which operates at the very low pressure of 90 bar with an unusually large catalyst volume, the catalyst being a cobalt-enhanced version of the classical iron catalyst. The prototype was commissioned 1985 at Nitrogen Products (formerly CIL) in Canada, followed by additional plants in China. A flow sheet is shown in Figure 68

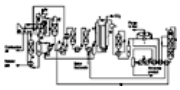

Figure 68. ICI AMV process

a) Desulfurization; b) Natural gas saturation; c) Process air compression; d1) Primary reformer; d2) Secondary Reformer; e) Boiler; f) High temperature shift; g) Low temperature shift; h) Selexol CO_2 removal; h1) CO_2 absorber; h2) Regenerator; i) Single stage compression; j) Methanation; k) Cooling and drying; l) Circulator; m) Hydrogen recovery; n) Ammonia converter; o) Refrigeration system

In the Canadian plant, only the air compressor is driven by a steam turbine, which receives the total steam generated in the plant and has an electric generator on the same shaft. All other consumers, including synthesis gas compressor, are driven by electric motors. Separate machines are used for make-up gas and recycle compression. The make-up gas compressor is located upstream of the methanator to make use of the compression heat to warm up the cold gas coming from the Selexol carbon dioxide scrubber.

A further key feature is that about half of the process steam is supplied by feed gas saturation. The synthesis converter is a three-bed design with quench between the first two beds and an exchanger after the second bed to raise the gas temperature of the feed to the first bed. Excess nitrogen and inerts (methane and argon) are removed by taking a purge gas stream from the circulator delivery and treating it in a cryogenic unit operating at loop pressure. The recovered hydrogen is returned to the circulator suction. Demonstrated efficiency is 28.5 GJ per tonne NH_3.

Foster Wheeler AM2 Process. The Foster Wheeler AM2 process [1033], [1087], also belongs to the group of processes that shift load from the primary to the secondary reformer, but differs from the preceding concepts in that only 20 – 50 % of the total feedstock is treated in the tubular primary reformer. The remaining feed is directly sent to the secondary (autothermal) reformer which operates with a high surplus of process air (up to 200 %) and a rather high methane slip of 2.75 % (dry basis). After conventional shift, further purification is performed by Selexol CO_2 removal, methanation, and molecular sieve drying. A cryogenic unit operating at synthesis pressure rejects the nitrogen surplus from the loop. An energy consumption of 29.3 GJ per tonne NH_3 is claimed.

Humphreys & Glasgow BYAS Process. The Humphreys & Glasgow (now Jacobs) BYAS process [1034], [1084], [1269] resembles the above-described processes in its principal process features: a considerable proportion of the feed is sent directly to the secondary reformer, bypassing the fired primary reformer; use of excess air in the secondary reformer; installation of a cryogenic unit as last step of make-up gas production to remove excess nitrogen, residual methane, and the majority of the argon. As a consequence the inert level in the loop can be kept rather low, with only a small withdrawal of purge gas. An energy consumption as low as 28.7 GJ per tonne NH_3 is claimed [1270]. The process is especially suited for revamps, where it allows plant capacity to be increased.

Jacobs Plus Ammonia Technology [1037] is especially tailored for small capacities in the 300 to 450 t/d range, with a load shift from primary to secondary reformer and use of excess process air. To produce a stoichiometric synthesis gas the surplus nitrogen has to be rejected in the final purification. This is done in a PSA unit, which receives the purge gas and part of the synthesis gas taken ahead of the methanation step. All nitrogen, methane, residual carbon oxides, and argon are adsorbed to give a stream of pure hydrogen. Hydrogen and the remainder of the synthesis gas downstream of methanation are mixed to achieve a 3:1 H_2:N_2 gas composition, with a lower inerts content than the synthesis gas after methanation. The consumption figure reported for a totally energy-balanced plant is 28.8 GJ per tonne NH_3, and with substantial steam export a value of 26.8 GJ per tonne NH_3 is claimed.

Montedison Low-Pressure Process. The Montedison low-pressure process [867], [1036], [1090], [1096] involves a split flow to two primary reformers. About 65 % of the feed – steam mixture flows conventionally through the radiant tubes of a fired primary reformer followed by a secondary reformer. The balance of the feed – steam mixture passes through the tubes of a vertical exchanger reformer. This exchanger reformer has a tubesheet for the catalyst tubes at the mixed feed inlet. There is no tubesheet at the bottom of the tubes, where the reformed gas mixes directly with the secondary reformer effluent. The combined streams flow on the shell side to heat the reformer tubes in a manner similar to that described for the M. W. Kellogg KRES reformer, see Section Processes Without a Fired Primary Reformer (Exchanger Reformer) and Feedstock Pretreatment and Raw Gas Production). The process air flow is stoichiometric. Synthesis is performed at 60 bar in a proprietary three-bed indirectly cooled converter with ammonia separation by water, from which ammonia is then recovered by distillation using low-grade heat. Other process steps are conventional. As driver of the process air compressor the installation of a gas turbine is suggested with use of the hot exhaust as preheated combustion air for the fired primary reformer. For this process, which has been tested in a 50 bar pilot plant, an energy consumption of 28.1 GJ per tonne NH_3 is claimed [1036].

Kellogg's LEAP Process. In the late 1970s Kellogg [288], [862], [867], [1089] proposed a process which extends the basic idea of the concept described above even further. The flow of the preheated gas stream mixture is split into three streams, with 47 % through catalyst tubes in the radiant section of the fired primary reformer, 12 % through catalyst tubes in the convection section, and 41 % through the tubes of an exchanger reformer heated by the effluent of a secondary reformer. It was intended to operate the ammonia synthesis at the pressure of the front end by using no synthesis gas compression or only a small booster. An enormous quantity of the classical ammonia synthesis catalyst would have been necessary, and for recovery of the ammonia from the loop a water wash with subsequent distillation was suggested, using low-level heat in an integrated absorption refrigerator. A consumption below 28 GJ per tonne NH_3 was calculated.

4.6.1.4.3. Processes Without a Fired Primary Reformer (Exchanger Reformer)

ICI LCA Process. The ICI LCA process [1091-1095] is a radical breakaway from the design philosophy of the highly integrated large plant used successfully for the last 25 years. Figure 69 shows a diagram of the so-called core unit which includes only the sections essential for ammonia production (up to 450 t/d). A separate utility section, shown in Figure 70, supplies refrigeration, steam, electricity, includes cooling and water-demineralization system, and, if needed, recovers pure carbon dioxide. Both figures show the configuration of the first two plants built at Severnside (UK) with a capacity of 450 t/d each.

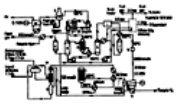

Figure 69. ICI LCA process (core unit)

a) Process air compressor; b) Start-up air heater; c) Hydrodesulfurization; d) Saturator; e) GHR; f) Secondary reformer; g) Shift converter; h) Desaturator; i) PSA system; j) Methanator; k) Gas dryer; l) Ammonia converter; m) Two-stage flash cooling (one stage shown); n) Chiller; o) Catchpot; p) Flash vessel q) Syngas compressor

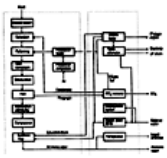

Figure 70. Arrangement of core units and utility section

Feed gas is purified in a hydrodesulfurization unit operating at lower than usual temperatures and passes through a saturator to supply a part of the process steam, while the balance is injected as steam. Heated in an inlet – outlet exchanger to 425 °C the mixed feed enters the ICI gas heated reformer (GHR) [461-463] at 41 bar, passing to the secondary reformer at 715 °C. The shell side entrance temperature of the GHR (secondary reformer exit) is 970 °C, falling to 540 °C at the exit of the GHR. Methane levels at the GHR exit and the secondary reformer are 25 % and 0.67 % respectively (dry basis). Overall steam to carbon ratio is 2.5 – 2.7. The gas, cooled to 265 °C in the inlet/outlet exchanger, enters a single-stage shift conversion reactor with a special copper – zinc – alumina-based catalyst that operates in quasi-isothermal fashion and is equipped with cooling tubes in which hot water circulates, whereby the absorbed heat is used for feed gas saturation, as described

above. CO_2 removal and further purification is effected by a PSA system, followed by methanation and drying.

Synthesis operates at 82 bar [1404] in a proprietary tubular converter loaded with a cobalt-enhanced formulation of the classical iron catalyst. Purge gas is recycled to the PSA unit, and pure CO_2 is recovered from the PSA waste gas by an aMDEA wash. Very little steam is generated in the synthesis loop and from waste gases and some natural gas in the utility boiler in the utility section (60 bar), and all drivers are electric.

The original intention was to design an ammonia plant which can compete with modern large capacity plants in consumption and specific investment, and, by means of lower energy integration, to achieve greater flexibility for start-up and reduced-load operation, needing minimum staffing. The basic plant features (GHR, isothermal shift, and synthesis) can in principal be applied to larger capacities. The flow sheet energy consumption is 29.3 GJ per tonne NH_3.

Kellogg, today Kellogg Brown and Root (KBR), offers several process schemes with use of its KRES exchanger reformer and its KAAP ammonia technology [467-470]. With including the Purifier (of C.F. Braun, later Braun and Root) KBR offers now the KAAPplusProcess [1329], [1330], [1379-1381] shown in Figure 71.

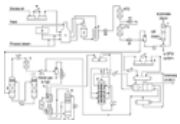

Figure 71. KBR KAAPplus Process

a) Air compressor; b) Sulfur removal; c) Process heater; d) Automatic reformer (ATR); e) Reforming exchanger (KRES); f) Condensate stripper; g) CO_2 absorber; h) Methanator; i) CO_2 stripper; j) Dryer; k) Expander; l) Feed/effluent exchanger; m) Condenser; n) Rectifier column; o) Synthesis gas compressor; p) KAAP ammonia converter; q) Refrigeration compressor; r) Refrigeration exchanger

Desulfurized gas is mixed with steam and then split into two streams in approximate ratio 2:1. These are separately heated in a fired heater. The smaller of the two enters the exchanger reformer at 550 – 550 °C, while the remainder is passed directly to the autothermal reformer at 600 – 640 °C. The exchanger reformer and the autothermal reformer use conventional nickel-based primary and secondary reforming catalysts, respectively. To satisfy the heat balance, the autothermal reformer operates with surplus of air. The required heat for the endothermic reaction in the tubes of the exchanger reformer comes from the gases on the shell side, comprising a mixture of the effluent from the autothermal reformer and the the gas emerging from the tubes. The shell side gas leaves the vessel at 40 bar. The purifier removes the nitrogen surplus together with residual methane and part of the argon.

Synthesis proceeds at about 90 bar in a four-bed radial-flow converter (Fig. 57) (hot-wall design) with interbed exchangers. The first bed is charged with conventional iron-based catalyst for bulk conversion and the others with Kellogg's high activity ruthenium-based catalyst, allowing an exit ammonia concentration in excess of 20 % to be obtained. The other process steps are more along traditional lines. The overall energy consumption claimed for this process can be as low as 27.2 GJ per tonne NH_3.

The LCA and KAAP*plus*Process are environmentally favorable because atmospheric emissions of both nitrogen oxides and carbon dioxide are dramatically reduced as there is no reforming furnace.

Chiyoda Process [1038]. In this process the traditional fired primary reformer is also replaced by an exchanger reformer and the heat balance requires excess air in the secondary reformer with the consequence of a cryogenic unit as final step in the make-up gas preparation to remove the surplus of nitrogen. Additionally, gas turbines are proposed as drivers for the process air compressor and synthesis gas compressor with the hot exhaust being used for steam generation and feed gas preheating.

Topsøe has described the lay-out for an ammonia plant based on fully autothemic reforming [1328].

4.6.1.4.4. Processes Without a Secondary Reformer (Nitrogen from Air Separation)

KTI PARC Process. The KTI PARC ammonia process [603], [1084], [1104], [1105], [1107-1110], [1260] uses the following process elements: air separation unit, classical primary reformer at 29 bar, standard HT shift, power generation in a Rankine cycle with CFC to generate electric power (optional), carbon dioxide removal (optional, only when pure CO_2 product is required), PSA, nitrogen addition, synthesis loop. In this concept four sections of the classical process sequence (secondary reforming, LT shift, CO_2 removal, methanation) can be replaced by the fully automatic high-efficiency PSA system, which has a proprietary configuration (UOP) in which nitrogen flushing enhances hydrogen recovery. The overall efficiency ranges from 29.3 to 31.8 GJ per tonne NH_3.

Linde LAC Process. The Linde LAC process [1112-1116] consists essentially of a hydrogen plant with only a PSA unit to purify the synthesis gas, a standard cryogenic nitrogen unit, and an ammonia synthesis loop. In principle it is similar to the PARC process, but designed for world-scale capacities. First application was for a 1350 t/d plant in India. Figure 72 compares the LAC process to a conventional one. If pure CO_2 is needed, it can be recovered by scrubbing the off-gas from the PSA unit at low pressure or, probably with better energy efficiency, by installing the CO_2 removal unit directly in the synthesis gas train ahead of the PSA system. The synthesis converter and loop are based on ICI and Casale know-how. According to Linde the process should consume about 28.5 GJ per tonne NH_3 or, with inclusion of pure CO_2 recovery, 29.3 GJ per tonne NH_3.

Figure 72. Comparison of Linde LAC process with a conventional process

Humphreys & Glasgow MDF Process (now Jacobs) [288], [665], [666], [1046], [1048], [1103], [1102], [1118], [1119]. This concept has a configuration similar to the Linde LAC process. Energy consumption with inclusion of pure CO_2 recovery (which is optional) is 32.8 GJ per tonne NH_3.

4.6.2. ***Ammonia Plants based on Partial Oxidation***

4.6.2.1. *Ammonia Plants Based on Heavy Hydrocarbons*

Although partial oxidation processes can gasify any hydrocarbon feedstock, on account of its higher energy consumption and investment costs, commercial use of this technology is restricted to the processing of higher hydrocarbons, often containing as much as 7 % sulfur. Where natural gas is unavailable or the heavy feedstock can be obtained at a competitive price, this gasification technology can be an economic choice.

There are two commercially proven partial oxidation routes for heavy feedstocks: the Shell process and the Texaco process. In contrast to the steam reforming, for which most contractors have their own proprietary technology for the individual process steps, the engineering firms which offer ammonia plants based on heavy hydrocarbons have often to rely for the individual process stages on different licensors. Lurgi, for example, has built very large capacity ammonia plants that use Shell's gasification process, its own proprietary version of the Rectisol process [1396], Linde's air separation and liquid nitrogen wash, and Topsøe's technology for synthesis converter and loop.

Independent, experienced engineering companies, not directly active in ammonia plant design may be used as general contractors to coordinate a number of subcontractors supplying the different technologies required. This is in line with the fact that the degree of energy integration is usually lower than in steam reforming technology, because in absence of a large fired furnace, there is no large amount of flue gas and consequently less waste-heat is available. Therefore, a separate auxiliary boiler is normally necessary to provide steam for mechanical energy and power generation. Nevertheless, some optimization has successfully reduced the overall energy consumption, for which in older installations values of around 38 GJ per tonne NH_3 were reported. More recent commercial bids quote values as low as 33.5 GJ per tonne NH_3.

The arguments presented above suggest describing the two principal routes, Shell and Texaco, which differ in the gasification process, rather than listing all individual contractor design approaches. Figure 73 shows the classical sequence of process steps for both cases.

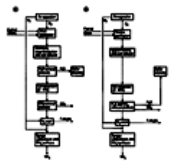

Figure 73. Flow diagrams of ammonia production from fuel oil or heavy residues by the Shell (A) and the Texaco (B) process (standard configuration)

Processes Using Shell Gasification (e.g., Lurgi) [471], [477-485], [492]. A cryogenic air separation plant provides oxygen for the gasification and the nitrogen for the liquid nitrogen wash and for supplying the stoichiometric amount for the synthesis of ammonia. Oil enters the alumina-lined gasification vessel through a central jet in the burner nozzle. A substantial pressure drop is needed to ensure atomization of the oil and proper mixing with oxygen, fed through the annulus between the jet and the outer case of the burner nozzle. The temperature in the gasification vessel (generator) is between 1200 and 1400 °C, and the pressure between 35 and 65 bar.

The hot gas contains soot, formed because of insufficient mixing of the reactants, and fly ash. A waste-heat boiler, a proprietary item of the Shell process, raises 100 bar steam and cools the gas to 340 °C. Soot is removed from the raw gas in a two-stage water wash. Older installations used an elaborate technique to remove the soot from the water by extraction with naphtha and light oil to form soot pellets which could be burnt or recycled to the feed oil. In newer installations the carbon – water slurry is filtered off in automatic filters, and the moist filter cake is subjected to a controlled oxidation in a multiple-hearth furnace.

A selective Rectisol unit with methanol of about –30 °C as solvent is used to remove H_2S and COS (together with some CO_2) to less than 0.1 %. (see also ⟶ Gas Production – Types of Processes). The removed sulfur-rich fraction is sent to a Claus plant for recovery of elemental sulfur or converted to sulfuric acid. The gas is heated subsequently by heat exchange, supplied with steam in a saturator, and then fed to shift conversion, which proceeds stepwise with intermediate heat removal. The gas is cooled by a direct water cooler, and the hot water is recycled to the saturator.

A second Rectisol wash stage follows to remove CO_2 by absorption at –65 °C in methanol, which is regenerated by flashing and stripping. Molecular sieve adsorption then removes residual traces of methanol and CO_2. To remove residual CO a liquid nitrogen wash is applied for final purification with the advantage of also lowering the argon content in the make-up gas, which is adjusted by nitrogen addition to the stoichiometric ratio $N_2 : H_2 = 1 : 3$. Converter and synthesis loop configuration depend on the licensor chosen. Plant descriptions are given in [1106, 1111, 1117, 1120].

Processes Using Texaco Gasification (e.g., Foster Wheeler, Linde, Uhde). Temperatures in the generator are similar to those in the Shell process; units with operating pressures up to 90 bar are in operation [422], [486], [487-491]. Some modern installations (e.g., Linde) use pumps for liquid oxygen instead of oxygen compressors. In contrast to the Shell arrangement, oxygen enters the gasifier through the central nozzle of the burner, and oil is fed through the annular space between central nozzle and outer burner tube. Instead of a waste-heat boiler a direct water quench is applied for cooling the raw synthesis gas, which is subsequently scrubbed first in a Venturi scrubber and then in a packed tower to remove the soot. Texaco also offers a version operating with a waste-heat boiler instead of a water quench. Although this is preferable when producing CO-rich synthesis gases (e.g., methanol or oxogas), quench is thought to be more economical when hydrogen-rich gases are manufactured. Soot recovery from the water is performed by extraction with naphtha. The soot – naphtha suspension is mixed with feed oil, and the naphtha is distilled off and recycled to the extraction stage. The shift reaction uses a cobalt – molybdenum – alumina catalyst [417], [418], [554], [555], [556], [638] which is not only sulfur-tolerant but also requires a minimum sulfur content in the gas for proper performance. The conversion is subdivided into stages with intermediate cooling by raising steam. The following Rectisol process has a somewhat more elaborate configuration than the version used in the Shell route. The large amount of carbon dioxide formed in shift conversion lowers the H_2S concentration in the sour gas, and for this reason a special concentration step is required for methanol regeneration to obtain pure CO_2 and a fraction sufficiently rich in H_2S for a Claus plant or a sulfuric acid plant. The remaining process steps are identical with the Shell route. Figure 74 gives an example of a Linde flow diagram. Descriptions of plants using the Texaco process can be found in [416], [418].

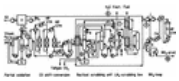

Figure 74. Ammonia production based on heavy fuel oil (Linde flow scheme with Texaco gasification)

a) Air separation unit; b) Soot extraction; c) CO_2 absorption; d) Methanol/H_2O distillation; e) Stripper; f) Hot regenerator; g) Refrigerant; h) Dryer; i) Liquid N_2 scrubber; j) Syngas compressor; k) NH_3 reactor

Material Balance

1		2		10	
Feed: heavy fuel oil		Oxygen		Ammonia	
m = 44 t/h		m = 49.6 t/h		m = 56.25 t/h	
p = 0.4 MPa		p = 9.0 MPa		p = 2.5 MPa	
t = 90 °C		t = 150 °C		t = 11 °C	
Composition, wt %		Purity			
C	84.6	O_2 = 98.5 vol %			
H	10.0				
O	0.4	balance: Ar and N_2			
N	0.4				
S	4.5				
Ash	0.1				

		3	4	5	6	7	8	9
H_2	vol %	43.0	60.8	96.3	0.9	0.1	–	75.00
CO	vol %	47.7	1.5	2.3	0.1	–	–	–
CO_2	vol %	6.9	35.8	–	98.8	67.4	–	–
N_2	vol %	0.1	0.1	0.2	0.1	2.5	100.00	25.00
CH_4	vol %	0.8	0.5	0.8	0.1	–	–	–
Ar	vol %	0.4	0.3	0.4	0.0	–	–	–
H_2S + COS	vol %	1.1	0.8	–	–	30.0	–	–
H_2O	vol %		0.2	–	–	–	–	–
kmol/h		5680	8250	5200	1800	200	1750	6650
p	MPa	8.37	7.65	7.30	0.15	0.25	7.4	7.1
t	°C	255	45	−63	−15	15	15	10

Topsøe Process. A concept using enriched air instead of pure oxygen and methanation instead of a liquid nitrogen wash was proposed by Topsøe [1121], [1122]. A Shell gasifier with a waste-heat boiler or a Texaco generator with a quench are equally well suited to this process. After soot removal, shift conversion is performed on a sulfur-tolerant catalyst in several beds with intermediate cooling, leaving a residual CO content of 0.6 mol %. An appropriate process (Rectisol or amine based) removes the sour gases H_2S, COS, and CO_2, and this is followed by methanation. Make-up gas drying, compression, and synthesis loop have no special features. The anticipated energy consumption is 34.8 GJ per tonne NH_3. A basically similar synthesis gas preparation, but based on gasification with pure oxygen, is already used in large commercial plant in Japan [1274].

Foster Wheeler Air Partial Oxidation Process [1033], [1123] is a proposed modification of the Texaco gasification process. It is intended to operate at 70 bar with highly preheated air (815 °C) instead of pure oxygen. The synthesis gas purification train comprises soot scrubbing followed by shift conversion, acid gas removal (for

example Selexol), and methanation. The gas is dried and finally fed to a cryogenic unit, which removes the surplus nitrogen by condensation together with methane, argon, and residual carbon monoxide. The rejected nitrogen is heated and expanded in a turbine, which helps to drive the air compressor.

A major aspect in the design concept is the separation of nitrogen and oxygen. A conventional air separation plant is based on the fractional distillation of oxygen and nitrogen, which differ in boiling point by only 13 °C. In the cryogenic unit for the Foster Wheeler process a lesser quantity of nitrogen is separated from hydrogen with a much higher boiling point difference (57 °C). According to Foster Wheeler this leads to considerable saving in capital investment and energy consumption compared to the traditional approach using pure oxygen from an air separation plant and a liquid nitrogen wash for gas purification. A figure of 35.6 – 37.6 GJ per tonne NH_3 is claimed for heavy oil feedstock and 31.4 – 32.7 for natural gas as feedstock. A similar variant also using air instead of pure oxygen is offered by Humphreys & Glasgow (now Jacobs) [1281].

4.6.2.2. *Ammonia Plants Using Coal as Feedstock*

In the early days the entire ammonia industry was based on coal feedstock. Today coal or coke (including coke oven gas) are used as feed for only a smaller part of world-wide ammonia production. In 1990, for example, only 13.5 % of the world ammonia capacity used this raw material [502]. A newer statistic estimates a figure of 19 % for 2001 [1387]. But with the enormous increase of the natural gas prices in the USA the share of coal in the feedstock pattern of ammonia might become larger in the future (see Chapter Economic Aspects).

Apart from a few plants operating in India and South Africa, the majority of coal-based ammonia plants are found in China. Commercially proven coal gasification processes are Lurgi (dry gasifier), British Gas/Lurgi (slagging gasifier), Winkler/HTW, Koppers – Totzek, Shell, Texaco, and Dow, [493], [494], [495], [505]. So far only the Koppers – Totzek, Texaco, and Lurgi processes have been used commercially for ammonia production [493], [498-501], [503], [1250, 1331, 1332, 1334], [1387]. The Shell process, demonstrated commercially in another application with a capacity equivalent to a world-scale ammonia plant, is also a potential candidate for ammonia production processes.

In recent years little development work has been done on complete ammonia plant concepts based on coal. The traditional leading ammonia contractors have to rely on proprietary processes licensed from different companies, which similarly tend not to have specific ammonia technology of their own. Again, compared to a steam reforming plant, the degree of integration is considerably lower; power generation facilities are usually separate. Thus it is difficult to identify specific ammonia processes for the individual contractors and the following descriptions serve as examples, without striving for completeness.

The Koppers – Totzek Process gasifies coal dust with oxygen in the temperature range 1500 – 1600 °C at about atmospheric pressure. For a more detailed description of the gasification, refer to Section Feedstock Pretreatment and Raw Gas Production and [504-510]. The cooled gas, free of coal dust and fly ash, contains about 60 % of CO. The next step is compression to about 30 bar, followed by sulfur removal at –38 °C with chilled methanol (Rectisol process). Steam is added for the shift conversion, carried out stepwise with intermediate heat removal and with a standard HTS catalyst. A second Rectisol stage, operating at –58 °C and 50 bar, removes the CO_2, and the final purification step is a liquid nitrogen wash. Any of the well known converter and synthesis loop concepts may be used, with no purge or minimal purge, due to the practically inert-free make-up gas. Several plants are operating in South Africa [1124], [1125] and India [1126].

The atmospheric-pressure gasification is a considerable disadvantage of this process route, which substantially increases equipment dimensions and costs, as well as the power required for synthesis gas compression. An energy input of 51.5 GJ per tonne NH_3 (LHV) has been reported. According to [537], the atmospheric ACGP gasifier could lower the consumption to 44 GJ per tonne NH_3 (HHV).

Lurgi Process. Lurgi [494], [495], [526, 527, 529, 530], [528, 1127] offers a concept using its proprietary Lurgi dry bottom gasifier, described in Section Feedstock Pretreatment and Raw Gas Production. The moving-bed generator, which can handle any sort of coal (ash content may exceed 30 %), operates at 30 bar, and the product gas contains up to 15 % CH_4, higher hydrocarbons, troublesome phenolic material, and tars. After washing with process condensates to remove ash and dust, the gas is cooled further with recovery of waste-heat. Several process steps treat the separated gas liquor to recover tar, phenols, and some ammonia. Shift conversion, Rectisol process, and liquid nitrogen wash are the further operations in the production of make-up gas. The liquid nitrogen wash produces a methane-rich fraction, which is separately processed in a steam reformer, and the reformed gas rejoins the main stream at the Rectisol unit for purification. The gasification has a power consumption of 32 – 34 GJ per tonne NH_3, and steam generation consumes 18 – 22 GJ per tonne NH_3, resulting in a total energy consumption of 50 – 56 GJ per tonne NH_3.

The Texaco Coal Gasification Process [495], [518-525], [1128-1131], [1272], [1273] (see Section Feedstock Pretreatment and Raw Gas Production) originates from Texaco's partial oxidation process for heavy oil fractions and processes a coal – water slurry containing 60 – 70 % coal. A lock hopper system removes ash and glassy slag as a suspension from the quench compartment of the generator. The process can handle bituminous and sub-bituminous coal but not lignite. The further gas purification steps used to arrive at pure make-up gas correspond to those described for an ammonia plant using the Texaco partial oxidation of heavy oil fractions.

Ube Industries commissioned a 1000 t/d ammonia plant in 1984 using Texaco's coal gasification process [1274], [1275]. An energy consumption of 44.3 GJ per tonne NH_3 is stated, which is lower than the 48.5 GJ per tonne NH_3 quoted for another Texaco coal gasification-based ammonia plant [494].

4.6.3. *Waste-Heat Boilers for High-Pressure Steam Generation*

Because of its great influence on reliability and efficiency of all ammonia plants a special section for the generation of high-pressure steam seems to be appropriate. The operating conditions for the boilers are more severe than those normally encountered in power plants; on account of the high pressure on both sides the heat transfer rates and thus the thermal stresses induced are much higher.

In steam reforming plants, for example, the temperature of the gas from the secondary reformer has to be reduced from 1000 °C to 350 °C before entering the HT shift vessel. In earlier plant generations two boilers were usually installed in series, with a bypass around the second to control the inlet temperature for the HT shift. Common practice for a long time was to use a boiler with water tube design. A famous example is the Kellogg bayonet-tube boiler, applied in more than 100 plants. Besause of size limitations two parallel units were installed. For sufficient natural water circulation these boilers needed a steam drum at a rather high elevation and a considerable number of downcomers (feed water) and risers (steam/water mixture). An alternative tube-bundle design which can directly substitute the bayonet-tube internals was recently developed. This concept uses twisted tubes [1378].

In contrast fire-tube boilers are much better suited for natural circulation and the steam drum can sit in a piggyback fashion right on top of the boiler. This makes it possible to provide each boiler with its own separate steam drum, which allows a greater flexibility in the plot plan. In a fire-tube boiler, the inlet tubesheet and the tubesheet welds are exposed to the extreme temperature of the reformed gas, which creates rather large temperature gradients and therefore high expansive stress. A positive feature of the design, however, is that debris in feed water (mainly magnetite particles spalling from the water side of the tubes) can collect at the bottom of the horizontally mounted vessel without creating diffficulties and are removed easily by blowdown. Water-tube boilers, especially bayonet-types, are very sensitive in this respect, because the deposits may collect in the lowest and most intensively heated part of the tube. In an extreme case of scaling, this may restrict the water flow to the point where boiling occurs irregularly (film boiling). The risk is overheating and tube failure.

The key factor which allowed the use of fire-tube boilers after the secondary reformer was the development of thin-tubesheet designs. Thick tubesheets in this kind of service are too rigid and have too high a temperature gradient, and the resultant stress on the tube-to-tubesheet welds can lead to cracks. The inherent flexibility of thin tubesheets assists in dispersing stresses and reduces the risk of fatigue failure of the tube-to-tubesheet welds and tubesheet-to-shell welds. In all the various designs of this concept, the tubesheet is only 20 – 30 mm thick. The hot inlet channel and the tubesheet are shielded by a refractory layer, and the tube inlets are protected by ferrules. In one concept (Uhde [840], [841], Steinmüller [1337]) the flexible tubesheet is anchored to and supported by the tubes to withstand the differential pressure, which poses some restriction on the tube length. The tubesheet of Babcock-Borsig (today Borsig) [837], [1338], [1339] is reinforced by stiffening plates on the back side (Fig. 75). Both solutions have full penetration tube-to-tubesheet welds. Steinmüller's waste-heat boiler product line is now owned by Borsig. A critical analysis of the two concepts is given in [1340].

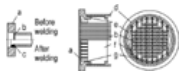

Figure 75. Reinforced tubesheet of the Borsig boiler

a) Tubesheeet; b) Tube; c) TIG welded root pass; d) Shell; e) Supporting ring; f) Stiffener plate; g) Anchor

In the synthesis loop boilers at the exit of the converter up to 50 % of the total steam is generated. As much as 1.5 t of steam per tonne ammonia, equivalent to 90 % of the reaction heat can be generated. For this service also fire-tube versions have been used, including Borsigs's thin-tubesheet design. But compared to the secondary reformer service, where the gas pressure is lower than the steam pressure, the conditions and stress pattern are different. In the synthesis loop boiler the opposite is the case with the result that the tubes are subjected to longitudinal compression instead of beeing under tension. Several failures in this application have

been been reported [842], [1341].

The thick-tubesheet concepts in various configurations are more generally accepted now. Proven U-type designs (Fig. 76) are available from Uhde [1342], [1353], Balcke – Dürr [833], [834], [1344], KBR [838], and Borsig [1347], [1345]. A horizontal synthesis waste-heat boiler was developed by Balcke – Dürr with straight tubes and thick tubesheets on both ends [835], [1346].

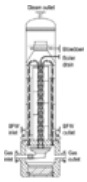

Figure 76. Uhde U-tube synthesis loop boiler

A special design is Borsig's hot/cold tubesheet. The hot and cold end of the tube are arranged alternately, so that a hot shank is always to a cold shank and vice versa. The advantage is that the tubesheet can kept below 380°C [836, 1318].

Waste-heat boilers in partial oxidation plants, which cool the exit gas of the generator from 1400 °C to around 350 °C, face additional difficulties. The gas contains soot and probably some fly ash particles. Very high gas velocities and appropriate design are necessary to prevent any deposition on the heat-exchange surfaces and to reduce the danger of attrition as well. A special design for Texaco gasifications (orginally by Schmidt'sche Heißdampf GmbH and Steinmüller GmbH) is offered now in an improved version (Fig. 77) by Borsig [1348]. Forced water circulation around the entrance nozzles helps to stand the high heat flux (700 kJ $m^{-2}s^{-1}$) at this location. The Shell process uses propretary designs (Fasel – Lentjes) [1349].

Figure 77. Borsig syngas waste-heat boiler for Texaco gasification

Additional information on waste-heat boilers is found in [762], [767], [404], [406], [1330]. Many papers on this subject were presented at the “AIChE Annual Symposia for Safety In Ammonia Plants and Related Facilities” [1350].

4.6.4. *Single-train Capacity Limitations – Mega-Ammonia Plants*

The principle of the economy of scale is well known in production processes. For ammonia plants a scale exponent 0.7 is widely used as a rule of thumb:

$$\mathrm{cost}_2 = \mathrm{cost}_1 \times (\mathrm{capacity}_2/\mathrm{capacity}_1)^{0.7}$$

The advantage of larger plants is not only the investment but also the reduction of fixed costs, such as labor, maintenance, and overhead expenses. But the simple formula should be handled with care because it is only valid as long as a given process configuration is scaled up and the dimensions of the equipment for very large capacities do not lead to disproportionate price increases. For example there may be fewer vendors or special fabrication procedures might become necessary. So the scale factor could increase for very large single-train capacities. Nevertheless the specific investment will decrease with the increase of capacity. On the other hand certain process developments like autothermal reforming and exchanger reforming could reduce the investment for the higher capacities.

For export-orientated plants at locations with cheap natural gas there is growing interest in very large capacities. Several studies [1317], [1355], [1381-1386] have been made by various contractors to investigate the feasibility and the size limitations for steam reforming plants. According to most of these studies a capacity of 5000 t/d seems to be possible. With a conventional process concept for the front end there is no restriction for producing syngas for 5000 t/d of ammonia. Steam reforming furnaces of the required size have already been built for other applications (methanol) and neccessary vessel diameters seem also to be fabricable. A limitation could be the synthesis compressor which can probably be designed for 4500 t/d only. The KBR KAAP process has there an advantage because of its low synthesis pressure. This also holds true for the Dual Pressure Process of Uhde, where just the loop which produces only two thirds of the ammonia is at high pressure whereas the once-through converter is at lower pressure. Another capacity limiting factor is pipe diameter in the synthesis section. The largest plant so far was built by Uhde in Saudi-Arabia. It has a capacity of 3300 t/d and came on-stream in 2006.

4.7. Modernization of Older Plants (Revamping)

With rising feedstock prices and hard competition in the market, many producers have looked for possibilities to “revamp” or modernize their older, less efficient plants so that they can stay competitive. Most revamp projects have been combined with a moderate capacity increase because some of the original equipment was oversized and only specific bottlenecks had to be eliminated, not entailing excessive cost. As the market possibilities for a company do not increase in steps of 1000 or 1500 t/d but slowly and continuously, such a moderate capacity addition will involve less risk and will be more economical than building a new plant.

For a revamp project first an updated base-line flow sheet of the existing plant should be prepared from which the proposed improvement can be measured [1002-1004]. Depending on the objective (energy saving and/or capacity increase) the following guidelines should be kept in mind: maximum use of capacity reserves in existing equipment; shifting duties from overtaxed units to oversized ones; if possible, simple modifications of existing equipment are preferable to replacement; the amount of additional equipment should be kept to a minimum [1003].

To give an exhaustive list or description of the individual modification options is beyond the scope of this article, but reviews on this subject and useful information are given in [1005-1017], [1098-1101], [1254], [1391], [1403]. Section Further Developments describes modifications that lower the energy consumption in newer plant generations compared to the first generation of single-train ammonia plants, and this also represents an overview of the revamp options for existing steam reforming plants.

Just a few of the frequently used revamp possibilities should be mentioned here. In steam reforming plants it is often possible to lower the steam/carbon ratio by using improved reforming catalysts and copper-promoted HT shift catalysts. More active LT shift catalysts lower the residual CO content, which will reduce H_2 loss (methanation) and inert content in the make-up gas. Drying of the make-up gas, addition of hydrogen recovery from purge gas, and installing a more effective CO_2 removal are other options. With the *aMDEA* system, which can be flexibly tailored to fit into existing process configurations, it is, for example, possible to simply replace the MEA solvent with the aMDEA solution, adjusting the activator concentration accordingly to achieve zero or only minor equipment modification. Also hot potash processes have been converted in this way to aMDEA.

Other measures, involving more additional hardware and engineering work, are introduction of combustion air preheating and reducing the primary reformer load. This latter option is used when the revamp objective is capacity increase and the primary reformer is identified as a bottleneck. One possibility is to increase the duty of the secondary reformer and use air in excess of the stoichiometric demand. Elegant variants of this principle are the *Jacobs BYAS process* [1034], [1084], [1269] and the *Foster Wheeler AM2 process* [1033], [1087], [1088]. Another method is to perform a part of the primary reforming in a pre-reformer [431-438] that uses low-level heat. Alternative methods to enlarge the reforming capacity make use of the process heat of the secondary reformer in an exchanger reformer such as *ICI's GHR* [1031] or *Kellogg's KRES* [469], [1398]. If oxygen is available, installation of a parallel autothermal reformer or a parallel *Uhde CAR unit* [472-476], (see also Section Feedstock Pretreatment and Raw Gas Production) could be considered. Description of executed revamp projects are given in [1270], [1019-1024].

Similarly, numerous modernization possibilities exist for partial oxidation plants, and they may even outnumber those for steam reforming plants. Common to both plant types is the potential for improvement of the synthesis loop and converter. Application of indirect cooling and smaller catalyst particles are frequently chosen to reduce energy consumption through lower pressure drop, reduced synthesis pressure, higher conversion, or a combination thereof. Apart from replacing existing ammonia converters, in situ modification of the internals of installed converters is a very economic approach. *Topsøe* [1011], [1026] mostly uses its Series 200 configuration; and *Ammonia Casale* [767-769], [1027], [1028], [813-816] its proprietary ACAR technology, with which more than 90 plants have been revamped so far.

With a “split-flow” configuration Kellogg proposed an in situ revamp option to change its four-bed quench converter into a two-stage intercooled converter (with two parallel beds for the second stage) using smaller catalyst particles of 1.5 – 3.5 mm and 3 – 6 mm [1293], [1294]. A promising loop modernization option for the future will be the *Kellogg KAAP process* [470], [1029], [1398].

The fact that more than 45 % of world ammonia plants are older than 30 – 35 years suggests that there is a major potential for revamp projects, even in plants which have already made modifications. Revamp histories are a constant topic at the “AIChE Annual Symposium Ammonia Plant and Related Facilities” [1350] and a wealth of information and practical experience can be found there.

A special revamping option is the integration of other processes into an ammonia plant [1318]. For example a CO production from a side stream upstream of the HT shift is possible [1356]. Hydrogen can also be produced from a side stream by using PSA. An example of coproduction of methanol is shown in Figure 78.

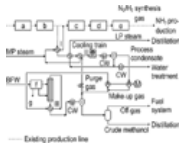

Figure 78. Methanol coproduction (side-stream loop)

a) Secondary reformer; b) Waste-heat boiler; c) HT shift; d) LT shift; e) Purification; f) Steam drum; g) Methanol converter; h) Catchpot; I) Syngas compressor; j) Let-down vessel

In the less integrated partial oxidation plants coproduction schemes are easier to be incorporated and there are several large installations which were directly designed to produce ammonia, methanol and hydrogen [1357], [1358].

4.8. Material Considerations for Equipment Fabrication

Hydrogen Attack. In several steps of the ammonia production process, especially in the synthesis section, the pressure shells of reaction vessels as well as the connecting pipes are in contact with hydrogen at elevated pressure and temperature with a potential risk of material deterioration [872-874].

Chemical Hydrogen Attack. Under certain conditions chemical hydrogen attack [875], [876], [878], [1408] can occur. Hydrogen diffuses into the steel and reacts with the carbon that is responsible for the strength of the material to form methane, which on account of its higher molecular volume cannot escape. The resulting pressure causes cavity growth along the grain boundaries, transforming the steel from a ductile to a brittle state. This may finally reach a point where the affected vessel or pipe ruptures, in most cases without any significant prior deformation. This phenomenon was already recognized and principally understood by BOSCH et al. [93] when they developed the first ammonia process. The resistance of steel against this sort of attack can be enhanced by alloy components which react with the carbon to form stable carbides (e.g., molybdenum, chromium, tungsten, and others). The rate of deterioration of the material depends on the pressure of the trapped methane, the creep rate of the material, and its grain structure. Areas highly susceptible to attack are those which have the greatest probability of containing unstable carbides, such as welding seams [877]. The type of carbides and their activity are strongly influenced by the quality of post-weld heat treatment (PWHT). The risk of attack may exist at quite moderate temperatures (ca. 200 °C) and a hydrogen partial pressure as low as 7 bar.

Numerous studies, experiments and careful investigations of failures have made it possible to largely prevent hydrogen attack in modern ammonia plants by proper selection of hydrogen-tolerant alloys with the appropriate content of metals that form stable alloys. Of great importance in this field was the work of NELSON [871], [878], [879], who summarized available experimental and operational experience in graphical form. These Nelson diagrams give the stability limits for various steels as a function of temperature and hydrogen partial pressure. In [880], [881], CLASS gives an extensive survey, still valid today, on this subject. Newer experience gained in industrial applications required several revisions of the original Nelson diagram. For example, 0.25 and 0.5 Mo steels are now regarded as ordinary nonalloyed steels with respect to their hydrogen resistance [875].

Physical Hydrogen Attack. A related phenomenon is physical hydrogen attack, which may happen simultaneously with chemical attack. It occurs when adsorbed molecular hydrogen dissociates at higher temperatures into atomic hydrogen, which can diffuse through the material structure. Wherever hydrogen atoms recombine to molecules in the material structure (at second-phase particles or material defects such as dislocations) internal stress becomes established within the material. The result is a progressive deterioration of the material that lowers its toughness until the affected piece of equipment cracks and ultimately ruptures.

The phenomenon is also referred to as hydrogen embrittlement. It is most likely to occur in welds that not received proper PWHT. Holding a weld and the heat-affected zone for a prolonged period at elevated temperature (an operation known as soaking) allows the majority of included hydrogen to diffuse out of the material. But this may not be sufficient if moisture was present during the original welding operation (for example if wet electrodes or hygroscopic fluxes were used), because traces of atomic hydrogen are formed by thermal decomposition of water under the intense heat of the welding procedure. Highly critical in this respect are dissimilar welds [1406], such as those between ferritic and austenitic steels [886], where the formation of martensite, which is sensitive to hydrogen attack, may increase the risk of brittle fracture.

At higher temperature and partial pressure, hydrogen is always soluble to a minor extent in construction steels. For this reason it is advisable not to cool vessels too rapidly when taking them out of service, and to hold them at atmospheric pressure for some hours at 300 °C so that the hydrogen can largely diffuse out (soaking). In contrast to the hydrogen attack described above this phenomenon is reversible.

It has been reported [882], [883] that hot-wall converters in which the pressure shell is in contact with ammonia-containing synthesis gas at 400 °C have developed cracks in circumferential welding seams to a depth nearly approaching the wall thickness in places. This was surprising because the operating conditions for the material (2.25 Cr/1 Mo) were well below the Nelson curve. One investigation [884] concluded that hydrogen attack had occurred by a special mechanism at temperatures lower than predicted by the Nelson diagram. Nitriding proceeding along microcracks could transform carbides normally stable against hydrogen into nitrides and carbonitrides to give free active carbon, which is hydrogenated to methane. High residual welding stress and internal pressure are considered to be essential for the propagation of the cracks. A rival theory [883], [885] attributes the damage to physical hydrogen attack resulting from the use of agglomerated hygroscopic flux in combination with insufficient soaking. However, for a longer time no problems were experienced for the same converter design in other cases where non-hygroscopic flux was used for welding and a more conservative vessel code was adopted, leading to less stress on account of thicker walls. But in the meantime also these converters developed cracks. So there are still questions left regarding the two theories.

Nitriding is a problem specific to the ammonia converter. It occurs in the presence of ammonia on the surface of steel at temperatures above 300 °C [888], [889-891]. With unalloyed and low-alloy steels, the nitride layer grows with time to a thickness of several millimeters. Austenitic steel, used for the converter basket, develops very thin but hard and brittle nitride layers, which tend to flake off. In the nitrided areas, the risk of formation of brittle surface cracks exists.

Temper Embrittlement. For heat-resistant steels long-term service at temperatures above 400 °C (e.g., high-pressure steam pipes) can lead to a decline in impact strength [1405]. Normally, transition temperatures (precipitous decline of notched bar impact values) of below 0 °C are encountered, but this can increase to 60 °C and more (temper embrittlement).

The susceptibility to temper embrittlement can be reduced by controlling the level of trace elements (Si, Mn, P, Sn) in steels [892], [893]. Vessels for which temper embrittlement is anticipated should not be pressurized below a certain temperature.

Metal Dusting. Metal dusting [404], [894-900] [1318], [1388] is a corrosion phenomenon which has come into focus again in the last few years with the introduction of exchanger reformer technology and the operation of steam superheaters in the hot process gas downstream of the secondary reformer waste-heat boiler. Conventional carburization is a familiar problem with high-temperature alloys in steam reforming furnaces caused by inward migration of carbon leading to formation of carbides in the metal matrix. It happens at high temperatures, typically above 800 °C, and the carbon originates from cracking of hydrocarbons. In contrast, metal dusting occurs at 500 – 800 °C on iron – nickel and iron – cobalt based alloys with gases containing carbon monoxide. The Boudouard reaction, strongly catalyzed by iron, nickel, and cobalt, is generally regarded as the source of the carbon in this case. It is assumed that thermodynamically favored sites exist for these elements at the surface and enhance the carbon deposition if the gas composition corresponds to a carbon activity >1 [898], [899].

As the name implies, the affected material disintegrates into fine metal and metal oxide particles mixed with carbon. Depending on the defects in a protective oxide film on the metal surface and the ability of the material to sustain this film, an induction period may be observed until metal dusting manifests itself as pitting or general attack. A possible mechanism was proposed by GRABKE [902] and HOCHMANN [903].

At least from a theoretical point of view, alloys formulated to form chromium, aluminum, or silicon oxide films should exhibit an increased resistance. Efforts to find solutions for this problem are continuing, but at present the following situation must be accepted: virtually all high-temperature alloys are vulnerable to metal dusting; higher steam/carbon ratios reduces this sort of corrosion; improvements may be expected by additional surface coating (for example with aluminum). With materials such as Inconel 601, 601H, 625 and similar alloys it is at least possible to reduce the attack to a level which is tolerable in practical operation. The active sites at the

metal surface which catalyze the Boudourd reaction can be poisoned by H_2S, thus inhibiting the initiation of metal dusting [899].

Hydrogen Sulfide Corrosion. Corrosion by hydrogen sulfide in partial oxidation plants can be controlled by the use of austenitic steels, but special care to ensure proper stress relief of welds is advisable to avoid stress corrosion cracking in these plants caused by traces of chlorine sometimes present in the feed oil.

Stress Corrosion Cracking. Stress corrosion cracking (SCC) of many steels in liquid ammonia is a peculiar phenomenon. It occurs at ambient temperature under pressure as well as in atmospheric storage tanks at –33 °C. Extensive studies [887], [901], [904-914] have defined the conditions under which SSC in liquid ammonia is likely to occur and how it may largely be prevented, but the mechanism is so far not fully understood. Preventive measures include maintaining a certain water content in the ammonia and excluding even traces of oxygen and carbon dioxide. Welds in pressurized vessels must be properly stress-relieved, and in atmospheric tanks it is important to select the appropriate welding electrodes, avoid excessive differences of the thickness of plates welded together and choose the correct geometry of welding seams.

[Top of Page]

5. Storage and Shipping

Producing and processing ammonia requires storage facilities to smooth out fluctuations in production and usage or shipments. If manufacture and use occur in separate locations, then appropriate transport must be arranged [1154]. This may be by ocean-going ships, river barges, rail or tank cars, or by pipeline. Ammonia is usually handled in liquid form, but in some cases delivery of ammonia vapor to downstream consumers on site may have some advantage due to savings of refrigeration energy in the ammonia plant. If there is an opportunity for on-site usage or marketing of aqueous ammonia, obtained as a byproduct of purge-gas scrubbing or, less frequently, deliberately produced, storage and handling facilities for this product will also be needed, but compared to liquid ammonia the demand is negligible.

Liquid ammonia is a liquefied gas. Its storage and distribution technologies therefore have much in common with other liquefied gases. Reference [402, vol. IV] summarizes the literature on storage, handling, and transportation of ammonia.

5.1. Storage

Three methods exist for storing liquid ammonia [1155], [1156]:

1. Pressure storage at ambient temperature in spherical or cylindrical pressure vessels having capacities up to about 1500 t
2. Atmospheric storage at – 33 °C in insulated cylindrical tanks for amounts to about 50 000 t per vessel
3. Reduced pressure storage at about 0 °C in insulated, usually spherical pressure vessels for quantities up to about 2500 t per sphere

The first two methods are preferred, and there is growing opinion that reduced-pressure storage is less attractive [402]. However, in many cases, the combination of atmospheric and pressure storage may be the most economic concept (Fig. 79). The determining factors for the type of storage — apart from the required size — are temperature and the quantity of ammonia flowing into and out of storage [1159].

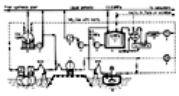

Figure 79. Ammonia terminal with loading and unloading facilities

a) Sphere at ambient temperature; b) Tank at ambient pressure (refrigerated); c) Pumps; d) Compressors; e) Air-cooled condenser; f) Barge with pressure tank; g) Barge for ammonia at ambient pressure; h) Booster; i) Rail car or truck; k) Heater

Characteristic features of the three types of storage are summarized in Table 30[404]. For pressurized storage, spheres with a capacity up to 1500 t have been constructed, in which case the ratio tonnes ammonia per tonne of steel will become about 6.5.

Table 30. Characteristic features of ammonia storage tanks

Type	Typical pressure, bar	Design temperature, °C	t ammonia per t steel	Capacity, t ammonia	Refrigeration compressor
Pressure storage*	16 – 18	ambient	2.8	< 270*	none
Semi-refrigerated storage	3 – 5	ca. 0	10	450 – 2700	single stage
Low-temperature storage	1.1 – 1.2	– 33	41 – 45	4500 – 45 000	two-stage

* Refers to cylindrical tanks ("bullets"); spherical vessels see Section Pressure Storage.

5.1.1. *Pressure Storage*

This system is especially suitable for:

- Storing small quantities of ammonia
- Balancing production variations with downstream units processing pressurized ammonia
- Loading and unloading trucks, tank cars, and marine vessels carrying pressurized ammonia
- Entrance to or exit from pipeline systems

Usually, cylindrical pressure vessels are designed for about 2.5 MPa. The larger spherical vessels are designed only for about 1.6 MPa to avoid wall thicknesses above 30 mm. A coat of reflecting paint or, frequently in hot climates, an outer covering of insulation may be used on the vessels to avoid solar radiation heating. Spraying the vessel with water is very effective against intense solar radiation but does stain and damage the paint. As a rule, liquid ammonia fed to storage from a synthesis loop carries inert gases with it. Besides the prescribed safety relief valves, pressure control is provided for the storage drum by controlled bleeding of the inert gases through a pressure-reducing valve, for example, into a water wash system (Fig. 79).

Sometimes high-strength or fine-grained steels are used in making pressure vessels. These may be susceptible to stress corrosion cracking by ammonia. Safely avoiding this hazard requires careful thermal stress relief after completing all welding on the vessel.

The shape of the vessel depends above all on its capacity:

Cylindrical, usually horizontal for up to about 150 t
Spherical vessels resting on tangentially arranged support columns but also more recently, for static and safety reasons, in a suitably shaped shallow depression for about 250 – 1500 t.

Ancillary equipment, designed for at least 2.5 MPa (25 bar): meters and flow controls for pressurized ammonia feed and effluent streams; centrifugal pumps for discharging into liquid ammonia supply piping and for liquid ammonia loading; equipment for safe pressure relief for ammonia vapor and inerts (see Fig. 79 and [1156]). Design of pressure storage tanks and the related safety aspects are discussed in [1158].

Stress corrosion cracking in pressurized ammonia vessels and tanks is a problem which has been discussed in many papers [901], [904-922]. The mechanism of this phenomenon, the influence of water, and the role of oxygen are not yet completely understood, in spite of extensive research. A review is given in [923]. As it is generally accepted that addition of water may inhibit stress corrosion [402], [914], [915] it has become a widely used practice to maintain a water content of 0.2 % in transport vessels [915]. Protection may also be achieved by aluminum or zinc metal spray coating [921], [923]. More recent research [919], [923], however, has

shown that water may not give complete protection.

The prevailing opinion was that stress corrosion should not occur in atmospheric storage tanks [1162]. Therefore, it was somewhat surprising when cases of stress corrosion in atmospheric ammonia tanks were reported [925-927]. Descriptions of further incidents, inspection techniques, and repair procedures can be found in [922], [928-935], [1160].

In 1995, the capital investment, including ancillary equipment, for a 1000-t pressurized ammonia storage facility amounted to about $ 3.5 × 10^6.

5.1.2. *Low-Temperature Storage*

Modern single-train plants need to have large-volume storage facilities available to compensate for ammonia production or consumption outages. Storage equivalent to about 20 days is customary for this. Moreover, transporting large amounts of ammonia by ship or pipeline and stockpiling to balance fluctuations in the ammonia market have great importance. Therefore, correspondingly large storage facilities are necessary at ship and pipeline loading and unloading points.

For comparable large storage volumes, the capital investment costs for atmospheric pressure storage are substantially lower than for pressure storage. In spite of higher energy costs for maintaining the pressure and for feed into and out of atmospheric storage, it is still more economical than pressure storage. This applies especially to storage of ammonia coming from the synthesis loop at low temperature and loading and unloading of refrigerated ships.

For atmospheric storage at – 33 °C single tanks with a capacity up to 50 000 t of ammonia are available [1160]. Design pressure is usually 1.1 – 1.5 bar (plus the static pressure of the ammonia). The cylindrical tanks have a flat bottom and a domed roof and are completely insulated. Refrigeration is provided by recompression of the boil-off, usually with two-stage reciprocating compressors. Incoming ammonia of ambient temperature is flashed to –33 °C before entering the tank, and the vapors from the flash vessel are also fed to the refrigeration unit. Refrigeration units have at least one stand-by unit, mostly powered by a diesel engine. Single-wall and double-wall tanks are used in the industry.

Single-Wall Tanks. The single-wall tank has a single-shell designed for the full operating pressure. Mats or panels of rock wool or foamed organics (e.g., in-situ applied polyurethane foam) are used for insulation of single wall tanks. The outside insulation must be completely vapor tight to avoid icing and requires the highest standards of construction and maintenance to avoid hazardous deterioration by meteorological influences. A metal sheet covering is normally applied, and sometimes a bond wall of reinforced concrete or steel is added [1163], [1164].

Double-wall tanks are known in various designs. In the simplest version an inner tank designed for storage temperature and pressure is surrounded by a second tank. The annular space between the two walls is filled with insulation material, for example, Perlite. The main purpose of the outer shell is to contain and hold the insulation. Today's usual practice is to design the outer shell to the same standard as the inner shell. This so-called double-integrity tank concept provides an additional safety measure as the outer tank can hold the full content if the inner shell fails.

Further safety provisions include surrounding the tanks by dikes or placing them in concrete basins to contain the liquid ammonia in the event of a total failure. Discussions of such secondary containment are found in [924], [1163], [1164], [1237].

Two principle “double integrity” designs are used as shown in Figure 80. In Figure 80 A and B the inner tank has a solid steel roof and is pressure-tight, whereas the gap between inner and outer shell, which can withstand the full hydrostatic pressure at operating temperature, is only covered by the water-tight insulation. The insulation for the inner shell may fill the total annular space [e.g., lightweight concrete [1165] or Perlite (B)] or may consist of an organic foam attached to the outside of the inner shell (A).

In the design according to Figure 80 C the outer tank is pressure-tight. The inner tank has a suspended roof and the insulation fills the annular gap, which contains an atmosphere of ammonia vapor.

Because of the static loading, cellular glass or lightweight concrete [1165] is used for the floor insulation in all cold tank systems. A foundation as shown in Figure 80 A and B requires underground heating to prevent formation of a continuous ice sheet under the tank, which could lift it. If, as in Figure 80 C, the tank bottom is on piles above ground level, heating is unnecessary.

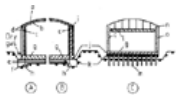

Figure 80. Arrangements of refrigerated ammonia tanks

a) Waterproof roofing; b) Steel roof; c) Inner steel shell; d) Foamed organic materials or rock wool; e) Heating coils; f) Stay bolts; g) Inner steel base; h) Concrete ring wall; i) Loose Perlite or lightweight concrete and rock wool; j) Outer steel shell and base; k) Cellular glass or lightweight concrete; l) Suspended deck with insulation; m) Pile foundation; n) Steel roof and shell or prestressed concrete lined for vapor containment; o) Loose Perlite or insulation on liner or in space

The tanks are fabricated by welding on site, from steels which retain their notch ductility strength at low temperature. Corresponding to the decreasing static pressure, steel plate thickness in the cylindrical shell is reduced stepwise from bottom to top. Technical details and design questions are treated in [402 Vol IV], [1161], [1237], [1242], [1243]. Influence of climatic conditions on design and operation is considered in [1244]. Specific cases and foundation problems are discussed in [1239-1242]. Retrofitting of existing storage tanks is described in [1245]. Maintenance and inspection procedures are covered in [1247-1253]. For stress corrosion occurred in atmospheric tanks see Section Pressure Storage.

A rough budget figure of the capital investment in 2005 for a double-integrity tank with a capacity of 45 000 t including cooling compressors is around $ 10 × 10^6. Inclusion of facilities to receive or deliver pressurized ammonia at ambient temperature would add a few millions more. Energy consumption would also increase considerably, to about 80 kWh/t ammonia throughput. In addition, substantial investment will be necessary for a complete terminal including, for example, a jetty and equipment for ship loading and facilities for loading rail and/or tank cars.

5.1.3. *Underground Storage*

For years, the liquefied petroleum gas (LPG) industry has used pressurized underground liquefied gas storage. This technique has been applied to ammonia also. DuPont has operated a rock cavern in the United States with a capacity of 20 000 t. Norsk Hydro has one in Norway at 50 000 t. Because of the contaminants occurring in liquid ammonia stored this way and the lack of suitable construction sites, no further storage facilities of this kind have been built for a long time.

5.1.4. *Storage of Aqueous Ammonia*

To allow storage and transportation at temperatures up to 35 °C, the concentration of aqueous ammonia should not exceed 25 %, because of its vapor pressure. A small facility uses mostly cylindrical vessels. The capacity of the transporting equipment determines the storage volume, e.g., at least 40 m^3 storage volume for 30 m^3 tank cars. For larger capacities, tanks are used. If it is necessary to avoid contaminating the aqueous ammonia by iron hydroxide, austenitic (stainless) steels may be used instead of the usual carbon steels.

5.2. Transportation

Transportation in Small Containers. The most common containers are:

- Cylindrical steel bottles and pressurized flasks for about 20 – 200 kg anhydrous ammonia to meet the requirements of laboratories, small refrigeration systems, and the like
- Polyethylene canisters, metal casks, and the like for 25 % aqueous ammonia.

Transportation in Trucks and Rail Cars with Capacities Normally up to 100 m^3; Jumbo Rail Cars Holding up to 150 m^3. The most widely used methods are:

- With pressure vessels for anhydrous ammonia (maximum allowable operating pressure about 2.5 MPa)
- With atmospheric pressure vessels for 25 % aqueous ammonia
- With vessels designed for elevated pressure for high-concentration aqueous ammonia (maximum allowable operating pressure in accordance with the ammonia content up to 1.6 MPa).

A modern rail car loading station for anhydrous and aqueous ammonia is described in [1167], [1168].

The distribution in rail cars and trucks primarily serves to supply smaller processing operations and wholesale merchants. However, rail transport of liquid ammonia may to some extent supplement large marine and pipeline shipments. Reference [1169] examines rail freight cost. Normally, shipping liquid ammonia by truck is used only where other means of transport are not available, e.g., in the agricultural practice of direct fertilization.

Shipping in Ocean-Going Vessels and River Barges. Regarding the transport volume, shipping of anhydrous ammonia is far more important than transport by rail. In 2004, for example, a total of 13 × 10^6 t of anhydrous ammonia was transported by ocean-going vessels (IFA statistics). Overseas shipping gained great momentum through exports from producers in countries with low natural gas prices or low-price policies [1181]. A comparison of shipping costs is given in [1173]; for more up-to-date information consult the journal *Nitrogen* (later *Nitrogen & Methanol*, today *Nitrogen and Syngas*), published by British Sulphur, which regularly reports shipping prices. Most river barges have loading capacities of 400 to 2500 t and mostly have refrigerated load, but a few are pressure vessels. Ocean-going vessels may transport as much as 50 000 t of fully refrigerated ammonia.

Pipelines. Transport of large volumes of ammonia by pipeline [1173], [1157] over great distances is far more economical than by river barge or rail. In the USA the MidAmerica Pipeline System transports ammonia to terminals in Kansas, Nebraska, and Iowa, which are intensive agricultural areas. The Gulf Central Pipeline connects the major producers along the Texas and Louisiana Gulf coast with terminals in Iowa, Illinois, Nebraska, and Missouri. [1189]The MidAmerica Pipeline, for example, has a peak delivery capacity of 8000 t/d for a number of destinations [1170], the hold up is about 20 000 t. Maintenance is described in [1171].

As ammonia is transported at a temperature of at least 2 °C it has to be warmed up at the supplier terminal and cooled down again to – 33 °C at the receiver terminal. Exact knowledge of the $p - V - T$ properties is important [1172]. Automatic lock valves are installed at intervals of 10 miles, so that the volume which can be released between two valves is limited to 400 t.

The world's longest ammonia pipeline has been in operation since 1983 in Russia, connecting the large production facilities Togliatti/Gordlovka with the terminals Grigorowski/Odessa at the Black Sea over a distance of 2424 km [1166]. Apart from some short pipeline connections, most of them shorter than 50 km, there is no significant pipeline transport in Europe, where ammonia is predominantly further processed in downstream facilities on site and where no widespread direct application to agricultural crops exists.

[Top of Page]

6. Quality Specifications and Analysis

The quality of the ammonia product depends to some extent on the operating conditions of the production plant and storage. For example, water content from a synthesis loop receiving a dry make-up gas is about zero, whereas plants which receive the feedgas after a methanation without further drying may give a water content of 0.1 – 0.2 %. Oil may be introduced by the seal oil of the synthesis compressor, but on account of its low solubility in liquid ammonia it usually settles out on storage, so that only a minor concentration will remain in commercial deliveries. There are two commercial qualities of anhydrous ammonia: commercial (technical) grade (ammonia as received from production and storage) and refrigeration grade (technical product purified by distillation [1197] or molecular sieve adsorption). Table 31 lists the commercial specifications. The relevant standards are US Specification OA-445a, Supplement 5 (1963) and, in Germany, DIN 8960 (1972) for refrigeration-grade ammonia. To inhibit stress corrosion cracking a water content of at least 0.2 % for shipped and pipelined liquid ammonia is generally recommended and is mandatory in the USA.

Table 31. Minimum quality requirements for ammonia

Quality		Commercial grade		Refrigeration grade	
		USA	Germany	USA	Germany
Purity	wt %, min	99.5	99.5	99.98	99.98 *
Water	wt %, max	0.5	0.2	0.015	0.2
Inerts **	mL/g, max	not spec'd	not spec'd	0.1	0.08
Oil	ppm by wt	5.0	5.0	3.0	not spec'd
Free of H_2S, pyridine, and naphthene					

* Allowable boiling point change on vaporization of 5 – 97 % of the test sample, 0.9 °C

** The noncondensable gases dissolved in ammonia are H_2, N_2, CH_4, and Ar. Their amounts depend on the methods of synthesis and storage. The inerts amount to about 50 mL/kg for atmospheric storage.

Various concentrations and purities of *aqueous ammonia* are on the market. Mostly, the concentration is 25 – 30 % NH_3 and the iron content less than 10 ppm. Shipping in pressure vessels is necessary for ammonia contents above 25 % because of its elevated vapor pressure. For more stringent purity requirements for aqueous ammonia, the containers should be made of seawater-resistant aluminum (magnesium alloyed) or austenitic steels.

Analysis. Ammonia is readily detectable in air in the range of a few parts per million by its characteristic odor and alkaline reaction. Specific indicators, such as Nessler's reagent (HgI in KOH), can detect ammonia in a concentration of 1 ppm. For the quantitative determination of ammonia in air, synthesis gas, and aqueous solutions, individual (manual) and continuous (recorded) analyses can be made (for a measurement station for automatic determination of ammonium/ammonia, see [1198]). The methods used include, among others:

Acidimetry and volumetric analysis by absorption
Gas chromatography
Infrared absorption
Thermal conductivity measurement
Electrical conductivity measurement
Measurement of heat of neutralization
Density measurement (for aqueous ammonia)

Normally, the water content of liquid ammonia is determined volumetrically as the ammonia-containing residue on evaporation or gravimetrically by fully vaporizing the ammonia sample and absorbing the water on KOH.

The oil content of liquid ammonia can be tested gravimetrically by first evaporating the ammonia liquid and then concentrating the ether extract of the residue. Iron, aluminum, calcium (ammonia catalyst) and other impurities can then be determined in the ether-insoluble residue.

The inert gas content is analyzed volumetrically after the vaporized ammonia has been absorbed in water. Then the inert gas composition is analyzed chromatographically.

[Top of Page]

7. Environmental, Safety, and Health Aspects

7.1. Environmental Aspects of Ammonia Production and Handling

Measured by its overall environmental impact — air, water and soil pollution, materials and energy consumption — ammonia production is a rather clean technology, with low emissions, low energy consumption due to high-efficiency process design, and no severe cross-media dilemmas, in which improving one environmental effect worsens another. Typical emissions are given in EFMA's publication Production of Ammonia in the series “Best Available Techniques and Control in the European Fertilizer Industry” [1174].

Steam Reforming Ammonia Plants. Table 32 summarizes the emission values for steam reforming ammonia plants.

Table 32. Emissions from steam reforming ammonia plants

			Existing plants	New plants
Emission to air	NO_x (as NO_2)*	ppmv (mL/m^3)	150	75
		mg/Nm^3	300	150
		kg/t NH_3	0.9	0.45
Emission to water	NH_3/NH_4 (as N)	kg/t NH_3	0.1	0.1
Waste material**		kg/t NH_3	< 0.2	< 0.2

* At 3 % O_2.

** Spent catalysts.

The source of the NO_x emissions is the flue gas of the fired primary reformer, and in plants without a fired tubular reformer NO_x is emitted from fired heaters and auxiliary boilers, but in considerably smaller quantities. However, NO_x emissions from ammonia production, compared to the total amount from human activities, is in fact a marginal quantity. Only about 0.16 % of the anthropogenic NO_x emissions come from ammonia production. Emissions to water, generally originating from the condensate from the condensation of surplus process steam ahead of the CO_2 removal system, can largely be avoided. Minor concentrations of methanol and amines can be removed and recycled to the reformer feed by stripping with process steam, and the stripped condensate can be recycled to the boiler feed water after polishing with ion exchangers. Steam stripping without recycling to the reformer would produce a cross media effect, because in this case the pollutants would be transferred from water to air. Another possibility is use of the process condensate for feed gas saturation [458].

Partial oxidation ammonia plants have the same emission sources except for the primary reformer flue gas. The plants have an auxiliary boiler to generate steam for power production and fired heaters, which on account of the sulfur content of the fuel oil release a flue gas containing SO_2 (< 1500 mg/m^3). Other possible emissions are H_2S (< 0.3 mL/m^3), CO (30 mL/m^3) and traces of dust. The NO_x content of the flue gas depends on the configuration of the auxiliary boiler and on the extent electric power generation on the site as opposed to outside supply. The total NO_x emission per tonne of product may be somewhat lower than for steam reforming plants.

Noise. An increasing awareness has developed for noise generation and emission from ammonia plants especially in the neighborhood of residential areas [1216]. Many investigations have dealt with noise generation and noise abatement in ammonia plants [1217-1221]. The following major sources can be identified: depressurizing of large gas quantities for control or venting, steam blowing, burner noise, resonance vibrations in the flue gas ducts, and noise from compressors, blowers, and pumps. Measures for noise reduction include installing low-noise let-down valves, use of silencers, sound-reducing enclosures for compressors or housing them in closed buildings.

Emission Limits and Guideline Values for Ammonia Production. There are two categories of regulations:

1. Legally binding emission values for certain pollutants associated with ammonia production
2. Guideline values not legally binding but providing the background for requirements laid down in individual operating permits

In Europe such legally binding emission levels relevant to ammonia production exist only in Germany. In the Netherlands and Germany limits for emissions from boilers also apply for ammonia plants, for example for the reformer furnace. The present limit in Germany is 200 mg NO_x/m^3 for furnaces up to 300 MW_{th}. Specific emission guideline values are laid down in the United Kingdom. In the other European countries no national emission limits or guidelines exist; individual operation permits are negotiated, usually orientated on other cases and countries. In the United States, for example, NO_x emission level values are categorized in the Clean Air Act, which defines a significance level for the source according to the total emission (10 to 100 t/a) on the one hand, and assigns a threshold limit to the geographical area (100 t/a to 10 t/a) on the other.

For ammonia in air the authorities in Germany require that the achievable minimum level should correspond to the state of the art; originally projected values in a draft of TA Luft [1215], were not subsequently adopted; some MIK values (maximum imission concentrations) are given in VDI Richtlinie 2310 (Sept. 1974). In Germany there is no maximum ammonia concentration laid down for wastewater, it is only referred to the general interdiction to introduce poisons or harmful substances into rivers and lakes. A concentration of 1.25 mg NH_3 per liter, stated to be harmful to fish, could serve as a guideline. In the “Catalog of Water Endangering Substances” (Germany) anhydrous ammonia is not listed, but aqueous ammonia is classified in the Water Endangering Group 1 as a weakly endangering substance.

To harmonize European regulations concerning emission value permits for industrial production, the EU commission has proposed a Council Directive Concerning Integrated Pollution Prevention Control (IPPC Directive). A key element is the concept of Best Available Techniques (BAT). This directive requests an exchange between national environmental authorities, industry and nongovernmental organizations with the aim of preparing IPPC BAT Reference Documents (BREFs), which could provide a common accepted basis for emission limit value regulations of the individual EU member states. A BREF for ammonia is being prepared to serve as pilot project. A previous BAT document on ammonia was published in 1990 [1175]. For noise emissions usually local permits are negotiated, whereby in most instances the contribution to the sound level in the surrounding residential areas will be limited. For the limitation of sound levels at working places within the plant occupational safety regulations are valid which also require ear protection above a certain level. The European Fertilizer Manufacture's Associatiom (EFMA) has published a series: “Best Available Techniques for Pollution Prevention and Control in the European Fertilizer Industry” Booklet No. 1 is “Production of Ammonia” [1174]. For most up-to-date immission regulations contact the relevant authorities in the different countries.

7.2. Safety Features

In ammonia production three potential hazard events can be identified: fire/explosion hazard from the hydrocarbon feed system; fire/explosion hazard due to leaks in the synthesis gas purification, compression, or synthesis section (75 % hydrogen); and toxic hazard from release of liquid ammonia from the synthesis loop. In addition there is also a potential toxic hazard in handling and storing of liquid ammonia. The long history of ammonia production since 1913 has demonstrated this production technology is a very safe operation. Despite the complexity, size and partially severe operating conditions of vessels and piping remarkably few failures occur. It is largely thanks to such forums as the annual AIChE Safety in Ammonia Plants and Related Facilities Symposium that failures, incidents, and accidents are reported and discussed including the manner how the problems were handled and dealt with. Thus the industry as a whole has the opportunity to learn from the specific experiences of its individual members. A short desciptions of a number of accidents is found in [1371], [1397]. The severe impacts of rare events with explosions seem to be confined to a radius of around 60 m [1174]. The long industrial experience is summarized in a number of national codes and standards which have to be applied for design, material selection, fabrication, operation, and periodic technical inspections of the equipment used. This is especially important for equipment operating at high temperatures and/or pressures. Apart from proper design, skilled and well-trained operators and effective and timely maintenance are essential for safe plant operation.

General practice today is to make so-called HAZOP (hazards and operability) studies with an experienced team consisting of operating personal, process engineers (also from contractors), experts in process control and safety experts (often independent consultats from outside the company) in which, following a very meticulous procedure apparatus after apparatus is checked for potential failure and risk possibilities together with a proposal for the appropriate remedies. HAZOP studies introduced by ICI [1359-1362] for general application in chemical plants proved to be very helpful not only for existing plants but also in the planning phase of new ones. It is especially important to define and configure trip systems and trip strategies for safe plant shut downs in case of offset conditions. A similar approach, called Reliability, Availability, Maintainability (RAM) was developed by KBR [1399]. A special point is the assessment of remaining service life of equipment because many ammonia plants operate beyond there originally anticipated lifetime, which was usually defined by economic criteria [1335], [1370]. In this context the question of adequate maintenance is very important and to which extent preventive maintenance is provided. Usually single-train ammonia plants are shut down every two to five years for some weeks for changing of catalyst, equipment inspections, maintenance and, if they fit in the time schedule, also smaller revamps. The time frame of unscheduled shut downs because of problems which need urgent repairs can be used for routine maintenance operations, too. The inspections are widely carried out with nondestructive testing [1402]. Further information is given in [1350]. Statistics of plant on-stream factors, shut-down frequencies and time, as well as an investigation of the causes are found in [1400], [1407].

An area which deserves special attention with respect to safety is the storage of liquid ammonia. In contrast to some other liquefied gases (e.g., LPG, LNG), ammonia is toxic and even a short exposure to concentrations of 2500 ppm may be fatal. The explosion hazard from air/ammonia mixtures is rather low, as the flammability limits [1177-1179], [1200], [1206] of 15 – 27 % are rather narrow. The ignition temperature is 651 °C. Ammonia vapor at the boiling point of – 33 °C has vapor density of ca. 70 % of that of ambient air. However, ammonia and air, under certain conditions, can form mixtures which are denser than air, because the mixture is at lower temperature due to evaporation of ammonia. On accidental release. the resulting cloud can contain a mist of liquid ammonia, and the density of the cloud may be greater than that of air [1176], [1177]. This behavior has to be taken into account in dispersion models.

Regulations and Guidelines. For maintaining safety in ammonia storage, transfer, and handling many regulations and guidelines exist in the various countries. It is beyond the scope of this article to list them completely and to check for the most up-to-date status. It is recommanded to consult the relevant governmental authorities of the individual countries and the relevant industrial codes, applying to fabrication and operation of equipment. An overview of such regulations and standards is given in [1335].

Ammonia Spills. In an ammonia spill a portion of the liquid ammonia released flashes as ammonia vapor. This instantaneous flash is followed by a period of slow evaporation of the remaining liquid ammonia [1176], [1180]. In a release from a refrigerated tank operating slightly above atmospheric pressure, the amount of the initial flash is only a few percent of the total, whereas in a release from a pressure vessel at about 9 bar and 24 °C approximately 20 % of the spilled ammonia would flash. Figures for the proportion flashed as a function of pressure and temperature of the spilled ammonia are given in [1180]. This reference also provides quantitative information on temperature development and evaporation rate of small and large ammonia pools at certain wind velocities and also takes radiation influences (sunny or overcast sky, day or night) into account.

Spill experiments on land and on water in various dimensions have been carried out by various companies and organizations. Underwater releases were also studied. Models have been developed to mathematically describe ammonia dispersion in such events [1182-1187].

A number of real incidents with tank cars, rail cars, tanks, loading/unloading of ships and barges, and pipeline transport in which major spills occurred is described in [1222-1229], [1246], [1278] and in an excellent review on safety in ammonia storage [1188]. Training of operating personal for ammonia handling with special regard to ammonia spills is important [1208], [1214], [1246].

7.3. Health Aspects and Toxicity of Ammonia

In ammonia production, storage, and handling the main potential health hazard is the toxicity of the product itself. For this reason this section concentrates on ammonia only. Other toxic substances such as carbon monoxide or traces of nickel carbonyl (which may be formed during shut down in the methanation stage) may be only a risk in maintenance operations and need appropriate protection provisions as well as blanketing or flushing with nitrogen.

Human Exposure. The threshold of perception of ammonia varies from person to person and may also be influenced by atmospheric conditions, values as low as 0.4 – 2 mg/m^3 (0.5 – 3 ppm) are reported in [1190], but 50 ppm may easily detected by everybody [1304]. Surveys [1191] found concentrations from 9 – 45 ppm in various plant areas. Though initially irritated, exposed persons may quickly become accustomed to these concentrations. Another report [1192] gives concentration limits for short time exposures as follows: 100 ppm (10 min); 75 ppm (30 min) 50 ppm (60 min). The time-weighted average Threshhold Limit Value (TLV) of the ACGIH is 25 ppm (or 18 mg/m^3) [1193]. This recommendation was supplemented by a value for short-time exposure: 35 ppm for 15 min. The MAK value is 50 ppm.

Exposure to higher ammonia concentration has the following effects: 50 – 72 ppm does not disturb respiration significantly [1305]; 100 ppm irritates the nose and throat and causes a burning sensation in the eyes and tachypnoe [1305]; 200 ppm will cause headache and nausea, in addition to the above symptoms [1306]; at 250 – 500 ppm tachypnoe and tachycardia [1306]; at 700 ppm, immediate onset of burning sensations in the eyes [1307]; 1000 ppm causes immediate coughing [1308]. The symptomatology of various exposure levels is also described in [1193-1195]. At 1700 ppm coughing with labored breathing, sometimes with momentary inability to breath (coughing of rescued persons may continue for hours). 2500 to 4500 ppm may be fatal after short exposure; 5000 ppm and higher causes death by respiratory arrest [1188].

The metabolism seems not to be significantly changed after exposure to 800 ppm [1308]. A discussion of metabolism and of acute and chronic health problems caused by ammonia can be found in [1196].

Toxicology. Ammonia is a strong local irritant. On mucous membranes alkaline ammonium hydroxide is formed, which dissolves cellular proteins and causes severe necrosis (corrosive effect).

The primary target organ is the pulmonary system, and the following symptoms can be observed: pharyngitis, laryngitis, tracheobronchitis, nausea, vomiting, increased salivation, reflectoric bradycardia, and life-threatening symptoms, such as edema of the glottis, laryngospasm, bronchospasm, and interstitial lung edema [1283].

Ammonia or ammonium hydroxide can penetrate the cornea rapidly, leading to keratitis, damage of the iris, cataract, and glaucoma [1284].

Oral ingestion of aqueous ammonia can corrode the mucous membranes of the oral cavity, pharynx, and esophagus and cause the shock syndrome, toxic hepatitis, and nephritis. Because of its corrosive action constrictions of the esophagus may result.

Ammonia is absorbed rapidly by the wet membranes of body surfaces as ammonium hydroxide, converted to urea, and excreted by the kidneys [1285]. The capacity of detoxification via urea is sufficient to eliminate the ammonium ion when ammonia is inhaled in nonirritating concentrations. The inhaled ammonia is partly neutralized by carbon dioxide present in the alveoli [1286]. Only a small fraction of the ammonia is exhaled unchanged by the lungs (12.3 % at an inhalation concentration of 230 ppm) [1287]. Repeated inhalation can cause a higher tolerance because the mucous membranes become increasingly resistant [1300]. Additional information on the toxicology of ammonia can be found in [1288-1292], [1295-1299], [1309-1315].

Carcinogenicity. Ammonia failed to produce an increase in the incidence of tumors in Sprague Dawley rats even when the protein ratio in the diet was increased or when urea was added [1301].

Lifetime ingestion of ammonium hydroxide in drinking water by mice was without any carcinogenic effects [1302].

Mutagenicity. Ammonia is not mutagenic in the Ames *Salmonella* system and in *Saccharomyces cerevisiae* [1303].

[Top of Page]

8. Uses

In 2003 about 83 % [1238] of ammonia production was consumed for fertilizers. Ammonia is either converted into solid fertilizers (urea; ammonium nitrate, phosphate, sulfate) or directly applied to arable soil.

The industrial use of ammonia is around 17 % [1238]. Actually every nitrogen atom in industrially produced chemicals compounds comes directly or indirectly from ammonia. An important use of the ammonia nitrogen, partly after conversion to nitric acid, is the production of plastics and fibers, such as polyamides, urea – formaldehyde – phenol resins, melamine-based resins, polyurethanes, and polyacrylonitrile. Another application is the manufacture of explosives, hydrazine, amines, amides, nitriles and other organic nitrogen compounds, which serve as intermediates for dyes and pharmaceuticals. Major inorganic products are nitric acid, sodium nitrate, sodium cyanide, ammonium chloride, and ammonium bicarbonate. Urea production consumed about 40 % of the ammonia produced in 1995.

In the environmental sector ammonia is used in various processes for removing SO_2 from flue gases of fossil-fuel power plants. The resulting ammonium sulfate is sold as fertilizer. In the selective catalytic reduction process (SCR) the NO_x in flue gases is reduced in a catalytic reaction of the nitrogen oxides with a stoichiometric amount of ammonia [1233-1236]. Also noncatalytic reduction is applied with ammonia or urea solutions.

Ammonia may also serve as a solvent in certain processes. Another application is the nitriding of steel. An old and still flourishing business is the use of ammonia as refrigerant, based on its low boiling point and its high heat of evaporation. For some time heavy competition came from chlorofluorocarbons (CFCs), but with increasing environmental concern regarding the application of CFCs ammonia's position is strengthening again. Ammonia is applied in a large number of industrial and commercial refrigeration units and air-conditioning installations [1-4]. In addition to the high specific refrigeration effect, the ammonia has the following advantages: it is noncorrosive; it tolerates moisture, dirt, and oil contaminants; it is cheap and there are many suppliers. A drawback is its toxicity.

The production of smaller volumes of hydrogen/nitrogen mixtures used as protective gases for chemical products [5] and for metal-working processes [6] by

decomposition of ammonia over iron- or nickel-based catalysts at 800 – 900 °C may be an economic alternative where production or purchase of pure hydrogen is too expensive (see also ⟶ Hydrogen – Hydrogen from Ammonia), [1199]. Energy-related applications of ammonia are proposed in [1235].

[Top of Page]

9. Economic Aspects

About 1.4 % of the world consumption of fossil energy (not including combustion of wood) goes into the production of ammonia. In developing countries, ammonia is generally one of the first products of industrialization. As 83 % of world nitrogen consumption is for fertilizers it might be expected that ammonia production should develop approximately in proportion to the growth of world population. This was roughly the case in the mid-1980s as can be seen from Figure 81 [406]; since then the rate of increase in production has been markedly slower than that of world population. This has been mainly for economic reasons in developing countries and for ecological reasons in industrialized countries.

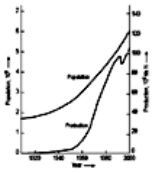

Figure 81. Development of ammonia production and world population

The political changes in the former Eastern Bloc caused a dramatic slump in 1991 – 1993. Table 33 shows the development from 1985 to 2006. Because of the surplus of capacity over demand from the mid-1980s to the mid-1990s the tendency to invest in new plants was rather small and revamping measures aiming at energy saving dominated. A retarding effect resulted also from the financial crises in Asia in the early 1990s associated with a retreat from goverment subsidizing.

Table 33. World ammonia supply/demand balance (10^6 t/a ammonia) [1201-1203, 1207], [1363-1365]

Year	1985	1990	1991	1992	1993	1994	1995	1998	2000	2001	2003	2004	2005	2006
Capacity	143.9	148,7	148.8	151.2	153.9	141.7	142.7	147.6	151.1	151.1	163.2	165.5	168.3	175.0
Demand	111.7	114.6	119.2	113.1	109.7	113,4	115.6	126.7	129.8	126.0	132.3	142.4	142.2	146.5
Capacity utilization, %	77.7	77.1	80.1	74.8	71.3	80.0	81.0	85.9	85.9	8.34	81.07	86.0	84.5	83.7

From 2002/2003 onwards new ammonia projects materialized in increasing capacity numbers (see Fig. 82). The required plant availabilities to satisfiy the expected demand are between 80 and 85 %. However, since the industry is capable to reach plant availabilities of 91 – 93 %, around 40 000 t/d of capacity is idle now. Much of this idle capacity is in the United States (because of the extreme increase of the natural gas price) and in Eastern Europe. The shift is to plants in areas with low gas cost (Middle East, Trinidad, Venezuela).

The future growth of the nitrogen demand supplied by synthetic ammonia is expected to be larger than the growth of the global population. Economic growth with the tendency of higher meat consumption and improvement of nutrition in developing countries are driving forces for increased fertilizer production. Meat consumption is presently increasing with around 3 % per year [1238], the nitrogen required to produce the animal feed is in the range of 20 – 30 % of the total nitrogen demand for fertilizers. The nonfertilizer use of nitrogen is governed by the general economic development and by environmental legislation [1365].

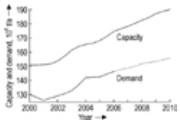

Figure 82. World ammonia capacity and demand

The geographical distribution of world ammonia capacity is shown in Figure 83. Europe and North America, which now together have a 26 % capacity share, lost their leading position (54 % in 1969) to Asia, which now accounts for 46 % (17 % in 1969), as may be seen from Figure 83.

Figure 83. Geographical distribution of world ammonia capacity [1364]

Acccording to the figures for 2002 [1366], the P. R. China is the largest producer with 36.6 × 10^6 t/a NH_3, followed by North America (USA and Canada) with 16.6 × 10^6 t/a, the FSU countries with 15.9 × 10^6 t/a, India 11.9 × 10^6 t/a and Western Europe 11.4 × 10^6 t/a.

Because of the steadily increasing natural gas prices in the United States and Western Europe ammonia imports from low-gas-price locations are constantly growing and a number of ammonia plants have been temporarily or permanently closed.

The major importers of ammonia are Western Europe (2002: 3.7 × 10^6 t) and the United States (2002: 5.6 × 10^6 t), whereas, on account of their enormous natural gas reserves, the FSU countries (4.5 × 10^6 t), Trinidad (3.5 × 10^6 t) and the countries in the Middle East, whose producers export predominantly to Asia, are dominant in ammonia export. Nearly three-quarters of world natural gas reserves (2003) are concentrated in two areas: in the former Soviet Union (32 %) and the Arabian Gulf (41 % including Iran). Iran, which owns 15 % of the world gas reserves, is so far not a major ammonia producer and not an ammonia exporter. An indication of the natural gas prices is given in Table 34.

Table 34. Natural gas prices at different locations around 2004 [1365], [1367], [1368]

Location	$/$10^6$ Btu*
USA	> 6.00
Canada	4.8
Western Europe	4.4
Trinidad	1.6
Venezuela	0.75
Argentina	1.25
Russia	0.9

*1 Btu (British thermal unit) = 1055 J.

For for comparative production cost evaluations these figures should be taken with some care, because they are statistically compiled and might not always be suited for an individual case or a concrete situation. In 2005 gas prices dramatically increased in the USA, where they reached in some areas values above \$ 9/10^6 Btu, and also in Europe the \$ 5 line was crossed.

The feedstock question is discussed in [1204]. Natural gas is by far the most economical feedstock for ammonia production, achieving the lowest energy consumption and requiring the lowest investment. This can be seen from Table 35, which compares ammonia production costs in North-West Europe for different feedstocks using today's best technological standards for each process.

Table 35. Ammonia production costs from various feedstocks (2004)[a]

Feedstock (Process)	Natural gas(steam reforming)	Vacuum residue (partial oxidation)	Coal(partial oxidation)
Feedstock price, \$/10^6 Btu[b] (LHV)	4.6*	2.7*	2.5*
Consumption, 10^6 Btu[b] per tonne NH_3	27	36	45
Feed + energy (fuel), \$ per tonne NH_3	124.2	97.2	112.5
Other cash costs, \$ per tonne NH_3	38.0	51.0	75.0
Total cash costs, \$ per tonne NH_3	162.2	148.2	187.5
Capital related costs[c], \$ per tonne NH_3	65.1	96.3	162,8
Total production costs, \$ per tonne NH_3	227.3	243.5	350.3
Investment[c], 10^6 \$	250	370	625

[a]The 2004 Feedstock prices are taken from the BP Statistical Review of World Energy 2005 [1205]. The price of the vaccuum residue is the heavy fuel oil price minus a deduction of \$ 50/t [1369].
[b]1 Btu (British thermal unit) = 1055 J.
[c]The investment for steam reforming is a budget figure based on recent estimates for a 2000 t/d plant inside battery limits and excludes any offsites, a 45 000 t ammonia tank with refrigeration is included. It is assumed that the utilities are supplied at battery limit in required specification. Investment figures for the vacuum residue and coal-based plants are derived from the steam reforming figure by the well-known investment relation factors. Capital related costs depend highly on the financing condition and on ROI (return on investment) objectives of the investor. So the figures for the capital related costs in the table assuming debt/equity ratio 60:40, depreciation 6%, 8 % interest on debts, 16 % ROI on equity, are only an example to illustrate this substantial contribution to the total costs.

Consumption figures for older plants can be considerably higher than those for modern ones as used in Table 35., so that the world-wide average figure for steam reforming plants is probably between 31 and 33 10^6 Btu per tonne NH_3. For the annual production 330 days of operation per year were assumed. For a specific plant the on-stream factor might be lower for market reasons or because of outages due to technical problems, which will increase the fixed costs per tonne.

With natural gas prices above \$ 9/10^6 Btu in the USA coal-based production would be competitive at full production costs [1411]. But the relative high investment cost for the coal-based plants create more difficulties for the project financing. Nevertheles there are already feasibility studies made in the USA for coal-based production using Illinois coal (\$ 30/t, \$ 1.30/10^6 Btu) [1372]. For coal as an alternative to natural gas see also [1374].

The total investment for a new ammonia production can vary considerably depending on the location and its specific conditions. In the example of Table 35. an industralized site is assumed where all the necessary off-site installations are already existing. For a remote site with heavy forest vegetation where all neccessary offsite facilities like workshops, fire fighting, ambulance, administration, and a few housing buildings and roads have to be included, the total investment for a 2000 t/d capacity could easily reach \$ 450 for a natural gas steam reforming plant.

A special case is the P. R. China where as a consequence of its large coal reserves and limited gas and oil reserves more than 60 % of the ammonia production uses coal, 20 % uses natural gas and the rest is oil-based [1370]. In the rest of the world 90 % of the ammonia production uses natural gas. For the future coal has prospects [1411, 1412] as can be seen from world reserves of fossil energy and their present consumption rate (Table 36).

Table 36. World reserves and consumption rate of fossil feestocks in 2004 [1368]

	Coal, t	Oil, t	Natural gas, Nm^3
Reserves	456×10^9	162×10^9	180×10^{12}
Consumption per year	2778×10^6	3868×10^6	2692×10^9
Years of expected supply	164	42	67

Only about 14 % of the ammonia is traded as such. Only for this proportion prices are regularly published in *Nitrogen + Syngas*, FINDS, by Ferticon, Green Market, IFA, EFMA, TFI statistics and other publications. Apart from some direct uses (3%) all ammonia is converteted into downstream products (Fig. 84).

Figure 84. Applications of ammonia

UAN = Urea – ammonium nitrate; DAP = Diammonium phosphate; MAP = Magnesium – ammonium phosphate

Even with inclusion of the shipping costs from the low-gas-price areas Trinidad (\$ 25/t) and from Russia/Ukraine (\$ 65-70 /t) to the US Gulf Coast these imports can easily compete with the domestic production as to be seen from Table 37.

Table 37. Ammonia cash costs 2005[*]

	USA	Trinidad	Russia	Arabian Gulf
Nat. gas price \$/10^6 Btu	9 – 10	1.6 – 2.0	0.9	0.5 – 1.0
Feedstock costs, \$ per tonne NH_3	310	50	30	16 – 31
Other costs, \$ per tonne NH_3	30	30	30	30
Freight, \$ per tonne NH_3	-	25 [1370] to USA	70 [1370] to USA	25 – 70 to Asia
Cash costs, \$ per tonne NH_3	309 – 340	105 – 155	130	71 – 131

*For simplicity reasons equal gas consumption and other costs are assumed.

Because of the economy of scale there is the tendency to choose a capacity of 2000 – 2200 t/d for new projects at low-gas-price locations and some investors are looking for even larger sizes. The largest single stream plant with 3300 t/d was built by Uhde in Saudi Arabia and started up in 2006. Technical feasability studies for 5000 t/d plants are made by various contractors (see Section Single-train Capacity Limitations – Mega-Ammonia Plants). But besides technical feasability there are questions about the economic risk and financing.

[Top of Page]

10. Future Perspectives

It is difficult to venture a prognosis for the future development of ammonia production technology. As about 85-87 % of the ammonia consumption goes into the manufacture of fertilizers, it is obvious that the future of the ammonia industry is very closely bound up with future fertilizer needs and the pattern of the world supply.

Nitrate pollution, which may become a problem in some countries, is sometimes used as an argument against the use of industrially produced “inorganic” or “synthetic” fertilizers and in favor of more “organic” or “natural” fertilizers. But this is a misconception, because degradation of biomass and manure also leads to nitrate run-off. In addition, the amount of this biological material is far to small to supply the increasing demand of the agriculture necessary to feed the growing world population. The solution to the run-off problem therefore can also not consist in cutting back fertilizer application, which would be disastrous in developing countries, but is likely to lie in a better management of the timing of fertilizer application, requiring a better understanding of the biological nitrogen cycle. Estimates of fixed nitrogen from natural sources and antropogenic activities on global basis are found in [1138-1142].

A fundamental question is, whether there are other options for the future than the present ammonia technology, which involves severe temperatures and pressures. The research for new approaches should strive for the following, admittedly rather ambitious and idealized objectives: ambient reaction conditions, less or no consumption of fossil energy, lower investment, simple application, even in remote developing areas, and wide capacity range. As the major demand is in agriculture, biological processes for in vivo conversion of molecular nitrogen into fixed nitrogen would probably be the first choice. Alternatively abiotic processes (in vitro) with homogeneous catalysis under mild reaction conditions could be discussed for ammonia production.

Biological Processes. Certain bacteria and blue-green algae are able to absorb atmospheric nitrogen, by themselves or in symbiosis with a host plant, and transform it into organic nitrogen compounds via intermediate ammonium ions, whereby the host plant indirectly supplies the large amount of energy needed. A well known example is the symbiotic relationship between legumes (e.g., peas, beans, soybeans, lupins) and Rhizobium bacteria, which settle in their root nodules. Ammonia synthesis is performed by the enzyme nitrogenase [1147-1149]. For the synthesis of the nitrogenase in the cell the so-called NIF gene is responsible. Apart from Rhizobium bacteria a number of other organisms have nitrogen fixation potential, living in symbiosis or providing the required energy from their own metabolism. Genetic engineering to extend biological nitrogen fixation to other plant groups pursues various routes [1150, 1151]:

1. Modifying Rhizobium to broaden its spectrum of host plants
2. Transferring the NIF gene to other bacteria which have a broader spectrum of host plants but have no natural nitrogen-fixing capability
3. Engineering soil bacteria to absorb and convert nitrogen to ammonia and release it to the soil [1143]
4. Inserting the NIF gene directly into plants

Especially the last option, to design nitrogen-fixing species of traditional crops such as wheat, rice, and corn, leads to the following questions: Would it be preferable to fix nitrogen in the leaves where energy is available or in the roots, to which carbohydrates have to be transported? If fixation is performed in the leaves, then how can the problem of the proximity of oxygen generated in the photosynthesis and oxygen-sensitive nitrogenase be dealt with? To what extent will the energy required by a plant to fix nitrogen affect overall productivity?

The last question has been discussed in some detail [1209]. For the conversion of molecular nitrogen into ammonia the following stoichiometric equation may be formulated:

$$0.5\,N_2 + 1.5\,H_2O \rightarrow NH_3 + 0.75\,O_2$$

$$\Delta F^0_{298} = +340\,\mathrm{kJ/mol}\,,\ \ \Delta H^0_{298} = +383\,\mathrm{kJ/mol}$$

The oxidation of glucose generates 3140 kJ/mol and with 100 % efficiency only about 0.11 mol of glucose would be required to produce one mole of ammonia. This ideal can, of course, never be achieved. In the molybdenum-containing nitrogenase a transfer of eight electrons is involved in one catalytic cycle:

$$N_2 + 8\,H^+ + 8\,e^- \rightarrow 2\,NH_3 + H_2$$

As seen from this equation two electrons are “wasted” in forming molecular hydrogen, which also seems to be unavoidable in vivo. This stoichiometry suggests an efficiency of 75 %, but if additional functions associated with fixation are taken into account, a maximum overall efficiency of 10 – 15 % is estimated for the enzyme, and even less for the bacteria as a whole [1210]. In investigations on Rizobium [1211], [1212] it was shown for the nodulated roots of peas that 12 g of glucose are consumed to produce 1 g of fixed nitrogen. Using this efficiency for the case where all the fixed nitrogen in the plant is supplied only via this route, a yield loss of 18 % would result. Nitrate, for example from a fertilizer, absorbed by the plant is transformed to ammonia involving a free energy change $\Delta F^0_{298} = +331$, which is comparable to the free energy change for converting molecular nitrogen to ammonia, as shown in the stoichiometric equation above.

This is a somewhat surprising result, which suggests that nitrogen supply with nitrate has an energetic advantage over the biological fixation process [1213]. A definite answer to this question will only be possible when genetically engineered nitrogen-fixing plants are available in sufficient amounts for field tests. There is no doubt that some day in the future genetic engineering will succeed in designing nitrogen-fixing plants for agricultural crops.

But until this will become a reality, application of selected natural nitrogen-fixing microorganisms to crops would be an intermediate solution to reduce industrial fertilizer consumption. In the United States field tests with *Azospirillium* applied to corn could demonstrate a considerable increase in harvest yield [1151]. Careful directed application of blue-green algae, (*Anabaena azolla*) and water fern (*Azolla pinata*) in tropical rice fields is also very promising [1152, 1153]. Unlike most diazotropic bacteria, blue-green algae are self-sufficient in that they themselves produce the energy necessary for the nitrogen fixation by photosynthesis.

Abiotic Processes [1209]. Possibilities of converting molecular nitrogen into ammonia in homogeneous solution by using organometallic complexes have been studied from around 1966 onwards. Numerous investigations resulted in the laboratory-scale synthesis of ammonia , but with rather low yields. The prospects of this route are not judged to be very promising in terms of energy consumption and with respect to the costs of the rather sophisticated catalyst systems. Photochemical methods [1144], [1145], [1146], to produce ammonia at ambient temperature and atmospheric pressure conditions in presence of a catalysts have up to now attained ammonia yields which are far too low to be economically attractive. For a detailed review of biological and abiotic nitrogen fixation, see ⟶ Nitrogen Fixation and [1209]; additional literature references can be found in [1140], [1141], [1142], [1147], [1148].

Conclusions. Based on the foregoing discussion and the evaluation of present technology (see also [406]) and research efforts one may formulate the scenario for the future development of ammonia production as follows:

1. For the major part, ammonia production in the next 15 to 20 years will rely on the classic ammonia synthesis reaction combining nitrogen and hydrogen using a catalyst at elevated temperature and pressure in a recycle process.
2. It is very likely that genetic engineering will succeed in modifying some classical crops for biological nitrogen fixation and that large-scale application will occur predominantly in areas with strongly growing populations to secure the increasing food demand. This development may be pushed by the fact that compared to the classical fertilizer route less capital and less energy would be needed. This might happen in the future but an estimate regarding extent and time horizon is difficult. But even with the introduction of this new approach, traditional ammonia synthesis will continue to operate in parallel, because it might be necessary to supplement the biological nitrogen fixation with classical fertilizers. In addition, the existing ammonia plants represent a considerable capital investment and a great number of them may reliably operate for at least another 20 to 30 years from a mere technical point of view.
3. Natural gas will remain the preferred feedstock of present ammonia production technology in the medium term (15 – 20 years) as may be assumed from the world energy balance shown in Table 26. Partial oxidation of heavy hydrocarbon residues will be limited to special cases, but coal gasification could have a renaissance within this period.
4. Technical ammonia production based on non-fossil resources, e.g. electrolysis (power from nuclear, solar, water or geothermic energy) or gasification of biomass in this time frame will play no role.

5. Ammonia technology will not change fundamentally, at least not in the next 10 to 15 years. Even if there are radical, unforeseeable developments, they will need some time to mature to commercial applicability. With the traditional concepts, the margins for additional improvements have become rather small after years of continuous research and development. Thus further progress may be confined to minor improvements of individual process steps, catalysts and equipment design.
6. The best energy figure so far achieved in commercial production is around 28 GJ per tonne NH_3. With the ruthenium-based catalysts M.W. Kellogg claims a figure of 27.2 GJ per tonne NH_3. Further reduction would need an even more active catalyst to operate at a much lower synthesis pressure, for example, at around 30 bar, which is the usual pressure level of steam reforming. The probability of finding such a catalyst is rather small after all the years of intensive research. Figures between 27 and 28 GJ per tonne NH_3 already correspond to a rather high efficiency of around 75 % with respect to the theoretical minimum of 20.9 GJ per tonne NH_3 (calculated as stoichiometric methane demand, see Section Minimum Energy Requirement for Steam Reforming Processes).
7. The bulk of ammonia production in the near future will still be produced in world-scale plants of capacity 1000 – 2000 t NH_3 per day. Small capacities will be limited to locations where special logistical, financial or feedstock conditions favor them. More plants in high-gas-cost areas will be closed and the gap will be filled by higher imports from low-price gas areas, where new plants, intended for export, probably also mega-plants, will be built.
8. Most newer technology developments will mainly reduce investment costs and increase operational reliability. Smaller integrated process units, such as exchanger reformers in various configurations, contribute to this reduction and will achieve additional savings by simplifying piping and instrumentation. Progress in instrumentation and computer control could reduce the effective overall energy consumption achieved over the course of a year by tuning plants continuously to optimal operating conditions.

[Top of Page]

References

1. K. Dietrich, *Kältetech. Klim.* **22** (1970) no. 6, 184.
2. H. L. v. Cube, *Kältetechnik* **16** (1964) no. 3, 76.
3. F. Özvegyi, *Kälte (Hamburg)* **23** (1970) no. 6, 298.
4. A. Miller, *Kälte (Hamburg)* **20** (1967) no. 6, 275.
5. H. Borrmann, *Chem. Ing. Tech.* **40** (1968) 1192.
6. C. Hollmann, *Gas Wärme* **9** (1960) 267.
7. L. Pauling: *The Nature of the Chemical Bond*, 3rd ed., Cornell University Press, Ithaca, N.Y., 1960. L. Pauling: *Die Natur der chemischen Bindung*, 3rd ed., Verlag Chemie, Weinheim 1968.
8. F. Hund, *Z. Phys.* **31** (1925) 95.
9. L. Haar, *J. Phys. Chem. Ref. Data* **7** (1978) no. 3, 635 – 792. Links
10. L. J. Christiansen, J. Kjaer: *Enthalpy Tables of Ideal Gases*, Haldor Topsøe, Copenhagen 1982, p. 1.
11. Dow Chemical Corp. *JANAF Thermodynamical Tables* 2nd ed., 1971, NSRDS NBS-37.
12. *VDI Forschungsh.* 1979, no. 596, VDI-Verlag Düsseldorf.
13. F. Din: *Thermodynamic Functions of Gases*, vol. 1, Butterworths, London 1956.
14. A. Nielsen: *An Investigation on Promoted Iron Catalysts for the Syntheses of Ammonia*, Gjellerup, Kopenhagen 1968.
15. M. P. Wukalowitsch *et al.*, *Teploenergetika (Moscow)* **7** (1960) no. 1, 63 – 69; *Chem. Ing. Tech.* **33** (1961) no. 4, 5, 291, 292.
16. R. Döring: *Thermodynamische Eigenschaften von Ammoniak (R 717)*, Verlag C. F. Müller, Karlsruhe 1978; *Ki Klima + Kälte Ing*. extra no. 5.
17. *Landolt-Börnstein*,vol. 2, part 1 – 10, Springer Verlag, Berlin-Heidelberg-New York 1971.
18. *Handbuch der Kältetechnik*, vol. **IV**, Springer Verlag, Berlin – Heidelberg – New York 1956.
19. A. Kumagai, T. Toriumi, *J. Chem. Eng. Data* **16** (1971) no. 3, 293. Links
20. R. Plank, *Kältetechnik* **12** (1960) 282 – 283.
21. W. K. Chung, *Can. J. Chem. Eng.* **55** (1977) no. 6, 707 – 711. Links
22. G. G. Fuller, *Ind. Eng. Chem. Fundam.* **15** (1976) no. 4, 254 – 257.
23. B. D. Djordjevic *et al.*, *Chem. Eng. Sci.* **32** (1977) no. 9, 1103 – 1107. Links
24. R. Wiebe, V. I. Gaddy, *J. Am. Chem. Soc.* **60** (1938) 2300. Links
25. E. P. Bartlett, H. L. Cupples, T. H. Tremearne, *J. Am. Chem. Soc.* **50** (1928) 1275. Links
26. E. P. Bartlett *et al.*, *J. Am. Chem. Soc.* **52** (1930) 1363. Links
27. B. H. Sage, R. H. Olds, W. N. Lacey, *Ind. Eng. Chem.* **40** (1948) 1453.
28. A. Michels *et al.*, *Appl. Sci. Res.* **A3** (1953) 1.
29. A. Michels *et al.*, *Appl. Sci. Res.* **A4** (1954) 180.
30. M. Zander, W. Thomas, *J. Chem. Eng. Data* **24** (1979) 1. Links
31. L. Haar, *J. Res. Natl. Bur. Stand. Sect. A* **72** (1968) 2.
32. E. R. Grahl, *Pet. Process.* **8** (1953) 562.
33. I. Hay, G. D. Honti, in G. D. Honti (ed.): *The Nitrogen Industry*, part 1, Akademiai Kiado, Budapest 1976, pp. 68 – 73.
34. R. G. Barile, G. Thodos, *Can. J. Chem. Eng.* **43** (1965) 137 – 142. Links
35. K. V. Reddy, A. Husain, *Ind. Eng. Chem. Process Des. Dev.* **19** (1980) 580 – 586.
36. H. Masuoka *et al.*, *J. Chem. Eng. Jpn.* **10** (1977) no. 3, 171 – 175. Links
37. A. T. Larson, C. A. Black, *J. Am. Chem. Soc.* **47** (1925) 1015. Links
38. A. Michels *et al.*, *Physica (Utrecht)* **16** (1950) 831 – 838.
39. A. Michels *et al.*, *Physica (Utrecht)* **25** (1959) 840.
40. G. Guerreri, *Hydrocarbon Process.* **49** (1970) no. 12, 74 – 76.
41. H. H. Reamer, B. H. Saga, *J. Chem. Eng. Data* **4** (1959) no. 2, 152. Links
42. R. Wiebe, T. H. Tremearne, *J. Am. Chem. Soc.* **55** (1933) 975. Links
43. F. Heise, *Ber. Bunsenges. Phys. Chem.* **76** (1972) 938.
44. H. Zeininger, *Chem. Ing. Tech.* **45** (1973) no. 17, 1067.
45. B. Lefrancois, C. Vaniscotte, *Genie Chim.* **83** (1960) no. 5, 139.
46. H. H. Reamer, B. H. Saga, *J. Chem. Eng. Data* **4** (1959) no. 4, 303. Links
47. B. I. Lee, M. G. Kesler, *AIChE J.* **21** (1975) no. 3, 510. Links
48. C. G. Alesandrini *et al.*, *Ind. Eng. Chem. Process Des. Dev.* **11** (1972) no. 2, 253.
49. H. Masuoka, *J. Chem. Eng. Jpn.* **10** (1977) no. 3, 171 – 175. Links
50. K. Konoki *et al.*, *J. Chem. Eng. Jpn.* **5** (1972) no. 2, 103 – 107. Links
51. G. Sarashina, *J. Chem. Eng. Jpn.* **7** (1974) no. 6, 421 – 425. Links
52. R. Heide, *Luft Kältetech.* **7** (1971) no. 3, 128 – 130.
53. I. Durant, *J. Chem. Soc.* 1934, no. 135, 730. Links
54. Yu. V. Efremov, I. F. Golubev, *Russ. J. Phys. Chem. Eng. (Engl. Transl.)* **36** (1962) no. 5, 521 – 522.
55. J. A. Jossi *et al.*, *AIChE J.* **8** (1962) no. 1, 59 – 63.
56. H. Shimotake, G. Thodos, *AIChE J.* **9** (1963) no. 1, 63.
57. B. K. Sun, T. S. Storwick, *J. Chem. Eng. Data* **24** (1979) no. 2, 88. Links
58. I. T. Perelsatein, E. B. Parushin, *Kholod. Tekh.*, 1980, no. 6, 34 – 37.
59. A. K. Pal, A. K. Barura, *J. Chem. Phys.* **47** (1967) no. 1, 216. Links

60. E. A. Mason, L. Mouchik, *J. Chem. Phys.* **36** (1962) no. 10, 2746. Links
61. A. B. Rakshit, C. S. Roy, A. K. Barua, *J. of Chemical Physics* **59** (1973) 3633 – 3688.
62. D. P. Needham, H. Ziebland, *Int. J. Heat Mass Transfer* **8** (1965) 1387 – 1414.
63. A. M. P. Tans, *Ind. Chem.* **39** (1963) no. 3, 141.
64. R. Afshar, S. Murad, S. C. Saxena, *Chem. Eng. Commun.* **10** (1981) 1 – 11. Links
65. D. A. Kouremenos, *Kältetechnik* **20** (1968) no. 10, 319.
66. A. O. S. Maczek, P. Gray, *Trans. Faraday Soc.* **66** (1970) 127.
67. P. Correia *et al.*, *Ber. Bunsenges. Phys. Chem.* **72** (1968) 393.
68. W. Inkofer *et al.*, *Ammonia Plant Saf.* **13** (1971) 67.
69. K. A. Hofmann: *Anorganische Chemie*, 20th ed., Vieweg & Sohn, Braunschweig 1969.
70. U. Müller, A. Greiner, *Chem. Tech. (Leipzig)* **18** (1966) no. 6, 327.
71. *Chem. Zentralbl.* 1966, 14 – 0884.
72. J. Jander: *Chemie in wasserfreiem flüssigem Ammoniak*, Vieweg & Sohn, Braunschweig – Interscience, New York 1963.
73. J. P. Lenieur *et al.*, *C. R. Acad. Sci. Ser. C* **268** (1969) no. 20, 1791.
74. E. Weiss, W. Buchnery cited in [69].
75. H. Smith: *Organic Reactions in Liquid Ammonia*, Vieweg & Sohn, Braunschweig – Interscience, New York–London 1963.
76. F. W. Bergstrom, *Ind. Eng. Chem.* **24** (1932) 57.
77. F. Ritter, F. Walter in K. Winnacker, E. Weingaertner (eds.): *Chemische Technologie*, **vol. 2**, Carl Hanser, München 1950 pp. 368.
78. J. Hess, *Chem. Ind. (Berlin)* **45** (1922) 538.
79. *Chem. Ind. (Berlin)* **48** (1925) 718.
80. G. Plumpe: *Die I. G. Farbenindustrie*, — Wirtschaft, Technik und Politik Dunker & Humblot, Berlin 1990.
81. P. Barton *et al.*, *Ind. Eng. Chem. Process Des. Dev.* **7** (1968) no. 3, 366.
82. A. Mittasch: Chemische Grundlegung der industriellen Ammoniakkatalyse, Ludwigshafen 1953.
83. *Ullmann*, 3rd ed., **vol. 3**, p. 544 – 546.
84. A. Mittasch, W. Frankenburger, *Z. Elektrochem. Angew. Phys. Chem.* **35** (1929) 920.
85. F. Haber, G. van Oordt, *Z. Anorg. Allg. Chem.* **43** (1904) 111; **44** (1905) 341; **47** (1905) 42.
86. F. Jost, *Z. Anorg. Allg. Chem.* **57** (1908) 414. Links
87. F. Haber, R. Le Rossignol, *Ber. Dtsch. Chem. Ges.* **40** (1907) 2144.
88. P. Krassa, *Chem. Ztg. Chem. Appar.* **90** (1966) 104.
89. F. Haber, *Z. Elektrochem. Angew. Phys. Chem.* **14** (1908) 181; **20** (1914) 597.
90. R. Le Rossignol, *Naturwissenschaften* **16** (1928) 1070.
91. B. Timm, *Chem. Ing. Tech.* **35** (1963) no. 12, 817.
92. F. A. Henglein, *Chem. Ztg.* **75** (1951) 345 – 350, 389 – 392, 407 – 410.
93. C. Bosch, Nobel prize address 1933.
94. K. Holdermann: *Im Banne der Chemie – Carl Bosch*, Econ Verlag, Düsseldorf 1953.
95. M. Appl in W. F. Furter (ed.): *A Century of Chemical Engineering*, Plenum Publ. Corp., 1982 p. 29.
96. M. Appl: *IIChE Conference*, Hyderabad, 1986; *Indian Chemical Engineer* **XXIX** (1987) no. 1, 7 – 29.
97. S. A. Tophamy: *Catalysis, Science and Technology*, vol.7, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo pp. 1 – 50.
98. A. Nagel et al.: *Stickstoff*, Schriftenreihe der BASF, 2nd ed., 1991.
99. M. Appl, *Nitrogen* **100** (1976) 47.
100. M. Appl: "The Haber-Bosch Heritage: The Ammonia Production Technology", *50th Anniversary of the IFA Technical Conference*, Sept. 1997, Sevilla.
101. L. J. Gillespie, J. A. Beattie, *Phys. Rev.* **36** (1930) 743.
102. F. Haber, S. Tamaru, *Z. Elektrochem. Angew. Phys. Chem.* **21** (1915) 191.
103. L. J. Gillespie, J. A. Beattie, *Phys. Rev.* **36** (1930) 1008.
104. V. G. Telegin, G. M. Lutskii, *J. Appl. Chem. USSR (Engl. Transl)* **37** (1964) 2271.
105. L. J. Christiansen in W. A. Nielsen (ed.): *Ammonia — Catalysis and Manufacture*, Springer, Heidelberg 1995 pp. 1 – 16.
106. I. Granet, *Pet. Refiner* **33** (1954) no. 5, 205.
107. I. Granet, P. Kass, *Pet. Refiner* **32** (1953) no. 3, 149.
108. A. K. Pal, A. K. Barua, *J. Chem. Phys.* **47** (1967) 116.
109. W. G. Frankenburger, *Catalysis 1954 – 1960* **3** (1955) 171.
110. R. Brill *et al.*, *Ber. Bunsenges. Phys. Chem.* **73** (1969) 999.
111. A. Ozaki, K. Aika: "The Synthesis of Ammonia by Heterogeneous Catalysis," in R. W. F. Hardy *et al.* (eds.): *A Treatise on Dinitrogen Fixation, Sect. I and II*, Wiley, New York 1979, pp. 169 – 247.
112. A. Ozaki, K. Aika: "Catalytic Activation of Dinitrogen," in R. H. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology*, **vol. 1**, Springer Verlag, Berlin–Heidelberg–New York 1981, pp. 87 – 158.
113. P. H. Emmett in E. Drauglis, R. I. Jaffee (eds.): *The Physical Basis for Heterogenous Catalysis*, Plenum Publ., New York 1975, p. 3.
114. K. Tamaru in J. Chatts *et al.* (eds.): *New Trends in the Chemistry of Nitrogen Fixation*, Academic Press, New York 1980, p. 13.
115. G. Ertl in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York and London 1991, pp. 109 – 131.
116. P. Stolze in A. Nielson (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, Heidelberg 1995, pp. 17 – 102.
117. G. A. Somorjai: *Chemistry in Two Dimensions: Surfaces*, Cornell Press, Ithaca 1981.
118. G. Ertl, J. Küppers: *Low Electrons and Surface Chemistry*, 2nd ed., VCH Verlagsgesellschaft, Weinheim 1985.
119. G. Ertl: "Kinetics of Chemical Processes on Well-defined Surfaces," in R. H. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology*, **vol. 4**, Springer-Verlag, Berlin–Heidelberg–New York 1983, pp. 257 – 282.
120. G. Ertl: "Zum Mechanismus der Ammoniaksynthese," *Nachr. Chem. Tech. Lab.* **31** (1983) no. 3, 178.
121. M. Boudart: "Kinetics and Mechanism of Ammonia Synthesis," *Catal. Rev. Sci. Eng.* **23** (1981) 1 – 15.
122. J. Schütze *et al.*: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 195 – 214.
123. G. A. Somorjai *et al.*: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 215 – 231.
124. R. Schlögl in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York and London 1991 pp. 19 – 108.
125. K. C. Wough *et al.*: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 295 – 301.
126. A. Nielsen, *Catal. Rev.* **4** (1970) no. 1, 1.
127. G. Gramatica, N. Pernicone in: J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York and London 1991, pp. 211 – 252.
128. J. P. Hansen in A. Nielson (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, Heidelberg 1995, pp. 149 – 190.
129. L. M. Aparicio, J. A. Dumesic: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 233 – 252.
130. A. Nielsen: "Ammonia Synthesis: Exploratory and Applied Research," *Catal. Rev. Sci. Eng.* **23** (1981) 17 – 51.
131. D. C. Silverman, M. Boudart, *J. Catal.* **77** (1982) 208. Links
132. C. Peters *et al.*, *Z. Elektrochem.* **64** (1960) 1194.
133. K. Schäfer, *Z. Elektrochem.* **64** (1960) 1190.
134. P. H. Emmett, S. Brunauer, *J. Am. Chem. Soc.* **59** (1937) 310; Links *J. Catal.* **15** (1969) 90. Links

135. S. Brunauer, P. H. Emmett, *J. Am. Chem. Soc.* **62** (1940) 1732. Links
136. D. C. Silverman, Dissertation, Stanford University 1976.
137. H. C. Chen, R. B. Anderson, *J. Colloid Interface Sci.* **38** (1972) 535; Links *J. Catal.* **28** (1973) 161. Links
138. G. Ertl, N. Thiele, *Appl. Surf. Sci.* **3** (1979) 99.
139. R. Hosemann, A. Preisinger, W. Vogel, *Ber. Bunsenges. Phys. Chem.* **70** (1966) 786. H. Ludwiszek, A. Preisinger, A. Fischer, R. Hosemann, A. Schönfel, W. Vogel, *J. Catal.* **51** (1977) 326. R. Buhl, A. Preisinger, *Surf. Sci.* **47** (1975) 344. Links
140. W. S. Borghard, M. Boudart, *J. Catal.* **80** (1983) 194. Links
141. G. Ertl, M. Huber, S. B. Lee, Z. Paal, M. Weiss, *Appl. Surf. Sci.* **8** (1981) 373.
142. J. J. F. Scholten *et al.*, *Trans. Faraday Soc.* **55** (1959) 2166.
143. B. Fastrup: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 273 – 283.
144. J. J. F. Scholten, Dissertation, Techn. Hochschule Delft 1959.
145. P. Mars et al. in J. H. de Boer (ed.): *The Mechanism of Heterogenous Catalysis*, Elsevier, Amsterdam 1960, p. 66.
146. K. Tanaka, A. Matsumaya, *J. Res. Inst. Catal. Hokkaido Univ.* **2** (1971) 87.
147. M. I. Temkin, *Adv. Catal.* **28** (1979) 173.
148. P. H. Emmett, S. Brunauer, *J. Am. Chem. Soc.* **56** (1934) 35. Links
149. H. S. Taylor, *J. Am. Chem. Soc.* **53** (1931) 578. Links
150. W. Frankenburger, *Z. Elektrochem.* **39** (1933) 45, 97, 269.
151. D. R. Strongin, G. A. Somorjai in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York 1991, pp. 133 – 178.
152. T. Nakata, S. Matsushita, *J. Phys. Chem.* **72** (1968) 458.
153. F. Boszo *et al.*, *J. Cat.* **49** (1977) 103.
154. G. Ertl *et al.*, *J. Vac. Sci. Technol.* **13** (1976) 314.
155. G. Ertl *et al.*, *Surf. Sci.* **114** (1982) 515. Links
156. G. Ertl *et al.*, *Surf. Sci.* **114** (1982) 527. Links
157. K. T. Kozhenova, M. Y. Kagan, *Zh. Fiz. Khim.* **14** (1940) 1250.
158. S. R. Bare, D. R. Strongin, G. A. Somorjai, *J. Phys. Chem.* **90** (1987) 4726.
159. D. R. Strongin *et al.*, *J. Catal.* **103** (1987) 289. Links
160. L. Falicov, G. A. Somorjai, *Proc. Natl. Akad. Sci. USA* **82** (1985) 2207.
161. M. Bowker: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 265 – 271.
162. C. Rettner, H. Stein, *Phys. Rev. Let.* **56** (1987) 2768.
163. C. Rettner, H. Stein, *J. Chem. Phys.* **87** (1987) 770. Links
164. J. K. Norskov: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 385 – 403.
165. M. Boudart: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 405 – 414.
166. M. Gunze *et al.*, *Phys. Rev. Lett.* **53** (1984) 850.
167. L. J. Whitman, C. E. Bartosch, W. Ho, G. Strasser, M. Grunze, *Phys: Rev. Lett.* **56** (1986) 1984.
168. L. J. Whitman, C. E. Bartosch, W. Ho, *J. Chem. Phys.* **85** (1986) 3688. Links
169. M. Grunze, G. Strasser, M. Golze, *Appl. Phys.* **A 44** (1987) 19.
170. G. Ertl: "Plenary Lecture," in T. Seiyama, K. Tanabe (eds.): New Horizons in Catalysis, Proc. Int. Congr. Catal. 7th 1980, Elsevier 1981, p. 21.
171. F. Bozso, G. Ertl, *J. Catal.* **49** (1977) 18; Links **50** (1977) 519.
172. R. Schlögl: "Ammonia Synthsis," in G. Ertl, H. Knözinger, J. Weitcamp (eds.): *Handbook of Heterogeneous Catalysis*, **vol. 4**, Wiley-VCH, Weinheim 1997, pp. 1698 – 1748.
173. N. D. Spencer, R. C. Schoonmaker, G. A. Somorjai, *J. Catal.* **74** (1982) 129. Links
174. R. Westrik, P. Zwieterring, *Proc. K. Ned. Akad. Wet. Ser. B: Palaeontol. Geol. Phys. Chem.* **56** (1953) 492.
175. R. Brill *et al.*, *Angew. Chem.* **79** (1967) 905. Links
176. H. Taube, Dissertation, Freie Universität Berlin 1968.
177. R. Brill, *Ber. Bunsenges. Phys. Chem.* **75** (1971) 455.
178. J. A. Dumesic, H. Topsøe, S. Khammouma, M. Boudart, *J. Catal.* **37** (1975) 503; Links see also [368].
179. W. A. Schmidt, *Angew. Chem.* **80** (1968) 151. Links
180. V. I. Shvachko *et al.*, *Kinet. Katal.* **7** (1966) 635, 734.
181. R. Brill *et al.*, *Z. Phys. Chem. (Wiesbaden)* **64** (1969) 215.
182. Y. Morikawa, A. Ozaki, *J. Catal.* **12** (1968) 145. Links
183. N. Takezawa, I. Toyoshima, *J. Catal.* **19** (1970) 271. Links
184. S. Carra, R. Ugo, *J. Catal.* **15** (1969) 435. Links
185. I. I. Tretyakov *et al.*, *Dokl. Akad. Nauk SSSR* **175** (1967) 1332.
186. G. Parravano, *J. Catal.* **8** (1967) 29. Links
187. M. I. Temkin *et al.*, *Kinet. Katal.* **4** (1963) 494.
188. S. R. Logan, J. Philp, *J. Catal.* **11** (1968) 1. Links
189. N. Takezawa, P. H. Emmett, *J. Catal.* **11** (1968) 131. Links
190. G. Schulz-Ekloff, *Ber. Bunsenges. Phys. Chem.* **75** (1971) 110.
191. A. Nielsen *et al.*, *J. Catal.* **3** (1964) 68. Links
192. P. E. H. Nielsen, unpublished work, Haldor Topsøe A/S, cited in [130].
193. D. A. King, F. Sebba, *J. Catal.* **4** (1965) 253. Links
194. N. Segal, F. Sebba, *J. Catal.* **8** (1967) 105. Links
195. S. R. Tennison in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York and London 1991, pp. 303 – 364.
196. K. Aika, K. Tamaru in W. A. Nielsen (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, Heidelberg 1995, pp. 103 – 148.
197. J. N. Nwalor, J. G. Goodwin, jr.: *Topics in Catalysis*, **vol. 1**, Baltzer AG, Basel 1994, p. 228 – 148.
198. K. Aika, A. Ozaki, *J. Catal.* **14** (1969) 311. Links
199. G. E. Moore, F. C. Unterwald, *J. Chem. Phys.* **48** (1968) 5409, 5393. Links
200. M. R. Hillis *et al.*, *Trans. Faraday Soc.* **62** (1966) 3570.
201. K. Azuma, *Nature (London)* **190** (1961) no. 4775, 530.
202. K. Aika, A. Ozaki, *J. Catal.* **16** (1970) 97. Links
203. H. Kubota, M. Shindo, *J. Chem. Eng. Jpn.* **23** (1959) 242.
204. M. I. Temkin, V. Pyzhev, *Acta Physicochim. URSS* **12** (1940) 327.
205. M. I. Temkin, *Zh. Fiz. Khim.* **24** (1950) 1312.
206. M. I. Temkin, N. M. Morozow, E. M. Shapatina, *Kinet. Catal. (Engl. Trans.)* **4** (1963) 565.
207. R. M. Adams, E. W. Comings, *Chem. Eng. Prog.* **49** (1953) 359. Links
208. P. H. Emmett, J. T. Kummer, *Ind. Eng. Chem.* **35** (1943) 677.
209. D. Annable, *Chem. Eng. Sci (Genie Chimique)* **1** (1952) no. 4, 145.
210. C. Bokhoven *et al.*, *Catalysis 1954 – 1960* **3** (1955) 265.

211. R. Brill, *J. Chem. Phys.* **19** (1951) 1047. Links
212. S. Kiperman, V. S. Granovskaya, *Zh. Fiz. Khim.* **26** (1952) 1615.
213. M. I. Temkin *et al.*, *Kinet. Katal.* **4** (1963) 224.
214. A. Ozaki *et al.*, *Proc. R. Soc. London Ser. A* **258** (1960) 47.
215. V. D. Livshits, I. P. Sidorov, *Zh. Fiz. Khim.* **26** (1952) 538.
216. K. Aika, A. Ozaki, *J. Catal.* **13** (1969) 232. Links
217. R. Brill, *J. Catal.* **16** (1970) 16. Links
218. ICI: *Catalyst Handbook with Special Reference to Unit Processes in Ammonia and Hydrogen Manufacture*, Wolfe Scientific Books, London 1970.
219. A. Nielsen, J. Kjaer, B. Hansen, *J. Catal.* **3** (1964) 68. Links
220. A. Ozaki, H. S. Taylor, M. Boudart, *Proc. R. Soc. London Ser. A* **258** (1960) 47.
221. A. Cappelli, A. Collina, *Inst. Chem. Eng. Symp. Ser.* **35** (1972) 10.
222. G. Buzzi-Ferraris *et al.*, *Eng. Chem. Sci.* **29** (1974) 1621.
223. U. Guacci *et al.*, *Ind. Eng. Chem.* **16** (1977) 166.
224. D. G. Dyson, J. M. Simson, *Ind. Eng. Chem. Fundam.* **7** (1968) 601.
225. E. Wicke, *Z. Elektrochem.* **60** (1956) 774.
226. M. Bowker, I. Parker, K. Waugh, *Appl. Catal.* **14** (1985) 101.
227. M. Bowker, I. Parker, K. Waugh, *Surface Science* **97** (1988) L 233.
228. P. Stolze, J. K. Norskov, *Phys Rev. Let.* **55** (1985) 2502.
229. P. Stolze, J. K. Norskov, *Surface Science* **189/190** (1987) 91; *Surface Science* **197** (1988) L 230; Links *J. Catal.* **110** (1988) 1. Links
230. P. Stolze, *Phys. Sci.* **36** (1987) 824.
231. J. A. Dumesic, A. A. Trevino, *J. Catal.* **116** (1989) 119. Links
232. A. Wheeler, *Adv. Catal.* **3** (1951) 249.
233. T. Akehata *et al.*, *Can. J. Chem. Eng.* **39** (1961) 127. Links
234. C. Bokhoven, W. van Raayen, *J. Phys. Chem.* **58** (1954) 471.
235. J. Hoogschagen, *Ind. Eng. Chem.* **47** (1955) 906.
236. C. Wagner, *Z. Phys. Chem. (Leipzig)* **193** (1943) 1.
237. E. W. Thiele, *Ind. Eng. Chem.* **31** (1939) 916.
238. D. G. Dyson, J. M. Simon, *Ind. Eng. Chem. Fundam.* **7** (1968) 605.
239. C. Chu, O. A. Hougen, *Chem. Eng. Sci.* **17** (1962) 167. Links
240. C. P. P. Singh, N. D. Saraf, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) 364.
241. A. Cappelli, A. Collina, *Inst. Chem. Eng. Symp. Ser.* **35** (1972) 10.
242. J. Kjaer: *Calculation of Ammonia Converters on an Electronic Digital Computer*, Akademisk Forlag, Kopenhagen 1963.
243. J. J. Hag, I. M. Palla, *Br. Chem. Eng.* **8** (1963) no. 3, 171.
244. J. Kubec, J. Burianova, Z. Burianec, *Int. Chem. Eng.* **14** (1974) no. 4, 629.
245. R. F. Baddour, R. L. T. Brian, B. A. Logeais, J. P. Eymery, *Chem. Eng. Sci.* **20** (1965) 281. Links R. L. T. Brian, R. F. Baddour, J. P. Eymery, *Chem. Eng. Sci.* **20** (1965) 297. Links
246. M. J. Shah, *Ind. Eng. Chem.* **59** (1967) no. 1, 73.
247. A. D. Stephens, *Chem. Eng. Sci.* **30** (1975) 11. Links
248. L. D. Gaines, *Ind. Eng. Chem. Process. Des. Dev.* **16** (1977) no. 3, 381.
249. T. Huberich, R. Krabetz in A. V. Slack (ed.): *Ammonia*, part III, Marcel Dekker, New York – Basel 1977.
250. A. Mittasch: *Geschichte der Ammoniaksynthese*, Verlag Chemie, Weinheim 1951.
251. A. Mittasch, *Adv. Catal.* **2** (1950) 82.
252. C. Bosch, *Z. Elektrochem. Angew. Phys. Chem.* **24** (1918) 361.
253. K. Aika, J. Yamaguchi, A. Ozaki, *Chem. Lett.* 1973, 161; cited in [112]. Links
254. S. R. Logan, C. Kemball, *Trans. Faraday Soc.* **56** (1960) 144.
255. E. P. Perman, *Proc. R. Soc. London* **76** (1905) 167.
256. F. Haber, G. von Oørdt, *Z. Anorg. Allg. Chem.* **43** (1905) 111.
257. W. Nernst, *Z. Elektrochem. Angew. Phys. Chem.* **13** (1907) 521.
258. BASF, DE 249447, 1910.
259. A. T. Larson, R. S. Tour, *Chem. Metall. Eng.* **26** (1922) 493, 555, 588, 647.
260. J. A. Almquist, E. D. Crittenden, *Ind. Eng. Chem.* **18** (1926) 1307.
261. G. L. Bridger *et al.*, *Chem. Eng. Prog.* **43** (1947) 291. Links
262. H. Hinrichs, *Br. Chem. Eng.* **12** (1967) 1745.
263. Indianapolis Center for Advanced Research, US 4235749, 1980.
264. Topsøe, GB 989242, 1961.
265. P. D. Rabina *et al.*, *Khim. Prom.* **5** (1969) 350.
266. Chevron Res. Comp., US 3653831, 1972.
267. J. L. Carter, C. G. Savini, US 3472794, 1969; SU 173204 (749 751) 23-4, 1961.
268. The Lummus Comp., US 3992328, 1976.
269. Ammonia Casale, IT-A 47920 A, 1979.
270. ICI, EP-A 7830276, 1979.
271. The Lummus Comp., US 3951862, 1976. Société Chimique de La Grande Paroisse, Azote et Produits Chimiques, US 3839229, 1974.
272. P. H. Emmett, S. Braunauer, *J. Am. Chem. Soc.* **59** (1937) 1553. Links
273. C. Peters, K. Schäfer, R. Krabetz, *Z. Elektrochem.* **64** (1960) 1194.
274. M. E. Dry, J. A. K. du Plessis, B. M. Leuteritz, *J. Catal.* **6** (1966) 194. Links
275. R. Krabetz, C. Peters, *Angew. Chem. Int. Ed. Engl.* **4** (1965) 341.
276. A. Kazusaka, I. Toyoshima, *Z. Phys. Chem. (Wiesbaden)* **128** (1982) 111.
277. L. M. Dmitrenko *et al.*, *Proc. Int. Congr. Catal. 4th 1968* **1** (1971) 404.
278. K. Karaslavova, M. D. Anastasov, *Geterog. Katal. Tr. Mezhdunar. Simp. 3rd 1975*, 1978, 297.
279. K. Z. Zakieva, P. D. Rabina, J. E. Zubova, I. E. Klaustova, N. Z. Pavlova, L. D. Kusnetzov, *Tr. Mosk. Khim. Tekhnol. Inst. im. D. I. Mendeleeva* 1973 no. 73, 120.
280. B. Aleksic, N. Jovanovic, A. Terlecki-Baricevic, *Geterog. Katal. Tr. Mezhdunar. Simp. 3rd 1975*, 1978, 147.
281. M. G. Berengarten *et al.*, *Kinet. Katal.* **15** (1974) 250.
282. Yu. N. Artyukh, M. T. Rusov, N. A. Boldyreva, *Kinet. Katal.* **8** (1967) 1319.
283. P. J. Smith, D. W. Taylor, D. A. Dowden, C. Kemball, D. Tayler, *Appl. Catal.* **3** (1982) 303.
284. I. D. Kusnetsov, *Chem. Tech. (Leipzig)* **15** (1963) 211.
285. J. Herrmann *et al.*, *Chem. Tech. (Leipzig)* **18** (1966) 472.
286. D. G. Ivanov, N. D. Anastasov, *Chem. Tech. (Leipzig)* **15** (1963) 229.

287. L. Sokol, *Chem. Tech. (Leipzig)* **15** (1963) 214.
288. U. Zardi, A. Antonini, *Nitrogen* **122** (1979) 33.
289. M. C. Sze, *Hydrocarbon Process.* **56** (1977) no. 12, 127.
290. N. Pernicone, G. Liberti, G. Servi in C. Eyraud, M. Escoubes (eds.): *Progress in Vacuum Mikrobalance Techniques*, **vol. 3**, Heyden, London 1975, p. 304.
291. A. Mittasch, E. Keunecke, *Z. Elektrochem. Angew. Phys. Chem.* **38** (1932) 666.
292. H. Uchida, I. Terao, K. Ogawa, *Bull. Chem. Soc. Jpn.* **37** (1964) 653. Links
293. N. Pernicone, F. Traina: *Preparation of Catalysts*, **vol. 2**, Elsevier, Amsterdam 1979, p. 321.
294. B. S. Clausen, S. Mørup, H. Topsøe, R. Candis, E. J. Jensen, A. Baranski, A. Pattek, *J. Phys. (Orsay, Fr.)* **37** (1976) C6-245.
295. P. D. Rabina, T. Y. Malysheva, L. D. Kusnetzov, V. A. Batyrev, *Kinet. Catal. (Engl. Transl.)* **11** (1970) 1030.
296. A. Mittasch, E. Keunecke, *Z. Elektrochem. Angew. Phys. Chem.* **38** (1932) 666.
297. R. Westrik, *J. Chem. Phys.* **21** (1953) 2049.
298. M. E. Dry, L. C. Ferreira, *J. Catal.* **7** (1967) 352. Links
299. F. Garbassi, G. Fagherazzi, M. C. Calcaterra, *J. Catal.* **26** (1972) 338. Links
300. A. C. Turnock, H. P. Eugster, *J. Petrol.* **3** (1962) 533.
301. R. Brill, *Allg. Prakt. Chem.* **17** (1966) 94.
302. R. Hosemann *et al.*, *Ber. Bunsenges. Phys. Chem.* **70** (1966) 769.
303. L. v. Bogdandy *et al.*, *Ber. Bunsenges. Phys. Chem.* **67** (1963) 958.
304. F. J. Brykowski (ed.): *Ammonia and Synthesis Gas*, Noyes Data Corp., Park Ridge, N.J., 1981, p. 280.
305. N. Pernicone *et al.*, *Catal. Proc. Int. Congr. 5th 1972* 1973, 1241.
306. V. Solbakken, A. Solbakken, P. H. Emmett, *J. Catal.* **15** (1969) 90. Links
307. The Gas Council London, DE-OS 1931758, 1969 (T. Nicklin, D. Oyden).
308. Y. Sasa, M. Uda, I. Toyoshima, *Chem. Lett.* 1982, no. 12, 2011. Links
309. S. K. Egeubaev *et al.*, *Kinet. Katal.* **6** (1965) no. 4, 676; also see [126] and [277]
310. M. M. Ivanov *et al.*, *Kinet. Katal.* **9** (1968) 1239; *Kinet. Katal.* **10** (1969) 349.
311. A. Ozaki *et al.*, *Catal. Proc. Int. Congr. 5th 1972* **2** (1973) 1251.
312. E. K. Enikeev, A. V. Krylova, *Kinet. Katal.* **3** (1962) 116; *Chem. Abstr.* **57** (1962) 7965.
313. G. Ertl, M. Weiss, S. B. Lee, *Chem. Phys. Lett.* **60** (1979) 391. Links
314. M. C. Tsai *et al.*, *Surf. Sci.* **155** (1985) 387. Links
315. J. A. Dumesic, Dissertation, Stanford University 1974.
316. T. Yoshioka, J. Koezuka, I. Toyoshima, *J. Catal.* **14** (1969) 281. Links
317. Y. E. Sinyak, Thesis, Inst. of Chemical Physics Akad. Sci. USSR, Moscow 1960; cited in [318].
318. Haldor Topsøe, ZA 645279, 1964; also see [326].
319. I. P. Nukhlenov *et al.*, *J. Appl. Chem. USSR (Engl. Transl.)* **37** (1964) 248.
320. V. N. Anokhin *et al.*, *J. Appl. Chem. USSR (Engl. Transl.)* **35** (1962) 29.
321. A. Baranski *et al.*, *J. Catal.* **26** (1972) 286. Links
322. A. Baranski *et al.*, *Archivum Hutniktva* **25** (1980) 153.
323. A. Baranski *et al.*, *Appl. Catal.* **3** (1982) 207.
324. A. Baranski *et al.*, *Appl. Catal.* **19** (1985) 417.
325. A. Pattek-Janczyk *et al.*, *Appl. Catal.* **6** (1983) 35.
326. V. V k, *Chem. Ing. Tech.* **45** (1973) 608.
327. K. Feind, *Chem. Ing. Techn.* **38** (1966) 1081.
328. Chemie Linz, US 3965246, 1976.
329. Wargons Aktiebolag, Haldor Topsøe, US 3243386, 1961.
330. Kuhlmann, GB 1080838, 1964.
331. Kuhlmann, CH 434218, 1964.
332. Haldor Topsøe, GB 989242, 1961.
333. ICI, BE 576059, 1959.
334. IG Farbenind., DE 748620, 1938.
335. I. V. Nicolescu *et al.*, *Genie Chim.* **82** (1959) 33.
336. I. V. Nicolescu *et al.*, *Chem. Tech. (Leipzig)* **15** (1963) 226.
337. S.I.R.I. Società Italiana Ricerche Industriali, US 4073749, 1978.
338. ICI, GB 1484864, 1977.
339. J. E. Jarvan, H. U. Larsen, unpublished work, (Haldor Topsøe A/S) cited in [130].
340. A. T. Larson *et al.*, *Chem. Trade J. Chem. Eng.* **100** (1937) 403.
341. G. M. Schwab (ed.): *Handbuch der Katalyse*, **vol. 5: Heterogene Katalyse II**, Springer-Verlag, Wien 1957, p. 566.
342. A. M. Rubinštejn *et al.*, *Izv. Akad. Nauk SSSR Ser. Khim.* 1966, 1707; *Chem. Zentralbl.* 1968, 8 – 0639.
343. A. M. Rubinštejn *et al.*, *Kinet. Katal.* **6** (1965) 285; *Chem. Zentralbl.* 1965, 42 – 0622.
344. P. Bussiere, R. Dutartre, G. A. Martin, J. P. Mathieu, *C. R. Hebd. Seances Acad. Sci. Ser. C* **280** (1975) 1133. R. Dutartre, M. Primet, G. A. Martin, *React. Kinet. Catal. Lett.* **3** (1975) 249.
345. H. Topsøe, J. A. Dumesic, E. C. Derouane, B. S. Clausen, S. Mørup, J. Villadsen, N. Topsøe in: *Preparation of Catalysts*, **vol. 2**, Elsevier, Amsterdam 1979, p. 365. H. Topsøe et al. in T. Seiyama, K. Tanabe (eds.), New Horizons in Catalysis, Proc. Int. Congr. Catal. 7th 1980, Elsevier 1981.
346. M. Boudart, A. Delbouille, J. A. Dumesic, S. Khammouna, H. Topsøe, *J. Catal.* **37** (1975) 486. Links J. A. Dumesic, H. Topsøe, S. Khammouna, M. Boudart, *J. Catal.* **37** (1975) 503. Links J. A. Dumesic, H. Topsøe, M. Boudart, *J. Catal.* **37** (1975) 513. Links
347. The Lummus Comp., GB 1529823, 1978.
348. IG Farbenind., DE 554855, 1928; FR 611139, 1935.
349. N. W. Sseljakow, SU 48210, 1936; *Chem. Zentralbl.* **1937** II, 119.
350. N. N. Pschenitzyn *et al.*, *Zh. Prikl. Khim. (Leningrad)* **13** (1940) 76; *Chem. Zentralbl.* 1940 II, 1932.
351. Österreichische Stickstoffwerke, AT 252957, 1964.
352. M. V. Tovlin *et al.*, *Kinet. Katal.* **7** (1966) 749; *Chem. Zentralbl.* 1966, 42 – 0597.
353. H. Flood, D. G. Hill, *Z. Elektrochem.* **61** (1957) 18.
354. J. Valcha, *Chem. Tech. (Leipzig)* **15** (1963) 222.
355. Österreichische Stickstoffwerke, GB 833878, 1958.
356. D. B. Christozvonov *et al.*, *Int. Chem. Eng.* **9** (1969) 387.
357. SU 169077, 1963.
358. H. Uchida, N. Todo, *Bull. Chem. Soc. Jpn.* **29** (1956) 20; also see [356]. Links
359. J. A. Burnett *et al.*, *Ind. Eng. Chem.* **45** (1953) 1678.
360. Z. N. Bardik *et al.*, *Khim. Promst. (Moscow)* **42** (1966) 351.
361. CS 117658, 1966 (L. Sokol).

362. P. A. Strelzov *et al.*, *Khim. Promst. (Moscow)* **43** (1967) no. 1, 46.
363. T. L. Joseph, *Trans. AJME* **120** (1936) 72.
364. P. K. Stangway, H. U. Ross, *Trans. AJME* **242** (1968) 1981.
365. M. Moukassi *et al.*, *Metall. Trans.* **B14** (1983) 125.
366. Y. K. Rao *et al.*, *Metall. Trans.* **B10** (1979) 243.
367. J. O. Edström, *J. ISI* **175** (1953) 289.
368. H. Amariglio, G. Rambeau, in G. C. Bond et al. (eds.): Proc. Int. Congr. Catal. 6th 1976, Chemical Society London 1977, p. 1113.
369. A. Baranski, M. Lagan, A. Pattek, A. Reizer, L. J. Christiansen, H. Topsøe in: *Preparation of Catalysts*, **vol. 2**, Elsevier, Amsterdam 1979, p. 353.
370. R. Royen, G. H. Langhans, *Z. Anorg. Allg. Chem.* **315** (1962) 1.
371. P. H. Emmett, S. Brunauer, *J. Am. Chem. Soc.* **52** (1930) 2682; Links *J. Am. Chem. Soc.* **62** (1940) 1732. Links
372. I. A. Smirnov *et al.*, *Kinet. Katal.* **6** (1965) 351.
373. I. A. Smirnov, *Kinet. Katal.* **7** (1966) 107.
374. P. E. Højlund Nielsen, paper at the ACS meeting, Washington 1983.
375. P. E. Højlund Nielsen in W. A. Nielsen (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, Heidelberg 1995, pp. 191 – 199.
376. P. E. Højlund Nielsen in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York 1991, pp. 285 – 301.
377. P. Stolze, *Phys. Sci.* **36** (1983) 824.
378. P. Stolze, J. K. Nørskov, *J. Vac. Sci. Technol.* **A5** (1987) 581.
379. T. Kirkerød, P. Skaugset, Abstracts of IV Nordic Symp. on Catalysts, Trondheim 1991.
380. I. P. Sidorov, K. E. Istomina, *Chem. Abstr.* **54** (1960) 6049.
381. H. Hinrichs, *Chem. Ing. Tech.* **40** (1968) 723.
382. L. N. Marakhovets *et al.*, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **15** (1972) no. 5, 735.
383. R. Brill *et al.*, *Ber. Bunsenges. Phys. Chem.* **72** (1968) 1218.
384. J. D. Bulatnikova *et al.*, *Zh. Fiz. Khim.* **32** (1958) 2717.
385. H. J. Hansen in V. Sauchelli (ed.): *Fertilizer Nitrogen*, its Chemistry and Technology, Reinhold Publ. Co., New York 1964.
386. A. Ozaki, *Acc. Chem. Res.* **14** (1981) 16. Links
387. F. Haber, *Z. Elektrochem.* **16** (1910) 244.
388. F. F. Gadallah et al. in G. Poncelet et al. (eds.): *Preparation of Catalysts*, **vol. 3**, Elsevier, Amsterdam 1983.
389. GB 1 468 441, 1977.
390. GB 174 079, 1985.
391. L. Volpe, M. Boudart, *J. Phys. Chem.* **90** (1986) 4874.
392. O. Glemser, *Naturwissenschaften* **37** (1950) 539. Links
393. D. J. Lowe, B. E. Smith, R. N. F. Thornely in P. M. Harrison (ed.): *Metalloproteins*, Part 1, Verlag Chemie, Weinheim 1985, p. 207.
394. A. Ozaki, K. Urabe, K. Shimazaki, S. Sumiya: *Preparation of Catalysts*, vol. 2: Scientific Basis for the Preparation of Heterogenous Catalysts, Elsevier, Amsterdam 1979, p. 381.
395. *Nitrogen* **193** (1991) 17 – 21.
396. W. K. Taylor, S. A. Hall, D. G. Heath, ICI/CFDC Technical Symp., Shanghai 1989.
397. J. P. Shirez, J. R. LeBlanc, Kellogg Ammonia Club Meeting, San Francisco 1989.
398. British Petroleum Co., US 4163775, 1979.
399. H.-D. Müller, *Z. Phys. Chem. (Leipzig)* **255** (1974) no. 3, 486.
400. S. Peter, H. Wenzel, *Chem. Ing. Tech.* **45** (1973) no. 9 + 10, 573.
401. E. Comandini, A. Passariello, U. Zardi, *Ammonia Plant Saf.* **23** (1981) 44.
402. A. V. Slack, G. Russel James (eds.): *Ammonia*, "part I – IV", Marcel Dekker, New York 1973 /74.
403. S. Strelzoff: *Technology and Manufacture of Ammonia*, Wiley-Interscience, New York 1981.
404. I. Dybkjaer in W. A. Nielsen (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, Heidelberg 1995, pp. 199 – 327.
405. C. W. Hooper in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis*, Plenum Press, New York 1991, pp. 253 – 283.
406. M. Appl: *Ammonia, Methanol Hydrogen, Carbon Monoxide — Modern Production Technologies*, CRU, London 1997.
407. *Nitrogen* **168** (1987) 29 – 37.
408. *Nitrogen* **186** (1990) 21 – 30.
409. D. Wagner, J. F. Meckel, K. H. Laue in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 471.
410. *Nitrogen* 1975, no. 97, 35.
411. T. Grundt, K. Christiansen in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 73.
412. H. Wendt, *Chem. Ing. Tech.* **56** (1984) no. 4, 265.
413. G. A. Crawford, S. Benzimra, *Fertilizer '85 Conference (British Sulphur)*, London, Feb. 1985.
414. K. J. Mundo, *Chem. Ing. Tech.* **45** (1973) no. 10 a, 632.
415. T. A. Czuppon, L. J. Buividas, *Hydrocarbon Process.* **58** (1979) no. 9, 197.
416. J. Morrison, *Oil Gas J.* **67** (1969) Feb. 24, 76.
417. W. Auer, *Erdöl Kohle Erdgas Petrochem.* **24** (1971) 145.
418. D. F. Balz, H. F. Gettert, K. H. Gründler, *Plant Oper. Prog.* **2** (1983) no. 1, 47.
419. *Nitrogen* **226** (1997) 43 – 46.
420. M. Appl, H. Gössling, *Chem. Ztg.* **96** (1972) 135.
421. H. Fricke, *Chem. Ztg.* **96** (1972) 123.
422. R. Lohmüller, *Chem. Ing. Tech.* **56** (1984) no. 3, 203.
423. C. I. Kuhre, C. I. Shearer, *Hydrocarbon Process.* **50** (1971) no. 12, 113.
424. L. J. Buividas, J. A. Finneran, O. J. Quartulli, *Chem. Eng. Prog.* **70** (1974) no. 10, 21; Links *Ammonia Plant Saf.* **17** (1975) 4.
425. J. Thuillier, F. Pons, *Ammonia Plant Saf.* **20** (1978) 89.
426. R. G. Cockerham, G. Percival, *Trans. Inst. Gas. Eng.* **107** (1957/58) 390/433.
427. H. Jockel, B. E. Triebskorn, *Hydrocarbon Process.* **52** (1973) no. 1, 93.
428. W. Rall, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) no. 5, 351.
429. H. Jockel, *GWF Gas Wasserfach* **110** (1969) no. 21, 561.
430. T. Ishiguro, *Hydrocarbon Process.* **47** (1968) no. 2, 87.
431. N N. Clark., W. G. S. Henson, 33rd AIChE Ammonia Safety Symp., Denver 1988.
432. B. J. Cromarty, *Nitrogen* **191** (1991) 30 – 34.
433. R. Vannby, S. Winter-Madsen, 36th AIChE Ammonia Safety Symp., Los Angeles 1991.
434. B. J. Cromarty, B. J. Crewdson, *The Application of Pre-reforming in the Ammonia and Hydrogen Industries*, ICI Catalco Tech. Paper, 59W/0590/0/CAT2.
435. K. Elkins, *Feedstock Flexibility Options*, ICI Catalco Tech. Paper, 313/035/0REF.
436. P. W. Farnell, *Pre-Reforming — a Retrofit Case Study*, ICI Catalco Tech. Paper, 291W/025/0/IMTOF.
437. W. D. Verduijn, 37th AIChE Ammonia Safety Symp., San Antonio 1992.
438. S. E. Nielsen, I. Dybkaer, 41th AIChE Ammonia Safety Symp., Boston 1996.

439. T. Tomita, M. Kitugawa, *Chem. Ing. Tech.* **49** (1977) no. 6, 469.
440. *Nitrogen* **144** (1983) 33.
441. T. Tomita et al., Special Paper 9, 11th World Petroleum Congress, London 1983.
442. O. J. Quartulli, *Hydrocarbon Process.* **44** (1965) no. 4, 151.
443. J. Davies, D. A. Lihou, *Chem. Process Eng. (London)* **52** (1971) no. 4, 71.
444. H. D. Marsch, H. J. Herbort, *Hydrocarbon Process.* **61** (1982) no. 6, 101.
445. *Nitrogen* **214** (1995) 38 – 56.
446. I. Dybkær, *Fuel Processing Technology* **42** (1995) 85 – 107. Links
447. *Nitrogen* **195** (1992) 22 – 36.
448. A. I. Forster, B. J. Cromarty, *The Theory and Practice of Steam Reforming*, ICI/Catalco/KTI/UOP, 3rd Int. Seminar on Hydrogen Plant Operation, Chicago 1995.
449. *Nitrogen* **166** (1987) 24 – 31.
450. *Nitrogen* **167** (1987) 31 – 36.
451. G. W. Bridger, W. Wyrwas, *Chem. Process Eng. (London)* **48** (1967) no. 9, 101.
452. F. Marschner, H. J. Renner, *Hydrocarbon Process.* **61** (1982) no. 4, 176.
453. J. R. Rostrup-Nielsen, *Ammonia Plant Saf.* **15** (1973) 82.
454. G. W. Bridger, *Oil Gas J.* **74** (Feb. 16, 1976) 73; Links *Ammonia Plant Saf.* **18** (1976) 24.
455. D. P. Rounthwaite, *Plant Oper. Prog.* **2** (1983) no. 2, 127.
456. *Nitrogen* **174** (1988) 23 – 24.
457. B. I. Grotz, *Hydrocarbon Process.* **46** (1967) no. 4, 197.
458. J. G. Livingstone, A. Pinto, *Chem. Eng. Prog.* **79** (1983) no. 5, 62.
459. *Chem. Eng. (N.Y.)* **73** (1966) 24.
460. *Hydrocarbon Process.* **63** (1984) no. 4, 103.
461. *Nitrogen* **178** (1989) 30 – 39.
462. K. J. Elkins, I. C. Jeffery, D. Kitchen, A. Pinto, “Nitrogen '91”, *International Conference (British Sulphur)*, Copenhagen, June 4 – 6, 1991.
463. P. M. Armitage, J. Elkins, D. Kitchen, AIChE Ammonia Safety Symp., Los Angeles 1991.
464. D. Kitchen, K. Mansfield, Eurogas 92, Trondheim 1992.
465. G. E. Weismantel, L. Ricci, *Chem. Eng. (N.Y.)* **86** (1979) no. 21, 57.
466. R. Schneider, Kellogg Ammonia Club Meeting, San Diego 1990.
467. J. R. LeBlanc, Asia Nitrogen: British Sulphur Intern. Conf., Singapore 1996.
468. *Ammonia*, Kellogg brochure, HG 1/96.
469. J. R. LeBlanc, R. Schneider, K. W. Wright, 40th AIChE Ammonia Safety Symp., Tucson 1985.
470. T. A. Czuppon, S. A. Knez, 41th AIChE Ammonia Safety Symp., Boston 1996.
471. G. J. van den Berg *et al.*, *Hydrocarbon Process.* **45** (1966) no. 5, 193.
472. H. D. Marsch, N. Thiagarajan, 33rd AIChE Ammonia Safety Symp., Denver 1988.
473. H. J. Herbort, Uhde Ammonia Symp., Dortmund 1992.
474. N. Thiagarajan, Uhde Ammonia Symp., Dortmund 1992.
475. H. D. Marsch, N. Thiagarajan, 37th AIChE Ammonia Safety Symp., San Antonio 1992.
476. *Nitrogen* **214** (1996) 46.
477. L. W. ter Haar, *Ind. Chim. Belge* **33** (1968) 655.
478. P. D. Becker *et al.*, *Chem. Process Eng. (London)* **52** (1971) no. 11, 59.
479. C. L. Reed, C. J. Kuhre, *Hydrocarbon Process.* **58** (1979) no. 9, 191.
480. *Lurgi Gas Production Technology: The Shell Process* Lurgi comp. brochure 189e/6.92/2.20.
481. C. Higmann, *Perspectives and Experience with Partial Oxidation of Heavy Residues*, Lurgi company brochure.
482. *Integrated Gasification Combined Cycle Process*, Lurgi information paper, 1995.
483. W. Liebner, N. Hauser, ERPI Conference on Power Generation and the Environment, London 1990.
484. C. Higmann, G. Grünfelder, Conference on Gasification Power Plants, San Francisco 1984.
485. W. Soyez, 33rd AIChE Ammonia Safety Symp., Denver 1988.
486. *Nitrogen* **83** (1973) 40.
487. C. P. Marion, J. R. Muenger, *Energy Prog.* **1** (1981) no. 1 – 4, 27.
488. *Uhde's Design of the Texaco Oil Gasification*, Krupp – Uhde comp. brochure TOPG 2A/AB, 1995.
489. H. J. Madsack, *Hydrocarbon Process.* **61** (1982) no. 7, 169.
490. S. Strelzoff, *Hydrocarbon Process.* **53** (1974) no. 12, 79.
491. *Hydrocarbon Process.* **58** (1979) no. 4, 168.
492. *Hydrocarbon Process.* **63** (1984) no. 4, 90.
493. *Nitrogen* **126** (1980) 32.
494. *Nitrogen* **161** (1986) 23 – 27.
495. E. Supp: *How to Produce Methanol from Coal*, Springer-Verlag, Heidelberg 1989.
496. *Ammonia from Coal*, Tennessee Valley Authority, Muscle Shoals, Ala. 1979.
497. *Nitrogen* **226** (1997) 47 – 56.
498. L. J. Buividas, *Chem. Eng. Prog.* **77** (1981) no. 5, 44; *Ammonia Plant Saf.* **23** (1981) 67.
499. *Ammonia from Coal*, Bull. Y-143, Tennessee Valley Authority, Muscle Shoals, Ala. 1979.
500. D. A. Waitzman, *Chem. Eng. (N.Y.)* **85** (1978) 69.
501. F. Brown, *Hydrocarbon Process.* **56** (1977) no. 11, 361.
502. The Ultimate in Ammonia/Urea Studies, Fertecon 1991.
503. H. Teggers, H. Jüntgen, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) no. 4, 163.
504. H. Staege, *Tech. Mitt. Krupp Werksber.* **40** (1982) no. 1, 1.
505. *Nitrogen* **126** (1980) 32 – 39.
506. *Hydrocarbon Process.* **63** (1984) no. 4, 94.
507. H. J. Michaels, H. F. Leonard, *Chem. Eng. Prog.* **74** (1978) no. 8, 85.
508. H. Staege, *Hydrocarbon Process.* **61** (1982) no. 3, 92.
509. J. E. Franzen, E. K. Goeke, *Hydrocarbon Process.* **55** (1976) no. 11, 134.
510. A. D. Engelbrecht, Tennessee Valley Authority Symposium in Muscle Shoals, Alabama 1979.
511. *Nitrogen* 1985, no. 156, 35.
512. *Hydrocarbon Process.* **63** (1984) no. 4, 96.
513. M. J. van der Burgt, *Hydrocarbon Process.* **58** (1979) no. 1, 161.
514. C. A. Bayens, 36th AIChE Ammonia Safety Symposium, Los Angeles 1991.
515. *Clean Coal Technology*, International Power Generation, 1990.

516. M. J. van der Burgt, J. E. Naber, *Ind. Chem. Bull.* (1983) 104 – 105.
517. J. N. Mahagaokar, J. N. Phillips, A. Kreweinghaus, 10th ERPI Conf. on Coal Gasification, 1986.
518. T. W. Nurse, *5th International Symposium Large Chemical Plants*, Antwerpen 1982.
519. E. T. Child, Tennessee Valley Authority Symposium in Muscle Shoals, Alabama 1979.
520. B. Cornils *et al.*, *Hydrocarbon Process.* **60** (1981) no. 1, 149.
521. B. Cornils *et al.*, *Chem. Ing. Tech.* **52** (1980) no. 1, 12.
522. *Hydrocarbon Process.* **63** (1984) no. 4, 97.
523. W. Konkol *et al.*, *Hydrocarbon Process.* **61** (1982) no. 3, 97.
524. D. E. Nichols, P. C.Williamson in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 53.
525. B. Cornils *et al.*, *CEER Chem. Econ. Eng. Rev.* **12** (1980) no. 6 – 7, 7.
526. T. J. Pollaert, *Chem. Eng. Prog.* **74** (1978) no. 8, 95.
527. *Hydrocarbon Process.* **60** (1980) no. 11, 130.
528. M. K. Schad, C. F. Hafke, *Chem. Eng. Prog.* **79** (1983) no. 5, 45.
529. H. Hiller, *Erdöl Kohle Erdgas Petrochem.* **28** (1975) no. 2, 74.
530. P. D. Becker, Tennessee Valley Authority Symposium in Muscle Shoals, Alabama 1979.
531. *Hydrocarbon Process.* **61** (1982) no. 4, 133.
532. W. Schäfer *et al.*, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) no. 12, 557.
533. *Hydrocarbon Process.* **61** (1982) no. 4, 151.
534. K. A. Theis, E. Nitschke, *Hydrocarbon Process.* **61** (1982) no. 9, 233. *Hydrocarbon Process.* **63** (1984) no. 4, 93.
535. P. Speich, H. Teggers, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) no. 8, 376.
536. J. M. Moe, *Chem. Eng. Prog.* **58** (1962) no. 3, 33. Links
537. W. L. E. Davey, E. L. Taylor, M. D. Newton, P. S. Larsen, P. S. Weitzel, "Asia Nitrogen," *Br. Sulphur Conf.*, Kuala Lumpur, Feb. 22 – 24, 1998, Proceedings of the Conf., vol. 2, p. 165.
538. *Nitrogen* **40** (1966) 28.
539. A. P. Ting, Shen-Wu-Wan, *Chem. Eng. (N.Y.)* **76** (1969) no. 11, 185.
540. J. F. Lombard, *Hydrocarbon Process.* **48** (Aug. 1969) 111.
541. L. Lloyd, M. V. Twigg, *Nitrogen* 1979, no. 118, 30.
542. *Nitrogen* 1981, no. 133, 27.
543. P. N. Hawker, *Hydrocarbon Process.* **61** (1982) no. 4, 183.
544. P. S. Pedersen, J. H. Carstensen, J. Boghild-Hansen, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
545. R. E. Stockwell, *FINDS V* (1990).
546. S. J. Catchpole *et al.*, *Modern Catalyst Systems for Increased Ammonia Production and Efficiency*, ICI Catalco Tech. Paper.
547. H. Roos, H. Wanjek, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
548. J. R. Rostrup-Nielsen *et al.*, 37th AIChE Ammonia Safety Symp., San Antonio 1992.
549. J. B. Hansen, P. S. Pedersen, J. H. Carstensen, 33rd AIChE Ammonia Safety Symp., Denver 1988.
550. J. M. Stell *et al.*, *The Operation of High, Low and Intermediate Temperature Catalysts*, ICI Catalco/KTI/UOP Hydrogen Plant Seminar, Chicago 1995.
551. J. Ilg, B. Kandziora, 41st AIChE Ammonia Safety Symp., Boston 1996.
552. I. Dybkjaer in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 503.
553. J. D. Rankin, J. G. Livingstone, *Ammonia Plant Saf.* **23** (1981) 203.
554. *Hydrocarbon Process.* **61** (1982) no. 4, 154.
555. I. Dybkjaer, H. Bohlbro, *Ammonia Plant Saf.* **21** (1979) 145.
556. I. Dybkjaer, in: *Ammonia from Coal Symposium*, TVA Muscle Schoals Alabama, USA p. 133.
557. P. W. Young, B. C. Clark, *Ammonia Plant Saf.* **15** (1973) 18.
558. W. C. Lundberg, *Ammonia Plant Saf.* **21** (1979) 105.
559. *Nitrogen* **72** (1971) 34.
560. J. Y. Livingston, *Hydrocarbon Process.* **50** (Jan. 1971) 126.
561. H. W. Neukermans, J. P. Schurmans, *AIChE Symp. Saf. Ammonia Plants*, Montreal 1981.
562. H. W. Schmidt, *Chem. Ing. Tech.* **40** (1968) 425.
563. C. A. Vancini: *Synthesis of Ammonia*, Macmillian, London 1971, p. 249.
564. H. W. Schmidt, H. J. Henrici, *Chem. Ztg.* **96** (1972) no. 3, 154.
565. *Hydrocarbon Process.* **63** (1984) no. 4, 57 – 64.
566. A. G. Eickmeyer in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 2**, Marcel Dekker, New York 1974.
567. H. E. Benson, R. W. Parrish, *Hydrocarbon Process.* **53** (1974) no. 4, 81.
568. H. E. Benson in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 2**, Marcel Dekker, New York 1974, p. 159.
569. B. S. Grover, E. S. Holmes: "Nitrogen 86," *Br. Sulphur Conf.*, Amsterdam 1986, Proceedings of the Conf., p. 101.
570. R. K. Bartoo, S. J. Ruzicka: "Fertilizer '83," *Br. Sulphur Corp. 7th Int. Conf.*, London 1983, Proceedings of the Conf., p. 129.
571. US 4035 166, 1977 (F. C. van Hecke).
572. L. C. Crabs, R. Pouillard, F. C. van Hecke, 24th AIChE Ammonia Safety Symp., San Francisco 1989.
573. *Nitrogen* **180** (1989) 20 – 30.
574. C. C. Song *et al.*, 36th AIChE Ammonia Safety Symp., Los Angeles 1991.
575. *Hydrocarbon Process.* **61** (1982) no. 4, 95 – 102.
576. R. N. Maddox, M. D. Burns, *Oil Gas J.* **66** (1968) 4 91.
577. G. Giammarco, in A. V. Slack (ed.): *Ammonia*, **vol. 12**, Marcel Decker New York 1974, p. 171.
578. L. Tomasi, *Nitrogen* **199** (1992) 35.
579. K. E. Zarker in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 2**, Marcel Dekker, New York 1974 p. 219.
580. S. Linsmayer, *Chem. Tech. (Leipzig)* **24** (1972) no. 2, 74.
581. G. Hochgesand, *Ind. Eng. Chem.* **62** (1970) 7 – 37.
582. L. Dailey, in A. V. Slack, G. J. James (eds.): *Ammonia*, Marcel Decker New York 1974.
583. W. V. Korf, K. Thormann, K. Braßler, *Erdöl Kohle Erdgas, Petroch.* **16** (1963) 2, 94.
584. K. Elberling, W. Gabriel, *Chem. Tech. (Leipzig)* **29** (1977) no. 1, 43.
585. T. M. Gemborys, 39th AIChE Ammonia Safety Symp., Vancouver 1994.
586. M. J. Mitariten, C. M. Wolf, T. J. DePaola, ICI Catalco/UOP Hydrogen Plant Seminar, Chicago 1995.
587. P. Clough: "Asia Nitrogen," *Br. Sulphur Conf.*, Bali 1994, Proceedings of the Conf., p. 191 – 193.
588. J. N. Iyengar, D. E. Keene, 38th AIChE Ammonia Safety Symp., Orlando 1993.
589. US 4702 898 (B. S. Grover).
590. R. K. Bartoo, T. M. Gemborys: "Nitrogen 91," *Br.Sulphur Conf.*, Copenhagen 1991, Proceedings of the Conf., p. 127 – 139.
591. R. K. Bartoo, *Removing Acid Gas by the Benfield Process*, UOP Comp brochure, GP 5132–1M-11/93.
592. R. E. Meissner, U. Wagner, *Oil Gas J.* **81** (Feb. 7, 1983) 55. Links

593. K. Volkamer, E. Wagner, F. Schubert, *Plant Oper. Prog.* **1** (1982) no. 2, 134.
594. K. Volkamer, U. Wagner: "Fertilizer '83," *Br. Sulphur Corp. 7th Int. Conf.*, London 1983, Proceedings of the Conf., p. 139.
595. W. Gerhard, W. Hefner, 33rd AIChE Ammonia Safety Symp., Denver 1988.
596. D. W. Stanbridge, Y. Ide, W. Hefner, AIChE Ammonia Safety Symp., Anaheim 1984.
597. R. Welker, R. Hugo, R. Meissner, 41st AIChE Ammonia Safety Symp., Boston 1996.
598. R. Welker, R. Hugo, B. Büchele: Asia Nitrogen 96, Singapore 1996.
599. G. Ripperger, J. C. Stover, AIChE Spring National Meeting, Boston 1996.
600. J. E. Nobles, N. L. Shay: "Nitrogen 88," *Br. Sulphur Int. Conf.*, Geneva 1988.
601. W. Hefner, R. E. Meissner, 37th AIChE Ammonia Safety Symp., San Antonio 1992.
602. K. J. Stokes, *Ammonia Plant Saf.* **22** (1980) 178.
603. K. J. Stokes, *Nitrogen* **131** (1981) 35. K. J. Stokes in A. I. More (ed.): *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 525.
604. J. P. Klein, *Erdöl Kohle Erdgas Petrochem.* **23** (1970) no. 2, 84.
605. D. K. Judd, *Hydrocarbon Process.* **57** (1978) no. 4, 122.
606. C. G. Swanson, *Ammonia Plant Saf.* **21** (1979) 152.
607. C. G. Swanson, F. C. Burkhard, *Ammonia Plant Saf.* **24** (1984) 16.
608. V. A. Shah, *Energy Progress* **8** (1988) 67 – 70.
609. R. J. Hernandez, T. L. Huurdemann: "Nitrogen 88," *Br. Sulphur Int. Conf.*, Geneva 1988.
610. T. L. Huurdemann, V. A. Shah, 34th AIChE Ammonia Safety Symp., San Franzisco 1989.
611. V. A. Shah, *Hydrocarbon Process.* **67** (1988).
612. R. J. Hernandez, T. L. Huurdemann, *Chem. Eng.* **5** (1989) 154 – 156.
613. *Fertilizer Focos* **5** (1988) 27.
614. J. L. Lewis, H. A. Truby, M. B. Pascoo, *Oil Gas J.* **72** (June 24, 1974) 120. Links
615. J. L. Lewis, H. A. Truby, M. B. Pascoo, *Hydrocarbon Process.* **58** (1979) no. 4, 112.
616. R. W. Bucklin, R. L. Schendel, *Energy Progr.* **4** (1984) no. 3, 137.
617. J. P. Cook in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 2**, Marcel Dekker, New York 1974, p. 171
618. B. Sehrt, P. Polster, *Chem. Tech. (Leipzig)* **32** (1980) no. 7, 345.
619. *The Rectisol Process for Gas Purification*, Lurgi brochure 1676 e/5.95/10.
620. *Rectisol for Gas Treating*, Lurgi Express Information O 1051/12.72.
621. G. Ranke, Linde Berichte aus Wissenschaft und Technik.
622. H. Haase, *Chem. Anlagen Verfahren* **10** (1970) 59.
623. B. G. Goar, *Oil Gas J.* **69** (July 12, 1971) 75.
624. A. L. Kohl, F. C. Riesenfeld: *Gas Purification*, Gulf Publ. Co., Houston, Tex. 1979.
625. G. Hochgesand, *Chem. Ing. Techn.* **40** (1968) 43.
626. S. Strelzoff, *Chem. Eng. (N.Y.)* **82** (1975) no. 19, 115.
627. D. Werner, *Chem. Ing. Techn.* **53** (1981) no. 2, 73.
628. K. G.Christensen, *Hydrocarbon Process.* **57** (1978) no. 2, 125.
629. H. Thirkell in A. V. Slack G. J. James (eds.): *Ammonia*, **vol. 2**, Marcel Dekker, New York 1974, p. 117.
630. S. Strelzow: *Technology and Manufacture of Ammonia*, Wiley, New York 1981, p. 193.
631. F. C. Brown, C. L. Leci, *Proc. Fertilizer Soc.* **210** (1982) 1.
632. *Nitrogen* **180** (1989) 20.
633. *Nitrogen* **229** (1997) 37 – 51.
634. K. F. Butwell *et al.*, *Chem. Eng. Prog.* **69** (1973) no. 2, 57 – 61. Links
635. *Nitrogen* **96** (1975) 33.
636. *Nitrogen* **102** (1976) 40.
637. K. F. Butwell, D. J. Kubek, P. W. Sigmund, *Ammonia Plant Saf.* **21** (1979) 156; *Chem. Eng. Prog.* **75** (1979) no. 2, 75. K. F. Butwell, D. J. Kubek, *Hydrocarbon Process.* **56** (1977) no. 10, 173.
638. BASF, DE-AS 1250792, 1959.
639. B. W. Burklow, R. L. Coleman, *Ammonia Plant Saf.* **19** (1977) 21.
640. D. W. Allen, W. H. Yen, *Ammonia Plant Saf.* **15** (1973) 96.
641. A. J. M. Janssen, N. Siraa, J. M. Blanken, *Ammonia Plant Saf.* **23** (1981) 19.
642. *Chemical Economy and Engineering Review*, **12** (6 – 7), 33 – 36, (Jun/Jul 1980).
643. C. M. Buckthorp, *Nitrogen* **113** (1978) 34.
644. *Nitrogen* **123** (1980) 39.
645. J. C. Bonacci, T. G. Otchy, *Ammonia Plant Saf.* **20** (1978) 165.
646. J. H. Colby, G. A. White, P. N. Notwick, *Ammonia Plant Saf.* **21** (1979) 138.
647. *Nitrogen* **197** (1992) 18 – 25.
648. P. Soregaard-Andersen, O. Hansen, 36th AIChE Ammonia Safety Symp., Los Angeles 1991.
649. *Nitrogen* **217** (1995) 41 – 48.
650. US 3442 613, 1969 (B. J. Grotz).
651. *Nitrogen* **144** (1983) 30.
652. B. J. Grotz, *Nitrogen* **100** (1976) 30.
653. *Nitrogen* **182** (1989) 25.
654. B. J. Grotz, G. Good, (1980) *Chem. Age (London)* (November, 14), 18.
655. C. K. Wilson *et al.*, *Nitrogen* **151** (1984) 31.
656. W. Scholz, *DECHEMA-Monogr.* **58** (1968) 31.
657. W. Förg, *Chem. Anlagen + Verfahren* 1970 (March), 33.
658. *Nitrogen* **95** (1975) 38.
659. S. Hahesa *et al.*, *Saf. Air Ammonia Plants* **8** (1966) 14.
660. S. R. Krolikowski, *Chem. Eng. (London)* 1965 no. 186, 40.
661. F. Corr, F. Dropp, E. Rudelstorfer, *Hydrocarbon Process.* **58** (1979) no. 3, 119.
662. D. H. Werner, G. A. Schlichthärle, *Ammonia Plant Saf.* **22** (1980) 12.
663. *Nitrogen* **121** (1979) 37.
664. J. L. Heck, *Oil Gas J.* **78** (Feb. 11, 1980) 122. Links
665. P. Taffe, J. Joseph, *Chem. Age* 1978 (October) 14.
666. P. R. Savage, *Chem. Eng. (N.Y.)* **82** (1978) no. 25, 68D.
667. R. Rehder, P. Stead, FAI Seminar 1985, The Fertilizer Association of India, New Dehli, Tech II-2/1.
668. W. F. van Weenen, J. Tielroy, *Nitrogen* **127** (1980) 38.
669. *Oil Gas J.* **19** (1981) 270.

670. W. F. van Weenen, J. Tielroy, *Chem. Age India* **31** (1980) no. 12, Dev-2/1.
671. J. Voogd, FAI Seminar 1985, The Fertilizer Association of India, New Dehli, Tech 11–1/1.
672. D. J. Carra, R. A. McAllister, *Chem. Eng. (N.Y.)* **70** (1963) 62.
673. "Compressors" in *Encyclopedia of Chemical Engineering*, **Vol. 10**, Dekker, New York 1979, 157 – 409.
674. E. E. Ludwig, "Compressors" in *Applied Design for Chemical and Petrochemical Plants*, **Vol. 3**, Gulf. Houston, 1983, pp. 251 – 396.
675. G. A. J. Begg, *Chem. Process Eng. (London)* **49** (1968) no. 1, 58.
676. F. Fraschetti *et al.*, *Quad. Pignone* 1967, no. 9, 5.
677. C. A. Vancini: *La Sintesi dell Amoniaca*, Hoepli, Milano 1961, p. 497.
678. J. L. Kennedy, *Oil Gas J.* **65** 1967, no. 46, 105; no. 48, 95; no. 1, 72; no. 51, 76. *Oil Gas J.* **66** 1968, no. 4, 76.
679. *Chem. Eng. (N.Y.)* **73** (1966) no. 9, 126.
680. P. Kerklo, *Hydrocarbon Process.* **61** (1982) no. 10, 112.
681. H. Strassmann, Uhde Ammonia Symp., Dortmund 1992.
682. C. Dickinson, *Why Dry Running Gas Seals?* J. Crane Inc. Comp. Publ.
683. W. Koch, *Dry Gas Seal — Principles — Capabilities — Application*, J. Crane Inc. Comp. Publ.
684. S. Gray, B. Baxter, G. Jones, "Magnetic Bearings can Increase Availability, Reduce Oil and Maintenance Costs," Power Engineering, 1990.
685. R. R. Poricha, D. G. Rao, *Technology (Sindri, India)* **3** (1966) 96.
686. H. Förster, *Chem. Tech. (Leipzig)* **23** (1971) 157.
687. *Hydrocarbon Process. Pet. Refiner* **45** (1966) no. 5, 179.
688. *Eur. Chem. News* **11** (1967) no. 270, 26.
689. W. Plötner, *Chem. Tech. (Leipzig)* **24** (1972) 324.
690. A. F. Wilck, *Energy Tech.* **23** (1971) no. 5, 161.
691. I. M. Kalnin, *Luft Kältetech.* **8** (1972) no. 3, 142.
692. W. L. Luther, *Ingenieur Digest* **10** (1971) no. 1, 60.
693. F. Fraschetti, U. Filippini, P. L. Ferrara, *Quad. Pignone* **9** (1967) 5.
694. G. A. J. Begg, *Chem. Proc. Eng. (London)* **49** (1968) no. 1, 58.
695. A. Vitti, *Het Ingenieursblad* **40** (1971) 619.
696. P. L. Ferrara, A. Tesei, *Quad. Pignone* **25** (1978) 131.
697. J. Salviani et al., in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 2** , Marcel Dekker, New York 1974, p. 1.
698. *Nitrogen* **55** (1968) 37.
699. S. Labrow, *Chem. Proc. Eng. (London)* **49** (1968) no. 1, 55.
700. H. Foerster, *Chem. Technol.* **23** (1971) 93.
701. H. Foerster, *Chem. Technol.* **23** (1971) 157.
702. W. Plötner, *Chem. Technol.* **24** (1972) 324.
703. *Nitrogen* **181** (1981) 23.
704. H. E. Gallus, *Brennst. Wärme Kraft* **23** (1971) no. 4, 172.
705. S. D. Caplow, S. A. Bresler, *Chem. Eng. (New York)* **74** (1967) no. 7, 103.
706. K. J. Stokes, *Chem. Eng. Prog.* **75** (1979) no. 7, 88.
707. GB 1134 621, 1968 (D. R. Twist, D. W. Stanbridge).
708. S. Ujii, M. Ikeda, *Hydrocarbon Process.* **60** (1981) no. 7, 94.
709. W. Rall, 35th AIChE Ammonia Safety Symp., San Diego 1990.
710. R. F. Baddour *et al.*, *Chem. Eng. Sci.* **20** (1965) 281. Links
711. C. van Heerden, *Ind. Eng. Chem.* **45** (1953) 1242.
712. H. Inoue, T. Komiya, *Int. Chem. Eng.* **8** (1968) no. 4, 749.
713. H. Bakemeier *et al.*, *Chem. Ing. Tech.* **37** (1965) 427.
714. I. Porubszky *et al.*, *Br. Chem. Eng.* **14** (1969) no. 4, 495.
715. F. Horn, *Chem. Ing. Tech.* **42** (1970) no. 8, 561.
716. V. Hlavacek, *Ind. Eng. Chem.* **62** (1970) no. 7, 8.
717. J. Kjaer: *Measurement and Calculation of Temperature and Conversion in Fixed-bed Catalytic Reactors*, Gjellerup, Kopenhagen 1958.
718. G. F. Froment, *Chem. Ing. Tech.* **46** (1974) no. 9, 374.
719. J. E. Jarvan, *Ber. Bunsenges. Phys. Chemie* **74** (1970) no. 2, 74.
720. J. Kjaer, Computer Methods in Catalytic Reactor Calculations, Haldor Topsøe, Vedaek 1972.
721. J. E. Jarvan, *Oil Gas J.* **76** (1978) no. 5, 178.
722. J. E. Jarvan, *Oil Gas J.* **76** (1978) no. 3, 51. Links
723. C. P. P. Singh, D. N. Saraf, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) no. 3, 364.
724. J. Simiceanu, C. Petrila, A. Pop, *Chem. Tech. (Leipzig)* **35** (1983) no. 12, 628.
725. K. Lukas, D. Gelbin, *Chem. Eng. (London)* **7** (1962) 336.
726. M. J. Shah, *Ind. Eng. Chem.* **59** (1967) 72.
727. L. D. Gaines, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) no. 3, 381.
728. L. D. Gaines, *Chem. Eng. Sci.* **34** (1979) 37. Links
729. K. V. Reddy, A. Husain, *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) no. 3, 359.
730. L. M. Shipman, J. B. Hickman, *Chem Eng. Progr.* **64** (1968) 59.
731. H. Hinrichs, J. Niedetzky, *Chem. Ing. Tech.* **34** (1962) 88.
732. R. F. Baddour *et al.*, *Chem. Eng. Sci.* **20** (1965) 281. Links
733. P. L. Brian, R. F. Baddour, J. P. Emery, *Chem. Eng. Sci.* **20** (1965) 297. Links
734. G. A. Almasy, P. Jedlovsky, *Chem. Eng. (London)* **12** (1967) 1219.
735. I. Porubsky, E. Simonyi, G. Ladanyi, *Chem. Eng. (London)* **14** (1969) 495.
736. A. Murase, H. L. Roberts, A. O. Converse, *Ind. Eng. Chem. Process Des. Dev.* **9** (1970) 503.
737. D. Balzer *et al.*, *Chem. Technik (Leipzig)* **23** (1971) 513.
738. D. R. Levin, R. Lavie, *Ind. Chem. Symp. Ser.* **87** (1984) 393.
739. A. Nielsen: *Advances in Catalysis*, **vol. V** , Academic Press, New York 1953, p. 1.
740. L. Fodor, *Chim. Ind. Genie Chim.* **104** (1971) 1002.
741. J. B. Allen, *Chem. Proc. Eng. (London)* **46** (1965) 473.
742. L. B. Hein, *Chem. Eng. Progr.* **48** (1952) 412.
743. C. A. Vancini: *Synthesis of Ammonia*, Macmillan, London 1971.
744. I. Hay, G. D. Honti: *The Nitrogen Industry*, **vol. 1**, Akademiai Kiado, Budapest 1976.
745. *Nitrogen* **140** (1982) 30.
746. U. Zardi, *Hydrocarbon Process.* **61** (1982) no. 8, 129.

747. C. A. Vancini: *Synthesis of Ammonia*, Macmillan, London 1971, p. 336, 240, 244.
748. A. V. Slack, G. J. James, *Ammonia*, Marcel Dekker, New York 1974, p. 310.
749. J. J. Hay, I. M. Pallai, *Br. Chem. Eng.* **8** (1963) 171.
750. P. L. T. Brian *et al.*, *Chem. Eng. Sci.* **20** (1965) 297. Links
751. W. Hennel, *Chem. Tech. (Leipzig)* **15** (1963) 293.
752. US 2953 371, 1958 (A. Christensen, R. D. Rayfield).
753. US 3041 151, 1962 (A. Christensen).
754. US 3050 377, 1962 (A. Christensen).
755. US 2861 873, 1976 (G. A. Worn).
756. US 3032 139, 1976 (M. Vilceanu, C. Bors).
757. L. B. Hein, *Chem. Eng. Prog.* **48** (1952) no. 8, 412. Links
758. N. S. Zayarnyi, *Int. Chem. Eng.* **2** (1962) no. 3, 378.
759. A. Murase *et al.*, *Ind. Eng. Chem. Process Des. Dev.* **9** (1970) no. 4, 503.
760. J. B. Allen, *Chem. Process Eng. (London)* **46** (1965) no. 9, 473.
761. H. W. Graeve, *Chem. Eng. Prog.* **77** (1981) no. 10, 54; *Ammonia Plant Saf.* **23** (1981) 78.
762. *Nitrogen* **81** (1973) 29, 37.
763. H. Hinrichs, H. Lehner, *Chem. Anlagen + Verfahren* 1972, no. 6, 55.
764. F. Förster, *Chem. Eng. (N.Y.)* **87** (1980) no. 9, 62.
765. *Nitrogen* **101** (1976) 42.
766. Topsøe Topics, June 1976.
767. U. Zardi, E. Commandini, C. Gallazzi in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 173.
768. Ammonia Casale, DE-OS 3146778, 1981 (U. Zardi, E. Commandini).
769. E. Commandini, U. Zardi: "Fertilizer '83", *Br. Sulphur Corp. 7th Int. Conf.*, London 1983, Proceedings of the Conf., p. 179.
770. O. J. Quartulli, G. A. Wagner, *Hydrocarbon Process.* **57** (1978) no. 12, 115.
771. S. E. Handman, J. R. LeBlanc, *Chem. Eng. Prog.* **79** (1983) no. 5, 56.
772. W. A. Glover, J. P. Yoars, *Hydrocarbon Process.* **52** (1973) no. 4, 165; *Ammonia Plant Saf.* **15** (1973) 77.
773. *Nitrogen* **82** (1973) 34.
774. M. W. Kellogg, US 4298589, 1981 (J. R. LeBlanc, R. B. Peterson).
775. P. Lesur, *Nitrogen* **108** (1977) 29.
776. *Nitrogen* **75** (1972) no. 75, 33.
777. US 3475 136, 1969 (G. P. Eschenbrenner, C. A. Honigsberg).
778. S. Strelzow: *Technology and Manufacture of Ammonia*, Wiley, New York 1981, p. 34.
779. US 3694 169, 1972 (R. Fawcett, A. W. Smith, D. Westwood).
780. D. E. Riddler, 16th AIChE Ammonia Safety Symp., Atlantic City 1971.
781. US 3633 179, 1972 (D. D. Metha, E. J. Miller).
782. R. Harmon, W. H. Isalski, *Ammonia Plant Saf.* **23** (1981) 39.
783. I. Dybkaer in H. I. Lasa (ed.): *Chemical Reactor Design and Technology*, Nasa ASI Series E: Applied Sciences, no. 110, Nijhoff Publ., Dordrecht 1986.
784. V. Vek, 7th Int. Symp. Large Chem. Plants, Brugge 1988, p. 77.
785. V. Vek, *Chem. Ing. Tech.* **45** (1973) 608.
786. V. Vek, *Ind Eng. Chem. Process Des. Dev.* **16** (1977) 412.
787. DE 3334 777, 1984 (K. Ohasaki, J. Zanma, H. Watanabe).
788. C. A. Combs, *Ammonia Plant Saf.* **23** (1981) 32.
789. *Nitrogen* **31** (1964) 22.
790. T. Wett, *Oil Gas J.* **69** (1971) 70.
791. *Chem. Eng (New York)* **78** (1971) 90.
792. US 3372 988, 1968 (H. J. Hansen).
793. A. Nielsen, 16th AIChE Ammonia Safety Symp., Atlantic City 1971.
794. A. V. Slack in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 3**, Marcel Dekker, New York 1974, p. 345.
795. F. Horn, L. Küchler, *Chem. Ing. Tech.* **31** (1959) 1.
796. A. V. Slack in A. V. Slack, G. J. James (eds.): *Ammonia*, **vol. 3**, Marcel Dekker, New York 1974, p. 354.
797. US 3754 078, 1975 (H. Hinrichs *et al.*).
798. US 3918 918, 1975 (H. B. Kohn, G. Friedman).
799. H. Bakemeier, R. Krabetz, *Chem. Ing. Tech.* **34** (1962) 1.
800. US 3567 404, 1971 (L. C. Axelrod *et al.*).
801. G. P. Eschenbrenner, G. A. Wagner, 16th AIChE Ammonia Safety Symp., Atlantic City 1971.
802. R. Jackson, *Chem. Eng. Sci.* **19** (1964) 19, 253. Links
803. C. A. Vancini: *Synthesis of Ammonia*, Macmillan, London 1971 p. 232.
804. US 2898 183, 1959 (G. Fauser).
805. S. Strelzow: *Technology and Manufacture of Ammonia*, Wiley, New York 1981, p. 27.
806. D. C. Dyson *et al.*, *Canadian J. Chem. Eng. Sci.* **45** (1967) 310.
807. I. Dybkjaer, E. A. Gam, 29th AIChE Ammonia Safety Symp., San Francisco 1994.
808. I. Dybkjaer, E. A. Gam, *CEER Chem. Econ. Eng. Rev.* **16** (1984) 29.
809. Haldor Topsøe, US 4181 701, 1980 (E. A. Gam).
810. Haldor Topsøe, US 2710 247, 1981.
811. I. Dybkjaer, J. E. Jarvan, Nitrogen 97, British Sulphur Conference, Geneva 1997.
812. L. D. Gaines, *Ind. Eng. Chem. Process Des. Dev.* **16** (1977) no. 3, 381.
813. *Nitrogen* **169** (1987) 33.
814. E. Commandini, U. Zardi, Fertilizer Latin America, Int. Conf. British Sulphur , Caracas 1989.
815. L. Sutherland, B. Wallace, 33rd AIChE Ammonia Safety Symp., Denver 1988.
816. G. Pagani, U. Zardi, FAI Seminar 1988, The Fertilizer Society of India, New Dehli, S II/1–19.
817. K. A. Clayton, N. Shannahan, B. Wallace, 35th AIChE Ammonia Safety Symp., San Diego 1990.
818. N. Shannahan, *Hydrocarbon Process* **68** (1989) 60.
819. GB 1134 621, 1989 (D. R. Twist, D. W. Stanbridge).
820. W. A. Glover, J. P. Yoars, *Hydrocarbon Process* **52** (1973) 165.
821. US 3721 532, 1973 (L. E. Wright, A. E. Pickford).
822. US 3851 046, 1975 (L. E. Wright, A. E. Pickford).
823. US 4 744 966, 1988 (B. J. Grotz).

824. EP 268 469, 1988 (B. J. Grotz).
825. K. C. Wilson, B. J. Grotz, *J. Tech. Dev.* **14** (1990) 54.
826. B. J. Grotz, L. Grisolia, *Nitrogen* **199** (1992) 39.
827. H. Stahl, *Symposium of the Fertilizer Society*, London 1982, Proceedings, p. 61.
828. US 4452760, 1984 (R. B. Peterson, R. Finello, G. A. Denavit).
829. US 3892 535, 1975 (W. Hennel, C. Sobolewsk).
830. I. Dybkjaer in A. Nielsen: *Ammonia — Catalysis and Manufacture*, Springer-Verlag, New York 1995, p. 251.
831. Ammonia — Energy Integration in Ammonia Plants, Uhde Engineering News 2–91, Hi 111 9 1000 91, 1991.
832. M. J. P. Bogart in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 141.
833. G. R. Prescott *et al.*, 33rd AIChE Ammonia Safety Symp., Denver 1988.
834. NH_3—Gas—Synthesis, Balke-Dürr comp. brochure, CBA 100o-7.91, 1991.
835. M. Podhorsky *et al.*, 40th AIChE Ammonia Safety Symp., Tucson 1995; T. Timbres *et al.*, 34th AIChE Ammonia Safety Symp, San Francisco 1989.
836. Synloop Wast Heat Boiler in Ammonia Plants — Unique Hot/Cold Tubesheet Design, Babcock Borsig comp. brochure.
837. K. Nassauer, M. Fix, AFA/Abu — Qir Ammonia/Urea Technol. Symp., 1996.
838. T. Timbres *et al.*, 34th AIChE Ammonia Safety Symp, San Francisco 1989.
839. M. J. P. Bogart, *Plant Oper. Prog.* **1** (1982) no. 3, 147.
840. H. W. Graewe, 25th AIChE Ammonia Safety Symp., Portland 1980.
841. H. W. Graewe, *Chem. Eng. Progr.* **77** (1981) no. 10, 54.
842. H. Lachmann, D. Fromm, 32nd AIChE Ammonia Safety Symp., Minneapolis 1987.
843. D. L. Johnson, T. M. O'Connor, 31st AIChE Ammonia Safety Symp., Boston 1986.
844. *Nitrogen* **176** (1988) 25.
845. M. J. P. Bogart, *Hydrocarbon Process.* **57** (1978) no. 4, 145.
846. I. Dybkjaer in A. Nielsen: *Ammonia — Catalysis and Manufacture*, Springer-Verlag, New York 1995, p. 231.
847. W. Malewski, *Chem. Ztg.* **95** (1971) 186.
848. A. Finn, *Nitrogen* **175** (1988) 25 – 32.
849. W. H. Isalski, *Nitrogen* **152** (1984) 100.
850. R. Harmon, W. H. Isalski, 25th AIChE Ammonia Safety Symp., Portland 1980.
851. C. A. Combs, 25th AIChE Ammonia Safety Symp., Portland 1980.
852. R. Fabian, D. Tilman, *Linde Ber. Techn. Wiss.* **59** (1986) 6.
853. K. Bohlscheid, *Chem. Prod.* **8** (1979) no. 3.
854. A. K. Fritzsche, R. A. Narayan, *CEER Chem. Econ. Eng. Rev.* **19** (1987) 19.
855. G. Holldorff, *Hydrocarbon Process.* **58** (1979) no. 7, 149.
856. W. A. Koros, R. A. Narayan, *Chem. Eng. Progr.* (1995) 68 – 81.
857. *Hyrocarbon Process.* **62** (1983) 43 – 62.
858. G. Q. Miller, M. J. Mitariten, ICI Catalco/KTI/UOP Hydrogen Plant Seminar, Chicago 1995.
859. *Nitrogen* **121** (1979) 37 – 43.
860. G. Low, *Nitrogen* **147** (1984).
861. J. J. Sheridan, III, *et al.*, *Less Common Met.* **89** (1983). 447.
862. Pullman Inc., DE-OS 2741851, 1978 (C. L. Becker).
863. O. J. Quartulli, J. B. Fleming, J. A. Finneran, *Hydrocarbon Process.* **47** (1968) 153.
864. O. J. Quartulli, J. B. Fleming, J. A. Finneran, *Nitrogen* **58** (1969) 25.
865. I. Dybkjaer in A. Nielsen (ed.): *Ammonia — Catalysis and Manufacture*, Springer-Verlag, New York 1995, p. 226, 227.
866. G. Pagani, U. Zardi, *Hydrocarbon Process.* **51** (1972) no. 7, 106 – 110.
867. F. Saviano, V. Lagana, P. Bisi, *Hydrocarbon Process.* **60** (1981) no. 7, 99.
868. H. Hinrichs, *Chem. Ztg. Chem. Appar.* **86** (1962) 223.
869. H. Neth *et al.*, *Chem. Eng. Prog.* **78** (1982) no. 7, 69.
870. *Chem. Week* **116** (Feb. 19, 1975) 29.
871. API Publication 941, 2nd ed., American Petroleum Institute, Washington, D.C., June 1977.
872. B. Szantay, E. Jahab in G. D. Honti (ed.): *The Nitrogen Industry*, Akademiai Kiado, Budapest 1976, p. 701 – 706.
873. C. A. Vancini: *La Sintesi dell Amoniaca*, Hoepli, Milano 1961, p. 769.
874. *Nitrogen* **16** (1962) 48.
875. Steel for Hydrogen Service at Elevated Temperatures and Pressures in Petrochemical Refineries and Petrochemical Plants, API Publ. 941, Am. Petrol. Inst. 1990.
876. W. Hausmann *et al.*, *Stahl Eisen* **107** (1987) no. 12, 45 – 53.
877. B. Granville, *Welding in the World* **31** (1993) no. 5, 308.
878. W. A. Bonner, *Hydrocarbon Process.* (1951) no. 5, 165.
879. G. A. Nelson, *Transactions of ASME* **56** (1977) 205 – 213.
880. I. Class, *Stahl Eisen* **80** (1960) 11, 17.
881. I. Class, *Stahl Eisen* **85** (1965) 149.
882. A. Heuser, 37th AIChE Ammonia Safety Symp., Los Angeles 1991.
883. G. R. Prescott, 37th AIChE Ammonia Safety Symp., Los Angeles 1991.
884. G. H. Wagner, A. Heuser, G. Heinke, 37th AIChE Ammonia Safety Symp., Los Angeles 1991.
885. G. R. Prescott, B. J. Grotz, 39th AIChE Ammonia Safety Symp., Vancouver 1994.
886. G. R. Prescott, *Plant Oper. Prog.* **1** (1982) no. 2, 94.
887. A. W. Loginow, E. H. Phelps, *Corrosion-NACE* **31** (1975) no. 11, 404.
888. H. D. Marsch, *Plant Oper. Prog.* **1** (1982) no. 3, 152.
889. U. Jäkel, W. Schwenk, *Werkst. Korros.* **22** (1971) no. 1, 1.
890. S. Y. Sathe, T. M. O'Connor 32nd AIChE Ammonia Safety Symp., Denver, 1988.
891. C. A. van Grieken, 33rd AIChE Ammonia Safety Symp., Minneapolis, 1987.
892. Y. Murakami, T. Nomura, J. Watanabe, *MPC/ASTM Symposium on the Application of 2 1/4 Cr-1 Mo Steel for Thick Wall Pressure Vessels*, Denver, Col. 1980. J. Watanabe *et al.*, 29th Petroleum Mechanical Engineering Conf., Dallas, Tex. 1974. G. Grote, *Chem. Ing. Tech.* **55** (1983) no. 2, 93.
893. R. Bruscato, *Weld. J. (Miami)* **49** (1970) no. 4, 148.
894. J. A. Richardson, *Nitrogen* **205** (1993) 49 – 52.
895. H. Stahl, S. G. Thomson, 40th AIChE Ammonia Safety Symp., Tucson 1995.
896. T. Shibasaki *et al.*, 40th AIChE Ammonia Safety Symp., Tucson 1995.
897. R. J. Gommans, T. L. Huurdeman, 39th AIChE Ammonia Safety Symp., Vancouver 1994.
898. J. DeJaeger, L. Guns, J. Korkhaus, 39th AIChE Ammonia Safety Symp., Vancouver 1994.
899. J. A. Richardson, Boudouard Carbon and Metal Dusting, ICI Catalco Tech. paper.

900. O. J. Dunmore, *Proc. of UK Corrosion Conf.*, 1982.
901. A. W. Loginow, E. H. Phelps, *Corrosion (Houston)* **18** (1962) 299.
902. H. Grabke, R. Krajak, J. C. Nava Paz, *Corrosion Science* **35** (1993) 1141. Links
903. R. F. Hochmann, 4th Int. Congr. on Metal Corrosion, Nat. Corr. Engineers, 1972.
904. L. Lunde, *Ammonia Plant Saf.* **24** (1984) 154. J. M. Blanken, *Ammonia Plant Saf.* **24** (1984) 140.
905. *Plant Oper. Prog.* **2** (1983) no. 3, 247.
906. Dechema-Werkstoff-Tabelle/Chem. Beständigkeit, keyword “Ammoniak,” sheet 4, Nov. 1978.
907. I. Class, K. Gering, *Werkst. Korros.* **25** (1974) no. 5, 314.
908. A. W. Loginow, *Mater. Perform.* **15** (1976) no. 6, 33.
909. B. E. Wilde, *Corrosion-NACE* **37** (1981) no. 3, 131.
910. R. S. Brown, *Plant Oper. Prog.* **1** (1982) no. 2, 97.
911. T. Kawamoto *et al.*, *IHI Eng. Rev.* **10** (1977) no. 4, 17.
912. N. K. Roy, *Fert. Technol.* **18** (1981) no. 1+2, 1.
913. P. B. Ludwigsen, H. Arup, *Corrosion (Houston)* **32** (1976) no. 11, 430.
914. E. H. Phelps, *Ammonia Plant Saf.* **16** (1974) 32.
915. E. H. Phelps, 18th AIChE Ammonia Safety Symp., Vancouver 1973.
916. E. A. Olsen, 13th AIChE Ammonia Safety Symp., Montreal 1968.
917. E. H. Phelps, 16th AIChE Ammonia Safety Symp., Atlantic City 1971.
918. H. Arup, 21st AIChE Ammonia Safety Symp., Atlantic City 1976.
919. L. Lunde, R. Nyborg, 27th AIChE Ammonia Safety Symp., Boston 1986; *Plant/Oper. Progr.* **6** (1967) 11 – 16.
920. J. D. Stephens, F. Vidalin, 32nd AIChE Ammonia Safety Symp., Minneapolis 1987.
921. L. Lunde, R. Nyborg, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
922. J. Blanken, 22nd AIChE Ammonia Safety Symp., Denver 1983;
923. L. Lunde, R. Nyborg, The Fertilizer Society, London 1991, Proc. no. 307.
924. A. Haslam, P. Brook, H. Isalski, L. Lunde, *Hydrocarbon Process.* **55** (1976) no. 1, 103.
925. J. R. Byrne, F. E. Moir, R. D. Williams, 33rd AIChE Ammonia Safety Symp., Denver 1988.
926. M. Appl *et al.*, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
927. R. A. Selva, A. H. Heuser, 21st AIChE Ammonia Safety Symp.,
928. K. A. van Krieken, 16th AIChE Ammonia Safety Symp., Atlantic City 1976.
929. A. Cracknell, 24th AIChE Ammonia Safety Symp., San Francisco 1979.
930. R. S. Brown, *Plant/Oper. Progr.* **1** (1982) 97.
931. D. C. Guth, D. A. Clark, 29th AIChE Ammonia Safety Symp., San Francisco 1994.
932. D. C. Guth, D. A. Clark, *Plant/Oper. Progr.* **4** (1985) 16.
933. S. Hewerdine, The Fertilizer Society, London 1991, Conf. Proc. no. 308, p. 1.
934. B. G. Burke, D. E. Moore, 21st AIChE Ammonia Safety Symp., San Francisco 1989.
935. M. J. Conley, S. Angelsen, D. Williams, 35th AIChE Ammonia Safety Symp., San Diego 1990.
936. *Nitrogen* **102** (1976) 35.
937. R. Banks, *Chem. Eng. (N.Y.)* **84** (1977) no. 21, 90.
938. A. A. Haslam, W. H. Isalski, *Ammonia Plant Saf.* **17** (1975) 80.
939. R. Harmon in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 113.
940. R. Banks, *Ammonia Plant Saf.* **20** (1978) 79.
941. *Oil Gas J.* **77** (March 5, 1979) 182.
942. D. L. MacLean, C. E. Prince, Y. C. Chae, *Chem. Eng. Prog.* **76** (1980) no. 3, 98; *Ammonia Plant Saf.* **22** (1980) 1.
943. Y. C. Chae, G. S. Legendre, J. M. van Gelder in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 457.
944. *Nitrogen* **130** (1981) 40.
945. *Nitrogen* **136** (1982) 29.
946. R. L. Schendel, C. L. Mariz, J. Y. Mak, *Hydrocarbon Process.* **62** (1983) no. 8, 58.
947. D. L. MacLean, D. J. Stockey, T. R. Metzger, *Hydrocarbon Process.* **62** (1983) no. 8, 47.
948. H. Knieriem, Jr., *Hydrocarbon Process.* **59** (1980) no. 7, 65.
949. M. D. Rosenzweig, *Chem. Eng. (N.Y.)* **88** (1981) no. 24, 62.
950. G. Schulz, H. Michele, U. Werner, *Chem. Ing. Tech.* **54** (1982) no. 4, 351.
951. Union Carbide Corp., US 4077780, 1976.
952. *Chem. Eng. (London)* 1979, no. 345, 395.
953. K. Knoblauch *et al.*, *Erdöl Kohle Erdgas Petrochem.* **32** (1979) 551.
954. J. G. Santangelo, G. T. Chen, *Chemtech* **13** (1983) no. 10, 621.
955. R. J. Berry, *Chem. Eng. (N.Y.)* **86** (1979) no. 15, 62.
956. W. H. Isalski, G. J. Ashton in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 125.
957. S. Lynn, C. Alesandrini, *Ammonia Plant Saf.* **16** (1974) 80.
958. G. P. Williams, W. W. Hoehing, *Chem. Eng. Prog.* **79** (1983) no. 3, 11.
959. O. J. Quartulli, W. Turner, *Nitrogen* **80** (1972) 28.
960. O. J. Quartulli, W. Turner, *Nitrogen* **81** (1973) 32.
961. L. Silberring, *Nitrogen* **120** (1979) 35.
962. P. Hinchley, *Chem. Eng. (N.Y.)* **86** (1979) no. 17, 120. P. Hinchley, *Proc. Inst. Mech. Eng.* **193** (1979) no. 8.
963. H. Lachmann, *Ammonia Plant Saf.* **23** (1981) 51.
964. H. Weber *et al.*, *Chem. Ing. Tech.* **56** (1984) no. 5, 356.
965. *Nitrogen* **162** (1986).
966. J. G. Livingstone, A. Pinto, 27th AIChE Ammonia Safety Symposium, Los Angeles 1982.
967. W. K. Taylor, A. Pinto, 31st AIChE Ammonia Safety Symposium, Boston 1986.
968. “Ammonia – Uhde's low energy technology”, *Uhde engineering news 1 – 91 (1991).*
969. O. J. Quartulli, D. Wagener, *Erdöl Kohle Erdgas Petrochem.* **26** (1973) no. 4, 192.
970. U. Zardi, A. Antonini, *Nitrogen* **122** (1979) 33.
971. G. D. Honti, 4th Int. Conf on Fertilizer Technol., London 1991, Conf. Proc., p. 1.
972. F. Saviano, W. Lagana, P. Bisi, *Hydrocarbon Process.* **60** (1981) 99.
973. I. Dybkjaer, ECN Europ. Chemical News: Fertilizers 83 (suppl.), 1983, p.15.
974. G. R. James, K. J. Stokes, *Chem. Eng. Progr.* **62** (1984) 81.
975. J. M. Blanken, 33rd AIChE Ammonia Safety Symp., Denver 1988.
976. F. C. Brown, *Nitrogen* **100** (1976) 65.

977. K. J. Mundo, *Erdöl, Kohle, Erdgas Petrochemie* **31** (1978) 74.
978. *Nitrogen* **182** (1989) 25.
979. J. B. LeBlanc, M. N. Shah, L. J. Buividas, *Hydrocarbon Process.* **59** (1980) no. 4, 68-G.
980. *Hydrocarbon Process.* **62** (1983) no. 11, 81.
981. P. Tjissen, 21nd AIChE Ammonia Safety Symp., Atlantic City 1976.
982. G. R. Nieman, L. C. Daigre, III, 18th AIChE Ammonia Safety Symp., Vancouver 1973.
983. C. C. Yost, C. R. Curtis, C. J. Ryskamp, 24th AIChE Ammonia Safety Symp., San Francisco 1979.
984. *Nitrogen* **65** (1970) 32.
985. T. L. Huurdeman, 33rd AIChE Ammonia Safety Symp., Denver 1988.
986. R. L. Allen, Jr., G. A. Moser, 36th AIChE Ammonia Safety Symp., Los Angeles 1991.
987. D. Dekmush *et al.*, 37th AIChE Ammonia Safety Symp., San Antonio 1992.
988. E. Futterer, E. Pattas, *Chem. Ztg.* **98** (1974) no. 9, 438.
989. S. Madhaven, *Plant /Oper. Progr.* **3** (1984) no. 1, 14.
990. S. M. Solomon, 29th AIChE Ammonia Safety Symp., San Francisco 1984.
991. S C. Moore, T. M. Piper, C. C. Chen, 30th AIChE Ammonia Safety Symp., Seattle 1985.
992. *Nitrogen* **169** (1987) 35.
993. S. Mani, S. K. Shoor, H. S. Pedersen, 33rd AIChE Ammonia Safety Symp., Denver 1988; *Plant /Oper. Progr.* **8** (1989) 33.
994. G. Grossman, J. Dejaeger, 37th AIChE Ammonia Safety Symp., Los Angeles 1991.
995. K.-J. Mundo, *Chem. Anlagen + Verfahren* 1972, no. 6, 49.
996. I. Dybkjaer in A. Nielsen (ed.): *Ammonia Catalysis and Manufacture*, Springer, New York 1995 p. 262, 263.
997. I. Dybkjaer in A. Nielsen (ed.): *Ammonia Catalysis and Manufacture*, Springer, New York 1995 p. 258 – 261; 269 – 271 .
998. P. Radke, *Nitrogen* **225** (1997) 27.
999. J. S. Campbell, J. W. Marshall, *Nitrogen* 1976, no. 103, 33.
1000. L. J. Buividas, *Hydrocarbon Process.* **58** (1979) no. 5, 257.
1001. G. D. Honti in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 1.
1002. D. C. Thomson, 4th Int. Conf. on Fertilizer Technol., London 1991, Conf. Proc., part 1, p. 1.
1003. J. A. Tonna, F. C. Brown, T. W. Nurse, *Nitrogen 91: Brit. Sulphur Conf.*, Copenhagen 1991, p. 127 – 139.
1004. S. Madhavan, B. Landry, *Nitrogen 91: Brit. Sulphur Conf.*, Copenhagen 1991, p. 127 – 139.
1005. G. R. James, 32nd AIChE Ammonia Safety Symp., Minneapolis 1987. USA 1981. G. R. James, K. J. Stokes, *Chem. Eng. Prog.* **80** (1984) no. 6, 33.
1006. R. W. Parrish, *Process Improvements*, Gulf Coast Ammonia Producers Meeting, Baton Rouge 1991.
1007. J. R. LeBlanc, *Hydrocarbon Process.* **65** (1986) no. 8, 39 – 44.
1008. M. Jung, *Nitrogen* **191** (1991) 42 – 52.
1009. *Nitrogen* **141** (1983) 38.
1010. F. C. Brown, *ECN Europ. Chem. News, Proc. Rev.* **8** (1982).
1011. I. Dybkjaer, Fertilizer Latin America, Int. Conf. Brit Sulphur, Caracas 1989.
1012. A. Nielsen *et al.*, *Plant/Oper. Progr.* **1** (1982) no. 3, 186.
1013. H. Graewe, *Nitrogen 88: 12th Int. Conf. Brit. Sulphur*, Geneva 1988, p. 77.
1014. A. M. Dark, E. A. Stallworthy, *Nitrogen* **153** (1985) 25
1015. D. Singh, *Process Plant Eng.* **73** (1986) Jan — March.
1016. J. R. LeBlanc, D. O. Moore, R. V. Schneider, III, *Oil Gas J.* **80** (1982) 115. Links
1017. R. Darjat, J. R. LeBlanc, Nitrogen 88: 12th Int. Conf. Brit. Sulphur, Geneva 1988, p. 77.
1018. I. Dybkjaer, *Eur. Chem. News* 1983, Feb. 21, 15 (Fertilizers '83 Suppl.).
1019. H. Bendix, L. Lenz, 33rd AIChE Ammonia Safety Symp., Denver 1988.
1020. D. L. MacLean, C. E. Prince, Y. C. Chae, 24th AIChE Ammonia Safety Symp., San Francisco 1979; *Chem. Eng. Progr.* 1980, 98 – 104.
1021. R. G. Howerton, 24th AIChE Ammonia Safety Symp., San Francisco 1979.
1022. N. W. Patel, S. I. Wang, K. J. Kittelstad, 33rd AIChE Ammonia Safety Symp., Denver 1988.
1023. M. Tsujimoto *et al.*, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
1024. M. Tsujimoto *et al.*, *Nitrogen 91: Brit. Sulphur Conf.*, Copenhagen 1991, p. 127 – 139.
1025. A. Pinto, P. L. Rogerson, *Chem. Eng. Prog.* **73** (1977) no. 7, 95. Links
1026. R. L. Newland, J. D. Pierce, D. M. Borzik, 32nd AIChE Ammonia Safety Symp., Minneapolis 1987.
1027. *Fertilizer News* (1987) (December) 19.
1028. F. C. Brown, U. Zardi, G. Pangani, *Nitrogen 88: 12th Brit. Sulphur Int. Conf.*, Geneva 1988, p. 1159 – 139.
1029. T. Czuppon *et al.*, 38th AIChE Ammonia Safety Symp., Orlando 1993.
1030. K. J. Mundo, *Erdöl Kohle Erdgas Petrochem.* **31** (1978) no. 2, 74.
1031. D. Kitchen, A. Pinto, 35th AIChE Ammonia Safety Symp., San Diego 1990.
1032. A. Nielsen *et al.*, *Plant Oper. Prog.* **1** (1982) no. 3, 186.
1033. P. H. Brook: "Fertilizer '83," *Br. Sulphur Corp. 7th Int. Conf.*, London 1983, Proceedings of the Conf., p. 159.
1034. F. C. Brown, *Eur. Chem. News* 1982, no. 15, 10 (Process Review Suppl.).
1035. Pullman Inc., US 4079017 (D. B. Crawford, C. L. Becker, J. R. LeBlanc); US 4162290, 1979 (D. B. Crawford, C. L. Becker, J. R. LeBlanc).
1036. G. Pagani, G. Brusasco, G. Gramatica in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 195.
1037. *Jacobs Ammonia Technology*, Jacobs company brochure, 1996.
1038. T. Miyasugi *et al.*, *Chem. Eng. Prog.* **80** (1984) no. 7, 41. T. Miyasugi *et al.*, *Ammonia Plant Saf.* **24** (1984) 64.
1039. S. Uji, M. Ikeda, *Hydrocarbon Process.* **60** (1981) no. 7, 94.
1040. E. Nobles, J. C. Stover, *Chem. Eng. Prog.* **80** (1984) no. 1, 81. E. Nobles, J. C. Stover, *Ammonia Plant Saf.* **24** (1984) 41.
1041. *Chem. Eng. (N.Y.)* **86** (1979) no. 26, 88.
1042. *Oilweek (Calgary, Alberta)* (May 23, 1983) 12.
1043. *Chem. Week* **134** (March 21, 1984) 15.
1044. F. C. Brown in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 39.
1045. L. J. Ricci, *Chem. Eng. (N.Y.)* **86** (1979) no. 3, 50, 54.
1046. V. Pachaiyappan, *Fertilizer News* 1979, 41.
1047. *Oil Gas* J. 76 (1978) no. 49, 34.
1048. P. Taffe, *Chem. Age (London)* **117** (Dec. 15, 1978) 8.
1049. *CEER Chem. Econ. Eng. Rev.* **11** (1979) no. 5, 24.
1050. I. Dybkjaer in A. Nielsen (ed.): *Ammonia Catalysis and Manufacture*, Springer, New York 1995 p. 262, 263.
1051. I. Dybkjaer, Fertilizer Latin America Int. Conf., Caracas 1989; Brit. Sulphur Corp. Proc., vol. 1, p. 77.
1052. I. Dybkjaer, FAI Seminar 1990, The Fertilizer Assoc. of India, New Delhi, p. SIII–1.
1053. T. Bajpai, *Nitrogen 88, Brit. Sulphur 12th Int. Conf.*, Geneva 1988.

1054. T. S. Hariharan, *J. Tech. Dev.* **7** (1987) no. 4, 42.
1055. S. R. Sahore, T. S. Krishnan, *Fertilizer News* **15** (1989).
1056. I. Dybkjaer, *Fertilizer Industry Ann. Rev.* **XIII** (1990) 42a.
1057. I. Dybkjaer, IFA Technical Conf., The Hague 1992.
1058. I. Dybkjaer, IFA – FADINAP Regional Conf. for Asia and Pacific, Bali 1992
1059. *Hydrocarbon Process.* **62** (1983) no. 11, 79.
1060. F. Brown, Proceedings no. 218, The Fertilizer Society, London 1983.
1061. *Hydrocarbon Process.* **60** (1981) no. 11, 132.
1062. *The Uhde Reformer: High Pressure, High Temperature Service*Uhde Brochure Hi 18 1500 11/1991.
1063. Energy Efficient Ammonia Production *Uhde Eng. News.* **3** (1991)
1064. Ammonia Plant – Energy Integration *Uhde Eng. News.* **2** (1991).
1065. *Uhde's Ammonia Technology* Uhde Brochure Hi 18 1500 11/1991; RRD, 1992.
1066. Ammonia, Uhde's Low Energy Technology *Uhde Eng. News.* **1** (1991).
1067. R. Hakmann, FAI Seminar 1990, The Fertilizer Assoc. of India, New Delhi, p. SIII–3.
1068. J. Dejaeger, E. Das, 38th AIChE Ammonia Safety Symp., Orlando 1993.
1069. B. J. Grotz, *Nitrogen* **100** (1976) 71.
1070. P. A. Ruziska, C. C. Song, 29th AIChE Ammonia Safety Symp., San Francisco 1984.
1071. P. A. Ruziska, P. Dranze, C. C. Song, Nitrogen 86, Brit. Sulphur conf., Amsterdam 1986.
1072. B. Grotz, G. Good, *Chem. Age (London)* **121** (Nov. 14, 1980) S 18.
1073. *Nitrogen* **144** (1983) 30.
1074. ICI, EP-A 49967, 1982 (A. Pinto).
1075. J. G. Livingstone, A. Pinto, 27th AIChE Ammonia Safety Symp., Los Angeles 1982.
1076. Fertilizer Industry Annual Review, vol. XI, 1988
1077. S. A. Topham, S. A. Hall, D. G. Heath, ICI/CFDC Tech. Symp., Shanghai 1989.
1078. W. K. Taylor, A. Pinto, *Commisioning C–I–Ls Ammonia Plant*, 31st AIChE Ammonia Safety Symp., Boston 1986; *Plant/Operations Prog.* **6** (1987) 106 – 111.
1079. M. P. Robert, C. W. Hooper, ICI/CFDC Tech. Symp., Shanghai 1989.
1080. K. J. Elkins *et al.*, Asia Nitrogen, Int. Conf., (Brit. Sulphur) Bali 1994.
1081. K. J. Elkins *et al.*, *ICIs AMV Ammonia Technology* ICI Catalco tech. paper 246W/126/3/AMM.
1082. *Nitrogen* **162** (1989) 32 – 37.
1083. J. G. Livingstone, A. Pinto, Fertilizer 83, Int. Conf. (Brit. Sulphur), London 1988.
1084. *Nitrogen* **141** (1983) 37.
1085. *Chem. Week* **132** (Jan. 5, 1983) 23.
1086. *Hydrocarbon Process.* **62** (1983) no. 11, 80.
1087. D. L. Banquy, *Ammonia Plant Saf.* **24** (1984) 8.
1088. Foster Wheeler, US 4296 085, 1981 (D. L. Banquy).
1089. Pullman Inc., US 4148866, 1979 (C. L. Becker); US 4153673, 1979 (C. L. Becker).
1090. *Chem. Eng. (N.Y.)* **88** (1981) no. 10, 33.
1091. *Nitrogen* **178** (1989) 30 – 39.
1092. J. M. Halstead, A. Pinto, FAI Seminar, Delhi 1988.
1093. A. Pinto, J. M. S. Moss, T. C. Hicks, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
1094. K. Elkins, I. R. Barton, 39th AIChE Ammonia Safety Symp., Vancouver 1994; *Operational Performance of the ICI Leading Concept Ammonia (LCA) Process*, ICI Catalco tech. paper 274W/126/3/LCA.
1095. T. C. Hicks, A. Pinto, *Fertilizer News* (1989) 37.
1096. P. Conolly, *Chem. Age (London)* **122** (Feb. 6, 1981) 12. *Eur. Chem. News* **36** (1981) no. 967, 16.
1097. V. Lagana, *Chem. Eng. (N.Y.)* **85** (1978) no. 1, 37.
1098. C. L. Becker in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 537.
1099. J. R. LeBlanc *et al.*, *Oil Gas J.* **80** (1982) no. 38, 115. Links
1100. S. I. Wang, N. M. Patel, *Ammonia Plant Saf.* **24** (1984) 1. S. I. Wang, N. M. Patel, *Plant Oper. Prog.* **3** (1984) no. 2, 101.
1101. G. Low, *Nitrogen* **147** (1984) 1 – 32 (Suppl. “Revamping Ammonia Plants”).
1102. P. R. Savage, *Chem. Eng. (N.Y.)* **85** (1978) no. 25, 68 D.
1103. M. Spear, *Chem. Eng. (London)* 1979, no. 340, 29.
1104. W. F. van Weenen, J. Tielroy, *Nitrogen* **127** (1980) 38.
1105. *Oil Gas J.* **79** (1981) no. 18, 270.
1106. W. Armbruster, *Erdöl Kohle Erdgas Petrochem.* **33** (1980) no. 3, 118.
1107. S. Ratan, K. S. Jungerhans, KTI Symp., Scheveningen 1991.
1108. J. J. Westenbrink, J. Voogd, Nitrogen 86, Brit. Sulphur Int. Conf., Amsterdam 1986.
1109. J. J. Westenbrink, J. Voogd, FAI Seminar, New Delhi 1995.
1110. W. F. van Weenen, J. Tielroy, *Proc. Fertilizer Soc. (London)* (1980) no. 191, 1.
1111. G. Kammholz, G.-A. Müller, *Erdöl Kohle Erdgas Petrochem.* **26** (1973) no. 12, 695.
1112. *Nitrogen* **208** (1994) 44 – 49.
1113. M. Lembeck, Asia Nitrogen 96: Brit. Sulphur Int. Conf., Singapore 1996.
1114. M. Lembeck, *Fertilizer News* (1993) 15.
1115. *The Linde Ammonia Concept*, Linde Comp. publ. 1995.
1116. J. Ilg, B. Kandziora, 41st AIChE Ammonia Safety Symp., Boston 1996.
1117. H. E. Butzert, *Chem. Eng. Prog.* **72** (1976) no. 1, 56; Links *Ammonia Plant Saf.* **18** (1976) 78.
1118. *ECN Euro Chem News* **20** (1978) 39.
1119. F. C. Brown, Proc. Dev. Symp. 1978, The Inst. of Engineers, p. 1.
1120. E. Supp, *Nitrogen* **109** (1977) 36.
1121. *CEER Chem. Econ. Eng. Rev.* **12** (1980) no. 6 – 7, 33.
1122. J. Dybkjaer, J. Hansen *Chem. Age India* **31** (1980) (12): DEV. 3/1
1123. G. F. Skinner in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 491.
1124. R. A. Sharpe, *Hydrocarbon Process.* **55** (1976) no. 11, 171.
1125. L. J. Partridge, *Chem. Eng. Prog.* **72** (1976) no. 8, 57; *Ammonia Plant Saf.* **18** (1976) 73.
1126. S. McQueen, *Chem. Eng. (N.Y.)* **85** (1978) no. 25, 68 H.
1127. Lurgi Express Information Ö 1323/5.79, Ammonia Plant Based on Coal.
1128. *Chem. Eng. News* **57** (1977) no. 23, 27.
1129. G. W. Alves, D. A. Waitzmann, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 2, 70.
1130. D. Netzer, J. Moe, *Chem. Eng. (N.Y.)* **84** (1977) no. 23, 129.

1131. U. Buskies, F. Summers, *Erdöl Kohle Erdgas Petrochem.* **31** (1978) no. 10, 474.
1132. H. Schmidt-Traub, H. C. Pohl, *Chem. Ing. Tech.* **55** (1983) no. 11, 850. *Hydrocarbon Process.* **63** (1984) no. 4, 95.
1133. S. Weems, D. H. Ball, D. E. Griffin, *Chem. Eng. Prog.* **75** (1979) no. 5, 64; *Ammonia Plant Saf.* **21** (1979) 39.
1134. G. Collier, J. D. Voelkers, D. E. Griffin, *Ammonia Plant Saf.* **22** (1980) 206.
1135. F. Yazhari, *Hydrocarbon Process.* **61** (1982) no. 5, 187.
1136. H. Uchida, M. Kuraishi, *Bull. Chem. Soc. Jpn.* **23** (1955) 106.
1137. H. Hinrichs *et al.*, *DECHEMA Monogr.* **68** (1971) 493 – 500.
1138. K. Isermann in A. I. More (ed.), *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 571.
1139. Nitrex Statistik 1980, Nitrex AG, Zürich.
1140. A. Quispel: *The Biology of Nitrogen Fixation*, Elsevier/North Holland Publ. Comp., Amsterdam – Oxford – New York 1974.
1141. Emerging Technologies no. 1: Nitrogen Fixation. An Analysis of Present Research and Implications for the Future. Technical Insights Inc., Fort Lee, New Jersey.
1142. R. W. F. Hardy, U. D. Havelka, *Science (Washington, D.C.)* 1975, no. 4188, 633.
1143. J. R. Postgate: "Fertilizer '83," *Br. Sulfur Corp. 7th Int. Conf.*, London 1983, Proceedings of the Conf., p. 119.
1144. *Chem. Eng. News* **55** (1977) no. 40, 19. Links
1145. *Nitrogen* **111** (1978) 41.
1146. G. N. Schrauzer, T. D. Guth, *J. Am. Chem. Soc.* **99** (1977) 7189. Links
1147. *Nitrogen* **144** (1983) 32.
1148. *The Energetics of Biological Nitrogen Fixation*, Workshop Summaries I, Plant, Physiology, Am. Soc. of Plant Physiologists, Rockville, Maryland, USA.
1149. D. F. Shanmugam, F. O'Gara, D. Andersen, R. C. Valentine: "Biological Nitrogen Fixation," *Am. Rev. Plant Physiol.* **29** (1978) 263.
1150. H. J. Rehm: "Beiträge der modernen Biotechnologie und Biochemie zur Ernährung," *VCI-Schriftenreihe Chemie + Fortschritt* 1983, no. 3, 18.
1151. *Nitrogen* **146** (1983) 24.
1152. M. Tamaguchi: "Biological Nitrogen Fixation in Flooded Rice Fields," in Nitrogen and Rice, International Rice Research Institute, Manila, Philippines, 1979, p. 193 – 204.
1153. K. Gopalakrishna Pillai, D. B. B. Chaudary, K. Krishnamurty: "Bio-Fertilizers in Rice Culture – Problems and Prospects for Large Scale Adaptation," *Fert. News* **25** (1980) 40 – 45.
1154. W. v. d. Heuvel, G. v. d. Lindenbergh, *Ingenieursblad* **43** (1974) no. 18, 540 – 546.
1155. C. C. Hale, *Nitrogen* **119** (1979) 30 – 36.
1156. C. C. Hale, *Nitrogen* **125** (1980) 5.
1157. C. C. Hale, *Nitrogen 88: 12th Brit. Sulphur Int. Conf.*, Geneva 1988.
1158. I. K. Suri, R. K. Bohla, *Fertilizer News* (1986) (May), 52
1159. T. Huberich, *Plant Oper. Prog.* **1** (1982) no. 2, 117 – 122.
1160. C. C. Hale, *Ammonia Plant Saf.* **16** (1974) 23; **21** (1979) 61; **24** (1984) 181. C. C. Hale, *Nitrogen* **150** (1984) 27.
1161. Code of Practice for the Large Scale Storage of Fully Refrigerated Anhydrous Ammonia in the UK, Chemical Industries Association, London 1975.
1162. H. Arup, *Ammonia Plant Saf.* **19** (1977) 73.
1163. J. J. Aarts, D. M. Morrison, *Ammonia Plant Saf.* **23** (1981) 124.
1164. N. A. Hendricks, *Ammonia Plant Saf.* **21** (1979) 69.
1165. K. Feind, *Ammonia Plant Saf.* **20** (1978) 46.
1166. V. V. Kharlamov, Yu. M. Tsymbal, *Zh. Vses. Khim. Ova.* **28** (1983) no. 4, 433 – 438.
1167. G. Schlichthärle, T. Huberich, *Plant Oper. Prog.* **2** (1983) , no. 3, 165 – 167.
1168. A. A. Arseneaux, 29th Ammonia Safety Symp., San Francisco 1984.
1169. W. R. Southard et al., Interstate Commerce Comission, U.S. Govt., Informal Study of Freight Rate Structure Fertilizer and Fertilizer Products, vol. II.
1170. G. V. Rohleder, 13th Ammonia Safety Symp., Montreal 1968.
1171. W. A. Inkofer, 13th Ammonia Safety Symp., Montreal 1968.
1172. W. A. Inkofer, G. M. Wilson, J. E. Adams, 15th Ammonia Safety Symp., Denver 1970.
1173. T. P. Hignett, *Transportation and Storage of Ammonia*, Fertilizer Industry Round Table, Washington 1979.
1174. *Best Available Techniques for Pollution Prevention and Control in the European Fertilizer Industry*,Booklet 1: Production of Ammonia, European Fertilizer Manufacturers Assoc., 1995
1175. *Environmental Quality of Life*, Technical Note on the best Available Technology not Entailing Excessive Costs for Ammonia Production, EC Report EUR 13002 EN, 1990.
1176. J. M. Blanken, 24th Ammonia Safety Symp., San Francisco 1979. *Chem. Eng. Progr.* **76** (1980) 89 – 104.
1177. R. F. Griffiths, G. D. Kaiser, *J. Hazard Mater* **6** (1982) 197. Links
1178. C. Charp, Agricult. Anhydrous Ammonia Technol. Use Proc. Symp., San Louis 1966, p. 21.
1179. G. F. P. Harris, P. E. MacDermott, *Inst. Chem. Eng. Symp. Ser.* **49** (1977) 29.
1180. W. L. Ball, 14th AIChE Ammonia Safety Symp., Portland 1969.
1181. R. F. Schrader, *Nitrogen* **117** (1979) 26.
1182. A. Resplandy, *Chim. Ind. Genie Chim.* **102** (1969) 691.
1183. H. C. Goldwire, *Chem. Eng. Progr.* **82** (1986) 35.
1184. P. K. Raj, J. H. Hagopian, A. S. Kalekar, 19th AIChE Ammonia Safety Symp., Salt Lake City 1969.
1185. J. C. Statharas *et al.*, *Process Safety Progr.* **10** (1993) 118.
1186. K. P. Raj, R. C. Reid, *Environ. Sci. Technol.* **12** (1978) 1422. Links
1187. National Research Council Panel on Response to Causalities Involving Shipborne Hazards, AD-A075203, 1979, p. 1.
1188. A. Nielsen: *Ammonia — Catalysis and Manufacture*, Springer, New York 1995.
1189. *Nitrogen* **140** (1982) 20.
1190. M. Y. Nuttonson, PB–209478, 1 (1972)
1191. National Inst. for Occupational Safety and Health, Rockville 1974, PB-246669, p. 1.
1192. National Research Council, Committee on Toxicology, Washington, PB-244336, p. 1.
1193. L. Legters, (1980), AD-A094501 p. 1
1194. J. E. Lessenger, *Plant/Oper. Progr.* **4** (1985) 20.
1195. J. C. Barber, 22nd AIChE Ammonia Safety Symp., Denver 1977.
1196. J. E. Ryer-Powder, 35rd AIChE Ammonia Safety Symposium, San Diego 1990.
1197. T. F. Kohlmeyer, *Gas Wärme Int.* **19** (1970) no. 1, 15 – 20.
1198. R. Leschbaer, H. Schumann, *DECHEMA-Monogr.* **80** (1976) 1639 – 1669, part 2, 747 – 748.
1199. G. Kurz, *Proc. Eng.* (1976) Oct., 97.
1200. W. J. De Coursey *et al.*, *Can. J. Chem. Eng.* **40** (1962) 203. Links
1201. M. N. Park, *Nitrogen 88: Brit. Sulphur International Conference*, Geneva Mar 1988.
1202. "World Ammonia Production, Consumption and Trade", Statistical supplement, *Nitrogen* **217** (1995) 22.
1203. "Five Year Outlook for Ammonia 1995 – 2000", British Sulphur.
1204. M. Appl, *Ammonia*, Methanol, Hydrogen, Carbon monoxide — Modern Production Technologies CRU Publishing Ltd, London 1997 63 – 65.

1205. “BP Statistical Review of the World's Energy”, British Petroleum Co., June 1995.
1206. E. Banik, *Explosivstoffe* **5** (1957) no. 2, 29.
1207. F. J. Wollner, *Ammonia Casale's 75th Anniversary Symposium*, Lugano, Switzerland, Nov. 1996.
1208. C. C. Hale, W. H. Lichtenberg, *Ammonia Plant Saf.* **22** (1980) 35.
1209. G. J. Leigh, in *Catalytic Ammonia Synthesis*, ed. by J. R. Jennings, 1991, 365 – 387.
1210. J. S. Pathe, C. A. Atkins, R. M. Rumbird, in A. H. Gibson, W. E. Newton (eds.): *Current Perspectives in Nitrogen Fixation*, Australian Academy of Science, Canberra 1981.
1211. V. P. Gutschick, Long term strategies for supplying nitrogen to crops; Informal report LA-6700–19S, Los Alamos Scientific Laboratory, Los Alamos, NM, USA, 1977.
1212. R. F. Michin, S. J. Pate, *J. Exp. Bot.* **24** (1973) 259.
1213. M. J. Merrik, *J. R. Agric. Soc. Engl.* **147** (1986) 202.
1214. P. J. Baldock, *Chem. Eng. Prog. Loss Prev.* **13** (1980) 35 – 42.
1215. TA, Technische Anleitung zur Reinhaltung der Luft vom Dezember 1983.
1216. D. Eckhold *et al.*, *Chem. Tech. (Leipzig)* **24** (1972) no. 2, 92.
1217. L. V. Caserta, *Chem. Eng. Prog.* **68** (1972) no. 5, 41.
1218. T. Dear, *Ammonia Plant Saf.* **16** (1974) 108.
1219. R. Werchan, *Ammonia Plant Saf.* **18** (1976) 87.
1220. J. G. Seebold, *Hydrocarbon Process.* **51** (1972) no. 3, 97.
1221. E. E. Allen, *Gas World* **172** (1970) no. 4501, 431.
1222. H. v. Bell, *Chem. Eng. Prog.* **78** (1982) no. 2, 74 – 77.
1223. K. A. Vick, J. B. Witthaus *et al.*, *Chem. Eng. Prog. Techn. Man.* **22** (1980) 54 – 62.
1224. B. H. Winegar, *Chem. Eng. Prog. Tech. Man.* **22** (1980) 226 – 230.
1225. F. J. Heller, *Chem. Eng. Prog. Tech. Man.* **23** (1981) 132 – 145.
1226. R. J. Eiber, *Chem. Eng. Prog. Tech. Man.* **23** (1981) 146 – 156.
1227. J. J. O'Driscol, *Ammonia Plant Saf.* **17** (1975) 119 – 122.
1228. R. F. Griffiths, G. D. Kaiser, *J. Hazard. Mater.* **6** (1982) no. 1/2, 197 – 212. Links
1229. P. P. K. Raj, R. C. Reid, *Environ. Sci. Technol.* **12** (1978) no. 13, 1422 – 1425. Links
1230. L. J. Buividas in A. I. More (ed.): *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981* 1982, 601.
1231. *Nitrogen* **147** (1984) 3.
1232. *Nitrogen* **140** (1982) 5.
1233. T. Wakabayashi, *Ammonia Plant Saf.* **20** (1978) 17.
1234. *Erdöl Kohle Erdgas Petroch.* **37** (1984) no. 4, 143. *Chem. Ing. Tech.* **56** (1984) no. 8, A 407.
1235. H. J. Bomelburg, *Plant Oper. Prog.* **1** (1982) no. 3, 175.
1236. *Chem. Ing. Tech.* **56** (1984) no. 4, A 154.
1237. J. M. Shah, *Plant Oper. Prog.* **1** (1982) no. 2, 90.
1238. Yara Fertilizer Industry Handbook, 31. May 2005, www.yara.com
1239. E. T. Comeau, M. L. Weber, 21st AIChE Ammonia Safety Symp., Atlantic City 1976.
1240. K. A. Wick, J. B. Withaus, H. C. Mayo, 21st AIChE Ammonia Safety Symp., Atlantic City 1976.
1241. O. A. Martinez, S. Madhavan, D. J. Kellett, 27th AIChE Ammonia Safety Symp., Boston 1986; *Plant Oper. Prog.* **6** (1967) 129
1242. J. R. Thomson, 34th AIChE Ammonia Safety Symp., San Francisco 1989.
1243. J. R. Thomson, R. N. Carnegie, 33rd AIChE Ammonia Safety Symp., Denver 1988.
1244. C. C. Hale, *Plant Oper. Prog.* **1** (1982) no. 2, 107.
1245. C. C. Hale, J. A. Josefson, D. E. Mattick, 29th AIChE Ammonia Safety Symp., San Francisco 1984.
1246. M. L. Greiner, *Plant Oper. Prog.* **3** (1984) no. 2, 66; *Ammonia Plant Saf.* **24** (1984) 109.
1247. P. N. Arunachalam, *Fertilizer News* **57** (1986).
1248. I. Dayasagan, *Fertilizer News* **60** (1986).
1249. P. P. Briggs, J. M. Richards, III, E. G. Fiesinger, 30th AIChE Ammonia Safety Symp., Seattle 1985.
1250. F. Prasek, 32nd AIChE Ammonia Safety Symp., Minneapolis 1987.
1251. R. C. A. Wiltzen, 35th AIChE Ammonia Safety Symp., San Diego 1990.
1252. R. H. Squire, 35th AIChE Ammonia Safety Symp., San Diego 1990.
1253. S. B. Ali, R. E. Smallwood, 35th AIChE Ammonia Safety Symp., San Diego 1990.
1254. J. R. LeBlanc, *Hydrocarbon Process.* **63** (1984) no. 7, 69.
1255. *Nitrogen* **189** (1989) 25.
1256. J. R. LeBlanc, *Chem. Econ. Eng. Rev.* **18** (1986) no. 5, 22.
1257. *Fertilizer Focus* **4** (1987) no. 10, 40.
1258. J. R. LeBlanc, *Energy Prog.* **5** (1985) no. 1, 4.
1259. M. Appl: *Ammonia, Methanol, Hydrogen, Carbon Monoxide — Modern production Technologies* CRU Publ., London 1997, p. 44, 45.
1260. W. F. van Weenen, J. Tielrooy, *32nd Ann. Meet., The Fertilizer Industry Round Table*, USA 1982, p. 268.
1261. K. C. Wilson *et al.*, *Nitrogen* **151** (1984) 31.
1262. K. C. Wilson, B. J. Grotz, J. Richez, Nitrogen 86, Brit. Sulphur Conf., Amsterdam 1986.
1263. B. J. Grotz, *Nitrogen* **217** (1995) 41 – 48.
1264. K. G. Christensen, B. J. Grotz, K. G. Gosnell, *Nitrogen* **181** (1989) 31 – 36.
1265. W. Glover, J. P. Yoars, *Hydrocarbon Process.* **52** (1973) no. 4, 165.
1266. W. Glover, J. P. Yoars, 17th AIChE Ammonia Safety Symp., Minneapolis 1972.
1267. K. G. Christensen, B. J. Grotz, K. G. Gosnell, Fertilizer Latin America Int. Conf., Caracas 1989; Brit. Sulphur Corp. Proc., vol. 1.
1268. B. J. Grotz, L. Grisolia, *Nitrogen* **199** (1992) 39.
1269. Humphreys & Glasgow Ltd., GB-A 2126208 A, 1983 (C. L. Winter).
1270. F. C. Brown, C. Topham, Proc. Europ. Conf. on Energy Efficient Prod. of Fertilizers, Bristol 1990, p. 1.
1271. M. W. Kellogg, EP-A 106076, 1983 (C. van Dijk *et al.*)
1272. D. A. Waitzman et al., Fall Annual Meeting of the AIChE, Washington, D.C., Oct. 30, 1983.
1273. T. Matsunami, paper presented to the Internat. Fertilizer Industry Assoc., Johannesburg, South Africa, March 1983.
1274. T. Sueyama, T. Tsujino, Fertilizer 85: Brit. Sulphur Int. Conf., London 1985.
1275. *Ubes Texaco Process Coal Gasification Plant*, Ube comp. brochure 1984.
1276. ICI, EP-A 93502, 1983 (A. Pinto).
1277. J. R. LeBlanc, AIChE Symposium on Safety in Ammonia Plants, San Francisco, Calif., 1984.
1278. J. E. Lessenger, *Plant Oper. Prog.* **4** (1985) no. 1, 20.
1279. I. Dybkjaer, E. A. Gam, AIChE Symposium on Safety in Ammonia Plants, San Francisco, Calif., 1984.
1280. R. K. Bartoo, *Chem. Eng. Prog.* **80** (1984) no. 10, 35

1281. Humphreys & Glasgow Ltd., GB-A 2126573 A, 1983 (F. C. Brown).
1282. J. G. Ommen *et al.*, *J. Catal.* **38** (1975) 120.
1283. M. Coplin, *Lancet* **241** (1949) 95.
1284. W. M. Grant, *Arch. Ophthal.* **44** (1950) 399.
1285. T. Sollmann, *A Manual of Pharmacology*, 6th ed., Saunders, Philadelphia 1944.
1286. R. C. Wands, in G. D. Clayton, F. E. Clayton: *Patty's Industrial Hygiene and Toxicology*, **vol. 2 B**, Wiley Interscience, New York 1981, p. 3045 – 3052.
1287. K. B. Lehmann, *Arch. Hyg.* **17** (1893) 329.
1288. I. M. Alpator, *Gig. Tr. Prof. Zabol.* **2** (1964) 14. Links
1289. D. P. Stombaugh, H. S. Teague, W. L. Roller, *I. Anim. Sci.* **6** (1969) 844.
1290. H. F. Smith, *Amer. Industr. Hyg. Quart.* **17** (1956) 145.
1291. C. P. Carpenter, H. F. Smith, Jr., U. C. Pozzani, *J. Industr. Hyg.* **31** (1949) 343.
1292. V. I. Mikhailow, *Probl. Kosmich. Biol. Akad. Nauk SSSR* **4** (1965) 531.
1293. The M. W. Kellogg ammonia converter retrofit, Kellogg company brochure HG/2.5M/3–88 (1988)
1294. R. G. Howerton, S. A. Noe, *33rd AIChE Ammonia Safety Symposium*, Denver, Oct 1988.
1295. E. M. Boyd, M. L. McLachlin, W. F. Perry, *J. Industr. Hyg.* **26** (1944) 29.
1296. S. D. Silver, F. P. McGrath, *I. Industr. Hyg.* **30** (1948) 7.
1297. K. B. Lehmann, *Arch. Hyg.* **5** (1886) 68.
1298. I. H. Weatherby, *Proe. Soc. Exp. Biol.* **81** (1952) 300.
1299. R. A. Coon, R. A. Jones, L. J. Jenkins, I. Siegel, *Toxicol. Appl. Pharmacol* **16** (1970) 646. Links
1300. P. Trendelenburg in A. Heffter: *Handbuch der experimentellen Pharmakologie*, Vol. I, p. 470, Springer, Berlin 1923.
1301. D. C. Topping, W. J. Visek, *J. Nat.* **106** (1976) 1583.
1302. B. Toth, *Int. J. Cancer* **9** (1972) 109. Links
1303. Litten Bionetics, *NTIS, PB-245*, 506 (1975) Washington D.C.
1304. M. B. Jacobs: *The analytical Toxicology of Industrial Inorganic poisons*, **509** Interscience, Wiley N.Y. (1967).
1305. I. M. Alpator, *Prom. Toksikol. i. Klinika Prof. Zabol. Khim. Etiol.* **200** (1962) .
1306. L. Matt, *Med. Inaugural Diss*, Würzburg (1889).
1307. A. C. S. Fieldner, S. H. Katz, S. P. Kinney, *in Noxious Gases*, **125** 2nd rev. ed. Reinhold, N.Y. (1943).
1308. L. Silvermaur, L. Whittenberger, I. Muller, *I. Industr. Hyg.* **31** (1949) 74.
1309. K. S. Warren, S. Schenker, *J. Lab. Clin. Med.* **64** (1964) 442. Links
1310. W. D. Lotspeich, *Amer. J. Physiol.* **206** (1965) 1135.
1311. P. Vinay, E. Alignet, C. Pichette, M. Watford, G. Lemieux, A. Gougoux, *Kidney Intern.* **17** (1980) 312.
1312. D. Z. Levine, L. A. Nash, *Amer. J. Physiol.* **225** (1973) 380.
1313. K. S. Warren, *J. Clin. Invest.* **37** (1958) 497. Links
1314. J. Oliver, E. Bourke, *Clin. Sci. Molec. Med.* **48** (1975) 515.
1315. J. Yoshida, K. Nakame, R. Nakamura, *Nippon Chikusangakukaiko* **28** (1957) 185.
1316. G. Ertl et al., *Surface Science* **114** (1982) 527.
1317. J. Abughazaleh, J. Gosnell, D. P. Mann, R. Strait: 2005 KBR Indonesian Ammonia Seminar, 20-23 February 2005, Jakarta, Indonesia.
1318. M. Appl: *Ammonia, Principles and Industrial Practice*, WILEY-VCH, Weinheim-New York 1999.
1319. Syngas Technologies Supplement to Nitrogen & Methanol, 2005.
1320. *Nitrogen & Methanol* **270** (2004) 37.
1321. N. Ringer, *paper presented at Süd-Chemie Catalyst Symposium*, Manama, Bahrain, June 2004.
1322. N. Ringer, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1323. L. Huazhang *et al.*, US 5,846 507 (1998) Zhejiang University of Technology, China.
1324. J.-C. Park, D. Kim, C.-S. Lee, D.-K. Kim, *Bull. Korean Chem. Soc.* **20** (1999) no. 9, 1005–1009.
1325. S. Bencic: "*Ammonia synthesis promoted by iron catalysts*", Literature report, Michigan State University (April 2001).
1326. C. J. H. Jacobson, S. E. Nielsen, *paper presented at 46th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2001.
1327. J. Brightling, P. Farnell, C. Forster, F. Beyer, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1328. I. Dybkjaer, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1329. X. Shen, X. Xu, W. Feng, S. Singh, A. Malhotra, *Nitrogen & Methanol*, **274** (2005) 39.
1330. A. Malhotra, L. Hackemesser, *paper presented at 46th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Montreal 2002.
1331. P. V. Broadhurst, P. K. Ingram, *paper presented at Nitrogen 2005, International Conference and Exhibition*, Bucharest, Romania 27 February – 2 March 2005.
1332. M. P. Sukumaran, *Nitrogen & Methanol* **264** (2003) 19–28.
1333. *Nitrogen & Methanol* **275** (2005) 38–47.
1334. *Nitrogen & Methanol* **264** (2003) 12–18.
1335. *Nitrogen & Methanol* **264** (2003) 37–45.
1336. E. West-Toolsee, E. Filippi, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1337. Plant and process enineering equipment – Components and plant circuitry, Steimüller company brochure P 8604-06-0510 (1989).
1338. K. Nassauer: "*Process gas waste heat boilers*", Borsig AG, company brochure 290151 AC.
1339. Waste heat recovery for reformed gas and synthesis gas cooling in modern ammonia plants, Borsig AG company brochure.
1340. Process Gas Waste Heat recovery, Borsig AG company brochure.
1341. *Nitrogen & Methanol* **268** (2004), 38 (references).
1342. Ammonia – Energy Integration in Ammonia Plants, Uhde Engineering News 2-91, Hi 111 9 1000 91, 1991.
1343. Ammonia, Uhde company brochure F 110e/2000 09/2004 DÖ/Hi (09.08.2004).
1344. Balcke-Dürr, Company Leaflet 31.07.1990.
1345. Synloop Waste Heat Boiler in Ammonia Plants, Borsig AG company brochure.
1346. M. Podhorski: "Hydraulic expansion of tubes", *paper presented at International Conference on Expanded and Rolled Joint Technology*, Toronto 1993.
1347. Synloop Waste Heat Boiler in Ammonia Plants-Unique Hot/Cold Tube Sheet design, Babcock-Borsig AG company brochure.
1348. Synthesis Gas Cooler Downstream of Partial Oxidation of Oil or Natural Gas, Borsig AG company brochure.
1349. Shell POX Waste heat boilers, Standard Fasel-Lentjes, Company brochure 1996.
1350. *Ammonia Plants and Related Facilities Symposia – Proceedings from the last 50 Years*, 3 CD-ROMs, American Institute of Chemical Engineers 2005, ISBN 0-8169-0995-4.
1351. *Nitrogen & Methanol*, **258** (2002) 47–48.
1352. J. Larsen, D. Lippman, C. W. Hooper, *Nitrogen & Methanol*, **253** (2001) 41–46.
1353. D. Lippmann: "Large-scale ammonia plant technology", *paper presented at 4th Uhde Fertilizer Symposium*, Dortmund June 2002.
1354. J. Rüther, J. Larsen, D. Lippman, D. Claes, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1355. J. S. Larsen, D. Lippman, *paper presented at 46th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego, 2002.
1356. Carbon monoxide production technologies, KTI Newsletter (Winter 1987).
1357. H. Jungfer: "Synthesegas aus Raffinerierückständen, Konzeption und Betriebsergebnisse von Lindeanlagen zur Partiellen Oxidation", *Linde-Bericht aus*

Technik und Wissenschaft **57** (1985) 15–20 (ISBN 0024-3728).
1358. Lurgi References – Gas and Synthesis Gas Technology, Lurgi brochure 1569e/9.93/4.10.
1359. C. D. Swann, M. L Preston, *Loss Prev. Proc. Ind.* **8** (1995) 6, 349–353.
1360. R. D. Turney, *TransChem.* **68** B (1990) 12–16.
1361. R. C. McConnel, *paper presented at 36th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Los Angeles 1991, *Ammonia Plant Saf.* **32** (1992) 104.
1362. A. L. Ormond, *J. Chem. A. Loss Prev. Bulletin* **126** (1995).
1363. IFA – Production and International Trade – July 2005.
1364. J. Gosnell, *paper presented at Hydrogen Conference, Argonne National Laboratory*, 13 October 2005.
1365. *Yara Fertilizer Handbook*, 31 May 2005.
1366. D. A. Kramer: "*Nitrogen*", in Mineral Commodity Profiles, US Department of the Interior, US Geological Survey, Open file report 2004-1290.
1367. S. R. Wilson: Agriculture as a Producer & Consumer of Energy, Farm Foundation & USDA Office of Energy Policy, CF Industries, June 24-25 2004.
1368. BP Global-Reports and publications – Statistical Review of World Energy 2005.
1369. M. J. McGrath, E. J. Houde, *paper presented at AIChE Spring 1999 Meeting*, 14-18 March 1999 Houston.
1370. C. E. Jaske, *paper presented at 47th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego 2002.
1371. V. Pattabathula, *FINDS*, A Stokes Engineering Publication, Volume XX, Number 3, Third Quarter 2005.
1372. More Ammonia from US Coke and Coal, *FINDS*, A Stokes Engineering Publication Volume XX, Number 3, Third Quarter 2005.
1373. Ch. J. Muhlenforth, *FINDS*, A Stokes Engineering Publication Volume XVII, Number 4, Fourth Quarter 2002.
1374. R. Strait, *FINDS*, A Stokes Engineering Publication, Volume XVII, Number 1, First Quarter 2002.
1375. K. Blanchard, St. Noe, E. Plaxco, *FINDS*, A Stokes Engineering Publication, Volume XV Number 4, Fourth Quarter 2000.
1376. Ch. J. Muhlenforth, *FINDS*, A Stokes Engineering Publication, Volume XV, Number 3, Third Quarter 2000.
1377. "A balancing act", *Nitrogen & Methanol* **273** (2005) 39–44.
1378. "New ammonia WHB internals raise capacity, enhance reliability", *Nitrogen & Methanol* **266** (2003) 41–47.
1379. J. Gosnell, *paper presented at Nitrogen 2000, Vienna (March 2000), International Conference & Exhibition*, British Sulphur Publishing.
1380. J. Gosnell, *paper presented at Gulf Nitrogen 2002, Doha (March 2002), International Conference & Exhibition*, British Sulphur Publishing.
1381. "Mega-ammonia round-up", *Nitrogen & Methanol*, **258** (2002) 39–48.
1382. J. Abughazaleh, J. Gosnell, R. Strait, *paper presented at 47th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego 2002.
1383. W. E. Davey, T. Wurzel, E. Filippi, *Nitrogen & Methanol* **262** (2003) 41–47.
1384. S. E. Nielsen, *paper presented at 46th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Montreal 2002.
1385. J. Larsen, D. Lippmann, C. W. Hooper, *Nitrogen & Methanol* **253** (2001) 41–46.
1386. W. E. Davey, T. Wurzel, E. Filippi, *paper presented at Nitrogen 2003, Warsaw (February 2003), International Conference & Exhibition*, British Sulphur Publishing.
1387. "Coal as feedstock" *Nitrogen 2003, Warsaw (February 2003), International Conference & Exhibition*, British Sulphur Publishing.
1388. "GHR – out of the wood yet" *Nitrogen & Methanol* **266** (2003) 33–40.
1389. "Is there any real competition for iron", *Nitrogen & Methanol* **257** (2002) 34–39.
1390. J. Brightling, *Nitrogen & Methanol* **256** (2002) 29–39.
1391. T. Evans, *Nitrogen & Methanol* **232** (1998) 41–52.
1392. R. Strait, *Nitrogen & Methanol* **238** (1999) 37–43.
1393. R. Strait, *paper presented at Nitrogen '99, Caracas, Venezuela (March 1999) International Conference & Exhibition*, British Sulphur Publishing.
1394. "Maximising output from a single train", *Nitrogen & Methanol* **238** (1999) 44–48.
1395. "The chemical bearings of the ammonia process" *Nitrogen & Methanol* **244** (2000) 31–39.
1396. "By no means a foregone conclusion", *Nitrogen & Methanol* **252** (2001) 33–51.
1397. V. Pattabathula, B. Rani, D. H. Timbres, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1398. A. Malhotra, P. Kramer, S. Singh, *paper presented at 48th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Orlando 2003.
1399. D. Bourgois, M. Felscher, W. Moore, *paper presented at 48th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Orlando 2003.
1400. G. P. Williams: "Safety Performance in Ammonia, Methanol and Urea Plants 1999-2001 Surveys", *paper presented at 48th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Orlando 2003.
1401. G. P. Williams: "*World-wide Ammonia Benchmarking Study 2000-2001: Multi-client Benchmarking Report*" Plant Survey International, Inc 10813 Bland Ridge Court, Petersburg, VA 23805, USA.
1402. M. Dressel, M. Heinke, U. Steinhoff, *paper presented at 35th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego 1990.
1403. S. E. Nielsen, P. Vaug, *paper presented at 49th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Denver 2004.
1404. D. Hooper, *paper presented at 47th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego 2002.
1405. G. R. Precott, *paper presented at 47th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, San Diego 2002.
1406. J. B. Sievert, *paper presented at 46th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Montreal 2002.
1407. G. P. Williams: "Safety Performance in Ammonia Plants, 1997 – 1998 Surveys", *Plant Survey International, Inc., paper presented at 45th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Tuscon 2000.
1408. G. R. Prescott, B. Shannon, *paper presented at 45th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Tuscon 2000.
1409. G. P. Williams, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1410. N. Ringer, M. Michel, R. Stockwell, *paper presented at 50th AIChE Symposium on Safety in Ammonia Plants and Related Facilities*, Toronto 2005.
1411. M. Appl, "Is Coal the Future Feedstock for Ammonia", 2nd Casale Symposium for Customers and Licensees, Lugano 30 May – 2 June 2006.
1412. *Nitrogen+Syngas*, **280** (2006) 13–15.
1413. I. Dybkjaer, *Fuel Processing Technology*, **42** (1995) 85–107. Links
1414. S. Fritsch "Synthesis gas Processing Comparison of four synthesis gas routes", *Krupp-Unde fertilizer symposium*, Dortmund, June 11–13.
1415. E. Filippi, F. Di Muzio, E. Rizzi. 2nd Casale Symposium for Customers and Licensees, Lugano 30 May – 2 June 2006.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Ammonium Compounds

Ammonium Bromide → Bromine Compounds; Ammonium Fluoride → Fluorine Compounds, Inorganic; Ammonium Iodide → Iodine and Iodine Compounds; Ammonium Phosphates → Phosphate Fertilizers

Karl-Heinz Zapp, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 and 2)

Karl-Heinz Wostbrock, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 3 and 4)

Manfred Schäfer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 3)

Kimihiko Sato, Asahi Glass Company Ltd., Yokohama, Japan (Sections 3.2.1 and 3.5 in part)

Herbert Seiter, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Werner Zwick, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Ruthild Creutziger, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 5)

Herbert Leiter, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 5)

1. Ammonium Nitrate 1
1.1. Physical and Chemical Properties . 2
1.2. Production 3
1.2.1. From Ammonia and Nitric Acid . . . 3
1.2.2. Conversion of Calcium Nitrate Tetrahydrate 6
1.3. Granulation 7
1.4. Grain Stabilization and Surface Treatment (Conditioning) 8
1.5. Transport Regulations 9
1.6. Safety 9
1.7. Uses 10
1.8. Economic Aspects 11
2. Ammonium Sulfate 11
2.1. Properties 11
2.2. Production 12
2.2.1. From Coke-Oven Gas 12
2.2.2. From Ammonia and Sulfuric Acid . . 12
2.2.3. Coproduct in Organic Syntheses . . . 13
2.2.4. From Gypsum 13
2.2.5. Other Processes 13
2.3. Granulation and Storage 14
2.4. Uses 14
2.5. Economic Aspects 14
3. Ammonium Chloride 14
3.1. Properties 15
3.2. Production 16
3.2.1. Modified Solvay (Ammonia – Soda Ash) Process 16
3.2.2. Direct Reaction Between HCl and NH_3 19
3.2.3. Reaction of Reciprocal Pairs of Salts . 19
3.3. Corrosion 19
3.4. Packaging, Storage, Shipping, and Handling 20
3.5. Uses 20
4. Ammonium Carbonates 20
4.1. Ammonium Hydrogencarbonate . . 21
4.2. Ammonium Carbamate 21
4.3. Ammonium Carbonate 22
4.4. Storage, Transport, and Handling . 23
4.5. Uses 23
5. Toxicology and Occupational Health 23
6. References 24

1. Ammonium Nitrate

Ammonium nitrate [*6484-52-2*], NH_4NO_3, is a major chemical product. It is produced most frequently by neutralization of nitric acid with ammonia and is mainly processed into high-quality fertilizers. As a straight fertilizer, it accounts for 24 % of world consumption of nitrogen fertilizers and, in addition, is present in many blended and complex fertilizers, thus making an important contribution to feeding the world's population. Ammonium nitrate is also used as an oxidizing agent and a constituent of many explosives.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_243

1.1. Physical and Chemical Properties

Ammonium nitrate is a colorless salt, M_r 80.05, d_4^{20} 1.725, specific heat capacity 1.70 J g^{-1} K^{-1} between 0 and 31 °C; *mp* 169.6 – 170 °C. The reaction forming ammonium nitrate from ammonia and nitric acid is highly exothermic:

$$NH_3\,(g) + HNO_3\,(l) \longrightarrow NH_4NO_3\,(s) \quad \Delta H = -146\,\text{kJ/mol}$$

The salt has five modifications:

I	169.6	to	125.2 °C	cubic
II	125.2	to	84.2 °C	tetragonal
III	84.2	to	32.3 °C	orthorhombic
IV	32.3	to	– 16.9 °C	orthorhombic
V	below		– 16.9 °C	tetragonal

The transition at 32.3 °C is of particular significance for storage of fertilizers containing ammonium nitrate: on passing through the transition repeatedly, the fertilizer granules lose strength and finally disintegrate because of the differing densities.

Ammonium nitrate dissolves readily in water; in addition, the salt is hygroscopic. Therefore, care must be taken to avoid moisture during transport and storage. When the salt dissolves in water, heat is absorbed; therefore, ammonium nitrate can be used in freezing mixtures, for example, in mixtures with sodium chloride and ice. The heat of solution in an almost infinite quantity of water is + 26.4 kJ/mol at 18 °C, the integral heat of solution to saturation is + 16.75 kJ/mol, and the heat of solution in saturated solution is + 15 kJ/mol. The physical properties of ammonium nitrate solutions important for concentration by evaporation on an industrial scale are shown in Tables 1, 2, and 3 and Figures 1 and 2 (see below, and next page).

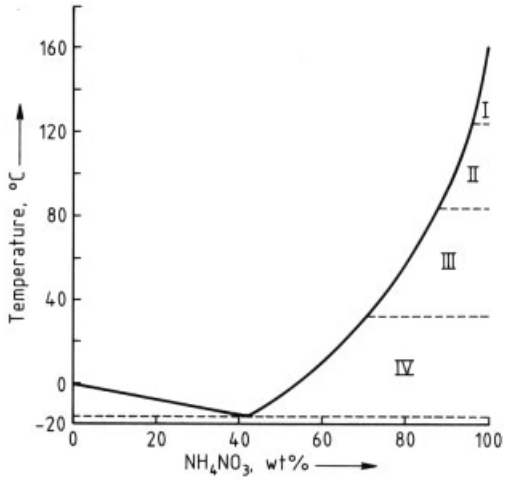

Figure 1. The system $NH_4NO_3 - H_2O$
Roman numerals denote the modifications

The solubility of ammonium nitrate in some nonaqueous solvents is considerable. The salt dissolves in liquid NH_3 and vigorously absorbs NH_3 to form solutions known as Divers liquid. Aqueous ammonium nitrate solutions 50 – 70 % by weight vigorously absorb NH_3 and may be used for stripping NH_3 from gases. Solutions such as these are also used for ammoniating superphosphate. The salt dissolves in methanol to give solutions of about 20 % at 30 °C and about 40 % at 60 °C. Its solubility in ethanol is about

Table 1. Densities of NH_4NO_3 solutions

t, °C / *c*, wt %	20 %	30 %	40 %	50 %	60 %	70 %	80 %	90 %	94 %	97 %	99 %
20	1.0830	1.1275	1.1750	1.2250	1.2785	–	–	–	–	–	–
40	1.0725	1.1160	1.1630	1.2130	1.2660	1.3220	–	–	–	–	–
60	1.0620	1.1045	1.1510	1.2005	1.2525	1.3090	1.3685	–	–	–	–
80	1.0550	1.0935	1.1390	1.1875	1.2395	1.2960	1.3555	–	–	–	–
100	1.0410	1.0820	1.1270	1.1745	1.2265	1.2825	1.3420	1.4075	–	–	–
120							1.3285	1.3930	1.4210	–	–
140								1.3785	1.4065	1.4285	–
160									1.3940	1.4165	1.4325
180										1.4060	1.4225
200											1.4121
220											1.4030

Table 2. Boiling points of aqueous ammonium nitrate solutions at atmospheric pressure

c, wt %	60	80	90	94	96	98	99
bp, °C	113.5	128.5	147	165	182	203	222

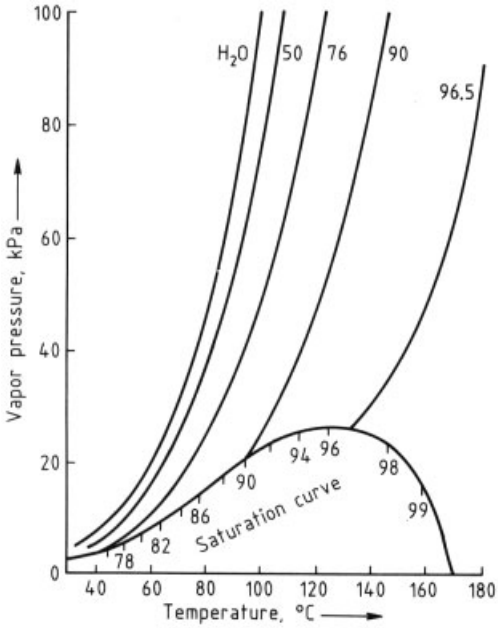

Figure 2. Vapor pressures of ammonium nitrate solutions
The saturation curve shows the vapor pressures of saturated solutions. The other curves show the vapor pressure of unsaturated solutions of definite composition. The salt contents are indicated in weight percent.

Table 3. Molar and specific heat capacities of ammonium nitrate solutions

NH_4NO_3,	Heat capacities,	
	$J\,mol^{-1}\,K^{-1}$	$J\,g^{-1}\,K^{-1}$
2.9	320.8	4.038
9.1	309.6	3.870
15.1	294.4	3.678
28.6	241.6	3.021
47.1	233.6	2.916
64	204.0	2.552

4 % at 20 °C and is minimal in acetone. Ammonium nitrate is insoluble in ether.

Ammonium nitrate is a powerful oxidizing agent. It is stable at normal temperature and pressure. On heating, it decomposes above 170 °C to gases. These reactions are promoted especially by small quantities of chlorine or free acid.

1.2. Production

Ammonium nitrate is produced mainly from ammonia and nitric acid. Ammonium nitrate is also formed in the production of nitrogen – phosphorus (NP) and nitrogen – phosphorus – potassium (NPK) fertilizers on decomposition of crude phosphate with nitric acid, the salt generally remaining in the fertilizers. In Europe the reaction of calcium nitrate with NH_3 and CO_2 is also used to produce NH_4NO_3.

1.2.1. From Ammonia and Nitric Acid

The neutralization of 45 – 65 % HNO_3 with gaseous NH_3 is accompanied by the release of 100 – 115 J per mole of NH_4NO_3 [6]. In most processes this considerable heat of reaction is used for partial or complete evaporation of the water. Depending on the pressure and the concentration of the nitric acid, 95 – 97 % solutions of ammonium nitrate can be obtained. During neutralization, the components must be mixed quickly and thoroughly in the reactor to avoid local overheating, losses of nitrogen, and decomposition of the ammonium nitrate.

One design of a neutralization reactor is shown in Figure 3 [7]. Ammonia and acid enter the inner tube, the actual neutralization zone, through pipes a and b. The neutralizer is filled to

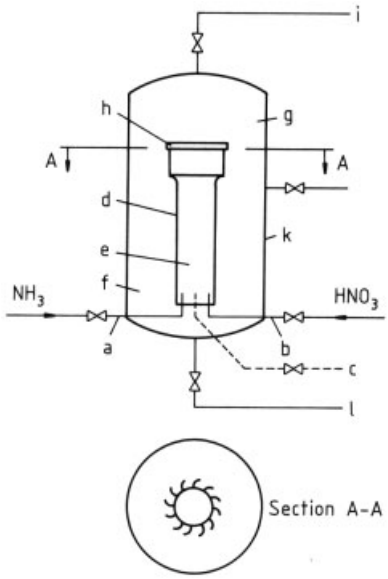

Figure 3. Neutralizer (Soc. Belge de l'Azote) [7]
A) Filling level; a) Ammonia feed pipe; b) Nitric acid feed pipe; c) Optional steam feed pipe; d) Inner tube; e) Inner compartment; f) Outer compartment; g) Steam chamber; h) Neutralizer top; i) Steam outlet; k) Container; l) Solution outlet

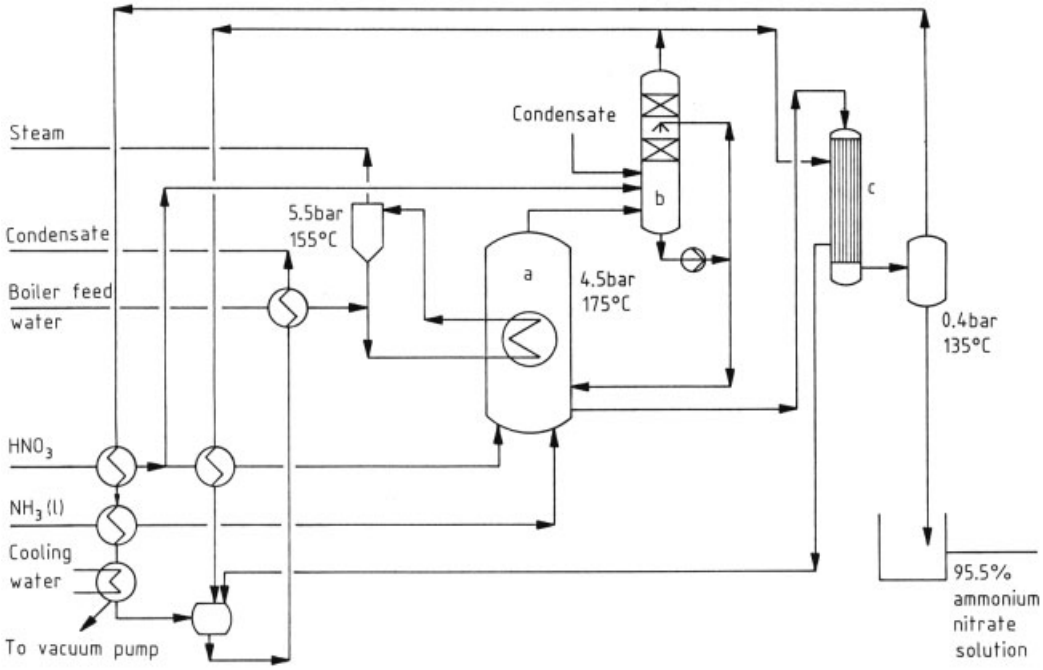

Figure 4. The UCB process
a) Reactor; b) Scrubber; c) Falling film evaporator

mark A with ammonium nitrate solution. While the solution boils vigorously in the outer compartment f, boiling is suppressed in the inner compartment e because of a slight excess pressure. The higher the temperature in the neutralizer (to 180 °C in pressure processes), the more important it is to maintain an exact pH value and to avoid introducing chlorides, heavy metals, and organic compounds. In addition, the quantity of ammonium nitrate at elevated temperature is kept as small as possible. The neutralizers and evaporators are generally constructed from Cr – Ni stainless steels.

A reduced-pressure process is operated by Chemico [6]. Normal-pressure installations use the Uhde process [8], the SBA (Soc. Belge de l'Azote) process [9], etc.; they and are characterized by low reaction temperatures and minimal corrosion. The heat can be best utilized if neutralization is carried out under pressure [10]. Steam can be generated in some processes. Operation under pressure goes back to FAUSER [11]. The pressure processes differ in removal of the heat of reaction.

In the *UCB process* (Fig. 4) [12], a heat exchanger in the pressure reactor uses a part of the heat of reaction to make steam.

Ammonia and 52 – 63 % HNO_3 are preheated and sprayed into the sump of the reactor. The pressure in the reactor is approximately 0.45 MPa (4.5 bar), the temperature is 170 – 180 °C, and the pH is 3 – 5. The pH is kept in this range by controlling the ratio of reactants. The heat exchanger cools the reaction mixture, and neutralization follows a stable course. The 75 – 80 % NH_4NO_3 solution leaving the reactor is concentrated to 95 % by evaporation in a falling film evaporator. The heat of reaction generates

1) process steam out of the evaporating water from the nitric acid. This is used in the process to preheat boiler feed water and nitric acid and to operate the falling film evaporator.
2) pure steam up to 0.55 MPa (5.5 bar) in the heat exchanger. This can be fed into the steam pool and used for other purposes.

The pH 3 – 5 reduces nitrogen losses into the process steam. The working conditions are adjusted

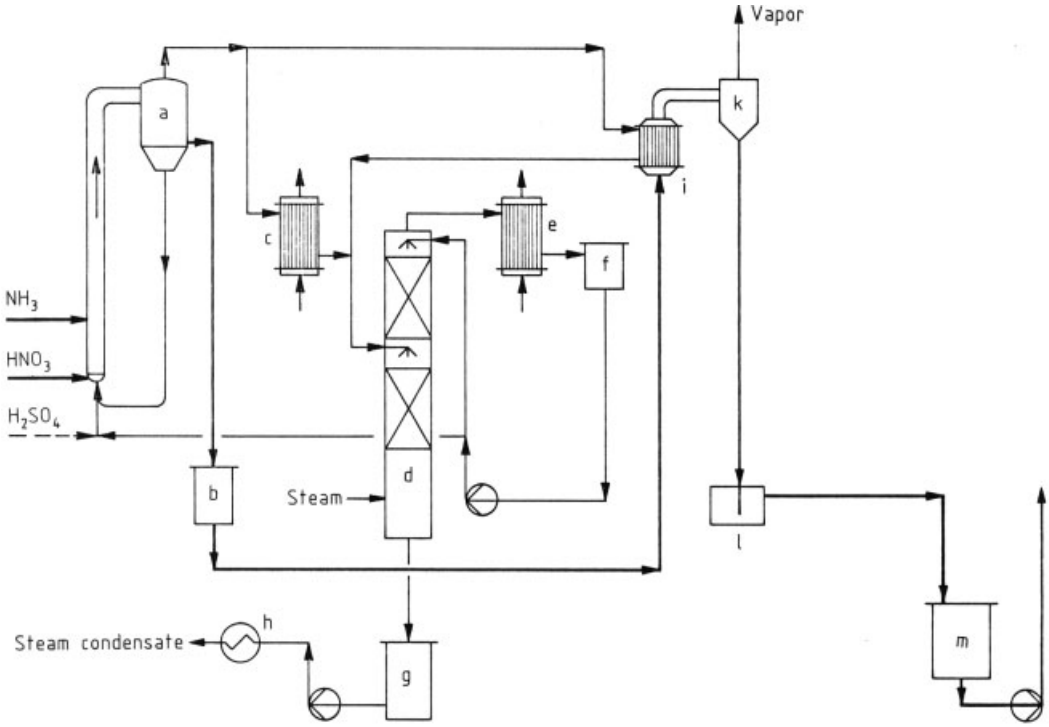

Figure 5. The Stamicarbon process
a) Neutralizer; b) Intermediate tank; c) Surplus steam condenser; d) Ammonia scrubber; e) Condenser; f) Dilute ammonia solution tank; g) Condensate tank; h) Cooler; i) Evaporator; k) Separator; l) Seal pot; m) Storage tank for 95 % ammonium nitrate solution

so that excess process steam does not accumulate.

Another process that works under pressure is the *Stamicarbon process* (Fig. 5) [13]. The neutralizer is a loop reactor that opens into a separator. The reaction solution is circulated without a pump by heat generated.

Nitric acid (60 wt%), preheated ammonia, and a small quantity of sulfuric acid are introduced at the lower end of the loop. The reactor operates at 0.4 MPa (4 bar) and 178 °C. The ammonium nitrate solution formed in the reactor has a concentration of 78 %. The steam removed at the top of the separator is passed through a mist eliminator and is mainly used to concentrate the NH_4NO_3 solution to 95 % in a vacuum evaporator. Excess steam is condensed, and ammonia is recovered from the condensate and returned to the reactor. In a second evaporator, the concentration can be increased to 98 – 99.5 % using fresh steam. The temperature of the ammonium nitrate solution is kept below 180 °C throughout neutralization and evaporation.

The *NSM/Norsk Hydro pressure process* (Fig. 6) [14] uses preheated ammonia and nitric acid.

The pressure in the reactor is between 0.4 and 0.5 MPa (about 4.5 bar), and the temperature ranges from 170 to 180 °C, conditions corresponding to a 70 – 80 % solution. Forced circulation and a thermal siphon effect circulate the solution through the reactor. Some of the heat of reaction is used to generate pure steam in an external boiler; some vaporizes water in the reactor, producing process steam, which is used to concentrate the ammonium nitrate solution to 95 %. The ammonia losses are kept small by washing the process steam with nitric acid, which is added to a circulating ammonium nitrate solution. Further concentration of the NH_4NO_3 solution, to 99.5 %, is carried out with steam in a special vacuum evaporator.

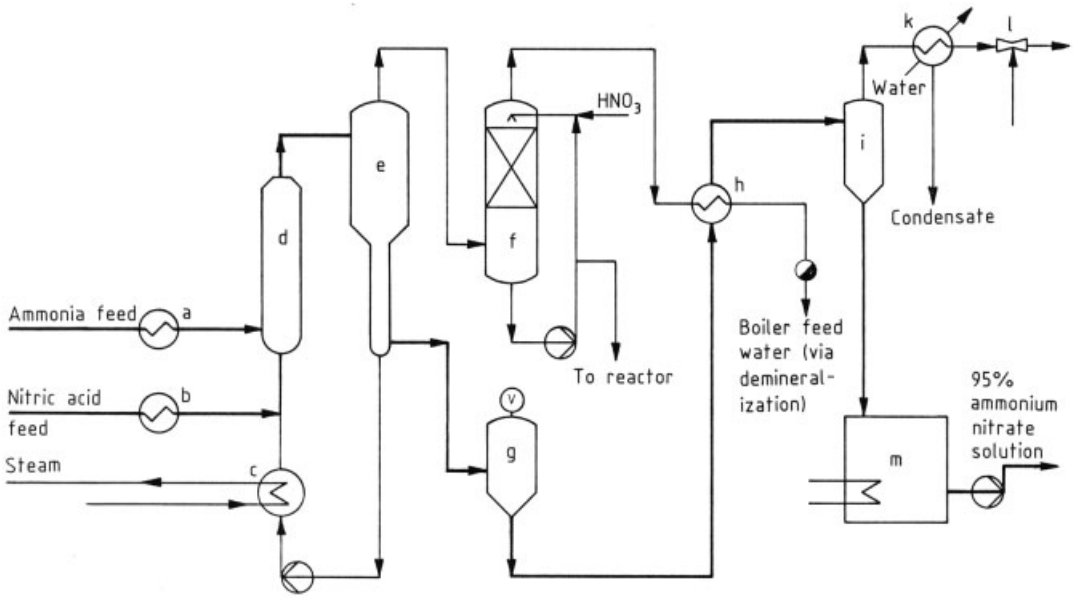

Figure 6. NSM/Norsk Hydro process
a) Ammonia evaporator/superheater; b) Nitric acid preheater; c) Boiler; d) Reactor; e) Reactor separator; f) Scrubber; g) Flash-tank; h) Evaporator; i) Separator; k) Condenser; l) Ejector; m) Tank

In the United States the *Stengel process* is significant [15], [16]. It produces anhydrous ammonium nitrate directly.

The ammonia and the ca. 58 % nitric acid are preheated with fresh steam and fed into a packed vertical tube reactor at 0.35 MPa (3.5 bar) and 240 °C. The mixture of NH_4NO_3 and steam is expanded into a vacuum in a centrifugal separator. After stripping with hot air, a 99.8 % NH_4NO_3 melt is discharged, onto a cooled steel belt, solidified, and then broken up or granulated. The steam is removed at the top.

In all these processes the maintenance of the desired pH is most important. Where the reaction temperatures are below 170 °C, the pH is kept between 2.4 and 4 to minimize nitrogen losses. In pressure neutralizers, a higher pH, 4.6 to 5.4, is necessary on account of the higher temperatures and greater danger of decomposition.

1.2.2. Conversion of Calcium Nitrate Tetrahydrate

The production of nitrophosphate fertilizers by digestion of crude phosphate with nitric acid in the Odda process produces as a byproduct calcium nitrate tetrahydrate, $Ca(NO_3)_2 \cdot 4\,H_2O$, in considerable quantities (→ Fertilizers). The output of nitrophosphates is increasing, while demand for calcium nitrate is declining. In a process introduced some years ago, the calcium nitrate tetrahydrate is treated with ammonia and carbon dioxide to form ammonium nitrate and calcium carbonate:

$$Ca(NO_3)_2 \cdot 4\,H_2O\,(s) + 2\,NH_3\,(g) + CO_2\,(g) \longrightarrow 2\,NH_4NO_3\,(aq) + CaCO_3\,(s) + 3\,H_2O$$

$$\Delta H = -126\,\text{kJ/mol}$$

The accompanying heat is adequate to evaporate all the water. However, a direct procedure is not possible because of the unfavorable equilibrium at elevated temperatures. In the *BASF process* (Fig. 7), heat removal is separated from the reaction of $Ca(NO_3)_2$ with $(NH_4)_2CO_3$: NH_3 and CO_2 are dissolved in a circulated NH_4NO_3 solution, and the heat given off is removed. The calcium nitrate tetrahydrate is also dissolved in an NH_4NO_3 solution. The two solutions are then reacted at ca. 50 °C; the heat produced is minimal.

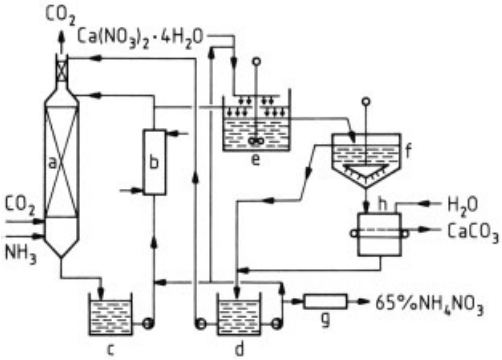

Figure 7. BASF process for conversion of calcium nitrate
a) Absorption tower; b) Condenser; c) NH_4NO_3–$(NH_4)_2CO_3$ solution; d) Cloudy NH_4NO_3 solution; e) Reaction vessel; f) Settling vessel; g) Classification filter; h) Belt filter

The grain size of the calcite precipitated can be influenced by the way the reactants are introduced. After reaction the approx. 65 % NH_4NO_3 solution is separated from $CaCO_3$ on a belt filter and concentrated by evaporation. The $CaCO_3$ still contains small amounts of ammonium compounds and phosphate and is mostly used for the production of calcium ammonium nitrate. If the calcium nitrate is suitably prepared before conversion, quite pure $CaCO_3$ may be produced.

For directly converting calcium nitrate and at the same time removing the heat of reaction, Hoechst developed a special vertical reactor [17]. The gaseous CO_2 is introduced at the bottom, and the ammonia is introduced in three zones, each cooled with water.

1.3. Granulation

Pure ammonium nitrate has a nitrogen content of 35 %. Commercial fertilizer ammonium nitrate is produced in several countries with an N content of 34.5 % or 33.5 %. It generally contains additives to stabilize the grain and improve storage properties.

Various processes have been developed for granulation [18], [19]. *Prilling* is preferred for large plants producing upwards of 1000 t per day, while the other processes are also suitable for lower outputs, around 250 t per day. Prilling starts from melts containing around 0.5 % water. Round or rectangular spray towers up to 70 m tall are used. The towers are concrete, steel, aluminum, etc. Steel frameworks and concrete must be coated with protective paints. Figure 8 [20] shows a diagram of an ammonium nitrate prilling plant.

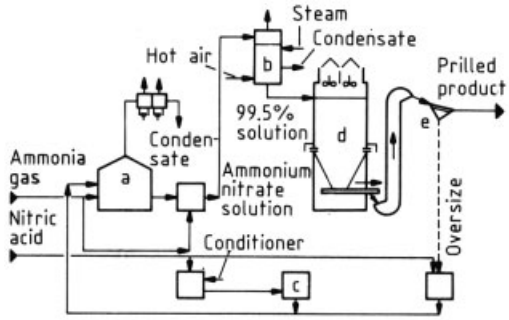

Figure 8. ICI process with prilling tower
a) Ammonium nitrate neutralizers; b) Falling film evaporator; c) Filter; d) Prilling tower; e) Screen

The solution is sprayed through one-component nozzles, perforated plates, or perforated centrifuges at the top of the tower. Cold air is drawn through the tower countercurrently to remove the heat evolved on crystallization. As the droplets fall through the tower, they solidify into round granules, which are discharged at the foot of the prilling tower and subsequently cooled and screened.

There is no need for the granulate to be dried if 99.5 % NH_4NO_3 melt is sprayed. The prilled grain has a 2 – 3 mm diameter, thus smaller than from most other processes. Grain size is affected to a certain extent by the shape of the nozzle, by the perforation cross sections and the rotational speed of the spray centrifuge, and by the viscosity of the melt.

Serious emission problems attributable to ammonium nitrate fumes are often encountered in prilling on account of the large air throughput. Modern plants are generally equipped with special emission control systems. In the CFCA shroud system [21], for example, the cooling air is divided by baffle plates or tubes in the prilling tower into two streams. Only the inner stream, containing most of the pollutants, is cleaned. Usually the off-gases from the neutralization and evaporation stages are also cleaned.

The universally applicable recycle salt process (Fig. 9), developed by BASF in 1928 [22] for *granulating* calcium ammonium nitrate, produces 2 – 4 mm granules. The process can be modified, for example, by arranging a revolving

drum after the paddle screw to improve granulation, by carrying out drying or cooling in a fluidized bed, or by subjecting the granulate to hot or cold screening.

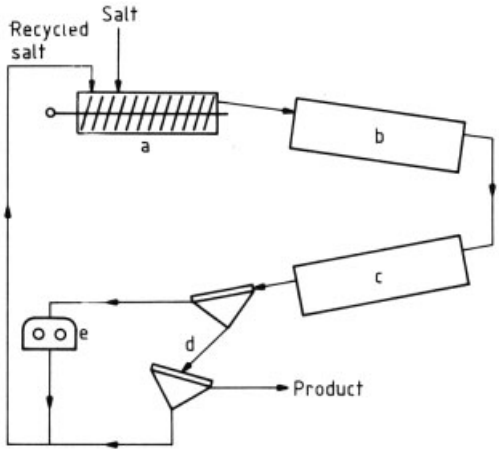

Figure 9. Recycle salt process
a) Granulation screw; b) Drying drum; c) Cooling drum; d) Screens; e) Grinder

The *pan granulation process* has been further developed, e.g., by TVA and Norsk Hydro. Figure 10 [23] shows the plan of the Norsk Hydro process. In the sloping, rotating pan granulator, the 99.5 % ammonium nitrate melt is sprayed onto a moving solid bed and solidifies on the cold particles. The growing granules are graded in the rotating pan and, after reaching the required grain size, roll over the edge. Grain size may be varied by changing the operating parameters: a normal fertilizer grain diameter of 2 – 4 mm and a grain size of 7 – 11 mm for forest fertilization can be made in the same granulator.

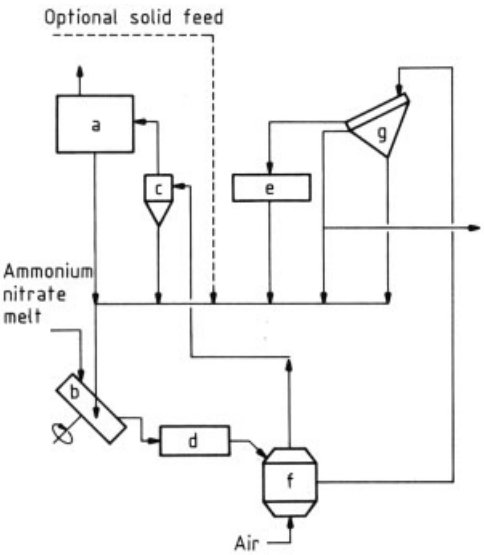

Figure 10. Norsk Hydro pan granulation process
a) Dust filter; b) Pan; c) Cyclone; d) Polishing drum; e) Crusher; f) Cooler; g) Screen

Ammonium nitrate can also be granulated in a revolving drum, known as a spherodizer, or in a fluidized bed (e.g., Fisons Nitro-Top process) [24].

In addition to the production of ammonium nitrate containing 33.5 % or 34.5 % of N, each of these processes is also suitable for the production of calcium ammonium nitrate containing 26 % – 28 % N.

The explosives industry requires a porous, low-density (LD) ammonium nitrate that can be impregnated with oil. This LD ammonium nitrate can be produced by spraying with hot air in two-component nozzles into a revolving drum heated with hot air. However, porous LD ammonium nitrate can also be produced by prilling [25]. An ammonium nitrate melt of relatively low concentration (96 %) is sprayed in the prilling tower, and the prilled product is then carefully dried with hot air in two successive dryers. The product contains 0.2 % of water and has a powder density of 800 g/L and a porosity of 33 vol%.

1.4. Grain Stabilization and Surface Treatment (Conditioning)

The crystal transition point at 32 °C and the hygroscopicity adversely affect the storability of the granulated ammonium nitrate. A temperature of 32 °C can often be exceeded, particularly during transport and storage in warm countries: the granules lose strength and finally disintegrate into dust. The modification change and other solid-phase reactions can be inhibited by low moisture content and additives that act as internal stabilizers. The additives are added to the melt before it is sprayed. Examples of additives are anhydrous $CaSO_4$ [26], $H_3BO_3 + (NH_4)_2HPO_4 + (NH_4)_2SO_4$ [27], ammonium and potassium polyphosphate [28], and nucleating agents, such as silica gel [29], metal oxides, or kaolin [30]. $Mg(NO_3)_2$ [31] and $Al_2(SO_4)_3$ [32] are important stabilizers. In gen-

eral, the stabilizers also inhibit the reaction between NH_4NO_3 and $CaCO_3$, which forms the highly hygroscopic calcium nitrate.

Usually, the ammonium nitrate granules are prevented from caking during storage by a final surface treatment (coating). The earlier practice of powdering with inorganic substances to separate the grains mechanically has mostly been replaced by treatment of the grain with effective organic surfactants, such as cationic long-chain fatty amines or anionic alkylarylsulfonates. Anticaking effectiveness is enhanced by addition of nonionic organic substances, such as polyethylene (PE) waxes [33] or paraffins. The fatty amines are applied as melts or oil solutions. One process [34] is multistage: in the first stage the fertilizer granules are coated with a uniform, compact layer of molten $C_{12}-C_{18}$ alkylamine. In the second stage they are cooled to convert the amine into the solid state. In the third and final stage, they are provided with a thin coating of mineral oil.

1.5. Transport Regulations

National and international regulations governing the transport of ammonium nitrate and its mixtures are being based to an ever-increasing extent on the classification of the "United Nations Committee of Exports on the Transport of Dangerous Goods" [35].

A number of ammonium nitrate-based products are to be transported according to regulations for Class 5.1, e.g., technical ammonium nitrate UN No. 1942 or the various ammonium nitrate fertilizers UN No. 2067 – 2070. Main criteria for this class are concentrations of the components and a strict limitation of combustible substances. Mixtures with high concentrations of ammonium nitrate or combustible materials are not allowed to be transported unless they are clearly defined as explosives, Class 1. Mixtures with ammonium nitrate contents below the limits are considered nonhazardous, with the exception of one group that may be subject to self-sustaining decomposition (smoldering) when heated. This group of AN fertilizers is classified in Class 9 (UN No. 2071).

Regulations for the marine transport of the above classes are given in the IMDG Code (Inter-Governmental Maritime Consultant Organisation, London) on pages 5014, 5015, and 9013.

The international regulations for road and rail transport, ADR (Accord européen relatif au transport international des marchandises dangereuses par route) and RID (Règlement international concernant le transport des marchandises dangereuses par chemins de fer) have brought the ammonium nitrate products also into Class 5.1, Subdivision 6. The criteria of composition correspond closely to those of the IMDG Code.

Transport on inland waterways is regulated by ADN (Accord européen relatif au transport international des marchandises dangereuses par voie de navigation interieure) and for the Rhine by ADNR. Although ammonium nitrate products are still placed in a special class (Class III c), these regulations are being redrafted to harmonize the classification.

1.6. Safety

Although stable under normal conditions, ammonium nitrate undergoes a number of decomposition reactions at elevated temperature [36], [37]. The following are significant:

1) Endothermic dissociation and lowering of the pH above 169 °C.

$$NH_4NO_3 \longrightarrow HNO_3 + NH_3 \quad \Delta H = +175\ \text{kJ/mol}$$

2) Exothermic elimination of N_2O on careful heating at 200 °C.

$$NH_4NO_3 \longrightarrow N_2O + 2\,H_2O \quad \Delta H = -37\ \text{kJ/mol}$$

3) Exothermic elimination of N_2 and NO_2 above 230 °C.

$$4\,NH_4NO_3 \longrightarrow 3\,N_2 + 2\,NO_2 + 8\,H_2O \quad \Delta H = -102\ \text{kJ/mol}$$

4) Exothermic elimination of nitrogen and oxygen, accompanied by detonation.

$$NH_4NO_3 \longrightarrow N_2 + 1/2\,O_2 + 2\,H_2O \quad \Delta H = -118.5\ \text{kJ/mol}$$

Pure ammonium nitrate, highly concentrated hot solutions of ammonium nitrate, some mixtures of ammonium nitrate, and fertilizer ammonium nitrate, unless stabilized, are included

among explosives detonated by shock waves. Although the heat released is small in comparison with that of an explosive such as hexogen, there are serious risks attending the storage of large quantities. For example, hydrogen ions, chlorides, and heavy metals catalyze the decomposition. The heating of contaminated and tamped ammonium nitrate is particularly dangerous. Following the disasters in Brest and Texas City (1947) [38], [39], where whole ships' cargoes of wax-coated fertilizer ammonium nitrate exploded as a result of fires, the content of inflammable substance has been universally limited to 0.2 % or 0.4 %. In the Federal Republic of Germany, special procedures have been introduced under the *Arbeitsstoffverordnung* (Working Materials Act) [40] for storage, loading, and transport of ammonium nitrate and products containing ammonium nitrate. For example, the storage of ammonium nitrate capable of detonation is permitted only in small batches in specially equipped storage sites. There are similar regulations in many other European countries, whereas in the United States, France, Norway, and England the storage of relatively large amounts is permitted under some conditions.

In Germany inert materials, such as limestone powder, dolomite, or precipitated calcium carbonate, are added to ammonium nitrate for fertilizer use (calcium ammonium nitrate). Such fertilizers, containing no more than 80 % ammonium nitrate, no more than 0.4 % combustible constituents, and no less than 18 % magnesium or calcium carbonate, rate as nondetonatable. The sensitivity of ammonium nitrate and its mixtures to shock is determined by tests in which a sample tamped in a closed steel tube is exposed to a shock [36], [40], [41]. In addition, oil retention [41], a measure of porosity, is a basis for assessing the safety of ammonium nitrate.

To summarize:

1) If products containing ammonium nitrate have solidified during storage, they should not be broken up by blasting
2) Products containing ammonium nitrate must be stored separately from oxidizable, inflammable materials
3) If ammonium nitrate must be heated (during production, evaporation, etc.), the quantity must be kept small and catalysts avoided

1.7. Uses

Ammonium nitrate (AN) is used mainly as fertilizer, either pure, diluted, or in multinutrient mixtures. It is used with urea in liquid fertilizers, which are important in the United States, Eastern Europe, and France.

In the United States, the United Kingdom, and France, ammonium nitrate having a nitrogen content of 33.5 % and higher is used in large quantities in agriculture. A fertilizer containing 32.5 % N is also used in the United States. In the Federal Republic of Germany, ammonium nitrate is used in mixtures with lime, dolomite, ammonium sulfate, or potash. Among these calcium ammonium nitrate is particularly important (→ Fertilizers).

Calcium Ammonium Nitrate (CAN). Ammonium nitrate solution (ca. 95 – 97 %) can be granulated with finely divided calcium carbonate from crushed limestone or from the conversion of calcium nitrate. The granules obtained are dried, cooled, screened, and treated to avoid caking. The reaction of ammonium nitrate and limestone to form hygroscopic calcium nitrate is prevented by additives, such as $(NH_4)_2SO_4$, $MgSO_4$, and $FeSO_4$. In Germany, the nitrogen-content of calcium ammonium nitrate was increased in stages from the original level of 20.5 % N to the present level of 27.5 % N. The upper limit under the regulations of the European Economic Community is 28 % N.

Ammonium Sulfate Nitrate (ASN). The mixed sulfate – nitrate fertilizer is produced by adding ammonium sulfate to approximately 95 % ammonium nitrate solution [42] or by neutralizing HNO_3–H_2SO_4 mixtures with ammonia [43]. A slightly hygroscopic product is obtained after granulation. It is a mixture of the double salt $2\,NH_4NO_3 \cdot (NH_4)_2SO_4$ and a little ammonium sulfate [44], containing 26 % N for < 45 % ammonium nitrate. The mixture tends to harden during storage as a result of further reactions. This can be prevented by addition of salts of Mg, Fe, or Al [45].

Potassium Ammonium Nitrate. Potassium ammonium nitrate is produced in the same way

Table 4. Production of fertilizer NH_4NO_3, 10^6 t N, 1981

	Production	Import	Export
Western Europe	4.8	1.0	0.9
Eastern Europe	6.3	–	0.4
North America	1.9	0.1	0.1
South America	0.3	–	–
Africa	0.5	0.2	–
Asia and Oceania	1.0	0.1	–
Total	14.8	1.4	1.4

as ammonium sulfate nitrate, except that a potassium salt (the chloride or sulfate) is added to give, e.g., 20–0–20 ($N–P_2O_5–K_2O$) fertilizer.

Nitromagnesia. A fertilizer can be produced from ammonium nitrate, ammonium sulfate, and magnesium compounds, such as dolomite, magnesium carbonate, or magnesium sulfate. For example, such a fertilizer may contain 20 % N, 8 % MgO, and generally 0.2 % Cu.

Other Uses. Safety explosives for mining are produced from ammonium nitrate (AN). It is suitable by virtue of its low explosion temperature. In mixtures with NaCl, the explosion temperature is not high enough to set off the much-feared fire damp. For safety explosives, ammonium nitrate must have the following qualities:

NH_4NO_3 content		99.0 %
Ignition residue	max.	0.2 %
Chloride	max.	0.1 %
Nitrite	max.	0.1 %
Sulfate	max.	0.1 %
Water insoluble	max.	0.1 %

In many cases, safety explosives are also based on the reciprocal salt pair $NaNO_3 + NH_4Cl \rightleftharpoons NH_4NO_3 + NaCl$. Where greater explosive power is required, as in mining (rock explosives), porous prilled ammonium nitrate containing approximately 6 % diesel oil is used (ANFO = ammonium nitrate fuel oil) (→ Explosives).

To a lesser extent, ammonium nitrate is used to produce dinitrogen monoxide, N_2O. The salt must be 99.5 % NH_4NO_3 and be free of organic materials, iron, chlorides, and sulfates.

1.8. Economic Aspects

Table 4 provides a synopsis ofthe production of ammonium nitrate for fertilizers in various countries.

2. Ammonium Sulfate

Ammonium sulfate [*7783-20-2*], $(NH_4)_2SO_4$, was produced industrially in the 19th century, mainly from the ammonia in coke-oven gas. At the beginning of this century ammonia was synthesized on an industrial scale, and ammonium sulfate became a widely used fertilizer. Its importance, however, has steadily declined with the development of more concentrated nitrogen fertilizers. Small quantities are used for industrial purposes. Since about 1960, ammonium sulfate has been produced to an increasing extent as a coproduct in organic syntheses. In some countries ammonium sulfate is produced from gypsum.

2.1. Properties

Ammonium sulfate, M_r 132.14, d_4^{20} 1.77, average specific heat capacity 1.423 $J\,g^{-1}\,K^{-1}$ between 2 and 55 °C, rhombic bipyramidal. The size and habit of the crystals can be affected significantly by substances in the crystallizing solution, and this can be an important factor in commercial production.

Ammonium sulfate cannot be melted at atmospheric pressure without decomposition, releasing ammonia and leaving bisulfate. However, the ammonia vapor pressure of pure, anhydrous ammonium sulfate is effectively zero up to 80 °C. Above 300 °C, decomposition gives N_2, SO_2, SO_3, and H_2O in addition to ammonia.

The solubility of ammonium sulfate in water is shown in Figure 11. The salt does not form hydrates. The heat of solution on dissolution of 1 mol of salt in 400 mol of water is + 9.92 kJ at 18 °C, the integral heat of solution is + 6.57 kJ/mol at 30 °C, and the differential heat of solution for saturated solution is + 6.07 kJ/mol at 30 °C. Ammonium sulfate deliquesces only above 80 % relative humidity; therefore, the salt can be stored in dry air. The solubility of ammonium sulfate is reduced considerably by addition of ammonia: At 10 °C, from 73 g $(NH_4)_2SO_4$ in 100 g of water, nearly linearly, to 18 g salt in 100 g of 24.5 % aqueous ammonia.

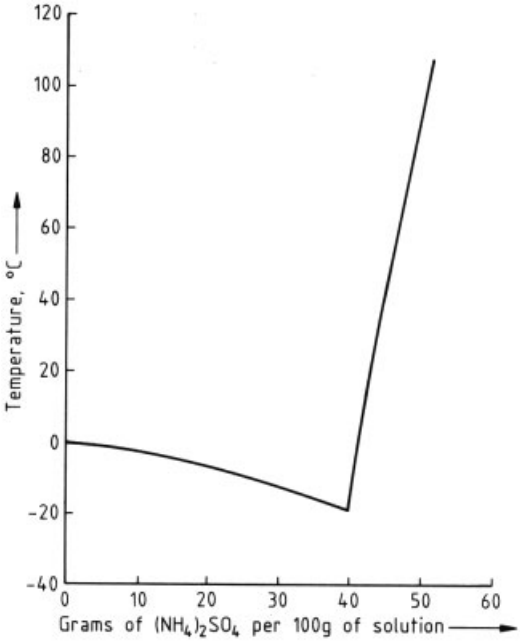

Figure 11. The system $(NH_4)_2SO_4$–H_2O

Calcium sulfate is approximately twice as soluble in ammonium sulfate solution as in water, which favors reaction of gypsum or anhydrite with ammonium carbonate – ammonium sulfate solutions. The salt is insoluble in the usual organic solvents. Dissolution of ammonium sulfate in aqueous solutions of ethanol, propanol, butanol, acetone, pyridine, etc., causes two phases to form, one aqueous and the other the organic solvent.

Special Cr – Ni stainless steel is resistant to ammonium sulfate solutions and to solutions of the salt in sulfuric acid. Ammoniacal solutions do not attack iron or aluminum. If the solutions contain substances that corrode stainless steel (e.g., Cl^-), the vessels can be lined with acid-resistant bricks.

2.2. Production

Ammonium sulfate is produced from

1) Coke-oven gas
2) Ammonia and sulfuric acid
3) Organic syntheses, such as the production of caprolactam
4) Gypsum, ammonia, and carbon dioxide

2.2.1. From Coke-Oven Gas

The production of ammonium sulfate from coke-oven gas has diminished considerably over the past twenty years, because of the partial closure of steel mills and the development of coking processes producing less ammonium sulfate [46].

In the direct process the unpurified coke-oven gas is introduced into sulfuric acid to yield ammonium sulfate contaminated by colored tar products. In the indirect process the ammonia is washed out of the coke-oven gas with water, released with a lime suspension, and finally introduced into sulfuric acid.

2.2.2. From Ammonia and Sulfuric Acid

The heat of reaction between ammonia and sulfuric acid is sufficient to evaporate all the water if an acid concentration of 70 % or higher is used.

$$2\,NH_3\,(g) + H_2SO_4\,(l) \longrightarrow (NH_4)_2SO_4\,(s) \qquad \Delta H = -274\,\text{kJ/mol}$$

Today, the reaction is usually carried out in saturators, devices that are derived from the evaporation crystallizers commonly used earlier [47–49].

In the saturator process (Fig. 12), neutralization and crystallization are carried out in one and the same apparatus. The sulfuric acid is delivered to the suction side and the ammonia to the pressure side of the forced circulation pump. Crystallization of the metastable solution gives particle sizes generally between 0.5 and

3 mm. The salt is continuously discharged at the lower end of the saturator. The salt is separated in centrifuges, dried, and cooled. The mother liquor is returned to the saturator. Impurities in the sulfuric acid can adversely affect crystallization. Small quantities of phosphoric acid, urea, or inorganic salts are added to promote crystal growth.

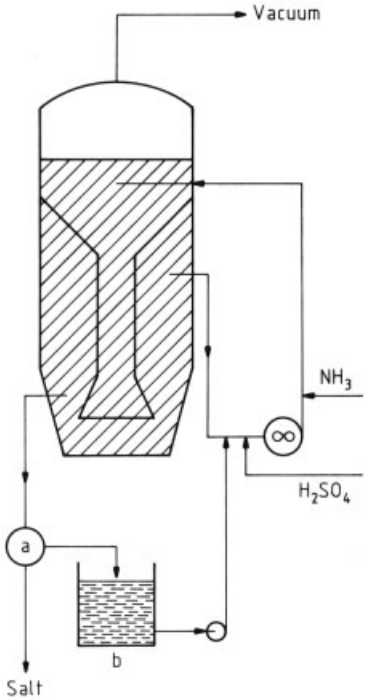

Figure 12. Ammonium sulfate saturator
a) Centrifuge; b) Mother liquor tank

2.2.3. Coproduct in Organic Syntheses

Ammonium sulfate is coproduct in the production of synthetic-fiber intermediates, such as caprolactam, acrylonitrile, and methyl methacrylate, and in the production of formic acid and acrylamide. The most important source is the production of caprolactam, which is required for nylon 6. The conventional caprolactam process (→ Caprolactam) produces 2.5 – 4.5 t of ammonium sulfate per tonne of lactam. Of that, 0.3 – 2.8 t is formed in the hydroxylamine – oximation stage and approximately 1.7 t in the Beckmann rearrangement stage. Crystallizers are used in the evaporation of ammonium sulfate solutions that result from caprolactam production. The processes developed by BASF [50], [51], Inventa [50], [52], [53], Toray Industries [54], Snia Viscosa [50], Techni-Chem [52], Kanebo [54], [55], and DSM [56] have reduced the 2.5 – 4.5 t, in some cases, to 1.7 – 1.8 t per tonne of lactam.

2.2.4. From Gypsum

Both anhydrite and gypsum react with NH_3 and CO_2:

$$CaSO_4(\cdot 2\,H_2O) + (NH_4)_2CO_3 \longrightarrow (NH_4)_2SO_4 + CaCO_3(+2\,H_2O)$$

The process, developed by BASF during the First World War, has retained its importance where sulfuric acid is in short supply, including India, Pakistan, and Turkey.

Finely ground gypsum is reacted with ammonium carbonate solution in a cascade of stirred vessels. The residence time is several hours. The reaction mixture of calcium carbonate and ammonium sulfate solution is filtered on rotary vacuum filters. The washed calcium carbonate can be used for fertilizing lime, the production of calcium ammonium nitrate, raw material for glass, or filler for rubber or PVC [57]. The slightly cloudy ammonium sulfate solution is filtered in filter presses, acidified with H_2SO_4, and processed in multistage evaporation crystallizers to form coarse-grained ammonium sulfate. (For details and a flowsheet, see [58]) An alternate method for converting gypsum into ammonium sulfate and calcium carbonate is the Continental Engineering Process [59], in which NH_3 and CO_2 are directly introduced into a gypsum slurry in a tall, cylindrical stirred vessel.

In principle, byproduct gypsum from the phosphoric acid process may be used too. However, it may contain phosphoric acid, silicates, fluorosilicates, and organic substances, which make filtering difficult, and may require preliminary washing.

2.2.5. Other Processes

There are several obsolete processes that start from SO_2 and atmospheric oxygen [60], [61]. There has been interest in processes using am-

monia to remove SO_2 from power-station exhaust. For example, the Walther process [62] produces a concentrated ammonium sulfite solution by two-stage washing of the dust-free exhaust. The solution is oxidized with atmospheric oxygen and spray-dried to ammonium sulfate powder, which is subsequently granulated [63].

2.3. Granulation and Storage

Typical crystallizers generally produce a 0.5- to 3-mm grain that can be dried to a water content of less than 0.1 %. The free acid content should not exceed 0.03 %, and the N content should not be below 21 %. The grains are protected against caking, generally by small quantities (0.1 %) of surfactants, and are storable.

For use in bulk blending, a procedure for producing the blended fertilizers widely used in the United States, the ammonium sulfate is generally granulated. This is done by compacting, pan granulation, or use of the TVA Ammoniator-Granulator [64]. Granulation can be improved by the use of additives, e.g., calcium salt [65]. The mixed granulation of concentrated urea solutions with ammonium sulfate crystals, e.g., in a pan granulator, gives a hard granulate of high nitrogen content (30 – 43 % N) suitable for bulk blending [66].

2.4. Uses

Almost all of the ammonium sulfate (AS) produced is used as fertilizer; very little is used industrially. In industrial countries ammonium sulfate is almost always a coproduct or byproduct, and it can only be sold as a fertilizer, mostly in the developing countries (Table 5). In industrial countries its low nitrogen content makes transportation cost per unit of nitrogen higher than for other nitrogen fertilizers. In Africa and Asia ammonium sulfate is used especially to fertilize rice, tea, and rubber. In Europe, the United States, and Brazil, it is often a component in blended and complex fertilizers (→ Fertilizers).

Table 5. World consumption of ammonium sulfate as fertilizer in 1983, 1000 t N

Western Europe	285
Eastern Europe	818
North America	146
South America	419
Asia	824
Africa	98
Others	30
Total	2620

In industry $(NH_4)_2SO_4$ is used for the production of persulfates, flameproofing agents, and fire-extinguishing powders; in tanning; in the photographic, textile, and glass industries; and as a nutrient for yeast and bacterial cultures.

2.5. Economic Aspects

World ammonium sulfate capacity is shown in Table 6 [46]. Table 7 [46] shows the worldwide consumption of ammonium sulfate, which stagnated in 1976. As can be seen from the tables, production capacities are by no means fully utilized.

The transportation costs for shipping ammonium sulfate from the countries in which it is produced, the industrialized countries, into the countries in which it is consumed, the developing countries, reduces any profit. Production in some countries, e.g., China, Mexico, and Spain, has been based on a political decision – to increase domestic employment or preserve limited foreign exchange. In addition, worldwide consumption has been decreasing steadily. These three facts have been the driving force in the industrialized countries behind reducing the coproduction of ammonium sulfate in the production of synthetic fibers. If efforts to reduce sulfur emissions produce additional amounts of ammonium sulfate, which also will have to be sold in the developing world, the pressure on profits will increase further.

3. Ammonium Chloride

Although ammonium chloride [*12125-02-9*], NH_4Cl, M_r 53.49, occurs naturally in volcanic material, production from natural sources is of no significance. The history of industrial NH_4Cl production is intertwined with the development of the soda industry and the large-scale production of synthetic NH_3. The original motivation for producing ammonium chloride was to use it as fertilizer, but it is now used for a wide range of purposes. Ammonium chloride processes are

Table 6. World ammonium sulfate capacity, 1981/82, 1000 t N

	Synthetic	Coproduct	Byproduct and others	Total
World	2900	2289	982	6171
Western Europe	561	913	158	1632
Eastern Europe	210	631	402	1243
North America	219	264	225	708
Central America	395	36	26	457
South America	74	32	1	107
Africa	40	–	39	79
Asia	1387	413	75	1875
Oceania	14	–	56	70

Table 7. Worldwide consumption of ammonium sulfate, 1000 t N

	1976/77	1977/78	1978/79	1979/80	1980/81
World	2814.7	2890.8	2796.0	2916.3	2828.4
Western Europe	461.6	434.0	468.5	471.5	430.8
Eastern Europe	729.9	718.9	684.9	661.7	653.5
North America	208.8	178.7	152.2	172.2	191.2
Central America	325.6	304.8	310.8	341.0	353.4
South America	173.9	226.4	211.0	233.8	238.8
Africa	130.9	138.0	144.3	141.3	128.2
Asia	759.3	858.7	802.4	868.0	769.2
Oceania	24.7	31.3	21.9	26.8	63.3

of historical interest to chemical engineering because they are an early example of chemical processing involving all three phases: solid, liquid, and gas.

3.1. Properties

Solid ammonium chloride has a specific gravity d_4^{20} of 1.530. Its average specific heat $\overline{c}_p$ between 298 and 372 K is 1.63 kJ/kg.

Ammonium chloride has two modifications [67]. The transformation between the two is reversible at 457.6 K (184.5 °C):

$$\alpha\text{-}NH_4Cl \text{ (cubic, CsCl type)} \leftrightharpoons$$
$$\beta\text{-}NH_4Cl \text{ (cubic, NaCl type)} \rightleftharpoons$$
$$\Delta H = +4.3 \text{ kJ/mol}$$

The α modification is the one stable at room temperature. β-NH_4Cl melts at 793.2 K under 3.45 MPa; it sublimes at atmospheric pressure. In fact, NH_4Cl is quite volatile at lower temperatures, dissociating into NH_3 and HCl:

T, K	523.2	543.2	563.2	583.2	603.2	611.2
p, kPa	6.6	13.0	24.7	45.5	81.4	101.3

The solubility of NH_4Cl in water increases with temperature

T, K	273.2	293.2	313.2	333.2	353.2	373.2	389.2
c, wt %	22.9	27.2	31.5	35.6	39.7	43.6	46.6

The integral heat of solution to saturation is + 15.7 kJ/mol, and the differential heat of solution at saturation is + 15.2 kJ/mol. The solubility in water may be increased by adding ammonia. NaCl tends to salt NH_4Cl out of aqueous ammoniacal solution [68], [69]. The partial pressures of saturated NH_4Cl solutions show that NH_4Cl is weakly hygroscopic

T, K	283.2	293.2	303.2	313.2	323.2	389.2
p, kPa	1.0	1.9	3.3	5.4	8.8	101.3

Moisture causes product caking; however, whether moisture contents less than 0.1 wt % alone are responsible for the caking of ammonium chloride, is questionable. Another possible explanation is sublimation.

Ammonium chloride is very soluble in liquid NH_3 but virtually insoluble in acetone and pyridine. At 292.7 K its solubility in methanol is 3.24 wt %, and its solubility in ethanol is 0.64 wt %.

The crystal form obtained from aqueous solutions can be affected by other substances. This was exploited to produce large crystals, and it is

a help in determining impurities arising during NH_4Cl production [70], [71].

3.2. Production

Ammonium chloride is produced commercially by two processes:

1) Modified Solvay process (ammonia – soda ash process or ASAP)
2) Direct reaction between HCl and NH_3

A third process, the reaction of reciprocal pairs of salts [72], has not yet been commercially applied, but is still of interest.

3.2.1. Modified Solvay (Ammonia – Soda Ash) Process

In the Solvay process, ammonia and carbon dioxide are dissolved in aqueous sodium chloride to produce sparingly soluble sodium bicarbonate, which is calcined to sodium carbonate. The ammonia is recovered from the mother liquor by reaction with lime, this reaction also producing the calcium chloride. The source of the lime and the carbon dioxide is limestone. The net reaction is the conversion of the feedstocks rock salt and limestone into sodium carbonate and the byproduct calcium chloride.

Water, carbon dioxide, and ammonia are added only to the extent necessary to compensate for plant losses. The feedstock and product dictate the location of a Solvay plant. For this reason, the plants normally function as largely independent units (→ Sodium Carbonates).

The modified Solvay process (ammonium chloride – soda ash process) is one of the oldest examples of integrated industrial production of two substances. It differs from the Solvay process in that ammonium chloride is also precipitated from the mother liquor. The net reaction is

$$2\,NH_3 + CO_2 + H_2O + 2\,NaCl \longrightarrow 2\,NH_4Cl + Na_2CO_3$$

The amounts of ammonium chloride and soda ash produced are almost equal: two moles (107 g) of NH_4Cl are produced for each mole (106 g) of Na_2CO_3. The modified process, requiring external NH_3 and CO_2, must be incorporated into an integrated system of plants.

Furthermore, the treatment of ammonia-containing waste gases is carried out for environmental reasons, rather than to reduce NH_3 and CO_2 losses. Finally, the energy balance in the modified process is entirely different from that in the traditional Solvay process because of the different feedstocks and products.

Process Description. As a rule, the manufacture of NH_4Cl and Na_2CO_3 is carried out in a continuously operated recycle process in which the concentrations within the cycle may be adjusted according to whether ammonium chloride or sodium carbonate is the primary product [76]. BASF originally developed the process in which NH_4Cl is the primary product (Fig. 13).

The cycle begins with the addition of ammonia and carbon dioxide to the aqueous recycle solution. The ammonium bicarbonate intermediate reacts further with the sodium chloride present in the solution to yield ammonium chloride and sparingly soluble sodium bicarbonate. After thickening, the latter is separated out with a centrifuge; it is washed during centrifugation. It is calcined to produce sodium carbonate.

The sodium bicarbonate remaining in the mother liquor is decomposed by heating to 337.2 K. The very soluble sodium carbonate does not interfere with the crystallization of ammonium chloride. The high temperature of the solution at this point is utilized to dissolve the rock salt feedstock rapidly. This rock salt is introduced via a series of mixing tanks, the quantity dictated by the material balance over the entire cycle. Any insoluble impurities in the rock salt are filtered off. The major impurities are sulfates and other salts of iron, calcium, and magnesium. Disposal of the residues represent an increasing problem. Therefore, pure salt is replacing rock salt as raw material.

Ammonium chloride is recovered from the clear solution in a two-stage crystallization by indirect cooling. The suspension from the crystallizer is thickened in hydrocyclones, and the solid is separated and washed in a centrifuge. The solid from the centrifuge containing 5 – 7 wt % water is dried in a current drier to a moisture content less than 0.1 %. The final product can be coated with an anticaking agent. After the removal of the ammonium chloride, the cycle is complete, and the mother liquor can be reused.

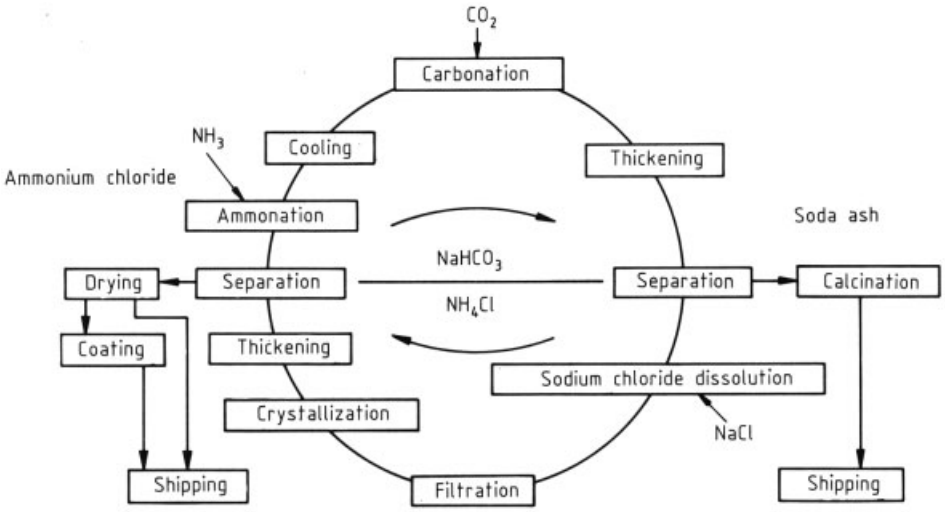

Figure 13. Ammonium chloride – soda ash process

A process that produces soda ash as the primary product has been developed by Asahi Glass [77], [78].

Ammonia is added to the mother liquor prior to the crystallization of ammonium chloride. Crushed raw salt is added with or without washing and dissolved in the solution. The solution is then cooled below 20 °C to crystallize ammonium chloride. The crystals are separated with a centrifugal separator, and the mother liquor is recycled to the carbonation section. The ammonium chloride crystals are dried to a moisture content below 0.3 % in a rotary or fluidized bed drier. Crystal sizes can be kept between 1 and 2 mm for granule or about 200 µm in diameter for high-grade compound fertilizer.

In the carbonation section, the solution is sent into a carbonating tower for the precipitation of sodium bicarbonate through the reaction with blown-in carbon dioxide at 30 – 40 °C. The resulting slurry is separated into sodium bicarbonate crystals and the mother liquor. The mother liquor is sent to the ammonia absorption section. The separated crystals are calcined and converted to dense soda ash.

This process is capable of producing bicarbonate as pure as that from the conventional Solvay process and fertilizer-grade ammonium chloride. The purity of ammonium chloride exceeds 97 wt %.

As shown in Figure 14, the Asahi Glass process includes all the steps shown in Figure 13; however, the sequence differs. The ammonia is added after separation of the sodium bicarbonate. The heat of solution of ammonia serves to provide the heat needed to decompose the residual sodium bicarbonate. However, the energy saving is offset to some extent by the higher energy need in the crystallization because ammonium chloride is more soluble in the ammonia-rich solution [68], [69]. A washing step for the rock salt feed, prior to solution, avoids the filtration before crystallization. With or without washing this process utilizes the raw salt to the fullest extent, an important consideration in Japan where the salt must be imported.

Equipment Variations. Little is known about the process details of the individual process variants. Nevertheless, general characteristics of the process technology, adequate for plant selection, can be provided.

The individual steps of several other important process variants are described below.

If the feed gases are available at sufficiently high pressure, they can be mixed with the recycle solution by gas – liquid contacting devices, such as jet injectors. In the case of ammonia the amount of excess gas is reduced, and the solution is homogenized rapidly. In the case of CO_2 the nucleation rate and crystal growth of $NaHCO_3$ are improved.

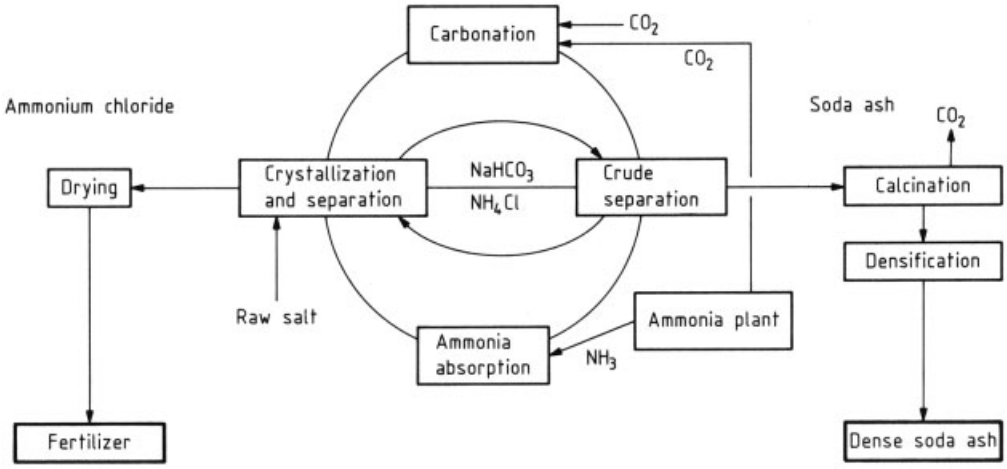

Figure 14. Soda ash – ammonium chloride coproduction process

Further parameters influencing the size of the $NaHCO_3$ crystals are temperature and residence time in the carbonation step. The various carbonation columns used in the production of $NaHCO_3$ differ in the number and type of internal parts such as mixing devices and additional heat exchangers, the positioning and type of the CO_2 nozzles, and the volume needed to produce a given $NaHCO_3$ equivalent. The large $NaHCO_3$ crystals may be separated and washed using a band filter, or after thickening they may be separated by centrifugation.

The important differences in NH_4Cl crystallization lie in the number of stages and type of cooling. Direct cooling during the crystallization, in comparison with indirect cooling, uses more energy and requires the use of additional equipment for refrigerant condensation and recovery. However, it avoids problems associated with indirect cooling, i.e., the operation of heat exchangers, from which a NaCl crust must be periodically removed. Special plate heat exchangers have been developed to solve the fouling problem, and these are to be preferred to shell and tube heat exchangers and coil coolers.

$NaHCO_3$ and NH_4Cl crystallization and removal are the keys to high conversion per pass and product purity.

Ammonium chloride is frequently dried in a drum drier instead of a pneumatic drier.

Pollution Problems. The pollution problems associated with plant operation are not critical. By careful operation the production of wastewater streams can be almost entirely eliminated. However, because of its high ammonia content, any wastewater produced will require treatment.

The scrubbing of waste gases with hydrochloric acid to produce additional ammonium chloride offers a neat solution to the problem of gaseous emissions. Generally ammonia conversions of 98 % and rock salt conversions of 95 % are achieved, the exact values depending on the type of process and the environmental measures.

Grades and Capacity. After closure of several production plants due to reduced market demand, the capacity for NH_4Cl by the modified Solvay process for the western hemisphere and Japan was estimated at ca. 100000 t/a in 1997. Only limited information is available on total capacities and applied technology in China. In Asia, a granular product containing more than 97 % NH_4Cl is usually produced. The ammonia nitrogen content of this fertilizer grade NH_4Cl always exceeds 25 wt %. These plants are primarily for the production of soda ash.

However, if the production of NH_4Cl is of primary interest, the process can be carried out in such a way as to give purities exceeding 99.7 wt % NH_4Cl with < 0.25 wt % NaCl and < 3 ppm Fe, a quality adequate for most industrial purposes. Reagent-grade NH_4Cl can be produced on a large scale with special operat-

ing techniques but without additional processing steps. This ammonium chloride contains < 0.01 wt % NaCl and corresponds to the ACS specification [79].

3.2.2. Direct Reaction Between HCl and NH_3

The synthesis of NH_4Cl from HCl and NH_3 is economically attractive if HCl is available as feedstock at low or no cost.

The *Engeclor process* [80], developed by the Brazilian company Engeclor, carries out the reaction in aqueous solution:

Ammonia is fed into the conical section of a saturator while HCl diluted with air is passed into the NH_4Cl suspension. The reaction occurs at 353 K, under reduced pressure, and with excess NH_3 (pH 8). The suspension is drawn off from the base of the saturator and thickened in hydrocyclones. The NH_4Cl is separated in a centrifuge and dried. The mother liquor is recycled to the saturator. The waste gases from the saturator are scrubbed with water.

Schemes have been proposed whereby gaseous feedstocks are fed into nonaqueous solvents. The heat of reaction

$$NH_3(g) + HCl\ (g) \rightleftharpoons NH_4Cl\ (s)\ \Delta H = -176\ \text{kJ/mol}$$

is removed by the evaporation of the solvents, which are subsequently condensed [81]. Fluidized bed processes have also been described [82]. To avoid aerosol formation, carbon dioxide is recommended as carrier gas in the fluidized bed [102].

The annual world production of NH_4Cl via the $HCl - NH_3$ process was estimated to be 50000 t in 1997. Ammonium chloride produced by this method generally contains < 0.1 wt % NaCl. The metal content, and especially the heavy metal content, depends on the plant. As a rule the levels are higher than those encountered in the modified Solvay process.

3.2.3. Reaction of Reciprocal Pairs of Salts

The reaction of reciprocal salts [72] is still of interest. Suitable pairs of salts seem to be $(NH_4)_2SO_4 - KCl$, which gives $NH_4Cl - K_2SO_4$ [73–75][103][104] and $(NH_4)_2SO_4 - NaCl$ which gives $NH_4Cl - Na_2SO_4$ [106]. The basic problem with these salts is the lower purity of NH_4Cl due to the high content of sulfate. NH_4Cl and KNO_3 are obtained by reaction of NH_4NO_3 and KCl. Different process modifications are described in [105, 107–109]. High-purity KNO_3 can be used in manufacturing explosives, and NH_4Cl with a purity of 96 % min. can be used as a fertilizer. The preparation of $NaNO_3$ and NH_4Cl from NaCl and NH_4NO_3 has been investigated [110], [111].

3.3. Corrosion

Ammonium chloride is corrosive, as gas, as solid, or in solution, because (1) it is acidic, (2) it complexes metal ions, and (3) it contains the corrosive chloride ion. Therefore, corrosion protection is one of the key factors for production reliability and product quality. Several forms of corrosion are observed.

For solutions, the extent of corrosion largely depends on the NH_4Cl concentration, the pH, and the temperature. Equipment made from iron, aluminum, lead, or nonferrous metals is especially prone to stress corrosion cracking.

Older plants were constructed with rubber-lined pipes and containers that had an additional ceramic coating. Today special steels or plastics are usual. Low carbon, pure austenitic, or austenitic – ferritic steel, for example, that designated as SS 360 in the United States, is used up to 313 K. In borderline cases Hastelloy or Ti [83] is used.

The steel must be correctly worked. For example, welding must be carried out under a protective gas, and welded seams always should be posttreated. Up to 333 K polyethylene or fiberglass-reinforced polypropylene can be used where abrasion and erosion are absent. Heat exchangers of the graphite material Diabon have proved satisfactory. Cloths made from polyacrylonitrile are most suitable for filtration. Protection against external corrosion may also be necessary. Protective coatings with an epoxide or phenoxy resin base have proved effective. Such coatings are effective also for the internal protection of vessels or tanks not exposed to much mechanical wear and tear, such as storage silos.

3.4. Packaging, Storage, Shipping, and Handling

On account of its corrosive nature and its tendency to cake, NH_4Cl is best packed in sacks, or in the case of "flowable" goods in big bags, of paper or polyethylene rather than plastic or metal drums.

Despite its low water content (< 0.1 wt %) ammonium chloride tends to cake. There are two methods to enhance the "flowability." First, additives such as soda or B_2O_3 serve to separate the individual particles and preferentially absorb any water present. Product coated with such materials retains the ability to flow freely for a limited storage period and can be stored in silos on a temporary basis. Second, individual particles are encapsulated with hydrophobic chemicals. Fatty acid amines, e.g., stearylamine, and their acetate and hydrochloride derivatives are commonly used.

No special precautions beyond those usual for chemicals are needed for shipping ammonium chloride. Ammonium chloride should not be allowed to come into contact with nitrates or strong acids or alkalies. It is not covered by industrial chemical legislation although European Economic Community guidelines require labeling as a weak poison [84]. Ammonium chloride for the food industry must meet the specifications of the *European Pharmaceutical Handbook* [85] in Europe and those of the Food and Nutrition Board [86] in the United States. The latter restrict the use of anticaking agents.

3.5. Uses

Ammonium chloride is widely used as an effective nitrogen fertilizer for paddy and upland rice, wheat, and other crops in Japan, China, and South East Asia [77]. Most of the annual production of NH_4Cl in Japan is used as high-grade compound fertilizers: chloro-ammonium phosphate, chloro-potash-ammonium phosphate, magnesia-chloro-potash-ammonium phosphate, and nitrogen-potash. Further fertilizer uses are limited by the acidity and high chlorine content.

The industrial uses of technical-grade ammonium chloride are, in order of importance: solid electrolytes in dry cell batteries, one component of quarrying explosives (an especially fine form is used), hardeners for formaldehyde-based adhesives, one component of etching solutions in the manufacture of printed circuit boards, and a component, along with zinc chloride, of fluxes in tin and zinc plating. Minor uses include a rapid fixer additive in photography, cleaner additives, and a nutrient in yeast cultures. Ammonium chloride is also used in tanning, refining of precious metals, textile printing and dying, and in the rubber industry. The addition of ammonium chloride to tiles and bricks prior to firing was proposed. The ammonium chloride serves to control product porosity and accelerates the firing process [87]. The use of NH_4Cl as hardener for formaldehyde-based adhesives has been reduced significantly due to substitution by chloride-free products.

High-purity ammonium chloride is employed in the food and pharmaceutical industries and in a few chemical syntheses.

4. Ammonium Carbonates

Salt of hartshorn was first mentioned at the beginning of the 14th century in English manuscripts. It was produced by dry distillation of antlers, hooves, and leather and consisted of a solid mixture of NH_4HCO_3, $NH_4CO_2NH_2$, and $(NH_4)_2CO_3 \cdot H_2O$. This salt of hartshorn was used as an expanding agent for certain baked goods on account of its decomposition into entirely gaseous products and good handling properties. It was manufactured on a semi-industrial scale at the beginning of the 19th century.

Following the development of the industrial synthesis of ammonia, BASF started large-scale production of ammonium carbonate in 1919. Today, the term salt of hartshorn refers to a mixture of pure NH_4HCO_3 and $NH_4CO_2NH_2$ containing 32.5 wt % NH_3. The following compounds appear in the ternary system NH_3–CO_2–H_2O [88], [89]:

NH_4HCO_3	ammonium hydrogencarbonate
$(NH_4)_2CO_3 \cdot H_2O$	ammonium carbonate
$(NH_4)_2CO_3 \cdot 2\,NH_4HCO_3 \cdot H_2O$	ammonium sesquicarbonate
$NH_4CO_2NH_2$	ammonium carbamate

All these substances can be produced from NH_3 and CO_2. The high purity of the feedstocks

assures adequate product purity without further processing, provided the materials of construction are suitable.

4.1. Ammonium Hydrogencarbonate

Ammonium hydrogencarbonate [*1066-33-7*], ammonium bicarbonate, NH_4HCO_3, M_r 79.06, ϱ 1.586 g/cm^3. The decomposition pressure of ammonium bicarbonate increases rapidly with temperature [90]:

T, K	298.6	307.4	313.9	318.2	323.2	329.0	332.5
p, kPa	7.85	16.26	26.79	37.06	52.65	82.11	108.64

The solubility of ammonium bicarbonate in water increases with temperature [90]

T, K	273.2	283.2	293.2	303.2	313.2	323.2	333.2
c, wt %	10.6	13.9	17.8	22.1	26.8	31.6	37.2

The reaction forming NH_4HCO_3 is exothermic:

$$NH_3\,(g) + CO_2\,(g) + H_2O\,(l) \rightleftharpoons NH_4HCO_3\,(s) \qquad \Delta H = -126.5\,\text{kJ/mol}$$

In the $NH_3 - CO_2 - H_2O$ system, ammonium bicarbonate is the only compound soluble in water without decomposition. Upon dissolution of one mole of salt in 6 – 8 L H_2O the heat of solution is + 26.4 kJ at 288.2 K. The dissociation constant in aqueous solution is $K_d = 1.45$ [91] for the reaction $NH_4HCO_3 \rightleftharpoons NH_4^+ + HCO_3^-$.

Ammonium bicarbonate is best produced in aqueous solution, be it continuously or batch. The process is relatively easy to control. The crystallization is the critical stage. The heat of reaction must be removed, while the residence time, concentration, and temperature profile have to be controlled to yield large, easily separable crystals.

The flow diagram of a continuous process is shown in Figure 15. Ammonia is added to the mother liquor, which is kept cool, to achieve the desired NH_3 concentration of slightly above 10 wt %. Carbon dioxide is added and allowed to react in an absorption column. The resulting solution, which has been warmed by the reaction, is cooled indirectly in a crystallizer. The crystal suspension is withdrawn, thickened in hydrocyclones, and passed to a centrifuge, where the solid is separated. This solid is pneumatically dried, cooled, and conditioned. The mother liquor is recycled. If stainless steel equipment is used, the product meets the specifications of the *European Pharmaceutical Handbook* [92] and the *Food Chemical Codex* [93] without further purification.

The world production in 1997 was estimated at ca. 100000 t/a, almost half of which was produced in the western hemisphere and the remainder in Asia. The process variants of individual producers usually differ in the absorption/crystallization equipment, and this also results in different particle size distributions. The theory of CO_2 and NH_3 adsorption is discussed in [94], and crystallization with direct cooling is described in [95].

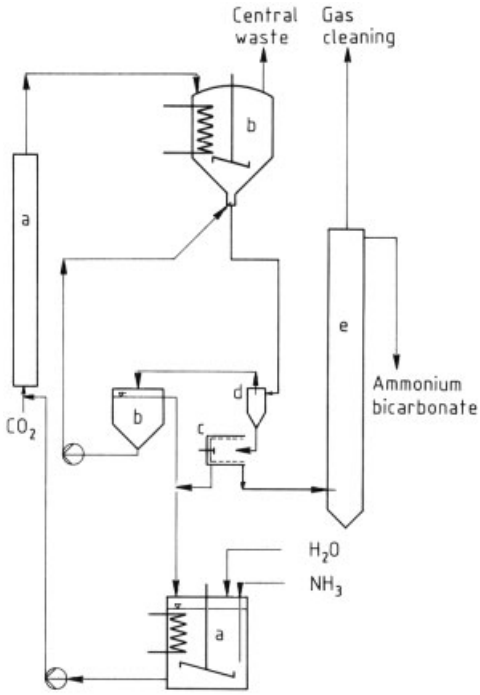

Figure 15. Production of ammonium bicarbonate
a) Gas absorption; b) Crystallization; c) Centrifugation; d) Thickening; e) Drying

4.2. Ammonium Carbamate

Ammonium carbamate [*1111-78-0*], $NH_4CO_2NH_2$, M_r 78.08, melts in a sealed tube at 425.2 K.

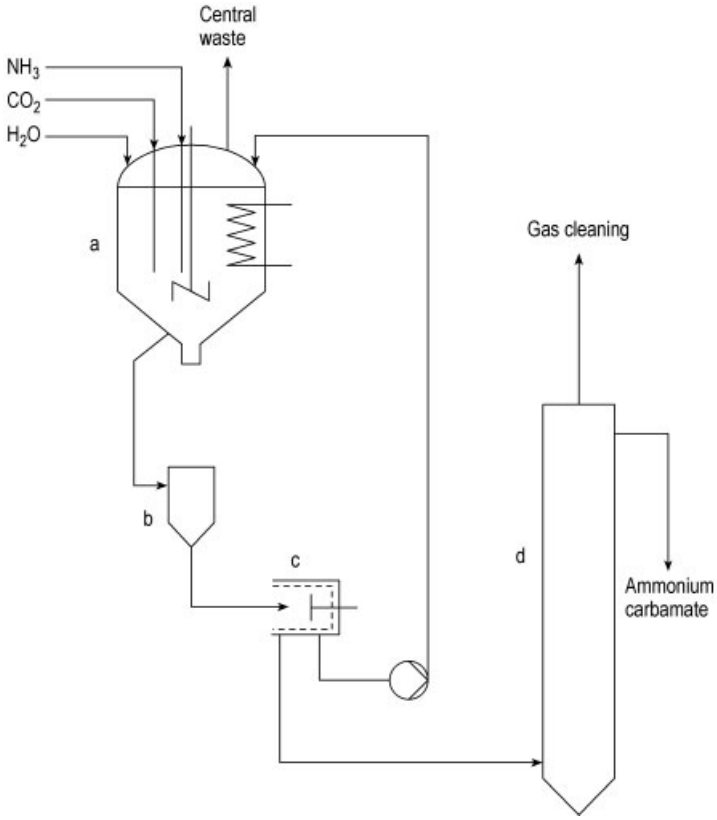

Figure 16. Production of ammonium carbamate
a) Crystallizer; b) Thickening; c) Centrifuge; d) Drying

The dissociation pressures show that the compound is quite volatile and thermally unstable [90].

T, K	283.19	288.10	291.10	294.41	298.10
p, kPa	38.0	55.3	68.50	88.50	114.80
T, K	299.97	304.10	309.07	313.05	318.02
p, kPa	130.40	173.70	242.70	315.30	431.00

The solid consists of dimers [96]. The heat of solution in water is + 15.9 kJ/mol.

Ammonium carbamate is manufactured by two methods. In the first process, which directly gives dry ammonium carbamate, predried NH_3 and CO_2 are fed into a heat exchanger, where they react and are cooled. The carbamate condenses within the pipes and is removed later by partial melting. This carbamate is sold as carbamate rocks.

The second process (Fig. 16), also semibatch, uses a concentrated $(NH_4)_2CO_3$ solution feed. Ammonia and carbon dioxide are fed into the solution, which is kept cool. The carbamate crystals are separated batchwise by centrifugation and dried pneumatically or under a protective CO_2 atmosphere.

A fluidized-bed process has been proposed for carbamate production [97].

Ignoring the large-scale production associated with urea manufacture (→ Urea), the world carbamate capacity in 1997 was estimated at 4000 t/a.

4.3. Ammonium Carbonate

Ammonium carbonate [*506-87-6*], $(NH_4)_2CO_3 \cdot H_2O$, M_r 114.11, crystallizes as flat, columnar, prismatic crystals or as elongated flakes. The carbonate melts at 316.2 K, simultaneously decomposing to the carbamate and sesquicar-

bonate. $(NH_4)_2CO_3 \cdot H_2O$ is only formed under precisely defined conditions, which are difficult to achieve in any production plant. Therefore usually equimolar mixtures of ammonium carbamate and ammonium hydrogencarbonate are used and sold as ammonium carbonate. These mixtures contain the same amount of ammonia as $(NH_4)_2CO_3 \cdot H_2O$. The world carbonate capacity in 1997 was estimated at 7000 t/a.

4.4. Storage, Transport, and Handling

The ammonium carbonates show noticeable dissociation pressures at temperatures as low as 293 K. Both carbamate and carbonate decompose to yield the bicarbonate, resulting in product losses and composition changes. For this reason the substances should be stored in sealed containers and kept as cool as possible. In sealed containers decomposition occurs only to a limited extent because progressive decomposition is prevented by the establishment of equilibrium. The rate of decomposition of NH_4HCO_3 is also a function of particle size [95]. Ammonium hydrogencarbonate has a tendency to cake, and its flowability is maintained by anticaking additives such as corn flour, saccharose, MgO, or $MgCO_3$.

There are, in general, no special requirements with regard to transport, safety precautions, and handling.

4.5. Uses

Ammonium bicarbonate is mainly used as an expanding agent for certain baked goods (→ Bread and Other Baked Products); it is also used in buffer solutions for neutralization of acids when additional anions are undesirable. Further uses include an ammonization agent for the humic acid in turf, a nitrogen source for yeast cultures, and blowing agents for foam rubber and poly(vinyl chloride). Minor uses include smelling salts and formaldehyde binders in laminates. Ammonium hydrogencarbonate is also used in the production of textiles, ceramics, pigments [98], and leather.

Ammonium carbamate is used as an insecticide for cereals, a neutralization agent in the chemical industry, and a feedstock for ammonium carbonate production.

Ammonium carbonate is used in the manufacture of catalysts [99–101], as blowing agent in foam rubber and foam plastic, and as an additive in photographic developers. In a number of countries it is also used as an expanding agent in certain baked goods.

5. Toxicology and Occupational Health

Ammonium chloride is harmful by ingestion; the oral LD_{50} in rats is 1410 mg/kg [112]. The i.v. LD_{50} in mice is 7 – 10 mmol/kg [113]. Ammonium chloride causes no skin irritation, but it is irritating to the eyes [114]. Ammonium chloride is not sensitizing in animals [115]. Oral treatment of various animal species with daily doses of 500 – 1000 mg/kg causes metabolic acidosis [116–118]: plasma and urinary pH and dioxide-combining power decrease, whereas chloride, urea nitrogen, and solids concentrations in the blood and gluconeogenesis increase. The Ames test [119][120] with *S. typhimurium* and *E. coli* and the micronucleus test [121] in mice did not show mutagenic effects. Ammonium chloride was not carcinogenic in rats and mice after chronic exposure in the diet and in drinking water [122–125].

Ammonium bicarbonate is harmful by ingestion. The LD_{50} (rat, oral) is ca. 1570 mg/kg [126] and the LD_{50} (mouse, i.v.) is 3.1 mmol/kg [127]. Rats fed for 8 days at a daily dose of 474 mg/kg showed only a slight increase in the ammonia content of the urine [128]. No mutagenic effects were detected in the Ames test with six different strains of *S. typhimurium* and in the chromosomal aberration test in vitro with a Chinese hamster fibroblast cell line [129].

Ammonium carbonate is virtually nontoxic after a single oral uptake. The LD_{50} (rat, oral) is 2150 mg/kg [126] and the LD_{50} (mouse, i.v.) is 1.02 mmol/kg [127]. Rats fed for 5 weeks with 5 % ammonium carbonate in the diet showed depressed growth and elevated urea concentrations in the blood [130]. Chinchilla rabbits were given 100 – 200 mg/kg ammonium carbonate by gavage or in their drinking water for 5 to 26 months [131], [132]. The treatment cycle was

for three weeks with the test compound followed by one week without. When ammonium carbonate was applied in the drinking water, the only notable effect was a parathyroid hypertrophy. Given by gavage ammonium carbonate leads to enlargement of adrenals, ovaries, mammary glands, and womb, as well as lactation and proliferation of ovarian follicles and corpora lutea. However, because these effects were found only in the gavage study, the relevance to humans seems to be doubtful.

Ammonium carbamate is harmful by ingestion. The LD_{50} (rat, oral) is in the range 690 – 1470 mg/kg [126] and the LD_{50} (mouse, i.v.) is 0.99 mmol/kg [127]. It is not irritating to rabbit skin, but irritates the rabbit eye [133]. No mutagenic effects were detected in the Ames test using four different strains of *S. typhimurium* [134]. Because ammonium carbamate is quite volatile and thermally unstable, forming ammonia and carbon dioxide, the MAK value of ammonia of 20 ppm (TWA 25 ppm) must be taken into account.

6. References

General References

1. G. D. Honti: *The Nitrogen Industry,* Akadémiai Kiadó, Budapest 1976.
2. J. L. Chadwick: *Fertilizers: Nitrogen, Interim Report No. 127 A 1, Private Report by Process Economics Program,* Menlo Park, Calif., 1980.
3. L. Medard: *Les Explosifs Occasionnels,* **vol. 2,** Technique et Documentation, Paris 1979.
4. *Winnacker-Küchler,* **vol. 2.**
5. G. S. Scott, R. L. Grant: *Ammonium Nitrate, Its Properties and Fire Explosion Hazards,* U.S. Bureau of Mines, J.C. 7463 (1948).
6. Chemico, *Nitrogen* 1967, no. 49, 27 – 31.
7. Soc. Belge de l'Azote, DE 971418, 1951.
8. J. Morrison, *Oil Gas Int.* **8** (1968) no. 9, 93.
9. Soc. Belge de l'Azote, *Nitrogen* 1960, no. 5, 28 – 31.
10. D. Wagener, K. H. Laue, *Chem. Ing. Tech.* **50** (1978) no. 6, 421 – 425.
11. G. Fauser, DE 590469, 1931.
12. A. David, C. Parmentier, J. Passelecq, *Hydrocarbon Process.* **57** (1978) no. 11, 169 – 175.
13. Stamicarbon, *Fertilizer Know-How from A to Z.*
14. NSM/Norsk Hydro, *A New Low Energy Ammonium Nitrate Process,* Nov. 1980.
15. L. A. Stengel, US 2568901, 1947 (and later patents).
16. A. S. Hester, J. J. Dorsay, J. T. Kaufman, *Ind. Eng. Chem.* **46** (1954) 622 – 632. J. J. Dorsay, *Ind. Eng. Chem.* **47** (1955) 11 – 17.
17. G. Langhans, B. Bieniok, *ISMA 1976 Technical Conference The Hague 1976,* Elsevier, Amsterdam 1976, preprints pp. 215 – 233.
18. I. W. McCamy, M. M. Norton: "Production of Granular Urea, Ammonium Nitrate and Ammonium Polyphosphate – a Process Review," in A. I. More (ed.): *Granular Fertilizers and their Production,* Pap. Int. Conf., The British Sulphur Corporation Ltd., London 1977.
19. *Nitrogen* 1978, no. 116, 34 – 39.
20. *Nitrogen* 1968, no. 56, 27 – 29.
21. *Nitrogen* 1977, no. 107, 34 – 39.
22. BASF, DE 691686, 1928.
23. *Nitrogen* 1975, no. 95, 31 – 36.
24. *Nitrogen* 1973, no. 86, 33 – 37.
25. R. Collins, *Nitrogen* 1979, no. 118, 35 – 37.
26. Oesterreichische Stickstoffwerke AG, AT 306754, 1972.
27. Mississippi Chemical Corp., US 3630712, 1971.
28. Veba-Chemie AG, US 4001377, 1977.
29. Stamicarbon, NL 6709519, 1969.
30. Fisons Fertilizer Ltd., GB 1189448, 1970.
31. R. W. R. Carter, A. G. Roberts, *Proc. Fert. Soc.* 1969, no. 110, 31 – 32.
32. C. Sjölin, *J. Agric. Food Chem.* **19** (1971) no. 1, 83 – 95.
33. BASF, DE 1905834, 1969.
34. NSM, US 4150965, 1979.
35. Library of Congress Catalog No. 81-82269, ISBN 0-940394-00-9 (1981).
36. G. Hansen, W. Berthold, *Chem. Ztg.* **96** (1972) 449 – 455.
37. M. Berthelot, *Rev. Sci.* **6** (1883) 8.
38. V. J. Clancey, *RARDE Memor.* 1962, no. 10, 48 – 63.
39. *Chem. Age* **61** (1949) 485 – 490.
40. Verordnung über gefährliche Arbeitsstoffe in der Fassung vom 11. 2. 1982, Anhang II, Nr. 11, Bundesarbeitsblatt 4/1982.
41. EG-Richtlinie (80/876/EWG) vom 15. 7. 1980, Amtsblatt der Europäischen Gemeinschaften Nr. L 250, 7 – 11.
42. BASF, DE 355037, 1919.

43. Gew. Victor, DE 650381, 1929; DE 614705, 1931.
44. E. Jänecke, *Z. Anorg. Allg. Chem.* **160** (1927) 171 – 184.
45. I. G. Farbenind., DE 612708, 1930.
46. *Nitrogen* 1983, no. 141, 26 – 30.
47. H. Svanoe, *Ind. Eng. Chem.* **32** (1940) no. 5, 636 – 639.
48. Th. Messing, *Chem. Ztg.* **91** (1967) no. 24, 963 – 967.
49. Standard-Messo, DE-AS 1260440, 1964.
50. *Eur. Chem. News* 1976, no. 734, 24 – 26.
51. BASF, GB 1372108, 1971; DE-AS 2508247, 1975.
52. *Eur. Chem. News* Caprolactam Supplement, 2 May 1969.
53. *Eur. Chem. News Suppl.* 1972, no. 552, 3 – 4.
54. *Chem. Eng. News* **51** (1973) no. 15, 14 – 15.
55. Kanebo Ltd., US 3689477, 1971.
56. S. J. Loyson, G. H. J. Nunnink, *Hydrocarbon Process.* **51** (1972) no. 11, 92 – 94.
57. H. Hofmann, W. Witte, J. Schlicke, H. J. Wassermann, *Silikattechnik* **32** (1981) no. 9, 269 – 271.
58. *Ullmann,* 4th ed., **vol. 7,** 523 – 524.
59. J. F. Witte, J. J. De Wit, R. M. Voncken, *Chem. Weekbl.* **65** (1969) no. 37, 25 – 30.
60. *Ullmann,* 3rd ed., **vol. 3,** p. 618.
61. *Chem. Eng. News* **38** (1960) no. 35, 44.
62. H. Bechthold, *Tech. Mitt. Krupp Werksber.* **39** (1981) no. 2, 43 – 48.
63. Buckau-Walther, DE-OS 3108986, 1981 (H. Bechthold).
64. *Nitrogen* 1967, no. 45, 33.
65. Kali und Salz, DE 2603917, 1976 (S. Luther, W. v. Maessenhausen).
66. G. C. Hicks, J. M. Stinson, *Ind. Eng. Chem. Process Des. Dev.* **14** (1975) no. 3, 269 – 276.
67. *Gmelin,* 8th ed., System 23, 150 – 191.
68. E. Weitz, *Z. Elektroch.* **31** (1925) 546.
69. E. Weitz, H. Stamm, *Ber. Dtsch. Chem. Ges.* **61** (1928) 1144.
70. A. W. Bamforth, *Chem. Eng. (London)* 1974, 455 – 457.
71. E. V. Khamskii, N. S. Dregubskii, *J. Appl. Chem. USSR* **46** (1973) 2449 – 2455.
72. V. E. Pikalow et al., *Khimiko-Farmatsevticheskii Zhurnal* **10** (1976) no. 8, 111 – 112; *Pharm. Chem. J.* **10** (1976) 1095 – 1096.
73. Allied Chemical Corp., US 3595609, 1969.
74. Kali und Salz AG, DE 2142114, 1971.
75. Dan Kako Co Ltd., JP 4717693, 1971.
76. Z. Rant, *Die Erzeugung von Soda,* F. Enke Verlag, Stuttgart 1968, pp. 136 – 137.
77. Asahi Glass, JP 200173, 1953 (S. Uemura et al.).
78. T. Miyata, *Chem. Ind. (London)* 1983, 21 Feb., 142 – 145; K. Tsunashima, K. Nakaya, The New Asahi (NA) Process for Synthetic Soda Ash Production, Fifth Ind. Miner. Int. Cong. Madrid, Spain, 25 – 28 April 1982, H 1 – H 9.
79. ACS, Reagent Chemicals, 6th ed., ACS, Washington, D.C., 1981.
80. A. W. Bamforth, S. R. S. Sastry, *Chem. Process Engng.* **53** (1972) no. 2, 72 – 74.
81. F. Shadman, A. D. Randolph, *AIChE J.* **24** (1978) 782 – 788.
82. V. G. Shlyakhtor, V. V. Streitsor, *Theor. Found. Chem. Engng.* **8** (1974) no. 1, 133 – 136.
83. A. K. Gorbachev et al., *The Soviet Chem. Ind.* **10** (1978) 1004 – 1005.
84. Guideline 67/5481 of the Commission of the EEC, 5th Adaption.
85. *European Pharmaceutical Handbook*, vol. 1, Deutscher Apotheken Verlag, Stuttgart 1974.
86. *Food Chemical Codex,* 3rd ed., National Academy Press, Washington, D.C., 1981.
87. Studie van de Relaties tussen chemische, fysische en mineralogische Kenmerken van de Boomse Klin en van der Verhittingsprodukten, Katholische Universität Leuven, Doctor Jos. Decleer 1983, 113 – 136.
88. E. Terres, H. Weiser, *Z. Elektroch.* **27** (1921) 177 – 244.
89. E. Jänecke, *Z. Elektrochem.* **35** (1929) 723 – 727.
90. *Gmelin,* 8th ed., System 23, 42 – 66.
91. G. M. Marion, G. R. Dutt, *Soil. Sci. Soc. Am. Proc.* **38** (1974) 889 – 891.
92. *European Pharmaceutical Handbook,* **vol. 1,** Deutscher Apotheken Verlag, Stuttgart 1974.
93. *Food Chemical Codex,* 4th ed., National Academy Press, Washington, D.C., 1996, p. 25 – 26.
94. V. Rod, M. Rylek, *Collect. Czech. Chem. Commun.* **39** (1974) 1996 – 2006.
95. J. P. Usyukin, *Khim. Prom.* **8** (1972) 610 – 611; *Soviet Chem. Ind.* **4** (1973) 267 – 270.
96. A. Gieren et al., *Angew. Chem.* **85** (1973) 308 – 309.
97. BASF, DE-OS 3346719, 1985 (K. Kinkel).
98. C. B. Pimentel, *Rev. Quim. Ind.* **39** (1970) 453.
99. Shell Oil Co., US 5234477, 1993.
100. Platinum Plus Inc., WO 9504211, 1995.
101. Meidensha Corp., JP 10 033947, 1998.
102. BASF AG, EP 718238-A1, 1996.
103. S. Deng, CN 1144772 A, 1997.
104. Dalian Chem. Res. Inst. Min. Chem. Ind., CN 1131640-A, 1996.

105. Project and Development India Ltd., IN 170975-A, 1992.
106. Broul, CS 8702-618-A, 1988.
107. Rubezahn, SU 1393-791-A, 1989.
108. Pot-Nitrate Mfr. Pty., ZA 8805-130-A, 1989.
109. Chem Factory Tulufa, CN 8601-635-A, 1987.
110. Dongshen Inst. Fine Chem., *Huaxue Shijie* **38** (1997) no. 5, 237 – 240.
111. Chem Factory Tulufa, CN 8603-687-A, 1987.
112. BASF AG, unpubl. results, 1983.
113. D. P. Stombaugh et al., *Anim. Sci.* **6** (1969) 844.
114. BASF AG, unpubl. results, 1969.
115. Hoechst AG, unpubl. results, 1987; cited in Hoechst GDS, April 14th 1994.
116. W. D. Lotspeich, *Am. J. Phys.* **206** (1965) 1135.
117. P. Vinay et al., *Kidney Int.* **17** (1980) 312.
118. D. Z. Levine, *Am. J. Physiol.* **6** (1973) 380.
119. Hoechst AG, unpubl. results (87.0392) 1987; cited in Hoechst GDS, April 14th 1994.
120. M. Ishidate et al., *Food Chem. Toxicol.* **22** (1984) 623 – 636.
121. M. Hyashi et al., *Food Chem. Toxicol.* **26** (1988) 487 – 500.
122. T. Fujii et al., *Food Chem. Toxicol.* **25** (1987) 359 – 362.
123. S. Fukushima et al., *Cancer Research* **46** (1986) 1623 – 1626.
124. M. Arai et al.: Symposia on the 4th Int. Cancer Congress, Budapest, vol. 1 (120), 1986, p. 21 – 26.
125. A. Flasks, B. D. Clayson, *Br. J. Cancer* **31** (1975) 585 – 587.
126. BASF AG, unpubl. results, 1989.
127. R. P. Wilson et al., *Am. J. Vet. Res.* **29** (1968) 897.
128. J. Oliver, E. Bourke, *Clin. Sci. Mol. Med.* **48** (1975) 515.
129. M. Ishidate ,Jr., et al., *Food Chem. Toxicol.* **22** (1984) 623.
130. J. S. Finlayson, C. A. Baumann, *J. Nutrition* **59** (1956) 211.
131. I. G. Fazekas, *Orvosi Hetilap* **90** (1949) 777; cited also in WHO Tech. Rep., Ser. No. 683, 1982.
132. I. G. Fazekas, *Endokrinologie* **32** (1954) 45. Cited also in WHO Tech. Rep., Ser. No. 683, 1982.
133. BASF AG, unpubl. results, 1958.
134. BASF AG, unpubl. results, 1991.

Ammonium Phosphate → Phosphate Fertilizers
Amorphous Metals → Metallic Glasses
Amyl Alcohols → Pentanols
Amylase → Enzymes

Analeptics

Alfred Hofmeister, Chemische Werke Minden GmbH, Minden/Westphalia, Federal Republic of Germany

Analeptics are pharmaceuticals for treatment of depressed excitability of the central nervous system. They comprise a varied group of substances, whose use in medicine is decreasing in importance continually. This fact can be explained by the increasing awareness that analeptics, as far as they are used as psychostimulants (e.g., amphetamines), frequently lead to abuse and drug dependence. Therefore, these drugs are controlled by narcotics laws, or they have been withdrawn completely from the market. However, this has not prevented their extensive use on the drug scene.

Analeptics have only one indication, namely, acute intoxication, such as hypnotic poisoning or postnarcotic complications with lesion of the respiratory and circulatory centers. For these cases, the following preparations are still available:

Bemegride [*64-65-3*], 4-ethyl-4-methyl-2,6-piperidinedione, $C_8H_{13}NO_2$, M_r 155.19, *mp* 127 °C. Trade names: Mikedimide (Panray), Eukraton (Nordmark), Malysol (Arco, Switzerland), Megimide (Nicholas). Bemegride is used in place of picrotoxin in acute poisoning with hypnotics.

Crotetamide [*6168-76-9*], *N*-(1-dimethylcarbamoylpropyl)-*N*-ethylcrotonamide, $C_{12}H_{22}N_2O_2$, M_r 226.31, *bp* 132 – 134 °C (30 Pa). Trade name (in combination with cropropamide [*633-47-6*]): Micoren (Ciba-Geigy).

Fominoben hydrochloride [*18053-31-1*], 3′-chloro-α-[methyl[(morpholinocarbonyl)-methyl]-amino]-*o*-benzotoluidide · HCl, $C_{21}H_{25}Cl_2N_3O_3$, M_r 438.36, *mp* 206 – 208 °C (decomp.). Trade name: Noleptan (Thomae), Oleptan (Bender, Austria), Terion (Lusofarmaco, Italy).

Amiphenazol hydrochloride [*490-55-1*], 5-phenyl-2,4-thiazole diamine · HCl, $C_9H_{10}ClN_3S$, M_r 259.78. Trade name: Daptazile (Nicholas).

Crotetamide, fominoben, and amiphenazol are used especially for treating acute respiratory disorders. Today the alkaloid lobelin [*90-69-7*] is seldom used.

Derivatives of amphetamine, which in fact are also analeptics, now are used only as appetite depressants, whereas derivatives of ephedrine are used mainly in combination drugs as bronchospasmolytics. The strongly psychostimulating drugs, such as *methamphetamine hydrochloride* [*537-46-2*], trade name Pervitin (Temmler), and *methylphenidate hydrochloride* [*113-45-1*], trade name Ritalin (Ciba), are considered correctly to be narcotics. *Pemolin* [*2152-34-3*], trade name Tradon (Beiersdorf), is still available as a mild stimulant.

The stimulating drugs used in psychiatry, for example, for depressions with retarded activity, are classified as psycho-active drugs, in particular, as the so-called thymoleptics. The antihypotonics used for treating low blood pressure are not classified as analeptics.

The methods available to modern intensive medicine have made the analeptics, which were much used previously, largely unnecessary. Attempts to deal with widespread misuse – by narcotics laws, obligatory prescriptions, and withdrawal of some preparations from the market – have been partially successful.

References

1. D. R. Curtis: "The Pharmacology of Central and Peripheral Inhibition," *Pharmacol. Rev.* **15** (1963) 333 – 364.
2. D. W. Esplin, B. Zablocka-Esplin: "Mechanism of Action of Convulsants," in: H. H. Jasper, A. A. Ward, Jr., A. Pope (ed.): Basic Mechanism of the Epilepsies, Little, Brown & Co., Boston 1969, pp. 167 – 183.
3. F. Hahn: "Analeptics", Pharmacol. Rev. 12 (1960) 446 – 530.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_267

Analgesics and Antipyretics

Elmar Friderichs, Grünenthal GmbH, Center of Research, Aachen, Germany
Thomas Christoph, Grünenthal GmbH, Center of Research, Aachen, Germany
Helmut Buschmann, Laboratorios Dr. Esteve S.A., Barcelona, España

1. Introduction 1
2. Nonsteroidal Anti-inflammatory Drugs (NSAIDs, COX Inhibitors) 3
2.1. Salicylates 7
2.2. para-Aminophenol Derivatives 8
2.3. Anthranilates 10
2.4. Arylacetic Acids 12
2.5. Arylpropionic Acids 19
2.6. Pyrazolinone Derivatives 27
2.7. Acidic Enolic Compounds 28
2.7.1. Pyrazolidine-3,5-diones 28
2.7.2. Arylsulfonamides (Oxicames) . . 30
2.8. Other Structures 32
2.9. COX-2 Inhibitors 33
2.9.1. Nonselective COX-2 Inhibitors . 33
2.9.2. Selective COX-2 Inhibitors (Coxibs) 35
3. Centrally Acting Analgesics . . 42
3.1. Opioids 42
3.1.1. Opioids in Clinical Use 49
3.1.1.1. Morphine and Morphinane Derivatives 49
3.1.1.2. Piperidine Derivatives 56
3.1.1.2.1. Meperidine and Congeners 56
3.1.1.3. Fentanyl and Congeners 58
3.1.1.4. Methadone and Congeners 60
3.1.2. Other Structures 64
3.1.2.1. Mixed Opioid Agonist – Antagonists and Partial Agonists 66
3.1.2.2. Opioid Antagonists 72
3.2. Antineuropathic Analgesics and Other Non-opioid Compounds . 74
3.2.1. Neuropathic Pain Treatment . . . 74
3.2.2. Individual Compounds 76
4. References 85

1. Introduction

Compounds that are used for the treatment of pain, inflammation, and fever are classified into two main groups according to their mode of action: nonsteroidal anti-inflammatory drugs (NSAIDs) and so-called centrally acting analgesics.

Nonsteroidal anti-inflammatory drugs are compounds with a predominantly peripheral mechanism of action. Besides inhibition of pain they exert a more or less pronounced anti-inflammatory and fever-reducing effect. In the body key peripheral mediators of pain, inflammation, and fever are prostaglandins (→ Prostaglandins). NSAIDs and related "peripheral" or "weak analgesics" act via inhibition of cyclooxygenase. Cyclooxygenase is the key enzyme of prostaglandin synthesis and its inhibition may be the most important, but not the exclusive mechanism of action of NSAIDs. The supposed additional mechanisms are widely unknown.

The group of *centrally acting analgesics* is dominated by the opioid compounds. They act mainly within the central nervous system, but more recent investigations indicate that an additional peripheral action component may exist. The endogenous opioid system is the most important pain control system within the body and opioids are powerful tools for the treatment of severe pain situations. Central neurotransmitters like noradrenaline and serotonin are likewise involved in pain inhibition. Directly acting compounds like, e.g., the α_2-adrenergic agonist clonidine as well as indirectly acting inhibitors of noradrenaline and serotonin reuptake (mainly used as antidepressants) have a definite value as primary analgesics or as co-analgesics in combination with opioids.

Acute pain is an alarm signal to alert the organism to noxious tissue damage, irritation, or injury. Ongoing, persistent or chronic pain may

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_269.pub2

lose this function and may become a mere burden to the patient. Pain does not only indicate physical suffering, but involves a strong mental and emotional component. According to the definition of the International Association for the Study of Pain "it is an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage".

The nervous system contains an afferent fiber system that is specialized for registration, transmission, and processing of nociceptive information. In the periphery small unmyelinated nerve endings, called nociceptors, are activated by mechanical, thermal, or chemical stimuli. One type, the mechano-heat nociceptor is specialized for thermal and mechanical stimulation, whereas the so-called polymodal receptor is less specific and responds to all kinds of mechanical, thermal, or chemical irritation. Mechano-heat nociceptors have a high stimulation threshold and respond only to strong stimuli. They are linked to small myelinated and rapidly conducting Aδ fibers which produce a sharp and well located pain sensation. The sharp pain stimulus activates withdrawal reflexes and has to protect the body against further damage. The polymodal receptors respond to weaker stimuli including endogenous compounds like histamine, bradykinin, and prostaglandins. These receptors are mediators of pain and inflammation and are released by tissue damage. The impulses of polymodal nociceptors are transmitted more slowly via thicker, unmyelinated C-fibers and provoke a dull, aching, and less precisely localized pain sensation. The nerve fibers originating from nociceptors are named primary afferent fibers. They have their cell body in the dorsal root ganglia and project to the dorsal root of the spinal cord where they form synapses with fibers of the ascending spino-thalamic pathway. Most of the spino-thalamic pain fibers pass the thalamus and project to the limbic system and the cortex. The limbic system is responsible for the emotional component of pain whereas in the cortex realization and localization of the origin of pain takes place. The spino-thalamic fibers form connections with the formatio reticularis of the brain stem, inducing the various vegetative reactions associated with the pain perception such as sweating, palpitation, nausea, and blood pressure increase.

Inhibitory control of the pain process can take place in the periphery at the level of the nociceptors or the primary afferents. These are the regions where local anesthetics and NSAIDs act. Central pain inhibition can take place at the level of the spinal cord, the thalamus, cortex, and the limbic system. The pain transmission at the level of the spinal cord is under inhibitory control of descending fibers, which originate in different parts of brain stem, mid brain, and cortex. They contain opioid synapses as well as noradrenergic and serotoninergic synapses and represent the descending spinal pain inhibitory system. Inhibitory opioid synapses are not only found in the spinal cord but are strategically located along the whole pain pathway. Opioid receptors are found in the cortex, limbic system, and formatio reticularis and this explains why opioids not only inhibit pain perception but also reduce the emotional component of pain suffering and mitigate the vegetative pain symptoms.

According to the clinical situation, pain can be classified as acute and chronic pain. *Acute pain* is induced by trauma, tissue damage, or disease and has a well defined localization and time course. Acute pain normally responds well to NSAIDs and opioid analgesics. *Chronic pain* is more complex in nature and often contains an inflammatory component. The association with trauma and lesion can be lost and then pain no longer fulfills its alarming function. Chronic pain is often a heavy burden to the patient and induces physical, psychological, and social deterioration. Chronic pain treatment is an important medical challenge.

Physiologically, pain can be differentiated into nociceptive and neuropathic pain. Nociceptive pain results from activation of nociceptors by acute or chronic stimulation and is not associated with damage of nerves. Neuropathic or neurogenic pain in contrast is the result of damage or dysfunction of the peripheral nerves or of the central parts of the pain processing system. Examples of neuropathic pain of peripheral origin are deafferentiation pain (stump pain), diabetic neuropathy, post herpetic neuralgia, trigeminal neuralgia, and sympathetically maintained pain like the complex regional pain syndrome. Central neuropathic pain may result from lesions of the spinal cord or brain lesions following stroke. In chronic cancer pain, nociceptive pain is often associated with a neuropathic component in-

duced by neoplastic nerve infiltration or compression of the nerve by tumor growth.

Cancer pain is one of the most important forms of chronic severe pain and guidelines for the treatment of cancer pain were published by the WHO in 1986 (Fig. 1).

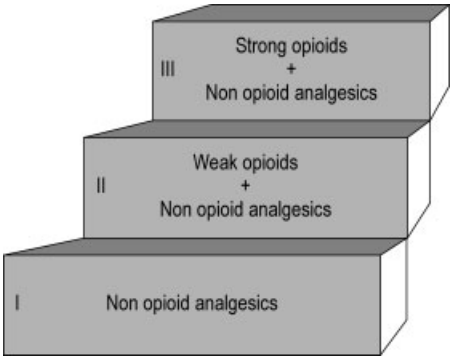

Figure 1. WHO-Guidelines for Cancer pain treatment ("WHO-Ladder")

Meanwhile these guidelines have become the standard not only for treatment of chronic malignant pain but also for benign chronic and acute pain. The WHO proposal is often described as "treatment by mouth, by the clock, and by the ladder". That means that for repeated application of analgesics the oral route should be used. The application interval should be regular and selection of compounds should follow the increase of potency and efficacy as indicated in the three-step analgesic ladder. The first step starts with the single use of a non-opioid analgesic. If pain control is insufficient, a weak opioid may be added. If this is not effective enough, the weak opioid should be replaced by a strong opioid and the non-opioid may be omitted. Each stage of the treatment may be supplemented by the use of co-analgesics and other pharmacological and non-pharmacological treatments to improve pain control and to reduce side effects.

2. Nonsteroidal Anti-inflammatory Drugs (NSAIDs, COX Inhibitors)

In 1899 acetylsalicylic acid, the first nonsteroidal anti-inflammatory drug was registered under the name aspirin. Acetylsalicylic acid is still the best-known nonsteroidal anti-inflammatory drug and is used in almost every household for the treatment of mild to moderate pain states and fever. In 1971, it was shown that the analgesic action of nonsteroidal anti-inflammatory drugs is due to the inhibition of the enzymatic production of prostaglandins. Cyclooxygenase (COX) is the key enzyme in the conversion of arachidonic acid derived from lipids of the cell membrane to prostaglandins and other eicosanoids (Fig. 2).

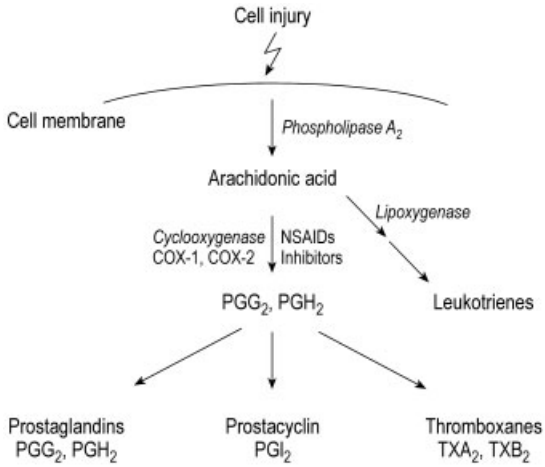

Figure 2. Simple scheme of the biosynthesis of prostaglandins (PG), thromboxanes (TX), and leukotrienes from arachidonic acid consequent to cell injury (modified after [6])

The eicosanoids are important mediators in pain and inflammation leading to hyperalgesia by sensitization of nerve fibers and fever. Furthermore they fulfill important roles in the protection of the gastric mucosa, in platelet aggregation, and maintenance of normal kidney function [11]. The diversity of physiological functions of the eicosanoids is reflected by a variety of different receptors. Five main receptor types have been described, designated DP, FP, IP, TP, and EP, which show the greatest apparent affinity for PGD, PGF, PGI_2, TXA_2, and PGE, respectively [12]. PGE_2, the eicosanoid which plays a key role in pain perception, exerts this and other functions via the four subtypes of the EP receptors, EP1 – 4.

COX exists in two different isoforms, COX-1 and COX-2, characterized by different expression patterns. Both forms are encoded by two genes which have been cloned and sequenced

[13–16]. The cloning of the two isoforms led to further characterization of the enzymes by means of a whole range of molecular biology tools from knock-out mice [17–19] to protein structures [20, 21].

COX-1 is constitutively expressed in many cells of the body and responsible mainly for the production of eicosanoids serving normal physiological functions. One important physiological role is the protection of the gastric mucosa. Pain relief by inhibition of COX-1 is therefore often accompanied with gastrointestinal side effects.

COX-2 expression is induced during inflammation and is thought to be responsible for the production of eicosanoids in inflammatory conditions related to fever and pain. Furthermore, COX-2 is expressed in the central nervous system and might play a direct role in central pain processing.

Many nonsteroidal anti-inflammatory drugs of different chemical structures have been introduced for the treatment of inflammatory and painful conditions. Many years of clinical experience with these drugs have shown that there is no induction of tolerance and dependence and no respiratory depression as seen with opioids. The major side effects of these compounds with COX-1 specificity or balanced COX-1 and COX-2 inhibition are damage of the gastric mucosa, prolongation of bleeding time, and renal failure. The discovery of the inducible isoform led to the identification of drugs that show a stronger inhibition of COX-2 compared to COX-1 (Table 1).

The characteristic differences in expression of the two COX isoforms suggest a potential for new drugs addressing inflammatory and painful conditions specifically via inhibition of COX-2. The analgesic and anti-inflammatory potential of selective COX-2 inhibitors should come without the well known gastrointestinal side effects of classical NSAIDs targeting COX-1 alone or in combination with COX-2.

The first generation of COX-2 inhibitors like meloxicam still shows inhibition of COX-1 in physiological concentrations and hence, still generates gastrointestinal side effects of COX-1 inhibition although to a lower degree. The second generation of COX-2 inhibitors reached the market in 1999 with celecoxib and rofecoxib and shows a more than 100-fold selectivity for COX-2 compared to COX-1 (see also → Prostaglandins, Chap. 3). As expected, the gastrointestinal side effects of these drugs were shown to be clearly lower than those of classical nonsteroidal anti-inflammatory drugs like naproxen.

The ongoing research on the COX isoforms revealed that the strict differentiation of COX-1 as the constitutive enzyme and source of physiological (good) prostaglandins and COX-2 as the inflammation- and pain-mediated induced form of the enzyme, producing the "bad" prostaglandins is no longer valid. It was shown that COX-1 expression is also subject to regulatory processes and can be increased in inflammatory conditions [25]. COX-2 expression on the other hand is increased in the gastric mucosa by inflammatory stimuli and Helicobacter pylori infection [26, 27]. Finally, analysis of COX-1 and COX-2 knock-out mice indicated that both isoforms can contribute to an inflammatory response and have significant roles in the maintenance of physiological homeostasis and carcinogenesis [28]. These data suggest a more complex situation than the initial idea on the roles of COX-1 in normal and of COX-2 in inflammatory conditions, respectively.

Nevertheless, the COX-2 inhibitors showed a clearly improved gastrointestinal side effect profile and up to the year 2000 the newly developed coxibs seemed to displace the classical COX-1 selective NSAIDs from the market. The situation totally changed when the results of ongoing clinical studies and case reports revealed that COX-2 inhibitors seem to increase the risk of adverse thromboembolic events. The case reports came from patients with connective tissue diseases who developed arterial thrombosis after starting celecoxib treatment [29]. An increased number of cardiovascular events as compared to naproxen were seen in a large-scale randomized controlled trial [30], which investigated the efficacy and tolerability of rofecoxib (the VIGOR trial).

The concern about the risk associated with the COX-2 inhibitors increased rapidly and in September 2004 rofecoxib was withdrawn from the market because a trial, designed to test the hypothesis that COX-2 inhibitors would prevent recurrent colonic polyps, indicated an increased cardiovascular toxicity [31]. A similar trial with celecoxib, performed at the National Cancer Institute was stopped when an indepen-

Table 1. COX-2 specific inhibitors

Group	Compound	COX-2 selectivity	Reference
First generation		(< 100-fold)	
	etodolac	10	[22]
	meloxicam	10	[22, 23]
	nimesulide	5 – 100	[22, 23]
Second generation		(100 – 1000-fold)	
	celecoxib	375	[24]
	rofecoxib	800	[561]

dent panel of cardiovascular experts reviewed the data and also found a greater incidence of cardiovascular events among patients treated with celecoxib [32]. Negative data on cardiovascular tolerability were also seen with valdecoxib and its intravenous prodrug, parecoxib, given for pain treatment to patients recovering from coronary-artery bypass surgery [33]. In the beginning, the mechanism of the cardiovascular toxicity of the COX-2 inhibitors was unclear, but it emerged that it was a group effect of the coxibs and the most likely mechanism is a reduced production of the antithrombotic eicosanoid prostacyclin without changing the production of the prothrombotic compound thromboxane [34].

Besides the cardiovascular intolerability, the COX-2 inhibitor valdecoxib showed an increased risk of serious and potentially fatal skin reactions which was the reason for the suspension of marketing autorization. The remaining COX-2 inhibitors, etoricoxib and lumiracoxib, although having in principle a similar cardiovascular and skin tolerability risk, seem to be better tolerated and are still on the market, but doctors are advised to use COX-2 inhibitors only after careful consideration, and to treat their patients with the shortest course and the lowest possible dose of these compounds.

The unexpected tolerability problems with the COX-2 inhibitors have strongly reduced the enthusiasm originally evoked by these compounds and have re-established the therapeutic reputation of the COX-1 inhibitors and of compounds with a balanced COX-1 and COX-2 inhibition, although the gastrointestinal side effects remain a serious problem. Therefore there is increased need for new anti-inflammatory principles with better gastrointestinal tolerability and research activities have been focussed on new compounds combining COX-inhibition with other analgesic and anti-inflammatory principles such as inhibition of 5-lipoxygenase (5-LO) or release of nitric oxide (NO).

Dual COX and 5-LO Inhibitors. This new class of compounds is reported to have COX-2 and 5-lipoxygenase inhibiting properties [35]. The compounds were effective in several animal models of arthritis and were devoid of ulcerogenic properties. One of these compounds, S-2474, ((E)-5-(3,5-di-tert-butyl-4-hydroxybenzylidene)-2-ethyl-1,2-isothiazolidine-1,1-dioxide), which has a γ-sultam skeleton, additionally inhibits the production of interleukin-2 (IL-2) in in vitro assays. S-2474 was selected as development candidate and is now under clinical investigation [36].

ML-3000 ((2,2-dimethyl-6-(4-chlorophenyl)-7-phenyl-2,3-dihydro-1H-pyrrolizine-5-yl)-acetic acid) is a nonoxidant dual cyclooxygenase and 5-lipoxygenase inhibitor and has been compared with indometacin in a number of experimental models of inflammation and pain [37]. Corresponding to animal data the compound showed a wide range of activities in Phase II clinical studies, including anti-inflammatory, analgesic, antiplatelet and antiasthmatic properties [38, 39].

NO Releasing COX Inhibitors. Nitric oxide (NO) releasing non-steroidal anti-inflammatory drugs are a new and promising goup of analgesic and anti-inflammatory compounds. NO release is found as a mechanistically not yet resolved property of a few new COX-1 inhibitors such as amtolmetin [40] or it can be induced by adding a NO releasing moiety (e.g., nitroxybutyl) to existing COX-inhibitors [41]. The most advanced compounds are NO-naproxen and NO-flurbiprofen (NCX-2216).

Preclinical and clinical studies confirm that both NO-release and COX-1 or COX-2 inhibi-

tion take place in vivo, causing less adverse effects on the gastrointestinal tract than conventional NSAIDs and coxibs. In addition, the compounds reduce systemic blood pressure. Possible indications for NO-flurbiprofen are, besides pain inhibition and anti-inflammation, urinary incontinence, Alzheimer's disease, osteoporosis and Paget's disease [42]. A topical formulation is under development for the treatment of dermatological disorders such as contact urticaria.

A further group of NO-donating drugs (e.g., NCX-4016) has been obtained by coupling NO to aspirin. These compounds retain the analgesic, anti-inflammatory and antithrombotic properties of aspirin but show an increased gastrointestinal tolerability [43]. Animal studies have shown that NO-aspirins and the NO-donating NSAIDs maintain gastric mucosal blood flow and reduce leucocyte-endothelial cell adherence.

Up to 2006, no NO-COX inhibitor compound has yet reached the market.

Aspirin as an Antithrombotic Drug. Aspirin, the archetype of the COX inhibitors, is still one of the most interesting compounds of this group and has reached additional importance in the last decades due to its inhibition of platelet aggregation. It is used in low dose [44] in the treatment of acute and the prevention of chronic myocardial infarction and stroke (→ Cardiovascular Drugs).

Classification of COX Inhibitors. COX inhibitors, as it is done here, can be classified according to their chemical structure.

Classification According to Inhibition Specificity. A more functional classification is related to the specificity for inhibition of the COX-1 or COX-2 isoforms of the enzyme. The inhibition of the COX isoenzyme is strongly dependent on the test system used. Inhibition of COX can be quantified in recombinant or natural enzyme preparations, cellular systems, isolated human cell populations, such as platelets (COX-1) and white blood cells (COX-2), or in ex vivo stimulated whole blood samples. The more physiological the situation is, the smaller is the selectivity of most of the COX-2 selective inhibitors. The standard test for comparison is considered to be a whole blood assay which mimics in vivo conditions like plasma binding [45]. It is commonly accepted that reasonable variations in COX inhibition is found in different laboratories. Therefore, whenever possible, data of several compounds, generated within the same test situation, should be compared to each other.

Using this classification based on isoenzyme selectivity, there is besides low-dose aspirin no COX-1 selective inhibitor on the market. Instead, most of the classical NSAIDs belong to the group of nonselective COX inhibitors.

- *Selective COX-1 inhibitors* (low-dose aspirin)
- *Nonselective COX inhibitors* (e.g., high-dose aspirin, indometacin, and other NSAIDs)
- *Preferential COX-2 inhibitors* (e.g., meloxicam)
- *Highly selective COX-2 inhibitors* (e.g., celecoxib and other second- and third-generation COX-2 inhibitors, now called coxibs)

Classification According to Drug – Protein Interactions. Another classification is based on the mode of interaction between the inhibitor drug and the enzyme [46].

- *Irreversible inhibitors of COX-1 (aspirin) and COX-2 (APHS).* Aspirin and APHS [47] acetylate the amino acid serine in the catalytic center of the enzyme to a stable ester derivative so that the endogenic arachidonic acid is prevented from interaction with the enzyme.
- *Reversible, competitive inhibitors of COX-1 and COX-2.* Inhibitors such as ibuprofen, piroxicam, or mefenamic acid block the enzyme by competing with arachidonic acid for binding at the catalytic center.
- *Slow, time-dependent, reversible inhibitors of COX-1 and COX-2.* Compounds such as indometacin and flurbiprofen act by ionic interaction between the acidic carboxyl group of the inhibitor and the basic arginine residue of the enzyme. This seems to influence the helix D of the protein followed by a significant loss of flexibility of the enzyme protein.
- *Slow, time-dependent, irreversible inhibitors of COX-2.* The coxibs such as celecoxib, rofecoxib and lumiracoxib induce a strong, time-dependent inhibition of COX-2, but show only a minor competitive inhibition of COX-

1, which is of no relevance compared to the COX-2 inhibition.

COX Inhibitors in Clinical Use

2.1. Salicylates

Acetylsalicylic acid [*50-78-2*], aspirin, 2-acetoxybenzoic acid, $C_9H_8O_4$, M_r 180.04, *mp* 135 °C (for properties and synthesis, see → Salicylic Acid).

Synthesis [48, 49]:

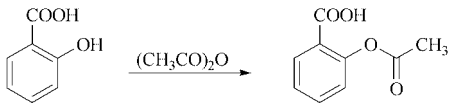

Acetylsalicylic acid

Clinical use: Acetylsalicylic acid is the prototype of a nonsteroidal anti-inflammatory drug and is used in a large number of inflammatory and pain indications including musculoskeletal, soft-tissue and joint disorders, headache, dysmenorrhoea, and fever [50]. In addition acetylsalicylic acid is used as an antiplatelet drug for the prevention of myocardial infarction and stroke [44] and in combination with thrombolytics in the acute treatment of myocardial infarction.

Depending on the assay system, acetylsalicylic acid shows a balanced inhibition of COX-1 and COX-2 [23] or a selectivity towards COX-1 [22]. In addition to COX inhibition acetylsalicylic acid modulates the activities of several cellular kinases [51], which may contribute to its anti-inflammatory effects.

The inhibition of COX-1 and COX-2 by acetylsalicylic acid does not follow a competitive mechanism like with other nonsteroidal anti-inflammatory drugs but rather is due to a covalent enzyme inhibition via acetylation. After absorption, acetylsalicylic acid is hydrolyzed to salicylate which itself still shows some COX inhibition activity. Both compounds are bound to plasma proteins to 80–90 %. The plasma elimination half-life is about 15 min for acetylsalicylic acid and between 3 and 22 h for salicylate, depending on the dose [52]. Acetylsalicylic acid is usually administered orally (0.5–8 g/d) for pain and inflammation and for antiplatelet therapy (75–100 mg/d). It is also available in rectal and topical formulations and as soluble lysine derivative for intravenous or intramuscular application. Acetylsalicylic acid is often used in multi-drug preparations. The main side effects are gastrointestinal disorders. The use in children is limited due to the risk of Reye's syndrome [53]. The lithium, magnesium, calcium, and aluminum salts of acetylsalicylic acid are used in some special preparations.

Trade names: Aspirine (France), Aspirin, Aspisol (Germany), Aspro (UK), Alka Seltzer (USA).

Diflunisal [*22494-42-4*], 2′4′-difluoro-4-hydroxybiphenyl-3-carboxylic acid, $C_{13}H_8F_2O_3$, M_r 250.20, *mp* 210–211 °C (also reported as 212–213 °C).

Synthesis [48, 54–58, 505]: The diazotation of 2,4-difluoroaniline with isoamyl nitrite and condensation with anisole gives 4-(2,4-difluorophenyl)anisole, which is hydrolyzed with HI in refluxing acetic acid yielding 4-(2,4-difluorophenyl)phenol. Finally this compound is carboxylated with K_2CO_3 and CO_2 at 175 °C and 90 bar, followed by an acidification.

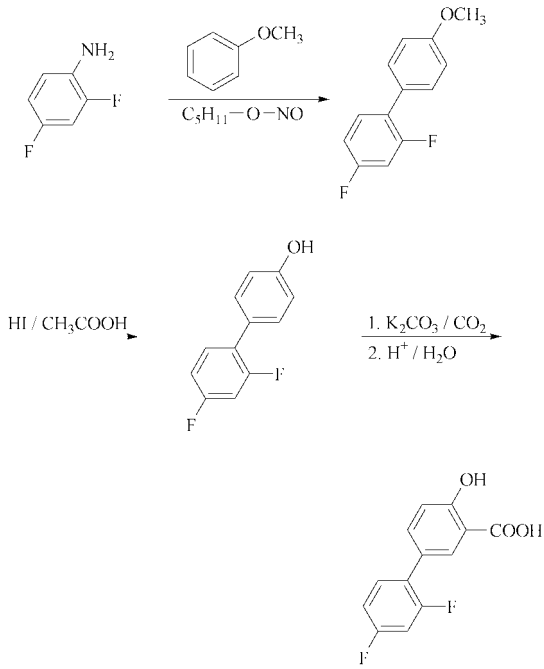

Diflunisal

Clinical use: Diflunisal [60] is a nonsteroidal anti-inflammatory drug used in the treatment of mild to moderate pain including osteoarthritis, rheumatoid arthritis, and primary dysmenorrhoea. It is used as base or lysine- or arginine-salt for oral or parenteral application.

Diflunisal is a weak inhibitor of both COX-1 and COX-2 [61].

Peak plasma concentrations are reached within 2 to 3 h after oral dosing. Diflunisal is strongly bound to plasma protein (> 99 %), has a long elimination half-life (8–12 h) and nonlinear elimination kinetics. Hence, it is used with an initial loading dose (1000 mg) and a lower maintenance dose (500–1000 mg/d).

The main side effects are gastrointestinal disturbances, headache and rash.

Trade names: Dolobis (France), Dolobid (Italy, UK, USA).

Ethenzamide [*938-73-8*], 2-ethoxybenzamide, $C_9H_{11}NO_2$, M_r 165.19, *mp* 132–134 °C.

Synthesis: Salicylamide is ethylated with diethyl sulfate [48, 62].

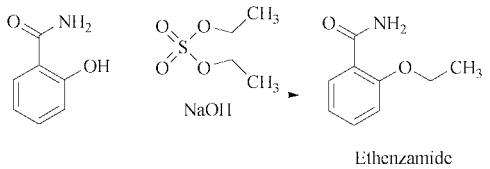

Clinical use: Ethenzamide is a nonsteroidal anti-inflammatory drug used mainly in combination with other ingredients for the treatment of mild to moderate pain including musculoskeletal and joint disorders.

Trade names: Trancalgyl (France).

Salicylamide [*65-45-2*], 2-hydroxybenzamide, $C_7H_7NO_2$, M_r 137.14, *mp* 140 °C.

Synthesis [48, 63]:

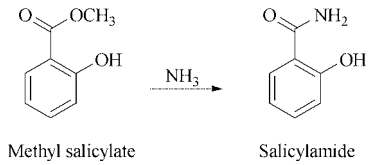

Clinical use: Salicylamide shows analgesic and antipyretic efficacy and was used in multidrug combinations for the treatment of a variety of mild pain conditions including musculoskeletal, soft-tissue, and joint disorders. It is now mostly substituted by newer COX inhibitors.

Trade names: Percutalgine (France), Intralgin (UK), Anabar (USA).

2.2. para-Aminophenol Derivatives

Paracetamol [*103-90-2*], acetaminophen, *N*-(4-hydroxyphenyl)acetamide, $C_8H_9NO_2$, M_r 151.16, *mp* 168–169 °C.

Synthesis:

a) classical route [48, 64]:

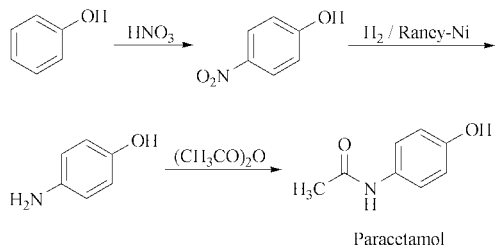

b) Hoechst-Celanese process:

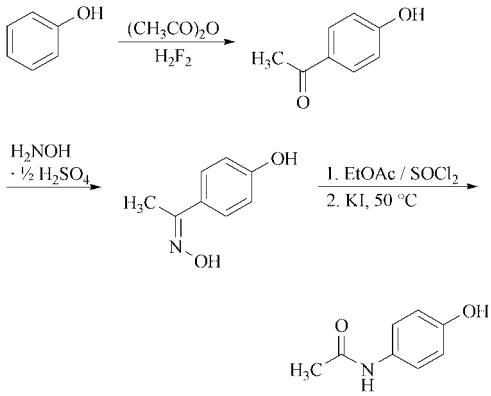

Clinical use [65]: Paracetamol has analgesic and antipyretic properties, but no relevant anti-inflammatory action. It is used for the treatment of various mild to moderate pain conditions and to reduce fever. Despite a long therapeutic use, the mechanism of action of paracetamol is unclear. The compound shows some weak inhibition of COX-1 and COX-2,

and there are speculations that a third COX isoenzyme, named COX-3, may be inhibited [66]. Recently, a new and unexpected mechanism through which paracetamol could exert its analgesic effect was detected [67, 68]. Following deacetylation to its metabolite p-aminophenol, paracetamol is conjugated with arachidonic acid to give a compound named AM404. The metabolite AM404 affects several important targets in the nervous system involved in pain. AM404 inhibits purified COX-1 and COX-2 and prostaglandin formation in lipopolysaccharide-stimulated macrophages. AM404 is also a potent activator of the ion channel $TRPV_1$ (transient receptor potential cation channel, former classified as the vanilloid receptor VR_1) and has an effect on cannabinoid CB_1 receptors. Both $TRPV_1$ and CB_1 receptors are involved in pain and in thermoregulatory pathways.

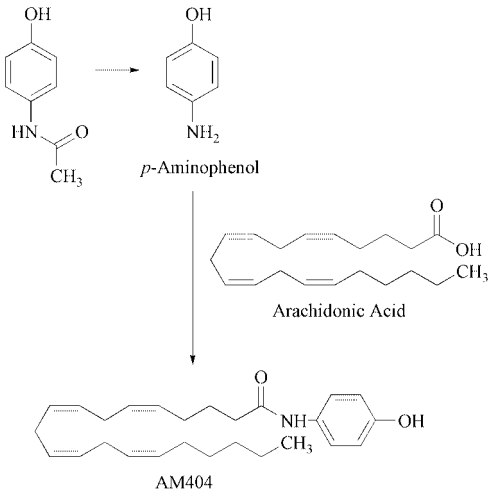

Paracetamol is one of the most popular analgesics as single drug or in multi-ingredient preparations, often in combination with NSAIDs or weak opioids. It is used orally or rectally as suppositories, the oral dose range is 500–1000 mg every 4–5 h up to 4 g daily. Side effects are rare and may include hematological reactions, leucopenia, agranulocytosis and other hypersensitivity reactions. Paracetamol has a narrow therapeutic dose range and overdosage induces severe liver and renal damage [69] via accumulation of a toxic metabolite, *N*-acetylbenzoquinoneimine (NABQI). Acetylcysteine or methionine, which increase glutathione conjugation of the metabolite, are used as antidote.

Paracetamol is not soluble in aqueous solutions and cannot be given parenterally. A soluble glycine prodrug derivative has been developed for parenteral use (propacetamol, see below).

Trade names: Benuron (Germany, Switzerland), Dafalgan (France) Tylenol (Austria, Canada, USA, Spain), Alvedon (Norway, Sweden, UK).

Combination with codeine: Co-Tylenol (USA).

Combination with tramadol: Zaldiar (Germany, Austria, France, Spain, Switzerland), Tramacet (UK), Ultracet (USA).

Paracetamol is also used in many generic formulations.

Propacetamol [*66532-85-2*], 4-(acetylamino)phenyl *N*,*N*-diethylglycinate, $C_{14}H_2ON_2O_3$, M_r 264.33, is a soluble glycine prodrug derivative of paracetamol. It is used as hydrochloride ([*66532-86-3*], $C_{14}H_2ON_2O_3 \cdot HCl$, M_r 300.79), for intramuscular or intravenous application and is rapidly metabolized to free paracetamol [70].

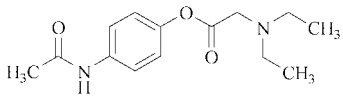

Trade name: Pro-Dafalgan (Belgium, France), Pro-Efferalgan (Italy)

Phenidine [*62-44-2*], phenacetin, *N*-(4-ethoxyphenyl)acetamide, $C_{10}H_{13}NO_2$, M_r 179.22, *mp* 134–135 °C.

Synthesis [48]:

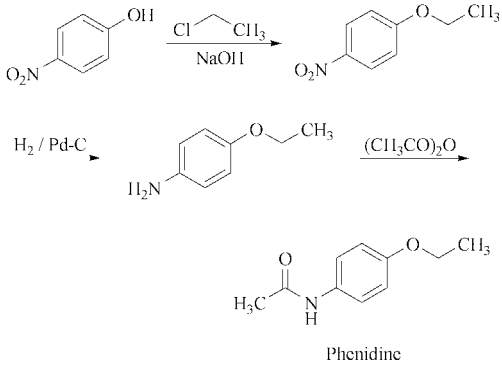

Clinical use: Phenidine is a weak analgesic, antipyretic compound without anti-inflammatory action. It has been used in combination with other compounds like aspirin, caffeine, or codeine, but due to hematological and nephrotoxic side effects [71] has been withdrawn from the market and substituted by the less toxic paracetamol.

2.3. Anthranilates

Etofenamate [*30544-47-9*], 2-(2-hydroxyethoxy)ethyl 2-(3-trifluoromethylphenylamino) benzoate, $C_{18}H_{18}F_3NO_4$, M_r 369.34, pale yellow viscous oil, thermolabile at 180 °C, *bp* (133.32 Pa) 130–135 °C.

Synthesis: Esterification of the potassium salt of flufenamic acid (see page 10) with 2-(2-chloroethoxy)ethanol in dimethyl formamide as solvent yields etofenamate [48, 72].

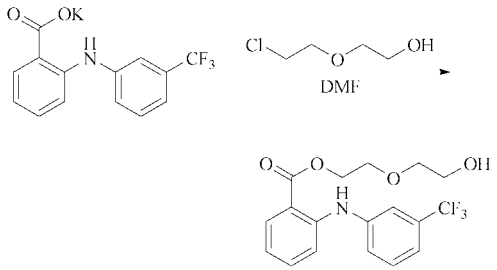

Etofenamate

Clinical use: Etofenamate [73] is a nonsteroidal anti-inflammatory drug which was used mainly as a topical formulation (500–1300 mg/d) and by intramuscular injection (1 g/d) for the treatment of joint, musculoskeletal and soft-tissue disorders.

Flufenamic acid [*530-78-9*], flufenamate, 2-[3-(trifluoromethyl)phenyl]aminobenzoic acid, $C_{14}H_{10}F_3NO_2$, M_r 281.23, *mp* 124-125 °C (also reported as 134-136 °C); aluminum salt (3 : 1) [*16449-54-0*], $C_{42}H_{27}AlF_9N_3O_6$, M_r 867.66.

Synthesis: 2-Chlorobenzoic acid is reacted with 3-trifluoromethylaniline in the presence of copper and potassium carbonate [48, 74, 75].

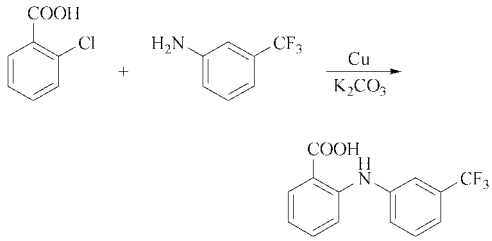

Flufenamic acid

Clinical use: Flufenamate is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain of musculoskeletal, joint or soft-tissue origin. It was marketed in a variety of topical formulations alone or in combination with other ingredients.

Flufenamate is not recommended in patients with acute porphyria and was associated with acute proctocolitis [76].

In addition to COX inhibition flufenamate like other fenamates modifies several ion channel functions, e.g., inhibition of nonselective cation conductance [77], calcium-activated chloride channels [78], voltage-gated calcium channels and potassium channels [79, 80] and induces blocking of gap junctions [81]. The relevance of these activities for the analgesic and anti-inflammatory potential of fenamates is unknown.

Meclofenamic acid [*644-62-2*], meclofenamate, 2-[(2,6-dichloro-3-methylphenyl)amino]benzoic acid, $C_{14}H_{11}Cl_2NO_2$, M_r 296.15, *mp* 257–259 °C; monosodium salt monohydrate [*6385-02-0*], $C_{14}H_{10}Cl_2NNaO_2 \cdot H_2O$, M_r 336.15, *mp* 289–291 °C.

Synthesis: Meclofenamic acid is obtained by condensation of 2-bromobenzoic acid with 2,6-dichloro-3-methylaniline using $CuBr_2$ in diethyleneglycol dimethyl ether containing *N*-ethylmorpholine, and heating at 145–155 °C [48, 82–84] .

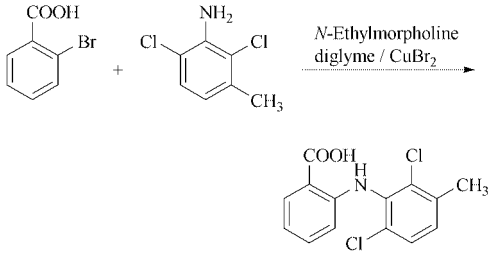

Meclofenamic acid

Clinical use: Meclofenamate [85] is a non-steroidal anti-inflammatory drug used for the treatment of mild to moderate pain, musculoskeletal, and joint disorders like rheumatoid arthritis and osteoarthritis as well as dysmenorrhoea. Meclofenamate inhibits COX-1 and COX-2 and modifies ion channels [86, 87] (see page 10).

Meclofenamate is administered orally (300–400 mg/d).

Trade names: Lenidolor (Italy), Meclomen (Austria, Spain, Switzerland, USA).

Mefenamic acid [*61-68-7*], 2-[(2,3-dimethylphenyl)amino]benzoic acid, $C_{15}H_{15}NO_2$, M_r 241.29, *mp* 230–231 °C; monosodium salt $C_{15}H_{14}NNaO_2$, M_r 263.27.

Synthesis [48, 88]:

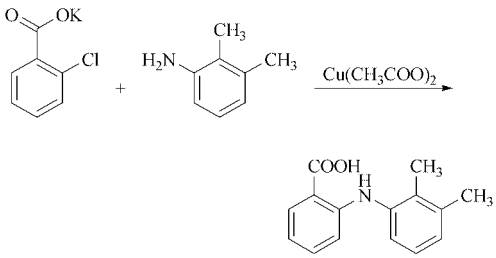

Mefenamic acid

Clinical use: Mefenamic acid is a non-steroidal anti-inflammatory drug which is used for the treatment of mild to moderate pain conditions, musculoskeletal and joint disorders like rheumatoid arthritis and osteoarthritis, and dysmenorrhoea. [89]. Mefenamic acid inhibits both COX isoforms with some preference for COX-2 and modifies ion channels [90] (see page 10).

Mefenamic acid is given orally (1500 mg/d maximal dose). The main side effects concern the gastrointestinal system and include diarrhea [91].

Trade names: Ponstyl (France), Ponstan (UK), Ponstel (USA).

Niflumic acid [*4394-00-7*], 2-(3-trifluoromethylphenylamino)nicotinic acid, 2-[[3-(trifluoromethyl)phenyl]amino]-3-pyridinecarboxylic acid, $C_{13}H_9F_3N_2O_2$, M_r 282.22, *mp* 204 °C.

Synthesis: Condensation of 2-chloronicotinic acid with 3-trifluoromethylaniline or reaction of 2-aminonicotinic acid with 1-bromo-3-trifluoromethylbenzene yields niflumic acid [48, 92].

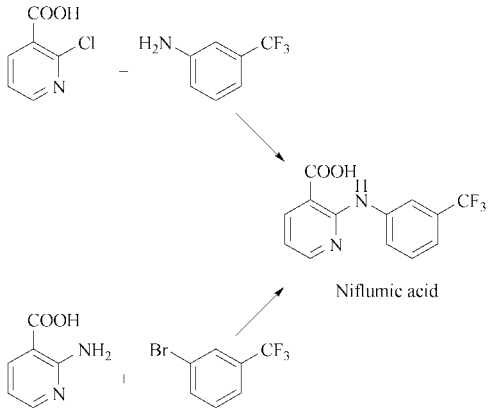

Niflumic acid

Clinical use: Niflumic acid [93] is a non-steroidal anti-inflammatory drug used for the treatment of inflammation and pain in musculoskeletal and joint disorders like rheumatoid arthritis as well as traumatic and postoperative pain. Niflumic acid is used in oral, rectal, or topical applications (up to 750 mg/d).

The morpholinoethyl ester morniflumate, which is used in topical formulations, inhibits both cyclooxygenase and 5-lipoxygenase which suggests an additional component of anti-inflammatory activity [94].

Trade names: Niflurid, Niflugel (Belgium, France, Switzerland), Actol (Austria, Sweden).

Tolfenamic acid [*13710-19-5*], 2-[(3-Chloro-2-methylphenyl)amino]benzoic acid, *N*-(3-chloro-*o*-tolyl)anthranilic acid; *N*-(2-methyl-3-chlorophenyl)anthranilic acid, $C_{14}H_{12}ClNO_2$, M_r 261.70, *mp* 207.0–207.5 °C.

Synthesis: Tolfenamic acid is obtained by condensation of 2-chlorobenzoic acid with 3-chloro-2-methyl-phenylamine using $CuBr_2$ in diethylenglycol dimethyl ether [95, 96].

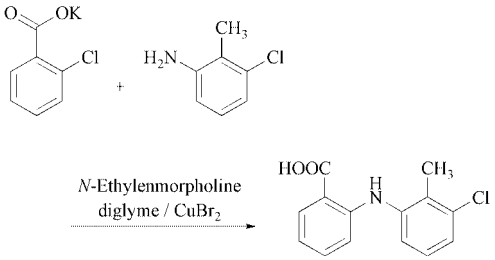

Tolfenamic acid

Clinical use: Tolfenamic acid is a non-steroidal anti-inflammatory drug [97] used for the treatment of rheumatoid arthritis [98], osteoarthritis, ankylosing spondylitis, trauma pain, and dysmenorrhoea [99]. It is also used for the prophylactic and acute treatment of migraine [100]. The compound is further marketed for veterinary use.

Trade name: Clotam (Denmark).

2.4. Arylacetic Acids

Acemetacin [*53164-05-9*], carboxymethyl [1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indol-3-yl]-acetate, $C_{21}H_{18}ClNO_6$, M_r 415.82, *mp* 150–153 °C (fine pale yellow crystals).

Synthesis [48, 101] (see on top, right column).

In addition, the reaction product of indomethacin (see page 14) with benzyl bromoacetate can be hydrogenated to acemetacin.

Clinical use: Acemetacin [102] is a non-steroidal anti-inflammatory drug acting directly and via its major metabolite indomethacin. Acemetacin is used in chronic joint pain as well as in postoperative pain. Oral dosing is recommended between 120 and 360 mg daily.

Trade names: Rantudil (Germany), Emflex (UK).

Acemetacin is also used in several generic formulations.

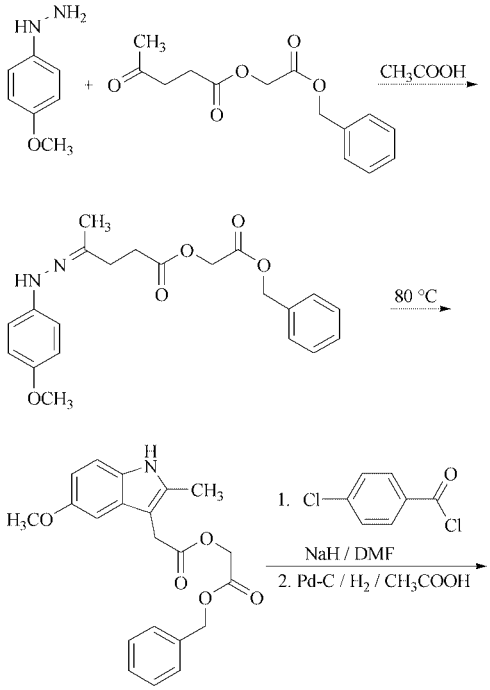

Acemetacin

Proglumetacin [*57132-53-3*], 1-(4-Chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indole-3-acetic acid 2-[4-[3[4-(benzoylamino)-5-(dipropylamino)-1,5-dioxopentyl]oxy]propyl]-1-piperanzinyl]ethyl ester, *rac-N*-[2-[1-(*p*-chlorobenzoyl)-5-methoxy-2-methyl-3-indolylacetoxy]ethyl]-*N′*-[3-(*N*-benzoyl-*N′*,*N′*-di-*n*-propyl-DL-isoglutaminoyl)oxypropyl]piperazine, $C_{46}H_{58}ClN_5O_8$, *Mr* 844.43, dimaleate, 4-Benzamido-5-(dipropylamino)-5-oxopentanoic acid 3-[4-[2-[2-[1-(chlorobenzoyl)-5-methoxy-2-methyl-1 *H*-indol-3-yl]acetoxy]ethyl]piperazin-1-yl]propyl ester dimaleate [*59209-40-4*], $C_{46}H_{58}ClN_5O_8$ $2C_4H_4O_4$, M_r 1076.58, *mp* 146–148°C.

Synthesis: The reaction of *N*-benzoyl-*N′*, *N′*-di-*n*-propyl-DL-isoglutamin (proglumide)

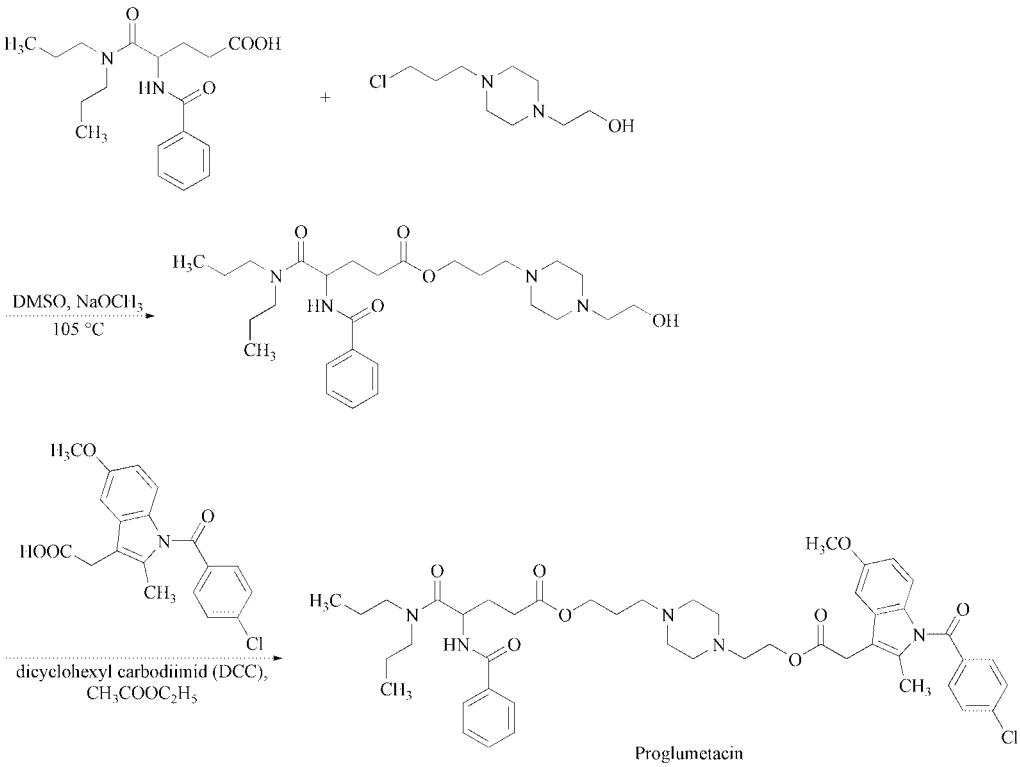

with 1-(3-chloropropyl)-4-(2-hydroxyethyl)-piperazine obtained from 1-(2-hydroxyethyl)piperazine and 1-bromo-3-chloropropane by means of sodium methanolate in DMSO at 105°C gives *N*-[3-(*N*-benzoyl-*N′*, *N′*-di-*n*-propyl-DL-isoglutaminyl)oxypropyl]-*N′*-(2-hydroxyethyl)piperazine, which is then condensated with Indomethacin (1-(*p*-chlorobenzoyl)-5-methoxy-2-methylindole-3-acetic acid) by means of dicyclohexylcabodiimide and $NaHCO_3$ in ethyl acetate [103, 104] (see on top).

Proglumetacine is a prodrug of acemetacine [105] and is used for the treatment of osteoarthritis, rheumatoid arthritis, gout, ankylosing spondylitis, and pain [106].

Trade names: Protaxon (Gemany), Proxil (Italy).

Bufexamac [*2438-72-4*], 2-(4-butoxyphenyl)-*N*-hydroxyacetamide, $C_{12}H_{17}NO_3$, M_r 223.12, *mp* 153–155 °C.

Synthesis [48, 107] (see right column).

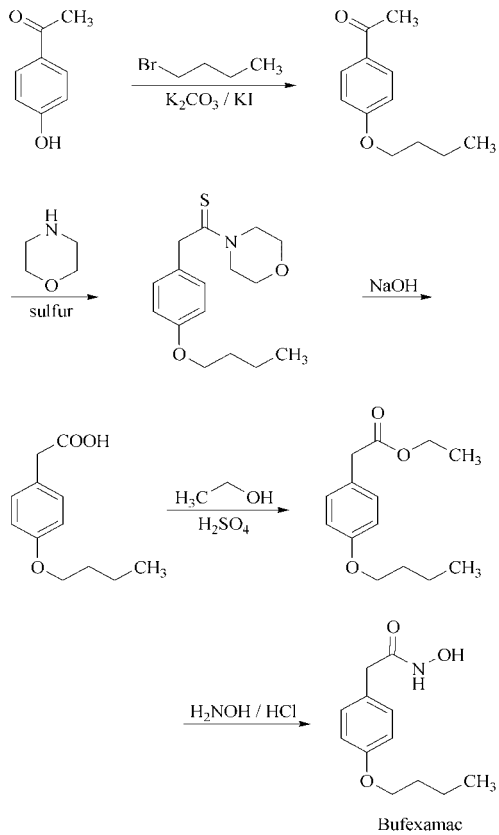

Clinical use: Bufexamac is a nonsteroidal anti-inflammatory drug used in topical formulations for mild skin disorders and as suppositories (250–500 mg/d) for hemorrhoids.

Trade names: Duradermal (Germany), Parfenac (France, Germany).

Diclofenac [*15307-86-5*], [2-(2,6-dichlorophenylamino)phenyl]acetic acid, $C_{14}H_{11}Cl_2NO_2$, M_r 296.15, *mp* 156–158 °C; sodium salt [*15307-79-6*], $C_{14}H_{10}Cl_2NNaO_2$, M_r 318.13, *mp* 283–285 °C.

Synthesis [108]:

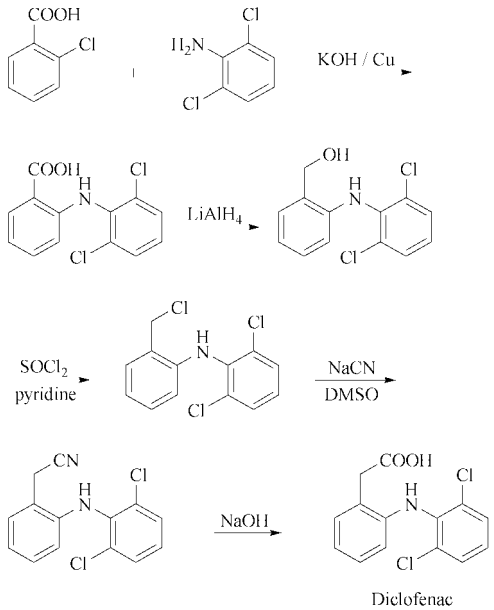

Clinical use: Diclofenac [109] is a nonsteroidal anti-inflammatory drug with balanced COX-1 and COX-2 inhibition [22]. It is commonly used for a variety of inflammatory and pain conditions such as musculoskeletal and joint disorders, periarticular disorders, soft-tissue disorders, renal colic, acute gout, dysmenorrhoea, and postoperative pain.

The plasma protein binding of diclofenac is greater than 99.5 % and the plasma elimination half-life is between 1 and 2 h.

Diclofenac is used mainly as the sodium salt orally or parenterally (75–150 mg/d) and as ophthalmic solution. Topical formulations may contain the diethylammonium or epolamine salt. Diclofenac is combined with misoprostol to reduce gastrointestinal side effects, which are the main adverse events and with weak opioids (codeine) to increase the analgesic potential.

Trade names: Allvoran (Germany), Diclac (Germany), Voltaren (Germany, Italy, USA), Voltarène (France), Voltarol (UK).

Combination with misoprostol: Arthotec (Germany).

Combination with codeine: Combaren.

Diclofenac is also used in many generic formulations.

Indomethacin [*53-86-1*], [1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indol-3-yl]-acetic acid, $C_{19}H_{16}ClNO_4$, M_r 357.79, *mp* 153–154 °C (crystals exhibiting polymorphism, *mp* for another form is 162 °C); sodium trihydrate [*74252-25-8*], $C_{19}H_{15}ClNNaO_4 \cdot 3H_2O$, M_r 433.82, pale yellow crystalline powder.

Synthesis [48, 110] (see next page).

Clinical use: Indomethacin is a nonsteroidal anti-inflammatory drug commonly used for the treatment of mild to moderate pain. Indomethacin is used in acute and chronic pain states like rheumatoid arthritis, osteoarthritis, joint and soft-tissue pain, dental pain, postoperative pain and dysmenorrhoea. Indomethacin is a COX-1 selective inhibitor with 10- to 60-fold selectivity compared to COX-2 [22, 23]. It is administered orally, rectally or topically (50–150 mg/d, maximal daily dose 200 mg) as well as ophthalmic solution.

The major side effects of indomethacin are gastrointestinal irritations and inflammation and central nervous system disturbances like depression, drowsiness, tinnitus, and convulsions. Indomethacin is absorbed after oral application reaching peak plasma concentrations after 2 h. It is bound to about 99 % to plasma proteins and shows a variable terminal half-life from 2.6 to 11.2 h in adults and up to 30 h in neonates.

Indomethacin is available as sodium, meglumine, or L-arginine salt or as prodrug (proglumetacin maleate).

Trade names: Elmetacin, Inflam, Wydora (Germany), Indocid (France, UK), Indocin (USA), Indolgina (Italy), Indocolir eye drops (Germany).

Prodrugs of indomethacin are acemetacin (see Section page 12) and proglumetacin maleate: Protaxon (Germany).

a) Merck & Co. process [111]:

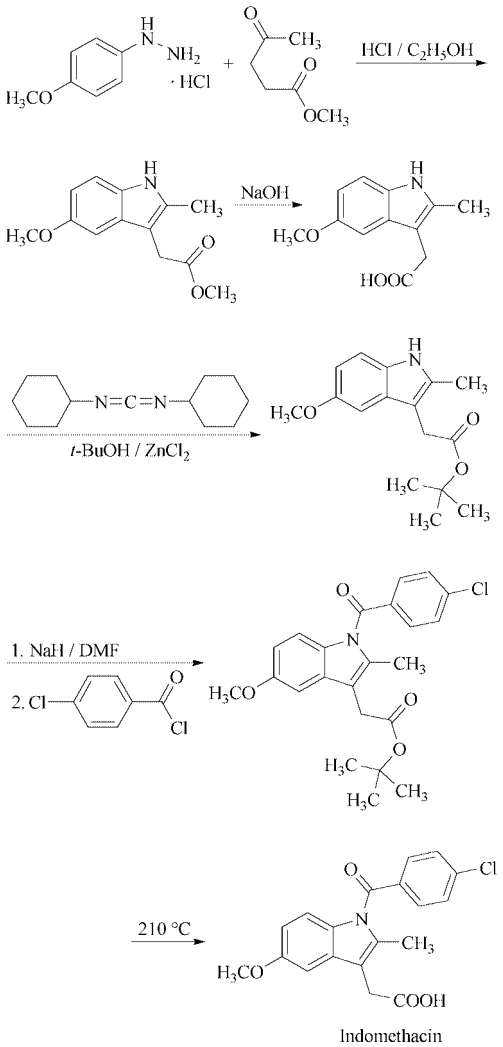

b) Sumitomo process [112]:

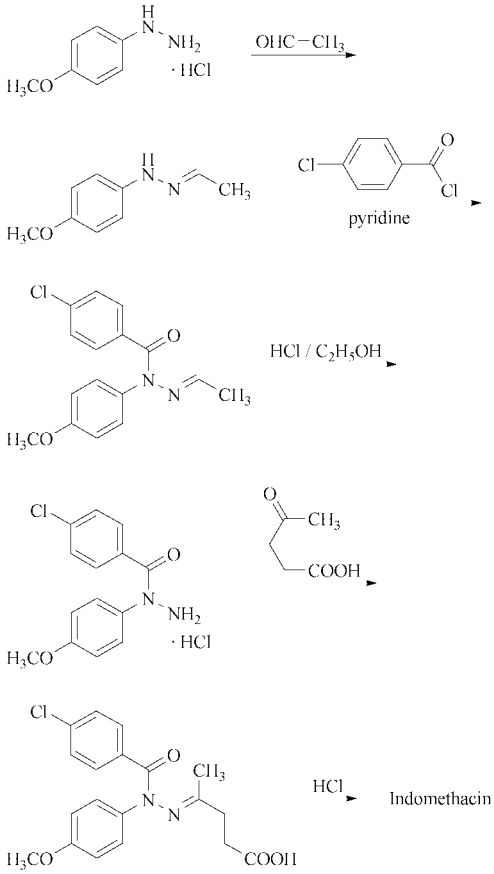

Lonazolac [*53808-88-1*], [3-(4-chlorophenyl-1*H*-pyrazol-4-yl]acetic acid, $C_{17}H_{13}ClN_2O_2$, M_r 312.75, *mp* 150–151 °C; calcium salt (2 : 1) [*75821-71-5*], $C_{34}H_{24}CaCl_2N_4O_4$, M_r 663.57, *mp* 270–290 °C (decomp.).

Synthesis: The pyrazole-4-carbaldehyde synthesized according to the Vilsmeier synthesis is reduced to the alcohol, which is chlorinated. The chloro derivative is reacted with sodium cyanide to give the nitrile, which is hydrolyzed to lonazolac. The calcium salt, slightly soluble in water, is formed by adding calcium chloride to the free acid [48, 113–115].

Clinical use: Lonazolac [85] is a nonsteroidal anti-inflammatory drug used for the treatment of acute inflammatory pain conditions of joint and soft-tissue disorders as well as posttraumatic and postoperative pain. Lonazolac is used as calcium salt and is given orally (600 mg/d, initial dose up to 900 mg/d) or rectally (800 mg/d).

Trade names: Argun (Germany), Irritren (Austria, Belgium, Switzerland).

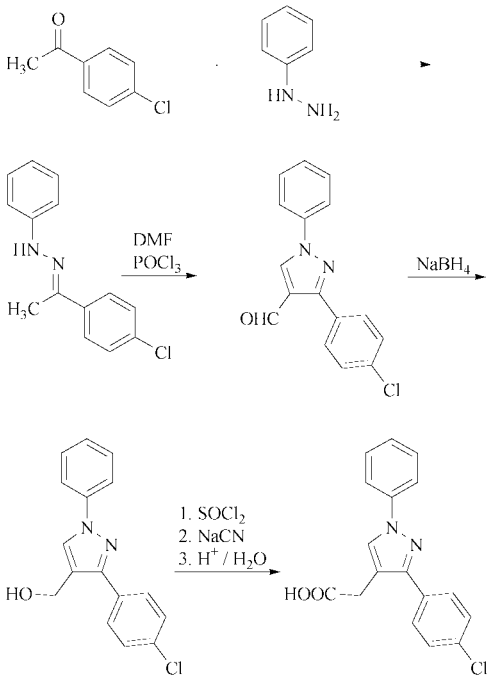

Lonazolac

Sulindac [*38194-50-2*], (*Z*)-5-fluoro-2-methyl-1-[4-(methylsulfinyl)phenyl]methylene-1*H*-indene-3-acetic acid, $C_{20}H_{17}FO_3S$, M_r 356.41, *mp* 182–185 °C (decomp.).

Synthesis [48, 116, 117] (see right column).

Clinical use: Sulindac [118] is a nonsteroidal anti-inflammatory drug used in the treatment of mild to moderate pain including musculoskeletal and joint disorders like rheumatoid arthritis, osteoarthritis, and gout. The main side effects are gastrointestinal disturbances and kidney stones [119]. Sulindac induces no relevant COX inhibition whereas the active metabolite sulindac sulfide inhibits both isoenzymes with some COX-1 preference [120], indicating that the pharmacological activity of sulindac probably results from its sulfide metabolite. Another metabolite, sulindac sulfone, induces apoptosis in tumor cells and sulindac is extensively studied for cancer treatment [121]. Sulindac is administered orally or rectally (200–400 mg/d).

Trade names: Arthrocine (France), Clinoril (Italy, UK, USA).

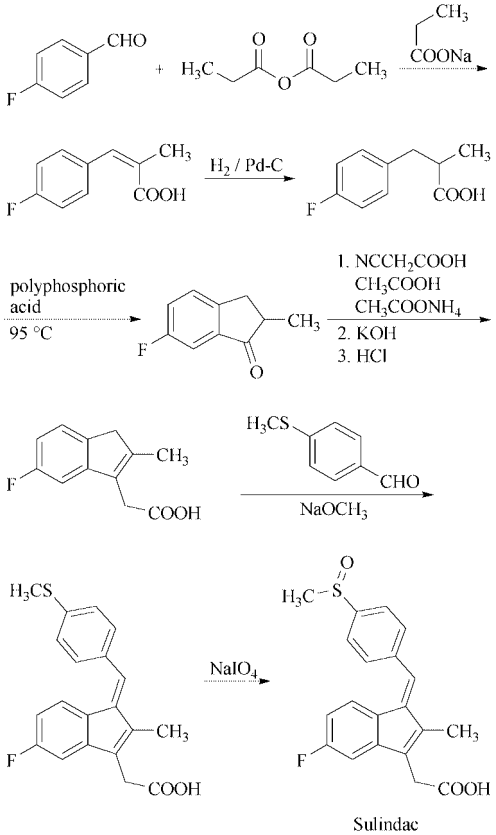

Sulindac

Tolmetin [*26171-23-3*], [1-methyl-5-(4-methylbenzoyl)-1*H*-pyrrol-2-yl]acetic acid, $C_{15}H_{15}NO_3$, M_r 257.28, *mp* 155–157 °C; sodium salt [*35711-34-3*], $C_{15}H_{14}NNaO_3$, M_r 279.27; sodium salt dihydrate [*64490-92-2*], $C_{15}H_{14}NNaO_3 \cdot 2\ H_2O$, M_r 315.30.

Synthesis [48, 122, 123] (see next page, left column).

Clinical use: Tolmetin [124] is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain states in musculoskeletal, soft-tissue, and joint disorders like rheumatoid arthritis, osteoarthritis, and gout as well as juvenile rheumatoid arthritis. Tolmetin inhibits both isoforms of cyclooxygenase with some preference for COX-1 [125]. Tolmetin is administered orally, rectally, or topically (600–1800 mg/d). The peak plasma concentrations are reached within 30 to 60 min. Tolmetin shows a high plasma protein binding of 99 % and a biphasic plasma half-life of 1 to 2 and 5 h, respectively.

Trade name: Tolectin (Austria, Switzerland, UK, USA).

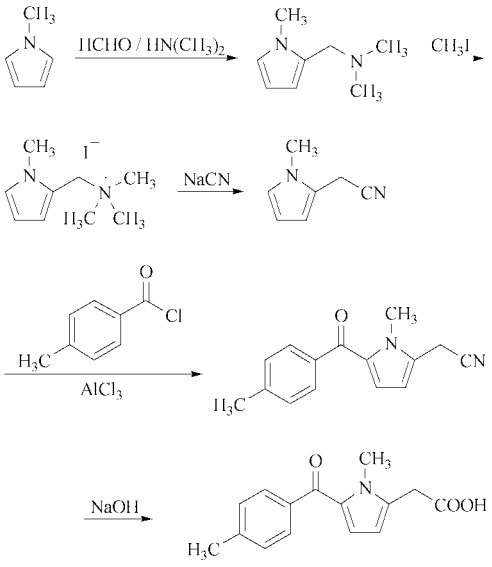

Tolmetin

Amtolmetin guacil (AMG, ester prodrug of tolmetin) [*87344-06-7*], 2-[2-[1-methyl-2-(4-methylbenzoyl)pyrrol-5-yl]acetamido]acetic acid 2-methoxyphenyl ester, 2-[2-[1-methyl-5-(4-methylbenzoyl)pyrrol-2-yl]-acetamido]acetic acid 2-methoxyphenyl ester, N-[(1-methyl-5-p-toluoylpyrrol-2-yl)acetyl]-glycine 2-methoxyphenyl ester, 1-methyl-5-p-toluoylpyrrole-2-acetamidoacetic acid guaicyl ester, $C_{24}H_{24}N_2O_5$, M_r 420.46, *mp* not reported.

Synthesis: The condensation of 1-methyl-5-(4-methylbenzoyl)pyrrole-2-acetic acid with glycine ethyl ester in the presence of carbonyldiimidazole and triethylamine in tetrahydrofuran gives the corresponding ethyl 2-[2-(1-methyl-5-(4-methylbenzoyl)-1*H*-pyrrol-2-yl)acetamido]acetate, which is hydrolyzed with NaOH in tetrahydrofuran – water yielding 2-[2-[1-methyl-5-(4-methylbenzoyl)pyrrol-2-yl]-acetamido]acetic acid. Finally, this compound is esterified with 2-methoxyphenol (guaiacol) in hot tetrahydrofuran with carbonyldiimidazole as catalyst [126].

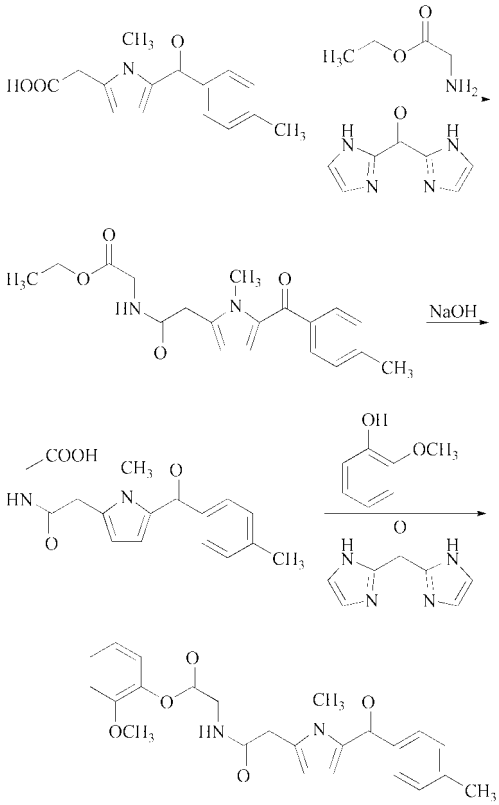

Amtolmetin guacil

Clinical use: Amtolmetin guacil (AMG) is a non-steroidal anti-inflammatory drug developed by Sigma-Tau for the treatment of pain and inflammation. It acts similarly to tolmetin, and in preclinical trials displayed long-lasting anti-inflammatory and analgesic activity. Amtolmetin has a good gastrointestinal tolerability in humans [127] and does not induce gastric ulcers in the rat [128]. The good gastrointestinal tolerability and even a gastroprotective effect was related to an increase of endogenous NO via stimulation of inducible NO synthase in the gastric mucosa [129].

Trade name: Eufans (Italy).

Mofezolac [*78967-07-4*], 3,4-bis(4-methoxyphenyl)-5-isoxazoleacetic acid, 3,4-di-(*p*-methoxyphenyl)isoxazole-5-acetic acid, $C_{19}H_{17}NO_5$, M_r 339.35, *mp* 147.5 °C.

Synthesis:

a) The reaction of deoxyanisoin with hydroxylamine in methanol gives the corresponding 1,2-bis(4-methoxyphenyl)ethanone oxime, which is cyclized with ethyl acetate by means of *n*-butyllithium in tetrahydrofuran yielding 3,4-di(4-methoxyphenyl)-5-methylisoxazole. Finally, this compound is condensed with CO_2 with *n*-butyllithium as catalyst in tetrahydrofuran to yield mofezolac [130, 131]. The synthesis with $ClCO_2C_2H_5$ instead of CO_2 is described in [132, 133].

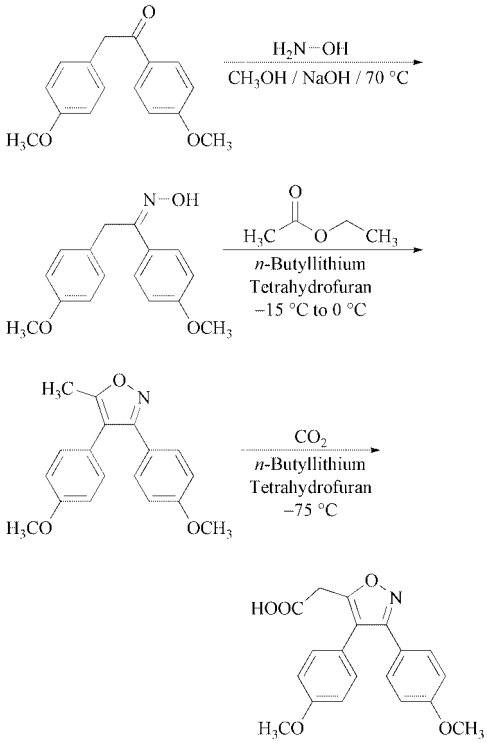

Mofezolac

b) Alternatively the corresponding 1,2-bis(4-methoxyphenyl)ethanone oxime can be condensed with (*E*)-methyl 3-methoxyacrylate with the aid of potassium *tert*-butyrate in *tert*-butanol to yield (*E*)-methyl 4,5-bis(4-methoxyphenyl)-5-oxopent-3-enoate, which is cyclized with hydroxyl amine. Subsequent treatment with oxygen in acetic acid yields mofezolac [132–135].

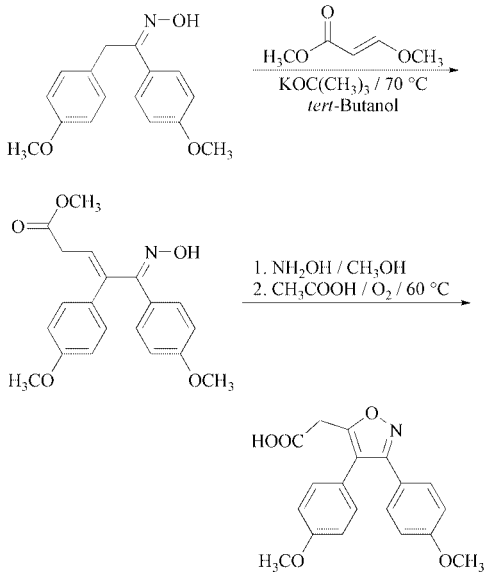

Clinical use: Yoshitomi (now Mitsubishi Pharma) has launched mofezolac, a COX-1 inhibitor [136] with preferential COX-1 inhibition [137] for the treatment of arthritis-related pain in Japan.

The compound is used in oral doses of 75 mg for the treatment of periarthritis, lumbago, neck-shoulder-arm syndrome and pain after surgery, trauma, or dental extraction.

Trade name: Disopain (Japan).

Bromfenac [*91714-94-2*], 2-amino-3-(4-bromobenzoyl)benzeneacetic acid, 2-[2-amino-3-(4-bromobenzoyl)phenyl]acetic acid, $C_{15}H_{12}BrNO_3$, M_r 334.17, monosodium salt [*91714-93-1*], $C_{15}H_{11}BrNNaO_3$, M_r 356.15, *mp* 284–286 °C (dec.), monosodium salt sesquihydrate [*120638-55-3*], $C_{15}H_{11}BrNNaO_3$ 3/2 H_2O, M_r 766.35.

Synthesis: The cyclization of 2-amino-4'-bromobenzophenone with ethyl 2-(methylthio)acetate with *tert*-butyl hypochlorite as catalyst in dichloromethane at 70 °C gives 7-(4-bromobenzoyl)-3-(methylthio)-2,3-dihydro-1*H*-indol-2-one, which is desulfurized by treatment with Raney nickel in THF yielding 7-(4-bromobenzoyl)-2,3-dihydro-1*H*-indol-2-one. Finally, this compound is hydrolyzed with refluxing 3 M aqueous NaOH and acidified with concentrated HCl [138–141].

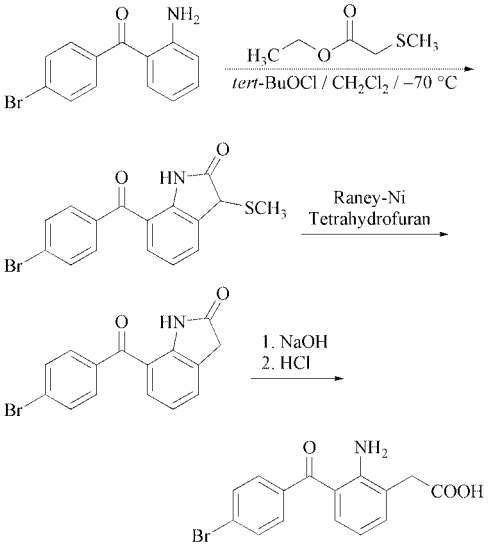

Bromfenac

Clinical use: Bromfenac, used as the sodium salt, is a non-steroidal anti-inflammatory drug acting via inhibition of COX-1 and COX-2. It was launched in Japan in 2000 by Senju for topical treatment of ocular inflammation [142] and has been approved in the United States in 2006 for the treatment of pain following cataract surgery. It was used for the short-term treatment of acute pain, but it was withdrawn for this indication in June 1998 because of several post-marketing reports of severe hepatic failure [143–145].

The drug is rapidly absorbed and excreted. The drug's long duration of anti-inflammatory action despite its short half-life deserves further investigation.

Trade name: Xibrom (Japan, USA).

Nabumetone [*42924-53-8*], 4-(6-methoxy-naphthalen-2-yl)butan-2-one, $C_{15}H_{16}O_2$, M_r 228.29, *mp* 80 °C.

Synthesis: Condensation of 6-methoxy-2-naphthaldehyde with acetone in an aqueous solution of NaOH gives 4-(6-methoxy-2-naphthyl)-3-buten-2-one, which is reduced with H_2 over Pd-C in ethyl acetate [48, 146–149].

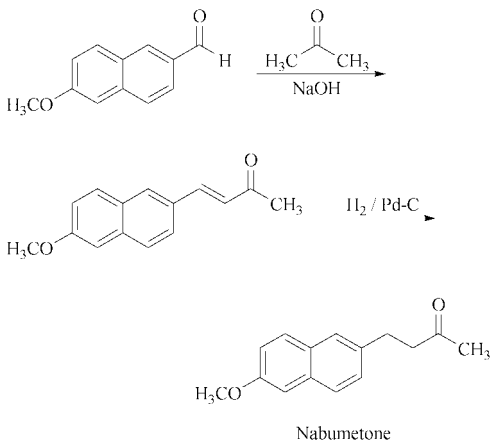

Nabumetone

Several additional ways for the synthesis of nabumetone are described. A short, simple, and economical process for large scale preparation avoiding hazardous chemicals or tedious work up procedures has been published [149].

Clinical use: Nabumetone [150] undergoes rapid first-pass metabolism to the active metabolite 6-methoxy-2-naphthylacetic acid (6-MNA). 6-MNA shows a balanced inhibition of both COX-1 and COX-2 [22]. Nabumetone is used for the treatment of pain and inflammation associated with osteoarthritis and rheumatoid arthritis. It shows a 99 % binding to plasma protein and an elimination half-life of up to 22 h with marked individual differences. Nabumetone is given orally (500–2000 mg/d).

Trade names: Listran (Spain), Nabuser (Italy), Relafen (USA), Relifex (UK).

2.5. Arylpropionic Acids

Flurbiprofen [*5104-49-4*], 2-(3′-fluorobiphenyl-4-yl)propionic acid, 2-fluoro-α-methyl[1,1′-biphenyl]-4-acetic acid, $C_{15}H_{13}FO_2$, M_r 244.26, *mp* 110–111 °C; sodium salt [*56767-76-1*], $C_{15}H_{12}FNaO_2$, M_r 266.25.

Synthesis [48, 151–154]:

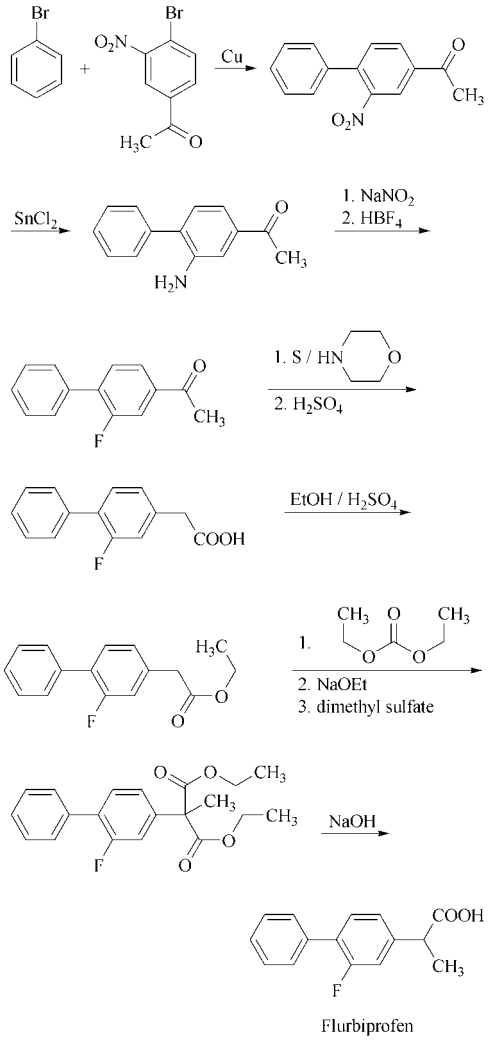

a)

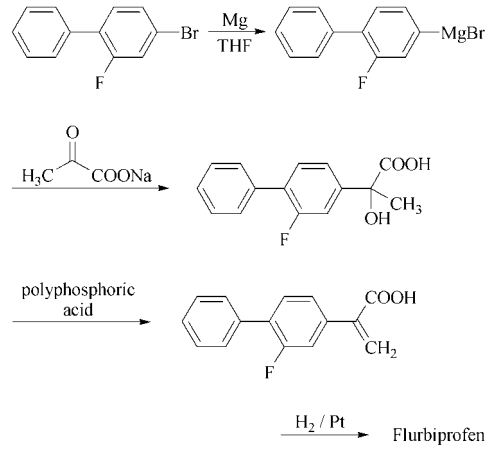

b)

Clinical use: Flurbiprofen is a nonsteroidal anti-inflammatory drug used for the treatment of pain and inflammation associated with musculoskeletal and joint disorders as well as neuralgia, dysmenorrhoea, and postoperative pain. Flurbiprofen is used as free acid or sodium salt and given orally or rectally (150–200 mg/d, max. 300 mg/d) and as ophthalmic solutions. Peak plasma concentration appears 1 to 2 h after oral application. The plasma protein binding is about 99.5 % and the elimination half-life is in the range of 3–5 h.

Use of flurbiprofen as antiplatelet agent after myocardial infarction is discussed [155].

In addition to COX inhibition, flurbiprofen induces weak inhibition of the transcription factors NFκ-B and AP-1 and inhibits cell proliferation, that is why its use in the treatment of cancer has been investigated [156].

Trade names: Ansaid (USA), Antadys, Cebutid (France), Dobedan (Germany), Froben (UK), Ocufen (UK, USA). Ocuflur eye drops (Germany)

Ibuprofen [*15687-27-1*], 2-(4-isobutylphenyl)propionic acid, $C_{18}H_{18}O_2$, M_r 206.28, *mp* 75–77 °C.

Synthesis [48, 157, 158] (see next page).

Several alternative processes are described in the literature.

Clinical use: Ibuprofen [159] is a nonsteroidal anti-inflammatory drug, commonly used for the treatment of mild to moderate pain. It is used in conditions like rheumatoid arthritis, osteoarthritis, joint and soft-tissue pain, dental pain, postoperative pain, dysmenorrhoea, and headache, including acute migraine attacks. Ibuprofen is given by oral, rectal, or topical administration (800–2400 mg/d) and with a lower dose (40 mg kg^{-1} d^{-1}) for the treatment of fever in children.

Ibuprofen shows the typical side effects of nonsteroidal anti-inflammatory drugs but seems to be better tolerated than other nonsteroidal anti-inflammatory drugs. This may be due to the balanced inhibition of COX-1 and COX-2 as seen in cultured human cells [23].

Ibuprofen is absorbed within 1 to 2 h after oral application, bound to plasma protein up to 99 % and shows a plasma half-life of about 2 h. Ibuprofen is a racemate, the active enantiomer, the S(+)-enantiomer, is available in some coun-

a) Boots process (industrial process):

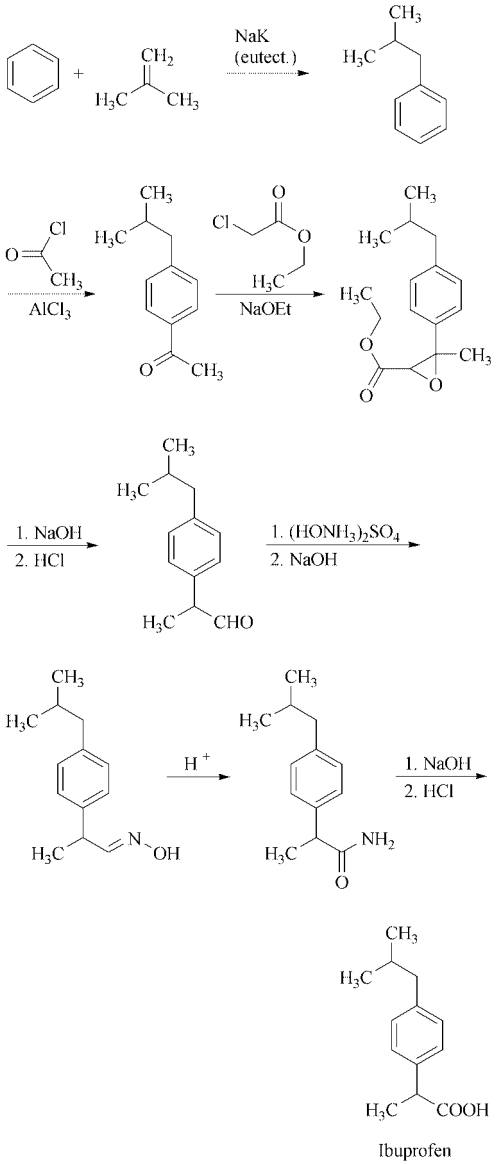

b) Boots-Hoechst-Celanese process:

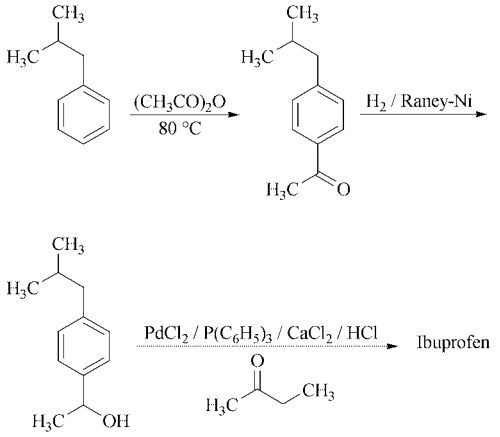

tries (dexibuprofen). Ibuprofen is given as free base or a variety of salts, esters, and other derivatives. Low-dose ibuprofen is available without prescription and has become one of the most popular OTC analgesics.

Trade names: Actren, Anco (Germany), Brufen (France, Italy, UK), Ibuprox (Spain), Imbun (Germany), Motrin (USA).

Combination with hydrocodone: Vicoprofen (USA).

Combination with oxycodone: Combunox (USA).

Ibuprofen is also marketed in many generic formulations.

Ketoprofen [*22071-15-4*], 2-(3-benzoylphenyl)propionic acid, 3-benzoyl-α-methylbenzeneacetic acid, $C_{16}H_{14}O_3$, M_r 254.28, *mp* 94 °C; lysine salt (1 : 1) [*57469-78-0*], $C_{16}H_{14}O_3 \cdot C_6H_{14}N_2O_2$, M_r 400.47; sodium salt [*57495-14-4*], $C_{16}H_{13}NaO_3$, M_r 276.27.

Synthesis [48, 160] (see next page).

Clinical use: Ketoprofen [161, 162] is a nonsteroidal anti-inflammatory drug used for the treatment of a variety of acute and chronic pain and inflammatory conditions including rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, postoperative pain, and dysmenorrhoea. It is administered orally, rectally, topically, or intramuscular (100–200 mg/d, maximal dose 300 mg/d) as sodium or lysine salt. Ketoprofen is given as racemate. The pharmacologically active isomer is mainly the S(+)-enantiomer, which is available in some countries as the trometamol (2-amino-2-(hydroxymethyl)-1,3-propanediol) salt. As compared to the racemate, absorption of the *S*(+)-isomer is said to be faster, leading to an earlier onset of action [163].

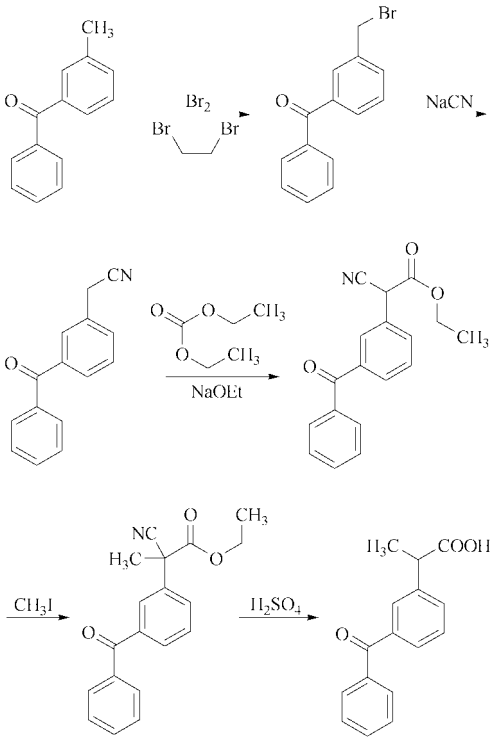

Ketoprofen

The peak plasma concentration after oral administration occurs within 2 h. Ketoprofen is bound to plasma protein up to 99 % and shows a plasma elimination half-life of 1.5 to 4 h.

Trade names: Ketocid (UK), Profénid (France), Orudis (Germany, Italy, USA).

S(+)-ketoprofen: Keral (UK).

Naproxen [*22204-53-1*], (+)-(*S*)-2-(6-methoxynaphthalen-2-yl)propionic acid, $C_{14}H_{14}O_3$, M_r 230.26, *mp* 155.3 °C (also reported as 152–154 °C), $[\alpha]_D$ + 66° (*c* = 1, $CHCl_3$); sodium salt [*26159-34-2*], $C_{14}H_{13}NaO_3$, M_r 252.25, *mp* 244–246 °C, $[\alpha]_D$ − 11° (*c* = 1, $CHCl_3$).

Synthesis [48, 164–166]:

For racemic resolution of naproxen the use of cinchonidine, *N*-alkyl-D-glucamine, dehydroabietylamine or (*S*)-α-phenylethylamine is described. For enzymatic cleavage of esters of racemic naproxen, cloned esterases, which are cheap and easy to produce, have been developed (production in a 100 t/a process is planned). Several other synthetic routes exist.

a) First large-scale manufacturing process of Syntex [167]:

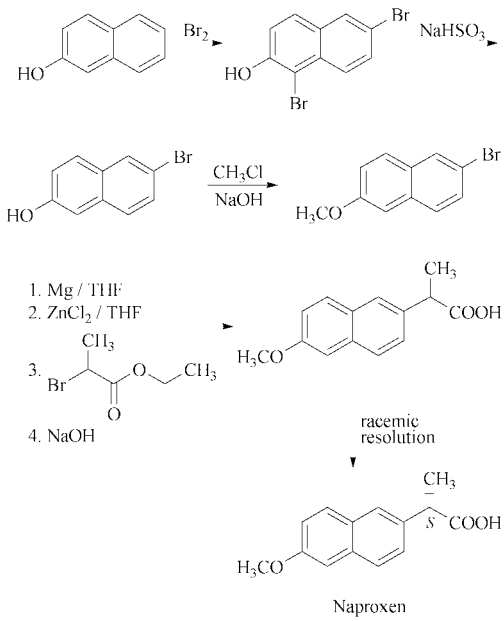

Naproxen

b) Stereospecific Syntex process [168]:
The stereospecific Syntex process is an example using chiral technology to produce enantiomerically pure naproxen:

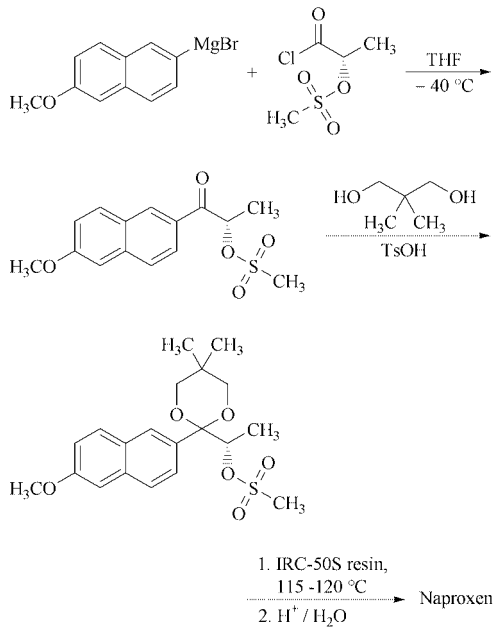

The asymmetric hydrogenation of 2-(6-methoxy-2-naphthyl)acrylic acid using ruthenium-2,2′-bis(diphenylphosphino)-1,1′-binaphthyl (BINAP) complexes yields also enantiomerically pure naproxen.

Clinical use: Naproxen [169] is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain and inflammatory pain conditions like rheumatoid arthritis, osteoarthritis, soft-tissue disorders, postoperative pain, and dysmenorrhoea. In contrast to most other COX inhibitors, which are racemic mixtures, naproxen has been developed and is clinically used only as the (*S*)(+)-enantiomer. Naproxen is given orally or rectally (up to 1250 mg/d). The major side effects are gastrointestinal disturbances. To decrease gastrointestinal side effects, fixed combinations wih the proton pump inhibitor lansoprazole were developed.

Naproxen is available as free base, as sodium salt, and in combination with misoprostol for the reduction of the gastrointestinal side effects. For topical use, morpholinyl- and methylpiperazinylacyloxyalkyl prodrugs have been developed [170].

Trade names: Aleve (France, Germany), Anoprox (USA), Apranax (France, Germany), Naprosyn (UK, USA), Proxen (Germany), Synflex (UK).

Combination with lansoprazole: Prevacid, NapraPAC (USA).

Naproxen is also marketed in several generic formulations.

Tiaprofenic acid [*33005-95-7*], 2-(5-benzoylthiophen-2-yl)propionic acid, $C_{14}H_{12}O_3S$, M_r 260.31, *mp* 96 °C.

Synthesis [48, 171] (see right column, top).

Clinical use: Tiaprofenic acid [172] is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain states in musculoskeletal, soft-tissue, and joint disorders as well as for postoperative pain treatment. Tiaprofenic acid is given as oral or rectal administration (600 mg/d) and as intramuscular injection of the trometamol (2-amino-2-(hydroxymethyl)-1,3-propanediol) salt. The main side effects are urinary tract symptoms like cystitis and bladder irritation [173].

Trade name: Surgam (France, Germany, UK).

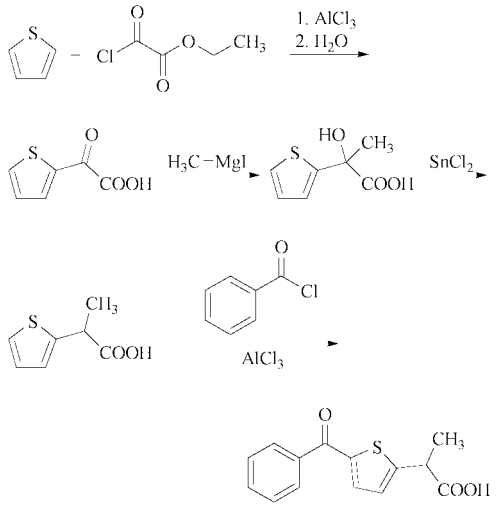

Tiaprofenic acid

Zaltoprofen [*7411-43-6*], 2-(10-oxo-10,11-dihydrodibenzo[*b*,*f*]thiepin-2-yl)-propionic acid, α-methyl-10-oxo-10,11-dihydrodibenzo[*b*,*f*]thiepin-2-acetic acid, 10,11-dihydro-α-methyl-10-oxodibenzo[*bf*]thiepin-2-acetic acid, $C_{17}H_{14}O_3S$, M_r 298.36, *mp* 131-133 °C; (*S*)-enantiomer [*89482-01-9*], *mp* 129.5-131 °C), $[\alpha]_D^{25}$ +32.4° ($c = 1$, chloroform).

Synthesis: The cyclization of 5-(1-cyanoethyl)-2-(phenylthio)phenylacetic acid in presence of polyphosphoric acid at 120 °C gives 2-(10-oxo-10,11-dihydrodibenzo[*b*,*f*]thiepin-2-yl)propionitrile, which is then hydrolyzed with KOH in refluxing ethanol – water to zaltoprofen [174, 175].

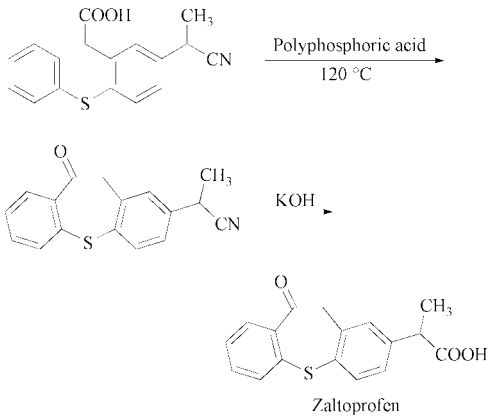

Zaltoprofen

Clinical use: Zaltoprofen is a non-steroidal anti-inflammatory drug originated by Nippon Chemiphar and jointly developed with Zeria. It has been available on the market in Japan since 1993 for the relief of pain and inflammation resulting from arthritis deformations, periarthritis of the shoulder, neck-shoulder-arm syndrome, rheumatoid arthritis, lumbago, post-surgery pain, trauma, and tooth extraction and used in oral doses of 80 mg [176]. According to studies, the analgesic effects of zaltoprofen may involve the inhibition of bradykinin-2 receptor-mediated responses in primary afferent neurons [177, 178]. Although zaltoprofen is marketed as a racemate the anti-inflammatory activity resides in the (*S*)-enantiomer.

Trade name: Peon, Soleton (Japan).

Pirprofen [*31793-07-4*], *rac*-3-chloro-4-(2,5-dihydro-1*H*-pyrrol-1-yl)-α-methylbenzeneacetic acid, *rac*-3-chloro-4-(3-pyrrolin-1-yl)hydratropic acid, $C_{13}H_{14}ClNO_2$, M_r 251.71, *mp* 98-100 °C.

Synthesis:

a) Ethyl 4-nitrophenylacetate is methylated with NaH – CH_3I in dimethylformamide – toluene to form ethyl 2-(4-nitrophenyl)-propionate, which is reduced with H_2 over Pd/C in EtOH – H_2O to ethyl 2-(4-aminophenyl)propionate. This product is acetylated with acetic anhydride to yield ethyl 2-(acetylaminophenyl)propionate. Alternatively, ethyl 4-aminophenylacetate can be acetylated with acetic anhydride to ethyl 4-acetylaminophenylacetate followed by methylation with CH_3I and sodium in liquid NH_3 to give ethyl 2-(4-acetylaminophenyl)propionate [179, 180]. Ethyl 2-(4-acetylaminophenyl)propionate is chlorinated with Cl_2 in acetic acid to ethyl 2-(3-chloro-4-acetylaminophenyl)propionate. This intermediate is hydrolyzed with HCl in EtOH

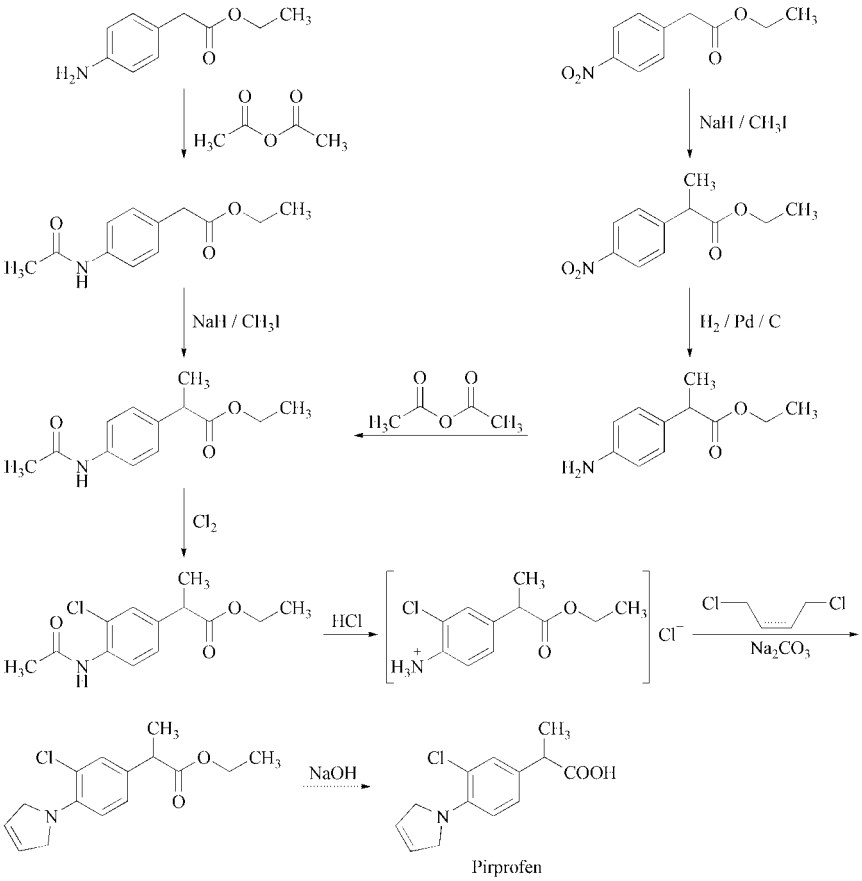

to give ethyl 2-(3-chloro-4-aminophenyl)-propionate hydrochloride. The last step is the reaction of the preceding intermediate with *cis*-1,4-dichloro (or dibromo)-2-butene in dimethylformamide and Na_2CO_3 to ethyl 2-(3-chloro-4-(3-pyrrolinyl)phenyl)-propionate, followed by hydrolysis with KOH in EtOH – H_2O (see previous page).

b) 2,4-Dichloronitrobenzene is reacted with sodium diethyl methylmalonate in dimethylformamide to yield diethyl 2-(4-nitro-3-chlorophenyl)-2-methylmalonate, followed by hydrolysis, decarboxylation and esterification to form ethyl 2-(3-chloro-4-nitrophenyl)propionate. The latter is reduced as in route (a) to ethyl 2-(3-chloro-4-aminophenyl)propionate hydrochloride. Alternatively, diethyl 2-(4-nitro-3-chlorophenyl)-2-methylmalonate can be hydrogenated first to diethyl 2-(4-amino-3-chlorophenyl)-2-methylmalonate, and then hydrolyzed, decarboxylated and esterified again to form ethyl 2-(3-chloro-4-aminophenyl)propionate hydrochloride [179, 180] (see below).

Clinical use: Pirprofen is a nonsteroidal anti-inflammatory drug [181] and can be used for the treatment of post-operative pain, oncological pain, episodic headache attacks, and acute migraine [182]. There are reports on hepatic and renal intolerability [183–186] which precluded the further use.

Trade names: Rengasil (France, Belgium, Netherlands, Chile), Seflenyl.

Fenoprofen [*31879-05-7*], *rac*-α-methyl-3-phenoxybenzeneacetic acid, *rac-m*-phenoxyhydratropic acid, α-(d,l)-2-(3-phenoxyphenyl)propionic acid, $C_{15}H_{14}O_3$, M_r 242.27, viscous oil, $bp_{0.11}$ (0.147 hPa) 168-171 °C; calcium salt dihydrate, [*53746-45-5*], $C_{30}H_{26}CaO_6 \cdot 2H_2O$, M_r 558.64, *mp* 68-171 °C.

Synthesis: The condensation of 3-hydroxyacetophenone with bromobenzene yields 3-phenoxyacetophenone, which is reduced with sodium borohydride to 1-(3-phenoxyphenyl)ethanol. Bromination with phosphorus(III)bromide yields 1-(1-bromoethyl)-3-phenoxybenzene. After reaction with sodium cyanide the resulting 2-(3-phenoxyphenyl)propanenitrile is hydrolyzed to fenoprofen [187, 188]; for alternative syntheses see [189–191].

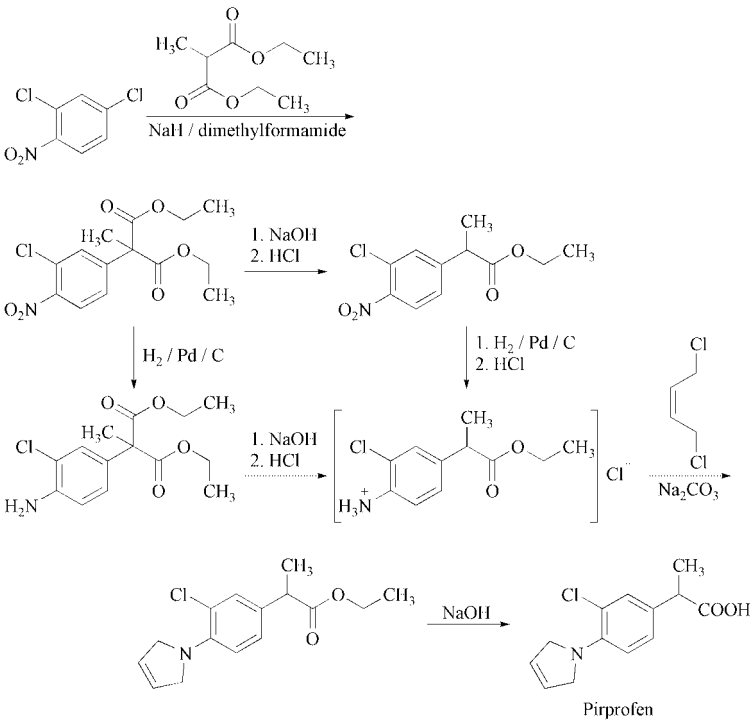

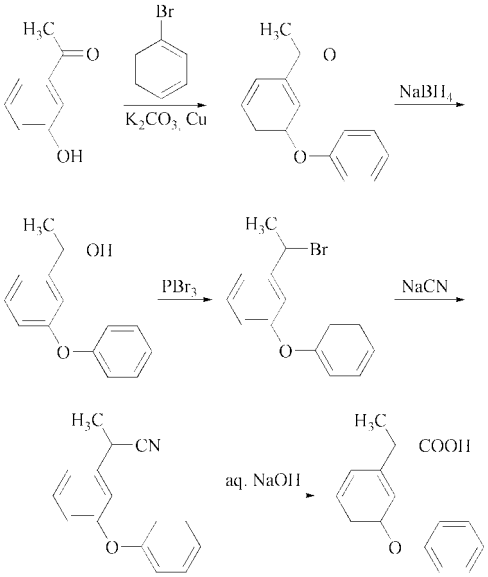

Fenoprofen

Clinical use: Fenoprofen [192–194] is a nonsteroidal anti-inflammatory drug used in oral doses of 300 mg as the calcium salt or in doses of 200–400 mg as free base for the treatment of moderate to severe pain and symptomatic treatment of acute and chronic rheumatoid arthritis and osteoarthritis. Enteric coated tablets of the Ca salt were developed to reduce gastrointestinal microbleeding [195]. The use of fenoprofen was associated with renal impairment [196] and allergic skin inflammation [197].

Trade names: Fepron (Italy), Nalfon (USA), Nalgesic (France).

Vedaprofen [*71109-09-6*], *rac*-4-cyclohexyl-α-methyl-1-naphthaleneacetic acid, *rac*-2-(4-cyclohexylnaphthalen-1-yl)propanoic acid, $C_{19}H_{22}O_2$, M_r 282.38, *mp* 150 °C.

Synthesis: The reaction of 1-cyclohexylnaphthalene and paraformaldehyde in the presence of concentrated hydrochloric acid and H_3PO_4 at 85 °C yields 1-chloromethyl-4-cyclohexylnaphthalene, which is transformed to (4-cyclohexylnaphthalen-1-yl)-acetonitrile by means of sodium cyanide in water acetone mixture as solvent. The hydrolysis of the nitrile with potassium hydroxide yields the sodium salt of (4-cyclohexylnaphthalen-1-yl)-acetic acid. After treatment with hydrochloric acid the free acid is converted to the corresponding (4-cyclohexyl-naphthalen-1-yl)-acetic acid methyl ester by means of acetyl chloride. Methylation with methyl iodide in the presence of *n*-butyllithium yields the desired vedaprofen after the standard work-up procedure [198].

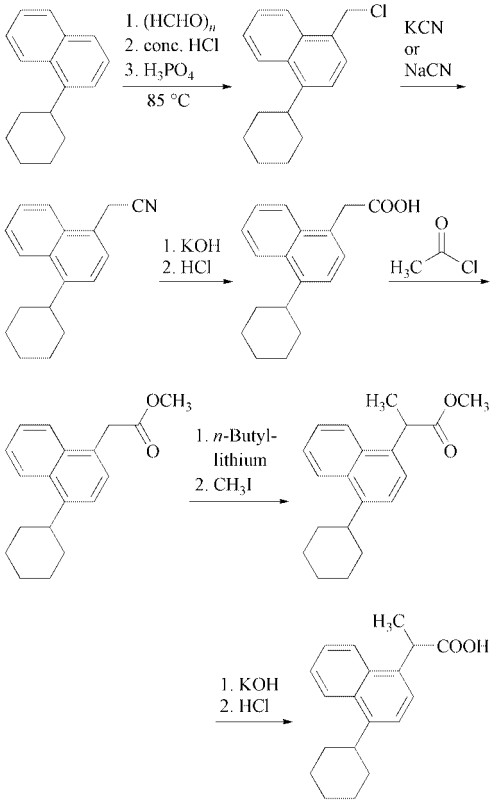

Vedaprofen

Clinical use: Vedaprofen is a nonsteroidal anti-inflammatory drug with antipyretic and analgesic properties, based on inhibition of prostaglandin synthesis. Both enantiomers contribute to the biological and therapeutic actions of vedaprofen. Although vedaprofen was initially developed for human use, it is is now mainly used as a veterinary drug as oral gel [199] for the reduction of inflammation and pain associated with musculo-skeletal disorders in dogs [200] and horses and parenterally as the meglumine salt for the acute treatment of colic pain in horses [201]. In horses vedaprofen is administered orally at an initial dose of 2 mg/kg, followed by 1 mg/kg every 12 h for up to 14 days.

Trade name: Quadrisol.

Nepafenac [*78281-72-8*], 2-(2-amino-3-benzoylphenyl)acetamide, 2-amino-3-benzoylbenzeneacetamide, $C_{15}H_{14}N_2O_2$, M_r 254.29, *mp* 178.5-180 °C.

Synthesis: The reaction of (2-aminophenyl)-phenylmethanone and thiocarbamic acid *S*-methyl ester in the presence of 2,2-dimethylpropionyl chloride in methylene chloride at -65 °C and subsequent treatment with triethylamine at the same temperature yields the corresponding condensation product 2-(2-amino-3-benzoyl-phenyl)-2-methylsulfanylacetamide. Finally, nepafenac is obtained by reduction with Raney nickel in tetrahydrofuran [202, 203].

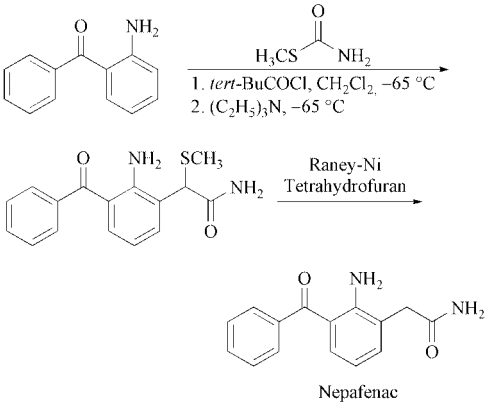

Clinical use: Nepafenac is a nonsteroidal antiinflammatory drug, that was launched by Alcon in 2005 for the treatment of pain and inflammation associated with cataract surgery, because it rapidly penetrates ocular tissues, and it is used locally as 1 % ophthalmic suspensions [204].

Nepafenac is an amide analog and a prodrug of amfenac (2-amino-3-benzoylbenzeneacetic acid), which is a COX-1 and COX-2 inhibitor [205]. Nepafenac is the first ophthalmic NSAID prodrug to receive FDA approval. It had originally been developed at Wyeth Consumer Healthcare [204, 206, 207].

Trade name: Nevanac.

Amfenac [*51579-82-9*], 2-amino-3-benzoylbenzeneacetic acid, 2-amino-3-benzoylphenylacetic acid, $C_{15}H_{13}NO_3$, M_r 255.28, *mp* 121-123 °C (dec.); sodium salt monohydrate [*61618-27-7*], $C_{15}H_{12}NNaO_3\ H_2O$, M_r 295.27, *mp* 254-255.5 °C.

Synthesis: The reaction of 1-aminoindolin-2-one with phenylacetone in presence of acetic acid in refluxing ethanol gives 1-(α-methylphenethylidieneimino)indolin-2-one, which by reaction with refluxing ethanolic hydrogen chloride affords ethyl α-(2-methyl-3-phenylindol-7-yl)acetate. The ozonolysis of this intermediate in acetic acid yields ethyl 2-acetamido-3-benzoylphenyl acetate, which is cyclized by refluxing with HCl in acetic acid to give 7-benzoylindolin-2-one. Alternatively, the hydrolysis of the ester ethyl α-(2-methyl-3-phenylindol-7-yl)acetate with KOH in refluxing water affords the corresponding acid, which can be ozonolyzed as before yielding 2-acetamido-3-benzoylphenylacetic acid. This acid can be cyclized to 7-benzoyl-1,3-dihydro-indol-2-one by refluxing with HCl in acetic acid as before [208–210].

2.6. Pyrazolinone Derivatives

Metamizol [*50567-35-6*], dipyrone[(1,5-dimethyl-3-oxo-2-phenyl-2,3-dihydro-1*H*-pyrazol-4-yl)methylamino]-methanesulfonic acid, $C_{13}H_{17}N_3O_4S$, M_r 311.36; sodium salt [*68-89-3*], $C_{13}H_{16}N_3NaO_4S$, M_r 333.34; sodium salt monohydrate [*50567-35-6*], $C_{13}H_{16}N_3NaO_4S \cdot H_2O$, M_r 351.36.

Synthesis [48, 211, 212] (see next page).

Clinical use: Metamizol [213, 214] is the water soluble sodium sulfonate of amidopyrine. After oral application it is rapidly hydrolyzed to 4-methylaminoantipyrine and metabolized to various active metabolites [215]. Metamizol has strong analgesic, spasmolytic, and antipyretic, but no anti-inflammatory properties. The exact mechanism of action is unknown but may include inhibition of prostaglandin synthesis. Inhibition of both COX isoenzymes by metamizol has been demonstrated only when very high concentrations of this drug were used, thus questioning the relevance of this activity.

COX inhibition may be induced by the metabolites, but this is not yet systematically investigated. Metamizol is used for the treatment of medium to severe pain, often in combination with opioids, for fever reduction and for the treatment of colic pain. It is given by mouth in doses of 500 mg up to 4 g daily, by intravenous or by rectal application. Metamizol is relatively free of acute side effects but in rare cases may cause severe and life-threatening allergic

reactions like agranulocytosis, allergic skin reactions, and allergic shock. The compound is not used in the UK, USA, and in Scandinavian countries.

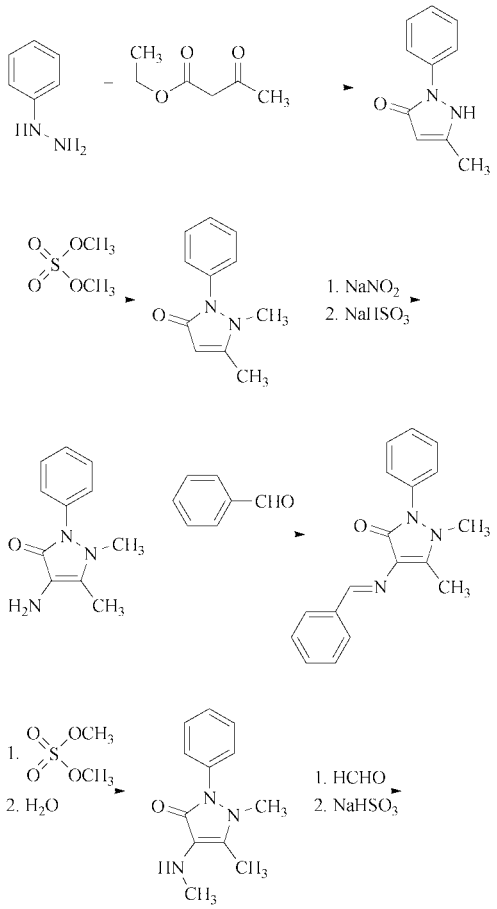

Metamizol sodium

Trade names: Analgin, Novalgin (Germany, Austria, Switzerland).

Metamizol is also marketed in several generic formulations.

Propyphenazone [*479-92-5*], 1,2-dihydro-1,5-dimethyl-4-(1-methylethyl)-2-phenyl-3*H*-pyrazol-3-one, $C_{14}H_{18}N_2O_3$, M_r 230.31, *mp* 103 °C.

Synthesis [48, 216, 217]:

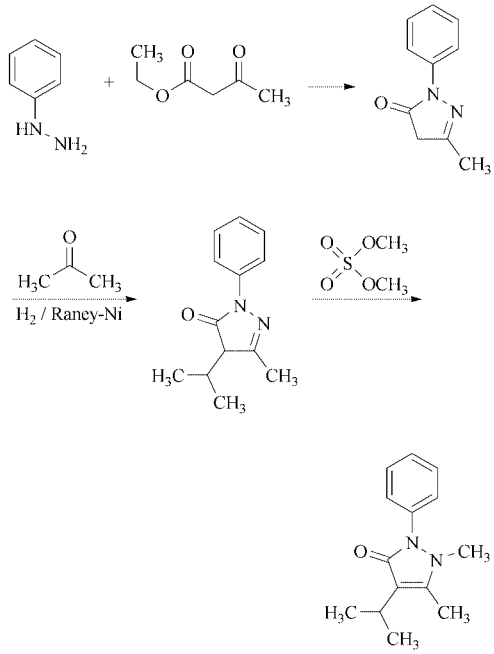

Propyphenazone

Clinical use: Propyphenazone [218, 219] is a derivative of phenazone and has a similar analgesic and antipyretic action. It is used as a monocompound as well as in multi-ingredient preparations.

Trade name: Demex (Germany).

Multi-ingredient preparation: Saridon (Germany).

2.7. Acidic Enolic Compounds

2.7.1. Pyrazolidine-3,5-diones

Kebuzone [*853-34-9*], 4-(3-oxobutyl)-1,2-diphenyl-pyrazolidine-3,5-dione, $C_{19}H_{18}N_2O_3$, M_r 322.36, *mp* 115.5–116.5 °C or 127.5–128.5 °C, depending on crystal form.

Synthesis [158, 220, 221]:

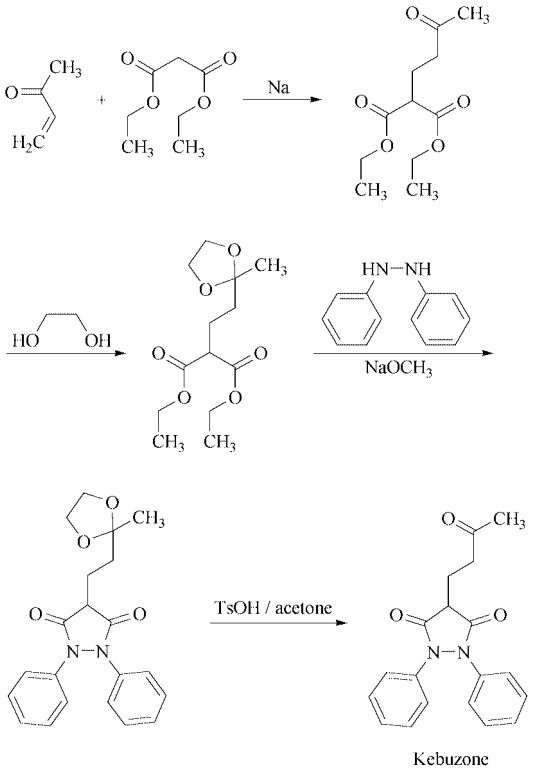

Kebuzone

Clinical use: Kebuzone [222] is a non-steroidal anti-inflammatory drug used for the treatment of acute and chronic pain and inflammation states like musculoskeletal, joint, and soft-tissue disorders. Kebuzone is administered orally, rectally, or intramuscular (up to 1500 mg/d initial dose, 250–500 mg/d maintenance dose).

Trade names: Ketazon (Austria), Chetopir (Italy).

Mofebutazone [*2210-63-1*], 4-butyl-1-phenylpyrazolidine-3,5-dione, $C_{13}H_{16}N_2O_2$, M_r 232.28, *mp* 102–103 °C; sodium salt [*41468-34-2*], $C_{13}H_{15}N_2NaO_2$, M_r 254.27.

Synthesis [48, 223, 224]:

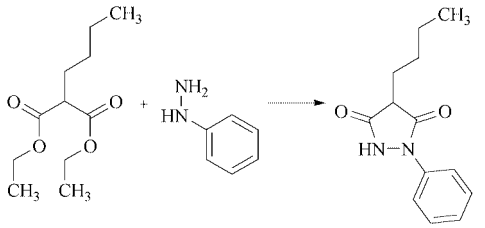

Mofebutazone

Clinical use: Mofebutazone [225] is a non-steroidal anti-inflammatory drug used as sodium salt for the treatment of mild to moderate pain including inflammatory and degenerative rheumatic disorders and musculoskeletal pain. Mofebutazone can be given by oral, rectal (900–1200 mg/d), or intramuscular administration (650 mg/d).

Trade name: Mofesal (Germany).

Oxyphenbutazone [*129-20-4*], oxyphenylbutazone, 4-butyl-1-(4-hydroxyphenyl)-2-phenylpyrazolidine-3,5-dione, $C_{19}H_2ON_2O_3$, M_r 324.37, *mp* 124–125 °C; monohydrate [*7081-38-1*], $C_{19}H_{20}N_2O_3 \cdot H_2O$, M_r 342.40, *mp* 96 °C.

Synthesis [48, 226]:

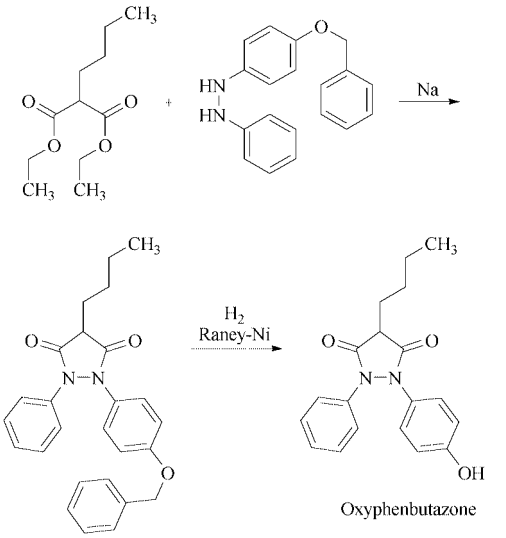

Oxyphenbutazone

Clinical use: Oxyphenbutazone is a non-steroidal anti-inflammatory drug, which was used by oral, rectal, or topical administration (400 - 600 mg/d) for the acute treatment of ankylosing spondylitis, chronic polyarthritis, and gout. Because of a high incidence of severe side effects including disturbances of the hematopoietic system like agranulocytosis and aplastic anemia [233] the compound is no longer used. Oxyphenbutazone is a metabolite of phenylbutazone.

Phenylbutazone [*50-33-9*], 4-butyl-1,2-diphenylpyrazolidine-3,5-dione, $C_{19}H_{20}N_2O_2$,

M_r 308.37, *mp* 105 °C; sodium salt [*129-18-0*], $C_{19}H_{19}N_2NaO_2$, M_r 331.36; calcium salt (2 : 1) [*36298-23-4*], $C_{38}H_{36}CaN_4O_4$, M_r 656.83; piperazine salt (1 : 1) [*4985-25-5*], $C_{19}H_{20}N_2O_2 \cdot C_4H_{10}N_2$, M_r 395.51, *mp* 140–141 °C.

Synthesis [48, 228]:

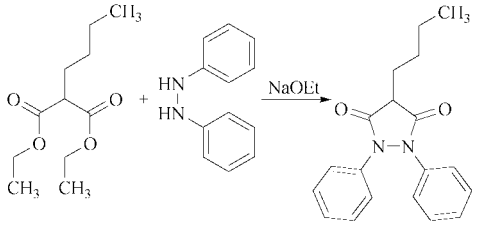

Phenylbutazone

Clinical use: Phenylbutazone [229] is a nonsteroidal anti-inflammatory drug used for the acute treatment of ankylosing spondylitis, chronic polyarthritis, and gout. Because of severe side effects including disturbances of the hematopoietic system like agranulocytosis and aplastic anemia [230] the use is limited to conditions in which other nonsteroidal anti-inflammatory drugs do not show sufficient efficacy. Phenylbutazone is given as oral, rectal, intramuscular or topical formulation (up to 600 mg/d initial dose, up to 400 mg/d maintenance dose). The peak plasma concentration is reached 2 h after oral application. Phenylbutazone is bound to 98 % to plasma protein. Oxyphenbutazone is formed as an active metabolite of phenylbutazone.

Trade names: Butazolidin (France, Germany, UK, USA), Ambene (Germany).

2.7.2. Arylsulfonamides (Oxicames)

The oxicams are a group of nonsteroidal anti-inflammatory agents containing a 2*H*-1,2-benzothiazine-3-carboxamide 1,1-dioxide moiety as common structural entity (for Meloxicam see Section 2.9.1).

Isoxicam [*34552-84-6*], 4-hydroxy-2-methyl-1,1-dioxo-1,2-dihydro-1λ^6-benzo[*e*][1,2]thiazine-3-carboxylic acid isoxazo-3-ylamide, 4-hydroxy-2-methyl-*N*-(5-methyl-3-isoxazolyl)-2*H*-1,2-benzothiazine-3-carboxamide 1,1-dioxide, $C_{14}H_{13}N_3O_5S$, M_r 335.34, *mp* 265–271 °C (decomp.); sodium salt, $C_{14}H_{12}N_3NaO_5S$, M_r 343.29, *mp* 270–272 °C.

Synthesis [48, 231–233]:

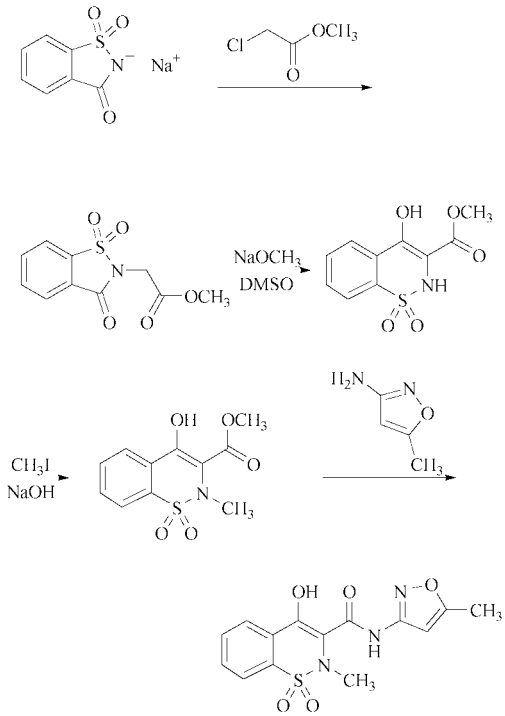

Isoxicam

Clinical use: Isoxicam [234] is a nonsteroidal anti-inflammatory drug withdrawn from the market as a consequence of reports of fatal skin reaction.

Lornoxicam [*70374-39-9*], 6-chloro-4-hydroxy-2-methyl-1,1-dioxo-1,2-dihydro-1λ^6-thieno[2,3-*e*][1,2]thiazine-3-carboxylic acid pyridin-2-ylamide, 6-chloro-4-hydroxy-2-methyl-*N*-(2-pyridyl)-2*H*-thieno[2,3-*e*]-1,2-thiazine-3-carboxamide 1,1-dioxide, $C_{13}H_{10}ClN_3O_4S_2$, M_r 371.81, *mp* 225–230 °C (decomp.).

Synthesis: Sulfonation of 2,5-dichlorothiophene with $ClSO_3H/SOCl_2$ gives 2,5-dichlorothiophene-3-sulfonyl chloride, which by reaction with methylamine in $CHCl_3$ yields the corresponding methylamide. The carboxylation with butyl lithium and CO_2 in ether affords 5-chloro-3-(*N*-methylsulfamoyl)thiophene-2-carboxylic acid,

which is esterified with PCl_5 and methanol to the methyl ester. Condensation with methyl iodoacetate by means of NaH in DMF gives methyl 5-chloro-3-[*N*-(methoxycarbonylmethyl)-*N*-methylsulfamoyl]thiophene-2-carboxylate, which is cyclized with sodium methoxide in methanol yielding methyl 6-chloro-4-hydroxy-2-methyl-2*H*-thieno[2,3-*e*]-1,2-thiazine-3-carboxylate-1,1-dioxide. Finally this compound is treated with 2-aminopyridine in refluxing xylene [235].

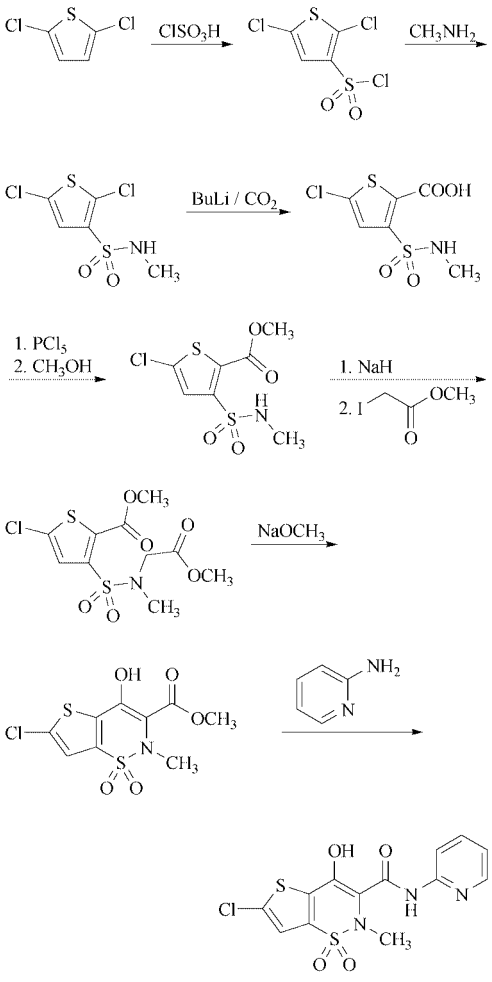

Lornoxicam

Clinical use: Lornoxicam [236] is a non-steroidal anti-inflammatory drug with a strong and balanced inhibition of COX-1 and COX-2 in cultured human cells [23]. In addition, lornoxicam inhibits inducible nitric oxide synthase and interleukin-6, both of which could contribute to its potent anti-inflammatory and analgesic action [237]. It is used orally (8–24 mg/d) for the treatment of mild to moderate pain including postoperative pain, rheumatoid arthritis, osteoarthritis, and ankylosing spondylitis [238]. Lornoxicam shows a high binding to plasma protein and in contrast to other oxicams a short plasma half-life of 3 to 5 h [236].

Trade names: Telos (Germany), Taigalor (Italy), Xefo (Austria, Denmark, Italy, Switzerland, UK).

Piroxicam [*36322-90-4*], 4-hydroxy-2-methyl-1,1-dioxo-1,2-dihydro-1λ^6-benzo-[*e*][1,2]thiazine-3-carboxylic acid pyridin-2-ylamide, 4-hydroxy-2-methyl-*N*-2-pyridinyl-2*H*-1,2-benzothiazine-3-carboxamide 1,1-dioxide, $C_{15}H_{13}N_3O_4S$, M_r 331.35, *mp* 198–200 °C.

Synthesis [239–243, 251]:

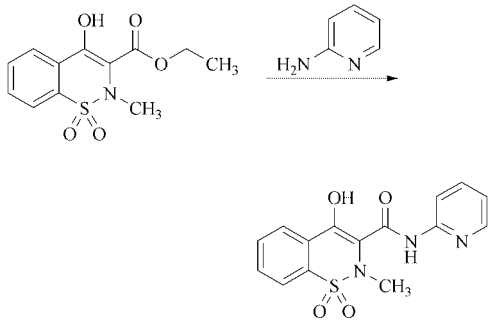

Piroxicam

Clinical use: Piroxicam [244] is a non-steroidal anti-inflammatory drug used for the treatment of mild to moderate acute and chronic pain and inflammation including musculoskeletal, soft-tissue, and joint disorders like ankylosing spondylitis, chronic polyarthritis, and gout [245]. Piroxicam shows a 600-fold selectivity for COX-1 compared to COX-2 in cultured animal cells [246]. Piroxicam is administered orally, rectally, intramuscular, or topically (10–30 mg/d, maximal initial dose 40 mg/d) as the free base, as complex with beta-cyclodextrine, and as cinnamate or pivalate. After oral administration, piroxicam reaches peak plasma concentration after 3 to 5 h, shows a 99 % binding to plasma protein and a long half-life of about 50 h.

Trade names: Brexidol (EU, USA), Brexin (Franc, Italy), Feldéne (France), Felden (Germany), Feldene (UK, USA).

Piroxicam is also marketed in several generic formulations.

Tenoxicam [*59804-37-4*], 4-hydroxy-2-methyl-1,1-dioxo-1,2-dihydro-1λ^6-thieno[2,3-*e*][1,2]thiazine-3-carboxylic acid pyridin-2-ylamide, 4-hydroxy-2-methyl-*N*-2-pyridinyl-2*H*-thieno[2,3-*e*]-thiazine-3-carboxamide 1,1-dioxide, $C_{13}H_{11}N_3O_4S_2$, M_r 337.38, *mp* 209–213 °C (decomp.).

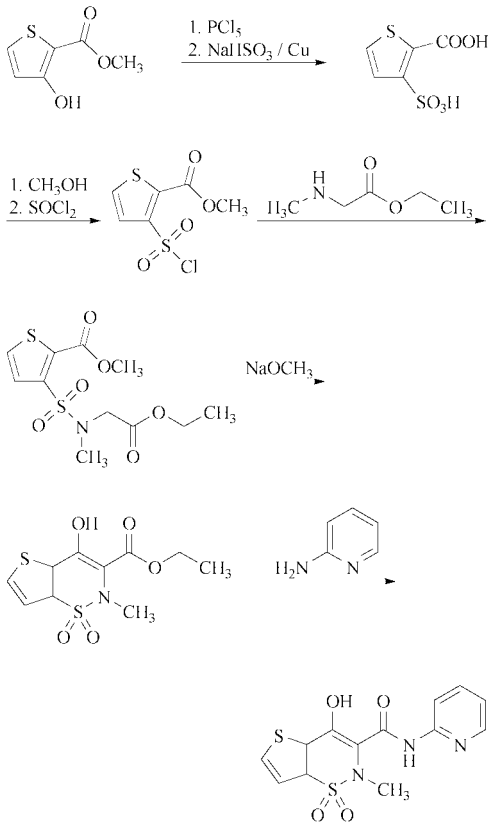

Tenoxicam

Synthesis: The reaction of methyl 3-hydroxythiophene-2-carboxylate with PCl_5 in refluxing CCl_4 gives 3-chlorothiophene-2-carboxylic acid, which by treatment with $NaHSO_3$ and Cu in aqueous alkaline solution at 143 °C in a pressure vessel is converted into 3-sulfothiophene-2-carboxylic acid. Its esterification with refluxing methanol affords methyl-3-sulfothiophene-2-carboxylate, which by reaction with refluxing $SOCl_2$ yields methyl-3-chlorosulfonylthiophene-2-carboxylate. The following condensation with sarcosine ethyl ester in hot $CHCl_3$ gives 3-(*N*-ethoxycarbonylmethyl-*N*-methylsulfamoyl)thiophene-2-carboxylate, which is cyclized by treatment with sodium methoxide in refluxing methanol affording 3-ethoxycarbonyl-4-hydroxy-2-methyl-2*H*-thieno-[2,3-*e*]-1,2-thiazine 1,1-dioxide. Finally this compound is condensed with 2-aminopyridine in refluxing toluene [48, 243, 247].

Clinical use: Tenoxicam [248] is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain states in musculoskeletal, soft-tissue, and joint disorders like rheumatoid arthritis, osteoarthritis and gout. Tenoxicam shows a balanced inhibition of both COX-1 and COX-2 in cultured human cells [23]. Tenoxicam is administered orally, rectally, or intramuscular (20 mg/d, maximal dose 40 mg/d). Peak plasma concentration appears within 1 to 6 h after administration depending on food intake. Tenoxicam is bound to plasma protein to 98.5 % and shows a long plasma elimination half-life of 70 to 90 h.

Trade names: Mobiflex (UK), Tilcotil (France, Italy, Switzerland).

2.8. Other Structures

Ketorolac [*74103-06-3*], 5-benzoyl-2,3-dihydro-1*H*-pyrrolizine-1-carboxylic acid, $C_{15}H_{13}NO_3$, M_r 255.27, *mp* 160–161 °C; tromethamine salt (1 : 1) [*74103-07-4*], $C_{15}H_{13}NO_3 \cdot C_4H_{11}NO_3$, M_r 376.41, (+)-form *mp* 174 °C, $[\alpha]_D$ + 173° (*c* = 1, CH_3OH), (−)-form *mp* 169–170 °C, $[\alpha]_D$ − 176° (*c* = 1, CH_3OH); monosodium salt [*110618-38-7*], $C_{15}H_{12}NNaO_3$, M_r 277.26.

Synthesis: Benzoylation of 2-methylthiopyrrole with *N*,*N*-dimethylbenzamide in the presence of $POCl_3$ in refluxing CH_2Cl_2 gives 5-benzoyl-2-methylthiopyrrole, which is condensed with spiro[2,5]-5,7-dioxa-6,6-dimethyloctane-4,8-dione by means of NaH in DMF. Oxidation of this product with *m*-chloroperbenzoic acid in CH_2Cl_2 affords the sulfone, which is submitted to methanolysis with methanol and HCl giving 1-(3,3-dimethoxycarbonylpropyl)-2-methanesulfonyl-5-benzoylpyrrole. Cycliza-

tion with NaH in DMF yields dimethyl-5-benzoyl-1,2-dihydro-3*H*-pyrrolo[1,2-α]pyrrole-1,1-dicarboxylate, which is finally hydrolyzed and decarboxylated with KOH in refluxing methanol [48, 249–253].

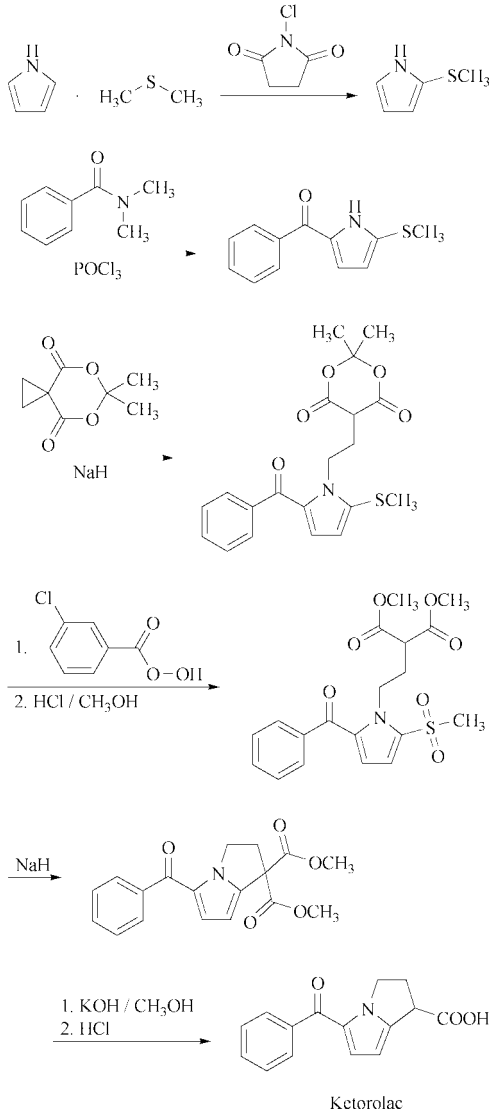

Clinical use: Ketorolac [254] is a nonsteroidal anti-inflammatory drug mainly used for the treatment of moderate to severe postoperative pain. Ketorolac shows a balanced inhibition of COX-1 and COX-2 in cultured human cells [23]. Ketorolac is mostly used as the tromethamine salt. Due to a number of severe side effects including gastrointestinal disturbances, impairment of liver functions, renal failures, skin irritations, and other hypersensitivity reaction it has been withdrawn in many countries.

Trade names: Acular (UK, USA), Droal (Spain), Tora-Dol (France, Italy), Toradol (UK, USA).

Oxaprozine [*21256-18-8*], 3-(4,5-diphenyl-oxazol-2-yl)propionic acid, $C_{18}H_{15}NO_3$, M_r 293.32, *mp* 160.5–161.5 °C.

Synthesis [255–257]:

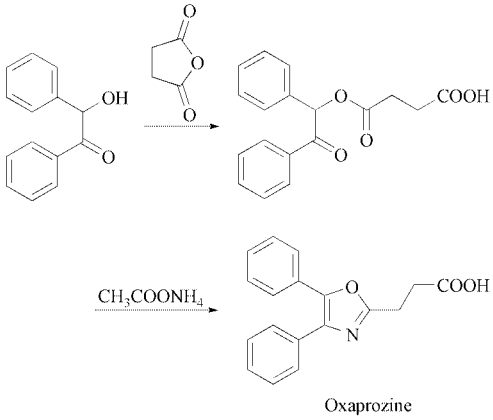

Clinical use: Oxaprozine [257] is a nonsteroidal anti-inflammatory drug used for the treatment of mild to moderate pain including rheumatoid arthritis and osteoarthritis. Oxaprozine shows slow elimination kinetics with an elimination half-life of about 24 h. It is given orally (600–1200 mg/d, maximum dose 1800 mg/d).

Trade names: Deflam (South Africa), Daypro (USA).

2.9. COX-2 Inhibitors

2.9.1. Nonselective COX-2 Inhibitors

Etodolac [*41340-25-4*], (1,8-diethyl-1,3,4,9-tetrahydropyrano[3,4-*b*]indol-1-yl)-acetic acid, $C_{17}H_{21}NO_3$, M_r 287.35, *mp* 145–148 °C.

Synthesis [258–261]:

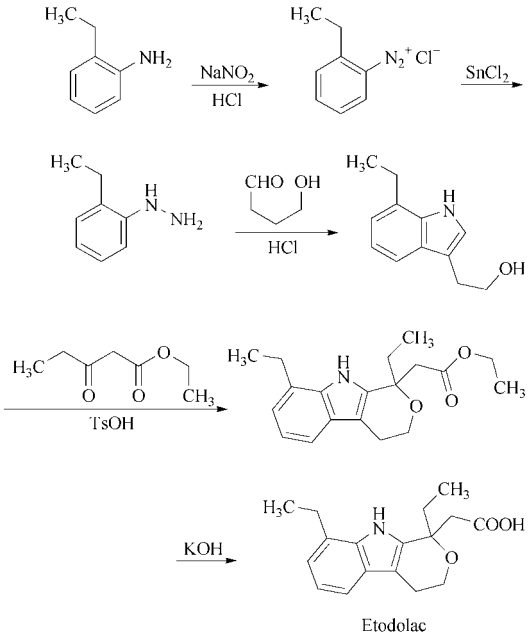

Clinical use: Etodolac [262] shows a ten-fold selectivity for COX-2 compared to COX-1 in human whole blood and belongs to the first generation of COX-2 inhibitors [22]. Etodolac is a racemate with an active (*S*)-enantiomer and an inactive (*R*)-enantiomer. It is used for the treatment of mild to moderate acute and chronic pain including rheumatoid arthritis and osteoarthritis. Clinical data indicate fewer gastrointestinal side-effects in comparison to naproxen [51, 264]. Etodolac is used in oral slow and immediate release preparations and applied in doses of 600–1200 mg/d. Peak plasma concentrations are reached within 2 h. Etodolac shows a 99 % binding to plasma protein and an elimination half-life of about 7 h.

Trade name: Lodine (USA, UK, France).

Meloxicam [*71125-38-7*], 4-hydroxy-2-methyl-1,1-dioxo-1,2-dihydro-1λ^6-benzo[*e*][1,2]thiazine-3-carboxylic acid (5-methyl thiazol-2-yl)amide, 4-hydroxy-2-methyl-*N*-(5-methyl-2-thiazoyl)-2*H*-1,2-benzothiazine-3-carboxamide 1,1-dioxide, $C_{14}H_{13}N_3O_4S_2$, M_r 351.40, *mp* 264 °C (decomp.).

Synthesis: Reaction of benzothiazolo-3(2*H*)-one-1,1-dioxide with methyl chloroacetate gives the methyl 2(3*H*)-acetate derivative, which is isomerized with sodium methoxide in toluene-*tert*-butanol yielding methyl 4-hydroxy-2*H*-1,2-benzothiazine-3-carboxylate-1,1-dioxide. Subsequent methylation with methyl iodide in methanol yields the 2-methyl compound. Finally this compound is treated with 2-amino-5-methylthiazole in xylene [48, 265, 266].

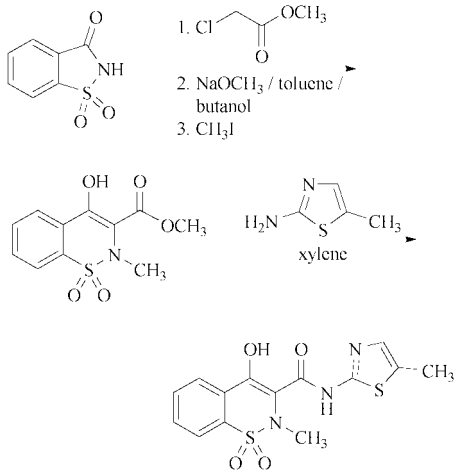

Meloxicam

Clinical use: Meloxicam [267] shows a ten-fold selectivity for COX-2 compared to COX-1 in human whole blood and belongs to the first generation of COX-2 selective drugs [22]. Meloxicam is used for the acute and chronic treatment of mild to moderate pain including arthritis and ankylosing spondylitis. Meloxicam shows central antinociceptive effects in rats which seems to be independent from the COX inhibiting properties [268].

Meloxicam is given by oral administration (7.5–15 mg/d). It shows an elimination half-life of 20 h.

Trade names: Mobec (Germany), Mobic (Italy, UK, USA), Movalis (Spain).

Nimesulide [*51803-78-2*], *N*-(4-nitro-2-phenoxyphenyl)methanesulfonamide, $C_{13}H_{12}N_2O_5S$, M_r 308.05, *mp* 143–144.5 °C.

Synthesis [48, 269]:

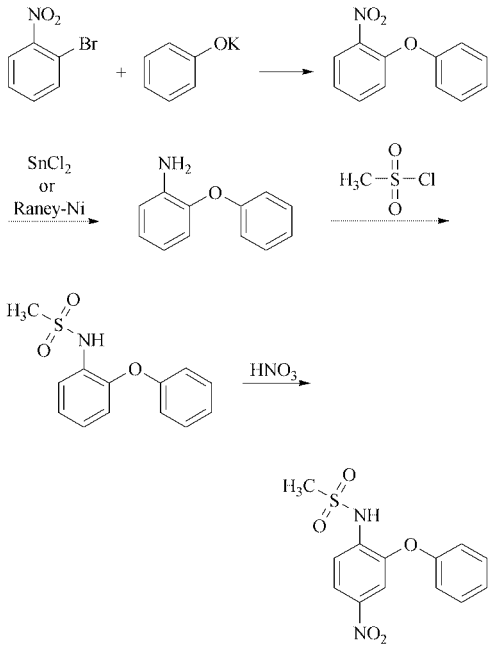

Nimesulide

Clinical use: Nimesulide [270] is a first-generation COX-2 inhibitor with a 5- to 100-fold selectivity for COX-2 compared to COX-1, depending on the assay system [22, 23]. It is used for the short-term treatment of inflammatory conditions, fever, and pain, including musculoskeletal and joint disorders. In addition to COX inhibition nimesulide reduces the production of the pro-inflammatory cytokine TNF-α under inflammatory conditions [271].

Nimesulide is available in oral or rectal formulations (up to 400 mg/d).

The use of nimesulide has been reported to be associated in some cases with hepatic failure [272, 273] and the compound was withdrawn from the market in several countries.

Trade names: Aulin (Italy, Switzerland), Nexen (France), Guaxan (Spain, withdrawn).

2.9.2. Selective COX-2 Inhibitors (Coxibs)

Celecoxib [*169590-42-5*], 4-(5-*p*-tolyl-3-trifluoromethylpyrazol-1-yl)-benzenesulfonamide, $C_{17}H_{14}F_3N_3O_2S$, M_r 381.37, *mp* 157–159 °C (pale yellow solid).

Synthesis: Condensation of 4-methylacetophenone with ethyl trifluoroacetate in refluxing methanol using sodium methoxide as catalyst gives 4,4,4-trifluoro-1-(4-methylphenyl)butane-1,3-dione, which is cyclized with 4-hydrazinophenylsulfonamide in refluxing ethanol [488, 493].

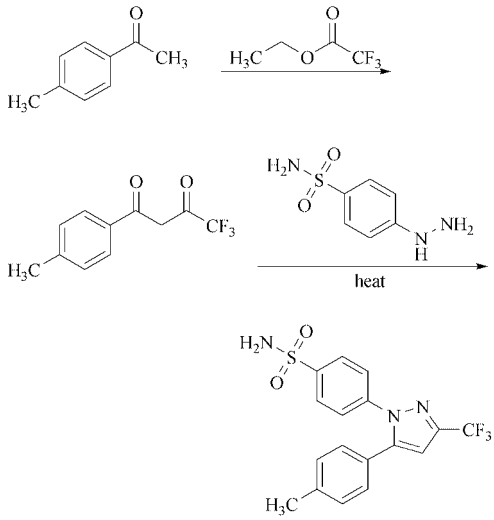

Celecoxib

Clinical use: Celecoxib [166] is a second-generation selective COX-2 inhibitor and was the first drug of this group to reach the market. The selectivity for COX-2 compared to COX-1 is about 375-fold in human recombinant enzyme preparations and about eight-fold in a whole blood assay. Celecoxib has been approved in 1999 in the USA and is now used for the treatment of rheumatoid arthritis and osteoarthritis, ankylosing spondylitis, acute pain and primary dysmenorrhoea all over the world.

Celecoxib is further approved for the use in familial adenomatous polyposis in the US and leads to a reduction in the number of colorectal polyps in these patients [276].

The use of celecoxib is associated with a reduced incidence of gastroduodenal ulcers in comparison to naproxen [277], ibuprofen, and diclofenac [278] in arthritic patients. Because of a greater COX-1 inhibiting component as compared to the newer highly COX-2 selective compounds such as valdecoxib, rofecoxib, and paricoxib, cardiovascular side effects of celecoxib are less prominent and the compound avoided withdrawal from the market [279]. Nevertheless

celecoxib should be used with appropriate precaution.

Plasma peak concentrations are achieved within 2 h and the elimination half-life is about 12 h. Celecoxib is given orally (200–400 mg/d).

Trade names: Celebrex (EU, USA), Onsenal (EU).

Etoricoxib [*202409-33-4*], [*202409-40-3*] (mono HCl), 5-chloro-3-[4-(methyl-sulfonyl)phenyl]-2-(6-methylpyridin-3-yl)pyridine, 5-chloro-6′-methyl-3-(4-[methylsulfonyl)phenyl]-2,3′-bipyridine, $C_{18}H_{15}ClN_2O_2S$, M_r 358.848, crystals, *mp* 271.5-138.1 °C (136.7 °C DSC onset).

Synthesis [280–284]:

a) Synthesis via dichloropyridine:
   i) The bromination of 5-chloro-2-hydroxypyridine with Br_2 in acetic acid gives 3-bromo-5-chloro-2-hydroxypyridine, which is treated with benzyl bromide and Ag_2CO_3 in hot benzene to yield the benzyl ether 2-(benzyloxy)-3-bromo-5-chloropyridine. Condensation of 2-(benzyloxy)-3-bromo-5-chloropyridine with 4-(methylsulfanyl)phenylboronic acid with the aid of $Pd(PPh_3)_4$ and Na_2CO_3 in refluxing ethanol/benzene affords 2-(benzyloxy)-5-chloro-3-[4-(methylsulfanyl)phenyl]pyridine, which is oxidized with OsO_4 and sodium sulfite to furnish sulfone 2-(benzyloxy)-5-chloro-3-[4-(methylsulfonyl)phenyl]-pyridine. Treatment of 2-(benzyloxy)-5-chloro-3-[4-(methylsulfonyl)phenyl]pyridine with trifluoroacetic acid (TFA) provides the 2-hydroxypyridine derivative, which is reacted with $POCl_3$ to yield 2,5-dichloro-3-[4-(methylsulfonyl)phenyl]-pyridine.
   ii) Bromination of 2-amino-5-chloropyridine with Br_2 in acetic acid provides 2-amino-3-bromo-5-chloropyridine, which is condensed with 4-(methylsulfanyl)phenylboronic acid in the presence of $Pd(PPh_3)_4$ and Na_2CO_3 in refluxing ethanol/benzene to give 5-chloro-3-[4-(methylthio)phenyl]pyridin-2-amine. Oxidation of 5-chloro-3-[4-(methylthio)phenyl]pyridin-2-amine with OsO_4 as before yields the sulfone 5-chloro-3-[4-(methylsulfonyl)phenyl]-pyridin-2-amine, which is converted into 2,5-dichloro-3-[4-(methylsulfonyl)phenyl]pyridine by treatment first with $NaNO_2$ and HCl and then chlorination with $POCl_3$ [280, 281].
   iii) 2,5-dichloro-3-[4-(methylsulfonyl)phenyl]-pyridine is condensed with either trimethyl(6-methyl-3-pyridyl)tin or the boronate ester lithium salt by means of $Pd(PPh_3)_4$ to afford etoricoxib .The metalated pyridine 2-methyl-5-(trimethylstannyl)pyridine is obtained by esterification of 3-hydroxy-6-methylpyridine with trifluoromethanesulfonic (triflic) anhydride to give the corresponding triflate 6-methylpyridin-3-yl trifluoromethanesulfonate, which is treated with hexamethylditin to afford the target tin intermediate 2-methyl-5-(trimethylstannyl)pyridine. The boronate lithium salt is prepared by treatment of 5-bromo-2-methylpyridine with *n*-butyl lithium followed by addition of triisopropyl borate [281] (see next page).

b) In addition, etoricoxib is obtained by several related ways via the ketosulfone key intermediate 1-(6-methylpyridin-3-yl)-2-[4-(methylsulfonyl)phenyl]ethanone, which can also be synthesized by several different routes:
   i) Reaction of ketosulfone 1-(6-methylpyridin-3-yl)-2-(4-[methylsulfonyl)phenyl]ethanone with 2-chloro-1,3-bis(dimethylamino)trimethinium hexafluorophosphate salt in the presence of an equimolar amount of potassium *tert*-butoxide, followed by treatment with acetic/trifluoroacetic acid and then heating at reflux with an excess of ammonium hydroxide. The 2-chloro-1,3-bis(dimethylamino)trimethinium hexafluorophosphate salt is obtained by reaction of chloroacetic acid with hot dimethylformamide and $POCl_3$, and then the reaction mixture is treated with 5 N NaOH and hexafluorophosphoric acid in water.
   ii) Cyclization of ketosulfone 1-(6-methylpyridin-3-yl)-2-(4-[methylsulfonyl)phenyl]ethanone with 2-chloromalondialdehyde (A: X=OH), or by cyclization of ketosulfone 1-(6-methylpyridin-3-yl)-2-[4-(methylsulfo-

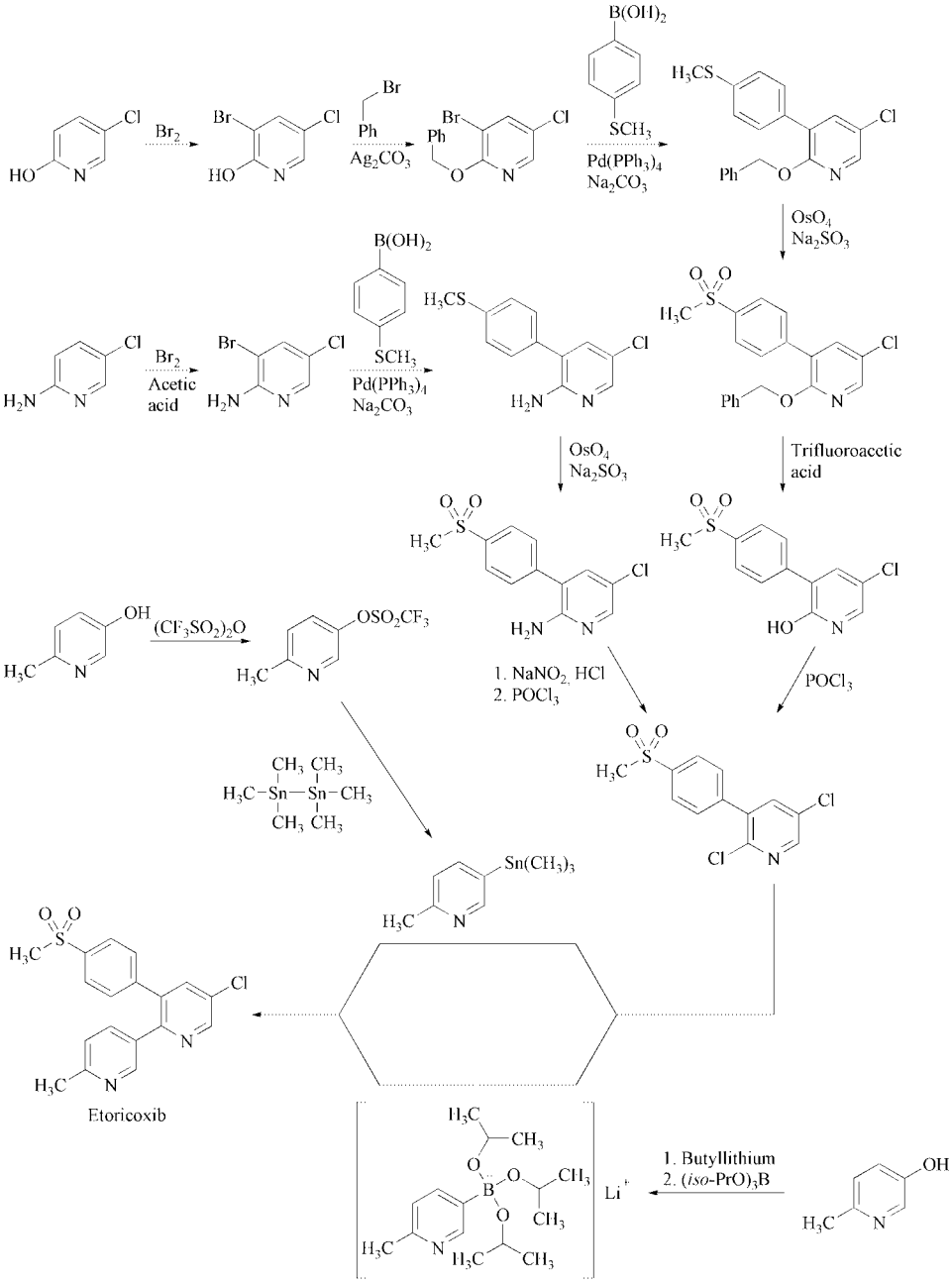

nyl)phenyl]ethanone with aminoacrolein (B: X=NH_2) in the absence of ammonium acetate. The cyclization of the lithium enolate of ketosulfone 1-(6-methylpyridin-3-yl)-2-(4-(methylsulfonyl)phenyl)ethanone with 2,3-dichloroacrolein (C: X=Cl) – obtained by treatment of chloromalondialdehyde with oxalyl chloride and dimethyl formamide in toluene – followed by reaction with ammonium acetate or anhydrous ammonia yield also the desired etoricoxib.

iii) Etoricoxib can also be obtained by cyclization of the aniline derivative 2-chloro-3-

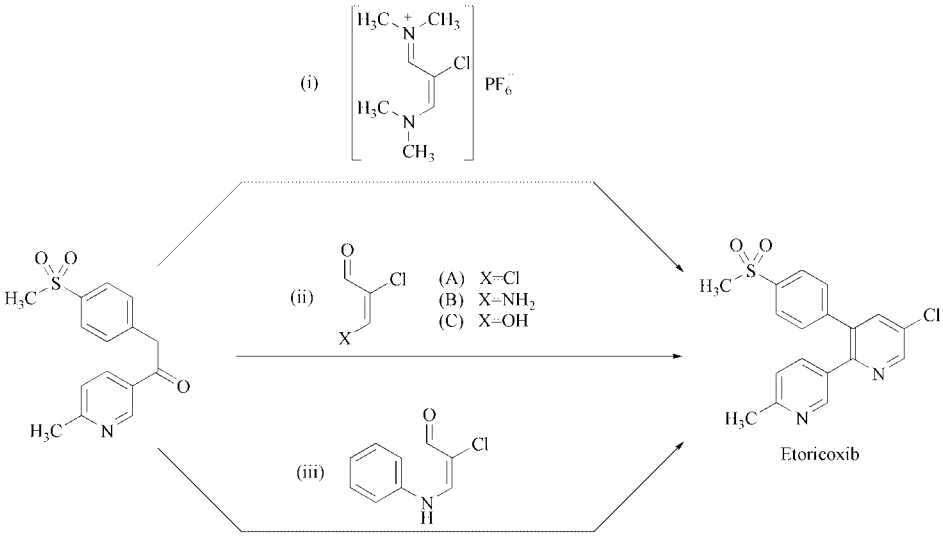

phenylaminopropenal and ammonium acetate in hot propionic acid [280–284].

Clinical use: Etoricoxib [280, 285] is a third-generation COX-2 inhibitor with a 100-fold selectivity for COX-2 in a whole blood assay. The selectivity ratio is higher than with other coxibs such as rofecoxib, valdecoxib, or celecoxib.

The compound was shown to be active in rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, postoperative pain, and in chronic low back pain. Further positive results were seen in patients with acute gout and against primary dysmenorrhoea [286].

Etoricoxib is used in oral doses of 60–120 mg/d. The compound is rapidly absorbed and peak plasma concentrations are achieved within 1.0 to 1.5 h after administration in healthy volunteers [287]. The elimination half-life is about 15 h. Etoricoxib is metabolized to about 60 % by members of the cytochrome P450 3A family [288]. In vitro studies give no evidence for active metabolites with respect to COX-1 and COX-2 inhibition [289].

Etoricoxib is generally well tolerated and induces less gastrointestinal side effects than conventional COX-1 inhibitors or non-COX selective NSAIDs [290]. Renal tolerability is similar to conventional COX-1 inhibitors [291]. According to its COX-2 selectivity, etoricoxib has an increased risk of cardiovascular side effects [292] and should only be used with appropriate precautions.

Trade name: Arcoxia (EU, USA).

Lumiracoxib [*220991-20-8*], 2-[2-(2-chloro-6-fluorophenylamino)-5-methylphenyl]acetic acid, $C_{15}H_{13}ClFNO_2$, M_r 293.72, *mp* 158-159 °C.

Synthesis [61, 294, 295]: The partial reduction of 4-methylanisole with sodium in liquid ammonia/THF/ethanol gives the enol ether 3-methoxy-6-methylcyclohexa-1,4-diene, which is condensed with 2-chloro-6-fluoroaniline in presence of $TiCl_4$ in chlorobenzene/THF to yield (*E*)-2-chloro-6-fluoro-*N*-(4-methylcyclohex-2-enylidene)benzenamine, which, without isolation, is aromatized with I_2 in AcOH/THF to provide *N*-(2-chloro-6-fluorophenyl)-*N*-(4-methylphenyl)amine. The acylation of this intermediate with 2-chloroacetyl chloride at 90 °C affords the 2-chloro-*N*-(2-chloro-6-fluorophenyl)-*N*-*p*-tolylacetamide, which is cyclized by means of $AlCl_3$ by heating at 160 °C to afford 1-(2-chloro-6-fluorophenyl)-5-methylindolin-2-one. Finally, this compound is hydrolyzed with NaOH in refluxing ethanol/water and acidified with 1 *N* HCl to give lumiracoxib. Alternatively, the intermediate *N*-(2-chloro-6-fluorophenyl)-*N*-(4-methylphenyl)amine can also be obtained by condensation of 2-chloro-*N*-(4-methylphenyl)acetamide with 2-chloro-6-fluorophenol using K_2CO_3 in isopropanol to yield 2-(2-chloro-6-fluorophenoxy)-*N*-(4-methylphenyl)acetamide, which is treated with MeONa in methanol to obtain the target secondary amine Lumiracoxib [61, 294].

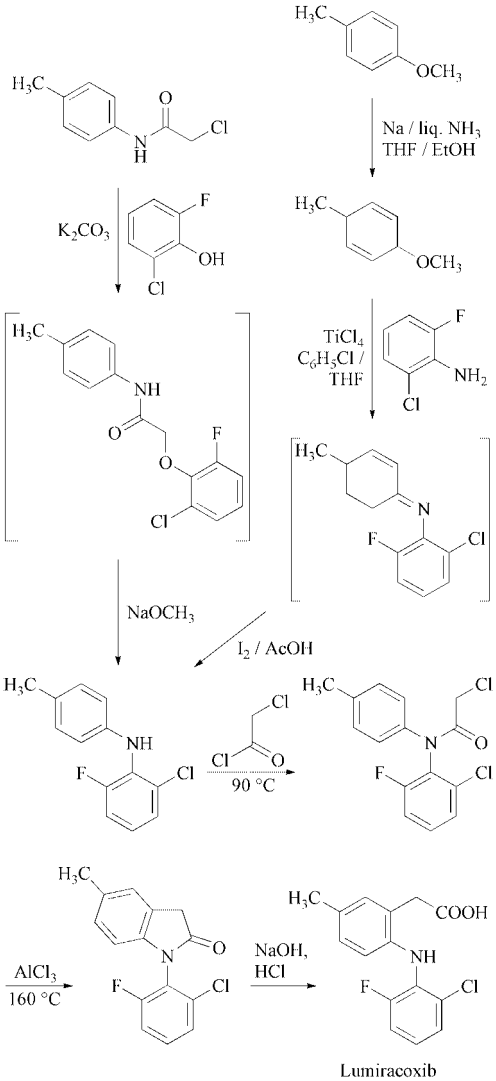

An alternative synthesis of lumiracoxib involves the coupling of *N*,*N*-dimethyl-5-methyl-2-iodophenylacetamide with 2-chloro-6-fluoroaniline in the presence of copper powder, copper(I) iodide and anhydrous potassium carbonate, and refluxing in xylene for 48 h. This synthetic route includes the following steps: The reduction of 2-iodo-5-methylbenzoic acid with BH_3/THF in THF gives 2-iodo-5-methylbenzyl alcohol, which is treated with refluxing 48 % HBr to yield the benzyl bromide. Reaction of the benzyl bromide with NaCN in ethanol/water affords the phenylacetonitrile, which is hydrolyzed with NaOH in refluxing EtOH/water to provide the phenylacetic acid. Reaction of 2-(2-iodo-5-methylphenyl)acetic acid with $SOCl_2$ in refluxing dichloromethane gives the corresponding acyl chloride, which is treated with dimethylamine in diethyl ether/THF to yield 2-(2-iodo-5-methylphenyl)-*N*,*N*-dimethylacetamide. Condensation of this intermediate with 2-chloro-6-fluoroaniline by means of Cu powder, Cu_2I_2 and K_2CO_3 in refluxing xylene affords 2-[2-(2-chloro-6-fluorophenylamino)-5-methylphenyl]-*N*,*N*-dimethylacetamide, which is finally hydrolyzed with NaOH in refluxing butanol/water to give lumiracoxib [61, 294, 295].

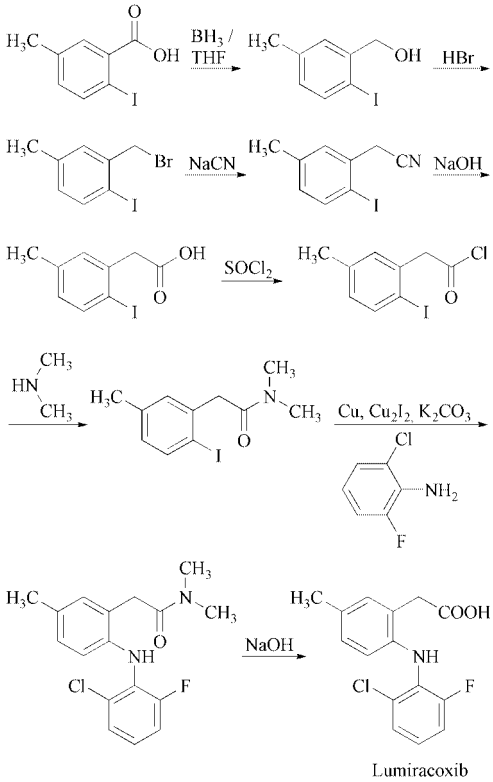

Clinical use: Lumiracoxib [296] is a novel COX inhibitor with the highest COX-2 selectivity of all coxibs. The compound is structurally related to the phenylacetic acid COX inhibitors, has weakly acidic properties and is devoid of the sulfonamide or methylsulfone group of the other coxibs. The compound showed clinical efficacy against osteoarthritis [297] and rheumatoid arthritis [298] and inhibited acute pain related to primary dysmenorrhoea, dental or orthopedic surgery, or tension-type headache [299].

The compound is well tolerated and gastrointestinal side effects were significantly lower than after treatment with COX-1 inhibitors like ibuprofen or naproxen [300, 301] The meta-analysis of all clinical trials in osteoarthritis and rheumatoid arthritis gave no indication of an increased risk of cardiovascular side effects [302]. Liver function tests showed a moderate and reversible increase in liver enzymes [303].

Lumiracoxib has a high oral bioavailability of about 74 %, is rapidly absorbed, reaches its peak plasma concentrations within 2 h and is strongly bound to plasma protein. The compound is strongly metabolized, is devoid of active metabolites and has a plasma elimination half-life of about 4 h [304]. Despite the short half-life, a once-daily administration in doses of 100–400 mg is recommended. A dose of 100 mg is used for the symptomatic relief of osteoarthritis and 400 mg for short-term relief of moderate to severe pain, associated with dental or orthopedic surgery. As with other COX-2 inhibitors, the compound should be used with appropriate precaution.

Trade name: Prexige (EU, USA).

Parecoxib [*198470-84-7*], (free acid) [*198470-85-8*], *N*-(4-(5-methyl-3-phenyl-isoxazol-4-yl)phen-ylsulfonyl)propionamide sodium salt $C_{19}H_{17}N_2O_4SNa$, M_r 392.409, crystals, *mp* 271.5–272.7 °C.

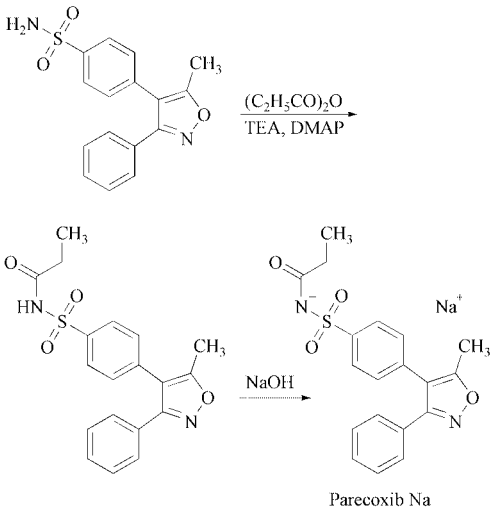

Synthesis: The acylation of 4-(5-methyl-3-phenylisoxazol-4-yl)benzenesulfonamide (valdecoxib), with propionic anhydride in a solution of triethanolamine (TEA) and 4-dimethylaminophenol (DMAP) in tetrahydrofuran gives *N*-[4-(5-methyl-3-phenylisoxazol-4-yl)phenylsulfonyl]propionamide, which is treated with NaOH in ethanol to give the sodium salt of parecoxib [79, 306, 307].

Clinical use: Parecoxib [77, 78, 80] is a third-generation COX-2 inhibitor. Parecoxib is a prodrug of valdecoxib with aqueous solubility sufficient for the use in parenteral formulations. The compound induced potent pain inhibition in dental, gynecological and orthopedic surgery and produced less gastrointestinal side effects than conventional COX-1 inhibitors [81].

Parecoxib is rapidly hydrolyzed in the liver to its active metabolite valdecoxib. Plasma peak concentrations for valdecoxib are achieved within 1.1 to 3.5 and 0.27 to 2 h after i.m. and i.v. administration, respectively. The elimination half-life for parecoxib is 15 to 35 min and 5 min for i.m. and i.v. administration, respectively. Metabolism of parecoxib follows the metabolism of the active metabolite valdecoxib which is a substrate for CYP3A4 and CYP2C9.

Parecoxib was used in doses of 20 and 40 mg of its sodium salt in the short-term treatment of postoperative pain.

Parecoxib together with valdecoxib was withdrawn from the market because of an increased risk of cardiovascular side effects [312, 313].

Trade names: Dynastat, Rayzon, Xapit (compound withdrawn from EU and US markets).

Rofecoxib [*162011-90-7*], 4-[4-(methylsulfonyl)phenyl]-3-phenylfuran-2(5*H*)-one, $C_{17}H_{14}O_4S$, M_r 314.36.

Synthesis: Condensation of phenylacetic acid with ethyl bromoacetate in the presence of triethylamine in THF yields ethyl 2-(phenylacetoxy)acetate, which is cyclized to the hydroxyfuranone using potassium *tert*-butoxide in *tert*-butanol. The reaction with trifluorosulfonyl anhydride and diisopropylethylamine in CH_2Cl_2 affords the corresponding trifluorosulfonate, which by reaction with LiBr in hot acetone yields the bromofuranone. The coupling with 4-(methylsulfanyl)phenylboronic acid with Na_2CO_3 and $Pd[(C_6H_5)_3P]_4$ as catalysts in hot toluene gives 4-[4-(methylsulfanyl)phenyl]-3-phenylfuran-2(5*H*)-one, which is finally oxi-

dized with $2KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$ (oxone) [314].

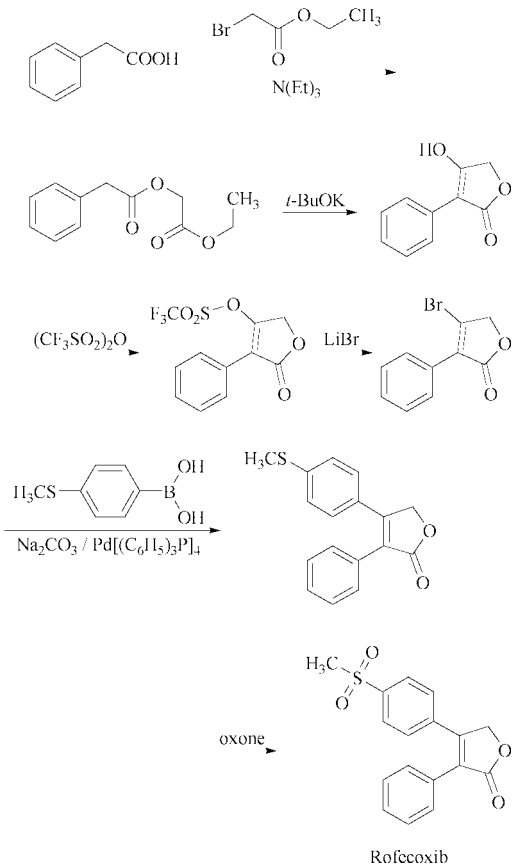

Rofecoxib can also synthesized by several different routes [156]. In 2001 a highly efficient synthesis of rofecoxib was described [316]. As illustrated in the following scheme, the acetophenone (**i**) is prepared according to the literature by Friedel–Crafts acylation with thioanisole. Oxidation with magnesium monoperoxyphthalate hexahydrate (MMPP) affords the sulfone (**ii**), which is reacted with bromine in chloroform in the presence of a trace amount of $AlCl_3$ to give (**iii**). Bromoketone (**iii**) is then coupled and cyclized in a second step, in a one-pot procedure with phenylacetic acid. The mixture of bromoacetophenone (**iii**) and phenylacetic acid in acetonitrile was then treated with triethylamine at room temperature to provide the ester intermediate, subsequent cooling and addition of dibutylurea (DBU) effected the cyclization to provide rofecoxib as the final product.

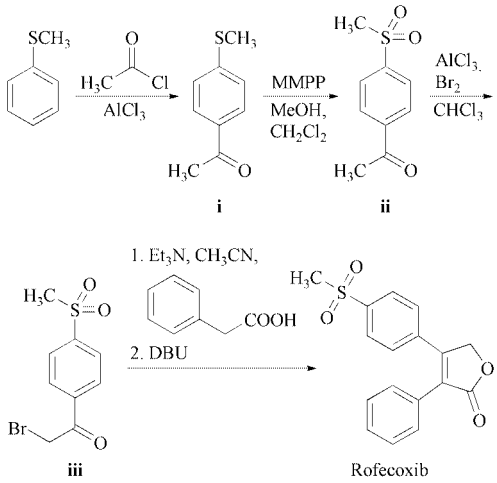

Clinical use: Rofecoxib [156, 317, 318] is a second-generation selective COX-2 inhibitor. It was the second COX-2 selective drug to reach the market in 1999. The selectivity for COX-2 compared to COX-1 is more than 800-fold in cellular assays and more than 10-fold in whole blood assays [319].

Rofecoxib is used for treatment of rheumatoid arthritis, osteoarthritis (12.5–25 mg/d) and pain of different origin such as postoperative pain, dental pain, and primary dysmenorrhoea (50 mg/d).

Rofecoxib reaches peak plasma concentrations between 2 to 9 h after oral administration. It is bound about 87 % to plasma protein and has an elimination half-life of about 17 h. Therefore a once daily dosing is appropriate. The main metabolites formed in the liver are the *cis*-dihydro and *trans*-dihydro derivatives of rofecoxib. Rofecoxib is excreted in the form of its metabolites mainly in the urine (72 %) with some unchanged drug excreted in the feces (14 %).

Rofecoxib shows significantly less gastrointestinal toxicity than ibuprofen in studies with osteoarthritis patients [320] and compared to naproxen in patients with rheumatoid arthritis [321]. Rofecoxib does not prolong bleeding time.

Rofecoxib was withdrawn from the market at the end of 2004 after it had been shown that use for a period of more than 18 month increased

the risk of heart attack and stroke [31]. It is being discussed with the health authorities that a relaunch should be possible under appropriate safety regulations [322].

Trade name: Vioxx (EU, USA; currently withdrawn from the market).

Valdecoxib [*181695-72-7*], 4-(5-methyl-3-phenylisoxazol-4-yl)benzenesulfonamide, $C_{16}H_{14}N_2O_3S$, M_r 314.366; *mp* 172-173 °C.

Synthesis: Deoxybenzoin is converted to the corresponding oxime by treatment with hydroxylamine under basic conditions with sodium acetate in aqueous ethanol or in toluene in presence of potassium hydroxide in absolute ethanol. The treatment of the oxime under nitrogen with two equivalents of butyllithium in tetrahydrofuran is followed by cyclization in ethyl acetate or acetic anhydride to the isoxazoline derivative. Finally, treatment of the isoxazoline with cold chlorosulfuric acid followed by reaction of the intermediate with aqueous ammonia affords the desired product [79–81, 156, 316, 323].

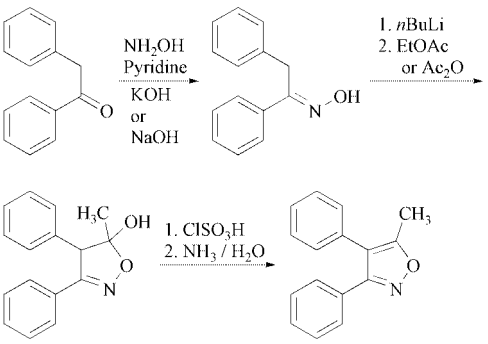

Clinical use: Valdecoxib [79, 324–326] is a third-generation COX-2 inhibitor. It shows about 30-fold selectivity for COX-2 in a whole-blood assay. The compound has a very low solubility in water and can only be administered orally. For parenteral administration, a water-soluble prodrug derivative, parecoxib, was developed in parallel. Valdecoxib showed strong analgesic and anti-inflammatory properties and was developed for the treatment of osteoarthritis, rheumatoid arthritis [327] and acute and chronic pain of various origin [328] including migraine headache [329].

Valdecoxib is converted in rodents and dogs, and in a low abundance in humans, by hydroxylation of the methyl group to an active metabolite (4-(5-hydroxymethyl-3-phenylisoxazol-4-yl)benzenesulfonamide).

Valdecoxib induced severe allergic skin reactions [330, 331] which, together with an increased risk of cardiovascular side effects [332] was the reason why the compound, together with its prodrug derivative parecoxib, was withdrawn from the market in 2004.

Trade name: Bextra (EU, USA; currently withdrawn from the market).

3. Centrally Acting Analgesics

3.1. Opioids

The term "centrally acting analgesics" is used for compounds which inhibit the pain reaction predominantly within the central nervous system. Most of them induce a very powerful pain inhibition and are thus named synonymously "strong analgesics". The most important representatives of this group are the opioids [333]. In respect to structural features [334] opioids can be separated into three groups:

- The first group contains the natural products morphine, codeine, and thebaine, which are isolated from opium. The group contains in addition various semisynthetic derivatives of the three compounds, which are prepared by chemical modifications of these natural products.
- The second group encompasses fully synthetic compounds which often have a total different chemical structure than the semisynthetic analogues, but interact with the same opioid receptors and show the same spectrum of analgesia and side effects as the natural compounds. For both groups together the older name "opiates" is still in use.
- The third group consists of naturally occurring and synthetic peptides with opioid-like properties. The opioid peptides were discovered during the search for endogenous ligands of the opioid receptors and share the same action and side effect profile with the nonpeptidic compounds, but they are not in clinical use.

Opioids interact with specific receptors (receptor ligands) and they can act as pure agonists, as partial agonists (agonists with a reduced intrinsic activity) or as antagonist (binding without intrinsic activity). Only few compounds of the opioid family are selective ligands for a single type of opioid receptor (Table 3). Most of them bind with similar or different affinities to more than one receptor type and the pharmacological effect is the result of the combined effects on all receptors involved. Opioids may act as a full or partial agonist on one receptor type and as an antagonist on the other. These compounds are named mixed agonist – antagonists and they are an important subgroup of opioid analgesics (see Section 3.1.2.1). Opioid analgesics are the most important drugs for the treatment of moderate, severe, and very severe pain.

Humans and most animal species are equipped with opioid receptors and endogenous ligands. This endogenous opioid system [336] is widely distributed within the body, it is phyllogenetically very old and is expressed in all vertebrates. A high density of opioid receptors is found in the brain [337] and in the spinal cord, where it is involved in pain inhibition and additionally in many other central regulatory processes. In addition to the CNS localization, opioid receptors are expressed in many peripheral organs [338]. Of great importance are the opioid receptors of the gastrointestinal system, which regulate stomach emptying, gut motility, and intestinal fluid secretion. Opioid receptors are found in cells of the immune system and peripheral opioids seem to be involved in the regulation of inflammatory and immunological processes [339]. In addition to the outstanding role in the central pain inhibition, action at *peripheral* opioid receptors, which are expressed in high quantities during inflammation and immune stimulation, may add to central pain inhibition [340].

Opioid Receptor Types. Differences in analgesic activity and in the side effect profile of synthetic opioids reinforced speculations that more than one type of opioid receptor exists and is involved in the analgesic activity of these compounds. Martin and coworkers investigated in 1960 these differences in a specially developed test model, the chronic spinal dog [341] and postulated three types of opioid receptors, the μ-receptor (ligand = morphine), the κ-receptor (ligand= ketazocine) and the σ-receptor (ligand = SKF 100 81). This was later confirmed by binding experiments with radioactive-labeled ligands and by the different binding and action profiles of the endogenous opioid peptides, the enkephalins and endorphines [342]. Up to now, three types of receptors, the μ-, κ-, and δ-opioid receptor have been confirmed [343, 344], the σ-receptor is no longer considered to be an opioid receptor.

More recently, a fourth opioid receptor type, named opioid receptor like (ORL1) receptor has been identified [345]. The receptor was detected as a c-DNA, which coded for a protein with opioid receptor-like properties. Within a short time, the endogenous ligand, a peptide named nociceptin, was isolated, which, depending on the place and route of application, induced pronociceptive or anti-nociceptive actions. The full spectrum of biological activity of nociceptin and the physiological role of the ORL-1 opioid receptor in pain processing is still under investigation [367].

Subtypes of the Different Opioid Receptors. Corresponding to other receptor systems, binding studies and functional investigations indicate that subtypes of opioid receptors exist [346, 347]. Within the μ-receptor two subtypes, the μ-1 and μ-2 receptor have been described. It was postulated by some investigators [347] that analgesia and opioid side effects are differently distributed between μ-receptor subtypes, which should allow separation of analgesia from the unwanted opioid side effects. According to Pasternak and Wood [347], analgesia should be mediated by the μ-1 receptor site, whereas respiratory depression and addiction is mediated via the μ-2 receptor subtype. But in contrast to binding experiments, the functional separation of the μ-1 and μ-2 subtype is more equivocal and a clear separation of analgesia from respiratory depression and addiction potential had never been found within the μ-opioids. Therefore, a differentiation of μ-opioid analgesics according to subtype specificity is no longer maintained.

The subtypes of μ-opioid receptors could not be confirmed in cloning experiments [348]. Therefore, possible heterogeneity of opioid receptor subtypes must result from a later modification which is independent from the gene level.

Possible variations could include splice variants, receptor association, posttranslational modifications (e.g., glycosidation) or coupling with different transduction mechanisms.

Endogenous Opioid Peptides. For each class of opioid receptors endogenous ligands of peptidic structure exist, which show a distribution within the body corresponding to that of the receptor. Three distinct families of ligands with various members have been identified and named endorphins, enkephalins, and dynorphins (Table 2). The opioid peptides are split off in the organism from specific precursor polypeptides. Enkephalins are derived from proenkephalin and interact with μ- and δ-opioid receptors. The precursor molecule of endorphin is pro-opiomelanocortin (POMC), which additionally releases peptidic hormones such as α-MSH, ACTH and β-LPH. Endorphins are specific ligands of the μ-opioid receptor type. Dynorphins interact with the κ-opioid receptor and are derived from pro-dynorphin. ORL binding sites have recently been identified together with its endogenous peptidic ligand named nociceptin (Table 2).

Various further natural and synthetic opioid peptides, acting as agonists or antagonist at the different opioid receptors have been discovered. They have the same effect and side effect profile as their nonpeptidic counterparts. Opioid peptides are only used experimentally and no compound has reached clinical application.

Molecular Pharmacology of Opioid Receptors. The first successfully cloned opioid receptor was the δ-opioid receptor of the mouse, which was in parallel described by the groups of Kieffer [349] and Evans [350]. Both used neuroblastoma-glioma hybridoma cells which expressed a high density of δ-receptors in their membrane. The isolated receptor protein consisted of 372 amino acids and had a molecular mass of 40 644. The amino acid sequence showed a partial overlap with the receptors of somatostatin (37 %), angiotensin (31 %) and interleukine-8 (22 %). Similar to the somatostatin receptor, seven hydrophobic domains were identified by which the receptor is inserted into the lipid bilayer of the cell membrane. Shortly thereafter the rat and human δ-receptors were cloned. With the same technique a novel opioid-like cDNA was isolated [351], which coded for an unknown opioid receptor. The new receptor had many similarities with the classical opioid receptors [352] and was added as the forth member to the opioid receptor family under the name ORL-1. All four opioid receptors have been identified in humans. All together, they show a high degree of structural identity, which corresponds to their widely overlapping biological functions [348, 353].

All opioid receptors belong to the group of pertussis toxin sensitive G-protein-coupled receptors of the rhodopsin family with seven transmembrane spanning hydrophobic domains. The N-terminal part is oriented to the outer side of the cell membrane and is involved in the selection and binding of the receptor specific ligands. Studies with chimeric or point mutated receptors indicate that predominantly the second and third extracellular loop determines receptor selectivity. The N-terminal sequence contains several free amino groups which can be conjugated with sugar residues. The carboxy-terminal is directed into the inner part of the cell and is involved in the signal transduction cascade. The carboxy-terminal contains groups which can be phosphorylated and which are involved in receptor internalization and inactivation. The seven transmembrane regions are connected by extracellular and intracellular loop regions of different length. Comparing the sequence of the μ-, δ-, and κ-receptor reveals that the highest degree of similarity is located in the transmembrane regions and in the intracellular loop, whereas the external loops and both terminal regions are more heterogeneous. The external loops and the terminal part are involved in the selection and binding of the ligands and contain the structural elements, which determine the receptor selectivity.

Transduction Mechanisms of Opioid Receptor Interaction. As described above, all four receptor types belong to the group of G-protein-coupled receptors (GPCRs) [354]. Agonistic binding at the receptor induces association of the α-, β-, and γ- subunit of the G-protein which triggers several biochemical reactions within the cell [355–358]. Most important for the pharmacological effects of the opioids are:

Table 2. Mammalian endogenous opioid peptides

Precursor	Endogenous peptide	Amino acid sequence
Pro-opiomelanocortin	β-endorphin	YGGFMTSEKSQTPLVTLFKNAIIKNAYKKGE
Pro-enkephalin	[MET]enkephalin	YGGFM
	[LEU]enkephalin	YGGFL
Pro-dynorphin	dynorphin A	YGGFLRRIRPKLKWDNQ
	dynorphin B	YGGFLRRQFKVVT
	α-neoendorphin	YGGFLRKYPK
	β-neoendorphin	YGGFLRKYP
Pro-nociceptin/OFQ [a]	Nociceptin	FGGFTGARKSARKLANQ

[a] OFQ = orphanin FQ.

- Activation of a hyperpolarizing K^+ channel (inward rectifying K^+ channel)
- Inactivation of voltage-dependent Ca^{2+} channels (N-, P- and R-type)
- Inhibition of adenylate cyclase

A further relevant mechanism involved in the inhibition of synaptic transmission by opioids is a direct impairment of the exocytotic release of neurotransmitters, induced by stabilization of the presynaptic membrane.

Additional actions with only partially understood relevance are activation of phospholipases (PLH2, PLC7), activation of MAP kinases and activation of some voltage dependent Ca^{2+} channels (L-type and T-type).

As a result of these actions opioids inhibit neurotransmission at the presynaptic and postsynaptic site. The presynaptic inhibition mostly depends on the direct inhibitory effect on transmitter exocytosis from membrane-associated storage vessels. This direct effect is increased by the inhibition of Ca^{2+} channels, since Ca^{2+} ions trigger the transmitter release. Activation of K^+ ions induces membrane hyperpolarization and this is the most important action component for postsynaptic inhibition.

Activation of K^+ channels, inactivation of Ca^{2+} channels and direct inhibition of neurotransmitter release are powerful mechanisms by which opioids inhibit the neuronal transmission of the pain signal.

Opioid receptors are not only located at excitatory synapses, but are likewise expressed at inhibitory neurons. At these synapses, opioids inhibit the transmission of the inhibitory signal, and this inhibition of inhibition as the result, induces excitation and an increased release of neurotransmitters in the innervated neuron.This explains why opioids, in addition to their prominent inhibitory actions, have some stimulant effects, too.

Relation of the Biological Effect to the Opioid Receptor Type. The peptide structure of the four different opioid receptors is very similar [348] and this is the reason why most of the opioid receptor ligands show affinity for more than one type of receptor. In most cases, one type is preferred and the action and side effect profile is dominated by the properties of this receptor. A huge chemical effort was necessary to develop receptor-specific "selective" ligands, but nowadays several selective agonists and antagonists for each type of opioid receptor exist (Table 3).

μ-Opioid Receptors. Binding experiments and investigations in pain models have shown that predominantly the μ-receptor and to a lesser degree the κ-and d-receptors mediate pain inhibition. Most of the clinically used opioid analgesics [359] have prevalence for the μ-opioid receptor and this confirms that μ-receptor activation is the common pain inhibiting mechanism of these compounds. Since morphine is the prototype, they are called "morphine-like"analgesics. Their side effect profile includes respiratory depression, constipation as well as addiction and dependence. All these effects, corresponding to analgesia, are mediated via μ-opioid receptor activation. This is the reason why the primary goal of the development of synthetic opioids, the separation of analgesia from the addiction and dependence potential, has never been achieved. The involvement of the same receptor population in analgesia and side effect profile of morphine and morphine-like opioids was impressively confirmed in μ-knockout mice [360,

Table 3. Opioid receptor specifity

Receptor type	Selective ligands		Nonselective ligands	
	Agonist	Antagonist	Agonist	Antagonist
MOR	morphine	CTOP	levorphanol	naloxone
μ-	fentanyl		etorphine	naltrexone
	methadone			
	DAMGO			
KOR	enadoline	nor-BNI	levorphanol	naloxone
κ-	U50,488		etorphine	naltrexone
	dynorphin			
DOR	DPDPE	naltrindol	levorphanol	naloxone
δ-	SNC-80		etorphine	naltrexone

MOR = mu-opioid receptor. KOR = kappa-opioid receptor. DOR = delta-opioid receptor. DAMGO = [D-Ala^2, MePhe^4, Gly(ol)^5]enkephalin. Enadoline = (5*R*)-(5α,7α,8β-(-)-*N*-methyl-*N*-(7-[1-pyrrolidinyl]-1-oxaspiro[4,5]*dec*-8-yl)-4-benzofuranacetamide, previously CI977. U50,488 = *trans*-3,4-dichloro-*N*-methyl-*N*-[2-(1-pyrrolidinyl)-cyclohexyl]-benzeneacetamide methanesulfonate. DPDPE = [D-Pen^2, D-Pen^5]enkephalin. SNC-80 = (+)-4-[(α*R*)-α-((2*S*,5*R*)-4-allyl-2,5-dimethyl-1-piperazinyl)-3-methoxybenzyl]-*N*,*N*-diethylbenzamide. CTOP = D-Phe-Cys-Tyr-D-Trp-Orn-Thr-Pen-Thr-NH_2. nor-BNI = nor-binaltorphimine.

361]. In these animals, morphine did neither show an analgesic effect nor side effects. There was no respiratory depression and no inhibition of gastrointestinal motility and secretion. In behavioral models in μ-knockouts, morphine did not induce liking or any other signs of addiction and repeated treatment with morphine did not produce signs of tolerance and physical dependence.

***κ**-Opioid Receptors.* Activation of κ-receptors induces a clinically relevant pain inhibition [362, 363], which seems to be less efficacious than μ-receptor mediated analgesia. κ-Receptor interaction is a common action component of the so-called partial opioid agonists or agonists – antagonists, such as pentazocine, nalbuphine, and butorphanol (see Section 3.1.1.1). The most prominent side effects of κ-activation are sedation and diuresis. In contrast to μ-agonists, which induce well-being and euphoria, activation of κ-receptors in the limbic system induces dysphoria and other unpleasant psychic effects, such as hallucinations and spatial disorientation. The dose range of analgesia and psychic side effects partly overlap and this interferes with the medical use of κ-receptor agonists. In compounds with a marked κ-agonistic component (cyclazocine, nalorphine) the dysphoric side effects are so prominent that the compounds could not be used for therapeutic purposes. This may explain why despite an intensive search for κ-selective analgesics, no selective κ-agonist has ever reached the clinical use.

***δ**-Opioid Receptors.* In contrast to μ- and κ-receptors, indications for a prominent role of δ-receptors in the pain process are less obvious [364, 365]. There is a broad but still controversial discussion concerning a genuine analgesic effect mediated by δ-1 or δ-2 receptors, because analgesia is scarcely observed in compounds with pure δ-agonistic activity.

ORL-1 Receptors. The ORL-1 receptor [352, 366] differs in its peptide sequence from the classical opioid receptors. In contrast to these, ORL-1 receptor activation at least at the supraspinal level has a pronociceptive effect and induces pain. Supraspinally induced release of spinal CCK-8, NMDA or PGE-2 is discussed as the mechanism of pronociceptive activity. Spinal application of nociceptine incontrary induces anti-nociception. With the aid of nonpeptidic ORL-1 agonists and antagonists, which are currently under development, it will be possible to elucidate the contribution of the ORL-1 opioid system to pain processing and to other physiological processes [367].

Location of Opioid Receptors in the Pain Pathway. The endogenous opioid system [368] is the most important component of the pain inhibitory system of the body. Opioids act at different levels of the pain pathway and their action interferes with different aspects of the pain processing [369, 370].

Opioids induce:

- Inhibition of the transmission of the pain signal

- Inhibition of the emotional aspect of pain
- Inhibition of pain realization

Pain processing is inhibited at the spinal and supraspinal level. Spinal opioid receptors are located pre- and postsynaptically at interneurons of the *substantia gelatinosa* of the dorsal horn. The opioid interneurons inhibit the release of excitatory transmitters and reduce the transmission of the pain signal from the primary afferents to the secondary neurons of the spinal ascending pain pathway. Supraspinally, opioids are located in different regions of the brainstem, in the periaquaeductal grey matter, in the limbic system, in the thalamic nuclei, in the basal ganglia, and in the cortex. Cortical and thalamic localization is involved in the perception of the pain stimulus, cortical regions in addition in the localization of the origin of the pain focus.

Opioids in the different parts of the limbic system suppress the emotional component of pain and the pain suffering. In the *formatio reticularis*, they inhibit the pain induced activation of autonomous functions, such as increase in respiration, increase in blood pressure and sweating. Inhibitory actions at the *formatio reticularis* are the origin of the most prominent side effects of the opioids, like respiratory depression, bradycardia and the centrally mediated part of the gastrointestinal inhibition.

In addition to the inhibitory effect at the ascending pain transmission, opioids activate a descending pain inhibitory system, which originates from different centers of the pons and medulla, such as *nucleus coeruleus*, areas of the periaquaeductal grey matter, and areas of the raphe nuclei. The descending nerve fibers terminate at the spinal interneurons in noradrenergic and serotoninergic inhibitory synapses, which suppress the ascending pain signal. Thus, opioids inhibit the spinal pain processing by two mechanisms, one is a direct pre- and postsynaptic inhibition of the ascending pain pathway and the other is a centrally-mediated activation of the descending pain inhibitory system.

Use of Opioids in Pain Treatment. Opioids are used for the treatment of moderate to severe or very severe pain of acute or chronic type [371]. Nearly all forms of pain are sensitive to opioid treatment and in contrast to traditional opinions even neuropathic pain is reasonably sensitive to higher doses of opioids. This was clearly shown in well-controlled clinical studies [372]. The most important use of opioids in acute pain treatment is postoperative pain, whereas treatment of cancer pain, often accompanied by a neuropathic pain component, is the classical domain of *chronic* opioid treatment.

Opioid Side Effects.

Respiratory Depression. Opioids induce respiratory depression via inhibition of the respiratory center of the *medulla oblongata*, which respond to the pCO_2 content of the blood [373, 374]. The inhibitory effect is more prominent in respect to the respiratory frequency than to the volume of respiration. At higher opioid dosages, respiration becomes irregular and grasping occurs. In awake persons, respiratory depression by opioids can be voluntarily compensated over a broad dose range. Respiratory impairment is not a prominent feature in awake pain patients, because pain itself is a strong stimulus for respiration.

Opioid-induced respiratory depression is augmented by other CNS-depressant compounds such as sedatives and hypnotics. Respiratory depression becomes an important side effect when opioids are used for postoperative pain treatment, because the anesthetic agent and most adjuncts of anesthesia induce a long lasting depressant effect on respiration, which can increase the opioid effects up to respiratory arrest. Therefore a careful supervision of respiration during the postoperative period is mandatory [375]. Opioid-induced respiratory depression can be interrupted by the opioid antagonist naloxone. Naloxone has a short duration of action and repeated applications may be necessary to successfully counteract the effect of longer acting opioids. Highly potent opioids induce severe respiratory depression in the higher dose range. In addition, they induce a stiffness of the chest musculature [376], called chest rigidity or "wooden chest", which is mediated via stimulation of dopamine release in the *nucleus caudatus*. Chest rigidity further increases the respiration impairment by these compounds. Therefore higher doses of potent opioids such as fentanyl and analogues, as used for anesthesia, need artificial or at least assisted ventilation.

Cardiovascular Effects. Nearly all opioids induce bradycardia [377], most likely mediated

via central stimulation of the vagus nerve. The cardiovascular depression of most opioids is moderate and only the stronger opioids of the fentanyl group induce a more prominent effect. Morphine and some analogues induce a non-opioid receptor mediated release of histamine, which can result in a decrease in blood pressure and a compensatory tachycardia.

Gastrointestinal Effects. Opioids induce an inhibitory effect on gastrointestinal motility and fluid secretion [378]. The effect is peripherally and centrally mediated. The peripheral part is related to μ- and κ-receptors at intestinal organs, which are densely equipped with opioid receptors. The receptors are located at parasympathic ganglions and inhibit the release of acetylcholine, which stimulates the contraction of the smooth muscles. Inhibition of the intestinal fluid secretion is mediated via inhibition of adenylate cyclase. The intestinal effects of opioids extend to all parts of the gut and results in inhibition of stomach emptying and inhibition of secretion and motility in the duodenum, jejunum, colon, and rectum.

Reduced motility and secretion can let to constipation, which is the most prominent side effect of chronic opioid treatment [379]. Opioid-induced constipation can increase up to megacolon or paralytic ileus. Therefore chronic opioid treatment should be accompanied by concomitant use of laxatives. Besides their peripheral actions, opioids are involved in the **central** regulation of intestinal functions which is located in the *formatio reticularis*. This explains why the intestinal side effects of opioids are not restricted to the more hydrophilic compounds such as morphine, but are seen likewise in the more centrally active lipophilic analogues. During chronic opioid treatment a varying degree of tolerance towards the intestinal side effects may occur.

The intestinal inhibitory action of opioids can be used for the treatment of diarrhea [380]. The clinically most important antidiarrheal opioid is loperamide [381]. After oral administration, loperamide acts locally within the gastrointestinal tract. After parenteral administration, the compound is rapidly inactivated and does not reach the CNS. Therefore loperamide does not show the typical central opioid side effects, has no analgesic action and has no abuse potential.

Emetic Activity. Nausea and emesis are common unpleasant side effects of opioids [382, 383]. They are most intensively experienced at the beginning of the treatment. During chronic application, tolerance may occur, which reduces the emetic sequelae. Nausea and emesis are induced via activation of chemoreceptors, which are located in the trigger zone of the area postrema of the *formatio reticularis.* The receptors are located at the tissue surface and contact the circulating blood. Thus the emetic effect of opioids is not mediated centrally, which means after penetration of the blood – brain barrier, but rather peripherally via the part of the compound, which is distributed in the circulating blood.

After passage through the blood – brain barrier, opioids have an antiemetic effect [384]. The emesis inhibition is induced via blockade of an emesis center located in a more central area of the *formatio reticularis*. This explains why the emetic effect of opioids is most prominent immediately after application, especially after rapid intravenous administration and is reduced or terminated when the compound has reached the CNS. The more hydrophilic opioids such as morphine have stronger emetic side effects than lipophilic compounds like methadone or fentanyl [385], which are rapidly transported into the CNS.

Tolerance and Dependence. μ-Opioid compounds induce a feeling of well being and euphoria, which is mediated by the release of dopamine within the limbic system. κ-Opioids induce an opposite effect with dysphoria, disorientation, and hallucinations [386]. Repeated activation of the μ-opioid rewarding system may induce a psychological dependence, which leads to addiction and a compulsory behavior of drug seeking [387]. In addition, higher opioid dosages, as used for nonmedicinal purposes, induces tolerance and as a consequence a further increase of the dose to achieve the intended effect. In the course of the tolerance development opioid users becomes physically dependent on the supply of the compound [388] and suspension of the treatment or blocking of the opioid receptors with an antagonist induces a withdrawal reaction, characterized by strong dysphoria, restlessness, pain, and various symptoms of autonomic dysregulation, like diarrhea, shivering, chills and cardiovascular collapse.

The euphorigenic effect of opioids, the "opioid kick", is more intensely induced by lipophilic compounds such as diacetylmorphine (heroin), which rapidly penetrates into the CNS. The feeling of euphoria on the one hand and the absence of well-being on the other hand is more prominent when brain concentrations of the opioid change rapidly. This intensifies drug seeking behavior and psychological as well as physical dependence.

In contrast to recreational use, chronic pain treatment with opioids afforts constant dosing and has only a limited risk to induce psychological dependence and drug addiction [389]. A proper dosing during chronic pain treatment can often postpone tolerance development for longer time periods. The most important precaution to avoid tolerance and dependence development is to ensure constant plasma levels of the opioid, which have to be high enough to give complete pain relief. This can be reasonably achieved by oral administration of retarded formulations or by an opioid patch [390]. Breakthrough pain, which is often induced by a fluctuation in pain intensity, should be rapidly compensated by additional treatment with an immediate release formulation of the same or a similar opioid.

3.1.1. Opioids in Clinical Use

3.1.1.1. Morphine and Morphinane Derivatives

Morphine [*57-27-2*], (5α,6α)-7,8-didehydro-4,5-epoxy-17-methylmorphinan-3,6-diol, $C_{17}H_{19}NO_3$, M_r 285.34, *mp* 254 °C (decomp., 197 °C is reported for a metastable phase), $[\alpha]_D^{25}$ $-132°$ (c = 1, CH_3OH); hydrochloride [*52-26-6*], $C_{17}H_{19}NO_3 \cdot HCl$, M_r 321.80, *mp* 200 °C (trihydrate), $[\alpha]_D^{25}$ $-113.5°$ (c = 2.2, water); sulfate (2 : 1) [*64-31-3*], $C_{17}H_{19}NO_3 \cdot 1/2\ H_2SO_4$, M_r 668.76; sulfate pentahydrate [*6211-15-0*], $C_{17}H_{19}NO_3 \cdot 1/2\ H_2SO_4 \cdot 5H_2O$, M_r 758.85, *mp* 250 °C (decomp.), $[\alpha]_D^{25}$ $-108.7°$ ($c = 4$, water). Morphine is the main alkaloid of opium, isolated from the milky liquid of the poppy seed capsule [108].

Preparation: Morphine is obtained by extraction of poppy capsules or opium (opium contains 9–14 % morphine, depending on the source) with water, precipitation with aqueous Na_2CO_3-solution, washing of the precipitate with ethanol and dissolving in diluted acetic acid [392–395].

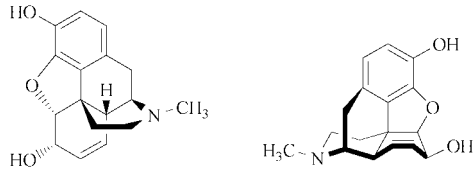

Morphine in the conventional presentation (left) and in a stereoformula according to IUPAC rules (right)

The total synthesis of (-)-morphine had been a challenging target for organic chemists for many decades [316, 334, 396]. Although a number of successful syntheses were developed, only a few could produce the compound in an enantio- and diastereo-controlled manner [397–399].

Morphine biosynthesis: The studies on the biosynthesis of morphine have been performed on cell cultures mainly of *Coptis japonica* and species of *Thalictrum*. The overall biosynthetic pathway from tyrosine to morphine is depicted in Figure 3.

Opioid receptor binding: Morphine has a high (nanomolar) binding affinity for the μ-opioid receptor. The affinity for the δ- and κ-opioid receptor is at least 10-fold lower.

Clinical use: Morphine is a very potent analgesic and still the "gold standard" for the treatment of severe, acute, and chronic pain [400]. The compound is used in various salt forms, preferentially as the hydrochloride or sulfate. Morphine can be administered orally or parenterally. Oral administration is preferred for chronic pain treatment and various slow-release forms have been developed to reduce the application frequency to 1–2 times per day. Parenterally, morphine is used in doses of 10 mg mostly for postoperative pain and self-administration devices have been developed for patient-controlled analgesia (PCA).

Other morphine preparations are under evaluation for local administration, e.g., inhalation via the broncho-pulmonal way.

Pharmacokinetic properties: Morphine is extensively metabolized by glucuronidation at both hydroxyl groups at positions 3 and 6 and by *N*-demethylation [401, 402]. The 3-glucuronide has a minimal opioid receptor binding affinity

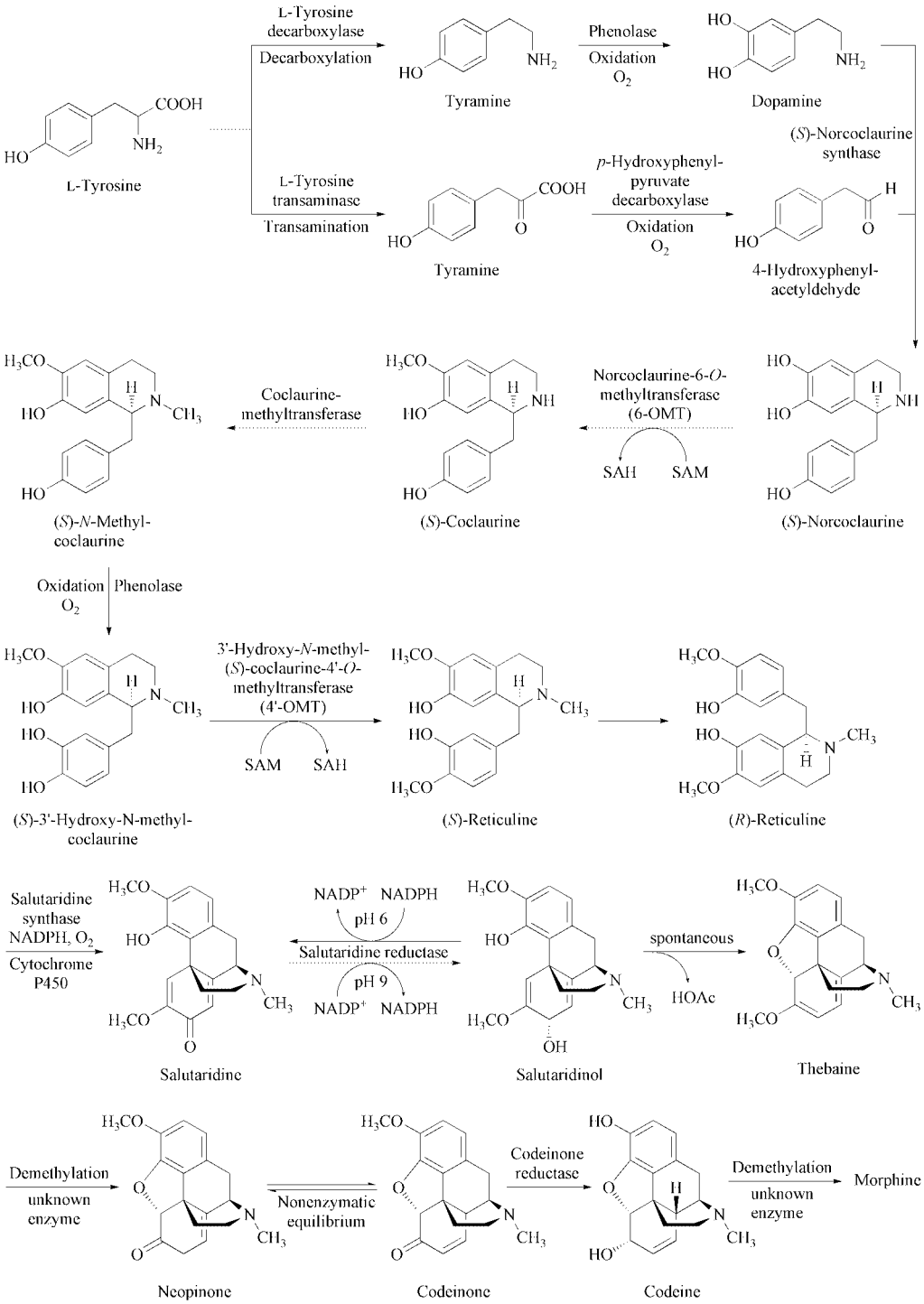

Figure 3. The overall biosynthetic pathway from tyrosine to morphine

and is devoid of analgesic action, whereas the morphine-6-glucuronide has a similar binding affinity as morphine, is analgesically active and seems to be involved in pain inhibition during chronic oral treatment with morphine [403]. Despite the polar sugar residue the glucuronide can pass the blood – brain-barrier. The morphine-6-glucuronide is being tested as an independent opioid analgesic. The *N*-demethyl derivative normorphine is analgesically active, but has a lower μ-receptor affinity, a lower potency and a shorter duration of action than morphine. Normorphine is not in clinical use.

The main metabolites of morphine are shown in Figure 4.

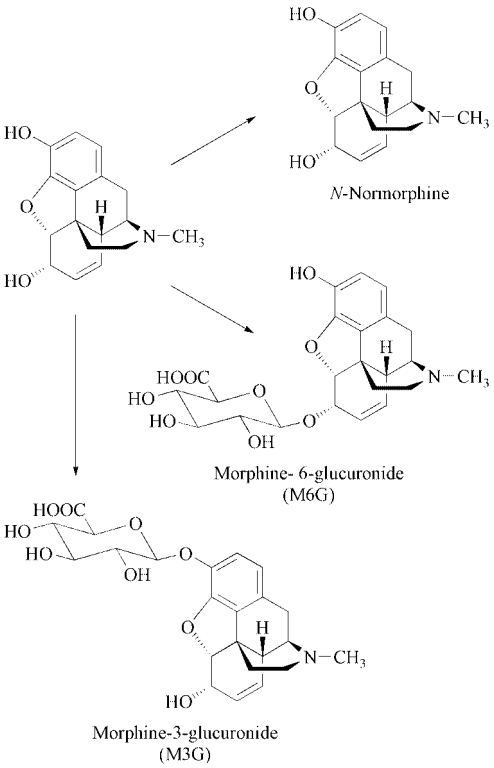

Figure 4. Metabolic pathway of morphine

Side effects: Morphine induces a variety of centrally and peripherally mediated side effects. Most important after parenteral administration is respiratory depression, especially in the postoperative situation. Chronic oral administration induces constipation and chronic treatment with oral morphine has to be supplemented with laxatives. Other frequent side effects are nausea, vomiting, dizziness, and sedation.

Morphine is a controlled substance. It has a high euphorigenic potential and is liable to abuse and dependence. The euphorigenic effect is less expressed in the context of pain treatment and tolerance and dependence can largely be avoided by appropriate dosing and application intervals, securing constant and pain-appropriate plasma levels of the compound.

Trade names: M-Long (Germany), MST continuous (Germany, UK), Capros (Germany), Kapanol (Germany, USA), Moscontin (France), Infumorph (USA), Oxamorph (USA).

Morphine is also marketed in many generic formulations.

Codeine [*76-57-3*], $C_{18}H_{21}NO_3$, M_r 299.36, *mp* 154–156 °C (monohydrate from water or diluted alcohol), anhydrous form sublimes at 140–145 °C and 133.32 Pa, $[\alpha]_D^{25}$ −136° (c = 2, CH_3CH_2OH); monohydrate [*6059-47-8*], $C_{18}H_{21}NO_3 \cdot H_2O$, M_r 317.38, *mp* 154–156 °C; hydrochloride dihydrate $C_{18}H_{21}NO_3 \cdot HCl \cdot 2\ H_2O$, M_r 371.86, *mp* 280 °C (decomp.), $[\alpha]_D$ − 108°; hydrobromide dihydrate [*125-25-7*], $C_{18}H_{21}NO_3 \cdot HBr \cdot 2\ H_2O$, M_r 416.31, *mp* 190–192 °C (anhydrous), $[\alpha]_D$ − 96.6°; phosphate [*52-28-8*].

Preparation: Codeine is extracted from opium (concentration in opium 0.7 to 2.5 %, depending on the source), but mostly prepared by methylation of morphine in a phase transfer reaction [404, 405].

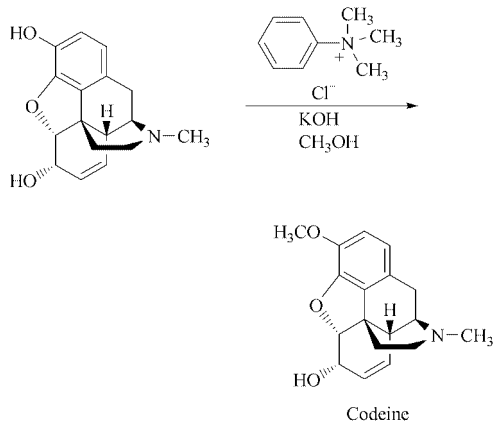

Opioid receptor binding: Codeine has a low affinity at μ-, δ-, and κ-opioid receptors and the in vivo effects are predominantly induced by metabolization to morphine [55].

Clinical use: Codeine [407] is an O-methyl analogue of morphine. It has a morphine-like action profile but a 5–10 times lower potency. Codeine has a good oral availability and oral administration forms are used for the treatment of mild to moderate pain and as antitussive (→ Cough Remedies, Section 2.1.1). In therapeutic dosages, codeine produces fewer opioid-type side effects than morphine and has only a low risk of abuse. The lower abuse and dependence can be explained by the fact that the opioid action is only available after metabolic activation. High codeine doses can induce excitement and convulsions.

The compound is extensively metabolized by *O*- and *N*-demethylation followed by glucuronidation. The main metabolites are norcodeine, morphine, and hydrocodeine and their glucuronides. There are indications [408] that the analgesic effect is reduced in poor metabolizers with a low CYP2D6 activity.

Codeine is used in the form of different salts like hydrochloride, phosphate and sulfate. To increase the duration of action, slow-release preparations have been developed. Codeine salts are frequently combined with non-opioid analgesics such as paracetamol, acetylsalicylic acid, or diclofenac.

Trade names: Codipront (Germany), Codicaps (Germany), Codimal (USA).

Combination with paracetamol: Co-Tylenol (USA).

Combination with Diclofenac: Combaren (USA).

Codeine is also marketed in many generic formulations.

Ethylmorphine [*76-58-4*], (5α,6α)-7,8-didehydro-4,5-epoxy-3-ethoxy-17-methylmorphinan-6-ol, $C_{19}H_{23}NO_3$, M_r 313.40, *mp* 199–201 °C; hydrochloride dihydrate [*6746-59-4*], $C_{19}H_{23}NO_3 \cdot HCl \cdot H_2O$ M_r 385.89, *mp* 123 °C (decomp.), anhydrous form melts at 170 °C (decomp.).

Synthesis: Ethylmorphine is synthesized by ethylation of morphine with ethyl benzenesulfonate [48, 409, 410].

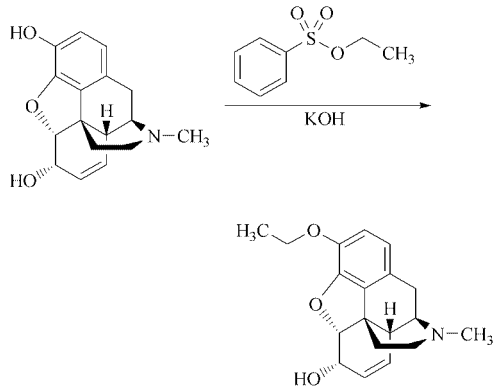

Ethylmorphine

Clinical use: Ethylmorphine (→ Cough Remedies, Section 2.1.1) is the ethyl congener of codeine and has a similar therapeutic application. It has a low opioid receptor affinity and is metabolized to the active metabolites morphine and normorphine. Ethylmorphine can be used as cough suppressant, medium-potent analgesic, and anti-diarrheal agent. Today the compound is widely out of therapeutic use, but is often used as clinical reagent to characterize the metabolic N-dealkylation activity of CYP3A4 [411, 412].

Trade names: Codethyline (Belgium), Trachyl (France), Collins Elixir (UK).

Diacetylmorphine [*561-27-2*], diamorphine, heroin, $C_{21}H_{23}NO_5$, M_r 369.41, *mp* 173 °C, $[\alpha]_D^{25}$ −166° (c = 1.49, CH_3OH); hydrochloride monohydrate [*561-27-2*], $C_{21}H_{23}NO_5 \cdot HCl \cdot H_2O$, M_r 423.89, *mp* 243–244 °C, $[\alpha]_D^{25}$ −156° (c = 1.044).

Synthesis: Diamorphine is prepared by acetylation of morphine with acetic anhydride [413].

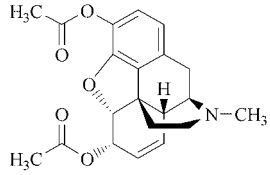

Diacetylmorphine

Opioid receptor binding: Diamorphine [414] has a 10- to 100-fold lower μ-opioid receptor binding affinity than morphine. The relevant opioid properties originate from the high μ-receptor

affinity of the metabolites 6-acetylmorphine and morphine [415].

Clinical use: Diamorphine [416] is a lipophilic morphine derivative which rapidly penetrates into the central nervous system, where 6-acetylmorphine and morphine are released [417, 418]. The rapid brain access induces an immediate onset of action and this seems to be the reason for the strong euphorigenic effect and the high abuse potential. Because of the widespread abuse, therapeutic application is prohibited in many countries including Germany and the US. In other countries like the UK, parenteral heroin is used for the relief of severe pain, especially in terminal illness. The compound is more potent than morphine, but has a similar action and opioid side effect profile.

Trade name: Diagesil (UK).

Dihydrocodeine [*125-28-0*], (5α,6α)-4,5-epoxy-3-methoxy-17-methylmorphinan-6-ol, $C_{18}H_{23}NO_3$, M_r 301.38, *mp* 112–113 °C; tartrate (1 : 1) [*5965-13-9*], $C_{18}H_{23}NO_3 \cdot C_4H_6O_6$, M_r 451.47, *mp* 192–193 °C (commercial medicinal grade usually melts at 186–190 °C, $[\alpha]_D^{25}$ −72° to −75° (*c* = 1.0, water).

Synthesis: Hydrogenation of codeine yields dihydrocodeine [48, 419, 420].

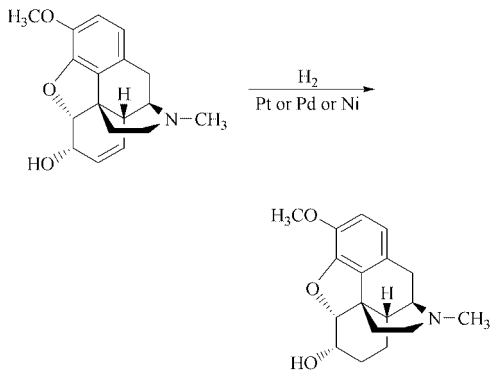

Dihydrocodeine

Opioid receptor binding: Dihydrocodeine has a low μ-opioid receptor binding affinity and opioid properties are mostly due to metabolic activation to the demethyl derivative dihydromorphine.

Clinical use: Dihydrocodeine has codeine-like analgesic and antitussive properties and is used for the treatment of moderate to severe pain [421] and as antitussive [422] (→ Cough Remedies, Section 2.1.1). Dihydrocodeine is mostly used as oral immediate- or sustained-release formulations [423]. For pain treatment the dose range is 30–80 mg, for cough inhibition doses are in the range of 10 mg.

Dihydrocodeine induces morphine-like side effects, but the intensity is less pronounced. Chronic treatment may produce dependence, and abuse by opioid addicts has been reported.

Trade names: Paracodin (Germany), Dicodin (France), DHC Continus (UK), Synalgos (USA).

Etorphine [*14521-96-1*], [5α,7α(*R*)]-4,5-epoxy-3-hydroxy-6-methoxy-α,17-dimethyl-α-propyl-6,14-ethenomorphinan-7-methanol, $C_{25}H_{33}NO_4$, M_r 411.53, *mp* 214–217 °C; hydrochloride [*13764-49-3*], $C_{25}H_{33}NO_4 \cdot HCl$, M_r 447.99, *mp* 266–267 °C.

Synthesis: Starting from thebaine etorphine can be synthesized in a similar way as buprenorphine (see page 66) [424–427].

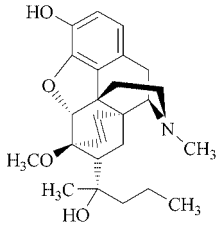

Opioid receptor binding: Etorphine [428] has a high affinity and selectivity for the μ-opioid receptor.

Clinical use: Etorphine [429] is an extremely strong and rapid-acting μ-opioid with a potency 400–1000-fold higher than morphine. The compound is very rapidly absorbed through skin and mucosa and extraordinary care is necessary to avoid contamination during medical use. Naloxone or diprenorphine can be used as antagonist. Etorphine induces strong central nervous depression and is mostly used in veterinary practice for immobilization or pain treatment of big animals [430]. Etorphine is only used parenterally.

Trade name: Immobilon (UK).

Hydrocodone [*76-42-6*], (5α)-4,5-epoxy-3-methoxy-17-methylmorphinan-6-one,

$C_{18}H_{21}NO_3$, M_r 299.36, *mp* 198 °C; hydrochloride [*25968-91-6*], $C_{18}H_{21}NO_3 \cdot HCl$, M_r 335.83; hydrochloride monohydrate [*124-90-3*], $C_{18}H_{21}NO_3 \cdot HCl \cdot H_2O$, M_r 353.85, *mp* 185–186 °C (decomp.), $[\alpha]_D^{27}$ −130° (*c* = 2.877); bitartrate hemipentahydrate [*34195-34-1*], $C_{18}H_{21}NO_3 \cdot C_4H_6O_6 \cdot 5/2\ H_2O$, M_r 494.5, *mp* 118–128 °C.

Synthesis: Palladium or platinum catalyzed isomerization of codeine yields hydrocodone [48, 431, 432].

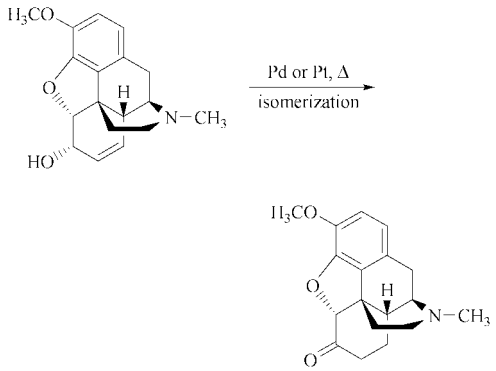

Hydrocodone

Opioid receptor binding: Hydrocodone has a μ-opioid receptor binding affinity that is 10- to 100-fold higher than that of codeine [433]. In contrast to codeine, the compound itself has a relevant μ-opioid receptor affinity and is, in addition to the active metabolite hydromorphone, responsible for analgesia and other opioid properties.

Clinical use: Hydrocodone [420] is an opioid related to codeine with analgesic and marked antitussive properties. The compound is more potent than codeine and is used in oral doses of 5 to 10 mg for the treatment of moderate to moderately severe pain. Combinations with paracetamol or other COX inhibitors are common, as well as multi-ingredient preparations as antitussives. Hydrocodone induces side effects similar to codeine and morphine and has an appreciable abuse and dependence potential [434].

Trade names: Dicodid (Germany, UK), Vicodin (USA).

Combination with paracetamol: Zydone (USA), combination with ibuprofen: Vicoprofen (USA).

Hydromorphone [*466-99-9*], (5α)-4,5-epoxy-3-hydroxy-17-methylmorphinan-6-one, $C_{17}H_{19}NO_3$, M_r 285.34, *mp* 266–267 °C; monohydrochloride [*71-68-1*], $C_{17}H_{19}NO_3 \cdot HCl$, M_r 321.80.

Synthesis [48, 435, 436]:

Morphine is hydrogenated over a palladium catalyst, and the resulting dihydromorphine is oxidized with benzophenone and potassium *tert*-butoxide. Alternative oxidants are cyclohexanone with aluminum tri(*tert*-butoxide) or aluminum triphenoxide.

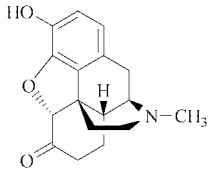

Opioid receptor binding: Hydromorphone has a high affinity and selectivity for the μ-opioid receptor, the μ-affinity is about 10-fold higher as compared to morphine.

Clinical use: Hydromorphone [437] is a morphine derivative with an analgesic potency higher than morphine. It is used as hydrochloride parenterally or orally in the dose range of 2–4 mg to treat moderate to severe pain [438] or as a syrup for cough inhibition. Hydromorphone shows the typical morphine-like side effect profile and has a relatively high potential for addiction and dependence [439].

Trade name: Dilaudid (Germany, UK, USA), Palladone (USA).

Levorphanol [*77-07-6*], 17-methylmorphinan-3-ol, (−)-3-hydroxy-*N*-methylmorphinan, $C_{17}H_{23}NO$, M_r 425.48, *mp* 198–199 °C, $[\alpha]_D^{20}$ −56° (*c* = 3, CH_3CH_2OH); tartrate dihydrate [*5985-38-6*], $C_{17}H_{23}NO \cdot C_4H_6O_6 \cdot 2\ H_2O$, M_r 257.37, *mp* 113–115 °C (anhydrous *mp* 206–208 °C), $[\alpha]_D^{20}$ −14° (*c* = 3, water).

Synthesis: The analgesic activity of racemorphan is due to the (−) isomer, levorphanol, which is obtained by resolving the racemate with (+)-D-tartaric acid. The resolution can also be performed on the intermediate 1-(4-methoxy benzyl)-1,2,3,4,5,6,7,8-octahydro isoquinoline (**1**) prior to N-methylation [48, 440–443].

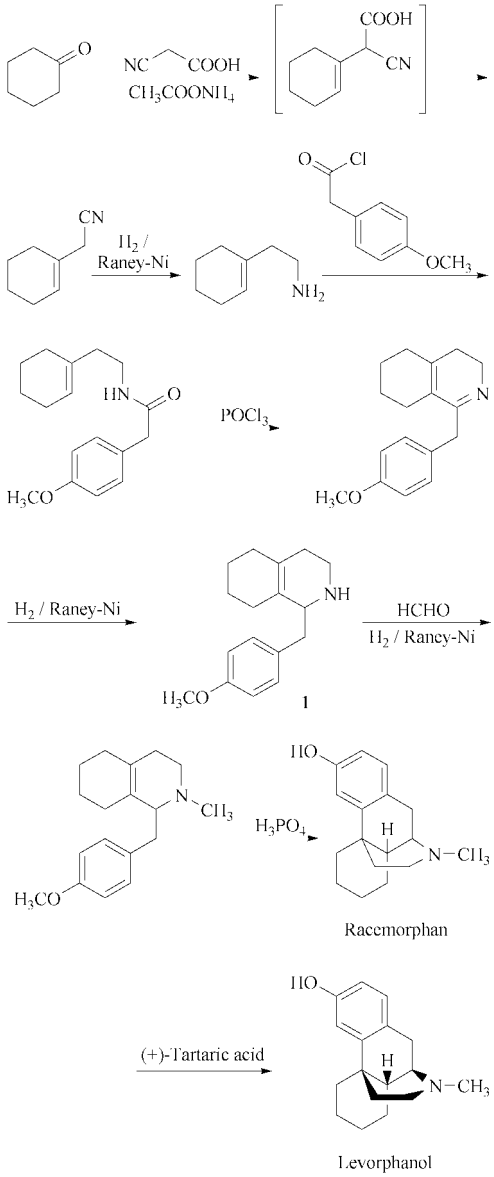

Opioid receptor binding: Levorphanol is a μ-selective synthetic opioid with a higher receptor affinity than morphine [225].

Clinical use: Levorphanol [445] is a strong analgesic used as tartrate for parenteral and oral administration. Parenterally, levorphanol is 4–5 times more potent than morphine. After oral administration, the compound has a long duration of action (up to 8 h). Levorphanol has a morphine-type side effect profile with prominent respiratory depression in the high dose range. The compound induces morphine-type addiction and dependence [446].

Trade name: Levo-Dromoran (USA).

Oxycodone [*76-42-6*], (5α)-4,5-epoxy-14-hydroxy-3-methoxy-17-methylmorphinan-6-one, 14-hydroxydihydrocodeinone, dihydrone, dihydrohydroxycodeinone, $C_{18}H_{21}NO_4$, M_r 315.36, *mp* 218-220 °C; hydrochloride [*124-90-3*], $C_{18}H_{21}NO_4 \cdot HCl$, M_r 351.81, *mp* 270–272 °C (decomp.), $[\alpha]_D^{20}$ −125° (c = 2.5, water).

Synthesis [48, 447, 448]: Thebaine is oxidized with hydrogene peroxide to 14-hydroxycodeinone [449, 450], which is hydrogenated directly or via its oxime, or its bromination products to oxycodone. The reduction of 14-hydroxycodeinone can also be performed with sodium hydrosulfite. Alternatively, 14-hydroxycodeinone is prepared by oxidation of codeine (see next page, top).

Opioid receptor binding: Oxycodone [433] is a μ-selective opioid with a 10-fold higher receptor affinity than codeine. Both the parent compound and the high affinity metabolite oxymorphone mediate the opioid effects of the compound [451].

Clinical use: Oxycodone [452] is an orally active opioid analgesic for the treatment of moderate to moderately severe pain. The compound is used as hydrochloride in a dose range of 5–10 mg and is often combined with paracetamol or other COX-inhibitors. Higher dosages and oral slow-release preparations [453] are used for the control of severe pain. Oxycodone has a morphine-like side effect profile. Respiratory depression has been found in children. The compound has a relevant abuse and dependence potential and illicit use of the retarded preparations is reported. To obviate misuse, abuse deterrent formulations (ADF) and combinations with naloxone and naltrexone are under development.

Trade names: Eukodal (Germany), Eubine (France), Proladone (UK),Roxicodone, OxyContine (USA), Oxygesic (Germany)

Combination with ibuprofene: Combunox (USA).

Oxymorphone [*76-41-5*], (5α)-4,5-epoxy-3,14-hydroxy-17-methylmorphinan-6-one, dihydrohydroxymorphinone, dihydro-14-hydroxymorphinone, 14-hydroxydihydromorphinone, $C_{17}H_{19}NO_4$, M_r 301.34, *mp*

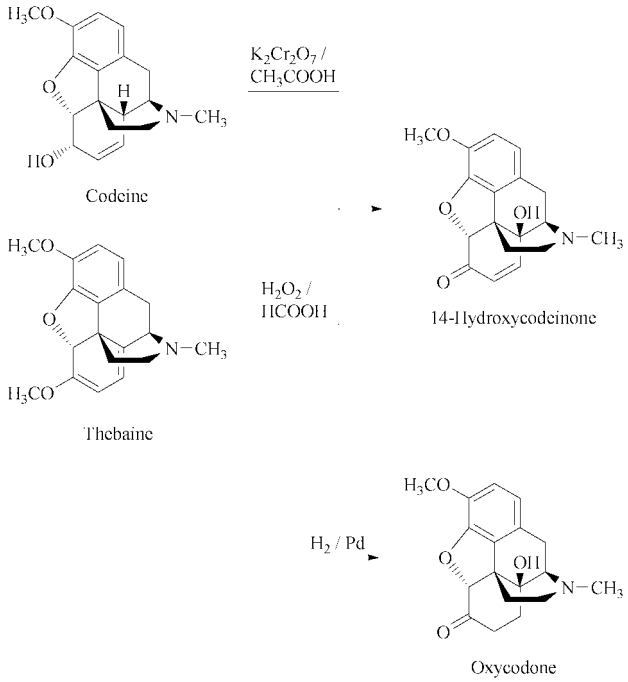

248–249 °C (decomp.); hydrochloride [*357-07-3*], $C_{17}H_{19}NO_4 \cdot HCl$, M_r 337.80.

Synthesis [454] Oxycodone is hydrolyzed with boiling concentrated hydrobromic acid.

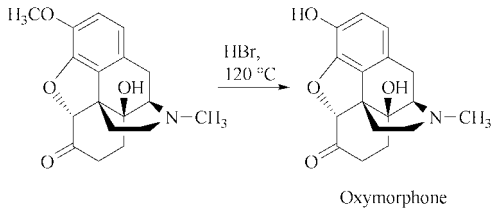

Opioid receptor binding: Oxymorphone is a μ-selective opioid with a binding affinity in the range of morphine.

Clinical use: Oxymorphone is a synthetic morphine derivative with a potency 5–10 times higher than that of morphine. The compound is used as hydrochloride in parenteral and rectal application forms (suppositories) and as oral sustained-release formulations to treat moderate to severe pain [455]. Oxymorphone has no cough suppressant activity. Oxymorphone has a morphine-type side effect profile and can induce addiction and dependence [456].

Trade name: Numorphan (USA, Canada).

3.1.1.2. Piperidine Derivatives

3.1.1.2.1. Meperidine and Congeners

(for Diphenoxylate, Loperamide, and Piritramide see Section 3.1.1.4)

Pethidine [*57-42-1*], meperidine, ethyl 1-methyl-4-phenylpiperidine-4-carboxylate, $C_{15}H_{21}NO_2$, M_r 247.33; hydrochloride [*50-13-5*], $C_{15}H_{21}NO_2 \cdot HCl$, M_r 283.80, *mp* 186–189 °C.

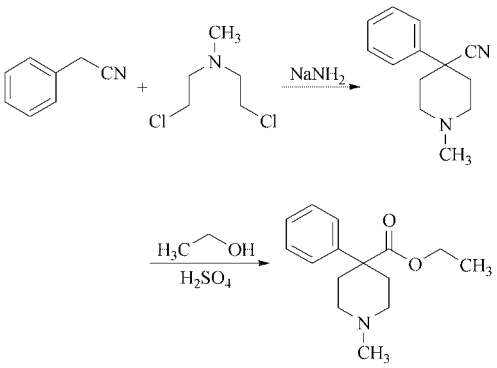

Synthesis [48, 457–460]:

a) The original synthesis involved condensation of benzyl cyanide with *N*,*N*-bis(2-chloroethyl)-*N*-methyl-amine, which is a skin irritant and a carcinogen.
b) Another synthesis begins with pyridine-4-carboxylic acid.

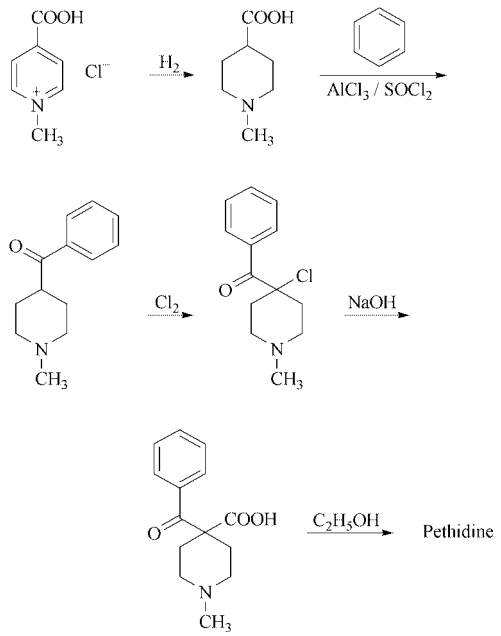

Opioid receptor binding: Pethidine is a μ-selective synthetic opioid with an intermediate receptor affinity [461].

Clinical use: Pethidine [462] is a synthetic analgesic with μ-receptor preference similar to morphine. Pethidine, introduced in 1939, was the first fully synthetic analgesic with morphine-like activity. It has a lower potency and a shorter duration of action than morphine and is used as hydrochloride for the treatment of moderate to severe pain. It is preferentially used in parenteral formulations, perioperative and obstetrical analgesia and as adjunct to anesthesia.

Pethidine has a morphine-type side effect profile with a lower incidence of constipation. Higher doses induce central stimulation and antimuscarinic effects accompanied by pupil dilatation. In even larger doses toxic symptoms such as muscular twitching, tremor, mental confusion, and convulsions occur, which are partly attributed to the more toxic metabolite norpethidine [463]. Severe side effects including coma and cyanosis have been observed in combination with MAO-inhibitors [464]. Pethidine induces morphine-type tolerance and dependence and addicts using high doses of pethidine have an increased risk of excitatory side effects.

Due to the high risk of severe side effects the clinical use of the compound has strongly decreased [465–467].

Trade names: Dolantin (Germany), Dolosal (France), Demerol (USA).

Ketobemidone [*469-79-4*], *N*-(3-hydroxyphenyl)-*N*-(1-methylpiperidin-4-yl)-propionamide, $C_{15}H_{21}NO_2$, M_r 247.34, *mp* 150–151 °C; hydrochloride [*5965-49-1*], $C_{15}H_{21}NO_2 \cdot HCl$, M_r 283.80, *mp* 201–202 °C.

Synthesis: The cyano group of 4-cyano-4-(3-methoxyphenyl)-1-methylpiperidine is converted into a ketoethyl group by reaction with ethylmagnesium bromide. Subsequent ether cleavage by means of HBr yields ketobemidone [48, 459, 468, 469].

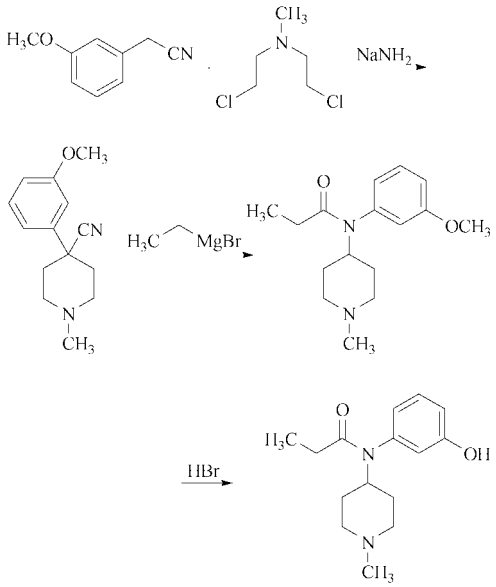

Ketobemidone

Opioid receptor affinity: Ketobemidone is a μ-selective synthetic opioid with a receptor affinity similar to morphine [470].

Clinical use: Ketobemidone is a synthetic opioid of the pethidine group with an analgesic potency in the range of morphine [471]. The

compound has a weak NMDA blocking effect [472], which may contribute to the analgesic efficacy. It can be given orally, by injection or rectally; usual doses are 5–10 mg. Ketobemidone has a morphine-like side effect and abuse and dependence potential [473].

Trade names: Cliradon (Germany, out of use), Ketogan (Sweden, Norway).

3.1.1.3. Fentanyl and Congeners

Fentanyl [*437-38-7*], *N*-(1-phenethylpiperidin-4-yl)-*N*-phenylpropionamide, $C_{22}H_{28}N_2O$, M_r 336.47, *mp* 83–85 °C; citrate (1 : 1) [*990-73-8*], $C_{22}H_{28}N_2O \cdot C_6H_8O_7$, M_r 528.60, *mp* 149–151 °C.

Synthesis [48, 474] (see right column).

Opioid receptor binding: Fentanyl is a μ-selective potent opioid with a similar receptor binding affinity as morphine. The higher in vivo potency of fentanyl results from its higher lipophilicity.

Clinical use: Fentanyl is a potent synthetic opioid related to pethidine [475] and has an action profile similar to morphine. Fentanyl citrate is used for the inhibition of acute and chronic severe pain, and especially as adjunct to anesthesia and as a primary anesthetic for induction and maintenance of anesthesia [476]. In combination with neuroleptics or tranquilizers it induces a surgical state named neuroleptanalgesia in which surgery can be performed in an awake, but calm and pain-free status. Fentanyl is orally bioavailable and can be administered orally (special forms exist for the oral-transmucosal route), parenterally, and transdermally as patch formulations. A formerly used less safe reservoir patch formulation has now been substituted by a new matrix patch. Fentanyl is a highly lipophilic compound and is bound to plasma proteins to about 80 %. The compound has a rapid onset and a relatively short duration of action, presumably due to tissue redistribution. The elimination half-life is longer than that of morphine.

The side effect profile is typical of potent μ-opioids with respiratory depression, increased muscle tone (chest wall rigidity during fentanyl anesthesia), strong sedation, and emesis being most prominent. Adverse reactions can be antagonized with naloxone.

Fentanyl and fentanyl derivatives (so-called designer drugs) have an essential abuse potential [477] and induce a morphine type of physical dependence.

Trade names: Actic (USA), Duragesic (USA), Durogesic, Durogesic-SMAT (Germany), Durogesic-Dtrans (UK), Fentanyl Janssen (Germany), Sublimaze (UK, USA).

Combination with Droperidol: Thalamonal.

Transdermal fentanyl patches are also marketed in generic formulations.

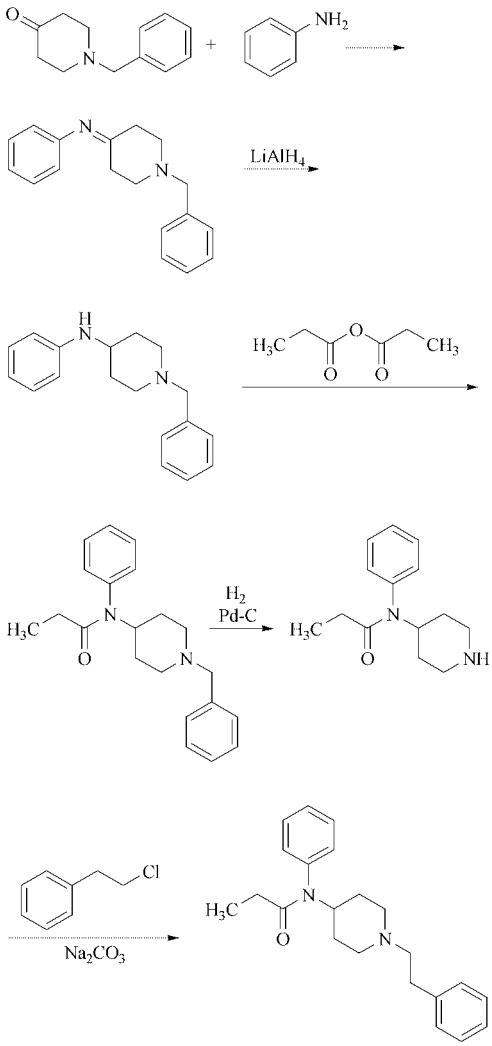

Fentanyl

Alfentanil [*71195-58-9*], *N*-{1-[2-(4-ethyl-5-oxo-4,5-dihydrotetrazol-1-yl)ethyl]-4-methoxymethylpiperidin-4-yl}-*N*-phenylpropionamide, $C_{21}H_{32}N_6O_3$, M_r 416,25; hydrochloride monohydrate [*70879-28-6*], $C_{21}H_{32}N_6O_3 \cdot HCl \cdot H_2O$, M_r 470.99, *mp* 138.4–140.8 °C.

Synthesis: The cyclization of ethyl isocyanate with sodium azide in the presence of $AlCl_3$ in refluxing THF gives 1-ethyl-1,4-dihydro-5*H*-tetrazol-5-one, which is alkylated with 1-chloro-2-bromoethane using Na_2CO_3 and KI in refluxing 4-methyl-2-pentanone to afford 1-ethyl-4-(2-chloroethyl)1,4-dihydro-5*H*-tetrazol-5-one (**2**) [48, 478–480].

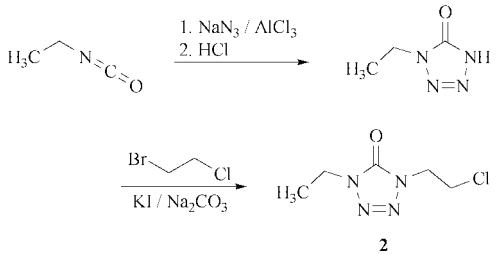

N-(4-methoxymethyl-4-piperidinyl)-*N*-phenylpropionamide (**3**) is synthesized according the following scheme starting from 1-benzyl-4-piperidone:

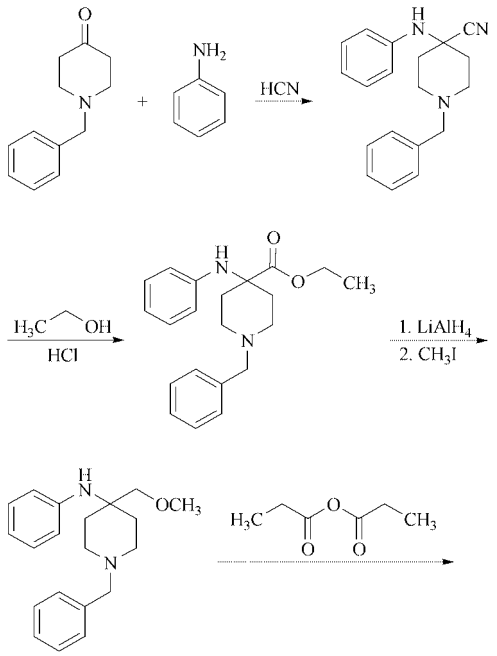

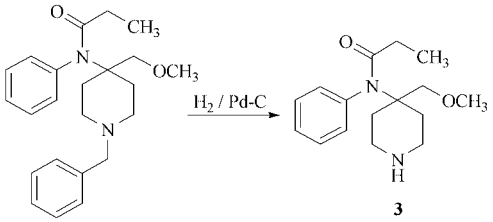

Finally the tetrazole derivative **2** is condensed with the propionamide **3** in the presence of Na_2CO_3 and KI as catalysts in refluxing 4-methyl-2-pentanone.

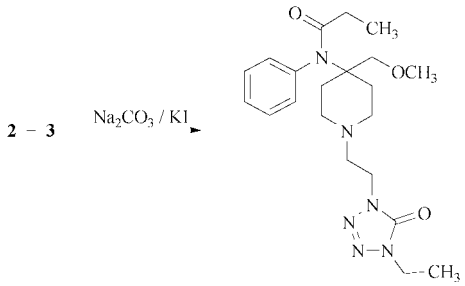

Alfentanil

Opioid receptor binding: Alfentanil is a μ-selective opioid [481] with a receptor affinity in the range of morphine and fentanyl.

Clinical use: Alfentanil is a potent opioid analgesic with a rapid onset and a shorter duration of action than fentanyl [482]. The i. v. formulation (dose range 1–4 mg) is mainly used perioperatively as strong analgesic, supplement to general anesthesia, or as a primary anesthetic [483]. Alfentanil has a strong respiratory depressant action and high doses induce chest wall rigidity. The compound has a μ-type addiction and dependence potential.

Trade names: Rapifen (Germany, France, UK), Alfenta (USA).

Remifentanil [*132875-61-7*], methyl 1-(2-methoxycarbonylethyl)-4-(phenylpropionylamino)piperidine-4-carboxylate, methyl 4-(methoxycarbonyl)-4-[(1-oxopropyl)phenylamino]-1-piperidinepropanoate, $C_{20}H_{28}N_2O_5$, M_r 376.45; monohydrochloride [*132539-07-2*], $C_{20}H_{28}N_2O_5 \cdot HCl$, M_r 412.91; oxalate (1 : 1) [*132875-62-8*], $C_{20}H_{28}N_2O_5 \cdot C_2H_2O_4$, M_r 466.49.

Synthesis: Remifentanil is prepared by condensation of methyl 4-(phenylpropionylamino) piperidine-4-carboxylate with methyl acrylate in hot acetonitrile [48, 484–486].

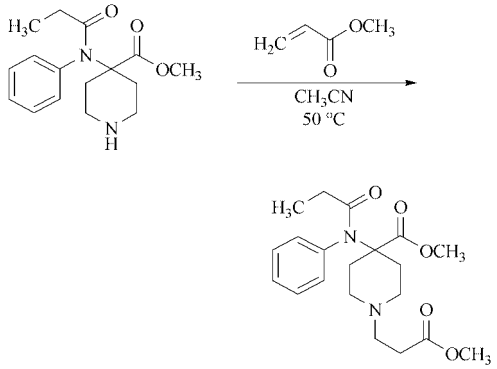

Remifentanil

Clinical use: Remifentanil [487, 488] is a very short-acting and potent μ-opioid agonist with strong analgesic and anesthetic properties. It has been developed as an ultra-short anesthetic and adjunct to general anesthesia. [489, 490]. It affords potent intraoperative analgesia and has a sparing effect on concomitantly used sedatives and hypnotics [491]. Remifentanil is given as intravenous short infusion in doses of 0.5–1 μg/kg or as continuous infusions in the range of 0.0025–2 μg $kg^{-1}min^{-1}$. It is rapidly inactivated by plasma and tissue esterases [492]. The terminal elimination half-life is 10–20 min.

Remifentanil has a μ-opioid type side effect profile with strong CNS and respiratory depressant properties and a morphine-like addiction and dependence potential.

Trade name: Ultiva (Germany, UK, USA).

Sufentanil [*56030-54-7*], *N*-[4-(methoxymethyl)-1-[2-(2-thienyl)ethyl]-4-piperidinyl]-*N*-phenylpropanamide, *N*-[4-methoxymethyl-1-(2-thiophen-2-yl-ethyl)piperidin-4-yl]-*N*-phenylpropionamide, $C_{22}H_{30}N_2O_3S$, M_r 386.56, *mp* 96.6 °C; citrate (1 : 1) [*60561-17-3*], $C_{22}H_{30}N_2O_3S \cdot C_6H_8O_7$, M_r 578.68.

Synthesis: Sufentanil is obtained by condensation of *N*-(4-methoxymethylpiperidin-4-yl)-*N*-phenylpropionamide with 2-thiophen-2-yl-ethyl methanesulfonate [48, 493–496].

Opioid receptor binding: Sufentanil is a μ-selective opioid with an about 10-fold higher receptor affinity than fentanyl [497].

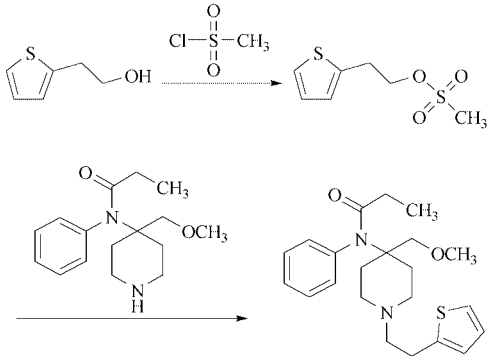

Sufentanil

Clinical use: Sufentanil [498] is a very potent and short-acting fentanyl derivative with preference for the μ-opioid receptor. It has strong sedating and analgesic properties and is used predominantly as an adjunct in anesthesia or as a primary anesthetic. Anesthetic doses induce respiratory depression and chest wall rigidity [499] which requires assisted ventilation. For pain treatment intravenous or epidural on-demand procedures are in use and a patch formulation for pain relief up to seven days is under development. Following parenteral administration sufentanil has a rapid onset and a short duration of action. The compound is very lipophilic and has a high plasma protein binding of ≈ 90 %. The short duration of action is more dependent on redistribution than on metabolic inactivation. Doses up to 8 μg/kg are adequate for pain treatment and higher doses up to 30 μg/kg for surgery [500]. Sufentanil has a morphine-type side effect profile and induces severe respiratory depression and chest wall rigidity in anesthetic dosages [501]. High dose levels have been associated with seizures. Prolonged use may induce tolerance and dependence.

Trade name: Sufenta (Belgium, Germany, France, Norway, USA).

3.1.1.4. Methadone and Congeners

Levomethadone [*125-58-6*], (*R*)-6-dimethylamino-4,4-diphenylheptan-3-one,

$C_{21}H_{27}NO$, M_r 309.45, *mp* 98–100 °C (*mp* racemic form [*57-42-1*] 78–79 °C), $[\alpha]_D^{20}$ −32° (*c* = 1.8, CH_3CH_2OH), hydrochloride [*5967-73-7*], $C_{21}H_{27}NO \cdot HCl$, M_r 345.91, *mp* 240–241 °C (*mp* racemic form [*1095-90-5*] 231 °C), $[\alpha]_D^{20}$ −169° (*c* = 2.0, water).

Synthesis [502–509]:

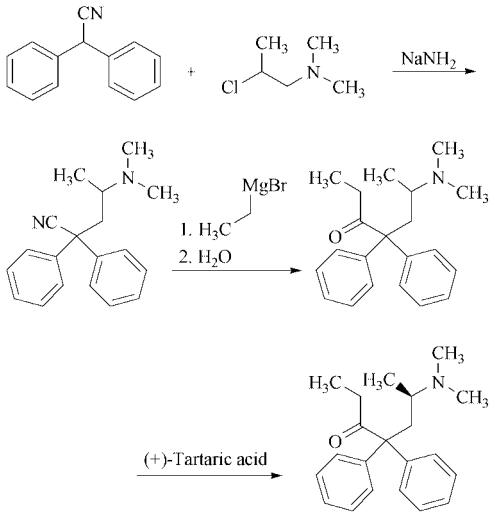

Levomethadone

Opioid receptor binding: Levomethadone is the μ-selective levorotatory enantiomer of racemic methadone [510]. It has an opioid receptor affinity in the range of morphine.

Clinical use: Levomethadone [511] is similar to morphine in potency and side effect profile. The compound has a high oral availability and a long duration of action [12, 513]. Methadone is extensively protein-bound and accumulation can occur during chronic administration. For pain treatment, the active dose range is 2.5–10 mg. Methadone can be given orally or by subcutaneous or intramuscular administration (2.5–10 mg in analgesia), but causes pain at the injection site.

The long duration of action makes the compound suitable for the substitution treatment of opioid addiction [514]. Whereas levomethadone is preferred for pain treatment, oral racemic methadone is widely used for the substitution treatment of opioid addiction [515].

Levomethadone has a morphine-like side effect profile with stronger respiratory depression and less sedation than morphine [516]. The compound has a morphine-type abuse and dependence potential. Because of its slow elimination withdrawal reactions are more protracted and less severe than with morphine [517].

Trade names: Eptadone (Italy), Dolophine (USA), Metasedin (Spain),Physeptone (UK), l-Polamidon (Germany).

Levomethadyl acetate [*34433-66-4*], l-α-acetylmethadol (LAAM), 4-dimethylamino-1-ethyl-2,2-diphenylpentyl acetate, [*S*-(*R**,*R**)]-β-[2-(dimethylamino)propyl]-α-ethyl-β-phenylbenzeneethanol acetate, $C_{23}H_{31}NO_2$, M_r 353.50; hydrochloride $C_{23}H_{31}NO_2 \cdot HCl$, M_r 389.96, *mp* 215 °C, $[\alpha]_D^{25}$ −60° (*c* = 0.2, water).

Synthesis: Levomethadyl acetate is prepared by reduction of dextromethadone and subsequent acylation [518].

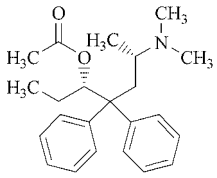

Opioid receptor binding: Levomethadyl acetate has a moderate affinity for opioid receptor binding with selectivity for the μ-type. Higher binding affinity is induced by the active metabolites l-alpha-noracetylmethadol, normethadol and methadol [519].

Clinical use: Levomethadyl acetate is an orally active methadone derivative with a slow onset and a long duration of action. The compound is extensively metabolized and active metabolites contribute to the long duration of action. Levomethadyl acetate was used mainly for substitution treatment of opioid dependence [520]. Oral doses from 20 up to 140 mg are given every 2–3 d. Supplementation with short acting opioids is necessary during the first days because of the delayed onset of action.

Levomethadyl acetate induces opioid-type side effects with respiratory depression, bradycardia, and impairment of cardiac contractility [521]. The compound inhibits cardiac electric conductivity and increases the QT interval in the electrocardiogram [522] and may induce ventricular arrhythmias and Torsade de pointes [523]. Because of the cardiovascular side effects

(Q/T-interval prolongation), the compound was recently withdrawn from the market.

Trade name: Orlaam (USA).

Dextromoramide [*357-56-2*], (+)-(*S*)-3-methyl-4-morpholin-4-yl-2,2-diphenyl-1-pyrrolidin-1-yl-butan-1-one, $C_{25}H_{32}N_2O_2$, M_r 392.55, *mp* 183–184 °C, $[\alpha]_D^{25}$ +25.5° (*c* = 5, benzene), $[\alpha]_D^{25}$ +16° (*c* = 5, ethanol), D-tartrate [*2922-44-3*], $C_{25}H_{32}N_2O_2 \cdot C_4H_6O_6$, M_r 542.64, *mp* 189–192 °C, $[\alpha]_D^{25}$ +25.5° (*c* = 5, water), $[\alpha]_D^{25}$ +30.5° (*c* = 5, CH_3OH).

Synthesis [48, 524]:

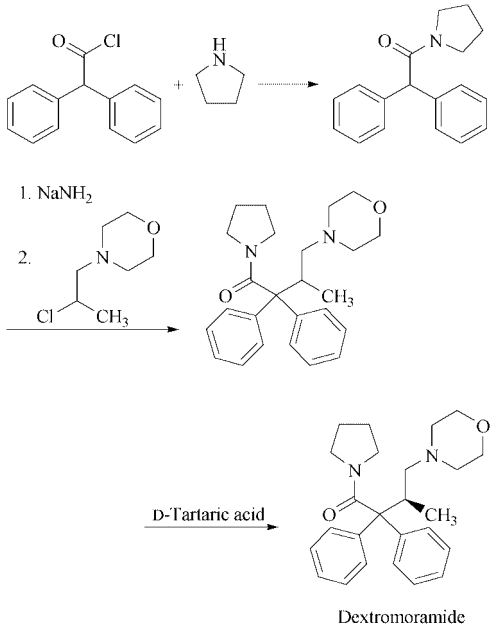

Opioid receptor binding: Dextromoramide is a μ-selective opioid with a higher receptor affinity than morphine.

Clinical use: Dextromoramide tartrate [525] is a strong opioid, structurally related to methadone and is used in the treatment of severe pain. Dextromoramide has a high oral availability and is administered orally or rectally. The parenteral potency is in the range of morphine, but the duration of action is shorter. The compound has a morphine-type abuse and dependence potential.

Trade name: Palfium (France, UK, Netherlands).

Dextropropoxyphene [*469-62-5*], 1-benzyl-3-dimethylamino-2-methyl-1-phenylpropionate, [*S*-(*R**,*S**)]-α-[2-(dimethylamino)-1-methylethyl]-α-phenylbenzeneethanol propanoate (ester), $C_{22}H_{29}NO_2$, M_r 339.47, *mp* 75–76 °C, $[\alpha]_D^{25}$ +67.3° (*c* = 0.6, $CHCl_3$); hydrochloride [*1639-60-7*], $C_{22}H_{29}NO_2 \cdot HCl$, M_r 375.93, *mp* 163–168.5 °C, $[\alpha]_D^{25}$ +59.8° (*c* = 0.6, water).

Synthesis: 4-Dimethylamino-3-methyl-1,2-diphenyl-butan-2-ol is formed by Grignard reaction of 3-dimethylamino-2-methyl-1-phenylpropan-1-one with benzylmagnesium chloride. The preferred product is the α-diastereomer (75 % α-form = 1*SR*,2*RS* isomer, 15 % β-form). The α-form crystallizes and the diastereomeric β-form remains in solution, because of its better solubility. Racemic resolution to obtain the analgetically (+)-enantiomer can be performed on the pure α-Grignard product via fractional crystallization of the salts with D-camphorsulfonic acid. Alternatively the resolution can be achieved by treating the racemic Mannich product 3-dimethylamino-2-methyl-1-phenylpropan-1-one with (−)-dibenzoyltartaric acid in acetone as solvent [582, 585–587] .

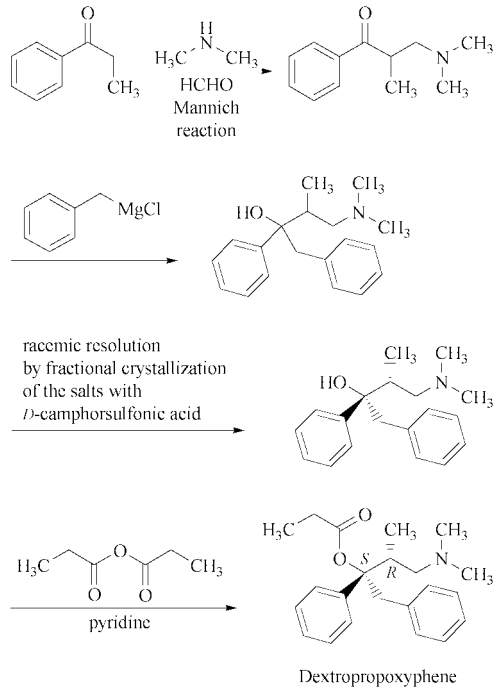

Opioid receptor binding: Dextropropoxyphene has a μ-opioid receptor binding affinity lower than morphine. Binding at other opioid receptors is even weaker.

Clinical use: Dextropropoxyphene is a moderately potent opioid analgesic often combined with paracetamol or acetylsalicylic acid. As hydrochloride or napsylate it is used orally for the treatment of mild to moderate pain [529]. Adverse reactions in the therapeutic range are mild and include drowsiness, dizziness, sedation, and nausea. Overdosage can induce serious adverse events including profound sedation, respiratory depression, cardiovascular disturbances, convulsions, and psychotic reactions, often with fatal outcome [530]. Dextropropoxyphene has a relatively low abuse liability. Abuse by injection is impeded by severe irritations at the injection side.

Because of severe and often fatal intoxications the compound is more and more substituted by other weak opioids.

Trade names: Antalvic (France), Darvon (USA), Develin (Germany), Doloxene (UK).

Diphenoxylate [*915-30-0*], ethyl 1-(3-cyano-3,3-diphenylpropyl)-4-phenylpiperidine-4-carboxylate, $C_{30}H_{32}N_2O_2$, M_r 452.59; hydrochloride [*3810-80-8*], $C_{30}H_{32}N_2O_2 \cdot HCl$, M_r 489.06, *mp* 220.5–222.0 °C.

Synthesis: The reaction of ethyl 4-phenylpiperidine-4-carboxylate with 4-bromo-2,2-diphenylbutyronitrile yields diphenoxylate [48, 531].

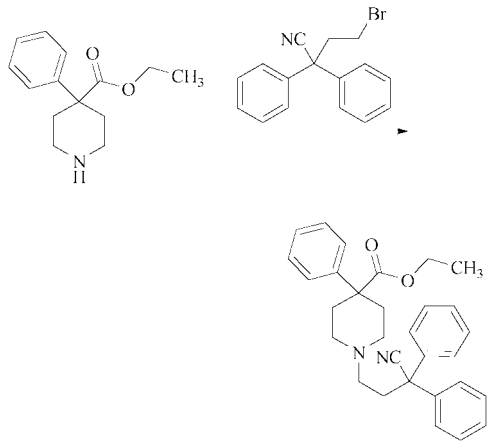

Diphenoxylate

Alternatively, diphenylacetonitrile can be condensed with ethyl 1-(2-chloro-ethyl)-4-phenylpiperidine-4-carboxylate in the presence of sodium amide.

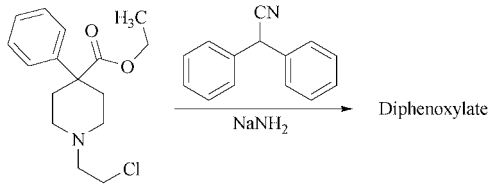

Opioid receptor binding: Diphenoxylate is the active metabolite of difenoxine [532] and has a high affinity and selectivity for the μ-type of opioid receptor.

Clinical use: Diphenoxylate [533] is a synthetic pethidine derivative with a limited access to the brain and marginal analgesic activity. After oral administration, it reduces intestinal motility and secretion and is used for the treatment of diarrhea. Side effects include dizziness, drowsiness, euphoria, anorexia, intestinal paralysis. Prolonged use of high dosages may induce morphine-like physical dependence. To avoid abuse, diphenoxylate is routinely combined with small amounts of atropine.

The compound is now substituted by loperamide, which has less CNS side effects.

Trade names: Reasec (Germany, Italy), Diaserd (France), Tropergen (UK), Lomotil (USA).

Loperamide [*53179-11-6*], 4-[4-(4-chlorophenyl)-4-hydroxypiperidin-1-yl]-*N*,*N*-dimethyl-2,2-diphenylbutyramide, 4-(4-chlorophenyl)-4-hydroxy-*N*,*N*-dimethyl-α, α-diphenyl-1-piperidinebutanamide, $C_{29}H_{33}ClN_2O_2$, M_r 477.04; monohydrochloride [*34552-83-5*], $C_{29}H_{33}ClN_2O_2 \cdot HCl$, M_r 513.51, *mp* 222–223 °C (decomp.).

Synthesis [48, 534, 535] (see next page).

Opioid receptor binding: Loperamide has a morphine-like affinity and selectivity for the μ-opioid receptor [536].

Clinical use: Loperamide [537] is a synthetic opioid subjected to extensive first-pass metabolization after oral administration. Therefore little intact drug reaches the systemic circulation and the central nervous system. Loperamide has no centrally mediated analgesic efficacy, but may have a clinically useful analgesic

effect via peripheral opioid receptors [539]. Systemic and central opioid side effects are widely missing. Orally administered loperamide acts locally in the gut by inhibition of intestinal motility and secretion [540]. Besides the strong μ-opioid action, calcium and calmoduline antagonism are involved in the antidiarrheal activity. The compound is used in doses of 4–8 mg for the treatment of acute and chronic diarrhea and for the management of colostomies and ileostomies [538]. Adverse effects include nausea, dry mouth, dizziness. High doses can induce toxic megacolon and paralytic ileus. The compound has no abuse and dependence potential and is meanwhile available as over the counter (OTC) product [542].

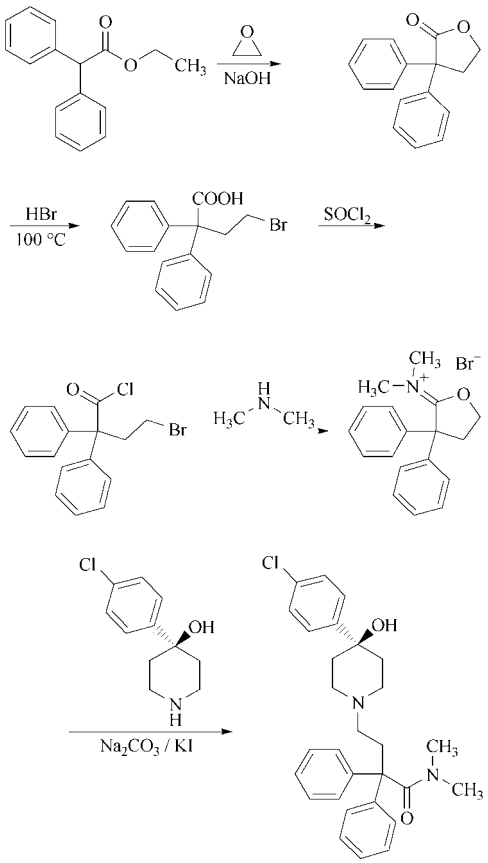

Loperamide

Trade names: Imodium (Germany, Belgium, USA, France), Arret (UK).

Loperamide is also marketed in many generic formulations.

Piritramide [*302-41-0*], 1′-(3-cyano-3,3-diphenylpropyl)-[1,4′]bipiperidinyl-4′-carboxamide, $C_{27}H_{34}N_4O$, M_r 430.59, *mp* 149–150 °C.

Synthesis: Piritramide is prepared by condensation of 4-piperidinopiperidine-4-carboxamide with 3,3-diphenyl-3-cyanopropylbromide [48, 543].

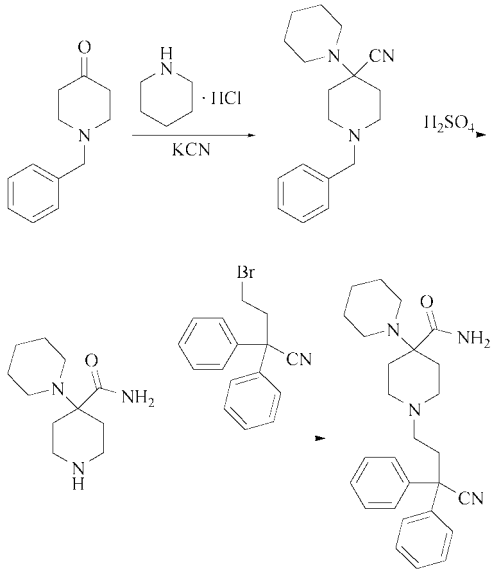

Piritramide

Opioid receptor binding: Piritramide is a μ-opioid with morphine-like affinity and receptor selectivity.

Clinical use: Piritramide [544] is a synthetic opioid analgesic used for the treatment of acute, preferentially postoperative pain. It is given in doses of 15 mg by the intramuscular, subcutaneous, or intravenous route. Piritramide has a morphine-like effect – side effect profile.

Trade names: Dipidolor (Germany, Belgium, Netherlands), Piridolan (Sweden).

3.1.2. Other Structures

Tilidine [*51931-66-9*], ethyl *trans*-2-dimethylamino-1-phenylcyclohex-3-ene-carboxylate, $C_{17}H_{23}NO_2$, M_r 273.17, *mp* 34 °C, *bp* (133.32 Pa) 95.5–96 °C; hydrochloride [*27107-79-5*], $C_{17}H_{23}NO_2 \cdot HCl$, M_r

309.84, *mp* 159 °C; hydrochloride hemihydrate $C_{17}H_{23}NO_2 \cdot HCl \cdot 1/2\ H_2O$, M_r 318.80, *mp* 125 °C.

Synthesis: Reaction of buta-1,3-dienyl-dimethylamine 4 with ethyl atropate 5 yields tilidine as a *cis*/*trans* mixture (*trans* : *cis* = 2 : 3). Most of the analgetically inactive *cis* isomer is separated as a zinc complex and the *trans* isomer is isolated as the hydrochloride. The cis isomer can be epimerized to the *trans* form by treating the epimeric mixture with acid [48, 545–548].

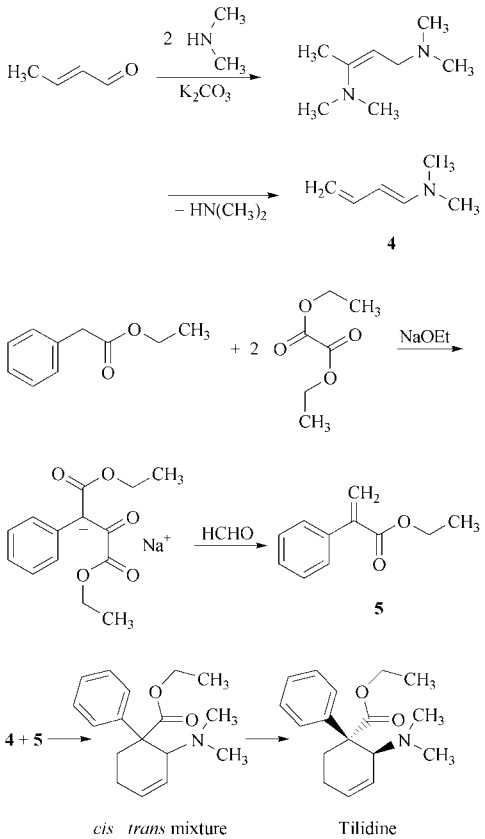

Opioid Receptor Binding: Tilidine itself has a low opioid receptor binding affinity, whereas the active metabolites nortilidine and bisnortilidine have a high affinity and selectivity for the μ-type of opioid receptor [549].

Clinical use: Tilidine [550] is rapidly metabolized to the active metabolites nortilidine and bisnortilidine. It can be given orally and parenterally and is used in single doses of 50–75 mg for the treatment of moderate to severe pain. Tilidine has a substantial abuse potential. To avoid parenteral misuse, oral preparations are combined in some countries with a small amount of naloxone hydrochloride. This is intended to induce withdrawal symptoms after illicit parenteral administration. After oral administration, the naloxone component is metabolically inactivated and does not inhibit analgesia. After parenteral administration the naloxone component is active and blocks the tilidine effect and can induce withdrawal reactions [551]. The combination with naloxone, in contrast to pure tildine, is available under normal prescription.

Trade names: Valoron (Germany, Switzerland), Tilidate (Spain).

Combination with naloxone: Tilidin-N, Valoron N.

Tramadol [*27203-92-5*], *cis*-2-dimethylaminomethyl-1-(3-methoxyphenyl)cyclohexanol, $C_{16}H_{25}NO_2$, M_r 263.38; hydrochloride [*36282-47-0*], $C_{16}H_{25}NO_2 \cdot HCl$, M_r 299.84, *mp* 180–181 °C.

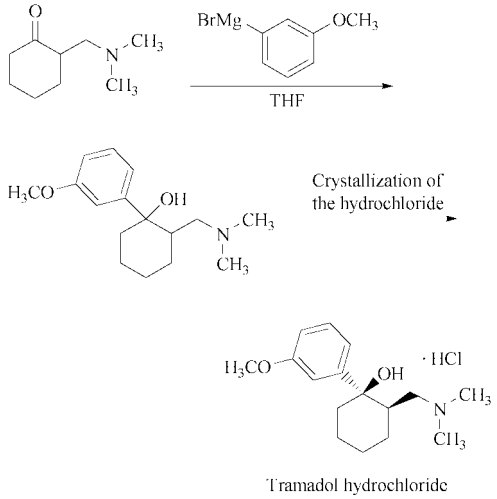

Synthesis: Grignard reaction of 2-(dimethylaminomethyl)cyclohexanone (obtained by Mannich reaction of cyclohexanone, formaldehyde, and dimethylamine hydrochloride) and the Grignard reagent of 3-bromoanisole yields tramadol as a *cis*/*trans* mixture (*cis* : *trans* = 85 : 15). Tramadol (*cis* isomer) is separated from the reaction mixture by crystallization of the hydrochloride

salt. A crystallization process of the hydrate or hydrobromide salts is also described. The *trans* isomer can be epimerized to the *cis* isomer by strong acids [48, 552, 553].

Opioid receptor affinity: Tramadol [554] itself has a weak opioid receptor affinity, the active metabolite O-demethyltramadol shows μ-selective binding with an affinity about ten times lower than morphine.

Clinical use: Tramadol hydrochloride [555] is a centrally acting analgesic with a μ-opioid and nonopioid action component [556]. The nonopioid component acts spinally via inhibition of noradrenaline (NA) and serotonin (5HT) reuptake. Tramadol is a racemic mixture of two optical enantiomers and uptake inhibition and opioid properties are differentially distributed between the enantiomers [557]. Opioid receptor affinity is low, but like codeine, tramadol is biotransformed to the O-demethyl derivative (M1), which has an essentially higher μ-opioid receptor affinity and may be responsible for the opioid properties of the compound. Tramadol is marketed worldwide and has become one of the most important centrally acting analgesics [558]. It is used in single doses of 50–100 mg for the treatment of acute and chronic medium to severe pain. The compound has a high oral bioavailability and can be given by mouth, rectally, or by intramuscular, subcutaneous, or slow intravenous injection or infusion. Typical side effects are nausea, sweating, and dizziness. High doses may have a pro-convulsive effect. Tramadol shows a reduced level of opioid side effects like respiratory depression and constipation. It has a very limited abuse potential and in most countries is not subject to narcotic control.

Tramadol is used as a mono-compound or as a combination with paracetamol [559].

Trade names: Adolonta (Spain), Contramal (Belgium, France, Italy), Nobligan (Skandinavia), Tramal, Tramal Long (Germany, Switzerland, Austria), Tramacet (UK), Ultram (USA), Ultracet (USA), Xprim (France), Zydol (UK, Ireland).

Combination with paracetamol: Ultracet (USA), Zaldiar (Germany, Austria, Switzerland).

Tramadol is also marketed in many generic formulations.

3.1.2.1. Mixed Opioid Agonist – Antagonists and Partial Agonists

Mixed agonist – antagonists are opioid compounds that have in vivo relevant affinity for more than one type of opioid receptors, commonly for the μ- and κ-types. The compounds of this group are historically derived from the antagonist nalorphine. At the μ-receptor they act as antagonists or partial agonists with low intrinsic activity. At the κ-receptor they act mostly as agonists. These substances are used as analgesics and the κ-agonistic effect and/or the partial μ-agonistic effect is responsible for their analgesic efficacy. The side effect profile and the misuse potential depend on the relative prevalence of either component in the individual compound. The κ-component effects dysphoria, hallucinations, and other psychotic reactions, but induces a low level of respiratory depression and addiction. Compounds with a marked κ-agonistic component such as pentazocine are meanwhile withdrawn from the market because of their unpleasant side effect profile. The μ-component is responsible for respiratory depression, euphoria, and dependence potential. Compounds with a more μ-antagonistic profile have a low abuse potential and are not under narcotic drug control, compounds with a stronger μ-agonistic component are scheduled. Application of a mixed agonist-antagonist in an opioid pretreated patient may reduce analgesia and induce a withdrawal reaction.

Buprenorphine [*52485-79-7*], 17-cyclopropylmethyl-α-(1,1-dimethylethyl)-4,5α-epoxy-18,19-dihydro-3-hydroxy-6-methoxy-α-methyl-6,14-ethanomorphinan-7-methanol, $C_{29}H_{41}NO_4$, M_r 467.30, *mp* 209 °C; hydrochloride [*53152-21-9*], $C_{29}H_{41}NO_4 \cdot HCl$, M_r 504.10.

Synthesis [48, 424–427] (see next page).

Opioid receptor binding: Buprenorphine [560] has a mixed agonistic – antagonistic action profile with a high affinity at the μ-, κ-, and δ-opioid receptor. A ca. 100-fold lower affinity was observed at the ORL1-receptor. The compound shows a slow receptor dissociation which can explain some peculiarities in the pharmacological actions [561].

Clinical use: Buprenorphine [562] is used against moderate to severe pain, mostly for

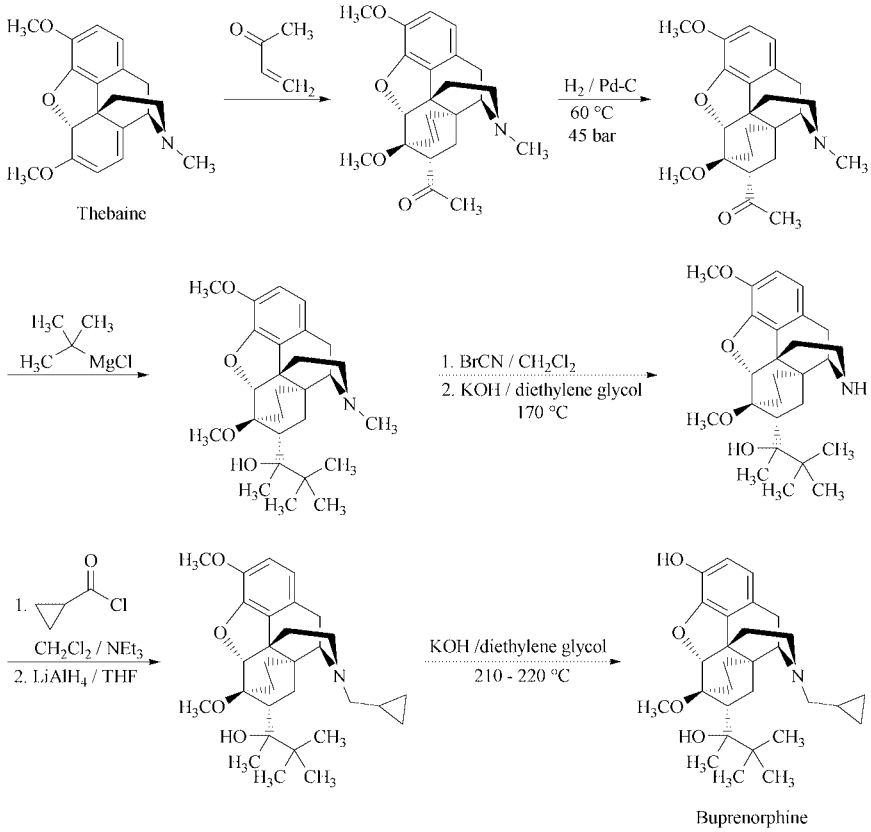

chronic pain treatment. The potency is about 20–30 times higher than that of morphine. Buprenorphine may be used for anesthesia premedication or as adjunct to anesthetics. The compound has a long duration and a slow offset of action and is therefore suited for the substitution treatment of opioid addiction [563, 564]. Due to its partial agonistic properties buprenorphine can act in combination with full agonists as an antagonist, reducing their effect and precipitating a withdrawal reaction in opioid dependent persons. Antagonistic properties are seen in doses much higher than the analgesic dose range. Therefore no special precaution is necessary when changing the treatment with a standard opioid agonist to buprenorphine or vice versa [565]. Buprenorphine is given parenterally, orally (sublingual), or by the transcutaneous route as a patch. The doses for slow intravenous or intramuscular administration are 300–600 μg, the sublingual doses are 200–400 μg, both given every 6–8 h. Buprenorphine patches are available as a matrix patch formulations with release rates of 35, 52.5, and 70 μg/h.

The use of buprenorphine for opioid addiction becomes more and more important. For this indication, buprenorphine is used in doses up to 24 mg sublingually alone or in combination with naloxone.

Adverse reactions of buprenorphine are typical of μ-opioids and include respiratory depression, drowsiness, nausea, and vomiting. Respiratory depression often occurs delayed, is long lasting, and may need higher doses of naloxone [566, 567]. Buprenorphine has a limited abuse potential and withdrawal reactions, due to slow receptor dissociation, are mild and delayed.

Trade names: Buprenex (USA), Buprex, Subutex (Germany, France, USA), Temgesic (Germany, France, UK), Transtec (Germany, Austria, Switzerland.

Combination with naloxone: Suboxone (USA).

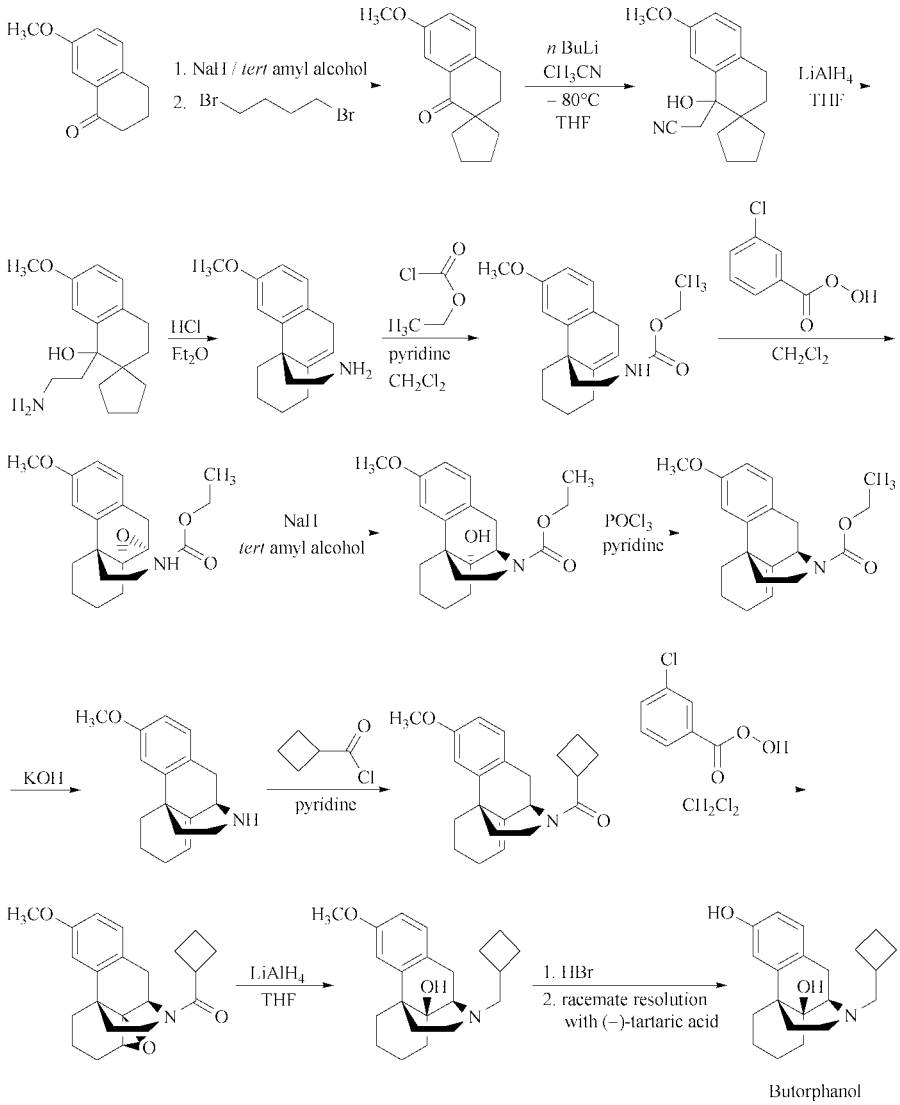

Butorphanol [*42408-82-2*], 11-cyclobutylmethyl-1,2,3,4,9,10-hexahydro-4α,10-propanophenanthrene-6,10α-diol, $C_{21}H_{29}NO_2$, M_r 327.22, *mp* 215-217 °C, $[\alpha]_D$ – 70° (c = 0.1, CH_3OH); tartrate [*58786-99-5*], $C_{21}H_{29}NO_2 \cdot C_4H_6O_6$, M_r 477.55, *mp* 217–219 °C, $[\alpha]_D$ – 64.0° (c = 0.4, CH_3OH).

Synthesis [48, 568–570] (see on top).

Opioid receptor binding: Butorphanol [571] is a mixed agonist – antagonist opioid with high μ- and κ-receptor affinity and full agonistic activity at the κ-receptor and partial agonistic – antagonistic effect at the μ-receptor.

Clinical use: Butorphanol [572, 573] is a fairly potent opioid analgesic. It is used for the treatment of moderate to severe pain, for migraine and headache [574] and as an adjunct to anesthesia. Butorphanol is orally inactive but can be given by the nasal route [574]. The usual administration is via the intramuscular or intravenous route. The intramuscular doses are 1–4 mg every 3–4 h, the intravenous doses are 0.5–2 mg. Nasal doses are ≈ 1 mg per spray in each nostril.

The plasma elimination half-life of butorphanol is in the range of 3 h. Butorphanol [571]

has a side effect profile combining morphine- and pentazocine-like symptoms. They include drowsiness, weakness, sweating, feelings of floating, and nausea. It has respiratory depressant properties similar to morphine but with a ceiling effect. As antidote naloxone can be used. Overt hallucinations or other psychotic effects are rare and less often reported than with pentazocine. The compound has a low abuse potential and has not been submitted to narcotic control.

Trade name: Stadol (USA).

Dezocine [*53648-55-8*], 15-amino-1-methyltricyclo[7.5.1.0^{127,255}]pentadeca-2,4,6-trien-4-ol, [5*R*-(5α,11α,13*S**)]-13-amino-5,6,8,9,10,11,12-octahydro-5-methyl-5,11-methanobenzocyclodecen-3-ol, $C_{16}H_{23}NO$, M_r 245.36; hydrobromide [*57236-36-9*], $C_{16}H_{23}NO \cdot HBr$, M_r 326.27, *mp* 269–270 °C.

Synthesis [48, 575, 576]:

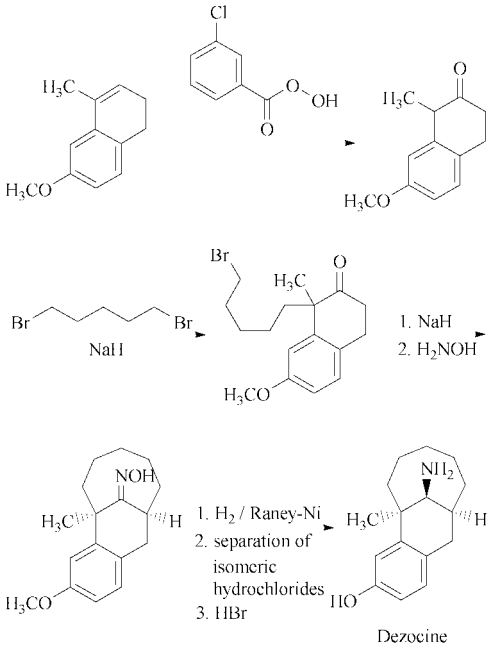

Opioid receptor binding: Dezocine [577] is a mixed agonist – antagonist with binding affinity to the μ-opioid receptor in the range of morphine. The δ- and κ-affinity is 10–100-fold lower [578].

Clinical use: Dezocine [579, 580] is a medium-potent opioid and is suited for the treatment of moderate to severe pain. It is given by injection in the dose range of 5–20 mg. Side effects include nausea, vomiting, and drowsiness. Overdosing can be treated with naloxone. In opioid pretreated patients, dezocine may precipitate withdrawal symptoms.

Trade name: Dalgan (USA, no longer marketed).

Meptazinol [*54340-58-8*], 3-(3-ethyl-1-methylazepan-3-yl)phenol, 3-(3-ethyl-hexahydro-1-methyl-1*H*-azepin-3-yl)phenol, $C_{15}H_{23}NO$, M_r 233.35, *mp* 127.5–133 °C; hydrochloride [*59263-76-2*], $C_{15}H_{23}NO \cdot HCl$, M_r 269.81.

Synthesis [48, 581, 582]:

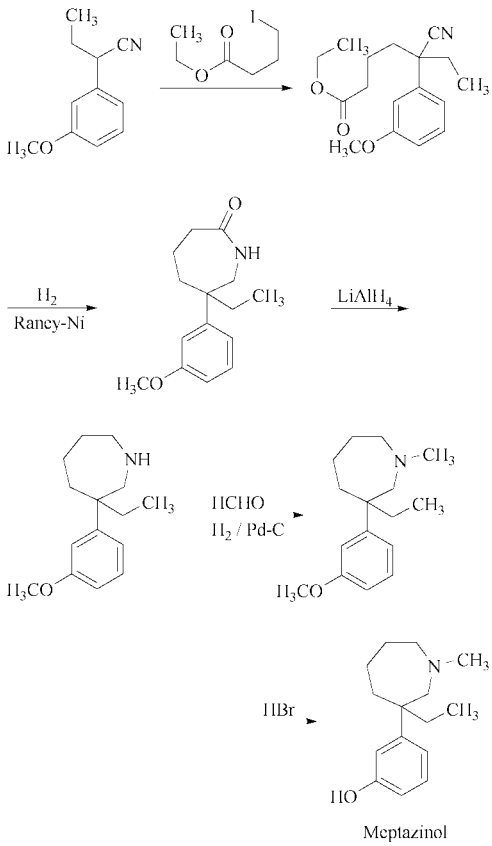

Opioid receptor binding: Meptazinol [583] is a partial μ-agonist with a high μ-receptor affinity. A selective action at a special subtype of μ-opioid receptor (μ-1 subtype, [584]) is controversial.

Clinical use: Meptazinol [585] is used parenterally (50–100 mg) or orally to treat moderate or moderately severe pain. The compound has a shorter duration of action than morphine. The most common side effects are nausea, vomiting, and dizziness. Sweating and hypotension can also occur. Meptazinol has opioid antagonistic properties and can induce withdrawal in opioid-dependent persons. The compound has a very low dependence potential and is not under narcotic drug control. The compound is reported to be relatively free of respiratory depressant activity, which was attributed to selective binding to a μ-1 subtype of the opioid receptor [584]. This is not confirmed by others and the effect may be alternatively explained by a cholinergic action component [583] of the compound. In accordance with a low κ-receptor affinity, the incidence of psychotomimetic actions and hallucinations is low.

Trade name: Meptid (Germany, UK).

Nalbuphine [*20594-83-6*], (5α6α)-17-(cyclobutylmethyl)-4,5-epoxymorphinan-3,6,14-triol, $C_{21}H_{27}NO_4$, M_r 357.44, *mp* 230.5 °C; hydrochloride [*23277-43-2*], $C_{21}H_{27}NO_4 \cdot HCl$, M_r 393.91, *mp* 291–292 °C (decomp.).

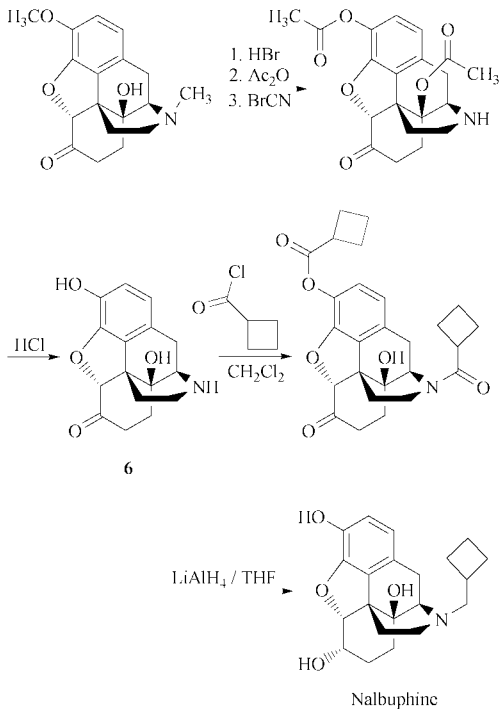

Synthesis: Starting material for the nalbuphine synthesis is oxycodone (see page 55). After ether cleavage to oxymorphone the product is acylated and the *N*-methyl group is removed by treatment with cyanogen bromide. The acetyl groups are hydrolyzed with dilute hydrochloric acid. The resulting 14-hydroxydihydronormorphinone 6 is N-alkylated with cyclobutylmethyl bromide, and the carbonyl group at C-6 is reduced [48, 586, 587].

Opioid receptor binding: Nalbuphine is a mixed agonist – antagonist opioid with a high affinity at the μ- and κ-opioid receptor. At the κ-receptor the compound is an agonist, at the μ-receptor a partial agonist with a very low intrinsic activity, thus acting more as a μ-antagonist [588].

Clinical use: Nalbuphine [589, 590] is an opioid analgesic with a mixed agonist – antagonist action profile. Analgesia is limited by a "ceiling" effect. The compound has a low oral availability and is used as hydrochloride only for parenteral application to treat moderate to severe pain and as adjunct to anesthesia. The dose for pain relief is 10–20 mg. The most frequent side effects are drowsiness, sweating, nausea, and vomiting. Hallucinations and psychotomimetic reactions are less frequent than with pentazocine, reflecting a relative weak κ-agonistic component of the compound. Nalbuphine induced respiratory depression is less severe than with morphine and is limited by a ceiling effect [591]. The dependence potential is low and nalbuphine is not subject to narcotic control.

Trade name: Nubain (Germany, France, UK, USA).

Nalorphine [*62-67-9*], (5α,6α)-7,8-didehydro-4,5-epoxy-17-(2-propenyl)morphinan-3,6-diol, $C_{19}H_{21}NO_3$, M_r 311.38, *mp* 208–209 °C, $[\alpha]_D^{25}$ −155.3° (c = 3, CH_3OH); hydrochloride [*57-29-4*], $C_{19}H_{21}NO_3 \cdot HCl$, M_r 347.84, *mp* 260–263 °C; hydrobromide [*1041-90-3*], $C_{19}H_{21}NO_3 \cdot HBr$, M_r 392.29, *mp* 258–259 °C (decomp.).

Synthesis: Diacetylmorphine (heroin) is demethylated with cyanogen bromide and hydrolyzed to normorphine, which is alkylated with allyl bromide to give nalorphine [48, 592].

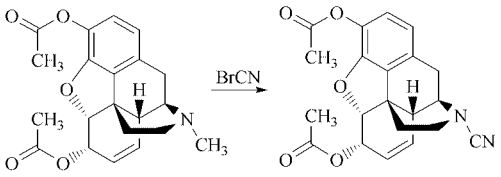

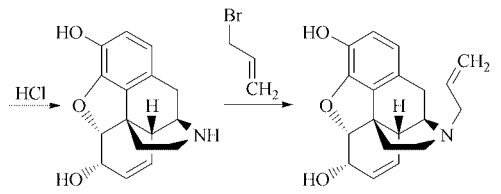

Opioid receptor binding: Nalorphine [113] is a mixed agonist – antagonist with a high affinity and low intrinsic action at the μ- and κ-opioid receptors.

Clinical use: Despite a predominant antagonistic action profile nalorphine has substantial analgesic activity and was the first opioid combining analgesic activity with an antagonistic action. The compound is no longer used because of κ-agonistic type adverse reactions such as anxiety, hallucinations, and dysphoric mood alterations.

Pentazocine [*359-83-1*], 6,11-dimethyl-3-(3-methylbut-2-enyl)-1,2,3,4,5,6-hexahydro-2,6-methanobenzo[*d*]azocin-8-ol, (2α,6α,11*R**)-1,2,3,4,5,6-hexahydro-6,11-dimethyl-3-(3-methyl-2-butenyl)-2,6-methano-3-benzazocin-8-ol, $C_{19}H_{27}NO$, M_r 285.42, *mp* 145–148 °C; hydrochloride [*2276-52-0*], $C_{19}H_{27}NO \cdot HCl$, M_r 321.89; lactate (1 : 1) [*17146-95-1*], $C_{19}H_{27}NO \cdot C_3H_6O_3$, M_r 375.51.

Synthesis [48, 594, 595] (see right column).

Opioid receptor binding: Pentazocine [596] is a mixed opioid agonist – antagonist with agonistic effects at the κ- and partial antagonistic effects at the μ-type of opioid receptor.

Clinical use: Pentazocine is a potent analgesic with therapeutic dosages in the range of 25–100 mg. It is orally bioavailable and was used as hydrochloride or lactate in oral, parenteral, and rectal formulations for the treatment of moderate to severe pain. Pentazocine induces morphine-type side effects like dizziness, nausea, vomiting, sedation, and sweating. In addition, it can cause κ-agonist type psychotomimetic effects like hallucinations, nightmares and thought disturbances. Respiratory depression is weaker than with morphine and is subject to a "ceiling" effect.

The compound has a relevant abuse potential [597] and high doses may produce dependence of the morphine type. Pentazocine effects can be antagonized by naloxone.

Because of its side effect profile, the compound is no longer used.

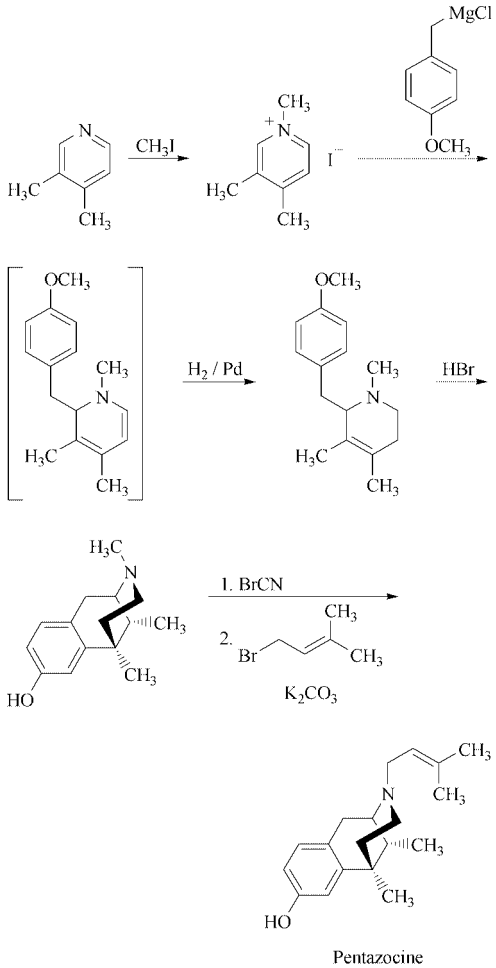

3.1.2.2. Opioid Antagonists [598]

Opioid and narcotic antagonists (Table 2) are compounds with binding affinity at opioid receptors, but without intrinsic activity. They only block the receptor without inducing a genuine effect but are able to inhibit or reverse the actions of opioid agonists or partial agonists. Antagonists may or may not be selective for one opioid receptor type. Only antagonists with a high affinity and/or selectivity for the μ-receptor have clinical application. They are used to terminate medically desired effects of μ-opioid such as anesthesia, to reverse side effects like respiratory depression or to treat opioid intoxications. Acute administration of antagonists precipitates withdrawal reactions in opioid-dependent subjects. In the course of long-term treatment orally active antagonists are used for detoxification of opioid addicts.

Antagonists with δ- or κ-receptor selectivity have no clinical application but are widely used in opioid research.

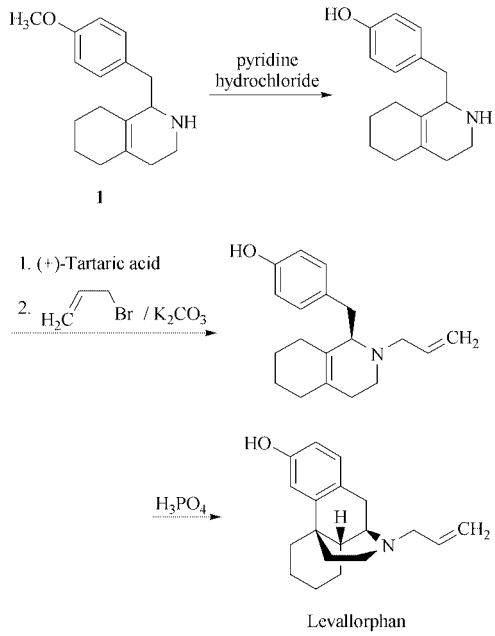

Levallorphan [*152-02-3*], 17-(2-propenyl)-morphinan-3-ol, 1-*N*-allyl-3-hydroxymorphinane, $C_{19}H_{25}NO$, M_r 283.42, *mp* 180–182 °C, $[\alpha]_D^{25}$ −88.9° (CH_3OH); hydrogen tartrate (1 : 1) [*71-82-9*], $C_{19}H_{25}NO \cdot C_4H_6O_6$, M_r 433.50, *mp* 176–177 °C, $[\alpha]_D^{25}$ −39° (water).

Synthesis: Starting material for the levallorphan synthesis is 1-(4-methoxybenzyl)-1,2,3,4,5,6,7,8-octahydro isoquinoline (**1**) (preparation see page 54) [48, 599–602].

Opioid receptor binding: Levallorphan has a high binding affinity but a low intrinsic activity at the μ-opioid receptor.

Analgesic efficacy and clinical use: Levallorphan [603] is an opioid antagonist with practically no analgesic action. It has been used as one of the first relative pure antagonists for the treatment of opioid overdosage, to reverse opioid central depression and to antagonize opioid-induced respiratory impairment [604]. The compound is now replaced by naloxone [605].

Naloxone [*465-65-6*], (5α)-4,5-epoxy-3,14-dihydroxy-17-(2-propenyl)morphinan-6-one, $C_{19}H_{21}NO_4$, M_r 327.37, *mp* 184 °C (also reported as 177–178 °C), $[\alpha]_D^{20}$ −194.5° (*c* = 0.93, $CHCl_3$); hydrochloride [*357-08-4*], $C_{19}H_{21}NO_4 \cdot HCl$, M_r 363.84, *mp* 200–205 °C.

Synthesis: 14-Hydroxydihydronormorphinone 6 (see page 70) is N-alkylated with allyl bromide [48, 606, 607].

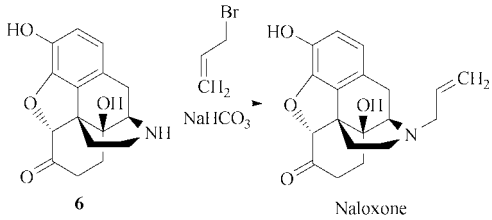

Opioid receptor binding: Naloxone [608] is a pure opioid antagonist with a high affinity and a limited selectivity for the μ-receptor.

Efficacy and clinical use: Naloxone [609, 610] has no analgesic activity. The compound is the standard antidote to treat opioid adverse reactions, opioid overdoses, or to stop an intended use of an opioid compound. Typical indications are inhibition of opioid-induced respiratory depression, termination of opioid anesthesia or protection of neonates following opioid treatment during labor. Naloxone has a short duration of action and repetitive administration may be necessary to antagonize longer acting agonists. To avoid parenteral misuse of non-scheduled oral opioid formulations (tilidine, pentazcocine), a small amount of naloxone is added which is orally inactivated, but is fully active after parenteral administration.

Naloxone is orally inactive and is only used parenterally in single or repetitive doses of 0.4–2 mg up to a total dose of 10 mg, as an intravenous bolus injection or by infusion. The compound is more potent against pure opioid agonists than against mixed agonist – antagonists. Caution should be used in opioid-dependent persons or in persons under high-dose opioid treatment, as naloxone may precipitate an acute withdrawal reaction. Naloxone is relatively free of side effects. Nausea, vomiting, and convulsions have occasionally been reported.

Trade names: Narcan (France, UK, USA), Narcanti (Germany, Austria).

Naltrexone [*16590-41-3*], (5α)-17-(cyclopropylmethyl)-4,5-epoxy-3,14-dihydroxy morphinan-6-one, $C_{20}H_{23}NO_4$, M_r 341.40, *mp* 168–170 °C; hydrochloride [*16676-29-2*], $C_{20}H_{23}NO_4 \cdot HCl$, M_r 377.87, *mp* 274–276 °C.

Synthesis:

a) Naltrexone is prepared in the same way as naloxone starting from oxycodone via the intermediate **6** except that the nitrogen atom is alkylated with cyclopropyl bromide [48, 611].

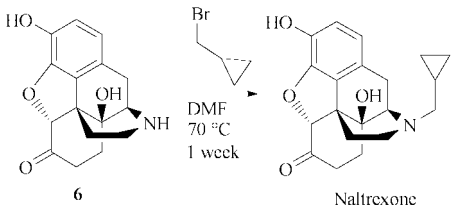

b)

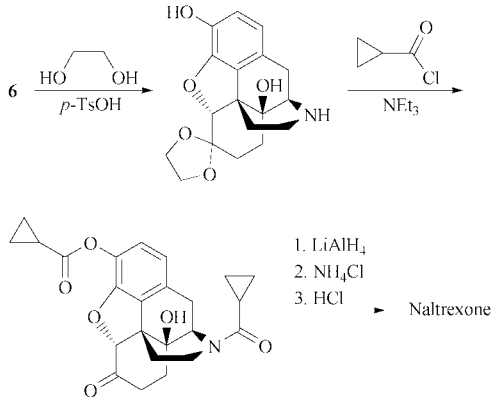

Opioid receptor binding: Naltrexone has a high affinity and selectivity for the μ-opioid receptor [612].

Efficacy and clinical use: Naltrexone [613, 614] is a pure opioid antagonist and has no analgesic activity. Naltrexone has a higher intravenous potency and longer duration of action than naloxone. It has a higher oral bioavailability and is given by mouth to treat opioid dependence and to maintain abstinence during opioid detoxification [615]. In opioid-dependent persons, naltrexone induces an acute withdrawal reaction.

Trade names: Nalorex (France, UK), Nemexin (Germany), Revia (USA), Trexan (USA).

N-Methylnaltrexone bromide [*073232-52-7*], 17-(cyclopropylmethyl)-4,5a-epoxy-3,14-dihydroxy-17-methyl-6-oxomorphinanium bromide, $C_{21}H_{26}BrNO_4$, M_r 436.34, *mp* 251 °C, $[\alpha]D^{25}$ - ?° (c = 1.49, CH_3OH); N-methylnaltrexone iodide [*073232-53-8*], $C_{21}H_{26}INO_4$, M_r 483.34.

Synthesis: Methylnaltrexone is obtained by quaternization of *N*-(cyclopropylmethyl)-noroxymorphone (naltrexone) with methyl bromide in acetone/dimethylformamide [616]:

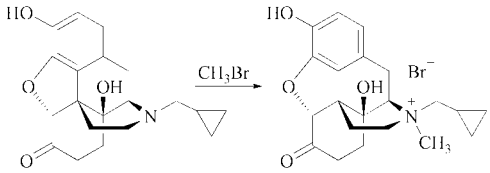

Efficacy and intended clinical use: *N*-Methylnaltrexone [617] is the polar derivative of naltrexone, which does not reach the central nervous system but can block intestinal opioid receptors. The compound is under development to counteract opioid induced bowel dysfunction such as constipation and megacolon [618].

Nalmefene [*055096-26-9*], *N*-(cyclopropylmethyl)-4,5α-6-methylenemorphinan-3,1β-diol, $C_{21}H_{25}NO_3$, M_r 339.43, *mp* 105–110 °C; hydrochloride [*58895-64-0*], $C_{21}H_{26}ClNO_3$, M_r 375.89, *mp* 188–190 °C.

Synthesis: Nalmefene is synthesized by a Wittig reaction of naltrexone with triphenylmethylphosphonium bromide in DMSO under basic catalysis of NaH.

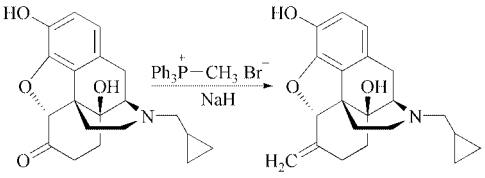

Nalmefene

Nalmefene is a 6-methylene analogue of naltrexone. It is a pure opioid receptor antagonist and has a high affinity for the κ-opioid receptor. Nalmefene has a long duration of action and is used in oral and parenteral formulations.

Nalmefene is used to reverse opioid intoxication [619] and for the treatment of opioid dependence [620]. The compound is in clinical development for the treatment of alcoholism [621], to support smoking cessation and to control pathological gambling [622].

Trade name: Revex (USA).

3.2. Antineuropathic Analgesics and Other Non-opioid Compounds

3.2.1. Neuropathic Pain Treatment [623–629, 658, 659]

Definition: Neuropathic pain results from damage to the nervous system due to many diverse processes. The damage causes persistent, distressing pain, that is reputedly unresponsive to conventional analgesics. The International Association for the Study of Pain (IASP) defines neuropathic pain as pain initiated or caused by a primary lesion or dysfunction in the nervous system. This malfunction of the nervous system can occur at any point along the sensory pathway (peripheral nerve, dorsal root ganglion, and CNS). The term peripheral neuropathic pain is usually used to refer to abnormalities in the peripheral nervous system. Central pain is retained as the term when the lesion or dysfunction occurs in CNS.

Table 4. Etiological classification of neuropathic pain [623]

Injury	Examples
Infection/inflammation	
	• postherpetic neuralgia • HIV
Trauma	
	• post-amputation phantom limb pain • spinal cord injury • surgery
Ischemia	
	• diabetic neuropathy • central post-stroke pain
Compression	
	• sciatica • carpal tunnel • trigeminal neuralgia
Neoplasm	
	• invasion of neural structures • compression by tumor expansion
Iatrogenic	
	• vinca alkaloids • antiretrovirals • post-irradiation
Idiopathic	
	• multiple sclerosis • trigeminal neuralgia • complex regional pain syndrome

Classification: Neuropathic pain has been traditionally classified according to its etiology (Table 1) and tends to be chronic and unremitting. Patients with neuropathic pain are often difficult to treat, the pain is sometimes relatively unresponsive to conventional analgesics and thus a wide variety of drugs from many classes have been suggested for this indication. Table 4. indicates that neuropathic pain refers to a group of

painful disorders characterized by pain due to dysfunction or disease of the nervous system at a peripheral level, a central level, or both. In contrast to nociceptive pain, which results from the activation of pain receptors (nociceptors), neuropathic pain is the result of injury to the pain-conducting nervous system.

Symptoms: Neuropathic pain is a complex entity with many symptoms and signs that fluctuate, in number and intensity, with time. The descriptors associated with neuropathic pain have been described qualitatively as burning, stabbing, shooting, aching and electric shock-like, amongst others. Pain may be continuous or paroxysmal and felt superficially or deeply. It may also occur spontaneously, independent of a stimulus. Sensory changes in the injured area are common. There may be areas of sensory loss but also of altered sensation to cutaneous stimuli. Non-noxious stimuli, e.g. stroking, may elicit pain (allodynia). Mildly painful stimuli may cause excessive pain (hyperalgesia) and repetitive stimulation can cause an explosive build-up of pain that lasts well beyond the stimulus (hyperpathia). These sensory abnormalities can be classified by type of stimulus – thermal, mechanical, and chemical. Mechanically evoked events can be further subdivided into dynamic (brush-stroke evoked), static (pressure-evoked) and punctate.

Current Treatment Options in Neuropathic Pain. There are several rationales to guide treatment of patients with neuropathic pain. The most widely used at present is based on the etiology and anatomical distribution of the pain. Others have suggested a symptom-based approach on the premise that certain symptoms might respond best to specific classes of drugs. In clinical practise it is shown that one mechanism may lead to different symptoms and that a similar pain symptom or deficit may be produced by several different underlying mechanisms. At present, there is no single treatment that adequately and predictably controls neuropathic pain, nor does there seem to be any way of preventing the development of neuropathic pain following nerve injury. Management by a combination of both pharmacological and non-pharmacological approaches, ideally in a multidisciplinary pain clinic setting, seems to offer the best treatment options. Other interventions such as stimulation, blockade, or destruction of nerves and psychological treatments are also part of the armament for treating neuropathic pain. Table 5. gives an overview on the current available evidence-based treatment options for neuropathic pain.

Table 5. Evidence-based treatment options for neuropathic pain [623]

Drug class	Proposed mechanism of action	Examples
Tricyclic antidepressants	• unspecific sodium channel blockade • inhibition of norepinephrine and serotonin reuptake • blockade of α-adrenergic receptors	• amitryptiline • nortriptyline • desipramine • imipramine
Serotonin and/or norepinephrine reuptake inhibitors	blockade of serotonin and/or norepinephrine reuptake	• duloxetine • venlafaxine
Anticonvulsants	• unselective block of sodium channels	• benzodiazepines • carbamazepine • oxacarbazepine • phenytoin • lamotrigin • topiramate • zonisamide
Anticonvulsants	• selective block of calcium channels	• gabapentin • pregabalin
Local anesthetics (type 1b antiarrythmic drugs)	• unspecific sodium (and/or calcium) channel blockade	• lidocaine • mexiletine • bupivacaine
Drugs for the treatment of neurological disorders	• antagonism of the NMDA receptor	• memantine • dextromethorphan • ketamine
Opioids	• opioid agonism	• blockade of norepinephrine reuptake • increased serotonin release • morphine • tramadol • methadone

3.2.2. Individual Compounds

Baclofen [*1134-47-0*], β-(aminomethyl)-4-chlorobenzenepropanoic acid, δ-(aminomethyl)-*p*-chlorohydrocinnamic acid, γ-amino-β-(*p*-chlorophenyl)butyric acid, $C_{10}H_{12}ClNO_2$, M_r 213.66, *mp* 206-208 °C; hydrochloride, $C_{10}H_{13}Cl_2N$, M_r 250.12, *mp* 179-181°C.

Synthesis:

a.) Baclofen is synthesized by reaction of 4-chlorobenzaldehyde with ethyl acetonacetonate under catalysis of sodium ethoxide. The resulting condensation product is hydrolyzed to 3-(4-chlorophenyl)glutaric acid. Cyclization to 3-(4-chlorophenyl)glutaric anhydride and subsequent aminolysis yields 3-(4-chlorophenyl)glutarimide, which is finally transformed to baclofen [630, 631]

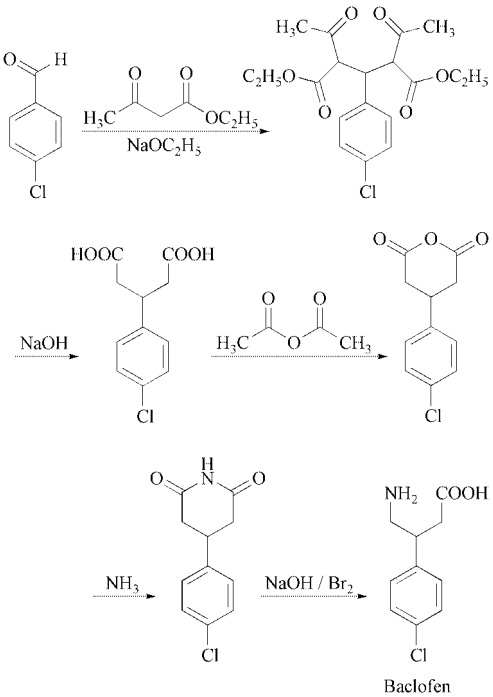

b.) Alternatively, baclofen can be obtained by reaction of ethyl-4-chlorocinnamate with nitromethane followed by hydrogenation and ester cleavage [630, 631].

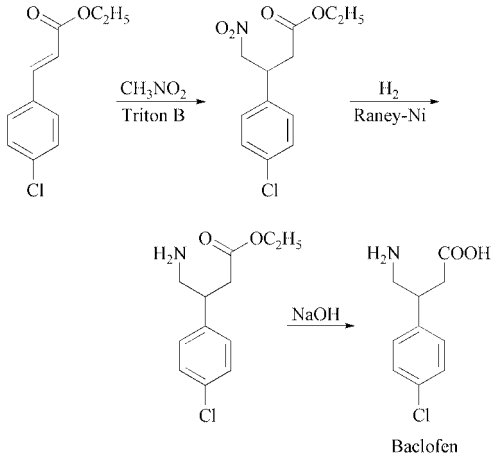

Clinical use: Baclofen, a $GABA_B$ agonist, was initially launched in the 1970s for the oral or parenteral treatment of voluntary muscle spasticity due to cerebrovascular accidents, cerebral palsy, meningitis, multiple sclerosis (MS), spinal lesions, and head injury. Baclofen mediates its effect directly via the $GABA_B$ receptor. An analgesic action of baclofen administered systemically was first described more than 25 years ago and subsequently confirmed by many different groups. Despite the large volume of preclinical data that is now available, the use of baclofen as an analgesic in humans has been limited owing, in part, to rapid tolerance and adverse effects following systemic administration. Possibly, the only type of pain for which baclofen can be used systemically is trigeminal neuralgia. However, when administered intrathecally, baclofen is able to reduce central pain such as occurs in stroke [632].

Trade names: Gabalon, Lioresal (Germany, UK, Italy), Liorésal (France).

Carbamazepine [*298-46-4*], dibenzo[*b*,*f*]-azepine-5-carboxamide, $C_{15}H_{12}N_2O$, M_r 236.27, *mp* 204–206 °C (also reported as 190–193 °C).

Synthesis [633]:

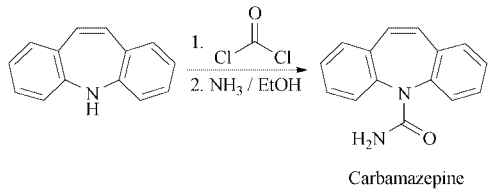

Alternatively, 5*H*-dibenzo[*b*,*f*]azepine can be treated directly with potassium cyanate yielding carbamazepine [634].

Clinical use: Carbamazepine is an orally active anticonvulsant. It is used additionally in the treatment of manic depression and as an analgesic in trigeminal neuralgia and other neuropathic pain conditions [635]. Common side effects are dizziness, drowsiness, ataxia, visual disturbances, dry mouth, abdominal pain, nausea, and vomiting. Other adverse reactions include headache, arrhythmias, and heart block.

Trade names: Tegretal (Germany, Italy), Tegretol (UK, France, USA).

Clonidine [*4205-90-7*], (2,6-dichlorophenyl)-(4,5-dihydro-1*H*-imidazol-2-yl)amine, $C_9H_9Cl_2N_3$, M_r 230.09, *mp* 139 °C; monohydrochloride [*4205-91-8*], $C_9H_9Cl_2N_3 \cdot HCl$, M_r 266.55, *mp* 305 °C.

Synthesis [48, 636, 637]:

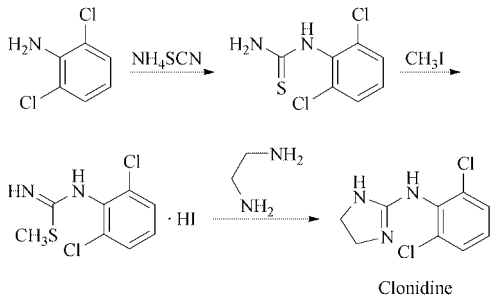

Clinical use: Clonidine is an α_2-adrenergic agonist and is primarily used in the cardiovascular field for blood pressure reduction (→ Antihypertensives, Section 5.1.2) [638]. The compound induces analgesia via central α_2-receptor interaction and can be used orally, parenterally or epidurally for pain treatment, often in combination with opioids or local anesthetics [639]. Epidural clonidine – opioid combinations are preferentially used for neuropathically maintained cancer pain. The compound is additionally used for migraine prophylaxis and to reduce opioid and alcohol withdrawal symptoms [640]. Clonidine frequently induces side effects like hypotension, sedation, drowsiness, dry mouth, and constipation. The compound is well absorbed orally.

Trade names: Duraclon (USA), Catapresan (Germany), Catapres (USA).

Duloxetine [*116539-59-4*], (+)-(*S*)-*N*-methyl-*N*-[3-(naphthalen-1-yloxy)-3-(2-thienyl)propyl]amine, (*S*)-*N*-methyl-γ-(1-naphthalenoxy)-2-thiophenepropanamine $C_{18}H_{19}NOS$, M_r 297.41, *mp* 118–122 °C, $[\alpha]_D^{20}$ +122 ° (*c* = 1.0, CH_3OH); hydrochloride [*136434-34-9*], $C_{18}H_{20}ClNOS_5$, M_r 333.88, *mp* 163–167 °C.

Synthesis: Reaction of 2-acetylthiophene with paraformaldehyde and dimethylamine in ethanol gives 3-(dimethylamino)-1-(2-thienyl)-1-propanone, which is enantioselectively reduced with a 2:1 complex of (2*R*,3*S*)-4-(dimethylamino)-3-methyl-1,2-diphenyl-2-butanol and $LiAlH_4$ in toluene to yield (*S*)-3-(dimethylamino)-1-(2-thienyl)-1-propanol. The condensation of (*S*)-3-(dimethylamino)-1-(2-thienyl)-1-propanol with 1-fluoronaphthalene catalyzed by NaH in DMSO affords the corresponding naphthyl ether (*S*)-*N*,*N*-dimethyl-3-(naphthalen-1-yloxy)-3-(thiophen-2-yl)propan-1-amine, which is finally monodemethylated with 2,2,2-trichloroethyl chloroformate and zinc in toluene and treated with oxalic acid [641, 642].

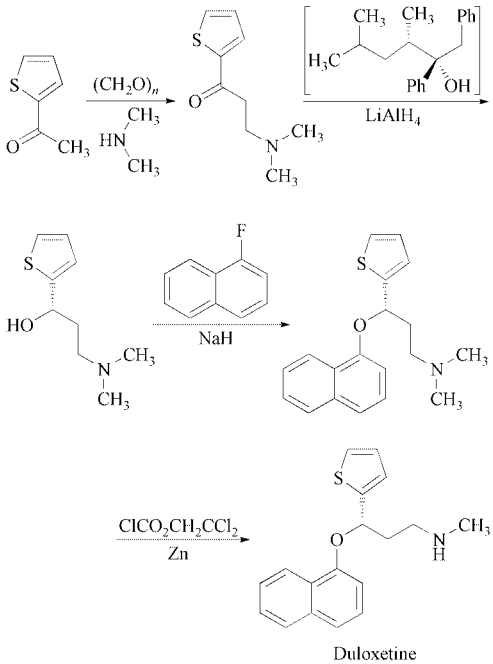

The ketone intermediate 3-(dimethylamino)-1-(2-thienyl)-1-propanone can also be re-

duced with $NaBH_4$ in ethanol to the racemic alcohol 3-(dimethylamino)-1-(thiophen-2-yl)-propan-1-ol, which is submitted to optical resolution with (*S*)-(+)-mandelic acid to provide the (*S*)-enantiomer [641, 642].

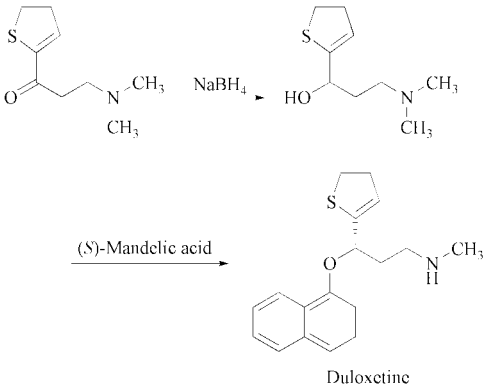

Reaction of thiophene-2-carboxylic acid with oxalyl chloride and triphenylphosphine gives the corresponding acyl chloride thiophene-2-carbonyl chloride, which is condensed with vinyltributylstannane, yielding 1-(2-thienyl)-2-propen-1-one. The addition of HCl to the double bond of 1-(2-thienyl)-2-propen-1-one affords 3-chloro-1-(2-thienyl)-1-propanone, which is reduced with BH_3 and the chiral (*R*)-oxazaborolidine catalyst in THF to give (*S*)-3-chloro-1-(2-thienyl)-1-propanol. Treatment of (*S*)-3-chloro-1-(2-thienyl)-1-propanol with NaI in acetone affords the (*S*)-3-iodopropanol derivative, which is condensed with methylamine in THF to provide (*S*)-3-(methylamino)-1-(2-thienyl)-1-propanol. Finally, this compound is condensed with 1-fluoronaphthalene by means of NaH in DMA [641, 642].

The Friedel – Crafts condensation of thiophene with 3-chloropropionyl chloride in presence of $SnCl_4$ in benzene gives the aforementioned 3-chloro-1-(2-thienyl)-1-propanone, which is reduced with $NaBH_4$ in ethanol to yield the racemic alcohol 3-chloro-1-(thiophen-2-yl)propan-1-ol. Optical resolution of 3-chloro-1-(thiophen-2-yl)propan-1-ol performed by means of immobilized CALB (Novozyme 435, Novo-Nordisk A/S) affords the enantiomer (*S*)-3-chloro-1-(thiophen-2-yl)propan-1-ol [643].

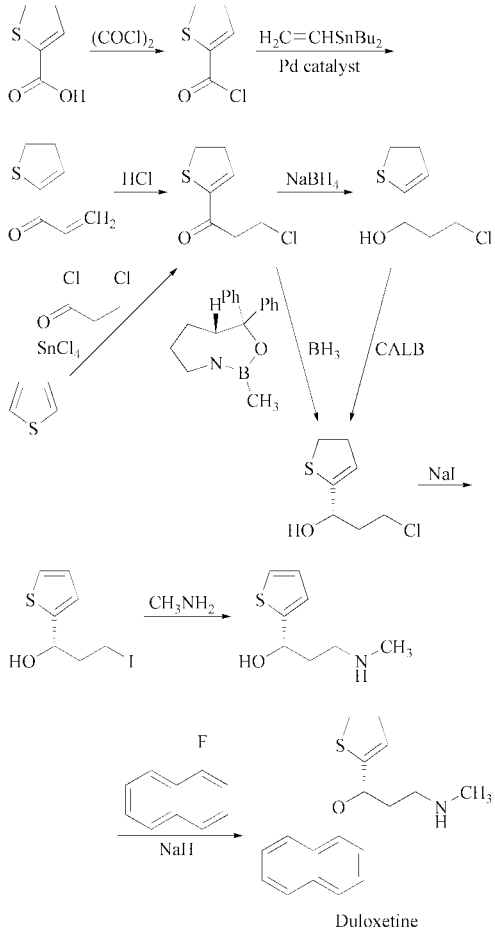

Clinical use: Duloxetine hydrochloride is an inhibitor of 5-HT reuptake and norepinephrine reuptake. The compound was launched in 2004 by originator Lilly for the treatment of major depression, urinary incontinence, and pain caused by diabetic peripheral neuropathy [643–645].

The drug is also undergoing phase III clinical trials at Lilly for the treatment of generalized anxiety and fibromyalgia.

Trade names: Cymbalta (Eli Lilly, UK, France, Italy, Germany), Yentrere (Eli Lilly, USA, UK, France, Italy, Germany), Arichaim (Boehringer Ingelheim, UK, France, Italy, Germany), Xeristar (Boehringer Ingelheim, UK, France, Italy, Germany).

Flupirtine [*56995-20-1*], ethyl[2-amino-6-(4-fluorobenzylamino)pyridin-3-yl]carbamate, $C_{15}H_{17}FN_4O_2$, M_r 304.32, *mp* 115–116 °C; monohydrochloride [*33400-45-2*], $C_{15}H_{17}FN_4O_2 \cdot HCl$, M_r 340.79, *mp* 214–215 °C; maleate (1 : 1) [*75507-68-5*], $C_{15}H_{17}FN_4O_2 \cdot C_4H_4O_4$, M_r 420.39, *mp* 175.5–176 °C.

Synthesis [48, 646–649]:

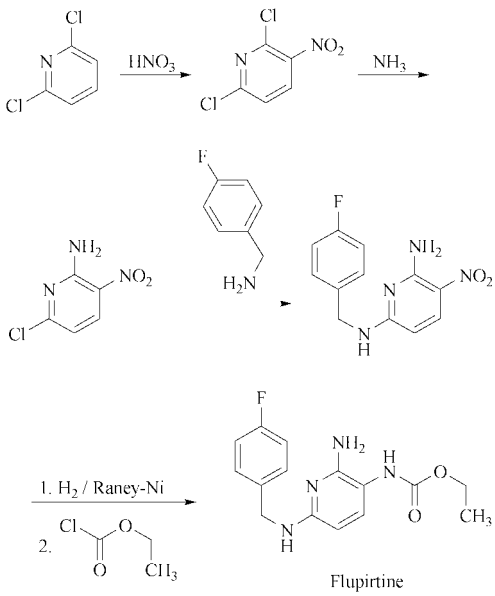

Clinical use: Flupirtine maleate [650] is an analgesic with an unknown mode of action. It is used in single doses of 100 mg orally and 150 mg rectally for the treatment of different pain conditions.

Trade name: Katadolon (Germany).

Gabapentin

Synthesis [651, 653–655]: In the original synthesis (Goedecke) cyclohexenone is reacted with ethyl cyanoacetate in the presence of ammonia to yield the Guareschi salt, which is hydrolyzed and decarboxylated to give 1,1-cyclohexanediacetic acid which is transformed by to the corresponding anhydride with acetic anhydride. This anhydride is treated with methanol to yield the half ester 2-[1-(methoxycarbonyl)cyclohexyl]acetic acid, which is subjected to a Curtius type rearrangement to give the isocyanate 2-[1-(isocyanatomethyl)cyclohexyl]acetic acid. The desired compound is obtained by hydrolysis of 2-[1-(isocyanatomethyl)cyclohexyl]acetic acid with HCl, followed by hydrochloric salt removal via anion exchange [651, 652].

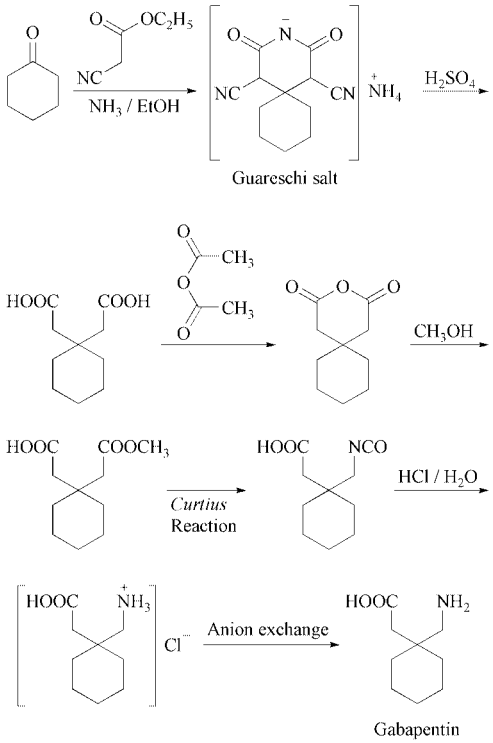

The anhydride can also be treated alternatively with *N*-hydroxylamine to afford the *N*-hydroxyimide, which is converted to the *N*-benzenesulfonyloxylimide by reaction with benzenesulfonyl chloride. Subsequent reaction of the *N*-benzenesulfonyloxylimide with triethylamine in methanol gives the urethane ester methyl 1-(methoxycarbonyl)cyclohexanecarboxylate, which is submitted to HCl hydrolysis to afford the hydrochloride of gabapentin. The free amino acid is obtained from the hydrochloride via anion exchange.

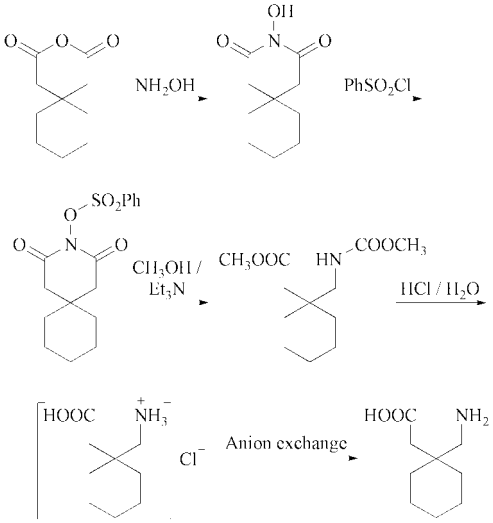

The anhydride can also be treated with ammonia to yield cyclohexane-1,1-diacetic acid monoamide. In this route the compound is submitted to a Hoffman rearrangement to provide the target amino acid.

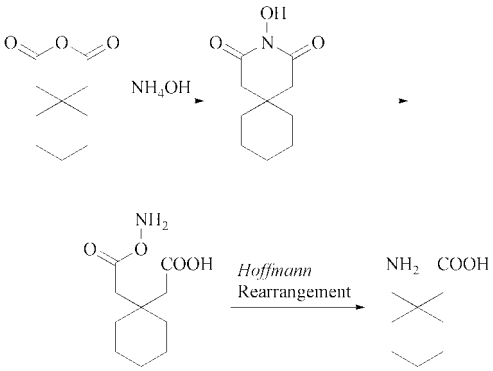

The reaction of 2-[1-(carbamoylmethyl)cyclohexyl]acetic acid with NaOCl and NaOH (a Hofmann rearrangement) gives the isocyanate intermediate, which without isolation hydrolyzes to the target gabapentin.

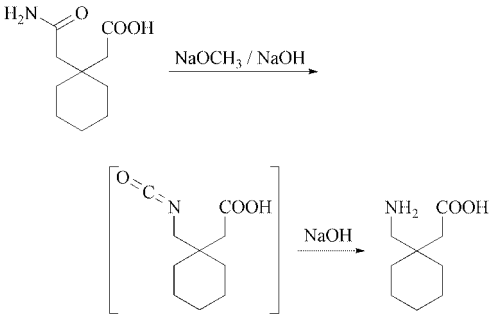

The condensation of cyclohexene-1-carbaldehyde with benzoylhydrazide in dichloromethane gives the corresponding hydrazone (*E*)-*N′*-(cyclohexenylmethylene)benzohydrazide, which is condensated under reductive conditions with trichloroacetyl chloride under catalysis of $BH_3/(CH_3)_2NH$ and Ms-OH (methanesulfonic acid) in diethyl ether to yield the adduct *N′*-(cyclohexenylmethyl)-*N′*-(2,2,2-trichloroacetyl)benzohydrazide. The cyclization of this adduct by means of CuCl and tetramethylenediamine (TMDA) in hot acetonitrile affords the spiro pyrrolidone, which is dechlorinated and debenzamidated by treatment with wet Raney nickel in ethanol at 110 °C to provide 2-azaspiro[4,5]decan-3-one. Finally, this compound is hydrolyzed with hot aqueous HCl to form the target gabapentin [653–655].

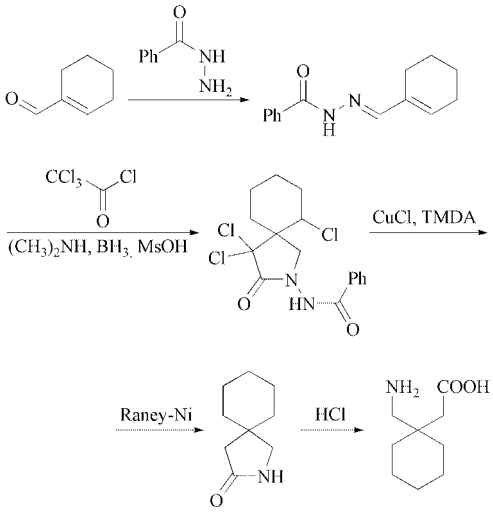

Clinical use: Gabapentin was developed as a structural GABA analogue but it has no direct GABAergic action and it does not affect GABA uptake or metabolism. Preliminary evidence points to the possible effect of gabapentin on the $\alpha 2\delta$ type of calcium channels.

The complex pharmacological mechanism by which gabapentin and the related compound pregabalin exert their clinical effect in epilepsy and neuropathic pain, remains unclear [656, 657].

Gabapentin has a clearly demonstrated efficacy in relieving pain and associated symp-

toms in patients with postdiabetic and postherpetic neuropathy. The starting dosage which can be used is 300 mg three times a day. Increase by 200 mg increments to 200–400 mg three to four times a day is recommended. As main side effects, sedation, dizziness, fatique, and pedal edema are reported [658, 659].

Trade name: Neurontin (Germany).

Gabapentin is also used in many generic formulations.

Lamotrigine [*084057-84-1*], 6-(2,3-dichlorophenyl)-1,2,4-triazine-3,5-diamine, 3,5-diamino-6-(2,3-dichlorophenyl)-1,2,4-triazine $C_9H_7Cl_2N_5$, M_r 256.09, *mp* 216-218 °C.

Synthesis: The reaction of the Grignard compound of 2,3-dichloroiodobenzene with CO_2 in diethyl ether gives 2,3-dichlorobenzoic acid, which is converted to the corresponding acyl chloride by refluxing with $SOCl_2$. The reaction of 2,3-dichlorobenzoyl chloride with cuprous cyanide and KI in refluxing xylene yields 2,3-dichlorobenzoyl cyanide. Finally, this compound is cyclized with aminoguanidine in DMSO to yield lamotrigine [660–663].

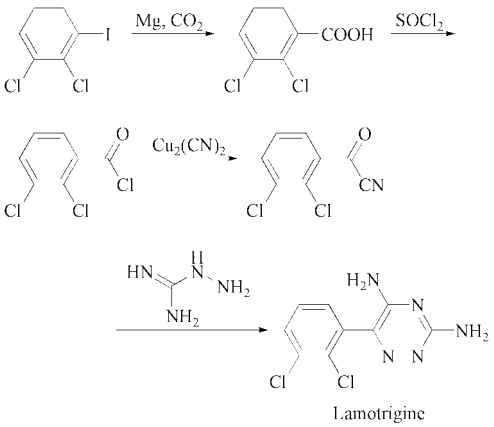

Condensation of 2,3-dichlorobenzoyl cyanide with aminoguanidine in the presence of polyphosphoric acid (PPA) as catalyst in hot acetonitrile yields 2,3-dichlorobenzoylcyanide amidinohydrazone, which is cyclized to the target 1,2,4-triazine by heating in refluxing propanol with or without DMSO.

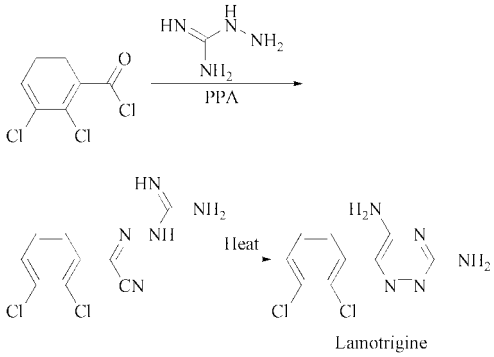

Lamotrigine can also be obtained by the following route: Hydrogenation of 2,3-dichloronitrobenzene with H_2 over Raney nickel in methanol gives 2,3-dichloroaniline, which is diazotized with $NaNO_2$ and HCl and treated with CuCN to yield 2,3-dichlorobenzonitrile. Hydrolysis of the nitrile with NaOH in refluxing methanol/water affords 2,3-dichlorobenzoic acid, which is treated with hot $SOCl_2$ to provide the corresponding acyl chloride. Reaction of 2,3-dichlorobenzoyl chloride with CuCN and KI in refluxing chlorobenzene gives 2,3-dichlorobenzoyl cyanide, which is condensed with aminoguanidine by means of H_2SO_4/TsOH in hot toluene to yield 2,3-dichlorobenzoyl cyanide amidinohydrazone. Finally, this compound is cyclized by treatment with NaOMe in refluxing methanol to yield lamotrigine.

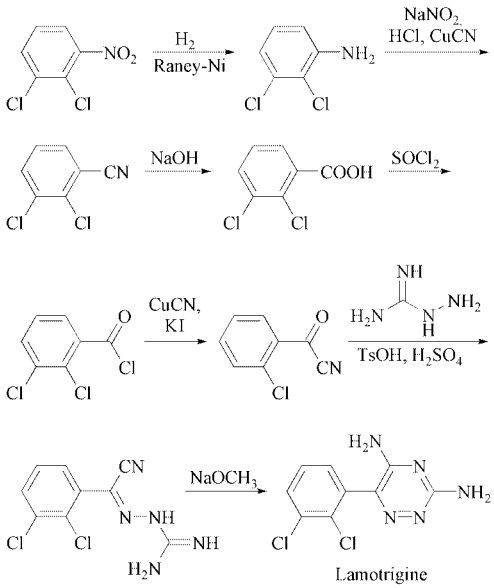

Clinical use: Lamotrigine is a phenyltriazine derivative and one of the newer antiepileptic agents, that stabilize neural membranes by blocking the activation of voltage-sensitive sodium channels and inhibiting the presynaptic release of glutamate. Lamotrigine was initially approved in the USA as add-on therapeutic for partial complex seizures. At doses of 50 to 400 mg/d lamotrigine has demonstrated efficacy in relieving pain in patients with trigeminal neuralgia refractory to other treatments, such as carbamazepine, phenytoin, or both. Pain reduction usually occurs at oral doses of more than 200 mg daily. However, lamotrigine must be dosed very slowly during the first six weeks of treatment to prevent a toxic rash. Recent negative results from a clinical trial indicated that more well-designed trials are needed to better define the role of lamotrigine as an analgesic drug. Main side effects are rash, dizziness, unsteadiness, and drowsiness [658, 659, 664].

Trade name: Lamictal (Germany, UK, France, Italy, USA).

Nefopam [*13669-70-0*], 5-methyl-1-phenyl-3,4,5,6-tetrahydro-1*H*-benzo[*f*][1,4]-oxazocine, $C_{17}H_{19}NO$, M_r 253.34; hydrochloride [*23327-57-3*], $C_{17}H_{19}NO \cdot HCl$, M_r 289.81, *mp* 238–242 °C.

Synthesis [48, 665, 666]:

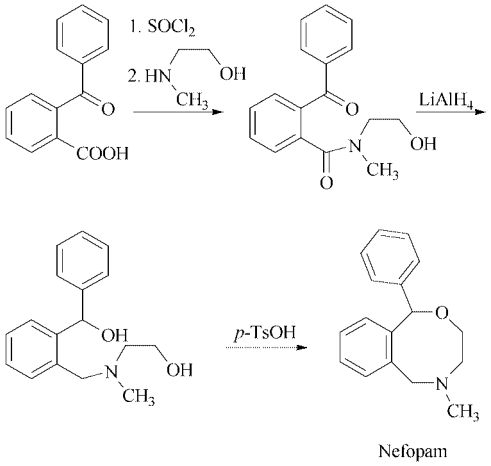

Alternative route:

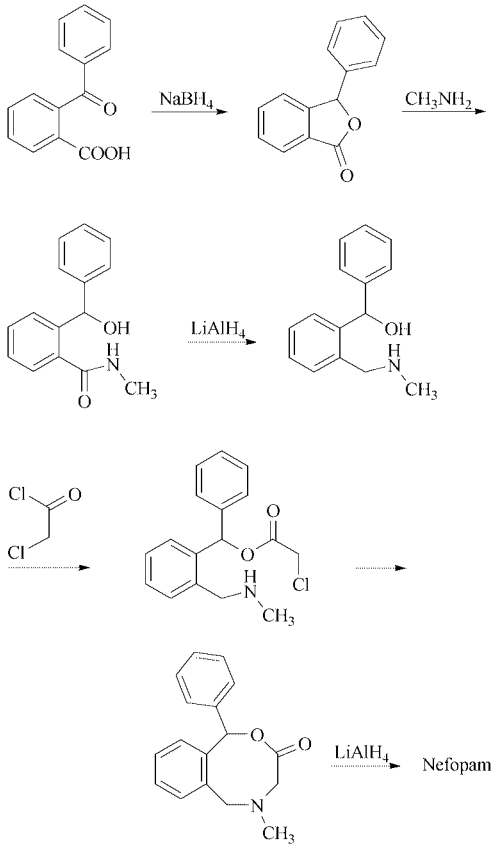

Clinical use: Nefopam [667] is a centrally acting analgesic without opioid properties. It induces inhibition of NA and 5-HT reuptake and has antimuscarinic and sympatomimetic actions which all might be relevant for its analgesic action. Nefopam is used orally and parenterally for the treatment of moderately severe pain. Side effects are nausea, vomiting, sweating, nervousness, tachycardia, and occasionally convulsions. Nefopam should not be used in combination with MAO inhibitors.

Trade names: Ajan (Germany), Acupan (Belgium, France, Italy, UK).

Phenytoin [*561-27-2*], diphenylhydantoin, 5,5-diphenyl-2,4-imidazolidinedione $C_{15}H_{12}N_2O_2$, M_r 252.27, *mp* 295-298 °C; monosodium salt [*690-93-3*], $C_{15}H_{11}N_2NaO_2$, M_r 274.26.

Synthesis:

a) Phenytoin is synthesized by cyclization of benzil with urea [668, 670].

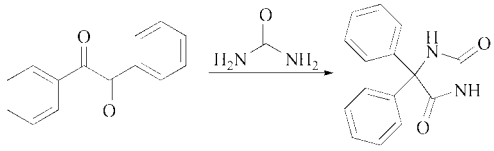

Phenytoin

b) Alternatively phenytoin can also obtained by using benzophenone as a starting material [669, 670]:

Phenytoin

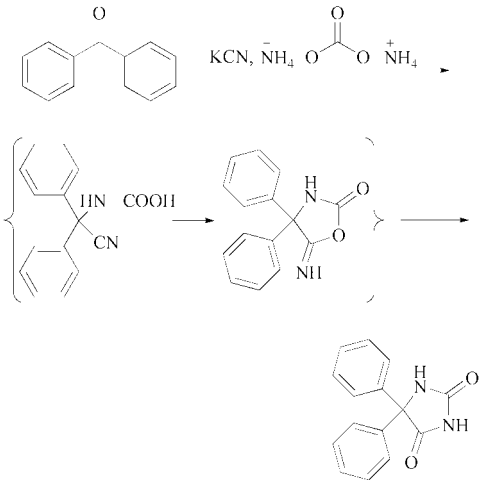

Phenytoin

Clinical use: Phenytoin became the first anticonvulsant to be used to treat neuropathic pain. The first clinical trials started in 1977. Current data indicate that the analgesic effect of phenytoin is achieved through the blockage of sodium channels, inhibition of presynaptic glutamate release and suppression of spontaneous neuronal ectopic discharges. Phenytoin is also used as a co-analgesic in combination therapy. As an oral starting dose, 200 mg at bedtime is recommended. An increase by 100 mg increments to 300–400 mg daily divided in 1–2 doses is possible. Main side effects are slurred speech, rash, unsteadiness, confusion, or blurry vision [658, 659].

Trade names: Epanutin, Phenhydan, Zentropil (Germany), Dihydan (France), Epanutin (UK), Dilantin (USA).

Topiramate [*097240-79-4*], 2,3;4,5-bis-*O*-(1-methylethylidene)-β-D-fructopyranose sulfamate, 2,3;:4,5-bis-*O*-(1-methylethylidene)-1-*O*-sulfamoyl-β-D-fructopyranose, $C_{12}H_{21}NO_8S$, M_r 339.36, *mp* 122-126 °C, $[\alpha]_D^{25}$ −34 ° (*c* = 0.4, CH_3OH).

Synthesis: Topiramate is synthesized by two methods [671, 672]: D-Fructose is reacted with acetone to produce the thermodynamically more stable bisacetonide. This bisacetonide is then condensed with sulfamoyl chloride in the presence of sodium hydride. Alternatively, the intermediate bisacetonide is reacted with sulfuryl chloride and pyridine in methylene chloride and then with sodium azide in acetonitrile to furnish the azido sulfate, which is reduced with copper powder in methanol [671, 672]

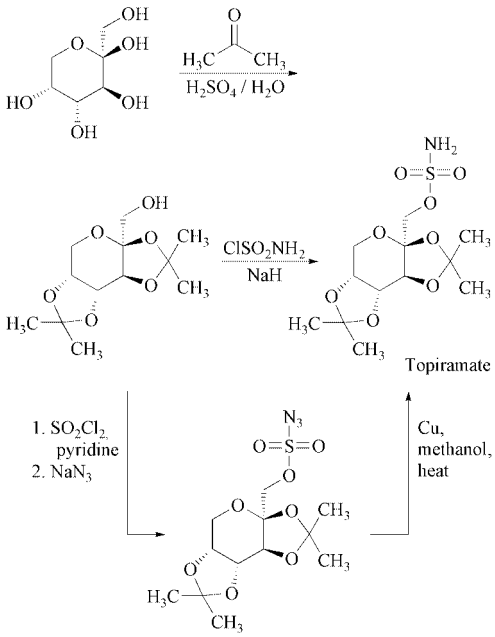

Topiramate

Alternatively the reaction of diacetone fructose with sulfamide in presence of of pyridine in xylene at 128-133 °C yields the target topiramate [673, 674].

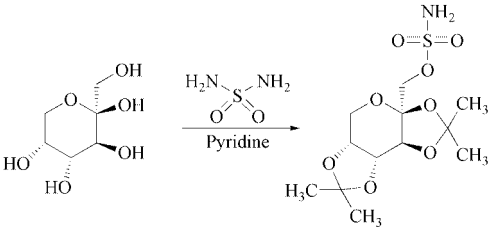

Clinical use: Topimarate is a sulfamate-substituted derivative of D-fructose with the ability to block sodium channels, enhance GABA activity by interacting with a nonbenzodiazepine site of $GABA_A$ receptors, and selectively block AMPA/kainate glutamate receptors [658].

The empirical use of topimarate as an antinociceptive drug in humans was established before systematic, planned research on animal models of pain was initiated. Topiramate was initially launched in 1995 as monotherapy and adjunctive therapy for the treatment of epilepsy, specifically for the treatment of partial seizures, Lennox–Gastaut syndrome, and primary generalized tonic clonic seizures in patients inadequately controlled by other anticonvulsants. Topiramate is also used for the prophylaxis of migraine headaches.

The oral starting dose of topimarate is 25–50 mg at bedtime and can be increased by 25 mg increments weekly to 100–200 mg twice a day. Main reported side effects include confusion, weakness, weight loss, and kidney stones.

The compound had been under development for the treatment of neuropathic pain, bipolar disorders, obesity, pathological gambling, abuse and dependency, and for smoking cessation.

Trade names: Epitomax, Topamax, Topimax.

Ziconitide [*097240-79-4*], H-Cys-Lys-Gly-Lys-Gly-Ala-Lys-Cys-Ser-Arg-Leu-Met-Tyr-Asp-Cys-Cys-Thr-Gly-Ser-Cys-Arg-Ser-Gly-Lys-Cys-NH_2 cyclic S-3.1-S-3.16:S-3.8-S-3.20:S-3-15-S-3.25-tris(disulfide), omega-conotoxin MVIIA (reduced), cyclic (1-16),(8-20),(15-25)-tris(disulfide), $C_{102}H_{172}N_{36}O_{32}S_7$, M_r 2639.14.

Synthesis: Ziconitide is the synthetic form of the conotoxin ω-conopeptide MVIIA [675, 676] (see below).

Clinical use: Ziconotide is the synthetic equivalent of a naturally occurring conopeptide found in the marine snail *Conus magus*. Ziconotide is an N-type calcium channel blocker, that was launched in the USA by Elan in 2005 as a solution for intrathecal infusion for the management of severe, chronic pain in patients who are refractory to other treatments [675, 676]. Regulatory approval has been obtained for the product in the EU. The drug is administered through appropriate programmable microinfusion pumps, that can be implanted or external, and which release the drug into the fluid surrounding the spinal cord.

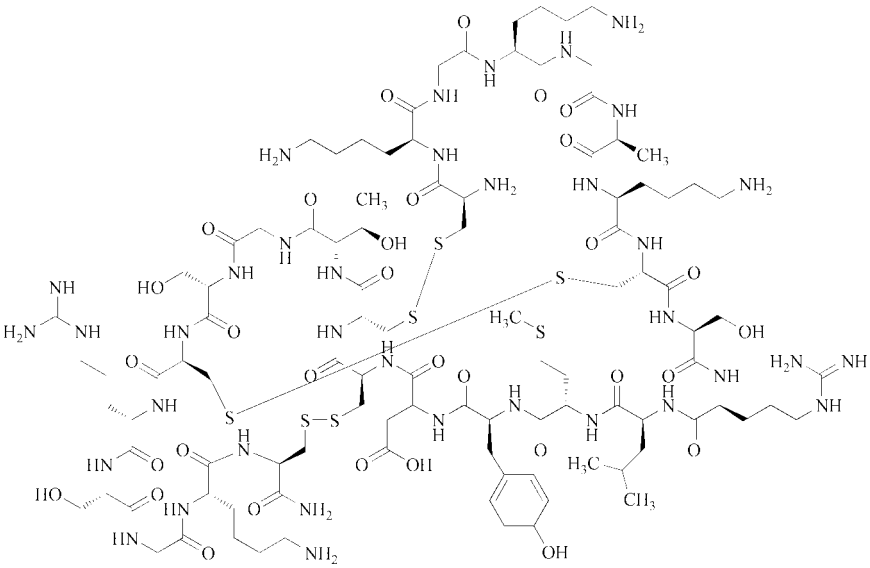

Ziconotide binds to N-type calcium channels located on the primary nociceptive (Aγ and C) afferent nerves in the superficial layers (rexed laminae I and II) of the dorsal horn in the spinal cord. Although the mechanism of action of the drug has not been established in humans, results in animals suggest that blocking of N-type calcium channels leads to a blockade of excitatory neurotransmitter release from the primary afferent nerve terminals which induces antinociception.

In July 2001, ziconotide received orphan drug designation from the EMEA for the treatment of chronic pain requiring intraspinal analgesia. The compound had also been under evaluation at Elan for the treatment of head injury.

Intended trade name: Prialt.

4. References

General References

1. J. G. Hardman (eds.) et al.: *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 9th ed., McGraw Hill, New York 1995, pp. 521–555, 557–600, 617–657.
2. J. E. F. Reynolds (eds.) et al.: *Martindale The Extra Pharmacopoeia*, 31st ed., The Royal Pharmaceutical Society, London 1996.
3. A. Herz (ed.): *Handbook of Experimental Pharmacology*, vol. 104/I, Opioids I., vol. 104/II, Opioids II, Springer Verlag, Berlin 1993.
4. T. Dickenson, J. Besson (eds.): *Handbook of Experimental Pharmacology. The Pharmacology of Pain*, vol. 130, Springer Verlag, Berlin 1997.
5. M. Zenz, I. Jurna (eds.): *Lehrbuch der Schmerztherapie. Grundlage, Theorie und Praxis für Aus- und Weiterbildung*. Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart 1993.
6. J. J. Bonica (ed.): *The Management of Pain*, 2nd ed., Lea & Febiger, 1990.
7. S. Budavari (ed.): *The Merck Index*, 12th ed., Merck & Co. Inc., 1996.
8. P. D. Wall, R. Melzack (eds.): *Textbook of Pain*, 3rd ed., Churchill Livingstone, New York 1994.
9. J. R. Vane et al., "Cyclooxygenases 1 and 2," *Annu. Rev. Pharmacol. Toxicol.* **38** (1998) 97–120.
10. C. J. Hawkey: "Cox-2 Inhibitors", *Lancet* **353** (1999) 307–314.

Specific References

11. J. R. Vane et al., *Ann. Rev. Pharmacol. Toxicol.* **38** (1998) 97 120.
12. K. L. Pierce et al., *TIPS* **16** (1995) 253 256.
13. C. D. Funk, et al., *FASEB J.* **5** (1991) 2304–2312.
14. D. A. Kujubu et al., *J. Biol. Chem.* **266** (1991) 12866–72.
15. T. Hla, K. Neilson, *Proc. Natl. Acad. Sci. U S A.* **89** (1992) 7384–8.
16. M. K. O'Banion, et al., *Proc. Natl. Acad. Sci. U S A.* **89** (1992) 4888–4892.
17. R. Langenbach et al., *Ann. N. Y. Acad. Sci.* **889** (1999) 52–61.
18. S. G. Morham et al., *Cell* **83** (1995) 473–482.
19. J. E. Dinchuk et al., *Nature* **378** (1995) 406–409.
20. D. Picot et al., *Nature* **367** (1994) 243–249.
21. J. R. Kiefer et al., *Nature* **405** (2000) 97–101.
22. J. R. Vane et al., *Annu. Rev. Pharmacol. Toxicol.* **38** (1998) 97–120.
23. J. Berg et al., *J. Pharmacol. Toxicol. Meth.* **37** (1997) 179–186.
24. A. Graul et al., *Drugs Future* **22** (1997) 711–714.
25. L. J. Crofford et al., *J. Clin. Invest.* **93** (1994) 1095–101.
26. H. Sawaoka et al., *Prostaglandins Leukot. Essent. Fatty Acids* **59** (1998) 313–316.
27. C. J. McCarthy et al., *Am. J. Gastroenterol.* **94** (1999) 1218–1223.
28. R. Langenbach et al., *Ann. N.Y. Acad. Sci.* **889** (1999) 52–61.
29. L. J. Crofford, *Curr. Pharm. Des.* **17** (2000) 1725–36.
30. C. Bombardier et al., *N. Engl. J. Med.* **343** (2000) 1520–1528.
31. R. S. Bresalier et al., *N. Engl. J .Med.* **352** (2005) 1092–1102.
32. S. D. Solomon et al., *N. Engl. J. Med.* **352** (2005) 1071–1180.
33. N. A. Nussmeier et al., *N. Engl. J. Med.* **352** (2005) 1081–1091.
34. B. F. McAdam et al., *Proc. Natl. Acad. Sci. U.S.A.* **96** (1999) 272–277.
35. X. Leval et al., *Curr. Med. Chem.* **9** (2002) 941–962.
36. M. Inagaki et al., *J. Med. Chem.* **43** (2000)2040-2048.
37. S. Tries et al., *Inflamm. Res.* **51** (2002) 135–143.
38. S. Laufer et al., *Arzneimittelforschung* **44** (1994) 1329–1333.
39. B. Chin, J. Wallace, *Curr. Opin. Investig. Drugs* **1** (1999) 148–152.

40. C. Pisano et al., *Dig. Dis. Sci.* **44** (1999) 713–724.
41. S. Fiorucci, E. Antonelli, *Curr. Top. Med. Chem.* **5** (2005) 487–492.
42. R. Scatena, *Curr. Opin. Investig. Drugs* **5** (2004) 551–556.
43. J. L. Wallace et al., *Am. J. Physiol. Gastrointest. Liver Physiol.* **286** (2004) G76–81.
44. C. Patrono, *N. Engl. J. Med.* **330** (1994) 1287–1294.
45. Patrignani et al., *Br. J. Pharmacol.* **118** (1996) 1285–1293.
46. R. G. Kurumbail et al., *Nature* **384** (1996) 644–648.
47. A. S. Kalgutkar et al. *Science* **280** (1998) 1268–1270.
48. A. Kleemann, J. Engel, B. Kutscher, D. Reichert: *Pharmaceutical Substances, Synthesis, Patents, Applications*, Thieme, Stuttgart – New York, 1999 (available in print and CD-ROM).
49. N. Kuhnert, *Chem. Unserer Zeit* **33** (1999) 213–220.
50. *Am. J. Med.* **74** (1983) no. 6A, 1–109.
51. I. Tegeder et al., *FASEB J.* **15** (2001) 2057–2072.
52. C. J. Needs, P. M. Brooks, *Clin. Pharmacokinet.* **10** (1985) 164–177.
53. R. J. Waldmann et al., *JAMA* **247** (1982) 3089–3094.
54. Merck & Co., US 4 225 730, 1980, (H. Jones, R. Houser.
55. For alternative synthesis: US 3 992 495, (Merck & Co.).
56. For alternative synthesis: US 4 131 618, (Merck & Co.).
57. J. Hannah et al., *J. Med. Chem.* **21** (1978) 1093.
58. E. Arrigoni-Martelli, *Drugs Future* **3** (1978) 28–33.
59. J. Hannah et al., *J. Med. Chem.* **21** (1978) 1093–1100.
60. M. Cotton, R. A. Hux, *Anal. Profiles Drug Subst.* **14** (1985) 491–526.
61. J.M. Young et al., *Inflamm Res.* **45** (1996) 246–253.
62. Lunbeck, GB 656 746, 1948.
63. D. S. Hoffenberg, C. R. Hauser, *J. Org. Chem.* **20** (1955) 1496–1500.
64. Warner-Lambert, US 2 998 450, 1961, (G. Wilbert, J. De Angelis).
65. B. Ameer et al., *Ann. Int. Med.* **87** (1977) 202–209.
66. D. A. Willoughby et al., *Lancet* **355** (2000) 646–648.
67. P. Kirkpatrick, *Nature Reviews Drug Discovery* **4** (2005) 883.
68. E. D. Hogestatt et al., *J. Biol. Chem.* **36** (2005) 31405–31412.
69. R. K. Lewis, F. P. Paloucek, *Clin. Pharm.* **10** (1991) 765–774.
70. M. Depre et al., *Fundam. Clin.Pharmacol.* **6** (1992) 259–262.
71. U. C. Dubach et al., *N. Engl. J. Med.* **308** (1983) 357–362.
72. K. H. Boltze, H. Kreisfeld, *Arzneim.-Forsch./Drug Res.* **27** (1977) no. 1, 1300–1312.
73. R. Coletta et al., *Int. J. Clin. Pharmacol. Res.* **8** (1988) 295–298.
74. Parke Davis, FR 1 341 M, 1961.
75. R. B. Moffett, B. D. Aspergen, *J. Am. Chem. Soc.* **82** (1960) 1605.
76. S. Ravi et al., *Postgrad. Med. J.* **62** (1986) 773–776.
77. H. Gögelein et al., *FEBS Lett.* **268** (1990) 79–82.
78. M. M. White et al., *Mol. Pharmacol.* **37** (1990) 720–724.
79. L. Sorbera et al., *Drugs of the Future* **26** (2001) 133–140.
80. G. J. Grover et al., *J. Pharmacol. Exp. Ther.* **269** (1994) 536–540.
81. Y. T. Lee, Q. Wang, *Eur. J. Pharmacol.* **378** (1999) 349–356.
82. Parke Davis, US 3 313 848, 1967; DE 1 149 015, 1961.
83. P. F. Juby et al., *J. Med. Chem.* **11** (1968) 111–117.
84. E. Arrigoni-Martelli, *Drugs Future* **3** (1978) 307–310.
85. J. R. McLean, M. I. Geuckman, *Arzneimittel-Forsch.* **33** (1983) 627–631.
86. A. Peretz et al. *Mol. Pharmacol.* **67** (2005) 1053–1066.
87. Y. T. Lee, Q. Wang, *Eur. J. Pharmacol.* **378** (1999) 349–356.
88. Parke Davis, US 3 138 636, 1964; DE 1 163 846, 1961.
89. G. R. Singh, P. Choudhury, *Indian Pediatr.* **26** (1989) 577–579.
90. M. Ottolia, L. Toro, *Biophys. J.* **67** (1994) 2272–2279.
91. J. S. Marks, M. H. Gleeson, *Br. Med. J.* **4** (1975) 442.
92. Labs. U.P.S.A., US 3 415 834, 1968; US 3 337 570, 1967.
93. J. Auclair et al., *Curr. Ther. Res.* **46** (1989) 782–788.

94. M. Civelli et al., *Agents Actions* **33** (1991) 233–239.
95. Park Davis, US 3313848 (A. Scherer, F. W. Short), 1967
96. M. Zovko, B. Zorc, M. J.-M. Takac, B. Metelko, P. Novac, *Croatica Chemica Acta* **76** (2003), 335–341.
97. T. Corell, *Pharmacol. Toxicol. 75 Suppl* **2** (1994) 14–21.
98. H. Isomaki, *Pharmacol. Toxicol.* 75 Suppl **2** (1994) 64–65.
99. J. Delgado et al. *Pharmacol. Toxicol. 75 Suppl* **2** (1994) 89–91.
100. P. E. Hansen, *Pharmacol. Toxicol.* 75 Suppl **2** (1994) 81–82.
101. K. H. Boltze et al., *Arzneim.-Forsch./Drug Res.* **30** (1977) no. 2 1314–1325.
102. H. Jacobi, H. D. Dell, *Arzneimittel-Forsch.* **30** (1980) 1348–1362.
103. Rotta Research Lab., DOS 2 535 799, 1975.
104. Rotta Research Lab., US 3 985 878, 1974.
105. F. Makovec et al., *Arzneimittelforschung* **37** (1987) 806–813.
106. V. Fossaluzza et al., *Minerva Med.* **74** (1983) 723–726.
107. N. P. Buu-Hoi et al., *Hebd. Seances Acad. Sci.* **261** (1965) 2259.
108. P. Moser et al., *J. Med. Chem.* **33** (1990) 2358–2368. Link
109. P. A. Todd, E. M. Sorkin, *Drugs* **35** (1988) 244–285.
110. T. Y. Shen et al., *J. Am. Chem. Soc.* **85** (1963) 488–489.
111. Merck & Co., US 3 161 654, 1964, DE 1 232 150, 1962.
112. H. Yamamoto et al., *Chem. Pharm Bull.* **16** (1968) no. 17, 647.
113. Byk Gulden, US 4 4 325 962, 1982; DE 1 946 370, 1969.
114. G. Rainer et al., *Arzneim.-Forsch./Drug Res.* **31** (1981) 649–655.
115. B. Unterhalt, *Drugs Future* **7** (1982) 110–111.
116. Merck & Co., US 3 882 239, 1975; DE 1 039 426, 1970.
117. R. F. Shuman et al., *J. Org. Chem.* **42** (1977) 1914–1919.
118. R. N. Brogden et al., *Drugs* **16** (1987) 97–114.
119. A. Whelton et al., *JAMA* **249** (1983) 2892.
120. T. D. Warner et al., *Proc. Natl. Acad. Sci. U.S.A.* **96** (1999) 7563–7568.
121. C. Haanen, *Curr. Opin. Investig. Drugs* **2** (2001) 677–683
122. McNeil, US 3 752 826, 1973; GB 1 428 272, 1973.
123. J. R. Carson et al., *J. Med. Chem.* **14** (1971) 646–647.
124. R. N. Brogden, *Drugs* **15** (1978) 429–450.
125. C. Brideau et al., *Inflamm Res.* **45** (1996) 68–74.
126. E. Arrigoni-Martelli, J. Prous, J. Castaner, *Drugs of the Future* **14** (1989) 963.
127. G. Biasi, R. Marcolongo, *Minerva Med.* **92** (2001) 315–324.
128. G. Coruzzi et al., *Curr. Drug Targets Inflamm. Allergy.* **3** (2004) 43–61.
129. G. Coruzzi et al., *Dig. Liver Dis.* **34** (2002) 403–410.
130. J. Prous, J. Castaner, *Drugs of the Future* **13** (1988) 1050.
131. CDC Life Sciences Inc., US 4 327 222, 1982 (R. G. Micetich et al.).
132. Teiho Pharmaceuticals, EP 454 871, 1990 (H. Tanaka, Y. Hagiwara, H. Kajitani, H. Yasumoto).
133. Taiho Pharmaceuticals, EP 464 218, 1991 (H. Tanaka, Y. Hagiwara, H. Kajitani, H. Yasumoto).
134. CDC Life Science, EP 26928, 1981 (R. G. Micetich et al.).
135. T. Ushio et al, *Iykuhin Kenkyu* **22** (1991) 892.
136. N. Ono et al., *Nippon Yakurigaku Zasshi.* **95** (1990) 63–81.
137. K. Goto et al., *Prostaglandins Other Lipid Mediat.* **56** (1998) 245–254.
138. D. A. Walch et al., *J. Med. Chem.* **27** (1984) 1379–1388.
139. US 4 126 635, 1977 (H. Robins et al.).
140. A.H. Robinson Co., Inc. US 4 568 695, 1983 (H. Robins et al.).
141. J. Prous, J. Castaner, *Drugs of the Future* **13** (1988) 943.
142. T. Ogawa et al. *Nippon Ganka Gakkai Zasshi* **99** (1995) 406–411.
143. R. J. Fontana et al., *Liver Transpl. Surg.* **5** (1999) 480–484.
144. E. B. Hunter et al., *Am. J. Gastroenterol.* **94** (1999) 2299–2301.
145. P. L. Moses et al., *Am. J. Gastroenterol.* **94** (1999) 1393–1396.
146. Beecham, US 4 420 639, 1983; DE 2 463 219, 1974.
147. A. C. Goudie et al., *J. Med. Chem.* **12** (1978) 1260–1264.
148. M. Neumann, *Drugs Future* **6** (1981) 35–36.
149. C. Prabhakar et al., *Org. Process Res. Dev.* **3** (1999) 121–125.
150. H. A. Friedel et al., *Drugs* **45** (1993) 131–156.
151. Degussa, US 3 513 171, 1970; DE 1 670 522, 1966.
152. *Drugs Future* **8** (1983) 773–775.

153. W. v. Bebenburg, et al., *Chem. Ztg.* **105** (1981) 217–219.
154. W. v. Bebenburg, G. Steinmetz, K. Thiele, *Chem. Ztg.* **103** (1979) 387–399.
155. M. L. Brochier, *Eur. Heart J.* **14** (1993) 951–957.
156. W. J. Wechter et al., *Cancer Res.* **60** (2000) 2203–2208.
157. J. M. Mayer, B. Testa, *Drugs Future* **22** (1997) 1347–1366.
158. M. Cleij et al., *J. Org. Chem.* **64** (1999) 5029–5035.
159. M. Busson, *J. Int. Med. Res.* **14** (1986) 53–62.
160. Rhône-Poulence, US 3 641 127, 1972.
161. G. G. Liversidge, *Anal. Profiles Drug Subst.* **10** (1981) 443–471.
162. J. L. Hommeril et al., *Br. J. Anaesth.* **72** (1994) 383–387.
163. P. Patrignani et al., *J Physiol. Pharmacol.* **48** (1997) 623–631.
164. R. L. Dorfman, *Arzneim.-Forsch./Drug Res.* **25** (1975) 278–281.
165. T. Y. Shen, *Angew. Chem. Int. Ed. Engl.* **11** (1972) 460.
166. H. R. Bellur, J. R. Ahuja, D. G. Kulkarni, *Tetrahedron: Asymmetry* **3** (1992) 163–192.
167. Syntex, US 3 663 584, 1972.
168. Syntex, US 4 605 758, 1986.
169. P. A. Todd, S. P. Clissold, *Drugs* **40** (1990) 91–137.
170. J. Rautio et al., *J. Med. Chem.* **43** (2000) 1489–1494.
171. Roussel-Uclaf, DOS 2 055 264, 1970.
172. G. L. Plosker, A. J. Wagstaff, *Drugs* **50** (1995) 1050–1075.
173. F. G. Mayall et al., *Br. Med. J.* **309** (1994) 599–600.
174. J. Prous, J. Castaner, *Drugs of the Future* **14** (1989) 302.
175. Nippon Chemiphar, US 4 247 706, 1981 (Y. Fujimoto, S. Yamabe).
176. M. Hatori, S. Kokubun, *Curr. Med. Res. Opin.* **14** (1998) 79–87.
177. N. Murakami et al., *Jpn. J. Pharmacol.* **72** (1996) 29–37.
178. H. B. Tang et al., *Neuropharmacology* **48** (2005) 1035–1042.
179. R. W. Carney et al., *Experientia* **29** (1973), 938.
180. Ciba, US 3 641 040, 1972 (R. W. Carney, G. DeStevens).
181. P. A. Todd, R. Beresford, *Drugs.* **32** (1986) 509–537.
182. M. Guidotti et al., *J. Int. Med Res.* **17** (1989) 48–54.
183. J. Pirotte et al., *Gastroenterol. Clin. Biol.* **11** (1987) 520.
184. J. R. Hannequin et al., *Rev. Rheum. Mal. Osteoartic.* **55** (1988) 983–988.
185. A. C. Depla et al., *Neth. J. Med.* **37** (1990) 32–36.
186. M. Eugene et al., *Rev. Med. Interna* **8** (1987) 115.
187. Ely Lilly, DOS 1 941 625, 1969.
188. Ely Lilly, US 3 600 437, 1971 (W. S. Marshall).
189. Mitsubishi Petrochemical, DOS 2 646 792, 1976 (M. Takeda, M. Uchide, H. Masayuka).
190. Nissin Flour Milling, US 4 016 196, 1977 (K. Katsura Kogure, T. Noriyoshi Sueda, T. Youziro Yoshino, K. Kunio Nakagawa).
191. Sagami, DAS 2 709 504, 1977 (G. Tsuchihashi, K. Ogura).
192. C. M. Gruber Jr., *J. Rheumatol.* **2** (1976) 8–17.
193. R. N. Brogden et al., *Drugs* **13** (1977) 241–265.
194. R. N. Brogden et al., *Drugs* **21** (1981) 1–22.
195. J. R. Ryan et al., *Clin. Pharmacol. Ther.* **42** (1987) 28–32.
196. M. L. Wendland et al., *Mayo Clin. Proc.* **55** (1980) 103–107.
197. J. S. Stotts et al., *J. Am. Acad. Dermatol*. **18**(4 Pt 1) (1988) 755–757.
198. CERM, US 4 218 473, 1980 (J. J. Godfroid, E. Steiner).
199. M. Hoeijmakers et al., *J. Vet. Pharmacol. Ther.* **28** (2005) 305–312.
200. T. Nell et al., *J. Small Anim. Pract.* **43** (2002) 208–212.
201. P. Lees et al., *J. Vet. Pharmacol. Ther.* **22** (1999) 96–106.
202. D. A. Walsh et al., *J. Med. Chem.* **33** (1990), 2296–2304.
203. Alcon, WO 09932104, 2000 (J. H. Yanni, M. R. Hellberg).
204. T. L. Ke et al., *Inflammation* **24** (2000) 371–384.
205. K. Takahashi et al., *Invest. Ophthalmol. Vis. Sci.* **44** (2003) 409–415.
206. D. A. Gamache et al., *Inflammation* **24** (2000) 357.
207. R. Linstrom, T. Kim, *Curr. Med. Res. Opin.* **22** (2006) 397–404.
208. W. J. Welstead et al., *J. Med. Chem.* **22** (1979), 1074–1079.
209. Robins Co., Inc., DE 2 324 768, 1973 (W. Welstead, H. W. Moran)
210. Robins Co., Inc., US 1972 254 285, 1972 (W. Welstead, H. W. Moran)

211. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **1**, Verlag Chemie, Weinheim 1972, pp. 171–172.
212. I.G. Farben, DRP 476 663, 1922.
213. Z. Fendrich, *Cas. Lek. Cesk.* **139** (2000) 440–444.
214. F. Arellano, J A. Sacristan, *Eur. J. Clin. Pharmacol.* **38** (1990) 617–619.
215. M. Levy et al., *Clin. Pharmacokinet.* **28** (1995) 216–234.
216. Hoffmann-La Roche, DRP 565 799, 1931.
217. Riedel-deHaen, DE 962 254, 1954.
218. K. Brune, *Agents Actions* **25** Suppl. (1988) 9–19.
219. E. Gores et al., *Med. Monatsschr. Pharm.* **27** (2004) 72–76.
220. R. Denss et al., *Helv. Chim. Acta* **40** (1957) 402.
221. M. Kühn et al., *J. Prakt. Chem.* **156** (1940) 103.
222. Horakova et al., *Pharmacotherapeutica* (1963) 335–350.
223. J. Büchi et al., *Helv. Chim. Acta* **36** (1953) 75.
224. Comm. Farmaceutica Milanese, GB 839 057, 1957.
225. D. Loew et al., *Z. Rheumatol.* **44** (1985) 186–192.
226. Geigi, US 2 745 783, 1956.
227. L. E. Westerholm, *Br. Med. J.* **3** (1973) 339–343.
228. Geigi, US 2 562 830, 1951.
229. S. L. Ali, *Anal. Profiles Drug Subs.* **11** (1982) 483–521.
230. G. A. Faich, *Pharmacotherapy* **7** (1987) 25–27.
231. Warner-Lambert, DOS 2 208 351, 1972.
232. J. G. Lombardino, E. H. Wisemann, *J. Med. Chem.* **14** (1971) 973.
233. H. Zinnes et al., *J. Med. Chem.* **25** (1982) 12.
234. W. W. Downie, *Semin. Arthritis Rheum.* **12** (1982) Suppl. 2, 170–174.
235. *Drugs Future* **17** (1992) 683–686.
236. T. Pruss et al., *Postgrad. Med. J.* **66** (1990) Suppl. 4818–821.
237. T. Christoph et al., *Inflamm. Res.* **44** (1995) Suppl. 3 Abstr. W16/17.
238. J. A. Balfour et al., *Drugs* **51** (1996) 639–657.
239. Pfizer, US 3 591 584, 1971; US 4 469 866, 1984.
240. J. G. Lombardino et al., *J. Chiaini, J. Med. Chem.* **5** (1973) 493–496.
241. E. H. Wisemann et al., *Arzneim.-Forsch./Drug Res.* **26** (1976) 1300–1303.
242. L. Castaner, E. Arrigoni-Martelli, *Drugs Future* **2** (1977) 124–127.
243. Hoffmann-La Roche, DOS 2 537 070, 1975.
244. R. N. Brogden et al., *Drugs* **22** (1981) 165–187.
245. R. N. Brogden et al., *Drugs* **28** (1984) 292–323.
246. J. A. Mitchell et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 11693–11697.
247. E. Arrigoni-Martelli, *Drugs Future* **7** (1982) 493–494.
248. P. A. Todd, S. P. Clissold, *Drugs* **41** (1991) 625–646.
249. Synthex, US 4 347 186, 1982; US 4 458 081, 1984; US 4 873 340, 1989.
250. E. Arrigoni-Martelli, *Drugs Future* **8** (1983) 871–874.
251. A. Guzman et al., *J. Med. Chem.* **29** (1986) 589–591.
252. J. M. Muchowski et al., *J. Med. Chem.* **28** (1985) 1037–1049.
253. F. F. Franco et al., *J. Org. Chem.* **47** (1982) 1682–1688.
254. M. M. T. Buckley, R. N. Brogden, *Drugs* **39** (1990) 86–109.
255. John Wyeth and Brothers, Ltd, US 3578671, 1971 (K. Brown).
256. E. Arrigoni-Martelli, *Drugs of the Future* **3** (1978) 539–542.
257. L. G. Miller, *Clin. Pharm.* **11** (1992) 591–603.
258. American Home Products, US 3 843 681, 1974.
259. L. Castaner, E. Arrigoni-Martelli, *Drugs Future* **2** (1977) 21–23.
260. C. A. Demerson, et al., *J. Med. Chem.* **19** (1976) 391–395.
261. C. A. Demerson et al., *J. Med. Chem.* **18** (1975) 189.
262. N. Bellamy, *Inflammopharmacology* **5** (1997) 139–152.
263. A. S. Taha et al., *Gut* **30** (1989) A751.
264. G. Bianchi Porro, et al., *J. Intern. Med.* **229** (1991) 5–8.
265. Thomae GmbH, DE 2 756 113, 1979.
266. G. Trummlitz, G. Engelhardt, U. Busch, *Drugs Future* **14** (1989) 1047–1048.
267. G. Engelhardt et al., *Inflamm. Res.* **44** (1995) 423–433.
268. J. A. Lopez-Garcia et al., *Neuroreport* **9** (1998) 647–651.
269. Riker, US 3 840 597, 1974; DOS 2 333 643, 1973.
270. R. Davis, R. N. Brogden, *Drugs* **48** (1994) 431–454.
271. A. Azab et al., *Life Sci.* **63** (1998) PL 323–327.

272. P. A. McCormick et al., *Lancet* **353** (1999) 40–41.
273. C. Ferreiro et al., *Gastroenterol. Hepatol.* **23** (2000) 428–430.
274. *Drugs Future* **22** (1997) 711–714.
275. Searle & Co., US 5 521 207; EP 731 795; WO 9 515 316 (J. J. Talley, T. D. Penning, P. W. Collins et al.).
276. G. Steinbach et al., *N. Engl. J. Med.* **342** (2000) 1946–1952.
277. J. L. Goldstein et al., *Am. J. Gastroenterol.* **96** (2001) 1019–1027
278. F. E. Silverstein et al., *JAMA* **284** (2000) 1247–1255.
279. J. M. Brophy, *Expert Opin. Drug. Saf.* **4** (2005) 1005–1015.
280. L.A. Sorbera et al., *Drugs of the Future* **26** (2001) 346–353.
281. R.W. Friesen et al., *Biorg. Med. Chem. Lett.* **8** (1999) 2777–2782.
282. I. W. Davies et al., *J. Org. Chem.* **65** (2000) 8415–8420.
283. I. W. Davies et al., *J. Org. Chem.* **65** (2000) 8571–8574.
284. J. F. Marcoux et al., *Org. Lett.* **2** (2000) 2339–2341.
285. A. Dallob et al., *J Clin Pharmacol.* **43** (2003) 573–585.
286. D. J. Cochrane et al., *Drugs.* **62** (2002) 2637–2651.
287. A. D. Rodrigues et al., *Drug Metab. Dispos.* **31** (2003) 224–232.
288. K. Kassahun et al., *Drug Metab. Dispos.* **29** (2001) 813–820.
289. N. Chauret et al., *Bioorg. Med. Chem. Lett.* **11** (2001) 1059–1062.
290. D. R. Ramey et al., *Curr. Med. Res. Opin.* **21** (2005) 715–722.
291. S. P. Curtis et al., *Clin Ther.* **26** (2004) 70–83.
292. S. Aldington et al., *N. Z. Med. J.* **118** (2005) U1684.
293. L. A. Sorbera et al., *Drugs of the Future* **27** (2002) 740–747.
294. Novartis, WO 2001 023346, 2001 (M. Murat, T. Allemendiger, J. Calienni, J. Cercus, O. Loiseleur).
295. Novartis, WO 1999 011605, 1999 (L. Mcquire, B. Mugrage, J. Van Duzer).
296. K. A. Lyseng-Williamson et al., *Drugs* **64** (2004) 2237–2246.
297. F. Berenbaum et al., *J. Int. Med. Res.* **33** (2005) 21–41.
298. P. Geusens et al., *Int. J. Clin. Pract.* **58** (2004) 1033–1041.
299. E. Packman et al., *Headache* **45** (2005) 1163–1170.
300. J. A. Rosenberg et al., *Nat. Clin. Pract. Gastroenterol. Hepatol.* **2** (2005) 14–15.
301. C. J. Hawkey et al., *Clin. Gastroenterol. Hepatol.* **4** (2006) 57–66.
302. P. Matchaba et al., *Clin. Ther.* **27** (2005) 1196–214.
303. H. Tannenbaum et al., *Ann. Rheum. Dis.* **63** (2004) 1419–1426.
304. C. M. Rordorf et al., *Clin. Pharmacokinet.* **44** (2005) 1247–1266.
305. Pfizer Inc., WO 2005123701, 2005 (L. J. Letendre, C. K. Snoddy, G. H. Klemm, J. P. Lawson, M. J. Bauer).
306. J. J. Talley et al., *J. Med. Chem.* **43** (2000) 1661–1663.
307. Searle & Co. WO 9515316;
US 5 521 207;
EP 731 795 1995 (J. J -Talley et al.).
308. S. M. Cheer, K. L. Goa, *Drugs* **61**(2001) 1133–1141.
309. A. W Gotta, *Idrugs* **4** (2001) 939–944.
310. A. S. Dalpiaz, D. Peterson, *Expert. Rev. Neurother.* **4** (2004) 165–177.
311. S. Aldington et al., *N. Z. Med. J.* **118** (2005) U1755.
312. C. D. Furberg et al., *Circulation* **111** (2005) 249.
313. E. G. Harks et al., *J. Pharmacol. Exp. Ther.* **298** (2001) 1033–1041.
314. L. A. Sorbera et al. *Drugs Future* **23** (1998) 1287–1296.
315. L. A. Sorbera et al., *Drugs of the Future* **23** (1998) 1287–1296.
316. M. Therien et al., *Synthesis* (2001) 1778–1779.
317. L. J. Scott, H. M. Lamb *Drugs.* **58** (1999) 499–505.
318. A. J. Matheson, D. P. Figgitt, *Drugs.* **61** (2001) 833–865.
319. C. C. Chan et al., *J. Pharmacol. Exp. Ther.* **290** (1999) 551–560.
320. L. Laine et al., *Gastroenterology* **117** (1999) 776–783.
321. C. Bombardier et al., *N. Engl. J. Med.* **343** (2000) 1520–1528.
322. M. Wadman, *Nature* **433** (2005) 790.
323. J. J. Talley in F. D. King A. W. Oxford (eds.): *Progress in Medicinal Chemistry-Vol 36*, Elsevier, Amsterdam 1999.
324. D. Ormrod et al., *Drugs* **62** (2002) 2059–2071.
325. M. L. Chavez, C. J. De Korte *Clin Ther.* **25** (2003) 817–851.
326. C. Fenton et al. *Drugs.* **64** (2004)1231–1261.
327. J. E. Edwards et al. *Pain* **111** (2004) 286–296.

328. G. P.Joshi *Expert Rev. Neurother.* **5** (2005) 11–24.
329. D. Kudrow et al. *Headache* **45** (2005) 1151–1162.
330. C. Talhari et al., *J. Allergy Clin. Immunol.* **115** (2005) 1089–1090.
331. F. L. Byerly et al., *Burns* **31** (2005) 383–387.
332. N. A. Nussmeier et al. *N. Engl. J. Med.* **352** (2005) 1081–1091.
333. A. F. Casy, R. T. Parfitt: *Opioid Analgesics, Chemistry and Receptors*, Plenum Press, New York 1986.
334. A. F. Casy, R. T. Parfitt, *Opioid Analgesics. Chemistry and Receptors*, Plenum Press, New York -London 1986.
335. A. F. Casy, R. T. Parfitt, *Opioid Analgesics. Chemistry and Receptors*, Plenum Press, New York -London 1986.
336. H. Akil et al. *Annu. Rev. Neurosci.* **7** (1984) 223–255.
337. A. Mansour et al., *Trends Neurosci.* **18** (1995) 22–29.
338. A. Herz J. Neural. *Transm. Suppl.* **18** (1983) 227–233.
339. G. B. Stefano, et al. *Crit. Rev. Immunol.* **16** (1996) 109–144.
340. C. Stein et al.,in C. Stein (ed.): *Opioids in Pain Control. Basic and Clinical Aspects*, Cambridge University Press, Cambridge 1999, pp. 96–108.
341. W.R. Martin et al., *J. Pharmacol. Exp Ther.* **197** (1976) 517–5132.
342. C. J. Fowler G. L. Fraser *Neurochem. Int.* **24** (1994) 401–426.
343. W. R. Martin, *Pharmacol. Rev.* **35** (1983) 283–323.
344. B. N. Dhawan et al., *Pharmacol Rev.* **48** (1996) 567–592.
345. J. C. Meunier et al., *Nature* **377** (1995) 532–535.
346. P. L. Wood, *Neuropharmacology* **21** (1982) 487–497.
347. G. W. Pasternak, P. J. Wood, *Life Sci.* **38** (1986) 1889–1898.
348. G. Gaveriaux-Ruff, B. Kieffer, in C. Stein, (ed.): *Opioids in Pain Control. Basic and Clinical Aspects*, Cambridge University Press, Cambridge 1999, pp. 1–20.
349. B. L. Kieffer et al., *Proc. Natl. Acad. Sci. U S A.* **89** (1992) 12048–12052.
350. C. J. Evans et al., *Science,* **258** (1992) 1952–1955.
351. R. K. Reinscheid et al. *Science,* **270** (1995) 792–794.
352. G. Calo' et al. *Peptides,* **21** (2000) 935–947.
353. R. J. Knapp et al., *FASEB J.* **9** (1995) 516–525.
354. S. R. Childers *Life Sci.* **48** (1991) 1991–2003.
355. I. McFadzean, *Neuropeptides* **11** (1988) 173–180.
356. W. F. Simonds, *Endocr. Rev.* **9** (1988) 200–212.
357. A. D. Blake et al. *Channels* **5** (1997) 231–235.
358. P. Y. Law, H. H. Loh, *J. Pharmacol. Exp. Ther.* **289** (1999) 607–624.
359. N. I. Cherny, *Drugs* **51** (1996) 713–737.
360. H. W. Matthes et al., *Nature* **383** (1996) 819–823.
361. B. L. Kieffer, *Trends Pharmacol. Sci.* **20** (1999) 19–26.
362. D. I. Scopes, *Drug of the Future* **18** (1993) 933–947.
363. A. Barber, R. Gottschlich, *Exp. Opin. Invest. Drugs* **6** (1997) 1354–1368.
364. G. Dondio et al., *Exp. Opin. Ther. Patents* **7** (1997) 1075–1098.
365. M. A. Scheideler, *Curr. Opin. CPNS Invest Drugs* **2** (2000) 171–177.
366. C. Mollereau, L. Mouledous, *Peptides.* **21** (2000) 907–917.
367. R. Bertorelli et al., *Trends Pharmacol. Sci.* **21** (2000) 233–234.
368. A. Mansour et al., *Trends Neurosci.* **18** (1995) 22–29.
369. J. Lipp, *Clin. Neuropharmacol.* **14** (1991) 131–147.
370. T. L Yaksh, *Acta Anaesthesiol. Scand.* **41** (1997) 94–111.
371. C. Stein, *Opioids in Pain Control. Basic and Clinical Aspects*, Cambridge University Press, Cambridge 1999.
372. C. P. Watson, *Clin. J. Pain*, **16** (2 Suppl.) (2000) S49–55.
373. R. C. Etches et al., *Can. J. Anaesth.* **36** (1989) 165–185.
374. J. E. Shook et al., *Am. Rev. Respir. Dis.* **142** (1990) 895–909.
375. M. F. Mulroy, *Reg. Anesth.* **21** (6 Suppl.) (1996) 89–93.
376. F. W. Jackson, *Gastroenterology* **106** (1994) 820–821.
377. T. A. Bowdle, *Drug Saf.* **19** (1998) 173–189.
378. W. Kromer, *Dig. Dis.* **8** (1990) 361–373.
379. I. Mancini, E. Bruera, *Support Care Cancer* **6** (1998) 356–364.
380. A. De Luca, I. M. Coupar, *Pharmacol. Ther.* **69** (1996) 103–115.
381. R. C. Heel et al. *Drugs.* **15** (1978) 33–52.
382. E. Campora et al., *J. Pain Symptom Manage.* **6** (1991) 428–430.
383. R. Aparasu et al., *J. Pain Symptom Manage.* **18** (1999) 280–288.

384. J. P. Blancquaert et al. *Eur. J. Pharmacol.* **128** (1986) 143–150.
385. N. M. Barnes et al., *Neuropharmacology* **30** (1991) 1073–1083.
386. A. Pfeiffer et al., *Science* **233** (1986) 774–776.
387. R. Brown, R. Lo, *Aust. N. Z. .J Ment. Health Nurs.* **9** (2000) 65–74.
388. D. A. Taylor, W. W. Fleming, *Pharmacol. Exp. Ther.* **297** (2001) 11–18.
389. H. A. Heit, *Eur. J. Pain* **5** Suppl A (2001) 27–29.
390. R. F. Reder, *Eur. J. Pain.* **5** Suppl A (2001) 109–111.
391. F. J. Muhtadi, *Anal. Profiles Drug Subs.* **17** (1988) 259–366.
392. Knoll, DOS 2 726 925, 1977.
393. T. Hudlicky et al.: "A historical perspective of morphine synthesis," in: Atta-ur-Rahman (ed.): *Studies in Natural products Chemistry*, Elsevier, Amsterdam 1996, pp. 43–116.
394. D. Trauner et al., *Synthesis* **1983** 203–204.
395. J. Mulzer, D. Trauner, *Chirality* **11** (1999) 475–482.
396. J. Frackenpohl, *Chem. u. Zeit* **34** (2000) 99–112.
397. G. J. Meuzelaar et al., *Eur. J. Org. Chem.* (1999) 2315–2321.
398. J. Mulzer, D. Trauner, *Chirality* **11** (1999) 475–482.
399. H. Nagata, et al. *Chem. Commun.* (2001) 1094–1095.
400. H. J. McQuay, *B. J. Anaest.* **63** (1989) 213–226.
401. L L. Christrup, *Acta Anaesthesiol. Scand.* **41** (1997) 116–122.
402. R. Osborne et al., *Clin. Pharmacol. Ther.* **47** (1990) 12–19.
403. J. Lotsch, G. Geisslinger, *Clin. Pharmacokinet.* **40** (2001) 485–499.
404. G. Ehrhart, H. Ruschig: *Arzneimittel*, Entwicklung, Wirkung, Darstellung, vol. **1**, Verlag Chemie, Weinheim 1972, pp. 117–118.
405. Boehringer, DRP 247 180, 1912.
406. B. L. Posey, N. S. Kimble, *J. Anal. Toxicol.* **8** (1984) 68–74.
407. F. J. Muhtadi, M. M. A. Hassan, *Anal. Profiles Drug Subs.* **10** (1981) 93–138.
408. Q. Y. Yue et al., *Ther. Drug Monit.* **19** (1997) 539–542.
409. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **1**, Verlag Chemie, Weinheim 1972, 118.
410. E. Merck AG, DRP 131 980, 1902.
411. E. Tani et al., *Arzneimittelforschung* **39** (1989) 1399–1402.
412. V. Ozdemir et al., *Pharmacogenetics* **10** (2000) 373–388.
413. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **1**, Verlag Chemie, Weinheim 1972, pp. 79, 115.
414. C.E. Inturrisi et al., *Life Sci. 33 Suppl* **1** (1983) 773–776.
415. J. G. Umans, C. E. Inturrisi, *J. Pharmacol. Exp. Ther.* **218** (1981) 409–415.
416. D. K. Wyatt, L. T. Grady, *Anal. Profiles Drug Subst.* **10** (1981) 357–403.
417. U. Boerner et al., *Drug Metab. Rev.* **4** (1975) 39–73.
418. C. E. Inturrisi et al., *N. Engl. J. Med.* **310** (1984) 1213–1217.
419. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **1**, Verlag Chemie, Weinheim 1972, p. 180.
420. A. Stein, *Pharmazie* **10** (1955) 180–186.
421. J. E. Edwards et al. *Cochrane Database Syst. Rev.* (2000) CD 2760.
422. H. Matthys et al., *Schweiz. Med. Wochenschr.* **115** (1985) 307–311.
423. R. S. Lloyd et al., *Curr. Med. Res. Opin.* **13** (1992) 37–48.
424. Reckitt & Colman, DE 1 620 206, 1966.
425. Reckitt & Sons Ltd., US 3 433 791, 1969.
426. K. W. Bentley, D. G. Hardy, *Proc. Chem. Soc. (London)* **1963**, 220.
427. W. G. Dorner, *Chem. Unserer Zeit* **15** (1986) 33–46.
428. K. O. Lee et al., *Eur. J. Pharmacol.* **378** (1999) 323–330.
429. G. F. Blane, *J. Pharm. Pharmacol.* **19** (1967) 367–373.
430. T. Alford et al., *J. Am. Vet. Med. Ass.* **164** (1974) 702–705.
431. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **1**, Verlag Chemie, Weinheim 1972, pp. 119–120.
432. Knoll, DRP 607 931, 1935.
433. Z. R. Chen et al., *Life Sci.* **48** (1991) 2165–2171.
434. A. B. Morrison, *Can. Med. Assoc J.* **120** (1979) 1338.
435. Merck, US 2 715 626, 1955.
436. H. Rapoport, R. Naumann, E. R. Bissel, R. M. Bonner, *J. Org. Chem.* **15** (1950) 1103–1107.
437. G. Lindena et al., *Schmerz* **12** (1998) 195–204.
438. D. E. Moulin et al., *Lancet* **337** (1991) 465–468.
439. J. L. Hill, J. P. Zacny, *Psychopharmacology (Berlin).* **152** (2000) 31–39.

440. G. Ehrhart, H. Ruschig: *Arzneimittel*, Entwicklung, Wirkung, Darstellung, vol. **1**, Verlag Chemie, Weinheim 1972, pp. 130–131.
441. O. Schnider, A. Grüssner, *Helv. Chim. Acta* **34** (1951) 2211–2217.
442. O. Schnider, J. Hellerbach, *Helv. Chim. Acta* **33** (1950) 1437–1448.
443. R. Grewe, *Naturwissenschaften* **33** (1946) 333–337.
444. S. R. Childers, *Life Sci.* **48** (1991) 1991–2003.
445. L. L. Randall, G. Lehmann, *J. Pharmacol. Exp. Ther.* **99** (1950) 163.
446. S. W. Coniam, *Anaesthesia* **46** (1991) 518.
447. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol. **I**, Verlag Chemie, Weinheim 1972, 118.
448. E. Merck, DRP 411 530, 1925.
449. K. W. Bentley: *The Chemistry of Morphine Alkaloids*, Oxford 1954.
450. F. M. Hauser et al., *J. Med. Chem.* **17** (1974) 1117.
451. J. Cleary et al., *J. Pharmacol. Exp. Ther.* **271** (1994) 1528–1534.
452. E. Kalzo, A. Vanino, *Clin Pharmacol. Ther.* **47** (1990) 639–646.
453. F. Kaiko et al., *Clin. Pharmacol. Ther.* **59** (1996) 52–61.
454. U. Weiss, *J. Am. Chem. Soc.* **77** (1955) 5891–5892.
455. R. S. Sinatra, D. M. Harrison, *Clin. Pharm.* **8** (1989) 541–544.
456. V. S. Copland et al., *Am. J. Vet. Res.* **48** (1987) 1626–1630.
457. Winthrop, US 2 167 351, 1939, (O. Eisleb).
458. I. G. Farben, DE 679 281, 1937.
459. O. Eisleb, *Chem. Ber.* **74** (1941) 1433–1450.
460. E. E. Smissman, G. Hite, *J. Am. Chem. Soc.* **81** (1959) 1201–1203.
461. C. B. Pert, S. H. Snyder, *J. Med. Chem.* **19** (1976) 1248–1250.
462. N. P. Fish, N. J. De-Angelis, *Anal. Profiles Drug Subs.* **1** (1972) 175–205.
463. K. Jiraki, *Am. J. Forensic Med. Pathol.* **13** (1992) 42–43.
464. D. Meyer, V. Halfin, *J. Clin. Psychopharmacol.* **1** (1981) 319–321.
465. C. F. Seifert, S. Kennedy, *Pharmacotherapy* **24** (2004) 776–783.
466. J. G. Stevenson et al., *Jt. Comm. J. Qual. Saf.* **30** (2004) 277–281, 233.
467. A. B. O'Connor, et al., *Am. J. Med.* **118** (2005) 885–889.
468. I.G. Farben, DRP 752 755, 1942.
469. H. Kägi, K. Miescher, *Helv. Chim. Acta* **32** (1949) 2489–2507.
470. C. B. Christensen, *Pharmacol. Toxicol.* **73** (1993) 344–345.
471. L. Jylli et al., *Acta Anaesthesiol. Scand.* **48** (2004) 1256–1259.
472. S. Andersen et al., *Pain* **67** (1996) 369–374.
473. P. Peltola, *P. Soisalo, AMA Arch Intern Med.* **101** (1958) 741–746.
474. Janssen, US 3 164 600, 1965; US 3 141 823, 1964.
475. R. Hess et al., *J. Pharmacol. Exp. Ther.* **179** (1971) 474–484.
476. M. A. Clotz, M. C. Nahata, *Clin. Pharm.* **10** (1991) 581–593.
477. J. F. Buchanan, C. R. Brown, *Med. Toxicol.* **3** (1988) 1–17.
478. Janssen, DE-OS 2 819 837, 1978.
479. F. Janssen et al., *J. Med. Chem.* **29** (1986) 2290–2297.
480. S. J. Hopkins, *Drugs Future* **6** (1981) 335–337.
481. R. F. Cookson et al., *Br. J. Anaesth. 55 Suppl* **2** (1983) 147S–155S.
482. C. J. E. Niemegeers, P. A. J. Janssen, *Drug. Dev. Res.* **1** (1981) 83.
483. G. E. Larijani, M. E. Goldberg, *Clin. Pharm.* **6** (1987) 275–282.
484. Glaxo, EP 383 579, 1990.
485. *Drugs Future* **19** (1994) 1088–1092.
486. K. Flick et al., *Arzneim.-Forsch./Drug Res.* **28** (1978) no. 1, 107–113.
487. W. E. Hoffman et al., *Anesthesiology* **79** (1993) 107–113.
488. T. E. Egan, *Clin. Pharmacokinet.* **29** (1995) 80–94.
489. F. Servin, *Eur. J. Anaesthesiol.* **15** Suppl.(1997) 41–44.
490. C. G. Haigh, *Minerva Anesthesiol.* **66** (2000) 414–416.
491. J. Cohen, D. l. Royston, *Curr. Opin. Crit. Care* **7** (2001) 227–231.
492. L. G. Michelsen et al., *J. Clin. Anesth.* **8** (1996) 679–682.
493. Janssen, US 3 998 834, 1976.
494. L. Castaner, E. Arrigoni-Martelli, *Drugs Future* **2** (1977) 334–337.
495. W. F. M. Van Bever, et al., *Arzneim.-Forsch./Drug Res.* **26** (1976) 1548–1556.
496. P. G. H. Van Daele et al., *Arzneim.-Forsch./Drug Res.* **26** (1976) 1521–1531.
497. J. E. Leysen et al., *Eur. J. Pharmacol.* **87** (1983) 209–225.
498. I. P. Monk et al., *Drugs* **36** (1988) 286–313.
499. M. Goldberg et al., *Anesthesiology* **63** (1985) 199–201.

500. J. A. Grass, *J. Pain Symptom Manage.* **7** (1992) 271–286.
501. M. Goldberg et al., *Anesthesiology* **63** (1985) 199–201.
502. Farbw. Hoechst, DBP 865 314, 1941.
503. Merck & Co., US 2 644 010, 1953.
504. Abott, US 2 983 757, 1961.
505. E. M. Schultz et al., *J. Am. Chem. Soc.* **69** (1947) 2454–2459.
506. N. R. Easton et al., *J. Am. Chem. Soc.* **69** (1947) 2941–2942.
507. A. A. Larsen et al., *J. Am. Chem. Soc.* **70** (1948) 4195–4196.
508. E. E. Howe, M. Sletzinger, *J. Am. Chem. Soc.* **71** (1949) 2935.
509. M. Bockmühl, G. Ehrhart, *Justus Liebigs Ann. Chem.* **561** (1948) 52–85.
510. S. K. Sim, *Can. Med. Assoc. J.* **109** (1973) 615–619.
511. R. H. Biskara, *Anal. Profiles Drug Subs.* **3** (1974) 365–439.
512. G. D. Olson et al., *Clin. Pharmacol. Ther.* **21** (1977) 147–157.
513. M. P.xs Davis, D. Walsh, *Support Care Cancer* **9** (2001) 73–83.
514. H. Joseph et al., *Mt. Sinai J. Med.* **67** (2000) 347–364.
515. M. Farell et al., *Br. Med. J.* **309** (1994) 997–1001.
516. M. J. Kreek, *JAMA.* **223** (1973) 665–668.
517. J. Lowinson et al., *Curr. Psychiatr. Ther.* **18** (1978) 165–172.
518. F. I. Carroll et al., *J. Org. Chem.* **41** (1950) 3521–3524.
519. S. A. Walczak et al., *Eur. J. Pharmacol.* **72** (1981) 343–349.
520. J. D. Blaine et al. *Ann. N. Y. Acad. Sci.* **362** (1981) 101–115.
521. A. Wolven et al., *NIDA Res. Monogr.* (1976) 29–38.
522. J. Kang et al., *Eur. J. Pharmacol.* **458** (2003) 25–29.
523. R. L. Deamer et al., *J. Addict Dis.* **20** (2001) 7–14.
524. Janssen, DE 1 117 126; 1956; GB 822 055, 1956.
525. A. J. Grace, *Br. Med. J.* **281** (1980) 1285.
526. A. Pohland, H. R. Sullivan, *J. Am. Chem. Soc.* **75** (1953) 4458–4461; **77** (1955) 3400–3401.
527. A. Pohland et al., *J. Org. Chem.* **28** (1963) 2483–2484.
528. H. R. Sullivan et al., *J. Org. Chem.* **28** (1963) 2381–2385.
529. D. Boven et al., *Hum. Toxicol.* **3** (1984) Suppl., 1S–238S.
530. A. A. Lawson, D. B. Northridge, *Med. Toxicol. Adverse Drug Exp.* **2** (1987) 430–444.
531. Janssen, US 2 898 340, 1959; Searle, US 4 086 234, 1978.
532. C. J. Niemegeers et al., *Arzneimittelforschung* **22** (1972) 516–518.
533. D. D. Hung, *Anal. Profiles Drug. Subst.* **7** (1978) 149–169.
534. Janssen, US 3 714 159, 1973; ED-OS 2 126 559, 1971.
535. R. A. Stokbroekx et al., *J. Med. Chem.* **16** (1973) 782–786.
536. M. R. Dashwood et al., *Prog. Clin. Biol. Res.* **328** (1990) 165–169.
537. R. C. Heel et al., *Drugs* **15** (1978) 33–52.
538. R. C. Heel et al., *Drugs.* **15** (1978) 33–52.
539. D. L. DeHaven-Hudkins, et al., *Life Sci.* **71** (2002) 2787–2796.
540. F. Awouters et al., *Dig. Dis. Sci.* **38** (1993) 977–995.
541. R. C. Heel et al., *Drugs.* **15** (1978) 33–52.
542. C. D. Ericsson P. C. Johnson, *Am. J. Med.* **88**(6A) (1990) 10S–14S.
543. Janssen, DE 1 238 472, 1961.
544. F. Weyne et al., *Acta anaesth. belg.* **19** (1968) 33–45.
545. Gödecke, DE 1 518 959, 1965; Warner-Lambert, US 3 557 126, 1969.
546. L. E. Overman et al., *J. Org. Chem.* **44** (1979) 4183–4185.
547. G. Satzinger, *Justus Liebigs Ann. Chem.* **758** (1972) 43–64, 65–71; **728** (1969) 64–87.
548. G. Satzinger et al., *Pharm. Ind.* **40** (1978) 657–664.
549. R. Schulz et al., *Naunyn Schmiedebergs Arch. Pharmacol.* **304** (1978) 89–93.
550. M. Hermann et al., *Arzneimittel-Forsch.* **20** (1970) 977–983.
551. K. O. Vollmer, *Fortschr. Med.* **106** (1988) 593–596.
552. Grünenthal, GB 997 399, 1964; US 3 564 100, 1971 (E. Frankus, K. Flick).
553. K. Flick et al., *Arzneim.-Forsch./Drug Res.* **28** (1978) no. 1, 107–113.
554. M. C. Frink et al., *Arzneimittelforschung* **46** (1996) 1029–1036.
555. E. Friderichs et al., *Arzneimittel-Forsch.* **28** (1978) 122–134.
556. R. B. Raffa, E. Friderichs, *Pain Rev.* **3** (1996) 249–271.
557. R. B. Raffa et al., *J. Pharmacol. Exp. Ther.* **267** (1993) 331–340.
558. C. R. Lee et al., *Drugs* **46** (1993) 313–340.
559. K. McClellan, L. J. Scott, *Drugs* **63** (2003) 1079–1088.

560. P,Huang et al., *J. Pharmacol. Exp. Ther.* **297** (2001) 688–695.
561. B. Kögel et al. *Eur. J. Pain.* **9** (2005) 599–611.
562. R. C. Heel et al., *Drugs.* **17** (1979) 81–110.
563. J. Martin, *J.Psychoactive Drugs Suppl* **2** (2004) 129–137.
564. S. Sung, J. M. Conry, *Ann. Pharmacother.* **40** (2006) 501–505.
565. T. J. Gal, *Clin. Pharmacol. Ther.* **45** (1989) 66–71.
566. N. Griessinger et al., *Curr. Med. Res Opin.* **21** (2005) 1147–1156.
567. A. Dahan et al., *Br. J. Anaesth.* (2006) **96**, 627–632.
568. Bristol-Myers, US 3 775 414, 1973.
569. J. Monkovic et al., *J. Am. Chem. Soc.* **95** (1973) 7910.
570. J. Monkovic et al., *J. Am. Chem. Soc.* **100** (1978) 4609–4610.
571. C. E. Rosow, *Acute Care 12 Suppl* **1** (1988) 2–7.
572. R. C. Heel et al., *Drugs* **16** (1978) 473–505.
573. B. Ameer, F. J. Salter, *Am. J. Hosp. Pharm.* **36** (1979) 1683–1691.
574. R. V. Homan, *Am. Fam. Physician* **49** (1994) 188–192.
575. American Home, BE 776 173, 1971; DE 2 159 324, 1971.
576. M. E. Freed et al., *J. Med. Chem.* **16** (1973) 595–599.
577. J. C. Chen et al., *Life Sci.* **52** (1993) 389–396.
578. J. J. O'Brien, et al., *Drugs* **38** (1989) 226–248.
579. J. L. Malis, *J. Pharmacol. Exp. Ther.* **194** (1975) 488–498.
580. J. J. O'Brien, P. Benfield, *Drugs* **38** (1989) 226–248.
581. Wyeth, DE-OS 1 941 534, 1969; GB 1 285 025, 1969.
582. G. Bradley et al., *Eur. J. Med. Chem.* **15** (1980) 375.
583. B. Holmes, A. Ward, *Drugs* **30** (1985) 285–312.
584. G. W. Pasternak et al., *Postgrad. Med. J. 61 Suppl* **2** (1985) 5–12.
585. B. Holmes, A. Ward, *Drugs* **30** (1985) 285–312.
586. Endo Laboratories, GB 1 119 270, 1968; US 3 332 950, 1967.
587. L. Castaner, P. J. Roberts, *Drugs Future* **2** (1977) 613–615.
588. E. B. De Souza, et al., *J. Pharmacol. Exp. Ther.* **244** (1988) 391–402.
589. J. K. Errick, R. C. Heel, *Drugs* **26** (1983) 191–211.
590. W. K. Schmidt, *Drug Alcohol Depend.* **14** (1985) 339–362.
591. G. C. Pugh et al., *Br. J. Anaesth.* **62** (1989) 601–609.
592. Merck & Co., US 2 364 833, 1944; US 2 891 954, 1959.
593. M. U. Zubair et al., *Anal. Profiles Drug Subs.* **18** (1989) 195–219.
594. Sterling Drug, BE 611 000, 1961.
595. S. Archer et al., *J. Med. Chem.* **7** (1964) 123–127.
596. R. N. Brogden et al., *Drugs* **5** (1973) 6–91.
597. D. A. Reed, S. H. Schnoll, *JAMA* **256** (1986) 2562–2564.
598. H. Schmidthammer, *Prog. Med. Chem.* **35** (1998) 83–132.
599. G. Ehrhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung*, vol.**1**, Verlag Chemie, Weinheim 1972, pp. 131–132.
600. J. Hellerbach et al., *Helv. Chim. Acta* **39** (1956) 429–440.
601. O. Schnider, A. Grüssner, *Helv. Chim. Acta* **24** (1951) 2211–2217.
602. D. M. Paton, *Drugs Future* **4** (1979) 438–441.
603. W. Leimgruber et al., *Adv. Biochem. Psychopharmacol.* **8** (1973) 45–50.
604. F. F. Foldes et al., *Can. Anaesth. Soc. J.* **16** (1969) 151–161.
605. J. M. Evans et al., *Anaesthesia* **29** (1974) 721–727.
606. US 3 254 088, 1966; DE 1 183 508, 1962 (M. J. Lewenstein).
607. R. A. Olofson et al., *Tetrahedron Lett.* 1977 1567–1577.
608. R. Simantov, S. H. Snyder, *Biochem. Soc. Trans.* **5** (1977) 62–65.
609. K. A. Handal et al., *Ann. Emerg. Med.* **12** (1983) 438–445.
610. P. M. Goodrich, *AANA J.* **58** (1990) 14–16.
611. Endo Laboratories, GB 1 119 270, 1968; US 3 332 950, 1967.
612. V. Höllt, A. Herz, *Fed. Proc.* **37** (1978) 158–161.
613. B. L. Crabtree, *Clin. Pharm.* **3** (1984) 273–280.
614. J. P. Gonzalez, R. N. Brogden, *Drugs* **35** (1988) 192–213.
615. B. A. Judson et al., *Drug Alcohol Depend.* **7** (1981) 325–346.
616. Boehringer Ingelheim Pharma GmbH & Co. KG, US 4176186, 1979 (L. Goldberg, H. Merz, K. Stockhans).
617. C. S. Yuan, J. F. Foss, *Neuropharmacology* **38** (1999) 425–432.
618. C. S. Yuan et al., *J. Pharmacol. Exp. Ther.* **300** (2002) 118–123.

619. D. S. Wang et al., *J. Emerg. Med.* **16** (1998):471–475.
620. H. E. Jones et al., *Drug Alcohol Depend.* **60** (2000) 29–37.
621. L. C. Costantini et al., *Int. J. Pharm.* **283** (2004) 35–44.
622. J. E. Grant et al., *Am. J. Psychiatry* **163** (2006) 303–312.
623. K. Hempenstall, A. S. Rice, *Curr. Opin. Investig. Drugs* **3** (220) 441–448.
624. R. H. Dworkin et al., *Arch. Neurol.* **60** (2003) 1524–1534.
625. H. Ueda, *Pharmacol Therap.* **109** (2006) 57–77.
626. D. B. Gordon, G. Love, *Pain Manage. Nurs.* **5** (2004) 19–33.
627. G. T. Carter, M. D. Sullivan, *Curr. Opin. Investig. Drugs* **3** (2002) 454–458.
628. J. Zhou, *Drugs of the Future* **29** (2004) 1235–1244.
629. J. Lai et al., *Ann. Rev. Pharmacol. Toxicol.* **44** (2004) 371–397.
630. Ciba, US 3 471 548. 1969.
631. Ciba, US 3 978 216, 1972.
632. N. G. Bowery, *Curr. Opin. Pharmacol.* **6** (2006) 37–43.
633. Geigi, US 2 948 718, 1960 (W. Schindler).
634. DD 133 052, 1977 (R. Müller).
635. D. Soyka et al., *Schmerz* **12** (1998) 419–427.
636. Boehringer Ingelheim, DE 1 303 141, 1961; US 3 236 857, 1966.
637. VEB Arzneimittelwerke Dresden, DE-AS 1 770 874, 1968.
638. H. Schmitt, *Handb. Exp. Pharmacol.* **39** (1977) 299–396.
639. C. Lund et al., *Br. J. Anaesth.* **63** (1989) 516–519.
640. D. R. Jasinki et al., *Arch. Gen. Psychiatry* **42** (1985) 1063–1066.
641. J. Deeter et al., *Tetrahedron Lett.* **31** (1990) 7101.
642. L. A. Sorbera et al., *Drugs of the Future* **25** (2000) 907–916.
643. H. Liu et al., *Chirality* **12** (2000) 26–29.
644. S. Attila, E. Leinonen, *Curr. Opin. Investig.Drugs* **3** (2002) 1217–1221.
645. N. Pitsikas, *Curr. Opin. Investig. Drugs* **1** (2000) 116–121.
646. Degussa, DE 1 795 858, 1968; US 3 513 171, 1970.
647. *Drugs Future* **8** (1983) 773–775.
648. W. von Bebenburg et al., *Chem. Ztg.* **103** (1979) 387–399.
649. W. von Bebenburg et al., *Chem. Ztg.* **105** (1981) 217–219.
650. H. A. Friedel, A. Fitton, *Drugs* **45** (1993) 548–569.
651. Pfizer, US 4152326, (J. Hartenstein, G. Satziger).
652. Pfizer, US 4024175, (G. Satzinger, J. Hartenstein, M. Hermann, W. Heldt).
653. Zambon Group, WO 2005 044 779, 2005 (K. Arrighi, F. Corcella, G. Marchioro, A. Nicoli, M. Paiocchi, M. Villa).
654. Taro Pharmaceutical Industries Ltd., WO 2003 089403, 2003
655. R. Cagnoli et al., *Tetrahedron* **59** (2003) 9951.
656. G. J. Sills, *Curr. Opin. Pharmacol.* **6** (2006) 108–113.
657. J. K. Baille, *I. Power Curr. Opin. Investig. Drugs* **7** (2006) 33–39.
658. I. W. Tremont-Lukats et al., *Drugs* **60** (2000) 1029–1052.
659. J. Devulder et al., *Acta Neurol. Belg.* **102** (2002) 97–103.
660. Wellcome Foundation, EP 21 12, 1981 (M. G. Baxter, et al.).
661. Wellcome, US 4 560 687, 1985 (M. G. Baxter, A. R. Elphick, A. A. Miller, D. A. Sawyer).
662. Wellcome, US 4 602 017, 1986 (M. G. Baxter, A. R. Elphick, A. A. Miller, D. A. Sawyer).
663. E. Eisenberg et al., *Rev. Neurotherapeutics* **5** (2005) 729–735.
664. M. Guven, H. Bozdemir, J. Gunay, Y. Sarica, J. Kahraman, F. Koc, The actions of lamotrigine and levetiracetam on the conduction properties of isolated rat sciatic nerve, *Eur. J. Pharmacol.* **553** (2006), 129–134.
665. Rexall, US 3 487 153, 1969.
666. Riker, US 3 830 803, 1974.
667. R. C. Heel et al., *Drugs* **19** (1980) 249–267.
668. Park Davis, US 2 409 754, 1946 (H. R. Henze).
669. H. Biltz et al., *Ber. Dtsch. Chem. Ges.* **41** (1908) 1391.
670. Warner-Lambert WO 99/48876, 1999 (R. L. Bosch)
671. B. D. Maryanoff, B. L. Margul, *Drugs of the Future* **14** (1989) 342.
672. B. E. Maryanoff, et al., *Med. Chem.* **30** (1987) 880–887.
673. Ortho-McNeil Pharmaceuticals Inc., US 4513006, 1985 (B. E. Maryanoff, J. F. Gardocki).
674. Ortho-McNeil Pharmaceuticals Inc., WO 2004 078769, 2004 (J. E. Berkner, S. Duncan)
675. C. E. Heading, *IDrugs* **4** (2001) 339–350.
676. C. E. Heading, *Curr. Opin. Investig. Drugs* **3** (2002) 915–920.
677. S. Tries et al., *Inflamm. Res.* **51** (2002) 129–134.

678. J. J.Talley, .*J Med. Chem.* **43** (2000) 775–777.
679. US 2 806 033, 1955 (M. J. Lewenstein, U. Weiss).
680. T. Appelboom, P. Franchimont, *Adv. Therapy* **11** (1994) 228–234.
681. Troponwerke, US 3 692 818, 1972, DE 1 939 112, 1969.
682. Syntex, US 3 896 157, 1975.
683. Geigi, US 3 558 690, 1971.

Anesthetics, General

→ *Local Anesthetics is a separate keyword.*

Hartmund Wollweber, Bayer AG, Wuppertal, Federal Republic of Germany

1. Introduction 1
2. Inhalation Anesthetics 3
2.1. Nitrous Oxide 3
2.2. Hydrocarbons 3
2.3. Halogenated Hydrocarbons 3
2.4. Ethers 4
2.5. Halogenated Ethers 5
3. Intravenous Anesthetics 6
3.1. Barbituric Acids 6
3.2. Thiobarbituric Acids 7
3.3. Miscellaneous Compounds 8
4. Combination Anesthetics 9
4.1. Sleep-Inducing Anesthetics 9
4.2. Benzodiazepines 10
4.3. Steroid Derivatives 11
4.4. Neuroleptanalgesics 12
5. References 13

1. Introduction

General anesthetics are agents that temporarily produce a stupor- or sleeplike condition during which "painless" operation is possible. In this condition of general anesthesia (narcosis), consciousness, pain sensitivity, defense reflexes, and muscular reflexes are depressed to a large extent, whereas the vital centers of the medulla oblongata, particularly those controlling respiration and circulation, are relatively unaffected. In practice, the conditions depend entirely on the dosage of the anesthetic. Agents with too narrow a therapeutic range, i.e., too narrow a margin of safety between anesthetic dose and lethal dose, may not be employed.

The numerous theories regarding anesthesia either deal with the effectiveness or distribution of anesthetics within the body or speculate on their modes of action. Because inert gases, even the noble gases, are also effective anesthetics, the mode of action is believed to be a purely physical, reversible effect on the lipid-containing cellular membrane of the neurons. In principle, this affects all body cells. Only the higher sensitivity of the cells of the central nervous system (CNS) makes the desired gradual anesthetization possible with an appropriate dosage of the anesthetic agent.

Four different stages of anesthesia are distinguished:

1) The analgesic stage, permitting slightly painful procedures, such as changing dressings.
2) The excitation stage, which must be overcome quickly by premedication or additional anesthetic.
3) The tolerance stage, in which the centers of the cerebrum and spinal cord are inhibited, allowing surgical operations to be carried out.
4) The asphyxia stage, characterized by inhibition of the autonomic centers and the medulla oblongata (the vasomotor and respiratory centers), extreme muscle relaxation, and cessation of thoracic respiration. In practice, this stage must be counteracted immediately by appropriate measures, such as artificial respiration and/or infusion therapy; indeed, this stage should never be reached during anesthesia.

The tolerance stage can be subdivided into

a) Paralysis of the spinal cord and mesencephalon, slight relaxation of the muscles
b) Drop in blood pressure and body temperature, relaxation of the skeletal musculature
c) Relaxation of the abdominal musculature, onset of respiratory depression
d) Relaxation of the nonstriated musculature, lack of reflexes, difficult and deep breathing, abdominal breathing only, further drop

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_289

in blood pressure and body temperature, onset of hypoxemia, cyanosis

During emergence from the tolerance stage, the patient passes back through the excitation and analgesia stages.

Effective anesthesia is required to bring about

1) Unconsciousness
2) Complete analgesia and absence of pain reflexes
3) Relaxation of the muscles

The patient must receive sufficient anesthetic to fulfill the third requirement; in practice, this usually means overdosage regarding the first two requirements. Consequently, a combined technique is currently in use: other pharmaceutical agents, generally sedatives, neuroleptics, analgesics, and muscle relaxants, are used to support the anesthesia, particularly in the initial phase.

Depending on the duration of their action, anesthetic agents are termed long acting, short acting, or very short acting. Depending on how they are administered, they also can be divided into inhalation, intravenous, and combination anesthetics.

Inhalation (Pulmonary) Anesthetics. A large number of gases or vapors that are inhaled in a mixture with air or oxygen produce unconsciousness and analgesia. For practical purposes, however, only a few have proved efficacious.

Inhalation anesthetics are characterized by great stability. They are taken up quickly in the lungs and to a large extent are eliminated rapidly and mostly unchanged through the lungs (only minimally in the urine). They can be withdrawn immediately should an untoward incident occur; i.e., they can be controlled. Their main disadvantages are that they often cause a pronounced excitation stage and that many patients experience discomfort (postnarcotic nausea, vomiting). With inflammable gases the danger of explosion is always present.

Gas and vapor anesthetics exhibit different distributions between the lipophilic and hydrophilic phases and so have different solubilities in blood and different affinities to the lipid-rich nervous tissue of the CNS. The gases are less soluble in blood and require a higher partial pressure to maintain the tolerance stage. The steep concentration gradient makes it possible to induce and withdraw anesthesia quickly, thus providing better control of the narcosis.

The "minimum alveolar concentration" (MAC), a concept introduced in 1965, is a measure of the potency of inhalation anesthetics. It is expressed as the concentration in the alveolar air that prevents 50 % of patients from responding to pain stimuli. The MAC levels are changed by premedication agents or adjuncts, and anesthetic dosage always must be adjusted accordingly. However, the concentration of an anesthetic agent cannot be increased arbitrarily because of the narrow therapeutic range, which for most agents is ≈ 1.5.

Intravenous (Parapulmonary) Anesthetics. Metabolization and distribution gradually render intravenous agents ineffective; i.e., they are subject to processes over which the anesthesiologist has no influence. Further, these agents cannot be controlled once they have been administered. However, they are psychologically well tolerated by patients, and they are quite useful for brief narcosis and as induction agents for inhalation anesthesia, providing the basic anesthesia. Only intravenous anesthetics can be used when the airways must be maintained completely patent.

Combination Anesthesia. Inhalation or intravenous anesthetics, used singly, depend on dosage to achieve the conditions necessary for surgery: autonomic depression, sleep, analgesia, and muscle relaxation. Nowadays, preference is given to combinations of medications, which broaden the therapeutic range. Moreover, with several substances administered at minimal dosages the particular goals of anesthesia can be achieved.

Sleep-inducing anesthetics. To induce narcosis and regulate sleep, the short-acting barbiturates (the induction agents) – etomidate, steroid anesthetics, and some benzodiazepine derivatives – are used.

Neuroleptanalgesics. A stage that approximates that reached during the first stage of anesthesia can be attained by simultaneous administration of a potent analgesic and a neuroleptic agent. In this state of so-called neurolep-

tanalgesia, there are complete analgesia, psychic and motoric depression, and amnesia. The autonomic nervous system is depressed, whereas all important reflexes are retained. In general, the patient remains conscious, making this type of narcosis particularly suitable for operations in which active participation on the part of the patient is necessary, for example, brain, ear, and laryngeal operations. If higher doses of analgesics are given, artificial respiration is necessary. Commonly, the patient is premedicated with muscle relaxants, such as succinylcholine, with atropine, and even with anesthetics, such as propanidid or N_2O–O_2.

Neuroleptanalgesia is used for gentle narcosis, for lengthy operations, or for patients at risk in senium or with cardiovascular diseases.

Relaxants. Muscular relaxation is necessary for anesthesia and the operation. Some anesthetics produce muscle relaxation by affecting the central regulation of the tonus. However, peripheral muscle relaxation is achieved only by inhibiting the neurotransmitter acetylcholine with depolarizing substances or substances that block the acetylcholine receptor in the motor end plate. Such relaxants (→ Skeletal Muscle Relaxants) include tubocurarine, alcuronium, gallamine, pancuronium, decamethonium or suxamethonium, neostigmine, and pyridostigmine. Premedication with the anticholinergics atropine and scopolamine is common practice. Administered intramuscularly, these agents inhibit salivary and gastric secretions (prophylaxis against aspiration), prevent bradycardia, and provide protection against acute vagal reflexes.

2. Inhalation Anesthetics

2.1. Nitrous Oxide

Nitrous oxide [*10024-97-2*], nitrogen monoxide or laughing gas, N_2O, M_r 44.02, is a colorless gas with a slightly sweet odor and taste, *mp* –90.81 °C, *bp* −88.46 °C. At a pressure of 1 bar (100 kPa) one volume of water dissolves one volume of N_2O at 5 °C and 0.596 volumes at 25 °C. Nitrous oxide also is soluble in alcohol and ether. For manufacture → Nitric Acid, Nitrous Acid, and Nitrogen Oxides.

Nitrous oxide exerts a shallow anesthetic effect, MAC > 80. Although deep anesthesia cannot be achieved with it alone, it is used mixed with oxygen and ether or halothane as a basic anesthetic for all types of surgery because of its minimal toxicity. Further, N_2O does not irritate the airways. Nitrous oxide is always mixed with oxygen (20 %, in dentistry 20 – 50 %) to avoid anoxia. The use of nitrous oxide considerably reduces the requirements for other anesthetics and analgesics.

Trade Names. Stickoxydul (Hoechst) and others.

2.2. Hydrocarbons

Hydrocarbons have an anesthetic effect. However, despite their good anesthetic qualities and low toxicities, they are now rarely used because of the high risk of explosion. After serious explosions during anesthesia, ethylene and acetylene are no longer employed in Germany.

Cyclopropane [*75-19-4*], trimethylene, USP XIX, C_3H_6, M_r 42.08, is a colorless, odorless, inflammable gas, *mp* -127 °C, *bp* −33 °C. At room temperature it liquefies at 4 – 6 bar (400 – 600 kPa). The explosive range is between 2.4 and 10.3 vol % in air and between 12.5 and 60.0 vol % in oxygen. The blood – gas partition coefficient at 37 °C is 0.46. Cyclopropane is mixed with oxygen (15 – 30 % cyclopropane) for medical purposes. In contrast to nitrous oxide, cyclopropane (MAC 9.2) at concentrations of 20 % can produce anesthesia. It can be controlled readily and has a broad therapeutic range. Cyclopropane increases the sensitivity of the heart to catecholamines.

Cyclopropane is manufactured commercially from 1,3-dichloropropane with zinc in the presence of sodium iodide.

Trade Names. Cyclopropane (ICI; Mallinckrodt; Squibb). Cyclopropane USP is available in sealed orange metal cylinders or in chrome-plated cylinders with orange labels.

2.3. Halogenated Hydrocarbons

Low-boiling halogenated hydrocarbons exhibit anesthetic activity. Compared to hydrocarbons, they are more potent but also more toxic. Anesthetics previously in common use – chloroform,

carbon tetrachloride, and ethyl chloride – are now never or hardly ever employed because of their hepatotoxic and carcinogenic side effects. In addition, tribromoethyl alcohol is no longer employed in the major industrial nations. The chlorinated hydrocarbons increase the sensitivity of the heart to catecholamines.

Trichloroethylene [*79-01-6*], CCl_2=CHCl, M_r 131.4, is a colorless, noninflammable liquid with a characteristic chloroform-like odor, *bp* 87.2 °C. (For further information → Chlorinated Hydrocarbons.) Trichloroethylene is highly hepatotoxic and nephrotoxic and causes hepatocellular carcinomas in mice. In the presence of alkali, trichloroethylene decomposes to highly toxic dichloroacetylene and HCl. For anesthetic purposes, the substance is stabilized with thymol and mixed with a blue coloring agent.

Indication. For inhalation analgesia in obstetrics and for minor surgical procedures, but only with a special inhaler. Trichloroethylene (MAC 0.17) should not be used in a closed-circuit apparatus. Trichloroethylene is better tolerated than chloroform.

Trade Names. Anamenth (Brunnengräber) contains menthol as a stabilizer; Chlorylen (Schering); Ecrylène (Robert & Carrière, Paris); Narkosid (Heyl); Trethylene (Davies-Rose-Hoyt); Trichloran (Merck); Trilene (ICI; Ayerst).

Halothane [*151-67-7*], 2-bromo-2-chloro1,1,1-trifluoroethane, $CF_3CHBrCl$, M_r 197.39, is a colorless, noninflammable liquid with a characteristic sweet odor, *bp* 50.2 °C. Halothane is soluble in water to the extent of 0.345 wt %, miscible with ether, and sensitive to light. Anesthetic halothane contains 0.01 % thymol as a stabilizer.

Halothane is manufactured by chlorination of CF_3CH_2Br or bromination of CF_3CH_2Cl in the gas phase at 250 – 475 °C [2] or by the rearrangement [3]

$$CBrF_2CHClF \xrightarrow[50\,^\circ C]{AlCl_3} CF_3CHBrCl$$

Indication. Because of its potent and rapid effect, halothane (MAC 0.77) is used in special vaporizers mixed with oxygen and water. The high lipid – blood partition coefficient leads to considerable accumulation in tissue and fat. Halothane reduces the peripheral vascular resistance, increases the possibility of arrythmias, and has a negative effect on liver function. On repeated use it is hepatotoxic.

Trade Names. Fluothane (ICI Pharma); Halothan "Hoechst" (Hoechst); Halan (Arzneimittelwerk Dresden); Halovis (Vister, Italy); Narcotan (Spofa, Prague); Rhodialothan (Rhodia Pharma).

Teflurane [*124-72-1*], 2-bromo-1,1,1,2-tetrafluoroethane, CF_3CHFBr, M_r 180.95, a nonexplosive, noninflammable gas, *bp* 8 °C. Its anesthetic qualities are similar to those of halothane [4]. Teflurane is prepared by bromination of 1,1,1,2-tetrafluoroethane at temperatures above 400 °C [5].

Trade Name. Teflurane (Dow; Abbott)

2.4. Ethers

The ethers have a broad anesthetic range; i.e., the difference between the dose needed to produce the tolerance stage and the lethal dose is large. They have enjoyed a certain popularity as inhalation anesthetics. The best known, diethyl ether, usually is combined with other agents. There is always the danger of explosion in mixtures with oxygen or air. The solubility of ether in blood and tissue results in a 15-min induction period, which is accompanied by irritation of the mucous membranes and airways and by salivation. The recovery phase is long, and ether is no longer in routine use by anesthesiologists, despite its favorable effects on the bronchial musculature and cardiocirculatory system and its relatively good relaxant effect on the muscles.

Diethyl ether [*60-29-7*], H_5C_2-O-C_2H_5, M_r 74.12, is a mobile, colorless, inflammable liquid with characteristic odor, *bp* 34.6 °C. Mixtures of diethyl ether and air are explosive. Diethyl ether is miscible with short-chain alcohols, chloroform, and petroleum ether. An aqueous solution saturated with ether contains 8.43 wt % ether at 15 °C and 6.05 wt % at 25 °C. Ether saturated with water contains 1.2 wt % water at 20 °C (MAC 1.9). For details on preparation and purification → Ethers, Aliphatic.

Trade Names. Aether puriss, pro narcosi "Hoechst" (Hoechst); Aether puriss, pro narcosi "Bonz" (Bonz).

Divinyl ether [*109-93-3*], vinyl ether, CH_2=CH-O-CH=CH_2, M_r 70.09, is a mobile, colorless, inflammable liquid with a characteristic sweet odor, *bp* 28.4 °C. Approximately 4 % ethanol is added as a stabilizer. Divinyl ether may contain 0.0077 % of a stabilizer, phenyl-α-naphthylamine, that produces a slightly reddish fluorescence.

Divinyl ether is prepared by the reaction of liquefied alkali hydroxide or sodium alkoxide on β,β'-dihalodiethyl ether [6] or by pyrolysis of α,α'-dichlorodiethyl ether in the gas phase at 600 – 800 °C in the presence of barium chloride catalyst [7].

Indication. Divinyl ether is used for minor surgery, obstetrics, dentistry, and inducing anesthesia. A certain degree of hepatotoxicity rules out long-term use. Hypersecretion has also been observed. Divinyl ether increases and decreases in tissue levels faster than diethyl ether.

Trade Names. Vinydan (Lundbeck) contains *N*-phenylnaphthylamine as a stabilizer and absolute alcohol; Venesthene (May & Baker; Merck Sharp & Dohme); Vinether (Robert & Carrière, Paris); Ethydan (Lundbeck) is a 1 : 3 mixture of divinyl ether and diethyl ether.

2.5. Halogenated Ethers

The polyfluorinated ethers have analgesic and muscle-relaxing properties but are more difficult to control. In addition, some are readily inflammable. They fall between the ethers and the halogenated hydrocarbons in their effect on the heart. The importance of the fluorinated ethers varies from country to country. In 1985, enflurane was the most important. Some fluorinated ethers – such as sevoflurane, 1,1,1,3,3,3-hexafluoro-2-(fluoromethoxy)-propane [8] (Baxter); aliflurane, 2-chloro-1-methoxy-1,2,3,3-tetrafluorocyclopropane [9] (W. R. Grace); dioxychlorane, 4,5-dichloro-2,2-difluoro-1,3-dioxolane [10] (Ayerst); and thiomethoxyflurane, 2,2-dichloro-1,1-difluoro-1-methylmercaptoethane [11] (Ohio Medical Products) – are still being tested clinically.

Enflurane [*13838-16-9*], 2-chloro-1-(difluoromethoxy)-1,1,2-trifluoroethane, $CHClFCF_2$-O-CHF_2, M_r 184.5, *bp* 56.5 °C, is a colorless, noninflammable liquid with a pleasant odor. Enflurane (MAC 1.68) is used in special vaporizers and, depending on the stage of anesthesia, is mixed with various amounts of oxygen and nitrous oxide. Its advantage over halothane is the more rapid induction and withdrawal. Arrythmias are observed less frequently, which makes the combination with catecholamines less a problem.

Enflurane is prepared by dichlorination of the methyl group of 1-chloro-1,2,2-trifluoro-2-methoxyethane and replacement of these two chlorine atoms by fluorine [12], [13]:

$$CH_3-O-CF_2CHClF \xrightarrow[h\nu]{Cl_2} CHCl_2-O-CF_2CHClF$$
$$\xrightarrow[SbCl_5]{H_2F_2} CHF_2-O-CF_2CHClF + 2\,HCl$$

Trade Names. Ethrane (Deutsche Abbott); Ethrane (Ohio Medical Products).

Methoxyflurane [*76-38-0*], 2,2-dichloro-1,1- difluoro-1-methoxyethane, $CHCl_2CF_2$-O-CH_3, M_r 164.98, *bp* 104.65 °C, is a colorless liquid with a fruity odor. It is produced industrially by the addition of methanol to 1,1-dichloro-2,2-difluoroethylene in the presence of sodium methoxide [14].

Methoxyflurane is used with special vaporizers and, depending on the depth of anesthesia desired, is mixed with various amounts of oxygen and nitrous oxide. Its high lipid solubility gives it the lowest MAC, 0.2, found among inhalation anesthetics in use. It does not sensitize the heart to catecholamines; otherwise, the majority of its side effects are those of halothane (see page 4). Oxalate deposits and nephrotoxic fluoride concentrations are undesirable metabolic products of methoxyflurane.

Methoxyflurane may not be administered to patients with severe liver damage or renal insufficiency or those concomitantly receiving aminoglycoside antibiotics and tetracyclines.

Trade Names. Penthrane (Abbott) contains 0.01 % butylated hydroxytoluene as an antoxidant; Analgizer (Abbott); Methofane (Pitman-Moore).

Isoflurane [*26675-46-7*], 1-chloro-1-(difluoro- methoxy)−2,2,2-trifluoroethane, $CF_3CHCl\text{-}O\text{-}CHF_2$, M_r 184.5, is a colorless, noninflammable liquid with a faint odor, *bp* 48.5 °C. Isoflurane is prepared by chlorination of 2,2,2-trifluoroethoxydifluoromethane, itself obtained by alkylation of trifluoroethanol with difluorochloromethane [12], [15]:

$$CF_3CH_2OH + HCClF_2 \xrightarrow[80-95\,^\circ C]{KOH} CF_3CH_2-O-CHF_2 \xrightarrow{Cl_2} CF_3CHCl-O-CHF_2$$

Isoflurane (MAC 1.3) has a low blood – gas distribution coefficient, which leads to rapid recovery from the anesthesia. Myocardial functions remain intact, and as the depth of anesthesia increases, vasodilation progressively develops [16].
Trade Name. Forane (Ohio Medical Products).

Fluroxene [*406-90-6*], 2,2,2-trifluoro-1-vinyloxyethane, $CF_3CH_2\text{-}O\text{-}CH{=}CH_2$, M_r 126.8, *bp* 42.5 °C, is a colorless liquid with a pungent odor. It is inflammable and, at concentrations above 4 %, explosive in air. Fluroxene (MAC 3.4) generally is used in a mixture with oxygen and nitrous oxide. Several cases of liver damage have been observed.

The compound is prepared by catalytic vinylation of 2,2,2-trifluoroethanol in the presence of alkali-metal trifluoroethoxide [17]:

$$CF_3CH_2OH + CH{\equiv}CH \longrightarrow CF_3CH_2\text{-}O\text{-}CH{=}CH_2$$

Another synthesis is based on acetaldehyde-bis(trifluoroethylacetal), from which trifluoroethanol is split off in the presence of montmorillonite catalysts [18]:

$$(CF_3CH_2O)_2CHCH_3 \longrightarrow CF_3CH_2\text{-}O\text{-}CH{=}CH_2 + CF_3CH_2OH$$

Trade Name. Fluoromar (Ohio Medical Products).

3. Intravenous Anesthetics

Intravenous anesthetics, like inhalation anesthetics, can be used alone. Most have side effects. The majority cause respiratory depression. Administration of oxygen is recommended, particularly with barbiturates and thiobarbiturates. Intraarterial injection must be avoided because of the danger of tissue necrosis.

3.1. Barbituric Acids

The 5,5-disubstituted barbiturates have attained an extremely broad range of use as anesthetics and also are employed in other therapeutic areas (→ Hypnotics,→ Sedatives, and→ Antiepileptics).

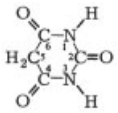

Barbituric acid

The effective anesthetics are methylated at position 1, and usually one of the substituents at position 5 is unsaturated. The barbiturates are so-called haloencephalic anesthetics; i.e., the reversible inhibition of activity of all excitable tissue is particularly marked in the central nervous system. The redistribution of the lipophilic barbiturates from the well-perfused cerebral tissue into the peripheral storage depots of muscle and fatty tissue is more important for the cessation of anesthesia than is metabolism. Balanced distribution between plasma and fatty tissue is attained only after a relatively long period, e.g., for the thiobarbiturates, only after 1.5 – 2 h. Repeated administration of barbiturates carries with it the risk of accumulation.

Only a few of the many barbiturates have pharmacokinetic properties that make them suitable for use. The barbiturates are poor analgesics and muscle relaxants. The tolerance stage of anesthesia is reached quickly and can be maintained with an inhalation anesthetic. The barbiturates are employed generally for minor surgery and as induction agents for anesthesia. Phenobarbital is used only for premedication.

The following classification is based on a British evaluation of barbiturate and thiobarbiturate anesthetics [19]; in other countries, assessment is not necessarily identical:

Good, equally effective: thiopental, thiamylal, thiobutabarbital, thialbarbital

Insufficient, frequent side effects: hexobarbital, methitural, buthalital
Some advantages, some side effects: methohexital, narcobarbital

Methohexital sodium [*309-36-4*], sodium 5-allyl-1-methyl-5-(1-methyl-2-pentynyl)-barbiturate, M_r 284.30. The free acid [*18652-93-2*] forms colorless crystals, *mp* 60 – 64 °C. It is prepared by condensing diethyl allyl(1-methyl-2-pentynyl)malonate with *N*-methylurea [20].

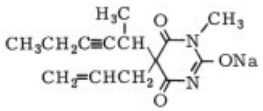

Methohexital sodium is a short-acting anesthetic administered intravenously. It is very potent and has the shortest duration of action among the barbiturates used. A disturbing factor is that rather frequently it causes hiccups directly after injection. Patients are reported to recover rapidly from anesthesia. Abuse can lead to addiction.
Trade Names. Brevane (Elanco, USA); Brevital Sodium (Eli Lilly).

Hexobarbital [*56-29-1*], 5-(cyclohexen-1-yl)-1,5-dimethylbarbituric acid, M_r 236.26. The acid forms colorless crystals that turn pink on exposure to light, *mp* 146 °C. A 10 % aqueous solution of the sodium salt [*50-09-9*], M_r 258.25, has a pH of 11.5.

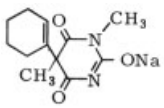

Hexobarbital sodium is prepared by condensation of methyl(cyclohexen-1-yl)cyanoacetic ester with dicyanodiamide to 5-(cyclohexen-1-yl)-5-methyl-2,6-diimino-3-cyanobarbituric acid, followed by methylation with dimethyl sulfate and hydrolysis with 25 % sulfuric acid [21].

Hexobarbital sodium is an intravenous anesthetic for short procedures and combination anesthesia. Hexobarbital is tolerated by tissue much better than thiopental and has proved efficacious for almost 50 years, at least for induction of anesthesia. Abuse can lead to addiction.
Trade Names (ampule with dry substance): Evipan-Natrium (Bayer); Evipal Sodium (Winthrop-Stearns); Cyclonal Sodium (May & Baker); Dorico Solubile (Winthrop); Hexanastab (Boots); Toleran inj. (Kwizda).

3.2. Thiobarbituric Acids

The anesthetic effect of the thiobarbituric acids differs from that of the barbiturates in that they produce a somewhat longer period of sleep after withdrawal of medication. Emergence is calmer and quicker. The duration of anesthesia can be lengthened considerably in patients having only limited fat deposits outside of the central nervous system (i.e., asthenic persons) or having fat deposits saturated by several injections of thiobarbiturates. The thiobarbiturates depress respiration more than the barbiturates and are less well tolerated by the veins.

The thiobarbituric acids are more acidic than the barbituric acids. They dissolve in aqueous alkali, but the heterocyclic ring opens after a time. They are prepared by condensation of substituted malonic esters, malononitrile, malononitrile esters, or malonamide esters with thioureas or *N*-alkylthioureas, whereby the imino derivatives initially produced are hydrolyzed to thiobarbituric acids [22].

Thiopental sodium [*71-73-8*], the sodium salt of 5-ethyl-5-(1-methylbutyl)thiobarbituric acid [*76-75-5*], M_r 264.35, a yellowish-white, crystalline, slightly bitter powder, is very soluble in water and soluble in alcohol. The compound is hygroscopic and has a weak unpleasant odor. The free acid forms colorless crystals, *mp* 158 – 159 °C. A 10 % solution of the sodium salt has a pH of 10.6.

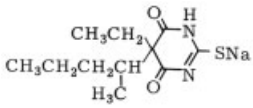

It is prepared by condensation of thiourea with a disubstituted malonic ester [23].

Thiopental sodium is a short-acting intravenous anesthetic used for brief surgical procedures and induction of anesthesia. It is a potent,

rapidly effective hypnotic. Sleep ensues within 30 – 60 s. A disadvantage is the lengthy hangover. In the United States and Great Britain it is the most widely used injection anesthetic. Abuse can lead to addiction.

Thiopental sodium is used in a mixture with approximately 6 % sodium carbonate.

Trade Names. Farmotal (Farmitalia); Hypnostan (Leiras, Turku); Intraval (May & Baker); Leopental (Leo); Nesdomal (Specia); Pentothal (Abbott); Pentothal-Natrium (Deutsche Abbott); Thio-Barbityral (Amino, Neuenhof, Switzerland); Thionembutal (Abbott); Thiobarbital (Miro, Palma de Mallorca); Thiopental "Lentia" (Hormonchemie, Munich); Trapanal (Promonta; Byk Gulden).

Thiobutabarbital sodium [*947-08-0*], the sodium salt of 5-*sec*-butyl-5-ethylthiobarbituric acid, M_r 241.32, is a yellowish-white amorphous powder soluble in water and alcohol. The colorless crystals of the free acid [*2095-57-0*] melt at 163 – 165 °C. Thiobutabarbital is prepared from thiourea and a malonic ester [24].

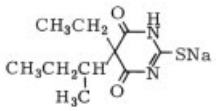

It is a short-acting intravenous anesthetic. It is the least potent of the thiobarbiturates and therefore causes the least number of complications during anesthesia. The compound is available as 5 % and 10 % solutions.

Trade Names. Inaktin (Byk Gulden); Narkothion (Sanabo); Brevinarcon (Arzneimittelwerk Dresden).

Thiamylal [*77-27-0*], 5-allyl-5-(1-methylbutyl)-2-thiobarbituric acid, M_r 254.36, *mp* 132 – 133 °C. The sodium salt [*337-47-3*], M_r 276.36, is an amorphous powder soluble in water. It is prepared by condensation of thiourea with a malonic ester [25], [26].

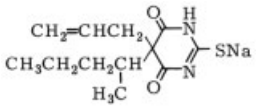

The properties of this rapidly effective anesthetic are similar to those of thiopental.

Trade Names. Surital (Parke-Davis); Thiosecional (Eli Lilly); Bio-Tal (Philips Roxane, veterinary anesthetic).

Buthalital sodium [*510-90-7*], the sodium salt of 5-allyl-5-isobutyl-2-thiobarbituric acid, M_r 262.31, is a water-soluble amorphous powder. The free acid [*468-65-5*], M_r 240.33, melts at 147 °C. Buthalital sodium is prepared by condensation of thiourea with a malonic ester [27], [28].

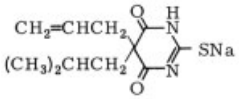

Buthalital sodium is eliminated so rapidly that slow injection does not produce anesthesia [29]. For pharmacological properties, see [30].

Trade Names. Baytinal (Bayer); Transithal (May & Baker); Ulbreval (Wyeth).

Thialbarbital [*467-36-7*], 5-allyl-5-(cyclohexen-2-yl)-2-thiobarbituric acid, M_r 264.36, *mp* 79 – 81 °C. The sodium salt [*3546-29-0*], M_r 286.34, is a pale yellowish, water-soluble, amorphous powder.

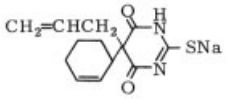

Thialbarbital is prepared from thiourea and a malonic ester [31].

Trade Name. Kemithal (ICI; Avlon).

3.3. Miscellaneous Compounds

A miscellaneous group mainly includes compounds that are readily soluble in lipids and that must be brought into aqueous solution with solubilizers. They are generally short- and very short-acting anesthetics used for surgical intervention and induction of inhalation anesthesia.

Propanidid [*1421-14-3*], propyl 3-methoxy-4-(*N,N*-diethylcarbamoylmethoxy)-phenylacetate, M_r 337.42, is a colorless to pale yellowish water-insoluble oil. It is prepared from the propyl ester of sodium homovanillic acid and *N,N*-diethylchloroacetamide [32].

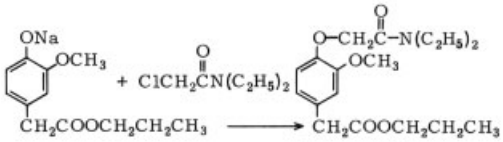

Propanidid

Propanidid is used as a very short- to short-acting intravenous anesthetic for minor surgery and gynecology. It is characterized by rapid onset of and quick recovery from anesthesia, producing surgical anesthesia for 3 – 4 min. Repeated injections can be used to prolong the anesthesia. Postanesthetic sleep does not occur because the propyl ester is hydrolyzed rapidly in the patient to form inactive 3-methoxy-4-(*N,N*-diethylcarbamoylmethoxy)phenylacetic acid. After 20 – 30 min the patients generally are completely responsive [33].

The commercial 5 % aqueous solution contains 20 % of the solubilizer Cremophor EL (polyethoxylated castor oil) [34]. Histamine release is sometimes observed, in part because of the Cremophor EL.

Trade Names. Epontol (Bayer; Theraplix); Propanidid (Riker); Fabontal (Bayer); Propantan (Leiras, Turku); Sombevrin (Gedeon Richter).

Ketamine [*6740-88-1*], 2-(*o*-chlorophenyl)-2-methylaminocyclohexan-1-one, M_r 237.74. The hydrochloride [*1867-66-9*] forms colorless crystals, *mp* 262 – 263 °C. Its solubility in water at 20 °C is 20 g per 100 mL.

It is prepared by bromination of *o*-chlorophenyl cyclopentyl ketone, which is then reacted with methylamine to give the methylimino alcohol. Thermolysis of the imino hydrochloride yields ketamine by ring expansion [35]:

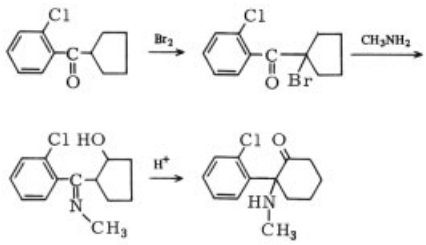

Ketamine

Ketamine is a short-acting anesthetic effective for 5 – 30 min, depending on the amount injected. It is suitable for diagnostic purposes and for surgical procedures that do not require muscle relaxation. The occasional hallucinations that occur during anesthesia suggest a chemical relationship to phencyclidine, which was used as a short-acting anesthetic until identified as a dangerous drug of abuse.

The mode of action of ketamine differs from that of the barbiturates. It suppresses the activities of the cerebral cortex (consciousness) and the thalamic pain pathways (analgesia). Parts of the upper brainstem and the limbic system are not affected (so-called dissociative analgesia). The patient exhibits a characteristic superficial sleep with complete elimination of pain. Because of the psychomotor side effects, combination with neuroleptics and tranquilizers is necessary. Ketamine is used especially in pediatrics. The usual preparations are 0.1 % and 0.5 % solutions.

Trade Names. Ketanest (Parke-Davis); Ketaject (Bristol); Ketalar (Parke-Davis; Sankyo); Ketaset (Bristol); Vetalar (Parke-Davis, veterinary anesthetic).

4. Combination Anesthetics

Combination anesthetics are used only in combinations to produce balanced anesthesia (see page 2). Chemically and pharmacologically they belong to many different groups.

4.1. Sleep-Inducing Anesthetics

Etomidate [*33125-97-2*], ethyl (*R*)-(+)-1(α-methylbenzyl)-imidazole-5-carboxylate, M_r 224.28, is a white to yellowish crystalline or amorphous powder, *mp* ca. 65 – 70 °C. It is insoluble in water at pH 7, but soluble in acidic aqueous solutions at pH < 3. Etomidate is soluble in propylene glycol and readily soluble in alcohol.

The preparation of etomidate involves a modification of Jones' synthesis: α-phenylethylamine is alkylated with ethyl chloroacetate to the *N*-(α-phenylethyl)-glycine ester, whose NH group is formylated. This product is condensed with methyl formate and cyclized

with HCl-KSCN to yield ethyl 2-mercapto-1-(α-methylbenzyl)-5-imidazolecarboxylate, which is then oxidatively desulfurized [36]. The (*R*) isomer is obtained by separating the racemic acid with (*R*)-(+)-1-phenylethylamine.

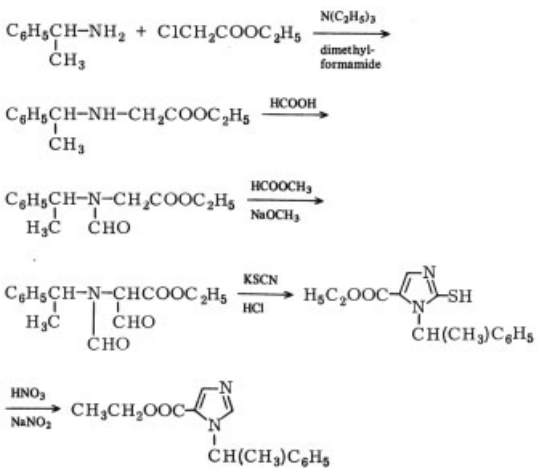

Etomidate

Etomidate is administered intravenously as a short-acting anesthetic for the induction of lengthy anesthesia, especially for the induction of neuroleptanalgesia (Section 4.4) and inhalation anesthesia [37], [38]. Etomidate produces 3 – 5 min of sleep.

Trade Names. Hypnomidate (Janssen; Johnson & Johnson; Sanwakagaku Kankyusho; Abbott) as a 0.2 % solution; Radenarcon (Arzneimittelwerk Dresden).

Sodium oxybate [*502-85-2*], sodium γ-hydroxybutyrate, $HOCH_2CH_2CH_2COONa$, M_r 126.09, is recrystallized from alcohol and is soluble in water. It is prepared by saponification of γ-butyrolactone [39].

The substance is related to γ-aminobutyric acid, a cerebral metabolite that is supposed to be sleep inducing. When given i.p., it produces a shallow anesthesia (tolerance stage a: paralysis of the spinal cord and mesencephalon, slight relaxation of the muscles, see Chap. 1) that resembles natural sleep as seen via the electroencephalograph. It is employed as an anesthetic during labor and in combination anesthesia with nitrous oxide, barbiturates, or neuroleptanalgesics.

The sodium salt is available as a 24 % aqueous solution.

Trade Names. Somsanit (Kühler Chemie); Gamma-OH (Ebifarm); Anetamin (Sankyo); Gioron (Kaken-Ono, Japan).

Disoprofol [*2078-54-8*], 2,6-diisopropylphenol, *bp* 136 °C at 40 mbar, *mp* 19 °C, belongs to a new group of anesthetics. Its effects are similar to those of Althesin (see page 11) and thiopental (see page 7). However, it is more effective and may be given repeatedly without lengthening the reapplication interval. It has been in phase II of clinical trials since 1982 [40].

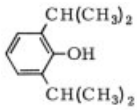

4.2. Benzodiazepines

Among the benzodiazepines, diazepam and flunitrazepam were used because of their sedative, anxiolytic, muscle relaxant, and anesthetic potentiating effects; initially, they were administered perorally but later also by injection. The frequency of pain at the injection site and of postoperative phlebothromboses led to the development of the water-soluble midazolam.

Midazolam maleate [*59467-70-8*], 8-chloro-6-(2-fluorophenyl)–1-methyl-4*H*-imidazo[1,5-a]-[1,4]benzodiazepine maleate [*59467-94-6*], M_r 441.59, *mp* 148 – 151 °C, is a stable, water-soluble powder. The solubility in water depends on pH: ≈ 85 mg/mL at pH 2.7 and 0.3 mg/mL at pH 7.6. The free base, M_r 325.78, melts at 152 – 154 °C. The maleate is subject to reversible ring opening. Below pH 4 the ring is open; above pH 4 the cyclic form is present. The anesthetic formulation is a buffered aqueous solution containing 2.5 mg/mL at pH 3.5.

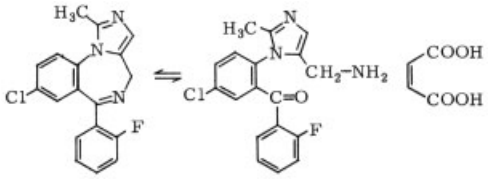

The preparation starts with 7-chloro-5-(2-fluorophenyl)−1,3-dihydro-2*H*-1,4-benzodiaze-pinone [41]. For the literature, also see [42].

Midazolam is approximately twice as active as diazepam, causes less pain at the injection site, and has a shorter half-life than diazepam. Side effects: dose-dependent cerebral depression with tranquilization, sedation, and dryness. Reduction in blood pressure, respiratory depression, and cardiovascular effects were slight.

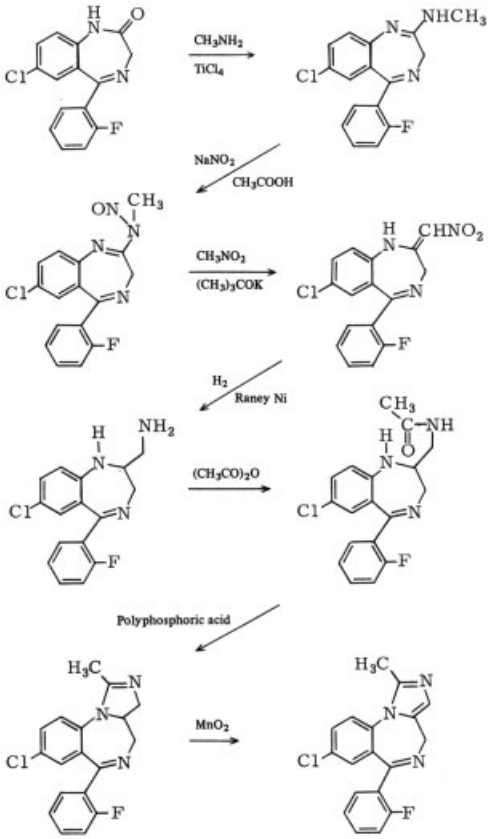

Trade Names. Dormicum (Hoffmann-La Roche, Switzerland); Hypnovel (Hoffmann-La Roche, United Kingdom).

4.3. Steroid Derivatives

A number of pregnane derivatives with oxygen substituents on C-3, C-20, and C-21 exhibit anesthetic properties without hormonal activity [43]. The preparations are used for brief narcosis and for induction of lengthy anesthesia. The first commercially available as an anesthetic was 21-hydroxypregnane-3,20-dione, as the water-soluble sodium hemisuccinate, under the generic name *hydroxidione* and by the trade names Presuren (Schering) and Viadril (Boehringer Mannheim; Pfizer).

The compound has been withdrawn in Great Britain and other countries, for, even when strongly diluted, it irritates the venous wall, increasing the risk of thrombosis, and elicits a long recovery phase. It is characterized by a late onset of the anesthesia after a relatively long, excitation-free induction period of 5 min. The induction period depends on the hydrolysis of the hemisuccinate ester to the active steroid alcohol. *Viadril G,* in which the hemisuccinate is combined with glycine, is supposed to be better tolerated. Most of the disadvantages can be avoided by dissolving the active pregnane derivative in Cremophor EL (BASF) [45].

A marked oversensitivity of patients to Althesin [45] led to the development of the water-soluble steroid *minaxolone* [44] (Glaxo-Allenbury), which is still being tested clinically.

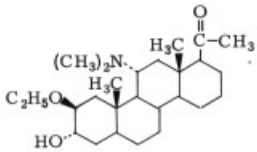

Minaxolone

Alphaxalone [*23930-19-0*], 3α-hydroxy-5αpregnane-11,20-dione, M_r 332.49, forms colorless prisms, *mp* 172 – 174 °C, $[\alpha]_D^{26}$ +113.4 (c = 1.2 mg/mL in $CHCl_3$). It is prepared by reduction of 3,11,20-trioxo-5-α-pregnane with trimethyl phosphite in the presence of an iridium catalyst. Only this reducing agent forms the axial alcohol (3α) [46], [47].

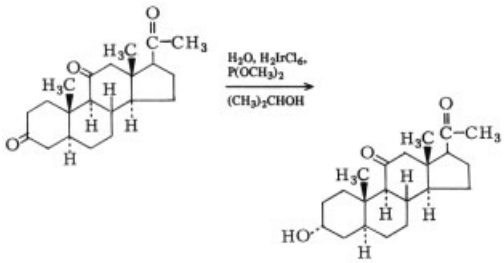

Alphaxalone

Alphaxalone is the more potent component of *Althesin*. Addition of alphadolone increases the total amount of active steroidal anesthetic in the Cremophor solution.

Alphadolone [*14107-37-0*], 3α,21-dihydroxy-5α-pregnane-11,20-dione-21-acetate, M_r 390.52, forms colorless crystals, *mp* 175 – 177 °C, $[\alpha]_D^{26}$ + 97 (c = 1.02 mg/mL in $CHCl_3$). To prepare alphadolone, alphaxalone is reacted with lead tetraacetate in the presence of a Lewis acid, such as BF_3 etherate [47].

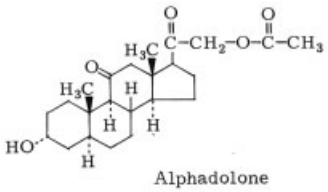

Alphadolone

Althesin is an effective intravenous anesthetic for short surgical procedures or for induction of lengthy anesthesia [45]. Doses of 0.050 – 0.075 mL per kg of body weight produce anesthesia in adults for 5 – 20 min and analgesia to surgical stimuli for 2 – 5 min. Althesin causes cardiodepressive side effects. Although it is used in countries in which French pharmaceuticals dominate, it is hardly used in Germany.

Trade Names. Althesin and Saffan (Glaxo); Aurantex (Glaxo, Bad Oldestoe).

4.4. Neuroleptanalgesics

Even with adequate somnolence and analgesia, surgical stress factors can be suppressed only if the activity of the deeper cerebral regions is subdued also. This is accomplished with small doses of neuroleptics (→ Neuropharmacology) to induce the so-called neuroleptanalgesia (see page 2). A potent analgesic is used in conjunction with a neuroleptic, the dosage usually adjusted for each individual. The first neuroleptanalgesic was the "coctail lytique" of Laborit and Huguenard, a mixture of pethidine as the analgesic, chlorpromazine as the neuroleptic, and promethazine as an antihistamine with pronounced sedative effect. In the course of further development, two combinations of neuroleptic and very potent analgesic emerged:

1) Haloperidol and phenoperidine (≈ 70 times more active than pethidine)
2) Droperidol and fentanyl (≈ 100 times more active than morphine)

Only the second is still important. A newer development is alfentanyl, *N*-[1-[2-(4-ethyl-4,5-dihydro-5-oxo-1*H*-tetrazol-1-yl)-ethyl]-4-(methoxymethyl)-4-piperidinyl]-*N*-phenylpropanamide hydrochloride (Johnson & Johnson), which was developed from fentanyl and is an effective anesthetic analgesic in phase III of clinical trials [48].

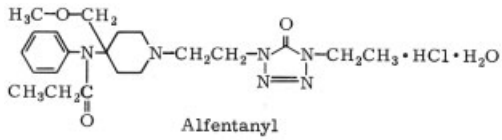

Alfentanyl

Droperidol [*548-73-2*], 1-[3-(4-fluorobenzoyl)-propyl]-4-(2-oxo-1-benzimidazolinyl)-1,2,3,6-tetrahydropyridine, M_r 379.44. The monohydrate melts at 145 – 146.5 °C. It is prepared by alkylating (2-oxo-1-benzimidazolinyl)-1,2,3,6-tetrahydropyridine with 4-fluoro-γ-chlorobutyrophenone [49].

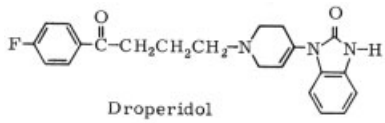

Droperidol

Droperidol is a very potent neuroleptic from the phenylbutyrophenone group. It is never used in outpatient practice. To induce narcosis, 15 – 25 mg is injected, followed by further doses as needed. Narcosis begins rapidly, and the duration of action is short. Extrapyramidal symptoms are absent to a large extent. For anesthesia, the analgesic fentanyl is administered simultaneously.

Trade Names. Dehydrobenzperidol (Janssen, Düsseldorf); Thalamonal, Innovar in the USA, (Janssen) is a combination of droperidol with fentanyl. (One milliliter contains 0.05 mg of fentanyl base and 2.5 mg dehydrobenzperidol.)

Fentanyl [*437-38-7*], *N*-phenyl-*N*-[1-(2-phenylethyl)-4-piperidinyl]-propanamide, M_r 336.46, forms colorless crystals, *mp* 83 – 84 °C. The bitter-tasting dihydrogen citrate [*990-73-8*],

mp 149 – 151 °C, is soluble in water (1 g per 40 mL at 20 °C) and methanol. For preparation → Analgesics and Antipyretics.

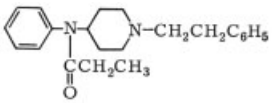

Fentanyl belongs to the group of base-substituted propionanilide analgesics [50] and is a typical morphine agonist. Because of the danger of addiction, the short-acting fentanyl is used exclusively for anesthesia. A dose of just 0.2 – 0.3 mg produces complete short-term analgesia in adults. Because it depresses respiration, it is necessary to control breathing during neuroleptanalgesia. Afterward, respiratory depression can be alleviated by a morphine antagonist such as 1-*N*-allyl-3-hydroxymorphinan (Lorfan).

Trade Names. Fentanyl (Janssen, Düsseldorf), Sublimaze (Janssen); Thalamonal, see trade names for Droperidol.

5. References

General References

1. M. E. Greig in V. E. Grill (ed.): *Theories of Anaesthesia in Pharmacology and Medicine,* McGraw-Hill, New York 1954. H. Kilian, H. Weese: *Die Narkose,* Thieme Verlag, Stuttgart 1954. M. B. Chenoweth (ed.): "Modern Inhalation Anaesthetics," *Handbuch der experimentellen Pharmakologie,* Heffter-Heubner, New Series, vol. XXX, Springer Verlag, Berlin 1972. AMA, Drug Evaluations by American Medical Assoc., 3rd ed., Publ. Sciences Group, Acton, Mass., 1977, pp. 285 – 299. J. W. Dundee, W. C. McCaughy: "Drugs in Anaesthetic Practice," in G. S. Avery: Drug Treatment, Adv. Press, Sidney 1976, pp. 215 – 248. J. W. Dundee, G. M. Wyant: *Intravenous Anaesthesia,* Churchill & Livingstone, Edinburgh 1974. J. W. Dundee: "New Intravenous Anaesthetics", Br. J. Anaesth. 51 (1979) 641 – 648. K. A. Lehmann: "Narkose – angewandte Toxikologie?", Pharm. Unserer Zeit 11 (1982) 18 – 32. B. Helwig, M. Helwig: *Moderne Arzneimittel,* 5th ed., Wissenschaftl. Verlags GmbH., Stuttgart 1982, pp. 233 – 248. Martindale: *The Extra Pharmacopoeia,* 28th ed., The Pharmaceutical Press, London 1982, pp. 740 – 761. H. Benzer, R. Frey, W. Hügin, O. Mayrhofer: *Lehrbuch der Anaesthesiologie, Reanimation und Intensivtherapie,* Springer Verlag, Berlin 1977. R. Dudziak: *Lehrbuch der Anästhesiologie,* F. K. Schattauer Verlag, Stuttgart – New York 1980. T. C. Gray, J. F. Nunn, J. E. Utting: *General Anaesthesia,* Butterworths, London 1980. C. Nemes, M. Niemer, G. Noack: *Datenbuch Anästhesiologie,* G. Fischer Verlag, Stuttgart – New York 1979.

Specific References

2. ICI, GB 767779, 1954.
3. Hoechst, US 2956624, 1960.
4. J. F. Artusio, J. Weingram, Y. J. Sohn, *Anesth. Analg. (Cleveland)* **46** (1967) 657. G. W. Black, R. S. Carke, P. J. Howard, H. McCullough, *Br. J. Anaesth.* **41** (1969) 288.
5. Dow Chemical, US 2971990, 1957.
6. Merck Sharp & Dohme, US 2021872,1931. Squibb, US 2136387, 1936.
7. Consortium für elektrochem. Ind., US 2832807, 1953.
8. *Drugs Future* **1** (1976) 307, 7 (1982) 440.
9. *Drugs Future* **2** (1977) 153.
10. *Drugs Future* **5** (1980) 488.
11. *Drugs Future* **3** (1978) 323.
12. R. C. Terrell, L. Speers et al., *J. Med. Chem.* **14** (1971) 517.
13. R. C. Terell, L. Speers et al., *J. Med. Chem.* **15** (1972) 604. Air Reduction Co., DE 1 138 406,1969; US 3 469 011,1969; US 3527813,1970.
14. W. T. Miller, Jr., E. W. Fager, P. H. Griswold, *J. Am. Chem. Soc.* **70** (1948) 431. J. D. Park, . C. H. Snow, J. R. Laiker, *J. Am. Chem. Soc.* **73**(1951) 861. Dow Chemical, GB 928786, 1960.
15. Air Reduction Co., DE 1 814 962, 1969; US 3 535 388, 1967; US 3535425, 1967. Hoechst, DE 2344442, 1973.
16. W. C. Stevens, T. H. Cromwell et al., *Anesthesiology* **35** (1971) 8 – 53.
17. Air Reduction Co., US 2830007, 1958.
18. Air Reduction Co., US 2870218, 1959.
19. J. W. Dundee, *Br. J. Anaesth.* **51** (1979) 641.
20. Eli Lilly, US 2872448, 1956.
21. IG-Farbenind., DE 590175, 1933. *Beilstein* **24 (2)**, 294.
22. D. L. Tabern, E. H. Volwiler, *J. Am. Chem. Soc.* **57** (1935) 1961. F. S. Crossley, *J. Org. Chem.* **5** (1940) 238.

23. E. Miller, J. C. Munch, F. S. Crossley, W. H. Hartung, *J. Am. Chem. Soc.* **58** (1936) 1090. Abbott, US 2876225,1956.
24. Parke-Davis, US 2153711, 1934. Abbott, US 2153729, 1934.
25. Lilly, GB 613704, 1968. Abbott, US 2876226, 1959.
26. J. K. Abe, T. Ischisaka, Y. Tsukamato, *Yakugaku Zasshi* **75** (1955) 891.
27. E. Miller, J. C. Munch, F. S. Crossley, W. H. Hartung, *J. Am. Chem. Soc.* **58** (1936) 1090.
28. IG-Farbenind., DE 692928, 1935.
29. H. Weese, F. H. Koss, *Dtsch. Med. Wochenschr.* **79** (1954) 601.
30. B. E. Schildt, E. E. Schildt, *Acta Pharmacol. Toxicol.* **19** (1962) 377.
31. Abbott, US 2153730, 1936.
32. Bayer, DE 1134981, 1960.
33. R. Hiltmann, H. Wollweber, F. Hoffmeister, W. Wirth, *Anaesthesiol. Wiederbeleb.* **4** (1961) 61.
34. BASF, BE 624258, 1963.
35. C. L. Stevens, J. L. Keundt, M. E. Munk, M. D. Pillai, *J. Org. Chem.* **30** (1965) 2967. Parke-Davis, BE 634208, 1963; US 3254124, 1966. E. F. Domino, P. Chodoff, G. Lorssen, *Clin. Pharmacol. Ther.* **6** (1966) 279. G. C. Chen, C. R. Ensor, B. Bohner, *J. Pharmacol. Exp. Ther.* **152** (1966) 332.
36. G. Jones, *J. Am. Chem. Soc.* **71** (1949) 644. F. Godefroi, P. A. J. Janssen et al. *J. Med. Chem.* **8** (1965) 220. Janssen Pharmaceutica, DE-OS 2609573,1976.
37. P. A. J. Janssen, C. J. E. Niemegeers et al. *Arzneim. Forsch.* **21** (1971) 1234. D. H. Kettler, H. Sonntag et al., *Anaesthesist* **23** (1974) 116.
38. *Annu. Drugs Data Rep.* **2** (1979/80) 88. *Drugs Future* **1** (1976) 461.
39. C. S. Marvel, E. R. Birkhumer, *J. Am. Chem. Soc.* **51** (1926) 260.
40. *Drugs Future* **7** (1982) 156.
41. Hoffmann-La Roche, DE-OS 2540522, 1975.
42. *Drugs Future* **3** (1978) 822, **4** (1979) 852, **5** (1980) 585, **6** (1981) 736.
43. H. Selye, *J. Immunol.* **41** (1941) 259. S. Y. P'an, G. D. Laubach, H. W. Rudel, *Science (Washington D.C.)* **122** (1955)78. Pfizer, US 2708651,1955.
44. *Drugs Future* **5** (1980) 614.
45. A. Benk, G. Gogolate, Ch. Stumpf, Ch. Tschekaloff, *Postgrad. Med. J. Suppl.* **2** (1972) 120 – 123. K. J. Child, J. P. Currie et al., *Br. J. Anaesth.* **43** (1971) 2 – 24. E. J. Stokes, R. E. M. Thomson et al., *Lancet* 1972, 88.
46. W. Nagate, C. Tamm, T. Reichstein, *Helv. Chim. Acta* **42** (1959) 1399.
47. P. A. Browne, D. N. Kirk, *J. Chem. Soc. C* **1969,** 1653. Glaxo, DE 2030402, 1971.
48. *Drugs Future* **6** (1982) 355.
49. Janssen Pharmaceutica, US 3141823, 1964.
50. W. B. Wright, H. J. Brabander, R. A. Hardy, *J. Am. Chem. Soc.* **81** (1959) 1518, *J. Org. Chem.* **26** (1961), 476, 485; H. Wollweber, *Eur. J. Med. Chem. Chim. Ther.* **17** (1982) 125. W. Geiger, H. Wollweber, *Eur. J. Med. Chem. Chim. Ther.* **17** (1982) 207.

Anhydrite → Calcium Sulfate

Aniline

THOMAS KAHL, BASF AG, Ludwigshafen, Germany (Chaps. 1 – 7)

KAI-WILFRID SCHRÖDER, BASF AG, Ludwigshafen, Germany (Chaps. 1 – 7)

F. R. LAWRENCE, E. I. du Pont de Nemours & Co., Deepwater, New Jersey 08023, United States (Chap. 8)

W. J. MARSHALL, E. I. du Pont de Nemours & Co., Deepwater, New Jersey 08023, United States (Chap. 8)

HARTMUT HÖKE, Weinheim, Germany (Chaps. 9, 10)

RUDOLF JÄCKH, BASF AG, Ludwigshafen, Germany (Chaps. 9, 10)

1. Introduction 1
2. Physical and Chemical Properties 1
3. Production 3
3.1. Raw Materials 3
3.2. Processes 3
3.2.1. Catalytic Hydrogenation of Nitrobenzene 4
3.2.1.1. Catalytic Vapor-Phase Hydrogenation 4
3.2.1.2. Catalytic Liquid-Phase Hydrogenation 5
3.2.2. Reduction of Nitrobenzene with Iron and Iron Salts 5
3.2.3. Amination of Phenol 6
4. Quality Specifications 6
5. Handling, Storage and Transportation 6
6. Uses 7
6.1. Methylene Diphenylene Isocyanate (MDI) 7
6.2. Rubber Processing Chemicals . . 7
6.3. Dyes and Pigments 8
6.4. Agricultural Chemicals 8
6.5. Pharmaceuticals 8
6.6. Cyclohexylamine/Dicyclohexylamine 8
6.7. Miscellaneous 8
7. Economic Aspects 8
8. Derivatives 9
9. Toxicology 12
9.1. Toxicokinetic Properties and Metabolism 12
9.2. Toxicological Properties 13
10. Occupational Health 14
11. References 14

1. Introduction

Aniline is the parent molecule of the family of aromatic amines (→ Amines, Aromatic). First isolated in the early 19th century (O. UNVERDORBEN 1826, F. RUNGE 1834, C. J. FRITZSCHE and N. N. ZININ 1841), the structure of the substance was finally proven in 1843 by A. W. VON HOFMANN, who was able to obtain aniline by the reduction of nitrobenzene.

The first technically applicable process (Bechamp process, see Chap. 3.2.2) for the production of aniline was developed as early as 1854.

Over the last 145 years aniline has become one of the hundred most important building blocks in chemistry. Aniline is used as an intermediate in many different fields of application as, for example, isocyanates, rubber processing chemicals, dyes and pigments, agricultural chemicals and pharmaceuticals.

2. Physical and Chemical Properties

Physical Properties. Aniline [*62-53-3*], $C_6H_5NH_2$, aminobenzene, benzamine (CAS name), phenylamine, aminophen when freshly distilled, is a colorless oily liquid with a characteristic sweet aminelike odor. Aniline turns to brown on exposure to air and light. Aniline is partially soluble in water and miscible with most organic solvents (e.g. ethanol, acetone, and benzene). Important physical properties of aniline are as follows [1]:

Molecular mass (M_r)	93.13
Boiling point (101.3 kPa)	184.4 °C
(4.4 kPa)	92 °C
(1.2 kPa)	71 °C

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_303

Solidification point	− 6.2 °C
Flash point (DIN 51758)	76 °C
Explosion limits	
Lower	1.2 vol %
Upper	11.0 vol %
Ignition temperature (DIN 51758)	540 °C
Vapor pressure (20 °C)	0.05 kPa
(100 °C)	7.0 kPa
Saturation concentration (air, 20 °C)	390 mL/m^3
(air, 30 °C)	880 mL/m^3
Odor threshold	0.5 mL/m^3
Density vapor (at *bp*, air = 1)	3.2
Density liquid (20 °C)	1.022 g/cm^3
(60 °C)	0.990 g/cm^3
Refractive index n_D^{20}	1.5863
Viscosity (20 °C)	4.35 mPa s
(60 °C)	1.62 mPa s
Solubility (20 °C)	
Aniline in water	3.6 %
Water in aniline	5.5 %
pH value (3.6 g aniline per liter, 20 °C)	8.8
log p_{ow} (distrib. coeff. *n*-octanol/water)	0.91
Specific heat (25 °C)	2.1 Jg^{-1}K^{-1}
Heat of vaporization (at *bp*)	478.6 J/g
Heat of combustion	36.4 kJ/g
Critical temperature	425.6 °C
Critical pressure	5.30 MPa
Dielectric constant (25 °C)	6.89

Chemical Properties. The chemistry of aniline is determined by the primary amino group (Fig. 1) and by the activated benzene nucleus (Fig. 2) [2].

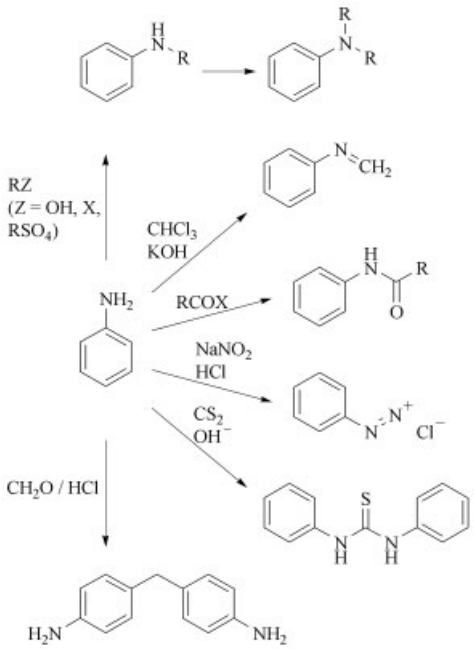

Figure 1. Reactions of the amino group

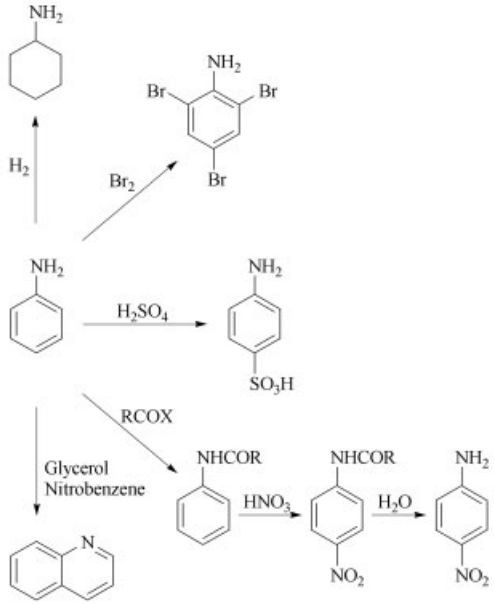

Figure 2. Reactions of the aromatic ring

The lone electron pair on the nitrogen of the amino group is partially delocalized to the aromatic ring system [3]. Therefore aniline is only a weak base ($pK_a = 4.60$) as compared with aliphatic amines such as cyclohexylamine ($pK_a = 10.64$). Nevertheless aniline forms stable, water-soluble salts with strong mineral acids (e.g., hydrochloric or sulfuric acid). The formation of aniline sulfate can be used to protect the amino group of aniline against oxidation during subsequent synthesis steps.

Various alkylating agents (like alcohols, alkylhalides, alkylsulfates, and olefins) alkylate the amino group to form *N*-alkyl and *N,N*-dialkylaniline in Lewis acid catalyzed reactions at elevated temperature and pressure. The reaction of aniline with chloroform in a caustic potash solution produces phenylisocyanide. *N,N*-Diphenylthiourea is the product of the reaction of aniline and carbon disulfide in an alkaline alcohol solution. A characteristic reaction of primary aromatic amines like aniline is the formation of diazonium salts with nitrous acid. Aniline undergoes condensation reactions with alkylaldehydes. The most important reaction is the formation of 4,4′-methylenedianiline (MDA) by condensation of aniline with formaldehyde in the presence of hydrochloric acid. Aniline reacts with organic acids and acid derivatives to

form amides. Aniline undergoes a variety of oxidation reactions, depending on the oxidizing agent and the conditions. Mononuclear products like *p*-benzoquinone (MnO_2, H_2SO_4), bimolecular compounds such as azobenzene (air) or polymeric substances like aniline black (Cu or V, $Na_2Cr_2O_7$ or $NaClO_3$) may be formed (see Fig. 2).

The aromatic ring of aniline is catalytically hydrogenated to give cyclohexylamine and/or dicyclohexylamine at elevated temperature and pressure, depending on conditions and type of catalyst. The presence of the amino group activates the ortho and para positions of the aromatic ring and, as a result, aniline reacts readily with electrophilic substances. For example, the bromation of aniline yields 1,3,5-tribromoaniline under mild conditions. The reaction with sulfuric acid at high temperature forms almost exclusively 4-aminobenzene sulfonic acid (4-sulfanilic acid). Nitration of aniline (after protecting the amino group against oxidation by the formation of acetanilide) forms the 4-nitro compound. Numerous N-heterocyclic compounds can be obtained from aniline. For example quinoline is produced in the Skraup synthesis by condensation of aniline with glycerol or acrylaldehyde (acrolein) in nitrobenzene.

3. Production

3.1. Raw Materials

Most commercially used synthesis of aniline start from benzene. There is some literature about direct amination of benzene, but the high temperature and pressure required and the need to use an extreme excess of ammonia never allowed the development of an economic process [4]. DuPont and Mitsui Toatsu examined the oxidative amination process of benzene forming water as a byproduct [5].

Up to now, all technically applied aniline syntheses use an indirect way to produce aniline from benzene. In all cases a kind of derivatization is included as an intermediate step. In this step one of the two direct precursors of aniline is formed: nitrobenzene or phenol.

Nitrobenzene (→ Nitro Compounds, Aromatic). Nitrobenzene is manufactured commercially by the direct nitration of benzene in liquid phase using a mixture of nitric and sulfuric acid (mixed acid, nitrating acid). There are basically two thermodynamic ways for the nitration of benzene:

isothermal reaction
adiabatic reaction

In the isothermal process the nitration of benzene takes part in a reaction cascade of reactors (stirred cylindrical reactor or tubular reactor) at a temperature of 50 to 100 °C and ambient pressure. The temperature is maintained by internal and/or external cooling with cooling water. An advantage of the isothermal process, caused by the low reaction temperature, is the very low formation rate of byproducts (nitrophenols, picric acid) [6]. The nitrating acid is a mixture of fuming sulfuric acid (oleum) and concentrated nitric acid. In the adiabatic process the nitration of benzene takes part in a cascade of stirred reactors or a jet impingement reactor [7] at a temperature of 90 to 190 °C and ambient pressure. As nitrating acid a large volume of 65 to 85 % sulfuric acid and concentrated nitric acid are used. An advantage of the adiabatic process is that the reaction heat of the nitration, heating up the reaction mixture, is used to reconcentrated the spent sulfuric acid. This is done by azeotropic distillation of water/benzene, by stripping water with an inert gas, or distillation of the spent acid under reduced pressure.

Phenol (→ Phenol). The cumene oxidation (Hock process) process is the most important commercial process to synthesize phenol.

Cumene is prepared by alkylation benzene with propene using a carrier catalyst containing absorbed phosphoric acid, at temperature of ca. 230 °C and at a pressure of 3500 kPa [8]. Another process is the toluene oxidation route via benzoic acid.

3.2. Processes

Nitrobenzene is used as raw material for aniline production by all world producers with the exception of *Mitsui Petrochemicals Ind.* (Japan) who additionally uses phenol as starting material and *Aristech Chemical Corp.* (United States), who only uses the phenol route [9].

3.2.1. Catalytic Hydrogenation of Nitrobenzene

The highly exothermic catalytic hydrogenation ($\Delta H = -544$ kJ/mol at 200 °C) of nitrobenzene is performed both in the vapor and in the liquid phase in commercially used processes (see Fig. 3).

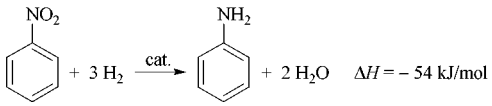

The exchange and utilization of the heat of reaction is a crucial point for all processes using nitrobenzene as a raw material [10].

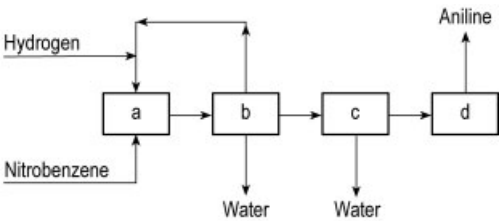

Figure 3. Simplified flow sheet of the reaction route nitrobenzene to aniline
a) Hydrogenation; b) Separation; c) Dewatering; d) Rectification

3.2.1.1. Catalytic Vapor-Phase Hydrogenation

Nitrobenzene is hydrogenated to aniline, usually in more than 99 % yield, using fixed-bed or fluidized-bed vapor-phase processes. The most effective catalysts for the gas-phase hydrogenation of nitrobenzene seem to be copper or palladium on activated carbon or an oxidic support, in combination with other metals (Pb, V, P, Cr) as modifiers or promoters in order to achieve high activity and selectivity.

In the *Lonza process*, which is operated by First Chemical Corp., a homogenized feed of hydrogen and nitrobenzene is passed over a fixed-bed catalyst of copper on pumice with an inlet temperature of about 200 °C (example: 215 °C). The homogenization is done by spraying the nitrobenzene with the help of the fresh hydrogen feed into the heated (example: 240 °C) circulating gas stream at a fixed position. The molar ratio of nitrobenzene feed to total hydrogen is about 1 : 100 at the reactor inlet. The reaction products leave the reactor with a temperature of more than 300 °C (example: 325 °C). The heat of hydrogenation is used for steam production and for heating up the recycle gas stream. The reactor outlet is cooled down further in a condenser; excess hydrogen, crude aniline and water are separated. Aniline is purified by distillation [11, 12, 14].

Bayer operates conventional fixed-bed reactors using a palladium catalyst on a alumina support, modified in its activity by the addition of vanadium and lead [13] and claims the adiabatic hydrogenation of nitrobenzene over a fixed-bed catalyst of 1.5 to 4 wt % palladium on coke with 0.1 to 2 wt % lead as a modifier to reduce aromatic ring hydrogenation. At a pressure of 100 – 700 kPa a mixture of vaporized nitrobenzene and hydrogen in a molar ratio of 1 : 120 to 1 : 200 is fed to the adiabatic reactor with an inlet temperature of 250 – 350 °C. The height of the catalysts' bed in the reactor is 0.1 to 1. 0 m The reaction products leave the reactor without cooling at a maximum temperature of 460 °C. After leaving the reactor, the heat of reaction is used for the production of high pressure steam. A production unit can be built of several serial and/or parallel adiabatic reactors. After cooling down to 140 to 180 °C, the outlet of the final reactor is fed to a separation unit, where after further cooling crude aniline, crude wastewater, and the recycled hydrogen are separated under pressure. The crude aniline is purified by distillation [15–18].

BASF operates a vapor-phase, fluidized-bed process [19]. Nitrobenzene is partially evaporated by atomizing it with the aid of a hot stream of gas consisting substantially of hydrogen. The stream of gas is circulated in the presence of a fluidized catalyst, the reaction products are condensed, and aniline is separated from the isolated crude reaction products. One type of preferred catalyst is copper (≈ 15 wt %) on a silica support promoted with chromium, zinc, and barium. The two phase mixture of nitrobenzene and hydrogen is injected through nozzles located at several heights in the fluidized bed and the hydrogenation is carried out at 250 – 300 °C and 400 – 1000 kPa in the presence of excess hydrogen. The hot product gas is cooled by passing it through a heat exchanger, and aniline is iso-

lated in a liquid – gas separator. The reaction heat is used for steam production. For catalyst regeneration, after flushing the whole system with nitrogen, the organic material deposited on the catalyst surface is burned off at 200 to 250 °C with air. After completion of the regeneration and subsequent replacement of the air in the system with nitrogen, the catalyst is activated again by reducing the copper oxide to copper with hydrogen at 200 to 300 °C [20], [21].

3.2.1.2. Catalytic Liquid-Phase Hydrogenation

The industrial aniline processes of *ICI* and *DuPont* involve hydrogenation in the liquid phase. Liquid-phase hydrogenation processes are operated at 90 – 200 °C and 100 – 600 kPa. The liquid phase reaction may be carried out in slurry or in fluidized-bed reactors. Conversion of nitrobenzene is normally complete after a single reactor pass with yields of 98 to 99 %.

In the 1960s, *ICI* developed a continuous, liquid phase hydrogenation process, that used aniline as the solvent in a proportion > 95 wt % of the liquid phase. By operating at or near the boiling point of the solvent (usually at pressures < 100 kPa), some or all of the heat of the reaction is dissipated by allowing the reaction mixture to evaporate. Water is removed with the effluent vapors and sufficient aniline is returned to the reaction vessel to maintain steady state conditions. One preferred catalyst is finely divided nickel on kieselguhr [22].

DuPont hydrogenates nitrobenzene in liquid phase using a platinum – palladium catalyst on a carbon support with iron as modifier. The modifier provides good catalyst life, high activity, and protection against hydrogenation of the aromatic ring. The continuous process uses a plug-flow reactor that achieves essentially quantitative yields, and the product exiting the reactor is virtually free of nitrobenzene [23], [24].

A comparison of the catalytic liquid-phase and vapor-phase hydrogenation of nitrobenzene shows virtually no differences in yield and product quality for both processes. The liquid-phase process has the advantage of a higher space-time yield and no need for a recycle gas loop (lower energy requirement), the vapor-phase process has the advantage of a very effective utilization of the heat of reaction (steam production), no need for a product-catalyst separation and longer catalyst live.

3.2.2. Reduction of Nitrobenzene with Iron and Iron Salts

A historical variation of the nitrobenzene route is the *Béchamp process*, which uses iron and iron(II) chloride for the reduction.

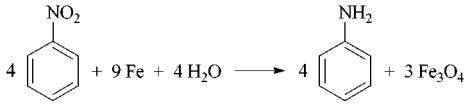

This process is more than one hundred years old, but still used in two *Bayer* aniline plants. Nowadays the product of interest is no longer the aniline, but the colored iron oxide pigments, that are formed as byproducts. In the Béchamp process nitrobenzene is reduced in an agitated reaction vessel with iron(II) chloride solution and ground iron fillings. The reactor is filled with the total amount of water required for the reaction (aniline water from aniline distillation), 20 % of the iron, the total amount of catalyst needed, and about 5 % to 10 % of the total nitrobenzene feed. Under intense agitation the reactor content is heated up to reflux. After the reduction has started the remaining nitrobenzene and iron are added slowly to avoid excessive temperature and pressure buildup. For the completion of the reduction the vessel is heated up to 100 °C for two more hours after the end of the iron/nitrobenzene addition. The reaction is completed in about 8 to 10 h [25].

The reaction mixture is neutralized with lime; it is then transferred to a separator, and the organic phase containing the aniline is withdrawn. Aniline is recovered from the organic phase by water stripping and distillation. Residual aniline is recovered from the material remaining in the separator before the iron oxide powder slurry is processed into a fine particle, colored pigment. The color of the iron oxide byproduct can be controlled by additives to the reaction medium, by the use of different types of iron, and by subsequent calcination conditions [26].

3.2.3. Amination of Phenol

In the commercial phenol route developed by *Halcon*, phenol is aminated in the vapor phase using ammonia in the presence of a silica – alumina catalyst.

$\Delta H = -544$ kJ/mol

The reaction is mildly exothermic ($\Delta H = -8.4$ kJ/mol) and reversible, so high conversion is obtained only by the use of excess ammonia (mole ratio of 20 : 1) and a low reaction temperature, which also reduces the dissociation of ammonia. Byproduct impurities include diphenylamine, triphenylamine and carbazole. Their formation is also inhibited by the use of excess ammonia. Yields based on phenol and ammonia are $\geqq 96$ % and 80 %, respectively [28].

In the process (Fig. 4) phenol and fresh and recycle ammonia are vaporized separately (to prevent yield losses) and combined in the fixed bed amination reactor (a) containing the silica – alumina catalyst. After the reaction at 370 °C and 1.7 MPa, the gas is cooled, partly condensed and the excess ammonia is recovered in a separation column (b), compressed and recycled. The condensation product is passed through a drying column to remove water and then through a finishing column to separate aniline from residual phenol and impurities in vacuum (less than 80 kPa). The phenol, containing some aniline (azeotropic mixture) is recycled [29], [30].

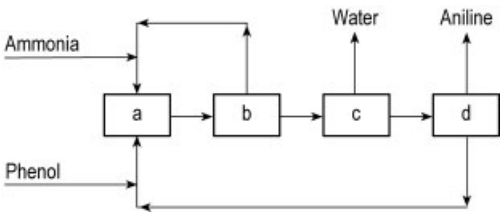

Figure 4. Simplified flow sheet of the reaction route phenol to aniline
a) Amination; b) Ammonia recovery; c) Dewatering; d) Rectification

Comparison of the phenol route to the nitrobenzene route, both starting from benzene, shows four steps for the former (cumene to phenol involves two steps via the intermediate cumene hydroperoxide) versus two for the latter. The nitrobenzene route has an overall advantage in yield and lower total energy requirements. The phenol route has an advantage in prolonged catalyst life and product quality. It is the preferred method if low cost phenol is available (phenol producers).

4. Quality Specifications

The common standard specification of commercial-grade aniline is listed below:

Aniline	min. 99.9 %
Nitrobenzene	max. 2 ppm
Water	max. 500 ppm
Color	max. 100 APHA

For special applications the concentration of trace impurities like cyclohexylamine, cyclohexanol, cyclohexanone, toluidines, phenol, phenylcyclohexylamines, and dicyclohexylamine may be specified as well.

The analysis of the specified organic impurities is normally carried out by gas chromatography using a nonpolar stationary phase (dimethylpolysiloxane) and a flame ionization detector.

The water concentration in aniline is determined with the coulometric Karl-Fischer titration.

The color of the aniline is determined spectrophotometrically. Freshly distilled aniline is a colorless liquid (< 10 APHA) which becomes brownish by exposure to light and air.

5. Handling, Storage and Transportation

Aniline is slightly corrosive to some types of metal. So all amphoteric materials such as aluminum, copper, tin, zinc, and alloys containing one of these metals (brass, bronze) are not suitable for the handling of aniline, as they are corroded by it. For normal applications carbon steel or cast iron are appropriate materials for the aniline handling or storage. Only if discoloration

must be kept to minimum, aniline should be stored and transported in stainless steel equipment with proper nitrogen blanketing. Suitable gasket materials include for example graphite, polytetrafluoroethylene or poly(vinylidene fluoride), best reinforced with glass fiber.

Aniline is transported in drums, containers for the transportation on rail, water and street, railway tank cars and tank trucks as well as in tanks of ships for river and sea transportation.

Aniline is classified as a hazardous chemical substance for transportation (UN no. 1547). Because of its chemical properties it is classified as a poisonous liquid (class 6.1 12 b) class b (dangerous substance). Therefore, aniline must be transported following the Hazardous Materials Transportation Act regulations regarding loading, handling, and labeling.

6. Uses

In 1996 aniline was known to be used in more than 300 different end products (see Table 1). Thereof the polyurethane building block 4,4-methylene-di-paraphenylene isocyanate (MDI), accounting for 2/3 of the total demand, was of paramount importance. With an expected growth well above the increase of the average global gross domestic product, MDI in future will extent its position as dominating outlet for aniline even further.

Table 1. Worldwide consumption of aniline by application

Application	Percentage of consumption
MDI	65 %
Rubber processing chemicals	18 %
Dyes and pigments	5 %
Agricultural chemicals	4 %
Pharmaceuticals	2 %
Cyclohexylamine/dicyclohexylamine	2 %
Miscellaneous	4 %

6.1. Methylene Diphenylene Isocyanate (MDI)

MDI is one of the principal isocyanates that is reacted with alcohols such as polyols and polyetherols to produce polyurethanes (PU).

MDI based PU systems are mainly used in construction, furniture, automotive industry, and insulation. Due to the wide spectrum of traditional end uses and new fields of application MDI-PU will continue to show a strong growth especially in emerging markets.

Worldwide production capacity for MDI in 1996 was 2.35×10^6 t. Biggest producer was Bayer (capacity: 580×10^3 t), followed by ICI, BASF, and, Dow. Whereas ICI and BASF are fully back-integrated (i.e., self-sufficient), Bayer and Dow to a certain extent depend on aniline supplies from third parties.

After double digit growth rates in the mid 1990s, future use of aniline for MDI is projected to grow at a more modest rate of 6 – 8 % per year.

A typical MDI process used by commercial producers is as follows:

A mixture of polymeric methylenephenylamines is prepared by condensing aniline with formaldehyde in the presence of HCl. The reaction involves intermediate formation of amine hydrochlorides, which subsequently are neutralized with caustic soda. The reaction usually is run with an excess of aniline at subatmospheric pressure and temperatures between 70 and 105 °C. The yield is about 96 %.

The polyamine product is converted to a crude mixture of isocyanates by reaction with phosgene in a solvent. Reaction temperature is about 120 °C, pressure is 345 kPa, and yield ca. 97 %. Chlorobenzene is removed by distillation. The crude is separated by vacuum distillation into pure MDI and polymeric MDI. Hydrogen chloride is produced as a byproduct of the reaction.

6.2. Rubber Processing Chemicals

In the first decade of this century pure aniline was used as a vulcanization accelerator. The call for higher effectiveness and safer handling led to the development of aniline based mercaptothiazole and sulfenic amide accelerators (→ Rubber, 4. Chemicals and Additives, Chap. 2.3.1.) which at present account for ca. 80 % of all vulcanization accelerators used worldwide.

Of even bigger importance for aniline within the rubber processing chemicals sector are antidegradants (i.e., antioxidants, antiozonants) such as paraphenylenediamines (PPD), quinolines, and diphenylamine (→ Rubber,

4. Chemicals and Additives, Chap. 3.2.1.). Aniline is feedstock to roughly 70 % of all antidegradants consumed worldwide.

Future use of aniline in the rubber industry is forecasted to increase by 2 – 3 % per year, equivalent to the worldwide growth in demand for synthetic and natural rubber.

6.3. Dyes and Pigments

Intermediates for dyes and pigments cover more than 50 % of all known formulations using aniline as a raw material, albeit most of them are of minor importance. The largest classes are the mono-, di-, and triazo compounds. Production of dyes and pigments has shifted toward Asian countries, such as China and India. However, some world-scale dyes and pigments plants in Europe and NAFTA (United States, Canada, Mexico) are still using aniline, e.g., for the production of indigo. Since the industrial realization of the chemical indigo syntheses at the end of the 19th century by BASF, indigo (→ Indigo and Indigo Colorants) is the most important dye based on aniline.

Worldwide growth for this application is expected to be at a low rate of 1 – 2 % per year.

6.4. Agricultural Chemicals

More than 40 active substances for pesticides (herbicides, fungicides, and insecticides) use aniline as raw material. Most important are amide and urea herbicides. NAFTA-located companies stand for more than 50 % of the world′s aniline consumption for agricultural chemicals. Aniline-based active substances are predominantly in the later stage of their life cycle and are about to be substituted. Global consumption therefore is forecasted to decrease by 1 – 2 % per year.

6.5. Pharmaceuticals

As a stable, although small outlet, aniline is mainly used for preparation of analgesics, antipyretics, antiallergics, and vitamins. A certain growth in demand for drugs is expected for the next five years, counterbalanced by a decreasing aniline consumption for production of vitamins where the aniline route is being abandoned in favor of fermentation technology.

6.6. Cyclohexylamine/Dicyclohexylamine

In the 1990s aniline became the raw material of choice for the production of these amines. Both are predominantly used as intermediates for sulfenamide vulcanization accelerators. Cyclohexylamine is also used in large quantities as corrosion inhibitor and feedstock for cyclamates.

Whereas the usage of these amines in accelerators and water treatment is expected to grow at a rate of ca. 3 – 4 % per year, the consumption of cyclohexylamine for production of cyclamates has fluctuated considerably and is difficult to predict. A boost in demand can be expected should the FDA lift the ban on the usage of cyclamates as artificial sweeteners in the United States.

6.7. Miscellaneous

Among various other applications (production of alkylanilines, optical brighteners, sulfonic acids, etc.) the consumption of aniline for synthesis of aramid fibers is worth mentioning. Aramids are suitable for a wide spectrum of applications, mainly as substitutes for steel, showing equivalent reinforcement characteristics at far lower weight. Because of additional use in other applications, e.g., the replacement of asbestos in brake linings, healthy growth rates for aramids can be expected.

7. Economic Aspects

A breakdown of aniline consumption by region is given in Table 2.

Table 2. Aniline consumption by region

Region	Percentage of consumption
Europe	39 %
Americas	34 %
Asia	27 %

Much of the growth in demand for aniline in recent years has originated in Asia, where extensive use of MDI based polyurethanes in construction and consumer goods was the driver for the entire urethanes sector. A substantial part of Asia′s demand for urethanes and urethane raw materials, however, was imported from Europe and the United States. It is safe to say that Asia will see stronger growth in demand for aniline because major MDI producers are pushing ahead with plans to build large plants in the region.

World aniline production capacity was estimated at ca. 2.3×10^6 t for 1996. Capacity by geographic area in 10^3 t was:

Europe	960
Americas	820
Japan	270
Asia exc. Japan	250
Africa	7

In 1996 worldwide more than 40 companies produced aniline. The leading manufacturers were (capacity in 10^3 t):

ICI incl. joint ventures	460
Bayer	345
BASF incl. joint ventures	305
Mitsui	125
DuPont	118
First Chemical Corp.	131
Aristech	91
MCHZ	75
Anilina de Portugal	60

BASF was the sole company with production facilities in Europe, the United States, and Asia. Besides ICI and Bayer with sites in Europe and the United States all other companies were restricted to one region or one site only.

Expansions that have taken place since 1996 and formally announced projects taken into account, annual world capacity will grow to more than 3.50×10^6 t until 2005, with ICI and BASF leading the competition by a wide margin.

8. Derivatives

Table 3 shows substituted anilines, their properties, and their uses.

***o*-Chloroaniline** [*95-51-2*], 1-amino-2-chlorobenzene, $ClC_6H_4NH_2$, is a water-white to tan liquid, M_r 127.6, Tagliabue open cup (TOC) flash point 90 °C, *bp* 208 °C, *fp* −2.3 °C. *o*-Chloroaniline is *used* as an intermediate for rubber chemicals, pigments, pesticides, and dyes. *Production* is by low-pressure hydrogenation of 2-chloronitrobenzene with noble metal and/or noble metal sulfide catalysts at temperatures of ca. 50–100 °C in inert organic solvents. The yield is almost quantitative.

***m*-Chloroaniline** [*108-42-9*], 1-amino-3-chlorobenzene, $ClC_6H_4NH_2$, is a water-white to light-amber liquid, M_r 127.6, flash point (TOC) 90 °C, *bp* 230 °C, *fp* −10.5 °C. This compound is *used* as an intermediate for pesticides, pharmaceuticals, and dyes. It is *produced* by low-pressure hydrogenation of 3-chloronitrobenzene in the liquid phase with noble metal and/or noble metal sulfide catalysts. The addition of metal oxides to the reaction helps avoid dehalogenation. The yield is approximately 98 %.

***p*-Chloroaniline** [*106-47-8*], 1-amino-4-chlorobenzene, $ClC_6H_4NH_2$, is a white to light amber, crystalline solid at room temperature, M_r 127.6, *bp* 230 °C, *mp* 70 °C. It is *used* as an intermediate for pesticides, pharmaceuticals, pigments, and dyes. *Production* is similar to that of *m*-chloroaniline [31].

3,4-Dichloroaniline [*95-76-1*], 1-amino-3,4- dichlorobenzene, $Cl_2C_6H_3NH_2$, is a gray to dark-brown crystalline solid at room temperature, M_r 162.0, *fp* 66–71 °C, *bp* 272 °C. It is *used* as an intermediate for pesticides and dyes and is *produced* by catalytic hydrogenation of 3,4-dichloronitrobenzene with noble metal catalysts under pressure. Various types of additives prevent dehalogenation during production and reactors must be fabricated with special steel alloys to inhibit corrosion [31], [32].

***N,N*-Dimethylaniline** [*121-69-7*], *N,N*-dimethylphenylamine, $(H_3C)_2NC_6H_5$, is a light yellow (straw colored) to brown oily liquid, freezing to a crystalline solid at low temperatures, M_r 121.2, *fp* 2.1 °C, flash point, Tagliabue closed cup (TCC) 63 °C, *bp* 193 °C. It is *used* as a polymerization catalyst and as an intermediate for pharmaceuticals and dyes. Dimethylaniline is *produced* from aniline and methanol under pressure in the presence of acidic catalysts or by passing dimethyl ether and aniline vapor over

Table 3. Substituted anilines

Class and name	Formula	CAS registry number	M_r	Appearance [a]	*mp*, °C	*bp*, °C	d_4^{20} [b]	n_D^{20} [b]	Solubility [a]	Use [c]
Salts										
Aniline hydrochloride	$C_6H_5NH_2HCl$	[*142-04-1*]	129.60	w. cr.	198	245	1.2215^{4}		s. w, alc	
Aniline sulfate	$(C_6H_5NH_2)_2H_2SO_4$	[*542-16-5*]	284.34	w. cr.	dec.		1.377^{4}		s. w	
Chloro										
o-Chloroaniline	$ClC_6H_4NH_2$	[*95-51-2*]	127.58	c. liq.	−14	209	1.2125	1.5881	s. alc, eth, be, ace	(1)
m-Chloroaniline		[*108-42-9*]	127.58	c. liq.	−10	230	1.2161	1.5941	s. alc, eth, be	(1)
p-Chloroaniline		[*106-47-8*]	127.58	c. cr.	72.5	232	1.429^{19}	1.5546^{87}	s. hot w, alc, eth, ace	(1) (8)
2,3-Dichloroaniline	$Cl_2C_6H_3NH_2$	[*608-27-5*]	162.02	cr.	24	252	1.383^{25}		s. alc, eth, ace, be	
2,4-Dichloroaniline		[*554-00-7*]	162.02	cr.	63	245	1.567		s.s. w; s. alc, eth	
2,5-Dichloroaniline		[*95-82-9*]	162.02	cr.	50	251			s. alc, eth, be	(1)
2,6-Dichloroaniline		[*608-31-1*]	162.02	cr.	39				s. alc, eth	
3,4-Dichloroaniline		[*95-76-1*]	162.02	w. cr.	72	272	1.33^{80}		s. alc, eth	(1) (8)
3,5-Dichloroaniline		[*626-43-7*]	162.02	cr.	51	260			s. alc, eth, be	
2,4,5-Trichloroaniline	$Cl_3C_6H_2NH_2$	[*636-30-6*]	196.46	cr.	97	270			s. alc, eth	
Alkyl										
o-Toluidine	$H_3CC_6H_4NH_2$	[*95-53-4*]	107.17	y. liq.	−16.4	200	0.9994	1.5725	s.s. w; s. alc, eth	(1)
m-Toluidine		[*108-44-1*]	107.17	c. liq.	−30	203	0.9889	1.5681^{22}	s.s. w; s. alc, eth, ace, be	(1)
p-Toluidine		[*106-49-0*]	107.17	w. cr.	44–45	200.6	0.9619	1.5636	s.s. w; s. alc, eth, ace, be	(1)
2,3-Xylidine	$(H_3C)_2C_6H_3NH_2$	[*87-59-2*]	121.20	liq.	<−15	221–222	0.9931	1.5684	s. alc, eth	
2,4-Xylidine		[*95-68-1*]	121.20	liq.	−14	215	0.9723	1.5569	s. alc, eth, be	
2,5-Xylidine		[*95-78-3*]	121.20	oily liq.	16	214	0.9790^{21}	1.5593	s. eth	(1)
2,6-Xylidine		[*87-62-7*]	121.20	liq.	11	216	0.9842	1.5610	s. alc, eth	(1)
3,4-Xylidine		[*95-64-7*]	121.20	cr.	51	226	1.076^{18}		s. eth	(4)
3,5-Xylidine		[*108-69-0*]	121.20	oily liq.	9.8	220–221	0.9706	1.5581	s. eth	(1)
Alkoxy										
o-Anisidine	$H_3COC_6H_4NH_2$	[*90-04-0*]	123.17	y. liq.	6.2	225	1.0923	1.5793	s. alc, eth, ace, be; s.s. w	(1)
m-Anisidine		[*536-90-3*]	123.17	oily liq.	<−12	251	1.096	1.5794	s. alc, eth, ace, be; s.s. w	
p-Anisidine		[*104-94-9*]	123.17	w. cr.	57.2	243	1.0605^{67}	1.5559^{67}	s. alc, eth, ace, be; s.s. w	(1)
o-Phenetidine	$H_5C_2OC_6H_4NH_2$	[*94-70-2*]	137.20	oily liq.	<−21	232		1.5560	s. alc, eth,	(1) (2) (3)
p-Phenetidine		[*156-43-4*]	137.20	liq.	2–3	254	1.0652^{15}	1.5528	s. alc, eth; s.s. w	(1)
p-Cresidine	$H_3CO(CH_3)-C_6H_3NH_2$	[*120-71-8*]	137.20	w. cr.	52	235			s. alc, eth, ace, be; s.s. w.	(1)
N-Alkyl, *N*-Aryl										

Table 3. (Continued)

Class and name	Formula	CAS registry number	M_r	Appearance [a]	*mp*, °C	*bp*, °C	d_4^{20} [b]	n_D^{20} [b]	Solubility [a]	Use [c]
N-Methylaniline	$C_6H_5NHCH_3$	[*100-61-8*]	107.17	y. liq.	−57	196	0.9891	1.5684	s. alc; s.s. w	
N,*N*-Dimethylaniline	$C_6H_5N(CH_3)_2$	[*121-69-7*]	121.20	y. liq.	2	194	0.9557	1.5582	s. alc, eth, ace, be	(1)
N-Ethylaniline	$C_6H_5NHC_2H_5$	[*103-69-5*]	121.20	liq.	−63.5	205	0.9625	1.5559	s. alc, eth, ace, be	(1) (5)
N,*N*-Diethylaniline	$C_6H_5N(C_2H_5)_2$	[*91-66-7*]	149.26	y. liq.	−38.8	216	0.9351	1.5409	s.s. alc, eth, ace	(1)
N-Ethyl-*N*-benzylaniline	$C_6H_5N(C_2H_5)-CH_2C_6H_5$	[*92-59-1*]	211.20	y. liq.		314	1.034		s. alc, eth	(1)
N,*N*-Diphenylamine	$C_6H_5NHC_6H_5$	[*122-39-4*]	169.24	w. cr.	54–55	302	1.159		s. alc, eth, ace, be	(3)
N-Acyl										
Formanilide	C_6H_5NHCOH	[*103-70-8*]	121.15	w. cr.	50	271	1.1437^{17}		s. alc; s.s. w	(1) (6)
Acetanilide	$C_6H_5NHCOCH_3$	[*103-84-4*]	135.18	cr.	114	304	1.219^{15}		s. alc, eth, ace, be, hot w	(1) (3) (6)
Acetoacetanilide	$C_6H_5NHCOCH_2COCH_3$	[*102-01-2*]	177.22	w. cr.	86				s. alc, eth, ace, be	(1)
Nitro										
o-Nitroaniline	$O_2NC_6H_4NH_2$	[*88-74-4*]	138.14	y. cr.	71–72	284	1.442^{15}		s. alc, eth, ace, be, hot w	(1)
m-Nitroaniline		[*99-09-2*]	138.14	y. cr.	114	306			s. alc, eth, ace; s.s. w	(1)
p-Nitroaniline		[*100-01-6*]	138.14	y. cr.	148–149	332	1.424		s.s. alc	(1) (3)
2,4-Dinitroaniline	$(O_2N)_2C_6H_3NH_2$	[*97-02-9*]	183.14	y. cr.	187–188		1.615^{14}			(1)
2,4,6-Trinitroaniline	$(O_2N)_3C_6H_2NH_2$	[*489-98-5*]	228.12	y. cr.	192–195		1.762^{12}			(7)
Sulfonated										
Orthanilic acid (*o*)	$H_2NC_6H_4SO_3H$	[*88-21-1*]	173.9	cr.	>320					(1)
Methanilic acid (*m*)		[*121-47-1*]	173.9	w. cr.	dec.				s.s. alc, w	(1)
Sulfanilic acid (*p*)		[*121-57-3*]	173.9	w. cr.	288		1.485^{25}		s.s. w	(1) (6)

[a] c. = colorless; w. = white; y. = yellow; liq. = liquid; cr. = crystals; dec. = decomposes; s. = soluble; s.s. = slightly soluble; w = water; alc = alcohol; eth = ether; ace = acetone; be = benzene

[b] Density and refractive index superscripts refer to temperature (°C)

[c] (1) dyes, pigments; (2) synthetic sweetener; (3) rubber antioxidant, accelerator; (4) vitamin B_2 (riboflavin); (5) explosive stabilizer; (6) medication (analgesic), pharmaceuticals; (7) explosive; (8) agricultural chemicals

highly activated aluminum oxide at 230 – 295 °C [33].

***N,N*-Diethylaniline** [*91-66-7*], *N*-phenyldiethylamine, $(H_5C_2)_2NC_6H_5$, is a yellow to brown, oily liquid, M_r 149.2, flash point (TOC) 85 °C, *bp* 216 °C. *Uses* are similar to those for the dimethyl analogue. Its *production* is analogous to methods used for dimethylaniline production except that ethanol instead of methanol is reacted with aniline.

***N*-Phenylacetamide** [*103-84-4*], acetanilide, acetylaniline, $C_6H_5NHCOCH_3$, is a colorless, glossy, crystalline material, M_r 135.2, *fp* 114 °C, *bp* 304 °C. It is *used* as a dye intermediate, plasticizer and stabilizer for cellulose esters, and as an intermediate for the synthesis of sulfonamides. It is *produced* by heating aniline with acetic acid, acetic anhydride, or acetyl chloride [34].

2,4-Xylidine [*95-68-1*], *m*-xylidine, 2,4-dimethylaniline, $(H_3C)_2C_6H_3NH_2$, is a colorless to tan, oily liquid, M_r 121.2, flash point (TCC) > 93 °C, *bp* 213 °C. It is *used* as an intermediate for photographic chemicals, pesticides, and dyes. It is *produced* by the catalytic hydrogenation of the corresponding nitro derivative.

2,6-Xylidine [*87-62-7*], 2,6-dimethylaniline, $H_2NC_6H_3(CH_3)_2$, is a colorless to reddish-yellow, clear liquid, M_r 121.2, flash point (TCC) > 94 °C, *bp* 217 °C. This compound is *used* as an intermediate for dyes, pesticides, and pharmaceuticals. *Production* is by the catalytic hydrogenation of the corresponding nitro derivative [35].

***N*-Phenylaminobenzene** [*122-39-4*], diphenylamine, $C_6H_5NHC_6H_5$, is isolated in the form of colorless leaflets, M_r 212.2, flash point (TCC) 153 °C, *bp* 302 °C, *mp* 53 °C. Diphenylamine is *used* as a stabilizer for elastomers, nitrocellulose, and nitroglycerine and as an intermediate for dyes. It is *produced* by passing aniline over various types of acidic catalysts at elevated temperatures and pressures [36].

9. Toxicology

9.1. Toxicokinetic Properties and Metabolism

Aniline, a weakly alkaline liquid, is readily absorbed into the circulation after oral ingestion, inhalation and dermal contact. In human volunteers, more than 90 % of the inhaled aniline vapors (5 – 30 mg/m^3) were absorbed in the state of rest [37–39]. The percutaneous uptake from the vapor phase accounted for 25 – 30 % of the total incorporation in normally dressed individuals at 25 °C and 35 % relative air humidity (estimated absorption rate: 0.2 – 0.4 µg cm^{-2}h^{-1}), but increased by 21 and 29 % when the temperature was elevated by 5 °C and the humidity from 35 to 75 %, respectively [38]. Likewise, when applied as liquid to human skin from a drained gauze (concentration 10 mg/cm^2), skin absorption of aniline was between 0.2 and 0.7 mg cm^{-2}h^{-1} but could reach up to 3.5 mg cm^{-2}h^{-1} on highly moistened skin [38], [39], also temperature appeared to be a factor.

Aniline undergoes rapid oxidation, mainly in the liver, but also in other organs like the intestine and erythrocytes. Three primary transformation reactions compete with each other and are expressed to varying degree in different species and individuals:

1) *N*-Hydroxylation
2) (Ring) hydroxylation
3) *N*-Acetylation followed by *p*-(ring) hydroxylation

In secondary steps, the hydroxyl intermediates are rapidly conjugated, largely to sulfate and glucuronic acid and excreted, mainly in the urine [40–43]. In humans, the half-life of aniline is ca. 3.5 h [39].

The primary conversion products, mainly phenylhydroxylamine and *p*-aminophenol as well as their oxidized forms nitrosobenzene and *p*-iminoquinone, resulting from reactions 1 and 2, are regarded as toxification steps to biologically active compounds (see below), whereas *N*-acetylation may be considered as a detoxification step, which is followed by *p*-hydroxylation to *N*-acetyl-*p*-aminophenol. *N*-Acetyl transferase is congenitally expressed to varying extent in humans ("strong and weak

acetylators"; see below); this is a reason for different individual susceptibilities.

Certain metabolites, such as nitrosobenzene, are coupled to thiols, especially glutathione; the quantities of aniline-protein conjugates, especially aniline-Hb adducts in blood, are diagnostic tools for the estimation of aniline exposure and body burden [44], [45].

9.2. Toxicological Properties

Central Nervous System Effects. Aniline may produce an euphoric state through its action on the central nervous system. At very high exposure, a narcotic effect may result and also cardiac dysfunction, coma, convulsions, and respiratory paralysis.

Methemoglobinemia and Hemolysis. Methemoglobinemia and hemolysis are typical aniline-related effects; they are frequently observed also with many other related aromatic amines: Primary oxidation products (phenyl hydroxylamine, *p*-aminophenol) are regarded as intermediate active agents and show in the presence of oxygen the propensity to compete with redox cycles (for a review see [46]).

Cats and humans appear to be the most responsive species and more susceptible than rats and mice [47–49]. In healthy adult humans, ca. 25 mg aniline per individual, orally ingested for 3 consecutive days, was found to be a critical threshold dose which causes measurable elevation of metHb; the highest noneffective dose in rats was reported as 20 mg/kg [50]. Up to ca. 15 % of metHb are normally tolerated in humans; doses of 65 mg per person may transiently exceed the 15 % limit [50].

One sequel of metHb formation may be the destruction and hemolysis of erythrocytes, associated with the formation of "Heinz bodies" in red blood cells and a hemolytic anemia with secondary effects on other organs.

A deficiency in oxygen supply follows the depletion of normal Hb to metHb and causes, depending on the dose, hypoxemia, anoxia, bluish discoloration to pronounced cyanosis, headache, weakness, drowsiness, shortness of breath, and unconsciousness.

In one acute case, a massive suicidal dose of 60 mL orally, the "hemolytic crisis" caused death by organ failure of heart, liver, kidney, and lung [51].

Early reports say that 6-h exposures to 40 – 53 mL/m^3 of aniline vapors (150 – 200 mg/m^3) were tolerated by workers without noticeable symptoms [52], but in other cases slight discomfort or dysfunction were produced already at 7 – 26 mL/m^3 after several hours; 105 – 160 mL/m^3 were not tolerable for more than 1 h without the health status being severely disturbed [52].

The ratio of toxifying and detoxifying reactions has implications for the individual susceptibility: ca. 50 % of Europeans are congenitally deficient in hepatic N-acyl transferase: in these individuals, the formation of inactive acetanilide is delayed in favor of phenylhydroxylamine and nitrosobenzene. Even at the industrial limit concentration of 2 mL/m^3, the sensitive persons show higher mean levels of metHb (1.5 versus 0.9 %) and of aniline-Hb adducts than similarly exposed persons with a high acetylating capacity [52, p. 7].

In general, methemoglobinemia will reverse spontaneously when exposure ceases. Cyanosis, however, needs medical treatment with supportive measures, e.g., oxygen inhalation and bed rest. In most cases, normal body functions will be restored after 3 weeks without irreversible damage [53].

Genotoxicity/Carcinogenicity. Under specific activating test conditions, aniline may exhibit certain forms of genotoxicity: standard Ames assays are negative, unless the incubation mixture is coactivated with norharmane. In mammalian cell assays for DNA damage, gene mutation and clastogenicity, and also in vivo, varying results were obtained. However, aniline does not appear to be a potent genotoxic agent [52].

In the early years of industrial production of aniline-based dyes, an increased incidence of urinary bladder cancer was ascribed to aniline exposure [54], [46]. However, it was recognized later that contaminations of β-naphthylamine or benzidine probably were the active agents, aniline itself was mainly considered noncarcinogenic.

Most early experimental investigations failed to support carcinogenicity (for a detailed review on early studies: see [59]. However,

in more recent rodent long-term feeding studies with aniline HCl, the incidence of spleen sarcomas was significantly increased in rats (100 – 300 mg $kg^{-1}d^{-1}$); the lower levels (10 – 30 mg $kg^{-1}d^{-1}$) were still associated with hemato- and splenotoxicity and few spleen tumors [55–58]. It was therefore assumed that the production of spleen sarcomas was a secondary response, initiated by the hemolytic toxicity and associated with the decay and phagocytosis of damaged erythrocytes in spleen sinuses, including an iron overload in this target organ. This is supported by the fact that the spleen is a very unique target organ for carcinogenic effects even with strongly genotoxic compounds, whereas spleen enlargement, hemosiderosis (iron deposition) and splenic lipid peroxidation [47] are typically induced by aniline-related methemoglobinemia. Tumorigenesis of typical genotoxic aromatic amines (→ Amines, Aromatic, Chap. 5.1.) is usually seen at other tumor sites than spleen [52].

Studies of individuals exposed to aniline and not to other aromatic amines have shown little evidence of increased tumor risk: among 1223 men producing or using aniline one bladder cancer was recorded vs. 0.83 expected from the general population incidence [59].

Reproduction Toxicity. Although aniline can pass the placental barrier, studies in rats and mice failed to provide evidence for the induction of malformations in the fetus. Slight embryo retardations, fetal hematological effects, and increased liver weights were recorded at the highest dose of 100 mg kg^{-1} d^{-1}. Also maternal effects occurred at that level. The no observed effect level was 10 mg kg^{-1} d^{-1} [60–62].

10. Occupational Health

Based on absorption properties (see Chap. 9), a TLV and a MAK value (maximal concentration at the workplace) of 2 ppm (8 mg/m^3) with a skin notation (potential of toxicologically relevant cutaneous absorption) were established in the United States and Germany, respectively [63], [52].

In Germany, because of the limited data base for embryotoxicity, the MAK Commission has allocated aniline to group D (no final conclusion on the prenatal risk under MAK value conditions) [52].

11. References

1. "Aniline" BASF Safety Data Sheet (91/155/EWG).
2. D. Barton, W. D. Ollis: *Comprehensive Organic Chemistry,* 1st ed., **vol. 2,** Pergamon Press, Oxford 1979, pp. 131 – 184.
3. S. Patai: *The Chemistry of the Amino Group,* 1st ed., Interscience Publ., London 1968, pp. 161 – 205.
4. DuPont, US 3 832 364, 1972.
5. DuPont, US 4 001 260, 1973.
6. *Kirk-Othmer,* 4th ed., **vol. 17,** pp. 133 – 152.
7. A. Guenkel, T. Maloney, *ACS Symp. Ser.* **623** (1996), pp. 223 – 233.
8. *Kirk-Othmer,* 4th ed., **vol. 7,** pp. 730 – 736.
9. *Winnacker-Küchler Chemische Technologie,* 4th ed., **vol. 6,** pp. 205 – 214.
10. D. N. Rihani, T. K. Narayanan, L. K. Doraiswamy, *I&EC Process. Des. Develop.* **4** (1965) 403 – 410.
11. Lonza/First Chemical Co., *Hydrocarbon Process* **59** (Nov. 1979) no. 11, 136.
12. Lonza, US 3 636 152, 1969.
13. Bayer, EP 011 090, 1979.
14. First Chem. Corp., US 4 740 621, 1986.
15. Bayer, US 5 304 525, 1990.
16. Bayer, EP 696 574, 1995; Bayer, EP 696 573, 1995.
17. Bayer, EP 748 790, 1996; Bayer, EP 748 789, 1996.
18. Chemopetrol, GB 1 452 466, 1975.
19. American Cyanamide, US 28 910 994, 1955.
20. BASF, US 3136818, 1961.
21. BASF, DE 4110457, 1991.
22. ICI, US 3270057, 1964.
23. DuPont, US 3 499 034, 1966.
24. DuPont, US 4 185 036, 1977.
25. *Winnacker-Küchler Chemische Technologie,* 4th ed., **vol. 4,** pp. 170 – 171.
26. F. A. Lowenheim, M. K. Moran: *Industrial Chemicals,* 4th ed., J. Wiley & Sons, New York 1975, pp. 109 – 111.
27. Halcon International, Inc., US 3272865, 1966 (R. S. Barker).
28. Halcon International, Inc., US 3860650, 1975 (M. Becker, S. Khoobiar).
29. Scientific Design, *Hydrocarbon Process.* **59** (1979) 137.

30. J. A. Kent (ed.): *Riegel's Handbook of Industrial Chemistry,* 9th ed., van Nostrand Reinhold, New York 1992, p. 1136.
31. DuPont, US 3 361 819, 1968 (J. R. Kosak, L. Spiegler).
32. Dow, US 3 067 253, 1962 (A. J. Dietzler, T. R. Kell).
33. Biller, Michaelis, US 2991311, 1961 (M. Thoma).
34. British Industrial Solvents, US 2 462 221, 1949 (E. S. Pemberton).
35. Standard Oil Development, US 2 421 608, 1947; US 2481245, 1949.
36. Amer. Cyanamid, US 2 968 676, 1961 (A. G. Potter, R. G. Weyker).
37. T. Dutkiewicz, *Med. Pracy* **12** (1961) 1.
38. T. Dutkiewicz, J. Piotrowski, *Pure Appl. Chem.* **3** (1961) 319.
39. J. Piotrowski, *Pracov. Lék.* **24** (1972) 94.
40. J. Piotrowski, *J. Hyg. Epidem.* **1** (1957) 23.
41. J. Kao, J. Faulkner, J. W. Bridges, *Drug Metab. Dispos.* **6** (1978) 549.
42. J. N. Smith, R. T. Williams, *Biochem. J.* **44** (1949) 242.
43. D. V. Parke, *Biochem. J.* **77** (1960) 493.
44. W. Albrecht, H.-G. Neumann, *Arch. Toxicol.* **57** (1985) 1.
45. G. Birner, H.-G. Neumann, *Arch. Toxicol.* **62** (1988) 110.
46. M. Kiese: *Methemoglobinemia: A Comprehensive Treatise,* CRC Press, Inc., Cleveland 1974.
47. M. F. Khan et al., *Toxicol. Lett.* **92** (1997) 31 – 37.
48. D. Lester, *J. Pharmacol. Exp. Ther.* **77** (1943) 154.
49. Y. Henderson, H. W. Haggard, cited in: F. Flury, F. Zernik: *Schaedliche Gase*, Springer-Verlag, Berlin 1931.
50. F. P. Jenkins et al., *Fd. Cosmet. Toxicol.* **10** (1972) 671.
51. S. Janik-Kurylcio et al., *Pol. Tyg. Lék.* **28** (1973) 1241.
52. Deutsche Forschungsgemeinschaft (DFG) (ed.): "Anilin", in: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-Arbeitsmedizinische Begründung von MAK-Werten,* VCH Verlagsgesellschaft, Weinheim 1992.
53. R. Hopmann, H. Bürschaper, *Arch. Toxicol.* **15** (1955) 340.
54. M. Mackenzie, *Med. Tms. (London)* **1** (1862) 239.
55. Chem. Ind. Inst. Toxicol. (CIIT): "Week Chronic Toxicity Study in Rats: Aniline Hydrochloride," *Final Rep. 104*-Res. Triangle Park, NC/USA (1982).
56. Nat. Cancer Inst. (NCI): *NCI-Bioassay Report No. 130*, US-DHEW, Bethesda, MD/USA (1978).
57. D. G. Goodman, J. M. Ward, W. D. Reichardt, "Splenic Fibrosis and Sarcomas in F344 Rats Fed Diets Containing Aniline Hydrochloride p-Chloroaniline, Azobenzene, o-Toluidine Hydrochloride, 4,4′-Sulfonyldianiline, or D & C Red No. 9,"*JNCI.* **75** (1984) no. 1.
58. M. A. Weinberger, R. H. Albert, St. B. Montgomery, "Splenotoxicity Associated With Splenic Sarcomas in Rats Fed High Doses of D & C Red No. 9 or Aniline Hydrochloride," *JNCI.* **75** (1985) no. 4.
59. IARC: "Aniline", in: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* **vol. 27,** Intern. Agency for Res. on Cancer, Lyon 1982.
60. C. J. Price et al., *Toxicol. Appl. Pharmacol.* **77** (1985) 465.
61. V. J. Piccirillo et al., NTIS, PB83-257600, Borriston Laboratories, Inc., MD/USA (1983)
62. B. D. Hardin et al., *Teratog. Carcinog. Mutag.* **7** (1987) 29.
63. Am. Conf. of Governmental Industrial Hygienists (ACGIH, ed.): *List of Threshold Limit Values (TLV),* Cincinnati, Ohio 1997.

Anilinesulfonic Acids → Benzenesulfonic Acids and Their Derivatives
Animal Feeds → Foods, 1. Survey

Anti-Obecity Drugs

WILLIAM J. HOULIHAN, Sandoz Research Institute, East Hanover, New Jersey 07936, United States

1. Introduction 1
2. Animal and Human Test Procedures . 1
3. Mechanism of Action 1
4. Clinical Side Effects 2
5. Phenethylamine Derivatives 2
5.1. β-Phenylisopropylamines 3
5.2. β-Phenyl-*tert*-butylamines 4
5.3. β-Phenyl-β-oxyisopropylamines . . . 4
5.4. 1-Methyl-2-phenylmorpholines 5
6. Other Compounds 6
7. Economic Aspects 7
8. References 7

1. Introduction

Anorexiants are substances that depress the desire for food. They are also referred to as anorectic agents, anorexigenics, and appetite suppressants. Because the main therapeutic use of these substances is to control weight, they are also called antiobesity agents. This term, however, is too broad when only the suppression of the desire for food is meant, for antiobesity drugs include not only anorexiants but also metabolism and biosynthesis modifiers [1].

The control of weight or obesity by drug therapy rather than dieting alone remains controversial because of the potential health risks and the limited efficacy of most of the available drugs. However, because many people find dieting difficult and seek some supportive measure, the use of drugs in a reducing program as an adjunct to diet modification and exercise is well established. A report on obesity in the United Kingdom [2] documents the adverse effects of obesity when coupled with such medical problems as coronary heart disease, hypertension, diabetes, gall bladder disease, respiratory disease, varicose veins, and certain forms of cancer.

2. Animal and Human Test Procedures

The main techniques used to search for anorexiants in animal models are the free feeding and operant behavioral methods [3]. Both measure the effectiveness of a compound at reducing food consumption in animals made hungry by dietary restrictions.

In the *free feeding* procedure the food intake is measured over a period of time. The animals then are given a drug and the effect on food intake is determined. Agents that reduce intake are potential anorexiants. Because this technique cannot distinguish true anorectic effects from conflicting pharmacologic properties, more refined models involving measurement of *operant behavior* have been designed. Animals are trained to respond to a food presentation by lever pressing, runway performance, and various shock schedules [4]. These methods have verified anorectic efficacy in a number of phenethylamine derivatives but are less successful in defining unique compounds with confirmed activity in humans [5].

Human testing of potential anorexiants involves measuring weight loss after drug therapy has been initiated. In addition the measurement of the thickness of the underarm skin-fold and subjective reports from the patient regarding desire for food also are used frequently. Despite many clinical references on anorectic agents, a standard testing procedure has not yet been designed [6].

3. Mechanism of Action

The regions of the brain that are involved in central control of appetite are thought to be mainly in the hypothalamus. Control of food intake is considered to involve a feeding center and a satiety center. Electrical stimulation of the ventrolateral hypothalamus causes feeding in sated animals; whereas stimulation of the ventromedial hypothalamus inhibits ongoing feeding [7], [8].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_313

At the molecular level a multiplicity of neurotransmitters are involved in the control of appetite. These substances can be classified broadly as neuropeptides and monoamines, with the latter class being investigated the most. The satiety and feeding centers appear to be controlled mainly by norepinephrine [*51-41-2*] and serotonin [*50-67-9*].

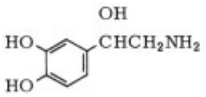

Norepinephrine

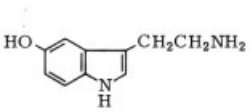

Serotonin

Commercial anorectic agents all appear to operate through a mechanism that involves the control of reuptake or release of these monoamines from the appropriate area of the hypothalamus [8], [9].

4. Clinical Side Effects

The drugs used to control appetite are taken orally one to three times daily, usually before meals. Medical practice in the United States calls for a maximum usage of twelve weeks; the majority of drugs appear to lose effectiveness after this time [10]. Because the neurotransmitters that control appetite also are involved in other biochemical pathways, drugs that affect them give rise to a variety of side effects. These side effects can be grouped into four major categories:

– Central nervous system (CNS)
– Cardiovascular
– Gastrointestinal
– Abuse and/or dependence

All the anorectic drugs except fenfluramine excite the *central nervous system* and give rise to restlessness, insomnia, tremor, euphoria, and headache. *Cardiovascular problems* include elevated blood pressure, palpitations, and rapid heart action (tachycardia). *Gastrointestinal irritation* can cause nausea, vomiting, dry mouth, unpleasant taste, diarrhea, and constipation.

Abuse and dependence of a physiologic or psychological origin are a particular problem with those compounds closely related to amphetamine. Symptoms of chronic abuse include prolonged insomnia, severe skin diseases, baldness, and dramatic personality changes resulting in schizophrenia-like behavior. Withdrawal by addicted patients is characterized by profound fatigue and mental depression. The seriousness of the abuse and dependency with these drugs has led various United States regulatory agencies to control them under the Bureau of Narcotics and Dangerous Drugs (BNDD). To assist the physician the drugs have been classified under BNDD Schedules I – IV, with Schedule IV substances being the least prone to cause abuse and dependency problems [11].

5. Phenethylamine Derivatives

The β-phenethylamine framework is found in all but one of the clinically useful anorectic drugs. This same unit is also found in the neurotransmitter

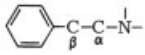

norepinephrine (see Chap. 3). Modification of the molecule by introducing one or two methyl groups in the α position has led to the β-phenylisopropylamines (Section 5.1) and β-phenyl-*tert*-butylamines (Section 5.2). Additional modification at the β position with a hydroxyl or ketone group gives the β-phenyl-β-oxyisopropylamines (Section 5.3), and cyclization of the oxygen and nitrogen atoms by a two-carbon bridge results in the 1-methyl-2-phenylmorpholines (Section 5.4). The (+) isomer of those compounds containing a chiral α carbon atom is the more active anorectic agent. Extensive studies have demonstrated that a Cl or CF_3 group in the ring or a group on the nitrogen atom that can be metabolized to the NH_2 derivative gives rise to clinically useful drugs [12].

The first phenethylamine derivative to be used as an anorexiant was ($\pm$)-amphetamine [13]. Subsequently, studies showed that the (+) isomer, dextroamphetamine, has fewer side effects and it has become used more widely. Combinations of phenethylamines with sedatives and tranquilizers have been prescribed in an attempt to alleviate the undesirable stimulant effects. Slow-release and resin-bound forms also have been marketed to overcome the same problem.

5.1. β-Phenylisopropylamines

Drugs containing β-phenylisopropylamine or its structural framework constitute the largest group of anorexiants on the market.

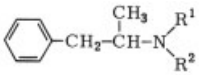

The general procedure for preparing these compounds involves reaction of a phenylacetone with hydroxylamine followed by catalytic hydrogenation, or reaction of the ketone with formamide (Leuckart-Wallach reaction) and subsequent acid hydrolysis (Eq. 1).

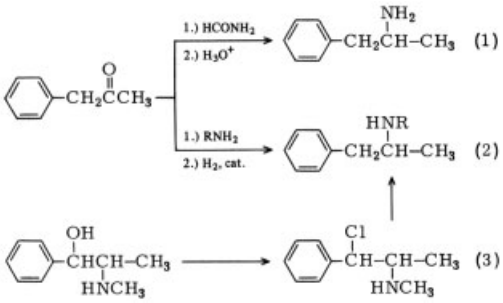

The *N*-alkyl derivatives are made by reductive aminoalkylation of phenylacetone by catalytic hydrogenation (Eq. 2). The optically active forms can be obtained by resolution with an optically active acid or by starting with (+)- or (-)-ephedrine, conversion to a chloroamine, and then catalytic hydrogenation (Eq. 3).

Amphetamine [*300-62-9*], benzedrine, (±)-α-methylbenzeneethanamine, $R^1 = R^2 = H$, $C_9H_{13}N$, M_r 135.21, is a colorless liquid, *bp* 200 – 201 °C; hydrochloride, *mp* 148 – 149 °C.

Amphetamine is prepared by treating phenylacetone with formic acid and ammonia under the Leuckart-Wallach conditions (Eq. 1) [14].

Trade names: Benzedrine (SKF), Obetrol-10 and Obetrol-20 (combined with dextroamphetamine; Schein), Biphetamine (resin combined with dextroamphetamine, Pennwalt), all United States; Adiparthrol (Medial, Switzerland).

Recommended adult dosage is 10 – 30 mg/d; BNDD Schedule II (United States).

Dextroamphetamine [*51-64-9*], (+)-α-methylbenzeneethanamine, $R^1 = R^2 = H$, $C_9H_{13}N$, M_r 135.21, is a colorless liquid, *bp* 80 °C (1.6 kPa), $[\alpha]_D^{15}$ +35.6° (neat); 0.5 sulfate, *mp* > 300 °C, $[\alpha]_D^{20}$ +21.8 to 24° (*c* = 5 g/100 mL H_2O).

The compound can be prepared by optical resolution of amphetamine (see above) or directly from (-)-norephedrine by conversion to the (+)-chloroamine followed by catalytic hydrogenation (Eq. 3) [14].

Trade names: Dexedrine (SKF, United States, Federal Republic of Germany, Switzerland), Obetrol-10 and Obetrol-20 (combined with amphetamine; Schein, United States).

Recommended adult dosage is 10 – 30 mg/d; BNDD Schedule II (United States). Dextroamphetamine also is approved for use in narcolepsy and for control of hyperkinetic children [15].

Methamphetamine [*537-46-2*], desoxyephedrine, *S*-(+)-*N*, α-dimethlybenzeneethanamine, $R^1 = H$, $R^2 = CH_3$, $C_{10}H_{15}N$, M_r 149.24; hydrochloride, *mp* 171 – 175 °C, $[\alpha]_D^{20}$ +16.0 to 19.2° (*c* = 5 g/100 mL H_2O) is prepared from phenylacetone and methlyamine (Eq. 2) or from (-)-ephedrine (Eq. 3) [16].

Trade names: Desoxyn (Abbott), Obedrin (Beecham-Massengill), both United States; Pervitin (Temmler-Werke, Federal Republic of Germany).

Recommended adult dosage is 7.5 – 15 mg/d; BNDD Schedule II (United States).

Benzphetamine [*156-08-1*], (+)-*N*,α-dimethyl-*N*-(phenylmethyl)benzeneethanamine, $R^1 = CH_3$, $R^2 = C_6H_5CH_2$, $C_{17}H_{21}N$, has M_r 239.35, *bp* 127 °C (3.5 Pa); hydrochloride, *mp* 129 – 130 °C, $[\alpha]_D^{20}$ +22 to 26° (*c* = 5 mg/100 mL H_2O).

Benzphetamine can be prepared by treating methamphetamine (see above) with a benzylhalide in the presence of a base [17].

Trade names: Didrex (Upjohn, United States, United Kingdom), Inapetyl (Upjohn, France).

Recommended adult dosage is 25 – 150 mg/d; BNDD Schedule III (United States).

Fenfluramine [*458-24-2*], *N*-ethyl-α-methyl-3-(trifluoromethyl)benzeneethanamine, $R^1 = H$, $R^2 = C_2H_5$, 3-CF_3, $C_{12}H_{16}F_3N$, has

M_r 231.27, *bp* 108 °C (1.7 kPa); hydrochloride, *mp* 166 °C.

Fenfluramine is prepared by reductive alkylation of norfenfluramine with acetaldehyde [18]. The nor compound is obtained by catalytic hydrogenation of the oxime made from 3-trifluoromethylphenyl acetone (Eq. 2).

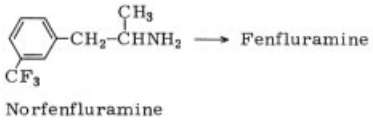

Trade names: Pondomin (Robbins, United States), Ponderal (Servier, France), Ponderax (Boehringer Ingelheim, Federal Republic of Germany).

Recommended adult dosage is 60 mg/d; BNDD Schedule IV (United States).

Fenfluramine is the only marketed anorectic agent that exerts its appetite depressant effect via a serotoninrgic mechanism (see Chap. 3). Its major metabolite, norfenfluramine, is believed to be the main active form in humans [19].

Propylhexedrine [*101-40-6*], cyclexedrine, *N,α*dimethylcyclohexaneethanamine, $C_{10}H_{21}N$, has M_r 155.28, *bp* 92 – 93 °C (2.7 kPa); (-)-hydrochloride, *mp* 138 – 140 °C, $[\alpha]_D^{20}$ −14.2° (H_2O); (±)-hydrochloride, *mp* 127 – 128 °C.

Propylhexedrine is prepared by catalytic hydrogenation of (±)- or (-)-ephedrine in a strong mineral acid [20].

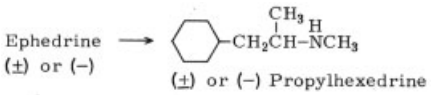

Trade names: Eventin [(-) isomer, Chemische Werke Minden]; Eggobasin, [(±) isomer, Fahlberg-List]; both Federal Republic of Germany.

The compound is marketed in the United States as an inhalant for nasal congestion and not as an anorectic agent.

5.2. β-Phenyl-*tert*-butylamines

The β-phenyl-*tert*-butylamines appear to offer no advantages over the β-phenylisopropylamines (Section 5.1). These *tert*-butylamines are known to induce pulmonary hypertension and their use in medical practice is declining.

Phentermine [*122-09-8*], α,α-dimethylbenzeneethanamine, $C_{10}H_{15}N$, M_r 149.23, is a colorless liquid, *bp* 100 °C (0.36 kPa); hydrochloride, *mp* 198 °C.

Phentermine is prepared by converting the condensation product of benzaldehyde and nitropropane to the corresponding aminoalcohol followed by transformation to the chloroamine and then catalytic hydrogenation (as in Eq. 3) [21].

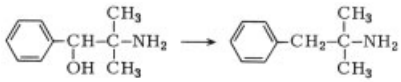

Trade names: Ionamin (resin; Pennwalt), Fastin (Beecham), Adipex (Lemmon), Oby-Trim (Obetrol), Teramine (Legere), Phentermine (Schein, Rugby), all United States.

Recommended adult dosage is 20 – 40 mg/d; BNDD Schedule IV (United States).

Phentermine has not been marketed in the Federal Republic of Germany since 1968 because of suspected pulmonary hypertension.

Phenpentermine [*434-43-5*], pentorex, α, α,β-trimethylbenzeneethanamine, $C_{11}H_{17}N$, has M_r 161.35; hydrochloride, *mp* 164 – 166 °C; hydrogen tartrate [*22876-60-4*], *mp* 160 – 162 °C, $[\alpha]_D^{20}$ +13.4° (c = 0.8 g/100 mL H_2O).

Phenpentermine is prepared by *N*-formylating the alcohol obtained from 3-phenyl-2-butanone and methylmagnesium bromide; the intermediate is hydrolyzed in acid [22].

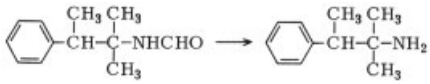

Trade name: Modatrop (Nordmark Werke, Federal Republic of Germany).

5.3. β-Phenyl-β-oxyisopropylamines

These drugs are the safest and most widely used anorexiants on the market.

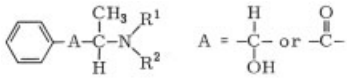

Ephedrine [*299-42-3*], (-) isomer, [*R*-(*R*∗, *S*∗)]-α-[1-(methlyamino)ethyl]benzenemethanol, A = HCOH, R^1 = H, R^2 = CH_3, $C_{10}H_{15}NO$, has M_r 165.23, *mp* 33 – 39 °C; hydrochloride [*50-98-6*], *mp* 217 – 220 °C, $[\alpha]_D^{20}$ −34.0 to −35.0 °(c = 5 g/100 mL H_2O).

(-)-Ephedrine is obtained by the catalytic hydrogenation of (-)-1-phenyl-2-methlyaminopropan-1-one. The latter substance is prepared by the optical resolution of the (±)-ketoamine [23].

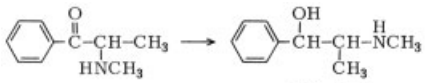

Ephedrine

Trade name: Alfabet (Zeppenfeldt, Federal Republic of Germany).

Norephedrine [*14838-15-4*], phenylpropanolamine (PPA), (*R*∗,*S*∗)-(±)-α-(1-aminoethyl)benzenemethanol, A = HCOH, R^1 = R^2 = H, $C_9H_{13}NO$, has M_r 151.18; hydrochloride [*154-41-6*], *mp* 190 – 194 °C.

Phenylpropanolamine is prepared by converting propiophenone to the α-oximino derivative with an alkyl nitrite, followed by catalytic hydrogenation, using a Pd or Pt catalyst [24].

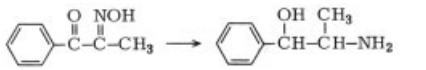

Phenylpropanolamine

Phenylpropanolamine is the only FDA approved (1978) appetite suppressant that is sold without a prescription on the United States market. In 1983, there were 35 brands of PPA available as a single drug entity (100 – 150 mg) or mixed with the central nervous system stimulant caffeine (50 – 200 mg), with the local anesthetic benzocaine, or with such bulk products as methyl or carboxymethlycellulose. Because PPA is so widely available, it has been abused, necessitating tighter controls on availability [25]. Phenylpropanolamine also has been used medically since the 1950s, primarily as a nasal decongestant and ingredient in cold remedies.

Norpseudoephedrine [*492-39-7*], [*36393-56-3*], cathine, (+) isomer, [*R*-(*R*∗,*R*∗)]-α-(1-aminoethyl)benzenemethanol, A = HCOH, R^1 = R^2 = H, $C_9H_{13}NO$, has M_r 151.18, *mp* 77 – 79 °C; hydrochloride *mp* 180 – 182 °C, $[\alpha]_D^{20}$ +42.5 to 44.0 ° (c = 5 mg/100 mL H_2O).

(+)-Norpseudoephedrine can be obtained by resolving the (±) form with d-mandelic acid or by isomerization of (-)-norephedrine [26].

Trade names: Adiposetten (Reiss), Boxogetten (Boxberger), both Federal Republic of Germany; Minusin (Doetsch-Grether, Switzerland).

Diethylpropion [*90-84-6*], amfepranon, α-(diethylamino)-phenyl-1-propanone, A = CO, R^1 = R^2 = C_2H_5, $C_{13}H_{19}NO$, has M_r 205.30, *bp* 140 °C (0.1 kPa); hydrochloride [*134-80-5*], *mp* 172 – 176 °C.

Diethylpropion is prepared by α-bromination of propiophenone followed by reaction with diethylamine [27].

Diethylpropion is the anorectic prescription drug of choice in the United States. It possesses central nervous system stimulant effects that are considerably weaker than those of amphetamine and its close analogs. Drug abuse and dependence is also low with diethylpropion.

Trade names: Tenuate (Merrell Dow, United States, Federal Republic of Germany, United Kingdom), Regenon (Temmler Werke, Federal Republic of Germany), diethylpropion HCl (Schein, Rugby, United States), Tepanil (Riker, United States).

Recommended adult dosage is 75 mg/d; BNDD Schedule IV (United States).

5.4. 1-Methyl-2-phenylmorpholines

The clinical side effects and efficacy of the 1-methyl-2-phenylmorpholines are similar to those of the amphetamines, and the probability of drug abuse is quite high.

All of these anorectics exist as *threo* isomers.

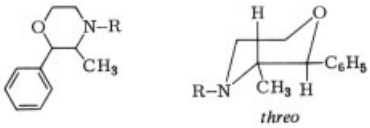

Phenmetrazine [*134-49-6*], *trans*-3-methyl-2-phenylmorpholine, R = H, $C_{11}H_{15}NO$, has M_r 177.50, *bp* 138 – 140 °C (0.04 kPa); hydrochloride [*1707-14-8*], *mp* 182 °C.

Phenmetrazine is prepared by reacting α-bromopropiophenone with *N*-benzylethanolamine, catalytic reduction of the intermediate ketone to *N*-hydroxyethyl norephedrine or norpseudoephedrine, and then cyclization with concentrated sulfuric acid [28].

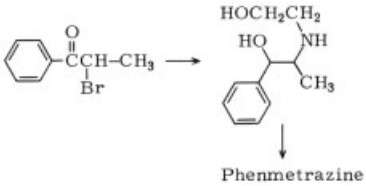

Trade names: Preludin (Boehringer-Ingelheim, United States), Cafilon [in combination with 8-chlorotheophyllinate and phenbutrazate (see below), Ravensberg, Federal Republic of Germany].

Recommended adult dosage is 50 – 75 mg/d; BNDD Schedule II (United States).

Phendimetrazine [*634-03-7*], *L*-(+)−3*R*-*trans*-3,4-dimethyl-2-phenylmorpholine, R = CH_3, $C_{12}H_{17}NO$, has M_r 191.26, *bp* 134 – 135 °C (0.21 kPa); hydrochloride, *mp* 208 °C, $[\alpha]_D^{20}$ + 35.7 ° (H_2O); hydrogen tartrate [*50-58-8*].

The compound can be prepared by the same procedure used to prepare phenmetrazineexcept that *N*-methlyethanolamine is reacted with α-bromopropiophenone [29].

Trade names: Plegine (Ayerst), Prelu-2 (Boehringer-Ingelheim), Bontril (Carnrick), Melfiat (Reid-Provident), all United States.

Recommended adult dosage is 30 mg/d; BNDD Schedule III (United States).

Phenbutrazate [*4378-36-3*], α-ethyl-2-(3-methyl-2-phenyl-4-morpholinyl)ethyl ester benzeneacetic acid, R = $CH_2CH_2OCOCH(C_2H_5)C_6H_5$, $C_{23}H_{29}NO_3$, has M_r 367.47; hydrochloride [*6474-85-7*], *mp* 154 °C.

Phenbutrazate is prepared by condensing *N*-hydroxyethylphenmetrazine (R = CH_2CH_2OH) with 2-phenylbutyric acid. The *N*-hydroxyethyl compound is obtained in the same manner as phendimetrazine [30].

Trade name: Cafilon [in combination with 8-chlorotheophyllinate and phenmetrazine (see above), Ravensberg, Federal Republic of Germany].

6. Other Compounds

A variety of compounds that do not possess the phenethylamine framework has been studied as anorectic agents; to date, however, only mazindol has reached the marketplace [12].

Mazindol [*22232-71-9*], 5-(4-chlorophenyl)-2,5-dihydro-3*H*-imidazo[2,1-*a*]isoindol-5-ol, $C_{16}H_{13}ClN_2O$, has M_r 284.74, *mp* 198 – 199 °C.

Mazindol is prepared by treating the dilithio reagent, obtained from 2-phenylimidazoline, with methyl 4-chlorobenzoate [31].

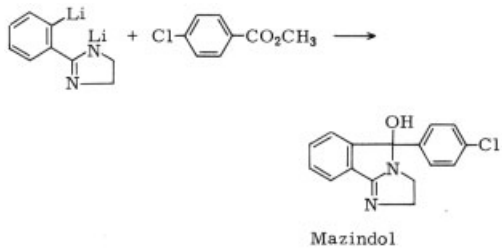

Although mazindol is unrelated structurally to the phenethylamine anorectic agents, it possesses a similar medical profile in humans. The most common side effects observed are dry mouth, tachycardia, constipation, nervousness, and insomnia. The mechanism responsible for the anorectic effects of mazindol differs from that for amphetamine and probably involves effects on both norepinephrine and serotonin [32].

Trade names: Sanorex (Sandoz), Mazanor (Wyeth), both United States; Teronac (Wander, Federal Republic of Germany).

Recommended adult dosage is 2 – 3 mg/d; BNDD Schedule IV (United States).

Table 1. Sales of prescription anorexiants on the United States market between October 1983 and October 1984

Anorexiant (BNDD schedule)	Trade name(s)	Sales, $\$ \times 10^3$	% of Market
Phentermine (IV)	Ionamin, Fastin, Adipex-P, generic	27868	36.7
Diethylpropion (IV)	Tenuate, Tepanil, generic	17970	23.7
Benzphetamine (III)	Didrex	7282	9.6
Phendimetrazine (III)	Prelu-2, Melfiat, Plegine, Bontril, generic	4899	6.5
Phenmetrazine (III)	Preludin	3977	5.3
Methamphetamine (II)	Desoxyn	3681	4.9
Mazindol (IV)	Sanorex, Mazanor	2928	3.9
Dextroamphetamine and amphetamine (II)	Biphetamine	2919	3.9
Dextroamphetamine (II)	Dexedrine	2858	3.9
Fenfluramine (IV)	Pondimin	393	0.4

7. Economic Aspects

The total sales of prescription anorexiants in the United States between October 1983 and October 1984 were $ 76 188 000. Products containing anorexiants classified as BNDD Schedule III or IV accounted for 82.3 % of the sales, while the more restricted drugs of Schedule II totaled 17.7 %. Table 1 gives United States sales figures for individual prescription anorexiants [33].

8. References

1. A. C. Sullivan, H. W. Baruth, L. Cheng, *Ann. Rep. Med. Chem.* **15** (1980) 1972.
2. "Obesity: A Report of the Royal College of Physicians," *J. R. Coll. Physicians London* **17** (1983) 6 – 65.
3. G. A. Heise: "Animal Techniques for Evaluating Anorexigenic Agents," in J. H. Nodine, P. E. Siegler (eds.): *Pharmacologic Techniques in Drug Evaluation,* Year Book Medical Publishers, Chicago 1964, pp. 279 – 282.
4. R. T. Kelleher, L. Cook, *J. Exp. Anal. Behav.* **2** (1959) 267.
5. R. P. Maickel, J. E. Zabik, *Life Sci.* **21** (1977) 173.
6. H. A. Levy, G. Faludi: "Clinical Techniques for Evaluating Anorexigenic Agents," in J. H. Nodine, P. E. Siegler (eds.): *Pharmacologic Techniques in Drug Evaluation*, Year Book Medical Publishers, Chicago 1964, pp. 283 – 290.
7. S. Garattini, R. Samanin (eds.): *Anorectic Agents: Mechanism of Action and Tolerance,* Raven Press, New York 1981.
8. J. A. Morley, A. S. Levine, *Lancet* 1983, 398.
9. S. Garattini, R. Samanin (eds.): *Central Mechanisms of Anorectic Drugs,* Raven Press, New York 1978.
10. A. C. Sullivan, K. Comai, *Int. J. Obes.* **2** (1978) 167.
11. J. Caldwell (ed.): *Amphetamines and Related Stimulants: Chemical, Biological and Sociological Aspects,* CRC Press, Boca Raton, Fla., 1980.
12. W. J. Houlihan, R. G. Babington: "Anorexigenics," in M. E. Wolff (ed.): *Burger's Medicinal Chemistry,* Part 3, J. Wiley & Sons, New York 1981, pp. 1069 – 1108.
13. M. F. Lesses, A. Meyerson, *N. Engl. J. Med.* **108** (1938) 528. M. H. Nathanson, *J. Am. Med. Assoc.* **218** (1938) 119.
14. W. B. Hartung, J. C. Munch, *J. Am. Chem. Soc.* **53** (1931) 1875.
15. *Physicians Desk Reference,* 37th ed., Medical Economics Co., Litton Industries, Oradell, N. J., 1983, p. 1898.
16. H. Emde, *Helv. Chim. Acta* **12** (1929) 365.
17. Upjohn Co., US 2789138, 1957 (R. V. Heinzelman, B. D. Aspergren).
18. Science Union et Cie., US 3 198383, 1965 (L. G. Beregi et al.).
19. R. M. Pinder, R. N. Grogden, P. R. Sawyer, T. M. Speight, G. S. Avery, *Drugs* **10** (1975) 241.
20. Knoll, DE 949657, 1956 (H. Metzger).
21. Wyeth, US 2590079, 1952 (L. L. Abell, W. F. Bruce, J. Seifter).
22. Nordmark-Werke, FR 2594M, 1962.
23. H. Takamatsu, *Yakugaku Zasshi* **76** (1956) 1221.
24. W. Hartung et al., *J. Am. Chem. Soc.* **74** (1952) 5927.
25. B. Calesnick, *Am. Fam. Physician* **26** (1982) 206.
26. C. I. Jarowski, W. H. Hartung, *J. Org. Chem.* **8** (1943) 564.

27. Temmler-Werke, US 3001910, 1958 (F. Keil, W. Dobke).
28. Boehringer Ingelheim, US 2835669, 1958 (O. Thoma).
29. Boehringer Ingelheim, US 2997469, 1961 (A. Boehringer, E. Boehringer, I. Liebrecht, J. Liebrecht et al.).
30. Ravensberg-Chemie, US 3018222, 1962 (H. Siemer, O. Hengen).
31. W. J. Houlihan, V. A. Parrino, *J. Org. Chem.* **47** (1982) 5177.
32. R. G. Engstrom, L. A. Kelly, J. H. Gogerty, *Arch. Int. Pharmacodyn.* **214** (1975) 308.
33. *Drugstore and Hospital Sales,* 1984 , IMS America, Ambler, Pa., 1984.

Antiulcer Drugs

This article discusses drugs used in acid-related diseases of the gastrointestinal tract.

Wolfgang Kromer, Byk Gulden, Konstanz, Federal Republic of Germany

Stefan Postius, Byk Gulden, Konstanz, Federal Republic of Germany

Uwe Krüger, Byk Gulden, Konstanz, Federal Republic of Germany

1. Introduction 1
2. Pathophysiology and Therapeutic Concepts 1
3. Drugs 2
3.1. Proton Pump Inhibitors 2
3.2. H_2-Receptor Antagonists 5
3.3. Antacids 8
3.4. Miscellaneous Drugs 10
3.4.1. Prostaglandin Derivates 10
3.4.2. Prokinetic Drugs 11
3.4.3. Antimuscarinic Drugs 12
3.4.4. Sucralfate 12
3.4.5. Bismuth Compounds 13
4. Prospect for Future Developments . 14
4.1. Reversibly Binding Acid Pump Antagonists 14
4.2. Drugs Directed Against *Helicobacter pylori* 14
5. Relative Drug Assessment 15
6. References 15

1. Introduction

Acid-related diseases of the upper gastrointestinal tract (GI-tract) comprise all the diseases of the esophagus, stomach, and duodenum that can be beneficially affected by neutralizing gastric acid or by inhibiting its secretion. These diseases include gastroesophageal reflux disease (GERD) [1], gastritis [2], gastric and duodenal ulcers (GU and DU [3]), drug-induced ulcers [4], Zollinger – Ellison syndrome [5], and stress ulcers [6]. However, "acid related" does not mean that all these diseases are causally related to acid hypersecretion. They may in fact go along with normal or even subnormal acid secretion. The complex pathophysiology is based on an imbalance between aggressive and defensive factors and will be discussed in Chapter 2.

Acid-related disorders of the GI-tract still constitute a major national-economical issue in western countries. The "crude lifetime ulcer prevalence" has been estimated in the United States to be between 8 and 30 % [7]. This principal notion also applies to Germany where, in 1995, 239×10^6 daily doses of antiulcer drugs [H_2-receptor antagonists (H_2-RAs), proton pump inhibitors (PPIs), etc.] were prescribed [8]. Three major risk factors for peptic ulcers and related disorders have been identified by Kurata and Nogawa [9]: infection by *Helicobacter pylori* (H.p.), use of nonsteroidal antiinflammatory drugs (NSAIDs) such as diclofenac, and cigarette smoking. H.p. infection is predominantly acquired in childhood, but the incidence of infections during childhood was considered to have decreased over the past decades being in line with a cohort phenomenon [10], [11]. Thus, as far as peptic ulcers caused by H.p. infection are concerned, their prevalence was expected to decrease in the future because transmission should become less likely under improved hygienic conditions in industrialized countries [12]. On the other hand, Hornemann et al. [13] found an age-dependent increase in the prevalence of H.p. infection in healthy German children between birth and 17 years of age admitted to hospital for minor surgical procedures in 1983 and 1984. Consequently, these authors questioned a hygiene-related cohort effect and interpreted their data as possibly pointing to an oral-oral route of infection as opposed to a fecal-oral route.

2. Pathophysiology and Therapeutic Concepts

Although the dictum "no acid – no ulcer" [14] is still valid in the majority of peptic ulcer cases,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_321

nowadays much more is known about their complex etiology and pathophysiology [15]. In fact, gastric acid and pepsin are the most important aggressive factors acting in concert with toxic factors of H.p. or with either NSAIDs, stress or smoking, but they are physiologically counterbalanced by defensive factors such as mucus and bicarbonate, epithelial cell lining and repair mechanisms, and mucosal blood flow [15], [16]. Consequently, although inhibition of acid secretion will be therapeutically effective even in cases with normal acid secretion, eradication of H.p. by antibiotics or improvement of mucosal defense, for example, by sucralfate and antacids, will either eliminate H.p. as the cause of the disease, or shift the balance in favor of defensive factors. Bismuth compounds do a little of both. Prostaglandin (PG) derivatives help preventing NSAID-induced ulcers by "gastroprotective" (misleadingly also called "cytoprotective") mechanisms but heal peptic ulcers only when given in antisecretory doses [17], [18]. Ulcer pain may be triggered by exposure of the inflamed mucosa to gastric acid [19] and will therefore be alleviated by antisecretory or acid-neutralizing drugs, apart from their healing effect.

The most effective target of acid-inhibitory drugs is the H^+/K^+-ATPase (acid pump; proton pump) of the parietal cell's apical membrane [20]. PPIs (pantoprazole, omeprazole, lansoprazole, see Section 3.1) block the final step in acid production (Fig. 1) and therefore display a higher antisecretory efficacy and clinical reliability than H_2-RAs (see Section 3.2) or, even more so, M_1-antimuscarinics. H_2-RAs block receptors for histamine located upstream on the basolateral membrane of the parietal cell (H_2-receptor), M_1-antimuscarinics block receptors for actetylcholine on either the histamine-liberating ECL-cell or intramural neurons (M_1-receptors). Consequently, both of these leave other pathways of acid stimulation unaffected. The only antimuscarinic antiulcer drug still in use is pirenzepine (see Section 3.4.3) which seems outdated due to its comparatively low efficacy combined with unwanted effects. Antacids (see Section 3.3), PG derivatives (see Section 3.4.1), sucralfate (see Section 3.4.4) and bismuth compounds (see Section 3.4.5) display complex actions partially related to acid, mucosal defensive mechanisms, and H.p. (see Section 4.2). Figure 1 provides an overview. Finally, the prokinetic drugs cisapride, metoclopramide, and domperidone facilitate propulsive motility of the GI-tract thus enhancing oral to aboral movement of contents to prevent duodeno-gastric reflux of bile acids or gastro-esophageal reflux of gastric acid (see Section 3.4.2).

3. Drugs

This section lays special emphasis on proton pump inhibitors as these drugs represent the latest development in the antisecretory therapy of peptic ulcer disease, succeeding H_2-receptor antagonists. In general, information on chemical and pharmaceutical drug properties, synonyms, trade names, dosages etc. beyond the information given below may be obtained from [22–26]. In the following sections, the chemical names of individual drugs are indicated according to the WHO nomenclature.

3.1. Proton Pump Inhibitors

Chemically, proton pump inhibitors (PPIs) consist of a substituted benzimidazole moiety linked by an essential sulfinylmethyl group to a substituted pyridine ring [20], [27]. They are acid-labile prodrugs that have first to be activated by acid in order to gain pharmacological activity but will subsequently be degraded to inactive products (see below). They are therefore administered as enteric-coated formulations (which protects the prodrug from acid in the gastric lumen), become absorbed from the intestine and penetrate into the acid compartment of the gastric parietal cell from the blood. The pH in the order of 1 present inside the canaliculus of a secreting parietal cell [20] causes protonation of the pyridine nitrogen (pK_a about 4) and thereby drug accumulation by a factor of more than 1000 relative to the blood. This protonation goes along with a complex, intramolecular rearrangement of the prodrug structure to form a thiophilic cyclic sulfenamide. The latter is a highly reactive compound and would be further degraded to inactive products (see above) if it were not generated in the vicinity of the extracytoplasmic portion of the proton pump where it immediately and covalently reacts with thiol groups of

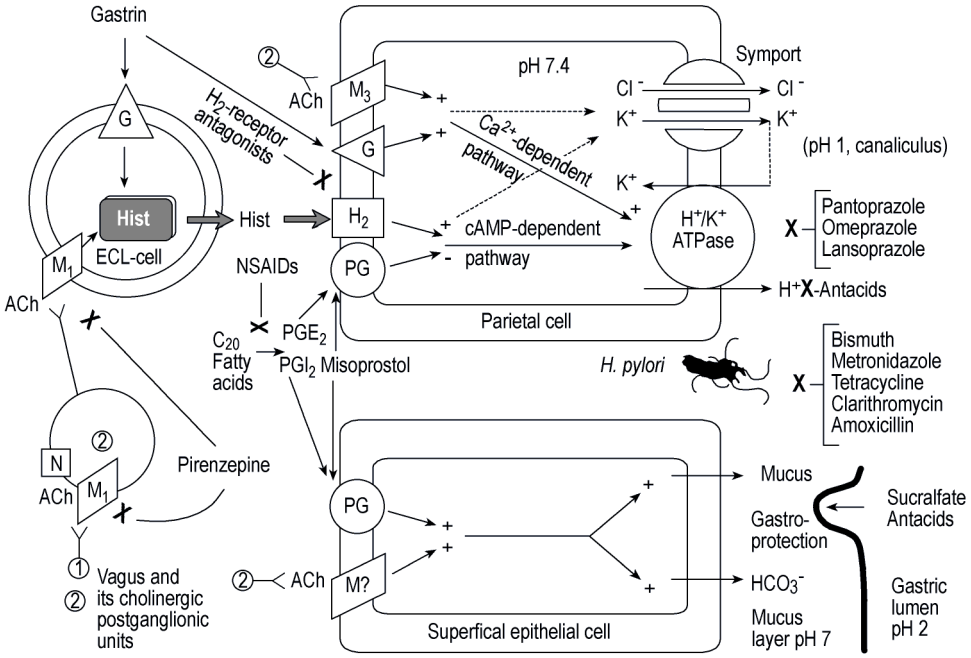

Figure 1. Simplified and incomplete illustration of the interplay between various aggressive and defensive gastric mucosal factors, their physiological regulation, and the different drug targets in peptic ulcer therapy.
ACh = acetylcholine; ECL = enterochromaffine-like; G = gastrin receptor; Hist = histamine; H_2 = histamine H_2 -receptor; M_1 = muscarinic M_1 -receptor; N = nicotinic receptor; NSAIDs = nonsteroidal antiinflammatory drugs; PG = prostaglandin receptor.
X indicates inhibition of the respective target.
After BRUNTON [21], slightly modified.

the proton pump. The result is a mixed disulfide of the active drug and the enzyme protein [28], [29]. The disulfide structure arrests the catalytic cycle of the proton pump [20] and thus inhibits acid secretion. Although this constitutes an irreversible mode of inhibition in a chemical sense, acid inhibition by PPIs is biologically reversible due to the turnover of both the proton pump ($t_{1/2}$ approx. 72 or 54 h for the rat proton pump or its alpha subunit; [30], [31]) and the parietal cell (turnover time in the rat ca. 164 d [32]).

It has been estimated that about one-third of the parietal cell mass is in its resting state even during meal stimulation [33] and therefore not available for inhibition by PPIs [34]. These resting cells are liable to endogenous stimulation after the drug has been eliminated from blood circulation and thus mainly determine the recovery of acid secretion following a first drug administration. It also follows that PPIs achieve their steady state effect only over about 3 days of daily administration, and any co-administration of receptor antagonists (be it H_2- or M_1-blockers) with PPIs is contraindicated. In fact, the latter would lose instead of gain efficacy because receptor blockade would prevent a certain percentage of parietal cells from being stimulated for acid secretion thus kept in a resting state not allowing for PPI activation.

Cyclic sulfenamides generated by acid from the inactive prodrugs (for review, see [27]) are specific SH-reagents which can modify any thiol group and should therefore be prevented from unwanted reactions at any targets apart from the parietal cell's proton pump in order to avoid side effects as much as possible. Thus, the main aspect governing selection of candidate compounds during drug development was their pH-dependent activation profile. As discussed in detail by KROMER et al. [37], PPIs should be acid-activated as fast as possible at low pH values of about 1 in order to potently inhibit the pari-

etal cell's proton pump. However, they should be activated as slowly as possible at pH values well above 3 to avoid unwanted SH-reactions in moderately acidic tissue compartments like lysosomes and other acidic vesicles. "Fast" and "slow" activation is always meant relative to the serum elimination half-life $t_{1/2}$. For example, omeprazole is as quickly activated at pH 5 (where activation is unwanted) as it is eliminated from serum ($t_{1/2}$ approx. 1 h), whereas the PPI pantoprazole is eliminated from blood by a factor of about 3 faster than it is activated at pH values in the order of 5 [37]. For lansoprazole and rabeprazole, see [35], [36]. All of these drugs are acid-activated at pH 1 with activation half-lifes in the order of a few minutes compared to their serum elimination half-lifes of about 1 hour. They all display therefore similar antisecretory potencies and efficacies [38–44] which solely depend on the activation rate within the acid space of the parietal cell. For the same reason, 20 mg omeprazole achieved lower healing rates than 40 mg of either omeprazole or pantoprazole in a number of studies [45–51]. This data does not conflict at all with the failure of other studies [52], [53] to statistically distinguish 40 mg pantoprazole from 20 mg omeprazole since the two doses represent the upper part of the dose – response relationship. Consequently, they are separated by only a small (but clinically relevant) difference in healing rates confounded by comparatively large variations between different patient samples. It has therefore been judged from a medical rather than a statistical point of view that 20 mg omeprazole daily being underdosed and 40 mg daily of both omeprazole and pantoprazole being the correct and desirable dose for acute ulcer healing [54], [55]. Figure 2 demonstrates the congruence between animal data, human pharmacology data, and selected clinical data (see above) and shows the equipotency and equiefficacy of pantoprazole compared to omeprazole on a milligram basis.

A further goal pursued during selection of PPIs was to minimize potential interactions with cytochrome P450 in order to avoid unwanted drug interactions in the liver. Following phase I oxidation in the endoplasmic reticulum of the liver cell, pantoprazole undergoes cytosolic phase II sulfate conjugation which is independent of cytochrome P450 and not subject to drug interactions [57]. By contrast, phase II conjugation is lacking in case of omeprazole, which in addition has an about seven-fold higher affinity to human cytochrome P450 than pantoprazole [58]. The two differences, i.e., lower affinity to cytochrome P450 and phase II conjugation of pantoprazole as opposed to omeprazole explain well that pantoprazole did not display any metabolic drug interactions in man [59], [60], whereas omeprazole did in a number of studies [61–70], for review, see [71].

Another goal considered during PPI selection and development was to obtain linear pharmacokinetics both with respect to dose increments and repeated dosing over time. This goal has been achieved with pantoprazole [72–74] but not with omeprazole [75].

All of the differences in the aforementioned drug properties depend on the different substitution patterns of the two ring systems of the PPIs (Table 1). Regarding the pH-selectivity in the PPI's chemical activation (see above), "the difluoromethoxy group present in pantoprazole slows reaction at neutral pH by electron withdrawal, which decreases the pK_a of the benzimidazole N and reduces its protonation at all but highly acidic pH values" [20]. The higher lipophilicity of omeprazole explains its higher volume of distribution (i.e., a higher percentage shifted from serum into tissues) and, consequently, its somewhat lower maximum serum concentration (C_{max}) and area under the serum concentration – time curve (AUC) (→ Pharmaceutical Dosage Forms, Chap. 1. on bioavailability) compared to the same dose of pantoprazole [37], [76]. The biological implications of these differences have been discussed in [37]. The individual PPIs will be presented below in alphabetical order.

Table 1. Chemical structures of four proton pump inhibitors

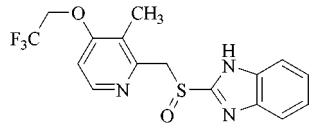

Lansoprazole

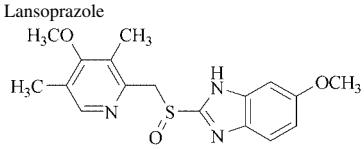

Omeprazole

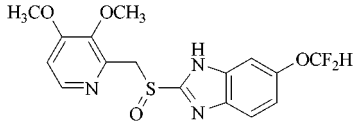

Pantoprazole

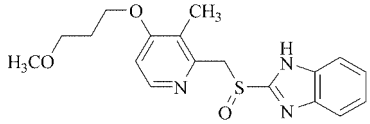

Pariprazole/Rabeprazole/E3810

Lansoprazole [*103577-45-3*], 2-[[[3-methyl-4-(2,2,2-trifluoroethoxy)-2-pyridyl]-methyl]sulfinyl]benzimidazole (WHO), $C_{16}H_{14}F_3N_3O_2S$, M_r 369.37 [77], available as capsules containing either 15 or 30 mg (free acid) of the racemic mixture (chiral center at the asymmetric sulfinyl group). Both the (+)- and (−)-enantiomers of lansoprazole are converted to the optically inactive cyclic sulfenamide and are therefore of the same antisecretory potency in vitro [78].

Trade names: Agopton (Takeda) and in addition various licensed products under different trade names (see [23]).

Omeprazole [*73590-58-6*], 5-methoxy-2-[[(4-methoxy-3,5-dimethyl-2-pyridyl)methyl]-sufinyl]benzimidazole (WHO), $C_{17}H_{19}N_3O_3S$, M_r 345.42 [52], available as capsules containing either 10, 20, or 40 mg and as a lyophylisate for i.v. infusion containing 40 mg of the racemic mixture (40 mg being equivalent to 42.6 mg of the sodium salt). The omeprazole enantiomers can be separated but both of them display about the same antisecretory potency in vitro [79], [80].

Trade names: The racemic mixture is produced by Astra/Hässle under trade names Antra and Losec, and is in addition distributed by various licensees under different trade names (see [23]).

Omeprazole was the first PPI in clinical use and has therefore attained the largest drug safety data base. Omeprazole has a remarkable safety profile with most adverse events in the range of that of placebo treatment [81–83]. Challenges had been hypergastrinemia (also seen with the other PPIs) with subsequent induction of carcinoids in rats chronically exposed to high toxicological doses [84], gastric bacterial overgrowth due to anacidity [85], impaired absorption of vitamin B_{12} and iron [86], induction of cytochrome P450 IA2 which both generates and detoxifies mutagens [87], [88]; suspected genotoxicity [89], and very rare events not conclusively related to omeprazole treatment (impaired vision and hearing, inflammatory and immunologic events like fever, vasculitis and colitis). All of these issues have been reviewed by experts and considered to be of either no relevance or even unrelated to omeprazole treatment [86], [90–96]. The aforementioned adverse events published as case reports seem to be extremely rare relative to the large number of treatment courses with omeprazole of well above 100 million. Some of these adverse effects, however, disappeared upon cessation of omeprazole therapy and reappeared upon rechallenge suggesting a causal relationship. These include cases of myopathy [97], neuromyositis [98], erythema nodosum [99], arthralgia [100], pityriasis rosea-like skin eruption [101], hemolytic anemia [102], interstitial nephritis [103], [104], CNS-symptoms [105], [106], and acute gout [107].

Pantoprazole [*102625-70-7*], 5-(difluoromethoxy)-2-[[(3,4-dimethoxy-2-pyridyl)methyl]-sulfinyl]benzimidazole (WHO), sodium salt $C_{16}H_{15}F_2N_3NaO_4S$, M_r 405.38, [27], [53], [56], [108–111]. Figure 3 shows the synthesis of the sodium salt, Figure 4 its chemical activation. Pantoprazole is available as enteric coated tablets containing 45.1 mg of the sodium salt sesquihydrate equivalent to 40 mg of the racemic mixture of the free anhydrous acid, and as a lyophylisate for i.v. infusion. The (+)- and (−)-enantiomers do not differ in their antisecretory potencies in animals. An oral 20 mg formulation for GERD/maintenance therapy has been approved recently.

Trade names: Pantozol (Byk Gulden) and others; various licensed products under different trade names (see [23]).

3.2. H_2-Receptor Antagonists

H_2-receptor antagonists (H_2-RAs) are structural analogues of histamine that – at least those in clinical use – competitively and surmountably

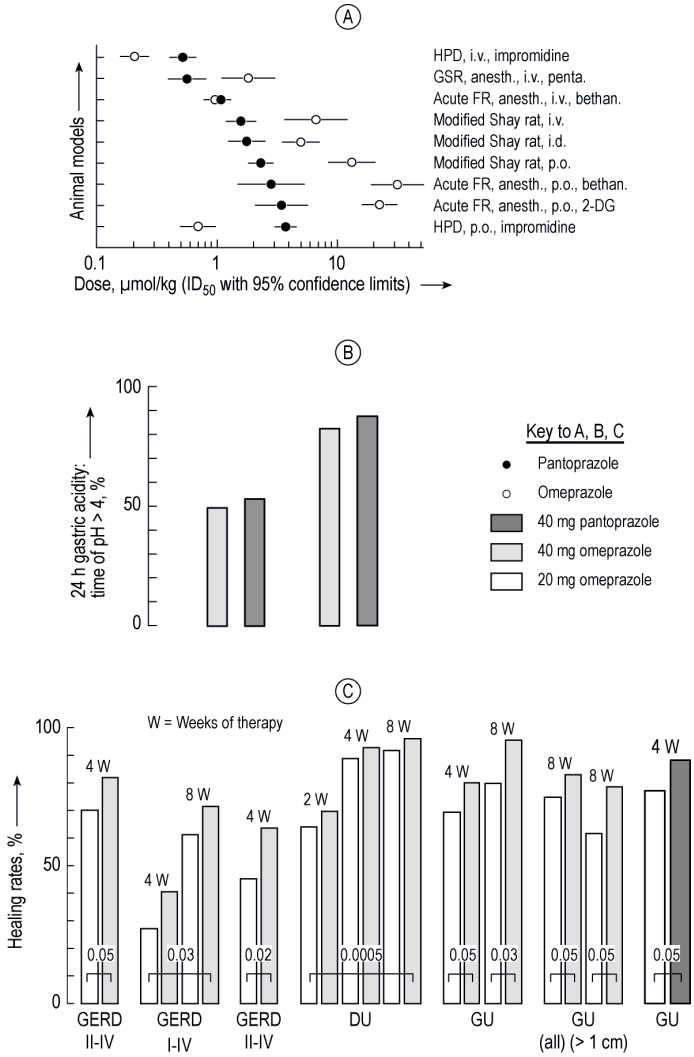

Figure 2. Congruence between animal data, human pharmacology data and clinical data with respect to equipotency and equiefficacy of two gastric proton pump inhibitors.A) Inhibition of gastric acid secretion in various animal models: Inhibitory doses, ID_{50} values (µmol/kg; 1 µmol = 0.35 mg of omeprazole or 0.38 mg of pantoprazole, free acid). HPD = Heidenhain pouch dog, FR = acute fistula rat, GSR = Ghosh-Schild rat, 2-DG = 2-deoxy-D-glucose. The routes of administration refer to test drugs, not secretagogues (impromidine, penta. = pentagastrin, bethan. = bethanechol). For further information, see [56].B) 24 h-pH-metry in healthy volunteers (time of intragastric pH > 4). All doses 40 mg orally, daily over 7 d Left pair of columns: Koop et al. [40]. Right pair of columns: Brunner et al. [44]. The comparison between the two studies shows that, although PPIs may vary in their efficacy depending on the random sample, omeprazole and pantoprazole proved equipotent when compared intraindividually.C) Healing rates in gastroesophageal reflux disease (GERD), duodenal ulcer (DU), and gastric ulcer (GU) in a comparison between daily oral doses of either 20 or 40 mg omeprazole or 40 mg pantoprazole. Selected studies that show a clear cut advantage of the 40 mg dose. From left to right: Hetzel et al. [45]; Laursen et al. [46]; Bate et al. [47]; Lauritsen et al. [48]; Walan et al. [49]; Valenzuela et al. [50]; Witzel et al. [51]. The specifications "all" and "> 1 cm" below the sixth group of columns [50] refer to the ulcers evaluated. p-Values are indicated at the bottom of columns compared to each other. All of the differences between the two doses shown in the figure are statistically significant irrespective of the particular drug involved. It is important to note that the failure of other studies (not shown here; see [50] and [53]) to demonstrate any significant difference between the healing rates upon 40 mg pantoprazole compared to 20 mg omeprazole does not conflict with the aforementioned superiority of the 40 mg dosage. In fact, 40 mg of either drug represent a ceiling dose separated by a comparatively small difference from the healing rate upon 20 mg. Thus, both of the two doses are localized in the upper portion of the dose response curve. Because of the large interindividual variation seen in clinical studies, the difference will be statistically significant in one, but statistically insignificant in another study.

block the H_2-receptors located on the basolateral membrane of the parietal cell. Histamine is the major secretagogue in humans [33], [112]. An imidazole ring as part of the structure was originally regarded essential for specific H_2-antagonism but this was refuted by the newer H_2-RAs (see [113]). They show a similarly favorable safety profile as the PPIs [114] but face the problem of about 10 % resistant peptic ulcers [115] and loss of effectiveness by about 50 % over a 4-week treatment course (tolerance development [116]). Potential reasons have been discussed in [117], [33]. Table 2 shows the chemical structures of the H_2-RAs. In contrast to cimetidine (imidazole ring), the newer compounds contain thiazole (famotidine and nizatidine), furan (ranitidine) or piperidine and benzene (roxatidine) ring systems. This helps to avoid unwanted cytochrome P450, i.e., metabolic drug interactions [118]. The individual drugs are discussed in the following in alphabetical order.

Table 2. Chemical structures of H_2-receptor antagonists. Note the different ring systems. The imidazole ring is mainly responsible for metabolic drug interactions [118].

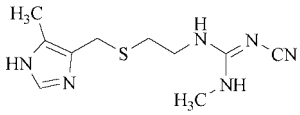

Cimetidine

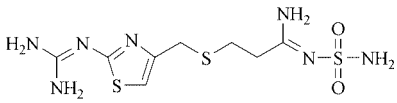

Famotidine

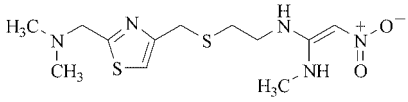

Nizatidine

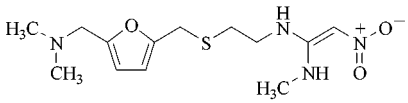

Ranitidine

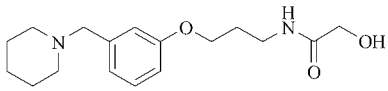

Roxatidine

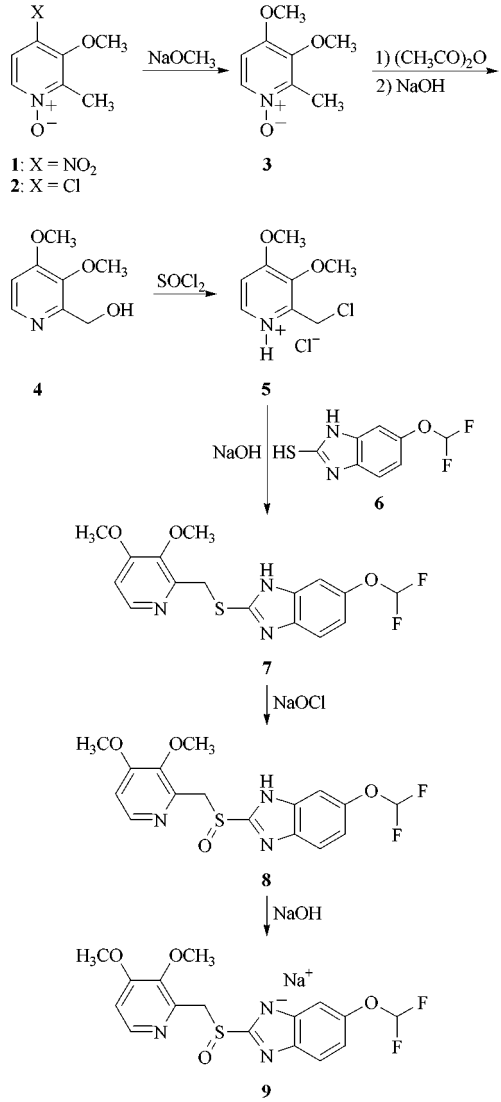

Figure 3. Synthesis of pantoprazole sodium as an example of PPIs [108]

Cimetidine [*51481-61-9*], 1-cyano-2-methyl-3-[2-[[(5-methylimidazol-4-yl)methyl]-thio]ethyl]guanidine (WHO) [119], $C_{10}H_{16}N_6S$, M_r 252.34, *mp* 141 – 143 °C, solubility in water at 37 °C 1.14 wt % increased by dilute HCl, pK_a-values 6.8 and 7.1. Cimetidine is a crystalline powder with bitter taste and characteristic odor. Citmetidine was the first of the H_2-RAs and revolutionized peptic ulcer therapy when introduced in the 1970s [120].

Trade names: Tagamet (SmithKline Beecham) and numerous others (see [23]).

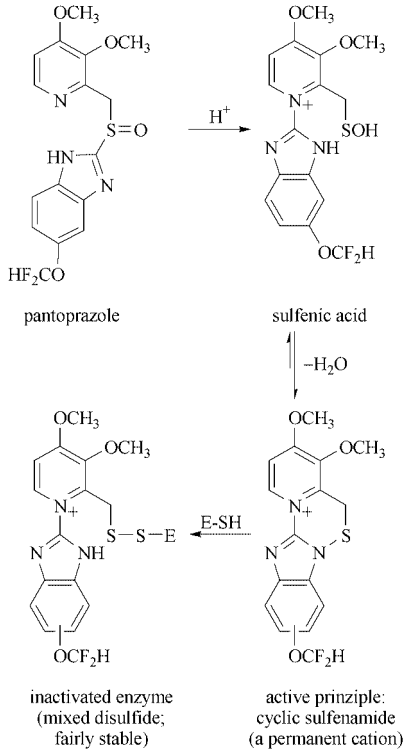

Figure 4. The chemical activation pathway of substituted benzimidazoles exemplified by pantoprazole (according to [27])E = enzyme

Famotidine [*76824-35-6*], [1-amino-3-[[[2-[(diaminomethylene)-amino]-4-thiazolyl]-methyl]thio]propylidene]-sulfamide (WHO) [121], $C_8H_{15}N_7O_2S_3$, M_r 337.43, *mp* 163 – 164 °C, solubility in water at 20 °C 0.1 wt %.

Trade names: Pepdul (Merck Sharp & Dohme) and numerous others (see [23]).

Nizatidine [*76963-41-2*], *N*-[2-[[[2-[(dimethylamino)methyl]4-thiazolyl]methyl]-thio]ethyl]*N'*-methyl-2-nitro-1,1-ethenediamine (WHO) [122], $C_{12}H_{21}N_5O_2S_2$, M_r 331.45, *mp* 130 – 132 °C, solubility in water 0.1 – 0.33 wt %, pK_a-values 2.1 and 6.8.

Trade names: Nizax (Lilly) and others (see [23]).

Ranitidine [*66357-35-5*], *N*-[2-[[5-[(dimethylamino)-methyl]furfuryl]thio]ethyl]*N'*-methyl-2-nitro-1,1-ethenediamine (WHO) [123], $C_{13}H_{22}N_4O_3S$, M_r 314.41, *mp* 69 – 70 °C, freely soluble in water. Ranitidine was the second H_2-receptor antagonist introduced in the early 1980s. It further optimized peptic ulcer therapy by minimizing unwanted interactions with androgen receptors and cytochrome P450 (see above). It has a higher affinity and, hence, greater potency compared to cimetidine which, however, is important only in the context of the relative drug safety margins.

Trade names: Zantic (Glaxo) and numerous others (see [23]). Ranitidine is also available as Raniditine bismuth citrate used in H.p.-positive ulcer patients (see Section 3.4.5).

Roxatidine [*78273-80-0*], *N*-[3-[(α-piperidino-*m*-tolyl)oxy]-propyl]glycolamide (WHO) [124], $C_{17}H_{26}N_2O_3$, M_r 306.41, *mp* 59 – 60 °C.

Trade name: Roxit (Albert-Roussel).

3.3. Antacids

Starting with naturally occurring carbonates centuries ago and later followed by potash combined with bismuth more than one century ago, antacids have since been developed and widely used over decades (see [125]). They are better than placebos and as effective as cimetidine in DU healing but less effective in GU healing [125]. Available both as liquids and tablets, they are nowadays usually taken 1 h after each meal and at bedtime in a total daily dose equivalent to about 400 mmol neutralizing capacity. This should be sufficient for all patients while minimizing adverse effects (see below). Such regimens elevate intragastric pH above 3.5 for about 2 h. Water-insoluble, aluminum – magnesium-containing antacids (like magaldrate) leave the stomach more slowly with the solid meal portion and are therefore longer acting. In addition to acid neutralization, these aluminum-containing antacids enhance the production of mucosal prostaglandins and possibly SH-containing compounds to protect the mucosa from injury ("gastroprotection") [125]. In therapy and prevention of mild GERD, acid neutralization, subsequent partial pepsin inactivation, and binding of bile acids and lysolecithin may act in concert (see [125]). Antacids display a favorable cost : benefit ratio in mild diseases.

More severe cases of acid related disorders require PPIs (see Section 3.1).

Although antacids did not display any teratogenic effects in animals and may therefore be used in pregnancy- and delivery-associated GERD (see [125]), they can interfere with iron absorption and may facilitate metabolic alcalosis and fluid overload. The latter effects depend on their bicarbonate and sodium ion content which is considered obsolete. In general, low sodium products should be preferred. Particulate antacids may also cause pneumonia upon aspiration and have been substituted for prevention of reflux complications largely by antisecretagogues. Nonulcer dyspepsia (NUD) and prophylaxis of NSAID-induced ulcers (see misoprostol, Section 3.4.1) are no recommended indications for antacids. Apart from the aforementioned adverse effects, aluminum-containing antacids can bind tetracyclines and many other drugs and thus prevent their absorption [21]. Aluminum-containing antacids may also complex phosphate (to cause osteomalacia in the worst case) and constipate the patient. The latter effect is counterbalanced by magnesium [21] contained in combination formulations such as magaldrate (see [126]), because magnesium salts in the intestine increase GI motility and cause osmotic diarrhea [127]. HCl converts insoluble aluminum oxide and hydroxide into more readily soluble and absorbable aluminum chloride. However, the causal relationship between aluminum intake with antacids and Alzheimer's disease is unlikely, while chronic intake of high doses may cause CNS toxicity in patients with renal insufficiency [128], [129].

Antacids containing carbonate, namely $NaHCO_3$ or $CaCO_3$, produce CO_2 with HCl and cause therefore gastric flatulence, belching and acid reflux. Na^+ may cause problems in hypertensive patients, Ca^{2+} adversely stimulates gastrin and HCl secretion ("acid rebound"). Those salts of carbonic acid are considered obsolete even in antacid mixtures although they are still used to some extent. Modern alternatives are the poorly soluble bases magnesium and aluminum hydroxide. If the two are incorporated together into one tridimensional lattice as represented by magaldrate (see below), they behave as a monosubstance buffering the pH between about 3 and 5 [130].

For tabulated overviews of antacids, see [21], [131], [132]. The indicated in vitro neutralization capacities do not necessarily reflect the relative therapeutic efficacies which are also affected by pH-dependent dissolution, viscosity, gastroprotective effects, etc. (see magaldrate). Antacid formulations may contain many additives such as simethicone (carminative contained in Riopan; see [131]); these are not considered here. Only the major antacid components which influence gastric acidity will be discussed below in alphabetical order.

Aluminum hydroxide [*21645-51-2*], $Al(OH)_3$, M_r 77.99, insoluble in water, soluble in dilute mineral acids and alkali hydroxides (4 % suspension in water), white, amorphous powder. For other relevant properties and production, see → Aluminum Oxide, Chap. 1.1. Available as $Al(OH)_3 \cdot nH_2O$ (algedrate) and algedrate hexitol complex. Depending on gastric acidity, stepwise buffering proceeds according to

$$Al(OH)_3 + 3\,HCl \rightleftharpoons Al(OH)_2Cl + H_2O + 2\,HCl$$
$$\rightleftharpoons Al(OH)Cl_2 + 2\,H_2O + HCl \rightleftharpoons AlCl_3 + 3\,H_2O$$

Pure $Al(OH)_3 \cdot nH_2O$ is not very stable, it ages at room temperature upon loss of water (formation of aluminum oxide) and neutralizing capacity. Aluminum hydroxide is preferably used as a combination preparation with magnesium hydroxide to counterbalance gut motility effects. This mixture also displays increased stability. Generated from aluminum hydroxide and gastric acid, aluminum chloride binds phosphate as insoluble $AlPO_4$ which is then excreted with feces. Aluminum hydroxide is therefore also used at mealtime in patients on dialysis to reduce phosphate intake. Magnesium containing antacids (see below) are contraindicated in this condition to avoid dangerous hypermagnesemia [132].

Trade names: Actal and numerous others (see [23]).

Aluminum phosphate [*7784-30-7*], $AlPO_4$, M_r 121.95, white powder, insoluble in water, soluble in concentrated mineral acids, pH 5.5 – 6.5 (4 % suspension in water). Available as $AlPO_4$ gel or $AlPO_4 \cdot 2\,H_2O$. For other relevant properties and production, see → Phosphoric Acid and Phosphates, Chap. 3.3.6.

Trade names: Phosphalugel (Boehringer Ingelheim) and others (see [23]).

Magnesium hydroxide [*1309-42-8*], $Mg(OH)_2$, M_r 58.32, amorphous powder, virtually insoluble in water (pH 9.5 – 10.5), soluble in dilute acids. For other relevant properties and production, see → Magnesium Compounds, Chap. 4. Available in combination preparations with aluminum hydroxide and under numerous trade names (see [23]). The equation for neutralization is

$$Mg(OH)_2 + 2\,HCl \rightleftharpoons MgCl_2 + 2\,H_2O$$

For potential side effects, see general discussion of antacids.

Magaldrate (hydroxymagnesium aluminate) [*74978-16-8*], $Al_5Mg_{10}(OH)_{31}(SO_4)_2 \cdot nH_2O$, insoluble in water, white, odorless crystalline powder, soluble in dilute mineral acids. Magaldrate [133] favorably combines the strongly basic properties of magnesium located in the crystal core with the less basic properties of aluminum ions surrounding the former in a layered lattice (Fig. 5). The magaldrate layered lattice consists of recurrent stacks of octahedral basic layers with an excess of positive charges $[Al_5Mg_{10}(OH)_{30}]^{5+}$ on the outer gel surface [134], [135]. These charges electrostatically interact with negative charges of the gastric mucus layer to favorably increase the antacid's retention time in the stomach [136].

Trade names: Riopan (Byk Gulden) and numerous others (see [23]).

In these brucite-like layers, about one third of the HO – Mg – OH units are substituted by positively charged $HO - Al^+ - OH$ units, counterbalanced by SO_4^{2-} and OH^- anions in the aqueous interlayer [137–139]. Compared to antacids composed of mixed hydroxides of aluminum and magnesium, magaldrate has a low aluminum content per milliliter and releases less aluminum because it rebinds aluminum ions in exchange for the more basic magnesium ions provided by an excess of antacid [140], [141]. A further advantage of the layered lattice is that its inner surface is not accessible to aluminum-binding proteins that otherwise reduce the buffering capacity of aluminum-containing conventional antacids [142]. In fact, although the in vitro acid-neutralizing capacity of magaldrate is similar to that of other antacids, its in vivo neutralizing capacity in humans is significantly higher [142] due to the exclusion of proteins [133]. However, magaldrate forms a gel with a large inner surface confining water spaces readily accessible to HCl. The alternation of magnesium and aluminum hydroxides within the layered lattice prevents losses of buffering capacity by aging. Magaldrate thereby shows prolonged acid neutralization up to the optimum of pH 4 [133]. In contrast, an aging phenomenon is found in aluminum hydroxide gels which is due to transition of aluminum hydroxide to aluminum oxide upon loss of water.

The magaldrate gel binds cholic acids like an anion exchanger, and it binds lysolecithin and pepsin on its outer surface [143]. It provides balanced, opposing actions of aluminum and magnesium on gut motility [21].

3.4. Miscellaneous Drugs

3.4.1. Prostaglandin Derivates
(→ Prostaglandins)

Prostaglandin (PG) derivatives exert gastroprotective effects at subantisecretory doses [144] but have the disadvantage of frequent side effects (mostly diarrhea and cramps) when administered at higher, antisecretory doses [128]. Only the latter accelerate healing of GU and DU. Prostaglandins seem to be somewhat inferior to H_2-receptor antagonists [145], particularly with respect to pain relief. Prostaglandins are abortifacient and are contraindicated during pregnancy and lactation. Low doses of PG derivatives have proved valuable in the prophylaxis of NSAID-induced peptic ulcers [21], [132]. Apart from the PG-E_1 derivative misoprostol (see below), other PG-E_1 (rioprostil) and PG-E_2 derivatives (arboprostil, enprostil, trimoprostil) had been under study [21], [146].

Misoprostol [*59122-46-2*], 15-deoxy-16-hydroxy-16-methyl-PG-E_1, (±)-Methyl-(1*R*,2*R*,3*R*)-3-hydroxy-2-[(*E*)-(4*RS*)-4-hydroxy-4-methyl-1-octenyl]-5-oxocyclopentaneheptanoate (WHO) [147], $C_{22}H_{38}O_5$, M_r 382.54, light yellow oil, slightly soluble in water. Misoprostol is approved for prophylaxis of

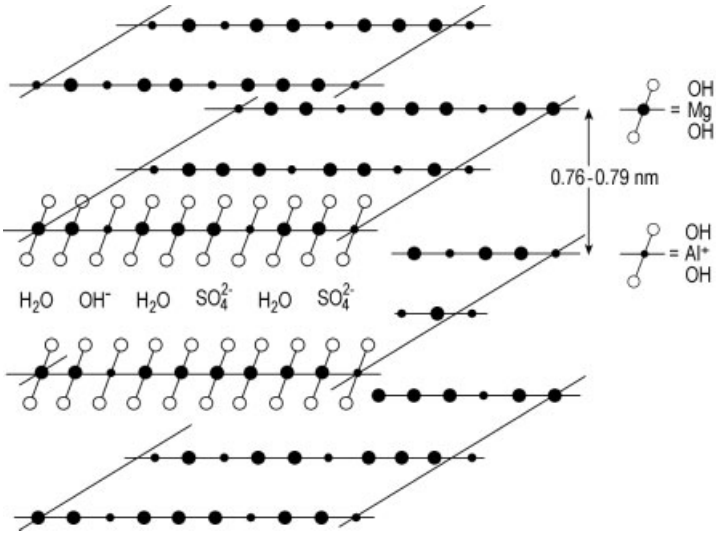

Figure 5. Schematic drawing of the layered lattice of magaldrate, a special example of an antacid (according to [134])

NSAID-induced ulcers and the treatment of DU and GU regardless of the limitations mentioned above.

Trade names: Cytotec (Heumann) and others (see [23]).

3.4.2. Prokinetic Drugs

Reflux esophagitis (GERD) is caused by reflux of gastric acid and therefore treated not only by PPIs and antacids but also by prokinetic drugs which facilitate oral to aboral clearing of the esophagus and emptying of the stomach [148]. Prokinetic drugs do so by alternately decreasing the frequency of intermittent lower esophageal sphincter relaxations, and increasing the amplitude and frequency of esophageal as well as gastric peristaltic contractions. Two pharmacological principles are used: (1) blockade of peripheral dopamine D_2-receptors by domperidone or metoclopramide which may lead to disinhibition of motility (although its mediation by D_2-receptor blockade is controversial), and (2) activation of peripheral 5-HT_4-receptors by either metoclopramide or cisapride. The latter interaction triggers the release of acetylcholine that subsequently stimulates motor neurons [149], [150]. Side effects relate primarily to either blockade of dopamine receptors (extrapyramidal motor and prolaktin effects as well as diarrhea; [151]) or release of acetylcholine (cramps, diarrhea; [152]). Domperidone has the advantage over metoclopramide of not significantly entering the central nervous system (CNS), thus avoiding central side effects. Although cisapride concentrations in rat brain can reach one-third to half of the plasma concentration, it does not block dopamine receptors and is therefore devoid of extrapyramidal motor and prolaktin effects [152]. Drug interactions by prokinetic drugs have been reviewed in [153].

Cisapride [*81098-60-4*], *cis*-4-amino-5-chloro-*N*-[1-[3-(*p*-fluorophenoxy)-propyl]-3-methoxy-4-piperidyl]-*o*-anisamide (WHO) [152], $C_{23}H_{29}ClFN_3O_4$, M_r 465.95, available as monohydrate, *mp* 109.8 °C is, like metoclopramide, a substituted benzamide.

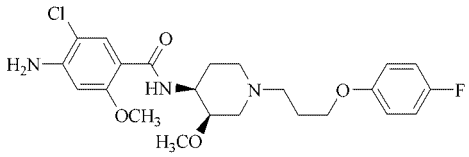

Trade names: Propulsin (Janssen) and others (see [23]).

Domperidone [*57808-66-9*], 5-chloro-1-[1-[3-(2-oxo-1-benzimidazolinyl)propyl]-4-piperidyl]-2-benzimidazolinone (WHO) [154],

$C_{22}H_{24}ClN_5O_2$, M_r 425.92, *mp* 242.5 °C, available as maleate (M_r 544.06).

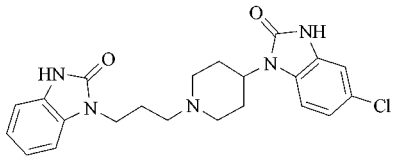

Trade names: Motilium (Byk Gulden) and others (see [23]).

Metoclopramide [*364-62-5*], 4-amino-5-chloro-*N*-[2-(diethylamino)-ethyl]-*o*-anisamide (WHO) [155], $C_{14}H_{22}ClN_3O_2$, M_r 299.81, *mp* 146.5 – 148 °C, solubility in water at 25 °C 0.02 wt %, pK_a-values 7.3 and 9.0. Metoclopramide is available as hydrochloride monohydrate (M_r 354.28) as a 10 % aqueous solution of pH 4.5 – 6.0.

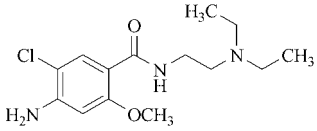

Trade names: Gastrosil (Heumann) and many others (see [23]).

3.4.3. Antimuscarinic Drugs

The only antimuscarinic drug still used in peptic ulcer disease is pirenzepine which has limited selectivity for M_1-muscarinic receptors present on both intramural neurons and paracrine, histamine-releasing cells (i.e., ECL-cells) of the gastric mucosa [156], [157].

Pirenzepine [*28797-61-7*], 5,11-dihydro-11-[(4-methyl-1-piperazinyl)acetyl]-6*H*-pyrido[2,3-b][1,4]benzodiazepin-6-one (WHO) [158], $C_{19}H_{21}N_5O_2$, M_r 351.42, pK_a values 2.1 and 8.1. Pirenzepine is available as dihydrochloride monohydrate, M_r 442.36, soluble in water. Because of its unfavorable balance between weak antisecretory and frequent anticholinergic adverse effects (dry mouth, blurred vision), pirenzepine is considered obsolete [128].

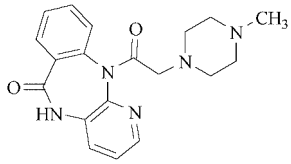

Trade names: Gastrozepin (Thomae) and many others (see [23]).

3.4.4. Sucralfate

Sucralfate [*54182-58-0*], sucrose hydrogen sulfate basic aluminum salt (WHO), $C_{12}H_{54}Al_{16}O_{75}S_8$, M_r 2086.74, insoluble in water, soluble in dilute HCl or NaOH, pK_a 0.43 – 1.19.

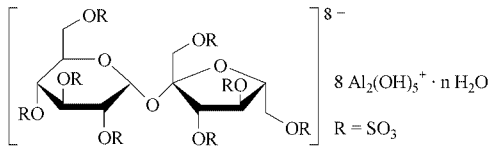

Sucralfate was first introduced in Japan in the late 1960s. Numerous active components of sucralfate have been described which act in concert to make sucralfate a gastroprotective antiulcer drug [159], [160] that even displays "cytoprotective" actions (i.e., on a cellular level) in vitro [161]. Maintenance of mucosal blood flow ensures rapid repair of mucosal damage; increased mucosal bicarbonate and mucus secretion protects the mucosa; and binding of fibroblast growth factor and epidermal growth factor to the sucralfate gel adhering to the injured mucosa enhances angiogenesis and epithelization of ulcerated areas [160]. Stimulation of PG biosynthesis and release of somatostatin may contribute to these actions and may further explain reduction of both parietal cell responsiveness and ulcer relapse rate following ulcer healing by sucralfate [159]. These effects may also explain reduced acid secretion under sucralfate therapy [162]. In addition, sucralfate may interfere with the attachment of H.p. to the gastric epithelium and may thereby suppress to some extent (but not eradicate) H.p. infection [159], [162]. Healing rates with sucralfate are similar to, and ulcer relapse rates are even lower than those after, H_2-receptor antagonists [132],

[163]. Tolerability is excellent, with aluminum-induced constipation being the most prominent side effect seen in 2–4 % of the patients [132], [164].

Trade names: Ulcogant (Merck) and many others (see [23]).

3.4.5. Bismuth Compounds

A variety of bismuth compounds are available as antiulcer drugs.

Bismuth aluminate [*12284-76-3*], $Bi_2(Al_2O_4)_3$.

Trade name: Noemin (Trommsdorf) and others (see [23]).

Bismuth citrate is on the market as a colloidal ammonium–potassium salt which is a mixture of the corresponding salts of the bismuth citrate complex $[BiC_6H_4O_7]^-$ (K or $NH_4)^+$ (see below) and citric acid $[C_6H_5O_7]^{3-}$ 3(K or $NH_4)^+$. The so-called "colloidal bismuth subcitrate, CBS" in fact contains the aforementioned bismuth citrate complex [165] in which H^+ of one citric acid carboxyl group is substituted by either K^+ or NH_4^+. These units are further complexed to a colloidal material [166]. Bismuth citrate is highly soluble in water.

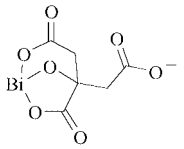

Trade names: Telen (Yamanouchi) and others (see [23]).

Bismuth subgallate [*99-26-3*], $C_7H_7BiO_7$, lipid solubility leads to a high rate of absorption, in contrast to the other salts (but see bismuth citrate).

Trade names: Bismofalk (Falk) and others (see [23]).

Bismuth subnitrate [*1304-85-4*], $BiO(NO_3)\cdot H_2O$.

Trade name: Angass (Medice).

Bismuth subsalicylate [*14882-18-9*], $C_7H_5BiO_4$.

Trade names: Jatrox (Procter & Gamble) and others (see [23]).

Mode of Action. These colloidal, basic bismuth preparations are, with the exception of bismuth citrate, almost insoluble in water. Depending on the actual pH, they form either colloidal solutions, highly hydrated gels, or thin films [167]. These cover the ulcerated areas, predominantly due to formation of bismuth oxychloride BiOCl, but also more complex structures containing free COOH and BiO^+ groups. These structures are precipitated from bismuth salts by gastric acid. Bismuth citrate is highly soluble in water but precipitates in an acid medium when the solubility product of its free acid (formed according to $R\text{–}CH_2CO_2^- + H^+ \rightleftharpoons R\text{–}CH_2CO_2H$) is exceeded [168]. Apart from the tenacious layer formed on the mucosa at low pH (see above), bismuth salts display a variety of additional action components that may be partially related to their reaction with thiol groups [169] and proteins [170], [171]. These additional effects include stimulation of PGE_2 biosynthesis and bicarbonate secretion, and either stimulation of mucus secretion or inhibition of its enzymatic degradation. Bismuth compounds also depress pepsin activity by either chief cell inhibition or pepsin complexation, and inhibit many enzymes of H.p., thus limiting mucus degradation. Bismuth salts bind and thereby inactivate bile acids at low pH, and accumulate epidermal growth factor in the ulcer area which may accelerate healing due to its trophic actions. Bismuth salts exert antimicrobial activity, also in H.p. infections [172]. The above pharmacodynamics of bismuth have been reviewed in [168], [173], with special reference to bismuth citrate, and in general in [132]. Substantial bismuth concentrations in the mucus or within the mucosa are observed for only a few hours following administration. Therefore, the low ulcer relapse rate with bismuth compounds may partially be due to trace amounts of bismuth systematically absorbed and, due to slow excretion, still available in the gastric tissue over up to 3 months. Nonetheless, safety of therapy with bismuth compounds seems to be excellent with only mild and infrequent side effects [174]. Probably due to its pK_a value in the order of 3

and, therefore, its comparatively fair ability to penetrate the stomach wall under acidic conditions as an uncharged molecule, bismuth citrate may produce higher (although transient) peak plasma concentrations of bismuth than some of the other bismuth salts which are unstable under these conditions. However, no major toxic effects have been reported [174].

Ranitidine bismuth citrate (RBC) [*128345-62-0*], $C_{19}H_{27}BiN_4O_{10}S$, M_r 712.48, is a novel salt of the base ranitidine and the monobasic acid bismuth citrate (note that a widely used nomenclature is subcitrate instead of citrate, see above and [165]). Ranitidine bismuth citrate forms a suspension in water but is soluble at low pH values to release its components. According to [175] RBC inactivates pepsin and H.p. more efficiently than an admixture of its components due to its better solubility. However, as discussed in page 13, bismuth citrate released at low pH will again precipitate as soon as the solubility product of its free acid is exceeded. Ranitidine bismuth citrate has been developed for combination therapy with antibiotics for eradication of H.p. Depending on the co-prescription, eradication ranged from 15 to 92 % in an intention-to-treat analysis [176]. It is claimed to combine the antisecretory action of ranitidine with the antimicrobial and gastroprotective actions of bismuth.

Trade name: Pylorid (GlaxoWellcome).

4. Prospect for Future Developments

4.1. Reversibly Binding Acid Pump Antagonists

In contrast to PPIs which bind covalently to the proton (acid) pump (see Section 3.1), acid pump antagonists (APAs) bind reversibly and compete with potassium on the luminal (extracytoplasmic) side of the H^+/K^+-ATPase to interrupt its catalytic cycle [20]. This reversible blockade not only regards the activated pumps already incorporated into the apical membrane, but also the resting pumps [177] still present in cytoplasmic vesicles. The latter are devoid of any proton gradient, they are unable to activate PPIs for this reason, and probably amount to about one-third of the total pumps in a parietal cell during meal stimulation [33], [177]. APAs will therefore result in a more rapid and complete inhibition of acid secretion than PPIs on a first administration. The duration of this inhibition by APAs corresponds to the drug's actual plasma concentration as opposed to a prolonged inhibition with covalently binding PPIs [178], [179]. APAs with different chemical structures have been synthesized (for example, imidazopyridines and acylquinolines; see [178]) but not one has reached the market so far. Although these weak nitrogen bases, like the PPIs, accumulate within the acidic space of the parietal cell, their selectivity for the acid pump is based on a specific fit between their chemical structure and that of the acid pump, being in line with a competitive mode of action.

Pumaprazole [*158364-59-1*], methyl 2-[[(2,3-dimethylimidazo[1,2-a]pyridin-8-yl)amino]methyl]-3-methylcarbanilate (WHO), $C_{19}H_{22}N_4O_2$. Pumaprazole has been abandoned in favor of a pharmacologically better follow up compound presently under development at Byk Gulden. The WHO decided to label the APAs by the stem syllable "-prazole" in order "to maintain the number of stems at a manageable level", irrespective of the completely different (i.e., reversible vs. covalent) modes of drug – target interactions by APAs and PPIs (WHO decision communicated on Nov. 30, 1995). Unfortunately, this pharmacologically misleading but legally binding decision will confuse the rational use and differential perception of APAs vs. PPIs in the future and counteracts the effort for an unambiguous terminology in pharmacology. It should be noted that the terms APAs and PPIs have been coined specifically to make a clearly recognizable distinction between the two chemically and mechanistically different classes of drugs.

4.2. Drugs Directed Against *Helicobacter pylori*

Helicobacter pylori (H.p.) has been recognized as a causal factor of most peptic ulcers (see Chap. 2). While companies worldwide try to develop antimicrobials both specific for H.p. and effective as a monotherapy, current therapy still takes advantage of the available reper-

toire of antibiotics [180–183] (→ Antibiotics). Although it has been claimed that antibiotics alone can eradicate H.p. infection and heal DU [184], present eradication regimes include one of the PPIs (see [180–183]) to enhance the effect of the antibiotics by a mode of interaction still under debate [185]. One-week triple therapies consisting of a PPI in combination with two of the three antibiotics amoxicillin (→ Antibiotics, Chap. 4.), clarithromycin, and metronidazole (→ Imidazole and Derivatives, Chap. 4.3.) are the therapies of choice according to the Maastricht consensus report [186].

5. Relative Drug Assessment

BURGET et al. [187] have published a model predicting the healing rates in DU, increasing up to an optimum defined by both a pH threshold and the percentage time the intragastric pH was above this threshold during a 24-h-period. A similar relationship in GERD treatment was later reviewed by HUNT [188]. It generally means that the longer the duration and the higher the degree of pH elevation by a certain drug is, the better the healing rate will be, up to an optimum which depends on the particular disease (GERD, GU, DU, Zollinger – Ellison Syndrome). This clearly favors the PPIs (see Section 3.1) for treatment of the H.p.-unrelated reflux esophagitis [189], [190]. The superiority of PPIs also shows in treatment of GU and DU with the limitation that any therapy eradicating H.p. (see Section 4.2) or suppressing H.p. in addition to gastroprotective effects (see Sections 3.4.4 and 3.4.5) seems to reduce the ulcer relapse rate compared to exclusively antisecretory drugs [191]. An interesting observation in this context is that eradication of H.p. may provoke GERD [192] or DU [193]. This may be due to the loss of neutralizing (NH_3) or antisecretory compounds that had been produced by H.p. in the patient's stomach before eradication therapy [194]. Obviously, this observation of DU induction by H.p. eradication therapy relates to the well known, low percentage of peptic ulcer cases causally independent of H.p. infection.

NSAID-induced peptic ulcers [195] can be prevented by prophylactic treatment with subantisecretory doses of misoprostol (see Section 3.4.1) but less so with H_2-receptor antagonists [196]. Of course, also PPIs are very effective prophylactic drugs [197]. Antacids have maintained their place in "common heartburn", gastritis, and less severe cases of GERD or GU and DU [132], although they constitute an inferior treatment option. Their major application will only be as cheap self-medication remedies.

6. References

1. P. J. Kahrilas: "Gastroesophageal reflux disease," in G. Friedman, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 31 – 44.
2. P. Malfertheiner, J. W. Freston: "Helicobacter pylori-induced gastritis, ulceration and neoplasia," in G. Friedman, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 21 – 29.
3. C. W. Howden: "Gastric and Duodenal ulcers," in G. Friedmann, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 45 – 51.
4. N. M. Agrawal, E. Z. Dajani: "Drug-induced Ulcerogenesis," in G. Friedman, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 55 – 63.
5. P. N. Maton: "Management of acid hypersecretion in patients with Zollinger-Ellison syndrome," in G. Friedman, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 65 – 73.
6. D. J. Cook: "Stress Ulcer Prophylaxis," in G. Friedman, E. D. Jacobson, R. W. McCallum (eds.): *Gastrointestinal Pharmacology and Therapeutics,* Lippincott-Raven Publishers, Philadelphia 1997, pp. 75 – 78.
7. J. H. Kurata, *Gastroenterology* **96** (1989) 569 – 580.
8. K. H. Holtermüller: "Magen-Darm-Mittel," in U. Schwabe, D. Paffrath (eds.): *Arzneiverordnungs-Report '96,* Gustav Fischer Verlag, Stuttgart 1996, pp. 305 – 324.

9. J. H. Kurata, A. N. Nogawa, *J. Clin. Gastroenterol.* **24** (1997) 2 – 17.
10. N. Banatvala, K. Mayo, F. Mégraud, R. Jennings, J. J. Deeks, R. A. Feldmann, *J. Infect. Dis.* **168** (1993) 219 – 221.
11. D. J. E. Cullen, B. J. Collins, K. J. Christiansen, J. Epis, J. R. Warren, I. Surveyor, K. J. Cullen, *Gut* **34** (1993) 1681 – 1682.
12. P. Sahay, A. T. R. Axon, *Helicobacter* **1** (1996) 175 – 182.
13. F. Hornemann, M. Nilius, P. Malfertheiner, P. Bartmann, *Helicobacter* **2** (1997) 176 – 179.
14. K. Schwarz, *Beitr. Klin. Chir.* **67** (1910) 96 – 128.
15. A. H. Soll: "Gastric duodenal and stress ulcer," in M. H. Sleisenger, J. S. Fordtran (eds.): *Gastrointestinal Disease,* 5th ed., **vol. 1,** W. B. Saunders Company, Philadelphia 1993, pp. 580 – 679.
16. J. L. Wallace, C. J. Bell, Current opinion in *Gastroenterology* **12** (1996) 503 – 511.
17. S. J. Konturek, *Scand. J. Gastroenterol.* **25** (1990) Suppl 174, 15 – 28.
18. M. Guslandi, *Drugs* **41** (1991) 507 – 513.
19. K. E. L. McColl, G. M. Fullarton, *Gut* **34** (1993) 1300 – 1302.
20. G. Sachs, J. M. Shin, C. Briving, B. Wallmark, S. Hersey, *Annu. Rev. Pharmacol. Toxicol.* **35** (1995) 277 – 305.
21. L. L. Burnton: "Agents for control of gastric acidity and treatment of peptic ulcers," in J. G. Hardman, L. E. Limbird, P. B. Molinoff, R. W. Ruddon, A. G. Goodman (eds.): *Goodman & Gilman's The Pharmacological Basis of Therapeutics,* 9th ed., McGraw-Hill, New York 1996, pp. 901 – 915.
22. E. Haen, W. Forth: *Wirkstoffprofile der Arzneimitteltherapie,* Ecomed Verlagsgesellschaft, Aulendorf 1997.
23. ABDATA Pharma-Daten-Service: "List of Pharmaceutical Substances," Eschborn/Taunus, Germany, 1997.
24. J. E. F. Reynolds (ed.): *Martindale. The Extra Pharmacopoeia,* 31st ed., Royal Pharmaceutical Society, London 1996.
25. *Merck Index,* 12th edition, Merck & Co., Inc., Whitehouse Station N.J. 1996.
26. *Index Nominum International Drug Directory,* 16th ed., Medpharm Scientific Publishers, Stuttgart 1995.
27. W. Kromer, *Digestion* **56** (1995) 443 – 454.
28. V. Figala et al., *J. Chem. Soc. Chem. Commun.* (1986) 125 – 127.
29. P. Lindberg et al., *J. Med. Chem.* **29** (1986) 1327 – 1329.
30. K. Gedda, D. Scott, M. Besancon, G. Sachs, *Gastroenterology* **106** (1994) Suppl., A79.
31. K. Gedda et al., *Gastroenterology* **109** (1995) 1134 – 1141.
32. H. Li, H. F. Helander, *Scand. J. Gastroenterol.* **30** (1995) 205 – 209.
33. B. I. Hirschowitz et al., *Dig. Dis. Sci.* **40** (1995) Suppl 2, 3S – 23S.
34. S. Postius, U. Bräuer, W. Kromer, *Life Sci.* **49** (1991) 1047 – 1052.
35. L. B. Barradell, D. Faulds, D. McTavish, *Drugs* **44** (1992) 225 – 250.
36. S. Yasuda et al., *Int. J. Clin. Pharmacol. Ther.* **32** (1994) 466 – 473.
37. W. Kromer, U. Krüger, R. Huber, M. Hartmann, V. W. Steinijans, *Pharmacology* **56** (1998) 57 – 70.
38. M. Hell, K. Eggert, G. Brunner, F. W. Schmidt, *Z. Gastroenterol.* **9** (1989) 562.
39. H. G. Dammann et al., *Aliment. Pharmacol. Ther.* **7** (1993) 191 – 196.
40. H. Koop et al., *Gut* **35** (1994) Suppl 4, A79.
41. H. Koop et al., *Eur. J. Gastroenterol. Hepatol.* **8** (1996) 915 – 918.
42. H. E. Scholtz, B. H. Meyer, H. G. Luus, *S. Afr. Med. J.* **85** (1995) 915.
43. M. Hartmann et al., *Aliment. Pharmacol. Ther.* **10** (1996) 359 – 366.
44. G. Brunner, H. Danz-Neef, C. Athmann, N. Samayoa, *Can. J. Gastroenterol.* **11** (1997) Suppl A, S2.
45. D. J. Hetzel et al., *Gastroenterology* **95** (1988) 903 – 912.
46. L. S. Laursen et al., *Gastroenterology* **102 (II)** (1992) Suppl., A110.
47. C. M. Bate et al., *Aliment. Pharmacol. Ther.* **7** (1993) 501 – 507.
48. K. Lauritsen et al., *Eur. J. Gastroenterol. Hepatol.* **4** (1992) 995 – 1000.
49. A. Walan et al., *N. Engl. J. Med.* **320** (1989) 69 – 75.
50. J. E. Valenzuela et al., *Am. J. Gastroenterol.* **91** (1996) 2516 – 2522.
51. L. Witzel, H. Gütz, W. Hüttermann, W. Schepps, *Aliment. Pharmacol. Ther.* **9** (1995) 19 – 24.
52. M. I. Wilde, D. McTavish, *Drugs* **48** (1994) 91 – 132.
53. A. Fitton, L. Wiseman, *Drugs* **51** (1995) 460 – 482.
54. H.-G. Dammann, *J. Pharmacol. Ther.* **4** (1995) 114 – 116.
55. P. Müller, B. Simon, *Münch. med. Wschr.* **138** (1996) 802 – 804.

56. W. Kromer et al., *J. Pharmacol. Exp. Ther.* **254** (1990) 129 – 135.
57. U. A. Meyer, *Yale J. Biol. Med.* **69** (1996) 203 – 209.
58. W. A. Simon, *Naunyn-Schmiedebergs Arch. Pharmacol.* **353** (1996) Suppl, R10.
59. V. W. Steinijans et al., *Int. J. Clin. Pharmacol. Ther.* **34** (1996) S 31 – S 50.
60. H. J. Koch et al., *Int. J. Clin. Pharmacol. Ther.* **34** (1996) 420 – 423.
61. R. Gugler, J. C. Jensen, *Gastroenterology* **89** (1985) 1235 – 1241.
62. R. Gugler, J. C. Jensen, *Pharmacol. Ther.* **33** (1987) 133 – 137.
63. T. Andersson, C. Cederberg, G. Edvardsson, A. Heggelund, P. Lundborg, *Clin. Pharmacol. Ther.* **47** (1990) 79 – 85.
64. T. Andersson et al., *Eur. J. Clin. Pharmacol.* **39** (1990) 51 – 54.
65. Y. Caraco, T. Tateishi, A. J. J. Wood, *Clin. Pharmacol. Ther.* **58** (1995) 62 – 72.
66. M. U. R. Naidu et al., *Drug Invest.* **7** (1994) 8 – 12.
67. P. J. Prichard et al., *Br. J. Clin. Pharmacol.* **24** (1987) 543 – 545.
68. S. Toon, B. L. Holt, F. G. P. Mullins, A. Khan, *J. Pharm. Pharmacol.* **47** (1995) 85 – 88.
69. T. Sutfin et al., *Ther. Drug Monit.* **11** (1989) 176 – 184.
70. P. Unge et al., *Br. J. Clin. Pharmacol.* **34** (1992) 509 – 512.
71. K. Zech et al., *J. Clin. Pharmacol. Ther.* **34** (1996) Suppl 1, S3 – S6.
72. B. Simon et al., *Aliment. Pharmacol. Ther.* **4** (1990) 373 – 379.
73. H. Bliesath, R. Huber, M. Hartmann, R. Lühmann, W. Wurst, *Int. J. Clin. Pharmacol. Ther.* **32** (1994) 44 – 50.
74. R. Huber et al., *Int. J. Clin. Pharmacol. Ther.* **34** (1996) 185 – 194.
75. T. Andersson, C. Cederberg, A. Heggelund, P. Lundborg, *Drug Invest.* **3** (1991) 45 – 51.
76. T. Andersson, C. Cederberg, C.-G. Regårdh, I. Skåndberg, *Eur. J. Clin. Pharmacol.* **39** (1990) 195 – 197.
77. A. E. Zimmermann, B. G. Katona, *Pharmacotherapy* **17** (1997) 308 – 326.
78. H. Nagaya, N. Inatomi, A. Nohara, H. Satoh, *Biochem. Pharmacol.* **42** (1991) 1875 – 1878.
79. S. Allenmark, B. Bomgren, H. Boren, P.-O. Lagerström, *Anal. Biochem.* **136** (1984) 293 – 297.
80. P. Erlandsson, R. Isaksson, P. Lorentzon, P. Lindberg, *J. Chromatogr.* **532** (1990) 305 – 319.
81. S. Joelson et al., *Digestion* **51** (1992) Suppl 1, 93 – 101.
82. R. Arnold, *Aliment. Pharmacol. Ther.* **8** (1994) Suppl 1, 65 – 70.
83. N. D. Yeomans, *Adverse Drug React. Toxicol. Rev.* **13** (1994) 145 – 156.
84. L. Ekman et al., *Scand. J. Gastroenterol.* **20** (1985) Suppl 108, 53 – 69.
85. B. K. Sharma et al., *Br. Med. J.* **289** (1984) 717 – 719.
86. H. Koop, *Aliment. Pharmacol. Ther.* **6** (1992) 399 – 406.
87. D. Diaz et al., *Gastroenterology* **99** (1990) 737 – 747.
88. W. M. McDonnell, J. M. Scheiman, P. G. Traber, *Gastroenterology* **103** (1992) 1509 – 1516.
89. B. Burlinson, S. H. Morriss, D. G. Gatehouse, D. J. Tweats, *Lancet* **335** (1990) 419.
90. W. Creutzfeldt, *Drug Saf.* **10** (1994) 66 – 82.
91. W. Creutzfeldt, *Z. Gastrenterol.* **33** (1995) Suppl 1, 12 – 21.
92. W. C. Creutzfeldt, A. Blum, *Lancet* **343** (1994) 1098.
93. D. Colin-Jones, *Lancet* **343** (1994) 1369.
94. S. Holt, Zhu Zhao-Hua, R. E. Powers, *Gastroenterology* **101** (1991) 650 – 656.
95. W. Schepp, V. Schusdziarra, *Z. Gastroenterol.* **28** (1990) 644 – 645.
96. J. Fryklund, A.-K. Falknäs, H. F. Helander, *Scand. J. Gastroenterol.* **27** (1992) 521 – 528.
97. F. J. Garrote et al., *Lancet* **340** (1992) 672.
98. F. Sarrot-Reynauld, H. Charras-De Matteis, C. Massot, *Reactions* **648** (1997) 10.
99. R. M. Ricci, K. C. Deering, *Cutis* **57** (1996) 434.
100. M. Beutler, K. Hartmann, M. Kuhn, J. Gartmann, *Br. Med. J.* **309** (1994) 1620.
101. C. Buckley, *Br. J. Dermatol.* **135** (1996) 660 – 661.
102. H. Amjad, W. Beckley, *Blood* **86** (1995) Suppl 1, 846A.
103. S. J. Ruffenach, M. S. Siskind, Y. H. H. Lien, *Am. J. Med.* **93** (1992) 472 – 473.
104. M. Assouad, S. L. Vicks, M. V. Pokroy, P. J. Willcourt, *Lancet* **344** (1994) 549.
105. E. J. M. Meeuwisse et al., *Reactions* **639** (1997) 10.
106. J. M. Prignet, E. Chauveau, J. L. Duval, D. Kunkel, *Semain des Hospitaux* **72** (1996) 19 – 20.
107. A. Kraus, L. F. Flores Suarez, *Lancet* **345** (1995) 461 – 462.
108. W. Kromer, B. Kohl, *Drugs Future* **15** (1990) 801 – 804.

109. W. Kromer, S. Gönne, R. Riedel, S. Postius, *Pharmacology* **41** (1990) 333 – 337.
110. R. Huber et al., *Aliment. Pharmacol. Ther.* **9** (1995) 363 – 378.
111. V. W. Steinijans (ed.), *Int. J. Pharmacol. Ther.* **34** (1996) Suppl 1.
112. J. W. Black, W. A. M. Duncan, C. J. Durant, C. R. Ganellin, E. M. Parsons, *Nature* **236** (1972) 385 – 390.
113. K.-Fr. Sewing, *Scand. J. Gastroenterol.* **23** (1988) Suppl 146, 73 – 77.
114. J. G. Mills et al., *Aliment. Pharmacol. Ther.* **11** (1997) 129 – 137.
115. G. Bianchi Porro, F. Parente, *Scand. J. Gastroenterol.* **23** (1988) Suppl 153, 81 – 88.
116. C. U. Nwoloko, J. T. L. Smith, A. Sawyerr, R. E. Pounder, *Aliment. Pharmacol. Ther.* **4** (1990) Suppl 1, 29 – 46.
117. C. H. Wilder-Smith, H. S. Merki, *Scand. J. Gastroenterol.* **27** (1992) Suppl 193, 14 – 19.
118. M. Murray, *Drug Metab. Rev.* **18** (1987) 55 – 81.
119. R. N. Brogden, R. C. Heel, T. M. Speight, G. S. Avery, *Drugs* **15** (1978) 93 – 131.
120. M. Freemantle, *C&E News* (1998) Jan. 5, 31 – 32.
121. D. M. Campoli-Richards, S. P. Clissold, *Drugs* **32** (1986) 197 – 221.
122. A. H. Price, R. N. Brogden, *Drugs* **36** (1988) 521 – 539.
123. S. M. Grant, H. D. Langtry, R. N. Brogden, *Drugs* **37** (1989) 801 – 870.
124. D. Murdoch, D. McTavish, *Drugs* **42** (1991) 240 – 260.
125. C.-K. Ching, S.-K. Lam, *Drugs* **47** (1994) 305 – 317.
126. R. P. Walt, M. J. S. Langman, *Drugs* **42** (1991) 205 – 212.
127. K. D. Fine, G. J. Krejs, J. S. Fordtran: "Diarrhea," in M. H. Sleisenger, J. S. Fordtran (eds.): *Gastrointestinal Disease,* W. B. Saunders, Philadelphia 1993, pp. 1043 – 1072.
128. D. W. Piper, *Drug Saf.* **12** (1995) 120 – 138.
129. W. Forth, *Dt. Ärztebl.* **84** (1987) C-1316 – 1317.
130. I. Szelenyi, *Drugs News Perspect.* **3** (1990) 396 – 402.
131. H. Arnott, *On Contin. Pract.* **18** (1991) 36 – 44.
132. Van Ness et al. (eds.): *Handbook of Gastrointestinal Drug Therapy,* 2nd ed., Little, Brown and Company, Boston 1995.
133. W. Kurtz, *Fortsch. Med.* **111** (1993) 93 – 96.
134. U. Krüger, *Gastro-Entero-Hepatologie* **3** (1985) Suppl 1, 5 – 7.
135. U. Assmus, A. Schwind, *Pharma International* **5** (1990) 172 – 178.
136. J. Monés et al., *Gut* **32** (1991) 147 – 150.
137. G. Brown, M. C. Gastuche, *Clay Miner.* **7** (1967) 193 – 201.
138. H. F. W. Taylor, *Mineral. Mag.* **39** (1973) 377 – 389.
139. L. R. Gutermann, C. J. Falzone, G. E. Wilson, *J. Pharm. Sci.* **75** (1986) 502 – 505.
140. H. C. Dollinger, E. Holzberg: "Absorption von Aluminium und Magnesium unter Magaldrat und anderen magnesium-aluminiumhaltigen Antazida," in R. Arnold, W. H. Häcki (eds.): *Magaldrat,* Verlag Christian M. Silinsky, Nürnberg 1985, pp. 98 – 103.
141. B. Winterberg et al., *J. Trace Elem. Electrolytes Health Dis.* **1** (1987) 69 – 72.
142. W. H. Häcki, F. Conti, M. Uehliger: "Diskrepanz zwischen In-vivo- und In-vitro-Säurebindungskapazitäten von Magaldrat (Riopan), Gelusil und Maaloxan," in R. Arnold, W. H. Häcki (eds.): *Magaldrat,* Verlag Christian M. Silinsky, Nürnberg 1985, pp. 57 – 61.
143. C. Baur, A. Becker, R. Linder, T. Schwan, *Drug Res.* **31** (1981) 504 – 507.
144. J. L. Wallace, A. W. Tigley, *Aliment. Pharmacol. Ther.* **9** (1995) 227 – 235.
145. S. J. Sontag, *Drugs* **32** (1986) 445 – 457.
146. K. L. Goa, J. P. Monk, *Drugs* **34** (1987) 539 – 559.
147. J. P. Monk, S. P. Clissold, *Drugs* **33** (1987) 1 – 30.
148. B. Ramirez, J. E. Richter, *Aliment. Pharmacol. Ther.* **7** (1993) 5 – 20.
149. M. R. Briejer, L. M. A. Akkermans, J. A. J. Schuurkes, *Pharmacol. Res.* **47** (1995) 631 – 651.
150. K.-A. Gwee, N. W. Read, *Aliment. Pharmacol. Ther.* **8** (1994) 105 – 118.
151. F. Mitchelson, *Drugs* **43** (1992) 295 – 315.
152. L. R. Wiseman, D. Faulds, *Drugs* **47** (1994) 116 – 152.
153. J. M. C. Greif, D. Rowbotham, *Clin. Pharmacokinet.* **27** (1994) 447 – 461.
154. M. C. Champion, *Pharma* **19** (1988) 499 – 505.
155. R. A. Harrington et al., *Drugs* **25** (1983) 451 – 494.
156. E. C. Hulme, N. J. M. Birdsall, N. J. Buckley, *Annu. Rev. Pharmacol. Toxicol.* **30** (1990) 633 – 673.
157. W. Kromer, E. Baron, R. Boer, M. Eltze, *Naunyn-Schmiedebergs Arch. Pharmacol.* **343** (1991) 7 – 13.

158. A. A. Carmine, R. N. Brogden, *Drugs* **30** (1985) 85 – 126.
159. W. D. W. Rees, *Scand. J. Gastroenterol.* **27** (1992) Suppl 191, 4 – 6.
160. S. Szabo, *Scand. J. Gastroenterol* **26** (1991) Suppl 185, 7 – 12.
161. M. Romano, M. Razandi, K. J. Ivey, *Dig. Dis. Sci.* **35** (1990) 467 – 476.
162. S. Banerjee et al., *Gastroenterology* **110** (1996) 717 – 724.
163. T. E. Bynum, *Scand. J. Gastroenterol.* **27** (1992) Suppl 191, 1 – 3.
164. I. N. Marks, *Scand. J. Gastroenterol.* **26** (1991) Suppl 185, 36 – 42.
165. W. A. Herrmann, E. Herdtweck, L. Pajdla, *Inorg. Chem.* **30** (1991) 2579 – 2581.
166. E. Asato, K. Katsura, M. Mikuriya, T. Fujii, J. Reedijk, *Inorg. Chem.* **32** (1993) 5322 – 5329.
167. R. Raedsch, *Dtsch. med. Wschr.* **116** (1991) 821 – 824.
168. A. J. Wagstaff, P. Benfield, J. P. Monk, *Drugs* **36** (1988) 132 – 157.
169. W. Beil, S. Bierbaum, K.-F. Sewing, *Pharmacology* **47** (1993) 135 – 140.
170. W. Beil, S. Bierbaum, K.-F. Sewing, *Pharmacology* **47** (1993) 141 – 144.
171. J. Wieriks et al., *Scand. J. Gastroenterol.* **17** (1982) Suppl 80, 11 – 16.
172. J. R. Lambert, P. Midolo, *Aliment. Pharmacol. Ther.* **11** (1997) Suppl 1, 27 – 33.
173. S. P. Lee, *Scand. J. Gastroenterol.* **26** (1991) Suppl 185, 1 – 6.
174. L. A. Tillman, F. M. Drake, J. S. Dixon, J. R. Wood, *Aliment. Pharmacol. Ther.* **10** (1996) 459 – 467.
175. A. A. McColm et al., *Aliment. Pharmacol. Ther.* **10** (1996) 241 – 250.
176. J. W. Wyeth et al., *Aliment. Pharmacol. Ther.* **10** (1996) 623 – 630.
177. K. B. Munson, G. Sachs, *Biochemistry* **27** (1988) 3932 – 3938.
178. A. J. Pope, M. E. Parsons, *Trends Pharmacol. Sci.* **14** (1993) 323 – 325.
179. W. Wurst, M. Hartmann, *Yale J. Biol. Med.* **69** (1996) 233 – 243.
180. A. T. R. Axon, *Scand. J. Gastroenterol.* **29** (1994) Suppl 201, 16 – 23.
181. A. Hackelsberger, P. Malfertheiner, *Drug Saf.* **15** (1996) 30 – 52.
182. J. G. Pentson, *Aliment. Pharmacol. Ther.* **10** (1996) 469 – 486.
183. F. Mégraud, *Aliment. Pharmacol. Ther.* **11** (1997) Suppl 1, 43 – 53.
184. S.-K. Lam et al., *Gut* **41** (1997) 43 – 48.
185. W. L. Peterson, *Aliment. Pharmacol. Ther.* **11** (1997) Suppl 1, 21 – 25.
186. EHPSG (The European Helicobacter Pylori Study Group) *Gut* **41** (1997) 8 – 13.
187. D. W. Burget, S. G. Chiverton, R. H. Hunt, *Gastroenterology* **99** (1990) 345 – 351.
188. R. H. Hunt, *Aliment. Pharmacol. Ther.* **9** (1995) Suppl 1, 3 – 7.
189. E. C. Klinkenberg-Knol, H. P. M. Festen, S. G. M. Meuwissen, *Drugs* **49** (1995) 695 – 710.
190. N. Chiba, Ch. J. De Gara, J. M. Wilkinson, R. H. Hunt, *Gastroenterology* **112** (1997) 1798 – 1810.
191. M. G. Korman, *Scand. J. Gastroenterol.* **30** (1995) Suppl. 208, 21 – 23.
192. J. Labenz et al., *Gastroenterology* **112** (1997) 1442 – 1447.
193. T. Pohle, E. Hilker, W. Domschke, *Z. Gastroenterol.* **35** (1997) 563 – 565.
194. E. F. Verdú et al., *Gut* **37** (1995) 743 – 748.
195. J. M. Scheiman, *Dig. Dis.* **12** (1994) 210 – 222.
196. M. Koch et al., *Dig. Dis.* **13** (1995) Suppl 1, 62 – 74.
197. A. Garner, H. Fadlallah, *Exp. Opin. Invest. Drugs* **6** (1997) 885 – 893.

Aldols → Aldehydes, Aliphatic and Araliphatic
Alfol Process → Fatty Alcohols
Algin → Polysaccharides
Alkali and Chlorine Products → Chlorine
Alkali and Chlorine Products → Sodium Hydroxide

Anthelmintics

PETER ANDREWS, Bayer AG, Wuppertal, Federal Republic of Germany

ACHIM HARDER, Bayer AG, Wuppertal, Federal Republic of Germany

1. Introduction 1
2. Parasitic Worms 1
2.1. Trematodes (Flukes) 2
2.1.1. Blood Flukes 2
2.1.2. Lung Flukes 2
2.1.3. Liver Flukes 3
2.1.4. Giant Intestinal Fluke 3
2.1.5. Lesser Intestinal Flukes 3
2.2. Cestodes (Tapeworms) 4
2.2.1. Broad or Fish Tapeworm 4
2.2.2. Beef Tapeworm 4
2.2.3. Pork Tapeworm 5
2.2.4. Dwarf Tapeworm 5
2.2.5. Dog Tapeworm 5
2.3. Nematodes (Roundworms) 5
2.3.1. Roundworm 6
2.3.2. Hookworms 6
2.3.3. Pinworm 6
2.3.4. Whipworm 7
2.3.5. Threadworm 7
2.3.6. Trichina 7
2.3.7. Larvae of Toxocara canis, Toxocara cati, Ancylostoma brasiliense and Ancylostoma caninum 8
2.3.8. Guinea Worm 8
2.3.9. Lymphatic Filariae 8
2.3.10. Convoluted Filaria 8
2.3.11. African Eye Worm 9
3. Anthelmintic Drugs 9
4. References 15

1. Introduction

Parasitism is a special form of intimate relationship between two species. One species, the host, is to some degree injured through the activities of the other species, the parasite. The parasite need not be parasitic through all stages of its existence. Parasites living within the host are distinguished as endoparasites, e.g., helminths, from those found on the surface of the body, the ectoparasites, e.g., arthropods. There are forms of close biologic relationships other than parasitism. Commensalism denotes an association beneficial to one partner and at least not disadvantageous to the other. Mutualism (or symbiosis) denotes an association that is mutually beneficial.

This article describes only the helminth infections of humans. Infections caused by other parasitic organisms, such as protozoa and ectoparasites, are described elsewhere (→ Chemotherapeutics, → Dermatotherapeutic Agents), as are the veterinary parasitoses (→ Veterinary Drugs).

This article is not intended to be a guide to therapy, and mention of a drug is not to be taken as an endorsement. Chapter 2 describes the helminth parasites of humans and the diseases they cause, and Chapter 3 describes the drugs that may be used to control the parasites.

2. Parasitic Worms

Helminths are the wormlike organisms classified in two phyla, the flatworms (Platyhelminthes), comprising the flukes (or trematodes) and the tapeworms (or cestodes), and the roundworms (Nemathelminthes). The roundworms are also called nematodes. Therefore, "helminth" refers to three morphologically different groups: flukes, tapeworms, and roundworms. Most flatworms are hermaphroditic, having both male and female reproductive organs within the same individual, and are parasites. In contrast, the roundworms have separate sexes and are overwhelmingly free living. Nevertheless, a number of roundworm species are parasitic, affecting humans, animals, and plants.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_329

2.1. Trematodes (Flukes)

2.1.1. Blood Flukes (*Schistosoma* spp.)

Three species of schistosomes cause most human schistosomiasis infections: *Schistosoma mansoni* and *S. japonicum* (locally also *S. intercalatum* and *S. mekongi*) cause the intestinal form of the disease, whereas *S. haematobium* (rarely *S. mattheii*) causes the urinary form.

Distribution. *S. haematobium:* all over Africa, in the Near East (Syria to Iran and Yemen), islands off East Africa. *S. mansoni:* throughout most of Africa southeast of a line from Gambia to Libya, Arabian peninsula, Brazil, Venezuela, Surinam, some Antilles Islands. *S. japonicum:* southern half of continental China, the Philippines; however, virtually extinct in Japan. *S. mekongi:* Laos, Cambodia. *S. intercalatum:* Cameroun to the Congo Basin. *S. mattheii:* restricted to southern Africa. About 270 million people are infected [13]. Annually, about one million deaths are caused by schistosomiasis [14].

Development. The schistosomes differ in many aspects from all other human flukes. They are elongate, wormlike, and live within the blood vessels. They have separate sexes. The adult parasites *S. mansoni* and *S. japonicum* are found as pairs in the mesenteric veins, whereas *S. haematobium* inhabits the veins of the vesical plexus. Adult males are 10 – 15 mm long and 0.5 – 1 mm wide. Their lateral body seams are folded downwards, forming a tube, the gynecophoric canal, which holds the female, which can be up to 28 mm long. The females deposit eggs, a few dozen to several thousand per day, depending on the species. About three fourths of the eggs laid are retained in the tissues of the body – mainly the liver and the intestinal and bladder walls – where they die within three weeks. The other eggs pass through the tissues into the lumen of the intestine or the bladder. They contain a larva, the miracidium, which hatches as soon as the eggs are voided into fresh water. The miracidium swims through the water in search of an appropriate snail (its intermediate host), penetrates the snail's tissues, and begins to multiply. After 4 – 7 weeks another type of larva, the cercaria, is released from the snail. These cercariae locate human skin exposed to water and penetrate it. Within another 6 – 12 weeks the juvenile schistosomes migrate through the vascular system and lungs to their predilected sites, mature, pair, and start laying eggs. The normal lifespan of schistosomes is 2 – 7 years. However, some parasites can survive and produce eggs for as long as thirty years.

Clinical Features. Severe itching (cercarial dermatitis) may occur at the site of skin penetration, and fever may occur during the migration phase of the parasite. In established infections, the egg incites an inflammatory reaction around the site where it is deposited in the tissues. This leads to fibrosis and calcification of the bladder, hematuria, uretric and renal involvement, and bladder cancer (*S. haematobium*). In the intestinal form of the disease (*S. mansoni* and *S. japonicum*) symptoms encountered are diarrhea, weakness, abdominal pain, splenomegaly, hepatic and portal fibrosis, portal hypertension, ascites, and carcinoma of the colon. Severe infections with *S. japonicum* can interfere with normal growth and mental development.

Therapy: praziquantel, oxamniquine, metrifonate, niridazole, hycanthone.

2.1.2. Lung Flukes (*Paragonimus* spp.)

Distribution. *P. westermani:* East Asia (Korea, Japan, China, Philippines, Indonesia). *P. kellicotti:* North America. *P. africanus:* Gabon to Zaire; several other species ranging from Mexico to Peru. About 20 million people are infected [15].

Development. The adult parasites measure 8 – 16 mm by 4 – 6 mm. They are found in pulmonary cysts that are formed by destroyed tissue and that connect to the respiratory passages of the lung. Eggs are voided with the sputum or the feces. Once eggs reach fresh water, a larva, the miracidium, develops within 2 – 4 weeks. The miracidium hatches from the egg, actively penetrates into specific aquatic snails, and multiplies. After 8 – 10 weeks, cercariae begin to be released from the snail. They locate crabs or crayfish, penetrate them, and encyst in the musculature as metacercariae. The crustaceans can also

become infected by feeding on infected snails. Humans acquire the infection by ingesting raw or improperly cooked crab meat or juice. The juvenile flukes excyst in the small intestine, penetrate the gut wall, and migrate through the diaphragm into the pleural cavity, where they penetrate the serosal layers of the lungs. Finally they arrive in the vicinity of the bronchioles. Within six weeks they develop into adult worms in tissue capsules that rupture and thus connect with the air ducts. These parasites can live for up to twenty years. Many mammals can act as carriers for human lung flukes. However, their importance as reservoir hosts is largely unknown.

Clinical Features. Mild infections: fever and cough. Severe infections: dry cough, violent blood spitting, weight loss, fever, pleural effusion, chest pain.

Therapy: praziquantel, bithionol.

2.1.3. Liver Flukes

Infections are caused by *Clonorchis sinensis, Opisthorchis viverrini*, or *O. felineus.* Infections with the cosmopolitan liver flukes of ruminants (*Fasciola hepatica, Dicrocoelium dendriticum*) may be important locally.

Distribution. *C. sinensis:* Korea, China, Southeast Asia, Japan, Taiwan, Philippines, Indonesia. *O. viverrini:* Laos, Thailand, Indian subcontinent. *O. felineus:* USSR and some European foci. About 50 million people are infected [15].

Development. Adult *C. sinensis* and *O. viverrini* (10 – 20 and 8 – 12 mm long, respectively) live in the intrahepatic bile ducts. Their eggs contain fully developed larvae, the miracidia, which are voided with the feces and ingested by aquatic snails. The parasitic larva (miracidium) hatches in the intestine of the snail, enters its tissues, and starts multiplying. After some time cercariae are released into the water. They infect cyprinid fish percutaneously, encyst in the fish musculature, and become infective metacercariae in about four weeks. Infection is acquired through the consumption of raw or improperly cooked fish. The young flukes excyst in the small intestine, migrate into the bile ducts, and mature within four weeks. They may then live for several decades. Dogs, cats, and pigs are important reservoir hosts.

Clinical Features. Mild infections: generally asymptomatic. Severe infections: dilation and fibrosis of the bile ducts, hemorrhage, digestive disturbances. Chronic infections: epigastric pain, anorexia, cholangitis, liver necrosis, bile duct and pancreas carcinoma, intrahepatic calculi.

Therapy: praziquantel, bithionol.

2.1.4. Giant Intestinal Fluke (*Fasciolopsis buski*)

Distribution. South China, Southeast Asia, Indonesia, and the Indian subcontinent. Two to ten million people are infected [15].

Development. The eggs of the giant intestinal fluke are voided with the feces. In fresh water a miracidium develops in 3 – 7 weeks, hatches, locates an aquatic snail, penetrates it, and begins multiplying. After 4 – 8 weeks cercariae begin to be released, and they encyst as metacercariae on aquatic plants. Infection is acquired when such plants (water caltrop, *Trapa* spp.; water chestnut, *Eleocharis* sp.) are peeled with the teeth. The metacercarial cysts are thus freed, are swallowed, and develop to sexually mature flukes (5 – 7 cm long) in the small intestine within about three months. Pigs, dogs, and rabbits are important reservoir hosts.

Clinical features. Mild infections are often asymptomatic. In severe infections there are generalized toxic and allergic symptoms, diarrhea, weakness, malabsorption, edema, ascites, abdominal pain, and gastrointestinal hemorrhage.

Therapy: praziquantel, niclosamide.

2.1.5. Lesser Intestinal Flukes

Several species of these small flukes (2 – 7 mm in length) are found in humans: *Echinostoma ilocanum, E. lindoense, Heterophyes heterophyes,*

Metagonimus yokogawai, Gastrodiscoides hominis, Watsonius watsoni.

Distribution. East and South Asia, West Africa, Mediterranean countries, especially Egypt (*Heterophyes*). About 16 million people are infected [13].

Development. These parasites live in the small intestine. They have typical trematode life cycles, with a single intermediate host (*Gastrodiscoides* and *Watsonius*; metacercariae on aquatic plants) or two intermediate hosts. The second host can be fish (*Heterophyes* and *Metagonimus*) or snails and mussels (*Echinostoma*). Infection is acquired through the consumption of raw or improperly cooked plants, fish, or mollusks. The lesser intestinal flukes live and produce eggs for several months.

Clinical Features. Mild infections are often asymptomatic or show only nonspecific symptoms: diarrhea, abdominal discomfort, and headache.

Therapy: praziquantel, thiabendazole, niclosamide, bephenium.

2.2. Cestodes (Tapeworms)

2.2.1. Broad or Fish Tapeworm (*Diphyllobothrium latum*)

Distribution. Northern temperate zone (Finland, USSR, Alaska, Canada; small foci in northern Italy, Switzerland, and Chile). *D. pacificum:* coastal regions of Peru. There are several related species that infect humans (e.g., *Diplogonoporus* in Japan). About 2 – 9 million people are infected [6].

Development. *D. latum* is a typical ribbonlike tapeworm of extreme length (up to 10 – 15 m) living in the small intestine. After the eggs are voided with the feces into fresh water, a larva, the coracidium, is released. Larvae are ingested by small copepod crustaceans. When these are ingested by cyprinid fish, the parasites penetrate into the musculature of the fish and develop into the infective stage, the plerocercoid. Humans acquire the infection by consuming raw or improperly cooked fish.

Clinical Features. Infections are generally asymptomatic. When they do appear, symptoms include abdominal discomfort, dizziness, and fatigue. About 4 % of those infected develop clinical symptoms of a vitamin B_{12} deficiency indistinguishable from pernicious anemia.

Therapy: praziquantel, niclosamide.

Several related species that cannot develop to maturity in humans, but do so in other mammals, may infect humans accidentally. They can proliferate and cause painful inflammatory tissue reactions. This rare condition, sparganosis, can be treated only by surgical removal. Praziquantel may prove to be effective.

2.2.2. Beef Tapeworm (*Taenia saginata* = *Taeniarhynchus saginatus*)

Distribution. Cosmopolitan, about 77 million people are infected [13].

Development. Adult beef tapeworms generally measure 3 – 10 m. Each day about ten proglottids of the ribbonlike body, the strobila, are shed, containing several hundred thousand eggs. When these eggs are ingested by cattle, a larva, the oncosphere, hatches from the egg and reaches the musculature (presumably by the circulatory system), where an infective larva, the cysticercus, develops within 10 – 12 weeks. Cattle may carry infective cysticerci in their muscles up to three years. Humans acquire the infection by consuming raw or improperly cooked beef. Within 3 – 4 months the parasite reaches maturity in the intestine. It may live up to twenty years.

Clinical Features. Infections often go unnoticed. In symptomatic cases: vague abdominal pain, nausea, weakness, loss of weight, increased or decreased appetite, headache.

Therapy: niclosamide, praziquantel, mebendazole.

2.2.3. Pork Tapeworm (*Taenia solium*)

Distribution. Almost cosmopolitan, wherever people eat improperly cooked pork. About 3 – 5 million people are infected with adult *T. solium* [15].

Development, clinical features, and therapy are all similar to those of the beef tapeworm.

Cysticercosis. Although pigs usually are the intermediate hosts in the life cycle of the pork tapeworm, humans can also become infected with the larvae of this parasite by ingesting the eggs, which remain viable in the soil for many weeks. The resulting infection, *cysticercosis,* occurs in Central and South America, South and East Asia, East Africa, and eastern Europe and may afflict more people than are infected with the adult tapeworm [15]. This condition is often accompanied by myositis and high fever. When the brain is involved (neurocysticercosis, developing in 1 – 5 % of the infected persons), this condition is often accompanied by epileptic attacks, meningoencephalitis, and intracranial hypertension. In South America alone, there are about 350000 cases [16].

Formerly, the only *treatment* for cysticercosis was surgical removal of larval cysts.

Therapy: praziquantel, niclosamide, mebendazole.

2.2.4. Dwarf Tapeworm (*Hymenolepis nana*)

Distribution. Cosmopolitan, about 45 – 50 million people, mainly children, are infected [15].

Development. The dwarf tapeworm is unique in that the adult can develop following ingestion of the eggs by humans. The larva, the oncosphere, that hatches from the egg develops to the cysticercoid stage in the intestinal mucosa within 4 – 7 days. Alternatively, infections may be acquired through accidental ingestion of grain beetles containing infective cysticercoid larvae. The adult tapeworm, developing from the cysticercoid to maturity in 1 – 2 weeks, grows up to 40 cm long.

Clinical Features. Mild infections are asymptomatic. Severe infections: abdominal pain, diarrhea, headache, dizziness, anorexia.

Therapy: praziquantel, niclosamide.

2.2.5. Dog Tapeworm (*Echinococcus granulosus, E. multilocularis*)

Distribution. *E. granulosus:* cosmopolitan, especially in sheep-rearing countries. *E. multilocularis:* small foci in North, Central, South, and East Europe, northern United States. About 100000 people are infected [15].

Development. Many mammals (*humans,* sheep, goats, cattle, pigs, dromedaries, horses, rodents) serve as intermediate hosts of these small (3 – 5 mm long) tapeworms that live as adults in the intestine of dogs or other canines. When humans ingest the eggs, larvae hatch, penetrate the gut wall, and generally lodge in the liver or lungs, although all other tissues can also be affected. The larvae grow extremely quickly, forming unilocular (*E. granulosus*) or multilocular (*E. multilocularis*) cysts. The cysts contain many protoscolices, which mature to adult worms when offal containing cystic material is fed to dogs.

Clinical Features. Cysts are often asymptomatic until they grow large, but then the symptoms become progressively severe. Symptoms depend on the location of the cysts. Liver cysts can cause symptoms resembling a mucoid carcinoma. Lung cysts give rise to coughing and chest pain, whereas cerebral cysts can cause serious neurological damage.

Therapy: surgical removal, in inoperable cases, mebendazole, which does not kill the parasite but arrests the progressive course of the disease.

2.3. Nematodes (Roundworms)

The parasitic nematodes are conveniently divided into two groups: the intestinal nematodes, including the roundworm, hookworm, pinworm, whipworm, threadworm, and trichina, and the

extra-intestinal nematodes, including the guinea worm and several species of filariae. The second group live in the tissues of fluids of the body and require an arthropod vector for the completion of their life cycles.

2.3.1. Roundworm (*Ascaris lumbricoides*)

Distribution. Cosmopolitan, common in humid tropical climates. About 1300 million people are infected [13].

Development. This large parasite (20 – 40 cm long) lives in the small intestine and feeds on gut contents. Each female produces over 200000 eggs per day. The eggs are passed with the feces, become infective in about one month, and can remain infective in the soil for up to fifteen years. After infective eggs have been ingested, e.g., with vegetables, larvae are released, penetrate the duodenal wall, and migrate through the liver and the blood vessels to the lungs. There they break through the alveolae into the air passages, ascend the trachea, are swallowed, and attain maturity in the gut about 8 – 10 weeks later. The females live about one year.

Clinical Features. The majority of infections are light and asymptomatic. Larvae in the lungs may cause pneumonitis (cough, wheezing, dyspnea). Complications are intestinal obstruction and worm irritation after medication, causing hepatic duct obstruction, appendicitis, and intestinal perforation.

Therapy: levamisole, pyrantel, piperazine, mebendazole, thiabendazole, albendazole. Ivermectin is in clinical development

2.3.2. Hookworms

Two species, *Ancylostoma duodenale* (common or Old World hookworm) and *Necator americanus* (American or New World hookworm), parasitize humans, causing one of the most important tropical diseases.

Distribution. *N. americanus:* the Americas, tropical Africa, South and East Asia. *A. duodenale:* from the Mediterranean through India to China and Southeast Asia, Brazil. *A. ceylanicum* is of local importance. About 930 million people are infected and about 60000 deaths are caused by hookworm infections each year [13], [14].

Development. The mature worms ($\approx$ 10 mm long) attach to the gut mucosa and suck blood (*Necator* 5 – 10 times less than *Ancylostoma*). Each female produces 10000 (*Necator*) or 20000 (*Ancylostoma*) eggs per day. These are voided with the feces and release larvae that develop in the soil within seven days to the infective stage, which is able to survive up to one month. Infection is percutaneous (*Ancylostoma* also oral). The larvae migrate to the heart and lungs. Subsequent development is as described for the roundworm (*Ascaris*).

Clinical Features. Mild, cutaneous reaction to the infective larvae and lung reactions (cough, wheezing) to the migratory larvae. Chronic phase: epigastric pain, tenderness, peptic symptoms, iron-deficiency anemia, protein-loss enteropathy, hypoproteinemia. Severe infections: high-output cardiac-failure dyspnea, edema. The symptoms of hookworm infections are often aggravated by malnutrition and secondary bacterial or other parasitic infections.

Therapy: pyrantel, thiabendazole, mebendazole, bephenium, levamisole, tetrachloroethylene, bitoscanate.

2.3.3. Pinworm (*Enterobius vermicularis*)

Distribution. Cosmopolitan, common in the developed countries of the northern hemisphere. About 350 million people are infected [13].

Development. The pinworm, which is more commensal than parasitic, lives in the lumen of the colon and feeds on gut contents. Gravid females migrate out of the anus to deposit their eggs (5000 – 10000), which are fully embryonated and infective within a few hours. Infections and reinfections are acquired through accidental oral intake of eggs. Larval development occurs in the intestine within about six weeks.

Clinical features. Perianal itching. Complications caused by ectopic migration are rare.

Therapy: pyrantel, piperazine, mebendazole, pyrvinium.

2.3.4. Whipworm (*Trichuris trichiura*)

Distribution. Cosmopolitan, especially common in warm and humid climates. About 690 million people are infected [13].

Development. The parasites (3 – 5 mm long) live attached to the cecal mucosa. The eggs are voided with the feces and require a period of at least three weeks to become infective. Eggs remain infective in the soil for over one year. The larvae hatch after eggs have been ingested, e.g., with vegetables or dirt, and mature in the intestine within 2 – 3 months. The adults live 3 – 10 years.

Clinical Features. Infections are generally asymptomatic; occasionally, abdominal discomfort and, in children, diarrhea occur.

Therapy: oxantel, mebendazole, thiabendazole, albendazole. Ivermectin is in clinical development.

2.3.5. Threadworm (*Strongyloides stercoralis*)

Distribution. Tropical and subtropical climates, extending into southern Europe and southern United States. About 80 million people are infected [15].

Development. The female worms live attached to the mucosa of the small intestine, on which they feed. They produce, usually parthogenetically, larvae that are infective shortly after passage of the feces and infect humans percutaneously. Their further development to maturity takes about 17 days and resembles that of hookworms. Autoinfection is possible when larvae mature in the gut and infections may be perpetuated up to twenty years. When environmental conditions are favorable, threadworms can exist for some time as free-living nematodes, completing several generations in the soil.

Clinical Features. Infections may be asymptomatic, but gastrointestinal symptoms (diarrhea, katarrhal enteritis, central epigastric pain) are common. In cases of severe infection the prognosis is poor, and the mortality rate is high.

Therapy: thiabendazole, mebendazole, albendazole. Ivermectin is in clinical development.

Infections with other nematodes (e.g., *Angiostrongylus, Anisakis, Capillaria, Dipetalonema, Gnathostoma, Oesophagostomum, Trichostrongylus*) are less important and generally respond to thiabendazole and/or mebendazole. Thiabendazole is also effective against migrating larvae of nematodes that normally mature in other mammals (cutaneous and visceral larva migrans). About 5 – 10 million people are infected [15].

2.3.6. Trichina (*Trichinella spiralis*)

Distribution. Cosmopolitan except Australia. About 46 million people are infected [13].

Development. Infection is initiated by the consumption of raw or improperly cooked meat (especially pork) containing the encysted larvae. These excyst in the gut and mature to adult worms in the gut wall within 3 – 4 days. Sexes mate in the gut lumen, and each female then produces up to 2000 larvae within two weeks. The larvae penetrate the gut wall, enter the lymphatic vessels, disseminate via the circulatory system throughout the body, develop in the muscle, and finally encyst in muscle fibers.

Clinical Features. Infections are generally asymptomatic during the intestinal phase. The muscle infection causes myositis, eosinophilia, leukocytosis, and occasional fever.

Therapy: thiabendazole, mebendazole, albendazole. Ivermectin is in clinical development.

2.3.7. Larvae of Toxocara canis, Toxocara cati, Ancylostoma brasiliense and Ancylostoma caninum

Distribution. Cosmopolitan.

Development. Humans, especially children are the "wrong hosts" for these larvae. Human infection are rare and occur through oral ingestion of larvae following contact with contaminated faeces of *Toxocara canis* infected dogs or *T. cati* infected cats or with contaminated soil, sand, or playgrounds. The larvae of *Ancylostoma brasiliense* or *Ancylostoma caninum* penetrate the skin ("creeping eruption") following contact with contaminated dog or cat faeces.

Clinical Features. The disease is determined by the site where larvae reside. Larvae migrantes viscerales of *T. canis* and *cati* are mostly observed in liver and lungs, occasionally in eyes or brain. Larvae of *Ancylostoma brasiliense* are called larvae migrantes cutaneae, larvae of *Ancylostoma caninum* are called epidermis larvae.

Therapy: Thiabendazole, diethylcarbamazine.

2.3.8. Guinea Worm (*Dracunculus medinensis*)

Distribution. Focally in West, North, and East Africa, the Middle East, from Iran through Pakistan into Indonesia. About 50 – 100 million people are infected [15].

Development. The parasites develop in the body cavity and deeper connective tissue for about 12 – 14 months. The mature females then migrate to the subcutaneous tissues, mainly of the legs. There they cause blisters and ulceration, through which large numbers of larvae are discharged when the skin contacts water. If taken up by copepod crustaceans, the larvae develop in the body cavity and become infective in 2 – 3 weeks. Humans acquire the infection by drinking water containing infected copepods. Dogs are known reservoir hosts.

Clinical Features. Infections are usually asymptomatic for one year. Specific signs of the disease are painful blisters and ulcerations of the skin where the females penetrate the skin. Secondary bacterial infections often cause additional complications.

Therapy: Careful mechanical extraction, thiabendazole. Both niridazole and metronidazole reduce the inflammatory tissue reactions without affecting the parasites.

2.3.9. Lymphatic Filariae (*Wuchereria bancrofti, Brugia malayi*)

Distribution. *Wuchereria:* throughout the tropics and subtropics. *Brugia malayi:* from India to Korea and Indonesia. *B. timori:* Timor. About 380 million people are infected [13].

Development. The adult parasites, the macrofilariae, live in lymphatic ducts and nodes. The females deposit eggs, from which larvae, the microfilariae, hatch. The larvae are distributed with the blood throughout the body but congregate in the peripheral vasculature during the night. From there they are taken up by mosquitoes and mature to infective larvae in their musculature. They then enter the proboscis of the mosquito, and when it takes another blood meal they enter the human skin through the puncture. In the new host, they mature and mate, and the females start producing the next generation of microfilariae.

Clinical Features. Infections generally are asymptomatic, but chronic lesions, presumably of an allergic nature, can cause lymphangitis and obstruction of lymphatic drainage, leading to elephantiasis and hydrocele.

Therapy: diethylcarbamazine.

2.3.10. Convoluted Filaria (*Onchocerca volvulus*)

Distribution. Africa south of the Sahara, Yemen, Mexico to Venezuela. About 40 million people are infected [14].

Development. The adult worms (macrofilariae) live in subcutaneous nodules. The larvae (microfilariae) collect in the superficial layers of the skin and are transmitted by black flies (*Simulium*), in which they develop to the infective stage. Humans are infected when the fly takes another blood meal.

Clinical Features. Microfilariae may cause chronic cutaneous lesions. Invasion of the eye: conjunctivitis, keratitis, iridocyclitis, chorioretinitis, glaucoma, opacity of the lens, leading to optic atrophy and blindness.

Therapy: diethylcarbamazine (microfilaricidal), suramin (macrofilaricidal), ivermectin (microfilacricidal).

2.3.11. African Eye Worm (*Loa loa*)

Distribution. Rain forests of West and Central Africa. More than 10 million people are infected [15].

Development. The adult parasites (macrofilariae) actively migrate through the subcutaneous tissue. The microfilariae are found in the peripheral blood during the day and are transmitted by flies (*Chrysops*).

Clinical Features. Allergic swelling of the skin.

Therapy: diethylcarbamazine.

3. Anthelmintic Drugs

The anthelmintic drugs are listed alphabetically. The *mechanism of action* of the older anthelmintics can be found in [17]. The numerous *trade names* for the older drugs are given for some twenty countries in [18]. A number of trade names are given in [19]. The national drug directories can also be consulted.

Albendazole [*54965–21–8*], methyl-(5-(propylthio)-2-benzimidazol)carbamate, $C_{12}H_{15}N_3O_2S$, M_r 265.3, *mp* 208 – 210 °C, is a colorless crystalline solid. It is insoluble in water, sparingly soluble in dimethylformamide or dimethylsulfoxide, and soluble in formic acid.

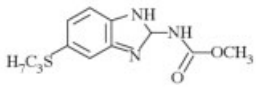

The compound is synthesized in a five step synthesis. 4-chloro-2-nitroaniline is first acetylated with acetic anhydride to give 1-acetamido-4-chloro-2-nitrobenzene. Treatment of this intermediate compound with potassium thiocyanate furnishes the key intermediate 1-acetamido-2-nitro-4-thiocyanatobenzene. Conversion of the thiocyanato group into the required *n*-propylthio analogue, with a simultaneous conversion of the acetamido group to the free amine, is effected by treating the last intermediate with 1-bromopropane in the presence of a base. Further reduction of the nitro group provides the diamine, which is subsequently ring closed to albendazole [20].

Albendazole is used against nematode infections: *Ascaris*, *Necator, Ancylostoma*, *Trichuris, Enterobius*, and systemic nematodes such as *Trichinella spiralis*, *Gnathostoma spinigerum*, and larval *Angiostrongylus cantonensis*. In addition it is used against the larval stages of the cestodes *Echinococcus granulosus* and *E. multilocularis* and for treatment of neurocysticercosis caused by *Taenia solium* [21].

Occasional *adverse effects* include fever, reversible leucopenia, headache, nausea, dizziness, sleeplessness, epigastric pain and diarrhoea. When higher dosages are given over a long period elevated levels of transaminases are observed, seldom hepatitis and occasionally alopecia. Contraindications are pregnancy and liver cirrhosis. Albendazole interferes selectively with the polymerization of β-tubulin to microtubuli. The microtubules play a crucial role in several important cell functions such as material transport within cells and neurotransmission [22].

Trade names: Eskadole (Smith Kline Beecham), Valbazen (Pfizer).

Bephenium [*3818-50-6*], benzyldimethyl-(2-phenoxyethyl)ammonium 3-hydroxy-2-naphthoate, $C_{28}H_{29}NO_4$, M_r 443.5, *mp* 170 – 171 °C, is a bitter tasting, odorless, yellow crystalline solid. It is insoluble in water and slightly soluble in ethanol.

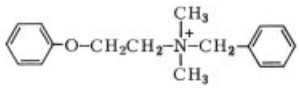

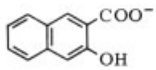

The compound is synthesized by the reaction of dimethyl(2-phenoxyethyl)amine with benzyl halide [23].

Bephenium can be used against *Ascaris, Ancylostoma*, and *Trichostrongylus* infections; it is less effective against *Necator.* Occasional adverse reactions occur: nausea, diarrhea, vomiting, headache, vertigo.

Trade names: Alcopar(a) (Wellcome, Tanabe); Befen (Andromaco); Lecibis (Andromaco, Columbia); Debephenium (ICN Usafarma); Fedal Uncin (Hosbon-Fedal); Hebe (Farmacon); Befeval (Valmorca); Befenium (Laquifa).

Bithionol [*97-18-7*], 2,2′-thiobis(4,6-dichlorophenol), $C_{12}H_6Cl_4O_2S$, M_r 356.1, *mp* 188 °C, is a tasteless, colorless, crystalline solid. It is insoluble in water but soluble in ethanol.

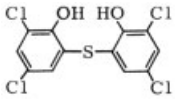

Bithionol is synthesized by the reaction of chlorophenol with sulfur dichloride and sulfuryl chloride [24].

Bithionol can be used against liver, lung, and intestinal flukes. Frequent adverse reactions are abdominal pain, diarrhea, anorexia, nausea, vomiting, dizziness, headache, skin rashes, urticaria. Hepatic and renal involvement are rare complications.

Trade names: Actamer (Monsanto); Bitin (Tanabe); Bitin S (Tanabe), which contains the sulfoxide of bithionol.

Diethylcarbamazine [*1642-54-2*], *N,N*-diethyl-4-methyl-1-piperazinecarboxamide citrate, $C_{16}H_{29}N_3O_8$, M_r 391.4, *mp*141 – 143 °C, is a white, odorless, crystalline solid with a bitter acid taste. It is freely soluble in water or hot ethanol but insoluble in acetone, chloroform, ether, or dioxane.

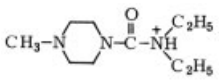

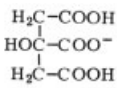

The compound is synthesized by the reaction of diethylcarbamoyl chloride with methylpiperazine [25].

Diethylcarbamazine dihydrogen citrate is used against the microfilarial stage of filariae. It also kills the adult filariae of *Wuchereria, Brugia,* and *Loa.* Occasional adverse reactions are anorexia, nausea, vomiting, headache, and drowsiness. Allergic reactions occur, especially in onchocercosis patients: intense pruritus, edema, lymphadenitis, dermatitis, papular rash, fever, tachycardia.

The most likely mechanism of action is a rapid effect on the surface of the microfilariae. Previously hidden antigenic determinants become accessible to the immune system of the host, which then eliminates the microfilariae from the circulating blood [8].

Trade names: Hetrazan and Hetrazan 1949 (Lederle); Banocide (Wellcome); Notezine (Specia); Supatonin (Tanabe); Carbilazine (Willows Francis); Caricide (Am. Cyanamid).

Ivermectin (synonyms: 5-*O*-Demethyl-22,23-dihydroavermectin A_{1a}, 22,23-Dihydroabamectin, 22,23-Dihydroavermectin B_1) *Ivermectin* B_{1a} [*71827–03–7*], $C_{48}H_{74}O_{15}$, M_r 891.1, *mp* 155 – 157 °C; ivermectin B_{1b} [*70209–81–3*], $C_{47}H_{72}O_{15}$, M_r 877.1, is a white powder and forms crystals from ethanol – water. It is soluble in methyl ethyl ketone, propylene glycol, and poly(ethylene glycol).

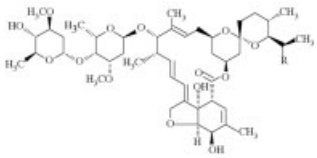

Ivermectin

Ivermectin is a mixture of at least 80 % ivermectin B_{1a} (R = CH_3) and not more than 20 % ivermectin B_{1b} (R = H).

Ivermectin is a semisynthetic compound derived from abamectin (avermectin B_1) a fermentation product of *Streptomyces avermitilis* by selective hydration with chlorotris(triphenylphosphin)rhodium (I) (Wilkinson catalyst) [26–28].

In *human medicine* ivermectin is the drug of choice for onchocerciasis (river blindness) although it only kills *Onchocerca volvulus* micro-

filariae and not the adults. Although the drug has no curative effect, a single dose of 0.2 mg per kilogram body weight every 6 – 12 months controls the disease [29]. The drug is under clinical development for treatment of intestinal nematode infections (*Ascaris*, *Enterobius*, *Trichuris*, *Strongyloides*, and *Trichinella spiralis*).

In *veterinary medicine* ivermectin is used in cattle against a wide variety of nematodes and arthropods. It is inactive against cestodes and trematodes.

Occasional *adverse effects* in humans include headache, fever, pruritus, oedema, and arthralgies. However, as ivermectin is used only 1 – 2 times a year at low dosages, side effects are rare. Veterinary usage has shown that collies are particular sensitive in developing gabergic side effects (CNS depression with dizziness, ataxia, tremor, salivation, mydriasis, and coma with exitus).

Presumably, ivermectin exerts its mode of action by interfering with GABA, glutamate or glycine gated chloride channels, all of which are inhibitory neurotransmitters [30–32]. This action leads to an increased permeation of chloride ions, followed by a hyperpolarization of nerve-muscle membranes and a flaccid paralysis of the parasites. The action of ivermectin can be antagonised by bicuculline and picrotoxin [33].

Trade names: Mectizan in France (MSD), Ivomec (MSDAGVET).

Levamisole [*16595-80-5*], L(−)−2,3,5,6-tetrahydro-6-phenylimidazo[2,1-*b*]thiazole hydrochloride, $C_{11}H_{12}N_2S \cdot HCl$, M_r 240.8, *mp*227 – 229 °C, is a colorless, odorless, crystalline solid. It is soluble in water and slightly soluble in chloroform.

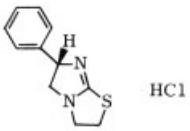

The compound is synthesized by the reaction of 4-phenyl-2-thioimidazoline with dibromoethane [34].

Levamisole is used against roundworm (*Ascaris*) infections. It is less effective against *Ancylostoma, Necator*, or *Strongyloides.* Occasional adverse reactions are nausea, vomiting, abdominal discomfort, headache, dizziness, hypertension.

Levamisole causes rapid spastic contraction in nematodes by persistently depolarizing the muscle-cell membrane. It also acts as a ganglion-stimulating compound [35]. The paralyzed nematodes are then eliminated from the intestine. Levamisole is likely to be hydrolyzed to L(−)−2-oxo-3-(2-mercaptoethyl)−5-phenylimidazoline under alkaline conditions. This metabolite is a strong and stereospecific inhibitor of the enzyme fumarate reductase through its interaction with −SH groups of the active center.

Trade names: Ascaridil, Decaris, and Stimamizol (Johnson & Johnson, Janssen); Ketrax (ICI); Solaskil (Specia).

Mebendazole [*31431-39-7*], methyl(5-benzoyl-1*H*-benzimidazol-2-yl)carbamate, $C_{16}H_{13}N_3O_3$, M_r 295.3, *mp* 288.5 °C, is an off-white amorphous powder. It is insoluble in water, sparingly soluble in dimethylformamide or dimethyl sulfoxide, and soluble in formic acid.

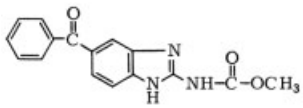

The compound is synthesized by the reaction of 3,4-diaminobenzophenone hydrochloride with *N*-carboxymethyl-*S*-methylisothiourea [36].

Mebendazole is used against nematode infections: *Ascaris, Trichuris, Enterobius, Ancylostoma, Necator, Capillaria.* It is also effective against tapeworms (*Taenia*). Occasionally adverse reactions, abdominal pain and diarrhea, occur. A rare complication is leukopenia. Pregnancy is a contraindication.

Mebendazole interferes with glucose uptake by nematodes and cestodes in vivo and in vitro. This interference occurs by the selective interaction with intracellular tubulin and subsequent inhibition of the assembly of microtubules. The microtubules participate in several important cell functions, e.g., the transport of materials within cells [37].

Trade name [18]: Vermox (Janssen, Ortho).

Metrifonate [*52-68-6*], *O,O*-dimethyl-2,2,2-trichloro-1-hydroxyethylphosphonate, $C_4H_8Cl_3O_4P$, M_r 257.4, *mp* 83 – 84 °C, is a colorless, crystalline solid. It is soluble in water, chloroform, or ether and very slightly soluble in pentane or hexane.

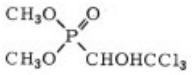

The compound is synthesized by reaction of chloral with dimethyl phosphite [38].

Metrifonate is used against the blood fluke *Schistosoma haematobium.* The occasional adverse reactions are nausea, vomiting, bronchospasm, weakness, diarrhea, and abdominal pain.

In vivo, metrifonate rapidly rearranges to *O,O*-dimethyl *O*-(2,2-dichlorovinyl) phosphate, which is a potent inhibitor of schistosome acetylcholinesterase. This paralyzes the parasites because of the accumulation of acetylcholine, which functions as inhibitory transmitter [39].

Trade name: Bilarcil (Bayer).

Niclosamide [*50-65-7*], 2′,5-dichloro-4′-nitrosalicylanilide, $C_{13}H_8Cl_2N_2O_4$, M_r 327.5, *mp* 225 – 230 °C, is a tasteless, pale yellow, crystalline solid. It is almost insoluble in water; sparingly soluble in ethanol, chloroform, or ether; and soluble in acetone.

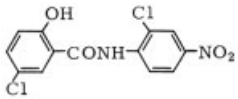

The compound is synthesized by condensation of 5-chlorosalicylic acid with 2-chloro-4-nitroaniline [40].

Niclosamide is used against all kinds of tapeworms. It is also effective against intestinal flukes. Occasional adverse reactions are abdominal pain and nausea.

Niclosamide causes uncoupling of oxidative phosphorylation and respiration. It also activates ATPases, and this disturbs the energy balance of cestodes, which degrade endogenous glycogen, and impairs glucose uptake. Protein synthesis is also affected; e.g., a trypsin inhibitor is no longer formed, which facilitates proteolytic attack of the tapeworm by host digestive enzymes [41].

Trade names: Yomesan, Cestocide, and Nasemo (Bayer); Niclocide (Miles); Trédémine (Roger Bellon); Radeverm (VEB Arzneimittelwerk Dresden); Sulqui (Ella); Atenase (ICN Usafarma); Teniarene (Amsa); Teniamida (Bial); Fedal Telmin (Hosbon-Fedal); Copharten (Cophar). Some preparations contain niclosamide monohydrate.

Niridazole [*61-57-4*], 1-(5-nitro-2-thiazolyl)-2-imidazolidinone, $C_6H_6N_4O_3S$, M_r 214.2, *mp* 260 – 262 °C, is a yellow, crystalline solid. It is almost insoluble in water and most organic solvents, but it is soluble in dimethylformamide.

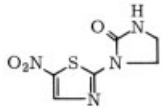

The compound is synthesized by reaction of 2-amino-5-nitrothiazole and β-chloroethylisocyanate [42].

Niridazole can be used against blood flukes, especially *Schistosoma haematobium.* Tolerance and efficacy are reduced in *S. mansoni* and especially in *S. japonicum* infections. It is also used in *Dracunculus* infections. Immunosuppression, vomiting, cramps, dizziness, and headache are among the frequent adverse reactions. Occasional adverse reactions are diarrhea, electrocardiographic changes, rash, insomnia, and paresthesia. Psychosis, hemolytic anemia, and convulsions are rare. Contraindications are impaired liver function, glucose-6-phosphatedehydrogenase deficiency, epilepsy, and severe heart diseases.

Niridazole causes a depletion of glycogen in schistosomes by inducing a reduced rate of conversion of active glycogen phosphorylase to its inactive form. This is achieved through inhibition of the enzyme phosphorylase phosphatase, which normally inactivates glycogen phosphorylase. It is possible that the active moiety is not niridazole, but its 5-imino analog, which can be formed by schistosomes in vitro under anaerobic conditions [43].

Trade names: Ambilhar (Ciba-Geigy); Ambilhar Ciba (Biogalenica Q Farm).

Oxamniquine [*21738-42-1*], 1,2,3,4-tetrahydro-2-[(isopropylamino)methyl]−7-nitro-6-quinolinemethanol, $C_{14}H_{21}N_3O_3$, M_r 279.3, *mp* 151 – 152 °C, is a light orange, crystalline solid. It is slightly soluble in water and soluble in methanol, acetone, or chloroform.

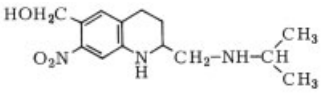

The compound is manufactured by fermentative oxidation of its 6-methyl analog by *Aspergillus scleroticum* [44].

Oxamniquine is used against the blood fluke *Schistosoma mansoni.* The adverse reactions observed are vertigo, vomiting, nausea, abdominal pain, anorexia, dizziness, drowsiness, and somnolence. Hallucinations are occasional, and epileptiform convulsions are rare. Intramuscular injection causes severe pain at the site of injection.

The antischistosomal effect of oxamniquine as a function of time correlates well with the activity of the enzyme ornithine-δ-transaminase. The enzyme may interfere with normal arginine formation and the maintenance of a normal nitrogen balance in the parasite [43].

Trade names: Mansil, Vansil, and Vancil (all Pfizer).

Oxantel [*68813-55-8*], (*E*)−3-[2-(1,4,5,6-tetrahydro-1-methyl-2-pyrimidinyl)ethenyl]-phenol 4,4′-methylenebis(3-hydroxy-2-naphthalenecarboxylate), $C_{36}H_{32}N_2O_7$, M_r 604.7, *mp* 207 – 208 °C (hydrochloride), is a yellow crystalline solid. It is practically insoluble in water.

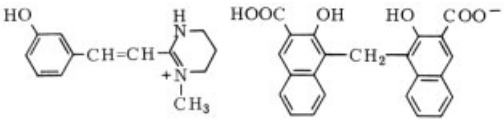

It is synthesized by condensation of 3-hydroxybenzaldehyde and 1,2-dimethyl-1,4,5,6-tetrahydropyrimidine in refluxing ethyl formate [45].

The cation is the active part of the compound. Oxantel is used against whipworm (*Trichuris*). No adverse reactions to this relatively new drug have been found, but it can be expected to behave much like the closely related pyrantel.

Trade name: Telopar (Pfizer).

Piperazine [*110-85-0*], piperazine hexahydrate, $C_4H_{10}N_2 \cdot 6\,H_2O$, M_r 194.2, *mp* 43 – 45 °C, is a colorless salty tasting crystalline solid. It is soluble in water and insoluble in ether.

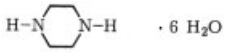

Piperazine is used against roundworms (*Ascaris*) and pinworms (*Enterobius*). Adverse reactions are nausea, vomiting, diarrhea, abdominal pain, headache, and paresthesia. Urticaria may occur occasionally. Contraindications are impaired renal or hepatic function. Various salts are also used: Calcium ethylenediaminetetraacetate (edetate), calcium citrate, citrate, tartrate, maleate, phosphate, adipate, sebacate, sulfate. Piperazine increases the resting potential of the somatic musculature of nematodes, especially in the syncytial region, by increasing the permeability of the membrane to chloride ions. This results in flaccid paralysis of the parasites, which are expelled from the intestine [17].

Trade names [18]: Uvilon (Bayer); Antepar 1953 (Wellcome); Helmazine (Midy); Multifuge (Glaxo); Neox (Rocador).

Praziquantel [*55268-74-1*], 2-cyclohexylcarbonyl-1,2,3,6,7,11 *b*-hexahydro-4*H*-pyrazino[2,1-*a*]isoquinolin-4-one, $C_{19}H_{24}N_2O_2$, M_r 312.4, *mp* 136 – 140 °C, is a bitter tasting, colorless, crystalline solid. It is sparingly soluble in water and soluble in ethanol, chloroform, or dimethyl sulfoxide.

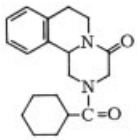

The compound is synthesized by reacting 1,2,3,6,7,11 *b*-hexahydropyrazino[2,1-*a*]isoquinolin-4-one with cyclohexylcarbonyl chloride in chloroform [46].

Praziquantel is used against all infections caused by trematode parasites (blood, liver, lung, and intestinal flukes) or adult tapeworms. It is also effective in cysticercosis. Occasional adverse reactions are epigastric pain, nausea, dizzi-

ness, diarrhea, and vomiting. Rare adverse reactions are skin manifestations, headache, and tiredness.

Praziquantel causes a very rapid increase in the permeability of membranes to divalent cations. In the musculature of trematodes and cestodes, this permeability increase results in a calcium concentration increase, which initiates rapid contracture and subsequent spastic contraction and paralysis. In the tegument, praziquantel causes rapid and progressive vacuolization, which leads to partial disintegration of the body surface of the parasites. This, in turn, renders them susceptible to attacks by the host defense system and its digestive enzymes [47].

Trade names: Biltricide (Bayer, Miles); Cestox, Cesol, and Cisticid (all E. Merck).

Pyrantel [*22204-24-6*], (*E*)−1,4,5,6-tetrahydro-1-methyl-2-[2-(2-thienyl)vinyl]-pyrimidinium 4,4′-methylenebis(3-hydroxy-2-naphthalenecarboxylate), $C_{34}H_{30}N_2O_6S$, M_r 594.7, *mp* 178 – 179 °C (free base), is a tasteless, odorless, yellow, crystalline solid. It is insoluble in water, slightly soluble in dimethylformamide, and soluble in dimethyl sulfoxide.

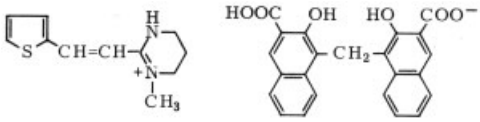

The compound is synthesized by reacting 2-thiophenealdehyde with 1,2-dimethyltetrahydropyrimidine [48].

The cation is the active part of the compound. Pyrantel is used against nematodes: *Ascaris, Enterobius, Ancylostoma*, and *Necator.* Adverse reactions are anorexia, abdominal cramps, diarrhea, nausea, vomiting. Occasional adverse reactions are headache, dizziness, drowsiness, and skin rash. Impaired hepatic function is a contraindication.

Pyrantel is a depolarizing neuromuscular blocking agent. In nematodes it causes a slowly developing contraction and paralysis. The immobilized parasites are then eliminated from the intestine. The neuromuscular junction of *Ascaris* is 100-fold more sensitive to pyrantel than to acetylcholine [49].

Trade names: Lombpiareu (Areu); Aut (Elea); Aguipiran (Aguilar); Tamoa (North Medicamenta); Antiminth (Roerig); Combantrin, Trilombrin, Cobantril, and Helmex (Pfizer); Tricocel and Piranver (ICN Usafarma); Verdal (Columbia); Perverme (Biofarma); Piranver and Piranver F (ICN Farmaceutica).

Pyrvinium [*3546-41-6*], 6-(dimethyl-amino)-2-[2-(2,5-dimethyl-1-phenyl-1*H*-pyrrol-3-yl)-ethenyl]−1-methylquinolinium 4,4′-methylenebis(3-hydroxy-2-naphthalenecarboxylate), $C_{75}H_{70}N_6O_6$, M_r 1151.4, *mp* 210 – 215 °C, is a tasteless, odorless, bright orange to black, crystalline solid. It is insoluble in water or ether, slightly soluble in chloroform, and soluble in glacial acetic acid.

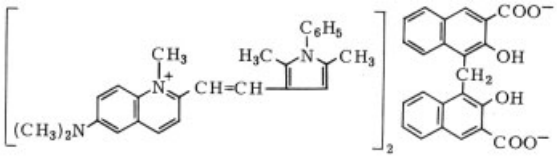

The compound is synthesized by the reaction of 1,2-dimethyl-6-dimethylaminoquinoline iodide with 2,5-dimethyl-1-phenyl-3-pyrrolaldehyde [50].

The cation is the active part of the compound. Pyrvinium is used against pinworms (*Enterobius*). Occasional adverse reactions are blood dyscrasia and folic acid deficiency. Rare adverse reactions are rash, vomiting, convulsion, shock, photosensitivity, and headache.

Trade names: Vanquil, Vanpar, Vanquin, Povan(yl) Molevac, Povan 1959, Paquil, Polyquil, and Povanyl (all Parke-Davis and Substantia).

Suramin [*129-46-4*], hexasodium 8,8′-[carbonylbis[imino-3,1-phenylenecarbonyl-imino(4-methyl-3,1-phenylene)carbonyl-imino]]bis-1,3,5-naphthalenetrisulfonate, $C_{51}H_{34}N_6Na_6O_{23}S_6$, M_r 1429.2, is a pinkish white, white, or cream-colored, odorless, hygroscopic powder with a slightly bitter taste. It is soluble in water and insoluble in chloroform or ether.

The compound is synthesized by condensation of 1-naphthylamine-4,6,8-trisulfonic acid with 3-nitro-4-methylbenzoyl chloride, reduction of the product, condensation with 3-nitrobenzoyl chloride, renewed reduction, and final condensation with phosgene [51].

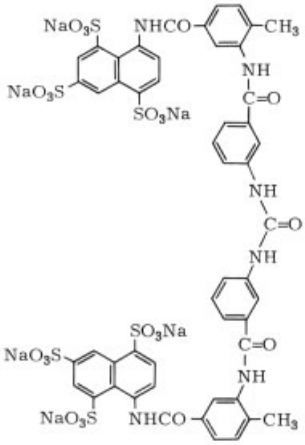

Suramin is used as a macrofilaricide against *Onchocerca.* Frequent adverse reactions are nausea, vomiting, colic, and urticaria. Occasionally, kidney damage, blood dyscrasias, shock, and optic atrophy occur. Allergic reactions to filarial proteins (pruritus, rash, fever, edema, hyperesthesia, photophobia, lachrymation) are also encountered.

The mechanism of action is not well understood. Suramin inhibits many different enzyme systems by binding to free cationic amino acid residues in the area of the active center. This appears to lead to a specific, but slight, interference with the DNA-RNA replication mechanism. The inhibition of protein kinase, an enzyme intimately involved in intracellular regulation, also may be involved in the effects of suramin [52].

Trade names: Germanin (Bayer); Antrypol (ICI); Moranyl (Specia).

Thiabendazole [*148-79-8*], 2-(4-thiazolyl)−1*H*-benzimidazole, $C_{10}H_7N_3S$, M_r 201.3, *mp* 304 – 305 °C, is a tasteless, colorless, crystalline solid. It is insoluble in water, slightly soluble in ethanol or chloroform, and soluble in acetone, dimethyl sulfoxide, or dimethylformamide.

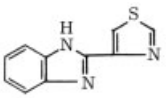

The compound is synthesized by heating 4-thiazolecarboxamide with *o*-phenylenediamine in polyphosphoric acid [53].

Thiabendazole is used against nematodes and nematode larvae: *Enterobius, Ascaris, Strongyloides, Necator, Trichinella, Ancylostoma, Capillaria, Dracunculus*, and *Trichostrongylus.* It is also effective against intestinal flukes. Frequent adverse reactions are dizziness, anorexia, vomiting, and epigastric distress. Diarrhea, fever, headache, skin rashes, hallucinations, and olfactory disturbances occur occasionally. Tinnitus, hypotension, and syncope are rare complications. The mechanism of action is not yet understood. Thiabendazole inhibits fumarate reductase, an enzyme essential for mitochondrial energy production in many nematodes, but this cannot be the primary or sole mechanism. Possibly thiabendazole, like mebendazole, affects tubulin polymerization, but there is no good experimental evidence for this [43].

Trade names: Mintezol and Minzolum (Merck Sharp & Dohme).

4. References

1. H. Loewe: "Anthelminthica," in G. Ehrhardt, R. Ruschig (eds.): *Arzneimittel, Entwicklung, Wirkung, Darstellung,* vol. 5, Verlag Chemie, Weinheim 1972, pp. 11 – 110.
2. R. B. Borrows: "Human and Veterinary Anthelmintics," *Progr. Drug. Res.* **17** (1973) 108 – 209.
3. P. J. Islip: "Anthelmintic Agents," in M. E. Wolff (ed.): *Burger's Medicinal Chemistry,* vol. 2, J. Wiley & Sons, New York 1979, pp. 481 – 530.
4. R. Cavier (ed.): "Chemotherapy of Helminthiasis," *Int. Encycl. Pharmacol. Ther.,* section 64, vol. 1, Pergamon, Oxford 1973.
5. "Intestinal Protozoan and Helminthic Infections," *WHO Tech. Rep. Ser.* 1981, no. 666.
6. "Parasitic Zoonoses," *WHO Tech. Rep. Ser.* 1979, no. 637.
7. M. A. Gemmell, P. P. Johnstone: "Cestodes," *Antibiot. Chemother.* (Basel) 30 (1981) 54 – 114.
8. F. Hawking: "Chemotherapy of Filariasis," *Antibiot. Chemother.* (Basel) 30 (1981) 135 – 162.
9. E. A. Malek (ed.): *Snail-Transmitted Parasitic Diseases,* CRC Press, Boca Raton, Florida, 1980.

10. A. E. R. Taylor, R. Muller (eds.): "The Relevance of Parasitology to Human Welfare Today," *Symp. Br. Soc. Parasitol.,* vol. 16, Blackwell Scientific Publications, Oxford 1978.
11. E. K. Markell, M. Voge: *Medical Parasitology,* Saunders, Philadelphia 1981.
12. P. E. C. Manson-Bahr, F. I. C. Apted: *Manson's Tropical Diseases,* Baillière Findall, London 1982.
13. W. Peters, in [10], pp. 25 – 40.
14. J. Walsh, K. S. Warren, *N. Engl. J. Med.* **301** (1979) 967 – 974.
15. D. Stürchler: *Endemiegebiete tropischer Infektionskrankheiten,* Huber, Bern 1981.
16. H. Schenone in A. Flisser et al. (eds.): *Cysticercosis, Present State of Knowledge and Perspectives,* Academic Press, New York 1982, pp. 25 – 38.
17. J. del Castillo in M. Florkin, B. T. Scheer (eds.): *Chemical Zoology,* **vol. 3,** Academic Press, New York 1969, pp. 521 – 554.
18. Chemindex International 1983, IMS A.G., Zug,Switzerland.
19. Index Nominum 1984, International Drug Directory, Laboratory of the Swiss Pharmaceutical Society, Zürich, Switzerland.
20. L. B. Townsend, D. S. Wise: "The synthesis and chemistry of certain anthelmintic benzimidazoles," *Parasitology Today* **6** (1990) 107 – 112.
21. G. C. Cook: "Use of benzimidazole chemotherapy in human helminthiases : indications and efficacy,″ *Parasitology Today* **6** (1990) 133 – 136.
22. E. Lacey: "Mode of action of benzimidazoles," *Parasitology Today* **6** (1990) 112 – 115.
23. Burroughs Wellcome, US 2918401, 1959 (C. Copp).
24. I. G. Farbenind., DE 583055, 1933 (F. Muth).
25. American Cyanamid, US 2467893, 1949 (S. Kushner, L. Brancone).
26. J. C. Chabala et al., *J. Med. Chem.* **23** (1980) 1134 – 1136.
27. H. Mrozik et al., *Tetrahedron Lett.* **23** (1982) 2377.
28. M. E. Jung, *Tetrahedron Lett.* **28** (1987) 5977.
29. H. Bradshaw, *Parasitology Today* **5** (1989) 63 – 64.
30. D. F. Cully et al., *Nature (London)* **371** (1994) 707 – 712.
31. D. F. Cully et al., *J. Biol. Chem.* **271** (1996) 20187 – 20191.
32. W. Forth in W. Forth, D. Henschler, W. Rummel, K. Starke, (eds): Anthelmintika, *Pharmakologie und Toxikologie,* Wissenschaftsverlag 1993, 711 – 712.
33. S. S. Pong, C. C. Wang, L .C. Fritz, *J. Neurochem.* **34** (1980) 351 – 358.
34. Janssen, US 3274209, 1966 (A. H. M. Raeymaekers, D. C. J. C. Thienpont, P. J. A. W. Demoen).
35. P. A. J. Janssen: "The Levamisole Story," *Prog. Drug Res.* **20** (1976) 347 – 383.
36. Janssen, US 3657267, 1972 (J. L. H. Van Gelder, L. F. C. Roevens, A. H. M. Raeymaekers).
37. H. Van den Bossche, F. Rochette, C. Hörig: "Mebendazole and Related Anthelmintics," *Adv. Pharmacol. Chemother.* **19** (1982) 67 – 128.
38. Bayer, US 2701225, 1955 (W. Lorenz).
39. B. Holmgren, I. Nordgren, M. Sandoz, A. Sundwall: "Metrifonate. Summary of Toxicological and Pharmacological Information Available," *Arch. Toxicol.* **41** (1978) 3 – 29.
40. Bayer, GB 824345, 1959.
41. P. Andrews, J. Thyssen, D. Lorke: "The Biology and Toxicology of Molluscicides, Bayluscide," *Pharmacol. Ther.* **19** (1983) 245 – 295.
42. Ciba, BE 632989, 1963.
43. P. Andrews, A. Haberkorn, H. Thomas: "Antiparasitic Drugs: Mechanism of Action, Pharmacokinetics, and in Vitro and in Vivo Assays of Drug Activity," in V. Lorian (ed.): *Antibiotics in Laboratory Medicine,* 2nd ed., Williams and Wilkins, Baltimore 1985, Chap. 9.
44. H. C. Richards, R. Forster, *Nature (London)* **222** (1969) 581 – 582.
45. Pfizer, ZA 6804589, 1967; FR 1584069, 1967 (J. W. McFarland).
46. E. Merck, US 4113867, 1978.
47. P. Andrews, H. Thomas, R. Pohlke, J. Seubert: "Praziquantel," *Med. Res. Rev.* **3** (1983) 147 – 200.
48. Pfizer, BE 658987, 1965.
49. M. L. Aubry, P. Cowell, M. J. Davey, S. Shevde: "Aspects of the Pharmacology of a New Anthelmintic: Pyrantel," *Br. J. Pharmacol.* **38** (1970) 332 – 344.
50. Eastman Kodak, US 2515912, 1950 (E. Van Lare, L. G. S. Brooker).
51. I. G. Farbenind., GB 224849, 1924.
52. F. Hawking: "Suramin: With Special Reference to Onchocerciasis," *Adv. Pharmacol. Chemother.* **15** (1978) 289 – 322.
53. Merck & Co., US 3017415, 1962 (L. H. Sarett, H. D. Brown).

Anthracene

Gerd Collin, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany
Hartmut Höke, Weinheim, Federal Republic of Germany
Jörg Talbiersky, Dorsten, Federal Republic of Germany

1. **Physical Properties** 1
2. **Chemical Properties** 1
3. **Production** 1
4. **Analysis** 2
5. **Uses** 2
6. **Toxicology and Environmental Aspects** 2
7. **References** 3

Anthracene [*120-12-7*], $C_{14}H_{10}$, was discovered in coal tar by J. Dumas and H. A. Laurent in 1832.

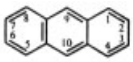

1. Physical Properties

Anthracene, $C_{14}H_{10}$, M_r 178.24, *mp* 218 °C, *bp* 340 °C (at 101.3 kPa), $d_4^{25} = 1.252$ colorless plates with blue-violet fluorescence, sublimes readily and forms a continuous series of mixed crystals with phenanthrene and carbazole in binary and ternary systems. Anthracene is slightly soluble in benzene, chloroform, and carbon disulfide; less soluble in diethyl ether and ethyl alcohol; and almost insoluble in water.

Specific heat capacity	1164 J/kg (at 24 °C)
Heat of fusion	162 kJ/kg
Heat of vaporization	294 kJ/kg
Heat of combustion	40110 kJ/kg (at 25 °C)

2. Chemical Properties

Anthracene is converted to dianthracene by irradiation with UV light in the absence of oxygen. In the presence of oxygen anthracene is converted to a peroxide by addition of oxygen to the 9,10-positions. Pyrocondensation gives 2,9-bianthryl and 9,10-dihydroanthracene [5]. Anthracene is readily hydrogenated to 9,10-dihydroanthracene. With homogeneous ruthenium catalysts, anthracene is hydrogenated to 1,2,3,4-tetrahydroanthracene [6]; further hydrogenation yields 1,2,3,4,5,6,7,8-octahydroanthracene. Oxidation of anthracene in the liquid or gas phase gives anthraquinone (→ Anthraquinone). Anthracene is primarily halogenated and nitrated in the 9- and 10-positions, and sulfonated in the 1- or 2-position, depending on the reaction conditions. Anthracene undergoes Diels – Alder addition of maleic anhydride and other dienophiles in the 9- and 10-positions.

With magnesium metal in THF anthracene undergoes an addition reaction to form a magnesium – anthracene adduct from which highly reactive magnesium can be obtained that can be converted to magnesium hydride and Grignard reagents [7].

3. Production

High-temperature coal tar contains, on average, 1.5 % of anthracene. Continuous tar distillation concentrates the anthracene in the fraction boiling between 300 and 360 °C (anthracene oil I). This oil amounts to 15 – 20 % of the raw tar and contains about 7 % anthracene; it is cooled to 20 – 30 °C and then centrifuged. About 10–15 % of the crude anthracene oil I is obtained as a yellow-green crystalline material with an anthracene content of 20 – 35 % (anthracene residues). These residues, which also contain 40 – 50 % phenanthrene and 12–20 %

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_343.pub2

carbazole, are crystallized or distilled. Distillation yields a crude product containing 45 – 55 % anthracene (50s anthracene). A 50s anthracene with an improved quality can be produced by two-step continuous column distillation [8, 9]. The crude 50s anthracene then is purified by recrystallization from selective solvents and by distillation. Repeated recrystallization from a three- to fourfold excess of pyridine produces anthracene (95 % purity) from the 50s anthracene in 80 % yield relative to the anthracene content. Other selective solvents suitable for recrystallization are acetone [10 – 14], acetophenone [15], benzene – methanol [16], dialkyl sulfoxides [17], dialkylformamides [18], *N*-methylpyrrolidone [19], dimethylacetamide [20], and propylene carbonate [21].

High yield of anthracene (96 %) and low energy consumption can be achieved by solvent-free, high-temperature melt crystallization of anthracene residues or anthracene oil I [22, 23].

Anthracene can be separated from the higher-boiling carbazole by distillation with lower-boiling hydrocarbon fractions as reflux medium [24, 25] or by azeotropic distillation with ethylene glycol [26, 27]. Azeotropic distillation also is used to separate the anthracene-accompanying tetracene and to obtain very pure anthracene for scintillation counting [28].

In principle, anthracene is also available from the high-boiling oil of coal liquefaction [29].

Carbazole-free crude anthracene can be prepared by distillation and crystallization of residues obtained from the pyrolysis of hydrocarbon fractions to olefins; however, the anthracene content of these pyrolysis oils generally is less than 1 % [30]. The anthracene yield can be increased by selective hydrodealkylation of the aromatic fraction containing methylphenanthrenes and methylanthracenes [31].

Anthracene can be synthesized by hydrogenation of phenanthrene, isomerization of the resulting *sym*-octahydrophenanthrene to *sym*-octahydroanthracene, and subsequent dehydrogenation [32 – 35]. A mixture of *sym*-octahydroanthracene and *sym*-octahydrophenanthrene is obtained by catalytic disproportionation of tetralin [36 – 38]. In addition, anthracene is formed from diphenylmethane in the presence of HF/BF_3 at 80 °C [39, 40] and by thermal reaction of *o*-methyldiphenylmethane at ca. 600 °C [41]. These syntheses have no commercial importance to date because sufficient anthracene is obtained from coal tar.

4. Analysis

Anthracene is determined quantitatively in the presence of other coal tar constituents by oxidation with chromic acid in glacial acetic acid to produce anthraquinone; by Diels – Alder reaction with maleic anhydride [42, 43]; by gas chromatography [44]; by UV spectroscopy [45, 46]; or by luminescence spectroscopy [47].

5. Uses

On a commercial scale, about 10 000 t/a of anthracene is used, almost exclusively for the production of anthraquinone (→ Anthraquinone). The use of anthracene as a plasticizer for thermosetting resins and as a light stabilizer for polymers has been proposed. The use of anthracene derivatives as building blocks for engineering plastics was also suggested. The polyether of 9,10-dihydoxyanthracene is an example [48]. Anthracene is a crystalline organic photoconductor that can be used in electrophotography. In its purest form, anthracene is frequently employed as a scintillant to detect high-energy radiation, e.g., in nuclear physics.

6. Toxicology and Environmental Aspects

Toxicology. As a polycyclic aromatic hydrocarbon, anthracene is suspected to be carcinogenic [49, 49, 50, 50, 51, 51 – 56]. This earlier experience involving workers is based on crude anthracene that was contaminated with various other polycyclic aromatic hydrocarbons. Pure anthracene, however, has no appreciable carcinogenic effect under experimental conditions [54, 57 – 59]. This is underlined by consistently negative findings in numerous in vitro and in vivo genotoxicity tests [58, 59]. Only subcutaneous injections of an oily solution containing 20 mg of anthracene, given 33 times at the rate of one per week, resulted in local development of fibroma, to some extent with sarcoma-like excrescences [54, 55]. An epicutaneous tumor initiation test conducted over 35 weeks on mice

Table 1. Photoinduced anthracene toxicity

Test organism	Effective concentration		Exposure to UV light		Ref.
	Type (total test duration)	Value, mg/L	Intensity, μW/cm^2	Time	
Selenastrum capricornutum (alga)	NOEC (36 h)	0.0015 – 0.003	A 410 – 765	22 h	[94]
Daphnia magna (crustacea)	NOEC (21 d)	ca. 0.002	A 117	16 h/d	[95, 96]
Juvenile bluegill (freshwater fish)	LC_{100} (24 – 72 h)	0.012	sunlight (Oct.)	day-night cycle	[93]
Juvenile bluegill (freshwater fish)	LC_{50} (96 h)	0.0012	B 170	96 h	[92]
Fathead minnow larvae (freshwater fish)	LC_{50} (40 h)	0.0054	A 95	16 h	[97]
Fathead minnow (freshwater fish)	NOEC (9 weeks)	<0.0066	none		[90]

with pure anthracene and phorbol ester as promoter resulted in papilloma in a few cases (4 out of 28 animals) [60]. Anthracene is classified as "not classifiable as to its carcinogenicity to humans" by IARC (Category 3) and by EPA (Group D) [59].

Peroral application of 1.7 g/kg of pure anthracene has no lethal effect on mice [61]. Anthracene is absorbed percutaneously: after topical application of a ^{14}C-labeled solution in hexane or acetone (ca. 9 μg/cm^3) to rat skin, some 50 % was absorbed in 6 d (ca. 29 % was recovered from the urine, ca. 22 % from the feces, and ca. 1 % from tissue, mainly the liver and kidneys); after 1 d 20 % of the dose was already present in the urine (ca. 17 %) and feces (ca. 3 %) [62]. Anthracene can sensitize the skin locally to light [63]. Gerarde proposes a TLV of 0.1 mg/m^3 [56].

Environmental Fate and Toxicity. Anthracene is inherently biodegradable in water [64 – 67] and soil [68 – 72]. Most of the elimination half-lives found in experimental soil and water systems are below 180 d. In water – sediment mesocosm systems involving photolytic and biological processes, short residence times were observed with elimination rates of >98 to >99.5 % for the parent substance from the water phase within 17 and 40 d, respectively [73 – 75]. Several studies using [9-^{14}C]-radiolabeled anthracene gave evidence of partial mineralization of the parent compound during the observation period [64, 76, 77].

Photodegradation in air and water is rapid; measured elimination half-lives were between 2 and 4 h in air [78, 79] and less than 1 to ca. 40 h in water systems [64, 80 – 82].

The moderate to high water-partition coefficient (log P_{ow} 3.9 – 4.5) [74, 83, 84] indicates significant bioconcentration potential. This is low to moderate in fish [64, 74, 85 – 90] and daphnia [64, 87, 88, 91] with bioconcentration factors (BCF) generally ranging between ca. 700 and 5000.

Photoinduced aquatic toxicity of anthracene in aquatic organisms may be 2 – 3 orders of magnitude higher than in the absence of UV light and appears to depend on the accumulated body burden, as fish previously exposed to anthracene in the dark were void of any symptoms, but exhibited severe signs of distress or showed increased mortality when placed in clean water in the presence of UV light [92]. But also without prior or concurrent UV irradiation, chronic toxic effects of anthracene are reported to affect reproduction of fish at low concentrations [90].

Acute lethal concentrations (LC_{50}) and no-observed-effect concentrations (NOEC) are below 0.1 and 0.01 mg/L, respectively, and significantly lower than the water solubility of anthracene (see Table 1).

7. References

General References

1. *Beilstein*, **5**, 657; **5 (1)**, 321; **5 (2)**, 569; **5 (3)**, 2123.
2. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung*, Verlag Glückauf, Essen 1951, p. 174–181.
3. H.-G. Franck, G. Collin: *Steinkohlenteer*, Springer Verlag, Berlin - Heidelberg - New York 1968, p. 56–58 136–137.

4. Rütgerswerke, *Erzeugnisse aus Steinkohlenteer*, Frankfurt/Main 1958.

Specific References

5. S. E. Stein, L. L. Griffith, R. Billmers, R. H. Chen, *J. Org. Chem.* **52** (1987) 1582–1592.
6. Rütgerswerke, DE 3303742, 1983 (B. Fell, G. Maletz).
7. Studiengesellschaft Kohle mbH, U.S. 5069894, 1991 (B. Bogdanovic).
8. Rütgers AG, EU-PS 799813, 1997 (E. Fuhrmann, J. Talbiersky, W. Erdmann, F. Alsmeier).
9. J. Talbiersky, *Erdöl Erdgas Kohle* **9** (1998) 436–439.
10. Ukrainian Scientific-Research Institute of Coal Chemistry, SU 386889, 1971 (V. E. Privalov, K. A. Belov, E. I. Vail, Y. T. Rezunenko).
11. B. Karabon, A. Zin, K. Mnich, *Koks Smola Gaz* **19** (1974) 33–37.
12. M. S. Litvinenko, L. D. Gluzman, A. A. Rok, S. N. Kipat, V. P. Bogunets, N. I. Zhuravskaya, R. M. Zil'berman, L. I. Didenko, *Koks Khim.* 1976, no. 3 33–36
Coke Chem. 1976, no. 3 38–42.
13. L. S. Kuznetsova, V. I. Borodin, A. A. Stepanenko, A. A. Tolochko, V. N. Nazarov, *Koks Khim.* 1978, no. 1 38–40
Coke Chem. 1978, no. 1 50–52.
14. B. Karabon, A. Zin, *Koks Smola Gaz* **26** (1980) 95–98.
15. Rütgerswerke, DE 2020973, 1970 (H. Buffleb, J. Falenski, H.-G. Franck, J. Turowski, F. Melichar, G. Collin, M. Zander).
16. I. Jurkiewicz, K. Wiszniowski, *Koks Smola Gaz* **5** (1960) 117–121.
17. B. Marciniak, *Mol. Cryst. Liq. Cryst.* **162 B** (1988) 301–313.
18. Union Rhein. Braunkohlen Kraftstoff, DE 1046002, 1957 (F. Hübenett, G. Altena).
19. J. Polaczek, Z. Lisicki, T. Tecza, *Przem. Chem.* **60** (1981) 169–171.
20. Charkovskij politechnitsheskij institut imeni V. T. Lenina, Ukrainskij nautshnoissledovatelskij Uglechimit sheskij institut, DE-OS 2626361, 1976 (V. J. Privalov, J. I. Vail, L. S. Kusnezova, K. A. Belov, I. M. Nosalevich, I. V. Romanov).
21. G. Kesicka, S. Trybuta, *Przem. Chem.* **65** (1986) 145–148.
22. Rütgers AG, DE 10103208, 2002 (R. Bischof, K. Dialer, A. Nikzad, T. Büttner, E. Fuhrmann, U. Knips, D. Scholl, B. Vierhaus, F. Alsmeier, W. Bergins, S. Giertler, K. Stolzenberg, J. Talbiersky).
23. D. Scholl, H. Appler, V. Marx, E. Fuhrmann, J. Talbiersky, *Erdöl Erdgas Kohle* **10** (2004) 336–339.
24. A. Zboril, J. Ruzek, CS 150737, 1971.
25. J. Komurka, I. Koropecky, V. Vlcek, J. Jelinek, J. Mixa, J. Ruzek, CS 164058, 1973.
26. Soc. Chim. Gerland, FR 976 773, 1948 (A. J. M. Saunier).
27. Chem. Systems, DE-OS 2123342, 1971 (G. M. Sugarman).
28. R. Sizmann, *Angew. Chem.* **71** (1959) 243–245.
29. E. Wolowski, H. Hosang, *Chemie-Technik* **8** (1980) 395–399
30. Rütgerswerke, DE 2000351, 1970 (H. Buffleb, H.-G. Franck, J. Turowski, G. Collin, M. Zander).
31. Rütgerswerke, DE 1906807, 1969 (H. Buffleb, H.-G. Franck, R. Oberkobusch, J. Turowski, G. Collin, M. Zander).
32. Koppers Co., US 3389188, 1966 (W. A. Michalowicz).
33. G. Kölling, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) 726–729.
34. K. Handrick, W. Hodek, F. Mensch, *DGMK-Compendium 78/79*, **2** (1978) 1089–1106.
35. Bergwerksverband, DE 2952062, 1979 (K. Handrick, G. Kölling, F. Mensch).
36. Tetralin, DE 333158, 1919.
37. Sun Oil, US 3320332, 1964 (A. Schneider).
38. Sun Oil, DE-OS 1493030, 1965 (R. D. Bushick, G. Mills).
39. F. Meyer, D. Hausigk, G. Kölling, *Liebigs Ann. Chem.* **736** (1970) 140–141.
40. Bergwerksverband, DE 1951127, 1969 (F. Meyer, G. Kölling, D. Hausigk).
41. Bayer, DE-OS 2218004, 1972 (H. Wolz, R. Wenzel, M. Martin, G. Scharfe).
42. H. P. Kaufmann, J. Baltes, *Fette Seifen* **43** (1936) 93–95.
43. E. Funakubo, Y. Matsumoto, M. Fujiura, I. Kawanshi, S. Hiroike, *Brennst. Chem.* **40** (1959) 377–383.
44. M. Zander, H. D. Sauerland, *Erdöl Kohle Erdgas Petrochem.* **25** (1972) 526–530.
45. H. Schmidt, *Erdöl Kohle Erdgas Petrochem.* **19** (1966) 275–278.
46. F. P. Hazlett, R. B. Hannan, J. H. Wells, *Anal. Chem.* **22** (1950) 1132–1136.
47. M. Zander, *Angew. Chem. Int. Ed. Engl.* **4** (1965) 930–938.
48. V. Sinigersky, K. Müllen, M. Klapper, I. Schopov, *Macromol. Chem. Phys.* **201** (2000) 1134–1140.

49. K. B. Lehmann, F. Flury: *Toxicology and Hygiene of Industrial Solvents*, Williams and Wilkins Co., Baltimore 1943.
50. W. C. Hueper: *Occupational Tumors and Allied Diseases*, C. C. Thomas, Springfield 1942.
51. National Institut for Occupational Safety and Health (NIOSH): Registry of Toxic Effects of Chemical Substances, U.S. Government Printing Office, Washington, D.C., July 1982, p. 1277.
52. N. I. Sax: *Cancer-Causing Chemicals*, Van Nostrand Reinhold Comp., New York 1981, p. 288.
53. N. I. Sax: *Dangerous Properties of Industrial Materials*, 5th ed., Van Nostrand Reinhold Comp., New York 1981, p. 288.
54. D. Schmähl, Z. Krebsforsch. **60** (1955) 697–710.
55. H. Druckrey. D. Schmähl, *Naturwissenschaften* **42** (1955) 159–160.
56. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons*, Elsevier, Amsterdam 1960, pp. 240–248.
57. E. L. Kennaway, *J. Ind. Hyg*. **5** (1924) 462.
58. WHO: Selected Non-heterocyclic Polycyclic Aromatic Hydrocarbons, Environmental Health Criteria (EHC) 202, International Programme on Chemical Safety (IPCS), Geneva, 1998.
59. U.S. EPA: Anthracene, IRIS (Integrated Risk Information System), last update 01 April 1997 [http://www.epa.gov/iris/subst/0434.htm]
60. J. D. Scribner. *J. Natl. Canc. Inst*. **50** (1973) 1717–1719.
61. P. A. Nagornyi. *Gig. Tr. Prof. Zabol*. **13** (1969) 59–61.
62. J. J. Yang, T. A. Roy, C. R. Mackerer. *Toxicol. Ind. Health* **2** (1986) 79–84.
63. I. E. Kochevar, R. B. Armstrong, J. Einbinder et al., *Photochem. Photobiol*. **36** (1982) 65–69
64. G. R. Southworth, *ASTM Spec. Tech. Publ.* **667** (1979) 359–380.
65. A. Männisto, E. S. Melin, J. A. Puhakka, J. F. Ferguson, *Polycyclic Arom. Compd*. **11** (1996) 27–34.
66. H. H. Tabak, S. A. Quave, C. I. Mashni, E. F. Barth, J. Water Pollut. *Control Fed*. **53** (1981) 1503–1518.
67. T. Nishihara, S. Hasebe, J. Nishikawa, M. Kondo, *J. Appl. Microbiol.* **82** (1997) 441–447.
68. K. S. Park, R. C. Sims, R. R. Dupont, W. J. Doucette, J. E. Matthews, *Environ. Toxicol. Chem*. **9** (1990) 187–195.
69. M. P. Coover, R. C. Sims, *Hazard. Waste Hazard. Mater*. **4** (1987) 69–82.
70. S. R. Wild, K. C. Jones, *Environ. Toxicol. Chem.* **12** (1993) 5–12.
71. K. A. Reilley, M. K. Banks, A. P. Schwab, *J. Environ. Qual.* **25** (1996) 212–219.
72. J. D. Goodin, M. D. Webber, *J. Environ. Qual.* **24** (1995) 271–278.
73. R. F. Lee, W. S. Gardner, J. W. Anderson et al., *Environ. Sci. Technol.* **12** (1978) 832–838.
74. P. Lu, R. L. Metcalf, E. M. Carlson, *Environ. Health Persp.* **24** (1978) 201–208.
75. K. T. J. Bestari, R. D. Robinson, K. R. Solomon et al., *Environ. Toxicol. Chem.* **17** (1998) 2359–2368.
76. W. S. Gardner, R. F. Lee; K. R. Tenore, L. W. Smith, *Water Air Soil Pollut.* **11** (1979) 339–347.
77. S. E. Herbes, L. R. Schwall, *Appl. Environ. Microbiol.* **35** (1978) 306–316.
78. H. W. Biermann, H. MacLeod, R. Atkinson et al., *Environ. Sci. Technol.* **19** (1985) 244–248.
79. R. Atkinson, D. L. Baulch, R. A. Cox et al., *J. Phys. Chem. Ref. Data Suppl. III* **18** (1985) 881–1097.
80. R. G. Zepp, P. F. Schlotzhauer in P. W. Jones, P, Leber (eds.): *Polynuclear Aromatic Hydrocarbons*, 141–158, Ann Arbor Sci. Publ., Ann Arbor, MI, 1979.
81. P. F. Landrum, S. M. Bartell, J. P. Giesy et al., *Ecotox. Environ. Safety* **8** (1984) 183–201.
82. J. W. Bowling, G. Leversee, P. F. Landrum, J. P. Giesy, *Aquat. Toxicol.* **3** (1983) 79–90.
83. M. M. Miller, P. Wasik, G.-L. Huang et al., *Environ. Sci. Technol.* **19** (1985) 522–529.
84. S. W. Karickhoff, D. S. Brown, T. A. Scott, *Water Res.* **13** (1979) 241–248.
85. A. Spacie, P. F. Landrum, G. J. Leversee, *Ecotoxicol. Environ. Safety* **7** (1983) 330–334.
86. NITE: Anthracene. Japan Chemicals Inventories, National Institute of Technology/Japan, 2005 [http://www.safe.nite.go.jp/].
87. J. L. Newsted, J. P. Giesy, *Environ. Toxicol. Chem.* **6** (1987) 445–461.
88. J. T. Oris, A. T. Hall, J. D. Tylka, *Environ. Toxicol. Chem.* **9** (1990) 575–584.
89. P. de Voogt, B. van Hattum, P. Leonards et al., *Aquatic Tox.* **20** (1991) 169–194.
90. A. T. Hall, J. T. Oris, *Aquat. Toxicol.* **19** (1991) 249–264
91. S. E. Herbes, G. F. Risi, *Bull. Environ. Contam. Tox.* **19** (1978) 147–155.
92. J. T. Oris and J. P. Giesy, *Aquat. Toxicol.* **6** (1985) 133–146.

93. J. W. Bowling, G. Leversee, P. F. Landrum, J. P. Giesy, *Aquat. Toxicol.* **3** (1983) 79–90.
94. W. R. Gala, J. P. Giesy, *Aquat. Environ. Contam. Toxicol.* **23** (1992) 316–323.
95. L. L. Holst, Giesy, *Toxicol. Chem.* **8** (1989) 933–942.
96. J. A. Foran, L. L. Holst, J. P. Giesy, *Environ. Toxicol. Chem.* **10** (1991) 425–427.
97. J. T. Oris, J. P. Giesy, *Chemosphere* **16** (1987) 1395–1404.

Anthraquinone

AXEL VOGEL, Bayer AG, Leverkusen, Federal Republic of Germany

1. **Introduction** 1
2. **Properties** 1
3. **Production** 3
3.1. **Oxidation of Anthracene with Chromic Acid** 4
3.2. **Vapor-Phase Oxidation of Anthracene with Air** 4
3.3. **Naphthalene Process** 5
3.4. **Synthesis from Phthalic Anhydride and Benzene** 6
3.5. **Styrene Process** 6
3.6. **Environmental Considerations** 7
4. **Purification, Quality Requirements, and Analysis** 7
5. **Uses** 7
6. **Economic Aspects** 8
7. **Toxicology** 8
8. **References** 8

1. Introduction

Anthraquinone [*84-65-1*] is the parent compound for a large palette of anthraquinone dyes and so is the most important starting material in their production. Furthermore, anthraquinone is gaining importance as a catalyst in the pulping of wood.

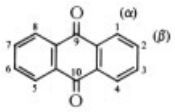

1

In 1835, anthraquinone was prepared for the first time by LAURENT, via oxidation of anthracene, which at first went largely unnoticed. Its special importance for the dye industry was recognized only in 1868 when GRAEBE and LIEBERMANN prepared anthracene from alizarin (1,2-dihydroxyanthraquinone) and, in turn, alizarin via anthraquinone. This laid the groundwork for the synthesis route of this – at the time – important dye. Since then, science and technology have extended the chemistry of anthraquinone swiftly, first by the discovery of new mordant dyes and acid dyes, later by the introduction of vat dyes and disperse dyes, and, since 1950, by the anthraquinone reactive dyes. At the same time, anthraquinone chemistry has acted as a stimulus for the rapid extension of the production processes of chemicals, such as oleum and chlorine.

Since 1970, the industrial chemistry of anthraquinone has been undergoing a revolution worldwide. Initiated by a growing shortage of anthracene accompanied by an increasing demand, new procedures of synthesis using naphthalene or styrene as starting materials were developed. Moreover, further processing of anthraquinone is changing; α-substituted anthraquinones should no longer be prepared by mercury-catalyzed sulfonation but by new nitration procedures.

A completely new use for anthraquinone was established by HOLTON in 1977, when he discovered the catalytic effect of anthraquinone on the alkaline pulping of wood. This branch of anthraquinone chemistry is experiencing a rapid development and has good chances of outdistancing the needs of the dye industry. The demand for anthraquinone has risen continuously since 1870. At present the annual demand is $\approx 30\,000$ t.

2. Properties

Physical Properties. Anthraquinone [*84-65-1*], $C_{14}H_8O_2$, M_r 208.20, *mp* 287 °C, *bp* 377 °C, crystallizes as thin, light yellow to pale yellow needles; it is almost colorless if finely divided, as when precipitated from sulfuric acid. The faint colors of the crystals are effected strongly by the granular structure and in commercial products often are covered up by trace impurities.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_347

Crystallographic class: monoclinic, space group $P2_1/a$. Dimensions of the unit cell: $a = 1.583$ nm, $b = 0.397$ nm, $c = 0.789$ nm, $\beta = 102.5$

Density of the crystals at 20 °C: 1.438 g/cm^3, of the melt (at 293 °C): 1.067 g/cm^3

Vapor pressure: 10.0 kPa at 286 °C, 20.0 kPa at 300 °C, 48.0 kPa at 340 °C

Molar heat capacity at constant pressure c_p (solid) 265 J mol^{-1} K^{-1}

Enthalpy of formation −179 kJ/mol

Enthalpy of combustion at 25 °C: 6449 kJ/mol

Melting enthalpy at 287 °C: 32.57 kJ/mol

Flash point 185 °C

The UV spectrum exhibits four bands – 251 nm (ε_{max} 54000), 279 nm (ε_{max} 17600), 321 nm (ε_{max} 4800), and 377 nm (ε_{max} 110) – each with pronounced vibrational structure. Unlike solutions of anthracene, most solutions of anthraquinone do not show fluorescence. The IR spectrum has very few bands (characteristic bands at $\nu_{C\rightarrow O} = 1680\ cm^{-1}$) because of the high symmetry of the compound.

Anthraquinone is sublimed easily without decomposition, a property exploited industrially for purification.

Anthraquinone is practically insoluble in water (0.006 g/L at 50 °C) and nearly insoluble to sparingly soluble in organic solvents at room temperature. However, solubility in organic solvents increases with temperature (in ethanol 0.05 g per 100 g ethanol dissolves at 18 °C, 2.25 g per 100 g at 78 °C; in toluene 0.19 g per 100 g toluene at 15 °C, 2.56 g at 100 °C). Therefore anthraquinone generally can be recrystallized from high-boiling solvents, particularly from the polar solvents pyridine, aniline, nitrobenzene, formamide, or dimethylformamide. Anthraquinone dissolves in 90 % sulfuric acid giving a yellow to orange solution, and in oleum (20 % SO_3), giving a red solution. It can be precipitated from these solutions in a finely divided form by dilution with water. The solubility increases with the concentration of the sulfuric acid. Anthraquinone behaves similarly in hydrofluoric acid; however, the solubility in almost 100 % phosphoric acid is relatively small.

Anthraquinone can form explosive dust mixtures with air. Friction causes anthraquinone to become electrostatically charged.

Chemical Properties. Anthraquinone exhibits an extraordinary thermal stability, even in the presence of oxidizing agents. Therefore it can be prepared by oxidizing anthracene in the vapor phase. A yield of 90 % is obtained if conversion is complete.

Anthraquinone is attacked by oxidizing agents only under vigorous conditions. For example, it can be converted to alizarin by fusion with sodium hydroxide and sodium chlorate. Anthraquinone can be converted to the polyhydroxyanthraquinones alizarin, quinizarin, anthrarufin, and purpurine (→ Anthraquinone Dyes and Intermediates, Chap. 2.6.) by oxidation with concentrated sulfuric acid in the presence of boric acid, persulfates, or metaarsenic acid at temperatures generally above 200 °C. The oxidation can be carried out with oleum at a lower temperature that depends on the SO_3 content. Under the usual conditions, concentrated nitric acid does not oxidize anthraquinone, nitration taking place instead. Several mixtures of anthraquinone with 94 – 100 % nitric acid with specific anthraquinone concentrations are explosive.

On the other hand, anthraquinone can be reduced easily by ordinary reducing agents or by hydrogen in the presence of a catalyst. The standard potential (in aqueous ethanolic hydrochloric acid) for the anthraquinone–anthrahydroquinone equilibrium is only 0.155 V. Especially the quinone nucleus is reduced. Sodium hyposulfite in alkaline solution reduces anthraquinone to a blood-red sodium salt of anthrahydroquinone, which rearranges to its tautomer oxanthrone (9-keto-10-hydroxy-9,10-dihydroanthracene) in acidic solution. More vigorous reduction (tin – hydrochloric acid or catalytic hydrogenation at high pressure or high temperature) leads to anthrone (**2**) [*90-44-8*], which is in equilibrium with its tautomer anthrol (**3**) (9-hydroxyanthracene) [*529-86-2*], and then to anthracene and hydroanthracene.

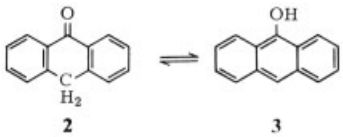

If zinc or sodium hydroxide is used, dianthranol is obtained via dimerization.

The reduction can be restricted to the two outer nuclei by choice of reaction conditions. 1,2,3,4-Tetrahydroanthraquinone [*28758-94-3*] or octahydroanthraquinone [*23585-26-4*] can be obtained. Reactions of anthraquinone with the usual ketone reagents take place with great difficulty or not at all.

If anthraquinone is heated with glycerol, iron powder, and sulfuric acid, condensation takes place at positions 1 and 9, providing the commercially important benzanthrone [*82-05-3*] (→ Anthraquinone Dyes and Intermediates, Chap. 2.12.1.).

Chlorine halogenates anthraquinone in oleum in steps, depending on the proportion used, mainly at the α position. Heating anthraquinone with antimony pentachloride or a similar chlorine-releasing reagent produces chlorinated products, from monochloro- up to perchloroanthraquinone.

Sulfuric acid or oleum can sulfonate and/or hydroxylate anthraquinone. Concentrated sulfuric acid sulfonates only above 200 °C; usually, anthraquinone is sulfonated with 20 – 45 % oleum at 140 – 160 °C. The major reaction products are β-sulfonic acids, i.e., anthraquinone-2-sulfonic acid [*84-48-0*], as well as anthraquinone-2,6-disulfonic acid [*14486-58-9*], and anthraquinone-2,7-disulfonic acid [*84-49-1*]. In the presence of mercury, however, the sulfonation takes place almost exclusively at the α position, providing anthraquinone-1-sulfonic acid [*82-49-5*], anthraquinone-1,5-disulfonic acid [*117-14-6*], and anthraquinone-1,8-disulfonic acid [*82-48-4*].

The nitration of anthraquinone requires vigorous conditions, and provides mainly α-nitroanthraquinones along with 10 – 20 % β-nitro- and α,β-dinitroanthraquinones. Both mono- and dinitration take place, so that impure nitration products are obtained. These can be separated into 1-nitroanthraquinone, 1,5-dinitroanthraquinone, and 1,8-dinitroanthraquinone only by extensive processing. Trinitration of anthraquinone does not take place.

1-Nitroanthraquinone [*82-34-8*] can be prepared by nitrating anthraquinone in mixed acid (H_2SO_4-HNO_3) with excess nitric acid, in nearly 30 mol 94 – 98 % nitric acid, or with an almost stoichiometric amount of the mixed acid in the presence of chlorinated aliphatic hydrocarbons. The crude 1-nitroanthraquinone is purified by recrystallization and, if necessary, by distillation; likewise it is processed directly to crude 1-aminoanthraquinone [*82-45-1*], which is then purified by distillation.

1,5- and 1,8-dinitroanthraquinone are prepared by exhaustive nitration of anthraquinone in excess, nearly anhydrous mixed acid. Pure 1,5-dinitroanthraquinone [*82-35-9*] and 1,8-dinitroanthraquinone [*129-39-5*] are obtained by recrystallizing the crude dinitroanthraquinone mixtures from nitrobenzene (→ Anthraquinone Dyes and Intermediates, Chap. 2.4.).

3. Production

Processes that are used in the industrial production of anthraquinone:

1) Oxidation of anthracene with chromic acid
2) Vapor-phase oxidation of anthracene with air
3) Naphthalene process
4) Synthesis from phthalic anhydride and benzene

Anthraquinone is produced from anthracene, where this is available from coal tar, either by oxidation with chromic acid in 48 % sulfuric acid or by oxidation with air in the vapor phase. The oxidation with chromic acid is competitive, provided that the chromium(III) sulfate lye formed can be processed to tanning agents. Anthracene with purity of $\geqq$ 94 % is required for both oxidation processes; crude anthracene from coal tar must be purified by recrystallization.

About 85 % of world production is based on the oxidation of anthracene. Since the mid-1970s, anthracene production has fallen continuously, creating a supply shortage. Therefore, the complex naphthalene process is gaining in importance. There is an adequate supply of naphthalene in coal tar. If necessary, additional naphthalene can be isolated from the residual oils of gasoline reforming, a process common in the United States.

In the new naphthalene processes developed in Japan by Kawasaki and in Europe by Bayer anthraquinone is synthesized in three steps: Naphthoquinone is prepared by vapor-phase oxidation with air. Butadiene is added to this naphthoquinone in a Diels-Alder reaction. The tetrahydroanthraquinone formed is oxydehydrogenated. In this naphthalene process a significant

amount of phthalic anhydride byproduct is produced.

In the synthesis of anthraquinone from phthalic anhydride and benzene approximately 1.4 t of aluminum chloride and 4 – 6 t of sulfuric acid per ton of anthraquinone are used. This procedure is used in areas where anthracene is not available in sufficient amounts from coal tar. It may become important again, as a result of the shortage of anthracene. Phthalic anhydride has become available at low prices and in sufficient quantities following the introduction of the economical oxidation of naphthalene or *o*-xylene with air.

In the early 1970s another interesting anthraquinone process was developed by BASF. Styrene is first dimerized to 1-methyl-3-phenylindane in the presence of an acid catalyst, which is then converted to anthraquinone in the vapor phase by oxidation with air.

3.1. Oxidation of Anthracene with Chromic Acid

The classic process of oxidizing anthracene with chromic acid is still carried out industrially on a large scale. In combination with the manufacture of tanning agents this process involves virtually no expense for the oxidizing agent. The batch process is still in use.

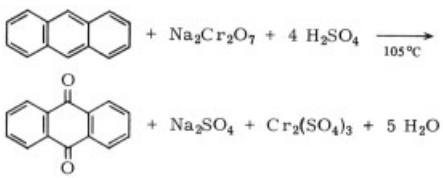

Pure, finely pulverized 94 – 95 % anthracene (2600 kg) is pasted with water (9000 kg) in a 30000-L mixing vessel. A wetting agent is added. For oxidation, portions of 48 % sulfuric acid and 20 % sodium dichromate solution are added alternately; the reaction begins at 50 – 60 °C. The temperature increases to 100 – 105 °C but should not be allowed to increase any further. The temperature is controlled by regulating the amount of reagents added. Foaming is severe. A total of 10200 kg of 48 % sulfuric acid and 23 500 kg of 20 % sodium dichromate solution are used. At the end of the reaction there should be a small excess of oxidizing agent, which is reduced with hydrogen sulfite. The oxidation takes 30 – 36 h [4].

The reaction mixture is filtered through a suction filter while being stirred, washed, suspended with water, and filtered into a chambered rotary filter. A yield of ≈ 3000 kg of 95 % crude anthraquinone is obtained. The product is dried, or it is recrystallized from nitrobenzene, in which case the retained water is evaporated first. A purity of 99 % is achieved.

3.2. Vapor-Phase Oxidation of Anthracene with Air

The vapor-phase oxidation of anthracene with air was described first by A. Wolf in 1916. The catalysts are similar to those for the oxidation of naphthalene or *o*-xylene to phthalic anhydride: vanadium compounds, mainly iron vanadate [5] or vanadic acid doped with substoichiometric amounts of alkali metal or alkaline-earth metal ions [6].

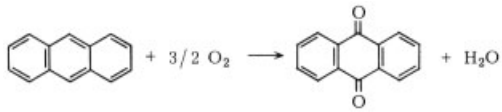

The conversion of the anthracene is nearly quantitative. The major byproduct is a small quantity of phthalic anhydride, which is separated easily. The advantage of this procedure is that it produces anthraquinone of 99 % purity. It can be converted further without purification.

Details on a plant constructed by Aziendi Colori Nazionali Affini, Milan, were published [7]:

Anthracene is evaporated with a preheated air–water vapor mixture, and the gas current is mixed carefully with more air (danger of explosion!) and led into a catalytic furnace from the bottom at 325 °C. The catalytic furnace consists of 1400 L of catalyst piled in layers; temperature is controlled by tubes in which pressurized water circulates through these layers. The lower part of the furnace, where the reaction takes place, is cooled to 390 °C; the upper part is heated to 339 °C. The catalyst consists of iron vanadate prepared from ammonium vanadate and iron(III) chloride. The air flow rate is 2150 m^3/h, each

cubic meter containing 20 g of 94 % anthracene. The gases leaving the catalytic furnace first go through heat exchangers and then through cooling towers, cooling chambers, and dust filters. The anthraquinone produced has an average purity of 99.6 %

The iron vanadate catalyst is prepared as follows [5]: 2.4 kg of red iron(III) oxide is dissolved by heating it in 9.1 L of 30 % hydrochloric acid, 9.6 L of water, and 120 g of potassium sulfate. The resulting solution is mixed with a solution of 30.1 kg of ammonium vanadate in 600 L of water at 60 °C. The solution is made alkaline with ammonia and the water evaporated until precipitation begins. The pregnant iron vanadate solution is evaporated onto pumice at 110 – 130 °C in a heatable ball mill. The catalyst is calcined at 330 – 350 °C. The catalyst should be yellow brown but not dark brown.

After a period of time the activity of the catalyst falls off, resulting in a decline in the yield of anthraquinone and an increase in phthalic anhydride. To keep the activity of the catalyst constant or to reactivate it, a small amount of ammonia is added to the oxidation air or the exhausted catalyst is washed with a dilute alkali solution [8].

3.3. Naphthalene Process

The naphthalene process developed by Kawasaki Kasei Chemicals [9] consists of three steps. In the first step naphthalene is oxidized in the gas phase with air on a vanadium pentoxide catalyst to produce the relatively unstable naphthoquinone (**4**). Large amounts of phthalic anhydride form at the same time. Because most of the naphthalene reacts, recycling of unreacted naphthaline is not necessary [10–13].

The hot reaction mixture is precooled in a gas cooler and then quenched and washed with water in a tower. The naphthoquinone is obtained mixed with phthalic acid anhydride (or phthalic acid) as a suspension in water. The naphthoquinone is extracted with an aromatic solvent, for example xylene or tetrahydroanthraquinone, to separate it from the phthalic acid. The small residue of acid is carefully removed by washing with dilute alkali [14–16].

In the second step the naphthoquinone is

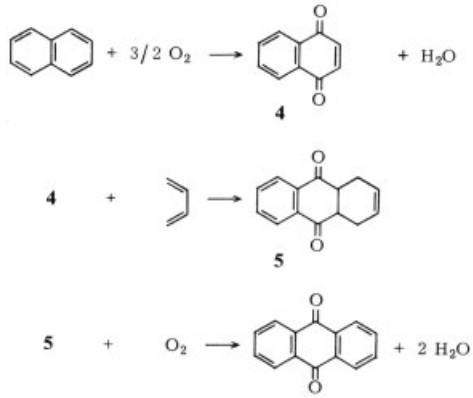

reacted with butadiene in a Diels-Alder reaction to give 1,4,4 *a*,10 *a*-tetrahydroanthraquinone (**5**). The Diels-Alder product is separated from unreacted naphthoquinone by extraction with aqueous alkali, which converts the quinone into a water-soluble alkali-metal salt. The organic phase, which contains naphthoquinone, is recycled to the first step [10], [11], [17].

In the third step the aqueous tetrahydroanthraquinone solution is reacted with air. The anthraquinone that forms is insoluble in aqueous alkali and can be isolated by filtration. The aqueous alkali is recycled into the tetrahydroanthraquinone extraction step [18]. The anthraquinone can be purified further by vacuum distillation [9].

Phthalic acid is isolated from the aqueous phthalic acid solution by crystallization and filtration. It is heated to convert it to phthalic acid anhydride. The phthalic acid anhydride can be further purified by distillation.

Kawasaki Kasei Chemicals operates a plant based on the naphthalene process that produces 3000 t/a. The plant has been in operation since 1980.

In Bayer's variation of the naphthalene process, naphthalene is oxidized with a mixture of air and recycled reaction gas on a vanadium pentoxide catalyst to give naphthoquinone, phthalic acid anhydride and unreacted naphthalene [19–21]. The products of the oxidation along with unreacted naphthalene are quenched and condensed into a liquid reaction mixture, which is reacted without purification with butadiene [22], [23]. The Diels-Alder product mixture that contains tetrahydroanthraquinone is reacted with air. The tetrahydroanthraquinone is

oxydehydrogenated to give anthraquinone. The naphthalene is removed by distillation [23–25]. The two end products, anthraquinone and phthalic acid anhydride, are separated by fractional distillation [26].

3.4. Synthesis from Phthalic Anhydride and Benzene

In the first step o-benzoylbenzoic acid (**6**) is prepared from phthalic anhydride and benzene by a Friedel-Crafts reaction. In the second step the *o*-benzoylbenzoic acid is cyclized to anthraquinone by heating with concentrated sulfuric acid. The primary product of the Friedel-Crafts reaction is an aluminum chloride complex of the *o*-benzoylbenzoic acid, which can harden easily to form a compact mass in the reaction vessel. Several methods have been proposed to overcome this problem. Based on a process described by G. HELLER already in 1906, condensation is carried out in excess benzene to give yields of more than 95 %. Patents issued to Klipstein & Sons and the I. G. Farbenindustrie from 1923 to 1927 describe a solvent-free process carried out in ball mills. The problem posed by the reaction mixture's baking can be reduced by adding ethylene glycol or 1,4-butanediol, which causes a reduction in the necessary excess benzene or makes a solvent-free process possible.

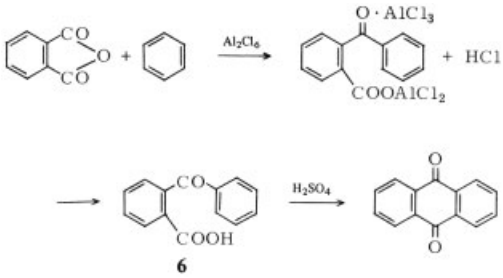

Phthalic anhydride, benzene, and aluminum chloride in a molar ratio of 1 : 1 : 2 are allowed to react below 45 °C in iron ball mills fitted with hollow axles for the addition of benzene and the removal of the hydrogen chloride formed during the reaction, or in a vessel equipped with a strong horizontal agitator, possibly propelled from the bottom. The reaction mass, which is heterogeneous at first, liquefies, then gradually becomes viscous as hydrogen chloride is continuously lost, and eventually forms a powder. During the reaction there is extensive foaming, producing several times the original volume, a fact that must be considered when the charge is measured out. The reaction is complete after 1 mol of hydrogen chloride per mol phthalic anhydride has been evolved. The reaction product is placed into dilute acid. The precipitated *o*-benzoylbenzoic acid is separated, washed, and dried, providing a yield of more than 95 %.

Pure *o*-benzoylbenzoic acid [*85-52-9*] crystallizes from water with 1 mol of water of hydration in triclinic crystals, *mp* 94 °C. The anhydrous acid crystallizes from xylene, *mp* 127 °C.

In the second step it is also possible to start with the aluminum chloride complex formed in the first step and cleave 1 mol of hydrogen chloride by heating to 200 °C. A good yield of anthraquinone is obtained. In practice, however, the condensation of the precipitated dried *o*-benzoylbenzoic acid with sulfuric acid is preferred.

The acid is dissolved in three to four times its weight in 95 – 98 % sulfuric acid and heated to 115 – 140 °C. The duration of heating depends on the quantity and concentration of the sulfuric acid and on the temperature. After completion of the reaction the anthraquinone is precipitated with water, filtered, and dried. A portion of the product also can be obtained in coarsely crystallized form by cooling the sulfuric acid solution directly. The yield is almost quantitative.

The condensation solution can be used directly for the manufacture of benzanthrone or anthraquinonesulfonic acids.

3.5. Styrene Process

This process was developed in pilot by BASF.

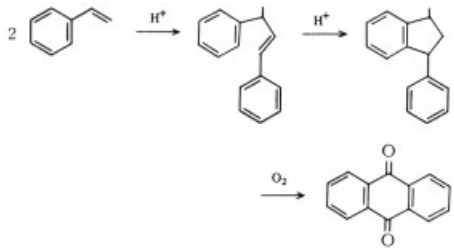

Styrene dimerizes in the presence of acid catalysts, such as sulfuric acid [27], [28], phosphoric acid, or boric acid [29], yielding primarily 1,3-diphenyl-2-butene, which cyclizes further on the same catalyst to 1-methyl-3-phenylindane [*6416-39-3*]. The yields obtained in this step are 85 – 90 %.

1-Methyl-3-phenylindane is converted directly to anthraquinone by oxidation with air in the vapor phase [30].

Basically, the same catalysts are used as those employed in the gas-phase oxidation of anthracene with air, i.e., vanadium compounds, primarily vanadium pentoxide in combination with other oxides, such as thallium oxide and antimony oxide [31].

Yields of ≈ 77 % are obtained in the oxidation stage. Byproducts are still attached to the precipitated reaction product, and further purification is necessary.

3.6. Environmental Considerations

Chromium salts should not get into the waste water from the chromic acid process. The chromium(III) sulfate lye is disposed of by processing it to tanning agents. The anthraquinone must be carefully washed free of chromium with small amounts of wash water.

In the gas-phase oxidation of anthracene with air and the naphthalene process, on the other hand, the emphasis is placed on purifying the large quantity of waste air.

The production of anthraquinone via the Friedel-Crafts reaction from phthalic anhydride and benzene is accompanied by a release of waste water containing 1.4 t of aluminum chloride and 4 – 6 t of sulfuric acid per ton of anthraquinone. Moreover, carcinogenic benzene must be eliminated from the waste air by thermal posttreatment.

4. Purification, Quality Requirements, and Analysis

Anthraquinone is most often purified by recrystallization, preferably from nitrobenzene. Other purification methods include sublimation or vacuum destillation.

At present, a degree of purity of at least 99 % is required for the major uses of anthraquinone, i.e., production of nitroanthraquinones and anthraquinonesulfonic acids. A content of 98 – 99 % is sufficient for use as catalyst in wood pulping. The admission of anthraquinone by the U.S. Food & Drug Administration (FDA) as catalyst in the manufacture of food packing paper requires a content of at least 98 % [32].

Today the main method of analyzing the content is gas chromatography: The other components and solvents, which are present in ppm quantities, are detected directly. High-pressure liquid chromatography is another suitable method of analysis.

According to the classic analytical processes the content is determined (1) by reduction to soluble anthrahydroquinone, filtration, back-oxidation, and weighing out the filtered anthraquinone or (2) by separating the anthraquinone by column chromatography and analyzing the eluate by UV spectroscopy.

5. Uses

Anthraquinone serves as the basis for the production of a large number of acid and base dyes, vat dyes, disperse dyes, and reactive dyes (→ Anthraquinone Dyes and Intermediates).

Anthraquinone-1-sulfonic acid, anthraquinone-2-sulfonic acid, anthraquinone-1,5-disulfonic acid, and anthraquinone-1,8-disulfonic acid as well as 1-nitroanthraquinone, 1,5-dinitroanthraquinone, 1,8-dinitroanthraquinone, anthrone, and benzanthrone are prepared in the first processing step. For environmental reasons the production of anthraquinone-2-sulfonic acid and anthraquinone-1,5- and anthraquinone-1,8-disulfonic acids is being replaced increasingly by the production of 1-nitroanthraquinone and 1,5- and 1,8-dinitroanthraquinone.

In the second stage of processing, the sulfonic or nitro groups generally are replaced by ammonia, amine, chlorine, alcoholate, phenolate, or other nucleophilic agents or are reduced to an amino group. Some of the anthraquinone dyes, however, are prepared from anthraquinone derivatives. These are synthesized not from anthraquinone, but from smaller molecules.

On a small scale, anthraquinone is used in the manufacture of denaturants to protect seed

from crows, in the modification of stand oils, to brighten up colophoniy, and in the manufacture of hydrogen peroxide.

The paper and pulp industry is another rapidly expanding market for anthraquinone [33]. Anthraquinone is a redox catalyst in the production of pulp; the intermediate anthrahydroquinone catalyzes the alkaline hydrolysis of the polymeric lignin. When 0.03 – 0.05 % anthraquinone is added, the yield of pulp is increased 2.5 – 4 % and the cooking time is reduced up to 30 % [34].

The worldwide production capacity is adequate to cover the actual demand for anthraquinone [35].

6. Economic Aspects

The production capacity for anthraquinone worldwide is at present 34000 t/a. About 50 % of this capacity is in Western Europe.

The most important manufacturers of anthraquinone and their production capacities are listed in Table 1.

Table 1. Anthraquinone production capacities [32], [35]

Western Europe	
Bayer	11000 t/a
ICI	2500 t/a
Ciba-Geigy	2200 t/a
ACNA	1000 t/a
Ugine Kuhlmann	800 t/a
L. B. Holliday	200 t/a
Yorkshire Chem.	200 t/a
	17900 t/a
United States	
Toms River Chem.	1500 t/a
American Cyanamid	400 t/a
	1900 t/a
Japan	
Kawasaki Kasei	3000 t/a
Nihon Joryn Kogyo	2000 t/a
	5000 t/a
Eastern Europe	
USSR	2600 t/a
Poland	1800 t/a
Czechoslovakia	1200 t/a
	5600 t/a
Other Countries	
Indian Dyestuff	3000 t/a
Amar Dye (India)	250 t/a
China	400 t/a
	3650 t/a
Worldwide	34050 t/a

7. Toxicology

Based on present knowledge, anthraquinone, unlike the "emodins" (hydroxyanthraquinone glycosides) from anthraquinone drugs [36], is biologically completely "inert," i.e., inactive, presumably as a consequence of its insolubility in water and lipids.

There is no toxicologic information about anthraquinone. Even in the *Toxic Substances List* [37] no LD_{50} is found for anthraquinone. In contrast to benzoquinone, which causes severe local irritation and is included in the list of occupational hazards because of its damaging effect on the cornea, a fine dust of anthraquinone to which emulgators were added had no greater effect on the eyes of rabbits than a talcum suspension, as was shown in a BASF study.

An MAK value of 10 mg/m^3 was recommended [38]; therefore, anthraquinone dust is considered the same as inert dust.

8. References

General References

1. *Beilstein,* **VII,** 781, **VII (1),** 407, **VII (2),** 709, **VII (3),** 4059, **VII (4),** 2556.
2. E. de Barry Barnett: *Anthracene and Anthraquinone,* Baillière, Tindall & Co., London 1921.
3. J. Houben: *Das Anthracen und die Anthrachinone,* Thieme Verlag, Leipzig 1929.

Specific References

4. Fiat Final Rep. no. 1313, vol. 2, p. 19.
5. Fiat Final Rep. no. 1313, vol. 1, p. 332.
6. Ciba, DE 1016694, 1954.
7. Bios Misc. Rep. no. 104.
8. Ciba, DE 1020617, 1954.
9. Kawasaki Kasei Chemicals Ltd., Research and Development in Japan awarded the Okochi Memorial Price, 1980, Okochi Memorial Foundation.
10. Kawasaki, JP 5108256, 1974.
11. Kawasaki, JP 5108257, 1974.
12. Kawasaki, JP 5322559, 1978.
13. Kawasaki, DE 3033341, 1980.
14. Kawasaki, JP 5251356, 1975.
15. Kawasaki, GB 2039897, 1978.
16. Kawasaki, JP 5422246, 1979.
17. Kawasaki, US 4412954, 1981.
18. Kawasaki, JP 5652434, 1981.
19. Bayer, DE 2532422, 1975.

20. Bayer, DE 2453232, 1974.
21. Bayer, DE 2532365, 1975.
22. Bayer, DE 2532388, 1975.
23. Bayer, DE 2218316, 1972.
24. Bayer, DE 2245555, 1972.
25. Bayer, US 4284576, 1975.
26. Bayer, DE 2532450, 1975.
27. J. Risi, D. Gauvin, *Can. J. Res. Sect. B.* **14** (1936) 255.
28. P. E. Spoerri, M. J. Rosen, *J. Am. Chem. Soc.* **72** (1950) 4918.
29. BASF, DE 2064099, 1970.
30. BASF, DE 1934063, 1969.
31. BASF, DE 2135421, 1971.
32. A. Budzinski, *Chem. Ind. (Düsseldorf)* **33** (1981) 332.
33. H. H. Holton, *Pulp Pap. Can.* **78** (1977) 10.
34. G. Gellerstedt, *Kem. Tidskr.* **58** (1979) no. 2.
35. *Eur. Chem.* 1984, no. 27, 429.
36. H. Auterhoff, *Arzneim. Forsch.* **3** (1953) 23.
37. US-Department of Health; *Toxic Substances List,* Nat. Instit. for Occup. Safety and Health, Rockville, Maryland 1972.
38. V. H. Volodishenko, *Biol. Abstr.* **53** (1972) no. 63983.

Anthraquinone Dyes and Intermediates

Hans-Samuel Bien, Bayer AG, Leverkusen, Federal Republic of Germany

Josef Stawitz, Bayer AG, Leverkusen, Federal Republic of Germany

Klaus Wunderlich, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction and History 2
2. Methods of Production 2
2.1. Anthraquinonesulfonic Acids . . . 3
2.2. Alkyl- and Arylanthraquinones . 4
2.3. Haloanthraquinones 4
2.4. Nitroanthraquinones 6
2.5. Aminoanthraquinones 7
2.5.1. General Aspects 7
2.5.1.1. Replacement Reactions 8
2.5.1.2. Modifications of the Amino Functions 10
2.5.1.3. Substitutions on the Aminoanthraquinone Nucleus 11
2.5.2. Individual Aminoanthraquinones . . 13
2.5.3. Diazotization of Aminoanthraquinones 18
2.6. Hydroxyanthraquinones, Alkoxy- and Aryloxyanthraquinones . . . 18
2.6.1. General Aspects 18
2.6.1.1. Synthesis of the Nucleus 18
2.6.1.2. Replacement Reactions 19
2.6.1.3. Conversions of the Hydroxy Group 20
2.6.1.4. Substitutions on the Hydroxyanthraquinone Nucleus 20
2.6.2. Individual Hydroxyanthraquinones 22
2.7. Mercaptoanthraquinones 26
2.7.1. General Aspects 27
2.7.2. Individual Mercapto Compounds and Thioethers 27
2.8. Anthraquinonesulfones 27
2.9. Anthraquinone Aldehydes and Their Derivatives 28
2.10. Anthraquinonecarboxylic Acids and Carboxylic Acid Derivatives 29
2.10.1. General Aspects 29
2.10.2. Individual Anthraquinonecarboxylic Acids 30
2.11. Anthraquinone Derivatives with Condensed Rings 31
2.11.1. Anthraquinone Derivatives Containing Imidazole, Oxazole, and Thiazole Rings 31
2.11.2. Anthraquinonecarbazole Derivatives 32
2.11.3. Anthraquinone Derivatives with Acridone Substructure (Phthaloylacridones) 35
2.11.4. Anthraquinone Derivatives with Phenazine Substructure 37
2.12. Anthraquinone Derivatives with Meso Rings (1,9-Cyclo-10-anthrones) 38
2.12.1. Benzanthrone and Its Derivatives . 38
2.12.2. Pyrazolanthrone and Its Derivatives 40
2.12.3. 1,9-Anthrapyrimidine and Its Derivatives (1,9(*N*-)-Pyrimidinoanthrone-10) 41
2.12.4. 1,9-Anthrapyridone and Its Derivatives 42
2.12.5. Dibenzpyrenequinones 42
2.12.6. Anthanthrones 43
2.12.7. Pyranthrone and Flavanthrone . . . 44
2.12.8. Violanthrone and Isoviolanthrone . 44
2.12.9. Acedianthrone 46
3. Dye Classes 46
3.1. Neutral Dyes 47
3.1.1. Disperse Dyes 47
3.1.1.1. Dyes for Polyester Fibers 47
3.1.1.2. Dyes for Cellulose Ester and Synthetic Polyamide Fibers 49
3.1.1.3. Transfer Dyes 50
3.1.1.4. Dyes for Cotton – Polyester Fabrics 50
3.1.2. Dyes Soluble in Organic Solvents . 51
3.1.3. Vat Dyes 51
3.1.4. Pigments 56
3.2. Anionic Dyes 57
3.2.1. Acid Dyes 57
3.2.2. Direct Dyes 59
3.2.3. Reactive Dyes 60
3.3. Cationic Dyes 61
4. References 61

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_355

1. Introduction and History

The elucidation of the structure of alizarine by C. Graebe and C. Liebermann in 1868 marks the beginning of anthraquinone chemistry. After Caro, Graebe, and Liebermann had worked out a commercial synthesis for alizarine, major efforts, especially by R. Bohn (Badische Anilin- & Sodafabrik) and R. E. Schmidt (Farbenfabriken Bayer, formerly Friedrich Bayer & Co.) were directed toward the manufacture of hydroxyanthraquinones. A sizeable collection of red, blue, and green mordant dyes was assembled over a span of 20 years. The turn of the century saw the development of the first acidic anthraquinone dyes that could be used to dye wool in fast red to green shades without the need for pretreatment with mordants.

In 1901 a new era of anthraquinone chemistry started with the discovery of the mercury-catalyzed α-sulfonation of anthraquinone. (Substituted anthraquinones are usually designated as α for substitution at positions 1, 4, 5, or 8 and as β for substitution at positions 2, 3, 6, or 7. For example, α,α'-anthraquinonedisulfonic acid is a mixture of anthraquinone-1,5 and 1,8-disulfonic acids.) About 80 % of the anthraquinone dyes are still prepared via anthraquinonesulfonic acids, even though emphasis is shifting at present toward synthesis via nitroanthraquinones. Important early reactions of anthraquinone chemistry include the Fischer reaction (replacement of a sulfonic acid group by chlorine) and high-pressure fusion with lime (replacement of sulfonic acid by a hydroxy group).

Simultaneously, new fundamental developments took place in the area of anthraquinone vat dyes, e.g., indanthrone, flavanthrone, violanthrone, benzoylaminoanthraquinones, anthrimides, and anthrimidecarbazoles. Subsequent work concentrated on closing gaps in the offerings of vat and acid dyes and testing of potential applications in nontextile dye applications.

The introduction of synthetic fibers, such as polyester, polyamide, and polyacrylonitrile fibers, initiated a new phase in the development of anthraquinone dyes. This led to a number of substitution products of anthraquinone that were useful as disperse dyes or as basic dyes. With the appearance of the first fiber-reactive dyes, work on acid anthraquinone dyes was intensified.

A new application for vat dyes constituted their use as organic pigments. Based on these insights, a whole new series of specialty pigments was developed.

The synthesis of an anthraquinone dye generally involves a large number of steps. For example, the preparation of a vat dye may require as many as eight steps and may necessitate the preparation of seven different anthraquinone intermediates.

Anthraquinone dyes are surpassed in importance only by azo dyes. The more complicated syntheses and lower tinctorial strengths of the anthraquinone dyes make production costs higher than for azo dyes. Therefore anthraquinone dyes tend to be used only when the required properties and color fastness are extreme or when the colors desired are not obtained easily with azo dyes (especially bright blue to turquoise shades). Use of a dye in vat dyeing requires the presence of special structural features (quinoid or indigoid). Therefore most vat dyes are derived from anthraquinone.

2. Methods of Production

The production of anthraquinone intermediates and anthraquinone dyes generally proceeds from a few key products generated by electrophilic substitution of unsubstituted anthraquinone or by synthesis of the nucleus. The major methods employed to prepare anthraquinone derivatives substituted in the α position are sulfonation and nitration. Preparation of β-substituted anthraquinones and of quinizarin (1,4-dihydroxyanthraquinone) generally is accomplished by synthesis of the nucleus starting from phthalic anhydride and a benzene derivative.

Until fairly recently, preparation of almost 80 % of all important anthraquinones was based on anthraquinonesulfonic acids. However, nitration of anthraquinone is gaining in importance presently, a development triggered mainly by environmental considerations (e.g., production of large volumes of waste dilute acids during anthraquinone sulfonation). However, great progress has been made recently in solving the problems associated with sulfonation.

Preparation of nearly all important anthraquinones starts from the following key inter-

mediates: anthraquinonesulfonic acids, nitroanthraquinones, and the products of nucleus synthesis, 1,4-dihydroxy-, 2-methyl-, or 2-chloroanthraquinone. The only exceptions are derivatives with condensed rings, e.g., benzanthrone and derived products, which are prepared directly from anthraquinone via anthrone.

2.1. Anthraquinonesulfonic Acids

Anthraquinonesulfonic acids generally are obtained by sulfonation. Additional options, such as displacement of substituents (e.g., chlorine atoms or nitro groups) by sulfite or oxidation of mercapto groups, are of little importance but are used for the preparation of some substituted anthraquinonesulfonic acids.

The sulfonic acid group in anthraquinone enters readily into nucleophilic exchange reactions, whereby the α position is much more reactive than the β position. Sulfonic acid groups are exchanged readily against amino, alkylamino, hydroxy, and alkoxy groups. Exchange against chlorine atoms also proceeds smoothly (Fischer synthesis, see page 4).

Reduction also may be used to remove the sulfonic acid group from certain substituted anthraquinones, as, for example, from 1-amino-4-arylaminoanthraquinone-2-sulfonic acid.

Desulfonation of the α position may be accomplished by heating in 70 – 90 % sulfuric acid in the presence of mercury. This method is used to convert unnecessary α,α'-disulfonic acids back to anthraquinones [15], [16].

Sulfonation of the β position of anthraquinone is at present of little importance, because β-substituted compounds (e.g., 2-aminoanthraquinone) are prepared nowadays from 2-chloroanthraquinone instead of the sulfonic acid.

Process Details. Sulfonation of anthraquinones requires extreme conditions ($\approx 150\,^{\circ}$C, 20 % oleum), which yield either mono- or disulfonic acids, depending on the method of operation. Whereas nitration and halogenation preferentially substitute the α position, direct sulfonation attacks the β position, yielding anthraquinone-2-sulfonic acid or the 2,6- or 2,7-disulfonic acid derivatives.

If mercury or mercury compounds are added as catalysts, sulfonation proceeds under much milder conditions and substitution occurs preferentially at the α position. The mercury can be reclaimed during workup of the sulfonation mixtures [17–25].

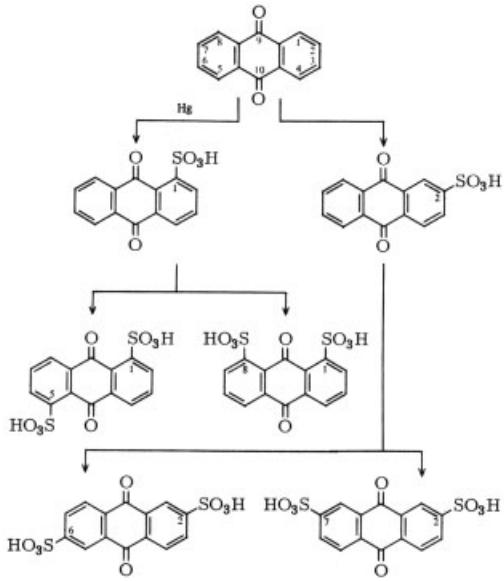

Although other catalysts, such as palladium, ruthenium, or rhodium, are also suitable for sulfonating anthraquinone at the α position, they are not used industrially [26–28].

Anthraquinone-1-sulfonic acid [*82-49-5*], K salt, also called diamant salt, M_r 288.3, is prepared by sulfonation in the presence of, e.g., 0.5 % mercury in 20 % oleum, an amount of SO_3 adequate for the sulfonation of about 60 % of the charge. The reaction is controlled in such a manner that about 50 % of the anthraquinone remains unsulfonated. After mercury is precipitated as the sulfide, the unreacted anthraquinone is precipitated by dilution and is collected by filtration. After washing and drying, it is recirculated as ‘reclaimed quinone‘. Mercury is removed completely from the filtrate, and the anthraquinone-1-sulfonic acid is then precipitated as the potassium salt [23].

Anthraquinone-2-sulfonic acid [*84-48-0*], Na salt, also called silver salt, M_r 288.3, is prepared by sulfonation in the absence of mercury (e.g., according to the procedure cited in [3,

p. 8]). The acid also can be obtained from the 1,6- (or 1,7-) disulfonic acid by desulfonation at the α position in dilute sulfuric acid [15].

Anthraquinone-1,5-disulfonic acid [*117-14-6*], Thiol Acid, M_r 368.3, and anthraquinone-1,8-disulfonic acid [*82-48-4*], Azine Acid, M_r 368.3, are produced by disulfonation under conditions similar to those for monosulfonation except for an increased concentration of SO_3. After sulfonation of the anthraquinone is complete, the 1,5-acid is precipitated by slow addition of dilute sulfuric acid, collected, and washed with dilute sulfuric acid. The filter cake is dissolved in hot water, the solution treated with charcoal, and the anthraquinone-1,5-disulfonic acid precipitated from the filtrate as its disodium salt by addition of saturated sodium chloride solution. The 1,8-acid is isolated as the potassium salt from the sulfuric acid mother liquors and the wash acids (see also [1] and [3, p. 10]). Mercury can be removed in a similar manner as given for anthraquinone-1-sulfonic acid (see, e.g., [23]).

2.2. Alkyl- and Arylanthraquinones

Alkyl- and arylanthraquinones are obtained from the corresponding benzene derivatives and phthalic acid anhydride via the benzoylbenzoic acids.

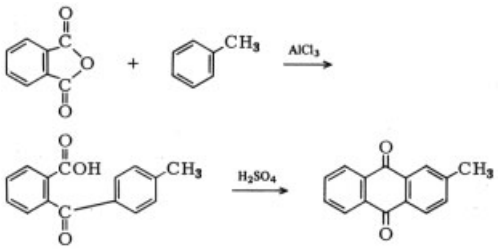

2-Methylanthraquinone is of great importance for the synthesis of dyes. 2-Alkylanthraquinones (e.g., 2-ethyl-, 2-isoamyl-, or 2-*tert*amylanthraquinone) are used as oxygen carriers in the preparation of hydrogen peroxides (→ Peroxo Compounds, Inorganic). If α-haloanthraquinones are heated in the presence of copper in inert, high-boiling solvents, dimerization (Ullmann reaction) yields the corresponding 1,1′-dianthraquinonyls. This reaction is used, for instance, in the preparation of starting materials for the synthesis of pyranthrone and flavanthrone (see Section 2.12.7).

2-Methylanthraquinone [*84-54-8*], M_r 222.2, *mp* 182 – 183 °C, is prepared from phthalic anhydride and toluene (see also [1], [2, p. 10], [5, p. 40]).

2-Ethylanthraquinone [*84-51-5*], M_r 235.3, *mp* 108 °C, is prepared from phthalic anhydride and ethylbenzene [29].

2.3. Haloanthraquinones

Haloanthraquinones are important intermediates for all classes of dyes, especially for vat and disperse dyes.

Unsubstituted halogenated anthraquinones are prepared mainly by replacement reactions or by synthesis of the nucleus. Only the chloro derivatives are of practical importance. Aminoanthraquinones, hydroxyanthraquinones, phenoxyanthraquinones, and anthrimides (dianthraquinonylamines) may be obtained by replacing the chlorine atom. The β-chloroanthraquinones are much more suitable for preparing β-aminoanthraquinones than the corresponding sulfonic acids. In contrast to the 1,5- and the 1,8-dichloroanthraquinones, a single chlorine atom of 2,3-dichloroanthraquinone can be replaced by ammonia to give 2-amino-3-chloroanthraquinone in good yield [5, p. 30].

Haloanthraquinones with additional substituents commonly are prepared by direct halogenation of the corresponding anthraquinones in water, hydrochloric acid, sulfuric acid, or organic solvents. Electron-withdrawing substituents, such as sulfonic acid or nitro groups, direct the halogenation to the other nucleus.

Anthraquinones with electron-donating substituents, such as amino or hydroxy groups, permit selective halogenation of one or the other nucleus by appropriate choice of reaction conditions. Only the chloro and bromo derivatives are commercial products.

Process Details. Direct chlorination of anthraquinone with chlorine or sulfuryl chloride, occurs preferentially at the α position;

however, the resulting product mixtures are difficult to separate. Only 1,4,5,8-tetrachloroanthraquinone of about 60 % purity [30] and 1-chloro-2methylanthraquinone are prepared commercially by this method [2, p. 60].

Replacement of sulfonic acid groups by chlorine is called the Fischer reaction in honor of its discoverer, A. Fischer. Sulfonic acid groups in both the α and β position are replaced.

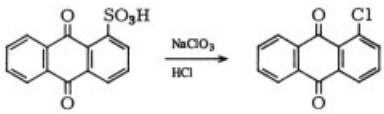

The reaction is carried out by slowly adding an aqueous solution of sodium chlorate to a hot hydrochloric acid solution of the anthraquinonesulfonic acid. The resulting, only slightly soluble chloroanthraquinones are isolated by filtration. The almost quantitative reaction is also used for the identification of anthraquinonesulfonic acids. The Fischer reaction of disulfonic acids can be directed in such a way that mostly monochloroanthraquinonesulfonic acids form.

A nitro group can be replaced by a chlorine atom at high temperature, generally above 200 °C, using chlorine or ionic chlorides, such as lithium chloride, in organic solvents [33]. A proposed industrial method is the reaction of chlorine with nitroanthraquinone in the melt [31]. Diluents for the high-melting nitroanthraquinones are lower melting compounds, such as chloroanthraquinone [32], trichlorobenzene [110], phthalic anhydride [42], and salt melts [34]. Chloroanthraquinones can be purified [35] and separated [36] by high-temperature distillation.

1,4-Dichloroanthraquinone may be obtained by replacing the hydroxy groups of 1,4-dihydroxyanthraquinone (quinizarin) with chlorine by heating with thionyl chloride [37] or with mixtures of phosphoryl chloride and phosphorus pentachloride [38].

The replacement of amino groups by chlorine through diazotization and treatment with cuprous chloride (Sandmeyer reaction) is used commercially only in special cases (see Section 2.5).

A few chloroanthraquinones can be obtained by synthesis of the nucleus via the corresponding benzoylbenzoic acids. Because β-anthraquinonesulfonic acids react slowly in the Fischer reaction, industrial production of 2-chloroanthraquinone starts generally from phthalic anhydride and chlorobenzene [2, pp. 8, 11].

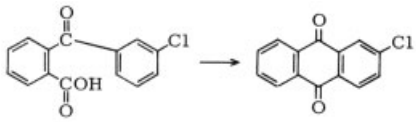

1-Chloroanthraquinone [*82-44-0*], M_r 242.7, *mp* 162 °C, is prepared from anthraquinone-1-sulfonic acid [4, p. 29], [39], or from 1-nitroanthraquinone [31], [32]. Purification by high-temperature distillation [35], [36].

2-Chloroanthraquinone [*131-09-9*], M_r 242.7, *mp* 210 °C, is prepared by synthesis of the nucleus starting with phthalic anhydride and chlorobenzene [2, pp. 8, 11].

1,5-Dichloroanthraquinone [*82-46-2*], M_r 277.1, *mp* 251 °C, is prepared by Fischer reaction from anthraquinone-1,5-disulfonic acid, by addition of chlorobenzenes [40], or from 1,5-dinitroanthraquinone [32] by replacement of NO_2 .

1,8-Dichloroanthraquinone [*82-43-9*], M_r 277.1, *mp* 203 °C, is prepared from anthraquinone-1,8-disulfonic acid [3, p. 15] or from 1,8-dinitroanthraquinone [32].

1,4-Dichloroanthraquinone [*602-25-5*], M_r 277.1, *mp* 187 °C, is prepared from 1,4-dihydroxyanthraquinone [37], [38] or by synthesis of the nucleus from phthalide and 1,4-dichlorobenzene [41].

2,3-Dichloroanthraquinone [*84-45-7*], M_r 277.1, is prepared by synthesis of the nucleus from phthalic anhydride and 1,2-dichlorobenzene [2, p. 10].

1,4,5,8-Tetrachloroanthraquinone [*81-58-3*], M_r 346.0, *mp* 342 °C, is prepared by chlorination of anthraquinone in sulfuric acid [30] or by sulfonation of 1,8-dichloroanthraquinone followed by Fischer reaction [43].

1-Chloro-2-methylanthraquinone [*129-35-1*], M_r 256.7, *mp* 171 °C, is prepared from 2-methylanthraquinone by direct chlorination in oleum [2, p. 60] or organic solvents [44].

2-Chloro-3-methylanthraquinone [*14684-09-4*], M_r 256.7, is prepared by synthesis of the nucleus from phthalic anhydride and 2-chlorotoluene [45].

2.4. Nitroanthraquinones

Nitration of anthraquinones has considerable industrial importance. Examples of compounds that may be nitrated are anthraquinone and the halo-, hydroxy-, amino-, and acylaminoanthraquinones. As a rule the nitro group attacks at the α position; choice of reaction conditions frequently permits mono- or dinitration.

Prior to the discovery of α-sulfonation of anthraquinone, nitration was the only useful method for preparing α-substituted anthraquinones. The nitro group of α-nitroanthraquinones may be replaced in a manner similar to the sulfonic acid moiety, e.g., by chlorine atoms and amino, hydroxy, alkoxy, or mercapto groups. Reduction readily yields aminoanthraquinones. Nitration of anthraquinone has gained increasing importance because of environmental considerations, this method offering an economical alternative to α-sulfonation.

1-Nitro-2-methylanthraquinone is the starting material for 1-amino-4-bromo-2-methylanthraquinone, 1-aminoanthraquinone-2-aldehyde, and 1-aminoanthraquinone-2-carboxylic acid derivatives.

Process Details. Substituted anthraquinones usually are nitrated in concentrated sulfuric acid using stoichiometric quantities of nitric acid or nitric – sulfuric mixed acid. This method always produces considerable amounts of dinitro compound in addition to the mononitro compound. Even the best methodology permits preparation of only 75 – 80 % of the theoretical amount of 1-nitroanthraquinone. Mononitration in highly concentrated nitric acid also has been developed to an acceptable procedure by avoiding concentration levels that can lead to explosions [46–49]. Dinitration can be retarded significantly during nitration with mixed acids by increasing the water content [50–52]. Other procedures specify the presence of organic solvents, such as halogenated hydrocarbons [54], [55], [56], sulfolane (tetrahydrothiophene-1,1-dioxide) [57], hydrofluoric acid [58], or phosphoric acid [59]. Some of these methods can be adapted to continuous processing.

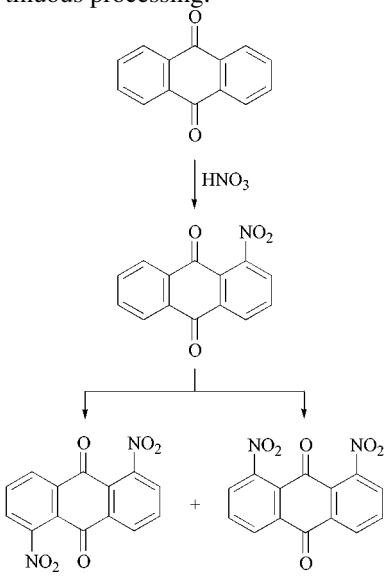

Mononitration of 2-methylanthraquinone leads to 2-methyl-1-nitroanthraquinone [2, p. 13]. Nitration of anthraquinone-1-sulfonic acid produces a mixture of 1-nitroanthraquinone-5-sulfonic acid and 1-nitroanthraquinone-8-sulfonic acid, whereas nitration of anthraquinone-2-sulfonic acid produces a mixture of 1-nitroanthraquinone-6-sulfonic acid and 1-nitroanthraquinone-7-sulfonic acid.

Halogen exchange by nitrite is an unimportant method of preparing nitroanthraquinones. There are only a few patents directed toward the synthesis of more highly substituted nitroanthraquinones, e.g., [82].

Finally, preparations of nitroanthraquinones by means of nuclear syntheses, for instance, by addition of butadiene to 5-nitronaphthoquinone (or its oxime), have also been proposed [83], [84].

Several methods have been developed to remove 2-nitroanthraquinone and dinitroanthraquinones from 1-nitroanthraquinone. The chemical purification processes are based on the treatment of the crude product with sodium sulfite

[60] or alkaline reagents, such as amines [61], alcoholates [62], or hydrazine [63], which cause preferential conversion of the contaminants to compounds that can be separated easily on the basis of their different solubilities. The more economical physical separation methods incorporate prepurification into the nitration process (e.g., crystallization from nitric acid [64]), or the isolated crude 1-nitroanthraquinone is purified with organic solvents, such as nitrobenzene [65] or amides [66]. Pure 1-nitroanthraquinone is then obtained by high-temperature distillation [67], [68].

Purification may also be deferred to the aminoanthraquinone stage (see page 13.

Modification of nitration conditions permits *dinitration* of anthraquinone. Nitration in highly concentrated nitric acid, as cited in [69], is for safety reasons unsuitable for industrial production [70]. However, addition of sulfuric acid makes it possible to work with a margin of safety [71]. The preferred method for dinitration is still the use of mixed nitric – sulfuric acid [72], [73]. However, nitration in organic solvents [74] or hydrofluoric acid [75] have been proposed also.

Apart from approximately equal portions of 1,5- and 1,8-dinitroanthraquinones, dinitration produces α,β-dinitroanthraquinone. The direct isolation of 1,5- and 1,8-dinitroanthraquinone is possible from the nitration mixtures [76] or, starting from isolated crude dinitroanthraquinone mixtures, by fractional crystallization from organic solvents, e.g., nitrobenzene [77], *N*-methylpyrrolidone [78], or high-boiling ketones [79]. Other procedures that simply separate the α,β-dinitroanthraquinones have been described also [80].

1-Nitroanthraquinone [*82-34-8*], M_r 253.2, *mp* 228 °C, is prepared by nitration of anthraquinone in nitric acid (danger of explosion!) [46–49], sulfuric acid [50–52], organic solvents [54–57], phosphoric acid [59], or hydrogen fluoride [58].

Purification from organic solvents by addition of bases, such as amines [61], alcoholates [62], or hydrazine [63], or from water by addition of sodium sulfite [60]. 1-Nitroanthraquinone is crystallized from nitric acid [64], from organic solvents, for example, nitrobenzene [65], or from amides [66]. The resulting 1-nitroanthraquinone may be finally purified by high-temperature distillation [67], [68].

1,5-Dinitro- and 1,8-dinitroanthraquinones [*82-35-9*], [*129-39-5*], M_r 298.2, are prepared by nitration of anthraquinone with nitric acid in sulfuric acid [72], [73], if necessary by addition of organic solvents [74], or in hydrogen fluoride [75].

For isolation and purification of mixtures of 1,5- (and 1,8-) dinitroanthraquinone, see [80]. Isolation of 1,5- (and 1,8-) dinitroanthraquinone is described in [77–79].

1-Nitroanthraquinone-5- (and -8-) sulfonic acids [*82-50-8*], [*129-37-3*], M_r 333.3, are prepared by nitration of anthraquinone-1-sulfonic acid [81].

2-Methyl-1-nitroanthraquinone [*129-15-7*], M_r 267.7, is prepared by nitration of 2-methylanthraquinone [2, p. 13].

2.5. Aminoanthraquinones

2.5.1. General Aspects

Aminoanthraquinones, especially 1-amino- and 1,5-diaminoanthraquinone, are key products for essentially all classes of anthraquinone dyes. Important production methods are the replacement of sulfonic acid and nitro groups or of halogen atoms by ammonia or primary or secondary amines. With 1,4-dihydroxy-, 1,4-aminohydroxy-, and 1,4-diaminoanthraquinones, the replacement of hydroxy and amino groups is also successful. Primary aminoanthraquinones are also prepared by reduction of nitroanthraquinones. Modifications of the amino functions also have industrial importance. These include alkylation, arylation, acylation, and hydrolysis of acylaminoanthraquinones. The choice of production method for a desired aminoanthraquinone depends on the position and type of the amino function, as well as on the availability of starting materials.

A special method for the production of 1-aminoanthraquinone consists in the catalytic hydrogenation of 5-nitro-1,4,4 a,9 a-tetrahydroanthraquinone, which is obtained by reacting butadiene and 5-nitronaphthoquinone (see [85]).

Aminoanthraquinones may be purified by recrystallization, distillation, or sublimation. However, purification is preferred at the stage of the nitroanthraquinones or the anthraquinonesulfonic acids.

In most situations the presence of the amino group facilitates the introduction of further substituents into the same ring. Reaction media and conditions as well as type and position of the amino group determine the position of substitution.

In addition to conversion of existing amino groups to other amino functions, the amino groups also may be replaced by halogen or hydroxy substituents through diazotization followed by the Sandmeyer reaction.

2.5.1.1. Replacement Reactions

$SO_3H \rightarrow NHR$ (R = H or alkyl). The replacement of sulfonic acid groups by amino or alkylamino groups can be carried out successfully with both unsubstituted and further substituted anthraquinonesulfonic acids. Depending on the choice of reaction conditions, diaminoanthraquinones and aminoanthraquinonesulfonic acids may be obtained from disulfonic acids. The α- and β-sulfonic acids react both with ammonia. In the reaction with alkylamines, only the α derivatives give commercially important products. Yields are determined in part by the number of sulfonic acid groups and in part by the type of amine and are on the order of 70 – 90 %.

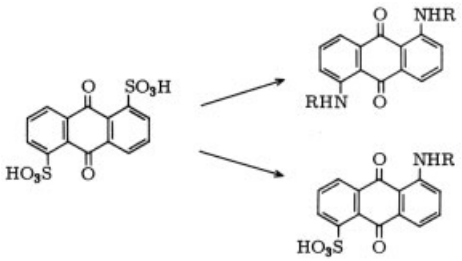

The exchange reactions are carried out in aqueous medium at 120 – 200 °C and under pressure. To avoid attachment of displaced sulfite, an oxidizing agent, e.g., *m*-nitrobenzenesulfonic acid, is usually added.

Halogen $\rightarrow NH_2$ or NHR. Of great importance industrially is the replacement of chlorine and bromine atoms in the α or β positions by amines in substituted and unsubstituted haloanthraquinones. Reagents for this purpose are ammonia, aliphatic and aromatic amines, aminoanthraquinones, and amides, such as *p*-toluenesulfonamide or phthalimide. Because of the increased reactivity of the α positions, the halogen atoms at the α positions of α,β-dihaloanthraquinones may be reacted selectively.

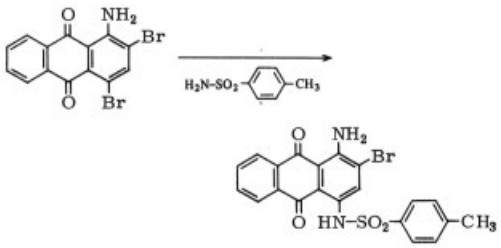

The partial replacement of halogen in α,α- and β,β-*dihaloanthraquinones,* for example, in 2,3-dichloroanthraquinone, is also a known process.

Haloanthraquinones without sulfonic acid groups are reacted in the presence of excess amine or in such organic solvents as alcohol, halobenzene, nitrobenzene, naphthalene, or *N*-ethylcarbazole. Reaction temperatures range from 100 to 250 °C, and the reaction is carried out in the presence of acid-neutralizing agents. Alkali carbonates, acetates, and stearates are suitable for this purpose. The reaction is accelerated by the addition of copper catalysts. In many cases, for example, in the preparation of anthrimides, the copper catalyst is necessary.

Acylation of *1-amino-4-haloanthraquinones* may increase the reactivity of the halogen atom.

Haloanthraquinones with sulfonic acid groups generally are reacted with amines in water or water–alcohol mixtures at temperatures up to 100 °C in the presence of alkali carbonates, alkali hydrogen carbonates, acetates, or hydroxides. The addition of a copper catalyst is absolutely essential in these reactions. Its action is reputed to be enhanced by adding Fe(II) salts [86].

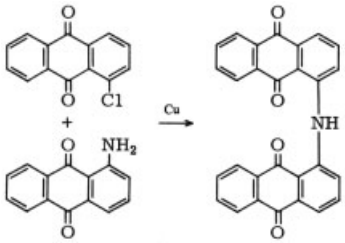

Of great industrial importance are the reactions of 1-amino-4-bromoanthraquinone-2-sulfonic acid (bromamine acid). Yields depend strongly on the choice of the reaction conditions and the reactivity of the amine. Strongly basic amines require the addition of alkali-metal hydroxides. Sterically hindered amines, such as 2,4,6-trimethylaniline, react very slowly.

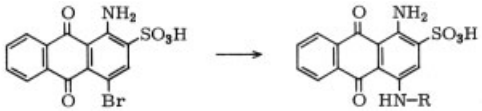

Amines with very low basicity are unreactive or react slowly. A side reaction of the bromamine acid reaction is the reductive elimination of the halogen atom and formation of 1-aminoanthraquinone-2-sulfonic acid or the replacement by a hydroxy group to form 1-amino-4-hydroxyanthraquinone-2-sulfonic acid.

$NO_2 \rightarrow NH_2$ or NHR. Aminoanthraquinones with primary amino groups may be prepared by replacement reactions or, even more simply, by reduction of the corresponding nitroanthraquinones. The industrial reducing agents are usually sodium sulfide or sodium hydrogen sulfide [87]. Reduction also may be accomplished with sulfur dioxide in ≈ 60 % sulfuric acid in the presence of catalytic quantities of iodine [88]. Other methods, particularly for the preparation of 1-aminoanthraquinone, include reduction by hydrazine in aqueous sodium hydroxide [89], by metals such as iron in sulfuric acid [90], and by hydrogenation in organic solvents in the presence of palladium or Raney nickel [91]. If catalytic hydrogenation is carried out in dilute sodium hydroxide solution, a purification effect may take place if the 1-nitroanthraquinone is contaminated with dinitroanthraquinone. A stepwise reduction with purification at the hydroxylamino stage has been proposed also [92]. Highly purified 1-aminoanthraquinone is obtained by high-temperature distillation [68].

Replacement of nitro groups by amino groups using ammonia leads to good yields and high quality 1-aminoanthraquinone and 1,5-diaminoanthraquinone. The reaction commonly is carried out at 140 – 170 °C, optionally under pressure, in organic solvents, such as alcohols [93], amides [94], xylene [95], chlorobenzene [96], or tetramethylene sulfone [97]. The use of ammonium salts or urea in place of ammonia has also been proposed [94].

Replacement of nitro groups by alkylamines or arylamines is favored in the production of secondary amines. Replacement reactions with 1-aminoanthraquinone to form anthrimides are also successful [98].

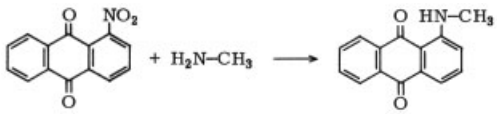

When dinitroanthraquinones are used as starting materials, the choice of reaction conditions makes a partial exchange possible. The reaction is carried out in water or organic solvents, or in an excess of the amine.

$OH \rightarrow NH_2$ or NHR (R = H, alkyl, aryl). The replacement of the hydroxy group by the amino group is limited to 1,4-dihydroxy- and 1-amino-4-hydroxyanthraquinone. It occurs via the 2,3-dihydro compounds and is reversible.

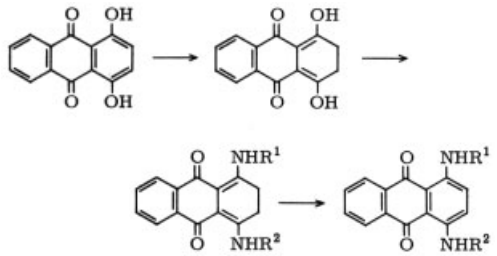

R = H, Alkyl, Aryl

Exchange reactions are carried out using ammonia, aliphatic and aromatic amines, and mixtures of amines. The starting material is the isolated 2,3-dihydro compound, generally as a mixture with dihydroxyanthraquinone. Alterna-

tively, the intermediate may be prepared in-situ by addition of reducing agents such as zinc–hydrochloric acid or iron – acetic acid. The reaction media are water, alcohol, or excess amine. The primary product, 2,3-dihydro-1,4-diaminoanthraquinone, is oxidized to 1,4-diaminoanthraquinone either during the reaction mediated by dihydroxyanthraquinone or afterwards, by reaction with air. It is also possible to isolate the dihydro compound and then oxidize it with oleum, manganese dioxide – sulfuric acid, or nitrobenzene in the presence of piperidine.

In 1,4-dihydroxyanthraquinone, both hydroxy groups are replaced in an excess amine. In the presence of additional diluents, such as alcohols, generally only one hydroxy group is exchanged. As expected, sterically hindered amines are not very reactive.

$NH_2 \rightarrow NHR$. The amino groups of 1,4-diaminoanthraquinones may be replaced by other amino groups via the 2,3-dihydro compounds.

OR → NHR. Replacement of ether functions by amino groups occasionally offers advantages over the other methods, especially in reactions with reducing amines.

2.5.1.2. Modifications of the Amino Functions

$NH_2 \rightarrow NHR$ (R = alkyl). The preparation of alkylaminoanthraquinones by alkylation of aminoanthraquinones can be achieved by commonly employed methods, e.g., with dimethyl sulfate, alkyl halides, or esters of toluenesulfonic acid. A special process is the methylation with methanol in sulfuric acid [108]. When used on aminohydroxy- and diaminoanthraquinones this process gives mixed products of interest to dyers.

$NH_2 \rightarrow NHR$ (R = aryl). The preparation of simple arylanthraquinones by arylation of aminoanthraquinones with aromatic halides, such as chlorobenzenes or chloronaphthalenes, has no industrial importance. However, arylation is used for the preparation of dianthraquinonylamines and similar compounds, which are classified under the term *anthrimides,* e.g., dianthrimide, see page 14. In this case aminoanthraquinones are reacted with chloroanthraquinones or nitroanthraquinones. The reaction is carried out in high-boiling solvents, such as naphthalene or nitrobenzene, at temperatures of 150 – 250 °C. Certain cases do not require a solvent (so-called "baking process").

Acid-binding agents used in the preparation of anthrimides include alkali carbonates, acetates, oleates, stearates, and alkaline-earth oxides. Catalysts in the reactions of haloanthraquinones are copper and copper salts. Nitroanthraquinones also may be reacted without a catalyst. Anthrimides may be isolated by filtration or distilling the solvent away. Inorganic salts are removed by boiling with water or dilute mineral acids.

A special reaction for the preparation of anthrimides is the oxidation of aminoanthraquinones in aqueous sulfuric acid: 1-aminoanthraquinone yields aminopolyanthrimides [99]. Mixtures of 1-aminoanthraquinone and 1,4-diaminoanthraquinone yield diaminodianthrimide [100].

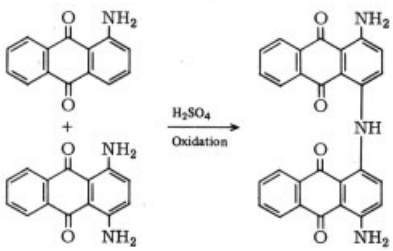

Various anthrimides are suitable as vat dyes, but they find their main use in the preparation of anthrimidecarbazoles (see Section 2.11.2).

$NH_2 \rightarrow NH$ – Heteroaryl. Of importance here are mostly the triazinylaminoanthraquinones, where R^1 and/or R^2 may constitute anthraquinonylamino residues.

Due to the high reactivity of chlorotriazines, one does not require a catalyst or, acid-binding agents. In addition to high boiling organic solvents such as nitrobenzene, phenols have proved valuable in the preparation of trisanthraquinonylaminotriazines [101].

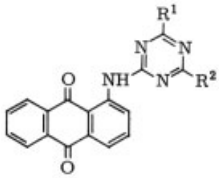

The reaction may be accelerated by adding tertiary amines, and in special cases by adding catalytic quantities of a Lewis acid such as $AlCl_3$ or $FeCl_3$.

$NH_2 \rightarrow$ NH-CO-R. The most important method of preparing acylaminoanthraquinones is the acylation of aminoanthraquinones with acid chlorides, preferably in an organic solvent, such as nitrobenzene or dichlorobenzene. A further method is the reaction of aminoanthraquinones with carboxylic acids or their anhydrides in oleum. Depending on the position of the amino groups and the nature of the acid chloride, diaminoanthraquinones also may be acylated in one position only, giving 60 – 90 % yields:

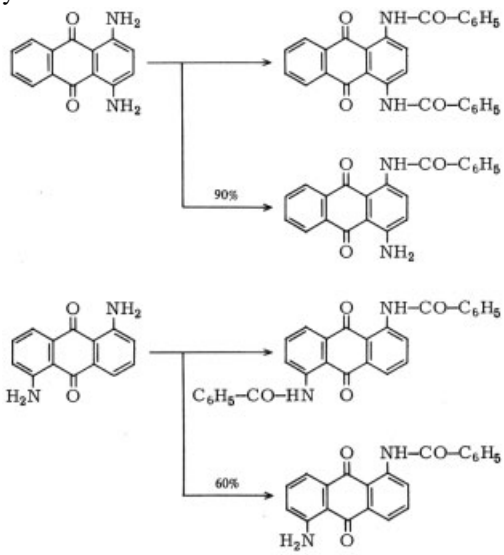

NHR $\rightarrow NH_2$. Secondary or tertiary alkyl aminoanthraquinones are converted into aminoanthraquinones by removal of the alkyl residue under the influence of acid or alkaline agents. This method is also applicable to cycloalkyl- [102] and aralkylaminoanthraquinones and facilitates the preparation of certain aminoanthraquinones that are accessible only with difficulty. In the reaction of sterically hindered arylamines with 1-amino-4-bromoanthraquinone-2-sulfonic acid (bromamine acid), this cleavage occurs as an undesirable side reaction, forming 1,4-diaminoanthraquinone-2-sulfonic acid.

NH-CO-R and NH-SO_2-R $\rightarrow NH_2$. Acylaminoanthraquinones are hydrolyzed in dilute to concentrated sulfuric acid as well as in 2 – 10 % aqueous sodium hydroxide at temperatures ranging from 20 to 150 °C. The sulfonamides are more stable in alkaline medium than in an acidic medium. They are especially easily cleaved in concentrated sulfuric acid. Because arylsulfonylaminoanthraquinones are smoothly prepared from haloanthraquinones (see page 19), this method is suitable for the production of aminoanthraquinones otherwise difficult to prepare. Hydrolysis of arylamino groups depends also on additional substituents. For instance, alkaline hydrolysis of 1,5-bis(benzoylamino)-4-hydroxyanthraquinone causes hydrolysis of the 5-benzoylamino group, whereas acidic medium causes cleavage of the 1-benzoylamino group.

2.5.1.3. Substitutions on the Aminoanthraquinone Nucleus

Nucleus H $\rightarrow SO_3H$. 1-Aminoanthraquinone- 2-sulfonic acids are prepared by sulfonation of aminoanthraquinones in oleum or chlorosulfonic acid. In order to suppress the sulfonation and oxidation of the 4 position as much as possible, chlorosulfonic acid in organic solvents or, alternatively, oleum in the presence of sodium sulfate is used. When 1,4-diaminoanthraquinone is sulfonated in oleum, addition of boric acid (to form cyclic boric acid amides) directs the sulfonic acid group into the 6 position. Sulfonation of β-aminoanthraquinones has no commercial importance.

An alternative method for the introduction of sulfonic acid groups consists of the treatment of α-aminoanthraquinone with aqueous alkali-metal sulfites.

Nucleus H $\rightarrow$ Halogen. Aminoanthraquinones may be halogenated with the usual halogenation agents such as chlorine, bromine, and sulfuryl chloride. Suggested reaction media

are hydrochloric acid, oleum, glacial acetic acid, sodium chloride–aluminum chloride melts, and inert solvents, such as chlorobenzene, nitrobenzene, halogenated hydrocarbons, and amides. The reaction of 1-aminoanthraquinone with chlorine or sulfuryl chloride [1] in inert solvents such as nitrobenzene leads to 1-amino-2,4-dichloroanthraquinone, whereas addition of tetraalkylureas [103] leads to 1-amino-2,3,4-trichloroanthraquinone. However, if 1-amino- or 1-phthalimidoanthraquinone is chlorinated in oleum, a good yield of 1-amino-5,8-dichloroanthraquinones is obtained, if necessary after hydrolysis of intermediates:

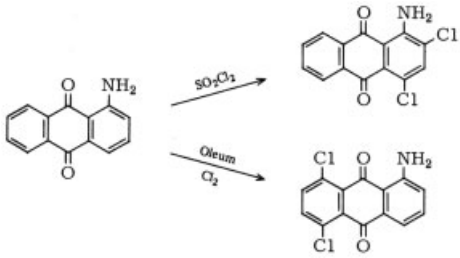

Use of 1-benzoylaminoanthraquinone as the starting material generally leads to substitution at the 4 position. Hydrolysis then yields the 1-amino-4-chloroanthraquinone. This compound may be obtained in better yields by chlorination of 1-formamidinoanthraquinone in oleum followed by hydrolytic cleavage of the protective group.

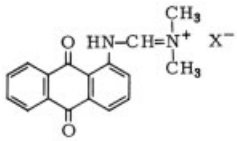

Bromination of 1-aminoanthraquinone in hydrochloric acid yields 1-amino-2,4-dibromoanthraquinone.

Chlorination of 1,4-diaminoanthraquinone in organic solvents with chlorine or sulfuryl chloride yields 1,4-diamino-2,3-dichloroanthraquinone; in boric acid–oleum, the 5,8-dichloro derivative is formed. In hydrochloric acid, 1-amino-2-methylanthraquinone and 1-aminoanthraquinone-2-sulfonic acid are brominated at the 4 position.

The 1-alkylaminoanthraquinones behave differently from the 1-aminoanthraquinones upon bromination in hydrochloric acid. In this case no substitution occurs at the 2 position, but the 1-alkylamino-4-bromoanthraquinones are formed in more than 90 % yield. By contrast, chlorination of 1-methylaminoanthraquinone in weak oleum [104] will give, in a manner similar to that of 1-aminoanthraquinone, the 5,8-dichloro-1-methylamino compound.

Nucleus $H \rightarrow CH_3$. Starting from 1-aminoanthraquinones, reaction with formaldehyde yields 1-amino-2-methylanthraquinones. The reaction proceeds in aqueous alkaline solutions via the leuco form of the aminoanthraquinone [115].

Nucleus $H \rightarrow NO_2$. In the nitration of aminoanthraquinones, which usually is carried out in sulfuric acid, the amino group must be protected by acylation or by formation of cyclic sulfimides. Location and number of the inserted nitro groups are determined to some extent by the nature of the protective group. For example, 1-ethoxycarbonylaminoanthraquinone yields the 1-amino-2,4-dinitroanthraquinone, whereas 1-phthalimidoanthraquinone yields the 4,5- (or 4,8-) dinitro derivatives [105]. If 1,4-diaminoanthraquinone is used as the starting material, 1,4-diamino-2-nitroanthraquinone is formed via the intermediate 1,4-bis(benzoylamino)anthraquinone. The nitro group enters the 5 position if the intermediate is the cyclic disulfimide **1**.

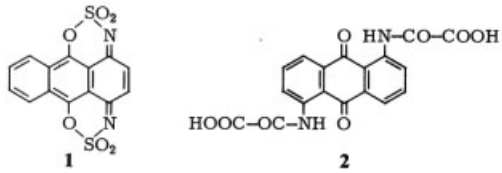

The dioxamic acid (**2**) of the 1,5-diaminoanthraquinone yields 1,5-diamino-4,8-dinitroanthraquinone after nitration and hydrolysis [106].

$SO_3H \rightarrow H$. In certain 4-substituted 1-aminoanthraquinone-2-sulfonic acids, replacement of the sulfonic acid by hydrogen is feasible by using reducing agents in aqueous alkaline medium. Sodium dithionite and glucose have been applied successfully here. This procedure is used in the preparation of 1-alkylamino- (or 1-arylamino-) 4-aminoanthraquinone:

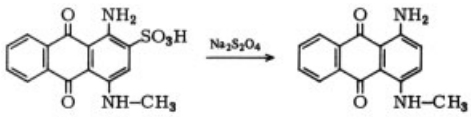

The procedure has certain advantages for the preparation of 1,4-diaminoanthraquinones with sensitive amino residues. An electrochemical reductive desulfonation of β-anthraquinonesulfonic acids has been proposed also [107].

2.5.2. Individual Aminoanthraquinones

1-Aminoanthraquinone [*82-45-1*], M_r 223.2, *mp* 252 – 253 °C, is prepared from anthraquinone-1-sulfonic acid in aqueous ammonia by replacement of SO_3H (see page 8), [5, p. 22], from 1-nitroanthraquinone in organic solvents by replacement of NO_2 by ammonia (see page 9), [93–97]; by reduction of 1-nitroanthraquinone with sodium sulfide in water [87] or in organic solvents [109], with hydrazine hydrate [89], with metal powders [90], or by catalytic hydrogenation. This last reaction can be carried out in organic solvents [91], in sodium hydroxide solution [92], or in dilute sulfuric acid [111]. Purification is carried out by recrystallization from organic solvents [112], [113] or sulfuric acid [114]. The greatest purity is obtained by distillation [68].

2-Aminoanthraquinone [*117-79-3*], M_r 223.2, *mp* 302 – 303 °C. Older methods used anthraquinone-2-sulfonic acid as the starting material, replacing the SO_3H group. The preferred method is replacement of the chlorine in 2-chloroanthraquinone (see page 8), [2, p. 20].

1-Amino-2-methylanthraquinone [*82-28-0*], M_r 237.3, *mp* 202 °C, is prepared by nitration of 2-methylanthraquinone [2, p. 13] followed by reduction or by methylation of 1-aminoanthraquinone with formaldehyde [115].

1,2-Diaminoanthraquinone [*1758-68-5*], M_r 238.2, *mp* 303 – 304 °C, is prepared from 1-aminoanthraquinone-2-sulfonic acid by SO_3H exchange (see page 8, [5, p. 98].

1,4-Diaminoanthraquinone [*128-95-0*], M_r 238.2, *mp* 268 °C, is prepared by hydroxyl replacement from 1,4-dihydroxyanthraquinone via the 2,3-dihydro compound and subsequent oxidation of the intermediate 2,3-dihydro-1,4-diaminoanthraquinone [5, p. 51].

1,5-Diaminoanthraquinone [*129-44-2*], M_r 238.2, *mp* 319 °C, is prepared by SO_3H exchange from anthraquinone-1,5-disulfonic acid (see page 8), [3, p. 14], [5, p. 43], by replacing NO_2 in 1,5-dinitroanthraquinone (see page 9), [116], or by reducing 1,5-dinitroanthraquinone with sodium sulfide or by catalytic hydrogenation [117].

1,8-Diaminoanthraquinone [*129-42-0*], M_r 238.2, *mp* 262 °C, is prepared by the methods given for 1,5-diaminoanthraquinone.

1,4,5,8-Tetraaminoanthraquinone [*2475-45-8*], M_r 268.3, *mp* 332 °C, is prepared from anthraquinone-1,5-dioxamic acid by nitration, hydrolysis, and reduction [3, p. 53].

1-Methylaminoanthraquinone [*82-38-2*], M_r 237.3, *mp* 170 °C, is prepared from anthraquinone-1-sulfonic acid by replacing SO_3H (see page 8), [3, p. 18], from 1-chloroanthraquinone by replacing chlorine (see page 8), [118], or from 1-nitroanthraquinone by replacing NO_2 (see page 9) [119], [120].

1-Isopropylaminoanthraquinone [*27354-18-3*], M_r 265.3, *mp* 187 °C, is prepared by replacement of SO_3H [121] or NO_2 [120], [121].

1,4-Bis(methylamino)anthraquinone [*2475–44–7*], M_r 266.3, is prepared by replacement of the hydroxy groups from 1,4-dihydroxyanthraquinone (see page 9) via the 2,3-dihydro compound [122].

1-Methylamino-4-(3-dimethylaminopropylamino)anthraquinone, (**3**), M_r 337.4, is prepared from 4-bromo-1-methylaminoanthraquinone (see page 8), [123].

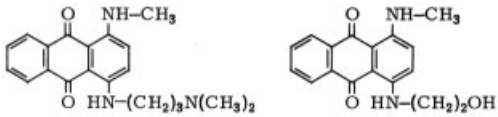

3 4

1-Methylamino-4-(2-hydroxyethylamino)anthraquinone [*2475-46-9*], (**4**), M_r 296.3, is prepared (in admixture with symmetrical and half-reacted exchange products) by replacement of NH_2 from 2,3-dihydro-1,4-diaminoanthraquinone with a mixture of methylamine and ethanolamine [124].

1-Amino-4-methylaminoanthraquinone [*1220–94–6*], M_r 252.2, is prepared from 1,4-diaminoanthraquinone by partial methylation [125], from 1-amino-4-methylaminoanthraquinone-2-sulfonic acid by reductive SO_3H cleavage (see page 12), [126], from 2,3-dihydro-1,4-dihydroxyanthraquinone by hydroxyl replacement (see page 9) [122], or from 2,3-dihydro-1,4-diaminoanthraquinone by replacement of NH_2 [127] (in admixture with 1,4-bismethylaminoanthraquinone).

1-Cyclohexylaminoanthraquinone [*1096-48-6*], M_r 305.4, is prepared by replacing SO_3H, chlorine, or NO_2 [119].

1,5-Bis-(2-nitrophenylamino)anthraquinone [*21982-51-4*], (**5**), M_r 480.4, is prepared from 1,5-diaminoanthraquinone and 2-chloronitrobenzene [128].

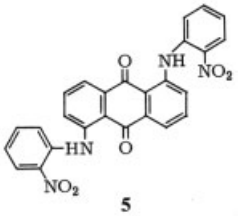

5

1-Benzoylaminoanthraquinone [*3571-23-1*], M_r 327.3, *mp* 248 – 249 °C, is prepared from 1-aminoanthraquinone and benzoyl chloride [129] or from 1-chloroanthraquinone and benzamide [130].

1,4-Bis-(benzoylamino)anthraquinone [*2987-68-0*], Indanthren Red 5 GK, M_r 446.4, is prepared from 1,4-diaminoanthraquinone and benzoyl chloride [4, p. 40].

1,5-Bis(benzoylamino)anthraquinone [*82-18-8*], Indanthren Yellow GK, M_r 446.4, *mp* 350 °C, is prepared from 1,5-diaminoanthraquinone with benzoyl chloride [131] or from 1,5-dichloroanthraquinone with benzamide [130], [132].

1-Amino-4-benzoylaminoanthraquinone [*81-46-9*], M_r 342.3, is prepared from 1,4-diaminoanthraquinone with benzoyl chloride [3, p. 4].

1-Amino-5-benzoylaminoanthraquinone [*117-06-6*], M_r 342.3, *mp* 244 – 245 °C, is prepared from 1,5-diaminoanthraquinone with benzoyl chloride [3, p. 4] or from 1-chloro-5-benzoylaminoanthraquinone with arylsulfamides and subsequent hydrolysis [133].

1,1′-Dianthraquinonylamine [*82-22-4*], dianthrimide, (**6**), M_r 429.4, is prepared by reacting 1-aminoanthraquinone with 1-chloroanthraquinone in nitrobenzene [4, p. 29] or water [134]. Also from 1-aminoanthraquinone and 1-nitroanthraquinone [98].

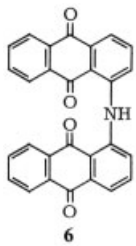

6

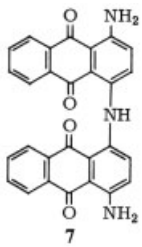

7

4,4′-Diamino-1,1′-dianthraquinonylamine [*128-87-0*], (**7**), M_r 459.4, is prepared by nitration of dianthrimide followed by reduction [4, p. 30], [5, p. 138] or by oxidation of a mixture of 1-aminoanthraquinone and 1,4-diaminoanthraquinone in aqueous sulfuric acid [100].

4,4′-Bis(benzoylamino)–1,1′-dianthraquinonylamine [*128-79-0*], (**8**), M_r 667.7, is prepared from 4,4′-diamino-1,1′-dianthrimide with benzoyl chloride [4, p. 31], [5, p. 139].

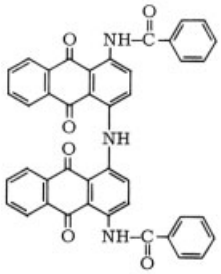

8

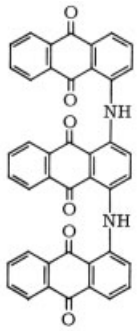

9

1,4-Bis-(1-anthraquinonylamino)anthraquinone [*116-76-7*], 1,1′,4′,1″-trianthrimide, (**9**), M_r 650.7, is prepared from 1 mol of 1,4-diaminoanthraquinone and 2 mol of 1-chloroanthraquinone [5, p. 102], [134].

***N,N′*-Bis-(1-anthraquinonyl)isophthalamide** [*3627-47-2*], Indanthren Yellow 5 GK, (**10**), M_r 576.6, is prepared from 1-aminoanthraquinone and isophthaloyl chloride [12, p. 596].

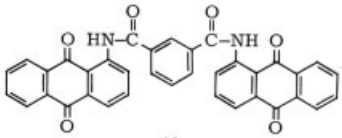

10

1-Amino-4-chloroanthraquinone [*2872-47-1*], M_r 257.7, *mp* 179 – 180 °C, is prepared by chlorination of 1-benzoylaminoanthraquinone with sulfuryl chloride [135] and subsequent hydrolysis or by chlorination of anthraquinonyl-1-formamidinium salt followed by hydrolysis [136].

2-Amino-3-chloroanthraquinone [*84-46-8*], M_r 257.7, *mp* 280 – 283 °C, is prepared from 2,3-dichloroanthraquinone by chlorine replacement (see page 8) [5, p. 30].

1-Amino-4-chloro-2-methylanthraquinone [*3225-97-6*], M_r 271.7, *mp* 255 – 256 °C, is prepared from 1-amino-2-methylanthraquinone by chlorination with sulfuryl chloride [137] or chlorine [138].

1-Benzoylamino-4-chloroanthraquinone [*81-45-8*], M_r 361.8, is prepared from 1-aminoanthraquinone by benzoylation in nitrobenzene [3, p. 13] and subsequent chlorination with sulfuryl chloride.

1-Benzoylamino-5-chloroanthraquinone [*117-05-5*], M_r 361.8, is prepared from 1-amino-5-benzoylaminoanthraquinone by diazotization and Sandmeyer reaction with Cu_2Cl_2 or by benzoylation of 1-amino-5-chloroanthraquinone in *o*-dichlorobenzene using benzoyl chloride [4, p. 16].

1-Amino-2,4-dichloroanthraquinone [*13432-32-1*], M_r 292.1, *mp* 217 – 219 °C, is prepared by chlorination of 1-aminoanthraquinone in nitrobenzene [139].

1-Amino-5,8-dichloroanthraquinone [*3223-94-7*], M_r 292.1, *mp* 199 °C, is prepared by chlorination of 1-aminoanthraquinone in chlorosulfonic acid or in oleum with a low percentage of SO_3 [140].

1,4-Diamino-2,3-dichloroanthraquinone [*81-42-5*], M_r 307.1, is prepared by chlorination of 1,4-diaminoanthraquinone or leuco-1,4-diaminoanthraquinone [4, p. 18] with sulfuryl chloride.

1-Amino-2,4-dibromoanthraquinone [*81-49-2*], M_r 381.0, *mp* 221 °C, is prepared from 1-aminoanthraquinone by bromination in dilute mineral acids [3, p. 6].

4-Bromo-1-methylaminoanthraquinone [*128-93-8*], M_r 316.1, *mp* 194 °C, is prepared from 1-methylaminoanthraquinone by bromination [141].

4-Bromo-1-isopropylaminoanthraquinone [*23573-29-7*], M_r 344.2, is prepared from 1-isopropylaminoanthraquinone by bromination [141].

4-Bromo-1-cyclohexylaminoanthraquinone [*14233-28-4*], M_r 384.4, is prepared from 1-cyclohexylaminoanthraquinone by bromination [141].

1-Amino-2-bromo-4-(dimethylaminopropylamino)anthraquinone [*51818-35-0*], (**11**), M_r 402.3, is prepared from 1-amino-2,4-dibromoanthraquinone and dimethylaminopropylamine by bromine exchange [142].

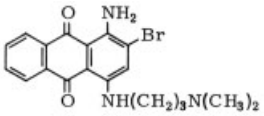

11

1,4-Diamino-2-nitroanthraquinone, M_r 283.2, is prepared from 1,4-bis(benzoylamino)anthraquinone by nitration in a solvent and subsequent hydrolysis [143].

1,4-Diamino-5-nitroanthraquinone [*82-33-7*], M_r 283.2, is prepared from 1,4-diaminoanthraquinone via the cyclic disulfimide (the structure is shown in page 12), which is nitrated in sulfuric acid and then hydrolyzed [3, p. 20].

1,5-Diamino-4,8-dinitroanthraquinone [*10262-79-0*], M_r 328.3, is prepared from anthraquinone-1,5-bis-oxamidic acid by nitration and hydrolysis [106] or from anthraquinone-1,5-bis(formamidinium chloride) by nitration and hydrolysis in dilute sulfuric acid [144].

1,4-Diamino-2-bromo-5-nitroanthraquinone, M_r 362.1, is prepared from 1-amino-2,4-dibromo-5-nitroanthraquinone and *p*-toluenesulfonamide followed by hydrolysis [145].

1-Aminoanthraquinone-2-sulfonic acid [*83-62-5*], M_r 303.3, is prepared from 1-aminoanthraquinone by sulfonation with $ClSO_3H$ in an organic solvent (if nitrobenzene, danger of explosion) [5, p. 214], [146], [147], [149], or with oleum in the presence of Na_2SO_4 [148].

1-Amino-4-bromoanthraquinone-2-sulfonic acid [*116-81-4*], bromamine acid, (**12**), M_r 382.2, is prepared from 1-aminoanthraquinone-2-sulfonic acid by bromination in water [5, p. 214], [151] or in dilute acid [149], [150]. Alternatively, it can be prepared by a one-pot process: sulfonation and bromination in sulfuric acid without isolation of intermediates [152].

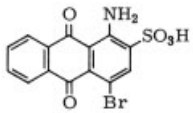

12

1,4-Diaminoanthraquinone-6-sulfonic acid [*64910-84-5*], M_r 318.3, is prepared by sulfonation of 1,4-diaminoanthraquinone in oleum in the presence of boric acid.

1-Aminoanthraquinone-2,5- (or -2,8-) disulfonic acid [*58294-46-5*], M_r 383.4, is prepared from 1-aminoanthraquinone-5- (or -8-) sulfonic acid by sulfonation [145], or from 1-cyclohexylaminoanthraquinone-5-sulfonic acid in oleum [102].

1-Aminoanthraquinone-2,6- (or -2,7-) disulfonic acid, M_r 383.4, is prepared from 1-nitroanthraquinone-6- (or -7-) sulfonic acid and Na_2S by formation of 1-amino-2-mercaptoanthraquinone-6- (or -7-) sulfonic acid followed by oxidation [153].

1-Isopropylaminoanthraquinone-5-sulfonic acid [*33175-83-6*], M_r 345.4, is prepared from anthraquinone-1,5-disulfonic acid by replacement of one SO_3H [154].

1-Cyclohexylaminoanthraquinone-5-sulfonic acid [*33175-82-5*], M_r 385.4, is prepared from anthraquinone-1,5-disulfonic acid by replacement of one SO_3H [155].

1,4-Bis-(*p*-toluidino)anthraquinone-2′,2″-disulfonic acid [*3443-90-1*], (**13**), M_r 578.6, is prepared from 1,4-dihydroxyanthraquinone and *p*-toluidine by replacement of hydroxyl followed by sulfonation [5, p. 215].

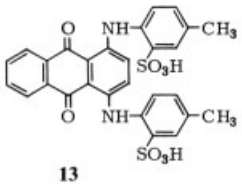

13

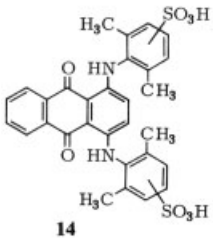

14

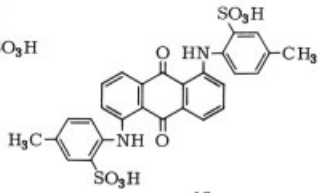

15

1,4-Bis-(2,6-dimethylanilino)anthraquinone-x′,x″-disulfonic acid, (**14**), M_r 606.7, is prepared from 1,4-dihydroxyanthraquinone and 2,6-dimethylaniline followed by sulfonation [156].

1,5-Bis-(*p*-toluidino)anthraquinone-2′,2″-disulfonic acid [*117-04-4*], (**15**), M_r 578.6, is prepared by replacement of NO_2 [157] or Cl [158] followed by sulfonation [159].

1,8-Bis-(*p*-toluidino)anthraquinone-2′,2″-disulfonic acid, (**16**), M_r 578.6, is prepared by replacement of NO_2 [157] or Cl followed by sulfonation [159].

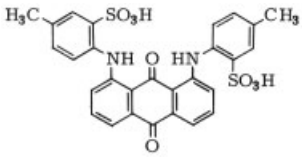

16

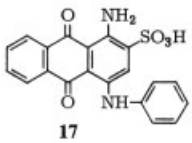

17

1-Amino-4-anilinoanthraquinone-2-sulfonic acid [*2786-71-2*], (**17**), M_r 394.4, is prepared from bromamine acid and aniline [3, p. 41]. Compound **17** is a blue acid dye (e.g., Alizarine Saphirol A).

1-Amino-4-(*p*-methylaminomethylanilino)anthraquinone-2-sulfonic acid [*64135-01-9*], (**18**), M_r 437.5, is prepared from bromamine acid and *N*-(4-amino-benzyl) methylamine [160].

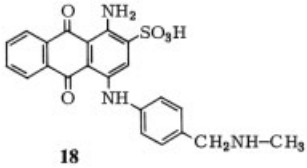

18

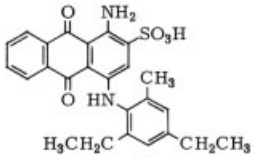

19

1-Amino-4-(2,4-diethyl-6-methylphenylamino)anthraquinone-2-sulfonic acid [*20074-70-8*], (**19**), M_r 464.5, is prepared from bromamine acid and 2,4-diethyl-6-methylaniline [161].

1-Amino-4-(4′-acetylaminoanilino)anthraquinone-2-sulfonic acid, a blue acid dye, (**20**), M_r 451.1, is prepared from bromamine acid and 4-acetylaminoaniline [2, p. 135].

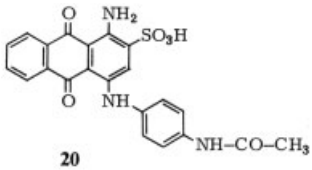

20

1-Amino-4-cyclohexylaminoanthraquinone-2-sulfonic acid [*5617-28-7*], is a brilliant blue acid dye (e.g., Alizarin Brilliant Pure Blue R), (**21**), M_r 400.5; it is prepared from bromamine acid and cyclohexylamine in the presence of NaOH [3, p. 30].

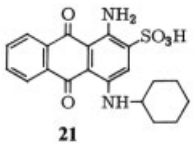

21

4-Bromo-1-isopropylaminoanthraquinone-5-sulfonic acid, (**22**), M_r 424.3, is prepared by bromination [397].

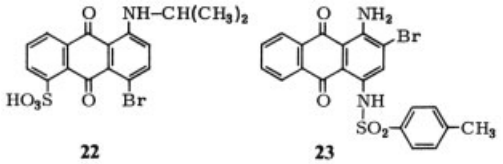

1-Amino-2-bromo-4-(4′-methylphenylsulfonamido)anthraquinone [*26868-32-6*], (**23**), M_r 471.3, is prepared from 1-amino-2,4-dibromoanthraquinone and *p*-toluenesulfonamide [162].

2.5.3. Diazotization of Aminoanthraquinones

Aminoanthraquinones may be diazotized by the usual methods in sulfuric acid. On an industrial scale, the diazotization of an aminoanthraquinone is accomplished by the conversion of 1-amino-5-benzoylaminoanthraquinone into 1-chloro-5-benzoylaminoanthraquinone by diazotization followed by the Sandmeyer reaction.

Azo derivatives of anthraquinone, which may be obtained by coupling diazotized aminoanthraquinones, are cited as colorants for a variety of applications [163]. Among the *hydrazino derivatives* only anthraquinone-1-hydrazinosulfonic acid is worth mentioning because it is obtained as an intermediate in the synthesis of pyrazoloanthrone (see Section 2.12.2).

2.6. Hydroxyanthraquinones, Alkoxy- and Aryloxyanthraquinones

Hydroxyanthraquinones are prepared primarily by synthesis of the nucleus, by exchange reactions, and by ether cleavage of alkoxy- (or aryloxy-) anthraquinones. The most important example of nucleus synthesis is the industrial production of 1,4-dihydroxyanthraquinone (quinizarin). Examples for exchange reactions are the conversion of anthraquinone-α,α'-disulfonic acids into α,α'-dihydroxyanthraquinones and the direct hydrolysis of a halogen in the 4 position in 1-amino-2,4-dihaloanthraquinone using boric–sulfuric acid. As the direct replacement of the nitro group by a hydroxy group frequently leads to side reactions, nitroanthraquinones are converted to hydroxyanthraquinones most readily via the corresponding alkyl or aryl ethers. In some cases, under rather special reaction conditions, direct replacement of one or more nitro groups is also possible. Replacement of halogen by hydroxy groups in otherwise unsubstituted haloanthraquinones is best carried out via the corresponding anthraquinone ethers. In individual cases the hydroxy groups also may be inserted at defined positions by oxidation.

Amino groups in 1,4-diamino- (or aminohydroxy-) anthraquinones are converted readily into hydroxy groups by reduction in alkaline medium via the 2,3-dihydro compounds or in an acid medium by oxidation via the quinoneimines.

Halogenation, nitration, and sulfonation of hydroxyanthraquinones present no special difficulties. Modification of the hydroxy group (boric acid esters, ethers) alters the mode of substitution. Derivatives of the hydroxy groups frequently enable a different or more selective substitution than the free hydroxy compounds.

2.6.1. General Aspects

2.6.1.1. Synthesis of the Nucleus

The preparation of quinizarin from phthalic anhydride and *p*-chlorophenol in boric–sulfuric acid has considerable industrial importance. The boric acid removes the quinizarin formed by formation of a stable boric acid ester which changes the equilibrium and, thus, enables the complete replacement of chlorine atoms by the hydroxy group.

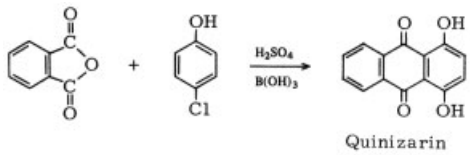

Other methods, such as condensation of maleic anhydride with naphthohydroquinone [164] and the reaction of phthalic anhydride

with hydroquinone in molten aluminum chloride/sodium chloride, have not yet reached any level of importance.

2.6.1.2. Replacement Reactions

H → OH. Hydroxy groups may be introduced into anthraquinones in acid or alkaline media. Suitable oxidizing agents are manganese dioxide, potassium peroxodisulfate, potassium chlorate, or sodium nitrite. Whereas the synthesis of 1,2-dihydroxyanthraquinone (alizarine) by oxidation of anthraquinone requires an alkaline medium, preparation of 1,2,4-trihydroxyanthraquinone (purpurin) from 1,2- or 1,4-dihydroxyanthraquinone is carried out in sulfuric acid solution by oxidation with manganese dioxide. If 1,8-dihydroxyanthraquinone is oxidized with oleum containing a high percentage of SO_3 in the presence of boric acid, oxidation occurs in the 4 position to form 1,4,5-trihydroxyanthraquinone.

SO_3H → OR (R = H, alkyl). Hydroxy and alkoxy groups may be introduced by reactions similar to those employed for amino group displacement of sulfonic acid groups. For example, anthraquinone-α-sulfonic acids are converted by heating the anthraquinonesulfonic acid with an aqueous suspension of calcium hydroxide and magnesium chloride under pressure at 200–250 °C, to form α-hydroxyanthraquinones, e.g.,

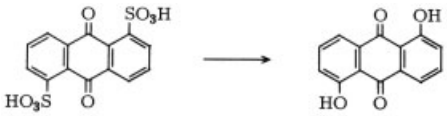

However, sulfonic acid groups in the β position react only in an alkali melt at temperatures of 150–200 °C. If oxidizing agents are added, the α position is also hydroxylated, e.g., synthesis of alizarine from anthraquinone-2-sulfonic acid:

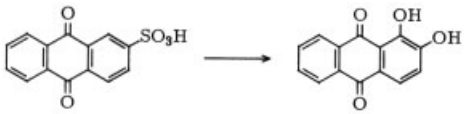

Further, anthraquinonesulfonic acids may be converted into alkoxyanthraquinones by alcoholic alkali at 50–120 °C. If higher temperatures are used, hydroxyanthraquinones are formed also. Anthraquinone aryl ethers are not prepared readily by this method. Instead, they are made from haloanthraquinones.

Halogen → OR (R = H, alkyl, aryl). Apart from the synthesis of quinizarin, direct replacement of halogen by hydroxy groups is of industrial importance only for replacing halogen in 1-amino-4-haloanthraquinones, e.g., **24** → **25** in the presence of boric–sulfuric acid. In all other cases, the haloanthraquinones are first converted with alcoholate or phenolate into the corresponding alkoxy- or phenoxyanthraquinones. This procedure is suitable for replacing halogens in both the α and the β position. However, for the preparation of β-alkoxyanthraquinones, it is recommended to generate first the β-phenoxy compound and then to subject this compound to an ether exchange reaction, e.g., **25** → **26** → **27**.

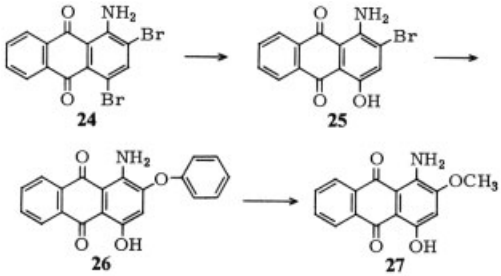

The corresponding hydroxy compounds are then accessible by acid hydrolysis of the alkoxy derivatives or by alkaline hydrolysis of the nitrated phenoxy derivatives.

NO_2 → OR (R=H, alkyl, aryl). Similar to halogen atoms, the nitro group may be replaced by hydroxy groups and ether residues. However, because direct replacement of hydroxy groups causes side reactions, only the reactions with alcoholate or phenolate, leading to the corresponding anthraquinone ethers, are of industrial importance. Alkoxyanthraquinones can then be converted to the hydroxyanthraquinones by acid cleavage.

These methods are important, especially for the preparation of 1,5- and 1,8-dialkoxy-, diphenoxy-, and dihydroxyanthraquinones, because industrial synthesis of anthraquinone intermediates is shifting more and more to the

use of nitroanthraquinones. In several cases, it is recommended that a sequence of reactions be done on the dialkoxyanthraquinones and the ether scission be performed at a later stage. In the case of 1,5-dihydroxy-4,8-dinitroanthraquinone, the replacement of one nitro group with boric–sulfuric acid presents no special difficulties, whereas the same method fails in the isomeric 1,8 compound.

NHR → OR (R=H, alkyl, aryl). The conversion of aminoanthraquinones to hydroxyanthraquinones is of little importance. Usual methods may be used to convert the amino group to the hydroxy group by means of diazotization and boiling of the reaction mixture.

The oxidative and the reductive degradation of 1,4-diamino- (or 1-amino-4-hydroxy)anthraquinones to the corresponding 1,4-dihydroxyanthraquinones is used to elucidate the structure of certain anthraquinone dyes. The reductive alkaline degradation, e.g., by treatment with warm sodium dithionite–sodium hydroxide, is used industrially for the manufacture of leuco-1,4,5,8-tetrahydroxyanthraquinone from 4,8-diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid:

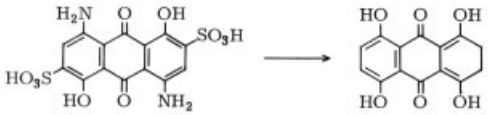

2.6.1.3. Conversions of the Hydroxy Group

OH → Alkoxy, Alkoxy → OH. The *O*-alkylation of hydroxyanthraquinone may be carried out in the presence of acid-neutralizing agents using dimethyl sulfate or esters of *p*-toluenesulfonic acid. In the reverse reaction, anthraquinone alkyl ethers may be hydrolyzed with mineral acids to give hydroxyanthraquinones [165].

2.6.1.4. Substitutions on the Hydroxyanthraquinone Nucleus

Nucleus H → Alkyl, Aryl. α-Hydroxyanthraquinones react with aldehydes in aqueous alkaline sodium dithionite solutions, thereby alkylating the β position. Reaction of benzaldehyde with quinizarin, after conversion into the leuco form, yields 2-benzyl-1,4-dihydroxyanthraquinone [166], [167]:

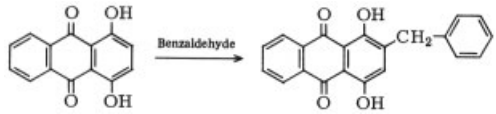

Furthermore, hydroxyanthraquinones are capable of undergoing the Mannich reaction. Quinizarin, formaldehyde, and piperidine react, for instance, under formation of 1,4-dihydroxy-2-(piperidinomethyl)anthraquinone [168]:

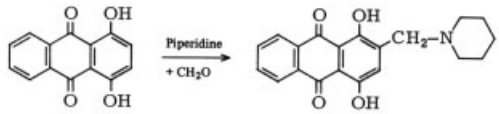

When the hydrogen atoms in the β position are activated sufficiently, direct arylation of 1,4-substituted anthraquinones is also possible. This reaction is important in preparing arylated diaminodihydroxyanthraquinones, which can be accomplished by three methods: Addition to anthraquinonediquinone, aminohydroxyanthraquinone sulfonic acids, or hydroxynitroanthraquinones.

Generally applicable is the arylation of diquinones and diquinoneimines of 1,4,5,8-aminohydroxyanthraquinones:

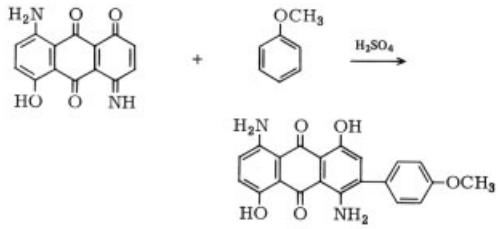

28

Special methods are available for preparing arylated 4,8-diamino-1,5-dihydroxyanthraquinones, e.g., **28**, where the starting materials 4,8-diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid (**29**) or 4,8-dinitro-1,5-dihydroxyanthraquinone (**30**) determine the position of the substituent [169]. Both reactions are carried out at low temperature in boric–sulfuric

acid. This requires that in the following step the sulfonic acid groups be cleaved [170] or the remaining nitro groups be reduced [171]:

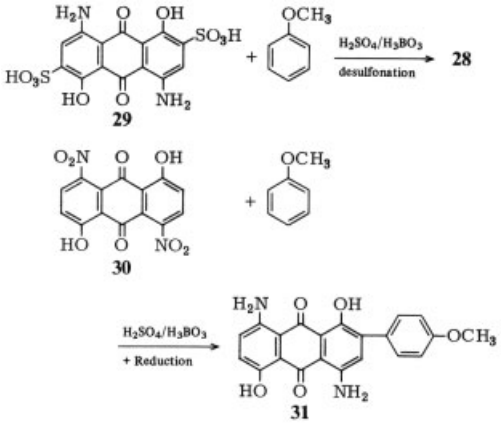

Nucleus H → SO_3H. Hydroxyanthraquinones are sulfonated in oleum. When the 1,4-dihydroxyanthraquinone is used as the starting material, addition of boric acid blocks sulfonation of the 2 position, and instead the sulfonic acid group enters the 5 or 6 position [3], [172]. In 1,5- (or 1,8-) dihydroxyanthraquinones, addition of sodium sulfate to the oleum may prevent an excessive sulfonation and the oxidation to higher hydroxylated products.

Nucleus H → Halogen. In general the choice of the halogenating agent has no influence on the substitution. Bromination is often preferred because of the greater ease in replacing bromine.

As hydroxy and amino groups generally enhance the reactivity towards halogenation, boric acid esters and boric ester amides of the 1,4-dihydroxy- (or hydroxyamino-) anthraquinones have a special function because they prevent substitution in the 2 and 3 positions. For example, quinizarin is halogenated in the 2 position in nitrobenzene and in the 5,8 positions in boric–sulfuric acid. The amino group has an additional effect on halogenation of aminohydroxyanthraquinones because halogenation is directed by either the hydroxy or the amino group, depending on the acidity of the medium. 1-Amino-4-hydroxyanthraquinone is halogenated at the 2 position in organic solvents or hydrochloric acid, in the 3 position in oleum or chlorosulfonic acid [173], and in the 5,8 positions in boric acid – oleum [174].

1-Alkylamino-5-hydroxyanthraquinones are halogenated preferentially in the 8 position in oleum [175] and only in the 4 position in hydrochloric acid [176].

A uniform monohalogenation product is not obtained from α-diaminodihydroxyanthraquinone. Choice of reaction conditions makes it possible to influence the isomer ratio (halogen next to amino or next to hydroxy groups).

1,5-Dihydroxy-4,8-dinitroanthraquinone is halogenated cleanly in the 2,6 positions in aqueous alkali [177]. In boric–sulfuric acid and in the presence of a 1,3-alkanediol, one of the nitro groups is replaced by a halogen [178].

Nucleus H → NO_2 . The rules for the course of the nitration of hydroxyanthraquinones are similar to those for the halogenation. Quinizarin in sulfuric acid or organic solvents is nitrated in the 2 position, whereas in boric acid–oleum one gets only nitration in the 5 position. A different behavior is shown by the boric or arsonic acid esters of 1,2-dihydroxyanthraquinone (alizarine). Here nitration in sulfuric acid yields the 3- and 4-nitro derivatives. 1,8-Dihydroxyanthraquinone in sulfuric acid is dinitrated up to 80 % in the 4 and 5 positions, whereas the isomeric 1,5-dihydroxy compound affords only unsatisfactory yields of 1,5-dihydroxy-4,8-dinitroanthraquinone.

The α-nitration is more successful with dimethoxyanthraquinones and yet more successful with diphenoxyanthraquinones. In the last case the benzene rings are also nitrated in their 2 and 4 positions, which then facilitates ether hydrolysis in alkaline medium:

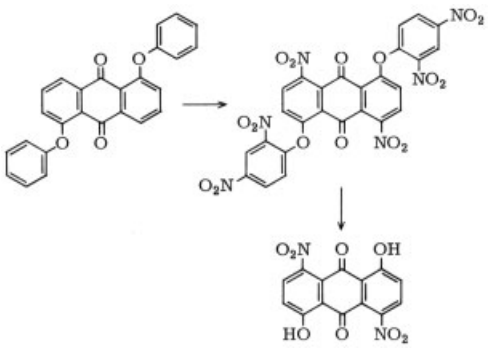

Halogen → SO_3H. The remaining chlorine atom in 1,4-diamino-3-chloro-2-phenoxyanthraquinone (obtained from 1,4-diamino-2,3-dichloroanthraquinone by phenolate melt) is replaced readily in alkali sulfite to yield the sulfonic acid.

Halogen → NHR. Halohydroxyanthraquinones can be converted into the corresponding aminohydroxyanthraquinones by the methods cited under aminoanthraquinones (see page 8).

NO_2 → NHR (R = H, alkyl, aryl). The reduction of hydroxynitroanthraquinones is possible using the usual methods. A special industrial process is a reduction with elemental sulfur in oleum.

The nitro group of 1-hydroxy-4-nitroanthraquinone is replaced easily by amino residues. In α-dihydroxydinitroanthraquinones, a stepwise substitution presents no special difficulties. In 1,8-dihydroxy-4,5-dinitroanthraquinone, the reaction with aromatic amines in the presence of boric acid takes a different course. The product is 5,8-diarylamino-1-hydroxy-4-nitroanthraquinone [179].

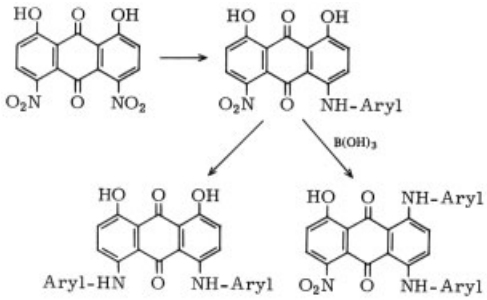

A special reaction constitutes the conversion of α-nitroanthraquinones into the corresponding 1-amino-4-hydroxyanthraquinones mediated by the action of sulfur or other reducing agents, e.g., $FeSO_4$, in oleum or sulfuric acid monohydrate, if necessary with addition of boric acid. This reaction is accompanied by a number of side reactions. Satisfactory yields and purity levels are achieved only in exceptional cases [180], [181]; see also [13, pp. 199 – 203].

2.6.2. Individual Hydroxyanthraquinones

1,2-Dihydroxyanthraquinone [*8005-35-4*], alizarine, (**32**), M_r 240.2, *mp* 289 – 290 °C, is prepared from anthraquinone by alkali melt [182] or from anthraquinone-2-sulfonic acid with 50 % aqueous sodium hydroxide and sodium nitrate at 190 – 200 °C [3, p. 39], [5, p. 44].

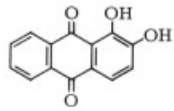

32

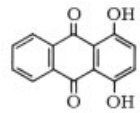

33

1,4-Dihydroxyanthraquinone [*81-64-1*], quinizarin, (**33**), M_r 240.2, *mp* 198 °C, is prepared from phthalic anhydride and *p*-chlorophenol by heating in oleum in the presence of boric acid [1], [12]. Improved methods are cited in [183], [184].

1,5-Dihydroxyanthraquinone [*117-12-4*], anthrarufin, (**34**), M_r 240.2, *mp* 280 °C, is prepared from 1,5-anthraquinonedisulfonic acid by replacing the SO_3H groups [3, p. 16], [12] and from 1,5-dinitroanthraquinone by directly replacing the nitro groups, e.g., in sulfolane with calcium oxide [185] or, more effectively, via 1,5-dimethoxyanthraquinone and subsequent hydrolysis of the ether linkage [186].

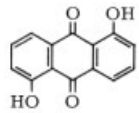

34

1,8-Dihydroxyanthraquinone [*117-10-2*], chrysazin, (**35**), M_r 240.2, *mp* 193 °C, is prepared by replacing SO_3H [5, p. 46], by directly replacing nitro groups in sulfolane in the presence of calcium oxide [185], or, better, via 1,8-dimethoxyanthraquinone and subsequent hydrolysis of the ether [186].

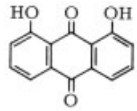

35

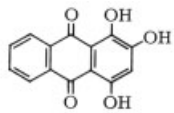

36

1,2,4-Trihydroxyanthraquinone [*81-54-9*], purpurin, (**36**), M_r 256.2, *mp* 257 – 259 °C, is prepared from alizarine [5, p. 57], [187] or from quinizarin by oxidation [188].

1,4,5-Trihydroxyanthraquinone [*2961-04-8*], M_r 256.2, *mp* 270 – 271 °C, is prepared from 1,8-dihydroxyanthraquinone with 80 % oleum and boric acid [189].

1,4,5,8-Tetrahydroxyanthraquinone [*81-60-7*], M_r 272.2, *mp* above 300 °C, is prepared from 4,8-diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid by treatment with a mixture of sodium dithionite–sodium hydroxide in aqueous medium [12].

1,5-Dimethoxyanthraquinone [*6448-90-4*], (**37**), M_r 268.3, *mp* 236 °C, is prepared from anthraquinone-1,5-disulfonic acid with methanolic alkali-metal hydroxide [190] or from 1,5-dinitroanthraquinone with methanolic alkali-metal hydroxide in the presence of amidosulfonic acid [191] or by passing air through the reaction mixture [192], [193].

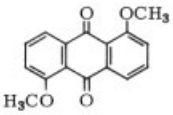

37

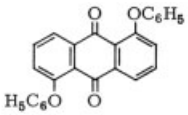

38

1,5-Diphenoxyanthraquinone [*82-21-3*], (**38**), M_r 392.4, *mp* 215 °C, is prepared from 1,5-dichloroanthraquinone and sodium phenolate without solvent [4, p. 8] or with an inert solvent [194] or from 1,5-dinitroanthraquinone by reaction with alkali phenolate in phenol [195] or an inert solvent [196], [197].

1,8-Dimethoxyanthraquinone [*6407-55-2*], M_r 268.3, *mp* 212 °C, is prepared from anthraquinone-1,8-disulfonic acid with methanolic alkali-metal hydroxide [190], from 1,8-dinitroanthraquinone with methanolic alkali-metal hydroxide in the presence of amidosulfonic acid [191], by passing air through the reaction mixture [192], [193].

1,8-Diphenoxyanthraquinone [*82-17-7*], M_r 392, *mp* 189 – 190 °C, is prepared from 1,8-dichloroanthraquinone [198] or 1,8-dinitroanthraquinone [195], [197] by reaction with alkali phenolate.

5,8-Dichloro-1,4-dihydroxyanthraquinone [*2832-30-6*], (**39**), M_r 309.1, *mp* 266 °C, is prepared from 3,6-dichlorophthalic anhydride by synthesis of the nucleus [199] or from quinizarin by chlorination [174].

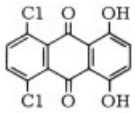

39

2-Bromo-1,4-dihydroxyanthraquinone [*81-52-7*], (**40**), M_r 319.1, sublimation at 300 °C, is prepared from quinizarin by bromination in nitrobenzene [3, p. 66].

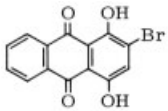

40

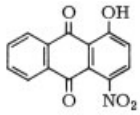

41

1-Hydroxy-4-nitroanthraquinone [*81-65-2*], (**41**), M_r 269.2, *mp* 268 °C, is prepared from 1-hydroxyanthraquinone in sulfuric acid monohydrate with nitric–sulfuric acid [200].

1,4-Dihydroxy-5-nitroanthraquinone [*27573-16-6*] M_r 285.2, *mp* 244 – 245 °C, is prepared from 1-hydroxy-10-chloroanthraquinone-(4,9) [201].

1,5-Dihydroxy-4,8-dinitroanthraquinone [*128-91-6*], M_r 330.2, is prepared from 1,5-diphenoxyanthraquinone by nitration followed by alkaline hydrolysis [4, p. 8].

1,8-Dihydroxy-4,5-dinitroanthraquinone [*81-55-0*], M_r 330.2, is prepared from 1,8-diphenoxyanthraquinone by nitration followed by alkaline hydrolysis or, alternatively, by nitration of 1,8-dihydroxyanthraquinone in oleum in the presence of boric acid and isolation after the sulfuric acid concentration has been adjusted to 80 – 100 % [202] or, using the same procedure,

from 1,8-dimethoxyanthraquinone followed by ether cleavage.

1-Amino-4-hydroxyanthraquinone [*116-85-8*], (**42**), M_r 239.2, *mp* 208 °C, is prepared from 1-amino-4-hydroxyanthraquinone-3-sulfonic acid by splitting off the SO_3H group [203], from 1-benzoylamino-4-chloroanthraquinone by treatment with boric–sulfuric acid, or from 1-nitroanthraquinone by reduction with $FeCl_2$ in sulfuric acid and rearrangement of the hydroxylamine stage [180], from 1,4-diaminoanthraquinone with MnO_2 in sulfuric acid [204], or from 1-hydroxyanthraquinone by nitration followed by reduction with Na_2S [5, p. 200].

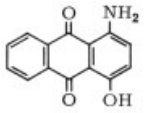

42

3-Amino-1,2-dihydroxyanthraquinone [*3963-78-8*], M_r 255.2, *mp* above 300 °C, is prepared from the 3-nitro compound by reduction with Na_2S [3, p. 3].

4-Amino-1,3-dihydroxyanthraquinone [*81-51-6*], M_r 255.2, is prepared from 1,2,4-trihydroxyanthraquinone (purpurin) and NH_3 [13, p. 163].

5-Amino-1,4-dihydroxyanthraquinone [*23215-11-4*], M_r 255.2, is prepared from the 5-nitro compound by reduction with SO_2 [205].

3-Amino-2-hydroxyanthraquinone [*117-77-1*], (**43**), M_r 239.2, is prepared from 5-benzoylbenzoxazolone-2′-carboxylic acid [2, p. 14].

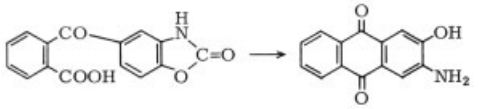

43

4,8-Diamino-1,5-dihydroxyanthraquinone [*145-49-3*], M_r 270.2, is prepared from 1,5-dihydroxy-4,8-dinitroanthraquinone by reduction with sodium sulfide; from 1,5-diphenoxyanthraquinone by nitration, cleavage, and reduction [4, p. 8]; a process improvement is described in [206]; or from 4,8-diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid by desulfonation [207].

4,5-Diamino-1,8-dihydroxyanthraquinone [*128-94-9*], M_r 270.2, is prepared from 1,8-dihydroxy-4,5-dinitroanthraquinone by reduction [208] or by cleavage of the sulfonic acid groups from the corresponding 2,7-disulfonic acid [207].

1-Hydroxy-5-isopropylaminoanthraquinone [*33902-49-7*], (**44**), M_r 281.3, is prepared from 1-isopropylaminoanthraquinone-5-sulfonic acid by replacement of SO_3H [175], [176].

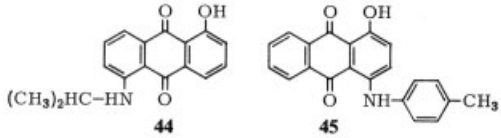

44 **45**

1-Hydroxy-4-(*p*-methylphenylamino)anthraquinone [*81-48-1*], (**45**), M_r 329.4, is prepared from the leuco form of quinizarin and *p*-toluidine [209], [3, p. 48] or from 4-halo-1-hydroxyanthraquinone and *p*-toluidine.

1-Hydroxy-5,8-bis-(*p*-methylphenylamino)anthraquinone [*4392-68-1*], M_r 434.5, is prepared from 1,4,5-trihydroxyanthraquinone and *p*-toluidine [3, p. 33] by replacing hydroxyl or from 5,8-dichloro-1-hydroxyanthraquinone [210] by chlorine replacement.

1,4-Dihydroxy-5,8-bis(phenylamino)anthraquinone, (**46**), M_r 422.4, is prepared from 5,8-dichloro-1,4-dihydroxyanthraquinone and aniline [211] or from leuco-tetrahydroxyanthraquinone and aniline in water [212] or in pure aniline [213] with the addition of boric acid.

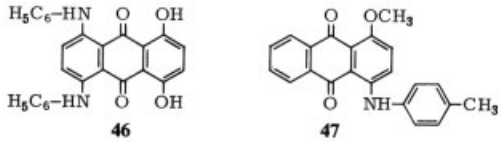

46 **47**

1-Methoxy-4-(*p*-methylphenylamino)anthraquinone, (**47**), M_r 343.4, is prepared by methylation of 1-hydroxy-4-(*p*-methylphenylamino)anthraquinone.

1,5-Dibenzoylamino-4-hydroxyanthraquinone [*6370-96-3*], M_r 462.4, is prepared by oxidation of 1,5-dibenzoylaminoanthraquinone with MnO_2 [214].

8-Benzoylamino-1,5-dihydroxy-4-(*p*-methoxybenzoylamino)anthraquinone, (**48**), M_r 508.4, is prepared by mixed acylation of 4,8-diamino-1,5-dihydroxyanthraquinone with *p*-methoxybenzoyl chloride and benzoyl chloride [4, p. 8].

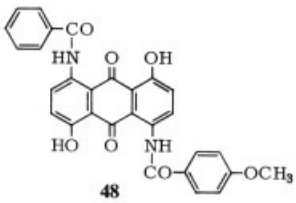

48

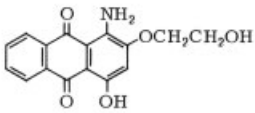

49

1-Amino-4-hydroxy-2-hydroxyethoxyanthraquinone [*17869-07-7*], (**49**), M_r 299.3, is prepared from 1-amino-4-hydroxy-2-phenoxyanthraquinone by replacement of the phenoxy group [215], from the 2-chloro compound by replacement of chlorine [216], or from the 1,4-dihydroxy derivative with ammonia [217].

1,4-Diamino-2,3-diphenoxyanthraquinone [*6408-72-6*], (**50**), M_r 422.4, is prepared from the 2,3-dichloro compound and phenol by chlorine replacement [3, p. 49].

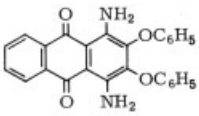

50

1-Amino-4-hydroxy-2-phenoxyanthraquinone [*17418-58-5*], (**51**), M_r 331.3, is prepared from 1-amino-2-chloro- (or -2-bromo-) 4-hydroxyanthraquinone and phenolate. This process was improved [218].

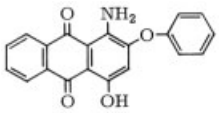

51

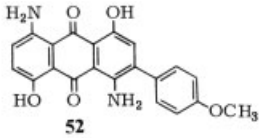

52

4,8-Diamino-1,5-dihydroxy-3-(*p*-methoxyphenyl)anthraquinone, (**52**), (for the elucidation of the structure see [169]), M_r 377.3, is prepared from 4,8-diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid by addition of anisole [219] and subsequent SO_3H removal [170] or from 4,8-diamino-1,5-dihydroxyanthraquinone via the iminoquinone and reaction with anisole [220].

4,8-Diamino-1,5-dihydroxy-2-(*p*-methoxyphenyl)anthraquinone [*4702-64-1*], (for the elucidation of the structure see [169]), M_r 377.3, is prepared from 1,5-dihydroxy-4,8-dinitroanthraquinone and anisole in boric–sulfuric acid followed by reduction [171].

1-Amino-5,8-dichloro-4-hydroxyanthraquinone [*2832-17-9*], (**53**), M_r 308.1, is prepared from 1-amino-4-hydroxyanthraquinone by chlorination in oleum–boric acid [180] or from 1-amino-4,5,8-trichloroanthraquinone with boric–sulfuric acid [221].

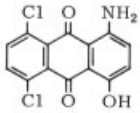

53

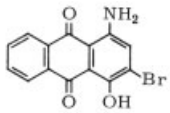

54

1-Amino-3-bromo-4-hydroxyanthraquinone, (**54**), M_r 318.1, is prepared from 1-amino-4-hydroxyanthraquinone and bromine in chlorosulfonic acid [173].

1-Amino-2-bromo-4-hydroxyanthraquinone [*116-82-5*], M_r 318.1, is prepared from 1-amino-2,4-dibromoanthraquinone with boric–sulfuric acid [3, p. 5] or in a one-pot process from 1-aminoanthraquinone (bromination – hydrolysis) [222–224].

8-Bromo-1-hydroxy-5-isopropylamino-anthraquinone, (**55**), M_r 360.2, is prepared by bromination [175].

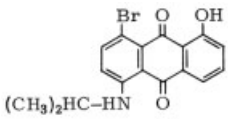

55

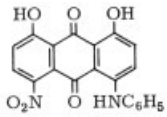

56

1,8-Dihydroxy-5-nitro-4-phenylamino-anthraquinone, (**56**), M_r 376.3, is prepared from 1,8-dihydroxy-4,5-dinitroanthraquinone and aniline [225].

1,2-Dihydroxyanthraquinone-3-sulfonic acid [*83-61-4*], (**57**), M_r 320.3, is prepared from 1,2-dihydroxyanthraquinone with fuming sulfuric acid [226].

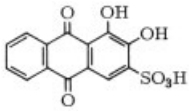

57

1,4-Dihydroxyanthraquinone-2-sulfonic acid [*145-48-2*], M_r 320.3, is prepared from 1,4-dihydroxyanthraquinone with sodium hydrogen sulfite and oxidizing agents [3, p. 65] or by sulfonation in oleum [227].

1,5-Dihydroxyanthraquinone-2,6-disulfonic acid [*6492-85-9*], M_r 400.3, is prepared from 1,5-dihydroxyanthraquinone with oleum–Na_2SO_4 [228].

1,8-Dihydroxyanthraquinone-2,7-disulfonic acid, M_r 400.3, is prepared by heating 1,8-dihydroxyanthraquinone in oleum [229].

1,5-Dihydroxy-4,8-dinitroanthraquinone-2,6-disulfonic acid [*6449-09-8*], (**58**), M_r 490.3, is prepared from 1,5-dihydroxyanthraquinone-2,6-disulfonic acid by nitration in sulfuric acid [228].

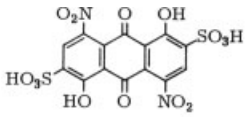

58

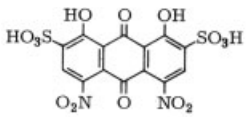

59

1,8-Dihydroxy-4,5-dinitroanthraquinone-2,7-disulfonic acid [*128-90-5*], (**59**), M_r 490.3, is prepared from 1,8-dihydroxyanthraquinone-2,7-disulfonic acid by nitration [229].

4,8-Diamino-1,5-dihydroxyanthraquinone-2-sulfonic acid [*5138-23-8*], (**60**), M_r 350.3, is prepared by removal of one SO_3H group from the 2,6-disulfonic acid [207], [230].

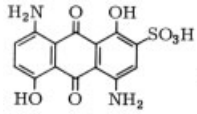

60

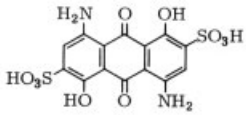

61

4,8-Diamino-1,5-dihydroxyanthraquinone-2,6-disulfonic acid [*128-86-9*], (**61**), M_r 430.3, (the sodium salt is the dye Acilan Sapphirol B), is prepared from 1,5-dihydroxyanthraquinone by sulfonation, nitration, and reduction [1, p. 607].

4,5-Diamino-1,8-dihydroxyanthraquinone- 2,7-disulfonic acid, M_r 430.3, is prepared from 1,8-dihydroxyanthraquinone by sulfonation, nitration, and reduction [229].

2.7. Mercaptoanthraquinones

Mercaptoanthraquinones are of little commercial importance. However, they appear as unisolated intermediates. They are obtained by replacement of halogen [231] or sulfonic acid groups with alkali sulfides, by addition of alkali sulfides to 1-aminoanthraquinones [232], or by cleavage of thiocyanatoanthraquinones [233].

More important are the anthraquinone thioethers, which are obtained by usual methods as replacement of halogen, nitro, or sulfonic acid groups; by modification of mercaptoanthraquinones; or by addition reactions to hydroxy- and aminoanthraquinones or to quinoneimine derivatives [234]. 1-Hydroxy-4-nitroanthraquinones react especially well under simultaneous reduction of the nitro group [235].

2.7.1. General Aspects

$SO_3H \rightarrow SR$ (R = alkyl, aryl). The exchange of sulfonic acid groups against mercapto groups is especially suitable for the preparation of alkylmercaptoanthraquinones. The preparative reactions are carried out in aqueous alkaline medium at temperatures up to 100 °C.

Halogen → SR (R = H, alkyl, aryl). Halogen displacement is used to prepare mercapto-, alkylmercapto-, and arylmercaptoanthraquinones. Suitable reaction media are organic solvents: halogenated aromatic hydrocarbons and especially amides in the presence of acid binding agents, such as alkali – metal carbonates and tertiary amines.

$NO_2 \rightarrow SR$. Displacement of nitro groups occurs under conditions similar to those for the replacement of halogens, insofar as addition reactions do not predominate [236].

H → SR. The addition of mercaptans to amino- or hydroxyanthraquinones or to 1-hydroxy-4-nitroanthraquinones [235] that occurs via reductive steps has no industrial importance.

SH → SR. Mercaptoanthraquinones may be converted to alkylmercaptoanthraquinones with the usual alkylating agents under alkaline conditions [237].

2.7.2. Individual Mercapto Compounds and Thioethers

1,4-Diamino-2-mercaptoanthraquinone, (**62**), M_r 270.3, is prepared from 1,4-diamino-2-bromoanthraquinone and sodium disulfide followed by reduction [238].

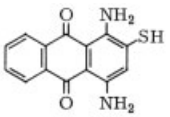

62

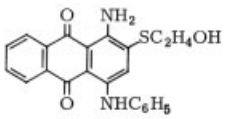

63

1-Amino-2-(2-hydroxyethylmercapto)-4-phenylaminoanthraquinone, (**63**), M_r 390.4, is prepared by displacement of SO_3H from the corresponding anthraquinone-2-sulfonic acid [239].

1,4-Dihydroxy-2-phenylmercaptoanthraquinone, (**64**), M_r 348.4, is prepared from 1,4-dihydroxyanthraquinone and thiophenol [240].

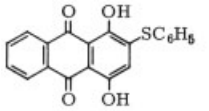

64

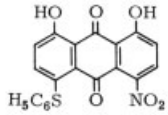

65

1,8-Dihydroxy-4-nitro-5-phenylmercaptoanthraquinone, (**65**), M_r 393.4, is prepared from 1,8-dihydroxy-4,5-dinitroanthraquinone and thiophenol [236].

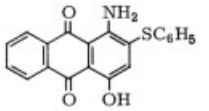

66

1-Amino-4-hydroxy-2-phenylmercaptoanthraquinone, (**66**), M_r 347.4, is prepared by displacement of bromine from 1-amino-2-bromo-4-hydroxyanthraquinone and thiophenol [241].

1-Amino-2-mercaptoanthraquinone-6-sulfonic acid, (**67**), M_r 335.4, is prepared from 1-nitroanthraquinone-6-sulfonic acid with aqueous sodium sulfide [242].

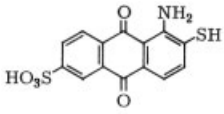

67

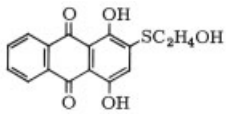

68

1,4-Dihydroxy-2-(2-hydroxyethylmercapto)anthraquinone [*3319-54-8*], (**68**), M_r 316.3, is prepared from 2-nitro-, 2-halo-, or 2-sulfo-1,4-dihydroxyanthraquinone with 2-mercaptoethanol [243].

2.8. Anthraquinonesulfones

Alkyl and aryl sulfone derivatives of anthraquinone have reached a certain amount of importance only as polyester dyes resistant to sublimation and as dyes for mixed polyester – cotton

fabrics. They are obtained by displacement of the halo [244] or nitro group with sulfinates, by sulfinic acid addition to quinoneimines [245], or by oxidation of the corresponding thioethers with chromium(VI) oxide or hydrogen peroxide in such solvents as glacial acetic acid or sulfuric acid [246].

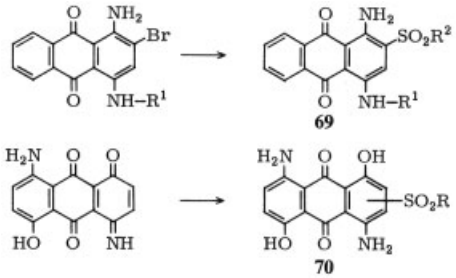

1-Amino-2-phenylsulfonyl-4-*p*-tosylaminoanthraquinone (**69** with $R^1 = -SO_2C_6H_4CH_3$, $R^2 = -C_6H_5$), M_r 456.6, is prepared from the 2-bromo compound by halogen displacement with benzene sulfinate [244].

1-Amino-2-ethylsulfonyl-4-(*p*-isopropylphenylamino)anthraquinone [*4839-56-9*], (**69** with $R^1 = -C_6H_4CH(CH_3)_2$, $R^2 = -C_2H_5$), M_r 448.5, is prepared from the 2-bromo compound by halogen displacement with ethyl sulfinate [237].

4,8-Diamino-1,5-dihydroxy-2- (and -3-) phenylsulfonylanthraquinone (**70** with $R = C_6H_5$), M_r 410.4, is prepared from 4,8-diamino-1,5-dihydroxyanthraquinone via the quinoneimine [245].

2.9. Anthraquinone Aldehydes and Their Derivatives

Only anthraquinone-2-aldehydes and their derivatives have industrial importance. Starting materials are mainly methylanthraquinones, which are converted by direct oxidation with manganese dioxide or chromic acid [248] or via the dihalomethyl compounds [249] into the aldehydes. In industrial processes, 1-amino-2-methylanthraquinone is oxidized with nitrobenzene in the presence of aniline. The azomethine formed in this step can be hydrolyzed easily to the aldehyde but usually is used directly in further reactions.

1-Aminoanthraquinone-2-aldehyde also may be prepared from 2-methyl-1-nitroanthraquinone by reaction with 65 % oleum followed by reduction of the resulting isoxazole [250].

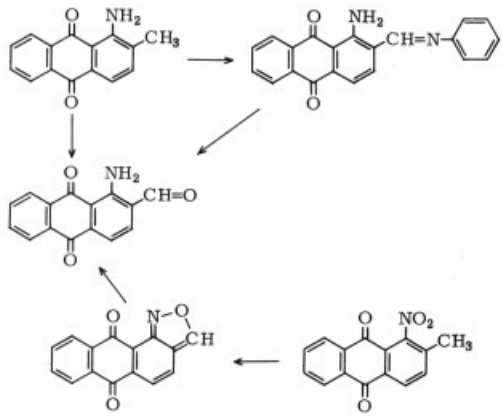

1-Aminoanthraquinone-2-aldehyde, M_r 251.2, is prepared by oxidation of 1-amino-2-methylanthraquinone with nitrobenzene in the presence of aniline and potassium carbonate, followed by hydrolysis of the 1-aminoanthraquinone-2-aldehyde anil in sulfuric acid [251], or by action of 65 % oleum on 2-methyl-1-nitroanthraquinone followed by reduction [250].

1-Amino-4-benzoylaminoanthraquinone-2-aldehyde anil, (**71**), M_r 445.5, is prepared from 1-amino-4-benzoylamino-2-methylanthraquinone by oxidation in nitrobenzene in the presence of aniline and potassium carbonate [252].

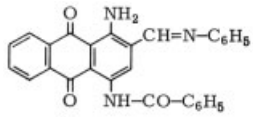

The azine of 1-amino-4-benzoylaminoanthraquinone-2-aldehyde, (**72**), M_r 736.6, is prepared from the aldehyde anil in glacial acetic acid with hydrazine hydrate [252].

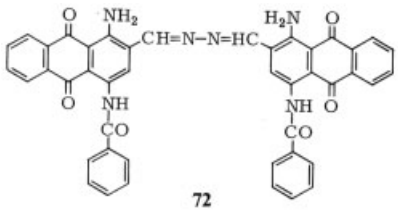

72

2.10. Anthraquinonecarboxylic Acids and Carboxylic Acid Derivatives

Whereas carboxylic acid and carbonyl chloride groups linked directly to the anthraquinone nucleus rarely appear outside the realm of dyestuff intermediates, the carboxylic esters, amides, imides, azoles, and nitriles exist in dyestuffs.

Anthraquinonecarboxylic acids may be obtained by ring closure of the corresponding *o*-benzoylbenzoic acids, by oxidation of methylanthraquinones, or by hydrolysis of nitriles. The derivatives are prepared from the carboxylic acids or nitriles by the usual methods. The nitrile group may be introduced by displacing the halogen or sulfonic acid group, or by addition.

2.10.1. General Aspects

Synthesis of Nucleus. Benzophenone dicarboxylic acids may be cyclized in weak oleum at 110–120 °C. This is the method used to prepare anthraquinone-2-carboxylic acid and 3-chloroanthraquinone-2-carboxylic acid.

$CH_3 \rightarrow COOH$. 2-Methylanthraquinone can be converted into anthraquinone-2-carboxylic acid by an oxidizing agent, such as chromic acid, in dilute sulfuric acid. Oxidation of the isoxazole obtained from 2-methyl-1-nitroanthraquinone yields the 1-nitroanthraquinone-2-carboxylic acid.

$SO_3H \rightarrow CN$. 1,4-Diaminoanthraquinone-2-sulfonic acid is converted to 1,4-diamino-2,3-dicyanoanthraquinone by reaction with alkali cyanides in aqueous alkaline medium and in the presence of an oxidant, such as manganese dioxide. From 1,4-diaminoanthraquinone-2,3-disulfonic acids are also formed the corresponding 2,3-dicyano compounds.

A related reaction is the formation of 1,4-diaminoanthraquinone-2,3-*N*-alkyliminodicarboxylic imides (**73**) from 1-amino-4-nitroanthraquinone-2-carboxylic amides on reaction with cyanides followed by reduction [253].

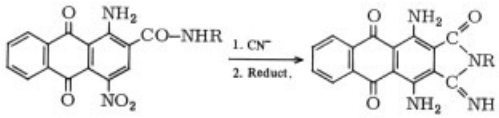

73

Halogen $\rightarrow$ CN. A general method for preparing cyanoanthraquinones is the displacement of halogen with CuCN in an organic solvent, such as benzyl cyanide, dimethylformamide, or quinoline.

$CN \rightarrow CONH_2 \rightarrow COOH$. Cyanoanthraquinones are hydrolyzed by the action of mineral acids or alkalies to yield the corresponding carboxamides or the free carboxylic acids, depending on the conditions.

Iminoimides (**74**), imides (**75**) [254], anhydrides, or free dicarboxylic acids are formed from 2,3-dicyanoanthraquinones. Reaction with alcoholic alkali forms the corresponding methoxyimino imides (**74,** R=CH_3) [255].

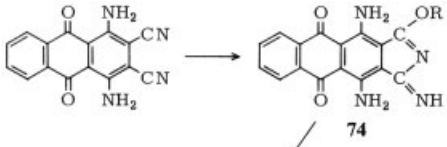

74

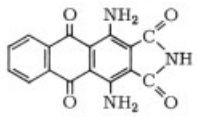

75

These hydrolysis products may be reacted further with amines.

$COOH \rightarrow COCl \rightarrow COOR$ or $CONR_2$. Anthraquinonecarboxylic acids may be converted by the usual methods (reaction with thionyl chloride or phosphorus halides) to their

acid chlorides and further to carboxylic esters and amides. Depending on their substituents, the carboxylic arylamides and hydrazides may be converted to azoles. For example, anthraquinonedicarboxylic acid dihydrazides are converted into the corresponding oxazoles.

$H \rightarrow NO_2$ or Halogen. 1-Aminoanthraquinone-2-carboxylic acid may be nitrated or halogenated in the 4 position in sulfuric acid [256]. However, under these conditions, anthraquinone-2-carboxylic acid is nitrated or dichlorinated in the 5,8 positions.

$Cl \rightarrow NHR$, $NO_2 \rightarrow SO_3H$, NHR. Carboxyl groups and their derivatives facilitate the replacement of nitro groups or halogen atoms by a variety of substituents.

$SO_3H \rightarrow Cl$. The sulfonic acid group of 2-carboxyanthraquinone-1-sulfonic acid, which is obtained from 1-nitroanthraquinone-2-carboxylic acid, is replaced easily by halogen.

2.10.2. Individual Anthraquinonecarboxylic Acids

Anthraquinone-2-carboxylic acid [*117-78-2*], M_r 252.5, is prepared from 4′-methylbenzophenone-2-carboxylic acid by oxidation with potassium permanganate and ring closure with 5 % oleum [3, pp. 8, 12].

1-Chloroanthraquinone-2-carboxylic acid [*82-23-5*], M_r 286.7, is prepared by Fischer reaction from 2-carboxyanthraquinone-1-sulfonic acid [257].

3-Chloroanthraquinone-2-carboxylic acid [*84-32-2*], M_r 286.7, is prepared from 3′-chloro-4′-methylbenzophenone-2-carboxylic acid by oxidation with potassium permanganate in alkaline solution, followed by ring closure in oleum [4, p. 19].

1-Nitroanthraquinone-2-carboxylic acid [*128-67-6*], (**76**), M_r 297.2, is prepared from 2-methyl-1-nitroanthraquinone by oxidation in sulfuric acid with sodium dichromate [2, p. 13], [5, p. 69] or in nitric acid with chromic acid [258].

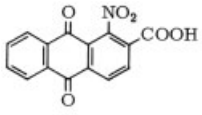

76

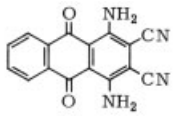

77

1,4-Diamino-2,3-dicyanoanthraquinone [*81-41-4*], (**77**), M_r 288.2, is prepared from 1,4-diaminoanthraquinone-2-sulfonic acid [259], [260] or from 1,4-diaminoanthraquinone-2,3-disulfonic acid in aqueous alkaline solution [261] or in amides [262] by reaction with sodium cyanide.

1-Amino-2-cyano-4-cyclohexylamino-anthraquinone, (**78**), M_r 345.4, is prepared by displacement of SO_3H with potassium cyanide in aqueous solution [3, p. 56].

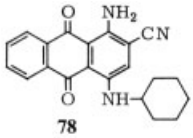

78

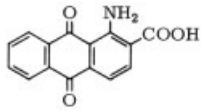

79

1-Aminoanthraquinone-2-carboxylic acid [*82-24-6*], (**79**), M_r 267.2, is prepared from 1-nitroanthraquinone-2-carboxylic acid with ammonia [2, p. 17] or from 2-methyl-1-nitroanthraquinone [263].

3-Aminoanthraquinone-2-carboxylic acid [*4831-47-4*], M_r 267.2, is prepared from 3-chloroanthraquinone-2-carboxylic acid with ammonia [4, p. 19].

1-Amino-4-cyclohexylaminoanthraquinone-2carboxamide [*3443-93-4*], (**80**), M_r 363.4, is prepared by partial hydrolysis of the nitrile [3, p. 57].

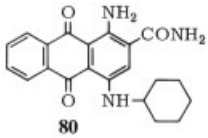

80

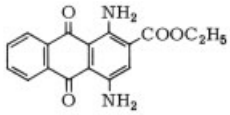

81

Ethyl-1,4-diaminoanthraquinone-2-carboxylate, (**81**), M_r 310.3, is prepared from 1-

amino-4-nitroanthraquinone-2-carbonyl chloride and ethanol followed by reduction [264].

1-Amino-4-nitroanthraquinone-2-carboxylic acid [*2058-02-8*], (**82**), M_r 312.2, is prepared by nitration of 1-aminoanthraquinone-2-carboxylic acid in sulfuric acid–formaldehyde [2, p. 17].

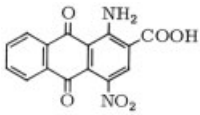

82

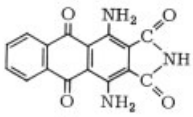

83

1,4-Diaminoanthraquinone-2,3-dicarbonyl imide [*128-81-4*], (**83**), M_r 307.3, is prepared from the corresponding dinitrile [254] or from 1,4-diaminoanthraquinone-2-carboxamide and NaCN in dimethylformamide with aeration [265].

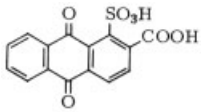

84

2-Carboxyanthraquinone-1-sulfonic acid, (**84**), M_r 332.3, is prepared from 1-nitroanthraquinone-2-carboxylic acid with Na_2SO_3 in aqueous solution [257].

2.11. Anthraquinone Derivatives with Condensed Rings

Anthraquinone derivatives with carbocyclic condensed rings, e.g., (**85**), have not yet become important as dye building blocks; however, anthraquinone structures with heterocyclic condensed rings are important components of many vat dyes and also of disperse dyes and pigments. Phthaloyl carbazoles, especially the anthrimide carbazoles, such as **86**, represent by far the most important class of compounds. Other heterocyclic systems used industrially include imidazole, oxazole, and thiazole derivatives (**87**); the phthaloyl acridones (anthraquinone acridone, **88**); and the azine derivatives of anthraquinone (indanthrone, **89**).

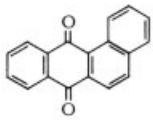

85

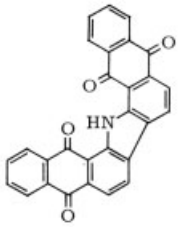

86

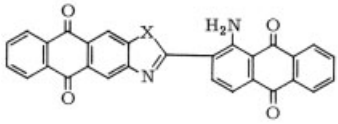

X = NH, O, S **87**

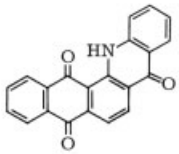

88

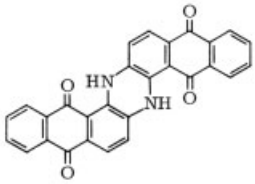

89

2.11.1. Anthraquinone Derivatives Containing Imidazole, Oxazole, and Thiazole Rings

Anthraquinone dyes with condensed imidazole, oxazole, or thiazole rings are prepared by dehydrating ring closure from 2-aroylaminoanthraquinones that carry an amino, hydroxy, or mercapto group in positions 1 or 3.

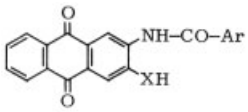

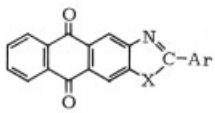

X = NH, O, S

Anthraquinone derivatives with condensed imidazole rings are the least important. The bisimidazole **90** prepared from 1,2-diaminoanthraquinone and terephthalyl chloride [266], [267] is a yellow vat dye.

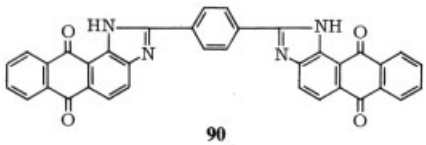

90

More important are the oxazole and thiazole derivatives. These are prepared from 2-hydroxy- (or 2-halo-) 3-aminoanthraquinones and 1-nitro- (or 1-amino-) anthraquinone-2-carboxylic acid [268] or 1-amino-4-nitroanthraquinone-2-carboxylic acid.

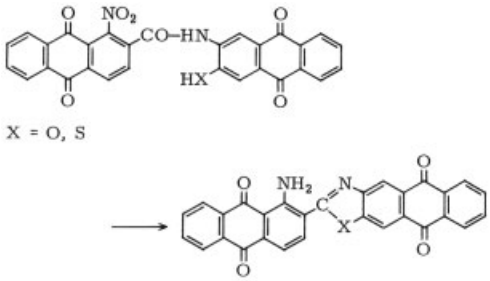

Whereas the oxazole derivatives are obtained directly from the 3-amino-2-hydroxyanthraquinone by acylation followed by ring closure with sulfuric acid [269], the preparation of the thiazoles requires acylation of 3-chloro-2-aminoanthraquinone followed by displacement of chlorine by a mercapto group using sodium sulfide, with simultaneous reduction of the 1-nitro group. Concentrated sulfuric acid is used for the ring closure in the same manner as for the oxazoles.

2-(1-Amino-2-anthraquinonyl)anthra [2,3-*d*-] oxazole-5,10-quinone [*2379-79-5*], Indanthren Red FBB, (**91**), M_r 470.4, is prepared by reacting 1-nitroanthraquinone-2-carbonyl chloride and 2-amino-3-hydroxyanthraquinone, followed by displacement of the nitro group with ammonia and ring closure [2, p. 15], [269].

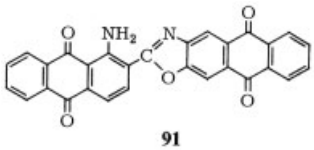

91

The following are the corresponding thiazole compounds: Indanthren Rubine B [*6371-49-9*], (**92**), [2, p. 15, 16], M_r 486.5; Indanthren Blue CLG [*6371-50-2*], (**93**), [2, p. 18], [270], M_r 605.6; and Indanthren Blue CLB [*6492-78-0*], (**94**), [2, p. 18], [271], M_r 673.6. Compound **94** is prepared from 1-amino-4-nitroanthraquinone-2-carbonyl chloride and 2-amino-3-chloroanthraquinone and Na_2S-sulfur.

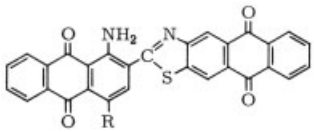

92, R = H

93, R = -NH-CO-

94, R = -NH-CO-

2,2′-Bisanthra [2,1-*d*-] thiazole-6,11-quinonyl [*6451-12-3*], Algol Yellow 8 G, (**95**), M_r 528.6, is prepared from 2-amino-1-chloroanthraquinone by replacing the chlorine atom by a mercapto group, followed by reaction with glyoxal sulfate in concentrated sulfuric acid [272], [2, p. 19].

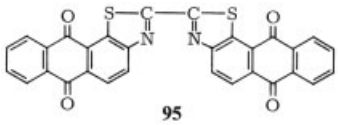

95

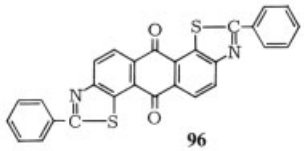

96

2,8-Diphenylthiazolo [5′,4′ : 7,8] anthra-[2,1-*d*-] thiazole-6,12-quinone [*129-09-9*], Algol Yellow GC, (**96**), M_r 474.6, is prepared from 2,6-diaminoanthraquinone and benzaldehyde [273] or benzotrichloride [2, p. 59] and sulfur in molten naphthalene.

2.11.2. Anthraquinonecarbazole Derivatives

The following description uses the designations common in industrial circles. The carbazole of 1,1′-dianthrimide is referred to as 1,1′-dianthrimide-2,2′-carbazole. Other terms for this carbazole would be: 1,2 : 7,8-diphthaloylcarbazole, 16*H*-dinaphtho [2,3-*a* : 2′,3′-*i*] carbazole-5,10,15,17-tetraone, and 16*H*-(bis{anthraceno-[1,2-*b* : 2′,1′-*d*] pyrrole}) 5,17 : 10,15-bisquinone.

1-Arylaminoanthraquinones may be condensed to coeramidonine derivatives, or they may be subjected to a dehydrogenating cyclization to give carbazole derivatives.

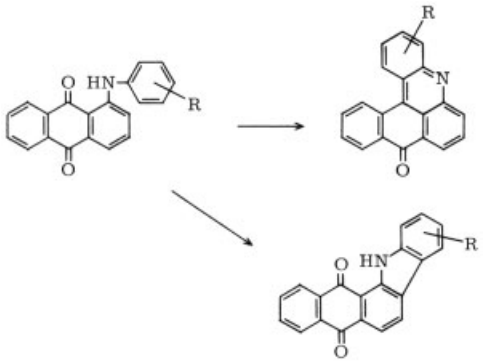

Coeramidonines are generally obtained on heating in aqueous sulfuric acid. When simple 1-arylaminoanthraquinones are treated with aluminum chloride, frequently both types of products are obtained [274]. Naphthylaminoanthraquinones and anthrimides tend to form carbazole derivatives. The reaction proceeds via dihydrocarbazoles, which are converted to the corresponding carbazoles by subsequent oxidation with air, sodium dichromate, manganese dioxide, sodium chlorate, sodium nitrite, or sodium hypochlorite. The dehydrogenating cyclization is generally termed *carbazolyation*.

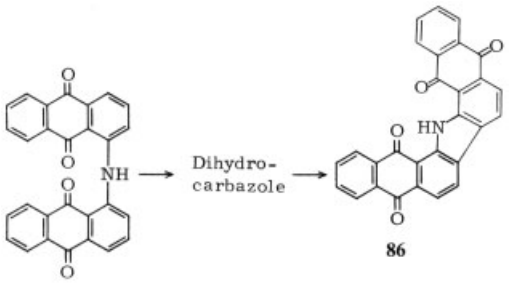

86

Among the carbazole derivatives that are not derived from anthrimides, Alizarin Light Brown BL (**97**) is an example. It is prepared from 1-benzoylamino-2-methyl-4-β-naphthylaminoanthraquinone by ring closure in concentrated sulfuric acid with simultaneous sulfonation [275].

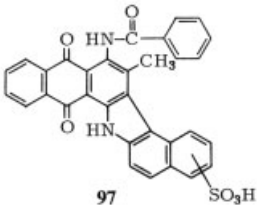

97

The vast majority of industrial carbazoles are derivatives of α,α'-anthrimides. They include some of the most important anthraquinone dyes with condensed heterocyclic rings. They are used mainly as vat dyes. They have good to excellent coloristic properties in this application and are used for yellow, orange, brown, olive, and gray colors. The particular shade depends on the number of linked anthraquinone moieties, on whether the linkage is linear or branched, and on the nature of the substituents. They are prepared industrially by carbazolyation of di-, tri-, tetra-, and pentanthrimides. Even if carbazolyation products of the higher anthrimides are generally cited as pure single compounds, carbazolyation reaction in many polyanthrimides by the usual methods must be considered incomplete.

In its broadest sense, the term anthrimide includes secondary amines, where anthraquinone has been replaced by anthraquinone derivatives such as benzanthrone, anthrapyrimidine, pyrazolanthrone, and phthaloylacridone.

Carbazolyation of anthrimide can be carried out with strong protic acids, metal halides, or alkali–metal hydroxides. The acid most commonly used is concentrated sulfuric acid, but chlorosulfonic acid or weak oleum are used also [276]. Anhydrous hydrogen fluoride in the presence of manganese dioxide also has been proposed [277]. Of the recommended metal halides ($AlCl_3$, $FeCl_3$ [278], $ZnCl_2$, $TiCl_4$, $SnCl_4$, etc.) the most frequently used is aluminum chloride. Carbazolyation in pure molten aluminum chloride has been used successfully but leads to rather contaminated products because of the temperatures required (200 – 280 °C) [279] and is not practical because of the difficulties in mixing the components intimately. The physical and chemical properties of the melts have been modified by a variety of additives. In the so-called acid melts, alkali–metal halides (NaCl, NaF, KCl), sulfur dioxide [281], sodium hydrogen sulfite,

organic solvents [282], nitrobenzene (danger of explosion!) [283], nitriles [285], or acid chlorides [286] have been proposed. The so-called alkaline aluminum chloride melts are obtained by addition of excess pyridine, picolines, quinolines, isoquinolines, or dimethylaniline [287]. Aluminum chloride–pyridine complexes in a molar ratio of $\geqq 1:1$ [288], possibly in the presence of organic solvents, and with aluminum chloride–ammonia complexes, are equivalent in their mode of action to the acid melts [289]. Carbazolyation with titanium tetrachloride in organic solvents, such as halobenzenes [290] or, sulfolane [291], have been described. Finally, alkaline carbazolyation in potassium hydroxide melts, possibly with addition of sodium hydroxide, has been reported also [292].

Each of the carbazolyation agents possesses specific properties and modes of action. The choice of an optimal carbazolyation process depends largely on the structure of the anthrimide and the nature of the substituents.

The following carbazolyation agents are used commercially:

1) *Sulfuric acid:* benzoylamino groups in the α-position and an acridone substructure facilitate carbazolyation in concentrated sulfuric acid. For example, the 4,4′-, 4,5′-, and 5,5′-dibenzoylamino-1,1′-dianthrimides; 4,5-dibenzoylamino-8-methoxy-1,1′-dianthrimide; and the anthrimide prepared from 1-amino-5-benzoylaminoanthraquinone and the 2,5,7-trichloro-3,4-phthaloylacridone are carbazolated in this way. The reactions are carried out by adding the anthrimide to sulfuric acid and stirring at 20–40 °C until reaction is complete, which is generally indicated by a change in the color of the solution. The melt is then poured into water, and the dihydrocarbazole oxidized. Posttreatment with sodium dichromate, sodium nitrite, sodium persulfate, sodium chlorate, or sodium hypochlorite facilitates a certain degree of purification by oxidizing impurities.
2) *Aluminum chloride–alkali metal halide:* mixtures of aluminum chloride and dry sodium chloride (weight ratio $\approx 4:1$) form relatively low melting eutectics. Such melts are used mainly for the carbazolyation of 1,1′-dianthrimide, 1,1′ : 5,1″-trianthrimide, and 1,1′ : 4,1″ : 5,1‴ : 8,1⁗-pentanthrimide at 140–220 °C. Stirring the melt into a solution of sodium hydroxide or hydrochloric acid, followed by oxidation with sodium dichromate, for example, yields the dyes.
3) *Aluminum chloride–sulfur dioxide:* melts of aluminum chloride and sulfur dioxide, to which sodium chloride can be added if necessary, have low viscosity and permit carbazolyation under mild conditions (20–100 °C). This method is superior to other carbazolyation processes in some cases, for instance, in treating 4′,4″-dibenzoylamino-1,1′ : 5,1″-trianthrimide. The reagent is prepared by passing sulfur dioxide into a melt of aluminum chloride and sodium chloride (weight ratio $\approx 4:1$) until the melt has low viscosity. After the anthrimide is added, the mixture is heated to the temperature required for carbazolyation. Workup involves stirring into sodium hydroxide solution.
4) *Aluminum chloride–pyridine (or methylpyridines):* carbazolyation in aluminum chloride–pyridine melts (excess pyridine) is applicable to many substituted and unsubstituted anthrimides. Industrially, this method is applied mainly to 1,1′ : 4,1″-trianthrimide, 5-benzoylamino-1,1′ : 4,1″-trianthrimide, 5-(benzanthron-3-ylamino)–1,1′-dianthrimide, and 1-benzoylamino-4,1′ : 5,1″ : 8,1‴-tetraanthrimide. In this process, aluminum chloride is added to pyridine (weight ratio $\approx 1:5$) first, and then the anthrimide is added. After heating to the temperature required for the carbazolyation, 100–170 °C, and stirring until reaction is completed, workup is carried out by adding to sodium hydroxide solution, and, if necessary, oxidative cleanup. The type of organic base (pyridine, methylpyridine, quinoline, etc.) selected and the reaction conditions exert a decisive influence on the quality of the dye.

Some important members of the anthraquinonecarbazole series are listed below.

1,1′-Dianthrimide-2,2′-carbazole [*4229-15-6*], Indanthren Yellow FFRK, M_r 427.4 (for formula see above), is prepared from

1,1′dianthrimide in $AlCl_3$-NaCl melts [4, p. 53] or in $AlCl_3$ – pyridine melts [287].

1,1′ : 5,1″-Trianthrimide-2,2′ : 6,2″-carbazole [*2172-33-0*], Indanthren Yellow 3 R, M_r 646.6, is prepared from 1,1′ : 5,1″-trianthrimide in $AlCl_3$-NaCl melts [2, p. 121].

1,1′ : 4,1″-Trianthrimide-2,2′ : 3,2″-carbazole [*2475-33-4*], (Indanthren Brown BR, *M* 646.6, is prepared from 1,1′ : 4,1″-trianthrimide in $AlCl_3$ – pyridine [5, p. 101].

1,1′ : 4,1″ : 5,1‴ : 8,1⁗-Pentanthrimide-2,2′ : 3,2′ : 6,2‴ : 7,2⁗-carbazole [*14999-97-4*], Indanthren Khaki GG, M_r 1085.0 (for formula see Section 3.1.3, **210**), is prepared from the corresponding pentanthrimide with aluminum chloride [278] or with aluminum chloride–sodium chloride [5, p. 133], [293].

5,5′-Bisbenzoylamino-1,1′-dianthrimide-2,2′carbazole [*128-70-1*], Indanthren Golden-Orange 3 G, M_r 665.6, is prepared from 5,5′-bisbenzoylamino-1,1′-dianthrimide by ring closure in concentrated sulfuric acid [4, p. 15].

4,5′-Bisbenzoylamino-1,1′-dianthrimide-2,2′carbazole [*131-92-0*], Indanthren Brown R, M_r 665.6, is prepared from 4,5′-bisbenzoylamino-1,1′-dianthrimide by ring closure in concentrated sulfuric acid [4, p. 10].

4,4′-Bisbenzoylamino-1,1′-dianthrimide-2,2′-carbazole [*2379-81-9*], Indanthren Olive R, M_r 665.6, is prepared from 4,4′-bisbenzoylamino-1,1′-dianthrimide by ring closure in concentrated sulfuric acid [4, p. 29].

4,4″-Bisbenzoylamino-1,1′ : 5,1 ″-trianthrimide-2,2′ : 6,2 ″-carbazole [*6487-07-6*], Indanthren Red Brown GR, (**98**), M_r 884.8, is prepared from the corresponding trianthrimide in aluminum chloride–sulfur dioxide [4, p. 42], [281].

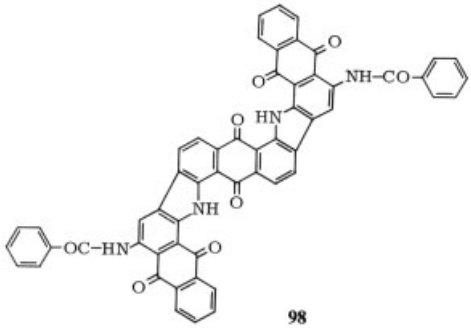

98

2.11.3. Anthraquinone Derivatives with Acridone Substructure (Phthaloylacridones)

The red to blue anthraquinone dyes with acridone substructure are all derived from 3,4-phthaloylacridone (**99**). Their use has been mainly as vat dyes.

The basic structure provides red-violet dyes, but with rather weak tinctorial strength. The 3,4-phthaloylacridones are prepared from the 1-arylaminoanthraquinone-*o*-carboxylic acids **100** or **101** by acid-catalyzed ring closure using sulfuric acid, weak oleum, chloro- or fluorosulfonic acid, polyphosphoric acid, or $AlCl_3$ or to a lesser extent phosgene, acetyl chloride, or phosphoryl chloride.

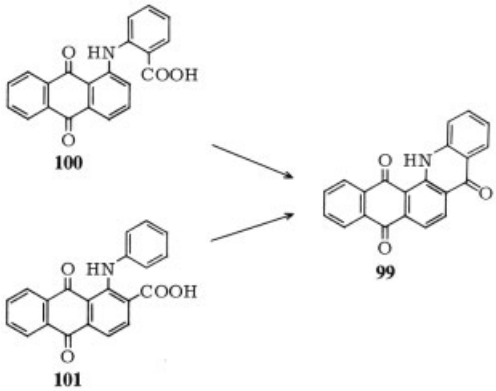

Special methods for ring closure are the chlorination of 1-anilino-2-methylanthraquinone, which yields 1,2,5,6,7,8-hexachloro-3,4-phthaloylacridone [294], [295], and the acridone ring closure that occurs on vatting of 1-anilinoanthraquinone-2′-carboxylate esters [296].

The phthaloylacridones are chlorinated easily. Depending on the chlorination conditions, the unsubstituted 3,4-phthaloylacridone may be chlorinated in the 2,5,7 positions when Cl_2–nitrobenzene–I_2 is used or in the 5,7,8 positions when Cl_2–chlorosulfonic acid is used [297]. The chlorine in position 2 of 2,5,7-trichloro-3,4-phthaloylacridone is replaced easily by amines. Therefore, 2,5,7-trichloro-3,4-phthaloylacridone reacts with aminoanthraquinones to form anthrimides [298].

Bromamine acid and (substituted) anthranilic acid are used in the preparation of (substituted) 2-amino-3,4-phthaloylacridone (**102**). Ring closure in acid medium (oleum or chlorosulfonic acid) and cleavage of the sulfonic acid group yield 2-amino-3,4-phthaloylacridone.

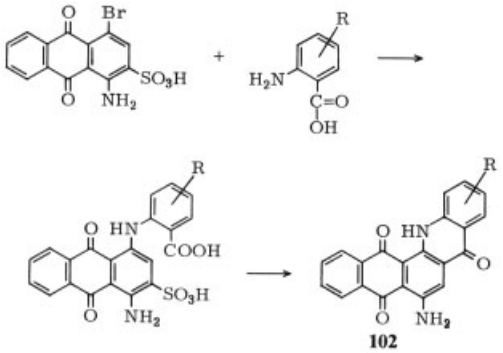

3,4-Phthaloylacridone, (**103**), M_r 325.3, is prepared from 1-anilinoanthraquinone-2-carboxylic acid [299], [300] or from 1-(2′-carboxyphenylamino)anthraquinone [301].

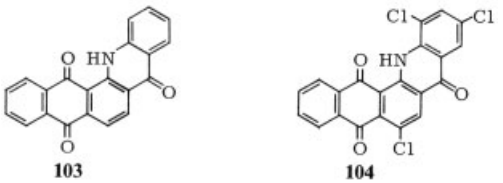

2,5,7-Trichloro-3,4-phthaloylacridone [*6373–31–5*], Indanthren Red Violet RRK, (**104**), M_r 428.6, is prepared by chlorination of 3,4-phthaloylacridone in nitrobenzene with sulfuryl chloride [302].

5,7,8-Trichloro-3,4-phthaloylacridone, (**105**), M_r 428.6, is prepared by chlorination of 3,4-phthaloylacridone in chlorosulfonic acid in the presence of sulfur [297] or from the 5,7-dichloro compound [303].

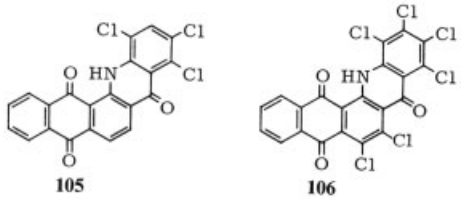

1,2,5,6,7,8-Hexachloro-3,4-phthaloylacridone [*1328-35-4*], (**106**), Indanthren Pink B, M_r 532.0, is prepared by chlorination of 1-anilino-2-methylanthraquinone in trichlorobenzene followed by hydrolysis [294].

6-Chloro-2-amino-3,4-phthaloylacridone [*6219-98-3*], (**107**), Indanthren Turquoise Blue GK, M_r 374.8, is prepared from 1-amino-4(2′-carboxy-5′-chlorophenylamino)anthraquinone-2-sulfonic acid by ring closure in weak oleum and cleavage of the sulfonic acid group [4, p. 21].

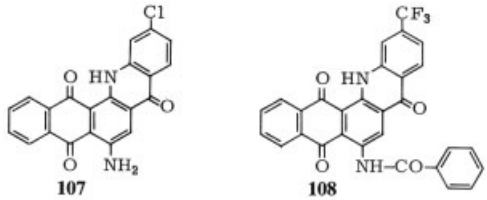

2-Benzoylamino-6-trifluoromethyl-3,4-phthaloylacridone [*6219-97-2*], Indanthren Blue HCGK, (**108**), M_r 512.4, is prepared from bromamine acid (**12**) and 4-trifluoromethylanthranilic acid followed by ring closure and desulfonation in sulfuric acid monohydrate [4, p. 34], [5, p. 143].

1,2 : 5,6-Diphthaloylacridone [*10142-57-1*], Indanthren Orange F 3 R, (**109**), M_r 455.4, is prepared from 1-nitro- (or chloro-) anthraquinone-2-carboxylic acid and 2-aminoanthraquinone in dichlorobenzene followed by ring closure in sulfuric acid [304].

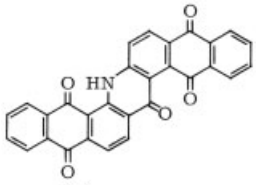

109

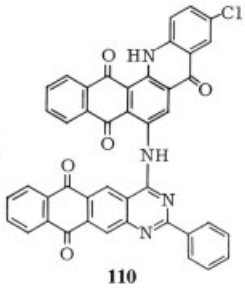

110

Indanthren Green 4 G [*6661-46-7*], (**110**), M_r 709.1, is prepared from 2-amino-7-chloro-3,4-phthaloylacridone and 2-phenyl-4-chloro-6,7-phthaloylquinazoline, which itself is derived from 3-aminoanthraquinone-2-carboxylic acid, benzoic acid, and ammonia [4, p. 19], [305].

2.11.4. Anthraquinone Derivatives with Phenazine Substructure

Among the anthraquinone derivatives with phenazine substructure, the *indanthrones* are an important class of vat dyes. The blue color of the unsubstituted unit may be shifted to green by appropriate substitutions. Several synthetic pathways are available:

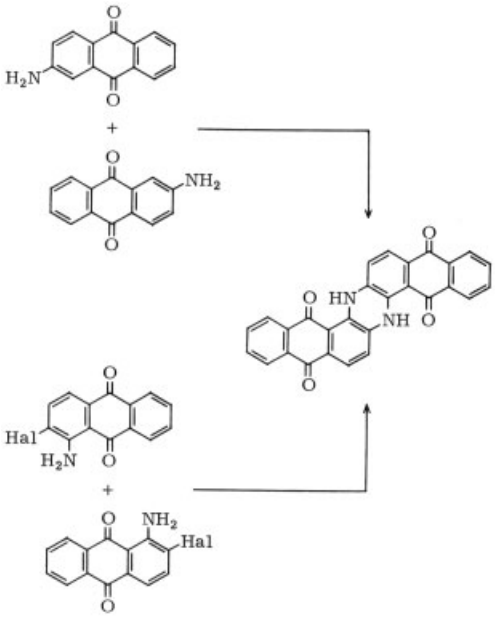

1) Aminoanthraquinones are dimerized to indanthrones in alkali melts (above 200 °C) in the presence of such oxidizing agents as sodium nitrite or air. The oldest and most important method for preparing unsubstituted indanthrone is the potash fusion of 2-aminoanthraquinone. A related process is the oxidation of the leuco-sulfuric ester of 2-aminoanthrahydroquinones in sodium hydroxide with potassium permanganate [306]. However, this preparation method is useful only for the preparation of leuco-sulfuric esters of indanthrones.
   Better suited for the oxidative dimerization of 1-aminoanthraquinone to indanthrone is a melt of potassium phenolate at 200 – 210 °C in the presence of oxidizing agents [308] or passage of air in the presence of catalysts, such as MnO_2 or $FeCl_3$ [309]. Dimerization of 1- (or 2-) aminoanthraquinone to indanthrone can be achieved with good yields at 115 – 125 °C in mixtures of dimethyl sulfoxide [310] or tetramethylurea [311] and 50 % aqueous KOH.
2) 2-Amino-1-halo- (or 1-amino-2-halo-) anthraquinones condense to indanthrones when heated in high-boiling solvents, such as nitrobenzene or naphthalene, in the presence of copper or copper salts and acid-binding agents (Kugel's indanthrone synthesis) [312]. This method, however, is important only in the preparation of substituted indanthrones.

Indanthrone [*81-77-6*], Indanthren Blue RS and R, (**111**), M_r 442.4, is prepared from 2- (or 1-) aminoanthraquinone by potash fusion [1, p. 712], in potassium phenolate [307], [309], or in 50 % potassium hydroxide [211], [310].

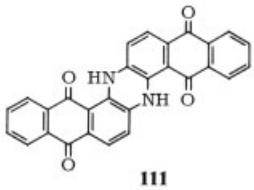

111

Chlorination of indanthrone in sulfuric acid under addition of 1 % or 5 % manganese dioxide (5 – 8 % or 13.8 – 14 % Cl) at 50 – 55 °C affords

dyes of slightly greener shades (Indanthren Blue GCD or BC).

3,3′-Dichloro-4,4′-diaminoindanthrone [*1328-41-2*], Indanthren Green BB, (**112**), M_r 541.4, is prepared from 1,4-diamino-2,3-dichloroanthraquinone in naphthalene in the presence of copper powder [313].

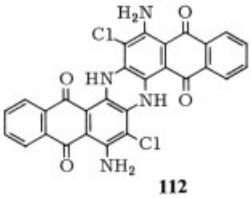

112

Indanthrone-3,3′-dicarboxylic acid, Cibanone Blue 2 G, (**113**), M_r 530.4, is prepared from2-aminoanthraquinone-3-carboxylic acid by potash fusion [314].

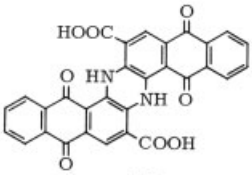

113

Leuco-sulfuric ester of 3,3′-dichloroindanthrone [*2519-28-0*], Anthrasol Blue IBC, (**114**), M_r 839.7, is prepared from the leuco-sulfuric ester of 2-acetylamino-3-chloroanthraquinone [315].

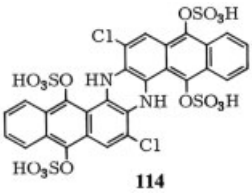

114

2.12. Anthraquinone Derivatives with Meso Rings (1,9-Cyclo-10-anthrones)

This class includes anthraquinone derivatives that contain a condensed carbon or heterocyclic ring linking the α position and the neighboring meso position bridging the 1 and 9 positions. (These derivatives are named in the following discussion by their "traditional" nomenclature, which is still the one in common use in industrial dye circles. For example, the compound called benzanthrone in the "traditional" nomenclature would be named as 7-oxo-7*H*-benz [*d,e*]anthracene by the new nomenclature rules).

Such ring systems have been prepared in nearly every feasible combination, but only the few listed below have attained importance in the commercial synthesis of dyes.

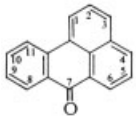

Benzanthrone

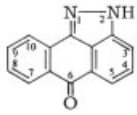

Pyrazolanthrone

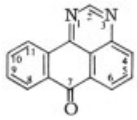

Anthrapyrimidine

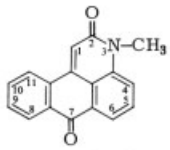

N-Methylanthrapyridone

Without additional substituents, these compounds possess a relatively faint yellow color, and have a yellowish green fluorescence in organic solvents or sulfuric acid. This group also includes a number of polycyclic compounds with additional side and/or meso-ring systems, such as acedianthrone, anthanthrone (**145**, for formula see Section 2.12.6), pyranthrone (**149**, for formula see Section 2.12.7), violanthrone (**151**, for formula see Section 2.12.8), and dibenzpyrenequinone.

2.12.1. Benzanthrone and Its Derivatives

Benzanthrone is an important intermediate for the synthesis of vat dyes. It is prepared from anthrone (**122**), glycerol, and sulfuric acid under conditions similar to those employed in Skraup's quinoline synthesis. Industrial synthesis commonly starts directly from anthraquinone, which is first reduced to anthrone by an appropriate reducing agent (e.g., iron) without isolation of the intermediate.

Dyes of the basic benzanthrone system are of little importance. Commercial dyes are almost always benzanthrones with additional fused ring systems derived from benzanthrone and its simple substituted derivatives.

Benzanthrone is substituted easily. Electrophilic reagents usually attack the 3 position first and then the 9 position, whereas nucleophilic substitution attacks the 4 position or the 6 position. Therefore halogenation of benzanthrones first yields the 3-halo derivatives; more vigorous reaction conditions yield the 3,9-disubstituted compounds. Direct synthesis is used to generate derivatives bearing halogen in other positions and to obtain 4-methylbenzanthrone.

For halobenzanthrones, displacement reactions are carried out most easily with the halogen in positions 3 or 9. Condensation with aminoanthraquinone in inert solvents in the presence of acid-binding agents and a copper catalyst will yield condensation products of the anthrimide type. Valuable vat dyes, e.g., **116,** are derived by periring-closure of 3-(1-anthraquinonylamino)-benzanthrone (**115**) using a potash melt [316] or pyridine – aluminum chloride.

The bromine atoms of 3,9-dibromobenzanthrone can be replaced stepwise.

Pyrazolanthrones can be reacted with halobenzanthrones in a manner similar to 1-aminoanthraquinone [317].

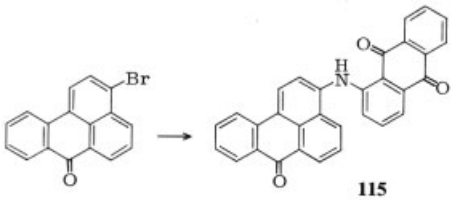

115

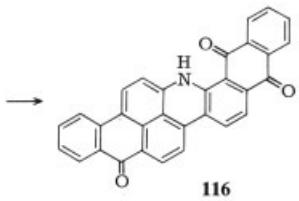

116

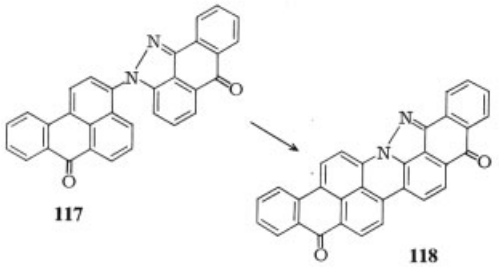

117 **118**

3-Bromobenzanthrone reacts with sodium sulfide to yield the mercaptan, which may be alkylated with chloroacetic acid to give the thioglycolic acid derivative. This product may then be converted to an indigoid dye by oxidation and potash fusion [318]:

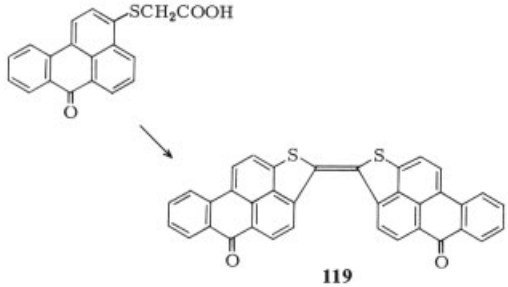

119

Treatment of 3-bromobenzanthrone with copper cyanide in an organic solvent yields 3-cyanobenzanthrone, which may be hydrolyzed to benzanthrone-3-carboxylic acid. This compound may then be converted further via the acid chloride and condensation with 1-aminoanthraquinone to the carboxamide (**120**), which in an alkali melt forms a ring by decarboxylation to give "Isoimide Green" (**121**) [319]:

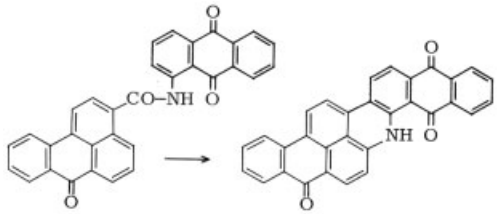

120 **121**

Mild treatment of benzanthrone with alcoholic alkalies leads to dimerization, giving 4,4′-bibenzanthronyl (**129**), an important intermediate for the preparation of violanthrone dyes (see Section 2.12.8).

Benzanthrone [*82-05-3*], (**123**), M_r 230.3, *mp* 170 °C, is prepared commercially by reacting anthraquinone with glycerol, sulfuric acid, and a reducing agent, usually iron. This involves reduction of anthraquinone to anthrone (**122**), which condenses immediately with the acrolein formed by dehydration of glycerol, and is finally oxidized by the sulfuric acid [1], [2, p. 65], [320]. An improved process is described in [321], [322]. More recently, the use of phos-

phorus or a phosphorus compound, e.g., sodium hypophosphite, has been proposed as the reducing agent [323]. Electrochemical procedures have been also suggested [324], [325]. It is possible to start with anthrone prepared in a separate step, e.g., by hydrogenation [326]. Benzanthrone is best purified by sublimation.

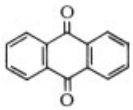

122

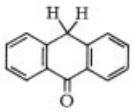

123

4-Methylbenzanthrone [*6409-46-7*], (**124**), M_r 244.3, is prepared from 2-methylanthraquinone via methylanthrone with glycerol in sulfuric acid [327].

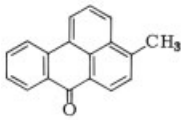

124

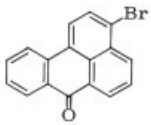

125

3-Bromobenzanthrone [*81-96-9*], (**125**), M_r 309.2, is prepared by bromination in hydrochloric acid [2, p. 71]; process improvement is described in [328].

3,9-Dibromobenzanthrone [*81-98-1*], (**126**), M_r 388.1, is prepared by bromination in chlorosulfonic acid [2, p. 76], [12], [329].

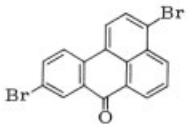

126

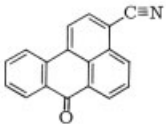

127

3-Cyanobenzanthrone, (**127**), M_r 255.3, is prepared from 3-bromobenzanthrone with CuCN in pyridine [330].

Benzanthrone-3-carboxylic acid, (**128**), M_r 274.3, is prepared from the nitrile by hydrolysis [331].

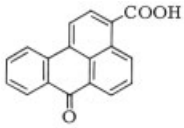

128

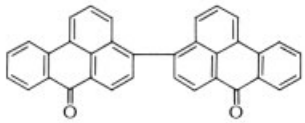

129

4,4′-Bibenzanthronyl [*116-90-5*], (**129**), M_r 458.5, is prepared by mild alcoholic alkali melt of benzanthrone (below 200 °C in the presence of excess alcohol) [1]. It is an intermediate in the synthesis of violanthrone dyes (see Section 2.12.8).

3-(1-Anthraquinonylamino)benzanthrone [*81-94-7*], (**115**), M_r 451.5, is prepared from 1-aminoanthraquinone and 3-halobenzanthrone [2, p. 71], [332].

Indanthren Olive Green B [*3271-76-9*], Imide Green, (**116**), M_r 449.5, is prepared from 3-(anthraquinonyl-(1)-amino)benzanthrone by alkali melt [2, p. 71], [316].

***N*-[Benzanthronyl-(3)]pyrazolanthrone** (**117**), M_r 448.5, is prepared from pyrazolanthrone and 3-bromobenzanthrone in nitrobenzene [317].

Indanthren Navy Blue R [*6247-39-8*], (**118**), M_r 466.5, is prepared from **117** by mild alkali fusion [317].

Additional compounds are given in the section on vat dyes, Section 3.1.3.

2.12.2. Pyrazolanthrone and Its Derivatives

1,9-Pyrazolanthrones are generated easily from the 1-hydrazinoanthraquinones [333] by ring closure with dehydration. This may be accomplished in either concentrated sulfuric acid at temperatures up to 100 °C or by heating in aqueous hydrochloric acid.

Hydrazinoanthraquinones are obtained from 1-aminoanthraquinones by diazotization, reduction with sodium hydrogen sulfite to the hydrazinosulfonic acid, and hydrochloric acid hydrolysis. Other synthesis options are displacement of halogen or sulfonic acid groups by hydrazine.

Only the unsubstituted pyrazolanthrone (**130**) is of commercial importance for the production of benzanthrone dyes obtained by reaction with 3-bromobenzanthrone and subsequent ring closure. It is also the starting material for the production of 2,2′-bipyrazolanthronyl (**131**). Alkylation of this substance then yields the corresponding diethyl derivative (Indanthren Rubin R, **132**).

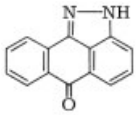

130

1,9-Pyrazolanthrone [*129-56-6*], (**130**), M_r 220.2, *mp* 285 – 286 °C, is prepared by diazotization of 1-aminoanthraquinone, reaction of the diazonium salt with sodium hydrogen sulfite, and cyclization of the resulting hydrazinosulfonic acid [5, p. 160].

2,2′-Bipyrazolanthronyl [*129-54-4*], (**131**), M_r 438.5, is prepared from pyrazolanthrone in an alkali melt [13, p. 345], [334].

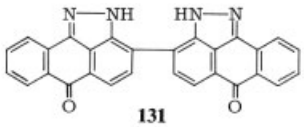

131

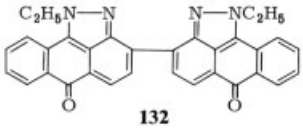

132

***N,N′*,-Diethyl-2,2′-bipyrazolanthronyl** [*4203-77-4*], (**132**), Indanthren Rubin R, M_r 494.6, is prepared from 2,2′-bipyrazolanthronyl by alkylation with diethyl sulfate after treatment with potassium hydroxide solution [5, p. 159]. For structure elucidation see [11, C.I. 70320].

2.12.3. 1,9-Anthrapyrimidine and Its Derivatives (1,9(*N*-)-Pyrimidinoanthrone-10)

The 1,9-anthrapyrimidines (**133**) are obtained easily from the corresponding 1-aminoanthraquinones by reaction with formamide [337] or with aqueous solutions of formaldehyde and ammonia in the presence of nitrobenzene sulfonic acid as the oxidizing agent [335], [336].

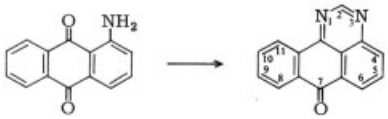

133

An equally smooth process starts by transforming 1-aminoanthraquinone into its formamidinium chloride by reaction with dimethylformamide and thionyl chloride, followed by ring closure in the presence of ammonium salts of weak acids [338]:

Only the 6- and 8-amino derivatives of 1,9-anthrapyrimidines are used as intermediates to produce greenish-yellow vat dyes. Unsubstituted 1,9-anthrapyrimidine is faint yellow, the 6-amino derivative is yellow with a green cast, and the 8-amino derivative is orange red.

6-Aminoanthrapyrimidine, (**134**), M_r 247.3, is prepared from 1,4-diaminoanthraquinone with aqueous formaldehyde and ammonia in the presence of nitrobenzenesulfonic acid [2, p. 88].

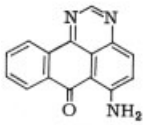

134

8-Aminoanthrapyrimidine, M_r 247.3, is prepared from 1,5-diaminoanthraquinone-2-sulfonic acid with aqueous formaldehyde and ammonia in the presence of nitrobenzenesulfonic acid, followed by splitting off the sulfonic acid group [2, p. 94]. In contrast to 1,4-diami-

noanthraquinone, reaction at only one side of the 1,5-diaminoanthraquinone is not assured.

6-(4′-Chlorobenzoylamino)anthrapyrimidine [*4216-00-6*], Indanthren Yellow 7 GK, M_r 385.8, is prepared by reacting 6-aminoanthrapyrimidine with 4-chlorobenzoyl chloride in *o*-dichlorobenzene [337] or from *N,N*-dimethyl-*N′*-[4-(4′-chlorobenzoylamino)anthraquinonyl-(1)] formamidinium chloride and ammonium acetate in *N*-methylpyrrolidone [338].

1,9-Anthrapyrimidine-2-carboxylic acid, (**135**), M_r 276.2, is prepared from 1-amino-2-methylanthraquinone with formamide or with formaldehyde–ammonia (addition of an oxidizing agent), followed by oxidation of the 2-methyl derivative to the 2-carboxylic acid [339].

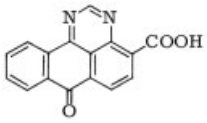

135

Conversion into the acid chloride and condensation with aminoanthraquinone affords fast yellow vat dyes and pigments [340], [341].

2.12.4. 1,9-Anthrapyridone and Its Derivatives

1-Acetylaminoanthraquinones are converted to anthrapyridones by alkaline reagents [342] and by direct heating [343], which is accompanied by loss of water:

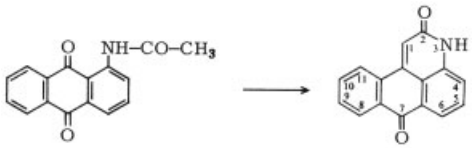

Acetylation and subsequent ring closure may also be combined by heating the aminoanthraquinone with acetic anhydride in the presence of alkali-metal acetates [344].

Only the 6-arylamino derivatives of 3-methylanthrapyridone and 1-ethoxycarbonyl-4methylanthrapyridone are of commercial interest. Starting from 4-bromo-1-methylaminoanthraquinone, the 6-bromo-3-methylanthrapyridone is obtained by acetylation and alkali-catalyzed ring closure [5, p. 217]. In this compound, the halogen may be easily displaced by amine functions [344]. Subsequent sulfonation yields lightfast red acid dyes (Alizarine Rubinoles).

6-Bromo-3-methylanthrapyridone, (**136**), M_r 340.2. For preparation see preceding paragraph [5, p. 217].

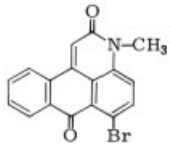

136

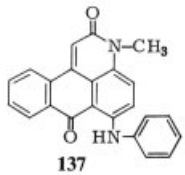

137

6-Anilino-3-methylanthrapyridone, M_r 352.4, is prepared from 6-bromo-3-methylanthrapyridone and aniline [3, p. 47].

6-Anilino-3-methylanthrapyridone-4′-sulfonic acid, (**138**), M_r 432.5, is prepared by condensation of 6-bromo-3-methylanthrapyridone with aniline, followed by monosulfonation [345].

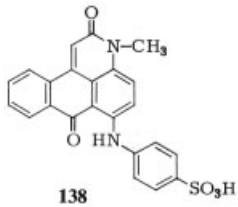

138

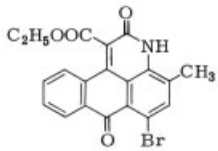

139

1-Ethoxycarbonyl-6-bromo-4-methylanthrapyridone, (**139**), M_r 412.3, is prepared from 1-amino-4-bromo-2-methylanthraquinone and diethyl malonate [346].

2.12.5. Dibenzpyrenequinones

According to their chemical structure, dibenzpyrenequinones should be classified as benzanthrone derivatives, but conventionally they are treated as a separate group.

Although the yield of benzanthrone by dehydrogenating ring closure of 1-benzoylnaphthalene in an aerated aluminum chloride – sodium chloride melt is too low [13, pp. 315 – 317], these conditions are suitable for the double ring closure of 1,5- (or 1,4-) dibenzoylnaphthalene [348] to yield 7,14- (or 5,8-) dibenzpyrenequinone, (**141**), (**140**), [347].

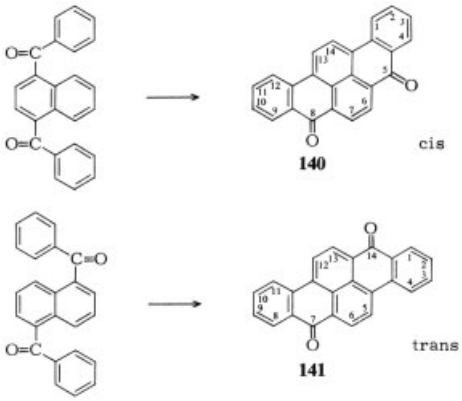

The 7,14-dibenzpyrenequinone is used commercially as a golden-yellow vat dye. Another, now obsolete, method for preparing this compound is fusion of benzanthrone with benzoyl chloride and aluminum chloride [349] and, if necessary, with aeration with oxygen [350].

7,14-Dibenzpyrenequinone [*128-66-5*], Indanthren Golden Yellow GK (**141**), M_r 332.4, is prepared from 1,5-dibenzoylnaphthalene (itself prepared from naphthalene with benzoyl chloride – $AlCl_3$) by fusion with $AlCl_3$-NaCl at 160 °C and aeration with oxygen [5, p. 123], [347], [348].

Dibromo-7,14-dibenzpyrenequinone [*1324-11-4*], Indanthren Golden Yellow RK, M_r 490.2, is prepared by bromination in the melt [5, p. 121].

2.12.6. Anthanthrones

The intensely orange anthanthrone (**145**), which has little fiber affinity when used in its unsubstituted form as a vat dye, is produced by double ring closure of 1,1′-binaphthyl-8,8′-dicarboxylic acid (**144**) in sulfuric acid [351]. Anthanthrone is prepared commercially from naphthostyril (**142**), which is hydrolyzed to **143**, diazotized, and dimerized, losing N_2 [352].

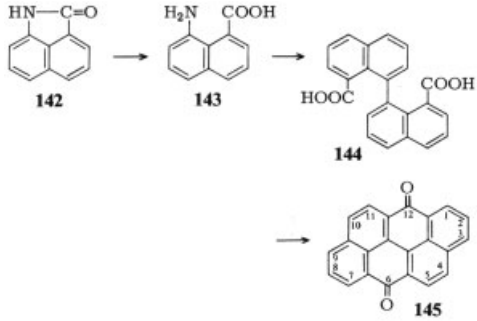

In sulfuric acid, anthanthrone may be dichlorinated or dibrominated in the 4,10 positions. This procedure yields a dye with greater affinity and brighter, more intense red shades. In commercial processes halogenation is carried out immediately following the cyclization of **144**without isolation of intermediates.

4,10-Dichloroanthanthrone [*1324-02-3*], Indanthren Brilliant Orange GK, M_r 375.2, is prepared from 1,1′-binaphthyl-8,8′-dicarboxylic acid in sulfuric acid at 50 °C, followed by chlorination in the presence of $FeSO_4$ [5, p. 96], or from anthanthrone by sparging with chlorine in the presence of iodine in sulfuric acid or organic solvents [353].

4,10-Dibromoanthanthrone [*4378-61-4*], Indanthren Brilliant Orange RK, M_r 464.1, is prepared from 1,1′-binaphthyl-8,8′-dicarboxylic acid in sulfuric acid monohydrate at 35 °C, followed by bromination in the presence of iodine [5, p. 88]. It is also used as a pigment.

4,10-Bis-(4′-benzoylaminoanthraquinonyl-1′-amino)anthanthrone [*6049-19-0*], Indanthren Grey BG, (**146**), M_r 987.0, is prepared from 4,10-dibromoanthanthrone and 1-amino-4-benzoylaminoanthraquinone in naphthalene [354].

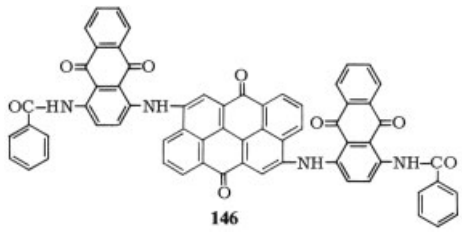

2.12.7. Pyranthrone and Flavanthrone

Pyranthrone [*128-70-1*], (**149**), M_r 406.4, is prepared by double ring closure of 2,2′-dimethyl-1,1′-bianthraquinonyl (**148**), which is obtained from 1-chloro-2-methylanthraquinone (**147**) by heating with copper powder in *o*-dichlorobenzene – pyridine. Ring closure occurs under alkaline conditions, e.g., KOH – isobutanol at 105 °C [2, p. 60]; process improvements are reported in [355], [356]. The unsubstituted pyranthrone and its halogen derivatives are used as vat dyes and pigments.

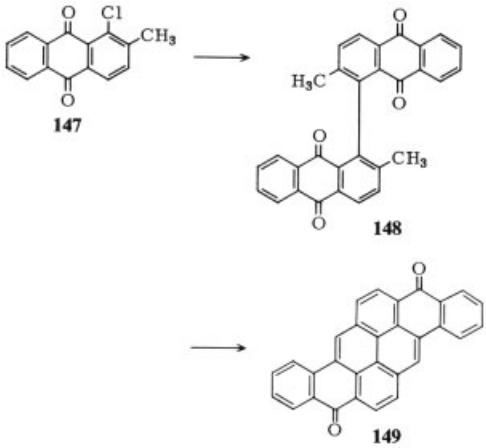

Brominated pyranthrones (the dibromo derivative [*1324-35-2*], M_r 564.3, is Indanthren Orange RRTS [2, p. 61]; a mixture of di- and tribromo derivatives is Indanthren Orange 4 R [2, p. 62]) are prepared from pyranthrone by bromination in chlorosulfonic acid in the presence of iodine or sulfur.

Flavanthrone [*475-71-8*], (**150**), M_r 408.4. Strictly speaking, flavanthrone should not be classified with the polycarbocyclic anthraquinone derivatives, but it is described here because it is so similar to pyranthrone.

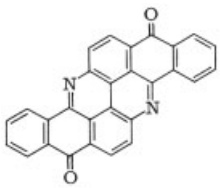

Flavanthrone is prepared from 1-chloro-2-phthalimidoanthraquinone via the 2,2′-diphthalimido-1,1′-bianthraquinonyl, using copper in boiling trichlorobenzene. The alkaline ring closure is accomplished by boiling with excess 3 – 5 % sodium hydroxide, which cleaves both phthalic acid residues simultaneously [5, p. 174]. Only the unsubstituted flavanthrone is used as a vat dye and pigment.

2.12.8. Violanthrone and Isoviolanthrone

Both violanthrone (**151**) and isoviolanthrone (**152**) are formally derived by symmetrical or unsymmetrical condensation of two molecules of benzanthrone. Both polycyclic quinones have intense blue color and are used either by themselves or in the form of simple derivatives as fast violet to green or black vat dyes and as pigments. In contrast to isoviolanthrone, unsubstituted violanthrone is available as a commercial dye.

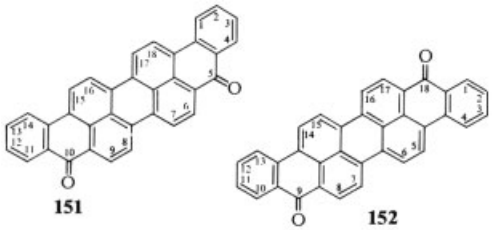

Violanthrone may be brominated or chlorinated in chlorosulfonic acid or organic solvents. Both reactions are used to obtain other hues. If sulfur is added during halogenation in chlorosulfonic acid, derivatives containing sulfur and halogen are obtained.

Nitration of violanthrone yields a mixture of mono- and dinitroviolanthrones, which are reduced in the vat dyeing process to (di)aminoviolanthrones.

Alternatively, reduction of the nitro group may be carried out during the original dye synthesis.

Posttreatment of the resulting greenish fabric color with an oxidizing agent, such as hypochlorite, produces fast, deep black colors.

Violanthrone [*116-71-2*], Indanthren Dark Blue BOA, (**151**), M_r 456.5, is generated by potash fusion of benzanthrone [11], [2, p. 75].

Direct potash fusion leads to side products, such as isoviolanthrone and 4-hydroxybenzanthrone. The crude product may be used directly or in mixtures with other dyes in vat dyeing. It is also used to generate violanthrone derivatives. Several processes have been developed to decrease the amounts of undesirable side products of the direct potash fusion of benzanthrone or to purify the bibenzanthronyl or the crude violanthrone [357–359]. Especially effective is the addition of high-boiling solvents such as naphthalene, the so-called low carbazole anthracene residues, or sodium acetate to the melt. The quality of the products prepared by these methods is satisfactory for the production of almost all violanthrone dyes.

Violanthrone of high purity is obtained from 4,4′-bibenzanthronyl (**153**) by alkaline or acid ring closure, preferably in the presence of oxidizing agents [359], [360]. If this ring closure is carried out in sulfuric acid with an excess of manganese dioxide, the 16,17-violanthronequinone (**155**) is obtained. This can be reduced readily with sodium hydrogen sulfite to 16,17-dihydroxyviolanthrone (**156**). Subsequent alkylation of the hydroxy groups yields very fast navy-blue to brilliant-green vat dyes. The dimethyl ether (**157**) is the well-known dye Caledon Jade Green or Indanthren Brilliant Green B and FFB (extremely pure form).

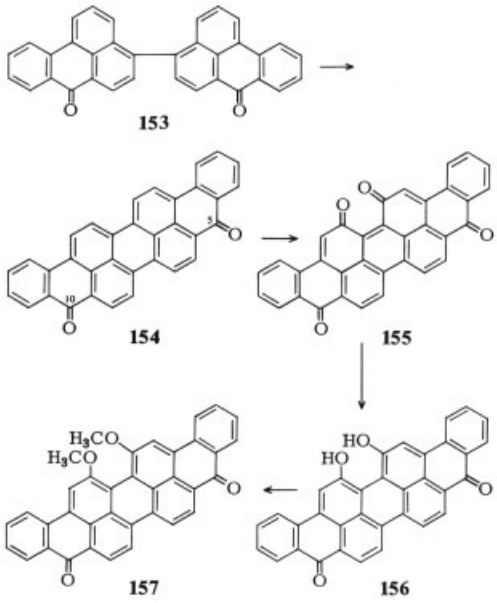

3,12,16,17-Tetrachloroviolanthrone [*84617-19-6*], Indanthren Navy Blue RB, M_r 594.3, is prepared by chlorination in trichlorobenzene [4, p. 27].

Bromoviolanthrones, with one bromine atom (Indanthren Navy Blue BRF, M_r 535.4) or with two bromine atoms (Indanthren Navy Blue BF [*1328-18-3*], M_r 614.3), are prepared by bromination of violanthrone in chlorosulfonic acid in the presence of sulfur [361] or antimony [362].

Nitroviolanthrone [*57455-91-1*], Indanthren Black BB, M_r 501.5, is prepared by nitration of violanthrone [363].

16,17-Dimethoxyviolanthrone [*128-58-5*], Indanthren Brilliant Green FFB, Caledon Jade Green, (**157**), M_r 516.6, is prepared from 4,4′-bibenzanthronyl by ring closure and oxidation followed by reduction and ether formation [1], [5, p. 81], [364].

Isoviolanthrone [*128-64-3*], (**152**), M_r 456.5. Formerly prepared from 3-halobenzanthrone by alcoholic potash melt. Isomer-free isoviolanthrone is obtained by an alcoholic potash melt of 3,3′-dibenzanthronyl sulfide [365] or by the one-pot method [366].

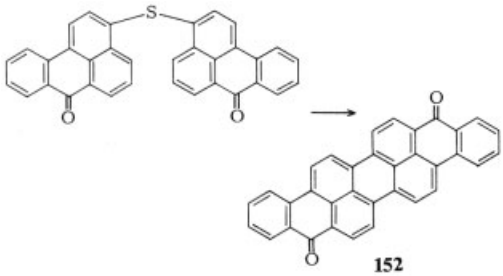

Dichlorination of isoviolanthrone with sulfuryl chloride, or alternatively bromination in chlorosulfonic acid, affords brilliant violet vat dyes and pigments.

Dichloroisoviolanthrone [*1324-55-6*], (Indanthren Brilliant Violet RR and 4 R with chlorine in the 6,15 positions), M_r 524.4, is prepared from isoviolanthrone by chlorination with sulfuryl chloride [367].

Tribromoisoviolanthrone [*1324-17-0*], Indanthren Brilliant Violet 3 B and F 3 B, M_r 693.2, is prepared by bromination of isoviolanthrone in chlorosulfonic acid in the presence of sulfur [4, p. 9].

2.12.9. Acedianthrone

The unusual ring system of acedianthrone [*129-68-0*], (**158**), is generated from anthrone by condensation with glyoxal sulfate [368], followed by alkaline oxidation of the glyoxal anthrone [368] and subsequent ring closure catalyzed by sulfuric acid, or by direct fusion of the glyoxal anthrone with $AlCl_3$-NaCl [369], [370].

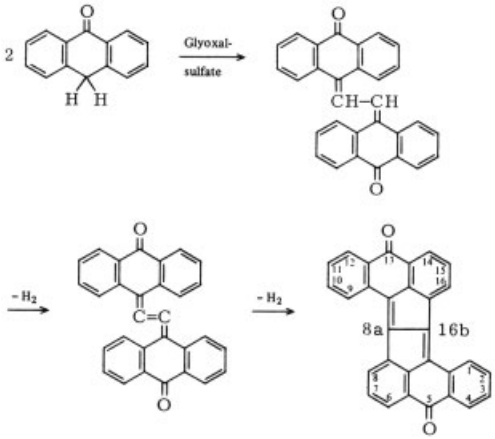

158

This process has found industrial application in the production of the dichloro compound from 2-chloro-9-anthrone.

Dichloroacedianthrone [*6424-51-7*], Indanthren Red Brown RR, (**159**), M_r 475.3, is prepared from 2-chloroanthrone and glyoxal sulfate in chlorobenzene via 2,2′-dichloroanthroneethylene, followed by ring closure in sulfuric acid [4, p. 47], [5, p. 149].

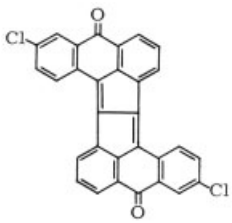

159

3. Dye Classes

The introduction of auxochromes into the almost colorless anthraquinone permits the tailoring of compounds to cover nearly all shades of dye colors: types and positions of the substituents in the molecule determine the hue. As a rule the bathochromic shift in simple anthraquinones increases with increasing basicity of the substituents. This generalization is clearly shown by the wavelengths of the longest wavelength absorption maximum of anthraquinones monosubstituted in the 1-position [371]:

Subst.	λ [mμ]
H	327
Cl	337
OCH_3	380
OH	405
$NHCOCH_3$	410
$NHCOC_6H_5$	415
SCH_3	438
NH_2	465
$N(CH_3)_2$	504
$NHCH_3$	508
NHC_6H_5	508

The wavelengths of the absorption maxima of hydroxy- and aminoanthraquinones are a function of position and number of OH and NH_2 groups [371]:

Position:	2	1	1,2	1,5	1,8	1,4
OH λ [mμ]	365	405	416	428	430	476
NH_2 λ [mμ]	410	465	480	480	492	550

The large shift from the 1,8- to the 1,4-disubstituted anthraquinones, which even exceeds that between the α- and the β-substituted compounds, is remarkable. For α-substituents, additional changes in shade and color fastness occur on formation of hydrogen bonds to the neighboring carbonyl group.

Alteration of color in isomeric compounds can be demonstrated by comparing various diaminohydroxyanthraquinones:

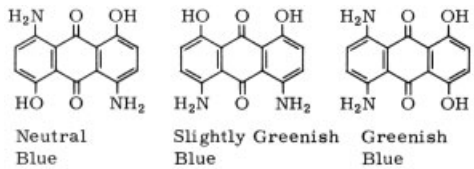

The effect of β-substituents on the hue is best demonstrated by the 1,4-diaminoanthraquinones:

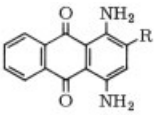

R = OCH_3, Red Violet
= OC_6H_5, Slightly Red Violet
= H, Violet
= $SO_2OC_6H_5$, Red Blue
= $COOCH_3$, Slightly Red Blue
= NO_2, Blue Green

Furthermore, the shade may be modified strongly by ring closure and condensation reactions:

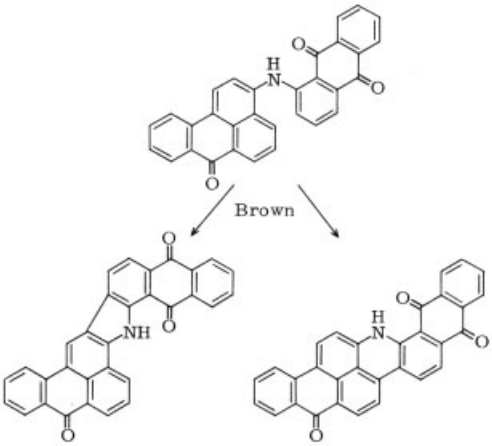

Yellow Brown Olive Green

Not every colored anthraquinone derivative is a dye. In addition to tinctorial strength and depending on the intended use, such other properties as affinity to the fiber and resistance to atmospheric conditions must be present also. The choice of a particular textile dye is determined also by its application properties and its fastness.

The three major dye types, i.e., neutral, anionic, and cationic dyes, are subdivided by use.

Neutral dyes: disperse dyes, vat dyes, pigments, and dyes soluble in organic solvents.
Anionic dyes: acid dyes, direct dyes, and reactive dyes.
Cationic dyes: dyes for polyacrylonitrile fibers and other fibers modified to contain acid groups.

The Colour Index (C.I.) numbers given for the individual dyes in the following sections correspond to the numbering system of the Colour Index [11] where additional references can be found.

3.1. Neutral Dyes

3.1.1. Disperse Dyes

The disperse dyes are water-insoluble, colloidally dispersed materials that are used to dye synthetics, including polyester, polyamide, and acetate fibers having 2.5 or 3 acetate groups per glucose subunit. Disperse dyes are used for dyeing polyacrylonitrile fibers when good leveling of light shades is required. Although disperse dyes were developed many decades ago for coloring cellulose acetate fibers, the most important application is to dye polyester fibers.

The simple anthraquinone dyes, containing hydroxy or amino functions as their major auxochromes, are used for brilliant red to blue shades. Yellow and orange shades are obtained by other dye classes. Choosing appropriate substituents allows the best dyes for various fibers and colors to be obtained.

3.1.1.1. Dyes for Polyester Fibers

Anthraquinone dyes for polyester fibers can be classified into five basic types:

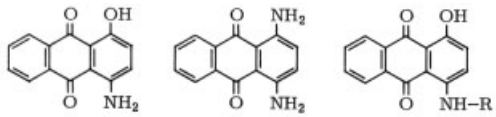

Yellow Red to Rubine | Violet to Blue Green | Red Blue to Blue

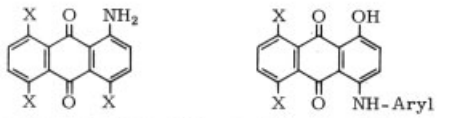

1 X = NH_2, 2 X = OH Blue to Green Blue | 1 X = OH, 1 X = NO_2 Green Blue

The most important factors in the selection of anthraquinone dyes are affinity, light fastness, and resistance to sublimation. Large substituents

in the side chain tend to improve the sublimation fastness considerably. On the other hand, increased sublimation resistance generally decreases the affinity characteristics of the dye. The affinity is greatly affected by the position and the hydrophilic characteristics of the substituents. Mixtures of appropriate dyes or the presence of contaminants formed during synthesis may increase affinity by synergistic action. Introduction of negatively charged substituents, for instance, carboxylic esters, halogen, or sulfone groups, may improve light fastness.

1-Amino-4-hydroxyanthraquinones (for preparation see page 19) possess good light fastness and affinity for polyester fibers. They are bright, red dyes, whose brilliance may be improved significantly by introduction of ether groups ortho to the amino groups. The aliphatic ethers surpass the aromatic ethers with respect to light fastness and are much more yellow and somewhat brighter. Additional substituents in the side chains may improve sublimation resistance. Isomeric compounds with alkyl or aryl ether groups in addition to hydroxyl entities are of little commercial interest because their shades are more blue and duller.

Examples are **160** and **161** both bright yellow reds [372], [373]; **162**, bright blue red (R = H: C.I. Disperse Red 60, C.I. 60756 [*17418-58-5*], e.g., Resolin Red FB, Bayer); for **b** see [374]; for **c** see [375].

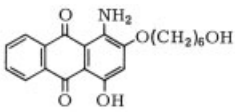

160 Bright Yellow Red

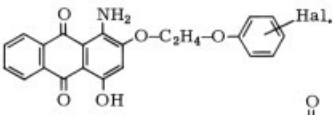

161 Bright Yellow Red

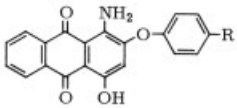

162 Bright Blue Red

a R = H

b R = 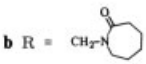

c R = 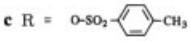

1,4-Diaminoanthraquinones (for preparation see page 9). The poor light fastness of 1,4-diaminoanthraquinone can be improved by appropriate negatively charged substitutents. The 2-sulfophenyl ester is a sublimation-resistant, brilliant blue dye with a reddish cast and good light fastness. The introduction of a nitro group into the 2 position shifts the shade into bluish green and improves resistance to sublimation and fading. Introduction of chlorine atoms into the β position considerably improves light fastness with little affect on the basic sublimation characteristics.

β-Phenoxy groups in the 2,3 positions shift the shade to a bright, somewhat reddish violet with good stability to fading and sublimation. A very bright turquoise color with excellent light fastness and good resistance to sublimation are properties of the 2,3-dicarboximides.

Examples are **167a**, C.I. Disperse Violet 28, 61102 [*81-42-5*] (e.g. Resolin Violet RL, Bayer); **163b**, a red violet dye [3, p. 48]; **164a**, blue green [376]; **167b** brilliant red blue [377]. Compounds **165** and **166** are both turquoise blue dyes [378], [379].

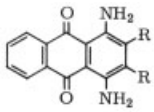

163

a R = Cl Disperse Violet

b R = 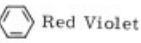 Red Violet

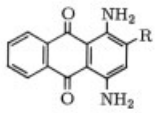

164

a R = NO_2 Blue Green

b R = 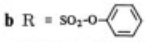

Brilliant Red Blue

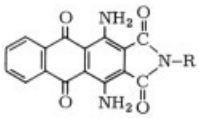

165 Turquoise Blue

R = H, Alkyl, Hydroxyalkyl

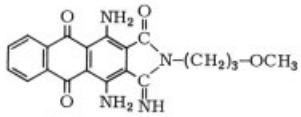

166 Turquoise Blue

***N*-Substituted 4-Amino-1-hydroxyanthraquinones.** Alkyl- (or aryl-) aminohydroxyanthraquinones, which show good affinity and light fastness and give violet to blue shades, generally do not satisfy the requirements with

respect to sublimation fastness. Compared to the dyes of the tetrasubstituted series, they are of lower tinctorial strength. Such substituents as carboxylic esters, arylsulfonic esters, amides, hydroxyethylether, and methoxy groups in arylamino compounds improve sublimation resistance, whereas optimized blends prevent lowering of affinity.

Examples are **167a**, C.I. Disperse Violet 27, 60724 [*19286-75-0*]; **167b**, C.I. Disperse Blue 72, 60725 [*81-48-1*], and **167c**, a violet dye [380].

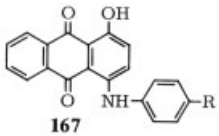

167

a R = H C.I. Disperse Blue 27
b R = CH_3 C.I. Disperse Blue 72
c R = $O{-}SO_2CH_3$ Violet

Diaminodihydroxyanthraquinones (for preparation see page 24). Derivatives of α-diaminodihydroxyanthraquinones are the most important disperse dyes with respect to shade and affinity. The dye properties may be optimized by introducing suitable substituents, selecting the positions of the isomers, and blending. Some of these properties can be illustrated with the three basic structures shown below:

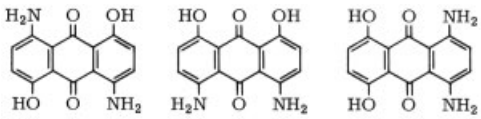

168 **169** **170**

The bathochromic shift from **168** to **170** affects the color; the light fastness increases in the order of **169** → **170** → **168**; and the affinity increases from **168** → **170** → **169**. The sublimation fastness of all three is moderate.

Halo, alkoxy, hydroxyaryl, and phenylmercapto derivatives have attained commercial importance. Compared to substitution next to the hydroxy group, substitution next to the amino group leads to brighter dyes and improved affinity.

Examples are **171**, C.I. Disperse Blue 56, 63285 [*12217-79-7*] (e.g., Resolin Blue FBL, Bayer) and **172,** C.I. Disperse Blue 73, 63265 [*12222-75-2*].

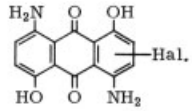

171 C.I. Disperse Blue 56

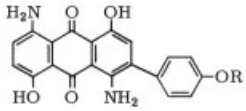

172 C.I. Disperse Blue 73
R = H, CH_3

Nitroarylaminodihydroxyanthraquinones. These dyes are valued because of their greenish-blue shade and their good fastness to light and sublimation. Reduction improves their affinity somewhat but decreases light fastness.

Example : **173** is a blue dye of this kind [381].

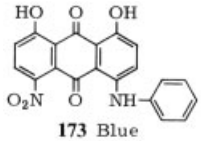

173 Blue

3.1.1.2. Dyes for Cellulose Ester and Synthetic Polyamide Fibers

The first disperse dyes (Celliton) were developed for dyeing cellulose fibers, but the importance of these diminished considerably when other synthetic fibers appeared on the market. Synthetic polyamide fibers could be dyed with dyes used for acetate fibers: very few new dyes had to be developed specifically for polyamide fibers.

The basic type of dye closely resembles that used for dyeing polyester fibers, but the selection of compounds is based on other criteria. Requirements regarding sublimation resistance are not as stringent, whereas fastness to ozone, exhaust gases, and washing are important. Substitution by amino, and especially by alkylamino, groups tends to decrease light fastness of the dyes in polyester fibers. This is not the case for acetate and polyamide fibers.

The orange derivatives of 1-aminoanthraquinone are of little importance in this context because of their low tinctorial strength. This contrasts with the 1-amino-4-hydroxyanthraquinone derivatives, which provide brilliant red dyes. The most important are dyes derived from 1,4-diaminoanthraquinone, with shades ranging

from violet to greenish blue. Affinity may be enhanced vastly by blending similar compounds.

Examples of this class of dyes are **174,** a brilliant red dye [382]; **175,** C.I. Disperse Blue 14, 61500 [*2475-44-7*] [3, p. 54] (e.g., Celliton Fast Blue, BASF); **176,** a brillant blue dye [383]; **177,** C.I. Disperse Blue 31, 64505 [*1328-23-0*] [3, p. 53] (e.g., Celliton Blue 3 G, BASF); **178,** C.I. Disperse Blue 7, 62500 [*3179-90-6*] [5, p. 201] (e.g., Celliton Blue Green B, BASF).

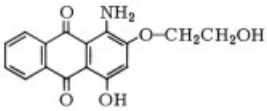

174 Brilliant Red

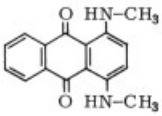

175 C.I. Disperse Blue 14

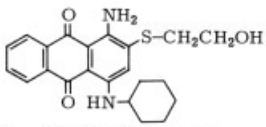

176 Brilliant Blue

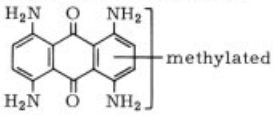

177 C.I. Disperse Blue 31

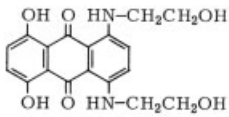

178 C.I. Disperse Blue 7

3.1.1.3. Transfer Dyes

In transfer printing the dye is supplied in the form of a coating on transfer paper. The fabric is pressed closely against the paper, and the dye is sublimed at ≈ 200 °C and diffuses into fibers. This process is used primarily for printing on polyester fabrics. Originally it relied on available disperse dyes with good sublimation characteristics. Dyes, specifically developed for this process have also appeared on the market (see also [6, vol. VIII, pp. 191 ff.]).

Examples are **179,** C.I. Disperse Red 60, 60756 [*17418-58-5*], and **180,** C.I. Disperse Blue 26, 63305 [*3860-63-7*].

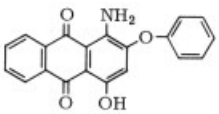

179 Disperse Red 60

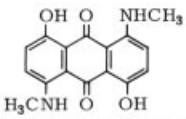

180 Disperse Blue 26

3.1.1.4. Dyes for Cotton – Polyester Fabrics

Anthraquinone dyes of medium molecular mass are suitable for direct printing and dyeing of cellulose fibers, especially cotton – polyester fabrics pretreated with water. Most of these dyes are classified as disperse dyes having excellent resistance to sublimation, but vat dyes of low molecular mass are included also. The dyes are applied generally together with higher boiling, water-miscible solvents (glycol and glycol derivatives [384] or boric acid esters of species with one to six hydroxy groups [385]) to fabrics preswollen with water. Heat treatment at ≈ 200 °C evaporates the water, and the dye enters the fiber via its solution phase. The polyester component of the fabric is dyed simultaneously.

Examples: Compounds **181** [384] and **182** [386] are blue dyes; **183** [387] is green.

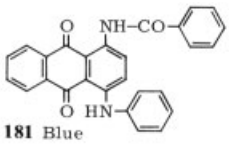

181 Blue

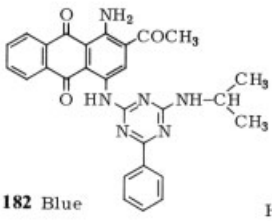

182 Blue

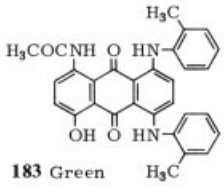

183 Green

Reactive disperse dyes also make it possible to dye cotton–polyester fabrics in equal shades. Fluorotriazine reactive dyes have been proposed for this application [388–390].

Examples: Compound **184** [388] is a blue, and **185** [389] a red dye.

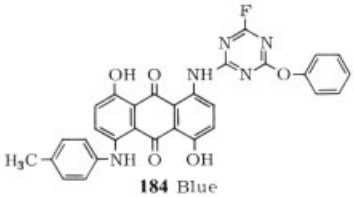

184 Blue

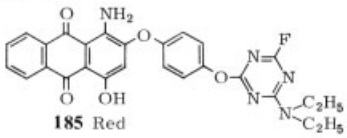

185 Red

3.1.2. Dyes Soluble in Organic Solvents

Anthraquinone dyes of relatively simple structure are used for coloring gasoline, oil, and plastics. For example, the highly soluble bis(alkylamino)anthraquinones are suitable additives for gasoline, whereas quinizarin and its derivatives are used for marking heating oils. Note: addition of alkali causes a color change.

Examples: Compound **186** is a blue dye for the coloring of hydrocarbons, including gasoline [391]. Compound **187** is a mixture of blue dyes used to color gasoline and mineral oils [392].

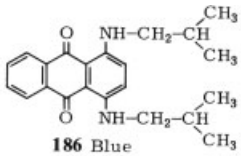

186 Blue

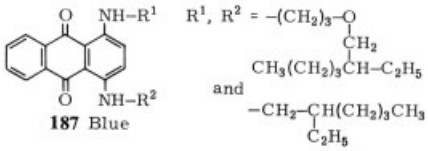

187 Blue

Dyeing of amorphous thermoplastics – polymethacrylate, (modified) polystyrene, or polycarbonate – requires dyes soluble in these materials so that transparency can be retained. However, such dyes also are used in combination with titanium dioxide or other materials to provide body colors for the thermoplastics. Initially compounds with relatively simple structures were used, many of them drawn from stocks of existing intermediates. Since then new products designed to satisfy the specific requirements have become available.

Examples are **188**, C.I. Solvent Red 111, 60505 [*82-38-2*]; **189**, C.I. Solvent Red 52, 68210 [*81-39-0*]; **190**, C.I. Solvent Violet 13, 60725 [*81-48-1*]; **191**, C.I. Solvent Green 3, 61565 [*128-80-3*].

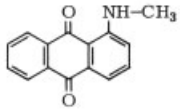

188 C.I. Solvent Red 111

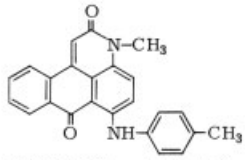

189 C.I. Solvent Red 52

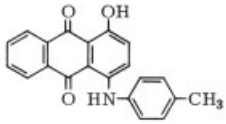

190 C.I. Solvent Violet 13

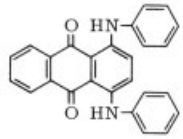

191 C.I. Solvent Green 3

Many relatively simple amino- and hydroxyanthraquinone derivatives show a high degree of order in liquid crystalline systems and are therefore suitable as dichroic dyes for guest – host displays. Several new dyes for this application have been described.

Example: Compound **192** is a blue dichroic dye [393].

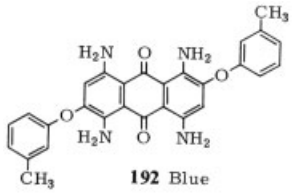

192 Blue

3.1.3. Vat Dyes

Vat dyes have been used for many decades to color cotton and other cellulose fibers. Despite their high cost and their muted colors, these dyes are extremely important for certain textiles because of their superior fastness. Very few new vat dyes have been developed over the past few years.

Water-insoluble vat dyes are converted to soluble "anthrahydroquinones" by reducing agents, such as sodium dithionite (hydrosulfite) in the presence of sodium hydroxide. The sodium salts of these mostly deep colored leuco compounds penetrate cellulose fibers. The insoluble dye is attached firmly to the fiber after reoxidation (→ Textile Dyeing). Representative of a special form are the water-soluble sulfuric acid es-

ters of the "anthrahydroquinone" compounds (→ Leuco Esters of Vat Dyes).

On the basis of their chemical constitutions the anthraquinoid vat dyes may be classified in the following major groups:

acylaminoanthraquinones
anthraquinoneazoles
anthrimides and other linked anthraquinones
anthrimidocarbazoles
phthaloylacridones
benzanthrone dyes
indanthrones
other polycondensed ring systems

Acylaminoanthraquinones. Acylation of aminoanthraquinones with benzoic acid or benzoyl chloride, for example, affords vat dyes with satisfactory affinity for cellulose fibers. The simplest dyes of this type, 1,4- (and 1,5-) dibenzoylaminoanthraquinones, are no longer important. Bridging linkages, such as the dicarboxylic acids oxalic or phthalic acid, permit coupling of two anthraquinone units; three aminoanthraquinones may be combined by use of a triazine, such as cyanuric chloride.

Acylation of 1-amino- (or 1,5-diamino) anthraquinones yields yellow vat dyes and affords red to ruby colored dyes when 1,4-diaminoanthraquinones are used. Use of 4,8-diamino-1,5-dihydroxyanthraquinones gives violet to blue dyes. The relatively low light fastness of the yellow acylaminoanthraquinones may be improved greatly by using azodiphenyl-4,4′-dicarboxylic acid. All acylaminoanthraquinones are relatively sensitive to atmospheric conditions, except the yellow acylation product obtained from anthrapyrimidinecarboxylic acid and 1-aminoanthraquinone.

Examples of this class of dyes are **193**, C.I. Vat Violet 15, 63355 [*6370-58-7*] [4, p. 8] (e.g., Indanthren Brilliant Violet BBK, Bayer); **194** C.I. Vat Yellow 12, 65405 [*6370-75-8*] [4, p. 56] (e.g., Indanthren Yellow 3 GF, Bayer); **195**, C.I. Vat Orange 17, 65415 [*6370-77-0*] [394] (e.g., Indanthren Orange GG, Bayer); **196**, C.I. Vat Yellow 10, 65430 [*2379-76-2*] [5, p. 178] (e.g., Indanthren Yellow GGF, Cassella); **197**, C.I. Vat Yellow 20, 68420 [*4216-01-7*] [395] (e.g., Indanthren Yellow 4 GF, BASF); **198**, C.I. Vat Red 28, 65710 [*6370-82-7*] [10, p. 353] (e.g., Cibanone Red G, Ciba-Geigy).

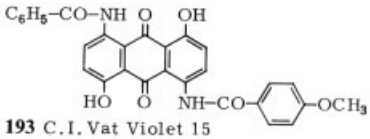

193 C.I. Vat Violet 15

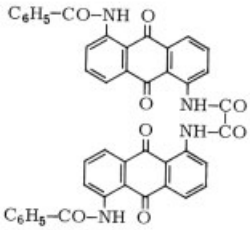

194 C.I. Vat Yellow 12

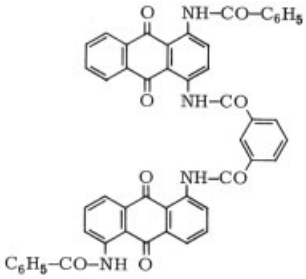

195 C.I. Vat Orange 17

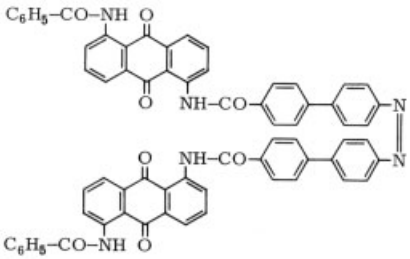

196 C.I. Vat Yellow 10

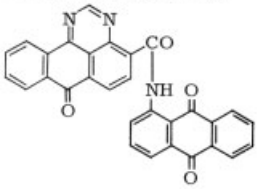

197 C.I. Vat Yellow 20

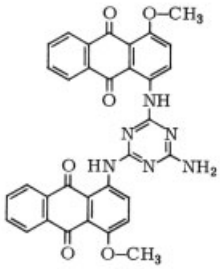

198 C.I. Vat Red 28

Anthraquinoneazoles. In contrast to the older yellow Algol dyes, which contain two thiazole rings (see page 32), the red to blue oxazoles and thiazoles derived from 1-aminoanthraquinone-2-carboxylic acid and 3-amino-2-hydroxy- (or -mercapto-) anthraquinones exhibit good light fastness. The good fastness to atmospheric conditions and chlorine of the blue derivatives of 1-amino-4-aroylaminoanthraquinone-2-carboxylic acid deserves special mention.

Examples are **199**, C.I. Vat Red 10, 67000 [*2379-79-5*] [5, p. 152] (e.g., Indanthren Red FBB, BASF); **200**, C.I. Vat Blue 30, 67110 [*6492-78-0*] [2, p. 17], [5, p. 72] (e.g., Indanthren Blue CLB, BASF).

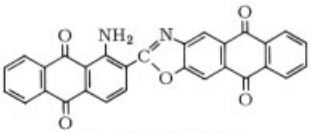

199 C.I. Vat Red 10

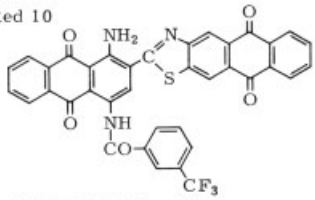

200 C.I. Vat Blue 30

Anthrimides and Other Linked Anthraquinones. Among the anthrimides (dianthraquinonylamines), only the α,β derivatives have reached a limited importance as vat dyes. Coupling two anthraquinone molecules via functional derivatives of the 2-aldehyde (or 2-carboxy) group offers another type of building block for vat dyes. Such compounds, e.g., 1-aminoanthraquinones, are linked in the 2 position via an azine or oxadiazole group, and all have good fastness.

Examples are **201**, C.I. Vat Violet 16, 65020 [*4003-36-5*] [4, p. 14] (e.g., Indanthren Corinth RK, Bayer); **202**, C.I. Vat Red 18, 60705 [*6409-68-3*] [2, p. 64] (e.g., Indanthren Bordeaux B, BASF); **203**, C.I. Vat Blue 64, 66730 [*15935-52-1*] [396] (e.g., Indanthren Blue ER, Bayer); **204**, C.I. Vat Red 13, 70320 [*4203-77-4*] [5, p. 159] (e.g., Indanthren Rubine R, Cassella).

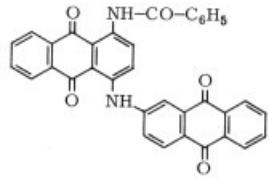

201 C.I. Vat Violet 16

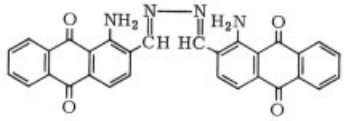

202 C.I. Vat Red 18

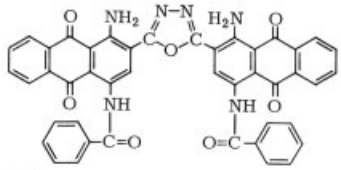

203 C.I. Vat Blue 64

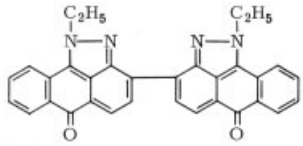

204 C.I. Vat Red 13

Anthrimide Carbazoles. Fast vat dyes are produced by carbazole ring closure from α,α-dianthraquinonylamines (anthrimides). The shade is determined by the number and position of the carbazole systems and by additional substituents, especially acylamino or alkoxy groups. Anthraquinone carbazoles make it possible to dye cellulose fibers in level, very fast yellow, orange, brown, gray, and olive shades. However, this series lacks dyes with bright shades.

Examples include **205**, C.I. Vat Yellow 28, 69000 [*4229-15-6*] [4, p. 53] (e.g., Indanthren Yellow FFRK, Hoechst); **206**, C.I. Vat Orange 15, 69025 [*2379-78-4*] [4, p. 15], [5, p. 119] (e.g., Indanthren Golden Orange 3 G, Bayer); **207**, C.I. Vat Brown 3, 69015 [*131-92-0*] [4, p. 10] (e.g., Indanthren Brown FFR, Bayer); **208**, C.I. Vat Black 27, 69005 [*2379-81-9*] [4, p. 29], [5, p. 135] (e.g., Indanthren Olive R, Bayer); **209**, C.I. Vat Brown 1, 70800 [*2475-33-4*] [5, p. 101] (e.g., Indanthren Brown BR, Bayer); **210**, C.I. Vat Green 8, 71050 [*14999-97-4*] [4,

p. 23] [5, p. 129] (e.g., Indanthren Khaki GG, Hoechst).

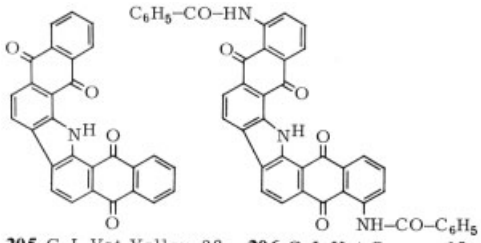

205 C.I. Vat Yellow 28 **206** C.I. Vat Orange 15

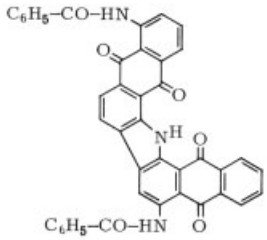

207 C.I. Vat Brown 3

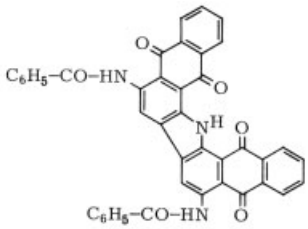

208 C.I. Vat Black 27

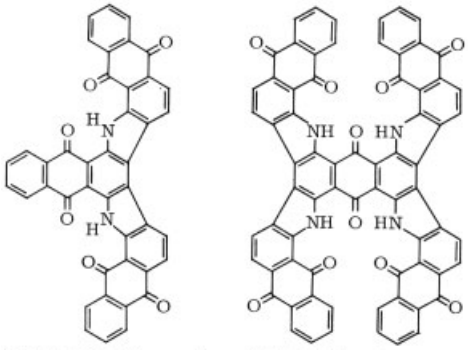

209 C.I. Vat Brown 1 **210** C.I. Vat Green 8

Phthaloylacridones. Depending on the substituent in the 2 position, the 3,4-phthaloylacridones supply shades varying between red and green. The acridone vat dyes combine good light fastness with somewhat poorer wash fastness. Dyes of this series are especially suited for printing purposes because of the ease of vatting.

Examples are **211**, C.I. Vat Blue 21, 67920 [*6219-97-2*] [4, p. 34], [5, p. 143] (e.g., Indanthren Blue HCGK, Hoechst); **212**, C.I. Vat Green 12, 70700 [*6661-46-7*] [4, p. 19].

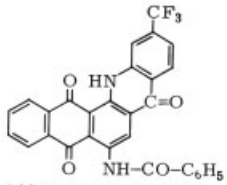

211 C.I. Vat Blue 21

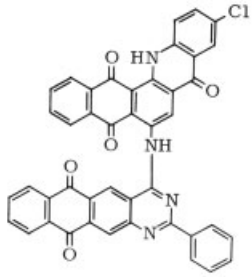

212 C.I. Vat Green 12

Benzanthrone Dyes. Vat dyes derived from benzanthrone can be subdivided into two major groups: the peri ring-closure products of 3-anthraquinonylaminobenzanthrone, referred to as 'imide-green' dyes, and the dyes of the violanthrone and the isoviolanthrone series.

The first class gives the muted colors olive green, olive, khaki, and grey and excels in its resistance to light and atmospheric conditions. The blue color of violanthrone is shifted to a brilliant green by introducing two alkoxy groups into the 16 and 17 positions. An additional shift may be achieved by halogenation.

A redder and brighter shade is obtained with isoviolanthrone. Its halogenation products are marketed as brilliant violet dyes. Related to violanthrone is the alkaline ring closure product of the 3-pyrazolanthronylbenzanthrone, which is used as a navy blue vat dye.

Examples are **213**, C.I. Vat Green 3, 69500 [*3271-76-9*] [2, p. 71] (e.g., Indanthren Olive Green B, BASF); **214**, C.I. Vat Black 25, 69525 [*4395-53-3*] [2, p. 76] (e.g., Indanthren Olive T, BASF); **215**, C.I. Vat Blue 25, 70500 [*6247-39-8*] [4, p. 26] (e.g., Indanthren Navy Blue R, Hoechst); **216**, C.I. Vat Blue 20, 59800 [*116-71-2*] [5, p. 108] (e.g., Indanthren Dark Blue BOA, BASF); **217**, C.I. Vat Green 1, 59825 [*128-58-5*] [2, p. 69], [5, p. 83] (e.g., Indanthren Brilliant Green FFB, BASF); **218**, C.I. Vat Violet 1, 60010 [*1324-55-6*] [367] (e.g., Indanthren

Brilliant Violet RR, BASF); **219**, C.I. Vat Blue 26, 60015 [*4430-55-1*] [414] (e.g., Indanthren Cyanine B, BASF).

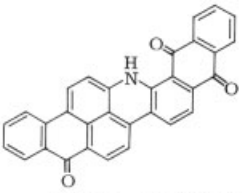

213 C.I. Vat Green 3

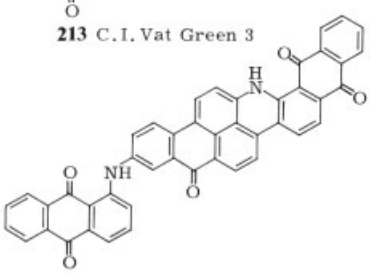

214 C.I. Vat Black 25

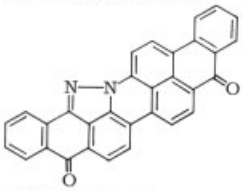

215 C.I. Vat Blue 25

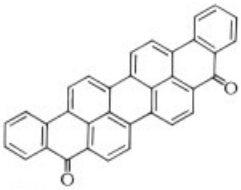

216 C.I. Vat Blue 20

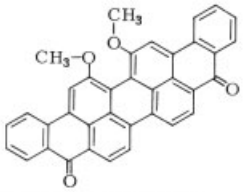

217 C.I. Vat Green 1

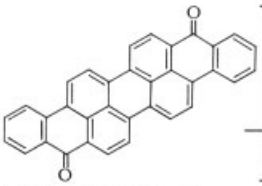

218 C.I. Vat Violet 1

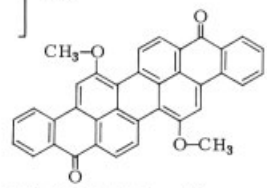

219 C.I. Vat Blue 26

Indanthrones. The blue indanthrone was the first synthetic vat dye of the anthraquinone series. Because of its excellent fastness and bright colors it has remained the most important vat dye for a long time despite its low resistance to chlorine. Its chlorine resistance can be improved somewhat by post-halogenation. Introduction of hydroxy or amino groups shifts the shade to green.

Examples are **220**, C.I. Vat Blue 4, 69800 [*81-77-6*] [2, p. 52], [5, p. 73] (e.g., Indanthren Blue RS, BASF); **221**, C.I. Vat Green 11, 69850 [*1328-41-2*] [4, p. 18] (e.g., Indanthren Green BB, Bayer).

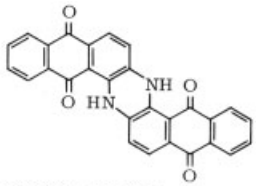

220 C.I. Vat Blue 4

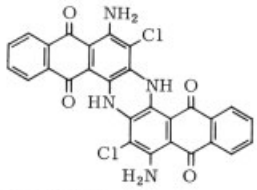

221 C.I. Vat Green 11

Highly Condensed Ring Systems. A valuable supplement to the anthraquinone vat dyes is found in the more highly condensed carbocyclic quinones dibenzpyrenequinone, anthanthrone, and pyranthrone. These substances are yellow to red vat dyes without additional auxochromic substituents. Halogenation may improve their substantivity and change their shades. The halo derivatives can be converted to anthrimide-like compounds by reaction with aminoanthraquinones. These are in themselves useful as vat dyes and dye cotton in brown to grey shades but can be subjected to further carbazolyation.

Examples are **222**, C.I. Vat Orange 1, 59105 [*1324-11-4*] [2, pp. 117, 119] (e.g., Indanthren Golden Yellow RK, Hoechst); **223**, C.I. Vat Orange 3, 59300 [*4378-61-4*] [2, p. 129], [5, p. 90] (e.g., Indanthren Brilliant Orange RK, Cassella); **224**, C.I. Vat Black 29, 65225 [*6049-19-0*] [2, p. 23] (e.g., Indanthren Grey BG, Bayer); **225**, C.I. Vat Orange 2, 59705 [*1324-35-2*] [2, p. 61], [5, p. 116] (e.g., Indanthren Orange RRTS, BASF); **226**, C.I. Vat Brown 45, 59500

[*6424-51-7*] [4, p. 47], [5, p. 149] (e.g., Indanthren Red Brown RR, Cassella); **227**, C.I. Vat Orange 9, 59700 [*128-70-1*] [2, p. 60], [5, p. 114] (e.g., Indanthren Golden Orange G, BASF).

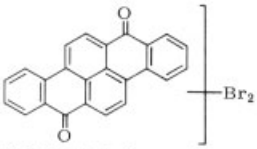
222 C.I. Vat Orange 1

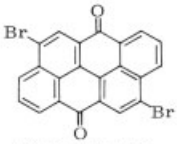
223 C.I. Vat Orange 3

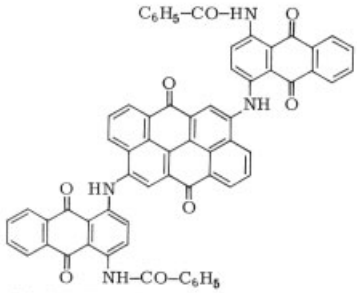
224 C.I. Vat Black 29

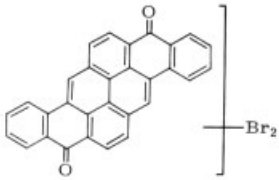
225 C.I. Vat Orange 2

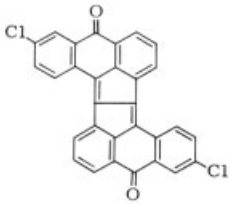
226 C.I. Vat Brown 45

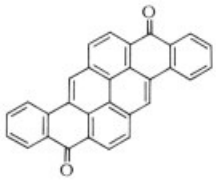
227 C.I. Vat Orange 9

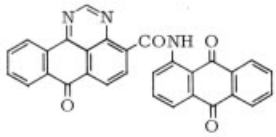
228 C.I. Pigment Yellow 108

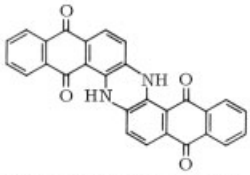
229 C.I. Pigment Blue 60

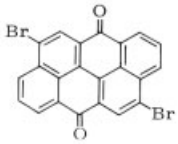
230 C.I. Pigment Red 168

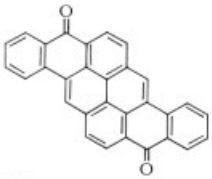
231 C.I. Pigment Orange 40

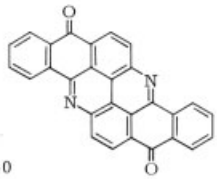
232 C.I. Pigment Yellow 24

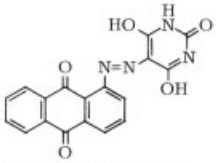
233 Yellow Pigment

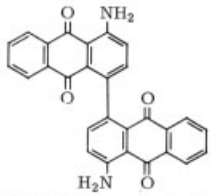
234 C.I. Pigment Red 177

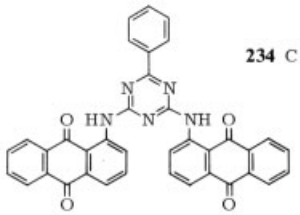
235 Yellow Pigment

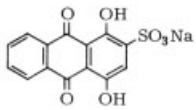
236 C.I. Pigment Violet 5:1

3.1.4. Pigments

Because of their low solubility in organic media, most vat dyes should also be usable as pigments. However, for this type of use they must be prepared with a high degree of purity and with special physical properties (crystal modification, particle size distribution, etc.). The relatively high costs of synthesizing anthraquinone pigments restrict their use to special applications where requirements for fastness are stringent. Only a few of the known vat dyes satisfy these requirements. Apart from some acylaminoanthraquinones, mostly dyes with polycondensed ring systems, including indanthrones, anthanthrones, pyranthrones, flavanthrone, and their halogen derivatives, have some importance as pigments. Beyond that a number of spe-

cialty anthraquinone dyes have been developed as pigments over the past 25 years: for example, anthraquinone-azo, bianthraquinonyl, and anthraquinonylaminotriazine derivatives.

Only of little interest at present are the colored lakes, especially aluminum lakes of the hydroxyanthraquinones, such as alizarine, purpurin, and quinizarin and their sulfonic acids. (For preparation see Chap. 2; for formulation and application → Pigments, Organic, and [398].)

Examples are, **228** C.I. Pigment Yellow 108, 68420 [*4216-01-7*]; **229** C.I. Pigment Blue 60, 69800 [*81-77-6*]; **230** C.I. Pigment Red 168, 59300 [*4378-61-4*]; **231** C.I. Pigment Orange 40, 59700 [*128-70-1*]; **231** C.I. Pigment Yellow 24, 70600 [*475-71-8*]; **233**, yellow pigment [399]; **234**, C.I. Pigment Red 177, 65300 [*4051-63-2*]; **235**, yellow pigment [400]; **236** Aluminum Lake, C.I. Pigment Violet 5 : 1; 58055 : 1 [*145-48-2*].

3.2. Anionic Dyes

The anionic anthraquinone dyes may be classified by type of application and dyeing procedure into the following groups:

acid dyes
direct dyes
reactive dyes

3.2.1. Acid Dyes

Acid dyes are used for dyeing wool, synthetic polyamides, and silk in aqueous media. They may be subdivided into the following basic types:

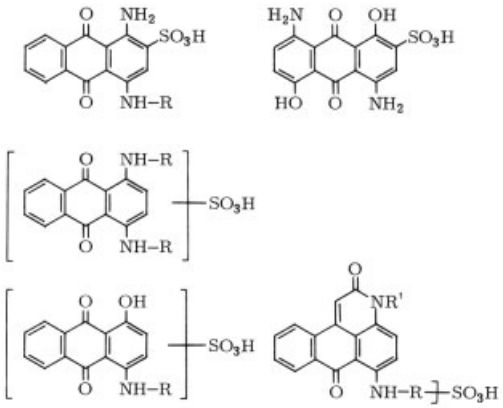

The majority of the acid anthraquinone dyes available commercially give bright blue shades not obtainable with azo dyes. The red and yellow anthraquinone dyes are of little importance. Dyes of green shades obtained by combining yellow and blue dyes possess mostly slight wash fastness. Here, the uniformly dyeing green dyes of the anthraquinone series have proved their special value. The acid anthraquinone dyes are classified for particular applications according to their leveling characteristics, light fastness, and wash fastness. Once synthetic polyamide fibers were introduced in the market appropriate types were selected from the existing stocks. Special acid dyes have been developed for polyamides.

1-Aminoanthraquinone-2-sulfonic Acids. Condensation of bromamine acid (1-amino-4-bromoanthraquinone-2-sulfonic acid, see page 16) with aromatic or cycloaliphatic amines is used to produce a large number of blue acid dyes. The shade, leveling characteristics, wash fastness, and light fastness may be varied over a wide range by choosing particular amines. Cycloaliphatic amines provide the same brightness as aliphatic dyes but impart greater light fastness. Arylamines substituted with alkyl, halogen, aryl, aryloxy, or sulfonic ester groups yield dyes with better wash fastness and affinity in neutral media but less uniform leveling. Substituents at the *o* position cause a hypochromic shift and increase brilliance.

The arylamino residue may be altered subsequently by sulfonation, halogenation, acylation, or by the Einhorn reaction. Substitution of anthraquinone in the 5, 6, 7, or 8 positions offers an additional possibility to change the characteristics of the dye. For instance, halogen atoms and sulfonic acid groups cause bathochromic effects that are most pronounced when the substituents are introduced in the β position. The solubility of the 2,6- (or 2,7-) disulfonic acids is higher than that of the 2,5- (or 2,8-) series.

Examples are **237,** C.I. Acid Blue 25, 62055 [*6408-78-2*] [3, p. 41] (e.g., Acilan Direct Blue A, Bayer); **238,** C.I. Acid Blue 62, 62045 [*4368-56-3*] [3, p. 30] (e.g., Alizarine Brilliant Blue R, Bayer); **239,** C.I. Acid Blue 129, 62058 [*6397-02-0*] [401]; **240,** C.I. Acid Blue 40,

62125 [*6424-85-7*] [2, p. 135] (e.g., Anthralan Blue G, Hoechst); **241** [402] and **242** [404] are brilliant blue dyes.

Diaminodihydroxyanthraquinonesulfonic Acids. These dyes belong to the oldest synthetic acid wool dyes, but their importance has decreased considerably. An example for this class is **243,** C.I. Acid Blue 43, 63000 [*2150-60-9*] [3, p. 42] (e.g., Acilan Saphirol SE, Bayer).

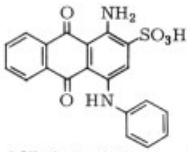

237 C.I. Acid Blue 25

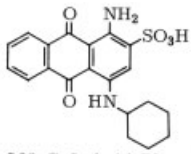

238 C.I. Acid Blue 62

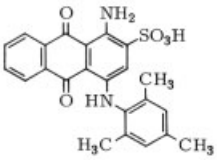

239 C.I. Acid Blue 129

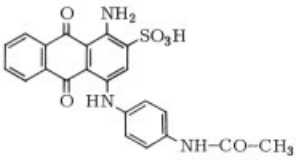

240 C.I. Acid Blue 40

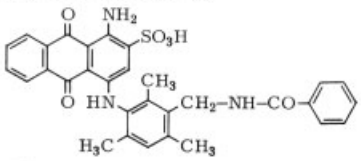

241 Brilliant Blue

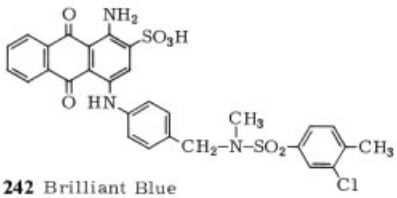

242 Brilliant Blue

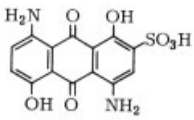

243 C.I. Acid Blue 43

1,4-Diaminoanthraquinones with External Sulfonic Acid Groups. A common feature of dyes of this group is their manufacture by sulfonation of the corresponding dye base derived from quinizarin or haloaminoanthraquinones. The reaction products of quinizarin with aromatic or araliphatic amines predominate in number and importance. The introduction of hydroxy groups into the 5 or the 5,8 positions brings about the expected bathochromic shift. Wash fastness and leveling properties may be altered by substitution. The shade can be varied from brilliant blue to green by appropriate amines. Araliphatic and cycloaliphatic amines lead to brilliant blue shades. A similar effect is exhibited by sterically hindered aromatic amines such as mesidine. These products are more light fast than the derivatives of bromamine acid. Among the unsymmetrically substituted dyes derived from 1-amino- (or 1-alkylamino-) 4-haloanthraquinones, the 1-alkylaminoanthraquinones have the lower light fastness. Exceptions are the derivatives of the 1-*sec*-alkylamino-4-haloanthraquinones. Introduction of alkoxy- (or aryloxy) groups into position 2 shifts the shade of the 4-arylamino-1-aminoanthraquinones to a bright violet.

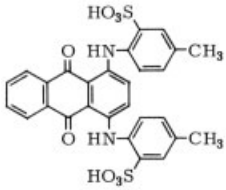

244 C.I. Acid Green 25

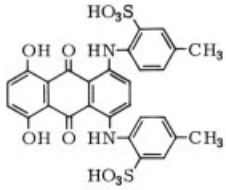

245 C.I. Acid Green 41

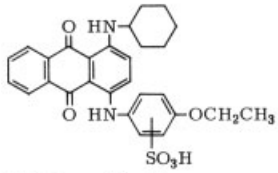

246 Greenish Blue

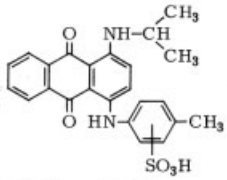

247 Greenish Blue

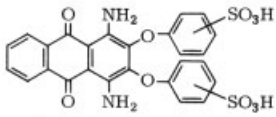

248 C.I. Acid Violet 42

Examples are **244**, C.I. Acid Green 25, 61570 [*4403-90-1*] [5, p. 215] (e.g., Alizarine Cyanine Green G, Bayer); **245**, C.I. Acid Green 41, 62560

[*4430-16-4*] [3, p. 35]; **246** [406] and **247** [407] are greenish-blue dyes; **248**, C.I. Acid Violet 42, 62026 [*6408-73-7*] [3, p. 48] (e.g., Supracen Violet 3 R, Bayer).

1-Amino-4-hydroxyanthraquinones with External Sulfonic Acid Groups. By one-sided reaction of quinizarin with arylamines, followed by sulfonation, violet leveling dyes are obtained. Derivatives of 1-amino-4-hydroxy-2-phenoxyanthraquinones were developed specially for polyamide fibers.

Examples are **249,** C.I. Acid Violet 43, 60730 [*4430-18-6*] [3, p. 48] (e.g., Supracen Violet 3 B, Bayer); **250** bluish brilliant red dye [408].

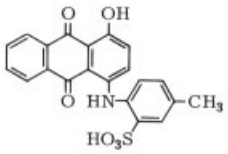

249 C.I. Acid Violet 43

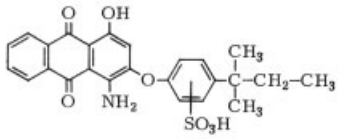

250 Bluish Brilliant Red

Anthrapyridones. Of the anthraquinone dyes with meso rings only species sulfonated in the arylamino or phenoxy groups derived from 6-arylamino-3-methyl-, 6-arylamino-1-carbethoxy-4-methyl-, and 6-arylamino-1-benzoyl-4-phenoxy-anthrapyridones are known to have been used commercially. They once were very important as light-fast, red, leveling dyes but have lost their position with the availability of the fast, red, azo dyes.

An example is **251,** C.I. Acid Red 82, 68205 [*2611-80-5*] [3, p. 40].

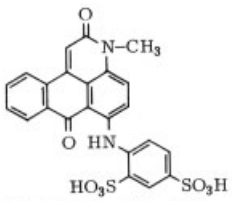

251 C.I. Acid Red 82

Other Acid Anthraquinone Dyes. In addition to the dyes in the preceding classes, a whole series of specially developed products is available. For instance, derivatives of the anthrimide or carbazole series are known to be very light-fast gray and brown wool dyes. The post-sulfonation products of 1,5- (and 1,8-) diarylaminoanthraquinones are violet dyes commonly applied as mixtures.

An example is **252,** C.I. Acid Black 48, 65005 [*1328-24-1*] [5, p. 216] (e.g., Alizarine Light Grey BBLW, BBL, Bayer).

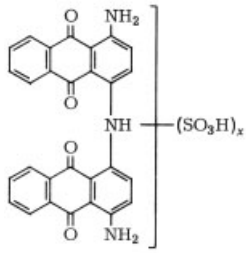

252 C.I. Acid Black 48

3.2.2. Direct Dyes

Compared to direct azo dyes, the direct anthraquinone dyes have lower tinctorial strengths and are therefore far less economical to use. They have lost most of their importance. Only a few special green dyes have retained their importance. Direct green cotton dyes can be produced by coupling a blue bromamine acid dye and a yellow azo dye via ureido or diaminotriazine bridges.

An example is compound **253,** C.I. Direct Green 28, 14155 [*6471-09-6*].

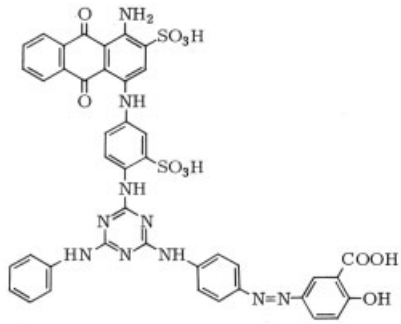

253 C.I. Direct Green 28

3.2.3. Reactive Dyes

The shades of the commercial reactive anthraquinone dyes range generally only from violet to blue. These dyes are used specifically for the production of fast brilliant blue shades on cellulose fibers. Their synthesis begins with basic dyes linked to the reactive component via an amino group. These amino or alkylamino group may be positioned directly on the anthraquinone nucleus (rare) or on an aryl or cycloalkyl residue. Another possible linkage involves aminomethylarylaminoanthraquinones.

Nearly all reactive anthraquinone dyes are derivatives of bromamine acid (see Section 3.2.1). To obtain a sufficiently soluble dye, at least one additional sulfonic acid group must be present in the molecule. This may be positioned directly on the anthraquinone nucleus, in the 5 to 8 positions, or may be linked to the arylamino substituent. It is also possible to introduce one or two additional sulfonic acid groups together with a reactive group.

A large number of fiber-reactive components has been proposed for the preparation of reactive anthraquinone dyes (→ Reactive Dyes). Reactive carboxylic acid derivatives, such as 2,3-dichloroquinoxaline-6-carbonyl chloride and α-bromoacryloyl chloride, or reactive chloro or fluoro heterocyclics, such as derivatives of triazine or pyrimidine, have become important industrially.

Dyes with reactive vinylsulfonyl or β-sulfatoethylsulfonyl groups are constructed on a somewhat different principle.

The various options in printing or continuous dyeing are affected more by the choice of the reactive group than by the chromophoric system.

Examples of this class of dyes are **254,** C.I. Reactive Blue 2, 61211 [*12236-82-7*] (e.g., Procion Blue H-B); **255,** C.I. Reactive Blue 4, 61205 [*13324-20-4*] (e.g., Procion Blue MX-R); **256,** reddish brilliant blue [409]; **257,** brilliant blue [410]; **258,** blue [411]; **259,** brilliant blue [412]; **260,** C.I. Reactive Blue 19, 61200 [*2580-78-1*] (e.g., Remazol Brilliant Blue R, Hoechst); **261,** reddish brilliant blue for wool [413].

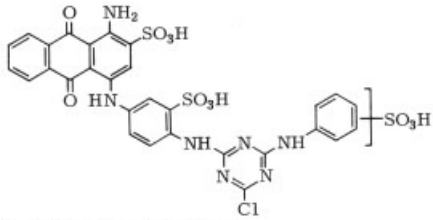

254 C.I. Reactive Blue 2

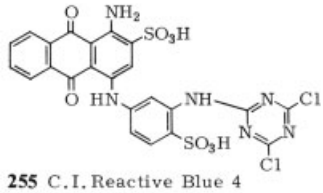

255 C.I. Reactive Blue 4

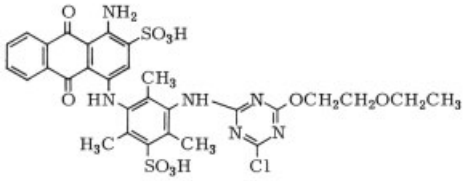

256 Reddish Brilliant Blue

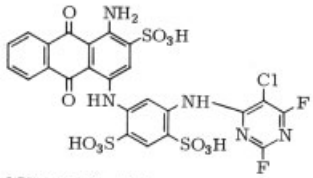

257 Brilliant Blue

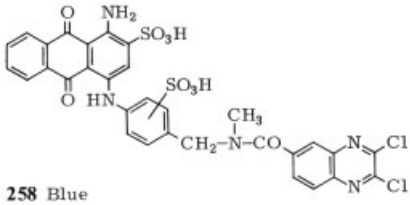

258 Blue

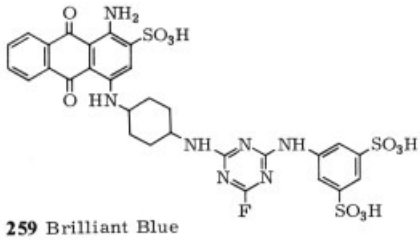

259 Brilliant Blue

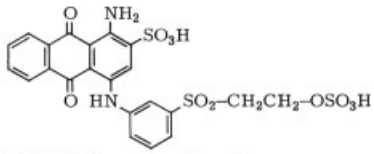

260 C.I. Reactive Blue 19

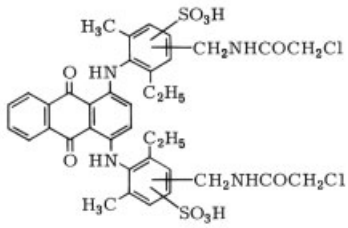

261 Reddish Brilliant Blue

3.3. Cationic Dyes

Water-soluble cationic anthraquinone dyes have become valuable for dyeing polyacrylonitrile fibers. Dyes with external ammonium groups are especially important. Dyes with quaternary ammonium groups as well as salts of sufficiently basic amino derivatives are used also. The majority of commercial dyes are alkylamino- or arylaminoanthraquinones that carry an ammonium group on the alkyl group, such as trialkylammonium, cycloammonium, or hydrazonium residue. A protonated dialkylamino group also may be present. Most of these dyes are derived from 1,4-diaminoanthraquinone.

Examples are **262,** a blue dye [307], **263,** a reddish-blue dye [284], **264,** a greenish-blue dye [307], and compound **265,** a blue dye used for wet-spun fibers while they are still gels [280].

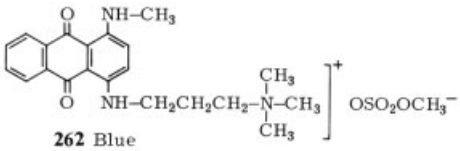

262 Blue

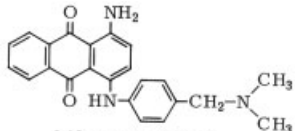

263 Reddish Blue

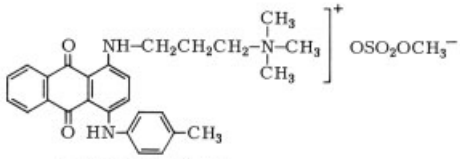

264 Greenish Blue

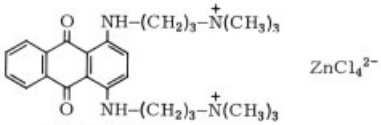

265 Blue

4. References

General References

1. *Ullmann,* 3rd ed., **vol. 3,** p. 662 – 732.
2. Bios Final Report 987.
3. Bios Final Report 1484.
4. Bios Final Report 1493.
5. Fiat Final Report 1313 II.
6. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. I – VIII,** Academic Press, New York – London 1952 – 1978.
7. J. Houben: *Das Anthracen und die Anthrachinone,* G. Thieme, Leipzig 1929.
8. N. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* Akademie-Verlag, Berlin 1966.
9. P. Rys, H. Zollinger: *Leitfaden der Farbstoffchemie,* Verlag Chemie, Weinheim 1970.
10. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer, Berlin – Göttingen – Heidelberg 1964.
11. Colour Index, 3rd ed., vol. 1 – 5, Soc. of Dyers & Colourists, Bradford/England 1971.
12. *Ullmann* 4th ed., **vol. 7,** p. 585 – 646.
13. *Houben-Weyl,* **4/7,** part 3 c.
14. P. F. Gordon, P. Gregory: *Organic Chemistry in Colour,* Springer, Berlin–Heidelberg–New York 1983.

Specific References

15. Bayer, DE 160104, 1903; *Friedländer,* vol. 8, p. 236.

16. Bayer, DE-OS 3106933, 1981 (K. Ebke, J. Ohm, J. Schroeder).
17. BASF, DE-OS 1593761, 1967 (H.-J. Nebel).
18. BASF, DE-OS 2102037, 1971 (H. Hiller, W. Jentzsch).
19. Bayer, DE-OS 2124261, 1971 (R. Schmitz).
20. Sumitomo Chem., DE-OS 2214948, 1971 (E. Hongo et al.).
21. Mitsubishi Chem., JP-KK 46-22335, 1969.
22. SU 138614, 1959 (W. A. Iwanowa et al.).
23. Bayer, DE-OS 2163674, 1971 (R. Schmitz, C. Wittig).
24. ICI, GB 1420191, 1971 (N. Ackerley, R. Price).
25. ICI, GB 1416678, 1971 (A. Bennie, R. T. Clarke, T. Hollis).
26. Bayer, DE-OS 2041547, 1970 (R. Schmitz, K. Alberti).
27. Agency of Ind. Sci. Tech., JP-Kokai 51-100064, 1975.
28. Agency of Ind. Sci. Tech., JP-Kokai 51-100063, 1975.
29. Reynolds Metals, US 2871244, 1955 (J. Kamlet).
30. Bayer, DE 228901, 1909; *Friedländer,* vol. 10, p. 578.
31. Bayer, DE-OS 2654650, 1976 (H. Seidler, N. Majer, H. Judat).
32. Bayer, DE-OS 2455587, 1974 (N. Majer et al.).
33. Bayer, DE-OS 2522177, 1975 (K.-J. Reubke).
34. Inst. f. org. Zwischenprod. u. Farbst., SU 178390, 1963 (N. S. Dokunichin et al.).
35. Bayer, DE-OS 2458022, 1974 (B. Thelen et al.).
36. Bayer, DE-OS 2452014, 1974 (N. Majer et al.).
37. ICI, DE-OS 2240518, 1971 (D. A. S. Phillips).
38. Ciba-Geigy, DE-OS 1768593, 1967 (M. Grelat).
39. Du Pont, US 3378572, 1964 (R. S. Wilder).
40. Du Pont, US 2417027, 1943 (V. Weinmayr).
41. IG-Farbenind., DE 677327, 1935.
42. Bayer, DE-OS 2720965, 1977 (H. Herzog, W. Hohmann, H. Seidler).
43. H. Schilling, *Chem. Ber.* **46** (1913) 1066.
44. AG für Anilin-Fabrikation, DE 269249, 1913; *Friedländer,* vol. 11, p. 548. AG für Anilin-Fabrikation, DE 293156, 1913; *Friedländer,* vol. 12, p. 410.
45. Du Pont, US 2033363, 1934 (J. M. Tinker, V. M. Weinmayr).
46. Sandoz, DE-OS 2162538, 1970 (I. Toth).
47. Sandoz, DE-OS 2227340, 1971 (I. Toth).
48. Bayer, DE-OS 2232464, 1972 (W. Auge, K.-W. Thiem, R. Neeff).
49. Bayer, DE-OS 2620486, 1976 (B. Thelen, W. Auge, K.-W. Thiem).
50. BASF, DE-OS 2039822, 1970 (E. Hartwig, O. Ackermann, H. Eilingsfeld).
51. Sumitomo Chem., DE-OS 2727587, 1976 (A. Fukasawa et al.).
52. Bayer, DE-OS 2646649, 1976 (W. Hohmann, K. Wunderlich, H. Seidler).
53. Sumitomo Chem., DE-OS 2751666, 1977 (A. Fukasawa, S. Masaki, N. Serizawa).
54. Bayer, DE-OS 2233185, 1972 (A. Vogel).
55. Iwaki Seiyaku, DE-OS 2419726, 1973 (T. Okada, K. Naito, T. Kikuchi).
56. Mitsui Toatsu Chem., JP-Kokai 54-19958, 1977.
57. Ciba-Geigy, DE-OS 2349753, 1972 (Z. Scha, T. Somlo).
58. Bayer, DE-OS 2219216, 1972 (E. Klauke, R. Schmitz, H.-S. Bien).
59. Sandoz, DE-OS 2103360, 1970 (W. Frey, I. Toth).
60. Sandoz, DE-OS 2206960, 1971 (F. Müller, R. Winkler).
61. BASF, DE-OS 2232446, 1972 (K.-H. Bantel, H. Eilingsfeld).
62. BASF, DE-OS 2233076, 1972 (K.-H. Bantel, H. Eilingsfeld, G. Stökelmann).
63. BASF, DE-OS 2449219, 1974 (G. Epple).
64. Bayer, DE-OS 2343978, 1973 (K.-W. Thiem, W. Auge, R. Neeff).
65. Bayer, DE-OS 2654649, 1976 (K. Wunderlich et al.).
66. BASF, DE-OS 2200088, 1972 (H. Eilingsfeld, O. Ackermann).
67. Bayer, DE-OS 2256644, 1972 (W. Auge et al.).
68. Sumitomo Chem., DE-OS 3029302, 1979 (M. Takahashi et al.).
69. Bayer, DE-OS 2351590, 1973 (W. Hohmann).
70. BASF, DE-OS 2853920, 1978 (S. Mensch, W. Elser).
71. Bayer, DE-OS 2637732, 1976 (W. Hohmann, K. Wunderlich).
72. BASF, DE-OS 2459164, 1974 (D. Lach, H. Eilingsfeld, G. Stöckelmann).
73. Mitsubishi Chem., JP-Kokai 54-100 362, 1978.
74. Bayer, DE-OS 2346317, 1973 (A. Vogel).
75. Inst. der Industrie für org. Chem., PL 52206, 1964 (S. Galinowski, S. Swiatly).
76. Bayer, DE-OS 2400164, 1974 (A. Vogel).
77. Bayer, DE-OS 2637733, 1976 (K. Wunderlich, W. Hohmann, H.-S. Bien).
78. BASF, DE-OS 2545699, 1975 (H. Bruenemann, H. Eilingsfeld, D. Lach).
79. Mitsui Toatsu Chem., JP-Kokai 53-44551, 1976.

80. BASF, DE-OS 2524747, 1975 (H. Eilingsfeld, D. Lach).
81. Sandoz, DE-OS 2222638, 1971 (R. Winkler).
82. ICI, DE-OS 2227766, 1971 (J. Cheetham).
83. Ciba-Geigy, DE-OS 2412171, 1973 (T. Somlo, J. Murphy).
84. Mitsui Toatsu Chem., DE-OS 2539631, 1974 (Y. Torisu, S. Kaba, K. Mukai).
85. Mitsui Toatsu Chem., DE-OS 2514445, 1974 (A. Iwamura et al.).
86. Nippon Kayaku, JP-KK 47-36855, 1968.
87. Sumitomo Chem., DE-OS 2920441, 1978 (A. Fukasawa et al.).
88. Bayer, DE-OS 1543605, 1966 (H. Pelster et al.).
89. BASF, DE-OS 2452413, 1974 (G. Epple).
90. BASF, DE-OS 2164458, 1971 (K.-H. Bantel, H. Eilingsfeld).
91. Sumitomo Chem., DE-OS 2715072, 1976 (M. Yoshimura et al.).
92. Mitsui Toatsu Chem., DE-OS 2557441, 1974 (Y. Hirai et al.).
93. Sumitomo Chem., JP-Kokai 50-49267, 1973.
94. Sandoz, DE-OS 2211411, 1971 (M. Aerberli, I. Toth).
95. Bayer, DE-OS 2314218, 1973 (K.-W. Thiem, W. Auge, R. Neeff).
96. Sumitomo Chem., JP-Kokai 51-32552, 1974.
97. Bayer, DE-OS 2526651, 1975 (K.-J. Reubke, H.-S. Bien).
98. Bayer, EP 49381, 1980 (K.-J. Reubke, J. Stawitz).
99. IG-Farbenind., DE 555937, 1930 (R. E. Schmidt, K. Bamberger); *Friedländer,* **vol. 19,** p. 2028.
100. Bayer, EP 81169, 1981 (J. Stawitz).
101. IG-Farbenind., DE 590163, 1930 (F. Wieners); *Friedländer,* **vol. 19,** p. 2045.
102. BASF, DE-AS 1205550, 1963 (W. Braun, M. Ruske).
103. Ciba-Geigy, BE 774901, 1971.
104. Bayer, DE-AS 1154490, 1962 (H. Vollmann, W. Hohmann, F. Baumann).
105. Scottish Dyes, GB 230116, 1923 (G. Beckett, J. Thomas).
106. Hoechst, DE 158076, 1900; *Friedländer,* vol. 7, p. 776.
107. Sandoz, DE-OS 1927785, 1968 (P. Buecheler).
108. Bayer, DE 288825, 1914; *Friedländer,* vol. 12, p. 414.
109. Sumitomo Chem., JP-Kokai 53-44550, 1976.
110. BASF, DE 252578, 1911; *Friedländer,* vol. 11, p. 545.
111. Mitsubishi Chem., JP-Kokai 51-23250, 1974.
112. SU 585156, 1976 (P. D. Jakuchny, T. I. Komarenko, D. F. Schurygina).
113. Mitsui Toatsu Chem., JP-Kokai 50-64257, 1973.
114. Mitsui Toatsu Chem., JP-Kokai 50-134026, 1974.
115. Kuhlmann, DE 586515, 1932; *Friedländer,* vol. 20, p. 1302.
116. Bayer, DE-OS 2827197, 1978 (R. Braden et al.).
117. Sandoz, DE-OS 2340114, 1972 (F. Krenmüller et al.).
118. Bayer, DE 144634, 1901; *Friedländer*, vol. 7, p. 201.
119. Ciba-Geigy, DE-OS 2541663, 1974 (Z. Seha).
120. Bayer, EP 63298, 1981 (K.-J. Reubke).
121. IG-Farbenind., DE 736901, 1937 (W. Zerweck, E. Heinrich).
122. H. Dreyfus, FR 602882, 1924.
123. Bayer, DE-AS 1150652, 1955 (G. Gehrke, L. Nüßler).
124. IG-Farbenind., DE 722593, 1933.
125. Allied Chem. & Dye Corp., US 2443899, 1943 (A. V. Erkkila, R. C. Hoare).
126. ICI, DE 511320, 1928; *Friedländer,* vol. 17, p. 1191.
127. Nat. Aniline, US 2207045, 1938 (R. S. Wilder).
128. Bayer, DE-OS 1644607, 1966 (K.-H. Peters, R. Neeff).
129. Bayer, DE 225232, 1908; *Friedländer,* vol. 9, p. 1197.
130. Du Pont, US 2346726, 1942 (E. C. Buxbaum).
131. R. Wedekind, GB 14476, 1911; *Friedländer,* vol. 11, p. 627.
132. Ciba, BE 719645, 1967.
133. IG-Farbenind., DE 623069, 1931 (K. Köberle); *Friedländer,* **vol. 21,** p. 1085.
134. Du Pont, US 3040063, 1960 (R. L. Walker).
135. IG-Farbenind., DE 518406, 1926 (W. Hartmann); *Friedländer,* **vol. 17,** p. 1211.
136. Bayer, DE-AS 1228274, 1965 (H. Leister, H. Vollmann, H.-S. Bien)
137. BASF, DE 158951, 1903; *Friedländer,* vol. 8, p. 277.
138. Ciba, BE 670204, 1964.
139. Ciba Geigy, DE-OS 2109058, 1970 (E. Mörgeli).
140. Bayer, DE-AS 1151517, 1962 (H. Vollmann, W. Hohmann, F. Baumann).
141. Ciba, DE-AS 1176668, 1962 (P. Sutter).
142. Du Pont, US 2716655, 1954 (S. N. Boyd).
143. Bayer, DE 267445, 1912; *Friedländer,* vol. 11, p. 562.
144. BASF, DE-AS 1221240, 1964 (H. Eilingsfeld).
145. Hoechst, DE-AS 1174925, 1961 (O. Fuchs, H. Rentél).

146. Nat. Aniline, US 2251688, 1938 (J. Ogilvie, R. X. Hoare).
147. Du Pont, US 2135346, 1937 (H. R. Lee, D. C. Klein).
148. IG-Farbenind., DE 484997, 1927 (K. Weinand); *Friedländer,* **vol. 16,** p. 1248.
149. Sumitomo Chem., JP-Kokai 58-131964, 1982.
150. Sumitomo Chem., JP-Kokai 57-77663, 1980.
151. ICI, GB 1291255, 1968 (J. H. Adam, D. N. Marsh).
152. Bayer, DE-OS 2740885, 1977 (R. Muders et al.).
153. Bayer, DE-AS 1226598, 1964 (F. Baumann, H. Vollmann, H.-J. Schulz).
154. Bayer, DE-OS 1768152, 1968 (K. Wunderlich, H.-S. Bien).
155. Bayer, DE-AS 1222051, 1965 (H.-J. Schulz, H.-S. Bien).
156. Sandoz, DE 631518, 1934; *Friedländer,* vol. 23, p. 952.
157. BASF, DE 106227, 1898; *Friedländer,* vol. 5, p. 307.
158. Nat. Aniline, US 2210517, 1937 (R. S. Wilder).
159. BASF, DE 108274, 1898; *Friedländer,* vol. 5, p. 311.
160. Bayer, GB 1014055, 1961 (J. Singer, H. W. Schwechten).
161. Sandoz, FR 1509724, 1966 (J. Günthard).
162. Eastman Kodak, BE 661209, 1960.
163. BASF, DE-AS 1227583, 1963 (W. Braun, I. Paetzke); BASF, DE-OS 1960100, 1969 (P. Dimroth, E. Schefczik); BASF, DE-OS 2060557, 1970 (P. Dimroth, G. Henning); GAF, GB 1062388, 1965.
164. IG-Farbenind., DE 503717, 1926 (K. Schirmacher, K. Zahn, H. Vollmann); *Friedländer,* **vol. 17,** p. 695.
165. Sandoz, DE-OS 2224793, 1971 (R. Winkler).
166. Kuhlmann, DE 583871, 1931; *Friedländer,* vol. 20, p. 1299.
167. BASF, DE-OS 1644439, 1965 (W. Braun, K. Maier).
168. Hoechst, FR 1363216, 1962.
169. K. Venkataraman, *Indian. J. Chem.* **9** (Oct. 1971) 1060 – 1063.
170. IG-Farbenind., DE 456235, 1925 (R. E. Schmidt); *Friedländer,* **vol. 15,** p. 674.
171. Bayer, BE 627010, 1962 (K. Wunderlich, H.-S. Bien).
172. ICI, GB 1471265, 1973 (R. T. Clarke, T. J. Smith, D. A. Stewart).
173. Bayer, DE-AS 1184879, 1962 (K. Klemm, G. Gehrke).
174. Bayer, DE-AS 1199279, 1963 (H.-S. Bien, W. Hohmann, H. Vollmann).
175. Bayer, DE-OS 1493739, 1965 (W. Hohmann, H. Vollmann, H.-S. Bien).
176. Bayer, DE-OS 1932646, 1969 (W. Hohmann, K. Wunderlich, H.-S. Bien).
177. Ciba, DE-AS 1134087, 1960 (P. Rhyner).
178. Sumitomo Chem., JP-KK 47-37252, 1970.
179. Bayer, DE-AS 1277475, 1963 (G. Gehrke).
180. BASF, DE-OS 1593780, 1967 (H.-J. Sturm, G. Steinhoff).
181. Bayer, EP 69910, 1981 (V. Hederich, G. Gehrke).
182. BASF, DE 186526, 1904; *Friedländer,* vol. 8, p. 237.
183. Ciba-Geigy, DE-OS 2758397, 1976 (M. Grélat).
184. Bayer, DE-OS 2830554, 1978 (R. Schmitz).
185. Mitsui Toatsu, JP-Kokai 54-22357, 1977.
186. Sumitomo Chem., JP-Kokai 49-5429, 1972.
187. BASF, DE-AS 1165180, 1961 (K. Scherf).
188. Bayer, DE 81960, 1893; *Friedländer,* vol. 4, p. 274.
189. Bayer, DE 161026, 1904; *Friedländer,* vol. 8, p. 256.
190. Bayer, DE 156762, 1903; *Friedländer,* vol. 8, p. 240.
191. Bayer, DE-OS 2903851, 1979 (H. Seidler, G. Gehrke).
192. Sumitomo Chem., JP-Kokai 50-131962, 1974.
193. Mitsubishi Chem. Ind., DE-OS 2607036, 1975 (Y. Kimura et al.).
194. Toms River Chem., US 4255342, 1979 (A. D. Olin).
195. Mitsubishi Chem. Ind., JP-Kokai 54-3051, 1977.
196. Nippon Kayaku, JP-Kokai 53-37650, 1976.
197. Bayer, DE-OS 2855939, 1978 (H. Herzog, G. Gehrke).
198. ACNA, FR 1352537, 1962 (A. Crotti, P. Mezzacappa).
199. ICI, DE 568311, 1929; *Friedländer,* vol. 19, p. 1926.
200. Bayer, DE 163042, 1904; *Friedländer,* vol. 8, p. 268.
201. Bayer, DE-OS 1543619, 1966 (R. Schmitz, H. Leister, H.-S. Bien).
202. Bayer, DE-OS 2 909481, 1979 (W. Steinbeck, G. Gehrke).
203. IG-Farbenind., DE 568760, 1925 (R. E. Schmidt); *Friedländer,* **vol. 18,** p. 1277.
204. IG-Farbenind., DE 554647, 1930 (R. E. Schmidt, K. Bamberger); *Friedländer,* **vol. 19,** p. 1949.
205. Bayer, DE-OS 1543605, 1966 (H. Pelster et al.).
206. Mitsui Toatsu, JP-Kokai 55-33407, 1978.

207. Ciba, DE 589074, 1931; *Friedländer,* vol. 19, p. 2001.
208. Bayer, DE 100138, 1897; *Friedländer,* vol. 5, p. 245.
209. Du Pont, US 2419405, 1943 (D. X. Klein).
210. Bayer, FR 1445843, 1964 (W. Hohmann, H.-S. Bien).
211. Bayer, FR 1468502, 1965 (W. Hohmann et al.).
212. Ciba, CH 146772, 1928.
213. IG-Farbenind., DE 445846, 1925 (C. Weinand); *Friedländer,* **vol. 15,** p. 677.
214. Bayer, DE 238488, 1910; *Friedländer,* vol. 10, p. 646.
215. Bayer, DE-OS 1939095, 1969 (P. Wegner et al.).
216. Ciba, GB 1085685, 1965.
217. Nippon Kayaku, JP-Kokai 49-17425, 1972.
218. Ciba-Geigy, EP 71576, 1981 (P. Kniel).
219. IG-Farbenind., DE 445269, 1925 (R. E. Schmidt, A. Jacobi); *Friedländer,* **vol. 15,** p. 671.
220. Bayer, DE-AS 1228734, 1963 (K. Klemm, G. Gehrke).
221. Bayer, GB 1029448, 1964 (W. Hohmann).
222. BASF, DE-AS 2713575, 1977 (J. Redeker, H. Hiller, E. Spohler).
223. Sumitomo Chem., DE-OS 2817890, 1977 (M. Nishikuri, A. Takeshita, H. Kenmochi).
224. Mitsui Toatsu Chem., JP-Kokai 56-123955, 1980.
225. Ciba, DE-AS 1065959, 1954 (P. Grossmann, W. Jenny, W. Kern).
226. Österr. Alizarinfabr.-Ges., DE 3565, 1878; *Friedländer,* vol. 1, p. 310.
227. American Aniline Products, US 3389151, 1966 (D. A. Zanella).
228. Bayer, DE 96364, 1897; *Friedländer,* vol. 5, p. 246.
229. Bayer, DE 100136, 1897; *Friedländer,* vol. 5, p. 247.
230. IG-Farbenind, DE 632 911, 1932 (K. Weinand, C. Bamberger); *Friedländer,* **vol. 21,** p. 1038.
231. Ciba, DE 841313, 1944 (P. Grossmann).
232. Chem. Fabrik Griesheim-Elektron, DE 290084, 1914; *Friedländer,* vol. 12, p. 439.
233. Bayer, DE 208640, 1907; *Friedländer,* vol. 9, p. 701.
234. Geigy, FR 1522933, 1966.
235. Bayer, DE-AS 1180473, 1962 (H.-S. Bien, K. Wunderlich, F. Baumann).
236. Bayer, DE-AS 1201933, 1962 (H.-S. Bien, K. Wunderlich, F. Baumann).
237. ICI, DE-OS 2734828, 1976 (A. J. Logan, R. W. Kenyon).
238. IG-Farbenind., DE 651104, 1935 (W. Zerweck, W. Kunze); *Friedländer,* **vol. 24,** p. 991.
239. Celanese, US 2640059, 1949 (V. S. Salvin, E. F. Landau).
240. Bayer, DE-AS 1190123, 1962 (H.-S. Bien, K. Wunderlich, F. Baumann).
241. Allied Chem. Corp., US 3018154, 1959 (J. F. Downey, R. C. Hoare).
242. Bayer, DE-AS 1226598, 1964 (F. Baumann, H. Vollmann, H.-J. Schulz).
243. Nippon Kayaku, JP-KK 70-7037, 1967.
244. Bayer, DE-OS 1644578 (Example 104), 1964 (R. Neeff et al.).
245. Bayer, DE-OS 1293363, 1964 (R. Neeff et al.).
246. Mitsubishi Chem., JP-Kokai 55-73735, 1978.
247. Hoechst, BE 647171, 1963.
248. AG für Anilin-Fabrikation, DE 267081, 1912; *Friedländer,* vol. 11, p. 593.
249. Hoechst, DE 361043, 1920; *Friedländer,* vol. 14, p. 858.
250. IG-Farbenind., DE 533249, 1926 (K. Wilke); *Friedländer,* **vol. 18,** p. 1245.
251. Cassella, DE 346188, 1915; *Friedländer,* vol. 13, p. 395.
252. Bayer, DE 938435, 1953.
253. BASF, DE-AS 1918696, 1969 (E. Hartwig).
254. Mitsubishi Chem. Ind., JP-KK 47-26413, 1968.
255. Bayer, GB 879240, 1957.
256. BASF, DE-OS 1668870, 1968 (K. Maier).
257. American Cyanamid, US 2499003, 1946 (M. Scalera).
258. BASF, DE 229394, 1909; *Friedländer,* vol. 10, p. 601.
259. BASF, DE-AS 1108704, 1959 (W. Braun, M. Ruske).
260. Sumitomo Chem., JP-Kokai 57-176944, 1981.
261. Bayer, DE 935669, 1953.
262. Bayer, DE-OS 2931981, 1979 (F. W. Kröck, R. Neeff, H. Scheiter).
263. IG-Farbenind., DE 464863, 1926 (K. Wilke); *Friedländer,* **vol. 16,** p. 1229.
264. BASF, DE-AS 1025079, 1955 (E. Anton, K. Saftien).
265. BASF, DE-AS 1250031, 1963 (E. Hartwig, W. Braun).
266. IG-Farbenind., DE 615756, 1934 (G. Kränzlein, M. Corell, W. Schaich); *Friedländer,* **vol. 22,** p.1078.
267. BASF, DE-AS 1100209, 1958 (H. Geeren, F. Ebel, W. Braun).
268. IG-Farbenind., DE 475687, 1926 (M. A. Kunz, G. v. Rosenberg, E. Goffarjé); *Friedländer,* **vol. 16,** p. 1341.
269. IG-Farbenind., DE 604279, 1932 (P. Nawiasky et al.); *Friedländer,* **vol. 21,** p. 1091.

270. IG-Farbenind., DE 623028, 1934 (E. Berthold, J. Müller); *Friedländer,* **vol. 22,** p. 1081.
271. IG-Farbenind., DE 692750, 1938 (H. Schlichenmaier, L. Berlin, E. Berthold).
272. BASF, DE 421236, 1922 (A. Lüttringhaus, F. Kačev); *Friedländer,* **vol. 15,** p. 689.
273. AG für Anilin-Fabrikation, DE 232711, 1910; *Friedländer,* vol. 10, p. 731.
274. J. Arient, V. Slavik, *Collect. Czech. Chem. Commun.* **34** (1969) 3576.
275. Sandoz, DE 593867, 1932; *Friedländer,* vol. 20, p. 1344.
276. Bayer, DE 239544, 1910; *Friedländer,* vol. 10, p. 638.
277. Du Pont, US 2258394, 1939 (J. M. Tinker et al.).
278. Hoechst, DE 251350, 1911; *Friedländer,* vol. 11, p. 618.
279. A. K. Wick, *Helv. Chim. Acta* **54** (1971) 769.
280. ICI, GB 2048963, 1979 (P. Gregory, M. Yelland).
281. IG-Farbenind., DE 746546, 1937 (F. Wieners, W. Mieg); *Friedländer,* Suppl. vol. I/2, p. 392.
282. Ciba-Geiby, CH 528576, 1969 (A. Wick).
283. Kuhlmann, DE 566708, 1930; *Friedländer,* vol. 19, p. 2120.
284. Bayer, DE 963502, 1955 (H. W. Schwechten, O. Bayer).
285. IG-Farbenind., DE 696370, 1936 (W. Burneleit, W. Mieg, F. Wieners).
286. Bayer, DE 630218, 1934 (W. Mieg, F. Wieners); *Friedländer,* **vol. 23,** p. 1020.
287. IG-Farbenind., DE 451495, 1925 (W. Mieg); *Friedländer,* **vol. 16,** p. 1345.
288. Ciba, DE-OS 1813729, 1967 (A. Wick).
289. Kuhlmann, FR 733440, 1931.
290. Výzkumný ústav organ. syntéz, DE-OS 1817388, 1968 (J. Arient et al.).
291. Ciba-Geigy, EP 49873, 1980 (Z. Seha).
292. Hoechst, DE 208969, 1908; *Friedländer,* vol. 9, p. 776.
293. Ciba, CH 191015, 1936.
294. IG-Farbenind., DE 525666, 1929 (W. Bruck); *Friedländer,* **vol. 18,** p. 1310.
295. IG-Farbenind., DE 501746, 1929 (W. Bruck); *Friedländer,* **vol. 17,** p. 1229.
296. BASF, DE 246966, 1911; *Friedländer,* vol. 10, p. 725.
297. IG-Farbenind., DE 560236, 1929 (M. A. Kunz, E. Berthold, K. Köberle); *Friedländer,* **vol. 19,** p. 2088.
298. Bayer, DE 678499, 1937 (W. Bauer); *Friedländer,* Suppl. vol. I/2, p. 382.
299. BASF, DE 237236, 1910; *Friedländer,* vol. 10, p. 708.
300. American Cyanamid, US 2492802, 1945 (H. Z. Lecher, W. S. Forster).
301. IG-Farbenind., DE 590579, 1932 (E. Kramer); *Friedländer,* **vol. 20,** p. 1293.
302. BASF, DE 258561, 1910; *Friedländer,* vol. 11, p. 671.
303. IG-Farbenind., DE 555967, 1929 (M. A. Kunz, E. Berthold, K. Köberle); *Friedländer,* **vol. 19,** p. 2082.
304. BASF, DE 279867, 1913; *Friedländer,* vol. 12, p. 445.
305. IG-Farbenind., DE 661152, 1936 (F. Baumann, H. W. Schwechten); *Friedländer,* **vol. 25,** p. 773.
306. IG-Farbenind., DE 470809, 1926 (K. Schirmacher, W. Schaich, A. Wolfram); *Friedländer,* **vol. 16,** p. 1316.
307. Bayer, DE-AS 1150652, 1955 (G. Gehrke, L. Nüßler).
308. Ciba, DE-AS 1060526, 1953 (P. Sutter, W. Fioroni).
309. Bayer, DE 904926 (Examples 5, 6), 1951 (H. Thielert, F. Baumann).
310. Cassella, DE-AS 1215284, 1962 (W. Zerweck, E. Schwamberger).
311. Cassella, DE-AS 1264648, 1963 (W. Zerweck, E. Schwamberger).
312. Bayer, DE 158287, 1903; *Friedländer,* vol. 8, p. 341.
313. Bayer, DE 193121, 1906; *Friedländer,* vol. 9, p. 783.
314. Ciba, DE 423311, 1922; *Friedländer,* vol. 15, p. 714.
315. IG-Farbenind., DE 580013, 1932; *Friedländer,* vol. 20, p. 1377.
316. BASF, DE 212471, 1908; *Friedländer,* vol. 9, p. 834.
317. IG-Farbenind., DE 490723, 1926 (K. Wilke); *Friedländer,* **vol. 16,** p. 1368.
318. IG-Farbenind., DE 483154, 1927 (A. Lüttringhaus, P. Nawiasky, A. Ehrhardt); *Friedländer,* **vol. 16,** p. 1490.
319. IG-Farbenind., DE 722868, 1937 (H. Scheyer, H. Ritter); *Friedländer,* Suppl. vol. I/2, p. 451.
320. Du Pont, GB 364042, 1930.
321. Ciba-Geigy, EP 10525, 1978 (H. Jäger).
322. SU 401130, 1971 (E. M. Natanson et al.)
323. Bayer, DE-OS 2823160, 1978 (J. Schroeder).
324. Ciba-Geigy, EP 22062, 1979 (J. Bersier, H. Jäger, H. Schwander).
325. Ciba-Geigy, EP 60437, 1981 (H. Jäger et al.).
326. Bayer, DE-OS 2830456, 1978 (B. Schroeder, R. Neeff, R. Braden).
327. BASF, DE 200335, 1905; *Friedländer,* vol. 9, p. 817.

328. BASF, DE-OS 2631853, 1976 (A. Schumacher, K. E. Kling).
329. Du Pont, US 2353049, 1942 (H. R. Lee, C. F. Belcher).
330. Kalle, DE 467118, 1924 (M. P. Schmidt, W. Neugebauer); *Friedländer,* **vol. 16,** p. 1453.
331. Hoechst, DE 420412, 1923 (A. Wolfram); *Friedländer,* **vol. 15,** p. 736.
332. Bayer, DE 200014, 1907; *Friedländer,* vol. 9, p. 760.
333. Bayer, DE 171293, 1904; *Friedländer,* vol. 8, p. 304.
334. Chem. Fabrik Griesheim-Elektron, DE 255641, 1912; *Friedländer,* vol. 11, p. 583.
335. Bayer, DE 220314, 1908; *Friedländer,* vol. 9, p. 742.
336. IG-Farbenind., DE 711775, 1938 (E. Berthold, W. Roland); *Friedländer,* Suppl. vol. I/2, p. 280.
337. IG-Farbenind., DE 633207, 1931 (M. A. Kunz, K. Köberle); *Friedländer,* **vol. 21,** p. 1143.
338. BASF, DE-AS 1159456, 1960 (H. Weidinger, H. Eilingsfeld, G. Haese).
339. Kawasaki Kasei, JP-Kokai 52-95678, 1976.
340. BASF, DE-AS 1215843, 1962 (M. Bertl, F. Graser).
341. BASF, DE-OS 2300019, 1973 (H. Hiller, W. Jentzsch, A. Schuhmacher).
342. Bayer, DE 203752, 1907; *Friedländer,* vol. 9, p. 735.
343. BASF, DE 216597, 1907; *Friedländer,* vol. 9, p. 738.
344. Bayer, DE 201904, 1907; *Friedländer,* vol. 9, p. 736.
345. Bayer, DE 233126, 1909; *Friedländer,* vol. 10, p. 609.
346. Sandoz, DE 578995, 1930; *Friedländer,* vol. 19, p.1964.
347. IG-Farbenind., DE 518316, 1927 (G. Kränzlein, H. Vollmann); *Friedländer,* **vol. 17,** p. 1301.
348. IG-Farbenind., DE 576253, 1927 (R. Sedlmayer, W. Eckert); *Friedländer,* **vol. 18,** p. 592.
349. Hoechst, DE 412053, 1922 (G. Kränzlein, M. Corell, R. Sedlmayr); *Friedländer,* **vol. 15,** p. 731.
350. Hoechst, DE 423720, 1924 (G. Kränzlein, R. Sedlmayr); *Friedländer,* **vol. 15,** p. 733.
351. L. Kalb, DE 280787, 1913; *Friedländer,* vol. 12, p. 498.
352. Cassella, DE 445390, 1925 (R. Herz, W. Zerweck); *Friedländer,* **vol. 15,** p. 300.
353. Cassella, DE 458598, 1925 (R. Herz, W. Zerweck); *Friedländer,* **vol. 16,** p. 1410.
354. IG-Farbenind., DE 485961, 1927 (R. Heidenreich); *Friedländer,* **vol. 16,** p. 1422.
355. BASF, DE-OS 1951708, 1969 (G. Bock).
356. BASF, DE-OS 2115093, 1971 (F. Graser).
357. Du Pont, US 2872459, 1956 (A. A. Baum).
358. GAF, DE-AS 1068687, 1956 (T. A. Martin, D. I. Randall, J. Taras).
359. Du Pont, US 2388743, 1944 (E. T. Howell).
360. IG-Farbenind., DE 450999, 1925 (A. Lüttringhaus, H. Neresheimer, H. J. Emmer); *Friedländer,* **vol. 15,** p. 721.
361. IG-Farbenind., DE 608442, 1933 (H. Wolff, W. Mieg); *Friedländer,* **vol. 21,** p. 1130.
362. IG-Farbenind., DE 595461, 1929 (M. A. Kunz, K. Köberle, E. Berthold); *Friedländer,* **vol. 20,** p. 1370.
363. BASF, DE 185222, 1904; *Friedländer,* vol. 9, p. 830.
364. Scottish Dyes, DE 417068, 1920; *Friedländer,* vol. 15, p. 760.
365. IG-Farbenind., DE 448262, 1924 (O. Braunsdorf, P. Nawiasky, E. Holzapfel); *Friedländer,* **vol. 15,** p. 728.
366. BASF, DE-AS 2704964, 1977 (W. S. Schweckendiek, A. Schuhmacher, H. Hiller).
367. BASF, DE 217570, 1909; *Friedländer,* vol. 9, p. 827.
368. IG-Farbenind., DE 453768, 1925 (R. Berliner, B. Stein, W. Trautner); *Friedländer,* **vol. 16,** p. 1283. IG-Farbenind., DE 470501, 1926 (H. Scheyer); *Friedländer,* **vol. 16,** p. 1283.
369. IG-Farbenind., DE 550712, 1930 (H. Scheyer); *Friedländer,* **vol. 19,** p. 2149.
370. IG-Farbenind., DE 576466, 1931 (H. Scheyer); *Friedländer,* **vol. 20,** p. 1426.
371. H. Labhart, *Helv. Chim. Acta* **40** (1957) 1410.
372. BASF, DE-AS 1209680, 1962 (K. Maier).
373. Bayer, DE-OS 2531557, 1975 (V. Hederich, H.-S. Bien, G. Gehrke).
374. Bayer, FR 1503492, 1965.
375. Sumitomo Chem., FR 1497689, 1965.
376. Bayer, DE-AS 1105837, 1958 (G. Gehrke).
377. Bayer, DE-AS 1644587, 1965 (M. Groll, K. Wunderlich, H.-S. Bien).
378. Du Pont, US 2628963, 1951 (J. F. Laucius, S. B. Speck).
379. BASF, DE-AS 1918696, 1969 (E. Hartwig).
380. Geigy, BE 650734, 1963.
381. Ciba, DE-AS 1065959, 1954 (P. Grossmann, W. Jenny, W. Kern).
382. Ciba, FR 1594324, 1968.
383. Sandoz, FR 1490805, 1965 (M. F. Müller).
384. Du Pont, DE-AS 1811796, 1968 (J. Blackwell, R. E. Starn, W. H. Gumprecht).

385. BASF, DE-OS 2524243, 1975 (H. Schwab, K. Oppenlaender, A. Blum).
386. BASF, DE-OS 2654434, 1976 (W. Elser, H. Eilingsfeld, G. Meyer).
387. BASF, DE-OS 2651975, 1976 (G. Epple, W. Elser).
388. Bayer, DE-OS 2918881, 1979 (W. Harms et al.).
389. Mitsubishi Chem. Ind., DE-OS 3218957, 1981 (T. Niwa, T. Hihara).
390. Mitsubishi Chem. Ind., JP-Kokai 55-164250, 1979.
391. Bayer, DE 849158, 1944 (K. Bähr).
392. BASF, DE-OS 1644482, 1967 (F. Graser, G. Riedel).
393. Bayer, EP 93367, 1982 (M. Blunck et al.).
394. IG-Farbenind., DE 653386, 1935 (W. Mieg, F. Wieners); *Friedländer,* **vol. 24,** p. 862.
395. IG-Farbenind., DE 696423, 1935 (K. Köberle); *Friedländer,* Suppl. vol. I/2, p. 239.
396. Bayer, DE 911493, 1952 (H. W. Schwechten).
397. Bayer, BE 729177, 1968 (K. Wunderlich, H.-S. Bien).
398. *Ullmann,* 4th ed., **vol. 18,** p. 661.
399. BASF, DE-AS 1544372, DE-AS 1544374, 1965 (W. Braun, I. Peatzke).
400. Ciba, DE-AS 1283542, 1961 (E. Mörgeli).
401. Sandoz, US 2121928, 1934 (A. Peter).
402. Ciba-Geigy, DE-OS 2850996, 1977 (R. Lacroix, J.-M. Adam, J. Vincze).
403. Sandoz, CH 214909, 1939.
404. Bayer, GB 1099756, 1966 (J. Singer).
405. Geigy, BE 713252, 1967.
406. Sandoz, FR 1455722, 1964 (M. J. Günthard).
407. Geigy, FR 1461074, 1964.
408. Crompton & Knowles, FR 1478769, 1965 (M. R. W. Eltonhead).
409. Ciba, DE-OS 1925515, 1968 (K. Seitz, G. Klahre, H. U. Schütz).
410. Bayer, DE-OS 1644204, 1967 (H.-S. Bien, E. Klauke).
411. Bayer, DE-AS 1264647, 1961 (J. Singer, H. W. Schwechten).
412. Bayer, DE-OS 2652119, 1976 (W. Harms, K. Wunderlich, K. v. Oertzen).
413. Bayer, DE-OS 3142852, 1981 (K. v. Oertzen).
414. Scottish Dyes, FR 543910, 1921.

Anthraquinonesulfonic Acids → Anthraquinone Dyes and Intermediates

Antiallergic Agents

Norbert Höfgen, Arzneimittelwerk Dresden GmbH, Radebeul, Federal Republic of Germany

Sonja Beckh, Medizinische Klinik 3, Pneumologie, Klinikum Nürnberg Nord, Nürnberg, Federal Republic of Germany

Istvan Szelenyi, Arzneimittelwerk Dresden GmbH, Radebeul, Federal Republic of Germany

Pal L. Bölcskei, Semmelweis Medical University, Budapest, Hungary

1. Introduction 2
2. Antihistamines 3
2.1. First Generation Antihistamines . . 3
2.1.1. Ethylenediamines 4
2.1.2. Ethanolamine Derivatives 6
2.1.3. Alkylamines 7
2.1.4. Piperazines 9
2.1.5. Piperidines 11
2.1.6. Phenothiazines 11
2.1.7. Other Tricyclic Systems 13
2.2. Second-Generation Antihistamines 15
2.2.1. Ethanolamines 15
2.2.2. Alkylamines 15
2.2.3. Piperazines 15
2.2.4. Piperidines 16
2.2.5. Phthalazinones 18
2.2.6. Others 18
2.2.7. Combinations of Antihistamines and Sympathomimetic Agents 18
2.3. Inhibitors of Histamine Synthesis . 19
3. Mast Cell Stabilizing Agents 19
4. Allergen Preparations for Desensitization 20
4.1. Properties of Allergens 20
4.2. Raw Materials 22
4.3. Allergen Standardization 22
4.4. Allergen Vaccines for Immunotherapy 24
4.5. Storage 24
4.6. Uses 24
4.6.1. Diagnostic Use 25
4.6.2. Therapeutic Use—Immunotherapy . . 25
4.6.3. Legal Aspects, Quality Requirements, and Safety 26
5. Anti-IgE-Omalizumab 26
5.1. IgE and its Role for the Allergic Response 27
5.2. IgE-Mediated Allergic Diseases . . . 27
5.3. Properties of Omalizumab 28
5.3.1. Chemistry and Production of Omalizumab 28
5.3.2. Pharmacology and Effect of Omalizumab 30
5.3.3. Pharmacokinetics 30
5.4. Indication for the Teatment with Omalizumab, Guidelines for Use . . 30
5.5. Application, Dosage, Costs 31
5.6. Side Effects 31
5.7. Legal Aspects, Safety 32
5.8. Possible Applications in the Future 32
6. Acknowledgement 32
7. References 32

Abstract

Allergic reactions cause a number of acute and chronic diseases. The *antigen* or *allergen* causing the allergic reaction is usually a protein, a polysaccharide, or a low molecular mass compound (hapten) bound to an endogenous protein. Antigens induce the synthesis of antibodies, which are serum proteins, referred to as immunoglobulins (Igs). Antigen cross-linking of the IgE molecules leads to cellular responses involving release of preformed mediators, e.g., histamine. *Antihistamines* or histamine H_1 receptor antagonists are used in the treatment of allergic diseases. In contrast to the first-generation or classical antihistamines the second-generation antihistamines do not have the side effects sedation, fatigue, and drowsiness. Mast cells are the predominant storage site for histamine in most tissues. Cromolyn sodium or disodium cromoglycate (DSCG) probably

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_419.pub2

acts on certain types of chloride channels expressed in mast cells and sensory nerves.

Allergen immunotherapy is a well established and accepted treatment of allergic diseases. Immunogenic tolerance is achieved by applying gradually increasing quantities of an allergen vaccine to the sensitized patient. Allergen preparations derive from reference standards containing defined amounts of relevant allergens. For efficient and well-tolerated therapy extracts are physically or chemically modified or conjugated. Recombinant DNA technology offers a new standard of qualitiy by selecting the disease causing major allergens.

The development of an anti-IgE antibody (omalizumab, Xolair^{TM}) offers a unique and novel treatment not only for therapeutic but also for preventive applications in IgE-mediated diseases. Omalizumab is a DNA-derived monoclonal humanized antibody which is well tolerated by the immune system. Omalizumab binds in form of complexes to free IgE in the serum and interacts with any kind of IgE, but not with other immunoglobulines. Omalizumab leads to a downregulation of IgE receptors on mast cells and tissue cells in target organs. The substance showed immunomodulatory effects on dendritic cells and reduced significantly the number of eosinophils. Effects of treatment on allergic rhinitis and asthma have been evaluated in several studies.

1. Introduction

Allergic diseases are worldwide on the increase. Studies in several countries show that about 15 – 20 % of all adults suffer from one of the many different allergic disorders. Allergic reactions cause a number of acute and chronic diseases, such as conjunctivitis, rhinitis, dermatitis, gastroenteritis, asthma, and anaphylactic shock. Allergic responses are immediate reactions to specific foreign substances to which individuals have developed hypersensitivity. Sensitizing antigens come from a variety of environmental sources: they can be airborne (e.g., pollens, molds, animals dander and urine, house dust mites), taken up orally (e.g., foods, drugs) or be injected (insect bites, drugs). The *antigen* or *allergen* is usually a protein, a polysaccharide or a low molecular mass compound (hapten) bound to an endogenous protein to form an antigenic complex. Antigens induce the synthesis of antibodies. In sensitized individuals, the subsequent exposure to the antigen results in the initiation of a series of cellular reactions that lead to the symptoms observed in allergy. Antibodies are serum proteins, referred to as immunoglobulins (Ig) such as IgG, IgM, IgA, and IgE. Allergic responses have been divided into four general categories [6]. Type I, or *anaphylactic reactions* are mediated by IgE antibodies (reagin). The main targets of this type of reaction are the skin (urticaria, atopic dermatitis), the gastrointestinal tract (food allergies), the eyes (conjunctivitis), the respiratory tract (rhinitis, asthma) and the vascular system (anaphylactic shock). Type II *or cytolytic reactions* are mediated by IgG and IgM (mainly blood transfusion- and drug-induced hemolytic reactions). Type III or *Arthus phenomenon* is mediated by IgG (e.g., serum sickness); the type IV or *delayed-hypersensitivity reactions* are mediated by macrophages and T lymphocytes (e.g., contact dermatitis).

Allergy, in a strict sense, is an IgE-mediated inflammation. As a part of an allergic response to an antigen, reaginic antibodies (IgE) are generated and bound to the surface of mast cells and basophils via high affinity Fc receptors (FcεRI) that are specific for IgE [7]. Antigen cross-linking of the IgE-molecules leads to cellular responses involving release of preformed mediators (e.g., histamine), lipid mediator formation and release, and cytokin generation. Mast cells with their mediators can be regarded as central to the initiation and mediation of the early phase of allergic inflammation and may also be responsible for the initiation of chronic allergic inflammation. Recent investigations have shown that many other cells with their mediators, first of all, eosinophils and lymphocytes are also involved in the maintenance and progress of allergic inflammation. Histamine, lipid mediators, bradykinin, enzymes, and cytokins are among the many important chemical mediators characterized so far.

This article discusses drugs for the treatment of allergic disorders already available on the market or undergoing advanced clinical trials.

2. Antihistamines

Histamine [*51-45-6*], a potent effector agent in allergic diseases, was first synthesized in 1907 by WINDAUS and VOGT [8] without being aware of its pharmacological activities. DALE and LAIDLAW [9] suggested in 1910 that histamine might be responsible for anaphylaxis, and that its pharmacological activity resembled that of many tissue extracts. In 1927 BEST and coworkers [10] were able to show that histamine was present in fresh samples of liver and lung. This was followed by demonstration of its presence in a variety of other tissues, hence the name *histamine* after the Greek word for tissues, *histos*. Every mammalian tissue that contains histamine is capable of synthesizing it from histidine by action of L-histidine decarboxylase. Mast cells are the predominant storage site for histamine in most tissues, in the blood, histamine is stored in the basophil granulocytes. These cells synthesize histamine and store it in their secretory granules. Once released, histamine is metabolized via two major pathways: (1) methylation of the ring nitrogen in position 1, catalyzed by histamine-*N*-methyl-transferase; and (2) oxidative deamination by diamine oxidase.

Three types of histamine receptors exist: H_1, H_2, and H_3. Symptoms of allergy and anaphylaxis are mainly mediated via stimulation of histamine H_1 receptors. Other effects, notably gastric acid secretion and increase in cardiac rate, are the result of H_2 receptor activation. Some responses, such as the hypotension that results from vasodilatation, are mediated by both H_1 and H_2 receptors. H_3 receptors are presynaptic receptors on histaminergic nerve terminals. *Antihistamines* or histamine H_1 receptor antagonists are used in the treatment of allergic diseases. Antihistamines can be divided into two groups: the so-called classical or sedating and the modern, nonsedating antihistamines. All first-generation or classical antihistamines show the major side effects sedation, fatigue, and drowsiness to a varying extent. This can be put down to the fact that they interact with the H_1 receptors in the brain. Other side effects that apparently are caused by the antimuscarinic action of some of the first-generation antihistamines include dryness of the mouth, diplopia, and dysuria. All these effects are not observed with the modern, the second-generation H_1 receptor antagonists [11]. Following the introduction of the nonsedating antihistamines, sedating histamine H_1 receptor antagonists have been replaced almost completely by the new antihistamines. Members of the first generation of antihistamines have nowadays only a relatively limited therapeutic value. They are often given as multi-ingredient preparations for the treatment of cough and common cold. In some cases, an attempt has been made to find new indications for the first-generation antihistamines.

2.1. First Generation Antihistamines

Research on antihistamines was initiated in 1933 by FOURNEAU and BOVET [12], who observed that 2-(*N*-piperidinomethyl)-1,4-benzodioxane protected animals from histamine-induced bronchoconstriction.

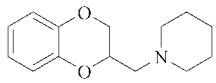

The first recorded evidence of a histamine antagonist originates from BOVET and STAUB [13] from the Pasteur Institute who noticed that a compound designated F 929, (5-isopropyl-2-methyl-phenoxyethyl)-diethylamine, attenuated some of the actions of histamine.

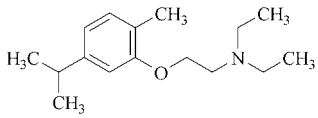

This compound protected guinea pigs against lethal doses of histamine, and lessened the symptoms of anaphylactic shock. Isosteric replacement of the ether oxygen by an amino group led to F 1571.

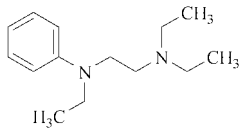

This first ethylenediamine derivative was synthesized by FOURNEAU and its pharmacological actions were described by STAUB [14] in 1937,

who concluded that an ethylenediamine moiety is probably the basic structure of antihistamines. In fact, most H_1 antagonists have a tertiary amino group linked by a two or three atom chain to two aromatic substituents and conform to the general formula:

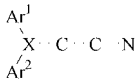

where Ar is aryl and X is nitrogen or a –C–O– ether linkage to the β-aminoethyl side chain [15, 16]. Sometimes the two aromatic rings are bridged, as in the tricyclic derivatives, or the ethylamine group may be part of a ring structure. Other variations also are possible [17].

2.1.1. Ethylenediamines

In 1942, phenbenzamine was introduced as the first clinically useful antihistamine [18].

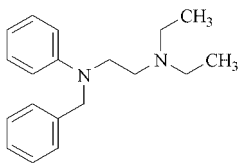

Phenbenzamine

Synthesis of Ethylenediamines. Three different methods of synthesizing ethylenediamines are widely used [19]:

1) Reaction of tertiary chloroethylamines with secondary amines in the presence of a base, e.g., NaH, $NaNH_2$, $NaOC_2H_5$, or KOH.
2) Alkylation of secondary amines, such as *N*-(2-dimethylaminoethyl)aniline with benzyl chlorides in the presence of a strong base.
3) Reaction of reactive heterocyclic halogen compounds, such as 2-bromopyridine with secondary amines, such as *N*-(2-dimethylaminoethyl)-aniline, also in the presence of a strong base.

Antazoline [*91-75-8*], 4,5-dihydro-*N*-phenyl-*N*-(phenylmethyl)-1*H*-imidazole-2-methanamine, $C_{17}H_{19}N_3$, M_r 165.3. Hydrochloride [*2508-72-7*], *mp* 237 – 241 °C; sulfate [*24359-81-7*].

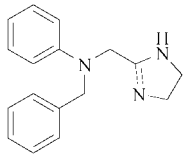

Antazoline is used for local application to the eye in allergic conjunctivitis.

Preparation: [20].

Trade name: Antistin (CibaVision); often combined with vasoconstrictor agents such as tetryzoline.

Chlorpyramine [*59-32-5*], *N*-[(4-chlorophenyl)methyl]-*N′*,*N′*-dimethyl-*N*-2-pyridinyl-1,2-ethanediamine, $C_{16}H_{20}ClN_3$, M_r 289.8. Hydrochloride [*6170-42-9*], *mp* 172 – 174 °C.

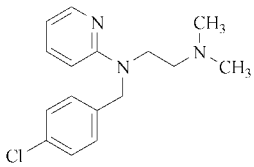

Preparation: [21].

Trade names: Synopen (Geigy), Chloropyribenzamine (Polfa); several multi-ingredient preparations.

Chlorothen [*148-65-2*], *N*-[(5-chloro-2-thienyl)methyl]-*N′*,*N′*-dimethyl-*N*-2-pyridinyl-1,2-ethanediamine, $C_{14}H_{18}ClN_3S$, M_r 295.8. Citrate [*148-64-1*], hydrochloride [*135-35-3*].

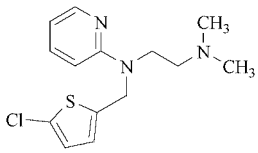

Chlorothen is synthesized by condensation of 5-chloro-2-thienylchloride and *N*,*N*-dimethyl-*N′*-(2-pyridinyl)ethylenediamine in the presence of sodium or potassium amide [22].

Trade name: Tagathen (Lederle).

Histapyrrodine [*493-80-1*], *N*-phenyl-*N*-(phenylmethyl)-1-pyrrolidineethanamine, $C_{19}H_{24}N_2$, M_r 280.4. Hydrochloride, *mp* 196 – 197 °C.

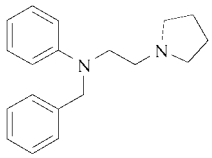

Preparation: [23].
Trade names: Domistan (Servier), Luvistin (Boehringer Ingelheim), Calcistin (Galenus).

Mepyramine [*91-84-9*], *N*-[(4-methoxyphenyl)methyl]-*N′*,*N′*-dimethyl-*N*-2-pyridinyl-1,2-ethanediamine, $C_{17}H_{23}N_3O$, M_r 285.4. Hydrochloride [*6036-95-9*], *mp* 143 °C; maleate [*59-33-6*].

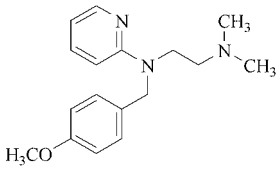

Preparation: [24].
Trade names: Neo-Antergan (Rhône-Poulenc; Merck USA), Neo-Bridal (Bayer), Pyramal (Roxane). Several multi-ingredient preparations.

Methapyrilen [*91-80-5*], *N*,*N*-dimethyl-*N′*-2-pyridinyl-*N′*-(2-thienylmethyl)-1,2-ethanediamine, $C_{14}H_{19}N_3S$, M_r 261.4. Hydrochloride [*135-23-9*], *mp* 162 °C; fumarate [*33032-12-1*].

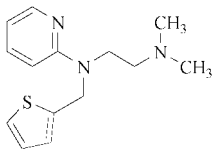

Methapyrilene is synthesized by heating a 2-thienyl halide with an alkali metal salt of *N*,*N*-dimethyl-*N′*-(2-pyridinyl)-1,2-ethanediamine [25].

Trade names: Thenylene (Abbott), Semikon (Beecham), Thionylan (Lilly).

Phenbenzamine [*961-71-7*], *N*,*N*-dimethyl-*N′*-phenyl-*N′*-(phenylmethyl)-1,2-ethanediamine, $C_{17}H_{22}N_2$, M_r 254.4. (Structural formula above.) Monohydrochloride [*2045-52-5*], *mp* 210 – 211 °C; dihydrochloride [*64335-73-5*]; citrate [*5956-46-7*].
Preparation: [26].
Trade names: Antergan (Rhône-Poulenc), Bridal (Bayer), Several multi-ingredient preparations.

Thenyldiamine [*91-79-2*] is the 3-thenyl isomer of methapyrilen. Hydrochloride [*958-93-0*]
Preparation: [27].
Trade names: Thenfadil (Sterling-Wintrop).

Thonzylamine [*91-85-0*], *N*-[(4-methoxyphenyl)methyl]-*N′*,*N′*-dimethyl-*N*-2-pyrimidinyl-1,2-ethanediamine, $C_{16}H_{22}N_4O$, M_r 287.3. Hydrochloride [*63-56-9*], *mp* 173 – 176 °C.

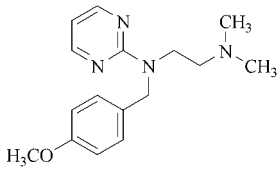

Thonzylamine is synthesized by treating the sodium salt of 2-(4-methoxybenzyl)aminopyrimidine with *N*,*N*-dimethyl-2-chloroethylamine [28].
Trade names: Tonamil (Ecobi).

Tripelenamine [*91-81-6*], *N*,*N*-dimethyl-*N′*-(phenylmethyl)-*N′*-2-pyridinyl-1,2-ethanediamine, $C_{16}H_{21}N_3$, M_r 255.3. Hydrochloride [*154-69-8*], *mp* 192 – 193 °C. Citrate [*6138-56-3*], *mp* 106 – 110 °C.

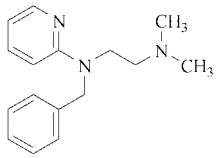

Preparation: [24, 29].
Trade names: Pyribenzamine (Ciba – Geigy), Azaron (Chefaro).

2.1.2. Ethanolamine Derivatives

Derivatives of aminoalkyl ethers were preferentially synthesized in the 1940s and 1950s in the United States. The most frequently used representative of this series is diphenhydramine that has been introduced into therapy in 1946.

In general, the histamine antagonistic potency of these basic ethers is lower than that of the ethylenediamines. Beside their antihistaminic activity, they have antimuscarinic and pronounced central sedative effects and are thus suited as sleeping aids. Because of their antiemetic properties they are used for prevention and treatment of motion sickness, emesis. Furthermore, these ethanolamine derivatives are commonly included in preparations used for the treatment of cold and cough.

Synthesis of Ethanolamines. In all known synthetic methods a benzhydryl unit is used as starting material [30]. Two synthetic routes are normally followed:

1) The sodium salt of a benzhydrol, prepared with NaH, $NaNH_2$, or $NaOC_2H_5$, is treated with an aminoethyl halide, e.g., dimethylaminoethyl chloride.
2) Sometimes a benzhydryl halide is condensed with an aminoethanol in the presence of potassium carbonate.

Bromdiphenhydramine [*118-23-0*], bromazine, 2-[(4-bromophenyl)phenylmethoxy]-*N*,*N*-dimethylethanamine, $C_{17}H_{20}BrNO$, M_r 334.3. Hydrochloride [*1808-12-4*], *mp* 144 – 145 °C.

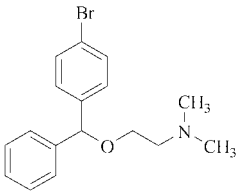

Preparation: [31].

Trade names: Ambodryl (Parke – Davis), multi-ingredient preparations such as Ambenyl.

Carbinoxamine [*486-16-8*], 2-[(4-chlorophenyl)-2-pyridinylmethoxy]-*N*,*N*-dimethylethanamine, $C_{16}H_{19}ClN_2O$, M_r 290.8. Hydrochloride, *mp* 162 – 164 °C; maleate [*3505-38-2*].

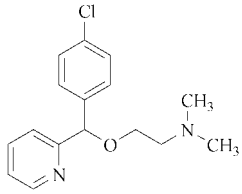

Preparation: [32].

Trade names: Allergefon (Lafon), Clistin (McNeil), Ciberon (Taisho), Polistin T-Caps (Trommsdorff). Several multi-ingredient preparations.

Clemastine [*15686-51-8*], meclastine, 2-[2-[1(4-chlorophenyl)-1-phenylethoxy]ethyl]-1-methylpyrrolidine, $C_{21}H_{26}ClNO$, M_r 343.9. Hydrochloride, *mp* 152 – 155 °C; fumarate [*14976-57-9*].

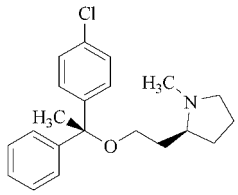

Preparation: [33].

Trade names: Tavegil (Sandoz – Novartis), Telgin-G (Takeda). Some multi-ingredient preparations such as Dexa Tavegil.

Diphenhydramine [*58-73-1*], 2-(diphenylmethoxy)-*N*,*N*-dimethylethanamine, $C_{17}H_{21}NO$, M_r 255.3. Hydrochloride [*147-24-0*], *mp* 166 – 170 °C.

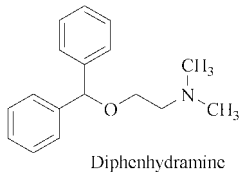

Preparation: [30]. To counteract the sedating effects, diphenhydramine is combined with 8-chlorotheophylline (Dimenhydrinate).

Trade names: Benadryl (Parke – Davis), Benadryl N (Warner – Lambert), Fenylhist (Mallard). Several multi-ingredient preparations.

Doxylamine [*469-21-6*], *N*,*N*-dimethyl-2-[1-phenyl-1-(2-pyridinyl)ethoxy]ethanamine, $C_{17}H_{22}N_2O$, M_r 270.4. Succinate [*562-10-7*], *mp* 100 – 104 °C.

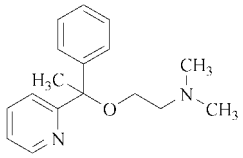

Doxylamine is synthesized from phenyl-2-pyridinylmethyl carbinol and dimethylaminoethyl chloride in the presence of $NaNH_2$ in xylene [34]. It exhibits pronounced sedative effects. Therefore, doxylamine is mainly used as a sedative, a sleeping aid.

Trade names: Mereprine (Merrell), Hoggar N (Stada), Unisom (Pfizer).

Ebramine [*3565-72-8*], 2-[(4-bromophenyl)-1-phenylethoxy]-*N*,*N*-dimethylethanamine, $C_{18}H_{22}BrNO$. Hydrochloride [*13977-28-1*], *mp* 150 – 152 °C.

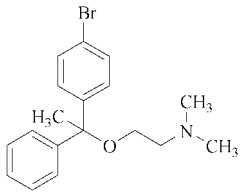

Preparation: [35].

Trade names: Bromadryl (Spofa).

Mephenhydramine [*3572-74-5*], 2-(1,1-diphenylethoxy)-*N*,*N*-dimethylethanamine, $C_{18}H_{23}NO$, M_r 269.4. Hydrochloride, *mp* 168 °C.

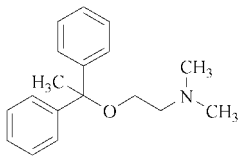

Compared with diphenhydramine the antihistaminic activity of mephenhydramin has been enhanced by the introduction of an additional methyl group.

Preparation: [36].

Trade names: Alphadryl (Spofa). Several multi-ingredient preparations.

Phenyltoloxamine [*92-12-6*], *N*,*N*-dimethyl-2-[2-(phenylmethyl)phenoxy]ethanamine, $C_{17}H_{21}NO$, M_r 255.4. Citrate [*1176-08-5*], *mp* 138 – 140 °C.

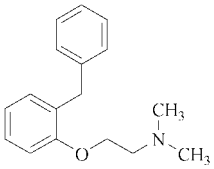

Preparation: [37].

Trade names: Phenyltoloxamine is available only in multi-ingredient preparations, e.g., Codipront (Mack Illertissen).

Trimethobenzamide [*138-56-7*], *N*-[[4-[2-(dimethylamino)ethoxy]phenyl]methyl]-3,4,5-trimethoxybenzamide, $C_{21}H_{28}N_2O_5$, M_r 388.5. Hydrochloride [*554-92-7*].

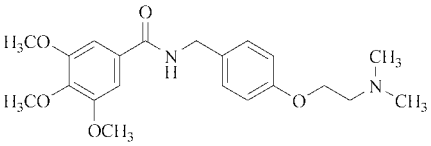

Preparation: [38].

Trade names: Anaus (Molteni), Tigan (Smith Kline Beecham).

2.1.3. Alkylamines

At the end of the 1940s, it was observed that appropriately substituted aminopropanes exhibit antihistaminic activities [39, 40]. Substituted aminopropanes have been introduced into therapy as antihistamines in the early 1950s. One of the remarkable features of these alkylamine derivatives is their strong and long-lasting antihistaminic activity. They are only moderately sedating and show some antimuscarinic properties. Among the stereoisomers, the dextrorotary form (D-form) is the eutomer (i.e., enantiomer with the higher bilological activity) [41].

Brompheniramine [*86-22-6*], 3-(4-bromophenyl)-3-(2-pyridinyl)-*N*,*N*-dimethylpropanamine, $C_{16}H_{19}BrN_2$, M_r 319.2. Maleate [*980-71-2*], *mp* 132 – 134 °C.

Preparation [42]:

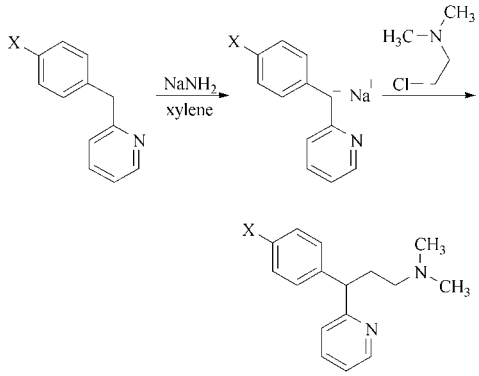

where X = Br (bromopheniramine), Cl (chloropheniramine) or H (pheniramine).

Similarly to other pheniramine derivatives, brompheniramine can also be administered parenterally.

Trade names: Ilvin (Merck), Dimegan (Dexo). Several multi-ingredient preparations.

Chlorpheniramine [*132-22-9*], 3-(4-chlorophenyl)-3-(2-pyridinyl)-*N*,*N*-dimethylpropanamine, $C_{16}H_{19}ClN_2$, M_r 274.8. Maleate [*113-92-8*], *mp* 130 – 135 °C. Synthesis [42], formula see page 8. The racemate can be separated into the enantiomers; the D-form, dexchlorpheniramine, is active [43].

Trade names: Teldrin (SKF), D-form: Polaronil (Schering), Polaramin (Schering), Phenamin (Nycomed). Several multi-ingredient preparations, e.g., Codicaps (Thiemann).

Dimetindene [*5636-83-9*], *N*,*N*-dimethyl-3[1-(2-pyridinyl)ethyl]-1*H*-indene-2-ethanamine, $C_{20}H_{24}N_2$, M_r 292.4. Maleate [*3614-69-5*], *mp* 159 – 161 °C.

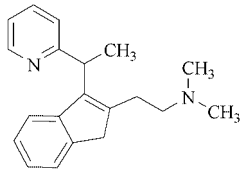

Preparation: Benzylmalonic acid is esterified with dihydropyran and reacted with sodium hydride and 2-dimethylaminoethyl chloride in toluene to give ditetrahydropyranyl-2-benzyl-2-(2-dimethylaminoethyl)malonate, which is cyclized with the help of polyphosphoric acid to 2-(2-dimethylaminoethyl)-indan-1-one. This compound reacts with 2-ethylpyridine and phenyllithium to give the carbinol, which is dehydrated by heating to dimetindene [44].

Trade names: Fenistil (Zyma), Forhistal (Ciba). Several multi-ingredient preparations.

Pheniramine [*86-21-5*], *N*,*N*-dimethyl-γ-phenyl-2-pyridinepropanamine, $C_{16}H_{20}N_2$, M_r 240.3. Maleate [*132-20-7*], *mp* 107 °C. (Formula and preparation see page 8).

Trade name: Avil (Hoechst). Several multi-ingredient preparations.

Tolpropamine [*5632-44-0*], 3-phenyl-3-(4-methylphenyl)-*N*,*N*-dimethylpropanamine, $C_{18}H_{23}N$, M_r 253.4. Hydrochloride [*3339-11-5*].

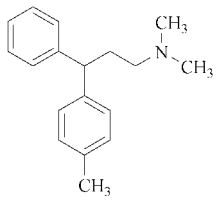

Preparation: [41].

Trade name: Pragman Gelee (Milanfarma).

Triprolidine [*486-12-4*], (*E*)-2-[1-(4-methylphenyl)-3-(1-pyrrolidinyl)-1-propenyl]pyridine, $C_{19}H_{22}N_2$, M_r 278.4. Hydrochloride [*550-70-9*], *mp* 116 – 118 °C; hydrochloride monohydrate [*6138-79-0*]. Synthesis [45]:

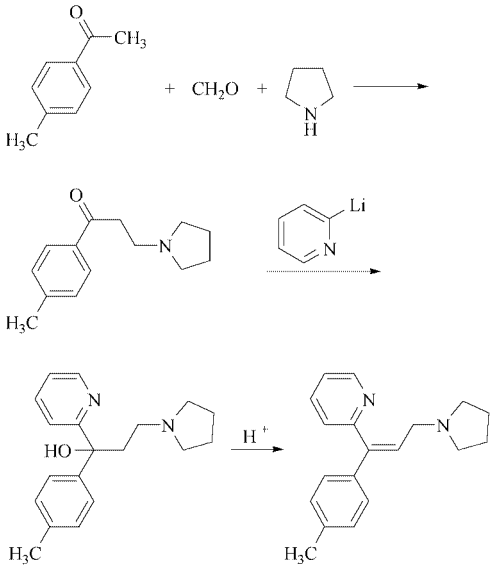

Triprolide

Trade names: Pro-Actidil (Wellcome), Venen (Tanabe). Several multi-ingredient preparations, e.g., Actifed (Warner – Lambert).

2.1.4. Piperazines

Buclizine [*82-95-1*], 1-[(4-chlorophenyl)-phenylmethyl]-4-[4-(1,1-dimethylethyl)phenyl]methyl]piperazine, $C_{28}H_{33}ClN_2$, M_r 433.0. Dihydrochloride [*29-74-8*].

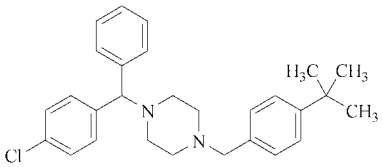

Preparation: [46]. Buclizine is mainly used in the prevention of motion sickness and in combination with analgesics for treatment of migraine.

Trade names: Buclifen (Pfizer), Posdel (UCB). Several multi-ingredient preparations, e.g., Migralave N (Tremmler Pharma).

Chlorcyclizine [*82-93-9*], 1-[(4-chlorophenyl)phenylmethyl]-4-methylpiperazine, $C_{18}H_{21}ClN_2$, M_r 300.8. Hydrochloride [*1620-21-9*], *mp* 216 °C.

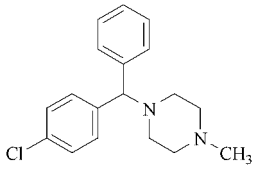

Preparation: [47].

Trade names: Di-Paralene (Abbott), Histantin (Wellcome).

Cinnarizine [*298-57-7*], 1-(diphenylmethyl)-4-(3-phenyl-2-propenyl)piperazine, $C_{26}H_{28}N_2$, M_r 368.5.

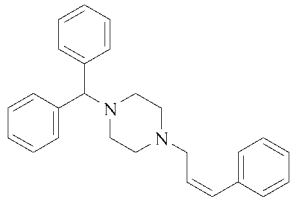

Preparation: Cinnarinzine is prepared by alkylation of 1-diphenylmethyl piperazine with 3-phenyl-2-propenyl chloride [48]. In addition to its antihistaminic activity, cinnarizine possesses Ca^{2+} channel blocking activity. Its vasodilator effect probably results from blocking the Ca^{2+} channels in the vascular smooth muscle cells. Therefore, cinnarizine is used in peripheral and central vascular diseases. Labyrinth disorders (e.g. vertigo, Ménière syndrome) are also often treated with cinnarizine.

Trade names: Stutgeron (Janssen), Aplactan (Eisai), Mitronal (Searle).

Clocinizine [*298-55-5*], 1-[(4-chlorophenyl)phenylmethyl]-4-(3-phenyl-2-propenyl)piperazine, $C_{26}H_{27}ClN_2$, M_r 403.0.

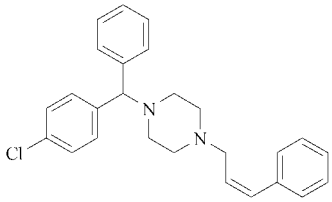

Preparation: [49].

Trade names: Clocinizine is used only in multi-ingredient preparations, e.g., Denoral (Rorer).

Cyclizine [*82-92-8*], 1-diphenylmethyl-4-methylpiperazine, $C_{18}H_{22}N_2$, M_r 266.4, *mp* 105 – 107 °C. Hydrochloride [*303-25-3*].

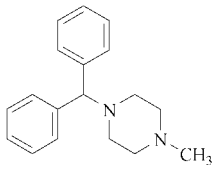

Preparation: Cyclizine is prepared by reaction of benzhydryl chloride and *N*-methyl piperazine [50]. It can be used not only in the treatment of allergic diseases, but also in the management of postoperative, irradiation or drug-induced vomiting. Cyclizine dependence has been suggested to occur when it is used in combination with analgesics in long-term therapy. For this reason, several countries have removed cyclicine from analgesic combination preparations.

Trade names: Marziné (Burroughs Wellcome). Some multi-ingredient preparations, e.g., Cyclimorph.

Flunarizine [*52468-60-7*], (*E*)-1-[bis(4-fluorophenyl)methyl]-4-(3-phenyl-2-propenyl)-piperazine, $C_{26}H_{26}F_2N_2$, M_r 404.5. Dihydrochloride [*30484-77-6*].

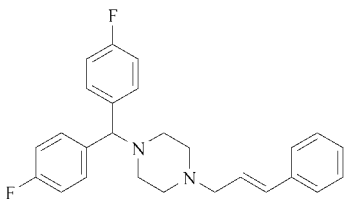

Flunarizine is the difluorinated derivative of cinnarizine. Thus, its preparation and therapeutic use are identical to those of cinnarizine.

Trade names: Sibelium (Janssen), Fluxarten (Beecham).

Hydroxyzine [*68-88-2*], 2-[2-[4-[4-(chlorophenyl)phenylmethyl]-1-piperazinyl]-ethoxy]ethanol, $C_{21}H_{27}ClN_2O_2$, M_r 410.0. Dihydrochloride [*2192-20-3*], *mp* 193 °C.

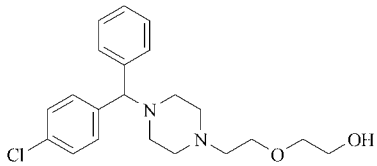

Preparation: [51]. Hydroxyzine has antihistaminic, antimuscarinic, and central nervous properties. Hydroxizine is used not only in the treatment of allergic diseases, but also as an anxyolitic.

Trade names: Atarax (Pfizer), Orgatrax (Organon).

Meclozine [*569-65-3*], mezlizine, 1-[(4-chlorophenyl)phenylmethyl]-4-[3-methylphenyl)methyl]piperazine, $C_{25}H_{27}ClN_2$, M_r 391.0. Dihydrochloride monohydrate [*31884-77-2*].

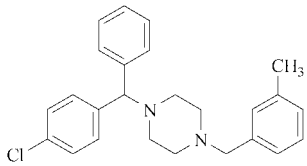

Preparation: The synthesis of meclozine is very similar to that of cyclizine.

Trade names: Postafen (UCB), Bonine (Pfizer), Veritab (Vista).

Niaprazine [*27367-90-4*], *N*-[3-[4-(4-fluorophenyl)-1-piperazinyl]-1-methylpropyl]-3-pyridinecarboxamide, $C_{20}H_{25}FN_4O$, M_r 356.4.

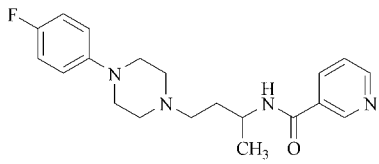

Preparation: [52].

Trade names: Nopron (Carrion).

Oxatomide [*60607-34-3*], 1-[3-[4-(diphenylmethyl)-1-piperazinyl]propyl]-1,3-dihydro-

2*H*-benzimidazol-2-one, $C_{27}H_{30}N_4O$, M_r 426.2.

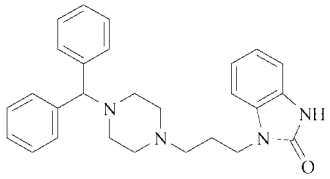

Preparation: Oxatomide is synthesized by alkylation of 1-diphenylmethylpiperazine with 1-(3-chloropropyl)-1,3-dihydro-2*H*-benzimidazol-2-one in the presence of Na_2CO_3 [53]. In addition to its histamine H_1 receptor antagonistic effect, oxatomide may have mast-cell stabilizing activities [54].

Trade names: Tinset (Janssen), Celtect (Kyowa).

2.1.5. Piperidines

Bamipine [*4945-47-5*], 1-methyl-*N*-phenyl-*N*-(phenylmethyl)-4-piperidinamine, $C_{19}H_{24}N_2$, M_r 280.4. Dihydrochloride [*61732-85-2*].

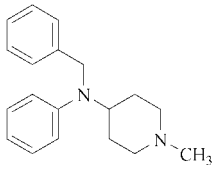

Preparation: [55].

Trade names: Soventol (Knoll).

Diphenylpyraline [*147-20-6*], 4-(diphenylmethoxy)-1-methylpiperidine, $C_{19}H_{23}NO$, M_r 281.4. Hydrochloride [*132-18-3*].

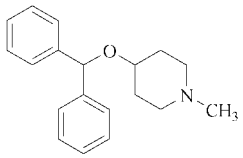

Preparation: [160].

Trade names: Arbid N (Bayer), Lergobine (3M). Some multi-ingredient preparations.

Pipoxizine [*55837-21-3*], 2-[2-[2-[4-(Diphenylmethylene)piperidinyl]ethoxy]ethoxy]ethanol, $C_{24}H_{31}NO_3$, M_r 381.3. Hydrochloride, *mp* 144 – 145 °C. Synthesis [56]:

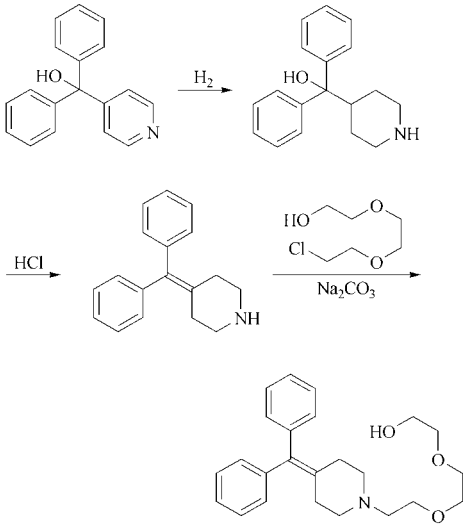

Pipoxizine

Trade names: Respacal (UCB).

Thenalidine [*86-12-4*], 1-methyl-*N*-phenyl-*N*-(2-thienylmethyl)-4-piperidinamine. $C_{17}H_{22}N_2S$, M_r 286.5. Tartrate, *mp* 170 – 172 °C.

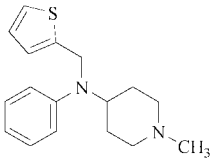

Preparation: Thenalidine is synthesized by alkylation of 1-methyl-4-phenylaminopiperidine with 2-chloromethyl thiophene [57].

Trade names: Sandosten (Sandoz).

2.1.6. Phenothiazines

Alimemazone [*84-96-8*], 10-[3-(dimethylamino)-2-methylpropyl]phenothiazine, $C_{18}H_{22}N_2S$, M_r 298.5. Tartrate [*4330-99-8*].

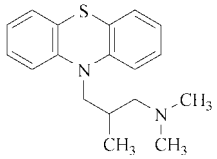

Preparation: [58].
Trade names: Theralene (Rhône – Poulenc), Temaril (Allergan).

Dimethothiazine [*7456-24-8*], 10-[2-(dimethylamino)-propyl]-*N*,*N*-dimethyl-10*H*-phenothiazine-2-sulfonamide, $C_{19}H_{25}N_3O_2S_2$, M_r 391.6. Methanesulfonate [*7455-39-2*].

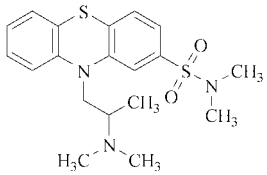

Preparation: [59].
Trade names: Banistyl (M & B), Migristene (Specia).

Dimelazine [*15302-12-2*], 10-[(1,3-dimethyl-3-pyrrolidinyl)methyl]-10*H*-phenothiazine, $C_{19}H_{22}N_2S$, M_r 310.1.

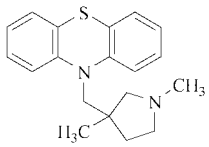

Trade names: Centrophene (Biosedra), Tacaryl (Westwood).

Dioxopromethazone [*13754-56-8*], 10-[2-(dimethylamino)propyl]-10*H*-phenothiazine-5,5-dioxide, $C_{17}H_{20}N_2O_2S$, M_r 316.2.

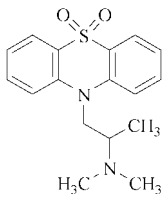

Trade names: Prothanon (LAW).

Hydroxyethylpromethazine [*7647-63-4*], (2-hydroxyethyl)dimethyl-(1-methyl-2-phenothiazi-10-ylethyl)ammonium chloride, $C_{19}H_{25}ClN_2OS$, M_r 364.9.

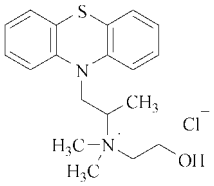

Preparation: [60].
Trade names: Aprobit (Kabi).

Isothipendyl [*482-15-5*], 10-(2-dimethylamino-2-methylethyl)-10*H*-pyrido[3,2-*b*][1,4]-benzothiazine, $C_{16}H_{19}N_3S$, M_r 285.4. Hydrochloride [*1225-60-1*] *mp* 222 – 223 °C. Preparation [61]:

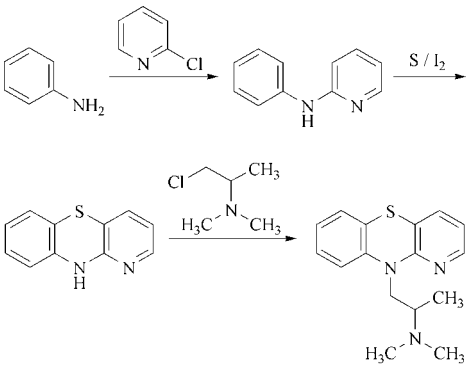

Isothipendyl

Isothipendyl is more active than promethazine and less sedating [62].
Trade names: Theruhistin (Ayerst).

Mequitazine [*29216-28-2*], 10-(1-azabicyclo[2.2.2]oct-3-ylmethyl)-10*H*-phenothiazine, $C_{20}H_{22}N_2S$, M_r 322.5.

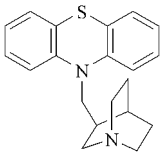

Preparation: [63]. Mequitazine has mild central sedating effects.

Trade names: Metaplexan (Rhône – Poulenc – Rorer), Butix (Fabre).

Methdilazine [*1982-37-2*], 10-[(1-methyl-3-pyrrolidinyl)methyl]-10*H*-phenothiazine, $C_{18}H_{20}N_2S$, M_r 296.4. Hydrochloride [*1229-35-2*], *mp* 187 – 189 °C.

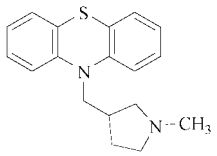

Preparation: [64].

Trade names: Dilosyn (British Drug Houses), Tacaryl (Westwood).

Oxomemazine [*3689-50-7*], 10-(3-dimethylamino-2-methylpropyl)phenothiazine-5,5-dioxide, $C_{18}H_{22}N_2O_2S$, M_r 330.5. Hydrochloride [*4784-40-1*].

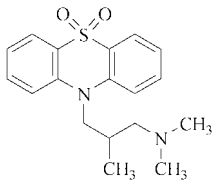

Preparation: [65].

Trade names: Doxergan (Rhône – Poulenc).

Promethazine [*60-87-7*], 10-(2-dimethylamino-propyl)phenothiazine, $C_{17}H_{20}N_2S$, M_r 284.4. Hydrochloride [*58-33-3*] *mp* 244 – 246 °C.

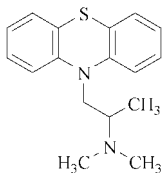

Preparation: [66]. Promethazine is synthesized from 10-phenothiazine-2-propyl chloride and dimethylamine in the presence of Cu. In addition to its antihistaminic activity, it has marked central sedative properties.

Trade names: Atosil (Bayer), Phenergan (Rhône – Poulenc, Wyeth), Hiberna (Yoshitomi). Several multi-ingredient preparations.

Propiomazine [*362-29-8*], 1-[2-(dimethylamino)propyl]-10*H*-phenothiazin-2-yl]-1-propanone, $C_{20}H_{24}N_2OS$, M_r 340.5. Hydrochloride [*1240-15-9*].

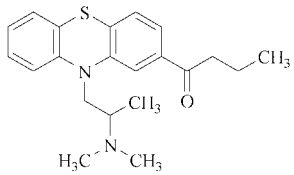

Preparation: [67].

Trade names: Propavan (Kabi), Indorm (Wyeth).

Thiazinamium [*2338-21-8*], trimethyl-(1-methyl-2-phenothiazin-10-ylethyl)ammonium methyl sulfate, $C_{19}H_{26}N_2O_4S_2$, M_r 410.6.

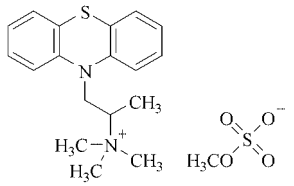

Preparation: Thiazinamium is prepared by reacting promethazine with dimethyl sulfate [68].

Trade names: Padisal (Bayer), Multergan (Specia).

2.1.7. Other Tricyclic Systems

Azatadine [*3964-81-6*], 6,11-dihydro-11-(1-methyl-4-piperidylidene)-5*H*-benzo[5,6]cyclohepta[1,2-*b*]pyridine, $C_{20}H_{22}N_2$, M_r 290.4, *mp* 124 – 126 °C. Dimaleate [*3978-86-7*]. Synthesis [69]:

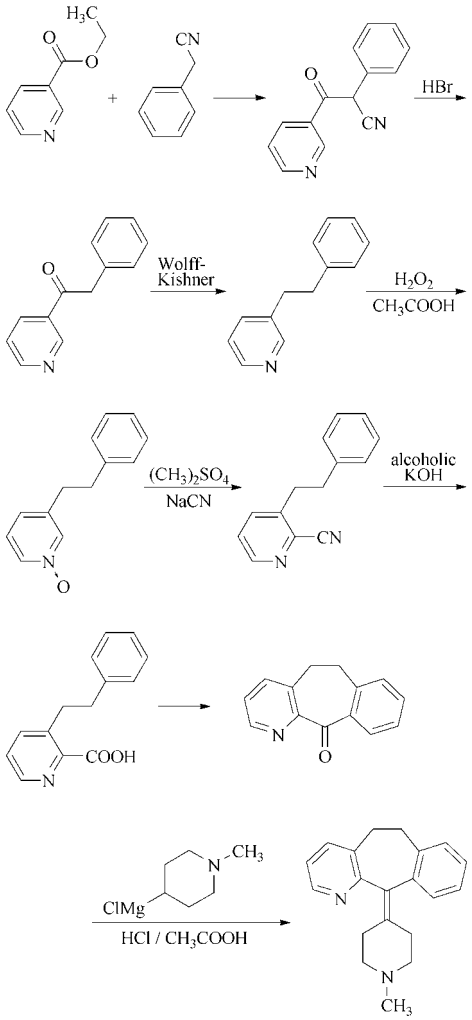

Azatadine

Trade names: Idulian (Unilabo), Zadine (Schering).

Clobenzepam [*1159-93-9*], 7-chloro-10-(2-dimethylaminoethyl)-5,10-dihydro-11*H*-dibenzo[*b*,*e*][1,4]diazepin-11-one, $C_{17}H_{18}ClN_3O$, M_r 315.8, Hydrochloride [*2726-03-6*], *mp* 225 – 233 °C.

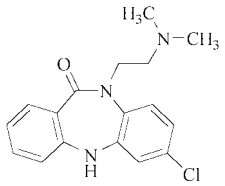

Preparation: [70].

Trade names: Tarpan (Wander).

Cyproheptadine [*129-03-3*], 4-(5*H*-dibenzo[*a*,*d*]cyclohepten-5-ylidene)-1-methylpiperidine, $C_{21}H_{21}N$, M_r 287.4. Hydrochloride [*969-33-5*], *mp* 252 – 254 °C. In addition to its antiallergic action cyproheptadine also exhibits an appetite-stimulating effect. For this reason, it is often used for promotion of weight gain.

Preparation [71]:

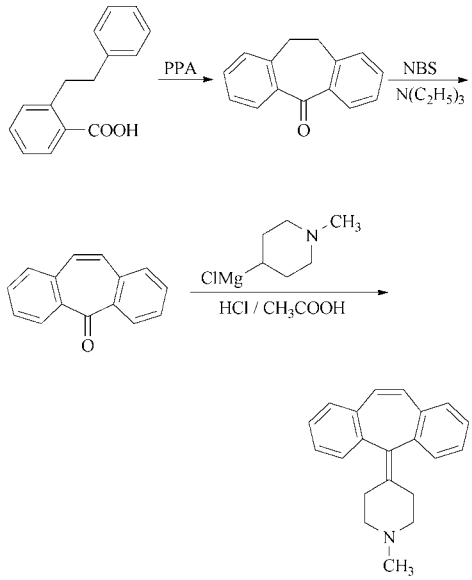

Cyproheptadiene

Trade names: Nuran (Frosst), Periactin (MSD), Peritol (EGYT). Some multi-ingredient preparations.

Ketotifen [*34580-13-7*], 4,9-dihydro-4-(1-methyl-4-piperidylidene)-10*H*-benzo[4,5]cyclohepta[1,2-*b*]thiophene-10(9*H*)-one, $C_{19}H_{19}NOS$, M_r 309.2. Hydrogen fumarate [*34580-14-8*], *mp* 192 °C.

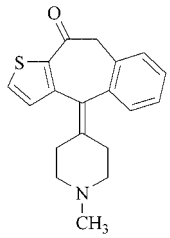

Preparation: [71].
Trade names: Zaditen (Sandoz – Novartis). Some multi-ingredient preparations.

Mebhydroline [*524-81-2*], 2,3,4,5-tetrahydro-2-methyl-5-(phenylmethyl)-1*H*-pyrido[4,3-*p*]indole, $C_{19}H_{20}N_2$, M_r 276.4. 1,5-Naphthalendisulfonate [*6153-33-9*].

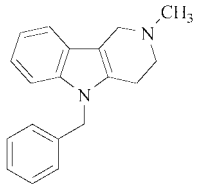

Preparation: [72].
Trade names: Omeril (Bayer).

Phenindamine [*82-88-2*], 2,3,4,9-tetrahydro-2-methyl-9-phenyl-1*H*-indeno[2,1-*c*]pyridine, $C_{19}H_{19}N$, M_r 261.4. Hydrogen tartrate [*569-59-5*], *mp* 165 – 167 °C.

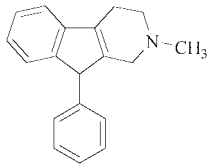

Preparation: [73].
Trade names: Pernovin (Chinoin), Thephorin (Sinclair).

Pimethixene [*314-03-4*], 1-methyl-4-(9*H*-thioxanthen-9-ylidene)piperidine, $C_{19}H_{19}NS$, M_r 293.4, *mp* 120 – 121 °C. Synthesis [74]:

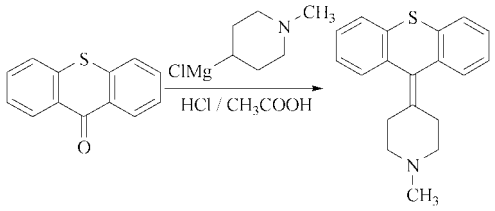

Pimethixene

Trade names: Muricalm (Sandoz – Novartis).

2.2. Second-Generation Antihistamines

Antihistamines of the second generation are generally free of sedation. However, this effect is dose-dependent. At higher doses and, consequently, at higher plasma levels of the drugs, mild sedative effects can be observed.

2.2.1. Ethanolamines

Setastine [*64294-95-7*], 1-[2[1-(4-chlorophenyl)-1-phenylethoxy]ethyl]hexahydro-1*H*-azepine, $C_{22}H_{28}ClNO$, M_r 357.9. Hydrochloride [*59767-13-4*].

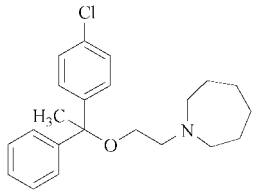

Preparation: [75].
Trade names: Loderix (EGIS).

2.2.2. Alkylamines

Acrivastine [*87848-99-5*], (*E*,*E*)-3-[6-[1-(4-methylphenyl)-3-(1-pyrrolidinyl)-1-propenyl]--2-pyridinyl]propenoic acid, $C_{22}H_{24}N_2O_2$, M_r 348.4.

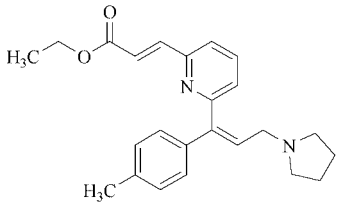

Preparation: [76].
Trade names: Semprex (Glaxo – Wellcome).

2.2.3. Piperazines

Cetirizine [*83881-51-0*], 2-[4-[1-(4-chlorophenyl)phenylmethyl]-1-piperazinyl]ethoxy]acetic acid, $C_{21}H_{25}ClN_2O_3$, M_r 388.5. Dihydrochloride [*83881-52-1*].

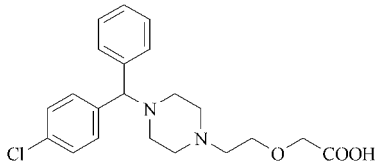

Preparation: [77]. Cetirizine is a metabolite of hydroxizine with a long plasma half-life.

Trade names: Zyrtec (UCB), Reactine (Pfizer).

2.2.4. Piperidines

Astemizole [*68844-77-9*], 1-(4-fluorophenyl)-methyl-*N*-[1-[2-(4-methoxyphenyl)ethyl]-4-piperidinyl]-1*H*-benzimidazol-2-amine, $C_{28}H_{31}FN_4O$, M_r 458.3, *mp* 149 °C. Synthesis [78]:

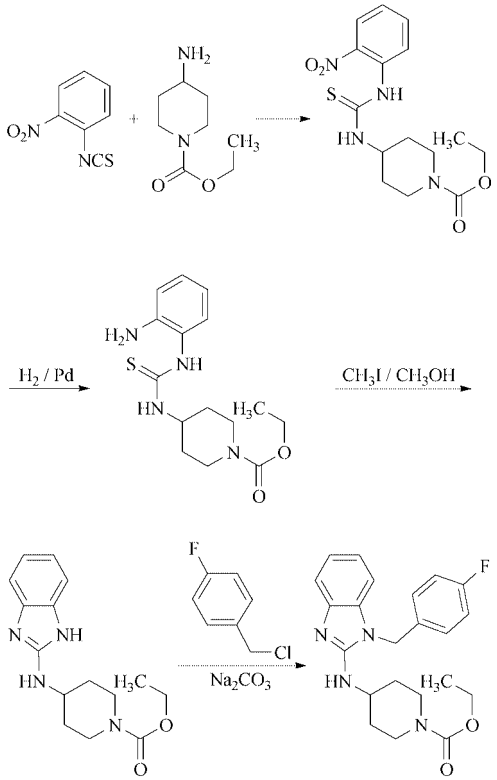

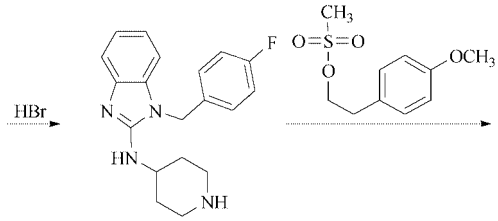

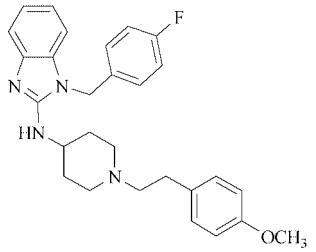

Astemizole

Astemizole is a strong histamine H1-receptor antagonist with an extremely long elimination half-life. Prolongation of QTc interval and cardiac arrhythmias can be associated with its use [79].

Trade names: Histmanol (Janssen), Cilergil (Cilag), Almizol (Nobel).

Betotastine Besilate [*190786-44-8*], (+)-(*S*)-4-[4-[1-(4-chlorophenyl)-1-(2-pyridyl)-methoxy]piperidin-1-yl]butanoic acid monobenzenesulfonate, $C_{21}H_{25}ClN_2O_3 \cdot C_6H_6O_3S$, M_r 547.1.

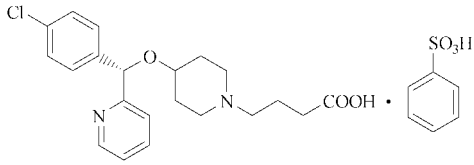

Preparation: [80]. Under development by Tanabe Seiyaku and Ube.

Ebastine [*90729-43-4*], 1-[4-(1,1-dimethylethyl)phenyl]-4-[4-(diphenylmethoxy)-1-piperidinyl]-1-butanone, $C_{32}H_{39}NO_2$, M_r 469.0.

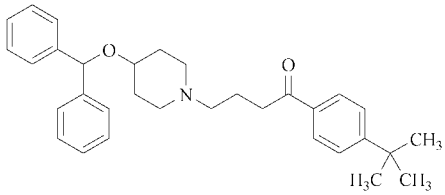

Preparation: [81].
Trade names: Ebastel (Almirall).

Fexofenadine [*153439-40-8*], 2-methyl-2-[4-[1-hydroxy-4-[4-(1,1-diphenylhydroxymethyl)-1-piperidinyl]butyl]phenyl]propionic acid, $C_{32}H_{39}NO_4$, M_r 501.2.

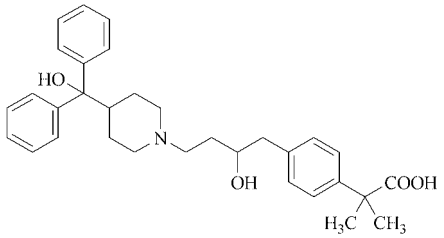

Fexofenadine is a metabolite of terfenadine (see page 17) and is free of cardiovascular side effects [82].
Trade names: Allegra (Hoechst Marion Roussel).

Levocabastine [*79516-68-0*], 1-[4-cyano-4-(4-fluorophenyl)cyclohexyl]-3-methyl-4-phenyl-4-piperidinic acid, $C_{26}H_{26}FN_2O_2$, M_r 417.2. Hydrochloride [*79547-78-7*].

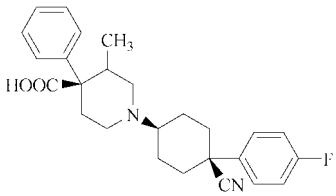

Trade names: Livocab (Janssen), Levocon (Johnson & Johnson).

Mizolastine [*108612-45-9*], 1-(4-fluorobenzyl)-2-[4-[*N*-(3,4-dihydro-4-oxopyrimidin-2-yl)-*N*-methylamino]piperidin-1-yl]benzimidazole, $C_{24}H_{25}FN_6O$, M_r 432.5.

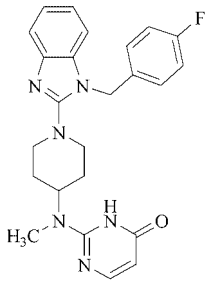

Preparation: [83]. Mizolastine is a potent nonsedating H_1-receptor antagonist providing satisfactory symptom relief in allergic rhinitis and urticaria [84].
Trade names: Mizollen (Synthelabo).

Noberastine [*110588-56-2*], 3-[(5-methyl-2-furanyl)methyl]-*N*-4-piperidinyl-3*H*-imidazo[4,5-*b*]pyridin-2-amine, $C_{17}H_{21}N_5O$, M_r 311.1.

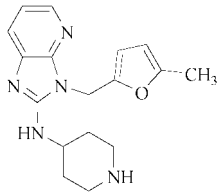

Under development by Cilag (phase III). During clinical trials noberastine did not produce any sedative effects.

Terfenadine [*50679-08-8*], 4-(1,1-dimethylethyl)-1-[1-hydroxy-4-[4-(1,1-diphenylhydroxymethyl)-1-piperidinyl]butyl]benzene, $C_{32}H_{41}NO_2$, M_r 471.2.

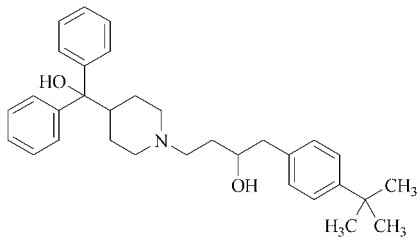

Due to its arrhythmogenic activity, terfenadine has been replaced by its metabolite, fexofenadine (see page 17) [85].

Trade name: Teldane (Hoechst Marion Roussel).

2.2.5. Phthalazinones

Azelastine [*58581-89-8*], 4-[(4-chlorophenyl)methyl]-2-(hexahydro-1-methyl-1*H*-azepin-4-yl)-1(2*H*)-phthalazinone, $C_{22}H_{24}ClN_3O$, M_r 381.7. Hydrochloride [*37932-96-0*], *mp* 225 – 229 °C. Synthesis [86]:

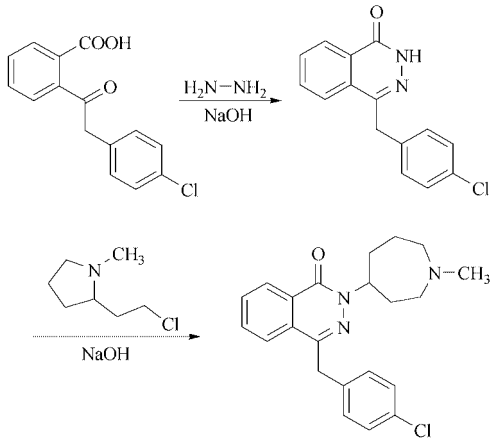

Azelastine

Azelastine has potent, long-lasting histamine H_1-receptor blocking activity and several other interesting anti-inflammatory properties related to its antiallergic efficacy [86].

Trade names: Allergodil, Rhinolast (ASTA Medica).

2.2.6. Others

Emedastine [*87233-61-2*], 1-(2-ethoxyethyl)--2-(hexahydro-4-methyl-1*H*-1,4-diazepin-1-yl]-1*H*-benzimidazole, $C_{17}H_{26}N_4O$, M_r 302.2.

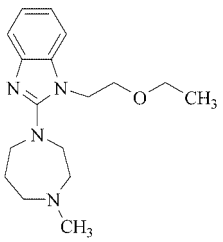

Trade names: Daren (Kanebo), Remicut (Kowa).

Epinastine [*80012-43-7*], 9,13*b*-dihydro-1*H*-dibenzo[*c*,*f*]imidazo[1,5-*a*]azepin-3-amine, $C_{16}H_{15}N_3$, M_r 249.1.

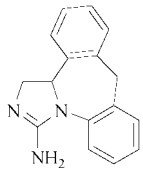

Preparation: [87].

Trade name: Alesion (Boehringer Ingelheim).

Loratadine [*79794-75-5*], 8-chloro-6,11-dihydro-11-(1-ethoxycarboxy-4-piperidinylidene)-5*H*-benzo[5,6]cyclohepta[1,2-*b*]-pyridine, $C_{22}H_{23}ClN_2O_2$, M_r 382.5.

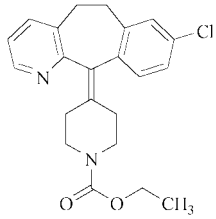

Loratadine is a selective antihistamine for the relief of symptoms of allergic rhinoconjunctivitis and chronic urticaria [88]. In 1997, loratadine has been shown to produced cardiac arrhythmias probably via blocking cardiac K^+ channels [89]. However, the concentrations at which this blockade occurred were much higher that the observed therapeutic plasma levels [90].

Trade name: Claritin (Schering – Plough), Lisino (Essex Pharma).

2.2.7. Combinations of Antihistamines and Sympathomimetic Agents

Antihistamines are often combined with sympathomimetic agents. For oral administration, indirectly acting sympathomimetics (e.g., pseudoephedrine) are preferred. Combinations for topical applications contain α-adrenergic agonists

(e.g., oxymetazoline). Recently it has been demonstrated that pseudoephedrine might increase the stroke risk. As regards combinations for topical applications, it should be kept in mind that long-term application of α-adrenergic compounds may cause serious undesired effects (rebound effect, nasal congestion).

2.3. Inhibitors of Histamine Synthesis

Tritoqualine [*14504-73-5*], 7-amino-4,5,6-triethoxy-3-(5,6,7,8-tetrahydro-4-methoxy-6-methyl-1,3-dioxolo[4,5-*g*]-isoquinolin-5-yl)--1(3*H*)isobenzofuranone, $C_{26}H_{32}N_2O_8$, M_r 500.5.

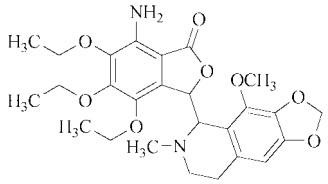

Preparation: [91].
Trade names: Inhibostamin (Zyma), Livalfa (Mitsubishi).

3. Mast Cell Stabilizing Agents

The prototype drug, disodium cromoglycate resulted from research on khellin, a chromone 2-carboxylic acid derived from the Middle Eastern herb *Ammi visnaga* [92]. Cromolyn sodium or disodium cromoglycate (DSCG) was synthesized in 1965 and introduced as a prophylactic, nonbronchodilating anti-inflammatory drug into the therapy of allergic disorders. Although DSCG was initially thought to be merely a mast cell stabilizer, subsequent pharmacological studies showed that it has multiple effects. Probably, it acts on certain types of chloride channels expressed in mast cells and sensory nerves [93]. Penner et al. [94] described the existence of a small conductance chloride channel in rat peritoneal mast cells. Activation of this channel causes an influx of chloride ions and results in hyperpolarization of the membrane which then drives a calcium influx, leading to degranulation. Disodium cromoglycate blocks these intermediate conductance channels which are activated after immunological stimulation [95]. Since cromones are practically not absorbed in the gastrointestinal tract, they must be administered by inhalation. All attempts to develop orally active cromones have been unsuccessful until now.

Amlexanox [*68302-57-8*], 2-amino-7-(1-methylethyl)-5-oxo-5*H*-[1]benzopyrano[2,3-*b*]pyridine-3-carboxylic acid, $C_{16}H_{14}N_2O_4$, M_r 298.2.

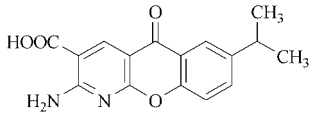

Amlexanox acts in two ways. Besides its membrane-stabilizing property amlexanox also inhibits leucotriene-mediated effects [101].
Preparation: [100].
Trade name: Solfa (Takeda).

Disodium cromoglycate [*15826-37-6*], DSCG, cromolyn sodium, 5,5′-[(2-hydroxy-1,3-propanediyl)bis(oxy)]bis[4-oxo-4*H*-1-benzopyran-2-carboxylic acid]disodium salt, $C_{23}H_{14}Na_2O_{11}$, M_r 512.3.

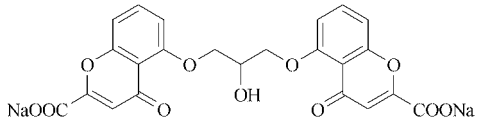

Preparation: [96].
Trade names: Intal, Adrane (Fisons).

Lodoxamide [*53882-12-5*], 2,2′-[(2-chloro-5-cyano-1,3-phenylene)diimino]bis[2-oxoacetic acid], $C_{11}H_6ClN_3O_6$, M_r 311.6.

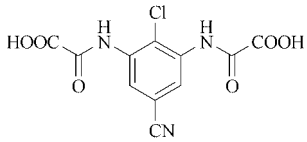

Trometamol (2-amino-2-(hydroxymethyl)-1,3-propanediol) salt [*63610-09-3*]. Lodoxamide is an antiallergic drug acting as a mast-cell stabi-

lizer, which is effective in the treatment of allergic conjunctivitis [99].

Trade names: Alomide (Alcon), Lomide (Pharmacia & Upjohn).

Nedocromil [*69049-73-6*], 9-ethyl-6,9-dihydro-4,6-dioxo-10-propyl-4*H*-pyrano[3,2-*g*]quinoline-2,8-dicarboxylic acid, $C_{19}H_{17}NO_7$, M_r 371.1. Disodium salt [*69049-74-7*].

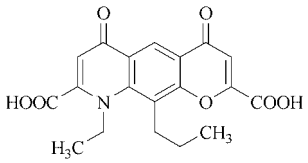

Nedocromil sodium acts mainly by inhibiting the release of inflammatory mediators. It has a stabilizing action on mast cells similar to that of DSCG. Nedocromil is capable of inhibiting chloride ion flux in mast cells, epithelial cells, and neurons. This feature may explain why it can prevent responses such as mast cell degranulation and neuronal activation [97]. Nedocromil is given by inhalation in the prophylactic control of asthma and rhinitis [98].

Trade name: Tilade (Rhône – Poulenc – Rorer).

4. Allergen Preparations for Desensitization

The aim of allergen immunotherapy (hyposensitization, desensitization) is to achieve allergen tolerance in the atopic patient by administering gradually increasing quantities of an allergen vaccine. This treatment was first used by Noon [102] and Freeman [103] in 1911 for allergic rhinitis on an empirical basis. Since then, immunotherapy has become a widely accepted method of treatment in IgE-mediated allergic diseases, caused by various seasonal or perennial inhalant allergens.

Since the 1970s venom immunotherapy is the standard of care for hymenoptera sting-induced systemic allergic reactions [104]. Venom immuntherapy results in significantly lower risk of systemic reactions even 10 to 20 years after therapy is stopped [105, 106]. Knowledge has been gained about the mechanism of action of immunotherapy. Like vaccines allergen extracts act as immune modifiers, so it seems justified to call them allergen vaccines. The important role of T_H1 and T_H2 lymphocytes has been explored as well as the cytokine regulation of the immune response [107]. Immunotherapy is the only treatment that may affect the natural course of allergic diseases, especially the chronification of inflammation of the respiratory tract [108 – 111]. Successful immunotherapy is dependent on the quality of allergen vaccines. The vaccines have to be standardized and manufactured properly with batch-to-batch consistency.

The ideal allergen vaccines permit a maximum efficacy while minimizing the incidence of adverse reactions. The pharmaceutical industry is constantly developing and investigating safer and at the same time equally or more effective forms of therapies. The molecular allergen characterization by recombinant DNA technology leads to novel forms of diagnostics and immunotherapy based on recombinant allergens [112 – 114]. Recombinant allergens identify exactly those IgE-binding molecules responsible for the allergic reaction.

4.1. Properties of Allergens

Allergenic components are most often of a proteinaceous nature. Their molecular weight ranges between 5 and 70 kD. Pollens of grasses, trees, and weeds are the main source of disease-causing allergens, followed by dust mites, molds, pets, and food. The allergic reaction to insect stings represents a special form, because it will not elicit any chronic inflammatory organic disease, but usually lead to a local or even systemic life-threatening anaphylactic shock. Some molecules bind as haptens to human proteins and lead only as a protein complex to allergic reactions.

Those allergenic parts to which most patients respond with formation of antibodies are called major allergens, those less often answered by an immunologic response minor allergens. A special and important kind of allergens are the so-called profilines. They are highly cross-reactive proteins with a similar amino acid structure found in related or nonbotanical related plants, food, or even in human cells. Patients who are sensitized to profilines are allergic to a broad

Table 1. Physicochemical properties of a choice of purified allergens (Comprehensive list under [117])

Species	Common name	Allergen	Comment	Aa_length
Aero Allergen Plants				
Ambrosia artemisifolia	Short ragweed	Amb a 6	Lipid transfer protein	118
Artemisia vulgaris	Mugwort	Art v 1	Major allergen precursor	132
Betula pendula	European white birch	Bet v 1	Isoform at 8	160
Betula pendula	European white birch	Bet v 2	Profilin	133
Corylus avellana	European hazelnut	Cor a 2	Profilin	131
Cryptomeria japonica	Japanese cedar	Cry j 1	Pectate lyase	374
Lollium perenne	Perennial ryegrass	Lol p 5.0102	Ribonuclease	308
Olea europea	Olive tree	Ole e 2	Profilin 3	134
Parietaria judaica	Weed	Par j 3	Profilin 2	131
Phleum pratense	Common timothy	Phl p 13	Polygalactorunase	394
Zea mays	Corn	Zea m 1	Beta-expansin 1 a	269
Aero Allergen Animal				
Canis familiaris	Dog	Can f 2	Lipocalin	180
Equus caballus	Horse	Equ c 3	Preproalbumin	607
Felis catus	Cat	Fel d 1	Beta chain	20
Rattus norvegicus	Rat	Rat n 1	Alpha-2u-globulin	179
Aero Allergen Mite				
Acarus siro	Mite	Aca s 13	Lipid binding protein	64
Dermatophagoides farinae	House dust mite	Der f 1	Cysteine proteinase	321
Dermatophagoides pteronyssinus	House dust mite	Der p 3	Trypsin-like protein	261
Dermatophagoides pteronyssinus	House dust mite	Der p 4	Alpha-Amylase	496
Lepidoglyphus destructor	Storage mite	Lep d 10	Tropomyosin	284
Tyrophagus putrescentiae	Dust mite	Tyr p 2	Group 2 allergen	141
Aero Allergen Fungi				
Alternaria alternata	Fungus	Alt a 1	Major allergen 1	157
Aspergillus fumigatus	Fungus	Asp f 2	Metalloprotease	310
Apergillus niger	Fungus	Asp n 14	Xylosidase	804
Davidiella tassiana	Fungus	Cla h 4	60S acidic ribosomal protein P2	111
Aero Allergen Insect				
Blattella germanica	German cockroach	Bla g 2	Aspartic protease-like	352
Periplaneta americana	American cockroach	Per a 7.0102	Tropomyosin	284
Food Allergen Plant				
Apium graveolens	Celery	Api g 1	PRB Bet v 1 family	154
Arachis hypogaea	Peanut	Ara h 4	Glycinin	530
Corylus avellana	European hazelnut	Cor a 1.0403	Cor a 1 (Bet v 1 homologue)	161
Daucus carota	Carrot	Dau c 4	Profilin	134
Glycine max	Soybean	Gly m 3	Profilin1	131
Malus x domestica	Apple	Mal d 1	PR-10 (Bet v 1 homologue)	159
Persea americana	Avocado	Pers a 1	Endochitinase	326
Prunus avium	Cherry	Pru av 3	Lipid transfer protein	117
Solanum tuberosum	Potato	Sola t 4	Serine protease inhibitor 7	221
Vitis sp.	Grape	Vit v 1	Nonspecific lipid-transfer protein P4	37
Venom Salivary Allergens				
Apis cerana	Indian honeybee	Api c 1	Phospholipase A 2	134
Polistes annularis	Paper wasp	Pol a 2	Hyaluronidase	367
Solenopsis invicta	Red fire ant	Sol i 4	Venom allergen IV	117
Vespula maculifrons	Wasp	Ves m 1	Phospholipase A1	300
Contact Allergens				
Hevea brasiliensis	Para rubber tree	Hev b 8	Profilin, chain A	131
Hevea brasiliensis	Para rubber tree	Hev b 6	Prohevein	204

spectrum of allergens. The fragments of allergens which bind to the T helper cells or B cells are called epitopes. Molecules of major antigens may possess 2–5 major epitopes, the total number of epitopes per allergen molecule may be much higher. Each epitope includes 8–15 amino acid residues. The allergen molecules have no physical or chemical features which distinguish them from any other proteinic antigens except that they are relatively small, and in some individuals give rise to initiation and maintenance of IgE production. Several hundred major allergen

molecules have been isolated and characterized up to now, and they have been shown to possess different biological properties, such as enzymes or to be involved in transport, reproduction or secretion [115, 116]. A comprehensive list of allergens whose amino acid sequences are known at present can be found in [117]. The nomenclature follows the recommendation of the WHO. The name of an allergen is formed out of the first three letters of the latin word, followed by the first letter of the accompanying adjective and then numbered. Some of the most common allergens are listed in Table 1.

4.2. Raw Materials

Raw materials for allergen immunotherapy are the biologic materials containing the allergenic components. The allergen raw material should be selected from relevant sources. Instructions for harvesting, storage, extraction, and purification have been described by the International Union of Immunologic Societies (IUIS) [118], in the Nordic Guidelines [119] and by the European Union [120] in order to achieve good batch-to-batch consistency of allergenic potency. Manufacturing of allergen extracts should be based on The Code of Good Manufacturing Practice (GMP) [121]. Figure 1 shows the production process of allergenic products.

4.3. Allergen Standardization

A prerequisite for standardization of allergen preparations extracted from well-defined source materials is that the content of major and minor allergens and the relation between these allergens is constant from batch to batch [122]. The most common vaccines used in clinical allergy practice are now available as standardized products. However, for some vaccines standardization is neither feasible nor economically possible. Several reference standards of widespread allergens containing defined amounts of relevant allergens are maintained under stable conditions [104]. These standards have been produced as part of the WHO/IUIS allergen standardization program for a number of vaccines [123]. Based on the precision of methods used for estimation of relative potency, the FDA and the Nordic Council on Medicine require a maximum variation in activity between batches of the same allergen from 50 to 200 % of the labeled value [104].

Methods of Standardization. The former classification in protein nitrogen units or weight/volume estimates is no longer used. Standardized products have been tested for potency both in vivo and in vitro using methods for detection of IgE antibodies to allergens. Using the technique of skin testing, allergen vaccines can be defined in biological units. The size of the allergen induced wheal is compared to a histamine reference solution (1 mg/mL). The allergen solution which produces a medium wheal in sensitized persons receives one HEP/mL (histamine equivalence prick) [119]. For representative results, however, this method depends on the availability of allergic and, moreover, comparable patients. The inhibition of the binding capacity of IgE antibodies is determined with methods derived from radioallergosorbent test (RAST)-inhibition. Those tests for measurement of allergenic potency are required by the European Pharmacopoeia [115]. International reference extracts are withheld by the WHO. Manufacturers rely on definite values for quality controls. One international reference value is defined as 100 000 international units (IU). Standardized extracts show improved consistency to nonstandardized ones. This holds true both for extracts produced by different manufacturers and different batches produced by the same manufacturer [124]. The biologic effects of standardized extracts have been tested and thus the range of major allergen levels is much narrower than in nonstandardized products. However, some variation occurs with in vivo and in vitro tests, and it may be difficult to compare total allergenic potency of vaccines produced by different manufacturers. The composition of the vaccine can be determined by methods such as enzyme-linked immunosorbent assay (ELISA) techniques, sodium dodecyl sulfate – polyacrylamide gel (SDS – PAGE) electrophoresis, isoelectric focusing, IgE immunoblotting and crossed immunoelectrophoresis/crossed radioimmunoelectrophoresis (CIE/CRIE) [125].

Current Analysis Techniques. DNA and protein analysis offers excellent opportunities for standardization. The selection of the relevant

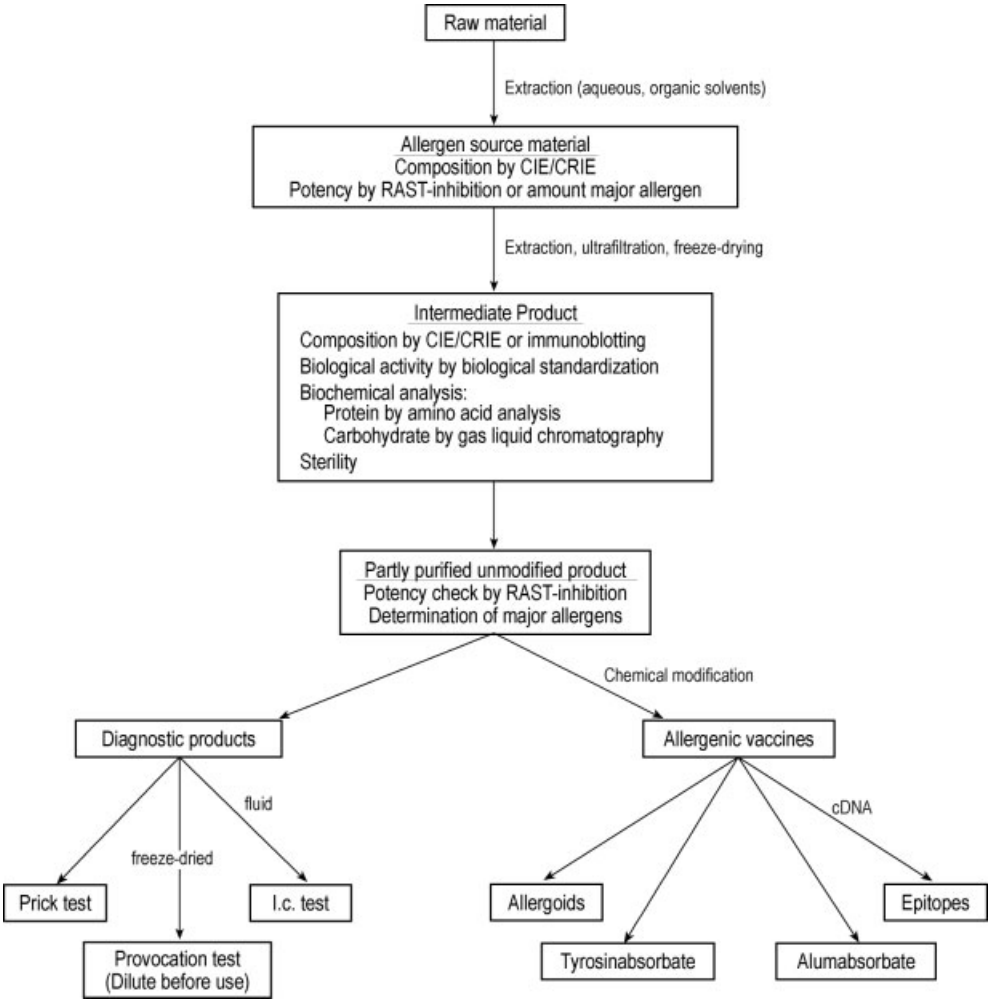

Figure 1. Production process of allergenic products according to [122]
CIE = crossed immunoelectrophoresis; CRIE = crossed radioimmunoelectrophoresis; RAST = radioallergosorbent test

disease-causing major allergens is best achieved by recombinant DNA technology [126]. The most important allergens such as pollens, mites, animal dander and insects have been cloned, sequenced and expressed. They show comparable IgE antibody binding to their natural counterparts and excellent reactivity in skin and in in vitro testing [127, 128]. With DNA and protein analysis allergen vaccines are characterized in terms of content of major allergen (nanograms or micrograms), and the consistency of each lot can be accurately monitored. Other techniques include high-performance liquid chromatography, capillary electrophoresis, and mass spectrometry [124].

Units. A critical problem in determining total allergenic potency, whether based on biologic or on in vitro assays, is the use of varying units over the years. Only two systems have been supported by authorities, i.e., the Nordic Biological Unit (BU/mL, former HEP) and the U.S. Bioequivalent Allergy Unit (BAU/mL). Both measuring methods are based on the skin sensitivity of sample patients clinically sensitive to the allergens in question. The unit BU/mL is similar and reproducible in different regions of Europe. 1 BU of common inhalant allergen has been shown to be equal to 1 ng ($\pm$ a factor of 2) of the most important major allergens [129]. In practice manufacturers use different terms for concentrations: TE

(therapeutical unit–concentration in three steps from 50 to 5000 TE/mL) is common besides SQ-E (standardized qualitiy units–concentration in four steps from 100 to 100 000 SQ-E/mL), AUeq (allergy units equivalent–concentration 20 000 AUeq/mL), and SU (standardized units–concentration in three steps from 300 to 2000 SU/mL). All extracts with common allergens are standardized according the content of the major allergen. Yet there are no obligatory prescriptions for the total major allergen dose. Concentrations differ between manufacturers and therefore a switch among products should be avoided.

4.4. Allergen Vaccines for Immunotherapy

Allergen vaccines are manufactured in different preparations. Formerly, aqueous allergen vaccines were most commonly employed for antiallergic immunotherapy, mainly for venom and inhalant allergies. Depot and modified vaccines have been developed in an attempt to make immunotherapy more effective and reduce side effects. The principle of preparing modified vaccines is to reduce or remove allergenicity, e.g., the capacity to induce IgE-mediated reaction. At the same time the immunogenicity, which is the capacity to modulate the immune system and maintain clinical efficacy should be preserved or increased.

Modification. For safer and more efficient therapy, extracts are prepared as depot vaccines in different varieties. *Physical modification* means the adsorption of allergens on aluminum, calcium phosphate, or tyrosine, or trapping of allergens in liposomes. Allergenic reactivity concerning adverse reactions can be reduced by *chemical modification* of allergens to the so-called allergoids, such as formaldehyde-, glutaraldehyde- and alginate-modified vaccines. Combinations of physically and chemically modified vaccines include tyrosine-adsorbed glutaraldehyde-modified vaccines and aluminum hydroxide adsorbed- formaldehyde-modified vaccines. A new technique is the *conjugation of allergen* to immunostimulatory bacterial molecules. Monophosphoryl-Lipid A (MPL®) has been introduced as such a novel adjuvant since 2001. The preparation derived from the lipopolysaccharide of *Salmonella Minnesota* R595 and has Th1-inducing effects [130, 131]. This effect is long-lasting and thus the number of vaccine injections can be significantly reduced.

Allergen mixtures present several problems. Excessive dilution by multiple allergens may result in suboptimal doses of individual allergens. Further, the potency of individual allergens may deteriorate more rapidly when diluted or mixed with other allergen vaccines. Some allergens possess enzymatic activity which can alter the composition of other allergens [104]. Allergen vaccines should be distributed as either vaccines from a single source material or mixtures of related cross-reacting allergen vaccines. For any mixture of allergen vaccine stability data and data on clinical efficacy should be available [104].

4.5. Storage

Allergen vaccines are protein solutions that are labile at elevated temperatures. Therefore, sterilization can only be achieved by filtration. Solutions may be destroyed by microbial and enzymatic degradation, so they routinely contain preservatives and should be stored in a refrigerator between 2 and 8 °C. The most common preservative is phenol in concentrations of 0.2–0.5 %.

Concentrated aqueous extracts in 50 % glycerin that are only to be used for skin prick testing remain stable for about three years if stored at 4 °C. Diluted extracts lose their potency more rapidly [132]. To avoid loss of potency by adsorption of active components on the surfaces of containers and syringes, Tween 20 (an ethoxylated sorbitan ester) and human serum albumin have been employed. Vials should be labeled in accordance with the requirements of regulatory authorities. Labeling includes a designation of relevant units (see Units) obtained by an approved method.

4.6. Uses

Allergenic extracts are used for diagnosis and as vaccines for treatment.

4.6.1. Diagnostic Use

Skin testing with concentrated solutions of allergens is used to detect the individual sensitivities of an allergic patient. The most common techniques are the prick test and the intracutaneous test. The *prick test* is performed by puncturing the skin with a sterile lancet through a small drop of the allergenic extract. *The intracutaneous (i.c.) test* consists of injecting a minute amount (ca. 0.02 mL) of an allergen extract into the superficial layers of the skin. This method is more sensitive than the prick test, and the concentration of the test solution must be 10 – 100 fold lower than that for prick testing. In cases of sensitivity both tests will result in the development of a wheal and erythema reaction within 15–20 min. To confirm clinical relevance of an allergen in selected cases the nasal and/or bronchial provocation test is performed. A small amount of adequately diluted allergenic extract is instilled into the nose or inhaled as nebulized aerosol. A positive reaction is indicated by typical allergic symptoms and can be quantified by special measurement procedures.

4.6.2. Therapeutic Use—Immunotherapy

Immunotherapy has been accepted as an efficient treatment for allergic rhinitis and hay fever since the early 20th century. Besides allergen avoidance immunotherapy is the only chance that may modify the natural course of the disease, e.g., the development from rhinitis to asthma. Nevertheless, different points of view exist concerning immunotherapy and asthma. According to the consensus report on diagnosis and management of asthma of the National Institutes of Health immunotherapy for asthma is only rarely used in the United Kingdom or even not recommended following the British Thoracic Society guidelines [133, 134]. Immunotherapy treatment of allergic diseases including asthma is widely accepted in the USA and in European countries except Scandinavia.

Immunotherapy is absolutely indicated for patients who had severe IgE-mediated anaphylactic reactions to bee and wasp stings [135]. Hymenoptera venom is the only allergen for which an efficacious maintenance dose has been securely established (e.g., 100 μg of purified venom, corresponding to about two bee stings or five to ten wasp stings) [136]. Immunization with venoms regularly results in essentially complete protection [105, 106, 137].

Mechanism of Immunotherapy. In former years the efficacy of immunotherapy was measured by decrease of IgE or increase in IgG antibodies. In the mid-1990s it has been suggested that immunomodulation plays a major role in immunotherapy. Immunotherapy alters the balance of cytokines released from T helper lymphocytes in the respiratory tract, with a shift from the T_H2 cells—that release interleukin-4 (IL-4) and interleukin-5 (IL-5) in association with allergic inflammation—to T_H1 cells that release γ-interferon, which inhibit T_H2 cells [136, 138]. Successful immunotherapy might then be associated with a shift in IL-4/γ-interferon production either as a consequence of down regulation T_H2 cells responses or increased T_H1 cells responses. IL-10 has been found to play an important role. Subsequent to immunotherapy allergen-specific IL-10 will increase even in the respiratory mucosa. IL-10 mediates the immunotolerogenic effects by inducing the B cells to produce IgG4, by inhibition of IgE-dependent mast-cell activation and inhibition of eosinophil cytokine production and suppression of IL-5 [139]. Different mechanisms may be valid in venom immunotherapy in nonatopic persons and in immunotherapy with inhalant allergens in atopic patients. Until now no measurable parameters of patient characteristics or symptom severity have been found that can predict or correlate with a successful outcome [132].

Administration. Immunotherapy is started with very small amounts of allergen vaccine injected into the arm every two weeks. The allergen concentration is slowly increased until the maximum tolerated dose is reached. The intervals of injection may then be prolonged to about four weeks. A maintenance dose of approximately 5 to 20 μg of major allergen (ragweed, grass, mite, cat) is associated with significant improvement in patient symptom scores [124]. In venom immunotherapy the usually recommended maintenance dose is 100 μg of venom protein. In pollen immunotherapy a reduced dose should be given during season to avoid serious reactions. Local reactions occur very often at the injection site, early reactions

within 20–30 min, some later than 30 min. Systemic reactions with generalized symptoms are the worse the earlier they occur. To minimize the risk of immunotherapy each patient should stay under supervision for at least 30 min after injection and every immunotherapy-practicing doctor has to be trained in the treatment of allergic shock and resuscitation [140].

New ways of immunotherapy, for instance oral, sublingual or local application of allergen vaccine have been tested . The sublingual way seems to be the most promising for allergic rhinitis and asthma. Efficacy and safety have been proven for children and adults in many trials. Compared to the subcutaneous route a much higher amount of allergen is necessary [139, 141]. A disadvantage might be that there is less control by the treating physician because patients take the needed dose at home. Moreover there are no experiences about long-time efficacy comparable to the decades of subcutaneous application.

Contraindications for immunotherapy include malignant and in immunodeficiency diseases, treatment with β-blockers, poor compliance, severe asthma, and significant cardiovascular diseases. During pregnancy treatment should not be started but may be continued.

4.6.3. Legal Aspects, Quality Requirements, and Safety

Only standardized extracts should be used for diagnosis and treatment. Potency should be quoted in allergy units or in micrograms of a major constituent whenever possible [137]. In the USA the registration of allergenic products is controlled by the FDA Center for Biologics Evaluation and Review (CBER), where a new Division of Allergenic Products and Parasitology (DAPP) deals with development, controlling, use, and safety of allergenic products [142]. Since 1998 in the EU allergenic extracts are registered by the Paul-Ehrlich-Institut (PEI–Bundesamt für Sera und Impfstoffe) which is also a WHO-Collaboration Center. The PEI is responsible for the examination and registration of allergenic extracts for diagnosis and therapy. Allergen products should be manufactured under conditions which comply with GMP.

Guidelines for immunotherapy with inhalant allergens and venoms have been published in the past ten years by the WHO, the European Academy of Allergy and Clinical Immunology (EAACI), the International Consensus Report on Asthma, the Global Strategy for Asthma Management and Prevention, the International Consensus Report on Rhinitis, the British Society for Allergy and Clinical Immunology (BSACI), the American Academy of Allergy, Asthma and Immunology (AAAAI), the American College of Allergy, Asthma and Immunology (ACAAI) , the International Union of Immunological Societies (IUIS) and the Deutsche Gesellschaft für Allergologie und Immunologie (DGAI) [104, 121, 124, 134, 135, 140, 143]. In immunotherapy the incidence of severe reactions is low. The mortality rate during immunotherapy is about one death per 2×10^6 injections [136]. A study carried out in the Mayo Clinic over a ten year period with 79 953 injections showed 0.137 % systemic reactions in an average time of 35 min after injection. The systemic reactions were equally frequent during the rising and maintenance phases [144]. Bernstein et al. reported in 2004 similar rates of fatal reactions to diagnostic or therapeutic procedures as published in previous surveys [145]. The PEI analyzed severe reactions to diagnostic and therapeutic allergen extracts in Germany in the time period from 1991 to 2000. The authors conclude that the benefits of allergy diagnosis and immunotherapy outweigh the risk of adverse reactions [146]. Diagnosis and therapy of allergic diseases should only be practiced by specially trained physicians. Extreme caution is necessary in treating any asthmatic individual [121, 143].

5. Anti-IgE-Omalizumab

The knowledge about mechanisms of the allergic reaction has opened new strategies for treatment of allergic diseases. A novel therapeutic principle in treating IgE-mediated allergic diseases forms the introducing of anti-IgE antibodies. The substance (former name rhuMAb-E25, new nomenclature omalizumab) is a recombinant humanized monoclonal antibody (see → Monoclonal Antibodies). The trade name is Xolair (Novartis).

5.1. IgE and its Role for the Allergic Response

The IgE was first identified in the 1960s. In the normal individual there are usually low levels of circulating IgE. The IgE is part of the physiologic mucosal defence in parasitic infections. In allergic patients TH-2 lymphocytes release cytokines under the influence of allergens which induce B cells to produce allergen-specific IgE. The priming of B cells is located mainly in the lymphatic tissue of the gastrointestinal and respiratory tract. The IgE molecule (Fig. 2) is built out of two identical heavy (H) and two light (L) chains, which are separated in variable (V) and constant (C) parts. All four chains form a Y-like molecule, whose upper part consists of two double and variable H and L chains. This part binds the specific antigen and is called Fab_2 (fragment antigen binding).The opposite site – called Fc (fragment crystallisable)–is formed out of two H chains and attaches to the cell-bound receptors. There are high- (FcεRI) and low-affinity (FcεRII) receptors on mast cells, basophils, other blood cells, and tissue cells in target organs. High-affinity receptors allow a sensitization of the individual even with low amounts of allergen. This effect lasts more longer than the normal two to three days half-life of the free circulating IgE. The FcεR receptor consists of an IgE-binding a chain, one β chain and two γ chains (Fig. 3). In an allergic individual the first contact with an allergen leads to a specific sensitization and an IgE formation. In further contacts with the disease-causing allergen the allergen proteins bind to the receptor–IgE complex in form of a cross-linking. By this reaction basophils degranulate immediately and release preformed substances such as histamine and TNF-α. On mast cells and tissue cells the bridging reaction starts to synthesize other substances such as leukotrienes, cytokines and prostaglandins, which are the main mediators of the late allergic response [4, 5].

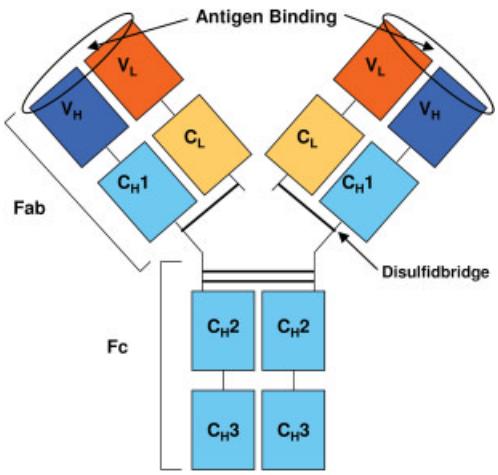

Figure 2. Structure of the IgE molecule (V=variable, C=constant, H=heavy, L=light). With courtesy of S. Ewert, Novartis.

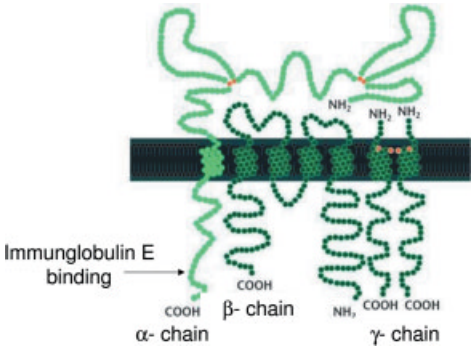

Figure 3. The IgE receptor on basophiles. Modified according to [5].

5.2. IgE-Mediated Allergic Diseases

In sensitized individuals allergens cause intermittent or chronic diseases of the skin, eyes, airways, gastrointestinal tract and sometimes other organs. Allergic rhinitis and asthma are the most common allergen-induced diseases, mainly due to inhalant and less often to food allergens. It is estimated that 2–15 % of the European population suffer from asthma, in some countries allergy may affect over 50 % of children [147]. The release of IgE and other mediators of the allergic reaction cause a constant and chronic inflammation, which in the case of asthma will lead to often hospitalization, destruction of lung parenchyma and loss of quality of life. In Great Britain every second patient suffers from severe asthma symptoms. More than two millions of patients have a bad controlled asthma despite of comprehensive therapy [148]. Another important entity of allergic diseases are the anaphylactic IgE-mediated disorders. The individ-

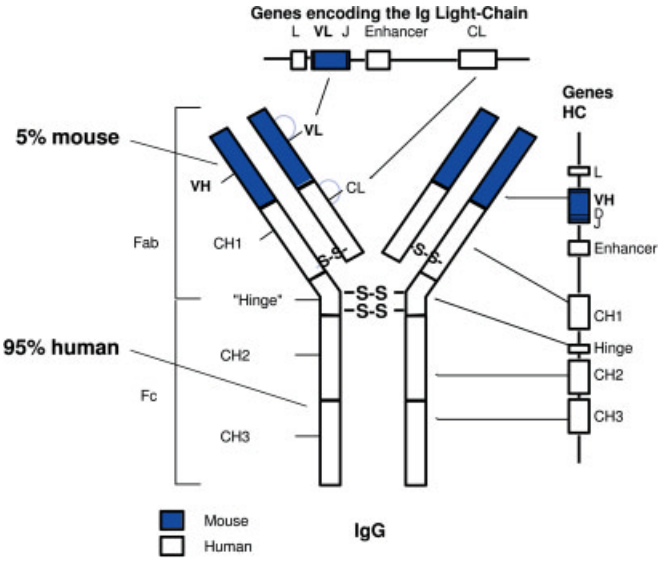

Figure 4. Molecular structure of monoclonal antibody. Genes encoding the Ig L and H chains are shown on top and on the right side, respectively. With courtesy of B. Gysin, Novartis.

ual has been sensitized by an insect sting without having any following chronic disease. But the next venom sting may lead to an immediate life threatening anaphylactic reaction. A similar reaction may occur in patients who have been sensitized to some kind of drugs for instance penicillin.

5.3. Properties of Omalizumab

In the year 1992 a recombinant monoclonal anti-IgE antibody was cloned for the first time. A decade passed till the substance was ready for the commercial use in human beings. Omalizumab is a recombinant DNA-derived humanized monoclonal antibody (see → Monoclonal Antibodies) that selectively binds to human IgE. Omalizumab does not bind to other immunoglobulins such as IgG or IgA.

5.3.1. Chemistry and Production of Omalizumab

The first step in the generation of anti-IgE antibodies has been immunizing mouse with human IgE. The DNA of selected murine anti-IgE producing B cells and the DNA of human immunoglobulin IgG producing B cells are matched together and form a recombinant DNA. The selected hybridoma cell line produces the specific monoclonal antibody which subsequently was cloned into an industrial producer cell line. This cell line produces an anti-IgE antibody which is only 5 % of murine and 95 % of human origine (Fig. 4). The humanizing of the anti-IgE antibody achieves tolerance of the human immune system. The manufacturing process takes place in large bioreactors (Fig. 5). The final stage of the cultivation procedure needs fourteen days (Fig. 6). Purification is performed by several downstream processing steps (Tab. 2). Important factors of the produced antibody solution such as identity, content, purity, stability, biological activity and sterility are checked with

Figure 5. Pilot-scale Bioreactors at Novartis. With courtesy of B. Gysin, Novartis.

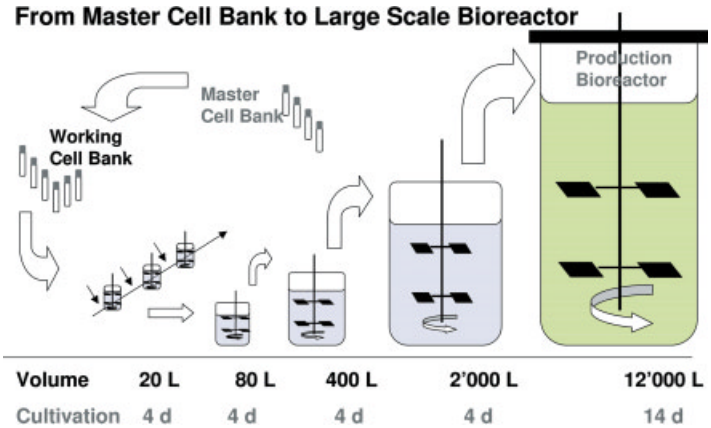

Figure 6. Cultivation of monoclonal antibodies. The process from master cell bank to large-scale bioreactor is shown. With courtesy of B. Gysin, Novartis.

Table 2. Downstream processing of monoclonal antibodies. With courtesy of B. Gysin, Nova.

MAB XY	Reasons for STEP
Spin filter or centrifugation and/or expanded-bed cation exchange	Cell-product separation, separation of product from cell debris and low-molecular-weight compounds
Virus inactivation	Inactivation (low pH) and removal (filtration)
Protein A	IgG selection
Anion exchange	Removal of DNA
Cation exchange	Removal of variants and contaminating protein A
Ultrafiltration	Concentration
Filling freezing	Storage

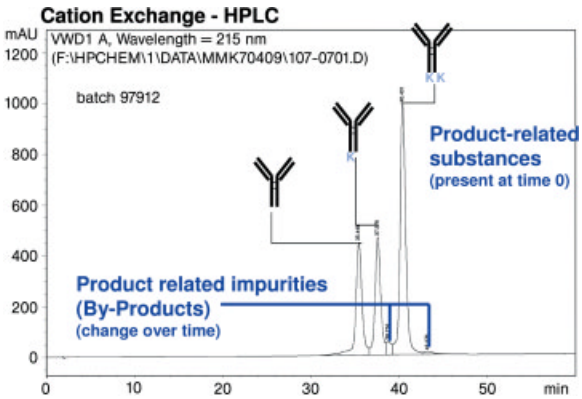

Figure 7. The analytical characterization by cation-exchange HPLC reveals the molecular heterogeneity of monoclonal antibodies. With courtesy of B. Gysin, Novartis.

Table 3. Analytical methods for biopharmaceuticals. With courtesy of B. Gysin, Novartis.

Identity	
	• MALDI-TOF mass spectrometry • Amino acid sequencing • Isoelectric focussing (IEF & cIEF), • Western blotting
Content	
	• Amino acid analysis • Size-exclusion chromatography (SEC) • Affinity chromatography • UV Spectroscopy
By- and degradation products	
	• HPLC (size exclusion, ion exchange, hydrophobic interaction, reversed phase) • CE • SDS PAGE
Process related impurities	
	• ELISA for protein A • Host cell protein assay
Aggregates	
	• SEC, LLS
Thermodynamic stability	Differential scanning calorimetry (DSC)
Secondary structure	Fourier transform infrared (FTIR), circular dichroism (CD), second derivative UV absorbance spectroscopy
Adventitious agents	Endotoxins, viruses, DNA, mycoplasms
Biological activity	ELISA, cell assays, Biacore
Sterility	MLT etc

different analytical methods (Tab. 3). Changes of the molecular heterogeneity that may occur over time are analyzed with cation-exchange HPLC (Fig. 7) The produced anti-IgE antibody is not specific for a definite allergen-induced IgE. It binds to any IgE in human blood and on cells.

5.3.2. Pharmacology and Effect of Omalizumab

There are four effects of omalizumab [149, 149 – 152]:

1) Omalizumab binds to the Cε3 domain of the IgE molecule which aggregates with the high-affinity receptor on mast cells and basophils. Thus the binding of the IgE molecule to the cell receptors is blocked and the release of mediators inhibited. Omalizumab does not interact with already bound IgE on FcεR receptors, therefore no degranulation or forming of mediators will be induced.
2) Omalizumab binds to the free IgE and forms molecular complexes. The relation of the free IgE to the concentration of omalizumab determinates the forming of trimers or hexamers (Fig. 8). The molecular complexes will have a molecular weight from 490 to 1000 kD. The application of omalizumab immediately leads to a reduction of the free IgE, but the values of total IgE, which include the measuring of the complexes will rise. The complexes have no pathogenic meaning. The most important step for suppressing the allergic reaction is the decrease of the free IgE.
3) Under the treatment with omalizumab the number of receptors on basophils and mast cells decreases significantly. Thus the capacity of IgE binding is reduced. This effect will last for several months.
4) Omalizumab plays also an important role in immunomodulation. The number of receptors on other cells involved in the allergic reaction such as tissue and dendritic cells decreases as well. This effect will have a long-lasting suppression of the induction and stimulation of the T_H2 response. Recent work showed significant reduction of nasal and bronchial eosinophils, bronchial mast cells, T cells, and B cells [149].

Though the free IgE decreases immediately after application of omalizumab the clinical effect will only occur after several weeks. In all treated allergic individuals a release of symptoms will be achieved within 16 weeks. The efficacy of omalizumab is only warranted under application of the dose, symptoms of allergic disease will relapse after end of therapy.

5.3.3. Pharmacokinetics

The humanizing of the murine antibody assures immunogenic tolerance by the human immune system. The average absolute bioavailability of omalizumab after subcutaneous administration is 62 %. In a clinical study of adults with allergic rhinitis omalizumab was slowly absorbed after subcutaneous administration of 0.15 mg/kg and resulted in peak serum concentrations of 2 mg/L after 14 days. The substance is cleared slowly from circulation, with a mean $\pm$S.D. apparent clearance of 2.4$\pm$1.1 mL/L/kg/d, resulting in a terminal elimination half-life of one to four weeks [150]. Omalizumab has an IgG structure and is delayed eliminated by the reticuloendothelial system of liver endothel cells. The half-life of omalizumab will be 19 to 22 days on average. After withdrawal of omalizumab the former IgE level will be reached within 12 months [149]. There is no rebound effect. The IgE – omalizumab complexes are not pathogenic, they do not accumulate in the kidney and they do not bind complement.

5.4. Indication for the Teatment with Omalizumab, Guidelines for Use

Immunoglobulin E plays a key role in allergic diseases. Omalizumab offers the opportunities to stop the cascade of the allergic reaction in a very early stage. Omalizumab inhibits the early and late reaction of inflammation. There might be a wide range of indications but till now mainly the treatment of patients with allergic asthma and rhinitis [149, 152] has been evaluated in studies. In spite of full therapy there remain worldwide about 26 % (Europe) to 42 % (UK) of patients with insufficient or bad controlled asthma. Those instable patients are at an increased risk of hospitalization, visits in emergency rooms or asthma death. In several studies including

patients with medium or severe asthma omalizumab reduced significantly exacerbations and hospitalization, the dose of additional inhalant glucocorticosteroids could be decreased [153, 154]. Omalizumab was found to be an effective add-on therapy for difficult-to-treat asthma patients, who have an important unmet medical need despite best available therapy. Quality of life ameliorated significantly [155]. The efficacy of omalizumab was the better the severe the degree of asthma [149, 153, 154].

The recommendation of the EMEA [156] for the registration of omalizumab follows the indications approved till present:

→ Omalizumab is indicated as add-on therapy in adult and adolescent patients (12 years of age and above) with severe persistent allergic asthma who have a positive skin test or in vitro reactivity to a perennial aeroallergen and who have reduced lung function ($FEV_1 < 80\,\%$) as well as frequent daytime symptoms or nighttime awakenings and who have had multiple documented severe asthma exacerbations despite daily high-dose inhaled corticosteroids, plus a long-acting inhaled beta-2 agonist.
→ Omalizumab treatment should only be considered for patients with convincing IgE-mediated asthma
→ It is proposed that omalizumab is initiated by physicians experienced in the diagnosis and treatment of severe persistent asthma

5.5. Application, Dosage, Costs

The concentration of the free IgE has to fall under 20.8 IU/mL (50 ng/mL) before expecting any clinical effect. Omalizumab has to be given in a relation of 15 to 20:1 to the absolute number of IgE molecules. Before starting treatment the needed omalizumab concentration is measured out of the total IgE concentration and the body weight. The dosage has to be determined individually for each patient, at least 0.016 mg/kg/IgE [IU/mL] every four weeks [150]. The maximum dosage is 750 mg every four weeks. The dosage is limited to a maximum of 700 kU/mL IgE and to a maximum body weight of 130 kg, because in these cases the calculated amount of the substance can no more be injected. The substance is supplied as a lyophilized sterile powder in a single-use, 5-mL vial designed to deliver 150 mg (1.2 mL) of drug after reconstitution. The lyophilized product takes about 15 – 20 min to dissolve while shaking the vial. The prepared solution should be used within eight hours when stored in the vial at 2 – 8 °C or within four hours when stored at room temperature. Reconstituted omalizumab is for single use only, it contains no preservatives. The recommended dosage per injection ranges between 150 mg and 375 mg omalizumab every two to four weeks according the necessary amount of the substance. The substance has to be given subcutaneously. More than one injection site should be used if the dosage exceeds 150 mg [150]. Total IgE levels remain elevated approximately for up to one year after ending of treatment. Dosage determination during therapy has always to be based on the initial IgE level. Significant changes in body weight must be taken into account. Omalizumab costs about 540 dollars (460 €) per vial, for a one-year therapy a sum of about 6000 to 10 000 dollars will be estimated. Yet evaluating this sum one has to consider that 5 to 10 % percent of patients with the most severe form of asthma contribute some 50 % of the total costs associated with asthma [157].

5.6. Side Effects

Placebo controlled studies have shown a very good tolerability of omalizumab. In a study with children occurred mainly urticaria (1.3 %), itching, rash, and pain on the site of the injection

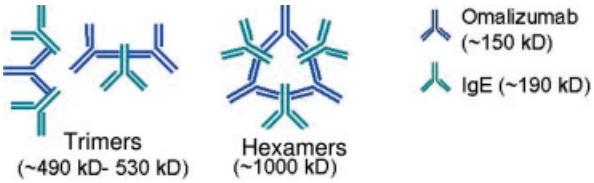

Figure 8. Omalizumab-IgE complexes. With courtesy of Novartis.

(each 0.4 %) [158]. In several controlled studies with adults there were no differences in adverse events between the omalizumab and control groups [149]. Till present in clinical studies three not severe anaphylactic reactions (< 0,1 % of treated patients) were observed. Symptoms were urticaria, throat, or tongue edema [150]. There are no special substance related severe side effects. The IgE–omalizumab complexes are small, no serum sickness or immune complex nephropathy will be induced. A small percentage (< 0,1 %) of patients may develop antibodies to omalizumab.

5.7. Legal Aspects, Safety

Omalizumab is currently approved for moderate-to-severe persistent allergic asthma in the USA, Canada, and Brazil, for moderate persistent allergic asthma in Australia, and for severe persistent allergic asthma in New Zealand [149]. Till the end of 2005 omalizumab will be registered in Germany. The official use of the substance is restricted for definite criteria following the guidelines (Section 5.4). Detailed conditions for the use of this product will be described in the Summary of Product Characteristics (SPC) which will be published in the European Public Assessment Report (EPAR) and will be available in all official European languages after the marketing authorization has been granted by the European Commission [156]. The Global Initiative for Asthma (GINA) recommends treatment with omalizumab in her recent guidelines [148]. Because of the limited data regarding the safety and effectiveness in children, the official use of omalizumab is recommended for patients 12 years of age or older. Safety and efficacy have not been established for patients with a body weight or a serum IgE outside the recommended range. During pregnancy omalizumab should only be applied when clearly needed. In breast-feeding women caution is recommended, because in animal studies omalizumab was excreted in the milk [150].

5.8. Possible Applications in the Future

Omalizumab is an innovative and unique form of treatment, because it offers preventive therapy of IgE-mediated disorders. Future aspects for indication with omalizumab could be seasonal allergic airways disease, atopic eczema with high IgE values, venom allergy, food allergy, and desensitization in high-reactive patients. In a study of patients with peanut allergy the treatment with anti-IgE (TNX-901, similar to omalizumab) gave a protection against small amounts of unintended ingestions of peanuts [159]. The therapy will not be given widespread because of the relatively high costs, but further experiences will be gathered within the next years. Omalizumab will surely offer new interesting perspectives of antiallergic therapy.

6. Acknowledgement

Thanks to Drs. S. Ewert and B. Gysin from Novartis for giving scientific advice.

7. References

General References

1. J. Ring, G. Burg: *New trends in allergy. II.*, Springer Verlag, Berlin – Heidelberg – New York 1986.
2. T. Holgate, M. K. Church: *Allergy*, Mosby-Wolfe, London 1995.
3. G. Schultze-Werninghaus, T. Fuchs, C. Bachert, U. Wahn: *Manuale allergologicum*, Dustri, München-Orlando 2004.
4. G. Schultze-Werninghaus, T. Fuchs, C. Bachert, U. Wahn. *Manuale allergologicum*. Dustri-Verlag, München-Orlando 2004.
5. N. Mygind, R. Dahl, S. Pedersen, K. Thestrup-Pedersen. *Allergologie*. Blackwell Berlin-Wien 1998.

Specific References

6. R. R. A. Coombs, P. G. H. Gell: "Classification of allergic reactions responsible for clinical hypersensitivity and disease," in P. G. H. Gell, R. R. A. Coombs, P J. Lachmann (eds.): *Clinical Aspects of Immunology*, Blackwell Scientific Publications, Oxford 1975, p. 761.
7. I. Tegeder in K. Brune (ed.): *Pharmacologie und Toxikologie Transparent*, Chapman & Hall, London 1996, pp. 32 – 35.

8. A. Windaus, W. Vogt, *Ber. Dtsch. Chem. Ges.* **3** (1907) 3691 – 3695.
9. H. H. Dale, P. P. Laidlaw, *J. Physiol. (London)* **41** (1910) 318.
10. C. H. Best, H. H. Dale, H. W. Dudley, W. V. Thorpe, *J. Physiol.* **62** (1927) 397 – 417.
11. P. G. Bardin, S. T. Holgate: "Antihistamines," in C. P. Page, W. J. Metzger (eds.): *Drugs and the Lung*, Raven Press, New York 1994, pp. 257 – 280.
12. E. Fourneau, D. Bovet, *Arch. Int. Pharmacodyn. Ther.* **46** (1933) 178.
13. D. Bovet, A. M. Staub, *C. R. Séances Soc. Biol. Ses. Fil.* **124** (1937) 547.
14. A. M. Staub, *Ann. Inst. Pasteur Paris* **63** (1939) 400, 420, 485.
15. H. Giertz, L. Flohe: "Mediatoren der Entzündung. Pharmakotherapie der Allergie; Arzneimittelallergie," in W. Forth, D. Henschler, W. Rummel (eds.): *Allgemeine und spezielle Pharmakologie und Toxikologie*, BI Wissenschaftsverlag, Mannheim – Wien – Zürich 1987, pp. 176 – 215.
16. A. F. Casy: "Antagonists of H1 receptors of histamine: recent developments," in B. Uvnäs (ed.): *Histamine and Histamine Antagonists*, Springer Verlag, Berlin – Heidelberg – New York 1991, pp. 549 – 572.
17. C. R. Ganellin, M. E. Parsons (eds.): *Pharmacology of Histamine Receptors*, Wright, Boston 1982.
18. B. N. Halpern, *Arch Int. Pharmacodyn.* **68** (1942) 339 – 408.
19. G. Ehrhart, E. Lindner: "Antihistaminica," in G. Ehrhart, H. Ruschig (eds.): *Arzneimittel*, 2nd ed., vol. **1**, Verlag Chemie, Weinheim 1972.
20. Ciba, US 2 449 241, 1945.
21. Merck & Co., US 2 607 778 ,1952.
22. R. C. Clapp et al., *J. Am. Chem. Soc.* **69** (1947) 1549.
23. BASF, GB 659 730, 1949.
24. C. P. Hutterer, C. Djerassi, W. L. Beears, R. L. Mayer, C. R. Scholz, *J. Am. Chem. Soc.* **68** (1946) 1999.
25. Monsanto, US 2 581 868, 1946.
    A. W. Weston, *J. Am. Chem. Soc.* **69** (1947) 980.
26. Rhône-Poulenc, FR 913161, 1941.
27. E. Campaigne, W. E. Lesner, *J. Am. Chem. Soc.* **79** (1949) 33.
28. Pyridium Corp., US 2 465 865, 1946.
29. Ciba, US 2 406 594, 1943.
30. G. R. Rieveschl, W. F. Huber, *Abstr. 109th Meet. Am. Chem. Soc.* 1946, 50 K.
31. Parke-Davis, US 2 527 963, 1948.
32. McNeil, US 2 800 485, 1955.
33. A. Ebnoether, H. P. Weber, *Helv. Chim. Acta* **7** (1976) 2462.
    F. E. Roth, *Chemotherapia* **3** (1961) 120.
34. N. Sperber et al., *J. Am. Chem. Soc.* **71** (1949) 887.
35. L. Novak, M. Protiva, *Coll. Czech. Chem. Commun.* **24** (1959) 3966.
36. M. Protiva et al., *Chem. Listy* **43** (1949) 257.
37. L. C. Cheney et al., *J. Am. Chem. Soc.* **71** (1949) 60.
38. Hoffmann-La Roche, US 2 879 293 ,1959.
39. A. La Belle, R. Tislow, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **7** (1948) 236.
40. R. Tislow et al., *Fed. Proc. Fed. A. Soc. Exp. Biol.* **8** (1949) 338.
41. Hoechst, DE 925 468 ,1955.
    F. E. Roth, *Chemotherapia* **3** (1961) 120.
42. Schering, US 2 567 245, 1948.
43. Schering, US 3 061 517, 1962.
44. Ciba, US 3 060 186, 1960.
    C. F. Huebner et al., *J. Am. Chem. Soc.* **82** (1960) 2077.
45. Burroughs Wellcome, US 2 712 023, 1952.
46. UCB, US 2 709 169 ,1955.
47. Burroughs Wellcome, US 2 630 435, 1948.
    R. Baltzly et al., *J. Org. Chem.* **14** (1949) 775.
48. Janssen, US 2 882 271, 1957.
49. Janssen, GB 809 760 ,1959.
50. K. E. Hamlin et al., *J. Am. Chem. Soc.* **71** (1949) 2731.
51. UCB, US 2 899 436 ,1959.
52. Cerm, DE 1 957 371 ,1970.
53. Janssen Pharm., BE 852 405, 1977.
    R. De Beule et al., *Acta Allergol.* **32** (1977) 278.
    P. Dockx et al., *Curr. Ther. Res.* **29** (1981) 510.
54. M. K. Church et al., *Agents Actions* **10** (1980) 4.
55. Knoll, US 2 683 714 ,1954.
56. UCB, BE 748 568, 1970.
    A. Brunner et al., *Lung* **154** (1977) 143.
57. Sandoz, US 2 717 251, 1952; US 2 757 175, 1954.
58. Rhône-Poulenc, US 2 837 518, 1958.
59. Rhône-Poulenc, GB 814 512 ,1959.
60. Actiebolaget Recip., GB 881 379 ,1961.
61. Degussa, DE 964 050, 1954; DE 1 001 684, 1954.
    W. A. Schuler, H. Klebe, *Justus Liebigs Ann. Chem.* **653** (1962) 172.
62. A. von Schlichtegroll, *Arzneim. Forsch.* **8** (1958) 489.
63. Sogeras, US 3 987 042 ,1970.

64. Mead Johnson, US 2945855, 1958.
Y. H. Wu, R. F. Feldkamp, *J. Org. Chem.* **26** (1961) 1529.
65. Rhône-Poulenc, US 2 972 612 ,1961.
66. Rhône-Poulenc, US 2 530 451, 1947.
Searle, US 2 687 414, 1948.
P. Charpentier, *C. R. Hebd. Séances Acad. Sci.* **225** (1947) 306; **232** (1951) 415.
67. E. Schmitt et al., *Bull. Soc. Chim. France* **1957**, 1474.
68. Rhône-Poulenc, GB 641 452 ,1950.
69. Schering, US 3 301 863, 1965;
US 3326924, 1963;
US 3 357 986, 1966;
US 3 366 635, 1963.
70. Wander, DE 1 263 774, 1960.
F. Hunziker et al., *Arzneim. Forsch.* **13** (1963) 324.
71. Sandoz, CH 544 113, 1973; DE-OS 2 111 071, 1971; DE-OS 2 144 490, 1972; DE-OS 2 302 944, 1973; DE-OS 2 302 970, 1973; DE-OS 2 605 980, 1976.
E. Waldvogel et al., *Helv. Chim. Acta* **59** (1976) 866.
72. Bayer, GB 721 171 ,1954.
73. Hoffmann-La Roche, US 2 470 108, 1949; US 2 470 109, 1949 (J. T. Plati, W. Wenner).
J. T. Plati, W. Wenner, *J. Org. Chem.* **15** (1950) 209; **20** (1955) 1412.
74. Sandoz, GB 942 876, 1963.
75. Egis, DE 2 528 194 ,1974.
76. Wellcome, EP 85 959 ,1983.
77. UCB, EP 58 146 ,1981.
78. Janssen Pharm., EP 5 318, 1979.
J. van Wauwe et al., *Arch. Int. Pharmacodyn. Ther.* **251** (1981) 39.
79. F. E. R. Simons, M. S. Kesselmann et al., *Lancet* (1988) no. 1, 624.
80. Ube, EP 335 586 ,1996.
81. Almirall, EP 134 124 ,1983.
82. J. A. Hey, W. Kreutner et al., *Allergy* **52** (1997) suppl. 37, 208.
83. Synthelabo, EP 217 700 ,1990.
84. J. Bousquet et al., *Allergy* **51** (1996) 251 – 256.
85. A. Genovese, G. Sparado, *Allergy* **52** (1997) suppl. 34, 67 – 78.
86. Asta, GB 1 377 231, 1974.
K. Tasake, M. Akagi, *Arzneim. Forsch.* **29** (1979) 488.
B. Fischer, W. Schmutzler, *Arzneim. Forsch.* **31 (II)** (1981) 1193.
I. Szelenyi et al.: "Azelastine: A Multifaceted Drug for Asthma Therapy," in G. P. Anderson, I. D. Chapman, J. Morley (eds.): *New Drugs for Asthma Therapy*, Birkhäuser, Basel – Boston – Berlin 1991, pp. 295 – 311.
D. McTavish, E. M. Sorkin, *Drugs* **38** (1989) 778 – 800.
87. Boehringer-Ingelheim, DE 3 008 944 ,1980.
88. S. P. Clissold et al., *Drugs* **37** (1989) 42 – 57.
89. M. Lindquist, I. R. Edwards, *Lancet* **349** (1997) 1322.
R. Caballero et al., *Br. J. Pharmacol.* **122** (1997) 796 – 798.
90. E. Delpon, C. Valenzuela et al., *Anales de la Real Academia de Farmacia* **62** (1996) 25 – 60.
91. Lab. de Rech. Biol. Laborec., FR 1 295 309 ,1962.
92. J. S. G. Cox et al., *Adv. Drug Res.* **5** (1970) 115.
93. M. J. Parnham, *Pulm. Pharmacol.* **9** (1996) 95 – 105.
94. R. Penner, G. Matthews, E. Neher, *Nature* **334** (1988) 499.
95. M. Reinsprecht et al., *Biochem. Biophys. Res. Comm.* **188** (1992) 957.
96. Fisons, US 3 419 578, 1968.
97. E. W. F. W. Alton, A. A. Norris, *J. Allergy Clin. Immunol.* **98** (1996) 102.
A. C. Myers et al., *Allergy Clin. Immunol.* **98** (1996) 107.
98. S. T. Holgate, *Eur. J. Resp. Dis.* **69** (1986) Suppl. 147, 149 – 159.
99. G. Ciprandi et al., *J. Allergy* **51** (1996) 946 – 951.
100. Takeda, US 4 299 963 ,1978.
101. H. Kuriki, Y. Maki, *Jpn. Pharmacol. Ther.* **16** (1988) 283.
102. L. Noon, *Lancet* 1911, no. 1, 1572.
103. J. Freeman, *Lancet* 1911, no. 2, 814.
104. J. Bousquet, R.F. Lockey, H.J. Malling, WHO position paper, *Allergy* **53** (1998) 1–42.
105. D.B.K. Golden, A. Kagey-Sobotka, P.S. Norman, R.G. Hamilton, L.M. Lichtenstein, *N. Engl. J. Med.* **351** (2004) 668–674.
106. J. E. Moffitt, D.B.K. Golden, R.E. Reisman, R.L. Lee, R.N. Nicklas, *J. Allergy Clin. Immunol.* **114** (2004) 869–886.
107. S.J. Till, J.N. Francis, K.N. Aria, F.R.C. Path, S.R. Durham, *J. Allergy Clin. Immunol.* **113** (2004) 1025–1034.
108. P. S. Creticos, C. E. Reed, P. S. Norman et al., *N. Engl. J. Med.* **334** (1996) 501 – 506.
109. J. Bousquet, H. Dhivert, F. B. Michel, *J. Allergy* **49** (1994) 31 – 36.
110. P.S. Norman, *J. Allergy Clin. Immunol.* **113** (2004) 1013–1023.

111. S.R. Durham, S.W. Walker, E.-M. Varga, M.R. Jacobson, F. O'Brien, *N. Engl. J. Med.* **341** (1999) 468–475.
112. R. Valenta, J. Lidholm, V. Niederberger, B. Hayek, D. Kraft, H. Grönlund, *Clin. Exp. Allergy* **29** (1999) 896–904.
113. R. Valenta, D. Kraft, *Curr. Opin. Immunol.* **14** (2002) 718–727.
114. L. Kazemi-Shirazi et al., *Int. Arch. Allergy Immunol.* **127** (2002) 259–268.
115. H. Løwenstein, *Clin. Exp. Allergy* **21** (1991) suppl. 1, 227 – 231. (1994)
116. S. T. Holgate, M. K. Church, A. Kapp (eds.): *Allergology*, Ullstein Mosby, Berlin/Wiesbaden 1996, pp. 9 – 12.
117. www.allergenonline.com
118. H. Løwenstein: "Report on behalf of the International Union of Immunological Societies (IUIS) Allergen Standardization Subcommittee", vol. **78**, Arb. Paul-Ehrlich-Inst. Georg Speyer Haus Ferdinand Blum Inst., Frankfurt/M 1983, p. 41 – 48.
119. Nordic Council on Medicines: Registration of allergen preparations, Nordic Guidelines, NLN Publication No. 23, 2nd ed., Uppsala 1989.
120. H.-J. Malling, B. Weeke, *J. Allergy* **48** (1993) suppl. 14, 9 – 35.
121. S. Dreborg, R. Einarsson, *J. Allergy* **47** (1992) 418 – 423.
122. P. S. Norman: "WHO-IUIS International Standards: Advantages of these Extracts," in R. Kurth, D. Haustein (eds.): *Regulatory Control and Standardization of Allergenic Extracts, 7th International Paul-Ehrlich-Seminar 1993*, vol. **87**, Gustav Fischer Verlag, Stuttgart – New York 1994, pp. 59 – 64.
123. Allergen products (Producta allergenica), European Pharmacopoeia 1997, pp. 1063 – 1068.
124. American Academy of Allergy, Asthma and Immunology (AAAAI), *J. Allergy Clin. Immunol.* **99** (1997) 583 – 586.
125. H. Løwenstein: "Methods Used to Develop Standards," in R. Kurth, D. Haustein (eds.): *Regulatory Control and Standardization of Allergenic Extracts, 7th International Paul-Ehrlich-Seminar 1993*, vol **87**, Gustav Fischer Verlag, Stuttgart – New York 1994, pp. 49 – 57.
126. K. Westritschnig, D. Kraft, R. Valenta. *J. Lab. Med.* **26** (2002) 120–129.
127. J. Bousquet: "Clinical Use of Recombinant Allergens and Epitopes," in R. Kurth, D. Haustein (eds.): *Regulatory Control and Standardization of Allergenic Extracts, 7th International Paul-Ehrlich-Seminar 1993*, vol. **87**, Gustav Fischer Verlag, Stuttgart – New York 1994, pp. 257 – 262.
128. M.D. Chapman et al., *J Allergy Clin. Immunol.* **106** (2000) 409–418.
129. S. Dreborg: "Importance of Standardized Allergens in the Diagnosis of Atopy and Atopic Diseases," in R. Kurth, D. Haustein (eds.): *Regulatory Control and Standardization of Allergenic Extracts, 7th International Paul-Ehrlich-Seminar 1993*, vol. **87**, Gustav Fischer Verlag, Stuttgart – New York 1994, pp. 181 – 187.
130. L.K.J. Drachenberg, A.W. Wheeler, P. Stuebner, F. Horak, *Allergy* **56** (2001) 498–505.
131. L.K.J. Drachenberg, A.W. Wheeler, P. Stuebner, F. Horak, *Allergy* **56** (2001) 498–505.
132. W. F. Schoenwetter, *Postgraduate Medicine* **100** (1996) 123 – 135.
133. P. J. Barnes, *N. Engl. J. Med.* **334** (1996) 531 – 532.
134. BSACI Position Paper on Allergen Immunotherapy, *Clin. Exp. Allergy* **23** (1993) suppl. 3, 1 – 44.
135. *Lancet* (1989) no. 1, 259 – 261.
136. F. Bonifazi, M. B. Bilò, *J. Allergy* **52** (1997) 698 – 710.
137. P. S. Norman, T. E. van Metre Jr., *J. Allergy Clin. Immunol.* **85** (1990) 522 – 525.
138. A. W. Wheeler, K.-J. Drachenberg, *J. Allergy* **52** (1997) 602 – 612.
139. T.B. Casale, *J. Allergy Clin. Immunol.* **113** (2004) 1036–1039.
140. AAAI Board of Directors, *J. Allergy Clin. Immunol.* **93** (1994) 811 – 812.
141. G.W. Canonica, G. Passalacqua, *J. Allergy Clin. Immunol.* **111** (2003) 437–448.
142. D. T. Liu: "Regulation of Allergenic products in the U.S.A.; CBER Initiatives," in R. Kurth, D. Haustein (eds.): *Regulatory Control and Standardization of Allergenic Extracts, 7th International Paul-Ehrlich-Seminar 1993*, vol. **87**, Gustav Fischer Verlag, Stuttgart – New York 1994, pp. 7 – 12.
143. J. Kleine-Tebbe et al., Position Paper of the DGAI, *Allergo J.* **9** (2000) 317–324.
144. M. A. Valyasevi, M. W. Yocum, V. A. Gosselin, L. W. Hunt, *J. Allergy Clin. Immunol.* **99** (1997) 66.
145. D.I. Bernstein, W. Wanner, L. Borrish, G.M. Liss, *J. Allergy Clin. Immunol.* **113** (2004) 1129–1136.

146. U. Lüderitz-Püchel, B. Keller-Stanislawski, D. Haustein, Bundesgesundheitsbl.-Gesundheitsforsch.-Gesundheitsschutz, **44** (2001) 709–718.
147. J. Bousquet, I.J. Ansotegui, R. van Ree, P.G. Burney, T. Zuberbier, P. van Cauwenberge, *Allergy* **59** (2004) 1–4.
148. www.asthma.org.uk
149. S.T. Holgate, R. Djukanovic, T. Casale, J. Bousquet, *Clin Exp Allergy* **35** (2005) 408–416.
150. C.G. Ruffin, B.E. Busch, *Am. J. Health-Syst. Pharm.* **61** (2004) 1449–1459.
151. S. Holgate, T. Casale, S. Wenzel, J. Bousquet, Y. Deniz, C. Reisner, *J. Allergy Clin. Immunol.* **115** (2005) 459–465.
152. M.V. Kopp, J. Kühr, *Allergo J.* **13** (2004) 129–134.
153. J. Bousquet, S. Wenzel, S. Holgate, W. Lumry, P. Freeman, H. Fox, *Chest* **125** (2004) 1378–1386.
154. J. Bousquet et al., *Allergy* **60** (2005) 302–308.
155. M. Humbert et al., *Allergy* **60** (2005) 309–316.
156. EMEA, Committee for Medicinal Products for Human Use (CHMP) Summary of Opinion, July 2005.
157. P.J. Barnes, B. Jonsson, J.B. Klim, *Eur. Respir. J.* **9** (1996) 636–642.
158. R.F. Lemanske, A. Nayak, M. McAlary, F. Everhard, A. Fowler Taylor, N. Gupta, *Pediatrics* **110** (2002) e55.
159. D.Y.M. Leung, et al., *N. Engl. J. Med.* **348** (2003) 986–993.
160. Nopko, US 2 479 843 ,1949.

Related Articles → Antiasthmatic Agents is a separate keyword.

Antiarrhythmic Drugs

Gerda von Philipsborn, Knoll AG, Ludwigshafen, Federal Republic of Germany

Anton Oberdorf, Knoll AG, Ludwigshafen, Federal Republic of Germany

Fritz Binnig, Knoll AG, Ludwigshafen, Federal Republic of Germany

Albrecht Franke , BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Antiarrhythmic Drugs 1
1.1. Physiology 1
1.2. Disorders of Cardiac Rhythm (Arrhythmias) 2
1.3. Cardiac Action Potentials 3
1.4. General Aspects of Pharmacology . 4
1.5. General Aspects of Chemistry 7
2. Individual Antiarrhythmic Drugs . 7
2.1. Class I Drugs: Sodium Channel Blockers 7
2.1.1. Class I a Drugs 7
2.1.2. Class I b Drugs 8
2.1.3. Class I c Drugs 9
2.2. Class II Drugs: β-Receptor Blocking Agents 10
2.3. Class III Drugs: Agents Increasing the Action Potential Duration 11
2.4. Class IV Drugs: Calcium Channel Blockers 11
2.5. New Antiarrhythmic Drugs 12
3. References 13

1. Antiarrhythmic Drugs

1.1. Physiology

Unimpaired functioning of the heart is the most important requirement for the normal supply of blood to all regions of the organism. The heart rhythm is determined by the body's changing requirements for blood. This constant adjustment is controlled by venous return from the body's periphery, which influences both the stroke volume of the individual heart beats and the contractile force of the ventricular muscle. Heart rate and contractile force are controlled by the autonomic nervous system.

The individual heart beat begins with the electrical excitation of the *sinoatrial node* (SA node), a group of special pacemaker cells situated at the entrance of the superior vena cava. The SA node generates impulses automatically; their frequency is controlled largely by the autonomic nervous system, although it also can be affected by hormones, e.g., adrenaline, by metabolic processes, and by wall tension. The normal frequency of SA node excitation is about 70 – 80 impulses per minute. Within the heart, excitation spreads via the conduction system from the SA node to the atrial muscles and also to the atrioventricular node, passes through it, and travels along special fibers within the ventricular walls (*His-Purkinje system*), thereby reaching the ventricular muscle (working myocardium).

The *atrioventricular node* (AV node) is a secondary pacemaker; its natural frequency is lower than that of the SA node (its intrinsic rhythm is about 40 – 60 impulses per minute). Accordingly, the higher frequency of the SA node determines the normal heart beat. However, if the SA node fails, the AV node can take over the pacemaker function. A similar relationship exists between the AV node and the Purkinje fibers. They also are capable of spontaneous rhythmic activity if the AV node fails; however, their frequency is at 15 – 40 impulses per minute.

Contraction of the heart muscle therefore is the result of electrical, biochemical, and mechanical processes in the cell membranes and within the myocardial cells (electromechanical coupling). Blood is pumped by the contraction of the heart wall muscles in coordination with the functioning of the various heart valves. All processes associated with the excitation and contraction of the heart are reversible. All the abovementioned phenomena and the automaticity of the pacemakers, particularly the SA node, therefore enable the heart to function rhythmically. For more information on the physiology of excitation and contraction of the heart, see [1–4].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_439

1.2. Disorders of Cardiac Rhythm (Arrhythmias)

Diagnosis. Arrhythmias differ greatly in their etiology and symptoms. The most important diagnostic tool is the *electrocardiogram* (ECG).

The electrocardiogram reflects electrical potential changes in the heart as a function of time. The potential differences originate from the so-called "action potentials" at all excitable cardiac cells (see Section 1.3); they are measured at various sites on the body's surface. For example, the ECG in Figure 1 has been obtained from electrodes placed on the right arm and the left leg. A series of waves, designated P, Q, R, S, and T, is detected. The P – Q interval is the period between the beginning of the contraction of the atria and the beginning of the contraction of the ventricles. The QRS complex is caused by the depolarization of the ventricles. The Q – T interval reflects the duration of ventricular contraction.

A more accurate diagnosis of cardiac arrhythmias is made possible by *His bundle electrography.* In this method, the electrical activity of the heart is detected by intracardiac electrodes, which are inserted intravenously [5], [6]. In addition, intermittent arrhythmias may be detected by *24-hour long-term electrocardiography* (Holter monitoring) [7]. In certain types of arrhythmias, auscultation of the heart sounds as well as palpation of the pulse in arteries and neck veins also provide information on the presence and type of arrhythmia.

Arrhythmias are classified in two main groups according to their etiology: disorders of impulse generation and disturbances of impulse conduction.

Disorders of Impulse Generation. *Nomotopic disorders* involve the normal pacemaker, i.e., the sinoatrial node. They occur either as *sinus tachycardia* (increased rate) or *sinus bradycardia* (reduced rate). Irregular functioning of the sinus node also may cause sinus arrhythmias. Impulse generation disorders are called *heterotopic* when the impulse originates not in the sinoatrial node but in one of the centers subordinate to it: in the AV node, at certain foci in the ventricular conduction system with escape rhythms, or even at other sites of the heart muscle. They trigger individual, premature heart beats, which are called *extrasystoles.* These may occur either sporadically or frequently, irregularly or firmly coupled to the normal rhythm of the heart. *Polytopic extrasystoles* originate at several different foci simultaneously. Depending on the anatomic position of the impulse generation, extrasystoles are described as supraventricular or ventricular. *Supraventricular extrasystoles* are triggered at points above the His-Purkinje system. They include sinoidal, atrial, and atrioventricular extrasystoles. *Ventricular extrasystoles* originate in the His-Purkinje system or in the ventricular myocardium.

Paroxysmal Tachycardia. Heterotopic disorders of impulse generation also include sudden and more or less prolonged phases of a highly accelerated heart rate, which are known as paroxysmal tachycardias. Rapid regular impulse generation (250 – 300/min) in the atrium is known as *atrial flutter.* In this case, impulse conduction to the ventricles is disturbed. The same applies to *atrial fibrillation,* which is characterized by high-frequency ($>$ 400/min) and irregular impulse generation. Because the ventricles may be refractory to most of these impulses, their activity becomes irregular; this results in *absolute arrhythmia.*

Similar impulse-generation disorders may also originate in the ventricles, namely *ventricular flutter* (with a frequency in most instances above 220/min) and *ventricular fibrillation* in which irregular multifocal impulses occur simultaneously at high frequency. In ventricular flutter, the heart is still able to pump a small volume of blood; in ventricular fibrillation, however, circulatory arrest occurs.

Impulse-Conduction Disturbances. Impulse transmission to lower sections of the heart can be delayed or blocked at different points of the conduction system. Atrial or ventricular systoles can be blocked either partially or completely. In *total sinoatrial block,* sinus excitation does not reach the atria and the ventricles at all. Consequently, systoles cease until another pacemaker situated at a lower level becomes active as the automaticity center.

Intra-atrial disturbances affect impulse propagation within the atrial musculature. They

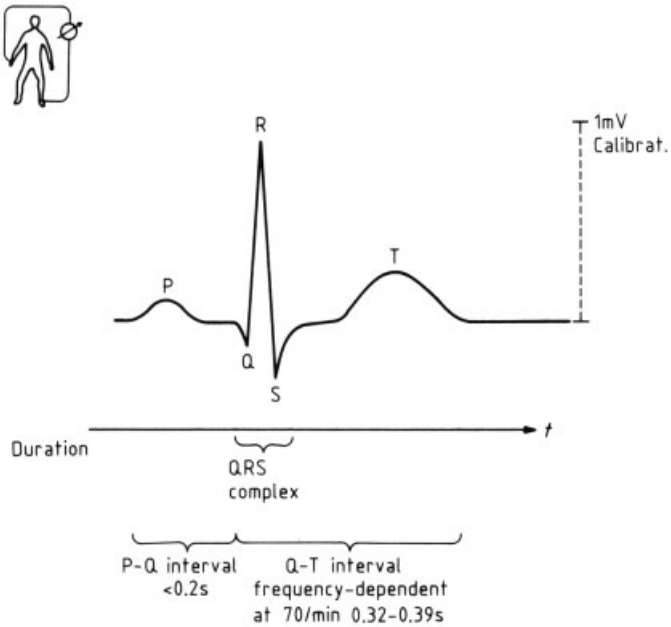

Figure 1. Normal ECG with bipolar recording from the body surface in the direction of the long axis of the heart
The times below the ECG curve are important limiting values for the durations of the parts of the curve [1].

are less significant provided regular conduction to the ventricles remains intact.

However, *atrioventricular block* (AV block), in which conduction between atria and ventricles is disturbed, is extremely important. Atrioventricular block is characterized according to either location or degree. In *first-degree AV block,* conduction is still regular but in some cases delayed by several times the normal duration. In *second-degree AV block* (incomplete AV block), conduction is interrupted sporadically or frequently. In *third-degree AV block* (total heart block) there is no conduction of excitation between atria and ventricles. Both beat independently of one another, the atria generally at the sinus rate and the ventricles about once every two seconds.

In *intraventricular disturbances,* impulse conduction within the ventricular walls is delayed or interrupted.

For more detailed information on causes and basic diseases, syndromes, diagnosis, therapy, and prognosis of cardiac arrhythmias, see [5–7], [9].

1.3. Cardiac Action Potentials

In the heart, as in other excitable tissue, rest and excitation are accompanied by electrical phenomena, the so-called action potentials, which are caused by ionic currents. The flow of ions into and out of the cardiac cells is controlled by changes in cell membrane permeability to sodium, potassium, and calcium ions. For details, see [1–4], [6–9].

Figure 2 shows transmembrane action potentials of normal, nonautomatic cardiac cells, such as ventricular or atrial muscle cells (V), and of sinoatrial nodal cells (SA). Usually, five different phases are distinguished [10].

In *phase 0* rapid depolarization occurs on excitation. The difference in membrane potential decreases and the inside of the cell becomes positive in relation to the outside during the so-called overshoot. The ionic mechanism for phase 0 is an increase in sodium conductance.

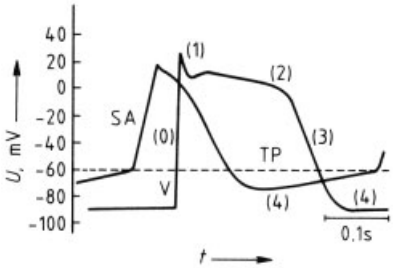

Figure 2. Action potentials of a ventricular muscle cell (V) and a sinoatrial nodal cell (SA).
TP = Threshold potential

The first phase of repolarization (*phase 1*) is quite rapid and is followed by a period of slow repolarization (*phase 2* or plateau). Repolarization speeds up again in *phase 3* and completes the process by returning the transmembrane potential to the resting potential. The ionic mechanisms for repolarization are an increase in calcium conductance (slow inward flow) and a decrease in potassium conductance. In ventricular muscle cells, *phase 4* is characterized by an electrical quiescence in which depolarization currents and repolarization currents are in balance.

The action potentials of automatic cells, such as those of the sinoatrial node, differ conspicuously from those of ventricular muscle cells: First, the maximum diastolic potential reached in these cells is less negative (−70 mV). Second, the automatic SA nodal cells exhibit spontaneous diastolic depolarization. This is shown by the gradual increase of transmembrane action potential in phase 4. When the potential difference reaches the threshold potential (TP; about −50 mV in SA nodal cells), rapid depolarization (phase 0) occurs. This mechanism, by which the heart generates its own impulse, is called autorhythmicity. Finally, upstroke velocity (phase 0) and amplitude are distinctly smaller.

Further parameters of the transmembrane action potential that are important in a discussion of the mechanism of action of antiarrhythmic drugs are:

1) *Action potential amplitude:* the total potential change occurring during phase 0
2) *Action potential duration* (APD): the period from the start of phase 0 to the end of phase 3
3) *Conduction velocity:* the speed at which a stimulus or action potential is propagated
4) *Automaticity:* determined by the slope of phase 4 (discussed previously)
5) *Maximum upstroke velocity* (V/s): the steepest slope of rise during phase 0
6) *Excitability:* dependent on the level of the threshold potential
7) *Effective refractory period* (ERP): period of repolarization during which no normal action potential can be elicited, although a weak, nonpropagated action potential can arise
8) *Relative refractory period* (RRP): period of repolarization in which no propagated action potential can arise

These parameters can be altered experimentally by antiarrhythmic drugs. The therapeutic antiarrhythmic activity of such drugs in humans is associated with these changes also.

The most important antiarrhythmics are those that inhibit pathologically accelerated impulse generation and/or propagation, that is, any so-called *tachyarrhythmia.* Such compounds are also termed *antifibrillatory drugs.*

Bradyarrhythmias are disturbances of cardiac rhythm caused by delayed or blocked impulse generation and/or propagation; they are rare. The bioelectrical phenomena of heart function insure that the heart always responds to the more rapid pacemaker. Whenever impulse generation in the SA node is delayed, normal cardiac function can be restored by implanting an *electrical pacemaker.* In emergencies, but not in long-term therapy, sympathomimetic noradrenaline and parasympatholytic atropine can be administered parenterally. Oral β-sympathomimetics, such as isoprenaline and orciprenaline, may be used until a pacemaker is implanted.

1.4. General Aspects of Pharmacology

Antiarrhythmic agents act by changing the electrical properties of the membrane of the heart cell [11].

Figure 3 shows the characteristic changes caused by various class I antiarrhythmic agents in the ventricular action potential (Fig. 3 A), the electrocardiogram (Fig. 3 B), the duration of the effective refractory period (ERP), the relative refractory period (RRP), and the total action potential duration (APD; Fig. 3 C).

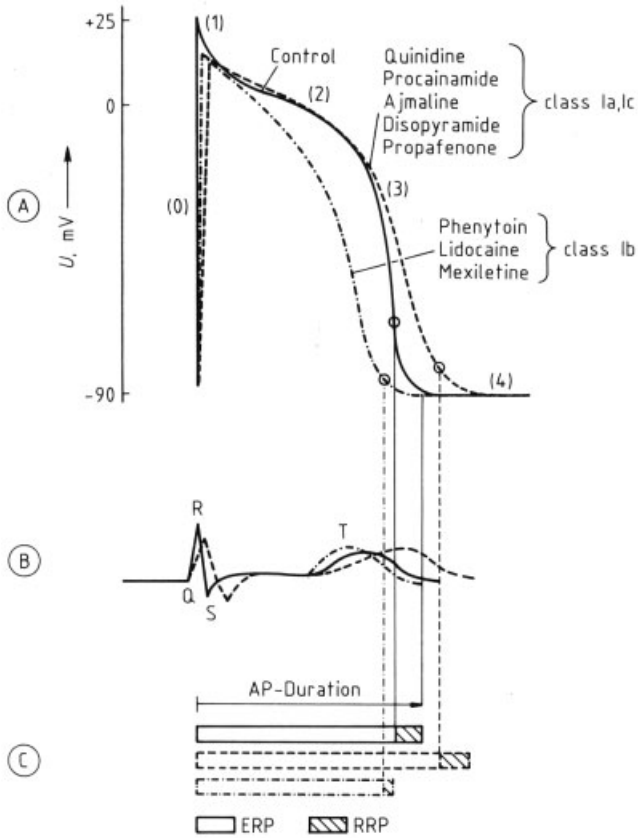

Figure 3. Effects of various antiarrhythmic drugs on the ventricular action potential, the electrocardiogram, and the durations of the effective refractory period (ERP), the relative refractory period (RRP), and the total action potential duration (APD) [11].
The arabic numerals in parentheses indicate phases of the action potential1.3. The circles represent the level of repolarization at which the fiber becomes reexcitable.

The classification of antiarrhythmic agents by Vaughan Williams [12–14] (Table 1), is comprised of four classes and three subgroups of class I; it is based primarily on electrophysiological effects (class II is the exception).

Several of the drugs act in more than one of the four ways; therefore, the classification does not so much categorize the agents as it describes four ways in which abnormal cardiac rhythms can be corrected or prevented.

Class I: Sodium Channel Blockers. All class I antiarrhythmic agents (I a, I b, and I c) decrease the rate of rise (upstroke velocity) of phase 0 of the action potential by inhibiting the rapid sodium influx. Under suitable conditions these agents also can block the sodium channels of nerve fibers, which explains the local-anesthetic effect of class I agents.

Under block, the sodium channels may recover quickly or slowly, depending on the class of antiarrhythmic drug [15]. With *class I b agents,* such as lidocaine, phenytoin, tocainide, mexiletine, or sparteine, recovery time is less than 1 s. Therefore, this class of drugs can block premature and high-frequency excitation selectively. With *class I a* or *class I c agents,* however, the recovery time of the sodium channels is longer (1 – 250 s); therefore, the blocking effect also affects normal-frequency action potentials. This difference also is detected by ECG. In the healthy heart, class I b antiarrhythmic agents in therapeutic doses do not change the conduction times, whereas class I a and class I c agents

Table 1. Classification of antiarrhythmic drugs [14] , pp. 232, 234 [15]

	Generic name	Mode of action	ECG changes ∗			
			Heart rate	P–Q ∗∗ interval	QRS ∗∗ duration	Q–T ∗∗ interval
Class I a:	quinidine procainamide disopyramide	sodium channel blockade by classes I a, I b, I c; rate of rise of phase 0	inc.	(inc.) (dec.)	inc.	inc.
Class I b:	lidocaine phenytoin tocainide sparteine mexiletine	of action potential: delayed by I a, I b, I c; action potential duration: shortened by I b, not changed or prolonged	dec.	-	-	-
Class I c:	ajmaline prajmalium aprindine lorcainide flecainide propafenone	by I a and I c	dec.	inc.	inc.	-
Class II:	propranolol and all other β-blockers	β-receptor blockade	dec.	inc.	-	-
Class III:	amiodarone sotalol bretylium	prolongation of action potential duration	dec.	-	-	inc.
Class IV:	verapamil gallopamil diltiazem bepridil	calcium channel blockade; action potential (SA node and AV node) duration and plateau shortened	dec.	inc.	-	-

∗ dec., decrease; inc., increase; -, no or little change; ( ), change not pronounced.
∗∗ for definition, see 1.2.

lengthen the ECG times (Fig. 3). Some class I a agents have an additional atropine-like effect. Occasionally this increases frequency and shortens the P – Q interval.

The duration of the action potential and the Q – T interval in the ECG are lengthened by class I a agents (quinidine, procainamide, disopyramide) and by some class I c agents (ajmaline, propafenone) but are shortened by class I b agents (phenytoin, lidocaine, mexiletine). Phenytoin acts during repolarization: the membrane potential at which the fibers become reexcitable is shifted to a more negative value and the effective refractory period is lengthened with respect to total action potential duration.

Class II: β-Receptor Blocking Agents. Two types of receptors exist in the sympathetic nervous system: α- and β-receptors [16]. Stimulation of β-receptors produces increased sympathetic activity of the heart, increasing contractility, oxygen consumption, and heart rate and accelerating impulse propagation.

These effects are specifically, competitively, and reversibly blocked by β-receptor blockers. In addition to this specific effect, high doses of β-blockers also block membrane sodium channels (membrane stabilizing effect). In antiarrhythmic therapy, β-blockers are used primarily for tachyarrhythmias resulting from sympathetic hyperfunction or from increased catecholamine release [17] (→ β-Receptor Blocking Agents).

Class III: Agents That Increase Action Potential Duration. Class III antiarrhythmic agents lengthen the duration of the action potential without changing its rate of rise. The refractory period of the heart fibers is therefore prolonged. Currently, only three agents of this class are known: amiodarone, sotalol, and bretylium.

Class IV: Calcium Channel Blockers. Calcium channel blockers also are known as calcium antagonists. In various excitable cells these agents block the slow influx of calcium ions through the cell membrane during excitation; the flow of sodium ions is not influenced. Calcium antagonists are applied chiefly in coronary heart disease. Some of these substances also are used in antiarrhythmic therapy: verapamil, gal-

lopamil, diltiazem, and bepridil. Their antiarrhythmic effect is based on prolongation of the impulse conduction in the AV node and to a lower degree on inhibition of the impulse generation in the sinus node. They act especially on tachyarrhythmias that originate in the atria (see also → Calcium Antagonists).

1.5. General Aspects of Chemistry

Class I antiarrhythmic agents belong to two groups: N-substituted carboxamides and ajmaline derivatives. The group of N-substituted carboxamides includes lidocaine, tocainide, procainamide, disopyramide, lorcainide, flecainide, bunaftine, and encainide; mexiletine usually is included because of its similar structure. Antiarrhythmic drugs of the second group include the ajmaline-type alkaloids and such derivatives as prajmalium, detajmium, and lorajmine, as well as quinidine and sparteine. Some other substances, e.g., phenytoin and aprindine, do not fit into either group.

All N-substituted carboxamides have three features in common: a lipophilic aromatic group, an aliphatic spacer group, and an amino substituent. *Lipophilicity* is crucial for nonspecific interaction with the alkyl chains of the membrane's phospholipids. Together with the amino group, which can be protonated at physiological pH values, this seems to be the molecular requirement for antiarrhythmic activity in this group of compounds. The molecular interaction between these drugs and the membrane and quantitative structure – activity relations of antiarrhythmic agents are described in more detail in [18].

The class II antiarrhythmic agents (β-receptor blockers) virtually all belong to the group of aryloxyaminopropanols (→ β-Receptor Blocking Agents). As mentioned above, these compounds can even initiate a membrane-stabilizing effect if administered in relatively high concentrations. Predominantly the L-isomers of the β-receptor blockers show β-adrenolytic activity; the membrane-stabilizing effect is observed in both isomers and depends on the lipophilicity of the compounds [19], [20].

Structure – activity relationships are not evident in class III and class IV antiarrhythmic agents.

2. Individual Antiarrhythmic Drugs

[6], [7], [9], [15], [21–24]

2.1. Class I Drugs: Sodium Channel Blockers

2.1.1. Class I a Drugs

Quinidine [*56-54-2*], (+)-(αS)-α-(6-methoxy-4-quinolyl)-α-[(2*R*,4*S*,5*R*)-(vinylquinuclidin-2-yl)]methanol, $C_{20}H_{24}N_2O_2$, M_r 324.4, *mp* 174 – 175 °C.

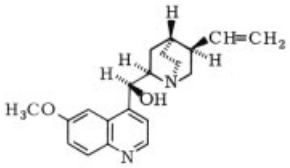

Quinidine is one of the most widely applied antiarrhythmics and has been used for 30 years. It is used primarily against atrial flutter and fibrillation and against heterotopic extrasystoles.

Quinidine generally is administered orally, although slow intravenous administration also is possible [25]. Quinidine is the prototype of class I antiarrhythmics; therefore, substances in this group also are called "quinidine-like." However, it differs, for example, from lidocaine by the following effects: it increases the duration of the action potential, in the ECG it increases QRS duration and Q – T interval, and the refractory period is markedly prolonged. Quinidine can accelerate the impulse conduction rate in the AV node (atropine-like effect). Side effects include gastrointestinal disturbances and allergic reactions.

Trade Names: Chinidin, Kinidin (Astra Chemicals) contain quinidine hydrogen sulfate [*747-45-5*]; Duraquin (Parke-Davis), Quinaglute (Berlex) contain quinidine gluconate [*7054-25-3*].

Procainamide [*51-06-9*], 4-amino-*N*-(2-diethylaminoethyl)benzamide, $C_{13}H_{21}N_3O$, M_r 235.3; the hydrochloride [*614-39-1*] melts at 165 – 169 °C.

In the United States procainamide is one of the most frequently used antiarrhythmics; in Europe, it is of minor importance. Its mode of action resembles that of quinidine. The substance may be administered parenterally and orally but is only short acting. Side effects with longterm therapy are gastrointestinal disturbances, agranulocytosis, and lupus erythematosus. An antiarrhythmic metabolite, acecainide page 13, was detected in humans.

Trade Names: Novocamid (Hoechst), Pronestyl (Squibb) both contain procainamide hydrochloride.

Disopyramide [*3737-09-5*], (±)−4-diisopro-pylamino-2- phenyl-2- (2-pyridyl) butyramide, $C_{21}H_{29}N_3O$, M_r 339.5, *mp* 94.5 °C.

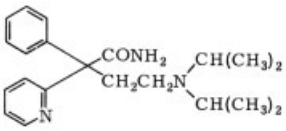

Disopyramide has a pharmacological profile similar to that of quinidine and procainamide [26–29]. Clinically, it is active against most forms of arrhythmias (supraventricular and ventricular). However, the substance also can produce serious side effects, such as negative inotropic and hemodynamic effects, increased oxygen consumption, and increased infarct size, as well as anticholinergic activity.

Trade Names: Norpace (Searle), Rythmodan (Roussel), Rythmodul (Albert Roussel), all contain disopyramide phosphate [*22059-60-5*].

2.1.2. Class I b Drugs

Lidocaine [*137-58-6*], lignocaine, *N*-(2,6-xylyl)-diethylaminoacetamide, $C_{14}H_{22}N_2O$, M_r 234.4, *mp* 68 – 69 °C.

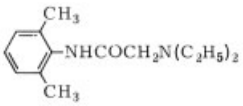

Lidocaine is used both as a local anesthetic and as an antiarrhythmic. As an antiarrhythmic it is administered only parenterally as an infusion because it has a short half-life. The most important side effects are disturbances of the central nervous system.

Trade Names: Xylocain, Xylocaine, Xylocard (Astra Chemicals); all preparations contain lidocaine hydrochloride [*73-78-9*].

Phenytoin [*57-41-0*], 5,5-diphenylhydantoin, $C_{15}H_{12}N_2O_2$, M_r 252.3, *mp* 295 – 298 °C.

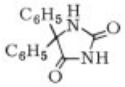

Phenytoin is used primarily as an antiepileptic. The compound finds its chief use as an antiarrhythmic in cases of cardiac glycoside overdosage. The effect of the substance is difficult to control because of its long half-life. In addition, it has strong cardiac and extracardiac side effects (gingival hyperplasia, osteomalacia, anemia, hypertrichosis).

Trade Names: Dilantin, Epanutin (Parke Davis), Phenydan (Desitin), Zentropil (Nordmark-Werke); all preparations contain phenytoin sodium [*630-93-3*].

Tocainide [*41708-72-9*], (±)−2-amino-*N*-(2,6-xylyl)propionamide, $C_{11}H_{16}N_2O$, M_r 192.1; the hydrochloride [*35891-93-1*] melts at 246 – 247.5 °C.

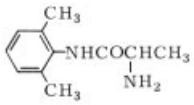

Tocainide corresponds to lidocaine in its action profile and side effects. In contrast to lidocaine it can be taken orally because it lacks two ethyl groups that contribute to lidocaine's first-pass hepatic degradation after oral administration [27], [30], [31]. The substance has been on the market in most European countries since 1981 – 1982.

Trade Names: Tonocard, Xyloctan (Astra Chemicals); all preparations contain tocainide hydrochloride.

Mexiletine [*31828-71-4*], (±)-1-methyl-2-(2,6-xylyloxy)ethylamine, $C_{11}H_{17}NO$, M_r 179.3; the hydrochloride [*5370-01-4*] melts at 203 – 205 °C.

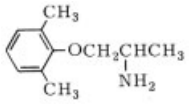

Mexiletine is effective both parenterally and orally [32], [33]. Because the substance has a large distribution volume, a high initial dose is necessary. Side effects include bradycardia and disturbances of the central nervous system.

Trade Name: Mexitil (Boehringer Ingelheim) contains mexiletine hydrochloride.

Sparteine [*90-39-1*], dodecahydro-7,14-methano-2*H*, 6*H*-dipyrido-[1,2-*a* : 1′,2′-*e*][1,5]diazocine, $C_{15}H_{26}N_2$, M_r 234.4, *bp* 173 °C (at 11 kPa).

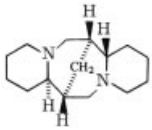

Sparteine is effective against sinoatrial and atrial tachycardias, atrial extrasystoles, and also against ventricular arrhythmias [34]. Despite prolongation of the myocardial refractory period in vitro, the ECG times are not prolonged in vivo.

Pregnancy in the last trimester is a contraindication because the substance can induce labor through increased smooth muscle tone.

Trade Name: Depasan (Giulini) contains sparteine sulfate [*299-39-8*].

2.1.3. Class I c Drugs

Ajmaline [*4360-12-7*], (17*R*,21*R*)-ajmalan-17,21-diol, $C_{20}H_{26}N_2O_2$, M_r 326.4, *mp* 195 °C (decomp.).

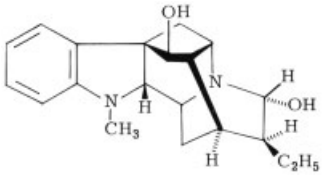

Ajmaline is active against both ventricular and supraventricular arrhythmias [35–37]. Unlike quinidine, procainamide, and disopyramide, ajmaline has no anticholinergic activity and shows only a weak negative inotropic effect. The preferred route of application is slow intravenous infusion because oral bioavailability varies greatly among individuals. Side effects include disturbances of the central nervous system and intrahepatic cholestasis.

Trade Names: Cardiorythmine (Servier), Gilurytmal (Giulini).

Prajmalium bitartrate [*2589-47-1*], 4-propylajmalium hydrogen tartrate, $C_{27}H_{38}N_2O_8$, M_r 518.6, *mp* 133 °C (decomp.).

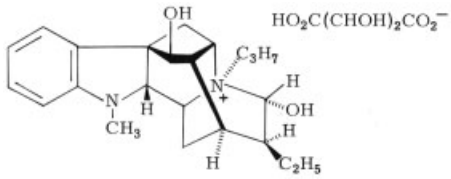

With at least five times the potency of ajmaline, prajmalium is a most effective antiarrhythmic, and it is superior to ajmaline in its reliable oral effect and sustained action [37–39]. The activity profile of prajmalium is very similar to that of ajmaline, although the sodium channels are blocked much longer by it than by any other class I antiarrhythmic [15].

Trade Name: Neo-Gilurytmal (Giulini).

Detajmium bitartrate [*53862-81-0*], 4-(3-dimethylamino-2-hydroxypropyl)ajmaline hydrogen tartrate, $C_{31}H_{47}N_3O_9$, M_r 605.7, *mp* 123 – 125 °C.

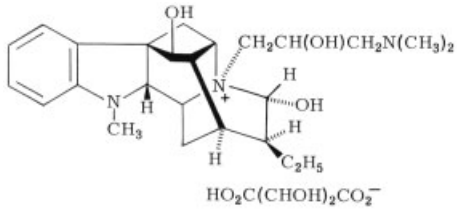

Detajmium is comparable to ajmaline but shows better oral activity. In potency it ranges between ajmaline and prajmalium [40].

Trade Name: Tachmalor (VEB Arzneimittelwerk, Dresden).

Aprindine [*37640-71-4*], *N*-(3-diethylaminopropyl)-*N*-indan-2-ylaniline, $C_{22}H_{30}N_2$, M_r 322.5; the hydrochloride [*33237-74-0*] melts at 120 °C.

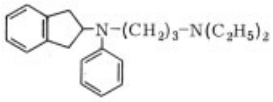

Aprindine has a sustained activity. Because of serious side effects (agranulocytosis), aprindine should only be used for life-threatening arrhythmias otherwise refractory to therapy [27].

Trade Names: Amidonal (Madaus), Fibocil (Lilly), Fiboran (Christiaens); all preparations contain aprindine hydrochloride.

Lorcainide [*59729-31-6*], 4′-chloro-*N*-(1-isopropyl-4-piperidyl)–2-phenylacetanilide, $C_{22}H_{27}ClN_2O$, M_r 370.9; the hydrochloride [*58934-46-6*] melts at 263 °C.

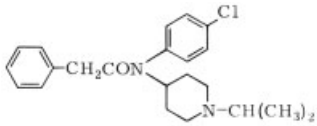

Like lidocaine, lorcainide acts primarily on ventricular arrhythmias but is stronger and also orally effective [41], [42]. Side effects include insomnia and gastrointestinal disturbances.

Trade Name: Remivox (Janssen) contains lorcainide hydrochloride.

Flecainide [*54143-55-4*], (±)-*N*-(2-piperidylmethyl)–2,5-bis-(2,2,2-trifluoroethoxy)benzamide, $C_{17}H_{20}F_6N_2O_3$, M_r 414.4; the hydrochloride [*54143-55-4*] melts at 228 – 229 °C.

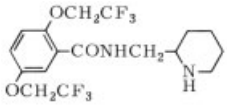

Flecainide is effective if administered intravenously and orally; its effect is long-lasting. The drug combines the modes of action of class I a and I b drugs with those of class III. Flecainide shows few side effects; it does not affect the central nervous system [29], [43]. The substance has been used increasingly since it was first put on the market in 1982.

Trade Name: Tambocor (Kettelhack Riker) contains flecainide acetate [*54143-56-5*].

Propafenone [*54063-53-5*], (±)–2′-(2-hydroxy-3-propylaminopropoxy)–3-phenylpropiophenone, $C_{21}H_{27}NO_3$, M_r 341.5; the hydrochloride [*34183-23-8*] melts at 173 – 174 °C.

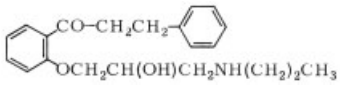

Propafenone has not only pronounced class I effects but also class II (structure related) and class IV activities [44–46]. Accordingly, it has a broad spectrum of activity against ventricular and supraventricular arrhythmias. Propafenone can be administered intravenously and orally; it is well suited for long-term anitiarrhythmic treatment [9] and has been increasingly used since it was first marketed in Germany in 1978.

Trade Name: Rytmonorm (Knoll) contains propafenone hydrochloride.

2.2. Class II Drugs: β-Receptor Blocking Agents

For more information, → β-Receptor Blocking Agents.

Propranolol [*525-66-6*], (±)–1-isopropylamino-3-(1-naphthyloxy)-propan-2-ol, $C_{16}H_{21}NO_2$, M_r 259.3; the hydrochloride [*318-98-9*] melts at 163 – 164 °C,

Major indications of propranolol, and of many other β-blockers, include coronary heart disease, hypertension, and cardiac arrhythmias. These β-blocking agents are mainly used in arrhythmias accompanied by increased catecholamine levels, e.g., excitement, stress, or hyperthyroidism [17].

Trade Names: Dociton, Inderal (ICI) contain propranolol hydrochloride.

2.3. Class III Drugs: Agents Increasing the Action Potential Duration

Amiodarone [*1951-25-3*], (2-butylbenzofuran-3-yl)-[4-(2-diethylaminoethoxy)−3,5-diiodophenyl]ketone, $C_{25}H_{29}I_2NO_3$, M_r 645.3; the hydrochloride [*19774-82-4*] melts at 161 °C.

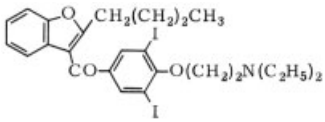

Amiodarone was introduced as a coronary therapeutic agent more than 20 years ago. Later, its antiarrhythmic activity was discovered [47–49]. Amiodarone increases the duration of the refractory period and is active against supraventricular as well as ventricular arrhythmias. Its therapeutic spectrum is relatively broad. The full activity of amiodarone only develops after 10 – 15 d of treatment. The drug is active up to 45 d after it is discontinued. Side effects include: microcrystalline deposits in the cornea in 90 % of the cases, impaired thyroid function, and photosensitivity.

Trade Names: Cordarex, Cordarone (Labaz), both contain amiodarone hydrochloride.

Sotalol [*3930-20-9*], (±)−4′-(1-hydroxy-2-isopropylaminoethyl)methanesulfoanilide, $C_{12}H_{20}N_2O_3S$, M_r 272.4; the hydrochloride [*959-24-0*] melts at 207 °C.

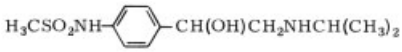

Sotalol acts as an antiarrhythmic by β-receptor blockade as well as by prolonging the action potential duration in atrium and ventricle and the refractory period in atrium and AV node [48]. Activity against supraventricular and ventricular arrhythmias was demonstrated. The substance has the typical side effects of β-blockers.

Trade Names: Beta-Cardone (Duncan, Flockhart), Sotalex (Lappe), Sotazide (Bristol-Meyers Pharmaceuticals), all preparations contain sotalol hydrochloride.

Bretylium tosylate [*61-75-6*], 2-bromobenzyl (ethyl) dimethylammonium toluene-4-sulfonate, $C_{18}H_{24}BrNO_3S$, M_r 414.4, *mp* 97 – 99 °C.

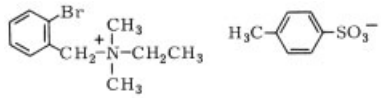

Bretylium tosylate is reported to be effective in the acute control of recurrent ventricular tachycardia and even fibrillation [48], [50]. The drug is contraindicated in arrhythmias caused by digitalis intoxication.

Trade Names: Bretylol (American Critical Care), Bretylate (Wellcome).

2.4. Class IV Drugs: Calcium Channel Blockers

The main application of calcium antagonists is the treatment of coronary heart disease. The substances mentioned below also are active against arrhythmias.

Verapamil [*52-53-9*], (±)−5-[*N*-(3,4-dimethoxyphenethyl)-*N*-methylamino]−2-(3,4-dimethoxyphenyl)−2-isopropylvaleronitrile, $C_{27}H_{38}N_2O_4$, M_r 454.6; the hydrochloride [*152-11-4*] melts at 139.5 – 140.5 °C.

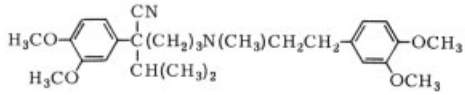

Verapamil, the classic calcium antagonist, has a negative inotropic, anti-ischemic, and conduction-delaying effect on the heart [27], [51–54]. Its antiarrhythmic effect is based primarily on a prolongation of impulse conduction time in the AV node and a reduction of frequency of impulse generation in the SA node. Verapamil has a very strong effect against paroxysmal supraventricular arrhythmias, although activity against ischemic ventricular arrhythmias has been demonstrated also [51]. The most important side effect is decreased blood pressure.

Trade Names: Isoptin (Knoll), Securon (Knoll Ltd.), Calan (Searle), Cardibeltin (Pharma Schwarz), Cordilox (Abbott); all preparations contain verapamil hydrochloride.

Gallopamil [*16662-47-8*], (±)−5-[*N*-(3,4-dimethoxyphenethyl)-*N*-methylamino]−2-(3,4,5-trimethoxyphenyl)−2-isopropylvaleronitrile, $C_{28}H_{40}N_2O_5$, M_r 484.7; the hydrochloride [*16662-46-7*] melts at 145 – 148 °C.

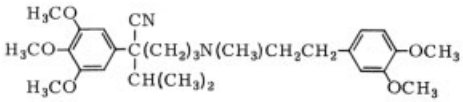

Gallopamil has a distinctly stronger effect than verapamil but a similar activity profile [54], [55].

Trade Name: Procorum (Chemische Werke Minden); contains gallopamil hydrochloride.

Bepridil [*74764-40-2*], *N*-(3-isobutoxy-2-pyrrolidin-1-ylpropyl)-*N*-phenylbenzylamine, $C_{24}H_{34}N_2O$, M_r 366.3; the hydrochloride monohydrate melts at 88 – 90 °C.

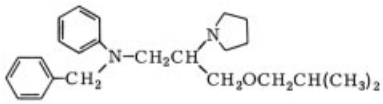

In addition to class IV effects, bepridil also shows class I effects. The substance is active against both supraventricular and ventricular arrhythmias [56], [57].

Trade Name: Cordium (Lab. Mauverney); contains bepridil hydrochloride monohydrate.

Diltiazem [*42399-41-7*], *cis*-(+)−3-acetoxy-5-(2-dimethylamino-ethyl)−2,3-dihydro-2-(4-methoxyphenyl)−1,5-benzothiazepin-4(5*H*)-one, $C_{22}H_{26}N_2O_4S$, M_r 414.5, *mp* 212 °C (decomp.).

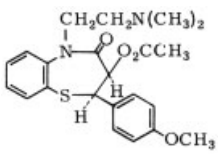

Like verapamil and gallopamil, diltiazem has both vascular (antihypertensive) and cardiac (antianginal) activities and shows antiarrhythmic activity similar to these drugs [58], [59].

Trade Names: Dilzem (Gödecke), Herbesser (Tanabe), Cardizem (Marion); all preparations contain diltiazem hydrochloride [*33286-22-5*].

2.5. New Antiarrhythmic Drugs

Only limited clinical experience exists for these drugs.

Encainide [*37612-13-8*], (±)−2′-[2-(1-methyl-2-piperidyl)ethyl]-*p*-anisanilide, $C_{22}H_{28}N_2O_2$, M_r 352.2, *mp* 131.5 – 132.5 °C.

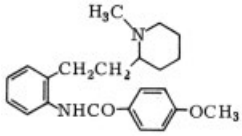

Encainide is a membrane-active class I c agent with complex metabolism and clinical pharmacology [60], [61]. To date, its effectiveness has been proved against ventricular arrhythmias. Side effects include central nervous system activity.

Trade Name: Enkade (Bristol-Meyers).

Bunaftine [*32421-46-8*], *N*-butyl-*N*-(2-diethylaminoethyl)-1-naphthamide, $C_{21}H_{30}N_2O$, M_r 326.7.

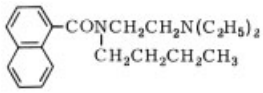

Bunaftine is said to have a "quinidine-like" effect.

Trade Name: Meregon (Malesci) contains bunaftine citrate.

Moracizine [*31883-05-3*], ethyl[10-(3-morpholinopropionyl)phenothiazin-2-yl]carbamate, $C_{22}H_{25}N_3O_4S$, M_r 427.5, *mp* 156 – 157 °C.

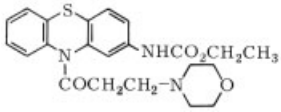

Moracizine is said to have an activity similar to but stronger than quinidine. Gastrointestinal and neurological side effects are observed rarely [27], [62].

Trade Name: Ethmozine (Riga Pharmaceutical "Alaine Pharm," USSR).

Lorajmine [*47562-08-3*], ajmaline-17-chloroacetate, $C_{22}H_{27}ClN_2O_3$, M_r 402.9.

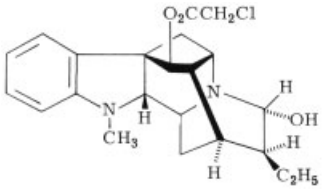

In electrophysiological studies, lorajmine has been shown to exert a lidocaine-like (class I b) rather than an ajmaline-like (class I c) effect [63].

Trade Names: Ritmos Elle (Inverni della Beffa), Viaductor (Lab. Servier); both contain lorajmine hydrochloride [*40819-93-0*].

Acecainide [*34118-92-8*], 4-acetylamino-*N*-[2-(diethylamino)ethyl]benzamide, $C_{15}H_{23}N_3O_2$, M_r 315.65; the hydrochloride melts at 190 – 193 °C.

Acecainide, a main metabolite of procainamide, is a class I a antiarrhythmic agent [64].

Trade Name: Napa (Medco).

3. References

1. H. Antoni in R. F. Schmidt, G. Thew (eds.): *Human Physiology,* Springer Verlag, Berlin-Heidelberg-New York 1983, pp. 358 – 396.
2. A. C. Guyton: *Basic Human Physiology,* Saunders, Philadelphia-London-Toronto 1977.
3. E. B. Mason: *Human Physiology,* Benjamin-Cummings Publishing Co., Menlo Park 1983.
4. R. M. Berne (ed.): "The Heart",*Handbook of Physiology* , Sect. 2, vol. 1, Am. Physiol. Soc., Bethesda 1979.
5. R. Thorspecken, P. Hassenstein: *Rhythmusstörungen des Herzens,* Thieme Verlag, Stuttgart 1975.
6. H. Roskamm, H. Reindell (eds.): *Herzkrankheiten,* Springer Verlag, Berlin-Heidelberg-New York 1982.
7. B. Lüderitz (ed.): *Ventrikuläre Herzrhythmusstörungen,* Springer Verlag, Berlin-Heidelberg-New York 1981.
8. D. T. Mason, A. N. Demaria, E. A. Amsterdam, R. Zelis, R. A. Massumi, *Drugs* **5** (1973) 292 – 317.
9. B. Lüderitz: " Herzrhythmusstörungen", *Handbuch der inneren Medizin,* Springer Verlag, Berlin-Heidelberg-New York 1983.
10. P. H. Morgan, I. W. Mathison, *J. Pharm. Sci.* **65** (1976) no. 4, 467.
11. L. S. Gettes, *Am. J. Cardiol.* **28** (1971) 528.
12. E. M. Vaughan Williams in L. Szekeres (ed.): *Pharmacology of Antiarrhythmic Agents,* Pergamon Press, Oxford-New York 1981, pp. 125 – 150.
13. E. M. Vaughan Williams, *J. Clin. Pharmacol.* **24** (1984) 129 – 147.
14. E. M. Vaughan Williams, *Eur. Heart J.* **5** (1984) 96 – 98.
15. P. Honerjäger in C.-J. Estler (ed.): *Lehrbuch der allgemeinen und systematischen Pharmakologie und Toxikologie,* F. K. Schattauer Verlag, Stuttgart-New York 1983, p. 236.
16. R. P. Ahlquist, *Am. J. Physiol.* **153** (1984) 586.
17. D. G. McDevitt, *Drugs* **17** (1979) 267 – 288.
18. R. E. Thomas in M. Wolff (ed.): *Burger's Medical Chemistry,* 4th edition, part III, Wiley & Sons, New York 1981, pp. 81 – 102.
19. D. Hellenbrecht et al., *Naunyn Schmiedebergs Arch. Pharmacol.* **277** (1973) 211.
20. D. Hellenbrecht et al., *Eur. J. Pharmacol.* **29** (1974) 223.
21. O. Kraupp in W. Forth, D. Henschler, W. Rummel (eds.): *Allgemeine und spezielle Pharmakologie und Toxikologie,* 4th ed., Bibliographisches Institut, Mannheim-Wien-Zürich 1983, pp. 192 – 212.
22. B. N. Singh in L. Szekeres (ed.): *Pharmacology of Antiarrhythmic Agents,* Pergamon Press, Oxford-New York 1981, pp. 151 – 194.
23. B. R. Lucchesi and E. S. Patterson in M. J. Antonaccio (ed.): *Cardiovascular Pharmacology,* Raven Press, New York 1984, pp. 329 – 414.
24. C. Kumana, J. Hamer in J. Hamer (ed.): *Drugs for Heart Disease,* Chapman & Hall, London 1979, pp. 44 – 147.
25. C. D. Swerdlow, J. O. Yu, E. Jacobson, S. Mann, R. A. Winkle, J. G. Griffin, D. Ross, J. W. Mason, *Am. J. Med.* **75** (1983) 36 – 42.

26. R. G. Heel, R. N. Brogden, T. M. Speight, G. S. Avery, *Drugs* **15** (1978) 331 – 368.
27. D. P. Zipes, P. J. Troup, *Am. J. Cardiol.* **41** (1978) 1005 – 1024.
28. A. Wirtzfeld (ed.): *"Disopyramid,"* Band III, Übersicht über die internationale Literatur 1977 – 1980, Excerpta Medica, Amsterdam 1981.
29. R. W. F. Campbell in J. Coltart, D. E. Jewitt (eds.): *Recent Developments in Cardiovascular Drugs,* Churchill Livingstone, Edinburgh 1982, pp. 94 – 100.
30. L. Mogensen in J. Coltart, D. E. Jewitt (eds.): *Recent Developments in Cardiovascular Drugs,* Churchill Livingstone, Edinburgh 1982, pp. 118 – 127.
31. B. Holmes, R. N. Brogden, R. G. Heel, T. M. Speight, G. S. Avery, *Drugs* **26** (1983) 93 – 123.
32. E. Sandøe, D. G. Julian, J. W. Bell (eds.): "*Management of ventricular tachycardia – role of mexiletine*", Proceedings of a Symposium held in Copenhagen, 25th – 27th May, 1978, Excerpta Medica, Amsterdam 1978.
33. C. Y. C. Chew, J. Collett, B. N. Singh, *Drugs* **17** (1979) 161 – 181.
34. J. Senges, L. Ehe, *Naunyn Schmiedebergs Arch. Pharmacol.* **280** (1973) 253 – 264.
35. H. Kleinsorge, *Med. Klin. (Munich)* **54** (1959) 409 – 416.
36. R. Bojorges, G. Pastelin, S. Sanchez-Perez, R. Mendez, E. Kabela, *J. Pharmacol. Exp. Ther.* **193** (1975) 182 – 193.
37. D. L. D. Keefe, R. E. Kates, D. C. Harrison, *Drugs* **22** (1981) 363 – 400.
38. H. Homburger, H. Antoni, *Arzneim. Forsch.* **24** (1974) 545 – 549.
39. E. Schaumlöffel, *Med. Welt* **25** (1974) 2008 – 2014.
40. W. Sziegoleit, H. W. Crodel, W. Förster, *Dtsch. Gesundheitswes.* **30** (1975) 2489 – 2491.
41. E. Carmelit, P. A. Janssen, R. Marsboom, J. M. Van Nueten, R. Xhonneux, *Arch. Int. Pharmacodyn. Ther.* **231** (1978) 104 – 130.
42. W. K. Amery, T. Aerts in A. Scriabine (ed.): *New Drugs Annual – Cardiovascular Drugs,* Raven Press, New York 1983, pp. 109 – 132.
43. F. Bender, G. Cronheim (eds.): *Flecainid: Experimentelle Befunde und klinische Erfahrungen,* Fischer Verlag, Stuttgart-New York 1982.
44. H. Hochrein, H.-J. Hapke, O. A. Beck (eds.): *Fortschritte in der Pharmakotherapie von Herzrhythmusstörungen, I. Internat. Propafenon-Symposium, 14th and 15th November 1975, in Berlin,* Fischer Verlag, Stuttgart–New York 1979.
45. M. Schlepper, B. Olsson (eds.): *Cardiac Arrhythmias, Proceedings 1st International Rytmonorm-Congress,* Springer Verlag, Berlin-Heidelberg-New York 1983.
46. J. D. Dukes, E. M. Vaughan Williams, *Eur. Heart J.* **5** (1984) 115 – 125.
47. P. Touboul, G. Kirkorian, G. Atallah in J. Coltart, D. E. Jewitt (eds.): *Recent Developments in Cardiovascular Drugs,* Churchill Livingstone, Edinburgh 1982, pp. 110 – 117.
48. R. S. Bexton, A. J. Camm, *Pharmacol. Ther.* **17** (1982) 315 – 355.
49. J. W. Mason, L. M. Hondeghem, B. G. Katzung, *Pflügers Arch.* **396** (1983) 79 – 81.
50. E. Patterson, B. R. Lucchesi, *Am. Heart J.* **106** (1983) no. 2, 426 – 431.
51. G. G. Belz, F. Bender: *Therapie der Herzrhythmusstörungen mit Verapamil Isoptin®,* Fischer Verlag, Stuttgart 1974.
52. B. N. Singh, G. Ellrodt , C. T. Peter, *Drugs* **16** (1978) 169 – 197.
53. A. Zanchetti, D. M. Krikler: "*Calcium antagonism in cardiovascular therapy: Experience with verapamil,*" International Symposium Florence 2nd – 4th Oct. 1980, Excerpta Medica, Amsterdam (1981).
54. A. Fleckenstein: *Calcium Antagonism in Heart and Smooth Muscle,* Wiley-Interscience, New York 1983.
55. H. U. Lehmann, C. Taegner-Torge, E. Witt, O. A. Beck, H. Hochrein, *Angiocardiology* **2** (1979) 103 – 111.
56. E. Winslow, K. A. Kane, *J. Cardiovasc. Pharmacol.* **3** (1981) 655 – 667.
57. J. M. Davy, J. F. Lainé, C. Sebag, G. Motté, *Circulation* **68** (1983) Suppl. III, abstr. 1241.
58. F. Bender, K. Greef (eds.): *"Calciumantagonisten zur Behandlung der Angina pectoris, Hypertonie und Arrhythmie,"* 1. Dilzem® Symposium Copenhagen 25th – 27th June 1981, Excerpta Medica, Amsterdam 1982.
59. A. Fleckenstein, K. Hashimoto, M. Herrmann, A. Schwartz, L. Seipel (eds.): *New Calcium Antagonists, Recent Development and Prospects, Diltiazem Workshop* held on May 10th, 1982 in Freiburg, Fischer Verlag, Stuttgart 1983.
60. R. A. Winkle, J. W. Mason, R. E. Kates, D. C. Harrison in J. Coltart, D. E. Jewitt (eds.): *Recent Developments of Cardiovascular Drugs,* Churchill Livingstone, Edinburgh 1982, pp. 102 – 109.

61. L. B. Mitchell, R. A. Winkle in A. Scriabine (ed.): *New Drugs Annual – Cardiovascular Drugs,* Raven Press, New York 1983, pp. 93 – 107.
62. N. N. Kipshidze, *Med. Actual* **15** (1979) No. 7, 320 – 323.
63. G. G. Gensini, R. F. Knope, A. E. Kelly (eds.): *Concepts of the Mechanism and Treatment of Arrhythmias,* Futura Publ. Company, Mount Kisco 1974.
64. E. E. Bagwell, T. Walle, D. E. Drayer, M. M. Reidenberg, J. K. Pruett, *J. Pharmacol. Exp. Ther.* **197** (1976) 38 – 48.

Antiasthmatic Agents

Hans Michael Jennewein, Boehringer Ingelheim Pharma KG, Ingelheim/Rhein, Federal Republic of Germany

Jorge Casals-Stenzel, Gauting-Königswiesen, Federal Republic of Germany

Kurt Schromm, Boehringer Ingelheim Pharma KG, Ingelheim/Rhein, Federal Republic of Germany

1. Introduction 1
2. Bronchodilators (Relievers) 3
2.1. Adrenoceptor Agonists 3
2.1.1. Aminocatechol Derivatives 3
2.1.2. Resorcinol Derivatives 6
2.1.3. Saligenin Derivatives 8
2.1.4. Further Arylalkylamines 9
2.2. Xanthine Derivatives 13
2.3. Anticholinergic Drugs 16
3. Antiinflammatory Drugs or Controllers (Glucocorticosteroids) 17
4. Antiallergic Therapy 18
5. Antileukotrienes 19
6. Newer Concepts in Asthma Therapy 20
7. References 20

1. Introduction

Asthma therapy is based on the complex pathophysiology of this disease. Because of the variety of stimuli and the multiple factors that are involved in the induction and development of asthma, its treatment involves a multiplicity of therapeutic approaches and drugs [1], [2].

Bronchial asthma is generally understood to be a chronic inflammatory pulmonary disease characterized by recurrent episodes of wheezy labored breathing with prolonged expiration accompanied by dry coughing and tough mucus. These symptoms are the clinical expression of a narrowing of the bronchi (bronchoconstriction), a bronchial mucosal thickening (edema, eosinophilic bronchial infiltration), bronchial wall remodeling, and an excessive mucus production with plugging of the conducting airways in the lungs. This more or less generalized airway changes result in an increased irritability or sensitivity of the bronchi (bronchial hyperreactivity) to a wide variety of stimuli (irritants), allergic and nonallergic. The obstruction is usually reversible, either spontaneously or in response to appropriate therapy.

Depending on its dynamics bronchial asthma is clinically categorized into *acute* (attacks, status asthmaticus) and *chronic* (persistent) asthma. It is further subdivided into *extrinsic* and *intrinsic* asthma depending on the nature of its causes.

Extrinsic or allergic asthma (50 % of cases) generally occurs in childhood or young adulthood (under the age of 40) and relates to such extrinsic factors as common allergens (pollens, house dust, animal dander, inhalants, and some foods and drugs) or occupationally encountered dust (sawdust, flour). The house dust mite *Dermatophagoides pteronyssinus* is often involved. Extrinsic asthma can be controlled effectively with drug therapy.

Intrinsic or nonallergic asthma generally occurs in middle-aged patients who have no family history of allergies and exhibit negative allergy tests. It relates to intrinsic factors that are not well understood, e.g., respiratory tract infections, exercise, emotional stress, changes in weather and temperature, cold air, and inhalation of such air pollutants as sulfur dioxide and other chemicals, e.g., diisocyanates, poly(vinyl chloride), phthalic anhydride, or ethylenediamine. Patients with intrinsic asthma do not respond well to drug therapy and may progressively develop a chronic form of asthma.

Bronchial hyperreactivity, the fundamental feature of asthma, has been attributed to a pathophysiologic mechanism, which involves dysfunction of sympathetic nerves mediated by β-adrenergic receptors, hyperfunction of the parasympathetic nerves (efferent fibers of the vagus nerves), and altered intracellular cyclic adenosine monophosphate (cAMP) – cyclic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_453

guanosine monophosphate (cGMP) ratios in bronchial smooth muscles (effector cells) or in mast cells [3–5]. Immunologic mechanisms (antigen – antibody reactions) also are involved. Remodeling of the bronchial wall due to the chronic inflammatory event may also contribute to airway hyperreactivity.

The smooth muscle of the airway is controlled by the autonomic nervous system and bronchoactive chemical mediators. The airways are supplied with both *sympathetic* and *parasympathetic* nerve fibers from the autonomic nervous system. At the bronchial smooth muscle preferentially β_2-adrenoreceptors mediate relaxation of the muscle, i.e., dilation of the bronchi. Stimulation of the parasympathetic nerve fibers (*Nervus vagus*) or the administration of acetylcholine causes bronchoconstriction.

The bronchial smooth muscle also responds to bronchoactive chemical mediators, including such bronchoconstrictors as histamine, serotonin, neurokinins, and prostaglandins (PG), e.g., PGF_{2a}. The leukotrienes C_4, D_4, and E_4, are also included in this group. Bronchodilating mediators are, e.g., PGE_1 and PGI_2 . All these mediators act via specific receptors on the bronchial smooth muscle. The chemical mediators capable of causing bronchoconstriction generally are released by precursors stored either in static cells fixed within tissue (the mast cells) or in mobile blood cells (the eosinophilic leukocytes) infiltrating the bronchial mucosa. The classic example of such a mediator release is the allergic challenge in allergic or extrinsic asthma, in which the allergen interacts with a specific immunoglobin E (IgE) antibody (antigen – antibody reaction, type I allergic reaction) fixed to the surface of mast cells. It is also possible that mediators are released by other nonallergic stimuli in nonallergic or intrinsic asthma.

The binding of the mediators histamine, leukotriene, PG, etc., or of β-adrenoceptor agonists to the corresponding receptors on the smooth muscle cells triggers intracellular changes in the concentrations of cAMP or cGMP. There is an optimal ratio of cAMP to cGMP. An increase in cAMP concentration or a fall in that of cGMP, triggered by β_2 -adrenergic sympathomimetics, PGE_1 , or phosphodiesterase inhibitors, such as theophylline, induces relaxation of smooth muscle. Conversely, a fall in cAMP concentration or a rise in that of cGMP triggered by vagal stimulation, cholinergic agonists, histamine, β-adrenergic antagonists, or $PGF_{2\alpha}$, causes contraction. Calcium plays an important role in the mast cell, where it interacts with cAMP in the regulation of granule release. The release of mediators (e.g., histamine and leukotrienes) is stimulated by a flow of calcium into the cell, possibly initiated by the IgE – antigen reaction at the cell surface.

Eosinophilic bronchial infiltration is one of the hallmarks of asthma. This infiltration is obviously caused by a dominant T lymphocyte helper cell subpopulation (so called T_{H2} cells → Blood, Chap. 2.1.3.), which, via secretion of various cytokines (i.e., IL-5), facilitates this eosinophilic inflammation. Eosinophils are thought to be the most important inflammatory cells in asthma.

Taking into consideration the pathophysiology of bronchial asthma, the therapy of this disease can be directed to several sites along the trigger-to-target organ response pathway. A causal therapy would include the prevention of exposure to precipitating factors, on the one hand, and the prevention of mast cell stimulation by hyposensitization or desensitization procedures (immunotherapy) on the other. Causal therapy is uncertain and difficult to accomplish. Current antiasthmatic therapy is therefore to a great extent symptomatic and its principal aim is to keep airway patency at the optimal level for each particular patient. The symptomatic therapy of asthma consists of bronchodilation by modification of the end-organ response, i.e., the smooth muscle response or by blockage of the cholinergic reflex pathways (vagal reflex) and of antiallergic therapy, including anti-inflammatory treatment, inhibition of antibody formation (immunosuppression), and prevention of mast cell degranulation or mediator antagonism. These symptomatic therapeutic measures are achieved by five main groups of substances: sympathomimetic agents, methylxanthines, anticholinergic drugs, corticosteroids, and disodium cromoglycate and related drugs. In the last few years new types of possible antiasthmatic agents have been proposed: leukotriene antagonists or inhibitors of leukotriene synthesis [6], [7].

2. Bronchodilators (Relievers)

This type of antiasthmatic drug acts on the bronchial smooth muscle and induces a relaxation of the airway muscle which leads to desired bronchodilation thus relieving the most prominent symptom of asthma. The term bronchodilator drug might be expected to cover any drug that causes dilation of the bronchi, but by common usage it is restricted to drugs with a fairly rapid action. Because this sort of therapy causes a more of less immediate relieve, therapy with bronchodilators is also known under the term reliever therapy.

2.1. Adrenoceptor Agonists

β-Receptor stimulation produces bronchodilation. This can be achieved by a wide variety of agents including those which release aminocatechols producing α and β effects (e.g., ephedrine), direct β stimulants which stimulate both β_1 and β_2 receptors (e.g. isoprenaline and adrenaline, which also has α effects), and drugs that are more selective in their stimulation of β_2 receptors (e.g., fenoterol, clenbuterol, and salbutamol) [8]. The development of relatively specific agents for β_2 stimulation left little place for ephedrine or other nonselective drugs that stimulate both β_1 and β_2 receptors and have adverse effects on the heart caused by β_1 stimulation (tachycardia, palpitations). The β_2-adrenoceptor agonistic bronchodilators can be used by inhalation, by oral, or by intravenous administration. In clinical practice the main problem in their use lies in the cardiovascular side effects and the skeletal muscle tremor caused by overdosage. Nevertheless, they are the drugs of choice in the treatment of acute bronchospasm. Long-acting β_2-adrenoceptor agonists have been introduced. They are especially useful in the treatment of nocturnal asthma attacks.

The parent compound of the sympathomimetic agents is 2-phenylethylamine. Substitutions on the benzene ring, on the α- and β-carbon atoms, or on the terminal amino group yield a wide range of compounds with sympathomimetic activity. They can be classified into aminocatechols, resorcinol derivatives, and saligenin derivatives depending on their chemical structure. Further substitutions on the benzene ring lead to other very active compounds, e.g., the long-acting clenbuterol, procaterol, or bitolterol.

Aminocatechols are largely eliminated by neuronal and extraneuronal uptake mechanisms at the sympathomimetic nerve terminals. Non-aminocatechols are resistant to both uptake mechanisms. The aminocatechols (adrenaline, isoprenaline, isoetharine, and rimiterol) containing a 3,4-benzenediol ring are also metabolized by the enzyme catechol-*O*-methyltransferase (COMT) to relatively inert 3-methoxy metabolites. Non-aminocatechols (resorcinol and saligenin derivatives) are not metabolized by this enzyme and thus their half-lifes are prolonged. Another widely distributed enzyme, monoamine oxidase (MAO), cleaves the aminocatechol between the α-carbon and the amino group. The removal of one or both of the hydroxyl groups from the benzene ring prevents the action of COMT, and substitution of the α-carbon atom and the amino group blocks oxidation by MAO.

2.1.1. Aminocatechol Derivatives

Epinephrine [*51-43-4*], adrenaline, 1-(3,4-dihydroxyphenyl)-2-(methylamino)ethanol, free base, $C_9H_{13}NO_3$, M_r 183.21, *mp* 211 – 212 °C (L-form); hydrochloride [*51-42-3*], $C_9H_{13}NO_3 \cdot HCl$, M_r 219.67, *mp* 147 – 154 °C (L-form); for pharmacology, see [9].

A naturally occurring hormone, L-epinephrine is produced by the medulla of the adrenal gland and released when the body is under stress (→ Hormones). It is the model for all other sympathomimetic drugs, but its clinical use has been largely supplanted.

Synthesis [10], [11]: see next page.

Trade names: Suprarenin (Hoechst, Germany), Medihaler Epi (Riker, United Kingdom, USA), Adrenalin (Parke-Davis, USA).

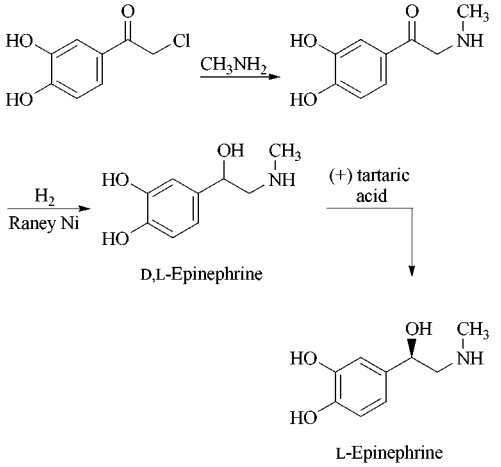

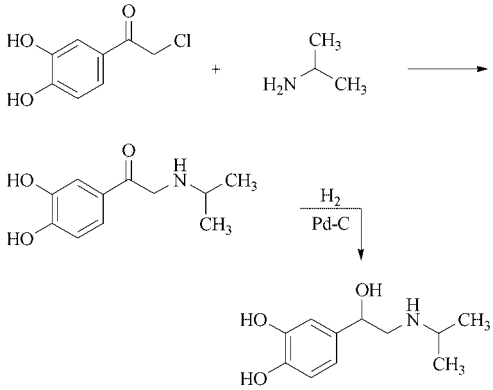

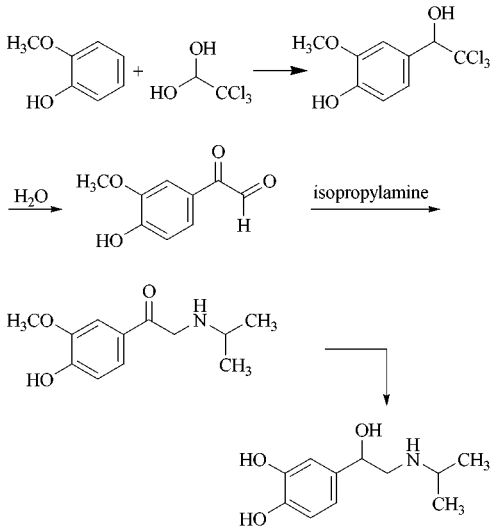

Isoproterenol [*149-53-1*], isoprenaline, 1-(3,4-dihydroxyphenyl)-2-(isopropylamino)ethanol, free base, $C_{11}H_{17}NO_3$, M_r 211.26, *mp* 155.5 °C (racemic form); hydrochloride [*949-36-0*], $C_{11}H_{17}NO_3 \cdot HCl$, M_r 247.72, *mp* 170 – 171 °C (racemic form); sulfate dihydrate [*6078-56-4*], $C_{11}H_{17}NO_3 \cdot 1/2\,H_2SO_4 \cdot 2\,H_2O$, M_r 296.33, *mp*128 °C (racemic form, some decomp.); for pharmacology, see [12].

Clinical use has been all but abandoned in Europe because isoproterenol is not selective for β_2 receptors.

Synthesis [13]: see right column, above.

Alternative synthesis [14]: see right column, centre.

Trade names: Aludrin (Boehringer Ingelheim, FRG), Aleudrin (Lewis, United Kingdom), Isomenyl (Kaken, Japan), Isuprel (Winthrop, USA).

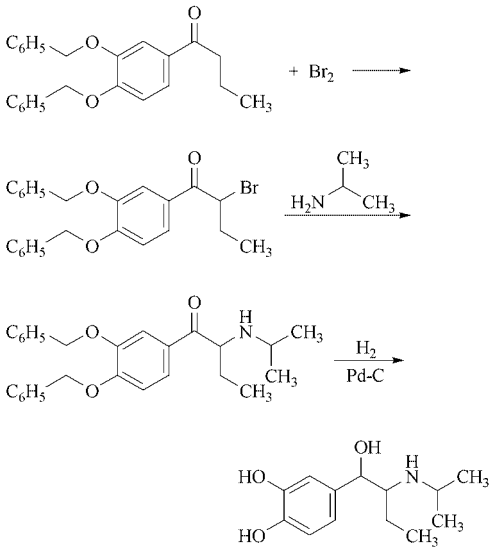

Isoetharine [*530-08-5*], 1-(3,4-dihydroxyphenyl)-2-(isopropylamino)butanol, free base, $C_{13}H_{21}NO_3$, M_r 239.32, hydrochloride [*50-96-4*], $C_{13}H_{21}NO_3 \cdot HCl$, M_r 275.78, *mp* 212 – 213 °C (decomp.); for pharmacology, see [15].

The clinical use of isoetharine is declining because of its low β_2 specificity, but it is still used in combinations.

Synthesis [16]: see right column, below.

Trade names: Asthmalitan (Kettelhack-Riker, FRG), Numotac (Riker, United Kingdom), Bronkosol (Breon, USA).

Protokylol [*136-70-9*], 1-(3,4-dihydroxyphenyl)-2-(α-methyl-3,4-methylenedioxyphenethylamino)ethanol, free base, $C_{18}H_{21}NO_5$, M_r 331.37, *mp* 163 °C; hydrochloride [*136-69-6*], $C_{18}H_{21}NO_5 \cdot HCl$, M_r 367.83, *mp* 126 – 127 °C; for pharmacology, see [17]:

Synthesis [18]:

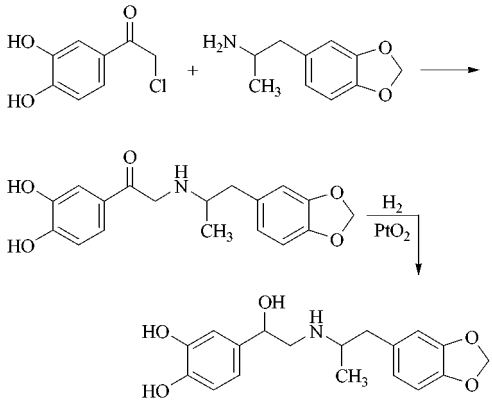

Trade names: Caytine (Chugai, Japan), Ventaire (Marion, USA).

Rimiterol [*32953-89-2*], 1-(3,4-dihydroxyphenyl)-1-(2-piperidinyl)methanol, free base, $C_{12}H_{17}NO_3$, M_r 223.27, *mp* 203 – 204 °C; hydrobromide [*31842-61-2*], $C_{12}H_{17}NO_3 \cdot HBr$, M_r 304.19, *mp* 220 °C (decomp.); for pharmacology, see [19].

Rimiterol is not in wide-spread clinical use.

Synthesis [20]:

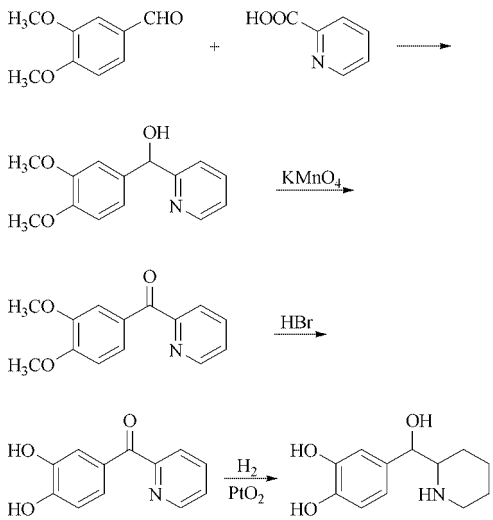

Trade name: Pulmadil (Riker, United Kingdom).

Trimethoquinol [*30418-38-3*], tetroquinol, (−)-1-(3,4,5-trimethoxybenzyl)-1,2,3,4-tetrahydro-6,7-isoquinolinediol, free base, $C_{19}H_{21}NO_5$, M_r 343.38, *mp* 125 – 126 °C; hydrochloride [*18559-59-6*], $C_{19}H_{21}NO_5 \cdot HCl$, M_r 379.84; for pharmacology, see [21].

Trimethoquinol is a β-sympathomimetic agent that inhibits the aggregation of platelets.

Synthesis of D,L-trimethoquinol [22]:

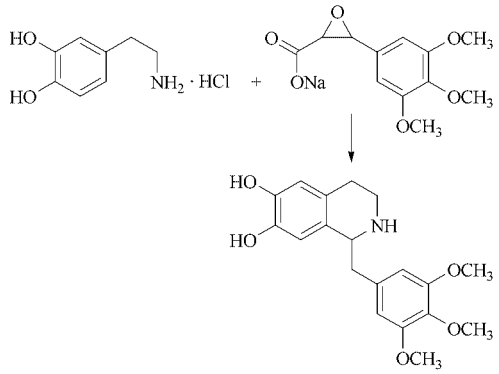

The racemic mixture is resolved using (+)-tartaric acid to yield L-trimethoquinol.

Trade names: Inolin (Tanabe, Japan), Vems (ISF, Italy).

Hexoprenaline [*3215-70-1*], *N,N′*-bis[2-(3,4-dihydroxyphenyl)-2-hydroxyethyl]hexamethylenediamine, free base, $C_{22}H_{32}N_2O_6$, M_r 420.51, *mp* 162 – 165 °C; dihydrochloride [*4323-43-7*], $C_{22}H_{32}N_2O_6 \cdot 2\,HCl$, M_r 493.43, *mp* 197.5 – 198 °C; sulfate [*32266-10-7*], $C_{22}H_{32}N_2O_6 \cdot H_2SO_4$, M_r 518.59, *mp* 222 – 228 °C; for pharmacology, see [23].

Synthesis [24]: see next page.

Trade names: Etoscol (Byk Gulden, Germany), Etoscol (Morishita, Japan).

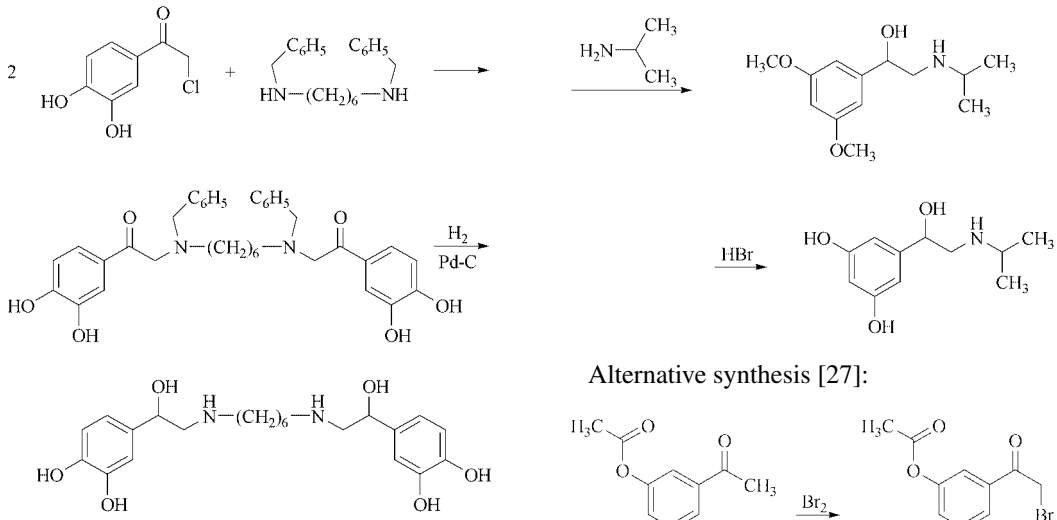

2.1.2. Resorcinol Derivatives

Metaproterenol [*586-06-1*], orciprenaline, 1-(3,5-dihydroxyphenyl)-2-(isopropylamino)-ethanol, free base, $C_{11}H_{17}NO_3$, M_r 211.26, *mp* 100 °C (racemic form); hydrochloride [*7104-40-7*], $C_{11}H_{17}NO_3 \cdot HCl$, M_r 247.72, *mp* 147 °C (racemic form); sulfate [*31023-56-0*], $C_{11}H_{17}NO_3 \cdot 1/2\,H_2SO_4$, M_r 260.30, *mp* 202 – 203 °C (racemic form); for pharmacology, see [25].

Metaproterenol was the first β-sympathomimetic drug with metabolic stability to achieve wide clinical use.

Synthesis [26]:

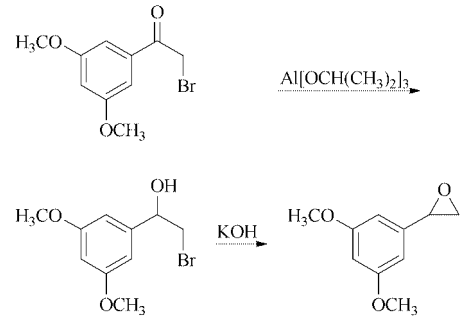

Alternative synthesis [27]:

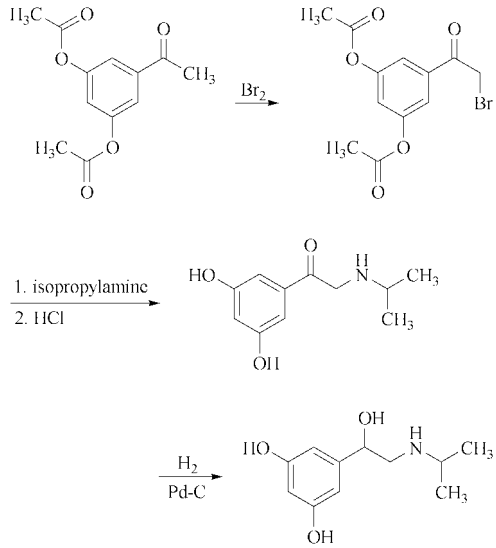

Trade names: Alupent (Boehringer Ingelheim, Germany, France, United Kingdom, USA), Alotec (Boehringer-Tanabe, Japan).

Terbutaline [*23031-25-6*], 1-(3,5-dihydroxyphenyl)-2-(*tert*-butylamino)ethanol, free base, $C_{12}H_{19}NO_3$, M_r 225.29; sulfate [*23031-32-5*], $C_{12}H_{19}NO_3 \cdot 1/2\,H_2SO_4$, M_r 274.33, *mp* 246 – 248 °C; for pharmacology, see [28].

Terbutaline is a specific β_2-sympathomimetic agent in wide clinical use.

Synthesis [29]: see next page.

Trade names: Bricanyl (Astra, Germany, United Kingdom, USA), Bricanyl (Astra-Injisawa, Japan), Terbasmin (Erba, Italy).

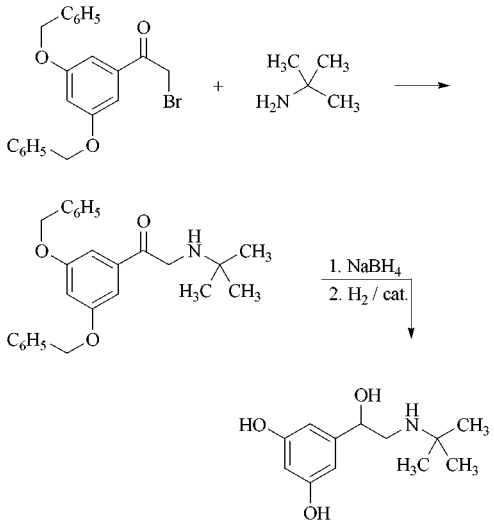

Fenoterol [*13392-18-2*], 1-(3,5-dihydroxyphenyl)-2-(4-hydroxy-α-methylphenethylamino)ethanol, free base, $C_{17}H_{21}NO_4$, M_r 303.36; hydrobromide [*1944-12-3*], $C_{17}H_{21}NO_4 \cdot HBr$, M_r 384.28, *mp* 232 – 233 °C; for pharmacology, see [30].

Fenoterol is a selective β_2-sympathomimetic agent that is in wide clinical use in Europe.

Synthesis [31]:

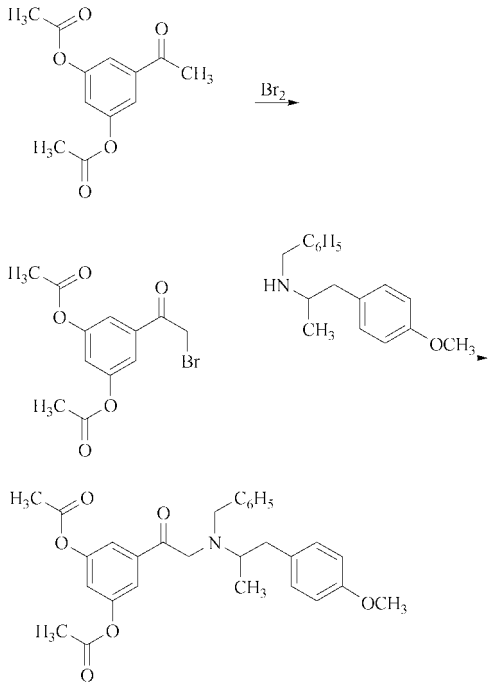

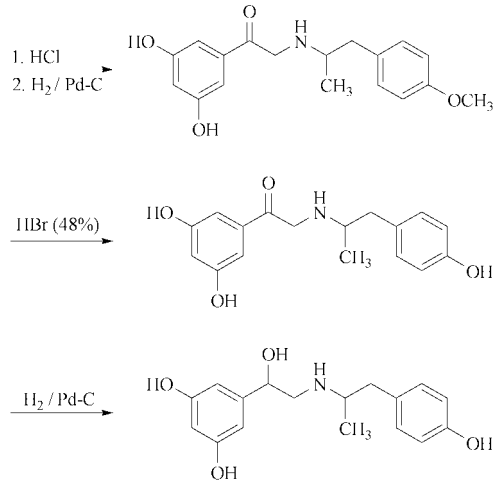

Trade names: Berotec (Boehringer Ingelheim, Germany, Switzerland, Sweden, the Netherlands, Canada), Berotec (Warner-Lambert, United Kingdom, Ireland), Respilac (Angeli, Italy).

Reproterol [*54063-54-6*], 7-[-3-[[2-(3,5-dihydroxyphenyl)-2-hydroxyethyl]amino]propyl]-3,7-dihydro-1,3-dimethyl-1*H*-purine-2,6-dione, free base, $C_{18}H_{23}N_5O_5$, M_r 389.41; hydrochloride [*13055-82-8*], $C_{18}H_{23}N_5O_5 \cdot HCl$, M_r 425.88, *mp* 249 – 250 °C; for pharmacology, see [32].

Reproterol is structurally a combination of a β-sympathomimetic compound and theophylline, but it shows only sympathomimetic activity.

Synthesis [33]:

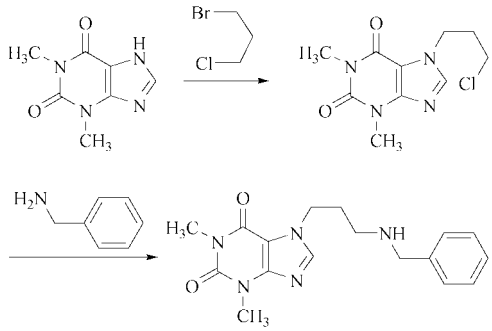

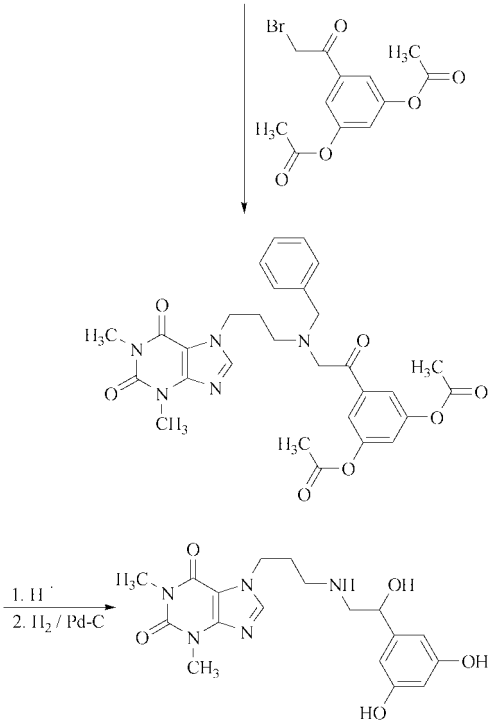

Trade names: Bronchospasmin (Chemiewerke Homburg, Germany), Bronchospasmin (Farmades, Japan), Bronchodil (Keymer, United Kingdom).

2.1.3. Saligenin Derivatives

Salbutamol [*18559-94-9*], albuterol, 2-(*tert*-butylamino)-1-(4-hydroxy-3-hydroxymethylphenyl)ethanol, free base, $C_{13}H_{21}NO_3$, M_r 239.32, *mp* 151 °C; sulfate [*51022-70-9*], $C_{13}H_{21}NO_3 \cdot 1/2\,H_2SO_4$, M_r 288.36; for pharmacology, see [34].

Salbutamol is a specific β_2-stimulant in wide clinical use in Europe and Japan.

Synthesis [35]:

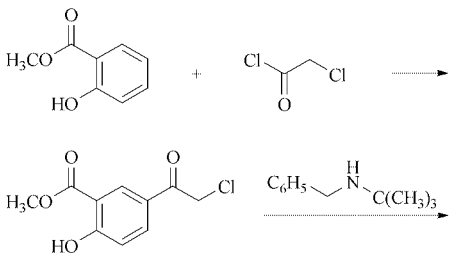

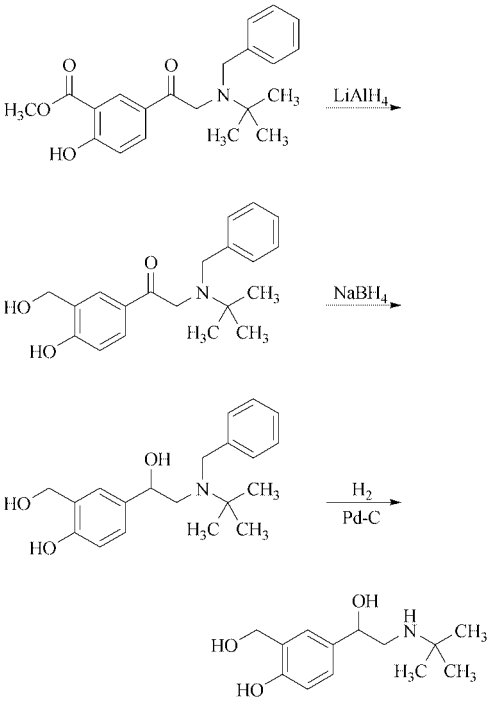

Alternative synthesis [36]:

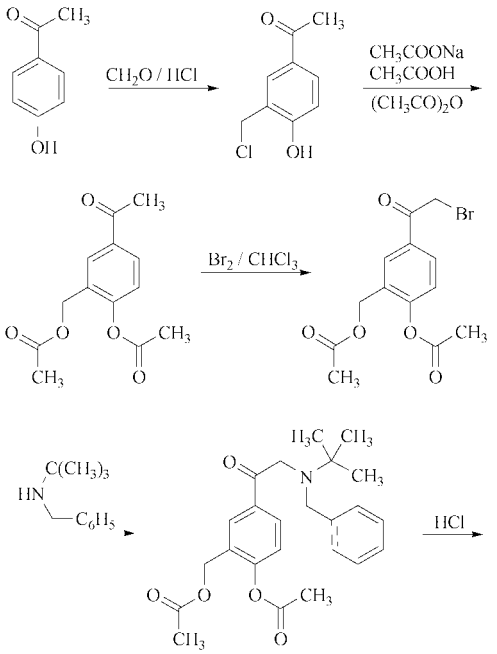

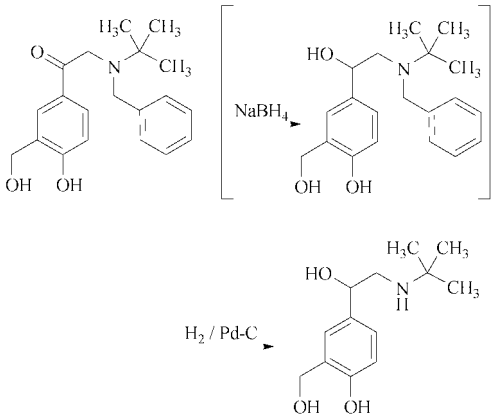

Trade names: Sultanol (Glaxo, Germany, Japan), Ventolin (Glaxo, France; Allen & Hanburys, United Kingdom; Sankyo, Japan), Broncovaleas (Valeas, Italy).

Salmeterol [*89365-50-4*], (±)-4-hydroxy-α^1-[[[6-(4-phenylbutoxy)hexyl]amino]methyl]-1,3-benzenedimethanol, 1-hydroxy-2-naphthoic acid salt [*94749-08-3*], $C_{25}H_{37}NO_4C_{11}H_8O_3$, M_r 603.75, *mp* 137 – 138 °C, for pharmacology see [90], [91]. Salmeterol is a long-acting β_2-agonist for patients with chronic asthma, useful also in controlling persistent nocturnal asthma. Synthesis [92]:

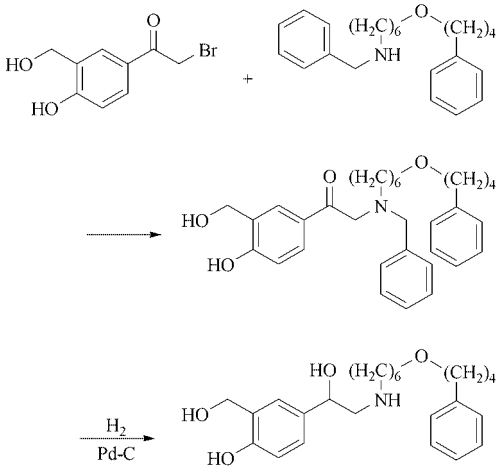

Trade name: Serevent (Glaxo Wellcome).

Pirbuterol [*38677-81-5*], α^6-[(*tert*-butylamino)methyl]-3-hydroxy-2,6-pyridinedimethanol, dihydrochloride [*38029-10-6*], $C_{12}H_{20}N_2O_3 \cdot 2HCl$, M_r 312.2, *mp* 182 °C (decomp.), acetate [*65652-44-0*], for pharmacology see [93]. Pirbuterol is a β_2-agonist in the management of disorders with reversible airways obstruction. It is given by inhalation and orally. Synthesis [94]:

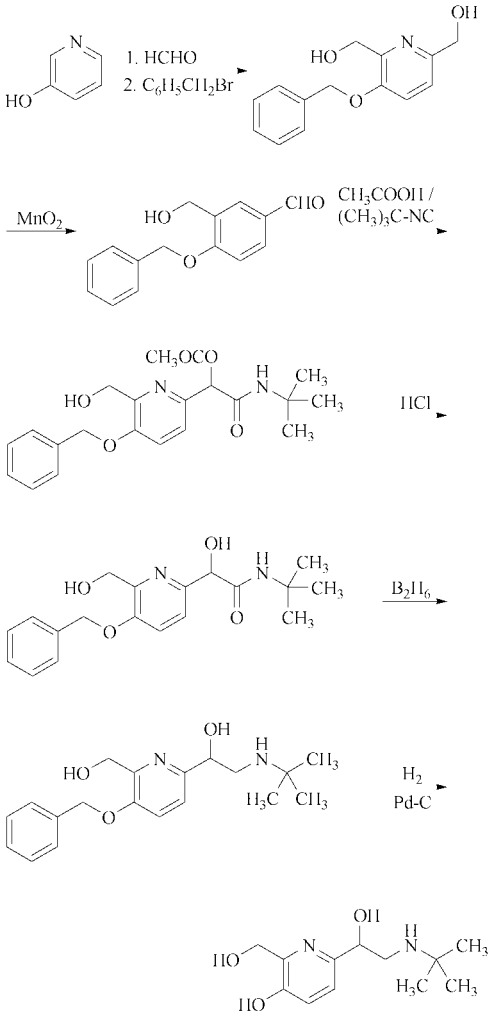

Trade names: Exirec, Spirolair, Noxair, Zeisin (3M Medical).

2.1.4. Further Arylalkylamines

Modern β_2-specific sympathomimetics with different substituents on the aromatic ring are described in this group along with ephedrine and thiadrine.

Ephedrine [*299-42-3*], 2-methylamino-1-phenylpropanol, free base, $C_{10}H_{15}NO$, M_r 165.24, *mp* 79 °C (racemic form); hydrochloride [*50-98-6*], $C_{10}H_{15}NO \cdot HCl$, M_r 201.70, *mp* 187 – 188 °C; sulfate [*134-72-5*], $C_{10}H_{15}NO \cdot 1/2\,H_2SO_4$, M_r 214.28, *mp* 247 °C; for pharmacology, see [37].

Ephedrine is a naturally occurring alkaloid drug (→ Alkaloids) derived from *Ephedra equistina* and *Ephedra vulgaris*. It can be prepared synthetically. It is used mainly in combination with other agents.

Synthesis of L-(−)-ephedrine [38]:

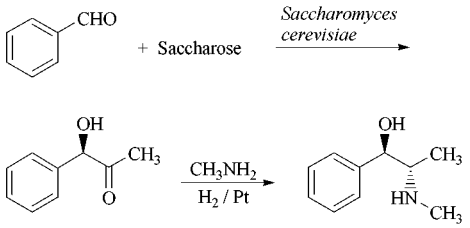

Trade names: Ephetonin (E. Merck, Germany), Ephedrin (Knoll, Germany), Spaneph (Smith Kline & French, United Kingdom), Ephedrine "Nagai" (Dainippon, Japan), Calcidrine (Abbott, USA).

Thiadrine [*14007-67-1*], 3,4-dimethyl-2-imino-5-phenylthiazolidine, free base, $C_{11}H_{14}N_2S$, M_r 206.31; thiocyanate, $C_{11}H_{14}N_2S \cdot HSCN$, M_r 265.40, *mp* 190 – 192 °C; for pharmacology, see [39].

Thiadrine is not in wide-spread clinical use as a bronchodilator.

Synthesis [40]:

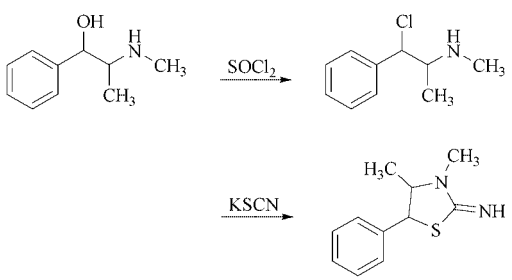

Trade name: Priatan (Chem. Werke Minden, Germany).

Clenbuterol [*37148-27-9*], 1-(4-amino-3,5-dichlorophenyl)-2-(*tert*-butylamino)ethanol, free base, $C_{12}H_{18}Cl_2N_2O$, M_r 277.20, *mp* 109.1 °C; hydrochloride [*21898-19-1*], $C_{12}H_{18}Cl_2N_2O \cdot HCl$, M_r 313.66, *mp* 174.0 – 175.5 °C; for pharmacology, see [41].

Clenbuterol is a long-acting, potent, and specific β_2 stimulator that is used orally.

Synthesis [42]:

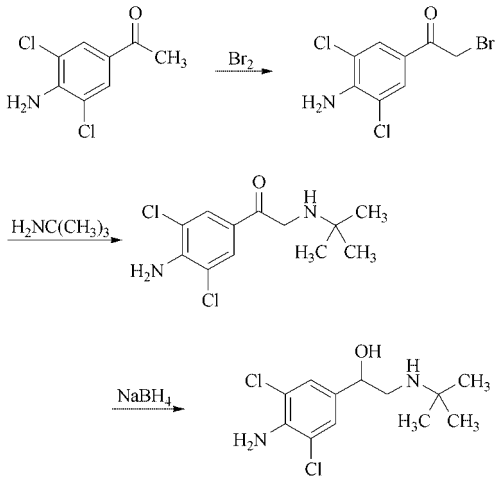

Alternative synthesis [43]:

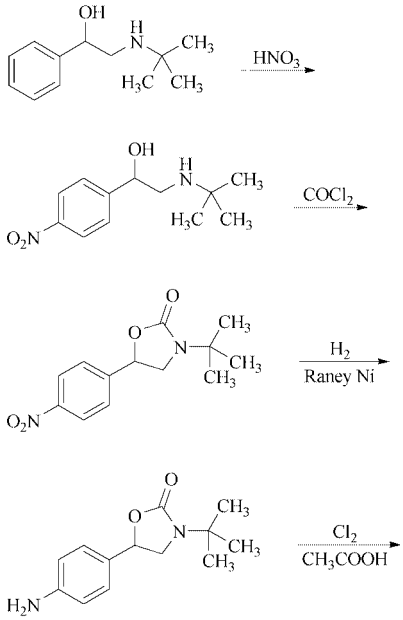

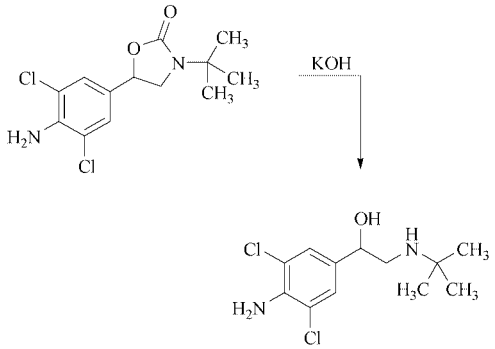

Trade name: Spiropent (Dr. Karl Thomae, Germany).

Mabuterol [*56341-08-3*], 1-(4-amino-3-chloro-5-trifluoromethylphenyl)-2-(*tert*-butylamino)ethanol, hydrochloride [*54240-36-7*], $C_{13}H_{18}ClF_3N_2O \cdot HCl$, M_r 347.21 for pharmacology see [44]. Mabuterol is an orally active β_2-adrenergic agonist related to Clenbuterol. Synthesis [45]:

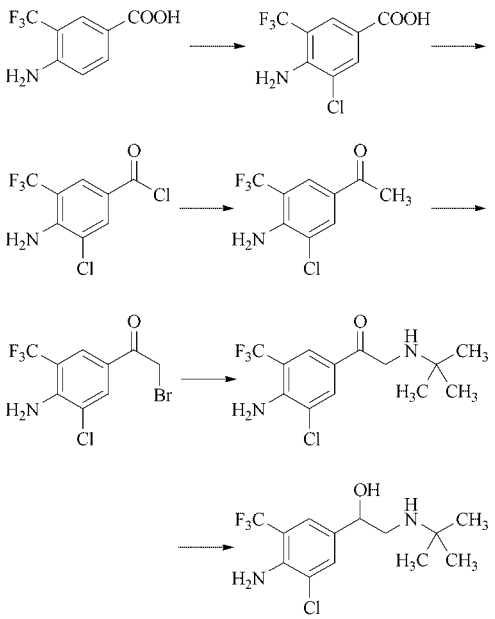

Trade name: Broncholin (Kaken Pharmaceutical).

Tulobuterol [*41570-61-0*], 2-(*tert*-butylamino)-1-(2-chlorophenyl)ethanol, free base, $C_{12}H_{18}ClNO$, M_r 227.74, *mp* 89 – 91 °C; hydrochloride [*56776-01-3*], $C_{12}H_{18}ClNO \cdot HCl$, M_r 264.20, *mp* 165 – 170 °C; for pharmacology, see [46].

Synthesis [47]:

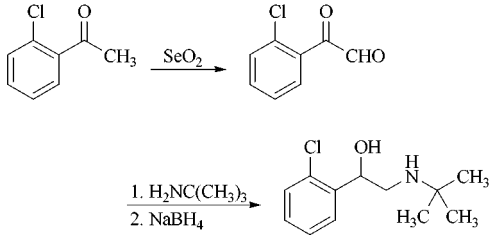

Trade names: Berachin (Toyo Tanabe, Japan), Hokunalin (Hokuriku Senjaku, Japan).

Carbuterol [*34866-47-2*], 1-(5-[2-(*tert*-butylamino)-1-hydroxyethyl]-2-hydroxyphenyl)urea, free base, $C_{13}H_{21}N_3O_3$, M_r 267.33, *mp* 205 – 207 °C; hydrochloride [*34866-46-1*], $C_{13}H_{21}N_3O_3 \cdot HCl$, M_r 303.79, *mp* 209 – 210 °C; for pharmacology, see [48].

Synthesis [49]:

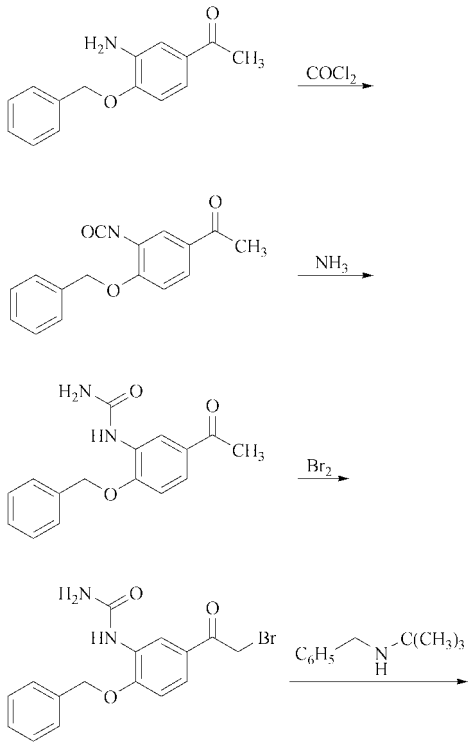

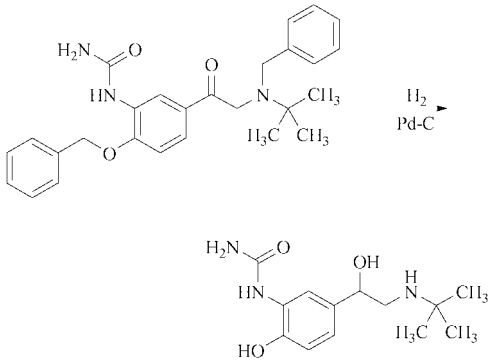

Trade names: Bronsecur (Smith Kline & French, Belgium, the Netherlands, South Africa), Pirem (Gödecke/ Sasse, FRG).

Formoterol [*73573-87-2*], (*RR,SS*)-(±)-*N*-[2-hydroxy-5-[1-hydroxy-2-[[2-(4-methoxyphenyl)-1-methylethyl]amino]ethyl]phenyl]-formamide, semifumarate [*43229-80-7*], $C_{21}H_{26}N_2O_6$, M_r 402.21, for pharmacology see [50], [51]. Formoterol is a potent bronchoselective β-agonist with long duration of action given by inhalation.

Synthesis [52]:

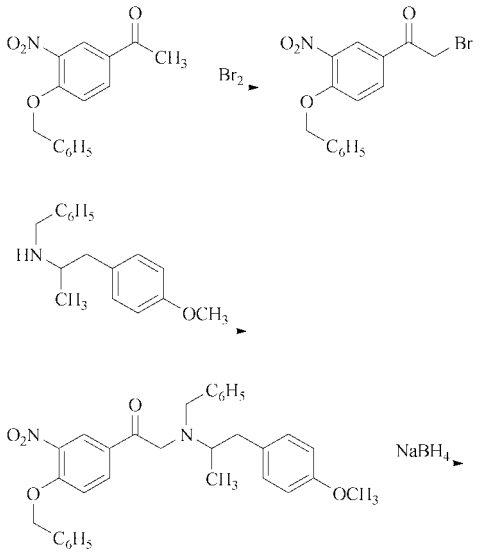

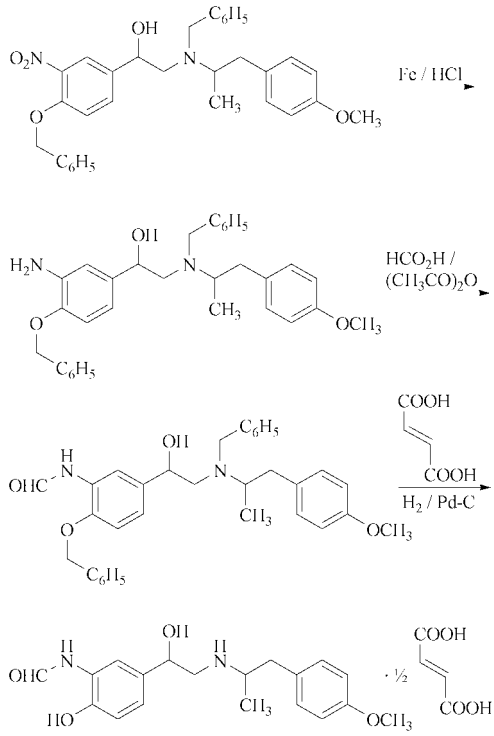

Trade names: Foradil (Novartis Pharma), Atock (Rhône Poulenc), Oxis (Astra/pharma stern.)

Procaterol [*72332-33-3*], 8-hydroxy-5-[1-hydroxy-2-(isopropylamino)butyl]-2(1*H*)-quinolinone, $C_{16}H_{22}N_2O_3$, M_r 290.37, *mp* 141 – 142 °C; hydrochloride [*62929-91-3*], $C_{16}H_{22}N_2O_3 \cdot HCl \cdot 1/2\,H_2O$, M_r 335.83, *mp* 193 – 197 °C (decomp.); for pharmacology, see [53].

Synthesis [54]: see left column of next page, above.

Trade name: Meptin (Otsuka, Japan).

Bitolterol [*30392-40-6*], 4-[2-(*tert*-butylamino)-1-hydroxyethyl]-1,2-phenylene-bis(*p*-toluate), free base, $C_{28}H_{31}NO_5$, M_r 461.56, *mp* 80 – 84 °C; mesylate [*30392-41-7*], $C_{28}H_{31}NO_5 \cdot CH_4O_3S$, M_r 557.67, *mp* 170 – 172 °C; for pharmacology, see [55].

Synthesis [56]: see left column of next page, below.

Trade names: Effectin (Shionogi, Japan), Tornalate (Sterling Drug, USA).

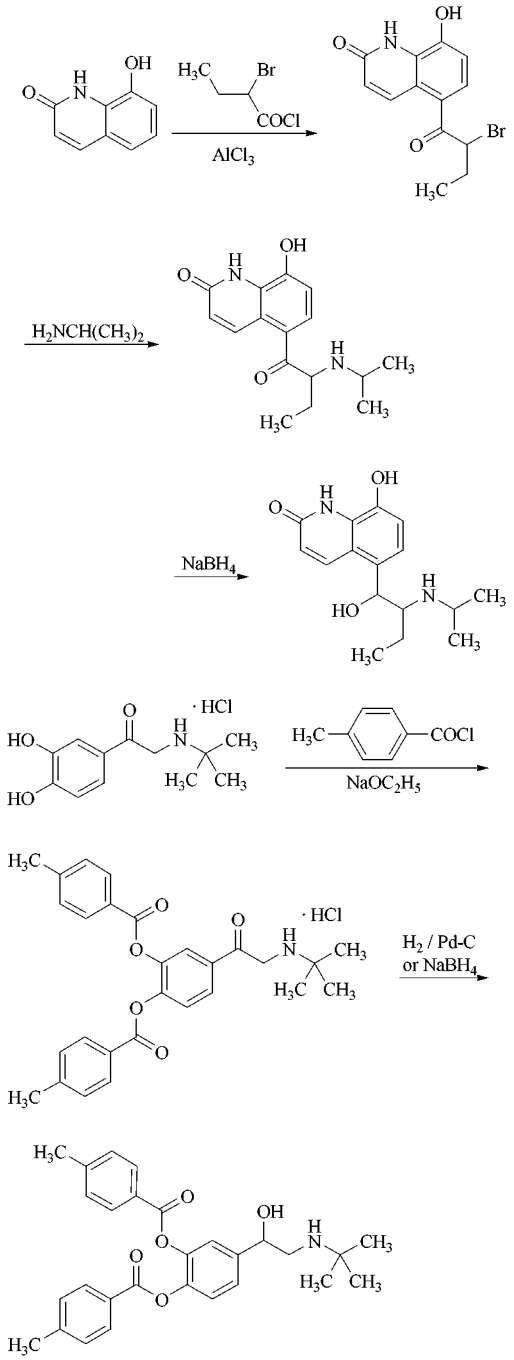

Bambuterol [*81732-65-2*], (±)5-[2-(*tert*-butylamino)-1-hydroxyethyl]-*m*-phenylene-bis-(dimethylcarbamate), $C_{18}H_{29}N_3O_5$, M_r 367.45, hydrochloride [*81732-46-9*], $C_{18}H_{29}N_3O_5 \cdot HCl$, M_r 403.9, for pharmacology see [57], [58]. Bambuterol is an inactive prodrug of terbutaline (see page 6). It has a prolonged duration of action when administered orally.

Synthesis [59]:

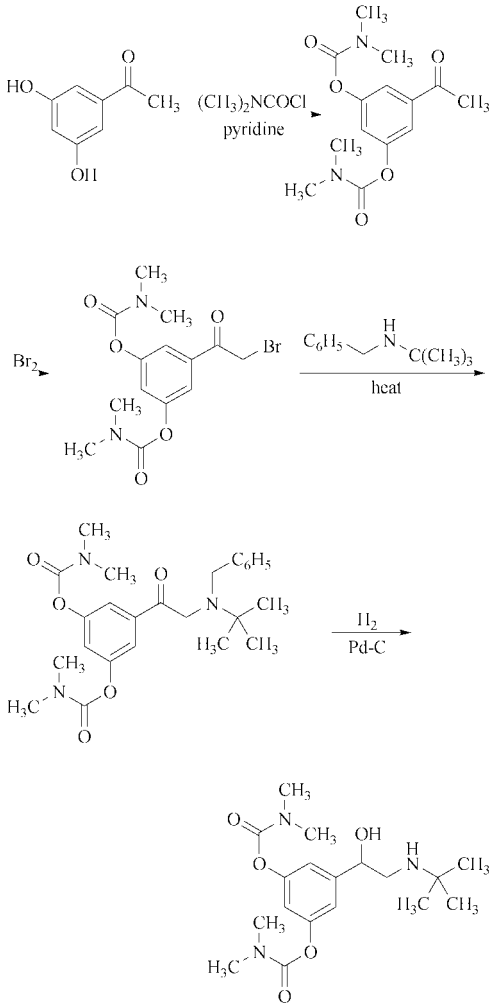

Trade name: Bambec (Astra/pharma stern, Germany).

2.2. Xanthine Derivatives

This group includes the familiar compounds present in the stimulant beverages, i.e., caffeine (coffee, tea, cola), theobromine (cocoa), and theophylline (tea) [60–62]. All are methylated derivatives of the natural metabolite xanthine, a

precursor of uric acid. Apart from their stimulating effects, methylxanthines possess a number of properties that are clinically useful (diuretic effect, increase in cardiac output, vasodilation). Most xanthines have bronchodilator activity, but only some are employed therapeutically for this purpose.

The following processes contribute to the mechanism of action of theophylline. Phosphodiesterase inhibition enhances monophosphate (cAMP). The intracellular metabolism and concentration of calcium in bronchial smooth muscle (bronchodilation) and in the mast cells (inhibition of mediator release) change consecutively. Adenosine receptor antagonism also may contribute to the relaxation of bronchial smooth muscle and to the inhibition of histamine release from mast cells; adenosine is known to increase this release. Theophylline also influences prostaglandin synthesis.

Theophylline preparations can be administered intravenously, orally, or rectally. They would be ineffective and irritating if administered as aerosols. The major pharmacological problem with these agents is the enormous individual variability in their rates of absorption and metabolism. The unpredictability of its levels in plasma after oral administration and the low therapeutic index of theophylline make it difficult to handle, and careful monitoring of plasma levels may be required. Theophylline must be used in different preparations because its solubility in water is poor (about 1 %). Therefore, theophylline itself is not suitable for intravenous use, but is sufficiently soluble for oral administration. Theophylline is prepared as a soluble salt of ethylenediamine or of choline (choline theophyllinate) for intravenous use. True theophylline derivatives are 7-substituted compounds, e.g., dihydroxypropyltheophylline (proxyphylline). These are less potent bronchodilators than theophylline, but they cause less intestinal irritation than theophylline or its salts and are much used in Europe.

Theophylline [*58-55-9*], 3,7-dihydro-1,3-dimethyl-1*H*-purine-2,6-dione, $C_7H_8N_4O_2$, M_r 180.17, *mp* 270 – 274 °C.

Synthesis [63], [64]: see next column, above.

Trade names: Solosin (Cassella-Riedel, Germany), Neulin (Riker, United Kingdom), Phyllin

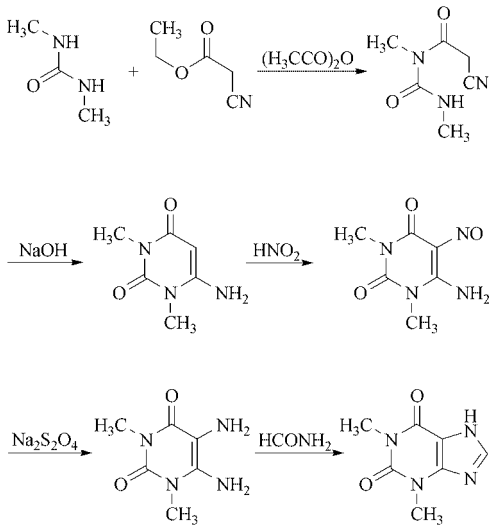

(Rona, United Kingdom), Theophyllol (Sankyo, Japan), Broncodyl (Breon, USA), Blixophyllin (Berlex, USA).

Theophylline ethylenediamine [*317-34-0*] is a salt of theophylline with ethylenediamine.

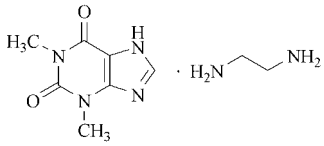

Trade names: Aminophylline (Promonta, Germany), Aminophylline (ICN; Searle; Wyeth; USA), Euphyllin (Byk Gulden, Germany), Cardophylin (Fisons, United Kingdom), Neophyllin (Eisai, Japan).

Etofylline [*519-37-9*], 3,7-dihydro-7-(2-hydroxyethyl)-1,3-dimethyl-1*H*-purine-2,6-dione, $C_9H_{12}N_4O_3$, M_r 224.22, *mp* 158 °C; for pharmacology, see [65].

Synthesis [66], [67]: see next page.

Trade names: Cordalin (Chemiewerke Homburg, Germany), Oxyphylline (Amido, France), Oxyphylline (Sankyo, Japan), Mesotin (Malescil, Italy).

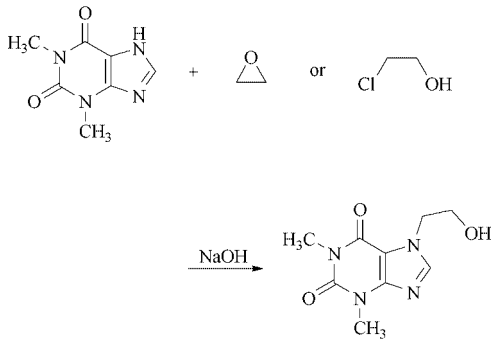

Proxyphylline [*603-00-9*], 3,7-dihydro-7-(2-hydroxypropyl)-1,3-dimethyl-1*H*-purine-2,6-dione, $C_{10}H_{14}N_4O_3$, M_r 238.25, *mp* 135 – 136 °C; for pharmacology, see [68].

Synthesis [69]:

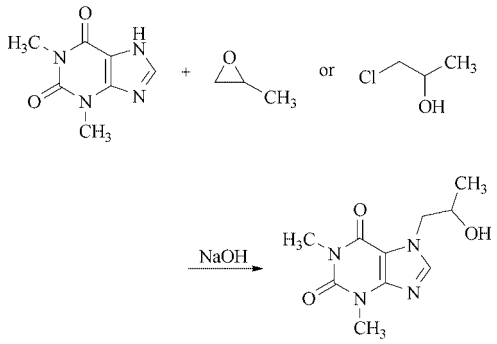

Trade names: Spantin (Pharmacia, Germany), Spasmolysin (Kade, Germany), Brontyl (Reckitt & Colman, United Kingdom), Monophylline (Yoshitomi, Japan).

Diprophylline [*479-18-5*], dyphylline, 7-(2,3-dihydroxypropyl)-3,7-dihydro-1,3-dimethyl-1*H*-purine-2,6-dione, $C_{10}H_{14}N_4O_4$, M_r 254.25, *mp* 158 °C.

Synthesis [67], [70]:

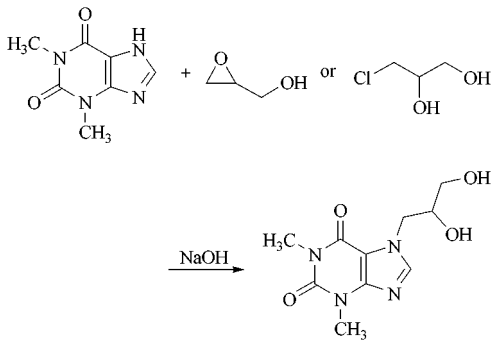

Trade names: Asthmolysin (Kade, Germany), Neutraphylline (Houdé, France), Neutraphylline (Cox-Continental, United Kingdom), Dihydrophylline (Tokyo Hosei, Japan), Neothylline (Lemmon, USA).

Acephylline piperazine [*18833-13-1*], acepifylline, piperazine salt of 1,2,3,6-tetrahydro-1,3-dimethyl-2,6-dioxo-7-purineacetic acid, $C_9H_{10}N_4O_4C_2H_5N$, M_r 281.27.

Synthesis [71]: The compound is synthesized by reacting theophylline with chloroacetic acid.

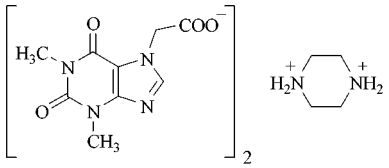

Trade name: Etaphydel (Delalande, Germany).

Bamifylline [*2016-63-9*], 3,7-dihydro-7-[*N*-ethyl-2-(hydroxyethylamino)ethyl]-8-benzyl-1,3-dimethyl-1*H*-purine-2,6-dione, free base, $C_{20}H_{27}N_5O_3$, M_r 385.47, *mp* 80 – 85.5 °C; hydrochloride [*20684-06-4*], $C_{20}H_{27}N_5O_3 \cdot HCl$, M_r 421.93, *mp* 185 – 186 °C; for pharmacology, see [72].

Synthesis [73]:

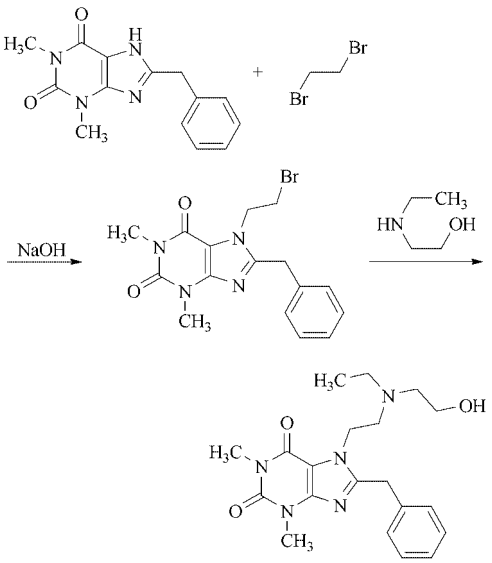

Trade name: Trentadil (Sedaph, France), Trentadil (Armour, United Kingdom).

2.3. Anticholinergic Drugs

The rationale for using anticholinergic drugs in the treatment of asthma is based on an irritant – bronchoconstrictor vagal reflex [74]. The allergen combines with antibodies on the surface of the mast cells releasing mediators that act on nervous receptors in the epithelium and elicit a reflex bronchoconstriction via the vagus nerves (parasympathetic nerves). Compounds, such as atropine (→ Alkaloids), which block transmission at parasympathetic neuroeffector junctions, inhibit vagally induced smooth muscle contraction and secretory activity. The clinical use of atropine has been discouraged because of fears of side effects (an increase in the sputum viscosity and a possible decrease in the ciliary activity). However, the clinical use of this type of compound has been reassessed in response to the introduction of ipratropium bromide [75] and oxitropium bromide. These drugs are inhaled. A long acting anticholinergic drug is tiotropiumbromide, which is in an advanced phase of development.

Atropine [*51-55-8*], D,L-hyoscyamine, 1αH,5αH-tropan-3α-ol (±)-tropate, *endo*-(±)-α-(hydroxymethyl)phenylacetic acid 8-methyl-8-azabicyclo[3.2.1]oct-3-yl ester, free base, $C_{17}H_{23}NO_3$, M_r 289.38, *mp* 114 – 116 °C; sulfate [*55-48-1*], $C_{17}H_{23}NO_3 \cdot 1/2\,H_2SO_4$, M_r 338.42, *mp* 190 – 194 °C; methylnitrate [*52-88-0*], $C_{17}H_{23}NO_3 \cdot CH_3NO_3$, M_r 366.42, *mp* 163 °C; for pharmacology, see [9].

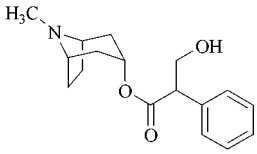

Atropine is an alkaloid drug (→ Alkaloids) found in the plants *Atropa belladona, Datura stramonium*, and *Hyoscyamus niger.* L-Hyoscyamine is obtained by extraction from these plants and then transformed into the racemic form using an alkaline ethanol solution [76].

Trade names: Bronchovycrin (Endopharm, Germany), Perasthman Inhalat (Bissantz, Germany).

Ipratropium bromide [*22254-24-6*], (8*r*)-3α-hydroxy-8-isopropyl-1αH,5αH-tropanium bromide (±)-tropate, $C_{20}H_{30}BrNO_3$, M_r 412.38, *mp* 230 – 232 °C; for pharmacology, see [75], [77].

Synthesis [78]:

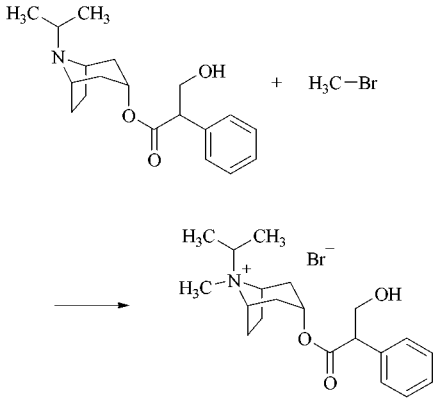

Trade name: Atrovent (Boehringer Ingelheim, Austria, Belgium, Germany, United Kingdom, Italy, Mexico, the Netherlands, Spain).

Oxitropium bromide [*30286-75-0*], 7(*S*)-(1α,2β,4β,5α,7β)-9-ethyl-7-(3-hydroxy-1-oxo-2phenylpropoxy)-9-methyl-3-oxa-9-azoniatricyclo[3.3.1.0^{2,4}]nonane bromide, $C_{19}H_{26}BrNO_4$, M_r 412.33, *mp* about 198 °C.

Synthesis [79]:

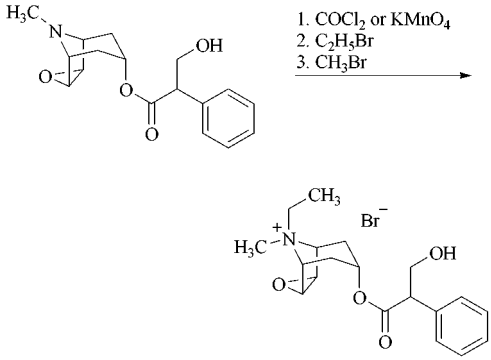

Trade name: Ventilat (Dieckmann, Germany), Tersigat (Laboratoires Francais de Thérapeutique, France).

Tiotropium bromide [*136310-93-5*] [7(*S*)-(1α,2β,4β,5α,7β]-7-[2-hydroxy-2,2-di(2-thienyl)acetoxy]-9,9-dimethyl-3-oxa-9-azoniatricyclo[3.3.1.0^{2,4}]nonane bromide, $C_{19}H_{22}BrNO_4S_2$, M_r 472.41, *mp* 226 – 227 °C (decomp.), bromide hydrate [*139404-48-1*], for pharmacology, see [80]

Synthesis [81]:

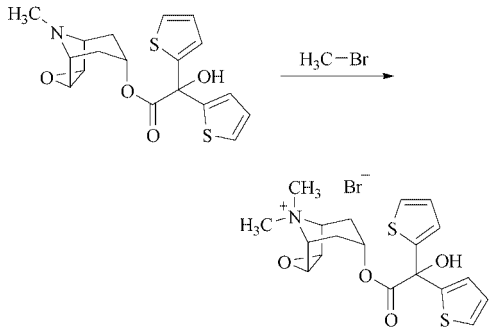

3. Antiinflammatory Drugs or Controllers (Glucocorticosteroids)

Glucocorticosteroids are the most potent agents available for the treatment of asthma [82]. When given orally, these substances cause adrenal suppression, metabolic side effects like obesity, osteoporosis, hyperglycemia, glycosuria, and disturbance of growth. However, when administered topically via inhalation glucocorticoid therapy can avoid these systemic side effect to a great extent. Therefore glucocorticoids given by inhalation have become the primary treatment of asthma. The mechanism of action of this type of compound is not fully understood, but it includes interaction with an intracellular glucocorticoid receptor [83]. The glucocorticoid receptor binds the steroid at its *C*-terminal part. After nuclear translocation the middle part of the receptor, which forms so-called "zinc fingers", interacts with DNA. An *N*-terminal domain then induces transcriptional transactivation of genes. In its inactivated form the glucocorticoid receptor is bound to a protein complex that includes heat shock proteins, immunophilline proteins, and others. The binding site of the DNA is termed glucocorticoid response elements (GRE). The interaction of the steroid receptor complex with GREs change the rate of transcription of steroid responsive genes resulting in either induction or repression of genes. Thereby production of inflammatory cytokines, enzymes, antiinflammatory proteins, receptors and adhesion molecules may be influenced. So for example β_2-receptors are upregulated which improves the efficacy of β-receptor agonists. This is known as the permissive effect of glucocorticoids.

Oral corticosteroid therapy suffers from a major drawback. The high dosage of steroids used to control inflammatory and immunologic reactions leads inevitably to unwanted, disturbing influences on metabolism; moreover withdrawal symptoms occur after long-term administration. Every effort should be made to avoid such long-term oral use. However, in severe case this cannot always be avoided. For intravenous administration during severe acute asthma (attacks), hydrocortisone (→ Hormones) is the preparation of choice, among other drugs. For oral corticosteroid therapy, prednisolone is preferred. For topical or inhalative treatment, aerosol preparations of beclomethasone dipropionate, betamethasone valerate, flunisolide, budenoside, and triamcinolone acetonide (→ Hormones) are available. In some countries fluticasone has already been introduced. Inhalative administration minimizes the metabolic side effects of the glucocorticoids. Side effects, like candidiasis of the mouth and hoarseness of the throat are usually mild and can be tolerated.

Beclomethasone [*4419-39-0*], 9α-chloro-11β,17α,21-trihydroxy-16β-methylpregna-1,4-diene-3,20-dione; 17,21-dipropionate [*5534-09-8*], $C_{28}H_{27}ClO_7$, M_r 510.97, *mp* 117 – 120 °C. Beclomethasone dipropionate is claimed to exert a topical effect on the lungs without significant systemic activity. It is used by inhalation, generally from a metered aerosol, for the prophylaxis of the symptoms of asthma [84], [85]. Chemical structure:

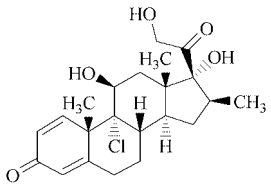

Trade names: Sanasthmax, Sanasthmyl (Glaxo, UK).

Flunisolide [*3385-03-3*], 6α-fluoro-11β,21-dihydroxy-16α,17α-isopropyldioxypregna-1,4-diene-3,20-dione, $C_{24}H_{31}FO_6$, M_r 434.5, hemihydrate [*77326-96-6*]. Flunisolide hemihydrate is administered by inhalation from metered aerosol and used in the management of asthma. Chemical structure:

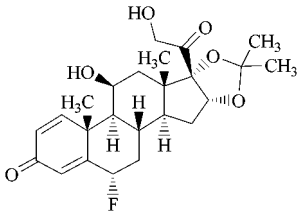

Trade name: Inhacort (Boehringer Ingelheim, Germany).

Budesonide [*51333-22-3*], an epimeric mixture of α- and β-propylforms of 16α,17α-butylidendioxy-11β-21dihydroxy-pregna-1,4-diene, $C_{25}H_{34}O_6$, M_r 430.54. Comparative investigations of the clinical efficacy/safety of budesonide and beclomethasone dipropionate administered by inhalation in patients with chronic bronchial asthma showed no differences [86], [87]. Nebulized budenoside was effective in severe childhood asthma [88]. Chemical structure:

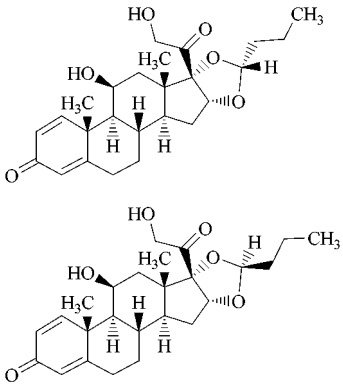

Trade name: Pulmicort (Astra/Pharma stern, Germany).

Fluticasone propionate [*80474-14-2*], *S*-fluoromethyl-6α,9α-difluoro-11β-hydroxy-16α-methyl-17α-propionyloxy-3-oxoandrosta-1,4-diene-17β-carbothioate, $C_{25}H_{31}F_3O_5S$, M_r 500.58, *mp* 272 – 273 °C (decomp.). Fluticasone propionate is claimed to exert a topical effect on the lungs without significant systemic effects, due to its low systemic bioavailability [89]. Chemical structure:

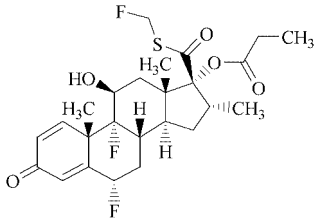

Trade name: Flutide (Glaxo Wellcome).

4. Antiallergic Therapy

Antiallergic therapy, including the use of corticosteroids, is an alternative to bronchodilator therapy in the treatment of asthma. The objectives of antiallergic therapy are the inhibition of mediator release, e.g., by disodium cromoglycate or related compounds (→ Antiallergic Agents, Chap. 3.) or by corticosteroids (→ Hormones, Chap. 4.), and the antagonism of mediators, e.g., by antihistamines or by the inhibition of either the activity or the synthesis of leukotrienes. The classical antihistamines are irrelevant to asthma therapy, presumably because of the presence and importance of other strong mediators. Newer antihistaminic agents, such as ketotifen [*34580-13-7*], have been employed in the treatment of asthma (→ Antiallergic Agents, Chap. 2.1.7.). Antagonists of cysteinyl-leukotrienes (i.e., leukotriene D_4 antagonists, zafirlukast, montelukast) or inhibitors of the enzyme 5-lipoxygenase are close to registration or already on the market (zileuton).

Disodium cromoglycate [*15826-37-6*] disodium 4,4'-dioxo-5,5'-(2-hydroxytrimethylenedioxy)di-14*H*-chromene-2-carboxylate, $C_{23}H_{14}Na_2O_{11}$, M_r 512.3, free acid [*16110-51-3*] is a relatively nontoxic agent used for the long-term prevention of many types of asthma [95]. It has no bronchodilator properties and is used in the prophylaxis of asthma.

It does not play a role in the treatment of acute asthmatic attacks including severe asthma [96], [97]. Chemical structure:

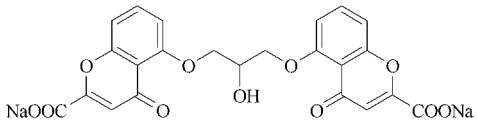

Disodium cromoglycate was developed from khellin, a natural substance that has been used for centuries to alleviate colic pains. It is a highly soluble powder that must be inhaled directly into the lung to be effective. Less than 1 % is absorbed when it is administered orally.

Trade name: Intal (Fisons).

Nedocromil sodium [*69049-74-7*], disodium 9-ethyl-6,9-dihydro-4,6-dioxo-10-propyl-4*H*-pyrano[3,2-*g*]quinoline-2,8-dicarboxylate, $C_{19}H_{15}Na_2O_7$, M_r 415.3; free acid [*69049-73-6*]. Nedocromil sodium has similar properties to sodium cromoglycate and can inhibit asthmatic responses provoked by allergen challenge and attenuate allergen-induced bronchial hyperresponsivness. It is rapidly absorbed from the lungs following inhalation and poorly absorbed from the gastrointestinal tract [98]. Chemical structure:

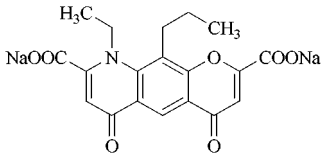

Trade name: Tilade (Fisons).

5. Antileukotrienes

It has been shown that leukotrienes (→ Prostaglandins, Chap. 2.) play a fundamental role in asthma. Especially the cystenyl leukotrienes are thought to be involved in the pathophysiology of asthma. The leukotriene antagonists montelukast and zafirlukast are the most prominent antileukotrienes which are registered or close to registration. Another possibility is to inhibit the synthesis of leukotrienes. The key enzyme is 5-lipoxygenase which, when inhibited, blocks the formation of leukotriene C_4, D_4, E_4 and of leukotriene B_4. Zileuton is the most advanced 5-lipoxygenase inhibitor, which has been introduced into the therapy of asthma in some countries.

Zafirlukast [*107753-78-6*], *N*-[4-[5-cyclopentyloxycarbonylamino)-1-methylindol-3-ylmethyl]-3-methoxybenzoyl]-2-methylbenzenesulfonamide, $C_{31}H_{33}N_3O_6S$, M_r 575.68. Chemical structure:

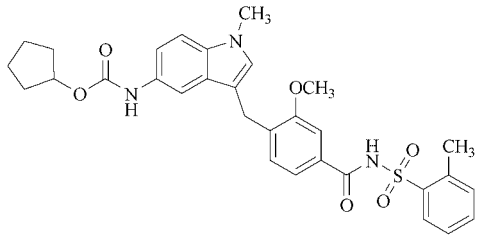

Zafirlukast is a potent LTC_4-, LTD_4- and LTD_4-antagonist and is clinically effective in patients with mild to moderate asthma [99], [100].

Trade name: Accolate (Zeneca).

Montelukast sodium [*151767-02-1*], sodium 2-[1-[1(*R*)-[3-[2(*E*)-(7-chloroquinolin-2-yl)vinyl]-phenyl]-3-[2-(1-hydroxy-1-methylethyl)phenyl]propylsulfanylmethyl]cyclopropyl]acetate, $C_{35}H_{35}ClNNaO_3S$, M_r 608.17, free acid [*158966-92-8*]. Chemical structure:

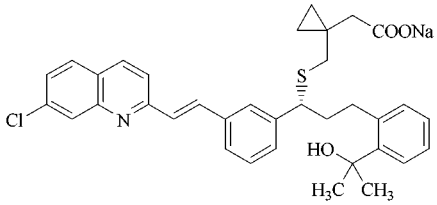

Montelukast sodium is a potent, selective competitive antagonist for cysteinyl leukotriene receptors such as LTD_4. It is orally active and shows therapeutic effects in patients with asthma [101].

Trade name: Singulair (Merck Frosst).

Zileuton [*111406-87-2*], (±)-*N*-(1-benzo[*b*]thien-2-ylethyl)-*N*-hydroxyurea, $C_{11}H_{12}N_2O_2S$, M_r 236.29. Chemical structure:

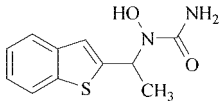

Zileuton is a potent and selective orally active 5-lipoxygenase inhibitor. It inhibited LTB_4 synthesis and was active in some inflammation models. Clinically it shows effects after acute and chronic administration in mild to moderate asthma [102], [103].

Trade name: Zyflo (Abbott).

6. Newer Concepts in Asthma Therapy

Various new and old approaches have been tried in the therapy of asthma. Calcium antagonists have no place in the routine management of asthma [7]. A substantial portion of ongoing research concentrates on phosphodiesterase inhibitors of type IV. The isoenzyme phosphodiesterase is located in inflammatory cells like the eosinophils which are important in asthma. There are also trials going on using humanized monoclonal antibodies against IgE, IL-5, IL-4, tumor necrosis factor α (TNFα) or adhesion molecules. Search for inhibitors of inducible NO synthetase is also continuing. Efforts are made to influence various transcription factors (i.e., NF_KB) or to find inhibitors of phospholipase A2 and antagonists for neurokinin receptors. Endothelin receptor antagonists are in development. Also potassium channel openers are still under evaluation. Time will show which of those various approaches will lead to novel therapies of asthma.

7. References

1. L. M. Lichtenstein, K. F. Austen: *Asthma: Physiology, Immunopharmacology and Treatment,* Academic Press, New York – San Francisco – London 1977.
2. M. E. Gershwin: *Bronchial Asthma, Principles of Diagnosis and Treatment,* Grune & Stratton, New York 1981.
3. A. Szentivanyi, *J. Allergy* **42** (1968) 203.
4. B. G. Simonsson, F. M. Jacobs, J. A. Nadel, *J. Clin. Invest.* **46** (1967) 1812.
5. W. M. Gold in J. A. Nadel (ed.): *Physiology and Pharmacology of the Airways,* Marcel Dekker, New York – Basel 1980, p. 123.
6. B. R. C. O'Driscoll, A. B. Kay, *Thorax* **37** (1982) 241.
7. E. Middleton, *J. Allergy Clin. Immunol.* **73** (1984) 643.
8. A. E. Tattersfield, *Pharmacol. Ther.* **17** (1982) 299.
9. A. Goodman Gilman, L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th ed., Macmillan Publ. Co., New York 1980.
10. F. Stolz, *Ber. Dtsch. Chem. Ges.* **37** (1904) 4149.
11. B. F. Tullar, *J. Am. Chem. Soc.* **70** (1948) 2067.
12. H. Konzett, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmacol.* **197** (1940) 27.
13. C. H. Boehringer Sohn, DE 723278, 1939.
14. D. Beke, O. Kovács, I. Fabricius, I. Lám, *Pharm. Zentralhalle Dtschl.* **92** (1953) 237.
15. A. M. Lands, F. P. Luduena, J. I. Grant, E. Ananenko, *J. Pharmacol. Exp. Ther.* **99** (1950) 45.
16. I.G. Farbenind., DE 638650, 1934.
17. R. A. Seibert, C. A. Handley, *J. Pharmacol. Exp. Ther.* **110** (1954) 304. A. Leslie, D. H. Simmons, *Am. J. Med. Sci.* **234** (1957) 321.
18. Lakeside Labs., US 2900415, 1959.
19. W. C. Bowman, I. W. Rodger, *Br. J. Pharmacol.* **45** (1972) 574.
20. G. H. Sankey, K. D. E. Whiting, *J. Heterocycl. Chem.* **9** (1972) 1049. Smith, Kline & French, US 3705169, 1972.
21. Y. Iwasowa, A. Kiyomoto, *Jpn. J. Pharmacol.* **17** (1967) 143.
22. Tanabe Sinjaku, GB 1114660, 1966.
23. R. M. Pinder, R. N. Brogden, T. M. Speight, G. S. Avery, *Drugs* **14** (1977) 1.
24. Oesterr. Stickstoffwerke, AT 241436, 1963.
25. A. Engelhardt, W. Hoefke, H. Wick, *Arzneim. Forsch.* **11** (1961) 521.
26. C. H. Boehringer Sohn, BE 611502, 1961.
27. C. H. Boehringer Sohn, DE 1275069, 1960.
28. J. Bergmann, H. Persson, K. Wetterlin, *Experientia* **25** (1969) 899.
29. Draco, GB 1199630, 1967.
30. H. P. Kuemmerle, G. Hitzenberger, G. M. Ling, H. Uehleke (eds.): "Berotec-Symposium 1971," *Int. J. Clin. Pharmacol. Ther. Toxicol.* **4** (1972) Suppl., 6 – 27.
31. C. H. Boehringer Sohn, DE 1286047, 1962.
32. S. Habersang, F. Leuschner, F. Stroman, A. Domenico, A. v. Schlichtegroll, *Arzneim. Forsch.* **27** (1977) 22.

33. K. H. Klingler, *Arzneim. Forsch.* **27** (1977) 4.
34. A. A. G. Lewis (ed.): "Salbutamol," *Postgrad. Med. J.* **47** (1971) Suppl., 11 – 34.
35. D. T. Collin, D. Hartley, D. Jack, L. H. C. Lunts, J. C. Press, A. C. Ritchie, P. Toon, *J. Med. Chem.* **13** (1970) 674.
36. Allen & Hanburys, GB 1200886, 1966.
37. D. M. Aviado: *Sympathomimetic Drugs,* Charles C. Thomas, Springfield, Ill., USA, 1970, p. 95.
38. R. H. F. Manske, T. B. Johnson, *J. Am. Chem. Soc.* **51** (1929) 580.
39. A. Kraushaar, *Ärztl. Wochenschr.* **5** (1950) 779.
40. P. Ballmer, CH 266997, 1948; US 2558068, 1949.
41. G. Engelhardt, *Arzneim. Forsch.* **26** (1976) 1404.
42. J. Keck, G. Krüger, K. Noll, H. Machleidt, *Arzneim. Forsch.* **22** (1972) 861.
43. Thomae, DE 1543928, 1966; DE 2157040, 1971.
44. G. Engelhardt et al., *Arzneimittel-Forsch.* **34** (1984) 1612.
45. J. Keck, G. Krüger, K. Noll, H. Machleidt, *Arzneim.-Forsch.* **22** (1972) 861.
46. S. Kubo, Y. Kasé, T. Mijata, G. Kito, I. Uesaka, *Arzneim. Forsch.* **25** (1975) 1028. S. Kubo, I. Uesaka, I. Matsubara, T. Ishihara, Y. Kasé, *Arzneim. Forsch.* **27** (1977) 1433.
47. Hokuriku Pharm. Co., JP 7170337, 1971.
48. J. R. Wardell, D. F. Colella, M. S. Shetzline, P. J. Fowler, *J. Pharmacol. Exp. Ther.* **189** (1974) 167.
49. C. Kaiser, D. F. Colella, M. S. Schwartz, E. Garrey, J. R. Wardell, *J. Med. Chem.* **17** (1974) 49.
50. H. Ida, *Arzneimittel-Forsch.* **26** (1976) 839, 1337.
51. D. Faulds et al., *Drugs* **42** (1991) 115.
52. Yamanouchi, DE 2 305 092, 1973, US 3 994 974, 1976.
53. S. Yamashita, M. Takai, Y. Yabuuchi, *J. Pharm. Pharmacol.* **30** (1978) 273.
54. S. Yoshizaki, K. Tanimura, S. Tamada, Y. Yabuuchi, K. Nakugawa, *J. Med. Chem.* **19** (1976) 1138.
55. H. Minatoya, B. F. Tullar, *Pharmacologist* **16** (1974) 211. H. Minatoya, *J. Pharmacol. Exp. Ther.* **206** (1979) 515.
56. Sterling Drug Inc., DE 2015573, 1970.
57. L. A. Svensson, *Pharm. Res.* 1985, 156.
58. O. A. Torsten, L. A. Svensson, *Pharm. Res.* 1984, 19.
59. A. B. Draco, EP 43 807, 1982, US 4 419 364 1983.
60. N. Svedmyr, *Scand. J. Respir. Dis.* **101** (1977) Suppl. 125.
61. J. W. Jenne in M. Stem (ed.): *New Directions in Asthma,* American College of Chest Physicians, Park Ridge, Ill., 1975, p. 391.
62. E. Schmid, M. von Bubnoff, R. Taugner, *Arzneim. Forsch.* **5** (1955) 194.
63. W. Traube, *Ber. Dtsch. Chem. Ges.* **33** (1900) 3035.
64. H. Bredereck, A. Edenhofer, *Chem. Ber.* **88** (1955) 1306.
65. A. Lespagnol, J. Vanlerenberghe, L. Masse, *Bull. Soc. Pharm. Lille* **2** (1948) 18.
66. L. Fabbrini, R. Cencioni, *Farmaco Ed. Sci.* **17** (1962) 660.
67. H. J. Roth, *Arch. Pharm. Ber. Dtsch. Pharm. Ges.* **292** (1959) 234.
68. C. Graffner, G. Johnsson, J. Sjögren, *Acta Pharm. Suec.* **10** (1973) 425.
69. Gane's Chem. Works, US 2715125, 1955.
70. State Univ. Iowa, US 2575344, 1951.
71. J. Blaisse, *Bull. Soc. Chim. Fr.* **16** (1949) 769.
72. A. Georges, J. Page, G. Duvernay, *Therapie* **17** (1962) 211.
73. A. Christiaens, BE 602888, 1961.
74. W. M. Gold, *Postgrad. Med. J.* **51** (1975) Suppl. 7, 53.
75. R. Bauer, S. Püschmann, H. Wick, *Arzneim. Forsch.* **26** (1976) 981.
76. F. Chemnitius, *J. Prakt. Chem.* **116** (1927) 276. W. Schneider, *Arch. Pharm. Ber. Dtsch. Pharm. Ges.* **284** (1951) 306.
77. R. Bauer, F.-J. Kuhn, K. Stockhaus, H. Wick, *Arzneim. Forsch.* **26** (1976) 974.
78. W. Schulz, R. Banholzer, K.-H. Pook, *Arzneim. Forsch.* **26** (1976) 960.
79. Boehringer Ingelheim GmbH, DE 1670048, 1966.
80. B. Disse et al., *Life Sci* **52** (1993) 537.
81. Boehringer Ingelheim, DE 3 931 041, 1991; EP 418 716, 1991.
82. J. V. Collins, T. J. H. Clark, D. Brown, J. Townsend, *Q. J. Med.* **44** (1975) 259.
83. P. J. Barnes, *J. Allerg. Clin. Immunol.* **97** (1996) 159.
84. J. M. Cox, R. M. Mowers, *DICP Ann. Pharmacother.* **23** (1989) 597.
85. W. Pedersen, D. Prahl, *Allergy* **42** (1987) 272.
86. R. F. Wiley et al., *Chest* **76** (1982) 61.
87. O. Sebroos et al., *Allergy* **49** (1994) 833.
88. F. E. McCarthy, *Lancet* **1989i,** 380.
89. S. M. Holliday, *Drugs* **47** (1994) 318.
90. M. Johnson, *Lung* **168** (1990) Suppl., 115.

91. R. N. Brogden, D. Faulds, *Drug* **42** (1991) 895.
92. Glaxo, US 4 992 474, 1991.
93. *Drugs of the Future II* (1977) 284.
94. Pfizer, DE 2 204 195, 1972, US 3 700 681, 1972.
95. M. K. Church, *Med. Actual* **14** (1978) 281.
96. S. Murphy, H. W. Kelly, *Drug Intell. Clin. Pharm.* **21** (1987) 22.
97. T. L. Petty et al., *Am. Rev. Respir. Dis.* **139** (1989) 694.
98. R. N. Brogden, E. M. Sorkin, *Drugs* **45** (1993) 693.
99. *Ann. Drug Data Rep.* **18** (1996) 789.
100. I. S. Kelloway, *Ann. Pharmacother.* **31** (1997) 1012.
101. *Drugs of the Future* **22** (1997) 1103.
102. *Drugs of the Future* **18** (1993) 616.
103. *Ann. Drug Data Rep.* **19** (1997) 609.

Antiatherosclerotic Agents → Cardiovascular Drugs
Antibacterial Chemotherapeutics → Chemotherapeutics

Antibiotics

Masaji Ohno, Faculty of Pharmaceutical Sciences, University of Tokyo, Tokyo, Japan (Chaps. 1 – 4)

Masami Otsuka, Faculty of Pharmaceutical Sciences, University of Tokyo, Tokyo, Japan (Chaps. 1 – 4)

Morimasa Yagisawa, Japan Antibiotics Research Association, Tokyo, Japan (Chaps. 1 – 4)

Shinichi Kondo, Institute of Microbial Chemistry, Tokyo, Japan (Chaps. 1 – 4)

Heinz Öppinger, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chaps. 5 – 7)

Hinrich Hoffmann, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chaps. 5 – 7)

Dieter Sukatsch, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chaps. 5 – 7)

Leo Hepner, L. Hepner and Associates, Ltd., London, United Kingdom (Chap. 8)

Celia Male, L. Hepner and Associates, Ltd., London, United Kingdom (Chap. 8)

1. Introduction 2
1.1. General Definition 2
1.2. Historical Development and Classification 2
1.3. Nomenclature 4
2. Chemotherapeutic Use of Antibiotics 4
2.1. Microbial Pathogens 4
2.2. Tumor Cells 4
2.3. Enzyme Inhibitors 4
2.4. Chemotherapeutic Uses 5
2.5. Use in Agriculture 5
2.6. Resistance 5
2.7. Units 5
2.8. Analysis 6
3. Classification of Antibiotics 6
3.1. β-Lactams 6
3.1.1. Natural Penicillins 6
3.1.2. Semisynthetic Penicillins 7
3.1.3. Natural Cephalosporins 8
3.1.4. Semisynthetic Cephalosporins 8
3.1.5. Cephamycins 9
3.1.6. 1-Oxacephems 11
3.1.7. Clavulanic Acids 12
3.1.8. Penems 13
3.1.9. Carbapenems 14
3.1.10. Nocardicins 16
3.1.11. Monobactams 16
3.2. Tetracyclines 17
3.2.1. Structure and Properties 17
3.2.2. Anhydrotetracyclines 17
3.2.3. Anthracyclines 17
3.3. Aminoglycosides 18
3.4. Nucleosides 24
3.4.1. *N*-Nucleosides 25
3.4.2. *C*-Nucleosides 26
3.4.3. Carbocyclic Nucleosides 27
3.4.4. An Exceptional Nucleoside 28
3.5. Macrolides 28
3.5.1. 12-Membered Ring Macrolides . . . 28
3.5.2. 14-Membered Ring Macrolides . . . 28
3.5.3. 16-Membered Ring Macrolides . . . 29
3.6. Ansamycins 30
3.7. Peptides 31
3.7.1. The Bleomycin Group 32
3.7.2. The Gramicidin Group 34
3.7.3. The Polymyxins 34
3.7.4. The Bacitracins 35
3.7.5. Large-Ring Peptide Antibiotics Containing Lactone Linkages 35
3.7.6. The Actinomycin Group 35
3.7.7. Other Peptide Antibiotics 35
3.8. Other Important Antibiotics and Intermediates 35
4. Individual Antibiotics Use 42
5. Fermentation 76
5.1. Screening 76
5.2. Selection, Mutation, and Maintenance of Strains 77
5.3. Process Development Leading to Large-Scale Production 78
5.4. Fermentation Technology 80
5.4.1. Maintenance of the Strain and Production of Inoculum 80
5.4.2. Treatment Before and During Fermentation 82
6. Isolation and Purification of Antibiotics; Quality Specifications 83
6.1. Isolation 83

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a02_467

6.2. Purification Techniques, Sterile End Products, Official Regulations 85
7. Analytical Measurements and Quality Control 87
7.1. Microbiological Analysis 88
7.2. Isotopically Labeled Antibiotics . 89
8. Economic Aspects 90
9. References 91

1. Introduction

1.1. General Definition

In 1942, WAKSMAN defined antibiotics as chemical substances produced by microorganisms and capable of inhibiting the growth of microorganisms [20]. Great effort has been devoted to the worldwide search for new antibiotics, and numerous compounds possessing various biological activities, that is, antibacterial, antiviral, antifungal, antitumor, and enzyme-inhibiting activities, have been discovered. These substances are mostly of microbial origin but are also semisynthetic in some cases. They have a wide variety of structural characteristics. Each is entered in the index of antibiotics. The area defined by the term "antibiotics" is therefore expanding, although WAKSMAN's original definition is still basically valid.

1.2. Historical Development and Classification

In 1877, PASTEUR observed that saprophytic bacteria inhibited the growth of pathogenic anthrax organisms. His was the first scientific description of the antagonism phenomenon. Production of a certain metabolic substance seemed to be responsible for the inhibition. PASTEUR suggested the therapeutic potential of this type of growth repression. VUILLEMIN used the term "antibiosis" to describe the inhibition of the growth of one organism by another. The potential utility of "bacteriotherapy" was recognized and enormous experimental efforts were made to investigate the antagonism phenomenon.

In 1894, METCHNIKOFF reported the repressive effect of *Pseudomonas* on *Vibrio cholerae.* From the culture of a *Penicillium,* GOSIO isolated an antibacterial crystalline substance, mycophenolic acid, in 1896. Other results also demonstrated the ability of various microbes to produce antibacterial substances [21].

It was in 1929 that FLEMING observed that a culture of a *Penicillium* inhibited the growth of bacteria [22]. He demonstrated the production of an antibacterial substance in the culture broth and named it penicillin. Although he suggested the promising therapeutic utility of penicillin, none of the attempts to isolate penicillin were successful and immediate attention was not attracted for the next decade. Instead, synthetic chemotherapeutics, such as sulfonamides, became objects of general interest after the discovery of prontosil by DOMAGK in 1935 [23]. The antibiotics, many of which were known even before the discovery of penicillin, remained without great importance for decades. Only the outbreak of the Second World War in 1939 led to an intense worldwide search for drugs to treat infections and wounds. Toward the end of the 1930s, FLOREY, CHAIN, and co-workers began to investigate penicillin in the course of their systematic study of antibacterial substances. They demonstrated the marked activity and therapeutic value of penicillin in 1940 and the "antibiotic era" began [24]. The production of penicillin had until that time been unsatisfactory, and favorable conditions for the effective formation of the antibiotic were explored in the United States. An active culture of the penicillin-producing organism was sought and submerged fermentation was developed. The use of lactose as a carbon source and the addition of cornsteep liquor to the nutrients were found effective. Irradiation of the culture with X-rays or ultraviolet light produced mutant strains. These findings set the stage for the industrial production of penicillin.

After the discovery of penicillin, BROTZU began a search for antibiotic-producing organisms and examined a culture of *Cephalosporium* spp. isolated from the sea near a sewage outlet in Sardinia. It secreted substances active against gram-positive bacteria. In September of 1948, BROTZU sent his organism to ABRAHAM at Oxford for detailed inspection. Several antibiotics were isolated from the culture and named cephalosporin [25]. Particular attention was at-

tracted by cephalosporin C, crystallized from the crude mixture of antibiotics, because of its low toxicity and its resistance to penicillinase.

Penicillins and cephalosporins, both of which possess a β-lactam ring as a structural characteristic, are designated β-lactam antibiotics. Extensive attempts to improve their antibacterial spectra through chemical modifications led to the development of many kinds of semisynthetic penicillins and cephalosporins. Moreover, nontraditional β-lactams have been discovered and commercialized [26–29].

Other novel antibiotics were discovered among the products of fungi, bacteria, and actinomycetes [30]. As a result of a search for water-soluble and heat-stable substances which would be active against gram-negative bacteria, WAKSMAN isolated actinomycin in 1940, streptothricin in 1942, and streptomycin in 1944 from cultures of actinomycetes. Various other antibiotics were isolated from microbes in France, Germany, Japan, the United Kingdom, the United States, and other countries.

Actinomycin is a peptide antibiotic effective against tumor cells as well as bacteria. Peptide antibiotics, one of the major antibiotics groups, possess diverse activities, i.e., antibacterial, antifungal, and antitumor activities. Their structures are varied and feature complicated modes of connection of often unusual amino acids. Ring peptides, linear peptides, lactonic peptides, and peptides containing hydroxy acids have been isolated. The group includes actinomycins, gramicidins, polymyxins, and colistins.

Streptomycin is used for infections of gram-positive and gram-negative bacteria and as a specific medicine for tuberculosis. Kanamycin, discovered by UMEZAWA in 1957, is especially effective against resistant bacteria. Paromomycin, spectinomycin, and ribostamycin were isolated later. These antibiotics are called aminoglycosides because their structural units are amino sugars, sugars, and amino acids. They are water soluble and basic in nature. The mechanism of resistance to aminoglycosides has been closely investigated and derivatives for use active against the resistant bacteria have been developed.

A yellow substance showing antibacterial activity was found in 1948 among products of *Streptomyces aureofaciens* and was named aureomycin. Terramycin was isolated from the fermentation broth of *Streptomyces rimosus.* The chemical and structural similarities of the two soon became apparent; they each have a linearly fused tetracyclic structure of six-membered rings. This parent skeleton is designated "tetracycline," and aureomycin and terramycin are now called chlortetracycline and oxytetracycline, respectively.

Chloramphenicol was first isolated as a product of *Streptomyces venezuelae;* it showed a broad antimicrobial spectrum. Chloramphenicol is an unusual natural product because it possesses both chloro and nitro groups in its structure. It is the only antibiotic produced commercially by an entirely chemical synthesis.

Macrolides are macrocyclic lactones to which sugars are attached. Various clinically important antibiotics are macrolides. Erythromycin, isolated in 1952, has two sugars connected to different sites on the 14-membered ring aglycone. Dimethylamino sugars are often found in macrolide antibiotics. Macrolides are classified as 12-, 14-, or 16-membered ring macrolides.

Polyene antibiotics have a conjugated olefinic structure in the macrocyclic lactone moiety, as is the case with amphotericin B, and sometimes lack the amino sugar moiety. Polyenes are produced mainly by *Streptomyces* and show antifungal activity. Ansamycin has an aliphatic "ansa" bridge spanning two nonadjacent positions of the aromatic system. Rifamycin is a representative ansamycin possessing a 1,4-naphthoquinone moiety. Benzenoid ansamycins having a hydroquinone moiety are also known.

In 1950, rhodomycin was isolated from the culture broth of *Streptomyces purpurascens* by BROCKMANN and his co-workers. Structural analysis disclosed that rhodomycin is a glycoside, combining an amino sugar and a 7,8,9,10-tetrahydro-5,12-naphthacenequinone moiety. Structurally similar antibiotics have since been discovered, and the generic name "anthracycline" has been assigned. The aglycone of anthracycline is called anthracyclinone. Daunorubicin and doxorubicin are representative anthracycline antitumor antibiotics. Aclacinomycin, recently isolated by UMEZAWA, possesses three sugars and is of interest because of its low toxicity.

The first nucleoside antibiotic, cordycepin, was isolated in 1950. Various nucleoside an-

tibiotics of unusual structure have subsequently been isolated and are important as antibacterials, antineoplastics, and agricultural chemicals.

In addition to the antibiotics previously described, various unclassified antibiotics have been isolated. Their structures are not always fully known, although they exhibit significant biological activity. The most important antibiotics and derivatives are discussed in this article. Some representative compounds that have recently been introduced in clinics are also described. These include the new semisynthetic penicillins and the cephalosporins. Clearly the field of antibiotics is experiencing dynamic growth.

1.3. Nomenclature

In principle, it is the privilege of the discoverer to name to his or her new antibiotic. Usually antibiotics are named after the producing organisms or some aspect of their chemical and biological nature. In accordance with the suggestions of the Nomenclature Committee of the American Society of Microbiology [31], names of antibiotics should be based on (a) the family to which the antibiotic belongs, (b) the chemical structure of the compound, or (c) some property of the antibiotic. If, for some reason, a name cannot be given to the new antibiotic, a code designation may be given.

2. Chemotherapeutic Use of Antibiotics

2.1. Microbial Pathogens

Microorganisms that can be treated by chemotherapy include bacteria, fungi (→ Antimycotics), viruses, rickettsia, and protozoa (→ Chemotherapeutics).

Outside the cytoplasmic membrane, bacteria have a rigid shell called the cell wall that is not seen in mammalian cells. The main constituent of the cell wall is peptidoglycan, a cross-linked structure of long parallel chains of polysaccharides and short peptide chains. Bacteria are separated into two classes based on the results of a Gram's stain. Gram-positive bacteria hold the color of the primary stain and gram-negative ones are decolorized and are stained by the counterstain. The structure and constituents of the cell walls of gram-positive and gram-negative bacteria are slightly, but distinctly, different.

Some antibiotics are effective only against gram-positive bacteria whereas others are active against gram-negative ones. Streptomycin and kanamycin are effective against mycobacteria as well as gram-positive and gram-negative bacteria. Chloramphenicol and tetracyclines are broad-spectrum antibiotics active against not only the usual bacteria but also rickettsia.

2.2. Tumor Cells

Certain antibiotics are effective for clinical treatment of cancer. Cancer cells function abnormally, escaping growth regulation. Antibiotics occupy an important position among the various agents for cancer chemotherapy. At present, doxorubicin, daunorubicin, mitomycin C, bleomycin, actinomycin D, chromomycin A_3, and neocarzinostatin are used clinically. They interact with DNA to inhibit polymerases of DNA and RNA, or to cause DNA strand breakage. Normal cells are also damaged to some extent; the selective toxicity is generally based on the unusually rapid multiplication of the tumor cells. Some of these antibiotics also possess antineoplastic activity because they inhibit the synthesis of DNA.

2.3. Enzyme Inhibitors

The action of antibiotics can be interpreted as a direct or indirect inhibition of certain enzyme systems. The activity of an antibiotic can be recognized only when its site of action is critically important to the maintenance of the life of the cell. However, various enzyme inhibitors of microbial origin have been found useful as medicines even though they do not exhibit antibiotic activity [32]. For example, pepstatin, an inhibitor of pepsin isolated by Umezawa in 1970 from *Streptomyces testaceus,* shows promise for the treatment of gastric ulcer, but its therapeutic effects are not based on any antibiotic action.

2.4. Chemotherapeutic Uses

Chemical and bacteriological diagnoses are especially important for successful chemotherapy because the choice of drug depends primarily on the sensitivity of the microorganism to the drug. The antimicrobial activity of an antibiotic is expressed as the minimum inhibitory concentration (MIC) measured by the dilution method. Antibiotics are administered by hypodermic, intramuscular, or intravenous injections, or as internal medicines. For external applications, they are given according to the nature of the antibiotic and the characteristics of the disease. Doses large enough to maintain a sufficient drug concentration in the blood and tissues are prescribed. Various undesirable side effects of antibiotics have been reported. One of the serious side effects is the allergic reaction to penicillins. Oto- and nephrotoxicity result from the long-term use of aminoglycosides in quantity. Aplastic anemia caused by chloramphenicol also has been reported. Antitumor and antiviral antibiotics are generally highly toxic.

2.5. Use in Agriculture

Although antibiotics were originally developed for use against microbial diseases in humans, they are also applicable to agriculture. Several antibiotics are used in the treatment of animal and plant diseases. Kasugamycin and blastcidin S are effective against rice blast disease. Polyoxins are selectively effective against certain species of phytopathogenic fungi. Penicillins, tetracyclines, mikamycins, erythromycins, tylosins, spiramycins, and thiopeptins are used in animal feeds to stimulate growth. The mechanism of growth enhancement induced by these antibacterial antibiotics is unknown. The use of antibiotics as an additive in animal fodder has led to considerable improvement in agricultural production. Antibiotics, on the one hand, prevent breeding diseases, afflictions, and mortality in young animals. On the other hand, antibiotics, even in very small quantities (on the order of 5 – 20 mg per kilogram of fodder mixture), enhance the growth and full utilization of fodder in, e.g., pigs and poultry. For all antibiotics that are used simultaneously as therapeutic agents and fodder additives, there is the danger of developing a resistance. Therefore, worldwide efforts are directed at avoiding the use of the therapeutically important antibiotics, particularly the penicillins and the tetracyclines, as feed additives. Instead, these antibiotics are used only for genuine veterinary treatment in compliance with certain controls and quarantine guarantees.

2.6. Resistance

An organism becomes resistant to an antibiotic if it survives continued contact. Antibiotics repress the growth of the sensitive organism in a culture, resulting in the survival of naturally resistant organisms. Microbial resistance can be acquired through a spontaneous or induced mutation. Various examples of cross-resistance have been observed and several cross-resistant groups of antibiotics are recognized. Combined use of two antibiotics retards the appearance of resistant organisms.

2.7. Units

The production, isolation, and processing of commercial products require careful control for all pharmaceuticals. Because of the extremely high sensitivity and the danger of diminished activity, this consideration has always been particularly important.

The Oxford unit (O. U.) is defined as the amount of penicillin that just prevents the growth of a certain *Staphylococcus aureus* species. Very pure crystalline penicillin salts generally have constant biological activities and the Oxford unit has been replaced by the international unit: 1 mg of pure benzylpenicillin sodium contains 1670 O. U.; the O. U. specific to this salt was declared to be the international unit (I. U., usually abbreviated U). Conversely, 0.6 μg of benzylpenicillin sodium has the activity of 1 I. U. Because the biological activity comes from the penicillin nucleus, the change to another cation leads to a change in activity proportional to the molecular mass. This change can be calculated. The activities of the chief penicillin salts are:

benzylpenicillin sodium	1670 U/mg
benzylpenicillin potassium	1598 U/mg
benzylpenicillin procaine	1011 U/mg
penicillin-2-hydroxyprocaine	1008 U/mg
penicillin-*N,N′*-dibenzyl-ethylenediamine	1213 U/mg
penicillin-*N*-ethylpiperidine	1328 U/mg.

The mass of 1 U, for benzylpenicillin sodium, 0.6 µg, is extremely small. The following larger units of mass are used in production and trade:

1 Mega U	$= 1 \times 10^6$ I. U.
	= 600 mg benzylpenicillin sodium
	= ca. 1 g benzylpenicillin procaine
1 Mio Mega U	$= 1 \times 10^{12}$ I. U.
	= 600 kg benzylpenicillin sodium
	= ca. 1 t benzylpenicillin procaine

The activities of some older penicillins are given in I. U.

phenoxymethylpenicillin:	1 mg free acid	1699 U
phenethicillin:	1 mg D-potassium salt	1476 U
	1 mg L-potassium salt	1470 U
penicillin O:	1 mg potassium salt	1612 U

All other penicillins that are used therapeutically can be made very pure and the preparations are dosed and traded in mass units (µg, mg, g, kg).

2.8. Analysis

The practical determination of active substances in penicillins and other antibiotics can be divided among three types of methods [25]:

1) Microbiological testing (see Chap. 7).
2) Determination of the contents by chemical or enzymatic conversion followed by a physical method, such as colorimetry.
3) Purely physical methods, such as UV or IR absorption.

3. Classification of Antibiotics

3.1. β-Lactams

The β-lactam group includes natural penicillins, semisynthetic penicillins, natural cephalosporins, semisyntheric cephalosporins, cephamycins, 1-oxacephems, clavulanic acids, penems, carbapenems, nocardicins, and monobactams.

3.1.1. Natural Penicillins

Penicillin was discovered in 1929 by Fleming [22]. At first it was obtained as a mixture of several similar compounds, but these were later separated from each other. The β-lactam structure of penicillin was proposed by Abraham and Chain and supported by Woodward, but it was opposed by those who believed in the alternative thiazolidine-oxazole structure [33]. The β-lactam structure was finally established by an X-ray crystallographic analysis performed by Hodgkin and Low [34]. Penicillins G, F, K, X, and N, dihydropenicillin F, and isopenicillin N have been isolated from the fermentation broths of *Penicillium notatum* or *P. chrysogenum*. These compounds differ only in the R moiety of structure **1**.

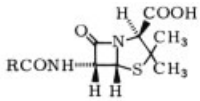

1

Name	R
Penicillin G (Benzylpenicillin) [61-33-6]	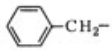
Penicillin F [118-53-6]	$CH_3CH_2CH{=}CHCH_2-$
Dihydropenicillin F [4493-18-9]	$CH_3(CH_2)_4-$
Penicillin K [525-97-3]	$CH_3(CH_2)_6-$
Penicillin X [525-91-7]	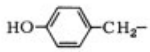
Penicillin N [525-94-0]	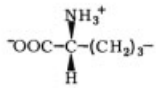
Isopenicillin [58678-43-6]	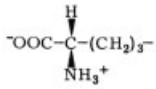

Of these, penicillin G shows good stability, activity, and rate of production by microorganisms. Total synthesis of penicillin V was achieved by Sheehan and Henery-Logan in 1957 [35]. Biogenic syntheses of penicillin – cephalosporin antibiotics also have been reported [36], [37].

3.1.2. Semisynthetic Penicillins

Several limitations have become apparent concerning the antibiotic activity of benzylpenicillin. This drug is not very active against gram-negative bacteria; it is inactivated by penicillinase produced by resistant organisms, and it is not suitable for oral administration because it breaks down under acidic conditions. Penicillins having different side chains have been made by adding appropriate precursors to the fermentation [33]. Various penicillins have been obtained biosynthetically. Among these is phenoxy-methylpenicillin, the first used by oral administration. In contrast, 6-aminopenicillanic acid (6-APA, **2**, R is NH_2) can be prepared by either enzymatic or chemical means. Penicillin amidase or penicillin acylase cleaves the side chain of penicillin to produce 6-APA. The amide bond of the side chain is also efficiently cleaved by treatment with phosphorus pentachloride [25, p. 27]. Penicillins with modified side chains (**2**) have been synthesized from 6-APA via the acyl chloride method, the ethyl or isobutyl chloroformate method, or the dicyclohexylcarbodiimide method in order to improve the antibacterial spectra and increase the stability against penicillinase [33, p. 59]. Ampicillin is active against gram-negative bacteria, and carbenicillin and sulbenicillin are effective against *Pseudomonas.* Ampicillin and amoxicillin are suitable for oral administration. Methicillin, oxacillin, cloxacillin, dicloxacillin, flucloxacillin, and naficillin are resistant to β-lactamase. Mecillinam has an unusual amidino side chain and is relatively stable and effective against gram-negative bacteria.

Modification of the carboxyl group has been found to be effective for the purpose of oral administration, and penicillin esters (**3**) have been developed [33, p. 59]. These are absorbed and hydrolyzed by the small intestine to release free acids of the parent penicillins.

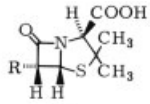

2

Name	R
Penicillin V (Phenoxymethylpenicillin) [87-08-1]	$C_6H_5OCH_2CONH-$ (biosynthetic)
Penicillin O [87-09-2]	$CH_2{=}CHCH_2SCH_2CONH-$
Phenethicillin [147-55-7]	$C_6H_5OC(CH_3)HCONH-$
Propicillin [551-27-9]	$C_6H_5OC(C_2H_5)HCONH-$
Phenbenicillin [1926-48-3]	$C_6H_5OC(C_6H_5)HCONH-$
Carbenicillin [4697-36-3]	$C_6H_5-C(H)(COOH)-CONH-$
Sulbenicillin [41744-40-5]	$C_6H_5-C(H)(SO_3H)-CONH-$
Ticarcillin [34787-01-4]	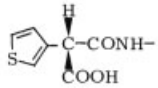
Methicillin [61-32-5]	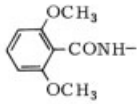
Nafcillin [147-52-4]	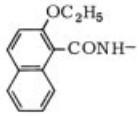
Oxacillin [66-79-5]	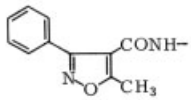
Cloxacillin [61-72-3]	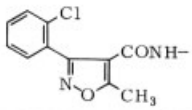

Dicloxacillin [3116-76-5]

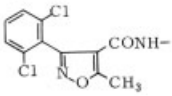

Flucloxacillin [5250-39-5]

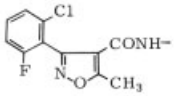

Ampicillin [69-53-4]

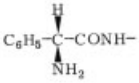

Apalcillin [63469-19-2]

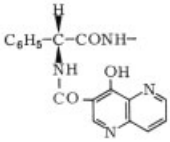

Mezlocillin [51481-65-3]

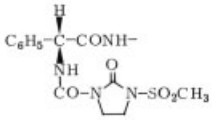

Piperacillin [61477-96-1]

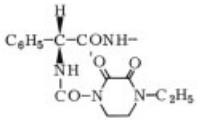

Amoxicillin [26787-78-0]

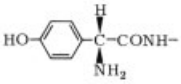

Cyclacillin [3485-14-1]

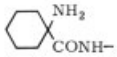

Hetacillin [3511-16-8]

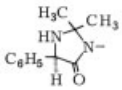

Mecillinam [32887-01-7]

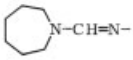

3.1.3. Natural Cephalosporins

The fermentation broth of *Cephalosporium* spp. isolated by BROTZU contained several antibiotics: cephalosporin P, penicillin N, and cephalosporin C. Cephalosporin P was shown to be an acidic steroidal substance. Cephalosporin C was active against gram-negative bacteria, resistant to β-lactamase, and much less toxic than penicillin. The chemical structure of cephalosporin C was determined by ABRAHAM and NEWTON [25]. It consists of 7-aminocephalosporanic acid (7-ACA) and D-α-aminoadipic acid (see **4**). Treatment of cephalosporin C with acetyl esterase yields deacetylcephalosporin C, which exhibits about 20 % of the antibacterial activity of cephalosporin C. The allylic acetoxy group of cephalosporin C can be hydrogenated in the presence of palladium – charcoal catalyst to yield deacetoxycephalosporin C, which shows 10 % of the activity of the parent cephalosporin C [25]. WOODWARD and his co-workers synthesized cephalosporin C in a fully stereospecific manner [38].

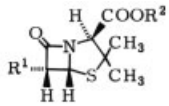

3

Name	R^1	R^2
Talampicillin [47747-56-8]		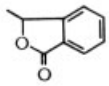
Pivampicillin [33817-20-8]	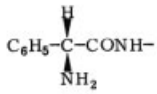	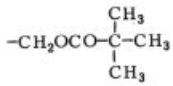
Bacampicillin [50972-17-3]	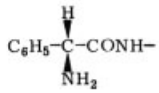	
Pivmecillinam [32886-97-8]	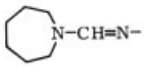	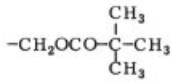

3.1.4. Semisynthetic Cephalosporins

Because the antibacterial activity of cephalosporin C itself is relatively low, the development of a more active derivative is desirable. The phosphorus pentachloride method has been applied to the cephalosporin system to produce 7-aminocephalosporanic acid (7-ACA) in high yield [25, p. 27]. The 3′-acetoxy group of cephalosporin is easily replaced by various nucleophiles [25, p. 134]. Modification

of the 7-amino group and the 3′ group make possible the various cephalosporin derivatives **4a** [33, p. 59]. Cephaloridine, cefazolin, and cefamandole are active against gram-negative bacteria. Cefuroxime, cefotaxime, and ceftizoxime have a methoxyimino group and a 2-aminothiazole ring in common and are resistant to β-lactamase. Cefoperazone is particularly active against *Pseudomonas*. All of these are used by injection. On the other hand, several cephalosporins are used only by oral administration. These include cephalexin, cephaloglycin, cefradine, cefadroxil, cefaclor, cefroxadine, and cefatrizine.

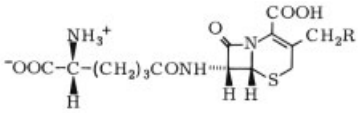

4

Name	R
Cephalosporin C [61-24-5]	$OCOCH_3$
Deacetylcephalosporin C [1476-46-6]	OH
Deacetoxycephalosporin C [26924-74-3]	H

Cephalosporins are also obtainable via the ring expansion reaction of penicillin sulfoxide first devised by MORIN [25, p. 183]. Thus, cephalexin is produced by chemical conversion of phenoxymethylpenicillin or benzylpenicillin.

3.1.5. Cephamycins

Substances similar to cephalosporin C were found among the products of various streptomycetes and were characterized by the presence of a 7α-methoxy group. They are named cephamycins after their cephem skeleton (see **5**) and their production by streptomycetes [39], [29, vol. 1, p. 199]. They are strongly resistant to β-lactamase and effective against gram-negative bacteria and bacteria that have acquired resistance to penicillins and cephalosporins. Semisynthetic cephamycins (**5a**) with improved activities are obtained by chemical transformations.

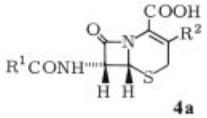

4a

Name	R^1	R^2
Cephalexin [15686-71-2]	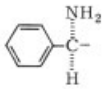	$-CH_3$
Cefaclor [53994-73-3]	"	$-Cl$
Cephaloglycin [3577-01-3]	"	$-CH_2OCOCH_3$
Cephradine [38821-53-3]	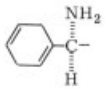	$-CH_3$
Cefroxadine [51762-05-1]	"	$-OCH_3$
Cefadroxil [50370-12-2]	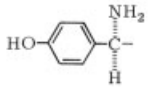	$-CH_3$
Cephapirin [21593-23-7]	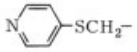	$-CH_2OCOCH_3$
Cephalothin [153-61-7]	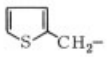	$-CH_2OCOCH_3$
Cephacetrile [10206-21-0]	$N{\equiv}CCH_2-$	$-CH_2OCOCH_3$
Cefsulodin [62587-73-9]		
Cephaloridine [50-59-9]	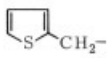	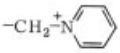
Cefatrizine [51627-14-6]	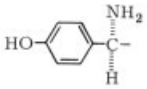	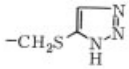

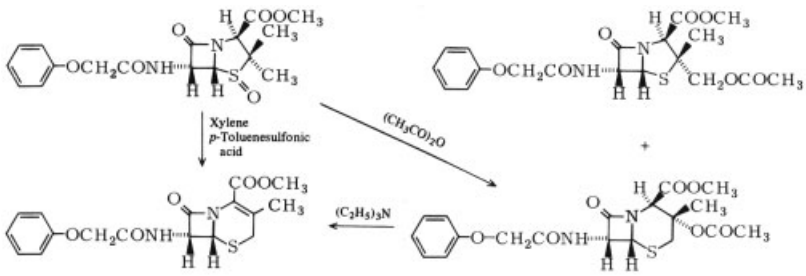

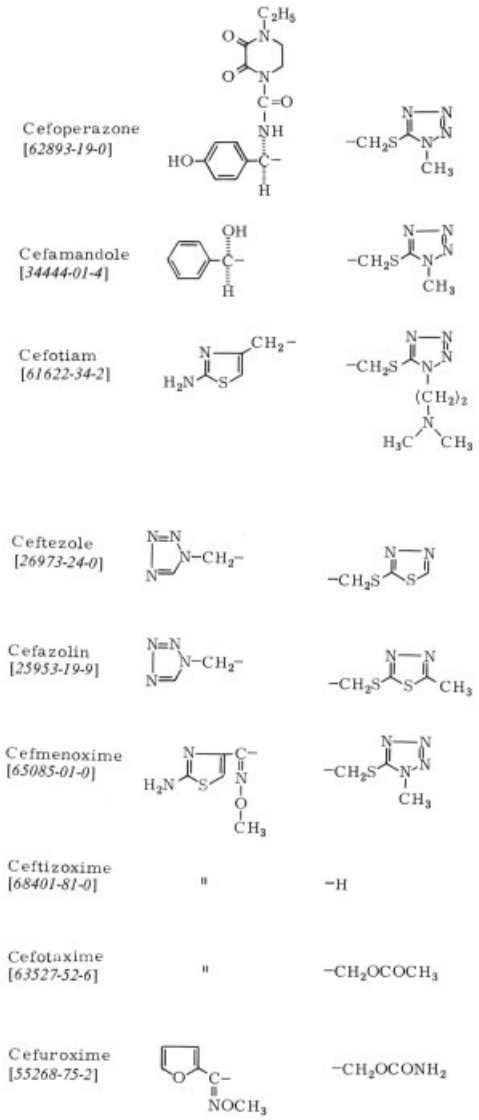

Cefoxitin, synthesized from cephamycin C, is particularly active against various gram-negative bacteria and stable to β-lactamase. Cefmetazole, produced from 7-ACA, has almost the same antibacterial spectrum as cefoxitin and maintains its high concentration in blood. Cefotetan is reported to be more active than cefoxitin against gram-negative bacteria (for structures see page 53).

Chemical modification of cephamycins requires special devices for the following reasons. 7-Aminocephamycinoic acid (7-ACMA; **6**), which corresponds to the 7-ACA of cephalosporins, is not easily isolated because of its instability.

It has methoxy and amino groups on the same carbon atom of the β-lactam ring and the elimination of the protonated amino group is quite facile, because of the electron-donating nature of the methoxy group. Moreover, the usual phosphorus pentachloride method cannot be applied to the side chain cleavage of cephamycin C because a strong N – P bond is formed by the reaction of the carbamate moiety of cephamycin C with phosphorus pentachloride [40]. Instead, exchange of the α-aminoadipoyl side chain for another acyl group is achieved by treating the fully protected cephamycin C with the appropriate acyl chloride in the presence of a neutral acid scavenger. This is followed by the simultaneous removal of the amino protective group and the α-aminoadipoyl group [41]. The side chain transformation is also effected using an acyl chloride and partially hydrated molecular sieves [42].

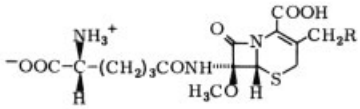

5

Name	R	Producing organism
7 α-Methoxy-cephalosporin C [32178-82-8]	$-OCOCH_3$	*Streptomyces lipmannii* *St. lactamdurans*
Cephamycin C [34279-51-1]	$-OCONH_2$	*St. clavuligerus* *St. lactamdurans* *St. jumonjiensis*
Cephamycin A [34279-78-2]	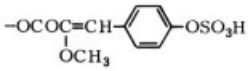	*St. griseus* *St. chartreusis* *St. cinnamonensis* *St. fimdriatus*
Cephamycin B [34279-77-1]	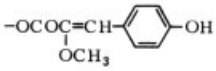	*St. halstedii* *St. rochei* *St. viridochromogenes*
C-2801X [62851-50-7]	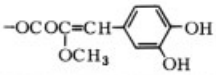	*St. heteromorphous* *St. panagensis*

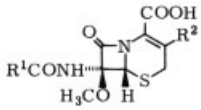

5a

Name	R^1	R^2
Cefoxitin [35607-66-0]	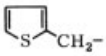	$-CH_2OCONH_2$
Cefmetazole [56796-20-4]	$N\equiv CCH_2SCH_2-$	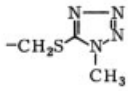
Cefotetan [69712-56-7]	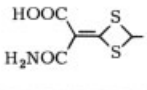	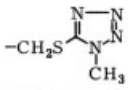

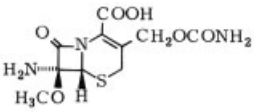

6

7-Aminocephamycinoic acid (7-ACMA) [62041-14-9]

Chemical conversion of cephamycin into 7-ACMA ester has been reported [43].

3.1.6. 1-Oxacephems

In addition to the modification of the side chains of natural β-lactam antibiotics, totally or partially synthetic nuclear analogs of penicillins

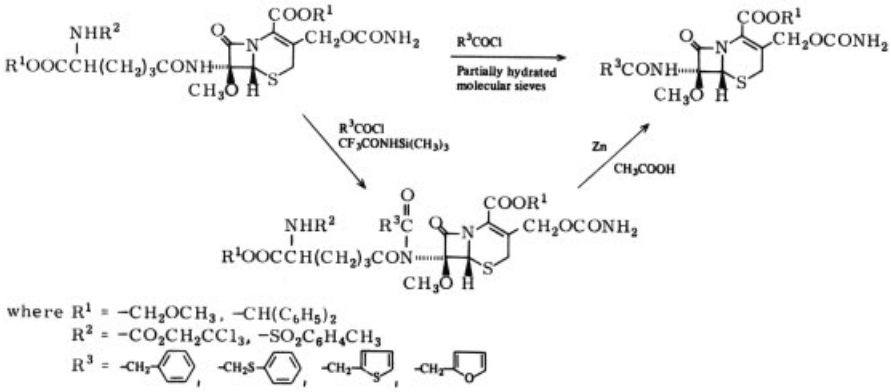

and cephalosporins have been explored extensively [28], [33, p. 59]. In 1974 WOLFE reported the first 1-oxacephem (**7**) derived from penicillin, but its antibacterial activity remains unknown because the amino and carboxy protective groups have not been removed. Racemic 1-oxacephalothin (**8**), synthesized by CHRISTENSEN and his co-workers in the same year, was found to be antibacterially active, suggesting that the sulfur atom is not always necessary for the expression of antibiotic activity. Racemic 1-oxacefamandole is twice as active as cefamandole and the activity of optically active 1-oxacephalothin (**8**) is four to eight times as high as that of cephalothin. NAGATA and his co-workers discovered latamoxef (**9**) (moxalactam, 6059-S), which exhibits strong activity against pathogenic anaerobes, such as *Bacteroides fragilis,* as well as gram-negative bacteria, including *Pseudomonas* [29, vol. 2, p. 1]. It is completely stable against various β-lactamases and has low toxicity. A high plasma-peak level and long duration are maintained. Latamoxef is a nuclear analog of cephamycin that has a 2-(4-hydroxyphenyl)malonylamino side chain and a 1-methyltetrazolylthio moiety; it is produced on an industrial scale by a totally chemical process starting with *epi*-penicillin *S*-oxide [44], [45].

3.1.7. Clavulanic Acids

In the course of screening the substances inhibiting β-lactamase, which is responsible for bacterial resistance to penicillins and cephalosporins, a potent β-lactamase inhibitor, clavulanic acid, was isolated from *Streptomyces clavuligerus* [46]. The antibiotic activity of clavulanic acid is not strong, but it has a broad antibacterial spectrum. Instead, it is effective synergistically when used with β-lactamase-sensitive penicillins and cephalosporins against β-lactamase-producing organisms. Clavulanic acid is characterized by its 1-oxadethiapenam ring system and lack of the side chain at position 6. Several congeners of clavulanic acid have been isolated [27], [29, vol. 2, p. 361], [47]. Clavam-2-carboxylic acid, 2-hydroxymethylclavam, and 2-formyloxymethylclavam exhibit antifungal activity [48].

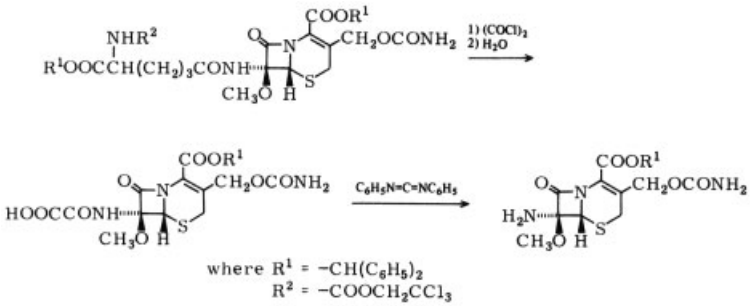

Name	Structure
Methyl (6*S*,7*S*)-3-methyl-8-oxo-7-phthalimido-5-oxa-1-azabicyclo[4.2.0]-oct-2-ene-2-carboxylate [*54997-17-0*]	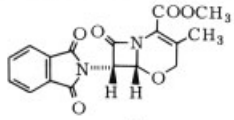 **7**
1-Oxacephalothin [*54214-83-4*]	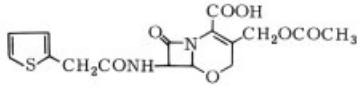 **8**
Latamoxef (Moxalactam, 6059-S) [*64952-97-2*]	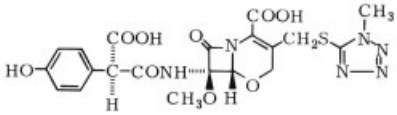 **9**

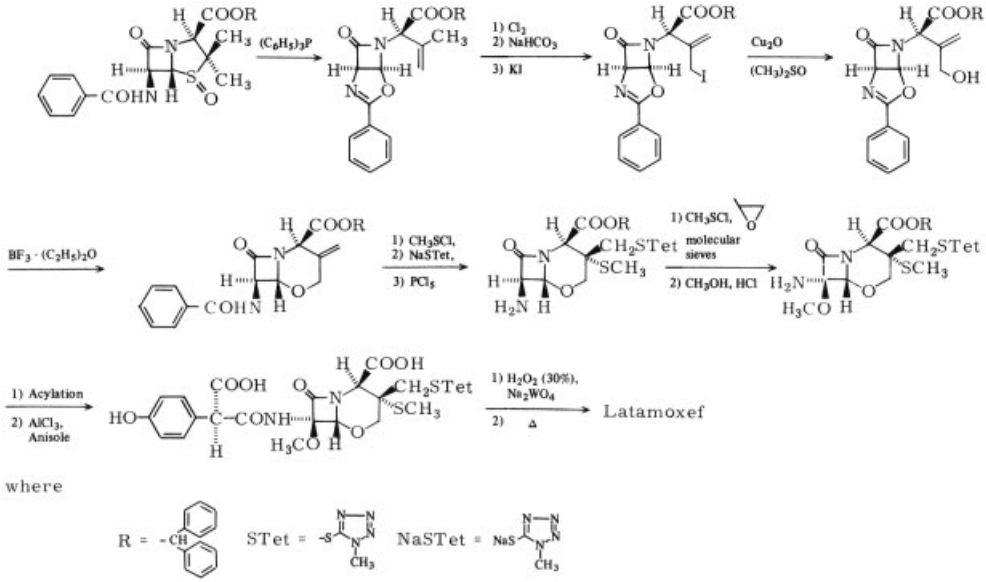

3.1.8. Penems

The penem ring system has not been found in nature; it has been designed artificially by WOODWARD [26, p. 167], [28], [29, vol. 2, p. 315]. That the antibacterial activity of β-lactam antibiotics is based on their ability to acylate enzymes is widely accepted. In penicillins, the rigid, nonplanar bicyclic system enhances the reactivity of the β-lactam ring by diminishing the delocalization of the unshared electron pair of the amide nitrogen onto the adjacent carbonyl group. On the other hand, in cephalosporins, where the β-lactam nitrogen is bonded almost planar, the double bond of the six-membered ring interacts with the unshared electrons of the β-lactam nitrogen, diminishing the delocalization to the amide carbonyl. Therefore, the β-lactam ring of cephalosporins is cleaved easily. Penems combine the two structural elements, the five-membered ring and the double bond. 6-Acylaminopenem-3-carboxylic acids (**10**), 6-

unsubstituted penem-3-carboxylic acids (**11**), and 6-alkylpenem-3-carboxylic acids (**12**) have been synthesized. The β-lactam moiety with two asymmetric centers derives from penicillin, and the five-membered ring fused with it is formed by the intramolecular Wittig reaction [29, vol. 2, p. 315]. Various 2-substituents, such as H, CH_3, C_2H_5, $CH_2C_6H_5$, and $SCH_2CH_2NHCOCH_3$, have been introduced. Although the activity of compound **10** with $R = CH_3$ was disappointing, presumably because of its low stability, 6-unsubstituted penems (**11**) exhibit powerful antibiotic activity [49]. 6-Monoalkylpenems (**12**) show interesting activity in general. The penems having bulky substituents at C-6 are biologically inactive because of the low reactivity of their β-lactam rings.

Name	Structure
Clavulanic acid [*58001-44-8*]	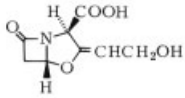
ß-Hydroxypropionyl clavulanic acid [*64675-12-3*]	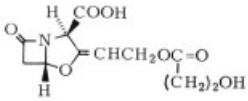
Clavam-2-carboxylic acid [*71657-61-9*]	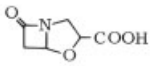
2-Hydroxymethyl-clavam [*66036-39-3*]	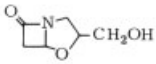
2-Formyloxymethyl-clavam [*66036-40-6*]	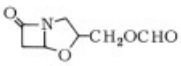

3.1.9. Carbapenems

Cabapenems are a family of antibiotics having the 1-azabicyclo[3.2.0]hept-2-ene system [27], [29, vol. 2, p. 227]. The first carbapenem antibiotic, thienamycin, was discovered at Merck in 1976 among the fermentation products of *Streptomyces cattleya* [50]. Antibiotics of this type have been isolated one after another in the search for inhibitors of bacterial cell wall synthesis and β-lactamase. From the fermentation broth of *Streptomyces olivaceus,* the Beecham group isolated olivanic acids MM4550, MM13902, MM17880, MM22380, MM22381, MM22382, and MM22383 [51–54]. Epithienamycin A, B, C, D, E, and F were found by the Merck group [55–58]. Some of olivanic acids and epithienamycins are identical. Olivanic acid MM4550 is identical to MC696-SY2-A found by UMEZAWA as a product of *Streptomyces fulvoviridis* [59]. The antibiotics designated PS-5, −6, and −7 were isolated by Sanraku-Ocean in collaboration with Panlabs from *Streptomyces cremeus,* subsp. *auratilis* A271 [60–63]. Carpetimycin A and B, reported by the Kowa Company, are products of *Streptomyces* spp. [64], [65]. Asparenomycin A was isolated by the Shionogi research group from *Streptomyces tokunonensis* and *Streptomyces argenteolus* [66].

Name	Structure
6-Acylaminopenem-3-carboxylic acid	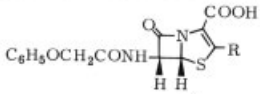 **10**
Penem-3-carboxylic acid	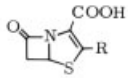 **11**
6-Alkylpenem-3-carboxylic acid	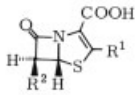 **12**

Carbapenems are classified into three classes according to the mode of substitution on the β-lactam ring, that is, *trans*-carbapenems (thienamycin, epithienamycins C and D, PS-5, F-6, and F-7), *cis*-carbapenems (epithienamycin A, B, E, and F, MC696-SY2-A, carpetimycin A and B), and ene-carbapenems (asparenomycin A, B, and C). The 5*R* configuration seems significant for biological activity. The instability of the carbapenems and low broth titer cause difficulties in the determination of the structure.

Structure	Name	R^1	R^2
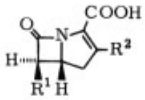	Thienamycin [59995-64-1]		$-SCH_2CH_2NH_2$
	Epithienamycin C Olivanic acid MM 22381 [63599-16-6]		$-SCH_2CH_2NHCOCH_3$
	Epithienamycin D Olivanic acid MM 22383 [65322-98-7]		$-SCH{=}CHNHCOCH_3$
	PS-5 [78856-77-6]	CH_3CH_2-	$-SCH_2CH_2NHCOCH_3$
	PS-6 [72615-19-1]	$(CH_3)_2CH-$	$-SCH_2CH_2NHCOCH_3$
	PS-7 [72615-18-0]	CH_3CH_2-	$-SCH{=}CHNHCOCH_3$
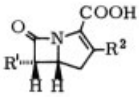	Epithienamycin A Olivanic acid MM 22380 [63582-78-5]		$-SCH_2CH_2NHCOCH_3$
	Epithienamycin B Olivanic acid MM 22382 [65376-20-7]		$-SCH{=}CHNHCOCH_3$
	Epithienamycin E Olivanic acid MM 13902 [79057-46-8]	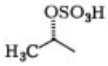	$-SCH{=}CHNHCOCH_3$
	Epithienamycin F Olivanic acid MM 17880 [79057-45-7]		$-SCH_2CH_2NHCOCH_3$
	Olivanic acid MM 4550 MC696-SY2-A [76985-32-5]		$-\overset{O}{\overset{\|\|}{S}}CH{=}CHNHCOCH_3$
	Carpetimycin A [76025-73-5]	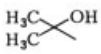	$-\overset{O}{\overset{\|\|}{S}}CH{=}CHNHCOCH_3$
	Carpetimycin B [76094-36-5]	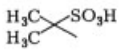	$-\overset{O}{\overset{\|\|}{S}}CH{=}CHNHCOCH_3$

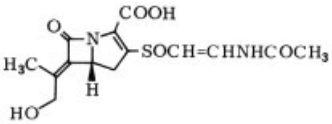

Asparenomycin
[76466-24-5]

Thienamycin is active against a wide range of gram-positive and gram-negative bacteria, including the ones resistant to conventional β-lactam antibiotics. Carpetimycins and asparenomycin are also effective against resistant bacteria.

Chemical modifications and a great deal of synthetic study of carbapenems have been un-

dertaken to improve the stability of the carbapenem skeleton and compensate for the low productivity of the microbes. This is one of the most important fields of antibiotics [67–80], [69, p. 1142].

Naturally occurring carbapenems have several functional groups that have been subjected to chemical modifications to improve their stability and antibacterial potency [54]. The aminoethylthio side chain plays an important role in extending the antibiotic activity, especially the antipseudomonal activity, and is also thought to be a cause of the instability of carbapenems, presumably by intramolecular aminolysis of the β-lactam ring. Derivatives of the aminoethylthio group, carboxyl group, and hydroxyethyl side chain are being sought.

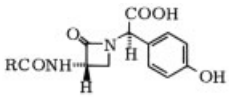

Name	R
Nocardicin A [39391-39-4]	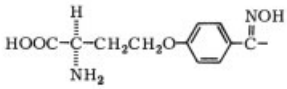
Nocardicin B [60134-71-6]	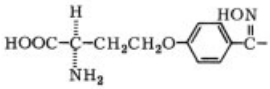
Nocardicin C [59511-12-5]	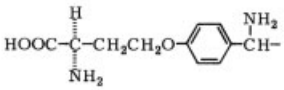
Nocardicin D [61425-17-0]	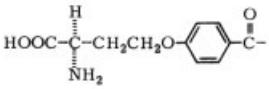
Nocardicin E [63555-59-9]	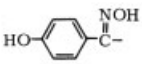
Nocardicin F [63598-46-9]	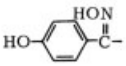
Nocardicin G [65309-11-7]	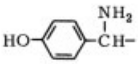

3.1.10. Nocardicins

A mutant strain of *Escherichia coli* showing specific supersensitivity to β-lactam antibiotics has been developed at the Fujisawa Research Laboratories and used to isolate nocardicins from *Nocardia uniformis* by a screening procedure [81]. The nocardicin structure has been elucidated by spectroscopic analysis and chemical degradation. The noncardicins are monocyclic β-lactam antibiotics [27], [29, vol. 2, p. 165], [47, p. 281]. Several congeners differing in the side chains have been isolated. Among them, significant activity is exhibited only by nocardicin A, which is active against gram-negative bacteria, especially *Pseudomonas aeruginosa, Proteus,* and *Neisseria,* but inactive against gram-positive bacteria.

3.1.11. Monobactams

Sulfazecin was isolated in 1981 by the Takeda group as a product of *Pseudomonas acidophila* by screening using organisms highly sensitive to β-lactams [82]. The structure was shown to be a monocyclic β-lactam. **Isosulfazecin**, a diastereomer of sulfazecin, also was isolated by the same group. In the same year, the Squibb group reported on a group of monocyclic β-lactams produced by *Agrobacterium, Chromobacterium,* and *Gluconobacter* [83]. A compound (SQ 26 445) identical to sulfazecin was included. Sykes proposed the name "monobactam" for compounds characterized by the 3-acylamino-2-oxoazetidine-1-sulfonic acid group. In monobactams, the β-lactam ring presumably is activated by the electronic effect of the sulfonate moiety alone, in contrast to the case of penicillins and cephalosporins. Because the antibacterial activity of sulfazecin is not satisfactory, many derivatives have been synthesized chemically [29, vol. 3, p. 339]. Among them **aztreonam** (SQ 26 776), synthesized from threonine, has been found highly effective [84].

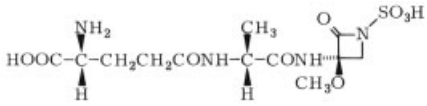

Sulfazecin [77912-79-9]

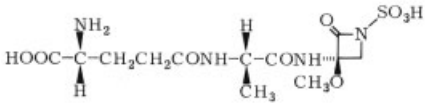

Isosulfazecin [77900-75-5]

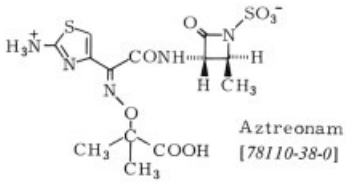

Aztreonam [78110-38-0]

3.2. Tetracyclines

The discovery of the tetracyclines, the first being aureomycin (7-chlorotetracycline), was one of the great successes of the worldwide screenings, i.e., testing of media samples and other materials, for the presence of antibiotic-producing microorganisms.

The first patents and publications of Lederle Laboratories [85], marked the beginning of an extensive stream of publications and patents that reflect the medical, industrial, and economic importance of the tetracyclines [86], [87].

3.2.1. Structure and Properties

The linear four-ring-system skeleton is characteristic of the tetracyclines (**13**) and has given the whole group its name. The strongly conjugated system of keto and enol groups is of particular significance for the biological activity. The structures of the first tetracyclines were elucidated and proved by synthetic work, e.g., that of MUXFELDT et al., shortly after their discovery and parallel to their clinical testing and industrial development. The chief tetracyclines are listed under structure **13** (see next page).

The tetracyclines are bright yellow compounds, amphoteric, and with the exception of rolitetracycline and similarly constructed derivates insoluble in water at the isoelectric point. Their salts, e.g., hydrochlorides, are soluble in water and can be administered either parenterally or orally, although the low pH of the solution causes some problems in the latter instance.

3.2.2. Anhydrotetracyclines

Tetracyclines are aromatized in ring C by dehydration with concentrated acids, e.g., aqueous hydrochloric acid or anhydrous hydrogen chloride in acetone, to form anhydrotetracyclines (**14**, wherein $R^1 - R^5$ are similar to those of **13**) [88–90]. These compounds have less biological activity than the starting compounds. Their formation, like that of the *epi*-anhydrotetracyclines, must be avoided because they are toxic to the kidneys. An interesting fact is that antibiotics with the anhydrotetracycline structure also are formed in nature and can be isolated from cultures of microorganisms, e.g., chelocardin from *Nocardia sulfurea* [91].

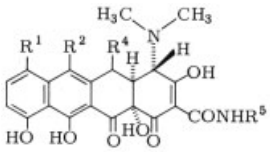

14

3.2.3. Anthracyclines

Structurally these antibiotics belong to the tetracyclines. They are characterized by the *p*-quinone structure of ring C in addition to the aromatic nuclei B and D (see **15**). Here $R^1 - R^5$ are simple substituents, such as H, OH, or CH_3. Only R^6 is a sugar or similarly complex group. Although anthracyclines show antibacterial activity, they have not been used as antibiotics because of their relatively high toxicity and strong side effects. The antitumor activity of rhodomycin was discovered by ARCAMONE et al. in 1961 [92], and various antitumor anthracyclines were subsequently isolated [93]. The most important anthracyclines are listed under structure **15**. Daunorubicin and doxorubicin are representative anthracyclines. Aclarubicin was found by UMEZAWA during a search for anthracyclines that might have lower cardiac toxicity

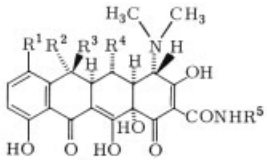

13

Name	Production	R^1	R^2	R^3	R^4	R^5
Tetracycline [6416-04-2]	*Streptomyces aureofaciens*	H	CH_3	OH	H	H
Chlortetracycline (aureomycin) [57-62-5]	*Streptomyces aureofaciens*	Cl	CH_3	OH	H	H
Demethylchlortetracycline (demeclocycline) [127-33-3]	*S. aureofaciens* *S. viridifaciens*	Cl	H	OH	H	H
Oxytetracycline [79-57-2]	*S. rimosus*	H	CH_3	OH	OH	H
Methacycline [914-00-1]	Semisynthetic	H	$=CH_2$		OH	H
Doxycycline [564-25-0]	Semisynthetic	H	CH_3	H	OH	H
Rolitetracycline [751-97-3]	Semisynthetic	H	CH_3	OH	H	CH_2N
Minocycline [10118-90-8]	Synthetic	$N(CH_3)_2$	H	H	H	H

than doxorubicin. Anthracyclines exert their effect by interacting with DNA, the primary cellular receptor [94].

3.3. Aminoglycosides

WAKSMAN initiated the screening of antibiotics and, after finding actinomycin and streptothricin, he discovered the first useful aminoglycoside, streptomycin, in 1944 [95–99]. After wide use of penicillin, streptomycin, chloramphenicol, and tetracycline, resistant organisms appeared in hospital patients. In 1957, staphylococci and gram-negative organisms resistant to all the known antibiotic drugs caused serious infections; kanamycin was discovered at that time by UMEZAWA and was introduced clinically. However, in 1965, kanamycin-resistant strains appeared. In 1967, the enzymatic mechanism of resistance to aminoglycoside antibiotics was elucidated. UMEZAWA suggested that 3′-phosphotransferase and 6′-acetyltransferase, which transferred the terminal phosphate of ATP to the 3′-hydroxyl group of kanamycin, neomycin, and paromomycin or the acetyl group of acetyl-CoA to the 6′-amino group, were involved in the mechanism of resistance [100], [101]. In order to prove this enzymatic mechanism of resistance conclusively, 3′-deoxykanamycin A and 3′,4′-dideoxykanamycin B were synthesized and used to demonstrate the inhibition of the growth of resistant strains [102]. This conclusively demonstrated the enzymatic mechanism of resistance. Effective derivatives were obtained not only by deoxygenation but also by modification of the 1-amino group, which was involved in binding to the enzymes.

More than 150 naturally occurring aminoglycosides have been isolated from culture filtrates of *Streptomyces, Streptoverticillium, Nocardia, Micromonospora, Streptoalloteichus, Dactylosporangium, Saccharopolyspora,* and other bacterial strains.

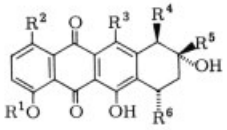

15

Name	R^1	R^2	R^3	R^4	R^5	R^6
Daunorubicin [*20830-81-3*]	CH_3	H	OH	H	$COCH_3$	A
Doxorubicin [*23214-92-8*]	CH_3	H	OH	H	$COCH_2OH$	A
Carminomycin I [*50935-04-1*]	H	H	OH	H	$COCH_3$	A
Baumycin A_2 [*64253-71-0*]	CH_3	H	OH	H	$COCH_3$	B ($R = CH_2OH$)
Baumycin B_2 [*64312-53-4*]	CH_3	H	OH	H	$COCH_3$	B ($R = COOH$)
Rhodomycin A [*1404-50-8*]	H	H	OH	C	CH_2CH_3	C
Rhodomycin B [*1404-52-0*]	H	H	OH	OH	CH_2CH_3	C
Aklavin [*60504-57-6*]	H	H	H	$COOCH_3$	CH_2CH_3	C
Cinerubin [*34044-10-5*]	H	OH	H	$COOCH_3$	CH_2CH_3	D
Aclarubicin [*57576-44-0*]	H	H	H	$COOCH_3$	CH_2CH_3	D

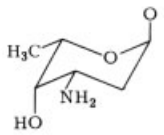

A

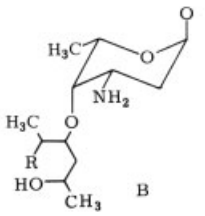

B

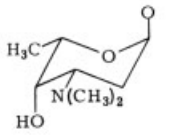

C

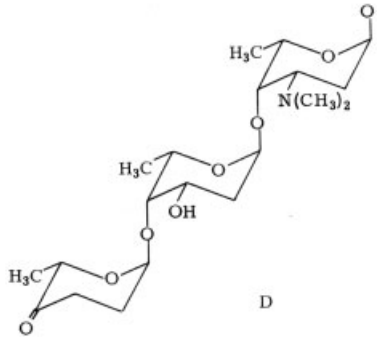

D

Naturally occurring aminoglycoside antibiotics

1) **Noncyclitol aminoglycosides**
   a) *Monosaccharide derivatives*
   3-amino-3-deoxy-D-glucose, nojirimycin, *N*-carbamoyl-D-glucosamine, streptozotocin, prumycin
   b) *Disaccharides* (trehalosamines)
   trehalosamine, mannosyl glucosaminide, 4-amino-4-deoxytrehalose
   c) *Diaminosorbitol aminoglycosides* (sorbistins)
   sorbistins A_1, A_2, B, D
   d) *Glycocinnamoylspermidines*
   LL-BM123β, γ_1, γ_2, glysperins A, B, C
2) **Aminoglycosides containing neutral cyclitols and monoaminocyclitols**
   kasugamycin, myomycins A, B, C, LL-BM782α_1, α_{1a}, α_2, minosaminomycin, LL-BM123α, validamycins A, B, C, D, E, F
3) **Aminoglycosides containing streptamine and related aminocyclitols**
   a) *Streptidine aminoglycosides* (streptomycins)
   streptomycin, mannosidostreptomycin, dihydrostreptomycin, hydroxystreptomycin (reticulin), *N*-demethylstreptomycin, mannosidohydroxystreptomycin, glebomycin (bluensomycin)
   b) *Actinamine aminoglycosides* (spectinomycins)
   spectinomycin (actinospectacin), dihydrospectinomycin
   c) *4-Substituted deoxystreptamine aminoglycosides* (neamines)
   neamine, paromamine, nebramine (nebramycin 8), lividamine, NK-1003, seldomycin factor 2, gentamines C_1, $C_{1\alpha}$, C_2, apramycin (nebramycin 2), oxyapramycin (nebramycin 7)
   d) *5-Substituted deoxystreptamine aminoglycosides* (destomycins)
   hygromycin B, destomycins A, B, C, A-396−1, SS-56-C, A16316-C
   e) *4,5-Disubstituted deoxystreptamine aminoglycosides* (neomycins)
   neomycins B, C, LP-B, LP-C, paromomycins I, II, lividomycins A, B, mannosyl paromomycin, ribostamycin, xylostasin, ribosyl paromamine (LL-BM408α), butirosins A, B, BU-1709E_1, E_2, BU-1975C_1, C_2
   f) *4,6-Disubstituted deoxystreptamine aminoglycosides* (kanamycins)
   kanamycins A, B, C, NK-1001, NK-1012−1, NK-1013−1, NK-1013−2, tobramycin (nebramycin 6), 6″-*O*-carbamoylkanamycin B (nebramycin 4), 6″-*O*-carbamoyltobramycin (nebramycin 5′), 2′-*N*-carbamoyltobramycin (nebramycin 11), 3″-deamino-3″-hydroxytobramycin (nebramycin 12), 6′-*N*-carbamoyltobramycin (nebramycin 13), gentamicins A, A_1, A_2, A_3, A_4, B, B_1, C_1, C_{1a}, C_2, C_{2a}, C_{2b} (sagamicin), X_2, JI-20A, JI-20B, G-418, I-1, 6′-*C*-methylgentamicin A (II-2), 6′-*C*-methylgentamicin A (III-1), VII-1, VII-3, VII-5, seldomycin factors 1, 3, 5, sisomicin, verdamicin, G-52, 66−40B, 66−40C, 66−40D, 3″-*N*-demethylsisomicin (66−40G)
4) **Aminoglycosides containing 1,4-diaminocyclitols**
   a) *Fortamine aminoglycosides* (fortimicins)
   fortimicins A, C, D, KG_3, 3-*O*-demethylfortimicin A, sporaricin A, 2″-*N*-carbamoylsporaricin A, 2″-*N*-formylsporaricin A, istamycins A (sannamycin A), A_1, A_2, B, B_1, C, C_1, dactimicin (SF-2052), 2″-*N*-formylfortimicin A
   b) *Non-glycine fortamine aminoglycosides*
   fortimicins B, E (AE, KH), AH, AI, AK (KI), AL, AM, AO, AP, AQ, AS, KE, KF, KG, KG_1, KG_2, KO_1, KQ, sporaricin B (KA-6606 II), KA-6606 V, VI, sannamycins B (KA-7038 II, istamycin A_0), C(KA-7038 VI), KA-7038 III, IV, V, VII, istamycin B_0, C_0

Kanamycins are produced by *Streptomyces kanamyceticus,* gentamicins by *Micromonospora purpurea*, and butirosins by *Bacillus circulans.* Many compounds analogous to aminoglycoside antibiotics are produced by the same strain. For example, more than 20 compounds structurally analogous to gentamicin have been isolated from a culture filtrate of a *Micromonospora* strain [103], [104].

Most aminoglycoside antibiotics that are important for chemotherapy contain 1,3- or 1,4-diaminocyclitols named actinamine,

2-deoxystreptamine, fortamine, or streptidine. Among these naturally occurring aminoglycosides, dihydrostreptomycin, kanamycin A, kanamycin B, lividomycin A, ribostamycin, sisomicin, spectinomycin, streptomycin, tobramycin, a mixture of gentamicins C_1, C_2, and C_{1a}, a mixture of neomycins B and C, and a mixture of paromomycins I and II are commercially available as chemotherapeutic agents useful in treating infections. Hygromycin B and destomycin A are used as animal anthelmintics. Kasugamycin and validamycin A are used for the prevention of plant diseases. Among resistant bacteria of clinical origin, the most important mechanism of resistance to aminoglycoside antibiotics is the inactivation by *O*-phosphorylation, *O*-nucleotidylation, or *N*-acetylation of specific sites of the antibiotic. The gene for these enzymes is located on a plasmid. Organisms with resistance resulting from permeability barriers to drugs have been isolated, but ribosomal resistance to aminoglycosides is very rare in organisms isolated clinically. Studies of the enzymatic mechanism of resistance to aminoglycosides have been reviewed extensively [105–108].

Semisynthetic Aminoglycosides have been made. Based on the enzymatic mechanism of resistance, studies of the chemical synthesis of derivatives that inhibit the growth of resistant strains have been initiated. 3′-Deoxykanamycin A has been synthesized and used to inhibit the growth of resistant strains having aminoglycoside-3′-phosphotransferase enzymes. Dibekacin (3′,4′-dideoxykanamycin B), synthesized from kanamycin B, shows a strong activity not only against resistant staphylococci and gram-negative organisms but also against *Pseudomonas* [102]. These results prove the enzymatic mechanism of resistance.

Streptomycin, the first aminoglycoside antibiotic, was discovered by WAKSMAN. This drug is produced by *Streptomyces griseus* and extracted from the culture filtrate by adsorption on a column of Amberlite IRC-50 resin. The hydrogen chloride – calcium chloride ($3\,HCl \cdot 1/2\,CaCl_2$) complex salt of streptomycin is easily crystallized from an anhydrous methanol solution. Streptomycin is also produced by several other strains: *Streptomyces bikiniensis, Streptomyces olivaceus, Streptomyces poonensis, Streptomyces mashuensis, Streptomyces galbus, Streptomyces rameus,* and *Streptomyces erythrochromogenes* subsp. *narutoensis.*

The early structural studies have been reviewed [109]. The two anomeric configurations were found to be α-L by application of Hudson's rules of isorotation and NMR spectral analysis. The absolute structure of streptomycin has been confirmed by X-ray analysis of its oxime selenate [110]. Streptomycin has been synthesized by oxidation of dihydrostreptomycin [111].

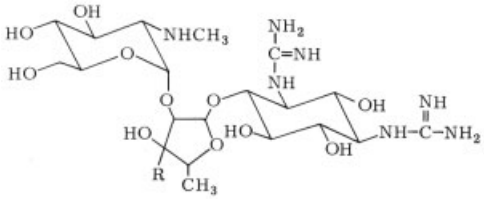

Streptomycin
[*57-92-1*]
R = CHO

Dihydrostreptomycin
[*128-46-1*]
R = CH_2OH

Spectinomycin (Actinospectacin, M-141) is produced by *Streptomyces spectabilis* and *Streptomyces flavopersicus.* It is also produced by *Streptomyces hygroscopicus* subsp. *sagamiensis.* Spectinomycin hexahydrate is crystallized from an aqueous acetone solution. This antibiotic is labile, especially in acidic solution.

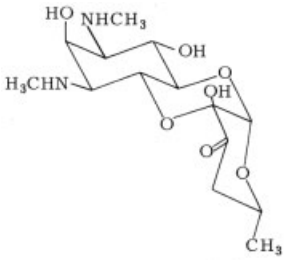

Spectinomycin (**16**)
[*1695-77-8*]

The structure of spectinomycin (**16**) was revealed by chemical studies, and its stereochemistry was determined by X-ray analysis of its dihydrobromide pentahydrate crystal [112]. To-

tal synthesis of spectinomycin has been accomplished [113], [114].

Neomycin (Fradiomycin), a mixture of **Neomycins B and C,** is produced by *Streptomyces fradiae* and by *Streptomyces albogriseolus*. It is marketed as a mixture that contains 85 – 90 % neomycin B [115]. Neomycins B and C are extremely stable in neutral or alkaline aqueous solution.

The final structure and stereochemistry of neomycins B and C were established in 1962. Neomycin C has also been synthesized also [116].

Paromomycin (Catenulin, Aminosidin, Hydroxymycin, Zygomycin A) is produced by *Streptomyces rimosus* subsp. *paromomycinus*; the structures of two isomers, paromomycins I and II, have been proposed. Paromomycin I is the main component of paromomycin preparations.

Catenulin, produced by *Streptomyces catenulae;* amminosidin, produced by *Streptomyces chrestomyceticus;* hydroxymycin, produced by *Streptomyces paucisporogenes;* and zygomycin A, produced by *Streptomyces pulveraceus,* are identical with paromomycin. Zygomycin A_1 is identical with paromomycin I and zygomycin A_2 is identical with paromomycin II.

The absolute configurations of the paromomycins have been determined along with those of other deoxystreptamine-containing aminoglycosides and dihydrostreptomycin [117], [118].

Ribostamycin (SF-733) is produced by *Streptomyces ribosidificus.* The free base is crystallized from methanol solution. The structure has been determined by chemical methods and total synthesis has been undertaken [119].

Kanamycin A is produced by *Streptomyces kanamyceticus.* The monosulfate monohydrate and the free base are crystallized from aqueous methanol. Kanamycin A is extremely stable in neutral or alkaline aqueous solutions.

The stereochemical structure of kanamycin A has been confirmed by X-ray analysis of its monosulfate monohydrate and monoselenate monohydrate crystals [120]. Total synthesis of kanamycin A has been achieved [121].

Kanamycin B (Bekanamycin, Aminodeoxykanamycin) is one of the two minor components that have been isolated from the culture filtrate of kanamycin-producing *Streptomyces kanamyceticus.* Kanamycin B has a retention factor (R_f) of 0.37, whereas kanamycin A, the major component, shows R_f 0.21 – 0.26. (R_f is a measure of the relative mobilities of substance and solvent in a chromatographic system.) The free base kanamycin B is crystallized from aqueous *N,N*-dimethylformamide.

Gentamicin (Gentamicin C Complex), a mixture of **Gentamicins C_1, C_{1a}, and C_2,** is the antibiotic complex produced by *Micromonospora purpurea* and *Micromonospora echinospora.* Gentamicins C_1 and C_2 are the principal products, and C_1 is itself a mixture of two major components designated C_1 and C_{1a}. Gentamicin C complex, which consists of the mixture of C_1 and C_{1a} (60 – 80 %) and of C_2 (20 – 40 %), has been used as a chemotherapeutic agent. The structures and the stereochemistry of the gentamicin C components have been reported [122].

Sisomicin (66 – 40, Rickamicin) is the major antibiotic produced by *Micromonospora inyoensis.* Its structure has been elucidated [123].

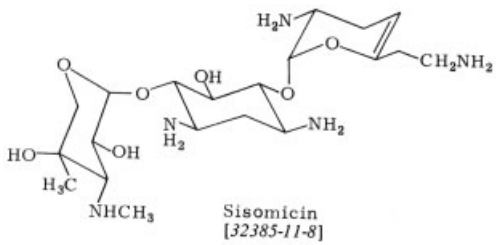

Dibekacin (3′,4′-Dideoxykanamycin B, DKB) was the first drug developed on the basis of the enzymatic mechanism of resistance to aminoglycosides. Dibekacin is synthesized from kanamycin B by the application of the Tipson-Cohen deoxygenation method after selective *N*- and *O*-protections [102]. A modified synthetic route via the 3′,4′-epoxy compound has given a high yield of more than 40 % on an industrial scale [124].

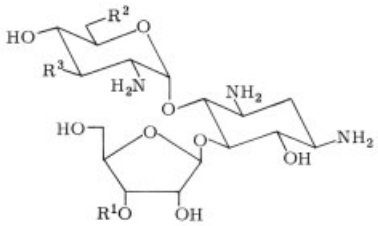

Name	R^1	R^2	R^3
Ribostamycin [*25546-65-0*]	H	NH_2	OH
Neomycin B [*119-04-0*]	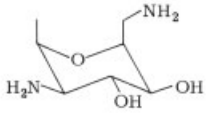	NH_2	OH
Paromomycin I [*7542-37-2*]	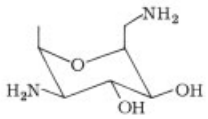	OH	OH
Lividomycin A [*36441-41-5*]	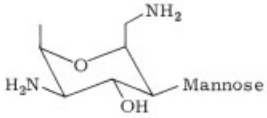	OH	H

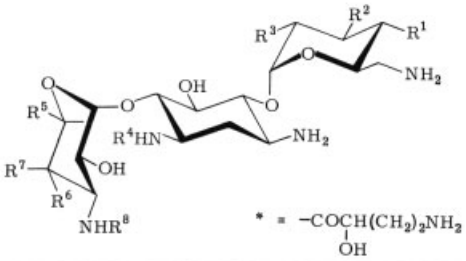

Name	R^1	R^2	R^3	R^4	R^5	R^6	R^7	R^8
Kanamycin [*59-01-8*]	OH	OH	OH	H	CH_2OH	OH	H	H
Bekanamycin [*4696-76-8*]	OH	OH	NH_2	H	CH_2OH	OH	H	H
Tobramycin [*32986-56-4*]	OH	H	NH_2	H	CH_2OH	OH	H	H
Gentamicin C_{1a} [*26098-04-4*]	H	H	NH_2	H	H	CH_3	OH	CH_3
Dibekacin [*34493-98-6*]	H	H	NH_2	H	CH_2OH	OH	H	H
Amikacin [*37517-28-5*]	OH	OH	OH	*	CH_2OH	OH	H	H

Amikacin (1-*N*-[(*S*-)-4-Amino-2-hydroxybutyryl]kanamycin, BB-K$_8$) is synthesized in 22 % yield from the 6′-*N*-protected kanamycin by selective 1-*N*-acylation with *N*-protected (*S*)-4-amino-2-hydroxybutyric acid using the active ester method [125]. The most important point of this synthesis is the selective protection of all of the amino groups except the 1-amino group. The 3,6′-di-*N*-protected kanamycin has been obtained in 95 % yield by selective *N*-protection using chelation with Co^{2+}, Ni^{2+}, and Cu^{2+} [126]. A new 3,6′,3″-tri-*N*-protection method using the selective 3″-*N*-trifluoroacetylation of the 3,6′-di-*N*-protected kanamycin with ethyl trifluoroacetate has been developed and the synthesis via 3,6′-di-*N*-(benzyloxycarbonyl)-3″-*N*-(trifluoroacetyl)kanamycin gives amikacin in a yield of more than 60 % [127].

The Fortimicin Group is a new type of deoxyaminoglycoside antibiotics, each consisting of glycine and a pseudodisaccharide, that have been found by screening. Fortimicin is produced by a *Micromonospora* species, sporaricin by *Saccharopolyspora hirsuta* subsp. *kobensis,* istamycin by *Streptomyces tenjimariensis,* sannamycin by *Streptomyces sannanensis,* and dactimicin by *Dactylosporangium matsuzakiense* [128–131], [128, p. 1061]. The structures of the fortimicin antibiotics are shown. These antibiotics strongly inhibit the growth of gram-positive and gram-negative bacteria but most *Pseudomonas* strains are resistant to them.

The axial amino group at C-1 in fortimicin A and istamycin A can be acetylated by aminoglycoside acetyltransferase(3)-I, but the equatorial amino group at C-1 in sporaricin A and istamycin B is scarcely acetylated. Other aminoglycoside-modifying enzymes that participate in the resistance to deoxystreptamine-containing aminoglycosides do not inactivate fortimicin-group antibiotics.

3-*O*-Demethyl derivatives of sporaricin A and istamycin B exhibit good activity not only against gram-positive and gram-negative bacteria but also against most *Pseudomonas* strains. These derivatives will be developed as valuable chemotherapeutic agents in the near future.

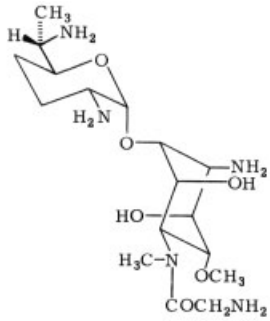

Fortimicin A
[55779-06-1]

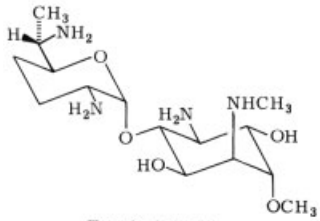

Fortimicin B
[54783-95-8]

3.4. Nucleosides

The biological effects associated with metabolic processes and specific enzyme control mechanisms are diverse in naturally occurring nucleosides and their synthetic analogs. Nucleosides exhibit several biological effects, including antibiotic, anticancer, and antiviral activities. They possess antimitotic and immunosuppressive activities and cardiovascular and other effects [132–134]. Moreover, it should be kept in mind that nucleoside analogs can assume other functional roles not as yet recognized, and that further therapeutic applications can be expected in the future. These analogs are obtained predominantly from microbial sources.

The nucleoside antibiotics consist of a heterocyclic base aglycone and a carbohydrate or a carbocyclic ring linked by a carbon – nitrogen (*N*-nucleoside) or a carbon – carbon bond (*C*-nucleoside). The nucleoside antibiotics fall somewhat outside the normal field of antibiotics with respect to their activity spectra and hence to their use. They are important for use against fungi, viruses, and certain types of cancer cells. Some typical nucleoside antibiotics are mentioned here.

3.4.1. *N*-Nucleosides

5-Azacytidine (**17**), a triazine analog of cytidine produced by *Streptoverticillium ladakanus*. It is active against some bacterial strains, Ehrlich ascitic tumor, leukemia L1210, and certain other leukemias. 5-Azacytidine also inhibits the DNA synthesis of bacteriophage T4 [135].

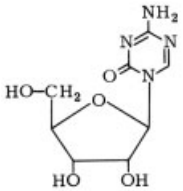

17
5-Azacytidine
[*320-67-2*]

Bredinin (**18**), produced by *Eupenicillium brefeldianum,* shows marked immunosuppressive activity in mice, interferes with replication of *Vaccinia* virus in vitro, and inhibits leukemia L 5178 cells and *Candida albicans* [136].

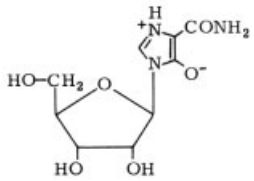

18
Bredinin
[*50924-49-7*]

Coformycin (**19a**) is isolated from *Nocardia interforma* along with formycin. Coformycin shows a synergistic effect with formycin on Yoshida rat sarcoma cells because of its strong inhibition of adenosine deaminase, which inactivates formycin. Coformycin, having a characteristic seven-membered ring base moiety, is thought to be a typical example of a "transition-state analog" in the adenosine deaminase reaction. 2′-Deoxycoformycin (**19b**) has also been isolated also [137–139].

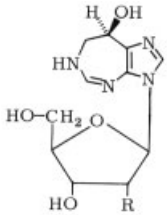

19a: R = OH
Coformycin
[*11033-22-0*]

19b: R = H
2′-Deoxycoformycin
[*53910-25-1*]

Cordycepin (**20**), 3′-deoxyadenosine, was one of the first nucleoside antibiotics isolated from *Cordyceps militaris.* It inhibits *Bacillus subtilis, Mycobacterium tuberculosis,* KB cell cultures, and Ehrlich ascites tumor cells.

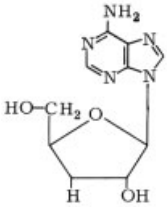

20
Cordycepin
[*73-03-0*]

Crotonoside (**21**), isolated from *Croton tiglium* seeds, acts as a vasopressor [140].

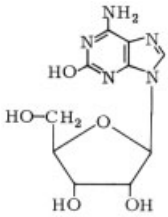

21
Crotonoside
[*1818-71-9*]

Nebularine (**22**), produced by the mushroom *Agaricus* (*Clitocybe*) *nebularis,* inhibits the growth of *Mycobacterium tuberculosis* and *Brucella abortus,* and is markedly cytotoxic to mammalian cells, whereas the purine base is relatively nontoxic. Nebularine is toxic to Sarcoma 180 cells [141], [142].

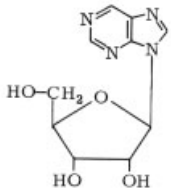

22
Nebularine
[*550-33-4*]

Toyocamycin (**23a**), produced by *Streptomyces toyocaensis,* strongly inhibits *Candida albicans, Trichophyton interdigitale,* and *Mycobacterium tuberculosis* and is also active against NF-sarcoma cells [143]. **Tubercidin** (**23b**) and **sangivamycin** (**23c**) also belong to this class [144].

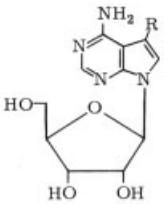

23a: R = CN Toyocamycin [*606-58-6*]
23b: R = H Tubercidin [*69-33-0*]
23c: R = $CONH_2$ Sangivamycin [*18417-89-5*]

Polyoxins A – O (**24**, **25**, **26**) are peptide nucleosides produced by *Streptomyces cacaoi*. They possess various heterocycles, e.g., uracil, thymine, 5-hydroxymethyluracil, uracil-5-carboxylic acid, or formylimidazolone. The compounds **24** and **26** are particularly active against sheath-blight in rice, *Pellicularia sasakii,* and are widely used as agricultural drugs [145–147].

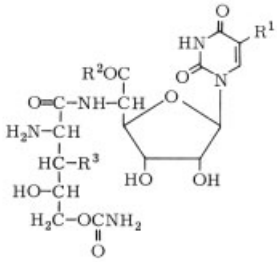

24

Name	R^1	R^2	R^3
Polyoxin A [*19396-03-3*]	CH_2OH	$CH_3-CH=$ 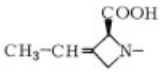	OH
Polyoxin B [*19396-06-6*]	CH_2OH	HO	OH
Polyoxin D [*22976-86-9*]	COOH	HO	OH
Polyoxin E [*22976-87-0*]	COOH	HO	H
Polyoxin F [*23116-76-9*]	COOH	$CH_3-CH=$ 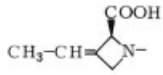	OH
Polyoxin G [*22976-88-1*]	CH_2OH	HO	H
Polyoxin H [*24695-54-3*]	CH_3	$CH_3-CH=$ 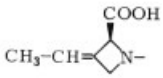	OH
Polyoxin J [*22976-89-2*]	CH_3	HO	OH
Polyoxin K [*22886-46-0*]	H	$CH_3-CH=$ 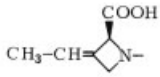	OH
Polyoxin L [*22976-90-5*]	H	HO	OH
Polyoxin M [*34718-88-2*]	H	HO	H

3.4.2. *C*-Nucleosides

Formycin (**27**) (**Formycin A**) is isolated from *Nocardia interforma* and from *Streptomyces lavendulae* [148], [149]. The antibiotic is effective against *Xanthomonas oryzae* and *Pellicularia filamentosa*. Its activity against Yoshida rat sarcoma cell is enhanced by coformycin. **Formycin B** (**28**) inhibits *Xanthomonas oryzae* and interferes with multipli-

cation of influenza A virus in the cells of chick chorioallantoic membrane [150]. **Oxoformycin B** (**29**) shows no activity against *Xanthomonas oryzae* [151].

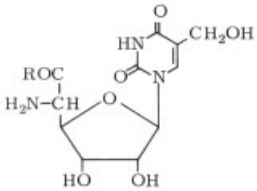

25

Polyoxin C [*21027-33-8*] R = OH

Polyoxin I [*22886-33-5*] R =

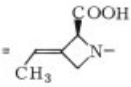

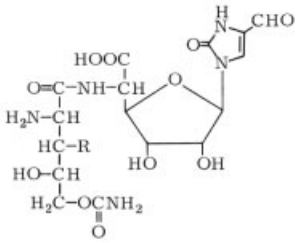

26

Polyoxin N [*37362-29-1*] R = OH

Polyoxin O [*37362-28-0*] R = H

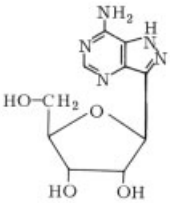

27
Formycin
[*6742-12-7*]

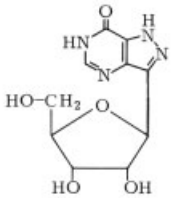

28
Formycin B
[*13877-76-4*]

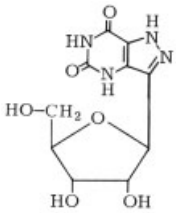

29
Oxoformycin B
[*19246-88-9*]

Showdomycin (**30**), isolated from *Streptomyces showdoensis,* is very active against *Streptococcus hemolyticus*. It is moderately active against other gram-positive and gram-negative bacteria and also effective against Ehrlich ascites tumor in mice and HeLa cells [152]. **Oxazinomycin** (**31**) belongs to this class of nucleosides [153].

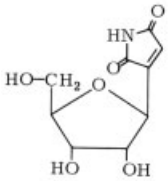

30
Showdomycin
[*16755-07-0*]

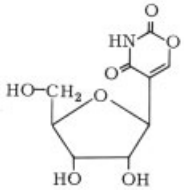

31
Oxazinomycin
[*32388-21-9*]

3.4.3. Carbocyclic Nucleosides

Since the pioneering synthesis of the racemic carbocyclic analog of adenosine by SHEALY and CLAYTON and the subsequent isolation of **aristeromycin** (**32**) from *Streptomyces citricolor*, the interest in this class of compounds has been renewed by the isolation of a new carbocyclic nucleoside, **neplanocin A** (**33**). The latter exhibits remarkable antitumor activity against L 1210 leukemia in mice, and its synthetic analogs are now being studied extensively [154–156].

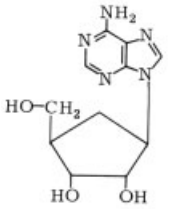

32
Aristeromycin
[*19186-33-5*]

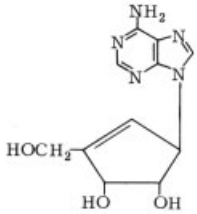

33
Neplanocin A
[*72877-50-0*]

3.4.4. An Exceptional Nucleoside

Blasticidin S (**34**) has a pyran ring as the sugar moiety and inhibits various gram-positive and gram-negative bacteria. It is particularly effective against *Pericularia oryzae* and is now used as an agricultural drug [157].

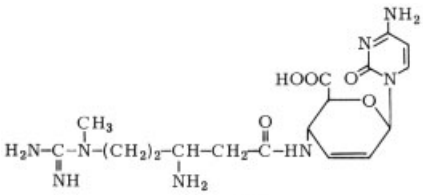

34
Blasticidin S
[*2079-00-7*]

3.5. Macrolides

This group of antibiotics is characterized by excellent antibacterial activity, particularly against gram-positive bacteria. Macrolides can be defined and distinguished from the other groups of antibiotics by the unique feature of their chemical structure. They are polyfunctional macrocyclic lactones and the majority of them contain at least one amino sugar moiety, which is the cause of the basicity of the molecules. Neutral macrolides containing only a neutral sugar moiety are also known. These antibiotics have become targets in the aldol strategy of organic synthesis to construct their polyhydroxy functions stereoselectively [158], [159]. The antibiotics are classified as either 12-, 14-, or 16-membered ring macrolides according to the ring size of the aglycone.

3.5.1. 12-Membered Ring Macrolides

Methymycin (**35**), produced by *Streptomyces venezuelae,* was first shown to be a 12-membered lactone, comprising the aglycone or methynolide and D-desosamine [160], [161]. **Neomethymycin** (**36**) has an isomeric structure.

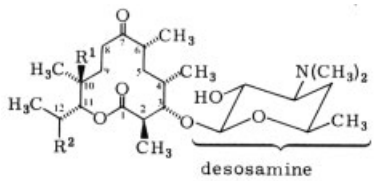

Methymycin (**35**) R^1 = OH, R^2 = H
[*497-72-3*]
Neomethymycin (**36**) R^1 = H, R^2 = OH
[*497-73-4*]

3.5.2. 14-Membered Ring Macrolides

The erythromycins, produced by *Streptomyces erythreus,* are clinically important macrolides and are the most widely investigated 14-membered ring macrolides [162]. Extensive chemical and X-ray crystallographic studies of **erythromycin A** (**37**) have established its structure as well as those of its minor components, **erythromycin B** (**38**), **erythromycin C** (**39**), and **erythromycin D** (**40**). Erythromycin is effective against streptococcal and pneumococcal infections. Derivatives of erythromycin, modified in the cladinose ring, the desosamine ring, and the aglycone moiety (especially at C-9 of the aglycone), have been described. Their characterization contributed greatly to understanding the chemistry and structure – activity relationships of the macrolide antibiotics [163].

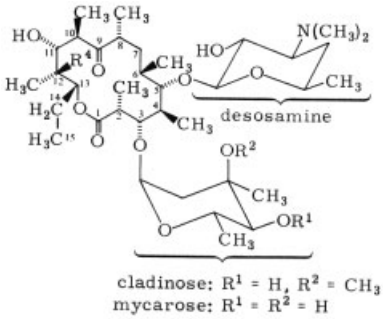

Erythromycin A (**37**) [*114-07-8*]
$R^1 = R^3 = H$, $R^2 = CH_3$, $R^4 = OH$

Erythromycin B (**38**) [*527-75-3*]
$R^1 = R^3 = R^4 = H$, $R^2 = CH_3$

Erythromycin C (**39**) [*1675-02-1*]
$R^1 = R^2 = R^3 = H$, $R^4 = OH$

Erythromycin D (**40**) [*33442-56-7*]
$R^1 = R^2 = R^3 = R^4 = H$

Picromycin (**41**), **narbomycin** (**42**), and **oleandomycin** (**43**) belong to the 14-membered lactones.

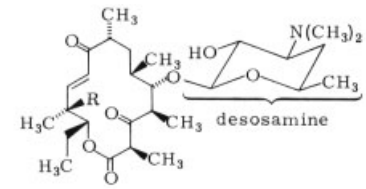

Picromycin (**41**) [*19721-56-3*] R = OH
Narbomycin (**42**) R = H

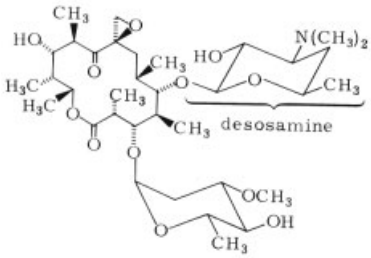

Oleandomycin (**43**) [*3922-90-5*]

3.5.3. 16-Membered Ring Macrolides

Stereochemical structures of carbomycins, leucomycins, spiramycins, and other macrolides have been disclosed by extensive chemical studies [164], [165]. These compounds have a formylmethyl group at C-6 and two conjugated double bonds in the 16-membered lactone ring. They differ only in the nature of the acyl substituents at C-3 and C-4. For instance, the leucomycin complex (kitasamycin), produced by *Streptomyces kitasatoensis* [166] is a mixture of ten similar components. These are **leucomycin A_1** (**44**, turimycin H_5), **leucomycin A_3** (**45**, josamycin, YL-704A_3 or platenomycin A_3), **leucomycin A_4** (**46**), **leucomycin A_5** (**47**), **leucomycin A_6** (**48**, YL-704B_3 or platenomycin B_3), **leucomycin A_7** (**49**), **leucomycin A_8** (**50**), **leucomycin A_9** (**51**), **leucomycin U** (**52**), and **leucomycin V** (**53**) [164], [167].

The spiramycins (foromacidins) produced by *Streptomyces ambofaciens* [168] have been separated into three components, namely **spiramycin I** (**54**), **spiramycin II** (**55**), and **spiramycin III** (**56**).

Many other macrolides, such as **rosamicin** [169], [170], **cirvamycin A_1** [171], [172], **juvenimicin A_2, juvenimicin A_4** [173], [174], **deltamycin** [175], **carbomycin** [176], [177], **angolamycin** [178], **tylosin** (**57**) [179–181], and the **mycinamicins** [182], [183], belong to this class.

Since 1950, the structures of more than 90 macrolides have been elucidated, and this knowledge has had a great impact on modern organic synthesis. The classification of the above-mentioned macrolides according to ring size is useful, but very schematic. There is a marked change in the antibiotic activity spectrum corresponding to changes in the ring size, even though many of the substituents and the degree of unsaturation differ considerably.

Macrolides with ca. 10 – 16 ring members are very strongly antibacterial, as are the smaller lactones. Because there is no cross-resistance among these macrolides, they are mainly used to treat bacterial infections that are resistant to other antibiotics.

Macrolides with ca. 16 – 40 ring members show very little activity against bacteria, often

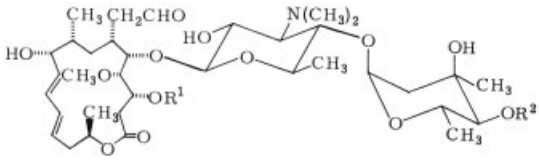

Leucomycin A_1 (Turimycin H_5) (**44**) $R^1 = H$, $R^2 = COCH_2CH(CH_3)_2$
[*16846-34-7*]

Leucomycin A_3 (Josamycin, YL-704 A_3 Turimycin A_5) (**45**) $R^1 = COCH_3$, $R^2 = COCH_2CH(CH_3)_2$
[*16846-24-5*]

Leucomycin A_4 (**46**) $R^1 = COCH_3$, $R^2 = COCH_2CH_2CH_3$
[*18361-46-1*]

Leucomycin A_5 (**47**) $R^1 = H$, $R^2 = COCH_2CH_2CH_3$
[*18361-45-0*]

Leucomycin A_6 (YL-704 B_3) (**48**) $R^1 = COCH_3$, $R^2 = COCH_2CH_3$
[*18361-48-3*]

Leucomycin A_7 (**49**) $R^1 = H$, $R^2 = COCH_2CH_3$
[*18361-47-2*]

Leucomycin A_8 (**50**) $R^1 = COCH_3$, $R^2 = COCH_3$
[*18361-50-7*]

Leucomycin A_9 (**51**) $R^1 = H$, $R^2 = COCH_3$
[*18361-49-4*]

Leucomycin U (**52**) $R^1 = COCH_3$, $R^2 = H$
[*31642-61-2*]

Leucomycin V (**53**) $R^1 = H$, $R^2 = H$
[*22875-15-6*]

none at all, but they are highly effective against fungi, yeasts, etc.

The very large macrolides are highly effective against not only fungi, but also viruses and tumors.

3.6. Ansamycins

The ansamycins are a clinically important class of antibiotics with a characteristic structure. They have an aliphatic "ansa" bridge that connects two nonadjacent positions of an aromatic system [184]. The name "ansamycin" is based on the term "ansa compounds" [185]. Ansamycins are classified into two groups based on the nature of the aromatic moiety, i.e., benzoquinoid and naphthoquinoid ansamycins. Geldanamycin and maytansinoids belong to the benzoquinoid ansamycins and are studied as potential antitumor agents. The naphthoquinoid ansamycins include rifamycins, tolypomycins, streptovaricins, halomicins, and naphthomycin. The naphthoquinoid ansamycins are the major group of known ansamycins.

The rifamycins, produced by *Nocardia mediterranei,* have great therapeutic value [186, 188]. Their chemistry is very similar to that of the macrolides. After many attempts to separate, isolate, and purify the naturally occurring rifamycins, **rifamycin B** (**58**), **rifamycin O** (**59**), and **rifamycin S** (**60**) were found among the fermentation products. Rifamycin B was moderately effective against gram-positive bacteria. The oxidation of rifamycin B gave rifamycin O, which can be hydrolyzed to the more active rifamycin S. The latter can be reduced to **rifamycin SV** (**61**) using ascorbic acid. Rifamycin SV is converted, through its formyl derivative (**62**), to the therapeutically important **rifampicin** (**63**). In order to obtain an antibiotic with a broader spectrum and good oral absorption characteristics, thousands of derivatives of rifampicin have been prepared. Ri-

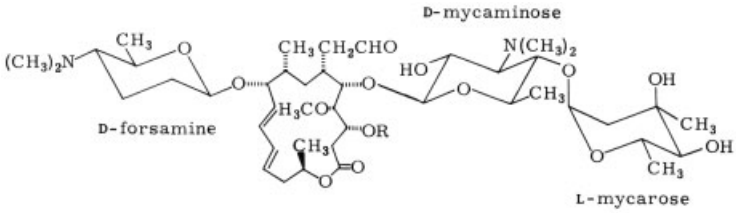

Spiramycin I (Foromacidin A) (**54**): R = H
[*24916-50-5*]

Spiramycin II (Foromacidin B) (**55**): R = $COCH_3$
[*24916-51-6*]

Spiramycin III (Foromacidin C) (**56**): R = $COCH_2CH_3$
[*24916-52-7*]

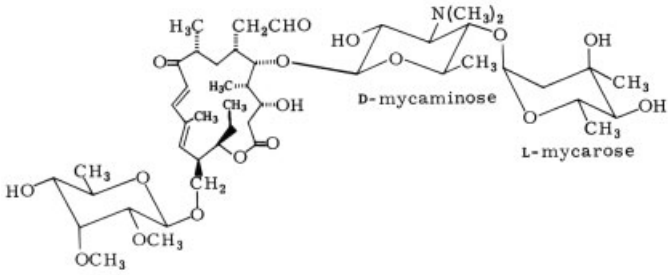

Tylosin (**57**)
[*1401-69-0*]

famycins B, O, and S all are used as starting materials for the modifications. The rifampicins have strong biological activity against gram-positive microorganisms and mycobacteria, particularly *Mycobacterium tuberculosis* [189–191]. Most of the ansamycins are weakly active against viruses, and certain derivatives, such as **3-formylrifamycin S** (**62**), have been found to be active against certain tumors. No cross-resistance of the ansamycins to most of the other antibiotics has been observed.

3.7. Peptides

The various important functions of the living organism are frequently mediated by oligopeptides and proteins, which exist in the most diverse structures. It is therefore not surprising that a large number of low molecular mass peptides, oligopeptides, and protein-like substances are found among the antibiotics of microbial origin. Although peptide antibiotics consist of amino acids linked by peptide bonds, they differ from the proteins and peptides of higher animals and plants in many respects [192]. The following characteristics frequently are found in the peptide antibiotics:

1) Molecular masses of the antibiotics are smaller (in the range of 500 – 1500) than those of peptide hormones, which are frequently much larger.
2) The antibiotics contain some uncommon amino acids that are not found in proteins and peptide hormones of animal or plant origin. The usual amino acids are infrequently detected or are found in modified forms.
3) Lipids and other moieties not of amino acid character are found in many peptide antibiotics.
4) The peptide antibiotics frequently contain D-amino acid residues, whereas peptides of

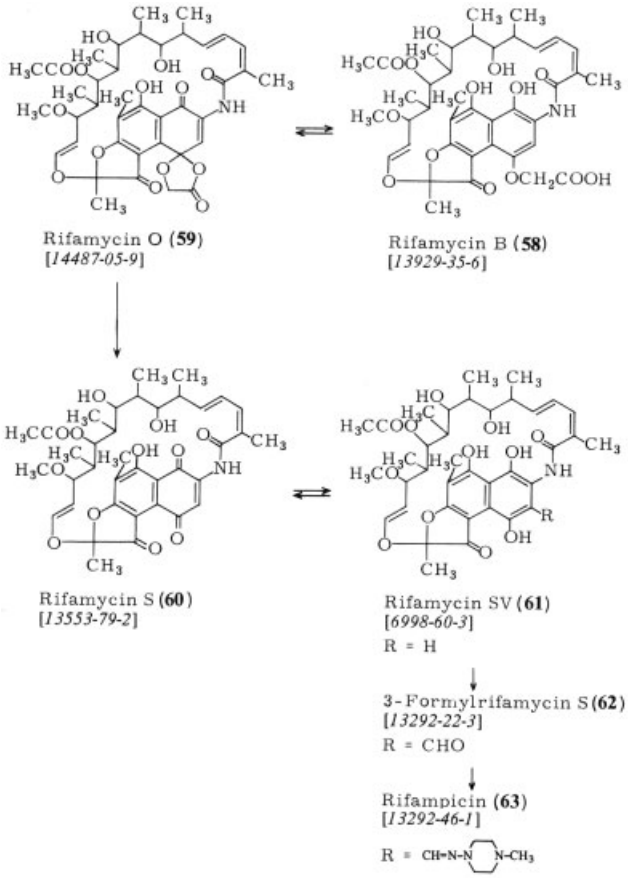

plant and animal origin consist solely of L-amino acid residues.

5) Virtually all of the peptide antibiotics resist hydrolysis by proteolytic enzymes, which are otherwise effective in hydrolyzing peptides of plant and animal origin.
6) The antibiotics are often cyclic peptides.
7) Families of closely related peptide antibiotics are frequently produced by the same microorganism.

3.7.1. The Bleomycin Group

The bleomycins, a group of glycopeptide antibiotics produced by *Streptomyces verticillus* [193], make up one of the most widely used groups of antitumor antibiotics, effective against squamous cell carcinoma and malignant lymphoma. Extensive degradation studies have shown the main structural features to be a peptide containing unusual amino acid residues and a disaccharide of uncommon sugars. The complete structure has been elucidated by chemical studies and X-ray crystallographic analysis of P-3A, a biosynthetic intermediate structurally related to bleomycin [194], [195]. This structure has been verified by the total synthesis of **bleomycin A_2** (**67**). The naturally occurring bleomycins are obtained in copper-chelated form as a mixture of congeners that differ only in the substituents at the *C* terminus of **bleomycinic acid** (**64**), the common structural unit. Among these are **bleomycin A_1** (**65**), **demethylbleomycin A_2** (**66**), **bleomycin A_2** (**67**), **bleomycin $A_{2\prime-a}$** (**68**), **bleomycin $A_{2\prime-b}$** (**69**), **bleomycin $A_{2\prime-c}$**(**70**), **bleomycin A_5** (**71**), **bleomycin A_6** (**72**), **bleomycin B_1** (**73**), **bleomycin B_2** (**74**), and **bleomycin B_4**

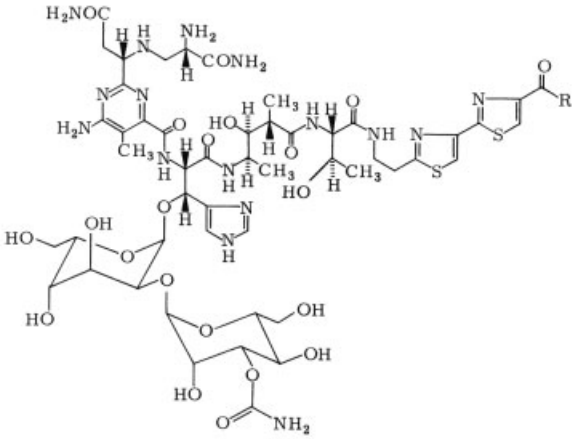

Bleomycinic acid (**64**) [*37364-66-2*]: R = OH

Bleomycin A_1 (**65**) [*58995-26-9*]: R = $NH(CH_2)_3\overset{O}{\overset{\|}{S}}CH_3$

Bleomycin demethyl A_2 (**66**) [*41089-03-6*]: R = $NH(CH_2)_3SCH_3$

Bleomycin A_2 (**67**) [*11116-31-7*]: R = $NH(CH_2)_3S^+(CH_3)_2$, X^-

Bleomycin $A_{2'}$ -a (**68**) [*73666-81-6*]: R = $NH(CH_2)_4NH_2$

Bleomycin $A_{2'}$ -b (**69**) [*41138-53-8*]: R = $NH(CH_2)_3NH_2$

Bleomycin $A_{2'}$ -c (**70**) [*62960-69-4*]: R = $NH(CH_2)_2$

Bleomycin A_5 (**71**) [*11116-32-8*]: R = $NH(CH_2)_3NH(CH_2)_4NH_2$

Bleomycin A_6 (**72**) [*37293-17-7*]: R = $NH(CH_2)_3NH(CH_2)_4NH(CH_2)_3NH_2$

Bleomycin B_1 (**73**) [*41138-54-9*]: R = NH_2

Bleomycin B_2 (**74**) [*9060-10-0*]: R = $NH(CH_2)_4NH\overset{NH}{\overset{\|}{C}}-NH_2$

Bleomycin B_4 (**75**) [*9060-11-1*]: R = $NH(CH_2)_4NH\overset{NH}{\overset{\|}{C}}NH(CH_2)_4NH\overset{NH}{\overset{\|}{C}}-NH_2$

Pepleomycin (**76**) [*68247-85-8*]: R = $NH(CH_2)_3NH-\overset{CH_3}{\underset{H}{C}}-C_6H_5$

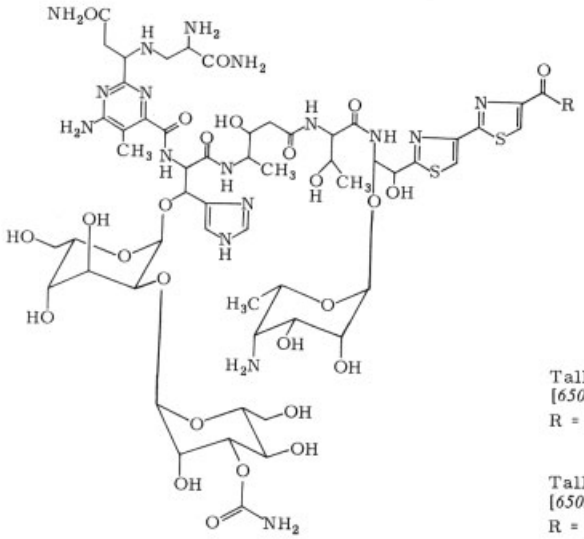

Tallysomycin A (**77**)
[*65057-90-1*]
R = $NH(CH_2)_3\underset{NH_2}{C}HCH_2CONH(CH_2)_3NH(CH_2)_4NH_2$

Tallysomycin B (**78**)
[*65057-91-2*]
R = $NH(CH_2)_3NH(CH_2)_4NH_2$

(**75**). Metal-free bleomycin can be prepared by treatment with hydrogen sulfide. A mixture of metal-free bleomycins consisting mainly of A_2 (55 – 70 %) and B_2 (25 – 32 %) has been used for clinical treatment because the mixture has an effect superior to that of A_2 alone on human squamous cell carcinoma. The copper ion in bleomycin is replaced by iron after penetration into cells. A bleomycin – iron complex that exerts antitumor activity is formed. More than 300 bleomycin analogs have been prepared by chemical modifications or fermentations. **Pepleomycin** (**76**), possessing improved properties, has been brought into clinical use [196] and the **tallysomycins A and B**, **77** and **78**, respectively, are similar glycopeptides [197], [198].

3.7.2. The Gramicidin Group

Generally speaking, peptide antibiotics with simple straight-chain structures play a minor role and are mentioned here only for review purposes. However, peptide antibiotics with cyclic structures, which often contain D-amino acids, are of considerable importance. The usual nomenclature is used here for the constituents of the amino acids: Ala = alanine; Asp = asparagine; Cys = cysteine; Dab = α,γ-diaminobutyric acid; Glu = glutamic acid; Gly = glycine; His = histidine; Ileu = isoleucine; Leu = leucine; Lys = lysine; Orn = ornithine; Phe = phenylalanine; Pro = proline; Thr = threonine; Trp = tryptophan; Tyr = tyrosine; Val = valine; Meval = *N*-methylvaline; Sar = sarcosine; Ser = serine; MOA = 6-methyloctanoic acid; IOA = isooctanoic acid. The arrow indicates the direction of the amide bond between the amino acids. Thus, the arrow begins where the carbonyl group attaches and ends where the amino group attaches (– CO – NH →).

The gramicidin group consists of a series of open chain peptides and a series of cyclic peptides. Clinically important **gramicidin S** (**79**) was isolated from a strain of *Bacillus brevis* [199].

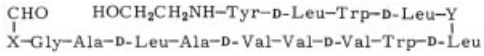

	X	Y
Valine-gramicidin A	Val	Trp
Isoleucine-gramicidin A	Ile	Trp
Valine-gramicidin B	Val	Phe
Isoleucine-gramicidin B	Ile	Phe
Valine-gramicidin C	Val	Tyr
Isoleucine-gramicidin C	Ile	Tyr

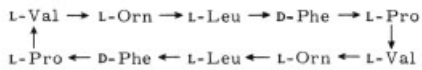

Gramicidin S (**79**)
[*113-73-5*]

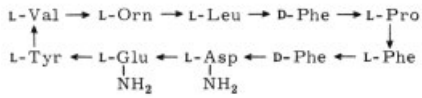

Tyrocidine A
[*1481-70-5*]

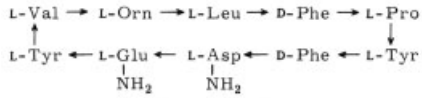

Tyrocidine B
[*865-28-1*]

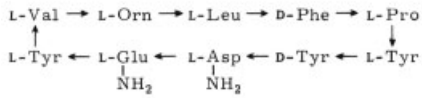

Tyrocidine C
[*3252-29-7*]

3.7.3. The Polymyxins

Among the polymyxins, **polymyxins B, D, and E** are of particular importance. The **colistins A and B** correspond to **polymyxins E_1 and E_2,** respectively, as shown in **80** [200]. These are active only against gram-negative bacteria.

80

	R	X	Y	Z
Polymyxin A_1 (M1) [*65454-50-4*]	MOA	D-Dab	D-Leu	Thr
Polymyxin A_2 (M2) [*65454-51-5*]	IOA	D-Dab	D-Leu	Thr
Polymyxin B_1 [*4135-11-9*]	MOA	Dab	D-Phe	Leu
Polymyxin B_2 [*34503-87-2*]	IOA	Dab	D-Phe	Leu
Polymyxin D_1 [*10072-50-1*]	MOA	D-Ser	D-Leu	Thr
Polymyxin D_2 [*34167-45-8*]	IOA	D-Ser	D-Leu	Thr
Polymyxin E_1 (colistin A) [*7722-44-3*]	MOA	Dab	D-Leu	Leu
Polymyxin E_2 (colistin B) [*7239-48-7*]	IOA	Dab	D-Leu	Leu

Dab = α, γ-diaminobutyric acid; MOA = (+)-6-methyloctanoic acid; IOA = isooctanoic acid

3.7.4. The Bacitracins

Bacitracin, produced by *Bacillus subtilis* [201] and *Bacillus licheniformis* [202], is active against gram-positive bacteria. After extensive study on the chemistry of bacitracins, a revised structure was proposed for **bacitracin A** (**81**) [203].

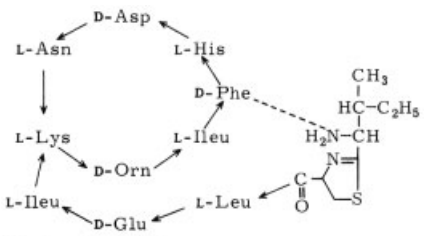

Bacitracin A (**81**)
[*1402-99-9*]

3.7.5. Large-Ring Peptide Antibiotics Containing Lactone Linkages

This important group can be classified with macrolide antibiotics because it contains a lactone moiety, but it is more properly regarded as a group of cyclic peptides. These peptides generally contain about 16- to 19-membered rings, including many amide bonds and lactone linkages. There are no C–C double bonds in the ring.

3.7.6. The Actinomycin Group

These cyclic peptide antibiotics are also known as chromopeptide antibiotics. The **actinomycins** (**82**), first isolated by WAKSMAN in 1940 from *Streptomyces antibioticus,* are of interest because they have found use in therapy of tumors, particularly Wilm's tumor (see next page).

They are also used as a biochemical tool because of their specific inhibition of DNA-primed RNA synthesis. Structurally they have two pentapeptide lactone rings attached to a phenoxazinone chromophore. Several of the actinomycins isolated were found to be identical and the nomenclature is confusing. For example, actinomycin D and actinomycin C_3 are also known as dactinomycin and cactinomycin, respectively [204].

3.7.7. Other Peptide Antibiotics

There are many peptide antibiotics derived from *Streptomyces* species. Among these are amphomycin, capreomycin, distamycin, the enduracidins, mikamycin, neocarzinostatin (antitumor), stendomycin, viomycin, and virginiamycin.

3.8. Other Important Antibiotics and Intermediates

Cycloheximide (**83**) is produced by *Streptomyces griseus* [205]. It is highly effective against fungi and is therefore used mainly for plant protection. The analogs streptovitacin, naramycin B, and streptimidone have also been isolated, and also are called glutarimide antibiotics because of their common structural moiety.

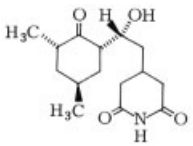

Cycloheximide (**83**)
[*66-81-9*]

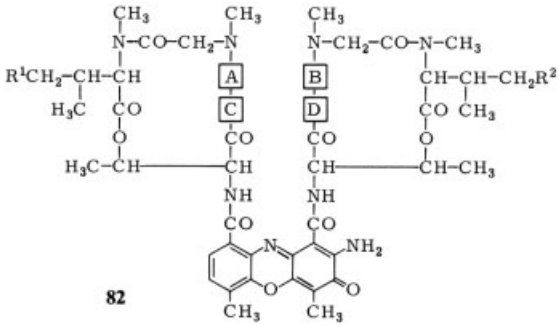

82

Name	A	B	C	D	R^1	R^2
Actinomycin C_2 [*2612-14-8*]	L-Pro	L-Pro	D-Val	D-allo-Ile	H	H
Actinomycin C_{2a} (VI) [*17914-41-9*]	L-Pro	L-Pro	D-allo-Ile	D-Val	H	H
Actinomycin C_3 (VII) [*6156-47-4*]	L-Pro	L-Pro	D-all-Ile	D-allo-Ile	H	H
Actinomycin D (C_1, IV) [*50-76-0*]	L-Pro	L-Pro	D-Val	D-Val	H	H
Actinomycin E_1 [*1402-41-1*]	L-Pro	L-Pro	D-allo-Ile	D-allo-Ile	H	CH_3
Actinomycin E_2 [*1402-42-2*]	L-Pro	L-Pro	D-allo-Ile	D-allo-Ile	CH_3	CH_3
Actinomycin F_3 [*1402-46-6*]	Sar	Sar	D-allo-Ile	D-allo-Ile	H	H
Actinomycin F_8 (II, A_{II}) [*32934-48-8*]	Sar	Sar	D-Val	D-Val	H	H
Actinomycin F_9 (III, A_{III}) [*60469-11-8*]	L-Pro	Sar	D-Val	D-Val	H	H
Actinomycin $X_{0\beta}$ (I) [*1402-60-4*]	L-Pro	Pro(4-OH)	D-Val	D-Val	H	H
Actinomycin X_2 (V, A_V, B_V) [*1402-61-5*]	L-Pro	Pro(4-O)	D-Val	D-Val	H	H

Cycloserine (**84**) (**oxymycin, seromycin, orientomycin**) is the simplest antibiotic, D-4amino-3-isoxazolidine, isolated from many *Streptomyces* species [206]. Cycloserine is now produced only synthetically and used particularly for tuberculosis of the lungs and for leprosy with *p*-aminosalicylic acid (PAS) or isonicotinic acid hydrazide (INH).

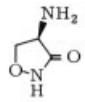

Cycloserine (**84**)
[*68-41-7*]

Variotin (**85**) (**pecilocin**) is produced by *Paecilomyces varioti* [207]. It is an oily, ester-like substance with an aromatic odor and is particularly active against fungi. Variotin is used against trichophytes.

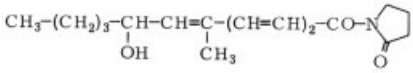

Variotin (**85**)
[*19504-77-9*]

Sarkomycin A (**86**) is produced by *Streptomyces erythrochromogenes* [208]. It is active not only as an antibiotic but also as an antitumor agent.

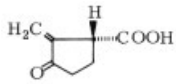

Sarkomycin (**86**)
[*489-21-4*]

Novobiocin, isolated from a culture filtrate of *Streptomyces nieveus,* has the structure **87**, consisting of the aglycone, novobiocic acid, and 3-*O*-carbamoylnoviose. The monosodium or calcium salt is used in therapy and is active mainly against gram-positive bacteria [209].

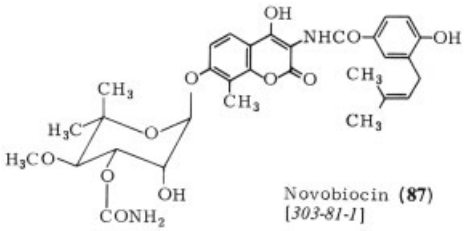

Novobiocin (**87**)
[*303-81-1*]

Griseofulvin (**88**) is produced by *Penicillium griseofulvum, P. janczewskii,* and *Nigrospora oryzae.* It is unique in possessing the spirocarbon moiety [210], [211]. Griseofulvin is very active against fungi, and it is used orally to treat fungal infections of human skin.

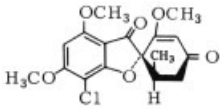

Griseofulvin (**88**)
[*126-07-8*]

Chloramphenicol (**89**), the first of the so-called broad-spectrum medicinal antibiotics, was originally obtained from *Streptomyces venezuelae* in 1947 [212]. It is now manufactured by a chemical process, and the parent compound and its esters are commercially available. Chloramphenicol is active against rickettsia, chlamydiae, and mycoplasmas, as well as a wide range of gram-positive and gram-negative bacteria. However, use is limited by the risk of bone marrow damage or aplastic anemia at too high or too prolonged an application [213].

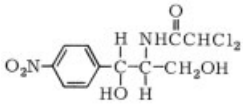

Chloramphenicol (**89**)
[*56-75-7*]

Mitomycins (**90–93**) are a group of unique chemical structures in which three different carcinostatic functions – aziridine, carbamate, and quinone – are arranged around a pyrro[1,2-*a*]indole nucleus [214]. The first mitomycins were discovered in 1956 by Hata in a culture filtrate of *Streptomyces caespitosus.* These compounds, designated mitomycins A and B, show highly potent antibacterial activity and moderate antitumor activity, but they are quite toxic in mice. In 1958, mitomycin C, an extremely valuable antitumor drug, was isolated from *Streptomyces caespitosus* [215], [216]. In 1960, another mitomycin, porfiromycin (**93**), was isolated from *Streptomyces ardus.*

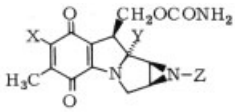

Name	X	Y	Z
Mitomycin A (**90**) [*4055-39-4*]	OCH_3	OCH_3	H
Mitomycin B (**91**) [*4055-40-7*]	OCH_3	OH	CH_3
Mitomycin C (**92**) [*50-07-7*]	NH_2	OCH_3	H
Porfiromycin (**93**) [*801-52-5*]	NH_2	OCH_3	CH_3

Fumagillin (**94**) is a useful polyene antibiotic; feeding it to honeybees with natural *Nosema apis* infections suppressed the disease and led to considerably increased honey production [217].

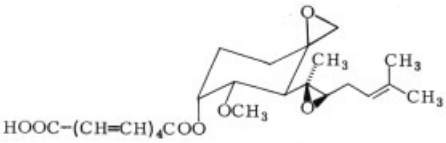

Fumagillin (**94**)
[*23110-15-8*]

Monensins (**95**) are useful polyether antibiotics that control infections of *Coccidia*. They are particularly important to the poultry industry [218].

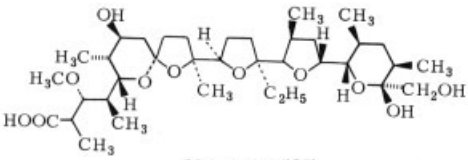

Monensin (**95**)
[*17090-79-8*]

Pyrrolnitrin (**96**), isolated from *Pseudomonas* species, is highly active against fungi, particularly trichophyte species.

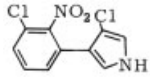

Pyrrolnitrin (**96**)
[*1018-71-9*]

Fosfomycin (**97**) is unique in possessing a simple epoxide ring and has a broad activity spectrum against gram-positive and gram-negative bacteria [219].

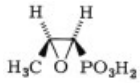

Fosfomycin (**97**)
[*23155-02-4*]

Fusidic acid (**98**) has a steroidal skeleton, but it is markedly different from the usual steroid hormones in biological activity. Fusidic acid has been isolated from *Fusidium coccineum* and is particularly active against *Staphylococcus, Clostridium, Neisserias, Corynebacterium diphtheriae,* and *Mycobacterium tuberculosis* [220].

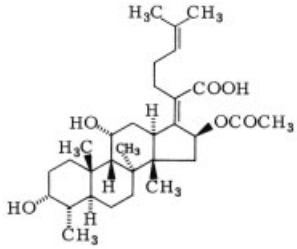

Fusidic acid (**98**)
[*6990-06-3*]

d-(*p*-Hydroxyphenyl)glycine (d-HPG) is widely used in large amounts as an important intermediate for the synthesis of amoxicillin and several other semisynthetic β-lactam antibiotics.

Industrially, D-HPG has been produced by resolving the racemic DL-HPG obtained by the

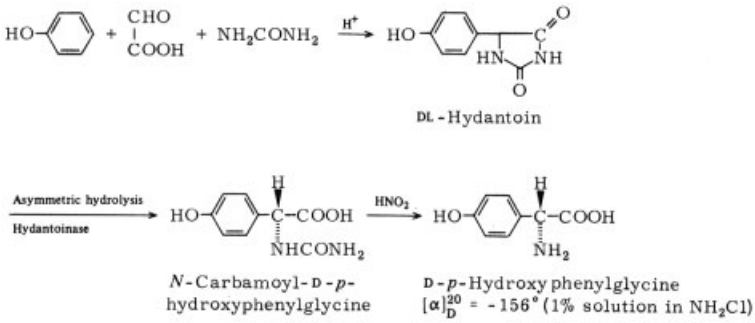

usual nonenzymatic synthesis. The optical resolution by means of fractional crystallization of the corresponding diastereomeric salt or by the predominant crystallization of the corresponding aromatic sulfonic acid salt are typical methods. However, these methods suffer disadvantages. The former requires multistep reactions and an expensive resolving agent.

The optical purity of the product is not very high. The yield obtained from one cycle of the latter resolution method is very low, and both methods require racemization and recyclization steps in the optical resolution process in order to increase the total yield of D-HPG.

The process developed by YAMADA et al. involves subjecting DL-5-(*p*-hydroxyphenyl)hydantoin (DL-HPH) to an enzymatic asymmetric hydrolysis utilizing microorganisms [221]. A new method of synthesizing DL-HPH has also been developed [222]. Thus, a high-purity D-HPG is produced industrially as shown in the sequence at the top of this page.

The substrate (DL-HPH) for the enzymatic reaction is synthesized starting with relatively inexpensive basic materials: phenol, glyoxylic acid, and urea. The enzymatic reaction hydrolyzes the hydantoin stereospecifically to afford the acid with the D configuration. The remaining substrate with the L configuration is automatically racemized under enzymatic reaction conditions. Consequently, DL-HPH can be quantitatively transformed into the desired acid, D-HPG. In this manner a very effective optical resolution is conducted kinetically. The D-*N*-carbamoyl(*p*-hydroxyphenyl)glycine thus otained is easily converted to D-HPG by treatment with nitrous acid.

As has been mentioned, the process for preparing D-HPG has been greatly simplified and provides D-HPG of very high optical purity when the enzymatic reaction is used.

This novel process has been developed industrially by Kanegafuchi Chemical Industry Company.

d-Phenylglycine is widely used in large quantity as the side chain that makes ampicillin and cephalexin orally acceptable. Industrially, D-phenylglycine has been produced by a conventional resolution method or by kinetic resolution with acylase. Another enzymatic procedure has been developed by YAMADA et al. [221], [222] as shown below.

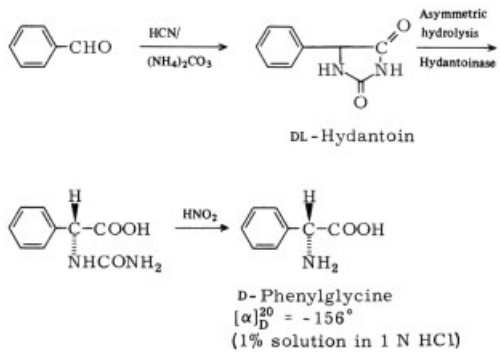

Enediynes. A class of antitumor antibiotics possessing characteristic enediynechromophores have been isolated [223]. These include neocarzinostatin (**240**) [224], [225], calicheamicins (**241**) [226], [227], esperamicins (**242**) [228], [229], dynemicin A (**243**) [230], C-1027 (**244**) [231], [232], kedarcidin (**245**) [233–236], and maduropeptin (**246**) [237].

Neocarzinostatin was isolated from *Streptomyces carzinostaticus* Var. F-41 as a complex of a chromophore and an apoprotein [224], [225].

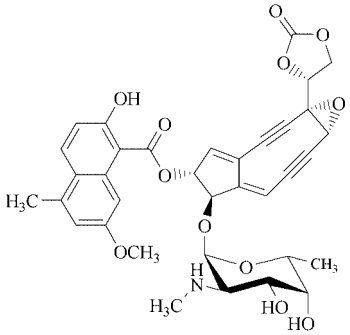

240 Neocarzinostain chromophore

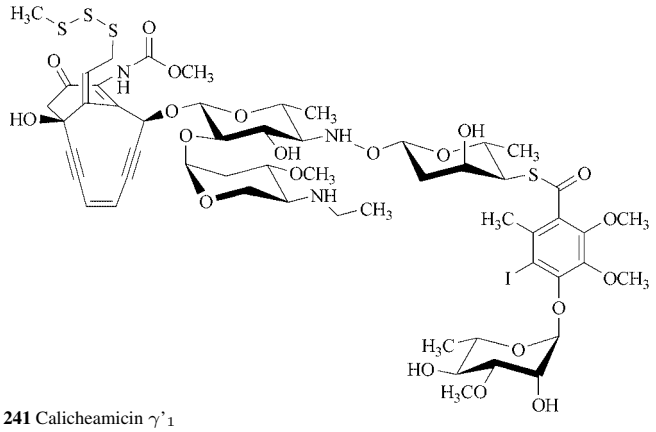

241 Calicheamicin γ'_1

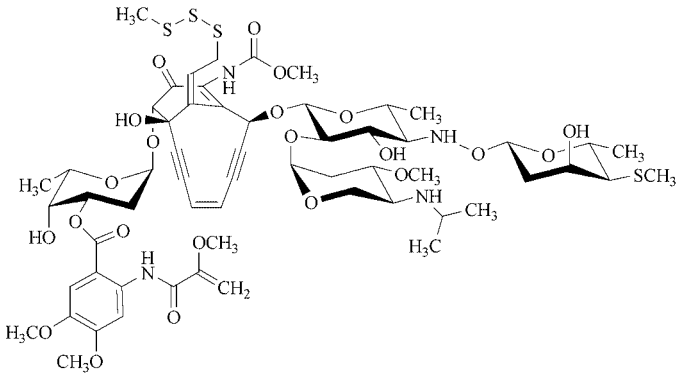

242 Esperamicin A_1

The role of the apoprotein is to transport and stabilize the chromophore that is responsible for the anticancer activity of neocarzinostatin coming mainly from its capability to cleave DNA. The DNA damage is initiated by nucleophilic attack at C12 of neocarzinostatin chromophore by a sulfur nucleophile leading to the formation of a labile cumulene intermediate that undergoes a facile cycloaromatization [238]. The resulting biradical abstracts hydrogen atom of the deoxyribose of DNA to induce degradation. A polymer conjugated derivative of neocarzinostatin was prepared and administered via the tumor-feeding artery showing increased stability in blood and the immunogenicity was much lower than the parental neocarzinostatin [239]. Similar biradical formation via the Bergman cy-

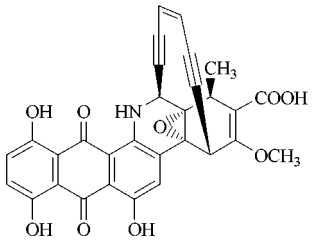

243 Dynemicin A

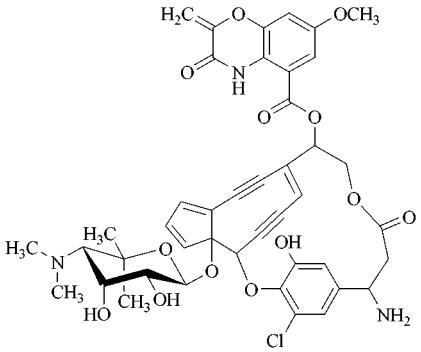

244 C-1027 chromophore

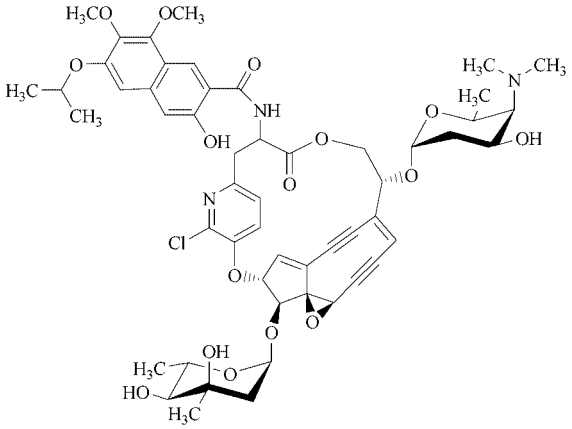

245 Kedarcidin chromophore

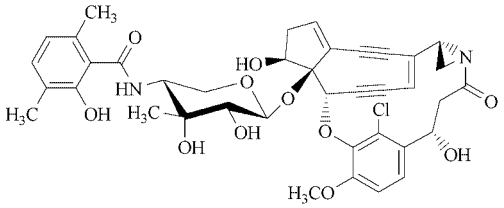

246 Maduropeptin chromophore

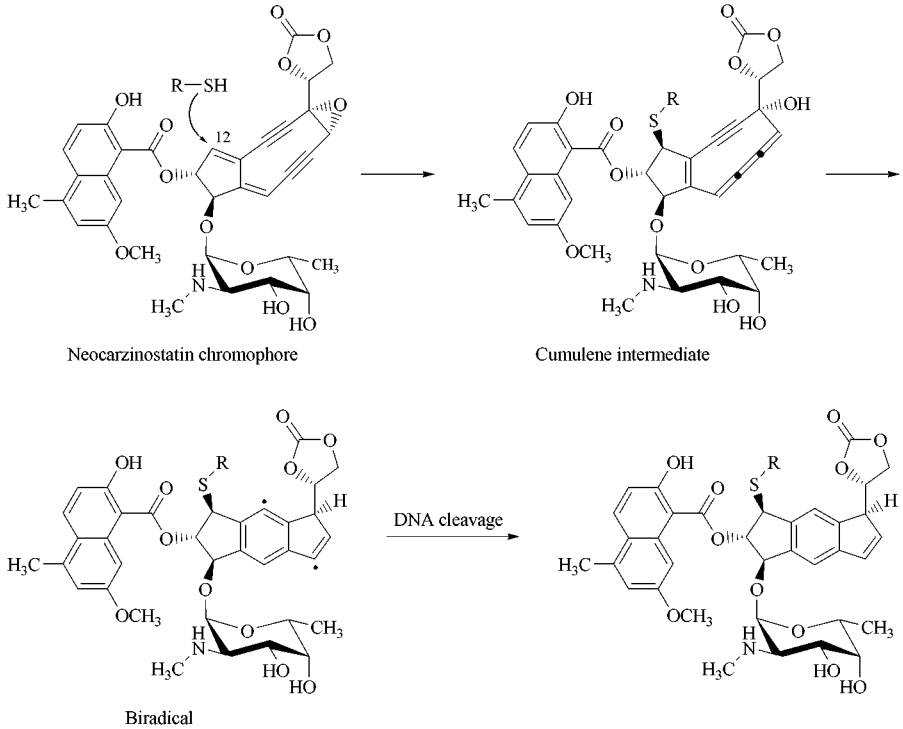

clization of the other enediyne class antibiotics has been proposed [223].

4. Individual Antibiotics Use

Benzylpenicillin potassium [*113-98-4*] (**99**), penicillin G potassium, $C_{16}H_{17}KN_2O_4S$, M_r 372.49.

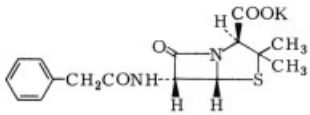

99

Benzylpenicillin potassium was the first crystalline penicillin produced on an industrial scale [22], [24]. It is produced in the pure state by the addition of phenylacetate to a culture of *Penicillium chrysogenum.* The crystalline benzylpenicillin potassium contains 1598 U/mg.

Benzylpenicillin procaine [*6130-64-9*] (**100**), penicillin G procaine, $C_{16}H_{18}N_2O_4S \cdot C_{13}H_{20}N_2O_2 \cdot H_2O$, M_r 588.73.

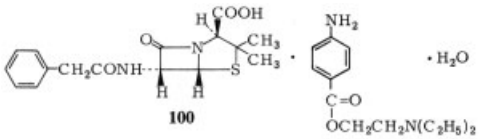

100

Benzylpenicillin procaine was developed as a dilatorily acting benzylpenicillin, only slightly soluble in water. It has been used with peanut oil or carboxymethylcellulose as an oil suspension or an aqueous suspension, respectively [240].

Benzylpenicillin benzathine [*41372-02-5*] (**101**), penicillin G benzathine, $(C_{16}H_{18}N_2O_4S)_2 \cdot C_{16}H_{20}N_2 \cdot 4H_2O$, M_r 981.21.

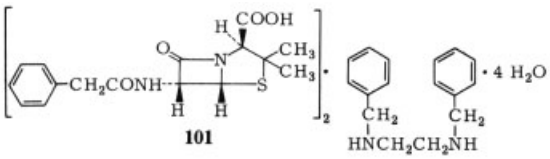

101

Benzylpenicillin benzathine was developed as a dilatorily acting benzylpenicillin that maintains an effective serum concentration for 2 days following a single intramuscular injection of 600000 U. It is also used as an orally active benzylpenicillin because it is only slightly affected by a patient's meal [241].

Phenoxymethylpenicillin [*87-08-1*], (**102**), penicillin V, $C_{16}H_{18}N_2O_5S$, M_r 350.40.

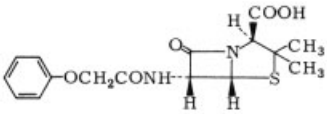

102

Phenoxymethylpenicillin was produced in the culture broth of *Penicillium chrysogenum* when phenoxyacetic acid was added to the medium at Biochemie in 1953 [242]. It is more stable against acid than benzylpenicillin and is used as an orally active penicillin. Its therapeutic applications are the same as those of benzylpenicillin.

Phenoxymethylpenicillin potassium [*132-98-9*], (**103**), penicillin V potassium, $C_{16}H_{17}KN_2O_5S$, M_r 388.49.

Phenoxymethylpenicillin potassium was first obtained at Lilly Research Laboratories in 1948,

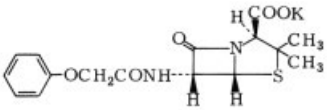

103

following the addition of phenoxyacetic acid to a culture of *Penicillium chrysogenum*. Its industrial-scale production was established in 1953 [243]. Its usefulness as an orally active penicillin was suggested by Spitzy et al. in 1955. It is less hygroscopic and much more stable against gastric acid than benzylpenicillin, and it has been used orally for therapy of infections caused by *Streptococcus, Staphylococcus,* and other gram-positive bacteria as well as *Neisseria* and *Leptospira*.

Phenethicillin potassium [*132-93-4*], (**104**), $C_{17}H_{19}KN_2O_5S$, M_r 402.52.

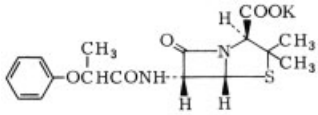

104

Phenethicillin was the first member of the semisynthetic penicillin class of antibiotics to be introduced in clinics. This drug was synthesized in 1960 by Bristol-Myers in collaboration with Beecham Research Laboratories starting with 6-aminopenicillanic acid [244]. It is as stable against gastric acid as phenoxymethylpenicillin and is used orally for therapy of gram-positive and *Neisseria* infections.

Propicillin [*551-27-9*], (**105**), $C_{18}H_{22}N_2O_5S$, M_r 378.45.

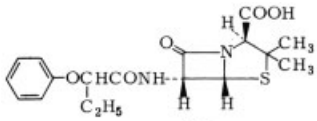

105

Propicillin was synthesized by a collaboration of Beecham Research Laboratories with Bristol-Myers Laboratories in 1961 starting with 6-aminopenicillanic acid [245]. It is stable against gastric acid and shows three- to fourfold higher serum concentrations than phenoxymethyl-penicillin when administered orally. Its antibacterial spectrum and activity are almost the same as those of phenoxymethylpenicillin. Propicillin has been used for therapy of infections caused by *Streptococcus, Staphylococcus,* and *Neisseria.*

Carbenicillin disodium [*4800-94-6*], (**106**), $C_{17}H_{16}N_2Na_2O_6S$, M_r 422.37.

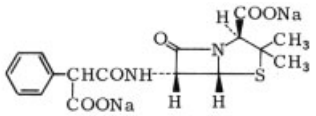

106

Carbenicillin was synthesized by Brain et al. of Beecham Research Laboratories in 1965. It was the first synthetic penicillin to show activity against *Pseudomonas aeruginosa* [246]. Although its activity against the microorganism is not strong (MIC = 25 – 100 µg/mL), it is widely

used against *P. aeruginosa* infections because of its low toxicity and the lack of other antibiotics suitable for use against this microorganism. Carbenicillin is mainly used clinically to treat urinary tract and respiratory tract infections and sepsis caused by *Proteus, Escherichia coli, Klebsiella,* and *Pseudomonas aeruginosa.*

Carbenicillin phenyl sodium [*21649-57-0*], (**107**), carfecillin, $C_{23}H_{21}N_2NaO_6S$, M_r 476.49.

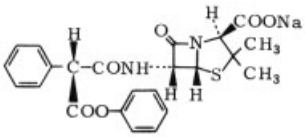

107

Carbenicillin phenyl was synthesized by Beecham Research Laboratories in 1966 as an orally active carbenicillin [247]. It is hydrolyzed to carbenicillin by intestinal esterase and thus acts the same when administered orally. The phenol produced by the hydrolysis is conjugated and excreted in urine.

Carbenicillin indanyl sodium [*26605-69-6*], (**108**), carindacillin, $C_{26}H_{25}N_2NaO_6S$, M_r 516.55.

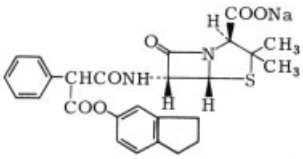

108

Carbenicillin indanyl was synthesized by Pfizer in 1972 as an orally active carbenicillin [248]. It shows strong activity against a variety of bacteria in vitro, and, when administered orally, it behaves as carbenicillin after being hydrolyzed by intestinal esterase.

Sulbenicillin [*34779-28-7*], (**109**), $C_{16}H_{18}N_2O_7S_2$, M_r 414.46.

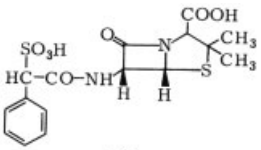

109

Sulbenicillin was synthesized by Takeda Chemical Industries in 1968 [249]. It shows slightly lower activity against gram-positive bacteria, almost the same activity against gram-negative bacteria, and slightly higher activity against anaerobic bacteria in comparison with carbenicillin. Sulbenicillin has been used by intravenous and intramuscular administration for therapy of sepsis, bacterial endocarditis, pyoderma, urinary tract and respiratory tract infections, and other infections caused by *Staphylococcus, Streptococcus, Klebsiella, Proteus, Enterobacter, Citrobacter, Escherichia coli* and *Haemophilus influenzae.*

Ticarcillin disodium [*4697-14-7*], (**110**), $C_{15}H_{14}N_2Na_2O_6S_2$, M_r 428.40.

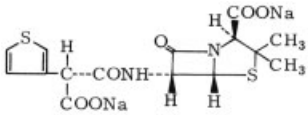

110

Ticarcillin was synthesized by Beecham Research Laboratories in 1971. The phenyl residue of carbenicillin was replaced by the 3-thienyl moiety [250]. Ticarcillin shows almost the same activity as carbenicillin against gram-positive bacteria and a twofold higher activity against gram-negative bacteria. Its in vivo activity against *Pseudomonas aeruginosa* infections in mice is fourfold higher than that of sulbenicillin. Ticarcillin is used to treat sepsis, urinary tract infections, and serious *Pseudomonas aeruginosa, Escherichia coli, Proteus,* and *Enterobacter* infections in leukemic and other cancer patients.

Methicillin sodium [*7246-14-2*], (**111**), meticillin, $C_{17}H_{19}N_2NaO_6S \cdot H_2O$, M_r 420.42.

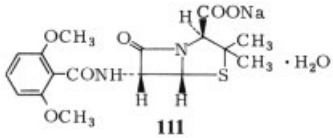

111

Methicillin was synthesized independently by Bristol-Myers Laboratories and Beecham Research Laboratories in 1960 [251]. It was the

first member of the penicillinase-stable semisynthetic penicillin class of antibiotics to be introduced clinically. This antibiotic is a parenteral penicillin having an antibacterial spectrum similar to that of benzylpenicillin. Although its activity against benzylpenicillin-sensitive bacteria is about 1/3 to 1/50 that of benzylpenicillin, it shows strong activity against benzylpenicillin-resistant strains because of its stability toward penicillinase. Methicillin has been used for therapy of respiratory tract and urinary tract infections, sepsis, and gynecological and other infections caused by benzylpenicillin-resistant bacteria.

Oxacillin sodium [*7240-38-2*], (**112**), $C_{19}H_{18}N_3NaO_5S \cdot H_2O$, M_r 441.43.

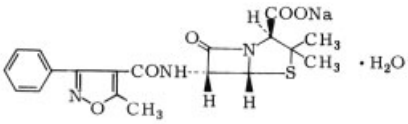

112

Oxacillin was synthesized by Bristol-Myers Laboratories in 1961 starting with 6-aminopenicillanic acid [252]. It was the first orally active and penicillinase-stable semisynthetic penicillin to be introduced clinically. Oxacillin is slightly less stable against gastric acid and shows a lower serum concentration than phenoxymethylpenicillin, but it is highly active against phenoxymethylpenicillin-resistant *Staphylococcus aureus.* Oxacillin has been used by oral and intramuscular administration for therapy of respiratory tract, urinary tract, gynecological, and other infections caused by benzylpenicillin-resistant bacteria.

Cloxacillin sodium [*7081-44-9*], (**113**), $C_{19}H_{17}ClN_3NaO_5S \cdot H_2O$, M_r 475.89.

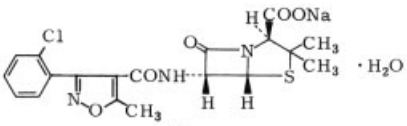

113

Cloxacillin was synthesized by Beecham Research Laboratories in 1962 starting with 6-aminopenicillanic acid [253]. It was the first semisynthetic penicillin with a halogen atom in the side chain to be used clinically. It shows 5- to 20-fold stronger activity than methicillin and is twice as active as oxacillin against *Staphylococcus aureus*, including benzylpenicillin-resistant strains. Cloxacillin is highly stable against penicillinase and well absorbed by oral administration. It has been widely used, by oral and intramuscular administration, to treat internal, surgical, gynecological, and other infections caused by both benzylpenicillin-sensitive and benzylpenicillin-resistant bacteria.

Dicloxacillin [*3116-76-5*], (**114**), $C_{19}H_{17}Cl_2N_3O_5S$, M_r 470.33.

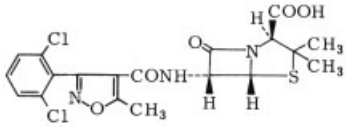

114

Dicloxacillin was synthesized by Bayer in 1965 starting with 6-aminopenicillanic acid [254]. It is a penicillinase-stable and orally active semisynthetic penicillin and shows higher and longer serum concentrations than cloxacillin when administered orally. Dicloxacillin is used orally, either alone or in combination with ampicillin, to treat various infections, including those caused by benzylpenicillin-resistant bacteria.

Flucloxacillin [*5250-39-5*], (**115**), floxacillin, $C_{19}H_{17}ClFN_3O_5S$, M_r 453.88.

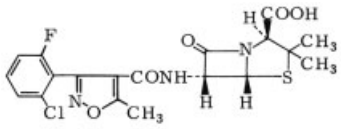

115

Flucloxacillin was synthesized by Beecham Research Laboratories in 1962 as a penicillinase-stable and orally active semisynthetic penicillin [255]. It shows almost the same activity as dicloxacillin, and it has slightly higher serum and tissue concentrations than dicloxacillin. This drug has been used to treat pyoderma, sepsis, and postoperative infections as well as ear and nose, respiratory tract, and other infections caused by *Staphylococcus* and *Streptococcus*, including benzylpenicillin-resistant strains.

Ampicillin [*69-53-4*], (**116**), aminobenzyl penicillin, $C_{16}H_{19}N_3O_4S$, M_r 349.41.

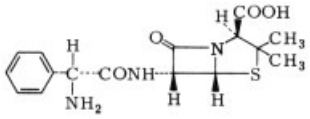

116

Ampicillin was synthesized by Beecham Research Laboratories in 1961 and evaluated for its anti-gram-negative activity by ROLINSON et al. [256]. It was the first semisynthetic penicillin showing strong activity against gram-negative bacilli. Although it is hydrolyzed by bacterial penicillinase, it is active against *Escherichia coli, Shigella, Proteus mirabilis,* and *Haemophilus influenzae* and is used very widely as an oral antibiotic. Ampicillin also is used by injection for serious infections.

Mezlocillin [*51481-65-3*], (**117**), $C_{21}H_{25}N_5O_8S_2$, M_r 539.59.

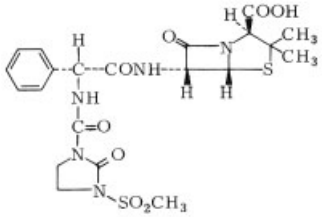

117

Mezlocillin was synthesized by Bayer in 1974 [257]. The NH_2 residue of ampicillin was acylated. Mezlocillin is a member of the so-called ureidopenicillins and shows two- to eightfold greater activity against *Citrobacter, Enterobacter, Klebsiella, Escherichia coli,* and *Haemophilus influenzae* than ampicillin, carbenicillin, or sulbenicillin. Its activity against gram-positive bacteria is almost the same as that of carbenicillin. Mezlocillin is used by intravenous administration for therapy of sepsis, meningitis, and respiratory tract, urinary tract, and abdominal infections.

Piperacillin sodium [*59703-84-3*], (**118**), $C_{23}H_{26}N_5NaO_7S$, M_r 539.55.

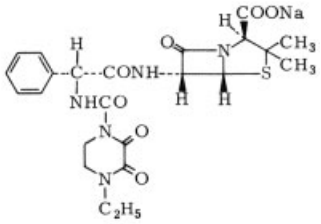

118

Piperacillin was synthesized by Toyama Chemicals Company in 1976 by acylation of the amino residue of ampicillin [258]. It shows excellent activity against a wide range of gram-positive and gram-negative bacteria and has been used as the most potent semisynthetic penicillin. Piperacillin is administered intravenously or intramuscularly for therapy of sepsis, meningitis, respiratory and urinary tract infections, and for abdominal infections caused by *Staphylococcus, Streptococcus, Escherichia coli, Klebsiella, Haemophilus influenzae, Serratia marcescens, Pseudomonas aeruginosa, Proteus, Enterobacter, Citrobacter,* and *Bacteroides.*

Amoxicillin [*61336-70-7*], (**119**), amoxycillin, $C_{16}H_{19}N_3O_5S \cdot 3\,H_2O$, M_r 419.46.

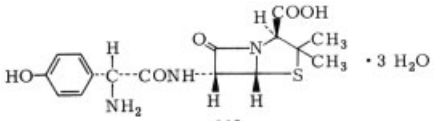

119

Amoxicillin was synthesized by Beecham Research Laboratories in 1968 [259]. A hydroxyl group was introduced on the benzene ring of ampicillin. Amoxicillin shows about a twofold higher serum concentration than ampicillin when administered orally. It was shown by double-blind comparative studies with ampicillin that amoxicillin was as effective as ampicillin when administered at half the dose of ampicillin.

Ciclacillin [*3485-14-1*], (**120**), cyclacillin, $C_{15}H_{23}N_3O_4S$, M_r 341.43.

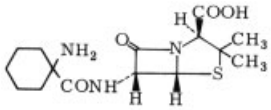

120

Ciclacillin was synthesized by Wyeth Laboratories in 1967 in the course of studies on the improvement of oral absorption of ampicillin [260]. Although its antibacterial activity is one-sixteenth to one-half that of ampicillin, it shows a four to tenfold higher oral absorption and higher urinary excretion. Ciclacillin shows less tendency than ampicillin to cause diarrhea and is used for therapy of pyoderma, wound infection, respiratory and urinary tract infections, as well as ear and nose, and other infections caused by *Staphylococcus, Streptococcus, Escherichia coli, Citrobacter, Klebsiella, Proteus,* and *Haemophilus influenzae.*

Hetacillin [*3511-16-8*], (**121**), $C_{19}H_{23}N_3O_4S$, M_r 389.48.

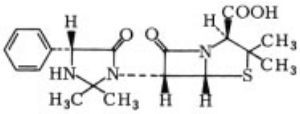

121

Hetacillin was synthesized by Bristol-Myers Laboratories in 1964. It shows activity as strong as that of ampicillin against a variety of gram-positive and gram-negative bacteria [261]. Hetacillin shows almost the same pharmacokinetic properties as ampicillin when administered orally and a later serum peak when given by intramuscular administration. When administered either orally or intramuscularly, it is hydrolyzed in vivo into ampicillin. Therefore, the clinical applications of hetacillin are the same as those of ampicillin.

Sulbactam sodium [*69388-84-7*], (**122**), $C_8H_{10}NNaO_5S$, M_r 255.22.

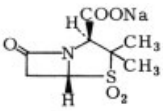

122

Sulbactam was synthesized by Pfizer Research Laboratories in 1977 in the course of screening for β-lactamase inhibitors [262]. It shows strong activity against penicillinase and moderate activity against cephalosporinase. Sulbactam itself shows activity against some gram-negative bacteria but no activity against most pathogenic bacteria. The use of sulbactam in combination with cefoperazone, which is partially hydrolyzed by penicillinase, is under study along with its use as an esterified complex with ampicillin (sultamicillin) for therapy of cefoperazone-ampicillin-resistant infections.

Talampicillin hydrochloride [*39878-70-1*], (**123**), ampicillin phthalidyl, $C_{24}H_{23}N_3O_6S \cdot HCl$, M_r 517.99.

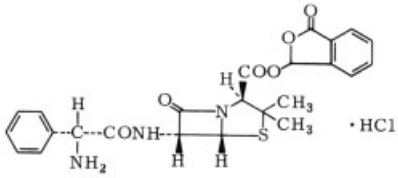

123

Talampicillin was synthesized by Yamanouchi Pharmaceutical Co. in 1971 by esterifying the carboxylic acid group of ampicillin in order to improve the oral absorption [263]. When administered orally, it is hydrolyzed to ampicillin by intestinal esterase. Its bioavailability is two or more times as high as that of ampicillin, and it is used to treat the same infections as those for which ampicillin is used orally but in doses only half or one-third as large.

Bacampicillin hydrochloride [*37661-08-8*], (**124**), ampicillin ethoxycarbonyloxyethyl hydrochloride, $C_{21}H_{27}N_3O_7S \cdot HCl$, M_r 501.99.

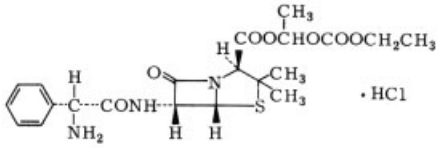

124

Bacampicillin was synthesized by Ekström et al. of Astra Läkemedel in 1975 [264]. It is more stable against acid than ampicillin and more rapidly absorbed orally; its absorption is less affected by a patient's most recent meal than is the case for ampicillin. Bacampicillin is hydrolyzed by intestinal esterase after oral administration and then behaves the same as ampicillin. In double-blind comparison studies, bacampicillin was shown to be as effective as ampicillin when administered at half the dose.

Pivmecillinam [*32886-97-8*], (**125**), amdinocillin pivoxil, $C_{21}H_{33}N_3O_5S$, M_r 439.58.

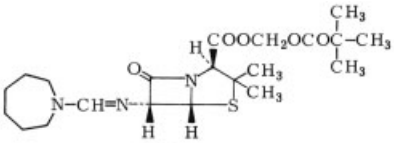

125

Pivmecillinam was found by a series of studies at Leo in 1969 [265]. It is a derivative of 6β-formamidinopenicillanic acid and shows strong activity against gram-negative bacilli. Pivmecillinam is hydrolyzed by intestinal esterase and acts similarly to mecillinam after oral administration. Mecillinam shows weaker activity than ampicillin against gram-positive cocci but much stronger activity against a wide range of gram-negative bacilli [266].

Cephalexin [*23325-78-2*], (**126**), cefalexin, $C_{16}H_{17}N_3O_4S \cdot H_2O$, M_r 365.41.

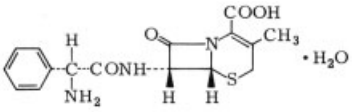

126

Cephalexin was first synthesized in 1967 by Glaxo Research Laboratories [267] and first produced on an industrial scale by Eli Lilly & Co. in the same year. It is a deacetoxylated derivative of cephaloglycin that is not metabolized in vivo. When administered orally, it shows a much higher serum concentration and much lower tendency to induce diarrhea than cephaloglycin. Cephalexin has been used widely and is the most popular orally active antibiotic in the world for treatment of respiratory tract, urinary tract, surgical, ear and nose, and other infections caused by *Staphylococcus, Streptococcus, Escherichia coli, Klebsiella, Enterobacter,* and *Proteus.*

Cefaclor [*70356-03-5*], (**127**), $C_{15}H_{14}ClN_3O_4S \cdot H_2O$, M_r 385.83.

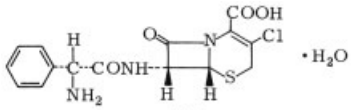

127

Cefaclor was found by Eli Lilly & Co. in 1976 in the course of studies on the improvement of synthetic procedures for cephalexin [268]. The intermediates with chlorine at the 3 position of the cephem nucleus have excellent antibacterial activity. Cefaclor was selected from among the various derivatives that differed at the 7 position because of its strong activity and high bioavailability following oral administration.

Cephaloglycin [*22202-75-1*], (**128**), cefaloglycin, $C_{18}H_{19}N_3O_6S \cdot 2\,H_2O$, M_r 441.46.

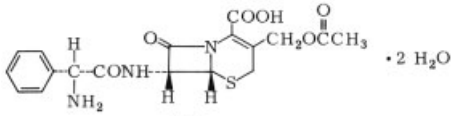

128

Cephaloglycin was synthesized by Lilly Research Laboratories in 1966 as the first member of the orally active cephalosporin C class of antibiotics [269]. Its activity against *Staphylococcus* and *Streptococcus* is one-fourth to one-half that of cephaloridin, a parenteral antibiotic, but against gram-negative bacteria it is almost the equal of cephaloridin. Cephaloglycin is readily absorbed by oral administration but is partially metabolized in vivo into deacetylcephaloglycin, which shows one-tenth to one-half as much activity. This antibiotic has been used in therapy of urinary tract infections and pyoderma, but it is being replaced by the newer orally active cephalosporins. Cephaloglycin often causes diarrhea, especially in children, and pediatric use is not allowed.

Cefadroxil [*66592-87-8*], (**129**), $C_{16}H_{17}N_3O_5S \cdot H_2O$, M_r 381.41.

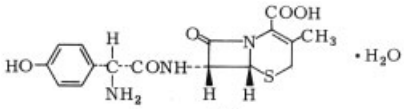

129

Cefadroxil was found by Bristol-Myers Co. in 1976 [270]. A hydroxyl group was attached to the benzene ring of cephalexin. Like amoxicillin and ampicillin, cefadroxil shows almost the same antibacterial activity spectrum as cephalexin and superior oral absorption. Its in vivo activity is four to six times greater than that of cephalexin, and its half-life in serum is twice as long.

Cephradine [*38821-53-3*], (**130**), cefradine, $C_{16}H_{19}N_3O_4S$, M_r 349.41.

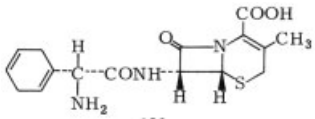

130

Cephradine was synthesized by the Squibb Institute of Medical Research in 1971 [271]. It shows almost the same antibacterial activity and pharmacokinetic properties as cephalexin. Cephradine has been used for therapy of urinary and respiratory tract infections caused by *Staphylococcus, Streptococcus, Escherichia coli, Klebsiella,* and *Proteus mirabilis.*

Cefroxadine [*51762-05-1*], (**131**), $C_{16}H_{19}N_3O_5S$, M_r 365.41.

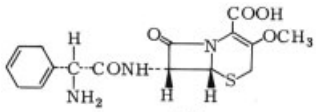

131

Cefroxadine was synthesized by Ciba-Geigy in 1972. A methoxyl group replaced the methyl group of cephradine at the 3 position of the cephem nucleus [272]. Cefroxadine shows stronger activities than cephalexin, especially bactericidal and bacteriolytic activities, and it has better oral absorption that is less affected by a recent meal. Cefroxadine shows less renal toxicity than cephalexin in toxicological studies using animals.

Cephapirin sodium [*24356-60-3*], (**132**), cefapirin, $C_{17}H_{16}N_3NaO_6S_2$, M_r 445.45.

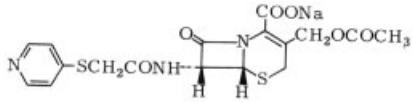

132

Cephapirin was synthesized by Bristol-Myers Laboratories in 1970 [273]. It shows almost the same in vitro antibacterial activity as cephalothin, but its in vivo effects are slightly greater than those of cephalothin. Like cephalothin, it is metabolized in vivo, and its deacetylated metabolite shows almost the same activity against gram-positive bacteria as cephalothin, but weaker activity against gram-negative bacteria. Cephapirin has been used for therapy of urinary tract infections and osteomyelitis caused by *Staphylococcus, Streptococcus,* and *Escherichia coli.*

Cephalothin sodium [*58-71-9*], (**133**), $C_{16}H_{15}N_2NaO_6S_2$, M_r 418.43.

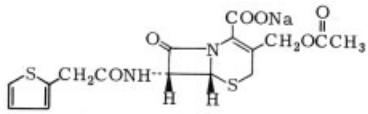

133

Cephalothin, along with cephaloridine, was the first of the synthetic cephalosporin C class antibiotics to be introduced clinically. It was synthesized from 7-amino-cephalosporanic acid by Lilly Research Laboratories in 1962 [274]. Cephalothin shows strong activity against gram-positive and gram-negative bacteria and *Leptospira,* including benzylpenicillin-resistant strains. It has been used intravenously and intramuscularly to treat a variety of infections caused by *Staphylococcus, Streptococcus, Escherichia coli,* and *Neisseria*. The drug is metabolized in vivo, and the metabolite, deacetylcephalothin, is almost inactive.

Cephacetrile sodium [*23239-41-0*], (**134**), cefacetrile, $C_{13}H_{12}N_3NaO_6S$, M_r 361.31.

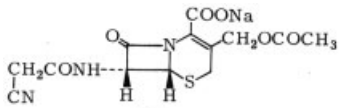

134

Cephacetrile was synthesized by Ciba-Geigy in 1970 [275]. It shows almost the same activity against gram-negative bacteria as cephalothin, but it has a higher activity against β-lactamase-producing *Escherichia coli.* Like cephalothin, it is metabolized partially by deacetylation in vivo. Cephacetrile has been used to treat sepsis and abdominal and respiratory tract infections caused by *Staphylococcus, Streptococcus, Escherichia coli,* and *Klebsiella pneumoniae,* but it is gradually being replaced clinically by the newer and more active cephalosporins.

Cefsulodin sodium [*52152-93-9*], (**135**), $C_{22}H_{19}N_4NaO_8S_2$, M_r 554.54.

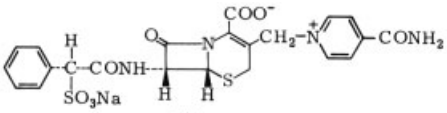

135

Cefsulodin was synthesized by Takeda Chemicals Industries in 1974 by introducing the sulfobenzyl group, the same moiety as in sulbenicillin, at the 7 position of the cephem nucleus [276]. Its side chain at the 3 position is similar to that of cephaloridine except for the carbamoyl group. The introduction of these hydrophilic groups increases the activity against *Pseudomonas aeruginosa*, but it markedly decreases it against gram-positive and other gram-negative bacteria. Therefore cefsulodin is used as a specific antibiotic against infections caused by the opportunistic pathogen *P. aeruginosa.*

Cephaloridine [*50-59-9*], (**136**), cefaloridine, $C_{19}H_{17}N_3O_4S_2$, M_r 415.49.

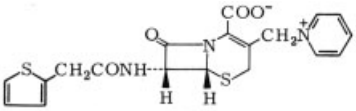

136

Cephaloridine, along with cephalothin, was the first member of the synthetic cephalosporin C class of antibiotic, to be introduced clinically. It was synthesized, starting with 7-aminocephalosporanic acid, by Glaxo Research Laboratories in 1962 [277]. This drug shows strong activity against gram-positive and gram-negative bacteria and *Leptospira*, including benzylpenicillin-resistant strains. It has been given widely by intravenous, intramuscular, and intraspinal injection to treat a variety of infections caused by *Staphylococcus, Streptococcus, Neisseria, Klebsiella, Escherichia coli,* and other pathogens. Because of its renal toxicity and the development of newer and more active synthetic cephalosporins, its use is declining.

Cefatrizine [*51627-14-6*], (**137**), $C_{18}H_{18}N_6O_5S_2$, M_r 462.51.

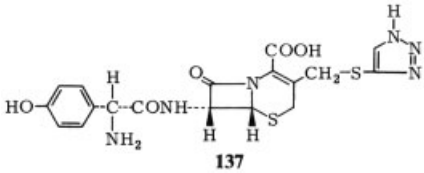

137

Cefatrizine was synthesized by Bristol-Myers Co. and Smith Klein & French Laboratories in 1974 [278]. It shows two to four times higher activity against gram-positive and four to eight times higher activity against gram-negative bacteria than cephalexin. Cefatrizine also shows excellent oral absorption, and its in vivo activity is 30 to 500 times higher than that of cephalexin.

Cefoperazone sodium [*62893-20-3*], (**138**), $C_{25}H_{26}N_9NaO_8S_2$, M_r 667.66.

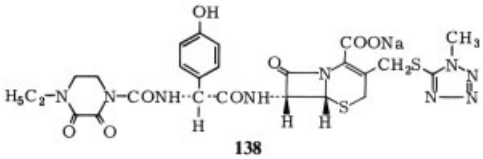

138

Cefoperazone was synthesized by Toyama Chemicals Co. in 1978. Except for the hydroxyl group, the side chain attached to the cephem nucleus is the same as that of piperacillin [279]. Cefoperazone shows excellent activity against gram-positive (except *Staphylococcus*) and gram- negative bacteria, including *Pseudomonas aeruginosa.* Its pharmacological characteristics are unique. Cefoperazone is excreted mainly in bile, and a concentration five to tenfold higher in bile than in serum is obtained. The transfer into cerebrospinal fluid is 10 – 30 % of the serum concentration; the half-life in serum is 2.0 – 2.6 h, and the degree of binding with serum protein is as high as 86.6 %.

Cefamandole [*34444-01-4*], (**139**), $C_{18}H_{18}N_6O_5S_2$, M_r 462.51.

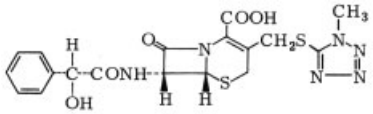

139

Cefamandole was synthesized by Eli Lilly & Co. in 1972 [280]. It shows strong activity against *Proteus* (indole-positive) species, *Enterobacter*, and *Citrobacter,* against which the earlier cephalosporins, such as cephalothin and cefazolin, are inactive. The nafate (sodium salt of the *O*-formyl ester) has been used in the United States and Europe, and the sodium salt of cefamandole has been used in Japan by injection.

Cefotiam hydrochloride [*66309-69-1*], (**140**), $C_{18}H_{23}N_9O_4S_3 \cdot 2\,HCl$, M_r 598.56.

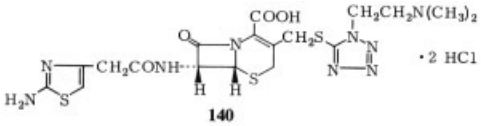

140

Cefotiam was synthesized by Takeda Chemicals Industries in 1977 as the first cephem derivative to introduce the aminothiazol group in the 7 side chain and the alkylated tetrazole group at the 3 side chain [281]. This drug shows greater antibacterial activity against gram-negative bacteria than the earlier cephalosporins, including *Enterobacter, Citrobacter,* and indole-positive *Proteus,* and it retains the same activity as they have against gram-positive organisms. Its excellent activity against *Klebsiella* and *Escherichia coli* is based on its ability to penetrate the cell. This is 3- to 70-fold greater than that of cephalothin or cefazolin.

Cefazolin sodium [*27164-46-1*], (**141**), $C_{14}H_{13}N_8NaO_4S_3$, M_r 476.50.

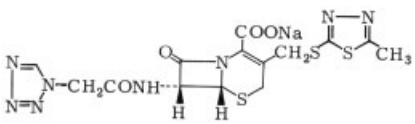

141

Cefazolin was synthesized by Fujisawa Pharmaceutical Co. in 1969 [282]. It was the first of the cephem antibiotics to introduce a thiadiazolylthiomethyl group at the 3 position and a tetrazole group at the 7 position in the side chain. Cefazolin is a parenteral cephem antibiotic showing better activity against gram-negative bacteria than cephalothin or cephaloridine. Its bactericidal activity, tissue distribution, and urinary excretion are excellent and it has wide clinical use.

Ceftizoxime sodium [*68401-82-1*], (**142**), $C_{13}H_{12}N_5NaO_5S_2$, M_r 405.39.

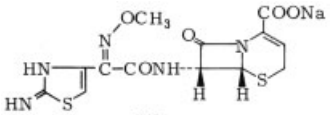

142

Ceftizoxime was synthesized by Fujisawa Pharmaceutical Industries in 1979 [283]. It possesses the (iminothiazolyl)methoxyiminomethyl group at the 7 position of the cephem nucleus, but there is no side chain at the 3 position. The compound shows excellent activity against gram-positive and gram-negative bacteria, behavior similar to that of cefotaxime. Unlike cefotaxime, however, it is not metabolized in vivo.

Cefotaxime sodium [*64485-93-4*], (**143**), $C_{16}H_{16}N_5NaO_7S_2$, M_r 477.45.

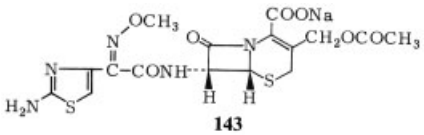

143

Cefotaxime was synthesized by Hoechst and Roussel-Uclaf in 1977 [284]. It was the first derivative of cephalosporin to introduce the methoxyimino and aminothiazole groups at the 7 position of the cephem nucleus. Although it shows unexpectedly low oral absorption, its excellent activity against a wide range of gram-positive and gram-negative organisms, including *Serratia, Enterobacter, Citrobacter,* and anaerobes, guided research and development of the newer synthetic cephems, the so-called third-generation cephalosporins.

Cefmenoxime hydrochloride [*75738-58-8*], (**144**), $(C_{16}H_{17}N_9O_5S_3)_2 \cdot HCl$, M_r 1059.59.

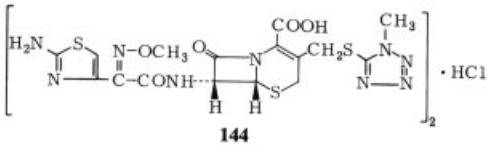

144

Cefmenoxime was synthesized by Takeda Pharmaceutical Co. in 1978 [285]. It contains aminothiazole and methoxyimino groups in the 7-acyl position and a methyltetrazole group at the 3 position of the cephem nucleus. Cefmenoxime has a wide activity spectrum against gram-positive and gram-negative bacteria, including anaerobes, except *Staphylococcus aureus.* Cefmenoxime is a member of the so-called third-generation cephalosporins, a recently introduced class of clinically important antibiotics.

Cefuroxime [*55268-75-2*], (**145**), $C_{16}H_{16}N_4O_8S$, M_r 424.39.

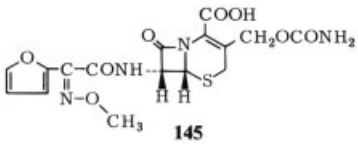

145

Cefuroxime was synthesized by Glaxo Laboratories in 1975 as the first cephem antibiotic with the methoxyimino group at the 7 position [286]. It is highly resistant to hydrolysis by cephalosporinase and is active against a variety of gram-negative bacteria, including indole-positive *Proteus, Enterobacter,* and *Citrobacter.* Cefuroxime is considered to be one of the so-called second-generation cephalosporins.

Ceftriaxone sodium [*74578-69-1*] (**146**), $C_{18}H_{16}N_8Na_2O_7S_3$, M_r 598.55.

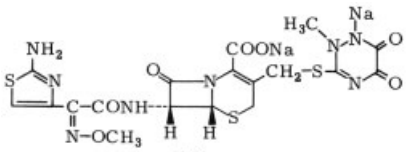

146

Ceftriaxone was synthesized by Hoffmann-La Roche in 1981 [287]. The triazinyl moiety was introduced at the 3 position of the cephem nucleus. The same side chain as possessed by cefotaxime and the other so-called third-generation cephalosporins was retained at the 7 position. The antibacterial activity of ceftiaxone is almost the same as that of cefotaxime in vitro, but its in vivo activity is 10 to 100 times higher. Its most characteristic property is its seven to eight hour half-life in serum, the longest among the known cephem antibiotics.

Ceftazidime [*72558-82-8*], (**147**), $C_{22}H_{22}N_6O_7S_2$, M_r 546.58.

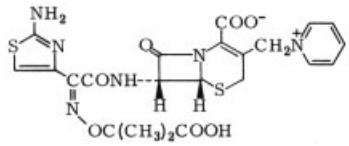

147

Ceftazidime was synthesized by Glaxo Laboratories in 1979 [288]. A new side chain, a carboxypropyloxyimino group with an aminothiazolyl group, was attached at the 7 position of the cephem nucleus. The compound shows excellent activity against *Pseudomonas aeruginosa,* indole-positive *Proteus* spp., *Enterobacter cloacae,* and *Serratia,* which are called opportunistic pathogens. Its activity against clinically isolated *P. aeruginosa* strains was found to be twofold higher than that of cefsulodin, a specific antibiotic widely used against that microorganism. Unlike cefsulodin, it shows balanced activity against other gram-negative bacteria and *Streptococcus.*

Cefoxitin [*35607-66-0*], (**148**), $C_{16}H_{17}N_3O_7S_2$, M_r 427.46.

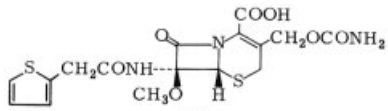

148

Cefoxitin was synthesized by Merck Sharp & Dohme Research Laboratories in 1972 as the first clinically applicable cephamycin; it was derived from a cephamycin component obtained from *Streptomyces lactamdurans* [289]. The drug is slightly hydrolyzed by β-lactamases

and shows great activity against gram-negative bacteria and anaerobes.

Cefmetazole [*56796-20-4*], (**149**), $C_{15}H_{17}N_7O_5S_3$, M_r 471.54.

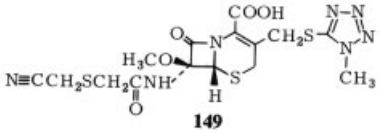

149

Cefmetazole was synthesized by Sankyo Co. in 1976 starting with a biosynthetic cephamycin [290]. It shows excellent activity against *Serratia* and *Proteus*, against which cefazolin is not active, and it has stronger activity than cefoxitin, another derivative of cephamycin. Cefmetazole is active against anaerobes and resistant to β-lactamase, but it is not active against *Pseudomonas aeruginosa.*

Cefotetan [*69712-56-7*] (**150**), $C_{17}H_{17}N_7O_8S_4$, M_r 575.62.

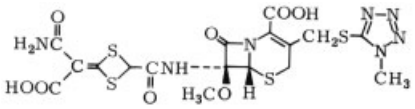

150

Cefotetan was synthesized by Yamanouchi Pharmaceutical Co. in 1979 starting with oganomycin, a cephamycin, produced by *Streptomyces oganoensis* YG19Z [291]. Its 7β side chain, 1,3-dithietan, is unique and contributes greatly to its strong activity against gram-negative bacteria, including *Serratia, Citrobacter, Enterobacter,* indole-positive *Proteus,* and anaerobes. Its half-life in the serum is as long as three hours, and about 90 % of it is excreted in the urine.

Latamoxef [*64953-12-4*], (**151**), disodium moxalactam, $C_{20}H_{18}N_6Na_2O_9S$, M_r 564.44.

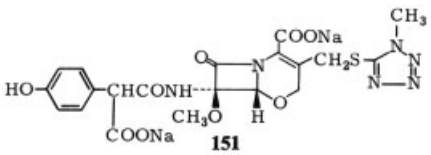

151

Latamoxef was synthesized by Shionogi Pharmaceuticals in 1975 starting with benzylpenicillin and using a novel drug design [292]. The oxacephem nucleus, in which the sulfur atom had been replaced by oxygen, was substituted with a methoxyl group at the 7α position, as in the cephamycins. A carboxyl moiety and a hydroxybenzyl group were added at the 7β position, as in carbenicillin, and a methyltetrazolylthiomethyl group was attached at the 3 position. These substitutions resulted in a strong activity against gram-negative bacteria and a high resistance to the action of β-lactamase, along with excellent activity against *Pseudomonas aeruginosa,* even though the compound has no activity against *Staphylococcus aureus.*

Clavulanic acid [*58001-44-8*], (**152**), $C_8H_9NO_5$, M_r 199.16.

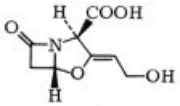

152

Clavulanic acid was found in the culture broth of *Streptomyces clavuligerus* by Beecham Research Laboratories in 1976 [293]. It was the first β-lactamase inhibitor. This antibiotic shows weak antibacterial activity against gram-positive and gram-negative organisms but strong inhibitory activity against the β-lactamase produced by ampicillin-resistant bacteria. Clavulanic acid is used orally in combination with amoxicillin and with ticarcillin by injection to enhance the activities of these antibiotics against resistant infections.

Imipenem [*64221-86-9*], (**153**), *N*-formimidoylthienamycin.

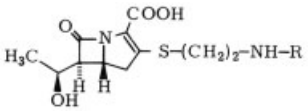

Thienamycin: R = H
$C_{11}H_{16}N_2O_4S$, M_r 272.33
N-Formimidoylthienamycin: R = -CH=NH
$C_{12}H_{17}N_3O_4S$, M_r 299.35

153

Thienamycin was found in the culture broth of *Streptomyces cattleya* by Merck Sharp & Dohme in 1976, as a very unstable substance. It has a unique carbapenem structure, like that of the olivanic acids found in *S. olivaceus* by Beecham Research Laboratories in 1979 [294]. Thienamycin shows excellent activity against a variety of pathogenic bacteria, including *Pseudomonas aeruginosa.* Its chemical stability has been improved by derivatization with the formimidoyl group, and its biological stability has been improved by combining it with cilastatin, an inhibitor of kidney dihydropeptidase. The combination drug imipenem – cilastatin is now under study to evaluate its clinical efficacy and safety.

Aztreonam [*78110-38-0*], (**154**), azthreonam, $C_{13}H_{17}N_5O_8S_2$, M_r 435.44.

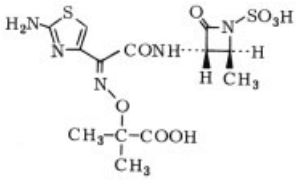

154

Aztreonam was synthesized by the Squibb Institute for Medical Research in 1981 starting with L-threonine. The synthesis was based on findings about bacterial β-lactam compounds of a monocyclic nature [295]. The β-lactam compounds, called monobactams, were isolated from *Chromobacterium violaceum, Agrobacterium radiobacter,* etc. Such monocyclic β-lactams of bacterial origin had previously been found independently in 1981 by Takeda Chemicals Industries in the culture broths of *Pseudomonas acidophila* and *P. mesoacidophila* and named sulfazecin and isosulfazecin, respectively [82] (see page 16). Aztreonam was selected from among hundreds of derivatives as a candidate for clinical trials because of its unique antibacterial spectrum and strong activity. This antibiotic shows excellent activity against a variety of gram-negative aerobic bacteria but no activity against gram-positive bacteria or anaerobes. Its efficacy and safety are now being clinically evaluated.

Tetracycline [*60-54-8*], (**155**), $C_{22}H_{24}N_2O_8$, M_r 444.45.

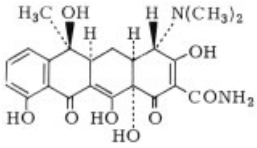

155

Tetracycline was first obtained from chlortetracycline by reductive dehalogenation at Lederle Laboratories and Pfizer in 1953 [296]. It was also obtained either by fermentation of the chlortetracycline-producing organism, *Streptomyces aureofaciens,* under conditions of chlorine limitation, or by fermentation of a mutant of the organism lacking the chlorinating enzyme. Tetracycline is more stable than chlortetracycline in aqueous solution. Its antimicrobial activity is the same as that of chlortetracycline and oxytetracycline, but its serum concentration after oral administration is considerably higher and more enduring.

Chlortetracycline hydrochloride [*64-72-2*], (**156**), $C_{22}H_{23}ClN_2O_8 \cdot HCl$, M_r 515.35.

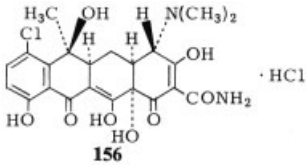

156

Chlortetracycline was found in the culture broth of *Streptomyces aureofaciens* by DUGGAR et al. of Lederle Laboratories in 1948 and named aureomycin [85]. It shows a wider range of antibiotic activity than the earlier antibiotics, penicillins, and streptomycins and as great as that of chloramphenicol. Its activity covers gram-positive and gram-negative bacteria as well as *Rickettsia* and *Chlamydiae.* Chlortetracycline has been replaced by other tetracyclines in clinical use and is used now used as a feed additive to promote the growth of livestock.

Demethylchlortetracycline [*127-33-3*], (**157**), demeclocycline, $C_{21}H_{21}ClN_2O_8$, M_r 464.86.

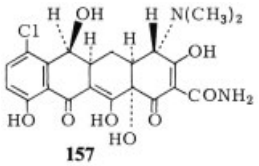

157

Demethylchlortetracycline was isolated from the culture broth of a mutant of *Streptomyces aureofaciens,* the chlortetracycline-producing strain, by Lederle Research Laboratories in 1957 [297]. It shows one and one-half to two times as much in vitro antimicrobial activity and in vivo protective effect as tetracycline. Its base and hydrochloride have been used orally and by topical application to treat infections caused by *Staphylococcus, Streptococcus, Rickettsia, Chlamydiae, Neisseria, Klebsiella, Proteus, Escherichia coli,* and *Haemophilus influenzae.*

Oxytetracycline [*6153-64-6*], (**158**), $C_{22}H_{24}N_2O_9 \cdot 2\,H_2O$, M_r 496.48.

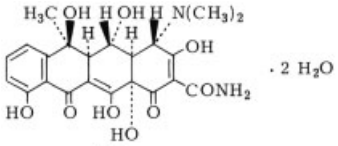

158

Oxytetracycline was found in the culture broth of *Streptomyces rimosus* by Pfizer in 1950 and named terramycin [298]. It is closely related to chlortetracycline (aureomycin), and its structure was elucidated in 1952. Oxytetracycline has had wide clinical use as a substitute for chlortetracycline, even after the introduction of tetracycline. Oxytetracycline is administered orally, topically, and parenterally.

Methacycline [*914-00-1*], (**159**), metacycline, $C_{22}H_{22}N_2O_8$, M_r 442.43.

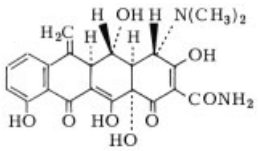

159

Methacycline was synthesized by Pfizer Research Laboratories in 1961 starting with oxytetracycline [299]. It shows two to fivefold greater in vitro antibacterial activity and in vivo protective activity than tetracycline. Methacycline has been given orally to treat infections by *Rickettsia, Chlamydiae, Staphyloccus, Streptococcus, Neisseria, Klebsiella, Proteus, Escherichia coli,* and *Haemophilus influenzae*.

Doxycycline [*17086-28-1*], (**160**), $C_{22}H_{24}N_2O_8 \cdot H_2O$, M_r 462.46.

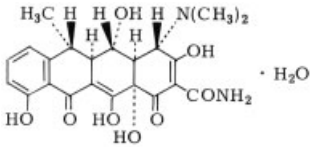

160

Doxycycline was synthesized by Pfizer starting with either tetracycline or oxytetracycline in 1958 [300]. It shows about fourfold higher antibacterial activity than tetracycline against a variety of pathogens. Doxycycline shows higher oral absorption than tetracycline, and its concentration in tissue is higher and maintained longer. Hyclate is a preparation of doxycycline containing one mole of hydrochloric acid and one-half mole each of ethanol and water per mole of doxycycline.

Rolitetracycline [*751-97-3*], (**161**), $C_{27}H_{33}N_3O_8$, M_r 527.58.

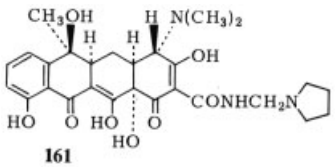

161

Rolitetracycline was synthesized in 1958. Bristol-Myers Laboratories prepared it from tetracycline by introducing a pyrrolidinylmethyl moiety [301]. This antibiotic is very soluble in water and more stable than tetracycline under acidic conditions. Rolitetracycline is used by intravenous injection, and its nitrate is used by both intravenous and intramuscular injections for therapy of the same infections as those treated by tetracycline.

Minocycline [*10118-90-8*], (**162**), $C_{23}H_{27}N_3O_7$, M_r 457.49.

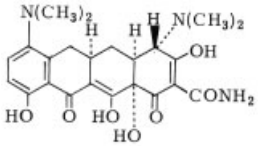

162

Minocycline was synthesized by American Cyanamid in 1966 in the course of studies on derivatives of 6-deoxytetracyclines [302]. It shows activity against tetracycline-resistant bacteria and higher activity than tetracycline against a variety of pathogens. Minocycline has a higher oral absorption than tetracycline, and its concentrations in several tissues are four to tenfold higher than that of tetracycline. It is widely used by oral administration, and it is also administered by drip infusion for serious infections.

Daunorubicin hydrochloride [*23541-50-6*], (**163**), $C_{27}H_{29}NO_{10} \cdot HCl$, M_r 563.99.

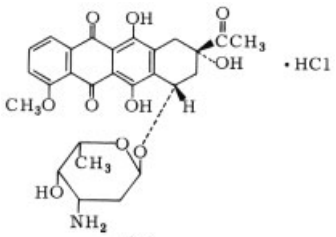

163

Daunorubicin was found independently by Farmitalia and Rhône Poulenc in 1963 in the mycelium of *Streptomyces peucetius* and the culture broth of *Streptomyces coeruleorubidus,* respectively [303]. It was the first anthracycline antibiotic clinically used for therapy of cancers, especially leukemia. Daunorubicin is used in combination with other anticancer drugs.

Doxorubicin [*23214-92-8*], (**164**), $C_{27}H_{29}NO_{11}$, M_r 543.53.

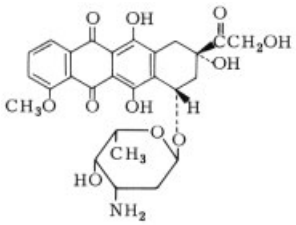

164

Doxorubicin was found in the culture broth of *Streptomyces peucetius* var. *cesius* by Farmitalia in 1967 in the course of studies of anthracycline antibiotics [304]. It shows stronger activity against a variety of tumors and leukemia than daunorubicin, and its clinical application in the therapy of cancer is wider than that of daunorubicin. Doxorubicin is sometimes called by its old generic name, adriamycin, especially in the medical field.

Aclarubicin [*57576-44-0*], (**165**), $C_{42}H_{53}NO_{15}$, M_r 811.89.

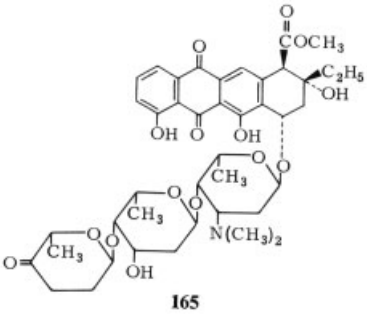

165

Aclarubicin was found in the culture broth of *Streptomyces galilaeus* MA144-M1 by Umezawa et al. of the Institute of Microbial Chemistry in 1975 [305]. It was produced along with structurally related compounds showing antileukemic activity and named aclacinomycin A. Sanraku-Ocean cooperated in isolating aclacinomycin A as a yellow crystalline powder and evaluated its strong antileukemic activity and low cardiac toxicity. Its generic name was changed to aclarubicin on the recommendation of the World Health Organization.

Kanamycin sulfate [*25389-94-0*], (**166**), $C_{18}H_{36}N_4O_{11} \cdot H_2SO_4$, M_r 582.59.

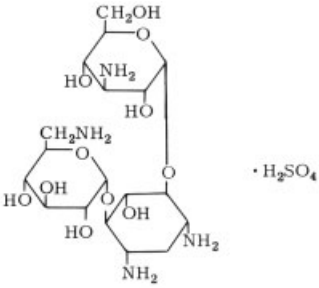

166

Kanamycin was found by UMEZAWA et al. in the culture broth of *Streptomyces kanamyceticus* in 1957 [306]. It is produced with other components, kanamycins B (bekanamycin) and C, which are also separated during the purification. The compound shows much lower toxicity than the earlier aminoglycosides, streptomycin and neomycin, and strong activity against a wide range of gram-positive and gram-negative bacteria, including *Mycobacterium.* Kanamycin has been used clinically for treatment of such serious infections as dysentery, salmonellosis, and tuberculosis.

Bekanamycin [*4696-76-8*], (**167**), kanamycin B, $C_{18}H_{37}N_5O_{10}$, M_r 483.52.

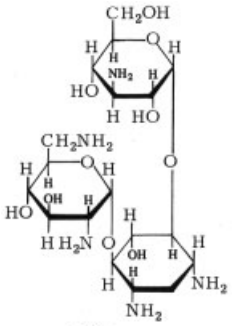

167

Bekanamycin, kanamycin B, was found in the culture broth of *Streptomyces kanamyceticus* by UMEZAWA et al. in 1957 [307]. It shows the same antibacterial spectrum as kanamycin but with stronger activity. The total synthesis of bekanamycin was completed by UMEZAWA et al. in 1968 and the knowledge gained from its synthesis was successfully applied to the synthesis of dibekacin.

Tobramycin [*32986-56-4*], (**168**), $C_{18}H_{37}N_5O_9$, M_r 467.52.

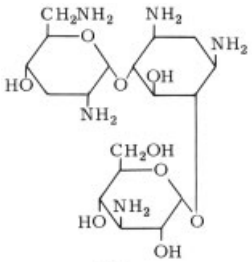

168

Tobramycin was found in 1967 by Eli Lilly Co. in the nebramycins complex that was produced in the culture broth of *Streptomyces tenebrarius* [308]. Structurally it is closely related to kanamycin, a naturally produced 3′-deoxy derivative of bekanamycin. The 3′-hydroxyl group was found to be the target of enzymatic phosphorylation by resistant bacteria. As expected, tobramycin shows strong activity against resistant bacteria, including *Pseudomonas aeruginosa,* having this phosphorylating enzyme.

Gentamicin sulfate [*1405-41-0*], (**169**), gentamicin.

Gentamicin C_1, $C_{21}H_{43}N_5O_7$, M_r 477.61.
Gentamicin C_2, $C_{20}H_{41}N_5O_7$, M_r 463.58.
Gentamicin C_{1a}, $C_{19}H_{39}N_5O_7$, M_r 449.55.

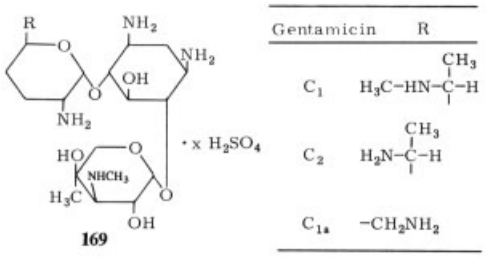

169

Gentamicin was found by Schering Plough Co. in the culture broths of *Micromonospora purpurea* and *M. echinospora* in 1963 [309]. It is a mixture of at least 16 structurally related components. The major components used in clinical preparations are gentamicins C_{1a}, C_1, and C_2. Gentamicin shows high activity against a variety of gram-positive and gram-negative bacteria, including *Pseudomonas aeruginosa, Proteus,* and *Serratia.* It has been widely used for the clinical treatment of serious infections. Gentamicin is used alone or in combination with β-lactam antibiotics and is being replaced gradually by the newer, less toxic aminoglycosides.

Dibekacin [*34493-98-6*], (**170**), $C_{18}H_{37}N_5O_8$, M_r 451.52.

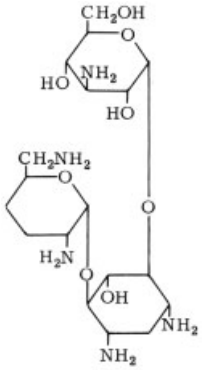

170

Dibekacin was synthesized in 1967 by UMEZAWA et al. by the removal of the 3′- and 4′-hydroxyl groups of kanamycin B [310]. Studies by the same workers on the mechanisms of bacterial resistance to kanamycin-group antibiotics preceded the discovery. Dibekacin shows excellent activity, as expected, against a variety of bacteria, including kanamycin-resistant strains. It shows higher activity than kanamycin against *Pseudomonas aeruginosa, Proteus,* and other pathogens.

Amikacin [*37517-28-5*], (**171**), $C_{22}H_{43}N_5O_{13}$, M_r 585.61.

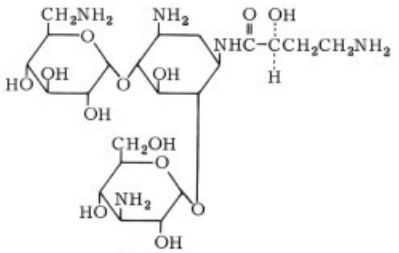

171

Amikacin was synthesized by KAWAGUCHI et al. of the Bristol-Banyu Research Institute in 1970 starting with kanamycin and the acyl moiety of butirosin [311]. Its design is based on knowledge of the mechanisms of bacterial resistance to kanamycin and related compounds in which the 3′-hydroxyl group of the antibiotic is phosphorylated enzymatically. The acyl moiety in butirosin prevents this enzymatic inactivation.

Micronomicin [*52093-21-7*], (**172**), $C_{20}H_{41}N_5O_7$, M_r 463.58.

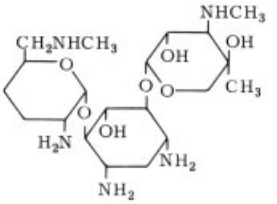

172

Micronomicin, formerly called sagamicin, was found in the culture broth of *Micromonospora sagamiensis* var. *nonreducans* by Kyowa Hakko Kogyo Company and Abbott Laboratories in 1974 [312]. The compound is identical with one of the minor components contained in gentamicin, gentamicin C_{2b}, but it is produced as a single component. Its amino group at the 6′-position is methylated and is not subject to the enzymatic acetylation caused by resistant bacteria, including *Pseudomonas aeruginosa.* Micronomicin is less toxic than gentamicin to the renal and aural systems.

Ribostamycin [*25546-65-0*], (**173**), $C_{17}H_{34}N_4O_{10}$, M_r 454.48.

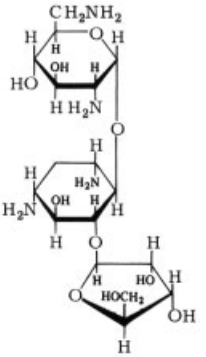

173

Ribostamycin was found in the culture broth of *Streptomyces ribosidificus* by Meiji Seika Kaisha in 1970 in the course of screening aminoglycoside antibiotics [313]. It is structurally related to neomycin but lacks the diaminoidose (glucose) moiety substituted on the ribose moiety. Ribostamycin is much less toxic than neomycin and shows strong activity against a variety of gram-positive and gram-negative bacteria, except *Pseudomonas aeruginosa.* It is used parenterally for therapy of urinary tract, respiratory tract, surgical, and other infections.

Neomycin sulfate [*1405-10-3*], (**174**), fradiomycin, $C_{23}H_{46}O_{13}N_6$, M_r 614.66.

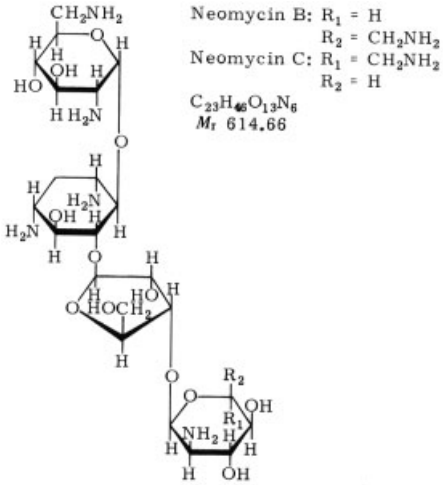

174

Neomycin was found independently in 1949, by UMEZAWA et al. and WAKSMAN et al. in the culture broth of *Streptomyces fradiae* [314]. It consists of two closely related components, B and C, and shows strong activity against a wide range of gram-positive and gram-negative bacteria, including *Serratia* and *Pseudomonas aeruginosa.* Because of its renal and ototoxicity, it is given orally or by topical application. Because it is not absorbed orally, as are other aminoglycoside antibiotics, it is used orally only for the purpose of suppressing intestinal flora, i.e., in treating dysentery, salmonellosis, and diarrhea in the pediatric field. Neomycin also has been used topically in the treatment of bacterial infections of the eye and skin.

Paromomycin sulfate [*1263-89-4*], (**175**), $C_{23}H_{45}N_5O_{14} \cdot xH_2SO_4$, M_r 615.64 (base).

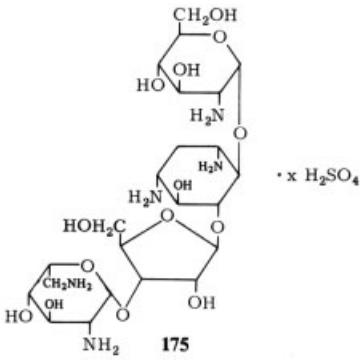

175

Paromomycin was found in the culture broth of *Streptomyces rimosus* forma *paromomycinus* by Parke Davis & Co. in 1959 [315]. In the same year it was found in the culture broth of *S. Crestomyceticus* by Farmitalia and named amminosidin. Paromomycin is related structurally to neomycin, but it has a hydroxyl group at the 6′ position, whereas neomycin has an amino group. Its antibacterial activity is weaker than that of neomycin, but its toxicity is much less. Paromomycin is used by intramuscular injection for therapy of respiratory, urinary, and surgical infections and by oral administration to treat dysentery and salmonellosis.

Streptomycin sulfate [*3810-74-0*], (**176**), $C_{42}H_{84}N_{14}O_{36}S_3$, M_r 1457.40.

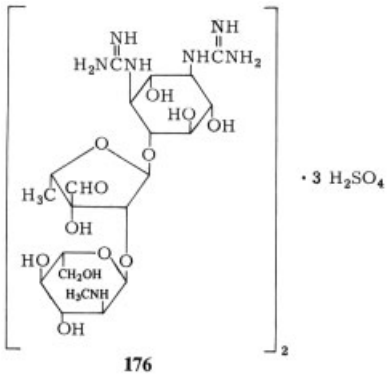

176

Streptomycin was found in the culture broth of *Streptomyces griseus* by S. A. WAKSMAN of Rutgers University in 1944; it was the second antibiotic introduced clinically (after penicillin) [316]. This drug is a water soluble, basic substance having an aminoglycoside structure and showing strong activity against a wide range of gram-positive and gram-negative bacteria including *Mycobacterium.* Streptomycin is the first choice among antituberculotic antibiotics and has been used for the therapy of *Spirochaeta* and *Treponema* infections.

Dihydrostreptomycin [*128-46-1*], (**177**), $C_{21}H_{41}N_7O_{12}$, M_r 583.60.

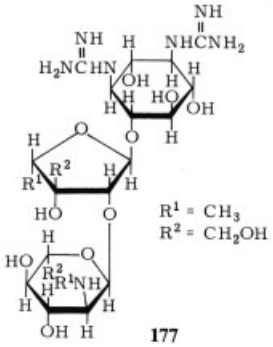

177

Dihydrostreptomycin was first synthesized by Parke Davis Co. in 1946 by the reduction of streptomycin [317]. Naturally occurring dihydrostreptomycin was found in the culture broth of *Streptomyces humidus* by Takeda Chemicals Industries in 1957. Its hydrochloride or sulfate is more easily crystallized and more stable under alkaline conditions than streptomycin. Dihydrostreptomycin has been used for therapy of tuberculosis, but because it has a higher ototoxicity than streptomycin its use is now restricted to animal therapy.

Destomycin A [*14918-35-5*], (**178**), $C_{20}H_{37}N_3O_{13}$, M_r 527.53.

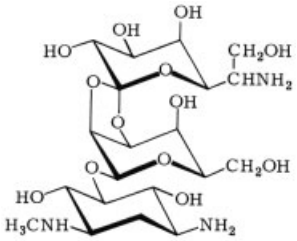

178

Destomycin A was found in the culture broth of *Streptomyces rimofaciens* by Meiji Seika Kaisha in 1965 in the course of screening for water-soluble and basic antibiotics, such as aminoglycosides [318]. It shows activity against a variety of gram-positive and gram-negative bacteria as well as fungi and, more interestingly, against helminths. Destomycin A has been used to treat helminth infections in swine and poultry.

Hygromycin B [*31282-04-9*], (**179**), $C_{20}H_{37}N_3O_{13}$, M_r 527.53.

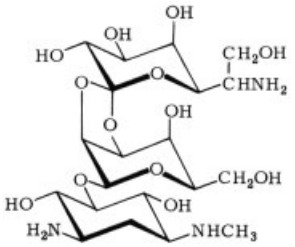

179

Hygromycin B was found in the culture broth of *Streptomyces hygroscopicus* by Lilly Research Laboratories in 1953, as a mixture with hygromycin A [319]. It showed activity against a variety of gram-positive and gram-negative bacteria as well as fungi. Hygromycin B has been used for therapy of helminthic infections in swine and poultry.

Apramycin [*41194-16-5*], (**180**), $C_{21}H_{41}N_5O_{11}$, M_r 539.59.

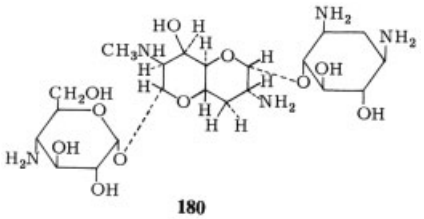

180

Apramycin was found in the culture broth of *Streptomyces tenebrarius* by Eli Lilly in 1967 as one of the eight components of the nebramycin complex [320]. Apramycin is used in feed additives.

Sisomicin [*32385-11-8*], (**181**), $C_{19}H_{37}N_5O_7$, M_r 447.54.

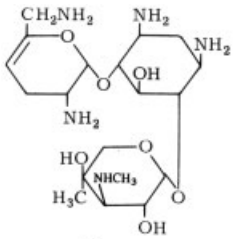

181

Sisomicin was found in the culture broth of *Micromonospora inyoensis* by Schering-Plough Co. in 1970, following the discovery of gentam-

icin by the same research group [321]. The structure and activity of sisomicin are very similar to those of gentamicin C_{1a}, the major component of the gentamicin complex. Sisomicin shows stronger bacterial activity and lower renal and ototoxicity than gentamicin C_{1a}.

Netilmicin sulfate [*56391-57-2*], (**182**), $(C_{21}H_{41}N_5O_7)_2 \cdot 5\,H_2SO_4$, M_r 1441.57.

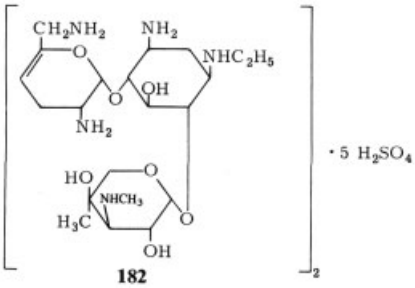

182

Netilmicin was synthesized by Schering-Plough Co. in 1976 [322]. The ethyl moiety was introduced at the 1-amino group of sisomicin. Its design was based on an understanding of the biochemical mechanisms of bacterial resistance to the gentamicin – sisomicin antibiotic group. The modification at the 1-amino group is known to prevent adenylation at the 2″-hydroxyl group and acetylation at the 3-amino group, and to deter acetylation at the 6′-amino residue. Netilmicin shows almost the same activity against a variety of gram-positive and gram-negative bacteria as sisomicin and strong activity against gentamicin – sisomicin-resistant bacteria. It is now under study for evaluation of its clinical efficacy and safety.

Spectinomycin hydrochloride [*22189-32-8*], (**183**), $C_{14}H_{24}N_2O_7 \cdot 2\,HCl \cdot 5\,H_2O$, M_r 495.36.

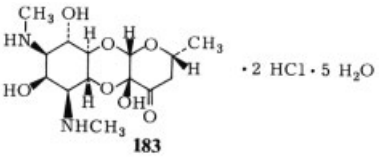

183

Spectinomycin was found in the culture broth of *Streptomyces spectabilis* by Upjohn Co. in 1961 [323]. This antibiotic was intended for use in treating various infections, as are the other aminoglycoside antibiotics, but its activity is insufficient for clinical efficacy except against gonorrhea. Spectinomycin does not develop crossresistance with any other antibiotic and shows low toxicity; it is used by deep intramuscular injection for "single-session," bolus-injection therapy of gonorrhea.

Astromicin sulfate [*72275-67-3*], (**184**), fortimicin A, $C_{17}H_{35}N_5O_6 \cdot 2\,H_2SO_4$, M_r 601.65.

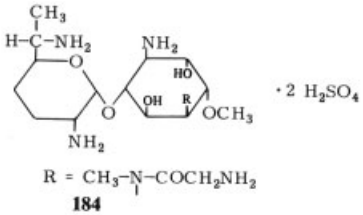

184

Astromicin was found in the culture broth of *Micromonospora olivoasterospora* by Nara et al. of Kyowa Hakko Kogyo Co. in 1976 [324]. It has a unique conformation with an acylated diamino inositol moiety different from other aminoglycoside antibiotics. Astromicin is produced with one major byproduct, fortimicin B, and several minor components.

Validamycin [*37248-47-8*], (**185**), validamycin A, $C_{20}H_{35}NO_{13}$, M_r 497.50.

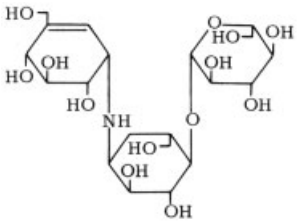

Validamycin A
185

Validamycin was found in the culture broth of *Streptomyces hygroscopicus* var. *limoneus* by Takeda Chemicals Industries in 1971 in the course of screening for substances active against sheath blight in the rice plant [325]. It consists of five components. The major component, validamycin A, was found to be the major contributor to the activity. Validamycin is active against a variety of phytopathogenic fungi, especially *Pel-*

licularia sasakii, and it has been used to protect rice plants against sheath blight.

Kasugamycin [*6980-18-3*], (**186**), $C_{14}H_{25}O_9N_3$, M_r 379.37.

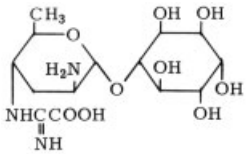

186

Kasugamycin was found in the culture broth of *Streptomyces kasugaensis* by UMEZAWA et al. in 1965 [326]. It has an aminocyclitol structure and shows strong activity against phytopathogenic fungi, especially *Pericularia oryzae,* the pathogen causing rice blast. This drug also shows activity against *Pseudomonas,* and its toxicity is very low; no mice died following intravenous injection of doses as high as two grams per kilogram. Kasugamycin has been used to protect rice plants against rice blast and for animal infections.

Polyoxin [*11113-80-7*], (**187**), polyoxin A [*19396-03-3*], $C_{23}H_{32}N_6O_{14}$, M_r 616.54, polyoxin B, [*19396-06-6*], $C_{17}H_{25}N_5O_{13}$, M_r 507.41.

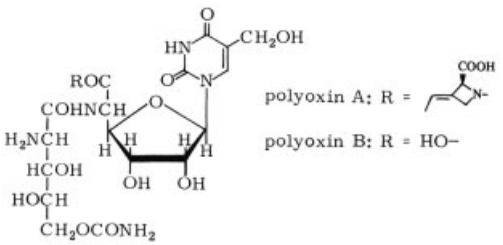

187

Polyoxin was found in the culture broth of *Streptomyces cacaoi* var. *asoensis* by SUZUKI et al. in 1965 [145–147]. It consists of several closely related components, A through O, and shows activity against phytopathogenic fungi by inhibition of cell-wall chitin synthesis. Polyoxin has been used in agriculture against fungal infections, especially *Alternaria* leaf spot in vegetables and fruits.

Blasticidin S [*2079-00-7*], (**34**), $C_{17}H_{26}N_8O_5$, M_r 422.45 (for structure, see page 28).

Blasticidin S was found in the culture broth of *Streptomyces griseochromogenes* by YONEHARA et al. of the University of Tokyo in 1958 [157]. It has a nucleoside-analogue structure and shows strong activity against phytopathogenic fungi, especially *Pericularia oryzae,* the pathogen causing rice blast. Blasticidin S has been used to protect rice plants.

Erythromycin [*114-07-8*], (**188**), $C_{37}H_{67}NO_{13}$, M_r 733.95.

Erythromycin, the first of the macrolide antibiotics, was found in the culture broth of *Streptomyces erythreus* by Eli Lilly Co. in 1952 [327].

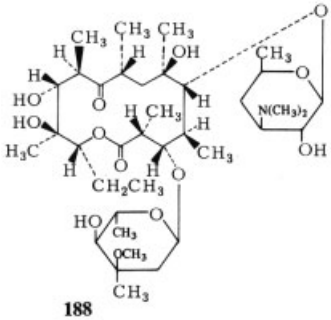

188

Its chemical structure and synthesis have been studied extensively by the WOODWARD group at Harvard University. Erythromycin shows activity against gram-positive bacteria and gram-negative cocci as well as *Mycoplasma* and *Leptospira.* Oral administration of its base and esters is used widely in clinics to treat respiratory tract and other infections. An ester, lactobionate, is used by drip infusion for serious infections (see next two paragraphs).

Erythromycin estolate [*3521-62-8*], (**189**), erythromycin 2′-propionate dodecyl sulfate (salt), $C_{40}H_{71}NO_{14} \cdot C_{12}H_{26}O_4S$, M_r 1056.41.

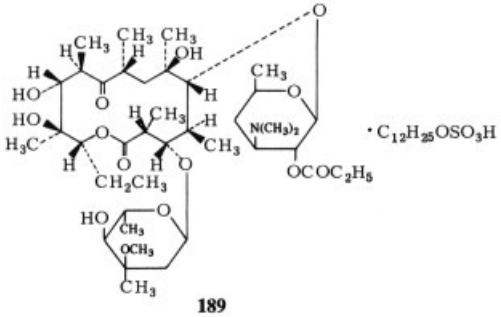

189

Erythromycin estolate was synthesized by the Lilly Research Laboratories in 1959 [328]. It shows a higher and longer lasting serum level than erythromycin and the other esters, ethylsuccinate and stearate, when administered orally. This antibiotic is stable against gastric acids, and after oral absorption it is hydrolyzed gradually into erythromycin. Long-term usage of erythromycin estolate is toxic to the liver. Erythromycin ethylsuccinate [*41342-53-4*], ($C_{43}H_{75}NO_{16}$, M_r 862.07) and erythromycin stearate [*643-22-1*] ($C_{37}H_{67}NO_{13} \cdot C_{18}H_{36}O_2$, M_r 1018.43) also have been used orally, and erythromycin lactobionate [*3847-29-8*] ($C_{37}H_{67}NO_{13} \cdot C_{12}H_{22}NO_{12}$, M_r 1092.25) has been used by intravenous administration and opthalmic topical application.

Oleandomycin phosphate [*7060-74-4*], (**190**), $C_{35}H_{61}NO_{12} \cdot H_3PO_4$, M_r 785.87.

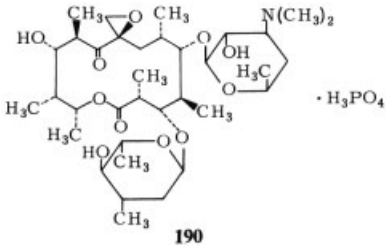

190

Oleandomycin is a macrolide antibiotic with a 14-membered lactone constituent. It was found in the culture broth of *Streptomyces antibioticus* by Pfizer Research Laboratories in 1954 [329]. Oleandomycin shows almost the same antimicrobial spectrum and activity as the other macrolide antibiotics. Its serum concentration is low, but its tissue concentration is high enough to provide a therapeutic effect. Oleandomycin phosphate is used orally or intravenously to treat pyoderma, sepsis, meningitis, surgical and abdominal infections, respiratory tract and urinary tract infections, and other infections caused by *Staphylococci, Streptococci, Corynebacterium, Neisseria,* and *Mycoplasma.* Oleandomycin is also used as a feed additive and therapeutic agent for animals.

Triacetyloleandomycin [*2751-09-9*], (**191**), troleandomycin, $C_{41}H_{67}NO_{15}$, M_r 813.99.

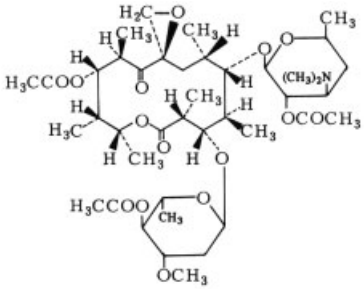

191

Triacetyloleandomycin was synthesized by Pfizer Research Laboratories in 1958 [330]. It shows a higher and longer-lasting serum level than oleandomycin when administered orally. Triacetyloleandomycin behaves as oleandomycin in vivo, following its hydrolysis by intestinal esterase. However, considerable amounts of the intermediates, the monoacetates and diacetates, are detected in the serum and urine. Prolonged use of triacetyloleandomycin causes damage to the liver as does erythromycin estolate.

Kitasamycin [*1392-21-8*], (**192**),

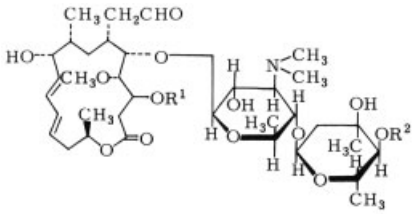

Leucomycin A_1: R^1 = H, R^2 = $COCH_2CH(CH_3)_2$
Leucomycin A_3: R^1 = $COCH_3$, R^2 = $COCH_2CH(CH_3)_2$
Leucomycin A_4: R^1 = $COCH_3$, R^2 = $COCH_2CH_2CH_3$
Leucomycin A_5: R^1 = H, R^2 = $COCH_2CH_2CH_3$
Leucomycin A_6: R^1 = $COCH_3$, R^2 = $COCH_2CH_3$
Leucomycin A_7: R^1 = H, R^2 = $COCH_2CH_3$

192

Kitasamycin, formerly called leucomycin, was found in the culture broth of *Streptoverticillium kitasatoensis* by Hata et al. of the Kitasato Institute, Japan, in 1953 [166]. This was the first macrolide antibiotic with a 16-membered lactone constituent. Kitasamycin is a complex of eight leucomycin A components and is used in its base and ester forms to treat gram-positive bacterial and gram-negative coccal infections, as well as infections of *Mycoplasma, Spirochaeta,* and *Treponema* by injection or by oral topical administration.

Josamycin [*56689-45-3*], (**193**), $C_{42}H_{69}NO_{15}$, M_r 828.02.

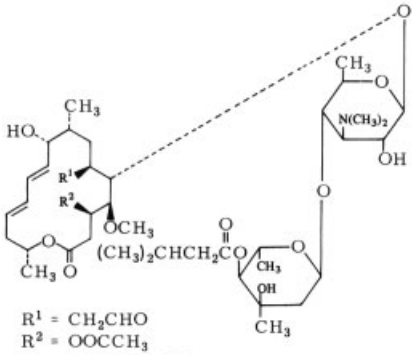

193

Josamycin was found in the culture broth of *Streptomyces narboensis* var. *josamyceticus* by Yamanouchi Pharmaceutical Co. and the Institute of Microbial Chemistry in 1964 [331]. It was the first macrolide antibiotic with a 16-membered lactone constituent to be prepared as a single compound. Its lactone structure was first thought to be 17-membered, but later this idea was revised and its identity with leucomycin A_3, a component of kitasamycin, was confirmed. Clinically, josamycin is given to children in the form of a base or propionate ester to avoid its bitter taste.

Spiramycin [*8025-81-8*], (**194**).

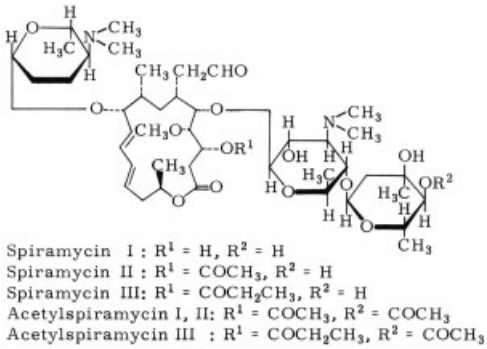

194

Spiramycin was found in the culture broth of *Streptomyces ambofaciens* by Rhône Poulenc in 1954 [168] and its acetate was synthesized by Kyowa Hakko Kogyo Co. in 1965. This antibiotic consists of three closely related macrolide components, I, II, and III, whose ratio is 2 : 1 : 1. Spiramycin shows almost the same antimicrobial spectrum and activity as the other macrolides, but its acetate, acetylspiramycin, has much better pharmacokinetic properties and activity in vivo.

Tylosin [*1401-69-0*], (**195**), $C_{46}H_{77}NO_{17}$, M_r 916.12.

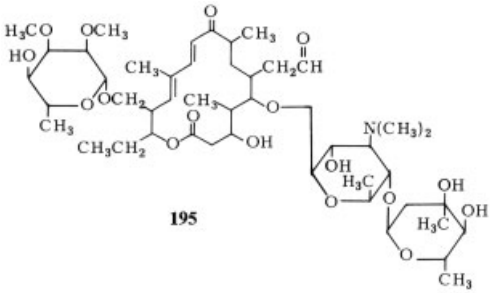

195

Tylosin was found in the culture broth of *Streptomyces fradiae* by Lilly Research Laboratories in 1959 [179–181]. It shows the same antimicrobial spectrum as the other macrolide antibiotics, and its activity against *Mycoplasma* is the widest and highest of any member of its group. Against gram-positive bacteria it is slightly weaker than erythromycin. Tylosin shows a very high and long-term tissue concentration when administered subcutaneously. It is used by injection or oral administration to treat *Mycoplasma* and gram-positive bacterial infections in poultry, swine, and other livestock.

Ivermectin [*70288-86-7*], (**196**), component B_{1a} [*70161-11-4*], component B_{1b} [*70209-81-3*].

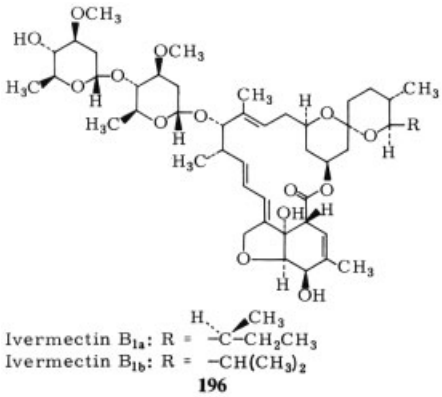

196

Ivermectin was synthesized starting with avermectin, which was found by Merck Sharp & Dohme in the culture broth of *Streptomyces*

avermitilis in 1977 [332]. Unlike other antibiotics, its activity is strictly against insects, mites, and animal parasites. Ivermectin has been used against pests, mites, and other parasites in domestic animals and livestock.

Midecamycin [*35457-80-8*], (**197**), $C_{41}H_{67}NO_{15}$, M_r 813.99.

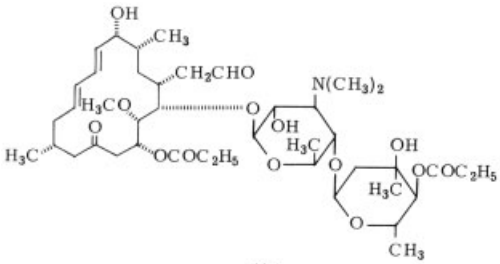

197

Midecamycin, a macrolide antibiotic having a 16-membered lactone ring, was found in the culture broth of *Streptomyces mycarofaciens* by Meiji Seika Kaisha in 1971 [333]. Under specific culture conditions, it is produced by the organism as a single component. Midecamycin shows almost the same antimicrobial spectrum and activity as kitasamycin. Although its serum and urine concentrations are low, it distributes in tissues at high concentration following oral administration.

Rifampicin [*13292-46-1*], (**198**), rifampin, $C_{43}H_{58}N_4O_{12}$, M_r 822.96.

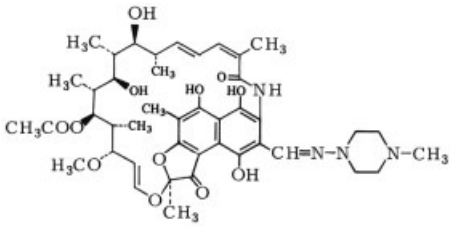

198

Rifampicin was synthesized from rifamycin O obtained from a culture of *Nocardia mediterranei* by Lepetit and Ciba Geigy in 1967 [334]. It shows strong activity against gram-positive bacteria and gram-negative cocci, including *Mycobacterium*, and against viruses. Its clinical application has been restricted to the therapy of tuberculosis caused by streptomycin-resistant tuberculotic bacteria, but recently its usefulness against infections by *Legionella* has been considered.

Bleomycin sulfate [*9041-93-4*], (**199**), variable composition.

Bleomycin was found by UMEZAWA et al. in the culture broth of *Streptomyces verticillus* B80-Z2 in 1965 as a mixture of 13 closely related components [193]. It is produced as a complex with copper (II) ion showing a blue color. Removal of the copper ion with 8-hydroxyquinoline gives a slightly yellowish powder that has the same anticancer activity as the complex but is less toxic. The major component of bleomycin sulfate is bleomycin A_2, which shows strong activity and low toxicity. It is effective against skin, head, neck, lung, and other cancers and malignant lymphoma.

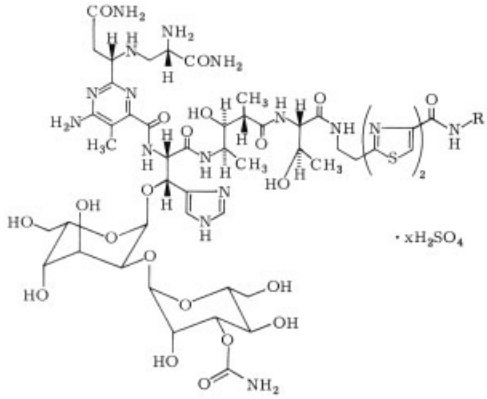

(Main component: Bleomycin A_2, in which R is $(CH_3)_2S^+CH_2CH_2CH_2-$)

199

Peplomycin sulfate [*70384-29-1*], (**200**), pepleomycin, $C_{61}H_{88}N_{18}O_{21}S_2 \cdot H_2SO_4$, M_r 1571.70.

Peplomycin was derived biosynthetically from *Streptomyces verticillus,* the bleomycin-producing organism, in the course of efforts to obtain bleomycin derivatives with high activity and low toxicity. This antibiotic was found by Nippon Kayaku Co. and the Institute of Microbial Chemistry, Japan, in 1974 [335]. Its anticancer effects appear earlier than those of bleomycin; consequently, the period of therapy is shortened. Peplomycin is also active against

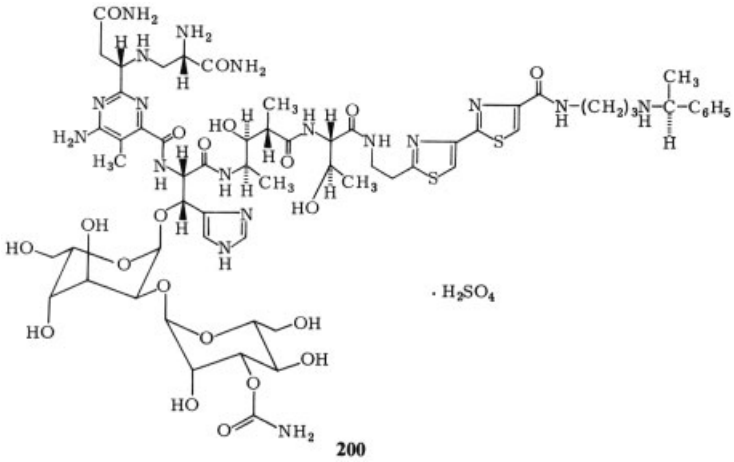

200

metastatic cancer of the lymph nodes and shows less pneumotoxicity than bleomycin.

Gramicidin S [*113-73-5*], (**201**), $C_{60}H_{92}N_{12}O_{10}$, M_r 1141.48.

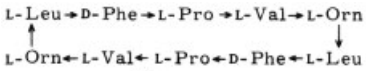

201

Gramicidin S was found in the culture broth of *Bacillus brevis* by GAUSE of the USSR Academy of Medical Sciences in 1944 [199]. It was later found by OTANI of Osaka University in the cells of the *Nagano* strain of the same species. The antibiotic is a basic peptide showing strong activity against gram-positive bacteria and considerable activity against gram-negative cocci and *Mycobacterium.* Gramicidin S is rather toxic but is not absorbed orally; it is used topically as an ointment or as eye or ear drops in combination with other antibacterial drugs.

Polymyxin B [*1404-26-8*], (**202**).

Polymyxin B was isolated by Wellcome Research Laboratories, in 1949, from the mixture of polymyxins A, B, C, and D produced by *Bacillus polymyxa* [336]. It was later separated into the major component, B_1, and the minor component, B_2. Polymyxin B is a basic polypeptide and shows strong activity against gram-negative bacteria, but its activity against gram-positive bacteria, *Mycobacterium,* and fungi is weak. Because of its toxicity, it is used carefully by intramuscular injection for resistant *Pseudomonas aeruginosa* infections, e.g., sepsis. Polymyxin B is used orally to sterilize the gut in leukemic patients, intraspinally for meningitis, or topically.

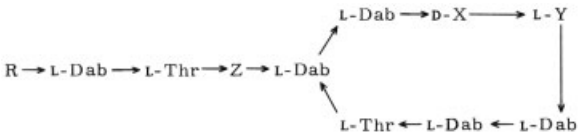

Dab = α, γ-diaminobutyric acid

Polymyxin B_1, [*4135-11-9*], $C_{56}H_{98}N_{16}O_{13}$: M_r *1203.50*, R = (+)-6-methyloctanoyl; X = phenylalanine; Y = leucine; Z = L-Dab

Polymyxin B_2, [*34503-87-2*], $C_{55}H_{96}N_{16}O_{13}$: M_r *1189.48*, R = 6-methylheptanoyl; X = phenylalanine; Y = leucine; Z = L-Dab

202

Bacitracin [*1405-87-4*], bacitracin A (the major component) [*22601-59-8*] (**203**), $C_{66}H_{103}N_{17}O_{16}S$, M_r 1422.73.

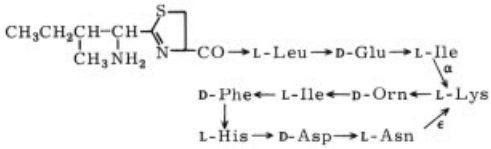

203

Bacitracin was found as a polypeptide complex in the culture broth of *Bacillus subtilis* and *B. licheniformis* by JOHNSON of Columbia University in 1945 [337]. It was first used as a mixture of at least nine bacitracin components. The structure of bacitracin A was determined in 1966 by RESSLER et al. [203].

Colistin sulfate [*1264-72-8*], (**204**), $C_{45}H_{85}O_{10}N_{13} \cdot 2.5\,H_2SO_4$, M_r 1213.45.

Colistin was found in the culture broth of *Bacillus polymyxa* var. *colistinus* by Kayaku Antibiotics Research in 1950 [338]. It is closely related to polymyxin and shows strong activity against gram-negative bacteria, including *Pseudomonas aeruginosa.* Colistin and its methanesulfonic acid derivative have been used to treat urinary tract infections caused by *Escherichia coli* and *P. aeruginosa.* They have also been used parenterally to treat dysentery and abdominal infections and topically for ophthalmic and otorhinolaryngological infections.

Colistinmethanesulfonate sodium (**205**) [*8068-28-8*], colistimethate sodium, $C_{58}H_{107}N_{16}Na_5O_{28}S_5$ (colistin A component), M_r 1749.84; [*21362-08-3*] $C_{57}H_{103}N_{16}Na_5O_{28}S_5$ (colistin B component), M_r 1735.82.

Colistinmethansulfonate was synthesized by Kayaku Antibiotics Research Co. in 1955 [339]. It is much less toxic than colistin and has been used intramuscularly and orally to treat infections caused by *Pseudomonas aeruginosa, Escherichia coli, Klebsiella pneumoniae, Enterobacter,* and *Shigella.*

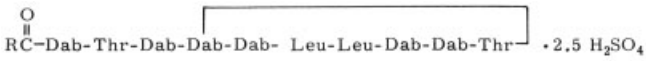

Dab is L-α, γ-diaminobutyric acid; R is 5-methylheptyl in colistin A and 5-methylhexyl in colistin B

204

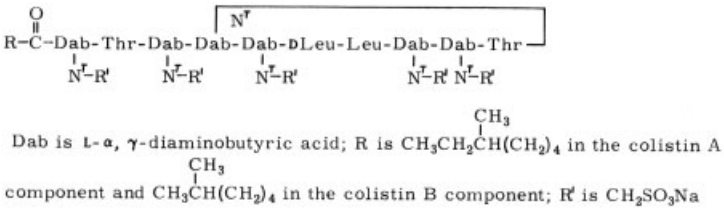

Dab is L-α, γ-diaminobutyric acid; R is $CH_3CH_2CH(CH_2)_4$ in the colistin A component and $CH_3CH(CH_2)_4$ in the colistin B component; R' is CH_2SO_3Na

205

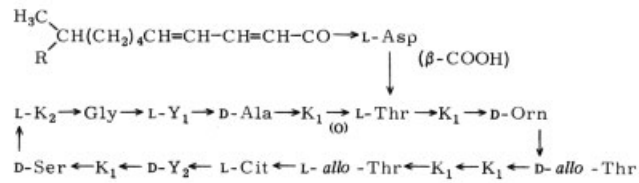

Enramycin A: R = CH_3
Enramycin B: R = C_2H_5
K_1: α-Amino-4-hydroxyphenylacetic acid
K_2: α-Amino-3,5-dichloro-4-hydroxyphenylacetic acid
Y_1: α(*S*)-Amino-β-4(*R*)-(2-iminoimidazolidinyl)propionic acid
Y_2: α(*R*)-Amino-β-4(*R*)-(2-iminoimidazolidinyl)propionic acid

206

Enramycin [*11115-82-5*], (**206**), enduracidin, A: $C_{107}H_{138}N_{25}O_{31}Cl_2$, M_r 2355.35; B: $C_{108}H_{140}N_{26}O_{31}Cl_2$, M_r 2369.38.

Enramycin, formerly called enduracidin, was found in the culture broth of *Streptomyces fungicidicus* by Takeda Chemicals Industries in 1967 [340]. This antibiotic shows strong activity against gram-positive bacteria but not against gram-negative bacteria. It is not subject to any of the mechanisms of resistance against other antibiotics and shows activity against otherwise resistant strains. Enramycin was evaluated for human therapy, but it is now used as a feed additive for growth promotion in poultry.

Mikamycin [*11006-76-1*], mikamycin B [*3131-03-1*] (**207**), $C_{45}H_{54}N_8O_{10}$, M_r 866.98.

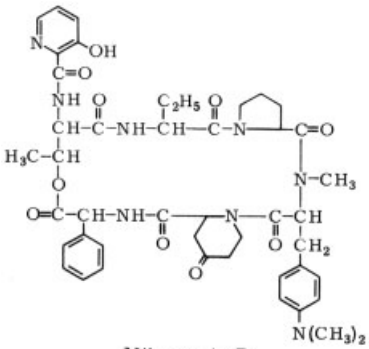

Mikamycin B

207

Mikamycin was found in the culture broth of *Streptomyces mitakaensis* by UMEZAWA et al. of the University of Tokyo in 1956 [341]. It is a mixture of two components, A and B, showing synergistic effects on each other. Staphylomycin, which was found by DE SOMER of Leuven University in 1955, is a closely related compound. Mikamycin shows strong activity against gram-positive bacteria, including strains resistant to other antibiotics. It has been used topically to treat *Staphylococcus* or *Streptococcus* infections of the skin, including burns.

Virginiamycin [*11006-76-1*], (**208**),

Virginiamycin, formerly called staphylomycin, was found in the culture broth of *Streptomyces virginiae* by DE SOMER et al. of Leuven University in 1955 [342]. The commercial product contains about 75 % of the fraction M_1 and 5 % of the fraction S. The structure of the minor component, M_2, has not yet been elucidated. Virginiamycin has very low toxicity and is not absorbed from the intestine when administered orally. It shows activity against gram-positive bacteria, and it has been used as a feed additive for growth promotion in swine, cattle, and other livestock.

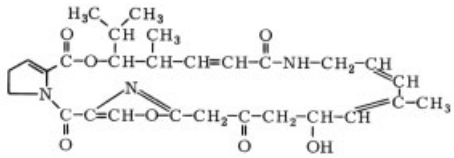

Factor S, $C_{42}H_{47}N_7O_{10}$, M_r 809.88

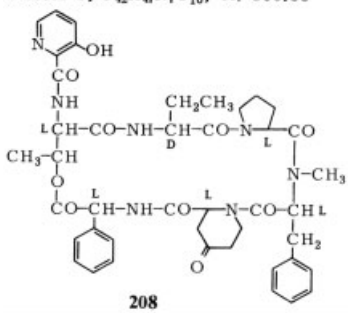

208

Capreomycin sulfate [*1405-37-4*], (**209**),

Capreomycin was found as an antituberculotic antibiotic by HERR et al. of Eli Lilly & Co. in the culture broth of *Streptomyces capreolus* NRRL2773 in 1962 [343]. It shows no cross-resistance with streptomycin and is used against tuberculosis caused by streptomycin-resistant *Mycobacterium* or for patients with adverse side effects caused by streptomycin (see next page).

Viomycin [*32988-50-4*], (**210**), $C_{25}H_{43}N_{13}O_{10}$, M_r 685.70.

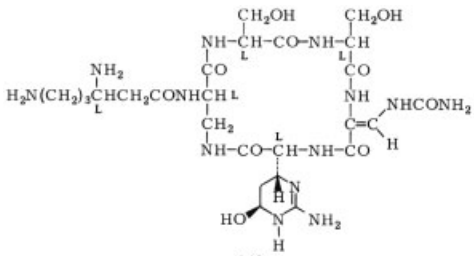

210

Viomycin was found independently in 1951, in the culture broth of *Streptomyces puniceus*

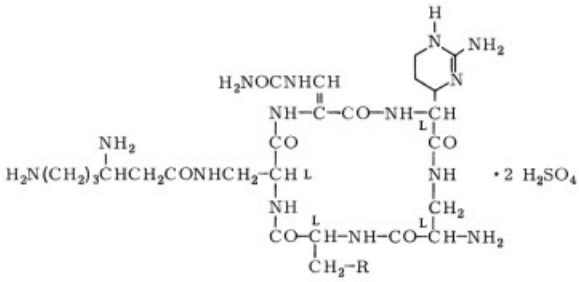

Capreomycin I A : R = OH; $C_{25}H_{44}N_{14}O_8 \cdot 2\ H_2SO_4$, M_r 864.87
Capreomycin I B : R = H; $C_{25}H_{44}N_{14}O_7 \cdot 2\ H_2SO_4$, M_r 848.87

209

by Pfizer Research Laboratories and in that of *S. floridae* by Parke Davis Co. [344]. It shows strong activity against *Mycobacterium* species and moderate activity against gram-positive and gram-negative bacteria. Viomycin has been used by intramuscular administration to treat tuberculosis, but because of its ototoxicity and renal toxicity it is being replaced by more active and less toxic drugs, such as enviomycin and other antituberculotic antibiotics.

Enviomycin [*33103-22-9*], (**211**), tuberactinomycin N, $C_{25}H_{43}N_{13}O_{10}$, M_r 685.70.

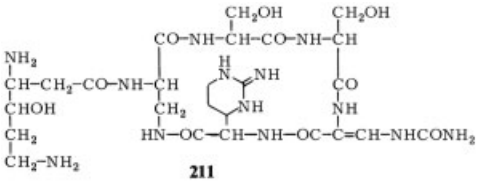

211

Enviomycin was found in the culture broth of *Streptomyces griseoverticillatus* var. *tuberacticus* by Toyo Jozo Co. in 1966 as a mixture of tuberactinomycins N and O [345]. It is a water-soluble, basic peptide closely related to viomycin and showing selective activity against *Mycobacterium* species about twice as strong as that of viomycin. Enviomycin, in combination with other antituberculotic drugs, is used to treat tuberculosis caused by streptomycin-resistant *Mycobacterium.*

Vancomycin [*1404-90-6*], (**212**), $C_{66}H_{75}Cl_2N_9O_{24}$, M_r 1449.29.

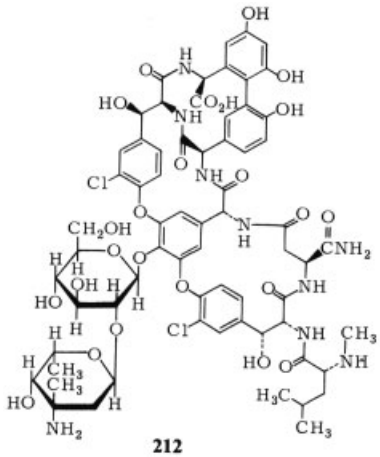

212

Vancomycin was found by Lilly Research Laboratories in 1956 in the culture broth of *Streptomyces orientalis* [346]. It is a large, molecular glycopeptide showing bactericidal activity against gram-positive bacteria and anaerobes. It is used to treat resistant infections of *Staphylococcus,* to sterilize the gut in the perioperation of bone-marrow transplantation, and in leukemic patients. Recently its efficacy has been demonstrated against pseudomembrane colitis, which is caused by *Clostridium difficile*.

Actinomycin D [*50-76-0*], (**213**), dactinomycin, $C_{62}H_{86}N_{12}O_{16}$, M_r 1255.45.

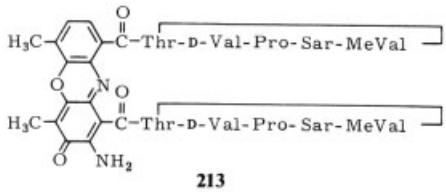

213

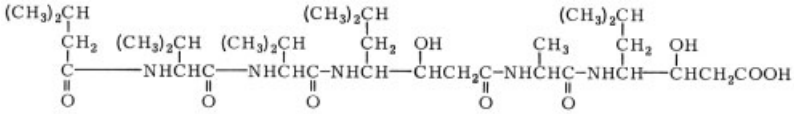

215

Actinomycins (mixtures of A, B, C, and other components) were first found by WAKSMAN et al. of Rutgers University in 1940 in the culture broth of *Streptomyces antibioticus* [347]. The mixture of the C components has been used as cactinomycin for anticancer therapy. Actinomycin D was found by the same group in 1954, obtained as a single component from *S. parvullus*. Its strong activity against Wilm's tumor and other cancers has been evaluated. Dactinomycin is the name recommended by INN and used in the United States.

Neocarzinostatin [*9014-02-2*], zinostatin.

Ala – Ala – Pro – Thr – Ala – Thr – Val – Thr – Pro – Ser – Ser – Gly – Leu – Ser – Asp – Gly – Thr – Val – Val – Lys – Val – Ala – Gly – Ala – Gly – Leu – Gln – Ala – Gly – Thr – Ala – Tyr – Asp – Val – Gly – Gln – Cys – Ala – Ser – Val – Asn – Thr – Gly – Val – Leu – Trp – Asn – Ser – Val – Thr – Ala – Ala – Gly – Ser – Ala – Cys – Asx – Pro – Ala – Asn – Phe – Ser – Leu – Thr – Val – Arg – Arg – Ser – Phe – Glu – Gly – Phe – Leu – Phe – Asp – Gly – Thr – Arg – Trp – Gly – Thr – Val – Asx – Cys – Thr – Thr – Ala – Ala – Cys – Gln – Val – Gly – Leu – Ser – Asp – Ala – Ala – Gly – Asp – Gly – Glu – Pro – Gly – Val – Ala – Ile – Ser – Phe – Asn

Neocarzinostatin was found in the culture broth of *Streptomyces carzinostaticus* by ISHIDA et al. of Tohoku University in 1957 [348]. It showed strong cytotoxicity against sarcoma 180 ascites, tumor cells, and leukemia SN-36. Neocarzinostatin was found to be an acidic peptide of M_r 10700, consisting of 109 amino acid residues. Recently the principal agent of anticancer and antibacterial activity in neocarzinostatin was found to be a small molecular chromophore with the peptide component playing a role by stabilizing the chromophore in vivo. Neocarzinostatin is used to treat leukemia and gastric and pancreatic cancer.

Bestatin [*58970-76-6*], (**214**), $C_{16}H_{24}N_2O_4$, M_r 308.38.

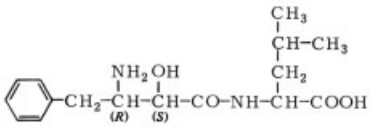

214

Bestatin was found in the culture broth of *Streptomyces olivoreticuli* in 1976 by UMEZAWA et al. in the course of screening specific enzyme inhibitors of microbial origin [349]. It shows inhibitory activity against specific exopeptidases, e.g., aminopeptidase B and leucine aminopeptidase. However, it does not act against aminopeptidase A, carboxypeptidases A and B, or endopeptidases. Bestatin acts as a bioresponse modifier and shows antitumor activity via stimulation of immuno response in the host. In combination with cytotoxic anticancer drugs, it is now under clinical evaluation for use in cancer therapy.

Pepstatin [*26305-03-3*], (**215**), $C_{34}H_{63}N_5O_9$, M_r 685.91.

Pepstatin was found in the culture broth of *Streptomyces testaceus* by UMEZAWA of the Institute of Microbial Chemistry in 1970 in the course of screening for specific enzyme inhibitors [350]. It shows strong inhibitory activity against several proteases, such as pepsin, cathepsin D, and renin, but not against trypsin, chymotrypsin, papain, etc. Its potential use against ulcers in humans and animals has been studied.

Monensin [*17090-79-8*], (**216**), $C_{36}H_{62}O_{11}$, M_r 670.89.

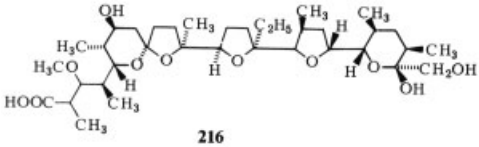

216

Monensin was found in the mycelium of *Streptomyces cinnamonensis* by Eli Lilly & Co. in 1967 in the course of screening for anticoccidial and growth-promoting substances for cattle, swine, and chickens [351]. It shows antiprotozoal, antibacterial, and antifungal activities, as well as anticoccidial activity. Monensin is widely used to treat such infections and as a feed additive to promote the growth of livestock.

Lasalocid [*25999-31-9*], (**217**), $C_{34}H_{54}O_8$, M_r 590.80.

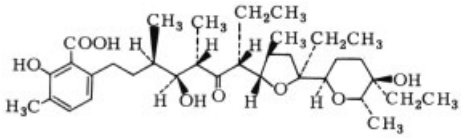

217

Lasalocid is one of the polyether ionophore antibiotics and was found in the culture broth of *Streptomyces lasaliensis* by Hoffmann-La Roche in 1951 [352]. It shows activity against gram-positive bacteria, including *Mycobacterium* and *Streptomyces*, and also against anaerobic bacteria, but it has no activity against gram-negative bacteria or fungi. Lasalocid also shows strong anticoccidial activity, and it stimulates propionic acid formation in the rumen. It is used for prophylaxis and for therapy of coccidial infections in poultry and as a feed additive for growth promotion in cattle.

Salinomycin [*53003-10-4*], (**218**), $C_{42}H_{70}O_{11}$, M_r 751.02.

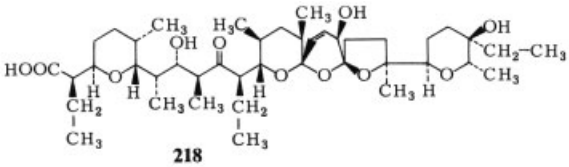

218

Salinomycin was found in the culture mycelium of *Streptomyces albus* by OTAKE et al. of the University of Tokyo in 1973 [353]. Like other polyether ionophore antibiotics, it shows activity against gram-positive bacteria, fungi, and *Coccidium*. Salinomycin has been used as a feed additive to protect poultry against coccidial infections.

Amphotericin B [*1397-89-3*], (**219**), $C_{47}H_{73}NO_{17}$, M_r 924.10.

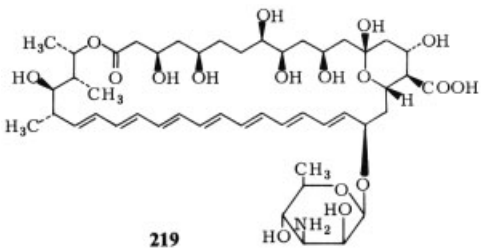

219

Amphotericin B was found in the mycelium of *Streptomyces nodosus* M-4575 by GOLD et al. of the Squibb Institute of Medical Research in 1956 [354]. It is produced with another polyene macrolide antibiotic, amphotericin A, and separated by solvent extraction. Amphotericin B shows strong antimycotic activity against a variety of fungi and pathogenic yeasts (*Candida*) and is used by injection and as a vaginal suppository.

Nystatin [*1400-61-9*], (**220**).

Nystatin A_1 [*34786-70-4*], $C_{47}H_{75}NO_{17}$, M_r 926.12.

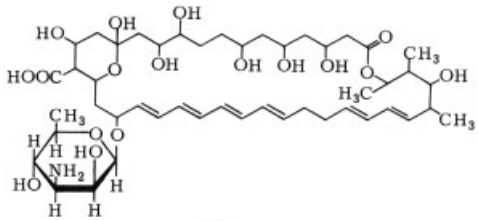

220

Nystatin was found in the mycelium of *Streptomyces noursei* in 1950 [355] and produced on an industrial scale by Squibb & Sons Co. in 1954. This antibiotic was used orally and topically as the first clinically applied polyene macrolide with antifungal properties. Nystatin shows activity against *Candida* and filamentous fungi and is used to treat *Candida* infections of the mouth, digestive organs, and vagina. The application of nystatin in combination with gentamicin and vancomycin to sterilize the gut in perioperation of bone-marrow transplantation has been developed recently.

Natamycin [*7681-93-8*], (**221**), pimaricin, $C_{33}H_{47}NO_{13}$, M_r 665.74.

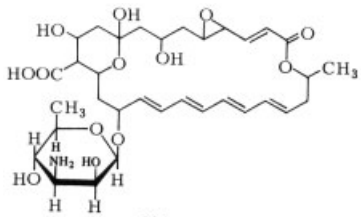

221

Natamycin was found in the culture mycelium of *Streptomyces natalensis* by Royal

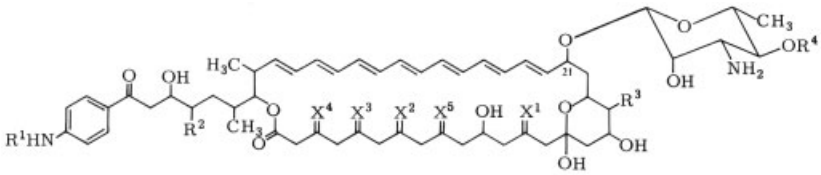

Antibiotic:	R^1	R^2	R^3	R^4	X^1	X^2	X^3	X^4	X^5
Trichomycin B:	H	H	COOH	H	H, OH	O	H, H	O	H, OH
Levorin A_2 (candicidin D):	H	CH_3	COOH	H	H, OH	O	H, H	O	H, OH
Hamycin (67-121 B):	H	H	COOH	H	H, OH	H, OH	H, OH	H, OH	H, OH

222

Dutch Yeast & Fermentation Industries and in that of *S. gilvosporeus* by American Cyanamid, independently, in 1957 [356]. It has a tetraene structure, and like other polyene macrolide antibiotics, it shows activity against pathogenic fungi. Natamycin has been used as a vaginal suppository for therapy of *Candida* infections.

Trichomycin [*1394-02-1*], (**222**), hachimycin.

Trichomycin (hachimycin in the INN) was found in the mycelium of *Streptomyces hachijoensis* by HOSOYA et al. of the University of Tokyo in 1952 [357]. It has a heptaene structure and once was considered identical with two other polyene macrolide antibiotics, candicidin (found by WAKSMAN et al. of Rutgers University in 1953 in *S. griseus*) and hamycin (found by THIRUMALACHAR et al. of Hindustan Antibiotics Research Institute in 1961 in *S. pimprina*). Its difference from them was shown by an HPLC comparison in 1980. Trichomycin shows activity against pathogenic *Candida, Trichomonas,* and *Trypanosoma* and has been used in vaginal applications to treat *Trichomonas vaginalis* infections.

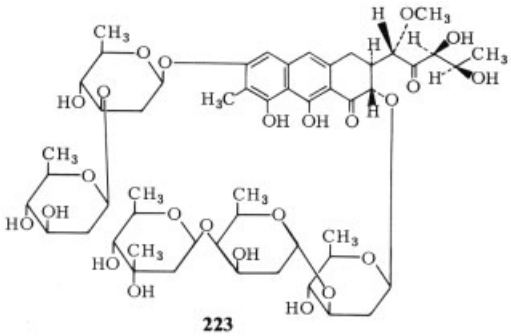

223

Mithramycin [*18378-89-7*], (**223**), plicamycin, $C_{52}H_{76}O_{24}$, M_r 1085.17.

Mithramycin, recently renamed plicamycin, was found in the culture broth of *Streptomyces argillaceus* and *S. tanashiensis* by Abbott Laboratories in 1952 [358]. It is structurally related to chromomycin A_3. Mithramycin shows strong inhibitory activity against malignant cells of human origin. It acts by inhibition of the DNA-directed RNA synthesis through binding with DNA. Mithramycin is used intravenously to treat cancers of the embryonal cells, seminoma, choriocarcinoma, etc.

Lincomycin [*154-21-2*], (**224**), $C_{18}H_{34}N_2O_6S$, M_r 406.55.

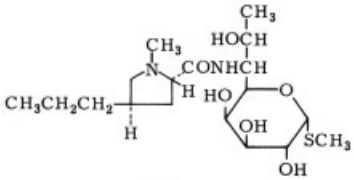

224

Lincomycin was found in the culture broth of *Streptomyces lincolnensis* var. *lincolnensis* by the Upjohn Co. in 1962 [359]. It shows antibacterial activity similar to that of the macrolide antibiotics and also shows excellent activity against anaerobic bacteria. Lincomycin is used clinically in combination with other classes of antibiotics for postoperative, gynecological, urinary tract, ear and nose, and other infections.

Clindamycin [*18323-44-9*], (**225**), 7(*S*)-chloro7-deoxylincomycin, $C_{18}H_{33}ClN_2O_5S$, M_r 424.99.

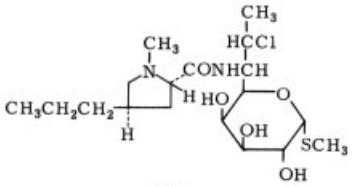

225

Clindamycin was synthesized at the Upjohn Co. in 1966 by introducing a chlorine atom at the 7 position of lincomycin [360]. It is used as an orally active lincomycin-group antibiotic with the same range of activity as lincomycin. Clindamycin shows excellent activity against gram-positive bacteria and anaerobes, and it is used with other antibiotics for complicated infections.

Clindamycin palmitate hydrochloride [*25507-04-4*], (**226**), $C_{34}H_{63}ClN_2O_6S \cdot HCl$, M_r 699.87.

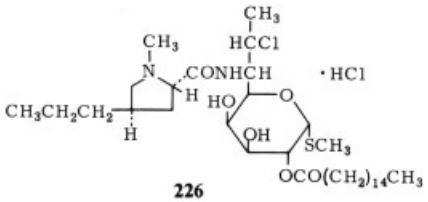

226

Clindamycin palmitate was synthesized by the Upjohn Co. in 1968 [361]. It has a much less bitter taste than clindamycin and is suitable for oral preparations for children. Clindamycin palmitate is very rapidly hydrolyzed into clindamycin after oral administration. Clindamycin phosphate [*24729-96-2*], $C_{18}H_{34}N_2O_8ClPS$, M_r 504.97, has been introduced clinically as an injectable clindamycin.

Flavophospholipol [*11015-37-5*], (**227**), moenomycin, bambermycin, $C_{69}H_{107}N_4O_{35}P$, M_r 1583.60.

Flavophospholipol, formerly called moenomycin, was found in the culture broth of *Streptomyces bambergiensis* by Hoechst in 1955 [362]. This drug is called bambermycin by INN and in the United States. It shows strong activity against gram-positive bacteria and weak activity against gram-negative bacteria. Flavophospholipol has been used as a feed additive for growth promotion in poultry.

Cycloserine [*68-41-7*], (**228**), $C_3H_6N_2O_2$, M_r 102.09.

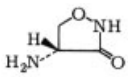

228

Cycloserine was found in the culture broth of *Streptomyces orchidaceus* by Commercial Solvent Co. [363]. It was developed in collaboration with Eli Lilly & Co. in 1955 and is manufac-

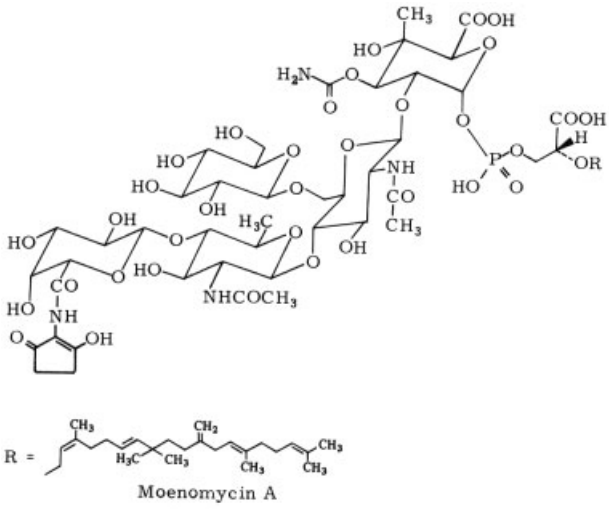

Moenomycin A

227

tured synthetically. Cycloserine shows weak activity against gram-positive and gram-negative bacteria, including *Mycobacterium*. Its activity against clinically isolated tuberculotic bacteria, including streptomycin- and viomycin-resistant strains, is five to ten micrograms per millimeter. Cycloserine is used to treat tuberculosis caused by organisms resistant to other antituberculotic antibiotics.

Pecilocin [*19504-77-9*], (**229**), variotin, $C_{17}H_{25}NO_3$, M_r 291.39.

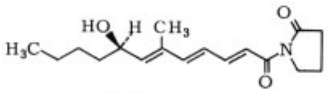

229

Pecilocin was discovered in the culture broth and mycelium of *Paecilomyces varioti* var. *antibioticus* by YONEHARA et al. of the University of Tokyo in 1959 [364]. It has an oily nature and shows activity against specific filamentous fungi. Pecilocin is used as an ointment or in ethanolic solution for the treatment of dermatomycoses caused by *Trichophyton, Microsporum,* and *Epidermophyton.*

Griseofulvin [*126-07-8*], (**230**), $C_{17}H_{17}ClO_6$, M_r 352.77.

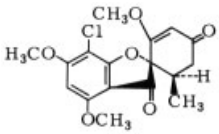

230

Griseofulvin was discovered in the mycelium of *Penicillium griseofulvum* and other *Penicillium* species by OXFORD et al. in 1939 [210]. Its application in the treatment of fungal infections was initiated by ICI in 1946. Griseofulvin is used clinically by topical and oral administration for therapy of dermatomycoses caused by *Trichophyton* and *Microsporum* species.

Chloramphenicol [*56-75-7*], (**231**), $C_{11}H_{12}Cl_2N_2O_5$, M_r 323.13.

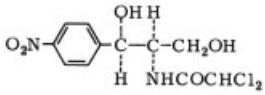

231

Chloramphenicol was found in the culture broth of *Streptomyces venezuelae* by EHRLICH et al. of Parke Davis & Co. in 1947 [212]. The effective industrial synthesis was developed in 1949, and the compound is manufactured synthetically. This was the third antibiotic to be introduced clinically and showed a wide range of activity against gram-positive and gram-negative bacteria, as well as *Treponema, Rickettsia,* and *Chlamydiae.* Chloramphenicol shows excellent activity in the treatment of dysentery and salmonellosis, and it is widely used against these diseases.

Chloramphenicol palmitate [*530-43-8*], (**232**), $C_{27}H_{42}Cl_2N_2O_6$, M_r 561.54.

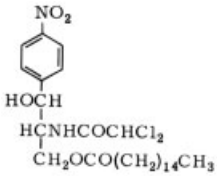

232

Chloramphenicol palmitate was synthesized by Parke Davis Co. in 1952 [365]. It has a much less bitter taste than chloramphenicol and is suitable for oral administration, especially for children. The palmitate shows a higher serum level after oral administration than chloramphenicol does and acts the same as chloramphenicol in vivo.

Chloramphenicol sodium succinate [*982-57-0*], $C_{15}H_{15}N_2O_8Cl_2Na$, M_r 445.19, is highly soluble in H_2O and can be given by intravenous injection.

Mitomycin C [*50-07-7*], (**233**), $C_{15}H_{18}N_4O_5$, M_r 334.33.

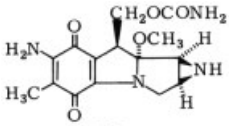

233

Mitomycin C was found in the culture broth of *Streptomyces caespitosus* by WAKAKI in 1958 [216]. This followed the discovery of mitomycins A and B, derived from the same organism by HATA of the Kitasato Institute in 1956. Mitomycin C shows strong activity against a variety of tumors and bacteria. Because it decreases the concentration of leukocytes, its use had been limited, but after 1975 its low-dosage use in anticancer therapy in combination with other anticancer drugs increased.

Pyrrolnitrin [*1018-71-9*], (**234**), $C_{10}H_6Cl_2N_2O_2$, M_r 257.08.

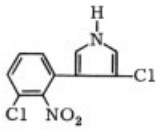

234

Pyrrolnitrin was found in the cells of *Pseudomonas pyrrocinia* grown in a medium containing a high concentration of inorganic phosphate. It was discovered by ARIMA of the University of Tokyo in collaboration with Fujisawa Pharmaceutical Industries Co. in 1965 [366]. The same compound was found soon after that by Eli Lilly & Co. in the cells of *P. aureofaciens.* Pyrrolnitrin shows strong activity against a variety of fungi and weak activity against gram-positive bacteria. It is used topically to treat dermatomycoses caused by *Trichophyton, Epidermophyton,* and *Microsporum.* The antibiotic is unstable in sunlight and cannot be used to protect plants against phytopathogenic fungi.

Fosfomycin [*23155-02-4*], (**235**), $C_3H_7O_4P$, M_r 138.06.

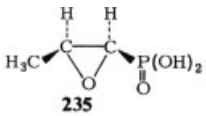

235

Fosfomycin was found in the culture broth of *Streptomyces fradiae* by Compania Espanola de Penicilina y Antibióticos and Merck Sharp & Dohme in 1967 [219]. Its chemical structure is simple and unique among antibiotics in having a C–P bond. Fosfomycin shows antibacterial activity against gram-positive and gram-negative organisms, including *Pseudomonas aeruginosa* and *Serratia marcescens*, and β-lactam-resistant *Staphylococcus aureus.* Its mechanism of action is probably the inhibition of cell-wall synthesis. It shows no cross-resistance with other classes of antibiotics.

Fusidic acid [*6990-06-3*], (**236**), $C_{31}H_{48}O_6$, M_r 516.72.

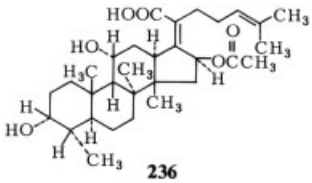

236

Fusidic acid was found in the culture broth of a fungus imperfectus, *Fusidium coccineum,* by Leo in 1962 [220]. It has a steroid structure but shows no hormonal activity. Fusidic acid shows very strong activity against *Staphylococcus aureus* and weak activity against other gram-positive bacteria and gram-negative cocci and *Mycobacterium.* Its clinical use is restricted to staphylococcal infections resistant to other classes of antibiotics.

Bicozamycin [*38129-37-2*], (**237**), bicyclomycin, $C_{12}H_{18}N_2O_7$, M_r 302.29.

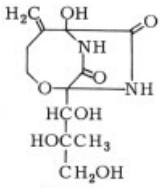

237

Bicozamycin, formerly called bicyclomycin, was found independently in 1972, in the culture broth of *Streptomyces sapporonensis* by Fujisawa Pharmaceuticals Industries and in that of *S. aizuensis* by MIYAMURA et al. of Niigata University [367]. It has a unique bicyclic structure and shows activity against *Klebsiella, Salmonella,* and *Shigella* but none against other gram-negative bacteria or gram-positive microorganisms. Bicozamycin is not absorbed

orally and shows very low toxicity; no mice died following its intravenous injection at a dose as high as two grams per kilogram. This drug is used orally to treat salmonellosis and dysentery.

Tiamulin [*55297-95-5*], (**238**), $C_{28}H_{47}NO_4S$, M_r 493.76.

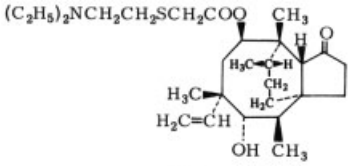

238

Tiamulin was synthesized by Sandoz Co. in 1973, starting from pleuromutilin produced by *Pleurotus mutilus* [368]. It has been developed by that company in collaboration with E. R. Squibb & Sons for use against animal infections. Tiamulin is used against dysentery and *Mycoplasma* infections in sheep, swine, cattle, and chickens.

Siccanin [*22733-60-4*], (**239**), $C_{22}H_{30}O_3$, M_r 342.48.

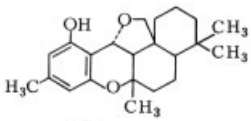

239

Siccanin was found in the culture broth of *Helminthosporium siccans* by Sankyo Co. in 1962 [369]. It shows specific and strong activity against the dermatophytes *Trichophyton, Epidermophyton,* and *Microsporum,* but it has almost no activity against other fungi. Siccanin has been used as a topical ointment or solution to treat dermatomycoses.

5. Fermentation

Fermentation is considered here from the following points of view:

1) Biological development, which includes screening and selection, mutation, and maintenance of the strain.
2) Process development leading to large-scale manufacture.
3) Improvements in fermentation technology.

5.1. Screening

Technical developments in the production of penicillin have given the field new momentum and have stimulated the search, not only for more efficient strains, but also for microorganisms that produce completely different antibiotics. This process is called screening because valuable antibiotic producers are separated from the large number of organisms found in nature. The screening process and the expected results are influenced by several factors.

Source of Sample. Worldwide screening endeavors to isolate the individual microorganisms not only from soil samples from different sources, but also from other microbe-containing materials. Samples from unusual sources often show the occurrence of selection and adaptation. For example, thermophilic microorganisms are examined in samples taken from deep caves, the sea bottom, hot springs, or geysers.

Examination Technique. There are several factors that determine the conditions under which a certain microorganism not only lives and grows but also efficiently produces its antibiotic. These factors include the composition and pH of the culture medium, the additives, the air supply, and the temperature. These factors are also of prime importance for any later industrial fermentation. The isolation and testing of the new antibiotic, first in vitro and then in animal experiments, and the indisputable proof that the new compound is not identical to one of the numerous known antibiotics are part of the examination technique.

Purpose of the Examination. In the early days of antibiotic screening, any organism that showed antibiotic activity was screened, but later definite objectives were set and appropriate examination techniques were developed. The factors mentioned in the previous paragraph narrow

the choice to certain bacteria and fungi. Further restrictions are brought about by the selection of organisms used to test the efficacy of the antibiotic. Such specific screening methods are used to find, e.g., antifungal agents, antibiotics active against cancer or viruses, or antibiotics effective against bacteria resistant to other antibiotics.

A general overview of the successes and failures experienced during the search for antibiotics has been presented [370]. GOULDEN [371] reported that in the United States from 1955 to 1966, about 90000 synthetic compounds, 20000 plant extracts, and 120000 culture solutions of microorganisms were tested against different types of neoplasms. About 1600 substances showed sufficient activity to justify their purification; 31 fermentation products reached the first clinical test, but only five of them got as far as the second step. Of these only two products, mithramycin and streptonigrin, are clinically used today. It can be concluded that, starting with a limited number of samples, the probability of obtaining a therapeutic agent is extremely low. This also applies to other screening objectives, e.g., the search for antibiotics more effective against tuberculosis, resistant microbes, or fungi and yeasts. In order to realize a definite goal, new test methods had to be introduced. Asteromycin was discovered in the process of introducing new tests against mycoplasmas. A search for antibiotics active against bacteriophages led to the discovery of a strain producing dextrochrysin. Dienomycin was found when testing nucleotides with Wood's reagent. The leucopeptines, which are active against phytopathogenic microorganisms, gram-positive bacteria, and mycobacteria, were discovered as a result of their antiplasmin activity. Although all these antibiotics have not been approved for use, they show the importance of new screening methods.

5.2. Selection, Mutation, and Maintenance of Strains

The biological production of antibiotics is carried out predominantly by microorganisms. The discovery and isolation of the microorganism are the first steps of a long process leading to the production of the antibiotic. A yield-improvment program, a very time-consuming process, is needed to raise the yield of the strain to an economic level [372]. This is done primarily by developing optimal cultivation conditions, keeping in mind that the deep-tank and submerged methods are the only ones technically applicable. Even so, the concentration of antibiotic in the culture medium is generally not enough to start production.

For this reason *selection* is necessary. A large number of single individuals belonging to a strain are isolated. These are bred, and the antibiotic production in the cultures is quantitatively measured. New individuals with good, average, poor, or even no productivity usually develop. Hence, selection is carried out from generation to generation in an effort to develop a strain with as high an antibiotic productivity as possible, one that produces few interfering byproducts (dyes, toxins, other antibiotics, etc.), and one that remains stable over a long period of time, i.e., one whose antibiotic production does not decline.

Another technique used to obtain improved strains is *mutation.* Cultures are exposed systematically to mutagens, such as ultraviolet radiation or specific chemical compounds. The dosage is chosen so that of a very large number of treated individuals only a few survive, and these are genetically altered. The mutants obtained in this way are generally valueless. In a few cases, however, it is possible to separate a single organism that possesses properties, such as increased antibiotic production and strain stability, superior to those of the untreated strain.

Penicillin production has been perfected to such an extent that today's industrial strains produce at least 35000 U/mL of culture medium. On a smaller scale, still higher production rates have been obtained. The *maintenance* of the strain, i.e., the production, choice, testing, and storage of efficient antibiotic producers, plays a very important part in the manufacture of antibiotics. The yield of antibiotic tends to decrease through many successive rounds of selection. This tendency must be monitored using tests in culture plates, which include methods using the agar diffusion test, photocytometry, and tests in shaken erlenmeyer flasks or in small fermenters with volumes ranging from 10 to 3000 L.

5.3. Process Development Leading to Large-Scale Production

After an efficient microorganism has been found in the laboratory, the strain must be brought into large-scale production. This process, known as scaling up, is undertaken in steps and poses tremendous technical problems. Several factors are important in scaling up.

Transition to Larger Volumes. Microbial growth, begun in the laboratory on culture plates, transferred subsequently to shaking flasks having a maximum volume of one liter, and later to small industrial fermenters having a working volume of 3000 L, must ultimately be carried out in production fermenters having a total capacity of about 150 m^3 or more. The main problem involved in this process of scaling up is to modify the fermentation conditions in such a way that the same yields are obtained in the larger fermenters as in the smaller ones.

Changes in the Fermenter Geometry. The structure of the fermenter, its construction, and its dimensions, greatly affect the yield obtained. The height versus the diameter of the fermenter, the stirring and aeration systems, the cooling system (jacket, spiral, or inserted cooling), and the protection of the inlets and outlets of the fermenter against infection (pressure sealing) are all important factors in the fermentation process.

Experiments in large-scale fermenters are time consuming and expensive. Critical comparisons of similar experiments conducted by different institutes or companies must take into account the fact that the research laboratories have only a few types of fermenters at their disposal. In addition, almost every researcher handles similar fermentation problems using different strains of a microorganism. Hence, a real comparison cannot be made and only reserved conclusions can be drawn.

Variations in the Culture Medium or Nutrient Solution. The culture medium influences the growth of the microorganism and, independently of this, the amount of antibiotic it produces. The growth media or nutrient solutions required for the prefermentation treatment, which primarily must support the rapid multiplication of the microorganisms as a monoculture, have a different composition than the nutrient solution used during fermentation. For example, in fermentation the carbon source should not be too plentiful. As a result of rapid consumption, nutrients must be resupplied to prevent a nutrient shortage.

In the choice and supply of nutrients, factors other than the achievement of an optimal antibiotic production must be considered. A nutrient suitable for improving the yield of an antibiotic may simultaneously hinder its recovery. Only an accurate comparison of the yield with the effort required, from the prefermentation treatment to the final product, can decide whether an apparently good fermentation raw material is also suitable for production. The addition or removal of certain substances has a direct effect on the antibiotic production. In the manufacture of penicillin, the addition of building blocks or precursors to the fermentation broth causes, depending on the addition, a preferred production of benzylpenicillin (addition of phenylacetic acid) or of phenoxymethylpenicillin (addition of phenoxyacetic acid). On the other hand, if chloride ions are largely removed from the culture medium, e.g., by pretreatment with silver salts [373], [374], or by ion exchange [375], the production of chlortetracycline is suppressed in favor of tetracycline. Certain inhibitors, e.g., inorganic additives, such as bromides and thiocyanates [376], and a great number of organic compounds, also suppress the production of chlortetracycline.

Variation of Other Fermentation Conditions. Strict control must be maintained during fermentation. The temperature, the pH (including the effects of nutrients and additives on the pH), the stirrer speed, the air supply, and the duration of fermentation must be monitored constantly.

Control of Fermentation by Means of Additives. A resting surface culture or a simple shaken culture contains a definite nutrient medium, which is required to support the growth of the microorganism. The growing culture eventually slows down and ceases growth, usually because the medium is spent. A submerged fermenter allows the sterile addition of additives during the course of fermentation. In this way important changes can be made. A

sterile air supply and its generally continuous distribution are vital to all aerobic microorganisms. Any interruption of the air supply must be avoided and the air must be evenly distributed throughout the fermenter. This is achieved by a mixing and air-distributing system. The air supply is often limited in large fermenters as a result of high viscosity and foaming, which makes the addition of *antifoam substances* necessary (oils, silicones). Mechanical foam destroyers can also be used but they consume large amounts of energy and are not applicable in production plants. The addition of *nutrients*, in portions or continuously, permits the supply of nutrients to be adjusted at each stage in the fermentation process. The added nutrients may be organic (e.g., sugar as the carbon source) or inorganic (e.g., ammonia as the nitrogen source).

The pH can be adjusted during fermentation by the addition of *acid or base*. However, experience has shown that the addition of certain nutrients causes a simultaneous change in the pH. Slower adjustment is physiologically preferable in this case. The addition of sugar often causes a fall in pH because of carbon dioxide formation; peptides and amino acids cause an increase in pH (via the formation of ammonia or other nitrogen bases). The addition of *building blocks, precursors, and inhibitors* during the course of fermentation has proved useful, especially for the production of penicillins and tetracyclines. In such cases, a single addition at the beginning of the fermentation procedure leads to a concentration toxic to the fungus, but because of its rapid consumption, the precursor concentration should be maintained at a certain level.

Measurement and Control Techniques; Analytical Measurements during Fermentation. The process of fermentation is relatively long, and the antibiotic production is very sensitive to disturbances. Precise analytical measurements and rapid and accurate control mechanisms are therefore required.

The monitoring of conditions during the course of fermentation can be divided into direct measurements in or at the fermenter and indirect testing in the laboratory of samples withdrawn at regular intervals and under sterile conditions.

The direct measurements are immediate and can sometimes be automated. For example, an electrode could be installed to monitor the foam level and automatically release an antifoam additive as required. The formation of foam is influenced, within certain limits, by changes in the air supply. Equipment to monitor and control the temperature (by adjusting the amount of cooling water) and the pH of the medium is common. Some other direct measurements are the determination of the weight of the full fermenter, e.g., with the help of a pressure gage (especially at the start of fermentation and at harvesting), the measurement and control of the stirring speed and of the air supply, and the determination of the partial pressures of oxygen and carbon dioxide in the fermenter and their concentrations in the exhaust gas.

Computer monitoring is of great help in industrial production because of the large number and size of the fermenters. The results of measurements on indirect samples, which are available after hours or a few days, also are fed into the computer.

The tendency today is to analyze numerous fermentation samples, taken at as short intervals as possible, using automatic analyzing instruments for sugar, nitrogen, phosphates, biomass, product, etc. Methods that permit the fully automatic withdrawal of samples during fermentation and the automatic transfer of the samples to different analyzers have been perfected.

Scale Down. When a fermentation procedure is carried out for the first time on an industrial scale, scale-up problems occur. After their start-up problems have been overcome, large fermenters often produce disproportionately larger amounts than the previously used smaller fermenters. It is difficult to explain this phenomenon when the same strain and approximately the same fermentation conditions are used. This leads to the scale-down problem, i.e., the problem of increasing the yield obtained from the smaller fermenter to that of the larger.

Solution of the scale-down problem is very important for the further development of a strain. A new, promising mutant or variant developed during a large-scale fermentation must first be evaluated in a small fermenter. The yield thus obtained must be comparable to the yield obtained using the original production strain, also in a small fermenter.

Continuous Fermentation. Continuous fermentation has proved to be feasible in breeding yeast, in the activated-sludge purification of waste water, and in the brewing of beer [377], [381]. Continuous industrial production of antibiotics suffers from several difficulties, and it has made little progress in displacing batch methods.

1) It is very difficult to keep the yield constant. The highly efficient strains used today tend to degenerate; i.e., the antibiotic production declines. This process makes the maintenance of the strain very important.
2) Maintaining the sterility of the fermentation environment and the additives is much more difficult than in a batch process.
3) A purely technical problem is the continuous accumulation of culture solutions; this generally necessitates continuous further processing. It therefore becomes necessary to convert a factory normally working only days into one working round-the-clock shifts.
4) The main saving in introducing a continuous process lies in the reduction of the volume of the equipment. However, the extent of space saving is directly proportional to the rate of the reaction. When the time required for fermentation is two weeks or less, then the volume of equipment required for a continuous process is scarcely less than that needed for a batch process. In addition, the construction of the equipment required for a continuous process is more complicated and more expensive.

Even in the case of antibiotics produced by fast-growing bacteria, e.g., tyrothricin [382] and gramicidin S [383], reasons 1, 2, and 3, along with higher equipment and factory costs, speak against continuous production.

5.4. Fermentation Technology

The essential prerequisites for the production of antibiotics using either submerged fermentation or other fermentation methods are the same.

1) A strain of microorganism should produce the desired antibiotic in satisfactory amounts and, as far as possible, without unwanted byproducts that are difficult to remove. It strain should be as stable as possible; i.e., the production of antibiotic should not decrease with time. The strain should also be resistant to other microorganisms, phages, etc.
2) Complete industrial facilities, which include laboratories for the preparation of inoculum and for the maintenance of the strain and vessels for the prefermentation treatment and for fermentation, must be available. The vessels must be equipped with devices such as temperature regulators, automatic foam destroyers, and appliances for the addition of nutrients and for the supply of sterile air. In addition, a sufficiently large recovery plant and enough storage space for raw materials, fermentation aids, intermediates, and finished products must be available.
3) The fermentation process and its optimal operation, the properties of the antibiotic formed, its isolation, and its efficient purification must be known in detail.
4) Analytical equipment and methods to monitor the operation of the fermentation and recovery processes and to control the raw materials, intermediates, and end products must be available.

Figure 1 shows a schematic outline of the large-scale production of penicillin, an example of a fermentation process.

5.4.1. Maintenance of the Strain and Production of Inoculum

The strain of microorganism is maintained as a pure culture in a microbiological laboratory. The underlying principle is preservation; i.e., a form of the microorganism that is as stable as possible must always be available. The microorganism is stored in a large number of small, separate ampules or vials that are used successively. Cultures of good colonies are constantly restarted so that the strain is never depleted.

If the microorganism forms spores, its storage is relatively easy. The spores, a resting form, are dried, usually mixed with sterile soil, and stored in ampules. Spores can be stored for months or years.

The application of frozen inoculum, stored in the vegetative state, has advantages. This form

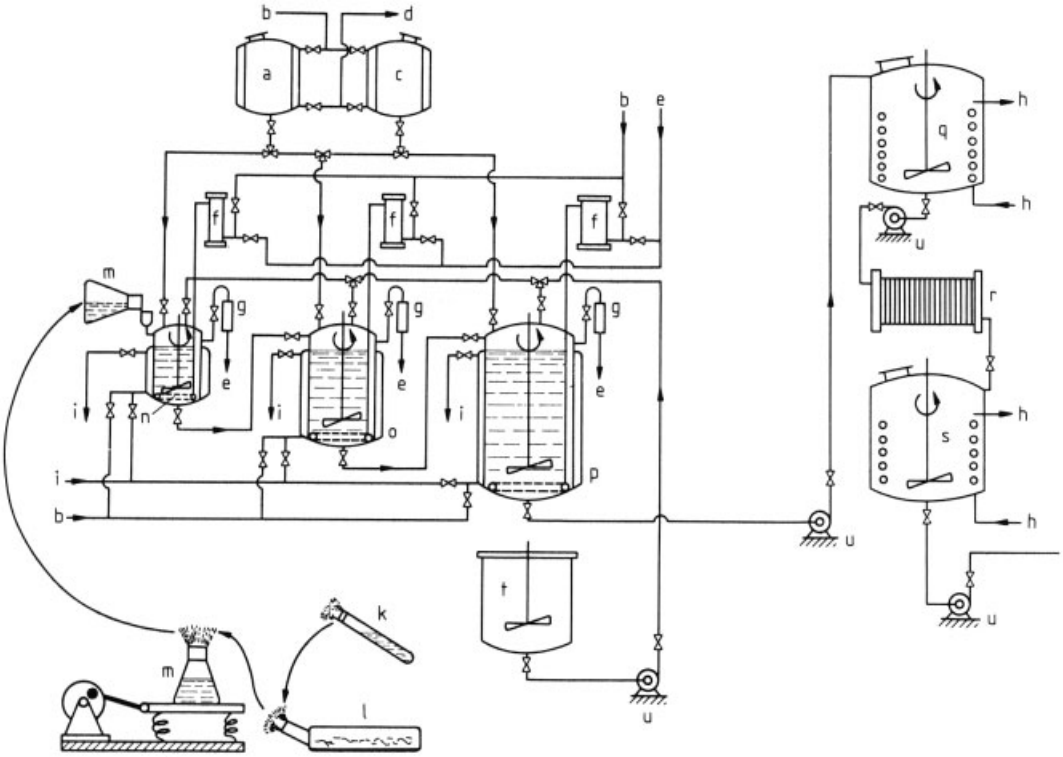

Figure 1. Schematic outline of the manufacture of penicillin
a) Antifoam substance; b) Steam; c) Precursors for penicillin formation; d) Condensate; e) Air; f) Air filter; g) Air-flow recorder; h) Cooling brine; i) Cooling water; k) Spore culture (filled into l); l) Fungal culture with spores; m) Spore suspension; n) Prefermenter (inoculation culture for o); o) Intermediate fermenter (inoculation culture for p); p) Production fermenter; q) Cooling tank; r) Filtration unit; s) Filtrate container; t) Starting vessel for nutrient solution; u) Pump.

is easy to prepare in large amounts and germination is no longer necessary. However, the storage times are limited.

The inoculum for penicillin fermentation is produced by placing the spore-containing soil in a sterile agar nutrient medium in Roux bottles and incubating at 24 °C. The spores germinate in one or two days (vegetative form). A rich mycelium network is formed, from which new spores develop in a few days. These young, freshly formed spores are removed from the fungal network under sterile conditions and with water or normal saline. They are then transferred to erlenmeyer flasks containing a suitable sterile nutrient solution. The suspension of spores is shaken at 24 °C, enabling them to undergo multiplication.

The inoculum is transferred to another shaken flask and is allowed to grow in nutrient solution until a submerged culture can be started. The next steps (Fig. 1) lead to rapid growth and an increase in volume until finally enough mass of mycelium is obtained to inoculate the production fermenter.

Besides breeding the inoculation material, the microbiology laboratory has the equally important task of guaranteeing the maintenance and care of the strain, which insures a steady supply of a microorganism with a constant efficiency. If the laboratory limited itself to breeding, storage, and regular reinoculation, the antibiotic activity of the fermentation cultures would very soon decrease because these highly productive strains tend to mutate and degenerate. To avoid a decrease in the antibiotic production in industrial fermentation, efficient strains must be subjected annually to several thousand single selections, and the resultant colonies must be tested. The single strains thereby isolated show considerable differences in their stability, i.e., in

their tendency to develop into good or bad producers or even into nonproducers. Only after the minimum number of selections necessary for the maintenance of a strain has been exceeded, do the chances of surpassing the efficiency of the original strain increase. Then an improvement in the factory productivity becomes possible.

5.4.2. Treatment Before and During Fermentation

The manufacture of antibiotics by means of fermentation is always carried out in closed, sterile vessels constructed of stainless steel or of steel lined with stainless steel. Figure 2 shows a typical construction.

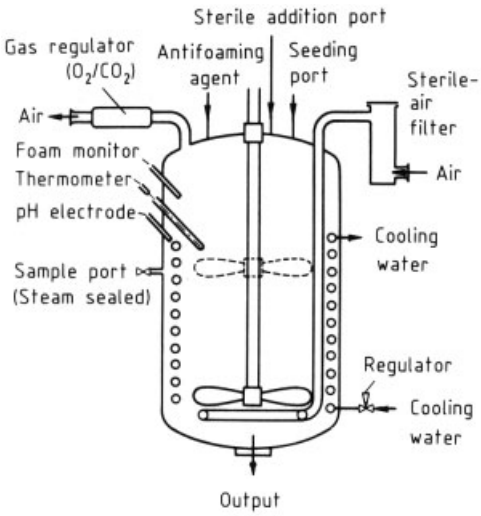

Figure 2. Schematic outline of a fermenter

The supply of sterile air for fermentation is very important. Foreign organisms are filtered from the air by means of glass wool, a filtering candle, or other methods.

The composition of the nutrient solutions must meet the nutritional needs of the microorganism; these needs vary depending on the stage of fermentation. The solutions are produced in separate vessels and are sterilized therein, in the fermenter itself, or in a continuous-flow heat exchanger. This heater is also used to sterilize the additives [380].

The amounts of raw material required may only be transported and stored in silos. Hydraulic transport and weighing with a pressure gage have replaced conventional methods, and only minor additives, e.g., trace element salts, are weighed in the normal way.

Balance Studies. The balance of energy and materials in the particular case of benzylpenicillin has been described [384] The manufacture of 100 kg of the sodium salt of penicillin requires 1.2 t carbohydrates, 60 kg animal and vegetable fats, 770 kg cornsteep liquor, 220 kg inorganic compounds (buffer, sources of sulfur and phosphorus), and 100 kg phenylacetic acid as precursors. The amount of product and the distribution of energy are shown in Table 1.

Table 1. Balance of energy and materials

	Mass, kg	Mass fraction, %	Energy distribution ∗, MJ
Raw materials	2350	–	–
Products			
benzylpenicillin sodium	100	4	2453
Fungal mycelium	825	35	11744
Remaining substance in culture medium	660	28	8051
Carbon dioxide ∗∗	765	33	13176
Total			35424

∗ as heat of formation or combustion.
∗∗ Heat of combustion released during transition to carbon dioxide (carried off with the spent air or cooling water).

The energy requirements for the production of 100 kg of benzylpenicillin sodium are:

Electrical energy:	10.8 GJ (mainly for stirring)
Steam:	4 t (sterilization, sealing)
Fermentation air:	50000 m^3 (at STP)
Cooling water:	900 m^3

Waste Materials. The accumulation of substantial amounts of fermentation waste materials, such as the fungal mycelium and the culture solution freed of antibiotic, is a real problem. After the removal of organic solutions, e.g., by distillation (stripping), the spent medium must be fed into a biological water treatment plant. Seepage is no longer allowed.

The fungal mycelium can be processed or disposed of in several ways.

1) It can be fed to animals, directly or after drying. The proceeds cover only a part of the costs, especially if the mycelium has been dried. Also, the presence of antibiotics in the

filter cake must be avoided, and filtering aids, e.g., kieselguhr or activated charcoal, may not be used.
2) It can be incinerated after the addition of liquid fuel; this is a clean but very expensive procedure.
3) It can be disposed of by dumping and humus production along with sludge depositing. This alternative often must be considered, although it entails high transportation costs.
4) It can be recycled. The mycelium can be used, directly or after intermediate processing, as a raw material for further fermentation. Mycelium is usually used as a nutrient for another microorganism. This approach is economically attractive (see Table 2 [385–392]).

Control of the Fermentation Process. At regular intervals of several hours, samples of the culture solution are withdrawn through the sample port for analysis. Important data are obtained by means of chemical, physical, and biological tests. The values are plotted and curves that present a good picture of the fermentation process are obtained.

The most important analyses are:

1) Determination of the amount of antibiotic. (Biological assay is described on 7.)
2) Determination of the weight of the mycelium as an indication of the growth of the microorganism. After inoculation of the fermenter and an initial slow phase, rapid multiplication occurs. This slows down later and finally almost comes to a standstill. The point in time at which the antibiotic production decreases and falls short of economic viability can be empirically determined. At about this point the fermentation is interrupted and the culture harvested.
3) Microscopic control of the growth of the microorganism.
4) Sterility tests, i.e., tests for the absence of foreign microorganisms.
5) Measurement and correction of the pH of the culture.
6) Determination of sugar. Figures 3 A and 3 B graphically show two possibilities. Figure 3 A shows the sugar consumption. The amount of sugar consumed, plotted against time, is the difference between the total amount of sugar added and the analytically determined sugar concentration at that particular time. Figure 3 B shows the sugar content of the nutrient solution. Here, the shape of the curve is a measure of the sugar consumption (provided no more sugar is added).
7) Determination of nitrogen in the mycelium and in the culture solution, possibly combined with the addition of a nitrogen-containing nutrient solution.

6. Isolation and Purification of Antibiotics; Quality Specifications

6.1. Isolation

When fermentation is completed, i.e., when a sufficiently high amount of antibiotic has accumulated in the culture solution, the antibiotic must be separated from the spent medium. The contents of the fermenter are transferred to a harvesting tank so that the fermenter can be turned immediately to the production of the next batch. The aeration is stopped, the solution cooled if necessary, and recovery of the product is begun as soon as possible. If permitted, a disinfectant, e.g., formaldehyde, is added or heat is applied to prevent further proliferation of the microorganism.

The recovery of the antibiotic can be carried out in several different ways, depending not only on its properties, but also on its subsequent processing.

Drying Process. Technically speaking, the easiest and cheapest process is to dry the entire culture, the culture filtrate, or the filter cake. Drying is employed on a large scale only in the manufacture of antibiotics used to supplement animal feed, e.g., tetracyclines and moenomycin (flavophospholipol, flavomycin) or salinomycin. Spray drying is the method most often used. Other methods, such as roller drying (possibly under vacuum) are also used. It is advisable to concentrate the solution, e.g., using a downdraft evaporator, before it is actually dried. In any case, the antibiotic must be resistant to higher temperatures. Because of its very short heating time, the spray-drying method is one of the most gentle procedures.

Table 2. Fermentation residues used as raw material for further fermentation

Fermentation residue	Condition	Raw material for subsequent fermentation	Reference
Penicillium (from benzylpenicillin)	moist mycelium	oxytetracycline	[385]
	moist mycelium	chlortetracycline	[386]
	moist mycelium	streptomycin, vitamin B_{12},	[387]
		or riboflavin	
	mycelium hydrolyzate	phenoxymethylpenicillin	
Penicillium or other mold mycelia	mycelium hydrolyzate	nutrient medium (e.g., for	[388]
from fermentation		*Lactobacillus bifidus*)	
Penicillium (from benzylpenicillin)	moist mycelium	calcium gluconate	[389]
Streptomyces (from tetracyclines)	moist mycelium	tetracyclines, etc.	[390]
Penicillium, *Aspergillus*, *Rhizopus*,	moist mycelium	nisin, in connection with lactic	[391]
or yeast		acid fermentation (silage)	
Penicillium, *Aspergillus*, *Actinomyces*, *Rhizopus*, yeasts, and activated sludge (from water treatment plants)	moist or dry mycelium	moenomycin (flavophospholipol)	[392]

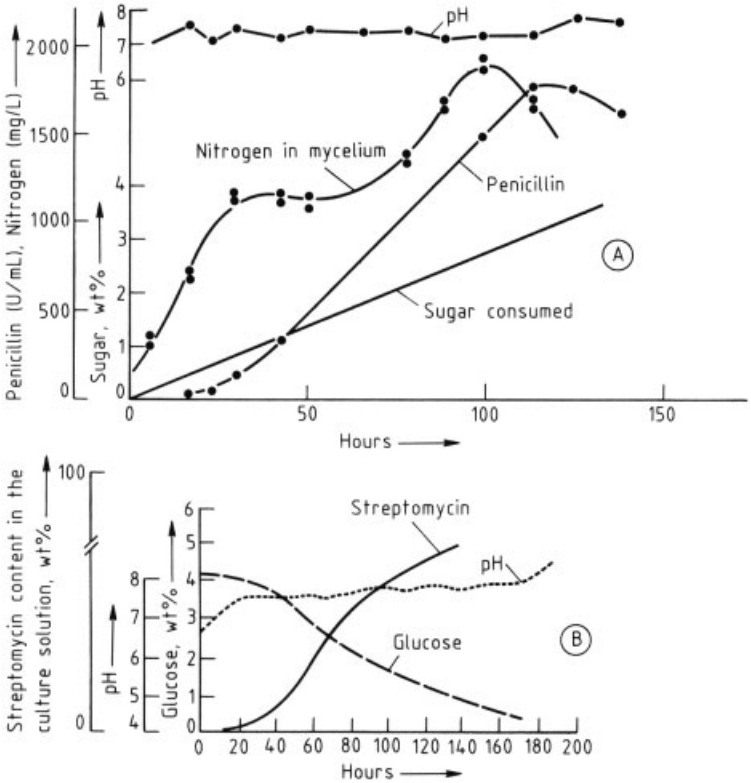

Figure 3. A. Penicillin formation with continuous glucose addition after development of the mycelium [393]
B. Streptomycin formation by *Streptomyces griseus* [394]

Filtration Followed by Extraction and Precipitation. The mycelium is separated from the liquid medium by passing the entire culture solution through a filter press, using filtering aids if necessary. A rotating filter can also be used e.g., an Oliver filter, which has three zones, intended for suction, washing, and peeling. If the culture solutions contain small amounts of mycelium, separation can also be carried out in a centrifuge.

Extraction is the method used to separate most antibiotics contained in the filtrate. A classic example is the extraction of benzylpenicillin (and phenoxymethylpenicillin) with butyl acetate (Fig. 4). It leads to a 120- to 150-fold enrichment. The penicillins are then precipitated from the extract as salts. Only those organic bases that preferentially form sparingly soluble salts with penicillin G or V but highly soluble salts with other penicillins can be used for precipitation from either water or organic solvents.

Tertiary morpholines, *N*-ethylhexahydropicoline and *N*-ethylpyrrolidine, besides *N*-ethylpiperidine, can be used for precipitation from butyl acetate, amyl acetate, and similar esters. *N*-Ethylhexamethyleneimine is used for precipitation from chloroform.

The salts thus obtained generally are easily crystallized easily and are quite stable in the dry state. They can be stored until further production steps are carried out to give the product that is used in clinical practice.

Filtration and Direct Precipitation. After filtration the aqueous culture filtrate can be subjected to direct precipitation. This method was once important in the isolation of streptomycin as a highly insoluble, colored salt, but this use has long been abandoned. Direct precipitation from the culture filtrate is of interest now because the amount of antibiotic produced by today's highly developed, efficient strains is so large that the traces of antibiotic remaining in the aqueous solution after precipitation are negligible. This method has acquired importance in the isolation of, e.g., tetracycline, which can be precipitated at its isoelectric point (pH 4.8) [395]. Another example is the direct precipitation of 5-hydroxytetracycline using a long-chain quaternary ammonium salt [396].

Filtration Followed by Extraction from the Filter Cake. If the antibiotic is present entirely or almost entirely in the mycelium, its isolation is greatly facilitated by filtration, which causes a considerable decrease in the volume of material. The moist or dry filter cake bearing the product is extracted with a solvent. The resulting solution is filtered again and processed further [397]. Examples are griseofulvin and moenomycin (flavophospholipol, flavomycin).

Adsorption Methods. Adsorption on activated charcoal following filtration is no longer used industrially. However, this method is used with some success in the developmental stages of new antibiotics.

Direct adsorption, e.g., on resins, without prior filtration, is still industrially important for the separation of such basic antibiotics as streptomycin, kanamycin, neomycin, and paromomycin. Filtration of the culture solution, especially because of the slimy substances produced by actinomycetes, is laborious and requireds large amounts of filtering aids. The real breakthrough came when the adsorption material (cation exchanger) was brought into contact with an ascending stream of the culture solution, without prior filtration. In this case the antibiotic molecules leave the solution and attach themselves to the surface of the adsorption material.

6.2. Purification Techniques, Sterile End Products, Official Regulations

Antibiotics are fermentation products and are isolated either as unfinished products or as intermediates, generally solid substances of limited stability. They are purified by methods normally employed in organic chemistry, which include chromatography, crystallization, and precipitation.

A major requirement is that the antibiotics intended for parenteral administration be free of pyrogens (fever-producing substances) and histamine-like compounds. These unwanted substances can be carried over from the fermentation, but such impurities must not appear in the final product. Special purification steps, e.g., treatment with elemental chlorine (destruction of pyrogens associated with streptomycin), filtration through activated charcoal, or deep-bed adsorption filtration, must be carried out. Precise

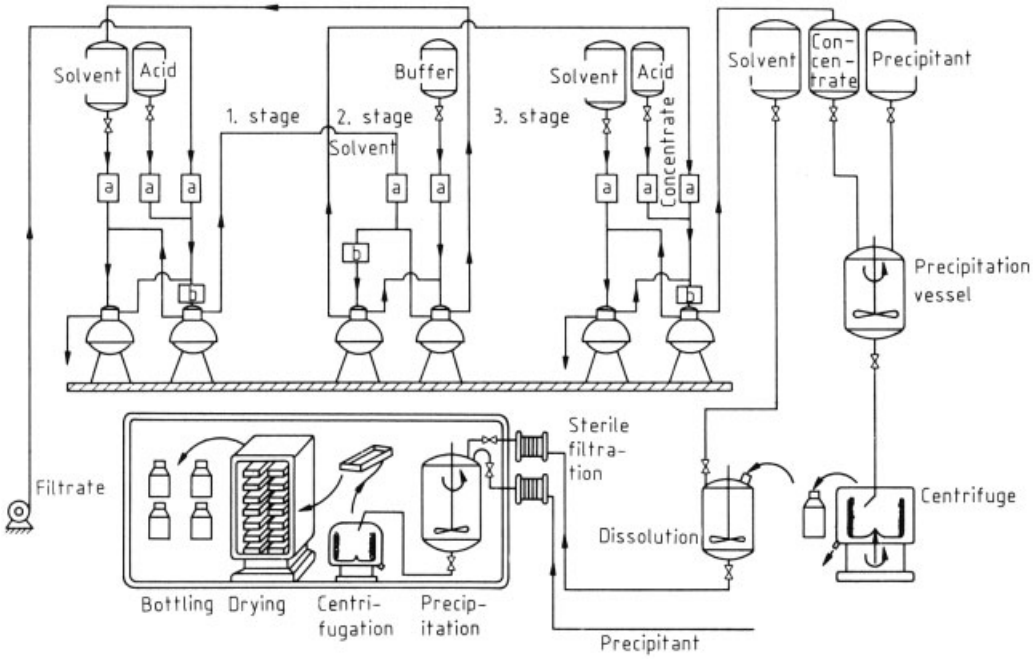

Figure 4. Penicillin extraction and the sterile final stage
a) Rotameter; b) Mixer

tests must confirm the absence of all unwanted substances.

In the manufacture, purification, and preparation of antibiotics, special measures also must be taken to prevent penicillin contamination. In the processing of active substances to give pharmaceuticals, a strict spatial separation is necessary to avoid mutual contamination through the air.

Even minute amounts of antibiotics, especially penicillin, can lead to sensitization in humans and to the formation of resistant microorganisms. These antibiotics are then ineffective in the treatment of diseases because either the patient is allergic or the pathogens have become resistant. Therefore, during the processing and filtering operations the air supply must be monitored very carefully to insure the protection of the operating personnel and to guarantee that no patient unintentionally receives even traces of penicillin with another drug.

Production of Sterile Bulk Drug Substances. If sterile bulk drug substances are required, then aseptic conditions must be maintained from the start of the process. Many antibiotics are chemically unstable and cannot tolerate sterilization by heat or other agents. Generally, the solution is sterilized before the final crystallization, precipitation, spray drying, or freeze drying. This is done by filtering through porcelain filters, sintered metal filters, layers of filter paper, or graded-porosity films.

Work is carried out in specially equipped sterile rooms, which are fully air conditioned with practically germ-free air. Air-filtering devices similar to those used for the production of fermentation air are used. The rooms are disinfected using, e.g., gaseous formaldehyde, before work is commenced. The floor is kept damp with a solution of a disinfectant (phenol, quaternary ammonium bases) mixed with glycerol, to control dust. The air pressure in these sterile rooms is higher than atmospheric pressure. This prevents the entrance of unclean air. One enters the sterile rooms only after carefully washing and donning sterile clothes and through an airlock equipped with UV lamps and foot mats soaked in disinfectant. Small objects (tools, etc.) are brought into the sterile room through smaller air locks, in which they are disinfected using intense UV radiation. Containers, e.g., stainless steel cans, are first washed thoroughly and then sterilized

in an autoclave. The autoclave is provided with one door that leads to the unsterile washing room and a second door leading to the sterile room.

The different steps in the course of further processing the antibiotics, such as centrifugation, drying, pulverization, sieving, and packaging, are performed in sterile glove boxes, which are provided with sterile air at a slightly elevated pressure and are equipped with UV irradiators.

Laminar-Flow Technique. In this technique, only one part of a clean or sterile room is maintained as a clean bench. This area is surrounded by a sterile box provided with a working access. To avoid the penetration of unclean air into this confined clean space, a displacing, turbulence-free air stream is created in the box. A continuous stream of sterile, filtered air enters the box from the top or from one side at a fixed speed. This air is distributed uniformly and is then sucked out the opposite side at the same speed. Special devices insure that there is minimum turbulence in the area of contact between the flowing, sterile air and the stationary, unclean air. Laminar-flow (LF) units can also be used for work with substances that should not escape to the outside world. In this case the air is recirculated through a filter.

Clean Packing. Antibiotics packed in bulk, in large containers, and intended for therapeutic use, usually are present in a sterile, highly pure state. Products such as tetracycline hydrochloride, intended for oral administration, must be very pure but not necessarily sterile.

Official Quality Requirements, Pharmacopoeias. The production process and the quality of the end product are subjected to rigorous official controls. The requirements have long been stipulated in the pharmacopoeia of each country and generally possess legal authority. The pharmacopoeias *European Pharmacopoeia, International Pharmacopoeia,* and *Compendium Medicamentorum* (standard pharmacopoeia for all Comecon states) are each valid in several countries.

In the United States, the influence exerted by the federal government goes beyond the determination of minimum quality standards for drugs. The Food and Drug Administration (FDA), attached to the U.S. Department of Health and Human Services, has published a Code of Federal Regulations (CFR) that is continuously supplemented with new regulations and reissued annually. The demands on the quality of a drug are substantially stricter and more comprehensive than those stipulated in the *U.S. Pharmacopeia.* In addition, detailed requirements have been established for the production and encapsulation techniques and rooms, the documentation, and the storage of raw materials, additives, intermediates, and end products. A detailed analysis of the starting materials, process controls, and tests of the end products and preparations must also be carried out. When a product not yet approved by the FDA is to be registered, preliminary tests must be conducted as well.

Food and Drug Administration officials have the right to inspect production plants regularly. Complaints can lead to the temporary closing of the plant and to the recall of certain preparations or particular batches of a preparation. In order to achieve a uniform standard in the production of drugs, the FDA [398] and the WHO [399–402] have elaborated and published fundamental rules that are now internationally called Regulations for Good Manufacturing Practices (GMPs). Drugs manufactured in countries outside the United States but imported into the United States are also subject to FDA regulations, including the GMPs, and FDA inspections. The FDA has published detailed rules in the *Federal Regulations* [403], especially for the registration of imported products.

Officials in other countries also demand a detailed description of the manufacture, quality, and safety of any drug they import. In many cases, sales depend on a prerequisite inspection of the factory, similar to the one conducted by the FDA. Canadian officials, American military forces purchasers, and the British Department of Health and Social Security all have this requirement.

7. Analytical Measurements and Quality Control

The analyses of antibiotics can be divided into two basic groups:

1) Tests during production, usually process surveillance and control.

2) Quality control, practiced as required by the WHO and the FDA. These tests have been routinely conducted by independent laboratories for a long time.

The end product, raw materials, and intermediates all are tested. The analytical measurements can be divided into chemical and physical tests, on the one hand, and biological tests, on the other. For the former, general methods used in synthetic organic chemistry are applied, along with some special methods that have been worked out for antibiotics. Fully or partially automatic techniques have been introduced to handle the large number of samples.

7.1. Microbiological Analysis

Biological Assay. Numerous microbiological methods are available to determine the amount of antibiotic present in a sample.

Agar Diffusion Test (Cylinder-Plate Method). See [404] for the original method; improvements are described in [405], [406].

A standard method has been published by the FDA [407]. For supplements, see [408] and [409].

Twenty milliliters of nutrient agar is placed in a flat-bottomed petri dish. After this solidifies, four milliliters of a second nutrient solution, seeded with the test bacteria, is poured evenly onto the first layer (at 48 °C). As soon as the second layer has solidified, six sterile stainless steel cylinders are placed on the agar, preferably using a cylinder-placing machine. To the open cylinders are added equal amounts of a standard penicillin solution containing 2.0, 1.5, 1.0, 0.5, and 0.25 U/mL. Samples of the test antibiotic solution are deposited analogously on other petri dishes.

The dishes are incubated at 37 °C for 16 – 18 h. During this time the penicillin diffuses out of the cylinders into the surrounding agar and suppresses the growth of the test organism. Thus, the cylinders are surrounded by clear zones, free of bacteria. The diameter of each zone provides an index of the activity of the penicillin preparation. The mean values obtained from 10 – 20 standard plates are used to draw a calibration curve, and the biological activity of the test solution in international units is determined using conversion tables.

Antibiotic Disk Method. This is a modification of the diffusion test. The method is widely used to determine whether a definite strain or a mixture of different microorganisms is sensitive or resistant to a given antibiotic [410–415].

The pathogen is freshly isolated from patients and used to inoculate a suitable nutrient agar plate. Filter paper disks 6 or 9 mm in diameter are placed on the petri dish before incubation. These disks are impregnated with a solution of the antibiotic. The amount is chosen so that the concentration of active substance present after diffusion into the agar medium corresponds to the level attainable in the patient (blood or tissue level). Test doses of 0.5 to 20 U are normal for a disk test of penicillins G or V.

In order to maintain a certain uniformity in the production of the nutrient medium for the disk method, the following directions for the preparation of peptone – casein agar have proved useful:

Peptone	6.0 g
Pancreatic – digested casein	4.0 g
Yeast extract	3.0 g
Meat extract	1.5 g
Dextrose	1.0 g
Agar	15.0 g

Table 3. Diameter of zones of inhibition for *Staphylococcus aureus* ATTC 6538 P

Antibiotic	Concentration per disk	Inhibition zone
Ampicillin	5 µg	26
Chloramphenicol	10 µg	20
Lincomycin	2 µg	19
Methicillin	10 µg	26
Novobiocin	10 µg	26
Oxacillin	10 µg	30
Penicillin (P)	0.5 I.U.	26
Penicillin (∗ P) [a]	20 I.U.	>40
Streptomycin [b]	10 µg	16
Tetracycline	10 µg	27

[a] Massive dose of penicillin (see Fig. 5);
[b] A pH of 8.0 is required for the optimal evaluation of the substance. Under these conditions the diameter of this zone is also 26 mm.

The components are dissolved in 1000 mL of distilled water and the pH of the liquid agar is adjusted to 6.55 after sterilization. If an agar plate thus prepared is incubated, the growth of

the microorganisms seeded on the plate can be observed from the turbidity of the agar surface. If the antibiotic on the disk is effective, a clear zone of inhibition forms surrounding the disk (Fig. 5). Table 3 shows the experimental values for the diameters of the zones of inhibition for *Staphylococcus aureus* ATTC 6538P, using the pH 6.55 nutrient medium described above.

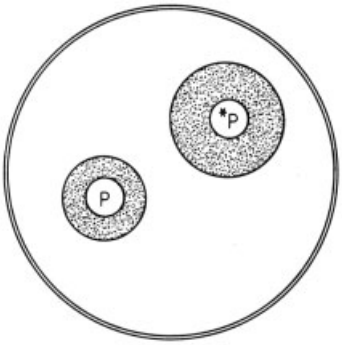

Figure 5. Antibiotic disk test
The diameter of the clear zone depends on the test dose of antibiotic: disk P (left) contained 0.5 U, disk ∗P (right) 20 U benzylpenicillin sodium.

Tube Dilution Method. Three milliliters of a nutrient solution is put into each of a row of tubes. Three milliliters of a penicillin solution with a dilution of 1 : 100 is pipetted into the first tube. After thorough mixing, three millimeters is removed and added to the next tube. After mixing, three millimeters is removed and added to the third tube, and so on. The tubes contain successively lower concentrations of the drug. If the initial penicillin solution had a 1 : 100 dilution, the first tube now contains a 1 : 200 dilution, the second a 1 : 400, and so on.

Each tube is inoculated with one drop of a day-old *staphylococcus* culture (test bacteria). After incubation for one day at 37 °C, the end point is determined, i.e., the lowest concentration of penicillin that prevents the development of turbidity. If tube 3, for example, is clear but tube 4 is turbid, the end point is calculated by multiplying the starting dilution of 1 : 100 by 2^3. Thus the bacteriostatic units are obtained, here 800 Bact. U. These can be converted into international units with the help of the penicillin sensitivity factor. For instance, if this factor is 0.04, then 25 Bact. U = 1 I. U. and the solution contains 800/25 = 32 I. U. [416–419].

Slight turbidity also can be caused by protein precipitation resulting from a change in pH. For this reason, it is advisable to add a pH indicator to the nutrient broth. Changes in the pH are then clearly visible. Phenol red and bromothymol blue generally are used. A critical assessment of the test is given in [408].

The tube dilution method is also used to determine the sensitivity of freshly isolated single strains or mixtures of pathogens to antibiotics. A comparison of the lowest inhibitory concentration of an antibiotic with the serum levels attainable in vivo indicates which antibiotic is most suitable for clinical administration.

Automatic Analyses. Automatic sample removal and preparation have speeded up analyses. However, in cases that involve special preparatory steps, such as dissolution, filtration, or extraction, full automation is not yet possible. Many references discussing automatic biotesting, are available [408], [409], [420–425].

The measurement of the clear zones in the agar diffusion test, which formerly was conducted visually, can now be performed objectively and automatically using commercially available instruments (scanning analysis systems). By connecting a laboratory calculator (e.g., HP 85) to such an instrument, the amount of antibiotic in a test solution can be calculated using reference standards and printed out directly.

7.2. Isotopically Labeled Antibiotics

Antibiotics containing a radioisotope at a definite position in the molecule are very important for scientific studies. Labeled substances can be used to trace:

1) Accumulation in specific tissues, e.g., tumors, for diagnostic purposes.
2) Metabolism of an antibiotic, i.e., tracing the metabolites and cleavage products in animal and human organs or excretions.
3) Determining the location and partial degradation of the antibiotic during further fermentation, processing, and purification steps.

Isotopically labeled antibiotics can be manufactured in two ways.

Fermentation Production. If appropriate isotopically labeled compounds are added to the culture solution during fermentation, a corresponding amount of labeled antibiotic is produced and can be isolated with the unlabeled antibiotic. This technique also is suitable for following labeled precursors, nutrients, salts, etc., during the fermentation procedure. In this way, insights into the mechanism of formation of the antibiotic in a microorganism or the mechanism of formation in the presence of an isolated enzyme system can be obtained. Benzylpenicillin, phenoxymethylpenicillin, and 6-aminopenicillanic acid have been labeled with ^{14}C and ^{35}S in this way [395, vol. 4, pp. 266, 296], [426]. Streptomycin has been labeled with ^{14}C and ^{3}H [395, vol. 4, p. 349].

Labeling. Acylation of 6-aminopenicillanic acid with a ^{14}C-labeled acid yields a product labeled in the side chain [427]. A subsequent isotopic exchange within the antibiotic molecule is also possible. However, only the easily removed ^{1}H atoms can be replaced by ^{2}H or ^{3}H.

8. Economic Aspects

Antibiotics find widespread use in human and veterinary medicine. As yet, agricultural usage is low and generally confined to Asia where they are used for antifungal treatment of rice plants, etc. Over 30 kt/a of antibiotics was produced worldwide in 1984. There are six main categories of antibiotics:

1) β-Lactam
2) Tetracycline
3) Macrolide
4) Peptide and glycopeptide
5) Aminoglycoside
6) Polyether

At least twenty other commercial antibiotics are not included within these six categories. They belong to a variety of chemical types, i.e., polyene, ansamycin, anthracycline, nucleoside, etc.

Over the past ten years, output has grown by approximately 4 % per year, with the most rapid growth in the β-lactams, macrolides, and polyethers. On the other hand, tetracyclines have presented a static or declining market, particularly for human therapy. Dollar volume sales have grown correspondingly, with a successful new human antibiotic product being defined as one commanding minimal sales of \$ 100 000 000 worldwide. β-Lactam sales account for at least half of the total human antibiotic market, which exceeds \$ 5 000 000 000.

All the categories except the polyethers find use in human medicine. The most important veterinary antibiotics belong to the tetracycline, macrolide, peptide, and polyether families although some β-lactams, aminoglycosides, and other antibiotics also have veterinary markets.

Worldwide, there are over seventy primary producers of antibiotics by fermentation. If companies involved in producing semisynthetic penicillins and cephalosporins from purchased parent antibiotics are included, the number is well over one hundred. Some companies specialize in the production of a single antibiotic, but more generally a number of different antibiotics are produced, e.g., benzylpenicillin, phenoxymethylpenicillin, cephalosporin C, oxytetracycline, and streptomycin. Large multinational pharmaceutical companies frequently operate a number of separate antibiotic fermentation plants in one or more countries. For technical reasons, it may not be possible to produce two different antibiotics in the same plant.

United States and European companies are active in all categories of antibiotics; Japanese, Chinese, and Korean producers have tended to specialize in the aminoglycosides, macrolides, anticancer drugs, semisynthetic second- and third-generation β-lactams, and agricultural antibiotics.

Some old antibiotics, which are no longer protected by patents, are traded in bulk at the prices quoted in the following paragraphs. The bulk products are purchased for use in specialities or conversion into semisynthetic drugs by companies that do not have their own fermentation facilities or whose fermentation capacities are not adequate to supply their growing needs. Bulk antibiotics are also purchased on tender by government agencies, charities, etc., for use in developing countries. There is usually a significant difference between a bulk price and that of a finished (branded or generic) speciality.

In the following paragraphs the estimated worldwide antibiotic output for the year 1985

is listed. The antibiotics are grouped into the six main categories plus "other antibiotics." The production figures include antibiotics for human and veterinary applications. The specific compounds are arranged alphabetically, not according to their commercial importance. The bulk prices are quoted only for the antibiotics that are traded; this price is much lower than the price of the finished specialty product.

β-Lactams. Total output is 10 – 20 kt/a. There are over 50 producers. The bulk price for benzylpenicillin is 25 – 30 $/kg. The following compounds are included: ampicillin, amoxycillin, carbenicillin, cefaclor, cefamandole, cefazolin, cefoperazone, cefotaxime, cefoxitin, ceftazidime, cefuroxime, cephadroxil, cephalexin, cephalosporin C, cephalothin, cephamycin C, cephradine, clavulanic acid, cloxacillin, dicloxacillin, flucloxacillin, oxacillin, benzylpenicillin, phenoxymethylpenicillin, piperacillin, and ticarcillin.

Tetracyclines. Total output is 5 – 10 kt/a. There are 30 – 40 producers. The bulk price for oxytetracycline is 25 – 30 $/kg. The following compounds are included: chlortetracycline, democlocycline, doxycycline, methacycline, minocycline, oxytetracycline, and tetracycline.

Macrolides. Total output is 3 – 5 kt/a. There are 20 – 30 producers. The bulk price for erythromycin base is 100 – 120 $/kg. The following compounds are included: erythromycin, ivermectin, josamycin, kitasamycin, midecamycin, milbemycin, miocamycin, oleandomycin, spiramycin, and tylosin.

Peptides and Glycopeptides. Total output is 2 – 3 kt/a. There are 10 – 20 producers. The bulk price for bacitracin is 15 $/kg. The following compounds are included: avoparcin, bacitracin, colistin, enramycin, gramicidin, nisin, polymixin, and thiopeptin.

Aminoglycosides. Total output was 1 – 2 kt/a. There are 20 – 30 producers. The bulk price for streptomycin is 30 $/kg. The following compounds are included: amikacin, apramycin, dibekacin, dihydrostreptomycin, gentamicin, hygromycin, kanamycin, lincomycin, neomycin, netilmicin, paromomycin, ribostamycin, sagamicin, sisomicin, streptomycin, and tobramycin.

Polyethers. Total output is 3 – 5 kt/a. There are 5 – 10 producers. The following compounds are included: laidlomycin, lasalocid, maduromycin, moenomycin, monensin, narasin, and salinomycin.

Other Antibiotics. Total output was 1 – 2 kt/a. There are 30 – 40 producers. The following compounds are included: amphotericin, anticancer (including bleomycin, daunorubicin, doxorubicin, epirubicin, and mitomycin), blasticidin, clindamycin, cycloserine, flavomycin, fusidic acid, griseofulvin, novobiocin, nystatin, pimaricin, pleuromutilin, pyrrolnitrin, rifampicin, spectinomycin, vancomycin, viomycin and virginiamycin.

9. References

General References

1. D. Gottlieb, P. D. Shaw (eds.): *Antibiotics,* vol. I. Mechanism of Action, Springer-Verlag, Berlin – Heidelberg – New York 1967.
2. D. Gottlieb, P. D. Shaw (eds.): *Antibiotics,* vol. II, Biosynthesis, Springer-Verlag, Berlin – Heidelberg – New York 1967.
3. J. W. Corcoran, F. E. Hahn (eds.): *Antibiotics,* vol. III, Mechanism of Action of Antimicrobial and Antitumor Agents, Springer-Verlag, Berlin – Heidelberg – New York 1975.
4. J. W. Corcoran (ed.): *Antibiotics,* vol. IV, Biosynthesis, Springer-Verlag, Berlin – Heidelberg – New York 1981.
5. F. E. Hahn (ed.): *Antibiotics,* vol. V, part 1, Mechanisms of Action of Antibiotic Agents, Springer-Verlag, Berlin – Heidelberg – New York 1979.
6. F. E. Hahn (ed.): *Antibiotics,* vol. V, part 2, Mechanisms of Action of Antieukaryotic and Antiviral Compounds, Springer-Verlag, Berlin – Heidelberg – New York, 1979.
7. J. S. Glasby: *Encyclopedia of Antibiotics,* 2nd ed., J. Wiley & Sons, Chichester – New York – Brisbane – Toronto 1979.
8. G. Lancini, F. Parenti: *Antibiotics. An Integrated View,* Springer-Verlag, New York – Heidelberg – Berlin 1982.

9. L. P. Garrod, H. P. Lambert, F. O'Grady: *Antibiotic and Chemotherapy,* Churchill Livingstone, Edinburgh – London – Melbourne – New York 1981.
10. E. F. Gale, E. Cundliffe, P. E. Reynolds, M. H. Richmond, M. J. Waring: *The Molecular Basis of Antibiotic Action,* 2nd ed., J. Wiley & Sons, London – New York – Sydney – Toronto 1981.
11. *Kirk-Othmer,* 3rd ed., **vol. 2,** 3.
12. J. Bérdy, A. Aszalos, M. Bostian, K. L. McNitt, *CRC Handbook of Antibiotic Compounds,* vol. I – X, CRC Press Inc., Boca Raton, Florida, 1980.
13. H. Umezawa (ed.): *Index of Antibiotics from Actinomycetes,* vol. I, II, Japan Scientific Societies Press, Tokyo, University Park Press, Baltimore 1978.
14. T. Korzylski, Z. Kowszyk-Gindifer, W. Kurylowicz: *Antibiotics Origin, Nature, and Properties,* American Society for Microbiology, Washington D.C., 1978.
15. M. J. Weinstein, G. H. Wagman: "Antibiotics Isolation, Separation and Purification," *J. Chromatogr. Libr.* **15** (1978).
16. R. Reiner: *Antibiotics,* Georg Thieme Verlag, Stuttgart 1982.
17. W. Kurylowicz (ed.): *Antibiotics (a Critical Review),* Polish Medical Publishers, Warsaw 1976.
18. P. Sammes (ed.): *Topics in Antibiotic Chemistry,* vol. I – VI, Ellis Horwood Ltd, 1980.
19. H. P. Kuemmerle (ed.): *Clinical Chemotherapy,* vol. I, II, III, Thieme-Stratton Inc., New York 1984.

Specific References

20. S. A. Waksman, *Antibiot. Chemother. (Basel, 1954 – 70)* **6** (1956) 90.
21. S. A. Waksman, *The Antibiotic Era,* The Waksman Foundation of Japan Inc., Tokyo 1975.
22. A. Fleming, *Br. J. Exp. Pathol.* **10** (1929) 226.
23. G. Domagk, *Dtsch. Med. Wochenschr.,* **61** (1935) 250.
24. E. B. Chain et al., *Lancet II* (1940) 226.
25. E. P. Abraham, P. B. Loder, in E. H. Flynn (ed.): *Cephalosporins and Penicillins,* Academic Press, New York 1972, p. 1.
26. W. Dürckheimer et al., *Angew. Chem. Int. Ed. Engl.* **24** (1985) 180.
27. R. D. G. Cooper in P. G. Sammes (ed.): *Topics in Antibiotic Chemistry,* **vol. 3,** Ellis Horwood Limited, Chichester 1980, p. 39.
28. F. A. Jung, W. R. Pilgrim, J. P. Poyser, P. J. Siret in P. G. Sammes (ed.): *Topics in Antibiotic Chemistry,* **vol. 4,** Ellis Horwood Limited, Chichester 1980, p. 11.
29. G. Albers-Schönberg et al. in R. B. Morin, M. Gorman (eds.): *Chemistry and Biology of β-Lactam Antibiotics,* vol. 1, vol. 2, vol. 3, Academic Press, New York 1982.
30. L. P. Garrod, H. P. Lambert, F. O'Grady: *Antibiotic and Chemotherapy,* 4th ed., Churchill Livingstone, Edinburgh 1973.
31. S. A. Waksman, *Jpn. J. Microbiol.* **10** (1966) 129.
32. H. Umezawa: *Enzyme Inhibitors of Microbial Origin,* University of Tokyo Press, Tokyo 1972.
33. E. P. Abraham in S. Mitsuhashi (ed.): *Beta-Lactam Antibiotics,* Japan Scientific Societies Press, Tokyo, and Springer-Verlag, Berlin – Heidelberg – New York 1981, p. 3.
34. D. C. Hodgkin, *Advancemt Sci.* **6** (1949) 85.
35. J. C. Sheehan, K. R. Henery-Logan, *J. Am. Chem. Soc.* **79** (1957) 1262; **81** (1959) 3089.
36. S. Nakatsuka, H. Tanino, Y. Kishi, *J. Am. Chem. Soc.* **97** (1975) 5008, 5010.
37. J. E. Baldwin, M. A. Christie, S. B. Haber, L. I. Kruse, *J. Am. Chem. Soc.* **98** (1976) 3045.
38. R. B. Woodward et al., *J. Am. Chem. Soc.* **88** (1966) 852.
39. E. O. Stapley, J. Birnbaum in M. Salton, G. D. Shockman (eds): *β-Lactam Antibiotics,* Academic Press, New York 1981, p. 327.
40. S. Karady et al., *Tetrahedron Lett.* 1976, 2401.
41. S. Karady et al., *J. Am. Chem. Soc.* **94** (1972) 1410.
42. L. M. Weinstock et al. *Tetrahedron Lett.* **1975,** 3979.
43. M. Shinozaki, N. Ishida, K. Iino, T. Hiraoka, *J. Chem. Soc. Chem. Commun.* 1978, 517.
44. Y. Hamashima, H. Matsumura, S. Matsuura, W. Nagara, M. Narisada, T. Yoshida in G. I. Gregory (ed.): *Recent Advances in the Chemistry of β-Lactam Antibiotics,* Special Publ. no. 38, The Royal Society of Chemistry, London 1981, p. 57.
45. W. Nagata in H. Nozaki (ed.): *Current Trends in Organic Synthesis,* Pergamon Press, Oxford 1983, p. 83.
46. A. G. Brown et al., *J. Antibiot.* **29** (1976) 668.
47. A. G. Brown et al., in J. Elks (ed.): *Recent Advances in the Chemistry of β-Lactam Antibiotics,* Special Publ. no. 28, The Chemical Society, London 1977, p. 295.

48. P. H. Bentley, P. D. Berry, G. Brooks, M. L. Gilpin, E. Hunt, I. I. Zomaya, *J. Chem. Soc. Chem. Comm.* 1977, 748.
49. M. Lang, K. Prasad, W. Holick, J. Gosteli, I. Ernest, R. B. Woodward, *J. Am. Chem. Soc.* **101** (1979) 6296.
50. J. S. Kahan et al., *Intersci. Conf. Antimicrob. Agents Chemother.,* 16th, Chicago, Pap. no. 227 (1976).
51. A. G. Brown, D. F. Corbett, A. J. Eglington, T. T. Howarth, *J. Antibiot.* **32** (1979) 961.
52. S. G. Box, J. D. Hood, S. R. Spear, *J. Antibiot.* **32** (1979) 1239.
53. D. F. Corbett, A. J. Eglington, T. T. Howarth, *J. Chem. Soc. Chem. Commun.* 1977, 953.
54. A. G. Brown, D. F. Corbett, A. J. Eglington, T. T. Howarth, *J. Chem. Soc. Chem. Comm.* 1977, 523.
55. E. O. Stapley et al., *Intersci. Conf. Antimicrob. Agents Chemother.,* 17th, New York, Pap. no. 80 (1977).
56. E. O. Stapley et al., *J. Antibiot.* **34** (1981) 628.
57. P. J. Cassidy et al., *Intersci. Conf. Antimicrob. Agents Chemother.,* 17th, New York, Pap. no. 81 (1977).
58. P. J. Cassidy et al., *J. Antibiot.* **34** (1981) 637.
59. K. Maeda et al., *J. Antibiot.* **30** (1977) 770.
60. K. Okamura et al., *J. Antibiot.* **31** (1978) 480.
61. K. Okamura et al., *J. Antibiot.* **32** (1979) 262.
62. K. Yamamoto, T. Yoshioka, Y. Kato, N. Shibamoto, Y. Shimauchi, T. Ishikura, *J. Antibiot.* **33** (1980) 796.
63. N. Shibamoto et al., *J. Antibiot.* **33** (1980) 1128.
64. T. Mori et al., *Intersci. Conf. Antimicrob. Agents Chemother.,* 20th, New Orleans, Pap. no. 165 (1980).
65. M. Nakayama et al., *J. Antibiot.* **33** (1980) 1388.
66. K. Tanaka et al. *J. Antibiot*. **34** (1981) 909.
67. D. B. R. Johnston, S. M. Schmitt, F. A. Bouffard, B. G. Christensen, *J. Am. Chem. Soc.* **100** (1978) 313.
68. F. A. Bouffard, D. B. R. Johnston, B. G. Christensen, *J. Org. Chem.* **45** (1980) 1130.
69. S. M. Schmitt, D. B. R. Johnston, B. G. Christensen, *J. Org. Chem.* **45** (1980) 1135; **45** (1980) 1142.
70. T. Kametani, S. P. Huang, S. Yokohama, Y. Suzuki, M. Ihara, *J. Am. Chem. Soc.* **102** (1980) 2060.
71. T. Kametani, T. Nagahara, Y. Suzuki, S. Yokohama, S. P. Huang, M. Ihara, *Heterocycles* **14** (1980) 403.
72. T. N. Salzmann, R. W. Ratcliffe, B. G. Christensen, F. A. Bouffard, *J. Am. Chem. Soc.* **102** (1980) 6161.
73. D. G. Melillo, I. Shinkai, T. Liu, K. Ryan, M. Sletzinger, *Tetrahedron Lett.* 1980, no. 21, 2783.
74. D. G. Melillo, T. Liu, K. Ryan, M. Sletzinger, I. Shinkai, *Tetrahedron Lett.* 1981, no. 22, 913.
75. R. J. Ponsford, R. Southgate, *J. Chem. Soc. Chem. Commun.* 1979, 846;1980, 1085.
76. M. Ohno, S. Kobayashi, T. Iimori, Y.-F. Wang, T. Izawa, *J. Am. Chem. Soc.* **103** (1981) 2405.
77. H. Kotani et al., *Agric. Biol. Chem.* **47** (1983) 1363.
78. S. Kobayashi, T. Iimori, T. Izawa, M. Ohno, *J. Am. Chem. Soc.* **103** (1981) 2406.
79. K. Okano, T. Izawa, M. Ohno, *Tetrahedron Lett.* 1983, no. 24, 217.
80. T. Iimori, Y. Takahashi, T. Izawa, S. Kobayashi, M. Ohno, *J. Am. Chem. Soc.* **105** (1983) 1659.
81. H. Aoki et al., *J. Antibiot.* **29** (1976) 492.
82. A. Imada, K. Kitano, K. Kintaka, M. Muroi, M. Asai, *Nature (London)* **289** (1981) 590.
83. R. B. Sykes et al. *Nature (London)* **291** (1981) 489.
84. C. M. Cimarusti et al., *J. Org. Chem.* **47** (1982) 179.
85. B. M. Duggar, *Ann. N.Y. Acad. Sci.* **51** (1948) 177.
86. T. Money, A. I. Scott, *Prog. Org. Chem.* **7** (1968) 1.
87. D. L. J. Clive, *Quart. Rev.* **22** (1968) 435.
88. C. W. Waller et al., *J. Am. Chem. Soc.* **74** (1952) 4980.
89. H. Muxfeldt, *Chem. Ber.* **92** (1959) 3122.
90. J. J. Goodman et al. *J. Bacteriol.* **69** (1955) 70.
91. A. Mitscher et al., *J. Am. Chem. Soc.* **92** (1970) 6070.
92. F. Arcamone, A. D. Di Marco, M. Gaetani, T. Scotti, *G. Microbiol.* **9** (1961) 83.
93. F. Arcamone in P. G. Sammes (ed.): *Topics in Antibiotic Chemistry,* **vol. 2,** Ellis Horwood Ltd., Chichester 1978, p. 99.
94. S. Neidle in P. G. Sammes (ed.): *Topics in Antibiotic Chemistry,* **vol. 2,** Ellis Horwood Ltd., Chichester 1978, p. 240.
95. S. A. Waksman: *Streptomycin,* The Williams and Wilkins Co., Baltimore 1949; *Neomycin,* Rutgers Univ. Press, New Brunswick, New Jersey, 1953.
96. L. Weinstein, N. J. Ehrenkranz in: H. Welch, F. Marti-Ibánez (eds.): *Streptomycin and DHS, Antibiotic Monographs,* no. 10, Medical Encyclopedia, New York 1958.

97. H. Schmidt et al., *Pharmazie* **23** (1968) 161.
98. J. S. Brimacombe, *Angew. Chem.* **83** (1971) 261.
99. H. Grisebach, R. Schmid, *Angew. Chem.* **84** (1972) 192.
100. H. Umezawa et al., *J. Antibiot. Ser. A* **20** (1967) 136.
101. H. Umezawa et al., *Science (Washington, D.C.)* **157** (1967) 1559.
102. H. Umezawa et al., *J. Antibiot.* **24** (1971) 485.
103. I. R. Hooper in H. Umezawa, I. R. Hooper (eds.): *Handbook of Experimental Pharmacology,* vol. 62, Springer Verlag, Heidelberg 1982.
104. S. Umezawa in S. Mitsuhashi (ed.): *Drug Action and Drug Resistance in Bacteria,* **vol. 2,** University of Tokyo Press, Tokyo 1975, p. 3.
105. J. Davies, D. I. Smith in M. Starr (ed.): *Annual Review of Microbiology,* **vol. 32,** Annual Reviews Inc., California 1978, p. 469.
106. H. Umezawa in R. S. Tipson, D. Horton (eds.): *Advances in Carbohydrate Chemistry and Biochemistry,* **vol. 30,** Academic Press, New York 1974, p. 183.
107. H. Umezawa, in S. Mitsuhashi (ed.): *Drug Action and Drug Resistance in Bacteria,* **vol. 2,** University of Tokyo Press, Tokyo 1975, p. 211.
108. H. Umezawa, S. Kondo in H. Umezawa, I. R. Hooper (eds.): *Handbook of Experimental Pharmacology,* vol. 62, Springer Verlag, Heidelberg 1982.
109. R. U. Lemieux, M. L. Wolfrom in W. W. Pigman, M. L. Wolfrom (eds.): *Advances in Carbohydrate Chemistry,* **vol. 3,** Academic Press, New York 1948, p. 337.
110. S. Neidle, D. Rogers, M. B. Hursthouse, *Tetrahedron Lett.* 1968, 4725.
111. S. Umezawa et al., *J. Antibiot.* **27** (1974) 997.
112. T. G. Cochran, D. J. Abraham, L. L. Martin, *J. Chem. Soc. Chem. Commun.* 1972, 494.
113. S. Hanessian, R. Roy, *J. Am. Chem. Soc.* **101** (1979) 5839.
114. D. R. White et al., *Tetrahedron Lett.* 1979, 2737.
115. K. L. Rinehart Jr.: *The Neomycins and Related Antibiotics,* J. Wiley and Sons, New York 1964.
116. S. Umezawa, Y. Nishimura, *J. Antibiot.* **30** (1977) 189.
117. M. Hichens, K. L. Rinehart Jr., *J. Am. Chem. Soc.* **85** (1963) 1547.
118. S. Tatsuoka, S. Horii, *Proc. Jpn. Acad.* **39** (1963) 314.
119. T. Ito et al., *Agric. Biol. Chem.* **34** (1970) 980.
120. G. Koyama et al., *Tetrahedron Lett.* 1968, 1875.
121. S. Umezawa et al., *J. Antibiot.* **21** (1968) 367.
122. D. J. Cooper et al., *J. Chem. Soc. C* 1971, 3126.
123. H. Reimann et al., *J. Org. Chem.,* **39** (1974) 1451.
124. S. Fukatsu, *J. Antibiot.* **32 (suppl),** (1979) S-178.
125. H. Kawaguchi et al., *J. Antibiot.* **25** (1972) 695.
126. T. L. Nagabushan et al., *J. Am. Chem. Soc.* **100** (1978) 5253.
127. T. Tsuchiya, Y. Takagi, S. Umezawa, *Tetrahedron Lett.* 1979, 4951.
128. T. Deushi et al., *J. Antibiot.* **32** (1979) 173; **32** (1979) 1061.
129. S. Inouye et al., *J. Antibiot.* **32** (1979) 1354.
130. T. Nara et al., *J. Antibiot.* **30** (1977) 533.
131. Y. Okami et al., *J. Antibiot.* **32** (1979) 964.
132. A. Bloch in E. J. Arièns (ed.): *Drug Design,* **vol. 4,** Academic Press, New York 1973, p. 286.
133. M. Ohno in J. M. Cassady, J. D. Douros (eds.): *Anticancer Agents Based on Natural Product Models,* Academic Press, New York 1980, p. 73.
134. A. Bloch in M. Grayson (ed.): *Antibiotics, Chemo-therapeutics, and Antibacterial Agents for Disease Control,* J. Wiley & Sons, New York 1982.
135. H. E. Bergy, R. R. Herr, *Antimicrob. Agents Chemother. (1961 – 70)* 1966, 625.
136. K. Mizuno, M. Tsujino, M. Takada, M. Hayashi, K. Atsumi, K. Asano, T. Matsuda, *J. Antibiot.* **27** (1974) 775.
137. T. Tsuruoka et al., *Meiji Seika Kenkyu Nempo* **9** (1967) 17.
138. P. W. K. Woo, H. W. Dion, et al., *J. Heterocycl. Chem.* **11** (1974) no. 4, 641 – 3.
139. H. Nakamura, G. Koyama, Y. Iitaka, M. Ohno, N. Yagisawa, S. Kondo, K. Maeda, H. Umezawa, *J. Am. Chem. Soc.* **96** (1974) 4327.
140. E. Cherbuliez, K. Bernhard, *Helv. Chim. Acta* **15** (1932) 464.
141. L. Ehrenberg et al., *Sven. Kem. Tidskr.* **58** (1946) 269.
142. K. Isono, S. Suzuki, *J. Antibiot. Ser. A* **13** (1960) 270.
143. H. Nishimura, K. Katagiri, K. Sato, M. Mayama, N. Shimaoka, *J. Antibiot. Ser. A* **9** (1956) 60.
144. K. G. Cunningham, S. A. Hutchinson, W. Manson, F. S. Spring, *J. Chem. Soc.* 1951, 2299.

145. K. Isono, K. Asahi, S. Suzuki, *J. Am. Chem. Soc.* **91** (1969) 7490.
146. K. Isono, S. Suzuki, *Heterocycles* **13** (1979) 333.
147. M. Uramoto, M. Matsuoka, J. G. Liehr, J. A. McCloskey, K. Isono, *Agric. Biol. Chem.* **45** (1981) 1901.
148. M. Hori, E. Ito, T. Takita, G. Koyama, T. Takeuchi, H. Umezawa, *J. Antibiot. Ser. A* **17** (1964) 96.
149. S. Aizawa, T. Hidaka, N. Otake, H. Yonehara, K. Isono, N. Igarashi, S. Suzuki, *Agric. Biol. Chem.* **29** (1965) 375.
150. G. Koyama, H. Umezawa, *J. Antibiot. Ser. A.* **18** (1965) 175.
151. T. Sawa, Y. Fukagawa, I. Homma, T. Wakashiro, T. Takeuchi, M. Hori, T. Komai, *J. Antibiot.* **21** (1968) 334.
152. H. Nishimura, M. Mayama, Y. Komatsu, H. Kato, N. Shimaoka, Y. Tanaka, *J. Antibiot. Ser. A* **17** (1964) 148.
153. T. Haneishi et al., *J. Antibiot.* **24** (1971) 797.
154. T. Kusaka, H. Yamamoto, M. Shibata, M. Muroi, T. Kishi, K. Mizuno, *J. Antibiot.* **21** (1968) 255.
155. S. Yaginuma, N. Muto, M. Tsujino, Y. Sudate, M. Hayashi, M. Otani, *J. Antibiot.* **34** (1981) 359.
156. M. Arita, K. Adachi, Y. Ito, H. Sawai, M. Ohno, *J. Am. Chem. Soc.* **105** (1983) 4049.
157. S. Takeuchi et al. *J. Antibiot. Ser. A* **11** (1958) 1.
158. S. Masamune, G. S. Bates, J. W. Corcoran, *Angew. Chem. Int. Ed. Engl.* **16** (1977) 585.
159. S. Masamune, J. W. Ellingboe, W. Choy, *J. Am. Chem. Soc.* **104** (1982) 5526.
160. M. N. Donin et al., *Antibiot. Annu.* 1953 – 1954, 179.
161. C. Djerassi, J. A. Zderic, *J. Am. Chem. Soc.* **78** (1956) 2907.
162. J. M. McGuire et al., *Antibiot. Chemother. (Washington D.C.)* **2** (1952) 281.
163. R. A. LeMahieu, J. F. Blount, R. W. Kierstead, *J. Antibiot.* **28** (1975) 705, and references cited therein.
164. S. Omura, A. Nakagawa, *J. Antibiot.* **28** (1975) 401.
165. S. Omura, A. Nakagawa, M. Machida, H. Imai, *Tetrahedron Lett.* 1977, 1045.
166. T. Hata et al., *J. Antibiot. Ser. A* **6** (1953) 87.
167. T. Watanabe, T. Fujii, K. Satake, *J. Biochem. (Tokyo)* **50** (1961) 197.
168. S. Pinnert-Sindico et al., *Antibiot. Annu.* **2** (1954 – 55) 274.
169. G. H. Wagman, J. A. Waitz, J. Marquez, A. Murawski, E. M. Oden, R. T. Testa, M. J. Weinstein, *J. Antibiot.* **25** (1972) 641.
170. H. Reimann, R. S. Jaret, *J. Chem. Soc. Chem. Commun.* 1972, 1270.
171. H. Koshiyama, H. Tsukiura, K. Fujisawa, M. Konishi, M. Hatori, K. Tomita, H. Kawaguchi, *J. Antibiot.* **22** (1969) 61.
172. H. Tsukiura, M. Konishi, M. Saka, T. Naito, H. Kawaguchi, *J. Antibiot.* **22** (1969) 89.
173. K. Hatano, E. Higashide, M. Shibata, *J. Antibiot.* **29** (1976) 1163.
174. T. Kishi, S. Harada, H. Yamana, A. Miyake, *J. Antibiot.* **29** (1976) 1171.
175. Y. Shimauchi, K. Kubo, K. Osumi, K. Okamura, Y. Fukagawa, T. Ishikura, *J. Antibiot.* **32** (1979) 878.
176. R. B. Woodward, *Angew. Chem.* **69** (1957) 50.
177. R. B. Woodward, L. S. Weiler, P. C. Dutta, *J. Am. Chem. Soc.* **87** (1965) 4662.
178. A. Kinumaki, M. Suzuki, *J. Antibiot.* **25** (1972) 480.
179. R. B. Morin, M. Gorman, P. L. Hamill, P. V. Demarco, *Tetrahedron Lett.* 1970, 4737.
180. S. Bhuwapathanapun, P. Gray, *J. Antibiot.* **30** (1977) 673.
181. S. Omura, H. Matsubara, A. Nakagawa, A. Furusaki, T. Matsumoto, *J. Antibiot.* **33** (1980) 915.
182. M. Hayashi, M. Ohno, S. Satoi, *J. Chem. Soc. Chem. Commun.* 1980, 119.
183. P. W. K. Woo, H. W. Dion, Q. R. Bartz, *Chem. Comun.* **86** (1964) 2724, 2726.
184. F. J. Antosz in M. Grayson (ed.): *Antibiotics, Chemotherapeutics and Antibacterial Agents for Disease Control,* J. Wiley & Sons, New York 1982, p. 76.
185. A. Lüttringhaus, H. Gralheer, *Justus Liebigs Ann. Chem.* **550** (1942) 67.
186. K. L. Rinehart Jr., *Acc. Chem. Res.* **5** (1972) 57.
187. P. Sensi, *Chim. Ind. (Milan)* **51** (1969) 811.
188. P. Sensi, *Res. Prog. Org. Biol. Med. Chem.* **1** (1964) 337.
189. M. T. Timbal, *Antibiot. Annu.* 1959 – 1960, 271.
190. N. Bergamini, G. Fowst, *Arzneim. Forsch.* **15** (1965) 951.
191. S. Riva, L. G. Silvestri, *Annu. Rev. Microbiol.* **26** (1972) 199.
192. D. Perlman in M. Grayson (ed.): *Antibiotics, Chemotherapeutics, and Antibacterial Agents for Disease Control,* J. Wiley & Sons, New York 1982, p. 210.
193. H. Umezawa, K. Maeda, T. Takeuchi, Y. Okami, *J. Antibiot. Ser. A.* **19** (1966) 200.

194. T. Takita, Y. Muraoka, T. Nakatani, A. Fujii, Y. Umezawa, H. Naganawa, H. Umezawa, *J. Antibiot.* **31** (1978) 801.
195. T. Takita et al., *Tetrahedron Lett.* 1982, no. 23, 521, and references cited therein.
196. H. Umezawa in J. M. Cassady, J. D. Douros (ed.): *Anticancer Agents Based on Natural Product Models,* Academic Press, New York 1980, p. 147.
197. H. Kawaguchi et al., *J. Antibiot.* **30** (1977) 779.
198. M. Konishi et al., *J. Antibiot.* **30** (1977) 789.
199. G. F. Gauze, M. G. Brazhnikova, *Lancet* **247** (1944) 715.
200. T. S. G. Jones, *Ann. N.Y. Acad. Sci.* **51** (1949) 909.
201. I. M. Lockhart, E. P. Abraham, *Biochem. J.* **58** (1954) 633.
202. A. Arriagada et al., *Br. J. Exp. Pathol.* **30** (1949) 425.
203. C. Ressler, D. V. Kashelikar, *J. Am. Chem. Soc.* **88** (1966) 2025.
204. H. Brockmann, *Angew. Chem.* **72** (1960) 939.
205. A. J. Whiffen, J. N. Bohono, R. L. Emerson, *J. Bacteriol.* **52** (1946) 610.
206. G. M. Schull, J. L. Sardinas, *Antibiot. Chemother.* **5** (1955) 398.
207. S. Takeuchi, H. Yonehara, H. Umezawa, *J. Antibiot. Ser. A* **12** (1959) 195.
208. H. Umezawa et al., *J. Antibiot. Ser. A* **6** (1953) 101.
209. H. Hoeksema, J. L. Johnson, J. W. Hinman, *J. Am. Chem. Soc.* **77** (1955) 6710.
210. A. E. Oxford, H. Raistrick, P. Simonart, *Biochem. J.* **33** (1939) 240.
211. J. F. Grove, J. MacMillan, T. P. C. Mulholl, M. A. T. Rogers, *J. Chem. Soc.* 1952, 3977.
212. J. Ehrlich, Q. R. Bartz, R. M. Smith, D. A. Joslyn, P. R. Burkholder, *Science (Washington D.C.)* **106** (1947) 417.
213. G. Keiser, *Dtsch. Med. Wochenschr.* **96** (1971) 1544.
214. K. Shirahata, N. Hirayama, *J. Am. Chem. Soc.* **105** (1983) 7199.
215. T. Hata, Y. Sano, R. Sugawara, A. Matsume, K. Kanamori, T. Shima, T. Hoshi, *J. Antibiot. Ser. A* **9** (1956) 141.
216. S. Wakaki, H. Marumo, K. Tomioka, G. Shimizu, E. Kato, H. Kamada, S. Kudo, Y. Fujimoto, *Antibiot. Chemother. (Washington D.C.)* **8** (1958) 228.
217. J. Vodrazka, *Pharm. Ind.* **32** (1970) 951.
218. A. Agtarap et al., *J. Am. Chem. Soc.* **89** (1967) 5737.
219. D. Hendlin et al., *Science (Washington D.C.)* **166** (1969) 122.
220. W. O. Godtfredsen, S. Jahnsen, H. Lorck, K. Roholt, L. Tybring, *Nature (London)* **193** (1962) 987.
221. S. Takahashi, T. Ohashi, Y. Kii, H. Kumagai, H. Yamada, *J. Ferment. Technol.* **57** (1979) 328.
222. T. Ohashi, S. Takahashi, T. Nagamichi, K. Yoneda, H. Yamada, *Agric. Biol. Chem.* **45** (1981) 831.
223. K. C. Nicolaou, W.-M. Dai, *Angew. Chem. Int. Ed. Engl.* **30** (1990) 1387.
224. N. Ishida, K. Miyazaki, K. Kumagai, M. Rikimaru, *J. Antibiot.* **18** (1965) 68.
225. K. Edo et al., *Tetrahedron Lett.* **26** (1985) 331.
226. M. D. Lee et al., *J. Am. Chem. Soc.* **109** (1987) 3464.
227. M. D. Lee et al., *J. Am. Chem. Soc.* **109** (1987) 3466.
228. J. Golik et al., *J. Am. Chem. Soc.* **109** (1987) 3461.
229. J. Golik et al., *J. Am. Chem. Soc.* **109** (1987) 3462.
230. M. Konishi et al., *J. Antibiot.* **42** (1989) 1449.
231. J. Hu et al., *J. Antibiot.* **41** (1988) 1575.
232. T. Otani, Y. Minami, T. Marunaka, R. Zhang, M.-Y. Xie, *J. Antibiot.* **41** (1988) 1580.
233. K. S. Lam et al., *J. Antibiot.* **44** (1991) 472.
234. S. J. Hofsteader, J. A. Matson, A. R. Malacko, H. Marquardt, *J. Antibiot.* **45** (1992) 1250.
235. N. Zein et al., *Proc. Natl. Acad. Sci. USA.* **90** (1993) 8009.
236. J. E. Leet et al., *J. Am. Chem. Soc.* **115** (1993) 8432.
237. M. Hanada et al., *J. Antibiot.* **44** (1991) 403.
238. A. G. Meyer, *Tetrahedron Lett.* **28** (1987) 4493.
239. H. Maeda, K. Edo, N. Ishida (eds.): *Neocarzinostatin. The Past, Presend, and Future of an Anticancer Drug,* Springer-Verlag, Heidelberg 1997.
240. N. P. Sullivan et al., *Science (Washington, D.C.)* **107** (1948) 169.
241. J. L. Szabo et al., *Antibiot. Chemother. (Washington, D.C.)* **1** (1951) 499.
242. E. Brandl et al., *Wien. Med. Wochenschr.* 1953, 602.
243. K. H. Spitzy, *Wien. Klin. Wochenschr.* **65** (1953) 583.
244. Y. G. Perron et al., *Antibiot. Ann.* 1959 – 1960, 107; *J. Am. Chem. Soc.* **82** (1960) 3934.
245. G. M. Williamson et al., *Lancet 1961,* **vol. I,** 847; M. Nagley, *Lancet 1961,* **vol. I,** 851.
246. E. T. Kundsen et al., *Br. Med. J.* **3** (1967) 75.

247. J. P. Clayton et al., *J. Med. Chem.* **18** (1975) 172.
248. K. Butler, *Del. Med. J.* **43** (1971) 366. A. R. English et al., *Antimicrob. Agents Chemother.* **1** (1972) 185.
249. K. Tsuchiya et al., *J. Antibiot.* **24** (1971) 607.
250. H. Neu, E. B. Windshell, *Appl. Microbiol.* **21** (1971) 66.
251. E. T. Kundsen, G. N. Rolinson, *Br. Med. J.* **2** (1960) 700.
252. F. P. Doyle et al., *Nature (London)* **192** (1961) 1183.
253. E. T. Kundsen et al., *Lancet 1962,* **vol. II,** 632.
254. H. Yoshioka et al., *J. Antibiot. Ser. B* **20** (1967) 34. I. P. Fomina et al., *Antibiotiki (Moscow)* **16** (1971) 153.
255. R. Sutherland et al., *Br. Med. J.* **4** (1970) 455.
256. G. N. Rolinson, S. Stevens, *Br. Med. J.* **2** (1961) 191.
257. H. B. Koenig et al., *Infection (Munich)* **5** (1977) 170.
258. K. Ueo et al., *Antimicrob. Agents Chemother.* **12** (1977) 455.
259. R. Sutherland et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1971, 411. R. Sutherland et al., *Br. Med. J.* **3** (1972) 13.
260. A. J. Gonzaga et al., *J. Infect. Dis.* **129** (1974) 545. H. C. Neu, *Drugs* **9** (1975) 81; *Int. J. Clin. Pharmacol. Biopharm.* **11** (1975) 132.
261. R. Sutherland et al., *Br. Med. J.* **2** (1967) 804. R. Kahrimanis, P. Pierpaoli, *N. Engl. J. Med.* **285** (1971) 236.
262. C. N. Baker et al., *Intersci. Conf. Antimicrob. Agents Chemother. 18th,* Abstract 293, Atlanta 1978.
263. Y. Shiobara et al., *J. Antibiot.* **27** (1974) 665.
264. M. Ehrnebo et al., *J. Pharmacokinet. Biopharm.* **7** (1979) 429.
265. K. Roholt, *J. Antimicrob. Chemother. (Suppl. B)* **3** (1977) 71.
266. F. Lund, L. Tybring, *Nature (London)* **236** (1972) 135.
267. C. W. Ryan et al., *J. Med. Chem.* **12** (1969) 310.
268. R. R. Chauvette, P. A. Pennington, *J. Med. Chem.* **18** (1975) 403.
269. J. L. Spencer et al., *J. Med. Chem.* **9** (1966) 746.
270. R. F. Buck, K. E. Price, *Antimicrob. Agents Chemother.* **11** (1977) 324.
271. J. E. Dolfini et al., *J. Med. Chem.* **14** (1971) 117.
272. O. Zak et al., *J. Antibiot.* **29** (1976) 653.
273. L. B. Crast et al., *J. Med. Chem.* **16** (1973) 1413.
274. R. R. Chauvette et al., *J. Am. Chem. Soc.* **84** (1962) 3401.
275. F. Knüsel et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1970, 140. H. C. Neu et al., *J. Antibiot.* **25** (1972) 400.
276. H. Nomura et al., *J. Med. Chem.* **17** (1974) 1312. K. Tsuchiya et al., *Antimicrob. Agents Chemother.* **13** (1978) 137.
277. G. T. Stewart et al., *Lancet (1964),* **vol. II,** 1305.
278. R. Del Busto et al., *Antimicrob. Agents Chemother.* **9** (1976) 397.
279. I. Saikawa et al., *Yakugaku Zasshi* **99** (1979) 929.
280. W. E. Wick, D. A. Preston, *Antimicrob. Agents Chemother.* **1** (1972) 221.
281. M. Numata et al., *J. Antibiot.* **31** (1978) 1262.
282. K. Kariyone et al., *J. Antibiot.* **23** (1970) 131.
283. T. Kamimura et al., *Antimicrob. Agents Chemother.* **16** (1979) 540.
284. R. Heymes et al., *Infection (Munich)* **5** (1977) 529.
285. K. Tsuchiya et al., *Antimicrob. Agents Chemother.* **19** (1981) 56.
286. R. N. Brogden et al., *Drugs* **17** (1979) 223. C. H. O'Callaghan et al., *Antimicrob. Agents Chemother.* **9** (1976) 511.
287. R. Reiner et al., *J. Antibiot.* **33** (1980) 783.
288. C. H. O'Callaghan et al., *Antimicrob. Agents Chemother.* **17** (1980) 876.
289. E. O. Stapley et al., *Antimicrob. Agents Chemother.* **2** (1972) 122.
290. H. Nakao et al., *J. Antibiot.* **29** (1976) 554.
291. M. Iwanami et al., *Chem. Pharm. Bull.* **28** (1980) 2629.
292. M. Narisada et al., *J. Med. Chem.* **22** (1979) 757.
293. T. T. Howarth et al., *J. Chem. Soc. Chem. Commun.* 1976, 266.
294. W. J. Leanza et al., *J. Med. Chem.* **22** (1979) 1435.
295. R. B. Sykes, I. Phillips, *J. Antimicrob. Chemother. (Suppl. E)* **8** (1981) 1.
296. J. H. Boothe et al., *J. Am. Chem. Soc.* **75** (1953) 4621. L. H. Conover et al., *J. Am. Chem. Soc.* **75** (1953) 4622.
297. J. R. D. McCormick et al., *J. Am. Chem. Soc.* **79** (1957) 4561.
298. A. C. Finlay et al., *Science (Washington, D.C.)* **111** (1950) 85. P. P. Regna, I. A. Solomons, *Ann. N.Y. Acad. Sci.* **53** (1950) 221.
299. R. K. Blackwood et al., *J. Am. Chem. Soc.* **83** (1961) 2773; **85** (1963) 3943.

300. C. R. Stephens et al., *J. Am. Chem. Soc.* **80** (1958) 5324. J. R. D. McCormick et al., *J. Am. Chem. Soc.* **82** (1960) 3381.
301. W. J. Gottstein et al., *J. Am. Chem. Soc.* **81** (1959) 1198.
302. G. S. Redin, *Antimicrob. Agents Chemother. (1961 – 70)* 1967, 371.
303. A. D. Di Marco et al., *Nature (London)* **201** (1964) 706. F. Arcamone et al., *J. Am. Chem. Soc.* **86** (1964) 5334.
304. F. Arcamone et al., *Tetrahedron Lett.* 1969, 1007.
305. T. Oki et al., *J. Antibiot.* **28** (1975) 830.
306. H. Umezawa et al., *J. Antibiot. Ser. A* **10** (1957) 181.
307. T. Wakazawa et al., *J. Antibiot. Ser. A* **14** (1961) 180, 187.
308. D. A. Preston, W. E. Wick, *Antimicrob. Agents Chemother.* 1971, 322.
309. M. J. Weinstein et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1964, 1. J. Black et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1964, 138.
310. H. Umezawa et al., *J. Antibiot.* **24** (1971) 485.
311. H. Kawaguchi, *J. Infect. Dis. (Suppl.)* **134** (1976) 242.
312. R. Okachi et al., *J. Antibiot.* **27** (1974) 793.
313. T. Shomura et al., *J. Antibiot.* **23** (1970) 155.
314. S. A. Waksman, H. A. Lechevalier, *Science (Washington, D.C.)* **109** (1949) 305.
315. T. H. Haskell et al., *J. Am. Chem. Soc.* **81** (1959) 3480, 3482.
316. A. Schatz, E. Bugie, S. A. Waksman, *Proc. Soc. Exp. Biol. Med.* **55** (1944) 66.
317. R. L. Peck et al., *J. Am. Chem. Soc.* **68** (1946) 1390.
318. S. Kondo et al., *J. Antibiot. Ser. A* **18** (1965) 38.
319. R. L. Mann, W. W. Broomer, *J. Am. Chem. Soc.* **80** (1958) 2714.
320. R. Q. Thompson, E. A. Presti, *Antimicrob. Agents Chemother. (1961–70)* 1967, 332.
321. J. A. Waitz et al., *Antimicrob. Agents Chemother.* **2** (1972) 431. C. C. Crowe, E. Sanders, *Antimicrob. Agents Chemother.* **3** (1973) 24.
322. S. A. Kabins et al., *Antimicrob. Agents Chemother.* **10** (1976) 139. J. J. Rahal et al., *Antimicrob. Agents Chemother.* **9** (1976) 595. V. Dhawan et al., *Antimicrob. Agents Chemother.* **11** (1977) 64.
323. D. J. Mason et al., *Antibiot. Chemother. (Washington, D.C.)* **11** (1961) 118.
324. R. Okachi et al., *J. Antibiot.* **30** (1977) 541.
325. S. Horii et al., *J. Antibiot.* **24** (1971) 57, 59.
326. H. Umezawa et al., *J. Antibiot. Ser. A* **18** (1965) 101.
327. J. M. McGuire et al., *Antibiot. Chemother. (Washington, D.C.)* **2** (1952) 281.
328. V. C. Stephens et al., *J. Am. Pharm. Assoc. Sci. Ed.* **48** (1959) 620.
329. B. A. Sobin et al., *Antibiot. Annu.* 1954 – 1955, 87.
330. W. D. Celmer et al., *Antibiot. Annu.* 1957 – 1958, 476.
331. T. Osono et al., *J. Antibiot.* **20** (1967) 174. K. Nitta et al., *J. Antibiot.* **20** (1967) 181.
332. J. C. Chabala et al., *J. Med. Chem.* **23** (1980) 1134.
333. T. Tsuruoka et al., *J. Antibiot.* **24** (1971) 319, 452.
334. P. Sensi et. al., *Antimicrob. Agents Chemother. (1961 – 70)* 1967, 699.
335. W. Tanaka et al., *Heterocycles* **13** (1979) 469.
336. W. Hausmann et al., *J. Am. Chem. Soc.* **76** (1954) 4892.
337. B. A. Johnson et al., *Science (Washington, D.C.)* **102** (1945) 376.
338. Y. Koyama et al., *J. Antibiot.* **3** (1950) 457.
339. S. Benjamin et al., *Antibiot. Annu.* 1959 – 1960, 41.
340. S. Tanayama et al., *J. Antibiot.* **21** (1968) 313.
341. M. Arai et al., *J. Antibiot. Ser. A* **9** (1956) 193.
342. P. J. Van Dijck et al., *Chemotherapy (Basel)* **14** (1969) 109. P. Crooy et al., *J. Antibiot.* **25** (1972) 371.
343. E. B. Herr Jr. et al., *Ann. N.Y. Acad. Sci.* **135** (1966) 940.
344. A. C. Finlay et al., *Am. Rev. Tuberc.* **63** (1951) 1. Q. R. Bartz et al., *Am. Rev. Tuberc.* **63** (1951) 4.
345. H. Yoshioka et al., *Tetrahedron Lett.* 1971, 2043.
346. M. H. McCormick et al., *Antibiot. Annu.* 1955 – 1956, 606.
347. S. A. Waksman et al., *Proc. Soc. Exp. Biol. Med.* **45** (1940) 609.
348. N. Ishida et al., *J. Antibiot. Ser. A* **18** (1965) 68.
349. H. Umezawa et al., *J. Antibiot.* **29** (1976) 97.
350. H. Umezawa et al., *J. Antibiot.* **23** (1970) 259.
351. M. E. Haney Jr. et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1967, 349.
352. J. Berger et al., *J. Am. Chem. Soc.* **73** (1951) 5295.
353. Y. Miyazaki, *J. Antibiot.* **27** (1974) 814.
354. W. Gold et al., *Antibiot. Annu.* 1955 – 1956, 579.
355. E. L. Hazen, *Science (Washington, D.C.)* **112** (1950) 423.

356. A. P. Struyk et al., *Antibiot. Annu.* 1957 – 1958, 878.
357. S. Hosoya et al., *Jpn. J. Exptl. Med.* **22** (1952) 505. S. Hosoya, *Chemotherapy (Tokyo)* **2** (1954) 1.
358. W. E. Grundy et al., *Antibiot. Chemother. (Washington, D.C.)* **3** (1953) 1215.
359. C. Lewis et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1963, 570.
360. B. J. Magerlein, *Antimicrob. Agents Chemother. (1961 – 70)* 1967, 727.
361. Upjohn Co., ZA 6802283, 1968. (W. Morozowich, A. A. Sinkula)
362. K. H. Wallhäuser et al., *Antimicrob. Agents Chemother. (1961 – 70)* 1965, 734.
363. R. L. Harned et al., *Antibiot. Chemother. (Washington, D.C.)* **5** (1955) 204.
364. S. Takeuchi et al. *J. Antibiot. Ser. A.* **12** (1959) 109.
365. A. J. Glazko et al., *Antibiot. Chemother. (Washington, D.C.)* **2** (1952) 234.
366. K. Arima et al., *Agric. Biol. Chem.* **28** (1964) 575. K. Arima et al., *J. Antibiot. Ser. A* **18** (1965) 201.
367. T. Kamiya et al., *J. Antibiot.* **25** (1972) 576.
368. H. Egger, H. Reinshagen, *J. Antibiot.* **29** (1976) 915.
369. K. Ishibashi, *J. Antibiot. Ser. A* **15** (1962) 161.
370. S. A. Waksman, *Dtsch. Apoth. Ztg.* **109** (1969) 1019.
371. S. A. Goulden, *Manuf. Chem. Aerosol News* **36** (1965) no. 4, 45.
372. C. T. Calan, *Process Biochem.* **7** (1972) no. 7, 29.
373. American Cyanamid, GB 773453, 1954.
374. Bristol-Myers, US 2970946, 1960.
375. American Cyanamid, US 2734018, 1953 (P. P. Minieri, H. Sokol, M. C. Firman).
376. Bristol Lab., US 2739924, 1953.
377. C. G. T. Evans, *Manuf. Chem.* **31** (1960) 5 – 9.
378. G. D. Wilkin, *Manuf. Chem.* **31** (1960) 329.
379. J. Málek, J. Hospodka, *Folia Microbiol. (Prague)* **5** (1960) 120.
380. J. Málek, Z. Fencl: *Theoretical and Methodological Basis of Continuous Culture of Microorganisms,* Publ. House of the Czechosl. Acad. of Sciences, Prague, and Academic Press, New York – London, Engl. ed. 1966.
381. A. L. Demain, C. L. Cooney, *Process Biochem.* **7** (1972) no. 7, 21.
382. W. Oberzill, N. Matsché, *Chem. Ing. Tech.* **43** (1971) 83.
383. H. W. Blanch, P. L. Rogers, *Biotechnol. Bioeng.* **13** (1971) 843.
384. R. Kreutzfeldt, *Angew. Chem. Int. Ed. Engl.* **6** (1967) 470.
385. M. J. Thirumalachar, R. S. Sukapure, P. W. Rahalkar, K. S. Gopalkrishnan, *Hind. Antibiot. Bull.* **5** (1962) 1.
386. B. N. Ganguli, V. M. Doctor, *Hind. Antibiot. Bull.* **7** (1964) no. 2, 85.
387. Distillers, GB 649818, 1948.
388. Benckiser, DE 1000572, 1952.
389. G. R. Ambekar, S. B. Thadani, *Appl. Microbiol.* **13** (1965) 713.
390. Ankerfarm, DE 1467764, 1965.
391. Benckiser, DE 1103735, 1958.
392. Hoechst, DE 1247549, 1965; NL-A 6602132, 1966; BE 677053, 1966.
393. P. Hosler, M. J. Johnson, *Ind. Eng. Chem.* **45** (1953) 871.
394. C. Rainbow, A. H. Rose: *Biochemistry in Industrial Microbiology,* Academic Press, New York 1963, p. 254.
395. A. Söder: "Tetracycline," in G. Ehrhart, H. Ruschig (eds.): *Arzneimittel, Entwicklung, Wirkung, Darstellung*, 2nd ed., vol. 4, Verlag Chemie, Weinheim 1972, p. 368.
396. R. V. Reeves, *Chem. Eng. (N.Y)* **59** (1952) Jan., 145.
397. W. Dürckheimer in [395] vol. 5, p. 302.
398. Bundesverband der Pharmazeutischen Industrie: "GMP-Regulations of Food and Drug Administration/USA," 15. Jan. 1971, including amendment of 2. Mar. 1971, *Pharm. Ind.* **33** (1971) 364.
399. WHO: "Draft Requirements for Good Manufacturing Practice in the Manufacture and Quality Control of Drugs and Pharmaceutical Specialities," World Health Organisation, Tech. Rep., Ser. 1969, no. 418, Annex 2.
400. WHO: "Quality and Control of Drugs," Official Records of the World Health Organisation, no. 176, Dec. 1972, Annex 12.
401. R. Marris, *Pharm. Ind.* **33** (1971) 749.
402. *Pharm. Ind.* **32** (1970) 813 – 819.
403. FDA, *Fed. Regist.* **37** (1972) no. 101, 10510.
404. E. P. Abraham, E. Chain et al., *Lancet 1941,* **vol. II,** 177, 189.
405. W. H. Schmidt et al., *J. Bacteriol.* **47** (1944) 199.
406. M. D. Reeves et al., *J. Bacteriol.* **49** (1945) 395.
407. FDA, *Fed. Regist.* **12** (1947), 4th Apr., 2215, 2217 – 2226.
408. H. Seyfarth, O. P. Ewald, *Pharm. Ind.* **34** (1972) 40.
409. K. H. Wallhäusser, *Pharm. Ind.* **34** (1972) 23.
410. T. Dimmling, *Ärztl. Wochenschr.* **8** (1953) no. 27, 633.

411. P. Naumann, *Antibiot. Chemother. (Basel)* **10** (1962) 1 – 93.
412. A. L. Barry, *Am. J. Med. Technol.* **30** (1964) 153, 333.
413. L. J. Griffith, C. G. Mullins, *Appl. Microbiol.* **16** (1968) 656.
414. L. G. Wayland, P. J. Weiss, *J. Pharm. Sci.* **57** (1968) 806.
415. K. H. Wallhäusser, *Ärztl. Lab.* **16** (1970) 150.
416. H. Knöll, *Pharmazie* **2** (1947) 392.
417. W. Irmer, *Dtsch. Med. Rundsch.* 1949, 123.
418. K. Irrgang, *Z. Naturforsch. B: Anorg. Chem. Org. Chem. Biochem. Biophys. Biol.* **5 B** (1950) 155.
419. G. Dorner, T. Lammers, *Med. Klin. (Munich)* **46** (1951) 522.
420. W. H. C. Shaw, R. E. Duncombe, *Analyst (London)* **88** (1963) 694.
421. T. A. Haney et al., *Ann. N.Y. Acad. Sci.* **87** (1960) 782; **93** (1962) 627.
422. K. Heil, V. Beitz, *Pharm. Ind.* **34** (1972) 37.
423. W. H. Shaw et al., *Ann. N.Y. Acad. Sci.* **130** (1965) 647.
424. D. A. Burns et al., *Biotechnol. Bioeng.* **11** (1969) 1011.
425. R. E. Hone, C. T. Rhodes, *Process Biochem.* **7** (1972) Feb., 27.
426. E. P. Abraham, G. G. F. Newton, *Adv. Chemother.* **2** (1966) 23.
427. D. E. Nettleton et al., *Int. J. Appl. Radiat. Isot.* **13** (1962) 259.

Antibodies, Monoclonal → Monoclonal Antibodies
Anticoagulants → Blood
Anticonvulsants → Antiepileptics
Anticorrosion Pigments → Pigments, Inorganic
Antidepressive Agents → Psychopharmacological Drugs

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Oral Antidiabetic Drugs

Standard Article

Hilmar Bischoff[1]

[1]Bayer HealthCare AG, Wuppertal, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a03_001.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Stimulators of Insulin Secretion
2.1. Sulfonylureas and Related Compounds
2.2. Benzoic Acid Derivatives
2.3. Amino Acid Derivatives
2.4. Mode of Action
2.5. New Insulinotropic Mechanisms
2.5.1. Dipeptidyl Peptidase IV Inhibition
2.5.2. Dipeptidyl Peptidase IV Inhibitors in Clinical Development
3. Biguanides
4. Inhibitors of Intestinal Carbohydrate Digestion
4.1. α-Glucosidase Inhibitors
4.1.1. Mode of Action
4.1.2. Pharmacology
4.1.3. Individual Compounds
4.2. Plant Fibers
5. Enhancers of Insulin Action
5.1. Mode of Action
5.2. Pharmacology of PPAR γ agonists
5.3. Thiazolidine-2,4-Dione Derivatives (Glitazones)

The review describes all oral drugs used for the treatment of type 2 diabetes mellitus which are marketed or in late clinical development. Stimulators of insulin secretion are the most prescribed class of antidiabetic drugs. These compounds increase insulin secretion from pancreatic β-cells via the closure of the ATP-sensitive K^+ channel. Sulfonylurea (SU) derivatives bind to the SU receptor on a regulatory protein of the ATP-sensitive K^+ channel. Since various binding sites exist, also other compounds like benzoic acid or amino acid derivatives are able to bind and to stimulate insulin secretion via the closure of the ATP-sensitive K+ channel. Very new mechanisms, not yet approved and still in late clinical development, prevent the degradation of glucagons-like-peptide-1 (GLP-1), an intestinal hormone, which stimulates insulin secretion by increasing intracellular cAMP levels of β-cells. Those compounds, which inhibit the GLP-1 degrading enzyme dipeptidyl peptidase IV (DPP IV), sustain the insulinotropic effect of GLP-1. The antidiabetic action of biguanides, only metformin is marketed today, is still not fully understood. Metformin improves insulin sensitivity probably by an increase in peripheral glucose utilization and additionally metformin reduces hepatic gluconeogenesis. Alpha-glucosidase inhibitors are enzyme inhibitors of small intestinal carbohydrate digesting glucosidases. They display their antidiabetic effect by delaying the postprandial (pp) carbohydrate absorption resulting in reduced pp hyperglycemia and -insulinemia, which finally improves insulin sensitivity. Enhancers of insulin action are ligands of the peroxisome proliferator activated receptor γ (PPAR γ). PPARs belong to the nuclear receptor family of ligand-activated transcription factors. The major antidiabetic and blood glucose lowering effect of the PPAR γ agonists is the improvement of insulin sensitivity, however, the mechanism is not fully elucidated yet.

[Top of Page]

1. Introduction

Type 2 diabetes or noninsulin dependent diabetes mellitus (NIDDM) is a chronic disease of impaired carbohydrate, lipid, and protein metabolism, resulting from inadequate insulin action. Patients may have diminished insulin secretion, i.e., insulin deficiency, or insulin is secreted but does not act properly because of impaired sensitivity of muscle and liver to insulin (insulin resistance) [1], [2]. The treatment with antidiabetic drugs is aimed at achieving glycemic control to prevent the development of diabetic complications: neuropathy, nephropathy, and retinopathy. Oral antidiabetic drugs reduce or normalize elevated concentrations of glucose in blood and urine. Antidiabetic drugs available today aim at various target tissues and can be characterized by their different mode of actions. Basically two major pharmacological principles can be distinguished: (1) stimulation of insulin secretion (A) and (2) mechanisms where no elevated plasma insulin levels are required for glycemic control (B – D):

A. Insulin secretagogues act on the β-cells of the pancreas and stimulate insulin secretion.
B. Biguanides affect the glucose metabolism in the liver and peripheral tissues (e.g., skeletal muscle).
C. Inhibitors of carbohydrate digestion are active in the small intestine and delay carbohydrate digestion and glucose absorption.
D. Insulin sensitizers primarily act on peripheral tissues and enhance insulin action.

[Top of Page]

2. Stimulators of Insulin Secretion

2.1. Sulfonylureas and Related Compounds

When testing sulfonamide derivatives for antibacterial activity LOUBATIERES recognized their blood glucose lowering potential already in 1944 [3]. However, it took more than 10 years before the first non-antibacterial compounds carbutamide (Boehringer Mannheim) [4] and tolbutamide (Hoechst) [5] (Table 1) were introduced into the market for the treatment of type 2 diabetic patients in 1956.

Sulfonylurea (SU) derivatives of the first generation were active only in high dosages of up to 2000 mg/d. Sulfonylureas of the “second generation” showed much higher pharmacological potency and were active in significantly lower doses [6]. The first one, glibenclamide (glyburide) (Table 1), was introduced in the market by Boehringer Mannheim and Hoechst in 1969 requiring daily doses of only 2.5 – 15 mg.

Although the efforts in pharmaceutical sulfonylurea research decreased in the 1970s, investigations in the last two decades have been concentrated on increasing, prolonging, or shortening the action of insulin secretagogues or accelerating their onset of action. Glimepiride (Table 1), a structure closely related to glibenclamide in which the phenyl substituent of glibenclamide is replaced by pyrroline, was introduced into the market in 1996. In contrast to glibenclamide, in several animal species glimepiride is characterized by a faster onset of blood glucose lowering activity, and a longer-lasting glucose lowering effect [7].

Also in clinical trials, glimepiride showed longer-lasting hypoglycemic effects than glibenclamide [8]. Due to these properties, glimepiride was introduced as a drug for once-daily treatment. Although all 14 SU compounds of Table 1 are available at least in parts of the world, more than 90 % of the SU world market is dominated by only five SU derivatives — glibenclamide/glyburide, gliclazide, glipizide, chlorpropamide, and glimepiride (Table 1).

Table 1. Examples of insulin secretion simulators (daily dose for humans and threshold dose for rabbits) and elimination half life $\tau_{1/2}$

Generic name *M_r, mp,* Ref.	Examples of commercial preparations formula	Daily dose, mg	Threshold dose, mg	$\tau_{1/2}$ h
Sulfonylureas				
Acetohexamide [968-81-0] $C_{15}H_{20}N_2O_4S$ 324.2, 188 – 190 °C, 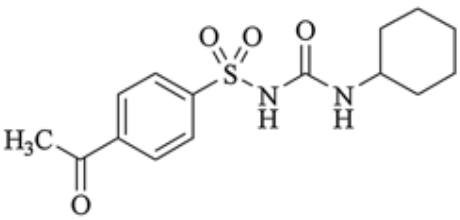 [9] *N*-(4-Acetylbenzenesulfonyl)-*N′*-cyclohexylurea	Dimelor (Lilly)	250 – 500	10	4 – 7
Carbutamide [	Nadisan (Boehringer Mannheim) Invenol	500 –	200	40

339-43-5] $C_{11}H_{17}N_3O_3S$ 271.3, 140 – 143 °C,	(Hoechst)	2000		

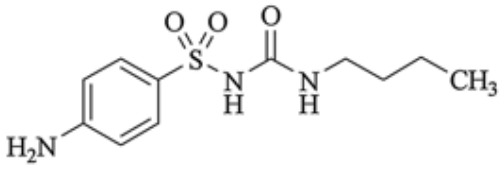

[10], [11]

N-Sulfanilyl-*N′*-*n*-butylurea

Chlorpropamide [94-20-2] $C_{10}H_{13}ClN_2O_3S$ 276.6, 126 – 128 °C,	Chloronase (Hoechst) Diabetoral (Boehringer Mannheim) Diabenese (Pfizer)	125 – 500	10	35

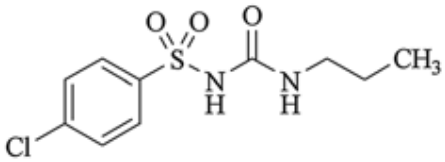

[5]

N-(4-Chlorobenzenesulfonyl)-*N′*-*n*-propylurea

Glibornuride [26944-48-9] $C_{18}H_{26}N_2O_4S$ 366.4, 192 – 195 °C,	Gluborid (Grünenthal) Glutril (Hoffmann-La Roche)	12.5 – 75	1	5 – 7

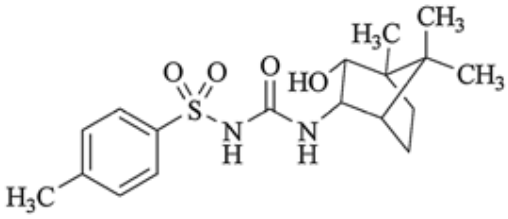

[12]

N-(Tosyl)-*N′*-(2-endo-hydroxy-3-endo-DL-bornyl)-urea

Glyburide (Glibenclamide) [10238-21-8] $C_{23}H_{28}ClN_3O_5S$ 495.0, 172 – 174 °C,	Euglucon (Boehringer Mannheim, Hoechst) Daonil (Hoechst)	1.75 – 20	0.01	5 – 8

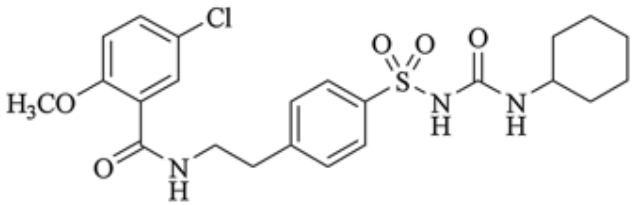

[6]

N-4-[2-(Chloro-2-methoxybenzamido)-ethyl]-benzenesulfonyl)-*N′*-cyclohexylurea

Gliclazide [21187-98-4] $C_{15}H_{21}N_3O_3S$ 323.2, 180 – 182 °C,	Diamicron (Science Union, Servier)	40 – 240	1.0	7

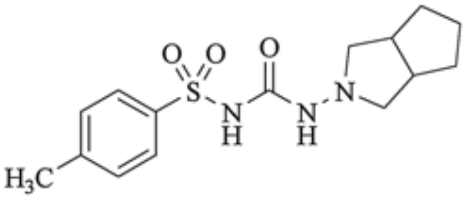

[13]

N-(Tosyl)-*N′*-(3-aza-bicyclo-[3.3.0.]-3-octyl)-urea

Glimepiride [93479-97-1] $C_{24}H_{34}N_4O_5S$ 490.62, 207 °C,	Amaryl (Hoechst/Sanofi-Aventis)	1 – 6	1.5 – 3.5

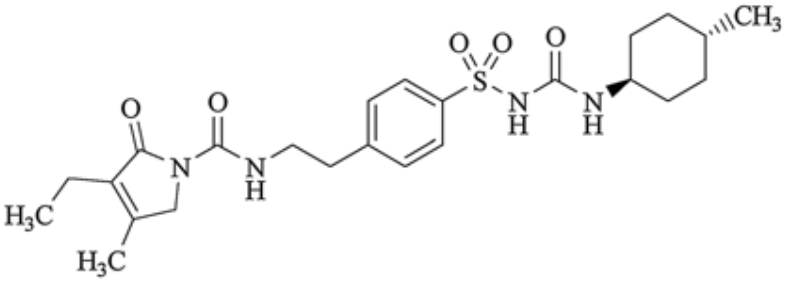

[14]

Trans-3-ethyl-2,5-dihydro-4-methyl-*N*-[2-[4-[[[[(4-methylcyclohexyl) amino]carbonyl]amino]sulfonyl] phenyl]ethyl]-2-oxo-1*H*-pyrrole-1-carboxamide

Glipizide [29094-61-9] $C_{21}H_{27}N_5O_4S$ 445.5, 208 – 209 °C,	Glibenese (Pfizer) Mini Diab (Carlo Erba)	2.5 – 20	0.02	2 – 4

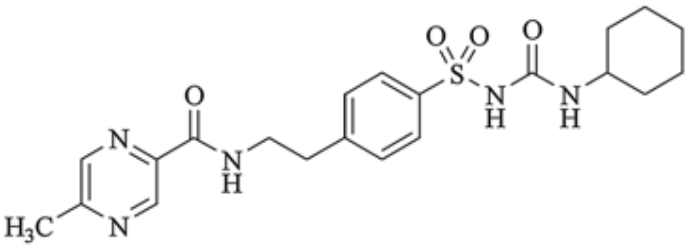

[15]

N-{4-[5-methylpyrazine-2-carboxamido)-ethyl]-benzenesulfonyl}-*N′*-cyclohexylurea

Gliquidone [33342-05-1] $C_{27}H_{33}N_3O_6S$ 527.3, 180 – 182 °C,	Glurenorm (Thomae)	15.0 – 120	2.5	2 – 4

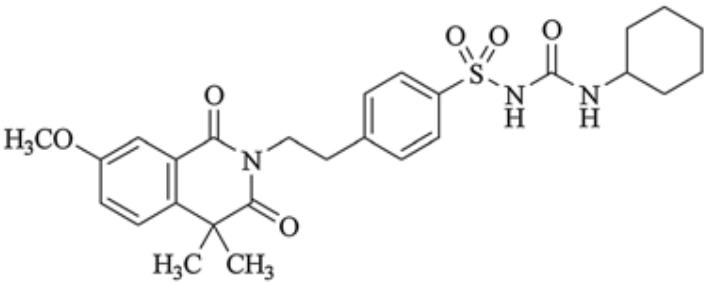

[16]

N-{4-[2-(3,4-Dihydro-7-methoxy-4,4-dimethyl-1,3-dioxa-2(1*H*)-isoquinolyl)-ethyl]-benzenesulfonyl}-*N*-cyclohexylurea

Glisoxepide [25046-79-1]	Pro-Diaban (Bayer, Schering)	2.0 – 16	0.05	2 – 4

$C_{20}H_{27}N_5O_5S$

449.5, 194 – 198 °C,

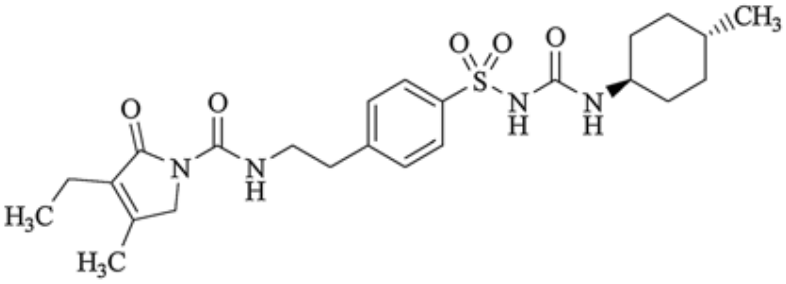

[17], [18]

4-{4-[β-(5-Methylisoxazolyl-3-carboxamido)-ethyl]-benzenesulfonyl}-1,1-hexamethylenesemicarbazide

Glymidine (Glycodiazine) [3459-20-9]	Redul S (Schering)	500 – 1500	20	4

$C_{13}H_{14}N_3O_4SNa$

331.2, 221 – 226 °C,

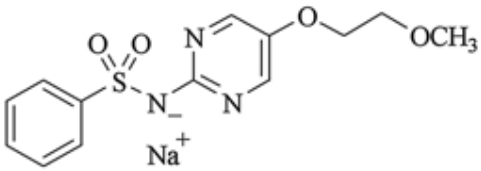

[19]

2-Benzenesulfonamido-5-(2-methoxyethoxy)-pyrimidine sodium

Tolazamide [1156-19-0]	Norglycin (Upjohn) Tolinase (Upjohn)	125 – 1000	5	7

$C_{14}H_{21}N_3O_3S$

311.4, 165 – 173 °C,

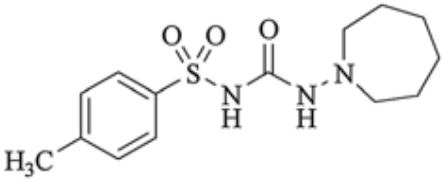

[20]

N-(Tosyl)-*N′*-(hexahydro-1*H*-azepin-1-yl)-urea

Tolbutamide [64-77-7]	Artosin (Boehringer Mannheim) Rastinon (Hoechst) Orinase (Upjohn)	500 – 3000	10	5 – 7

$C_{12}H_{18}N_2O_3S$

270.4, 127 – 129 °C,

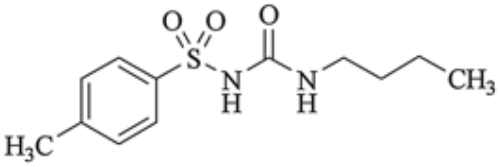

[5]

N-(Tosyl)-*N′*-*n*-butylurea

Tolcyclamide (Glyciclamide) [664-95-9]	Diaboral (Carlo Erba)	150 – 1000	10	6 – 7

$C_{14}H_{20}N_2O_3S$

296.4, 174 – 175 °C, 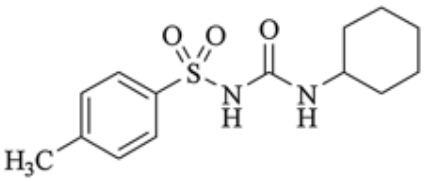 [5] *N*-(Tosyl)-*N′*-cyclohexylurea			
Benzoic Acid Derivative			
Repaglinide [135062-02-1] $C_{27}H_{36}N_2O_4$ 452.59, 126 – 131 °C, 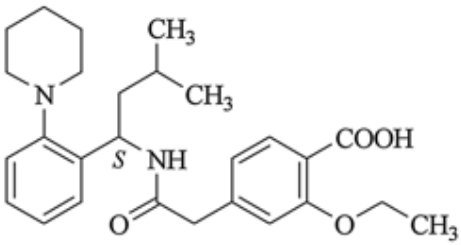 [21] (*S*)-2-Ethoxy-4-[2-[[3-methyl-1-[2-(1-piperidinyl)phenyl]butyl]amino]-2-oxoethyl]benzoic acid	Novonorm (Novo Nordisk)	1.5 – 16	0.5 – 1
Amino Acid Derivative			
Nateglinide [491828-09-2] $C_{19}H_{27}NO_3$ 317,43 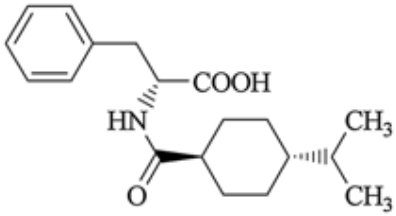 [22]	Starlix (Novartis)	180 – 360	1 – 2

2.2. Benzoic Acid Derivatives

Research on the structure-activity relationship of potent “second generation” SUs like gliblenclamide has revealed that the nonsulfonylurea moiety of glibenclamide also stimulated insulin secretion. The benzoic acid derivative meglitinide (HB 699) stimulated insulin release in vitro and in vivo, but the potency was only comparable to tolbutamide [8]. Repaglinide, the active (S)-enantiomer of AG-EE 388 ZW (Boehringer Ingelheim) (Table 1) is a benzoic acid derivative with remarkably higher potency, comparable with glibenclamide [23]. Due to its different pharmacokinetic behavior from glibenclamide, more rapid absorption, and shorter half-life, repaglinide showed interesting pharmacological properties in terms of quick insulin release and short-acting blood glucose lowering effects [24]. Repaglinide was introduced into the market in 1998.

2.3. Amino Acid Derivatives

Amino acid derivatives derived from D-phenylalanine are structurally a new type of insulin secretagogues [23].

Nateglinide (A-4166), Starlix, shows stereospecific activity, and it was found that the R-configuration was essential for hypoglycemic activity.

It is suggested that the phenylmethyl substituent of the amino acid residue is essential for activity. Variation of the amino acid moeity, e.g., glycine, alanine, phenylglycine, tyrosine, and others reduced the hypoglycemic activity [23]. Nateglinide produces a very quick insulin release followed by a rapid onset of blood glucose decrease and shows a short duration of action [25], [26]. Clinical trials have demonstrated mild blood glucose lowering activity after single doses of 60 – 180 mg. The first market introduction was in 1999 in Japan and in 2001 in USA.

Mitiglinide [145525-41-3] (KAD-1229, Kissei Pharmaceutical) is another non-sulfonylurea insulin secretagogue [27]. Mitiglinide is a benzylsuccinic acid derivative and in phase III of clinical development.

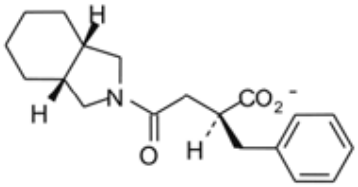

2.4. Mode of Action

Insulin secretion is physiologically stimulated by the binding of ATP to a cytosolic nucleotide binding site of the membrane bound ATP-sensitive K^+ channel which leads to closure of the K^+ channel. The inhibition of K^+ permeability depolarizes the plasma membrane, subsequently the voltage-dependent Ca^{2+} channel opens to promote the Ca^{2+} influx which finally results in insulin secretion [28].

The insulin secretagogues stimulate insulin secretion by closure of this ATP-sensitive K^+ channel of pancreatic β-cells. The KATP-channel consists of two subunits, a functional ion channel pore and a regulatory protein. The binding of the secretagogues is suggested to occur to the separate regulator protein containing the binding sites for sulfonylureas (sulfonylurea receptor 1, SUR1) but also other compounds [26], [27], [29].

2.5. New Insulinotropic Mechanisms

In 2005, a new antidiabetic principle was approved in the USA and introduced in the market: exendin-4 (Exenatide) [30], [31], an injectable analogue of the gastrointestinal peptide hormone glucagon-like peptide-1 (GLP-1) (see ⟶ Peptide and Protein Hormones). GLP-1 stimulates insulin secretion in a glucose-dependent manner with little risk of hypoglycemia via activation of the β-cell hormone receptor. Additionally, GLP-1 shows further important effects on β-cells, insulin biosynthesis and glucagon secretion, as well as on gastric motility. However, GLP-1 is unsuitable as a therapeutic agent, because of rapid enzymatic degradation by dipeptidyl peptidase IV (DPP IV also known as CD26) which results in a half-life of about two minutes [32].

2.5.1. *Dipeptidyl Peptidase IV Inhibition*

Very shortly after the discovery in 1993 that DPP IV inactivates GLP-1 [33], the inhibition of DPP IV was proposed as a new antidiabetic therapy using the attractive antidiabetic profile of GLP-1 [34]. Differently from injectable peptide hormones, all DPP IV inhibitors under development are orally available drugs.

DPP IV inhibitors sustain the plasma concentrations of GLP-1 by preventing the N-terminal degradation of GLP-1 [35], which results in significant enhancement of its insulinotropic activity, as well as improvement of islet function, which has been observed in long-term animal studies with DPP IV inhibitors [36], [37].

In general, good tolerability has been reported for DPP IV inhibitors in clinical studies. However, little data are available with respect to selectivity regarding other DPPs (e.g., DPP 8/9) [38], [39] and consequences of their inhibition are unknown. And because DPP IV cleaves and inactivates a variety of peptide hormones, special observation in long-term trials is needed for the safety evaluation of this class of drugs.

2.5.2. *Dipeptidyl Peptidase IV Inhibitors in Clinical Development*

The most advanced compounds in clinical phase III trials are vildagliptin [274901-16-5] (LAF-237, Novartis) [40]

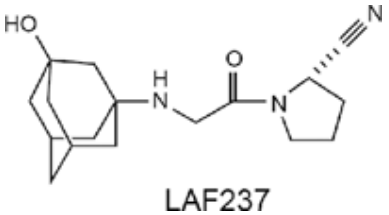

and *sitagliptin* [274901-16-5] (MK-0431, Merck & Co.) [41].

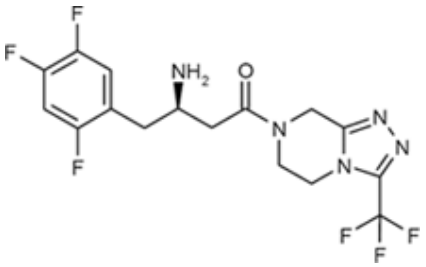

The only published long-term (52 weeks) trial results are available for vildagliptin (50 mg OD) in combination with metformin. Vildagliptin showed reductions in HbA_{1C} after 12 weeks (Δ = –0.7 %) and good tolerability after 52 weeks [42]. Further long-term clinical trials with both, vildagliptin and sitagliptin, are still ongoing. Although no long-term data are published with sitagliptin, in USA a new drug approval (NDA) for sitagliptin was filed in early 2006. It is too early for an evaluation of the

therapeutic efficacy, but in general, one may say DPP IV inhibitors reduce glycemia and glucagon levels, sustain insulin levels in patients, improve the glucose tolerance and as a glucose-dependent insulin secretagogue they avoid serious hypoglycemic events [43], [44].

[Top of Page]

3. Biguanides

The hypoglycemic activity of guanidine derivatives was firstly reported in 1918 [45]. The first blood glucose lowering diguanidine derivative decamethylenediguanidine (Synthalin A) was synthetized in the 1920s. Because of liver toxicity development of these compounds was discontinued. Only in the 1950s three biguanides were developed. The disubstituted dimethylbiguanide metformin [657-24-9] was introduced in 1957.

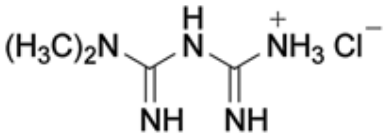

Monosubstituted compounds with a longer side chain, the phenylethylbiguanide *phenformin* [114-86-3] and the *N*-butylbiguanide *buformin* [692-13-7], were introduced in the same year and 1958, respectively [46].

Pharmacology and Mode of Action. Biguanides are ineffective in normal animals, unless very high doses are used and there is no risk of hypoglycemia. However, in diabetic animals blood glucose-lowering effects of biguanide can be demonstrated, depending on the model used. Biguanides do not stimulate insulin secretion, but need insulin for pharmacological hypoglycemic activity. In severe insulinopenic animal models, high doses have to be used and effects are modest. In mild diabetic animals or in insulin resistant animals, lower doses are already efficacious in lowering hyperglycemia and hyperinsulinemia [47].

The mechanism of action of biguanides is still not fully understood. Three major tissues have been identified as pharmacological sites of action: (1) the small intestinal wall, (2) the liver, and (3) peripheral tissues, mainly the skeletal muscle:

1. For the small intestine an inhibition of glucose absorption was described, however, this is, at least for metformin, of minor significance and not important for the blood glucose lowering effect. However, the intestinal glucose metabolization to lactate is stimulated and reduces the postprandial uptake of glucose by the liver.
2. Numerous studies have shown that biguanides inhibit hepatic gluconeogenesis and this may contribute to the blood glucose lowering effect, particularly in the fasting state. Again, metformin has probably less impact on gluconeogenesis than phenformin and buformin.
3. In the peripheral tissues, metformin increases the glucose disposal and utilization particularly in the skeletal muscle, which is probably the major contribution to the blood glucose lowering activity. In vitro studies using cell cultures have shown that metformin potentiates insulin action. In vivo studies in animals and diabetic patients have demonstrated that metformin reduces insulin resistance, at least in obese individuals [47], [48].
   Phenformin and buformin may cause lactic acidosis, which is the most serious adverse event of biguanides. Due to very serious cases of lactic acidosis both drugs were withdrawn from the markets in most parts of the world during the 1970s. Today they are available only in a few countries of Eastern Europe and South America. Owing to a more favorable pharmacokinetic profile of metformin, metformin associated lactic acidosis is very rare [48]. Therefore, only metformin is available today in all important countries. In the 1990s metformin experienced a renaissance in many countries, particularly in the USA, where metformin was approved for the treatment of type 2 diabetes only in 1994.

[Top of Page]

4. Inhibitors of Intestinal Carbohydrate Digestion

Delaying absorption as a therapeutic principle is well accepted for dietary therapy to avoid high postprandial blood glucose excursions [49]. Carbohydrates represent quantitatively the major portion of the human diet, in the form of mono-, di- and polysaccharides. However, only the relatively infrequently occurring monosaccharides such as glucose and fructose can be absorbed and are readily taken up from the small intestine. The more important dietary carbohydrate components, the complex starches and also sucrose have to be hydrolyzed enzymatically to their constituting monosaccharides by specific intestinal α-glucosidases (α-amylase, glucoamylase, dextrinase, maltase, isomaltase and sucrase) before they can be absorbed. The high density of intestinal membrane bound α-glucosidase enzymes in the jejunum and the high proportion of refined carbohydrates in the Western diet are the main reasons that the digestive process usually takes place very rapidly in the proximal small intestine. As a result, a high carbohydrate digestion rate leads to a rapid postprandial rise in blood glucose which is followed by a glucose stimulated insulin increase [50].

4.1. α-Glucosidase Inhibitors

α-Glucosidase inhibitors are of various chemical structures (polypeptides and carbohydrates). The most potent inhibitors are from a carbohydrate-derived backbone containing a nitrogen atom in the molecule. Polypeptides are less active.

4.1.1. *Mode of Action*

Essential for the action of the carbohydrate-type α-glucosidase inhibitors is their intramolecular nitrogen atom, which binds to the carbohydrate binding site of the intestinal α-glucosidases [51]. Compared with the oxygen atom of the α-glycosidic bond of dietary complex carbohydrates, the affinity of the nitrogen atom to the carbohydrate binding site of the enzyme is much

higher (factor 10^4 – 10^7). Despite their high affinity for these enzymes, the binding of α-glucosidase inhibitors is reversible and their inhibition kinetics are competitive [52].

4.1.2. *Pharmacology*

Pharmacological inhibition of intestinal α-glucosidases effectively retards the digestion of carbohydrates, slows down the absorption of glucose and avoids the often very high postprandial blood glucose peaks particularly in diabetic patients. As a result the postprandial glycemia and glucose stimulated hyperinsulinemia are flattened and the metabolic control will be improved. From the primary pharmacological effects, it can be concluded that α-glucosidase inhibitors act to reduce postprandial hyperglycemia without stimulating insulin secretion, thus avoiding any risk of hypoglycemia. Because of their unique mode of action in the small intestine, α-glucosidase inhibitors can be used in combination with all other antidiabetic treatments and act in an additive manner together with other antidiabetic drugs [50].

4.1.3. *Individual Compounds*

The first active compounds were inhibitors of the *soluble pancreatic α-amylase*. BAY D 7791 was a protein isolated from wheat flour [53]. This α-*amylase* inhibitor inhibited the postprandial blood glucose rise in oral loading tests with raw starch, but was much less effective when used together with cooked starch.

BAY E 4609, an α-amylase inhibitor of microbial origin, was characterized to be of carbohydrate nature and showed a significantly higher inhibitory potency [53]. However, also this compound demonstrated only modest efficacy in clinical trials. Because α-amylase inhibitors affect only starch digestion, the degradation of important oligo- and disaccharides by the membrane bound enzymes is not inhibited, the overall efficacy is only modest.

Further research aimed at inhibitors of the membrane bound α-*glucosidases*. This approach was very successful and lead to the discovery of acarbose (Table 2) [54], the first α-glucosidase inhibitor, which was introduced into the market for the treatment of type 2 diabetic patients in 1990 (in many countries also for type 1 in addition to insulin).

Table 2. α-Glucosidase inhibitors

Generic name M_r, *mp*, Ref.	Examples of commercial preparations formula	Daily dose
Acarbose [56180-94-0] $C_{25}H_{43}NO_{18}$	Glucobay, Precose, Prandase, Glucor (Bayer)	100 – 300 mg
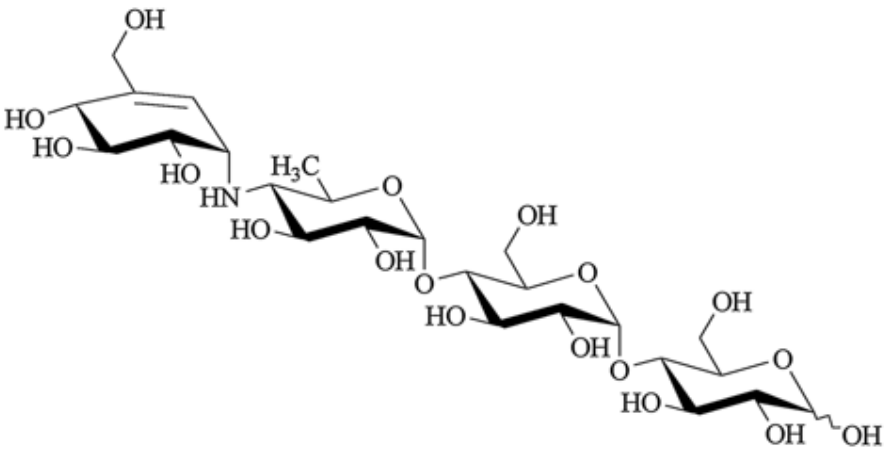		
645.61; amorph [55] *O*-{4,6-Dideoxy-4-[1(*S*)-(1,4,6/5)-4,5,6-trihydroxy-3-hydroxymethyl-2-cyclohexene-1-yl]amino- α-D-glucopyranosyl}-(1 ⟶ 4)-*O*- α-D-glucopyranosyl-(1 ⟶ 4)-D-glucose		
Miglitol [72432-03-2] $C_8H_{17}NO_5$	Diastabol (Sanofi) Glyset (Pfizer) Seibule (Sanwa Kagaku)	50 – 300 mg
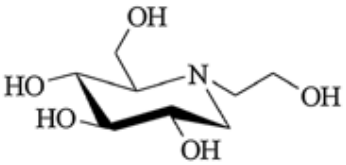		
207.23; 114 °		

C [56] 1,5-Dideoxy-1,5-[(2-hydroxyethyl)imino]-D-glucitol		
Voglibose [83480-29-9] $C_{10}H_{21}NO_7$	Basen (Takeda)	0.3 – 0.9 mg
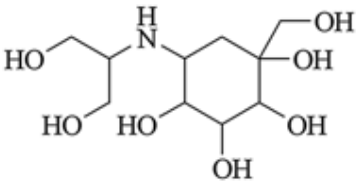		
267.28; 162 – 163 °C [57] 3,4-Dideoxy-4-[[2-hydroxy-1-(hydroxymethyl)ethyl]amino]-2-*C*-(hydroxy-methyl)-D-epiinositol		

Acarbose is a pseudotetrasaccharide of microbial origin, and was isolated from bacterial culture broths (*Actinoplanes* spec.) by Bayer [51], [52]. Acarbose consists of an acarviosine unit and maltose. The acarviosine nitrogen is responsible for the high affinity to the active binding site of α-glucosidases, however, the C – N linkage cannot be hydrolyzed and the enzymatic reaction stops.

Further research generated smaller molecules with different structures. The synthetic N-substituted valiolamine and the N-substituted 1-deoxynojirimycin derivatives showed even higher affinity to various α-glucosidases [51], [53].

The 1-deoxynojirimycin derivative *miglitol* (BAY M 1099) (Table 2) shows similar activity to acarbose with daily dosages of 50 – 300 mg [52].

Emiglitate (BAY O 1248) [80879-63-6], another deoxynojirimycin analogue is characterized by at least five-fold higher inhibitory potency than acarbose and additionally shows a long-lasting inhibitory effect [51], [52].

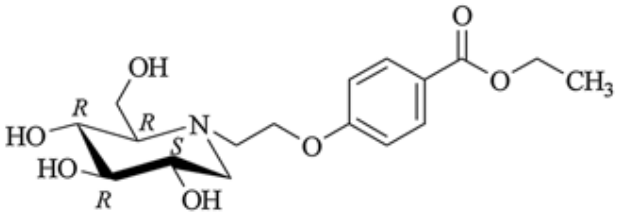

In animal studies emiglitate was still active up to 17 h after a single oral administration due to its pharmacokinetic behavior (formation of an active metabolite) [53], [58]. Both miglitol and emiglitate differ in their pharmacokinetic behavior from acarbose, which is poorly absorbed (1 – 2 %). The deoxynojirimycin derivatives are structurally related to monosaccharides and will be absorbed in the jejunum [58].

The most potent α-glucosidase inhibitor is the valiolamine derivative *voglibose* (Takeda) (Table 2). Compared with acarbose its inhibitory potency in vitro, depending on the α-glucosidase enzyme, is up to 200 times higher [51], [53]. In Japan, daily dosages are only 0.3 – 0.9 mg, however, therapeutic efficacy on HbA1c reduction could not be clearly demonstrated with those dosages. HbA1c, a glycated hemoglobin species, which occurs in blood, is a diagnostic measure for the non-enzymatic glycation process, and depends on the plasma glucose concentration. The process of protein glycation is discussed to be responsible for the development of diabetic complications (neuropathy, retinopathy, nephropathy). In Europe voglibose is under evaluation with distinctly higher daily dosages of 1.5 to 6 mg [59]. In contrast to acarbose, the smaller inhibitors of the monosaccharide type (miglitol, emiglitate, voglibose) do not inhibit pancreatic α-amylase [51], [53].

4.2. Plant Fibers

A more dietary approach for delaying absorption of nutrients is the use of undigestable carbohydrate polymers. Guar gum [9000-30-0] is a galactomannan obtained from the seeds of the guar plant (*Cyamopsis tetragonoloba*). Guar gum has no specific pharmacological site of action, but delays the absorption of digestible carbohydrates because it increases the viscosity in the small intestine and additionally hinders physically the enzymatic degradation of digestible carbohydrates. This approach is less effective in delaying the carbohydrate absorption than the pharmacological approach by the use of α-glucosidase inhibitors.

[Top of Page]

5. Enhancers of Insulin Action

The thiazolidine-2,4-dione derivatives (TZDs or glitazones) are insulin sensitizing compounds which improve glucose utilization without stimulating pancreatic insulin secretion. However, the thiazolidinediones are only active in conjunction with

insulin and show the best results in hyperinsulinemic, insulin resistant, obese individuals (animals and humans). The first compound *ciglitazone*, synthetized by Takeda and reported in 1982, was a result of screening of clofibrate analogues for hypolipidemic activity [60], [61].

5.1. Mode of Action

The mechanism of blood glucose lowering activity by thiazolidinedione derivatives has not been fully elucidated yet. Only minor or no effects are observed in normal animals or insulin sensitive diabetes models. In obese or hyperinsulinemic insulin resistant rodents thiazolidinediones lower insulin and triglyceride levels after treatment for 2 to 8 d. If hyperglycemia occurs, also the blood glucose concentration is reduced [23], [61].

Mechanistic in vitro studies with adipocytes have shown that the thiazolidinediones are ligands of the peroxisome proliferator activated receptor γ [62]. Peroxisome proliferator activated receptors (PPARs) belong to the nuclear receptor superfamily of ligand-activated transcription factors.

So far, three receptor subtypes have been identified: PPAR α, PPAR δ, and PPAR γ. The PPARs are believed to play a physiological role in the regulation of lipid metabolism [63], [64]. The in vivo antihyperglycemic activity of thiazolidinediones strongly correlates with their potency as PPAR γ agonists in vitro [65]. The mechanisms responsible for the blood glucose effects are controversally discussed, however, as long as there are no more mechanistic data available it is suggested that PPAR γ is a primary site of action of the thiazolidinediones [66].

5.2. Pharmacology of PPAR γ agonists

Ciglitazone reduced plasma triglycerides, hyperglycemia and hyperinsulinemia in several insulin resistant diabetic and non-diabetic rodent models after repeated oral dosing at doses of 100 – 300 mg/kg body weight, while having little or no effect in normal rodents or those with insulin deficient diabetes [61]. In long-term animal studies lens cataracts occurred and ciglitazone development was discontinued. Further synthetic research on thiazolidinedione analogues yielded compounds of distinctly higher in vitro and in vivo activity. Depending on the animal model, after repeated dosing troglitazone lowered hyperglycemia as well as levels of free fatty acids, triglycerides, and insulin at daily doses of 50 mg/kg p.o. or slightly lower [23], [67]. The glitazones are not only able to reduce insulin resistance, in vitro and in vivo studies have demonstrated increased synthesis of glucose transporter 1 (GLUT1) in various tissues. Little or no effect was observed on the insulin sensitive glucose transporter 4 (GLUT4), synthesis or insulin stimulated translocation of GLUT4, particularly in adipocytes, were only slightly increased. These effects may contribute to increased cellular glucose uptake [68].

In accordance with the PPAR γ mechanism, which primarily affects the adipose tissue, all obese animals gain weight under treatment with glitazones [64].

5.3. Thiazolidine-2,4-Dione Derivatives (Glitazones)

Clinical trials have confirmed the preclinical results. Best effects were observed in obese, highly insulin resistant type 2 diabetic patients and particularly in combination with insulin secretagogues (SU). However, the increase in body weight gain was also observed [64].

Troglitazone (Sankyo, see Table 3) was firstly approved 1995 in Japan and 1997 in USA and UK. It was launched in these countries in 1997. The recommended starting dose was 400 mg/d and the highest dose 600 mg. After being marketed, the daily practice revealed adverse events in the liver with cases of fatal liver failure. This lead to withdrawal of troglitazone from the market in the UK already after two months [69], in USA and Japan in the year 2000 [70].

Table 3. Thiazolidine-2.4-dione derivatives (TZDs)

Generic name M_r, *mp,* Ref.	Examples of commercial preparations formula	Daily dose
Pioglitazone [112529-15-4] $C_{19}H_{20}N_2O_3S \cdot HCl$	Actos (Ely Lilly) (Takeda)	15 – 45 mg
392.90; 183 – 184 °C [72] (±)-5-[[4-[2-(5-Ethyl-2-pyridinyl)ethoxy]phenyl]methyl]-2,4-thiazolidinedione, hydrochloride		
Rosiglitazone [122320-73-4] $C_{18}H_{19}N_3O_3S$	Avandia (SKB)	4 – 8 mg

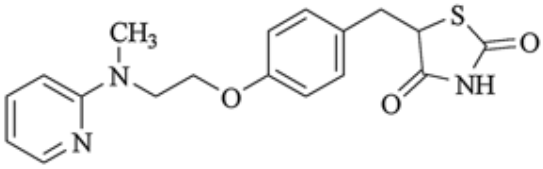

357.43; 122 – 123 °C

[73]

5-(4-[2-(Methylpyridin-2-ylamino)ethoxy]benzyl)thiazolidine-2,4-dione

Troglitazone [97322-87-7] Noscal (Sankyo) Rezulin (Warner-Lambert) 400 – 600 mg

$C_{24}H_{27}NO_5S$

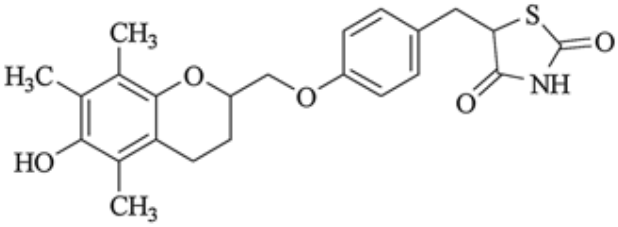

441.55; 184 – 186 °C

[74]

5-[[4-[(3,4-Dihydro-6-hydroxy-2,5,7,8-tetramethyl-2*H*-1-benzopyran-2-yl) methoxy]phenyl]methyl]-2,4-thiazolidinedione

---

Pioglitazone (Table 3) is approximately ten times more potent in both, animals, and humans than troglitazone. The highest potency shows *rosiglitazone* (Table 3), which is approximately 100-fold more active than troglitazone [69]. Both compounds showed better tolerability in clinical trials than troglitazone. Both drugs were introduced into the US market in 1999, and differently from troglitazone, for these TZDs, hepatotoxicity is less serious. However, both drugs cause fluid retention, therefore, heart failure could be a potential risk for special patients [71].

[Top of Page]

References

1. R. A. DeFronzo, *Diabetes* **37** (1988) 667 – 681. Links
2. A. J. Scheen, P. F. Lefèbvre in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 7 – 42.
3. C. R. Loubatières, *Soc. Biol. (Paris)* **138** (1944) 766 – 767.
4. H. Franke, J. Fuchs, *Dtsch. Med. Wochenschr.* **80** (1955) 1449 – 1452. Links
5. H. Ruschig *et al.*, *Arzneimittelforsch.* **8** (1958) 448 – 454.
6. W. Aumüller *et al.*, *Arzneim.-Forsch./Drug Res.* **16** (1966) 1640 – 1641.
7. K. Geisen, *Arzneim.-Forsch./Drug Res.* **38** (1988) 1120 – 1130.
8. C. M. Clark, A. W. Helmy, *Drugs of Today* **34** (1998) 401 – 408. Links
9. Hoechst, DE-AS 1135891, 1960; Eli Lilly, DE-AS 1177631, 1961.
10. E. Haack, *Arzneim.-Forsch.* **8** (1958) 444 – 448.
11. Boehringer Mannheim, DE-AS 1117103, 1953.
12. Hoffmann-La Roche, DE-AS 1695201, 1967.
13. Sciences Union et Cie., BE 693702, 1966; DD 113223, 1973.
14. R. Weyer, V. Hitzel, *Arzneim.-Forsch./Drug Res.* **38** (1988) 1079 – 1080.
15. V. Ambrogi *et al.*, *Arzneim.-Forsch.* **21** (1971) 200 – 204.
16. Thomae, DE-AS 2000339, 1970.
17. Bayer, DE-AS 1670952, 1967.
18. H. Plümpe, H. Horstmann, W. Puls, *Arzneim.-Forsch./Drug Res.* **24** (1974) 363 – 374.
19. K. Gutsche *et al.*, *Arzneim.-Forsch./Drug Res.* **14** (1964) 373 – 376.
20. Upjohn, US 3063903, 1961.
21. PCT Int.- Pat. Appl. WO 93/00337, 1993 (W. Grell *et al.*).
22. H. Shinkai *et al.*, *J. Med. Chem.* **32** (1989) 1436 – 1441. Links
23. H. Bischoff, H. E. Lebovitz in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 650 – 696.
24. J. A. Balfour, D. Faulds, *Drugs & Aging* **13** (1998) 173 – 180.

25. T. Ikenoue *et al.*, *Br. J. Pharmacol.* **120** (1997) 137 – 145. Links
26. A. H. Karara, B. E. Dunning, *J. Clin. Pharmacol.* **39** (1999) 172 – 179. Links
27. Y. Sunaga *et al.*, *Eur. J. Pharmacol.* **431** (2001) 119–125. Links
28. U. Panten, M. Schwanstecher, C. Schwanstecher in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 129 – 159.
29. A. M. Hansen *et al.*, *Br. J. Pharmacol.* **144** (2005) 551–557. Links
30. R. Göke *et al.*, *J. Biol. Chem.* **268** (1993) 19650–19655. Links
31. A. H. Barnett, *Drugs of Today* **41** (2005) 563–578. Links
32. M. Nauck, J. Meier, *Regul Pept* **128** (2005) 135–148. Links
33. R. Mentlein *et al.*, *Eur J Biochem* **214** (1993) 829–835. Links
34. C. F. Deacon *et al.*, *Diabetes* **44** (1995) 1126–1131. Links
35. C. F. Deacon *et al.*, *Diabetes* **47** (1998) 764–769. Links
36. J. A. Pospisilik *et al.*, *Diabetes* **51** (2002) 943–950. Links
37. M. K. Reimer *et al.*, *Eur. J. Endocrinol.* **146** (2002) 717–727. Links
38. C. A. Abbott *et al.*, *Eur. J. Biochem.* **267** (2000) 6140–6150. Links
39. C. Olson, N. Wagtmann, *Gene* **299** (2002) 185–193. Links
40. C. F. Deacon *et al.*, *Exp. Opin. Investig. Drugs* **13** (2004) 1091–1102.
41. L. A. Sorbera, J. Castaner, *Drugs Fut.* **30** (2005) 337–343.
42. B. Ahrén *et al.*, *Diabetes Care* **27** (2004) 2874–2880. Links
43. B. Ahrén *et al.*, *Diabetes Care* **25** (2002) 869–875. Links
44. B. Ahrén *et al.*, *J. Clin. Endocrinol. Metab.* **89** (2004) 2078–2084. Links
45. C. K. Watanabe, *J. Biol. Chem.* **33** (1918) 253. Links
46. R. Beckmann in H. Maske (ed.): *Handbook of Experimental Pharmacology*, vol. **29**, Springer, Berlin – Heidelberg – New York 1971, pp. 439 – 596.
47. N. F. Wiernsperger in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 305 – 358.
48. L. S. Hermann in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 373 – 407.
49. W. Creutzfeld, W. R. Fölsch (eds.): *Delaying Absorption as a Therapeutic Principle in Metabolic Disease*, Thieme Verlag, Stuttgart – New York 1983.
50. H. Bischoff, *Clin. Invest. Med.* **18** (1995) 303 – 311. Links
51. B. Junge, M. Matzke, J. Stoltefuss in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 411 – 482.
52. E. Truscheit *et al.*, *Prog. Clin. Biochem. Med.* **7** (1988) 17 – 99.
53. W. Puls in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 497 – 534.
54. W. Puls *et al.*, *Naturwissenschaften* **64** (1977) 536. Links
55. B. Junge *et al.*, *Carbohydr. Res.* **128** (1984) 235 – 268. Links
56. EP 49858, 1982 (prior DE 3038901, 1980) (G. Kinast, M. Schedel, W. Köbernick).
57. S. Horii *et al.*, *J. Med. Chem.* **29** (1986) 1038 – 1046. Links
58. H. P. Krause, H. J. Ahr in J. Kuhlmann, W. Puls (eds.): *Handbook of Experimental Pharmacology*, vol. **119**, Springer, Berlin – Heidelberg – New York 1996, pp. 541 – 555.
59. B. Göke, H. Fuder, *Digestion* **56** (1995) 493 – 501. Links
60. T. Sohda *et al.*, *Chem. Pharm. Bull.* **30** (1982) 3563. Links
61. A. Y. Chang, B. M. Wyse, B. J. Gilchrist, T. Peterson, A. R. Diani, *Diabetes* **32** (1983) 830 – 838. Links
62. M. J. Reginato, M. A. Lazar, *Trends Endocrinol. Metab.* **10** (1999) 9 – 13. Links
63. K. Schoonjans, B. Staels, J. Auwerx, *Biochim. Biophys. Acta* **1302** (1996) 93 – 109. Links
64. J. Auwerx, G. Martin, M. Guerre-Millo, B. Staels, *J. Mol. Med.* **74** (1996) 347 – 351. Links
65. T. M. Willson *et al.*, *J. Med. Chem.* **39** (1996) 665 – 668. Links
66. J. Berger *et al.*, *Endocrinology* **137** (1996) 4189 – 4195. Links
67. T. Fujiwara, S. Yoshioka, T. Yoshioka, I. Ushiyama, H. Horikoshi, *Diabetes* **37** (1988) 1549 – 1558. Links
68. B. Hulin, P. A. McCarthy, E. M. Gibbs, *Curr. Pharmac. Des.* **2** (1996) 85 – 102.
69. C. Day, *Diabetic Medicine* **16** (1999) 179 – 192. Links
70. J. Kohlroser *et al.*, *Am. J. Gastroenterol.* **95** (2000) 272–276. Links
71. A. J. Scheen, *Drug Saf.* **27** (2004) 841–856. Links
72. T. Sohda *et al.*, *Arzneim.-Forsch./Drug Res.* **40** (1990) 37 – 42.
73. B. C. Cantello *et al.*, *J. Chem. Soc. Perkin Trans. 1*, **22** (1994) 3319 – 3324.
74. T. Yoshioka *et al.*, *J. Med. Chem.* **32** (1989) 421 – 428. Links

[Top of Page]

Related Articles

Insulin ⟶ Peptide and Protein Hormones; GLP-1 ⟶ Peptide and Protein Hormones

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Antiemetics

GERHARD RACKUR, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany

1. **Tropane Alkaloids** 1
2. **Benzhydryl Derivatives** 1
3. **Phenothiazine Derivatives** 2
4. **Metoclopramide** 3
5. **References** 3

Nausea and vomiting (emesis) can be induced by certain types of drugs (emetics) or in most persons by motion (motion sickness). Nausea and vomiting can result from surgery, irradiation, cancer chemotherapy, pregnancy, and gastrointestinal disorders.

Emetics are drugs that induce vomiting. They are used only on orally intoxicated patients who for some reason cannot undergo lavage, a much safer method of emptying the stomach. However, even in such cases emetics have limited value because of their side effects. The most widely used emetic is apomorphine hydrochloride [*41372-20-7*], $C_{17}H_{18}ClNO_2$, M_r 303.79, which induces vomiting within a few minutes of subcutaneous injection. The usual dose is 5 mg.

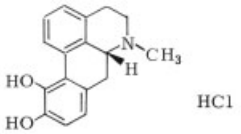

Apomorphine hydrochloride

Antiemetics are drugs that suppress nausea and vomiting. In many cases their mechanism of action is not fully understood. Generally antiemesis results from direct suppression of the chemoreceptor trigger zone, by action on the vestibular apparatus, or by local or central effects. The most widely used antiemetics are tropane alkaloids, benzhydryl derivatives, and phenothiazines.

1. Tropane Alkaloids

Tropane alkaloids were the first drugs used to prevent motion sickness.

Scopolamine [*51-34-3*] and L-hyoscyamine [*101-31-5*] are the most commonly used alkaloids (→ Alkaloids, Chap. 4.). A prophylactic dose of 0.1 mg scopolamine hydrobromide protects against the symptoms of motion sickness in 75 % of all cases. Side effects are rare at this dosage. Intramuscular injection of 0.2 mg can relieve motion sickness symptoms. L-Hyoscyamine acts in a similar way but is less potent.

2. Benzhydryl Derivatives

Chlorphenoxamine [*77-38-3*], 2-[1-(4-chlorophenyl)−1-phenylethoxy]-*N*,*N*-dimethylethyl-amine, $C_{18}H_{22}ClNO$, M_r 303.84, *bp* 150 – 155 °C (7 Pa). The hydrochloride [*562-09-4*], $C_{18}H_{23}Cl_2NO$, M_r 340.30, *mp* 128 °C, is soluble in water.

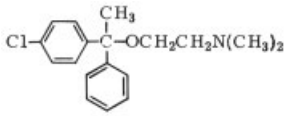

Chlorphenoxamine is prepared by action of methylmagnesium chloride on 4-chlorobenzophenone and subsequent alkylation of the carbinol with *N*,*N*-dimethyl-2-chloroethylamine [1].

Chlorphenoxamine is also used in combination with 8-chlorotheophylline.

Trade names: Rodavan (Asta), Phenoxene (Dow).

Cyclizine [*82-92-8*], 1-(diphenylmethyl)−4-methylpiperazine, $C_{18}H_{22}N_2$, M_r266.39, *mp* 106 °C, is soluble in ethanol and chloroform and partly soluble in water. The hydrochloride [*303-25-3*], $C_{18}H_{23}ClN_2$, M_r 302.85, *mp*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_009

285 °C (decomp.), is moderately soluble in water. The lactate [*5897-19-8*], $C_{21}H_{28}N_2O_3$, M_r 356.47, is suitable for intramuscular injection.

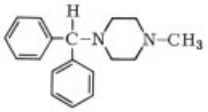

Synthesis: alkylation of *N*-methylpiperazine with benzhydryl chloride [2].

Trade names: Collox (Wellcome), Marezine (Wellcome).

Diphenhydramine [*58-73-1*], 2-diphenylmethoxy-*N*,*N*-dimethylethylamine, $C_{17}H_{21}NO$, M_r 255.53, *bp* 150 – 165 °C (0.270 kPa); hydrochloride [*147-24-0*], $C_{17}H_{22}ClNO$, M_r 291.82, *mp* 166 – 170 °C. One gram of diphenhydramine is soluble in 1 mL of water, 2 mL of ethanol, 2 mL of chloroform, or 50 mL of acetone; the compound is very slightly soluble in benzene or ether.

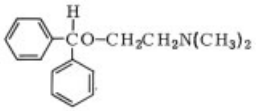

Synthesis: diphenylmethyl bromide, *N, N*dimethylaminoethanol, and sodium carbonate are heated in toluene [3].

Trade name: Benadryl (Parke-Davis).

As an antiemetic diphenhydramine is used most widely as the 8-chlorotheophyllinate: *Dimenhydrinate*
[*523-87-5*], $C_{17}H_{21}NO \cdot C_7H_7ClN_4O_2$, M_r 469.96, *mp* 102 – 107 °C, which is freely soluble in ethanol and chloroform, slightly soluble in water, and almost in-soluble in ether.

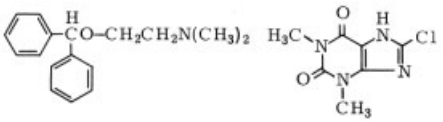

Dimenhydrinate is prepared by heating the components (1 : 1 mixture) in ethanol [4].

Trade names: Vomex (Endopharm), Dramamine (Searle).

Diphenidole [*972-02-1*], 1,1-diphenyl-4-piperidinobutanol, $C_{21}H_{27}NO$, M_r 309.43, *mp* 104 – 105 °C. The hydrochloride [*3254-89-5*], $C_{21}H_{28}ClNO$, M_r 345.91, is used for tablets and injections.

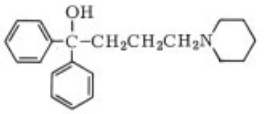

Diphenidole is prepared by reacting the Grignard reagent of 1-(3-chloropropyl)piperidine with benzophenone in tetrahydrofuran. The crude product is recrystallized from 2-propanol [5].

Trade names: Vontrol (Smith Kline & French), Cephadol (Nippon Shinyaku).

Meclozine [*569-65-3*], 1-[(4-chlorophenyl)-phenylmethyl]−4-[(3-methylphenyl)methyl]-piperazine, $C_{25}H_{27}ClN_2$, M_r 390.96, *bp* 230 °C (270 Pa). The dihydrochloride [*1104-22-9*], $C_{25}H_{29}Cl_3N_2$, M_r 463.90, *mp* 224 °C (decomp.), is readily soluble in chloroform and pyridine, slightly soluble in dilute acids and ethanol, and insoluble in water or ether.

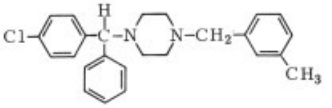

Meclozine is prepared by alkylation of (3-methylbenzyl)piperazine with 4-chlorobenzhydrylchloride [6].

Trade names: Antivert (Roerig), Bonamine (Pfizer).

3. Phenothiazine Derivatives

Many phenothiazines are antiemetics as well as neuroleptics (→ Phenothiazine and Derivatives).

Prochlorperazine [*58-38-8*], 2-chloro-10-[3- (4-methyl-1-piperazinyl)-propyl]−10*H*-pheno-thiazine (**1**, X = Cl), $C_{20}H_{24}ClN_3S$, M_r 373.94, *mp* 228 °C; for synthesis → Phenothiazine and Derivatives.

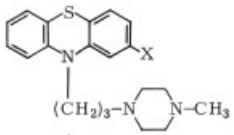

1

Usually the free base is administered rectally. The *ethanedisulfonate* of prochlorperazine [*1257-78-9*], $C_{22}H_{30}ClN_3O_6S_3$, M_r 564.13, and the dimaleate [*84-02-6*], $C_{28}H_{32}ClN_3O_8S$, M_r606.09, are used as oral antiemetics.

Trade name: Compazine (Smith Kline & French).

Thiethylperazine [*1420-55-9*], 2-ethylthio-10- [3-(4-methyl-1-piperazinyl)propyl]−10*H*-phenothiazine (**1**, X = SC_2H_5), $C_{22}H_{29}N_3S_2$, M_r 399.61, *bp* 227 °C (1 Pa), acts both as a peripheral and as a central antiemetic. The malate [*52239-63-1*], $C_{30}H_{41}N_3O_{10}S_2$, M_r 667.79, is suitable for intramuscular injection, whereas the maleate [*1179-69-7*], $C_{30}H_{37}N_3O_8S_2$, M_r 631.76 is an oral antiemetic.

Trade name: Torecan (Boehringer Ingelheim).

Trifluopromazine [*146-54-3*], 10-[3-(4-methyl-1-piperazinyl)-propyl]−2-trifluoromethyl-10*H*-phenothiazine (**1**, X = CF_3), $C_{21}H_{24}F_3N_3S$, M_r 407.49, *bp* 202 – 210 °C (80 Pa), is an oral antiemetic. For its preparation → Phenothiazine and Derivatives.

Trade name: Psyquil (Heyden).

Other phenothiazine derivatives that have their chief use as tranquilizers and neuroleptics are used as antiemetics occasionally: chlorpromazine, hydroxyzine, and haloperidole.

4. Metoclopramide

Metoclopramide [*364-62-5*], $C_{14}H_{22}ClN_3O_2$, M_r 299.81, is the most important antiemetic in Japan.

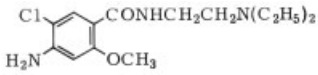

Metoclopramide is prepared by the reaction of *N,N*-diethylethylenediamine with 4-acetamido-5-chloro-2-methoxybenzoyl chloride [7].

Trade names: Cerucal (VEB Arzneimittelwerk), Maltyl (Diabetylinges.), Maxolon (Beecham), Metoclol (Toyama), Moriperan (Morishita), Peraprin (Taiyo), Pramiel (Teisan), Reliveran (Finadiet).

Monohydrochloride monohydrate [*54143-57-6*], $C_{14}H_{23}Cl_2N_2O_3$, M_r 338.26, *mp* 182.5 – 184 °C, is a colorless powder, soluble in water.

Trade names: Eucil (Farmasa), Maxeran (Nordic), Paspertin (Kali-Chemie), Plasil (Lepetit), Reglan (Robins).

Dihydrochloride monohydrate [*2576-84-3*], $C_{14}H_{24}Cl_3N_3O_2$, M_r 374.72, *mp* 145 °C (decomp.), is soluble in water and methanol and partially soluble in ethanol.

Trade names: Gastronerton (Dolorgiet), Imperan (Delagrange, Bender), Primperan (Fujisawa Yakuhin).

5. References

1. Asta Werke, US 2785202, 1957 (H. Arnold, N. Brock, E. Kühas).
2. R. Baltzly et al., *J. Org. Chem.* 14 (1949) 775. Burroughs Wellcome, US 2630435, 1953 (R. Baltzly, J. C. Castillo).
3. Parke-Davis, US 2421714, 2427878, 1947 (G. Rieveschl, Jr.).
4. J. W. Cusic, *Science* 109 (1949) 574. Searle, US 2499058, 2534813, 1950 (J. W. Cusic).
5. Ciba, US 2411664, 1946 (K. Miescher, A. Marxer).
6. H. G. Morren, BE 502889, 1951.
7. Soc. d'Études Sci. et Ind. de l'Île de France, BE 620543, 1962; US 3177252, 1965 (M. L. Thominet).

Antiepileptics

Klaus Unverferth, ASTA Medica Group, Corporate Research and Development, Dresden, Germany
Chris Rundfeldt, ASTA Medica Group, Corporate Research and Development, Dresden, Germany

1. Introduction 1
1.1. Classification 2
1.2. Treatment of Epilepsy 2
1.3. Discovery Strategies for New Antiepileptic Drugs 3
2. Antiepileptic Drugs 3
2.1. Phenytoin 4
2.2. Carbamazepine and Oxcarbazepine . 5
2.3. Valproic Acid 6
2.4. Ethosuximide and Trimethadione . . . 7
2.5. Phenobarbital and Primidone 7
2.6. Benzodiazepines 8
2.7. New Drugs 9
2.8. Other Antiepileptic Drugs 11
3. References 12

1. Introduction

Epilepsy is a condition of recurrent, paroxysmal seizures. The epileptic seizure represents an abnormal synchronized discharge among a large population of central neurons. Depending on the neuronal populations involved, a disturbance of movement, sensation, behavior, or consciousness can result from the discharges [1]. However, a seizure can only be seen as a symptom which principally can be induced in every individual if strong enough stimulation is applied. Epilepsy is characterized by the increased susceptibility to experience seizures upon exogenous or endogenous stimuli. Clinically, only a patient who has experienced three or more seizures is considered to be epileptic [2].

About 0.5 – 1 % of the world's population (i.e., more than 50×10^6 people) suffer from epilepsy, which makes it one of the most common neurological disorders; no major difference is found between developed and developing countries [3]. The disease usually starts in childhood with a second peak in elderly people, but people can become epileptic at any age. Once acquired, patients usually suffer from the disease throughout their lives [3].

The diagnosis of epilepsy is based on clinical observations of the seizures and on recording of changes in the electrical activity of the brain using scalp electrodes (electroencephalography, EEG). Since patients in most cases lose consciousness during seizures and since EEG recordings between seizures can be fully normal, a preliminary diagnosis is often based on seizure descriptions obtained from eye witnesses. Ancillary investigations exploring the etiology of the condition include different imaging techniques, seizure provocation experiments, and depth electrode recordings. Epileptic seizures can be distinguished from other convulsive-like states not involving synchronous activation of neuronal cells. Examples of these states include transient ischemic attacks or hyperventilation resulting in loss of righting reflexes due to fainting.

The pathophysiology of epilepsy is diverse, involving both intrinsic and extrinsic factors. Developmental or congenital anomalies such as neuronal dysplasia, vascular malformations, and cerebral atrophy as well as cerebral tumors, meningitis, encephalitis or, cerebrovascular diseases are possible intrinsic factors. Birth trauma, head injuries, anoxic disturbances induced by respiratory arrest, intoxication with heavy metals or other toxins and endocrine disorders such as uremia, hypoglycemia, or hypocalcemia are exogenous factors [4]. However, in most patients the epilepsy is of idiopathic origin. The precise mechanism involved in the increased susceptibility to synchronized discharges is still unknown. However, based on experiments with anticonvulsant and proconvulsant drugs and on biochemical findings from human epileptic tissue, several theories have been proposed. The increased excitability could be related to both neurotransmitter and ion channel function. An overactive excitatory and a reduced inhibitory neurotransmitter system has been found in these tissues. Reported changes in ion channel func-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_013

tion resulted in a diminished membrane potential and a defective repolarization of the action potential. Besides these functional results, several genetic changes have been reported in both epilepsy-prone animals and in subpopulations of humans with epilepsy leading to hereditary epileptic syndromes [5], [6].

1.1. Classification

Seizures are stereotyped within individuals, but vary considerably between patients. This necessitates a system of classification for both seizures (symptomatic classification) and epileptic syndromes (syndromatic classification). The most generally accepted classification system is that of the International League Against Epilepsy [7].

Classification of Seizures. Seizures are classified as to whether their onset is partial (focal) or generalized.

Partial seizures are further subdivided according to whether consciousness is retained throughout the seizure (simple partial seizures) or impaired at some point (complex partial seizures). Patients having *simple partial seizures* show jerks or convulsions on one extremity. During *complex partial* seizures, the patient is in a state of reduced consciousness, not fully recognizing his surroundings, experiencing hallucinations and performing automatic gestures, such as talking, fumbling with clothing, or walking around. Any partial seizure can become *secondarily generalized* into a tonic – clonic convulsion.

Generalized seizures are further subdivided according to whether only consciousness is lost (absence seizures) or muscle activity is involved (myoclonic, clonic, tonic, tonic – clonic, or atonic seizures). An *absence seizure* is characterized by a sudden interruption of consciousness lasting only a few seconds without any other symptoms or with only a quick jerk of the eyes or hands. *Myoclonic seizures* are characterized by sudden jerks on the musculature of the limbs, the head, or the whole body, appearing in connection with absence seizures or as a separate entity. During an *atonic seizure* patients exhibit a sudden loss of muscle tone resulting in a fall to the ground (drop attack). Consciousness may be lost but is immediately regained. *Tonic-clonic seizures* start with a sudden loss of consciousness and a fall to the ground; a few seconds of intense tonic spasms of major muscles are followed by generalized jerks of the whole body. These convulsions gradually cease, leaving the patient comatose and flaccid. A gradual recovery of consciousness follows, often combined with confusion and drowsiness. *Status epilepticus* is a condition in which seizures follow each other without a recovery of consciousness between them. Both absence seizures and convulsive seizures can occur as status epilepticus.

Classification of Syndromes. A classification of epileptic syndromes is based not only on seizure characteristics but on more general features of the disease such as age of onset, type of EEG abnormalities, associated neurological features, or possible cause of disease including genetic factors [8], [9]. The International League Against Epilepsy [7] has listed more than 30 different epileptic syndromes.

1.2. Treatment of Epilepsy

The treatment of epilepsy is based on pharmacotherapy to suppress the seizure attacks, behavioral therapy to avoid seizure initiating situations, and on surgical resection of the seizure focus. As yet no ideal and general treatment regime exists for epileptic patients; the therapy rather has to be adopted individually. With optimal usage of the aforementioned strategies, seizures can be fully controlled in about 70 – 80 % of patients, however, side effects often have to be accepted [10]. Although surgical resection of the hyperexcitable focus is considered the only available cure for epilepsy, only a small percentage of patients, i.e., those having a defined singular focus which can be resected without disturbing vital brain functions, is eligible for surgery. Based on this well-defined population, the success rate of surgery is high but serious side effects can occur such as unwanted changes in personality or loss of brain functions like short term memory or ability to speak [11]. Behavioral treatment is aimed at avoiding endogenous or exogenous stimuli and is therefore only of symptomatic nature. Generally, the same is also

true for pharmacotherapy, which is aimed at suppressing the seizure attacks. The pharmacotherapy must be adapted both to seizure type and epileptic syndrome.

Antiepileptics (anticonvulsants) can be divided into drugs active against generalized absence seizures, those against partial seizures, and in drugs with a broad spectrum of activity. As of 2000, a syndrome-related therapy is in its beginnings. Although for some syndromes a selective therapy exists, such strategy is still limited due to a lack of selective drugs for most syndromes. Since seizures are often resistant to pharmacotreatment, two or more anticonvulsant drugs are concomitantly administered to improve seizure control [2], [12].

1.3. Discovery Strategies for New Antiepileptic Drugs

Since the pathomechanisms of epilepsy are not fully understood, it is impossible to define a straightforward strategy to develop drugs aimed at a molecular target which is solely and critically involved in generation of seizures. Therefore, nonrational (older) strategies of drug development are still valid in this field [13], [14].

Serendipitous search for new drugs, based on anticonvulsant effects observed in in vivo or in vitro screening models of seizures, may lead to drugs with unknown modes of action but potent anticonvulsant activity. This approach, however, is limited by the animal models used [15]. If models are selected based on efficacy of known drugs, only drugs with a similar anticonvulsant profile might evolve. If models are selected to mimic the disease state or drug resistance [16], the predictive value might still be limited. A second possibility to find new marketable drugs is to optimize the structure of known anticonvulsants using predictive animal models and structure – activity relationship (SAR) modeling to improve anticonvulsant potency while reducing unwanted effects or to optimize the physicochemical behavior. This approach is hampered by the same restrictions as the first one, since it is also based on experimental model results.

Although a rational approach may be limited, the knowledge of mechanisms of action of anticonvulsant drugs and of possible pathomechanisms involved in epilepsy can generate several approaches. A straightforward possibility is to investigate the mechanisms of action of currently known antiepileptic drugs. Such an approach will lead to drug discovery targets which are not necessarily involved in epilepsy but in antiepileptic drug action. Examples of drugs are use-dependent sodium channel blockers, T-type calcium channel blockers, and drugs enhancing γ-aminobutyric acid (GABA) ergic or reducing excitatory glutamatergic neurotransmission. However, the resulting new drugs may not differ in therapeutic potential from the old ones.

Another approach is to investigate human tissue resected from epileptic foci during surgery or to analyze the genom of patients with hereditary epileptic syndromes to track down the mutated gene and consequently the misfunctional protein. In epileptic tissue, levels of neurotransmitters, densities of receptors, or gene expression levels can be compared with those in nonepileptic tissue. From this information, a theory can be put forward on how to compensate for the observed changes. Examples for such targets are among others (*R,S*)-2-amino-3-(3-hydroxy-3-methylisoxazol-4-yl)propionic acid (AMPA) antagonists [17] or potassium channel openers [18]. These targets can be used in modern automated screening systems. However, although this approach may lead to novel targets, these may only coexist with the real causes or the changes in gene expression may only be a compensation initiated by the seizures. This would render even high affinity ligands useless for the treatment of epilepsy. In summary, no superior strategy can be defined to develop new antiepileptic drugs. An approach which involves elements of several strategies may be best in developing new drugs which are superior to the existing ones both in regard to efficacy and tolerability.

2. Antiepileptic Drugs

Landmarks in Antiepileptic Drug Treatment:

1857	Bromides
1912	Phenobarbital
1938	Phenytoin
1958	Ethosuximide
1967	Carbamazepine
1974	Valproic acid
1989	Vigabatrin

1991	Lamotrigine
1993	Gabapentin
1995	Topiramate
1996	Tiagabine

Phenobarbital was the first effective organic antiepileptic agent and the starting point for structural variation leading to hydantoins (phenytoin), succinimides (ethosuximide), and oxazolidinediones (trimethadione). These active chemical classes exhibit certain structural similarities, they all have an imide group and a carbon atom bearing four nonhydrogen substituents. In spite of chemical similarities, the structural variation leads to different anticonvulsant effects. Like phenobarbital, phenytoin is highly effective in suppressing generalized tonic – clonic as well as simple complex partial seizures, and ethosuximide is effective in suppressing absence seizures. Although structural similarities exist between these classes, phenytoin and ethosuximide exert the activity through different mechanisms. Thus, phenytoin blocks voltage dependent Na^+-channels and ethosuximide blocks Ca^{2+}-T-channels.

Therefore, structure-activity relationships can be discussed only within well-defined chemical classes: hydantoins [19], dibenz[b,e]azepines [20], acetic acids [21], succinimides [22], oxazolidinediones [23], barbiturates [24], benzodiazepines [25] or for compounds of different chemical classes which exert anticonvulsant activity essentially through the same mechanism [26].

Current therapeutic use is the criterion for listing a compound in this article. Since the 1980s several new antiepileptic drugs have entered the market. At present the new drugs are an alternative to the old drugs for 25 % of patients with refractory epileptic seizures. The search for effective and cheap drugs which are well tolerated and devoid of side effects still persists. Books on the old and new antiepileptic drugs [27], [28] and reviews of the new antiepileptic drugs [29], [30] have appeared, additional mechanistic aspects are discussed in [31]. The market value of antiepileptics has grown significantly. Worth $\$560 \times 10^6$ in 1987, the market topped $\$2000 \times 10^6$ in 1996 and, if current growth rates continue, could exceed $\$4000 \times 10^6$ by 2000 (currently, 50 % of the antiepileptic drugs are sold in North America). Much of this spectacular growth is due to the premium-priced products entering the market, but the well-established and cheaper drugs (phenytoin, carbamazepine, and valproate) are the mainstay of treatment and still dominate the market [32].

2.1. Phenytoin

Phenytoin [*57-41-0*], 5,5-diphenyl-2,4-imidazolidinedione, $C_{15}H_{12}N_2O_2$, M_r 252.26, *mp* 295 – 298 °C; practically insoluble in water. 1 g dissolves in about 75 ml of ethanol or 30 ml of acetone.

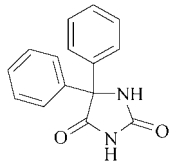

Synthesis:

1) Condensation of benzil with urea [33]

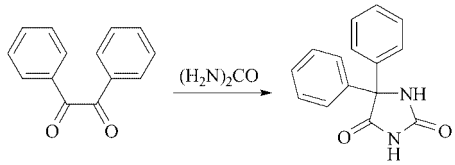

2) Bucherer reaction [34]

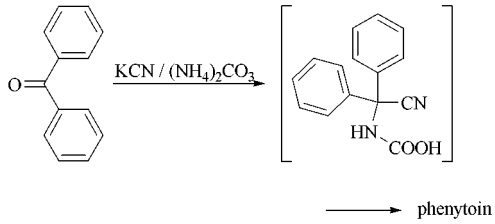

Trade names: Difhydan (Leopharm), Dihycon (CMC), Di-Hydan (Carrion), Dilantin (Parke – Davis), Ekko (Fleming), Hydantin (Lääkefarmos), Hydantol (Fujinaga – Sankyo), Lehydan (Abigo), Phenhydan (Desitin Arzneimittel), Phenytoin AWD (Arzneimittelwerk Dresden), Zentropil (Nordmark)

Phenytoin sodium [*630-93-3*], sodium 5,5-diphenylhydantoin, $C_{15}H_{11}NaN_2O_2$, M_r 274.26; 1 g dissolves in ca. 66 ml of water. The aqueous solution is turbid unless the pH is adjusted above 11.7, which is the pH of the saturated solution. 1 g dissolves in 10.5 mL of alcohol. Practically insoluble in ether and chloroform.

Trade names: Antisacer (Wander), Danten (McKesson), Diphantoin (Katwijk), Diphenin (Lannett), Epanutin (Parke – Davis), Minetoin (Helvepharm), Phenhydan (Desitin), Solantyl (Roussel – UCLAF), Zentropil (Nordmark).

Phenytoin and its sodium salt are effective in primarily and secondarily generalized tonic – clonic seizures and are also indicated in simple and complex partial seizures, but not in absence seizures. The intravenous form can be used in status epilepticus. The clinical use of phenytoin is complicated by its narrow therapeutic margin and its saturation kinetics. Side effects include nystagmus, ataxia, slowly progressing impairment of mental function, skin rash, gingival hyperplasia, and hirsutism. The main mechanism of action of phenytoin is based on the blockade of voltage-dependent Na^+-channels.

Phenytoin and its sodium salt are widely used in the USA (18.1 % of the sales in 1995, Western Europe 4.2 %)

Other Hydantoins.

Mephenytoin [*50-12-4*], 5-ethyl-3-methyl-5-phenyl-2,4-imidazolidinedione, can be substituted for phenytoin where the peripheral neuropathy and hirsutism create problems. Side effects, including rash, fever, and fatal blood dyscrasia, prevent the use as an anticonvulsant drug of first choice.

Trade name: Insulton (Sandoz).

Ethotoin [*86-35-1*], 3-ethyl-5-phenyl-2,4-imidazolidindione, has never achieved the status of a drug of choice.

Trade name: Peganone (Abbott).

2.2. Carbamazepine and Oxcarbazepine

Carbamazepine [*298-46-4*], 5*H*-dibenz[*b,f*]azepine-5-carboxamide, $C_{15}H_{12}N_2O$, M_r 236.26, *mp* 189 – 193 °C; soluble in chloroform, DMF, ethylene glycol monomethyl ether, or methanol, only slightly soluble in ethanol or glacial acetic acid, virtually insoluble in water.

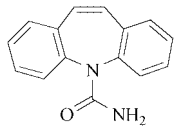

Synthesis: 5*H*-dibenz[*b,f*]azepine is first reacted with phosgene and then with ammonia [35].

Carbamazepine is the only antiepileptic combining structural characteristics of classical antiepileptics with those of polycyclic psychoactive agents, such as imipramine (→ Psychopharmacological Agents, Chap. 4.1.1.).

Carbamazepine is the drug of choice for complex partial seizures and is also effective against simple partial and generalized tonic – clonic seizures, but not against absence seizures. Carbamazepine also has an antineuralgic and a psychotropic effect, i.e., a positive influence on disturbed mood and behavior, and little or no detrimental effect on intellectual functions. Side effects include skin rash, headache, nausea, fatigue, vertigo, and ataxia.

The substance is considered to exert its main effect by inhibiting voltage-dependent Na^+-channels.

Carbamazepine is widely used worldwide (15.5 % of the sales in the USA in 1995, 34.7 % in Western Europe).

Trade names: Biston (Spofa), Calepsin (Orion), Carbagamma (Wörwag), Carbamazepin – GRY (GRY – Pharma), Carbamazepin Heumann (Heumann), Carbamazepin – neuraxpharm (neuraxpharm), Carbamazepin – ratiopharm (ratiopharm), Carbamazepin 200 Stada (Stadapharm), Carbelan (Elan), Epitol (Lemmon), Finlepsin (Arzneimittelwerk Dresden), Fokalepsin (Promonta Lundbeck), Sirtal (Sanofi Wintrop), Tegretol (Novartis), Timonil (Desitin Arzneimittel).

Oxcarbazepine [*28721-07-5*], 10,11-dihydro-10-oxo-5*H*-dibenz[*b,f*]azepin-5-carboxamide, $C_{15}H_{12}N_2O_2$, M_r 252.27, *mp* 215 – 216 °C; virtually insoluble in water.

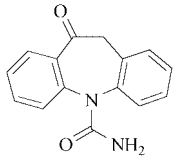

Synthesis: Oxcarbazepine can be obtained in two different ways.

1) Reaction of 10-methoxy-5*H*-dibenz[b,f]azepine (**1**) with phosgene gives the 5-chlorocarbonyl compound, treatment with NH_3 affords 10-methoxy-5*H*-dibenz[*b,f*]azepine-5-carboxamide (**2**), which is hydrolyzed with diluted HCl to oxcarbazepine [36].

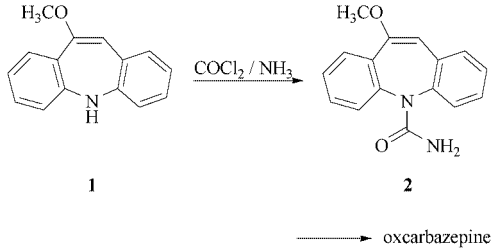

2) Nitration of 5-cyano-5*H*-dibenz[*b,f*]azepine (**3**) with $NaNO_3$ in acetic anhydride/acetic acid gives 5-cyano-10-nitro-5*H*-dibenz[*b,f*]azepine (**4**), which is treated with BF_3 and powdered iron in acetic acid [37].

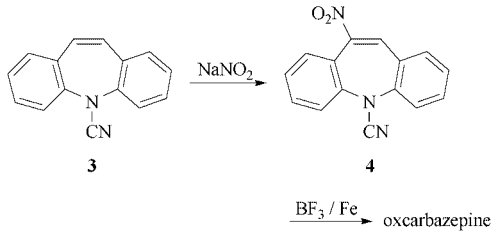

Oxcarbazepine is a new antiepileptic drug (launched in 1990). Compared with its parent drug, carbamazepine, it is metabolized via a different pathway and may have milder side effects. The efficacy of the two drugs seems to be similar.

Trade name: Trileptal (Novartis).

2.3. Valproic Acid

Valproic acid [*99-66-1*], 2-propylpentanoic acid, $C_8H_{16}O_2$, M_r 144.21; colorless liquid, *bp* 221 – 222 °C at 101.3 kPa, $n_D^{25} = 1.425$, $d_4^{25} = 0.904$; soluble in most organic solvents, including methanol, chloroform, and ether, solubility in water: 1.27 mg/mL.

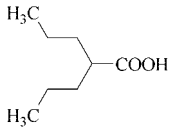

Trade names: Convulex (Promonta Lundbeck), Mylproin/Orfiril (Desitin Arzneimittel), Mylproin (ICI).

Sodium valproate [*1069-66-1*], $C_8H_{15}NaO_2$, M_r 166.20, colorless crystalline powder, hygroscopic, very soluble in water (ca. 0.66 g/mL of solution), soluble in ethanol (ca. 0.2 g/mL of solution), practically insoluble in chloroform and diethyl ether.

Trade names: Convulex (Byk – Gulden, Promonta Lundbeck), Depakene/Depakote/Valcote (Abbott), Depakine/Epilim/Ergenyl (Sanofi), Leptilan (Novartis), Orfiril (Desitin Arzneimittel).

Calcium valproate [*138995-18-3*], $C_{16}H_{30}CaO_4 \cdot 2H_2O$, M_r 362.5

Trade name: Convulsofin (Arzneimittelwerk Dresden).

Magnesium valproate [*62959-43-7*], $C_{16}H_{30}MgO_4$, M_r 310.73.

Trade name: Logical (Armstrong).

Synthesis: Diethyl malonate is alkylated with 1-bromopropane and then hydrolyzed and decarboxylated [38].

Valproic acid and its salts are major antiepileptic drugs for the treatment of absence, myoclonic, and generalized tonic – clonic seizures and also may be effective against complex partial seizures. The most common side effects are gastrointestinal disturbances and sedation. The most serious adverse effect is liver damage, although this is rare. Teratogenic effects, predominantly spina bifida, are reported frequently. The mechanism of action of valproic acid is not certain. The most cited mecha-

nisms are the increase in GABA function and the blockade of voltage-dependent Na^+-channels.

Valproic acid and its salts are the most widely used antiepileptic drugs worldwide (25.6 % of the sales in the USA in 1995, 31.5 % in Western Europe).

2.4. Ethosuximide and Trimethadione

Ethosuximide [*77-67-8*], 3-ethyl-3-methyl-2,5-pyrrolidinedione, $C_7H_{11}NO_2$, M_r 141.17, *mp* 64 – 65 °C; very soluble in water.

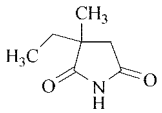

Synthesis: The condensation of butanone with cyano ethyl acetate and subsequent reaction with NaCN and H_2SO_4 gives 2-ethyl-2-methylsuccinic acid (**5**), which is ring-closed via the diammonium salt to yield ethosuximide [39].

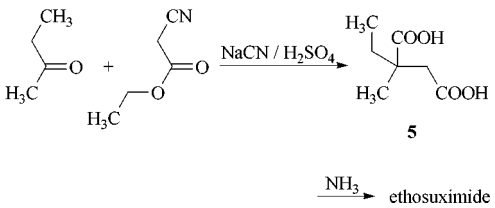

Ethosuximide is the most effective succinimide against absence seizures. It acts by blockade of T-type Ca^{2+}-channels in thalamic neurons.

Ethosuximide is a useful and unique antiepileptic drug, but for the most forms of absences valproic acid is the drug of first choice.

Trade names: Epileo Petitmal (Eisai), Pemal (Benzon), Petnidan (Desitin Arzneimittel), Suxilep (Parke – Davis, Jenapharm), Suxinutin/Zarontin (Parke – Davis).

Phensuximide [*86-34-0*], 1-methyl-3-phenyl-2,5-pyrrolidinedione

Trade name: Mirontin (Parke – Davis).

Methsuximide [*77-41-8*], 1,3-dimethyl-3-phenyl-2,5-pyrrolidinedione

Trade name: Cerontin (Parke – Davis).

Phensuximide and methsuximide have also some efficacy against complex partial seizures.

Trimethadione [*127-48-0*], 3,5,5-trimethyl-1,3-oxazolidine-2,4-dione, $C_6H_9NO_3$, M_r 143.14, *mp* 46 – 46.5 °C; slight camphorlike odor, burning faintly bitter taste; solubility in water about 5 %, freely soluble in chloroform, ethanol, and diethyl ether.

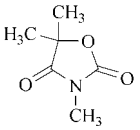

Synthesis: Ethyl 2-hydroxy-2-methylpropionate (**6**), obtained by reaction of acetone and KCN, is condensed with urea in the presence of sodium ethoxide to give 5,5-dimethyl-1,3-oxazolidine-2,4-dione (**7**), which is methylated with dimethyl sulfate to form trimethadione [40].

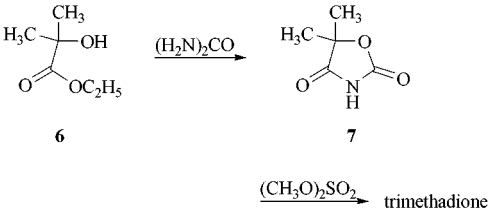

Trimethadione is effective against absence seizures, but its clinical use is very limited because of its toxicity.

Trade names: Absentol (Nourypharma), Epidione (Bellon), Ptimal (EGIS), Tridione/Trimedone (Abbott).

The share of succinimide and oxazolidine drugs in the world market is < 5 %.

2.5. Phenobarbital and Primidone

Phenobarbital [*50-06-6*], phenobarbitone, 5-ethyl-5-phenyl-2,4,6(1*H*,3*H*,5*H*)-pyrimidinetrione, $C_{12}H_{12}N_2O_3$, M_r 232.23; (synthesis and properties of phenobarbital and other barbiturates → Hypnotics, Chap. 5.1.).

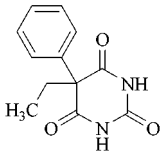

Phenobarbital is highly effective in suppressing generalized tonic – clonic as well as simple and complex partial seizures; however, its clinical use is limited because of its serious side effects: fatigue, drowsiness, irritability, and especially deterioration of higher cerebral functions (attention memory, and intellectual performance). Intravenous use in status epilepticus may lead to severe respiratory depression. Phenobarbital potentiates inhibitory GABAergic and inhibits excitatory glutamatergic neurotransmission through interaction with modulatory binding sites on the respective neurotransmitter receptors.

Trade names: Agrypnal (Eggochemia), Eskabarb (SKF), Lepinal (Arzneimittelwerk Dresden), Luminal (Winthrop, E. Merck, Desitin Arzneimittel), Phenobal (Fujinaga).

Methylphenobarbital [*115-38-8*], 5-ethyl-1-methyl-5-phenyl-2,4,6(1*H*,3*H*,5*H*)-pyrimidinetrione, is regarded as a prodrug for phenobarbital and has an advantage over phenobarbital because it produces plasma phenobarbital levels that vary in direct proportion to drug dose.

Trade name: Mebaral (Winthrop).

Primidone [*125-33-7*], 5-ethyldihydro-5-phenyl-4,6(1*H*,5*H*)-pyrimidinedione, $C_{12}H_{14}N_2O_2$, M_r 218.25, *mp* 281 – 282 °C; slightly bitter taste; sparingly soluble in water (0.6 g/L at 37 °C) and most organic solvents.

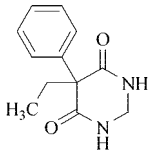

Synthesis: Primidone is obtained by reductive desulfurization of 2-thiophenobarbital with Raney nickel or zinc and formic acid, or by ring closure of ethylphenylmalonamide with formamide or formic acid [41].

Primidone is effective against most generalized and partial seizure types, but its clinical use is limited because of its serious side effects, which are similar to those of its main metabolite, phenobarbital. Although primidone is a prodrug for phenobarbital its antiepileptic properties are independent of phenobarbital.

Trade names: Liskantin (Desitin Arzneimittel), Mylepsin (ICI), Mylepsinum (Zeneca), Resimatil (Sanofi Winthrop), Mysoline (Wyeth – Ayerst), Sertan (Chinoin).

Phenobarbital and primidone are rarely used worldwide (share < 5 % of the sales of antiepileptics in 1995).

2.6. Benzodiazepines (→ Hypnotics, Chap. 5.4., → Sedatives, Chap. 3.1., → Psychopharmacological Agents, Chap. 4.1.)

Clonazepam [*1622-61-3*], 5-(2-chlorophenyl)-1,3-dihydro-7-nitro-2*H*-1,4-benzodiazepin-2-one, $C_{15}H_{10}ClN_3O_3$, M_r 315.72; *synthesis:* [42].

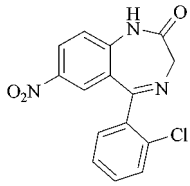

Trade names: Antelepsin (Arzneimittelwerk Dresden), Clonopin/ Rivotril (Roche).

The other 1,4-benzodiazepines, diazepam [*439-14-5*], 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2*H*-1,4-benzodiazepin-2-one, *trade name*: Valium (Roche), (→ Psychopharmacological Agents, Chap. 4.1.1.) and nitrazepam [*146-22-5*], 1,3-dihydro-7-nitro-5-phenyl-2*H*-1,4-benzodiazepin-2-one, *trade name*: Mogadon (Roche), (→ Hypnotics, Chap. 5.4.) as well as the 1,5-benzodiazepine clobazam [*22316-47-8*], 7-chloro-1-methyl-5-phenyl-1*H*-1,5-benzodiazepine-2,4(3*H*,5*H*)-dione, *trade name*: Frisium, (Hoechst), (→ Psychopharmacological Agents, Chap. 4.1.2.) have been in clinical use for a long time.

The major clinical use of the benzodiazepines in epilepsy is in the initial treatment of status epilepticus. They are also effective in a variety of

other seizure types, but tolerance to their effects usually develops early in the course of treatment. For this reason they are used mostly for adjunct therapy in refractory cases.

The antiepileptic activity of benzodiazepines is mediated primarily by an enhancement of GABAergic synaptic transmission.

The benzodiazepines are widely used (sales of clonazepam in USA in 1995 26.3 %, in Western Europe 2.2 %).

2.7. New Drugs

Since the 1980s several new drugs have been approved or are in the process of being approved, e.g., vigabatrin, felbamate, zonisamid, lamotrigine, gabapentin, topiramate, and tiagabine (oxcarbazepine see Section 2.2). At present, the main use of the new agents is in patients that are refractory to first-line drugs such as carbamazepine or valproate. Further studies are required to characterize their activity spectrum as well as their potential value in monotherapy [29], [30].

Vigabatrin [*60643-86-9*], 4-amino-5-hexenoic acid, $C_6H_{11}NO_2$, M_r 129.16, *mp* 171 – 177 °C; off-white crystals; solubility: water 335 mg/mL, ethanol 7.5 mg/mL, chloroform 0.1 mg/mL.

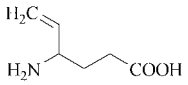

Synthesis: The reaction of 1,4-dichloro-2-butene with diethyl malonate in the presence of sodium ethoxide as catalyst in refluxing ethanol gives 1,1-bis(ethoxycarbonyl)-2-vinylcyclopropane (**8**), which by reaction with gaseous ammonia in DMF is converted into 3-carboxamido-5-vinyl-2-pyrrolidone (**9**). This compound is treated with HCl in refluxing acetic acid to yield vigabatrin [43].

Vigabatrin has a good efficacy profile in partial seizures. However, anticonvulsant efficacy is lost in up to 50 % of patients due to development of tolerance. Side effects are weight gain and effects on mood and behavior. It has been found that vigabatrin potentates the GABAergic transmission through inhibition of GABA-transaminase.

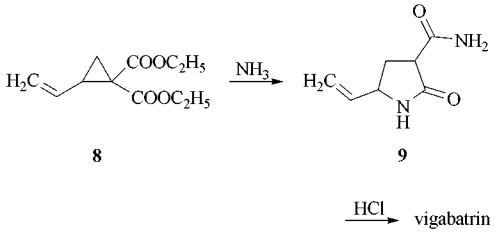

Trade name: Sabril (Merrell Dow).

Lamotrigine [*84057-84-1*], 3,5-diamino-6-(2,3-dichlorophenyl)-1,2,4-triazine, $C_9H_7Cl_2N_2$, M_r 256.09, *mp* 216 – 218 °C; off-white crystals; pk_a 5.5.

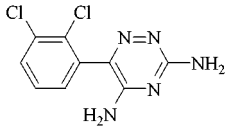

Synthesis: The reaction of 2,3-dichlorobenzoyl chloride (**10**) with $Cu(CN)_2$ and KI in refluxing xylene yields 2,3-dichlorobenzoylcyanide (**11**), which is cyclized with aminoguanidine in DMSO [44].

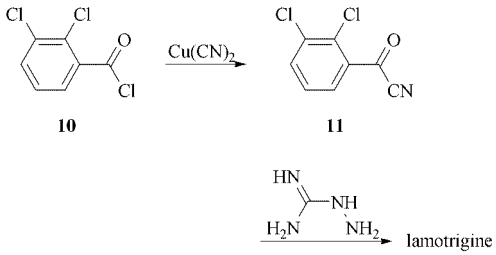

Lamotrigine is considered to be effective in different seizure types and it is well tolerated. The main mechanism of action is the blockade of voltage dependent Na^+-channels. Severe skin rashes can necessitate the stop of the treatment.

Trade name: Lamictal (Wellcome).

Gabapentin [*60142-96-3*], 1-(aminomethyl)cyclohexaneacetic acid, $C_9H_{17}NO_2$, M_r 171.24, *mp* 165 – 167 °C; off-white crystals; zwitterion at physiologic pH; highly water soluble.

Synthesis: The synthesis is started with the Guareschi salt (**12**) (obtained from cy-

clohexanone and cyanoacetate), which is hydrolyzed and decarboxylated to give 1,1-cyclohexanediacetic acid. The anhydride can be treated either with methanol to yield the half ester (**13**) or with hydroxylamine to afford the *N*-hydroximide (**14**). The half ester is converted to the azide and subjected to a Curtius-type rearrangement to give the isocyanate (**15**) which is hydrolyzed to gabapentin. Gabapentin is also obtained from the corresponding *N*-hydroxyimide (**14**) via a Lossen-type rearrangement by conversion of (**14**) to *N*-benzenesulfonyloximide and following reaction with triethylamine to the urethane ester (**16**) [45].

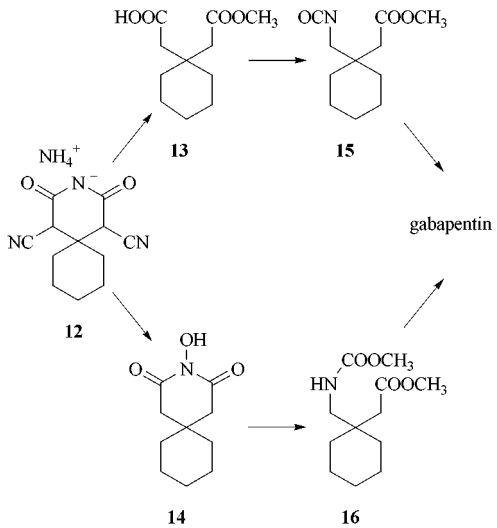

Gabapentin has a good tolerability. Dizziness, fatigue, somnolence, ataxia and tremor are relatively common. Its short half-life in the body is a disadvantage. Gabapentin binds with high affinity to a branched amino acid transporter. This may be the reason for increased GABA-levels in certain region of the brain during therapy with gabapentin.

Trade name: Neurontin (Warner – Lambert).

Topiramate [*97240-79-4*], 2,3:4,5-bis-*O*-(1-methylethylidene)-β-D-fructopyranose sulfamate, $C_{12}H_{21}NO_8S$, M_r 339.36, *mp* 125 – 126 °C; $[\alpha]_D^{24} = -34.0$; soluble in methanol.

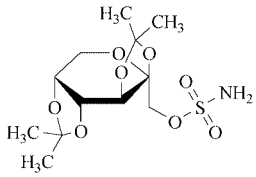

Synthesis: D-Fructose is reacted with acetone to produce the bisacetonide (**17**). This compound is then condensed with sulfamoyl chloride in presence of sodium hydride [46].

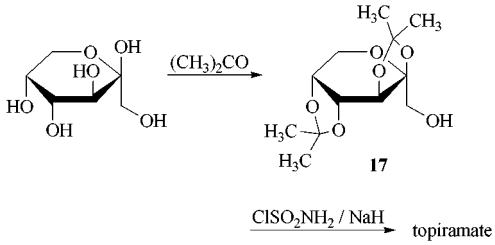

Topiramate has a good efficacy profile in partial seizures and probably a broad spectrum. Its mechanism of action includes blockade of voltage-dependent Na^+-channels, potentiation of GABA response, antagonism of Kainate/AMPA receptor sites, and inhibition of carbonic anhydrase. Clinically significant adverse effects associated with topiramate are fatigue, dizziness, somnolence, impaired concentration, ataxia, and weight loss.

Trade name: Topamax (Johnson & Johnson).

Tiagabine [*115103-54-3*], (*R*)-(−)-1-[4,4-bis(3-methyl-2-thienyl)-3-butenyl]-3-piperidinecarboxylic acid hydrochloride, $C_{20}H_{25}NO_2SO_2 \cdot HCl$, M_r 412.0, *mp* 193 – 195 °C; off-white, nonhygroscopic, crystalline powder; sparingly soluble in water.

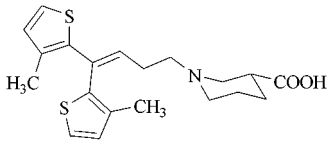

Synthesis: Bromobis(methylthienyl)butene (**18**) is reacted with *R*-(−)-ethyl-3-piperidine carboxylate (**19**) and subsequently hydrolyzed to give the free base which is converted to the hydrochloride [47].

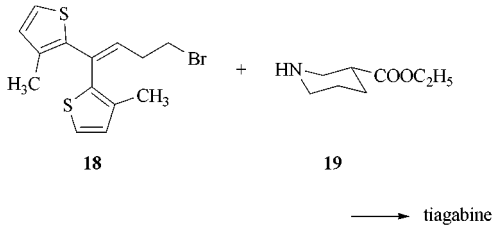

Tiagabine is considered to show a similar activity as vigabatrin or lamotrigine against refractory partial seizures in randomized-controlled add-on trials. Tiagabine exerts its effect by enhancement of GABAergic transmission through inhibition of GABA reuptake. Dizziness, headache, tremor, impaired concentration, and fatigue are common side effects.

Trade name: Gabitril (Abbott).

Felbamate [*25451-15-4*], 2-phenyl-1,3-propanediol dicarbamate, $C_{11}H_{14}N_2O_4$, M_r 238.24, *mp* 150.3 – 151.2 °C; white crystalline powder; solubility in water 0.33 mg/mL, in ethanol 5.0 mg/mL, and in DMF 333.4 mg/mL.

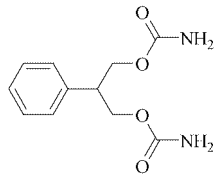

Synthesis: The synthesis of felbamate was first published in 1959. 2-Phenyl-1,3-propanediole is reacted with ethylcarbamate to yield felbamate [48].

Felbamate is effective in Lennox – Gastaut syndrome and it is considered to act by blockade of voltage-dependent Na^+-channels, antagonism of excitatory transmission, through an action on the glycine modulatory site of the NMDA receptor, and potentiation of GABA response. Felbamate, however, has been associated with a high risk of aplastic anemia and hepatotoxicity.

Trade names: Felbatol (Carter – Wallace), Taloxa (Schering – Plough).

Zonisamide [*68291-97-4*], 1,2-benzisoxazole-3-methane-sulfonamide, $C_8H_8N_2O_3S$, M_r 212,23, *mp* 164 – 168 °C; white to pale yellow crystals or crystalline powder; freely soluble in acetone, sparingly soluble in ethanol (95 %), soluble in water 0.8 mg/kg at pH 7.

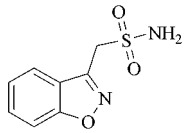

Synthesis: The reaction of 3-bromomethyl-1,2-benzisoxazole with sodium sulfite in methanol/water gives sodium 1,2-benzisoxazole-3-methanesulfonate, which is converted into 1,2-benzisoxazole-3-methanesulfonyl chloride. Treatment with NH_3 gives zonisamide [49].

Clinical trials have shown zonisamide to be a promising drug for treating a wide variety of seizures. Studies on zonisamide's cellular mechanism of action have demonstrated that the drug blocks voltage-sensitive Na^+- and Ca^{2+}-channels. Relatively common side effects are anorexia, ataxia, dizziness, fatigue, somnolence, impaired concentration, and confusion.

Trade names: Aleviatin (Warner – Lambert), Exceglan/Excegram (Dainippon).

2.8. Other Antiepileptic Drugs

Sulfonamides. The sulfonamides acetazolamide and sulthiam have been known as anticonvulsants since 1952 and 1960, respectively. In most patients, the development of tolerance limits their usefulness as antiepileptic drugs. Their anticonvulsant action is due to the inhibition of carbonic anhydrase.

Acetazolamide [*59-66-5*], *N*-[5-(aminosulfonyl)-1,3,4-thiadiazol-2-yl]acetamide, 2-acetylamino-1,3,4-thiadiazole-5-sulfonamide, $C_4H_6N_4O_3S_2$, M_r 222.25, (→ Diuretics, Chap. 2.).

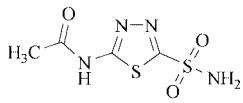

Trade name: Diamox (Lederle).

Sulthiam [*61-56-3*], 4-(tetrahydro-2*H*-1,2-thiazin-2-yl)benzenesulfonamide *S,S*-dioxide, $C_{10}H_{14}N_2O_4S_2$, M_r 290.37; (synthesis [50]).

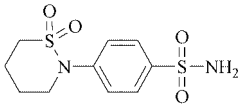

Trade name: Ospolot (Desitin Arzneimittel).

Bromides were first introduced as antiepileptic drugs into therapy in 1857. Side effects and ongoing difficulties in defining a therapeutic range of bromide therapy made their use become obsolete. Now these drugs encounter a kind of comeback because of their good activity against tonic – clonic seizures in refractory patients.

Potassium bromide, KBr, *trade name*: Dibro-Be mono (Dibropharm).

3. References

1. D. Janz, "Antiepileptic Drugs," *Handb. Exp. Pharmacol.* **74** ((1985) 3 – 34.
2. M. J. Eadie, "Antiepileptic Drugs," *Handb. Exp. Pharmacol.* **74** (1985) 765 – 790.
3. J. W. Sander, D. D. Shorvon, *J. Neurol. Neurosurg. Psychiatry* **50** (1987) 829 – 839.
4. J. M. Sutherland, H. Tait, M. J. Eadie (eds.): *The Epilepsies, Modern Diagnosis and Treatment,* 2nd ed., Churchill Livingstone, Edinburgh-London 1974.
5. M. A. Dichter, *Epilepsia* **30** (1989) Suppl.1, 829 – 839.
6. M. A. Dichter, Overview: The Neurobiology of Epilepsy, in J. Engel Jr., T. A. Pedley (eds.): *Epilepsy: A Comprehensive Textbook,* Lippincott-Raven Publishers, Philadelphia New York 1997.
7. Commission on Classification and Terminology of the International League Against Epilepsy, *Epilepsia* **22** (1981) 489.
8. C. Biervert et al., *Science* **279** (1998) 403 – 406.
9. H. Lerche, N. Mitrovic, F. Lehmann-Horn, *Fortschr. Neurol. Psychiatr.* **65** (1997) 481 – 488.
10. D. Schmidt, in M. N. G. Dukes (ed.): *Meyler's Side Effects of Drugs,* 12th ed., Elsevier Science, Amsterdam 1992 122 – 143.
11. H. G. Wieser, *Baillieres-Clin-Neurol.* **5** (1996) 849 – 875.
12. K. D. Laxer, *West. J. Med.* **161** (1994) 309 – 314.
13. H. J. Kupferberg, *Pharmac. Weekbl. Science ed* **143A** (1992) 132 – 138.
14. W. Löscher, D. Schmidt, *Epilepsy Res.* **17** (1994) 95 – 134.
15. W. Löscher, D. Schmidt, *Epilepsy Res.* **2** (1988) 145 – 181.
16. W. Löscher, C. Rundfeldt, *J. Pharmacol. Exp. Ther.* **258** (1991) 483 – 489.
17. M. A. Rogawski in T. W. Stone (ed.): *CNS Neurotransmitters and Neuromodulators: Glutamate,* **vol. 2,** CRC, Boca Raton, FL 1995, pp. 219 – 237.
18. C. Rundfeldt, *Europ. J. Pharmacol.* **336** (1997) 243 – 249.
19. W. J. Brouilette, G. B. Brown, T. M. DeLorey, G. Liang, *J. Pharm. Sci.* **79** (1990) 871 – 874.
20. V. Pestellini et al., *J. Med. Chem.* **23** (1988) 473 – 476.
21. A. Haj-Yehia, M. Bialer, *Pharmaceut. Res.* **6** (1989) 683 – 689.
22. J. Lange, W. Kazmierski, J. Daroszewski, *Pol. J. Pharmacol. Pharm.* **43** (1991) 71 – 77.
23. H. Schäfer, "Antiepileptic Drugs," *Handb. Exp. Pharmacol.* **74** (1985) 199 – 243.
24. J. A. Bikker, J. Kubanek, D. F. Weaver, *Epilepsia* **35** (1994) 411 – 425.
25. G. Klopman, R. Contreras, *Mol. Pharmacol* **27** (1985) 86 – 93.
26. K. Unverferth et al., *J. Med. Chem.* **1998** 63 – 73.
27. R. H. Levy, R. H. Mattson, B. S. Meldrum (eds.): *Antiepileptic Drugs,* 4th ed., Raven Press, New York 1995.
28. M. Dam, L. Gram (eds.): *Comprehensive Epiletology* Raven Press, New York 1991.
29. E. Perucca, *Br. J. Clin. Pharmacol.* **42** (1996) 531 – 543.
30. A. G. Marson, Z. A. Kadir, J. L. Hutton, D. W. Chadwick, *Epilepsia* **38** (1997) 859 – 880. D. W. Chadwick, *Epilepsia* **39** (1998) 799 – 803.
31. R. S. Fischer, J. T. Coyle: "Neurotransmitters and Epilepsy," in I. Bodis-Wollner, E. A. Zimmermann (eds.): *Frontiers of Clinical Neuroscience,* **vol. 11,** Wiley-Liss, New York 1991.
32. SCRIP Reports: "Epilepsy: A Mature Market or Long-term Prospect?" PJB Publications Ltd., Richmond 1997.
33. H. Biltz, K. Seydel, *Ber. Dtsch. Chem. Ges.* **44** (1911) 1 – 413.
34. Parke-Davis, US 2 409 754, 1946 (H. R. Hence).
35. Geigy, US 2948718, 1960 (W. Schindler).
36. Geigy, DE 2 011 087, 1970 (W. Schindler).
37. Geigy, EP 28 028, 1979 (A. Aufderhaar).

38. B. S. Burton, *Am. Chem. J.* **3** (1882)385 – 395. W. Keil, *Hoppe-Seyler's Z. Physiol. Chem.* **282** (1947) 137 – 142.
39. S. Sahay, G. Sircar, *J. Chem. Soc.* 1927, 1252 – 1256.
40. J. S. H. Davis, W. H. Hook, F. Long, *J. Chem. Soc.* 1950, 32 – 41.
41. ICI, GB 666 027, 1952 (W. R. Boon, H. C. Carrington, C. H. Vasey) .
42. L. H. Sternbach, R. I. Fryer, O. Keller, W. Metlesics, G. Sach, N. Steiger, *J. Med. Chem.* **6** (1963) 261 – 265.
43. Merrell, BE 873 766, 1978 (M. W. Gittos, G. J. Leterte).
44. Wellcome, GB 7 919 257, 1979 (M. G. Baxter, A. R. Elphick, A. A. Miller, D. A. Sawyer).
45. Gödecke, DE 2 460 891, 1974 (G. Satzinger, J. Hartenstein, M. Herrmann, W. Heldt). DE 2 611 690, 1977 (J. Hartenstein, G. Satzinger).
46. B. E. Maryanoff et al., *J. Med. Chem.* **30** (1987) 880 – 887.
47. K. E. Anderson et al., *J. Med. Chem.* **36** (1993) 1716 – 1725.
48. B. J. Ludwig, L. S. Powell, F. M. Berger, *J. Med. Chem.* **12** (1969) 462 – 472.
49. Dainippon, US 4 172 896, 1979 (H. Uno, O. Takatsuki). H. Uno, M Kurokawa, Y. Masuda, H. Nishimura, *J. Med. Chem.* **22** (1979) 180 – 183.
50. Bayer, DE 1 111 191, 1959 (B. Helferich, R. Behnisch, Wirth); US 2916489, 1959 (B. Helferich, R. Behnisch).

Antifouling Agents → Biocides

Antifreezes

Sidney F. Bosen, Norchem, Inc., Morris, Illinois 60450, United States

William A. Bowles, Norchem, Inc., Morris, Illinois 60450, United States

Emory A. Ford, Norchem, Inc., Morris, Illinois 60450, United States

Bruce D. Perlson, Norchem, Inc., Morris, Illinois 60450, United States

1. Introduction 1
2. Antifreeze Materials 1
3. Physical Properties 2
4. Corrosion Inhibitors 3
5. Production 4
6. Environmental Protection 4
7. Quality Specifications 5
8. Chemical Analysis 5
9. Storage and Transportation 6
10. Legal Aspects 6
11. Uses 6
12. Trade Names 7
13. Economic Aspects 7
14. Toxicology and Occupational Health . 8
15. References 8

1. Introduction

Antifreeze is a material that is added to water or water-containing liquids to lower their freezing points. The liquid whose freezing point is lowered is usually water. The largest single use of antifreeze is to protect internal combustion engines from the damage that can result to the engine and to the radiator when water freezes. Antifreeze finds additional smaller volume uses in refrigeration, heat transfer, and water heating systems, building air conditioners, solar energy units, automatic sprinklers, hydraulic systems, snow-melting equipment, ice skating rinks, deicing fluids, water-based paints, pharmaceutical products, and freeze-drying apparatus. Antifreeze has been produced in the United States since 1920, and in European countries for over 50 years.

The antifreezes produced for automobiles, trucks, and other vehicles mostly use formulations based on ethanediol, 1,2-ethanediol, ethylene glycol [*107-21-1*]. Applications of ethanediol as antifreeze in internal combustion engines are the major emphasis of this article (→ Ethylene Glycol).

2. Antifreeze Materials [1], [2]

Early in the development of internal combustion engines, *water* was selected as the coolant. The reasons for this selection were its availability, low cost, and good heat-transfer properties compared to other materials. However, water has a number of disadvantages:

1) Under normal operating conditions it may be corrosive to metals found in cooling systems.
2) It has a relatively high freezing point.
3) On freezing water expands its volume by 9 – 10 %. This change in volume is sufficient to rupture engine blocks, water pumps, and radiators.

These disadvantages may be overcome by adding antifreeze and corrosion inhibitors. Because water has a relatively low boiling point (100 °C), pressure caps are used to increase the boiling point of the coolant. However, the maximum pressure is limited by the materials of construction, particularly of the radiator, and by the solder used to seal joints in the radiator. The use of ethanediol as an antifreeze elevates the coolant's boiling point, circumventing the problem of using high pressures to achieve higher coolant temperatures.

During the 1920s, United States motor vehicle registration increased from 7.5×10^6 to 26.5×10^6 with a corresponding increase in the need for antifreeze. During this period, denatured *ethanol* (ethyl alcohol) became an important antifreeze because of its availability, low cost, and relatively good freezing point depressing efficiency in aqueous solutions. Although glycerol (1,2,3-trihydroxypropane) was used as

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_023

a substitute antifreeze, the amount available for this use varied considerably depending on the overall market for it. Calcium chloride and other salt solutions also were used as engine coolants in parts of the United States. However, because of the corrosiveness of the brine solutions, these did not become well-established antifreeze materials. In addition honey and sugar solutions and, occasionally, kerosene and other hydrocarbon oils were used as engine coolants. None of these materials gained widespread use. *Ethanediol* also made its appearance during the 1920s and proved to be an excellent antifreeze; however, initially, its availability limited its use.

Prior to World War II (1940), the principal antifreeze materials available were ethanol, synthetic *methanol,* and ethanediol, accounting for 90 % of the United States antifreeze consumption. The remaining requirements were served by wood byproduct methanol, glycerol, and 2-propanol (isopropyl alcohol). During World War II, ethanol, methanol, and ethanediol remained the dominant antifreeze materials; however, availability rather than cost or performance was the primary criterion for using a particular material.

Following World War II (1945), the increase in the number of vehicles in both the United States and Western Europe heightened the demand for antifreeze materials. During this period, coolants based on methanol and ethanediol were used almost exclusively. Methanol-based coolants reached maximum consumption in the early 1950s. By then car manufacturers had begun to install ethanediol in new car production, and by the early 1970s they were using only this material. Similarly, the entire market for replacement automotive engine coolants shifted to ethanediol. By 1960 ethanediol had increased its market share to > 80 % [3]. Methanol-based coolants, used in some stationary engines and required in older engines, hold a small and decreasing share of the coolant market.

The main factor that has displaced methanol coolants in favor of ethanediol in automotive use is the higher coolant operating temperatures required to achieve superior space heating in vehicle interiors during cold weather. Higher coolant operating temperatures also result in slight improvements in vehicle fuel efficiency.

Antifreezes based on *1,2-propanediol,* propylene glycol (→ Propanediols), are used as heat transfer fluids in systems where toxicity of the antifreeze poses a problem, e.g., in some active solar energy collectors, in cooling systems of engines operating potable water sources, and in freezing and refrigeration systems for food and dairy products. The use of 1,2-propanediol as an automotive antifreeze in some European countries (e.g., Switzerland) is more widespread than in the United States because of its reduced potential for poisoning humans, particularly children, by accidental ingestion.

1-Methoxy-2-propanol also has been used as an antifreeze material, principally in ebullient cooling systems and in some heavy-duty diesel engines. Because an azeotrope is formed between 1-methoxy-2-propanol and water, freeze protection also is provided in the vapor phase in ebullient cooling applications.

1-Methoxy-2-propanol-based antifreeze-shave applications in heavy-duty diesel engines and other industrial engines where coolant seep age into the crankcase is a problem [4]. Because of low volatility ethanediol remains in the crankcase, resulting in sludge and varnish formation and thereby causing accelerated bearing wear. On the other hand, 1-methoxy-2-propanol is more volatile and is more readily expelled along with water vapor through the crankcase vent system.

1-Methoxy-2-propanol also maintains lower metal temperatures than comparative glycol-based coolants [5]. This was an advantage in rotary engines. However, with the nearly complete disappearance of these engines, this advantage is no longer a major market factor.

3. Physical Properties

The physical properties of water and several common antifreeze materials are given in Table 1.

Antifreeze–water mixtures exhibit unusual freezing behavior compared to each pure material or to dilute aqueous solutions. When freezing begins, these mixtures do not form a solid – liquid equilibrium with the same chemical composition in both phases. The higher freezing component, water, starts to crystallize first. This increases the concentration of antifreeze in the liquid, lowering its freezing point. The mixture behaves as a slush over a much

Table 1. Physical properties of water and several antifreeze compounds [6]

	Water, H_2O [*7732-18-5*]	Methanol, CH_3OH [*67-56-1*]	1,2-Ethanediol, $(CH_2OH)_2$ [*107-21-1*]	1,2-Propanediol, $HOCH_2CHOHCH_3$ [*57-55-6*]	1-Methoxy-2-propanol, $CH_3OCH_2CHOHCH_3$ [*107-98-2*]
M_r	18.02	32.04	62.07	76.09	90.1
d_4^{20}	0.998	0.792	1.113	1.036	0.923
bp (0.1013 MPa), °C	100.0	64.7	197.5	187.4	120.1
fp, °C	0	−97.6	−12.6	−43.5	−96
Specific heat (20 °C), J g^{-1} K^{-1}	4.17	2.495	2.357	2.481	2.426
Flash point (open cup), °C ∗	–	12	117	102	38
Viscosity (20 °C), mPa · s	1.0	0.57	20.1	56.2	1.81

∗ DIN 58758

wider temperature range than either pure material.

Whereas the freezing point depression of antifreeze–water mixtures can be easily computed for dilute solutions, this is not the case for the corrosion-inhibited antifreezes that are in general use. Approximate freezing points can be estimated from data generated from mixtures of water and uninhibited antifreeze compounds (Fig. 1). Exact freezing points are best determined by actual measurement (a convenient method is given in ASTM Standard Test Method D 1177 [7]).

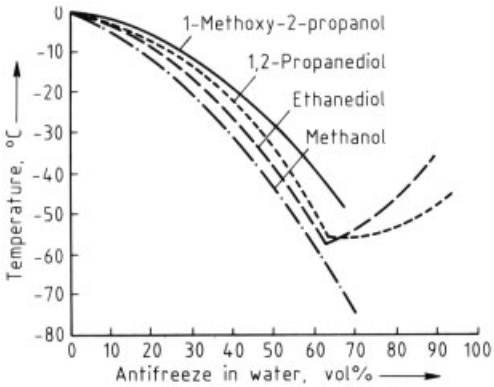

Figure 1. Freezing points of various antifreeze–water solutions [6]

Along with the freezing point, antifreeze materials affect the boiling point of water. Permanent antifreezes raise the boiling point of the mixture above 100 °C. Table 2 gives the boiling points of several uninhibited antifreeze–water mixtures [8]. Methanol, unlike the other anti freezes in the table, actually decreases the boiling point of the solution below that of water. Because corrosion inhibitors also increase the boiling point of water, exact boiling points must be determined on the actual mixture (ASTM Standard Test Method D 1120 [7]).

Other physical properties of antifreeze–water mixtures that impact equipment design, such as viscosity, thermal conductivity, specific heat, and specific gravity, can be obtained from product bulletins or engineering handbooks [9].

4. Corrosion Inhibitors

Corrosion protection of the engine and associated cooling/heating system is another important function of the formulated antifreeze. Because antifreeze – water solutions can be more corrosive than water alone, corrosion inhibitors are added to give efficient corrosion protection and also to hinder the oxidation of ethanediol to corrosive derivatives.

Solution corrosivity is a complex function of antifreeze base, type and combinations of metals used, nature of heat rejection surfaces, and cooling system temperatures. The quality of the water is also important. Anions of strong acids, such as chloride or sulfate, can increase the solution corrosivity.

In the United States coolant inhibitor formulations are made without regard to water quality. This permits the use of a wide variety of inorganic materials that might otherwise not be used, such as orthophosphate.

Formulations with low corrosiveness are produced by combining ethanediol with inhibitors from the following groups: phosphates, benzoates, sebacates, borates, nitrates, nitrites, silicates, or molybdates, or organic compounds, such as amines, benzothiazoles, benzotriazoles,

Table 2. Boiling points, °C, of mixtures of water and various uninhibited antifreezes (at 101.3 kPa) [8]

Antifreeze, %	Methanol	Ethanediol	1,2-Propanediol	1-Methoxy-2-propanol
10	91.7	101	101	100
20	86	102	101.5	100.5
30	82	103	102	101
40	79	104.5	102.5	102
50	76.5	107	104.5	102.5
60	74	110	107	103

or imidazoles (e.g., triethanolamine phosphate, sodium benzoate, sodium nitrate).

The amount of each is based on the formulator's experience as to what is necessary for the desired corrosion protection. General Motors' specification GM 6038-M (see Table 3) contains the typical inhibitors found in most antifreezes produced in the United States.

Table 3. Antifreeze composition specifications, General Motors

	GM 6038-M (parts by wt)	GM 6043-M (wt %)
Ethylene glycol [a]	1115.5	95.5
$NaNO_3$	2.5	0.1
$Na_2B_4O_7 \cdot 5H_2O$	11.5	0.4
$Na_2SiO_3 \cdot 5H_2O$	1.9	–
Liquid sodium silicate [b]	–	0.3
$Na_2MoO_4 \cdot 2H_2O$	–	0.2
$Na_3PO_4 \cdot 12H_2O$	5.0	–
H_3PO_4 (85 wt % solution)	–	0.15
NaOH	2.25	0.235
Sodium 2-mercapto-benzothiazole (50 wt % solution)	6.5	0.5
Sodium tolyltriazole (50 wt % solution)	–	0.2
Antifoam [c]	0.5	0.05
Dye [d]	0.05	0.005
Water	– [e]	2.3
Silicate antigel [f]	–	0.06

[a] 90 wt % minimum monoethylene glycol, 10 wt % maximum other glycols;
[b] Liquid sodium silicate contains a $SiO_2 : Na_2O$ ratio of 3.22 : 1, by weight;
[c] Pluronic L-61 or equivalent;
[d] Uranine C and Alizarine Cyanine Green G-extra or equivalents in a weight ratio of 2 and 3 parts, respectively;
[e] Water may be added to aid the dissolution of salts consistant with the composition maintaining a −37 °C, maximum, freezing point when mixed 50 vol % with distilled water;
[f] Union Carbide silicate stabilizer Y-9794 or equivalent.

In both Europe and the United States, some automotive manufacturers limit the use of materials, such as the combination of amines or nitrites, that may react to form carcinogenic compounds.

In Europe the coolant must be compatible with local hard water. Such materials as orthophosphate and, in some areas silicate, cannot meet this criterion. Benzoate is an alternative to orthophosphate. Amine salts are used only in some special applications where the specifications have not been changed to get improved formulations.

Inhibited antifreezes used in Japan are similar to those used in the United States, except triethanolamine [*tris*(2-hydroxyethyl)amine][*102-71-6*] is frequently used as the neutralizing base.

5. Production

Most antifreezes are formulated products. The major ingredient is the freezing point-depressing fluid. In automotive antifreeze concentrates, ethanediol represents greater than 94 wt % of the formulated product. The remaining ingredients are metal and glycol antioxidants, water, and minor amounts of other additives, e.g. to prevent foaming and to provide a distinctive color.

Antifreeze formulations are manufactured in a batch operation using a tank equipped with an agitator. The tank may also contain heating coils to facilitate the dissolution of additives. In the case of ethanediol-based antifreeze, the blending tank is usually constructed of mild steel or glass-fiber-reinforced resin. Rubber-lined tanks are incompatible with ethanediol.

Manufacturers of the antifreeze base may or may not have their own blending facilities. Many regional facilities offer blending and packaging services on a contract basis. This provides more economical distribution of antifreeze products.

6. Environmental Protection

Antifreeze materials are most likely to enter the aquatic environment after use. Both ethanediol and 1,2-propanediol have low toxicity to animals, plants, and a variety of microorganisms.

However, discharge of excessively high concentrations of glycols in localized areas of the environment should be avoided. Dilute aqueous solutions are rapidly degraded biologically to carbon dioxide and water, whereas concentrated solutions can overload the biologic capacity of the environment to degrade these materials. In summary dilute solutions of glycols present little threat to the environment, and therefore, glycol-based antifreezes represent minor disposal problems.

7. Quality Specifications

Ethanediol-based antifreeze is manufactured to meet either performance or composition specifications. *Performance specifications* define the attributes antifreeze must have, without restricting the composition of the product. Performance specifications normally include:

1) The acceptable maximum freezing points and minimum boiling points of the aqueous solutions
2) The minimum acceptable levels of corrosion performance in vehicle and/or laboratory tests
3) The acceptable effect on elastomers used in vehicle cooling systems
4) The acceptable effect on vehicle finishes

Some performance specifications do place limited restrictions on antifreeze composition, for example, by excluding or limiting the use of some additives, or specifying ranges of physicochemical properties.

Composition specifications prescribe the precise identity and amount of each chemical used to manufacture the antifreeze.

Antifreeze specifications and performance requirements are determined by three major purchasing groups: vehicle manufacturers, consumers, and government agencies. Vehicle manufacturers generally purchase antifreeze to meet their own performance specifications. General Motors has two composition specifications, GM 6038-M and GM 6043-M [10] (see Table 3). The latter specification was issued in 1983 to provide greater corrosion protection of aluminum cooling system parts. The performance specification designations (ca. 1983) of the major automobile manufacturers are:

AMC	AM 4085
BMW	BMW N 60069.0
Chrysler (USA)	MS 7170
Daimler Benz	DBL 7700
Fiat	5552311
Ford (USA)	ESE M 97 B 44-A
Ford (Europe)	SSM-97 B 9101-A
General Motors (USA)	GM 1825-M
General Motors (Europe)	GME L6 368
Honda	Z 009 – 75
Nissan	NES M 5059
Renault	41 – 01 – 001
Toyo Kogyo	TES K 221 – 78
Volkswagen	TL-VW 774
Volvo	STD 1285.05

In the United States, the American Society for Testing and Materials (ASTM) plays a prominent role in determining the quality specifications of antifreeze marketed to consumers. The ASTM Committee D 15 on Engine Coolants has established a Standard Specification for Ethylene Glycol Base Engine Coolant (Standard D-3306). Many states requiring registration of antifreeze products have incorporated this specification into legislation (see Chap. 10) governing the distribution and sale of antifreeze.

Both military and civilian agencies are involved in government purchases of antifreeze products. In the United States, national civilian government agency purchases are made using General Services Administration Federal Specification O-A-548 D, which is a performance specification. Military purchases are made using primarily composition specifications.

8. Chemical Analysis

Analysis of antifreezes provides information about either the bulk properties of the fluid or the individual components of the mixture. Recommended ASTM test methods [7] for the most common *bulk properties* are:

f_P	D 1177
	D 3321
bp	D 1120
Specific gravity	D 1122
pH	D 1287
Reserve alkalinity	D 1121
Foaming dependency	D 1881
Ash content	D 1119
Effect on automotive finishes	D 1882

Freezing point is probably the most important of the bulk properties. When exact freezing points are needed Test Method D 1177 is recommended, but if a freezing point within 1 °C of the ethanediol base fluid is all that is required, Test

Method D 3321 is used because it can be easily followed by persons with minimal training.

Solution pH, reserve alkalinity (buffer capacity), and ash content are gross measures of corrosion inhibitor levels. Quantitation of individual corrosion inhibitor compounds gives a clearer view of the condition of the fluid. Knowledge of the absence or presence of corrosion products is also useful, particularly in maintaining large-volume cooling systems of stationary engines. The analytical methods suggested for measurement of individual components are:

Water content of antifreeze concentrates	Karl Fisher titration [11]
Anion corrosion inhibitors	ion chromatography [12]
Ethylene glycol content of antifreeze concentrates	gas chromatography [13]
Boron, silicon, and metallic corrosion products	direct current plasma spectrometry [14]
Organic, copper, brass, and solder corrosion inhibitors	liquid chromatography [15]

9. Storage and Transportation

Ethanediol-based antifreeze concentrate has a flash point of approximately 120 °C and therefore should be handled as a flammable substance. Aqueous antifreeze solutions do not support combustion. Ethanediol has an ignition temperature of 402 °C, and may have been responsible for a vehicle fire resulting from ignition by a hot exhaust manifold [16].

Bulk quantities of ethanediol-based antifreeze concentrates are generally stored in carbon steel tanks. Storage in resin-coated, stainless steel, or aluminum containers is also satisfactory. The material should not be permitted to come in contact with strong oxidizing agents. Ethanediol is hygroscopic; therefore, long, open exposure to a moist atmosphere should be avoided to assure maintaining product quality specifications.

Smaller quantities of ethanediol-based antifreeze concentrate are stored in sealed carbon steel or high-density polyethylene containers. In these packages product integrity is maintained for an indefinite period.

Antifreeze is distributed in a wide variety of lot sizes ranging from barge lots (ca. 1×10^6 L) to packages as small as 1 quart (0.95 L). Packaging materials for ethanediol-based antifreeze include steel and high-density polyethylene containers. Both packaging materials are relatively impervious to chemical attack by ethanediol-based antifreeze and therefore permit long-term storage in the product distribution chain.

Although regulations exist relating to the packaging of ethanediol-based antifreeze products (see Chap. 10), no specific regulations apply in the United States to the transportation of these products over land or water.

10. Legal Aspects

During the last two decades, a number of United States government regulations have been enacted regarding the packaging, labeling, and registration of antifreezes [17]. Specific state registrations for antifreeze include minimum performance characteristics. Similar to these state registrations are the regulations that exist in Western Europe. These regulations will certainly impede the introduction of new antifreeze materials in the coming years and will probably insure continued use of ethanediol-based antifreezes.

11. Uses

The principal use for antifreeze is the protection of vehicle cooling systems. Other applications are in the cooling systems of stationary engines used in petroleum and natural gas pipelines and in other industrial systems that require freeze and corrosion protection of cooling/heating systems.

The principal reason for using antifreeze is to prevent the possibility of damage to liquid cooled/heating equipment that may intermittently or continuously operate below the freezing point of water. Factors that influence the choice of antifreeze material include:

1) The heat capacity of the fluid
2) The viscosity of the fluid
3) The metallurgy in the system
4) The nonmetallic components in the system
5) The flammability of the fluid
6) The toxicity of the fluid
7) Other special operating considerations

12. Trade Names

In Europe automotive antifreeze sold to consumers (replacement antifreeze) is distributed predominantly through vehicle maintenance centers, together with other maintenance and replacement parts. The products are sold under trade names of the manufacturers or under the specific brand names of trading companies and supermarkets. Some examples are: ARAL-Frostschutz (ARAL), DOW FROST (Dow Chemical), Fricofin (Fuchs Mineralölwerke), Frostschutz 500 (Mobil Oil), Frostschutz UT 184 (BP Chemie), Gefrierschutz-Glykol (Hüls), Genantin (Hoechst), Glycoshell (Shell-Chemie), Glysantin (BASF), and others.

Antifreeze also is offered under trade names of the automobile manufacturer or associated companies, e.g., Antigels (Renault), Frostschutz (Opel), Motorcraft (Ford), Frostschutz G 11 (VW).

In the United States and Canada, consumer antifreezes are marketed predominantly under trade names registered to ethanediol manufacturers. Examples of nationally distributed brands in the United States include: Dowguard (Dow Chemical), Peak (Norchem, Inc.), Prestone II (Union Carbide), Shellzone (Shell Chemical), Startex (Texaco Chemical), and Zerex (DuPont). In addition a large number of companies purchase ethanediol and market trade-named antifreezes on both a regional and a national basis. Vehicle manufacturers also purchase and distribute antifreeze bearing their own trade name.

Several United States glycol manufacturers are involved in marketing ethanediol-based antifreeze specifically for industrial, nonautomotive applications. Included in this group are: Dow Chemical (Ambitrol coolant), Norchem, Inc. (Norkool coolants), Texaco Chemical Co. (Jeffcool coolant), and Union Carbide (Ucartherm coolant). Other glycol-based fluids are sold for highly specialized applications, such as heavy-duty diesel engine coolants, e. g., Dowtherm 209 (Dow Chemical) and Compleat (Fleetguard); and for heat-transfer fluids for food and dairy applications, e.g., Dowfrost (Dow Chemical) and Ucar Food Freeze (Union Carbide).

13. Economic Aspects

Antifreeze represents a major end-use market for ethanediol. In the United States approximately 40 % of all ethanediol production is used to manufacture antifreeze [18].

Producers of ethanediol in the United States are (ca. 1983 capacities, 10^3 t/a) [19].

BASF Wyandotte	159
Celanese	227
Dow Chemical (USA)	322
Eastman Chemical	85
ICI American	181
Norchem, Inc.	91
Olin	18
PPG Industries	91
Shell Chemical	181
Texaco Chemical	227
Union Carbide	1316
Total	2898

Ethanediol production capacities for other countries are (ca. 1983, 10^3 t/a):

Canada	327 [19]	
Mexico	191	
Belgium	260	290 [20]
France	215	165
Federal Republic of Germany	325	315
Italy	106	60
Netherlands	304	334
Spain	68	69
Sweden	20	5
United Kingdom	230	255
Japan	733	
Total	2779	

In 1981, United States plants were estimated [21] to be operating at 68 % of nameplate capacity. New plant capacities have been announced in the Middle East, Canada, and several developing countries for startup in the mid-1980s.

The Chemical Specialties Manufacturer's Association annually tabulates the volume of antifreeze shipments from all United States manufacturers (Fig. 2) [22]. The leveling off of total ethanediol-based antifreeze shipments (solid line) is ascribed to decreasing new car production, longer new car maintenance intervals, and decreasing capacity of automobile cooling systems. The cooling system capacity is expected to continue decreasing as the automobile industry strives to improve fuel efficiency by reducing vehicle mass and engine size. In addition antifreeze consumption is strongly influenced by consumer

spending patterns. In recessionary periods consumers tend to defer purchases of nondurable, discretionary goods such as antifreeze.

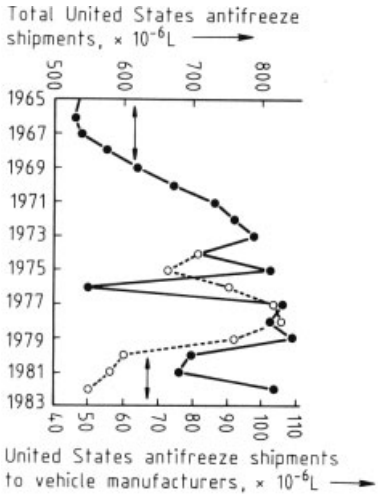

Figure 2. United States ethanediol-based antifreeze consumption

The manufacturing cost of ethanediol-based antifreeze is related directly to the cost of ethylene, presently the only commercially significant raw material for ethanediol. Table 4 shows the historic relationship between the market prices of ethanediol (industrial grade) and ethylene [21].

Table 4. Comparative prices of gaseous ethylene and ethanediol [21]

Year	Gaseous ethylene, average market price, U.S. Gulf Coast, ¢/kg	Industrial grade ethanediol, list price, delivered, U.S., ¢/kg
1973	7.3	19.8
1974	16.5	30.8
1976	26.0	51.8
1979	34.4	61.7
1981	58.6	79.3
1982	42.8	60.6

Because ethanediol-based antifreeze is sold primarily as retail merchandise, the market pricing structure is highly fragmented, particularly in the United States domestic market. Pricing is influenced heavily by advertising, other merchandising techniques, and channels of distribution.

14. Toxicology and Occupational Health

The toxicity of antifreeze materials is an important consideration because of their widespread use [23]. The antifreeze base is normally the principal determinant of the toxicity; however, such antifreeze additives as anticorrosives and antioxidants also can contribute to their overall toxicity. For example, arsenic containing compounds previously were used as anticorrosives.

Handling *ethanediol-based antifreeze* or aqueous antifreeze solutions presents no unusual hazard. *External exposure* to ethanediol presents little hazard; neither skin nor eyes are affected adversely by occasional exposure. Prolonged skin exposure causes a slight macerating action, and the sensitivity of humans to this action is highly variable. Rubber gloves are advised for those handling antifreeze or antifreeze solutions frequently. Safety glasses, face shields, protective clothing, and general-purpose respiration equipment are also recommended.

The most serious hazard of ethanediol-based antifreeze is *oral ingestion.* The acute oral toxicity is relatively low for a variety of small animal species. For humans the lethal dosage by ingestion is approximately 100 mL in adults. The lethal dosage decreases with decreasing body weight.

Inhalation of mists or vapors of hot ethanediol is hazardous [23]. However, at ambient temperatures the vapor pressure of ethanediol is sufficiently low that inhalation risks are minimal.

The toxicity of *1,2-propanediol* to large animals and humans is significantly lower than that of ethanediol. The lethal dosage for adult humans is estimated to be in excess of 1 L. The effects of skin exposure to 1,2-propanediol are minimal, with the only concern being dehydration. 1,2-Propanediol is not an irritant to eyes nor does it cause tissue damage. Inhalation of vapors of 1,2-propanediol appears to cause no toxic effects.

15. References

1. Frank Howard et al., *Automotive Antifreezes,* National Bureau of Standards Circular 576, U.S. Department of Commerce, U.S.

Government Printing Office, Washington, D.C., July 1956.
2. *Maintenance of Automotive Engine Cooling Systems,* Society of Automotive Engineers (SAE) Booklet TR-40, Warrendale, Pa.
3. Summary of Antifreeze Sales for 1959 and 1960, *Bulletin No. 173-61,* Chemical Specialties Manufacturer's Assoc., May 16, 1961.
4. *Dowtherm 209 Coolant Bulletin,* Dow Chemical, Midland, Mich., 1971.
5. G. A. Paul: "The Effect of Selected Coolants on Metal Temperatures in a Rotary Engine," Paper 741091, *SAE Automotive Engineering Meeting,* Toronto, Canada, Oct. 21 – 25, 1974.
6. *Ullmann, 4th ed.* **12** 205.
7. *ASTM Standards on Engine Coolants,* ASTM, Philadelphia, Pa., 1980.
8. J. Timmermans: *Physico-Chemical Constants of Binary Systems.* Interscience, New York 1960.
9. R. H. Perry, H. Chilton (ed.): *Chemical Engineers' Handbook,* 5th ed., McGraw Hill, New York 1973. J. A. Dean (ed.): *Lange's Handbook of Chemistry,* 12th ed., McGraw Hill, New York 1979. R. C. Weast, M. J. Astle (ed.): *Handbook of Chemistry and Physics,* 61st ed., CRC Press, Boca Raton, Fla. 1980.
10. *General Motor Engineering Standards,* "Automotive Engine Coolant Concentrate – Ethylene Glycol Type," GM 6038-M (March 1976); GM 6043-M (February 1983).
11. ASTM, Standard Test Method D-1123, part 30.
12. H. Small, T. S. Stevens, W. C. Baumanm, *Anal. Chem.* **43** (1975) 1801.
13. ASTM Standard Test Method E-202, part 30.
14. G. W. Johnson, H. E. Taylor, R. K. Shogerboe, *Spectrochim. Acta. Part B* **34** (1979) 197.
15. T. M. Schmitt, E. S. Muzher: *Talanta* **28** (1981) 777.
16. *Personal Communication,* W. E. Tudor, Pontiac Motor Division, General Motors, 1983.
17. E. Richard Humburg, *Chem. Times & Trends* **4** (Jan. 1981), no. 1, 35.
18. "Chemical Profiles," *Chem. Mark. Rep.* **214** (July 31, 1978).
19. S. A. Cogswell: *Chemical Economics Handbook,* SRI International, Menlo Park, Calif., 1983, Section 652.5030, pp. A-S.
20. "Petrochemicals," *Eur. Chem. News* **83** (Oct. 19, 1985).
21. *Personal Communication,* Northern Petrochemical Co., Marketing Department, 1983.
22. *Ethylene Glycol Based Antifreeze Sales Survey,* Chemical Specialities Manufacturers Association, Washington, D.C., 1983.
23. G. D. Clayton, F. E. Clayton (ed.): *Patty's Industrial Hygiene and Toxicology, vol 2,* Wiley-Interscience, New York 1962, pp. 3817 – 3832, 3852 – 3861, 4528 – 4541.

Antihistamines → Antiulcer Drugs
Antihistamines → Antiallergic Agents

Anti-inflammatory – Antirheumatic Drugs

Burkhard Hinz, Department of Experimental and Clinical Pharmacology and Toxicology, Friedrich Alexander University Erlangen-Nürnberg, Erlangen, Germany

Conrad P. Dorn, Jr., Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065, United States

Tsung Ying Shen, Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065, United States

Kay Brune, Department of Experimental and Clinical Pharmacology and Toxicology, Friedrich Alexander University Erlangen-Nürnberg, Erlangen, Germany

1. Introduction 1
2. Glucocorticoids 3
3. Nonsteroidal Anti-inflammatory Drugs and Selective COX-2 Inhibitors 4
3.1. Salicylates 5
3.2. Arylacetic Acids 6
3.3. Arylacetic Acid Prodrugs 10
3.4. Arylpropionic Acids 11
3.5. Acidic Enolic Compounds 15
3.6. Anthranilates 17
3.7. Nonacidic Drugs 18
3.8. Selective COX-2 Inhibitors 18
4. Antirheumatic Agents 19
4.1. Gold 19
4.2. Antimalarials 21
4.3. Sulfhydryl Compounds 21
4.4. Immunoregulants 22
5. Inhibitors of Tumor Necrosis Factor . 23
6. Natural Products 23
7. References 24

1. Introduction

The *inflammatory response to injury* is a normal host defense mechanism that serves to isolate and remove the damage. Early events in this process include release of mediators, such as histamine and serotonin that dilate the blood vessels and increase their permeability. Subsequently, fluid leaking into the surrounding tissue space causes inflammation (i.e., swelling, pain, redness, and heat). Included in this fluid exudate are proteins, such as fibrinogen, which, when converted to fibrin, helps to seal off the affected area. Also present are polymorphonuclear neutrophils (PMNs) and leukocytes that phagocytize the infectious or toxic agent.

The inflammatory process is triggered by several interrelated cascade systems in the body. These include:

1) The complement system
2) The plasmakinin system
3) The clotting system
4) The arachidonic acid cascade

The *complement system* represents a cascade of approximately 25 serum proteins. The complement system mediates lysis of antibody-coated targets (bacteria, viruses, cells), recruits inflammatory cells to the sites of inflammation, and increases the efficiency of phagocytosis through opsonization.

Activation and interaction of the *clotting* and *plasmakinin systems* of the body also contribute to the inflammatory process (→ Blood, Chap. 2.2.). Activation of the Hageman factor (clotting factor XII), in addition to its effects on the coagulation system of the host, in turn activates the circulating protein, prekallikrein, to its enzymatically active form, kallikrein. The subsequent cleavage of another plasma substrate, kinonogen, by kallikrein liberates bradykinin, a linear nonapeptide. This peptide is a potent vasodilator that also has the ability to increase blood vessel permeability. In addition, bradykinin acts as a mediator of pain (elicits a pain response).

The *arachidonic acid cascade* has been the most actively studied of all the physiologic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_033

components contributing to inflammation. Both cyclooxygenase and lipoxygenase pathway of arachidonic acid metabolism have been investigated for their effects on inflammation. Arachidonic acid is generated from cellular membrane phospholipids by the action of the enzyme phospholipase A_2. In the cyclooxygenase pathway, it is converted first to a short-lived endoperoxide and then to prostaglandins (→ Prostaglandins). In mammalian cells two cyclooxygenase (COX) enzymes exist which are encoded by different genes, but share a 60 % identity in amino acid sequence. COX-1 is constitutively expressed as a "housekeeping" enzyme in most tissues and mediates physiological responses such as regulation of renal and vascular homeostasis and cytoprotection of the stomach. In comparison, COX-2 is primarily considered as an inducible immediate-early gene product whose synthesis can be up-regulated by various proinflammatory agents, including endotoxin, cytokines, and mitogenes. COX-2 is the major isoform expressed by inflammatory cells and has, accordingly, been shown to release the high levels of prostanoids present under pathological conditions, such as acute and chronic inflammation [1]. The role of prostaglandins in inflammation is profound. Those of the E type increase blood vessel permeability and have the ability to sensitize various tissues (e.g., blood vessels, pain receptors) to the effects of other mediators, such as bradykinin.

The alternate pathway of arachidonic acid metabolism, the 5-lipoxygenase pathway, leads to the generation first of a hydroperoxide derivative, 5 HPETE, that stimulates histamine release from basophiles. Subsequent conversions lead to compounds known as leukotrienes that synergize with prostaglandins to cause increased blood vessel permeability and pain. One such leukotriene, LTB_4, is a potent chemotactic agent and a hyperalgesic agent (pain sensitizer) that shows additive effects with prostaglandins in pain mediation.

Much of the tissue destruction occurring during the inflammatory cycle can be attributed to substances released from the activated phagocytes that have been recruited to the site. These substances include free radicals and radical precursors, such as hydroxyl radicals and superoxide anion. Free radicals are nonspecific in their action and can cause destruction of membrane components, degradation of connective tissue, and depolymerization of hyaluronic acid leading to collagen damage. They may potentiate the action of proteolytic enzymes by oxidative destruction of naturally occurring inhibitors, such as α-1-proteinase inhibitor. In addition they feed the inflammatory cycle by further stimulation of leukocyte and macrophage functions.

Proteolytic enzymes, such as elastase and collagenase, are elaborated by PMNs at the inflammatory site either by active secretion or by death of the cell. Elastase is a neutral protease capable of degrading elastin and other connective tissue components, such as collagen, proteoglycan, and basement membrane. It may also interact with the complement and plasmakinin systems to liberate peptide mediators of inflammation, such as kinins, from soluble precursors. Collagenase catalyzes a specific cleavage of collagen and its precursor tropocollagen.

Furthermore, several *cytokines* (→ Blood, Chap. 2.1.3.) appear to play an essential role in the inflammatory process. Cytokines mediate effects induced by infectious agents, regulate lymphocyte growth, stimulate growth of precursor cell populations, and mediate inflammatory reactions. Interleukin (IL)-1 and tumor necrosis factor (TNF) possess very similar physiological actions in that both cytokines co-stimulate the process of antigen-induced activation of T cells, cause fever, and act on hepatocytes to induce the synthesis of plasma proteins, termed acute phase reactants. Under conditions of rheumatoid disease, IL-1 and TNF induce bone resorption by osteoclasts and increase proliferation of fibroblasts and synovial cells. Accordingly, increased levels of IL-1 and TNF have been detected in joint fluids from patients with rheumatoid arthritis. IL-6, which is produced by macrophages, monocytes, endothelial cells, and fibroblasts in response to IL-1 and TNF, stimulates the production of fibrinogen by hepatocytes and acts as a B cell growth factor. Other cytokines and growth factors contributing to the manifestation of the inflammatory response are IL-2, IL-8, and granulocyte-macrophage colony-stimulating factor (GM-CSF). Interestingly, some cytokines and growth factors possess anti-inflammatory activity (e.g., IL-10 and transforming factor (TGF)-β_1).

Despite the complexity and apparent propensity for endless self-perpetuation, most bouts of acute inflammation are self-limiting and result in

eventual resolution and healing. In some cases, however, persistent stimulation or immunologic derangement occurs, leading to a chronic and degenerative inflammation. Anti-inflammatory drugs are commonly used for these situations.

The most prevalent clinical inflammatory disorders are a variety of arthritic conditions, such as rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, gout, and systemic lupus erythematosus, which together affect about 7 % of the population worldwide.

Rheumatoid arthritis is a chronic inflammatory disease occurring in about 2 % of the population, with more women than men being affected. Most cases are relatively mild, and spontaneous remissions of the disease often occur. A minority of patients, however, develop severe and crippling deformities. Rheumatoid arthritis often affects peripheral joints, such as the knee, the wrist, and the fingers' proximal joints. The ankle, elbow, and hip may become involved but much less frequently. In rheumatoid arthritis, the membrane that lines the joint becomes inflamed and swollen, eventually covering and invading the cartilagenous surfaces of the joint. Progressive destruction of articular and periarticular structures follows, leading to loss of joint function.

The etiology of rheumatoid arthritis is largely unknown. One of the main characteristics of rheumatoid arthritis is an accumulation of T cells, plasma cells, macrophages, and fibroblast-like cells in inflamed joints. T lymphocytes specific for an unknown autoantigen have been suggested to play a crucial role in the induction and maintenance of synovial inflammation. However, the T cell centered hypothesis has been challenged by recent data showing that T cell targeted therapies were rather disappointing, particularly the use of anti-CD4 monoclonal antibodies.

Osteoarthritis, or degenerative joint disease, probably occurs as the result of chronic trauma or biomechanical stress. This disease also shows a predilection for certain joints, namely, the hip, the knee, and the fingers' distal joints. The elbow, shoulder, and ankle are rarely affected. Stress damage or simple "wear and tear" of the joint leads to chondrocyte (type of cartilage cell) damage and to the release of proteolytic enzymes that destroy the articular cartilage. The body's attempt to repair this damage causes formation of cartilagenous lumps that harden and become the osteophytes (bone knobs) characteristic of this disease.

Ankylosing spondylitis is a chronic and progressive inflammation that usually begins at the entheses (sites where muscle and tendon are joined to bone) of the sacroiliac joint and slowly spreads up the spinal column. Bony outgrowths gradually fuse one vertebral joint to the next, leading to restricted movement and eventual immobility of the spine. This disease occurs primarily in men, particularly those whose genetic make-up results in the presence of the HL-A W27 blood group antigen. Such individuals are approximately 600 times more likely to contract ankylosing spondylitis than those not carrying this antigen.

Gout is an arthritis of peripheral joints resulting from the deposition of sodium urate crystals in and about the joints and tendons. Ingestion of these crystals by PMNs results in the release of lysosomal enzymes that induce inflammation. Urate crystals can also activate the Hageman factor and the complement cascade. This condition is initially an acute problem, but repeated attacks lead to chronicity and deformation of the affected joints.

Systemic lupus erythematosus is an autoimmune disease of unknown origin resulting in a generalized inflammation of connective tissue throughout the body. It affects the joints, lungs, skin, heart, and kidneys and is primarily a disease of young women.

The anti-inflammatory drugs commonly used to treat the above conditions are generally divided into three classes:

1) Corticosteroids (→ Hormones)
2) Nonsteroidal anti-inflammatory – analgesic drugs (NSAIDs) and Selective COX-2 Inhibitors
3) Disease-modifying antirheumatic agents

2. Glucocorticoids

The glucocorticoids are oxygenated pregnenes bearing an α-ketol side chain at position 17. They include the natural adrenal cortical hormones, cortisol and cortisone, as well as a host of synthetic analogs. The potent and dramatic

anti-inflammatory effect of cortisone was first demonstrated in rheumatoid arthritic patients in 1948 and thus began a new era of antiarthritic therapy. In addition corticosteroids also are used to treat dermatitis and often preserve sight when used for ocular inflammation.

Molecular-biological studies performed in the 1990s have provided new insights into the mode of action of glucocorticoids in controlling inflammation. Glucocorticoids bind to a cytoplasmic receptor within the target cells thus inducing a rapid nuclear translocation of the activated glucocorticoid receptor-steroid complex into the nucleus. After binding to DNA at consensus sites (i.e., glucocorticoid response elements) in the promoter region of glucocorticoid-responsive genes glucocorticoids can either increase or decrease transcription of numerous genes involved in inflammatory reaction. Furthermore, glucocorticoid receptors directly interact with transcription factors activated under chronic inflammatory conditions. Accordingly, glucocorticoid receptor complexes inhibit the activated form of nuclear factor κB (NF-κB) via a direct protein-protein interaction thereby repressing gene transcription. As NF-κB induces many inflammatory genes encoding cytokines, chemokines, adhesion molcules, and the inflammatory enzymes inducible nitric oxide synthase (iNOS) and COX-2, the inactivation of activated NF-κB in the nucleus represents an important effect of glucocorticoids in repressing gene transcription of various proinflammatory mediators [2], [3].

The use of glucocorticoids is accompanied by serious *adverse effects.* Changes are seen in the metabolism of carbohydrate (increased gluconeogenesis and hyperglycemia), of protein (increased catabolism and negative nitrogen balance), and of fat (increased synthesis and deposition). Disturbances in mineral balance include sodium retention, potassium depletion, and loss of bone calcium. Effects on the CNS range from euphoria to mania. Finally, suppression of endocrine function, particularly of the pituitary and hypothalamus, occurs.

Glucocorticoids used for systemic or local anti-inflammatory activity include hydrocortisone, cortisone acetate, prednisolone, dexamethasone, betamethasone, and triamcinolone. Derivatization provides injectable compounds having a rapid onset of action. For this purpose, esters of the 21-hydroxy group with phosphoric or succinic acid are used as their sodium salts. Depot forms of glucocorticoid drugs can be obtained by preparing slowly hydrolyzable esters of the 21-hydroxy group with hindered acids, such as pivalic or *tert*-butylacetic acid.

Glucocorticoids used extensively for topical administration include hydrocortisone acetate, betamethasone acetate, and triamcinolone acetonide.

For structure, properties, and preparation → Hormones.

3. Nonsteroidal Anti-inflammatory Drugs and Selective COX-2 Inhibitors

Nonsteroidal anti-inflammatory drugs (NSAIDs) inhibit the activity of COX which catalyzes the synthesis of prostanoids by converting arachidonic acid and molecular oxygen into prostaglandin H_2, the common substrate for specific prostaglandin synthases [4]. Most of the NSAIDs inhibit both COX-1 and COX-2 at therapeutic doses, although they vary in their relative potencies against the two isozymes [5]. Whereas the mechanism-based side effects of NSAIDs (e.g., gastrointestinal toxicity, platelet dysfunctions) are due to a suppression of COX-1-derived prostanoids, compelling evidence suggests that inhibition of prostanoids produced by COX-2 can be ascribed to the anti-inflammatory, analgesic, and antipyretic effects of NSAIDs. Accordingly, treatment with NSAIDs with a higher affinity to the COX-1 isozyme (e.g., piroxicam, sulindac) results in increased gastrointestinal adverse reactions. Specific COX-2 inhibitors are expected to exert anti-inflammatory and analgesic effects without causing gastric ulcerogenic effects or platelet dysfunction. *Per definitionem*, a substance may be regarded as a specific COX-2 inhibitor if it causes no clinically meaningful COX-1 inhibition at maximal therapeutic doses. Such compounds usually reveal a more than 100-fold difference in the concentration that inhibits recombinant COX-2 versus COX-1 in respective biochemical in vitro assays [6]. X-ray crystallographic studies of the three-dimensional structures of COX-1 and COX-2 have provided more insight into how COX-2

specificity is achieved. Within the hydrophobic channel of the COX enzyme, a single amino acid difference in position 523 (isoleucin in COX-1, valin in COX-2) has been detected, that is critical for the COX-2 selectivity of several drugs. Accordingly, the smaller valin molecule in COX-2 gives access to a "side pocket", which has been proposed to be the binding site of COX-2 selective substances. Consequently, the total NSAID-binding site is about 25 % larger in COX-2 than in COX-1 [7]. Thus, the increased NSAID-binding pocket of the COX-2 isozyme can bind bulky inhibitors more readily than the COX-1 isoform. Celecoxib (SC58635) and rofecoxib (MK-966) are novel selective COX-2 inhibitors of the diarylheterocyclic family (see Section 3.8). The 4-methylsulfonylphenyl and 4-sulfonamoylphenyl groups of these compounds have been shown to interact with specific residues within the "side pocket" of the COX-2 isozyme. Both compounds have been shown to be effective analgesics in dental pain models, and effective anti-inflammatory and analgesic substances in patients with rheumatoid arthritis and osteoarthritis [8]. Celecoxib and rofecoxib have been associated with fewer adverse gastrointestinal effects than traditional NSAIDs. The first selective COX-2 inhibitor, celecoxib (SC58635) was approved in December 1998 by the U.S. Food and Drug Administration for rheumatoid arthritis and osteoarthritis. Rofecoxib became available in 1999 for the indications osteoarthritis and pain. COX-2 is also expressed under basal conditions in organs such as the ovary, uterus, brain, spinal cord, kidney, cartilage, and bone, suggesting that this isozyme may play a more complex physiological role than formerly expected. Accordingly, COX-2 may be implicated in physiological processes such as ovulation [9] and delivery [10], [11]. Moreover, COX-2 is induced in tissue on the edges of ulcers, and in animals studies selective COX-2 inhibitors have been shown to retard ulcer healing [12]. Thus, in patients with NSAID-associated ulcers, it will be obligatory to show whether effective ulcer healing occurs in those patients that switched to treatment with selective COX-2 inhibitors. Controlled clinical trials will gain insights into other possible side effects of selective COX-2 inhibitors in humans. On the other hand, the involvement of the COX-2 isozyme in pathological states suggests that selective COX-2 inhibitors may have further indications in conditions such as colonic polyposis, colorectal cancer, and Alzheimer's disease.

3.1. Salicylates

The therapeutic use of salicylates, in the form of extracts of willow bark, for an antipyretic effect dates back to early civilization. Elucidation of the chemical structure of the active ingredient, salicin, led to the use of sodium salicylate for the treatment of both rheumatic fever and gout in the 1870s. This was followed by the introduction of the less irritating *O*-acetyl derivative of salicylic acid, aspirin, in 1899. Acetylsalicylic acid served for many years as the cornerstone in the therapy of rheumatoid disease, being the safest, least expensive yet effective agent available. This efficacy is now attributed to the ability of aspirin to irreversibly inhibit arachidonic acid cyclooxygenase by transacetylation of this enzyme. However, the theory that suppression of prostaglandin biosynthesis accounts for the pharmacological actions and the side effects of NSAIDs has been questioned by comparing the actions of salicylate and aspirin [13]. Salicylate, being used as an NSAID for more than one century now, does not, unlike aspirin, inhibit COX-1 and COX-2 activity in vitro [5], [14]. However, despite lacking inhibitory activity against purified COX-1 or COX-2, salicylate exerts comparable anti-inflammatory properties as aspirin [15], and has considerable inhibitory effects on prostaglandin synthesis in intact cells [5], [16], [17] as well as in vivo [18] and ex vivo [19] at sites of inflammation. Several suggestions have been made to describe how salicylates exert their anti-inflammatory and side effects. One proposal is that salicylates act by inhibition of the transcription factor NF-κB [20]. However, relatively high concentrations of sodium salicylate (i.e., higher than that obtained after therapeutic doses) were required to provide significant effects. In 1997 is was demonstrated that sodium salicylate is an effective inhibitor of COX-2 activity at concentrations far below those required to inhibit NF-κB activation and it has been suggested that aspirin and sodium salicylate exerts their pharmacological effects via a weak competitive inhibition with arachidonic acid in the

active site of the enzyme, rather than via inhibition of NF-κB activation. Furthermore, in IL-1β-stimulated endothelial cells sodium salicylate was shown to inhibit COX-2 expression at pharmacological concentrations [16]. However, these results could not be reproduced in other studies [17], [21], [22], [23] despite the fact that much higher salicylate concentrations were investigated.

Years of chemical effort to modify the salicylic acid molecule in order to find a safer and more effective derivative have yielded only a few therapeutically useful salicylates. Newer salicylic acid derivatives such as trilisate, benorylate, and chlorthenoxazine are available for use primarily as acute analgesic – antipyretic agents. Diflunisal with a biphenyl structure is a salicylate analog possessing greater potency, better tolerance, and longer duration of action in anti-inflammatory and analgesic studies.

For the structure and syntheses of acetylsalicylic acid, benorylate, and diflunisal, → Analgesics and Antipyretics.

Acetylsalicylic Acid [*50-78-2*], 2-(acetyloxy)benzoic acid, aspirin, $C_9H_8O_4$, M_r 180.15, *mp* 135 °C.

Trade names: Acetylin (Bristol Myers Squibb, Germany), Acimetten (Pharmonta, Austria), Adprin (Pfeiffer, USA), Alka Seltzer (Bayer, Austria, Switzerland, Czech Republic), Angettes (Bristol-Myers, UK), Asaped (Sanofi-Winthrop, USA), Aspirin (Bayer, Germany), Aspro (Roche Nicholas, Germany), Colfarit (Bayer Pharma Deutschland, Germany), Zorprin (Knoll, USA).

Benorylate [*5003-48-5*], 4-acetamidophenyl 2-acetoxybenzoate, $C_{17}H_{15}NO_5$, M_r 313.32, *mp* 175 – 176 °C.

Trade names: Benoral (Sterling-Winthrop, UK), Benortan (Winthrop, The Netherlands), Duvium (Inpharzam, Belgium, Switzerland), Longalgic (Evans Medical, France), Salipran (Evans Medical, France), Spierifex (Winthrop, The Netherlands).

Benorylate is the acetaminophen ester of acetylsalicylic acid, designed to provide a less irritating form of aspirin. Gastrointestinal blood loss in patients taking benorylate is less than that during therapy with soluble aspirin. Benorylate has been found to be as effective as aspirin for the treatment of chronic rheumatic disorders.

Diflunisal [*22494-42-4*], 2′,4′-difluoro-4-hydroxy-(1,1′-biphenyl)-3-carboxylicacid, $C_{13}H_8F_2O_3$, M_r 250.20, *mp* 210 – 211 °C.

Trade names: Biartac (Merck Sharp & Dohme, Belgium), Diflonid (Dumex, Denmark, Sweden), Diflusal (Merck Sharp & Dohme, Belgium), Dolobid (Frosst, Canada; Logos, South Africa; Merck, USA; Morson, UK), Dolobis (M.S.D.-Chibret, France), Fluniget (Merck Sharp & Dohme, Austria).

Diflunisal is a salicylate with greater anti-inflammatory and analgesic potency and with a long plasma half-life that allows for twice-a-day dosage. Without an *O*-acetyl group it is a reversible inhibitor of cyclooxygenase. It is comparable in efficacy to aspirin in the treatment of osteoarthritis and rheumatoid arthritis but is much better tolerated.

3.2. Arylacetic Acids

Indomethacin [*53-86-1*], 1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indole-3-acetic acid, $C_{19}H_{16}ClNO_4$, M_r 357.81, *mp* 155 °C or 162 °C depending on crystal form.

Synthesis: acylation of sodium 2-(4-methoxyphenyl)hydrazine-1-sulfonate with 4-chlorobenzoyl chloride followed by heating yields 1-(4- chlorobenzoyl)-1-(4-methoxyphenyl)hydrazine. Condensation with levulinic acid in a Fischer indole synthesis affords indomethacin [24].

The introduction of indomethacin in 1963 represented a milestone in the therapy of rheumatic diseases. For 20 years this drug has served as the clinical standard against which new NSAIDs are evaluated. It has also been a valuable biochemical – pharmacological tool in the study of the inflammatory process. It is a potent inhibitor of prostaglandin biosynthesis.

Clinically it is indicated for the treatment of rheumatoid arthritis, ankylosing spondylitis, osteoarthritis, and gout. Its potency as an anti-inflammatory, analgesic, antipyretic agent is counterbalanced by accompanying adverse effects, the most common of which are gastrointestinal irritation and headache.

Trade names: Amuno (MSD Sharp & Dohme, Germany), Arthrexin (Lennon Generics, South Africa), Confortid (Dumex, Denmark, Sweden, Finland), Doctucid (Coctum, Greece), Dynamectin (Dynamed, South Africa), Flexidin (Mundipharma, Austria), Inacid (Merck Sharp & Dohme, Spain), Indobene (Merckle, Austria, CIS), Indocin (Merck, USA), Reumadolor (Bros, Greece), Zoflam (Norpharma, Denmark).

Attempts to develop a safer indomethacin have led to the introduction of the following four derivatives.

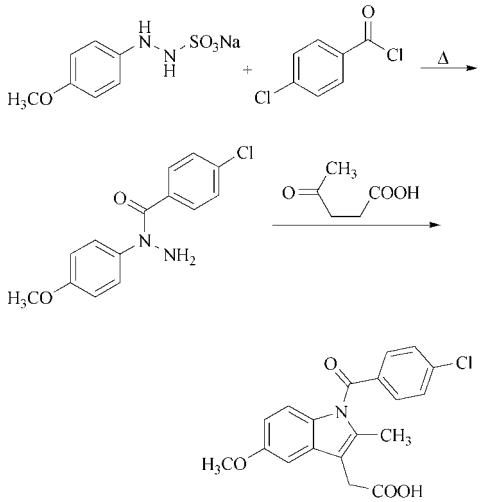

Indomethacin

Acemetacin [*53164-05-9*], {[1-(4-chlorobenzoyl)-5-methoxy-2-methylindol-3-yl]acetoxy}-acetic acid, $C_{21}H_{18}ClNO_6$, M_r 415.60, *mp* 149.5 – 150.5 °C

Synthesis: alkylation of indomethacin with benzyl bromoacetate in K_2CO_3/*N,N*-dimethylformamide gives the corresponding benzyl glycolate ester, which is hydrogenated over 10 % palladium on charcoal in acetic acid to yield acemetacin [25].

Acemetacin is used in the treatment of pain and restricted mobility resulting from chronic articular rheumatism, degenerative articular disease, gout, and inflammation of muscle, joints, and tendons. Acemetacin causes less gastrointestinal blood loss than indomethacin. Its anti-inflammatory activity results from liberation of the parent compound, indomethacin.

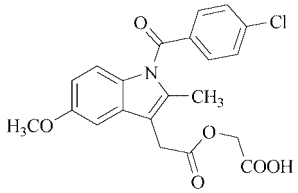

Acemetacin

Trade names: Acemetacin (Heumann, Stada, Germany), Acemix (Bioprogress, Italy), Altren (Rhône-Poulenc Rorer, Belgium), Emflex (Merck, UK), Rantudil (Bayer Pharma Deutschland, Germany).

Glucametacin [*52443-21-7*], 2-({[1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indol-3-yl]acetyl}amino)−2-deoxy-D-glucose, $C_{25}H_{27}ClN_2O_8$, M_r 518.96, *mp* 218 °C decomp.

Synthesis: glucametacin is prepared by acylation of D-glucosamine with indomethacin acid chloride (prepared from indomethacin and thionyl chloride) in the presence of sodium hydroxide [26].

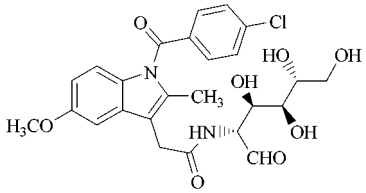

Glucametacin

Glucametacin is used to treat inflammatory and degenerative arthropathy. The compound is well tolerated with significantly less gastrointestinal distress than indomethacin. Glucametacin does not appear to be metabolized to indomethacin.

Trade names: Euminex (Asta Medica, Spain), Teoremin (Labofarma/Degussa, Brazil).

Oxametacine [*27035-30-9*], 1-(4-chlorobenzoyl)-*N*-hydroxy-5-methoxy-2-methyl-1*H*-indole-3-acetamide, $C_{19}H_{17}ClN_2O_4$, M_r 372.81, *mp* 181 – 182 °C

Synthesis: oxametacine is prepared by the acylation of hydroxylamine with indomethacin acid chloride [27].

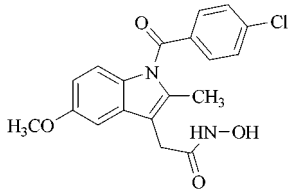

Oxametacin

Oxametacine at a dose of 100 mg is reported to be as effective as 50 mg of indomethacin in reducing pain and inflammation with a lower incidence of adverse effects.

Trade names: Flogar (UCB, Belgium; ABC, Portugal).

Proglumetacin maleate [*59209-40-4*], *N*-{2-[1-(4-chlorobenzoyl)–5-methoxy-2-methyl-3-indolylacetoxy]ethyl}-*N′*-[3-(*N*-benzoyl-*N′*,*N′*-di*n*-propyl-DL-isoglutaminyl)oxypropyl]piperazine (±)-dimaleate, $C_{46}H_{58}ClN_5O_8 \cdot 2\,C_4H_4O_4$, M_r 1075.98, *mp* 146 – 148 °C.

Synthesis: alkylation of the anticholinergic agent proglumide with *N*-(3-chloropropyl)-*N′*-(2-hydroxyethyl)piperazine using sodium methoxide in dimethyl sulfoxide gives *N*-[3-(*N*-benzoyl-*N′*,*N′*-di-*n*-propyl-DL-isoglutaminyl)oxypropyl]-*N′*-(2-hydroxyethyl)piperazine, which is then condensed with indomethacin using dicyclohexylcarbodiimide. The resulting ester is treated with maleic acid to give proglumetacin maleate [28].

Proglumetacin at a dose of 450 mg/d appears to be as effective as 150 mg/d of indomethacin in the treatment of a wide variety of rheumatic conditions, and is faster acting with significantly lower incidence and severity of adverse effects. It is metabolized to several derivatives, including indomethacin, in humans.

Trade names: Afloxan (Rotta Research, Italy), Protaxil (Rottapharm, Spain), Protaxon (Opfermann, Germany), Proxil (Rottapharm, Italy), Tolindol (La Meuse, Belgium).

Sulindac [*38194-50-2*], (*Z*)-5-fluoro-2-methyl-1-{[4-(methylsulfinyl)phenyl]-methylene}-1*H*-indene-3-acetic acid, $C_{20}H_{17}FO_3S$, M_r 356.42, *mp* 182 – 185 °C (decomp.).

Synthesis: Friedel – Crafts reaction of fluorobenzene and α-bromoisobutyryl bromide gives 5-fluoro-2-methylindan-1-one, which is treated with 4-methylthiobenzylmagnesium chloride to yield 5-fluoro-2-methyl-1-(4-methylthiobenzyl)indene. Condensation with glyoxylic acid in the presence of *N*-benzyltrimethylammonium hydroxide (Triton B) gives 3-carboxymethylene-5-fluoro-2-methyl-1-(4-methylthiobenzyl)indene, which is isomerized in acid to 5-fluoro-2-methyl-1-(4-methylthiobenzylidene)indene-3-acetic acid. Oxidation with hydrogen peroxide affords sulindac [29], [30].

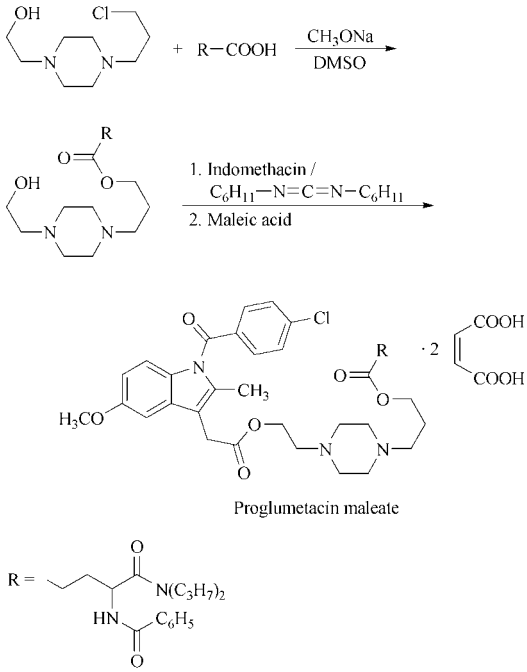

Sulindac is an indene isostere of indomethacin and is a reversible prodrug. Its anti-inflammatory activity depends on metabolic conversion to the corresponding sulfide. The long serum half-life of the sulfide metabolite allows it to be given only once or twice a day. Sulindac is much better tolerated than indomethacin in all five of its indications: rheumatoid arthritis, osteoarthritis, gout, ankylosing spondylitis, and acute painful shoulder. There is a low incidence of adverse effects, consisting chiefly of gastrointestinal disturbances.

Trade names: Aflodac (Biotekfarma, Italy), Algocetil (Francia Farm., Italy), Clinoril (Frosst, The Netherlands; Merck Sharp & Dohme, Aus-

tria, Belgium, Denmark, UK, Slovacian republic); Dorindac (Chibret, Portugal), Zirofalen (Farmalen, Greece).

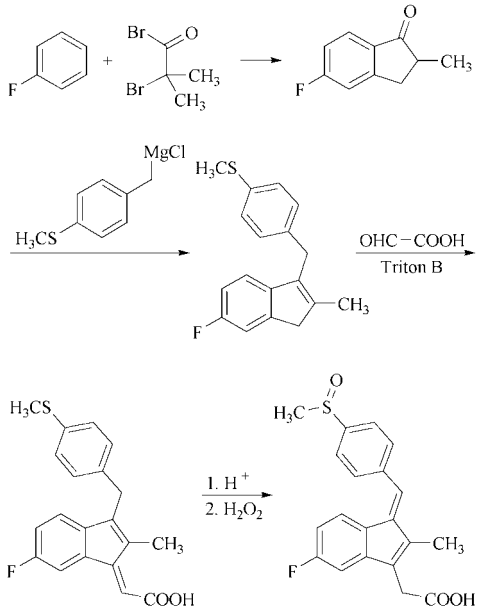

Sulindac

Tolmetin [*26171-23-3*] 1-methyl-5-(4-methylbenzoyl)-1*H*-pyrrole-2-acetic acid, $C_{15}H_{15}NO_3$, M_r 257.30, *mp* 155 – 157 °C.

Synthesis: Friedel – Crafts acylation (aluminum chloride – carbon disulfide) of ethyl 1-methyl-1*H*-pyrrole-2-acetate with 4-toluyl chloride gives ethyl 1-methyl-5-(4-methylbenzoyl)-1*H*-pyrrole-2-acetate, which on saponification yields tolmetin [31].

Tolmetin is an anti-inflammatory, analgesic – antipyretic agent indicated for acute and long-term treatment of rheumatoid arthritis. Its activity is comparable to aspirin and indomethacin, but it causes fewer and milder gastrointestinal disturbances than the former and fewer CNS disturbances than the latter. Other adverse effects include dizziness, tinnitus, edema, and rash. It is administered as its sodium salt dihydrate.

Trade names: Artrocaptin (Estedi, Spain), Tolectin (Cilag, Belgium; Janssen-Cilag, Austria; McNeil, USA).

Metiazinic acid [*13993-65-2*], 10-methylphenothiazine-2-aceticacid, $C_{15}H_{13}NO_2S$, M_r 271.34, *mp* 146 °C.

Synthesis: metiazinic acid is prepared via the Willgerodt reaction (sulfur and morpholine) from 2-acetyl-10-methylphenothiazine [32].

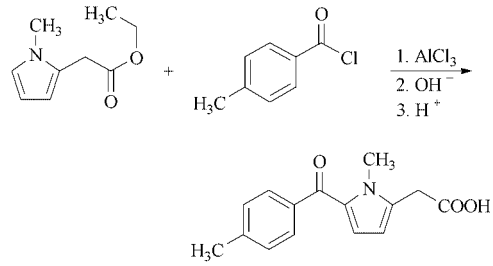

Tolmetin

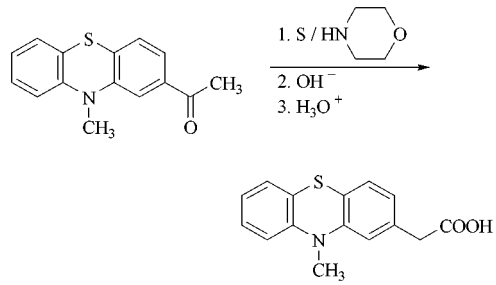

Metiazinic acid

Metiazinic acid has been beneficial to arthritis patients when given at a daily dose of 0.75 – 1.5 g/d.

Trade names: Roimal (Kantoishi, Japan), Soripal (Torii, Japan).

Alclofenac [*22131-79-9*], 3-chloro-4-(2-propenyloxy)benzeneacetic acid, $C_{11}H_{11}ClO_3$, M_r 226.66, *mp* 92 – 93 °C.

Synthesis: alkylation of ethyl 3-chloro-4-hydroxyphenylacetate with allyl bromide in acetone containing potassium carbonate followed by saponification of the ester group gives the product [33].

Alclofenac when used at 3 g/d in the treatment of rheumatoid arthritis is as effective as 75 mg/d of indomethacin with fewer adverse effects. In the treatment of osteoarthritis 1500 mg/d is as effective and is better tolerated than 300 mg/d of phenylbutazone. Alclofenac has been withdrawn from the market

in the United Kingdom because of skin rash and vasculitis (blood vessel inflammation).

Trade names: Mervan (Continental, Belgium).

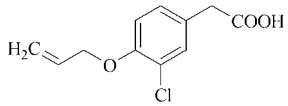

Alclofenac

Diclofenac sodium [*15307-79-6*], 2-[(2,6-dichlorophenyl)amino]benzeneaceticacid monosodium salt, $C_{14}H_{10}Cl_2NNaO_2$, M_r 282.68, *mp* 283 – 285 °C.

Synthesis: acylation of *N*-phenyl-2,6-dichloroaniline with chloroacetyl chloride gives the corresponding chloroacetanilide, which is fused with aluminum chloride to give 1-(2,6-dichlorophenyl)-2-indolinone. Hydrolysis of the indolinone with dilute aqueous-alcoholic sodium hydroxide affords the desired sodium salt directly [34].

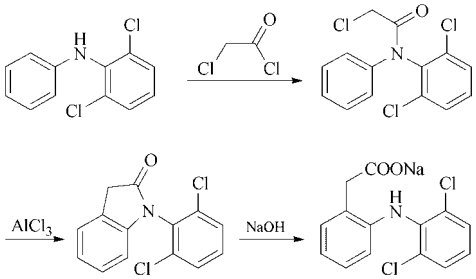

Diclofenac sodium

The structure of diclofenac is a hybrid of a fenamate and an arylacetic acid. The compound is used as its sodium salt for the symptomatic relief of rheumatoid arthritis and osteoarthritis, including degenerative joint disease of the hip. The recommended dose is 75 – 150 mg/d which is clinically equivalent to 3.6 g/d of aspirin. Gastrointestinal problems (ulceration and bleeding) and adverse CNS reactions (dizziness and headache) are the most commonly encountered adverse effects (for review see [35]).

Trade names: Arthrex (BASF GENERICS, Germany), Benfofen (Sanofi Winthrop, Germany), Diclac (Hexal, Germany), Diclobene (Merckle, CIS, Austria), Diclophlogont (Azupharma, Germany), Diclosifar (Siphar, Switzerland), Dolgit (Dolorgiet, Germany), Rewodina (Asta Medica AWD, Germany), Voltaren (Novartis Pharma, Germany), Arthrotec (Searle, UK, Italy).

Fentiazac [*18046-21-4*], 4-(4-chlorophenyl)- 2-phenyl-5-thiazoleaceticacid, $C_{17}H_{12}ClNO_2S$, M_r 329.81, *mp* 161 – 162 °C.

Synthesis: refluxing a mixture of 3-bromo-3-(4-chlorobenzoyl)propionic acid and thiobenzamide in ethanol gives ethyl 2-phenyl-4-(4-chlorophenyl)thiazole-5-acetate, which is then saponified to yield fentiazac [36].

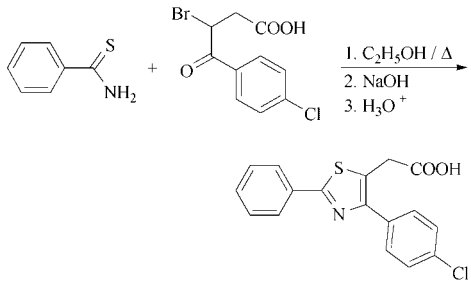

Fentiazac

Fentiazac is reported to have anti-inflammatory, analgesic, and antipyretic activity. It has been given once or twice a day at levels between 100 and 200 mg/dose in the treatment of postoperative pain, including that following orthopedic surgery. The most common adverse effect is gastrointestinal intolerance, including epigastric pain, nausea, and vomiting. Effects on the CNS, such as headache and dizziness, also have been reported.

Trade names: Atilan (Zambon, Brazil), Donorest (Fontoura-Wyeth, Brazil), Flogene (Polifarma, Italy), Norvedan (Boehringer Mannheim, Austria; LPB, Italy).

3.3. Arylacetic Acid Prodrugs

Metabolic precursors to arylacetic acids have been introduced into rheumatoid therapy in the form of 4-aryl-4-oxobutyric acids. ω-Phenyl fatty acids undergo normal two-carbon degradation of the side chain to give either benzoic or phenylacetic acids depending on the number of carbon atoms in the original molecule. Therefore, ω-phenyl-γ-oxoacids were presumed to behave similarly; e.g., 4-phenyl-4-oxobutyric and

6-phenyl-4-oxo-hexanoic acid are degraded in vivo to phenylacetic acid.

The oxoacids are not very potent inhibitors of cyclooxygenase and their in vivo activity is probably dependent on metabolic conversion to the corresponding phenylacetic acid.

Fenbufen [*36330-85-5*], γ-oxo(1,1′biphenyl)-4-butanoic acid, $C_{16}H_{14}O_3$, M_r 254.29, *mp* 185 – 187 °C.

Synthesis: Fenbufen is prepared by the Friedel-Crafts (aluminum chloride – nitrobenzene) acylation of biphenyl with succinic anhydride [37].

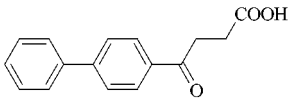

Fenbufen

Fenbufen has been found to be an effective, well-tolerated drug for the treatment of rheumatoid arthritis, osteoarthritis, and ankylosing spondylitis. The compound is metabolized in humans first to 4-hydroxy-4-biphenylbutyric acid (t_{max} 2.5 h) then to 4-biphenyl acetic acid (t_{max} 7.5 h). Both metabolites are more active than fenbufen itself and circulate for several hours ($t_{1/2}$ 10 h). This slow conversion of fenbufen to active metabolites having relatively long plasma half-lives allows for once a day dosing with this agent.

Trade names: Clincopal (Lederle, Spain), Lederfen (Lederle, UK), Napanol (Lederle, Japan).

3.4. Arylpropionic Acids

Ibuprofen [*15687-27-1*], α-methyl-4-(2-methylpropyl)benzeneaceticacid, $C_{13}H_{18}O_2$, M_r 206.27, *mp* 75 – 77 °C.

Synthesis: treatment of ethyl 4-isobutylphenylacetate and diethyl carbonate with sodium ethoxide gives diethyl 4-isobutylphenylmalonate, which is methylated using methyl iodide and sodium ethoxide. Saponification followed by decarboxylation of the resulting malonic acid derivative affords ibuprofen [38].

Ibuprofen is indicated for both acute and long-term management of rheumatoid arthritis, osteoarthritis, and psoriatic arthritis. Anti-inflammatory efficacy depends on relatively high dosage, 2.4 g/d. The drug is well tolerated, the chief adverse effect being a low incidence of gastrointestinal disturbance. The drug is rapidly absorbed and eliminated with a serum half-life of about 2 h. Ibuprofen does not interfere with the protein binding of warfarin and can be used safely in patients undergoing anticoagulant therapy. It is available as a nonprescription drug.

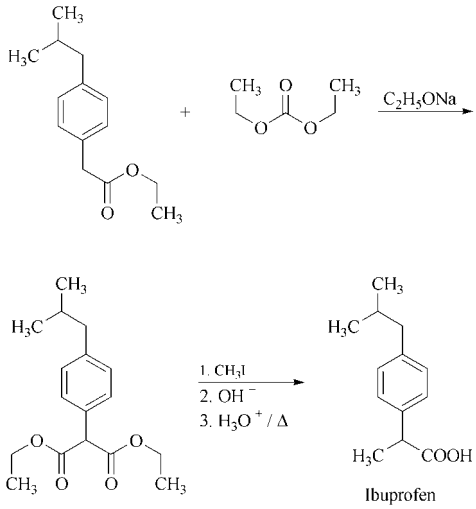

Trade names: Advil (Whitehall, USA; Whitehall-Robins, Canada); Aktren (Bayer Selbstmed, Germany), Algifor (Vifor, Switzerland), Ardinex (Boots, Finland), Brufen (Boots, Portugal; Knoll, Belgium), Dismenol (Merz Consumer Division, Germany), Dolgit (Dolorgiet, Germany), Dolo-Puren (Isis Puren, Germany), Dolormin (Woelm, Germany), Duafen (Pharmed, Austria), Haltran (Upjohn, USA), Ibufen (Amino, Switzerland), Ibunet (Nettopharma, Denmark), Rufen (Knoll, USA).

Naproxen [*22204-53-1*], (+)-6-methoxy-α-methyl-2-naphthaleneaceticacid, $C_{14}H_{14}O_3$, M_r 230.26, *mp* 155.3 °C, $[\alpha]_D$ +65.5° (*c* 1.0, $CHCl_3$).

Synthesis: Friedel - Crafts acylation (aluminum chloride - nitrobenzene) of β-naphthol methyl ether affords 2-acetyl-6-methoxynaphthalene, which, when treated with either dimethylsulfonium or dimethylsulfoxonium methylide, gives 2-(6-methoxynaphthalen-2-yl)propylene oxide. Treatment

of the latter with boron trifluoride etherate in tetrahydrofuran gives 2-(6-methoxynaphthalen-2-yl)propionaldehyde, which is oxidized using Jones reagent (4 M chromic acid) to yield the racemic 2-(6-methoxynaphthalen-2-yl)propionic acid. Resolution and isolation of the dextrorotatory enantiomer is accomplished via its cinchonidine salt [39].

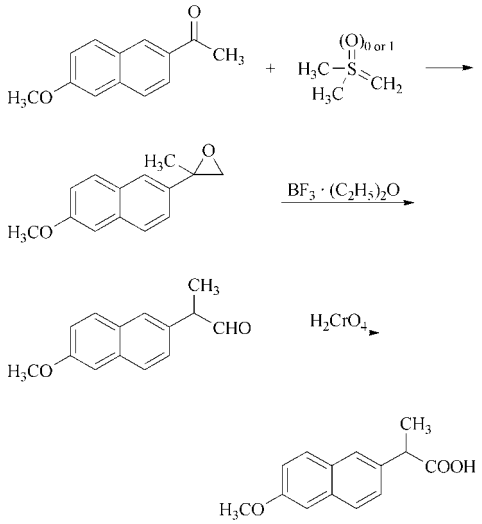

Naproxen

Naproxen is used in the treatment of rheumatoid arthritis, osteoarthritis, gout, and ankylosing spondylitis at a daily dose of 500 – 700 mg. Absorption from the gastrointestinal tract is virtually complete, and extensive binding (99 %) to serum albumin provides for a relatively long plasma half-life (13 h) and twice-a-day dosing. Naproxen is reasonably well-tolerated with gastrointestinal disturbances being the principal adverse effect. Its more rapidly absorbed sodium salt is marketed as an analgesic providing earlier pain relief.

Trade names: Anaprox (Syntex, Canada, USA), Apranax (Roche, France), Bonyl (Ercopharm, Denmark), Miranax (Syntex, Finland), Novo-Naprox (Novorpharm, Canada), Proxen (Hoffmann La Roche, Germany).

Ketoprofen [*22071-15-4*], 3-benzoyl-α-methylbenzeneaceticacid, $C_{16}H_{14}O_3$, M_r 254.29, *mp* 94 °C.

Synthesis: bromination (bromine – ultraviolet light) of 3-methylbenzophenone and nucleophilic substitution of the resulting bromomethyl compound with sodium cyanide gives 3-benzoylphenylacetonitrile. Treatment with diethyl carbonate and sodium ethoxide followed by alkylation of the resulting 2-cyano ester gives ethyl 2-(3-benzoylphenyl)-2-cyanopropionate. Acid hydrolysis and decarboxylation yields ketoprofen [40].

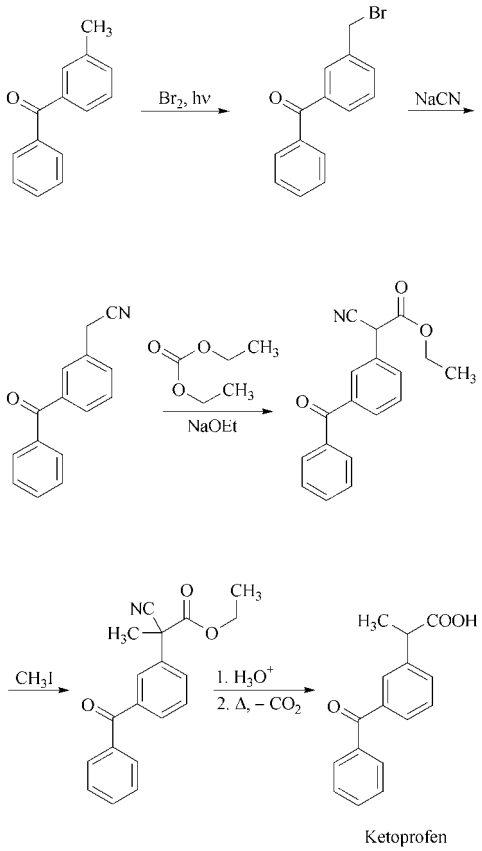

Ketoprofen

Ketoprofen is indicated in the treatment of rheumatoid arthritis, ankylosing spondylitis, and osteoarthritis at a daily dose of 150 – 300 mg. At 150 mg/d its clinical efficacy is comparable to a similar dose of indomethacin, but fewer and less severe adverse effects are observed.

Trade names: Alrheumat (Bayer, United Kingdom), Alrheumun (Bayer Pharma Deutschland, Germany), Gabrilen (Kreussler, Germany), Orudis (Rhône-Poulenc Rorer, Canada, Denmark; Wyeth-Ayerst, USA).

Fenoprofen [*31879-05-7*], α-methyl-3-phenoxybenzeneaceticacid, $C_{15}H_{14}O_3$, M_r 242.28, *bp* 168 – 171 °C (0.015 kPa).

Synthesis: sodium borohydride reduction of 3-phenoxyacetophenone followed by bromination of the resulting alcohol with PBr_3 gives α-methyl-3-phenoxybenzyl bromide. Reaction of this bromide with sodium cyanide in dimethyl sulfoxide gives the corresponding nitrile, which is hydrolyzed using sodium hydroxide. Acidification affords fenoprofen [41].

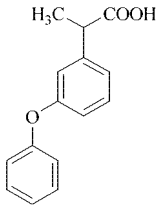

Fenoprofen

Fenoprofen is used as its calcium salt dihydrate in the treatment of rheumatoid arthritis and osteoarthritis at a daily dose of 1.2 – 3.0 g. The drug is rapidly absorbed and excreted with a plasma half-life of about 3 h despite being extensively bound (99 %) to plasma protein. Fenoprofen is well tolerated, with dyspepsia being the chief adverse effect.

Trade names: Fenopron (Lilly, South Africa, United Kingdom), Fepron (Lilly, Italy), Nalfon (Dista, USA, Austria, Canada), Nalgesic (Lilly, France), Progesic (Lilly, United Kingdom).

Flurbiprofen [*5104-49-4*], 2-fluoro-α-methyl(1,1′-biphenyl)-4-aceticacid, $C_{15}H_{13}FO_2$, M_r 244.27, *mp* 110 – 111 °C.

Synthesis: ethyl 2-fluoro-4-biphenylacetate, prepared from 4-acetyl-2-fluorobiphenyl via the Willgerodt reaction and subsequent esterification, is treated first with diethyl carbonate and sodium ethoxide, then with dimethyl sulfate to give diethyl 2-(2-fluoro-4-biphenylyl)-2-methyl malonate. Saponification followed by acidification affords the malonic acid, which is decarboxylated at 180 – 200 °C to yield flurbiprofen [42].

Flurbiprofen has been used at 150 – 300 mg/d in the treatment of rheumatoid arthritis, osteoarthritis, and ankylosing spondylitis. Gastrointestinal disturbances, particularly dyspepsia, are the principal adverse effects.

Trade names: Ansaid (Pharmacia & Upjohn, CIS; Upjohn, Canada, Czech Republic, Poland, USA), Cebutid (Knoll, France), Froben (Kanoldt, Germany; Ebewe, Austria; Boots, Spain, The Netherlands, Portugal), Tulip (Upjohn, Spain).

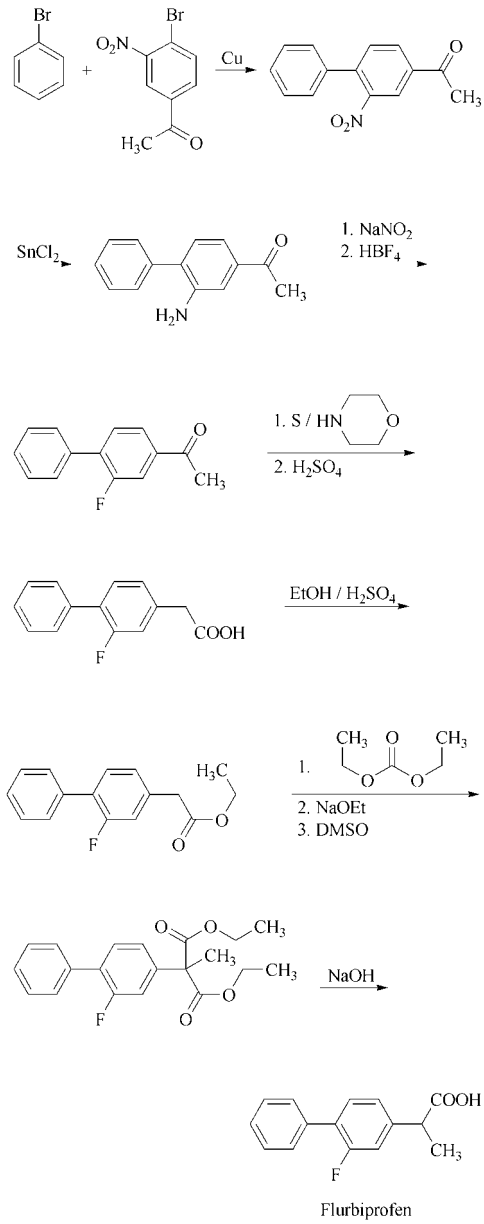

Flurbiprofen

Clidanac [*34148-01-1*], 6-chloro-5-cyclohexyl-2,3-dihydro-1*H*-indene-1-carboxylic acid, $C_{16}H_{19}ClO_2$, M_r 278.78, *mp* 150.5 – 152.5 °C.

Synthesis: formylation of cyclohexylbenzene using dichloromethyl methyl ether and titanium tetrachloride gives 4-cyclohexylbenzaldehyde, which is condensed with diethyl malonate to yield diethyl 4-cyclohexylbenzylidenemalonate. Treatment of the latter with potassium cyanide in ethanol at 70 °C gives ethyl 3-cyano-3-(4-cyclohexylphenyl)propionate, which is converted to 4-cyclohexylphenylsuccinic anhydride following acid hydrolysis and treatment with refluxing acetic anhydride. An intramolecular Friedel – Crafts acylation (aluminum chloride – methylene chloride) affords 5-cyclohexyl-3-oxo-1-indanecarboxylic acid, which is catalytically reduced using 10 % palladium on charcoal in the presence of perchloric acid to give 5-cyclohexyl-1-indanecarboxylic acid. Chlorination using *N*-chlorosuccinimide in *N*,*N*-dimethylformamide (DMF) gives clidenac [43].

Clidenac is used for the relief of pain and inflammation related to arthrosis deformans, periarthritis humeroscapularis, cervical syndrome, and lumbago.

Trade names: Britai (Bristol Banyu, Japan), Indanal (Takeda, Japan).

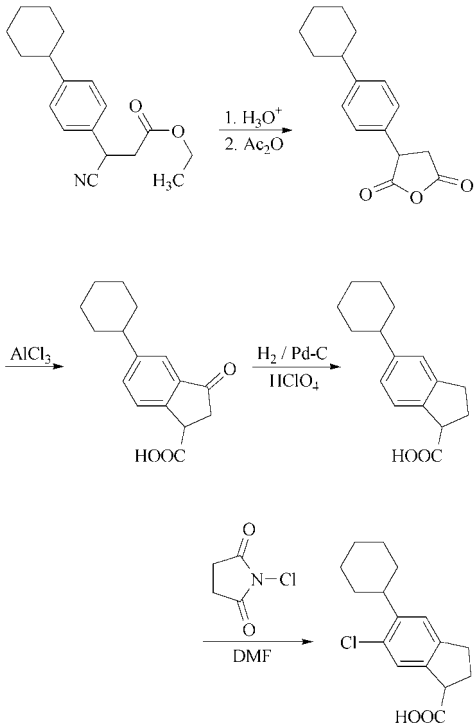

Clidanac

Pirprofen [*31793-07-4*], 3-chloro-4-(2,5-dihydro-1*H*-pyrrol-1-yl)-α-methylbenzeneacetic acid, $C_{13}H_{14}ClNO_2$, M_r 251.71, *mp* 98 – 100 °C.

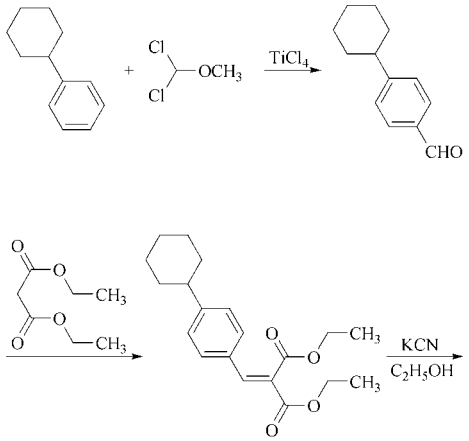

Synthesis: treatment of the sodium salt of diethyl methylmalonate with 2,4-dichloronitrobenzene yields diethyl (3-chloro-4-nitrophenyl)methylmalonate. Saponification, decarboxylation, and subsequent reesterification followed by catalytic reduction gives ethyl 4-amino-3-chloro-α-methylbenzeneacetate hydrochloride. Treatment of the latter with 1,4-dichloro-2-butene in the presence of sodium carbonate followed by saponification affords pirprofen [44].

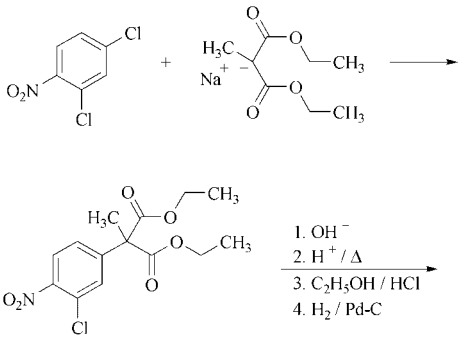

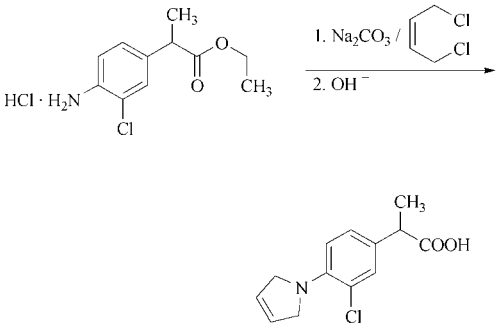

Pirprofen

Pirprofen has been used to treat rheumatoid arthritis, osteoarthritis, and ankylosing spondylitis. An optimal dosing regimen of 200 mg three times a day has been developed for maximal activity with minimal adverse effects. Pirprofen also is effective in relieving pain from malignant disease and oral surgery.

Trade name: Rengasil (Ciba, Greece), Seflenyl (Geigy, Argentina).

3.5. Acidic Enolic Compounds

Phenylbutazone [*50-33-9*], 4-*n*-butyl-1,2-diphenyl-3,5-pyrazolidinedione, $C_{19}H_{20}N_2O_2$, M_r 308.37, *mp* 105 °C.

Synthesis: heating diethyl-*n*-butylmalonate and 1,2-diphenylhydrazine in the presence of sodium ethoxide at 150 °C followed by acidification affords phenylbutazone in good yield. Alternatively, *n*-butylmalonyl chloride in pyridine and ether at 0 °C is treated with 1,2-diphenylhydrazine to give the desired product [45].

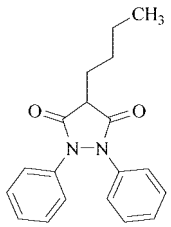

Phenylbutazone

Phenylbutazone, one of the earliest NSAIDs introduced, is now indicated for the symptomatic relief of rheumatoid arthritis, osteoarthritis, psoriatic arthritis, ankylosing spondylitis, gout, and acute superficial thrombophlebitis.

The gastrointestinal and bone marrow toxicity observed in its early use have been greatly reduced by lower dosage (300 mg/d). Nevertheless, it is used primarily where other drugs have failed and then only for short-term therapy. The drug has a long serum half-life of about 100 h. It is a moderately active cyclooxygenase inhibitor and it suppresses both spontaneous and chemotactic motility of neutrophils. In addition to the serious gastrointestinal and hematological adverse effects, sodium and water retention, rash, vertigo, and dermatitis are observed.

Trade names: Ambene (Merckle, Germany), Butazolidin (Novartis Pharma, Germany), Butazone (Major, USA), Butrex (Noristan, South Africa), Ticinil (Boehringer Ingelheim, Italy).

Oxyphenbutazone [*129-20-4*], 4-*n*-butyl-1-(4-hydroxyphenyl)-2-phenyl-3,5-pyrazolidinedione, $C_{19}H_{20}N_2O_3$, M_r 324.37, *mp* 124 – 125 °C.

Synthesis: heating 1-(4-benzyloxyphenyl)-2-phenylhydrazine with diethyl-*n*-butylmalonate in the presence of sodium ethoxide at 140 °C followed by acidification gives 1-(4-benzyloxyphenyl)-4-*n*-butyl-2-phenyl-pyrazolidine-3,5-dione, which is unblocked by catalytic reduction (Raney nickel) to yield oxyphenbutazone [46].

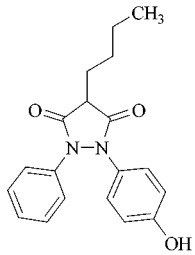

Oxyphenbutazone

Oxyphenbutazone is one of the metabolites formed in the liver following administration of phenylbutazone. Given orally it causes fewer gastrointestinal adverse effects than phenylbutazone and is used at 300 – 400 mg/d for the same indications as the parent drug.

Trade names: Diflamil (Belmac, Spain), Phlogont (Azupharma, Germany), Tabazone (Major, USA).

Apazone, azapropazone [*13539-59-8*], 5-(dimethylamino)-9-methyl-2-propyl-1*H*-pyrazolo[1,2*a*][1,2,4]benzotriazine-1,3-(2*H*)-dione, $C_{16}H_{20}N_4O_2$, M_r 300.37, *mp* 228 °C.

Synthesis: a mixture of 1,2-dihydro-3-dimethylamino-7-methylbenzotriazine, diethyl propylmalonate, and sodium ethoxide in xylene is heated over a period of time to 150 °C, cooled, and acidified to give the desired product [47].

Apazone is a complex pyrazolidinedione used at 600 – 1200 mg/d to achieve an anti-inflammatory effect comparable to phenylbutazone. It is tightly bound to plasma protein and may displace anticoagulants and hypoglycemic agents. Its plasma half-life is 8 – 12 h. It is well tolerated with only a low incidence of gastric disturbances. As with all pyrazolone derivatives, the potential for bone marrow suppression exists and careful monitoring of the patient's hematologic state is recommended.

Trade names: Cinnamin (Chemiphar, Japan), Prolixan (Siegfried, Greece), Rheumox (Wyeth, UK), Xani (Farmakos, Yugoslavia).

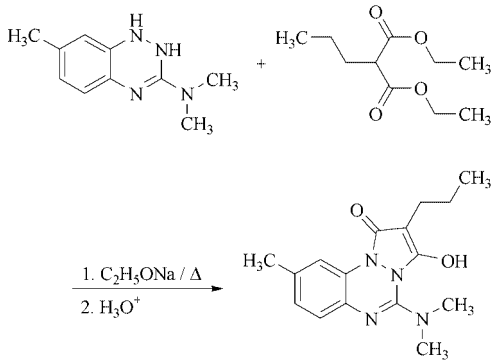

Apazone

Piroxicam [*36322-90-4*], 4-hydroxy-2-methyl-*N*-2-pyridinyl-2*H*-1,2-benzothiazine-3-carboxamide1,1-dioxide, $C_{15}H_{13}N_3O_4S$, M_r 331.35, *mp* 198 – 200 °C.

Synthesis: an improved procedure using 2-methoxyethyl 2-chloroacetate in place of methyl 2-chloroacetate for the alkylation of sodium saccharin has been described [48]. The resulting 2-methoxyethyl saccharin-2-acetate is treated with sodium 2-methoxyethoxide in dimethyl sulfoxide, then acidified to give 2-methoxyethyl 4-hydroxy-2*H*-1,2-benzothiazine-3-carboxylate 1,1-dioxide, which is *N*-alkylated with methyl iodide in acetone – aqueous sodium hydroxide. The resulting 2-methoxyethyl 4-hydroxy-2-methyl-2*H*-1,2-benzothiazine-3-carboxylate 1,1-dioxide is heated with 2-aminopyridine in xylene to give piroxicam [49].

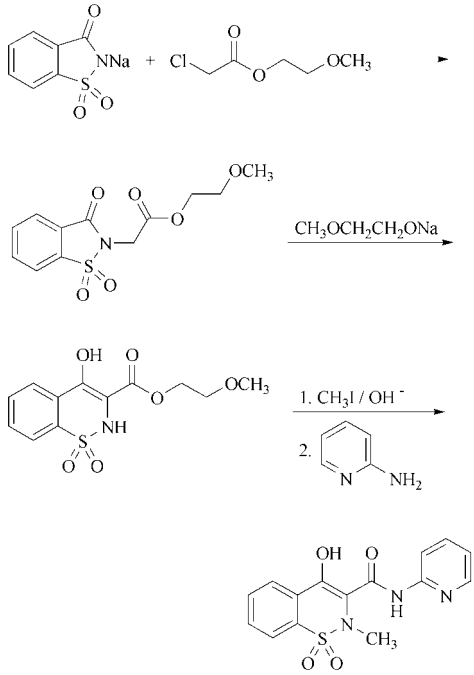

Piroxicam

Piroxicam, the first of a new class of NSAIDs known as "oxicams," was introduced in 1979 for the treatment of rheumatoid arthritis, osteoarthritis, ankylosing spondylitis, gout, and acute musculoskeletal disorders [48]. This compound was the result of a long and intensive effort to develop a potent NSAID having a novel structure and a plasma half-life sufficiently long to maintain continuous therapeutic blood levels on a once-a-day dosing regimen. Piroxicam is a potent competitive inhibitor of cyclooxygenase and lowers blood levels of prostaglandins in humans. Its ability to inhibit the infiltration of polymorphonuclear leukocytes into inflamed joints has been demonstrated in animals. Its potency and long plasma half-life of 40 h in humans allow for once-a-day dosing at 20 mg. A high incidence of gastric irritation results from this agent.

Trade names: Brexidol (Pharmacia & Upjohn, Germany), Brexin (Pharmacia & Up-

john, Austria), Doblexan (Organon, Spain), Durapirox (Durachemie, Germany), Felden (Mack/Pfizer, Germany), Piro-Phlogont (Azupharma, Germany), Ruvamed (Coup, Greece).

3.6. Anthranilates

Mefenamic acid [*61-68-7*], 2-[(2,3-dimethylphenyl)amino]benzoic acid, $C_{15}H_{15}NO_2$, M_r 241.28, *mp* 230 – 231 °C.

Synthesis: mefenamic acid is prepared via the Jourdan – Ullmann – Goldberg synthesis utilizing either anthranilic acid and 3-bromo-1,2-dimethylbenzene or 2,3-dimethylaniline and an *o*-halobenzoic acid in the presence of a copper catalyst and a proton acceptor [50].

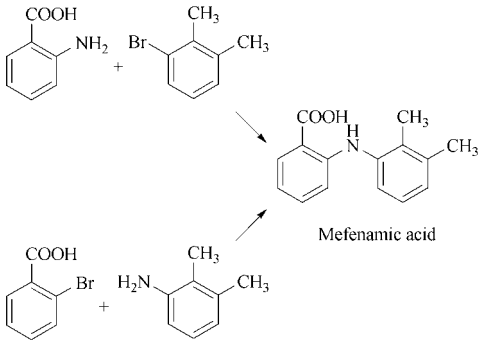

Mefenamic acid has mild anti-inflammatory properties and is used primarily as a short-term analgesic. Gastrointestinal disturbances, including possibly allergic diarrhea and potential renal toxicity, limit its use.

Trade names: Lysalgo (SIT, Italy), Opustan (Opus Pharm, UK), Parkemed (Parke Davis, Germany), Ponstan (Werner-Lambert, Switzerland), Ponstel (Parke Davis, USA), Pontal (Sankyo, Japan).

Flufenamic acid [*530-78-9*], 2-{[3-(trifluoromethyl)phenyl]amino}benzoicacid, $C_{14}H_{10}F_3NO_2$, M_r 281.24, *mp* 125 °C.

Synthesis: refluxing a mixture of 2-iodobenzoic acid and 3-trifluoromethylaniline in aqueous potassium carbonate containing copper-bronze as the catalyst affords, after acidification, the desired flufenamic acid [51].

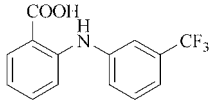

Flufenamic acid

Flufenamic acid is used at 600 – 800 mg/d to provide a beneficial therapeutic effect in chronic polyarthritis. The adverse effects most often encountered are gastrointestinal disturbances.

Trade names: Ansatin (Ono, Japan), Arlef (Sankyo, Japan), Felunamin (Hokuriko, Japan), Romafen (Biofarma, Turkey).

Meclofenamic Acid [*644-62-2*], 2-[(2,6-dichloro-3-methylphenyl)amino]benzoicacid, $C_{14}H_{11}Cl_2NO_2$, M_r 296.15, *mp* 257 – 259 °C.

Synthesis: heating a mixture of potassium 2-bromobenzoate, 2,6-dichloro-3-methylaniline, morpholine, diglyme (diethylene glycol dimethyl ether), and cupric bromide at 145 – 155 °C and subsequent acidification afford meclofenamic acid [52].

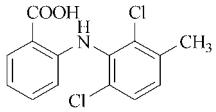

Meclofenamic acid

Meclofenamic acid is available as its sodium salt hydrate. This drug is indicated for the relief of acute and chronic rheumatoid arthritis and osteoarthritis. Gastrointestinal distress, including diarrhea, nausea, vomiting, and abdominal pain, is the most commonly reported adverse effect.

Trade names: Meclodol (Parke Davis, Italy), Meclomen (Parke Davis, Portugal, USA), Stadium (Menarini, Belgium).

Niflumic acid [*4394-00-7*], 2-{[3-(trifluoromethyl)phenyl]amino}-3-pyridinecarboxylic-acid, $C_{13}H_9F_3N_2O_2$, M_r 282.23, *mp* 204 °C.

Synthesis: heating a mixture of 2-chloronicotinic acid, 3-trifluoromethylaniline, and a catalytic amount of potassium iodide at 140 °C affords the desired niflumic acid [53].

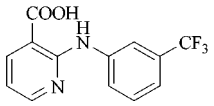

Niflumic acid

Niflumic acid has been used effectively to treat rheumatoid arthritis, psoriatic arthritis, and hypertrophic osteoarthritis of the hip and knee. Gastric complications are the chief adverse effects of this drug.

Trade names: Actol (Mayrhofer, Austria; Upsamedica, Spain), Donalgin (Gedeon Richter, Hungary), Livornex (Genepharm, Greece), Niflam (Upsamedica, Italy), Nifluril (UPSA, France; Upsamedica, Belgium, Switzerland, Portugal).

3.7. Nonacidic Drugs

Proquazone [*22760-18-5*], 7-methyl-1-(1-methylethyl)-4-phenyl-2(1*H*)-quinazolinone, $C_{18}H_{18}N_2O$, M_r 278.35, *mp* 137 – 138 °C.

Synthesis: proquazone is prepared by the alkylation of 2-amino-4-methylbenzophenone with 2-iodopropane in the presence of sodium carbonate to give 4-methyl-2-[(1-methylethyl)amino]benzophenone, which is condensed with urethane in the presence of zinc chloride at 190 °C [54].

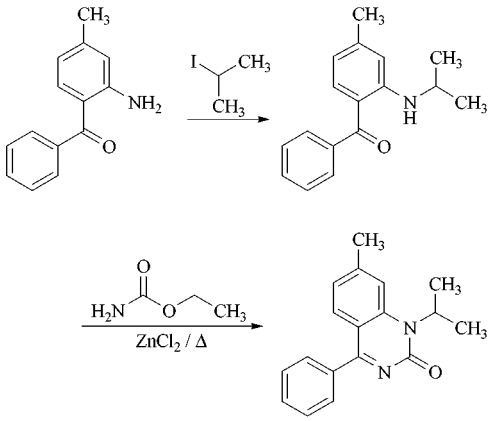

Proquazone

Proquazone has been used successfully to treat rheumatoid polyarthritis. At 900 mg/d it is comparable in effect to 150 mg/d of indomethacin. Gastrointestinal disturbances are the principal adverse effects observed with its use.

Trade names: Biarison (Swiss-Pharma, Egypt; Biochemie, Austria; ICN Alkaloida, Hungary; Sandoz, Finland, Turkey).

Ditazole [*18471-20-0*], 2,2′-[(4,5-diphenyl-2- oxazolyl)imino]diethanol, $C_{19}H_{20}N_2O_3$, M_r 324.38, *mp* 96 – 98 °C (monohydrate).

Synthesis: ditazole is prepared by heating 2-chloro-4,5-diphenyloxazole and diethanolamine in ethanol [55].

Ditazole is an anti-inflammatory analgesic agent that is of interest primarily for its ability to inhibit platelet aggregation. For example, it normalizes the enhanced platelet activity in cerebrovascular patients suffering transient ischemic attacks.

Trade name: Ageroplas (Lepori, Spain; Serono, Italy), Fendazol (Lepori, Portugal).

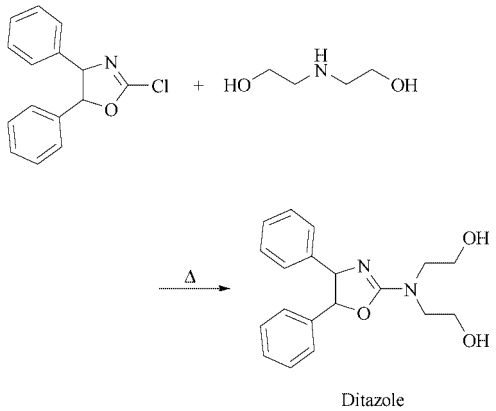

Ditazole

3.8. Selective COX-2 Inhibitors

Celecoxib [*169590-42-5*], 4-[5-(4-methylphenyl)-3-(trifluoromethylpyrazol-1-yl]benzenesulfonamide, $C_{17}H_{14}F_3N_3O_2S$, M_r 381.37, *mp* 157 – 159 °C.

Synthesis: Celecoxib is prepared by condensation of 4-methylacetophenone with ethyl trifluoroacetate to give 4,4,4-trifluoro-1-(4-methylphenyl)butane-1,3-dione, which is cyclized with 4-hydrazinophenylsulfonamide [56]:

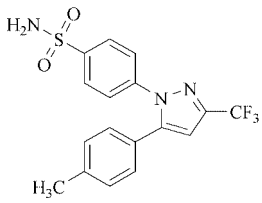

Celecoxib

Celecoxib (SC-58635) is a selective inhibitor of the COX-2 isozyme. Based on human recombinant enzyme assays celecoxib was shown to be 375-fold more selective for COX-2 (IC_{50} for COX-2: 4×10^{-8} M vs. IC_{50} for COX-1: 1.5×10^{-5} M). Celecoxib displays a plasma half-life of 10 – 12 h. Celecoxib is indicated for relief of the signs and symptoms of osteoarthritis (recommended oral dose is 200 mg per day administered as a single dose or as 100 mg twice per day) and rheumatoid arthritis in adults (recommended oral dose is 100 to 200 mg twice per day). In comparison to traditional NSAIDs, celecoxib causes fewer gastrointestinal side effects. Celecoxib has been associated with adverse effects such as headache, change in bowel habits, abdominal discomfort, and dizziness.

Trade name: Celebrex (G. D. Searle & Co., USA).

Rofecoxib [*162011-90-7*], 4-[4-(methylsulfonyl)phenyl]-3-phenylfuran-2(5*H*)-one, $C_{17}H_{14}O_4S$, M_r 314.36, *mp* 208 °C.

Synthesis: Rofecoxib can be obtained by different synthetic routes [57], e.g., by condensation of phenylacetic acid with ethyl bromoacetate to ethyl 2-phenylacetoxyacetate, which is then cyclized to a hydroxyfuranone. Subsequently, the hydroxyfuranone reacts with trifluoromethanesulfonic (triflic) anhydride to the corresponding triflate which reacts with LiBr to yield a bromofuranone. The bromofuranone is condensed with 4-(methylsulfanyl)phenylboronic acid to give 4-[4-(methylsulfanyl)phenyl]-3-phenylfuran-2(5*H*)-one which is finally oxidized to rofecoxib.

The selective COX-2 inhibitor rofecoxib is a methylsulphonylphenyl derivative. In comparison to celecoxib, rofecoxib is slightly more potent and has a longer half life of ca. 17 h. Based on human recombinant enzyme assays rofecoxib was shown to be > 800-fold more selective for COX-2 (IC_{50} for COX-2: 1.8×10^{-8} M vs. IC_{50} for COX-1: 1.5×10^{-5} M). Rofecoxib is indicated for relief of the signs and symptoms of osteoarthritis (recommended starting dose is 12.5 mg once daily, maximum recommended dose is 25 mg/d), for the management of acute pain in adults and for the treatment of primary dysmenorrhea (recommended initial doses are 50 mg once daily, use of rofecoxib for more than 5 d in management of pain has not been studied). Rofecoxib causes a significantly lower incidence of upper-gastrointestinal adverse effects (perforations, ulcers, and bleeding) than conventional NSAIDs. Most common adverse events associated with rofecoxib are diarrhoea, headache, nausea, and upper respiratory tract infection.

Trade name: Vioxx (Merck & Co., USA).

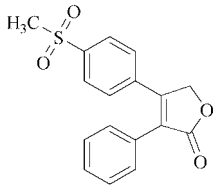

Rofecoxib

4. Antirheumatic Agents

This class of agents encompasses a wide variety of structural and pharmacologic types, including gold, antimalarials, sulfhydryl compounds, and immunoregulants (both immunosuppressants and immunostimulants). These drugs are characterized in many instances by a slow onset of activity and, in contrast to the anti-inflammatory agents discussed heretofore, do have a beneficial effect on the underlying disease process. The acronyms SAARDs (slow acting antirheumatic drugs) and DMARDs (disease modifying antirheumatic drugs) have been applied to these substances.

4.1. Gold

Until the advent of the orally administered auranofin, gold preparations used in the treatment of rheumatoid arthritis were administered parenterally. Ionic gold in its monovalent state is preferred. Gold(I) has a strong affinity for sulfur relative to other elements and the antirheumatic preparations used contain gold bound to an organosulfur carrier. Indeed, the ability of gold to bind to physiologic sulfhydryl groups is thought to be an important factor in its bioactivity. Nevertheless, the actual mechanism of action remains to be determined.

The following activities of gold related to inflammation have been observed:

1) protection of collagen against collagenase

2) inhibition of lysosomal enzymes
3) prevention of denaturation of macro-globulins and formation of antigen – antibody complexes
4) uncoupling of oxidative phosphorylation
5) inhibition of leukocyte chemotaxis

Chrysotherapy (treatment with gold salts) is successful in approximately 80 % of patients. Profound modification and remission of the disease, including retardation of the abnormal growth of synovial tissue over the joint surface, are observed. Unfortunately, a high incidence of adverse effects (25 – 50 %) counterbalances the beneficial effects. Cutaneous hypersensitivity and lesions of mucous membranes are most common. Blood disorders including thrombocytopenia, leukopenia, agranulocytosis, and aplastic anemia occur. Renal effects, as evidenced by proteinuria, are observed.

Aurothioglucose [*12192-57-3*], (1-thio-D-glucopyranosato)gold, $C_6H_{11}AuO_5S$, M_r 392.22.

Synthesis: gold thioglucose is prepared by adding a solution of gold bromide to an aqueous solution of thioglucose that contains sulfur dioxide. After heating, the product is precipitated by the addition of ethanol [58].

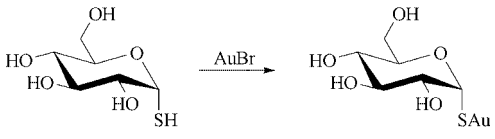

Aurothioglucose

Aurothioglucose is an antirheumatic used to treat active progressing rheumatoid arthritis and nondisseminated lupus erythematosus. The drug is administered at weekly intervals by intramuscular injection (10 mg, 25 mg, then 50 mg) until 800 mg to 1 g has been given. If improvement takes place, the drug is then administered at levels that balance the urinary excretion of gold. During this maintenance therapy the interval between injections is lengthened to 3 – 4 weeks.

Trade names: Aureotan (BYK Gulden, Germany), Solganal (Schering, USA, Yugoslavia).

Sodium aurothiomalate [*12244-57-4*], mercaptobutanedioic acid monogold(I) sodiumsalt, $C_4H_4AuNaO_4S/C_4H_3AuNa_2O_4S$.

Synthesis: a solution of thiomalic acid and 3 equivalents of sodium hydroxide are mixed with an aqueous suspension of gold(I) iodide. The product, a mixture of the mono- and disodium salts, is precipitated by the addition of ethanol [59].

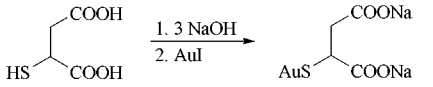

Sodium aurothiomalate

Sodium aurothiomalate is administered in the same manner and for the same indications as aurothioglucose.

Trade names: Aurolate (Pasadena, USA), Myocrisin (Rhône-Poulenc Rorer, Denmark, Sweden, Finland), Tauredon (BYK Gulden, Germany).

Auranofin [*34031-32-8*], (2,3,4,6-tetra-*O*-acetyl-1-thio-β-D-glucopyranosato-*S*)(triethylphosphine)gold, $C_{20}H_{34}AuO_9PS$, M_r 678.49, *mp* 110 – 111 °C.

Synthesis: ethanolic thiodiglycol is treated first with aqueous gold(I) acid chloride trihydrate, then with ethanolic triethylphosphine to give triethylphosphine gold(I) chloride, which is added to an aqueous solution of *S*-(2,3,4,6-tetra-*O*-acetylglucopyranosyl)pseudothiourea hydrobromide and potassium carbonate to give the desired auranofin [60].

Auranofin was the first orally active gold preparation available for treatment of rheumatoid arthritis. Although it is incompletely absorbed, administration of 3 mg twice a day has resulted in profound improvement in patients, approaching the efficacy of parenteral gold, after a period of 8 – 12 weeks. The principal adverse effects with auranofin are diarrhea and loose stools. Mucocutaneous effects, such as rash and pruritis, also were seen, as were abnormalities in taste; but in general, it seems less toxic than parenteral gold.

Trade names: Aktil (Lek, Yugoslavia), Ridaura (Smith Kline Beecham, USA), Ridauran (Robapharm, France).

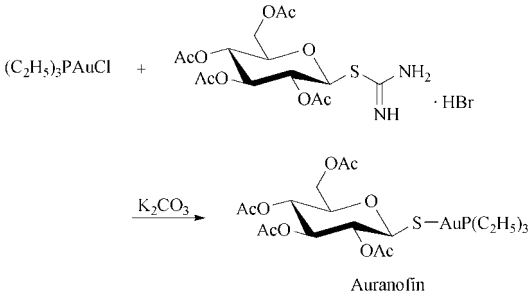

Auranofin

4.2. Antimalarials

Several antimalarial agents – chloroquine, hydroxychloroquine, and dapsone – possess clinically useful anti-inflammatory properties. Hydroxychloroquine is preferred over chloroquine for the treatment of rheumatoid, but not other types of arthritis. When used clinically at a low dose of 4 – 6 mg/kg, safety, moderate effectiveness, and a low patient dropout rate can be achieved. Retinopathy can be minimized by careful ophthalmologic monitoring. The clinical effect is delayed and good-to-moderate suppression of the disease is observed after 6 – 9 months of therapy. Hydroxychloroquine also is effective in treating systemic lupus erythematosus.

Dapsone is clinically effective in treating rheumatoid arthritis. Decreases in both erythrocyte sedimentation rate and plasma levels of *C*-reactive protein have been observed. Unfortunately, its propensity to induce blood disorders and its carcinogenic potential preclude its widespread use.

The mechanism of the antirheumatic action of antimalarials is still under investigation. Inhibition of various cellular functions of lymphocytes and macrophages has been observed.

For structure, properties, and preparation → Chemotherapeutics.

4.3. Sulfhydryl Compounds

d-Penicillamine [*52-67-5*], 3-mercapto-D-valine, $C_5H_{11}NO_2S$, M_r 149.21, *mp* 198.5 °C.

Synthesis: D-penicillamine can be synthesized in a multistep process that begins with heating isobutyraldehyde, pyridine, sulfur, and ammonia in benzene to form 5,5-dimethyl-2-isopropyl-Δ^3-thiazoline. Treatment with hydrogen cyanide gives 4-cyano-5,5-dimethyl-2-isopropylthiazolidine, which on acid hydrolysis gives D,L-penicillamine hydrochloride [61]. Resolution is accomplished by conversion of the racemate to D,L-3-formyl-2,2,5,5-tetramethylthiazolidine-4-carboxylic acid by treatment first with acetone, then with acetic formic anhydride. The enantiomers are separated in the usual manner, using, for example, L-lysine [62] or D-(−)-threo-1-(4-nitrophenyl)-2-aminopropane-1,3-diol [63]. Acidification liberates D-3-formyl-2,2,5,5-tetramethylthiazolidine-4-carboxylic acid, which is hydrolyzed with hydrochloric acid to yield D-penicillamine hydrochloride. Neutralization with ethanolic triethylamine affords D-penicillamine.

Improvements have also been made in the preparation of D-penicillamine from penicillins by conversion to the penicilloic acid, which when treated with concentrated hydrochloric acid and methanolic mercuric chloride gives D,L-penicillamine as its mercuric thiolate. Treatment with hydrogen sulfide in hydrochloric acid gives D,L-penicillamine hydrochloride [64], which is resolved as described previously.

D-Penicillamine is as effective as gold in bringing about remission in severe rheumatoid arthritis. As with gold it is administered in gradually increasing amounts over a long period of time (4 – 6 months) before beneficial effects are seen. Its mechanism of action is unknown. Speculations concerning the basis for its effectiveness include: copper chelation and mobilization as well as effects on collagen biosynthesis, lymphocyte and monocyte function, antibody response, and cellular immunity. The profound therapeutic benefits of D-penicillamine are offset by a high incidence of adverse reactions. Skin rashes similar to those seen during chrysotherapy constitute the most common adverse effect. Dyspepsia and hypogusia also are encountered. The most serious problem is bone marrow suppression, evidenced by thrombocytopenia and agranulocytosis. Renal toxicity with resultant albuminuria also is observed frequently.

Trade names: Artamin (Sanabo, Austria), Atamir (Sandoz, Denmark), Dimetylcystein (Lilly, Denmark), Metalcaptase (Knoll, South Africa), Reumacillin (Leiras, Finland), Trisorcin (Merckle, Germany), Trolovol (Asta Medica, Germany).

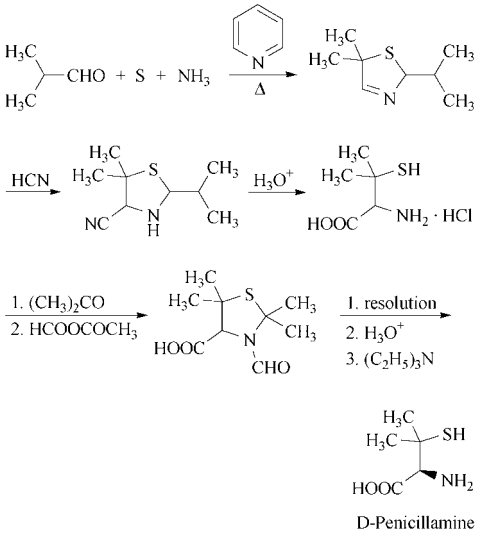

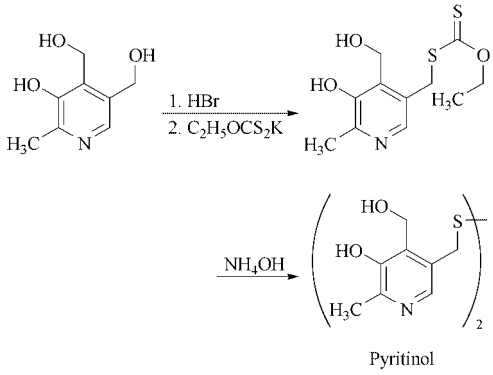

Pyritinol, pyrithioxin [*1098-97-1*], 3,3′-(dithiodimethylene)bis(5-hydroxy-6-methyl-4-pyridinemethanol), $C_{16}H_{20}N_2O_4S_2$, M_r 368.48, *mp* 218 – 220 °C.

Synthesis: treatment of pyridoxine with hydrobromic acid gives 4,5-bis(bromomethyl)-3-hydroxy-2-methylpyridinium bromide, which when treated in the cold with potassium ethyl xanthate gives ethyl 4-hydroxymethyl-3-hydroxy-2-methyl-3-pyridyl methylxanthate. Hydrolysis and oxidation are carried out in aqueous-alcoholic ammonia to give the disulfide pyritinol [65].

Pyritinol, the disulfide of pyridoxine-5-thiol, is marketed as a cerebral stimulant in Europe. Both the thiol and the disulfide have been shown to have D-penicillamine-like activity in the treatment of rheumatoid arthritis. Adverse effects are similar to those seen with D-penicillamine. Occasionally, pyritinol seems effective and better tolerated in some individuals who fail to respond to D-penicillamine.

Trade names: Ansefal (Anka, Turkey), Bonifen (Merck, Portugal), Enbol (Chugai, Japan), Encefabol (Merck, Brazil, Greece, Czech Republic), Encephabol (Merck, Austria, CIS, South Africa), Neuroxin (Yamanouchi, Japan).

4.4. Immunoregulants

The idea that autoimmune mechanisms might be involved in the pathogenesis of rheumatoid arthritis led to the use of immunosuppressive therapy as far back as the 1950s. More recently, its use in severe forms of the disease has been studied carefully.

The folic acid analogue *methotrexate* is a very promising drug for the treatment of rheumatoid arthritis. At doses used for the treatment of rheumatic diseases its principal mode of action probably relates to an inhibition of aminoimidazolecarboxamide ribonucleotide transformylase and thymidylate synthase, plus enhanced adenosine release.

The alkylating agent *cyclophosphamide* is sometimes used in rheumatoid arthritis. A more frequent use is precluded by its considerable toxicity. Cyclophosphamide nonspecifically kills cells by reacting with DNA and RNA molecules.

The immunosuppressive drug *azathioprine* is FDA-approved for treatment of rheumatoid arthritis. Azathioprine is believed to interfere with cell division by inhibiting metabolism and synthesis of proteins and DNA.

Novel therapeutic approaches for treatment of autoimmune rheumatic diseases include The immunosuppressive drug *cyclosporin A* and *leflunomide*. Cyclosporine A is an immunosuppressive drug used to prevent organ rejection which has been shown to provide benefit in rheumatoid arthritis. Most common adverse effects of this drug are renal damage, hypertension, and increased body hair. Leflunomide is an isoxazol derivative that has no struc-

tural relationship to other known immunoregulatory drugs [66], [67]. In vivo leflunomide is rapidly metabolized to its active form, a malononitrilamine termed A771726. The substance may serve as an immunomodulatory agent in rheumatoid arthritis by preferentially causing cell arrest of autoimmune lymphocytes through inhibition of the enzyme dihydroorotate dehydrogenase, which catalyzes a critical step in the production of uridine monophosphate (rUMP). A decrease in rUMP leads to cell cycle arrest in lymphocytes. Furthermore, leflunomide inhibits tyrosine kinase activation in T cells after stimulation of the IL-2 receptor. Tyrosine kinase inhibition has been proposed to partially explain the immunosuppressive action of leflumide.

5. Inhibitors of Tumor Necrosis Factor

Clinical trials suggest that *monoclonal antibodies* directed against TNF provide a novel approach for the treatment of rheumatoid arthritis [68], [69]. The proinflammatory cytokine TNF has been implicated in the pathogenesis of rheumatoid arthritis. TNF concentrations are increased in the synovial fluid of persons with active rheumatoid arthritis, and increased TNF plasma levels have been shown to be associated with joint pain. TNF increases synoviocyte proliferation, and triggers the release of secondary mediators involved in the recruitment of inflammatory cells during neoangiogenesis and in the process of joint destruction. Infliximab is a chimeric TNF monoclonal antibody (CA2) which was the first monoclonal antibody to be investigated for the treatment of rheumatoid arthritis. Etanercept is a recombinant version of the soluble p75 TNF receptor linked to the Fc portion of human immunoglobulin G1. Etanercept competitively inhibits the binding of TNF to its cell surface receptors, thereby inhibiting the biological activity of TNF in a reversible manner.

6. Natural Products

Colchicine [*477-27-0*], (*S*)-*N*-(5, 6, 7, 9-tetrahydro-1, 2, 3, 10-tetramethoxy-9-oxobenzo[*a*]heptalen-7-yl)acetamide, $C_{22}H_{25}NO_6$, M_r 399.43, *mp* 142 – 150 °C.

Production: colchicine is obtained by extracting the seeds and/or corm of *Colchicum autumnale* Linné [70].

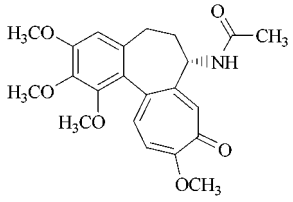

Colchicine

Colchicine is a specific drug for the treatment of gouty arthritis. It is especially useful in acute attacks but also is used prophylactically in conjunction with other drugs, such as allopurinol or phenylbutazone. The drug is believed to act by disrupting microtubular assembly in inflammatory cells, thereby impairing both their locomotion and their membrane function, including phagocytosis and the release of inflammatory mediators. The alkaloid is extremely toxic and must be used with care. Gastrointestinal adverse effects include nausea, vomiting, and diarrhea. Blood disorders, including agranulocytosis and aplastic anemia, sometimes are seen after prolonged administration of this drug.

Trade names: Colchicina (Smith Kline, Brazil), Colchicine (Various, USA), Colchicum-Dispert (Solvay, Poland), Colabid (Major, USA).

Orgotein [*9016-01-7*], M_r 34000, is the name applied to a group of metalloproteins obtained from various tissuesources, such as hemolyzed plasma-free bovine erythrocytes [71] or beef liver [72]. They are copper – zinc/protein chelates with molecular masses of about 34 000. These compounds have potent superoxide dismutase activity. When administered directly into the joint at 4 mg/week for 6 weeks, marked improvement in the mobility of acute rheumatoid patients occurred. Synovial fluid levels of rheumatoid factor and prostaglandin E_2 were reduced.

Trade names: Ontosein (Tedec-Meiji, Spain), Peroxinorm (Grünenthal, Austria), Serosod (Serono, Italy).

7. References

1. H. R. Herschman, "Prostaglandin synthase 2," *Biochim. Biophys. Acta* **1299** (1996) 125 – 140.
2. P. J. Barnes, I. Adcock, "Anti-inflammatory actions of steroids: molecular mechanisms," *Trends Pharmacol. Sci.* **14** (1993) 436 – 441.
3. P. J. Barnes, M. Karin, "Nuclear factor-κB: a pivotal transcription factor in chronic inflammatory diseases," *N. Engl. J. Med.* **336** (1997) 1066 – 1071.
4. W. L. Smith, L. J. Marnett, "Prostaglandin endoperoxide synthase: structure and catalysis," *Biochim. Biophys. Acta* **1083** (1991) 1 – 17.
5. J. A. Mitchell et al., "Selectivity of nonsteroidal antiinflammatory drugs as inhibitors of constitutive and inducible cyclooxygenase," *Proc. Natl. Acad. Sci. USA* **90** (1994) 11693 – 11697.
6. P. E. Lipsky, "Specific COX-2 inhibitors in arthritis, oncology, and beyond: where is the science headed?," *J. Rheumatol.* **26** (1999) Suppl. 56, 25 – 30.
7. C. Luong et al., "Flexibility of the NSAID binding site in the structure of human cyclooxygenase-2," *Nat. Struct. Biol.* **3** (1996) 927 – 933.
8. N. E. Lane, "Pain management in osteoarthritis: the role of COX-2 inhibitors," *J. Rheumatol.* **24** (1997) Suppl. 49, 20 – 24.
9. J. Sirois, M. Dore, "The late induction of prostaglandin G/H synthase-2 in equine preovulatory follicles supports its role as a determinant of the ovulatory process," *Endocrinology* **138** (1997) 4427 – 4434.
10. R. Sawdy et al., "Use of a cyclooxygenase type-2-selective nonsteroidal anti-inflammatory agent to prevent preterm delivery," *Lancet* **350** (1997) 265 – 266.
11. H. Lim et al., "Multiple female reproductive failures in cyclooxygenase 2-deficient mice," *Cell* **91** (1997) 197 – 208.
12. A. Schmassmann et al., "Effects of inhibition of prostaglandin endoperoxide synthase-2 in chronic gastro-intestinal ulcer models in rats," *Br. J. Pharmacol.* **123** (1998) 795 – 804.
13. G. Weissmann, "Prostaglandins as modulators rather than mediators of inflammation," *J. Lipid Mediat.* **6** (1993) 275 – 286.
14. J. R. Vane, "Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs," *Nat. New Biol.* **231** (1971) 232 – 235.
15. S. J. Preston et al., "Comparative analgesic and anti-inflammatory properties of sodium salicylate and acetylsalicylic acid (aspirin) in rheumatoid arthritis," *Br. J. Clin. Pharmacol.* **27** (1989) 607 – 611.
16. K. K. Wu et al., "Aspirin inhibits interleukin 1-induced prostaglandin H synthase expression in cultured endothelial cells," *Proc. Natl. Acad. Sci. USA* **88** (1991) 2384 – 2387.
17. J. A. Mitchell et al., "Sodium salicylate inhibits cyclo-oxygenase-2 activity independently of transcription factor (nuclear factor κB) activation: role of arachidonic acid," *Mol. Pharmacol.* **51** (1997) 907 – 912.
18. B. J. Whittle et al., "Selective inhibition of prostaglandin production in inflammatory exudates and gastric mucosa," *Nature* **284** (1980) 271 – 273.
19. G. A. Higgs, J. A. Salmon, B. Henderson, J. R. Vane, "Pharmacokinetics of aspirin and salicylate in relation to inhibition of arachidonate cyclooxygenase and antiinflammatory activity," *Proc. Natl. Acad. Sci. USA* **84** (1987) 1417 – 1420.
20. E. Kopp, S. Ghosh, "Inhibition of NF-κB by sodium salicylate and aspirin," *Science* **265** (1994) 956 – 959.
21. D. A. Kujubu, H. R. Herschman, "Dexamethasone inhibits mitogen induction of the TIS10 prostaglandin synthase/cyclooxygenase gene," *J. Biol. Chem.* **267** (1992) 7991 – 7994.
22. M. G. O'Sullivan, E. M. Huggins Jr, C. E. McCall, "Lipopolysaccharide-induced expression of prostaglandin H synthase-2 in alveolar macrophages is inhibited by dexamethasone but not by aspirin," *Biochem. Biophys. Res. Commun.* **191** (1993) 1294 – 1300.
23. M. Barrios-Rodiles, K. Keller, A. Belley, K. Chadee, "Nonsteroidal antiinflammatory drugs inhibit cyclooxygenase-2 enzyme activity but not mRNA expression in human macrophages," *Biochem. Biophys. Res. Commun.* **225** (1996) 896 – 900.
24. Merck & Co., US 3527796, 1970 (M. Sletzinger, J. Chemerda, G. Gal).
25. Tropenwerke Dinklage & Co., US 3910952, 1975 (K. Boltze, O. Brendler, H. Dell, H. Jacobis).
26. SIR Laboratorie Chimico Biologici SpA, DE 2223051, 1973 (D. Antonia, F. Ganzina, M. Magi, E. Serino et al.).
27. Instituto Biologico Chemiotherapico "ABC", US 3624103, 1971 (F. DeMartiis,

E. Arrigoni-Martelli, T. Tamietto).
28. Rotta Research Laboratorium SpA, US 3985878, 1976 (F. Makorec, P. Sevin, L. Rovati).
29. Merck & Co., US 3870753, 1975 (R. J. Tull, R. F. Czaja, R. F. Shuman, S. H. Pines).
30. Merck & Co., US 3994600, 1976 (R. J. Tull, R. F. Czaja, R. F. Shuman, S. H. Pines).
31. McNeil Laboratories Inc., FR 1574570, 1969 (J. R. Carson).
32. Rhône-Poulenc SA, US 34244748, 1969 (D. Farge, C. Jeanmart, M. N. Messer).
33. Madan AG, BE 704368, 1968 (N. P. Buu-Hoi, C. Gillet, G. Lambelin).
34. Geigy Chemical Corp., US 3558690, 1971 (A. Sallmann, R. Pfister).
35. P. D. Fowler: "Diclofenac sodium," in E. C. Huskisson (ed.): *Anti-rheumatic Drugs,* Praeger Publishers, New York, 1983, p. 117.
36. John Wyeth & Brother Ltd., US 3476766, 1969 (K. Brown).
37. American Cyanamid Co., DE 2147111, 1972 (A. S. Tomcufcik, R. G. Child, A. E. Sloboda).
38. Boots Pure Drug Co., Ltd., US 3385886, 1968 (J. S. Nicholson, S. S. Adams).
39. Syntex Corp., US 3637767, 1972 (F. S. Alvarez).
40. Rhône-Poulenc SA, US 3641127, 1972 (D. Farge, M. N. Messer, C. Moutonnier).
41. Eli Lilly & Co., US 3600437, 1971 (W. S. Marshall).
42. The Boots Co. Ltd., US 3755427, 1973 (S. S. Adams, J. Bernard, J. S. Nicholson, A. R. Blancafort).
43. Bristol-Myers Co., US 3565943, 1971 (P. F. Juby, R. A. Partyka, T. W. Hudyma).
44. Ciba Corp., US 3641040, 1972 (R. W. J. Carney, G. deStevens).
45. J. R. Geigy AG, US 2562830, 1951 (H. Stenzl).
46. J. R. Geigy AG, US 2745783, 1956 (F. Hafliger).
47. Siegfried AG Zofingen, US 3482024, 1969 (L. Molnar, T. Wagner-Jauregfg, U. Jahn, G. Mixich).
48. E. H. Wiseman, J. G. Lombardino in J. S. Bendra, D. Lednicer (ed.): *Chronicles of Drug Discovery,* **vol. 1,** J. Wiley & Sons, New York 1982, Chapter 8.
49. Pfizer Inc., US 4289879, 1981 (J. G. Lombardino).
50. Parke, Davis & Co., BE 605302, 1961 (R. A. Scherner).
51. J. H. Wilkinson, I. L. Finar, *J. Chem. Soc.* 1948 32 – 35.
52. Parke, Davis & Co., US 3313848, 1967 (R. A. Scherrer, F. W. Short).
53. Laboratories UPSA, US 3415834, 1968 (C. Hoffman, A. Faure).
54. Sandoz, Inc., US 3925548, 1975 (H. Oh).
55. Instituto Farmacologico Serono SpA, US 3557135, 1971 (E. Marchetti).
56. *Drugs of the Future* **22** (1997) 711 – 714.
57. *Drugs of the Future* **23** (1998) 1287 – 1296.
58. P. Lebeau (ed.): *Traite de Pharmacie Chimique II,* 4th ed., **vol. 2,** Masson et Cie, Paris 1956, p. 661.
59. Societe des Usines Chimiques Rhône-Poulenc, US 1994213, 1935 (M. Delepine).
60. SKF, US 3635945, 1972 (P. E. Nemeth, B. M. Sutton).
61. Degussa Normals Roessler, US 4060548, 1977 (F. Ansinger, H. Offermanns, M. Ghyczy).
62. Degussa Normals Roessler, US 3980665, 1976 (R. Fahnenstick, J. Heese, H. Offermanns).
63. Degussa Normals Roessler, US 3980666, 1976 (P. Scherberick).
64. Firma Heyl & Co. Chemisch-Pharmazeutische Fabrik, US 3894067, 1975 (M. Bock).
65. E. Merck AG, US 3010966, 1961 (O. Zima, G. Schorre).
66. R. I. Fox, "Mechanism of action of leflunomide in rheumatoid arthritis," *J. Rheumatol.* **25** (1998) Suppl. 53, 20 – 26.
67. B. Rozman, "Clinical experience with leflunomide in rheumatoid arthritis," *J. Rheumatol.* **25** (1998) Suppl. 53, 27 – 32.
68. L. W. Moreland, "Inhibitors of tumor necrosis factor for rheumatoid arthritis," *J. Rheumatol.* **26** (1999) Suppl. 57, 7 – 15.
69. E. C. Keystone, "The role of tumor necrosis factor antagonism in clinical practice," *J. Rheumatol.* **26** (1999) Suppl. 57, 22 – 28.
70. F. E. Hamerslag, *The Technology and Chemistry of Alkaloids,* D. Van Nostrand Co., Inc., New York 1950, pp. 75 – 78.
71. Diagnostic Data Inc., US 3579495, 1971 (W. Huber).
72. Diagnostic Data Inc., US 3624251, 1971 (W. Huber).

Antimicrobial Agents → Disinfectants
Antimicrobial Agents → Biocides
Antimitotic Agents → Cancer Chemotherapy

Antimony and Antimony Compounds

Sabina C. Grund, SMC-Metallurgical Consultant, Dorsten, Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15)

Kunibert Hanusch, Hüttenwerke Kayser AG, Lünen, Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 and 15)

Hans J. Breunig, Universität Bremen, Institut für Anorganische und Analytische Chemie, Bremen, Germany (Chap. 11.10)

Hans Uwe Wolf, Universität Ulm, Fakultät für Theoretische Medizin, Ulm, Germany (Chap. 16)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Occurrence 3
5. Beneficiation 4
6. Recovery of the Metal 6
6.1. Roasting 7
6.2. Reduction of Oxide to the Metal . . 8
6.3. Direct Reduction to the Metal . . . 10
6.4. Hydrometallurgical and Electrolytic Methods 10
6.5. Recovery from Byproducts 13
7. Refining 14
8. Fine Purification 16
9. Recovery of Antimony Oxide 18
10. Antimony Alloys and Intermetallic Compounds 18
11. Antimony Compounds 18
11.1. Antimony Chlorides 18
11.2. Antimony Fluorides 20
11.3. Antimony Tribromide 20
11.4. Antimony Triiodide 20
11.5. Antimony Oxides 20
11.6. Antimonic Acid and Antimonates . 21
11.7. Antimony Sulfides 21
11.8. Antimony Sulfate 22
11.9. Stibine 22
11.10. Organometallic Compounds 22
12. Chemical Analysis 24
13. Uses 24
14. Recycling 26
15. Economic Aspects 26
16. Toxicology 27
17. References 29

The article describes the properties, production, and usage of the element antimony and its primary compounds. Stibnite, which is referred to as crude antimony when its Sb_2S_3 content is above 90 %, has a low melting point and it can be extracted by melting (liquidation). Antimony metal is recovered from ore primarily by pyrometallurgical techniques. Because rich ores are becoming rare, greater recourse is being taken to intermediates in processing and metal industries. Hydrometallurgical processing is suitable for some ores containing precious metals and is used by the Sunshine Mining Co. (USA). Refining and fine purification techniques are used for production of pure antimony metal.There is growing demand for extremely pure antimony. Antimony oxide is being used as flame retardant in increasing quantities. Antimony is a component of many lead and tin alloys, which are important materials for making bearings and solders. Preparation, properties, and usage of the antimony halides, oxides, sulfides, and other compounds, as well as organometallic compounds are described.Other important aspects are recycling, economics, and toxicology.

1. Introduction [14, 15]

Antimony [*7440-36-0*], Sb, is an element of group 15 having atomic number 51 and atomic mass M_r 121.75. Two isotopes occur in nature, mass numbers 121 [*14265-72-6*] and 123 [*14119-16-5*]. The configuration in the outer electron shell is $5\,s^2 5\,p^3$. If these electrons are

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_055.pub2

donated, antimony is electropositive and trivalent, as in $SbCl_3$, or pentavalent, as in $SbCl_5$; if three electrons are added to the outer shell, the antimony is electronegative and trivalent, as in Li_3Sb.

Usually antimony is a white lustrous metal of average hardness, 3.0 on the Mohs scale. It is brittle and easy to pulverize. If it solidifies slowly, pure antimony has a foliated structure; if rapidly, a granular structure. It forms rhombohedral crystals.

Metallic antimony is the only stable allotrope. There are unstable forms: yellow antimony, black amorphous antimony, and what is known as explosive antimony. Yellow antimony is formed when air or oxygen is passed through liquid stibine. Black amorphous antimony is obtained by rapidly cooling antimony vapors and is also formed from yellow antimony at − 90 °C. At room temperature black antimony slowly reverts to metallic antimony, and at 400 °C reversion is spontaneous. Black amorphous antimony ignites spontaneously in air. Explosive antimony is obtained by electrolysis of antimony(III) chloride solution in hydrochloric acid at a high current density with an antimony anode and a platinum cathode. It consists of black amorphous antimony contaminated with antimony trichloride.

History [14, 16]. Antimony was known to the Chinese 5000 years ago. The ancient Egyptians were also acquainted with it, but they considered it to be a variety of lead, a belief that persisted until the 16th century. The first instructions on processing the Latin *stibium* or the Egyptian *stim* (antimony glance or stibnite) appeared at the beginning of our era. Pedanios Dioskurides and Pliny the Elder referred to antimony glance as platyophthalmon (eye dilator), calliblephary (eyelid cosmetic), and gynaikeios (feminine), because antimony(III) sulfide powder was then an ophthalmic ointment and a coveted cosmetic.

The word *Antimonium* seems to have been first used for antimony(III) sulfide in 1050 by Constantinus Africanus. Light was gradually shed on the chemistry of antimony and on antimony therapy by Paracelsus (1526 – 1541), Andreas Libavius (*Alchemia*, 1597, and the *Syntagmatis Arcanorum* volumes, 1613 – 1615), the Benedictine monk Basilius Valentinus (*Triumphwagen Antimonii*, 1604), N. Lemery (*Traité de l'Antimoine*, 1707), and other works of the iatrochemical epoch. At an early stage, alchemists understood the significance of antimony for separating gold from silver. The recovery of *antimonium crudum* from its ores was described in detail by Lazarus Ercker (1574) and Georgius Agricola (*De re metallica*, 1556).

2. Physical Properties [14, 16, 17]

Density	
solid at 20 °C	6.688 g/cm^3
liquid at 630.5 °C	6.55 g/cm^3
Melting point	630.5 °C
Heat of fusion	10.49 kJ/mol
Boiling point at 101.3 kPa	1325 °C
Tensile strength	10.8 N/mm^2
Modulus of elasticity	566 N/mm^2
Surface tension	
solid at 432 °C	317.2 mN/m
liquid at 630 °C	349 mN/m
liquid at 1200 °C	255 mN/m

The surface tension of the solid decreases nearly linearly with temperature. The slope at the melting point is − 0.07 mN m^{-1} K^{-1}.

Molar heat capacity at 630.5 °C	
solid	30.446 J mol^{-1} K^{-1}
liquid	31.401 J mol^{-1} K^{-1}
Coefficient of linear expansion between	
0 and 100 °C	10.8×10^{-6} °C^{-1}
Thermal conductivity at 0 °C	18.51 W m^{-1} K^{-1}
at 100 °C	16.58 W m^{-1} K^{-1}

Roughly 1/20th of the thermal conductivity of copper. The thermal conductivity of antimony depends on the grain size and the direction in the crystal.

Electrical resistivity at 0 °C	30.0×10^{-6} Ω cm
Molar susceptibility at 20 °C	-99.0×10^{-6}

Polycrystalline antimony is diamagnetic.

3. Chemical Properties [14, 16, 17]

Pure antimony does not change in air at room temperature, and it is not tarnished in humid air or pure water. If antimony is heated to redness in air, the molten metal ignites. Above 750 °C, steam oxidizes liquid antimony to antimony trioxide, and hydrogen is evolved. Antimony can be removed from lead – antimony alloys with steam. Antimony cannot be ignited in a current

of hydrogen. If it is heated to redness in a current of nitrogen, gray vapors are given off, which condense to amorphous antimony.

Neither solid nor liquid antimony dissolves nitrogen. Fluorine, chlorine, bromine, and iodine react violently with antimony, even at room temperature, to form trihalides. The reaction with chlorine yields either $SbCl_3$ or a mixture of $SbCl_3$ and $SbCl_5$. Antimony(III) sulfide is the product of the reaction with sulfur, hydrogen sulfide, or dry sulfur dioxide. In the presence of an oxidizing agent or sulfur, aqueous sodium sulfide dissolves antimony to yield sodium thioantimonite. Yellow ammonium sulfide and sulfur-containing potassium sulfide solutions also readily dissolve antimony.

Molten antimony reacts with phosphorus, arsenic, selenium, and tellurium but not with boron, carbon, and silicon. The eutectic with lead contains 13 wt % antimony and melts at 246 °C.

Antimony is resistant to concentrated hydrofluoric, dilute hydrochloric, and dilute nitric acids. It is readily soluble in a mixture of nitric and tartaric acids and in aqua regia. Phosphoric acid and a few organic acids also dissolve the metal, although acetic acid hardly belongs to this category. At room temperature it is not attacked by dilute or concentrated sulfuric acid. It is attacked at 90 – 95 °C by concentrated sulfuric acid, and sulfur dioxide is evolved.

Pure antimony is resistant to solutions of ammonium and alkali-metal hydroxides and to molten sodium carbonate. If heated to redness, it reacts with molten sodium or potassium hydroxide to form hydrogen gas and antimonites.

In the electrochemical series antimony falls between hydrogen and bismuth: H, Sb, Bi, As, Cu.

4. Occurrence

The average antimony content of the earth's crust has been estimated at $(2 - 5) \times 10^{-5}$ %. Most antimony lodes occur in areas of volcanic activity and frequently in volcanic rocks themselves. They are found predominantly in ancient formations ranging up to the Carboniferous. Antimony deposits are seldom of sedimentary or epigenetic origin.

The most important antimony ore is rhombic *antimony glance* (gray antimony, antimonite, or stibnite), Sb_2S_3. It contains 71.7 % antimony and occurs as characteristic black acicular crystals. Other ores include antimony oxides (valentinite, Sb_2O_3, rhombohedral; senarmontite, Sb_2O_3, cubic; cervantite, Sb_2O_4, orthorhombic), antimony hydroxides (stibiconite, $Sb_2O_4 \cdot H_2O$), antimony oxide sulfides (kermesite, pyrostibite, red antimony, 2 $Sb_2S_3 \cdot Sb_2O_3$), native antimony, and double sulfides (jamesonite, $Sb_2S_3 \cdot 2$ PbS, which may contain silver; tetrahedrite, $Sb_2S_3 \cdot 4$ Cu_2S_2; and livingstonite, 2 $Sb_2S_3 \cdot HgS$). More than 100 minerals containing antimony are known: there are eight antimonides, one sulfide (stibnite), 59 mixed antimonides and sulfides, and 46 oxide minerals [14, 18].

Simple antimony deposits consist mainly of stibnite or, rarely, native antimony in a siliceous gangue, commonly with some pyrite, and in places a little gold and small amounts of other metal sulfides, principally those of silver and mercury. Most of these geologically simple stibnite deposits do not individually contain more than several thousand tonnes of ore. Deposits of this type are found in the world's most productive antimony districts, including those of Bolivia, China, Mexico, and the Republic of South Africa [18, 19].

Complex antimony deposits consist of stibnite associated with pyrite, arsenopyrite, cinnabar, or scheelite, or of antimony sulfo salts with varying amounts of copper, lead, and silver, as well as common sulfides of these metals and zinc. Ores of the complex deposits are mined primarily for gold, lead, silver, tungsten, or zinc; antimony is recovered as byproduct. The tetrahedrite ores found in the Coer d'Alene district of Idaho are a typical example of a complex silver – copper ore which is mined primarily for its silver and copper content and from which antimony is recovered in a concentrate as a byproduct [18, 20].

Principal identified world antimony resources are estimated to be about $(3 - 5) \times 10^6$ t Sb. About 50 % of total world resources are located in China, mainly in Hunan province in the south. Further major antimony resources of about 200 000 to 300 000 t Sb content are located in Bolivia, South Africa, Mexico, and the former Soviet Union (Central Asia, the Ural

and Caucasus mountains). Deposits of about 100 000 t are located in Australia, Italy, former Yugoslavia, Malaysia, Thailand, Turkey, and the United States, while small deposits are found in Burma, France, Canada, Morocco, Austria, Peru, Spain, and Zimbabwe [18, 21].

Table 1 lists world mine production of antimony. Total mine production increased from 1993 (86 445 t) to 2001 (15 7000 t) and decreased again until 2003 (81 600 t).

Since 1987 China has been the dominant producer of antimony ores and concentrates. Its strong market position with a share of about 60 – 70 % of world mine production has remained almost constant. Other countries with major productions of primary antimony raw materials are Bolivia (7.1 %), the countries of the former Soviet Union (6.8 %) and South Africa (4.1 %).

The largest antimony mine in the world is the Murchinson Range, which lies in the Lowveld of the north-eastern Transvaal. Many types of minerals are present in the Murchinson belt, including antimony, gold, mercury, copper, zinc, tungsten, titaniferous magnetite, and emeralds. Only three of these commodities — antimony (with gold as a byproduct), gold (ex dumps), and emeralds — are currently exploited. The antimony deposit consists predominantly of stibnite (Sb_2S_3), which is mined down to 60 m underground [19].

In contrast to the large-scale mining in South Africa, most mines in Bolivia are far smaller, the largest being Chilcobija (850 000 t), Caracota (120 000 t) owned by Empressa Unificada SA (EMUSA), and San Luis (35 000 t). In Bolivia antimony ore is also mined down to 50 m below ground [25]. Other small antimony deposits which are mined economically are found in Turkey, where the 400 t Sb mine production is divided between many small ore bodies [26].

Other sources of antimony besides ores are intermediates in the processing of lead and copper ores: flue dusts, speiss, slags from lead refining processes, anode slimes and the electrolyte from the electrolytic copper refineries (Chapter 6.5). Used car batteries are another valuable resource for the recovery of antimony (Chapter 14).

5. Beneficiation

The beneficiation process used depends chiefly on the content of antimony and other substances in the ore. Another important factor is how the ores are dispersed in the rock. If an antimony ore is not closely interlaced with the gangue, hand-picking techniques or jigs suffice for separation [14].

If stibnite (Sb_2S_3) ore contains more than 90 % stibnite, it can be sold directly as Sb_2S_3 (crude antimony). At some mines in Bolivia and in the Republic of South Africa the high-grade sulfide ore is treated by hand-cobbing and sold as lump sulfide ore with 60 % Sb content. Stibnite has a low melting point and it can be extracted by melting if the ore contains 45 – 60 % antimony and is free from lead and arsenic. This technique is called liquidation. If the antimony content lies between 5 and 25 %, the antimony is concentrated in a flue dust by volatilizing roasting (see Chapter 6.1). Antimony ores may contain auriferous antimony sulfides and arsenopyrites. Beneficiating such ores may become profitable again because of their precious metal content [14, 20].

As a rule, low-grade and complex ores or ores in which the minerals are finely dispersed throughout the gangue are beneficiated by flotation, particularly if they contain precious metals [27–30].

Flotation. Figure 1 shows the flow sheet for the recovery of antimony and gold at Murchinson Consolidated in the Republic of South Africa. An integral part of this process are the different flotation operations, which produce a stibnite – gold concentrate as an intermediate product. Different flotation techniques have been in use at Consolidated Murchinson since its formation in 1934, each change in the process being due to variations in the grade of the mined ore and the minor elements accompanying antimony and gold.

Today the ore received by road from the operational shafts has a maximum particle size of approximately 0.4 m and a typical analysis of 1.8 % Sb, 0.2 % As, and 2 g/t Au. In the preliminary comminution circuit the final grind achieved is about 65 – 75 % passing 75 µm. The flotation circuit comprises four parallel rougher banks, four parallel scavenger banks, two cleaner banks, and a recleaner bank. The concentrate produced contains 58.0 % Sb, 25 –

Table 1. World mine production by country (in t Sb content) [20, 22–24]

	1993	1995	1997	1999	2001	2003
Morocco	168	170	160	250		
South Africa	4485	5537	5000	5278	4927	5310
Zimbabwe	62	37	5			
Bolivia	5556	6426	8700	2790	2264	2300
Canada	673	684	673	357	243	143
Guatemala	90	665	880			
Mexico	1494	1783	1909	126		
United States	266	262	356	450		
China	60 000	69 395*	95 000*	89 600	140 000	70 000
Thailand	620	230	60	59	40	40
Turkey	111	416	400	180	370	350
Australia	1700	900	1900	1679	1380	1300
Russia	7120	6000	6000	4000	4500	
Kirghizia	2500	1500	1200			
Tadzhikistan	1200	1000	1200	1800	2500	1800
World total	86 445	95 395	123 443	107 000	157 000	81 600

* Mine production in China varies from 60 000 to 120 000 t/a depending on the literature source.
** France, Austria, Italy, Spain, Czech Republic, Slovakia, Romania, former Yugoslavia, Romania.

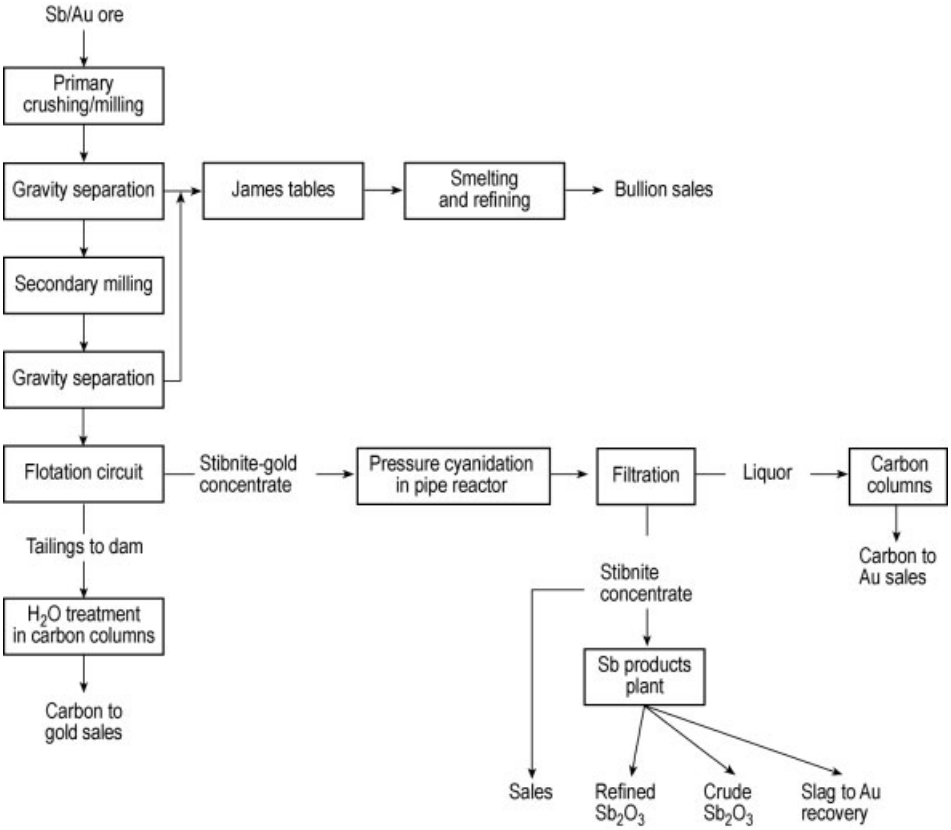

Figure 1. Schematic of Consolidated Murchinson's processes for the recovery of gold and antimony [19]

40 g/t Au, 0.17 % As, 0.18 % Pb, 0.11 % Ni, 0.17 % Cu, 1.9 % Fe, 22.3 % S, 9.9 % insoluble matter, 3.0 ppm Se, and 10 ppm Hg. The recovery of antimony of this grade varies between 80 and 88 % depending on the head grade and the type of mineralization. The stibnite – gold concentrate is leached by pressure cyanidation in a pipe reactor to separate the gold from the antimony. Crude and refined antimony oxide is recovered from the final antimony concentrate [19].

In the flotation circuit lead nitrate is used as an activator for antimony flotation, with sodium isobutyl xanthate as a primary collector. Aero promoter 208, a dithiophosphate, acts as a secondary collector for antimony and a promoter for gold flotation. The dithiophosphate has an alcohol base and doubles as a frother, although small amounts of other alcohol-based frothers are used occasionally. Dextrin is used as a gangue depressant. The ore is naturally alkaline owing to the presence of dolomite as a major gangue constituent. Sulfuric acid is therefore used to control the pH: 6.8 – 7.0 in the roughers, 6.5 – 6.8 in the scavengers, and 6 – 6.5 in the cleaners. Additionally, sodium cyanide is used to depress the flotation of arsenic minerals [19, 31].

Liquation. Ores with more than 40 – 60 % Sb_2S_3 and free from lead and arsenic often are beneficiated by heating. Because stibnite, Sb_2S_3, has an unusually low melting point, 546 – 548 °C, it can be separated from the gangue by melting (liquation). Stibnite more than 90 % pure is obtained and is referred to as *crude antimony*. It can be sold as such for producing antimony compounds, or it can be converted to antimony metal.

The commercial product should have a striated crystalline structure and a metallic gloss, and its color should be that of graphite. These properties are attained only if the molten antimony(III) sulfide is cooled very slowly. Ores with less than 40 – 45 % Sb_2S_3 are unsuitable for liquation because the large amount of gangue interferes and too much antimony is retained in the residue.

The ore is best fed in the form of walnut-sized lumps (1 – 4 cm). If the ore is too fine, the molten Sb_2S_3 does not flow. The most suitable temperature range is 550 – 660 °C because the volatilization of the Sb_2S_3 (*bp* 1000 °C) is prevented, and liquation is successful. Liquation is carried out in a reducing atmosphere in crucibles or in reverberatory furnaces. The main impurities in crude antimony are arsenic, lead, and iron sulfide.

6. Recovery of the Metal

Antimony metal is recovered from ore primarily by pyrometallurgical techniques. Either antimony(III) sulfide is converted into the oxide, which is then reduced, or the ore is partially roasted and allowed to react with sulfide to form the metal and sulfur dioxide.

Sulfide ores with antimony contents between 5 and 25 % are roasted to give volatile Sb_2O_3, which is reduced directly to the metal. In many smelters mixed oxide – sulfide ores are processed in water-jacketed furnaces together with recycled material and byproducts containing antimony. Reverberatory furnaces are used mostly for reducing rich oxide materials. The most suitable processes are listed in Table 2 for the various oxide and sulfide ores.

Table 2. Suitable processes for various antimony ores or concentrates

Nature of the ore	Antimony content, %	Process
Sulfide	< 20	volatilizing roasting
Sulfide	< 35	smelting in shaft furnaces
Sulfide	> 60	liquation and cyclone volatilizing
Oxide	< 30	smelting in reverberatory furnaces and shaft furnaces
Oxide	< 50	direct reduction
Mixed sulfide and oxide	–	smelting in shaft furnaces

Special processes are used if the ore is rich in precious metals. Hydrometallurgical techniques can be used to minimize losses of precious metals. For example, crude metal is obtained at the cathode in a hydrometallurgical process used by the Sunshine Mining Co. (Cœur d'Alène district of Idaho, USA). Because rich ores are becoming rare, greater recourse is being taken to intermediates in processing and metal industries. Also, more and more complex ores are being processed.

6.1. Roasting

The sulfide can be oxidized to antimony(III) oxide (volatilizing roasting) or to antimony(IV) oxide (nonvolatilizing or dead roasting). Furnace control in the production of the nonvolatile antimony(IV) oxide is comparatively simple, but the oxide is not separated from the residue. If the ore contains a high proportion of gangue, difficulties are encountered in reduction. As a result this process is used rarely today.

Volatilizing roasting has the advantage that volatile Sb_2O_3 is produced selectively in a 98 % yield and is separated from the residue, which contains the precious metals in a recoverable form. However, temperature and draft control in the furnace is difficult. The oxide forms between 290 and 340 °C in an oxidizing atmosphere, and the rate of reaction attains its maximum at 500 °C, where antimony(IV) oxide begins to form. The basic reaction is

$$2\,Sb_2S_3 + 9\,O_2 \rightarrow 2\,Sb_2O_3 + 6\,SO_2$$

If too much oxygen is available, antimony(IV) oxide is formed:

$$Sb_2S_3 + 5\,O_2 \rightarrow Sb_2O_4 + 3\,SO_2$$

Above 560 °C the rate of reaction drops considerably. During roasting, antimony(IV) oxide can react with antimony(III) sulfide to give antimony(III) oxide:

$$Sb_2S_3 + 9\,Sb_2O_4 \rightarrow 10\,Sb_2O_3 + 3\,SO_2$$

Therefore, the process must be engineered so that Sb_2O_3 is formed rapidly and preferentially. The temperature must be high enough to ensure adequate volatilization, and the oxygen supply must be kept low to inhibit the formation of antimony(IV) oxide. However, if the temperature is too high, part of the charge will melt, envelope the sulfide grains, and prevent their oxidation. During condensation of antimony(III) oxide, the oxygen content in the gas phase must be kept low enough to prevent oxidation of antimony(IV) oxide.

The temperature level is governed by the sulfide content of the ore. Low-grade ores can be roasted between 850 and 1000 °C. If rich ores are roasted, the upper limit is the melting point of the sulfide (546 °C). In practice, the temperature should not exceed 400 °C. The oxygen supply is controlled by mixing wood charcoal or coke breeze with the charge and by admitting only the amount of air required to form carbon monoxide and antimony(III) oxide. The carbon monoxide suppresses oxidation to antimony(IV) oxide. Despite everything, however, formation of antimony(IV) oxide cannot be entirely suppressed. Antimony(IV) oxide, lead(II) oxide, and flue dust are obtained in the first condenser. The antimony(III) oxide condenses in the second, and the very volatile arsenic(III) oxide condenses in the last. If the oxygen concentration in the roaster is too low, partial oxidation can occur and give a mixture of oxide and sulfide melting at 485 °C. If the concentration of oxygen is too high, arsenates and antimonates of lead, copper, and other metals can form and pass into the slag.

A grain size of 5 – 10 cm suffices in conventional processes because fresh surface is always exposed for reaction owing to volatilization of the trioxide. Very low grade ores must be ground more finely to separate the sulfide from the gangue. The yield depends on the process and the sulfide content of the ore, varying between 60 and 90 % or more.

Commercial antimony(III) oxide should be at least 99.5 % Sb_2O_3. If it is finely crystalline, adheres firmly to the fingers, and does not cake, it is considered of good quality. The oxide must be white; a reddish tinge indicates the presence of antimony(III) sulfide. The yellowish shade of selenium and lead(II) oxide is also undesirable. The arsenic content should be on the order of 0.1 %. Typical specifications and grain size distribution [32] are listed below.

Assay	99.5 %
Sb_2O_4	0.209 %
As_2O_3	0.11 %
PbO	0.15 %
Fe_2O_3	0.03 %
SnO	0.001 %
S	0.18 %
Solubility in tartaric acid	> 99.5 %
Solubility in water	< 0.4 %
Whiteness measured by reflectometry	> 90 %
Size distribution	
0 – 1.5 µm	60 %
1.5 – 3 µm	38 %
> 3 µm	2 %
> 44 µm	0.1 %
Apparent density	0.8 kg/L

Commercial refined antimony(III) oxide can be produced by a second roasting process in which the crude oxide is roasted again under controlled conditions. Especially the separation of arsenic and antimony can be an important factor for the production of high-quality antimony oxide. Whilst arsenic oxide has little value, there is a market for antimony oxide [19, 33–37].

Before *mercury-containing antimony ores* are processed, they generally must be roasted below 400 °C to volatilize the mercury. The temperature must be controlled accurately by regulating the admission of air to avoid overheating the charge and thus volatilizing antimony compounds. The roasted ore still contains antimony sulfide. The ore is removed from the furnace and cooled slowly under exclusion of air in special bunkers. The antimony is recovered by reduction or iron precipitation in open-hearth furnaces [38–40].

The *roasting furnaces* most commonly encountered today are rotary kilns. If the throughput is high, antimony(III) oxide yields of 95 – 98 % are attainable. The yield does not depend on the nature of the ore. The antimony content of sulfide ores does not affect the kiln's efficiency; therefore, sorting and flotation of low-grade sulfide ores are unnecessary. Rotary kilns are also suitable for oxide – sulfide ores, reducing the antimony initially to the metal. The metal volatilizes and is oxidized to antimony(III) oxide that is free of nonvolatile impurities and the off-gases produced on reduction. Some rotary kilns are heated not by burners but by hot air blown into them. The gas flow rate is significantly less than that in the old burner-heated furnaces, and less gangue is entrained in the gas stream. Nevertheless, hot-air heating does not entirely avoid temperature peaks, with melting of the charge.

Crude antimony oxide (81 % Sb, 0.3 % As, 0.15 % Ni) is produced by Consolidated Murchinson in the Republic of South Africa by a volatilizing roast of the sulfide concentrate in a rotary kiln. Temperatures in the kiln are in the range 1100 – 1200 °C so that the oxidation of Sb_2S_3 to Sb_2S_3 occurs partly in the gas phase. The conversion of stibnite to the oxide is carried out by the Chemetron Process, which has unique features embodied in the oxidation and cooling control. While the antimony oxide is removed from the kiln continuously, the nonvolatile elements remain in the kiln and are removed in weekly deslagging campaigns. Kiln-feed rates, burner positioning, and cooling rates are important process control parameters that affect product grade and recovery. Refined antimony trioxide of 99.5 % purity is produced by refuming of the crude oxide under rigidly controlled conditions [19, 37].

Cyclone smelting entirely eliminates melting of the charge. The antimony smelter at Vinto, Bolivia, was the first in the Western Hemisphere to use the cyclone furnace smelting process. Conventional antimony roasting consumes large amounts of coal, and the lack of domestic coal has favored the cyclone furnace. The feed contains at least 60 % antimony and less than 0.5 % lead plus arsenic. Antimony is volatilized as antimony sulfide, burned, and separated in a baghouse as antimony(III) oxide. The impure oxide is pelletized and melted in a rotary furnace. The cyclone slag contains about 1 % antimony [41]. (See also Section 6.2 and Fig. 2.)

6.2. Reduction of Oxide to the Metal

The oxides from volatilizing roasting, as well as other oxidized antimony materials, are reduced to the metal with carbonaceous materials. Occasionally oxide ores containing more than 50 % antimony are finely ground and also reduced. Sometimes the oxide must be treated beforehand to remove arsenic, but improved roasting techniques frequently make this unnecessary. The amount of carbon required for reduction depends on the composition of the oxide, ranging between 8 and 12 %. Pulverized charcoal, anthracite dust, or lean coke breeze are used [14].

$$Sb_2S_3 + 3\,C \rightarrow 2\,Sb + 3\,CO$$

$$Sb_2S_3 + 3\,CO \rightarrow 2\,Sb + 3\,CO_2$$

Different slag systems are in use. In the combined blast furnace/reverberatory smelting technology, as traditionally operated in China, a 40:30:20 $FeO:CaO:SiO_2$ slag is used. To prevent the loss of volatile Sb_2O_3, a sodium salt, such as soda (Na_2CO_3) or Glauber's salt (Na_2SO_4), is added to give a slag of low viscosity. Impurities such as copper and iron are converted to matte by adding sulfur, usually in the form of crude antimony. The slags can contain 5 – 20 % antimony, which must be recovered.

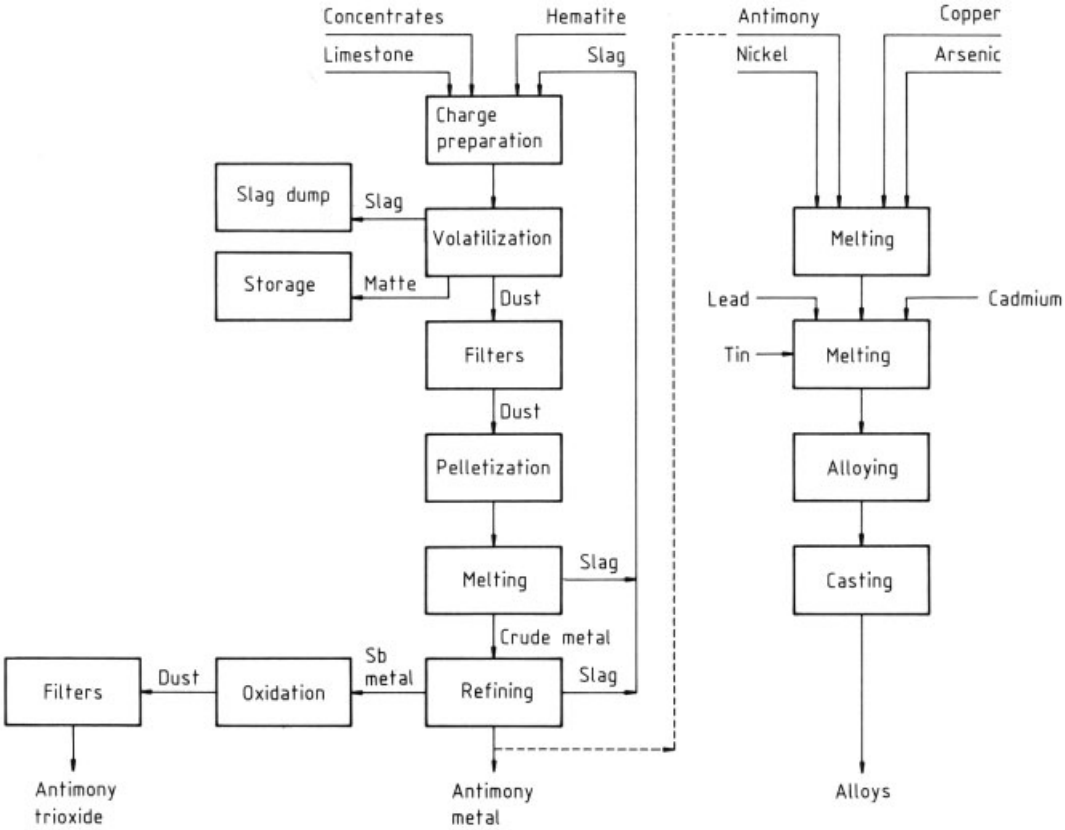

Figure 2. Antimony flowsheet for the smelter at Vinto, Bolivia [41]

Smelting of antimony oxide flue dust containing arsenic, such as is obtained by roasting ores, with sodium hydroxide produces metallic antimony.

$$2\,Sb_2O_3 + 3\,As_2O_3 + 18\,NaOH \rightarrow 4\,Sb + 6\,Na_3AsO_4 + 9\,H_2O$$

The arsenic dissolves in the alkaline slag [42–44].

The oxides are reduced in shaft, reverberatory, or short rotary furnaces. All types of furnaces require efficient precipitators and off-gas filters to remove volatilized antimony(III) oxide.

The pronounced reducing action in *shaft furnaces* does not produce pure metal but does yield slags containing a maximum of only 2.5 % antimony and usually less than 1 %. These slags can be used in glass manufacture. The furnaces have large height to diameter ratios, e.g., 6 : 1. The charge must be in the form of lumps or briquettes. The yield is between 95 and 98 %; the impure product contains 90 – 92 % antimony, 5 – 7 % iron, and arsenic and lead.

Reverberatory furnaces produce metal of comparatively high purity, but the slag is rich in antimony and requires further treatment. These furnaces are fed with high-grade ores that yield fairly small quantities of slag and with antimony-laden flue dust.

Stationary reverberatory furnaces have been largely replaced by *short rotary furnaces* in which the charge is very thoroughly mixed and rapidly melted. The rotary movement efficiently transfers heat to the charge. Tapholes are arranged at various heights in the furnace to allow molten metal, liquefied slag, and solid residues to be withdrawn. The great advantage of this process is that reduction in the furnace can be controlled so that the accompanying metals, such as lead, iron, and arsenic, are not reduced or are volatilized. This makes the refining simpler.

At the antimony smelter at Vinto, Bolivia, the oxide melt passes from the rotary furnace to reverberatory furnaces for reduction and final refining at 800 °C. Coal and sodium hydroxide are added. The final product contains 99.6 % antimony. If less pure, it can be alloyed with copper, nickel, or lead. The total antimony recovery rate for the smelter is 92 % [41]. The complete flowsheet is shown in Figure 2 [14, 45–49].

Mining, beneficiation, and metallurgical procedures have changed little in the past 35 years. No new technological developments that would affect productivity are foreseen for the future [50, 51]. Apart from some work on the thermodynamics of the slag compositions used in antimony smelting, most of the publications concerning smelting technology date from around 1950 [45–49, 52–57].

6.3. Direct Reduction to the Metal

In blast furnace smelting the charge is antimony(III) sulfide, either alone or mixed with antimony oxides. The molten sulfide is first oxidized by the air blast. The oxide formed reacts with unconverted sulfide to yield metallic antimony.

$$2\,Sb_2S_3 + 9\,O_2 \rightarrow 2\,Sb_2O_3 + 6\,SO_2$$

$$2\,Sb_2O_3 + Sb_2S_3 \rightarrow 6\,Sb + 3\,SO_2$$

$$3\,Sb_2S_3 + 9\,O_2 \rightarrow 6\,Sb + 9\,SO_2$$

In view of the many oxidation states through which the antimony must pass, this process is undoubtedly more complex than the equivalent lead and copper processes. It has been successfully practiced for many years in Mexico, Belgium, Great Britain, and Czechoslovakia. Best results are obtained with sulfide ores containing 25 – 30 % antimony. The main disadvantage of the direct process is that antimony compounds are volatilized; this has been offset by using high, slender smelting columns.

Standard practice favors low air pressures, e.g., 5.5 kPa at a flow rate of 1200 m^3/s. The antimony content of the mixed charge should be less than 20 %. Suitable fluxes are limestone, pyrite cinders, and ferrous slags. The required proportion of coke is about 15 %. The slag and the metal usually are separated in a heated forehearth.

The yield of antimony can be as high as 98 %, but not all of this is metal. Some leaves the shaft furnace as oxide, which must be trapped in filters. The metal is contaminated by iron, lead, and some copper and arsenic and must be refined (see Chap. 7) [19, 58].

6.4. Hydrometallurgical and Electrolytic Methods

Hydrometallurgical processing is suitable for some ores, particularly those containing gold, silver, and copper, and especially tetrahedrite, $Sb_2S_3 \cdot Cu_2S$ [59–62]. The high antimony content makes extracting the precious metals with cyanide difficult.

A number of methods have been proposed; however, most are plagued by unresolved problems. If sulfidic antimony ores are digested with *acids*, hydrogen sulfide is given off. Sulfuric acid does not digest the ores completely at 300 °C. Although hydrochloric acid is a better solvent and the antimony(III) chloride is adequately soluble, the problems involved in further processing have not been resolved. Cementation with iron, precipitation as the sulfide, and electrolysis have been investigated as means of separating the antimony. Explosive antimony is likely to form if electrolysis is used on antimony chlorides.

Antimony ores can be digested by alkalies. The mechanochemical leaching of antimony from Slovakian tetrahedrite concentrate in an attritor and the chemical leaching of the resulting solid residues were studied. The results indicate that recoveries of 52 – 100 % Sb, 60 – 85 % Hg and 50 – 96 % Ag can be achieved by using alkaline solutions of Na_2S for leaching of Sb and Hg. The dissolution of Sb from a Columbian stibnite concentrate in basic solutions of $Na_2S \cdot 9\,H_2O$ and NaOH gave a maximum Sb extraction of 94 % from a concentrate containing 30 % Sb with a mixture of 12.5 % Na_2S and 4.0 % NaOH [63].

The antimony can be reduced from the solutions electrolytically or by cementation on aluminum or zinc. Very pure rhombohedral antimony is obtained by electrolysis.

Any raw materials that are soluble in sodium sulfide, e.g., antimony glance or foundry prod-

ucts containing antimony sulfide, are suitable starting materials for antimony recovery by electrolysis. The solution is purified and then electrolyzed in diaphragm cells with iron or lead anodes and iron cathodes. The purity of the antimony is 96 – 99 %.

The source of all the valuable metals in the Sunshine Mine ore is an argentiferous tetrahedrite (a silver – copper – antimony sulfide), which cannot be broken down physically into its constituents. The tetrahedrite occurs reasonably free, and about 85 % can be floated into a high-grade concentrate after grinding to a relatively coarse size.

According to a publication from 1944 [64], the high antimony content precludes adoption of large-scale copper smelting for the concentrate, and lead smelting is just as unsuccessful because of the high antimony and copper contents and the presence of bismuth. Instead, the antimony is extracted from the argentiferous tetrahedrite by an alkaline sulfide solution. The residue consists of insoluble silver and copper sulfides. The antimony is recovered from the alkaline solution by electrolysis.

The steps are as follows:

1) The antimony is batch leached with hot concentrated sodium sulfide solution from milled tetrahedrite concentrate to give a solution containing about 250 g Na_2S and 50 g Sb per liter. The leaching time is 8 – 10 h at 100 – 103 °C.
2) After the residue settles, the supernatant solution is decanted. The residue is recovered by two-stage filtration with a repulping step between the stages. The pregnant solution is fed to the electrolytic section through two anolyte and two catholyte tanks.
3) The solution is batch electrolyzed in banks of diaphragm cells. The antimony is deposited at mild-steel cathodes. The electrolytic section consists of 96 cells in series, arranged in cascades of six.

During electrolysis, the antimony is deposited at the cathodes, and sulfur is concentrated in the anolyte, with the result that sodium polysulfide (Na_2S_3), sodium sulfite (Na_2SO_3), sodium thiosulfate ($Na_2S_2O_3$), and sodium sulfate (Na_2SO_4) form. The alkalinity of the anolyte decreases, whereas that of the catholyte increases.

Fresh anolyte is made up from a mixture of caustic soda solution and barren catholyte to be equivalent to 225 g NaOH per liter. It is discharged into the anolyte feed tanks in 7.5-t batches, and the same volume, having an alkalinity of about 175 g/L, is withdrawn to the regeneration section.

The fresh catholyte, also handled in batches, carries about 50 g of antimony and the equivalent of 250 g of NaOH per liter.

Catholyte is introduced near the bottom of each cell and withdrawn at the top. Because antimony depletion lowers the specific gravity, the weakest solution is drawn off and fed to the following cascade. The solution leaving the final cell in a cascade is pumped back to the catholyte feed tank, unless its antimony content has been reduced to the barren solution level of 8 – 10 g/L. The barren solution has an alkalinity equivalent of ≈ 280 g NaOH per liter. About 30 t of barren solution is withdrawn during each cycle, and the same amount of fresh pregnant solution is added to the feed tank.

The busbar current is 1500 A, producing a current density of 300 A/m^2. There is no gassing at the cathodes until the antimony content of the solution falls below 8 g/L. The voltage drop varies slowly between 2.5 and 3 V per cell, depending on the alkalinity of the anolyte, averaging about 2.65 V. The cathodes are stripped every three days.

The spent solutions from both the catholyte and anolyte cell banks are led to the regeneration circuit for treatment with barium sulfide solution. After the precipitated barium salts have been separated, the solutions are returned to the leaching circuit. The precipitated barium salts are mixed with coal and passed through a rotary kiln for reduction to a black ash containing barium sulfide.

The process is cyclic and regenerative, except that some barite must be supplied, and some caustic soda must be added to the anolyte to replace sodium lost with the copper residue.

The antimony is sold as unrefined granular cathode antimony containing 96 – 97 % antimony and 2 – 3 % arsenic. Copper, lead, iron, bismuth, and other elements make up less than 0.4 %. The small remainder is probably sodium sulfate. The flowsheet for the Sunshine Mining Co. electrolytic process is shown in Figure 3 [64].

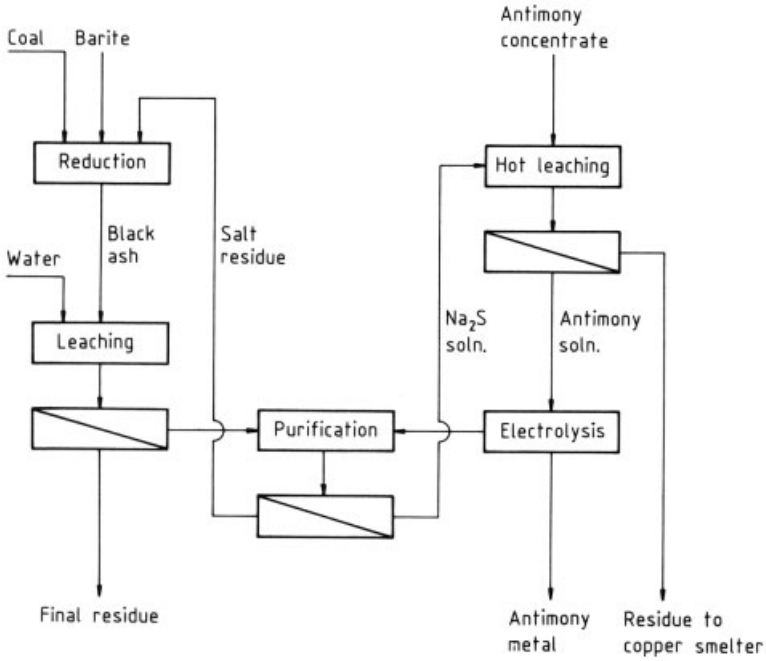

Figure 3. Flowsheet for the Sunshine electrolytic antimony plant

In a method devised by the U.S. Bureau of Mines [65], the concentrate containing precious metals is smelted together with some iron. The metal obtained is deposited electrolytically at the cathode and refined. As much as 80 – 90 % of the antimony is recovered as metal; 80 – 94 % of the gold and 30 – 48 % of the silver pass into the anode mud (Fig. 4).

In another process the ore is smelted with soda and tetrahedrite and leached with lime. The antimony passes into solution as sodium thioantimonate, Na_3SbS_4. The arsenic is precipitated. If the ratio of sodium to sulfur in the electrolyte is 2 : 1 and the current density on the order of 1000 A/m^2, no diaphragms are needed; the current efficiency is 60 % [64–66].

Solvent extraction methods can be utilized for the selective recovery of antimony from solutions. A suitable application could be the recovery of antimony from solutions with a high copper content such as the electrolytes in electrolytical copper refining plants (see Chapter 6.5) [67].

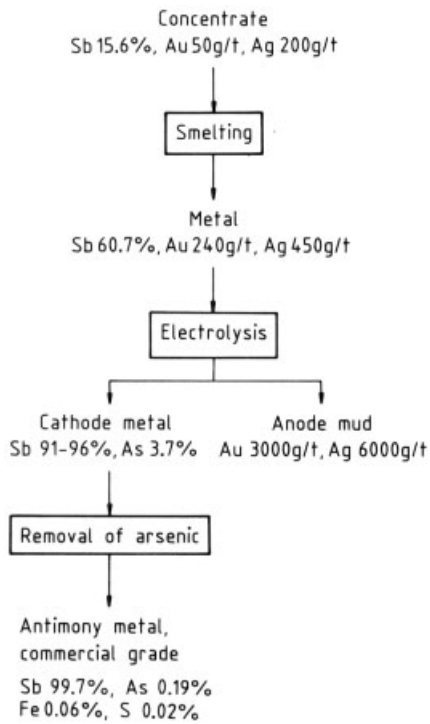

Figure 4. Beneficiation of a concentrate containing precious metals [65]

6.5. Recovery from Byproducts

Lead and copper ores often contain antimony which is concentrated in byproducts and removed during the recovery of the main metal. Different treatments of the byproducts for the recovery of antimony are reported.

In the pyrometallurgical processing of *antimony-containing lead ores* about 90 % of the antimony is recovered in the lead. During refining, the antimony is oxidized, usually in reverberatory furnaces, into an oxide slag, called antimony skimmings. After the entrained lead has been liquated, skimmings are smelted together with coal and fluxes, frequently in short rotary kilns, to yield a crude antimonial lead.

In the Harris refining process, lead bullion is treated at 400 °C with a melt of sodium chloride, sodium monoxide, and sodium nitrate. Arsenic, tin, and antimony are converted into slags. The antimony-rich slag from the last Harris refining step contains the antimony as sodium antimonate. This caustic slag is treated hydrometallurgically with a weak NaOH solution. While the residue contains about 45 – 50 % Sb and can be sold as white sodium antimonate for the production of antimony metal or compounds, the filtrate is evaporated and the concentrated soda recycled to the Harris process [68].

Particularly in the case of bismuth-rich lead, lead bullion is cast into anodes and refined by an electrolytical process. The antimony is retained in the anode slime, together with the other undissolved metals: copper, arsenic, bismuth, and precious metals. A typical anode slime from a lead refining electrolysis contains 6 % Ag, 30 % Sb, 15 % Bi, 12 % Pb, and some Cu [69]. The anode slime is smelted under reducing conditions to give a lead alloy. The antimony is then removed by selective oxidation and subsequently recovered from the flue dust by another reducing smelting operation, which can be followed by oxidation to produce refined antimony oxide [69–72]. A flow sheet for the treatment of lead anode slime is given in Figure 5.

About 95 % of the antimony in lead ores is recovered in the form of antimony alloys, which are usually diluted with refined lead and sold as is [73]. These alloys are used mainly in lead – acid storage batteries.

The recovery of copper from antimony-containing sulfidic copper ores includes the following process steps: copper matte smelting, matte converting, fire refining, anode casting, and electrolytic refining. During the pyrometallurgical operations some of the harmful impurities of copper smelting such as antimony, bismuth, arsenic, lead, and others are removed in slag or as flue dust. Of these impurities antimony is the most difficult to remove [74].

The reported distributions of antimony into the different product phases of the pyrometallurgical unit operations vary widely. In the matte smelting process in the Outokumpu flash smelting furnace, 26 % of the antimony reports to the matte, 45 % to the slag, 13 % to the flue dust, which is recycled to the furnace, and 16 % to the flue dust bleed off [69]. Depending on the converting technology, most of the antimony from the charged copper matte reports to the slag and the blister copper, while only a smaller fraction is found in the flue dust [69, 75–77].

Antimony in copper smelter flue dust can be recovered during dearsenification or kept in the metallurgical circuit by recycling the flue dust to the furnaces for further enrichment. The particular composition determines how these flue dusts are processed; there are no standard procedures [14]. Slags from the converting section are usually recycled to the matte smelting furnace. The direct recovery of valuable metals from converting slags by reduction with carbon has been suggested [75, 78].

Much research has been undertaken into separating antimony from the copper before it is electrolytically refined. Different fluxes (e.g., alkali metal carbonates) and the thermodynamical conditions which influence the distribution of antimony between slag and metal phase in copper fire refining have been studied in detail [77, 79–84]. Still an appreciable amount of antimony remains in the cast anodes. In the electrolytical refining about 50 % of the antimony reports to the anode slime, while the rest accumulates in the electrolyte. The presence of Sb^{V} ions in the electrolyte seems to increase the amount of floating slimes in the electrolytic cells and thus has a negative influence on cathode quality [85]. During the cleaning of the electrolyte bleed-off stream, antimony is precipitated together with copper in the copper electrowinning cells and recycled into the pyrometallurgical process section [86]. Antimony can also be separated from

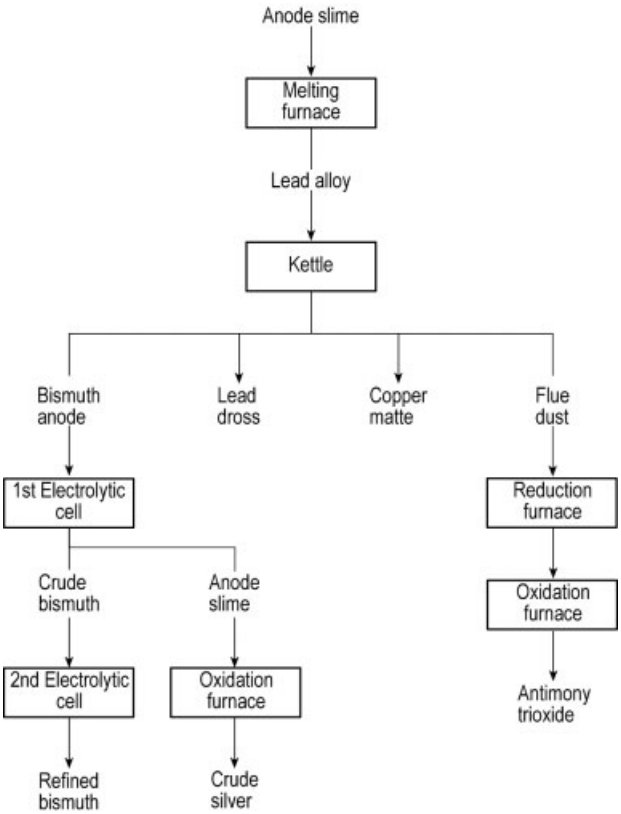

Figure 5. Basic flow sheet for lead anode slime treatment [69]

copper-containing solutions by solvent extraction or ion exchange [85, 87–93].

The chemical composition of anode slimes from electrolytical copper refineries varies over wide ranges (Table 3). The antimony content is up to 9.5 % and is generally higher when secondary raw materials are charged in the smelting section instead of concentrates. As numerous as the compositions of the slimes are the technologies which can be applied for their treatment [94–97]. The aim of the different treatment methods is mainly the recovery of precious metals Such as gold, silver, and the platinum group metals. Figure 6 shows a selection of different approaches for the treatment of anode slimes. The slags of the Doré furnace are one possibility for the recovery of antimony metal [94].

7. Refining

Antimony from the smelter usually contains iron, arsenic, tin, zinc, copper, lead, and bismuth. Several processes are used to produce a commercial grade containing $\geqq$ 99 % antimony and $\leqq$ 0.3 % arsenic. Iron, arsenic, and tin are removed by melting with caustic soda. A slow current of air is passed over the metal – caustic soda bath. Generally the caustic soda is renewed once during purification. At a bath temperature of 950 °C and a total air flow of 500 m^3/kg arsenic is reduced in two hours to a residual 0.05 % and iron is reduced to a residual 0.005 %. Sodium nitrate can be used instead of the combination of caustic soda and air.

Although the maximum permissible *arsenic* content is 0.3 %, a maximum of 0.1 % is often specified. The measures taken to ensure such a low content must begin in the earliest beneficiation stages (Chap. 5 and Section 6.1). The

Table 3. Anode slimes compositions in different refineries (in wt %) [96]

Refinery	Cu	Ag	Au	Pt	Pd	Se	Te	As	Bi	Sb	Sn	Pb	Fe	Ni
CCR (Noranda)	18.7	19.5	0.18			10.0	1.2	1.14	0.77	1.68		8.0		0.6
Chuquicamata	27	12.0	0.07			4		5		4				
Cu Refineries Pty.	27	9.0	0.07			0.7	0.03	5.8	0.25	0.45		7.5	0.15	0.40
ER & S, Australia	13	9.0	0.10		0.09	5.8	0.2	1.2	0.3	3.0	5.0	31.0		2.0
El Salvador , Chile	5	24.0	1.4			21		0.7		3				
Hibi Kyodo Co. Ltd., Tamano Smelter	21.5	9.6	0.50									22.7		
IMI Refineries, U.K.	14	5.5	0.07	0.004	0.008	2	0.6	3.5	0.5	3.5	5	22		9
Inco	21.0	6.37	0.12			8.4	1.8	0.50	0.14	0.09		1.7		17.6
Kidd Creek	26.0	12.7	0.15			19.5	0.05	0.45	0.28	0.09		22.0		
Metallurgie Hoboken Overpelt	17.4	17.2				5.1	0.95	6.8	0.74	6.8	0.35	23.2		
Mhangura Copper, Zimbabwe	2	62	1.0	0.05	0.1	10	1.0	0.16	0.13	0.07	0.15	1.8	0.1	0.3
Minero Peru	41.0	20.0	0.04			11	1.1							
Outokumpu	8.5	16.8	0.3			46						8.0		5.2
Palabora	53.4	7.8	0.33	0.05	0.07	3.6	2.2	0.15	0.01					4.5
Phelps Dodge, El Paso	27.1	12.2	0.12	0.0007	0.006	8.8	3.1	1.7		0.66		4.65	0.08	0.64
Rabak, Turkey	24.7	4.8	0.11	0.0005	0.002	7.9	3.0	2.3	0.5					0.05
Sarkuysan, Turkey	26.7	4.73	0.10			7.6	2.5	3.6		1.3		1.7		0.83
Southwire, Georgia	10.0	4.1	0.02	0.08	0.35	0.6		1.4	0.5	9.4	5.3	13.1		8.4

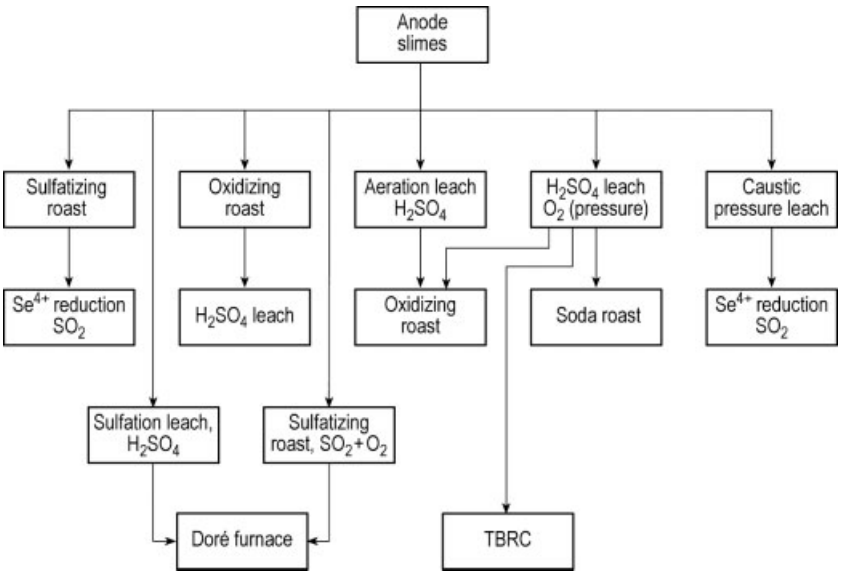

Figure 6. A selection of different approaches to the treatment of anode slimes [96]
TBRC = top-blown rotary converter

methods are well known, and some of them are standard practice [4, pp. 38 – 45].

To a great extent arsenic can be prevented by selective flotation from entering the concentrate [27, 98, 99]. About 88 % of the arsenic can be volatilized from concentrates and antimony oxide flue dust by roasting in the presence of steam at 475 – 500 °C. The associated antimony losses are limited to 10 – 45 %. If arsenic – antimony oxides are treated with hot caustic soda, the arsenic trioxide passes into solution as arsenite, whereas the antimony oxide is not changed. One to six kilograms of caustic soda is required for each kilogram of arsenic. The arsenic is completely separated from the antimony [4, p. 44].

Electrostatic precipitators offer means of separating arsenic from antimony at elevated temperature because gaseous arsenic trioxide can

pass through them, whereas solid antimony trioxide precipitates. The best results, only 0.66 – 0.79 % arsenic in the antimony trioxide, are obtained for a gas-inlet temperature of 620 – 650 °C and an outlet temperature of 430 – 470 °C.

The removal of *lead* from crude antimony is of less importance because most antimony produced is used in lead alloys. In any case, separation is difficult owing to the affinity between the two metals. Suitable methods are electrolysis, chloridization, and sulfidization.

Liquid antimony containing lead and arsenic can be refined in two subsequent stages. Lead, as well as iron, tin, and copper, are preferably removed in the first stage by means of sulfide-rich antimony slags. The best conditions for the removal of iron from liquid antimony are simultaneous oxidation and sulfidization with antimony slags containing 70 mol % Sb_2O_3. Arsenic generally does not react with Sb – S – O slags. A low refining temperature (e.g., 973K) is preferable. A possible second refining stage is the application of Na_2CO_3 – Sb_2O_3 slags, which preferentially dissolve arsenic [100, 101].

A number of methods have been proposed, and occasionally used, for electrolytic separation of lead from antimony on an industrial scale [4, 59]. Although pure antimony is obtained from alkaline thioantimonite solutions, the current efficiency is low, and there are difficulties, including the formation of an insoluble film at the anode. Under some circumstances, electrolysis of antimony chloride gives rise to explosive antimony.

However, the only electrolysis that gives better results involves antimony trifluoride. The crude metal can be refined in a sulfate – fluoride solution. Even if the percentage of lead is quite high, lead-free antimony is obtained at the cathode. The excess sulfuric acid in the electrolyte precipitates the lead as lead sulfate, which passes into the anode mud. The energy consumed is 320 kWh/t. Arsenic must be removed beforehand; otherwise, it also deposits on the cathode. Operational details are given in [60].

It is claimed [4, pp. 46, 47], [102] that antimony of 99.2 % purity is produced if 96.5 % antimony with lead as the main impurity is treated with chloride at 1100 °C. If antimony rich in lead is fused with a sulfur flux, the lead, as well as zinc, copper, and bismuth, is removed almost completely [4, p. 48]. This method is simpler and cheaper than electrolysis.

Electrolytic refining in molten sodium and potassium chloride at 750 – 800 °C is said to yield pure antimony. *Iron* can be reduced rapidly to a residual 50 – 90 ppm. Lead and copper are difficult to remove; therefore, the antimony should be purified pyrometallurgically beforehand.

8. Fine Purification [14]

The antimony used as a solvent for extremely pure silicon must also be extremely pure. Silver, aluminum, chromium, copper, iron, magnesium, manganese, and nickel can be removed almost completely by subliming commercial antimony (99.5 – 99.9 % Sb). However, almost all of the arsenic, lead, tin, and sulfur enters the sublimate, and the product is often contaminated with carbon and silicon. If commercial antimony is alloyed with 0.5 – 4 % manganese, the sublimate is free from tin and largely free from sulfur. If 1 – 2 % aluminum is used in addition to the manganese, the lead and arsenic contents are also reduced substantially. However, this procedure is not effective unless the rate of sublimation is reduced, to one third in the first case or one tenth in the second. This refining by sublimation is shown schematically in Figure 7.

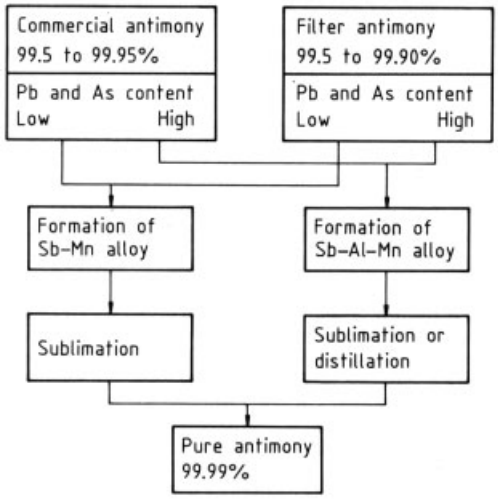

Figure 7. Sublimation refining

There is growing demand for extremely pure antimony. It is derived from previously purified metal or from antimony(III) oxide. (Table 4 shows typical compositions of the starting materials.) The impurities remaining are removed by physical, chemical, and metallurgical methods, including distillation [104], ion exchange [105], liquid – liquid extraction [106], precipitation of intermetallic compounds [8], and zone melting [107–113].

Table 4. Spectrographic analysis of antimony regulus and of the antimony trioxide obtained by oxidation [103]

Element	Contents, ppm unless otherwise indicated	
	Antimony regulus	Antimony trioxide
As	210	192
Pb	3600	380
Fe	206	158
Ni	85	15
Cu	90	10
Sn	22	12
Bi	17	7
Ag	12	< 1
S	230	106
Sb	99.5 %	> 83.38 %
Sb as Sb_2O_3	–	> 99.8 %

A method described by M. Tanenbaum involves dissolving antimony of 99.8 % purity in hydrochloric acid, repeatedly distilling the antimony trichloride obtained to remove the arsenic, and reducing the purified antimony trichloride with iron pentacarbonyl. Zone melting then gives a product of 99.999 % purity [112].

A process patented by M. Wilhelm uses hydrogen to reduce antimony pentachloride that has been purified by repeated distillation at 700 – 900 °C in an atmosphere of argon. The yield is $\leqq$ 70 % [113].

Antimony can be separated from numerous impurities by fractional distillation of the anhydrous trichloride or pentachloride [104]. Platinum, gold, silver, manganese, nickel, cobalt, and copper remain behind in the residue. If the heating is carried out carefully, the lead, bismuth, zinc, thallium, and cadmium also remain in the residue. Aluminum chloride, arsenic trichloride, and tin(IV) chloride pass completely into the first fractions, whereas tin(II) chloride remains behind. Currently only hydrochloric acid solutions are distilled. Under these circumstances, arsenic trichloride distills at temperatures below 110 °C. Because distillation above 110 °C results in considerable loss of antimony, the temperature must be kept slightly under 110 °C.

Ileva and Vigdorovich stressed the effectiveness of chemical methods in separating arsenic from antimony and in lowering antimony's heavy metal content. Antimony trichloride is prepared by dissolving technical-grade antimony in hydrochloric acid and is hydrolyzed to give antimony trioxide, which is reduced with pure sucrose. The metal is then purified by zone melting to give a product with 99.995 % antimony in 51 – 53 % yield [114].

One rather complex procedure consists of the following: (1) dissolving technical-grade antimony in hydrochloric acid, (2) purifying the antimony trichloride obtained by fractional distillation, (3) converting the purified antimony trichloride into antimony trisulfide, (4) dissolving the sulfide in pure hydrochloric acid, (5) distilling the antimony trichloride, (6) hydrolyzing the trichloride into antimony trioxide, and (7) reducing the oxide with hydrogen. The purity of the antimony exceeds 99.998 %, but the yield is low because of the large number of operations.

Very pure antimony trichloride can be purified further by hydrolysis to antimony trioxide with extremely pure water. The trioxide is reduced with purified hydrogen, first at about 600 °C and then at a somewhat higher temperature [8].

The recrystallization of chloroantimonic acid (H_2SbCl_6) efficiently removes final traces of iron [104].

Ion exchange removes cobalt, nickel, zinc, and lead [115, 116]. However, hydrolysis can result in less than quantitative elution.

Liquid – liquid extraction offers a means for refining antimony [106], e.g., by purifying acidic solutions of antimony with alkylphosphoric acid or separating copper, iron, and lead from hydrochloric acid solutions by extraction with ethyl acetate. Pentavalent antimony dissolved in hydrochloric acid can be extracted by dibutyl ether, amyl alcohol, hexyl alcohol, *n*-octyl alcohol, and diisoamyl ether. However, liquid – liquid extraction is seldom used in practice.

The effectiveness of zone melting, usually the last stage in fine purification, depends on the impurities present. Arsenic, tin, and bismuth have unfavorable distribution coefficients (Table 5) and are difficult to remove by zone melting.

Arsenic must be removed completely in the preceding purification steps. Lead and germanium are easier to remove. In one case, seven passages of the melted zone dropped the nickel, lead, silver, and copper contents to one tenth their original values [111, 112]. The impurity levels, in ppm, are reduced to < 0.5 Ag, < 5 As, < 1 Bi, 0.5 Cu, < 1 Fe, < 5 Ni, 2 Pb, and < 2 Sn [103].

Table 5. Distribution coefficients in antimony [103]

As	0.64	Pb	0.06
Sn	0.3	Ge	< 0.06
Bi	0.2		

Very pure antimony is produced in three grades: > 99.99 % (4N), > 99.999 % (5N), and > 99.9999 % (6N). Very pure antimony for electronic applications has to be at least of an 99.999 % Sb quality. Qualities of 99.999999 % Sb (Mikuni Smelting & Refing Co., Japan) and of 99.9999 % Sb (Preussag AG Metall, Cominco American Corp., USA) are also available.

9. Recovery of Antimony Oxide

Antimony(III) oxide is being used as flame retardant in an increasing quantities. It is produced mostly by roasting sulfidic antimony ores or other antimony-containing raw materials. The main techniques applied are described in Chapter 6.1. To recover antimony oxide of higher purity, especially for the separation of arsenic and antimony oxide, selective volatilization of the more volatile arsenic trioxide and selective condensation of the less volatile antimony tetroxide at high temperatures, leaving arsenic trioxide to condense out at lower temperatures, can be applied [19, 36]. Antimony trioxide can also be recovered from complex antimony sulfide ores by reaction with calcium chloride. The reaction takes place in an oxidizing atmosphere at about 500 °C.

$$Sb_2S_3 + 3\,CaCl_2 + 6\,O_2 \rightarrow 2\,SbCl_3 + 3\,CaSO_4$$

More than 90 % of the antimony is volatilized as $SbCl_3$, which can be purified by distillation and hydrolyzed to Sb_2O_3. The flow sheet is shown in Figure 8 [32].

10. Antimony Alloys and Intermetallic Compounds

Antimony is a component of many lead and tin alloys (→ Lead Alloys, → Tin, Tin Alloys, and Tin Compounds). These alloys are important materials for making bearings (→ Semiconductors) and solders (→ Lead Alloys).

The intermetallic compounds AlSb [*25152-52-7*], GaSb [*12064-03-8*], and InSb [*1312-41-0*] (→ Lead Alloys) are important. Zinc antimonide, ZnSb, has interesting semiconductor and thermoelectric properties; cadmium antimonide, CdSb, has interesting semiconductor properties. Antimony is used to dope very pure germanium, silicon, and gallium arsenide to produce *n*-type semiconductors.

11. Antimony Compounds

Antimony can be trivalent or pentavalent in compounds. Both exist in antimony(IV) oxide, Sb_2O_4, and presumably in the deep violet haloantimonates(IV), such as $Cs_2(SbCl_6)$. The pentavalent compounds, as well as most of the trivalent ones, are hydrolyzed by water to form antimony salts, e.g., SbOCl, and hydrated oxides. The only halides of antimony that have any industrial significance are the chlorides and fluorides.

Mixed antimony(V) chloride fluorides and bromide fluorides exist, but antimony pentabromide does not. The tribromide, $SbBr_3$, is chemically similar to the trichloride. Stable complex halides are known, mainly hexahaloantimony anions of the types $[SbX_6]^{3-}$ and $[SbX_6]^{-}$.

11.1. Antimony Chlorides

Antimony trichloride [*10025-91-9*] (antimony chloride, butter of antimony, caustic antimony), $SbCl_3$, M_r 228.1, *mp* 73 °C, *bp* 223 °C, ϱ 3.064 g/cm^3 at 26 °C, is soft, crystalline, colorless, and very hygroscopic. It is soluble in the usual organic solvents, including benzene, carbon disulfide, carbon tetrachloride, acetone, and ethanol. Only highly concentrated aqueous solutions are stable. Dilution causes antimony(III) chloride oxide, SbOCl, to precipitate. The addition of

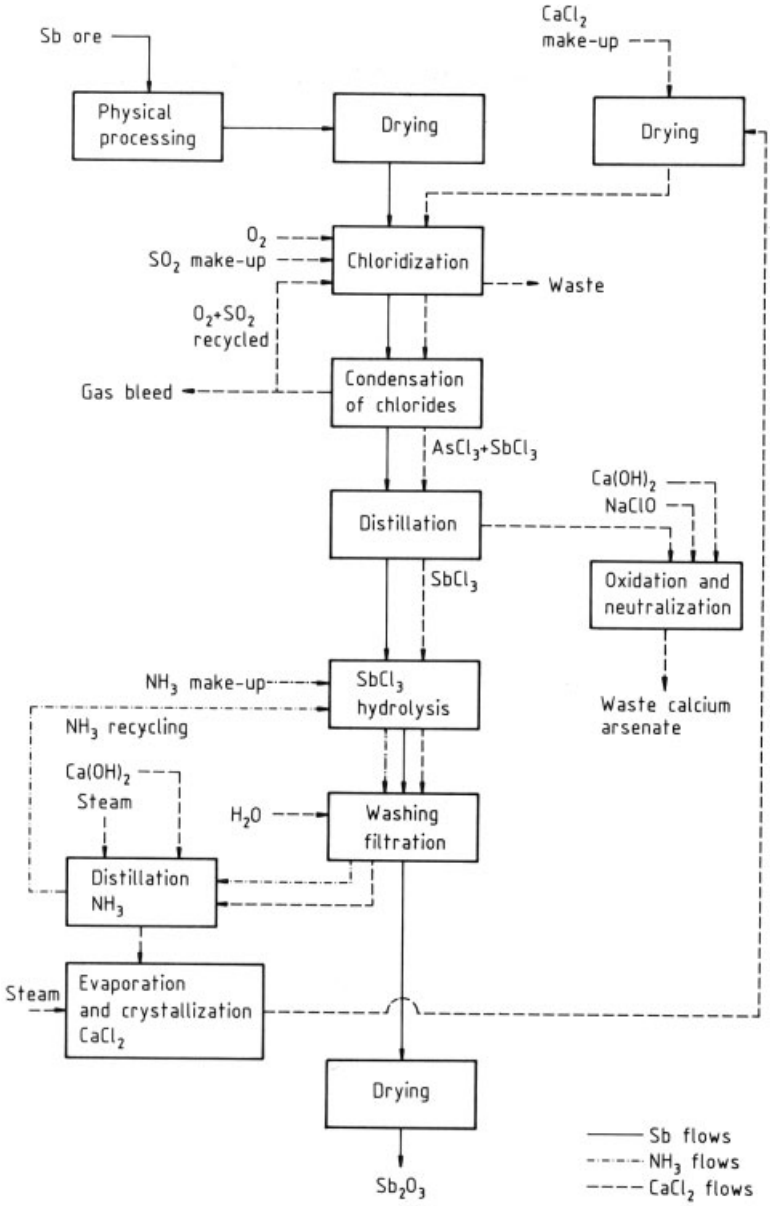

Figure 8. Chloridization process

more water causes further hydrolysis to take place, forming hydrated antimony(III) oxide. Antimony trichloride forms complexes with numerous compounds, some having a striking deep color.

Antimony trichloride is an intermediate in the production of other antimony compounds, a catalyst in organic chlorination and polymerization reactions, and a vitamin A reagent. Large quantities are consumed in rendering textiles, plastics, and other combustible substances flame retardant or self-extinguishing.

Although manufacture from antimony metal and chlorine is expensive, it is feasible on a small scale. The best procedure is to react chlorine with antimony metal in liquid antimony trichloride or concentrated hydrochloric acid. In the first case the temperature can be as high as

200 °C. Equally good results are obtained by dissolving antimony oxides in hot, concentrated hydrochloric acid. Distillation of the crude product or its solution in strong hydrochloric acid in the presence of some metallic antimony or iron (to reduce antimony(V) and iron(III) compounds) then yields pure antimony trichloride.

Stibnite is usually digested in concentrated hydrochloric acid. The antimony passes into solution, whereas the arsenic largely remains in the residue. Disadvantages are the noxious hydrogen sulfide given off and difficulties in heating the strongly acidic solutions.

Antimony chloride oxide (oxychloride), [*7791-08-4*], SbOCl, is a white crystalline powder. It is precipitated by hydrolysis when concentrated solutions of antimony trichloride in water or hydrochloric acid are diluted. Because other chloride oxides, e.g., $Sb_4O_5Cl_2$, also form, reaction conditions must be controlled carefully. Antimony chloride oxide is an intermediate in the production of antimony salts and is used in smoke-producing agents.

Antimony pentachloride [*7647-18-9*], $SbCl_5$, M_r 299.1, *mp* 4 °C, *bp* 140 °C, ϱ 2.34 g/cm^3 at 20 °C, is a pale yellow fuming liquid of unpleasant odor. It can be produced by saturating molten antimony trichloride with chlorine, followed by vacuum distillation. Other methods are the direct chlorination of antimony and the chlorination of antimony oxides or sulfides.

Antimony pentachloride, which easily dissociates and tends to chelate, acts as a chlorine donor in the chlorination of organic compounds. It is also used as a polymerization catalyst.

11.2. Antimony Fluorides

Antimony trifluoride [*7783-56-4*], SbF_3, M_r 178.8, *mp* 292 °C, *bp* 319 °C, ϱ 4.39 g/cm^3, is a colorless compound that crystallizes readily and volatilizes at room temperature. If it is exposed to atmospheric moisture, hydrogen fluoride is given off. In organic solvents it is less soluble than the chloride, but both dilute and concentrated aqueous solutions are stable. Antimony trifluoride forms complexes with inorganic compounds. It is an important fluorinating agent, replacing chlorine or bromine by fluorine in both inorganic and organic compounds. It is especially useful for converting acid chlorides to the fluorides. However, if the halogen group in organic compounds is not very reactive, fluorination is quite sluggish. The rate of reaction can be increased considerably by adding some antimony pentachloride or bromine, in order to form antimony(V) bromide fluorides. Antimony trifluoride is used in the production of pottery and porcelain, and potassium fluoroantimonate(III) is a mordant in dyeing.

Antimony pentafluoride [*7783-70-2*], SbF_5, M_r 216.8, ϱ 2.99 g/cm^3, is a colorless oily liquid at room temperature. Under atmospheric pressure at 7 °C it solidifies to a crystalline mass; it boils at 150 °C. The solid dissolves in water with a violent reaction. The pentafluoride carbonizes and fluorinates organic substances. The chlorine in chlorinated hydrocarbons is replaced by fluorine. Many inorganic substances form complexes with antimony pentafluoride.

Antimony pentafluoride is prepared by treating antimony pentachloride with excess anhydrous hydrogen fluoride. The antimony pentafluoride is separated from the volatile byproducts by fractional distillation. It can also be prepared by allowing fluorine to react with antimony powder or molten antimony trifluoride.

The main use of antimony pentafluoride is as a fluorinating agent.

11.3. Antimony Tribromide

Antimony tribromide [*7789-61-9*], $SbBr_3$, M_r 361.51, *mp* 96.6 °C, *bp* 280 °C, ϱ 4.15 g/cm^3, is a white solid. It is used as a mordant.

11.4. Antimony Triiodide

Antimony triiodide [*7790-44-5*], SbI_3, M_r 502.50, *mp* 167 °C, *bp* 401 °C, ϱ 4.78 g/cm^3, is a red solid that hydrolyzes readily and forms complex ions. It adds to ethers, aldehydes, and thiols. Antimony triiodide is prepared by reacting iodine with antimony or antimony sulfide in an organic solvent. The triiodide is used in medicine.

11.5. Antimony Oxides

Antimony trioxide [*1309-64-4*], Sb_2O_3, M_r 291.52, has two modifications: senarmontite be-

low 570 °C, ϱ 5.25 g/cm^3, and valentinite (white antimony, antimony bloom) above 570 °C, ϱ 5.7 g/cm^3. The melting point is 656 °C. The commercial product is a white crystalline powder with a density of 5.2 – 5.8 g/cm^3. On heating, it becomes yellow but turns white again on cooling. Antimony trioxide dissolves in acids and alkalies. For uses see Chapter 13.

Antimony trioxide can be produced by roasting antimony trisulfide, by burning antimony in air or oxygen, by alkaline hydrolysis of antimony halides (trichloride, tribromide, or triiodide), and by hydrolysis of antimony trisulfide with superheated steam. The production of antimony trioxide from antimony trisulfide during ore processing is described in Section 6.1.

Antimony tetraoxide [*1332-81-6*], Sb_2O_4, is a white nonvolatile powder that is soluble in alkalies but not in acids. Its density can differ considerably according to the duration of heating during preparation: 4.07 – 7.5 g/cm^3. Cervantite has a density of 4.07 g/cm^3. Antimony tetraoxide is obtained by roasting antimony trisulfide in a current of air or by heating antimony trioxide at 300 – 400 °C in air. The tetraoxide is used in ceramics.

Antimony pentoxide [*1314-60-9*], Sb_2O_5, is a whitish yellow powder. On heating, it first turns deep yellow; above 300 °C it turns from brownish yellow to brown but becomes paler above 380 °C. Its density ranges from 3.78 to 7.86 g/cm^3. Antimony pentoxide is sparingly soluble in water but readily soluble in alkalies.

The pentoxide can be produced by the action of strong oxidizing agents on metallic antimony or antimony trioxide. Hydrated antimony pentoxide can be produced by the hydrolysis of antimony pentachloride. The hydrate can be dried without difficulty.

11.6. Antimonic Acid and Antimonates

Many years ago antimonic acid was thought to have ortho, pyro, and meta forms, but these have all turned out to be only hydrated antimony oxide, primarily colloidal and differing in water content and solubility in acids and bases.

Potassium antimonate [*12208-13-8*], $K[Sb(OH)_6]$. Heating antimony pentoxide with excess potassium hydroxide produces rubbery, granular, or crystalline products. If the substance is recrystallized from a little water, deliquescent mammilated crystals form. Treated with copious cold water or boiled together with a small quantity of water, they are converted into potassium hexahydroxyantimonate, $K[Sb(OH)_6]$, which is sparingly soluble in cold water but somewhat more soluble in warm water. Solutions precipitate sodium as the sparingly soluble sodium hexahydroxyantimonate, $Na[Sb(OH)_6]$. This reaction is used for the detection of sodium.

Sodium hexahydroxyantimonate [*12339-41-2*] was formerly thought to be sodium pyroantimonate, $Na_2H_2Sb_2O_7 \cdot 5\ H_2O$. The hexahydroxyantimonate is prepared by melting antimony or antimony oxides with excess sodium nitrate. It is used as an opacifier for glass and enamel.

According to X-ray diffraction analysis, the hydrated antimonates(V) of magnesium, cobalt, and nickel are hexahydroxyantimonates, $[M(H_2O)_6][Sb(OH)_6]_2$, M = Mg, Co, or Ni.

11.7. Antimony Sulfides

Antimony trisulfide [*1345-04-6*] (stibnite), Sb_2S_3, M_r 339.70, *mp* 550 °C, ϱ 4.64 g/cm^3, has a black crystalline modification and an unstable red-orange modification. If both occur together, the mixture can be carmine or brownish red.

Antimony trisulfide is soluble in alkalies and in concentrated, but not dilute, hydrochloric acid. It also goes into solution if it is boiled in alkali carbonate solutions. However, it precipitates as a reddish-brown powder, of nondefinite oxygen content (kermesite), when the solution is cooled. It forms thioantimonates if it is dissolved in alkali-metal sulfides. Antimony trisulfide is a starting material for the production of antimony and antimony compounds. For specific uses see Chapter 13.

Antimony pentasulfide [*1315-04-4*], Sb_2S_5, M_r 403.82, is an orange to dark red powder. It is doubtful whether stoichiometric Sb_2S_5 exists, because the preparations contain free sulfur and have variable composition. Antimony pentasulfide is insoluble in water but soluble in alkalies and alkali-metal carbonates. Thioantimonates are formed if the pentasulfide is dissolved in alkali sulfides; the most important of these is Schlippe's salt [*1317-86-8*], sodium tetrathioantimonate, $Na_3SbS_4 \cdot 9\ H_2O$, which is

produced on a commercial scale by melting stibnite, sodium sulfide, and sulfur together. Antimony pentasulfide can be obtained from Schlippe's salt by reaction with hydrochloric acid.

11.8. Antimony Sulfate

Antimony sulfate [*7446-32-4*], $Sb_2(SO_4)_3$, M_r 531.74, ϱ 3.62 g/cm^3, is a hygroscopic compound in the form of needles with a silken sheen. When antimony or antimony trioxide, trisulfide, or oxychloride is dissolved in hot sulfuric acid, antimony sulfate or antimonyl salts are recovered. Antimony sulfate is used in the manufacture of explosives and in fireworks.

11.9. Stibine

Stibine [*7803-52-3*], antimony hydride, SbH_3, M_r 124.77, is a poisonous, foul-smelling gas of density 5.685 g/L (STP). The liquid boils at −88.5 °C. The gas is slightly soluble in water, soluble in alcohol, and very soluble in CS_2. Its formation is endothermic, ΔH_f = 140 kJ/mol, and it easily decomposes to hydrogen gas and antimony metal. Stibine is used to prepare *n*-type silicon semiconductors.

11.10. Organometallic Compounds

Organic compounds of antimony are known for the antimony oxidation states + 3 (SbR_3, R_2SbX $RSbX_2$, and + 5 (SbR_5, R_4SbX, R_3SbX_2, R_2SbX_3, $RSbX_4$), and the six-coordinate anions $[SbR_6]^-$). Organodistibines, $R_2Sb\text{-}SbR_2$, oligomeric and polymeric compounds with more than one Sb – Sb bond, $(RSb)_n$, and an aromatic organoantimony derivative, stibabenzene, have been described. A large number of derivatives of the stibonic acids, $RSbO(OH)_2$, and the stibinic acids, $R_2SbO(OH)$, have also been investigated [117–120].

Tertiary stibines, SbR_3, are usually prepared by Grignard alkylations of $SbCl_3$, by reaction of antimony trihalides with organolithium compounds, or by alkylation of Sb_2O_3 with trialkylaluminum in hexane. An example of the last is the preparation of $Sb(C_4H_9)_3$ [121]. Direct synthesis from Sb and organic halides, first used in 1850, is used to prepare $Sb(CF_3)_3$ from Sb and CF_3I at 170 °C [122]. Trialkylstibines are air-sensitive, the lower alkyl derivatives are even pyrophoric. $Sb(CH_3)_3$ is a liquid boiling at 79 – 81 °C under 101.3 kPa. $Sb(C_6H_5)_3$, *mp* 45 – 58 °C, *bp* 185 °C under 40 Pa, and other triarylstibines are more stable. The triorganylantimonys, SbR_3, react with halogens in an oxidative addition reaction to form triorganoantimony dihalides and with alkyl halides to produce the stibonium halides $[R_3SbR]^+X^-$. They also react with transition metal complexes to form R_3Sb – transition metal complexes containing metal – antimony bonds. The triorganylantimonys, especially $Sb(C_6H_5)_3$, are used as cocatalysts, antioxidants for certain organic compounds, corrosion inhibitors for steel, flame retardants for transparent resins, and an antifogging reagent in color film development [117].

The organoantimony(III) halides R_2SbX and $RSbX_2$ are made by pyrolysis of triorganoantimony dihalides and diorganoantimony trihalides, respectively, or by redistribution between SbX_3 and R_3Sb or R_2SbX. For example, $(C_6H_5)_2SbCl$ and $C_6H_5SbCl_2$ are synthesized by reaction of Ph_3Sb and $SbCl_3$ in the appropriate molar ratios [123]. $(C_6H_5)_2SbCl$, *mp* 68 – 70 °C is also formed from $SbCl_3$ and $Sn(C_6H_5)_4$ [124]. Only Grignard reagents with bulky alkyl groups, (R = *tert*-C_4H_9) are useful for stepwise alkylation reactions of $SbCl_3$ to give, for example, $(Me_3Si)_2CHSbCl_2$ [125] or tBu_2SbCl [126] selectively.

Addition of halide ions to organoantimony(III) halides results in the formation of $[R_2SbX_2]^-$, $[(RSbX_3)_2]^{2-}$ or $[RSbX_4]^{2-}$ [127]. Diorganoantimony cations, R_2Sb^+, are stabilized in adducts with donor molecules [128]. Diorganoantimony anions, R_2Sb^-, are formed by metalation of R_2SbH with BuLi [129].

R_2SbX and $RSbX_2$ compounds where X is a group other than halogen are also known. Hydrolysis of R_2SbCl leads to $(R_2Sb)_2O$, e.g., $[(C_6H_5)_2Sb]_2O$, *mp* 78 – 80 °C, and hydrolysis of $RSbCl_2$ leads to $(RSbO)_n$, e.g., $(RSbO)_4$ [R = $(Me_3Si)_2CH$] [130]. Esters of organic acids, such as $(C_6H_5)_2SbOOCCH_3$, *mp* 127 – 133 °C [131]; mercaptans, R_2SbSR or $RSb(SR)_2$; and

compounds with Sb–N, Sb–P, and Sb–metal bonds are known [117].

Pentamethylantimony, $Sb(CH_3)_5$, obtained from $(CH_3)_3SbBr_2$ and methyllithium, is a very reactive liquid, *bp* 130 – 131 °C under 101.3 kPa, which sometimes explodes during distillation [132]; the corresponding phenyl derivative, $Sb(C_6H_5)_5$, *mp* 164 – 172 °C, made by reacting $(C_6H_5)_4SbBr$ with phenyllithium, is much more stable [133].

Tetraalkyl- and tetraarylantimony compounds, R_4SbX, are known as stibonium salts. $(CH_3)_4SbF$, *mp* 50 °C, is made from $Sb(CH_3)_5$ and $(CH_3)_3SnF$ in ether [134], and $(C_6H_5)_4SbCl$, *mp* 202 – 205 °C, is the product of the reaction between $Sb(C_6H_5)_3$, $AlCl_3$, and C_6H_5Cl [135].

The alkoxides $(CH_3)_4SbOR$ have been prepared from $Sb(CH_3)_5$ [134]. $Sb(CH_3)_5$ can also be cleaved by mercaptans to form $(CH_3)_4SbSR$ [136] or by HN_3, HCN, or HSCN to form the corresponding tetramethylantimony azides, cyanides, or thiocyanates [137]. $Sb(C_6H_5)_5$ reacts with LiC_6H_5 in ether to form a white powder, $Li[Sb(C_6H_5)_6]$, which melts at 185 °C [133, 138].

Compounds of the type $RSbX_4$ (R = aryl) are prepared by chlorination of $RSbCl_2$ with Cl_2. They are usually converted to stibonic acids. $(C_6H_5)_2SbF_3$ is obtained by fluorination of $(C_6H_5)_2SbF$ with XeF_2 [139], and $(CH_3)_2SbCl_3$ is prepared by chlorination of $(CH_3)_2SbCl$. Other alkyl derivatives are obtained by the chlorination of distibines, $R_2Sb–SbR_2$ [140]. Trialkyl- and triarylantimony dihalides R_3SbX_2, prepared by the direct halogenation of SbR_3 compounds, are quite stable substances. They decompose on heating to form R_2SbX and RX [141].

Older references describing the existence of diazene analogues, RSb=SbR, are erroneous. However, recently a sterically protected distibene, RSb=SbR, R = 2,4,6-$[(Me_3Si)_2CH]_3C_6H_2$, was synthesized [142]. Compounds like $(CH_3Sb)_n$ are polymeric, and $(C_6H_5Sb)_6$, a molecule whose core is a six-membered ring of antimony atoms in a chair configuration, is hexameric [143]. Mono(*tert*butyl)antimony is tetrameric, $(tert\text{-}C_4H_9Sb)_4$ [126]. Compounds formed in the reactions of $RSbCl_2$ or $(RSbO)_n$ derivatives with primary arsines, or by reduction of arylarsonic acids or arylarsine oxides with antimony compounds, and containing As–Sb bonds and most probably polymeric were used as chemotherapeutics [144]. Tetraorganodistibines, $R_2Sb–SbR_2$, for example, $(CH_3)_2SbSb(CH_3)_2$ [145], are obtained by the action of organic free radicals upon metallic antimony mirrors. $(C_6H_5)_2SbSb(C_6H_5)_2$ is obtained from $Sb(C_6H_5)_3$ and sodium in liquid ammonia [146].

Stibabenzene, a reactive two-coordinate organoantimony compound, was the first example of a new class of organoantimony compounds [147]:

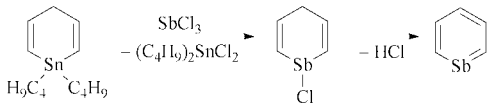

A distibabutadiene R'RC=Sb–Sb=CRR' (R = Me_3SiO, R' = 2,4,6-tBu$_3C_6H_2$) is formed by reaction of R'COCl with $Li[Sb(SiMe_3)_2]$ [148]

Stibonic acids, $RSbO(OH)_2$, R = aryl, are prepared from $SbCl_3$ by the diazo reaction with RN_2Cl and subsequent hydrolysis and are amorphous solids. Methylstibonic acid, $(CH_3SbO_2)_n \cdot m\ H_2O$ is formed by hydrolysis of $CH_3Sb(OCH_3)_4$ [149]. The dialkylstibinic acids, $R_2SbO(OH)$, are prepared by hydrolysis of the corresponding R_2SbCl_3 derivatives, whereas diphenylstibinic acid is made from $Sb(C_6H_5)_3$ by oxidative cleavage with NaOH and H_2O_2 [150]. The anhydride of diphenylstibonic acid is a cage compound $(Ph_2Sb)_4O_6$ [151]. *p*-Aminophenylstibonic acid is an ingredient of neostibosan, a drug for the treatment of oriental sore (leishmaniasis) [152]. Orthoesters of diarylstibinic acids are fungicides and bactericides, used in paints, plastics, and fibers [153].

Aliphatic and aromatic derivatives of stibine oxides, R_3SbO, and their hydrates, $R_3Sb(OH)_2$, are prepared by hydrolysis and subsequent dehydration of R_3SbCl_2. For example, $(CH_3)_3Sb(OH)_2$ is prepared by hydrolysis of $(CH_3)_3SbCl_2$. The hydrate loses water above 110 °C to give $(CH_3)_3SbO$ [154]. Triphenylstibine oxide $(C_6H_5)_3SbO$ is dimeric in the solid state [151].

12. Chemical Analysis

The presence of antimony can be demonstrated by the Marsh test. There are a number of classical gravimetric and volumetric methods for determining antimony. A popular colorimetric method is based on a complex between rhodamine B and antimony. Instrumental methods for traces of antimony include polarography, emission spectrography, neutron activation, and atomic absorption [155].

13. Uses

The uses of antimony can be divided into three categories: uses in alloys, uses in the oxide form, and other uses of antimony chemicals. For the U.S. market the reported industrial consumption of primary antimony by products is given in Table 6. About 50 % of the primary antimony production is used for flame retardants in plastics, while only about 20 % is used in metallic products. The former main application of antimony as an alloying element in lead alloys for the production of lead – acid batteries is declining, but most of the antimony still in use for this purpose originates from the recycling of used batteries, which is not included in the table. One product of the recycling process is a lead – antimony alloy, which is upgraded by the addition of pure lead and reused for the production of batteries.

Antimony Metal. Owing to its poor mechanical properties, antimony metal is not used directly but is used in numerous alloys, primarily with lead and tin.

Antimonial (Hard) Lead (see also → Bearing Materials). The addition of antimony to lead increases the tensile strength of cold-rolled lead from approximately 16.6 to approximately 32.1 MPa for an 8 % antimony alloy. Uses include storage-battery grids (7 – 12 % Sb), small-arms ammunition (12 – 15 % Sb), and cable sheathing and lead pipe (1 to 9 % Sb). Lead pipe is used in many applications that require acid-resistant pipe coils.

Antimonial lead (1.5 – 3 % Sb) is used in the grid plates, straps, and terminals of lead – acid batteries. The addition of antimony to lead improves the fluidity, making the casting of battery grids easier; moreover, the tensile strength is increased, and the thin grids are able to withstand the stresses imposed by the charging and discharging reactions. There is also an improvement in the electrochemical stability of the lead.

Table 6. Industrial consumption of primary antimony in the United States by product [20]

Product	Consumption	
	t/a	%
Metal products		
Antimonial lead[a]	1170	8.7
Bearing metal and bearings	45	0.3
Solder	226	1.7
Others[b]	1150	8.5
Total	2600	19.3
Nonmetal products		
Ammunition primers	20	0.1
Ceramics and glass	1080	8.0
Pigments	824	6.1
Plastics	1220	9.0
Others[c]	159	1.2
Total	3300	24.4
Flame retardants		
Adhesives	344	2.5
Plastics	6610	49.0
Rubber	342	2.5
Textiles	247	1.8
Others[d]	7	0.1
Total	7550	55.9
Grand total	13500	100

[a] Antimonial lead from secondary sources (recycling of lead batteries) not included.
[b] Includes ammunition, cable covering, castings, sheet and pipe, and type metal.
[c] Includes fireworks and rubber products.
[d] Includes paper and pigments.

Babitt Metals. Tin-based babbitts consist of ternary tin – antimony – copper alloys or quaternary tin – antimony – copper – lead alloys containing 4.5 – 14 % Sb and are used for low-load bearings that have good antiseizure properties and corrosion resistance but low fatigue strength.

Lead-based babbits are alloys of antimony, tin, and lead containing 8 – 15 % Sb, and may contain 1 – 3 % As (see → Soldering and Brazing). They are used for heavy-duty bearings where fatigue resistance is important, e.g., in railways.

Type Metals. Alloys of lead and antimony containing 2.5 – 25 % Sb are used in the printing industry. Antimony lowers the casting temperature, increases the hardness and minimises the shrinkage during freezing.

Brittania metal and pewter are alloys consisting primarily of tin with ca. 2 % Cu and 7 – 20 % Sb that are used in the manufacture of vases, lamps, candlesticks, tea and coffee services, etc. The antimony increases the hardness of the metal and permits a highly polished surface to be obtained.

Solders. Antimony is added to solders to increase the hardness of the tin – lead alloys. The Sb content is typically less than 1 % but may be as high as 13 % for certain applications such as electrical equipment, the end and side seams of cans, car radiators and other plumbing, and as a car-body filler for gaps in spot-welded seams.

Semiconductors. Increasing amounts of antimony are used as dopants in the manufacture of *n*-type semiconductors and III/V intermetallic semiconductor materials such as indium, aluminum, and gallium antimionide for the substrates of light-emitting diodes.

Antimony Trioxide. The uses of antimony trioxide can be divided into two broad categories: flame-retardant and non-flame-retardant. Antimony trioxide in combination with a halogen imparts flame-retardant properties to various materials. Ineffective on its own, the antimony must be intimately associated with the halogen, which is usually present as a chlorinated alkyd resin, a poly(vinyl chloride) resin, or a chlorinated plasticizer. A theory commonly used to explain this phenomenon recognizes that the primary mechanism concerns the generation of antimony trihalide and/or oxide halide. Antimony halides promote reactions that cause the formation of carbonaceous char rather than volatile gases. This char then acts as a heat shield that retards the breakdown of the plastics, thus preventing the release of flammable gases.

As an additive in plastics, antimony trioxide also acts as an opacifier. This is can be an desirable characteristic if colorant content is to be minimized. To minimize colorant requirements while retaining flame retardancy, the oxide particles should be 2 – 3 μm larger than the particles of standard pigments. Savings of up to 30 % in pigments can be achieved if the average grain size is increased from 1 to 3 μm. This explains the need for the fine control of the particle size during the final refining of the antimony trioxide. The market for antimony trioxide has expanded rapidly during since ca. 1970. *Flame-retardant applications* are as follows:

- Poly(vinyl chloride) for conveyor belts in mines, cable coatings, and coated fabrics for wall covering, cushion covers, etc.
- High-impact polystyrenes for television backs and domestic electrical appliances, and as housings for electrical equipment (including PCs)
- Polyethylene and polypropylene, primarily for wire sheathing and electrical conduits in buildings
- Polyamides and engineering plastics such as nylons for automotive uses, industrial equipment, and electrical moulded parts
- Unsaturated polyesters for applications such as building panels, automotive parts, and lifeboat hulls.

The major *non-flame-retardant* uses of antimony trioxide are as follows:

- In pigments, particularly in the manufacture of chromate pigments.
- As an opacifier for ceramic glazes and as a frit.
- In the refining of lead crystal glass and glass for television tubes. The oxide removes color and gas bubbles. Sodium antimonate is also used for these purposes.
- As a polymerization catalyst in the manufacture of polyester fibers.

Others. Antimony and antimony salts are used for a number of minor applications:

- Organic antimony compounds such as tartar emetic (potassium antimonyl tartrate) and fuadin are used in the treatment of certain tropical diseases such as bilharzia.
- In certain pesticides.
- Ammunition primers, flares, tracer shells, and fireworks (black powder).
- Vulcanization of red rubber compounds (SbS_5).
- Mordants
- As a lubricating agent in the manufacture of disk-brake pads and cutting disks.

14. Recycling

The main forms of *antimony scrap* are type metal, Babbitt metal, lead and tin waste, antimonial lead, and especially waste lead – acid batteries: the recycling of antimony is closely connected to lead recycling. The extent to which antimony is recycled has increased steadily over the last few decades, in proportion to the worldwide production of automobiles.

Antimony accounts for about 1.8 – 2 % of the total weight of an automobile battery, although the trend is towards using less. Today, waste batteries are separated mechanically into a plastics fraction, a metal fraction, and an oxide – sulfate sludge. The grid metal contains 87 – 88 % lead and 3 – 3.5 % antimony; the sludge, 70 – 76 % lead and 0.5 – 1 % antimony. Most refineries process these materials separately from ore or concentrate. The waste grid metal can be smelted in rotating reverberatory or rotary kilns to yield antimony – lead alloy. The sludge is treated with reagents that bind sulfur, e.g., iron filings and soda, and smelted under reducing conditions to antimonial lead in rotating reverberatory furnaces, short rotating furnaces, or in some cases rotating tube furnaces. The antimonial lead produced can be remelted together with lead containing a higher proportion of antimony to produce commercial alloy. Frequently, the antimony in the crude antimonial lead is converted by oxidation into an antimonial slag or skimmings, as in the refining of lead bullion. These oxidized products, essentially lead and antimony, are treated with fluxes and coal and smelted in short rotating or rotating reverberatory furnaces to yield lead containing a high proportion of antimony. Because battery alloys now contain less antimony, the proportion of antimony in recycled lead will be lower in the years to come [14].

Antimony used as antimony oxide is far more difficult to recycle and mostly ends up in household waste. A Japanese study found 7.6 ppm Sb in regular household waste. About 63 % of the antimony came from materials with a comparatively high antimony content: 30 % from curtains (2100 ppm Sb), 20 % from industrial textiles (250 ppm Sb), and another 10 % from bedding (150 ppm Sb) [156].

15. Economic Aspects

Antimony ores are mined in about 20 countries worldwide (see Table 1 and Chap. 4). In 2003, antimony mine production was estimated at about 81 600 t. Only three of the countries in which antimony is mined have mine production capacities that might influence the world market situation: China (86 %), Bolivia (2.8 %), and the Republic of South Africa (6.4 %).

The main capacities for the production of antimony metal and antimony oxide are also located in these countries. Further large production capacities can be found in Mexico, in the United States and in Australia. In 1991 the production capacities for antimony were estimated to be about 110 000 t Sb worldwide [18]. The antimony production in 1990 was 52 685 t Sb [157]. Because China's share of the world antimony production was about 40 % in 1990 [157] and has increased since then, these production figures are based on uncertain information and may differ considerably depending on the literature consulted.

Production from secondary raw materials still is of importance for the antimony market. The main source for antimony recycling are lead – acid storage batteries. In Europe about 30 % of the total antimony production originates from this source with a decreasing tendency [158]. In 1997 the smelter production of primary antimony in the United States was 26 700 t. At the same time the antimony production from secondary raw materials in the United States was 7550 t (22 %) [20].

Some important producers of antimony metal are as follows [158]:

Argentina	Calles y Chernitsky SA, Buenos Aires
Belgium	Campine SA, Brussels; NV Union Minieres SA (UM), Brussels
China	Dachang Ming Administration, Dachangzhen; Guang long Antimony Smelter, Hsikwangshan Antimony Works, Hsikwanshan Area; Hunan Non-Ferrous Metals Industry Corp, Changsha
Germany	Metaleurop GmbH, Hannover
France	Mines de la Lucette, Paris
Greece	Greek Metallurgical & Mining Co of Lavrion SA - Emmel, Lavrio
U.K.	MCP Wafer Technology Ltd, Milton Keynes
India	The Metal Powder Co Ltd, Thirumangalam
Japan	Dowa Mining Co Ltd, Tokyo; Nippon Rare Metal Inc, Yokohama; Nippon Seiko KK, Shimagawa-ku

Kyrgistan	Kadamdzhaisky Antimony Works, Oshsk
Netherlands	Billiton Whitemetaal BV, Naarden
Peru	Centromin Peru SA-Empresa Minera del Centro del Peru SA, Lima
Romania	Intreprindera Metalurgica de Metale Neferoase (Metallurgical Enterprise for Non-ferrous Metals), Copsa-Mica
Serbia	Trepca Works-Sour Rudarsko-Metalursko-Hemijski Kombinat Olova i Cinka, Titova; Zajaca Rudarsko Topionicarski Basen, Loznica
Slovakia	Vajskova Antimony Smelter, Vajskova
South Korea	Il Sung Antimony Co Ltd, Seoul
Turkey	Mutlu Akü ve Malzemeleri SAS, Istanbul; Ozdemir Antimuan Madenleri Ltd Sti, Istanbul
U.S.A.	Metalspecialties Inc, Fairfield; Ney Smelting & Refining, Brooklyn

Some important producers of antimony oxide are listed in the following:

Argentina	Calles y Chernitsky SA, Buenos Aires
Belgium	Campine SA, Brussels
Bolivia	Enaf-Empresa Metalurgica Vinto, Oruro
China	Guang Long Antimony Smelter
France	Mines de Lucetta, Paris
Italy	Ginatta SpA, Turin
Japan	Mitsui Mining & Smelting Co Ltd, Tokyo
Russia	Novosibirsk Rare Metal Works, Novosibirsk
Serbia	Zajaca Rudarsko Topinicarski Basen, Loznica
South Africa	Antimony Products (Pty) Ltd, Johannesburg
U.S.A	Amspec Chemical Corp, Gloucester City Asarco Inc, New York Ney Smelting & Refining, Brooklyn

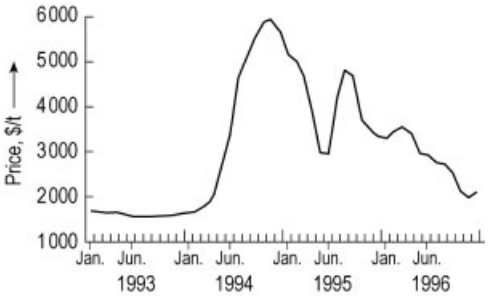

Figure 9. Prices for antimony metal [159]

Figure 9 shows the price ranges for antimony from 1993 to 1996. The price has undergone large variations, especially in 1994. The dramatic increase of the antimony price in 1994 from 1700 $/t in January to 5900 $/t in November was caused by difficulties in the supply from China, the states of the former Soviet Union, and Bolivia. Again China's role was the main factor influencing the antimony market. Shortly after the price peak in 1994/95 the antimony price again reached an average value of about 2000 $/t. Antimony production in China is expected to decrease by up to 40 % in 1997/98 due to the high production costs of 2500 to 2800 $/t Sb. The strong dependency of the antimony price on China has led to plans for the reopening of older antimony mines (e.g., Thomson Falls). In addition, a new antimony trioxide plant will be opened in Canada in 1998. New sources might result in a greater degree of stability in the antimony industry worldwide [18, 20, 160, 161].

Future growth and demand for antimony will continue to depend on trends in consumption of antimony trioxide in flame retardants, which account for 45 % of the world antimony demand. Other growth areas are the use of antimony trioxide in polymerization catalysts for PET production and of sodium antimonate as decolorizer in glass for cathode ray tubes [162].

16. Toxicology

Antimony, which is considered a nonessential element, is comparable in its toxicological behavior to arsenic and bismuth [163]. In analogy to arsenic, trivalent antimony compounds generally are more toxic than the pentavalent compounds [164]. Poisoning with antimony and its compounds can result from acute and chronic exposure, especially from exposure to airborne particles in the workplace; to a minor extent, exposure occurs through treatment of tropical diseases with antimony compounds.

Toxicokinetics. Most antimony compounds, mainly those with poor water solubility, are absorbed only slowly from the gastrointestinal tract. Trivalent compounds especially tend to accumulate in the human body, because they are excreted very slowly via urine and feces [164]. Antimony and its compounds react readily with mercapto groups in various cellular constituents, especially in enzymes, blocking their activity. After acute and chronic intoxications, the highest levels of antimony were found in lungs, liver and kidneys [165].

Acute Toxicity. Acute oral poisoning by antimony compounds after ingestion of contaminated drinks was observed in 150 children; the symptoms were nausea, diarrhea, and vomiting

[166]. Acute intoxications with antimony trioxide, Sb_2O_3, lead to vomiting, diarrhea and abdominal pains; in cases of subchronic and chronic intoxication symptoms such as nausea, cough, headache, as well as pains of the stomach and muscles were observed. Acute respiratory exposure of seven workers to 70 – 80 mg/m^3 of antimony trichloride, $SbCl_3$, resulted in irritation of the upper respiratory tract [167]; in five cases disturbances of the gastrointestinal function and loss of appetite were observed. Antimony pentachloride, $SbCl_5$, caused severe pulmonary edema in three cases, two of them being lethal [168]. Stibine, SbH_3, a highly toxic, relatively unstable gas with an unpleasant odor, causes symptoms of the central nervous and circulatory systems, such as nausea, vomiting, headache, slow breathing, weak pulse, hemolysis, hematuria, abdominal pain, and death [169, 170]. Stibine is generated if nascent hydrogen can react with antimony in acid, e.g., in lead – acid batteries, where antimony may be a component of the battery plates [171].

People who are exposed to antimony and its compounds often show transient pustular skin eruptions [172]. These "antimony spots" mainly occur in those regions of the skin that are exposed to heat and sweat. In another case, three employees exposed to antimony metal dust and antimony trioxide fumes in a melting process developed a generalized eruption of follicular papules and pustules [173].

Antimony trifluoride and pentafluoride are highly toxic and irritant to the skin; they cause eczematous eruptions and inflammation of mucous membranes of nose and throat.

Chronic Toxicity. Chronic poisoning after oral ingestion of antimony compounds has been observed in many cases of clinical treatment with antimony compounds. These chronic intoxications resulted in nausea and vomiting [174]; liver damage, with increased plasma values of aspartate-aminotransferase and alanine-aminotransferase [175]; changes in the electrocardiogram, indicating a cardiotoxic effect of these compounds [176]; and even sudden death [177].

Most cases of chronic respiratory intoxications result from exposure to airborne particles in the workplace. In a study involving 78 workers in antimony smelters, symptoms such as soreness, nosebleeds, rhinitis, pharyngitis, pneumonitis, and tracheitis could be observed after exposure times of up to five months, the air containing antimony trioxide, Sb_2O_3, in a concentration of 4.7 – 11.8 mg/m^3 [172]. Exposure to Sb_2O_3, which is used in the glass-producing industry as a refining agent and glass colorant, leads to significantly increased levels in the urine of 1.5 to 15.7 µg/L (median: 5.0 µg/L). Due to the fast renal excretion of antimony, its determination in urine is useful for biological monitoring [178]. Long-term exposures often result in lung X-rays that are similar to those seen in pneumoconiosis [179, 180]. Workers who had been exposed to Sb_2O_3 for up to 28 years had pneumoconiosis (21 %) and symptoms of emphysema (42 %) [181]. Brieger et al. [182] extensively examined 125 workers who had inhaled antimony trisulfide, Sb_2S_3, in air at concentrations of 0.6 – 5.5 mg/m^3 for 8 – 24 months. An increased morbidity and mortality were observed among these workers. Changes in the electrocardiogram were seen in 37 cases out of 75. In addition, a significantly higher incidence (6 %) of peptic ulcers occurred within the group of workers exposed to antimony compared to the incidence within the total plant population (1.5 %). During the period of this study, six workmen died suddenly, and two others died from chronic heart diseases. After extensive exposure to antimony fumes, workers showed gastrointestinal symptoms, such as diarrhea, vomiting, and abdominal convulsions [172].

Mutagenicity, Carcinogenicity. Antimony has been demonstrated to be mutagenic in bacteria and phages, and to induce chromosomal aberrations and abnormal cell divisions in animal and plant cells [183]. Based on the data of animal experiments, antimony trioxide is a suspected carcinogen in humans [184]. There are limited epidemiologic data concerning a possible carcinogenicity of antimony compounds in humans [185, 186].

Toxicological Data.

Antimony, Sb, MAK 0.5 mg/m^3 (measured as *einatembarer Staubanteil*, inhalable fraction of dust) [184]; limit value for drinking water 0.01 mg Sb/L [187]; LDLo 100 mg/kg (rat, intraperit.) [188].

Antimony hydride, stibine, SbH_3, MAK 0.1 ppm (0.5 mg/m^3), LCLo 92 ppm (guinea pig, inhalation, 1 h) [189].

Antimony trifluoride, SbF_3, TLV 0.5 mg/m^3 (as Sb), LDLo 100 mg/kg (guinea pig, oral) [190].

Antimony trichloride, $SbCl_3$, TLV 0.5 mg/m^3 (as Sb), TDLo 73 mg/m^3 (human, inhalation) [167].

Antimony pentachloride, $SbCl_5$, TLV 0.5 mg/m^3 (as Sb), LD_{50} 1115 mg/kg (rat, oral) [191], causes mutation in *Bacillus subtilis* at 30 µL/disk [192].

Antimony trioxide, Sb_2O_3, carcinogenic potential suspected for humans, TLV 0.5 mg/m^3 (as Sb), TCLo 4.2 mg/m^3 during 52 weeks, carcinogenic (rat, inhalation) [193].

Antimony pentoxide, Sb_2O_5, LDLo 4000 mg/kg (rat, intraperit.) [188].

Antimony trisulfide, Sb_2S_3, TLV 0.5 mg/m^3 (as Sb), TCLo 580 µg/m^3 during 35 weeks (humans, inhalation) [182], LDLo 1000 mg/kg (rat, intraperit.) [188].

Antimony pentasulfide, Sb_2S_5, TLV 0.5 mg/m^3 (as Sb), LD_{50} 1500 mg/kg (rat, intraperit.) [188].

17. References

General References

1. *Gmelin*, Sb (system no. 18), Main A3 (1950) 303 – 351.
2. V. Tafel: *Lehrbuch der Metallhüttenkunde*, vol. 2, Hirzel, Leipzig 1951.
3. C. Y. Wang: *Antimony*, C. Griffin & Co., London 1952.
4. W. Wendt: *Antimon und seine Verhüttung*, Deuticke, Vienna 1950.
5. R. C. Weast (ed.): *Handbook of Chemistry and Physics*, 56th ed., CRC Press, Cleveland, Ohio, 1975.
6. W. J. Maeck: *The Radiochemistry of Antimony*, U. S. At. Energy Comm. 1961.
7. F. Pawlek: *Metallhüttenkunde*, De Gruyter, Berlin – New York 1983.
8. *Ullmann*, 4th ed., **8**, pp. 1 – 18.
9. *Kirk-Othmer*, 2nd ed. **2**, pp. 562 – 588.
10. B. Tougarinoff: "Nouveaux métaux et matériaux," *Rev. Soc. R. Belge Ing. Ind.* 1966 no. 9, 10.
11. E. Bonnier, M. Charveriat: "Purification de l'antimoine par sublimation," *ATB Metall.* **5** (1965) 319.
12. *Ullmann*, 3rd ed., **3**, pp. 806 – 827.
13. A. F. Taggart: *Handbook of Mineral Dressings*, Ores and Industrial Minerals, J. Wiley & Sons, New York 1976.

Specific References

14. *Ullmann*, 5th ed., **vol. A3**, 55 – 76.
15. N. C. Norman (ed.): *Chemistry of Arsenic, Antimony and Bismuth*, Blackie Academic & Professional, London 1998.
16. *Römpps Chemie-Lexikon*, 8th ed., **vol. 1.**
17. D. L. Lide, Jr. Lide (eds.): *CRC Handbook of Chemistry and Physics*, 79th ed., CRC Press, Boca Raton, FL 1998.
18. T. O. Llewellyn: in "Antimony", *Minerals Yearbook*, **vol. 1: Metals and Minerals**, United States Department of the Interiour Buro of Mines, 1991, pp. 227 – 238.
19. D. R. Davis, D. B. Paterson, D. H. C. Griffiths: "Antimony in South Africa", *J. S. Afr. Inst. Min. Metall.* **86** (1986) no. 6, 173 – 193.
20. J. F. Carlin, Jr. "Antimony", *U.S. Geology Survey, Minerals Information*, 1997, pp. 1 – 6.
21. "Antimon — von der Augenschminke zum Flammschutzmittel", *Neue Züricher Zeitung* **129** (1989) June 7th.
22. *Metallstatistik* **84** (1997) 47, 492.
23. Societe de l'Industrie Minerale (ed.): *World Mining and Metals Yearbook*, Paris 1997.
24. United States Geological Survey through: www.indexmundi.com/en/commodities/minerals/antimony/antimony
25. G. Dill: "Antimon in Bolivien — Ein wenig bekanntes Element in einem wenig bekannten Land", *Bergbau* (1994) no. 11, 494 – 500.
26. M. Kemal, H. Köse, F. Simsir; "Der Antimonerzbau in der Türkei", *Erzmetall* **42** (1989) no. 1, 36 – 40.
27. E. W. Mayer in H. Schranz (ed.): *Flotation*, Hirzel, Leipzig 1931, p. 375.
28. U. S. Popova, R. L. Popov, V. G. Nesterov: "The Effect of Non-metalliferrous Components in Antimony Ores on the Flotation of Antimonite," *Sb. Nauchn. Tr. Sredneaziat. Nauchno Issled. Proektn. Inst. Tsvet. Metall.* 1972, no. 5, 75 – 80.
29. K. Kijayakumer, K. K. Majumdar, *J. Mines Met. Fuels* **20** (1972) no. 11, 342 – 346.
30. H. Grothe, L. Engel, H. Hock, K. Löhberg, K. Schönert: *Hüttentechnik*, **1**, Rowohlt, Hamburg-Reinbek 1972, p. 36.
31. P. M. Swash: "A mineralogical Examination of an Antimony-Bearing Gold Ore and its Benefication Products", *Mintek Report No. M281, 10th Nov. 1986*, Council for Mineral Technology, SA, pp. 2 – 7.

32. G. Morizot, J. M. Winter, G. Barbery: *Volatilization Chloridization with Calcium Chloride of Complex Sulphide Minerals and Concentrates*, published in Complex Sulphide Ores, papers presented at Complex Sulphide Ores Conference by the Inst. Min. Metall, Rome 1980, pp. 151 – 158.
33. G. Lindkvist, A. Hlolmström: "Roasting of Complex Concentrates with High Arsenic Contents", in H. Y. Sohn et al., (eds.): *Advances in Sulfide Smelting*, TMS 1983, pp. 451 – 472.
34. I. Imris et al.: "Behaviour of Antimony During Roasting of Tetrahedrite Concentrates", in *Extraction Metallurgy '85, Symposium, London 9 – 12 Sept. 1985* IMM, p. 1015 – 1033.
35. Ph. Cauwe, Ph. Minet, R. Sheridan: "Selective Roasting of Complex Sulfide Material", in: H. Y. Sohn et al. (eds.): *Advances in Sulfide Smelting*, The Metallurgical Society of AIME, 1983.
36. G. A. Brooks, W. J. Rankin, N. B. Gray: "Thermal Separation of Arsenic and Antimony Oxides", *Met. Trans. B* **25B** (1994) 873 – 884.
37. G. A. Brooks, W. J. Rankin, R. Papp: "The Treatment of Arsenic and Antimony Oxide Fumes", in M. Nilmani, T. Lehner, W. J. Rankin (eds.): *Pyrometallurgy for Complex Materials and Wastes*, The Minerals, Metals & Materials Society, Warrendale, PA 1994, pp. 277 – 287.
38. *Ann. Meet. Am. Ind. Hyg. Assoc.* **20**(1980) 1.
39. L. Komorova, I. Imris: "Recovery of Antimony from Tetrahedrite Calcines", in: *European Metal Conference EMC'94*, pp. 164 – 177.
40. P. Balaz, R. Kammel, M. Achimovicova: "Selektive hydrometallurgische Gewinnung von Sb, Hg und Ag aus mechano-chemisch behandelten Tetrahedritkonzentraten" (engl.), *Metall* **48** (1994) no. 3, 217 – 220.
41. D. A. Pazour, *World Min.* **33** (1980) no. 6, 42 – 47.
Min. Mag. **134** (1976) no. 5, 369 – 374.
42. H. W. Burkey, US 1654527, 1926;
US 1654528, 1926.
43. AB Metallreduction, NO 48804, 1929.
44. G. N. Kirsebom, GB 315811, 1929.
45. C. W. Jensen: "Antimony Smelting in Czechoslovakia", *The Mining Magazine* (1950) Oct. 217 – 220.
46. C. W. Jensen: "Antimony Smelting in Yugoslavia", *The Mining Magazine* (1951) Jan. 20 – 22.
47. "New Smelter to Produce most of U.S. Antimony" *Engineering and Mining Journal* **150** (1949) no. 7, 140 – 141.
48. J. B. Huttl: "Yellow Pine is Expanding its Output of Stratetic Metals", *Engineering and Mining Journal* **153** (1952) no. 5, 72 – 77.
49. R. L. Kulpaca, J. C. Archibald: "Mechanisition Program Results in Savings at Laredo Antimony Smelter", *JOM* (1953) Jun. 786 – 788.
50. T. Mederer, T. Wippermann: "Mineralogische und geochemische Zusammensetzung von Antimonschlacke und Antimonerzen und ihr Freisetzungsverhalten für Schwermetalle", *Z. angew. Geol.* **40** (1994) no. 1, 38 – 40.
51. N. Akkarapattanagoon, Y. Takeda, A. Yazawa: "Distribution of Impurities between Soda Slags and Molten Antimony", *Metallurgical Review of MMIJ* **7** (1990) no. 1, 31 – 47.
52. C. Y. Wang, G. C. Riddel: "Antimony — its Metallurgy and Refining in Recent Years", Proc. A.I.M.E. meeting, Cleveland, Oct. 1944, 446 – 461.
53. US 2 295 039, 1942.
54. W. H. Dennis: Antimony Production", *Mine and Quarry Engineering* (1949) Sept. 289 – 296.
55. C. Y. Wang: "Calculating Charges for the Precipitation Method of Antimony Smelting", *Engineering and Mining Journal* **150** (1949) no. 3.
56. W. Wendt: "Antimony Smelting", *Metal Industry* (1948) Oct. 303 – 305.
57. W. Wendt: "Antimony Smelting, the Development of the Blast Furnace Process", *Metal Industry* (1948) Oct. 329 – 330.
58. C. Hsiaohong, P. Rongquiu, H. Fucheng: "Study on the Bath Smelting for Antimony Sulfide Concentrates", *Erzmetall* **46** (1993) no. 6, 377 – 380.
59. N. P. Shashin in N. Muratsch (ed.): *Handbuch des Metallhüttenmannes*, **1**, VEB Verlag Technik, Berlin 1954.
60. C. L. Mantell: *Electrochemical Engineering*, 4th ed., McGraw-Hill, New York 1960, pp. 232 – 234, 238.
61. Chung Yu Wang, Chung Yu Wang, G. C. Riddell in O. M. Liddell (ed.): *Handbook of Nonferrous Metallurgy*, McGraw-Hill, New York 1945.
62. W. Schopper in G. Eger (ed.): *Handbuch der technischen Elektrochemie. Die technische Elektrolyse wäßriger Lösungen*, **vol. 1**, part 1, Akademische Verlags GmbH, Leipzig 1961.

63. L. A. Meza-Saucedo: "Hydrometallurgical Treatment of Colombian Stbnite Concentrate", in D. R. Gaskell, J. P. Hayer, J. E. Hoffmann, P. J. Macker (eds.): *The Reinhard Schumann International Symposium on Innovative Technology and Reactor Design in Extractive Metallurgy*, The Metallurgical Society, Warrendale, PA 1986, pp. 1003 – 1017.
64. W. C. Holmes, *Eng. Min. J.* **145** (1944) no. 3, 54 – 58.
65. J. Koster, M. B. Roger: "Electrolyte Recovery of Antimony from Antimonial Gold Ores," *Rep. Invest. U. S. Bur. Mines* RI 3491 (1946).
66. A. E. Albrethsen, M. L. Hollander: "The ASARCO Antimony Electrowinning Process," *TMS Pap. Sel.* **A 79-64** (1979) 9.
67. B. Balmforth, T. A. Leese, P. A. Tasker: "Applications of Selective Solvent Extraction in Waste Minimisation and Metal Recovery", in V. Ramachandran, C. C. Nesbitt (eds.): *Second International Symposium on Extraction and Processing for the Treatment and Minimization of Wastes*, The Minerals, Metals and Materials Society, Warrendale, PA 1996, 85 – 98.
68. P. van Negen, R. Maes, C. Coekelbergs: "The Hoboken Complex Lead-Copper Smelter A Survey of Main Technological Developments During the Last Decade", in T. S. Mackey, R. D. Prengaman (eds.): *Lead-Zinc '90*, The Minerals, Metals and Materials Society 1990.
69. J. Minoura, Y. Maeda: "Current Operation at Kosaka Smelter and Refinery", *Metallurgical Review of MMIJ* **1** (1984) no. 2, 138 – 156.
70. E. Freni, *Erzmetall* **23** (1970) 128 – 132.
71. Y. Sugawara, S. Ishii: "Current Operations at Takehara Refinery", *Metallurgical Review of MMIJ* **2** (1985) no. 1, 90 – 101.
72. G. Brecka, K. Hein, H. J. Lange, P. Paschen: "A Pyrometallurgical Alternative: the Refining Electrolysis of Lead and Solder", *JOM* (1997) no. 4, 62 – 65.
73. G. Schenker: "Lead Recycling from Battery Scrap and Other Raw Materials in Metaleurop's Lead Smelting Plant in Onker", in T. S. Mackey, R. D. Prengman (eds.): *Lead Zinc '90*, The Minerals, Metals and Materials Society, Warrendale, PA 1990.
74. T. Oishi et al.: "Removal of Antimony by Volatilization from Industrial Copper Matte and White Metal", *Metallurgical Review of MMIJ* **3** (1986) no. 2, 14 – 21.
75. Y. Mori, T. Kimura: "Cleaning of Copper Converter Slag by Coal Injection", *Metallurgical Review of MMIJ* **3** (1986) no. 3, 141 – 154.
76. A. K. Kyllo, G. G. Richards, D. Verhelst: "Analysis of Smelter Dusts" in V. Ramachandran, C. C. Nesbitt (eds.): *Second International Symposium on Extraction and Processing for the Treatment and Minimization of Wastes*, The Minerals, Metals & Materials Society, Warrendale, PA
77. B. Zhao, Q. Zhao, N. J. Themelis: "Removal of As, Sb and Bi from molten Copper by Injection of SF6", in T. Lehner (ed.): *Co-products and Minor Elements in Non-Ferrous Smelting*, The Minerals, Metals & Materials Society, Warrendale, PA 1995, pp. 39 – 52.
78. Y. Okajima, T. Kimura: "Recovery of Valuable Metals from Copper Converter Slag" *Metallurgical Review of MMIJ* **3** (1986) no. 2, 90 – 101.
79. I. V. Kojo et al.: "Distribution Equilibria of Lead, Arsenic and Antimony Between Copper and Sodium Silicate Slags", *Erzmetall* **40** (1987) no. 3, 138 – 144.
80. T. Nakamura, F. Noguchi, Y. Ueda: "Fire Refining of Crude Copper by Alkaline Carbonate Fluxes", *Metallurgical Review of MMIJ* **3** (1986) no. 2, 102 – 116.
81. T. Lehner: "Smelter Nasties — A Changing World", in M. Nilmani, T. Lehner, W. J. Rankin (eds.): *Pyrometallurgy for Complex Materials and Wastes*, The Minerals, Metals & Materials Society, Warrendale, PA 1994, pp. 177 – 191.
82. A. Yazawa, Y. Takeda: "Dissolution of Arsenic and Antimony in Slags", *Metallurgical Review of MMIJ* **3** (1986) 2, 117 – 130.
83. J. Gortais et al.: "Equilibrium Distribution of Fe, Ni, Sb and Sn Between Liquid Cu and a CaO-Rich Slag", *Met. Trans. B* **25B** (1994) Oct. 645 – 651.
84. R. U. Pagador, M. Hino: "Arsenic and Antimony Activities in Copper, Nickel and Their Mattes", *Metallurgical Review of MMIJ* **13** (1996) no. 1, 166 – 185.
85. S. Abe et al.: "A Method for Electrorefining Copper Anodes Contaminated with High Levels of As, Sb and Bi", *Metallurgical Review of MMIJ* **3** (1986) no. 3, 164 – 174.
86. A. K. Biswas, W. G. Davenport: *Extractive Metallurgy of Copper*, 3rd ed., Pergamon Press, Oxford–New York 1994.
87. K. Ando, N. Tsuchida: "Recovering Bi and Sb from Electrolyte in Copper Electrorefining", *JOM* (1997) Dec. 49 – 51.

88. D. S. Flett: "Solution Purification", *Hydrometallurgy* **30** (1992) 327 – 344.
89. W. Schwab, H. Kroke: "Purification of Copper Tankhouse Electrolytes by Solvent Extraction", in *Arsenic Metallurgy Fundamentals and Applications, TMS Annual Meeting, Jan. 1988*, pp. 249 – 262.
90. Y. Oda et al.: "Tamano Chelating Resin Adsorption Process", *Metallurgical Review of MMIJ* **3** (1986) no. 2, 155 – 163.
91. T. Nishimura, K. Tozawa: "Behaviour of Antimony and Arsenic in Sulfuric Acid Solution", *Metallurgical Review of MMIJ* **3** (1986) no. 2, 131 – 145.
92. H. K. Lin, X. Wu: "Behaviour of Antimony (III) during Copper Electrowinning in Chloride Solutions", *Met. Trans. B*, **27B** (1996) April, 157 – 162.
93. B. Balmforth, T. A. Leese, P. A. Tasker: "Applications of Selective Solvent Extractants in Waste Minimisation and Metal Recovery", in V. Ramachandran, C. C. Nesbitt (eds.): *Second International Symposium on Extraction and Processing for the Treatment and Minimization of Wastes*, The Minerals, Metals & Materials Society, Warrendale, PA 1996.
94. J. T. I. Poijärvi: "By-Products of Copper and Nickel Production", in T. Lehner (ed.): *Co-Products and Minor Elements in Non-Ferrous Smelting*, The Minerals, Metals & Materials Society, Warrendale, PA 1995, pp. 53 – 63.
95. M. A. Fernandez et al.: "Selective Leaching of Arsenic and Antimony Contained in the Anode Slimes from Copper Refining", *Hydrometallurgy* **41** (1996) 255 – 267.
96. W. C. Cooper: "The Treatment of Copper Refinery Anode Slimes", *JOM* (1990) Aug. 45 – 49.
97. J. E. Hoffmann: "The Wet Chlorination of Electrolytic refinery Slimes", *JOM* (1990) Aug. 50 – 54.
98. W. G. Clark, B. H. Moore, *Min. Mag.* **43** (1930) 58.
99. N. J. Tschumarow, G. M. Goutscharowa, *Tsvetn. Met.* **16** (1941) 43.
100. N. Akkarapattanaangoon, K. Itagi, A. Yazawa: "Phase Relations and Distribution of Minor Elements Between Sb-S-O Slags and Liquid Antimony", *Erzmetall* **43** (1990) no. 2, 63 – 69.
101. N. Akkarapattanaangoon, K. Itagi, A. Yazawa: "An Ionic Model for the Distribution of Minor Elements between Sb-S-O Slags and Metallic Antimony", *Erzmetall* **43** (1990) no. 3, 117 – 125.
102. T. Rondelli, IT 380529, 1940.
103. D. Jevtić, D. Vitorović, *Ind. Eng. Chem. Prod. Res. Dev.* **13** (1974) 275 – 279.
104. E. Groschuff, *Z. Anorg. Allg. Chem.* **103** (1918) 164 – 188.
105. Y. Y. Lure, N. A. Filpparu, *Zavod. Lab.* **14** (1948) 159.
106. G. R. Smithson et al., *J. Met.***17** (1966) 1037 – 1046.
107. W. G. Pfann: *Zone Melting*, 2nd ed., J. Wiley & Sons, New York 1966.
108. H. Schildknecht: *Zonenschmelzen*, Verlag Chemie, Weinheim 1964.
109. Zd. Trousil, *Chemie (Prag)* **9** (1957) 633.
110. H. Schell, *Z. Metallkd.* **46** (1955) 58.
111. M. Tanenbaum, A. J. Goss, W. G. Pfann, *J. Met.* **7** (1955) 297 – 303.
112. M. Tanenbaum, A. J. Goss, W. G. Pfann, *J. Met.* **6** (1954) 762 – 763.
113. M. Wilhelm, DE 1155914, 1964.
114. V. N. Vigdorovich, V. S. Ileva, L. Y. Krol, *Izv. Akad. Nauk SSSR Otd. Tekh. Nauk Metall. Topl.* 1960, no. 1, 44.
115. W. A. Kusnetzow: "Ion Exchange in Hydrometallurgy," (Russ.) in *Metallurgy of Alloys and Rare Metals 1967*, Natl. Inst. for Text Books and Tech. Information, Moscow 1968, pp. 81 – 105.
116. E. Scheffler, S. Ziegenbalg, *Freiberg, Forschungsh. B* **99** (1964) 43 – 63.
117. Gmelin, Sb (system no. 18), Org.-Sb 1 – 3 (1981 – 1982).
118. G. O. Doak, L. D. Freedman: *Organometallic Compounds of Arsenic, Antimony, and Bismuth*, Interscience, New York 1970.
119. S. Samaan: *Houben-Weyl*, Metallorganische Verbindungen As, Sb, Bi, XIII/8, G. Thieme Verlag, Stuttgart 1978.
120. S. Patai (ed.): *The Chemistry of Organic Arsenic*, Antimony and Bismuth Compounds, Wiley, Chichester 1994.
121. W. Stamm, A. Breindel, *Angew. Chem.* **76** (1964) 99.
122. J. W. Dale, H. J. Emeleus, R. N. Hazeldine, J. H. Mass, *J. Chem. Soc.* 1957, 3708 – 3713.
123. M. Nunn, D. B. Sowerby, D. M. Wesolek, *J. Organometal. Chem.* **251** (1983) C45 – C47.
124. Z. M. Manulkin, A. N. Tatarenko, F. Yu. Yusupov, *Dokl. Akad. Nauk SSSR* **88** (1953) 687 – 690.
125. H. J. Breunig, W. Kanig, A. W. Soltani-Neshan, *Polyhedron*, **2** (1983) 291 – 292.

126. K. Issleib, B. Hammann, L. Schmidt, *Z. Anorg. Allgem. Chem.* **339** (1965) 298 – 303.
127. W. S. Sheldrick, C. Martin, *Z. Naturforsch.* **46b** (1991) 639 – 646.
128. C. J. Carmalt, D. Walsh, A. H. Cowlwy, N. C. Norman, *Organometallics*, **16** (1997) 3597 – 3600.
129. H. J. Breunig, M. E. Ghesner, E. Lork, *Z. Anorg. Allg. Chem.* **631** (2005) 851 – 856.
130. H. J. Breunig, M. A. Mohammed, K. H. Ebert, *Z. Naturforsch.* **49b** (1994) 877 – 880.
131. H. Schmidt, *Justus Liebigs Ann. Chem.* **429** (1922) 123 – 152.
132. G. Wittig, K. Torssell, *Acta Chem. Scand.* **7** (1953) 1293 – 1301.
133. G. Wittig, K. Claus, *Justus Liebigs Ann. Chem.* **577** (1952) 26 – 39.
134. H. Schmidbaur, J. Weidlein, K. H. Mitschke, *Chem. Ber.* **102** (1964) 4136 – 4146.
135. G. O. Doak, G. G. Long, C. D. Freedman, *J. Organomet. Chem.* **12** (1968) 443 – 450.
136. H. Schmidbaur, K. H. Mitschke, *Chem. Ber.* **104** (1971) 1837 – 1841.
137. H. Schmidbaur, K. H. Mitschke, *Z. Anorg. Allgem. Chem.* **386** (1971) 139 – 146.
138. K. M. MacKay, D. B. Sowerby, W. C. Young, *Spectrochim. Acta* Part A 24 (1968) 611 – 631.
139. L. M. Yagupolskii, L. M. Popov, V. I. Kondeatenko, N. V. Korsunskii, B. L. Aleinikov, *Zh. Obshch. Khim.* **45** (1975) 459.
140. H. A. Meinema, *J. Organomet. Chem.* **51** (1973) 223 – 230.
141. G. O. Doak, G. G. Long, L. D. Freedman, *J. Organomet. Chem.* **4** (1965) 82 – 91.
142. N. Tokitoh et al., *J. Am. Chem. Soc.* **120** (1998) 209 – 213.
143. H. J. Breunig, K. Häberle, M. Dräger, T. Severengiz, *Angew. Chem.* **97** (1985) 62 – 63.
144. Hoechst, DE 269843, 1912; DE 269744, 1912; *Chem. Abstr.* **1914,** 2221; GB 17533, 1913; *Chem. Abstr.* **1914,** 991; GB 19778, 1913; *Chem. Abstr.* **1915,** 694.
145. F. A. Paneth, H. Loleit, *J. Chem. Soc.* 1935, 366 – 371.
146. W. Hewertson, H. R. Watson, *J. Chem. Soc.* 1962, 1490 – 1494.
147. A. J. Ashe, M. D. Gordson, *J. Am. Chem. Soc.* **94** (1972) 7596 – 7597.
148. P. B. Hitchcock, C. Jones, J. F. Nixon, *Angew. Chem.* **107** (1995) 522 – 523, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 492 – 493.
149. M. Wieber, J. Walz, *Z. Naturforsch.* **45b** (1990) 1615 – 1617.
150. G. O. Doak, J. M. Summy, *J. Organomet. Chem.* **55** (1973) 143 – 151.
151. J. Bordner, G. O. Doak, T. S. Everett, *J. Am. Chem. Soc.* **108** (1986) 4206 – 4213.
152. E. Beveridge in R. J. Schnitzer, F. Hawking (eds.): *Experimental Chemotherapy*, vol. 1, Academic Press, New York 1963.
153. M. & T. Chemicals, Inc., US 3287210, 1966 (J. R. Leebrink).
154. G. G. Long, G. O. Doak, L. D. Freedman, *J. Am. Chem. Soc.* **86** (1964) 209 – 213.
155. E. Merian: *Metalle in der Umwelt*,Verlag Chemie, Weinheim 1984, p. 310. C. Veillon in J. D. Wineforder (ed.): *Trace Analysis: Spectroscopic Methods for Elements*, J. Wiley & Sons, New York 1976, pp. 164 – 165.
156. J. Vehlow, L. Birnbaum, W. Köppel: "Arsen und Antimon in der Abfallverbrennung", *AJ* (1997) no. 3, 9 – 19.
157. F. Feichtinger: "Wichtige Metalle aus Südafrika: Antimon, Chrom, Mangan und Vanadium", *Metall* **47** (1993) no. 7, 678 – 681.
158. H. P. Münster (ed.): *Taschenbuch des Metallhandels*, 9th ed., Hüthig Verlag, Heidelberg 1997, pp. 156 – 159.
159. R. Serjeantson (ed.): *Metal Bulletin's Prices and Data 1997*, Metal Bulletin Books Ltd., Worcester Park, Surrey 1997, pp. 56 – 59.
160. "Viel weniger Antimon aus China im Jahr 1997", *Metall*, **51** (1997) no. 12, 689.
161. "Preisanstieg auf dem Antimonmarkt", *Metall*, **49** (1995) no. 6, 392.
162. Roskill: "The Economics of Antimony 1997", Roskill Information Services Ltd., 1998; see also http://www.roskill.com/roskill/scripts/search.asp
163. *Ullmann*, 4th ed., **8**, pp. 16 – 18.
164. L. S. Godman, A. Gilman: *The Pharmacological Basis of Therapeutics*, 5th ed., Macmillan Publishing Company, New York 1975, pp. 929 – 930.
165. H. Schicha, K. Kasperek, L. E. Feinendegen, V. Siller, H. J. Klein, *Beitr. Pathol.* **146(1)** (1972) 55 – 62.
166. N. Werrin, Hussock, *Food Drug Off. U.S. Quart. Bull.* **27** (1963) 38 – 45.
167. P. J. Taylor, *Br. J. Ind. Med.* **23** (1966) 318 – 321.
168. E. M. Cordasco, F. D. Stone, *Chest* **64** (1973) 182 – 185.
169. H. E. Stokinger, in G. D. Clayton, F. E. Clayton (eds.): Patty's Industrial Hygiene and Toxicology, 3rd ed., vol. 2 A, Wiley-Interscience, New York 1981, pp. 1505 – 1517.

170. Merck Index, 11th ed., Merck and Co., Inc., Rahway, N.J., 1989, p. 1387.
171. E. Browning: *Toxicity of Industrial Metals*, 2nd ed., Appleton-Century Crofts, New York 1969, pp. 23 – 38.
172. L. E. Renes, *Arch. Ind. Hyg. Occup. Med.* **7** (1953) 99 – 108.
173. G. P. White, C. G. Mathias, J. S. Davin, *J. Occup. Med.* **35** (1993) no. 4, 392 – 395.
174. M. R. Pedrique, S. Barbera, N. Ercoli, *Ann. Trop. Med. Parasitol.* **64** (1970) 225 – 261.
175. A. W. Woodruff, *Ann. N. Y. Acad. Sci.* **160** (1969) 650 – 655.
176. D. W. Sapire, N. H. Silverman, *South African Med. J.* **44** (1970) 948 – 950.
177. A. A. El Halawani, *Bull. Endemic Dis.* **10** (1968) 123 – 133.
178. R. Ludersdorf, A. Fuchs, P. Mayer, G. Skulsuksai, G. Schäcke, *Int. Arch. Occup. Environ. Health* **59** (1987) no. 5, 469 – 474.
179. R. C. Browne, *Br. J. Ind. Med.* **25** (1968) 187 – 193.
180. V. Potkonjak, M. Pavlovich, *Int. Arch. Occup. Environ. Health* **51** (1983) no. 3, 199 – 207.
181. I. Klucik, A. Juck, J. Gruberová, *Pracovni Lékarstvi* **14** (1962) 363 – 368 (in Czech; information from Scientific Reports of Industrial Hygiene and Occupational Diseases in Czechoslovakia **7**, 56; English summaries only).
182. H. Brieger, C. W. Semisch, J. Stasney, M. A. Piatnek, *Ind. Med. Surg.* **23** (1954) 521 – 523.
183. C. P. Flessel, *Adv. Exp. Med. Biol.* **91** (1977) 117 – 128.
184. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentrationen (MAK) 1996*, **part III B**, VCH Verlagsgesellschaft, Weinheim 1996.
185. T. M. Schnorr, K. Steenland, M. J. Thun, R. A. Rinsky, *Am. J. Ind. Med.* **27** (1995) no. 5, 759 – 770.
186. R. B. Hayes, Cancer Causes Control. 1997 May; 8(3): 371 – 85.
187. Trinkwasserverordnung of 5 Dec. 1990.
188. W. R. Bradley, W. G. Frederick, Ind. Med. 10, Ind. Hyg. Sect. **2** (1941) 15.
189. E. Browning: *Toxicity of Industrial Metals*,Butterworths, London 1961, p. 30.
190. Registry of Toxic Effects of Chemical Substances, NIOSH (ed.), Cincinnati, Ohio, 1976.
191. Registry of Toxic Effects of Chemical Substances, NIOSH (ed.), Cincinnati, Ohio, 1977.
192. N. Kanematsu, M. Hara, T. Kada, *Mutation Res.* **77** (1980) 109 – 116.
193. *Ann. Meet. Am. Ind. Hyg. Assoc.* **20** (1980) 1.

Antimycotics

Dermatotherapeutic Agents, Disinfectants, Fungicides, Agricultural are separate keywords

Axel Schmidt, Bayer AG, Wuppertal, Federal Republic of Germany

Frank-Ulrich Geschke, Bayer AG, Wuppertal, Federal Republic of Germany

1. Introduction 1
2. Azole Antimycotics 2
2.1. Azole Antimycotics for Systemic Treatment 3
2.1.1. Fluconazole 3
2.1.2. Itraconazole 4
2.1.3. Ketoconazole 5
2.1.4. Miconazole 7
2.2. Azole Antimycotics for Topical Treatment 9
2.2.1. Bifonazole 9
2.2.2. Butoconazole 10
2.2.3. Clotrimazole 11
2.2.4. Croconazole 12
2.2.5. Econazole 12
2.2.6. Fenticonazole 13
2.2.7. Isoconazole 14
2.2.8. Omoconazole 15
2.2.9. Oxiconazole 15
2.2.10. Sertaconazole 16
2.2.11. Sulconazole 17
2.2.12. Terconazole 17
2.2.13. Tioconazole 19
3. Polyene Antimycotics 19
3.1. Amphotericin B 20
3.2. Natamycin (Pimaricin) 22
3.3. Nystatin A1 23
4. Flucytosine 24
5. Griseofulvin 25
6. Ciclopirox 27
7. Thiocarbamates 28
7.1. Tolciclate 29
7.2. Tolnaftate 29
8. Allylamines 30
8.1. Naftifine 30
8.2. Terbinafine 31
9. Amorolfine 33
10. Unspecific Topical Antimycotics . 33
11. Recent Developments and Outlook 33
12. References 35

1. Introduction

The significant advances in antimycotic chemotherapy in the last decades are mainly due to the introduction of new, potent, and broadly active antimycotics to the market. The substances can be classified according to their mode of application into topical and systemic antimycotics. Topical antimycotics are applied to the skin, mucous membranes, and intravaginally in order to treat superficial mycoses locally. Peroral therapy and prophylaxis of intraluminal colonializations or infections of the gastrointestinal tract, e.g., with drops, syrups, or lozenges containing antimycotics that are not absorbed or are absorbed to an extremely small extent from the gastrointestinal tract (no or extremely low bioavailability) must also be regarded as topical therapy. Systemic antimycotics are applied parenterally and/or orally and show moderate to good bioavailability.

At present, numerous drugs for the topical treatment of superficial mycoses such as cutaneous and mucocutaneous infections are available, but there are only a limited number of drugs for the treatment of systemic mycoses. Additionally, resistance problems are arising, especially in the antimycotic chemotherapy of systemic mycoses, so that the medical need for new substances for this indication is great. Several "historical" compounds that show a relatively unselective antifungal activity, such as dyes, organic acids, mercury salts, and quaternary ammonium compounds, are sometimes still used in the treatment of superficial mycoses. Their therapeutic efficacy, tolerance, and patient acceptance are rated as low nowadays, and these substances are of decreasing relevance as much more selective antimycotics are available.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_077

Modern antimycotics are expected to be effective in vitro, in animal models, and in infected humans and animals against one or more of the following classes of pathogenic fungi: dermatophytes (e.g., *Trichophyton* and *Microsporum* species, *Epidermophyton floccosum*), yeasts (e.g., *Candida* and *Torulopsis* species, *Cryptococcus neoformans*), moulds (e.g., *Aspergillus fumigatus* and other *Aspergillus* species, *Fusarium* species, zygomycetes such as *Mucor, Rhizopas,* and *Absidia* species), and dimorphic (biphasic) fungi, which can express a yeast phase, which is mostly the pathogenic form in the infected host, and a mould phase, which is mostly the infective, saprophytic form in nature, e.g., in soils [1].

In addition to the very common superficial mycoses of the skin and mucous membranes, yeasts, moulds, and dimorphic fungi can also induce serious, life-threatening generalized opportunistic infections, especially in immunocompromised hosts. As the number of patients at high risk of systemic mycoses (e.g., patients after transplantations, under aggressive antineoplastic chemotherapy or in intensive care, AIDS patients) is increasing steadily, these infections are much more common nowadays [1]. Dermatophytes solely cause infections of the skin, hair, and nails, and are unable to induce systemic, generalized mycoses, even in immunocompromised hosts.

The in vitro antimycotic activity of the substances is assessed by the determination of MIC values (minimal inhibitory concentrations). This is the lowest concentration of a substance that inhibits the growth of a fungal isolate in vitro. The MICs of antimycotics can vary considerably – up to as much as several decades – depending on culture conditions such as medium composition, incubation temperature, inoculum size and preparation method. Therefore, a high degree of standardization must be maintained to obtain comparable results. Furthermore, there is often only a poor or no in vitro/in vivo correlation of MIC values with the clinical outcome.

Some antimycotics, including azole agents, also show a moderate to mostly low in vitro antibacterial activity which does not seem to be relevant for the clinical situation during therapy. Therefore, these minor antibacterial properties of the substances are not discussed in this chapter. Many formulations, especially of topically applicable antimycotics, contain other compounds such as urea, antibiotics, and steroids. These combinations are not extensively discussed here.

2. Azole Antimycotics

Azoles with antimycotic properties can be devided into imidazole and triazole derivatives. Both groups have the same mechanism of antifungal action. Imidazole derivatives with antimycotic activity were first synthesized at Bayer and Janssen and described in 1968/1969. Serval important commercial products have been developed for human and veterinary medicine and plant protection [2].

Low doses of azole antimycotics produce morphological changes in sensitive fungi. These are reflected in changes in the cell volume and plasma membrane [3] and are probably due to the interaction of the azoles with ergosterol biosynthesis [4–7]. Biochemically, azoles specifically inhibit fungal ergosterol biosynthesis by interference with fungal specific isoenzymes of the cytochrome P450-dependent 14α-demethylation of lanosterol or 24-methylenedihydrolanosterol, which is a key step in the biosynthesis of ergosterol. The accumulation of membrane-disturbing 14α-methylsterols seems to be the origin of the antifungal properties of azole compounds, as known from in vitro studies [8]. Ergosterol is an essential constituent of the fungal plasma membrane, helping to maintain the integrity of the cell membrane which has a barrier function. Ergosterol deficiency leads to loss of essential cytoplasmatic constituents through the plasma membrane, and results in cessation of growth and even death of the fungus. In all studies, an accumulation of sterols with a methyl group on C14, which are unsuitable as membrane components, was observed [9], [10]. At higher doses, azoles also interfere with synthesis of fatty acids and triglyceride [4]. The weak antibacterial activity of these substances seems to be based on another mechanism. However, MIC values of azole antimycotics strongly depend on the test method and are mostly of limited use in predicting clinical efficacy.

2.1. Azole Antimycotics for Systemic Treatment

2.1.1. Fluconazole

Fluconazole [*86386-73-4*], α-(2,4-difluorophenyl)-α-(1*H*-1,2,4-triazol-1-ylmethyl)-1*H*-1,2,4-triazole-1-ethanol, $C_{13}H_{12}F_2N_6O$, M_r 306.27, *mp* 138 – 140 °C, crystals from ethyl acetate/hexane.

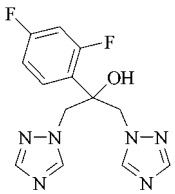

Synonym(s). UK 49858.

Solubility. Soluble in water.

Stability. The pure solid form is stable under normal storage conditions.

Description. The triazole antimycotic fluconazole was first synthesized at Pfizer Pharmaceuticals [11], [12].

Formulations. Capsules à 0.05, 0.1, 0.15, and 0.2 g. Suspensions (0.5 and 1 %) for oral application. Infusion bottles à 0.1, 0.2, and 0.4 g.

Trade Names. Biozolene (Bioindustria), Diflucan (Mason, Pfizer, Roerig), Dimycon (Alkaloid), Elazor (Sigma-Tau), Flavizol (Gador), Fungata (Mack), Mutum (Raffo), Triflucan (Pfizer).

Antimycotic Properties. Fluconazole is a broad spectrum triazole antimycotic with good in vitro and in vivo activity against pathogenic *Candida* species (apart from *Candida krusei*, which is a primary resistant *Candida* species), *Cryptococcus neoformans*, and dermatophytes [13]. *Aspergillus* species mostly show primary resistance towards fluconazole. Secondary resistance development can be observed after prolonged or repeated therapy of *Candida*-associated mycoses, especially in AIDS patients. In vitro susceptibilities of yeasts against fluconazole are reported in Table 1 [14].

Table 1. MIC values for fluoconazole

Fungus	No. isolates tested	MIC range, µg/mL
Candida albicans	443	0.125 – > 80
Candida kefyr	22	32 – 64
Candida lusitaniae	28	0.25 – 32
Candida parapsilosis	68	0.5 – 4
Candida tropicalis	88	0.5 – > 64
Trichosporon beigelii	15	1 – 64

Pharmacokinetics. The bioavailability after oral application ranges around 90 % [15]. The plasma concentrations and AUC (Area Under the Curve) are dose-proportional in the dose range of 50 – 400 mg/d [15]. After a single oral dose of 100 mg, a plasma concentration of 1.7 µg/L is achieved (6.7 µg/L after 400 mg) [16]. A steady state is mostly achieved after a therapy duration of 6 – 10 d. The serum protein binding of fluconazole is 12 % [17], and the plasma half-life 25 h. In *Cryptococcus neoformans* meningitis in AIDS patients, the liquor concentration was 80 % of the plasma concentration [18]. In vaginal secretions, the concentration of fluconazole was similar to the plasma concentration [19]; 60 – 75 % of the substance is eliminated by the kidneys in unmetabolized form. The elimination half-life is between 22 and 37 h [17].

Indications. Vaginal candidiasis [20] and systemic and mucocutaneous candidiasis [21–24]. In *Cryptococcus neoformans* meningitis the substance is mostly used in combination with amphotericin B and flucytosine [25]. Prophylaxis and preemptive therapy in AIDS patients, patients with leukemia, and patients with bone marrow transplants [26], [27].

Side Effects. Fluconazole is well tolerated.

CNS. Vertigo: 3.7 %, headache: 1.9 %.

GIT. Stomachache: 1.7 %, vomiting: 1.7 %, diarrhoea: 1.5 %.

Liver. Rise in transaminases (> 8 % of normal level): 1 %.

Skin. Exanthema: 1.8 %.

A Stevens – Johnson syndrome was observed after fluconazole therapy in an AIDS patient.

2.1.2. Itraconazole

Itraconazole [*84625-61-6*], 4-[4-[4-[4-[[2,4-Dichlorophenyl)-2-(1*H*-1,2,4-triazol-1-ylmethyl)-1,3-dioxolan-4-yl]-methoxy]phenyl]-1-piperazinyl]phenyl]-2,4-dihydro-2-(1-methylpropyl)-3*H*-1,2,4-triazol-3-one, $C_{35}H_{38}Cl_2N_8O_4$, M_r 705.64, *mp* 166.2 °C, crystals from toluene.

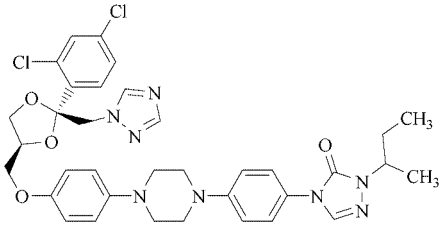

Synonym(s). R 51211.

Solubility. Itraconazole is a lipophilic substance which is practically insoluble in water and dilute acidic solutions. Partition coefficient in *n*-octanol/aqueous buffer of pH 8.1: 5.66.

Stability. The pure solid compund is stable under normal storage conditions.

Synthesis. Itraconazole is synthesized as follows [28], [29] (see top of next page):

Description. The triazole antimycotic itraconazole was first synthesized at Janssen Pharmaceuticals [30], [31].

Formulations. Capsules à 0.1 g.

Trade Names. Itrizole (Cilag), Sempera (Janssen), Siros (Cilag), Sporanox (Edward Keller, Janssen), Triasporin (Lifepharma).

Antimycotic Properties. Itraconazole is a broad-spectrum triazole antimycotic with good in vitro and in vivo activity against pathogenic dermatophytes, moulds (including *Aspergillus* species), *Candida* species, and dimorphic fungi such as *Blastomyces dermatitidis, Coccidioides immitis, Histoplasma capsulatum, Cryptococcus neoformans,* and *Sporothrix schenckii* [32–40]. MIC values of itraconazole are given in Table 2 [14].

Table 2. MIC values for itraconazole

Fungus	No. isolates tested	MIC range, µg/mL
Aspergillus fumigatus	96	0.001 – 10
Candida albicans	1605	0.063 – 128
Candida (Torulopsis) glabrata	4	2 – 128
Candida krusei	15	0.125 – 0.5
Candida lusitaniae	15	≤ 0.03 – 0.25
Candida parapsilosis	46	0.063 – > 128
Candida tropicalis	50	0.13 – > 128
Cryptococcus neoformans	118	0.001 – 0.5
Epidermophyton floccosum	5	0.063
Fusarium species	44	8 – > 32
Histoplasma capsulatum	10	0.063
Malassezia furfur	35	0.001 – 1
Microsporum species	10	0.063 – 0.25
Sporothrix schenckii	33	0.001 – 4
Trichophyton species	15	0.063 – 64
Trichosporon beigelii	15	≤ 0.03 – 0.25

Pharmacokinetics. The substance has a bioavailability after oral application of 40 % if taken on an empty stomach to 55 % if taken simultaneously with food [41]. The absorption of itraconazole after oral application is slow; peak plasma concentrations are reached 3 – 6 h after intake [41], [42]. Plasma concentrations are low, and a peak plasma concentration of 0.1 µg/L was observed after a single oral dose of 100 mg [42]. The serum protein binding of itraconazole is high; more than 99 % of the substance is bound to plasma proteins, mostly to albumin. The substance shows good tissue penetration, especially into the skin and nails [41]. Itraconazole is extensively metabolized in the liver, and more than 30 metabolites have been identified so far [43]. The elimination of the substance mostly occurs by the biliary-fecal route and renally in a metabolized form [44]. The elimination half-life ranges between 15 and 42 h, depending on the applied dose [41].

Indications. Therapy of dermatophytoses and onychomycoses [45–47] with a dosing regimen of 100 to 200 mg per day. Vaginal candidosis with 200 mg/d for 3 days [48]. Therapy of systemic infections such as candidosis and aspergillosis including disseminated infections, blastomycosis, coccidioidomycosis, histoplasmosis, and cryptococcosis with a dosing regimen of 100 to 400 (or up to 600) mg/d [49–51].

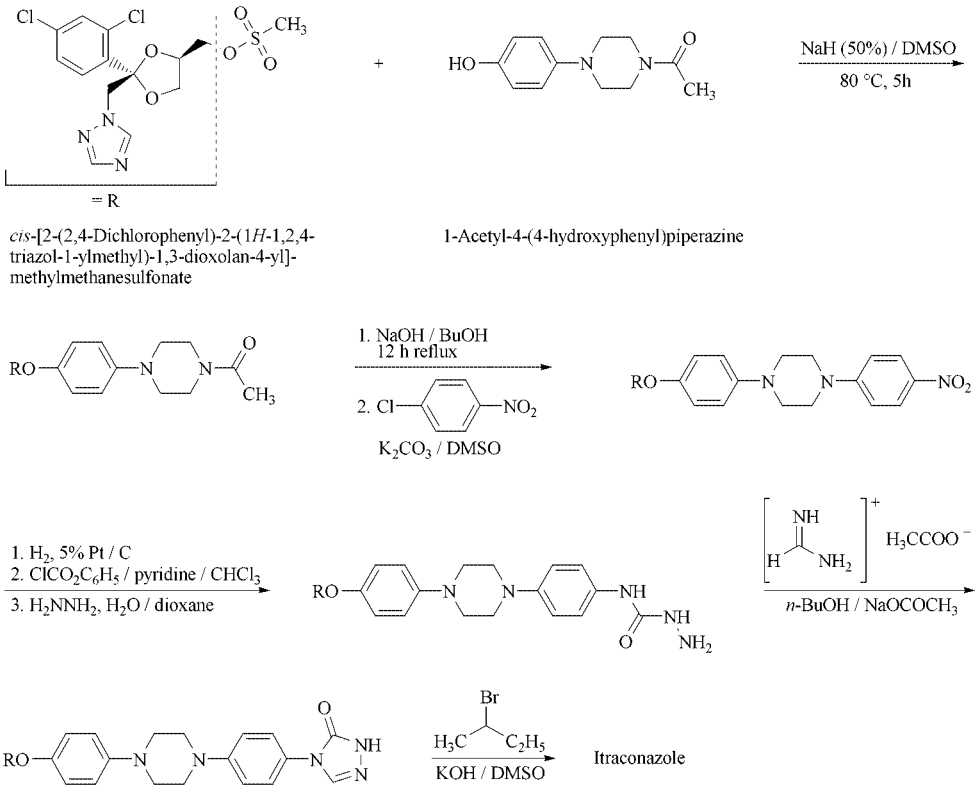

Especially in aspergillosis, an early onset of therapy is important for the clinical outcome.

Side Effects. Itraconazole is generally well tolerated.

CNS. Vertigo and headache: 10 %.

GIT. Stomachache, vomiting, and diarrhoea were observed.

Liver. Rise in transaminases and other reversible liver affections: 7 %.

Metabolism. Hypokaliemia and rise in serum triglycerides [52], [53].

2.1.3. Ketoconazole

Ketoconazole [*65277-42-1*], *cis*-1-acetyl-4-[4-[[2-(2,4-dichlorophenyl)-2-(1*H*-imidazol-1-ylmethyl)-1,3-dioxolan-4-yl]methoxy]phenyl]-piperazine, $C_{26}H_{28}Cl_2N_4O_4$, M_r 531.4, *mp* 146 °C, colorless crystals.

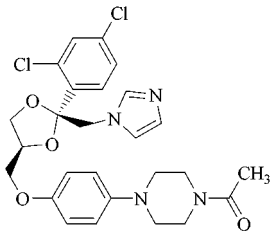

Synonym(s). R 41400.

Solubility. Ketoconazole is soluble in methanol, ethanol, and chloroform; slightly soluble in dilute acids; and barely soluble in water.

Stability. The pure solid compund is stable under normal storage conditions.

Synthesis. The synthesis of ketoconazole begins with the ketalization of 2,4-dichloroacetophenone (**1**) with glycerol. The intermediate (**2**) is not isolated but treated with

bromine to form the bromide (**3**). The benzoxylation of (**3**) gives (**4**) in the form of a *cis – trans* mixture. The *cis* form can be separated by crystallization. The reaction of (**4**) with imidazole in dimethylacetamide gives (**5**), which is then hydrolyzed to (**6**), which in turn is converted with methanesulfonyl chloride into the sulfonate (**7**). The reaction of (**7**) with the sodium salt of *N*-acetyl-*N*-*p*-hydroxyphenylpiperazine in DMSO gives ketoconazole.

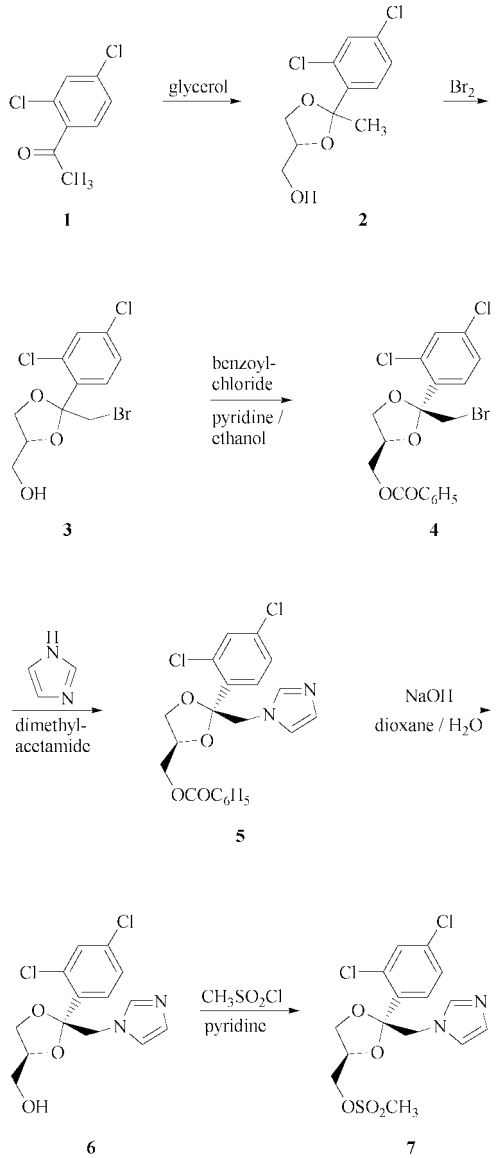

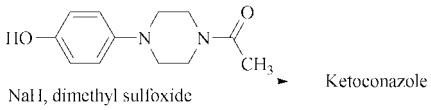

Description. The imidazole antimycotic ketoconazole was first synthesized at Janssen Pharmaceuticals [54–58].

Formulations. Tablets à 0.2 g; 2 % solutions, ointments, and cream preparations.

Trade Names. Candoral (Aché), Cetonax (Cilag), Fitonal (Disprovent), Fungarest (Janssen), Fungicil (Labinca), Fungo-Hubber (Hubber), Fungoral (Ilsan, Janssen, Leo), Ketazol (Exa), Ketoderm (Janssen), Ketoisdin (Isdin), Ketonan (IMA), Ketoral (Bilim), Micoral (Cassara), Micotek (Kressfor), Micoticum (Vita), Nizoral (Abic, Edward Keller, Janssen), Nizshampoo (Janssen), Oromycosal (Alkaloid), Oronazol (Krka), Panfungol (Esteve), Rofenid (Rhône-Poulenc), Terzolin (Janssen).

Antimycotic Properties. Ketoconazole shows a good efficacy against dermatophytes, dimorphic fungi (*Blastomyces dermatitidis, Coccidioides immitis, Histoplasma capsulatum*), and pathogenic *Candida* species [59], [60]. The activity against hyphomycetes and *Cryptococcus neoformans* is limited. MIC values of ketoconazole are summarized in Table 3. The in vitro activity of ketoconazole, like that of the other azoles, is strongly dependent on the test method used.

Table 3. MIC values for ketoconazole

Fungus	MIC range, µg/mL
Dermatophytes	0.04 – 10
Candida species	0.04 – 40
Cryptococcus neoformans	0.08 – 5
Aspergillus species	1.25 – 40
Dimorphic fungi	0.04 – 5
Chromomycetes	0.04 – 20

Pharmacokinetics. After oral administration, ketaconazole shows a variable degree of absorption of up to 70 %. The degree of absorption depends on the pH of the gastric fluid [61], [62]. Antacids and H_2-antagonists reduce

the absorption of the drug. Peak serum concentrations are reached 2 – 4 h after administration and are between 2.0 and 3,5 µg/mL for a dose of 200 mg (3 mg/kg), and between 3.5 and 7.0 µg/mL for a 400 mg dose. Serum concentrations of 10 – 20 µg/mL can be reached after administration of a single dose of 1200 mg [63]. The substance shows a binding potency to plasma proteins and erythrocytes of 99 % [61]. After vaginal application, 1 % of the substance is absorbed; after cutaneous application, no absorption could be measured [64]. Ketoconazole penetrates well into the stratum corneum of the skin and vaginal secretions [61]. The concentration of ketoconazole in urine is low, with less than 1 µg/mL after a daily dose of 400 mg. The drug diffuses only slowly into the cerebrospinal fluid, where it reaches about 5 % of the serum concentration. The elimination half-life ranges from 4 to 9 h. The substance is mainly eliminated by the biliary – fecal route in a metabolized, inactive form.

Indications. Systemic therapy of dermatomycoses, blastomycosis, coccidioidomycosis, and histoplasmosis. Because of the side effects of ketoconazole, the newer azoles such as fluconazole and itraconazole should be considered instead of ketoconazole for systemic treatment. Ketoconazole is especially indicated in topical therapy of dermatomycoses, pityriasis versicolor, seborrheoic dermatitis, and dandruff.

Side Effects. After oral administraition of ketaconazole, side effects are seen at up to 10 % (400 mg/d) to > 50 % (800 mg/d) [65].

CNS. Vertigo, dizziness, and headache.

GIT. Somachache, loss of appetite, vomiting, and diarrhoea.

Liver. Rise in transaminases and other liver affections which are mostly reversible, although fulminant hepatitis was observed in 1 in 10 000 cases.

Vessels. Angioedema. Further, skin reactions, insomnia, and lethargy can be observed. Gynaecomastia, oligospermia, loss of libido, and alopecia, which may occur, are due to an inhibition of testosterone and cortisol biosynthesis [66–69]. Daily doses of higher than 400 mg reduce the testosterone serum concentration.

2.1.4. Miconazole

Miconazole [*22916-47-8*], 1-[2-(2,4-dichlorophenyl)-2-[(2,4-dichlorophenyl)methoxy]ethyl]-1*H*-imidazole, $C_{18}H_{14}Cl_4N_2O$, M_r 415.92.

Miconazole nitrate, [*22832-87-7*], $C_{18}H_{15}Cl_4N_3O_4$, M_r 479.15, *mp* 184 – 185 °C, colorless crystals.

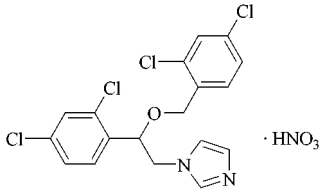

Synonym(s). R 18134; R 14889 (mononitrate).

Solubility. Miconazole is barely soluble in water (0.3 mg/mL) and slightly soluble to soluble in the many organic solvents and dilute inorganic acids.

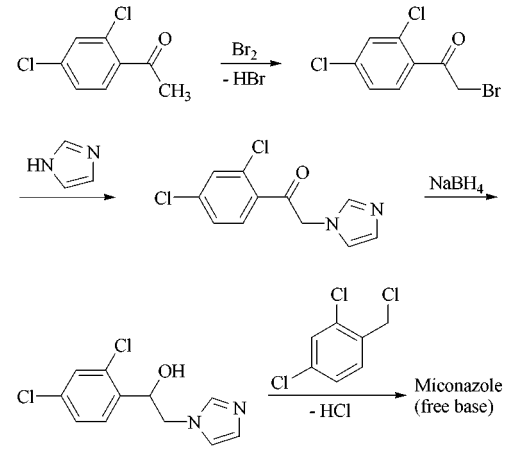

Stability. The pure solid compound is stable under normal storage conditions.

Synthesis. The synthesis [70], [71] begins with 2,4-dichloroacetophenone, which is first brominated to form the phenacyl bromide and then treated with imidazole to form the phenacylimidazole. Reduction with sodium borohydride and reaction of the resulting alcohol

with 2,4-dichlorobenzyl chloride yields miconazole.

Description. The imidazole antimycotic miconazole was first synthesized at Janssen Pharmaceuticals [72], [73].

Formulations. Ampulles à 0.2 g; lozenges à 0.25 g; 2 % solutions, lotions, powders, gels, ointments, and cream preparations; vaginal ovula.

Trade Names (Miconazole). Andergin (Pierell), Brentan (Janssen), Daktar (Janssen), Daktarin (Abic, Edward Keller, Janssen), Dumicoat (Dumex), Femeron (Janssen), Micofim (Elofar), Micotar (Dermapharm), Micotef (LPB), Monistat (Cilag, Ethnor, Janssen), Monistat I.V. (Janssen), Vodol (Searle).

Trade Names (Miconazole Nitrate). Aflorix (Gerardo Ramon), Albistat (Cilag), Aloid (Janssen), Andergin (Pierrel), Anotit (Janssen), Brenazol (DuraScan), Brentan (Janssen), Britane (Janssen), Canofite (Janssen), Conoderm (C-Vet), Conofite (Janssen, Pitman-Moore), Daktar (Janssen), Daktarin (Abic, Janssen), Decomyc (Merck), Deralbine (Andromaco), Dermacure (Janssen), Dermonistat (Ortho), Epi-Monistat (Cilag, Janssen), Florid (Janssen, Mochida), Fungiderm (Janssen), Funginazol (Morgens), Fungisdin (Esteve, Janssen), Fungucit (Iltas), Fungur (Salutas), Gyno-Daktar (Janssen), Gyno-Daktarin (Edward Keller, Janssen, Johnson & Johnson), Gyno-Monistat (Cilag, Janssen), Ipec (Byk-Gulden), Micatin (McNeil, Ortho), Micoderm (Italsuisse), Micogel (Cipla), Micogyn (Elofar), Miconal (Ecobi), Micotef (LPB), Monistat (Cilag, Janssen, Ortho), Monostat 7 (Ortho), Mykotral (Chephasaar), Prilagin (Gambar), Surolan (Janssen), Vodol (Searle).

Antimycotic Properties. Miconazole is a broad-spectrum imidazole antimycotic that generally shows activity against pathogenic dermatophytes (*Epidermophyton floccosum; Trichophyton* and *Microsporum* species), *Candida* species, moulds including *Aspergillus* species, and pathogenic dimorphic fungi [74–78]. MIC values are summarized in Table 4.

Table 4. MIC values for miconazole

Fungus	No. isolates tested	MIC range µg/mL
Aspergillus fumigatus	74	0.5 – 64
Blastomyces dermatitidis	57	0.001 – 2
Candida albicans	1815	0.016 – 100
Candida (Torulopsis) glabrata	224	0.016 – 64
Candida kefyr	33	0.016 – 4
Candida krusei	40	< 0.063 – 6.25
Candida lusitaniae	7	0.18 – 6.3
Candida parapsilosis	54	0.016 – 32
Candida tropicalis	172	0.016 – 33
Coccidioides immitis	99	0.063 – 6
Cryptococcus neoformans	69	0.063 – 25
Epidermophyton floccosum	19	0.06
Histoplasma capsulatum	48	0.05 – 2
Malassezia furfur	44	0.2 – 50
Microsporum canis	18	0.1 – 4
Pseudoallescheria boydii	42	< 0.016 – 4
Sporothrix schenckii	65	0.5 – 16
Trichophyton mentagrophytes	28	0.1 – 4
Trichophyton rubrum	63	0.001 – 32

Phamacokinetics. Miconazole is incompletely absorbed after oral application with an average bioavailability of 27 % [79]. The active drug is dissolved in Cremophor EL for intravenous administration. After intravenous administration of a single dose of 200 mg, a peak plasma concentration of 1.6 mg/L was observed. Infusion of 500 mg of miconazole briefly produces serum concentrations of 2 – 9 µg/mL, which, however, rapidly fall to under 0.2 µg/mL. Miconazole is barely absorbed after cutaneous and/or mucocutaneous application. The substance shows a plasma protein binding of 90 % and poor penetration into the liquor cerebrospinalis and other body fluids. Miconazole is extensively metabolized in the liver. The substance is eliminated predominantly by the biliary – fecal route as inactive metabolites. Elimination occurs in phases with a terminal elimination half-life of 24 h [79].

Indications. Especially topical therapy of skin infections caused by dermatophytes and yeasts, and mucocutaneous infections including *Candida* vulvovaginitis. Systemic, intravenous therapy especially of infections caused by *Pseudoallescheria boydii* (dosage up to 1.2 g 3 × per day) [80]. For treatment of other generalized my-

coses, mostly the newer, more effective, and less toxic azole antimycotics for systemic use, such as fluconazole and itraconazole, are indicated.

Side Effects. Multiple side effects have been observed after intravenous infusion of miconazole, which seem to be partly due to the formulation of the solvent, which contains the detergent Cremophor EL; mostly anaphylactic reactions, pruritus, dizziness, tachyarrhythmia, hyperaggregation of erythrocytes, anemia, thromocytosis, phlebitis, nausea, vomiting, hyperlipidemia, hyponatriemia, arthralgia, and flush [81].

2.2. Azole Antimycotics for Topical Treatment

2.2.1. Bifonazole

Bifonazole [*60628-96-8*], 1-([1,1'-biphenyl]-4-ylphenylmethyl)-1*H*-imidazole, $C_{22}H_{18}N_2$, M_r 310.4, *mp* 147 – 148 °C, colorless, odorless, tasteless crystals.

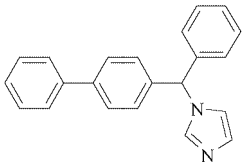

Synonym(s). Trifonazole, Bay h 4502.

Solubility. Bifonazole is soluble in glacial acetic acid (833 mg/mL), chloroform (344 mg/mL), benzene (38.2 mg/mL), methanol (36.5 mg/mL), ethanol (24.6 mg/mL), ethyl acetate (17.1 mg/mL), and acetonitrile (10.4 mg/mL). It is practically insoluble in water, but in Walpole's buffer, solubility is 3.75 mg/mL at pH 2.

Stability. Pure bifonazole is stable under normal storage conditions. In solution, the substance is stable under neutral or slightly alkaline conditions. It slowly decomposes in strongly alkaline or strongly acidic media. Bifonazole solutions are sensitive to light.

Synthesis. Bifonazole is synthesized as follows:

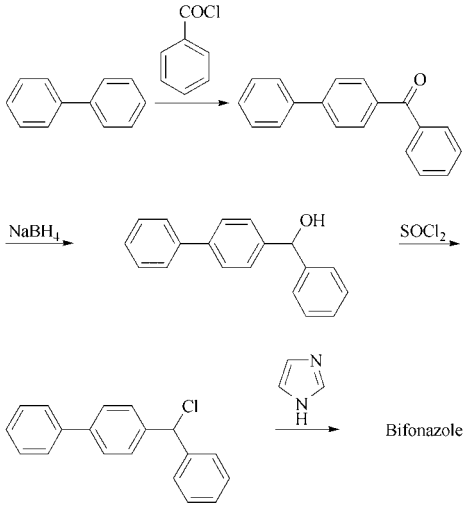

Description. The imidazole antimycotic bifonazole was first synthesized at Bayer Pharmaceuticals in 1974 [82–86].

Formulations. 1 % solutions, powders, gels, ointments, and cream preparations.

Trade Names. Amycor (Lipha), Azolmen (Menarini), Bifonazol (Bayropharm), Micofun (Incobra), Mycospor (Bayer, Kai Cheong), Mycosporin (Bayer).

Antimycotic Properties. Bifonazole is a broad-spectrum imidazole antimycotic with excellent in vitro activity against pathogenic dermatophytes, yeasts, moulds, and dimorphic fungi. MIC values of bifonazole are summarized in Table 5.

Table 5. MIC values for clotrimazole and bifonazole

Fungus	MIC range, µg/mL	
	clotrimazole	bifonazole
Dermatophytes	<0.01 – 2	0.01 – 1
Candida species	0.04 – 10	0.08 – 10
Cryptococcus neoformans	0.04 – 10	0.04 – 2.5
Pityrosporum species	2 – 4	0.5 – 1
Aspergillus species	0.1 – 1.25	0.5 – 1
Chromomycetes	0.04 – 2.5	<0.04 – 0.31
Dimorphic fungi	0.04 – 0.31	<0.04 – 0.08

Pharmacokinetics. Bifonazole is absorbed adequately after oral administration to humans and animals. The substance quickly induces ribosomal liver enzymes and is primarily metabolized to metabolites without antimycotic activity. Therefore, bifonazole can only be used as a topical antimycotic [87], [88]. The substance is minimally absorbed after topical application of 1 % formulations. Concentrations in blood serum remain below 5 ng/mL, even after long term treatment of a 200 cm^2 skin area. Concentrations in the stratum papillare of the skin reach 2 – 3 µg/g [88].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes, yeasts, and also chromomycetes.

Side Effects. Topical applications, including intravaginal application, are well tolerated. Skin irritations and allergic reactions are rarely observed and seem to be mainly due to the galenic formulation (e.g., fatty preparations on acute, inflammative lesions). Allergic reactions are rarely observed [89].

2.2.2. Butoconazole

Butoconazole [*64872-76-0*], (±)-1-[4-(4-chlorophenyl)-2-[(2,6-dichlorphenyl)thio]butyl]-1*H*-imidazole, $C_{19}H_{17}Cl_3N_2S$, M_r 411.78, *mp* 68 – 70.5 °C, crystals from cyclohexane.

Butoconazole nitrate [*64872-77-1*], $C_{19}H_{18}Cl_3N_2O_3S$, M_r 475.01, *mp* 162 – 163 °C, colorless blades from acetone/ethyl acetate.

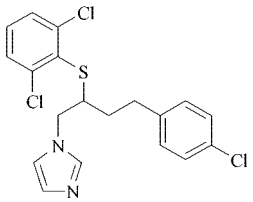

Solubility. For further information, see [90].

Stability. The pure solid compound is stable under normal storage conditions.

Synthesis [91]:

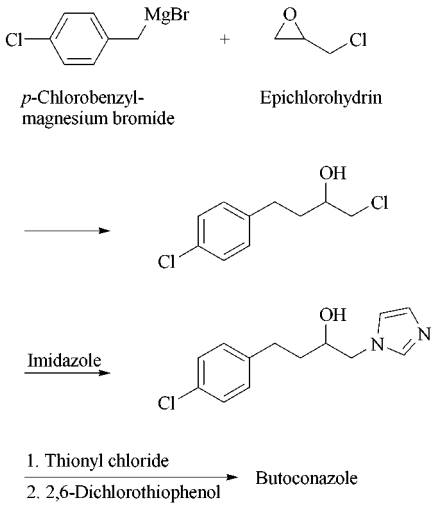

Description. The imidazole antimycotic butoconazole was first synthesized at Syntex Pharmaceuticals [92].

Formulations. 2 % cream preparations.

Trade Names (Butoconazole Nitrate): Femcosyn (Protochemie), Femstat (Syntex), Gynomyk (Cassenne).

Antimycotic Properties. Broad-spectrum imidazole antimycotic with in vitro and in vivo activitiy against pathogenic yeasts, dermatophytes, moulds, and dimorphic fungi [93], [94].

Pharmacokinetics. After intravaginal application of 100 mg of butoconazole nitrate, 5 % of the dose was found in the urine and feces [95].

Indications. Butoconazole is generally indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata* [96], [97].

Side Effects. Local irritations such as itching and burning sensations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation [98].

2.2.3. Clotrimazole

Clotrimazole [*23593-75-1*], 1-[(2-chlorophenyl)diphenylmethyl]-1*H*-imidazole, $C_{22}H_{17}ClN_2$, M_r 344.8, *mp* 143 – 144 °C, colorless, tasteless crystals.

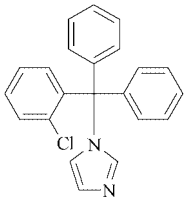

Synonym(s). Chlortritylimidazole, Bay 5097, FB b 5097, PCPIM.

Solubility. Clotrimazole is readily soluble in methanol, chloroform, and DMF (all > 100 mg/mL); slightly soluble in acetone (50 mg/mL), ethyl acetate (45 mg/mL), ethanol (95 mg/mL), and PEG 400 (60 mg/mL); and practically insoluble in water (< 0.01 mg/mL).

Stability. Pure clotrimazole is stable under normal storage conditions. The stability of the solution is pH dependent. Above pH 7 the solution is stable. In acid solution, clotrimazole undergoes gradual hydrolysis to form 2-chlorophenylbisphenylmethanol and imidazole. The hydrolysis half-life at pH 1 is about 170 h in ethanol/water (1/1) at room temperature.

Synthesis. The synthesis of clotrimazole begins with 2-chlorobenzyl chloride, which is converted to benzotrichloride. Friedel – Crafts reaction with benzene followed by reaction with imidazole forms clotrimazole [99].

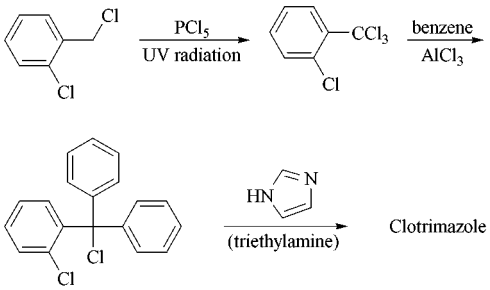

Description. The imidazole antimycotic clotrimazole was first synthesized at Bayer Pharmaceuticals in 1967 [100]. Ist antimycotic activity was first reported in 1969 [101], [102–104].

Formulations. Solutions, powders, gels, ointments, and cream preparations; lozenges and vaginal tablets.

Trade Names. Agisten (Agis), Akneclor (Spirig), Antifungol (Hexal), Antimicotico (Savoma), Antimyk Neu (Pfleger), Apocanda (Apogepha), Azutrimazol (Azuchemie), Canesten (Bayer, Bayropharm, Kai Cheong), Canifug (Wolf), Clazol (Ilsan), Clocim (Cimex), Clot-basan (Sagitta), Clotrifug (Wolff), Clotrimaderm (Taro), Clotrimix (Hosbon), clotri OPT (Optimed), Clotrizol (Jossa), Clozole (Jean-Marie), Contrafungin (PharmaGalen), Cutistad (Stada), Dignotrimazol (Dignos), Dolexalan (Wölfer), Durafungol (Durachemie), Empecid (Bayer, Yoshitomi), Eurosan (Mepha), Femcare (Schering-Plough), Fungiframan (Oftalmiso), Fungi-med (Permamed), Fungizid-ratiopharm (ratiopharm), Fungosten (Mulda), Fungotox (Mepha), Gilt (Lyssia), Gino-Canesten (Bayer), Gino-Clotrimix (Hosbon), Gino-Lotramina (Schering Corp.), Gromazol (Grossman), Gyne-Lotremin (Mason, Schering Corp., Schering-Plough), Jenamazol (Jenapharm), KadeFungin (Kade), Localicid (Dermapharm), Lotramina (Schering Corp.), Lotremin (Mason, Schering Corp.), Lotrimin (Schering Corp., Schering-Plough), Micomisan (Hosbon), Micoter (Cusi), Mono-Baycuten (Bayropharm), Mycelex (Bayer, Dome), Mycelex-7 (Bayer), Mycelex-G (Bayer, Dome), Myclo (Boehringer Ingelheim), Mycofug (Hermal), Myco-Hermal (Hermal, Jebsen), Mycoril (Remedica), Mycosporin (Bayer), Mycotrim (Lagap), Myko Crdes (Ichthyol), Mykofungin (Wyeth), Mykohaug C (Salutas), Ovis Neu (Aldenylchemie), Peckle (Hisamitsu), Pedisafe (Sagitta), Plimycol (Pliva), SD-Hermal (Hermal), Stiemazol (Stiefel), Tibatin (DAK), Tricosten (Farmion), Trimysten (Bellon), Uromykol (Hoyer).

Antimycotic Properties. Broad-spectrum imidazole antimycotic with good activity against yeasts, dermatophytes, moulds, and dimorphic fungi. MIC values of clotrimazole are summarized in Table 5.

Pharmacokinetics. Clotrimazole is absorbed adequately after oral administration to humans and animals. The substance quickly induces ribosomal liver enzymes and is primarily metabolized to metabolites without antimycotic activity. Therefore, clotrimazole can only be used as a topical antimycotic [105], [106]. The substance is minimally absorbed after the topical application of 1 % formulations. Concentrations in blood serum remain below 5 ng/mL, even after long-term treatment of a 200 cm^2 skin area. Concentrations in the stratum papillare of the skin reach 2 – 3 µg/g [106].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes, yeasts, and chromomycetes. In addition, clotrimazole is indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata.*

Side Effects. Topical applications, including intravaginal application, are well tolerated. Skin irritations and allergic reactions are rarely observed and seem to be mainly due to the galenic formulation (e.g. fatty preparations on acute, inflammative lesions). Allergic reactions are rarely observed [107].

2.2.4. Croconazole

Croconazole [*77175-51-0*], 1-[1-[2-(3-chlorobenzyloxy)phenyl]vinyl]imidazole, $C_{18}H_{15}ClN_2O$, M_r 310.78, *mp* 72 – 73 °C, white crystals.

Croconazole hydrochloride [*77174-66-4*], $C_{18}H_{16}Cl_2N_2O$, M_r 347.24, *mp* 148.5 – 150 °C, white crystals.

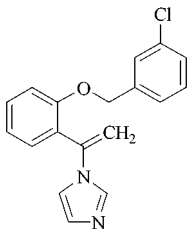

Synonym(s). Croconazole (free base); S 710674 (hydrochloride).

Solubility. Soluble in ethyl acetate.

Stability. The pure solid compound is stable under normal storage conditions.

Description. The imidazole antimycotic croconazole was first synthesized at Shionogi Pharmaceuticals [108], [109].

Formulations. 1 % creams and gel preparations.

Trade Names (Croconazole Hydrochloride). Pilzcin (Merz, Merz + Schoeller, Shionogi).

Antimycotic Properties. Broad-spectrum imidazole antimycotic with activity against almost all pathogenic fungi.

Pharmacokinetics. Croconazole is not absorbed after cutaneous application.

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts.

Side Effects. Local irritations and allergic reactions after topical treatment may occur in rare cases and are mainly due to the galenic formulation.

2.2.5. Econazole

Econazole [*27220-47-9*], 1-[2-[(4-chlorphenyl)-methoxy]-2-(2,4-dichlorophenyl)ethyl]-1*H*-imidazole, $C_{18}H_{15}Cl_3N_2O$, M_r 381.69, *mp* 86.8 °C, colorless crystals.

Econazole nitrate [*68797-31-9*], $C_{18}H_{16}Cl_3N_3O_4$, M_r 444.7, *mp* 164 – 165 °C, colorless crystals.

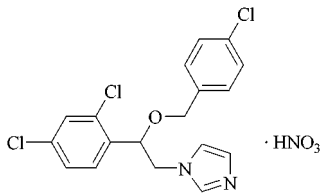

Synonym(s). R 14827 (mononitrate).

Solubility. Econazole is soluble in methanol, acetic acid, and PEG 400; less soluble in ethanol, acetone, chloroform, and butanol; and barely soluble in water, ether, cyclohexane, and hexane [110].

Stability. The solid substance is stable under usual storage conditions but should be stored in the dark.

Synthesis. The synthesis is almost the same as that of miconazole (see Section 2.1.4) [70], [71], [111], differing only in the last stage, in which the 1-[2,4-dichloro[β-hydroxy)phenethyl]imidazole intermediate is treated with 4-chlorobenzyl chloride.

Description. The imidazole antimycotic econazole was first synthesized at Janssen Pharmaceuticals [111] in 1969 [112], [113].

Formulations. 1 % solutions, powders, ointments, and cream preparations; vaginal ovula (150 mg).

Trade Names (Econazole). Pevalip (Janssen), Pevaryl (Cilag), Pevaryl Lipogel (Cilag, Janssen), Pevaryl lotion (Cilag), Pevaryl P.V. (Cilag).

Trade Names (Econazole Nitrate). Amicel (Salus), Chemionazolo (Brocchieri), Dermazol (CT), Dermazole (Lision Hong), Ecalin (JAKA-80), Ecodergin (Leben's), Eco Mi (Geymonat), Ecorex (Tosi), Ecostatin (Squibb, Westwood-Squibb), Ecotam (Alacan), Epi-Pevaryl (Cilag), Etramon (Johnson & Johnson), Fitonax (Cilag), Gyno-Pevaryl (Cilag, Edward Keller), Ifenec Zilliken), Micocert (Dexter), Micoespec (Centrum), Micofungal (Labopharma), Micogin (Crosara), Micogyn (Crosara), Micos (AGIPS), Micosten (Bergamon), Micostyl (Stiefel), Mitekol (Lek), Mykopevaryl (Cilag), Pargin (Gibipharma), Pevaryl (Cilag, Edward Keller, Smith Kline Beecham), Pevaryl P.V. (Edward Keller), Polycain (Taiho), Skilar (Bonomelli), Spectazole (Ortho).

Antimycotic Properties. Broad-spectrum imidazole antimycotic with activity against almost all species of pathogenic fungi [114], [115]. MIC values of econazole are summarized in Table 6 [14].

Table 6. MIC values for econazole

Fungus	No. isolates tested	MIC range µg/mL
Aspergillus fumigatus	24	0.15 – 25
Candida albicans	283	0.016 – 25
Candida (Torulopsis) glabrata	38	0.016 – 25
Candida kefyr	33	0.016 – 6.25
Candida krusei	34	0.125 – 12.5
Candida parapsilosis	32	0.016 – 25
Candida tropicalis	52	0.016 – 25
Malassezia furfur	19	0.0125 – 25

Pharmacokinetics. After vaginal application, 1 – 27 % of the applied dose was found in the urine [116], [117].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts [118]. In addition, econazole is indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata* [119], [120].

Side Effects. Local irritations such as itching and burning sensations and allergic reactions may occur in rare cases after topical therapy and are mainly due to the galenic formulation [121].

2.2.6. Fenticonazole

Fenticonazole [*72479-26-6*], 1-[2-(2,4-dichlorophenyl)-2-[[4-(phenylthio)phenyl]-methoxy]ethyl]-1*H*-imidazole, $C_{24}H_{20}Cl_2N_2OS$, M_r 455.4.

Fenticonazole nitrate [*73151-29-8*], $C_{24}H_{21}Cl_2N_3O_4S$, M_r 518.63, *mp* 136 °C, odorless white crystalline powder.

Synonym(s). Rec 15/1476 (mononitrate).

Solubility. Solubility at 20 °C: water < 0.1 mg/mL, diethyl ether < 0.1 mg/mL, ethanol 30 mg/mL, methanol 100 mg/mL, chloroform 300 mg/mL, DMF 600 mg/mL [122].

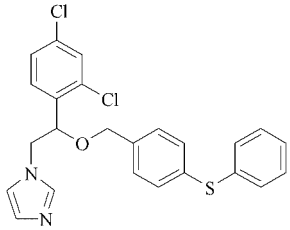

Stability. The pure solid compund is stable under normal storage conditions.

Description. The imidazole antimycotic fenticonazole was first synthesized at Recordati Pharmaceuticals [123], [124].

Formulations. 2 % solutions, ointments, and cream formulations; vaginal ovula (600 mg).

Trade Names (Fenticonazole Nitrate). Falvin (Farmades), Fenizolan (Organon), Fentiderm (Zyma), Fentigyn (Novartis), Lomexin (Grünenthal, Recordati), Mycodermil (Vifor Fribourg).

Antimycotic Properties. Broad-spectrum imidazole antimycotic with activity against almost all species of pathogenic fungi.

Pharmacokinetics. After topical application, the substance is not absorbed.

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts [125], [126]. In addition, fenticonazole is indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata* [127].

Side Effects. Local irritations such as itching and burning sensations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.7. Isoconazole

Isoconazole [*27523-40-6*], 1-[2-(2,4-dichlorophenyl)-2-[(2,6-dichlorophenyl)methoxy]ethyl]-1*H*-imidazole, $C_{18}H_{14}Cl_4N_2O$, M_r 416.1, white crystals.

Isoconazole nitrate [*24168-96-5*], $C_{18}H_{15}Cl_4N_3O_4$, M_r 479.1, *mp* 182 – 183 °C, colorless crystals.

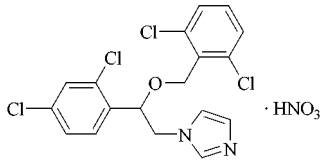

Synonym(s). R 15454 (mononitrate).

Solubility. Isoconazole is soluble in methanol, acetic acid, and PEG 400; less soluble in ethanol, acetone, chloroform, and butanol; and barely soluble in water, diethyl ether, cyclohexane, and hexane.

Stability. The solid substance is stable under normal storage conditions.

Synthesis. The synthesis of isoconazole [128] differs from that of miconazole (see Section 2.1.4) only in the last stage, in which 1-[2,4-dichloro(β-hydroxy)phenethyl]imidazole is treated with 2,6-dichlorobenzyl chloride in hot benzene – DMF in the presence of sodium hydride.

Description. The imidazole antimycotic isoconazole was first synthesized at Janssen Pharmaceuticals [128–131].

Formulations. 1 % solutions, ointments, and cream formulations.

Trade Names (Isoconazole Nitrate). Fazol (Bellon), Gyno-Travogen (Jebsen, Schering), Icaden (Schering), Isogyn Ginecologico (Crosara), Mupaten (Schering-Plough), Travogen (Jebsen, Schering), Travogyn (Schering).

Antimicrobial Properties. Broad-spectrum azole antimycotic with activity against almost all species of pathogenic fungi [132], [133].

Pharmacokinetics. After topical application, isoconazole remains practically unabsorbed. Application in an ethanol/propyleneglycol vehicle increases penetration into the stratum corneum of the skin significantly [134].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts. In addition, isoconazole is indicated for therapy of vulvovaginal mycoses causes by *Candida* species and *Candida (Torulopsis) glabrata* [135].

Side Effects. Local irritations such as itching and burning sensations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.8. Omoconazole

Omoconazole [*74512-12-2*], (*Z*)-1-[2-[2-(4-chlorophenoxy)ethoxy]-2-(2,4-dichlorophenyl)-1-methylethenyl]-1*H*-imidazole, $C_{20}H_{17}C_{13}N_2O_2$, M_r 423.73, *mp* 89 – 90 °C, crystals from ethyl acetate/hexane (1/4).

Omoconazole nitrate [*83621-06-1*], $C_{20}H_{18}Cl_3N_3O_5$, M_r 485.96, *mp* 118 – 120 °C (Büchi) / 122.5 (Mettler), crystals from ethyl acetate/ethanol.

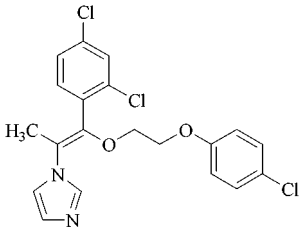

Synonym(s). CM 8282 (mononitrate).

Stability. The pure solid substance is stable under normal storage conditions.

Description. The imidazole antimycotic omoconazole was first synthesized at Siegfried Pharmaceuticals [136], [137]. Stereospecific synthesis [138].

Formulations. 1 % cream formulations.

Trade Names (Omoconazole Nitrate). Azameno (Kwizda, Wyeth), Fongamil (Biorga), Fongarex (Sanofi), Fongorex (Siegfried), Melur (Siegfried).

Antimycotic properties. Broad-spectrum imidazole antimycotic with activity against almost all pathogenic fungi [139].

Pharmacokinetics. After topical application, omoconazole is not absorbed.

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts.

Side Effects. Local irritations such as itching and burning sensations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.9. Oxiconazole

Oxiconazole [*64211-45-6*], (*Z*)-1-(2,4-dichlorophenyl)-2-(1*H*-imidazol-1-yl)ethanone *O*-[(2,4-dichlorophenyl)methyl]oxime, $C_{18}H_{13}Cl_4N_3O$, M_r 429.92.

Oxiconazole nitrate [*64211-46-7*], $C_{18}H_{14}Cl_4N_4O_4$, M_r 492.15, *mp* 137 – 138 °C, crystals from ethanol.

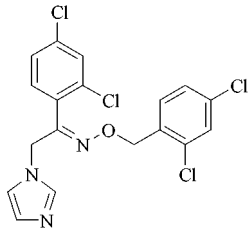

Synonym(s). Ro 13-8996/000 (free base); Ro 13-8996/001, Sgd 301-76 (mononitrate).

Stability. The pure solid substance is stable under normal storage conditions.

Synthesis. Oxiconazole is synthesized as follows [140], [141]:

Description. The imidazole antimycotic oxiconazole was synthesized at Roche and Siegfried Pharmaceuticals [142], [143].

Formulations. 1 % solutions, powders, ointments, and cream formulations. Vaginal tablets (600 mg).

Trade Names (Oxiconazole Nitrate). Gyno-Myfungar (Klinge), Myfungar (Klinge, Siegfried, Wyeth), Oceral (Roche), Oxistat (Glaxo).

Antimycotic Properties. Broad-specktrum imidazole antimycotic with activity against almost all species of pathogenic fungi [144].

Pharmacokinetics. After local application, oxiconazole penetrates well into the stratum corneum of the skin and the hair follicle [145].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts. In addition, oxiconazole is indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata* [146].

Side Effects. Local irritations such as itching and burning sensations, and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

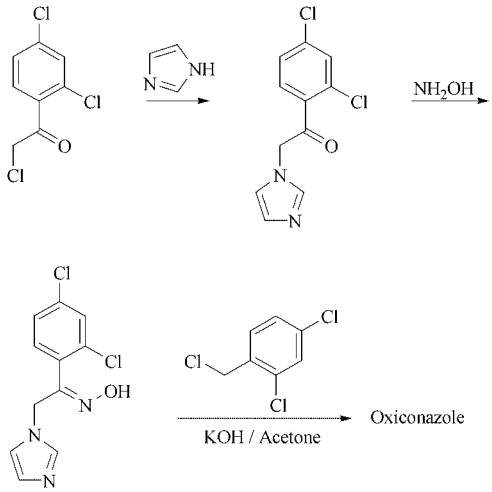

2.2.10. Sertaconazole

Sertaconazole [*99592-32-2*], (±)-1-[2,4-dichloro-β-[(7-chlorobenzo[*b*]thien-3-yl)methoxy]phenethyl]imidazole, $C_{20}H_{15}Cl_3N_2OS$, M_r 437.78.

Sertaconazole nitrate [*99592-39-9*], $C_{20}H_{15}Cl_3N_3O_4S$, M_r 501.01, *mp* 158 – 160 °C, odorless, white crystalline powder.

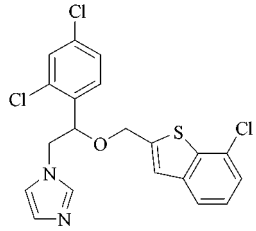

Synonym(s). FI-7045 (free base); FI-7056 (mononitrate).

Solubility. Fairly soluble in ethanol (1.7 %), chloroform (1.5 %); slightly soluble in acetone (0.95 %); very slightly soluble in *n*-octanol (0.069 %). Practically insoluble in water (< 0.01 %).

Stability. The pure solid substance is stable under normal storage conditions.

Synthesis. The Synthesis is described in [147].

Description. See [148], [149].

Formulations. 1 % cream formulations, gels, and ointments. New formulations for the treatment of vaginal candidiasis and oral formulations for the prophylaxis and treatment of mucocutaneous buccopharyngeal candidiasis are being developed.

Trade Names (Sertaconazole Nitrate). Dermofix (Ferrer), Dermoseptic (Smith Kline Beecham), Zalain (Robert).

Antimycotic Properties. Sertaconazole is a rather new broad-spectrum imidazole antimycotic with activity against almost all species of pathogenic fungi. It also has excellent activity against pathogenic yeasts [150–153].

Parmacokinetics. The substance is not absorbed after topical administration.

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as yeasts and dermatophytes [154]. New formulations for the treatment of vaginal mycoses are in development.

Side Effects. Local irritations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.11. Sulconazole

Sulconazole [*61318-90-9*], 1-[2-[[(4-chlorophenyl)methyl]thio]-2-(2,4-dichlorophenyl)ethyl]-1*H*-imidazole $C_{18}H_{15}Cl_3N_2S$, M_r 460.8.

Sulconazole nitrate [*61318-91-0*], $C_{18}H_{15}Cl_3N_3O_3S$, M_r 524.03, *mp* 130.5–132.0 °C, white or off-white crystalline powder.

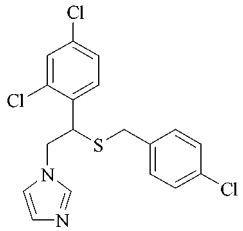

Synonym(s). RS-44872, RS-44872-00-10-3 (both mononitrate).

Solubility. 1 part in 3,333 (water), 1 part in 100 (ethanol), 1 part in 130 (acetone), 1 part in 333 (chloroform), 1 part in 2000 (dioxan), 1 part in 71 (methanol), 1 part in 286 (chloromethane), 1 part in 10 (pyridine), 1 part in 2000 (toluene).

Stability. The pure solid substance is stable under normal storage conditions, but shoud be protected from light.

Synthesis. [155]:

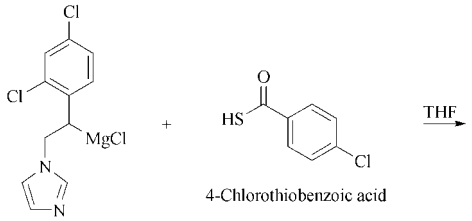

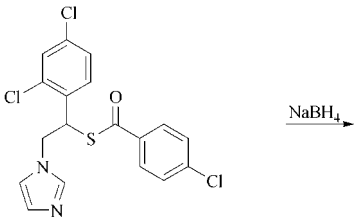

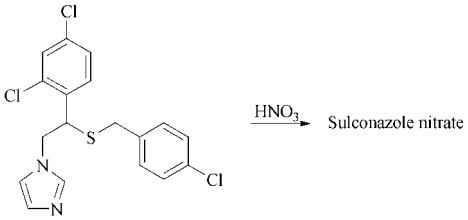

Description. See [156–158].

Formulations. 1 % solutions, creams, and ointments.

Trade Names (Sulconazole Mononitrate). Exelderm (Schwarz), MYK (Cassenne), MYK-1 (Will), Sulcosyn (Syntex), Suldicyn (Syntex).

Antimycotic Properties. Sulconazole is a broad-spectrum imidazole antimycotic with activity against almost all pathogenic fungi [159–161].

Pharmacokinetics. The substance is not absorbed after topical administration.

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts.

Side Effects. Local irritations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.12. Terconazole

Terconazole [*67915-31-5*], *cis*-1-[4-[[2-(2,4-dichlorophenyl)-2-(1*H*-1,2,4-triazol-1-ylmethyl)-1,3-dioxolan-4-yl]methoxy]phenyl]-4-(1-methylethyl)piperazine, $C_{26}H_{31}Cl_2N_5O_3$,

M_r 532.48, *mp* 126.3 °C, crystals from isopropyl ether.

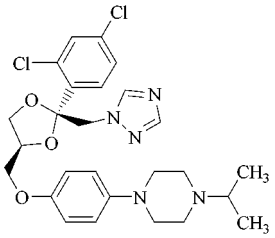

Synonym(s). Triaconazole, R 42470.

Stability. The pure solid substance is stable under normal storage conditions.

Synthesis. [162], [163]:

2,4-Dichloroacetophenone Glycerol

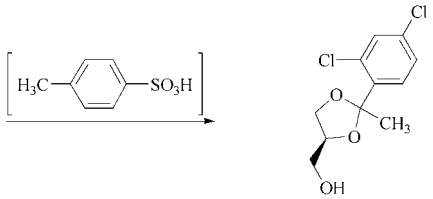

Dioxolane

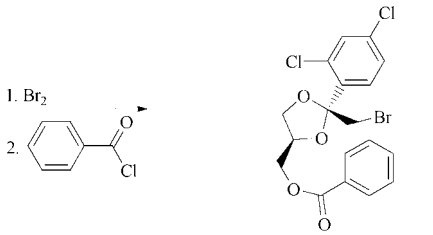

2-(Bromomethyl)-2-(2,4-dichlorophenyl)-4-(benzoyloxymethyl)-1,3-dioxolane

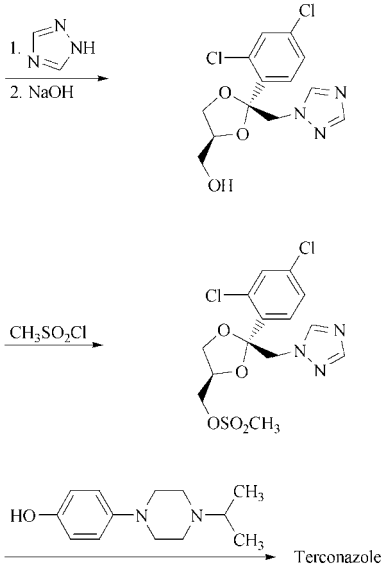

Description. The triazole antimycotic terconazole was first synthesized at Janssen Pharmaceuticals [164–166].

Formulations. 1 % solutions, powders, ointments, and cream formulations; vaginal pessar (80 mg).

Trade Names. Fungistat (Janssen), Gyno-Terazol (Cilag), Panlomyc (Janssen), Terazol (Cilag, Organon, Ortho), Terconal (Fisons), Tercospor (Cilag).

Antimycotic Properties. Terconazole is a broad-spectrum triazole antimycotic with activity against most species of pathogenic fungi [163].

Pharmacokinetics. After intravaginal application, 5 – 16 % of the dose is aborbed. The substance is extensively metabolized and eliminated via the urine and feces [163].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts. In addition, terconazole is indicated for theraphy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata* [167], [168].

Side Effects. *CNS.* After intravaginal application of terconazole, a flu syndrome (fever, headache, hypotension) was observed [169].

Local irritations and allergic reactions may occur in rare cases and are mainly due to the galenic formulation.

2.2.13. Tioconazole

Tioconazole [*65899-73-2*], 1-[2-[(2-chloro-3-thienyl)methyloxy]-2-(2,4-dichlorophenyl)ethyl]-1*H*-imidazole, $C_{16}H_{13}Cl_3N_2OS$, M_r 387.70, crystals.

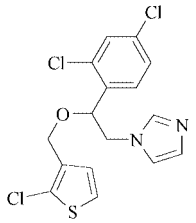

Synonym(s). TIO, IK 20 349.

Stability. The pure solid substance is stable under normal storage conditions.

Synthesis. [170]:

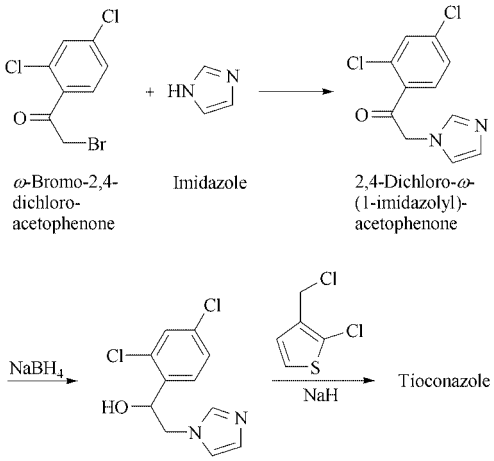

Description. The imidazole antimycotic tioconazole was first synthesized at Pfizer Pharmaceuticals [171], [172].

Formulations. 1 % solutions, lotions, ointments, and cream preparations for therapy of dermatomycoses; 28 % solutions for onychomycoses; 2 – 6 % cream preparations for vaginal candidiasis.

Trade Names. Dermo-Trosyd (Pfizer), Fungibacid (Asche), Gino-Tralen (Pfizer), Gyno-Trosyd (Mason, Pfizer), Tralen (Pfizer), Trosid (Pfizer), Trosyd (Mason, Pfizer, Roerig), Trosyl (Roerig, Novex, Pfizer), Vagistat (Bristol-Myers-Sqibb), Zoniden (Irbi).

Antimycotic Properties. Tioconazole is a broad-spectrum imidazole antimycotic with activity against almost all species of pathogenic fungi [173–175].

Pharmacokinetics. Tioconazole is normally not absorbed after topical application to the skin [173]. On vaginal application of an ovulum containing 300 mg tioconazole, a serum concentration of 21 µg/mL was observed after 8 h [174].

Indications. Topical treatment of mycoses of the skin induced or sustained by fungi such as dermatophytes and yeasts [176]. In addition, tioconazole is indicated for therapy of vulvovaginal mycoses caused by *Candida* species and *Candida (Torulopsis) glabrata*.

Side Effects. Local irritations such as itching, burning sensations, erythema, and dermatitis, and allergic reactions may occur in rare cases and are mainly due to the galenic formulation [177], [178].

3. Polyene Antimycotics

The mechanism of action of polyene antimycotics is based on the formation of a complex with ergosterol in the fungal plasma membrane. This complex causes changes in the permeability of the plasma membrane, loss of ions and low molcular mass cytoplasmatic components, and inhibition of glycolysis [179]. The effect of polyenes on fungi is antagonized by Ca^{2+}, Mg^{2+}, and sterols in vitro. Further, an antagonism between polyene and azole antimycotics has been observed in several cases, both in vitro

and in vivo. This antagonism can be explained in terms of both substance classes interfering with ergosterol biosynthesis.

The polyenes are primarily fungistatic at concentrations around the MIC level. In vitro fungicidal effects are achieved with 2, 4 or up to 10 times the MIC. The three relevant polyene antimycotics amphotericin B, natamycin (pimaricin), and nystatin show broad-spectrum in vitro antimycotic activity. Amphotericin B is the most potent of these three substances. However, in vivo activity is confined to yeasts, dimorphic (biphasic) fungi, and moulds, including *Aspergillus* species. Generally, the polyenes are less active to mostly ineffective and therefore generally not indicated in the treatment of dermatophytoses.

Polyenes are not absorbed sufficiently, if at all, on oral administration. Parenteral administration causes serious local inflammations. Nystatin and pimaricin are therefore unsuitable for intravenous administration.

3.1. Amphotericin B

Amphotericin B [*1397-89-3*], natural product, 38-membered cyclic lactone, isolated, e.g., from culture filtrates of *Streptomyces nodosus*. There are seven double bonds in the molecule. The keto group in position 1 is cyclized with the hydoxl group in position 35 to form a hemiacetal. The carbon atom at position 33 is attached to the dideoxy sugar mycosamine by a β-glucoside bond. $C_{47}H_{73}NO_{17}$, M_r 924.1, *mp*: does not have an exact melting point and decomposes above 170 °C, amorphous yellow powder.

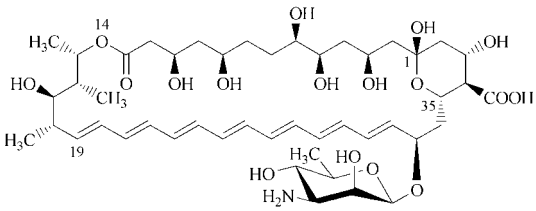

Synonym(s). RP 17774.

Solubility. Insoluble in water at pH 6 – 7. Solubility in water at pH 2 or pH 11: 0.1 mg/mL. Soluble in aqueous ethanol; DMF: 2 – 4 mg/mL; DMF + HCI: 60 – 80 mg/mL; DMSO: 30 – 40 mg/mL; insoluble in ether, ethyl acetate, or benzene. Soluble in aqueous sodium desoxycholate.

Stability. Amphotericin B is unstable to heat and prolonged exposure to light. It should be stored in the dark and kept refrigerated. In solution it decomposes at pH < 4 and > 10 [180]. Solutions show an optimum stability in a citrate – phosphate buffer at pH 5 – 7 [181–183].

Synthesis. Mostly fermentative from *Streptomyces* species [184]. First isolated from a culture of *Streptomyces nodosus* M4575, an isolate obtained from soil of the Orinoco river region of Venezuela.

Description. The first gereal characterization appeared in 1961 [185]. After some partial structures were determined, the complete structure was published in 1970 [186], and the total synthesis in 1987 [187], [188].

Formulations. For systemic treatment: Ampules à 0.05 g. Liposomal formulations with reduced toxicity [189–191] are available for systemic treatment. For local therapy, amphotericin B formulations such as cream preparations and ointments, lozenge formulations (à 0.1 and 0.01 g), and suspensions (100 mg/mL) are used. Also topical cream formulations additionally containing antibiotics (e.g., neomycine) and/or steroids (e.g., triamcinolone) are on the market.

Trade Names. AmBisome (Vestar), Ampho-Moronal (Bristol-Myers Squibb, Heyden, Squibb), Funganiline (Med. y Prod. Quim.), Fungilin (Bristol-Myers Squibb, Novo Nordisk, Squibb), Fungizone (Bristol-Myers Squibb, Squibb).

Antimycotic Properties. Amphotericin B is a broad-spectrum polyene antimycotic with activity against pathogenic yeasts, moulds, dimorphic fungi, and in vitro also against dermatophytes [192], [193]. MIC values of amphotericin B are summarized in Table 7.

Table 7. MIC values for amphotericin B, nystatin, and pimaricin

Fungus	MIC range, µg/mL		
	amphotericin B	nystatin	pimaricin
Candida species	0.1 – 0.5	1 – 5	1.5 – 10
Cryptococcus neoformans	0.1 – 0.5	0.5 – 4	6 – 12.5
Aspergillus species	0.1 – 0.5	1 – 10	3 – 12
Coccidioides immitis	0.1 – 0.5	0.5 – 1.5	2.5 – 25
Histoplasma capsulatum	0.1 – 1	0.5 – 1	3
Blastomyces species	0.1 – 0.5	0.5 – 1	1.5
Dermatophytes	2 – 4	2 – 10	1.2 – 100

Pharmacokinetics. Amphotericin B is scarcely absorbed after oral application. Therefore, amphtericin B has to be applied by the intravenous route for the treatment of systemic fungal infections [194]. Intravenous administration of 5 – 75 mg/d leads to serum concentrations of 0.14 – 2.39 mg/L 4 h after infusion [194]. The extravascular tissue distribution of amphotericin B is minimal because of cell membrane and lipoprotein binding [195]. The highest amphotericin B tissue concentrations are achieved in the liver and spleen, and also in the lungs and kidneys [196], [197]. Because of its poor penetration of the blood – brain barrier, meningitis and infections of the central nervous system can only be treated intralumbally and/or intrathecally [194]. The metabolization of amphotericin B is still unclear, and no metabolites have been charcterized so far [196]. Only 3 % of the administered dose is found in the urine [198], and 1 – 15 % of the daily dose is eliminated by the biliary – fecal route [197]. Much of the substance is trapped in the organs. Elimination occurs in two phases. The initial half-life is between 1 and 2 d, with a terminal half-life of 15 d [194].

Indications. Amphotericin B is effective in therapy of *Candida* infections, also including generalized and disseminated forms, caused by *Candida albicans* and other pathogenic *Candida* species [199]. Further, amphotericin B is indicated in the therapy of mycotic diseases such as histoplasmosis, sporotrichosis, cryptococcosis, blastomycosis, aspergillosis, zygomycosis, and coccidioidomycosis [200]. Although the substance shows an in vitro activity against dermatophytes, amphotericin B is not indicated in the therapy of dermatophytoses, mostly due to toxicity and stability problems and high cost. In the treatment of generalized mycoses, often a synergistic effect between amphotericin B and fluorocytosine [201], [202], and an antagonistic effect between amphotericin B and azole antimycotics is observed (the latter with the exception of CNS cryptococcosis due to *Cryptococcus neoformans*). Indications and efficacy of amphotericin B are summarized in Table 8. Amphotericin B can be given as a continuous drip infusion, even though serious, and in some cases, irreparable side effects, predominantly to the kidney tubulus, have been observed. Therefore, systemic treatment should be initiated only after exact diagnosis, and the patient should be kept under strict supervision. The optimum dosage for systemic therapy is 1 mg/kg every 24 to 48 h. Despite serious side effects, amphotericin B still remains a drug of choice for treating systemic mycoses.

Table 8. Indications and efficacy of amphotericin B at optimal dosage

Indication	Pathogen	Clinical efficacy
Candidosis	*Candida* species, mostly *C. albicans*	effective in > 50 % of cases
Aspergillosis	*Aspergillus* species except *A. nidulans*	uncertain
Zygomycosis	*Mucor* and *Absidia* species	uncertain
Blastomycoses	*Blastomyces dermatidis, Bl. brasiliensis*	effective in ≈ 50 % of cases
Coccidioidomycosis	*Coccidioides immitis*	effective
Histoplasmosis	*Histoplasma capsulatum*	usually effective

Side Effects. The following side effects have been observed under systemic therapy:

CNS. Fever, vomiting, and chills are observed in about 50 % and seem to be mediated by prostaglandines.

Heart. Rarely, arrythmia is observed under infusion [203].

Circulation. Collapse.

Vessels. Thrombophlebitis at the injection site is requently observed, liposomal formulations show a lower phlebotoxicity [204].

Blood/Spleen. Anemia and thrombocytopenia [204] are rarely observed.

Liver. In some cases, hepatotoxicity was observed.

Kidney. Amphotericin B shows a high degree of nephrotoxicity. Cumulative doses of 4 – 5 g generally cause reversible damage to the kidneys [205]. Higher doses mostly cause an irreversible kidney damage with persisting hypopotassemia, hematuria, proteinuria, and azotemia [203–206]. Therefore, serum potassium, magnesium, and creatinine levels have to be frequently monitored.

In the case of intralumbal and/or intrathecal administration, severe side effects of the central nervous system are common. Liposomal formulation of amphotericin B can reduce toxicity problems [207].

Topical application of amphotericin B is highly effective and well tolerated in the treatment of cutaneous and mucocutaneous candidoses. This is also true for inhalation and instillation. Local irritations and allergic reactions may occur in rare cases and are partly due to the galenic formulation.

3.2. Natamycin (Pimaricin)

Natamycin [*7681-93-8*], natural product isolated from culture filtrates of *Streptomyces natalensis* (original isolate from soil of Pietermaritzburg, South Africa) and *Streptomyces chattanoogensis* [208], [209], contains mycosamine as sugar component, like amphotericin B and nystatin. $C_{33}H_{47}NO_{13}$, M_r 665.7, *mp*: does not have an exact melting point and decomposes above 200 °C, colorless substance.

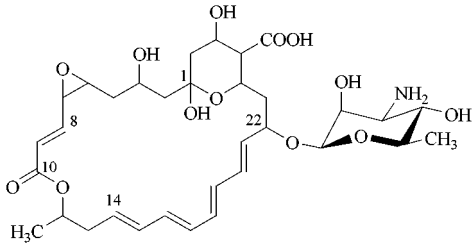

Synonym(s). Pimaricin, antibiotic A 5283, CL 12625, tennecetin.

Solubility. Natamycin is soluble in propylene glycol (20 mg/mL), DMF (50 mg/mL), and *N*-methylpyrrolidone (120 mg/mL); slightly soluble in methanol (2 mg/mL); and practically insoluble in water (0.05 mg/mL). Natamycin is amphoteric, dissolving in dilute acids and bases. However, these solutions are unstable.

Stability. Solutions or suspensions of natamycin are stable for several weeks at pH 5 – 7. The solutions can even be sterilized by heating at 110 °C for 20 min. However, they must be protected from light, which causes rapid decomposition [208], [210].

Synthesis. Fermentative from *Streptomyces* species.

Description. First isolated in 1957 [208], [211], [212].

Formulations. For therapy of human infections: Ointments, powders, creams, lotions, lozenges, dragees, suspensions, and vaginal tablets of different concentration. Also in combination with antibiotics and/or steroids (e.g., neomycine, hydrocortisone).

Trade Names. Deronga (Basotherm), Mycophyt (Gist-Brocades), Myprozine (Lederle), Natacyn (Alcon), Natafucin (Brocades), Pima-Biciron (Basotherm), Pimafucin (Basotherm, Beytout, Byk Procienx, Gist-Brocades), Pimagyn (Doetsch-Grether), Synogil (Basotherm).

Antimycotic Properties. Natamycin is a broad-spectrum polyene antimycotic, MIC values are summarized in Table 7.

Phamacokinetics. After topical application, no systemic absoption was detectable.

Indications. Cutaneous and mucocutaneous candidosis; in some cases, the substance was also successful for therapy of human dermatophytoses [213–215]. In veterinary medicin, natamycin is used for the therapy of dermatophytoses in cows (generally caused by *Trichophyton verrucosum*) and horses (generally caused by *Trichophyton equinum*) [216], [217]. As natamycin also has antibacterial properties, it is also used as a food preservative.

Side Effects. When applied topically, natamycin is highly effective and well tolerated in the treatment of cutaneous and mucocutaneous candidoses. Local irritations and allergic reactions may occur in rare cases and are mostly due to the galenic formulation. Especially in veterinary medicine, phototoxic effects were observed.

3.3. Nystatin A1

Nystatin (A1) [*1400-61-9*], three biologically active components – Nystatin A1, A2, and A3 – have been described. Nystatin A1 is a natural product isolated from culture filtrates of *Streptomyces noursei* [218], *Streptomyces aureus*, and other *Streptomyces* species. Like amphtericin B it is a 38-membered macrocyclic lactone, from which it differs solely in the position of the hydroxyl groups and in the absence of a double bond at C23 [219]. $C_{47}H_{75}NO_{17}$, M_r 926.1, *mp*: does not have a melting point, begins to decompose above 160 °C and decomposes above 250 °C without melting, yellow powder.

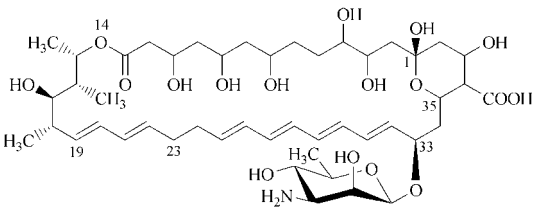

Synonym(s). Polyfungin A1.

Solubility. Nystatin A1 is soluble in pyridine, DMSO, DMF, ethylene glycol (8.7 µg/mL), methanol (11 µg/mL), ethanol (1.2 µg/mL), butanol, dioxane, and water (4.0 µg/mL). The substance is hygroscopic [220].

Stability. The substance is amphoteric, but aqueous and alkaline solutions are unstable. Nystatin shows optimum stability in phosphate – citrate buffers at pH 5 – 7. If kept refrigerated, the pure substance can be stored for several months without loss of activity. For stability, see [221].

Synthesis. Fermentative from *Streptomyces* species.

Description. See [222], [223].

Formulations. Tablets, dragées, lozenges, suspensions, drops, ointments, powders, gels, creams, ovulas, and vaginal tablets of different concentrations. There are also formulations available in combination with antibiotics and/or steroids such as tetracycline, neomycin, gramicidin, and cortisone. Dosages are often expressed in IE (100 000 IE = 22.73 mg nystatin A1).

Trade Names. Adiclair (Ardeypharm), Biofanal (Pfleger), Candex (Dome), Candio-Hermal (Hermal, Merck), Canstat (Lederle), Diastatin (Pfizer), Fungicidin (Spofa), Fungireduct (Azupharma), Herniocid (Mayrhofer), Korostatin (Holland-Rantos), Lederlind (Lederle), Lystin (Mekim), Mikostatin (Squibb), Moronal (Heyden, Squibb), Multilind (Bristol-Myers Squibb, Fair), Mycostatin (Bristol-Myers Squibb, Heyden, Sanofi Winthrop, Westwood-Squibb), Mykinac (NMC), Myco-Posterine N (Kade), Mykundex (Jossa), Nadostine (Pan-Well), Nilstat (Lederle), Nyaderm (Taro), Nysert (Norwich Eaton), Nystacid (Farmos Group), Nystaderm (Dermapharm), Nysta-Dome (Dome), Nystan (Squibb), Nystat-Rx (Pharma-Tek), Nystavescent (Squibb), Nystex (Savage), Oranyst (Taro), O-V Statin (Squibb), Restatin (Remedica), Rivostatin (Rivopharm) Stereomycin (Medica).

Antimycotic Properties. Nystatin A1 is a broad-spectrum polyene antimycotic with in vitro activity against pathogenic yeasts, moulds,

dermatophytes, and dimophic fungi. MIC values of nystatin are summarized in Table 12.

Pharmacokinetics. After topical application, no systemic absorption was detectable.

Indications. Nystatin A1 is especially indicated in the therapy of cutaneous and mucocutaneous infections caused by pathogenic yeasts.

Side Effects. When applied topically, nystatin A1 is highly effective and well tolerated in the treatment of cutaneous and mucocutaneous candidoses. Local irritations and allergic reactions may occur in rare cases and are mostly due to the galenic formulation. If topically applicated in the gastrointestinal tract, nausea, vomiting, and diarrhoea are rarely observed.

4. Flucytosine

Flucytosine [*2022-85-7*], 4-amino-5-fluoro-2(1*H*)-pyrimidinone, $C_4H_4FN_3O$, M_r 129.03, *mp* 297 °C (decomp.), colorless crystals.

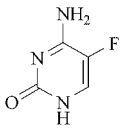

Synonym(s). Fluorocytosine, 5-Flurocytosine, 5-FC, Ro 2-9915.

Solubility. Flucytosine is readily soluble in water. It is basic and forms salts with acids.

Stability. Pure flucytosine as well as aqueous solutions, both acidic and basic, are stable. The substance should be stored in the dark.

Synthesis. The synthesis begins with *S*-ethylisothioureahydrobromide and ethyl-2-fluoro-3-methoxy-acrylate [224], [225]:

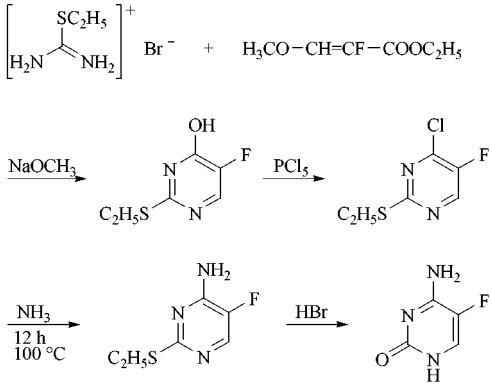

Description. The substance was initially synthesized at Hoffmann-La Roche Pharmaceuticals as an antimetabolite [224], [226–228].

Formulations. Infusion bottles à 0.1, 0.2, and 0.4 g. Capsules à 0.05, 0.1, 0.15, and 0.2 g. 0.5 and 1 % suspensions for oral application, 10 % suspension creams.

Trade Names. Alcoban (Roche), Alcobon (Roche), Ancobon (Roche), Ancotil (Edward Keller, Hoffmann-La Roche, Roche), Cocol (Horita, Tobishi).

Mechanism of Action. The mechanism of action of flucytosine is based on the incorporation of 5-fluorouracil, which is relatively selectively formed in fungal cells by deamination of flucytosine by the fungal cytosine-desaminase, into the messenger RNA and ribosomal RNA of the fungal cell. Fungal RNA containing 5-fluorouracil-riboside increasingly inhibits fungal protein biosynthesis [231–234].

Antimycotic Properties. Flucytosine has a narrow in vitro antimycotic activity spectrum that includes yeasts, *Cryptococcus neoformans*, as well as some *Aspergillus* strains and some chromomycetes. The activity spectrum and MIC values of flucytosine are summarized in Table 9 [231], [232]. Because the antimycotic action of flucytosine is antagonized by purines, pyrimidines, peptides, and some amino acids in vitro, only media free from these antagonists, e.g., yeast – nitrogen base medium containing dextrose, can be used for in vitro testing. Of the

fungus strains affected by flucytosine, more than 10 % can be expected to show primary resistance. The development of secondary resistance by sensitive fungi during treatment also occurs, especially in cases of underdosage or prolonged treatment [232], [234]. In vitro fungicidal effects can be obtained with > 10 times the MIC and a contact time of > 150 h.

Pharmacokinetics. Flucytosine can be administered orally, parenterally, and topically. After oral administration, the drug is absorbed rapidly and almost completely in the intestine (80 – 90 %) [235]. Average serum concentrations of 30 – 60 µg/mL are normally reached with a single oral dose of 35 mg/kg. In patients with normal kidney function, optimal continuous serum levels between 25 and 100 mg/L are achieved with a dose of 50 mg/kg every 6 h [236]. Flucytosine penetrates readily into the tissues, body fluids, and cerebrospinal fluid (CSF) [237]. The CSF level reaches approximately 70 – 80 % of blood levels [236]. The plasma protein binding of flucytosine is low (ca. 4 %) [238]. More than 90 % of the dose is eliminated via the kidneys in unchanged form. The elimination half-life is 3 – 6 h. In patients with severely impaired kidney function, the elimination half-life can rise to 4 d. In patients with restricted renal function, the daily dose of flucytosine therefore must be adjusted according to creatinine elimination. The compound can be readily eliminated by hemodialysis [231–233]. A small amount of flucytosine is metabolized to 5-fluorouracil, e.g., by the bacterial flora of the gut, which seems to be a reason for the hematotoxicity of flucytos e [239], [240].

Table 9. MIC values for flucytosine

Fungus	MIC range, µg/mL
Candida albicans	0.1 – 10 (> 100)
Candida tropicalis	0.1 – 100
Candida parapsilosis	0.1 – 0.5
Candida pseudotropicalis	0.1
Candida krusei	0.5 – 1
Candida (Torulopsis) glabrata	0.1 – 1
Cryptococcus neoformans	0.5 – 4 (16)
Aspergillus fumigatus	0.5 – 10
Aspergillus niger	0.5 – 1
Aspergillus flavus	1
Aspergillus nidulans	100
Phialophora species	1 – 10
Cladosporium species	1 – 10

Indications. Flucytosine has proved very effective in the treatment of *Candida* fungemia and systemic candidoses (but not mucocutaneous candidiasis) in combination with amphotericin B. Also, *Cryptococcus* infections, including *Cryptococcus neoformans* meningoencephalitis in immunocompromized patients (e.g., AIDS patients), can be treated with flucytosine in combination with amphotericin B (and in the case of *Cryptococcus neoformans* associated infections, additionally fluconazole). Further, flucytosine is indicated in the therapy of chromomycoses [241]. There can also be some activity of flucytosine in disseminated aspergilloses. Because of the rapid development of resistance, the substance should only be given in combination with other antimycotic agents, mostly amphotericin B [242].

Side Effects. Flucytosine is normally well tolerated provided dosage instructions are observed and precautionary measures are taken [232]. The following side effects have been observed during treatment with flucytosine: *Blood/spleen.* Neutropenia, leukopenia, thrombopenia [243].

GIT. Nausea, vomiting, and diarrhoea were observed in rare cases [243].

Liver. Liver malfunctions with reversible hepatomegalia and increase in serum transaminases; an extensive liver cell necrosis was reported after flucytosine therapy in two patients [243].

Skin. Allergic skin reactions were observed under therapy. Changes in blood counts can have serious consequences. Blood counts and liver function should therefore be checked regularly during therapy. Hematotoxic side effects were especially observed if serum levels exceeded 100 mg/L. Flucytosine can prolong engraftment in bone marrow transplantation patients.

5. Griseofulvin

Griseofulvin [*126-07-8*], 7-chloro-4,6-trimethoxy-6′-methylspiro[benzofuran-2(3*H*),1′-[2]cyclohexene]-3,4′-dione, $C_{17}H_{17}ClO_6$, M_r 352.8, *mp* 222 °C, colorless crystals.

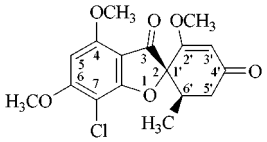

Synonym(s). Amudane, Curling Factor.

Solubility. Griseofulvin is soluble in methanol (2.5 g/mL), ethanol (1.5 g/mL), ether, and glacial acetic acid; quite soluble in DMF (150 mg/mL); slightly soluble in chloroform (25 mg/mL), ethyl acetate, acetone (33 mg/mL), and toluene; and insoluble in petroleum ether. The solubility in water is 40 µg/mL.

Stability. Griseofulvin is stable under normal storage conditions.

Synthesis. Griseofulvin is mostly produced by fermentation with strains of *Penicillium patulum*. After the first total synthesis in 1960 [244], other processes were published, but none of them has acquired industrial significance up to now.

Description. Griseofulvin was isolated from culture broths of *Penicillium griseofulvum* in 1938 and later also found in culture filtrates of *Penicillium nigricans* (= *P. janczewskii*), *P. patulum*, and other *Penicillium* species. The structure was determined in 1952. Griseofulvin was first used in plant protection and was introduced to medicine as an antimycotic in 1958 [245].

Formulations. Tablets à 125 and 500 mg. 5 % cream formulations.

Trade Names. B-GF (Biokema), Delmofulvina (Coli), Fulcin (ICI, Mason, Zeneca), Fulcin S (ICI), Fulvicin (Schering), Fulvicina (Interpharma), Fulvicin P/G (Schering-Plough), Fungivin (Nycomed), Gefulvine (Dif-Dogu), Greosin (Glaxo), Gricin (Arzneimittelwerke Dresden), Grifulin (Teva), Grifulvin V (McNeil, Ortho), Grisactin (Ayerst, Wyeth), Griséfuline (Sanofi Winthrop), Griseo (Ayerst), Griseoderm (Trinity), Griseomed (Waldheim), Griseostatin (Mason, Schering Corp.), Griseo von ct (ct-Arzneimittel), Grisfulvin (Protea), Grisol (Gebro), Grisovin (Glaxo), Grisovina Fp (Glaxo, Teofarma), Grisovin-FP (Glaxo), Grivate (Fujisawa), Grysio (Ayerst), Lamoryl (Leo, Lövens), Likuden (Hoechst), Norofulvin (Norbrook), Polygris (Schering Corp.), Sulvina (Dibios), Ultragris (Sidmak).

Mechanism of Action. In addition to the inhibition of protein synthesis resulting from the interference of griseofulvin with guanine during RNA synthesis, there may also be a specific inhibition of chitin synthesis. Chitin is an essential cell-wall constituent in dermatophytes. Inhibition of its synthesis would explain the characteristic morphological changes in the hyphae, called curling, that occur in dermatophytes in the presence of griseofulvine [231], [232], [246].

Antimycotic Properties. Griseofulvin has a narrow spectrum of antimycotic activity, including only dermatophytes. *Trichophyton* and *Microsporum* species and *Epidermophyton floccosum* are inhibited by concentrations of 0.1 – 0.5 µg/mL. The in vitro effect of griseofulvin depends primarily on the test system used and is confined to proliferating hyphae. Griseofulvin is fungistatic. Purines and pyrimidines are antagonistic in vitro [232], [247]. Acquired in vitro resistance towards griseofulvin develops gradually following a number of fungal passages [248], [249]. However, increases in resistance during therapy are rarely observed, even during long-term treatment. Although dermatophyte variants showing primary resistance have been described, they are very rare [250]. MIC values for griseofulvin are summarized in Table 10 [14].

Table 10. MIC values for griseofulvine

Fungus	No. isolates tested	MIC range µg/mL
Epidermophyton floccosum	6	3 – > 5
Microsporum canis	19	0.1 – > 18
Microsporum gypseum	6	0.5 – 3.13
Trichophyton mentagrophytes	57	0.1 – > 30
Trichophyton rubrum	558	0.1 – > 30
Trichophyton schoenleinii	170	0.1 – 25
Trichophyton tonsurans	25	0.1 – > 18
Trichophyton violaceum	345	0.1 – 21

Pharmacokinetics. The degree of absorption of orally administered griseofulvin varies from individual to individual from 25 to 75 % of the administered dose. The absorption is strongly dependent on the degree of micronization of the drug, a high fat content in the food also increases absorption. After the administration of a single dose of 0.25 – 0.5 g, serum concentrations reach 0.6 to 1.5 µg/mL. The serum half-life is about 20 h [231], [232], [251]. Griseofulvin accumulates during keratinization in human epidermis cells, hair, and nails. In the course of long-term griseofulvin treatment, concentrations up to 25 µg/g can be detected in the stratum corneum of the skin [232]. Griseofulvin concentrations of 0.2 to 0.3 µg/mL are found in the sweat [252]. Further, the substance shows high concentrations in liver, body fat, and muscles. Griseofulvin is metabolized in the liver and is mostly eliminated by the kidneys in a metabolized, inactive form. Up to 75 % of the quantity absorbed is excreted in the urine as inactive 6-demethylgriseofulvin. The concentration of unchanged griseofulvin in the urine is 1 – 2 µg/mL, corresponding to approximately 1 % of the dose [232], [253]. Elimination occurs biphasically with a terminal elimination half-life of 10 – 40 h [254].

Indications. The indications for griseofulvin are dermatophytoses and onychomycoses which do not respond to topical therapy. The optimal daily dose is 5 – 10 mg/kg body weight. The duration of treatment ranges from 3 to 5 weeks for dermatophytoses of the trunk and extremities, and from 5 to 10 weeks for infections of palms, soles, and scalp, but is more than 6 months and even up to 18 months for onychomycoses [231], [232]. Additional topical therapy increases the cure rate and reduces duration of treatment. Due to better tolerability and efficacy, newer antimycotic compounds such as azole antimycotics and allylamines have mostly replaced griseofulvin in the therapy of dermatomycoses and onychomycoses. Griseofulvin is inactive in mycoses caused by fungi other than dermatophytes.

Side Effects. *CNS.* Headache, depression, dizziness, sleeplessness.

Nerves. Paresthesia and neuropathia [255].

GIT. Dryness of the mouth, changes in taste, and other unspecific gastrointestinal disorders.

Skin. Minor allergic skin reactions, Lyell syndrome [256], erythema multiforme, exfoliative dermatitis, angioedema, photosensitization.

Occasionally, reversible hematopathies, proteinuria, and, especially in case of previous hepatopathy, increases in the liver enzyme values in the blood may be observed. A carcinogenic and embryotoxic potential of griseofulvin was observed in laboratory animals such as rats.

6. Ciclopirox

Ciclopirox [*29342-05-0*], 6-cyclohexyl-1-hydroxy-4-methyl-2(1*H*)-pyridinone, $C_{12}H_{17}NO_2$, M_r 207.27, *mp* 144 °C, colorless crystals.

Ciclopirox olamine [*41621-49-2*], 6-cyclohexyl-1-hydroxy-4-methyl-2(1*H*)-pyridinone olamine, $C_{14}H_{24}N_2O_3$, M_r 268.4, *mp* 143 °C, colorless crystals.

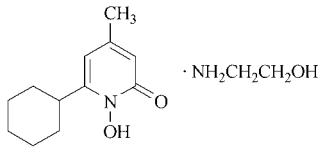

Synonym(s). HOE 296b (free base); Ciclopirox ethanolamine, Hoe 296 (olamine salt).

Stability. The substance is stable under normal storage conditions.

Synthesis. Base **9** is produced by heating 4-methyl-6-cyclohexyl-2-pyrone (**8**) [257] for 8 h at 80 °C with hdroxylamine hydrochloride and 2-aminopyridine [258] or from 5-cyclohexyl-3-methyl-5-oxopentenoate (**10**) [258] by reaction with hydroxylamine hydrochloride via the oxime **11**, followed by cyclization. The final step is to form the salt with ethanolamine.

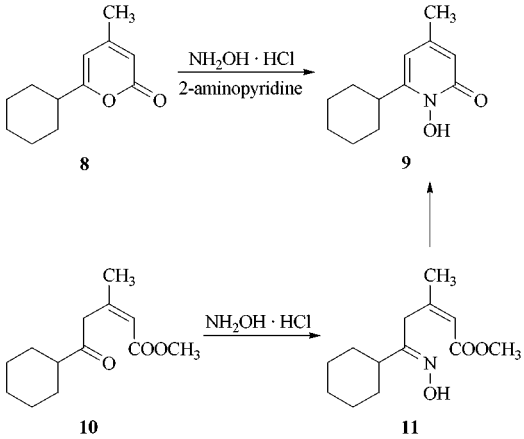

Description. The substance was first synthesized by Hoechst Pharmaceuticals in 1973 [258–262].

Formulations. 1 % solutions, ointments, and cream preparations.

Trade Names (Ciclopirox). Nagel Batrafen (Casella-Riedel).

Trade Names (Ciclopiroxolamine). Batrafen (Casella-Riedel, Hoechst, Knoll), Brumixol (Bruschettini), Ciclochem (Novag), Dafnegin (Poli), Fungiderm (Heilmittelwerke Wien), Fungowas (Wassermann), Loporox (Hoechst-Roussel), Miclast (Ellem), Micomicen (Vita), Micoxolamin (Delalande), (Sinbio), Obytin (Hoechst, Jugoremedija).

Mechanism of Action. Ciclopirox is a substituted pyridione derivative. Experiments show that the mechanism of action involves a disruption of transport mechanisms into the cell plasma [263], [264], especially the uptake of leucine [265]. Further, the uptake of other amino acids such as phenylalanine and lysine, as well as the uptake of potassium and phosphate ions is impaired. This all probably involves a change in membrane permeability, as reported for some other antimycotics, although ciclopirox does not show lytic activity [263], [264]. Consecutively, the substance significantly reduces fungal growth rates [266].

Antimycotic Properties. Ciclopirox has a broad in vitro antifungal activity spectrum, including dermatophytes, yeasts, and moulds. The most important fungi within the activity spectrum of ciclopirox and the corresponding MIC values are listed in Table 11 [263], [264], [267]. Ciclopirox is primarily fungistatic, but if the contact time is long enough and the concentration exceeds 20 µg/mL, it exerts a fungicidal effect on yeasts and dermatophytes [264]. In vitro, the effect is limited to proliferating fungi and is greatly dependent on the nutrient medium.

Table 11. MIC values for ciclopirox olamine

Species	MIC range, µg/mL
Dermatophytes	0.5 – 5
Candida species	0.5 – 5
Moulds	0.5 – 30

Pharmacokinetics. Ciclopirox penetrates rapidly into and through the skin and nail keratin. After dermal application of a 1 % cream formulation, about 1.3 % of the dose is absorbed. Serum levels range around 10 µg/L after application of a dose of 37 mg of ciclopirox [268]. After vaginal administration, 15 – 20 % of the dose is absorbed [269]. Ciclopirox penetrates well into the skin and nail matrix. Absorbed ciclopirox is mainly eliminated by the kidneys; 1.1 – 1.7 % of the topical dose is eliminated in the urine within four days. Only a fraction thereof is unaltered ciclopirox. Approximately 96 % of the absorbed drug is bound to human serum proteins [270].

Indications. Ciclopirox is indicated for the topical treatment of dermatophytoses and vaginal infections caused by *Candida* species [271].

Side Effects. Topically applied, ciclopirox is well tolerated. The incidence of local irritations such as itching and burning sensations range between 1 and 4 % and are often due to the galenic formulation.

7. Thiocarbamates

Thiocarbamates are inhibitors of fungal squalene epoxidase (see also Chap. 8). By this mechanism, tolciclate and tolnaftate inhibit biosynthesis of fungal ergosterol [272].

7.1. Tolciclate

Tolciclate [*50838-36-3*], methyl (3-methlphenyl)carbamothioic acid *O*-(1,2,3,4-tetrahydro-1,4-methanonaphtalen-6-yl) ester, $C_{20}H_{21}NOS$, M_r 323.46, *mp* 120 – 121 °C, white crystals from ethanol – diethylether.

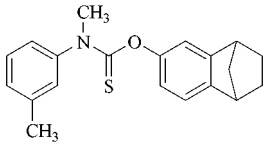

Synonym(s). KC 9147.

Solubility. The substance shows high liposolubility. Soluble in *n*-hexane (15 mg/mL) and *n*-octanol (24 mg/mL).

Stability. The substance is stable under normal storage conditions, but should be protected from light.

Synthesis. [273]:

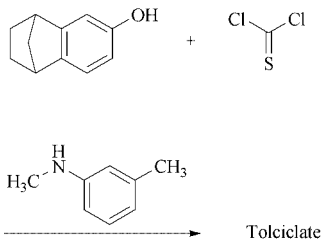

Description. Tolciclate was first synthesized at Carlo Erba Pharmaceuticals [274], [275].

Formulations. 1 % solutions, ointments, and cream formulations.

Trade Names. Fungifos (Basotherm), Kilmicen (Carlo Erba), Tolmicen (Carlo Erba, Wing Yee), Tolmicol (Carlo Erba).

Antimycotic Properties. In vitro, tolciclate has a narrow activity spectrum with good efficacy only towards dermatophytes. Isolates of susceptible species showing primary resistance or developing secondary resistance are rarely observed.

Indications. The indications for tolciclate are defined by its narrow spectrum and limited to the treatment of dermatophytoses. Because of inadequate penetration into the nail keratin, it is unlikely to be effective for the treatment of onychomycoses. On average, the efficacy of tolciclate in topical therapy of dermatomycoses is lower than that of most other antimycotics for topical treatment.

Side Effects. Tolciclate is well tolerated on the skin.

7.2. Tolnaftate

Tolnaftate [*2398-96-1*], (3-methylphenyl)carbamothioic acid *O*-2-naphthalenylmethylester, $C_{19}H_{17}NOS$, M_r 307.4, *mp* 110.5 – 111.5 °C, colorless crystals.

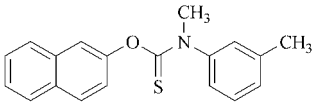

Synonym(s). Naphthiomate-T.

Solubility. The substance shows a good solubility in ethanol, ether, and glacial acetic acid; is quite soluble in DMF, chloroform, dichloromethane, and benzene; and practically insoluble in water.

Stability. Tolnaftate is stable under normal storage conditions, but schould be protected from light.

Synthesis. Tolnaftate is synthesized from β-naphthol, thiophosgene, and *N*-methl-*m*-tolulidine by the following pathway [276]. Another route, which does not use thiophosgene, has been described [277].

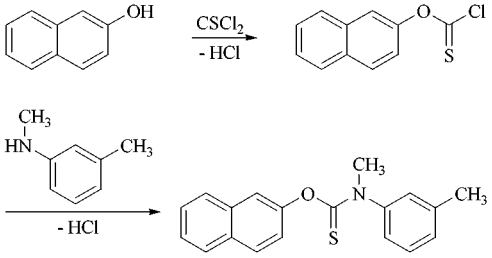

Description. Tolnaftate was first synthesized by Japan Soda and first described as a topical antimycotic in 1962 [278–281].

Formulations. 1 % solutions, ointments, and cream preparations.

Trade Names. Aftate (JDH, Plough), Alber-T (Hing Yip), Chinofungin (Chinoin), Chlorisept (Chinosol), Ezon-T (Yamanouchi), Focusan (Lundbeck), Footwork (Lederle), Hi-Alarzin (Hing Yip, Yamanouchi), Mikoderm (Akdeniz), NP-27 (Thompson), Pedesal (Krka), Pediderm (Nelson), Pedimycose (Scholl), Pitrex (Taro), Separin (Sumitomo), Sorgoa (Scheurich), Sporiderm (Cétrane), Sporiline (Schering-Plough), Tinacidin (Schering Corp.), Tinactin (Schering-Plough), Tinatox (Brenner), Tinavet (Schering), Tineafax (Wellcome), Ting (Fisons), Tonoftal (Schering), Zeasorb-AF (Stiefel).

Antimycotic Properties. In vitro, tolnaftate has a narrow activity spectrum with fungistatic effect on towards dermatophytes and a few moulds such as some isolates of selected *Aspergillus* species. The MIC values range between 0.1 and 3 µg/mL for susceptible fungi. Isolates of susceptible species showing primary resistance or developing secondary resistance are rarely observed. Tolnaftate is almost ineffective against yeasts and moulds.

Pharmacokinetics. Tolnaftate combines good penetration into the stratus corneum of the skin with poor penetration into nail keratin.

Indications. The indications for tolnaftate are defined by its narrow spectrum and limited to the treatment of dermatophytoses. Tolnaftate is effective when topically administered several times daily over a period of several weeks [231], [232]. Because of its inadequate penetration into nail keratin, tolnaftate is unlikely to be effective for the treatment of onychomycoses. On the average, the efficacy of tolnaftate in topical therapy of dermatomycoses is lower than that of most other antimycotics for topical treatment [282].

Side Effects. Tolnaftate is well tolerated on the skin. Erythemas and allergic reactions are very rare and might also be due to the galenic formulation. No other side effects are known [232], [283].

8. Allylamines

The allylamines are inhibitors of the enzyme squalene epoxidase. This results in an accumulation of squalene, which exhibits lethal effects towards the fungal organism at high concentrations. Thereby, the inhibition of squalene epoxidase blocks ergosterol biosynthesis in the fungal cell [284]. The essential portion of the allylamine molecule is the conjugated enzyme group with a *trans* configuration in the side chain [285]. The enzyme squalene epoxidase is not linked to the cytochrome P450 complex.

8.1. Naftifine

Naftifine [*65472-88-0*], (*E*)-*N*-Methyl-*N*-(3-phenyl-2-propenyl)-1-naphthalenemethenamine, $C_{21}H_{21}N$, M_r 287.40, *bp* 162 – 167 °C, colorless, viscous oil.

Naftifine hydrochloride [*65473-14-5*], $C_{21}H_{22}ClN$, M_r 323.9, *mp* 177 °C, crystals from propanol.

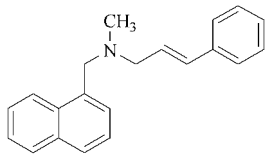

Synthesis. [286], [287]:

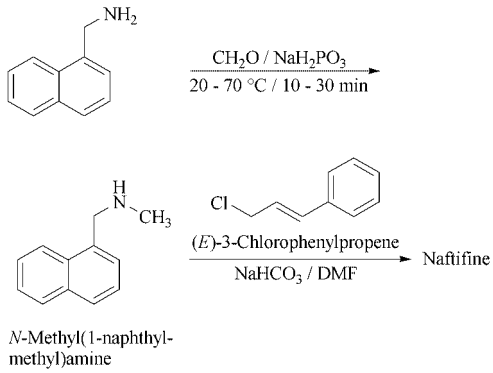

Description. See [288–290].

Formulations. 1 % creams, ointments, gels, and solutions.

Trade Names (Naftifine Hydrochloride). Exoderil (Allergan, Novartis, Rentschler, Schering), Fotimin (KRKA), Nafteryl (Schering), Naftin (Allergan).

Antimycotic Properties. Naftifine is fungicidal against dermatophytes such as *Epidermophyton floccosum, Trichophyton* and *Microsporum* species. Against pathogenic yeasts such as *Candida* species and moulds, it shows only an intermediate fungistatic activity in vitro [291], [292].

Pharmacokinetics. Naftifine penetrates well into the stratum corneum of the skin; 2 – 4 % of the topically administered dose were absorbed after administration of a 1 % gel preperation. After occlusion, an absorption of 6 % of the administered dose was observed [293].

Indications. Naftifine is primarily indicated in the therapy of superficial dermatomycoses [294–296]. It also shows some clinical efficacy in superficial infections caused by yeasts and moulds.

Side Effects. Naftifine is well tolerated although rare cases of local skin irritations and contact dermatitis have been found, which might also be due the galenic formulation [293].

8.2. Terbinafine

Terbinafine [*91161-71-6*], (*E*)-*N*-(6,6-dimethyl-2-hepten-4-ynyl)-*N*-methyl-1-naphthalenemethanamine, $C_{21}H_{25}N$, M_r 291.44, colorless crystals.

Terbinafine hydrochloride [*78628-80-4*], $C_{21}H_{26}ClN$, M_r 327.90, *mp* 195 – 198 °C, colorless crystals from 2-propanol diethyl ether, change in crystal structure begins at 150 °C.

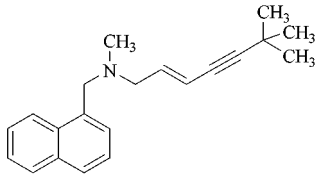

Synonym(s). SF 83627, SF 83-627 (both hydrochloride).

Solubility. Terbinafine is a highly lipophilic substance.

Stability. The substance is stable under normal storage conditions.

Synthesis. [297] (see top of next page):

Description. The substance was first synthesized at Sandoz Pharmaceuticals [298].

Formulations. 1 % cream and ointment preparations, and tablets à 250 mg.

Trade Names (Terbinafine). Lamisil (Edward Keller).

Trade Names (Terbinafine Hydrochloride). Lamisil (Sanabo, Sandoz-Wander).

Antimcotic Properties. Terbinafine shows a good and mostly fungicidal in vitro antimycotic activity against dermatophytes such as *Epidermophyton floccosum, Trichophyton,* and *Microsporum* species. The substance also shows a high but often fungistatic in vitro activity against isolates of various *Aspergillus* species. The activity against pathogenic yeasts is quite variable. Many isolates of *Cryptococcus neoformans* and *Candidia parapsilosis* show an in vitro sensitivity towards terbinafine, although isolates of *Candida albicans, Candida (Torulopsis) glabrata,* and *Candida tropicalis* are mostly resistant or have rather high MIC levels [299].

Pharmakocinetics. Terbinafine is well absorbed from the gastrointestinal tract with a bioavailability of 70 – 80 %. Peak plasma concentrations of 0.97 µg/mL were observed following a single oral dose of 250 mg within 2 h

Terbafin hydrochloride (Synthesis 1):

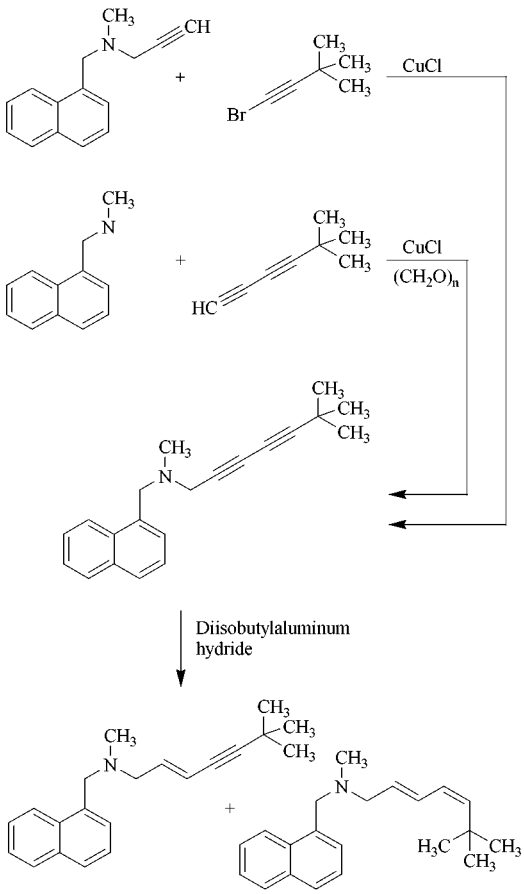

Terbafin hydrochloride (Synthesis 2):

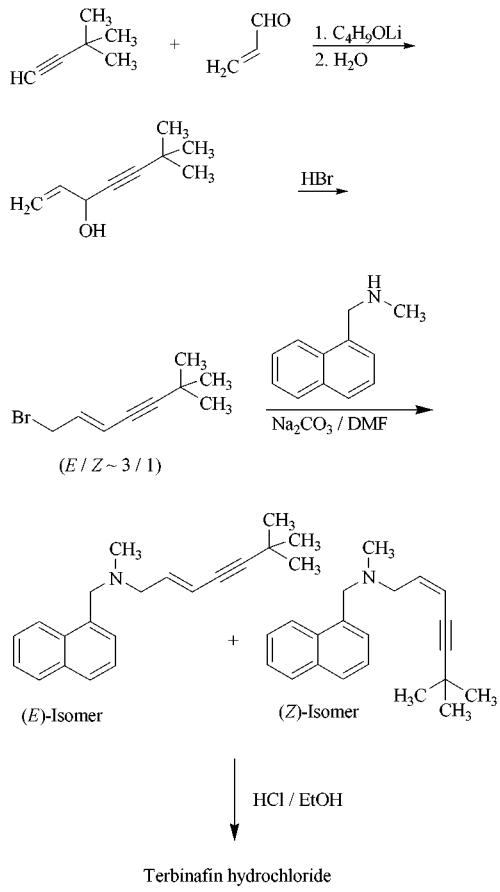

of administration. Steady state concentrations are reached within two weeks of therapy [300]. The substance is highly bound to plasma proteins. After oral administration, terbinafine distributes well into tissues, including skin, the nail plate, sweat, sebum, and hair. Terbinafine is metabolized in the liver to inactive metabolites which are excreted in the urine. Elimination occurs biphasically. Plasma elimination half-lifes of 11 – 17 h have been reported, as has a slower elimination half-life of 90 – 100 h [301]. After topical administration, the lipophilic substance shows a good penetration into the skin and nail matrix and is hardly absorbed (< 5 % of the administered dose) [302].

Indications. Oral and topical treatment of dermatophytoses and onychomycoses [303]. The substance also shows activity in the topical treatment of pityriasis versicolor and cutanous candidiasis. 250 mg of terbinafine are given once daily by mouth for 2 to 4 weeks for therapy of tinea cruris and tinea corporis. Especially for tinea pedis, treatment may have to be continued for up to 6 weeks [292].

Side Effects. The following side effects were observed after oral administration:

CNS. Headache.

GIT. Nausea, diarrhoea, anorexia, mild abdominal pain (5 %) [301].

Arthralgia, myalgia, and skin reactions such as rush and urticaria rarely occur. Severe

skin reactions including erythema multiforme (1:100,000), Stevens – Johnson syndrome and Lyell syndrome have occurred occasionally. Loss or disturbance of taste, which is mostly reversible [304], [305], and liver disfunction have been observed [306]. After topical administration, the substance is well tolerated, although side effects such as skin irritations (e.g., itching and burning sensations) and allergic reaction may occur, which may also often be due to the galenic formulation.

9. Amorolfine

Amorolfine [*78613-35-1*], (±)-*cis*-2,6-dimethyl-4-[2-methyl-3(*p-tert*-pentylphenyl)propyl]morpholine, $C_{21}H_{35}NO$, M_r 317.60, *bp* 134 °C.

Amorolfine hydrochloride [*78613-38-4*], $C_{21}H_{36}ClNO$, M_r 354.06.

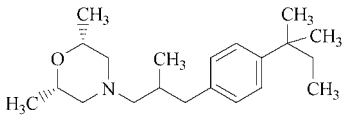

Synonym(s). Ro 14-4767/000 (free base); Ro 14-4767/002 (hydrochloride).

Description. The substance was first synthesized at Roche Pharmaceuticals [307], [308].

Formulations. 0.25 % cream preparations, 5 % nail lacquers, 50 mg vaginal tablets.

Trade Names (Amorolfine Hydrochloride). Loceryl (Roche, Sauter), Locéryl (Roche).

Mechanism of Action. Amorolfine is a morpholine derivative and selectively inhibits Δ^{14}-reduction and Δ^{8}-Δ^{7}-isomerase, which are required for Δ^{14}-reductase and Δ^{8}-Δ^{7}-isomerization, following C14 demethylation of lanosterol or 24-methlenedihydrolanosterol in the ergosterol biosnthesis pathway of the fungal cell [309].

Antimycotic Properties. Amorolfine is a broad-spectrum antimycotic [310] with in vitro activity against dermatophytes, pathogenic *Candida* species, and a variety of pathogenic dimophic fungi such as *Blastomyces dermatitidis, Histoplasma capsulatum*, and *Sporothrix schenckii*. The substance also shows some activity against various isolates of *Aspergillus* species.

Pharmacokinetics. Amorolfine is not absorbed after topical administration on normal skin. After application of a single vaginal tablet containing 50 mg of amorolfine, plasma concentrations between 27 and 83 ng/mL were observed [311].

Indications. For the treatment of skin infections, a 0.25 % cream preparation is applied once daily until clinical cure is achieved and then for a further 3 – 5 d [312]. For the treatment of nail mycoses caused by dermatophytes, yeasts, and moulds, a lacquer preparation containing 5 % amorolfine is painted onto the affected nail once or twice a week until the nail has regenerated. Treatment of onychomycoses normally needs to be continued for up to 1 year. Amorolfine is also active in the treatment of vaginal candidiasis.

Side Effects. Skin irritations, such as erythema, pruritus, or a burning sensation, rarely occur.

10. Unspecific Topical Antimycotics

Other unspecific antimycotics for topical use are listed in Table 12. These somewhat "historical" substances are of minor importance in clinical use nowadays, as highly effective and safe specific substances have become available.

11. Recent Developments and Outlook

The antimycotic market is clearly dominated by azoles, especially for the treatment of systemic mycoses. The latter market is led by fluconazole and itraconazole. Improved azoles are in clinical development (mostly phase I and II); pradimicins, aureobasidins, echinocandins, and pneumocandins for the treatment of systemic mycoses. The new azoles are reported to have

Table 12. Other unspecified topical antimycotics

Generic name	Structural formula	Activity spectrum
Quaternary ammonium compounds		
Benzalkonium chloride [*8001-54-5*]	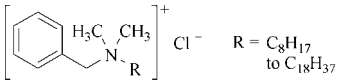	Dermatophytes, yeasts, moulds
Domiphen bromide [*13900-14-6*]	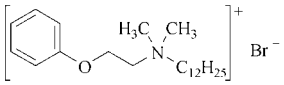	Dermatophytes, yeasts, moulds
Dibromosalane [*87-12-7*]	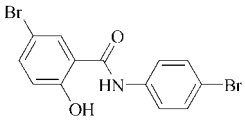	Dermatophytes, yeasts, gram-positive bacteria
Chloroquinaldol [*72-80-0*]	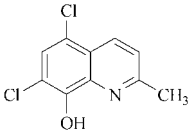	Dermatophytes, yeasts
Dichlorophene [*97-23-4*]	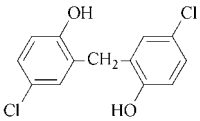	Dermatophytes, yeasts
Hexachlorophene [*70-30-4*]	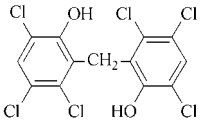	Dermatophytes, yeasts, gram-positive bacteria
Dibenzthione [*350-12-9*]	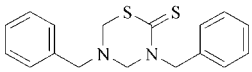	Dermatophytes, yeasts
Chlormidazole [*3689-76-7*]	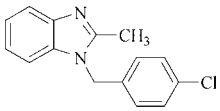	Dermatophytes, yeasts
Mercury compounds		
Merbromin (Mercurochrome) [*129-16-8*]	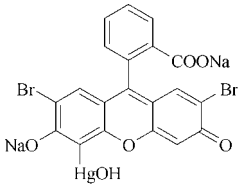	Dermatophytes, yeasts, bacteria
Thimerosal [*54-64-8*]	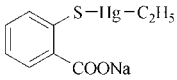	Dermatophytes, yeasts, bacteria
Triphenylmethane dyes		
Fuchsin (basic violet) [*632-99-5*]	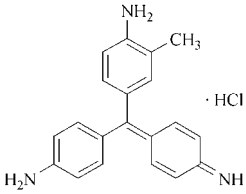	Dermatophytes, yeasts
Undecylenic acid [*112-38-9*]	$CH_2=CH(CH_2)_8COOH$	Dermatophytes, yeasts

activity against fluconazole-resistant strains and species as well as a broader spectrum of antifungal activity. Except the azoles, all other reported compounds habe a relatively low bioavailability and/or tolerability on the average which may limit their use to therapy of life threatening infections within the hospital segment. Production costs of all these compounds seem to be rather high as compared, e.g., to fluconazole.

Significant increase in research activities in antifungals for systemic application has been observed within the last few years in many pharmaceutical companies and institutes worldwide due to the high medical need and the fast growing market. Therefore, the chances for the development of new antifungal agents with new chemical structures and new mechanisms of action are improving.

12. References

1. K. J. Kwon-Chung et al.: *Medical Mycology,* Lea & Febiger, Philadelphia 1992.
2. K. H. Büchel et al., in J. S. Bindra et al. (eds.): *Chronicles of Drug Discovery,* **vol. 2,** J. Wiley, New York 1984.
3. H. Yamaguchi et al., *Sabouraudia* **17** (1979) 311.
4. H. Van den Bossche et al., *Biol. Interact.* **21** (1978) 59.
5. H. Van den Bossche et al., in W. Siegenthaler et al. (eds.): *Current chemotherapy,* Am. Soc. Microbiol., Washington, D.C. 1978 p. 228.
6. J. Haller, *Abstracts XII – Intl. Congr. Microbiol.,* Munich Sept. 1978, no. 38/3.
7. H. Van den Bossche et al., *Arch. Int. Physiol. Biochim.* **87** (1979) 849.
8. H. Van den Bossche in M. R. McGinnis (ed.): *Current Topics in Medical Mycology,* **vol. 1** Springer, New York 1985, p. 313.
9. W. R. Ness et al., *J. Biol. Chem.* **253** (1978) 6218.
10. K. E. Bloch, *CRC Crit. Rev. Biochem.* **7** (1979) 1.
11. Pfizer, GB 2 099 818, 1982.
12. Pfizer, US 4 404 216, 1983.
13. T. E. Rogers et al., *Antimicrob. Agents Chemother.* **30** (1986) 418.
14. M. R. McGinnis et al. in V. Lorian (ed.): *Antibiotics in Laboratory Medicine,* Williams & Wilkins, Baltimore 1996.
15. K. W. Brammer et al., *Rev. Infect. Dis.* **12** (1990) Suppl. 1, S 318.
16. J. E. Thorpe et al., *Antimicrob. Agents Chemother.* **34** (1990) 2032.
17. M. J. Humphrey et al., *Antimicrob. Agents Chemother.* **28** (1985) 648.
18. E. T. Houang et al., *Antimicrob. Agents Chemother.* **34** (1990) 909.
19. R. M. Tucker et al., *Antimicrob. Agents Chemother.* **32** (1988) 369.
20. K. W. Brammer, *Br. J. Obstet, Gyn.* **96** (1989) 226.
21. B. Dupont et al., *J. Med. Vet. Mycol.* **26** (1988) 67.
22. J. A. Como et al., *N. Engl. J. Med.* **33** (1994) 263.
23. J. E. Mangino et al., *Lancet* **345** (1995) 6.
24. J. H. Rex et al., *Antimicrob. Agents Chemother.* **39** (1995) 1.
25. J. E. Bennet et al., *N. Engl. J. Med.* **301** (1979) 126.
26. W. G. Powderly et al., *N. Engl. J. Med.* **326** (1992) 793.
27. M. Rozenberg-Arska et al., *J. Antimicrob. Chemother.* **27** (1991) 369.
28. J. Heeres et al., *J. Med. Chem.* **27** (1984) 894.
29. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 8,** Springer, Berlin – Heidelberg – New York 1993, p. 634.
30. Janssen, EP 6 711, 1980.
31. Janssen, US 4 267 179, 1981.
32. D. A. Borelli, *Rev. Infect. Dis.* **9** (1987) 57.
33. M. Bogers et al., *Rev. Infect. Dis.* **9** (1987) 33.
34. G. Cauwenbergh et al., *Drug Dev. Res.* **8** (1986) 317.
35. J. D. Cleary et al., *Ann. Pharmacother.* **26** (1992) 502.
36. D. W. Denning et al., *Arch. Intern. Med.* **149** (1989) 2301.
37. T. S. Jennings et al., *Ann. Pharmacother.* **27** (1993) 1206.
38. R. Negroni et al., *Rev. Infect. Dis.* **9** (1987) 47.
39. J. W. Van't Wout et al., *J. Infect.* **22** (1991) 45.
40. J. W. Van't Wout et al., *J. Infect.* **20** (1990) 147.
41. J. Heykants et al., *Mycoses* **32** (1989) Suppl. 1, 67.
42. T. C. Hardin et al., *Antimicrob. Chemother.* **32** (1988) 1310.
43. J. Heykants et al. in R. A. Fromtling (ed.): *Recent Trends in the Discovery, Development and Evaluation of Antifungal Agents,* J. R. Prous, Barcelona 1987, p. 223.
44. S. M. Grant et al., *Drugs* **37** (1989) 310.
45. J. Faegermann, *Mycoses* **31** (1988) 377.

46. E. Van Hecke et al., *Mycoses* **31** (1988) 641.
47. T. Piepponen et al., *J. Antimicrob. Chemother.* **29** (1992) 195.
48. G. E. Stein et al., *Antimicrob. Agents Chemother.* **37** (1993) 89.
49. W. E. Dismukes et al., *Am. J. Med.* **93** (1992) 489.
50. J. S. Hostetler et al., *Chemotherapy* **38** (1992) Suppl. 1, 12.
51. J. R. Graybill et al., *Am. J. Med.* **89** (1990) 282.
52. A. P. Lavrijsen et al., *Lancet* **340** (1992) 251.
53. R. M. Tucker et al., *J. Antimicrob. Chemother.* **26** (1990) 561.
54. Janssen Pharmaceutica, DE-OS 2 804 094, 1978 (J. V. Heeres et al.).
55. J. V. Heeres et al., *J. Med. Chem.* **22** (1979) 1003.
56. Janssen, DE 2 804 096, 1978.
57. Janssen, US 4 144 346, 1979.
58. Janssen, US 4 223 036, 1980.
59. N. R. Blatchford et al., *Lancet* **320** (1982) 770.
60. A. Tavihan et al., *Gastroenterology* **90** (1986) 443.
61. T. K. Daneshmend et al., *Clin. Pharmacokin.* **14** (1988) 13.
62. G. Lake-Bakaar et al., *Ann. Int. Med.* **109** (1988) 471.
63. T. K. Daneshmend et al., *Antimicrob. Agents Chemother.* **25** (1984) 1.
64. M. D. Ene et al., *Br. J. Clin. Pharmacol.* **17** (1984) 173.
65. A. M. Sugar et al., *Antimicrob. Agents Chemother.* **31** (1987) 1874.
66. G. Cauwenbergh, *Mycoses* **32** (1989) Suppl. 2, 59.
67. R. De Felice et al., *Antimicrob. Agents Chemother.* **19** (1981) 1073.
68. A. Pont et al., *Arch. Int. Med.* **142** (1982) 2137.
69. J. H. Lewis et al., *Gastroenterology* **86** (1984) 503.
70. Janssen Pharmaceutica, US 3 717 655, 1969 (E. F. Godefroi et al.).
71. E. F. Godefroi et al., *J. Med. Chem.* **12** (1969) 784.
72. Janssen, DE 1 940 388, 1970.
73. Janssen, US 3 717 655, 1973.
74. L. Brugmans et al., *Arch. Dermatol.* **102** (1970) 428.
75. R. Godts et al., *Arzneim. Forsch. / Drug Res.* **21** (1971) 256.
76. R. C. Heel et al., *Drugs* **19** (1980) 7.
77. J. R. Graybill et al., *Drugs* **25** (1983) 41.
78. A. Kucers et al., in A. Kucers et al., (eds.): *The Use of Antibiotics,* Lippincott, Philadelphia 1987, p. 1492.
79. T. K. Daneshmend et al., *Clin. Pharmacokin.* **8** (1983) 17.
80. D. A. Stevens, *Drugs* **26** (1983) 347.
81. V. Feinstein et al., *Ann. Intern. Med.* **93** (1980) 432.
82. Bayer, DE-OS 2 461 406, 1976 (E. Regel et al.).
83. Bayer, DE 2 461 406, 1976.
84. Bayer, FR 2 295 747, 1976.
85. Bayer, DE 2 643 563, 1978.
86. Bayer, US 4 118 487, 1978.
87. W. Ritter, *Proc. VII Int. Congr. ISHAM Palmerston North* (1982) 105.
88. H. Weber, *Proc. VII Int. Congr. ISHAM Palmerston North* (1982) 105.
89. S. Stettendorf, *Arzneim. Forsch. / Drug Res.* **33** (1983) 750.
90. T. Anik et al., *J. Pharm. Sci.* **70** (1981) 8.
91. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 7,** Springer, Berlin – Heidelberg – New York 1993, p. 581.
92. Syntex, US 4 078 071, 1978.
93. K. A. M. Walker et al., *J. Med. Chem.* **21** (1978) 840.
94. F. C. Odds et al., *J. Antimicrob. Chemother.* **14** (1984) 105.
95. W. Droegemueller et al., *Obstst. Gynecol.* **64** (1984) 530.
96. J. B. Jacobson et al., *Acta. Obstet. Gynecol. Scand.* **64** (1985) 241.
97. C. S. Brandbeer et al., *Genitourin. Med.* **61** (1985) 270.
98. W. A. Van Dyk, *J. Reproduct. Med.* **31** (1986) Suppl. 7, 662.
99. K. H. Büchel et al., *Arzneim. Forsch. / Drug Res.* **22** (1972) 1260.
100. Bayer, DE-OS 1 617 481, 1976 (K. H. Büchel et al.).
101. M. Plempel et al., *Dtsch. Med. Wochenschr.* **94** (1969) 1356.
102. Bayer, DE 1 940 626, DE 1 940 627, 1969.
103. Bayer, ZA 6 805 392, ZA 6 900 039, 1969.
104. Bayer, US 3 657 445, US 3 705 172, 1972.
105. W. Ritter, *Proc. VII Int. Congr. ISHAM Palmerston-North* (1982) 105.
106. H. Weber, *Proc. VII Int. Congr. ISHAM Palmerston-North* (1982) 105.
107. S. Stettendorf et al., *Arzneim. Forsch. / Drug Res.* **33** (1983) 750.
108. Shionogi, BE 883 665, 1980.
109. Shionogi, US 4 328 348, 1982.

110. F. Toshiaka et al., *Iyakuhin Kenkyu* **9** (1978) 448.
111. E. F. Godefroi et al., *J. Med. Chem.* **12** (1969) 784.
112. Janssen, DE 1 940 388, 1970.
113. Janssen, US 3 717 655, 1973.
114. V. Tullio et al., *Mycoses* **33** (1990) 257.
115. R. C. Heel et al., *Drugs* **16** (1978) 177.
116. W. Rindt et al., *Arzneim. Forsch. / Drug Res.* **29** (1979) 697.
117. A. Vukovich et al., *Clin. Pharmacol. Ther.* **21** (1977) 121.
118. L. E. Millikan et al., *J. Am. Acad. Dermatol.* **18** (1988) 52.
119. G. Gabriel et al., *Br. J. Vener. Dis.* **59** (1983) 56.
120. J. S. Bingham et al., *Br. J. Vener. Dis.* **57** (1981) 204.
121. J. E. Benett in L. S. Goodman et al. (eds.): *The Pharmacological Basis of Therapeutics,* Pergamon Press, New York 1990 p. 1165.
122. A. Tajana et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 2127.
123. Recordati, DE 2 917 244, 1979.
124. Recordati, US 4 221 803, 1980.
125. A. Finzi et al., *Mykosen* **29** (1986) 41.
126. E. M. Kokoschka et al., *Mykosen* **29** (1986) 45.
127. A. Gastaldi, *Curr. Ther. Res.* **38** (1985) 489.
128. E. F. Godefroi et al., *J. Med. Chem.* **12** (1969) 784.
129. Janssen, DE 1 940 388, 1970.
130. Janssen, US 3 717 655, 1973.
131. Janssen, US 3 839 574, 1974.
132. H. J. Kessler, *Arzneim. Forsch. / Drug Res.* **29** (1979) 1344.
133. H. J. Kessler et al., *Arzneim. Forsch. / Drug Res.* **29** (1979) 1352.
134. U. Täubler, *Arzneim. Forsch. / Drug Res.* **37** (1987) 461.
135. H. Wendt et al., *Arzneim. Forsch. / Drug Res.* **29** (1979) 846.
136. Siegfried, DE 2 839 388, 1980.
137. Siegfried, US 4 210 657, 1980.
138. Siegfried, US 4 554 356, 1985.
139. M. Mosse et al., *Pathol. Biol.* **34** (1986) 684.
140. G. Mixich et al., *Arzneim. Forsch. / Drug Res.* **29** (1979) 1510.
141. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 8,** Springer, Berlin – Heidelberg – New York 1993, p. 1262.
142. Siegfried, DE 2 657 578, 1977.
143. Siegfried, US 4 124 767, 1978.
144. A. Polak, *Arzneim. Forsch. / Drug Res.* **32** (1982) 17.
145. G. Stüttgen et al., *Mykosen* **28** (1985) 138.
146. W. H. Beggs, *IRCS Med. Sci.* **11** (1983) 677.
147. M. Raga et al., *J. Med. Chem. Chim. Ther.* **21** (1986) 329.
148. Ferrer, EP 151 477, 1985.
149. Ferrer, US 5 135 943, 1992.
150. E. Martin-Mezuelos et al., *Chemotherapy* **42** (1996) 112.
151. E. Drouhet et al., *Arzneim. Forsch. / Drug Res.* **42** (1992) 705.
152. C. Palacin et al., *Arzneim. Forsch. / Drug Res.* **42** (1992) 699.
153. C. Palacin et al., *Arzneim. Forsch. / Drug Res.* **42** (1992) 714.
154. O. Azcona et al., *Curr. Ther. Res.* **49** (1991) 1046.
155. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 9,** Springer, Berlin – Heidelberg – New York 1993, p. 691.
156. Syntex, DE 2 541 833, 1976.
157. Syntex, FR 2 285 126, 1976.
158. Syntex, US 4 038 409, US 4 039 677, 1977.
159. P. Benfield et al., *Drugs* **35** (1988) 143.
160. A. Yoshida et al., *Chemotherapy* **32** (1984) 477.
161. F. Iwata et al., *Shinkin to Shinkinsho* **25** (1984) 147.
162. J. Heeres et al., *J. Med. Chem.* **26** (1983) 611.
163. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 9,** Springer, Berlin – Heidelberg – New York 1993, p. 808.
164. Janssen, DE 2 804 096, 1978.
165. Janssen, US 4 144 346, 1979.
166. Janssen, US 4 223 036, 1980.
167. R. A. Fromtling, *Clin. Microbiol. Rev.* **1** (1988) 187.
168. A. Kjaeldgaard, *Pharmacotherapeutica* **4** (1986) 525.
169. U. M. Moebius, *Lancet* **332** (1988) 966.
170. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 9,** Springer, Berlin – Heidelberg – New York 1993, p. 944.
171. Pfizer, BE 841 309, 1976.
172. Pfizer, US 4 062 966, 1977.
173. S. P. Clissold et al., *Drugs* **31** (1986) 29.
174. S. Jevons, *Antimicrob. Agents Chemother.* **15** (1979) 597.
175. F. C. Odds, *J. Antimicrob. Chemother.* **6** (1980) 749.
176. Y. M. Clayton et al., *Clin. Exp. Dermatol.* **7** (1982) 543.

177. D. Brunelli et al., *Contact Dermatitis* **27** (1992) 120.
178. R. Izu et al., *Contact Dermatitis* **26** (1992) 130.
179. R. W. Holz in F. E. Hahn (ed.): *Antibiotics,* **vol. 5,** Springer, Berlin – Heidelberg – New York 1979, p. 313.
180. J. Vandeputte et al., *Antibiot. Anm.* (1955/56), 587.
181. J. M. T. Hamilton-Miller, *J. Pharm. Pharmacol.* **25** (1973) 401.
182. J. F. Gallelli, *Drug Intell.* **1** (1967) 102.
183. E. Borowski et al.: *Vth Intl. Congr. Biochem. Moscow 1961,* Pergamon Press, London 1961, p. 3.
184. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 7,** Springer, Berlin – Heidelberg – New York 1993, p. 237.
185. C. P. Schaffner et al., *Antibiot. Chemother.* **11** (1961) 724.
186. E. Borowski et al., *Tetrahedron Lett.* (1970) 3909.
187. K. C. Nicolaou et al., *Chem. Commun.* (1986) 413.
188. O. Mathieson, US 2 908 611, 1959.
189. R. Janknegt, *Clin. Pharmacokinet.* **23** (1992) 279.
190. H. Lackner et al., *Pediatrics* **89** (1992) 1259.
191. A. Zoubek et al., *Pediat. Hematol. Oncol.* **9** (1992) 187.
192. E. D. Ralph et al., *Antimicrob. Agents Chemother.* **35** (1991) 188.
193. C. E. Hughes et al., *Antimicrob. Agents Chemother.* **25** (1984) 560.
194. T. K. Daneshmend et al., *Clin. Pharmacokin.* **8** (1983) 17.
195. J. Brajtburg et al., *J. Infect. Dis.* **149** (1984) 986.
196. K. J. Christiansen et al., *J. Infect. Dis.* **152** (1985) 1037.
197. N. Collette et al., *Antimicrob. Agents Chemother.* **33** (1989) 362.
198. A. J. Atkinson et al., *Antimicrob. Agents Chemother.* **13** (1978) 271.
199. A. Kucers et al. in A. Kucers et al. (eds.): *The Use of Antibiotics,* J. B. Lippincott, Philadelphia 1987, p. 1441.
200. D. J. Drutz, *Drugs* **26** (1983) 337.
201. G. Medoff et al., *Proc. Natl. Acad. Sci. USA* **69** (1972) 196.
202. J. E. Bennett et al., *N. Engl. J. Med.* **301** (1979) 126.
203. H. A. Gallis et al., *Rev. Infect. Dis.* **12** (1990) 308.
204. F. Gigliotti et al., *J. Infect. Dis.* **156** (1987) 784.
205. M. S. Maddux et al., *Drug Intell. Clin. Pharm.* **14** (1980) 177.
206. W. T. Butler et al., *Ann. Intern. Med.* **61** (1964) 175.
207. G. Lopez-Berenstein et al., *J. Infect. Dis.* **151** (1985) 704.
208. A. P. Struyk et al., *Antibiot. Ann.* (1957/58) 878.
209. W. E. Meyer, *Chem. Commun.* (1968) 470.
210. T. Strittmatter, *Chem. Abstr.* **94** (1981) 214469.
211. Koninklijke Nederlandsche Gist en Spiritus-Fabriek, GB 844 289, 1960.
212. American Cyanamid, GB 846 933, 1960.
213. E. Skytte Christensen et al., *Acta Obstst. Gynecol. Scand.* **61** (1982) 325.
214. D. Kerridge, *Ad. Microb. Physiol.* **27** (1986) 1.
215. T. Wegemann, *Internist* **30** (1989) 46.
216. M. I. Maurad, *Assiut. Vet. Med.* **11** (1984) 236.
217. D. J. Sutton et al., *Vet. Rec.* **118** (1986) 27.
218. E. L. Hazen et al., *Proc. Soc. Exp. Biol. Med.* **76** (1951) 93.
219. C. N. Chong et al., *Tetrahedron Lett.* (1970) 5145.
220. A. C. Moffat: *Ckarke's Isolation and Identification of Drugs,* Pharmaceutical Press, London 1986.
221. J. R. Carlson et al., *Antibiot. Chemother.* **9** (1959) 139.
222. O. Mathieson, US 2 832 719, 1958.
223. American Canamid, US 3 517 100, 1970.
224. R. Duschinsky et al., *J. Am. Chem. Soc.* **79** (1957) 4559.
225. Hoffmann-La Roche, US 2 945 038, 1960 (R. Duschinsky et al.).
226. Hoffmann-La Roche, US 2 802 005, 1957.
227. Hoffmann-La Roche, US 2 945 038, 1960.
228. Hoffmann-La Roche, US 3 040 026,1962.
229. Hoffmann-La Roche, BE 628 615, 1963.
230. Hoffmann-La Roche, US 3 368 938, 1968.
231. D. C. E. Speller (ed.): *Antifungal Chemotherapy,* J. Wiley & Sons, New York 1980.
232. H. Otten et al.: *Antibiotika-Fibel,* Thieme, Stuttgart 1975, p. 666.
233. A. Polak et al., *Eur. J. Biochem.* **32** (1973) 276.
234. J. Schönbeck et al., *Sabouraudia* **11** (1973) 10.
235. R. E. Cutler et al., *Clin. Pharmacol. Ther.* **24** (1978) 333.
236. A. Polak, *Postgrad. Med. J.* **55** (1979) 667.

237. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 8,** Springer, Berlin – Heidelberg – New York 1993, p. 226.
238. T. K. Daneshmend et al., *Clin. Pharmacokin.* **8** (1983) 17.
239. R. B. Diasio et al., *Antimicrob. Agents Chemother.* **14** (1978) 903.
240. B. E. Harris et al., *Antimicrob. Agents Chemother.* **29** (1986) 44.
241. K. F. Wagner in G. L. Mandell et al. (eds.): *Principles and Practice of Infectious Diseases,* Churchill Livingstone, New York 1990, p. 1975
242. D. Amstrong in J. F. Ryley (ed.): *Chemotherapy of Fungal Diseases,* Springer, Berlin – Heidelberg – New York 1990, p. 439.
243. A. Kucers et al. in A. Kucers et al. (eds.): *The Use of Antibiotics,* J. B. Lippincott, Philadelphia 1987, p. 1534.
244. A. Brossi et al., *Helv. Chim. Acta.* **43** (1960) 1444.
245. Glaxo, US 3 069 328, US 3 069 329, 1962.
246. P. W. Brian, *Trans. Br. Mycol. Soc.* **43** (1960) 1.
247. W. Dittmar, *Mykosen* **9** (1966) 104.
248. R. Brehme et al., *Pharmazie* **34** (1979) 372.
249. T. Nogichi et al., *Antimicrob. Agents Chemother.* **1962** (1961 – 70) 259.
250. W. Adam et al., *Ärztl. Forsch.* **14** (1960) 144.
251. M. Rowland, *J. Pharm. Sci.* **57** (1968) 984.
252. M. Kraml, *Antibiot. Chemother.* **12** (1962) 239.
253. V. Shah, *J. Pharm. Sci.* **61** (1972) 634.
254. C. Lin et al., *Drug Metabol. Rev.* **4** (1975) 75.
255. B. R. Lecky, *Lancet* **336** (1990) 230.
256. G. Mion et al., *Lancet* **334** (1989) 1331.
257. G. Lohhaus et al., *Chem. Ber.* **100** (1967) 658.
258. Hoechst, DE-AS 2 214 608, 1973 (G. Lohhaus et al.).
259. Hoechst, ZA 6 906 039, 1970.
260. Hoechst, DE 2 214 608, 1973.
261. Hoechst, US 3 883 545, 1975.
262. Hoechst, DE 2 795 831, 1978.
263. W. Dittmar, *Proc. III. Intern. Congr. Trop. Der.,* Sao Paulo 1974.
264. K. Iwata et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 1323.
265. K. Sakurai et al., *Chemotherapy* **24** (1978) 68.
266. K. Iwata et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 1323.
267. W. Dittmar et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 1317.
268. H. M. Kellner et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 1337.
269. S. G. Jue et al., *Drugs* **29** (1985) 330.
270. H. G. Alpermann et al., *Arzneim. Forsch. / Drug Res.* **31** (1981) 1328.
271. H. G. Peil, *Arzneim. Forsch. / Drug Res.* **31** (1981) 1366.
272. N. S. Ryder et al., *Antimicrob. Agents Chemother.* **29** (1986) 858.
273. I. De Carneri et al., *Arzneim. Forsch. / Drug Res.* **26** (1976) 769.
274. Carlo Erba, DE 2 313 845, 1973.
275. Carlo Erba, US 3 855 263, 1974.
276. Nippon Soda K.K. Tokio, DE-AS 1 468 388, 1962 (M. K. Taoka et al.).
277. R. Brehme et al., *Pharmazie* **34** (1979) 372.
278. T. Nogichi et al., *Antimicrob. Agents Chemother.* (1962) 259.
279. Japan Soda, FR 1 337 797, 1963.
280. Japan Soda, BE 627 322, 1962.
281. Japan Soda, US 3 334 126, 1967.
282. S. C. Harvey in L. S. Goodman et al. (eds.): *The Pharmacological Basis of Therapeutics,* Macmillan Publ., New York 1985 p. 959.
283. A. H. Gould, *South. Med. J.* **59** (1966) 176.
284. G. Petranyi et al., *Science* **224** (1984) 1239.
285. A. Schütz et al., *J. Med. Chem.* **27** (1984) 1539.
286. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 8,** Springer, Berlin – Heidelberg – New York 1993, p. 1068.
287. H. Loibner et al., *Tetrahedron Lett.* **25** (1984) 2535.
288. Sandoz, BE 853 976, 1977.
289. Sandoz, US 4 282 251, 1981.
290. Lab. Frumtost-Prem S.A., ES 504 432, 1982.
291. A. Georgopoulos et al., *Antimicrob. Agents Chemother.* **19** (1981) 386.
292. G. Petranyi et al., *Antimicrob. Agents Chemother.* **19** (1981) 390.
293. J. P. Monk et al., *Drugs* **42** (1991) 659.
294. S. Nolting et al., *Mykosen* **28** (1985) 69.
295. L. E. Millikan et al., *J. Am. Acad. Dermatol.* **18** (1988) 52.
296. I. Effendy et al., *Therapiewoche* **36** (1986) 848.
297. F. von Bruchhausen et al. (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* **vol. 9,** Springer, Berlin – Heidelberg – New York 1993, p. 802.
298. Sandoz, EP 24 587, 1981.
299. Y. M. Clayton et al., *Br. J. Dermatol.* **130** (1994) Suppl. 43, 7.
300. J. C. Jensen, *Clin. Exp. Dermatol.* **14** (1989) 110.
301. J. A. Balfour et al., *Drugs* **43** (1992) 259.

302. P. J. Dykes et al., *J. Dermatol. Treat.* **1** (1990) Suppl. 2, 19.
303. T. C. Jones, *Br. J. Dermatol.* **132** (1995) 683.
304. L. Juhlin, *Lancet* **339** (1992) 1483.
305. J. P. Ottervanger et al., *Lancet* **340** (1992) 728.
306. G. Lowe et al., *B.M.J.* **306** (1993) 248.
307. Hoffmann-La Roche, DE 2 752 096, 1978.
308. Hoffmann-La Roche, US 4 202 894, 1980.
309. A. Polak in J. F. Ryley (ed.): *Chemotherapy of Fungal Diseases,* Springer, Berlin – Heidelberg – New York 1990, p. 153.
310. A. Polak in J. F. Ryley (ed.): *Chemotherapy of Fungal Diseases,* Springer, Berlin – Heidelberg – New York 1990, p. 505.
311. E. Rhode et al. in R. A. Fromtling (ed.): *Recent Trends in the Discovery, Development and Evaluation of Antifungal Agents,* J. R. Prous, Barcelona 1987, p. 575.
312. M. Haria et al., *Drugs* **49** (1995) 103.

Antioxidants

PETER P. KLEMCHUK, CIBA-GEIGY Corporation, Ardsley, New York 10502, United States

1. Introduction 1
2. Oxidation 2
3. Reaction Mechanisms 3
3.1. Autooxidation Mechanisms 3
3.2. Antioxidant Mechanisms 4
3.2.1. Radical Trapping 4
3.2.2. Peroxide Decomposing 5
3.3. Metal Deactivators 5
4. Antioxidant Classes 6
4.1. Hindered Phenols 6
4.2. Aromatic Amines 7
4.3. Organosulfur Compounds 7
4.4. Phosphorus Compounds 11
4.5. Hindered Amines 12
4.6. Metal Deactivators 13
5. Practices in Stabilization with Antioxidants 15
5.1. Stabilization of Food 15
5.2. Stabilization of Fuels 16
5.3. Stabilization of Lubricants 16
5.4. Stabilization of Polymers 16
5.4.1. Elastomers 17
5.4.2. Thermoplastics 17
5.4.3. Synthetic Fibers 18
5.4.4. Economic Aspects 18
6. Food Safety Considerations 18
7. Test Methods to Evaluate Polymer Thermal Oxidative Stability 19
7.1. Melt Processing 19
7.2. Oven Aging 19
7.3. Oxygen Uptake 19
7.4. Thermal Analytical Methods 20
7.5. Tests to Monitor Polymer Properties 20
8. References 21

1. Introduction

Organic matter has a strong tendency to react with oxygen and oxidize. This is true for most commercial organic materials, e.g., plastics, elastomers, fibers, fuels, lubricants, and foods. It is also true for humans, if aging is considered to be, at least in part, an oxidation process.

In 1861, A. W. VON HOFMANN first made the connection between oxygen and the deterioration of polymer properties in his studies on the aging of gutta-percha [1]. The concept of antioxidants and the discovery that small amounts of reducing agents could protect materials susceptible to oxidation were attributed to S. L. BIGELOW (1898) for sodium sulfite solution [2]; to P. SISLEY (1903 and 1904) for the stabilization of dyed silk to oxidation by hydroquinone (among other compounds); to A. LUMIÈRE, L. LUMIÈRE, and A. SEYEWETZ (1905) for the protection of photographic developers (these authors first used the term antioxidant); and to CH. MOUREU and CH. DUFRAISSE (1920) for the protection of acrolein.

Virtually all polymers that have become commercial products since these early years are dependent on stabilizing additives, antioxidants, which permit the polymers to withstand the oxidative stresses of processing and fabrication and to be used in so many different ways.

Oxidation of organic materials takes place by a number of processes:

autooxidation
biooxidation
combustion
photooxidation

The ultimate result of these processes is the conversion of carbon and hydrogen in organic compounds to carbon dioxide and water, with the accompanying release of large amounts of energy for each mole of carbon dioxide and water formed. In the autooxidation of commercial organic materials, oxidation to carbon dioxide and water is seldom complete.

This article primarily describes the *autooxidation* of organic materials and its prevention by antioxidants, with emphasis on the autooxidation of commercially produced polymers. *Biooxidation* and *combustion* are important processes but are not covered here. *Photooxidation* is discussed only briefly. Although conventional antioxidants, such as hindered phenols, thio-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_091

synergists, and phosphites, also are moderate stabilizers of organic materials against photo-oxidation, other, more effective additives, especially ultraviolet absorbers, hindered amines, and nickel chelates, are required to give adequate light stability.

Antioxidants, also called inhibitors (of oxidation), are organic compounds that are added to oxidizable organic materials to retard auto-oxidation and, in general, to prolong the useful life of the substrates. Relatively few chemical classes are effective as antioxidants. Those in common use are hindered phenols, secondary aromatic amines, certain sulfide esters, trivalent phosphorus compounds, hindered amines, metal dithiocarbamates, and metal dithiophosphates.

Another class of stabilizers, the metal deactivators, reduces markedly the rate of oxidation of polyolefin insulation on copper conductors.

Some antioxidants are capable of functioning alone as *primary stabilizers.* These are the hindered phenols, aromatic amines, metal dithiocarbamates and phosphates, and hindered amines. The phosphites and sulfide esters are *secondary stabilizers* and are used most commonly in conjunction with primary stabilizers. Stabilizer practices are discussed Chapter 5.

Many of the secondary aromatic amines derived from *p*-phenylenediamine are active as both antioxidants and antiozonants. *Antiozonants* are added to unsaturated elastomers to reduce the rate of ozone cracking that arises from the reaction of ozone with carbon – carbon double bonds in the polymer. This reaction ultimately leads to chain cleavage and to cracking of the elastomer when sufficient numbers of bonds are broken.

2. Oxidation

Oxygen is a key factor in oxidation. In the ground state, it exists as a biradical (**1**) and reacts readily with carbon free radicals, even at ambient temperatures, to yield *peroxy radicals*

$$\underset{\mathbf{1}}{\mathrm{RCH_2\cdot + \cdot O-O\cdot}} \longrightarrow \mathrm{RCH_2OO\cdot}$$

The reaction of carbon free radicals with oxygen has such a low-activation energy that it is the preferred pathway when not limited by oxygen diffusion. The resulting peroxy radicals abstract hydrogen from the substrate, generating another carbon free radical and *a hydroperoxide*:

$$\mathrm{RCH_2OO\cdot + RH} \longrightarrow \mathrm{CH_2OOH + R\cdot}$$

This reaction has severe consequences for the organic material because: (1) it is a chain reaction that has a significant length and can cause extensive localized oxidation before termination; (2) the hydroperoxide can yield *radical initiators* of oxidation by unimolecular and bimolecular processes:

$$\mathrm{RCH_2OOH} \longrightarrow \mathrm{RCH_2O\cdot + HO\cdot}$$

$$\mathrm{2\,RCH_2OOH} \longrightarrow \mathrm{RCH_2OO\cdot + RCH_2O\cdot + H_2O}$$

Unless these radical reactions are inhibited with antioxidants, the organic substance can quickly deteriorate. The physical properties and appearance of polymers and elastomers are reduced, fuels and lubricants sour, and foods lose their flavor and become rancid.

Ions of certain *metals,* e.g., copper, iron, and cobalt, which can exist in two oxidation states, act as catalysts to decompose hydroperoxides:

$$\begin{array}{l} \mathrm{M^{n+} + ROOH} \longrightarrow \mathrm{M^{(n+1)+} + RO\cdot + OH^-} \\ \mathrm{M^{(n+1)+} + ROOH} \longrightarrow \mathrm{M^{n+} + ROO\cdot + H^+} \\ \hline \mathrm{2\,ROOH} \longrightarrow \mathrm{RO\cdot + ROO\cdot + H_2O} \end{array}$$

Hydroperoxides are also susceptible to *photolytic cleavage* by near ultraviolet light (< 400 nm but particularly ca. 300 nm):

$$\mathrm{RCH_2OOH} \xrightarrow{h\nu} \mathrm{RCH_2O\cdot + HO\cdot}$$

Temperature plays a key role in the rate of oxidation of organic materials. Most materials are subjected, however briefly during some stage, to relatively high temperatures; in these intervals significant oxidation can occur and significant amounts of oxidation initiators can be formed. Commercial polymers (thermoplastics, elastomers, synthetic fibers) are dried, processed, and fabricated at elevated temperatures; lubricants are employed at elevated temperatures; foods are cooked at high temperatures. Therefore, processed and fabricated materials will contain significant, but varying, amounts of hydroperoxide and peroxide functionalities, depending on the degree to which the material is stabilized with antioxidants and on the degree of thermo-oxidative exposure.

With plastics and synthetic fibers, compounding to incorporate stabilizer, filler, lubricant, pigment, etc., requires several minutes exposure at 175 – 300 °C. Subsequent fabrication into final articles usually requires < 10 min at temperatures that vary according to the fabrication process:

Fabrication process	Temp., °C
Injection molding	175 – 340
Film blowing	175 – 260
Film extrusion	175 – 260
Blow molding	175 – 230
Fiber spinning	230 – 350
Rotomolding	280 – 340

Processing at high temperature can have a significant impact on a polymer. For example, polypropylene was extruded a total of five times at 260 °C in a 2.54 cm laboratory extruder. Table 1 shows the melt flow rates for pellets retained from the first, third, and fifth extrusions. The large increase in melt flow rate of the unstabilized sample reflects considerable polymer chain cleavage during processing. Melt flow rate determinations and viscosity measurements are convenient and sensitive methods for following changes in average molecular masses of polymers.

Table 1. Multiple extrusion of polypropylene at 260 °C

Antioxidant (AO) stabilizer	Melt flow rate (g/10 min) after extrusion number:			
	0	1	3	5
None	4∗	8	14	23
0.1 % phenolic AO	–	6	7	9

∗ With 0.5 % phenolic AO

The influence of processing on the thermal stability of a polymer can be seen in Table 2. polypropylene samples (0.6 mm plaques), after one, three, and five extrusions at 260 °C, were oven aged at 150 °C; the deleterious effect of additional processing on thermal stability is clearly evident.

Table 2. Influence of processing on thermal stability of polypropylene

Antioxidant (AO) stabilizer	Hours to failure after extrusion number:		
	1	3	5
None	<20	<20	<20
0.1 % phenolic AO	510	400	330

The *ease of oxidation* of organic materials is related to the tendency of the substrate to participate in free-radical reactions with oxygen. Substrates that contain carbon – carbon unsaturation with allylic hydrogen atoms are particularly prone to oxidation. For this reason, unsaturated polymers, notably unvulcanized elastomers, and foods with unsaturated fats and oils oxidize easily. The impact of unsaturation is shown by the following comparative polymer data (0.1 % phenolic AO, 100 °C, 0.6 mm plaques)

Polymer	Hours to failure	
polypropylene	10000 [a]	↓
Impact polystyrene	500 [b]	Increasing degree of unsaturation
Styrene – butadiene rubber	69 [c]	
Polybutadiene rubber	15 [c]	↓

[a] Disintegration.
[b] Color formation.
[c] Gel formation.

3. Reaction Mechanisms

3.1. Autooxidation Mechanisms

The work on autooxidation carried out at the Natural Rubber Producers' Research Association in England in the 1940s and early 1950s was a major contribution to understanding autooxidation and stabilization reactions. The work of Bateman [3], Bolland [4], Gee, and Ten Have, among others, showed the oxidation of hydrocarbon materials to:

1) be free radical in nature,
2) have a period of inhibition,
3) have the properties of an autocatalytic chain reaction,
4) yield hydroperoxides as primary oxidation products.

Autooxidations have three distinct steps:

Initiation:

$$\mathrm{ROOH} \xrightarrow{k_1} \mathrm{RO}\cdot + \mathrm{HO}\cdot$$

Propagation:

$$\mathrm{ROO\cdot(RO\cdot, HO\cdot) + RH \xrightarrow{k_2} ROOH + R\cdot} \quad (1)$$
$$\mathrm{R\cdot + O_2 \xrightarrow{k_3} ROO\cdot}$$

Termination:

$$\mathrm{2\,ROO\cdot \xrightarrow{k_4} [ROOOOR] \longrightarrow ROOR + O_2} \quad (2)$$
$$\mathrm{2\,R\cdot \xrightarrow{k_5} RR}$$
$$\mathrm{R\cdot + ROO\cdot \xrightarrow{k_6} ROOR}$$

In most instances, when the oxygen concentration is not limiting, termination involves mainly peroxy radicals.

In his work on the oxidation of ethyl linoleate, a model for unsaturation in elastomers and, incidentally, also for fatty foods, BOLLAND [5] identified hydroperoxides as the main products of oxidation. He also showed ethyl linoleate hydroperoxide to be an initiator of ethyl linoleate oxidation. The rate expression in this instance is:

$$\frac{\mathrm{d}c_{\mathrm{ROOH}}}{\mathrm{d}t} = k_2 c_{\mathrm{RH}} \sqrt{\frac{k_1 c_{\mathrm{ROOH}}}{k_4}}$$

The termination of secondary peroxy radicals proceeds through a tetroxide intermediate to yield *sec*-alcohols and ketones [6]:

$$\mathrm{2\,R_2CHOO\cdot \longrightarrow R_2CHOH + R_2C{=}O + O_2}$$

Tertiary peroxy radicals, on the other hand, do not terminate directly. In this case the metastable tetroxide intermediate decomposes to yield alkoxy radicals and oxygen [7–9]:

$$\mathrm{2\,ROO\cdot \longrightarrow [ROOOOR] \longrightarrow 2\,RO\cdot + O_2} \quad (3)$$

where R contains a *tert*-carbon atom attached to oxygen. With cumylperoxy radicals, about 10 % of termination takes place by Reaction (2) to yield dicumyl peroxide, and 90 % proceeds by Reaction (3) [9]. The cumylalkoxy radical reacts further and eventually terminates as a methylperoxy radical:

$$\mathrm{RO\cdot \longrightarrow CH_3\cdot + C_6H_5COCH_3}$$
$$\mathrm{CH_3\cdot + O_2 \longrightarrow CH_3O_2\cdot}$$
$$\mathrm{CH_3O_2\cdot + RO_2\cdot \longrightarrow CH_2O + ROH + O_2}$$

where R is $C_6H_5C(CH_3)_2$.

Oxidation can start in a hydrocarbon at ambient temperature because most commercial organic materials are subjected to high temperature, e.g., during processing, fuel and lubricant refining, and food preparation. Small but detectable quantities of peroxides form and serve as initiators of autooxidation. The main product of autooxidation, hydroperoxide, is an initiator of autooxidation. This is the reason autooxidation is autocatalytic.

The plot of oxygen uptake (Fig. 1) for high-density polyethylene (HDPE) at 100 °C exhibits the autocatalytic nature of autooxidation. During an induction period of about 35 h, little oxygen is absorbed. Once initiator concentration is sufficiently high and oxidation inhibitors, if any, are consumed, however, oxygen consumption increases rapidly to a steady rate. This behavior is characteristic of organic materials undergoing autooxidation.

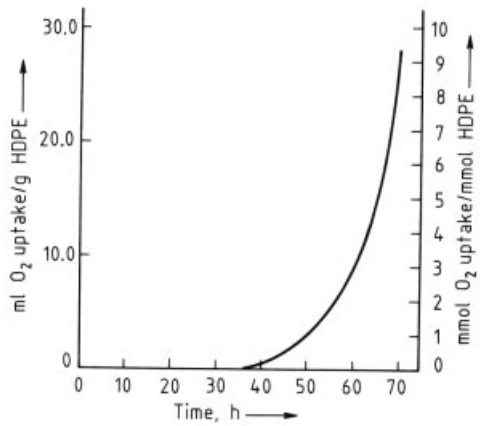

Figure 1. Oxygen uptake curve of Du Pont Alathon 5496 HDPE at 100 °C and 101.3 kPa

3.2. Antioxidant Mechanisms

Antioxidants are classified as either radical trapping (chain breaking) or peroxide decomposing, terms that describe the mechanisms by which they function.

3.2.1. Radical Trapping

Radical trapping antioxidants (AH) function by reaction with (trapping) the propagating *oxygen free radicals* in autooxidation:

$$ROO\cdot + AH \longrightarrow ROOH + A\cdot \quad (4)$$

The antioxidant thus competes with the organic substrate (RH) as a source of hydrogen for peroxy radicals in the usual autooxidation sequence (Eq. 1). For effective radical trapping antioxidants, Reaction (4) is much faster than Reaction (1). In addition, the radical A · must be stable and not abstract hydrogen atoms from the organic substrate; i.e., it should not function as a chain-transfer agent:

$$A\cdot + RH \not\longrightarrow AH + R\cdot \quad (5)$$

Otherwise, the antioxidant does not become a chain terminator but only transfers the propagation of oxidation from one molecule to another.

Radical trapping antioxidants, e.g., hindered phenols and secondary aromatic amines in particular, react with oxygen radicals (peroxy and alkoxy) almost exclusively. That they usually do not trap carbon free radicals efficiently is a distinct disadvantage. Trapping a *carbon free radical* (Eq. 6) is the preferred radical trapping reaction because it reduces to zero the oxidative chain length (the average number of oxygen molecules consumed in a chain reaction), and markedly reduces hydroperoxide formation. The usual radical-trapping reaction is represented by Equation (4).

$$R\cdot + AH \longrightarrow RH + A\cdot \quad (6)$$

A new class of radical-trapping stabilizers, the hindered amines trap alkyl radicals (see Section 4.5).

Even though they do not reduce the oxidative chain length to zero, as would be preferred, radical-trapping antioxidants are very effective in reducing the oxidative chain length. For example, 1,4-benzohydroquinone at 0.03 % (mole fraction of hydroquinone to ester 9×10^{-4}) reduced the oxidative chain length for the benzoylperoxide-initiated, uninhibited oxidation of ethyl linoleate from 241 to one [10]. This means that in the uninhibited reaction, 241 molecules of oxygen were consumed and 241 molecules of hydroperoxide were formed in the chain reaction following the generation of one initiating radical. With the antioxidant present, only one molecule of oxygen was consumed and one molecule of hydroperoxide was formed per initiating radical. Both the uninhibited and the inhibited oxidative chain lengths vary with the substrate; however, stabilization by antioxidants always results in significant lowering of chain lengths.

3.2.2. Peroxide Decomposing

Peroxide decomposing antioxidants reduce the hydroperoxidic and peroxidic products of oxidation to more innocuous compounds, usually alcohols or ethers. Because peroxides are initiators of oxidation, their removal from organic materials is important in the inhibition of oxidation. The two major classes of peroxide decomposing stabilizers are trivalent phosphorus compounds and thiosynergists (see Sections 4.3 and 4.4).

Trivalent phosphorus compounds– phosphites and phosphonites – function as peroxide decomposers by abstracting peroxidic oxygen from hydroperoxides and peroxides, and reducing them. The phosphorus compounds are oxidized to phosphates or phosphonates:

$$(RO)_3P + R'OOH \longrightarrow (RO)_3PO + R'OH$$

Thiosynergists, mainly esters of thiodipropionic acid, function mainly at elevated temperatures ($>100\,°C$) both by direct reaction with hydroperoxides and by the nonradical decomposition of hydroperoxides catalyzed by the acidic oxidation products of the thiosynergists. This latter reaction is similar to the acid-catalyzed decompositions of cumene hydroperoxide used to manufacture phenol and acetone (→ Acetone).

3.3. Metal Deactivators

Metal deactivators are effective in reducing the particularly deleterious effects that copper in wires and cables has on polyolefin insulation. Copper is a particularly severe agent for degrading polymers by thermal oxidation. Metal deactivators are postulated to function by deactivating copper to oxidation by complexation at the polymer – metal interface (see Section 4.6).

4. Antioxidant Classes

4.1. Hindered Phenols

Hindered phenols are a major class of antioxidants. They are presently used as radical-trapping agents in plastics, elastomers, synthetic fibers, fuels, lubricants, and foods. Steric hindrance is provided by bulky substituents in positions on the ring ortho to the hydroxyl group. These bulky substituents influence the specificity of the phenols by blocking phenoxyl radicals from abstracting hydrogen atoms from organic substrates.

2,6-Di-tert-butylphenol [*128-39-2*] is the most widely used functionality in commercial hindered phenolic antioxidants. This group is effective and easily made at moderate cost by the ortho alkylation of phenol by olefins using an aluminum alkoxide alkylation catalyst [11], (→ Acylation and Alkylation, Chap. 2.1.4.):

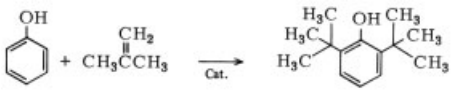

Structures of common phenolic antioxidants are included in Table 3. A more comprehensive list of phenolic and other antioxidants can be found in [12].

Substituents in the 4-position of 2,6-di-*tert*-butylphenol influence the solubility of the antioxidant. Moreover, high-molecular-mass substituents influence the migration and extractability of antioxidants, reducing their extraction from food-packaging plastics, and making such antioxidants suitable for use as indirect food additives. The 4-substituent can also influence the reactivity of the hindered phenol, making some compounds more suitable for certain substrates, such as unsaturated elastomers.

The mechanism of the reaction of hindered phenols with alkylperoxy radicals, as a model for chain propagating radicals, was investigated. In oxidizing organic substrates [13], [14]. In solution each hindered phenolic moiety consumed two peroxy radicals:

$$\mathrm{ROO}\cdot + \mathrm{AH} \longrightarrow \mathrm{ROOH} + \mathrm{A}\cdot \qquad (7)$$
$$\mathrm{ROO}\cdot + \mathrm{A}\cdot \longrightarrow \mathrm{ROOA}$$

4-Alkylperoxycyclohexadienones were isolated:

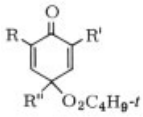

where R, R′, R″ are CH_3 or t-C_4H_9.

Compounds with 4-methyl substituents also yielded *stilbenequinone* derivatives (**15**), arising from the following reaction sequence:

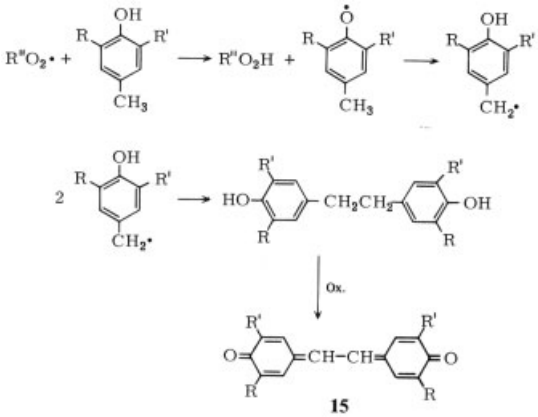

Stilbenequinones have very high-extinction coefficients in the visible portion of the spectrum, and can contribute significant color to the materials in which they are incorporated. Stilbenequinones are serious problems, particularly in plastics stabilization because amounts as small as 5×10^{-6} parts are sufficient to impart a yellow color to the plastic, which is objectionable in colorless or white plastic articles.

The antioxidant efficiency of phenolic antioxidants as well as the degree of chain transfer have been measured [14]. The *antioxidant efficiency* is defined as $2k_a/k_r$, where k_a is the rate constant for Equation (4) and k_r is the rate constant for Equation (1). Efficiencies were found to be 32 for 2,4,6-tri-*tert*-butylphenol and 33.5 for 2,6-di*tert*-butyl-4-methylphenol. No chain transfer (Eq. 5) was found with either antioxidant.

The 4-alkylperoxy-4-alkyl-2,6-di-*tert*-butylcyclohexadienones have been formed in the presence of relatively high concentrations of alkylperoxy radicals. Because that situation does not usually exist in inhibited oxidation of organic materials under conditions of use, the normal reaction in plastics is the disproportionation of two phenoxy radicals. This reaction regenerates starting phenol and quinone methide (**16**):

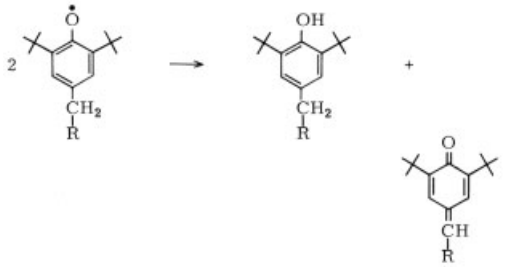

16

Quinone methides have limited stability and react to yield a complex mixture of products [15], [16].

4.2. Aromatic Amines

Secondary aromatic amines for use in elastomer stabilization were among the first antioxidants. The main classes are diarylamines, amine – ketone condensation products (e.g., aniline – acetone condensates), and substituted *p*-phenylenediamines. Most are colored and form highly colored oxidation products. Thus, their uses are generally limited to those for which discoloration is not a drawback, e.g., carbon-black-filled, vulcanized elastomers are a major application.

In spite of their long history of use, the *mechanisms of action* of aromatic amine antioxidants are not as well understood as are those of hindered phenolic antioxidants. To a great extent this is because of the mixture of oxidation products obtained from aromatic amines, and the complex composition of some of the aromatic amine antioxidants. Nevertheless, the general features of the mechanisms are known.

The secondary amines serve as a source of hydrogen atoms, which are abstracted by peroxy radicals (Eq. 4) to yield relatively stable aminyl radicals. The same competitive reactions as found with hindered phenols are the key to the activity of amines as antioxidants.

The *antioxidant efficiencies* of secondary aromatic amines were found to be comparable to those of hindered phenols in the sensitized oxidation of 9,10-dihydroanthracene [17]. The intermediate aminyl radicals had a low tendency to abstract hydrogen atoms from polymers. They were similar in this behavior to hindered phenoxyl radicals. Tertiary amines showed no oxidation retarding properties. *N,N′*-Di-*sec*butyl-*p*-phenylenediamine was particularly efficient in trapping peroxy radicals; it trapped two peroxy radicals per diamine molecule, leading to the formation of quinone imide.

In the diphenylamine-inhibited oxidation of cumene at 68.5 °C, diphenylaminyl radicals were converted in significant amount to nitroxyl radicals; about 20 % of the initial diphenylamine was isolated as diphenyl nitroxide [18].

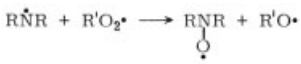

The *N*-oxyl may eventually trap an alkyl radical:

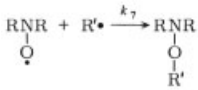

(8)

The ratio of the rate constants, k_8/k_7, is about ten in the oxidation of styrene:

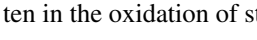

(9)

Reaction (9) therefore is favored in the presence of significant oxygen. However, in the virtual absence of oxygen, Reaction (8) is important, as evidenced by the ability of 4,4-dimethoxydiphenyl nitroxide to function as an inhibitor for styrene polymerization initiated by azobisisobutyronitrile (AIBN) at 60 °C [19].

4.3. Organosulfur Compounds

The usefulness of organosulfur compounds as stabilizers for organic materials was recognized early. Metal salts of dialkyldithiocarbamic acids (→ Dithiocarbamic Acid and Derivatives), zinc dialkyldithiophosphates, and esters (dilauryl, dimyristyl, distearyl) of thiodipropionic acid have become important commercial antioxidants for the stabilization of diverse materials: lubricants, elastomers, and plastics. Several comprehensive reviews have been published on these compounds as antioxidants [20–23].

Mechanisms of Action. Organosulfur antioxidants decompose hydroperoxides by *nonradical reactions.* The effectiveness of various

Table 3. Major commercial hindered phenolic antioxidants

Structure	Chemical name	CAS registry no.	Trade names
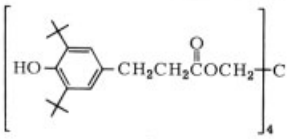 **2**	tetrakis [methylene (3,5-di-*tert*-butyl-4-hydroxyhydrocinnamate)] methane	[*6683-19-8*]	Irganox 1010
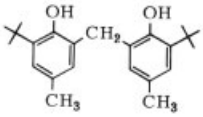 **3**	2,2′-methylenebis-(4-methyl-6-*tert*-butylphenol)	[*119-47-1*]	Cyanox 2246
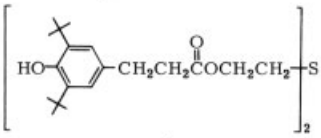 **4**		[*41484-35-9*]	Irganox 1035
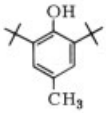 **5**	2,6-di-*tert*-butyl-4-methylphenol	[*128-37-0*]	Butylated hydroxytoluene (BHT)
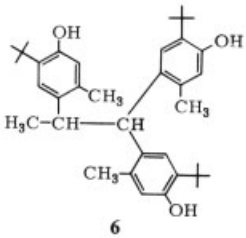 **6**		[*1843-03-4*]	Topanol CA
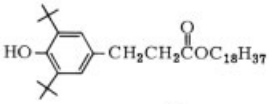 **7**		[*2082-79-3*]	Irganox 1076
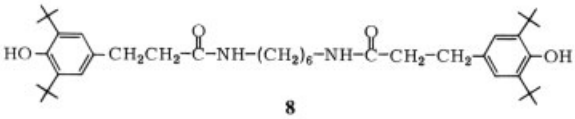 **8**	*N*,*N*′-1,6-hexamethylene-bis-3-(3,5-di-*tert*-butyl-4-hydroxyphenyl) propionamide	[*23128-74-7*]	Irganox 1098

Table 3. Continued.

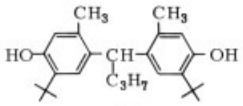 **9**	4,4′-Butylidenebis-(6-*tert*-butyl-3-methylphenol)	[*85-60-9*]	Santowhite powder
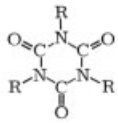 R = 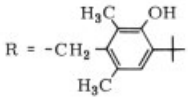 **10**		[*40601-76-1*]	Cyanox 1790
R = 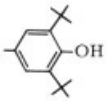 **11**		[*27676-62-6*]	Good-rite 3114
R = 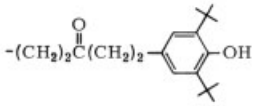 **12**		[*34137-09-2*]	Good-rite 3125
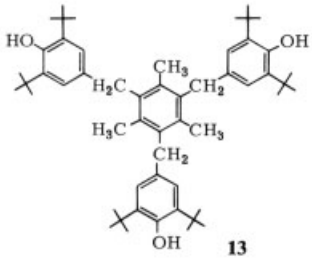 **13**		[*1709-70-2*]	Ethanox 330 Irganox 1330
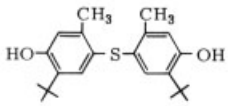 **14**	4,4′-Thiobis(2-*tert*-butyl-5-methylphenol)	[*96-69-5*]	Santonox R

organosulfur compounds, i.e., alkyl- and aryl-sulfides, alkyl- and arylthiols, and a zinc dialkyldithiophosphate in decomposing cumene hydroperoxide was determined in mineral oil at 150 °C [24]. The most effective compounds were zinc di(4-methylpentyl)-2-dithiophosphate, *n*-dodecylthiol, and a sulfonic acid. All reactions gave high yields of phenol, implying that they proceeded by acid catalysis [25].

Oxidation and decomposition products of the organosulfur compounds are responsible for inhibiting the oxidation of organic materials by decomposing the hydroperoxides. Sulfur dioxide, one of these terminal oxidation products, is a highly effective hydroperoxide decomposer [26]. Sulfides and disulfides react with hydroperoxides to yield sulfoxides and thiosulfinates:

$$RSR + R'OOH \longrightarrow RS(O)R + R'OH$$
$$RSSR + R'OOH \longrightarrow RS(O)SR + R'OH$$

Sulfenic acid, RS(O)H, and thiosulfoxylic acid, RSS(O)H, are decomposition products of sulfoxides and thiosulfinates. Each molecule of these acids is capable of decomposing many hydroperoxide molecules [27]. In addition higher oxidation products of sulfenic and sulfoxylic acids, i.e., sulfonic, thiosulfurous, thiosulfuric, and sulfuric acids, and sulfur dioxide are also catalytic hydroperoxide decomposers [27]. A large number of active hydroperoxide decomposers are generated by the oxidation of sulfides and disulfides with hydroperoxides.

Dilauryl thiodipropionate (**17**) [*123-28-4*], (DLTDP) is a common thiosynergist for polyolefin stabilization. Its mechanism of action is similar to relatively simple aromatic and aliphatic sulfide and disulfide model compounds [28]. The DLTDP sulfoxide (**17a**) was thermally unstable at 150 °C and decomposed to a gas (SO_2, H_2S, H_2O), a liquid (containing lauryl acrylate), and a solid (a complex mixture of several compounds, including the sulfenic acid, $H_{25}C_{12}O_2CCH_2CH_2SOH$) [28]. In addition to hydroperoxide decomposition, a second mechanism for DLTDP activity involves the reaction of DLTDP (**17**) with the peroxycyclohexadienone reaction product of a hindered phenol. The hindered phenol was regenerated in 41 % yield [28]:

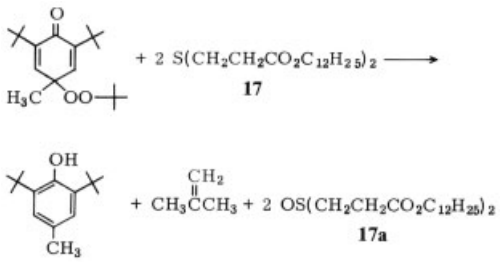

This mechanism can be important in oxidations where significant amounts of peroxycyclohexadienone products are formed from hindered phenols, and in the performance of 2,6-diphenyl-4-octadecyloxyphenol [29].

A third mechanism for thiosynergists comes from the observation that zinc dialkyldithiophosphates, zinc dialkylxanthates, and zinc dialkyldithiocarbamates, inhibit the azobisisobutyronitrile oxidation of squalene at 60 °C [30], [31]. The inhibition is the result of peroxy radical trapping involving electron transfer from sulfur in the zinc salts to the peroxy radicals. Disulfides and zinc peroxides are the products:

$$[(RO)_2P(S)S]_2Zn + 2\,R'OO\cdot \longrightarrow$$
$$(RO)_2P(S)S{-}S(S)P(OR)_2 + Zn(OOR')_2$$

Hydroperoxide decomposition is generally the main mechanism of stabilization by organosulfur compounds. The relative importance of other mechanisms, peroxy radical trapping and hindered phenol regeneration, depends on the stabilizers and circumstances of use.

Thiosynergism. A *synergism of action* exists between radical trapping antioxidants (hindered phenols or secondary aromatic amines) and peroxide-decomposing antioxidants (zinc dialkyldithiophosphate or dialkyl sulfide) in the stabilization of mineral oil at 155 °C and 180 °C [24]. Sulfoxides inhibit oxidation only after the absorption of a small amount of oxygen. Sulfoxides, RS(O)R, and thiosulfinates, RS(O)SR, were especially effective inhibitors for the oxidation of squalene at 75 °C [32].

Esters of thiodipropionic acid play a key role in the high-temperature stabilization of polyolefins by hindered phenols. Figure 2 shows the considerably longer lifetimes obtained with the combination of a hindered phenol and DLTDP [33]. All samples contained 0.4 wt% of stabilizer. Curves a and b depict the lifetimes obtained with DLTDP and hindered phenol, respectively; c is calculated from a and b; and d is a plot of the experimental values from actual compositions of hindered phenol and DLTDP tested at 140 °C. The markedly superior performance shown by the actual mixtures is a result of synergistic interaction between the two stabilizers, the one basically radical trapping, the other hydroperoxide decomposing.

Thioesters are effective primarily when polymers are in the solid state and are at temperatures < 160 °C. This is in contrast to phosphites (see Section 4.4), another class of hydroperoxide decomposers, which are effective in protecting polymers in the molten state during processing and fabrication, usually at temperatures $\gg 160$ °C.

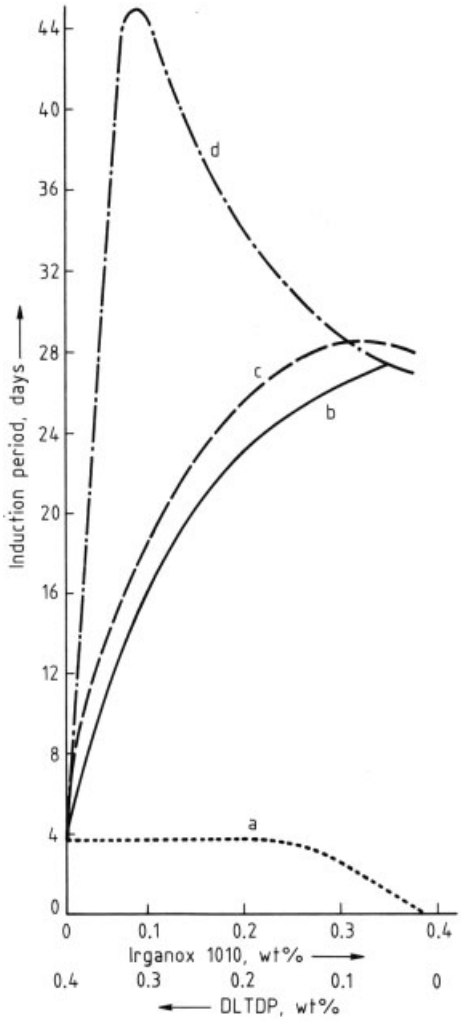

Figure 2. Synergists in heat stabilization
a) DLTDP (dilauryl thiodipropionate); b) Irganox 1010; c) Calculated; d) Experimental

4.4. Phosphorus Compounds

Trivalent phosphorus compounds, i.e., phosphites and phosphonites, are practically the only types of phosphorus compounds being used commercially as antioxidants. The only other major class is zinc dialkyldithiophosphate, which finds use primarily in lubricants. Its mode of action is discussed in Section 4.3.

Mechanisms of Action. Trivalent phosphorus antioxidants generally function by reducing hydroperoxides and peroxides to alcohols and ethers:

$$(RO)_3P + R'OOH \longrightarrow (RO)_3PO + R'OH$$

Neither the phosphate nor the alcohol is an oxidation initiator. The reaction between phosphites or phosphonites and peroxy compounds takes place at moderate temperature but, unfortunately, is stoichiometric. Therefore the effectiveness ceases once all the trivalent compound is oxidized.

Phosphites also function as alkoxy and peroxy radical trappers, as in reactions (10) and (11). Stabilization results if (**18**) is a hindered phenoxy radical with a low tendency to abstract hydrogen from polymer, i.e., if it is comparable in activity and stability to radicals generated from effective hindered phenolic antioxidants [34]:

$$(RO)_3P + R'OO\cdot \longrightarrow (RO)_3\dot{P}OOR' \qquad (10)$$
$$(RO)_3\dot{P}OOR' \longrightarrow (RO)_3P{=}O + R'O\cdot$$

$$R'O\cdot + (RO)_3 \longrightarrow (RO)_3\dot{P}OR' \qquad (11)$$
$$(RO)_3\dot{P}OR' \longrightarrow RO)_2POR' + RO\cdot$$

18

Certain cyclic phosphites, notably catechol phosphites, are catalytic hydroperoxide decomposers, their corresponding phosphates functioning as strong acids [35]:

where R is alkyl

Thiophosphites, thiophosphates, and easily hydrolyzable phosphates also are catalytic hydroperoxide decomposers. The thiophosphorus compounds are oxidized to acidic products that function as the active agents [36].

Uses. Phosphites, such as tris(nonylphenyl) phosphite (TNPP) (**19**), have been used commercially for about 40 years as stabilizers for unsaturated elastomers. In spite of limited stability to hydrolysis and attendant corrosion problems, TNPP continues to be used for low-cost antioxidant applications. More sophisticated trivalent phosphorus compounds have been marketed primarily to minimize molecular mass changes in polyolefins during compounding and fabrica-

Table 4. Influence of phenolic antioxidant, phosphite, and thiosynergist on polypropylene processing stability at 260°C

Multiple extrusions, Profax 6501	Melt flow (extrusions)			YI color 3 mm plaques (extrusions)		
	1st	3rd	5th	1st	3rd	5th
None	7 [a]	17	28	6	10	13
0.1 % AO 1 [b]	6	8	9	12	16	19
0.1 % AO 1 + 0.05 % P 1 [c]	4	6	9	12	13	16
0.1 % AO 1 + 0.05 % P 2 [d]	4	6	7	8	11	14
0.1 % AO 1 + 0.05 % S 1 [e]	6	8	9	10	13	16

[a] Melt flow was 3 for unprocessed polymer with 0.5 % BHT.
[b] Irganox 1010.
[c] Phosphite 168.
[d] Sandostab P-EPQ.
[e] Distearyl thiodipropionate.

tion. Several of these newer products are regulated by the FDA as indirect food additives for plastic food packaging. They are used with hindered phenolic antioxidants not only to minimize molecular mass changes, but also to reduce the color development on high-temperature processing and subsequent use that is usually associated with hindered phenols (Table 4; see above). The data in Table 5 show the reduction in concentration of phosphites during high-temperature processing of polypropylene. These results emphasize the sacrificial role of phosphite stabilizers and suggest they are not suitable candidates for long-term stabilization. However, by inhibiting polyolefin thermooxidative degradation during high-temperature processing, phosphites help prolong polymer life.

Table 5. Influence of processing on additive concentration; multiple extrusions at 260°C, Profax 6501

Stabilizers *	Additive conc., %	
	1st extrusion	5th extrusion
None	–	–
0.1 % AO 1	0.08	0.06
0.1 % AO 1 +	0.09	0.07
0.05 P1	0.04	0.02
0.1 % AO 1 +	0.09	0.05
0.05 % P2	0.04	<0.001
0.1 % AO 1 +	0.10	0.08
0.05 % S1	0.05	0.05

* Stabilizers identified in Table 4.

Table 6 (see next page) gives the major commercial trivalent phosphorus compounds used as antioxidants for plastics and rubbers.

4.5. Hindered Amines

Hindered amines are used primarily as light stabilizers but were also recognized early as stabilizers against thermooxidative degradation. The following data [37] summarize the antioxidant activity of hindered amine stabilizer in oven aged (120 °C) polypropylene multifilaments (8000/396 den):

Stabilizer	Days to embrittlement
Hindered amine stabilizer, O.1 %	47
Hindered amine stabilizer, 0.25 %	80
Hindered amine stabilizer, 0.5 %	110
Phenolic antioxidant-1, 0.1 %	14
Phenolic antioxidant-2, 0.1 %	20

In certain instances, particularly at moderate temperatures (< 120 °C), some hindered amines are more effective than hindered phenols for the stabilization of polypropylene stretched tapes [38]. These results are in line with the generally accepted *mechanisms of action* of hindered amines:

1) trapping of carbon free radicals by hindered amine nitroxyls [39]

2) trapping of peroxy radicals by hydroxyl amine ethers (reaction products from trapping carbon free radicals by hindered amine nitroxyls) with the regeneration of nitroxyls and formation of peroxides [40]

Table 6. Trivalent phosphorus antioxidants

Structure	CAS registry number	Trade names
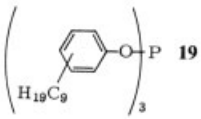 **19**	[*26523-78-4*]	TNPP
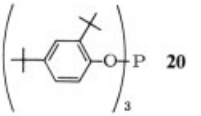 **20**	[*31570-04-4*]	Phosphite 168
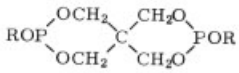 **21** R = $C_{18}H_{37}$, **22** R = 2,4-di-*tert*-butylphenyl	[*3806-34-6*]	Weston 618
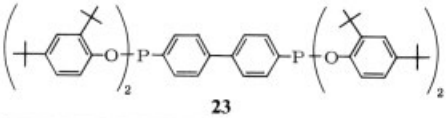 **23**	[*26741-53-7*]	Ultranox 626
	[*38613-77-3*]	Sandostab P-EPQ Irgafos P-EPQ

$$>NOR + R'O_2\cdot \longrightarrow >NO\cdot + R'OOR$$

3) complexing of hydroperoxides by hindered amines and hindered amine nitroxyls [41].

Trapping carbon free radicals is preferable to trapping peroxy radicals for purposes of stabilization because the former assures the oxidative chain length will be zero (no hydroperoxide formed), whereas trapping peroxy radicals yields one or more hydroperoxide products.

The complexation of hydroperoxides by stabilizer molecules is also a real advantage because it effectively increases the stabilizer concentration in the vicinity of the main species responsible for initiating oxidation.

Table 7 contains the main hindered amines available commercially.

4.6. Metal Deactivators

Metal deactivators were developed primarily to increase the lifespan of polyolefin insulation in contact with copper conductors in communications wire. Phenolic antioxidants alone are not effective in providing long-term stability. Copper is a prooxidant in polyolefin oxidation. It functions as a catalytic decomposer of hydroperoxides, generating alkoxy and peroxy radical initiators of oxidation:

$$\begin{array}{l} ROOH + Cu^{+} \longrightarrow RO\cdot + Cu^{2+} + OH^{-} \\ ROOH + Cu^{2+} \longrightarrow ROO\cdot + Cu^{+} + H^{+} \\ \hline 2\,ROOH \longrightarrow ROO\cdot + RO\cdot + H_2O \end{array}$$

In the absence of metal deactivators, copper penetrates polyethylene via oxidation reactions, yielding copper carboxylates among many products [42]. Metal deactivators markedly reduce the rate of copper penetration and are postulated to function by complexing copper at the polymer – metal interface and within the polymer [42]. Table 8 (see page after next) shows the severe effect of copper on the heat stability of polyethylene and the effectiveness of two different metal deactivators in reducing thermal degradation.

Table 7. Major commercial hindered amine stabilizers

Structure	CAS registry number	Trade names
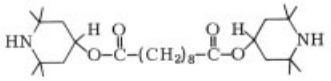 **24**	[*52829-07-9*]	Tinuvin 770
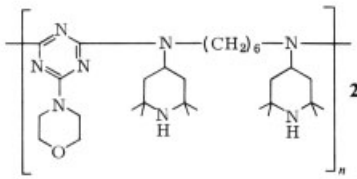 **25**	[*82451-48-7*]	Cyasorb UV-3346
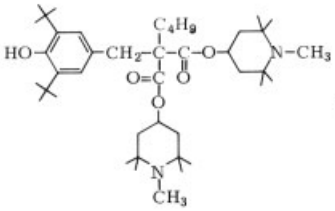 **26**	[*63843-89-0*]	Tinuvin 144
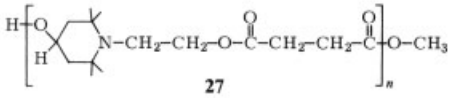 **27**	[*65447-77-0*]	Tinuvin 622LD
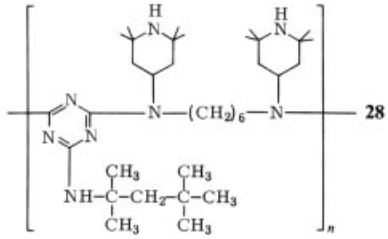 **28**	[*70624-18-9*]	Chimassorb 944LD
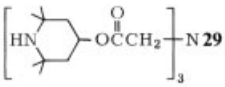 **29**	[*64022-57-7*]	Mark LA 55
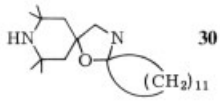 **30**	[*81406-61-3*]	Hostavin TMN 20
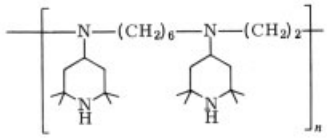 **31**	[*61269-61-2*]	Spinuvex A-36

Table 8. Thermal stability of poly(propylene – ethylene) compression molded plaques containing embedded copper screen

0.2 % Irganox1010 +	Hours to failure				
	120 °C		40 °C	150 °C	
	Cu	No Cu	Cu	No Cu	Cu
No metal deactivator	460	1710	360	850	100
0.2 % Irganox MD 1024	1320	1870	570	760	250
0.2 % Eastman OABH	1360	1780	500	760	250

The major commercially available metal deactivators are Irganox MD-1024 [*32687-78-8*], (**32**); Eastman Inhibitor OABH [*6629-10-3*], (**33**); Naugard XL-1 [*70331-94-1*], (**34**).

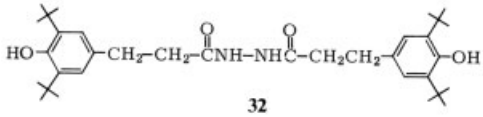

32

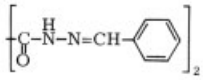

33

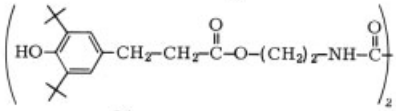

34

The amide or hydrazide functionalities complex copper, and are responsible for the inhibition activity. Normally, a combination of a hindered phenol and a metal deactivator is used to stabilize polyolefins in the presence of copper. However, because Irganox MD 1024 and Naugard XL 1 contain both phenolic and nitrogen groups, they can be used as the sole stabilizers for that application.

Metal deactivators are also used with hindered phenols for the stabilization of filled polyolefins, where metallic impurities in fillers adversely affect the thermooxidative stability of polymers [43].

5. Practices in Stabilization with Antioxidants

Specialized practices have evolved for each of the four major substrate classes that are stabilized with antioxidants:

1) foods,
2) fuels,
3) lubricants, and
4) polymers.

Some antioxidants are used in more than one substrate class. For the most part, however, each class has its own stabilizer types and amounts. Some antioxidant practices are common to all substrate classes: antioxidants for each class are selected on the basis of effectiveness, solubility, and cost. In the United States, antioxidants used in food and food-packaging plastics are limited to ones regulated by the Food and Drug Administration (FDA) for these uses.

5.1. Stabilization of Food

The antioxidants most commonly used to stabilize food against autooxidation are butylated hydroxytoluene (**5**), see Table 3; butylated hydroxyanisole (BHA) (**35**) [*25013-16-5*], mixture of 2- and 3-*tert*-butyl-4-methoxyphenol, Tenox BHA (Kodak); and *n*-propyl gallate (**36**) [*121-79-9*], 3,4,5-trihydroxybenzoic acid propyl ester, Tenox PG (Kodak).

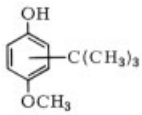

35

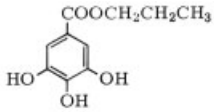

36

These antioxidants may be used directly in foods governed by FDA regulations. They are used to protect fats and oils during frying and retain activity during storage. In general they are used to prevent oxidation of foods, which results in deterioration of flavor and odor. Butylated hydroxytoluene is used in packaging materials for cookies, cold cereals, etc., where the relatively high-vapor pressure of BHT permits it to transfer from the packaging material to the contained food.

Use of BHT and BHA as direct food additives has been questioned as a result of published toxicity studies showing a higher than normal incidence of benign, precancerous, and cancerous tumors in test animals having been fed the antioxidants. However, detailed evaluations of the studies have always been resolved in favor of

the continued use of both antioxidants for food. A number of studies have supported the safety of BHT for the stabilization of food [44].

5.2. Stabilization of Fuels

Hydrocarbon fuels, gasoline, jet fuel, and fuel oil, are subject to autooxidation (→ Aviation Turbine Fuels, → Automotive Fuels). Autooxidation of fuels causes deposits that can foul carburetors, valves, pumps, and other precision equipment. Acidic oxidation products corrode critical metal parts in fuel systems. Hydroperoxide oxidation products are undesirable in gasoline because their rapid homolysis under combustion conditions leads to preignition, thereby lowering octane performance.

The addition of antioxidants to fuels effectively stabilizes them against autooxidation during storage. Both phenolic and aminic antioxidants are used, sometimes in combination. The main phenolic antioxidants for fuels are BHT (**5**); 2,4-dimethyl-6-*tert*-butylphenol, and 2,6-di*tert*-butylphenol. Among the aminic antioxidants, *N,N′*-di-*sec*-butyl-*p*-phenylenediamine has been of the greatest importance historically and continues to be used in fuels. However, other aminic antioxidants, some proprietary, including *p*-phenylenediamine with *N*-isopropyl or *N*-heptyl substituents, are also used.

Use levels of antioxidants for fuel stabilization are presently on the order of 0.5 – 1 kg per 1000 barrels of fuel (1 U.S. barrel = 159 L). This level is markedly lower than the 5 – 7 kg and more per 1000 barrels used in the past. The lower stabilizer requirement has come about because the composition of, and the production processes for fuels, especially gasoline, have increased the oxidative stability of base fuels.

5.3. Stabilization of Lubricants

Lubricating oils (→ Lubricant and Related Products) usually operate at relatively high temperature in gas turbines, steam turbines, jet engines, and crankcases of internal combustion engines. The performance requirements of stabilizers for these conditions are very demanding. Antioxidants minimize the deterioration of these lubricants by retarding viscosity increase, metal corrosion (usually controlled in conjunction with a corrosion inhibitor), and formation of acid, sludge, resins, and lacquers.

Mineral oil lubricants (gas turbines, steam turbines) are stabilized with 0.5 – 1.0 % of zinc dialkyldithiocarbamates in combination with 0.5 – 1.0 % of a phenolic antioxidant (BHT or 2,6-di-*tert*-butylphenol) or with an aminic antioxidant (octylated or nonylated diphenylamine, or phenyl α-naphthylamine or its derivatives).

Mineral crankcase oils may contain 0.5 – 1.0 % zinc dialkyldithiophosphates, which function as radical-trapping and peroxide-decomposing antioxidants in addition to their antiwear activity. These additives often are supplemented with sulfurized olefins (obtained by reacting sulfur with olefins), and aminic and phenolic antioxidants. The trend in the United States to unleaded gasolines and catalytic convertors has put pressure on the lubricant manufacturer to remove phosphorus-containing additives from crankcase oils, but so far with little success.

Mineral lubricants for diesel engines are stabilized with 0.5 – 1.0 % zinc diaryldithiophosphates, which are more thermally stable than the dialkyl derivatives and occasionally are supplemented with octylated or nonylated phenyl-α-naphthylamine.

Synthetic lubricants, both polyester and polyolefin types, are stabilized with 0.5 – 1.0 % alkylated phenyl-α-naphthylamine and/or alkylated diphenylamine.

5.4. Stabilization of Polymers

Virtually all commercial organic polymers, including elastomers, thermoplastics, synthetic fibers, and adhesives, are susceptible to oxidative degradation during processing and end use, with resulting reduction in physical properties and discoloration. Most polymers would not be useful without the small amounts of stabilizing additives, antioxidants, and processing stabilizers that are added to permit them to survive high-temperature processing and other thermooxidative conditions.

Antioxidant properties that influence performance include compatibility or solubility in the

polymer or other substrate, volatility, discoloration resistance, and thermal stability. Antioxidants are selected on the basis of how important these properties are for the specific application and how well antioxidants meet them.

Additional factors in selecting antioxidants include the desired lifetime, the product thickness (samples of small cross section are more difficult to stabilize than thick samples); other components of the formulation, i.e., fillers, pigments, and flame retardants that may influence the effectiveness of the antioxidant, use conditions, FDA status, and cost performance. The last is a key consideration because it is not the selling price of the antioxidant that should be determinant but the cost of stabilization to get the job done.

5.4.1. Elastomers (→ Rubber, 1.Survey)

Many different stabilization practices have evolved for polymers over the years. Elastomers, which commonly exhibit carbon – carbon unsaturation in the backbone, are particularly sensitive to oxidative degradation. They require stabilization during drying and storage in the uncompounded state as well as subsequently in the vulcanized forms. Both nonstaining and staining antioxidants are used. *Nonstaining, hindered phenolic antioxidants,* such as BHT (**5**), *phosphites*, such as tris(nonylphenyl phosphite) (**19**), and *nonstaining amines,* such as alkyl-substituted diphenylamines and alkyl-substituted phenyl-α-naphthylamines, are used when the final formulations are relatively light colored. However, because much of the production of elastomers is used in carbon-black-filled, vulcanized formulations, *staining antioxidants* may be used. These antioxidants contain dark-colored components capable of migrating into light-colored objects in contact with the vulcanized elastomer, causing objectionable discoloration or staining. Aryl- and alkyl-substituted *p*-phenylenediamine derivatives are commonly used as staining antioxidants for elastomers because they provide antioxidant as well as antiozonant activity. Other staining aminic antioxidants, such as aniline-acetone condensation products, also are used as antioxidants for elastomers.

Stabilization of elastomers against oxidation was historically among the earliest uses of polymer antioxidants [2]. Most were staining antioxidants could not be directly applied to other polymers. However, nonstaining antioxidants used in elastomers, hindered phenols, and phosphites particularly, and nonstaining amines, to a limited degree, are used also in other polymers.

5.4.2. Thermoplastics (→Thermoplastic Elastomers)

Stabilization in the early history of thermoplastics revolved around hindered phenols, such as BHT (**5**); 2,2′-methylenebis(4-methyl- 6-*tert*-butylphenol) (**3**), 4,4′-methylenebis(2,6-di-*tert*-butylphenol), 4,4′-thiobis(2-*tert*-butyl-5-methylphenol) (**14**), and other relatively simple phenolic antioxidants. These antioxidants were generally adequate for the limited stabilization demands of early thermoplastics. However, stabilization requirements became more demanding with the advent of polystyrene containing unsaturated elastomeric constituents for impact modification and especially with the introduction of polyolefins from Ziegler-Natta catalysis in the mid-1950s. More effective antioxidants were needed for these demanding applications. Today, sophisticated, relatively high-molecular-mass antioxidants, stabilize the bulk of commercial polymers (see Table 3). Although BHT continues to be used for the stabilization of thermoplastics, its relatively high-vapor pressure limits its use to low-demand applications.

Hydrocarbon thermoplastic polymers, which are the main consumers of antioxidants, *polystyrenes, polyolefins,* and *thermoplastic elastomers* are stabilized with phenolic antioxidants, phosphites or phosphonites, and thiosynergists; the composition and concentrations depend on life stage and end use, as shown in Table 9. These stabilizers are used to retard the loss of physical properties and to minimize discoloration of polymers that accompany thermooxidative degradation. The stability problems associated with these polymers are increased by recycling scrap from processing operations. Thus, some of the polymers may be processed one or more times, depending on the amount of scrap recycled.

Table 9. Practices in polymer stabilization

Life stage of polymer	Stabilizer	Usual concentration, ppm
Drying	antioxidant	<250
Storage	antioxidant	<250
Compounding	antioxidant	500 – 1000
	phosphite	500 – 1000
Fabrication	antioxidant	500 – 1000
	phosphite	500 – 1000
End-use		
low stress	same as compounding and fabrication	
thermally stressing	antioxidant	1000 – 5000
	thiosynergist	1000 – 5000
	phosphite	500 – 1000
weathering	antioxidant	0 – 1000
	phosphite	500 – 1000
	hindered amine	0 – 10000
	ultraviolet absorber	0 – 10000

Polyoxymethylene polymers (→ Polyoxymethylenes) are sensitive to hydrolysis so that in addition to 0.1 – 0.5 % hindered phenolic anti oxidants, 0.1 – 0.5 % of an acid – formaldehyde scavenger, such as cyanoguanidine, calcium ricinoleate, or terpolyamide, is also used.

Polyamides (→ Polyamides) discolor and their physical properties deteriorate as a result of thermal oxidation. Early in their history, polyamides were stabilized with copper halides and simple phosphites, sometimes even phosphorous acid. More recently, hindered phenols, notably *N,N′*−1,6-hexamethylene-bis-3-(3,5-di-*tert*-butyl- 4-hydroxyphenyl)propionamide (**8**), have been used at 0.1 – 0.5 % to stabilize nylon 6 and nylon 66. Polyamides with longer aliphatic chains, such as nylon 12, are stabilized with phenolic antioxidants, such as tetrakis[methylene(3,5-di-*tert*butyl-4-hydroxyhydrocinnamate)]methane (**2**).

Several polymers, notably *polyesters, acrylics,* and *polycarbonates,* are relatively thermally stable and require only small amounts of phosphites or hindered phenolic antioxidants to reduce discoloration.

5.4.3. Synthetic Fibers (→Fibers, 4. Synthetic Organic)

Synthetic fibers (polyamides, polyesters, and polyolefins) are stabilized in accordance with the practice for the corresponding thermoplastics, with the exception that the fibers are usually stabilized against photooxidation as well as thermooxidative degradation. Stabilization against photooxidation usually entails the use of hindered amines and ultraviolet absorbers as well as hindered phenolic antioxidants and phosphites.

5.4.4. Economic Aspects

The estimated consumption of antioxidants for plastics in the United States is given in Table 10 [45]. Table 11 shows the estimated consumption by plastic in the United States [46].

Table 10. Estimated consumption of antioxidants for plastics in the United States [45]

	$t \times 10^3$		
Antioxidant	1981	1982	1983
BHT	3.5	3.1	3.9
Other hindered phenols	5.6	5.2	7.3
Phosphites	3.2	2.9	4.5
Thioesters	2.7	2.4	3.1
Other	0.9	0.8	1.1
Total	15.9	14.4	19.9

	$\$ \times 10^6$		
Antioxidant	1981	1982	1983
BHT	12.8	10.2	15.9
Other hindered phenols	52.9	47.9	93.2
Phosphites	8.1	7.4	15.7
Thioesters	10.2	8.9	15.3
Other	4.0	3.4	6.7
Total	88.0	77.8	146.8

Table 11. Estimated consumption of antioxidants by plastic in the United States [46]

Plastic	Consumption, $t \times 10^3$		
	1981	1982	1983
ABS	4.6	3.7	4.4 (29 %)
Polyethylene	3.2	2.9	3.0 (20 %)
polypropylene	4.4	4.1	4.5 (29 %)
Polystyrene	2.2	1.9	2.0 (13 %)
Other	1.5	1.3	1.4 (9 %)
Total	15.9	13.9	15.3 (100 %)

6. Food Safety Considerations

The safety of antioxidants used as both direct and indirect food additives is monitored rigor-

ously by food agencies in various countries, such as the FDA in the United States. In order to be permitted for use as stabilizers for food packaging, the safety of an indirect food additive needs to be demonstrated in subacute feeding studies with test animals, usually rats and dogs, and by migration studies with food-simulating solvents. The latter studies permit the estimation of levels of additives that may migrate into various types of foods from plastic packaging materials during filling, sterilization, and storage.

These data and the use profiles of the plastic packaging materials permit the calculation of *estimated daily intakes* (EDI) of the additive. Safety assessment is made by examining the EDI in light of the results of the feeding studies, especially the no-effect levels established by the feeding studies. For more details of the food safety and toxicologic aspects of antioxidants, see [47], → Foods, 3. Food Additives.

7. Test Methods to Evaluate Polymer Thermal Oxidative Stability

Antioxidants are used selectively for specific polymers and for specific applications. Therefore, selecting the best antioxidant is dependent on the performance characteristics of the antioxidants, their limitations, the formulation methods, and the oxidation behavior of the organic material to be stabilized. Evaluations must be conducted on a small scale in a finite time to predict field results, sometimes of thirty or more years.

7.1. Melt Processing

Melt processing is an important operation in the preparation of polymer formulations and of molded objects. However, the temperatures and high-shear rates involved in this processing degrade many polymers. The efficacy of stabilizers in preventing the degradation is usually evaluated by repeated melt processing.

The polymeric material is extruded and injection or blow molded. Then the processed material is chopped and reprocessed anywhere from 2 to 20 times, depending on the polymer and the application. Changes in melt viscosity are usually monitored after specific numbers of processing cycles.

Generally, little or no change in melt viscosity on repeated processing is desired, and additives are evaluated relative to one another. In addition, the color of samples on repeated processing is of considerable importance. Additives that prevent discoloration and are not themselves discolored are especially desirable.

7.2. Oven Aging

Oven aging at elevated temperature is another important test method used in the plastics industry. By this method, oxidative degradation is accelerated and the results obtained in relatively short times. However, to minimize the risks of extrapolation, *temperature* for oven aging is selected as close to the end-use requirement as is feasible from a testing time point of view. Results from at least three test temperatures permit extrapolation of failure times in apparent Arrhenius plots to a use temperature at which a life span estimate is desired.

The two most important variables in oven aging are the temperature and *air velocity* in the oven. Tests carried out in ovens without air movement can give misleading results because additives of low-molecular mass do not volatilize in static air ovens to the degree that they do in actual use and to the degree that they do in ovens with significant air velocities. The measurement of oven air velocities is difficult because considerable variations exist depending on the location in an oven. One method of overcoming this difficulty involves continuous rotation of specimens within the oven during the test period.

7.3. Oxygen Uptake

Another useful method for studying accelerated thermal oxidative degradation of polymers is oxygen uptake. Polymer samples are enclosed in a system with an oxygen atmosphere, and the rate of oxygen consumption is measured. Oxygen uptake methods have the advantage of being relatively rapid but require more specialized equipment. In addition, the results from oxygen uptake investigations are more difficult to relate

to end-use conditions because of the problem in extrapolating results in an oxygen atmosphere to normal atmospheric conditions.

7.4. Thermal Analytical Methods

Differential scanning colorimetry (DSC) and thermal gravimetric analysis (TGA) are useful methods that can provide information relatively quickly on the thermal stability of polymers. However, these measurements are usually carried out at temperatures above 180 °C, where many polymers are molten. Extrapolation of the results to performance in solid polymers at ambient temperature is questionable. On the other hand, DSC is very useful for quality control purposes once a formulation has been selected. Also TGA is particularly applicable in assessing the thermal properties of polymers, polymer formulations, and additives. Measurements by TGA are carried out using both programmed temperature increases and isothermal conditions. Temperatures at 1 %, 5 %, 10 %, and 50 % loss in sample mass, for example, are noted in the programmed temperature mode. In the isothermal mode, time to these mass losses are recorded.

7.5. Tests to Monitor Polymer Properties

Color and physical property measurements are the two most important ways of assessing performance of stabilizer compositions.

Color. Discoloration of plastic articles is usually undesirable. Color is generally monitored initially and after various exposure periods. Color may be assessed visually or measured instrumentally. Instruments generally read three parameters: L, the brightness of the specimen; a, the redness or greenness of the specimen; and b, the blueness or yellowness of the specimen; and tristimulus values X, Y, and Z. Various relationships of L, a, and b have been used to express specimen color. Yellowness index, a commonly used expression for relative color levels, is calculated for transparent samples from the tristimulus values:

$$\frac{1.28X - 1.06Z}{Y} \times 100$$

Table 12. Specific test details for selected polymers

Polymer	Property	Test conditions
ABS	color color impact strength	repeated processing oven aging oven aging
Polyacetal	weight loss weight loss	TGA 200 – 250 °C oven aging
Polyamides	solution viscosity color tensile retention of fiber	oven aging oven aging oven aging
Polycarbonate	color impact strength	oven aging oven aging
Polyester	color impact strength	oven aging oven aging
Polyethylene, low density	M_r – melt index gloss and clarity of film	repeated processing film blowing
Polyethylene, low density, cross-linked	tensile strength and elongation	oven aging
Polyethylene high density	M_r – melt index physical integrity embrittlement	repeated processing oven aging to
polypropylene	M_r – melt index physical integrity	repeated processing oven aging to embrittlement
Polystyrene, impact	color color impact strength	repeated processing oven aging oven aging
Polyvinyl chloride	color	oven aging
Polybutadiene rubber	M_r – Mooney viscosity gel content	oven aging oven aging
Thermoplastic elastomers	color tensile and elongation	oven aging oven aging

Physical Properties. Among physical test methods commonly used for evaluating the oxidative stability of polymers are tensile strength and elongation, and impact strength. *Elongation* is a more useful measure of oxidative degradation than tensile strength. *Tensile strength* may not be markedly affected by sample embrittlement, whereas elongation is much more sensitive and can provide evidence of degradation more readily. However, because both measurements are made simultaneously, it is not necessary to choose one over the other prior to testing.

Impact strength is measured in different ways depending on the end-use performance of articles made from the polymers in question. Among the methods used are unnotched Izod, notched Izod, tensile-impact, chip impact, falling dart, and high-speed-impact testing. Each has disadvantages and no single ideal impact test method exists. However, high-speed-impact testing with instrumented tups has several key advantages and would very likely be more universally adopted were it not for the relatively high cost of equipment, in the neighborhood of $ 50000 – $ 60000.

Monitoring of molecular mass changes on oxidative degradation is also a frequently employed technique. The melt index determination (Section 7.1) is a standard technique in the plastics industry. Molecular mass is also monitored by solution viscosity measurements, less frequently by gel permeation chromatography, and, in the case of elastomers, by Mooney viscosity measurements.

In undergoing oxidation, hydrocarbon polymers develop carbonyl groups (i.e., ketones, aldehydes, acids, esters, etc.). The monitoring of carbonyl absorbance by infrared spectroscopy is a useful technique for following the course of polymer oxidation.

For the specific test used to assess various polymer properties, see Table 12.

8. References

1. A. W. von Hofmann, *J. Chem. Soc.* **13** (1861) 87.
2. A. Seyewetz, P. Sisley, *Bull. Soc. Chim. De France* **31** (1922) 672.
3. L. Bateman, *Q. Rev. Chem. Soc.* **8** (1954) 147.
4. J. L. Bolland, *Q. Rev. Chem. Soc.* **3** (1949) 1.
5. J. L. Bolland, *Proc. Royal Soc. London Ser. A* **186** (1946) 218.
6. G. L. Russell, *J. Am. Chem. Soc.* **79** (1957) 3871.
7. P. D. Bartlett, T. G. Traylor, *J. Am. Chem. Soc.* **85** (1963) 2407.
8. T. G. Traylor, *J. Am. Chem. Soc.* **85** (1963) 2411.
9. T. G. Traylor, C. A. Russell, *J. Am. Chem. Soc.* **87** (1965) 3698.
10. J. L. Bolland, P. Ten Have, *Trans. Faraday Soc.* **43** (1947) 201.
11. A. J. Kolka, J. P. Napolitano, A. H. Filbey, G. C. Ecke, *J. Org. Chem.* **22** (1957) 642.
12. T. J. Henman: *World Index of Polyolefin Stabilizers,* Kogan Page, London 1982.
13. A. F. Bickel, E. C. Kooyman, *J. Chem. Soc.* 1953, 3211.
14. A. F. Bickel, E. C. Kooyman, *J. Chem. Soc.* 1956, 2215.
15. J. Pospisil, *Adv. Polym. Sci.* **36** (1980) 69.
16. J. Pospisil in N. S. Allen (ed.): *Advances in Polymer Photochemistry-2,* Applied Science Publishers Ltd., Essex, England 1981, p. 53.
17. A. F. Bickel, E. C. Kooyman, *J. Chem. Soc.* 1957, 2217.
18. J. R. Thomas, C. A. Tolman, *J. Am. Chem. Soc.* **84** (1962) 2930.
19. I. T. Brownlie, K. U. Ingold, *Can. J. Chem.* **45** (1967) 2427.
20. J. R. Shelton in W. L. Hawkins (ed.): *Polymer Stabilization,* Wiley-Interscience, New York 1972, p. 84.
21. J. R. Shelton in G. Scott (ed.): *Developments in Polymer Stabilization-4,* Applied Science Publishers, Essex, England 1981, p. 23.
22. G. Scott in G. Scott (ed.): *Developments in Polymer Stabilization-6,* Applied Science Publishers, Essex, England 1983, p. 29.
23. S. Al-Malaika, K. Chakraborty, G. Scott in G. Scott (ed.): *Developments in Polymer Stabilization-6,* Applied Science Publishers, Essex, England 1983, p. 73.
24. G. W. Kennedy, W. L. Patterson, Jr., *Ind. Eng. Chem.* **48** (1956) 1917 – 1924.
25. H. Hock, S. Lang, *Ber. Dtsch. Chem. Ges.* **77 B** (1944) 257.
26. W. L. Hawkins, H. Sautter, *Chem. Ind.* 1962, 1825 – 1826.
27. J. R. Shelton, E. R. Harrington, *Rubber Chem. Technol.* **49** (1976) 147.
28. N. P. Neureiter, D. E. Bown, *Ind. Eng. Chem. Prod. Res. Dev.* **1** (1962) 236 – 241.
29. C. R. H. I. DeJonge, P. Hope in G. Scott (ed.): *Developments in Polymer Stabilization-3,* Applied Science Publishers Ltd., Essex, England, 1980, p. 21.
30. T. Colcough, J. I. Cunneen, *J. Chem. Soc.* 1964, 4790.
31. A. J. Burn, *Tetrahedron* **22** (1966) 2153 – 2161.
32. D. Barnard, L. Bateman, M. E. Cain, T. Colclough, J. I. Cuneen, *J. Chem. Soc.* 1961, 5339.
33. D. Rysavy, *Kunststoffe* **60** (1970) 120.
34. K. Schwetlick, *Pure Appl. Chem.* **55** (1983) 1629 – 1636.

35. K. J. Humphris, G. Scott, *J. Chem. Soc.* 1974, 617.
36. J. Holcik: *4th International Conference on Advances in the Stabilization and Controlled Degradation of Polymers,*Lucerne, Switzerland, June, 1982.
37. A. Tozzi, G. Cantatore, F. Masine, *Text. Res. J.* **48** (1978) 433.
38. CIBA-GEIGY, unpublished data.
39. K. Murayama, S. Morimura, Y. Toshioka, *Bull. Chem. Soc. Jpn.* **42** (1969) 1640.
40. Y. B. Shilov, F. T. Denisov, *Vysokomol. Soed. Ser. A* **16** (1974) 2313.
41. D. W. Grattan, A. H. Reddock, D. J. Carlsson, D. M. Wiles, *J. Polym. Sci., Polym. Lett. Ed.* **16** (1978) 143.
42. D. L. Allara, C. W. White, R. L. Meek, *J. Polym. Sci.* **14** (1976) 93.
43. H. Mueller in R. Gaechter, H. Mueller (ed.): *Plastic Additives,* Hanser-Verlag, München 1985, p. 75.
44. William H. Lederer: "Status of Butylated Hydroxylated Toluene," International Life Sciences Institute, Washington, D.C., Sept., 1982.
45. R. Lacallade, *Chem. Eng. News* 1983 (October 24) 18.
46. *Mod. Plast.* **1983** (September) 1967.
47. R. Leimgruber in R. Gaechter, H. Mueller (ed.): *Plastics Additives,* Hanser-Verlag, München 1985, p. 685.

Antiozonants → Rubber, 4. Chemicals
Antiperspirants → Skin Cosmetics
Antiphlogistic Agents → Anti-inflammatory–Antirheumatic Drugs
Antiprotozoal Chemotherapeutics → Chemotherapeutics
Antipyretics → Analgesics and Antipyretics
Antirheumatic Drugs → Anti-inflammatory–Antirheumatic Drugs
Antiseptics → Disinfectants
Antisera → Immunotherapy and Vaccines
Antispasmodics → Spasmolytics
Antistatic Agents → Plastics, Additives
Antistatic Agents → Textile Auxiliaries
Antitussives → Cough Remedies

Arsenic and Arsenic Compounds

Sabina C. Grund, SMC-Metallurgical Consultant, Dorsten, Federal Republic of Germany

Kunibert Hanusch, Hüttenwerke Kayser, Lünen, Federal Republic of Germany

Hans Uwe Wolf, Universität Ulm, Ulm, Federal Republic of Germany (Chap. 7)

1. Arsenic 1
1.1. Introduction 1
1.2. Physical Properties 2
1.3. Chemical Properties 2
1.4. Occurrence 3
1.5. Production 3
1.5.1. Production of Arsenic Trioxide . 3
1.5.1.1. Ore Dressing 4
1.5.1.2. Pretreatment of Arsenical Materials 5
1.5.1.2.1. Roasting 5
1.5.1.2.2. Alternative Pretreatments of Arsenical Materials 11
1.5.1.3. Production of Refined As_2O_3 . . 12
1.5.2. Production of the Metal 14
1.6. Treatment of Arsenic-Containing Byproducts 17
2. Arsenic Compounds 21
2.1. Arsenic Trioxide 22
2.2. Arsenous Acid 22
2.3. Arsenic Pentoxide 23
2.4. Arsenic Acid 23
2.5. Arsenic Sulfides 24
2.6. Arsenic Halides 25
2.7. Arsenic Hydride and Arsenides 26
2.8. Organoarsenic Compounds . . . 27
2.8.1. Aliphatic Organoarsenic Compounds 28
2.8.2. Aromatic Organoarsenic Compounds 28
3. Uses and Economic Aspects . . 29
4. Stabilization of Arsenic for Safe Disposal 30
5. Recycling 32
6. Safety Measures 32
7. Toxicology 32
8. References 36

1. Arsenic

1.1. Introduction [1], [2]

Arsenic, As [*7440-38-2*], belongs to group 15 of the periodic table and has an atomic number of 33 and A_r 74.92. In compounds, it has oxidation states of +3, +5, and −3.

Arsenic is a bright silver-gray metal. Its surface tarnishes in humid air. Arsenic forms trigonal crystals, which are brittle and of average hardness (3 – 4 on the Mohs scale).

In addition to the metallic modification, there are other modifications, namely, yellow arsenic and three amorphous forms (β, γ, δ) [3]. Of these, black arsenic (the β form) is the best known. Black arsenic is formed as a coating (arsenic mirror) when arsenic hydride is passed through an incandescent glass tube and also, together with other modifications, when the vapor is rapidly cooled. Yellow arsenic is formed by the sudden cooling of arsenic vapor and consists of transparent, waxy, regular crystals. It is unstable and changes into metallic arsenic on exposure to light or on gentle heating. All amorphous modifications change into the metallic form above 270 °C.

Another form, brown arsenic (ϱ 3.7 – 4.1 g/cm^3), is either a special modification or simply a more finely divided form. This brown modification is obtained in the reduction of solutions of arsenic trioxide in hydrochloric acid with tin(II) chloride or hypophosphorous acid.

History [1], [2], [4]. The natural sulfides, realgar and orpiment, were known in pre-Christian times and had already been named by Aristoteles and his pupil Theophrast. In the first century a.d., Plinius mentioned sandarac (realgar) as occurring in gold and silver mines. The color of the yellow sulfide gave it the name orpiment. The Greek name "arsenikon" (fearless, brave, manly), which was given to the sulfides on account of their reactivity with metals, derived from the custom prevailing at the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_113

time of classifying the metals as male or female. Subsequently, the name was applied to the metal itself. Many early authors described medicinal applications. DIOSKORIDES (first century A.D.) reported on the roasting of the sulfides and that they were used as remedies and artist's colors. In the fifth century A.D., OLYMPIODORUS described the production of the arsenic oxide As_2O_3. An incomplete description of metallic arsenic dating from the fourth century has also been found. The alchemists were more closely concerned with the sulfides of arsenic. They generally called the roasted product "white arsenic". Many considered the arsenic oxide, like sulfur, to be a basic constituent of metals. ALBERTUS MAGNUS apparently obtained the element in 1250; SCHROEDER described two methods of preparation in 1649. In 1733, BRANDT showed that white arsenic is the oxide of the metal. The alchemists were aware of the poisonousness of white arsenic (As_2O_3). Records show that the compounds of arsenic were extensively used by the poison brewers of the Middle Ages. In the Renaissance it was PARACELSUS who pioneered the use of arsenic compounds in medicine. Arsenic acid and arsenic hydride were discovered by SCHEELE in 1775.

1.2. Physical Properties [1], [2], [5]

Densities: metallic arsenic 5.72 g/cm^3 at 20 °C, yellow arsenic 2.03 g/cm^3
Melting point: 1090 K (817 °C) at 3.7 MPa
Sublimation point: 886 K (613 °C) at 0.1 MPa
Linear coefficient of thermal expansion: ca. $5 \times 10^{-6}\,K^{-1}$
Specific heat capacity: $0.329\,J\,g^{-1}\,K^{-1}$ at 291 K (18 °C), $0.344\,J\,g^{-1}\,K^{-1}$ on average between 273 K (0 °C) and 373 K (100 °C)
Electrical resistivity: $24 \times 10^{-6}\,\Omega \cdot cm$ at 273 K (0 °C)

The normal potential of arsenic with respect to the normal hydrogen electrode is ca. 0.24 V. Arsenic therefore comes between bismuth and copper in the electrochemical series.

1.3. Chemical Properties [1], [2], [5]

The various forms of arsenic have different behaviors in air. Whereas the amorphous forms remain unchanged for months in dry air, the crystallized α form turns black in a few days, as originally observed by BERZELIUS. Powdered, moist arsenic is oxidized to arsenic trioxide with evolution of heat. When heated in air, it burns with a bluish-white flame, forming dense vapors of arsenic trioxide. When heated, arsenic metal gives off a characteristic garliclike odor.

Concentrated nitric acid and aqua regia oxidize arsenic to arsenic acid; arsenic is oxidized to the +3 state by dilute nitric acid or concentrated sulfuric acid and by boiling alkali metal hydroxides in air. Hydrochloric acid has little effect on arsenic. Chlorine combines fierily with arsenic to form arsenic trichloride. When the metal is heated with sulfur, AsS, As_2S_3, or As_2S_5 is obtained, depending on the ratios used. A mixture of finely divided arsenic and potassium chlorate explodes on impact.

Arsenic combines with metals to form arsenides. When subjected to oxidizing roasting, arsenides give partly metal oxide and arsenous acid and partly basic arsenates. When subjected to chloridizing roasting, they give arsenic trichloride. On heating in the absence of air, a sublimate of arsenic metal is formed from the heavy metal arsenides, although the arsenic can only be partly removed even at very high temperatures.

With oxygen, arsenic forms three oxides, namely arsenic trioxide, As_2O_3; arsenic pentoxide, As_2O_5; and As_2O_4, which apparently contains both trivalent and pentavalent arsenic. The first two oxides may be regarded as acid anhydrides. Arsenous acid is derived from arsenic trioxide and can only exist in aqueous solution. Its well-known salts are the arsenates(III) (formerly arsenites). Arsenic acid is derived from arsenic pentoxide. Its salts, the arsenates, generally correspond to the phosphates in their stoichiometry. However, solid arsenic acids that correspond to the phosphorus acids are not known; instead, $H_3AsO_4 \cdot 1/2\,H_2O$ or $As_2O_5 \cdot 4\,H_2O$ crystallizes from arsenic acid solutions.

The highly poisonous arsenic hydride, AsH_3, is formed from arsenic compounds in acidic solution in the presence of strong reducing agents (e.g., Zn) or from suitable arsenides (e.g., As_2Zn_3) and acids.

1.4. Occurrence [1], [6–9]

Arsenic is widespread and can be detected in traces everywhere. With an abundance in the earth's crust of ca. $(2-10) \times 10^{-4}$ %, it is scarce but is found in higher concentrations in sulfide deposits (60 ppm or more). Worldwide most arsenic is produced as a byproduct of copper and lead smelting.

Native arsenic is found in many places, but only in small quantities, usually in ores containing gold, silver, cobalt, nickel, and antimony. Although many arsenic-bearing minerals are known, only a few occur in such quantities that they can be worked economically. The most important arsenic minerals are shown in Table 1. Other arsenides are chloanthite, $NiAs_2$; niccolite, NiAs; smaltite, $CoAs_2$; cobaltite, CoAsS; gersdorffite, NiAsS. Arsenic-bearing fahlores include inter alia tennantite, $4\,Cu_2S \cdot As_2S_3$; proustite, $3\,Ag_2S \cdot As_2S_3$. These minerals may be regarded as thioarsenites. Enargite, $3\,Cu_2S \cdot As_2S_5 = Cu_3AsS_4$, is a thioarsenate.

An extensive listing of all natural occurring arsenic minerals is given in [9].

1.5. Production

Most of the arsenic produced commercially accumulates as a byproduct in the smelting of nonferrous metal ores containing gold, silver, lead, copper, nickel, and cobalt. Today China is believed to be the only country that produces arsenic trioxide directly from arsenic ores [10].

Arsenopyrite often contains primary gold deposits. Fahlores and proustites always contain silver. Gersdorffite can accompany copper ores. Arsenopyrite and löllingite accompany pyrites [11]. Arsenic is mainly obtained from complex ores, such as enargite-containing copper, lead, and zinc ores. The arsenic content varies considerably.

Because of the volatility of arsenic sulfides and oxides, arsenic is mostly concentrated in the gas phase during roasting or smelting operations. A crude oxide is produced from the intermediate products, and, after refining, is reduced to the metal or is sold as is to the chemical industry for further treatment.

The falling world demand for arsenic oxide, the increasingly elaborate treatment of complex ores, the high toxicity, and environmental requirements have made arsenic-containing byproducts a problem for many producers. During the last decade research focussed on the stabilization of arsenic for safe disposal rather than on production technologies. In the early 1990s only China and Chile have countered the trend toward decreasing arsenic production [10]. Today arsenic is almost always an unwanted accompanying element in metallurgy.

1.5.1. Production of Arsenic Trioxide

History. Arsenic trioxide has been produced in China for 500 years in extremely simple plants, consisting of retorts with condensation chambers, from an ore containing 15 % As. However, it was not until the beginning of the 18th century that a smelter for native arsenic was set up in Germany at the instigation of J. v. SCHARFENBERG in Reichenstein, Silesia. Because the demand for As_2O_3 was relatively low, it was not until the 19th century that a second country, Great Britain, began producing As_2O_3. Thereafter, from the middle of the 19th century to 1901, Great Britain was the leading As_2O_3 producer. Production figures up to 1945 are only estimates. When legislation was introduced in various countries to avoid damage caused by release of As_2O_3-containing fumes, world production of As_2O_3 increased considerably. The production of native arsenic was given a fresh impetus by the appearance of the boll weevil in Mexico and the United States – damage from 1909 to 1923 was estimated at \$ 3 billion. Calcium arsenate was used to combat the boll weevil. The demand for arsenic became so great that it exceeded supply. New plants were established, particularly in the United States and Mexico. The largest plant was operated between 1932 and 1962 by Boliden AB in Rönnskär, Sweden.

Since the main use of arsenic trioxide as wood and plant preservative is declining (only the United States and developing countries such as Malaysia still allow the application of arsenic trioxide for this purpose), changes in the worldwide structure of arsenic trioxide production can be observed. While major producers of arsenic trioxide such as Boliden Mineral in

Table 1. Most important arsenic-bearing minerals

	CAS registry number	Mohs hardness	ϱ, g/cm^3	Crystal system	Arsenic content, %
Arsenopyrite (FeAsS)	[*1303-18-0*]	5.5 – 6	5.9 – 6.2	orthorh.	46
Löllingite ($FeAs_2$)	[*12255-65-1*]	5 – 5.5	7.1 – 7.3	orthorh.	73
Orpiment (As_2S_3)	[*12255-89-9*]	1.5 – 2	3.4 – 3.5	mon.	61
Realgar (AsS)	[*12044-30-3*]	1.5 – 2	3.4 – 3.6	mon.	70
Native arsenic	[*7440-38-2*]	3 – 4	5.6 – 5.8	trig.	90 – 100

Sweden and Asarco in the United States ceased production, mostly for environmental reasons, China emerged as the world's largest producer of both arsenic trioxide and arsenic metal, selling both commodities to the United States [10].

1.5.1.1. Ore Dressing

The mining and dressing of ores with the primary object of recovering arsenic minerals are of little if any importance. Most arsenic raw materials are byproducts from the dressing and smelting of arsenical ores. The raw materials differ widely in character: mostly they are sulfides, often with pyrites as the principal constituent. For dressing, mixed arsenide – sulfide concentrates of very high metal content are floated, whereas the concentration of pure arsenic minerals is of secondary importance and should be avoided altogether [1].

The non ferrous metal concentrates used for the recovery of arsenic have arsenic contents from < 1 % to, in exceptional cases, 10 %. Table 2 shows As distributions in the concentrates and in the waste products, based on the initial As concentration in the ore [1].

In the concentration of complex ores, most of the arsenic remains in the tailings. In dressing tin ore, only 7.8 % of the arsenic passes into the tin concentrate. In the case of copper and copper – zinc ores, the arsenic is concentrated in the copper concentrate to a level of around 30 % of the quantity originally present in the starting ore. The copper concentrate therefore contains 0.5 – 1 % As, and concentrates of the complex ores up to 5 – 8 % As [1], [12].

One example of copper ore in which a large part of the copper is bound to arsenic, as enargite or luzonite, $3\,Cu_2S \cdot As_2S_5$, is the Lepanto ore from the Philippines. The flotation product contains approximately 28 % Cu, 32 % S, and 9 % As. In most ores, however, only a small proportion of the copper is bound chemically to arsenic, and the copper concentrates have a considerably lower arsenic content. Enargite-containing copper concentrates are produced in Cerro de Pasco (Peru) and Butte, Montana (USA) [1].

Arsenopyrite often occurs in pyritic copper ores, for example, on the Iberian Peninsula, in the Balkan countries, in Sweden, and in New Brunswick, Canada. The arsenopyrite can be recovered by selective flotation. Copper pyrites are removed by flotation with higher xanthates, the other sulfide minerals being held back by lime. After acidifying and heating the pulp, the pyrites and lastly, after further acidification and activation with copper sulfate, the arsenopyrite are flotated. Alternatively, after flotation of the copper pyrites, the arsenopyrite may be selectively floated from the pyrites pulp after heating and activation with copper sulfate.

In many precious metal ores, gold and silver occur together with arsenopyrite (FeAsS), löllingite ($FeAs_2$), and pyrites or other sulfides. Although the metals are not chemically bound to the arsenic, instead occurring as native gold and in the form of nonarsenical minerals, separation by selective flotation is not satisfactory on account of a partially extensive intergrowth of the minerals. In such a case, a combined concentrate of sulfides and arsenides with an arsenic content from a few percent to more than 30 % is floated. Research is still focussed on improving the separation of refractory gold and arsenical minerals by means of ore dressing, but in the case of refractory Au ores the limit of flotation and other classical ore dressing techniques has been reached [13], [14].

An example of the dressing of a combined precious metal and arsenic ore is the ore from Salsigne, France: an arsenopyrite – pyrites ore containing precious metals in large quantities, some copper, and a little bismuth. A description

Table 2. Distribution of arsenic in the production of the concentrated raw materials (% of arsenic content in starting ore) [12]

	Complex ores	Copper and copper – zinc ores	Tin ore	Gold-bearing ore	Nickel – cobalt-bearing ore
Starting ore	100	100	100	100	100
Concentrates, total	15.3	41.0	17.4	35.0	1.9
Lead	7.8	–	–	–	–
Zinc	3.8	1.8	–	–	–
Copper	3.3	30.1	9.6	–	–
Pyritic	0.4	9.1	–	–	–
Tin	–	–	7.8	–	–
Gold-containing	–	–	–	35.0	–
Others	–	–	–	–	1.9
Waste products	84.7	59.0	82.6	65.0	98.1

of the flotation circuit is given in [15]. The metal concentrate obtained contains ca. 23 % As, 27 % S, 34 % Fe, 0.7 % Cu, 0.6 % Bi, 55 ppm Au and 115 ppm Ag [1], [15].

Lead and copper concentrates rich in arsenic are produced in Mexico, South America, and South West Africa; cobalt concentrates in North Africa; and precious metal concentrates in Canada, South America, and other regions [1].

A trend in mining and minerals processing is to produce so called dirtier concentrates from complex ores. Therefore, the copper and zinc industries are forced to cope with concentrates which are higher in arsenic content. At the same time, new environmental laws mean increased effort and therefore costs for avoiding the distribution of arsenic in byproducts and for stabilizing it for save disposal. The tolerable As content of zinc concentrates for hydrometallurgical zinc recovery (about 80 – 85 % of the total primary zinc recovery) is limited to about 0.8 % [16–18].

Nickel concentrate at the Luikonlahti Concentrator of Partek in Finland has varying As contents of up to 5 % As. To transform such high-arsenic Ni concentrates into a saleable product, research was undertaken to separate pentlandite from As-containing niccolite and gerdorffite by advanced flotation with a combination of grinding, aeration, and NaCN depression [19].

Arsenic ores worth extracting and usable in concentrated or unconcentrated form as starting material for the production of arsenic seem to be processed only in China [10].

1.5.1.2. Pretreatment of Arsenical Materials

Arsenic in nonferrous metal smelter feed constitutes a major source of concern since it increases production costs, interferes with metal extraction, lowers product purity and quality (e.g. deformability), represents an environmental hazard, and poses disposal problems. Therefore, arsenic should be separated as early as possible, which is also dictated by industrial safety regulations. Especially in pyrometallurgical processes, the formation of arsenic-rich phases (speisses) in the reducing smelting of arsenic-rich raw materials leads to intolerable losses because of their ability to dissolve large quantities of precious metals [1], [8], [20].

Processes for the dearsenifying pretreatment of arsenical materials are therefore of great interest and increasing importance in the nonferrous metals industries. Among various pyro- and hydrometallurgical options, dearsenifying roasting is quite widely used. For certain applications, hydrometallurgical processes such as oxidative pressure leaching or biochemical pretreatment are applicable alternatives.

1.5.1.2.1. Roasting

The separation of arsenic by roasting is complicated by the multivalency of arsenic, i.e., its ability to form more than one compound or complex with a given element. Arsenic in arsenopyrite can be transferred to the vapor phase as any of its volatile compounds: As_4S_4, As_2S_3, AsS, As_4O_6, As_4, As_2, etc. Besides $AsCl_3$, As_4O_6 possesses by far the highest vapor pressure. However, which compound will be predominant depends on the prevailing partial pressures of oxygen and sulfur.

The final product should contain so little arsenic that it can be further treated as a quasi-arsenic-free material. To obtain a high arsenic

removal rate and hence a high concentration of arsenic in as few processing steps as possible, the roasting process must be conducted such that only volatile arsenic compounds or As_4 or As_2 form. The formation of nonvolatile arsenic compounds such as As_2O_5 or arsenates, which remain in the roasted material, must be avoided. The valuable metals remain in the roasted material as sulfide or oxide, depending on the process conditions [1], [8], [21–23].

Although predominance area diagrams represent equilibrium conditions, they can be used to evaluate roasting phenomena. Figure 1 shows three faces of the quaternary system Fe – As – S – O at 798 K. The diagram reveals that the partial pressures of O_2, S_2, and As_4 are each limited by the precipitation of a condensed phase. The partial pressure of O_2 in the Fe – As – O system is limited by the reaction between As_4 and O_2 in the production of the solid pentoxide [22].

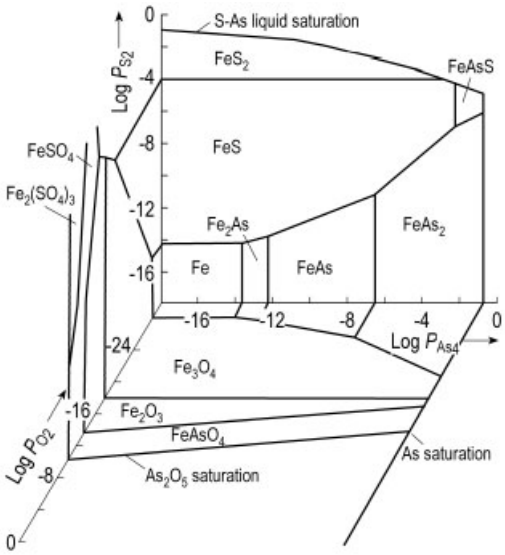

Figure 1. Predominance area diagram for the Fe – As – S – O system at 798 K [22]

Reviews of the roasting conditions reveal that inert, reducing, and oxidizing conditions can be used to eliminate arsenic from arsenical materials. To separate As from concentrates, oxidizing conditions appear to be the most promising. In this case, higher temperatures and higher oxygen partial pressures lead to the stability area of As_4O_6. Above a certain temperature that depends on the oxygen partial pressure, nonvolatile As_2O_5 and metal arsenates, which decompose only at high temperatures, become stable, and this leads to a poor yield of As in the flue dust. Higher As contents in the raw material result in a better dearsenification, and higher SO_2 partial pressures expand the stability field of As_4O_6 [8], [21–23].

At high SO_2 partial pressures, i.e., at low O_2 partial pressure in the roasting zone, volatile arsenic oxides are mainly formed:

$$4\,MAs + 5\,O_2 \longrightarrow 4\,MO + 2\,As_2O_3$$

At higher O_2 partial pressures the overall course of the reaction may be written as:

$$MAs + 2\,O_2 \longrightarrow MAsO_4$$

As_2S_2 and As_2S_3 are volatile at low temperatures. Arsenopyrite and similar compounds decompose:

$$FeAsS \longrightarrow FeS + As$$

At low oxygen partial pressure, particularly in the case of pyrites-containing materials, the sulfur partial pressure reaches such high levels that reactions of the following type attain practical significance in the gas phase:

$$2\,S_2 + As_4 \longrightarrow As_4S_4$$

$$As_4S_4 + 7\,O_2 \longrightarrow (As_2O_3)_2 + 4\,SO_2$$

In the dearsenifying roasting of raw materials containing arsenopyrite or other dissociable substances, the dissociation reaction and secondary reactions are important factors. The best results are obtained by roasting at low oxygen partial pressure. To convert all the arsenic into the trioxide, the oxidation of the dissociation products has to be completed at a relatively high oxygen partial pressure. Before cooling and purification, therefore, the roasting furnace gases are burnt under controlled conditions. If only partial desulfurization is necessary, a single-stage process is sufficient. For complete desulfurization, however, a two-stage process may be necessary [1].

If roasting is correctly controlled, arsenic can be separated in high yields from the most important arsenic minerals, namely arsenopyrite, FeAsS; löllingite, $FeAs_2$; and enargite,

$3Cu_2S \cdot As_2S_5$. Nickel and cobalt arsenide minerals are much more stable and therefore more difficult to roast. Fahlores and proustite are also very stable [1].

The following roasting processes are in use [8]:

- Partial roasting: Here the calcine still contains FeS, and there is no tendency for formation of stable metal arsenates. Arsenic elimination depends on the retention time. Multiple-hearth and single-stage fluidized-bed roasting are applicable.
- Magnetite roasting: In this case the calcine is in the Fe_3O_4 form, and arsenic elimination is satisfactory. The Boliden process is based on one-stage fluidized-bed roasting to magnetite.
- Dead roasting: The calcine is principally in the Fe_2O_3 form with residual sulfur. Arsenic can be eliminated by multiple-hearth roasting or two-stage fluidized-bed roasting, as in the BASF process.
- Reducing roasting: To prevent formation of iron(III) arsenate, roasting is sometimes carried out under reducing conditions, especially if the required degree of sulfur elimination is high, as in the dearsenicating roasting of complex tin concentrates. It has also been applied in the partial roasting of copper concentrates to obtain low residual As levels.
- Sulfidizing roasting: In the sulfidizing evaporation pretreatment process of Outokumpu, it was proposed to operate in the p_{S_2} range of 0.05 – 0.1 bar at 600 – 800 °C , which would result in 0.1 % As in the calcine, independently of the degree of desulfurization.

Roasting Plants. Any standard roasting furnace (multiple-hearth or fluidized-bed furnace, sintering apparatus or cyclone) may be used for roasting arsenical raw materials. Multiple-hearth furnaces and fluidized bed reactors are the most common aggregates for industrial dearsenifying roasting processes. Details of the roasting technology and a review of industrial applications are provided in [8], [22], [24].

Multiple-Hearth Furnaces. Where roasting is carried out in a multiple-hearth furnace (e.g., a Wedge furnace), the atmosphere in the material bed during the first part of the roasting process should be sufficiently low in oxygen to prevent formation of arsenates. In this region most of the arsenic compounds are released from the material being roasted and are then oxidized in the atmosphere richer in oxygen over the bed of material being roasted before dust separation and cooling. Before it is discharged, the roasted material encounters the oxygen-rich fresh gas through the quantity of which the sulfur content can be adjusted to the required level. On account of the long residence time, mass diffusion is not a problem, even in relatively large particles. However, the maximum temperatures have to be limited to avoid melting and contamination. Multiple-hearth furnaces operate under reduced pressure. The flue dust content of the off-gases is relatively low.

Fluidized-Bed Furnaces. Because of their high roasting capacity, low maintenance costs, and the high content of SO_2 in the off-gas, fluidized-bed roasters are increasingly being used for the roasting of sulfidic raw materials. In applications with the aim of a partial roasting procedure, e.g. for the subsequent matte smelting in the primary copper industry, multiple-hearth furnaces are still in use to about the same extent as fluidized-bed furnaces. Especially for the dead roasting of arsenical materials, where the major aim is to remove sulfur rather than arsenic, fluidized bed roasters are being increasingly preferred to multiple-hearth furnaces. Table 3 gives some details on industrial fluidized-bed roasting plants.

In fluidized-bed roasting, temperature and gas composition are easily kept constant at a level favoring separation of arsenic and subsequent complete desulfurization. The composition of the mineral, the arsenic content, and the grain size of the raw material are critical factors affecting the efficiency of arsenic separation during roasting. The roasted material is completely or partly transported with the off-gases to the gas-cleaning system, which makes gas cleaning difficult. During cooling of the gas, the arsenic trioxide can react with the metal oxides, so that the roasted material must be mechanically separated at elevated temperature. Even after this has been done, the gas still has a high flue dust content, which affects the separation of As_2O_3 and necessitates the installation of equipment for selective gas cleaning if the highest possible As_2O_3 content is to be obtained in the product.

Off-Gas Cleaning (Separation of Arsenic from Arsenic-Bearing Gases). The gas-cleaning system must provide the possibility to collect the

Table 3. Operating parameters for two-stage industrial roasting plants for arsenic containing gold ores [24]

Paramater	Giant Yellowknife	Cambell Red Lake	New Resort Mine
Feed rate, t/h	5.6	2.4	1.4
Feed condition	pulp, 76 % solids	pulp, 80 % solids	pulp, 70 % solids
Feed composition:			
Fe, %	23.0	23.0	21.8
As, %	7.6	9.0	6.75
S, %	17.5	0.5	6.6
Sb, %	0.18	0.5	
Au, tr oz/t [a]	1.6	5 – 10	6.6
Calcine production, t/h	4.2	1.7	1.1
Calcine composition:			
Fe, %	27.8	26.8	27.8
As, %	1.0	1.6	ca. 1
S, %	2.5	1.0	0.7
Sb, %	0.13		
Au, tr oz/t [a]	2.0	6.2 – 12.3	8.9
Operating temperature, 1st stage, °C	496	540 – 565	600 – 700
Roaster diameter, m			
1st stage	3.96	3.35	1.95
2nd stage	2.9	2.65	1.0
Fluidized-bed height, m	1.5	1.5	1
Air Inlet, m^3/h (STP)			
1st stage	4200	3000	2000
2nd stage	1540	1560	160
S removal, %	88	96	98
As removal,%	92	87	88
Calcine/feed ratio	0.77	0.7	0.74
Bed superficial velocity, m/s	0.27	0.28	0.6 – 0.66
Roaster loading coefficient, tph/m^2 [b]	0.45	0.27	0.46
Au recovery from calcine, %	94	96	y

[a] 1 troz (Troyounce) = 31.1 g; [b] tph/m^2 = specific throughput.

dust carry-over and the volatized arsenic separately. Because of the aim to gain a high concentration of arsenic in the product and the tendency of arsenic to react with with metal oxides under oxidizing conditions in the off-gas stream to form metal arsenates, off-gas cleaning is a critical unit operation in dearsenifying roasting processes.

The waste gases from the roasting reactors may first be burnt. They are then freed from dust in, for instance, a hot cyclone and a hot electrostatic precipitator above the condensation temperature of As_2O_3. To achieve sufficiently low levels of arsenic (1 mg/m^3) for use of the SO_2 in a sulfuric acid plant and to prevent the formation of metal arsenates, the oxygen potential should be as low and the temperature as high as possible in the primary gas cleaning step. The gases are then subsequently cooled to about 100 °C . Relatively pure As_2O_3 is deposited in bag filters.

Wet cleaning systems such as wet scrubbers, ionizing washers, and wet electrostatic precipitators may also be applied. In the wet gas cleaning systems, arsenic is washed out and eventually dissolves in the acidic solution [1], [8].

Arsenopyrite. For the roasting of arsenopyrite with the aim of arsenic recovery the roasting plant which was run by Boliden AB at Rönnskär is a good example, although production ceased in 1991 [10]. The multiple-bed furnace contained 9 – 11 hearths with a total hearth area of 230 m^2. The maximum roasting temperature was 500 – 800 °C . Most of the arsenic was removed in the upper hearths. Arsenic and arsenic sulfides were then allowed to burn in the roasting gas. After passing through a cyclone and a flue-dust filter, the roasted material was cooled in a metal air cooler to below the condensation temperature of As_2O_3 and purified in a hot precipitator with an exit temperature of 120 – 140 °C . Sulfuric acid was recovered from the gas containing 3 – 5 % SO_2 in a fine purification plant (washing tower or wet electrostatic precipitator). In the roasting of concentrates containing 12 %

As, the crude oxide that accumulated in the air cooler and the hot precipitator had an As_2O_3 content of 80 – 95 %. The arsenic content was lowered from 11 % in the arsenic concentrate to 1.37 % in the calcine [1], [25].

At Salsigne, France, a multiple-hearth furnace with 12 hearths and an inner diameter of 4 m was used to reduce the arsenic content from 32 % in the concentrate to 0.5 % in the calcine [15].

In the former Soviet Union an arsenopyrite/pyrites concentrate from Darunskii containing 9 – 13 % As and 32 – 40 % S is partially roasted at ca. 575 °C in a fluidized-bed furnace. After the roasting gas has been cleaned in a hot cyclone and a hot electrostatic precipitator (400 °C), the oxide is separated in a condensation chamber. The roasted material contains ca. 0.5 % As and 12 % S; the oxide has a purity of 92 – 97 % [1].

Arsenic-Containing Gold Concentrates. Refractory gold ores are characterized by poor metal extraction in direct cyanidation. Their refractory nature may be due to the presence of carbon and/or locking of gold and silver in arsenical and other sulfide minerals. Sulfidic arseniferrous gold ores exist in many parts of the world: in Canada (Giant Yellowknife, Campbell Red Lake), in the former Soviet Union (Darasun), in South Africa (Barberton Area in Eastern Transvaal, New Consort Mine), France (Salsigne), Greece (Olympias), Brazil (Sao Bento). Roasting of refractory gold concentrates has the objective to render the gold amenable to leaching. The roasting process should achieve:

1) The decomposition of arsenopyryite or enargite to produce hematite or copper sulfide, respectively, and fixation of arsenic as arsenic trioxide and/or iron arsenate
2) The fixation of sulfur values as sulfuric acid or as gypsum by scrubbing the sulfur dioxide and other gases

With arsenopyrite ores and concentrates, two-stage roasting in fluidized bed roasters has been practised on a commercial scale at Campbell Red Lake, Giant Yellow Knife, and New Consort Mine. During the first-stage roasting at 500 °C, arsenic elimination of ca. 65 – 70 % and pyrite conversions of 20 – 30 % are common. Arsenic is collected as arsenic trioxide. Air ratios are 60 – 80 % of the stoichiometric requirement. The second stage is carried out at slightly higher temperature of 530 °C with about 10 % excess air. Arsenic elimination in the final calcine is in the 90 – 95 % range, and sulfur removal in the 85 – 90 % (Table 14).

A number of different roasting technologies vary this two-stage scheme by fixing a large part of the arsenic as iron arsenate instead of As_2O_3 (METBA process); by applying a multiple-hearth furnace for the first stage (Mechim process); by roasting at higher temperatures (800 – 850 °C) in the first stage and thus oxidizing the pyrite to pyrrhotite or magnetite, which is further oxidized to hematite in the second stage (Boliden process and BASF process); and by extending the roasting process to a third stage (Montedison process) [8], [15], [23], [24], [26], [27], [28].

Arsenic-Containing Copper Concentrates. Copper smelters are the biggest producers of arsenic-containing intermediate products. The arsenic content of copper concentrates varies over a wide range, in many cases reaching 1 – 10 %. Pretreatment of copper concentrates by partial roasting to remove arsenic and adjust the sulfur content has long been applied commercially. Nowadays it has been replaced in many cases by the Outokumpu flash smelting furnace, in which the concentrate is roasted and the matte smelted at the same time (see Section 1.6). Usually the copper concentrate is roasted in multiple-hearth furnaces, and the process is controlled such that optimal conditions for removal of arsenic prevail in the upper hearths, whereas in the lower hearths the sulfur content of the material being roasted is reduced to the level required for the subsequent smelting of low-grade matte (e.g., 15 – 20 %).

The processes and plants used are much like those used for the roasting of arsenopyrite (see above). Roasting is exothermic, and the maximum roasting temperature can reach 500 – 800 °C, depending on the composition of the batch. The yield of arsenic also depends on the composition of the batch, the type and quantity of the arsenic minerals, the roasting temperature, and other factors. For high arsenic contents or when arsenopyrite or enargite are present, the yield can be more than 80 %. For lower arsenic contents the yield can be much lower because a

few tenths of a percent of arsenic always remain in the roasted material.

After the hot roasting gas has been cleaned in cyclones and dust-settling chambers, relatively small quantities of a crude arsenic heavily contaminated by flue dust and volatile substances are obtained by cooling the off-gas. The As_2O_3 content of the product is generally between 10 and 50 %. However, far higher contents can be obtained by using electric precipitators to remove dust from the hot gas. The dust left after cleaning of the hot gas is returned to the process.

The cleaned SO_2-containing gas still contains traces of arsenic and other impurities that have to be removed by washing with water or a circulating washing liquid containing sulfuric acid and by final cleaning in a wet electric precipitator. The sulfur is used to produce sulfuric acid. The arsenic-containing slurries that accumulate during wet cleaning and effluent treatment are collected and neutralized to recover valuable materials.

Examples for the partial roasting of copper concentrates are St. Joe's El Indio Mine, Chile [29], [30] (multiple-hearth roasting) and the roasting of Lepanto concentrates at Pasar, Philippines (fluidized-bed roasting).

Arsenic-Containing Pyrites Concentrates. Pyrites are roasted to generate sulfur dioxide for the production of sulfuric acid (→ Sulfur Dioxide, Chap. 5.). The roating residue, known as pyrite cinders, contains most of the impurities which were initially present in the pyrites. The composition of pyrites can vary over a wide range: Fe 42 – 49 %, S 45 – 52 %, As 0.05 – 0.4 %, Sb 0.01 – 0.05 %, Zn 0.2 – 4 %, Pb 0.1 – 1.5 %, Cu 0.04 – 1.5 %, Ag 10 – 60 ppm, Au 0.5 – 2 ppm.

The production of sulfuric acid implies the subsequent treatment of the pyrite cinders, since their impoundment is becoming more and more problematical. To convert pyrite cinders to iron ore the nonferrous metals must be removed to a large extent. The maximum levels of impurities tolerated in iron ore are S $\leq$ 0.5 %, As $\leq$ 0.05 %, Cu $\leq$ 0.1 %, Zn $\leq$ 0.1 %, Pb $\leq$ 0.02 % [8].

During cinders treatment there are no commercial facilities for the elimination of arsenic. It therefore has to be separated from the iron in the primary pyrite roasting stage. The roasting process must be conducted in such a way that as much arsenic as possible is removed. In addition the gas must be freed from arsenic completely before it is delivered to the sulfuric acid factory or otherwise used. Besides arsenic compounds, other constituents, such as selenium or mercury may be obtained at the same time.

The traditional process for the dearsenicating roasting of pyrites is multiple-bed roasting controlled so that the arsenic is released in the low-oxygen atmosphere of the upper beds, whereas the sulfur is released in the oxygen-rich atmosphere of the lower beds. The gases formed have a relatively low SO_2 content and such a high oxygen content that arsenic, sulfur, and the arsenic sulfides are completely oxidized. The As_2O_3 formed reacts further with Fe_2O_3 and oxygen to form iron arsenate. Therefore, most of the arsenic in the dust is in bound form and, in addition, because of the low total content, is an unsuitable starting material for arsenic. The multiple-bed roasting of pyrites has largely been replaced by fluidized-bed roasting.

The standard process for roasting pyrites in a fluidized bed forms Fe_2O_3 with excess air. This is unsuitable for roasting arsenic-bearing ores because the arsenic then is bound as iron arsenate in the roasted material. Two fundamentally different processes have been developed for the fluidized-bed roasting of arsenic-bearing pyrites: the two-stage BASF process and the single-stage Boliden process.

In the two-stage process developed by BASF, about one-half of the sulfur present is roasted off in the first reactor at around 900 °C at a low oxygen partial pressure. After solids and gas have been separated in a cyclone, roasting is completed in a second reactor around 800 °C in excess air. Air is added to the gas stream of the first reactor to burn the sulfides and other volatile substances still present. The gas is then cooled in a waste-heat boiler and cleaned in a hot precipitator and by washing. Some of the arsenic is present as iron arsenate in the dust precipitated, and some is washed out as As_2O_3 during wet cleaning [26], [31].

In the single-stage fluidized-bed roasting process developed by Boliden AB at Reymersholm, the pyrites are roasted in such a way that magnetite, Fe_3O_4, is formed from the iron in the ore at a low oxygen partial pressure. In contrast to Fe_2O_3, magnetite does not react with the As_2O_3 to form arsenate. A small amount of

iron arsenate is separated with the flue dust. The gas freed from the flue dust is mixed with air to burn the dissociation products. Then it is cooled in a waste-heat boiler, freed from dust in a hot precipitator, and subjected to wet gas cleaning which separates most of the arsenic as As_2O_3 [32], [33].

In the wet cleaning of the roasting gases, the gases are washed with water or circulating sulfuric acid solutions in washing towers, Venturi scrubbers, or other apparatus and finally are cleaned in wet precipitators. Separation of the arsenic by the wash liquid is almost complete, although neutralization, utilization, or storage of the arsenic removed presents problems. Normally the process water is circulated so that a high concentration of dissolved substances builds up; a side stream is purified. For example, the arsenic may be precipitated with lime, possibly after neutralization with alkalis. Because the wash liquids are strongly acidic and contain many impurities, the consumption of chemicals is high and the product impure. Selective precipitation may be carried out. At the Reymersholm works of Boliden AB, the As_2O_3 is continuously recovered from the circulating sulfuric acid wash solution by crystallization. The product obtained serves as raw material for the production of refined As_2O_3.

Other Arsenic Containing Concentrates and Materials. Besides the above arsenic-containing concentrates and ores, dearsenicating roasting is applied to a wide variety of nonferrous metal concentrates such as complex tin concentrates, cobalt arsenides, and lead or zinc concentrates. Roasting can also be used for the separation of arsenic from intermediate products from various smelting processes, e.g., tin-bearing speiss which was generated in the blast furnace smelting of Pb/Cu primary material at MHO [8], [34].

1.5.1.2.2. Alternative Pretreatments of Arsenical Materials

Apart from the well-known, classical roasting processes for the dearsenification of ores, concentrates, and arsenic-containing intermediate products, recently the focus has shifted more to hydrometallurgical alternatives. One possible reason is the lower number of product phases compared to pyrometallurgical operations and therefore better selectivity. Furthermore, the arsenic-containing products of most hydrometallurgical operations are relatively stable arsenic compounds, such as ferric arsenates, which are suitable for dumping. Given the rapidly shrinking market for As_2O_3 and the limited applications of arsenic metal, the separation of arsenic in combination with its stabilization is of increasing economical interest in many industrial applications.

Hydrometallurgical pretreatment technology is quite common in the recovery of gold from refractory gold ores. For copper concentrates, two processes based on sodium sulfide leaching have been described, one of which was operated at Equity Silver Mines. Their economical and industrial importance seems to be minor.

Hydrometallurgical Pretreatment of Refractory Gold Ores. Hydrometallurgical pretreatment of refractory gold ores or concentrates consists of an oxidative treatment to oxidize the sulfides, mainly pyrites and arsenopyrites, to sulfates. Thus, the formerly enclosed gold is made amenable to subsequent leaching.

Oxidative Pressure Leaching. The oxidative pressure leaching of refractory gold ores operates at 160 – 180 °C, a pressure of 1300 kPa and 40 – 45 % solids in the feed pulp (Fig. 2). The required retention time is 1.5 – 2 h and oxygen consumption 0.1 t per tonne of ore. The final acidity is 15 – 20 g/L H_2SO_4. More than 95 % of the sulfur is oxidised to sulfate, and iron and arsenic are precipitated according to the following simplified reaction equations:

$$FeS_2 + 15/4\,O_2 + 2H_2O \longrightarrow 1/2\,Fe_2O_3 + 2\,H_2SO_4$$

$$FeAsS + 11/4\,O_2 + 1/2\,H_2O \longrightarrow FeAsO_4 + 1/2\,H_2SO_4$$

The oxidative pressure leach may be preceded by a pre-leaching stage. After oxidation and subsequent solid – liquid separation, the gold can be recovered from the solids by cyanidation, while the solution is neutralized with limestone and the precipitate is dumped as tailings.

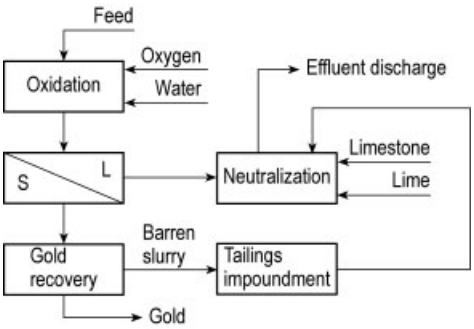

Figure 2. Flowsheet for oxidative pressure leaching of refractory gold feedstocks [35]

In Canada miniplant campaigns have been performed at the Sherrit Research Centre on various arseniferrous gold ores such as Campbell Red Lake, Giant Yellowknife, and Olympias, and have been translated to commercial plants at Key Lake and Sao Bento [8], [35].

A silver-rich copper concentrate containing 17 % Cu, 7000 g/t Ag, 19 g/t Au, 8 % Sb and 3 % As is treated as illustrated in Figure 3. At the heart of the process are two autoclaves for the removal of Sb and As. The arsenic is converted by treatment with CaO into insoluble Ca_3AsO_4. The production of calcium arsenate requires O_2 pressures of 1.5 MPa and a temperature of 190 °C . The calcium arsenate crystals are centrifuged off and dehydrated in a filter press. Production capacity is 600 t/a of calcium arsenate [36].

Biochemical Pretreatment. For the selective biochemical oxidation of gold ores the most widely used bacterium is *Thiobacillus ferrooxidans*, which oxidizes reduced sulfur compounds, such as mineral sulfides, and reduced iron to obtain energy for growth and other life processes. These bacteria, which thrive in acidic conditions, break down sulfide minerals, release metal into solution and, if pyrite is present, increase the acidity of the solution. For refractory ores, oxidation of the gold-bearing sulfides liberates the gold from the mineral matrix and makes it available for recovery (see also → Liquid – Solid Extraction, Chap. 3.3.3.).

Among many chemical reactions involved in the action of *T. ferrooxidans*, the reactions of pyrite are summarized in the following:

$$4\,FeS_2 + 14\,O_2 + 4\,H_2O \longrightarrow 4\,FeSO_4 + 4\,H_2SO_4$$

$$4\,FeSO_4 + O_2 + 2\,H_2SO_4 \longrightarrow 2\,Fe_2(SO_4)_3 + 2\,H_2O$$

$$FeS_2 + Fe_2(SO_4)_3 \longrightarrow 3\,FeSO_4 + 2\,S$$

$$2\,S + 2\,H_2O + 3\,O_2 \longrightarrow 2\,H_2SO_4$$

The oxidation of arsenopyrite can be represented by the following reaction in which pentavalent arsenic (arsenic acid) is formed:

$$4\,FeAsS + 13\,O_2 + 6\,H_2O \longrightarrow 4\,H_3AsO_4 + 4\,FeSO_4$$

Trivalent arsenic (arsenous acid) may be formed as an intermediate product, followed by oxidation of trivalent arsenic to the pentavalent form by iron(III). In the presence of iron(III) in solution, ferric arsenate may be formed from arsenic acid and precipitate.

A process for treating a refractory arsenical gold concentrate was operated on a commercial scale (2 t/d) in a pilot plant at Equity Silver Mines, British Columbia in 1984/1985 with recovery of gold in a cyanide circuit. Small commercial plants are reportedly in operation in South Africa and Zimbabwe [37].

Direct Hydrometallurgical Treatment of Arsenical Concentrates. A number of processes for the direct hydrometallurgical treatment of other concentrates and ores have been developed. These include Co – As concentrates as well as complex Cu – Co – As concentrates and Ni – As concentrates, which are treated by pressure oxidation or atmospheric leaching with HNO_3 oxidation, respectively. The Nitrox process was developed for the hydrometallurgical treatment of refractory gold concentrates and ores. It consists of a direct oxidation with nitric acid at atmospheric pressure and a temperature of 85 °C [8], [38–40].

1.5.1.3. Production of Refined As_2O_3

Preliminary Concentration [1]**.** The starting material for the industrial production of refined As_2O_3 is the arsenic-rich dust or slurry that accumulates as an intermediate product in the roasting and smelting of arsenic-bearing ores and concentrates. The bulk of this arsenic raw material comes from copper smelters, although some also comes from lead, cobalt, and

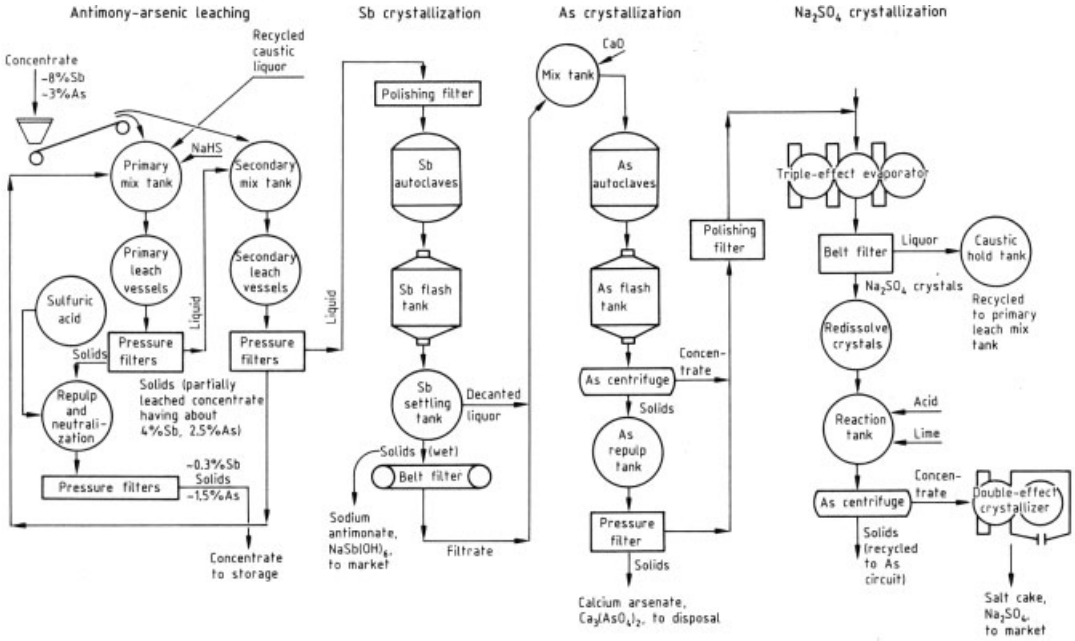

Figure 3. Simplified schematic flow sheet of leach plant at Equity Silver Mines, British Columbia

other smelters. Occasionally, in the former Soviet Union and possibly China, arsenic-rich ore is roasted with the primary object of recovering arsenic. All these raw materials of various origins and arsenic contents have in common that they are obtained as products of gas cleaning and also contain other condensation products and flue dusts. Arsenic is mainly present as As_2O_3, although some may also be bound as arsenite or arsenate. The arsenic content can vary enormously, depending on the ore used and the way in which roasting and gas cleaning are carried out. The gas-cleaning products often contain other valuable constituents and, because of this, must be worked up even if the arsenic content is very low. In such cases the materials are circulated until the arsenic content exceeds 10 %. Low-grade gas-cleaning products of low As_2O_3 content from copper and lead roasting plants contain flue dust from the furnace charge as their principal impurity, whereas the high-grade products mainly contain volatile impurities, such as lead, antimony, bismuth, and selenium. Most of the impurities are present as oxide or sulfate.

Low-grade starting materials are treated in a multistage process. In the first stage the material is roasted to produce ≥ 80 % As_2O_3. This is sublimed and separated from the roasting gases. The bulk of the impurities remains in the roasted material, which is returned to the original process or treated separately to recover valuable secondary products. To increase the yield of arsenic and to facilitate the movement of material through the roasting furnace, sulfides or charcoal are added as reducing agents to the furnace charge. The reducing atmosphere thus established in the material decomposes the metal arsenates and no more arsenate is formed. For example, dust from the gas-cleaning process is mixed with a few percent of iron pyrites concentrate, and the resulting mixture heated at 300 – 500 °C in oil-fired multiple-hearth furnaces. The arsenic-containing gas is cooled in a long brick pipe, passes through a flue-dust chamber, and then through several chambers in which the temperature falls from 220 °C to below 100 °C to separate the arsenic. Finally the gas is cleaned in bag filters. In this system the flue dust and gaseous impurities of low vapor pressure are separated preferably in the first part of the system, so that the gas-cleaning process is fractionated. In a critical temperature range above 200 °C, a relatively small quantity of impure, amorphous glassy ar-

senic trioxide accumulates. However, most of the arsenic trioxide obtained is crystalline. A product containing the high vapor pressure impurities is separated in the bag filter.

Some of the crude arsenic from the separating chambers containing more than 95 % As_2O_3 is sold, but most of it is refined further. The less pure products are returned to the process.

Refining. High-purity crude oxide, obtained either directly or via concentration processes, can be dry refined by sublimation or wet refined by dissolution and crystallization.

The widely used sublimation process is carried out by heating the crude arsenic in a reverberatory furnace. The gases are passed through a dust-settling chamber to a ca. 70-m-long system of, e.g., 39 brick arsenic separation chambers called kitchens. The kitchens are followed by a bag filter. The temperature in the dust-settling chamber is kept at ca. 295 °C, which is considerably above the condensation temperature of arsenic trioxide. The temperature of the bag filter is kept at 90 – 100 °C. A black, amorphous arsenic trioxide containing ca. 95 % As_2O_3 accumulates in the first kitchens. In the following kitchens, in the temperature range from 180 to 120 °C, most of the arsenic trioxide accumulates in a crystalline form analyzing more than 99 %. The product separated in the last kitchens below 120 °C and in the bag filters assays ca. 90 % As_2O_3. The low-purity products are returned to the process or, if they assay more than 95 % As_2O_3, are sold as crude arsenic. The high-purity products from the middle kitchens are marketed as white arsenic.

Several versions of the dry refining process, e.g., with repeated sublimation, have been described in the literature. The Rönnskär works have developed a process for the production of nondusting white arsenic by shock cooling of hot-cleaned sublimation gas [41].

The wet refining of crude oxide is based on the large temperature coefficient of the solubility of As_2O_3 in water (12.1 g/L at 0 °C, 18.1 g/L at 20 °C, 81.8 g/L at 98.5 °C), on the low solubility of the impurities, and on the the purification accompanying crystallization of the As_2O_3.

In the wet refining process, introduced at Rönnskär works as early as 1930, the crude oxide, assaying from 80 to 90 % As_2O_3, is brought in a closed system to the refining plant and pressure leached in steam-heated autoclaves with water or circulating solution. The As_2O_3 dissolves, whereas the impurities, such as oxides, sulfides, sulfates, and arsenates of lead, bismuth, antimony, iron, and other metals, as well as silica, selenium, and other substances, form a slightly soluble sludge, which is separated from the hot leaching solution. The solution is vacuum cooled, and the crystallization is controlled to give a relatively coarse product, which is separated by centrifuging from the solution, washed, dried, and packaged in a closed system. The mother liquor is recycled. The process is fully continuous and also satisfies stringent industrial hygiene requirements. The wet-refined arsenic contains more than 99 % As_2O_3 and is marketed as white arsenic, either in metal casks or in lined wooden casks.

At Nerco Con Mine in Canada, dumping ponds containing over 70 000 t of sludge from off-gas cleaning of the former roasting operation had to be treated for environmental reasons. The sludge contains 50 % arsenic trioxide, 0.70 tr oz/t gold and 1.20 tr oz/t silver. The material is extremely difficult to handle. It is toxic, very fine, with an average particle size of 85 % $< 5\,\mu m$, and forms a thixotropic mass containing 35 % moisture. The arsenic is extracted in a two-stage countercurrent leach at 95 °C. In the primary stage, 90 % of the available arsenic trioxide is extracted, and the remainder in the secondary stage. Key process developments are the addition of small amounts of hydrogen peroxide to enhance the rate of leaching, and clarification by thickening and filtration through powdered activated carbon. The hot pregnant liquor is subjected to four stages of evaporative cooling and crystallization. The plant at Nerco Con Mine was shut down in 1985 and reopened in 1987 with a newly developed carbon filtration system [42].

1.5.2. Production of the Metal

Arsenic metalis produced on a fairly wide scale as commercial arsenic metal, primarily as an additive for alloys, and to a lesser extent as high-purity arsenic for electronic applications.

Commercial-Grade Arsenic Metal [1]. The starting material for the production of commercial-grade arsenic metal is usually white

arsenic. However, the metal can also be produced from arsenopyrite or löllingite.

In the processing of As_2O_3 the oxide is normally reduced with carbon:

$$2\,As_2O_3 + 3\,C \longrightarrow As_4 + 3\,CO_2$$

The reaction is endothermic and is carried out at 500 – 800 °C. The elemental arsenic sublimes and is condensed out of the reaction gas by cooling.

There are several versions of the carbon reduction process. In one version a mixture of As_2O_3 with 25 % charcoal is placed in a horizontal steel retort jacketed with firebrick and heated to 650 – 700 °C in a gas-fired furnace. The reaction gas is collected in water-cooled condensers, in which the metal separates as crystals. After about 10 h, the heating is stopped, and, after the furnace has cooled, the metal can be removed from the condensers.

Similar processes are thought to have been used by several producers. For example, at the Rönnskär works of Boliden AB, steel retorts are provided with demountable air-cooled condensers. After the cold retorts have been charged, they are closed and heated to 700 – 800 °C in an electric furnace. The metal is separated in the condenser, while the off-gases are cleaned in a scrubber system. Most of the metal is coarsely crystalline. After size reduction and grading, the arsenic is packaged in metal casks. In the process used at the Reichenstein works in Silesia, As_2O_3 was vaporized by indirect heating in a gas-heated retort furnace, and the vapor was passed through a 3-m-tall incandescent coke layer into a tower reactor in which it was reduced to gaseous arsenic. The metal was separated in large condensers arranged in series.

Carbon monoxide [43] or ammonia [44] may also be used as reducing agents. Metallic arsenic may be obtained from arsenopyrite or löllingite by thermal dissociation:

$$4\,FeAsS \longrightarrow 4\,FeS + As_4$$

$$4\,FeAs_2 \longrightarrow 4\,FeAs + As_4$$

V. G. Rchiladze [45] described the theory of this process in detail and suggested how it may be carried out in practice. Arsenopyrite is placed in steel retorts lined with diabase or basalt and heated electrically in two stages to 700 – 900 °C. This causes the arsenic and a small quantity of sulfur compounds to volatilize. The metal vapor is condensed at 450 – 500 °C in an electrically heated steel condenser. The metal is scraped off the condenser walls and cooled in a closed system. The volatile sulfur compounds are separated from the off-gas at a lower temperature. The purity of the metal depends on the ore, specifically on the content of volatile impurities that condense under the same conditions as the arsenic. Some sulfur content is always present. In an experiment, metal containing more than 99 % As was obtained from arsenopyrite assaying from 25 to 45 % As. The principal impurities were sulfur and bismuth.

Practical experience in the production of metallic arsenic by reduction with carbon has shown that there are many difficulties. The danger to personnel and the risk of explosion are the two most important. In addition, As vapor is highly corrosive, so that the useful life of the plant is shortened significantly [46]. To overcome these problems, attempts are being made to produce metallic arsenic in other ways. Crystallized As can be obtained from $AsCl_3$, As_2S_3, or As_2O_3 by gas-phase reduction with hydrogen. The hydrogen required for reduction is produced in situ by the thermal dissociation of ammonia. The reduction is highly exothermic. This process is said to give an arsenic yield of 93 – 98 %. The off-gas contains 70 % N_2 and 30 % hydrogen. An advantage lies in a certain degree of refining that takes place. Table 4 shows one example of how impurities present in As_2O_3 affect the purity of the metallic arsenic. Figure 4 shows a schematic diagram for one such plant.

Table 4. Refining effect in the reduction of As_2O_3 with H_2

	As_2O_3, ppm	As, ppm
Ag	< 1	< 0.1
Al	7	1
Bi	2	2
Ca	9	3
Cd	< 1	< 1
Co	< 1	< 1
Cu	5	2
Mg	3	< 1
Mn	< 1	< 1
Na	20	<10
Ni	< 1	0.3
Pb	10	0.5
Sb	1000	0.2
Zn	3	3
Cr	< 1	< 1
Fe	20	8
Si	50	10
Sn	< 1	< 0.3
Te	<30	<10

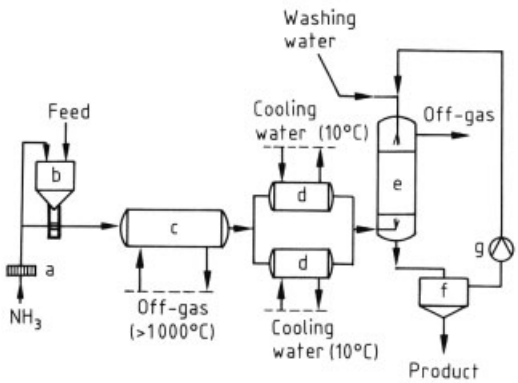

Figure 4. Plant for producing crystalline arsenic metal
a) Roots blower; b) Reservoir; c) Reactor; d) Condensers in parallel; e) Washing tower; f) Decanter; g) Pump

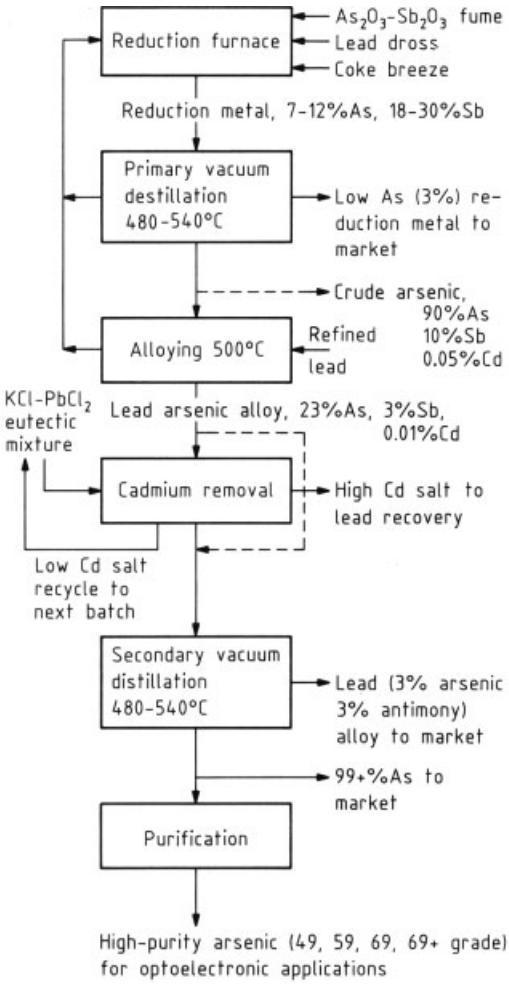

Figure 5. Flow sheet of the Cominco recovery process

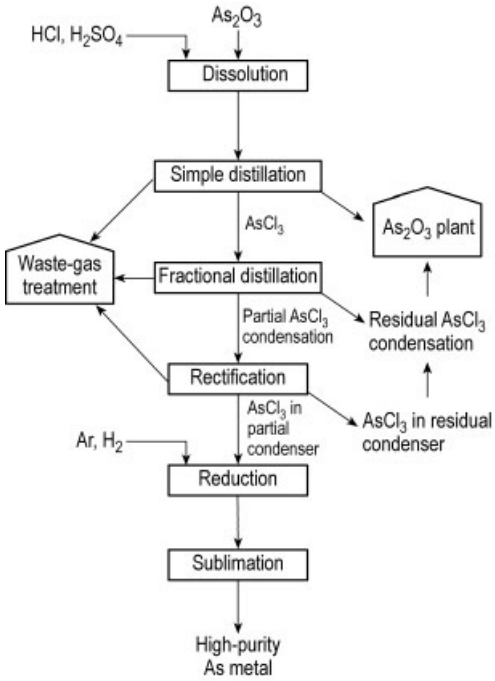

Figure 6. Flow sheet for the production of high-purity arsenic metal at Sumitomo Metal Mining Corporation [51]

Another way of obtaining metallic arsenic is to sublime a Pb – As – Sb alloy to give 99 % pure arsenic metal. Flue dust containing antimony and arsenic and a lead-containing residue from the electrolytic refining of Pb are reduced with carbon to a lead alloy containing 7 – 12 % As and 18 – 30 % Sb. A 90 % As containing 10 % Sb and 0.05 % Cd is produced in a first vacuum distillation stage carried out at 480 – 590 °C (Fig. 5).

A 99 % As suitable for the commercial market is obtained in a second vacuum distillation stage carried out after removal of the cadmium from the residue [47].

Commercial arsenic produced from As_2O_3 contains > 99 % As and as its main impurities antimony and oxygen. This quality is used primarily for the production of alloys.

High-Purity Arsenic. Arsenic is produced in various high-purity grades for special applications. A limited refining effect can be obtained by resubliming commercial metal in a stream of hydrogen. In general, however, various purification stages are combined in the production of high-purity metal.

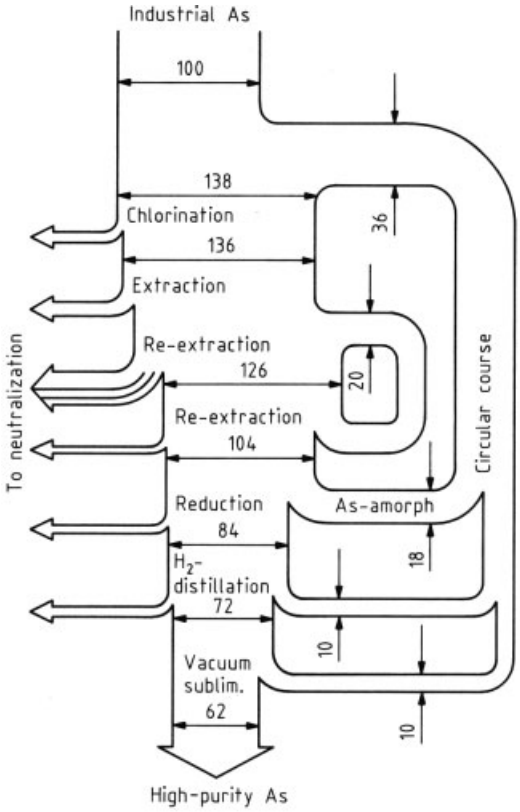

Figure 7. Material flow in the production of semiconductor-purity arsenic by the chloride method

Commercial arsenic is sublimed, reacted with oxygen to produce As_2O_3 which is recrystallized from aqueous hydrochloric acid, sublimed in a stream of hydrogen, and then reduced to the metal with hydrogen at 800 – 850 °C. The arsenic obtained has a purity > 99.999 % [48].

The process most often reported in the literature treats white arsenic with hydrochloric acid to form arsenic trichloride:

$$As_2O_3 + 6\,HCl \longrightarrow 2\,AsCl_3 + 3\,H_2O$$

The arsenic trichloride is purified by fractional distillation, possibly in combination with chemical methods. It subsequently is reduced by reaction with hydrogen at 800 – 850 °C, and the metal vapor is condensed in crystalline form at 400 – 450 °C.

$$4\,AsCl_3 + 6\,H_2 \longrightarrow As_4 + 6\,HCl$$

After sublimation, the metal is packaged in glass ampoules. An arsenic content of > 99.99 % can be obtained by this method.

Figure 6 shows the flow sheet for the production of high-purity arsenic at Sumitomo Metal Mining Corporation, and Figure 7 shows the distribution of arsenic in the process. The yield is 62 %. Table 5 lists the impurities present in the high purity arsenic after the individual process steps [1], [48–50], [51].

1.6. Treatment of Arsenic-Containing Byproducts

Arsenic and products containing arsenic are under an increasing barrage of criticism, because of their toxicity. This has produced a significant drop in demand. Increasing quantities of arsenic containing byproducts are accumulating as the production of heavy metals from complex and so-called dirtier concentrates increase. These products accumulate in the material circuits of the individual plant. Ultimately, arsenic-containing materials have to be removed from these circuits. The processes involved are expensive, and since the demand for arsenic is mostly inadequate, special measures must be taken to stabilize and concentrate these materials for disposal or storage.

Table 6 shows the distribution of arsenic in the different product phases of industrial smelting processes for the recovery of copper, zinc, and lead. Whereas in the copper smelting processes the main arsenic carriers are flue dusts and slags, in lead processing most of the arsenic reports to flue dusts and copper dross. In the hydrometallurgical processes for the recovery of zinc, the arsenic generally remains in the final iron residue (jarosite or goethite), which today still is mostly dumped in ponds.

It is common practice to return arsenic-containing phases to the smelting process several times to increase the arsenic concentration and to minimize the arsenic containing byproducts, which have to be further processed. The actual arsenic elimination from a smelter circuit depends on the fraction being recirculated.

In times of a stable or growing market for As_2O_3 and for arsenic metal the arsenic-containing byproducts from nonferrous metal smelters were a welcome source of these products. Today nonferrous metal smelters are under

Table 5. Content of impurities (wt %) in two samples of pure arsenic ∗

	Sample no.	Cu × 10^6	Al × 10^5	Fe × 10^5	Si × 10^5	Mg × 10^5	S × 10^5
After reduction	1	1	1	0.5	10	4	7
	2	1	0.5	1	5	2	20
After distillation	1	1	0.5	1	1	1	5
in hydrogen	2	1	2	1	2	2	2
After vacuum	1	1	0.5	1	2	1	2
distillation	2	1	1	1	1	1	2

∗ Content of Ni, Bi, Pb, Cr, and Mn in all samples, $<5 \times 10^{-6}$ %; Sb, In, Sn, and Cd, $<5 \times 10^{-5}$ %

Table 6. Distribution of arsenic in the products of industrial smelting processes for the recovery of copper, lead and zinc [8], [20], [52], [53]

Main element recovered	Process	Feed material	As distribution in products [a]	
Cu	flash smelting (e.g., Outokumpu furnace)	Cu concentrate	flue dust:	75 – 85
			slag:	5 – 20
			matte:	rest
Cu	electric furnace smelting	Cu concentrate	flue dust:	10 – 25
			slag:	50 – 70
			matte:	rest
Cu	reverbatory smelting	Cu concentrate (As in feed > 0.2 %)	flue dust:	50 – 70
			slag:	10 – 20
			matte:	rest
Cu	reverbatory smelting	Cu concentrate (As in feed < 0.2 %)	flue dust:	5 – 35
			slag:	15 – 55
			matte:	rest
Cu	Pierce Smith converter	Cu-matte from a reverbatory furnace	flue dust:	75 – 90
			slag:	5 – 25
			blister copper:	rest
Cu	Pierce Smith converter	Cu-matte from a flash smelting furnace	flue dust:	40 – 50
			slag:	20 – 30
			blister copper:	rest
Zn, Pb (via flue dust)	waelzprocess	lead shaft furnace slag, neutral leach residue [b], (EAF dusts [c])	waelz oxide:	10 – 23
			calcine:	77 – 90
Pb, Zn (via flue dust)	ausmelt processes	neutral leach residue [b], lead shaft furnace slag	flue dust:	94.6
			discard slag:	5.4
Zn, Pb	Imperial Smelting furnace	Zn and Pb concentrates, secondary Pb- and Zn-containing materials, e.g. neutral leach residue [b], lead shaft furnace slag	zinc:	5
			lead bullion:	36
			slag:	57
Zn, Pb	Imperial Smelting furnace	Zn and Pb concentrates, secondary Pb- and Zn-containing materials, e.g. neutral leach residue [b], lead shaft furnace slag	zinc:	< 1
			lead bullion:	11
			slag:	17
			seiss:	70
Pb	Pb shaft furnace	Pb concentrates, secondary Pb-containing materials	lead bullion:	68
			slag:	30
Pb	Pb shaft furnace	Pb concentrates, secondary Pb-containing materials	lead bullion:	22.5
			slag:	7.5
			speiss:	70
Pb	Pb refining (pyrometallurgical)	Pb bullion	Cu dross:	> 95 %
			lead:	rest
Zn (via flue dust)	slag fuming process	lead shaft furnace slag	flue dust:	90
			discard slag:	10
Zn (via flue dust)	electrothermic slag reduction	neutral leach residue [b]	flue dust:	80
			discard slag:	20

[a] Percentage of total As in feed material. [b] intermediate product in the hydrometallurgical processes for the recovery of zinc. [c] Electric arc furnace dust.

increasing pressure from environmental legislation and quality requirements to separate the arsenic from their production circuits and to convert it to saleable products or treat it for safe disposal with no economic advantage.

Arsenic in Copper Smelters. The recovery of copper from sufidic concentrates includes the following process steps: copper matte smelting, converting of the copper matte, fire refining, anode casting, and electrolytic refining. During the pyrometallurgical operations, arsenic is mostly distributed between matte and slag or blister copper and slag (Table 3) while smaller quantities remain in the copper and have to be removed from the circuit during electrolytic refining. Whereas flue dusts from the matte smelting operation are often recycled to the furnace for arsenic enrichment, flue dusts from the converting operation are mostly treated for arsenic removal [54].

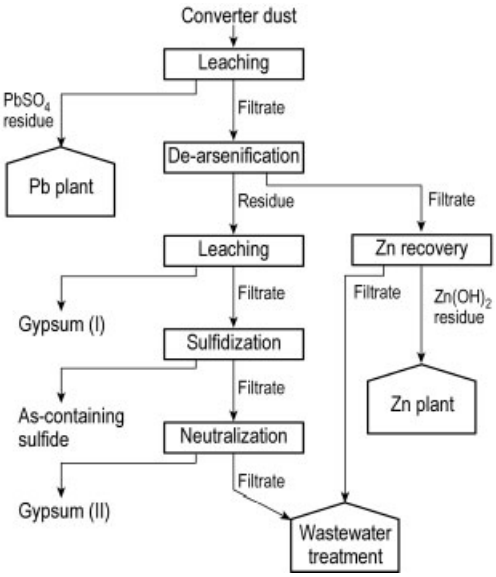

Figure 8. Flow sheet of the converter dust treatment operation at Saganoseki Smelter [56]

At Saganoseki Smelter, Japan the flue dusts from the converting section (5 – 8 % As, 20 – 25 % Pb, 1 – 4 % Cd, 1 – 2 % Fe, 2 – 4 % Bi, 10 – 16 % Zn, 4 – 6 % Cu) are hydrometallurgically converted to a $PbSO_4$ product which also contains Bi and Sb and is sent to a lead smelter, a Zn- and Cd-containing hydroxide product which is sent to a zinc smelter, and arsenic sulfide which is sufficiently stable for disposal.

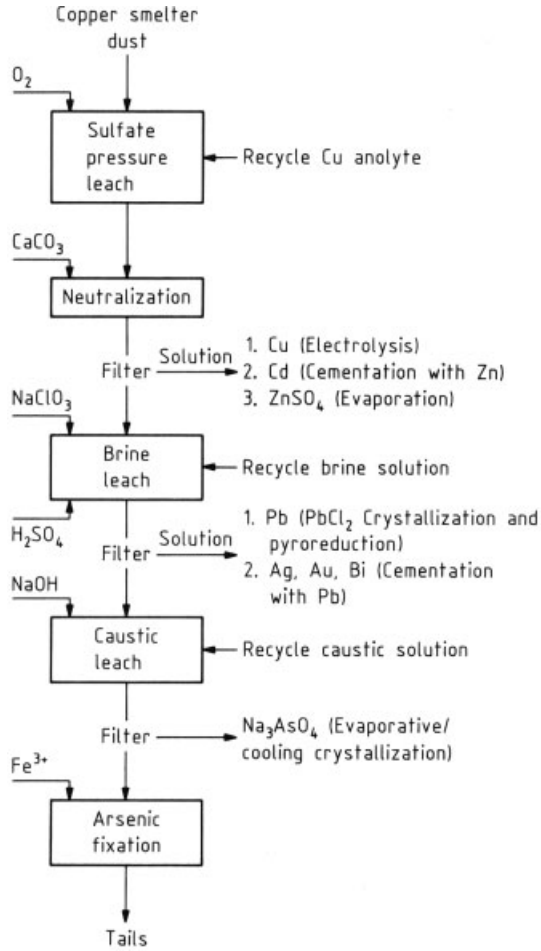

Figure 9. Process for recovering metal values from materials containing arsenic [55]

Figure 8 shows a flow sheet of the process at Saganoseki Smelter; the corresponding analyses of the phases are listed in Table 7. The distributions of the elements in the product phases are given in Table 8. Initially the converter dust is leached with 40 g/L H_2SO_4. Zn, Cu, and As are transferred into solution, while Pb, Bi, and Sb remain almost completely in the $PbSO_4$ residue. For the removal of As from the solution, the pH is raised to 4 and As is oxidized from the trivalent to the pentavalent state. The As/Fe molar ratio is adjusted to 1/1.2 in the solution to support the dearsenification. The remaining elements in the solution are precipitated as hydroxides by neu-

Table 7. Concentrations of the elements (wt %) during the Saganoseki operation for the treatment of converter flue dusts [56]

Phase	As	Cu	Cd	Zn	Bi	Sb	Pb
Converter dust	6.0	5.0	1.3	13.0	2.7	0.2	23.0
$PbSO_4$-residue	2.5	1.2	0.1	1.8	5.0	0.4	44.0
Gypsum (I)	0.3						
$Zn(OH)_2$ residue	2.2	9.9	2.9	28.6	0.1		
As sulfide residue	41.0	2.7	0.3		0.3		
Gypsum (II)	0.3	0.2	0.1	1.0			

Table 8. Percentage distributions of the elements during the Saganoseki operation for the treatment of converter flue dusts [56]

Phase	As	Cu	Cd	Zn	Bi	Sb	Pb
Converter dust	100	100	100	100	100	100	100
$PbSO_4$ residue	22	13	4	7	97	100	100
Gypsum (I)	1						
$Zn(OH)_2$ residue	15	82	92	91	2		
As sulfide residue	61	4	2		1		
Gypsum (II)	1	1	2	2			

tralization (pH 9 – 10) with sodium hydroxide. The As residue is leached again with H_2SO_4 and precipitated as arsenic sulfide by sulfidizing with sodium hydrogensulfide.

Figure 9 shows a multistage process in which Cu, Cd, and Zn are separated off in solution after a neutralization step. Arsenic is precipitated together with lead and the precious metals. Lead and the precious metals are dissolved in a further chloride – sulfate leaching step. During a subsequent alkaline leaching step, the As is dissolved as Na_3AsO_4 and then crystallized out. By adding iron(III) sulfates, the soluble arsenic constituents are precipitated. An iron arsenide residue suitable for disposal is obtained. Tests to determine the solubility of arsenic from such a residue in acidic solution revealed a content of < 0.4 ppm As after 5 d [55].

During the electrorefing of the copper anodes, As accumulates in the electrolyte. To reduce the impurity contents in the electrolyte, a bleed stream is diverted from the circuit and treated to recover Cu, Co, and Ni, usually by electrowinning. This leaves behind a sludge that typically contains 50 % Cu, 20 % As, 5 % Sb, and small amounts of Co, Ni and Bi.

At Sumitomo Metal Mining Corporation in Japan, this sludge is treated together with arsenic sulfide residue from the gas cleaning operation in the converter section. From these raw materials either marketable refined As_2O_3 or arsenic metal is produced, or copper arsenate is precipitated from a leach solution as an alternative product. The sludge from the electrolyte cleaning operation is leached with H_2SO_4, and selective leaching of the arsenic from the arsenic sulfide from the converter gas cleaning facilities is performed with caustic soda. Before neutralization, the two liquids are mixed to adjust the Cu/As ratio to 1.3 to 1.45. Copper and arsenic are precipitated as copper arsenate by neutralization with caustic soda [57].

Especially in Japan, where copper concentrate is mostly imported, the ability to process concentrates with high impurity contents might be an important factor in keeping raw material costs low and maintaining a stable position in the market.

Dearsenification of Lead – Copper Matte. A lead – copper matte containing 25.7 % Cu, 18.2 % Ni, 1.65 % Co, 27.2 % Pb, and 4.14 % As is leached with concentrated H_2SO_4 at 85 – 100 °C in the presence of MnO_2:

$$MS + MnO_2 + 2\,H_2SO_4 \longrightarrow MSO_4 + MnSO_4 + S + 2\,H_2O$$

(M = Cu, Ni, Co, and Pb)

The metals Ni, Co, Fe, and As dissolve, leaving a residue of elemental sulfur and $PbSO_4$. The residue is separated. As is precipitated as $FeAsO_4$ by addition of NaOH or $Ca(OH)_2$ at pH 4. The $FeAsO_4$ takes up 88 – 92 % of the arsenic. It consists of 24 % As, 22 % Fe, and 3 – 5 % Cu and is in a suitable form for safe disposal [58], [59]. As shown in Figure 10, the remaining solution is processed further to remove Co and Cu and recover MnO_2.

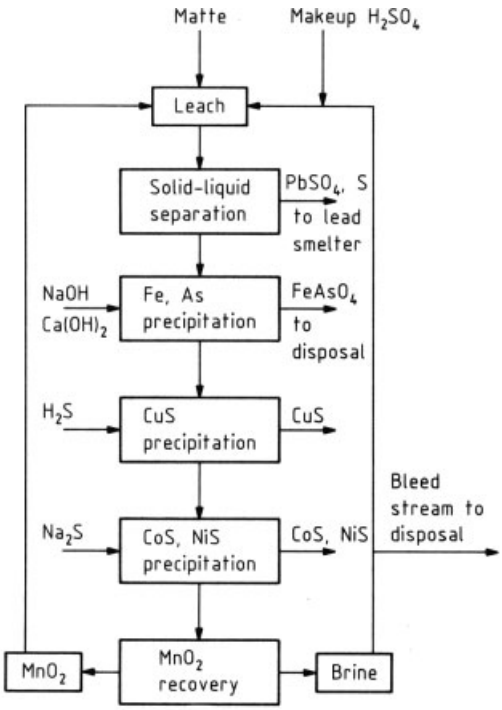

Figure 10. Flow diagram for recovery of cobalt, nickel, and copper from mattes [60]

Lead Smelter Flue Dusts. Arsenic-containing lead flue dusts can be leached with a 20 % H_2SO_4 at 90 °C. More than 90 % of the As, Zn, and Cd pass into solution. The residue rich in lead is returned to the lead production process. After removal of the indium from the solution by solvent extraction, the arsenic is precipitated with $Fe_2(SO_4)_3$ and $CaCO_3$ as an iron-containing arsenic residue suitable for storage in containers or for further treatment. Table 9 shows the typical composition of three different lead smelter flue dusts. Figure 11 shows a flowsheet [61]. Arsenic also can be precipitated with calcined zinc concentrates [59], H_2, Na_2S, or even NaHS. Either As_2S_3 or As_2S_5 is obtained [62].

Separation of Arsenic and Antimony from Flue Dusts. Since many flue dusts occurring in the nonferrous metals industries contain both arsenic and antimony, the question of the separation of these two metals arises if values of the two metals are comparable. Whereas arsenic now only has a small market and disposal is often preferred to a recovery from fumes, antimony still is of a noticeable economic value. Research is in progress to develop techniques for the separation of both metals. One possibility is the injection of oxygen into mixed arsenic/antimony oxide vapors to recover antimony oxide [63].

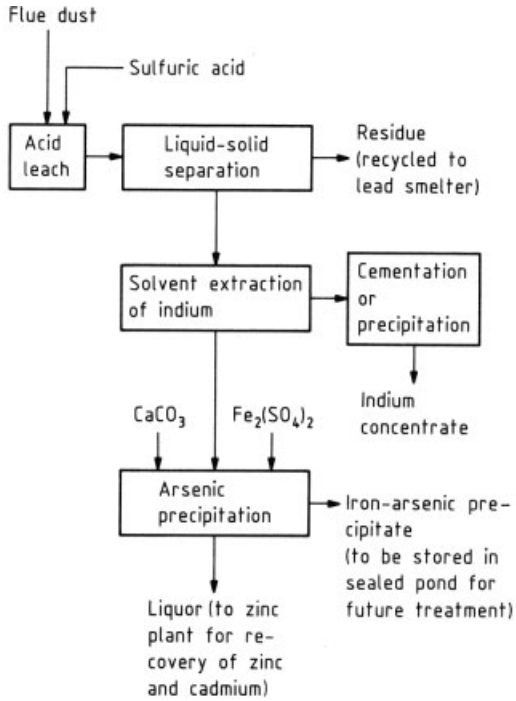

Figure 11. Flow sheet for treatment of lead smelter copper dross flue dust

Roasting of Tin Concentrates. In the roasting of tin ore concentrates in multiple-hearth furnaces, the atmosphere is regulated so that As volatises as As_4. One proposal for avoiding the production of nondisposable products is to adjust the oxygen and sulfur partial pressures so that the As_4 is converted by two additional burners, for example in a tube reactor, at 600 – 900 °C into disposable As_4S_4, which is recovered in the sump of a washer. The product is suitable for disposal or, if desired, may be worked up to give marketable products [64]

2. Arsenic Compounds

The most important use of arsenic compounds is the production of arsenic-containing wood

Table 9. Assays of lead smelter flue dust samples

Sample	Description	Assay, wt %								
		As	Pb	Zn	Cd	Sb	In	Cu	Fe	SO_4^{2-}
1	Composite sample	10.0	41.0	10.3	0.79	3.5	0.26	0.10	0.28	7.0
2	Dry sample	10.0	51.0	16.2	0.22	1.4	0.36	0.16	0.03	6.9
3	High-zinc sample	7.0	45.0	21.0	0.20	2.4	0.55	0.21	0.40	9.8

preservatives (see Chap. 3), although even here consumption is very much on the decline because of environmental pollution and toxicity [65–68].

2.1. Arsenic Trioxide

As_2O_3 [*1327-53-3*], M_r 197.8, is also known as white arsenic. It occurs in two crystalline and one amorphous modification. The octahedral (or cubic) modification, arsenolite, ϱ 3.86 g/cm^3, is stable at room temperature. It is formed on condensation of As_4O_6 vapor; As_4O_6 molecules are the lattice units. At temperatures above 221 °C, it changes into a monoclinic modification, claudetite, ϱ 4.15 g/cm^3. This modification is directly formed when condensation is carried out at temperatures above 221 °C. Condensation above 250 °C results generally in the formation of the amorphous, glassy phase, ϱ 3.70 g/cm^3, which devitrifies into the octahedral modification at room temperature.

In an open vessel, arsenic trioxide sublimes without melting above 135 °C. The monoclinic form melts at 312.3 °C under its own vapor pressure. Up to about 800 °C the vapor density corresponds to the formula As_4O_6; at 1800 °C, dissociation into As_2O_3 molecules is complete. The As_4O_6 molecules were also detected in nitrobenzene solutions.

Arsenic trioxide is moderately soluble in water. The glassy – amorphous As_2O_3 dissolves much more easily than the crystalline form; its solubility is 12 g/L at 2 °C and 114.6 g/L at 100 °C [69]. The solution has a sweetish taste with an unpleasant metallic tang. As little as 0.1 g can be fatal, if As_2O_3 enters the stomach.

For production of the oxide, see Section 1.5.1.

2.2. Arsenous Acid

Arsenous acid [*13464-58-9*] cannot be obtained from aqueous solutions of As_2O_3; instead, As_2O_3 is reprecipitated. However, in aqueous solution it exists in equilibrium with its dissociation products. It is still not clear what formula should be assigned to dissolved arsenous acid. In addition to the formulation as orthoacid, H_3AsO_3 or $As(OH)_3$, which is used here, the dissolved acid may also be formulated as metaacid, $HAsO_2$ or AsO(OH), or as hexahydroxoacid, $H_3[As(OH)_6]$. Arsenous acid is amphoteric, dissociating in two ways:

$$As(OH)_3 \rightleftharpoons AsO(OH)_2^- + H^+$$

$$As(OH)_3 \rightleftharpoons As(OH)_2^+ + OH^-$$

The dissociation constants are 6×10^{-10} and 1×10^{-14}, respectively. Accordingly, arsenous acid is a very weak acid, approximately equal in strength to boric acid, but it is a much weaker base. The As^{3+} ions occur in significant quantities only in very strongly acidic solutions.

Arsenates (III). The salts of arsenous acid are called arsenates(III). Formerly they were called arsenites. Some salts exist in the forms MH_2AsO_3, M_2HAsO_3, and M_3AsO_3, although most salts of arsenous acid are derived from the metaacid, $HAsO_2$, i.e., have the formula $MAsO_2$.

The alkali metal arsenates(III) dissolve easily in water, whereas the alkaline earth metal arsenates(III) are slightly soluble to insoluble. The compounds of arsenous acid with heavy metals are insoluble in water. Arsenous acid can be retained in considerable quantities by slimy metal oxide hydrates, such as iron oxide hydrate by adsorption.

The arsenates(III) are decomposed by strong acids because they are salts of a very weak acid.

Sodium arsenate(III), $NaAsO_2$ [*7631-89-2*], does not exist in crystallized form. Solutions of the salt are obtained by dissolving arsenic trioxide in sodium hydroxide. For manufacture, arsenic trioxide is mixed with caustic soda. After the addition of a little water, a vigorous reaction begins, spreading quickly through the mixture. Soda may also be mixed with the oxide and a little water, and the resulting mixture dried at elevated temperature in a furnace. In both cases, a water-soluble product is obtained after grinding. It is used in the manufacture of insecticides, herbicides, etc.

Other arsenates(III): Many complex metal arsenates(III) are known. They include *Scheele's green*, formerly used as an artist's color. It is a mixture of copper arsenates(III) to which formulas such as $Cu_3(AsO_3)_2$, $CuHAsO_3$, and $Cu(AsO_2)_2$ have been assigned. The double salt of copper arsenate(III) and copper acetate, $Cu(CH_3COO)_2 \cdot 3\,Cu(AsO_2)_2$, is *Schweinfurth green*, another artist's color. This use has declined enormously because of the poisonousness. Calcium and copper arsenates(III) were formerly used as "*copper-arsenic liquor*," to control pests in viticulture.

2.3. Arsenic Pentoxide

Arsenic pentoxide, As_2O_5 [*1303-28-2*], M_r 229.8, is a white glassy mass (ϱ 4.32 g/cm^3), which deliquesces in air to form arsenic acid. On heating, arsenic pentoxide begins to decompose near its melting point, ca. 300 °C. The vapor consists of arsenic trioxide and oxygen. The pentoxide can be prepared by reaction of arsenic trioxide with oxygen under pressure or by dehydration of crystalline arsenic acid at temperatures above 200 °C [70]. The crystal structure was unknown until recently. The monocrystals were grown under oxygen excess pressure: space group $P2_12_12_1$ [71], [72].

2.4. Arsenic Acid

Arsenic acid, H_3AsO_4 [*7778-39-4*], is known in the solid state only as the hemihydrate $H_3AsO_4 \cdot 0.5\,H_2O$ in the form of rhombic, deliquescent crystals. Solutions in water behave as a tribasic acid with the dissociation constants $K_1 = 5.6 \times 10^{-3}$, $K_2 = 1.7 \times 10^{-7}$, and $K_3 = 3 \times 10^{-12}$. Arsenic acid loses water when heated to 120 °C with the formation of pyroarsenic acid, $H_4As_2O_7$. At higher temperatures, more water is lost and metaarsenic acid, $HAsO_3$, forms [70]. If the protons of the acids are completely or partly replaced by metals, the arsenates(V), which generally correspond to the phosphates, are formed. In acidic solution, arsenic acid and its salts are strong oxidizing agents.

Production. Arsenic acid is produced solely by the oxidation of As_2O_3. Various processes for its production based mostly on oxidation with nitric acid, which gives a high yield, are described in the patent literature. Originally the apparatus used for those processes were usually made of stoneware but they are now made of alloys resistant to nitric acid.

The plant usually consists of so-called developers and a regenerator for nitric acid. The developers are ca. 4-m^3 vessels equipped with a high-speed stirrer, a steam coil, and a mechanical feeder for oxide. About 2000 L of 38 % HNO_3 is introduced initially, and the oxide is run in with strong heating until the nitric acid has been consumed. The oxides of nitrogen formed pass from the developer into the regenerator, where they are worked up to nitric acid in the presence of air and water. Up to 90 % of the NO and NO_2 formed are converted back into nitric acid. One developer can process 2000 kg of As_2O_3 into arsenic acid in 24 h.

Oxidation with sodium or calcium chlorate also enables a relatively large quantity of oxide to be oxidized in a short time. The oxide is suspended in water and a little arsenic acid is added to the resulting suspension. An aqueous solution of the chlorate is run in at boiling temperature. The reaction takes place smoothly while chlorate is uniformly run in. The input of chlorine must be reduced towards the end of the reaction because otherwise chlorine readily escapes. The arsenic acid obtained by the chlorate process always contains chloride, which causes problems in the production of lead arsenate.

Arsenic acid is used as a defoliant and as a starting material for important inorganic and organic arsenic compounds.

Arsenates(V). Arsenic acid gives three series of salts. They correspond generally to the phosphates in their compositions and solubilities. Many of them are also isomorphous with the phosphates, e.g., ammonium magnesium arsenate(V), $MgNH_4AsO_4 \cdot 6\,H_2O$, which changes on calcination into magnesium pyroarsenate(V), $Mg_2As_2O_7$, a form suitable for the quantitative determination of arsenic acid. Unlike silver phosphate, which is yellow, the silver salt Ag_3AsO_4 is brown.

Of the numerous salts of arsenic acid, the salts of potassium, sodium, calcium, and lead are important commercially. Even today they are used on a wide scale in pest control, although modern synthetic organic compositions have reduced the use of arsenates. In some countries, the use of arsenic compounds in plant protection agents has been banned. The same applies to metal arsenates(V) as active constituents of arsenic-containing wood preservatives [73], [74].

*Monopotassium arsenate(*V*),* KH_2AsO_4, M_r 180.04, *mp* 288 °C, ϱ 2.867 g/cm^3, was used in the manufacture of insecticides and in the preservation of skins. It is prepared by calcining a mixture of equal parts of arsenic oxide and potassium nitrate in cast-iron cylinders, leaching the melt with water, and concentrating the solution with arsenic acid. It forms white crystal crusts.

*Trisodium arsenate(*V*),* M_r 423.9, ϱ 1.759 g/cm^3, *mp* 86.3 °C, is obtained most easily by neutralizing arsenic acid with sodium carbonate or sodium hydroxide. Concentration by evaporation gives a marketable product having the formula $Na_3AsO_4 \cdot 12\,H_2O$.

Disodium arsenate(V), Na_2HAsO_4, M_r 185.9, ϱ 1.72 g/cm^3, is obtained by roasting a mixture of arsenic oxide with sodium carbonate and sodium nitrate in the presence of air. In addition, this compound can be easily obtained by calcining sodium arsenate(III) in the presence of air. Sodium arsenate(V) is used in the production of other arsenates, such as calcium and lead arsenates.

*Calcium arsenates(*V*).* There are three known calcium salts of arsenic acid, $Ca(H_2AsO_4)_2 \cdot H_2O$, $CaHAsO_4 \cdot H_2O$, and $Ca_3(AsO_4)_2 \cdot 3\,H_2O$. In addition, there are basic salts, for example, $Ca_5OH(AsO_4)_3$, which corresponds to apatite. Commercial preparations are based on the neutral salt. However, different manufacturers use different methods and do not always guarantee the production of a strictly neutral salt. Because of the risk of damage to plants that acid arsenic salts involve, a basic character is given to the commercial product during its production by using excess lime. Elevated temperature, abundant moisture, and atmospheric carbon dioxide cause calcium arsenate to convert into the plant-damaging acid arsenate:

$$Ca_3(AsO_4)_2 + CO_2 + H_2O \longrightarrow 2\,CaHAsO_4 + CaCO_3$$

The reason the acid salt is harmful is its greater solubility in water, 0.31 % As_2O_5 vs. 0.013 % for the neutral salt. The neutral salt begins to turn acid only after the excess lime added has been converted completely into carbonate. For the solubility in the presence of atmospheric carbon dioxide, see [68].

Calcium arsenate(V) is generally prepared by reacting milk of lime with arsenic acid. Maintaining a certain temperature and an experimentally determined feed rate and stirring speed, the arsenic acid can be run as a fine distribution into the milk of lime. The precipitate obtained is filtered, washed, and dried.

Calcium arsenate(V) is also obtained by heating calcium arsenate(III) in a stream of dry air.

Lead arsenate(V), $PbHAsO_4$, M_r 347.12, is a white water-insoluble powder of density ϱ 5.79 g/cm^3. It was marketed as a plant protection agent in the form of powders and pastes. It is produced mainly from lead acetate, lead nitrate, and lead oxide and disodium arsenate. The direct reaction of arsenic acid with litharge in the presence of nitric acid gives a dilead arsenate. Commercial lead arsenates contain mostly $PbHAsO_4$ along with varying amounts of $Pb_3(AsO_4)_2$. Accordingly, their lead content is higher and their As_2O_3 content lower than the levels calculated for $PbHAsO_4$.

Arsenic-containing plant protection agents have been banned in Europe. However, they are still used in East Asia [73], [74].

2.5. Arsenic Sulfides

Table 10 shows some physical data of the arsenic sulfides [69], [75].

Table 10. Physical data of arsenic sulfides

		Color	M_r °C	*mp,* °C	*bp,*
Arsenic disulfide (realgar)	As_4S_4	red	427.9	307	565
Arsenic(III) sulfide (orpiment)	As_4S_6	yellow	492.1	320	707
Arsenic(V) sulfide	As_4S_{10}	yellow	620.3	>95 decomp.	–

Arsenic disulfide, As_4S_4 (realgar, red arsenic glass, red glass, ruby arsenic, arsenic blend), occurs in ruby-red crystals or as an amorphous reddish mass. Its commercial value is determined less by the chemical composition than by its purity and fineness. Industrially produced red arsenic glass varies in its composition. Whereas the red glasses previously contained 70 % As and 30 % S, substantially corresponding to the formula AsS, they now contain around 61 – 64 % As and 39 – 36 % As. The red glass is marketed in finely divided form.

Red glass has long been produced on a large scale from arsenic ores and sulfur ores or from arsenopyrite containing iron pyrites. Another industrial process based on production from arsenous acid and sulfur has been perfected. Prolonged development problems were presented by discolored products that had to be converted into an attractive, marketable red glass. The red glass is produced in two separate steps. A red glass that does not yet have the desired composition and that needs refining in appearance is obtained in the first step by melting and sublimation. The intermediate product is therefore refined and is given the required color.

Production by melting a mixture of elemental arsenic and sulfur gives a red glass particularly fine in appearance but too expensive on account of the high price of the elemental arsenic. The processing of arsenopyrite and sulfur, although possible, is not usual.

Most red glass is used in tanneries as a depilatory in the manufacture of fine leather.

Arsenic(III) sulfide, As_4S_6, occurs as a yellow mineral known as orpiment. It can be produced by precipitation of compounds of trivalent arsenic with hydrogen sulfide. The colloidal solution of the arsenic trisulfide can be flocculated by hydrochloric acid, in which it is insoluble. However, it dissolves readily in basic reagents, above all in alkali-metal sulfide solutions, including ammonium sulfide, to form thioarsenites:

$$As_2S_3 + 3\,S^{2-} \longrightarrow 2\,[AsS_3]^{3-}$$

If yellow ammonium hydrogen sulfide is allowed to act on arsenic trisulfide, the dissolved sulfur promotes oxidation to thioarsenate:

$$As_2S_3 + 3\,S^{2-} + 2\,S \longrightarrow 2\,[AsS_4]^{3-}$$

Arsenic trisulfide is a lemon-yellow compound, which begins to sublime before melting, melts at 310 °C to form a red liquid, and boils at 707 °C without decomposing. It is used as an artist's color, marketed in pure form as king's yellow. The impure product obtained by fusing arsenic trioxide with sulfur is marketed as orpiment. Orpiment always contains unchanged arsenic trioxide and, unlike the pure sulfide, is poisonous.

Orpiment was once used for cosmetic purposes. It is now used in the semiconductor industry, for the production of IR-permeable windows, and as a pigment.

Arsenic(V) sulfide, As_4S_{10}, is produced by fusing stoichiometric quantities of As and S powder or by precipitation from highly acidic arsenate(V) solution with H_2S. At relatively high temperatures As_4S_{10} decomposes into arsenic(III) sulfide and sulfur.

2.6. Arsenic Halides

The known arsenic halides are listed in Table 11 [69], [75], [76].

All arsenic halides can be produced by direct combination of the components. They can also be obtained by halogenation of the oxides or sulfides. The bonds are hydrolyzed by water. Only arsenic(III) chloride is commercially significant.

Table 11. Physical data of arsenic halides

	Form	M_r	*mp*, °C	*bp*, °C
Arsenic(V) fluoride	colorless gas	169.9	−88.7	−58.2
Arsenic(V) chloride		252.2	−50 (decomp.)	–
Arsenic(V) oxychloride	colorless liquid	197.3	−78	–
Arsenic(III) fluoride	colorless liquid	131.9	− 8.5	60.4
Arsenic(III) chloride	colorless liquid	181.3	−13	130.2
Arsenic(III) bromide	yellow solid	314.6	31	221
Arsenic(III) iodide	red solid	455.6	140.7	ca. 400

Arsenic trichloride, $AsCl_3$, is a colorless oily liquid that fumes in air, ϱ 2.17 g/cm^3. By virtue of its low boiling point, $AsCl_3$ is easy to separate from $SbCl_3$ and the chlorides of other metals. $As(OH)_3$ forms during hydrolysis. The direct synthesis of $AsCl_3$ is uneconomical. An old commercial production method is the Glauber process using arsenic trioxide:

$$As_2O_3 + 6\,NaCl + 3\,H_2SO_4 \longrightarrow 2\,AsCl_3 + 3\,Na_2SO_4 + 3\,H_2O$$

However, $AsCl_3$ is easier to produce by reacting As_2O_3 with gaseous hydrogen chloride at 180 – 200 °C.

Arsenic(III) chloride is a starting material for numerous organoarsenic compounds and for the preparation of chlorine derivatives of the arsines. It is used as a dopant in the manufacture of semiconductors and in the production of high-purity arsenic metal (see page 16) [77].

2.7. Arsenic Hydride and Arsenides

Arsenic hydride, AsH_3, arsine, M_r 77.95, *bp* −55 °C, *mp* −116.3 °C, is the only known hydrogen compound of arsenic; it is a colorless gas with an unpleasant, garlic-like odor. It is extremely toxic: in many cases the inhalation of small amounts has caused death. Arsine does not ignite spontaneously but burns with a grayish-blue flame to form arsenic trioxide. If the flame is extinguished, for example, by holding a cold porcelain dish in it, arsenic is deposited as a black patch. The pure gas is stable at normal temperature. If arsenic hydride is passed through a red-hot glass tube, it decomposes into hydrogen and arsenic (arsenic mirror), 185.1 kJ being released per mole of arsenic. The arsenic is deposited just behind the heated area as a bright, metallic black mirror (Marsh test). When treated with hypochlorite solution or with alkaline hydrogen peroxide solution, the patch changes into soluble arsenate. Arsenic hydride has a strong reducing action. It precipitates metallic silver from silver nitrate solution. It reacts with solid silver nitrate to form the yellow compound $AsAg_3 \cdot 3\,AgNO_3$ (Gutzeit arsenic test).

Arsine cannot be prepared by direct synthesis. The most convenient method of production is the reaction between sodium arsenide and water or zinc arsenide and dilute hydrochloric acid:

$$As_2Zn_3 + 6\,HCl \longrightarrow 2\,AsH_3 + 3\,ZnCl_2$$

These reactions are fast and complete and have the advantage that the product is not contaminated with hydrogen. Arsenic hydride can also be produced by cathodic reduction or by reduction of arsenic compounds in acidic solution with zinc or magnesium. The last reaction is not quantitative because elemental arsenic, which does not react, precipitates. The product also has a high hydrogen content. Because arsenic is present as a common impurity in numerous metals suitable for the production of hydrogen, hydrogen produced by acid – metal reduction must be passed through a potassium permanganate solution to destroy any arsenic hydride present. Particular practical significance is attributed to the formation of arsenic hydride in the pickling of metals when arsenic-containing sulfuric acid is used.

Arsenic hydride is so poisonous that a few inhalations of air containing arsenic hydride can result in death. The MAK is 0.2 mg/m^3.

Arsenic hydride is used as a dopant in the semiconductor industry [77].

Arsenides. A number of arsenides occur in nature. Some have a defined composition, others are probably mixtures. Naturally occurring arsenides include Cu_3As (domeykite), $FeAs_2$

(löllingite), $NiAs_2$ (chloanthite), NiAs (niccolite), and $CoAs_2$ (smaltite). There are many minerals in which sulfur is partially replaced by arsenic.

The arsenides themselves have little commercial significance. Those produced are obtained by direct synthesis. The alkali metal arsenides are also obtained by reaction in liquid ammonia. However, in this case polyarsenides can form in addition to the simple arsenides. The arsenides of the alkali metals, alkaline earth metals, zinc, etc., are rapidly decomposed by water and acids with the formation of arsine. They may be regarded as derivatives of arsenic hydride. Most of the heavy metal arsenides, whose composition often does not reflect any simple relationship to arsenic hydride, are intermetallic phases and are highly resistant to acids. Gallium and indium arsenides are important in the semiconductor industry (see Section 3).

2.8. Organoarsenic Compounds

There are several known organoarsenic compounds. Intensive research began with the discovery by P. Ehrlich in 1908 of chemotherapeutic efficacy of organoarsenic compounds against trypanosomes and spirochetes (which cause sleeping sickness and syphilis). Aromatic organoarsenic compounds were successfully used for some time in the treatment of tropical diseases; compounds containing aliphatic organic groups were effective plant protection agents and pesticides. Since then the use of organoarsenic compounds has greatly diminished because of the introduction of antibiotics, such as penicillin, and the increased awareness of the environment [78]. Basic research, with a view to understanding the behavior of arsenic in the environment, has greatly increased over the last decade.

Monographs on organoarsenic compounds can be found in [4], [79–84]. Annual reviews are published in the Journal of Organometallic Chemistry [85], [86]. The recurring theme of these publications is the synthesis of new ring systems and their structural elucidation [87–94] and the refinement of analytical techniques [95–102]. A large number of publications are concerned with the metabolism and toxicity of organoarsenic compounds in organisms [103–112]. For chemotherapeutic uses, see [113], [114]. Today arsenic compounds are no longer used in the treatment of syphilis and are only rarely used in the treatment of amoebiasis, although they are still prescribed for the treatment of trypanosomal diseases (e.g., sleeping sickness), above all during advanced stages of the sickness [114]. Because the agricultural use of organoarsenic compounds as plant protection agents and pesticides always results in residues, the organoarsenic compounds have largely been replaced by metal-free compounds. In Europe agricultural use of arsenic compounds has been banned; in the USA, certain preparations are still accepted in some states, e.g., those of the Ansar series. In Eastern Asia preparations such as Neoasozin are essential rice fungicides. In wood preservatives, arsenic compounds are used only in compound preparations. In the Federal Republic of Germany, France, and Switzerland, their use is restricted or prohibited [73].

Nomenclature. Nomenclature is governed by the provisional IUPAC Rule D-5 [115], which suggests the following names (R = organic residue):

Arsines: the organic derivatives of arsenic hydride with the general formulas $RAsH_2$, R_2AsH, and R_3As (primary, secondary, and tertiary arsines).

Arsonium salts: $(R_4As)^+$, analogous to the ammonium salts.

Arsonous acid: 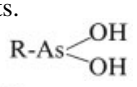

Arsinous acid: 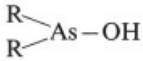

Oxoorganoarsines: compounds of the type $(R{-}As{=}O)_n$. The name "arsenoso" used by analogy to the nitroso compounds is misleading because, corresponding to the oligomeric form of these compounds, there are no analogies to the nitroso compounds [79].

Arseno compounds: the type R−As=As−R would correspond to the azo compounds; however, valence saturation is achieved by polymerization $(R{-}As)_n$ in the case of arsenic. This name has been retained by several authors in the interests of simplicity [79], [88].

Arsonic acid: 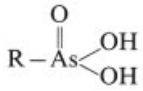

Arsinic acid: 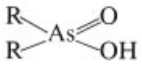

Arsorane: R_5As

Arsa: to be used as a prefix when arsenic is bound as part of a ring. Many compounds of this type are also called by trivial names taken from the *N*-heterocycles [79], [80].

The covalent/ionic bond problem is discussed in [79]. In the Anglo-Saxon literature, some of the nomenclature also follows the Chemical Abstracts rules [116].

2.8.1. Aliphatic Organoarsenic Compounds

Compounds of this class are still used as herbicides and fungicides in rice, cotton, fruit, and coffee plantations, particularly in Eastern Asia:

1) Salts of methanearsonic acid, $CH_3AsO(OH)_2$, especially the iron ammonium salt, Neoasozin, used as a fungicide in rice growing and the sodium, ammonium, and diethanolammonium salts used as herbicides in cotton growing.
2) Dimethylarsinic acid, $(CH_3)_2AsOOH$ cacodylic acid. Ansar 160, the sodium salt, is used as a total herbicide and desiccant. Generally it is produced by reaction of methyl halide with a salt of arsenous acid.
3) Methylarsine sulfide, $CH_3-As{=}S$, Urbasulf, was prepared by reduction of the sodium methylarsonate with SO_2, followed by reaction with H_2S. It is no longer on the market. Urbacid, methylarsine bis(dimethyldithiocarbamate), $CH_3-As[S-CS-N(CH_3)_2]_2$, formerly used in rice and coffee growing, is no longer available commercially. It was prepared from methyldichloroarsine and dimethyldithiocarbamate.

2.8.2. Aromatic Organoarsenic Compounds

These compounds were used for some time as chemotherapeutic agents for treating protozoal infections (→ Chemotherapeutics). Their importance has declined since the introduction of antibiotics. Preparation is based on two classic methods. Aniline is reacted with arsenic acid at 200 °C as in the sulfonation of organic compounds [117]:

$$C_6H_5NH_2 + H_3AsO_4 \longrightarrow H_2N-C_6H_4-AsO(OH)_2$$

Or, diazo compounds are reacted with sodium arsenate(III) [118–120]:

$$C_6H_5N_2Cl + Na_3AsO_3 \longrightarrow C_6H_5-AsO(ONa)_2 + N_2 + NaCl$$

This reaction was the more important commercially. Various modifications [121], e.g.,

$$C_6H_5NH_2 + HNO_2 + AsCl_3 + H_2O \longrightarrow C_6H_5-AsO_3H_2 + N_2 + 3\,HCl$$

have proved successful in special cases where the usual processes fail.

In one modern process [122], arsenic acid complexed with EDTA is added at 130 °C to a solution of excess aniline in perchloroethylene. The water of reaction and unreacted aniline are separated off, and the bis(4-aminophenyl)arsinic acid intermediate is converted by acidic hydrolysis into arsanilic acid, $NH_2C_6H_4-AsO(OH)_2$. Some arsonic acids are still used in raising livestock. 3-Nitro-4-hydroxyphenylarsonic acid is used as a feed additive against coccinid infections in poultry farming. In Germany, however, animal feed additives are controlled by legislation [123].

The main use of the arsonic acids was in their further processing to arsenobenzenes and "arsenic oxides" by reduction with SO_2, phosphorus trichloride, sodium dithionite, phosphorous acid, or $SnCl_2$. Reduction with zinc dust and hydrochloric acid gives the arsines, which are reoxidized simply on standing in air. For example, phenylarsine, $C_6H_5-AsH_2$, rapidly changes in air into the arseno compound $(C_6H_5-As)_n$. Further uptake of oxygen is considerably slower unless catalyzed, e.g., by iron.

The arsenic oxides $R-As{=}O$ are fairly stable. All compounds are oxidized to arsonic acid derivatives by relatively strong oxidants, including hydrogen peroxide, halogens, and Chloramine-T (sodium *p*-toluenechlorosulfonamide).

The aromatic arsonic acids are dibasic. Aqueous solutions of the monosodium salts are neutral to mildly acidic; those of the disodium salts

are moderately alkaline (pH 8 – 9). Calcium and magnesium salts are usually soluble in cold water, but when heated they precipitate as virtually insoluble deposits. This solubility in cold water may be used to separate arsonic acids from arsenic acid, whose calcium and magnesium salts precipitate from cold solutions. Heavy metal salts are insoluble. The arsonic acids generally crystallize well. The stability of an arsonic acid depends on the substituents on the benzene ring. A nitro group in the meta position destabilizes the bond to the arsono residue; thus, diazotized *m*-nitro-*p*-aminophenylarsonic acid loses arsenic on heating. The arsono residue can weaken the bond to the nitro group in the *meta* position, bringing about exchange with a hydroxyl group. Only a few arsonic acids have been used in therapy. Some arsonic acids form azo dyes that contain both arsonic acid and sulfonic acid groups. They are used in the analysis of metals. One example is arsenazo-III, 2,7-bis(*o*-arsonophenylazo)chromotropic acid [124].

Aromatic Arsenobenzenes. For some time arsenobenzenes were the only organic arsenic chemicals used in therapy. They are oxidized in the organism to arsenic oxides, the actual therapeutic agents. Probably the most important compound was 3,3′-diamino-4,4′-dihydroxyarsenobenzene, or Salvarsan, used to treat syphilis.

Properties. The aromatic arseno compounds are yellow substances. Compounds containing amino or hydroxyl groups are soluble in acids and alkalis. The introduction of formaldehyde sulfoxylate or formaldehyde hydrogen sulfite into the amino group makes the aromatic arseno compounds soluble in water. In aqueous solution the salts are hydrophilic colloids. The toxicities and effects of the arseno compounds depend on the degree of association and size of the molecule of the dissolved arsenobenzene. Minor changes in the production process can significantly affect the quality as a chemotherapeutic agent. The arsenobenzenes cannot be purified by crystallization or identified by a melting point. The original assumption that arsenobenzenes contained an arsenic – arsenic double bond had to be revised when cyclic and chain polymers were shown to be present [125].

Organic Oxoarsenic Compounds ("Arsenic Oxides"). The reduction of organoarsenic compounds can be controlled by using a suitable reducing agent so that reaction stops at the desired intermediate stage. This is not the case for oxidation. In the most widely used method for the production of organic oxoarsenic compounds from arsonic acids, the acid is directly reduced to the anhydride of the arsonous acid with SO_2. Hydrolysis of aryldihaloarsines is also a possibility.

Properties. The organic oxoarsenic compounds are the anhydrides of the arsonous acids. They are extremely poisonous, amphoteric substances sparingly soluble in water. They dissolve in acids and alkalis to form salts and can be precipitated from those solutions by carbon dioxide or ammonium chloride. The considerable significance attributed to sulfur-containing groups of the protein molecules in tissue and in ferments so far as biological functions are concerned and also the high affinity of arsenic oxides for cell thiols made it clear at a very early stage that the reaction between those two groups is responsible for the activity and toxicity of these compounds. The reaction takes place by formation of dithioarsenites:

$$R-As=O + 2\,R'SH \longrightarrow R-As\begin{matrix}\diagup SR' \\ \diagdown SR'\end{matrix} + H_2O$$

The resulting enzyme blockage can be reversed by a thiol.

3. Uses and Economic Aspects

The demand for metallic arsenic is limited. It is used in nonferrous alloys, and high-purity arsenic is used in electronic and semiconductor devices. The addition of ca. 0.5 % As to the lead grid in lead – acid storage batteries increases endurance and corrosion resistance. Additions of the same order to copper alloys improve high-temperature stability, recrystallization temperature, and corrosion resistance. The addition of up to 2 % As improves the sphericity of lead shot [1].

High-purity arsenic (at least 99.999 %) is used in electronics in conjunction with gallium or indium for the production of diodes (LED), infrared detectors, and lasers. Solar cells based on gallium arsenide have an efficiency of

20 % [126]. Among the semiconductors, silicon-based materials still dominate the market and will continue to do so in future because their production costs are far lower that those of the gallium arsenide semiconductors. Certain niches of properties which can not be provided by the silicon based semiconductors are filled by gallium arsenide, but although this marked is growing, it will be limited in the future [10], [127].

Most arsenic is used in the form of compounds, arsenic trioxide being the sole starting material. Arsenic compounds are used in the following fields:

1) Forestry: the use of arsenic trioxide as raw material for the production of wood preservatives today is its main application. The actual product is chromated copper arsenate (CCA), which is still hard to replace with other chemicals as far as effectiveness is concerned. A number of countries have forbidden the use of CCA as wood preservative The United States and Malaysia are the largest consumers. Because of the worldwide of environmental awareness and stricter regulations the almost monopoly consumer position of the United States makes the As_2O_3 market risky. In 1979 consumption of arsenic trioxide in agricultural chemicals in the United States was about 13500 t/a, but it dropped to 5000 t/a in 1991 [10], [16], [128], [129], [130]
2) Agriculture: herbicides and insecticides for cotton, coffee, and rice.
3) Industrial chemicals: electrolyte purification in the electrolytic recovery of zinc [131], metal pickling baths containing phosphoric acid.
4) Glass industry: fining agents and decolorizers [132].

The distribution of the arsenic use in the United States in the different fields of application are shown in Table 12. Overall consumption has remained constant at ca. 20 000 t/a for over ten years.

The wordwide production of arsenic trioxide is listed in Table 13. Total arsenic trioxide production underwent major changes since 1978. Production reached a peak in the mid-1980s and and has since been in decline, and at the same time the main producing countries changed completely. In 1978 Sweden, France, and Mexico were the largest producers; today's market is dominated by China, Chile, and the former Soviet Union. The main As_2O_3 producing companies are listed in the following:

China	not published
Chile	Cia Minera El Indio (CMEI)
Former Soviet Union	locations in the Urals, Altai, Kazahstan
Mexico	Industria Minera Mexico SA (IMM)
Philipines	not published
France	Societe Minere et Metallugique de Penarroya
Belgium	Metallurgie Hoboken-Overpelt (MHO)

In the 1980s Asarco in the United States stopped its As_2O_3 production, Boliden in Sweden shut down its production of arsenical products because of environmental problems, Salsigne in France declared bankruptcy, and China emerged as the world's largest producer of both arsenic trioxide and arsenic metal. Only China and Chile have countered the trend toward decreasing arsenic production [1], [10].

Arsenic metal accounts for only 3 % of the world demand for arsenic, mostly commercial grade (99 %). High-purity arsenic metal is produced by about ten companies. Furukawa Electric in Japan and Preussag in Germany are believed to be the largest producers, with reported capacities of 30 and 15 t/a, respectively [10].

Prices for arsenic trioxide are not readily available from published sources. As an average price for mexican As_2O_3 25 cents per pound can be found [10]. The price for arsenic metal was 55 – 60 cents per pound in October 1996 [135].

The largest consumers of arsenic trioxide are the United States, Malaysia, and the United Kingdom, in that order. Malaysia uses arsenic-based herbicides in its rubber plantations; the United Kingdom imports As_2O_3 and exports arsenic-containing pesticides and wood preservatives [133].

4. Stabilization of Arsenic for Safe Disposal

Because of the rapidly shrinking market for arsenic and its compounds, the focus of the research activities today lies more on the stabilization of arsenic-containing phases for safe disposal than on the production of arsenic-containing products. In some cases even As_2O_3-containing flue dusts that already have high arsenic contents are further processed for subse-

Table 12. Arsenic consumption in the United States in tons of As_2O_3 content per year [133], [10]

	1979		1991	
	t/a	%	t/a	%
Agriculture and forestry	13 500	70	5000	23
Industrial chemicals	4000	20	14 300	66
Glass industry	960	5	900	4
Alloys and electronics	600	3	1000	5
Others (pharmaceuticals, catalysts, feed additives)	360	2	400	2
Total	19 420	100	21 600	100

Table 13. Arsenic production in tons of As_2O_3 content per annum [133], [1], [10], [134]

	1978	1987	1991
Europe	22 189	32 178	16 060
Germany	563	360	360
Belgium		3500	3000
France	7280	10 000	3000
Portugal		218	200
Sweden	6706	10 000	2500
Spain			
Soviet Union	7640	8100	7000
Asia	551	15 536	15 900
Iran		1750	400
China		7000	10 000
Japan		500	500
Philippines		6286	5000
Africa	2407	1864	1804
Namibia		1864	1804
America	15 392	12 689	13 320
USA	7800 *		
Chile		3616	7000
Bolivia		132	460
Canada	72	2000	400
Mexico	6263	5304	4960
Peru	1257	1757	500
World total	40 283	62 387	47 084

* 1981.

quent disposal. Besides disposal in a highly concentrated form like As_2O_3 or even arsenic metal in special dumping areas, the most promising development is the stabilization as arsenate. The occurence of arsenates in nature and as ageing products on older arsenic-containing dumping sites suggests stability under weathering conditions. As an alternative to arsenates, the precipitation of sulfides has been tested and applied, although arsenic sulfides seem to be less stable. Especially Fe, Ca, Ba, Mg, and Sr arsenates are regarded as extremely stable. Today the preferred species is iron arsenate, which is not only among the most stable arsenates under weathering conditions and under the influence of atmospheric CO_2, but also the cheapest as far as additional chemicals are concerned.

The process parameters which most affect the stability of the arsenate precipitate are the temperature, the pH during precipitation, and the speed and method of addition of the precipitants. The stability of iron arsenate is further increased by a molar ratio Fe/As > 4 and by the presence of small quantities of Zn, Cu, and Cd. For such iron arsenates solubilities of < 0.2 mg/L As were measured in long-time tests (about three years). When precipitated under atmospheric pressure below 100 °C, amorphous iron arsenate is formed. Precipitation at 120 °C under the corresponding pressure leads to the formation of skorodite crystals with a structure similar to the naturally occuring mineral. The solubility of the crystallic iron arsenate 100 times lower than that of the amorphous phase. The production of iron arsenate from arsenic-

containing intermediate products in the nonferrous metals industry for subsequent disposal is in industrial use at various locations.

Stabilization of arsenic in pyrometallurgical slags has been investigated as an alternative to hydrometallurgical stabilization. Silicate slags seem to be less soluble than basic slags, but the attainable arsenic concentration in silicate slags (ca. 1 %, possibly up to 5 %) is fairly low [20], [54], [136–156].

5. Recycling

Recycling of arsenic-containing products to recover the arsenic content is presently of no economical interest. The increasing demand for GaAs in the electronic industry will lead to the question of the recovery of the gallium content from electronic scrap in the future because Ga is comparatively rare and therefore expensive. However, the aim of the recycling process as far as arsenic is concerned is stabilization and subsequent disposal. The smelting of GaAs scrap with a liquid copper alloy to collect arsenic while oxidizing the gallium into a soda/silica slag is one attempt to solve this future recycling problem [157].

6. Safety Measures

Although pure metallic arsenic is not poisonous, the occurrence of toxic arsenic oxides formed by oxidation in air is unavoidable. In handling or manufacture of arsenical materials, inhalation of dust and skin contact with arsenic-containing gases must be prevented. In the absence of adequate safety measures the following jobs involve a risk:

- metal recovery/contamination by arsenic
- roasting of iron pyrites
- processing of arsenical minerals
- production of arsenic-containing pharmaceuticals
- use in pyrotechnics
- burnishing of metal surfaces
- use of arsenic-containing raw materials in the glass industry
- repair or cleaning of flue-dust plants, filters, etc.
- use in wood preservatives.

All work involving arsenic or its compounds should be carried out in sealed containers. In filling and emptying of apparatus, adequate measures must be taken to ensure that the maximum allowable concentrations (MAK values) are not exceeded. A dust-free working environment should be ensured by maintaining reduced pressure in the apparatus or by an adequate exhaust system.

Sampling must be carried out only with the aid of suitable equipment, e.g., air locks, vacuum injectors, and sealed sampling vessels. Rooms used for work involving arsenical materials should be adequately ventilated and dust free. If there is a general danger of inhaling arsenical materials, e.g., during open sampling, respirators must be worn. In relatively heavily dust-laden atmospheres, special protective clothing should be worn to prevent any skin contact with arsenic. The arsenic content in the air of the working environment should be monitored continuously [158].

The MAK for AsH_3 in air is 0.16 mg/m^3 [158]. In Germany, arsenic trioxide, arsenic pentoxide, arsenous acid, arsenic acid, and their salts are included in the list of carcinogenic working materials. All work involving arsenic and its compounds is governed by the regulations for dangerous materials (TRK) [158].

In every country there are special provisions for packaging, identification, storage, and transportation of arsenical products. The dangers involved in the use of arsenical preparations are not treated as seriously in the United States as in the EEC, where generally speaking such preparations are completely banned.

7. Toxicology

Nearly all arsenic compounds are considerably toxic, especially the inorganic ones.

The trivalent arsenic compounds, such as arsenic trioxide, As_2O_3, arsenic trifluoride, AsF_3, arsenic trichloride, $AsCl_3$, and arsine, AsH_3, are markedly more toxic than the pentavalent compounds. Elemental arsenic itself is not appreciably toxic, but it is converted readily to toxic compounds in the organism. Pure arsenic(III) sulfide, As_2S_3 [*12255-89-9*], also seems to

have a comparatively low toxicity, but often the crude sulfide is contaminated by arsenic trioxide. In contrast, some pentavalent organic compounds have been used as therapeutics; these benzenearsonic derivatives contain a stable carbon – arsenic bond and are practically not converted to inorganic arsenicals [159].

Arsenicals can cause acute, subacute, and chronic poisoning. However, acute poisoning is rare in comparison to subacute or chronic poisoning.

Absorption. Arsenicals can be absorbed mainly by ingestion and inhalation, and also through the skin. They are widely distributed within 24 h into different tissues, mainly the liver, kidney, lung, spleen, bone marrow, skin, and, to a lesser extent, brain, heart, and uterus. Small quantities can be detected in hair and nails several months after the main portion of the arsenic has been eliminated from the body [160]. Inorganic arsenicals do not cross the blood – brain barrier but can pass across the mammalian placental membrane [161].

Distribution. After absorption, arsenic is distributed via the bloodstream to most tissues of the human body, preferring the liver, kidney, lung, intestinal wall, spleen, and heart, where it binds primarily to the SH groups of tissue proteins. Furthermore, it is incorporated into hairs and nails, and this can be used for the detection of an exposure to arsenical materials in the past. Arsenic passes the blood – brain barrier only in small amounts. It is incorporated into bones, where it replaces phosphorus and may remain for years [162].

Metabolism. Most arsenic compounds – trivalent or pentavalent, inorganic or organic – are transformed to an arsenic oxide form, which can react readily with proteins, especially with those containing two vicinal mercapto groups, thus forming stable cyclic structures. The inhibition of essential enzymes is considered to be the first step of cell damage [159]. The high affinity of the oxoarsenic moiety for the vicinal mercapto groups of 2,3-dimercapto-1-propanol [*59-52-9*] (BAL, British Anti Lewisite) and of 2,3-dimercapto-1-propanesulfonic acid [*74-61-3*] is used successfully in arsenic poisoning therapy to enhance excretion.

Excretion of arsenicals occurs mainly via urine, to a smaller extent via the feces [163], and by loss of hair and skin shedding. After the administration of a single dose, the concentration of arsenic in liver and kidney starts to decrease only after 24 h [163]. The concentration of arsenic in urine of unexposed persons is between zero and 0.22 mg/L, after industrial exposure between 0.04 and 0.9 mg/L [164]. After inhalation of arsenic dust, the excretion of methylated arsenic could be observed [165]; traces of arsenic can be exhaled as trimethylarsine [166].

Acute Toxicity. Arsenic poisoning mostly occurs by ingestion of inorganic trivalent compounds such as arsenic trioxide. The smallest doses of As_2O_3 reported to be fatal after ingestion are 70 – 180 mg [160]. Symptoms which can occur within a few minutes or after a delay of several hours are vomiting, diarrhea, damage of the intestinal tract, muscular cramps, facial edema, cardiac abnormalities, dehydration, and – as a consequence – shock. The inhalation of irritant arsenic compounds, such as arsenic trichloride or arsenical war gases, causes cough, dyspnea, pain in the chest, severe damage to the respiratory tract, headache, general weakness, nausea, vomiting, and colic diarrhea. Airborne dusts often result in irritation of exposed skin and mucous membranes such as bronchitis, conjunctivitis, laryngitis, and dermatitis [167]. Arsenic trichloride may be absorbed through the skin with fatal consequences [168].

Chronic Toxicity. Chronic arsenic poisoning can be caused by intake of food and water contaminated with arsenicals or by inhalation of airborne dusts during long-term occupational exposure, mainly in smelting plants of ores containing arsenic, insecticide factories, and among vineyard workers, even many years after termination of exposure. Symptoms of these chronic poisonings were symmetrical palmar and plantar hyperkeratosis, white striae of the finger nails, cardiovascular manifestations, myocardial ischemia, hypertension, liver dysfunctions, hematological changes, and vascular disorders resulting in gangrenes of the lower extremities (blackfoot disease). A variety of severe symptoms is treated below.

Skin lesions are frequently observed, e.g., in the production of insecticides [169] and in copper ore smelting plants [170]. Typical symptoms are hyperkeratosis, warts, melanosis, contact dermatitis, eczematoid features of varying degrees of severity, and vascular lesions. Hyperkeratosis often is characterized by thickening, drying, and cracking of the skin [171]. Arsenical melanosis is also a common sign after long-term exposure and can occur on the chest, abdomen, scrotum, and back. Further characteristic symptoms are depigmentation (leukodermia) of pigmented areas and white striae of the fingernails [172].

Effects on mucous membranes are conjunctivitis with swelling and pain, keratoconjunctivitis (resulting from exposure to the insecticide calcium arsenate, $Ca_3(AsO_4)_2$), irritation of the pharynx and bronchial passages, and irritation of the nose resulting in acute and chronic rhinitis. Among copper smelter workers, perforation of the nasal septum has been commonly observed [173].

Peripheral neuritis, with such symptoms as pain, difficulties in walking, burning and tenderness in the affected limbs, and later on severe weakness in both legs and feet, has been described after long-term exposure to arsenate dusts [171]. In contrast to the peripheral neuritis caused by lead, the arsenic-caused neuritis is painful; the nerves of these patients show neuronal degeneration [160]. Further symptoms, observed especially in cases of severe chronic intoxications, are headache, aphasia, drowsiness, disorientation, and changes in the personality.

Liver damages, especially liver cirrhosis, have been observed, e.g., in vintagers who used arsenical herbicides [174] and after long-term consumption of arsenic-contaminated homemade wine [175]. These liver disturbances, which often are complicated by the additional occurrence of chronic hepatitis, ascites, and enlarged liver, are probably not due to skin contact with or inhalation of arsenicals, but to ingestion of contaminated wine. The occurrence of liver damage among industrial workers is reported to be less frequent [167].

Circulatory system symptoms after chronic arsenical poisoning are abnormalities in the electrocardiogram [176] indicating a toxic myocardial effect, peripheral vascular disturbances, gangrene of the extremities, atrophic acrodermatitis, endoangitis [177], and the so-called blackfoot disease [178]. *Gastrointestinal* disturbances are less common among workers with long-term exposure to arsenicals [179]. During chronic intoxication by ingestion, digestive disturbances, nausea, vomiting, and loss of appetite are occasionally reported. *Hematological changes* are also less common in chronic poisoning after industrial airborne exposure. Symptoms like moderate anemia, leucopenia, and thrombocytopenia have been observed after long-term ingestion of inorganic arsenicals. However, these symptoms disappeared 2 – 3 weeks after termination of the arsenic ingestion [180].

Cancer. Workers in smelters or in insecticide factories exposed to arsenic trioxide have a significantly higher risk of lung cancer in comparison to persons not exposed , [182]. A significant excess of oral and pharyngeal cancer could be noticed among male textile workers in England and Wales engaged in fiber preparation [183]; this excess could be traced to arsenic-containing disinfectants used on sheep. Skin cancer as a result of ingesting of contaminated water, with multifocal lesions over the entire body, occurred in Cordoba, Argentina , Antofagasta, Chile , and in Taiwan [186]. The development of other types of tumors related to long-term exposure by ingestion, inhalation, or dermal contact has been observed frequently; however, in these cases the relationship between arsenic exposure and development of tumors could not be confirmed as yet unequivocally [164]. Arsenic should probably be considered as a cocarcinogen [187], [188], since it is able to inhibit DNA-repair enzymes but does not show mutagenicity [189].

Mutagenicity. The mutagenic action of inorganic and organic arsenicals has been tested on microorganisms (DNA damage in *Bacillus subtilis*) [190], Drosophila (increase in crossing-over frequencies) [191], and mammalian cells in vivo and in vitro. A significant increase of chromosomal aberrations could be demonstrated as a result of contact with arsenicals in lymphocytes from peripheral blood [192] and in fibroblast cultures of humans [193].

Reproductive Toxicity. *Teratogenic effects* of disodium arsenate [*7778-43-0*] , such as exen-

Table 14. Toxicological data of arsenic and some arsenic compounds

Compound	CAS registry number	Toxicol. parameter	Dose	Species	Route	Reference
Arsenic	[*7440-38-2*]	LDLo	20 mg/kg	rat	intramusc.	[212]
Arsenic acid, hemihydrate	[*7774-41-6*]	LD_{50}	6 mg/kg	rabbit	i.v.	[213]
Arsenic pentoxide	[*1303-28-2*]	LD_{50}	55 mg/kg	mouse	oral	[214]
Arsenic trichloride	[*7784-34-1*]	LDLo	100 mg/m^3, 1 h	cat	inhal.	[215]
Arsenic trifluoride	[*7784-35-2*]	LDLo	2000 mg/m^3, 10 min	mouse	inhal.	[216]
Arsenic trioxide	[*1327-53-3*]	LD_{50}	45 mg/kg	mouse	oral	[214]
		LD_{50}	20 mg/kg	rat	oral	[217]
		LDLo	4 mg/kg (as As)	rabbit	i.v.	[213]
Arsine	[*7784-42-1*]	TCLo	3 mL/m^3	human	inhal.	[160]
		LCLo	25 mL/m^3, 30 min	human	inhal.	[218]
		TCLo	93 µL/m^3	human	inhal.	[219]
Calcium arsenate	[*7778-44-1*]	LDLo	50 mg/kg	rabbit	oral	[220]
Disodium arsenate	[*7778-43-0*]	LDLo	30 mg/kg	rat	i.p.	[221]
Disodium arsenate, heptahydrate	[*10048-95-0*]	TDLo	43 mg/kg	mouse	i.p.	[222]
Lead arsenate	[*7645-25-2*]	LD_{50}	100 mg/kg	rat	oral	[223]
		LDLo	75 mg/kg	rabbit	oral	[220]
Sodium metaarsenite	[*7784-46-5*]	TDLo	40 mg/kg	mouse	oral	[224]

cephaly [194], cleft palate and lip, microanophthalmia, genitourinary abnormalities, ear deformities, and renal agenesis, were observed in hamster fetuses [195–197]. In the case of a 17-year-old girl, the ingestion of arsenic during the third trimester of her pregnancy caused the death of her baby (1100 g!) 11 h after birth; liver, kidney, and brain of the baby contained high concentrations of arsenic [198]. Among employees and persons living near a smelter, a significant decrease of the birth weight of the offspring, an increased frequency of spontaneous abortions, and a higher rate of malformations of the children born were observed [199], [200].

In animal studies, an effect of arsenic on *fertility* in males or females has not been demonstrated so far [201].

Arsine. Acute poisoning with arsine [*7784-42-1*], AsH_3, which is rather frequent among industrial workers as compared to poisoning with other arsenicals, results in general malaise, nausea, vomiting, abdominal cramps, coppery skin pigmentation, red staining of the conjunctiva [202], and leucocytosis [203]. A most characteristic symptom of acute arsine poisoning is rapid hemolytic anemia, which results in hemoglobinuria, decreased hemoglobin level, jaundice, and after 2 – 24 h, shock, oliguria, and anuria due to blockade of renal tubuli [204] by membrane residues of red blood cells. Other symptoms are pyrexia, edema of the lungs, delirium, and coma. Death after few days is mostly due to myocardial failure. In all cases of arsine poisoning an elevation of the T-wave in the electrocardiogram can be noticed. Depending on the severity of the intoxication, the time of the appearance of symptoms may vary between a few minutes and 24 h.

Chronic arsine poisoning shows a delayed onset of symptoms mentioned above, and in some cases peripheral neuritis [205], hemoglobin levels as low as 32 g/L, and basophilic stippling of red blood cells [206] have been observed.

Organic Arsenic Compounds. Compounds of the type $RAsCl_2$, R_2AsCl, and R_2AsCN have been developed and partly used as war gases during World War I. Examples are dichloro-2-chlorovinylarsine [*541-25-3*], lewisite, $ClCH{=}CHAsCl_2$, chlorodiphenylarsine, Clark I D., $(C_6H_5)_2AsCl$, and cyanodiphenylarsine, Clark II D., $(C_6H_5)_2AsCN$. They generally exert a high toxicity, which is directed against skin, eyes, respiratory tract, gastrointestinal tract, and central nervous system, showing symptoms similar to those described above [207].

Toxicological Data and Exposure Limits. Toxicological data of commonly used arsenicals are summarized in Table 14. The MAK value of arsine is 0.2 mg/m^3 (0.05 ppm) [208]. The MAK values for various arsenicals such as oxides and arsenic acids and their salts have not been established because these compounds have

been proved to cause malignant tumors in humans [208]. According to the IARC evaluation, arsenic and its compounds are carcinogenic to humans (Group I) [209], whereby the skin and lungs are proven target organs, while the liver, hematopoetic system, gastrointestinal tract and kidney are suspected target organs. The TRK values in Germany for all compounds with the exception of arsine is 0.1 mg/m^3 (as As) [208]. The unit risk value representing the additional relative carcinogenic risk caused by AS (based on life-long exposure to 1 µg As/m^3) has been established as 0.004 by the LAI [210].

In the United States, the TLV-TWA for arsenic and its compounds is 0.2 mg/m^3; the value for arsine is 0.2 mg/m^3 (0.05 ppm) [211]; in the United Kingdom, the MEL is 0.1 mg/m^3 (as As).

8. References

1. *Ullmann,* 5th ed., **vol. A3,** pp. 113 – 141.
2. *Römpps Chemie-Lexikon,* 8th ed., **vol. 1,** pp. 274 – 280.
3. H. Stöhr, *Z. Anorg. Allg. Chem.* **242** (1939) 138.
4. Ullmann, 4th ed., vol. 8, 46 – 66.
5. *Handbook of Chemistry and Physics,* 67th ed., CRC Press, Boca Raton, FL 1986.
6. *Technologie außergewöhnlicher Metalle,* Verlag Europa-Lehrmittel, Wuppertalpp. 7 – 8.
7. Felix A. Schaufelberger: "Arsen und das "Arsenpanorama" der Waliser Alpen", *Neue Zürcher Zeitung,* 18.1.1989, Nr. 14.
8. N. L. Piret, A. E. Melin: "An Engineering Approach to the Arsenic Problem in the Extraction of Non-Ferrous Metals", in M. Koch, J. C. Taylor (eds): *"Productivity and Technology in the Metallurgical Industries",* The Minerals, Metals & Materials Society (TMS), Warrendale, PA 1989, pp. 735 – 814.
9. V. L. Escobar Gonzales, A. J. Monhemius: "The Mineralogy of Arsenates Relating to Arsenic Impurity Control", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 405 – 454.
10. J. R. Loebenstein: "Arsenic", in: *Minerals Yearbook,* **vol. 1:** Metals and Minerals, United States Department of the Interior Bureau of Mines, 1991, pp. 239 – 244.
11. F. Pawlek: *Metallhüttenkunde,* De Gruyter, Berlin 1982, pp. 530 – 533.
12. L. K. Chuchalin, G. L. Pashkov, *Cvetnye Metally* 1980, no. 9, 16 – 18.
13. P. M. Swash: "A Mineralogical Investigation of Refractory Gold Ores and their Benefication, with special Reference to Arsenical Ores", *J. S. Afr. Inst. Min. Metall.* **88** (1988) no. 5, 173 – 180.
14. C. T. Zuniga: "Depression de Arsenopirata en un Cocentraado de Zinc", *Geomimet* **141** (1986) 76 – 96.
15. M. Lesoille: "Schema de Traitement de Concentres Arsenies Auriferes a l'usine Metallurgique de Salsignie", *Industrie Minerale-Mines et Carrieres-Les Techniques,* Sept. 1987, 321 – 235.
16. K. Neumann, S. Sudhölter: "Die Arsenproblematik in der Zinkgewinnung", unpublished Thesis at the Institute of NF Metals Processing, RWTH-Aachen, 1995.
17. R. M. Grant: "Zinc Concentrate and Processing Trends", in: "World Zinc '93", Proceedings of a Conference in Hobart Australia, October 10 – 13, The Australasian Institute of Mining and Metallurgy, 1993, 391 – 397.
18. B. M. Loftus et al.: "The Development of a Jet Reactor for the Leaching of Valuable Metals", *SAIMM* **97** (1997) no. 7, 285 – 288.
19. W. Qun, K. Heiskanen: "Separation of Pentlandite and Nickel Arsenide Minerals by Aeration Conditioning Flotation", *Int. J. Mineral Processing* **29** (1990) 99 – 109.
20. S. Sudhölter (Grund): "Die Rückstandsproblematik der hydrometallurgischen Zinkgewinnung – eine metallurgisch-wirtschaftliche Betrachtung unter besonderer Berücksichtigung des Arsens", Doctorates Thesis, RWTH-Aachen, 1998.
21. N. Chakraborti, D. C. Lynch: "Thermodynamics of Roasting Arsenopyrite", *Met. Trans. B.* **14B** (1983) June, 239 – 251.
22. D. C. Lynch, N. Chakraborti: "Strategies for Roasting arsenical Concentrates", in: Applied Mineralogy: Proceedings of the International Congress on applied Mineralogy in the Mineral Industries, 2nd 1984 (pub. 1985), 741 – 755.
23. G. Ramadori, R. K. Mishra: "Roasting of Arsenical Gold and Silver Bearing Minerals", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 161 – 172.

24. M. Stefanakis, et al.: "Dearsenifying and Gold Extraction from a Refractory Pyrite Concentrate" in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 173 – 198.
25. Olov Herneryd et al.: "Copper Smelting in Boliden's Rönnskär Works described", *JOM* (1954) 330 – 337.
26. *Sulphur* **no. 78,**1968,20 – 24.
27. G. Ramadorai: "Roasting Options for Inproved Gold-Silver recoveries", in D. M. Hausen et al. (eds.): *"Process Mineralogy XI-Characterisation of Metallurgical and Recyclable Products"* The Minerals, Metals & Materials Society, Warrendale, PA 1991, pp. 95 – 105.
28. G. Lindkvist, A. Holmström: "Roasting of Complex Concentrates with High Arsenic Contents", in H. Y. Sohn et al. (eds.): *"Advances in Sulfide Smelting",* TMS, 1983, pp. 451 – 472.
29. E. H. Smith et al.: "Selective Roasting to De-Arsenify Enargite/Pyrite Concentrate from St. Joe's El Indio Mine – From Pilot Plant to Commercial Operation", A. D. Zunkel et al. (eds.): *"Complex Sulfides Processing of Ores, Concentrates and By-Products",* The Minerals, Metals & Materials Society, Warrendale, PA 1985, pp. 421 – 444.
30. E. H. Smith, E. Paredes: How St. Joe Gold's El Indio Mine Has Become a Major Producer of High Quality Crude Arsenic Trioxide", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 145 – 160.
31. H. Wolf et al., *Chem. Ing. Tech.* **40** (1968) 441.
32. *Sulphur* **no. 84,**1969,30 – 32.
33. K. G. Görling et al., *Z. Erzbergbau Metallhüttenwesen* **12** (1959) 553 – 557.
34. H. Weigel, D. Zetsche: "Verfahren und Betriebsergebnisse der neuen ENAF-Zinnhütte in Vinto (Bolivien)", *Erzmetall* **27** (1974) no. 6, 257 – 312.
35. M. J. Collins et al.: "The Behaviour and Control of Arsenic in the Pressure Oxidation of Uranium and Gold Feedstocks", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 115 – 134.
36. St. Dayton: "Equity silver on line with leach plant. How purged from a high-silver copper concentrate before smelting," *Eng. Min. J.* **183** (Jan. 1982), 78 – 83.
37. R. W. Lawrence, P. B Marchant: Biochemical Pretreatment in Arsenical Gold Ore Processing", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 199 – 214.
38. K. P. V. Lei, T. G. Carnahan: "Effect of Silver on Oxidative Leaching of an Arsenical Copper Sulphide Concentrate", *TMS-AIME* **A88-4** (1988).
39. R. O. Dannenberg et al.: *"Recovery of Cobalt and Copper from Complex Sulphide Concentrates"* . US-Bureau of Mines, RI 9138, 1987.
40. G. van Weert: "The Nitrox-Process for Treating Gold Bearing Arsenopyrites", *TMS-AIME,* **A87-40** (1987).
41. J. F. Spisak: "Clean Production of High Purity Arsenic Trioxide from a Bolivian Tungsten Ore", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, p. 401.
42. D. H. Anthony: "Nerco Con Mine Arsenic Plant-Environmental Management through Resource Recovery", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, 1988, pp. 1135 – 144.
43. Patentaktiebolaget Gröndal-Ramén, DE 628554, 1937.
44. Boliden AB, DE-OS 1806647, 1968.
45. V. G. Rchiladze: *Arsen,* Moscow 1969.
46. Boliden AB, CA 863857, 1971 (S. T. Henriksson).
47. H. E. Hirsch: "The recovery of metallic arsenic in COMINCO's lead operations," *Proceedings of Lead-Zinc-Tin 80, committee at the 109th AIME Annual Meeting,* Las Vegas, Nevada, Febr. 1980, pp. 24 – 28.
48. D. Lezal et al., *Chem. Zvesti* **19** (1968) 620 – 627.
49. Siemens-Schuckert-Werke AG, DE 1181919, 1964.
50. A. S. Girenko, O. N. Kalashnik, V. A. Kirichenko, I. I. Skakovskii, A. M. Tuzovskii: "Producing arsenic of semiconductor purity by the chloride method," *Cvetnye Metally* 1972, no. 1, 59 – 62.

51. K. Toyabe et al.: "High-Purity Arsenic Metal Production at SMM's Nihama Copper Refinery", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 351 – 362.
52. M. Altepeter, S. James: "Proposed Treatment of Neutral Leach Residue at Big River Zinc", in R. G. Reddy, W. P. Imrie, P. B. Queneau (eds.): *"Residues and Effluents – Processing and Environmental Considerations",* The Minerals, Metals & Materials Society, Warrendale, PA 1991, pp. 449 – 459.
53. L. E. Deininger, R. W. Höhn: "Emissionsminderung einer Bleihütte durch Ersatz der Sinter- und Schachtofenanlage mit einem QSL-Reaktor", Abschlußbericht Umweltbundesamt Nr. 50441-5/149, Firmenschrift "Berzelius" Stolberg GmbH, Stolberg 1993.
54. M. Tomita, M. Higashi, S. Oto, H. Okamoto: "Hydrometallurgical Process of Copper Converter Dust", in R. G. Reddy, W. P. Imrie, P. B. Queneau (eds.): *"Residues and Effluents-Processing and environmental Considerations",* The Minerals, Metals & Materials Society, Warrendale, PA 1991, pp. 283 – 293.
55. Hazen Research, US 4244734, 1981 (J. E. Rupolds, E. L. Coltrinari).
56. M. Tomita et al.: "Hydrometallurgical Process of Copper Converter Dust", in R. G. Reddy et al. (eds.): *"Residues and Effluents-Processing and Environmental Considerations",* The Minerals, Metals & Materials Society, Warrendale, PA 1991, pp. 238 – 293.
57. T. Terayama: "Arsenic Recovery and High Purity Arsenic Metal Production from Arsenic Residues of Copper Smelting Processes", *Met. Rev. of MMIJ,* **6** (1989) no. 2, 1 – 16.
58. G. Sandberg, T. L. Hebble, D. L. Paulson: "Oxidative sulfuric acid leaching of lead smelter mattes," *Rep. Invest. US. Bur. Mines,* **RI 8371** (1979) 1 – 16.
59. "New USBM methods boost metals recovery in leaching processes in this month in mining," *Eng. Min. J.* **181** March 1980, 43.
60. E. I. Ponomarewa, W. D. Solowiewa, *Vestn. Akad. Nauk Kaz. SSR* 1974, no. 6, 52 – 55.
61. P. A. Bloom, J. H. Maysilles, H. Dolezal: "Hydrometallurgical treatment of arsenic-containing lead-smelter flue dust," *Rep. Invest. U.S. Bur. Mines,* **RI 8679** (1982) 1 – 12.
62. A. Metha: "A research investigation of new techniques for control of smelter arsenic flue dust wastes," EPA Grant R 804595010, Min. Res. Center Montana, Techn. Alumin. Foundation, Butte, Mont., 1977, sec. 4, pp. 1 – 43.
63. G. A. Brooks et al.: "The Treatment of Arsenic and Antimony Oxide Fumes", in N. Nilmani et al. (eds.): *Pyrometallurgy for Complex Materials and Wastes,* The Minerals, Metals & Materials Society, Warrendale, PA 1994, pp. 77 – 287.
64. Klöckner-Humboldt-Deutz, DE 3003635 A 1, 1981 (I. Barin, G. Michael, R. Hesse, S. Wirosoedirdio).
65. E. A. Woolson: "Arsenical Pesticides," *ACS Symp. Ser.* 1975, no. 7, 108 – 124.
66. R. S. Braman: "No Arsenical Pesticides," *ACS Symp. Ser.* 1974, no. 7, 1 – 13.
67. H. Berg, H. Sperlich: "Nachweis und Bestimmung von Organoarsenverbindungen auf Früchten," *Mitteilungsbl. GDCh Fachgruppe Lebensmittelchem. Gerichtl. Chem.* **28** (1974) 298.
68. R. G. Robins: "The Solubility of Metal Arsenates," *Met. Trans.* **12 B** (1981) 103.
69. *Handbook of Chemistry and Physics,* 63rd ed.,1982 – 1983, The Chemical Rubber Co., Cleveland.
70. *Kirk Othmer,* 3rd ed., **vol. 3,** p. 243 – 266.
71. M. Jansen, *Angew. Chem.* **89** (1977) 326.
72. M. Jansen, *Z. Anorg. Allg. Chem.* **411** (1978) 5.
73. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* vol. 4 + 5, Springer Verlag, Berlin York 1977.
74. Arbeitsstoffverordnung v. 29. 7. 1980, BGBl. I,p. 1071, p. 1536, 2159 and v. 11. 2. 1982, BGBl. I, p. 144.
75. D. D. Wagman et al., *NBS Tech. Note U.S.* 1968, no. 270-3, 95 – 98.
76. K. Seppelt, *Z. Anorg. Allg. Chem.* **434** (1977) 5; **439** (1978) 5.
77. *Winnacker-Küchler,* 4th ed., **vol. 3** (1983), p. 422.
78. Umweltbundesamt: *Umwelt- und Gesundheitskriterien für Arsen, Berichte 4/83* , E. Schmidt Verlag, Berlin 1983.
79. Houben-Weyl, vol. XIII/8, p. 33 – 443.
80. A. Tzschach, J. Heinicke: *Arsenheterocyclen,* Verlag für Grundstoffindustrie, Leipzig 1978.
81. P. Jutzi: "Neue Element-Kohlenstoff-(p-p)π-Bindungen", *Angew. Chem.* **87** (1975) 269 – 283.

82. F. Kober: "Chemie der Arsinooxine," *Chem. Ztg.* **101** (1977) 133 – 137.
83. G. O. Doak, L. D. Freedman: *Organometallic Compounds of As, Sb, Bi,* Wiley-Interscience, New York 1970.
84. F. G. Mann: *The Heterocyclic Derivatives of P, As, Sb, Bi,* Interscience Publ., New York 1970.
85. R. A. Zingaro, K. J. Irgolic, *J. Organomet. Chem.* **257** (1983) 67 – 104; **207** (1981) 121 – 184; **176** (1979) 245 – 306; **138** (1977) 125 – 156.
86. G. O. Doak, L. D. Freedman, *J. Organomet. Chem.* **106** (1976) 115 – 216; **89** (1975) 81 – 150; **68** (1974) 295 – 369.
87. A. L. Rheingold, P. J. Sullivan: "Crystal and Molecular Structure of Hexaphenylcyclohexaarsine c-$(AsPh)_6$," *Organometallics* **2** (1983) 327.
88. J. Ellermann, M. Lietz: "Chemie polyfunktionaler Liganden, 69. Drei Organoarsen-Käfigverbindungen mit Noradamantan-Struktur," *Z. Naturforsch. B* **36 B** (1981) 1532.
89. P. Maroni et al.: "Heterocyclic organoarsenic compounds, Part IX, 1,3-Diaza-2,4-diarsacyclobutanes derived from norephedrines. Evidence of oligermers," *J. Chem. Res. Synop.* 1981, no. 7, 176.
90. K. Kellner, H. J. Schultz, A. Tzschach: "Organo- arsen-Verbindungen XL. Synthesen und Reaktionen von Alpha-(arylsulfonamido)benzylphosphinen und -arsinen," *Z. Chem.* **20** (1980) 152.
91. F. Kober: "Adamantan-strukturierte Arsenverbindungen," *Chem. Ztg.* **103** (1979) 357.
92. F. D. Yambushev, V. I. Saviu: "Stereochemistry of organoarsenic compounds," *Usp. Khim.* **48** (1979) no. 6, 1093.
93. A. Otero, P. Royo: "Pentafluorophenylarsenic compounds," *J. Organomet. Chem.* **149** (1978) 315.
94. T. Kauffmann, R. Joussen, A. Woltermann: "New reagents, 5. Vielseitiges Prinzip zur Synthese funktionalisierter Kohlenwasserstoffe via Organoarsenverbindungen," *Angew. Chem.* **89** (1977) 759.
95. Y. Odanaka et al.: "Determination of Inorganic Arsenic and Methylarsenic Compounds by Gas Chromatography and Multiple Ion Detection Mass Spectrometry after Hydride Generation – Heptane Cold Trap Technique," *Anal. Chem.* **55** (1983) 929.
96. G. Ricci et al.: "Suitability of Various Filtering Media for the Collection and Determination of Organoarsenicals in Air," *ACS Symp. Ser.* 1981, no. 149, 383, No Chem. Hazards Workplace: Meas. Control.
97. W. A. Maher: "Determination of Inorganic and Methylated Arsenic Species in Marine Organisms and Sediments," *Anal. Chim. Acta* **126** (1981) 157.
98. G. Ricci et al.: "Ion Chromatography with Atomic Absorption Spectrometric Detection for Determination of Organic and Inorganic Arsenic Species," *Anal. Chem.* **53** (1981) 610.
99. J. P. Buchet, R. Lauwerys, H. Roels: "Comparison of Several Methods for the Determination of Arsenic Compounds in Water and Urine," *Int. Arch. Occup. Environ. Health* **46** (1980) 11.
100. F. E. Brinckman et al.: "Graphite Furnace Atomic Absorption Spectrophotometers as Automated Element-Specific Detectors for High-Pressure Liquid Chromatography. The Determination of Arsenite, Arsenate, Methylarsenic Acid and Dimethylarsonic Acid," *J. Chromatogr.* **191** (1980) 31.
101. C. H. Gast et al.: "Capabilities of On-Line Element-Specific Detection in High Performance Liquid Chromatography Using an Inductively Coupled Argon Plasma Emission Source Detector," *J. Chromatogr.* **185** (1979) 549.
102. R. S. Braman et al.: "Separation and Determination of Nanogram Amounts of Inorganic Arsenic and Methylarsenic Compounds," *Anal. Chem.* **49** (1977) 621.
103. G. Tiedemann, J. H. Einbrodt: "Zur Mutagenität anorganischer Arsenverbindungen," *Wiss. Umwelt* 1982, no. 3, 170.
104. G. Pershagen, B. Lind, N. E. Bjoerklund: "Lung Retention and Toxicity of some Inorganic Arsenic Compounds," *Environ. Res.* **29** (1982) 425.
105. R. A. Pyles, E. A. Woolson: "Quantitation and Characterization of Arsenic Compounds in Vegetables Grown in Arsenic Treated Soil," *J. Agric. Food Chem.* **30** (1982) 866.
106. K. J. Irgolic et al.: "Analytical and Biochemical Aspects of the Transformation of Arsenic and Selenium Compounds into Biomolecules," *NBS Spec. Publ.* 1981, no. 618, 244.
107. Y. Yamauchi, Y. Yamamura: "Dynamic Change of Inorganic Arsenic and Methylarsenic Compounds in Human Urine

after Oral Intake as Arsenic Trioxide," *Ind. Health* **17** (1979) 79.
108. R. L. Spehar et al.: "Comparative Toxicity of Arsenic Compounds and their Accumulation in Invertebrates and Fish," *Arch. Environ. Contam. Toxicol.* **9** (1980) 53.
109. A. E. Ledet, W. B. Buck: "Toxicity of Organic Arsenicals in Feedstuffs," *Hazard. Toxic Subst.* **2** (1978) 375.
110. W. B. Buck: "Toxicity of Inorganic and Aliphatic Organic Arsenicals," *Hazard. Toxic Subst.* **2** (1978) 357.
111. K. J. Irgolic et al.: "Characterization of Arsenic Compounds Formed by Daphnia magna and Tetraselmis chuii from Inorganic Arsenate," *EHP Environ. Health Perspect.* **19** (1977) 61.
112. A. Brockhaus, N. Manojlovic: "Arsenverbindungen in unserer Umwelt," *Lufthyg. Silikoseforsch.* **6** (1973) 49.
113. S. J. Zolotukhin: "Organic Arsenic Compounds in a Tropical Diseases Clinic," *Tr. Univ. Druzhby Nar. im. Patrisa Lumumby* **75** (1976) 232.
114. E. Schröder, C. Rufer, R. Schmiechen: *Arzneimittelchemie,* vol. 3, Thieme Verlag, Stuttgart 1976.
115. IUPAC Commision of the Nomenclature of Organic Compounds, Information Bulletin, I.U.P.A.C., Append. Tentative Nomenclature Symbols, Units, Standards, Section D, 1973.
116. *Chem. Abstr.* **56** (1962) 43 N.
117. A. Bechamp, *C.R. Séances Acad. Sci.* **56** (1863) 1172.
118. H. Bart, DE 250264, 1912; DE 254092, 1910; DE 254345, 1910; DE 266944, 1913.
119. H. Schmidt, *Justus Liebigs Ann. Chem.* **421** (1920) 159.
120. V. Heyden, DE 264924, 1913.
121. E. Scheller, DE 522829, 1926.
122. Rohm and Haas Co., US 3586707, 1971 (J. W. Nemec, H. R. Raterink, S. W. Wise).
123. BGBl. I, p. 352: Futtermittelverordnung v. 8. 4. 1981.
124. H. Berge, *Z. Chem.* **18** (1978) 168.
125. A. S. Levinson: "The Structure of Salvarsan and the Arsenic-Arsenic-Double Bond," *J. Chem. Educ.* **54** (1977) 98.
126. CAEN, 9 May 1977, 5.
127. F. Weinert, "Theoretische Grundlagen und Anwendungen zur Einkristallzüchtung von Galliumarsenid", Auszeichnungsantrag f. d. Paul-Grunfeld-Preis, Bergakademie Freiberg 1994.
128. A. D. J. Broome et al.: "The Production and Use of Arsenicals in Wood Preservation", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 491 – 506.
129. W. J. Baldwin: "The Use of Arsenic as a Woodpreservative – an Update", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, 1988, pp. 507 – 514.
130. P. M. Randall, A. S. C. Chen: "Waste Reduction Practices at two Chromated Copper Arsenate Wood-Treating Plants". *Journal of the Air & Waste Management Association,* **45** (1995) May 404 – 411.
131. K. Tozawa: "Comparison between Purification Process for Zinc Leach Solutions with Arsenic and Antimony Trioxides", *Hydrometallurgy,* **30** (1992) 445 – 461.
132. K. P. Hanke: "Blasenbildung in Glasschmelzen," *Umschau Wiss. Tech.* **72** (1972) 436.
133. U.S. Bureau of Mines, Bulletin 671, Mineral Facts and Problems, 1980.
134. Roskill: "The Economics of Arsenic 1988", 6th ed., Roskill Information Services Ltd.
135. R. Serjeantson (ed.): *"Metal Bulletin's Prices and Data 1997",* Metal Bulletin Books LTD, 1997, 59.
136. J. E. Hoffmann: "Remediating Copper Smelter Dusts: the Arsenic Problem", *JOM,* Juni (1994) 36 – 38.
137. X. Hartinger: *"Handbuch der Abwasser- und Recyclingtechnik",* 2nd ed., Carl Hanser Verlag München Wien, 1991, pp. 176 – 177.
138. V. L. E. Gonzales, A. J. Monhemius: "The Mineralogy of Arsenats Relating to Arsenic Impurity Control" in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Application", Phoenix, Arizona, 1988,* The Minerals, Metals & Materials Society, Warrendale, PA 1988, pp. 405 – 439.
139. F. J. G. Ugarte, A. J. Monhemius: "Charakterisation of High-Temperature Arsenic-Containing Residues from Hydrometallurgical Processes", *Hydrometallurgy* **30** (1992) 69 – 86.
140. R. G. Robins: "Arsenic Hydrometallurgy", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Application", Phoenix, Arizona, 1988,* The Minerals, Metals & Materials Society, Warrendale, PA 1988, pp. 215 – 247.

141. T. Nishimura, C. T. Itoh, K. Tozawa: "Stabilities and Solubilities of Metal Arsenites and Arsenates in Water and Effect of Sulphate and Carbonate Ions on their Solubilities", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *Proceedings of the Symposium "Arsenic Metallurgy Fundamentals and Application", Phoenix, Arizona, 1988,* The Minerals, Metals & Materials Society, Warrendale, PA 1988, pp. 77 – 98.
142. E. Krause, V. A. Ettel: "Ferric Arsenate Compounds: are they Environmentally Safe? Solubilities of Basic Ferric Arsenates", in "Impurity Control and Disposal", Proceedings of the 15th annual CIM Hydrometallurgy Meeting, Vancouver 1985, 5/1-5/20.
143. N. Papassiopi, M. Stefanakis, A. Kontopoulos: "Removal of Arsenic from Solutions by Precipitation as Ferric Arsenats", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Application", Phoenix, Arizona, 1988,* The Minerals, Metals & Materials Society, Warrendale, PA 1988, pp. 321 – 334.
144. G. B. Harris, S. Monette: "The Stability of Arsenic-Bearing Residues", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Application", Phoenix, Arizona, 1988,* The Minerals, Metals & Materials Society, Warrendale, PA 1988, pp. 469 – 488.
145. E. Krause, V. A. Ettel: "Solubilities and Stabilities of Ferric Arsenate Compounds", *Hydrometallurgy,* **22** (1989) 311 – 337.
146. E. Vircikova et al.: "Application of Precipitation Processes for Removing of As(V) from Solutions by Ca(III) and Fe(III) Ions", EMC'94, GDMB, Clausthal-Zellerfeld 1994, 190 – 202.
147. G. B. Harris, S. Monette: "The Disposal of Arsenical Solid Residues", in M. Koch, J. C. Taylor (eds.): *"Productivity and Technology in the Metallurgical Industries",* The Minerals, Metals & Materials Society, Warrendale, PA 1989, pp. 545 – 559.
148. G. van Weert, D. J. Droppert: "Aqueous Processing of Arsenic Trioxide to Crisalline Scorodite", *JOM* June (1994) 36 – 38.
149. N. Napasiopi et al.: Cristalline Ferric Arsenates; Structure and Stability Studies", EMC'94, GDMB, Clausthal-Zellerfeld 1994, 203 – 212.
150. J. Geldart, R. Williamson, P. Maltby: "Aqueous Pressure Oxidation as a Waste Treatment Process – Stabilizing Roaster Wastes", *Hydrometallurgy* **30** (1992) 29 – 44.
151. J. E. Litz: Flue Dusts: an Ideal Feed for Resource Recovery", in R. G. Reddy, W. P. Imrie, P. B. Queneau (eds.): *"Residues and Effluents-Processing and Environmental Considerations",* The Minerals, Metals & Materials Society, Warrendale, PA 1991, pp. 223 – 238.
152. D. Langbein et al.: "Das Arsenproblem in der NE-Metallurgie-Untersuchungen bei der Verarbeitung arsenhaltiger Zinnkonzentrate", *Neue Hütte,* **35** (1990) 16 – 20.
153. C. P. Broadbent: "Heavy Metal Immobilisation in Silicate Slags", in M. Nilmani et al. (eds.): *"Pyrometallurgy for Complex Materials and Wastes",* The Minerals, Metals & Materials Society, 1994, pp. 93 – 103.
154. S. Jahanishahi et al.: "The Safe Disposal of Toxic Elements in Slags", in M. Nilmani et al. (eds.): *"Pyrometallurgy for Complex Materials and Wastes",* The Minerals, Metals & Materials Society, 1994, pp. 105 – 119.
155. R. G. Robins: "Arsenic Hydrometallurgy", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 215 – 148.
156. W. Hopkin: "The Problem of Arsenic Disposal in Nonferrous Metals Extraction", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 363 – 384.
157. W. Bartlett: "Separation of As and Ga from Mixed GaAs Electronic Scrap", in R. G. Reddy, J. L. Hendrix, P. B. Queneau (eds.): *"Arsenic Metallurgy Fundamentals and Applications",* The Metallurgical Society, Warrendale, PA 1988, pp. 385 – 400.
158. "MAK- und BAT-Werte-Liste 1998", *Deutsche Forschungsgemeinschaft, Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 34,* WILEY-VCH, Weinheim 1998, pp. 26, 133.
159. L. S. Goodman, A. Gilman: *The Pharmacological Basis of Therapeutics,* 5th ed., MacMillan Publ. Co., New York 1975, p. 925.
160. B. L. Vallee, D. D. Ulmer, W. E. C. Wacker, *AMA Arch. Ind. Health* **21** (1960) 132.
161. H. P. Morris, E. P. Laug, H. J. Morris, R. L. Grant, *J. Pharmacol. Exp. Ther.* **64** (1938) 420.

162. M. J. Ellenhorn, D. G. Barceloux: *Medical Toxicology, Diagnosis and Treatment of Human Poisoning;* Elsevier, New York – Amsterdam – London, 1988.
163. F. T. Hunter, A. F. Kip, J. W. Irvine, *J. Pharmacol. Exp. Ther.* **76** (1942) 207.
164. IARC, International Agency for Research in Cancer, *Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Human,* **vol. 23,** Lyon, France, 1980, p. 39.
165. T. J. Smith, E. A. Crecelius, J. C. Reading, *EHP Environ. Health Perspect.* **19** (1977) 89.
166. H. A. Satterlee, *AMA Arch. Ind. Health* **17** (1958) 218.
167. K. Tsuchiya, N. Ishinishi, B. J. Fowler in: "Toxicology of Metals," EPA-600/1-77-022, *Environ. Health Effects Ser.* **2** (1977) 30.
168. S. Delepine, *J. Ind. Hyg.* **4** (1923) 346, 410.
169. P. Vigne, *Ann. Dermatol. Siphiligr.,* VIIe. Sér., 1 (1930) 1150.
170. I. Holmquist, *Acta Derm. Venereol.* **31**, Suppl. 26, (1951) 1.
171. R. M. Watrous, M. B. McCaughey, *Ind. Med.* **14** (1945) 639.
172. R. A. Mees, *JAMA J. Am. Med. Assoc.* **72** (1919) 1337.
173. S. S. Pinto, C. M. McGill, *Ind. Med. Surg.* **22** (1953) 281.
174. H. Lüchtrath, *Dtsch. Med. Wochenschr.* **97** (1972) 21.
175. K. H. Butzengeiger, *Klin. Wochenschr.* **19** (1940) 523.
176. F. S. Glazener, J. G. Ellis, P. K. Johnson, *Calif. Med.* **109** (1968) 158.
177. K. H. Butzengeiger, *Dtsch. Arch. Klin. Med.* **194** (1949) 1.
178. S. Yeh, *Natl. Cancer Inst. Monogr.* **10** (1963) 81.
179. W. D. Buchanan: *Toxicity of Arsenic Compounds,* Elsevier, Amsterdam 1962.
180. R. A. Kyle, G. L. Pease, *New Engl. J. Med.* **273** (1965) 18.
181. M. G. Ott, B. B. Holder, H. L. Gordon, *Arch. Environ. Health* **29** (1974) 250.
182. S. S. Pinto, V. Henderson, P. E. Enterline, *Arch. Environ. Health* **33** (1978) 325.
183. E. Moss, W. R. Lee, *Br. J. Ind. Med.* **31** (1974) 224.
184. L. Manzano, E. E. Tello, *Rev. Fac. Cienc. Med. Cordoba* **13** (1955) 133.
185. J. M. Borgoño, P. Vicent, H. Venturino, A. Infante, *EHP Environ. Health Perspect.* **19** (1977) 103.
186. W. P. Tseng, *EHP Environ. Health Perspect.* **19** (1977) 109.
187. O. Axelson, E. Dahlgren, C. D. Jansson, S. O. Rehnlund, *Br. J. Ind. Med.* **35** (1978) no. 1, 8 – 15.
188. K. Mabuchi, A. M. Lilienfeld, L. M. Snell, *Arch. Environ. Health.* **34** (1979) no. 5, 312 – 320.
189. T. G. Rossman, Metal Mutagenesis, in A. Goyer, M. G. Cherian (eds.): *Toxicology of Metals, Biochemical Aspects,* Springer-Verlag, Berlin – Heidelberg – New York, 1995
190. H. Nishioka, *Mutat. Res.* **31** (1975) 185.
191. G. W. R. Walker, A. M. Bradley, *Can. J. Gen. Cytol.* **11** (1969) 677.
192. J. K. Petres, K. Schmid-Ulrich, V. Wolf, *Dtsch. Med. Wochenschr.* **95** (1970) 79.
193. R. Happle, H. Hoehn, *Clin. Genet.* **4** (1973) 17.
194. V. H. Ferm, S. Carpenter, *J. Reprod. Fertil.* **17** (1968) 199.
195. V. H. Ferm, A. Saxon, *Experientia* **27** (1971) 1066.
196. V. H. Ferm, A. Saxon, B. W. Smith, *Arch. Environ. Health* **22** (1971) 557.
197. V. H. Ferm, *Rev. Environ. Health* **1** (1974) 238.
198. G. Luego, G. Cassady, P. Palmisano, *Am. J. Dis. Child.* **117** (1969) 328.
199. S. Nordström, L. Beckman, I. Nordensen, *Hereditas* **90** (1979) 291.
200. S. Nordström, L. Beckman, I. Nordensen, *Hereditas* **90** (1979) 297.
201. M. S. Golub, M. S. Macintosh, N. Baumrind, *J. Toxicol. Environ. Health. B. Crit. Rev.* **1** (1998) no. 3, 199 – 241.
202. D. Macaway, D. A. Stanley, *Br. J. Ind. Med.* **13** (1956) 217.
203. S. S. Pinto, S. J. Petronella, D. R. Johns, M. F. Arnold, *Arch. Ind. Hyg. Occup. Med.* **1** (1950) 437.
204. W. J. Levinsky, R. V. Smalley, P. N. Hillyer, R. L. Shindler, *Arch. Environ. Health* **20** (1970) 436.
205. S. F. Dudley, *J. Ind. Hyg.* **1** (1919) 215.
206. F. M. R. Bulmer, H. E. Rothwell, S. S. Polack, D. W. Stewart, *J. Ind. Hyg. Toxicol.* **22** (1940) 111.
207. F. Flury, F. Zernik: *Schädliche Gase,* Springer Verlag, Berlin 1931, reprint 1969.
208. Deutsche Forschungsgemeinschaft (eds.): *Maximale Arbeitsplatzkonzentrationen (MAK)* 1996, VCH Verlagsgesellsch., Weinheim 1996.
209. IARC, International Agency for Research in Cancer, *Overall evaluations of carcinogenicity: An updating of IARC monographs volumes 1 – 42,* Suppl. 7, Lyon, France, 1987.

210. LAI, Länderausschuß für Immissionsschutz, *Beurteilungsmaßstäbe zur Begrenzung des Krebsrisikos durch Luftverunreinigungen;* Ministerium für Umwelt, Raumordnung und Landwirtschaft des Landes Nordrhein-Westfalen, 1991.
211. ACGIH, American Conference of Governmental Industrial Hygienists, *Threshold Limit Values and Biological Exposure Indices,* Cincinnati, USA, 1991 – 1992.
212. Progress Report for Contract No. PH-43-64-886, National Cancer Institute, Institute of Chemical Biology, Univ. of San Francisco, Sept. 1970.
213. G. Joachimoglu, *Biochem. Z.* **70** (1915) 144.
214. G. Bünemann, W. Klosterkötter, *Int. Arch. Gewerbepathol. Gewerbehyg.* **20** (1963) 21.
215. F. Flury, *Z. Gesamte Exp. Med.* **13** (1921) 523.
216. National Defense Research Committee, Office of Scientific Research and Development, Progress Rep. NDCrc-132, Aug. 1942.
217. J. W. E. Harrison, E. W. Packman, D. D. Abbott, *AMA Arch. Ind. Health* **17** (1958) 118.
218. D. T. Teitelbaum, L. C. Kier, *Arch. Environ. Health* **19** (1969) 133.
219. G. G. Parish, R. Glass, R. Kimbrough, *Arch. Environ. Health* **34** (1979) 224.
220. C. W. Muehlberger, *J. Pharmacol. Exp. Ther.* **39** (1930) 246.
221. K. W. Franke, A. L. Moxon, *J. Pharmacol. Exp. Ther.* **58** (1936) 454.
222. R. D. Hood, S. L. Bishop, *Arch. Environ. Health* **24** (1972) 62.
223. *Assoc. Food Drug Off. U.S. Q. Bull.* **15** (1951) 122.
224. M. N. Baxley, R. D. Hood, G. C. Vedel, W. P. Harrison, G. M. Szczech, *Bull. Environ. Contam. Toxicol.* **26** (1981) 749.

Artificial Leather → Leather Imitates
Artificial Sweeteners → Sweeteners

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Artists' Colors

Standard Article

Hugo Müller[1], Wolfgang Müller[2], Manfred Wehner[3], Heike Liewald[4]

[1]formerly H. Schmincke & Co., GmbH & Co. KG, Feinste Künstlerfarben, Erkrath, Germany

[2]H. Schmincke & Co., GmbH & Co. KG, Feinste Künstlerfarben, Erkrath, Germany

[3]Gustav Grolman GmbH & Co. KG, Neuss, Germany

[4]Verband der Mineralfarbenindustrie e.V., Frankfurt/Main, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a03_143.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Organic Pigments
3. Inorganic Pigments
3.1. White Inorganic Pigments
3.2. Colored Inorganic Pigments
3.3. Black Pigments
3.4. Fillers
4. Dyes for Silk Painting
5. Binders
6. Quality Factors
7. Commercially Available Artists' Colors and Materials
7.1. Oil Colors
7.2. Acrylic Paints
7.3. Watercolor (Aquarell)
7.4. Pastels
7.5. Gouache and Distemper (Tempera)
7.6. Powdered Pigments
7.7. Silk Paints
7.8. Paints for Graphic Techniques
7.9. Other Artists' Materials

Artists'; colors represent just a small sector of the paint industry, the total invoiced sales in Europe were about 222.7 × 10^6 € in 2004, the production of artists' colors in Germany amounted to about 4180 t with a value of 67.1 × 10^6 € in 2004. Artists' colors are mixtures of three major components: pigments, binders, and fillers.

Pigments determine the hue, hiding power, and tinting strength of artists' colors. Binders enable the pigments to adhere to the substrate. Fillers can influence optical properties such as whiteness and hiding power, rheological behavior, mechanical properties, and resistance to weathering and chemicals.

The commercially available artists' colors and materials are the following types: oil colors, acrylic paints, watercolors, pastels, gouache and tempera, powdered pigments, silk paints, paints for graphic techniques, and others artists' materials such as primers and varnishes.

[Top of Page]

1. Introduction

Artists' colors represent just a small sector of the paint industry. In 2004 the production of artists' colors in Germany was about 4180 t with a value of 67.1 × 10^6 € and the total invoiced sales in Europe amounted to about 222.7 × 10^6 € [1], [2]. Artists' colors are mixtures of three major components: pigments, binders, and fillers.

Pigments are defined as colored, black, white, or fluorescent particulate organic or inorganic solids which usually are insoluble in, and essentially physically and chemically unaffected by, the vehicle or substrate in which they are incorporated. They alter appearance by selective absorption and/or scattering of light. Pigments retain a crystal or particulate structure throughout the coloration process [3], [4]. Pigments determine the hue, hiding power, and tinting strength of artists' colors. Natural inorganic pigments have been known since prehistoric times. The principal pigments used were based mainly on iron and manganese oxides, found in the soil as colored muds. Natural ocher was used in the Ice Age as a coloring material. Around 2000 B.C., natural ocher was burnt, sometimes in mixtures with manganese ores, to produce red, violet, and black pigments for pottery. Arsenic sulfide as the first clear yellow pigment, and ultramarine (lapis lazuli) and Egyptian blue cobalt aluminum spinel were the first blue pigments. Terra verte and malachite were the first green pigments. The first modern synthetic pigment was Prussian blue (1704), followed by the synthesis of cobalt blue (1777), Scheele's green, and chrome yellow (1778) [5]. The synthetic materials provide the artist with colorants that are relatively pure and reproducible in color and particle size.

A detailed description of the properties, production and uses of inorganic and organic pigments is given in ⟶ Pigments, Inorganic and ⟶ Pigments, Organic.

Pigments and dyes are categorized according to their generic name and chemical constitution in the Colour Index [6]. Important pigment properties include tinting strength, light fastness, weather resistance, hiding power, transparency, shade, dispersibility, and chroma (⟶ Pigments, Inorganic).

In former times, whole egg, egg yolk, and egg white were used as *binders*. In the Middle Ages it was already known that egg yolk act as emulsifier with oils or dissolved resins. Binders enable the pigments to adhere to the substrate.

Fillers can influence optical properties such as whiteness and hiding power (e.g., synthetic silicates and blanc fixe as partial replacements for white pigments), rheological behavior, mechanical properties, and resistance to weathering and chemicals.

[Top of Page]

2. Organic Pigments

Table 1 lists organic pigments used for artists' colors.

Table 1. Organic pigments used for artists' colors

C.I. name	C.I. number	Pigment type	Lightfastness [a]
Pigment Yellow 1	11680	monoazo	**
Pigment Yellow 3	11710	monoazo	**
Pigment Yellow 65	11740	monoazo	**
Pigment Yellow 74	11741	monoazo	**
Pigment Yellow 83	21108	disazo	**
Pigment Yellow 97	11767	monoazo	**
Pigment Yellow 108	68420	anthrapyrimidine	**
Pigment Yellow 150	12764	azo Ni-complex	***
Pigment Yellow 151	13980	benzimidazolone	***
Pigment Yellow 155		bisacetoacetarylide	***
Pigment Orange 5	12075	β-naphthol	**
Pigment Orange 34	21115	disazo	*
Pigment Orange 43	71105	perinone	***
Pigment Orange 62	11775	benzimidazolone	***
Pigment Orange 67	12915	pyrazoloquinazolone	***
Pigment Orange 73		diketopyrrolopyrrole	***
Pigment Red 9	12460	naphthol AS	**
Pigment Red 83 : 1	58000 : 1	alizarin crimson	*
Pigment Red 112	12370	naphthol AS	**
Pigment Red 122	73915	quinacridone	***
Pigment Red 144	20735	disazo condensation	**
Pigment Red 170	12475	naphthol AS	**
Pigment Red 176	12515	benzimidazolone	**
Pigment Red 179	71130	perylene	***
Pigment Red 188	12467	naphthol AS	**
Pigment Red 202		quinacridone	**

Pigment Red 253	12375	naphthol AS	**
Pigment Red 254	56110	diketopyrrolopyrrole	***
Pigment Brown 25	12510	benzimidazolone	***
Pigment Brown 41		disazo condensation	***
Pigment Violet 19	73900	quinacridone	***
Pigment Violet 23	51319	dioxazine	***
Pigment Violet 29	71129	perylene	**
Pigment Blue 15 : 3	74160	copper phthalocyanine	***
Pigment Blue 16	74100	phthalocyanine	***
Pigment Blue 60	69800	indanthrone	***
Pigment Green 7	74260	copper phthalocyanine	***
Pigment Green 36	74265	copper phthalocyanine	***

[a] see Table 5.

Azo pigments (⟶ Pigments, Organic – Azo Pigments*)* cover mainly the range of yellow, orange, red, violet, and brown shades.

Examples for monoazo pigments are C.I. Pigment Yellow 1, Pigment Orange 5, Pigment Red 112, Pigment Red 176, and disazo pigments include C.I. Pigment Yellow 83, Pigment Orange 34, and Pigment Red 144 (Table 1).

Polycyclic pigments used in artists' colors are:

- Phthalocyanine pigments (e.g., Pigment Blue 15 : 3, Pigment Green 36)
- Quinacridone pigments (e.g., Pigment Red 122, Pigment Violet 19)
- Perylene and perinone pigments (e.g., Pigment Violet 29, Pigment Orange 43)
- Isoindolinone and isoindoline pigments (e.g., Pigment Yellow 110)
- Anthrapyrimidine pigments (e.g., Pigment Yellow 108) Diketopyrrolopyrrole pigments (e.g., Pigment Red 254).

[Top of Page]

3. Inorganic Pigments

All inorganic pigments contain heavy metal constituents with the exception of titanium dioxide, carbon black, and ultramarine pigments.

Table 2 lists inorganic pigments used for artists' colors.

Table 2. Inorganic pigments used for artists' colors [a]

C.I. name	C.I. number	Common name	Composition
Pigment White 1	77597	lead white	$2PbCO_3 \cdot Pb(OH)_2$
Pigment White 4	77947	zinc white	ZnO
Pigment White 6	77811	rutile	TiO_2
Pigment White 18	77220	chalk	$CaCO_3$
Pigment White 19	77002	clay	$Al_2O_3 \cdot SiO_2$
Pigment White 22	77120	blanc fixe	$BaSO_4$
Pigment White 24	77002	alumina hydrate	$Al(OH)_3$
Pigment Yellow 35	77205	cadmium zinc sulfide	$(Cd,Zn)S$
Pigment Yellow 42	77492	synthetic hydrated iron oxide	$FeOOH$
Pigment Yellow 53	77788	nickel antimony titanium yellow (rutile)	$(Ti,Ni,Sb)O_2$
Pigment Yellow 184	-	bismuth vanadate/molybdate	$4BiVO_4 \cdot 3Bi_2MoO_6$
Pigment Orange 20	77202	cadmium sulfide/selenide	$Cd(S,Se)$
Pigment Red 101	77491	synthetic oxide	Fe_2O_3
Pigment Red 108	77202	cadmium sulfide/selenide	$Cd(S,Se)$
Pigment Brown 7	77491	sienna/umber	Fe_2O_3

Pigment Brown 24	77310	chromium antimony titanium (rutile)	$(Ti,Cr,Sb)O_2$
Pigment Brown 33	77503	zinc iron chromite brown (spinel)	$(Zn,Fe)(Fe,Cr)_2O_4$
Pigment Violet 14	77360	cobalt violet phosphate	$CO_3(PO_4)_2$
Pigment Violet 15	77007	ultramarine red/violet	Na,Si,Al,S
Pigment Violet 16	77742	manganese violet	$(NH_4)Mn(P_2O_7)$
Pigment Blue 27	77510	Prussian blue	$NH_4[Fe(II)Fe(III)(CN)_6] \cdot H_2O$
Pigment Blue 28	77346	cobalt blue (spinel)	$CoAl_2O_4$
Pigment Blue 29	77007	ultramarine blue	Na,Si,Al,S
Pigment Blue 33	77112	manganese blue	$BaMnO_4/BaSO_4$
Pigment Blue 35	77368	cerulean blue	Co_2SnO_4
Pigment Blue 36	77343	cobalt chromite blue (spinel)	$Co(Al,Cr)_2O_4$
Pigment Green 17	77288	chrome oxide green	Cr_2O_3
Pigment Green 18	77289	viridian	$Cr_2O_3 \cdot 2H_2O$
Pigment Green 19	77335	cobalt nickel zinc titanite green	$(Co,Ni,Zn)_2TiO_4$
Pigment Green 23	77009	green earth	iron(II) silicate with clay
Pigment Green 26	77344	cobalt chromite green (spinel)	$CoCr_2O_4$
Pigment Green 50	77377	cobalt titanate green (spinel)	Co_2TiO_4
Pigment Black 7	77266	carbon black	almost pure C
Pigment Black 9	77267	bone black	$Ca_3(PO_4)_2/C$
Pigment Black 11	77499	black iron oxide	Fe_3O_4

[a] All pigments have *** lightfastness (see Table 5).

3.1. White Inorganic Pigments

The largest group of the inorganic pigments are the white pigments, and they are among the most important in the artists' palette. The four major white pigments in use today are titanium dioxide (TiO_2, C.I. Pigment White 6), zinc white (ZnO, C.I. Pigment White 4), blanc fixe ($BaSO_4$, C.I. Pigment White 22), and white lead [$Pb(OH)_2 \cdot 2\ PbCO_3$, C.I. Pigment White 1].

White lead dates back to the early Greeks and was, until recently, the principal white oil-colors pigment because of its almost ideal properties when mixed with oil. White lead may discolor on exposure to hydrogen sulfide. Nowadays, especially because of the toxicity, the use of lead white is limited to restoration only.

Titanium dioxide (⟶ Pigments, Inorganic – Titanium Dioxide) occurs in nature in the modifications rutile, anatase, and brookite. Rutile and anatase are produced industrially as the most important pigments in terms of quantity [7]. Titanium dioxide (C.I. Pigment White 6) is of outstanding importance as a white pigment because of its scattering properties, its chemical stability, its biological inertness, and its lack of toxicity. The pigment is frequently coated with colorless organic or inorganic compounds of low solubility to improve its weather resistance, lightfastness, and dispersibility.

Zinc white is a fine white powder, which is amphoteric; it reacts with organic and inorganic acids, and also dissolves in alkalis to form zincates [8]. Acute toxicity: rat, oral, LD_{50} > 15 000 mg/kg [9].

Barium sulfate (blanc fixe) is of lower hiding power than white lead, but has resistance to hydrogen sulfide [10]. In artists' colors barium sulfate is often used to adjust the consistency of the liquid paint.

3.2. Colored Inorganic Pigments

Natural Earth Pigments. Today natural earth pigments used as artists' pigments are green earth (ferrous silicate with clay, C.I. Pigment Green 23) and natural iron oxides. Naturally occurring iron oxides and iron hydroxides have been used as pigments since prehistoric times. They were also used as colorants by the Egyptians, Greeks, and ancient Romans. Of the natural iron oxide pigments, hematite (α-Fe_2O_3) attained economic importance as a red pigment, goethite (α-FeOOH) as a yellow pigment (C.I. Pigment Yellow 43), and the umbers and siennas as brown pigments (C.I. Pigment Brown 7). Their poor tinting strength and less saturated color shade are due to their low purity and broader particle size distributions compared to synthetic pigments. Frequently clays or silica are present in natural earth pigments.

Metal Oxides. Synthetic iron oxide pigments are by far the most important group of colored pigments. They include α-FeOOH (iron oxide yellow, C.I. Pigment Yellow 42), α-Fe_2O_3 (iron oxide red, C.I. Pigment Red 101), and Fe_3O_4 (iron oxide black, C.I. Pigment Black 11). Synthetic iron oxides are produced under controlled conditions by a number of manufacturing processes. They are chemically purer and, because of their higher iron content, exhibit greater brightness than naturally occurring oxides. Depending on their crystalline structure and other physical parameters, iron oxide pigments range in color from various shades of red through ocher to black.

Chromium(III) oxide, which forms the basis of chromium oxide pigments, crystallizes in a corundum lattice [11]. Chromium oxide green pigments (Cr_2O_3, C.I. Pigment Green 17) contain only trivalent chrome. Acute toxicity: rat, oral, LD_{50} > 10 000 mg/kg [9]. Chromium oxide pigments yield a fairly dark olive-green shade. They are largely inert chemically and therefore possess good resistance properties.

Complex Inorganic Color (CIC) Pigments. The term “complex inorganic color pigment” refers to the fact that such pigments are a homogenous chemical phase and replaces the formerly used term “mixed-phase metal oxide pigment”, which gave the false impression that they are mixtures.

Nickel antimony titanium yellow $(Ti,Ni,Sb)O_2$ (C.I. Pigment Yellow 53, acute toxicity: rat, oral, LD_{50} > 5000 mg/kg [9]) and chromium antimony titanium yellow $(Ti,Cr,Sb)O_2$ (C.I. Pigment Brown 24, acute toxicity: rat, oral, LD_{50} > 5000 mg/kg [9]) are rutile pigments. The rutile lattice of titanium dioxide absorbs nickel oxide or chromium(III) oxide as coloring components and antimony(V) oxide to maintain an average cation valency of four.

The cobalt blue pigments are pigments with a spinel structure obtained by partial or complete replacement of metal ions of the $MgAl_2O_4$ spinel by cobalt and chromium (e.g., $CoAl_2O_4$, C.I. Pigment Blue 28, acute toxicity: rat, oral, LD_{50} > 10 000 mg/kg [9]). Incorporation of cobalt and nickel in the inverse titanium spinels MgO_2TiO_4 and Zn_2TiO_4 produces cobalt green pigments (e.g., Co_2TiO_4, C.I. Pigment Green 50, acute toxicity: rat, oral, LD_{50} > 10 000 mg/kg [9]).

All these pigments are chemically and thermally very stable and have outstanding lightfastness and resistance to temperature, chemicals and weathering.

Cadmium Pigments. All cadmium pigments are based on cadmium sulfide (CdS, C.I. Pigment Yellow 37) and exist in a highly stable hexagonal crystal form. Inclusion of zinc yields greenish yellow pigments [(Cd, Zn)S, C.I. Pigment Yellow 35] and inclusion of selenium changes the shades to orange [Cd(S, Se), C.I. Pigment Orange 20], red [Cd(S, Se), C.I. Pigment Red 108] and bordeaux. The terms cadmium yellow and cadmium red have become synonymous with brillant yellow and red shades. Especially because of their excellent lightfastness, cadmium pigments are appreciated as artists' colors. Acute toxicity: rat, oral, LD_{50} > 5000 mg/kg [9].

Ultramarine Pigments. Ultramarine blue is the synthetic equivalent of the naturally occurring semi-precious stone lapis lazuli. The unique blue color is due to polysulfides within a sodium aluminum lattice, $Na_6Al_6Si_6O_{24}(NaS_n)$ (n = 2 – 4). Violet (C.I. Pigment Violet 15) and pink (C.I. Pigment Red 259) ultramarines are derived from ultramarine blue (C.I. Pigment Blue 29) by further oxidation and ion exchange and have a very similar structure.

Ultramarines have excellent lightfastness and heat stability and, apart from their sensitivity to acid, good chemical resistance.

Manganese Violet Pigments. Manganese violet pigments are manganese ammonium pyrophosphates, and are red-shade, bright, clean violets ($NH_4MnP_2O_7$, C.I. Pigment Violet 16). Manganese violets have excellent lightfastness, good heat stability and, apart from their sensitivity to alkalis, good chemical resistance.

The excellent lightfastness and purity of shade, both in mass tone and pale shades, makes manganese violet ideal for use in artists' colors.

Iron Blue Pigments. The term iron blue pigment (C.I. Pigment Blue 27) has largely replaced a great number of older names (e.g., Prussian blue, Berlin blue, Milori blue). These names usually denoted insoluble pigments based on microcrystalline Fe (II)/Fe(III) cyano complexes with the formula $M^IFe^{II}Fe^{III}(CN)_6 \cdot H_2O$, where M^I is potassium, ammonium, or sodium, of which potassium is preferred because it produces excellent hues [11]. Acute toxicity: rat, oral, LD_{50} > 5000 mg/kg [9].

3.3. Black Pigments

Carbon black pigments (C.I. Pigment Blacks 6 and 7) are produced by thermal-oxidative dissociation of aromatic oils, e.g., by the lamp black, furnace black, and gas black processes (⟶ Carbon – Production Processes). A further black pigment used as an artists' color is bone black (C.I. Pigment Black 9), which essentially consists of carbon (10 – 20 %) and $Ca_3(PO_4)_2$ (70 – 80 %) [12].

3.4. Fillers

Fillers are used to modify or influence certain physical properties of paints. Organic pigments with high tinting strength often need a filler to adjust the consistency, hue, and hiding power.

Fillers of importance for artists' colors are natural barium sulfate (baryte), precipitated barium sulfate (blanc fixe), aluminum hydroxide, calcite, precipitated calcium carbonate, and clays.

[Top of Page]

4. Dyes for Silk Painting

Dyes as artists' colors are mainly used to color textiles, especially silk. Silk painting as a technique for producing exquisite articles of clothing and works of art is a tradition that dates back thousands of years, particularly in Asia.

Table 3 gives a list of dyes that are used for silk painting.

Table 3. Dyes for silk painting

C.I. name [a]	C.I. number	Chemical class	Reactive System	Lightfastness, 1/1 S.D. [b]
AY 61		azo		5 – 6
AO 95		disazo		4 – 5
ABr 413		azo metal complex		6 – 7
AB 83	42660	triphenylmethane		3 – 4
AG 81		anthraquinone		6 – 7
AB 222		azo (1 : 2 Cr complex)		7
BY 104		azo		6 – 7
RY 27		monoazo		6
RR 123		monoazo		4
RV 5		monoazo (metal complex)	vinylsulfonyl	6 – 7
RB 29		anthraquinone		5 – 6
RG 21		phthalocyanine		7
RB 113		azo (1 : 2 metal complex)	monochlorotriazinyl	6
DY 157	13965	monoazo		3 – 4
DB 199	sim. to 74190	phthalocyanine		7
DE 1168		azo		5

[a] First letter: A = acid, B = basic, R = reactive, D = direct; second letter: B = blue, Bl = black, Br = brown, G = green, O = orange, V = violet, Y = yellow.
[b] S.D. = standard depth (DIN 53235)

Reactive Dyes. Reactive dyes undergo true chemical bonding with the silk during fixing. Reactive dyes commonly used in silk dyes either contain sulfatoethylsulfone (SES) groups or organo-halogen anchor groups (mono- and dichlorotriazine groups, difluorchlorotriazine groups). With SES-type dyes, the reaction with the silk takes place with a vinylsulfone group that is formed from the SES group during thermal or alkaline fixing and quickly reacts with the amino groups of the silk fiber. Chloro-organic anchor groups react directly with the fiber, producing chloride as a reaction byproduct.

Acid Dyes. Acid dyes contain one or more anionic sulfonate and/or carboxylate groups and have high affinity towards wool and silk. Fixing mainly results from ionic reactions which take place with the basic or cationized groups in the fibers. Acid dyes can contain various classes of chromophores (azo, anthraquinone, triphenylmethane). Acid metal complex dyes are used for darker shades or where particularly high wetfastness properties are required and for the manufacture of silk screen dyes.

Direct Dyes. Direct (substantive) dyes are highly water soluble anionic dyes based on various classes of chromophores. They are mainly used to color cellulose-based fibers. Direct dyes are fixed onto silk by ionic reaction and hydrogen bonding.

[Top of Page]

5. Binders

For use as a paint or artists' colors pigments must be incorporated into a liquid or pastelike phase which allows it to adhere to a surface.

Besides modern synthetic binders classical, natural binders are used for artists' colors. The thinning power and the rheological characteristics of the paint are important for certain techniques like impasto or thin layer, and they depend on the type of vehicle [13]. Binders which are transparent and colorless and impart high fastness to all physical properties are preferred for artists' colors.

The binder incorporates the pigments and determines the physical and optical properties of the generated film. Physical properties are film strength, elasticity, gloss, adhesion to the ground, and resistance to light, weather, alkali, acid, and heat. Optical characteristics are transparency, hue, and refractive index, which determine the opacity of the pigment.

A paint generally consists of three components:

1. The film-former, a chemical that dries by polymerization, physically by evaporation of solvents, or by oxidation in air.
2. The solvent in which all other materials are dissolved or suspended if the binder is not liquid.
3. Small quantities of additives that perform special functions like accelerating or hindering functions, optimizing drying processes, optimizing rheology and surface properties, and controlling the open time.

Combinations of film-formers are also used. For example, the type binder for tempera paints is a combination of oil and

water-soluble gum arabic (or casein) emulsified with the aid of an emulsifying agent.

Table 4 lists types of binders which are used for artists' colors.

Table 4. Binders for artists' colors

Type of binder	Film former	Solvent	Thinner	Process
Natural wax	beeswax	none		physically
		turpentine	turpentine	evaporation
		white spirit	white spirit	evaporation
Natural drying oil	linseed, sunflower, safflor oil	oil	oil of turpentine, white spirit	chemical reaction with oxygen
Natural resin	dammar, mastic	oil of turpentine, white spirit	oil of turpentine, white spirit	evaporation
	shellac	ethanol	ethanol	evaporation
Synthetic resin	alkyd resins	oil of turpentine, white spirit	oil of turpentine, white spirit	chemical reaction with oxygen
	polymer dispersion	water	water	polymerization
Natural glue	animal-bone, skin, and casein glue	water	water	evaporation
	plant gum arabic and starch glue	water	water	evaporation
Synthetic glue	cellulose ether, carboxymethyl cellulose	water	water	evaporation
Natural mineral	lime, gypsum	water	water	chemical reaction with carbon dioxide (or physically with water); no film is formed
Synthetic mineral	water glass	water	water	chemical reaction with carbon dioxide to give silica; no film is formed

[Top of Page]

6. Quality Factors

The quality of artists' colors and the final fastness of the painting are determined jointly by the binders and pigments.

Pigments differ in their physicochemical and in their optical characteristics. Physically, pigments must withstand light, weather, and heat. Chemically they must resist the action of alkali and acids. To prevent migration, the pigments must be resistant to the binder, this mainly means insoluble in the binder.

Optical properties of pigments are shade, tinting strength, lightfastness, and hiding power. Pigments can fade or darken, mainly under the action of UV radiation, which can break certain chemical bonds or change the molecular constitution. A change in chemical structure means a change of absorption in the visible region of the spectrum and a subsequent loss of tint or change of hue. However, if the pigment absorbs UV light without any change of molecule structure, decomposition of the binder may be prevented [14], [15].

Three systems are used for defining lightfastness. The ASTM standard [16], defines five fastness classes, determined by exposure to xenon lamps or sunlight and color matching. In the blue wool scale method (DIN 53952) a range of eight wool strips are exposed to sunlight or a xenon lamp together with the paint, and after exposure the paint is evaluated according to the change on the wool scale. Colormakers use a star system to indicate the lightfastness of paints to the artist [17].

Table 5 compares the three scales.

Table 5. Comparison of lightfastness scales

ASTM class (D4303/93a)	Star scale	Blue wool scale (DIN 53952)
I	*** highest lightfastness	8 very excellent

		7 excellent
II	** very good lightfastness	7 excellent
		6 very good
III	* sufficient lightfastness	6 very good
		5 good
IV	o little lightfastness	4 quite good
		3 moderate
V		2 little
		1 very little

Opacity depends on the ability of a pigment to scatter light in a paint film. If the pigment does not scatter light, the layer is transparent. If light is strongly scattered, the film is opaque [13].

The light-scattering power of a pigment depends on its particle size and its refractive index relative to that of the medium in which it is dispersed.

[Top of Page]

7. Commercially Available Artists' Colors and Materials

7.1. Oil Colors

Paints made from vegetable oils (drying oils) have been known since the Middle Ages. In the mid-1500s, oil painting on canvas was the standard technique for artists' easel painting. Advantages of these paints include ease of manipulation due to the long open time, the ability to correct the painting after a couple of days, and the wide range of effects that can be produced (transparent to opaque in full range). Colors do not change to any great extent on drying, which takes place by reaction with atmospheric oxygen [15]. Then, the film becomes insoluble in white spirit or turpentine. The drying oil incorporates the pigment particles and forms a rigid and flexible film. The oil also acts as an adhesive, fixing the pigments on the ground. The optical efficacy is shown by the brillancy and the color of the pigments.

The best consistency for oil colors is a smooth, “short and buttery” paste which does not flow of its own accord and is without tacky, stringy, or long characteristics. The brush strokes must be visible in the applied paste.

The drying process of artists' oil colors is influenced by the type of oil and the pigment properties. The proportion of unsaturated fatty acids determines the drying properties. The drying properties increase with increasing content of linoleic and linolenic acids.

Properties of the dried oil color, which are also influenced by the pigment to some degree, include consistency, extent of oxidation, flexibility and hardness, which finally result in durability and stability to the environment and physical or chemical treatment.

Table 6 lists the most common drying oils for oil colors.

Table 6. Percentage fatty acid composition of drying/semi-drying oils used in artist's oil colors

Drying oil	Palmitic/stearic acid	Oleic acid	Linoleic acid	Linolenic acid
Linseed	9	22	17	52
Walnut	8	17	61	14
Soya	14	21	56	8
Safflower	10	13	77	
Poppy seed	10	16	75	
Sunflower	10	28	61	

Dryers influence the drying times of paints by compensating for the varying effects of the pigments. One of the most common dryer is cobalt naphthenate [17]. This dryer is a skin sensitizer, and the presence of more than 1 % of it in oil colors requires health-warning labeling.

Presently, two main types of oil colors are available commercially: traditional colors with drying oils as binders, and resin – oil colors containing dammar, mastic or acrylic resin. Some brands contain beeswax as a stabilizer for increased shelf life, film-forming properties, or consistency.

For further information on the composition of oil colors, see [18].

A new generation of water-miscible oil colors is now on the market. These oil colors can be diluted with water and are

suitable for persons sensitive to solvents. Instead of normal linseed oil, modified oils are commonly used, together with emulsifying agents.

7.2. Acrylic Paints

Artists' colors based on acrylic dispersions were first produced in the 1950s in the United States [20]. They are highly expressive and favored by many contemporary artists. Acrylics are used for impasto and flowing techniques like aquarelle, airbrush, fabric painting, mural painting, silkscreen painting, print making, sculpture, photo retouching, and hard-edge painting. They can be applied to paper, paperboard, canvas, wood, murals, leather, and textiles, which are usually treated with gesso, mediums, or additives. Acrylic paints are produced in various qualities and viscosities in dissolvers, roll mills, or ball mills and available in tubes or jars.

The paints consist of pigments (see Tables 1 and 2), fillers, acrylic dispersions, wetting agents, defoamers, coalescence additives, preservatives, and water [19]. For further specifications of acrylic paints, see [21], [22].

Acrylic dispersions are based on acrylic acid, methacrylic acid and their methyl, ethyl and butyl esters. The most frequently used dispersions for artists' paints contain mainly ethyl and methyl acrylate resins. They can be diluted with water and dry rapidly to a permanent water-insoluble film, which is soluble in certain organic solvents (e.g., dichloromethane).

During the drying process water evaporates and the dispersed acrylic particles coagulate.

This process depends on the minimum film forming temperature, a typical physical property of the resins. Drying temperatures of acrylic paints should not be lower than 8 °C and not exceed 15 – 30 °C [20]. Another important physical property for producers is the glass transition temperature. For mechanism of the coalescence process of acrylic dispersions, see [21].

Acrylic films are durable, flexible, do not discolor, and are easily overpainted. They can be of any thickness and have no tendency to crack. Like water colors, the appearance of the dry film differs from that of the freshly applied paint. Because of the quick drying process all brushes and tools have to be cleaned immediately after use.

An environmental risk of dispersions is posed by their content of emulsifiers. They are classified as slightly water polluting. Liquid paint has to be disposed of as hazardous waste. Preservatives are a human health risk if undiluted, but due to the concentration of 0.2 – 0.3 % in normal use, no health labeling is required.

A special line of acrylic colors is airbrush paint, which is liquid, also based on water, containing either acrylic dispersions or water-soluble acrylic resins and highly dispersed pigments.

7.3. Watercolor (Aquarelle)

Transparent watercolor painting is not associated with any particular period. KOLLER suggested that watercolor painting started in the 1400s [23].

Watercolor is a translucent paint which utilizes the brilliant white of the paper for its whites and pale tints. This definition is not absolutely correct, because the pigments and their typical characteristics influence the properties of the color. The pigments are not normally transparent but are applied in such thinned-out consistency that their effects are nearly as brilliant as those of the naturally transparent pigments. The ability of the paper to absorb and fix pigment particles in its interstices is equally important as the ability of the binder to bind the color to the ground; thus, watercolor is more like a stain [14].

Pigments for watercolors require the highest lightfastness, because there is no other light-absorbing medium on the surface of the paper. Pigments containing lead are not used for the production of watercolors.

Colormakers refer to the transparency and lightfastness of each shade in their product information sheets. For determination and definition of lightfastness, see [24].

SCHINK [25] classified the watercolors as *staining* and *nonstaining* colors. *Staining* colors are pigments of high adhesive power or of extremely fine grain size which stick to the paper or stain it strongly after application. *Nonstaining* colors can be washed off with a moist brush, cloth, or other tools. These pigments have weak adhesive power and good mixing properties with other pigments on the paper. Watercolors consist of finely dispersed pigments in a watery solution of gum arabic and/or dextrin, wetting agents (e.g., oxgall), and fillers. Sugar, glycerol, glucose syrup or different types of nontoxic glycols are used as softeners.

The combination of natural gum and softener determines the solubility of solid watercolors in pans. This system, especially the binder and fillers, guarantees the free flow of the paint, good wetting properties, and a stable pigment dispersion at high dilution [14]. Preservatives are added to prevent the growth of microorganisms.

The dispersing process is normally carried out in roll mills, which allow a high degree of dispersion of the pigment.

Solid watercolors in pans are manufactured by pouring the liquid mass into pans followed by drying or by extrusion of a high-viscosity paste, which is cut to shape.

7.4. Pastels

Compared to other painting mediums, pastels allow very pure and simple paintings — very similar to painting with pure pigments on a rough ground. The pigments adhere solely by adhesion without a binder.

Pastels are made from a colored watery paste, which is kneaded and then molded into a crayon shape. Only a weak binding

medium is used. To obtain a moldable paste, the pigments are ground together with clay or light spar and a solution of the binder.

Pastels are defined as soft or hard. Commonly gum tragacanth, gum arabic, sugar, gelatin, casein, aqueous oat extracts, and wax emulsions are used as binders.

Lead and cadmium pigments should not be used in pastel sticks due to the dust that can be generated during application and to avoid skin contact when artists blend the colors with their fingers.

The range of a complete pastel series starts with pure colors, followed by lighter and darker shades. FLATTMANN [26] gives an example of such a series of shades:

M: 80 % white added
K: 60 % white added
H: 40 % white added
F: 15 % white added
D: the pure color
C: 15 % black added
A: 40 % black added

The effect of pastel painting is achieved by diffusion of the light scattered from the surface. This results in an optical problem on fixation of the painting. Fixation reduces the opacity and diffusion of the surface-scattered light and so the painting becomes darker. The use of titanium white or zinc white in combination with the colored pigments, however, compensates for this process and allows easier formulation of the paint [26].

Fixatives for pastels, as described in [14], are ca. 1 – 2 % solutions of a natural resin, like mastic, dammar, Venetian turpentine, or shellac in highly volatile solvent such as diethyl ether, ethanol, or white spirit. Commercially available fixatives are solutions of yellowing-resistant ketones or acrylic resin in alcohol for spray application. In some cases, aqueous fixatives such as gelatin or casein can also be used.

7.5. Gouache and Distemper (Tempera)

Gouache and distemper are described together here because of their similar applications and compositions. The term *tempera* (from the Italian *temperare* = mixing) refers to thin, watery colors based on egg white or gum arabic, as well as for colors with egg yolk as binder [27]. Depending on the nature of the emulsion of aqueous binders and vegetable oils, tempera varies from opaque to transparent. The use and formulation of tempera in the early 1900s is referred to in [28], [29].

Tempera colors are insoluble, but this does not mean that they are absolutely waterproof. This permits overpainting without changing the underlying layers. Under completely dry conditions tempera colors are water-resistant. In contrast to oil paintings, which may become yellow or darken in course of time, tempera paintings remain unchanged. Properly painted tempera pictures are less likely to crack with age than oil paintings, but there are some other restrictions due to the inflexibility of the film of a tempera painting [13].

Tempera binders consist of linseed, sunflower, poppy, or other vegetable oils and/or solutions of dammar, mastic, or Venice turpentine emulsified in watery solutions of gum arabic, casein, or simply egg yolk or whole egg. Oil in water emulsions are also used.

Casein/oil tempera emulsions tend to turn yellow quickly; this can be avoided by adding saponified wax or Venice turpentine instead of oil [13]. Beeswax can also be used as a paste in white spirit (hydrophobic phase) or as emulsifying agent in saponified form [30]. Typical additives in tempera paints are further spikenard, clove oil, nontoxic glycols, emulsifying agents and preservatives. Glycerol and castor oil act as plasticizers. Tempera paints are produced in ball mills, dissolvers, or roll mills.

Tempera paints are applied to paper, paperboard, canvas, and wood which has been pretreated with a primer. Tempera painting is still often combined with oil painting, most like a preliminary coating (imprimatura) [14], [27]. The grounds are based on chalk, marble, or light spar as well as pure glue or half-oil grounds [30]. Tempera painting can be with or without varnish.

Gouache (from the Italian *guazzo* = puddle) paints consist of glue binder, pigment, opaque fillers, and water. They are opaque watercolors with the same basic binding medium and mainly the same pigments as aquarelles. Opacity is obtained by adding white pigments or calcium carbonate [31].

In the late 1700s in England, aquarelles was called “watercolor” and clear gouache was called “body color” [32]. In the 1700s and 1800s gouache painting played an important role in European art, especially in Italy and France [33]. Gouache paint is applied in thicker layers than aquarelles. The binding medium contains softeners, but the risk of cracking of the painting cannot be avoided, neither with synthetic resins (polyvinyl compounds) nor with gum arabic and dextrin.

Finger paints for children are water-miscible and nonwaterproof-drying colors which can be easily removed from paper and other grounds with water and tissues. Finger paints are subject to government regulations and must not contain toxic substances. A positive list of applicable binders, fillers, organic solvents, detergents, and embittering agents is part of European standards. Pigments and preservatives must be permitted for use in food or cosmetics. The migration of certain elements (Sb, As, Ba, Cd, Cr, Pb, Hg, and Se) is also restricted [34].

7.6. Powdered Pigments

The use of pure powdered pigments is one of the oldest methods of artwork. All pigments used in artists' colors are commercially available and can be used with different binders for self-made artists' colors.

Powdered pigments are mainly used in mural painting (fresco) and encaustic painting. There are three different mural-painting techniques [13]. The traditional fresco buon process involves painting on a wet, freshly prepared lime-plaster wall, with pigments dispersed in water only. The pigment particles are cemented on the lime surface in the same manner as the lime particles bond with each other and with sand. When the plaster dries, the calcium hydroxide reacts with atmospheric carbon dioxide to form insoluble calcium carbonate. Therefore, only lime-resistant pigments can be used for this technique, mainly inorganic pigments.

Fresco secco painting is performed on a finished, dried lime-plaster wall with pigments in an aqueous binding medium. The wall is impregnated with limewater, and the pigments are dispersed in water and a solution of casein. Secco painting must be applied in a thin layer, and the pigments must be resistant to lime.

The silicate technique eliminates the disadvantage of traditional fresco painting, that is, its sensitivity to atmospheric pollutants such as sulfuric acid. Potassium water glass is used as the binder. Only alkali-resistant pigments and water glass are used. Water glass reacts with the atmospheric carbon dioxide to give silica and calcium silicate, which form a water-permeable layer. The technique was discovered by KEIM in Munich, and “Keim's silicate paint” is still available today.

In the encaustic technique, the powdered pigments are mixed with beeswax and a small amount of dammar and melted to a pasty waxy paint. This is applied to canvas with special heated tools or to plaster-walls and results in a painting of high tinting strength and high fastness to humidity and air pollution.

7.7. Silk Paints

Paints Fixed by Ironing. Once the paint has dried, it is fixed to the silk with a hot iron at ca. 150 °C. The main advantage of these widely used iron-on paints is the simple method of fixing, which can be carried out at home, and the higher lightfastness of the pigments used in this type of silk paints. The disadvantages are limited flow behavior, poor solubility, and a somewhat stiffer texture of the silk afterwards.

Commercial iron-on paints consist mainly of aqueous pigment dispersions finished with dispersants and binders based on polyacrylics or polyurethane.

Dyes Fixed by Steam. After drying, these dyes are fixed in special steaming equipment. Those involved in silk screening as a hobby without access to such equipment can use a domestic steam pressure cooker as an alternative.

The main disadvantage associated with steaming lies in the method of fixing itself. The advantages of steam-fixed dyes are very good flow behavior, good solubility, high brilliance, good fastness, and the very soft texture of the painted silk. Silk dyes for steam fixing mainly consist of aqueous solutions of reactive or acid dye.

7.8. Paints for Graphic Techniques

Less common in the field of artistic techniques but nevertheless of importance are printed graphics.

Copperplate engraving is rarely used today, but wood- and linocut, etching, and lithography are still employed. For most of these applications the paint is based on oleic binders, and only for linocut is an aqueous paint used. Linoleum does not permit very delicate cutting, and it is characterized by blocky or posterlike forms. The technique has a certain popularity among artists and in schools for the production of more casual work. Many artists of the 1900s like Matisse, Picasso, Vlaminck and Rholfs preferred it to the more gnarled woodcut. Waterbased linocut paints are preferred for this application. The use of commercial printing inks is not recommended [13]. These waterbased paints are similar in composition to aquarelle and gouache paints, and certain properties are intermediate between the two, but the drying time of this type of paint is considerable longer than that of aquarelle or gouache. It contains organic and inorganic pigments, gum arabic, fillers, wetting agents, preservatives, glycerol, and nontoxic glycols. The characteristic effect of the colors is mainly semiopaque or semitransparent.

Lithography (⟶ Imaging Technology – Lithography (Planographic Printing, Offset)) requires various artists' materials: gum arabic, oil of turpentine, shellac, asphalt, rosin, talc, and others. Drawing materials are crayons, watercolor crayons, waxchalk, chalk, and India ink. India inks contain shellac (solved with borax in water), carbon black, and water. Others additionally contain beeswax, paraffin, or soap [35].

The printing ink was based on boiled linseed oil, linseed stand oil, and/or alkyd resins. Formerly artists produced the paint themselves from pigment and linseed oil. Today, offset printing oil is generally used.

7.9. Other Artists' Materials

Primers are used to seal porous surfaces before applying colors. Priming provides a smooth surface, which ensures that the subsequent paint layers adhere firmly. Primers are based on alkyd resins, acrylic emulsions, or linseed oil. They are pigmented with titanium dioxide or chalk. In oil painting, canvas, hardboard, bare wood, and other grounds are glue-sized before the primer is applied.

Solvents are used to dilute colors and clean up tools and materials. Solvents normally evaporate completely after application. The most widely used are white spirit, freshly distilled oil of turpentine, petroleum distillates, as well as brush cleaners and paint removers created as different mixtures by paint makers.

Mediums are used to improve the physical properties and behavior of a color. They are essential for certain techniques and are mixed in certain proportions with the paint. Mediums contain the same binder as the paint system or a compatible binder:

alkyd resins, dammar, acrylic resins or ketone resins for oil colors; acrylic emulsions for acrylic paints; gum arabic, starch, dextrin, and others for watercolors and gouaches. Some additionally contain thickening agents or other additives.

Varnishes provide a clear, transparent coat which gives substantial protection against environmental influence, but especially against dirt. The optical properties strongly influence the final appearance of a picture. Modern synthetic varnishes do not produce the same amount of gloss and color saturation as traditional varnishes. Molecular mass and refractive index are factors determining the optical properties of varnishes [36]. Changes in solubility, brittleness, crosslinking, and others result from unavoidable ageing. Resins for varnishes are mastic, dammar, acrylic, and ketone resins, and in some cases poly(vinyl acetate) and low molecular mass hydrocarbon resins. They are dissolved in white spirit, oil of turpentine, or ethanol.

Modern commercial varnishes contain UV-absorbing substances.

Retouching varnish can be applied after drying of oil paintings to give a temporary protection, but mainly for making a uniform surface between several layers of oil paint. After 10 or 12 months a normal picture or final varnish can be used to give the intended final protection. Beeswax or synthetic waxes and inorganic matting agents are added to varnishes for matte effects.

Fixatives are dilute solutions of acrylic and ketone resins in white spirit or of shellac in ethanol for fixation of drawings made with pencil, charcoal, chalk, pastel, and aquarell and gouache paintings.

Canvas is a flexible support woven from cotton, linen, or polyester yarn. The durability and stability of canvas depend on the type of yarn and the compactness and grain (texture) of the fabric. Normal canvas is already treated with an acrylic primer, but it can also be prepared in the traditional way with aqueous glue (size) followed by a stiff gesso which consists of chalk or light spar, linseed oil, and some aqueous glue.

Good quality watercolor *paper* is manufactured from cotton fibers, which impart strength and durability. It must be acid-free. Gelatin sizing gives a strong paper which is recommended for a wide range of water-based media (also gouache and acrylic).

Other types of papers of different weight and thickness, different sizing, grain, texture, as well as vellum and parchment, are available for gouache, acrylic, and oil techniques. Normally they should be prepared with gesso; oil paint in particular needs an isolating ground to avoid absorption of oil by the paper.

There is a wide range of artists' *brushes* for all painting techniques and materials, even for pastels. Most brushes are made from natural bristle, natural hair, or synthetic fibers intended to simulate the characteristics of the natural fibers. Aquarelle and wash brushes are maufactured from finest pure red sable, but ox, pony, and goat hair are also used. For oil, acrylic, or gouache painting the artist can use stroke brushes of different types of hair and a range of hog brushes.

Other accessories for artists are easels, palettes, palette knives, tiles, saucers, dippers, charcoal, rubbers, pencils, cutters, and tools for calligraphy, etching, lithography, and linocut.

[Top of Page]

References

1. Statistisches Bundesamt.
2. CEPE – European Council of the Paint, Printing Ink and Artists' Colours Industry 7, 2004
3. DCMA, *American Inkmaker*, June 1989.
4. ETAD/BCMA/VdMI/EPSOM Safe Handling of Pigments, European Edition 1995.
5. G. Buxbaum, *Industrial Inorganic Pigments*, VCH Verlagsgesellschaft, Weinheim, Germany 1993.
6. *Colour Index*, 3rd ed., vol. **3**, The Society of Dyers and Colorists, Bradford, The American Association of Textile Chemists and Colorists, 1982, p. 3318.
7. “Titandioxid”, *Römpps Chemie-Lexikon*, 9th ed., Thieme, Stuttgart 1992.
8. ISO 275; RAL 844: C 2, C 3 1974; ISO-DP 9238 1992.
9. Data from relevant safety data sheets.
10. “Bariumsulfat”, *Römpps Chemie-Lexikon*, 9th ed., Thieme, Stuttgart 1992.
11. G. Benzing *et al.*: *Pigmente und Farbstoffe für die Lackindustrie*, Expert Verlag, Ehningen 1992.
12. “Knochenkohle”, *Römpps Chemie-Lexikon*, 9th ed., Thieme, Stuttgart 1992.
13. R. Mayer: *The Artist's Handbook of Materials and Techniques*, 4th ed., R. Mayer Publications, New York 1985.
14. M. Doerner: *Malmaterial und seine Verwendung im Bilde*, Ferdinand Enke Verlag, Stuttgart 1994.
15. G. P. A. Turner: *Introduction to Paint Chemistry and Principles of Paint Technology*, 3rd ed., Chapman & Hall, London – New York 1988.
16. ASTM D 4303 /93a Standard Test Methods for Lightfastness of Pigments Used in Artists' Paints.
17. K. Wehlte: *Werkstoffe und Techniken der Malerei*, Otto Maier Verlag Ravensburg 1997.
18. ASTM D 4302 96a, Standard for Artist's Oil Paints.
19. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **IV**, Verlag W. A. Colomb in der H. Heenemann GmbH, Berlin-Oberschwandorf 1976.
20. D. Simmert: *Acrylharzkünstlerfarben*, Kunsttechnologie und Konservierung, "Heft 1", Wernersche Verlagsgesellschaft, Worms 1995.
21. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **I/3**, Verlag W. A. Colomb in der H. Heenemann GmbH, Berlin-

Oberschwandorf 1974.
22. ASTM D 5098, Standard Specification for Artists' Acrylic Emulsion Paints.
23. *Reclams Handbuch der künstlerischen Techniken*, vol. **1**, Phillip Reclam jun., Stuttgart 1998, p. 270.
24. ASTM D 5067, Standard Specification for Artists' Watercolor Paints.
25. C. Schink: *Mastering Colour and Design in Watercolour*, Watson-Guptill Publications, New York 1989.
26. A. Flattmann: *The Art of Pastel Painting*, Watson-Guphill Publications, New York 1987.
27. E. Reinkowski-Häfner: *Tempera, Kunsttechnologie und Konservierung*, "Heft 2", Wernersche Verlagsgesellschaft, Worms 1994.
28. R. H. Wackernagel: *Neues zur Maltechnik Wassily Kandinskys, Kunsttechnologie und Konservierung*, "Heft 2", 1997.
29. A. Hoffmeister-zur Nedden: *Zur Maltechnik von Paula Modersohn-Becker, Kunsttechnologie und Konservierung*, "Heft 2", 1997.
30. K. Wehlte: *Temperamalerei*, Otto-Maier-Verlag, Ravensburg 1961.
31. E. Berger: *Quellen und Technik der Fresko-, Oel- und Temperamalerei des Mittelalters*, Reprint Verlag, Vaduz 1982.
32. B. Cohn: *Wash and Gouache*, The Center for Conservation and Technical Studies, Fogg Art Museum and The Foundation of the American Institute for Conservation, Havard College, Cambridge, Mass. 1977.
33. C. Ehrenfort: Aquarell- und Gouachefarben, Diplomarbeit, Institut für Museumskunde an der Staatlichen Akademie der Bildenden Künste, Stuttgart 1993.
34. No. prEN 71-7: 1997 E, “Safety of Toys — Part 7: Finger Paints — Requirements and test methods.”
35. W. Dohmen: *Die Lithographie*, DuMont Buchverlag, Köln 1982.
36. E. Rene de la Rie, “The Influence of Varnishes on the Appearance of Paintings,” *Studies in Conservation* **32** (1987) 1 – 13.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Asbestos

SEYMOUR S. CHISSICK, King's College, London, United Kingdom

1.	**Introduction**	1
2.	**Crystal Structure**	2
2.1.	**Serpentine Minerals**	2
2.2.	**Amphibole Minerals**	4
3.	**Physical Properties**	5
4.	**Chemical Properties**	7
5.	**Occurrence and Extraction**	9
6.	**Uses and Asbestos Substitutes**	10
7.	**Economic Aspects**	12
8.	**Asbestos in the Environment and Waste Disposal**	12
9.	**Toxicology and Occupational Health**	14
9.1.	**Asbestos-Related Diseases**	14
9.1.1.	Inhalation of Asbestos Dust	14
9.1.2.	Pleural Thickening	15
9.1.3.	Asbestosis	15
9.1.4.	Cancer	15
9.1.5.	Mesothelioma	16
9.2.	**Occupational Health**	16
10.	**References**	17

1. Introduction

Asbestos (Greek: inextinguishable) is the common name of a number of naturally occurring, hydrated silicate minerals that possess a crystalline structure and low thermal conductivity. They are incombustible in air and separable into filaments. These minerals occur as extremely thin flexible fibers having great tensile strength and thermal stability, and they can be spun into yarn and made into textiles. There are some 3000 recorded uses of asbestos including "fireproof" textiles, paper and boards, clutch and brake linings, asbestos cement sheets and pipes, flooring and roofing products, electrical and thermal insulating materials, coatings, and heat shields.

Table 1. Classification of varieties of asbestos minerals

Group and Variety	Formula
Amphibole	
Anthophyllite	$Mg_7Si_8O_{22}(OH)_2$
Amosite	
(brown asbestos)	$(Fe^{2+}Mg)_7Si_8O_{22}(OH)_2$
Crocidolite	
(blue asbestos)	$Na_2Fe_2^{3+}(Fe^{2+}Mg)_3Si_8O_{22}(OH)_2$
Tremolite	$Ca_2Mg_5Si_8O_{22}(OH)_2$
Actinolite	$Ca_2(MgFe^{2+})_5Si_8O_{22}(OH)_2$
Serpentine	
Chrysotile	
(white asbestos)	$Mg_3[Si_2O_5](OH)_4$

The term asbestos was first used mineralogically in the middle of the 19th century with regard to a fibrous amphibole mineral discovered in the Italian Alps. The six major types of asbestos in use fall into two major groups: serpentine, having a layered silicate structure, and amphibole, having an SiO_4 chain structure. *Chrysotile* (white asbestos) is the sole member of the serpentine group. Chrysotile asbestos accounts for over 90 % of the world's asbestos production. Of the five members of the *amphibole group,* crocidolite (blue asbestos) and amosite (brown asbestos) are widely used commercially and tremolite and actinolite less so (Table 1).

Asbestos has been known and used in small amounts for thousands of years, but it was not widely used prior to the latter part of the 19th century [3]. Asbestos was first mined in the Quebec, Canada, chrysotile fields in 1878, followed by Russia in 1885 and South Africa in 1906.

Within 20 years of the first industrial production of asbestos, the public health hazards associated with asbestos started to come to light. Between 1890 and 1895, 16 out of 17 workers in a French asbestos weaving factory had died; by 1899, 11 men who had worked in an asbestos spinning factory in the United Kingdom had died at about the age of 30, having spent the whole of their working lives in this occupation (cording). The last of these deaths was reported to the Departmental Committee for Compensation for Industrial Diseases (U. K.) in 1906; this was the first recorded case of the disease that has become known as *asbestosis* [5]. The first complete description of asbestosis appeared in 1927 [6], [7]. Asbestosis and other asbestos-related diseases are treated in detail in Chapter 9.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_151

The widespread commercial use of asbestos over the past 100 years has led to its uncontrolled distribution throughout much of the industrialized world and its appearance in the general environment. The concentration of asbestos fibers in the urban atmosphere is actually in the lower part of the range 10 – 100 ng/m^3, but this can rise substantially in certain locations [8–12]. For example, in the vicinity of factories using asbestos, concentrations of asbestos fibers in the external atmosphere up to 5000 ng/m^3 have been reported. In buildings where asbestos is being worked with or damaged, the concentration of asbestos in the air can be as high as 800 ng/m^3 [11], [13], [14]. Low concentrations of asbestos are also found in water, beverages, and pharmaceutical and other medical products [9], [11], [15–23].

All commercial asbestos minerals are hydrous silicates of either the serpentine or the amphibole group. The serpentine minerals, which include chrysotile, have relatively simple chemical formulas; e.g., for chrysotile it is $Mg_3[Si_2O_5](OH)_4$. (See Figs.1,2,3) Amphiboles have a widely varying nature and they are best considered by the general formula $A_{0-1}X_2Y_5Z_8O_{22}(OH, F)_2$; the substituents are: A = sodium or potassium; X = sodium or calcium; Y = magnesium, iron, or aluminum; Z = aluminum or silicon. It is not uncommon that small amounts of elements such as chromium, nickel, and titanium occur; usually many other elements are present in trace amounts. Minerals can have wide ranges of composition but still possess their characteristic structure and general formulas.

Figure 1. Asbestos rock

Figure 2. Close-up of asbestos rock

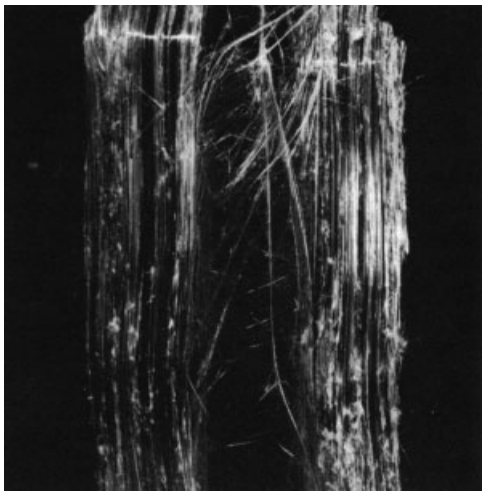

Figure 3. A piece of chrysotile asbestos rock showing the fibrous nature of the mineral
Reproduced by permission of the Asbestos Information Centre, London

2. Crystal Structure

2.1. Serpentine Minerals

Chrysotile is a serpentine mineral, and these have layered silicate structures. The layers are made up of silicon atoms, each surrounded by four oxygen atoms at the corners of an almost regular tetrahedron. The magnesium atoms are also surrounded by six oxygen atoms, each at the corner of an almost regular octahedron. The SiO_4 tetrahedra, lying on their triangular bases and with the apices all pointing in the same direction, are linked by sharing all basal oxygen

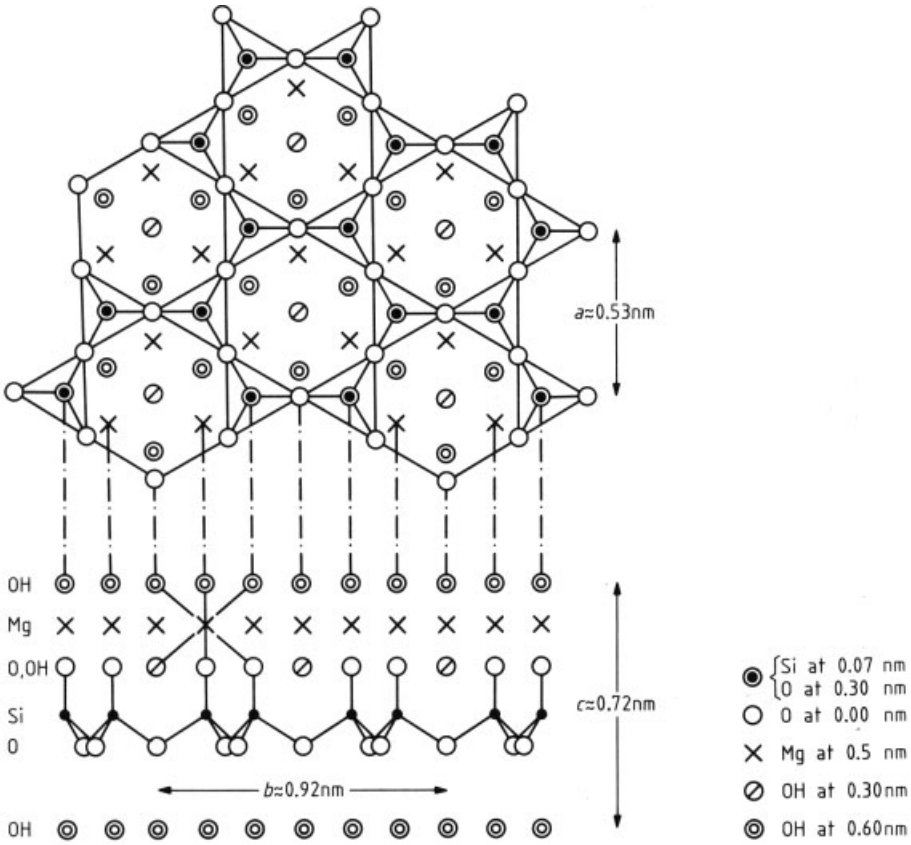

Figure 4. Plan and elevation view of idealized structure of serpentine minerals
Reproduced by permission [1]

atoms to form a continuous layer. Gaps in the plane formed by apical oxygen atoms are filled by OH ions to form a regular and approximately close-packed array of (O, OH) at this level. Lying above this plane is an array of magnesium atoms and above these a plane of OH ions such that the magnesium atoms are surrounded by octahedra of (O, OH) ions. Each Mg has a triangle of 3 (OH) above it and a triangle below of 2 (O) and 1 (OH), rotated by 60 degrees with respect to the first. The structure is best understood by the plan and elevation view in Figure 4. The chemical content of a unit cell is $Mg_3Si_2O_5(OH)_4$.

In building up the serpentine structure, composite layers of the type described above are superimposed; the interlayer distance is approximately 0.75 nm. However, one layer may be placed directly above another or there may be displacement or rotation between layers. This leads to the possibility of different stacking frequencies, different symmetries, and multilayer unit cells. One-, two-, three-, six-, and nine-layer unit cells are known. The symmetry may be trigonal, hexagonal, or monoclinic. Disordered stacking can also occur.

The structure described above is that of lizardite. These minerals are mostly fine grained and have platy morphology but some have lath-like structures with elongation parallel to the *a* axis.

The asbestiform morphology of chrysotile is not the result of an obviously layered crystal structure. High-resolution electron micrography shows that in chrysotile the structured layers are curved about the direction of the *a* axis to form either scrolls or concentric circled tubes [24]. Each scroll is formed from a closely connected double layer having magnesium hydroxide units on its external face and silica units on its inner face. The details of a small section of the scroll

show the structure of the double layer and of the unit cell based on $Mg_3(Si_2O_5)(OH)_4$ (Fig. 5).

The diameters of the tubes are about 20 nm. A sliver of chrysotile asbestos with a cross section of 0.1 mm^2 contains about 2×10^7 tubular fibrils all in approximately parallel orientation. It is therefore possible to strip very fine threads from an asbestos fiber bundle, each of which contains many thousands of fibrils.

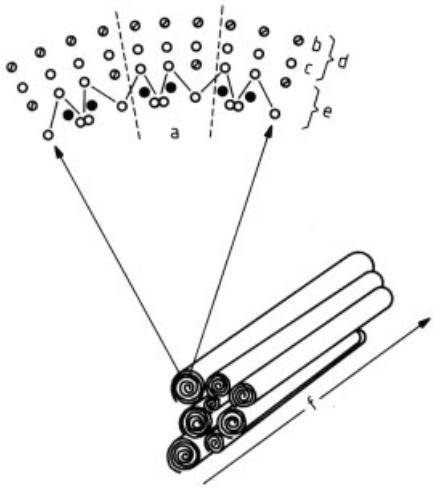

Figure 5. Schematic diagram of the structure of a chrysotile fiber formed of several scrolls of individual crystallites
a) Unit cell; b) Hydroxide ions; c) Magnesium ions; d) Magnesium hydroxide layer; e) Silica layer; f) Fiber axis Reproduced by permission [1]

The reason for the curving of the layers is the polar and composite structure of the fundamental sheet. The silicon component tends to have a smaller repeat dimension than the magnesium component, and the mismatch between the two can be overcome by bending, with the magnesium-containing octahedra on the outside of the curve.

The growth of very thin fibrils is more readily understood for chrysotile than for amphibole asbestos. When chrysotile tubes are formed, one radius of curvature may be the most stable, and this may impose an upper limit on the diameter of the tubes.

2.2. Amphibole Minerals

The fundamental unit of the amphibole structure is the chain of SiO_4 tetrahedra linked by corner oxygen atoms as shown in Figure 6. The characteristic chain formula is Si_4O_{11}, and the repeat distance along its length is about 0.53 nm. The chains are four tetrahedra wide and are very long. Fibrils of asbestiform amphiboles run parallel to the chain length.

In amphiboles the Si_4O_{11} chains are linked laterally by cations such as magnesium and calcium. In tremolite the magnesium ions link chains by strips of Mg(O,OH) octahedra. The calcium ions link neighboring chains, and they occur in distorted polyhedra of oxygen atoms.

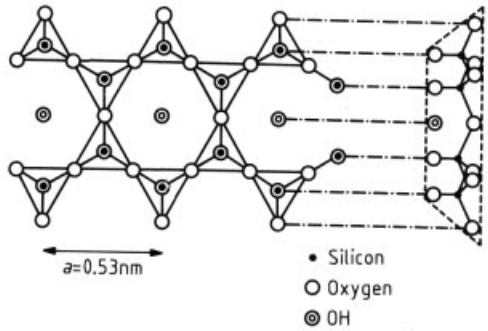

Figure 6. Plan and end view of an idealized amphibole chain (Si_4O_{11}) plus additional OH ions
Reproduced by permission [1]

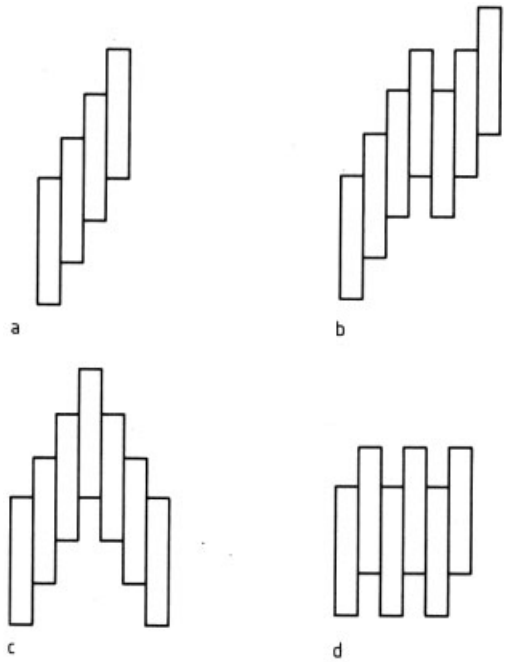

Figure 7. Schematic illustration of possible different modes of stacking of neighboring blocks of amphibole structure
a) Monoclinic; b) Monoclinic with stacking fault; c) Monoclinic twin; d) OrthorhombicReproduced by permission [1]

Table 2. Physical properties of asbestos varieties

Property	Chrysotile	Anthophyllite	Amosite	Crocidolite	Tremolite	Actinolite
Tensile strength, kg/cm^2	31×10^3	24×10^3	25×10^3	35×10^3	$<5\times10^3$	$<5\times10^3$
Young's modulus, kg/cm^2	1.65×10^6	1.58×10^6	1.65×10^6	1.9×10^6	–	–
Flexibility	good	fair to brittle	fair	good	brittle	fair to brittle
Specific gravity	2.55	2.85 – 3.1	3.43	3.37	2.9 – 3.2	3.0 – 3.2
Hardness (Mohs)	2.5 – 4.0	5.5 – 6.0	5.5 – 6.0	4.0	5.5	6.0
Specific heat, kJ g^{-1} K^{-1}	1.13	0.89	0.81	0.84	0.89	0.91
Resistivity, MΩ cm	0.003 – 0.15	2.5 – 7.5	up to 500	0.2 – 0.5	–	–
Magnetic susceptibility (mean χ_g at $H = 7.98\times10^5$ A/m)	5.3×10^{-6}	14.3×10^{-6}	78.7×10^{-6}	60.9×10^{-6}	–	–

Table 3. General optical properties of asbestos varieties

Asbestos	Color	Pleochroism	Birefringence	Orientation	Extinction
Chrysotile	white	nil	moderate, 1st order	length slow	parallel
Anthophyllite	white	nil	moderate, low 2nd order	length slow	parallel
Amosite	ash grey	nil	strong, 2nd order	length slow	parallel
Crocidolite	dark blue	grey-blue to dark blue	weak (masked)	length fast	parallel
Tremolite	white	nil	moderate, low 2nd order	length slow	oblique (max. 20 degrees)
Actinolite	dark green	pale green to deep green	moderate, low 2nd order	length slow	parallel

Most amphiboles are monoclinic as a result of the way in which successive chain units are stacked with respect to each other. However, anthophyllite is orthorhombic and the relation of its unit cell to that of a monoclinic structure is shown in Figure 7.

The chainlike structure of amphiboles affords good cleavages along (110) planes. If the chains themselves and their linkage across Y sites are considered the strongest points of the structure, paths of weakness can be traced as in Figure 8, and the average path can be taken as defining a likely cleavage plane.

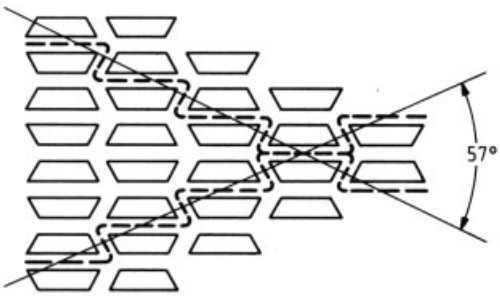

Figure 8. Schematic illustration of amphibole structure as seen along the z axis
Probable paths of weakness are shown, causing cleavages intersecting at approximately 57 degrees.Reproduced by permission [1]

The physical properties of asbestos do not result wholly from to the chainlike structure of amphiboles or their good cleavage. Many amphibole minerals are not asbestiform, although they possess chain structure and good cleavage.

3. Physical Properties

Some of the physical properties of the asbestos varieties are shown in Table 2.

Tensile Strength. One of the most important and most commonly quoted physical properties of asbestos fibers is tensile strength. Measurement is made on specimens 3 – 4 mm long (minimum), using for example the Marsh microtensile testing machine. Under the influence of heat, the tensile strength of the major types of asbestos fibers increases slightly before diminishing rapidly as structural changes and decomposition occurs (Fig. 9).

Surface Area. The most important parameter for industrial applications is surface area. Each process requires asbestos fibers with a critical degree of fiberization, which parallels surface area. For example, chrysotile fibers prepared for use in brake linings have a different degree of fiberization from those prepared for

making asbestos textiles. The required degree of fiberization is achieved by milling the raw asbestos; in multistage manufacturing processes account must be taken of any further fiberization that may occur.

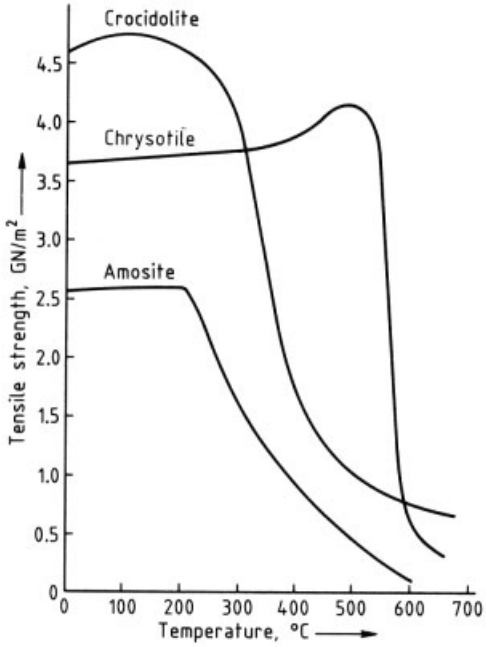

Figure 9. Tensile strength changes in heated asbestos fibers
Reproduced by permission [1]

Ideally, a fiber of cross section area *a* will split into *n* fibers of average area of cross section *a/n* without breaking into shorter lengths. Under these circumstances, the degree of fiberization has increased *n*-fold.

Raw amphibole asbestos, i.e., asbestos shipped to refineries from mines, has an absolute surface area of $6-20\times10^3\,cm^2/g$, whereas for chrysotile, the value may be up to $30\times10^3\,cm^2/g$. The absolute surface area of typical asbestos fibers used in manufacturing processes is $30-90\times10^3\,cm^2/g$, which can be raised to $140\times10^3\,cm^2/g$ by gassing out at high temperature.

Specific Gravity. For the determination of specific gravity, prepared fiber blocks of approximately $1\,cm^3$ are outgassed under water for several hours and then weighed in water.

Electrical Data. The only available electrical parameter on asbestos is volume resistivity, which is low for chrysotile and very variable for amphibole fibers.

Magnetic Susceptibility. Generally magnetic susceptibility is low, but there are considerable differences between that of the high-iron amphiboles and anthophyllite and between the amphiboles and chrysotile.

Optical Properties. General optical properties and refractive indices are compiled in Tables 3, 4.

Table 4. Refractive indices (20°C) of asbestos varieties

Asbestos	n_α	n_γ
Chrysotile		
Canada	1.537 – 1.554	1.554 – 1.557
USSR	1.543 – 1.550	1.551 – 1.555
South Africa	1.549 – 1.553	1.553 – 1.557
Italy	1.541 – 1.553	1.547 – 1.554
Anthophyllite	1.578 – 1.652	1.591 – 1.676
Amosite		
South Africa	1.670 – 1.675	1.683 – 1.694
E.Transvaal		
Crocidolite		
South Africa		
N.W. Cape	1.682 – 1.696	1.686 – 1.700
Transvaal	1.681 – 1.683	1.683 – 1.687
Australia	1.680 – 1.687	1.684 – 1.691
Bolivia	1.659	1.673
Tremolite	1.599 – 1.628	1.591 – 1.676
Actinolite	1.600 – 1.628	1.625 – 1.655

IR Spectroscopic Data. The use of IR spectroscopy in identification and structural studies of asbestos is relatively new. All types of asbestos exhibit strong absorptions around $1200-900\,cm^{-1}$ because of Si–O stretching vibrations and around $600-300\,cm^{-1}$ because of Si–O chain and ring vibrations. Additional peaks below $600\,cm^{-1}$ correspond to cation – oxygen vibrations. In theory, four cation sites for the amphiboles can be distinguished; however, in practice, it is difficult to detect more than two of these peaks. The O–H stretching vibrations are found in the region $3600-3700\,cm^{-1}$; their amplitude corresponds to the OH concentration of the sample.

The three major types of asbestos can readily be identified by peaks at $305\,cm^{-1}$ for chrysotile, $318-320\,cm^{-1}$ for crocidolite, and $330-332\,cm^{-1}$ for amosite. Principal absorption peaks for different types of asbestos are given in Table 5.

Table 5. Principal IR absorption peaks for asbestos minerals (wavenumber, cm^{-1}) Distinguishing strong (s) and week (w) peaks are indicated.

Vibrations	Chrysotile	Anthophyllite	Amosite	Crocidolite	Tremolite	Actinolite
O–H stretching vibrations	3700 (s) 3655 (w)	3680 3675	3660 (w) 3640 3620	3655 (w) 3640 3620	3680 (w) 3670 (w)	3680 (w) 3670 (w)
Si–O stretching vibrations	1080 1030 965	1095 1020 980	1129 1085 1000 890	1145 1110 995 898 880	1105 (s) 1065 (s) 998 (s) 950 (s) 922 (s)	1100 1040 998 953 922
Silicate chain and ring vibrations		780 755 712 670 (s)	775 700 636	780 695 635	755 685	758 685
Cation – oxygen stretching vibrations	610 438 410 305	495 455 (s)	500 480 426 330	545 505 450 320	508 461 390 360	510 460
NaO_8				270		

In environmental samples it is possible to detect 3 % of crocidolite in the presence of amosite. There is a difficulty when IR spectroscopy is used to assess airborne asbestos because a sample must be a minimum of 0.2 mg, which is a considerable weight by airborne sampling techniques.

The decomposition of asbestos by heat can also be followed by infrared spectroscopy as well as differential thermal analysis.

4. Chemical Properties

Thermal Decomposition (Table 6). All forms of asbestos break down progressively to simpler components (e.g., pyroxenes and silica) when heated to temperatures in the range 600 – 1000 °C (see also Fig. 10). Reactions of asbestos, as such, with other materials at high temperature therefore do not exist. In any reactions involving high-melting metals or fused salts, for example, sodium carbonate fusion, primary breakdown of the mineral is assumed to occur first, followed by reaction of the products with the fusing material. At lower temperature asbestos fibers resist attack by molten materials.

Asbestos fibers as such do not have a melting point; only the decomposition products melt. If the latter contain iron pyroxenes, fusion generally occurs between 1000 and 1100 °C, e.g., in the case of amosite and crocidolite. High-magnesium varieties, e.g., chrysotile and anthophyllite, decompose to give magnesium pyroxenes in the products, which melt at about 1450 °C.

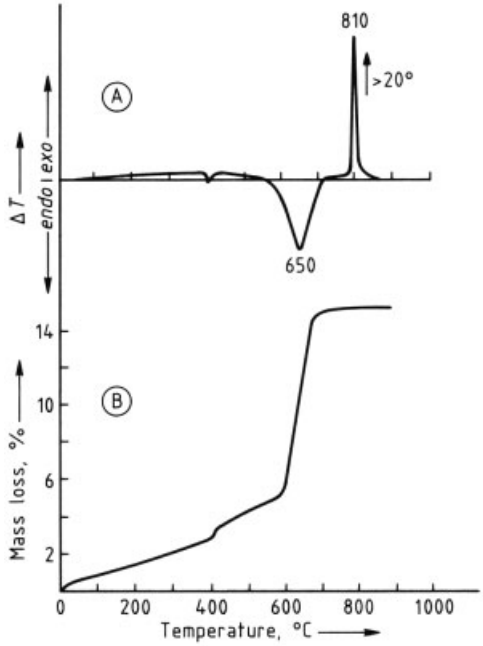

Figure 10. Typical thermal analysis of chrysotile
A. DTA (differential thermal analysis) at 10 °C/minB. TGA (thermogravimetry analysis) at 2.5 °C/minReproduced by permission [1]

Reactions with Acid and Alkali. Strong acids rapidly decompose chrysotile with removal of all the MgO and H_2O content. However, amphibole fibers show various degrees of resistance to reaction with acids. All varieties of

Table 6. Summary of the decomposition reactions of amphiboles

Type	Oxidizing conditions			Neutral conditions		
	Dehydrogenation	Dehydroxylation	Breakdown	Dehydrogenation	Dehydroxylation	Breakdown
Crocidolite	R[a] 400–600 °C P[b] 420 °C 1.9 % H_2O		P[b] 900 °C NaFe–py[d], hem[e], sil[g]	R[a] 550–700 °C P[a] 610 °C 1.9 % H_2O		P[e] 800 °C NaFe–py[d], mag[f], sil[g], glass phase
Prieskaite (fibrous actinolite)	R[a] 450–670 °C 1.9 % H_2O		R[a] 900–1010 °C CaMgFe–py[d], sil[g]	R[a] 400–850 °C (2)[c] P[b] 700 °C 1.9 % H_2O		P[b] 1040 °C CaMgFe–py[d], sil[g]
Amosite	R[a] 600–800 °C 0.9 % H_2O	R[a] 800–900 °C 1.0 % H_2O	R[a] 600–900 °C FeMg–py[d], mag[f], hem[e], sil[g]	R[a] 600–800 °C P[b] 780 °C 1.5 % H_2O	R[a] 600–800 °C 0.4 % H_2	R[a] 600–900 °C FeMg–py[d], sil[g]
Anthophyllite	R[a] 600–850 °C 1 % H_2O	R[a] 850–1000 °C P[b] 950 °C 2 % H_2O	P[b] 950 °C MgFe–py[d], sil[g] mag[f]	R[a] 800–1050 °C P[b] 950 °C 2.7 % H_2O	R[a] 800–1050 °C 0.2 % H_2	P[a] 950 °C MgFe–py[d], sil[g], mag[f]
Actinolite (semifibrous)	R[a] 620–960 °C 1.1 % H_2O	R 960–1080 °C P[b] 1040 °C 1.5 % H_2O	P[b] 1040 °C CaMgFe–py[d], sil[g]	R[a] 620–1080 °C P[b] 1040 °C 2.4 % H_2O	R[a] 620–1080 °C 0.2 % H_2	P[a] 1040 °C CaMgFe–py[d], sil[g]
Tremolite		R[a] 950–1040 °C P[b] 1040 °C 2 % H_2O	P[b] 1040 °C CaMgFe–py[d], sil[g]	R[a] 950–1050 °C P[b] 1040 °C 2 % H_2O		1040 °C CaMg–py[d], sil[g]

[a] R = temperature range.
[b] P = typical peak temperature (DTA).
[c] (2) Indicates two steps in dehydration.
[d] py = pyroxene.
[e] hem = hematite.
[f] mag = magnetite.
[g] sil = silica.

asbestos resist prolonged attack by strong alkali, e.g., 5 M sodium hydroxide solution, at least up to 100 °C. Typical acid resistance curves for asbestos fibers refluxed in 4 M hydrochloric acid are shown in Figure 11.

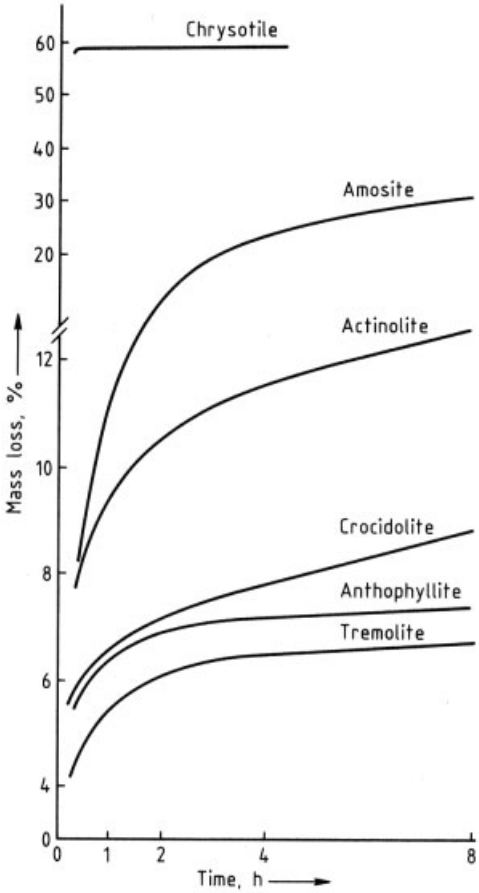

Figure 11. Typical acid resistance curves for asbestos fibers refluxed in 4 M hydrochloric acid
Reproduced by permission [1]

Surface Characteristics. The outer surface of chrysotile fibers consists largely of $Mg(OH)_2$ and behaves as such. The equilibrated pH in carbon dioxide-free distilled water is 10.33. The surface charge in water is positive at a pH lower than 11.8 and rises to a maximum at a pH of 3. With high acidities the charge falls off rapidly as magnesium ions are removed and the silicous surface is exposed. This positive charge renders chrysotile attractive to most other materials in solution that carry a negative charge. However, the surface of amphiboles is like that of silica and carries a small negative charge in water.

Chrysotile has a large specific surface area; this, together with its molecular configuration, enables it to absorb gases fairly readily, particularly highly polar molecules, such as water. In a liquid, absorption can be complicated by competition between the solvent and the solute. The affinity of chrysotile for liquid organic compounds decreases in the following order: ethanol, butanol, benzene, naphthalene, anthracene, hexane. The absorptivity of amphiboles apparently has not been as well studied.

5. Occurrence and Extraction

The three main types of asbestos that have had wide commercial exploitation are chrysotile, amosite, and crocidolite. Of these, chrysotile is by far the most abundant and most widely used. Three others in decreasing order of importance are anthophyllite, tremolite, and actinolite. The last material, although not uncommon as a mineral, is rare in asbestiform habit. Chrysotile exists as yellowish or greenish-white fibers, usually silky in nature. Crocidolite is blue and less silky, and amosite has white, grey, pale yellow, or pale brown fibers that are more brittle than those of the former two varieties. Amosite is not strictly a mineral name, but a commercial name derived from the initial letters of Asbestos Mines of South Africa. Amosite usually consists mainly of the amphibole mineral grunerite, but appreciable amounts of other amphibole fibers are present in some specimens. All forms of asbestos are metamorphic minerals that are closely associated with their parent rocks.

Chrysotile veins form an irregular lattice work in the host rock from which the mineral is obtained by removal of the whole rock, crushing, and extraction at the refinery. Open-pit methods have proved the most economical way of working chrysotile deposits (e.g., in Quebec). Underground mining, particularly block caving, is practiced, e.g., in the Soviet Union, China, and southern Africa; the depth of the mines may exceed 300 m. The mining of amphibole asbestos is almost completely confined to South Africa.

In Canada, over 95 % of all asbestos ore is mined in open-pit operations involving drilling and blasting. The broken ore is loaded into trucks

(35 – 200 t capacity) and hauled to the primary crusher. Overburden and waste rock removed in development of open pits is similarly loaded and hauled to waste dumps. Primary crushing reduces the ore to a size of about 15 cm. After screening the average is reduced to a maximum size of approximately 7.5 cm, followed by wet-rock storage or drying. The ore, delivered by conveyor, is dried in vertical fluidized-bed dryers or in rotary dryers. Using a drying time of 1 – 15 min and an inlet air temperature of 430 – 1000 °C results in a moisture loss of 3 – 18 wt %. The outlet gas is maintained at 60 – 180 °C to prevent damage to bags in dust collection units. The dry ore is conveyed to a storage building. The residual heat in the ore is sufficient for complete drying.

Ore from storage is run over shaking screens equipped with aspirator hoods at the discharge end. The shaking action separates the freed fibers from the rock; they are lifted by air suction and removed from the airstream by a cyclone separator. The rock is further crushed and screened to remove more fiber and the process repeated. The asbestos fibers are freed from dust, rocks, and other material by passage over cleaning screens. After grading the fibers are bagged: either loosely packed in multiwall paper bags or pressure packed in 45-kg units in multiwall paper, plastic, or coated plastic fiber bags. Further storage and shipment is by palletized units of one or more tons.

Because fiber is extracted by a completely dry process, dust is generated in various phases of the operation, and this must be well controlled. Methods include wetting, high-efficiency cyclones in the bag houses, wet scrubbers, and electrostatic precipitators.

6. Uses and Asbestos Substitutes

Asbestos has become very widely used because it is a low-cost material with very desirable chemical and physical properties. This unique combination makes the economical replacement of asbestos very difficult in many applications. Heat and fire resistance, the most widely known property, is not as great as is popularly believed. Asbestos cannot be classified as a refractory, but its properties are sufficient to withstand superheated steam and other high-temperature industrial environments. Whereas major loss of strength can occur at temperatures in the range 300 – 500 °C, useful performance can be obtained at higher temperatures, in some cases up to 1700 °C.

Other properties contribute to the value of asbestos as an industrial material: its resistance to chemical and biological attack make it valuable in applications involving hostile environments and in achievement of a useful service life. The friction and wear characteristics of chrysotile and its thermal decomposition product forsterite (a nonfibrous silicate) make chrysotile a widely used material in such applications as friction (clutch and brake linings), and bearings; the high length to width ratio of asbestos fibers and their tensile strength make them useful as mechanical reinforcers in both polymer- and cement-based products.

Heat Protection. Fire and heat protection wear is made from asbestos cloth; an aluminized surface is bonded to it by thermosetting resins to give a heat-reflecting surface.

Substitutes. Heat-resistant nylon fibers, e.g., Nomex (Du Pont), can be used in place of asbestos. These fibers also protect against chemical hazards (not strong acids) and can be laundered with little deterioration in properties. Nomex is three times more expensive than asbestos but it has a much longer useful life. Special wool blends, e.g., Multitect (Multifabs Ltd., Derby), can protect against molten steel splashes at 1500 °C. The material is resistant to chemical attack, can be laundered with no deterioration in properties, is surprisingly resistant to ignition and flame spread, and is priced competitively with asbestos-based products.

Bleached tetrafluoroethylene fibers, e.g., Teflon (Du Pont), can be used to protect against flash fires and have been used to make suits for Apollo astronauts and for missile fuel handling.

Fire Blankets, Curtains, and Aprons. In general, the materials detailed above are suitable.

Substitutes. A sandwich of ceramic fibers between layers of glass cloth (Marglass Ltd.) is effective against flame and molten metal hazards.

Electric Insulation. Asbestos is widely used in the form of paper, tape, cloth, and board for

electric insulation. Frequently it is applied as a felted material and as a filler for natural and synthetic insulating resins. Loss of strength can develop at above 200 °C if prolonged heating in air is involved.

Substitutes. Glass fabrics of all types (cloth, tape, tubes, cord) are suitable but cannot be used where severe flexing is involved. Polyimide and polyethersulfone are examples of high-temperature-resistant, thermosetting materials used as wire and cable coatings where their ability to stand high temperature, plus toughness and flexibility, justify the cost. Ceramic fiber cloth tape or sleeving may be used with fiberglass inserts at temperatures up to 530 °C.

Filters. Asbestos cloth is used for bulk liquid filters, e.g., for beer.

Substitutes. Expanded perlite and vermiculite products can substitute asbestos as a filter material. Teflon fluorocarbon fibers withstand temperatures up to 300°C, coupled with a corrosive chemical environment.

Ropes, Yarns, Tapes. Asbestos in the form of rope, yarn, and tape is used for many purposes.

Substitutes. Glass may be used, but it softens at about 300 °C. For higher temperatures, continuous ceramic and silica fibers can be used as suitable replacements if price permits.

Thermal Insulation. Asbestos insulation board, asbestos spray, and lagging fall into this group of products.

Substitutes. Glass fiber, mineral wool, vermiculite, perlite, and aluminosilicate ceramic fiber are all suitable. For high-temperature insulation vermiculite is normally bonded with a high-alumina cement or fire clay (satisfactory up to 1100 °C). Expanded perlite is used as an insulating cover on the surface of molten metal, occasionally up to 1100 °C but best below 925 °C because of increase in thermal conductivity of expanded perlite above this temperature.

Packing Materials. Woven and nonwoven asbestos blankets are used as dry packings, e.g., in prefabricated buildings, as furnace and kiln door seals, and as high-temperature caulking material.

Jointings. Asbestos can be bonded with various elastomers, e.g., rubber, polychloroprene, that are resistant to specific oils and solvents.

Gaskets. A sandwich of asbestos with paper or millboard is used between copper sheets in cylinder head gaskets. The material withstands hot gases, hot oil, and hot water – antifreeze mixtures.

Linings. A mixture of powdered diatomaceous earth and asbestos withstands a maximum temperature of 870 °C.

Substitutes. Almost all conventional refractory materials (insulating firebrick or calcined diatomaceous earth) or ceramic fibers are suitable.

Packed or Sprayed Products. Asbestos has superior chemical stability in some situations, e.g., in hot reducing atmospheres and under hot, stressed conditions.

Substitutes. Synthetic ceramic fiber products are suitable. Thermal stability is generally superior to asbestos and chemical stability usually good. Cost is higher than for asbestos but service life is increased. Less maintenance is required and there is increased process efficiency because of higher operating temperatures. Mineral fiber is cheap and mineral wool products are usually cost competitive with similar asbestos products, but mechanical and insulation properties may be inferior.

Millboard. The ease of shaping, insulating properties, impregnability with bonding agents or cement, ability to be wet molded, and compressibility make asbestos millboard one of the most versatile materials used in industry.

Substitutes. Millboards made from aluminosilicate fibers can provide direct replacement at extra cost.

Concrete Products. Asbestos cement is widely used in building and in industry. For high-density, high-surface building materials, approximately 12 – 15 wt% of chrysotile asbestos is normally added to hydrated Portland cement; the end product is compressed into flat or corrugated sheets or molded into other shapes. Asbestos wood (named because of its workability, not its content) contains approximately

25 % chrysotile fiber in hydrated Portland cement. The term "sprayed asbestos" generally refers to gun-applied hydrated cement containing about 60 % asbestos fiber. In industry, the major use is asbestos cement pipes for the transport of corrosive fluids.

Substitutes. Glass fiber-reinforced cement is made by adding approximately 5 % of relatively expensive alkali-resistant glass fibers to hydrated cement and other fillers. The resulting composite can be extruded, molded, or sprayed. It is marketed as CemFIL (Pilkingtons). Glass-reinforced gypsum concrete, and polyester are also available.

Friction Materials. Asbestos-based brake and clutch linings and pads are in widespread use. Automobile brake linings are usually molded from a mixture of short chrysotile fibers, phenolic resin, and fillers. Brake pads are made in this way, or a woven asbestos cloth, reinforced with brass wire and impregnated with phenolic resin, can be used.

Substitutes. Sintered metal-ceramic and metallized bonded products are available, but their high-thermal conductivity can adversely affect the brake fluid. Delaminated vermiculites maintain their strength at high temperature and are compatible with phenolic resins. Silicon nitride was used for the brake pads in prototypes of the supersonic passenger aircraft Concord at a cost of about U. S. $ 7000 per ton. Carbon – carbon composites were used in the Concords production. Steel, glass, mineral wool, and aluminosilicates have been used in phenolic binders, but none is as good as asbestos for such items as disk brake pads.

Dry Rubber Bearings. Asbestos cloth or yarn impregnated with phenolic resin is used in plain rolling bearings that are able to function effectively without lubrication (although they can also be lubricated). They are used, for example, in automotive steering columns, brake and clutch pedal pushers, bearings for textile spin-ning, frame spindles, food slicer bearings, copy machine rollers, large railway and marine bearings, or lock gates. The maximum operating temperature is 175 °C.

Substitutes. Polyester-bonded textile laminates with molybdenum disulfide or graphite and cellulose fabric-bonded phenolic laminates and polytetrafluoroethylene (PTFE) or graphite can be used up to 100 – 130 °C.

Composites. A thermosetting and sometimes a thermoplastic matrix (phenolic, silicone, or polyester) is reinforced by the addition of asbestos fibers and also sometimes by the addition of glass fibers, enhancing strength, stiffness, and toughness. The material is used in small machine parts, usually made by injection rolling.

Substitutes. Generally, glass fibers provide an adequate replacement for asbestos. Carbon fibers are used in special applications; they save weight but are expensive.

Miscellaneous. Asbestos felt impregnated with bitumen is used for roofing and damp courses. Floor tiles and high-temperature mastics (flexible fillers) may contain asbestos.

Substitutes. Conventional cellulose fiber papers and felts can be used in impregnated form. Glass fiber can replace asbestos for many applications. Mastics based on aluminosilicate fibers in colloidal silicate binder are used for refractory caulking.

7. Economic Aspects

The use of asbestos grew slowly during the early part of the 20th century and by 1930 a cumulative world total of approximately 5×10^9 kg of asbestos had been mined. Since the decade 1930 to 1940 production has accelerated such that the production for 1976 alone (ca. 5.6×10^9) exceeded the 1930 cumulative total. The U. S. Bureau of Mines has estimated that the world has a reserve base of 104×10^6 t of asbestos ore, 200×10^6 t of identified resources, and 15×10^6 t classified as hypothetical resources [4]. The world production of asbestos by principal producing countries is shown in Table 7.

8. Asbestos in the Environment and Waste Disposal

The amount of asbestos used in everyday appliances has been reduced. For example, asbestos blankets have been replaced by fiberglass ones and asbestos paper is no longer used in schools.

Table 7. World production of asbestos by principal producing countries

Year	Canada		South Africa		USSR		World total
	1000 t	%	1000 t	%	1000 t	%	1000 t
1948	590	63	38	4	–	–	–
1953	827	58	86	6	272	19	1420
1958	839	45	159	9	499	27	2055
1963	1157	46	187	8	685	27	2505
1968	1370	46	236	8	800	27	2985
1973	1690	43	334	8	1280	31	4179
1978	1422	30	257	5	1945	42	4673
1979	1493	30	249	5	2100	42	4992
1980	1323	27	277	6	2250	45	4988
1981	1133	24	273	6	2340	49	4817

However, the general population, without its knowledge, is still exposed to asbestos. One of the most likely ways this may occur is through drinking water [25]. Asbestos can enter the water supply in many ways. Erosion from industrial deposits of serpentine minerals or improper disposal of asbestos washings can be sources of asbestos. Also, some water pipes are made of asbestos cement; erosion of these pipes can also provide a source of asbestos [26]. Many gaskets used in the water system are made from asbestos.

The amount of asbestos in water is difficult to monitor. The fiber concentration has been found to vary from 0 to 130×10^6 fibers/L (visible by electron microscopy).

In the United States some people presumably consume at least 2×10^9 fibers or 710 µg of asbestos per year through their drinking water. The effects of these levels on health is not known. Asbestos has also been detected in wine.

According to the voluntary labelling scheme devised by the Asbestos Research Council and the Asbestos Information Committee (UK) products containing asbestos should be marked as such and precautions against inhaling the dust should be on the package. However, this is not done.

Asbestos is contained in many do-it-yourself products, for example, many plugging materials, asbestos cement sheets, acoustic tiles, and insulating board. Sawing and cutting of these sheets and tiles invariably releases dust into the atmosphere. The level of asbestos dust produced during these activities is many times that of the safety standard; also, unlike in professional practice, there is no damping of the dust to reduce levels.

It is not known how long an exposure is necessary to cause mesothelioma. The least time recorded by Department of Health and Social Security statistics is three weeks. Seventeen cases were recorded in 1967 – 1968 in which no one had worked with asbestos. One man with mesothelioma had been exposed for only one day while sawing asbestos sheets.

Do-it-yourself products probably do not contain crocidolite, but other types of asbestos are thought also to cause mesothelioma. In the United States four cases of meso-thelioma were found among the families of 1600 workers handling amosite. In the United Kingdom there are four cases known caused by chrysotile.

Asbestos is used in brake linings and clutches; there is an elevated level of airborne asbestos around braking vehicles. People who repair brakes and clutches may inhale large amounts of asbestos. Car brakes need to have the asbestos dust blown out of the drums regularly. According to various reports this dust consists of asbestos not in its normal state, but decomposed by the high temperature caused by friction on braking. However, there have been cases recorded of people with mesothelioma exposed in this way, namely a young boy in the United States who used to help his father clean out brake drums and who died from cancer at a young age.

Some roads and footpaths are surfaced with asbestos-bearing gravel, and this may produce airborne asbestos. Also, humidifiers may contain water with high levels of asbestos. There are asbestos-bearing paints and cigarettes with asbestos filters (reportedly no longer used in U. S. cigarettes). Even powdered papier-mâché

mixes, which are widely used in schools, have been found to contain asbestos.

However, most people are not engaged in the installation of asbestos products or other activities involving high asbestos levels. Ambient air therefore constitutes the greatest source of atmospheric exposure for the population at large.

Atmospheric Concentration. Atmospheric asbestos concentration data are scarce, because of the cost and the difficulty in obtaining it. The average reported concentration of asbestos in urban air in the United States varies between 0.09 and 70 ng/m^3 and between 0 and 2400 fibers per cubic meter. Using concentration data expressed in terms of mass (representing 203×10^6 people) the percentage of the urban population at risk to various atmospheric levels of asbestos in the United States has been calculated and is shown in Table 8. Assuming that humans breathe 15 m^3 of air daily the annual amounts of inhaled asbestos can be estimated from the ambient atmosphere for the United States urban population: 50 % of the urban population inhales 110 µg or more of asbestos per year, and 10.5 % inhales 230 µg or more, compared with the range 0.05 – 0.5 µg for rural populations. People who live or work in or near asbestos mines and factories are exposed to even higher levels of asbestos.

Disposal of Asbestos Waste. Asbestos is a rock mineral and not acutely toxic. There is no reason why it should not be disposed of by landfill. However, there must be safeguards, particularly to prevent dust formation. The asbestos must be buried at a sufficient depth to prevent the surface from drying out. This presently is practiced in many countries. Wet waste may be disposed of as a slurry at a dry waste site providing quantities are not excessive.

The asbestos waste is placed in a sealed impermeable bag or other sealed receptacle, marked "Asbestos Waste" or particularly "Blue Asbestos, Do Not Inhale Dust." The major principle of disposal is that no dust be emitted into the air during transport or dumping. In order to avoid the risk of the receptacles breaking, vehicles used for transporting the waste should have a low discharge point in relation to the ground. Ideally, the waste is deposited at the foot of the waste site, thus minimizing the need for further disturbance as the site is formed; finally, the waste is covered with at least 22 cm of consolidated earth or other suitable material. Covering should be done immediately in the case of waste derived from blue asbestos; no waste should be left uncovered at the end of a working day.

National or local regulations governing the disposal of asbestos waste usually exist. For example, in the United Kingdom notification of removal or deposit of asbestos waste is covered by Section 3 of the Deposit of Poisonous Waste Act 1972 and there is parallel legislation in other countries.

9. Toxicology and Occupational Health

9.1. Asbestos-Related Diseases

The mining and use of asbestos have always been associated with disease among the workers involved. However, because prior to 1900 the number of people involved was not large, there was little to draw attention to these hazards. It is now well established that asbestos fibers can produce a number of medical effects, the most common of which are pleural thickening, asbestosis, mesothelioma, and cancer. These asbestos-related diseases can appear from several years to tens of years after the first exposure to airborne asbestos dust; the former increasing incidence of cases in many countries could be regarded as a result of past dust exposure, which was obviously too high.

9.1.1. Inhalation of Asbestos Dust

When dust-laden air is inhaled, most large asbestos fibers (10 nm or larger in diameter) are trapped in the mucous membranes and hairs of the nasal passages. Fibers of smaller size may enter the bronchial tree and are mostly cleared by the cilia lining the respiratory tract, particularly if they are larger than 5 nm in diameter. Smaller fibers, particularly if less than 3 nm in diameter, may enter the terminal bronchioles and alveoli. These smaller fibers, especially those with a diameter of less than 1 nm, are considered to be the most significant in inducing biological effects in humans.

Table 8. Inhalation of asbestos from ambient air

Proportion of urban	Atmospheric concentration		Annual inhalation	
population consuming indicated level or more, %	Fibers∗/m^3	ng/m^3	Millions of fibers∗	µg
50	5000	20	27.4	110
25	7500	28	41.1	158
10	10500	40	57.5	220
5	12500	48	68.4	268
1	16000	64	87.6	350
Rural population	–	0.01 – 0.1	–	0.05 – 0.5

∗ Fibers visible in the electron microscope, assuming 250 fibers per nanogram of asbestos

Some of the needle sharp fibers may penetrate bronchiolar and alveolar walls to remain in the interstitial tissue, or they migrate to the pleural surface or enter the lymphatic vessels. A high proportion of asbestos fibers that are retained in the lung become coated with a ferroprotein complex and become asbestos bodies. These asbestos bodies remain dormant in the lungs for many years, often presenting giant multinucleoate cells. After this time, the ferroprotein coat can break down, releasing pieces of fiber; this causes the lung tissue to be destroyed and to be replaced with scar tissue. In some cases its changes are particularly marked in the pleura, which becomes thickened. This is the start of asbestosis.

9.1.2. Pleural Thickening

Pleural thickening has been observed in some asbestos workers and also in some people living near asbestos dumps. It is a thickening of the surface of the lungs that occurs after exposure to asbestos dust. In the absence of asbestosis, pleural thickening itself does not have any ill effects; it is not necessarily associated with lung cancer or mesothelioma.

9.1.3. Asbestosis

Asbestosis is a fibrosis of the lungs caused by inhalation of asbestos dust and is commonly found in asbestos workers. It is a nonmalignant, progressive, irreversible lung disease characterized by diffuse intestinal fibrosis [27]. The disease develops slowly and is detectable by a combination of clinical, radiographic, and lung function tests. In the early stage there is uncertainty about diagnosis, and the characteristic symptoms are progressive breathlessness and an unproductive cough. The fibrosis tends to increase despite the removal of the patient from further dust exposure. The risk of developing asbestosis apparently is related to the duration and level of asbestos dust exposure. Detectable early stages at which removal from further dust exposure markedly influences the course of the disease have not yet been found. Cases of asbestosis have occurred in people many years after their last known exposure to the dust.

9.1.4. Cancer

Cancer of the larynx, stomach, colon, and rectum may be associated with asbestos. There is also an increase in the incidence of bronchial tube cancer or lung cancer in people who have had asbestosis. It is not yet certain whether asbestos can contribute to cancer when asbestosis is not already present. Over half the British death certificates recording asbestosis also record a lung cancer.

The first reports of a relationship between asbestos dust exposure and *bronchial carcinoma* were made in 1935 [28], [29]; this was confirmed by subsequent studies [30]. All commercially used types of asbestos have been shown to cause an increase in the occurrence of bronchogenic carcinoma.

There is no difference in the clinical behavior of carcinomas in asbestos workers as compared with nonasbestos workers. Most bronchial carcinomas arise in the epithelium of the major bronchi. Asbestos is also reported to cause alimentary tract carcinomas (esophagus, stomach,

colon, and rectum) among asbestos workers. No firm data are available to link asbestos exposure to other neoplasms. Suggestions that carcinoma of the ovary might be asbestos related have not been proved.

Lung Cancer and Cigarette Smoking. In 1968, important new information on the relationship between asbestos exposure and cigarette smoking was published [31]. This showed that the risk of developing cancer for asbestos workers who smoked was 90 times greater than for nonsmoking asbestos workers (Table 9). For this reason health education and antismoking campaigns are of particular importance in the asbestos industry.

Cancers of the Alimentary Tract. Cancers of the alimentary tract occur about twice as regularly among heavily exposed asbestos workers as among the general population. All types of smoking, the consumption of alcoholic beverages, and the chewing of tobacco increase the risk of cancer of the oropharynx and the esophagus [32]. The consumption of beer has also been correlated with mortality from rectal cancer. Persons having close relatives with a cancer of the alimentary tract are also at an increased risk. However, refusing someone employment as an asbestos worker can only be justified when all these factors are present. Larynx carcinoma has also been related to asbestos exposure and alcohol consumption [32].

9.1.5. Mesothelioma

Mesothelioma is a malignant tumor, exceptionally rare in the general population; however, it occurs most often between the 5th and 8th decade of life among asbestos workers [33]. It affects the lining of the thoracic (pleural) or, more rarely, abdominal cavity (peritoneum). Mesothelioma is usually associated with exposure to blue asbestos but has also occurred among workers exposed to other types of asbestos. The duration of exposure required to cause mesothelioma may be brief and the disease can appear at some considerable time later. Most cases of mesothelioma are related to past exposure to asbestos; present evidence indicates that insulation workers as a group are at highest risk. Few cases of mesothelioma have been reported in which no exposure to asbestos can be traced.

Diffuse malignant mesothelioma is a rapidly progressing disease and there are no reliable estimates for a five-year survival rate. In 70 – 90 % of the cases, patients have shoulder or chest pain as a primary symptom. The pain becomes sharper as the tumor invades the chest wall. However, the pain is not aggravated by breathing or chest wall movement. Eventually, the pain is severe enough to interfere with sleep and to require narcotic medication. Weight loss, weakness, and fever are less frequent but become more common as the tumor progresses. Fluid may build up in the lungs and is sometimes bloody – this must be removed by aspiration.

Detection and identification of mesothelioma during life is difficult. Life expectancy of a person with pleural mesothelioma averages about one year with occasionally patients living several years.

In addition, peritoneal mesothelioma may occur; the earlier symptoms of this are vague abdominal discomfort, loss of appetite, and constipation. Intestinal obstruction may occur later but it may be the presenting complaint if the patient delays seeking medical attention. Physical examination is unrevealing in the early stages. As time goes by the presence of an abdominal mass becomes apparent but even at surgery diagnosis is elusive. The disease progresses more rapidly than pleural mesothelioma and death as a result of intestinal obstruction is common. No effective therapy is available, any treatment is supportive.

9.2. Occupational Health

Official exposure limits to asbestos dust levels have been established in many countries. The exposure limits tend to vary from one country to another: the TLV (threshold limit value, United States) is 0.2 fibers/cm^3 (crocidolite, $>5\,\mu$m diameter), 0.5 fibers/cm^3 (amosite, $>5\,\mu$m diameter), and 2 fibers/cm^3 (chrysotile and other forms, $>5\,\mu$m diameter); the CL (control limit, United Kingdom) is 1 fiber/cm^3 (chrysotile). No MAK (Maximale Arbeitsplatz Konzentration, Federal Republic of Germany) has been established; all asbestos forms are classified as

Table 9. Mortality of asbestos workers by smoking habit

Smoking habits	Number of men	Observed deaths	Expected deaths
Never smoked regularly	40	0	0.05
History of pipe or cigar smoking only	39	0	0.13
History of regular cigarette smoking	283	24	3.16

A1 (human carcinogens), the lowest technically feasible level (TRK) being at 1 fiber/cm^3.

The standard method for measuring asbestos dust is the membrane filter method. The air is sampled in the operator's breathing zone and a known volume is drawn through a membrane filter. The filter is then made transparent and the number of fibers deposited on the filter fitting a standard definition of size and shape are counted using a phase-contrast microscope.

Control of Asbestos in the Workplace. Exhaust extract ventilation equipment and protective clothing must be provided. Respirators must be approved by the appropriate authority and provide protection against the asbestos level encountered by the worker.

The masks providing the best protection are usually battery-powered, positive-pressure types. Overalls provided are usually of the one-piece boiler suit type with close-fitting neck, wrists, and ankles. The overalls are usually of a finely woven synthetic fiber, to which the asbestos dust does not adhere very readily. Ideally, the feet should be protected by Wellington boots worn with the overall trousers outside the boots. The boots should not be worn outside the work zone. Fingernails are a notorious trap for contaminants; for this reason gloves should be worn.

The machinery and the work areas themselves should be kept as clean and free from dust as possible. Measures should be taken to create as little dust as possible during the work process, e.g., by damping materials or by complete closure of the work area. If this is impracticable, a dust-control system is required.

Asbestos should be transported around the workplace in closed receptacles, and these receptacles should be treated as asbestos waste after use and placed in other closed receptacles for disposal. Whenever specially designed premises and equipment are used for asbestos work, it is relatively easy, although expensive, to maintain safe conditions. However, most of the danger lies in demolition or removal work, and here, too, the respirators and overalls previously described must be provided and used. Whenever crocidolite is present or expected to be present high-efficiency respirators must be used.

Workers Precautions

1) Cigarette smoking greatly increases the risk of lung cancer. Cessation should reduce the number of deaths in asbestos workers.
2) The history of lung cancer in a first-degree relative (parents, siblings, offspring) has been shown to confer the same amount of risk as cigarette smoking; both factors are synergistic. Perhaps these persons should be removed from asbestos exposure.
3) Prior occupational exposure to carcinogens affecting the lung may also increase the risk of lung cancer in an asbestos worker. Such carcinogens include arsenic, chloromethyl ethers, chromates, coal tar, petroleum and petroleum products, creosote, mustard gas, nickel, radium, and uranium.
4) Nonmalignant respiratory disease may increase the risk of lung cancer, e.g., pulmonary fibrosis or chronic bronchitis.

According to a Norwegian study (1975), vitamin A intake at all levels of cigarette smoking may decrease the risk of lung cancer. High doses of vitamin A and its synthetic analogues (retenoids) have been shown to prevent lung cancer in laboratory animals. Asbestos workers should make sure of taking their daily requirement of vitamin A.

10. References

1. L. Michaels, S. S. Chissick (eds.): *Asbestos – Properties, Applications and Hazards,* **1,** J.

Wiley & Sons, Chichester, England, 1979 (printed 1981).
2. S. S. Chissick, R. Derricott: *Asbestos – Properties, Applications and Hazards,* vol. 2, J. Wiley & Sons, Chichester, England, 1983.
3. G. D. Wright, *Am. Rev. Respir. Dis.* **2** (1969) 467 – 479.
4. "Asbestos Mining Industry Review, 1982" *Asbestos* 1983 (January) 5.
5. HMSO, *Report of the Departmental Committee for Compensation for Industrial Diseases,* Cmd. 3495, 3496, HMSO, London, 1970.
6. W. E. Cooke, *Br. Med. J.* **2** (1927) 1024 – 1025.
7. S. McDonald, *Br. Med. J.* **2** (1927) 1025 – 1026.
8. P. F. Holt, D. K. Young, *Atmos. Environ.* **7** (1973) 481 – 483.
9. P. Bogovski, J. C. Gilson, V. Timbrell, J. C. Wagner (ed.): *Biological Effects of Asbestos,* International Agency for Research on Cancer (IARC) Scientific Publications No. 8, International Agency for Research on Cancer, Lyon 1973, pp. 126 – 130.
10. P. Sebastien, J. Bignon: *Contribution à l'Etude de la Pollution Particulaire de l'Air Ambiant de la Ville de Paris par les Microfibrilles d'Amiante,* Rapport Final, Contract No. 206, Ministère de la Qualité de la Vie, Paris 1974.
11. P. Sebastien, J. Bignon, A. Gaudichet, G. Dufour, et al.: Les pollutions atmosphériques urbaines par l'asbeste, *Rev. Fr. Mal. Respir.* **4** (1976) 51 – 62.
12. I. J. Selikoff, W. J. Nicholson, A. M. Langer, *Arch. Environ. Health* **25** (1972) 1 – 13.
13. P. D. Nicholson, A. N. Rohl, I. Weisman: *Asbestos Contamination of the Air in Public Buildings,* EPA Contract No. 68-02-1346, U. S. Government Printing Office, Washington, D.C., 1975.
14. A. L. Richards, *Anal. Chem.* **45** (1973) 809 – 811.
15. H. M. Cunningham, R. D. Pontefract, *Nature (London)* **232** (1971) 332 – 333.
16. American Water Works Association, *Am. Water Works Assoc. J.* 1974 (September) 1 – 22.
17. P. M. Cook, G. E. Glass, J. H. Tucker, *Science* **185** (1974) 853 – 855.
18. R. L. Zielhuis (ed.): *Public Health Risks of Exposure to Asbestos,* European Economic Community, Directorate of Social Affairs, Contract No. 18-6- 1975, 1977.
19. R. R. Merliss, *J. Am. Med. Assoc.* **216** (1971) 2144.
20. A. H. Wolff, F. W. Oehme, *J. Am. Vet. Med. Assoc.* **164** (1974) 623 – 629.
21. H. J. Wehman, B. A. Plantholt, *Bull. Environ. Contam. Toxicol.* **11** (1974) 267 – 272.
22. W. J. Nicholson, C. J. Maggiore, I. J. Selikoff, *Science* **177** (1972) 171 – 173.
23. P. F. Holt, R. A. Lemen, *J. Am. Dent. Assoc.* **93** (1976) 221 – 222.
24. K. Yada, *Acta Crystallogr.* **23** (1967) 704 – 707.
25. I. M. Stewart, *Asbestos in the Water Supplies of Ten Regional Cities,* EPA Publication No. 560/ 6-76-017, Environmental Protection Agency, Washington, D.C., 1976.
26. American Water Works Association, *Am. Water Works Assoc. J.* **66** (1974) 1 – 22.
27. J. C. Wagner, *Ann. N.Y. Acad. Sci.* **132** (1965) 691 – 695.
28. K. M. Lynch, W. A. Smith, *Am. J. Cancer* **24** (1935) 56 – 64.
29. S. R. G. Coyne, *Tubercle* **17** (1935) 5.
30. R. Doll, *Br. J. Ind. Med.* **12** (1955) 81 – 86.
31. I. J. Selikoff, E. C. Hammond, J. Churg, *J. Amer. Med. Ass.* **204** (1968) 106 – 112.
32. J. F. Fravmeni, Jr. (ed.): *Persons at High Risk of Cancer,* Academic Press, New York 1975, pp. 139 – 148.
33. J. C. Wagner, C. A. Gleffs, P. Marchand , *Br. J. Ind. Med.* **17** (1960) 260 – 271.

Ascorbic Acid → **Vitamins**
Aspartic Acid → **Amino Acids**

Asphalt and Bitumen

Luke W. Corbett, Consultant – Asphalt Technology, Mountainside, New Jersey 07092, United States

Rolf Urban, Arbeitsgemeinschaft der Bitumen-Industrie (Arbit), Hamburg, Federal Republic of Germany (Section 9.3)

1. Introduction 1
2. Native Asphalt 2
3. Testing 3
4. Bitumen Production from Crude Oil . 5
4.1. Distillation 5
4.2. Solvent Separation Methods 5
4.3. Continuous Air Blowing 7
4.4. Manufacturing Routes for Cement Grades 7
4.5. Road Bitumen Compositions 8
4.6. Batch Air Blowing 8
5. Composition of Bitumen 9
5.1. Elemental Analysis and Molecular Mass 10
5.2. Physical, Structural, and Colloidal Aspects 11
6. Physical Properties 12
7. Chemical Reactions of Bitumen Surfaces 15
8. Handling and Storage 16
9. Uses of Bitumen 17
9.1. Bitumen in Road Building 17
9.2. Industrial Uses 17
9.3. Economic Aspects 18
10. Environmental and Workplace Exposure 19
11. References 20

1. Introduction

In the United States the terms asphalt and bitumen are synonymous. However, their usage often varies in other countries. The American Society for Testing and Materials (ASTM) defines bitumen as a generic class of amorphous, dark colored, cementitious substances, natural or manufactured, composed principally of high molecular mass hydrocarbons, soluble in carbon disulfide [11]. Asphalt also is defined as a cementitious material in which the predominating constituents are bitumens. The terms bituminous and asphaltic then refer to materials that contain or are treated with bitumen or asphalt.

European and Eastern countries prefer to use the term bitumen, whereas in the Americas asphalt is more commonly used, with the terms considered to be synonymous under the above definitions. In the Federal Republic of Germany terms and definitions relating to bitumen and bituminous preparations are compiled in DIN 55946. The Europeans also call bitumen – aggregate mixtures, as used in road paving, asphalt, which is not consistent with the ASTM definition. Thus some confusion exists, which cannot be resolved by this treatise, other than by adhering as closely as possible to overall usage and practice. In this article the term bitumen is used preferentially, although at times it may be used interchangeably with asphalt when citing certain published literature.

The primary source of bitumen or asphalt originates in or is derived from petroleum crude oil, commonly referred to as petroleum or crude. It occurs naturally in rock strata of certain geological formations, from which it is produced by drilling, then recovered by natural flow or by pumping. Subsequently it is submitted to refinery distillation which yields a residual fraction containing bitumen. In certain areas of the world petroleum crude oil has escaped through the earth's surface layers, then aged and weathered for long time periods. This results in a natural or native deposit of bitumen, generally mixed with mineral matter. Common usage dictates the term *native asphalt* in this case (Chap. 2).

This article provides a brief overview of the more important technical aspects of petroleum-derived bitumen. An enormous amount of research has been devoted to this subject. The references listed herewith only touch the fringes of the subject.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_169

Table 1. Three classes of native asphalt with typical properties and examples

Class	Nearly pure bitumen	Appreciable mineral matter	Small fraction of bitumen
Class name	asphaltite	lake asphalt	rock asphalt
Solubility in CS_2, %	98	53	7
Softening point, °C	195	93	–
Penetration at 25 °C	2	1	–
Specific gravity	1.07	1.40	2.30
Mineral matter	silica – carbonate	clay – silica	sandstone – limestone
Examples	Gilsonite, United States Bermudez, Venezuela	Trinidad, West Indies	Kentucky, United States Seyssel, France

2. Native Asphalt

The term native asphalt refers to a wide variety of materials found in nature and containing varying amounts of bitumen as defined by solubility in carbon disulfide. Native asphalt varies considerably in composition and properties. From a practical viewpoint, it is convenient to classify these asphalts into three groups according to their bitumen content, see Table 1.

There are many deposits of native asphalt occurring in various countries, the majority of which have little significant commercial usage. Some of the more commonly known sources are as follows.

Gilsonite. Native asphalt was discovered in Utah, United States, in 1882 and later developed commercially by S. H. Gilson, after whom it was named. This material, which varies somewhat from one vein to another, is characterized by its high softening point and low mineral matter content. For many years Gilsonite was extracted from mines by hand, whereas now it is recovered by means of high-pressure water jets. Its most common use has been in paints, primers, sealers, and floor tiles. More recently it has been coked in pilot tests to yield gaseous and liquid fuels.

Trinidad Asphalt. Trinidad Lake is situated on the island of Trinidad, West Indies. It covers a surface area of about 400000 m^2 and is about 90 m deep at its center; it has the form of an inverted cone. It was reportedly discovered in 1595 by Sir Walter Rayleigh and is estimated to contain about 6×10^6 t of asphalt, or pitch as it is often referred to.

The genesis of the lake was Cretaceous petroleum oil escaping in the Nariva time, forming a submarine seep on the sea bottom over a large uplift. The heavy oil and gas seeped through fissures and mixed with sediments, forming ultimately a lenselike body of asphalt heavy enough to remain at the sea bottom. During the unfolding and faulting, the asphalt rose some 1300 m through unconsolidated sediments to an exposed anticline.

The lake has been mined for many years by digging with mechanical machinery, transporting to stills where it is heated and screened, then run into wood or light steel drums. This material is known as refined Trinidad asphalt under the trade name Epuré and has been exported to many parts of the world. Epuré has found considerable use in European countries, where it is fluxed in minor proportions with petroleum asphalt cements. It is claimed by many to add service life when used in road surface mixes [12], [13].

Rock Asphalt and Oil Sands. Kentucky rock asphalt is found in deposits or beds 2 – 15 m thick in horizontal strata below an overburden of clay soil. It contains between 5 and 20 % bitumen and is recovered by simple quarrying, broken into lumps, and then crushed to a coarse powder. This product is then blended with a petroleum bitumen in order to upgrade the bitumen for use as a road mixture.

Athabaska sands from Alberta, Canada, are of special interest because of their reported reserve of over 300×10^9 barrels of oil. Currently the oil is being recovered on a sizeable scale by steam and kerosene extraction yielding a synthetic crude oil, termed syncrude (→ Tar Sands) [14]. The recovered oil is then processed using typical petroleum refinery methods. The oil sands can be classified as native bitumen or asphalt (small bitumen concentration). However,

the recovered syncrude is not suitable for manufacturing the normal bituminous products described in the following sections of this article.

3. Testing

The properties of bitumen are described by physical tests. Usually, consistency or some combination of consistency with other basic properties is measured. Testing and the control of properties through specifications are important in any discourse on bitumen. Only the most commonly used standardized tests are covered in this chapter; for more details, see the appropriate ASTM and DIN publications.

Softening point is a measure of the temperature at which a steel ball (9.5 mm diameter) passes through a disk of the sample and falls a distance of 25.4 mm (1 inch) when the specimen, ball, and bath of water or glycerol are heated at a specific rate (Fig. 1). Bitumen does not have a definite melting point but changes gradually from a solid to a fluid during an increase in temperature. The softening point as measured by this method approximates the temperature at which this takes place.

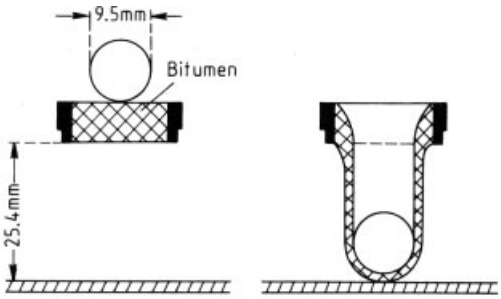

Figure 1. Softening point test (ring and ball)

Penetration involves the determination of the extent to which a standard needle penetrates a properly prepared sample of bitumen under specified conditions of temperature, load, and time (Fig. 2). The unit of penetration is 0.1 mm, which is generally omitted in favor of reporting just the measured number. The test range is from 0 to about 300. The harder materials are indicated by the lower numbers.

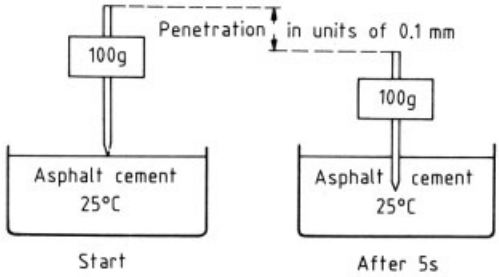

Figure 2. Penetration test (100 g/5s)

Viscosity is a measure of resistance to flow, that is, the internal friction of the bitumen. It is measured as the ratio of shear stress to shear rate and may be reported in different units depending on the apparatus used. Figure 3 illustrates the use of the vacuum capillary method (usually at 60 °C). On less viscous materials or at higher temperatures, viscosities are also determined using the Zeitfuch crossarm viscosimeter (capillary viscosimeter) [15]. Whereas the above test methods are routine, viscosity may also be determined at ambient or lower temperatures using more sophisticated instrumentation [16], [17]. In general,viscosity is perhaps the most important singular test quantity determined on any bitumen, mainly because it relates closely to both application and service conditions.

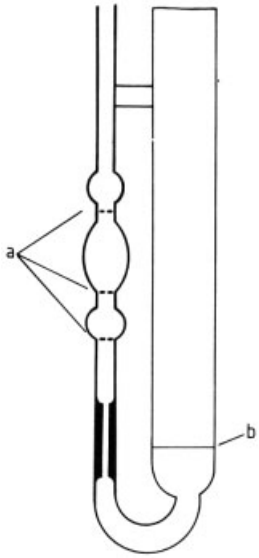

Figure 3. Vacuum capillary method for viscosity measurement (Cannon-Manning tube)
a) Timing marks; b) Filling line

A system for grading asphalt cement by viscosity was developed [18], [19]. It is based upon viscosity measurement at 60 °C. For example, an AC-20 grade is controlled to a viscosity of 200 Pa·s (± 20 %). Grades are then specified from AC-2.5 to AC-40. This system is almost universally used in the United States and is considered to be an improvement over the penetration grading system used in most other countries.

Flash point is defined as the lowest temperature at which a combustible or flammable material evolves sufficient vapor, under the conditions of the test, to form a flammable vapor – air mixture that flashes when ignited. On bitumen this is conveniently determined using the Cleveland open-cup method (ASTM 92–72, DIN 51376). Flash point measurements or values are related to the temperature at which the bitumen vapor may ignite under appropriate conditions. Flash test results, however, should not be interpreted as an indicator of safety control during storage or handling.

Ductility of a bituminous material is a measure of the distance in centimeters that it elongates before breaking when the two ends of a briquet specimen are pulled apart at a specified rate and temperature (Fig. 4). Ductility is a significant characteristic of bitumen but the presence or absence of ductility is more significant than the degree or the amount determined. Ductility does not measure any specific property but rather a combination of flow properties. It may be used in qualifying bitumen used in paving mixtures because it permits the mix to be slightly overstressed without cracking (see also [20]).

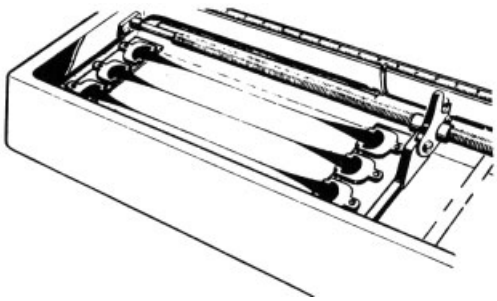

Figure 4. Ductility test

Fraass break point is a test applied to determine the temperature at which a layer of bitumen (0.5 mm) cracks or breaks under specified conditions of bending (Fig. 5). In principle it is a flexure test conceived to relate to ductility, but it is more adaptable to low-temperature testing. The reproducibility of the test is not good, but it has some relative value in judging low-temperature qualities.

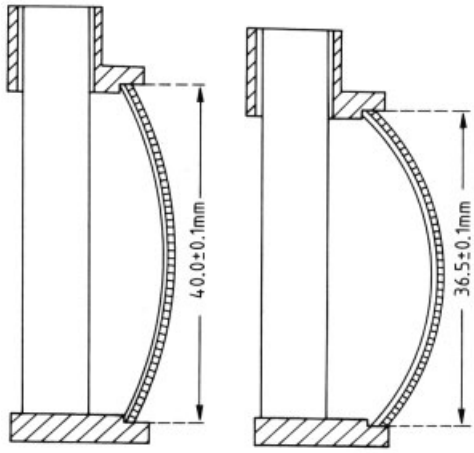

Figure 5. Break point (Fraass)

The *thin-film oven test* is used to measure the relative viscosity increase or penetration hardening of a bitumen exposed to 163 °C in a 3.2-mm film. The bitumen film is exposed in a pan on an oven shelf for a period of 5 h. The ratio of viscosity at 60 °C after the test to before the test is a measure of consistency change related to changes found when processing bitumen through the hot-mix cycle. Ductility minimum is also specified on the thin-film test residue, serving as a more critical requirement than when applied to the bitumen before the test. The *rolling thin-film oven test* and the *rotating flask test* are modifications of the above involving a rotating movement of the bitumen film during exposure [21]. They are considered to be equivalent in both test values and test significance.

Other tests are often specified but are not illustrated here. For example, solubility in trichloroethylene is usually required to be at least 99 % in order to insure purity. The Oliensis spot test [22] is universally required on bitumen used in the United States for state and federal highways. This would outlaw products that have been overheated during manufacture or handling, thus upsetting their compositional balance.

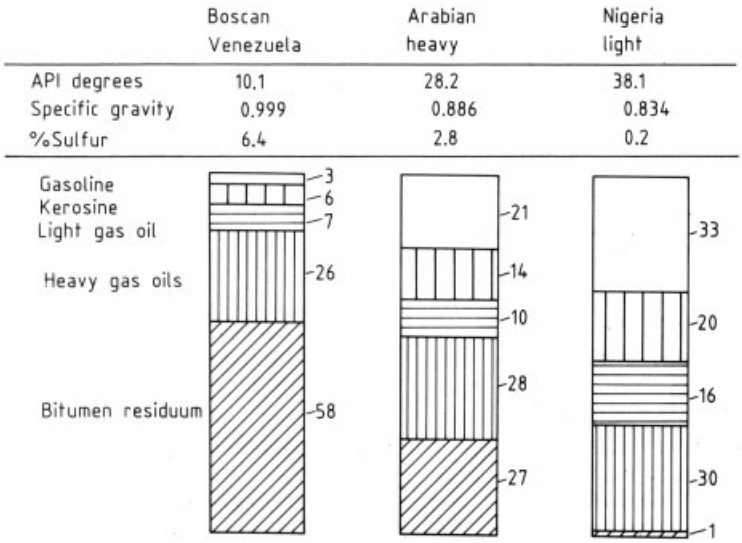

Figure 6. Composition of various petroleum crude oils (concentrations in vol %)

4. Bitumen Production from Crude Oil

The largest source of bitumen is crude petroleum oil [23]. Of the approximate 1500 crude oil fields and pipeline streams in the Western World, only about one fourth are suitable for the manufacture of bituminous products. The rest are normally used for the manufacture of fuel products or chemical specialities.

Figure 6 demonstrates the reason that some crudes are more suitable for bitumen manufacture than others [24]. Typically the lower API gravity crudes have the highest bitumen content and the lowest percentage of distillable fractions, that is, gasoline, kerosine, and gas oils. The high-gravity crudes typically contain less bitumen and higher percentages of distillable fractions. The low-gravity crudes (low API degree) are called heavy crudes, or sour crudes if their sulfur content is high.

4.1. Distillation

Distillation is a necessary first step in the separation of a bitumen-containing residuum from crude petroleum. This process is referred to as straight reduction and is illustrated in Figure 7. Distillation separates the more volatile fractions, generally requiring first an atmospheric fractionator and then a second vacuum stage in order to remove the distillable fractions from the bitumen-containing residuum. All of this is accomplished in a continuous process. The temperatures shown in Figure 7 approximate the boiling ranges of the fractions being separated. A final cut-point temperature of 450 °C is necessary to separate the bitumen-containing residuum from the crude oil. The temperatures reported are atmospheric equivalent vapor temperatures (AEVT), because reduced pressure is used at this point. The actual temperature needed is dependent upon the pressure and the nature or makeup of the crude oil being processed. As indicated by the heavy line in Figure 7 the most direct means of manufacturing is by straight reduction to bitumen of grade AC-2.5 to AC-40.

4.2. Solvent Separation Methods

Solvent treatment is often applied to a vacuum residuum in order to recover additional quantities of gas oil and lube distillates (deasphalted oil) and obtaining a bitumen fraction (precipitated asphalt) at the same time. This process is known as *solvent deasphalting* (SDA) and is commonly carried out using propane and/or butane as the solvent [25]. The *Residuum Oil Su-*

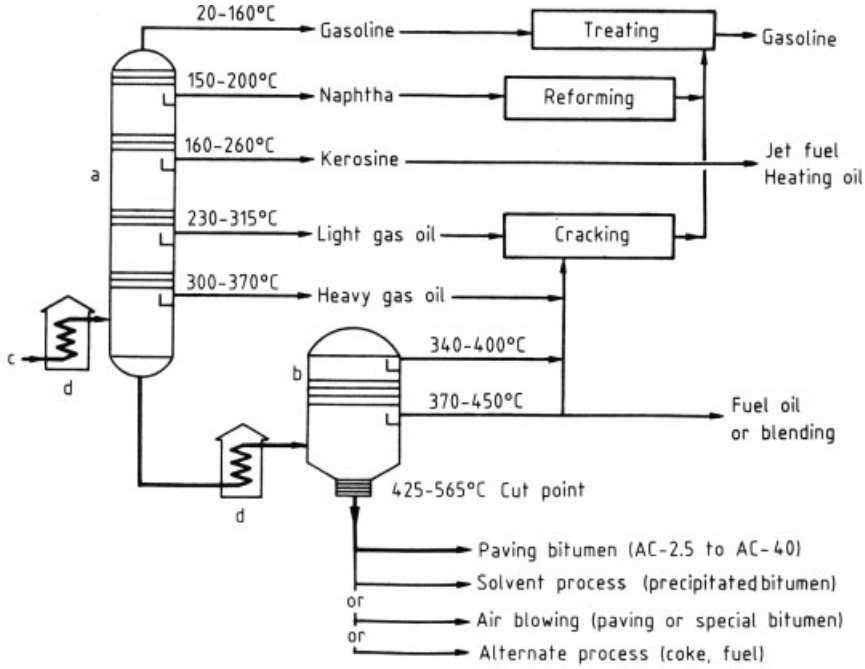

Figure 7. Distillation of petroleum crude oil as first step in bitumen refining
a) Atmospheric fractionation; b) Vacuum fractionation; c) Crude feed; d) Heater

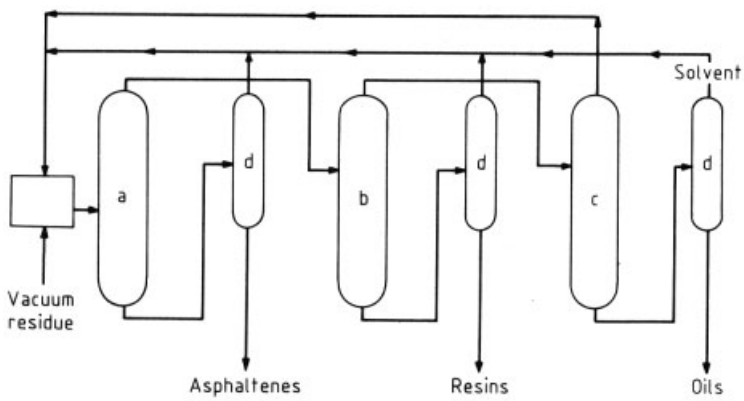

Figure 8. Supercritical extraction of bitumen residuum (reproduced with permission from Kerr McGee Ref. co.)
a) Reactor (temperature t_1); b) Reactor (temperature t_2); c) Reactor (temperature t_3); d) Separator

percritical Extraction (ROSE) process has been used to provide oil and bituminous fractions of different characteristics [26], [27]. In this process, illustrated in Figure 8, the residuum is contacted or mixed with the solvent at elevated temperature and pressure, then separated into an asphaltene concentrate and solvent. The extract fraction is then submitted to a second separator at higher temperature and pressure for separation of a resin fraction. In a similar manner the oil fraction is next separated. Both the SDA and ROSE processes provide bituminous fractions that can be used in the manufacture of bituminous products.

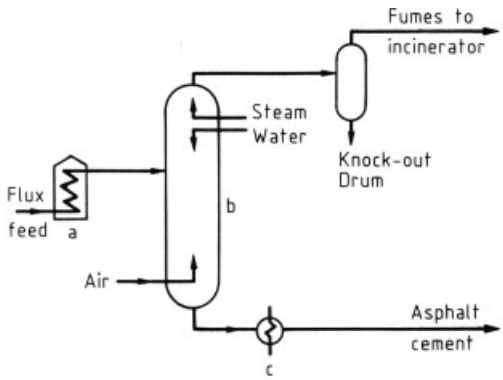

Figure 9. Continuous air blowing
a) Flux feed; b) Heat; c) Air; d) Steam; e) Water; f) Oxidation tower; g) Fumes to incinerator; h) Knock-out drum; i) Heat exchange; j) Asphalt cement

4.3. Continuous Air Blowing

The continuous air-blowing process is used in the manufacture of paving grades only when the available residuum feed must be increased in viscosityor when improved temperature susceptibility is required. As illustrated in Figure 9, this process involves the continuous processing of a vacuum residue (base stock or flux) through an oxidation tower held within the range of 230 – 270 °C while air is passed through the hot flux. This causes conversion within the flux, resulting in higher softening point, higher viscosity, and lower penetration. This process is occasionally used in European and Eastern countries for preparing cement grades, but it is seldom used in the Americas for that purpose.

4.4. Manufacturing Routes for Cement Grades

Bituminous cements may be prepared by any one of several routes depending on the crude being used and the processing applied [28]. If the crude feed permits, straight reduction to the desired grade is the most direct and most commonly practiced method. Blending or mixing of crude feeds is often practiced. If two or more crudes are processed separately with one yielding a low-viscosity residuum and the other a high-viscosity residuum, the products can be blended. If only a low-viscosity residuum is made, it can be blended with precipitated bitumen or it can be air-blown. If only a high-viscosity residuum is made, its viscosity can be regulated during distillation or it can be blended back to lower viscosity with gas oils or higher-

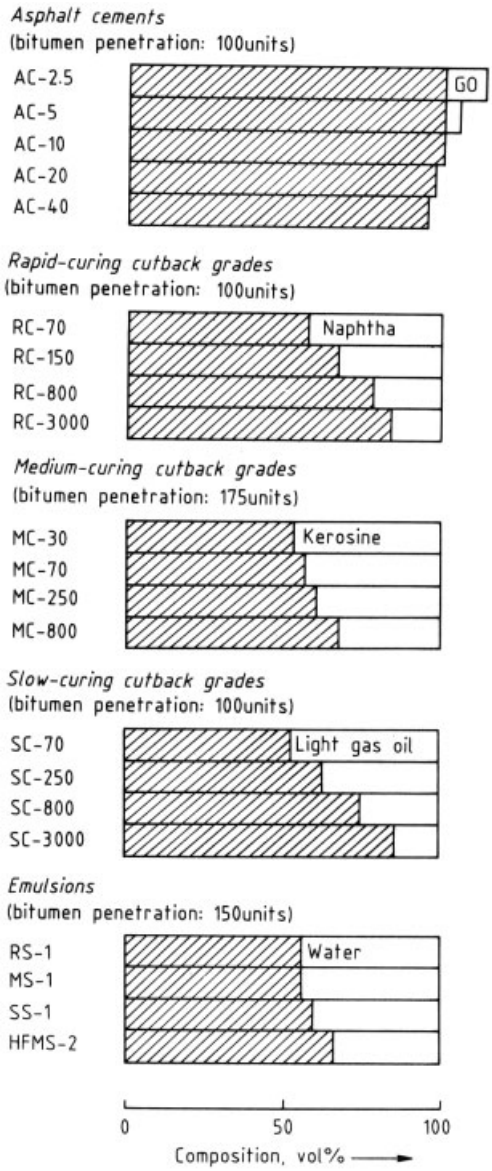

Figure 10. Composition of road bitumen (asphalt cements, cutback grades, and emulsions)
Hatched area: Bitumen; GO = gas oil; RS = rapid setting; MS = medium setting; SS = slow setting; HFMS = high-float medium setting

boiling fractions. Thus the refiner has several choices depending on his crude feed composition and the process facilities available.

4.5. Road Bitumen Compositions

In order to provide bituminous products of suitable handling and application characteristics, a great variety of types and viscosity grades are used. Some examples are compiled in Figure 10, which first shows the proportional amounts of gas oils (GO) needed or removed by distillation in order to meet the various viscosity grades of asphalt cement, i.e., AC-2.5 to AC-40 [29], [30].

The various cutback grades [31] are made by blending bitumen of specific penetration with naphtha, kerosine, and light gas oil. They have different curing properties (rapid curing, medium curing, slow curing), each in different viscosity grades to accomodate construction practices. Bitumen – water emulsions [32] are also made and used in rapid-, medium-, or slow-setting types, again to accommodate construction practices.

4.6. Batch Air Blowing

Batch air blowing is considered separately from the continuous process because it is used primarily in the manufacture of products that have a high softening point and are not used for paving. The geometry of the processing unit is the same as shown in Figure 9, but the flux feed-stock and the products obtained are different.

Air blowing is sometimes referred to as air conversion because chemical reactions take place, whereas in distillation the fractions are separated physically and do not react. To understand the air-conversion process better, reference is made to the major variables [33] involved (Table 2). Flux source and flux viscosity are the most important parameters, because they relate to the crude source being used and to its regulation to flux consistency during refinery distillation. Temperature, air-dispersion features, and catalysts are also control factors but are more nominal as variables. There are also five dependent variables involved, of which softening point and penetration are the most important.

Table 2. Major variables in air conversion of bitumen

Independent variables	Dependent variables
flux source or composition	softening point – penetration relationship
flux consistency	product composition
temperature of reaction	reaction rate
air rate or degree of dispersion	exothermic heat of reaction
use of catalysts	effluent losses

The batch process is best described graphically by following the temperature control as a function of time (Fig. 11). Product status and product endpoint are controlled by periodic determination of the softening point, as illustrated by the test values in parentheses in this figure. The reaction is exothermic with a differential heat of reaction of about 4 – 8 J per kilogram of bitumen per degree centigrade rise in softening point [34]. Generally this is more than needed to maintain a temperature within 200 to 270 °C throughout the blowing process; therefore, small amounts of water are injected for cooling as shown in Figure 11. The temperature of the reaction should generally be lowered toward the end of the process. This helps in stabilizing the endproduct, which otherwise tends to drop back in softening point. There are many good literature references on the subject of air blowing [35–37].

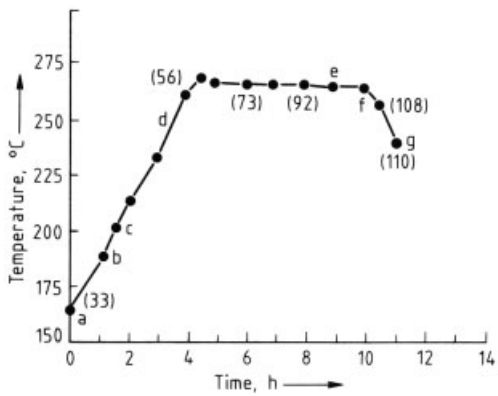

Figure 11. Batch air blowing
a) Begin charge and heat; b) End charge; c) Air injection start; d) Water injection start; e) Air injection reduced, hold test; f) Batch approved, circulate; g) Air and water off, unloadSoftening points of samples (in °C) are given in parentheses

The relationship between softening point and penetration is the most important consideration

in batch air blowing. This is largely dependent on the composition of the flux or feed-stock and its inherent qualities (see Chapter 5). Figure 12 illustrates the relationship between softening point and penetration that results when blowing two different fluxes. In this case the coating flux has a lower viscosity than the roofers flux, which accounts for differences in blow-down characteristics (relationship between softening point and penetration). This also shows that different grades of roofing bitumen may be obtained by stopping the reaction at appropriate softening point levels. Types I through IV are normally used for built-up roofing construction [38], whereas shingle coating is used in the manufacture of prepared composition roofing (for physical properties of various roofing grades, see Table 10,).

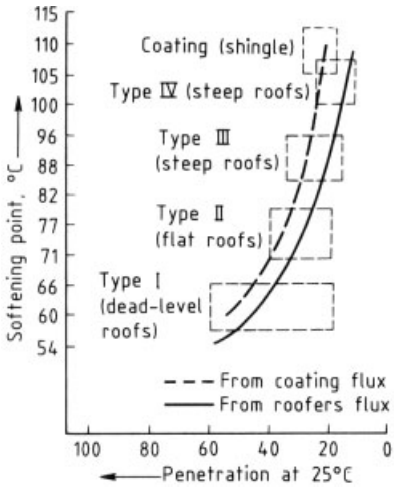

Figure 12. Blow-down curves for roofing bitumen

5. Composition of Bitumen

Composition is readily defined by reporting the mass fractions of the major hydrocarbon types or classes contained in bitumen. Many methods have been proposed for accomplishing this [39–44] , each of which attempts to resolve the overall complexity of the mixture. For the purpose of this discussion, a method now in ASTM Standards is explained in more detail [45]. This involves separating the bitumen into four generic fractions, using dispersion and precipitation in *n*-heptane, filtration, and elution – adsorption chromatography (Figure 13). As the result of many analyses of bitumen from many sources and modes of manufacture, it becomes apparent that all bitumens contain each of these four fractions, but in different proportions. The proportions also depend on the distillation cut-point and the crude from which the bitumen is derived. Figure 14 illustrates how each of the fractions may be distributed as dependent upon cut-point. It also shows that of the three petrolene fractions, saturates decrease in quantity as distillation cuts deeper. Polar aromatics increase at the same time.

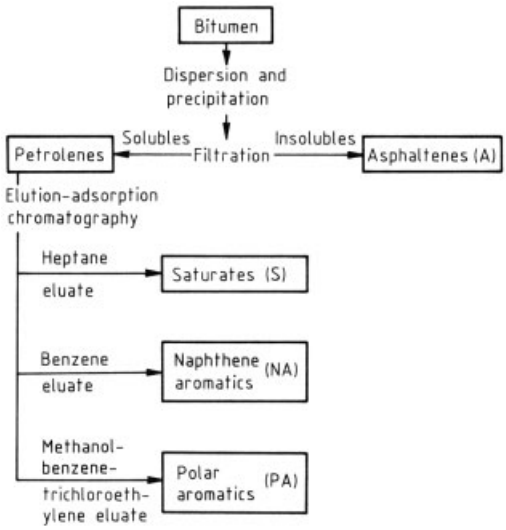

Figure 13. Scheme for separation of bitumen into four fractions

Another way of understanding the effect of distillation is to consider the composition of bitumen in various production stages: first the flux, then the paving bitumen, and finally the pitch (low penetration residuum; Table 3). As distillation progresses, the saturate and the naphthene aromatic fractions decrease, whereas the polar aromatic and asphaltene fractions increase. This shows that the first two fractions, on an average, are lower boiling than the polar aromatics, with distillation having no effect on the percentage of asphaltenes. Also, distillation increases the softening point and decreases the penetration of the bitumen residuum.

Bitumens from different crude sources have different compositions. This is illustrated in Table 4, where bitumens of the same viscosity but

Table 3. Effect of straight reduction on composition of a Venezuelan crude residuum

Fraction	Low-viscosity flux	Paving cement	Hard pitch
Saturates, wt %	16.7	13.0	3.7
Naphthene aromatics, wt %	41.4	34.7	27.2
Polar aromatics, wt %	29.8	37.4	45.8
Asphaltenes, wt %	11.5	14.6	23.0
Softening point, °C	28	46	85
Penetration at 25 °C	≧300	89	5
Kinematic viscosity at 98.9 °C, cST	390	–	–

from different crude sources are shown to contain different proportions of the four fractions. These differences are often related to the API gravity of the crude from which the bitumen is derived [28].

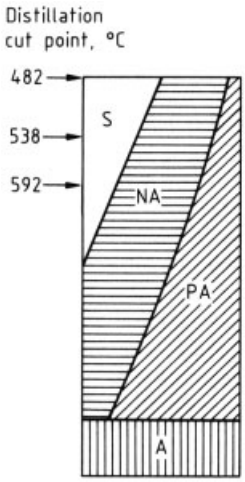

Figure 14. Distribution of fractions within bitumen
S = saturates; NA = naphthene aromatics; PA = polar aromatics; A=asphaltenes

As was indicated previously, air blowing or air conversion involves chemical changes. This is evidenced by a change in the proportion of the four fractions as the result of air blowing; see Figure 15. No significant change in the percentage of saturates is found, but there is a decrease in both naphthene aromatics and polar aromatics. Air blowing always results in a significant increase in asphaltenes. Therefore, the naphthene aromatics probably convert to polar aromatics and the polar aromatics in turn to asphaltenes. As would be expected when the total liquid fractions are decreased and the solid fractions increased, the softening point will increase and the penetration will decrease in the resulting composition.

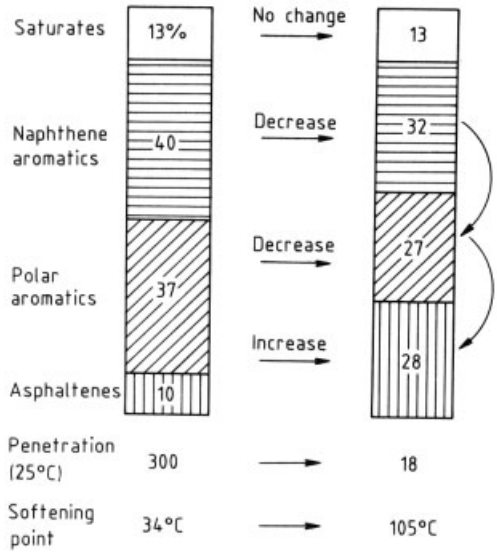

Figure 15. Chemical changes during air blowing

5.1. Elemental Analysis and Molecular Mass

The elemental analysis of a bitumen has little significance with respect to source, mode of manufacture, or service qualities. The same is true for its average molecular mass. Small differences can be measured, see Table 5, which serve the technologist in distinguishing petroleum-derived bitumen from bitumen of other sources [46]. For example, the carbon and hydrogen contents are always typically in the range given. Nitrogen and oxygen levels are always low and insignificant. The iron and nickel contents are

Table 4. Effect of crude source on composition of straight reduced asphalt cement AC-20

Crude	Arabian	United States	Venezuelan
Saturates, wt %	12.0	10.0	8.5
Naphthene aromatics, wt %	48.9	38.7	30.3
Polar aromatics, wt %	30.8	36.3	41.5
Asphaltenes, wt %	8.0	14.6	19.1
Viscosity at 60 °C, Pa · s	200	200	200
Viscosity at 135 °C, cSt	352	445	540
Penetration at 25 °C	50	85	130
Penetration ratio 4 °C/25 °C	24	30	41
RTFO viscosity at 60 °C, Pa · s	6100	7700	8650

Table 5. Elemental analysis and molecular mass characterization of flux and AC-20 grades

	Hawkins, United States		La Rose, Venezuela	
	flux	AC-20	flux	AC-20
Average M_r	910	1082	929	1038
Carbon, %	85.03	85.10	86.01	85.92
Hydrogen, %	9.95	10.02	10.40	10.55
Nitrogen, %	0.52	0.60	0.58	0.65
Oxygen, %	0.51	0.56	0.40	0.35
Sulfur, %	4.10	4.21	2.90	3.20
Iron, ppm	2	2	7	13
Nickel, ppm	15	17	58	71
Vanadium, ppm	20	24	417	525

also quite low, whereas the vanadium content seems to be characteristic of the origin of the crude. Bitumen from Venezuelan crudes always seems to have relatively high vanadium content, which on occasion has been used as a means of fingerprinting.

5.2. Physical, Structural, and Colloidal Aspects

The four fractions differ from each other significantly with respect to their physical characteristics. This is illustrated in Table 6, which shows that the saturates and naphthene aromatics have a relatively low density and low viscosity and thus may be classified as liquids. The polar aromatics and asphaltenes have a relatively high density and a very high viscosity, making them solids. The overall consistency of a bitumen therefore is quite dependent on the proportion of the fractions present.

Each fraction can be characterized structurally by identifying its carbon atom distribution [47] along with its molecular mass [48]. Carbon atom distribution is the normalized mass fraction of carbon atoms in each of the three types of hydrocarbon structure. Table 7 illustrates a typical example of an AC-20 bitumen; it shows that the saturates are largely made up of paraffins or paraffin chains on naphthene rings but contain no aromatic compounds. The naphthene aromatics also have a high percentage of paraffin carbons but these are present as side chains on naphthene and aromatic rings. In this case, presumably all molecules contain at least one aromatic ring; otherwise they would show up in the saturate fraction. The polar aromatic and asphaltene fractions have increasing amounts of aromatic carbon but each with appreciable amounts of paraffin carbon as side chains [49]. Minor percentages of sulfur, oxygen, nitrogen, and trace metals are also present, which are considered to have a minor effect on the structural character of the average molecule present. The range in composition generally found in an AC-20 bitumen along with the corresponding structural identities is summarized in Table 8.

Table 6. Physical characterization of the four generic fractions

Fraction	Density at 20 °C, g/cm^3	Viscosity at 60 °C, Pa · s	Physical state
Saturates	0.87	0.1	liquid
Naphthene aromatics	0.97	2.2	liquid
Polar aromatics	1.07	>100000	solid
Asphaltenes	1.15	not determinable	solid

Table 7. Typical carbon atom distribution and molecular mass of AC-20 fractions

Fraction	Carbon atom distribution in			Average M_r
	paraffin chains, %	naphthene rings, %	aromatic rings, %	
Saturates	69	31	0	610
Naphthene aromatics	41	34	25	725
Polar aromatics	36	26	38	1090
Asphaltenes	37	13	50	5300
Average	41	30	29	1275

Table 8. Range of composition and structural characterization of typical AC-20 fractions

Fraction	Mass fraction, %	Structures identified within fraction
Saturates	5 – 15	pure paraffins + pure naphthenes + mixed paraffin – naphthenes
Naphthene aromatics	30 – 45	mixed paraffin – naphthene – aromatics + sulfur derivatives
Polar aromatics	30 – 45	mixed paraffin – naphthene – aromatics in multiring structures + sulfur, oxygen, nitrogen derivatives
Asphaltenes	5 – 20	mixed paraffin – naphthene – aromatics in polycyclic structures + sulfur, oxygen, nitrogen derivatives

Whereas the distribution of the carbon atoms in each fraction helps in identifying the chemical structures, a colloidal description helps in explaining the role played by each fraction. Bitumens are considered to be colloids [50], [51], in which certain high molecular mass constituents are dispersed or peptized in another phase of lower molecular mass. Asphaltenes are generally considered to be the dispersed phase, with the mixture of saturates plus naphthene aromatics acting as the dispersing medium or the continuous phase. The polar aromatics may serve to protect the colloidal system by keeping the asphaltene micelles in a flocculated state. This is illustrated in Figure 16, in which the polar aromatics are shown to surround the asphaltene micelles closely. As a rule bitumen colloids are very stable and cannot be separated unless severely diluted by a nonpolar solvent, such as aliphatic C_3 to C_8 hydrocarbons. This is the initial step taken in most composition analyses, as well as in commercial solvent separation processes. Another practical aspect of the colloidal concept is in developing studies related to service characteristics. This includes adhesion to mineral aggregates, as in road building, and weathering when exposed in roofing applications.

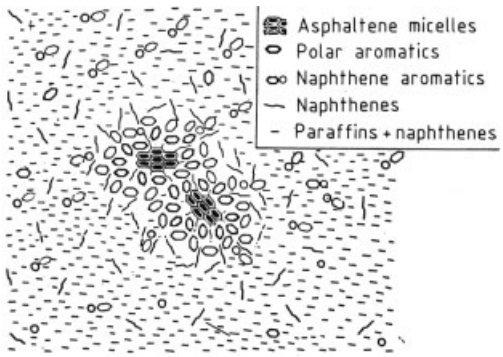

Figure 16. Flocculated asphaltene micelles

6. Physical Properties

Physical test characteristics of bitumen are of utmost importance, not only when handling or

Table 9. Physical test characteristics of straight reduced AC-20 bitumen from various crude sources

	Boscan, Venezuela	Hawkins, Texas	Arabian heavy, Saudi Arabia	Gach-Saran, Iran
Viscosity at 60 °C, Pa · s	200	200	200	200
Kinematic viscosity at 135 °C, cSt	495	451	430	406
Penetration at 25 °C (100/5)	130	80	80	51
Penetration at 4 °C (200/60)	42	26	25	17
Penetration at 0 °C (200/60)	31	19	18	9
Softening point (ring and ball), °C	53	49	47	48
Flash point (open cup), °C	255	320	300	290
Fraass break point, °C	−13	−7	−7	−6
Specific gravity at 25 °C	1.035	1.015	1.028	1.023
Thin-film oven test, loss, %	0.6	0.0	0.0	0.0
Penetration ratio at 25 °C	61	62	64	65
Viscosity ratio at 60 °C	4.8	2.3	2.7	2.2
Ductility at 25 °C	75	>100	>100	>100

compounding, but during its service life. The thermoplastic nature of bitumen dictates a control of temperature and viscosity throughout all stages of storage, movement, and compounding. In service, its flow properties, temperature susceptibility, low-temperature flexibility, and ductility contribute to the life of the product. The crude source from which the bitumen is derived affects all these properties. This is illustrated in Table 9, where straight reduced bitumens of the same viscosity at 60 °C but of different origin are compared. Temperature susceptibility (penetration ratio, viscosity ratio), high-temperature viscosity, softening point, and Fraass breaking point vary considerably. Properties such as hardening by the thin-film oven test and ductility relate to service qualities.

The relationship between viscosity at 60 °C and penetration at 25 °C is very useful when identifying differences between crude sources. For example, a bitumen from a single crude source at different viscosity levels will plot viscosity versus penetration as a straight line on a log-log basis. Bitumens from different sources will then show a nesting of lines at different levels and often at slightly different slopes. Considering the penetration at a given viscosity level (or viscosity at a given penetration level), a quality test characteristic dependent upon crude source becomes available [52].

Temperature susceptibility (change in consistency with temperature) is related to crude source. This is sometimes indicated by factors such as penetration index, penetration ratio, or penetration – viscosity index [53–57]. Thus it is not unusual to find that one of the above factors correlates with the penetration at 25 °C for a viscosity of 200 Pa · s at 60 °C as shown in Figure 17. Another example of this is shown in the data in Table 4 where the penetration ratio on another basis and the high-temperature viscosity directionally relate to the penetration at 200 Pa · s.

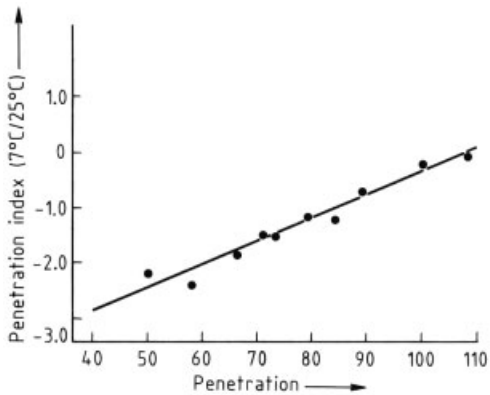

Figure 17. Penetration index (pen ratio 7 °C/25 °C) versus penetration at 25 °C for a viscosity of 200 Pa · s at 60 °C

Air-blown bitumens are represented by another set of physical properties; the bitumen grade must be identified and the softening point controlled. These bitumens are generally blown from a base residuum or a flux with a viscosity of 200 to 1000 cSt at 98.9 °C and from a crude source known to meet the end product requirements. The blowdown characteristics of such bitumens are illustrated in Figure 12 and described in Section 4.6. Table 10 shows the test limits for four grades or types of blown bitumen used in roofing applications. Blown bitumens are characterized by better temperature susceptibil-

Table 10. Physical test characteristics of air-blown bitumen for use in roofing applications

	Felt saturant	Dead level BU roofing	Steep BU roofing	Shingle coating
Softening point (ring and ball, °C)	52 – 58	58 – 66	85 – 96	102 – 113
Penetration at 0 °C (200/60)	> 13	> 3	> 6	> 6
Penetration at 25 °C (100/5)	45 – 75	18 – 60	15 – 35	17 – 26
Penetration at 46 °C (50/5)	<200	90 – 180	< 90	< 40
Ductility at 25 °C, cm	> 20	> 10	> 3	> 2.5
Flash point (open cup), °C	>260	>225	>225	>260
Solubility in trichloroethylene, wt %	> 99	> 99	> 99	> 99

ity (as indicated by closer limits on penetration at 0 and 46 °C) and lower ductility, compared with straight reduced bitumens. These properties contribute to better flow resistance and better outdoor weathering, qualities needed for roofing applications.

Rheological Behavior Bitumen is often described as a thermoplastic material that can repeatedly soften when heated and harden when cooled. Actually it is more accurate to describe bitumen as a viscoelastic substance, that is, one that acts as a viscous liquid at ambient and higher temperatures and as an elastic solid at lower temperatures. Such qualities are referred to as rheological character or rheological behavior; they are measured in relative units of flow and deformation [58]. To the engineer, this is often described in terms of mechanical properties, that is, strain responding to an applied stress. Mechanical properties are also conveniently described by the stiffness modulus [59], which in simple terms is the ratio of tensile stress to tensile strain. How well a road bitumen aggregate mixture resists stressing and seasonal effects in part depends on the inherent characteristics of the bitumen, or on the changes that occur during refining. The time of loading on a road surface and temperature are two important independent variables that determine the stiffness.

Figure 18 illustrates how a typical AC-20 bitumen would respond to different loadings and to temperature. The duration of loading in this case varies from a long loading time (approaching that of a standing load) to a very short loading time (100-km/h vehicular traffic). As would be expected, stiffness is higher at the shortest loading time, at lower temperature, and for higher viscosity grades of bitumen. Rheological characteristics of bitumens and mechanical properties of asphalt mixes have been the subject of much research [60–62]. This is attributed to the potential in better understanding road distress problems.

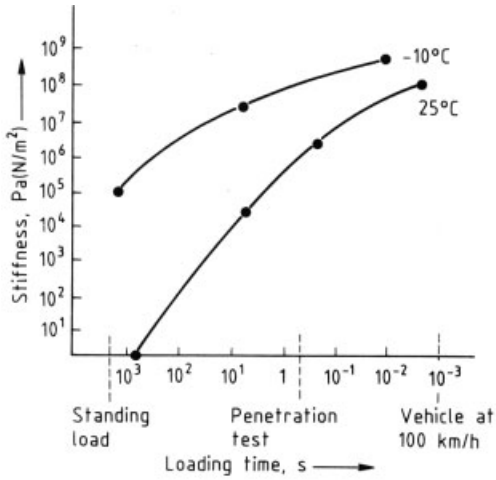

Figure 18. Bitumen stiffness as a function of loading time and temperature

The viscoelastic properties of the three petrolene fractions (as described in Chapter 5) are shown in Figure 19. Here the saturates, because of their makeup of saturated hydrocarbons, are the least susceptible to temperature change. The polar aromatics on the other hand are the most susceptible and, of course, have a much higher viscosity at 25 °C. The composite mixture of the three fractions along with the asphaltenes results in a typical viscosity as represented by the asterisk in Figure 19. This again supports the classification of saturates and naphthene aromatics as liquids and of the polar aromatics and asphaltenes as solids. Again, the rheology of bitumens may be attributed to the composite effect of the viscoelastic properties of the fractions making up the bitumens.

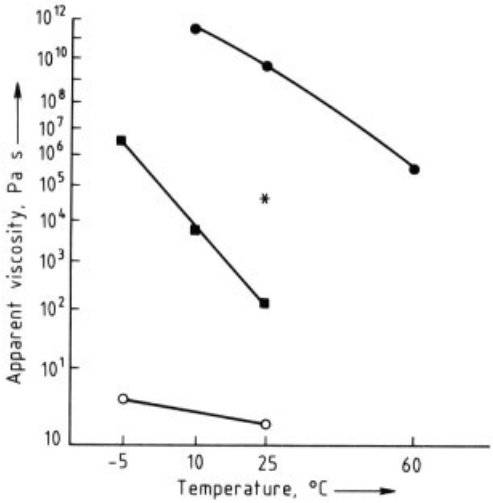

Figure 19. Viscosity as a function of temperature of the various petrolene fractions
• Polar aromatics; ■ Naphthene aromatics; o Saturates; ∗Viscosity of AC-20 bitumen at 25 °C

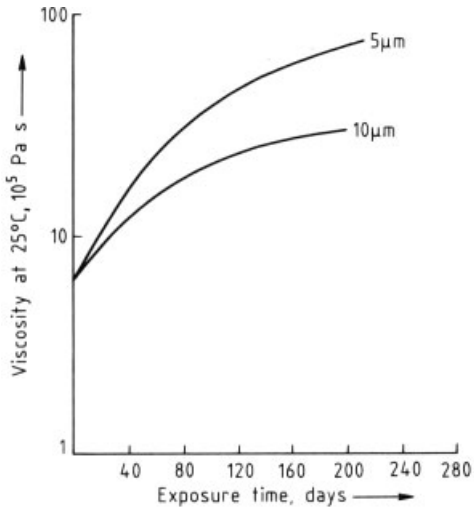

Figure 20. Viscosity increase of thin AC-40 bitumen films by exposure to air

7. Chemical Reactions of Bitumen Surfaces

In order to test the performance of bitumen surfaces, exposure to air, water, and chemicals is investigated.

Air. Under ambient conditions bitumen oxidizes slowly depending on the specific surface it exposes. The effect is always an increase in viscosity and a higher softening point. To control the surface effect, use is made of cover aggregate or films of appreciable thickness. For example, in road building the normal 10–30-μm films of bitumen in the aggregate mixtures are protected by closely compacting the mixture. In this way, only a very small fraction of the bitumen is exposed to atmospheric air and sunlight. In roofing and waterproofing applications, the exposed films are several millimeters thick, which permits gradual weathering of top layers or microsurfaces. Alternatively, even thick films may be protected by layers of sand or fine chipping; this is mostly practiced in European countries. In all these cases the rate of diffusion of air into the composite mixture or into the thick film is substantially reduced.

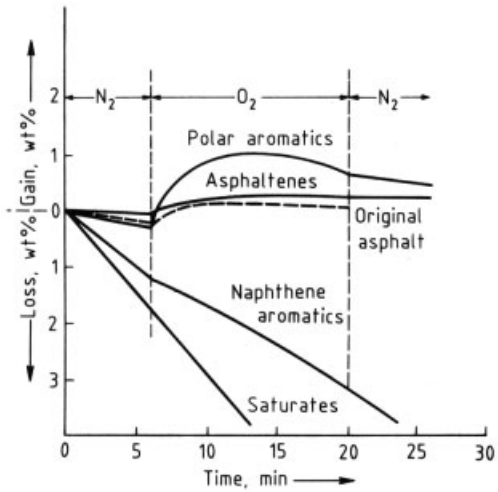

Figure 21. Gain-loss relationship of petrolene fractions in sub-micron films

The increase in viscosity of exposed thin films is illustrated in Figure 20. This shows that the thinner film (5 μm) increases faster in viscosity in the case of an AC-40 bitumen at room temperature [63]. In order to understand the mechanism of thin-film oxidation better, a submicron film of each of the four generic petrolene fractions (Chap. 5) was alternately exposed to nitrogen and oxygen at a temperature of 150 °C [64] (Fig. 21). This showed that under nitrogen a weight loss (volatilization) always takes place, whereas under oxygen a weight gain predomi-

nates for both the polar aromatics and the asphaltenes with only a slight effect on the saturates and naphthene aromatics. In the original bitumen there is first a loss under nitrogen and then a gain under oxygen, apparently a composite effect.

For specification purposes the thin-film oven test was developed [65] to show how the properties of bitumen change when a relatively thick film (3.2 mm) is exposed in a tray to 163 °C for 5 h. In the rolling thin-film oven test [66], film thickness and exposure time are reduced by providing a rolling action during the test (see also Chap. 3). Both tests were developed in order to provide a reasonable control of hardening or viscosity increase as the result of hot-mix processing of bitumen. These tests are now routinely being specified on paving grades of bitumen in the United States.

Water. When bitumen and water are in permanent contact, the solubility of water in the bitumen is of the order of 0.001 – 0.01 wt %. This is partly caused by the presence of small quantities of water-soluble inorganic salts or other filler materials. Ordinarily bitumen possesses good resistance to water, although it is not completely impermeable (ca. 1×10^{-10} g cm^{-1} h^{-1} Pa^{-1}). In practical applications this is negligible because thick films or layered membranes are used.

Chemicals. Ordinarily bitumen resists reaction with mild chemicals and all environmental effects other than those described above. Bitumen reacts with aggressive chemicals. Strong sulfuric, nitric, and hydrochloric acids react to form substitution products. Dilute acids sometimes cause bitumen to harden slowly, whereas dilute bases often cause emulsification as the result of a reaction with the acidic components in the bitumen. At moderate temperature potassium permanganate and other strong oxidizing agents react similarly to air oxidation. At high temperatures these agents have the effect of fast burning. Chlorine gas reacts with bitumen at 100 °C resulting in substitution, whereas at higher temperature hydrogen chloride is formed. At 150 °C and above sulfur reacts with perceptible rate with the liberation of hydrogen sulfide, similar to the reaction with oxygen [67].

Much work has been done using small amounts of catalysts during the air-blowing operation, generally at 225 °C and above. The purpose is to provide a higher penetration at a given softening point level, or to reduce the conversion time. Iron(III) chloride and phosphorus pentoxide [68] have been used commercially for this purpose, and other agents have been suggested. Some refer to these agents as chemical reactants rather than catalysts, because they are either consumed or are converted chemically.

8. Handling and Storage

Fluidity or viscosity is the primary consideration in pumping, transporting, and storing of bitumen. Viscosity is affected by temperature and by the type and grade involved. Figure 22 relates viscosity to temperature for two cement and two air-blown grades. Also shown is a line indicating the approximate design pumpability. At higher temperatures and lower viscosities pumping is optimized, whereas at lower temperatures and higher viscosities pumping efficiency decreases rapidly. Thus viscosity and its control by temperature is an important consideration in respect to all handling operations.

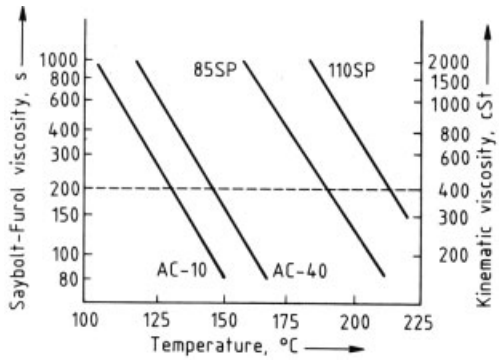

Figure 22. Viscosity as a function of temperature for asphalt cements (AC-10 and AC-40) and for air-blown bitumen (85 °C softening point and 110 °C softening point) – – – Design pumpability

If bitumen is held in hot storage, certain safety considerations are advisable. For example, all bitumens, whether straight reduced or air blown, emit small amounts of light hydrocarbons (C_1 to C_4 and above) into the head-space above the hot liquid. The amount of these light hydrocarbons

emitted depends largely on the storage temperature but also on practices during manufacture. The concentration of these hydrocarbons in the headspace above the liquid also depends on venting as well as loading and unloading practices. All of this becomes a safety consideration as dependent on the concentration of the hydrocarbons, the oxygen content in the headspace, and a possible source of ignition.

The oxygen content of the headspace above the bitumen also depends on air movement into and out of the tank. The ignition source can be extraneous or from pyrophoric iron sulfides that develop from corrosion on tank walls and ceilings. More on this subject may be found in the literature [69]. With storage temperatures being an important factor affecting headspace hydrocarbons, the maximum storage temperatures that are recommended in the literature for each type and grade of bitumen should be observed [70]. Blanketing of vapor spaces with either nitrogen or fuel gas has also been effective in reducing oxygen concentrations. In view of the above, the flash point of the bituminous material should be known, but it should not be the only consideration in safe handling.

9. Uses of Bitumen

Bitumen is a primary engineering material, functioning as a binder in road construction and as a weatherproofing membrane in roofing and structural applications. Its thermoplastic nature, its water resistance, and its adhesion toward most other substances make it unique as a building material. On a volume basis it nearly equals that of lumber used in the United States and surpasses that of either steel or Portland cement. Almost every home, building, or traffic area uses bitumen in some form or other. This attests to its reasonable cost and versatile nature.

9.1. Bitumen in Road Building

Bitumen serves primarily as a binder in asphalt compacted mixes, which in turn are widely used in many types of road, street, runway, revetment, and parking area applications. Asphalt hot mixes are typically prepared either by batch mixing plants or by continuous-flow drum mixing plants. This involves the selection and grading of mineral aggregates, passing them through a heated drum mixer during which the bitumen binder is added in controlled amounts (4 – 10 wt %). Mixing is carried out at about 150 – 160 °C, after which the hot mix is transported to road sites, where it is placed and compacted on prepared bases. Traffic loads and their volumes influence the road design, that is, the thickness and type of construction used. Over 90 % of all traffic-carrying areas today are constructed with asphalt mixes.

Table 11 lists specification requirements for three asphalt cement grades commonly used in the United States today. These requirements are published by the American Association of State Highway and Transportation Officials (AASHTO); ASTM and DIN 1995 also list specifications that are very similar to these.

9.2. Industrial Uses

There are over one hundred different industrial applications or products in which bitumen is used, more than can be described here. This section, therefore, is limited to three of the largest volume applications, each being in the roofing field.

In North America *asphalt composition shingles* are now used on more than 95 % of private dwellings. These shingles consist of a glass fiber or cellulosic felt backing, coated on both sides with a filled asphalt and surfaced on the weather side with mineral granules. Whereas most of the shingles used in past years were made with organic felts, today's manufacture largely involves the use of fiberglass.

Second in importance to the above is the use of bitumen in *asphalt built-up roofing systems.* These consist of successive layers of asphalt-saturated felt, bonded together with a hot mopping of air-blown roofing-grade asphalt (Table 12). This system is applied to roof inclines from dead level up to 50 %. Whereas the present specification practice is based on the use of grades of increasing softening point, a new system has been proposed [71] based on grade control by viscosity at 71 °C.

The development of *single-ply roofing materials* has resulted in a major change in roof-

Table 11. Viscosity-graded asphalt cement

	Viscosity grade		
	AC-5	AC-20	AC-40
Viscosity at 60 °C, Pa · s	50 ± 10	200 ± 40	400 ± 80
Kinematic viscosity at 135 °C, cSt	>175	>300	> 400
Penetration at 25 °C	>140	> 60	> 40
Flash point (open cup), °C	>177	>232	> 232
Solubility in trichloroethylene, wt %	> 99	> 99	> 99
Thin-film oven test			
viscosity at 60 °C, Pa · s	<200	<800	<1600
ductility at 25 °C, cm	>100	> 50	> 25
Spot test (when specified)	neg.	neg.	neg.

Table 12. Asphalt used in roofing

	Grade type *			
	I	II	III	IV
Softening point, °C	57 – 66	70 – 80	85 – 96	90 – 107
Flash point, °C	>225	>225	>225	>225
Penetration				
at 0 °C	> 3	> 6	> 6	> 6
at 25 °C	18 – 60	18 – 40	15 – 35	12 – 25
at 46 °C	<120	<100	< 90	< 75
Ductility at 25 °C	> 10	> 3	> 25	< 1.5
Solubility in trichloroethylene, wt %	> 99	> 99	> 99	> 99

* see Figure 12

ing design and application. This has involved a variety of materials, including polymers and rubbers, but also including those containing asphalt or bitumen. Those identified as modified asphalt, elastomer-modified asphalt or polymer-modified bitumen, are generally made into a membrane of about 4 mm in thickness. It consists of layers of glass fiber mats, glass fiber web, a polyester core, and the polymer-modified asphalt. The polymer is generally atactic polypropylene or styrene – butadiene – styrene block copolymer. This high-strength, flexible, excellent weathering membrane is packaged in rolls, from which it is placed as a single course on a prepared roof deck [72], see Figure 23. This illustrates an application technique that also requires fitting to roof appertunances (pertaining parts) with all joints and laps sealed. This innovative building material holds promise for better and longer lasting roofs on industrial and commercial type buildings.

9.3. Economic Aspects

The economic importance and the various uses of asphalt are described elsewhere in this article. Approximately 75 % is used as binder, mainly for the construction of roads, airports, and waterways. Approximately 20 % is processed into materials for roofs and related applications. The remainder is used for the production of construction proofing agents, paints and coatings, paper, rubber products, electrical cables, and other products of the electrical industry.

As a binder bitumen is almost indispensable for many industries. In the production of asphalt mixes the ratio of gravel, sand, and mineral flour to bitumen is about 20; one ton of such asphalt mixes requires about 9 L of heating oil during processing. The ratio of binder to matrix material (jute, glass fiber mats, glass fiber webs) is about one for single-ply roofing materials.

The consumption of bitumen increased enormously from the mid 1950s because of increased motorization; this changed sharply with the oil price shocks of 1974 and 1979. Asphalt demand in the industrialized countries then began

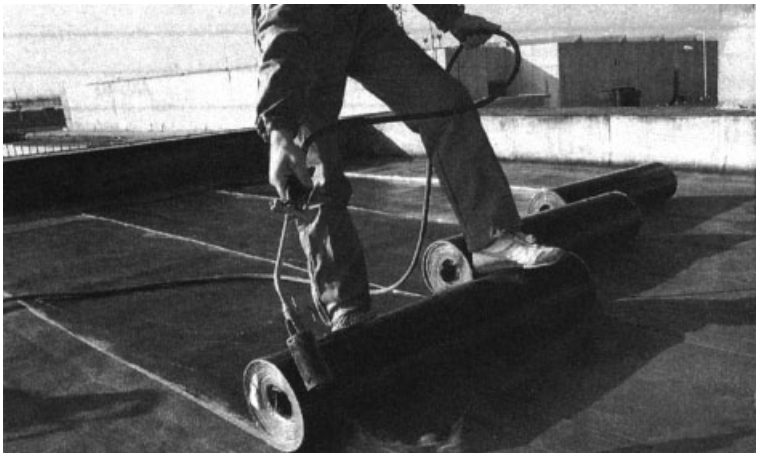

Figure 23. Application of single-ply roofing (reproduced with permission from Technicote Corp., Memphis, Tenn.)

to decrease because of the economic crisis and levelling-off of road construction and building. About 60×10^6 t of bitumen were consumed in the early 1970s, which was lowered to about 40×10^6 t in 1984. In 1982 bitumen consumption of some European countries was as follows (10^6 t):

Federal Republic of Germany	2.78
France	2.12
United Kingdom	1.96
Italy	1.40
Spain	0.85
Portugal	0.40
Sweden	0.44

In the United States the total consumption of bitumen increased since 1982 and reached 27.5×10^6 t in 1984.

10. Environmental and Workplace Exposure

The manufacture and use of straight reduced bitumen, cements, fluxes, or components does not present a problem with respect to air pollution or workplace exposure. At the refinery source each of these products is contained in tanks where vapor pressures are negligible and the emission of hydrocarbons to the atmosphere is insignificant. In the manufacture of hot mixes, blue-smoke emissions were investigated by the Asphalt Institute [73], reporting extremely low levels of potential pollutants (Table 13). All considered emissions (on a time-weighted average) were well within the air standards of the Environmental Protection Agency (EPA). Also, the standards of the Occupational Safety and Health Administration (OSHA) for workplace exposure are easily met, including the polynuclear aromatic compounds, sometimes considered important by health groups. This report concluded that "no significant air pollution or employee health problem can be expected or associated with emissions from hot preparation of asphalt paving mixtures."

Rapid- and medium-curing cutback asphalts intentionally contain from 17 to 42 % of volatile hydrocarbons in the C_7 to C_{15} range. These hydrocarbons come under the broad classification of volatile organic compounds and are considered in some states in the United States to be undesirable air pollutants. Therefore their usage is commonly restricted to the colder months of the year, November through April. This approach is gradually supplanting the old limitation of not more than 8 % of C_8 and higher aromatic compounds in the volatile portion of the cutback asphalt.

Fumes from the air-blowing process have at times constituted an odor problem. This can be resolved by incinerating all of the fumes from the process through a combustion furnace or incinerator where the burning temperature is 750 °C or higher. Some speculation still exists [74] that some of these fumes remain in the blown asphalt, and as such may be emitted when the asphalt is reheated in roofing kettles. In spite of this, an earlier survey [65] reported that petroleum work-

Table 13. Measured emissions at hot-mix asphalt plants compared with OSHA and EPA quality standards

	OSHA standards	EPA air standards	Measured emissions max. (8 samplings)
Carbon monoxide, ppm	50	9	6
Nitrogen dioxide, ppm	5	0.05	0.1
Sulfur dioxide, ppm	5	0.14	2
Hydrogen sulfide, ppm	20		1.5
Phenol, ppm	5		1
Ozone, ppm	0.1	0.08	0.1
Hydrocarbons ($C_7 - C_{14}$), ppm	500		1.5
Hydrocarbons ($C_2 - C_6$), ppm	10	0.24	1
Particulate matter, mg/m^3		0.365	7.2
Polynuclear aromatic compounds (total), mg/m^3	0.2		0.00036
Vanadium-V_2O_5 fume, mg/m^3	0.1		0.0001
Nickel and soluble nickel compounds as Ni, mg/m^3	1		0.00004
Cadmium fume, mg/m^3	0.1		0.00005
Lead and inorganic compounds, mg/m^3	0.2		0.00005

ers showed no greater incidence of disease than nonasphalt workers in similar environments.

11. References

General References

1. H. Abraham: *Asphalts and Allied Substances,* D. Van Nostrand Co. Inc., New York 1961.
2. J. Ph. Pfeiffer: *The Properties of Asphaltic Bitumen,* Elsevier, Amsterdam 1950.
3. R. N. Traxler: *Asphalt – Its Composition, Properties and Uses,* Reinhold Publ. Co., New York 1961.
4. A. J. Hoiberg: *Bituminous Materials: Asphalts, Tars and Pitches,* Interscience, New York 1964.
5. J. W. Bunger, N. C. Li: *Chemistry of Asphaltenes, Advances in Chemistry Series 195,* ACS, Washington, D.C., 1981.
6. "Road and Paving Materials," ASTM Annual Book of Standards, vol. 04.03, Philadelphia 1983.
7. "Roofing, Waterproofing and Bituminous Materials," ASTM Annual Book of Standards, vol. 04.04, ASTM, Philadelphia 1983.
8. DIN 1995: Berlin 1980.
9. American Association of State Highway and Transportation Officials *(AASHTO) Standard Specifications for Transportation Materials,* Washington, D.C., 1983.
10. Institute of Petroleum (IP) *Standards for Petroleum and Its Products,* Part 1, Elsevier, London 1969.

Specific References

11. *Standard Definitions of Terms Relating to Materials for Roads and Pavements,* ASTM D 04.03, Philadelphia 1984, ASTM D 8-83.
12. D. C. Broome, "The Development of Modern Asphalt Roads," *Surv. Munic. City Eng* . 1963, no. 13, 23.
13. I. Dussek: *Trinidad Lake Asphalt,* Handelsgesellschaft Wilh. Asche & Co., Bremen, 1981.
14. *Colliers Encyclopedia* **9** (1981). 144
15. ASTM D 2170-83.
16. D. F. Fink, R. L. Griffin, *ASTM Standard Technical Publication No. 309,* **79** (1961).
17. D. F. Fink, J. J. Heithaus, *ASTM Standard Technical Publication No. 309,* **95** (1961).
18. W. J. Halstead, J. Y. Welborn, *Hwy. Res. Board Record No. 350,* **1** (1971).
19. J. Skog, G. Sherman, *Hwy. Res. Board Record No. 350,* **26** (1971). J. Y. Welborn, *ASTM Standard Technical Publication 424,* **3** (1966).
20. *Bitumen, Teere, Asphalte, Peche* **23** (1972) 149, 508; **25** (1974) 21.
21. R. J. Schmidt, *ASTM Standard Technical Publication 532,* **52** (1972).
22. American Association for State Highway and Transportation Officials: *T 102-74, Standard Specifications for Transportation Materials,* Washington, D.C., 1983.
23. L. C. Krchma, D. W. Gagle, Proc. Assoc. Asphalt Paving Technol. **43 A** (1974) 25.
24. *Fisher/Tag Manual for Inspectors of Petroleum,* Fisher Scientific Co., Pittsburgh 1956, p. 8.
25. Foster-Wheeler, *Hydrocarbon Process* **52** (1973) 110.
26. J. A. Gearhart, S. R. Nelson: *Fifth Annual Industrial Energy Conservation Conference, Houston,* Tex., April 1983.
27. J. A. Gearhart, L. Garwin, *Hydrocarbon Process.* 55 (1976) no. 5, 125 – 128.

28. L. W. Corbett: *Transportation Res. Board 63rd Annual Meeting,* Washington, D.C., Jan. 1984.
29. *Annual Book of Standards,* vol. 04. 03, ASTM D 3381-81, ASTM, Philadelphia 1983.
30. *Standard Specifications for Transportation Materials,* (1983) AASHTO M 226-82.
31. *Annual Book of Standards,* vol. 04.03, ASTM D 2026-72, D 2027-76, D 2028-76, ASTM, Philadelphia 1983.
32. *Annual Book of Standards,* vol. 04.03, ASTM D 977–80, ASTM, Philadelphia 1983.
33. L. W. Corbett, *Ind. Eng. Chem. Des. Dev.* **14** (1975) 181.
34. D. B. Smith, H. E. Schweyer, *Ind. Eng. Chem. Prod. Res. Dev.* **2** (1963) 209.
35. H. Senolt, *Bitumen, Teere, Asphalte, Peche* **20** (1969) 563.
36. J. M. Goppel, J. Knotnerus: *Proc. 4th World Pet. Congress,* Sec. 111 (1955) 399.
37. J. D. Holmgren, Ph. D. Thesis, University of Florida, Florida, June 1954.
38. *Annual Book of Standards,* vol. 04.04, ASTM D 312–78, ASTM, Philadelphia 1983.
39. J. Marcusson, *Z. Angew. Chem.* **29** (1916).
40. A. J. Hoiberg, W. E. Garris, *Ind. Eng. Chem. (an. ed.)* **16** (1944) 294.
41. F. S. Rostler, H. W. Sternberg, *Ind. Eng. Chem.* **41** (1949) 598.
42. G. O'Donnell, *Anal. Chem.* **23** (1951) 894.
43. L. R. Kleinschmidt, *J. Res. Nat'l. Bur. Stand.* **54** (1955) 163.
44. L. W. Corbett, *Anal. Chem.* **41** (1969) 576. H. J. Neumann, *Erdoel Kohle* **4** (1981) 336.
45. *Annual Book of Standards,* vol. 04.04, ASTM D 4124–82, ASTM, Philadelphia 1983.
46. V. P. Puzinauskas, L. W. Corbett, *Asphalt Inst. Res. Ser. Res. Report* 78-1 (Jan. 1978), pp. 1 – 31.
47. L. W. Corbett, R. E. Swarbrick: *Dir. of Pet. Chem. Preprints,* ACS 11 (1966) pp. 13, 161 – 166.
48. Mechrolab Vapor Pressure Osmometer.
49. R. B. Williams, *ASTM STP No. 224* (1958).
50. J. Ph. Pfeiffer, R. N. J. Saal, *J. Phys. Chem.* **44** (1940) 139.
51. R. Dron et al., *Rapp. Rech. LPC* 1978, no. 75 (Mar.).
52. L. W. Corbett, H. E. Schwyer *ASTM Standard Technical Publication*, no. 532, **40,** 1942.
53. J. Ph. Pfeiffer, P. M. Van Doormaal, *Natl. Pet. News* **40** (1938) 40. E. C. Hughes, H. F. Hardman, *Proc. Assoc. Asphalt/Paving Technol.* **20** (1951) 1.
54. W. Heukelom, *Proc. Assoc. Asphalt/Paving Technol.* **42** (1973) 67.
55. W. J. Gaw, *Proc. Assoc. Asphalt/Paving Technol.* **47** (1978) 457.
56. J. A. Lefebvre, *Proc. Assoc. Asphalt/Paving Technol.* **39** (1970) 443.
57. V. P. Puzinauskas, *Proc. Assoc. Asphalt/Paving Technol.* **48** (1979) 646.
58. G. R. Dobson, *J. Inst. Pet. London* **58** (1972) 14.
59. C. Van der Poel, *J. Appl. Chem.* **4** (1954) 221.
60. M. S. Noureldin, P. G. Manke, *Proc. AAPT* **47** (1978) 270.
61. H. S. Pink et al., *Proc. Assoc. Asphalt/Paving Technol.* **49** (1980) 64.
62. F. Bonnaure et al., *Proc. Assoc. Asphalt/Paving Technol.* **46** (1977) 64.
63. J. W. Labout, W. P. Van Oort, *ACS Pet. Div.* 1954, 95.
64. W. H. King, L. W. Corbett, *Anal. Chem.* **41** (1969) 580.
65. R. H. Lewis, W. J. Halstead, *Public Roads* **24,** Apr. – June 1946, no. 8, 1.
66. R. J. Schmidt, *ASTM Standard Technical Publication No. 532,* **52** (1972) .
67. J. R. Tucker, H. E. Schweyer, *Ind. Eng. Chem. Prod. Res. Dev.* **4** (1965) 51.
68. A. J. Hoiberg, *Ind. Eng. Chem.* **19** (1980) 450.
69. L. H. Dimpfl, *Oil Gas J.* **76** (Dec. 29, 1980) 176.
70. *Asphalt Inst., Inf. Ser.* no. 180, Aug. 1981.
71. V. P. Puzinauskas, Paper No. 12, 56, *Proc. 7th Conf. on Roofing Technology,* Gaithersburg, Md., April 1983.
72. W. H. Gumpertz (ed.), *ASTM Standard Technical Publication 790,* 1982.
73. V. P. Puzinauskas, L. W. Corbett: *ACS, Div. Pet. Chem.,* Chicago, Ill., Aug. 1975.
74. R. W. Niemeier: *Spring National Meeting and Petroleum Exposition,* Houston, Tex., Mar. 1983.
75. C. H. Baylor, N. K. Weaver, *Med. Res. Rp. EA 7103, Am. Petr. Inst.*, Washington, D.C., 1971.

Astatine → Radionuclides
Asymmetric Synthesis → Optically Active Compounds
Atomic Absorption Spectroscopy → Atomic Spectroscopy

Automobile Exhaust Control

Martin Votsmeier, OMG AG & Co KG, Hanau, Germany
Thomas Kreuzer, OMG AG & Co KG, Hanau, Germany
Gerhard Lepperhoff, FEV Motorentechnik, Aachen, Germany

1. Introduction 1
2. Pollutant Formation and Limitation . 2
3. Developments in Engine Technology . 4
4. Catalytic Exhaust Aftertreatment, General Concepts 4
4.1. The Physical Design of the Catalytic Converter 5
4.2. The Washcoat 6
5. Catalytic Aftertreatment of Stoichiometric Exhaust Gas 7
5.1. Three-Way Catalysis 7
5.2. Oxygen Storage in Three-Way Catalysts 8
5.3. Precious Metals in Three Way Catalysis 9
5.4. Recent Developments in Three-Way Catalysis 10
6. Catalytic Aftertreatment of Lean Exhaust Gas 10
6.1. NO_x Removal from Lean Exhaust Gas 10
6.2. Catalytic Aftertreatment of Lean-Burn Gasoline Exhaust 11
6.3. Catalytic Aftertreatment of Diesel Exhaust 13
6.4. Removal of Particulate Emissions . . . 14
7. Conclusion and Outlook 14
8. References 15

1. Introduction

The widespread use of automobiles offers a degree of mobility to the individual that was unthinkable just a few centuries ago. However, with increasing car and truck population today it is obvious that the massive use of combustion-engine vehicles is not free of problems, one of the major problems being air pollution. In Germany in 1994 more than 50 % of nitrogen oxide and CO emissions and more than 40 % of hydrocarbon emissions originated from road traffic. A total of 4.1×10^6 t of carbon monoxide, 1.3×10^6 t of nitrogen oxides, and 0.74×10^6 t of hydrocarbons were emitted on German streets in 1994 [1]. The formation of the major pollutants in combustion engines and their impact on environmental and human health are discussed in Chapter 2.

Legal regulations limiting emissions of the major pollutants carbon monoxide, hydrocarbons, nitrous oxides, and diesel soot were passed in the 1970s in the USA and Japan and later also in Europe and most other parts of the world. As a consequence substantial improvement in air quality has been observed. For example, in California the number of smog-alert days decreased from 121 in 1977 before the introduction of catalytic converters to only one smog alert day in 1997, although car use has doubled since 1977. The 1996 regulations already reduced emissions per kilometer by 96 % compared to 1970 levels, with further improvements in 2000.

Chapter 3 gives a brief introduction to engine measures that allow reduction of emissions. While improved engine technology has led to some reduction in raw emissions, the dramatic reduction in tailpipe emissions, as reflected by the increasingly stringent legal limits, would not have been possible without the introduction of catalytic exhaust gas treatment systems [2–5]. Chapter 4 gives a general introduction to the working principle of catalytic exhaust gas treatment and especially focuses on the physical design of the catalyst system. The most effective type of catalyst system today is the three-way catalytic converter that simultaneously reduces concentrations of CO, NO, and hydrocarbons. Three-way catalysts are used with spark-ignition gasoline engines and allow the reduction of pollutant emissions by 96 % compared to the pollutant level before the introduction of catalytic converters. Three-way catalysts are described in Chapter 5.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_189

Carbon dioxide is an unavoidable combustion product of fossils fuels and the only way to reduce CO_2 emissions from internal combustion engines is to construct more energy efficient engines. Due to its role as a major greenhouse gas and due to concerns about global warming, energy efficiency of automobiles has become a major concern for engine design. Today the most fuel efficient engines run with excess air and therefore do not allow application of the proven three-way catalyst. This poses the challenge of developing catalyst concepts that allow fuel-efficient lean-burn operation of the engine while retaining the low pollutant levels reached by three-way catalysts. Aftertreatment systems for lean exhaust gas are treated in Chapter 6.

2. Pollutant Formation and Limitation

In a combustion engine a hydrocarbon fuel is oxidized by atmospheric oxygen. Under ideal conditions only CO_2, water, and heat would be generated (Eq. 1).

$$C_mH_n + (m + 0.25n)\,O_2 \longrightarrow mCO_2 + 0.5n\,H_2O \qquad (1)$$

Under engine conditions, however, ideal combustion is impossible because of the lack of chemical equilibrium and the inhomogeneous gas phase. Typical combustion products resulting from incomplete combustion are:

- Carbon monoxide (CO), a combustion intermediate
- Unburned and partially oxidized hydrocarbons (HC)
- Oxides of nitrogen (NO_x)
- Particulate matter

Carbon monoxide (CO) is formed as an intermediate reaction product during combustion of hydrocarbons. Incomplete combustion due to low combustion temperature, lack of oxygen, or too short a reaction time leads to the emission of CO, a poisonous gas that displaces oxygen from the blood.

Hydrocarbons (HC). Incomplete combustion of hydrocarbon fuels leads to the emissions of unburned and partially oxidized hydrocarbons [6].

In spark-ignition engines hydrocarbons in the exhaust originate mainly from regions in which the flame extinguishes or regions not reached by the flame (quench zones). Typical examples are layers near the combustion chamber walls and crevices, e.g., in the piston ring groves (wall quenching). With very lean mixtures, the flame can extinguish prematurely (flame quenching) at the end of the expansion because of low gas temperatures in the combustion chamber. In highly turbulent mixtures, exhaust gas can surround combustible gas zones before they are reached by the flame front. The HC emissions rate can be largely caused by the combustible gas losses in two-stroke gasoline engines and, to a lesser extent, in supercharged four-stroke engines.

In compression-ignition engines, HC emissions originate in the outer zones of the injection spray, where the fuel has thoroughly mixed with air to produce an nonignitable mixture. Emissions also result from the spray center, which has a very rich mixture, as well as from the chamber wall areas. Finally, near the end of the expansion, fuel vaporizes out of the injection nozzle.

The hydrocarbons emitted from internal combustion engines are always a mixture of various compounds: paraffins, olefins, acetylenes, and aromatics. The composition depends on the combustion system, the operating conditions of the engine, and the fuel.

Oxides of nitrogen (NO_x) are generated at combustion temperatures by oxidation either of atmospheric nitrogen or nitrogen contained in the fuel. During the combustion process mainly NO is formed; NO_2, which is more toxic than NO, is formed by the relatively slow oxidization of NO in the atmosphere. In combustion processes with an excess of air significant NO_2 can be expected. NO_x reacts with hydrocarbons in sunlight to form ozone and photochemical smog. NO_x can increase respiratory illnesses and is a contributor to acid rain. Tropospheric ozone causes breathing difficulties and damages plants. In the upper atmosphere NO_x is involved in a catalytic cycle that leads to the destruction of the ozone layer.

Three mechanisms of NO_x formation must be considered [7]:

1) Thermal NO_x
2) Prompt NO_x

3) Fuel NO_x

Fuel NO_x. Fossil fuels contain organic nitrogen compounds (e.g., amides, amines, heterocyclic compounds) in concentrations of $< 1\,\%$ (coals have higher concentrations). Equation (2) is an example of such a reaction pathway.

$$\mathrm{C}_n\mathrm{H}_m\mathrm{NH}_2 + \left(n + \frac{m}{4} + 1\right)\mathrm{O}_2 \longrightarrow \tag{2}$$

$$n\mathrm{CO}_2 + \left(\frac{m}{2} + 1\right)\mathrm{H}_2\mathrm{O} + \mathrm{NO} \tag{2}$$

Thermal NO_x. The direct oxidization of N_2 by oxygen at high temperatures, the so-called thermal NO_x mechanism, was first described by Zeldovic. At temperatures above 1300 °C NO is formed by the following reaction sequence:

$$\mathrm{O} + \mathrm{N}_2 \longrightarrow \mathrm{NO} + \mathrm{N} \tag{3}$$

$$\mathrm{N} + \mathrm{O}_2 \longrightarrow \mathrm{NO} + \mathrm{O} \tag{4}$$

In rich zones of the flames (oxygen deficiency) the following reaction takes place above 1300 °C:

$$\mathrm{N} + \mathrm{OH} \longrightarrow \mathrm{NO} + \mathrm{H} \tag{5}$$

The amount of NO formed largely depends on:

- The air-to-fuel ratio in the combustion zone
- The temperature in the combustion zones
- The time for which the reaction partners remain at combustion temperature

Prompt NO_x. Fenimore described a mechanism in which reactive radicals formed during hydrocarbon combustion react with molecular nitrogen to form NO:

$$\mathrm{CH} + \mathrm{N}_2 \longrightarrow \mathrm{HCN} + \mathrm{N} \tag{6}$$

$$\mathrm{C} + \mathrm{N}_2 \longrightarrow \mathrm{CN} + \mathrm{N} \tag{7}$$

$$\mathrm{CN} + \mathrm{H}_2 \longrightarrow \mathrm{HCN} + \mathrm{H} \tag{8}$$

$$\mathrm{CN} + \mathrm{H}_2\mathrm{O} \longrightarrow \mathrm{HCN} + \mathrm{OH} \tag{9}$$

$$\mathrm{HCN, CN} + \mathrm{O} \longrightarrow \mathrm{NO} + \text{organic residue} \tag{10}$$

This mechanism requires rich combustion conditions and therefore does not play a significant role in a well-controlled engine combustion processes.

Particulate Emissions. Particulates are tiny solid and liquid particles. They occur in the exhaust of diesel engines in sizes from 0.01 to 1 µm and consist of carbon-rich particles (soot), hydrocarbons, and sulfates. In a diesel engine, due to incomplete mixing, zones of air deficiency arise and lead to soot formation. In parallel, in zones with excess air, soot is partially burned. Soot emissions are strongly increased if the air – fuel mixture is enriched to near stochiometry.

Recent research has shown that especially ultrafine particles (< 100 nm in mean diameter) and nanosized particles (< 50 nm in mean diameter) are potentially dangerous due to their capability to enter deep into the respiratory tract. Furthermore, these small particles supposedly can penetrate cell membranes.

Carbon Dioxide. Carbon dioxide (CO_2), although the most commonly emitted combustion gas, is normally not considered as a pollutant since adverse health effects are not known. However, CO_2 should be regarded as a serious pollutant in terms of its global-warming potential. CO_2 has been identified as a greenhouse gas, i.e., a gas-phase species with strong infrared adsorption bands that lets short-wavelength light from the sun pass through but adsorbs the light reflected from the earth in the troposphere. Due to the large CO_2 emission rate and its long lifetime its contribution to the greenhouse effects is estimated to be 50 %. Since CO_2 is an intrinsic reaction product of fuel combustion, the only way to reduce CO_2 emissions is to improve fuel economy.

Sulfur Compounds. Any sulfur contained in the fuel is emitted through the tailpipe in the form of sulfur compounds, mostly as SO_2. Under atmospheric conditions the SO_2 is oxidized to SO_3, which leads to the formation of sulfuric acid. Sulfuric acid originating from the combustion of sulfur-containing fuels is a major contributor to acid rain. Currently the only feasible way to reduce emissions of sulfur compounds from internal combustion engines is to reduce the sulfur content of the fuel.

If the engine is operated under rich conditions SO_2 can be reduced to H_2S in the catalytic converter. Since H_2S has a characteristic unpleasant smell, it is the task of the engine manage-

ment system to avoid driving conditions that lead to the formation of significant concentrations of H_2S.

3. Developments in Engine Technology

Engine technology is a wide field and is not the main focus of this article. For this reason here only a few developments in engine technology that have direct impact on catalytic emission control are outlined. A more complete treatment of the development of engine technology can be found elsewhere [8].

Traditionally diesel engines show some advantage in fuel economy compared to spark-ignition gasoline engines. The fuel economy of diesel engines was recently improved even further by the introduction of high-pressure direct-injection technologies, so that today the diesel engine exhibits an advantage in fuel economy over the conventional gasoline engine of about 25 – 30 %.

Several effects contribute to the fact that conventional gasoline engines show poorer fuel economy than diesel engines, the major reason being that these engines must be run with a stochiometric air – fuel mixture and that at low load the air intake of the engine has to be adjusted to the fuel by placing a throttle in the air intake. Especially at low load the throttle creates a vacuum in the intake manifold that imposes extra work on the cylinders. Other effects that contribute to the limited efficiency of current gasoline engines are temperature loss to the walls and friction. Several possibilities exist to improve the fuel economy of gasoline fuel cars, all of which try to minimize the energy inefficiency caused by the operation of the throttle at low engine load:

- Downsizing
- Electromechanical valve control
- Direct injection

The operating points met in urban driving can be shifted towards the more fuel efficient high-load region by downsizing the engine. To retain the same overall maximum power the specific power of the downsized engine must be increased by supercharging the engine. Best fuel economy at low load is obtained if a turbocharger with variable compression ratio is used.

An alternative technology is electromechanical valve control. Here the valves can be controlled individually and independent of the cylinder movement so that the air intake can be controlled by the timing of valve opening instead of the energy-inefficient throttle. Both downsizing and electromechanical valve control operate the engine with a stochiometric air – fuel mixture and therefore can be used in combination with the well-established three-way catalyst. The improvement in fuel economy expected from these technologies is limited to about 15 % relative to conventional gasoline engines.

Ultimate improvements in fuel economy can probably only be achieved if direct injection or combinations of the above-mentioned technologies with direct injection are employed. In a gasoline direct-injection engine the fuel is directly injected into the cylinder. like in diesel engines. This allows operation of the engine with an oxygen-rich air – fuel mixture under partial load. In this way the disadvantages related to throttling the air intake can be completely avoided. Additional fuel economy is gained because modern gasoline direct-injection engines minimize wall heat losses and cylinder friction. Due to the lean operation at partial load, exhaust from a gasoline direct-injection engine can not be cleaned with a conventional three-way catalyst. New catalyst technologies must be developed for the treatment of the lean-burn direct-injection gasoline exhaust. Catalysts for the treatment of lean exhaust from gasoline lean-burn and diesel engines are described in Chapter 6.

4. Catalytic Exhaust Aftertreatment, General Concepts

Even though advanced engine technology offers substantial potential for the reduction of emissions of hazardous compounds, engine improvements alone do not allow a reduction of emissions to a level necessary to fulfill current and future emission legislation. For this reason aftertreatment of the raw exhaust gasses emitted by the engine is unavoidable. Under typical operation conditions the thermodynamic equilibria for the main reactions transforming exhaust gas

pollutants into harmless compounds are on the product side, so that automotive exhaust purification essentially is a kinetic problem. For this reason catalysts that enhance the reaction rates are the method of choice for exhaust aftertreatment [2–6].

4.1. The Physical Design of the Catalytic Converter

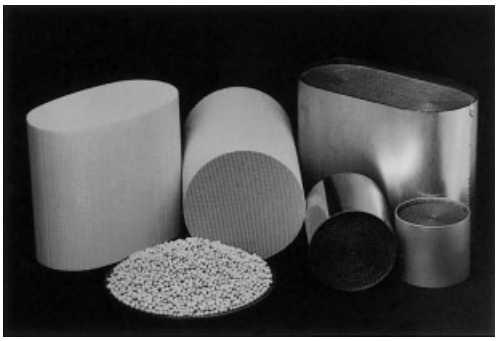

Figure 1. Various designs of emission control catalysts (beads, ceramic monoliths, metal monliths)

Exhaust gas catalysis is a heterogeneous surface process, and to reach acceptable conversion within a reasonable catalyst volume a carrier structure must be used that exposes the exhaust gas flow to a high geometric surface area. In the early days of exhaust gas catalysis pellets were used for this purpose. In recent years pellets have been completely replaced by monolithic honeycomb carrier structures. Monolithic converters consist of a single body that possesses a multitude of parallel channels, with the catalytically active material deposited along the walls of the channels. Figure 1 shows various designs of emission control catalysts. Honeycomb converters exhibit several advantages over pellet converters, the most important being the lower pressure drop and the reduced complexity of the converter housing design. A typical converter design with a ceramic monolith is shown in Figure 2.

Monolithic supports are produced either from ceramic or metal, each of which has specific advantages.

Ceramic monoliths are most commonly produced by an extrusion process from synthetic corderite $2\,MgO{\cdot}2\,Al_2O_3{\cdot}SiO_2$. This material has a number of advantageous properties for this application, such as high melting point ($>1300\,°C$), high mechanical strength, and most importantly a very low thermal expansion coefficient that results in excellent thermal shock resistance. Ceramic supports can be produced in a variety of different forms that differ in their outer dimensions, cell shape, cell density, and wall thickness. Today the most commonly used supports have square channels, a cell density of about 62 channels per square centimeter, and a wall thickness of about 0.15 mm. Thinner walls increase the open frontal area and accelerate heating up during vehicle start. Higher cell densities improve the geometrical contact surface within the channels. For these reasons there are ongoing efforts to develop supports with thinner walls and higher cell densities.

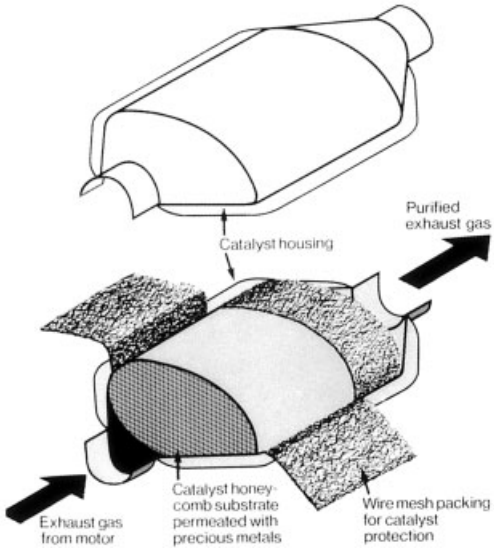

Figure 2. Design principle of a ceramic monolithic based converter for aftertreatment of exhaust gases

In the few last years monoliths made of metal were introduced to the market. Their main advantage is that they can be prepared with thinner walls. This allows higher cell densities with increased frontal area and thus reduced pressure drop. The low heat capacity of the thin metal walls and the higher thermal conductivity result in faster heat up compared to ceramic catalysts. However, metals are prone to corrosion under exhaust conditions, and their thermal expansion behavior is less compatible with the catalytic material.

4.2. The Washcoat

Ceramic and metallic monolith structures have a geometric surface area in the range of $2.0-4.0\,m^2$ per liter of support volume. This is much too low to adequately perform the catalytic conversion of the exhaust gas components. For this reason the monolith channel walls are coated with a thin layer of inorganic oxides with a very high internal surface area. This layer of carrier oxides is called the washcoat. The washcoat enhances the surface of the catalyst by a factor of about 10 000. Figure 3 shows individual channels of a ceramic monolith coated with washcoat. The main components of the washcoat are inorganic oxides and mixed oxides, and it also contains the catalytically active metals, mainly the platinum group metals (PGMs) Pt, Pd, and Rh, for which the oxides act as a support.

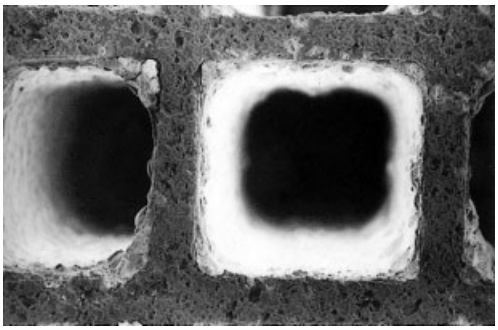

Figure 3. Close-up view of a washcoated ceramic monolith

The inorganic carrier oxides serve several functions vital for operation of the catalytic converter. First, they must maintain a high surface area even under severe aging conditions at temperatures above 1000 °C. Second, the carrier oxides have to prevent the catalytically active precious metal particles from sintering to larger particles at high temperatures. This is accomplished by tailoring the oxide composition such that the precious metal particles are tightly bonded on the carrier oxide surface. Figure 4 shows a high-resolution transmission electron micrograph of washcoat particles with finely dispersed PGM particles. Third, the carrier oxide improves the catalytic function of the precious metal particles by favorable electronic interaction between the precious metal clusters and the carrier oxide surface. To fulfill all these requirements modern washcoat systems are very complex chemical systems. The washcoat matrix consists of specifically designed mixed crystal lattices, and various chemically different phases are selectively combined to optimize performance and selectivity. The manufacturing process controls the optimum placement of the precious metals with respect to primary particle size and location on the different carrier phases.

Figure 4. Transmission electron micrograph of PGMs on a washcoat particle

Alumina is the most common carrier oxide used in automotive exhaust gas catalysis. Typically γ-, δ-, and θ-Al_2O_3 are used with internal surface areas of $50-250\,m^2\,g^{-1}$. Frequently the Al_2O_3 is doped with rare earth or alkaline earth elements. The role of the dopant is to increase the aging stability of the carrier and to optimize the carrier – precious metal interaction to prevent precious metal sintering and to enhance the catalytic activity of the precious metal particles. Besides alumina other common washcoat constituents are cerium and zirconium oxides. Cerium oxide mostly serves an oxygen storage component (see Section 5.2). Recently, zirconium cerium mixed oxides with optional additional dopants have been used.

The washcoat is usually prepared as an aqueous slurry and applied to the monolithic carrier by a dipping or spraying process followed by blowing out of excess washcoat slurry, drying, and calcination. Tight process and materials control is essential due to the high cost of the ingredients, especially the PGMs.

5. Catalytic Aftertreatment of Stoichiometric Exhaust Gas

Up to now the majority of passenger cars and light-duty vehicles, especially in the USA, are propelled by spark-ignition engines operated under stoichiometric conditions. Due to their large population and, compared to diesel engines, their high raw emissions these vehicles accounted for a large part of the air pollution in urban areas. Therefore, spark-ignition engines were the first to require catalytic exhaust aftertreatment to fulfill emissions regulations. Most of the commercial experience in automotive catalysis till now has been gained in the field of stochiometric spark-ignition engines.

Different concepts of catalytic pollution control for stoichiometric spark ignition engines have been used over time (Fig. 5). During an initial period following the introduction of precious metal exhaust gas catalysts in 1975, NO limits could be reached by engine measures such as exhaust gas recirculation, and legislation only required the treatment of CO and HC. At this time engines were run slightly rich to reduce NO emissions. Secondary air was injected in front of the catalyst (Fig. 5 A), and the catalyst was subjected to oxygen-rich exhaust gas so that CO and hydrocarbons could be oxidized by the following reactions:

$$2\,CO + O_2 \longrightarrow 2\,CO_2 \tag{11}$$

$$C_mH_n + (m + 0.25n)\,O_2 \longrightarrow m\,CO_2 + 0.5n\,H_2O \tag{12}$$

The next step in legislation in 1977 called for a reduction in nitrogen oxides. For this purpose a reduction catalyst was introduced that operated directly on the CO- and hydrocarbon-rich exhaust gas and removed nitrogen oxide by the reactions:

$$CO + NO \longrightarrow 0.5\,N_2 + CO_2 \tag{13}$$

$$C_mH_n + (2m + 0.5n)\,NO \longrightarrow m\,CO_2 + 0.5n\,H_2O + (m + 0.25n)\,N_2 \tag{14}$$

Secondary air was injected downstream of the reduction catalyst (Fig. 5 B), and CO and hydrocarbon emissions were reduced by an additional oxidation catalyst.

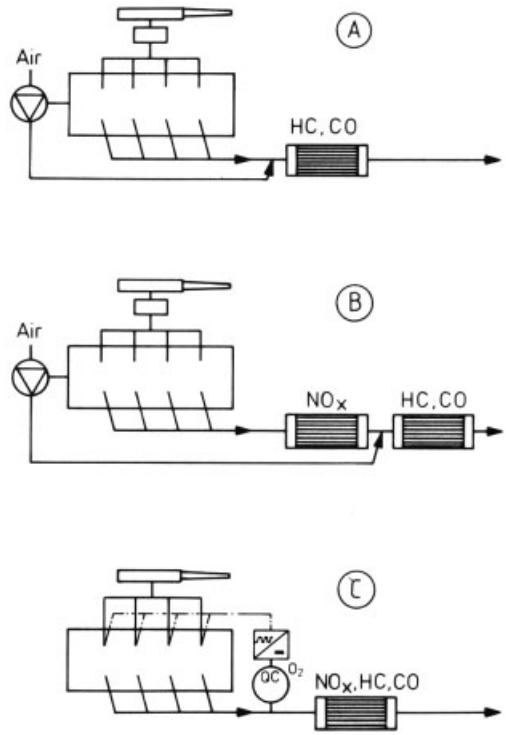

Figure 5. Processes for catalytic automotive-exhaust purification
A) Single-bed oxidation catalyst; B) Dual-bed catalysis; C) Three-way catalyst with closed-loop control

A further improvement in exhaust gas control technology for stoichiometric spark-ignition engines was achieved by the development of three-way catalysts that simultaneously convert all three regulated pollutants.

5.1. Three-Way Catalysis

In Figure 6 the conversion efficiencies of a typical closed-loop three-way catalyst (Fig. 5 C) for the major pollutants are plotted as a function of the normalized air to fuel (A/F) ratio lamba. Under oxygen-rich conditions ($\lambda > 1$) CO and hydrocarbon conversions are high, while NO conversion is prevented by insufficient supply of reductants. Under fuel-rich conditions ($\lambda < 1$) the catalyst shows high NO conversion but low conversion of CO and hydrocarbons. The conversion ranges for NO and CO/hydrocarbons overlap at $\lambda = 1$. In a small λ range around stoichiometry the catalyst exhibits acceptable conversion for all three pollutants. Efficient operation of the

three-way catalyst requires the engine to be run exactly within this λ window at all times and was therefore only made possible by the introduction of λ sensors and active-feedback λ control.

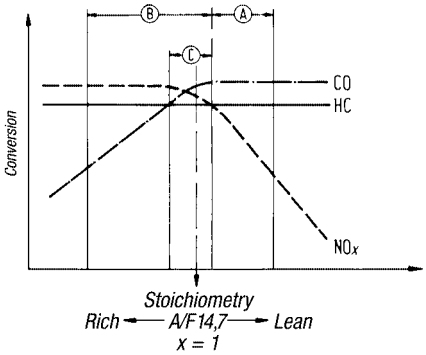

Figure 6. Range of λ values in which various catalytic aftertreatment concepts are operated
A) Single-bed oxidation catalyst; B) Dual-bed catalysis; C) Three-way catalyst with closed-loop control

The λ sensors applied for feedback control of A/F ratio are electrochemical solid-state sensors that measure the concentration of free oxygen after complete conversion of the exhaust gas mixture on a catalytically active layer coated on the sensor element. The active material of the sensor itself is the anion conductor ZrO_2, and the potential between a reference gas and the exhaust gas across the ZrO_2 is determined by the equilibrium concentration of free O_2 after complete conversion on the catalytic surface of the λ sensor [4]. The λ sensor gives a signal of about 50 mV at $\lambda > 1$, and about 800 mV for $\lambda < 1$ with a sharp jump at $\lambda = 1$. In a basic design of a three-way catalytic system, a λ sensor is situated in the exhaust gas before the catalytic converter and returns a feedback signal which is used to control the electronic fuel injection such that derivations from $\lambda = 1$ are compensated.

5.2. Oxygen Storage in Three-Way Catalysts

Due to the finite time response of the feedback control the A/F ratio in the feedback-controlled mode performs regular λ variations around $\lambda = 1$. These perturbations occur with a frequency around 1 Hz and an amplitude of about 1 – 10 % of the A/F set point. Comparison with the conversion plot in Figure 6 reveals that A/F variations of 1 – 10 % are large compared to the width of the three-way conversion window, and that during such perturbations the A/F value is outside the optimal range for most of the time.

For this reason three-way catalysts contain oxygen components that buffer the A/F variations. The compound most commonly used for this purpose is CeO_2. The redox transition between Ce^{IV} and Ce^{III} acts as an oxygen buffer [9]. During lean excursions Ce_2O_3 is oxidized by excess O_2, and this results in the desired correction of the gas phase A/F ratio towards stoichiometry.

$$Ce_2O_3 + 0.5\,O_2 \longrightarrow 2\,CeO_2 \tag{15}$$

In a susequent rich excursion the stored oxygen in turn is used to oxidize the excess CO and hydrocarbons, again shifting the gas phase composition towards stoichiometry.

$$2\,CeO_2 + CO \longrightarrow Ce_2O_3 + CO_2 \tag{16}$$

Later it was discovered that ZrO_2/CeO_2 mixed oxides show superior O^{2-} ion mobility than pure CeO_2 and hence have improved oxygen-storage behavior, and the oxygen-buffer performance can be further improved by doping with additional rare earth elements. For this reason in modern three-way catalysts pure CeO_2 has been widely replaced as an oxygen storage material by specifically designed mixed crystal lattices.

A further benefit of including CeO_2 or CeO_2/ZrO_2 mixed oxides in the washcoat is their ability to catalyze the water gas shift and steam reforming reactions:

$$CO + H_2O \longrightarrow CO_2 + H_2 \tag{17}$$

$$C_mH_n + (2m)\,H_2O \longrightarrow m\,CO_2 + (2m + 0.5n)\,H_2 \tag{18}$$

These reactions contribute to the removal of CO and hydrocarbons, and the H_2 acts as a very active reductant for NO:

$$NO + H_2 \longrightarrow 0.5\,N_2 + H_2O \tag{19}$$

Due to the presence of oxygen storage components λ fluctuations at the exit of the three-way catalyst are by far less pronounced than upstream of the catalyst. Consequently, a λ sensor downstream of the catalyst is perturbed to a

much lesser extent by λ fluctuations and allows more precise determination of the long-term average A/F ratio. For this reason modern engine management systems use a dual λ sensor concept. A front sensor located upstream of the catalyst is used for short-term λ feedback control, and a second, rear sensor located downstream of the catalyst in a secondary control loop controls the long-term average A/F ratio.

Current regulations require the performance of the exhaust gas converter system to be continuously monitored during vehicle operation. Ideally this on board diagnostics (OBD) requirement would be fulfilled by directly monitoring the output of one or several of the regulated pollutants. However, currently detectors for the major exhaust gas pollutants are expensive and voluminous devices, and there are no sensors available that from the point of view of cost, space requirements, and sensitivity are suitable for use in serial-production three-way catalytic converter systems. For this reason legislation currently accepts the monitoring of oxygen-storage capacity as an indicator of catalyst performance. The idea behind this is that oxygen-storage capacity is a key factor determining the performance of the three-way catalyst system and that major catalyst aging mechanisms impact oxygen-storage capacity in approximately the same way as the catalytic performance of the catalyst. Generally, on board diagnostics for the oxygen-storage capacity use a dual λ sensor configuration. A common scheme makes use of the natural λ fluctuations imposed by the A/F feedback control and determines the oxygen-storage performance from the comparison of up- and downstream modulation amplitude, phase shift, or related quantities.

5.3. Precious Metals in Three Way Catalysis

Much effort has been spent on replacing platinum group metals by less expensive and more abundant materials. Numerous publications describe the results obtained with catalysts that contain, for example, the oxides of Cu, Cr, Fe, Co, and Ni. However, until today no breakthrough has been achieved, and automotive exhaust treatment still relies on precious metals, mostly Pt, Pd, and Rh, as catalytically active components. Each of these three major precious metals has specific advantages and disadvantages.

Platinum and palladium are generally used for oxidation catalysts and to promote the oxidation reactions (11) and (12) in three-way catalysts. The characteristic advantage of Pt over Pd is its higher resistance to poisoning by sulfur and lead. This promoted the preferred use of Pt in the early days of exhaust gas catalysis.

The specific advantage of Pd is its higher resistance to high-temperature aging, especially under oxygen-rich conditions, and its better performance at lean cold start. Both make Pd the precious metal of choice for close-coupled start catalysts. One drawback associated with the use of Pd in three-way catalysts is that its three-way conversion window is generally narrower. This difficulty could be overcome by improved λ feedback control and specifically designed washcoat compositions.

Rhodium is employed in three-way and reduction catalysts due to its superior performance in promoting the NO reduction reactions (13) and (14) and its high selectivity for the formation of N_2 in these reactions, whereas Pt and Pd tend to form NH_3 and N_2O.

A combination of technical and price considerations led to the preferred use of different precious metals for different generations of three-way catalysts. In the early days of three-way catalysis there was no alternative to the application of Pt or Pt/Rh catalysts due to the contamination of the fuel with sulfur and lead. Starting from about 1980 preferences shifted towards Pd/Rh or even Pd-only systems. This can be explained by the increased fuel quality and by the relatively low price of Pd at that time. More recently, due to the high demand from the automotive catalyst industry, the cost advantage of Pd-based systems has disappeared, and interest has focused again on Pt-based systems. In general, the supply situation is considered more critical for Pd than for Pt, since Pd is a byproduct of Ni production, so that Pd production does not flexibly adjust to changes in Pd demand, while Pt production capacity, at least in the long term, is expected to increase according to the growing demand.

5.4. Recent Developments in Three-Way Catalysis

Today three-way catalysts, after having reached their operating temperature, show very high conversion of the major polutants CO, HC, and NO. More than 80 % of the overall emissions, especially of the hydrocarbon emissions, occur during the warm-up phase directly following the cold start of the vehicle. Consequently, the key to further reductions of overall emissions is to improve the cold-start performance of the catalytic converter.

One approach to accelerate the warm-up of the catalytic converter is to employ a start catalyst that is mounted very close to the outlet of the engine and is therefore heated very rapidly after starting the engine. This close-coupled catalyst can either be used in addition to or can even completely replace the traditional underfloor catalyst. Due to the high temperatures encountered in the close-coupled position, very temperature resistant washcoat formulations had to be developed for this application.

An alternative approach is to actively heat the catalyst, either electrically or by a burner system mounted in front of the catalyst. A general drawback of actively heated catalytic converters is the high cost associated with these systems.

A third approach to reduce hydrocarbon emissions during cold start is the application of hydrocarbon adsorption traps. Here during cold-start hydrocarbons are adsorbed by an adsorber material, generally a zeolite. When the catalytic converter has reached a higher temperature, the hydrocarbons are desorbed and converted on the catalyst. Hydrocarbon adsorber systems are frequently employed in combination with actively heated catalysts.

6. Catalytic Aftertreatment of Lean Exhaust Gas

Diesel engines and lean-burn gasoline engines become more and more important due to current efforts to increase fuel economy and reduce emissions of the greenhouse gas CO_2. The oxidation of the pollutants CO and hydrocarbons is in principal facilitated by the excess of oxygen in lean exhaust gas. However, NO reduction on Rh is inhibited by high oxygen concentrations. This makes NO_x removal the technical challenge in the treatment of lean exhaust gas. Several strategies have been developed to remove NO_x from lean exhaust gas. In principle these strategies can be applied for diesel and gasoline lean-burn engines. Consequently, first the different techniques for NO_x removal under lean conditions are discussed, followed by sections treating the specific problems and solutions associated with gasoline lean-burn and diesel engine operation, including the additional measures necessary to control soot emissions from diesel engines.

6.1. NO_x Removal from Lean Exhaust Gas

$DeNO_x$ Catalysts. Certain catalyst materials are known to selectively promote the reduction of NO by hydrocarbons even in the presence of excess oxygen. Two classes of $DeNO_x$ catalysts have been intensively studied: zeolites exchanged with transition metals (typically Cu/ZSM-5) and platinum group metals.

One major difficulty encountered with $DeNO_x$ catalysts is their relatively narrow temperature window of operation: At low temperatures conversion is too slow, while at higher temperature the selectivity is lost, and hydrocarbons are consumed by reaction with excess oxygen.

In the following list of starting compounds for direct dyes, recent results compiled from the patent literature have been taken into account in particular in the compounds listed under (b) and (e).

Further technical problems are caused by the limited temperature stability of current catalysts. This requires the catalysts to be positioned in an underfloor position and limits the light-off performance of the converter system.

Efficient conversion of NO_x requires the presence of HC in high enough concentrations. One option is to increase the hydrocarbon concentration in the raw exhaust by appropriate adjustment of the injection timing. This configuration exhibits severe problems in meeting HC limits, especially during start up of the vehicle, because it is not possible to use a start-up oxidation catalyst close to the manifold. Alternatively, injection of secondary fuel into the exhaust line upstream of the $DeNO_x$ catalyst can be used.

Due to the technical problems associated with the application of $DeNO_x$ catalysts the system is currently not widely used in serial applications. One exception are the first-generation gasoline direct injection vehicles that are sold in Japan, which use specially developed $DeNO_x$ catalysts based on iridium as the catalytically active material.

Selective Catalytic Reduction with Ammonia. Selective reduction of NO_x by NH_3 is a well-established process for the exhaust treatment of stationary sources. In this process NH_3 or an appropriate precursor is added to the exhaust gas, and a catalyst promotes the selective reduction of NO by NH_3 in the presence of excess O_2 by the following reaction [10]:

$$NO + NH_3 + 0.25\,O_2 \longrightarrow N_2 + 1.5\,H_2O \tag{20}$$

For automotive applications NH_3 is generated on board by catalytic or thermal decomposition of urea, a nontoxic and relatively easy to handle chemical. The amount of NH_3 added to the exhaust stream must be matched to the engine NO emission at each operating point. Vehicle tests have shown that this technology possesses promising potential, at least for applications with heavy-duty truck diesel engines.

Initially, the introduction of NH_3 SCR has met some resistance, mainly because of the need to handle and store a chemical on board, and because of the expenses associated with building a new infrastructure to distribute it. However, it is now becoming apparent that the advantages inherent to the system, mainly the possibility to reduce NO_x and simultaneously optimize fuel economy and particulate emissions, may well outweigh the disadvantages. It is expected that NH_3 SCR will be introduced into heavy duty truck serial production starting from about 2005. A typical system for onboard NH_3 selective catalytic reduction is shown in Figure 7.

The NO_x Adsorber Catalyst. The operating principle of NO_x absorbers is based on the fact that highly basic alkaline earth and alkali metal oxides form stable nitrates at temperatures below about 600 °C. In an NO_x adsorber/converter system NO is first oxidized to NO_2, which then reacts with the basic oxides in the adsorber to form a nitrate. Oxides frequently employed for this purpose are BaO and K_2O. For BaO the reaction reads:

$$BaO + 2\,NO_2 + \tfrac{1}{2}\,O_2 \longrightarrow Ba(NO_3)_2 \tag{21}$$

Of course, the storage capacity of the oxide is soon exhausted and the storage catalyst must be regenerated. For this purpose the engine is switched to rich operation for a few seconds. Finely divided precious metal on the adsorber catalyst promotes the reduction of nitrate by CO, hydrocarbon, or H_2 to form N_2:

$$Ba(NO_3) + 5\,CO \longrightarrow BaO + 5\,CO_2 + N_2 \tag{22}$$

$$Ba(NO_3)_2 + 5\,H_2 \longrightarrow BaO + 5\,H_2O + N_2 \tag{23}$$

In this way the oxide is reformed and can serve as NO_x adsorber during the following cycle of lean operation (Fig. 8). The regeneration cycle is triggered by the engine management computer system that monitors the NO_x level of the adsorber catalyst either by model-based computation of the total NO_x fed into the adsorber during each cycle or by applying a sensor downstream of the adsorber. For this reason the adsorber catalyst becomes an integral part of the engine- and fuel-management system.

Currently the major technical problem associated with the application of NO_x storage catalysts is the adsorption of sulfur oxides by the same mechanism as NO_2. Since the resulting sulfates, owing to their higher stability, are not destroyed in an average rich regeneration cycle, sulfur accumulates on the catalyst and lowers the NO_x adsorption capacity. For this reason the catalyst must be regenerated from sulfur from time to time, e.g., by running the engine rich at elevated temperatures for several minutes. The sulfur regeneration cycle is fuel-expensive, and since the execution frequency of the cycle is determined by the fuel sulfur content a high sulfur concentration directly translates into a corresponding fuel penalty during operation of the NO_x adsorber catalyst.

6.2. Catalytic Aftertreatment of Lean-Burn Gasoline Exhaust

Catalytic converters for direct-injection lean-burn engines are derived from three-way catalyst

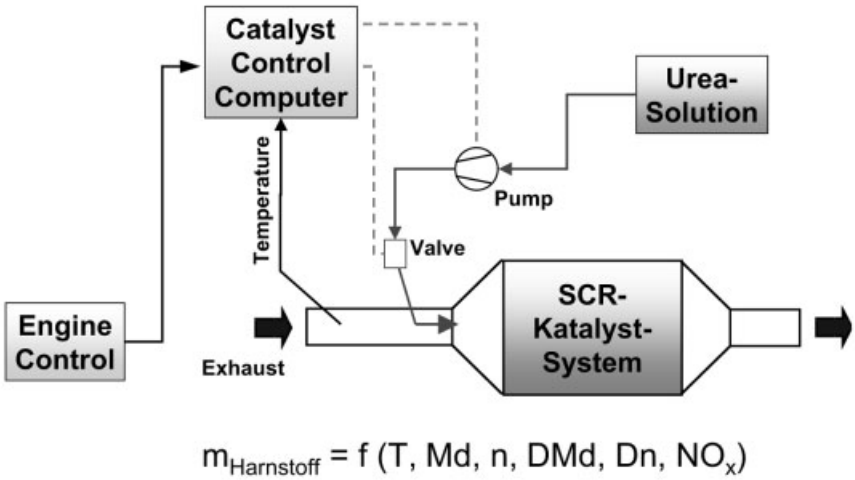

Figure 7. A typical system for NO_x control by area onboard SCR

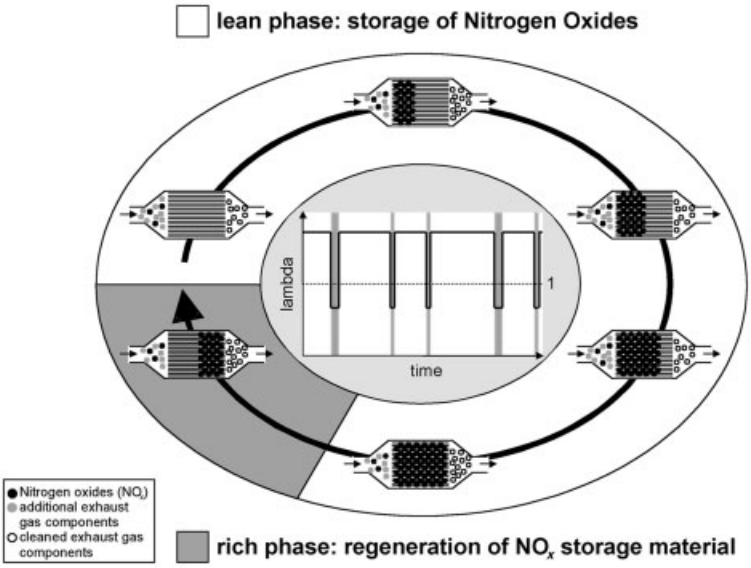

Figure 8. NO_x storage cycle

systems [11]. However, under the operating conditions when the engine is operated at lean burn the three-way converter does not convert NO. For this reason special measures must be taken to allow NO conversion under oxygen-rich conditions. The different strategies of NO treatment in lean exhaust gas are described in Section 6.1.

The first direct-injection lean-burn vehicles offered on the Japanese marked were equipped with special $DeNO_x$ catalysts (see Section 6.1) based on iridium as the catalytically active material. However, such systems do not exhibit high enough temperature stability to withstand the severe aging conditions during U.S. and especially European driving.

Currently, the preferred technology for NO_x treatment of gasoline direct-injection lean-burn engines is the NO_x absorber. A typical converter system for gasoline lean burn exhaust gas treatment is shown in Figure 9. The exhaust gas first

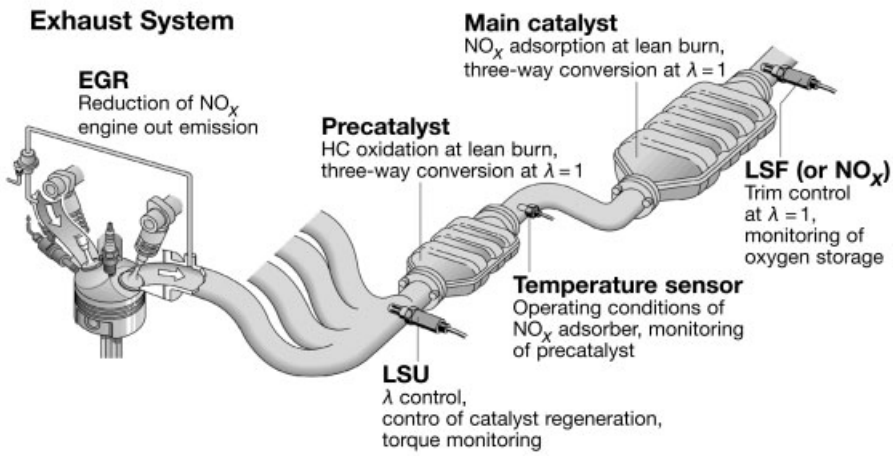

Figure 9. Aftertreatment system for a gasoline direct-injection engine

passes through a start catalyst that removes CO and hydrocarbons during light off and oxidizes NO to NO_2. The start catalyst is followed by the NO_x adsorber catalyst and a conventional three-way catalyst.

6.3. Catalytic Aftertreatment of Diesel Exhaust

Compared to gasoline engines, diesel catalytic exhaust treatment is complicated by two additional problems: the relatively high sulfur concentration in diesel fuel and the emission of soot and other particulate matter by diesel engines. For these reasons, and because of the intrinsically lower gaseous pollutant concentrations compared to spark ignition engines, catalytic aftertreatment for diesel engines was implemented much later than for spark-ignition vehicles. The only diesel exhaust technology widely used in serial applications today is the diesel oxidation catalyst. This technology has been generally applied to passenger cars in Europe and to some medium- and heavy-duty trucks in the USA since about 1991.

The function of a diesel oxidation catalyst is the reation of CO and hydrocarbons with the excess oxygen in the diesel exhaust (reactions 11 and 12). Additionally the mass of particulate matter is reduced by the oxidation of liquid and polynuclear aromatic hydrocarbons adsorbed on the particulates. Though diesel oxidation catalysts use washcoats that are similar to those in three-way catalysts, technical requirements are very different for diesel oxidation catalysts, and washcoat compositions optimized to meet these requirements have been developed.

The key is to design a catalyst that selectively catalyzes the oxidation of CO and hydrocarbons and simultaneously suppresses the oxidation of NO and SO_2. Especially SO_3 formation must be avoided because SO_3 deactivates alumina-based washcoats by the formation of sulfates and because SO_3 condenses as sulfuric acid on the particulates and thus increases their mass.

One additional problem faced in diesel exhaust catalysis is clogging of the monolith channels by accumulation of particulate matter. This problem is generally solved by applying monoliths with higher cell diameter.

Thermal aging is a less severe problem with diesel catalysts because of the intrinsically lower exhaust temperature of diesel engines. However, due to the low temperature of diesel exhaust, diesel oxidation catalysts must be optimized to become catalytically active at very low temperatures, frequently below 150 °C.

Until today none of the NO_x removal technologies described in Section 6.1 has found wide use in diesel vehicles. NO_x adsorber catalysts and $DeNO_x$ catalysts are very sensitive to fuel

sulfur and for this reason can not be used at current diesel sulfur levels. However, current diesel oxidation catalysts, depending on the design of the washcoat, show some degree of NO_x conversion. For this reason there is not a clear distinction between diesel oxidation catalysts and $DeNO_x$ catalysts.

Ammonia SCR is expected to be introduced for heavy duty trucks by about 2005. The advantage of this technology is that it allows the engine to be run under conditions where fuel economy is optimized and particulate emissions are reduced to a minimum.

6.4. Removal of Particulate Emissions

Besides NO, CO, and hydrocarbons diesel engines emit particulate matter as a fourth major regulated pollutant. In contrast to the catalytic removal of the gaseous pollutants removal of particulate matter is a physical separation process. The exhaust gas is forced to flow through a filter material with pores smaller than or similar in size to the particles, which are trapped within the pores or on the surface of the filter. The currently most common design for particulate filters is the wall-flow monolith design (Fig. 10). This design is derived from the flow-through monoliths used for catalytic exhaust treatment by plugging the channels alternatively at the inlet and at the outlet of the monolith. In this way the exhaust gas is forced to flow through the porous walls of the monolith and the particulate matter is deposited along the walls.

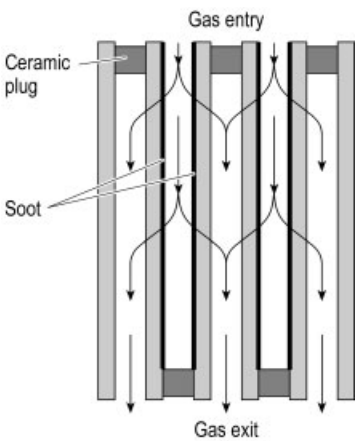

Figure 10. Wall-flow monolith particulate filter

Since accumulated particulate matter blocks the pores of the trap and causes excessive pressure drop, the filter must be regenerated by burning the trapped particulate matter with the oxygen present in diesel exhaust. Ignition of the trapped particulate matter requires temperatures above about 550 °C. Such temperatures are rarely reached in the exhaust of diesel vehicles. One solution to this problem is to apply filter traps that catalytically lower the ignition temperature of the soot. Regeneration of the particulate trap at temperatures below about 300 °C is achieved in this way. Alternatively fuel additives have been used that catalytically lower the ignition temperature of the deposited particulate matter. The extra effort involved in handling, storing, and dosing the extra fuel additive onboard the vehicle are obvious. As a further alternative so-called continuously regenerating trap (CRT) systems have been designed. Here the naturally occurring NO is oxidized by a strong oxidation catalyst to NO_2, which is used to continuously oxidize the deposited particulate matter on the filter trap.

7. Conclusion and Outlook

Catalytic exhaust gas treatment is a well-established technology that has allowed a significant reduction of air pollution. Since the first introduction of exhaust catalysts in the 1970s, aftertreatment technology has continuously improved, so that today automotive exhaust catalysis has become one of the most innovative fields of cataytic science and one of the most important applications of heterogeneous catalysis. Today, pollution from vehicles has been reduced by more than 90 % compared to the level before the introduction of catalytic exhaust control. The near future will see even further reductions.

A major remaining challenge for the automotive industry is the lowering of fuel consumption and CO_2 emissions. This is also a challenge for catalyst development, since the most fuel-efficient engines operate with excess air, and the conventional three-way catalyst does not allow NO_x reduction under these conditions. New technologies are being developed that allow fuel-efficient lean operation of the engine and at the same time retain the high emission standards of conventional three-way catalysts.

Early automotive catalysts were more or less add-ons to vehicles. Future emission control technologies will be significantly more complex. One driving force for this increased complexity is the need to apply storage-type catalysts. While conventional catalysts can be operated continuously, storage-type catalysts require cyclic operation with alternating storage and regeneration phases. This unsteady operation requires the catalytic system to be designed as an integral part of the overall fuel- and engine-management system. Consequently, automative emission control has evolved into an interdisciplinary field in which technological and commercial success can only be achieved in an inegrated development process with close cooperation of engine-design engineers, fuel-management experts, and catalysis specialists.

8. References

Specific References

1. http://www.umweltbundesamt.de
2. E. S. Lox, B. Engler: "Environmental Catalysis", G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* VCH, Weinheim, 1997
3. K. C. Taylor in A. Crucq, A. Frennet (eds.): *Catalysis and Automotive pollution Control,* VNR, New York; 1987
4. C. Hagelüken (ed.): *Autoabgaskatalysatoren* Expert Verlag, Renningen, 2001.
5. R. M. Heck, R. J. Farrauto: *Catalytic Air Pollution Control,* VNR, New York 1995.
6. N. A. Henein, D. J. Patterson: *Emissions from Combustion Engines and Their Control,* Ann Arbor Science, Ann Arbor, Michigan, 1972.
7. J. A. Miller, C. T. Bowman: "Mechanism and modeling of Nitrogen Chemistry in Combustion", *Prog. Energy Combust. Sci.* **15** (1989) 287 – 338.
8. Verbrennungsmotoren Band I und II, Vorlesungsdruck, Prof. Dr.-Ing. S. Pischinger, Lehrstuhl für Verbrennungskraftmaschinen, RWTH Aachen, 2002.
9. J. Kaspar, M. Graziani, P. Fornaserio: "Ceria-Containing Three-Way Catalysts", in K. A. Gscheidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths* **Vol. 29,** Elsevier Science, 2000.
10. J. Gieshoff, A. Schäfer-Sindlinger, P. C. Spurk, J. A. A. van den Tillaart, G. Carr: *Improved SCR for Heavy Duty Applications* SAE 2000-01-0189, Detroit, 2000.
11. W. Strehlau, J. Leyrer, E. S. Lox, T. Kreuzer, M. Mori, M. Hoffmann: *New Developments in Lean NOx Catalysis for Gasoline Fueled Passenger Cars in Europe,* SAE 962047, Detroit, 1996.

Aviation Turbine Fuels

GEOFF J. BISHOP, Shell Aviation Ltd., London, United Kingdom

1.	**History**	1
2.	**Fuel Types and Specifications**	1
2.1.	**Specification Requirements**	2
2.2.	**Fuel Properties**	2
2.2.1.	Bulk Properties	2
2.2.2.	Trace Properties	6
2.3.	**Nonspecification Properties**	8
3.	**Production**	9
3.1.	**Fuel**	9
3.2.	**Additives**	10
4.	**Handling, Storage, and Transportation**	11
5.	**Legal Aspects**	12
6.	**Environmental Aspects**	12
7.	**Economic Aspects**	12
8.	**Future Trends**	13
8.1.	**Petroleum-Derived Fuels**	13
8.2.	**Alternative Fuels**	13
9.	**References**	14

1. History

Turbine-powered aircraft were used by both *British* and *German* air forces for military operations during the latter part of World War II. The British engines were operated on illuminating kerosine containing 1 % light lubricating oil (jet engine oil) [1]. The German engines burned a wider boiling mixture of gasoline and heavier distillate fractions. Fuel development in Germany stopped at war's end. Continuing British research has tightened some requirements over the years, yet today's primary British specifications still describe kerosine-type fuels for both commercial and military aircraft, the main difference being in the additive content.

The initial United States military fuel was a low freezing point kerosine called JP-1, but its limited availability forced the development of a wider boiling fuel that included gasoline as well as kerosine fractions [2]. This "wide-cut" fuel called JP-4 was, until recently, the major military grade. However, safety concerns associated with its flammability led to its phase-out in favor of the kerosine based grade JP-8. The switch from JP-4 to JP-8 was largely completed by the end of 1996.

Commercial aircraft have been burning mostly kerosine-type fuels since 1953 when the early turboprop aircraft came into service, and this has remained the major civil fuel in the turbojets and later the turbofans which power most of the world's civil aircraft fleet. The principal grades are Jet A in the United States, TS-1 in Russia and Jet A-1 in the rest of the world. The wide-cut fuel Jet B, analogous to the military grade JP-4, is now rarely used in civil aviation.

2. Fuel Types and Specifications

The most widely used civil aviation turbine fuel grades are defined by a small number of key specifications as listed in Table 1. These specifications are similar in most respects although a small number of significant differences remain.

Table 1. Major civil jet fuel grades and their specifications

Grade	Area of usage	Major specification
Jet A-1	world-wide	ASTM D1655, DEF STAN 91-91 (DERD 2494)
Jet A	USA only	ASTM D 1655
TS-1	Russia and former Eastern Bloc countries	GOST 10227
Jet Fuel no.3	China	GB 6537
Jet A-1	Canada	CAN/CGSB-3.23

The US Jet A fuel, defined by specification ASTM D1655 [3], differs from Jet A-1 only in freezing point. The British DEF STAN 91-91 specification (formerly well known as DERD 2494 whilst under the jurisdiction of the UK Ministry of Defence) [4] is recognized as a civil jet fuel specification and is used as such in many countries around the world. The Russian GOST 10227 specification also defines a grade designated RT which is very similar to western Jet A-1, as is the Chinese no.3 Jet Fuel.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_201

In addition to these prime specifications, the International Air Transport Association (IATA) maintains fuel "Guidance Material" [5] for Jet A-1 and Jet B grades for the use of its member airlines. Because of the worldwide nature of commercial aviation, the close correspondence between specifications is a necessity to discourage the proliferation of many different national or local specifications. The major jet fuel suppliers subscribe to a document entitled "Aviation Fuel Requirements for Jointly Operated Systems" (AFQRJOS) which defines a "Check List" which comprises the most restrictive requirements of the ASTM D 1655, DEF STAN 91-91 fuel specifications and IATA Guidance Material. Thus fuel meeting the Check List automatically meets all the requirements of the parent specifications. This document was developed to cater for the situation where many different fuel suppliers deliver fuel into the common storage facilities which predominate at most international airports.

Further efforts are underway to minimize the differences between the major fuel specifications with the eventual aim of developing a single, international specification to define aviation turbine fuel. An International Civil Jet Fuel Specification Liaison Group was formed in 1989 to promote this activity and a growing number of national specification bodies are contributing to its progress.

2.1. Specification Requirements

Details of the major fuel specifications for Jet A-1 type fuel are shown in Table 2. Although test methods in most western countries are either identical or very similar, no international standardization exists and direct comparisons between limits may be misleading (refer to the notes applying to Table 2). Test method designations published by the principal standards organizations are given in Table 3.

2.2. Fuel Properties

Unlike most chemicals but like other petroleum fuels, aviation turbine fuels are not defined by composition, molecular structure, and purity. Instead their characteristics are limited by boundary conditions imposed by inspection tests. Some of these tests are fundamental, others empirical; many reflect a solution to past engine or aircraft problems. However, any fuel meeting all requirements is acceptable for aircraft operation. The specifications are dynamic, but changes are conservative because of the potential risk to aviation operations. Long equipment life ($\geqq 20$ years) and certification restrictions (see Chap. 5) also mitigate against drastic specification changes.

The individual specification properties or requirements can be divided rather arbitrarily into bulk and trace properties. *Bulk properties* are those requiring significant composition changes, involving more than 5 vol % of the total fuel, to effect a change in the property. *Trace properties* tend to respond to changes < 1 vol % or even 1 ppm or less. Changes in bulk properties affect product yield, whereas changes in trace properties do not. The grouping has no direct bearing on fuel performance. Instead it helps assess effects of crude composition and manufacturing steps on product yield [8].

2.2.1. Bulk Properties

Density. The mass – volume relationship is important because the quantity of fuel loaded onboard an aircraft is metered volumetrically, but fuel consumption is measured by mass. This relationship is also used extensively in commercial sales of fuel. Although in the past there have been several different ways of defining this relationship, such as specific gravity, relative density, or API gravity, all the major jet fuel specifications now define density at 15 °C in units of kg/m^3. Changes of density with temperature are given in [9]. Density is commonly measured using hydrometers or digital density meters. In-line density measurement is also possible and offers the opportunity for fuel delivery by mass instead of volume.

Specific Energy. Specific energy or heat of combustion is directly related to density. The fuel's specific energy, its density, the aircraft's performance characteristics, and the predicted flight conditions are employed to calculate the amount of fuel required for a given flight or conversely to estimate the aircraft's range with maximum fuel aboard. Specific energy is determined

Table 2. Major fuel specifications for civil Jet A-1 type fuel

	United States	United Kingdom	International	Russia [a]	China [b]
Specification	ASTM D1655	DEF STAN 91-91	Checklist	GOST 10227-86	GB 6537-94
Issue	96b	2	16	Amendment 4	effective June 95
Date	Jul. 96	May 96	Jun. 96	Apr. 96	Aug. 94
Grade	Jet A-1	Jet A-1	Jet A-1	TS-1	No.3 Jet Fuel
Composition					
Acidity, total, mg KOH/g (max)	0.10	0.015	0.015	0.009 [c]	0.015
Aromatics, vol % (max)	25	22.0 [d]	22.0 [d]	22 mass %	20.0
Olefins, vol % (max)	NR [e]	NR [e]	NR [e]	iodine no.2.5 max	5.0
Sulfur, total, mass % (max)	0.30	0.30	0.30	0.25	0.20
Sulfur, mercaptan (thiol), mass % (max)	0.003	0.003	0.003	0.005	0.002
or					
Doctor test	negative	negative	negative	NR [e]	negative
Volatility					
Distillation:					
Initial boiling point (IBP), °C (max)	NR [e]	report	report	150	report
10 % vol recovered at, °C (max)	205	205	205	165	205
20 % vol recovered at, °C (max)	NR [e]	NR [e]	report	NR [e]	report
50 % vol recovered at, °C (max)	report	report	report	195	232
90 % vol recovered at, °C (max)	report	report	report	230	report
Final boiling point (FBP), °C (max)	300	300	300	98 % (250 °C)	300
Residue, vol % (max)	1.5	1.5	1.5	NR [e]	1.5
Loss, vol % (max)	1.5	1.5	1.5	NR [e]	1.5
Flash point [f], °C (min)	38	40 (38)	40 (38)	33 (28)	38
Density at 15 °C , kg/m^3	775 – 840	775 – 840	775 – 840	775 (min) (20 °C)	775 – 830 (20 °C)
Fluidity					
Freezing point, °C (max)	– 47	– 47	– 47	– 50 [g]	– 47
Viscosity at 20 °C, mm^2/s (min)	NR	NR	NR	1.25	1.25
Viscosity at – 20 °C, mm^2/s (max)	8.0	8.0	8.0	8 (– 40 °C)	8.0
Combustion					
Specific energy (net), MJ/kg (min)	42.8	42.8	42.8	42.9	42.8
Smoke point, mm (min)	25	25	25	25	25
or					
Smoke point, mm (min)	18	19	19	NR [e]	20
and naphthalenes, vol % (max)	3.0	3.0	3.0	NR [e]	3.0
or					
Luminometer no. (min)	45	45	45	NR [e]	45
Corrosion					
Copper corrosion (max)	1	1	1	passes	1
Silver corrosion (max)	NR [e]	2 [h]	2	NR [e]	1
Stability					
Thermal stability:				not comparable	
JFTOT test temp., °C	260 [i]	260	260		260
Filter pressure drop, mm Hg (max)	25	25	25		3.3 kPa
Tube deposit (visual) rating	LT 3	LT 3	LT 3		LT 3
Peacock/abnormal deposits	fail	fail	fail		fail
Contaminants					
Copper content [j] (max)	NR [e]	NR [e]	NR [e]	NR [e]	150 µg/kg
Existent gum, mg/100mL (max)	7	7	7	5	7
Water reaction					
Interface rating (max)	1b	1b	1b	1	1b
Separation rating	NR [e]	NR [e]	NR [e]	NR [e]	report
Microseparometer (MSEP) (min)	NR [e]	85 (+ SDA = 70)	85 (+ SDA = 70)	NR [e]	NR [e]
Particulates, mg/L (max)	NR [e]	NR [e]	NR [e]	none	report
Ash content, mass % (max)	NR [e]	NR [e]	NR [e]	0.003	NR [e]
Electrical conductivity [k], pS/m	50 – 450	50 – 450	50 – 450	50 – 600	50 – 450 (20 °C)
Additives					
Antioxidant (hydroprocessed fuels)	optional	mandatory	mandatory	mandatory	optional
Antioxidant (other fuels)	optional	optional	optional	optional	optional
Metal deactivator	optional	optional	optional	NR [e]	NR [e]
Static dissipator	optional	mandatory	mandatory	optional	optional
Fuel system icing inhibitor	permitted	permitted	permitted	permitted	permitted
Lubricity improver	permitted	permitted	permitted	mandatory in HP	permitted

Table 2. Continued.

[a] Specification limits are in terms of Russian GOST test methods and may not be directly comparable with limits obtained using ASTM test methods.
[b] Specification limits are in terms of Chinese GB/T or SH/T test methods and may not be directly comparable with limits obtained using ASTM test methods.
[c] Value in terms of mg KOH/g calculated from GOST limit which is in terms of mg KOH/100 mL.
[d] 25.0 % max aromatics are permitted provided that hydrogen content is reported also.
[e] NR = no requirement under this heading.
[f] Flash point limits are in terms of ASTM D56 (TAG) test method; limits in parantheses are in terms of other, local, test methods.
[g] The GOST test method records wax appearence point on cooling which is 3 °C below the wax disappearance point on warming, as recorded in the ASTM method. Thus the specification limits for TS-1 and Jet A-1 are equivalent.
[h] It is planned to delete the silver corrosion clasification requirement by June 1997.
[i] If the test at 260 °C produces a failure, a pass at a test temperature of 245 °C is acceptable.
[j] Applies only if fuel produced by copper-sweetening process.
[k] Conductivity limits apply only if fuel contains static dissipator additive.

Table 3. National standard aviation fuel test methods

Test name	ASTM [3]	IP [6]	ISO [7]	GOST [a]
Flash point (TAG closed cup)	D56			6356
Distillation	D86	123	3405 : 1988	2177
Corrosion (copper strip)	D130	154	2160 : 1990	6321
Color (Saybolt)	D156			
Vapor pressure (Reid)	D323	69	3007 : 1986	1756
Existent gum	D381	131	6246 : 1981	1567
Kinematic viscosity	D445	71	3104 : 1994	33
Water reaction	D1094	289	6250 : 1982	27154
Sulfur (lamp method)	D1266	107		19121
Density (hyrometer)	D1298	160	3675 : 1993	3900
Aromatics	D1319	156	3837 : 1993	6994
Smoke point	D1322	57	3014 : 1993	4338
Sulfur (high temperature method)	D1552			
Luminometer no.	D1740			17750
Naphthalene hydrocarbons	D1840			17749
Particulate contamination (field method)	D2276	216		
Freezing point	D2386	16	3013 : 1974	5066
Sulfur (X-ray method)	D2622			
Electrical conductivity	D2624	274	6297:1984	25950
Sulfur, mercaptan	D3227	342	3012 : 1991	17323
Undissolved water	D3240			
Thermal oxidation stability (JFTOT)	D3241	323	6249 : 1984	17751
Total acidity	D3242	354		5985
Estimated heat of combustion	D3338			
Estimated hydrogen content	D3343			
Hydrogen content by NMR	D3701	338		
Peroxide no.	D3703			
Flash point (SetaFlash closed tester)	D3828	303	3679/3680 : 1983	
Water separation (MSEP)	D3948			
Density (Digital meter)	D4052	365		
Sulfur (ND-XRF method)	D4294			
Filter flow at low temperature	D4305			
Estimated heat of combustion	D4529	381	3648 : 1994	11065
Heat of combustion (bomb)	D4809			21261
Doctor test	D4952	30	5275 : 1979	
Vapor pressure (automatic)	D5190			
Vapor pressure (mini method)	D5191			
Particulate contamination (lab method)	D5452			
Sulfur (XRF method)	D5453	336	8754:1992	
Freezing point (auto/optical)	D5901			
Freezing point (auto/phase transition)	D5972			
Corrosion (silver strip)		227		
Copper content		225		
Sulfur (Wickbold combustion)		243	4260 : 1987	

[a] Although ASTM, IP and ISO methods are essentially identical, GOST methods may be quite different.

by measuring the total heat released by a unit mass of fuel and subtracting from this value the heat of condensation of the water vapor formed during the combustion. However, this measurement is time consuming, so several empirical relationships have been developed to calculate specific energy from combinations of other inspection tests. The most common is based on the product of aniline point and density, ASTM D 4529 [3]; others use the density, boiling range, and aromatic content, ASTM D 3338 [3], or the hydrogen content and density, IP 355 [6]. All three methods include corrections for fuel sulfur content. Although these calculation methods are used routinely, a method of measurement, ASTM D 4809 [3], is still specified as the referee method for use in cases of dispute.

Volatility. A fuel's volatility or its tendency to vaporize affects evaporation rate from tanks, including aircraft fuel tanks. Volatility can cause vapor lock in fuel systems when the combined fuel and vapor volume exceeds pump capacity; cavitation can result when the fuel vapor pressure exceeds the liquid system pressure between gear teeth or in plain bearings. Volatility also directly affects flammability and causes safety problems both inside and outside of tankage [10]. Because of the complex composition of turbine fuel, its volatility cannot be described by a single fundamental property. Therefore, volatility is normally expressed using a combination of boiling range, flash point, or vapor pressure.

The volatility of kerosine type fuels is controlled by flash point and distillation range (Table 2). Wide-cut fuel is regulated by a combination of distillation range and vapor pressure.

Combustion. Combustion in a modern jet engine is a highly turbulent, high-pressure, and high-intensity mixing process. Engine designers carefully control air flow, fuel flow, and fuel atomization to maximize combustion efficiency and to minimize smoke, carbon deposits, and the formation of partially oxidized gases, such as carbon monoxide. Fuel affects all these as well as the intensity of radiation emitted by the flame [11–13]. No single fuel property requires a greater degree of empiricism than the estimation of fuel combustion characteristics. *Smoke point,* and to a lesser extent the *luminometer,* have been favored because as actual combustion tests, they respond to a combination of other fuel properties including volatility, composition, and viscosity. However, a direct comparison of these laminar, wick-fed flames with the high pressure, turbulent combustor flame is not possible and thus other alternative controls also appear in the specifications. *Naphthalenes content* was added because multi-ring aromatics have shown adverse combustion effects, in terms of flame radiation and smoke formation, out of proportion to their concentration [14]. During the 1980s, the use of hydrogen content, which can be measured very accurately by NMR (ASTM D 3701), gained favor as a possible replacement for the less precise measurements of aromatics and naphthalenes contents. However, it was found that the total hydrogen content of the fuel could not predict reliably the combustion behavior of all fuels[15] and it is now unlikely that it will be adopted for civil fuels, although it continues to find limited use in certain military fuel specifications.

Total aromatic content (ASTM D1319, [3]), covered under "Elastomer Compatibility", does not correlate particularly well with engine combustion, possibly because of the particular separation and classification of hydrocarbons by the method [16].

In general, correlations between engine combustion performance and fuel inspection tests will always be difficult to establish because of the differing performance of the same fuel in different engines, the extreme expense of controlled engine tests, and the long time required to obtain engine experience in the commercial world, where only a limited quantity of minimum quality fuel is burned [16].

Elastomer Compatibility. A major requirement of jet fuels is their compatibility with materials of construction, particularly with the elastomers used in diaphragms, fuel tanks, fuel tank coatings and sealants, linings, and a wide variety of static and dynamic seals in all parts of the fuel system. The fuel property most indicative of elastomer compatibility has been *total aromatic content.*

Low-Temperature Operability. The extremely low temperatures at high altitudes can cause several different fuel problems. During fuel cooling, hydrocarbon molecules precipitate as wax crystals. *n*-Hydrocarbons are the most

insoluble, and in fact measurement of their type and concentration can indicate the insoluble fraction expected at low temperature [17].

The *freezing point* of the fuel is defined as the temperature at which the last wax crystal melts on warming the fuel following a specified cooling process. Traditionally, this has been determined visually but now several automated methods exist which employ variously, light scattering, phase transition measurement or flow through a fine filter (ASTM D 4305) [3].

With the increasing use of twin-engined aircraft on long over-water routes, the so-called extended operations (ETOPS) flights, and with the introduction of very long range aircraft such as the Boeing 747 – 400, there has been increasing focus on fuel low-temperature properties. Studies underway may question the usefulness of "freezing point" as currently measured and could lead to its replacement by a more meaningful, flow-based test.

Low fuel temperature can also affect engine starting, which depends heavily on fuel atomization. Adequate atomization is determined largely by fuel viscosity. Typical viscosities versus temperature are plotted in Figure 1. Because fuels within the same grade have the same viscosity-temperature slope, a single maximum viscosity at a standard temperature adequately controls this property.

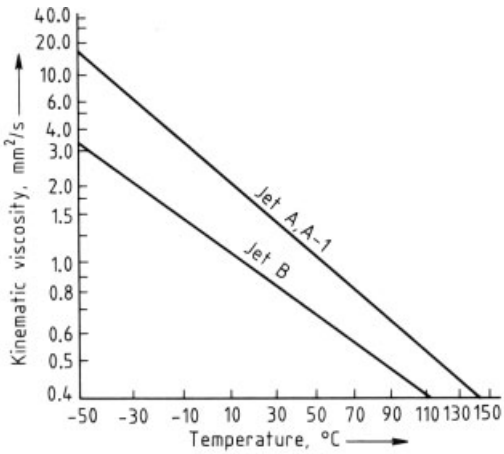

Figure 1. Typical viscosities vs. temperature [18]

2.2.2. Trace Properties

Instead of describing trace properties individually, a better overview is obtained by grouping these properties into the following performance categories.

Oxidative Stability. Jet fuel's exposure to oxidative conditions over a wide temperature and time range can result in a wide variety of problems [19].

Ambient temperature oxidation can form peroxides that degrade certain elastomers, e.g. Buna N, which are used in aircraft fuel systems. Long-term storage can cause soluble and insoluble oxidation products labeled "gums". Peroxide formation is usually seen in fuels which have been produced in the refinery by catalytic hydrogen processes (see Chap. 3) which remove certain trace compounds containing sulfur or nitrogen which are natural anti-oxidants. Anti-oxidancy is restored by the addition of oxidation inhibitor additives (Section 3.2). Gum formation is rare but is monitored and controlled by the Existent Gum test.

High-temperature oxidation is of major concern. Oxidation products – gums, lacquers and insoluble deposits – can form in the hot areas of an aircraft's fuel system causing a number of problems[20]. The heat exchanger of the fuel-cooled oil-cooler can lose efficiency; flow control valves in the fuel control unit can stick or jam, coke deposits can form in the burner feed arms which can, in the worst case, cause plugging of the fuel injector nozzles. These deposits represent such a small fraction of the fuel that the oxygen dissolved in the fuel is sufficient to cause their formation. Although deoxygenation is effective on a laboratory scale, in large scale practice it is impossible to reduce the oxygen content of the fuel to a level low enough to prevent oxidation. The problem is attacked by designing aircraft fuel systems to minimize hot spots; indeed, thermal management is one of the most important criteria in modern engine and aircraft design. In addition, the thermal stability of fuel supplies is controlled by a test in the fuel specification (Table 2) using an instrument called the Jet Fuel Thermal Oxidation Tester (JFTOT), which is a small-scale simulation of a heat exchanger in which formation of lacquer and deposit are measured (ASTM D 3241) [3].

Certain metals, notably copper, are oxidation catalysts and when dissolved in the fuel at extremely low levels, e.g. 20 µg/kg (ppb), can promote thermal instability and deposit formation. Traces of copper can be picked up by the fuel from copper-containing alloys sometimes encountered in the refinery or during transportation. The deleterious effects of this dissolved copper can be alleviated by dosing the fuel with a metal deactivator additive (Section 3.2).

Corrosivity. Corrosion of fuel system metals can be a problem. Copper or bronze in early fuel pump bearings and other components needed protection [1]. Silver surfaces in highly loaded portions of variable displacement pumps were attacked and ultimately removed by shear forces [21]. *Dihydrogen sulfide* and elemental *sulfur* were identified as culprits. As little as 0.5 mg/kg (ppm) of either material caused the difficulties. As a result, all the major fuel specifications contain a test for copper corrosion. The DEF STAN 91-91 (DERD 2494) specification has, in addition, a test for silver corrosion

Solutions to other corrosion difficulties have addressed the offending fuel constituent rather than the metal involved. *Thiols* cause cadmium corrosion [1] and are limited in fuel specifications. *Sulfur,* which includes all organosulfur compounds, causes high-temperature turbine blade corrosion, particularly in the presence of sodium [22]. However, such low sulfur concentrations cause corrosion that virtually complete sulfur removal from the fuel would be needed to solve the problem, and blade metallurgy was changed instead. Therefore the current sulfur specification maxima are not based on any corrosion criteria.

Organic acidity, primarily as naphthenic acids, has been blamed for some diesel engine corrosion, but the low-acidity level permitted by the jet fuel specifications appears to have prevented similar aviation difficulties. The primary role of organic acids is their involvement in the formation of surface-active agents as discussed in the next paragraph.

Water Separation. Free or immiscible *water* is a potential major hazard in aircraft. When cooled below 0 °C the resulting ice can block orifices, screens, or filters. In large quantities water will "put out the fire" by interrupting fuel flow. Water can lead to microbial growth in stagnant parts of the fuel system. Therefore, every effort is made to remove all immiscible water by special filter-separators before the aircraft is fueled. Satisfactory operation of these filters depends on the maintenance of internal hydrophobic surfaces. Trace levels of surfactants render these surfaces ineffective [23]. Although the organic acids normally in jet fuel are not strong surfactants, they react with sodium hydroxide or other basic materials. The resultant salts or soaps can have detrimental effects [1], [24], [25]. The lower organic acidity maximum in the British specification offers greater protection in this regard than does the higher level in ASTM D 1655, but surfactants are not limited to those formed by organic acids, and additional tests are used for surfactant detection and control.

The Water Reaction test gives a semiquantitative indication of insoluble interface material formed when fuel is shaken with buffered water. In a much more sensitive test a fuel–water emulsion is pumped through two layers of coated fiberglass from a filter-separator. The ability of the filter element to strip all water from the emulsion is measured.

The earliest test, the Separometer (ASTM D 2550 [3]), was developed to detect low levels of sodium sulfonate [33]; but this test as well as the later versions, the Miniseparometer (ASTM D 3602) and the Microseparometer (ASTM D 3948), are widely used in the United States and elsewhere to test for the presence of various surfactants. Because of the difficulties in correlating results from these tests with field performance, there is only limited industry agreement on the definition of acceptable fuel by these tests. ASTM D 2550 and D 3602 have now been discontinued.

Static Electricity. Whenever a hydrocarbon liquid flows with respect to another surface, an electrical charge is generated in the liquid and an equal but opposite charge is imposed on that surface. One set of charges is swept along with the fuel while the opposite charges, which accumulate along the wall surface, usually leak to ground (earth). This charge separation results in a rise in voltage in the moving fuel [26]. High speed pumping of aviation fuel through pipes and filters can generate very high charges and when the charged fuel enters a tank, a substan-

tial voltage difference may be produced between the surface of the liquid and the tank walls and this may result in a static discharge [27]. If the vapor space above the liquid is in the flammable range, such a spark discharge can cause a fire or explosion.

Charge build-up can be reduced by reducing the flow-rates of the fuel or designing and constructing facilities – pipework, filters, valves, etc. – to minimize charge generation and/or provide time for the charge to relax to ground before entering a tank [28]. Another approach is to add a static dissipater additive to the fuel (Section 3.2) to increase the electrical conductivity of the fuel to facilitate charge relaxation.

2.3. Nonspecification Properties

Some aviation fuel properties are not limited by specifications but are useful in equipment design or performance.

Absolute Vapor Pressure. Reid vapor pressure, specified for Jet B fuel, is not an absolute vapor pressure because it is measured with a vapor–liquid ratio of 10 : 1; it is determined only at 38.8 °C, and also it includes the partial pressures of water vapor and air. With low-vapor-pressure products, such as Jet A or A-1, absolute or true vapor pressure is used instead. Figure 2 shows typical true vapor pressure-temperature curves for different jet fuel grades [18].

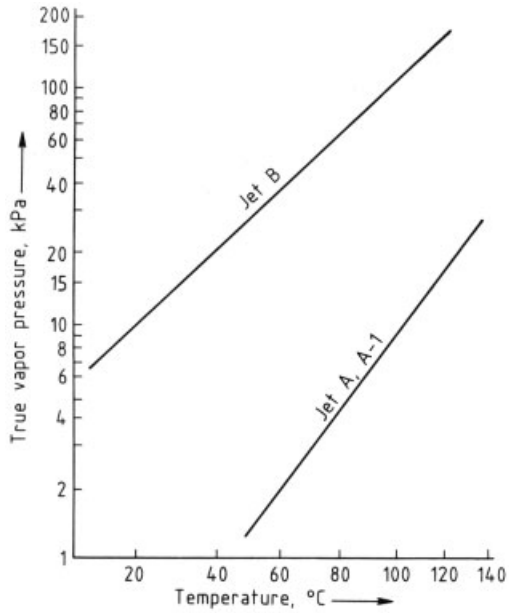

Figure 2. True vapor pressures [18]

Water Miscibility. Even the very low concentration of dissolved water in aviation turbine fuels is hazardous if this water separates and freezes in a critical aircraft location. Because dissolved water cannot be avoided in vented systems and is not removed by airport fuel handling equipment, aircraft fuel systems either heat the fuel above 0 °C or depend on anti-icing additives (see Section 3.2). Typical water saturation levels for commercial jet fuels are listed in Table 4 [18].

Table 4. Typical nonspecification properties [18]

	20 °C	40 °C	60 °C
Water solubility, vol %			
Jet A, A-1	0.0049	0.0088	0.0163
Jet B	0.0076	0.0138	0.0255
Specific heat, kJ kg^{-1} K^{-1}			
Jet A, A-1	1.95	2.04	2.13
Jet B	2.06	2.14	2.23
Thermal conductivity, Wm^{-1} K^{-1}			
Jet A, A-1	0.1155	0.1124	0.1085
Jet B	0.1155	0.1124	0.1085
Dielectric constant at 400 Hz			
Jet A, A-1	2.13	2.10	2.07
Jet B	2.06	2.03	2.00
Surface tension, mN/m			
Jet A, A-1	23.3	21.7	20.2
Jet B	21.8	20.5	18.4

Lubricity. Various aircraft fuel system components – fuel pumps, servo valves, fuel control units – rely on the fuel to lubricate their moving parts. The effectiveness of a jet fuel as a lubricant in such equipment is referred to as its "lubricity". Differences in component design and materials result in varying degrees of equipment sensitivity to fuel lubricity. Similarly, jet fuels vary in their level of lubricity. The chemical and physical properties of jet fuel cause it to be a relatively poor lubricating material under high temperature and load conditions where lubrication may depend on adsorbed films of trace polar compounds present in the fuel, so-called "boundary lubrication". Severe hydroprocessing (see Chap. 3) removes trace components, resulting in fuels which tend to have lower lubricity than straight-run or wet-treated fuels. This lack of lubricity can be remediated by the addition of corrosion inhibitor/lubricity improver additives (Section 3.2). Such additives are mandatory in most military jet fuel specifications but they have not found widespread use in civil aviation. It is generally acknowledged that the ultimate solution to the fuel lubricity "problem" is to design fuel system equipment such that it is insensitive to fuel lubricity. Civil fuel specifications do not contain a requirement for a minimum level of lubricity, not least because there is no method of test for this property which is universally accepted. One method, the Ball-on-Cylinder Lubricity Evaluator (BOCLE) ASTM D 5001 [3], can be used for identifying possible problem fuels and is used by equipment manufacturers to monitor special low-lubricity fluids used in development testing of fuel system equipment, and by the USAF to monitor the lubricity additive concentration in their fuel supplies.

Other Properties. Detailed data on other properties, including gas solubility, flammability, surface tension, heat of vaporization, enthalpy, specific heat, thermal conductivity, and dielectric constant, are contained in the *CRC Handbook of Aviation Fuel Properties* [18]. Some typical values for several properties are given in Table 4.

3. Production

3.1. Fuel

Since its earliest days virtually all aviation turbine fuel has been refined from selected petroleum crudes by atmospheric distillation followed by limited processing to reduce thiol or total sulfur. Crude selection is vital for such simple processing because the only other variable available for controlling bulk properties is adjusting the product's *boiling range*. Desirable crudes are low in sulfur (sweet crudes), low in *n*-hydrocarbons (nonwaxy crudes), low in aromatic content, and high in product boiling between 160 and 290 °C.

Adjusting either end of the boiling range has predictable results. Lowering the initial boiling point improves combustion and low-temperature properties but decreases handling safety (see Table 5). Raising the final boiling point degrades combustion and low-temperature performance and has no effect on handling safety. Because the crude also yields both lighter and heavier fractions, the demand for these products affects distillation cut points and therefore the volume of jet fuel available [29].

Table 5. Alteration of boiling range

Lowering initial temperature	Raising final temperature
Increases	
smoke point	density
hydrogen content	aromatic content
J/mass	naphthalenes
vapor pressure	viscosity
Decreases	
density	smoke point
flash point	hydrogen content
aromatic content	J/mass
naphthalenes	
freezing point	
J/vol	
viscosity	
Leads to	
better combustion quality	poorer combustion quality
better low-temperature properties	poorer low-temperature properties
poorer handling safety	unchanged handling safety
higher gravimetric specific energy	lower gravimetric specific energy
lower volumetric specific energy	higher volumetric specific energy

Total sulfur can be reduced by relatively low-pressure hydrotreating. Thiol sulfur is reduced

(sweetened) by low-pressure hydrotreating or wet chemical treating, such as Merox, Linde, Doctor, or Bender treating. Flash point may be adjusted so that the same product can be sold as either aviation turbine fuel or commercial illuminating or heating kerosine.

Some heavier fractions are hydrocracked to make jet fuel, but the 160 – 290 °C fraction from catalytic cracking is not suitable for jet fuel unless cracking is followed by high-pressure hydrotreating. A limited amount of kerosine is hydrotreated to reduce aromatic content and increase smoke point by aromatic saturation. In the United States the finished jet fuel is often filtered through towers of activated Attapulgus clay to adsorb surfactants and to remove particulate fines.

3.2. Additives

All aviation fuel additives must be tested and approved. The extent of testing is at the discretion of the aircraft and engine manufacturers who issue the approval. Additive composition plays a major part in approval procedures, with ash-forming materials receiving the greatest scrutiny. Recommended Practice D 4054 (ASTM) [3] covers the evaluation of compatibility between additives and aircraft fuel system materials. Approved additives and their permitted levels are normally listed in the fuel specifications. Unapproved additives cannot legally be used in aircraft operation.

Oxidation Inhibitors. Oxidation inhibitors prevent peroxide formation and improve storage stability in certain fuels (→ Antioxidants). They are generally secondary and tertiary hindered phenols and are listed in the fuel specification by chemical composition. They are effective primarily during storage, not during high-temperature use. Antioxidants are mandatory for hydroprocessed fuels in the major military fuel specifications and in civil Jet A-1 covered by the DEF STAN 91-91 (DERD 2494) specification and IATA and Joint System Checklist documents, but only optional in the ASTM D 1655 specification. Dosage rate is 17 – 24 mg/L. Annual consumption of oxidation inhibitors in aviation turbine fuel is difficult to estimate because the same additives are widely used in other petroleum products.

Metal Deactivators. Metal deactivators are intended to passivate trace metals by chelation (→ Antioxidants). The only approved deactivator is *N,N′*-disalicylidene-1,2-diaminopropane [*94-91-7*]. It is permitted up to 5.7 mg/L, but is rarely used.

Corrosion Inhibitors and Lubricity Improvers. Unlike the preceding additives, corrosion inhibitor formulations are proprietary. They usually contain dibasic acids that function as liquid phase rust inhibitors. The polar ends of the molecule adsorb onto metal surfaces and in ground handling and distribution systems, e.g., steel pipelines, they act to protect surfaces from corrosion. In aircraft fuel systems, the tendency to adsorb on to metal surfaces is utilized for lubricity improvement rather than corrosion inhibition. Because these additives are surface-active (commonly referred to as "surfactant"), they can display the adverse side-effect of impairing the water separation properties of the fuel. Furthermore, their surfactancy means that they can be lost from the fuel during distribution (as, of course, they should be if they are to protect metal surfaces from corrosion); consequently, if added for lubricity enhancement, these additives are often injected at the airport, just prior to aircraft fueling. Lists of approved additives are published in the Qualified Products Lists (QPLs) associated with the U.S. MIL-I-25017 and UK DEF STAN 68-251 specifications for this type of additive.

Static Dissipater Additives (SDAs), also referred to as Conductivity Improving Additives, are added to jet fuel to increase electrical conductivity, and thereby reduce the tendency of a fuel to accumulate dangerous electrostatic charge during handling [28]. There are only two additives approved for use in aviation fuel, Shell ASA-3 and Octel (formerly Du Pont) Stadis 450. Production of the Shell additive ceased in 1991 leaving Stadis 450 as the only additive currently available. The concentration is adjusted to obtain an electrical conductivity within the range 50 – 450 pS/m but with a maximum permitted dosage of 3 mg/L. A further addition up to 2 mg/L is permitted if conductivity has been

lost during fuel distribution. The minimum conductivity limit is set to ensure charge relaxation sufficient to alleviate static discharges, while a maximum limit is applied to avoid causing anomalies in aircraft fuel tank contents gauges which operate on an electrical capacitance principle. The use of SDA is mandatory in military JP-4 and JP-8 fuels and in civil Jet A-1 outside the United States ; its use is optional in Jet A fuel within the United States.

Anti-icing additive, also referred to as Fuel System Icing Inhibitor (FSII), prevents the freezing of trace levels of free water in fuel at the low temperatures encountered in aircraft fuel tanks at high altitudes and avoids blockage of fuel lines and filters by ice crystals.). Traditionally, ethylene glycol monomethyl ether (2-methoxyethanol) [*109-86-4*] , has been used for thispurpose but, owing to its toxicity, this has largely been superseded by diethylene glycol monomethyl ether (2-(2-methoxy)ethanol) [*111-77-3*]. Normal dosage rate is between 0.1 – 0.15 vol %.Such use is mandated in military fuels but in civil aviation its use is restricted to small aircraft which do not have heaters in their fuel systems. The additive, which is only sparingly soluble in the fuel, preferentially dissolves in any free water present (partition coefficient about 200 : 1) where it acts as an antifreeze (→ Antifreezes). This affinity for water can be problematic in fuel distribution systems and additive loss must be considered and possibly corrected. FSII has the additional property of acting as a biostat, preventing the growth of bacteria and fungi in fuel tanks (see below).

Biocides. Microorganisms that metabolize hydrocarbons can be present in both ground and aircraft fuel systems [1]. Certain anaerobic bacteria that reduce sulfate to hydrogen sulfide have caused copper and silver corrosion. The formation of slimes, fungal mats, and corrosive byproducts has caused wing tank corrosion and capacitance gage failures. The only generally approved fuel-soluble biocide is Biobor JF (Hammonds), a blend of 95 % dioxaborinanes and 5 % naphtha. Recommended dosage to kill microorganisms is 226 mg/L; maintenance level is 108 mg/L. An alternative material, which is approved for most applications, is Kathon FP1.5 (Rohm & Haas) which is based on a mixture of a methyl-isothiazolinone [*2682-20-4*] and a chloromethyl isothiazolinone [*26172-55-4*].Treatment levels are between 100 and 300 mg/L. Although treatment withbiocides kills the microbial growths in fuel tanks, subsequent tank cleaning is still essential to remove all the dead material which might otherwise lead to filter blockage.

Thermal Stability Additives. In the early 1990s, the USAF perceived that future aircraft designs would demand a fuel with enhanced thermal oxidative stability. They decided that the most economic way of achieving this would be to use fuel additives to increase the thermal stability of the existing fuel type, JP-8, rather than developing a special fuel. It was calculated that an increase of 100 °F (56 °C) in the thermal stability level of the fuel would be required and this led to the new fuel being referred to as "JP-8 + 100". The extensive research program to develop and test such additives was referred to as "HiTTS", standing for High Temperature Thermal Stability [30]. The main problems at high temperature are fouling of the injector feed arm and coking in the fuel supply rings to the augmenter (afterburner or reheat) section. An additive package comprising a detergent/dispersant, an antioxidant, and metal deactivator proved to be very effective in preventing deposit formation in the hot areas of existing aircraft fuel systems. Already the benefits in terms of reduced maintenance costs are being appreciated. However, because the dispersant/detergent element of the additive package is very surface active, it cannot be added to the fuel in storage because it would encourage water and dirt pick-up and also disarm the filter/water separators in the airport fueling system. Instead, the additive package is injected into the fuel as the aircraft is being fueled.

4. Handling, Storage, and Transportation

System Descriptions. It is essential to prevent aviation fuel from being contaminated with other petroleum products between the time it is manufactured in the refinery until the moment it is put on-board an aircraft. Consequently

fixed storage facilities are designed to keep aviation fuel physically isolated from other products. However, large pipelines, marine transports and sometimes even road tankers are multiproduct systems. Each transportation mode therefore has an elaborate protocol to protect the quality of all the products it carries. Pipelines carry compatible products both preceding and following jet fuel; the interface between products is cut into the non-aviation grade. Marine and road transports depend on cleaning procedures, with the amount and type of cleaning dependent on the preceding product. When aviation fuel has been transported through a non-dedicated (i.e., a multiproduct) system, samples are taken for recertification testing in which those properties most sensitive to contamination, e.g., flash point, freezing point, gum, etc., are retested and the results compared with the original values reported on the refinery Certificate of Quality to confirm that no inter-product contamination has occurred. At the airport, the fuel is delivered from storage to the aircraft either via underground pressurized piping (hydrant system) through mobile hydrant dispensers or by specialized tank trucks called fuelers. Both types of fueling vehicle are equipped with filter/water separators or filter monitors to ensure only clean, dry fuel is pumped into the aircraft.

Contamination Removal Equipment. Multiproduct transportation systems and long distances between refinery and airport furnish opportunities for water, rust, dust, surfactants, and other contaminants to enter the fuel. Specialized equipment and procedures remove these unwanted materials [31]. Water and solids settle in storage, assisted by suction lines that float just beneath the liquid level. Filter/water separators [32] and filter monitors [33] remove free water and fine solids, while activated Attapulgus clay filters can be used to adsorb surfactants.

The operation of this equipment is checked regularly with a variety of test equipment. Solids are detected and measured per ASTM D 2276 and ASTM D 3830 [3]. Surfactant removal by clay filters can be observed with the Microseparometer (ASTM D 3948 [3]). A variety of tests exist for the detection of low levels of undissolved or free water, with most of the tests containing a water-sensitive dye; ASTM D 3240 [3] measures water levels ≥ 1 ppm.

5. Legal Aspects

All aircraft, including all flight components, are certified for flight by a government agency which promulgates and enforces the certification rules. These rules have the force of law. Aircraft and engines use fuel meeting certain specifications during certification. These specifications become part of the aircraft's operating certificate. The fuel must therefore completely meet the certification specification even though a deviation may not affect operation adversely. Normally aircraft and engine manufacturers designate industry fuel specifications (e.g., ASTM D 1655 [3]), but manufacturers can certify to their own proprietary specification. In such a case any points of difference from the industry norms must be specified. Once certified, aircraft must be operated by the rules of the country in which they are registered.

6. Environmental Aspects

Aircraft exhaust emissions are controlled by regulations issued by the International Civil Aviation Organisation (ICAO). These regulations define the permitted levels of emissions of nitrogen oxides (NO_x), carbon monoxide, unburned hydrocarbons, and smoke for each engine type. At present, emissions are controlled only for the landing/take-off (LTO) cycle, which is defined as engine idle, taxi out, take-off, and climb to 900 m, descent from 900 m, taxi in, and idle. Exhaust emissions at altitude are not restricted although there are numerous studies underway to determine the effects of high altitude emissions on global warming and ozone depletion [34–36]. Future regulation is expected but this is likely to be met by further advances in combustion chamber design rather than from significant changes in the composition or properties of aviation fuel.

7. Economic Aspects

Compared with gasoline and diesel fuel production, aviation turbine fuel represents a small proportion of the crude barrel. In 1995, worldwide

production of kerosine was 285×10^6 t of which ca.190×10^6 t was for jet fuel with the remainder for non-aviation uses (heating, lighting, cooking, etc.). This total kerosine volume represents ca. 9 % of world oil consumption [37]. World air travel is predicted to grow, on average, by 5.5 % per year over the next decade and 4.9 % over the coming twenty years [38–40]. Highest growth rates are in the "Pacific Rim" countries. By 2006, air travel will rise by nearly 75 %; by 2016, it will reach two-and-a-half times current levels. Fuel volumes will also grow significantly but at a lower rate due to improvements in aircraft fuel efficiency. The United States is the largest single consumer of jet fuel with about 50 % of the total world volume; this share is expected to decrease in the future as the Far Eastern economies continue to grow.

A forecast of world aviation fuel demand in 10^6 t is given below [37]:

1990	1993	1995	2000	2005	2010	2014
161	175	189	255	330	420	500

8. Future Trends

8.1. Petroleum-Derived Fuels

Despite the predicted increase in demand for aviation fuel, there should be no overall supply problem. Both refining capacity and processing technology are available to supply these volumes. Whether the quality of jet fuel will change significantly in the future is open to debate. In the early 1980s, predictions tended to be gloomy with the expectation that fuel quality would inevitably deteriorate due to decreasing availability of sweet crudes and the increasing use of heavy and high sulfur crudes. This led to design studies aimed at developing engines which could cope with the resultant poorer combustion quality [41]. However, these predictions were not borne out during the 1980s and 1990s because new oilfields continued to be discovered. Advances in process technology will be able to maintain, or even improve on, fuel quality for the foreseeable future.

Conventional petroleum crude will continue to form the major source for jet fuel production well into the next century. Experimental production of jet fuel from oil shale in the United States during the 1980s, although successful, was terminated because it was economically non-viable in a world of continuing low crude oil prices. However, this technology remains on the shelf and could be resurrected should the economic climate change significantly. Production of jet fuel from tar sand derived crudes is practiced in Canada, but gives only a very small proportion of the total fuel production.

Synthesis of liquid fuels from natural gas can be achieved by the Shell Middle Distillate Synthesis process (SMDS) [42], [43]. Whilst producing kerosine of a very high quality in terms of combustion performance, the SMDS product itself is unlikely to be suitable as a jet fuel because its complete lack of aromatics causes concern in the areas of compatibility with elastomers and lubricity. However, the product remains an excellent blending component for upgrading lower quality conventional kerosines which would otherwise be unsuitable as jet fuel

8.2. Alternative Fuels

Kerosines derived from *coal* using the Fischer–Tropsch process (→ Coal Liquefaction, Chap. 2.2.) have been synthesized in South Africa by SASOL for a number of years. The process produces kerosine which could be considered suitable as jet fuel although it has never been supplied as such. Although energy intensive, increasing jet fuel demand in southern Africa has prompted the producers of this synthetic product to seek approval from the aircraft and engine manufacturers for its use as jet fuel. However, like SMDS kerosine, many of the coal-derived fractions contain no aromatics and will need to be blended with other streams before being acceptable. Although this production route has been confined to South Africa, at some time in the future it may be applied in other countries with large coal reserves, e.g., China. *Oxygenated liquids* like alcohols or esters have been examined as potential aviation fuels or fuel constituents. The pure materials penalize aircraft operation by their much lower specific energy content, whereas blends with petroleum fuels may suffer from water extractability and system material compatibility difficulties [29].

Thus the widespread adoption of oxygenated compounds, particularly methanol, as commercial aviation fuel appears unlikely.

It is often said that the long-term future may lie with *cryogenic fuels* such as liquid methane or liquid hydrogen. Successful flights have been made using methane in Russia, while the "Cryoplane" project [44] is aimed at proving the feasibility of a liquid-hydrogen powered aircraft. However, a major barrier to these cryogenic fuels is that they cannot be used in conventional aircraft, and a new generation of aircraft specifically designed for them would need to be developed. Furthermore, completely new production, storage, distribution, and handling systems would have to be constructed all around the world to ensure supply at all major airports. Proponents of cryogenic fuels tend to underestimate the capital investment involved in establishing a supply infrastructure and it is questionable whether these fuels will ever become commercially viable [45]. *Methane* may be more advantageously used as the starting material in the manufacture of conventional liquid hydrocarbon fuels rather than as a liquid fuel itself [46]. *Liquid hydrogen* has excellent combustion qualities and produces no CO_2 and, at first sight, appears to be more environmentally friendly than conventional kerosine-based jet fuel. However, unless it is produced via nuclear or hydroelectric power, its production from petroleum resources would continue to consume energy and produce CO_2. Moreover, combustion of hydrogen in aircraft engines would still not avoid the production of NO_x and, furthermore, produces 2.5 times as much water vapor as kerosine, which will also contribute to the greenhouse effect.

The massive investments required for new fuels together with the long lead time in design and certification will encourage the use of fuels similar to those currently used for the longest possible time.

9. References

1. M. Smith: *Aviation Fuels,* G. T. Foulis Ltd., Henley-on-the-Thames 1970, 495 pp.
2. V. B. Guthrie (ed.): *Petroleum Products Handbook,* McGraw-Hill, New York 1960, pp. 5-1 to 5-28.
3. *1997 Annual Book of ASTM Standards ,* **vols 05.01 – 05.03,** ASTM, West Conshohocken, PA 1997.
4. Defence Standard 91-91(DERD 2494), Issue 2, Ministry of Defence, London, 1996.
5. *IATA Guidance Material for Aviation Turbine Fuels.,* 2n ed., IATA Montreal 1994.
6. *Standard Methods for Analysis and Testing of Petroleum and Related Products,* Institute of Petroleum, London, 1997.
7. International Organisation for Standardisation, Case Postale 56, CH-1211 Geneva 20, Switzerland.
8. *Manual on the Significance of Tests for Petroleum Products, MNL 1* , 6th ed. , ASTM 1993, pp. 34 – 53.
9. *Petroleum Measurement Tables,* **vol. I – XII** ASTM, West Conshohocken, PA.
10. "Aviation Fuel Safety 1975," *CRC Report 482,* Co-ordinating Research Council, Atlanta 1975.
11. C. A. Moses, D. W. Naegeli: *American Society for Mechanical Engineers Preprint 79-GT-178,* March 1979.
12. J. M. Casper, E. E. Ekstedt, W. J. Dodd, M. W. Shayeson, *NASA Report CR-159855,* Cleveland Sept., 1980.
13. R. E. Vogel, D. L. Troth, A. J. Verdouw, *U.S. Air Force Report AFAPL-TR-79-2072,* Wright-Patterson, Ohio, April, 1980.
14. E. A. Droegemueller, R. K. Nelson: *SAE Preprint 47 A,* SAE, Warrendale, Pa., April 1958.
15. T. T. Bowden, J. H. Pearson, *The Effect of Fuel Composition Upon Combustion Performance in a Rolls-Royce Tyne Combustor, ASME Paper 85-GT-39,* ASME, March 1985.
16. National Research Council, *Aviation Turbine Fuels-An Assessment of Alternatives,* National Academy Press, Washington, D.C., 1982.
17. W. A. Affens, R. N. Hazlett, D. E. Smith, *NRL Report 8692,* Naval Research Lab., Washington, D.C., July, 1983.
18. "Aviation Fuel Properties," *CRC Report 530,* Coordinating Research Council, Atlanta 1983, 113 pp.
19. "CRC Literature Survey on the Thermal Oxidation Stability of Jet Fuel," *CRC Report 509,* Coordinating Research Council, Atlanta, April, 1979.
20. "Thermal Oxidation Stability of Aviation Turbine Fuels," *ASTM Monograph 1,* ASTM, West Conshohocken, PA 1991.
21. S. A. W. Thompson, S. R. Hills, N. P. H. Brown, A. Lewis, *J. Inst. Petr. (London)* **50** (1964) 72.

22. V. Patarini, N. S. Bornstein, M. A. DeCrescente: *ASME Preprint 78-GT-82,* American Society of Mechanical Engineers, New York April, 1978.
23. J. A. Bert, H. R. Porter: *Paper 11 a – 63,* American Petroleum Institute, Washington, D.C., May 13, 1963.
24. R. N. Hazlett, H. W. Carhart, *Filtr. Sep.* **9** (1972) 456.
25. "Water/Fuel Separation Characteristics: A Bibliography," *CRC Report 460,* Coordinating Research Council, Atlanta July, 1973.
26. W. M. Bustin, W. G. Dukek: *Electrostatic Hazards in the Petroleum Industry,* Research Studies Press, London 1983.
27. "Generation and Dissipation of Electrostatic Charge During Aircraft Fueling-A Selected Literature Survey," *CRC Report 466,* Coordinating Research Council, Atlanta 1964.
28. ASTM D 4865 -96: 1996.
29. K. H. Strauss: *AIAA Preprint AIAA-81-0770,* May, 1981.
30. S. P. Heneghan, S. Zabarnick, D. R . Ballal, W. E. Harrison, " JP-8 + 100: The Development of High-Thermal-Stability Jet Fuel," *ASME Journal of Energy Resources Technology* **118** (1996) September, 170 – 179.
31. *Manual of Aviation Fuel Quality Control Procedures, MNL 5,* 2nd ed.,ASTM, West Conshohocken, PA, 1995.
32. *Specifications and Qualification Procedures - Aviation Jet Fuel Filter/Separators, API Bulletin 1581,* 3rd ed., American Petroleum Institute, May 1989.
33. *Specifications and Qualification Procedures - Aviation Fuel Filter Monitors with Absorbent Type Elements,* Institute of Petroleum, January 1995.
34. L. J. Archer, *Aircraft Emissions and the Environment: CO_X, SO_X, HO_X and NO_X,* Oxford Institute for Energy Studies, 1993.
35. Air Transport and the Environment, IATA Seminar, Washington, March 1993.
36. Air Transport and Environmental Protection, Airbus Industrie, March 1993.
37. *Aviation Fuels Business Worldwide 1996,* Petroleum Economist, London April 1996.
38. "1997 Current Market Outlook," Boeing Commercial Airplane Group, March 1997.
39. Outlook for Commercial Aircraft 1995 – 2014, McDonnell-Douglas, August 1996.
40. Global Market Forecast 1995 – 2014, Airbus Industrie, March 1995.
41. J. P. Longwell, J. Grobman, *J. Eng. Power* **101** (1979) 161.
42. The Shell Middle Distillate Synthesis Process, Shell Selected Paper, November 1989.
43. P. J. A. Tijm, J. M. Marriott, The Markets for Shell Middle Distillate Synthesis Products, *Alternative Energy '95*, Vancouver, May 1995.
44. "The Light Stuff," *Chemistry in Britain* 1997 29 – 31.
45. A. M. Momenthy, Aviation Fuel in a Changing World, Boeing, January 1994.
46. O. Hadaller, A. M. Momenthy, , Fuel Reformulation and Jet Fuel, Boeing, November 1994.

Azelaic Acid → Dicarboxylic Acids, Aliphatic

Azine Dyes

Horst Berneth, Lanxess Deutschland GmbH, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Phenazine Dyes 5
2.1. Monoaminophenazine Dyes 5
2.2. Diaminophenazine Dyes 6
2.2.1. Eurhodine Dyes 6
2.2.2. Basic Safranine Dyes 6
2.2.3. Acid Safranine Dyes 8
2.2.4. Indulines 10
2.2.5. Nigrosines 11
2.2.6. Aniline Black 13
2.2.7. Oxidation Dyes 15
2.3. Aminohydroxy- and Monohydroxyphenazine Dyes 17
3. Oxazine Dyes 17
3.1. Monoaminophenoxazine Dyes 18
3.2. Diaminophenoxazine Dyes 19
3.3. Aminohydroxyphenoxazine Dyes . . 22
3.4. Mono- and Dihydroxyphenoxazine Dyes 23
4. Thiazine Dyes 23
4.1. Diaminophenothiazine Dyes 24
4.2. Aminohydroxyphenothiazine Dyes . 26
4.3. Hydroxyphenothiazine Dyes 27
5. Bisazine Dyes 27
5.1. Fluorindine Dyes 28
5.2. Dioxazine Dyes 28
5.3. Dithiazine Dyes 33
6. References 33

1. Introduction

The azine dyes include the phenazine, the oxazine, and the thiazine dyes. They correspond to the following general formula, the most important resonance formulas of which are shown:

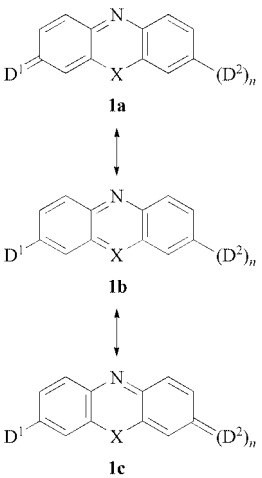

In these formulas D^1 and D^2 stand for an auxochrome group, usually an amino, arylamino, monoalkylamino, dialkylamino, or hydroxyl group; X stands for −O−, −S−, or −NR− (R = H, alkyl, aryl); and n is 0 or 1. Therefore, resonance formula **1** only has meaning when $n = 1$. The azine dyes include cationic, anionic, and neutral dyes, depending on the kinds of auxochrome groups and substituents. The multitude of possible resonance formulas is responsible for the deep and intense color of the azine dyes. The azine dyes have this in common with the methine dyes and the triphenylmethane dyes, and they all obey the same general theory of methine dyes [1].

Monoamino- and diamino-, monohydroxy- and dihydroxy-, and aminohydroxyazine dyes are differentiated according to the kind and number of auxochrome groups in formulas **1** – **1**. Because hydroxyazine dyes are present predominantly in the keto form (D^1 or D^2 is =O), they are referred to as phenazone, phenoxazone, and phenothiazone dyes.

It is clear from the general formulas above that, because of the −X− moiety, azine dyes are cyclized derivatives of indamine or indophenol dyes. This relationship is not only of a formal nature; the most important methods of preparation of the phenazine, oxazine, and thiazine dyes proceed via an intermediate indamine or indophenol dye. Either one component of these intermediates already contains the −XH group, and ring closure can take place immediately, or

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_213.pub2

the −XH group must first be introduced into the indamine or indophenol dye in a separate step. Oxazine dyes are prepared by the former procedure, thiazine dyes by the latter, and phenazine dyes by both.

Loveless, Brown, and Horrocks [2] investigated spectrophotometrically the formation of the phenazine ring via cyclization of *N*-(2-aminophenyl)-*p*-benzoquinonediimine **2**, the oxidation product of 2,4′-diaminodiphenylamine **2**. In the pH range 6 – 9, the amino group adds intramolecularly to the quinone ring, and 2-aminophenazine **2** is formed by oxidation. Above pH 10, hydrolysis of the imino group occurs, and *N*-(2-aminophenyl)-*p*-benzoquinoneimine **2** is formed. In this case, too, the amino group adds to the quinone ring, and 2-hydroxyphenazine **2** is obtained by oxidation.

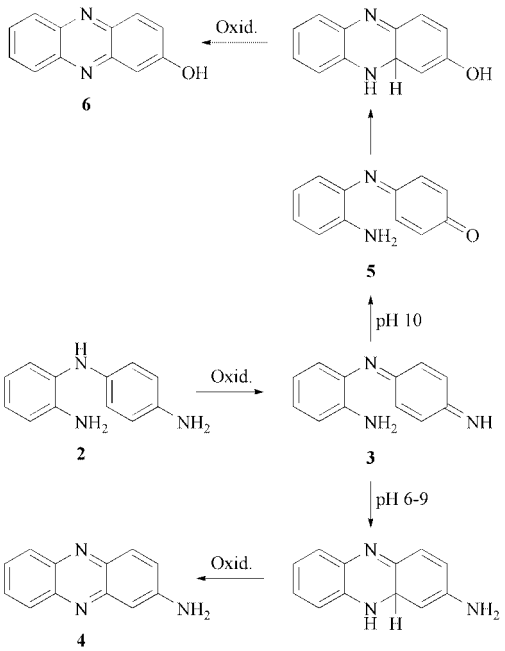

In an investigation of the reaction kinetics, Corbett [3] was able to establish the following mechanism of formation of phenosafranine **3** by the joint oxidation of *p*-phenylenediamine and aniline (see top of next page).

A rapid oxidation reaction is followed each time by the rate-determining nucleophilic addition of aniline to the quinoid system. The final addition step takes place intramolecularly, forming the phenazine ring system. In a similar system, the formation of indophenols in the Gibbs reaction, the initial step is a single electron transfer reaction between the quinoid system and the phenolate anion followed by fast combination of the radicals or by a chain reaction sequence [4].

A further method for preparation of phenazine dyes involves the reaction of azo compounds with aromatic amines. In this case, the formation of the indamine is preceded by a series of oxidation, reduction, and disproportionation reactions [5]. The first step is the exchange of the 1-amino group of 1-amino-4-benzeneazonaphthalene for the aniline moiety yielding **4**. Subsequently, an autocatalytic reaction begins according to the scheme on the bottom of the next page.

Common to both reaction mechanisms is the fact that two indamine intermediates (**3** and **3**, and **4** and **4**, respectively) precede the formation of the phenazine dyes **3** and **4**.

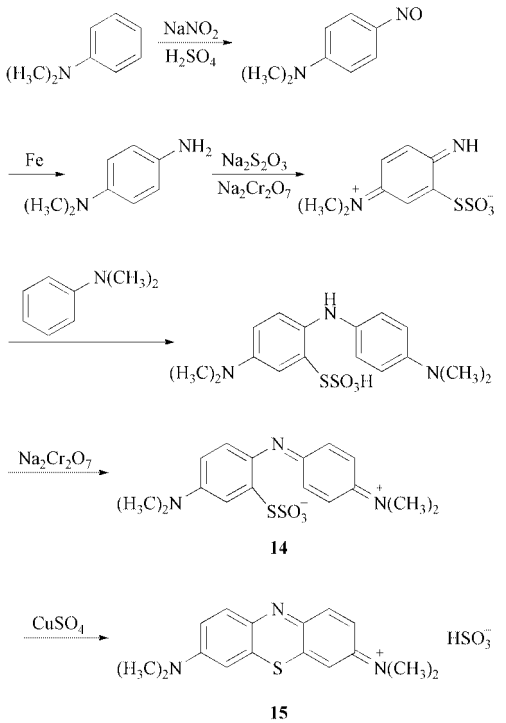

The ring-forming sulfur atom in the thiazine dyes is also introduced by addition to a quinoid intermediate. In addition to hydrogen sulfide and sulfides, sodium thiosulfate is used as sulfur donor in the most important industrial preparation procedure [6]; thiosulfonic acids are formed

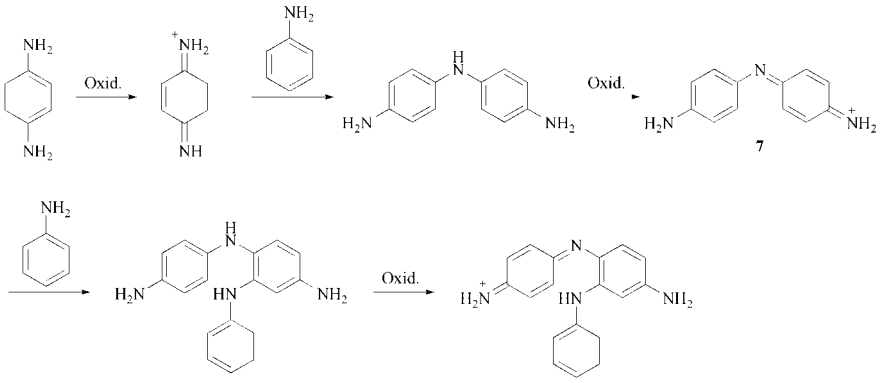
Oxid.
Oxid.
7

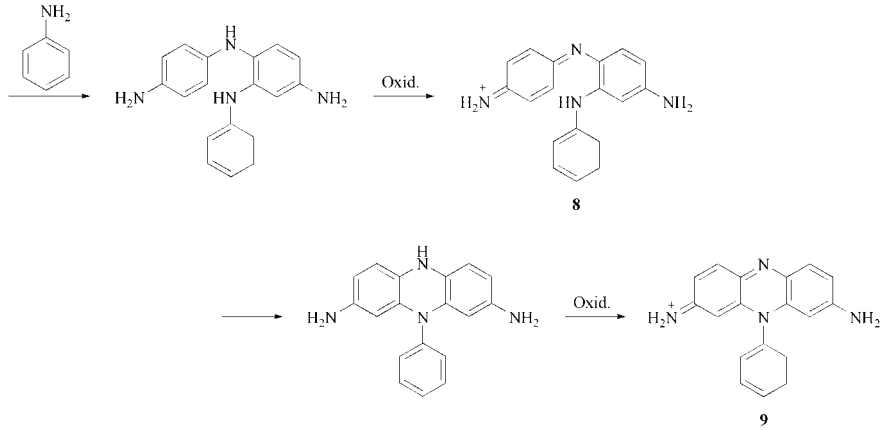
Oxid.
8
Oxid.
9

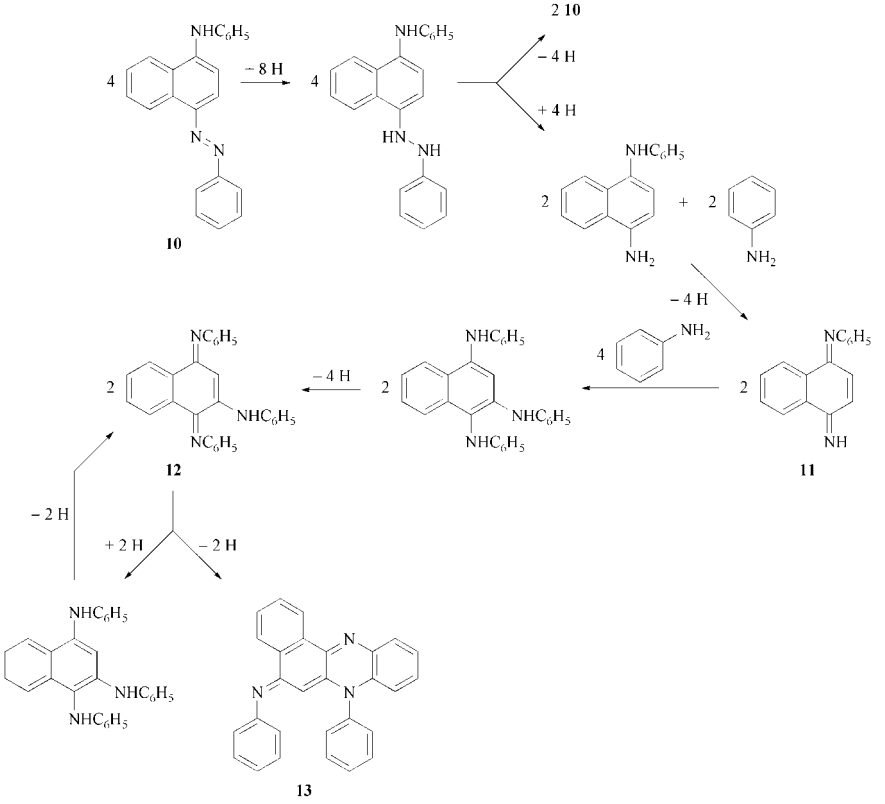
2 10
– 8 H
– 4 H
+ 4 H
10
– 4 H
11
– 4 H
12
– 2 H
+ 2 H
– 2 H
13

as intermediates, which are cyclized via the indamine **5** to the thiazine dye **5**. This reaction sequence was elucidated as early as 1885 by BERNTHSEN for Methylene Blue.

In the preparation procedures discussed so far, the indamine dye that is required as intermediate is always prepared by oxidation of 1,4-diamines. It can also be obtained by the condensation of nitroso derivatives of aromatic amines in acidic alcoholic solution. The reaction is especially used for the preparation of oxazine dyes. 3-Hydroxyanilines are used as aromatic amines that contain the ring-closing oxygen atom [7]. When the nitroso compound contains an alkoxyl group, cyclization can also take place with elimination of alcohol. This is the case for Basic Pure Blue G **6** [8], which is especially important industrially.

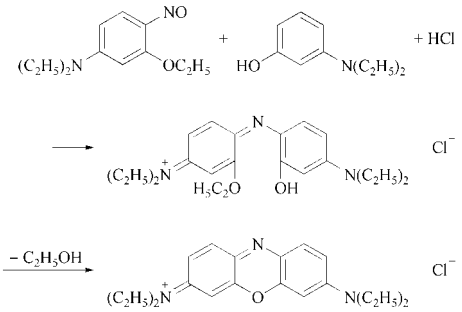

16

Azine dyes with hydroxyl groups as auxochromes are prepared analogously by cyclization of indophenols **7**, which are obtained by the action of aromatic amines on aromatic hydroxy compounds:

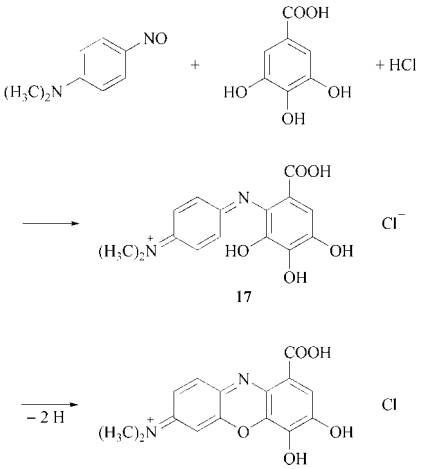

17

In dyes that contain only one auxochrome group, i.e., in which $n = 0$ in the general formula **1**, a second auxochrome group D^2 can be introduced if, instead of this group, a nucleophilically replaceable group is present, e.g., the sulfo group [9]:

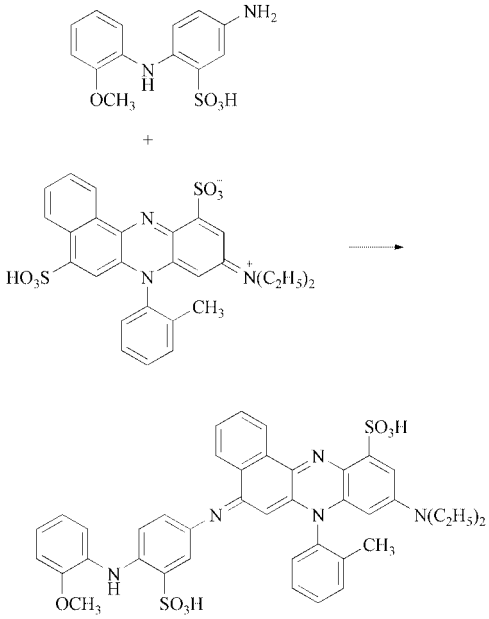

In the absence of such a leaving group, the introduction of the auxochromic group is carried out successfully by the use of oxidizing agents. Starting from phenazinium salts monoaminophenazine dyes [10] and even diaminophenazine dyes can be obtained in a similar way [11]:

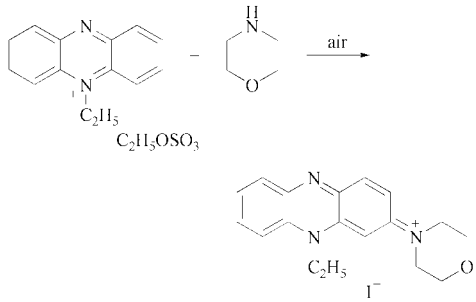

Starting from 10*H*-phenothiazine two (substituted) anilino groups can be introduced as auxochromes by using iodine as oxidizing agent yielding 3,7-bis(arylamino)phenothiazine dyes [12].

History. The oldest synthetic organic dyes belong to the group of azine dyes. In 1834, Runge found that aniline salt yielded dark green pigments on oxidation with dichromate or other oxidizing agents. Lightfood studied the oxidation of aniline in the presence of copper salt as oxygen-transfer agent and thus found in 1863 the first industrially useful procedure for the preparation of Aniline Black [*13007-86-8*].

In 1856, Perkin discovered mauveine **10** [13], which imparted to silk violet hues having a beauty so far unknown; it was also used to print postage stamps:

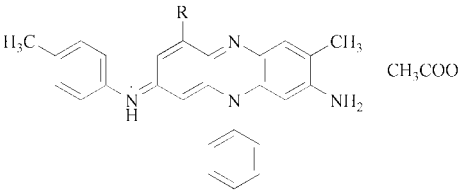

mixture with R = H, CH_3 **18**

The discovery of the induline dyes by Caro and Dale followed in 1863, and the discovery of the nigrosines by Coupier in 1867. Both groups of dyes represent extremely complex structural mixtures of phenazine dyes.

The first thiazine dye, named after its discoverer, was Lauth's Violet [*581-64-6*], which was obtained in 1876 by joint oxidation of 1,4-phenylenediamine and hydrogen sulfide with iron(III) chloride. However, the dye class achieved industrial importance only when CARO discovered Methylene Blue that same year at BASF.

Meldola's Blue [*7057-57-0*], the condensation product of nitrosodimethylaniline and 2-naphthol, was described in 1879; it is the oldest industrially produced oxazine dye. In 1928, Kränzlein, Greune, and Thiele discovered the group of dioxazine dyes by reaction of aromatic amines with chloranil followed by cyclization, the triphendioxazine base structure itself being known since 1890.

2. Phenazine Dyes

The basic aminophenazine dyes are suitable for dyeing wool, silk, and leather directly, and cotton mordanted with tannin or Katanol. They are only still of importance as leather dyes or for the dyeing of paper pulp. The basic aminophenazine dyes also include the nigrosine and the induline bases. The former are still widely used today to color phenolic resin moldings, typewriter ribbons, shoe polishes, and printing inks, as well as for the spin dyeing of polyacrylonitrile fibers.

The dyes that contain sulfo groups exhaust from neutral or acidic bath onto wool. They level satisfactorily; some are distinguished by good fastness to washing and to fulling and by adequate fastness to light. The acidic nigrosine and induline dyes are suitable for the dyeing of paper and leather.

The oxidation dyes are also classed with the phenazine dyes. They are formed when aromatic mono- and diamines applied to human hair or animal hairs or to cotton or synthetic fibers are treated with oxidizing agents.

2.1. Monoaminophenazine Dyes

Dyes that contain a naphthalene ring system are industrially useful. *Rosinduline dyes* bear the amino group on the naphtho ring, *isorosinduline dyes* on the benzo ring. Basic isorosinduline dyes, which are obtained by reaction of the nitroso derivatives of tertiary aromatic amines with *N*-alkyl- or *N*-aryl-2-naphthylamines, are suitable for the dyeing and printing of textiles containing acid groups. The dye **11**, obtained by heating 4-nitrosodiphenylamine and *N*-phenyl-2-naphthylamine in alcoholic hydrochloric acid, dyes polyacrylonitrile navy blue [14]. It can be used as a color base [15] or as a salt [16] for the spin dyeing of polymers or mixed polymers of acrylonitrile.

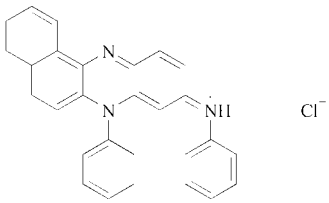

19

Basic rosindulines are proposed for the batch dyeing of polyester materials [17].

Induline Scarlet **12**, C.I. 50 080 [*2611-49-6*] (BASF, 1892), is prepared by heating 1-aminonaphthalene in phenol with the azo dye obtained from aniline and *N*-ethyl-*p*-toluidine [18].

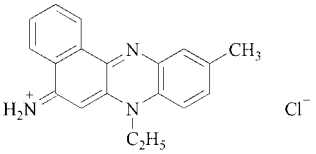

20

Acid rosinduline dyes are prepared by the action of oleum or chlorosulfonic acid on the color bases, e.g., *Azocarmin BX*, C.I. 50 090.

Acylated leuco derivatives of monoaminophenazine dyes are recommended as magenta developing components in photothermographic dry silver materials [10].

Acetone can be condensed with two equivalents of 2-amino-4-chlorodiphenylamine under oxidative conditions. Subsequent quaternization leads to **13**, which dyes polyacrylonitrile in a brilliant orange shade [19]:

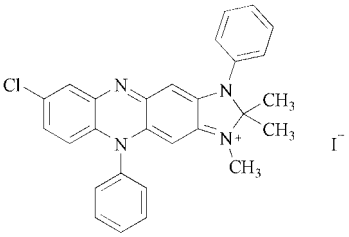

21

2.2. Diaminophenazine Dyes

The *eurhodines* and the *safranines* belong to the class of diaminophenazine dyes. The eurhodines bear a hydrogen atom or an alkyl group on the azine nitrogen atom; the safranines bear a phenyl group, which can be further substituted if necessary. If a naphthalene ring system is fused to the azine ring, the dyes are called benzophenylsafranines or, if naphthalene rings are fused to both sides, dibenzophenylsafranines. *Indulines* and *nigrosines* bear, in addition to the auxochrome groups, further phenylamino groups.

2.2.1. Eurhodine Dyes

Neutral Violet (**14**, R = H), C.I. 50 030 [*167382-91-4*], and *Neutral Red* (**14**, R = CH_3, *Toluene Red*), C.I. 50 040 [*553-24-2*] (Cassella), are seldom used these days. They are obtained by condensation of 4-nitrosodimethylaniline with 1,3-diaminobenzene or 1,3-diamino-4-methylbenzene [20].

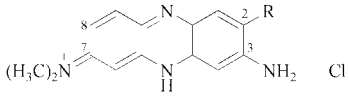

22

The formation of aggregates was investigated for Neutral Red [21]. The dye can be used for the detection of cells with phagocytic activity [22]. Dyes that bear β-chlorethyl groups instead of methyl groups have antitumor activity [23].

By reaction of Neutral Red with epichlorhydrin, the hydrogen atoms of the amino group are replaced by 2-hydroxy-3-chloropropyl groups. This compound dyes wool, polyamide, and cellulose, covalent bonds being formed [24]. In contrast, methyl halides alkylate the azine nitrogen atom and yield a dye which has improved fastness to alkali.

Dyes with additional alkyl-substituted amino groups in positions 2 and 8 show a significant bathochromic shift like indulines (see Section 2.2.4), but in contrast thereto exhibit high fluorescence and singlet oxygen quantum yields [25].

Dyes of the Neutral Red type, as well as basic safranine dyes, are suitable for fast dyeing of polyolefin materials in red, violet, and blue tints [26]. They can be used for dyeing human hair [27].

Acylated leuco derivatives of Neutral Red are disclosed for use in electrochromic recording materials [28].

2.2.2. Basic Safranine Dyes

Safranine T extra **15**, C.I. 50 240 [*477-73-6*], is offered for the dyeing of leather. Preparation

starts with the reduction of aminoazotoluene to give an equimolar mixture of 2-toluidine and 2-methyl-1,4-phenylenediamine, and is followed by oxidation in the presence of added aniline to give the azine dye.

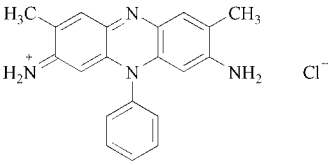

23

By the action of nitrite, it is converted to a blue diazonium salt, which on heating forms the 3,7-dihydroxy compound [29]. Safranine T can be used to remove lignin from mechanical wood pulp [30]. The dye is disclosed as colored charge-controlling agent in toners for electrophotography [31] and as photosensitizer in solar cells [32]. Safranine dyes are also suitable for use in photoelectric processes [33]. Furthermore, they can be used as desensitizers in direct-positive printing-out emulsions [34], for photopolymerization using visible light [35], as light filters with variable optical density [36], and as dye components in soluble complex-forming molecules used in affinity-specific separation of nucleic acids [37]. Safranine dyes act as redox indicators in the quantitative analysis of metal ions, e.g., ruthenium(III), antimony(III), silver, and organic substrates, e.g., ascorbic acid [38]. Oleophilic acylated leuco compounds of diaminophenazine dyes are suitable as color-forming agents in copying paper [39] and thermographic systems [40]. Basic safranines can also be used in laser technology [41] and in optical recording media with recording light having wavelengths shorter than 700 nm [42].

Monoamino- and diaminophenazine dyes bearing at least one phenyl residue on an amino group, in the form of their anhydro bases, are recommended for spin dyeing of polyacrylonitrile fibers in the gel stage [43].

Rhoduline Violet (Bayer, 1894), or *Leather Red Violet HM* **16**, C.I. 50 250 [*23279-70-1*] (Hoechst), is synthesized by reaction of 4-nitrosodimethylaniline with 3-amino-4-methyldiphenylamine.

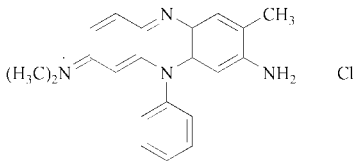

24

The amino group is diazotizable. By coupling with aromatic amines, phenols, and *N*-heterocycles, the safranine azo dyes are formed (Janus Black G, Janus Green B, Diazine Black K). The dyes can be used for dyeing polyacrylonitrile [44] and acid-modified polyester fibers [45] and as light filters with variable optical density [36].

Janus Black R **17**, C.I. 11 825 [*4443-99-6*], is formed by coupling diazotized diethylphenosafranine to phenol.

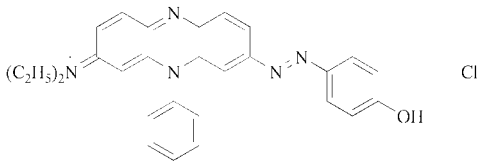

25

The dye was commercially available as Sumiacryl Black G (Sumitomo) and as Aizen Cathilon Grey BLH (Hodogaya) for the dyeing of polyacrylonitrile fibers.

By coupling the same diazo compound to *N*-methyl-*N*-cyanoethylaniline, Aizen Cathilon Navy Blue GLH (Hodogaya) is obtained, and by coupling to *N*,*N*-dicyanoethylaniline, Sumiacryl Navy Blue R (Sumitomo). These dyes have also been recommended for dyeing polyacrylonitrile.

Similar dyes are used for determination of antimony or thallium by extraction of the ion associate [46] and as thermosensitive color indicators [47]. They also serve as lustering agents in galvanic copper-plating baths [48] and as accessories in electrolytic tin plating [49]. Safranine azo dyes are also used in color photography [50]. By diazotizing Rhoduline Violet and boiling down the diazonium salt solution, polymeric phenazine dyes are obtained, which are likewise used as lustering agents in galvanic copper-plating baths [51].

2.2.3. Acid Safranine Dyes

The introduction of sulfo groups converts safranines whose amino groups are arylated into fast violet- to blue-coloring wool dyes, which are especially suitable for seawater-fast bathing suits and piece goods (women's fabrics). Today the importance of these dyes has decreased. They are synthesized from sulfo-containing aniline and 1,4-phenylenediamine derivatives by oxidation to the indamine and cyclization to the azine by using a primary aromatic amine. If 1,3-dianilinonaphthalene-8-sulfonic acid is used instead of the aniline component, an indamine is obtained that already contains the nitrogen atom required for the formation of the azine ring, so that in this case a third component is unnecessary. A further route to acid safranine dyes is the exchange of the sulfo group in the 6-position in isorosinduline di- and trisulfonic acids in neutral or weakly alkaline medium for the sulfonic acids of *N*-alkylated 1,4-phenylenediamines. The oxidative condensation of alkylated phenylenediamines with isorosinduline bases followed by sulfonation is also used.

An example for an acid safranine dye made from three components is *Acid Cyanine BF* **18** (Agfa, 1906) [51], C.I. 50 230 [*6448-97-1*].

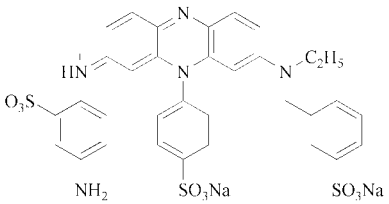

26

Preparation [52]**.** Neutral solutions containing the sodium salt of 61 kg of 4-nitro-4′-aminodiphenylamine-2-sulfonic acid in 500 L of water and the sodium salt of 57 kg of *N*-ethyl-*N*-benzylanilinesulfonic acid in 300 L water are combined in a 5000 L wooden vat equipped with a propeller stirrer. After the solution has been diluted with 1400 L of ice water, a solution containing 75.5 kg of sodium dichromate and 37.2 kg of sulfuric acid monohydrate in 4000 L water is added at 5 °C. After stirring the mixture for 15 min, a neutral solution containing the sodium salt of 30.5 kg of aniline-4-sulfonic acid in 290 L of water is added. The mixture is again stirred for 15 min, then heated to 40 °C and stirred for 15 min. After rapid heating of the mixture to 80 °C with live steam, 11 kg of sodium carbonate is added. The reaction mixture, cooled to 70 °C, is transferred to a 6000 L reduction vessel equipped with a propeller stirrer, mixed with 50 kg of iron filings, and stirred for 45 – 60 min. The alkaline spot test must show a brown rather a reddish blue spot. The reaction mixture is allowed to settle, then filtered through a filter press. The process of settling and filtration is repeated and the press is then washed, first with 1000 and then with 2000 L of water. The filtrate is salted out with 800 – 850 kg of common salt and, after stirring overnight at 18 – 20 °C, filtered under suction. After drying under vacuum, 180 – 190 kg of the dye is obtained, corresponding to 360 – 380 kg of commercial-type Acid Cyanine BF.

Other trade names were Acilan Fast Dark Blue BF and Aciderm Cyanine EZB (Bayer).

On this basis reactive dyes for cotton have been developed. Instead of aniline-4-sulfonic acid, 4-chloro- or 4-methoxyaniline-3-sulfonic acid has been used in this preparation procedure. The primary amino group is then allowed to react with cyanuric chloride, 2-methoxy-4,6-dichlorotriazine, or 2,6-dichloro-5-cyanopyrimidine [53]. The dyes obtained by reaction first with cyanuric chloride and then with 3-aminobenzenephosphonic acid or by reaction first with 2,4-dichlorotriazine and then with aliphatic or aromatic amines containing a sulfatoethylsulfonyl group are also recommended as reactive dyes for cotton [54].

Commercial products of this series were Procion Blue MX-7RX, PN PC, T-7RX (ICI).

The acid safranine dyes that are obtained by joint air oxidation of 1,3-bis(phenylamino)-naphthalene-8-sulfonic acid and 4-aminodiphenylamine-2-sulfonic acid or 1,4-diaminobenzene-2-sulfonic acid in the presence of copper oxide/ammonia still had significant sales about ten to fifteen years ago [55].

Supranol Blue BL (**19**, R = H) [*6378-88-7*] and *GL* (**19**, R = OCH_3) [*6856-08-2*], or *Wool Fast Blue BL* and *GL*, C.I. 50 315 and 50 320, were the most important dyes of this series. Of the latter about 40 t were produced in the Federal Republic of Germany in 1984.

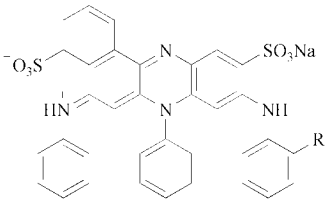

27

Preparation [55]**.** Weakly basic solutions containing 0.87 kmol of 1,3-bis(phenylamino)naphthalene-8-sulfonic acid in 1700 L of water and 0.9 kmol of 4-aminodiphenylamine-2-sulfonic acid in 1000 L of water are combined in a 10 000 L iron kettle equipped with a propeller stirrer (95 rpm), and a solution containing 35 kg of copper sulfate and 40 kg of 40 % ammonia in 130 L water is added, followed by 40 kg of ammonium chloride and 3500 L of ethanol. After dilution to 7000 L with water, air is blown through the mixture at 40 °C for 6 h. Then, the reaction mixture is allowed to cool to 25 °C and salted out with 2000 L of 25 % common salt solution. After the mixture is filtered through a suction filter under pressure, the press cake is washed twice with cold water. Yield: 500 kg = 1350 kg Supranol Blue BL, commercial type.

Other trade names were *Aciderm Blue EBL* (Bayer), *Benzyl Blue BL* [*6378-88-7*], *Orasol Blue BLN* [*52218-73-2*] (Ciba-Geigy), *Coomassie Blue BL* (ICI), and *Xylen Fast Blue BL* (Sandoz).

Supranol Blue BL can be converted to a spirit-soluble form by reaction with longer chained aliphatic amines. Such a dye salt was commercially available as *Irisol Fast Blue BLN* [*6378-88-7*] for dyeing nitrocellulose lacquers. This dye is also suitable for dyeing and printing of synthetic polyamides from organic solvents, in particular chlorinated hydrocarbons [56]. Supranol Blue GL is similarly used to prepare *Irisol Blue GLE*. The dyes can be reduced to leuco compounds, which can be used to check the packaging of foodstuffs stored under oxygen-free conditions with regard to imperviousness of the packaging [57].

Wool Fast Blue FFG **20**, C.I. 50 330 [*6837-47-4*], is prepared, analogously to Supranol Blue BL, from 1,3-bis(phenylamino)naphthalene-8-sulfonic acid and 4,4′-diamino-3′-methyldiphenylamine-2-sulfonic acid [58].

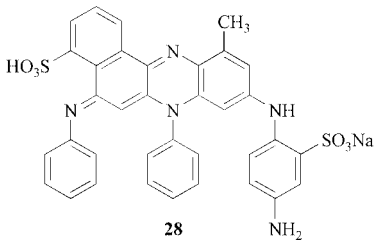

28

The primary amino group can react with cyanuric chloride [59] or with cyanuric fluoride and subsequently with aminobenzenesulfonic acid [60] to give reactive dyes which dye cotton in clear blue shades.

The condensation that takes place during the formation of *Acilan Fast Violet FB*, or *Wool Fast Violet B* **21**, C.I. 50 325 [*6837-46-3*], from 1,3-bis(phenylamino)naphthalene-8-sulfonic acid and 2-methoxy-1,4-phenylenediamine-5-sulfonic acid proceeds differently. In the cyclization step the methoxyl group is split off as methanol. Therefore, the dye can also be synthesized from 1,4-phenylenediamine-2-sulfonic acid.

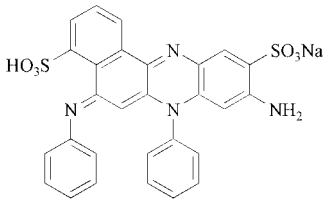

29

By reaction with longer chain aliphatic amines, the spirit-soluble dye *Irisol Fast Violet BBN* [*71839-03-7*] is obtained.

Dyes of the type of formula **19** or **21** can be used as main component in blue toners for the electrophotography [61] or as tinting dyes in fluorescent whitening.

Supranol Fast Cyanine B or *Indocyanine B* **22**, C.I. 50 335 [*6378-89-8*], (Agfa, 1902), is synthesized by the following reaction sequence: 2-phenylaminonaphthalene is condensed with 4-nitrosoethylbenzylaniline to give the indamine, which is oxidized with manganese dioxide to the isorosinduline (Blue I, **22**). Subsequent reaction with 4-aminodiethylaniline and 4-nitrosodiethylaniline gives the safranine (Blue II, **22**), which is then disulfonated to **22** [62]. The dye is used as an accessory for the galvanization of copper [63].

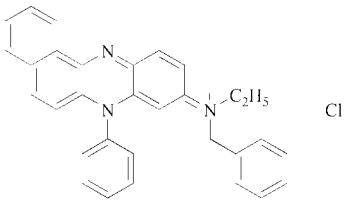

30

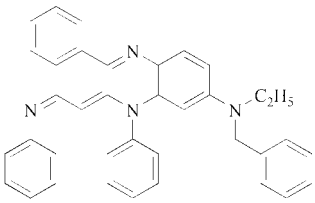

31

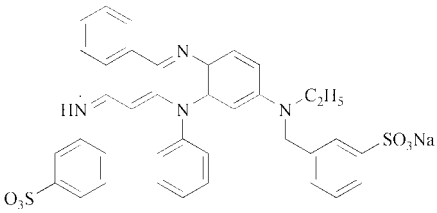

32

Fast Wool Blue EB **23**, C.I. 50 300 [*6471-74-5*], and *Supramine Blue EG* **23**, C.I. 50 310 [*6378-87-6*], are prepared according to the same method [64].

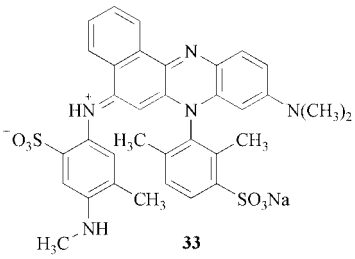

33

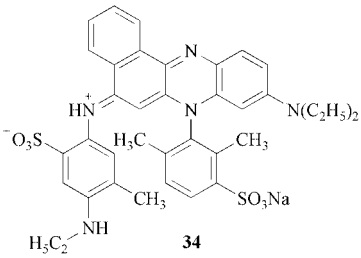

34

Supramine Blue EG was selected as Standard 4 for the eight-step ISO lightfastness test [65].

The dyes *Novazol Acid Blue GL* (**24**, R = C_2H_5) and Polar Blue G (**24**, R = $CH_2C_6H_5$), developed by Geigy, are obtained from diethylisorosinduline-1,6-disulfonic acid **24**, prepared according to [66], by exchange of the sulfonic group in the 6-position for 4-aminodiethylaniline-3-sulfonic acid and 4-aminoethylbenzylaniline-3-sulfonic acid [67], respectively. The second sulfo group in the 16-position can also be introduced by sulfonation in oleum [68]. This improves the alkali fastness and shifts the tint form a dull reddish blue to a pure greenish blue.

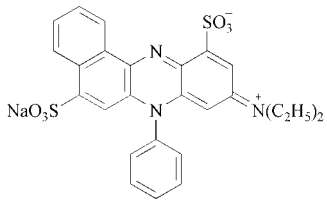

35

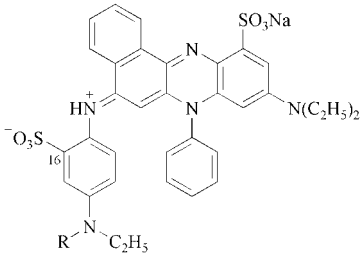

36

Acid safranine dyes, e.g., **23**, are recommended for use in inks for ink-jet printers [69]. They are useful as indicators in the quantitative analysis of metal ions such as As(III), Sb(III), Tl(I) and organic substrates like hydroquinones and ascorbic acid [70]. They have also been combined with monomeric acrylic acid derivatives or unsaturated polyesters, and deep-colored polymers have been prepared in this way [71].

2.2.4. Indulines

Induline dyes, the product of the reaction of aniline hydrochloride and aminoazobenzene, consist of a complex mixture of phenazine dyes, the composition of which continuously changes with progressive reaction. *Induline 6B* **26**, which is substituted with four aniline groups, is finally

formed. The reaction begins with the reduction of 4-aminoazobenzene to 1,4-phenylenediamine, which is oxidized to the quinone imine. Then, addition of aniline and oxidation to give the quinone alternate until 2,5-dianilinobenzoquinone-1,4-iminemonoanil **25** and -dianil (**25**; azophenine) are formed. After azine ring closure and oxidation, **25** gives *induline* (**25**; Hoechst, 1889), while **25** is converted to aminophenylaposafranine **25**.

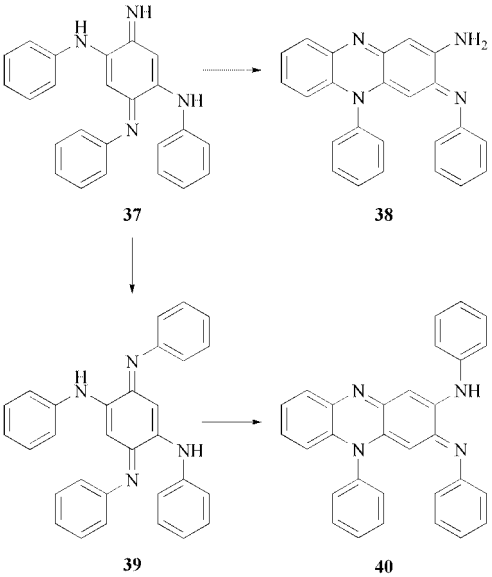

In addition, the following dyes could be identified in the induline melt: *Indamine Blue B, R*, *Nigrosine Base G* **26** (BASF), C.I. 50 204 [*6471-73-4*], *Induline Base 3B*, *Fast Blue Base RF* **26** (BASF), C.I. 50 400 [*8004-98-6*], *Induline Base N* (BASF), *Induline Base 6B* **26**, C.I. 50 400 [*8004-98-6*].

Induline Base 5BM consists of a mixture of the above-mentioned induline dyes, Induline base 6B being the main component.

Induline Base 6B was synthetically prepared by an independent route by condensation of 4,5-bis(phenylamino)-*o*-benzoquinone with 2-amino-5-chlorodiphenylamine followed by reaction with aniline hydrochloride [72].

Indulines which are soluble in benzene or spirit are synthesized by alkylation [73], e.g., with alkyl halides [74].

Induline bases are primarily used for dyeing of plastics. They are recommended as charge-controlling agents in positive toners for electrophotography [75]. Salts with oleic acid can be used in typewriter ribbons or carbon paper [76].

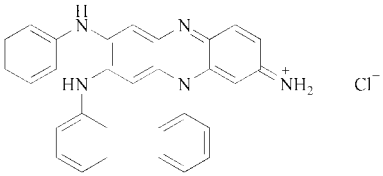

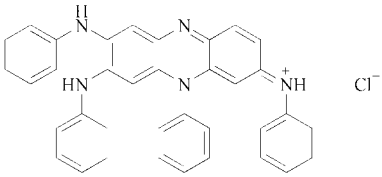

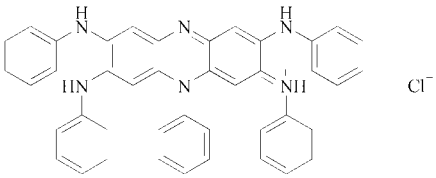

By sulfonation of the bases, acid induline dyes are obtained, which are used for the dyeing of leather. They are useful for improving the light stability of photographic paper [77] and for adjusting the contact angle of inks for ink-jet applications [78].

2.2.5. Nigrosines

Nigrosine dyes, the products of the action of nitrobenzene on aniline in the presence of iron(II) chloride and hydrochloric acid, also consist of a complex mixture of phenazine dyes. The thin-layer chromatogram shows a large number of components that are also present in induline, in particular Induline 6B **26**. In contrast to induline, however, the dye mainly consists of higher molecular mass phenazines, which remain at the starting spot of the chromatogram.

A distinction can be made between water-, ethanol-, and fat-soluble nigrosines. The ratio of aniline to nitrobenzene and the melting temperature differ only negligibly for the three kinds.

The crude melt, a mixture of nigrosine hydrochlorides, is offered as *Phenol Black L* for the dyeing of phenolic resins.

Preparation. In a 3000 L steel enamelled kettle equipped wth a stirring apparatus, an indirect heater, and a 3.0 MPa steam heat exchanger, 2000 L of aniline, 231 L of 30 % hydrochloric acid, and 132 kg of iron(II) chloride are heated over 2 – 3 h to 165 °C, and the water, together with azeotropically codistilled aniline, is distilled. Over 5 h, 588 L of nitrobenzene is introduced at 165 °C with stirring. Because of the exothermic reaction, the temperature increases to 170 – 180 °C; it is then adjusted to 186 °C over the course of 4 h, and maintained for 2 – 3 h at this temperature. The melt is subsequently diluted with 1400 L of aniline and drawn off into 4000 L of water which has been placed in a 6000 L steel enamelled kettle equipped with a stirring apparatus. The reaction kettle is washed with 100 L of aniline, and this, along with the aniline water from the distillation condenser, is added. The reaction mixture is now heated to 95 – 100 °C and stirred for 1 h at this temperature, and the charge is allowed to stand for 24 – 36 h to separate the phases. The lower layer is drawn off into a 4000 L paddle dryer; the upper aniline – water layer is fed to an aniline recovery unit. In the paddle dryer, the melt is heated in a vacuum of 27 kPa with steam of 100 kPa overpressure to 100 °C jacket temperature. After 2 – 3 h, the temperature is increased to 140 °C and maintained at this temperature for 18 h. The melt is cooled over 6 h, the vacuum lifted with nitrogen, and the paddle dryer emptied. Yield: 1150 kg Phenol Black L; 2700 kg of aniline is recovered.

The preparation of nigrosine dyes can also be carried out continuously by passing the reaction melt successively through two reaction zones at different temperatures. In the first zone, the melt is heated for 6 h at 150 – 160 °C, and in the second for 12 – 13 h at 180 – 200 °C [79]. The workup of the nigrosine melt can also be carried out continuously by mixing the aniline-containing dye melt with an aqueous medium and separating the aqueous phase from the aniline-containing phase in a suitable separation apparatus [80].

Removal of excess aniline can be carried out by grinding the melt with aqueous acid in a ball mill [81]. A pyrophoric nigrosine base can be freed of this disadvantageous property by dissolving it in aniline and introducing air for 16 h at 75 °C. Subsequently, the solution is evaporated to dryness under vacuum [82]. A nigrosine granulate is obtained if the nigrosine melt, prepared by reaction of aniline with *p-tert*-butylnitrobenzene at 180 – 185 °C, followed by treatment with caustic soda, is heated to 240 °C and extruded into cold water through a 3 mm nozzle [83]. By the addition of phthalic anhydride to the nigrosine melt, a readily dispersable, flowable form is obtained [84].

The replacement of 5 – 15 % of the aniline with *o*-nitrochlorobenzene supposedly results in better yields for the nigrosines and indulines [85].

The nigrosine bases serve for preparation of printing colors and for dyeing of plastics. They are used as broadband-absorbers in thermal optical limiters for laser application [86] and as additives in the dispersion polymerization of styrene and butyl methacrylate to yield monodisperse micrometer-size beads with narrow size distribution [87].

The preparation of water-soluble nigrosines starts from the crude nigrosine melt. In this case, the preparation is similar to that of Phenol Black L; however, before evaporation in a paddle dryer, the hydrochlorides are converted to the base by treatment with caustic soda. In order to facilitate separation of the aqueous phase, the viscous melt is first diluted with aniline.

By sulfonation of the thus-prepared Lake Black Crude, *Nigrosine WLF* and *W liq.* [*8005-03-06*] are obtained.

The dyes serve for the dyeing of paper and leather [88]. They are used as cheap darkening agent for olives [89]. Arylguanidine salts of sulfonated nigrosines, e.g., with di-*o*-tolylguanidine, afford ballpoint pen inks with black shades having high color strength. They exhibit infrared absorption [90].

Nigrosine dyes containing sulfonic acid groups can also be prepared by treating nigrosine bases with diazotized sulfanilic acid in the presence of pyridine, adjusting the pH to 10 by the addition of alkali, separating the pyridine base, and salting out the dye. This dye is used for inkjet printing [91]. Ammonium salts of the nigrosine sulfonic acids serve for the preparation of waterproof flexographic inks [92]. Solutions of

acid nigrosine dyes containing a polyvinylpyrrolidone resin are aqueous printing colors for ink-jet printing [93].

The ethanol-soluble nigrosines are the purified hydrochlorides of the nigrosine bases. Complete removal of the iron salts is a determining factor for the solubility in spirit. Important commercial products are *Brilliant Spirit Black RM*, *Lake Black H* (Lanxess), and *Basonyl Black X22* (BASF).

Preparation of Brilliant Spirit Black RM New. In a 1000 L cast-iron melting kettle equipped with a gate agitator, a condenser, and a heating coil (2.5 MPa steam), 570 kg of aniline, 182 kg of 30 % hydrochloric acid, 16 kg of dry iron(II) chloride, and 16 kg of iron turnings are slowly heated to 120 – 125 °C with stirring until the water is distilled off. Then, the reaction mixture is rapidly heated to 180 °C, 180 kg of nitrobenzene is added in a thin stream over 4 – 5 h at this temperature, and the mixture is maintained for an additional 3 h at 180 – 185 °C inner temperature. The reaction water together with some nitrobenzene and aniline is distilled off through a descending condenser. The nitrobenzene – aniline mixture, which separates from the water, is returned for reuse in the reaction process. After completion of the melting, the contents of the kettle are transferred to a 7000 L stone-lined cast-iron kettle equipped with a mechanical stirrer and containing 1500 L of water and 500 L of 30 % hydrochloric acid. The charge is stirred for 8 h at 95 °C and pressed hot in filter presses; the press cake is washed with water to remove the aniline, boiled with 1500 L of water and 75 kg of 30 % hydrochloric acid for 2 h, pressed out, and washed to neutrality. After drying under vacuum at 90 °C, 330 kg of Brilliant Spirit Black RM New is obtained. By neutralization of the salt liquor with lime followed by steam distillation, 340 kg of aniline is recovered.

An ethanol-soluble nigrosine granulate is prepared by first treating the washed, wet nigrosine paste with a paraffin wax melt in the presence of nonionic surfactants, then drying and grinding [94].

The fat-soluble brands consist of the nigrosine bases that have been obtained from the hydrochlorides by treatment with soda. Commercial products are *Nigrosine Base BA* (Lanxess).

By heating the nigrosine bases with fatty acids such as stearic acid or oleic acid, fat-soluble black brands are obtained. Important commercial products are *Nigrosine Black BA liq.*, *Nigrosine Black A liq.* (Lanxess), *Neptune Black X14*, *X16* (BASF).

Such dyes are useful for ink-jet application [95]. Salts of organic acids or acylated nigrosines are suitable as developers for electrostatic copying processes [96] and as colored charge-controlling agents in positive toners for electrophotography [97].

Nigrosine salts of long-chain alkyl sulfonic acids, prepared by reaction of nigrosine bases with these acids, are used for oil-soluble inks [98] or in microencapsulated form for pressure-sensitive copying material, which can be read in the infrared [99].

Induline and nigrosine dyes containing sulfonamide groups are prepared by reaction of induline and nigrosine sulfonyl chlorides with primary and/or secondary aliphatic or cycloaliphatic amines. They are suitable for the preparation of lack and printing dyes, ballpoint pen pastes, and mimeographic colors [100].

Spirit-soluble nigrosines, which are especially valuable for the preparation of inks and ballpoint pen pastes, are obtained by reaction of nigrosine bases with propylene oxide or butylene oxide [101].

2.2.6. Aniline Black

Aniline Black [*13007-86-8*] is the term for a group of dyes that are formed by oxidation of aniline, *o*-toluidine, *m*-toluidine, *p*-phenylenediamine, 4-aminodiphenylamine, and similar bases of the benzene series. The involved structures were the subject of several studies. It is believed that, starting from indamine structures such as emeraldine **27**, further oxidation and condensation with aniline form phenazine structures like **27** and **27** and finally insoluble polymeric compounds [102]. Aniline Black is mostly produced on the fiber itself in printing and in dyeing; for example, on cotton as well as on synthetic fibers, especially polyester fibers. In addition, for the dyeing of lacquers, plastics, and paper Aniline Black is produced in substance.

For the one-bath dyeing of cotton yarn, the fibrous material is treated with a solution of

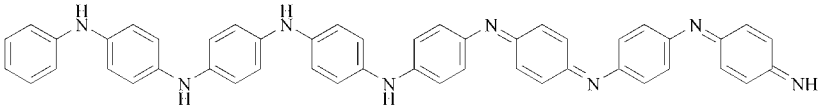

44

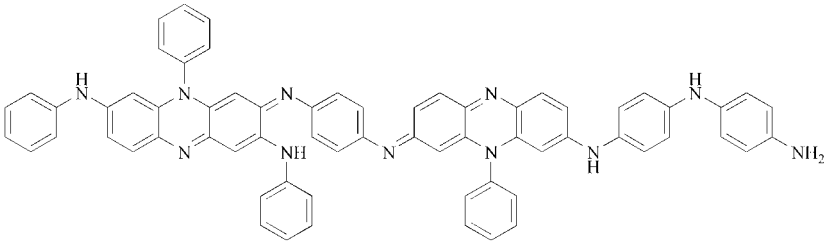

45

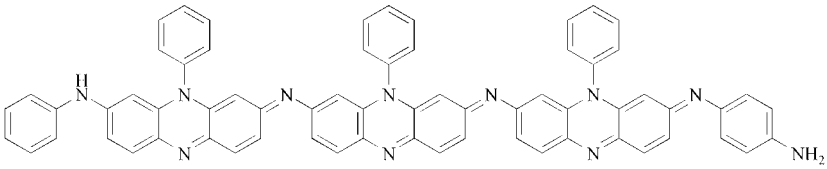

46

dichromate, aniline, hydrochloric acid, sulfuric acid, and copper sulfate, first in the cold, then at 70 °C. Finally, the material is washed and soaped. This black color greens under the influence of reducing agents. According to another dyeing process, the cotton fabric is placed in a bath containing aniline, aniline hydrochloride, sodium chlorate, ammonium chloride (as acid source), aluminum acetate, and a copper or vanadium salt as catalyst. The fabric is padded, dried, steamed, then possibly treated with chromium salts to complete the oxidation, and finally soaped. The colors thus obtained possess better fastness properties. To protect the fibers from damage by the mineral acids present, nonprecipitable proteins, urea, and dextrose have been recommended [103]. The general dyeing procedure is described in [104]. Improvements have been made by avoiding the steaming step, which results in energy conservation and reduction of air pollution by aniline vapor [105].

The production of black tints on polyester materials can be achieved with soluble derivatives of 4-aminodiphenylamine, e.g., azomethines of sulfobenzaldehydes (e.g., the former *Resolin Black Base*, Bayer), which are oxidized by quinonedichloroimide (e.g., the former *Resolin Black Developer*, Bayer) [106].

Helio Fast Black TW, a color-strong Aniline Black pigment, is obtained by replacing part of the aniline with *o*-toluidine. The pigment contains a significant amount of chromium, which comes from the oxidizing agent sodium dichromate; the amount can be decreased by stirring with 3 % hydrochloric acid.

Preparation. 3750 L of water and 436 L of 30 % hydrochloric acid are placed in a 8000 L enamelled kettle equipped with a mechanical stirrer, and 150 kg of aniline and 52 kg of *o*-toluidine are then added. With further stirring, 78 kg of sulfuric acid monohydrate is added, and the contents of the kettle are heated to exactly 27 °C. Starting at this temperature, a 24.5 % solution of sodium dichromate is added in several portions, and the solution is stirred each time for 15 min before the next amount follows. The temperature increases to 29 °C with the addition of 318 kg of sodium dichromate solution; further addition of three times 273 kg raises the temperature to 33, 39, and 45 °C. Then the pigment is immediately isolated in a filter press, washed with 3000 L of water, and dried by blowing. The dried pigment is mixed with 6000 L of deionized water and 500 L of 30 % hydrochloric acid and stirred overnight at 25 °C. The press cake is

washed with 8000 L of distilled water, and the pH of the mixture is raised to 5 with 5 kg of sodium carbonate. The isolated product is dried under vacuum at 90 – 95 °C. Yield: 280 kg Helio Fast Black TW.

An ash-poor Aniline Black is obtained by oxidizing aniline with potassium dichromate in the presence of copper(II) sulfate and sulfuric acid; the dichromate and sulfuric acid are added in three portions over 5 h at 50 °C [107]. The oxidation can also be effected with sodium hypochlorite instead of potassium dichromate [108]. Finally, a chromium-free Aniline Black can also be obtained by oxidation, first with H_2O_2 in the presence of copper sulfate, and subsequently with atmospheric oxygen [109].

A black pigment having a similar constitution is formed by oxidation of *o*-ethylaniline with sodium chlorate in acidic medium in the presence of copper sulfate [110].

Aniline Black is used in dry toners for electrophotographic laser printers [111], in pigment inks for ink jet printers [112], and bound to a protein in the production of an antiserum for cancer therapy in warm-blooded animals. Aniline Black is an electrical conductor with holes as the main charge carriers [113]. It can act as the cathode material of rechargeable lithium batteries [114].

2.2.7. Oxidation Dyes

Oxidation dyes are organic compounds that are converted into the dyeing substance on the substrate being dyed. They are used for dyeing furs because with these products the natural character of fur pelts can be especially well imitated or improved on. Today they are increasingly being displaced by acid dyes, but especially the luster and coverage of oxidation dyes are still superior.

In their fundamental patents Monnet in 1883 [115] and Erdmann in 1888 [116] made known a process for dyeing hair and feathers with aqueous or alcoholic solutions of aromatic diamines in which the dyeing substances, after being absorbed by the material, are oxidized with atmospheric oxygen, hydrogen peroxide, iron chloride, permanganates, hypochlorites, or dichromates. Because the small dye precursors penetrate deep into the hair and are oxidized there to larger colored compounds, the coloring is extremely durable to light and washing. The Ursol dyes, developed by Agfa, by Farbwerke Hoechst, and later by IG Farbenindustrie, were produced in the past, e.g., by BASF and are now produced by Dohmen. They also contain, besides diamines, aminophenols and their nitro and chloro derivatives as well as their salts with hydrochloric and sulfuric acids.

Fur Dyes. The most important dye is *Ursol D* or *Benzofur D* [*106-50-3*], 1,4-phenylenediamine. Other brand names were or are, e.g., Pelagol D, Pelagol Grey D, Peltol D, Tertral D, Durafur Black R, Fouramine D, Furro D, Futramine D, Nako D, Developer PF. On oxidation, it is converted via the quinoneimine into Bandrowski's Base **28**, X = NH), which, on further oxidation, forms phenazine dyes [117].

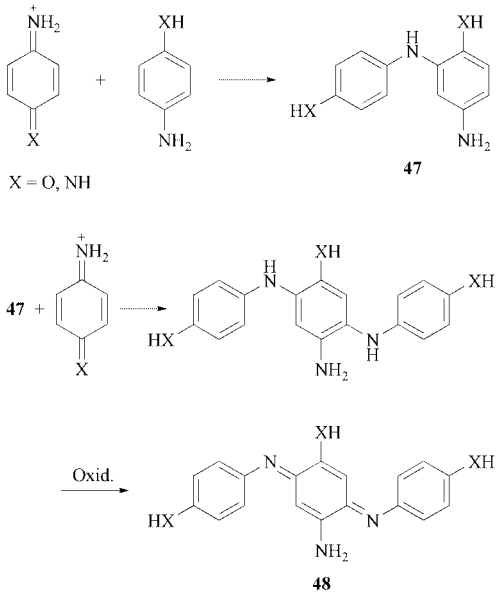

Ursol D is also present as a component in numerous other Ursol dyes, e.g., *Ursol A, DF, DG, NZ, NZD, RH, RHG, SK*, and *SKG*.

p-Aminophenol is commercially available as *Ursol P* or *Benzofur P* [*123-30-8*]. On oxidation, it forms first the quinonemonoimine, which, following an analogous path as described above, yields hydroxy compound **28** (X = O), which is analogous to Bandrowski's Base. This is converted into oxazine compounds on further oxidation [118]; cf. [4].

Some Ursol dyes, such as Ursol EG (*m*-aminophenol) [*591-27-5*], Ursol EWG (chlororesor-

cinol and resorcinol), Ursol ERN (1-naphthol) [*90-15-3*], and Ursol SLA (2,4-diaminoanisole sulfate) [*6219-67-6*], have inherently only a small dyeing ability, but in combination with other Ursol dyes effect a deepening as well as a shifting of the color shade; e.g., Ursol EG shifts gray to blue and black to olive brown. Only with such or similar combinations are pure blue or red shades available [119]. Substituents in the 4-position in such *m*-disubstituted benzenes stop further condensation on the level of indamines or indophenols **29**. Fur dyeing with simultaneous silvering of hair ends is achieved by treatment with metal salts, 4-amino-2-methoxydiphenylamine, and a condensation product of formaldehyde with resorcinol and cresolsulfonic acid [120].

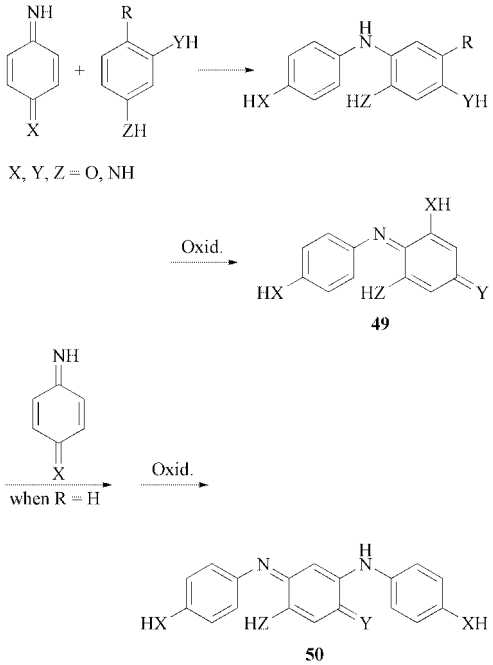

The *Ursatin dyes*, introduced commercially in 1935 by IG Wolfen, can also be used to effect combinations of shades. These are the salts of dihydroxynaphthalenes with aromatic or heterocyclic amines. *Ursatin Brown B* has, e.g., the following composition: 40 % of the salt consists of 1,5-dihydroxynaphthalene and toluylenediamine, and 60 % consists of a mixture of the sodium salts of the sulfonic acids obtained by sulfonation of coconut oil monoglyceric esters and naphthalene with fuming sulfuric acid (10 % SO_3). Important Ursatin dyes are *Ursatin Brown B, Yellowish Brown G, Blue RN, Violet B*, and *Red BB*.

Other fur assortments were *Nako* (Hoechst), *Durafur* (ICI), *Fourramine* (Francolor), *Eukamine* (Geigy).

Cosmetic Hair Dyes. For cosmetic hair dyes, the oxidation dyes are commercially available in the form of solutions or pastes. Ursol dyes used for this purpose include *Ursol D*, *Ursol 4G* (1-nitro-3,4-diaminobenzene hydrochloride), *Ursol 4R wet* (picramic acid), *Ursol 2G* (2,4-toluylenediamine), *Ursol P-Base, Ursol Brown RR* (2-nitro-1,4-phenylenediamine) [*5307-14-2*], *Ursol EG*, and *Ursol EWG*. For a historical review, see [121].

Azine and oxazine dyes are obtained on oxidation of equivalent amounts of diaminobenzenes or aminophenols and dihydroxybenzenes with hydrogen peroxide, the skin-irritating Bandrowski's Bases being avoided. Example: 2-methyl-1,4-phenylenediamine and resorcinol [122]. By preoxidation of a molar mixture of a 1,4-diamine (e.g., 2,5-toluylenediamine) and either a 1,3-diamine (e.g., 2,4-diaminoanisol), a 1,3-aminophenol, or a 1,3-dihydroxybenzene with 1 – 2 mol of O_2, a colored preparation is obtained [123] that produces permanent hair colors [124].

Commercial products of this series were *Ursol NZS* and *Ursol RHS* (BASF).

Aminophenols [125] and diaminophenol ethers [126] that bear a methyl group on the aromatic ring can be used as coupling components in hair coloring agents. 1-Methyl-4-isopropyl-2-hydroxybenzene is also recommended as a coupling substance [127]. The tolerance of skin toward the couplers based on phenylenediamine is improved by the introduction of alkyl ester groups into the benzene ring [128] or into the amino group [129] or by alkylation of the amino group [130]. The introduction of sulfo groups serves the same purpose. *N*,*N*-bis(2-hydroxyethyl)-*p*-phenylenediamine is used as a partial replacement for *p*-phenylenediamine [131]. Aminonaphtholmonosulfonic and -disulfonic acids are used as couplers [132].

4-Amino-3-methylpyrazolone hydrochloride and its *N*-substituted derivatives [133] as well as 3-aminopyrazoline [134] are physiologically compatible heterocyclic amines for dyeing hair.

An example of a heterocyclic hydroxy compound is 6-hydroxyquinoline [135]. 2,4,5,6-tetraminopyrimidine as developer and 2,6-diaminopyridine as coupler yield shades with high vividness and excellent fastness [136]. The same is true of hydroxyaminopyrimidines, e.g., 2-hydroxy-4,5,6-triaminopyrimidine [137], and diaminopyrazoles [138], which additionally have excellent physiological tolerance properties. 6-Amino-2,2-difluoro-1,3-benzodioxole is a coupler with improved fastness to washing and light and good substantivity [139].

Also suitable for hair dyeing are indamines, indoanilines, and indophenols, the initial reaction products of the hair dyes described above. Indamine salts are formed by the joint oxidation of aromatic diamines and aminohydroxy compounds [140]. Indamines with ring-cyclized amino groups are obtained by the use of 6-hydroxytetrahydrobenzoxazine [141]. The joint oxidation of phenylenediamine and phenols by ammonium persulfate affords the indaniline dyes [142]. Their leuco compounds, the aminohydroxydiphenylamines, are also suitable as hair dyes [143]. Indophenols are formed by the joint oxidation of aminophenols and phenols [144].

In the 1980s several compounds for cosmetic hair dyes have been banned in Europe for toxicological reasons, including 2,4-diaminotoluene and 2,4- and 2,5-diaminoanisole. Substitutes have been found, e.g., 2,4-diaminophenoxyethanol [145], 2-amino-4-β-hydroxyethylaminoanisole [146], 1,3-bis(2,4-diaminophenoxy)propane [147], 2-β-hydroxyethyl-*p*-phenylenediamine [148], and 3,4-diaminophenols bearing aminoalkyl or mono-, di-, or trihydroxyalkyl substituents on the N and O atoms [149].

A new study shows a significant increase of bladder cancer caused by the repeated use of permanent (oxidation) hair dyes over several years [150].

Because the use of hydrogen peroxide at high pH damages the hair and irritates the scalp, procedures using oxidative enzymes instead of peroxides have been developed which work at neutral pH values [151]. Oxidoreductase enzymes are recommended for the same reasons [152].

2.3. Aminohydroxy- and Monohydroxyphenazine Dyes

Phenazine dyes with an amino and a hydroxyl group as auxochromes are virtually completely present in the phenazone structure (**1**, where D^2 is =O). They are obtained by reaction of nitrosophenols with *m*-toluylenediamines that are alkylated in the position *para* to the methyl group to give indophenols, followed by oxidation to give the azine dye. Dyes of this kind having a naphthalene ring system are obtained if the isorosinduline dyes, prepared from alkyl- or aryl-2-naphthylamines and nitrosodialkylanilines, are treated with alkali under the action of oxygen. None of these dyes has industrial importance any more. However, phenazone compounds have been suggested for the delignification of lignocellulose materials [153]. Their use in recording materials has also been described [154].

3. Oxazine Dyes

The cationic mono- and diaminophenoxazine dyes are widely used to dye acid-modified synthetic fibers, especially polyacrylonitrile fibers containing acid groups [155]. Aminophenoxazine dyes are also suggested for dyeing polyolefin materials [156]. The earlier areas of application, the dyeing of tannin-mordanted cotton, silk, leather, and lacquers, are now only of minor importance.

The aminohydroxyphenoxazine dyes include the gallocyanines, which form light-stable metal complexes with chromium. They were mainly used for the dyeing and printing of wool, but also for the dyeing and printing of cotton; however, they have been replaced by acid anthraquinone dyes and by reactive dyes.

Diaminophenoxazine dyes have recently found application in laser technology. They complement the emission spectrum of rhodamine on the long-wavelength side [157].

Phenoxazine dyes function as desensitizers [158] and are used in optical recording media with recording light having wavelengths shorter than 700 nm [42]. Leuco phenoxazine dyes are used in acylated form as chromogenic substances in pressure-sensitive copying paper. In a photochemical oxidation step, an acyl group is

split off and blue dyes are formed. The benzoyl derivatives are preferably used [159]. However, the acyl moieties of aliphatic carboxylic acids are also used [160]. Suitable acylating agents include heterocyclic carboxylic acids derived from pyridine, furan, or thiophene [161]. These leuco derivatives can be used in electrically developable recording materials [162] and in dye transfer printing processes [163].

Aminophenoxazine dyes [164] and hydroxyphenoxazine dyes [165] are also suitable in the form of their leuco compounds as highly sensitive oxygen detectors. The dyes are also used for the analytical determination of ascorbic acid [166] and as anodic electrochromic materials in electrochromic devices such as rearview mirrors, sun roofs, or windows [167].

Furthermore, diaminophenoxazine dyes are used for the staining of preparations in microscopy. They have recently also become of interest in chemotherapy. Compounds such as **30** are thought to inhibit tumor growth and to be effective against tubercular bacilli. The activity of several dyes exceeds even that of streptomycin [168].

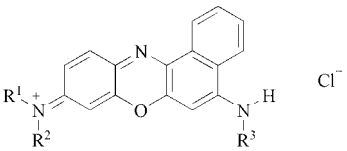

51

$R^1 = R^2 = H, CH_3$
R^3 = Benzyl, 4-Chlorophenyl, H

3.1. Monoaminophenoxazine Dyes

Monoaminophenoxazine dyes are synthesized by reaction of nitrosodialkylanilines with aromatic hydroxy compounds, in particular compounds of the naphthalene series.

Fast New Blue 3R **31**, C.I. 51 175, first described in 1879, is the reaction product of 2-naphthol with excess 4-nitrosodimethylaniline, which functions simultaneously as oxidizing agent. It is used for dyeing polyacrylonitrile and, mixed with brown or yellow basic dyes, for the production of black tints on leather.

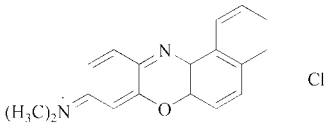

52

Preparation. 100 kg of dimethylaniline, dissolved in 260 kg of 30 % hydrochloric acid, is nitrosated with 58.4 kg of sodium nitrite in the form of a 40 % solution at 3 – 5 °C over 8 – 10 h. The reaction mixture is stirred for 1 h and allowed to stand overnight. It is then suction filtered and the product is washed with a dilute solution of common salt and finally with 200 L of ethanol. The nitroso compound, wet with ethanol, is made into a paste with 77 kg of zinc chloride and 250 – 300 L of ethanol and during 90 – 95 min added in 35 – 40 portions to a boiling solution containing 90 kg of 2-hydroxynaphthalene in 400 L of ethanol. The product is filtered hot through a closed suction filter under pressure, washed with a dilute solution of common salt containing zinc chloride, dried by suction, and made into a paste in a mixer with 13.2 kg of dextrin (to bring to commercial standard) and 5.4 kg of machine oil (to free from dust), and dried for 90 h at 55 – 60 °C.

Yield: 185 – 192 kg of commercial product. Moreover, the intermediate isolation of the nitroso compound can be eliminated [169].

The dye was offered as *Leather Dark Blue H3R* (Hoechst) and, for the dyeing of polyacrylonitrile, in dye mixtures. Furthermore, it is reported to be an intermediate hydrogen carrier for the determination of glucose-6-phosphate dehydrogenase [170] and a colored charge-control agent in toners for electrophotography [171]. The dye synthesized in a similar way from nitrosodiethylaniline is used in a solid laser composition [172].

Blue to black dyes suitable for dyeing cellulose acetate are obtained by condensation of nitrosodimethyl- or diethylaniline hydrochloride with 1,5-dihydroxynaphthalene in aqueous alcoholic solution containing at least 30 % water, preferably with the addition of manganese dioxide as catalyst [173]. The presence of water is necessary; otherwise, gray dyes are formed.

The position *para* to the nitrogen atom in the naphthalene ring is activated. By reaction

with amines, entry is achieved into the series of diaminophenoxazine dyes (see Section 3.2). By reaction with Michler's hydrol [bis(4-dimethylaminophenyl) methanol] *Fast New Blue R* (Bayer, 1892) is obtained. Sulfur-containing substituents can also be introduced into this position; e.g., Fast New Blue 3R can be treated with arylsulfinic acids, thiophenol, thiosalicylic acids, thioacetic acid, or thiourea. The initially formed leuco compounds are oxidized with iron chloride or nitric acid [174].

3.2. Diaminophenoxazine Dyes

Diaminophenoxazine dyes are synthesized by reaction of nitrosodialkylanilines or nitrosodialkylaminophenols with aromatic amines or aminophenols in alcoholic solutions or mineral acids. They are characterized by very clear hues.

Rhodulin Pure Blue 3G **32** (Bayer, 1916), C.I. 51 004 [*444-00-2*], is obtained by the action of 4-nitroso-1-diethylamino-3-ethoxybenzene on 3-diethylaminophenol in alcoholic hydrochloric acid, whereby the ethoxyl group is cleaved off.

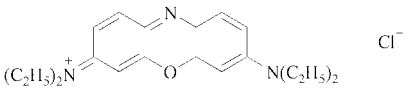

53

Rhodulin Pure Blue 3G, also called *Oxazine 725*, is at present the most important cationic oxazine dye and is represented in all large assortments of polyacrylonitrile dyes. Examples are *Aizen Cathilon Blue 5G* (Hodogaya), *Astrazon Blue BG* (DyStar), *Diacryl Brilliant Blue 5G-E* (Mitsubishi), *Sumiacryl Blue E-6G* (Sumitomo), *Maxilon Blue 5G* (Ciba), *Sevron Blue 5G* (Crompton & Knowles), and *Yoracryl Blue 5G* (Yorkshire).

A similar product was *Basacryl Blue MX-8GF* (BASF).

According to a newer preparation procedure for Rhodulin Pure Blue 3G, the intermediate isolation of the nitroso compound can be dispensed with, and the dye is converted into the hydrochloride by precipitation of the zinc as its carbonate.

Preparation [175]**.** 114 kg of a 40 % aqueous solution of sodium nitrite is introduced into a solution containing 125 kg of *N,N*-diethyl-*m*-aminophenetidine in 150 kg of ethanol and 168 kg of 30 % hydrochloric acid at 0 – 3 °C. After the solution has been stirred for 1 h, 7.5 kg of soda is added to raise the pH to 2 – 4. The suspension obtained is added in small portions to a boiling solution containing 100 kg of *m*-diethylaminophenol in 300 kg of ethanol. When the reaction is complete, the mixture is cooled to 50 °C, 50 kg of zinc chloride is added with stirring, the mixture is allowed to cool to 20 °C, and the dye is filtered off under suction in the form of the zinc chloride double salt and washed with ethanol. The wet suction-filter cake is suspended in 4000 L of water, and ammonium carbonate (ca. 90 kg) is added at 60 °C until the filtered solution gives no more zinc carbonate precipitate after further addition of ammonium carbonate. The solution is filtered at 60 °C and the readily soluble dye is isolated by spray drying.

Addition of acetic anhydride during the condensation step is advantageous [176]. The nitrosation and further condensation to give the oxazine dye can also be carried out in a cyclic ether (dioxane), in sulfolane, or in an alkyl-substituted amide (dimethylformamide) [177]. If the condensation is carried out in the presence of zinc chloride and zinc oxide in methyl glycol, the dye precipitates directly as the chlorozincate in very pure form [178]. Oxazine dyes like **32** can advantageously be isolated as bromide salts because of their low solubility. These low-salt products can be used directly for dyeing or for the preparation of liquid dye formulations, e.g., by reaction with potassium acetate [179]. If the condensation is done in a mixture of a water-miscible solvent (dimethylformamide) with water or an alcohol in the presence of nitric acid, pure dyes of high color strength are obtained as nitrate salts [180]. Tetrachloroferrate salts can be isolated in pure form and are recommended to avoid contamination of wastewater with zinc salts [181]. Liquid preparations can be obtained starting from chlorozincate salts by extraction of the salts of oxazine dyes with monocarbonic acids into glycol ethers, while the metal is kept in the water phase by use of a complexing agent [182].

Instead of 3-ethoxydiethylaniline, hydroxyalkyl ethers, prepared by reaction of 3-hydroxydiethylaniline with alkylene oxides, can be used. This reaction can be carried out as the first

step of a one-pot procedure leading to the finished dye [183]. The zinc-free dye is obtained by precipitation with urea or thiourea [184].

Rhoduline Pure Blue 3G tints are characterized by a frequently desired green shade in artificial light. The dye is therefore often used in mixtures for dyeing polyacrylonitrile fibers in blue [185] and black [186] tints. As the salt of the dialkyl ester of sulfosuccinic acid, Rhodulin Pure Blue 3G is used for dyeing acid-modified polyester fibers in chlorinated hydrocarbons [187]. The dye can also be used, especially in combination with yellow and red dyes, for dyeing human hair [188] or bones and tusks of vertebrates [189]. It may be useful for the preparation of infrared-fluorescent inks [190], as sensitizer [191], or as charge-controlling agent [192] or photoreceptor [193] in electrophotography. It can also be used as laser dye, e.g., in solid-state dye lasers [194] or as dye in a mono- or multilayer optical recording media using fluorescent light for reading [195]. As a salt with surface active anions, e.g., dodecylbenzenesulfonate, the dye is used in thermal transfer printing on acceptor sheets containing ion exchangers [196].

Capri Blue GON **33**, C.I. 51 015 [*7199-02-2*] [197], is obtained by the action of nitrosodimethylaniline hydrochloride on 3-diethylamino-1-hydroxy-4-methylbenzene. The dye is commercially available as Sevron Blue NF (Crompton & Knowles) for dyeing polyacrylonitrile.

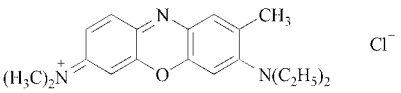

54

Diaminophenoxazine dyes whose amino groups bear a cyanoethyl group have proved to be especially valuable for dyeing polyacrylonitrile. For example, the dye prepared by condensation of 3-ethylmethylamino-4-methylphenol and 4-nitroso-*N*-ethyl-*N*-β-cyanoethylaniline hydrochloride in acetic acid and hydrochloric acid dyes polyacrylonitrile a clear blue [198]. Also worth mentioning are dyes prepared from *N*,*N*-bis(cyanoethyl)aniline [199] and from *N*-β-cyanoethyl-*N*-benzylaniline [200] according to the same procedure.

Commercial products of this series were *Remacryl Blue GGL* (Hoechst) and *Sevron Pure Blue 4G* (Du Pont).

Diaminophenoxazine dyes with primary and secondary amino groups are of particular industrial importance. Because of their low molecular mass, they migrate on polyacrylonitrile fibers and therefore result in especially level dyeing.

The dye **34** is prepared, e.g., by nitrosation of 3-hydroxydimethylaniline and condensation with 2,4-diaminoanisole [201] or with 2-methoxy-5-acetoxyaniline [202] in a polar, water-miscible solvent.

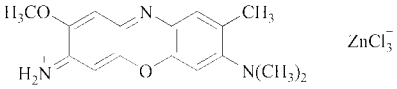

55

Another preparation procedure consists of the condensation in acetic acid of 3-hydroxydimethylaniline with the azo dye obtained by coupling diazotized *p*-sulfanilic acid to 2,5-dimethoxyaniline [203]. The indamine dye obtained by condensation of 3-methoxy-4-nitrosodimethylaniline with 2-amino-4-hydroxytoluene and precipitation as the chlorozincate complex can be isolated and subsequently converted into the oxazine dye **34** [204].

By mixing the dye with yellow and red low molecular mass cationic dyes, a wide palette of tints having high evenness is obtained [205].

Dyes have also been described that, instead of methoxyl groups, contain methyl groups in both nuclei [206]. The dye prepared from 4-nitroso-*N*,*N*-diethyl-*m*-phenetidine and 2-amino-4-hydroxytoluene is obtained by condensation in alcoholic solution in the presence of zinc powder [207]. 2-Methyl-4-nitrosomonoethylaniline, obtained by rearrangement of the corresponding nitrosamine in hydrochloric acid solution, affords, upon condensation with *m*-diethylaminophenol, a dye that colors polyacrylonitrile bright blue [208]. By joint oxidation of *p*-phenylenediamine and 2-amino-4-hydroxytoluene with $K_3[Fe(CN)_6]$ in ammonia solution, an oxazine dye is obtained that bears primary amino groups in the 3- and 6-positions [209].

Today the only commercially relevant product of this series is *Maxilon Blue M-G* (Ciba). Similar products were *Remacryl Blue B* [*51716-95-1*], *G* (Hoechst), *Remacryl Dark Blue E-GRL* [*64071-90-5*] (Hoechst), and *Maxilon Blue M-2G* (Ciba).

Oxazine dyes with cyclically bonded amino groups are suitable for dyeing polyacrylonitrile [210]. The dye **35** imparts a bright blue color to these fibers. It can also be used to dye hair [211].

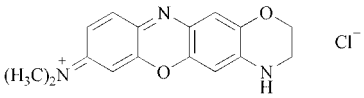

56

Oxazine dyes with both amino groups incorporated in rings are used as laser dyes because of their rigid molecular skeleton [212]. Derivatives with sulfoalkyl groups on the ring nitrogen atom are also known [213].

Furthermore, triaminooxazine dyes are known that bear amino groups in the 2-, 3-, and 6-positions [214]. Their leuco products can be acylated to a ring-closed derivative suitable for electrochromic recording materials [215].

Diaminophenoxazine dyes **36** in which one amino group is substituted by a phenyl group, and which are prepared from the correspondingly substituted diphenylamino derivatives, impart to polyacrylonitrile tints having low brightness but increased lightfastness [216].

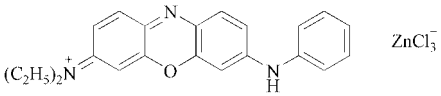

57

The dyes are suitable for transfer printing on polyacrylonitrile [217]. Their tetrafluoroborates can be used for the spin dyeing of the same fibrous material [218]. The action of alkalis on **36** produces a color base that can be used for spin dyeing [42] or transfer printing on textile material dyeable with cationic dyes, e.g., polyacrylonitrile [219], and on paper coated with an acidic layer [220]; the protonated dye is restored on the material. The color base can also be alkylated on the nitrogen atom by the action of dimethyl sulfate or ethylene oxide [221]. The same dye is obtained by starting from *N*-alkylated diphenylamino compounds [222]. Substituents in the *ortho* positions of the phenyl group, in particular methyl groups, improve the clearness of the phenylaminooxazine dyes, because the saddening effect of the phenyl group is reduced by steric hindrance [223].

A commercial product of this series is *Kayacryl Light Blue 4GSL* [*12221-43-1*] (Nippon Kayaku).

Nile Blue A **37**, C.I. 51 180 [*2381-85-3*] (BASF, 1888), is prepared by reaction of 4-nitroso-3-hydroxydiethylaniline with α-naphthylamine [224]. The dye is used in microscopy, immunoassays [225], and in laser technology [226]. It is also suitable for the prevention of deposits during the polymerization of vinyl chloride homo-, co-, and graft polymers [227].

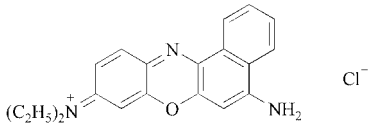

58

Nile Blue BX is the ethyl homologue of this dye.

Diaminophenoxazine dyes, above all of those with a naphthalene ring system, are also obtained by reaction of monoamino dyes, in particular *Fast New Blue 3R*, with aliphatic or aromatic amines.

Thus, *New Methylene Blue GG* **38**, C.I. 51 195 [*62770-30-3*] (Casella, 1890), is formed by oxidation of the product obtained by condensation of Fast New Blue 3R **31** and dimethylamine:

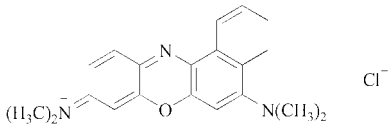

59

By condensation of 4-nitrosodimethylaniline hydrochloride with 3-hydroxydiphenylamine in a molar ratio of 2:1, a basic dye is obtained that contains both structural features of an oxazine and a phenazine dye, but not of a bis-azine dye (see Chap. 5) [228]. It dyes tannin-mordanted cotton black.

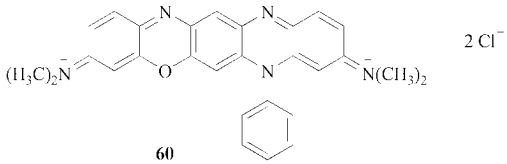

60

3.3. Aminohydroxyphenoxazine Dyes

Aminohydroxyphenoxazine dyes largely have the phenoxazone structure (**1**, where D^2 is =O). Hydroxy- and aminophenoxazones exhibit characteristic key bands between 6.0 and 6.5 mm in the IR spectrum [229]. In nature, they are present as orceine dyes in lichens [230]. Aminophenoxazone dyes are suitable for dyeing textile material made from polypropylene [231].

By the introduction of another hydroxyl group in the position neighboring the OH group that functions as an auxochrome, the dyes acquire the ability to form color lakes with chromium salts. They can be made from derivatives of gallic acid and nitrosodialkylaniline salts and are therefore known as gallocyanines. Gallocyanine dyes are used as chrome dyes on wool and cotton. Their industrial importance has decreased; however, they have been recommended for the dyeing of polyolefins in the presence of heavy-metal salts [232].

Gallocyanine, Brilliant Chrome Blue P, C.I. 51 030 [*1562-85-2*] (Sandoz, 1881), is prepared by the action of 4-nitrosodimethylaniline hydrochloride on gallic acid in methanol [233].

Coelestin Blue B **40**, C.I. 51 050 [*1562-90-9*] (Sandoz, Durand, Huguenin & Co., 1893), is correspondingly obtained by action of 4-nitrosodiethylaniline hydrochloride with gallamide [234].

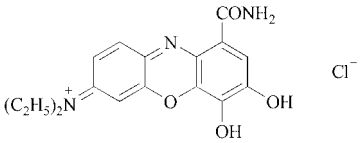

61

Preparation. In an enamelled 6000 L kettle equipped with a wooden gate stirrer, 110 kg of gallamide and 2000 L of alcohol (denatured, 94 %) are brought to a rapid boil. Over 10 h, 240 kg of nitrosodiethylaniline hydrochloride, dissolved in 1500 L of ethanol, is allowed to flow in. After addition of 21 kg of potassium carbonate, the reaction mixture is boiled for another 1 h and allowed to cool to 20 °C; 40 L of 20 % hydrochloric acid is added, and the mixture is pressed into a suction filter with a filtering stone bottom. After stirring with 500 L of 20 % hydrochloric acid, the mixture is filtered under suction until the product is well dried, and the product is further dried at 70 °C. Yield: 95 %. The ethanol is recovered by distillation.

The tetraphenylborates of dyes like **40** are recommended for use in electrophotographic toners [235].

Gallo Blue E, Prune Pure, C.I. 51 040 [*6416-51-9*] (Sandoz, 1887), is obtained by condensation of methyl gallate with the hydrochloride of nitrosodimethylaniline. The leuco compounds of these dyes were commercially available for some time as Mordant Violets; to improve their solubility, the dyes were also prepared as hydrogen sulfite compounds.

Delphin Blue **41**, C.I. 51 120 [*1323-96-2*], *Ultracyanol B* (Sandoz) is synthesized by blowing air through a hot mixture of gallocyanine hydrochloride and aniline, whereby the carboxyl group is exchanged by aniline, followed by sulfonation and conversion to the ammonium salt.

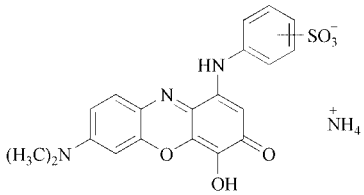

62

Chromazurin **42**, C.I. 51 125 [*1323-97-3*], *Chromazol Blue 5G* (ICI) is synthesized analogously but in the cold. Under these conditions the aniline group bonds to the position neighboring the carboxyl group, which is removed by subsequently heating to 100 °C.

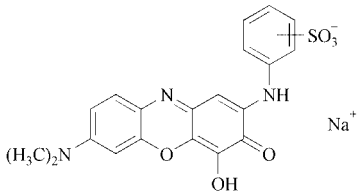

63

By heating Nile Blue A **37**, the phenoxazone dye Nile Red **43**, is obtained. It is used as a laser dye in the near-infrared (702 nm) [236]. Perfluoroalkyl groups can be introduced into dyes like **43** in *ortho* position to the oxo group by reaction with bis(perfluoroalkylcarbonyl) peroxides [237]; e.g., with bis(trifluoroacetyl) perox-

ide a trifluoromethyl-substituted dye is obtained. These dyes have increased stability and light-fastness. Dyes of this type can also be used as red fluorescent dyes in inks, fluorescent labeling agents, fluorescent collectors, and sensors [238].

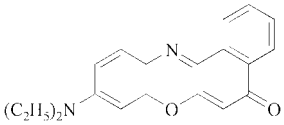

64

3.4. Mono- and Dihydroxyphenoxazine Dyes

Mono- and dihydroxyphenoxazine dyes occur in nature as orceine dyes [230]. Their use as food dyes is no longer allowed. The dyes are synthesized by reaction of nitrosophenols or nitrosonaphthols with phenols or naphthols. Another route is the reaction of 1,2-naphthoquinone-4-sulfonic acid with 1-amino-2-naphthol-6-sulfonic acid. *Alizarin Green G* **44**, C.I. 51 405 [*6492-63-3*], is thus obtained. It dyes chrome-mordanted wool a bluish-tinged green.

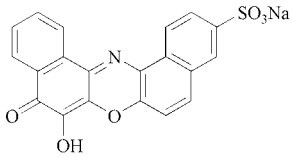

65

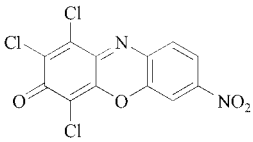

66

1,3,4-Trichloro-7-nitrophenoxazone-2 **44** [239] is a starting material for sulfur dyes. It is reduced and converted to the sulfur dye *Immedial Blue 5R* [*1327-96-4*] by polysulfide melt [240].

The dihydroxyphenoxazine dye resorufine serves as a chromogenic group for the analytical determination of a bacterial endotoxin [241]. Resorufin derivatives are used in diagnostics [242] and immunoassays [243].

4. Thiazine Dyes

The thiazine dyes are of minor importance today in the textile field because of their inferior fastness.

Silicon-containing diaminophenothiazine dyes, which can be prepared from *N,N*-dimethyl-*p*-phenylenediamine, *N*-(trimethylsilylpropyl)aniline, and thiosulfate by oxidation with alkali metal dichromate, are recommended for the dyeing of glass fibers [244].

Aminohydroxyphenothiazine dyes are present in the phenazone form (**1**, where D^2 is =S). If they contain a second hydroxyl group in the position *ortho* to the auxochromic OH group, they have mordant-dyeing properties.

Sulfur dyes that are prepared by fusion of 4-amino- and 4-nitrophenols with sulfur and sodium sulfide contain a thiazine ring system, which is, however, not in the form of the mesomeric azine system.

The thiazine dyes are used for dyeing paper and office supplies and also for the preparation of color lakes. The most important is Methylene Blue. In addition, the thiazine dyes have important photochemical applications. They are used as sensitizers in photopolymerization [245]. They are a component of a silver-free direct-positive color bleaching-out system, because they are reduced by the action of light and thereby bleach out the indigoid dyes also present [246]. A photochromic material consists of a polymer matrix, e.g., poly(vinyl acetate), a leuco phenothiazine dye, and a reducing agent. The dye is reoxidized on irradiation [247]. If the thiazine dye is bonded to a polymer chain, photo- and thermochromic polymers are obtained. The dye can be bonded to the polymer via acrylamide [248], chloromethylstyrene [249], or vinylpyridine [250]. Thiazine dyes inhibit the formation of polymer deposits in the preparation of vinyl chloride homo-, co-, and graft polymers [251]. Methylene Blue and other phenothiazine dyes are used as photoreducible dyes in galvanic cells [252]. They are also used in the leuco form as oxygen detectors for the detection of the undesired admission of air, e.g., in vacuum-packed foodstuffs [253]. Dissolved oxygen can be removed from alkaline water by treatment with leuco methylene blue [254]. Leuco dyes can also be used as anodic electrochromic materi-

als in electrochromic devices such as rearview mirrors, sun roofs, or windows [167].

The thiazine dyes have long been used for staining in medicine, bacteriology, and microscopy [255]. Besides Methylene Blue and its oxidation products Azur I and II, mixtures containing these dyes can be considered for microscopic investigations of cells, e.g., the mixture Methylene Blue – Eosine [256] or the mixture of Azur I or Azur II with Malachite Green and Fuchsine [257].

Thiazine dyes with a naphthalene ring system [258] are suitable for the differentiation of tissues in surgery and pathology, since they stain malignant tissue more strongly than normal tissue. For the diagnosis of cell samples by staining, radioactive tritium-labeled Methylene Blue can be used [259]. Thiazine dyes whose amino groups are replaced by haloalkyl groups have antimitotic activity [260].

4.1. Diaminophenothiazine Dyes

The reaction sequence that leads to the diaminophenothiazine dyes is described in Chapter 1. The most important dye, *Methylene Blue* **45**, C.I. 52 015 [*61-73-4*], is obtained by oxidation of 4-aminodimethylaniline in the presence of sodium thiosulfate to give the quinonediiminothiosulfonic acid, reaction with dimethylaniline, oxidation to the indamine, and cyclization to give the thiazine. The dye is commercially available as the hydrochloride and as the zinc chloride double salt.

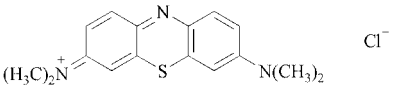

67

Preparation. To prepare the *nitroso compound*, 150 kg of dimethylaniline, 750 kg of ice, and 183 kg of sulfuric acid (78 %) are placed in a 4000 L brick-lined kettle equipped with a wooden strirrer, and the mixture is diluted with ice water to a volume of 1600 L. Then, 215 kg of 40 % sodium nitrite solution, together with 260 L of 25 % common salt solution, is allowed to flow in over 6 h at a temperature no higher than 0 °C. The mixture is subsequently stirred for 1 h and an additional 131 L of 78 % sulfuric acid is added. For the *reduction*, the charge is placed in a 2000 L kettle equipped with a heating jacket and treated at 25 – 40 °C with 1300 kg of iron filings added in 4 – 5 portions. After filtration, the solution is placed in a wooden vat having a capacity of 16 000 L.

For the *oxidation* to the quinonediiminothiosulfonic acid, the solution is first cooled to −1 °C and then 122 kg of 78 % sulfuric acid, diluted to 600 L with water, is added in 4 portions over 50 min as well as 120 kg of sodium dichromate dissolved in water (800 L solution) and 660 kg of 45 % sodium thiosulfate solution. The final temperature should be +1 °C. Then the mixture is stirred for 1 h and the temperature allowed to rise to 3 °C. For the conversion to the *indamine*, 700 L of an aqueous solution containing 255 kg of sodium dichromate is added at 3 – 4 °C, followed by 130 kg of dimethylaniline dissolved in 125 kg of 78 % sulfuric acid and 200 L of water. After 30 min, 68 kg of sodium dichromate, dissolved in 200 L of water and ice, is added again. Finally, the mixture is stirred for 45 min.

For the thiazine-ring *cyclization*, the reaction mixture is placed in a wooden vat having a capacity of 26 m^3. Then, 36 kg of sodium dichromate, dissolved in 24 L of water, and 18 kg of copper sulfate, dissolved in 100 L of water, are added, and the mixture is heated over 30 min to 90 – 92 °C by blowing in steam. Dye formation starts between 65 and 70 °C. Finally, the mixture is stirred for another 30 min at 90 – 92 °C, and the chromium sediment is then filtered off.

To prepare the *zinc chloride double salt*, the solution is treated at 45 °C with 40 L of 30 % hydrochloric acid, 300 L of 40 % zinc chloride solution, and 1000 kg of common salt solution, saturated at 45 °C. The mixture is allowed to cool to 20 °C and the product is filtered off and dried at 60 – 65 °C in a circulating-air dryer. Yield: 375 kg.

To isolate *Methylene Blue hydrochloride*, 40 L of 30 % hydrochloric acid and 2000 kg of saturated common salt solution are added to the dye solution after it has cooled to 45 °C, and the mixture is allowed to cool further to 25 °C. After filtration, the product is washed with a 2 % common salt solution.

Instead of sodium dichromate, manganese dioxide and catalytic amounts of copper sulfate

can be used for the oxidation [261]. Finally another procedure for the preparation of phenothiazine dyes has been reported in which the reaction is carried out with sodium thiosulfate in the presence of zinc chloride and aluminum sulfate, and the oxidation is carried out with sodium dichromate in the presence of copper sulfate in sulfuric acid [262].

Highly pure Methylene Blue can be obtained by extraction of impurities with chloroform from solutions of raw dye in borate puffer at pH 9.5 – 10 followed by acidification of the aqueous solution and isolation of the dye [263].

Methylene Blue is mainly used to dye paper and office supplies, but also to tone up silk colors. In addition, it is of some importance as a chemotherapeutic agent in human and animal medicine, as a stain in microscopy, as a hydrogen acceptor in enzymatic dehydrogenations, as a redox indicator in iodometry and titanometry, and as a cosmetic dye.

It was offered as *Leather Pure Blue HB* [*61-73-4*] (Hoechst) for dyeing leather and is a mixture component in Basic Leather Black brands. Methylene Blue can be adsorbed by active charcoal and then be used to remove BF_4^- anions from wastewater [264].

Methylene Blue has also been investigated as a dye in laser technology [41] and as an IR sensitizer generating $^1O_2$ useful in holographic registration material using the polymerization of acrylic monomers [265]. It sensibilizes the photochemical degradation of azo dyes with $^1O_2$ [266]. It is used as antidote against methemoglobin-forming poisons and as disinfectant in animal medicine.

10-Benzoylleucomethylene Blue [*1249-97-4*] **46** is obtained by reduction of Methylene Blue with hydrosulfite in aqueous alkaline medium, followed by benzoylation and then extraction of the crude product with toluene [267].

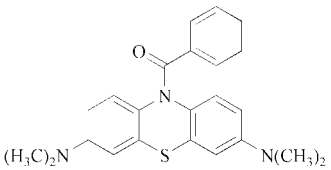

68

In place of the benzoyl group, the residues of halobenzoic acids [268], nitrobenzoic acids [269], aminobenzoic acids [270], sulfobenzoic acids [271], benzenedicarboxylic acids [272], naphthalenecarboxylic acids [273], isobutyric acid [274], as well as alkyl- and arylsulfonic acids [275] have been suggested. Reaction products with esters of chloroformic acid and isocyanates [276] can also be used. The lightfastness of the printing is improved by heteropolyacids of hexavalent molybdenum, such as phosphomolybdic acid [277].

Benzoylleucomethylene blue was used as a color-forming agent in colorless copying paper [278]. It is also recommended for documents to make them safe against forgery [279].

On contact with acid substances, such as attapulgite, montmorillonite, kaolinite, or bentonite, the color of Methylene Blue is slowly developed. The compound is also used as color-forming agent in thermoreactive paper in combination with aromatic hydroxy compounds serving as acidic substances in this case. The color reaction is effected by the action of heat, which causes the melting of the covering poly(vinyl alcohol) [280].

The mechanism of color-forming decomposition of *N*-benzoylleucomethylene blue during the copying process has been studied; the reaction is induced by photons and proceeds via free radicals [281].

Benzoylleucomethylene blue was also used in combination with other chromogenic compounds, e.g., with Crystal Violet lactone [282], rhodamineanilinolactam [283], spiro compounds [284], and fluoranes [285]. Its slow development compensates the bleaching of color formers with low lightfastness, e.g., Crystal Violet lactone. *N*-Alkyl- or arylsulfonylleucomethylene blue can be used in a dye sublimation sheet which uses nickel salicylates as developing agent on the receiver sheet [286].

A commercial product of this series was *React Blue G* (BASF).

Recording materials for electron beams contain benzoyl- or naphthoylleucomethylene blue in order to increase the sensitivity [287]. Electrochromic printing media contain *N*-benzoylleucomethylene blue and a bromide, which catalyzes the electrochemical oxidation of the leuco compound [271, 288]. Benzoyl leuco derivatives of Azur I bearing an additional benzoyl group on the methylamino group can also be used in the same manner [289]. *N*-

Acylated leucophenothiazine compounds also serve to stabilize polymeric hydrocarbons [290] and polychloroprenes [291].

Leucomethylene blue is used in oxygen detectors checking, e.g., nitrogen-blanketed or vacuum-packed perishable foodstuffs [292]. It is also recommended as oxygen-sensitive reactive material in optical data storage media such as CDs and DVDs which are intended to be usable (readable) only for a limited period of time [293].

The halogenation of Methylene Blue improves the lightfastness of the color on polyacrylonitrile [294]. The selenium analogue of Methylene Blue, prepared with ^{75}Se, is used for the imaging of epithelial corpuscle adomenas [295]. Inclusion complexation of Methylene Blue in β-cyclodextrin enhances the lightfastness even on textile fibers [296] and the fluorescence intensity by a factor of 3 – 5 [297].

Derivatives of Methylene Blue bearing, instead of one methyl group, a substituent which can react with amino, thiol, or carboxyl groups are applicable to photodynamic therapy of cancer or immunoassays utilizing chemoluminescence [298].

Methylene Blue NNX, *New Methylene Blue*, C.I. 52 030 [*6586-05-6*] (Cassella, 1881), is redder and somewhat faster to light than Methylene Blue [299]. It can be prepared analogously to Methylene Blue, starting from *N*-ethyl-2,5-diaminotoluene and *N*-ethyl-*o*-toluidine, and is suitable as a stain for microscopy.

Toluidine Blue O **47**, C.I. 52 040 [*6586-04-5*] (MLB, BASF, 1888), is prepared from *N*,*N*-dimethyl-*p*-phenylenediamine and *o*-toluidine and is used as a stain for microscopy [300].

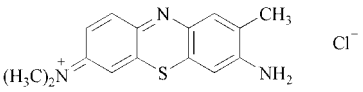

69

Methylene Green **48**, C.I. 52 020 [*6722-15-2*] (MLB, 1886), is obtained by action of nitric acid on Methylene Blue.

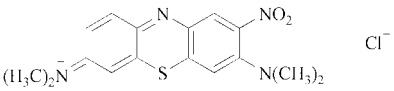

70

Several dyes that are important in microscopy are obtained by partial oxidative degradation of Methylene Blue, which removes some of the methyl groups. Cleavage of one methyl group gives Azur I (Azur B), of two methyl groups Azur A, and of three methyl groups Azur C. The oxidation is favorably carried out in an aqueous solution of potassium or sodium dichromate at 70 °C [301]. Highly purified rhodanides of these dyes can be made starting from the zinc chloride salts via the perchlorates [302]. Another dye for this purpose is Thionine or Lauth's Violet, C.I. 52 000 [*581-64-6*], with completely unsubstituted amino groups. A rapid method for separation and identification of these dyes is high-performance thin-layer chromatography [303].

Thiocarmin R [*6379-04-0*] (**49**; Cassella, 1890), is an acid wool dye from the series of diaminophenothiazine dyes [304]. It has been prepared, analogously to Lauth's Violet, by oxidation of 4-amino-*N*-ethyl-*N*-benzylaniline-3′-sulfonic acid in the presence of hydrogen sulfide and zinc chloride. In this case, too, better yields are obtained by using the thiosulfate method (see the discussion under Methylene Blue). According to the same preparation route, the thioninedisulfonic acid is obtained from 2,5-diaminobenzenesulfonic acid and 2-aminobenzenesulfonic acid.

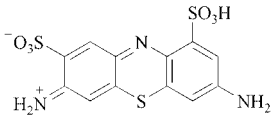

71

This dye [305] as well as cationic phenothiazine dyes can be used in photogalvanic cells [306].

4.2. Aminohydroxyphenothiazine Dyes

These dyes are essentially completely present in the phenazone form (**1**, where D^2 is =O). They are suitable for dyeing polypropylene textile materials [307]. By the introduction of a second hydroxyl group in the position neighboring the auxochromic OH group, the dyes acquire the ability to form color lakes.

Brilliant Alizarin Blue G **50** and R **50**, C.I. 52 055 [*6379-01-7*] and C.I. 52 060 [6379-02-8], dye chrome-mordanted wool green-tinged

and red-tinged blue, respectively. The dyes are obtained by reaction of the corresponding phenylenediaminothiosulfonic acids with 1,2-naphthoquinone-4-sulfonic acid and -4,6-disulfonic acid, respectively (Bayer, 1892) [308].

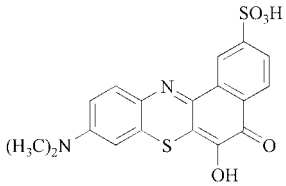

72

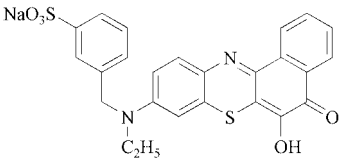

73

By reaction of 1,4-naphthoquinone with thiourea, cyclic thiocarbonate **51** is obtained, which can be treated with nitrosodiethylaniline to give thiophenoxazone dye **51**. The dye is also formed in a chromogenic developing process on image-forming irradiation of dialkyl-*p*-phenylenediamine in a silver-halide-containing gelatin layer [309].

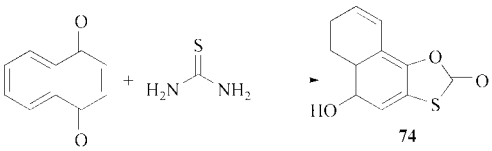

74

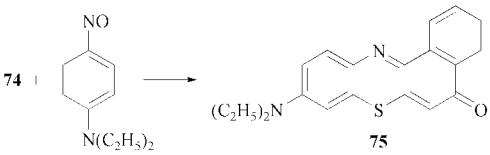

75

As a laser dye, it is tunable in the 660 – 700 nm range [310].

4.3. Hydroxyphenothiazine Dyes

Hydroxyphenothiazine dyes serve as starting material for sulfur dyes. 1,3,4-Trichloro-7-methylphenothiazine-2-one, obtained from chloranil and 1-methyl-4-amino-3-thiophenol, is the precursor for the former Immedial Reddish Brown CL3R.

5. Bisazine Dyes

By doubling of the general formula **1** of the azine dyes, one arrives at the general formula **52** of the bisazine dyes.

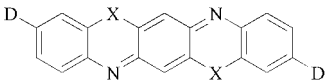

76

Whereas an uneven number of carbon atoms and heteroatoms is found between the two auxochrome groups D^1 and D^2 in the simple azine dyes, so that resonance formulas **1** – **1** are possible, the D groups in the bisazine dyes are separated by an even number of atoms, and therefore resonance in the sense of formula **1** is impossible.

The color of the bisazine dyes results from coupling of two trimethine chromophores according to **53**, which is a kind of a 1,4-didonor 2,5-diacceptor substituted benzene **53**. The mixing of their highest occupied and lowest unoccupied molecular orbitals, respectively, yields a significant bathochromic shift of the absorption band [311]. This shift and the color strength are increased by the condensed rings. Therefore, the groups D are not necessarily required and are lacking in a large number of important dyes.

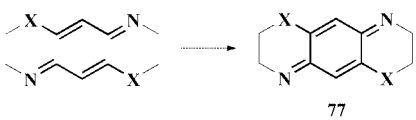

77

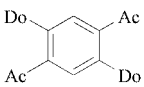

78

The bisazine dyes include the fluorindine dyes, the dioxazine dyes, and the dithiazine dyes. Of these, the dioxazine dyes have the greatest industrial importance.

5.1. Fluorindine Dyes

Fluorindine dyes are obtained by reaction of benzoquinone derivatives with phenylenediamine. 5,12-Dialkylfluorindines can be prepared by first adding two equivalents of *N*-alkyl-*o*-phenylenediamine to benzoquinone in the cold and then heating in acidic medium to form the azine ring [312]. The compounds are also obtained by condensing 2,5-dihydroxybenzoquinone with *N*-alkyl-*o*-phenylenediamines in pyridine at reflux temperature. The dyes are suitable for polyacrylonitrile [313]. They are quaternized by the action of alkylating agents to dyes like **54**, thereby becoming more soluble and electronically more similar to monoaminophenazine dyes. Polyacrylonitrile is dyed in blue tints [314].

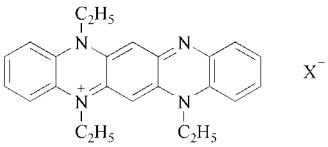

79

The alkylated diphenylfluorindines are also suitable for dyeing polyacrylonitrile [315]. Interesting derivatives of diphenylfluorindine are described in [316]. They are formed when 3-aminopyrenes are incorporated into azophenine compounds, and these are converted into fluorindines and made water-soluble by sulfonation. The products of high color strength dye cotton from a neutral bath in lightfast green tints. Fluorindine dyes that contain sulfo groups are obtained by reaction of *o*-phenylenediaminemono- or *o*-phenylenediaminedisulfonic acids with 2,5-dimethoxy-1,4-benzoquinone [317]. A violet pigment can be prepared by treating monobenzoyl-*o*-phenylenediamine with chloranil, followed by condensation to give the fluorindine [318].

5.2. Dioxazine Dyes

This class of dyes, which is especially important in dyeing cotton and as pigments, is prepared by reaction of benzoquinone derivatives, in particular tetrachlorobenzoquinone (chloranil), with aromatic amines. The components are first condensed in the presence of acid-binding agents to give 2,5-diarylamino-3,6-dichloro-1,4-benzoquinone. The cyclization is then carried out under oxidative conditions in high-boiling solvents in the presence of aromatic acid chlorides or in concentrated sulfuric acid [319]:

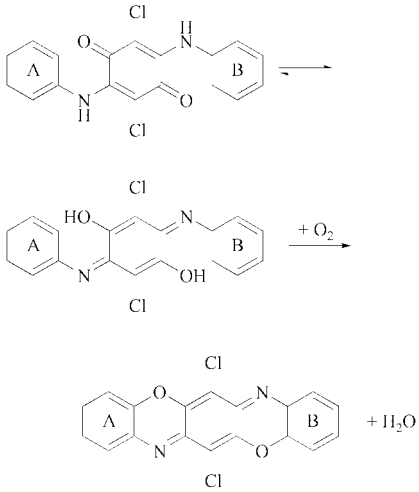

80

Instead of the widely used *o*-dichlorobenzene aryl alkyl ethers are recommended as halogen-free solvents [320].

Several modifications of the cyclization step allow the reaction to be run under milder conditions, even at room temperature: oleum [321], manganese dioxide in sulfuric acid [322], chlorine in oleum [323], iodine or an inorganic [324] or organic [325] iodine compound in oleum, bis(acetoxyiodo)benzene [326], perborate or percarbonate in sulfuric acid or oleum [327], electrochemical oxidation [328].

If the aromatic rings A and B do not contain, e.g., a sulfo group compounds **56** are water-insoluble and have pigment character. Of special importance is *Pigment Violet 23* **56**, C.I. 51 319 [*6358-30-1*], formed by condensation of chloranil with 3-amino-*N*-ethylcarbazole. Linear structure **56** was believed for long time to be the right one. But three-dimensional X-ray diffraction analysis revealed the angled structure **56** [329].

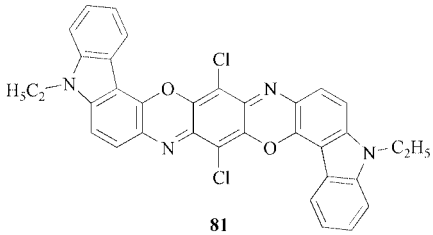

81

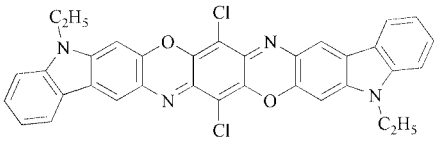

81a

Linear structure **56** is said to exhibit a blue shade and very poor application properties compared with the angled structure **56**.

Starting from 3-amino-2-methoxycarbazoles dioxazines having a linear **56** or angular **57** structure seem to be obtainable depending from the conditions used for the ring closing step, whereby the methoxyl groups are split off (→**56**) or retained in the molecule (→**57**) [330]. With such carbazoles bearing a long-chain alkyl group on the nitrogen atom, soluble carbazoledioxazine dyes have been synthesized by condensation with chloroanil and ring closure in 1-chloronaphthaline at 220 °C. The linear structure (analogous to **56**) of the molecule has been confirmed by ^{1}H NMR spectroscopy and X-ray crystallography [331]. Such long-chain substituted dyes are recommended as colorants in resins, inks, and for photoelectric materials [332].

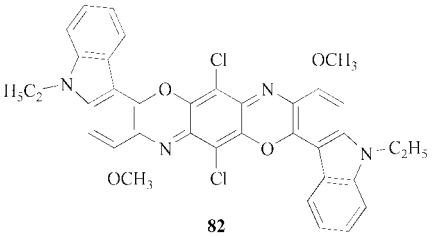

82

The preparation of Pigment Violet 23 is described below. The pigment serves for the dyeing of plastics, such as PVC, polyolefins, polystyrene, and rubber, as well for the preparation of printing colors and paints. It is also recommended for use in electrophotography [333], in organic electroluminescence devices [334], for optical data storage [335], or as shading agent together with copper phthalocyanine as light filter in liquid crystal displays [336]. To convert the crude dye to a clear, color-strength pigment, the aqueous suspension is heated in the presence of an aliphatic or alicyclic alcohol [337], an aromatic hydrocarbon [338], a ketone, or an acetic ester [339]. The pigment can be dispersed by dissolving it in a haloacetic acid and diluting it with water or a water-miscible solvent [340]. Dispersion can also be achieved by grinding in a vibrator mill [341]. The dispersability of the pigment can be improved by mixing it with a derivative of this pigment bearing amino or heterocyclic groups bound to the chromophore by spacer groups such as methylene or sulfonyl [342]. Mixtures with derivatives obtained by reaction of Pigment Violet 23 with chlorosulfonic acid followed by reaction with dimethylaminopropyleneamine result in paint or ink dispersions with low viscosity and high gloss [343].

Pigment Violet 23 as tailor-made preparations for different applications is commercially available in a great variety of assortments, e.g., *Hostaperm Violet BL*, *Hostatint Violet RL*, *Colanyl Violet RL 200* (Clariant), *Paliogen Violet L 5890*, *Xfast Violet ED 7575* (BASF), *Pigment Violet 23* (Sun-Chemicals).

The halogen-free pigment is also known. It is synthesized from *p*-benzoquinone and 3-amino-*N*-alkylcarbazoles under oxidative conditions [344] or from 2,5-dihalo- [345] or 2,5-dihydroxy-1,4-benzoquinone [346]. The monochloro pigment is analogously obtained from 2,3,5-trichloro-1,4-benzoquinone [347].

Water-soluble dyes are formed by sulfonation. They dye cotton directly in vivid violet and blue tints and are characterized by high lightfastness and high color strength. Because animal fibers are also deeply colored, the sulfonated dioxazine dyes are suitable in mixture with acid wool dyes for the dyeing of half-wool. These industrially important dyes are derived from 4-aminodiphenylamine or from multinuclear aromatic amines.

Sirius Light Blue FFRL, or *Remastral Blue FFRRL*, C.I. 51 320 [*1324-58-9*], is prepared by condensation of chloranil and 3-amino-*N*-ethylcarbazole followed by sulfonation [319].

Preparation. 1) *Condensation and oxazine-ring cyclization*. In a 2000 L enamelled kettle equipped with heater, stirrer, and condenser, 100 kg of 3-amino-*N*-ethylcarbazole, 42.5 kg of water-free sodium acetate, and 87.5 kg of chloranil are mixed by stirring in 1940 kg of *o*-dichlorbenzene at 50 – 55 °C. The mixture is stirred for 2 h at 60 – 68 °C, and the acetic acid is then distilled under vacuum at 115 °C. Subsequently, 45 kg of benzenesulfonyl chloride is added at 150 °C, and the solution is heated to 175 – 180 °C, stirred for 4 h at this temperature, allowed to cool, diluted with *o*-dichlorobenzene, and filtered at 100 °C through a pressure filter. The filter residue is washed with *o*-dichlorobenzene at 100 °C until a sample amount, boiled in *o*-dichlorobenzene, fluoresces red rather than blue. The *o*-dichlorobenzene is then removed by steam distillation and the product is washed with hot water and dried at 100 °C.

2) *Sulfonation*. 85 kg of ground dioxazine intermediate is added through a sieve over 2 h at room temperature to a mixture containing 1050 kg of 95.5 % sulfuric acid and 1050 kg of fuming sulfuric acid (29 % SO_3). The mixture is heated to 90 – 95 °C over 1 h and stirred for 8 h at this temperature. It is then allowed to cool to 25 – 30 °C. 1000 kg of water and 2000 kg of ice are added, the mixture is stirred for 24 h and filtered, and the product is washed with a 17 % solution of NaCl. The sulfonated product is mixed with 1000 kg of a 23 % solution of NaCl and made weakly alkaline with a solution of soda. After 2 h of stirring, the sodium salt is salted out with 100 kg of NaCl, filtered, and washed with 1500 L of a 23 % NaCl solution until the filtrate flows out with a weak blue color. The salt is dried at 60 – 70 °C.

The reaction of chloranil with 3-amino-*N*-ethylcarbazole is favorably carried out in the presence of a phase-transfer catalyst, e.g., tetrabutylammonium chloride [348]. Organic bases such as triethylamine, piperazine, and pyridine have also been suggested as acid-binding agents [349]. Benzenesulfonic acid and toluenesulfonic acid are also suitable as condensation agents [350]. Heterocyclic *N*-oxides such as 4-picoline *N*-oxide are advantageous because they produce less wastewater than arenesulfonyl chlorides, but high yields of pigments [351].

The unsulfonated color bases can be treated with *N*-hydroxymethylchloroacetamide to add one to four amidomethyl groups to the dye molecule. With pyridine, the chlorine substituent is exchanged for the pyridinium group to give cationic dyes that are suitable for spin dyeing of polyacrylonitrile [352]. By the action of oleum on these dyes, compounds are formed that bear cationic charges as well as sulfo groups and are suitable for dyeing natural and synthetic fibers as well as paper and leather [353]. The chloroacetyl group can also be cleaved off with hydrochloric acid and the aminomethylated dye can be treated with the reaction product formed from cyanuric fluoride and *o*-sulfanilic acid. The dye thus obtained reacts with natural and regenerated cellulose with formation of a covalent bond [354].

Sirius Light Blue FF2GL, or *Remastral Blue FF2GL* **58**, C.I. 51 300 [*6527-70-4*], is prepared from chloranil and 4-aminodiphenylamine-2-sulfonic acid [355].

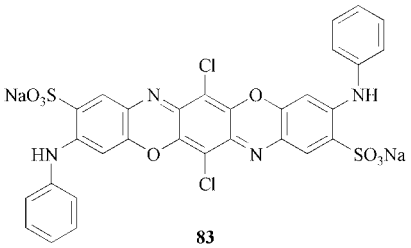

83

Other trade names were, *Solophenyl Brilliant Blue BL*, *Chlorantin Fast Brilliant Blue 2 GL* (Ciba-Geigy), *Tertro Direct Light Blue 2GLN* (Crompton & Knowles).

The corresponding dye obtained from 4-amino-4′-chlorodiphenylamine-2-sulfonic acid is *Sirius Light Blue FFB*. These dyes are also suitable for the batch dyeing of polyamides [356].

The reaction of chloranil with 1,4-phenylenediamine-2-sulfonic acid gives, after cyclization, a dioxazine dye with primary amino groups which dyes polyamide a clear blue [357]. By reaction with cyanuric chloride, it is converted to a reactive dye [358]. Introduction of substituted amino groups into the triazine rings gives dyes useful as direct dyes for textile fibers, leather, paper, and in ink-jet inks [359]. Two chromophores of this dioxazine dye can be combined by a bridging group formed by reaction with, e.g., phosgene or dicarboxylic acid dichlorides [360] or cyanuric chloride [361]. These dyes give a brilliant lightfast blue tint on cellulosic mate-

rial such as cotton and paper. They are also recommended as dichroitic dyes in light-polarizing films [362].

A sulfone residue can be introduced into these dyes in *ortho* position to one of the amino groups by reaction with sulfinic acids, e.g., acetamidobenzenesulfinic acid, in the presence of an oxidizing agent, e.g., potassium peroxodisulfate [363]. These dyes provide brilliant reddish blue shades on polyamide and wool.

Another starting material for the preparation of dioxazine reactive dyes is the product obtained from the reaction of 4,4′-diaminodiphenylamine disulfonic acid with chloranil. The primary amino groups react with cyanuric chloride to form **59** [364] (see first scheme on bottom of this page).

Another reactive group can be introduced into this dye by allowing the second chlorine atom of cyanuric chloride to react with the sulfate ester of 3-aminophenylhydroxyethylsulfone [365]. Commercial products of this series were *Procion Blue MX-G* [*72847-56-4*] (ICI), *Intracron Blue C-G* (Crompton & Knowles).

Dyes of this type in which rodlike substituents are fixed to the amino groups can be used as dichroitic dyes in polarizing films. They show excellent durability especially to heat and under moist, hot conditions [366].

Asymmetrical dioxazine dyes can be synthesized by subsequent condensation of chloranil with *p*-phenylenesulfonic acid and *p*-phenylene-2,6-disulfonic acid followed by ring closure. Reaction with cyanuric chloride and substitution of the remaining two chlorine atoms by metanilic acid and 2-(2′-hydroxyethylsulfonyl)ethylamine yields a reddish blue dye, which after esterification of the free hydroxyl group by 100 % sulfuric acid becomes a reactive dye, especially for cellulose fiber materials [367] (see second scheme on bottom of this page).

Dioxazine dyes that bear an aminoalkylamino group are starting materials for the preparation of reactive dyes such as **61**. The aliphatic amino group is reacted with cyanuric chloride [368] or with cyanuric fluoride [369], a second halogen atom on the triazine ring usually being exchanged for an anilinesulfonic acid. Higher brilliance, a less reddish blue shade, and a higher stability to alkali are said to be the advantages of these chromophores (see last scheme on bottom of this page).

Substitution of the third halogen atom by an amine [370] or by nicotinic acid [371] results in paper dyes and reactive dyes for cotton, re-

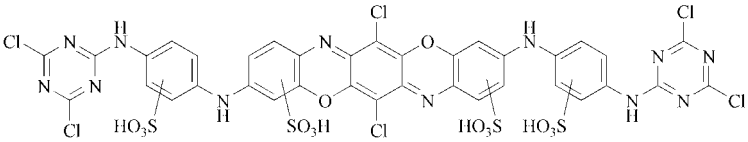

84

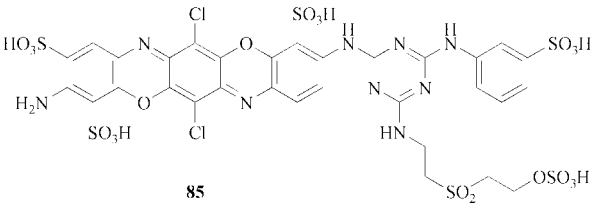

85

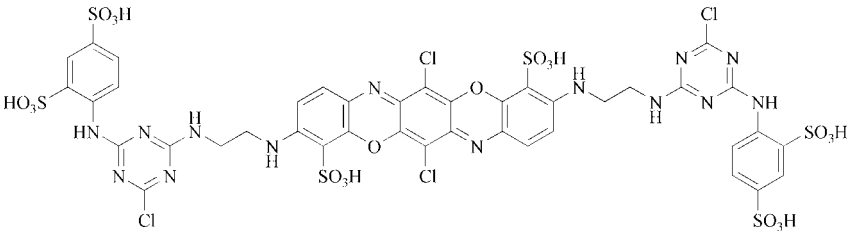

86

spectively, which can be used under neutral to acidic dyeing conditions [372]. *Procion Blue H-EG*, C.I. 51 307 [*77093-21-1*] (ICI, DyStar), *Kayacelon React Brilliant Blue CN-FL* (Nippon Kayaku) is a commercial dye of the latter type.

Oligomeric dyes can be synthesized by reaction of these aminoalkylamino groups with bifunctional acid chlorides, e.g., terephthalic acid dichloride [373]. These dyes can be used for dyeing and printing of cellulose material.

The precursors of these dyes can be desulfonated by heating in the presence of a mineral acid and subsequently alkylated. The cationic dyes thus obtained are suitable for dyeing paper [374]. Starting from γ-[2-(β-hydroxyethylsulfonyl)-4-aminoanilino]propyltrialkylammonium iodide, e.g., reaction with chloranil, ring closure, and esterification of the hydroxyl groups yields a cationic reactive dioxazine dye of high color strength which dyes cellulose, wool, or polyamide in fast reddish blue hues [375].

Red dioxazine dyes are obtained by condensation of 4-aminodiphenyl ether with chloranil, followed by either sulfonation of the dioxazine to give a polyamide dye [376] or by reaction of the dioxazine with *N*-hydroxymethylchloroacetamide and pyridine to give a cationic paper dye such as **62** [377].

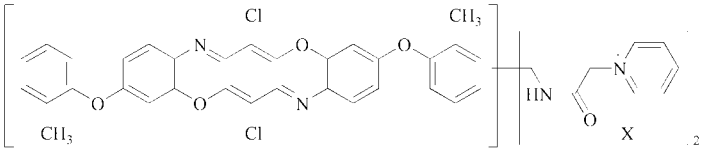

87

Condensation of chloranil with 3-sulfo-4-(ω-acylaminoalkoxy)anilines followed by hydrolysis of the amide and reaction with a fiber-reactive group, e.g., with cyanuric fluoride and than with anilinesulfonic acid [378], or condensation of chloranil with 3-sulfo-4-(ω-sulfatoethylsulfonylalkoxy)anilines [379] yields bluish red dyes for reactive dyeing of cellulose material. Chlorine-free magenta dyes useful for inkjet printing are synthesized from tetraphenoxybenzoquinone, obtained from chloranil and phenol, and 4-ethoxyaniline-3-sulfonic acid [380]. *Sirius Light Blue F3GL*, or *Remastral Blue F3GL* **63**, C.I. 51 310 [*30282-46-3*], is obtained by condensation of chloranil with aminopyrene and cyclization in the presence of *p*-toluenesulfonyl chloride and 1-chloronaphthalene at 260 °C. Sulfonation is carried out with a 15-fold excess of sulfuric acid monohydrate at 20 °C for 36 – 40 h [381].

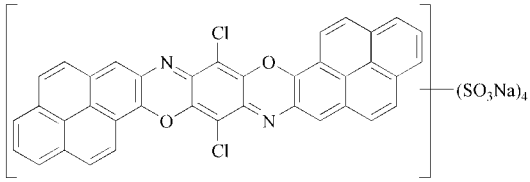

88

Sirius Light Violet FRL, or *Remastral Violet FRL*, C.I. 51 325 [*1324-43-2*], is obtained correspondingly from 2-aminofluorene [382].

If aromatic amines that bear an alkoxyl group in the position neighboring the amino group are used for the condensation, the oxazine-ring cyclization takes place with elimination of alcohol. Pigment dyes are obtained by condensation of chloranil with 1,4-dialkoxy-2-amino-5-aroylaminobenzenes followed by ring cyclization in concentrated sulfuric acid [383] with reacylation of saponified acylamino groups [384] or by ring cyclization in a solvent in the presence of sodium sulfate [385].

Commercial products of this class of dyes were *Irgazin Violet BLT* [*55177-94-1*] and *6RLT* [*12612-32-7*] (Ciba-Geigy).

Pigment dyes that bear acylamino groups instead of chlorine in the 3- and 6-positions are obtained by reaction of 2,5-dihalo-3,6-bis(acylamino)-1,4-benzoquinone with aromatic amines bearing an alkoxyl group in the position neighboring the amino group. Suitable acyl groups are the benzoyl group [386], the acetyl group [387], and the thiophenoyl group [388]. By exchange of the halogen atoms in the 3- and 6-positions for the cyano group, other valuable pigment dyes can be prepared [389]. Phenoxyl groups [390], carboxylic es-

ters, or carbonamide groups [391] also come into consideration as substituents in the 3- and 6-positions.

A commercial product of this series is *Cromophtal Violet B* [*57971-98-9*] (Ciba).

6-Sulfonyl-substituted dioxazine dyes with alkyl or aryl groups in the 3-position are recommended for dyeing paper or cotton as substantive [392] or reactive dyes [393].

Dioxazine dyes that bear alkoxyl groups [394], polyether groups [395], or phenoxyl groups [396] as substituents are suitable for the spin dyeing of polyesters. A color-strong, reddish violet vat dye is reported to be formed by treatment of triphendioxazine with $AlCl_3 \cdot 2 S_2Cl_2$ complex [397].

Dioxazine dyes, e.g., Pigment Violet 23, are suitable for information storage using the gallium arsenide semiconductor laser due to their absorption between 600 and 800 nm [398].

Musso et al. have investigated the formation of triphendioxazine from 2-hydroxy-*p*-benzoquinone and 2-aminophenol [399].

5.3. Dithiazine Dyes

Dithiazine dyes are suitable in their sulfonated form for the substantive dyeing of cotton [400]. To prepare the parent substances (**64** or **65**), two methods can be considered: the sulfurization of *N*,*N′*-diarylphenylenediamines and the reaction of phenothiazones with *o*-aminothiophenols. *N*,*N′*-diarylphenylenediamines are obtained by reaction of hydroquinone with aromatic amino compounds in the presence of sulfanilic acid at 240 °C.

Preparation. In a typical sulfurization, 52 g of *N*,*N′*-diphenyl-*p*-phenylenediamine is heated with 14.6 g of sulfur and 0.15 g of iodine for 4 h in a stream of air at 180 °C. The cold product is pulverized and extracted by boiling with 1 L of toluene. The violet dithiazine is extracted from the residue with 1,2-dichlorobenzene or pyridine.

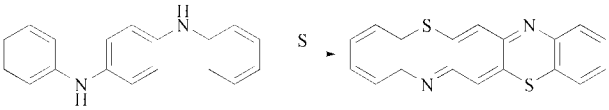

89

The reaction of phenothiazones with *o*-aminothiophenols proceeds according to the following equation:

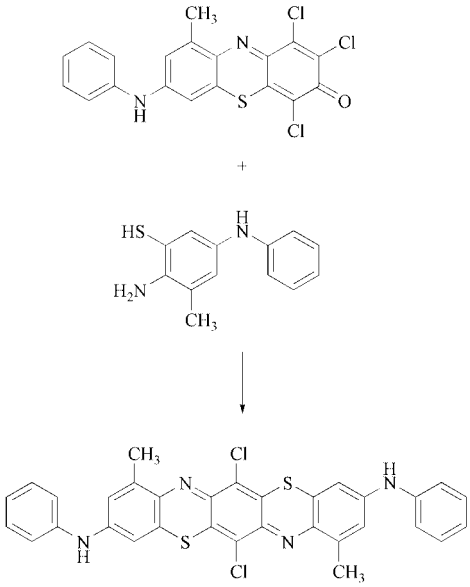

90

The diazine dyes thus obtained are sulfonated at 60 °C with sulfuric acid or oleum, depending on their constitution. The sulfonic acids dye cotton with reddish-violet to green tints.

6. References

1. S. Dähne, *Z. Chem.* **11** (1965) 441.
2. N. P. Loveless, K. C. Brown, R. H. Horrocks, *J. Org. Chem.* **46** (1981) 1182.
3. J. F. Corbett, *J. Soc. Dyers Colour.* **88** (1972) 438.
4. I. Pallagi, A. Toró, G. Horváth, *J. Org. Chem.* **64** (1999) 6530.
5. G. Schröder, W. Lüttke, *Justus Liebigs Ann. Chem.* **723**, (1969) 83; **736** (1970) 16.
6. *Friedländer*, vol. **1**, p. 247.
7. I. Sebe, I. Alexandrescu, *Rev. Cim. (Bucharest)* **14** (1990) 426.

8. H. Psaar, H. Heitzer, *Chem. Ber.* **102** (1969) 3603.
9. IG Farbenind., DE 638 338, 1933 (E. Huber, W. Vogt).
10. Minnesota Mining and Manufacturing Comp., EP 0 671 393, 1994 (D. McL.A. Grieve, A. W. Mott, R. J. D. Nairne, D. C. Bays, S. S. C. Poon, M. D. Attwood, A. C. Jackson).
11. V. N. Sil'nikov, N. P. Luk'yanchuk, G. V. Shishkin, *Isz. Akad. Ser. Khim* 1996, 2059.
12. E. Andreani, P. Costa Bizzarri, C. Della Casa, M. Fiorini, E. Salatelli, *J. Heterocycl. Chem.* **28** (1991) 295.
M. Mainwright, N. C. Grice, L. E. C. Pye, *Dyes and Pigments* **42** (1999) 45.
13. O. Meth-Cohn, M. Smith, *J. Chem. Soc. Perkin Trans. I* 1994, 5.
14. Hoechst, DE-OS 2 120 246, 1971 (R. Neeb, E. Mundlos, R. Mohr).
15. Hoechst, DE-OS 2 359 466, 1973 (M. Hähnke, T. Papenfuhs, W. Teige).
16. Hoechst, DE-OS 2 413 299, 1974 (M. Hähnke, T. Papenfuhs).
17. Du Pont, US 3 644 270, 1969 (G. D. Valiaveedan).
18. *Friedländer*, vol. **4**, p. 383.
19. Bayer, DE-OS 3 531 774, 1985 (G. Franke).
20. *Friedländer*, vol. **1**, p. 274.
J. Fernando, W. S. Morgan, J. H. Hausser, *J. Org. Chem.* **32** (1960) 1120.
21. O. Woerz, G. Scheibe, *Z. Naturforsch. B* **24** (1969) 381.
22. Reannal Finomvegyszergyár, WO 88/07583, 1987 (K. Polgár, G. Ábel, I. Harsányi, L. Nagy, S. Sipka, E. Vaszari, I. Daróczi, Z. Papp).
23. V. N. Konynkow et al., *Pharm. Chem. J. (Engl. Transl.)* **6** (1972) 429.
24. A. A. Charcharow, *Sowj. Betr. Faserforsch. Textiltech.* **3** (1966) no. **7**, 384.
25. D. F. Gloster, L. Cincotta, J. W. Foley, *J. Heterocyclic Chem.* **36** (1999) 25.
26. Ube Nitto Kasei, JP 12 948, 1962;
JP 12 950, 1962;
JP 12 953/66, 1962.
27. Ciba-Geigy, WO 9 515 144, 1993 (P. Möckli).
Henkel, DE-OS 19 746 137, 1997 (B. Meinigke, H. Möller).
28. IBM, US 4 478 687, 1983 (P. I. Gendler, R. J. Twieg).
29. H. Auterhoff, W. Wiemann, *Arch. Pharm. (Weinheim, Ger.)* **306** (1973) 481.
30. M. O. Domsjö AB, SE 408 718, 1977 (S. G. Lindenfors).
31. Canon, JP 60 195 553, 1984 (A. Yamamoto, H. Suematsu, K. Goto, T. Ikeda, E. Imai).
32. K. M. Gangotri, O. P. Regar, *Int. J. Energy Res.* **21** (1997) 1345.
33. Inst. of Phys. & Chem. Res., JP 52 092 885, 1976 (A. Yamada, M. Kaneko, H. Tsuchida).
Kagaku Res. Lab., JP 53 028 088, 1976 (A. Yamada, M. Kaneko).
Itek Corp., DE-OS 2 032 865, 1969 (R. F. Barlett).
34. Kodak, US 3 650 758, 1970 (P. B. Gilman, Jr.).
35. Hughes Aircraft, US 3 573 922, 1969 (J. B. Rust).
USM Corp., US 4 395 301, 1980 (H. Bauer, G. Piestert, H. G. Gilch).
BASF, DE-OS 3 248 246, 3 248 247, 1982 (B. Bronstert, G. Hoffmann, J. Lynch).
36. Amer. Optical Corp., US 3 660 299, 1969 (R. J. Hovey).
37. W. Müller, A. Eigel, H.-J. Schütz, H. Bünemann, DE-OS 2 712 344, 1977.
Boehringer, DE-OS 2 758 036, 1977 (W. Müller, H. Bünemann).
38. K. M. M. K. Prasad, N. Rao, *React. Kinet. Catal. Lett.* **56** (1995) 273.
K. M. M. K. Prasad, K. P. Subhashchandran, *Asian J. Chem.* **4** (1992) 715.
N. V. Rao, C. K. Sastri, *J. Indian Chem. Soc.* **64** (1987) 131.
39. Wiggins Teape, GB 1 271 289, 1969 (A. W. Hill).
Minnesota Mining and Manufactoring, WO 87 02 048, 1985 (A. G. Miller, R. J. Balchunis).
40. Hodogaya Chemical, JP 63 112 569, 1986 (M. Anzai, M. Yamaguchi, A. Utsunomiya, M. Gonda).
Minnesota Mining and Manufactoring, EP 124 296, 1982 (H. A. Brown, A. G. Miller, L. M. Leichter).
41. F. P. Schäfer, *Angew. Chem.* **82** (1970) 29, 37.
42. Ricoh, JP 10 291 368, 1997.
43. Hoechst, DE-OS 3 141 082, 1981 (M. Hähnke, W. Teige).
44. Hodogaya, JP 65-42 023, 1965.
45. Du Pont, FR 1 285 848, 1961 (D. C. Coe).
46. Inst. of Inorg. and Phys. Chem., SU 1 458 816, 1986 (I. K. Gusejnov, F. S. Mamedova, N. Ch. Rustanov);
SU 1 272 232, 1984 (C. A. Kerimova, I. K. Gusejnov).
47. Vickers PLC, EP 127 477, 1983 (C. W. Folkard, C. R. Millross).
48. Uemura Kogyo, JP 47 028 581, 1968.
RCA Corp., US 4 551 212, 1985 (S. T. Rao, L. Trager).
49. US Navy, US 3 585 114, 1969 (H. K. Wren, W. T. Hobson).
50. Ciba-Geigy, BE 874 474, 1978.

51. Schering, DE-OS 2 028 803, 1970 (H.-G. Jodt, W. Clauß, W. Dahms).
Matsushita, JP 56 062 983, 1979 (Y. Hata, Y. Okino, H. Ito).
52. *Friedländer*, vol. **9**, p. 263.
53. ICI, US 3 992 381, 1972;
DE-OS 2 438 036, 1973 (A. Crabtree).
54. Zeneca, EP 578 357, 1992 (K. Anderton).
55. *Friedländer*, vol. **4**, 426.
56. Bayer, DE-OS 1 941 105, 1969 (R. Kuth, D. Hildebrand, H.-S. Bien).
57. Mitsubishi, JP 55 042 068, 1978;
55 043 480, 1978 (Y. Yoshikawa, T. Nawata, M. Goto, Y. Fujii).
58. IG Farbenind., DE 611 966, 1932 (T. Nocken, W. Neelmeier).
59. ICI, DE-OS 2 305 990, 1972 (A. Crabtree).
60. ICI, DE-OS 2 603 398, 1975 (A. Crabtree, C. V. Stead).
61. Ricoh, JP 62 165 668, 1986 (N. Ide, K. Otaki, M. Tomita).
62. *Fiat Final Rep.*, 1313 **/2** (1948) 382.
63. M & T Chemicals Inc., US 3 966 565, 1972 (O. Kardos, D. A. Arcilesi, S. P. Valayil).
64. IG Farbenind., DE 738 049, 1935 (E. Huber).
65. F. J. Horton, D. Patterson, F. J. Rizzo, *J. Soc. Dyers Colour.* **92** (1976) 269.
66. *Friedländer*, vol. **5**, p. 358, 361.
67. Geigy, DE 504 331, 1926;
Friedländer, vol. **16**, p. 838.
68. Geigy, DE 531 977, 1927;
Friedländer, vol. **16**, p. 3020.
69. Canon, JP 58 176 276, 1982 (K. Arai, S. Koike, Y. Yokoyama, T. Sato, M. Kobayashi, T. Sakaeda).
70. N. V. Rao, I. J Kumari, *J. Inst. Chem. (India)* **59** (1987) 179.
N. V. Rao, P. V. Ramana, *Ind. J. Chem.* **21A** (1982) 214.
71. Nippon Kayaku, JP 48 097 943, 1972.
72. W. E. Solodar, A. R. Monahan, *Can. J. Chem.* **54** (1976) 2909.
73. Bashkir Inst. of Chemistry, SU 1 224 323, 1984 (G. A. Tolstikov, K. V. Rokitjanskij, U. M. Dshemilev, R. N. Fakhretdinov, A. L. Telin, G. B. Shamina).
74. Orient Kagaku Kogyo, JP 59 68 374, 1982.
75. Bayer, DE-OS 3 615 571, 1986 (R. Raue, H. Beecken).
76. Paper Centr. Res. Inst., SU 1 113 278, 1983 (V. E. Gorelik, F. V. Kaplun, E. I. Krylova, I. A. Moncenko, E. M. Orlova, M. S. Khajkin, G. B. Shtrejs);
SU 1 161 415, 1984 (M. S. Khajkin, L. A. Koslov, G. B. Shtrejs, Yu. S. Shimkin, A. M. Batvinkina, B. I. Shnaper, L. V. Tkacenko).
77. Fuji, JP 62 032 449, 1985.
78. Seiko Epson, EP 453 578, 1989 (H. Hayashi).
79. Orient Chemical Industries, JP 50 134 028, 1974;
US 4 056 530, 1976 (Y. Maekawa).
Bayer, DE-OS 3 000 529, 1980 (W. Maurer, J. Zimmer).
80. Bayer, EP 32 240, 1980 (W. Maurer, J. Zimmer).
81. Taoka Dyestuff, JP 50 130 817, 1974.
82. Amer. Cyanamid, DE-OS 3 018 842, 1980 (A. H. Granzow, E. Klingsberg).
83. Taoka Dyestuff, JP 50 148 427, 1974.
84. Aziende Colori Nazionali Affini, DE-OS 2 846 803, 1977 (M. Capaccioni, L. Gentile, A. Rusconi).
85. A. Mann et al., DD 64 761, 1967.
86. B. L. Justus, A, J Campillo, A. L. Huston, *Opt. Lett.* **19** (1994) 673.
87. D. Horák, F. Švek, J. M. J. Fréchet, *J. Polym. Sci., Part A: Polym. Chem.* **33** (1995) 2961.
88. M. Y. Melnik, SU 1 317 006, 1984.
Fur Ind. Res. Inst., SU 1 089 131, 1983 (L. S. Karbanovic, A. A. Anpilogova, N. V. Sepgejev, B. S. Grigorjev, P. T. Polusktov).
89. H. M. S. Ziena, M. M. Youssef, M. E. Aman, *Food Chem.* **60** (1997) 501.
90. Amer. Cyanamid, US 2 990 405, 1959 (F. L. Pepper, N. W. Floos).
91. Amer. Cyanamid, EP 28 699, 1979 (B. G. Buell).
92. Amer. Cyanamid, US 4 321 178, 1980, (E. L. Miller).
93. Whittaker, DE-OS 2 935 861, (1978) I. R. Mansukhani).
94. Perm. Chem. Works, SU 740 804, 1978 (S. N. Gusij, V. Ya. Pavlov, A. L. Sofronov, L. A. Spivak).
95. Exxon Research and Engineering, US 4 537 631, 1981 (T. M. Cooke, R. A. Lin, R. G. Whitfield, A. G. Hock).
96. Kodak, FR 2 010 819, 1968.
Canon, JP 76 040 806, 1970;
JP 76 047 046, 1970.
Monarch Marking Syst., US 3 908 063, 1974 (W. L. Rosendale, A. D. Lakes).
97. Fujitsu, JP 59 057 251, 1982.
Konishiroku, JP 60 192 958, 1984.
Mitsubishi Chem., EP 246 622, 1986 (T. Yoshitomi, H. Yushina, H. Horluchi, Y. Yamaguchi, K. Sekihara).
Orient Kagaku Kogyo, JP 02 102 269, 1988 (A. Shimada, M. Anzai).
J. Imaging Sci. **29** (1985) 209.
98. Pentel, JP 07 118 590, 1993.

99. BASF, EP 325 999, 1988 (G. Riedel, D. Hoffmann, H. Eisermann).
100. BASF, DE-OS 1 954 549, 1969 (A. Tartier).
101. Pentel, JP 57 078 456, 1980.
102. D. M. Molihner, R. N. Adams, W. J. Argersinger, Jr., *J. Amer. Chem. Soc.* **84** (1962) 3618.
R. Lantz, J. Gascon, *Bull. Soc. Chim. Fr.* 1965, 816.
R. V. Isaeva, T. I. Zverkova, T. M. Ivanova, E. I. Balabanov, E. I. Kovshev, *Electron. Org. Mater.* 1985, 138.
103. K. Schmidt, DE 683 951, 1934;
DE 734 702, 1935).
104. A. Schaeffer: *Handbuch der Färberei*, vol. **II**, Konradin-Verlag, R. Kohlhammer, Stuttgart, 1949, pp. 214 – 216.
Melliand Textilber. **31** (1950) 571.
105. H. A. Shah, A. K. Prassad, *Colourage* **36** (1989), 22.
106. Bayer, DE 1 444 234, 1962 (G. Breidbach, M. Schwarz, W. Kruckenberg).
107. Perm. Chem. Works, SU 709 648, 1977 (K. A. Leontjeva, G. A. Tolkaceva).
108. AS Azerb. Chloro-Org., SU 712 425, 1978 (B. A. Tagjev, A. V. Ragimov, V. Yu. Alinev, A. G. Mamedova).
109. Akad. Techn. Polnicza, PL 100 957, 1975 (K. Kozlowski, H. Maslowski, D. Jagodzinska, B. Czuprynski).
110. VEB Bitterfeld, DD 93 211, 1971 (F. Kleine, W. Simon).
111. Fuji Xerox Co., JP 61 117 566, 1984 (Y. Matsumura).
112. Seiko Epson, JP 04 161 467, 1990 (H. Nakamura, H. Niimura).
113. J. J. Langer, *Mater. Sci.* **10** (1984) 173.
114. BASF, DE-OS 3 244 900, 1982 (H. Naarmann, G. Köhler).
115. Monnet, FR 158 558, 1883.
116. *Friedländer*, vol. **2**, p. 498.
117. J. F. Corbett, *J. Chem. Soc. B* 1969, 818.
118. K. C. Brown, J. F. Corbett, *J. Chem. Soc. Perkin Trans.* **2** 1979, 308.
119. Wella, EP 687 669, 1995 (W. R. Balzer, A. Frank, A. Weinges).
120. Hoechst, DE-OS 2 909 034, 1979 (K. Rosenbusch).
121. J. F. Corbett, *Dyes and Pigments* **41** (1999) 127.
122. Schwarzkopf, BE 597 393, 1960.
123. J. F. Corbett, *J. Chem. Soc. B* 1969 827.
124. Schwarzkopf, BE 597 133, 1960.
125. Henkel, DE-OS 2 935 429, 1979 (G. Konrad, E. Lieske);
DE-OS 3 016 882, 1980 (G. Konrad, D. Rose, N. Maak).
126. Henkel, DE-OS 2 628 999, 1976 (D. Rose, E. Weinrich, E. Lieske).
127. Wella, DE-OS 2 840 830, 1978 (E. Konrad, H. Mager).
128. L'Oréal, DE-OS 2 737 138, 1976 (A. Bugaut, J.-J. Vandenbossche);
DE-OS 2 741 762, 1976 (A. Bugaut, A. R. Genet, K. G. Dossou).
Henkel, DE-OS 2 948 093, 1979 (D. Rose, E. Lieske).
129. L'Oréal, DE-OS 2 737 291, 1976 (A. Bugaut, P. Andrillon).
130. Goldwell, DE-OS 19 704 852, 1997 (F. Kufner, H. Lorenz).
131. Clairol, US 3 884 627, 1975 (F. Brody, S. Pohl).
132. Henkel, DE-OS 2 706 029, 1977 (D. Rose, P. Busch, E. Lieske).
133. Therachemie, DE-OS 1 492 198, 1965 (P. Berth, K.-J. Boosen).
Wella, DE-OS 19 619 112, 1996 (W. R. Balzer, S. Gerstung, I. Loewe).
134. Therachemie, DE-OS 1 792 526, 1968 (R. Maul, R. Rehberg).
135. Henkel, DE-OS 2 603 848, 1976 (D. Rose, E. Weinrich, E. Lieske).
136. Kao Corp., EP 376 078, 1988 (J. Kawase).
137. Henkel, DE-OS 4 330 711, 1993 (H. Hoeffkes, E. Lieske, D. Rose).
138. Wella, DE-OS 4 422 603, 1994 (W. R. Balzer, S. Gerstung, I. Loewe, A. Weinges, W. Balzer).
139. Henkel, DE-OS 19 718 534, 1997 (H. Hoeffkes, B. Meinigke, D. Rose).
140. L'Oréal, DE-OS 2 146 634, 1970;
DE-OS 2 146 635, 1970;
DE-OS 2 146 636, 1970 (G. Kalopissis, A. Bugaut, F. Estradier).
141. L'Oréal, DE-OS 2 032 342, 1969;
DE-OS 2 102 203, 1970;
DE-OS 2 201 670, 1971 (G. Kalopissis, A. Bugaut, F. Estradier).
142. L'Oréal, DE-OS 2 429 926, 1973;
DE-OS 2 429 959, 1973;
DE-OS 2 458 668, 1973 (G. Kalopissis, A. Bugaut, F. Estradier).
143. L'Oréal, DE-OS 2 234 476, 1971;
DE-OS 2 234 525, 1971;
DE-OS 2 507 566 – 2 507 569, 1974 (G. Kalopissis, A. Bugaut, F. Estradier).

144. L'Oréal, DE-OS 2 031 008, 1969;
DE-OS 2 163 636, 1970,
DE-OS 2 237 270, 1971 (G. Kalopissis, A. Bugaut, F. Estradier).
145. L'Oréal, US 4 125 367, 1976 (A. Bugaut, J.-J. Vandenbossche).
146. Wella, WO 80-01241, 1980 (E. Konrad, H. Husemeyer, H. Mager).
147. Henkel, US 4 314 809, 1982 (D. Rose, P. Busch, E. Lieske, G. Konrad).
148. Wella, GB 2 239 265, 1991, (H. Mager, R. Rau).
149. Schwarzkopf, DE-OS 19 826 456, 1998 (M. Akram, A. Kleen, W. Wolff).
150. M. Gago-Dominquez, J. E. Castelao, J.-M. Yuan, M. C. Yu, R. K. Ross, *Int. J. Cancer* **91** (2001) 575.
151. Revlon, US 3 251 742, 1962 (S. Soloway).
Perma S. A., EP 504 005, 1991 (M. Roure, P. Delattre, H. Froger).
Novo Nordisk, WO 97/37 633, 1996 (N. H. Sørensen).
152. L'Oréal, WO 9917733, 1997 (J. Cotteret, R. de la Mettrie, A. de Labbey, M. Maubru).
153. Ciba-Geigy, DE-OS 3 124 850, 1980 (J. Kelemen, H. Huber-Emden, P. M. Bersier, D. Werthemann).
154. Konica, JP 04 073 190, 1990;
JP 03 275 765, 1990 (O. Ishige, T. Masukawa).
AS Belo Phys. Org. Chem., SU 1 539 198, 1988 (Yu. A. Gudimenko, N. N. Aleksejev, G. A. Shishenko, V. E. Agabekov, V. S. Soldatov, L. V. Shibaeva).
155. Ciba, DE-AS 1 031 272, 1954 (J. Wegmann, P. Rhyner, H. Zollinger, H. W. Liechti).
Du Pont, US 2 741 605, 1954 (F. Zwügmeyer).
Bayer, DE-AS 1 030 302, 1956 (W. Müller, H. Psaar).
156. Nitto Spinning Co., JP 8311/66, 1962.
157. W. Schmidt et al., *Z. Naturforsch. A* **27A** (1972) 37.
Eastman Kodak, US 3 731 222, 1972 (K. H. Drexhage).
K. H. Drexhage in F. P. Schäfer: *Dye Lasers*, Springer Verlag 1973, p. 172 ff.
R. Raue, H. Harnisch, K. H. Drexhage, *Heterocycles* **21** (1984) 167.
158. M. Tamura, H. Hada, *Photograph. Sci. Eng.* **11** (1967) 82.
159. ICI, DE-OS 2 154 660, 1970 (A. H. Berrie, N. Huges).
Sumitomo, JP 76 29 045, 1972.
Fuji, JP 55 67 096, 1978.
Minnesota Mining and Manufacturing, EP 681 210, 1994 (D. McL. A. Grieve, J. A. Mooney, W. E. Bottomly, J. H. A. Stibbard, A. W. Mott, R. J. D. Nairne, D. C. Bays, S. S. C. Poon, R. J. Kenney, G. L. Elan, T. Ishida, D. C. Lynch, R. A. Mader, S. M. Simpson, K. M. Vogel).
160. Hodogaya, JP 48 23 828, 1971;
JP 53 41 323, 1976;
JP 53 144 339, 1977.
161. Fuji, JP 48 52 778, 1971.
Hodogaya, DE-OS 2 237 833, 1971 (Y. Miyazawa, M. Ozutsumi, K. Motohashi, S. Kimura).
Sumitomo, JP 76 02 848, 1972.
162. Matsushita Elec., JP 59 90 865, 1982.
Hilton Davis, EP 170 967, 1984 (W. Mo-Wie Hung).
163. Matsushita Elec., EP 97 493, 1982 (N. Taguchi, T. Shimizu, S. Hotta, W. Shimotsuma, S. Arai).
Hodogaya Chemical, JP 64 03 176, 1987.
164. Mitsubishi, JP 55 43 481, 1987.
165. Teijin, JP 56 65 072, 1979;
JP 56 132 560, 1980.
166. Miles Lab., US 4 141 688, 1977 (D. A. N. Morris, P. A. Rupchok, M. T. Skarstedt).
167. Donnelly, EP 612 826, 1993 (D. V. Varaprasad, A. Agrawi, M Zhao, P.-M. Allemand, C. A. Dornan, N. R. Lynam).
168. M. L. Crossley et al., *J. Am. Chem. Soc.* **74** (1952) 573.
R. C. Clapp et al., *J. Am. Chem. Soc.* **74** (1952) 1989.
Amer. Cyanamid, GB 680 496, 1950.
169. Hoechst, DE-OS 1 569 605, 1967 (N. Ottawa, G. Schäfer);
JP 77 96 94, 1969.
170. Miles Lab., US 3 732 147, 1970 (A. P. Fosker, P. J. Mill).
171. Ricoh, JP 27 6 86, 1986.
172. L. K. Denisov, SU 1 141 958, 1983.
173. C. S. Bedford, GB 428 143, 1933.
174. G. B. Afanas'eva et al., *Zh. Org. Khim.* **1** (1965) 1163.
175. Hoechst, DE-OS 1 569 604, 1966 (G. Schäfer, N. Ottawa);
JP 77 96 93, 1969.
176. PCUK, EP 38 736, 1980 (G. L. A. Belfort).
177. Ciba-Geigy, DE-OS 2 631 040, 1975 (E. Brunn).
178. BASF, DE-OS 2 206 508, 1972 (H. Baumann, U. Mayer).
179. Ciba-Geigy, EP 603 129, 1992 (H. Birri, G. Hanika, A. Nicopoulos).
180. Ciba Geigy, DE-OS 4 322 681, 1992 (G. Hanika, A. Nicopoulos).
181. BASF, EP 399 395, 1989 (J. P. Dix, G. Hansen, H. Kast).

182. BASF, EP 361 293, 1988 (J. P. Dix, G. Hansen, H. Kast).
183. Bayer, DE-OS 2 631 166, 1976 (H. Psaar).
Ciba-Geigy, EP 5 451, 1978 (P. Loew, R. Zink).
184. Bayer, DE-OS 2 353 987, 1973 (H. Psaar); DE-OS 2 631 207, 1976 (K.-F. Lehment, U. Trense).
185. Ciba-Geigy, CH 592 772, 1975 (J. Koller, S. Koller);
DE-OS 3 222 343, 1981 (P. Loew, R. Zink).
Sumitomo, JP 52 21 486, 1975.
Bayer, DE-OS 2 731 319, 1977 (H.-P. Kühlthau).
186. Sumitomo, JP 52 37 885, 1975.
187. Du Pont, US 4 063 889, 1975 (E. Kissa).
188. Ciba-Geigy, WO 95/01 772, 1993 (P. Möckli).
Wella, DE-OS 29 912 327, 1999.
189. Ciba-Geigy, DE-OS 4 326 460, 1992 (U. Lehmann).
190. IBM, EP 85 157, 1982 (M. A. Acitelli, R. F. Tynan, A. R. Wayson).
191. E. Balabanov, SU 972 457, 1980.
192. Bayer, US 4 847 177, 1987 (R. Raue, H. Psaar).
193. Canon, JP 63 74 068, 1986.
194. D. P. Pacheco, W. H. Russell, H. R. Aldag, *Solid State Lasers, SPIE Proceedings* **5332** (2004) 180.
195. Trid Store IP LLC, WO 00/48178, 1999 (M. Alperovich, H. Hamelech, I. Zuhl, E. Levich).
Trid Store IP LLC, WO 01/93258, 2000 (M. Alperovich, E. Kapinus, I. Zuhi, J. Malkin, E. Levich, V. Schwartz, S. Magnitskii, D. Pebalk, A. Tarasishin, D. Malakhov, V. Shubin).
196. Sony, EP 498 257, 1991 (K. Ito); WO 94/13 490, 1992 (K. Ito, T. Inoue, et al.).
197. *Friedländer*, vol. **3**, p. 382.
198. Hoechst, DE-OS 1 919 511, 1969 (L. Herz, R. Mohr, E. Mundlos, J. Ostermeier).
199. ICI, GB 1 358 270, 1971 (R. Baker);
GB 1 410 815, 1972 (J. L. Leng).
200. ICI, DE-OS 2 259 341, 1972 (R. Baker, J. L. Leng).
Mitsubishi, JP 49 109 683, 1973.
201. Ciba-Geigy, DE-OS 25 18 587, 1974 (P. Moser, A. Nicopoulos).
Hoechst, DE-OS 2 458 347, 1974 (R. Mohr, E. Fleckenstein).
202. Sandoz, DE-OS 2 701 677, 1976 (B. Gertisser, B. Henzi).
203. BASF, DE-OS 2 210 074, 1972 (H. Baumann, U. Mayer).
204. Hoechst, DE-OS 3 011 154, 1980 (R. Mohr, R. Neeb).
205. Ciba-Geigy, EP 57 457, 1981 (M. Herrmann, R. Jenny, M. Motter).
Hoechst, DE-OS 2 716 246, 1977 (N. Kollodzeiski, R. Löwenfeld, E. Mundlos, R. Mohr, K. Hohmann).
206. Ciba-Geigy, DE-OS 2 929 285, 1978 (P. Moser).
207. Hoechst, DE-OS 1 569 603, 1966 (G. Schäfer, A. Trombetta).
208. Hoechst, DE-OS 1 925 052, 1969 (N. Ottawa).
BASF, EP 47 429, 1980 (U. Mayer).
209. Gillette, GB 1 249 438, 1970 (J. F. Corbett, A. G. Fooks);
US 3 856 788, 1971 (J. F. Corbett, A. G. Fooks).
210. Sandoz, CH 619 476, 1976 (B. Gertisser, B. Henzi);
DE-OS 2 742 649, 1976 (B. Gertisser, B. Henzi).
211. L'Oréal, DE-OS 2 138 209, 1970 (A. Bugaut, M. Laudon).
212. US Department of Energy, US 5 149 807, 1992 (P. R. Hammond, G. F. Field);
US 4 622 400, 1983 (P. R. Hammond).
213. United States of America, US 5 792 389, 1993 (P. R. Hammond, J. F. Feeman, G. F. Field).
214. Hoechst, DE-OS 2 411 761, 1974 (E. Fleckenstein).
215. Hilton-Davis, US 4 549 192, 1984 (W. M. Hung).
216. Kodak, BE 661 783, 1965.
Hoechst, DE-OS 2 030 028, 1970 (E. Mundlos, R. Mohr, L. Herz);
DE-OS 2 065 887, 1970 (E. Mundlos, R. Mohr, L. Herz).
217. Nippon Kayaku, JP 51 53 086, 1974.
218. Ciba-Geigy, EP 55 223, 1980 (P. Loew).
219. Ciba-Geigy, US 4 137 042, 1979 (R. Défago, S. Koller, P. Moser).
220. BASF, EP 273 307, 1986 (J. P. Dix, K.-H. Etzbach, U. Mayer, R. Sens).
221. Nippon Kayaku, JP 48 97 930, 1972.
222. Sumitomo, JP 48 84 125, 1972.
Hodogaya, JP 73 32 764, 1970;
JP 49 60 334, 1972.
223. BASF, DE-OS 2 158 121, 1971 (H. Baumann, U. Mayer).
224. *Friedländer*, vol. **2**, p.173.
225. Anic S.p.A., US 4 463 099, 1981 (V. Baroncelli, C. Colapicchioni, I. Giannini, F. Porcelli).
Viomedics, EP 319 620, 1987 (S. Theodoropulos).
226. J. M. Yarborough, *Appl. Phys. Lett.* **24** (1974) 629.
J. B. Marling et al., *Appl. Opt.* **13** (1974) 2317.

227. Hoechst, DE-OS 2 919 258, 1979 (E. Fischer, J. Brandrup, J. Weinlich).
228. *Friedländer*, vol. **2** p. 184.
229. H. Musso et al., *Chem. Ber.* **90** (1957) 1814.
230. H. Schweppe, *Handbuch der Naturfarbstoffe*, ecomed Fachverlag, Landsberg 1992.
231. Hodogaya, JP 9076/65, 1961.
232. Ube Nitto Kasei, JP 13 596/66, 1962.
233. *Friedländer*, vol. **1**, p. 269.
234. *Friedländer*, vol. **4**, p. 485.
235. Ricoh, JP 62 278 572, 1986.
236. D. Basting, DE-OS 2 534 713, 1975.
N. N. Alekseev, SU 1 109 393, 1983.
237. Nippon Oils & Fats, JP 07 082 495, 1993.
238. Mitsubishi, JP 2000 234067, 1998.
239. IG Farbenind., DE 620 346 (G. Kalischer, W. Zerweck);
Friedländer, vol. **22**, p. 754.
240. IG Farbenind., DE 622 274 (W. Zerweck); *Friedländer*, vol. **22**, p. 796.
241. Seikagaku Kogyo, US 4 188 264, 1977 (S. Iwanaga, T. Morita, S. Nakamura, K. Takahashi, M. Niwa).
242. Boehringer Mannheim, DE-OS 3 534 927, 1985 (K. Mühlegger, H. v. d. Eltz);
DE-OS 3 526 566, 1985 (C. Klein, H. v. d. Eltz, H. Rupert, M. Junius). R. Herrmann, *Chimia*, **45** (1991) 58.
243. Boehringer Mannheim, DE-OS 3 526 565, 1985 (C. Klein, H.-G. Bath, H. Rupert).
244. Union Carbide, FR 1 205 992, 1958 (D. L. Bailey, R. M. Pike).
245. Hughes Aircraft, DE-OS 1 720 906, 1967 (J. B. Rust);
US 3 627 656, 1969
(L. J. Müller, J. B. Rust). G. G. Khort et al., SU 567 156, 1972.
Zentralinst. f. Schweißtechn., DD 121 329, 1975 (H. Schlegel, P. Möckel, K. Koch).
246. Little Inc., US 3 595 659, 1968 (J. C. Gerlach, I. D. Robinson, K. D. Jordan).
247. Amer. Opt. Corp., DE-OS 1 920 153, 1968.
248. Ag. Ind. Sc. Techn., JP 48 55 187, 1971;
JP 48 67 181, 1971.
Bur. Ind. Techn., JP 49 05 198, 1972.
249. Liko, JP 49 01 690, 1972.
250. Japan Kokai, JP 52 121 037, 1976.
K. Shigehara, H. Tsuchida, JP 52 121 036, 1976.
251. Hoechst, DE-OS 2 919 197, 1979 (E. Fischer, J. Brandrup, J. Weinlich).
252. Shell, JP 53 147 926, 1977;
GB 2 083 488, 1980 (J. C. Brokken-Zijp, M. J. v. d. Brink, P. A. J. M. Hendriks, J. H. H. Meurs).
The Trustees of Boston Univ., GB 1 591 941, 1977 (N. N. Lichtin, P. D. Wildes);
US 4 262 066, 1979 (R. K. Brenneman, N. N. Lichtin).
Kuraray, GB 2 064 863, 1979 (N. Michihiro).
253. Toppan Printing, JP 54 48 294, 1977.
Mitsubishi Gas, JP 54 70 887, 1977;
DE-OS 2 812 870, 1977 (Y. Yoshikawa, M. Goto, T. Nawata, Y. Kondo);
DE-OS 3 025 264, 1979 (Y. Yoshikawa, T. Nawata, M. Goto, Y. Fujii).
254. Nalco Chem., US 4 311 599, 1980 (M. Slovinsky).
255. Merck, DE-OS 3 112 442, 1981 (W. Fischer, R. Link).
256. A. Dousset, FR 2 027 498, 1969.
Miles Lab., US 4 392 864, 1982 (E. A. Helfrich, K. F. Yip).
257. Spitalul de Boli Infect. Col., DE-OS 1 914 675, 1969 (I. C. Marinescu).
258. Amer. Cyanamid, US 2 607 774, 1951 (R. C. Clapp).
259. Bio-Digital Sc., DE-OS 2 341 061, 1972 (J. H. Nodine, H. Waite, J. H. Waite, M. J. Fletcher).
260. Egyesült Gyógyszer és Tápszergyár, BE 728 439, 1968.
261. Amer. Cyanamid, US 4 212 971, 1979 (F. V. Randvere, L. M. Konzelman).
262. Interpr. Chimica Dudesti, RU 63 231, 1974 (I. G. Lukavchenko).
263. V. Kuban, I. Jancarova, *Chem. Listy* **79** (1985) 1200.
264. Hitachi, JP 55 94 680, 1979;
JP 55 94 681, 1979;
JP 55 94 682, 1979.
265. Etat Français, FR 2 570 517, 1984 (P. Gravey, J.-Y. Moisan, J.-P. Fouassier, D. Lognot).
266. *Zollinger, Color Chemistry*, 2nd ed., Weinheim, VCH Verlagsgesellschaft, 1991.
267. Mita Ind., JP 52 89 131, 1976.
268. National Cash Reg., DE-AS 1 065 114, 1956.
269. National Cash Reg., DE-AS 1 065 546, 1956.
270. Yamada Kagaku, JP 56 104 094, 1980.
271. IBM, US 4 309 255, 1980 (P. L. Gendler, B. D. Grant, C. D. Snyder).
272. Sumitomo, JP 50 20 809, 1973.
273. National Cash Reg., DE-AS 1 065 115, 1956.
274. Sumitomo, JP 70 24 676, 1966;
JP 73 29 729, 1966.
275. Hodogaya, JP 72 04 638, 1968;
JP 74 24 315, 1969.
Mita Ind., JP 52 105 931, 1976.
276. ICI, DE-OS 2 154 659, 1970 (N. Hughes, A. H. M. Renfrew).
277. Toppan Printing, JP 57 103 895, 1980.

278. National Cash Reg., GB 725 275, 1953. Wiggins Teape, GB 1 429 640, 1972 (J. F. McKellar, G. G. Warburton, D. M. Wood); DE-OS 2 402 384, 1973 (C. Srawley).
279. Aizen, JP 59 76 999, 1982.
280. Mita Ind., JP 79 09 062, 1971.
281. H. A. Potts et al., *J. Appl. Chem. Biotechnol.* **22** (1972) 651.
282. National Cash Reg., DE-OS 2 011 384, 1969 (P. S. Phillips jun.).
283. Kanzaki Paper, DE-OS 2 151 113, 1970 (Y. Takegawa, Y. Nishibayashi).
284. Yamamoto Kagaku, JP 12 248/71, 1966.
285. Oji Paper, JP 07 266 693, 1994.
286. Hodogaya, JP 04 28 595, 1990.
287. Fuji, DE-OS 2 144 284, 1970 (M. Hashiue, H. Yamashita, N. Sekikawa, S. Asaka).
288. *Chem. Abstr.* 1 156 458, 1980; EP 58 338, 1981 (W. E. Bernier). Mitsubishi Electric, JP 58 38 782, 1981.
289. Hilton-Davis, EP 170 188, 1984 (M. B. Gunn, W. Mo-Wei Hung).
290. Dainichi Nippon Cables, JP 50 158 638, 1974.
291. Dainichi Nippon Cables, JP 51 47 942, 1974; JP 51 112 857, 1975; JP 51 121 053, 1975.
292. Teijin, JP 86 043 661, 1986.
293. General Electric Comp., WO 03/090220, 2002 (K. A. Ezbiansky, D. R. Olson, P. Schottland, V. Thomas, M. B. Wisnudel).
294. Mitsubishi, DE-OS 2 354 084, 1972 (S. Inoue, H. Honda, F. Kitakyushu).
295. Hoechst, EP 68 425, 1981 (G. Waschulzik, H. Oehlschläger, H. Schuldes).
296. Mitsubishi Electr., JP 62 100 557, 1985. K. Enmanji, T. Andoh, *Nippon Kagaku Kaishi* 1987 1078.
297. C. Lee, Y. W. Sung, J. W. Park, *J. Phys. Chem.* B **103** (1999) 893.
298. Takeda Chem. Ind., EP 510 668, 1991 (H. Masuya, H. Shimadzu, T. Miyawaki, M. A. Motsenbocker).
299. *Friedländer*, vol. **3**, p. 360.
300. *Friedländer*, vol. **2**, p. 153.
301. Intreprind. Chim. Dudesti, RU 63 783, 1973 (K. A. Lobashov, E. M. Spektor).
302. Heyl Chemisch-pharmazeutische Fabrik, DE-OS 3 305 304, 1983 (S. Heydolph, W. Parr, E. Heyl).
303. B. M. van Liedekerke, A. P. De Leenheer, B. M. De Spiegeleer, *J. Chromatogr. Sci.*, **29** (1991) 49.
304. *Friedländer*, vol. **2**, p. 156.
305. W. J. Albery, P. N. Bartlett, A. W. Foulds, F. A. Souto-Bachiller, R. Whiteside, *J. Chem. Soc. Perkin Trans.* **2** 1981, 794.
306. Shell, GB 2 083 488, 1980 (J. C. Brokken-Zijp, M. J. van den Brink, P. A. J. M. Hendriks, J. H. H. Meurs).
307. Hodogaya, JP 9 076/65, 1961.
308. *Friedländer*, vol. **3**, p. 1011, 1016; vol. **4**, p. 456.
309. G. Mann, S. Hauptmann, H. Wilde, J. Lehmann, M. Naumann., DD 139 268, 1978; DD 139 269, 1978.
310. R. Raue, H. Harnisch, K. H. Drexhage, *Heterocycles* **21** (1984) 167.
311. S. Dähne, D. Leupold, *Angew. Chem.* **78** (1966) 1029. K. Elbl, C. Krieger, H. A. Staab, *Angew. Chem.* **98** (1986) 1024. R. Gompper, H.-U. Wagner, *Angew. Chem.* **100** (1988) 1492.
312. Eastman Kodak, DE-OS 1 569 711, 1965 (C. E. Osborne, E. R. Shelton).
313. Eastman Kodak, DE-OS 1 619 441, 1965 (J. M. Straley, R. C. Harris).
314. Eastman Kodak, DE-OS 1 619 442, 1965 (J. M. Straley, R. C. Harris).
315. Cassella, DE-AS 1 250 783, 1963 (K. Mix, H. Vollmann).
316. IG Farbenind., DE 659 592, 1933 (A. Wolfram, K. Bonstedt); *Friedländer*, vol. **24**, p. 567.
317. Centr. Med. Res. Inst., JP 49 109 399, 1973.
318. BASF, DE-OS 2 145 029, 1971 (P. Dimroth, H. Trapp).
319. Hoechst, DE 517 194, 1928 (G. Kränzlein, H. Greune, M. Thiele); *Friedländer*, vol. **17**, p. 942.
320. Clariant, EP 1 038 923, 1999 (G. Nagl, W. Bauer, P. Kempter).
321. Ciba-Geigy, EP 298 916, 1987 (H. Hahn, R. Blattner).
322. Bayer, DE-OS 3 510 613, 1985 (H. Dürscheid, W. Harms).
323. Bayer, EP 296 411, 1987 (J.-W Stawitz, W. Harms).
324. Bayer, DE-OS 3 339 923, 1983 (H. Jäger).
325. BASF, EP 311 969, 1987 (U. Nahr, M. Patsch).
326. M. Xia, C. Ye, *J. Chem. Res., Synop.* 1999, 618.
327. BASF, EP 400 429, 1989 (M. Patsch, C. Marschner).
328. Nippon Kayaku, JP 61 09 84, 1984.
329. E. Dietz, Proc. 11th Int. Color Symp., Montreux, 1991; K. Hunger, *Rev. Progr. Coloration* **29** (1999) 71; see also: Ciba Specialty Chemicals, EP 0 874 025, 1997 (F. Bäbler).

330. H. Nishi, Y. Hirasawa, K. Katahara, *Nippon Kagaku Kaishi* 1987, 911.
331. M. Ikeda, K. Kitahara, H. Nishi, K. Kozawa, T. Uchida, *J. Heterocyclic Chem.* **27** (1990) 1575; M. Ikeda, K. Kitahara, H. Nishi, *J. Heterocycl. Chem.* **29** (1992) 289.
332. Nippon Kayaku, JP 62 185 088, 1986.
333. Fuji Xerox, JP 05142 862, 1991.
334. Pioneer Electronic Corp., JP 07 090 254, 1993 (T. Murayama, S. Yamamura, M. Ikeda).
335. Ricoh, JP 58 132 231, 1982.
336. Toppan Printing Co. Ltd., JP 2001 124921, 1999.
337. Sumitomo, JP 58 120 673, 1982.
338. Sumitomo, JP 57 05 761, 1980.
339. Sumitomo, DE-OS 3 211 607, 1981 (I. Sawaguchi, Y. Hayashi, T. Yoshiaki).
340. Pentel, JP 52 00 935, 1975.
341. Hoechst, DE-AS 2 742 575, 1977 (E. Spietschka, M. Urban, E. Paulus).
342. Hoechst, EP 504 923, 1991 (E. Dietz, M. Urban).
Toyo Ink, JP 04 246 469, 1991; JP 06 049 386, 1992.
343. Sumitomo, EP 659 842, 1993 (H. Inoue, Y. Hayashi, H. Senba, K. Ito, M. Akita).
344. Sumitomo, JP 62 149 757, 1985.
345. Nippon Kayaku, JP 07 331 095, 1994 (M. Ikeda, M. Oonishi).
346. Bayer, DE-OS 3 444 886, 1984 (G. Franke, W. Harms).
347. Nippon Kayaku, JP 07 331 096, 1994 (M. Ikeda, M. Oonishi).
348. Sumitomo, JP 56 141 355, 1980.
349. Nippon Kayaku, JP 58 118 855, 1982. Sumitomo, JP 62 192 385, 1986.
350. Sandoz, DE-OS 3 239 321, 1981 (M. Deur).
351. Clariant, DE-OS 19 643 344, 1996 (G. Nagl, W. Bauer, M. Urban, D. Schnaitmann).
352. Ciba-Geigy, EP 83 309, 1981 (J.-M. Adam).
353. Ciba-Geigy, EP 14 678, 1979 (J.-M. Adam, P. Galafassi).
354. Bayer, EP 53 743, 1980 (W. Harms, K. Wunderlich).
355. Hoechst, GB 448 182, 1934.
GAF, US 2 134 505, 1934 (A. Brunner, H. Greune, M. Thiele, K. Thiess).
356. E. Schulz et al., DD 101 174, 1972.
357. ICI, DE-OS 2 122 262, 1970 (B. Parton).
358. ICI, DE-OS 2 124 080, 1970 (B. Parton).
359. Clariant, WO 00/71545, 1999 (L. Hasemann).
360. Bayer, DE-OS 3 832 531, 1988 (H. Henk, W. Harms, P. Wild).
361. Sandoz, DE-OS 3 905 673, 1989; DE-OS 4 005 551, 1989 (R. Pedrazzi).
Bayer, DE-OS 3 939 700, 1989 (H. Henk, W. Harms, P. Wild).
Bayer, DE-OS 4 101 067, 1991 (H. Jäger).
362. Bayer, DE-OS 4 000 481, 1990 (P. Wild, U. Claussen, F. W. Kröck).
363. Clariant, WO 99/51681, 1998 (G. Schofberger).
364. ICI, DE-OS 2 302 382, 1972 (B. Parton).
365. Sumitomo, JP 57 14 654, 1980.
366. Sumitomo, JP 2000 327936, 1999.
367. Sumitomo, EP 593 016, 1992 (K. Akahori, T. Omura, S. Kawabata, N. Harada).
368. ICI, DE-OS 2 503 611, 1974 (J. L. Leng, B. Parton, D. R. A. Ridyard, J. R. Lawson); DE-OS 2 823 828, 1977 (A. Greenall, J. L. Leng).
369. ICI, DE-OS 2 600 490, 1975 (J. L. Leng, D. W. Shaw).
370. ICI, EP 84 718, 1982 (A. H. M. Renfrew).
371. ICI, DE-OS 3 018 106, 1979 (H. G. Connor).
372. Sumitomo, JP 60 181 373, 1984.
373. Ciba-Geigy, EP 486 423, 1990 (U. Lauk).
374. ICI, EP 95 255, 1982 (K. Anderton, N. Hall, B. Parton, A. H. M. Renfrew, M. Shaw).
375. Hoechst, EP 355 735, 1988 (H. M. Büch, H. Springer).
376. Ciba-Geigy, DE-OS 2 506 098, 1974 (H. Schwander, A. Pugin, K. Burdeska).
377. Ciba-Geigy, EP 15 232, 1979 (J.-M. Adam, P. Galafassi).
378. Bayer, DE-OS 3 635 312, 1986 (W. Harms). Sumitomo, JA 02 238 064, 1989.
379. Bayer, EP 355 492, 1988 (W. Harms).
380. Canon, Mitsubishi, JA 62 190 276, 1986.
381. Hoechst, DE 606 672, 1932 (G. Kränzlein, H. Greune, W. Schultheiss); *Friedländer*, vol. **21**, p. 1187.
382. Hoechst, DE 637 020, 1933 (G. Kränzlein, H. Greune, W. Schultheiss); *Friedländer*, vol. **23**, p. 710.
383. BASF, DE-AS 1 034 294, 1955 (H. Fischer).
384. Geigy, DE-AS 1 088 636, 1958 (A. Pugin, E. Stocker).
385. Geigy, DE-AS 1 243 303, 1963 (A. Pugin, J. v. d. Crone).
386. Geigy, DE-AS 1 444 772, 1963 (A. Pugin, J. v. d. Crone, E. K. Burdeska).
Ciba-Geigy, DE-OS 2 156 685, 1970 (K. Burdeska, A. Pugin).
387. Ciba, DE-OS 2 047 332, 1969 (S. Hari, K. Ronco).
Ciba-Geigy, DE-OS 2 061 702, 1969 (K. Ronco, H. Tschudin).
388. Ciba-Geigy, DE-OS 1 644 672, 1966 (K. Ronco, I. Hari).

389. Ciba, DE-OS 1 929 507, 1968 (S. Hari, K. Ronco);
FR 2 010 714, 1968.
Mitsubishi, JP 52 30 823, 1975;
JP 52 134 634, 1976, JP 52 140 683.
390. Ciba-Geigy, DE-OS 2 322 892, 1972 (K. Burdeska, A. Pugin).
391. Geigy, DE-AS 1 231 370, 1962 (J. v. d. Crone, A. Pugin).
PCUK, FR 2 328 710, 1975 (A. Le Pape).
392. Bayer, WO 99/12937, 1997 (K. Kunde).
393. Bayer, WO 99/23170, 1997 (K. Kunde).
394. Ciba-Geigy, DE-OS 2 355 694, 1972 (A. Pugin, K. Burdeska, J. v. d. Crone).
395. Mitsubishi, DE-OS 2 733 539, 1976 (S. Imahori, Y. Murata, S. Suzuki, K. Yokohama);
JP 53 26 826, 1976;
JP 53 134 971, 1977;
JP 54 68 478, 1977.
396. Ciba-Geigy, DE-OS 2 157 519, 1970 (A. Pugin, K. Burdeska, J. v. d. Crone).
397. Du Pont, US 2 504 153, 1948 (G. B. Robbins).
398. Ricoh, JP 58 132 231, 1982.
399. H. Musso, D. Döpp, J. Kuhls, *Chem. Ber.* **98** (1965) 3937.
400. ICI, GB 698 200, 1951 (J. M. Fielden, D. G. Wilkinson).

Aziridines

Ulrich Steuerle, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 6 and 7)

Robert Feuerhake, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
3.1. Homopolymerization 2
3.2. Aziridine-Modified Polymers 2
3.3. Polyfunctional Aziridines 3
3.4. Miscellaneous 3
4. Production 4
4.1. The β-Chloroethylamine Process . . . 4
4.2. The Dow Process 4
4.3. The Wenker Process 4
4.4. Catalytic 2-Aminoethanol Dehydration Process 5
5. Environmental Protection 5
6. Uses 5
7. Economic Aspects 6
8. Toxicology and Occupational Health 6
8.1. Ethylenimine 7
8.2. Propylenimine 7
8.3. Aziridine Derivatives and Polyethylenimines 7
9. References 8

Aziridines are three-membered organic heterocycles which contain one nitrogen atom in the ring. All commercially available aziridine derivatives are made from ethylenimine (aziridine) and propylenimine (2-methylaziridine). Aziridines can undergo polymerization reactions, i.e., homopolymerization to form polyethyleneimines, or graft and copolymerization reactions to give aziridine-modified polymers. Polyfunctional aziridines are obtained by reaction of ethylenimine and propyleneimine and trifunctional arcylates. Ethylenimine and propylenime react with thiols and amines in ring-opening reactions, with oxiranes and isocyanates under preservation of the aziridine ring.Ethylenimine is produced industrially by the β-chloroethylamine process, the Dow process, the Wenker process and the catalytic 2-aminoethanol dehydration process. Aziridines, polyethylenimines and aziridine-modified polymers are used primarily in the coatings, paper, water treatment, textile, and petroleum industries. Aziridines are highly toxic. Their toxicity probably results from alteration of nucleic acids, proteins, and other important biochemicals by alkylation. Substances derived from aziridines but no longer containing the aziridine ring, e.g., polyethyleneimines, are normally of low toxicity.

1. Introduction

Aziridines are saturated three-membered ring compounds containing a single nitrogen atom in the ring. All commercially available aziridines and aziridine derivatives are made from ethylenimine (**1**) or propylenimine (**2**). Ethylenimine is also called aziridine [*151-56-4*] and propylenimine is called 2-methylaziridine [*75-55-8*].

1

2

Ethylenimine was first prepared in 1888 by Gabriel [4], who mistakenly called it vinylamine. He prepared the ethylenimine by reacting 2-bromoethylamine hydrobromide [*2576-47-8*] with silver oxide [*1301-96-8*] or potassium hydroxide [*1310-58-3*]. Propylenimine was first prepared in 1890 by Hirsch [5] from 2-bromopropylamine [*20888-43-1*] in a similar manner. Commercial production of ethylenimine was launched in Germany in 1938.

2. Physical Properties

Ethylenimine and propylenimine are clear, colorless liquids with amine-like odors. They are

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_239.pub2

miscible with water and most organic liquids. Pure aziridines are generally stored over solid sodium hydroxide [*1310-73-2*] and when stored in this manner are stable indefinitely. Other properties are shown in Table 1.

Table 1. Physical properties

	Ethylenimine	Propylenimine
M_r	43.07	57.10
Density at 25 °C, g/mL	0.832	0.802
bp, °C	57	66
mp, °C	−74	−65
Vapor pressure, 25 °C, kPa	28.5	18.7
Refractive index n_D^{25}	1.4123	1.4084
Viscosity, 25 °C, mPa · s	0.418	0.491
Flash point (closed cup), °C	−11	−4

3. Chemical Properties

Ethylenimine and propylenimine are highly reactive, versatile compounds that participate in two general types of reaction: (1) the ring-preserving reactions, involving alkylation or acylation of the ring nitrogen atom, and (2) the ring-opening reactions, involving protonation of the ring nitrogen followed by nucleophilic attack at one of the carbon atoms. Use of aziridines frequently is the most economical and efficient method of incorporating an ethylamine group into a polymer or into a complex organic molecule.

3.1. Homopolymerization

Polyethylenimine [*9002-98-6*] is produced by the homopolymerization of ethylenimine [6] [7] [8]. The reaction is catalyzed by acids, Lewis acids, or haloalkanes. The polymerization is usually carried out at 90 – 110 °C in water or in a variety of organic solvents. The average molecular mass of the polyethylenimine prepared as described above is 10 000 – 20 000. Higher molecular mass polymers are prepared by addition of a difunctional alkylating agent, such as chloromethyloxirane [*106-89-8*] or 1,2-dichloroethane [*107-06-2*]. Polyethylenimines with a higher average molecular mass can also be provided by ultrafiltration of polymers with a broad mass distribution [9, 10]. Likewise, polymers of lower molecular mass can be obtained by inclusion of a low molecular mass amine, such as 1,2-ethanediamine [*107-15-3*], during polymerization [11]. By using these techniques a range of molecular masses from 300 to 10^6 can be obtained. Cross-linking during the polymerization of ethylenimine in organic solvents leads to solid polyethylenimines [12]. Furthermore the polymerization process can be conducted on the surface of organic [13] or inorganic [14] materials, thus fixing the polyethylenimines to a support.

All of these polyethylenimines are highly branched polymers having roughly a spherical shape. The amine distribution is approximately 30 % primary, 40 % secondary, and 30 % tertiary. This distribution can be readily determined by ^{13}C NMR spectroscopy [15]. By using special techniques in the polymerization of ethylenimine, more linear [16] and more branched [17] polymeric structures can be obtained. Polyethylenimines can be easily modified by polymer-analogous reactions to increase their performance in certain fields of application.

Trade names: Lupasol, Polymin, Catiofast, Lugalvan (BASF), Epomin (Nippon Shokubai).

3.2. Aziridine-Modified Polymers

Ethylenimine and propylenimine graft copolymers can be obtained by reaction of ethylenimine with polymers either having carboxyl groups attached to the polymer backbone [18, 19] or having amines in their backbone [20]. This reaction improves the adhesion of the polymer to anionic surfaces or surfaces containing hydroxyl groups.

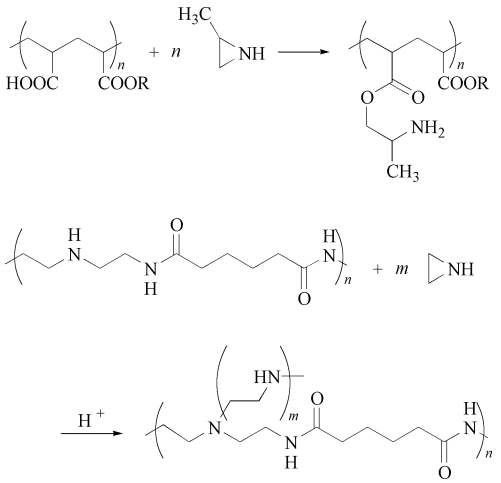

An alternative method of introducing polyethylenimine-type side chains into other polymers is by (co-)polymerizing monomers which are functionalized by oligoethylenimine side chains [21]. By this method, carboxylic acid-free polymers with high charge densities can be obtained.

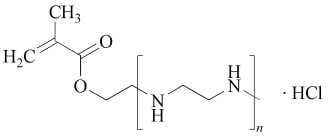

3.3. Polyfunctional Aziridines

Ethylenimine and propylenimine react with trifunctional acrylates, such as 2-ethyl-2-(hydroxymethyl)-1,3-propanediol triacrylate [*15625-89-5*] or 2,2-bis(hydroxymethyl)-1,3-propanediol triacrylate [*3524-68-3*], to produce trifunctional aziridines.

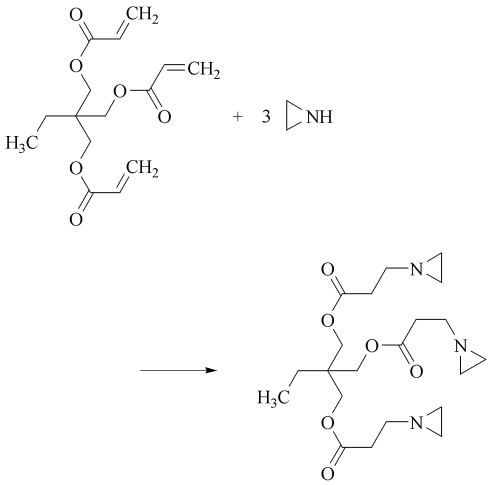

This reaction is normally carried out in the absence of solvent and does not require a catalyst. The resulting products are clear, slightly viscous liquids.

Trade names: Corial Härter AZ (BASF), Chemitite PZ (Nippon Shokubai), Xama (EIT), Crosslinker CX 100 (Zeneca), Ionac Pfaz (Sybron Chemicals). The cross-linkers are sold under their chemical names by Sogo Pharmaceutical.

3.4. Miscellaneous

Ethylenimine reacts with hydrogen sulfide to yield thiols and with thiols to give aminoethyl sulfides. This reaction is used commercially [22] to produce cysteamine (2-aminoethanethiol) (**3**) [*60-23-1*], a raw material for the synthesis of pharmaceuticals, and 2,2′-thiobisethylamine (**4**) [*871-76-1*].

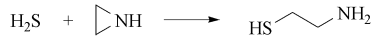

3

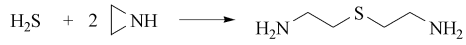

4

The reaction of ethylenimine with 2-mercaptoethanol to give 2-[(2-aminoethyl)thio]ethanol [*24304-84-5*] (**5**), an intermediate for dye production, and with sulfurous acid to give taurine (2-aminoethanesulfonic acid) [*107-35-7*] (**6**), an essential additive for mammalian nutrition, are further examples of industrially relevant ring-opening reactions of ethylenimine with sulfur nucleophiles.

The reaction of ethylenimine with amines provides an elegant synthetic route to asymmetrically substituted ethylenediamines, e.g. *N*,*N*-dimethyl-1,2-ethanediamine (**7**) [*108-00-9*] [23].

5

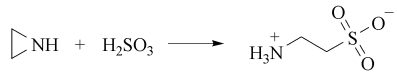

6

7

Ethylenimine and propylenimine react with oxiranes, opening the oxirane ring while leaving the aziridine ring intact. For example, 1-aziridine ethanol [*1072-52-2*], 1-(2-hydroxyethyl)ethylenimine, has been produced by the

reaction between ethylenimine and oxirane [*75-21-8*].

Another commercial application of ethylenimine is the reaction with isocyanates to give iminoureas. For example, ethylenimine reacts with 1,6-diisocyanatohexane [*822-06-0*]:

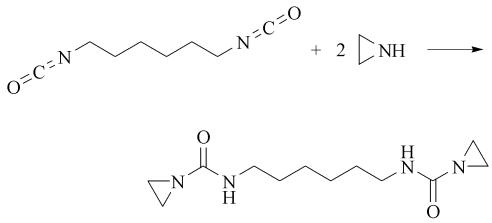

4. Production

Four industrial processes for ethylenimine production have been described in the literature: the β-chloroethylamine process, the Dow process, the Wenker process (BASF) and the catalytic 2-aminoethanol dehydration process (Nippon Shokubai). Only the two last-mentioned processes are still used today. The 2006 capacity for ethylenimine monomer worldwide is estimated to be about 9 000 t/a. Because of the toxicity and high reactivity of the monomer the two major producers, BASF and Nippon Shokubai, do not sell the monomer, but directly convert it to nontoxic polymers and intermediates.

4.1. The β-Chloroethylamine Process

From 1938 till 1963, ethylenimine was produced commercially in Germany by the reaction of 2-chloroethylamine hydrochloride [*870-24-6*] with sodium hydroxide. Disadvantages of this process are, similar to the Dow Process, the corrosivity of the chloride byproduct and the risk of contamination of the ethylenimine product with β-chloroethylamine. Since β-chloroethylamine can readily eliminate HCl, which then initiates uncontrolled polymerization of ethylenimine, this contamination must be carefully avoided.

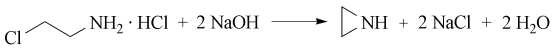

4.2. The Dow Process

In 1963 Dow Chemical began production of ethylenimine by a newly developed process that offered the possibility of significantly lower raw material costs [24]. This process is based on the reaction of 1,2-dichloroethane with excess ammonia [25]. Dow stopped production in 1978.

The advantage of this process is its low raw material cost. However, there are difficulties caused by impurities in the ethylenimine, high corrosion rates, and waste-stream disposal.

4.3. The Wenker Process

In 1935, Wenker [26] developed a process for making ethylenimine that is still the basis for most of the commercial production today. BASF produced ethylenimine according to the Wenker process in Ludwigshafen, Germany for more than 30 years safety without technical problems. The Wenker process is a two-step process, initially involving the reaction of 2-aminoethanol with sulfuric acid to give 2-aminoethyl hydrogensulfate [*926-39-6*] [27]:

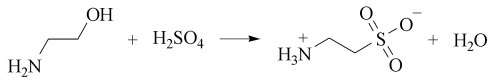

The product is a water-soluble, nonvolatile solid with properties similar to those of an amino acid. This ensures, that, in contrast to the β-chloroethylamine and Dow processes, no contamination of the ethylenimine occurs. The 2-aminoethyl hydrogensulfate reacts with aqueous sodium hydroxide to produce ethylenimine. The cyclization is preferably carried out under pressure at elevated temperatures [28, 29].

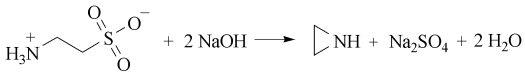

Originally yields of only 26 % were obtained, but yields of 85 – 90 % [30] are now attained. The advantage of this process is the production

of pure ethylenimine with no significant waste-disposal problems. The disadvantage is the relatively high raw material cost.

4.4. Catalytic 2-Aminoethanol Dehydration Process

Numerous attempts have been made to establish a direct industrial synthesis for ethylenimine starting from 2-aminoethanol [*141-43-5*] since the 1970s [31]. The key to an economic process was a suitable catalyst, which took a long time to develop. In 1990, Nippon Shokubai eventually started an industrial ethylenimine production via catalytic dehydration of 2-aminoethanol. The reaction is carried out at 350 – 450 °C, preferably under reduced pressure with a catalyst which has weakly acidic and weakly basic centers, e.g., SiO_2 doped with barium, cesium, and phosphorus. In this gas-phase reaction, selectivities up to 90 % at conversions from 40 – 80 % can be obtained. Unconverted 2-aminoethanol is recycled from the product mixture and used as feedstock [32 – 35].

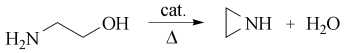

The advantage of the gas-phase process over other industrial processes is that it is a salt-free direct one-step synthesis based on 2-aminoethanol. Disadvantages are:

- Multistage distillation is required for the separation of ethylenimine from byproducts and for the recycling of 2-aminoethanol
- High energy demand
- Necessity for repeated regeneration of the catalyst due to deposition of coke and loss of phosphorus compounds [36]

5. Environmental Protection

The primary environmental concern is the prevention of aziridine vapors from entering the environment (see Chap. 8). Efficient scrubbers should be used on any vent line or hood exhaust where aziridine vapor may be present. An effective scrubbing medium is an aqueous solution of diammonium thiosulfate [*7783-18-8*].

The only waste stream from the Wenker process is an aqueous one, primarily containing sodium sulfate [*15124-09-1*] along with small amounts of various organic nitrogen compounds. This stream can be used by a kraft pulp mill as a source of sulfate, making the Wenker process essentially a non-waste-producing system. No waste streams are produced in the preparations of the principal aziridine derivatives, i.e., homopolymer, aziridine-modified polymers, and polyfunctional aziridines.

6. Uses

Aziridines are used primarily to manufacture products for the coatings, paper, water treatment, and petroleum industries.

In the *coatings* industry, polyfunctional aziridines are used to cross-link polymers with pendant carboxyl groups. These cross-linked polymers are used in a variety of performance coatings, such as those on exterior wood panels. Polyethylenimines are used as tie-coat adhesives, e.g., in the production of polypropylene laminated films [37], as adhesion promoters, and as pigment dispersants. Furthermore, polyethylenimines are excellent primers, e.g., for acrylate-based adhesives, and are used as amine components in epoxy and polyurethane resins.

In the manufacture of *paper*, polyethylenimine and aziridine-modified polymers are used as retention and drainage aids [38, 39]. Due to increasing rates of paper recycling and the trend towards closed water circuits in the paper industry, the role of polyethylenimines as fixing agents for soluble and insoluble contaminants (stickies) has gained importance [40].

Polyethylenimines are used in *water treatment* as clarifying aids [41] and are effective silica antiscaling agents [42]. Polyethylenimines can be used as *chelating agents* for heavy metal ions such as copper, rhodium, mercury and zinc. By carboxymethylation the complexing abilities of the polyethylenimines can be extended to other heavy metals and even alkaline earth metals. The enrichment of heavy metals from aqueous solutions can be conducted by membrane technology with polyethylenimine solutions [43, 44] or with solid polyethylenimines.

They are also used in the manufacture of *reverse osmosis membranes* [45].

In *textile* applications, polyethylenimines can be used in dyeing pretreatment and posttreatment of textiles to improve dye fixation and fastness. Furthermore, they improve the anti-static properties of hydrophobic fibers and enhance shrinkproofing [46].

Polyethylenimines also have applications in *metal surface treatment*. They are used as brightener components for galvanic baths: Zinc, tin, copper, and other metals including alloys as well as passivations and finishes [47].

In *air purification*, polyethylenimines can be used as adsorbents for acidic gases, ozone and aldehydes.

Polyethylenimines have several applications in the drilling and completion of *oil* and *gas* wells and in the subsequent production of oil. Polyethylenimines are used in conjunction with a sulfonated polymer, such as lignosulfonate or a condensed naphthalenesulfonate, to inhibit fluid loss from well cement [48]. They have also been used commercially as selective flocculants in drilling fluid and as demulsifiers for some crude oil emulsions.

An important emerging technology is the use of polyethylenimine to *immobilize enzymes* [49].

The high affinity of polyethylenimine for proteins such as hair or skin opens the door to a variety of *cosmetic and personal care applications*. For these purposes, the original high charge density of the polyethylenimines can be tuned to the desired degree by chemical modifications, e.g., alkoxylation.

Polyethylenimines can also be used as additives in *detergents* [50].

7. Economic Aspects

The cost of producing ethylenimine depends on the price of 2-aminoethanol and on the high capital requirements needed to build a safe and efficient ethylenimine plant. The cost of a plant is high because of both the highly toxic nature of the ethylenimine and the lack of gasket material that can withstand prolonged contact with ethylenimine. This means that a modern ethylenimine production facility will be almost entirely of welded construction, using gasketless pumps, valves, and flow meters. Therefore, ethylenimine is significantly more expensive to produce than monomers used in most other water-soluble polymers. The price per kilogram of polyethylenimine depends on quantity. Small quantities of polyethylenimine sold for 11 – 18 €/kg in the last few years.

Ethylenimine is produced commercially by BASF (Germany) and Nippon Shokubai Kagaku Kogyo (Japan). Propylenimine has become available from Aceto Corp./Arsynco (USA), Dixie Chemical Company (USA), and Whyte Chemicals (UK). Sybron Chemicals (USA) started up their own propylenimine plant for the production of propylenimine-based crosslinkers.

8. Toxicology and Occupational Health [51 – 54]

Like many strong alkylating agents, aziridines are highly toxic. Their toxicity probably results from alteration of nucleic acids, proteins, and other important biochemicals by alkylation.

At one time substances with two or more aziridine rings, such as triethylenemelamine (**8**) [*51-18-3*], were considered promising antineoplastic drugs.

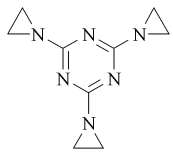

Many were prepared, a half dozen or more went to clinical trial, and some were used to treat cancer. However, although capable of inhibiting tumors and killing cancer cells, they were found to have serious side effects. These included severe damage to lymphatic cells and bone marrow and diminished fertility. Moreover, they often proved carcinogenic in animal tests. Because of these properties, they are rarely used as antineoplastic drugs today. However, the extensive testing of these agents, added to the testing of simple, commercially important aziridines, has provided more knowledge of the toxicological properties of aziridines than is available for most chemical classes. For example, the International Agency for Research on Cancer (IARC) has reviewed carcinogenic data of eleven aziridines

Table 2. Acute toxicity, doses in mg/kg

	EI	PI	Xama-7	Corcat P-600
Acute oral (LD_{50})	14	19	2000	7500
Acute dermal (LD_{50})	13	43	>3000	>3000
Eye irritation	severe	severe	moderate, transitory	irritant

[52]. For nine the data were sufficient to classify the substances as carcinogens. Clearly, any aziridines that have not been tested should be considered potential carcinogens.

8.1. Ethylenimine

In animal studies ethylenimine is highly toxic by ingestion, inhalation, skin or eye application, and injection. For example LD_{50} (oral, rat) is 14 mg/kg; LD_{50} (skin, rabbit) is 13 mg/kg; LC_{50} (10-min inhalation, mouse) is 2236 ppm. In inhalation studies, lethal air concentrations typically cause irritation of nose and eyes during exposure, with edema and other signs of lung injury becoming evident a few hours afterwards, and death ensuing a few days later.

To humans ethylenimine is a skin sensitizer and a severe irritant to eyes, nose, and throat. Transient CNS, liver, and kidney damage has been observed. Two deaths have been reported, one from inhalation and one from skin contact.

Ethylenimine is a mutagen in microbial and fruit fly tests. It proved carcinogenic by the oral and subcutaneous route in four mouse tests. However, an epidemiological study of 144 aziridine workers found no evidence of carcinogenicity.

Ethylenimine is classified by OSHA as a "cancer suspect" agent and detailed standards and work practices are prescribed to prevent any significant exposure by inhalation, ingestion, or skin or eye contact. The ACGIH recommends a TLV of 0.5 ppm, with skin contact to be avoided [53]. The DFG classifies ethylenimine as an A 2 carcinogen and a category 2 germ cell mutagen, requiring special protective and surveillance measures [54].

8.2. Propylenimine

The toxic properties of propylenimine are similar to those of ethylenimine. Limited studies in rats suggest it is one-fourth to one-eighth as toxic by inhalation, and this is part of the basis for the TLV of 2 ppm set by the ACGIH. In oral rat studies it produced a variety of malignant tumors. It is classified an A 2 suspected carcinogen by ACGIH, an A 2 proved animal carcinogen and a category 2 germ cell mutagen by DFG. It should be handled with the same care as ethylenimine.

8.3. Aziridine Derivatives and Polyethylenimines

Substances derived from aziridines but no longer containing the aziridine ring are normally of low toxicity. Examples are polymerized aziridines and polymers that have been modified by reaction with aziridines.

Polyethylenimine is used in personal care products and in food packaging materials. Table 2 shows toxicity data for ethylenimine (EI) and for propylenimine (PI) as well as for a typical polyfunctional aziridine (Xama-7) and a typical polyethylenimine (Corcat P-600).

FDA approval as an additive for food packaging on certain substrates exists for certain polyethylenimines (PEI; Lupasol grades) under the following parts:

- 173.357: PEI for enzyme immobilization
- 175.105: PEI for food packaging
- 175.320: PEI on polyolefin in contact with food
- 176.170: PEI in paper in contact with aqueous and fatty food
- 176.180: PEI in paper in contact with dry food
- 177.1200: PEI on cellophane in contact with food

9. References

General References

1. O. C. Dermer, G. E. Ham: *Ethylenimine and Other Aziridines*, Academic Press, New York and London 1969.
2. D. S. Zhuk, P. A. Gembitskii, V. A. Kargin, *Russian Chemical Reviews (Engl. Transl.)* **34** (1965) 515.
3. G. Scherr, U. Steuerle, R. Fikentscher: "Imines, Cyclic", in *Kirk-Othmer*, 4th ed., vol. **14**, pp. 2 – 40.

Specific References

4. S. Gabriel, *Ber. Dtsch. Chem. Ges.* **21** (1888) 1049.
5. P. Hirsch, *Ber. Dtsch. Chem. Ges.* **23** (1890) 964.
6. I.G. Farbenindustrie, US 2223930, 1940 (R. Griessbach, E. Meier, H. Wassenegger).
7. I.G. Farbenindustrie, US 2182306, 1936 (H. Ulrich, W. Harz).
8. Dow Chemical, US 3203910, 1962 (B. W. Wilson).
9. BASF Aktiengesellschaft, EP 873371, 1996 (U. Steuerle et al.).
10. BASF Aktiengesellschaft, EP 1177035, 2000 (U. Steuerle et al.).
11. Dow Chemical, US 3519687, 1970 (J. G. Schneider, C. R. Dick, G. E. Ham).
12. BASF Aktiengesellschaft, WO 97/40088, 1997 (U. Steuerle, W. Reuther, H. Meixner, M. Ehle, T. Greindl, R. Betz).
13. U.S. Secretary of Agriculture, US 3 885 069, 1975 (E. J. Roberts, S. P. Rowland).
14. BASF Aktiengesellschaft, DE 198 21 665 A1, 1998 (B. Camara, M. Patsch, U. Steuerle, J. Romeis).
15. T. St. Pierre, M. Geckle, *ACS Polym. Prepr.* **22** (1981) 128.
16. BASF Aktiengesellschaft, WO 97/21760, 1997 (U. Steuerle, W. Reuther, W. Harder).
17. Dow Chemical, US 3 565 941, 1968 (C. R. Dick, J. L. Potter, W. P. Coker).
18. National Starch & Chemical, US 3372149, 1968 (J. Fertig, E. D. Mazzarella, M. Skoultchi).
19. Dow Chemical, US 3280218, 1966 (D. J. Endsley, B. W. Wilson).
20. BASF Aktiengesellschaft, US 3642572, 1972 (H. Endres, R. Fikentscher, W. Maurer, E. Scharf, U. Soenksen).
21. BASF Aktiengesellschaft, DE 19955804, 1999 (J. Decker, S. Frenzel, M. Rübenacker, H. Meixner).
22. Dow Chemical, US 3463816, 1969 (G. F. Button, D. L. Childress).
23. BASF Aktiengesellschaft, EP 422 534, 1994 (A. Hohmann, W. Reuther, W. Bochnitschek).
24. J. H. Olin, *Ind. Eng. Chem.* **55** (1963) no. 9, 10.
25. Dow Chemical, US 3326895, 1967 (W. P. Coker).
26. H. Wenker, *J. Am. Chem. Soc.* **57** (1935) 2328.
27. Shell Oil Co., US 3153079, 1964 (E. R. A. Forshaw).
28. H. Kindler, W. Sanne, R. Sinn, A. Wittwer, *Chem. Ing. Tech.* **37** (1965) 400.
29. BASF Aktiengesellschaft, DE 1 302 658, 1971 (R. Sinn, W. Sanne, H. Kindler).
30. Chemirad, US 3326879, 1967 (R. G. Dunning).
31. A. A. Anderson, S. P. Simonyan, M. V. Shimanskaya, *Chem. Heterocycl. Compd. (Engl. Transl.)* **29** (1993) no. 10, 1134.
32. M. Ueshima, H. Tsuneki in S. Yoshida, N. Takezawa, T. Ono (eds.), *Catalytic Science and Technology*, vol.**1**, Kodansha, Tokyo, VCH Verlagsgesellschaft, Weinheim, Germany 1991 pp. 357 – 360.
33. M. Ueshima, et al. in L. Guczi, F. Solymosi, P. Tétényi (eds.): *New Frontiers in Catalysis*, Part C, Proceedings of the 10th International Congress on Catalysis, 19 – 24 July, 1992, Budapest, Hungary, pp. 2447 – 2450, Elsevier, 1993;
Chem. Abstr. **119** (1993) 273829z.
34. H. Tsuneki, Y. Shimasaki, K. Ariyosi, Y. Morimoto, M. Ueshima, *Nippon Kagaku Kaishi* **11** (1993) 1209 – 1216;
Chem. Abstr. **120** (1994) 55045v.
35. H. Tsuneki, *Applied Catalysis A: General* **221** (2001) pp. 209–217.
36. Nippon Shokubai, EP 385 714, 1990 (H. Tsuneki, K. Ariyoshi, A. Moriya, M. Ueshima).
37. A. M. DeRoo in I. Skeist (ed.): *Handbook of Adhesives*, 2nd ed., Van Nostrand Reinhold Co., New York 1977, pp. 592 – 596.
38. K. W. Britt, J. E. Unbehend, *Tappi* **63** (1980) 67.
39. D. Horn, F. Linhart in J. C. Roberts (ed.): *Paper Chemistry*, Blackie, London 1991, pp. 44 – 62.
40. H. Meixner, W. J. Auhorn, M. Gercke, *Das Papier* 1998, no. 10A, V36 – V41.
41. H. Yeh, M. M. Ghosh, *J. Am. Water Works Assoc.* **73** (1981) 211.
42. J. E. Harrar, F. E. Locke, C. H. Otto, Jr., L. E. Lorensen, S. B. Monaco, W. P. Frey, *Soc. Pet. Eng. J.* **22** (1982) 17.

43. DE 2 303 081, 1974 (E. Bayer, G. Lange).
44. K. Geckeler, G. Lange, H. Eberhardt, E. Bayer, *Pure Appl. Chem.* **52** (1980) no. 7, 1883.
45. H. K. Lonsdale, *J. Membrane Sci.* **10** (1982) 81.
46. Hoechst AG, US 5512064, 1996 (A. von der Eltz, J. Clauss, A. Schrell).
47. BASF Aktiengesellschaft, DE 2658934, 1978 (R. Fikentscher, K. Glaser, G. Gotsmann, W. Streit).
48. L. F. McKenzie, J. V. Eikerts, P. M. McElfresh, *Oil Gas J*. (1982) 146.
49. Miles Laboratories, US 4355105, 1982 (O. J. Lantero, Jr.,
50. Unilever PLC, EP 1003828, 1997 (E. Gutierrez, S.-R. Wu, U. Racherla, R. Vermeer).
51. H. J. Trochimowicz, G. L. Kennedy, Jr., N. D. Krivanek: "Aliphatic and Aromatic Nitrogen Compounds" in E. Bingham, B. Cohrssen, C. H. Powell (eds.): *Patty's Industrial Hygiene and Toxicology*, 5th ed., vol. 4, Wiley-Interscience, New York 2001, pp. 1107–1372.
52. IARC: *Evaluation of Carcinogenic Risk of Chemicals to Man*, vol. **9**, International Agency for Research on Cancer, Lyon 1975, pp. 31 – 109.
53. ACGIH (ed.): *Documentation of the Threshold Limit Values*, Cincinnati, Ohio 1983, pp. 186, 351.
54. DFG: *List of MAK and BAT Values - Maximum Concentrations and Biological Tolerance Values at the Workplace*, Wiley-VCH, Weinheim, Germany 2005, Report 41.

Azo Dyes

Klaus Hunger, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chaps. 1 – 4, 6 – 10)

Peter Mischke, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chaps. 1 – 4, 6 – 10)

Wolfgang Rieper, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chaps. 1 – 4, 6 – 10)

Roderich Raue, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

Klaus Kunde, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

Aloys Engel, DyStar Textilfarben GmbH & Co. Deutschland KG, Leverkusen, Federal Republic of Germany (Chap. 5)

1. Introduction 2
2. Production 3
2.1. Diazotization and Coupling 4
2.1.1. Diazotization 4
2.1.2. Coupling 7
2.2. Other Azo Group Syntheses 11
2.2.1. Condensation of Nitro Compounds with Amines 11
2.2.2. Reduction of Nitro Compounds . . 12
2.2.3. Oxidation of Amino Compounds . . 13
2.3. Synthesis from Azo Compounds . 13
2.3.1. Exposure of Concealed or Protected Amino Groups 13
2.3.2. Acylation of Amino Azo Compounds 14
2.3.3. Alkylation and Acylation of Phenolic Hydroxyl Groups 14
2.3.4. Metal-Complex Formation 15
2.4. Processes without Industrial Importance 15
2.4.1. Reduction of Diazonium Compounds 16
2.4.2. Oxidative Coupling 16
2.4.3. Dehydrogenation of Diarylhydrazines 17
2.4.4. Reaction of Arylhydrazines with Quinones 17
2.4.5. Condensation of Nitroso Compounds with Amines 18
2.4.6. Azo Compounds from Azido Compounds 18
2.5. Equipment for Industrial Production of Azo Dyes 19
3. Anionic Azo Dyes 20
3.1. Introduction 20
3.2. Acid Azo Dyes 20
3.2.1. Wool Dyes 21
3.2.1.1. Monoazo Dyes 22
3.2.1.2. Disazo Dyes 25
3.2.1.3. Chrome Dyes 28
3.2.1.4. Metal-Complex Dyes 30
3.2.2. Polyamide Dyes 31
3.2.3. Silk Dyes 33
3.2.4. Leather Dyes 34
3.2.5. Azo Food Dyes 35
4. Direct (Substantive) Dyes 37
4.1. Conventional Direct Dyes 38
4.1.1. Monoazo Dyes 38
4.1.2. Disazo Dyes 40
4.1.2.1. Primary Disazo Dyes 40
4.1.2.2. Secondary Disazo Dyes 43
4.1.3. Trisazo Dyes 43
4.1.4. Tetrakisazo Dyes 44
4.1.5. Condensation Dyes 45
4.1.6. Direct Dyes with a Urea Bridge . . 46
4.1.7. Triazinyl Dyes 46
4.1.8. Copper Complexes of Substantive Azo Dyes 48
4.2. Direct Dyes with Aftertreatment . 48
4.2.1. Aftertreatment with Cationic Auxiliaries 49
4.2.2. Aftertreatment with Formaldehyde 50
4.2.3. Diazotization Dyes 50
4.2.4. Aftertreatment with Metal Salts . . 50
5. Cationic Azo Dyes 51
5.1. Cationic Charge at the Coupling Component 52
5.1.1. Polyamines as Coupling Components 52
5.1.2. Heterocycles as Coupling Components 53
5.1.3. Coupling Components with Alkylammonium Groups 54

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_245

5.1.3.1. Coupling Components with Dialkylaminoalkyl Groups 54
5.1.3.2. Aromatically Linked Trialkylammonium Groups 54
5.1.3.3. Trialkylammoniumalkyl-Substituted Anilines 55
5.1.3.4. Other Trialkylammoniumalkyl-Substituted Coupling Components . 56
5.1.4. Coupling Components with Dialkylhydrazinium Groups 57
5.1.5. Coupling Components with Cyclic Ammonium Groups 57
5.1.5.1. Heterocyclically Linked Cyclic Ammonium Groups 57
5.1.5.2. Aliphatically Linked Cyclic Ammonium Groups 58
5.1.6. Coupling Components with Condensed Cyclic Ammonium Residues 59
5.1.7. Cyclic Ammonium Residues at the Azo Group 59
5.1.8. Coupling Components with two Different Cationic Residues 59
5.2. Cationic Charge in the Diazo Component 61
5.2.1. Diazo Components with Aminoalkyl Moieties 61
5.2.2. Diazo Components with Trialkylammonium Residues 61
5.2.2.1. Aromatically Linked Trialkylammonium Residues 62
5.2.2.2. Aliphatically Linked Trialkylammonium Residues 62
5.2.3. Diazo Components with Cyclic Ammonium Residues 64
5.2.3.1. Aromatically Linked Cyclic Ammonium Residues 64
5.2.3.2. Aliphatically Linked Cyclic Ammonium Residues 64
5.2.4. Diazo Components with two Different Cationic Residues 64
5.2.5. Diazo Components with Aromatic Condensed Cyclic Ammonium Residues 65
5.2.6. Cyclic Ammonium Compounds Linked in Meta Position to the Azo Group 66
5.3. Different Cationic Charges in Both the Coupling and the Diazo Component 66
5.4. Introduction of Cationic Substituents into Preformed Azo Dyes 66
5.5. Cationic Dyes with Sulfur or Phosphorus as Charge-Carrying Atoms 68
5.6. Dyes with Releasable Cationic Groups 68
6. Developing Dyes 69
6.1. Developing Dyes for Cotton 70
6.1.1. Developing Dyes for Dyeing 70
6.1.1.1. β-Naphthol Dyes 70
6.1.1.2. Naphtol AS Dyes 70
6.1.2. Developing Dyes for Printing 77
6.2. Developing Dyes for Animal Fibers 85
6.3. Developing Dyes for Hydrophobic Fibers 85
7. Disperse Azo Dyes 86
8. Azo Pigments 86
9. Alcohol- and Ester-Soluble Dyes . 86
10. Fat- and Oil-Soluble Dyes 87
11. References 89

1. Introduction [1–6]

Azo dyes are characterized by a chromophoric azo group −N=N−, whose nitrogen atoms are linked respectively to sp^2-hybridized carbon atoms. At least one of these carbon atoms belongs to an aromatic carbocycle (usually a benzene or naphthalene derivative) or heterocycle (e.g., pyrazolone, thiazole), whereas the second carbon atom adjoining the azo group may also be part of an enolizable aliphatic derivative (e.g., acetoacetic acid). The most common types of azo dyes can thus be summarized as follows: aryl−N=N−R, where R can be an aryl, heteroaryl, or −CH=C(OH)−alkyl.

Due to the simple nature of the synthesis, usually in aqueous medium, and the almost unlimited choice of starting products, an extremely wide variety of azo dyes is possible. The number of combinations is further increased by the fact that a dye molecule can contain several azo groups.

This diversity of inexpensively produced azo dyes permits a wide spectrum of shades and fastness properties suitable for use on a variety of substrates.

Naturally occurring azo dyes are unknown (some containing azoxy groups are known), but the azo dyes represent the most numerous and widely manufactured synthetic dyes.

Nomenclature. Depending on the number of azo groups, the azo dyes are called monoazo, disazo, trisazo, tetrakisazo, etc., dyes, and those with three or more azo groups, polyazo dyes.

The structure of an azo dye can be described by means of a structural formula (a), IUPAC nomenclature (b), or practical nomenclature (c), e.g.:

a)

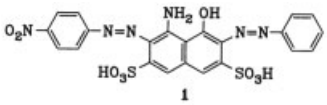

b) 1-hydroxy-2-phenylazo-7-(4′-nitrophenylazo)-8-aminonaphthalene-3,6-disulfonic acid

c) general formulation: D $-^1 \rightarrow$ K $\leftarrow^2-$ D, in words: *p*-nitroaniline $-^1 \rightarrow$ H acid $\leftarrow^2-$ aniline and in reaction equations:

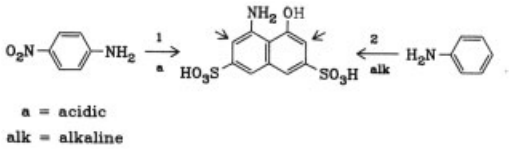

Designation (b) is less commonly used because of its complexity. The practical nomenclature (c) makes use of a genetic scheme of the technically most common synthesis, i.e., it characterizes the azo dye by the starting compounds (D = diazo component and K = coupling component) and the direction of coupling. This nomenclature is shorter and is commonly used in industry and technical literature. In the case of disazo and polyazo dyes, the coupling direction is specified by arrows, the sequence of operations by appropriate numbers above the arrows, and the coupling conditions by the abbreviations a (acid) and alk (alkaline). Additional arrows at the coupling components show the linkage points of both components across the azo groups. All industrially important dyes are also classified according to the Colour Index [7]. This catalog of dyes gives information about trade names, coloristic properties, and, if published, the chemical constitution. The dyes are classified according to color and chemical properties. Aside from its trade names, the azo dye in formula **1**, for example, can be found as C. I. Acid Black 1 or C. I. 20470.

Dyes listed as examples in the following sections are cited according to the internationally used designations in the Colour Index. Trade names are compiled in tabular form only for the individual dye classes; historical common names are for the most part included in the text.

Classification. Azo dyes can be classified either according to chemical guidelines (characteristic chemical groups) or by color aspects (application in dye works). A rigid assignment to one of the two classification systems is possible, but not advisable because of overlapping. There is therefore scarcely a chemical class of azo dyes that belongs to a single color class or vice versa. Moreover, many materials can be dyed with dyes from various color groups. In this article, a compromise between the two classification systems with emphasis on application in the dye works seemed to be the most appropriate.

The importance of the azo dyes is due to their mode of application and to the fact that they represent a clearly defined concept. They are briefly described here and dealt with more extensively under special headings: → Disperse Dyes; → Metal-Complex Dyes; → Pigments, Organic; → Reactive Dyes.

2. Production

For additional information see [1–6] and [8–12].

The processes that are important in the production of azo dyes are classified as follows:

Processes involving synthesis of the azo group:

- diazotization and coupling
- condensation of nitro compounds with amines
- reduction of nitro compounds
- oxidation of amino compounds

Syntheses with compounds already containing the azo group:

- exposure of concealed or protected amino groups
- acylation of aminoazo compounds
- alkylation and acylation of phenolic hydroxyl groups
- metal-complex formation

2.1. Diazotization and Coupling

Of all the industrially important processes azo coupling, during which the azo group is synthesized, is by far the most important. All other industrially employed processes are advantageous only where azo coupling cannot be utilized, for example, because of the unavailability of the starting compounds.

Azo coupling can be summarized as follows: an aromatic amine is linked to a nucleophilic partner RH (coupling component) in the presence of a nitrosyl-eliminating compound XNO to form an aromatic azo compound. The reaction equation is as follows:

$$\mathrm{Ar{-}NH_2 + XNO + RH \xrightarrow{-\,H_2O} Ar{-}N{=}N{-}R + HX} \qquad (1)$$

$$\mathrm{X = Cl,\ Br,\ NO_2,\ HSO_4}$$

Equation (1) describes the overall azo coupling reaction, which consists of two major steps: formation of the diazonium compound by means of the so-called diazotization and synthesis of the azo dye, the actual azo coupling.

2.1.1. Diazotization

Diazotization is defined as the reaction of primary aromatic amines with nitrites, preferably with sodium nitrite, in a usually aqueous mineral acid solution at around 0 °C, whereby the amine is converted into the corresponding diazonium salt:

$$\mathrm{Ar{-}NH_2 + 2\ HY + NaNO_2 \longrightarrow Ar{-}N{\equiv}\overset{+}{N}\ Y^- + NaY + 2\ H_2O} \qquad (2)$$

$$\mathrm{Y = Cl,\ HSO_4,\ etc.}$$

Weakly basic amines require a higher acid concentration, since diazoamino compounds Ar−N=N−HN−Ar may otherwise form. Very weakly basic amines, i.e., amines with several negative (−I/−M) substituents (e.g., tetrahalogen anilines, di- and trinitroanilines, halogennitroanilines) can only be diazotized after being dissolved in concentrated sulfuric acid and subsequently reacted with nitrosylsulfuric acid, HSO_4NO.

A further reason for using concentrated acids (concentrated sulfuric acid) is the fact that diazonium salts of weakly basic amines are readily hydrolyzable in dilute acids.

Diazotization of aromatic diamines gives bisdiazonium compounds and therefore is called bisdiazotization. The incorrect term "tetrazotization" is frequently found in the technical literature.

The essential step in diazotization is the electrophilic nitrosation of the amino group of the primary aromatic amine (**2**).

Formation of the diazonium ion (**4**) then follows via the diazohydroxide (**3**):

$$\underset{\mathbf{2}}{\mathrm{Ar{-}NH_2}} + \mathrm{X{-}N{=}O} \xrightarrow{-\,HX} \mathrm{Ar{-}\overset{H}{\overset{|}{N}}{-}N{=}O} \longrightarrow \underset{\mathbf{3}}{\mathrm{Ar{-}N{=}N{-}OH}}$$

$$\underset{OH^-}{\overset{H^+}{\rightleftharpoons}} \underset{\mathbf{4}}{\mathrm{Ar{-}N{\equiv}\overset{+}{N}}} + \mathrm{H_2O} \qquad (3)$$

$$\mathrm{X = Cl,\ Br,\ NO_2,\ HSO_4}$$

Reaction sequence (3) is a simplification; in fact, the reaction rate during diazotization is determined by the formation of the nitrosating agent (at low acidity) or nitrosation of the amine (at high acidity). In a strongly acid medium, i.e., in the absence of free amines, the protonated amine surprisingly becomes the nucleophile:

$$\mathrm{ArNH_3^+} \overset{NO^+}{\rightleftharpoons} \mathrm{Ar{-}\overset{+}{N}H_2{-}NO} + \mathrm{H^+}$$

$$\mathrm{B} \Downarrow \Uparrow \qquad (4)$$

$$\mathrm{ArNH{-}NO + HB^+ + H^+}$$

The equilibrium reactions preceding nitrosation require excess acid during diazotization leading to the formation of the nitrosating agent:

$$\mathrm{2\ HONO \rightleftharpoons N_2O_3 + H_2O} \qquad (5)$$

$$\mathrm{HONO + H_3O^+ \rightleftharpoons H_2\overset{+}{O}NO + H_2O} \qquad (6)$$

$$\mathrm{HONO + H^+ \rightleftharpoons NO^+ + H_2O} \qquad (7)$$

N_2O_3 is the nitrosating agent in a weakly acid medium (Eq. 5), H_2ONO in a more acid medium (Eq. 6), and the nitrosonium ion NO^+ in a strongly acid medium (Eq. 7).

An excess of sodium nitrite must be avoided, since nitrite can react with the coupling agent as well as with secondary or tertiary amines to form nitroso compounds during subsequent coupling. Excess nitrite can be detected with potassium

iodide–starch paper. It is eliminated by adding amidosulfuric acid or urea:

$$H_2NSO_3H + HNO_2 \longrightarrow H_2SO_4 + N_2 + H_2O$$

$$H_2NCONH_2 + HNO_2 \longrightarrow NH_3 + CO_2 + N_2 + H_2O$$

Diazo Components. Diazo components for the production of azo dyes can be divided into the following groups of aromatic amines:

Aniline and Substituted Anilines. Methyl, chlorine, nitro, methoxy, ethoxy, phenoxy, hydroxyl, carboxy, carbalkoxy, carbonamide, and sulfonic acid groups are primarily used as substituents.

Examples: Toluidines, nitroanilines, aminobenzoyl amides, and the sulfonic acids of these compounds.

Naphthylamines and Naphthylamine Sulfonic Acids. The following compounds, known by their common names, are examples of naphthylamine sulfonic acids:

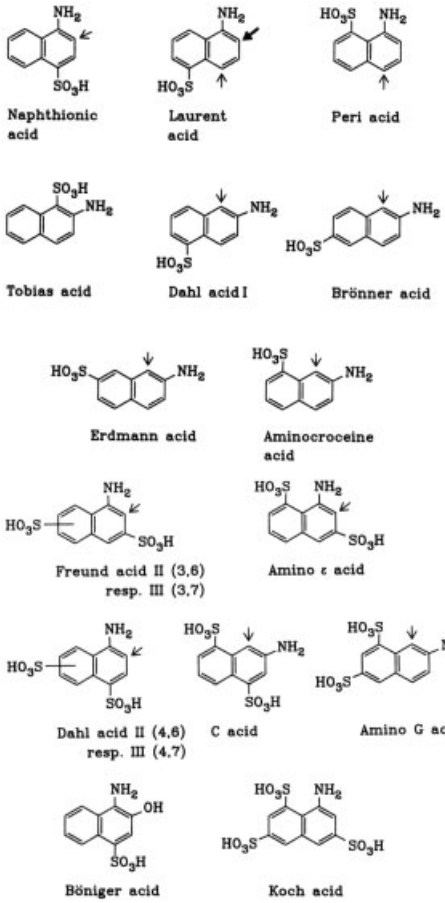

Many of these compounds can also be used as coupling components (see Section 2.1.2), as shown by the arrow indicating the coupling position.

Diamines $H_2N{-}A{-}NH_2$. The diamines can be divided into four groups:

a) A = phenylene or naphthylene radical, also substituted, e.g., by methyl, chlorine, nitro, methoxy, or sulfonic acid groups. Examples are:

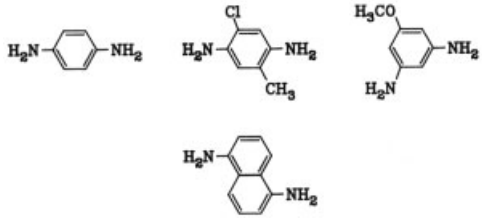

b) A = diphenyl radical, also substituted, mainly by methyl, chlorine, methoxy, carboxylic acid, or sulfonic acid groups. Examples are:

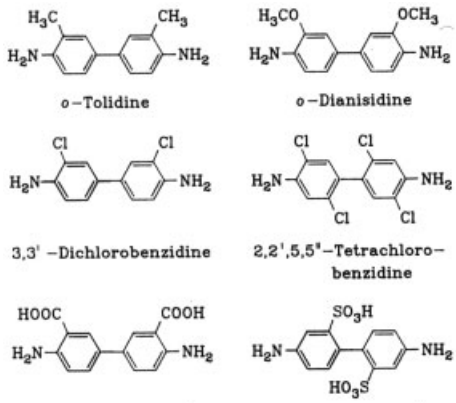

c) A = 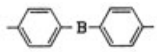

whereby the aromatic rings can also be substituted by methyl, chlorine, methoxy, carboxylic acid, or sulfonic acid groups. B = oxygen, sulfur, $-NH-$, $-SO_2-$, $-N{=}N-$, $-CH{=}CH-$, $-NHCO-$, $-NH-CO-HN-$. Examples are:

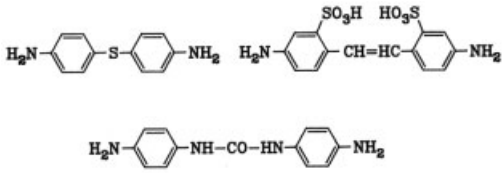

d) Heterocyclic amines and diamines. Examples are:

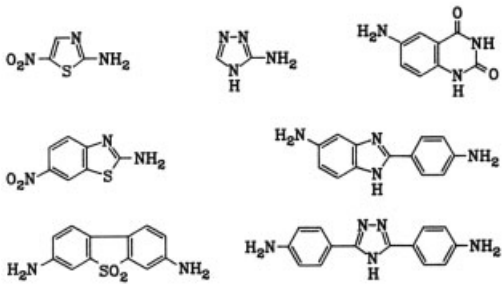

Diazotizing Methods. Considering the basicity and solubility of the amines being diazotized, the following diazotization methods find industrial use (for further details → Diazo Compounds):

a) *Direct Diazotization.* The primary aromatic amine is dissolved or suspended in aqueous hydrochloric or sulfuric acid to which an aqueous, concentrated sodium nitrite solution is added. An excess of 2.5 to 3 equivalents of acid per equivalent of amine is used. A temperature of 0 to 5 °C is maintained by adding ice.
b) *Indirect Diazotization.* Amines with sulfonic or carboxylic acid groups are often difficult to dissolve in diluted acid. Diazotization therefore takes place as follows: the amine is dissolved in water or weak alkalis, and the calculated amount of sodium nitrite solution is stirred into the ice-cooled acid solution already in the vessel. The acid can also be added to the amine–nitrite mixture already at hand.
c) *Diazotization of Weakly Basic Amines.* Weakly basic amines are dissolved in concentrated sulfuric acid and diazotized with nitrosylsulfuric acid, which is easily prepared from solid sodium nitrite and concentrated sulfuric acid.
d) *Diazotization in Organic Solvents.* The water-insoluble or only slightly soluble amine is dissolved in glacial acetic acid or other organic solvents and, where necessary, diluted with water. After the addition of acid it is diazotized in the usual manner with sodium nitrite solution. Nitrosylsulfuric acid, nitrosyl chloride, alkyl nitrites, or nitrous gases can also be used to eliminate the nitrite. Temperature, pH, and the concentration of the diazotizing solution often have a considerable effect on the progress of diazotization. Physical properties (distribution, particle size) and the addition of emulsifiers and dispersing agents influence the diazotization of slightly soluble amines.

Certain aromatic amines require special diazotizing processes:

1-Aminonaphthols, such as 1-amino-2-naphthol and several *1-aminonaphtholsulfonic acids*, such as 1-amino-2-naphthol-4-sulfonic acid, are oxidized to the respective quinone by nitrous acid. Diazotization can however take place under normal conditions in the presence of catalytic quantities of metal salts, such as copper or zinc salts.

o-Diamines, such as 1,2-phenylenediamine or 1,8-naphthylenediamine, undergo cyclization during the normal diazotization process to form triazoles:

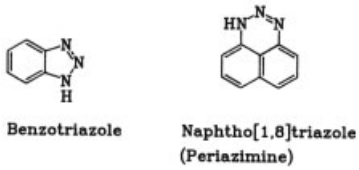

The desired bisdiazotization with *o*-phenylenediamine, as well as with *m*- and *p*-phenylenediamine, is achieved in glacial acetic acid with nitrosylsulfuric acid. 1,8-Naphthylenediamine can be bisdiazotized in excess acid.

Under normal diazotization conditions, various *ortho*-substituted aromatic amines also undergo more or less complete cyclization to form benzoheterocycles that are no longer accessible to subsequent coupling.

Examples are:

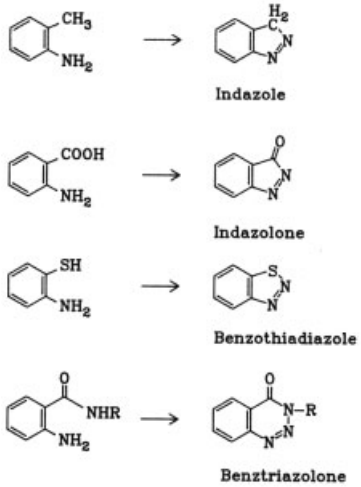

Diazonium compounds are generally stable only in aqueous solution at low temperatures. When heated, they frequently decompose by eliminating nitrogen to form the corresponding phenol. Some amines, however, can be diazotized at temperatures up to 40 °C. Light and heavy-metal ions also accelerate the decomposition of diazonium compounds. Diazotization, therefore, is usually carried out in wooden vats or iron stirring vessels with an acid-proof lining or rubber coating (see Section 2.5).

As solids most diazonium compounds are stable only to a degree and frequently sensitive to heat, impact, and shock (explosive) (→ Diazo Compounds).

2.1.2. Coupling

Coupling Sequence. The actual azo coupling reaction consists of an electrophilic substitution reaction of the diazonium compound with a nucleophilic partner (coupling component RH):

$$\mathrm{Ar-N^{+}\equiv NY^{-}+R-H \longrightarrow Ar{-}N=N-R+HY} \qquad (8)$$

Coupling components are aromatic systems with nucleophilic centers at the aromatic nucleus, particularly phenols, naphthols, and amines or enolizable compounds with reactive methylene groups. Phenols react as phenolates, naphthols as napththolates, and amines as free bases. According to Equation (8), free acid is formed during the coupling reaction. In order to maintain an optimal reaction sequence, the pH must be kept constant by adding alkalis or buffers. Coupling under strongly alkaline conditions is not possible. According to Equation (3), the diazonium compound undergoes a reverse reaction to form an *anti*-diazotate, which cannot be coupled. Under strongly alkaline conditions and with amines as coupling agents, the diazonium component may react with the amine nitrogen to form the diazoamino compound R−N=N−HN−Ar. Phenols and enols are therefore generally coupled in the weakly alkaline range (pH 7 – 9), amines mainly in the weakly acid range (pH 4 – 7).

In the production of developing dyes for textile printing the formation of diazoamino compounds is desired (see Section 6.1.2, → Diazo Compounds).

Substituents with donor effects in the aromatic nucleus of the coupling components increase the reactivity. Conversely, donor substituents on the lower the reactivity, whereas substituents with acceptor effects increase it. The coupling energy of substituted anilines as diazo components decreases as follows:

polynitroanilines > nitrochloroanilines > nitroanilines > chloroanilines > anilinesulfonic acids > aniline > anisidines > aminophenols.

The C-atom with the highest electron density is usually the preferred coupling position of a coupling component. Due to the directing influence of hydroxyl or amino groups in aromatic systems, coupling takes place at the *ortho* or *para* position. If these two positions are occupied, there is either no coupling or one of the substituents is exchanged. Coupling never occurs at the meta position in relation to the directing substituent. Components of the naphthalene groups generally couple more easily than benzene derivatives.

Amines, which tend to form diazoamino compounds, can easily be coupled at the *para* position if they have been reacted with formaldehyde and sodium hydrogen sulfite to form *N*-methylenesulfonate [13]:

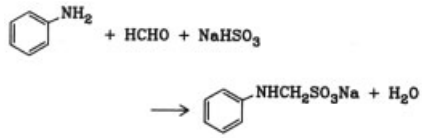

p-Aminoazo dyes are obtained by splitting methylenesulfonate from the coupling products under acid or alkaline conditions.

Other reaction conditions besides pH also play an important role in azo coupling. Increased temperature generally has a greater effect on the decomposition of the diazonium salt according to Equation (9)

$$R-N^{+}\equiv N\,Cl^{-}+H_2O \longrightarrow ROH+N_2+HCl \qquad (9)$$

than on the rate of coupling. The advantages are therefore limited. Coupling can be accelerated by raising the pH and the concentration of the reactants. In some cases the yield of azo dye can be increased by adding sodium chloride prior to the azo coupling. Some azo dyes can only be satisfactorily manufactured by adding pyridine or pyridine homologues; examples are the coupling of diazotized aminoazo compounds and aminodisazo compounds, as well as *o*-aminophenols for the production of complex forming dyes (→ Metal-Complex Dyes).

Pyridine acts as proton acceptor in the electrophilic coupling reaction and is especially beneficial when bulky substituents are present at the *ortho* or *peri* position in relation to the coupling position of the intermediate product or when the diazonium ion (e.g., in the case of *o*-diazophenols) exhibits a slight electrophilic reactivity.

Coupling Components. The most important coupling components can be divided into the following groups. Where there are several possibilities for coupling, the preferred coupling positions are marked by bold arrows and the other possible coupling positions by ordinary arrows.

Anilines, Diaminobenzenes. Examples are:

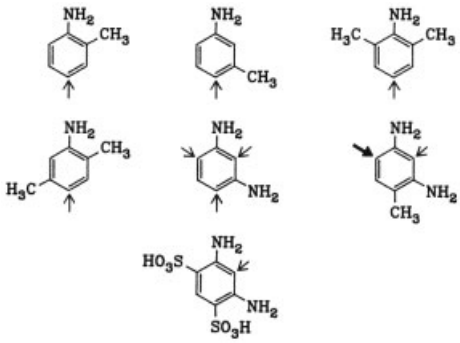

When reacted with aniline, diazotized aniline (benzenediazonium chloride) yields, apart from a small quantity of 4-aminoazobenzene (**5**), diazoaminobenzene (**6**) as main product:

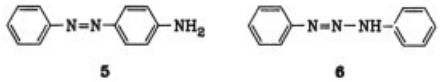

4-Aminoazobenzene (**5**) is obtained by coupling in a more strongly acid medium, but better still by heating diazoaminobenzene (**6**) in aniline with addition of aniline hydrochloride.

Electron donor substituents in the aniline, such as methyl or methoxy groups, especially at the meta-position, promote the readiness to couple; the inclination to couple thus increases in the sequence aniline $<$ *o*-toluidine $<$ *m*-toluidine $<$ *m*-anisidine $<$ cresidine $<$ 1-amino-2,5-dimethoxybenzene (aminohydroquinone dimethyl ether) to the extent that without formation of the diazoamino compound, the last three bases are attached almost quantitatively in the desired direction, i.e., at the 4-position in relation to the amino group.

m-Phenylene diamines couple to form monoazo dyes (chrysoidines, see Section 5.1.1) or disazo dyes.

Naphthylamines, Naphthylaminesulfonic Acids. Some examples are mentioned here; other compounds can be found among the diazo components (see Section 2.1.1), because most naphthylaminesulfonic acids are capable of being employed both for diazotization and for coupling purposes.

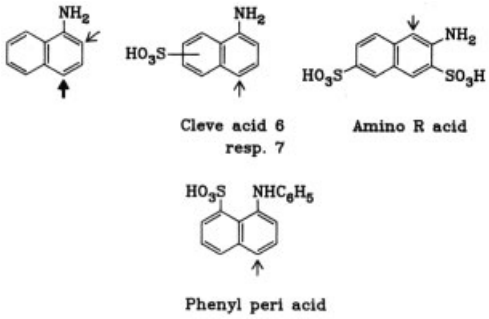

Whereas β-naphthylaminesulfonic acids always couple at the adjacent α-position, with the α-naphthylaminesulfonic acids, the coupling location is influenced by the position of the sulfonic acid group: 1,6-, 1,7-, and 1,8-naphthylaminesulfonic acids couple at the 4-

position; 1,5-naphthylaminesulfonic acid (Laurent acid) only couples with very "strong couplers" (e.g., 2,4-dinitroaniline), mainly at the 4-position. With diazotized aniline, chloroaniline, or diazotized anilinesulfonic acids the 2-aminoazo dyes are obtained, but with diazotized nitroaniline, mixtures of the 2- and 4-coupling products form:

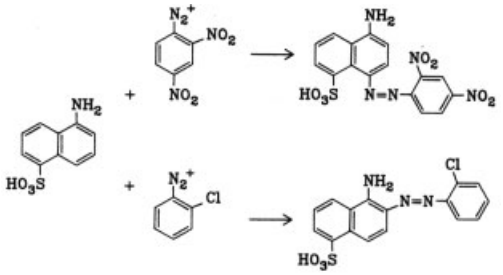

Phenols, Naphthols. Examples are:

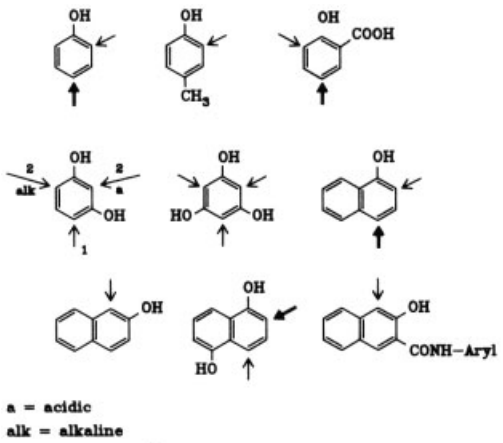

Phenols mainly couple at the 4-position, or at the 2-position if the 4-position is occupied. *p*-Hydroxybenzoic acid couples with elimination of CO_2; resorcinol couples twice: initially at the 4-position, with a second equivalent diazonium compound at the 2-position under acid conditions or at the 6-position under alkaline conditions. α-Naphthols mainly couple at the 4-position, in addition to which varying quantities of 2- and 2,4-coupling products are obtained, depending on the diazo component. β-Naphthol couples at the 1-position. Substituents in the 1-position, such as $-SO_3H$, $-COOH$, $-Cl$, $-CH_2OH$, or $-CH_2N(alkyl)_2$, may be eliminated during the coupling process. 1-Methyl-2-naphthol forms no azo dye.

Coupling components in this series that are of particular industrial importance are 2-hydroxynaphthalene-3-carboxylic acid arylamides, in particular anilides, which likewise couple at the 1-position (see Section 6.1.1.2).

Phenols and naphthols generally couple more easily and rapidly than amines; phenol-3-sulfonic acid can be coupled, but not aniline-3-sulfonic acid.

The readiness of the phenols to couple increases as the number of hydroxyl groups rises, as, for example, in the following series:

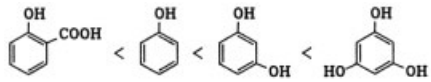

Naphtholsulfonic Acids. Examples are:

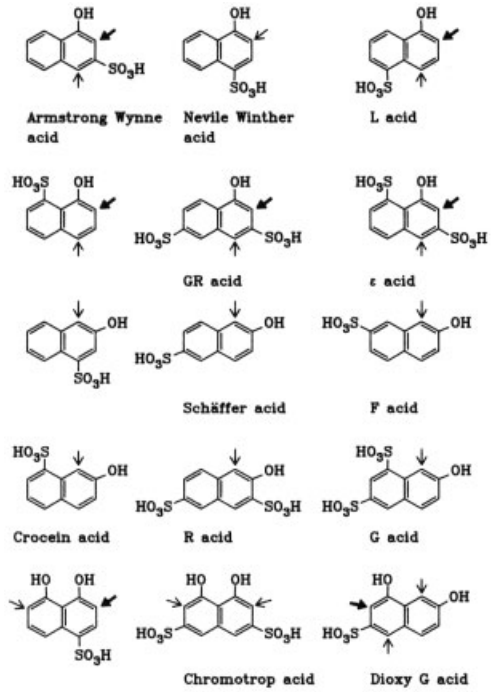

1-Naphtholsulfonic acids mainly couple at the 2-position. The 4-coupling products obtained as byproducts have to be carefully removed from the azo dyes, because unlike the 2-substitution products, their shade changes as a function of the pH value (shade intensification with rising pH due to formation of phenolate or

naphtholate resonance structures; see also Section 3.2.1).

Aminophenols, Aminophenolsulfonic Acids, Aminonaphtholsulfonic Acids. Examples are:

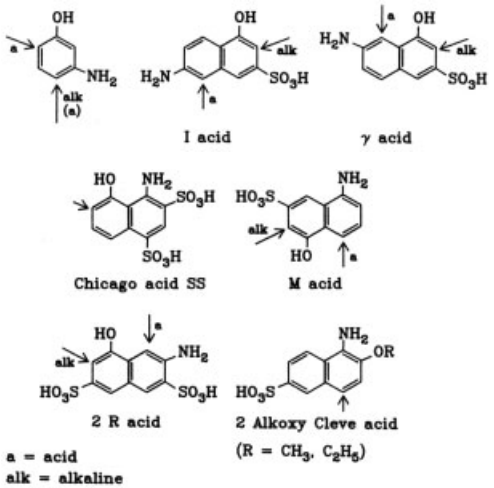

In an alkaline medium *m*-aminophenol couples at the *para* position in relation to the hydroxy group, in an acid solution *p*-hydroxy and *p*-aminoazo dyes are obtained side by side.

In the case of aminonaphtholsulfonic acids, orientation is greatly influenced by the pH value of the coupling medium. For instance, in an acid medium the azo group enters into the position *ortho* to the amino group, in an alkaline medium, in the position *ortho* to the hydroxyl group. 1-Amino-8-naphtholsulfonic acids with free *ortho* positions can couple with two equivalents of diazonium compound, coupling initially taking place in the acid medium at the o-amino position, followed by coupling in the alkaline range at the *o*-hydroxy position. Aminonaphtholsulfonic acids (*o*-hydroxyazo dyes) initially coupled in an alkaline medium cannot normally undergo further coupling to form disazo dyes:

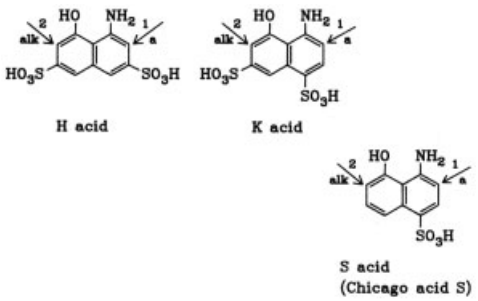

However, if this initial coupling is followed by a metal-complexing process, further coupling is possible in the neutral or alkaline range.

Compounds with Reactive Methylene Groups. In this series the coupling components with the greatest industrial importance are the *N*-acetoacetyl derivatives of aromatic amines (acetoacetic arylides):

$CH_3CO–CH_2CO–NH–Ar$

Anilines substituted by halogen, alkyl, alkoxy, nitro, and acylamino groups are the aromatic amines mainly suitable.

Example:

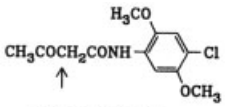

Bis(acetoacetylamino) derivatives, such as Naphtol AS–G, also enjoy practical importance.

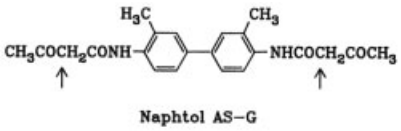

The following coupling components are also listed:

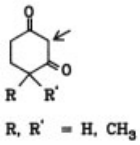

Heterocyclic Components. The most important compounds in this range are the 5-pyrazolones substituted at the 3-position.

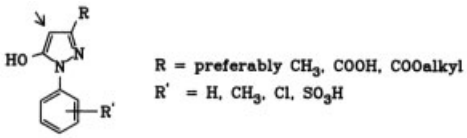

Further examples are:

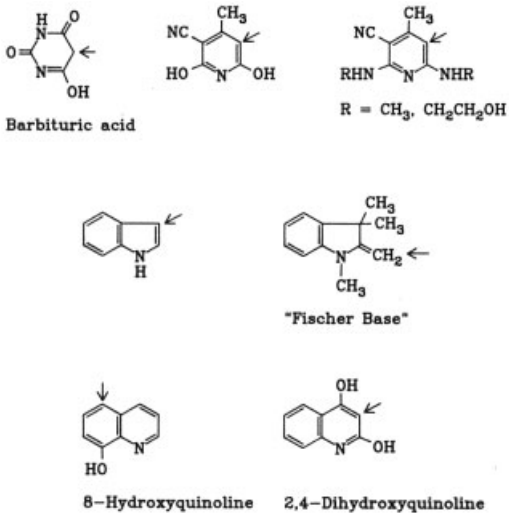

Azo Coupling in Practice. The optimum coupling conditions depend first of all on the nature of the diazo and coupling components used. The acid diazonium salt solution is generally allowed to enter the solution of the coupling component. Because, in addition to acid derived from the diazotization process, additional acid is released during coupling, the optimum pH value must be adhered to by adding bases. Alkali hydroxides, sodium carbonate, sodium hydrogen carbonate, ammonia, calcium carbonate, magnesium oxide etc., are used for this purpose. Such substances as sodium acetate or formiate (weakly acid) or sodium phosphate (weakly alkaline) can also be added to buffer the acid. These substances are added to the reaction mixture before, during, or after combining the components. Because at the feed point of the diazonium salt solution the pH value is always different than after thorough mixing has been completed, the type of stirrer and speed of stirring are also important in many instances. The sequence in which the two components are combined can also greatly influence the result.

The coupling reaction may be completed immediately after the components are mixed or after several hours. If the reaction requires a longer time, it is advisable to cool with ice and avoid exposure to dazzling light. In order to check whether excess diazonium compound is still present, a drop of reaction solution is spotted onto filter paper together with a component that couples easily (e.g., weakly alkaline H acid solution). If no coloration appears, the coupling is completed. The presence of unconsumed coupling component can be determined by spotting with a diazonium salt solution.

Attention must also be paid to the volume of coupling solution or suspension. With starting components of low solubility, the physical state is an important factor. In order to achieve complete reaction of diazo or coupling components with low solubility, it is often necessary to ensure that the reactants are distributed as finely as possible. This is carried out, for example, by adding dispersing agents to the diazo component or to the coupling mixture or by adding emulsifiers during acid precipitation of the coupling component prior to azo coupling. In each case, the most favorable reaction conditions must be precisely established and, because of the various influences that result in undesirable side effects, they must be carefully adhered to during manufacture.

See Section 2.5 for suitable apparatus.

2.2. Other Azo Group Syntheses (See also Chap. 2.4)

2.2.1. Condensation of Nitro Compounds with Amines [2, p. 339]

Aromatic azo and polyazo compounds can be produced by condensation of nitro compounds with amines in accordance with the following equation:

$$R^1\text{–}NO_2 + H_2N\text{–}R^2 \longrightarrow R^1\text{–}N=N\text{–}R^2 + H_2O\,(+O)$$

Because part of the amine is oxidized to the symmetrical azo compound $R^2-N=N-R^2$, an excess of amine is expedient. In some condensation processes it has been possible to observe the partial formation of azoxy bridges. Reaction takes place on heating the reactants for several hours to 40 – 120 °C in aqueous sodium hydroxide solution.

Substituents in the aniline component with electron-donor effects accelerate condensation; electron acceptors retard it. In the nitro component, the substituents have the opposite effect.

Of interest as regards the industrial manufacture of dye intermediates are condensation reactions of 4,4′-dinitrostilbene-2,2′-disulfonic acid

with aromatic amines. If specific sodium hydroxide solution concentrations, temperatures, and reaction times are precisely observed, it is possible with 1,4-diamines, 4-aminophenols, and their sulfonic and carboxylic acids to produce homogeneous monoazo and disazo dyes:

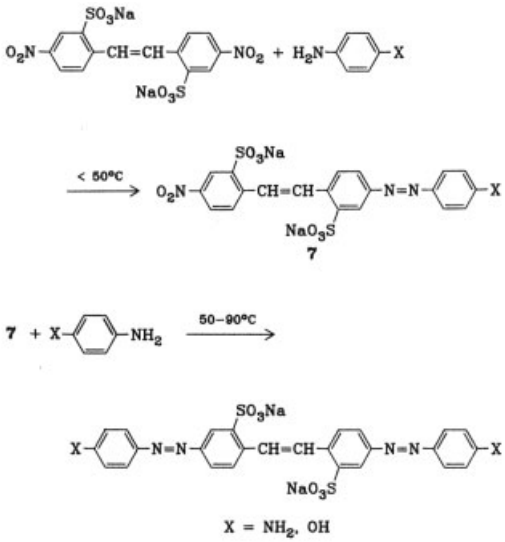

Of importance as regards the industrial production of the important direct cotton dyes for orange, scarlet, and brown shades are condensation reactions of 4,4′-dinitrostilbene-2,2′-disulfonic acid with aminoazo compounds (see Section 4.1.5). For example, by heating this acid with an equivalent of 4-aminoazobenzene-4′-sulfonic acid in aqueous sodium hydroxide solution, the disazo dye **8** is obtained:

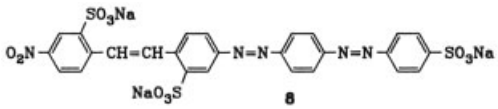

Condensation with two equivalents of 4-aminoazobenzene-4′-sulfonic acid at 105 °C in 5 % sodium hydroxide solution results in the corresponding tetrakisazo dye.

The dye synthesis is often followed by an oxidative aftertreatment (with chlorine bleaching liquor) or a reductive aftertreatment in alkaline medium (with sodium sulfide or glucose).

Instead of 4,4′-dinitrostilbene-2,2′disulfonic acid as starting product, use is frequently made of 4-nitrotoluene-2-sulfonic acid, which on being heated with sodium hydroxide solution is converted into dinitrodibenzyldisulfonic acid and dinitrostilbenesulfonic acid. All three nitro compounds condense in the presence of alkali, both with amino compounds and with each other. A number of direct cotton dyes (e.g., *C. I. Direct Yellow 21,* 40045 [*1325-46-8*]) are the result of complex condensation reactions of this type.

Simple azo compounds free of sulfonic acid groups can be produced with yields of 40 – 84 % by condensation of nitrobenzene with *o*- and *p*-anisidine, naphthylamines, and phenylenediamines without solvent in the presence of powdered sodium hydroxide at temperatures of 130 – 190 °C [14].

Because these reactions can proceed very vigorously, suitable precautionary measures are necessary.

2.2.2. Reduction of Nitro Compounds [2, p. 346]

During the reduction of aromatic nitro compounds, condensation of the nitro and hydroxylamino derivatives that occur as intermediate steps may result in the formation of azoxy compounds, which in a further reduction step can be converted into azo compounds:

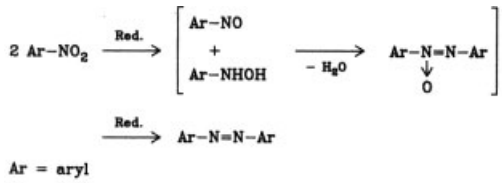

The method is only suitable for the synthesis of symmetrical azo compounds, for when two different nitro compounds are used, mixtures of all possible combinations are formed:

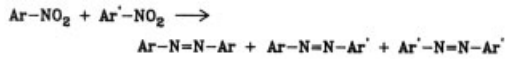

Glucose has proved especially suitable as a reducing agent for reductive linkage of aromatic nitro compounds, other reducing agents being alcohols, hydrazine, zinc, iron, and numerous inorganic salts. Only the reduction process with glucose has attained industrial significance in the synthesis of azo dyes, mainly for the linkage of nitroazo compounds under mild conditions. The

process generally takes place in a strongly alkaline, aqueous medium, less frequently in an alcoholic medium. For example, the red trisazo dye *C. I. Direct Red,* 25015 [*6391-13-5*] is obtained by coupling 2-methyl-5-nitroaniline to Nevile-Winther acid and subsequently treating with glucose and alkali:

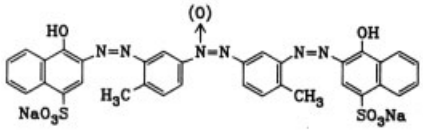

Reduction by glucose often does not proceed quantitatively up to the azo stage; instead the azo and azoxy compounds are mostly formed side by side. Where necessary, therefore, subsequent reduction of the azoxy compound to form the azo dye must follow.

2.2.3. Oxidation of Amino Compounds [2, p. 371]

Reaction of primary aromatic amines with suitable oxidizing agents results in symmetrical azo compounds, sometimes, by further oxidation of the azo stage, in the mixture with azoxy compounds:

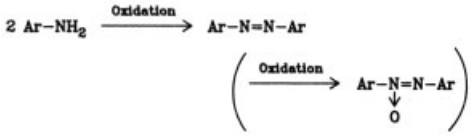

Hypochlorites are especially suitable as oxidizing agents, others being peroxy compounds (hydrogen peroxide, sodium perborate), oxygen or air in the presence of a pyridine – Cu(I)chloride catalyst, and chromic acid.

The main industrial application for the oxidative linkage of aromatic amines in the synthesis of azo dyes can be found in the thiazole range. The yellow cotton dye *C. I. Direct Yellow 28*, 19555 [*10114-47-3*], for example, is obtained by treating an aqueous solution of the sodium salt of the so-called dehydrothiotoluidinesulfonic acid (**9**) with sodium hypochlorite in an alkaline medium:

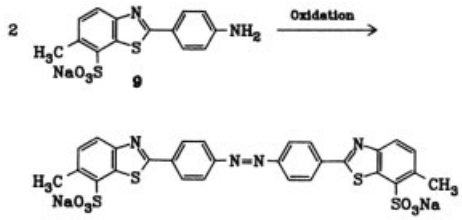

2.3. Synthesis from Azo Compounds

For specific technical purposes in the production of azo dyes, further reactions are carried out on the azo compound after introduction of the azo bridge:

- exposure of concealed or protected amino groups
- acylation of aminoazo compounds
- alkylation and acylation of phenolic hydroxyl groups
- metal-complex formation

2.3.1. Exposure of Concealed or Protected Amino Groups

The formation of an amino group in azo dyes is often necessary for the reaction with acid chlorides, e.g., for the production of reactive dyes, or for further diazotization and coupling. This is possible by reduction of nitro groups or by saponification of acylamino groups.

As the reduction of nitro groups must be carried out while maintaining the azo group, the number of suitable reducing agents is virtually restricted to sodium sulfide. Processing is carried out, depending on the nitro compound, in an aqueous alkaline solution or suspension at temperatures between 20 °C and 90 °C. A 10 – 30 % addition of alkali chloride has proved favorable, as a result of which most of the dye remains undissolved. After reduction the water-soluble sulfide oxidation products are separated, the dye is dissolved, and the sulfur is separated. The amino azo dye can be further processed straightaway in this form.

The acylamino groups, such as formyl, acetyl, or oxalylamino groups can be saponified in an acid or alkaline medium. As a rule 2 – 10 %

sulfuric acid or 2 – 6 % sodium hydroxide solution is used at temperatures up to the boiling point of the reaction mixtures.

2.3.2. Acylation of Amino Azo Compounds
[2, p. 390]; [5, vol. VI, p. 1]

Acylation reactions of aromatic amino compounds are important both for the production of reactive dyes, i.e., dyes that in the dyeing of textiles form a covalent bond with the fiber (→ Reactive Dyes), and for the synthesis of direct dyes (see Section 4), i.e., dyes that, because of their pronounced affinity for cellulosic fibers (substantivity), are adsorbed directly onto the cotton fiber from an aqueous solution.

Acylation agents that contain at least two reactive centers in the molecule and are capable of selective reactions under certain conditions, e.g., cyanuric chloride and cyanuric fluoride (**10**), 2,3-dichloroquinoxaline-6-carboxylic acid chloride (**11**) and chlorotrifluoropyrimidine (**12**), are used for the synthesis of a relatively large number of reactive dyes.

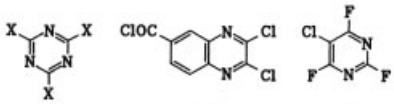

10 11 12

X = Cl or F

The reactive dyes most widely used so far are the chlorotriazinyl dyes, which can be produced on an industrial scale from dyes Ⓕ–NH_2 containing amino groups and cyanuric chloride, with cleavage of hydrogen chloride, in an aqueous organic medium at 0 – 5 °C and at a pH-value of 6.5 – 7.0. They can be condensed with an aliphatic or aromatic amine or even with a second dye Ⓕ–NH_2 at 40 to 50 °C to form the less reactive monochlorotriazinyl dyes:

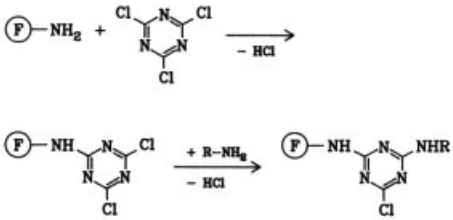

For the production of substantive dyes two identical or different aminoazo dyes are linked together by means of cyanuric chloride. The third chlorine atom remains either in the molecule or, preferably, is condensed with ammonia or simple aromatic amines at temperatures of 80 – 100 °C. The reaction with cyanuric chloride not only enables the inner-molecular blend to be varied at will but also considerably increases substantivity.

Phosgene, which is introduced into the solution of the amino azo compound at 20 – 35 °C and at a neutral pH, is used as a bifunctional acylation agent to link two amino azo compounds (see Section 4.1.6).

Numerous commercial dyes of a substantive nature have urea bridges of this kind, e.g., the direct dye *C. I. Direct Yellow 50,* 29025 [*8004-79-3*], which is produced from diazotized 2-aminonaphthalene-4,8-disulfonic acid and *m*-toluidine by reaction of the azo dye with phosgene:

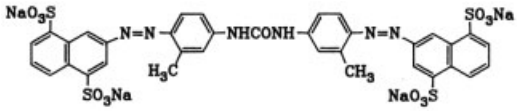

2.3.3. Alkylation and Acylation of Phenolic Hydroxyl Groups [2, p. 403]

Alkylation. To overcome the poor alkali fastness of *p*-hydroxyazo dyes (see Section 3.2.1) the phenolic OH groups can be converted into alkyl ether with such alkylation agents as dialkyl sulfate, alkyl halides, epoxides, etc. Depending on the water solubility of the *p*-hydroxyazo dyes, processing is carried out in weakly alkaline aqueous solution or in organic solvents, such as dimethylformamide or ethylene glycol dimethyl ether, under alkylation conditions, such as are usual with aromatic hydroxy compounds. The synthesis of *p*-alkoxyazo compounds by heating hydroxyazo compounds in an alcoholic medium with the addition of acid is of interest, e.g.:

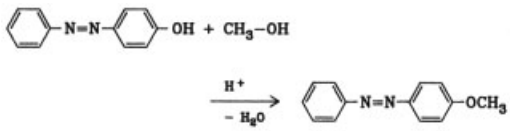

The reaction represents the inversion of the acid-catalyzed hydrolysis of azoaryl ethers.

On a large industrial scale hydroxyazo compounds are mainly reacted with ethyl chloride in aqueous alkaline solution (sodium carbonate, sodium hydroxide) by heating for several hours to temperatures of around 100 °C in a closed system.

Acylation. Here aryl sulfonation plays a particularly important role. In order to convert them into arylsulfonic acid esters, hydroxyazo dyes are reacted with arylsulfonic acid chlorides, mainly *p*-toluenesulfonic acid chloride or benzenesulfonic acid chloride in an aqueous alkaline medium. The formation of arylsulfonic acid esters not only overcomes the poor alkali fastness of *p*-hydroxyazo dyes but, with wool dyes, also improves the fastness to washing and milling and the neutral affinity.

2.3.4. Metal-Complex Formation

For more details → Metal-Complex Dyes and [15], [2, p. 434].

Azo dyes with certain structural groupings are capable of forming metal-complex compounds. Here even the coupling component, such as salicylic acid, can form a complex:

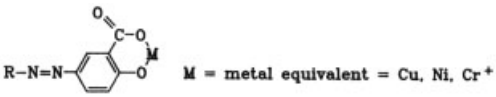

Technically far more important than these metal complexes, most of which are produced on the fiber (see chrome dyes, Section 3.2.1.3), are those dyes whose metal complexes are produced within the dye itself. Here the complex is always formed with the participation of the azo group. These dyes have complex-forming substituents in the *o,o′* position relative to the azo bridge and are, for instance, capable of forming 1 : 1 metal complexes of the general formula **13**.

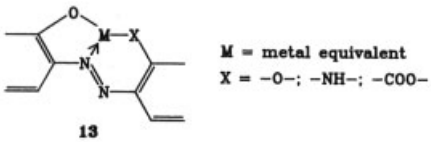

13

The azo dye anion always acts as a divalent ligand in threefold coordination. The remaining coordination positions of the metal ions are filled by addition of further ligands with lone electron pairs (e.g., H_2O, NH_3, amines) or by another azo dye anion (formation of 1 : 2 metal complexes).

Suitable complex-forming metals are in particular the trivalent chrome and trivalent cobalt (coordination number 6) and also the divalent copper and nickel (coordination number 4), whose complex dyes have great industrial importance.

One of the hydroxyl groups in the *ortho* position relative to the azo bridge can only be formed during metallizing by cleavage of an *o*-methoxy group. This demethylating metallizing process has been introduced on an industrial scale because of the wide variety of diazo components to choose from [15].

The addition of copper to the molecule in an oxidative process, in which a hydrogen atom in the *ortho* position relative to the azo group is converted into the appropriate hydroxyl group by means of oxidizing agents, such as hydrogen peroxide, has acquired technical significance [16]:

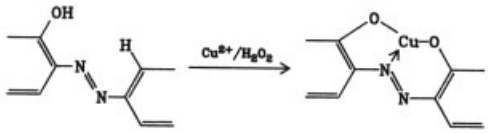

Chrome-complex dyes are best used as dyes for wool, polyamide, and leather and for coloring paints and plastics; copper complexes are used as cotton and leather dyes and also as pigments; and chrome and cobalt complexes, as spirit-soluble dyes.

2.4. Processes without Industrial Importance

Of the processes for producing azo dyes via synthesis of the azo group, the following are worth mentioning, although they have attained no industrial importance so far:

- reduction of diazonium compounds
- oxidative coupling
- dehydrogenation of diarylhydrazines
- reaction of arylhydrazines with quinones

– condensation of nitroso compounds with amines
– azo compounds from azido compounds

2.4.1. Reduction of Diazonium Compounds
[2, p. 320, 323]

When aromatic diazonium compounds are reacted with ammoniacal copper(I) salt solution, diarylazo compounds or diaryls are obtained with elimination of nitrogen

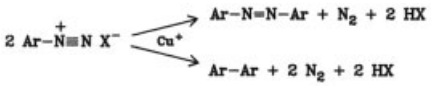

The course of the reaction depends on the type of substitution on the aromatic nucleus: whereas electron-attracting substituents result in diaryl formation, diazonium salts in which the substituents on the nucleus are electron donors can be converted to symmetrical azo compounds (Vorländer reaction) [17], [18]:

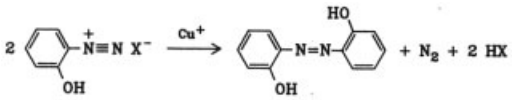

The action of sulfurous acid or its salts on diazonium compounds of the naphthalene series results in the formation of symmetrical azo compounds; the starting compounds may also contain electron-attracting substituents (Lange reaction) [19], e.g.:

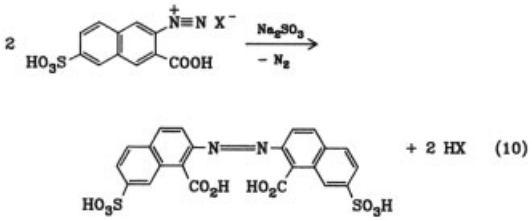

Studies by SUCKFÜLL and DITTMER have shown that in this process the diazonium salt reacts with the diazosulfonate formed from sulfite and diazonium ion [20]:

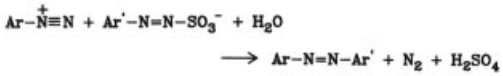

Whereas in the Lange reaction (see Eq. 10) the two aryl radicals are identical, SUCKFÜLL and DITTMER were able to expand the reaction of diazonium salt and diazosulfonate for the synthesis of asymmetrical azo compounds and, with certain restrictions, extend it to the azo series. The significance of the reaction lies in the access to azo compounds that are very difficult to synthesize by conventional means, e.g.:

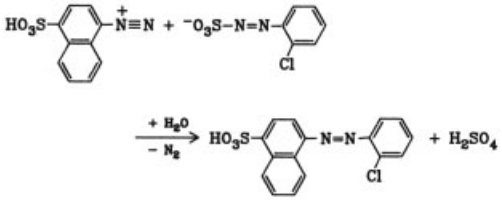

2.4.2. Oxidative Coupling

Of interest from the preparative and theoretical aspects is the oxidative azo coupling process discovered by S. HÜNIG, through which many heterocyclic azo compounds have become accessible [21].

A precondition of the reaction is the presence of a heterocyclic hydrazone, which couples with aromatic amines, phenols, or compounds with active methylene groups in the presence of dehydrogenating agents to form the corresponding azo compounds, e.g.:

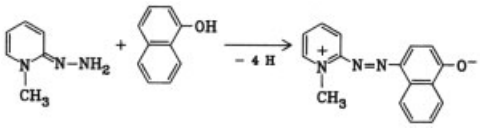

As regards the coupling components and the coupling location, the reaction resembles the azo coupling process. As electrophilic component, a diazonium ion **14** is suggested, which is formed from the hydrazone component through the action of the dehydrogenating agent and reacts by electrophilic substitution with the coupling component with the loss of two hydrogen atoms to form the azo compound.

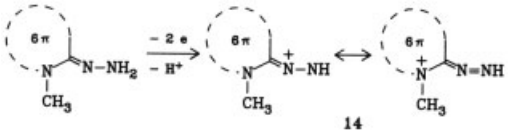

14

In the case of phenols and α-naphthols, the attack in principle always takes place at the *p*-position. Only when the *p*-position is occupied does coupling occur at the *o*-position. The optimum phenylazo compound yield (approx. 90 %) is attained at pH 9. The dehydrogenating agents mainly used for this purpose are potassium hexacyanoferrate(III), copper(II) sulfate, silver nitrate, sodium hypochlorite, or lead(IV) oxide.

The optimum pH range for oxidative coupling with aromatic amines and enamines is pH 1 – 3, diluted hydrochloric acid or glacial acetic acid being preferred as reagents. Well-tried dehydrogenating agents for this purpose are iron(III) chloride, red lead, lead(IV) oxide, and lead(IV) acetate.

In order to obtain an optimum yield, reagent and oxidizing agent should in each case be carefully matched to the coupling component used. Oxidative coupling has so far rarely been used for industrial production of azo dyes.

2.4.3. Dehydrogenation of Diarylhydrazines [2, p. 377]

N,N′-Diarylhydrazines can be dehydrogenated to form azo compounds:

$$\mathrm{Ar-NH-NH-Ar'} \xrightarrow{-2\mathrm{H}} \mathrm{Ar-N=N-Ar'}$$

Oxidizing agents frequently used are sodium hypochlorite or air in the alkaline, or iron(III) chloride or sodium dichromate in the acid medium, as well as nitrous acid. The method acquires importance when the azo compounds to be synthesized are not accessible by coupling or condensation reactions, or only with difficulty, and the relevant diarylhydrazines can be obtained by simple reactions, e.g.:

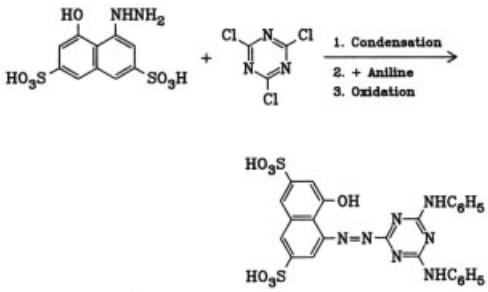

2.4.4. Reaction of Arylhydrazines with Quinones [2, p. 355]

Benzo- and naphthoquinones as well as quinones of some polycondensed aromatics react in an acid aqueous or aqueous alcoholic medium or in acetic acid to form hydroxyazo compounds, tautomers of quinone hydrazones, e.g.:

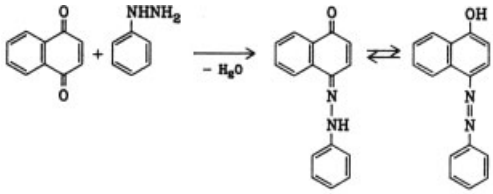

The reaction proceeds along the lines of this condensation process only if the oxidizing power of the quinone and the reduction potential of the hydrazine are not at such a high level that formation of hydroquinone takes place via a redox reaction. Quinones with a high oxidation potential (e.g., *p*-benzoquinone) can thus only be reacted with arylhydrazines (e.g., *o*- and *p*-nitrophenylhydrazine) and vice versa.

Of significance is the reaction of the quinones with arylhydrazines to form azo compounds, when the latter are difficult to obtain by azo coupling, e.g., 1-hydroxy-2-phenylazonaphthalene:

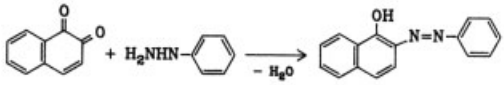

Instead of quinones, it is possible to react their monoximes with arylhydrazines, e.g.:

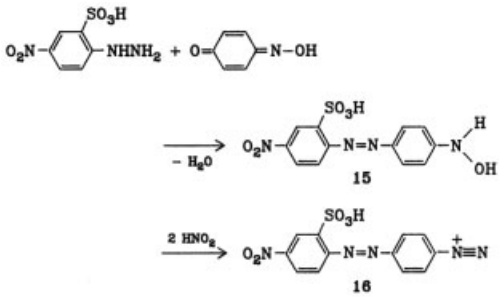

The hydroxylaminoazo compounds formed (**15**) react with nitrous acid to form azo diazo-

nium compounds (**16**), preliminary stages for the synthesis of polyazo dyes.

Arylhydrazines can also be converted into azo compounds by condensation with other dioxo compounds, provided this possibility is permitted by the keto-enolic tautomerism. An interesting example of this is the reaction of dioxosuccinic acid with two equivalents of phenylhydrazine-4-sulfonic acid, which with simultaneous formation of the pyrazolone ring results in tartrazine (*C. I. Acid Yellow 23,* 19140 [*1934-21-0*]), the oldest dye in this class:

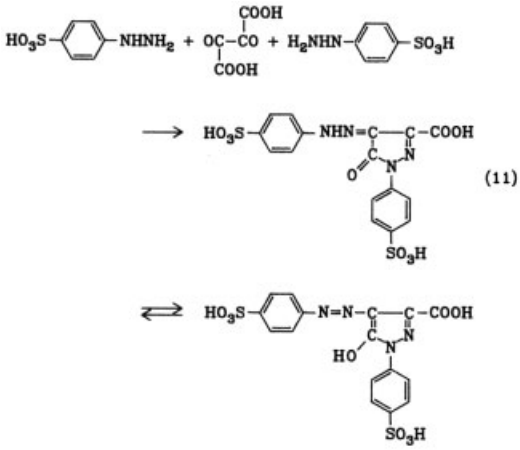

(11)

2.4.5. Condensation of Nitroso Compounds with Amines [2, p. 332]

Azo and polyazo compounds can be produced by condensation of amines with nitroso compounds, very good yields being frequently obtained:

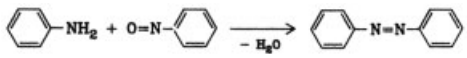

Depending on the substitution of the aromatic rings, the reactions are carried out at room temperature or at temperatures of up to about 100 °C. The reagent used is almost always acetic acid on its own or in mixtures with other organic solvents. Other reagents, such as water or alcohols, cause a sharp decrease in yield. The influence of substituents on the reaction rate is analogous to that in nitro – amine condensation (see Section 2.2.1).

Whereas nitroso – amine condensation is not generally possible in the naphthalene range, its use in the benzene range mainly serves the production of asymmetrically substituted azo compounds and the formation of polyazo compounds. It has no industrial importance in the production of azo dyes.

2.4.6. Azo Compounds from Azido Compounds

Compounds containing active methylene groups react with organic azido compounds in the presence of a base, with formal substitution of a diazo group for two hydrogen atoms (diazo group transfer) [22]:

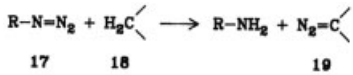

The resulting diazo compound (**19**) can be coupled with a second molecule of the methylene compound (**18**) to form a symmetrical azo compound (**20**):

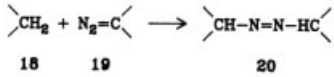

p-Toluenesulfonic acid azide (**17**, R = $CH_3C_6H_4SO_2-$), readily obtainable from *p*-toluenesulfonic acid chloride, has, in particular, proved to be a suitable azido compound.

The method permits the synthesis of a number of interesting heterocyclic azo compounds, e.g.:

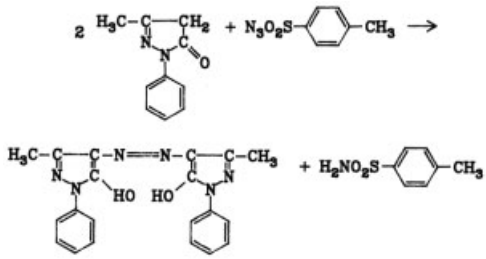

So far, however, it has been of no industrial importance.

2.5. Equipment for Industrial Production of Azo Dyes

Production of azo dyes on an industrial scale is still carried out on a batchwise basis.

The main items of equipment for industrial diazotization and coupling are the diazotizing vessel, the dissolving tank for dissolving the coupling component, and the central coupling tank with capacities between 25 and 80 m^3. Coupling batches of up to 8 kmol per batch can thus be processed.

Figure 1 illustrates the flow diagram and main items of equipment in a simple industrial azo dye synthesis.

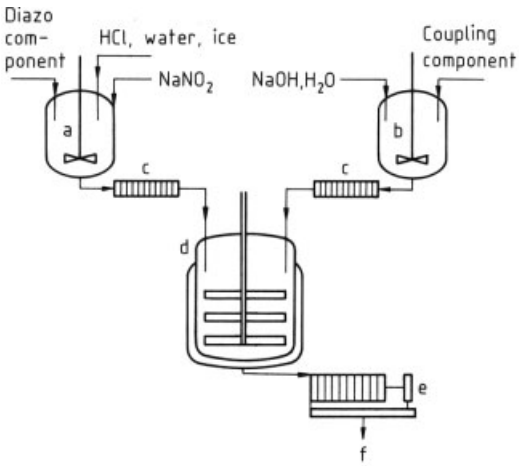

Figure 1. Equipment diagram for the production of an azo dye
a) Diazotizing tank; b) Dissolving tank; c) Clarifying press; d) Coupling vessel; e) Filter press; f) Drying and milling

In the simplest case, the solution of the coupling component is transferred to the coupling vessel after mechanical and adsorptive clarification by means of a clarifying press or a filter, and the diazonium salt solution, likewise after clarification, is added slowly above or below the surface of the coupling component solution. The solutions can also be added in the reverse sequence or both at the same time from the diazotizing and dissolving tank. The constancy of the reaction temperature and of a narrow pH range is a decisive factor as regards the coupling process and must be continuously monitored by means of suitable measuring devices. The coupling reaction can be followed in the coupling tank or other vessels by other reactions, such as saponification, esterification, acylation, or metal-complexing processes, which if necessary must be carried out under pressure and at higher temperatures.

Following the synthesis and any thermal aftertreatment process that may be required, azo dyes insoluble in water (e.g., pigments, disperse dyes) can be directly isolated by means of filter presses or suction filters. For large-volume products, preference is given to continuously operating rotary filters.

Water-soluble azo dyes (e.g., reactive dyes, acid dyes, or direct dyes) must be precipitated from the coupling solution (coupling liquor) by salting out or changing the pH. If in addition to the salt content other impurities in the dye are permissible, the spray drying process is used for separating water-soluble azo dyes from the reaction solution.

Depending on the requirements regarding corrosion resistance and ability to withstand compressive and thermal stress, the *materials* used for industrial equipment are iron, stainless steel, rubber-coated steel with or without an acid-proof lining, enameled steel, glass fiber reinforced synthetic resins, and wood.

In general, diazotizing, dissolving and coupling vessels are nowadays made from rubber-coated steel. The hard rubber lining can be exposed without damage to temperatures of up to 100 °C; contact with organic solvents should be avoided. Use has also been made of coupling tanks made from glass fiber reinforced synthetic resins (low weight, low cost, easy to repair), which are resistant to hydrochloric acid and can be exposed to temperatures of up to 100 °C. Wooden vats are still of importance in the industrial manufacture of azo dyes, primarily because of their corrosion resistance in aqueous systems, combined with low capital expenditure and repair costs. For reactions in an alkaline medium and in organic solvents, as well as for reactions under pressure, stainless steel tanks are employed. Stainless steel tanks and other items of equipment made from this material are not resistant to corrosion by mineral acid solutions. Stainless steel must also be avoided for reactions of weakly acid or salt-containing solutions carried out under pressure. Enameled equipment is used for reactions in acid media performed under pressure, but is not resistant to alkalis. Its advantage lies in its easy cleaning.

For reactions under pressure in an acid medium, rubber-coated and brick-lined steel tanks with agitators, heating coils, and thermometer tubes made from Hastelloy (nickel alloys with varying Mo, Cr, Mn, Cu, Si, Fe, and C contents) have also proved suitable. Brick-lined vessels are chiefly employed in large-capacity equipment, when the use of enameled vessels is dispensed with because of the high capital expenditure and repair costs involved.

In the industrial production of azo dyes in equipment that cannot be placed under pressure (wooden vats, drums made from synthetic resins), pumps are needed in order to convey the tank contents from vessel to vessel or via clarifying presses and filter presses. When steel drums are used, the solutions and suspensions are often transferred by air or nitrogen pressure.

3. Anionic Azo Dyes [1], [3], [4], [7], [5, vol. I, p. 480; vol. III, p. 268]

3.1. Introduction

Anionic dyes include many compounds from the most varied classes of dyes, which exhibit characteristic differences in structure (e.g., azoic, anthraquinone, triphenylmethane, and nitro dyes) but possess as a common feature water-solubilizing, ionic substituents. The anionic azoic dyes, which are discussed here, constitute the most widely used group of this class of dyes.

Most often sulfonic acid groups serve as hydrophilic substituents, because they are readily introduced and, as strong electrolytes, are completely dissociated in the acidity range used in the dyeing process. Almost invariably the products manufactured and employed are water-soluble sodium salts of the sulfonic acids.

Direct Dyes. In principle the anionic dyes also include direct dyes, but the latter, because of their characteristic structures, are used to dye cellulose-containing materials and go onto the fiber from a neutral dye bath. Because the classification used here is based chiefly on the dyes' application in the dyehouse, direct dyes are discussed in a separate chapter (see Chap. 4).

Reactive Dyes. From the chemical standpoint the group of anionic azo dyes includes a large proportion of the reactive dyes which, in addition to the usual structural characteristics, also contain certain reactive groups capable of reaction with functional groups of the fiber during the dyeing process, for example, with the hydroxyl groups of cellulose, the amino and mercapto groups of wool and silk, or the amino and carbonamide groups of synthetic polyamides. Owing to the covalent dye – fiber bond these dyeings usually exhibit good fastness to wet processing.

For structures of reactive dyes for wool and polyamide → Reactive Dyes.

Metal-Complex Dyes. The anionic azo dyes also including the metal-complex dyes, which are derived from *o,o′*-disubstituted azo dyes and which can be metallized not only on the fiber but whose metal-complex compounds are applied largely in the form of a complete dye (cf. Section 2.3.4). They serve for the dyeing of wool (see Section 3.2.1.4), silk and leather (Section 3.2.4) and polyamide (see Section 3.2.2), and also for cotton (see Section 4.1.8) and as alcohol- and ester-soluble dyes (see Section 6.1.2). This dye class, too, is discussed only briefly here; for details → Metal-Complex Dyes.

3.2. Acid Azo Dyes

This group of substances also includes dyes that are classified under other headings, → Anthraquinone Dyes and Intermediates, → Nitro and Nitroso Dyes; → Triarylmethane and Diarylmethane Dyes.

Azo dyes with relatively low molecular masses and one to three sulfonic acid groups serve as so-called acid azo dyes for dyeing and printing wool, polyamide, silk, and basic-modified acrylics and for dyeing leather, fur, paper, and food. The main area of application is the dyeing of wool and polyamide.

Individual dyes also play a part in the manufacture of color lakes, which are used to pigment printing inks and for coloring plastics.

Disazo and polyazo dyes containing sulfonic acid groups are also frequently used in the above applications. Many of them are also sold as substantive dyes for cotton because of their pro-

nounced affinity toward cellulosic fibers (see Chap. 4).

The term "acid dye" derives from the dyeing process, which is carried out in an acidic aqueous solution (pH 2 – 6).

Protein fibers contain amino and carboxyl groups, which in the isoelectric range (approximately pH 5) are ionized mostly to $-NH_3^+$ and $-COO^-$. In the acid dyebath the carboxyl ions are converted to undissociated carboxyl groups owing to the addition of acid HX (sulfuric or formic acid), which causes the positively charged wool ($H_3N^+-W-COOH$) to take up an equivalent amount of acid anions X^- (hydrogen sulfate, formate) [23]:

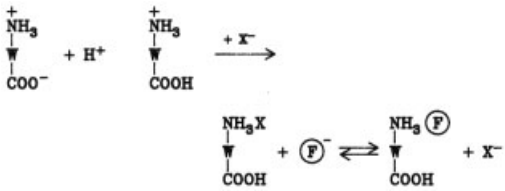

The actual dyeing process consists of a replacement of the absorbed acid anions X^- by the added dye anions $Ⓕ^-$, since the latter exhibit a much greater affinity for the substrate than the much smaller acid anions. Thus the dye is bonded to the wool not only by electrostatic attraction (salt formation) but also by its more or less strong affinity for the fiber.

Acid dyes are divided into three groups based on their differences in affinity, which is primarily a function of the molecular size.

1) *Leveling dyes* are relatively small molecules which form a saltlike bond with the protein fiber. This not very tight bond imparts good migration properties to such dyes on wool. The interplay between the bonded dye and the dye still in the dyebath results in very uniform, *level* dyeings. On the other hand they possess only poor wetfastness, so that the use of leveling dyes is severely limited.
2) *Milling dyes* are large-volume dye molecules, for which salt formation with the fiber plays only a secondary role and the adsorption forces between the hydrophobic regions of the dye molecule and those of the protein fiber predominate. The resulting tighter bond yields good wetfastness, but the levelness of the dyeings is not always satisfactory because of an inadequate migration property. In consequence the range of applications of milling dyes is also limited.
3) *Dyes with intermediate molecular size* not only form a saltlike bond with the wool fiber but are also bonded to the fiber by intermolecular forces and have properties lying in an intermediate position between those of the leveling and the milling dyes. They are used mainly where average requirements are placed on the wetfastness and levelness of the dyeing.

Concerning the possibility of influencing the rate of dyeing and the leveling power during the dyeing process by the addition of acid or color salt → Dyes, General Survey.

3.2.1. Wool Dyes

The standard dyes for wool are divided essentially into four groups according to their different dyeing behaviors (cf. Table 1).

Table 1. Names of important ranges of acid wool dyes

Manufacturer	Ranges			
	Group A	Group B	Group C	Group D
Bayer	Acilan	Supracen	Supranol	Supramin
Ciba-Geigy	Erio	Eriosin	Polar	
Hoechst	Amido	Lanaperl	Alphanol Fast	
ICI	Lissamin	Coomassie	Carbolan	
Sumitomo		Suminol leveling	Suminol milling	
Sandoz		Sandolan-E	Sandolan-N	

Group A contain the low-priced, usually older leveling dyes with moderate light fastness. These dyes are still used in larger quantities for cheap articles.

Group B also includes leveling dyes, which are applied in a strongly acid bath but which exhibit very good light fastness.

Group C contains the milling dyes, which are applied in a weakly acid to neutral bath.

Group D contains the acid dyes with good leveling properties, intermediate molecular sizes, and for the most part good light fastnesses.

The product ranges of the various dye manufacturers do not completely fulfill the demands made on the dyes in respect of the individual

fastnesses, and within a manufacturer's range the properties of the dyes can be individually specified.

In describing the separate dye types it is useful to distinguish between the monoazo and the disazo dyes, and to further break down the disazo dyes into primary and secondary types on the basis of chemico-structural principles.

In the industrial manufacture of acid azo dyes usually aniline derivatives are used as the diazo components and for disazo dyes also diaminodiphenyl derivatives and other bifunctional diamines. The coupling components for orange to blue shades are commonly aniline, naphthylamine, naphthol, and aminonaphthol derivatives, whereas phenylpyrazolones are much used for preparing dyes in the yellow and orange shades.

In regard to the stability of the shade to changes in pH, care must be taken that hydroxyl or amino groups lie adjacent to the azo group in order to form a hydrogen bridge with the latter. This hydrogen bridge prevents a dissociation of the hydroxyl groups or a protonation of the amino groups and hence avoids an unwanted change in shade in response to moderate pH shifts. Groups that are not in the *ortho* position must be alkylated or acylated (see Section 2.2.3).

3.2.1.1. Monoazo Dyes

Among the acid monoazo dyes are a number of much used wool dyes that possess no outstanding coloristic properties but that are distinguished by brilliance of shade, very good leveling power, and particularly low cost, while their wash fastness and light fastness meet only low to medium requirements. Partly influenced by the introduction of the International Woolmark to label high-quality wool articles, the fastness requirements have risen considerably in the last 20 years, which has necessitated the development and manufacture of particularly fast dyes. For this reason many, and above all the older, acid dyes have lost all importance for the dyeing of wool and are used today only for the coloring of paper, soaps, food, and cosmetics.

In the following discussion the acid monoazo dyes are classified according to the type of coupling component.

Aromatic Amines as Coupling Components. One of the oldest azo dyes is *C. I. Acid Yellow 36,* 13065 (metanil yellow) (**21**) [*587-98-4*].

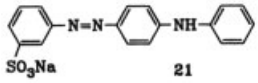

21

Because of the basic group suitable for salt formation the dye is not fast to acid, but it is still used today for dyeing wool and in special areas (leather and paper) primarily for price reasons.

Coupling H acid to 1-(phenylamino)naphthalene-8-sulfonic acid (*N*-phenylperic acid) yields *C. I. Acid Blue 92,*13390 (**22**) [*3861-73-2*].

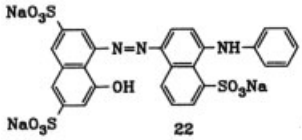

22

On wool the product turns out a very pure blue of good light fastness and with moderate wet fastness and adequate leveling power.

Naphthols and Naphtholsulfonic Acids as Coupling Components. In this series two important acid dyes with very similar structures are *C. I. Acid Red 88,* 15620 (**23**) [*1658-56-6*], derived from diazotized naphthionic acid and 2-naphthol, and *C. I. Acid Red13,* 16045 (**24**) [*2302-96-7*], from naphthionic acid and Schäffer's acid:

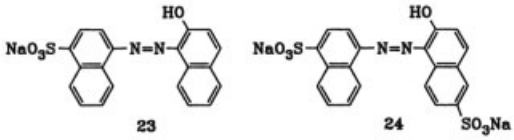

23 **24**

Both are all-purpose dyes which, because of their attractive red shades, are still in use today in many areas of textile dyeing and also for leather and paper dyes. Wool dyeings produced with these dyes exhibit moderate fastness levels.

A long-known, inexpensive, but only moderately fast dye is *C. I. Acid Orange 7,* 15510 (Orange II) (**25**) [*633-96-5*]. As a wool dye it is now of secondary importance, but is used in special areas such as leather dyeing, paper coloration, and the manufacture of color lakes.

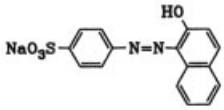

25

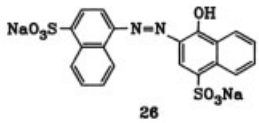

26

A previously much used red wool dye is *C. I. Acid Red 14,* 14720 (**26**) [*3567-69-9*], which is now used to a limited extent in wool dyeing for inexpensive articles.

The usually very low wetfastness of these simple acid azo dyes can be improved at the expense of leveling power by the introduction of special, large-volume substituents. Example: *C. I. Acid Orange 19,* 14690 (**27**) [*3058-98-8*], is a very important acid dye for wool and polyamide with good fastness properties and yet with good leveling power.

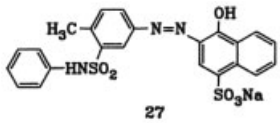

27

Aminonaphtholsulfonic Acid as the Coupling Component. Important coupling components are γ *acid* (2,8-aminonaphthol-6-sulfonic acid) and *H acid* (1,8-aminonaphthol-3,6-disulfonic acid), from which many important wool dyes are derived, some of which exhibit very good light and wetfastness.

Especially in acid coupling γ acid yields dyes with very good light fastness, which is presumably attributable to the formation of a hydrogen bridge between the azo bridge and the hydroxyl group *peri* to the azo bridge.

As examples are cited two outstandingly lightfast dyes of good leveling power whose dyeings on wool exhibit moderate wetfastness: *C. I. Acid Red 42,* 17070 (**28**) [*6245-60-9*], and *C. I. Acid Red 37,* 17045 (**29**) [*6360-07-2*].

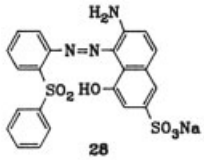

28

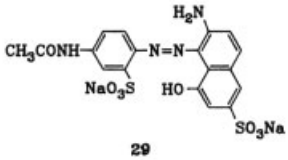

29

Through the choice of suitably substituted diazo components, bluish red to blue wool dyes with good light and wetfastnesses can be produced. Example: *C. I. Acid Violet 14,* 17080 (**30**) [*4404-39-1*].

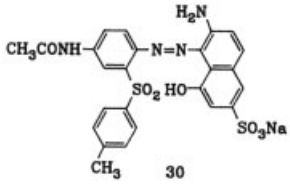

30

Of interest is the manufacture of 4-amino-3-(4′-toluenesulfonyl)acetanilide, which is used as the diazo component: *p*-phenylenediamine is oxidized with iron(III) chloride to 1,4-benzoquinonediimine, to which *p*-toluenesulfinic acid is added. The final step is partial acetylation:

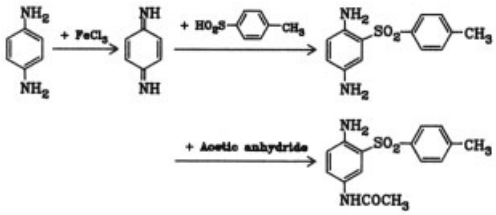

Further examples are: *C. I. Acid Red 32,* 17065 (**31**) [*6360-10-7*], and *C. I. Acid Blue 117,* 17055 (**32**) [*10169-12-7*].

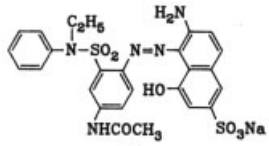

31

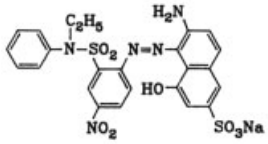

32

On wool both dyes exhibit quite good wetfastness and high leveling power.

In order to achieve valuable dyes with alkaline coupling to γ acid, the amino group must be acylated or arylated. Example: *C. I. Acid Red 68,* 17920 (**33**) [*6369-40-0*].

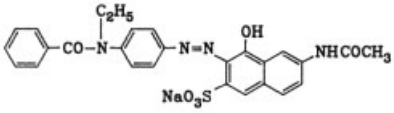

33

On wool it exhibits high leveling power and yields light- and wetfast dyeings.

N-arylation of γ acid results in a strong deepening of the shade. The usually brown to black dyes thus obtained are fast to light and milling and go onto wool and silk from a neutral liquor. Example: *C. I. Acid Brown 20,* 17640 (**34**) [*6369-33-1*].

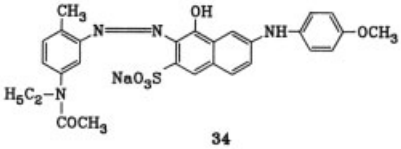

34

An important coupling component is H acid. Realization that acylation of its amino group substantially improves the light fastness and in addition yields dyes with good leveling power led to many valuable light- and wetfast products that even today occupy a firm position on the market. A relatively simply structured wool dye, *C. I. Acid Red 1,* 18050 (**35**) [*3734-67-6*] is obtained by the alkali coupling of diazotized aniline to *N*-acetyl H acid.

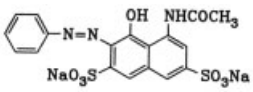

35

Valuable dyes can also be made with *N*-benzoyl H acid and *N*-toluenesulfonyl H acid.

An interesting development is the *Carbolans* of ICI. Their neutral affinity to wool and very good wetfastness properties are achieved by incorporating long hydrophobic hydrocarbon radicals [5, vol. III, p. 268]. Example: *C. I. Acid Red 138,* 18073 (**36**) [*15792-43-5*].

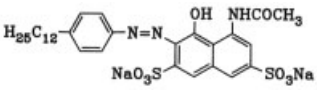

36

1-Phenyl-5-pyrazolones as Coupling Components. Especially lightfast yellow shades are obtained by using 1-phenyl-5-pyrazolones as coupling components. The first representative of the class to appear was Tartrazine, *C. I. Acid Yellow 23,* 19140 [*1934-21-0*], which is formed in a smooth reaction when phenylhydrazine-4-sulfonic acid is heated with dioxosuccinic acid (see Eq. 11, Section 2.4.4). Today the starting compound is 1-(phenyl-4′-sulfonic acid)-3-carboxy-5-pyrazolone, which can be obtained from oxaloacetic ester and phenylhydrazine-4-sulfonic acid and can be coupled with diazotized sulfanilic acid. Tartrazine is an intense yellow, moderately fast dye that is still important in wool dyeing but is also used in the dyeing of leather, the coloring of paper, soap, and food, and in the lake form as a pigment.

For price reasons the 1-aryl-3-methyl-5-pyrazolones and their derivatives are preferred to the corresponding 3-carboxypyrazolones. There is such a wide range of possible variations that dyes in this series extend from greenish yellow to reddish orange. Examples: *C. I. Acid Yellow 17,* 18965 (**37**) [*6359-98-4*].

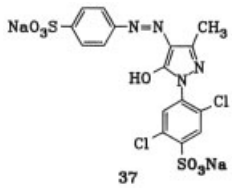

37

On wool the product yields a clear, superbly lightfast yellow with good to very good general fastness properties, but relatively poor milling fastness.

C. I. Acid Yellow 76, 18850 (**38**) [*6359-88-2*], is obtained by coupling diazotized 4-aminophenol to the pyrazolone component and then esterifying with *p*-toluenesulfonic acid chloride in an alkaline medium. The toluenesulfonic ester group substantially improves the fastness to milling and makes the shade obtained largely independent of pH; the light fastness is not quite as good as that of Acid Yellow 17.

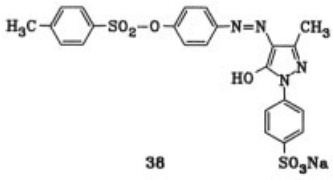

38

3.2.1.2. Disazo Dyes

Dyes with two azo groups are divided on chemico-structural principles into primary and secondary disazo dyes. Primary disazo dyes are manufactured either according to the scheme $D_1 \rightarrow K \leftarrow D_2$ from a bifunctional coupling component K and two identical or different diazo components D_1 and D_2, or according to the scheme $K_1 \leftarrow D \rightarrow K_2$ by coupling a bisdiazotized diamine D to the coupling components K_1 and K_2, which may be the same or different. Secondary disazo dyes are manufactured according to the scheme $D \rightarrow M \rightarrow K$ (M = middle component) from a diazotized aminoazo dye ($D \rightarrow M$) and a coupling component K (series coupling).

Primary disazo dyes of the type $D_1 \rightarrow K \leftarrow D_2$ are obtained by twofold coupling to, for example, resorcinol. These orange to brown dyes have poor light and wetfastness and are used only for inexpensive leather dyeings. Among the naphthalene derivatives H acid is one of the most important; it serves as a bifunctional coupling component K for the manufacture of many wool dyes now in use.

Dye synthesis following the above scheme generally permits little variation of the shade (black, dull brown, and blue), because completely conjugated chromophoric systems are formed.

One of the most important acid dyes is *C. I. Acid Black 1,* 20470 (**39**) [*1064-48-8*].

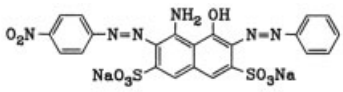

39

It dyes wool in blue-black shades with very good light fastness but only moderate wetfastness. Nevertheless, because of its high affinity and good leveling power it has remained an important product, which through shading with yellow, orange, or red forms the basis of most black acid dyes. To manufacture this type of dye, coupling is always first carried out in an acid medium with the more strongly electrophilic diazonium salt, here with diazotized 4-nitroaniline, then in the alkaline range with diazotized aniline:

$$\text{4-nitroaniline} \xrightarrow[\text{a}]{1} \text{H acid} \xleftarrow[\text{alk.}]{2} \text{aniline}$$

Abbreviated manufacturing instructions for C. I. Acid Black 1, 20470: Stir 260 kg of 4-nitroaniline with 1200 L of water and a 23 % solution of 132 kg sodium nitrite for about 1/2 h. Run this suspension into a mixture of 4000 L water, 530 L 31 % hydrochloric acid, 500 kg ice, and a 23 % solution of 1.2 kg sodium nitrite at no more than 10 °C. Add about 1000 kg of ice and then 2600 L of an aqueous solution of 485 kg H acid within a period of 1 – 2 h with thorough stirring. Coupling is complete at the end of 3 h.

Meanwhile diazotize 165 kg of aniline in 300 L water, 1500 kg ice, and 440 L 31 % hydrochloric acid by adding 124 kg of sodium nitrite as a 23 % solution. This yields a volume of 3000 L. Then dissolve the monoazo dye by adding 360 L of 30 % sodium hydroxide solution and cool with about 3000 kg of ice. To this run in the solution of diazotized aniline and then immediately 1800 L of 15 % soda solution within 2 – 3 min. The volume is 22000 – 23000 L. After 6 h add common salt (9.5 % calculated on the volume) to cause salting out and continue stirring overnight. Filter off the precipitated dye and dry; the yield is about 1900 kg.

Usually the acid coupling of 4-nitroaniline is not continued to completion, because this requires too much time. Consequently, the commercial product normally contains a small proportion of red monoazo dye, which together with the greenish black disazo dye yields the desired neutral black shade.

Primary Disazo Dyes of the Type $K_1 \leftarrow D \rightarrow K_2$ ($K_1 = K_2$ or $K_1 \neq K_2$). This series includes many milling dyes because of the molecular size achieved (cf. Table 1, group C). Depending on the type of coupling components K_1 and K_2, which may be phenols, pyrazolones, acetoacetic acid arylamides, or naph-

tholsulfonic acids, clear yellow to red shades are obtained. The preferred diazo components are diaminodiphenyl derivatives and similarly structured diamines. Once important dyes derived from unsubstituted benzidine are now of no further importance because of the latter's carcinogenic activity (→ Benzidine and Benzidine Derivatives). Because the coupling of bisdiazotized diamines to form monoazo and disazo dyes proceeds with different rates, asymmetrical disazo dyes are readily manufactured. For example, the milling and wetfast red wool dye *C. I. Acid Red 114,* 23635 (**40**) [*6459-94-5*], is manufactured by coupling bisdiazotized 3,3′-dimethylbenzidine with one equivalent of Gacid; the monoazo compound obtained is further coupled with phenol to yield the disazo dye, the *p*-hydroxyl group of which is subsequently esterified with toluenesulfonic acid to improve the fastness properties and pH stability.

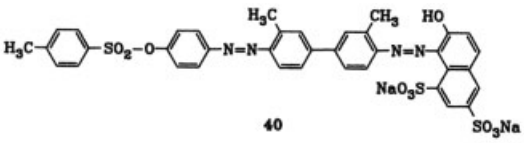

40

An important discovery was the realization that benzidines substituted in the 2,2′-position are outstandingly suitable for the manufacture of very wash-fast and milling-fast wool dyes, while the substantivity toward cellulosic fibers is reduced. A number of important acid wool dyes have been developed on this basis, for example, *C. I. Acid Yellow 44,* 23900 (**41**) [*2429-76-7*].

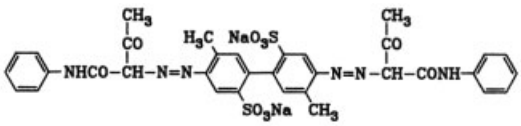

41

In the manufacture of this dye bisdiazotized 5,5′-dimethylbenzidine-2,2′-disulfonic acid is coupled to two equivalents of *N*-acetoacetylaniline. The clear, intense, greenish yellow dyeing obtained on wool is milling fast and exhibits excellent wetfastness but only moderate light fastness.

Other members of this series are *C. I. Acid Orange 56,* 22895:

benzidine-2,2′- ⟶ 1-phenyl-3-methyl-5-pyrazolone disulfonic acid ⟶2-naphthol

and *C. I. Acid Yellow 42,* 22910 [*6375-55-9*]

benzidine-2,2′- ⟶ 2 equivalents of 1-phenyl-3-methyl-disulfonic acid ⟶ 5-pyrazolone

Instructions for manufacturing C. I. Acid Yellow 42, 22910: Dissolve 250 kg of benzidine-2,2′-disulfonic acid in 2500 L of water at 50 °C that was made weakly alkaline with about 90 kg of soda. Then add 100 kg of sodium nitrite in the form of a 30 % solution. Run the entire solution into 4000 L of water at 25 °C, containing 350 L 31 % hydrochloric acid and 675 kg rock salt. The total volume is 7500 L, the temperature 30 °C. Stir for about 15 min during the diazotization process.

In the coupling vat add 264 kg of 1-phenyl-3-methyl-5-pyrazolone to 2000 L of water containing 80 kg of soda and dissolve at 80 °C. Then add another 80 kg of soda and cool with 2000 kg of ice to 20 °C. Run in the diazonium salt solution in about 30 min; no diazo reaction should occur during this procedure. Stir overnight, during which coupling takes place. Then draw off the dye and dry; the yield is about 720 kg.

In addition to benzidine derivatives another class of substances of interest as bifunctional diazo components consists of special diamines that can be manufactured through the reaction of aromatic amines with benzaldehyde or cyclohexanone, e.g.:

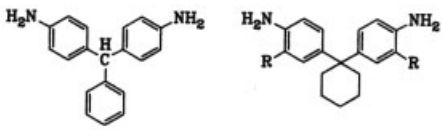

R = H, CH_3, OCH_3

Dyes based on these compounds possess, in addition to good light fastness, excellent wetfastness and are usually neutral dyeing on wool. This, although of no importance for dyeing pure wool, plays an important role in dyeing blended spun yarn and blended fabrics of wool and cotton or wool and viscose staple. The neutral dyeing acid dye can be used in combination with direct dyes (union wool recipes). Examples are *C. I. Acid Yellow 56,* 24825 (**42**) [*6548-24-9*],

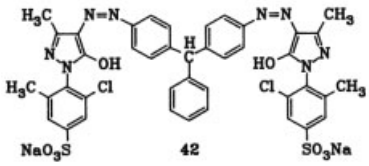

42

and *C. I. Acid Red 154,* 24800 (**43**; R = CH_3) [*6507-79-5*].

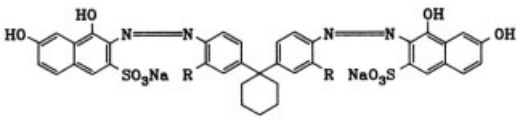

43

*C. I. Acid Red 134,*24810 (**43**; R = OCH_3) [*6459-69-4*], is another interesting dye for wool and polyamide.

Particularly good milling fastness is achieved by the introduction of 4,4′-diaminodiphenylthioether as the diazo component, for example in the yellow wool dye *C. I. Acid Yellow 38,* 25135 (**44**) [*13390-47-1*].

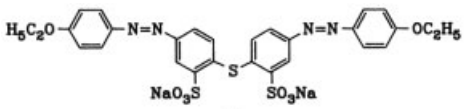

44

Secondary Disazo Dyes. The secondary disazo dyes, which can be manufactured by series coupling according to the scheme D → M → K (D = diazo component; M = middle component; K = coupling component), offer much greater possibilities for variation than the primary disazo dyes. This group includes red dyes and, above all, important navy blues and blacks. Some of the red dyes of this type are manufactured from aminoazobenzene, aminoazotoluene, or the corresponding sulfonic acids by diazotization and coupling to naphtholsulfonic acids.

One of the oldest disazo dyes is Sudan IV, *C. I. Acid Red 66,* 26905 (**45**) [*4196-99-0*], which today is scarcely of any importance.

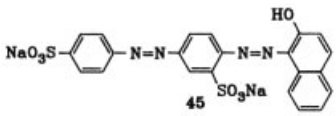

45

A product with a similar structure is *C. I. Acid Red 73,* 27290 (**46**) [*5413-75-2*], which is important today in wool dyeing and in paper coloration and leather dyeing because of its outstandingly clear shade and good leveling power.

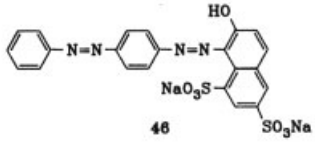

46

The most important dyes of the type D → M → K are, however, black and navy blue wool dyes, which contain as the middle component M chiefly 1-naphthylamine or 1-naphthylamine-7-sulfonic acid and as the coupling component K *N*-phenylperi acid [= (1-phenylamino)naphthalene-8-sulfonic acid], *N*-tolylperi acid, and naphthol or 1-naphthylamine derivatives.

Examples of these high-yield and very washfast and lightfast dyes are *C. I. Acid Black 24,* 26370 (**47**) [*3071-73-6*], which is manufactured by coupling diazotized 1-naphthylamine-5-sulfonic acid to 1-naphthylamine, further diazotization of the aminomonoazo dye and coupling to *N*-phenylperi acid; and *C. I. Acid Blue 113,* 26360 (**48**) [*3351-05-1*]. This blue dye is very important in the dyeing of wool and polyamide.

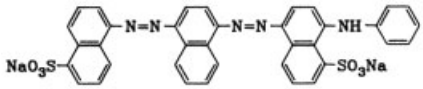

47

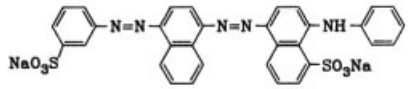

48

Manufacturing instructions for C. I. Acid Blue 113, 26360: Dissolve 250 kg of 100 % aniline-3-sulfonic acid in about 3000 L of neutral water, add about 1000 kg of ice and 400 L of 31 % hydrochloric acid, and diazotize at 10 °C with a 23 % solution of 100 kg sodium nitrite. Then dissolve 220 kg 100 % 1-naphthylamine in 2000 L water with 185 L 31 % hydrochloric acid at the boil. Quickly run the solution into the diazotized aniline-3-sulfonic acid to which 4000 kg of ice has been added. Then in about 45 min partly neutralize to about pH 4 with 400 L of 15 % sodium

hydroxide solution, temperature 10 °C, volume of 15000 L. After adding 25 % by volume rock salt, stir for 2 h and again add 360 L of 15 % sodium hydroxide solution.

After cooling to 0 °C with about 1000 kg of ice, add 120 kg of sodium nitrite as a 23 % solution and then quickly 600 L of 48.5 % sulfuric acid. After 4 h separate the diazonium salt of the aminomonoazo dye in a filter press and stir the press cake into 1500 L of water and 400 kg of ice.

Dissolve 438 kg of 100 % *N*-phenylperi acid in 2000 L of water at about 80 °C and then adjust to a volume of 8500 L and a temperature of 10 °C with about 1500 kg of ice and water. Finally add 230 L of 15 % sodium acetate solution and adjust to pH 3 – 4 with about 10 L of 31 % hydrochloric acid. Then simultaneously feed into this solution 400 L of 15 % sodium acetate solution and the suspension of the diazonium salt press cake. Stir thoroughly and adjust this dispersion to the alkaline range by adding 250 L of 30 % sodium hydroxide solution and cause to dissolve by heating to 80 °C. Add 10 % by volume rock salt and continue stirring for 3 h. Separate the dye in a filter press at 70 °C. The final volume is 19000 L; the yield is 800 kg of dried product.

A very wash-fast black wool dye is *C. I. Acid Black 26,* 27070 (**49**) [*6262-07-3*].

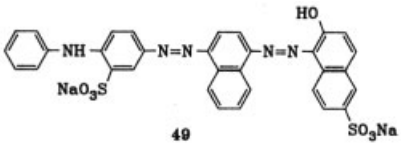

49

It contains 4-amino-diphenylamine-2-sulfonic acid as the diazo component, 1-naphthylamine as the middle component, and 2-naphthol-6-sulfonic acid as the final coupling component.

3.2.1.3. Chrome Dyes [3, p. 464], [15, p. 680]

Azo dyes with certain groupings can be converted on the fiber with chromium salts to chromium complexes that are more or less soluble (see Section 2.3.4). This "chroming" of largely acid monoazo dyes improves the light fastness of the dyeings and above all their wet-fastness, because the complex formation blocks the hydrophilic groups in the molecule. It also results in a deepening and dulling of the shade, so that these dyes do not yield brilliant colors. Consequently, for combination dyeing the range of chrome dyes has been supplemented with brilliant acid dyes, which cannot form chrome lakes and survive the chroming process without change.

Chrome dyes are applied to the fiber chiefly with the aid of the afterchroming method and special products also by the one-bath chroming method (metachrome process). For ecological reasons chrome dyes are of severely limited importance.

Dyes for the Afterchroming Method. In afterchroming the dye is allowed to go onto the fiber from an acid bath, and this is followed by treatment with dissolved alkali dichromate, which is reduced by the cystine group of the wool to trivalent chromium compounds. The chromium complex formed on the wool fiber contains two azo dye radicals for each metal atom (1 : 2 complex compound).

In the reversal of this process with so-called mordant dyes, first the chromium salt is applied to the fiber (premordanting) and then the dye. Mordant dyeing is essentially only of historical interest.

The afterchroming dyes have lost their former importance because of the considerable amount of time required by the dyeing process, but they are still used in wool dyeing because of their low price. For the names of important afterchroming dye ranges see Table 2.

Table 2. Important ranges of afterchroming dyes

Manufacturer	Range
Bayer	Diamantchrom
Ciba-Geigy	Eriochrom
Sandoz	Omegachrom

Furthermore, dyes of this type are still produced in large amounts in the COMECON countries and in the People's Republic of China.

In the yellow and orange types the salicylic acid grouping dominates. One of the most important chrome-developed yellow dyes is *C. I. Mordant Yellow 1,* 14025 (**50**) [*584-42-9*]. After chroming it yields a somewhat dull yellow with good fastness properties.

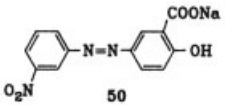

50

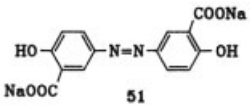

51

C. I. Mordant Yellow 5, 14130 (**51**) [*6054-98-4*], possesses two salicylic acid groups. After chroming the product is greenish yellow and exhibits very good general fastness properties. Because diazotized aminosalicylic acid is not sufficiently electrophilic, the dye is manufactured by coupling diazotized 5-amino-2-chlorobenzoic acid to salicylic acid and saponifying the activated chlorine atom.

The complex formation on the salicylic acid group has little influence on shade and light fastness, in contrast to *o,o′*-disubstituted arylazo compounds, on which complexing brings about a considerable deepening of shade and usually marked increase in light fastness. Examples: *C. I. Mordant Red 7,* 18760 (**52**) [*3618-63-1*], dyes wool a dull orange, which through chroming becomes a clear red with very good fastness properties.

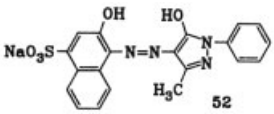

52

Red, blue, and primarily black afterchroming dyes contain naphthalene radicals in the diazo or coupling component.

A blue dye, for example, is *C. I. Mordant Blue 13,* 16680 (**53**) [*1058-92-0*].

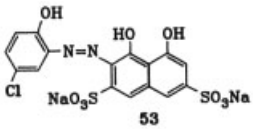

53

One of the most important black dyes is *C. I. Mordant Black 9,* 16500 (**54**) [*2052-25-7*].

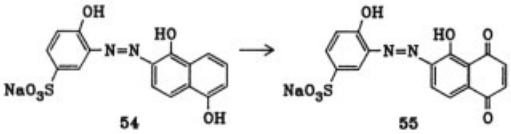

54 55

Oxidation of the dye during chroming with dichromate creates a naphthoquinone structure in the molecule (**55**). Two other important black dyes are *C. I. Mordant Black 3,* 14640 (**56**, R = H) [*3564-14-5*], and *C. I. Mordant Black 11,* 14645 (**56**, R = NO_2) [*1787-61-7*].

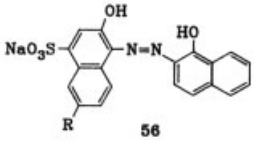

56

C. I. Mordant Black 11 is also used as an indicator in the complexometric titration of various bivalent metals with which the dye forms complexes [24].

The *o,o′*-dihydroxyaminoazo and *o*-hydroxy-*o′*-aminoazo compounds with phenols or aniline derivatives as coupling components yield largely brown shades on chroming; they include one of the most important brown dyes with very good fastness properties, *C. I. Mordant Brown 33,* 13250 (**57**) [*3618-62-0*].

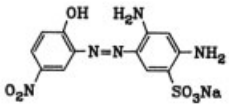

57

The rarely used chrome-developed disazo dyes mostly contain the salicylic acid grouping, which in the type D → M → K (see page 27) is usually introduced into the molecule in the form of aminosalicylic acid as the diazo component. Dyes of the type $K_1 \leftarrow D \rightarrow K_2$ (see page 25) are manufactured by coupling bisdiazotized benzidine derivatives (D) to two equivalents of salicylic acid ($K_1 = K_2$) or to one equivalent each of salicylic acid (K_1) and naphtholsulfonic acid (K_2).

Dyes for the One-Bath Chroming Method (Metachrome Process). In this method the metal complex is formed on the fiber by the simultaneous action of chrome dye and dichromate. Compared with the afterchroming method metachrome dyeing has the advantage of simpler application, but the number of usable dyes for this process is more severely limited. Whether a dye is suitable for the metachrome process depends on its hydrophilic character and its rate of complex formation [15, p. 680].

The method is used at present for wool dyeing primarily because of price considerations.

Examples of typical one-bath chroming dyes are: *C. I. Mordant Yellow 30,* 18710 (**58**) [*10482-43-6*], *C. I. Mordant Red 19,* 18735 (**59**) [*1934-24-3*], *C. I. Mordant Red 30,* 19360 (**60**) [*6359-71-3*], *C. I. Mordant Brown 48,* 11300 (**61**) [*6232-53-7*], and *C. I. Mordant Blue 7,* 17940 (**62**) [*3819-12-3*].

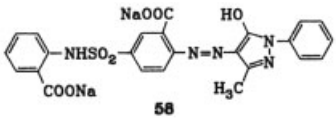

58

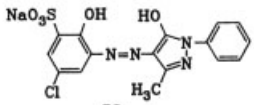

59

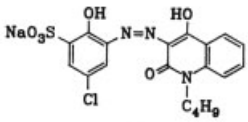

60

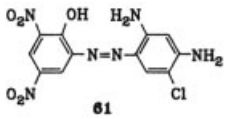

61

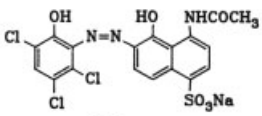

62

3.2.1.4. Metal-Complex Dyes

(cf. Section 2.3.4)

Only a short overview of the subject will be given here (for detailed treatment → Metal-Complex Dyes). Metal-complex dyes are manufactured in the complete dye form before the dyeing process and usually contain chromium or cobalt ions, rarely copper or iron ions, bound in the molecule. A distinction is made between 1 : 1 and 1 : 2 metal-complex dyes, depending on whether there is one metal ion for each dye molecule or for every two dye molecules. These products generally yield dyeings of high light and wetfastness but of somewhat muted shades.

1 : 1 Metal-Complex Dyes. Among these dyes primarily the 1 : 1 chromium complexes containing sulfonic acid groups have achieved commercial importance. They must be applied from a strongly acid bath, which imposes certain limits on their range of applications. The 1 : 1 metal complexes are not suitable for polyamide, which is partially decomposed under the dyeing conditions for these products. Their main area of application is in the dyeing of wool, but they are also suitable for leather dyeing. See Table 3 for a list of ranges.

Table 3. Important ranges of 1 : 1 metal-complex dyes

Manufacturer	Range
BASF	Palatin Fast
Ciba-Geigy	Neolan

The 1 : 1 metal-complex dyes are also produced in large amounts in the COMECON countries and in the People's Republic of China; manufacturers in, for example, India or Greece are of lesser significance.

The manufacture of 1 : 1 chrome complex dyes takes place in two operations: manufacture of the azo dye capable of chrome development and formation of the chromium complex. Similarly as for the one-bath chroming dyes, negatively substituted *o*-aminophenols but also *o*-aminonaphthols and their sulfonic acids are used as diazo components; 2-naphthol, 2-naphtholsulfonic acids, pyrazolones, acetoacetic acid aryl amides, and 2,4-dihydroxyquinoline serve as the coupling components.

1 : 2 Metal-Complex Dyes. Because of their complexed structure 1 : 2 metal-complex dyes exhibit anionic character. Those which have gained commercial importance are primarily the ones that are free of sulfonic acid groups and for which adequate water solubility is provided by nonionic, hydrophilic substituents, such as methylsulfone or sulfonamide groups [15, p. 689], [25].

The introduction of 1 : 2 metal-complex dyes, which are applied from a neutral to weakly

acid bath, represented a significant technical advance over the strong-acid-dyeing 1 : 1 chrome complex dyes. It has meant better protection of the fiber material, simplification of the dyeing process, and improvement of the fastness properties.

Because these dyeings, like all metal-complex dyeings, usually exhibit subdued shades, the 1 : 2 metal-complex dyes are combined with small amounts of acid dyes or suitable reactive dyes to brighten the tint.

Owing to their outstanding wetfastness, high light fastness, and good fiber levelness, the 1 : 2 metal-complex dyes are of major importance in dyeing wool and polyamide. Important ranges of these dyes are listed in Table 4.

Table 4. Important ranges of 1 : 2 metal-complex dyes

Manufacturer	Range
BASF	Ortolan
Bayer	Isolan
Ciba-Geigy	Irgalan
Sandoz	Lanasyn

The extent of variation is extremely large as a result of the great number of available diazo and coupling components, the choice of complex-forming metal (primarily chromium or cobalt), and the possibility of synthesizing 1 : 2 metal mixed complexes.

Substituted *o*-aminophenols and also anthranilic acid and its derivatives are the most important diazo components. The main coupling components are pyrazolones, acetoacetic acid arylamides, naphthols, N-substituted aminonaphthols, and naphthylamines.

For subdued shades the 1 : 2 metal-complex dyes with one or two sulfo groups continue to play an important role. Dyes with one such group contain the sulfonic acid group in one of the two parts bound to the azo group and complexed to chromium or cobalt. They are built up stepwise via the 1 : 1 metal-complex. Compared with the 1 : 2 metal-complex dyes with nonionic hydrophilic groups they exhibit better solubility in water but poorer leveling power, and therefore they are always used together with specially selected leveling auxiliaries. The ranges of these dyes (Table 5) contain primarily navy blue, black, dark brown, and olive shades.

Table 5. Important ranges of 1 : 2 metal-complex dyes containing monosulfonic and disulfonic acid groups

Manufacturer	Range
BASF	Acidol M
Bayer	Isolan S
Ciba-Geigy	Lanacron S
Hoechst	Azarin
ICI-Francolor	Neutrichrom S
Sandoz	Lanasyn S

The 1 : 2 metal-complex dyes with two sulfonic acid groups are a more recent development.

Contrary to the opinion frequently expressed in the literature that such dyes would exhibit inadequate levelness when applied in a weakly acid bath and inadequate stability to acid in a strongly acid dyebath, they are very well suited for dyeing wool and polyamide if certain pH conditions are observed. These products are very soluble in water, have high tinctorial strength, and are very high-yield dyes with a relatively simple structure (for further details → Metal-Complex Dyes).

3.2.2. Polyamide Dyes [5, vol. III, p. 276]

Synthetic polyamides have a structure similar to those of wool and silk but differ in having a low acid-binding power and in their capacity to dissolve nonpolar compounds. In consequence materials of polyamide can be dyed with disperse dyes as well as with selected acid dyes, including metal-complex dyes.

When choosing *acid dyes* for polyamides it must be borne in mind that owing to the lower acid-binding power dyes with two and more sulfonic acid groups in the molecule go onto the fiber much more slowly and to a much lower saturation value than dyes with one sulfonic acid group. Consequently, these dyes cannot be mixed or combined with one another, which has been taken into account in composing the product ranges.

Since acid dyes on polyamide behave as they do on wool in regard to leveling power, build-up, and fastness properties (see Section 3.2.1), they fall into two classes (cf. Table 6):

Group A: acid dyes with good leveling power and low substantivity for polyamide; they yield dyeings with reasonably good wetfastness.

Group B: acid dyes with lower leveling power, higher substantivity and high wetfastness standard on polyamide. Many of these acid dyes with a lower leveling power rather clearly reveal differences in fiber structure that may result from, for example, differences in the degree of drawing (so-called streakiness), so that it is usually necessary to add leveling and retarding auxiliaries (→ Textile Dyeing).

Table 6. Important ranges of polyamide dyes

Manufacturer	Range	
	Group A	Group B
BASF	Acidol	Acidol
Bayer	Telon Supra	Telon Fast
Ciba-Geigy	Erionyl Tectilon	Erionyl
Crompton & Knowles	Nylanthrene	Nylanthrene
ICI	Nylomine A Nylomine B	Nylomine C
Sandoz	Nylosan E	Nylosan F Nylosan N

The main area of application for the acid polyamide dyes characterized in group A is in carpet dyeing, but they are also used in other areas of textile dyeing where the fastness requirements are not too stringent. A few structures are presented below to illustrate the types of azo dyes used: *C. I. Acid Yellow 25,* 18835 (**63**) [*6359-85-9*], *C. I. Acid Orange 67,* (**64**) [*12220-06-3*], *C. I. Acid Red 42,* 17070 (**65**) [*6245-60-9*], *C. I. Acid Red 32,* 17065 (**66**) [*6360-10-7*].

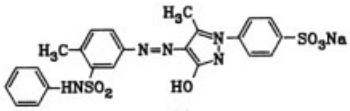

63

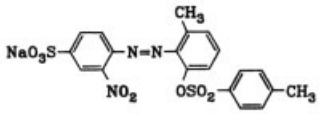

64

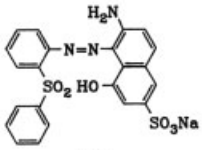

65

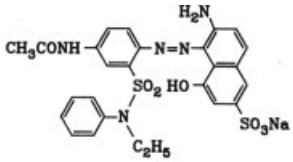

66

Group B dyes are used almost exclusively for clothing textiles for which more stringent requirements are placed on wetfastness. Some of the typical azo dyes of this type are *C. I. Acid Yellow 42*, 22910 (**67**) [*6375-55-9*], *C. I. Acid Yellow 65,* 14170 (**68**) [*6408-90-8*], *C. I. Acid Red 114,* 23635 (**69**) [*6459-94-5*], *C. I. Acid Blue 113,* 26360 (**70**) [*3351-05-1*].

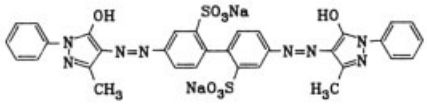

67

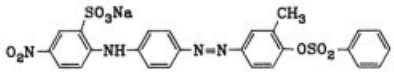

68

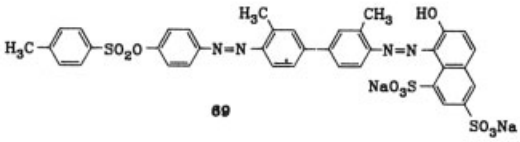

69

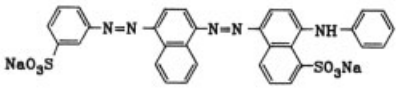

70

Other acid disazo dyes for polyamide are secondary disazo dyes (D → M → K) with one or two sulfonic acid groups bound to aromatic nuclei or with an external sulfate group, which is situated mainly in the diazo component D. Aniline, 1-naphthylamine, and their derivatives preferably serve as the middle component M, whereas the coupling components of the phenol and arylamine series common in disperse dyes are used as the final hydrophobic component K [26–30], as illustrated by the general structure **71**.

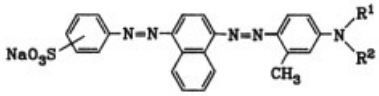

71

R^1 and R^2 = alkyl groups, partially substituted

The patent literature indicates that polyamide dyes of this type are generally characterized by high tinctorial strength and light fastness and a good capacity to compensate for nonuniformity in the material.

Because of increasing standards of wetfastness, lightfast *reactive dyes* are also being used in polyamide printing, especially to produce pastel shades. However, only selected representatives of the cotton reactive dyes so far offered on the market are suitable for polyamide. Essentially the products recommended for polyamide contain monosulfonic and disulfonic acid groups with β-sulfatoethylsulfone, α-bromoacrylamide, α,β-dibromopropionylamide, monochlorodialkylaminotriazine, and trihalopyrimidyl groups, particularly 2,4-difluoro-5-chloropyrimidyl groups (→ Reactive Dyes).

Because of their high fastness standards the *1 : 2 metal-complex dyes* have also acquired some importance in the dyeing of polyamide. They are used primarily in the clothing sector to achieve deep shades. The first complete range of these dyes for polyamide fibers was developed in 1940 – 1943 at the Ludwigshafen works of the then I. G. Farbenindustrie, but owing to the events of the day they never reached the market. The range consisted of dyes with and without nonionic, hydrophilic groups.

Later these dye types were further developed by other companies and uniform ranges were placed on the market. Dyes with nonionic, hydrophilic groups are offered for wool and polyamide dyeing (see page 30), while the use of water-insoluble 1 : 2 metal-complex dyes is restricted to polyamide dyeing; Table 7 shows ranges of these products. In a process comparable to disperse dyeing, these dyes are applied to polyamide from a finely dispersed aqueous suspension and yield extremely lightfast and wetfast dyeings (for further details → Metal-Complex Dyes).

Table 7. Important ranges of dispersed 1 : 2 metal-complex dyes

Manufacturer	Range
BASF	Vialon
Ciba-Geigy	Avilon
ICI-Francolor	Amichrom

3.2.3. Silk Dyes

Because natural silk, as a protein fiber of animal origin, resembles wool fiber in its chemical structure, it can be dyed with nearly all the classes of dyes used for wool. The choice of dyes depends essentially on the fastness properties required.

Of great importance for the dyeing of natural silk are selected members of the class of acid wool dyes (see Section 3.2.1). The occasionally inadequate wetfastness of these dyeings can be substantially improved by the proper aftertreatment (e.g., with potassium sodium tartrate and tannin).

Selected acid 1 : 1 chrome complex dyes, because of their good wetfastness and light fastness and their very good leveling power, and also 1 : 2 chrome complex and 1 : 2 cobalt complex dyes with hydrophilic groups (see page 30) have successfully come into use as silk dyes.

Furthermore, direct dyes (see Chap. 4) are among the most important silk dyes. Their fastness properties can be improved by aftertreatment with metallic salts or formaldehyde.

If suitably selected, cationic dyes (see Chap. 5) can also be used for dyeing weighted silk, toward which they exhibit special affinity, in addition to having good leveling power and very good wetfastness. On the other hand, the light fastness of these dyeings is not always satisfactory.

Reactive dyes have also found a place in silk dyeing. These dyeings are characterized by very fast and bright shades with exceptionally good wetfastness.

Initially the development of synthetic fibers greatly reduced the importance of silk dyeing. The processing of silk has undergone a marked increase owing to the growing quality consciousness of buyers, who appreciate the outstanding wear properties of silk, and because of the intensified activities of China on the world market.

In addition to China the countries that are still prominent in silk processing today are Japan and Italy.

3.2.4. Leather Dyes

Leather is produced by tanning skins of different origin and the tanning process has a decisive influence on the dyeability of the pretreated products. The differences in the dyeing behavior of leather of various origins that has been tanned in different ways are about as marked as those of the different textile fibers.

Many of the anionic, cationic, metal-complex, and direct dyes employed in the textile field are also used for dyeing leather. In addition ranges have been developed that are particularly suitable for dyeing leather. The choice of dye depends largely on the tanning method, the fastness properties required, and the end use, and a distinction is made, depending on the depth of penetration of the dyes, between surface dyeing, part-penetration dyeing, and full-penetration dyeing (for more details on the dyeing of leather → Leather).

The main shades used are different browns and blacks, as well as bordeaux, reds, yellows, blues, and greens. Most of the leather dyes belong to the azo class. They are marketed both as individual dyes and in the form of mixtures. The most important ranges are listed in Table 8. Because of the many different classes of dyes used, several ranges are available:

Group A includes both anionic and direct dyes. They are used mainly for dyeing chrome leather and alum-tanned glove leather. The basic difference between acid and direct dyes, which is pronounced in textile dyeing, is less marked in the dyeing of leather. Therefore, the two groups often tend to be combined under the term "anionic dyes". The dyeings obtained have good light fastness and good to very good wetfastness, rub fastness, and fastness to washing.

Group B contains acid 1 : 1 chrome-complex dyes as well as 1 : 2 chrome- and 1 : 2 cobalt-complex dyes with hydrophilic groups (see Section 3.2.1.4). Particularly stringent demands can be made on these dyes in terms of light, wet, and perspiration fastness and leveling properties. Some of the ranges used here are already familiar from the textile field.

Group C is composed of acid dyes and direct dyes. Unlike group A, they are mainly dye mixtures that give marked coloration of the leather as well as good surface coverage.

Group D consists of pigment preparations that are important for the treatment of leather and are applied by different methods (spraying, padding, curtain coating). The ranges contain both organic and inorganic pigments and mixtures of the two. They are marketed as binder-containing or binder-free, finely dispersed aqueous pigment preparations and have high covering power and in most cases very good light fastness.

The large number of *cationic* dyes suitable for dyeing leather cannot be grouped under one specific range. They have very good covering power and give a high yield and brilliant shades. Their light fastness is generally only moderate. They are used for vegetable tanned leathers and as a so-called top dyeing for chrome leather.

To obtain the good fastness to buffing that is particularly important in the dyeing of suede, special dyes are used which, in addition to adequately deep and level surface dyeing, also give good penetration so that no change in shade occurs when the leather is subsequently buffed.

Some typical *anionic* azo dyes for leather are: *C. I. Mordant Brown 33,* 13250 (**72**) [*3618-62-0*]: strong penetration, poor coverage;

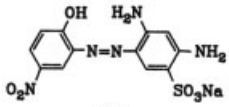

72

C. I. Acid Brown 105, 33530 (**73**) [*8003-78-9*]: poor penetration, good coverage;

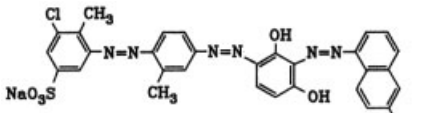

73

C. I. Acid Black 24, 26370 (**74**) [*3071-73-6*]: no penetration, good coverage;

Table 8. Names of important ranges of leather dyes and pigments

Manufacturer	Ranges			
	Group A	Group B	Group C	Group D
BASF	Luganil	Luganil C Eukesolar liq.	Lurazol Lumin	Lepton Helizarin Corial Fast
Bayer	Baygenal	Baygenal L	Aciderm Benzo Leather Chrome Leather	Eusin Eukanol Baysin Bayderm Isoderm
Ciba-Geigy	Sella Fast	Sellacron Neolan Lanacron Irgalan Irgaderm liq.	Sellaflor	Irgafin
FMC/ICI	Coriacide Inoderm			
Hoechst (with Cassella)	Coranil	Coranil Fast	Remaderm	Melustral Melustral Coating
Sandoz	Derma	Derma Light Sandoderm		Relca Relcasyn Relcasol

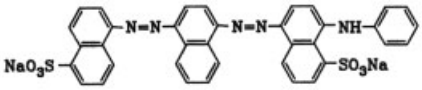

74

C. I. Acid Yellow 36, 13065 (**75**) [*587-98-4*] (metanil yellow): strong penetration, poor coverage;

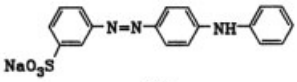

75

3.2.5. Azo Food Dyes

In terms of quantity and value food dyes play only a relatively minor role in the dye-making industry.

A brief survey is given here of the azo dyes that have been approved for food coloration. For a detailed treatment of the subject, including legislation and official regulations, → Foods, 3. Food Additives.

From the chemical viewpoint some of the most frequently used synthetic food dyes belong to the azo series.

The dyes approved for food coloration in the EEC are listed in the guidelines for colorants of 23. Oct. 1962, which are continually updated. The member states of the EEC must convert these guidelines to national law.

The azo dyes listed below bear, in addition to the L number of the Dye Commission of the Deutsche Forschungsgemeinschaft (DFG, German Research Association), an EEC number (E) under which they are registered in the EEC: *C. I. Food Yellow 4,* 19140, L-Yellow 2, E 102, Tartrazine, (**76**) [*1934-21-0*]

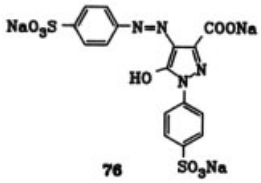

76

C. I. Food Yellow 3, 15985, L-Orange 2, E 110, Yellow Orange S, (**77**) [*2783-94-0*]

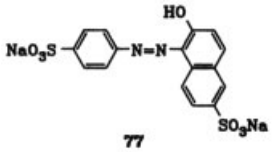

77

C. I. Food Red 9, 16185, L-Red 3, E 123, Amaranth S, (**78**) [*915-67-3*]

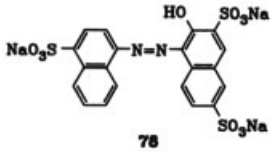

78

C. I. Food Red 7, 16255, L-Red 4, E 124, Ponceau 4 R, (**79**) [*2611-82-7*]

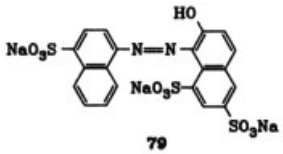

79

C. I. Food Red 3, 14720, L-Red 1, E 122, Azo Rubine, (**80**) [*3567-69-9*]

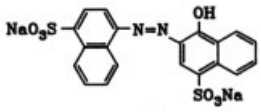

80

C. I. Food Black 1, 28440, L-Black 1, E 151, Brilliant Black BN, (**81**) [*2519-30-4*]

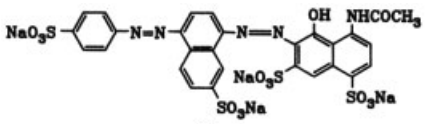

81

The use of dyes for pharmaceutical preparations is regulated in the EEC by guidelines from 1978 and 1981 and the Directive concerning Pharmaceuticals Dyes of 1982. Here practically the same dyes are permitted as for food.

For other areas of application different regulations apply. These regulations differ from one another in respect of the dyes permitted (for example, in the cosmetics sector the cosmetics guidelines and the Directive on Cosmetics).

The food dyes permitted in the USA are compiled in the Code of Federal Regulations, Vol. 21, Food and Drugs, Parts 1 to 99; continual amendments are published in the Federal Register.

A synthetic dye is permitted for dyeing food only if thorough toxicological studies reveal no danger of toxic effects to the consumer. Food dyes are among the food additives that have been subjected to the most thorough toxicological examinations.

Purity Requirements: Food dyes that are placed on the market must meet special purity requirements, which have been fixed by law for the Federal Republic of Germany in the Directive on Additives of 20. 12. 1977. According to this regulation the following substances must not be detectable in food dyes:Barium compounds that are soluble in dilute hydrochloric acid; cadmium, mercury, selenium, tellurium, thallium, uranium and chromates; polycyclic aromatic hydrocarbons with 3 or more fused nuclei; and 2-naphthylamine, benzidine, and 4-aminobiphenyl (xenylamine).

Upper limits have been specified for the following contaminants:

arsenic		max. 5 mg/kg
lead		max. 20 mg/kg
antimony, copper, chromium, zinc, barium sulfate	individually	max. 100 mg/kg
		max. 200 mg/kg
	together	
aromatic amines		max. 100 mg/kg
portions extractable with ethyl ether∗		max. 0.2 %
water-insoluble portions∗		max. 0.2 %

∗ Only in the dyes E 102 – 110, E 122 – 132, E 142 and E 154.

Furthermore, special purity requirements have been placed on individual dyes.

In the Federal Republic of Germany the analytical methods for testing the degree of purity have been published by the Dye Commission of the DFG.

In the USA the regulations for the purity of food dyes are specified in the Code of Federal Regulations. The Food and Drug Administration (FDA) is responsible for the registration and certification of food dyes.

4. Direct (Substantive) Dyes

Direct or substantive dyes are colored compounds that are mainly used to dye materials made from natural or regenerated cellulose (e.g., cotton, jute, viscose, or paper) without employing mordants as auxiliaries. The essential requirement for classification of a dye in this group is its substantivity, i.e., its absorption from an aqueous salt-containing solution onto cellulosic materials. Absorption onto cotton takes place in a neutral to soda alkaline medium and onto paper in a weakly acid to neutral medium.

Substantivity was initially attributed to secondary valence bonding between fiber and dye. The fact that coplanar molecules are always more substantive than nonplanar ones later led to the coplanarity theory with its assumption that coplanar dyes are in contact with the cellulose molecule along their entire length.

The presence of hydrogen bonds has also been expounded as a possible explanation for high affinity between fiber and dye [31]; however, such bonds are probably prevented by a water layer between fiber and dye [32].

The nature of substantivity has been very convincingly explained [33].

According to this explanation, single dye molecules are adsorbed by the intermicellary cavities of the cellulosic fibers and, unlike nonsubstantive dyes, they form aggregates in these cavities.

Because of their size, these aggregates can no longer be directly washed out with water, but only after further solvation has taken place.

Because direct dyes become aggregated in aqueous solutions at normal temperatures, substantivity often cannot take effect until the temperature has risen. Only then diffusion into the fiber is possible.

The tendency toward aggregation is therefore characteristic of substantive dyes, which also explains why coplanar dyes possess greater substantivity than nonplanar ones.

There is no exact delineation between substantive and nonsubstantive dyes, the boundaries between them are fluid.

Structural Characteristics. Azo dyes comprise the major proportion of the direct dyes; apart from these, only a few azine, phthalocyanine, and nonazo metal-complex dyes possess a certain significance. The following structural characteristics are necessary for high substantivity:

a) *Coplanarity of the Dye Molecules.* The importance of coplanarity is shown by the example of 4,4′-diaminodiphenyl derivatives. Whereas 3,3′-disubstituted diaminodiphenyls (e.g., *o*-tolidine or *o*-dianisidine) as diazo components produce high absorptive power in dyes, this substantivity is neutralized in 2,2′-disubstituted derivatives and coplanarity of these molecules is no longer possible.
b) *Long Chain of Conjugated Double Bonds.*Dye molecules of this kind (see also Schirm's rule [34]) can contain bridges, such as $-CH{=}CH-$, in addition to the aromatic rings, thus ensuring the development of resonance forms across the entire molecule. Exceptions to this rule are the bridges $-NHCONH-$, $-CONH-$, and $-NH-$, which increase substantivity in spite of the interruption in the conjugated system. In these groups, the ability to form hydrogen bonds is the decisive factor. Other groups causing an interruption in the conjugated system, e.g., $-CH_2-$, $-CH_2CH_2-$, $-CO-$, $-S-$, $-SO_2-$, and cyclohexylene, decisively lower substantivity. Only the substantivity of the linked segments remains.

On the basis of the structural characteristics described, direct dyes based on benzidine can be commended for their good substantivity. In fact this group was for a long time a significant constituent of all important direct dye ranges in the shade spectrum from red to brown, blue, green, and black.

Because of the carcinogenic properties of unsubstituted benzidine in humans about 15 – 20 years ago almost all major dye manufacturers in the most important western industrialized countries stopped production of dyes based on this substance.

This development led to an intensive search for substitute dyes, the results of which are reflected in particular in the patent literature. The significance of this development is also discernible in the number of direct dyes newly registered with the Colour Index between 1971 and 1982. The research efforts are especially apparent in the case of black, with 21 new registrations; yellow, brown, and blue, each with 19;

and red, with 17. However, of a total of 115 direct dyes, the constitution has as yet only been stated in 7 instances. It is therefore not possible to evaluate the patent literature results with regard to the industrial significance of the products.

Starting Compounds. Based on the requirements mentioned, the following starting compounds are especially suited for the synthesis of direct dyes:

a) 3,3′-Disubstituted 4,4′-diaminodiphenyl derivatives, but not 2,2′-disubstituted ones. These are, therefore, diamines with the structure

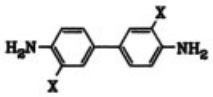

where $X = CH_3$, OCH_3, Cl, SO_3H. 1,4-Diaminophenyl derivatives also belong to this group.

b) Compounds of the type

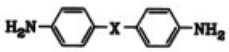

where $X = NH$, SO_2, $-N{=}N-$, $-CONH-$, $-CSNH-$, $-SO_2NH-$, $-CH{=}CH-$, $-HNCONH-$, $-HNCSNH-$, $-O_2SNHSO_2-$.

c) 1,4-Diaminonaphthalene and 1,5-diaminonaphthalene

d) *m*-Phenylenediamine and other 1,3-disubstituted diamines, e.g., 2,6-diaminotoluene-4-sulfonic acid or bis(3-aminophenyl)urea.

e) Heterocyclic diamines

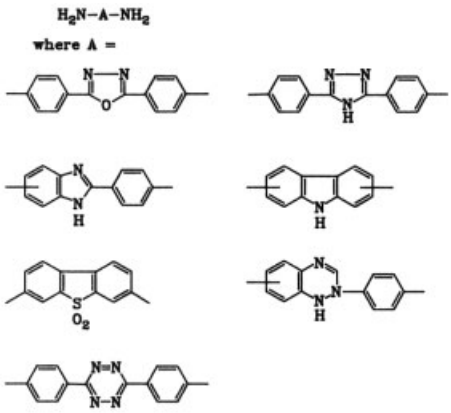

Because of their ease of application and moderate price, direct dyes still represent one of the largest groups of azo dyes, although their wetfastness, in particular, often only satisfies moderate to average requirements. The only slight decrease in consumption of direct dyes following the appearance of the wetfast reactive dyes is due to the markedly higher price of reactive dyes. For product ranges see Table 9.

Direct dyes are classified according to their method of application:

a) *Conventional Direct Dyes.* They include monoazo, disazo, trisazo, and tetrakisazo dyes. It is advantageous to subdivide them according to the nature of their chemical structure. Disazo dyes, for example, can be divided according to chemical synthesis principles into primary and secondary. Conventional azo direct dyes further include symmetric urea derivatives, dyes obtained by oxidation of amines, and triazinyl and copper-containing dyes.

b) *Direct Dyes with Aftertreatment.* This group includes direct dyes that after being applied to the fiber by the usual method undergo one of the following aftertreatments:
   Aftertreatment with cationic auxiliaries
   Aftertreatment with formaldehyde
   Diazotization of the dye on the fiber and coupling with suitable components (diazotization dyes)
   Aftertreatment with metal salts

4.1. Conventional Direct Dyes

4.1.1. Monoazo Dyes

Only a few monoazo dyes possess the necessary substantivity to render them suitable for direct dyeing of cellulosic fibers.

Of practical importance as diazo components are only the derivatives of 2-(4′-aminophenyl)-6-methylbenzthiazole, of the so-called dehydrothio-*p*-toluidine (**82**):

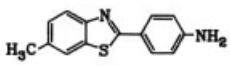

82

Table 9. Designations of important ranges of direct dyes (The names in brackets are the U.S. products of the relevant Western European company)

Manufacturer	Product range
Western Europe	
BASF (FRG)	Lusantin, (Phenamine)
Bayer (FRG)	Benzamin, Benzo, Benzocuprol, Sirius, Sirius Supra
Ciba-Geigy (Switzerland)	Cuprophenyl, Diphenyl, Solophenyl, (Diazophenyl)
Croda (United Kingdom)	Covazol, Diazin
Hoechst/Cassella (FRG)	Cotonerol, Diamin, Duasyn, (Rayon)
ICI (United Kingdom)	Chlorazol, Durazol
PCUK/ICI (France)	Cuprodiazol Light, Diazamine, Diazol, Diazol Fast, Diazol Light, Metadiazol
Rovira (Spain)	Hispadiazo, Hispaluz, Hispamin
Sandoz (Switzerland)	Cuprofix, Cuprofix C, Indosol, Solar, (Lumicrease, Pyrazol)
United States	
Amer. Cyanamid	Calcomine
Atlantic	Atlantic, Atlantic Diazo, Atlantic Direct, Atlantic Resin
Crompton & Knowles	Diazo, Direct-, Intrabond, Intralite, Intramet, Sol-Aqua-Fast, Superlitefast
Internat. Dyestuff	Elcofast, Elcomine
Org. Chem. Corp.	Orcoform, Orcolitefast, Orcomine
Japan	
Hodogaya	Aizen Direct, Aizen Primula, Direct
Mitsubishi	Diacotton, Diacupro, Dialuminous, Direct
Mitsui Toatsu	Direct, Suprazo
Nippon Kayaku	Kayafect, Kayaku Direct, Kayarus
Sumitomo	Direct, Japanol, Nippon, Sumilight, Sumilight Supra

The compound **82** is obtained by melting *p*-toluidine with sulfur at 130 – 230 °C. The derivatives (**83**) – (**86**) are used to produce monoazo direct dyes:

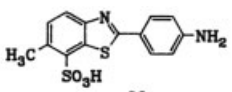

83

Dehydrothio–p–toluidine–sulfonic acid

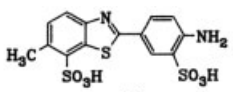

84

Dehydrothio–p–toluidine–disulfonic acid

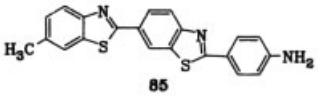

85

Primulin base

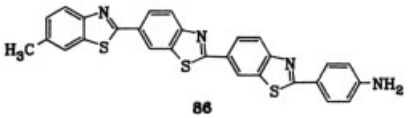

86

Whereas **83** and **84** are obtained by sulfonation of dehydrothio-*p*-toluidine (**82**) with oleum, compounds **85** and **86** are formed in a mixture with **82** during the sulfur melting process and are separated by vacuum distillation.

Acetoacetic acid arylides and pyrazolone derivatives are the coupling components mainly used. Naphtholsulfonic acids are of only minor importance here. Dehydrothio-*p*-toluidinesulfonic acid (**83**) and the sulfonated primulin base are also used as diazo components for certain orange disazo direct dyes.

Examples. Diazotization of dehydrothio-*p*-toluidinesulfonic acid (**84**) and coupling to acetoacetic acid *o*-anisidide results in the lightfast yellow *C. I. Direct Yellow 27,* 13950 (**87**) [*10190-68-8*].

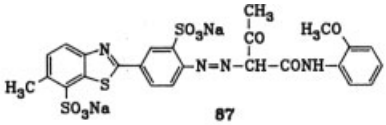

87

Coupling the diazotized monosulfonic acid of the primulin base to acetoacetic acid anilide produces the clear, greenish yellow *C. I. Direct Yellow 22,* 13925 (**88**) [*10190-69-9*].

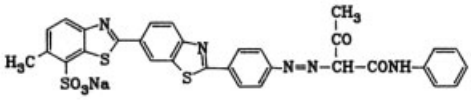

88

Substantive yellow monoazo dyes with very good light fastness are also obtained by oxidation of dehydrothio-*p*-toluidine and its deriva-

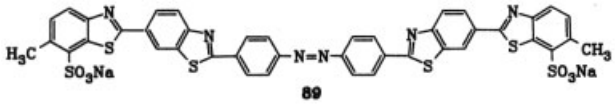

89

tives (see Section 2.2.3). Oxidation of the sodium salt of the monosulfonic acid of the primulin base in the alkaline pH range with sodium hypochlorite results, for example, in *C. I. Direct Yellow 29,* 19556 (**89**) [*6537-66-2*].

As with the corresponding oxidation of dehydrothio-*p*-toluidinesulfonic acid (**83**), which results in *C. I. Direct Yellow 28,* 19555 (see Section 2.2.3), in addition to the azo dye, diphenazines (**90**) are obtained as byproducts, e.g., which in some cases have to be separated

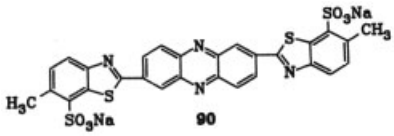

90

when they impair the shade and fastness properties of the dyes.

4.1.2. Disazo Dyes

4.1.2.1. Primary Disazo Dyes

Primary disazo dyes are obtained in accordance with the pattern $K_1 \leftarrow D \rightarrow K_2$ by bisdiazotization of a bifunctional diazo component D and coupling to two equivalents of coupling component K_1 and K_2, or according to the pattern $D_1 \rightarrow K \leftarrow D_2$ by coupling of two equivalents of diazo component (D_1 and D_2) to one equivalent of a bifunctional coupling component. Depending on whether K_1 and K_2 or D_1 and D_2 are the same or different, symmetrical or asymmetrical dyes are obtained.

Type $K_1 \leftarrow D \rightarrow K_1$ *Primary symmetrical disazo dyes* with benzidine as bisdiazo component were long considered to be the prototype of direct dyes. Development of this dye class commenced with congo red, *C. I. Direct Red 28,* 22120 (**91**) [*573-58-0*].

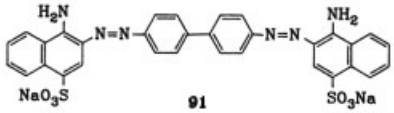

91

Both as regards the number of products on the market and the consumption in terms of volume, the disazo dyes of the bisdiazotized benzidine were for a long time the most important substantive dyes. Today, no individual chemical group enjoys such significance. Instead, besides the substituted 4,4′-diaminodiphenylene, several other groups, including heterocyclicones, have taken the place of benzidine.

o-Dianisidine as bisdiazo component has a certain significance in blue direct dyes of the congo red type. Coupling with two equivalents of H acid results in *C. I. Direct Blue 15,* 24400 (**92**) [*2429-74-5*].

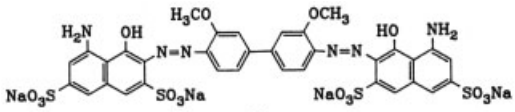

92

With chromotropic acid (1,8-dihydroxynaphthalene-3,6-disulfonic acid) as coupling component, *C. I. Direct Blue 10,* 24340 [*4198-19-0*], is obtained (see page 41).

A very clear pure blue is obtained through the alkaline coupling of bisdiazotized *o*-dianisidine to two equivalents of Chicago acid SS (1-amino-8-hydroxynaphthalene-2,4-disulfonic acid): *C. I. Direct Blue 1,* 24410 (**93**) [*2610-05-1*].

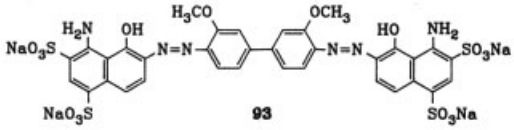

93

o-Tolidine is also occasionally encountered as diazo component, e.g., in *C. I. Direct Blue 14,* 23850 (**94**) [*72-57-1*].

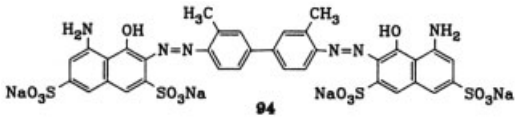

94

The dye enjoyed major significance and in addition was one of the first chemotherapeutics to be used as a remedy for protozoal diseases (Trypan Blue).

Other important diamines used for the synthesis of symmetrical substantive disazo dyes are 4,4′-diaminostilbene-2,2′-disulfonic acid (**95**) and 4,4′-diaminodiphenylurea-3,3′-dicarboxylic acid or -3,3′-disulfonic acid (**96**):

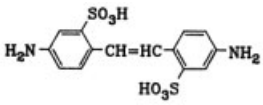

95

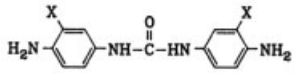

96

X = COOH or SO_3H

The diamine **95** is obtained by heating 4-nitrotoluene-2-sulfonic acid in diluted sodium hydroxide solution with sodium hypochlorite, 4,4′-dinitrostilbene-2,2′-disulfonic acid being formed in the process, and subsequent reduction with iron and hydrochloric acid. Compounds **96** can be produced from 4-nitroaniline-2-carboxylic acid or -2-sulfonic acid by treatment with phosgene and subsequent reduction. The sodium salt of 4,4′-dinitrostilbene-2,2′-disulfonic acid itself belongs to the group of stilbene dyes. It is mainly used as *C. I. Direct Yellow 11,* 40000 [*1325-37-7*] for the coloration of paper.

Stilbene dyes are different self-condensation products of 4-nitrotoluene-2-sulfonic acid produced with of sodium hydroxide solution. The various types are synthesized depending on the concentration of sodium hydroxide, the reaction temperature, and the reaction time. In some cases a subsequent reaction with arylamines is carried out.

An example of the combination with **95** is *C. I. Direct Yellow 12,* 24895 (**97**) [*2870-32-8*], also known as chrysophenine:

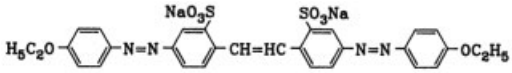

97

The product is obtained by coupling bis-diazotized 4,4′-diaminostilbene-2,2′-disulfonic acid with two equivalents of phenol and subsequent ethylation. It is a very high-strength reddish yellow with very good leveling power and average fastness properties, which is excellently suited for the dyeing of fabric blends, e.g., cotton – polyamide. For manufacturing specifications see below.

An example of the use of **96** as diamine is *C. I. Direct Red 75,* 25380 (**98**) [*2829-43-8*].

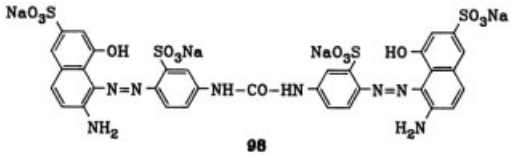

98

Other direct dyes with the diphenylurea group can be produced from aminomonoazo dyes by subsequent treatment with phosgene (see Sections 2.3.2 and 4.1.6).

Type $K_1 \leftarrow D \rightarrow K_2$ *Primary Asymmetrical Disazo Dyes.* Since the coupling reactions of substituted 4,4′-diaminodiphenylene to monoazo or disazo dye proceed at different rates, asymmetrical coupling reactions can be carried out. The first (acid coupling) proceeds from bis-diazotized *o*-dianisidine with one equivalent of a coupling component to a monoazo intermediate compound, which now together with a second (different) coupling component forms an asymmetrical disazo dye.

Example: *C. I. Direct Blue 22,* 24280 (**99**) [*2586-57-4*], is obtained by coupling *o*-dianisidine to Chicago acid SS and β-naphthol:

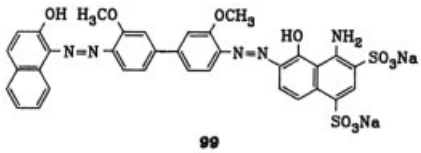

99

Production. The following abridged specifications are intended to illustrate processes for the

manufacture of symmetrical and asymmetrical disazo direct dyes.

C. I. Direct Yellow 12,

24895, (**97**) [*2870-32-8*]: 590 kg of 4,4′-diaminostilbene-2,2′-disulfonic acid are adjusted with water and ice to a volume of approximately 5500 L and a temperature of 5 °C and bisdiazotized with 220 kg sodium nitrite as 23 % solution; volume 6000 – 7000 L, temperature 20 – 22 °C.

Next, 312 kg of phenol are dissolved with 1500 L of water and 150 L of 30 % sodium hydroxide solution and 225 kg of soda are added. To this is transferred the above bisdiazotized solution. Then, 160 L of 30 % sodium hydroxide solution is added over 1 h. On the following morning, the solution is heated to 70 °C, and after addition of 240 L of 30 % hydrochloric acid, the dye is salted out with 12 % rock salt, calculated on the volume. Yield: 2090 kg.

Approx. 5500 kg of the damp press cake, corresponding to 2940 kg of 100 % product, 3500 kg of 94 % ethanol, 520 kg of soda, 870 L of 30 % sodium hydroxide solution, and 840 kg of ethyl chloride are kept in a closed stirrer vessel for 24 h at 100 °C (5 – 5.5 bar). The mixture is then cooled to 70 °C and transferred at its own pressure to a distilling vessel, from which the ethanol is distilled with direct steam via a column. After being cooled to 80 °C, the mixture is suction filtered and dried at 100 °C.

C. I. Direct Blue 15, 24400 (**92**) [*2429-74-5*]. 105 kg of dianisidine dihydrochloride is charged into 6000 L of water and 400 L of 31 % hydrochloric acid. Then 2000 kg of ice is added and the solution is bisdiazotized at 0 °C with 200 kg of sodium nitrite and stirred for 4 h.

To 210 kg of H acid as neutral solution 4000 L of a 20 % sodium carbonate solution is added and then 3000 kg of ice to cool to 0 °C. This solution is run in the bisdiazonium salt solution and stirred overnight.

Then the solution is heated to 80 °C, 4600 kg of sodium chloride is added, and the dye is filtered and dried in vacuo at 95 – 100 °C. Yield: 7600 kg.

Types $D_1 \rightarrow K \leftarrow D_1$ and $D_1 \rightarrow K \leftarrow D_2$

The most important coupling component for these types is the compound **100**, which is produced by phosgene treatment of the I acid.

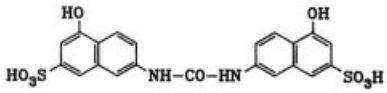

100

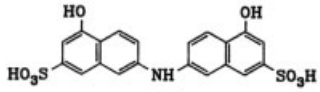

101

Dyes with a urea bridge, –HN–CO–NH–, are also obtained by phosgene treatment of aminoazo compounds (see Section 4.1.6). The so-called I acid imide (**101**), resorcinol and *m*-phenylenediamine are also of importance. Examples of dyes are: *C. I. Direct Orange 26,* 29150 (**102**) [*3626-36-6*],

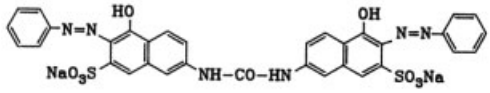

102

C. I. Direct Red 250, 29168 (**103**),

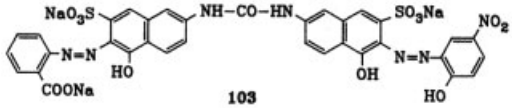

103

and *C. I. Direct Red 23,* 29160 (**104**) [*3441-14-3*].

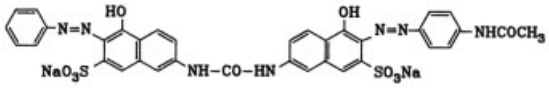

104

The dye **104** is obtained by simultaneous coupling of one equivalent each of diazotized aniline and 4-aminoacetanilide to I acid urea (**100**). Although it only possesses average fastness properties, it enjoys great importance and because of its good leveling power is used for dyeing union fabrics, polyamide, and chrome leather.

By coupling diazotized dehydrothio-*p*-toluidinesulfonic acid to resorcinol and subsequent coupling of diazotized aniline to the monoazo dye formed one obtains *C. I. Direct Orange 18,* 20215 (**105**) [*5915-59-3*]. The product likewise represents an important substantive dye.

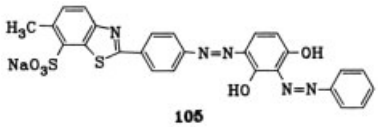

105

4.1.2.2. Secondary Disazo Dyes

Type D → M → K. For the manufacture of secondary disazo dyes, a diazotized amine D is coupled to an amine M with free *para* position, which in turn is diazotized and coupled to a coupling component K (series coupling).

Dyes of this type mostly have a straight-chain structure and often contain groups that increase substantivity, such as I acid, I acid derivatives, in particular with acylated amino groups, urea bridges, or benzoylamino groups. The shades in this range extend from orange through red, violet, and blue to black.

Examples: From 4-aminoazobenzene-4′-sulfonic acid and benzoyl I acid *C. I. Direct Red 81,* 28160 (**106**) [*2610-11-9*], is obtained. This dye represents an important red grade.

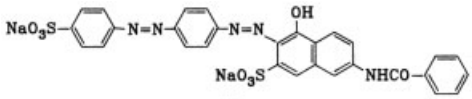

106

C. I. Direct Black 51, 27720 (**107**) [*3442-21-5*], is produced by coupling a mixture of diazotized 3- and 5-aminosalicylic acid to 1-naphthylamine and subsequent alkaline coupling to γ acid.

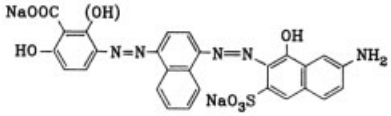

107

4.1.3. Trisazo Dyes

The trisazo dyes include in particular blue, green, and black grades. Benzidine, which here, too, used to be an important bisdiazo component, has been replaced by other diamines, such as *o*-tolidine, 4,4′-diaminobenzanilide (DABA), or 4,4′-diaminodiphenylamine-2-sulfonic acid (DADPS).

The most important synthesis principle in this range is the type $K_1 \leftarrow D_1 \rightarrow K_2 \leftarrow D_2$. The sequence of the three couplings depends on the nature of the components.

An example is *C. I. Direct Green 85,* 30387 (**108**).

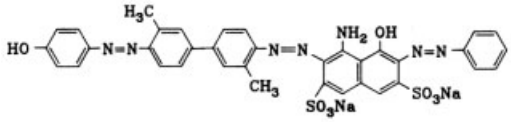

108

phenol $\xleftarrow{3}$ *o*-tolidine $\xrightarrow[a]{1}$ H acid $\xleftarrow[alk]{2}$ aniline

The reaction commences with stepwise coupling of the *o*-tolidine and then aniline with H acid, followed by coupling of this aminodisazo intermediate with phenol.

With *C. I. Direct Black 166,* 30026 (**109**), the first step is the acid coupling of 3,3′-diaminobenzanilide to H acid, followed by alkaline coupling of aniline to the monoazo dye and finally coupling of the disazo dye to *m*-phenylenediamine.

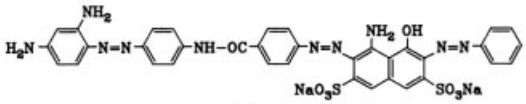

109

m-phenylenediamine $\xleftarrow{3}$ DABA $\xrightarrow[a]{1}$ H acid $\xleftarrow[alk]{2}$ aniline

C. I. Direct Black 150, 32010 (**110**) [*6897-38-7*], is produced by double (alkaline) coupling of bisdiazotized 4,4′-diaminodiphenylamine-2-sulfonic acid to γ acid and subsequent single coupling to *m*-phenylenediamine.

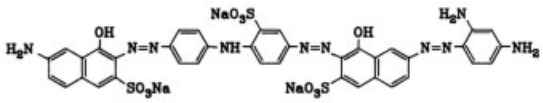

110

γ acid $\xleftarrow[alk]{1}$ DADPS $\xrightarrow[alk]{1}$ γ acid $\xrightarrow{2}$ *m*-phenylenediamine

Among the trisazo dyes, the synthesis principle of series coupling of the type $D \rightarrow M_1$

$\rightarrow M_2 \rightarrow K$ is also of importance. This group contains a number of dyes with good to very good general fastness properties.

Anilinesulfonic acids are chiefly used as starting component D, naphthylamine and Cleve acids as middle components M_1 and M_2, and I acid, its *N*-phenyl derivatives, H acid, γ acid, and their derivatives as final component K. This type mainly possesses blue and green shades. Example: *C. I. Direct Blue 78,* 34200 (**111**) [*2503-73-3*].

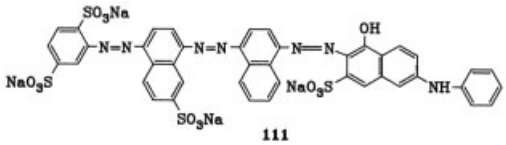

111

2-aminobenzene-1,4-disulfonic acid $\rightarrow$ Cleve acid-7 $\rightarrow$ 1-naphthylamine $\rightarrow$ *N*-phenyl I acid

In the manufacture of trisazo dyes, good yield and purity during final coupling are often obtained only in the presence of pyridine or other bases as coupling accelerators [35]. Intermediate isolation and separation of impurities prior to continuation of coupling are also frequently necessary.

The use of pyridine during final coupling (pyridine coupling) is illustrated in the following abridged manufacturing specification for *C. I. Direct Green 33,* 34270:

2-aminonaphthalene-8-sulfonic acid $\rightarrow$ 3,5-xylidine $\rightarrow$ 2-ethoxy Cleve acid-6 $\rightarrow$ *N*-acetyl H acid

Final Diazotization: The disazo dye produced by the usual method aminocroceic acid $\rightarrow$ sym. *m*-xylidene $\rightarrow$ 2-ethoxy-Cleve acid-6 (batch size: 81 kg aminocroceic acid 100 %) is stirred in the form of a press cake in 600 L of water with 500 L of 50 % acetic acid. Then 26.5 kg of sodium nitrite in the form of a 30 % solution is added at 25 °C. After stirring overnight, a reddish-brown solution is obtained. *Final coupling:* To 104 kg of H acid 100 %, dissolved in 500 L of water, are added 32 L of 40 % sodium hydroxide solution and 17 kg of soda; the mixture is heated to 70 °C and 50 kg of acetic acid anhydride is added. After acetylation is completed, the mixture is transferred to 1500 L of pure pyridine and 2000 kg of ice. The diazotizing solution is added at 0 – 12 °C over the course of 1/2 h. After a further hour, 375 kg of sodium hydrogen carbonate is added and after 1 h has expired, the mixture is heated to 60 °C by indirect steam. Following the addition of 8 % rock salt, the temperature is maintained at 60 °C for a further hour, and the dye is filtered and dried at 110 °C (drum drier or circulating air cabinet). Then yield is approx. 360 kg.

Other possible combinations are only of minor importance compared with the trisazo dyes manufactured according to the synthesis principles mentioned.

4.1.4. Tetrakisazo Dyes

Numerous syntheses are possible for these dyes, which all yield only dark shades. Here, too, only a few types are of industrial importance.

A further extension to the trisazo dyes on the principle of series coupling $D \rightarrow M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow K$ offers no advantages, because the intermediate isolation that is frequently necessary leads to yield losses and a chain extension is therefore ruled out on economic grounds. This is also reflected in the number of polyazo dyes given in the Colour Index [7, vol. 4, p. 4325]. Although 78 tetrakisazo dyes with eleven different synthesis principles are listed, only 14 dyes with five and more azo groups are mentioned, two of which are specified with eight azo groups (see top of next page).

An example of type $D \rightarrow M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow K$ is C. I. Direct Dye, 35850 [*8002-98-0*] (brown).

sulfanilic acid $\rightarrow$ *m*-toluidine $\rightarrow$ Cleve acid-6 $\rightarrow$ *m*-toluidine $\rightarrow$ 1-nitrophenylene-2,4-diamine

A number of important direct dyes are derived from type $K_1 \leftarrow M_1 \leftarrow D \rightarrow M_1 \rightarrow K_1$: C. I. Direct Black 22, 35435, (**112**) [*6473-13-8*] is produced via the primary disazo dye from 4,4′-diaminodiphenylamine-2-sulfonic acid and two equivalents of γ acid. This dye is bisdiazotized and coupled to two equivalents of *m*-phenylenediamine.

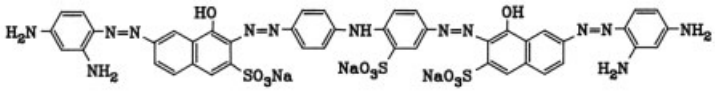

112

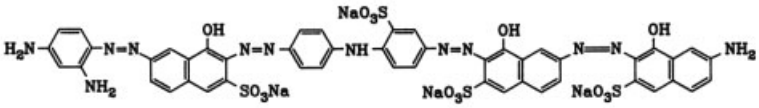

113

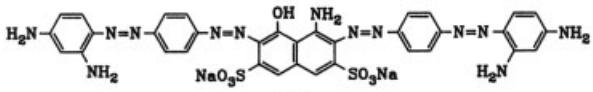

114

Pattern:

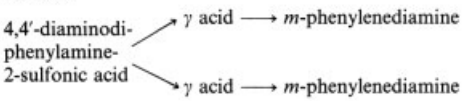

Another example is *C. I. Direct Black 151,* 35436 (**113**)

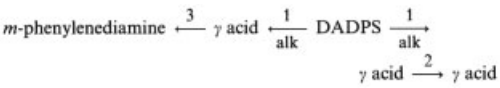

Here, too, the disazo dye is initially prepared from bisdiazotized 4,4′-diaminodiphenylamine-2-sulfonic acid with 2 equivalents of γ acid. Coupling is then again carried out stepwise to γ acid and finally to *m*-phenylenediamine.

C. I. Direct Black 19, 35255 (**114**) [*6428-31-5*], is also an important dye. It is obtained by coupling two equivalents of *p*-nitraniline to H acid, sulfhydrate reduction of the two nitro groups, and coupling of the bis-diazotized intermediate compound with two equivalents of *m*-phenylenediamine (see top of page 45).

4.1.5. Condensation Dyes

These dyes are produced by condensation of nitro compounds (see also Section 2.2.1).

Alkaline condensation of dinitrostilbenedisulfonic acid with aminoazo compounds produces a number of highly fast substantive dyes with industrial importance, especially in the shades orange, scarlet, and brown [36]. Example: *C. I. Direct Orange 39,* 40215 [*1325-54-8*], is obtained by condensation of 4-aminoazobenzene-4′-sulfonic acid with dinitrostilbenedisulfonic acid in aqueous sodium hydroxide solution. The main product formed is the tetrakisazo dye of the stilbene-2,2′-disulfonic acid (**115**), as well as the corresponding azoxy compounds. No uniform product is obtained, since part of the aminoazo compound is consumed as reducing agent. Condensation dyes are often purified by aftertreatment with reducing agents or such as glucose or sodium sulfide. Abridged specification: 468 kg of 100 % 4,4′-dinitrostilbene-2,2′-disulfonic acid as sodium salt is stirred with 1000 L of water and adjusted to a weakly alkaline pH value with sodium hydroxide solution. Then 440 L of 30 % sodium hydroxide solution is added and 346 kg of 100 % 4-aminazobenzene-4′-sulfonic acid introduced as sodium salt. After the contents of the tank have been diluted to 4500 L, the temperature is raised and maintained for 14 h at 100 °C. The condensation product is transferred into 10000 L of water and 39 kg of sodium sulfide is then added. The temperature is raised to 85 °C and maintained there for 1/2 h; then salting-out carried out with 12 % rock salt, calculated on the total volume. Isolation is carried out at 70 °C using a filter press. The dye is given a final washing

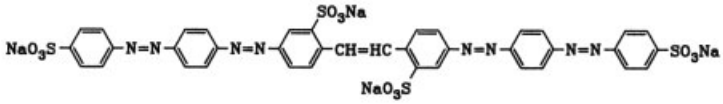

115

with 9000 L of a 10 % rock salt solution heated to 50 °C. The yield is approx. 780 kg dry.

Table 10 shows further examples of condensation dyes obtained from the corresponding aminoazo compounds and 4,4′-dinitrostilbene-2,2′-disulfonic acid.

4.1.6. Direct Dyes with a Urea Bridge

By the end of the 19th century, BASF had discovered that the reaction of aminoazo dyes with phosgene in an aqueous solution in the presence of soda resulted in valuable symmetrical urea azo dyes [37]:

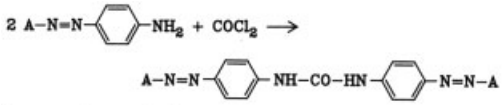

A = azo dye radicals

The urea bridge can also be introduced into direct dyes with 4,4′-diaminodiphenylurea derivatives by means of bisdiazotization and coupling (see (**98**) in page 40). *C. I. Direct Yellow 50,* 29025 (**116**) [*3214-47-9*]

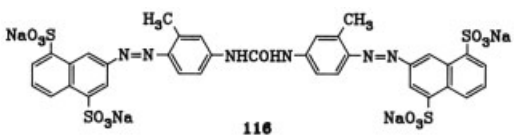

116

is obtained, for example, by phosgene treatment of two equivalents of aminoazo dye prepared from diazotized 2-aminonaphthalene-4,8-disulfonic acid coupled to *m*-toluidine.

Other examples of symmetrical direct dyes with a urea bridge are shown in Table 11.

Aminodisazo and aminotrisazo dyes can also be treated with phosgene, largely red and brown shades being obtained.

Example: *C. I. Direct Red 80,* 35780 [*2610-10-8*], is rendered accessible by phosgene treatment of the aminodisazo dye **117**:

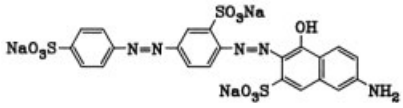

117

By mixing two different aminoazo compounds A–NH_2 and B–NH_2 and subsequent phosgene treatment, the asymmetrical diphenylurea direct dye **118** is obtained in addition to the two symmetrical members.

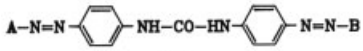

118

There are a number of yellow substantive dyes synthesized by this principle. Example: *C. I. Direct Yellow 41,* 29005 (**119**) [*8005-53-6*], is obtained by phosgene treatment of an equimolar mixture of bases A (sulfanilic acid → cresidine) and B (*p*-nitroaniline → salicylic acid and reduction of the NO_2 group). The mixture additionally contains the two corresponding symmetrical dyes.

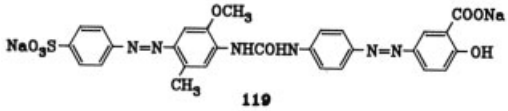

119

4.1.7. Triazinyl Dyes

The principle of linking two or three azo dyes by means of a triazine ring was established by Ciba patents [38]. The triazine bridge increases the substantivity of the dyes, in a similar manner to the −NH−CO−NH− group. Production is based on cyanuric chloride, the three chlorine atoms of which can be replaced in stages by nucleophilic radicals under different conditions (see Section 2.3.2).

The procedure that can be adopted is for intermediate products with amino groups (e.g., H

Table 10. Examples of condensation dyes made from aminoazo compounds and 4,4′-dinitrostilbene-2,2′-disulfonic acid

Aminoazo compound	Condensation dye	Aftertreatment with
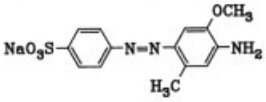	C.I. Direct Orange 40, 40265 [*1325-62-8*]	Na_2S
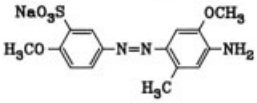	C.I. Direct Red 76, 40270 [*1325-63-9*]	glucose
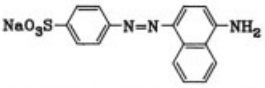	C.I. Direct Brown 78, 40290 [*1325-65-1*]	–

Table 11. Aminoazo compounds for symmetrical direct dyes with a urea bridge

Aminoazo dye	Direct dye
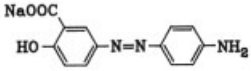	C.I. Direct Yellow 26, 25300 [*2829-42-7*]
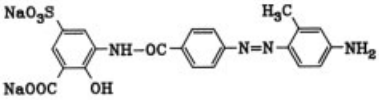	C.I. Direct Yellow 33, 29020 [*6420-28-6*]
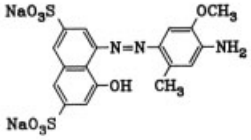	C.I. Direct Red 79, 29065 [*1937-34-4*]

acid) to be condensed with cyanuric chloride and the reaction product subsequently attached to the diazo component, or aminoazo dyes are reacted directly with cyanuric chloride. In the relevant commercial azo dyes, the third chlorine atom in the cyanuric chloride is usually reacted with aniline or ammonia and less frequently left unchanged.

Of interest, here, are dyes that yield green shades through combination of a blue and yellow component.

Example: *C. I. Direct Green 26,* 34045 (**120**) [*6388-26-7*].

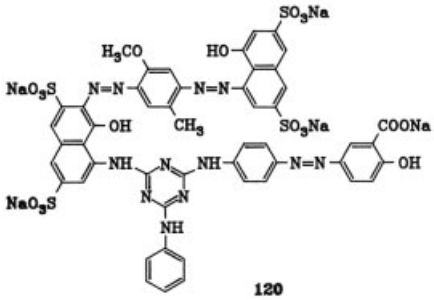

120

The blue component (H acid → cresidine → H acid) is linked with the triazine ring by means of the amino group of the H acid; after reduction of the nitro group, the yellow component (*p*-nitroaniline → salicylic acid) replaces the second chlorine atom of cyanuric chloride and aniline the third chlorine atom. The dye consti-

tutes a very pure, clear, bluish green with very good fastness properties.

4.1.8. Copper Complexes of Substantive Azo Dyes

Under certain structural conditions, azo dyes are capable of forming metal complexes (see Section 2.3.4). A process for the complexing of *o,o′*-dihydroxyazo dyes with copper by reaction of these dyes with copper sulfate in a weakly acid solution was described for the first time in 1915 [39].

The introduction of copper into conventional substantive dyes often increases their light and wetfastness considerably, since complexing blocks the hydrophilic groups. Copper complexes of *o,o′*-disubstituted azo dyes are chiefly of interest for direct dyes. Substituents especially capable of complexing are hydroxyl, methoxy, carboxy, and carbomethoxy groups. BASF developed the oxidative copper treatment process in which on the basis of *o*-hydroxyazo dyes, copper complexes of the corresponding *o,o′*-dihydroxyazo dyes are obtained [40].

The possibility of first converting suitable azo compounds into copper complexes, and subsequently using them as coupling components for the synthesis of azo direct dyes, was also described.

As in the case of the metal-free direct dyes, a large diversity of structures is encountered; here, too, there are direct dyes with pyrazolones, substituted 4,4′-diphenylamines, or the 1,3,5-triazine ring, urea and stilbene derivatives, and condensation dyes.

For each dye it is necessary to establish the most favorable conditions under which copper-containing products with improved properties are obtained. Because none of the known complexing processes provides optimum results in all instances, the production of certain dyes or complexing variations is protected by a large number of patents.

Especially for ecological reasons, however, the importance of copper complex dyes is decreasing. This applies in particular to the copper aftertreatment.

A demethylating copper treatment of dyes synthesized with *o*-dianisidine results in light-fast blue grades, such as *C. I. Direct Blue 98,* 23155 [*6656-03-7*], a copper complex of **121**;

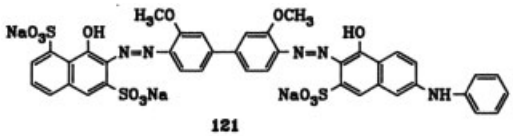

121

and *C. I. Direct Blue 84,* 23610 [*13569-92-1*], a copper complex of **122**.

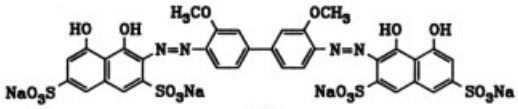

122

The following additional examples of substantive copper complex dyes can be mentioned: *C. I. Direct Violet 47,* 25410 [*13011-70-6*], a copper complex of **123**;

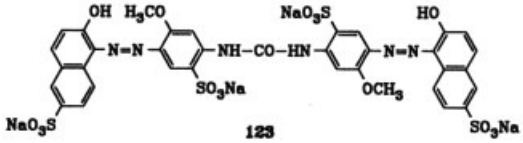

123

and *C. I. Direct Green 23,* 31985, a copper complex of **124**.

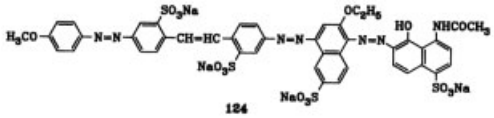

124

4.2. Direct Dyes with Aftertreatment

For supplementary literature see [41] and [42, vol. 2, p. 46]

In spite of the reactive dyes with their outstanding wetfastness, the direct dyes continue to hold a large share of the market for inexpensive cellulose and paper dyes. Apart from the new developments to replace the old established benzidine dyes, considerable efforts are being made to improve the fastness properties of direct dyes by aftertreatment. The various aftertreatment methods are described below in the order of their present importance.

4.2.1. Aftertreatment with Cationic Auxiliaries

Aftertreatment of substantive dyeings with organic, cationic substances has lately begun to gain increasingly in importance. Improvements are obtained in particular in wetfastness properties, especially fastness to water, washing, and wet pressing, as well as fastness to perspiration and cross-dyeing.

The cationic compounds precipitate the (anionic) dyes from their aqueous solutions and on the fiber form higher molecular mass, scarcely soluble, saltlike compounds with the dyes. Removal of the latter from the fiber is thus made more difficult. A number of companies have compiled their own ranges of this type of aftertreatment agent.

Cationic textile auxiliaries can be divided into two groups: quaternary ammonium compounds and cationic formaldehyde condensation resins.

Quaternary Ammonium Compounds. These contain aliphatic and/or cycloaliphatic radicals (R^1 to R^4 in Eq. 12), of which at least one represents a long-chain alkyl radical (with more than five C atoms). The compounds react with the dyes in accordance with the following pattern:

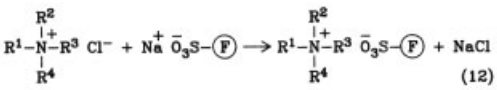

(12)

(F) = dye radical

Use is made, for example, of cetyldimethylbenzylammonium chloride and pyridinium compounds of the type

X^- R = long-chain alkyl radical
X = Cl, SO_4H, etc.

In addition, the quaternary ammonium compounds mostly possess strong affinity for the cellulosic fiber, whereby the insoluble reaction product is anchored even more firmly with the dye on the fiber. This high affinity can, however, also result in nonuniform improvements in fastness properties. This behavior also prevents the use of circulating liquor dyeing machines, because even with addition of aftertreatment agent in portions (e.g., in the case of wound packages) no completely uniform aftertreatment can be guaranteed.

The action of quaternary ammonium compounds on the dyed fiber in many cases produces a change of shade.

A further disadvantage is the possible elimination of aliphatic amines, e.g., in an alkaline medium in the presence of anionic substances or at excessive drying temperatures. This can cause an unpleasant (fishy) odor to arise. In such cases, an acid aftertreatment is necessary.

Cationic Formaldehyde Condensation Resins. An appreciable improvement in wetfastness properties is obtained by aftertreating the dyeings with polymeric condensation resins containing strongly basic groups. This leads to the formation of insoluble salts from resin cation and dye anion.

The most important resins of this type are obtained from dicyanodiamide (**125**) by condensation with formaldehyde and hydrochloric acid:

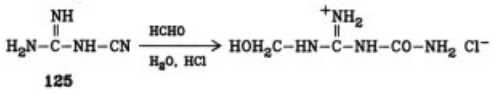

125

Suitable blending of ammonium salts of fairly strongly basic amines (e.g., melamine or guanidine) and acid condensation with formaldehyde enables favorable conditions to be obtained with regard to solubility and cross-linking of the cationic condensation resins.

The water-soluble resins are inexpensive, and compared with quaternary ammonium compounds relatively small quantities often suffice to achieve a marked improvement in wet fastness and fastness to perspiration. A disadvantage is a certain influence on light fastness. This may be lowered by one to two steps on the eight-step blue scale, when these resins are used on materials dyed by substantive dyes. Formaldehyde condensation resins are therefore used especially for articles in which wet fastness and fastness to perspiration are important, but light fastness is less crucial, e.g., lining fabrics.

4.2.2. Aftertreatment with Formaldehyde

The wetfastness properties of substantive azo dyes containing free amino or hydroxyl groups as terminal groups (final component of coupling, e.g., resorcinol or *m*-phenylenediamine) can be improved by aftertreating the dyed fiber material with aqueous formaldehyde solution. The conditions for the presence of free amino groups can also be satisfied by subsequent reduction of nitro groups or hydrolysis of acylamino groups. During reaction with formaldehyde, methylene bridges are formed between two dye molecules; i.e., an enlargement of the molecule occurs and hence an improvement in wetfastness, whereas light fastness is not influenced.

Because of a lack of storage stability in dyeings treated by this method, the formaldehyde aftertreatment is mainly confined to dark shades (brown and black).

The first patent on this aftertreatment method was granted to Geigy as early as 1899 [43]. Although the basic, simple process has remained virtually unchanged, its importance has lately declined sharply for ecological reasons.

4.2.3. Diazotization Dyes

Direct dyes containing one or more diazotizable amino groups in the dye molecule permit further diazotization on the fiber and subsequent coupling with a "developer." β-Naphthol can be used as a developer for orange, red, brown, blue, and black shades, 1,3-phenylenediamine and 2,4-diaminotoluene for brown, gray, and black shades.

The shade of the dye applied to the material usually changes considerably as a result of this aftertreatment. wetfastness and fastness to perspiration are greatly improved, light fastness remains unchanged. The amino group needed for diazotization is introduced by final coupling with an amine, an aminonaphtholsulfonic acid, or an amino-acylaminonaphtholsulfonic acid or by reduction of a nitro or hydrolysis of an acylamino group. "Development" results in enlargement of the dye molecule without further solubilizing groups being added.

The time-consuming process necessitates careful handling. Exposure to sunlight or partial drying of the diazotized dyeing should thus be avoided.

Diazotization dyes now enjoy only minor importance.

Example: *C. I. Direct Red 145,* 17805 (**126**) [*6771-94-4*], yields a clear yellowish red when developed with β-naphthol.

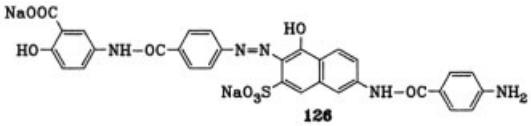

126

4.2.4. Aftertreatment with Metal Salts

This method is dealt with briefly, mainly on historical grounds, because for reasons connected with wastewater pollution and because of certain performance drawbacks it is hardly used any longer.

Metal complexing (see Section 4.1.8) cannot be carried out until after the fiber has been dyed. This aftertreatment, mainly with copper salts (sulfate, acetate), takes place under mild conditions, so that, for example, no metal complexing with an *o*-methoxyazo group occurs. Suitable complexing systems, on the other hand, are *o,o′*-dihydroxy-, *o*-hydroxy-*o′*-carboxy-, *o*-hydroxy-*o′*-carbomethoxy-, and *o*-carboxy-*o′*-aminoazo groups.

4,4′-Diaminodiphenyl-3,3′-dicarboxylic acid (**127**), 4,4′-diaminodiphenyl-3,3′-bisoxyacetic acid (**128**), and 4,4′-diaminodiphenylurea-3,3′-dicarboxylic acid (**129**) have proved successful as diamino components for the synthesis of disazo copper-aftertreated dyes.

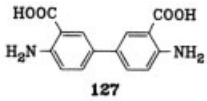

127

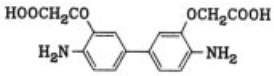

128

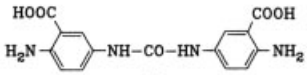

129

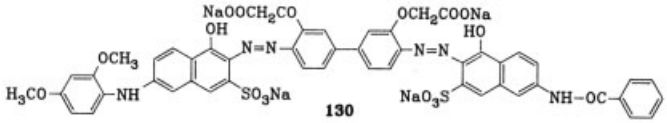

130

Metal complexing results in blocking of solubilizing groups; aggregation increases, the dye becomes less soluble, and an improvement in wet fastness occurs. Although the light fastness is generally also improved by this aftertreatment, fastness of the dyeings to perspiration is frequently reduced. The shades generally become duller and flatter. Catalytic action of traces of heavy-metal ions that are present can cause fiber tendering to occur in the presence of detergents containing oxidizing agents (peroxides).

An example of a direct dye suitable for copper aftertreatment is *C. I. Direct Blue 167,* 24560 (**130**), see top of the page.

5. Cationic Azo Dyes

Cationic azo dyes carry a positive charge in the colored portion of the molecule. The salt-forming counterion is, in most cases, the colorless anion of a low molecular mass inorganic or organic acid. The positive charge is generally carried by nitrogen, but there are also dyes in which this function is taken on by sulfur or phosphorus (see Section 5.5).

The positive charge may be either localized at an ammonium group (see, e.g., **131**)

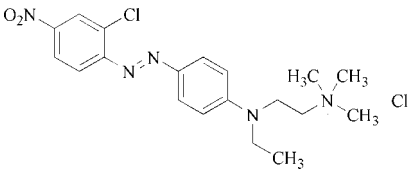

131

or delocalized across the dye cation. The transition between the two dye classes is continuous. At one extreme, this leads to the dark blue compound **132**, which is no longer considered an azo dye but classified as a diazacyanine dye (→ Methine Dyes and Pigments).

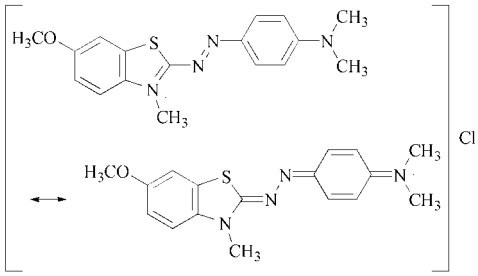

132 At an intermediate position, compound **133** has its positive charge much less delocalized because of the high energy of the quinoid resonance extreme **133b**.

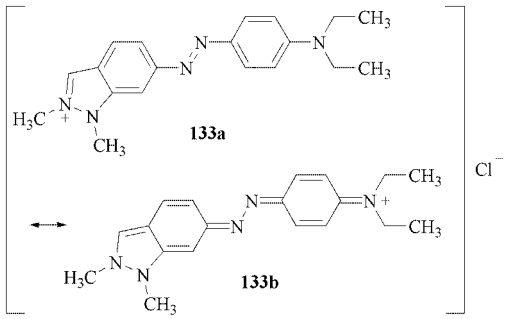

Finally, in compound **134** the structure does not permit a charge delocalization from the benzimidazolium ion to the back ring azo dye substituent.

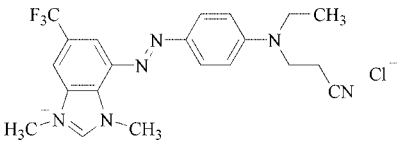

134

Both **133** and **134** confer a similar orange-red color to polyacrylonitrile and are classified as cationic azo dyes.

History. Basic dyes of the azo class are among the earliest known synthetic dyes. They were used originally for dyeing cotton mordanted with tannin and potassium antimonyl tar-

trate and wool from neutral solution. They continue to occupy a place of minor importance for dyeing leather, paper, plastics, and waxes, and as constituents of graphic arts colors.

The preparation of the first cationic azo dye, Vesuvin, was described by C. Martius in 1863. It is obtained by coupling diazotized *m*-phenylenediamine to an excess of the same amine. An analogous dye from toluylenediamine was reported by P. Gries in 1878. Chrysoidines, coupling products of aniline or toluidines to *m*-phenylenediamine or toluylenediamine, were reported by H. Caro in 1875 and O. N. Witt in 1876. These were followed in 1886 by cationic safranine azo dyes, obtained by coupling diazotized diethylphenosafranine to phenols or aromatic amines (→ Azine Dyes), and by dyes produced by coupling diazotized aminophenyltrimethylammonium chloride to phenols and aromatic amines. Both groups formed the basis of the Janus dye product line, which gained great importance for a time in dyeing weighted silk.

When polyacrylonitrile fibers appeared on the market, an intense research effort in the area of cationic azo dyes was stimulated worldwide at all the leading dye plants. These materials now occupy a place of importance in all significant product lines.

Azo dyes with several cationic charges, which are substantive dyes for cellulose, are increasingly being used for coloring bleached sulfite cellulose.

The cationic charge may be introduced into the dye molecule via either the diazo component or the coupling component.

5.1. Cationic Charge at the Coupling Component

Amino groups, amidine residues, trialkylammonium, or cyclic ammonium groups may serve as carriers of the cationic charge in the coupling components.

5.1.1. Polyamines as Coupling Components

Chrysoidines are obtained by coupling diazonium salts of aniline, toluidines, or mixtures thereof to 1,3-phenylenediamine, 2,4- or 2,6-toluylenediamine, or mixtures of these diamines. These dyes confer muted yellow to orange shades to paper, leather, and polyacrylonitrile fibers. When they are mixed with malachite green and fuchsin, medium-fast black shades are obtained on polyacrylonitriles. A mixture of chrysoidine with Crystal Violet or Victoria Pure Blue is used to adjust the color of nigrosine hectograph inks [44].

Examples are *C. I. Basic Orange 1,* 11320 (**135**) [*4438-16-8*] the azo dye from diazotized aniline and 2,4-diaminotoluene,

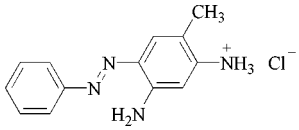

135

and *C. I. Basic Orange 2,* 11270 (**136**) [*532-82-1*] the dye from aniline and *m*-phenylenediamine.

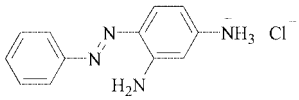

136

The dye *Astrachrysoidine RR* is a mixture of several compounds.

A mixture of 13 g aniline, 12.5 g *o*-toluidine, and 3.0 g *p*-toluidine is diazotized in aqueous hydrochloric acid solution at 0 °C using 66.5 mL of a 30 g/100 mL sodium nitrite solution. Excess nitrite is destroyed prior to coupling using sulfamic acid. The diazotized solution is added in a thin stream and at 0 °C to a solution of 16.5 g of *m*-phenylenediamine and 18.5 g of 2,4-diaminotoluene in 1000 mL of water. The coupling solution is then adjusted to pH 3 – 3.5 by addition of dilute aqueous sodium acetate. After completed coupling, the pH of the reaction mixture is adjusted with hydrochloric acid using congo red indicator. The precipitated dye is collected by filtration and dried to yield 75 g of a dark orange powder.

Chrysoidine dye salts of dodecylbenzenesulfonic acid are soluble in glycols and glycol ethers and are used in the production of inks, printing inks, and varnishes [45].

Mono- and disazodyes from diazotized anilines and 2,4,6-triaminotoluene dye wool and cotton [46].

Vesuvin and *Bismarck Brown* products are obtained when *m*-phenylenediamine, 2,4- and 2,6-diaminotoluene, or mixtures thereof are treated with nitrite in acidic solution, or when a mixture of the amines and nitrite is acidified. The final products consist of mixtures of monoazo, disazo, and polyazo dyes, with the disazo species probably predominating in commercial products. In the presence of sodium chloride, the dye is salted out immediately after its formation and thereby the formation of polyazo dyes is prevented.

Astravesuvin R (**137**), *C. I. Basic Brown 4*, 21010, serves as an example:

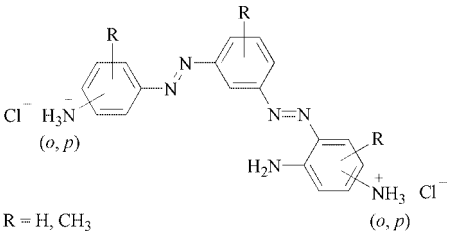

137

A mixture of 2,4- and 2,6-diaminotoluene (28.5 g, molar ratio 2 : 3) and 25 g of *m*-phenylenediamine is dissolved in 1200 mL water at room temperature and mixed with 67 mL of 30 g/100 mL aqueous sodium nitrite. After addition of 450 g of sodium chloride and 800 g of ice, which results in a temperature of − 10 to − 12 °C, 86 mL of concentrated hydrochloric acid is added rapidly. Finally, the pH is adjusted to 3 – 3.5 by the addition of dilute sodium acetate solution. After several hours, the coupling solution is adjusted to pH 1 and the precipitated dye is collected by vacuum filtration and dried to yield 90 g of a yellow-brown powder.

Liquid Vesuvin dye solutions may be obtained by diazotization and coupling of aromatic diamines [47–49] or, if desired, also in mixture with aromatic monoamines [50] in carboxylic acid solutions. The dye mixtures can be alkylated with alkylene oxides [51].

m-Phenylenediamine (105 g) is added to a mixture of 360 g glacial acetic acid, 80 g water, and 60 g ethylene glycol followed by stirring and warming to dissolve. The temperature is lowered to 8 – 10 °C and, at this temperature, 47 g sodium nitrite is added in portions over a period of 1.5 h. After addition is completed, the mixture is stirred for 1/2 h, 10 g *m*-phenylenediamine is added, and stirring is continued for another 3 h at 8 – 13 °C. Finally, the reaction is warmed to 35 – 40 °C for 1 h followed by chilling of the dye solution.

Vesuvines are used chiefly for dyeing paper containing wood pulp.

5.1.2. Heterocycles as Coupling Components

If heterocyclic polyamines, e.g., 2,4-diamino-6-hydroxypyrimidine, are used as coupling components with aromatic diazonium compounds, dyes are obtained that color polyacrylonitrile in lightfast yellow shades and that are characterized by excellent leveling properties [52]. An example is **138** [*6979-64-2*].

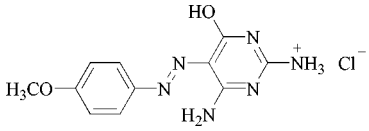

138

Coupling Components with Amidine Moieties. Well-leveling azo dyes are also obtained when pyrazolones substituted in the 1-position by an amidine residue are used as coupling components. The diazo components are predominantly substituted by methyl and methoxy groups [53]. An example is **139** [*68936-08-3*].

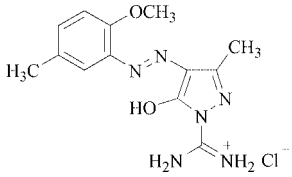

139

By reaction of these dyes with formaldehyde and acetone, condensation products suitable for wetfast dyeing of cotton are obtained [54].

5.1.3. Coupling Components with Alkylammonium Groups

Numerous cationic azo dyes contain a dialkylaminoalkyl or a trialkylammonium group as the charge carrier in the coupling component.

5.1.3.1. Coupling Components with Dialkylaminoalkyl Groups

Dyes bearing one or more dialkylaminoalkyl substituents have found favor for coloring paper because they behave as substantive dyes which, when properly substituted, result in a practically colorless waste liquor. A suitable starting material for the preparation of such dyes is the reaction product of 2 mol dialkylaminopropylamine with 1 mol cyanuric chloride. This may be converted to an acetoacetarylide, which,

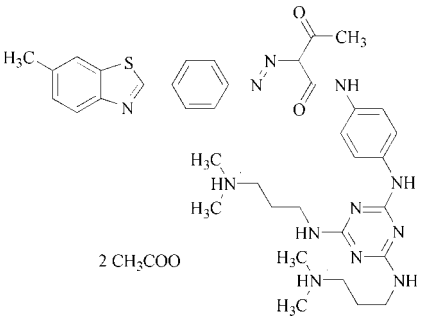

140

when coupled with diazotized 2-(4′-aminophenyl)-5-methylbenzothiazole, yields the yellow dye **140** [*91458-38-7*] with greenish cast suitable for coloring paper [55].

By reaction of a similar triazine component with J acid (1-hydroxy-6-aminonaphthalene-3-sulfonic acid), followed by coupling with diazotized 4-aminoazobenzene, the cationic substantive azo dye **141** [*71032-95-6*] is obtained, which colors paper in red hues.

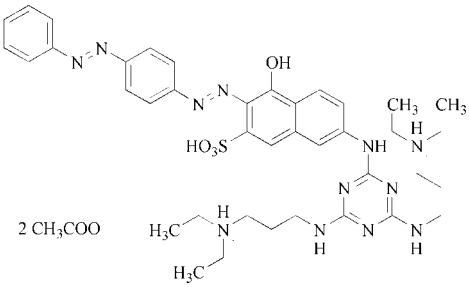

141

Copper complexes of these materials or of the analogous γ acid (1-hydroxy-7-aminonaphthalene-3-sulfonic acid) dyes confer gray-violet [56] or blue hues [57], [58] to paper, respectively, or dye particles of glass in navy blue tones [59]. A *commercial product* of this series is Cartasol Red K-2B (Clariant).

These kinds of triazine moieties may also be part of pyrazolone and pyridone coupling components [60], [61].

The dialkylaminoalkyl residue may also be connected directly to the phenyl residue of the acetoacetanilide. By coupling with diazotized 2-(4-aminophenyl)benzimidazole, a yellow dye suitable for paper is formed [62].

The dialkylaminoalkyl group may be connected to the N-1 atom of 3-methylpyrazolone [63], to the amide nitrogen atom of *N*-phenylpyrazolone-3-carbonamide [64], to the N-2 atom of 3-cyano-2,4,6-trisaminopyridine [65], or to the amino group of 2-amino-4,6-dihydroxypyridine [66].

3- or 6-hydroxy-2-naphthoic acid amides bearing dialkylaminoalkyl residues on the nitrogen atom lead to orange dyes for polyacrylonitrile or paper when coupled with diazotized aniline [67], [68], or to red dyes, soluble in glycol ethers and useful for ink-jet printing, when coupled with diazotized 3-amino-4-methoxybenzenesulfonic acid-*N,N*-diethylamide [69], or to substantive red dyes when the diazo component is a 2-(4′-aminophenyl)benzotriazole [70]. Dialkylaminoalkyl amides of J acid have been used to prepare red dyes for polymeric materials, especially paper, with 4-acetylaminoaniline as diazo component [71], [72].

5.1.3.2. Aromatically Linked Trialkylammonium Groups

Exhaustive alkylation of aminonaphthols, such as 2-amino-7-naphthol, and coupling with diazotized aromatic amines generates yellow, orange, red, and brown cationic azo dyes [73], [74]. Compound **142** dyes human hair and polyacrylonitrile in brown shades [75].

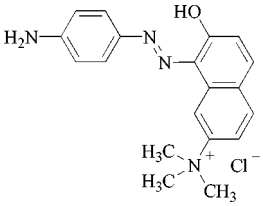

142

Exhaustively alkylated 4-aminodiphenylamine is also suitable as a coupling component. Dyes (e.g., **143**) [*41025-69-8*] obtained by reaction with diazotized nitroanilines confer fast yellow, orange, and red shades to polyacrylonitrile [76].

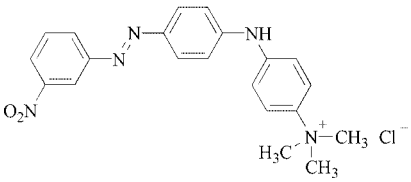

143

By exhaustive alkylation of 3-aminophenylpyrazolone compounds, a trimethylammonium moiety may be introduced into this coupling component. The disazo dye **144** [*86565-98-2*] colors paper in orange shades [78] (see bottom of page).

The trialkylammoniumphenyl group may also be part of acetoacetic acid arylides [77] or be connected to the coupling component via an ether link (see **145**) [*84041-74-7*] [79].

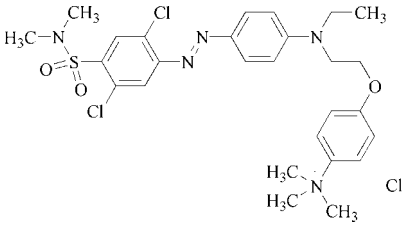

145

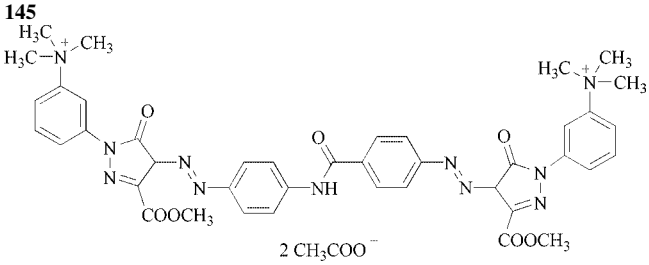

144

5.1.3.3. Trialkylammoniumalkyl-Substituted Anilines

Dialkylanilines that carry a cationic substituent at one of the alkyl groups are among the most important coupling components of cationic azo dyes.

By reaction of *N*-ethyl-*N*-(2-chloroethyl)aniline or of *N*-ethyl-*N*-(2-chloroethyl)-*m*-toluidine with trimethylamine, ammonium salts are obtained that, upon coupling with diazotized aromatic amines, yield a large number of valuable dyes for coloring polyacrylonitrile.

C. I. Basic Red 18, 11 085 (**146**) [*25198-22-5*] is derived from the diazo component 2-chloro-4-nitroaniline [80].

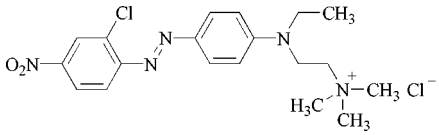

146

It is a component of all major product lines used for dyeing polyacrylonitrile.

Commercial products are Astrazon Red GTL (DyStar), Diacryl Red GTL-N (Mitsubishi), Sumiacryl Red F-GTL (Sumitomo), and Taiacryl Red GTL (T&T Ind.).

The corresponding dye from 2,6-dichloro-4-nitroaniline as the diazo component is also of commercial importance.

Other suitable diazo components are 4-nitroaniline (orange), 2,6-dichloro-4-nitroaniline (yellow-brown) [81], 2-cyano-4-nitroaniline (red) [82], 2,6-dichloro-4-(*N,N*-dimethylsulfamoyl)aniline (orange), 4-aminobenzophenone (yellow), 2,4,6-trichloroaniline (yellow), 2,4,6-tribromoaniline

(yellow), 2,6-dibromo-4-nitroaniline (yellow brown) [83], and 2-cyano-5-chloroaniline (orange).

Heterocyclic diazo components include 2-aminothiazole substituted with ester groups [84], and the coupling product of diazotized 4-nitroaniline with 2-amino-3-cyanothiophene [85]. An especially favored starting material is 3-phenyl-5-amino-1,2,4-thiadiazole, which is also used in the production of disperse dyes. It yields very bright dyes for coloring polyacrylonitrile [80]. An example is **147** [*85283-77-8*].

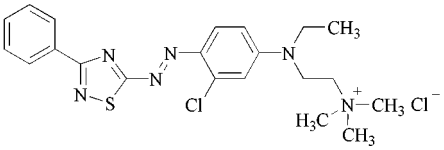

147

A commercial product is Astrazon Red BBL (DyStar).

By coupling aminobenzothiazoles to the quaternary ammonium bases red cationic dyes (e.g., **148**) are obtained [86].

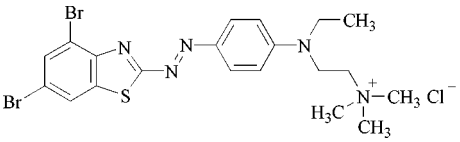

148

This is to be contrasted with the blue diazacyanines, which derive from the quaternized benzothiazole residue.

Dye affinity to polyacrylonitrile and, thereby, the dyeing rate may be enhanced by introducing an aralkyl residue [87] or an aryloxyalkyl residue [88] in place of an alkyl group at the trialkylammonium group.

The solubility of these dyes is significantly enhanced by introduction of alkenyl [89], hydroxypropyl [90], or polyetheralkyl [91] moieties into the trialkylammonium group. The trialkylammonium group may also be a component of a heterocyclic ring [92]. When the trialkylammonium residue contains an alkyl chain of more than 11 carbon atoms, the dyes become soluble in aliphatic and alicyclic hydrocarbons and can be used for the production of printing inks [93]. Also *C. I. Basic Black 8* is recommended as a dye for ink-jet printing [94]. The quaternization to the trialkylammonium residue may also be carried out using dialkyl phosphites or phosphonates [95].

With aromatic diamines as the diazo components, such as 4,4′-diaminodiphenylsulfone, basic disazo dyes are obtained which are suitable for dyeing acid-modified polyamide materials [96].

The diamine of dye **144** leads to cationic disazo dyes that color paper in brilliant orange hues [97].

5.1.3.4. Other Trialkylammoniumalkyl-Substituted Coupling Components

The coupling component 2,6-dihydroxypyridine, already an acknowledged component of high-intensity chromophores in disperse dyes, has also been applied successfully to the synthesis of cationic azo dyes. The trialkylammoniumalkyl group may be linked either to the pyridine nitrogen [98] (**149**) [*71873-54-6*] or to C-3 of the pyridine ring [99] (**150**).

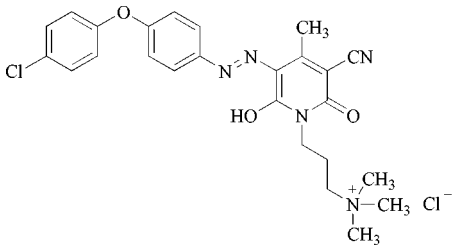

149

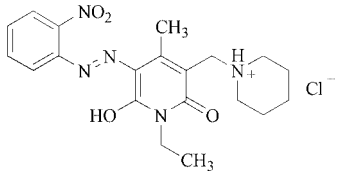

150

Dyes with cationically substituted derivatives of 2-hydroxy-6-aminopyridine [100], of 2,6-diaminopyridine [101], and of 5-hydroxy-3-methyl-1-phenylpyrazolone [102] as the coupling components have also been described. Coupling components based on 4-hydroxynaphthalimide, which carry the trialkylammoniumalkyl substituent at the imide nitro-

gen, have been used for the synthesis of red cationic azo dyes [103].

Aromatic coupling components may have the trialkylammoniumalkyl linked via a carboxamide group [104], e.g., **151**.

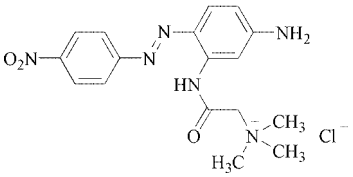

151

Also described as linking groups are ether [105], thioether [106], carbonyl [107], carboxylic ester [108], carbonamide [109–112] and sulfonamide [113] moieties.

Monofluortriazine reactive dyes, in which a trialkylammoniumalkyl group is bound to the heterocycle by an amino group are recommended for blends of polyacrylonitrile with wool, synthetic polyamides or cellulosics [114–116].

5.1.4. Coupling Components with Dialkylhydrazinium Groups

A number of valuable cationic azo dyes contain the dialkylhydrazinium group as the charged moiety of the coupling component. The aromatic amines used as the diazo components are the same as those described for the dyes mentioned in Section 5.1.3.3. The red dye **152** [*4531-45-7*] is formed with 2-chloro-4-nitroaniline:

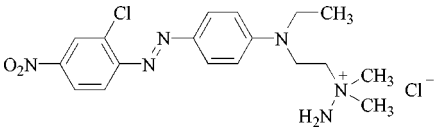

152

A yellow dye is obtained with 1-amino-2,6-dichloro-4-(*N,N*-dimethylsulfamoyl)benzene, an orange dye with 1-amino-2,5-dichloro-4-(*N,N*-dimethylsulfamoyl)benzene, a yellow brown dye with 1-amino-2,6-dichloro-4-nitrobenzene, a ruby dye with 1-amino-2-cyano-4-nitrobenzene, and a violet species with 1-amino-2,4-dinitro-6-bromobenzene [117].

Red dyes are obtained with 3-phenyl-5-amino-1,2,4-thiadiazole as the diazo component [118].

5.1.5. Coupling Components with Cyclic Ammonium Groups

By reaction of heteroaromatics, such as pyridine, with intermediates that contain anionically displaceable groups, coupling components are obtained that derive their positive charge from a cyclic ammonium group. Alternately, the heterocycle may be linked to the coupling component by other means for later quaternization.

5.1.5.1. Heterocyclically Linked Cyclic Ammonium Groups

Reaction of chloroacetamide with pyridine, followed by condensation with acetoacetic esters in alcoholic sodium hydroxide, leads to a 2,6-dihydroxy-4-methylpyridine that is substituted in the 3-position by a pyridinium moiety. If this is coupled with diazotized aromatic or heterocyclic amines, yellow to red cationic azo dyes (e.g., **153**) [*92691-25-3*] are obtained [119], [120].

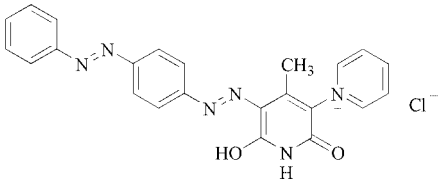

153

With aromatic diamines and triamines as the diazo components, cationic substantive disazo and trisazo dyes (e.g., **154**) [*62073-65-8*] are obtained, which are suitable for bulk dyeing of paper [121], [122], [123] or ink-jet printing [124].

A *commercial product* is Cartasol Yellow K-GL (Clariant).

Also known are 2,6-dihydroxypyridine dyes that carry a quaternized benzothiazole [125], benzimidazole [126], or imidazole [127], [128] moiety in the 3- or 4-position.

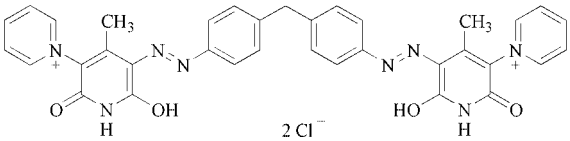

154

5.1.5.2. Aliphatically Linked Cyclic Ammonium Groups

By reaction of *N*-ethyl-*N*-chloroethylaniline, *N,N*-bis(chloroethyl)aniline, or *N*-ethyl-*N*-chloroethyl-*m*-toluidine with heteroaromatics, especially pyridine, coupling components are obtained that carry cyclic ammonium groups.

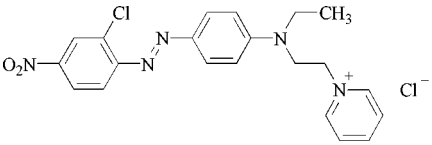

155

The red cationic azo dye **155** [*36986-04-6*], obtained with 2-chloro-4-nitroaniline [129], is represented in numerous cationic dye *product lines*: e.g., Maxilon Red 2GL (Ciba), and Yoracryl Red 2G (YDC).

Introduction of a second pyridiniumalkyl residue (**156**) [*24447-84-5*] leads to a brightening of the hue to orange.

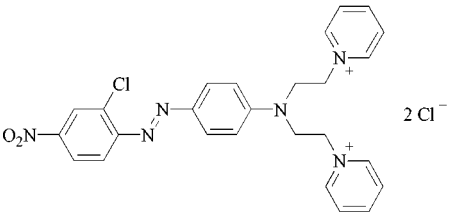

156

Here again, all compounds cited in Sections 5.1.3.3 and 5.1.4 are used as the diazo components. In addition, the following have been described as diazo components: 2,5-dichloro-4-nitroaniline (red) [130], 2-chloro-5-trifluoromethylaniline (yellow) [131], 4-nitro-2-trifluoromethylaniline (red) [132], 2-amino-5-trifluoromethyl-1,3,4-thiadiazole (red) [133], 3-methylmercapto-5-amino-1,2,4-thiadiazole (red) [134], 3-methyl-4-nitro-5-aminoisothiazole (blue violet) [135], 2-amino-6-chlorobenzothiazole (red) [136], 3-amino-5-nitro-7-bromobenzisothiazole (blue with a red cast) [137], and 2-amino-4-chloro-3-cyano-5-formylthiophne (blue) [138].

The solubility of these dyes is enhanced when picolines are used for quaternization in place of pyridine [139].

Concentrated aqueous solutions of these dyes are obtained by dissolving the bicarbonate salts of the dyes in aqueous organic acids, e.g., acetic acid [140].

If monoalkylanilines are reacted with vinylheterocycles, especially vinylpyridine, and if the heterocyclic nitrogen is then quaternized, coupling components are obtained that carry the pyridinium component by way of a carbon linkage [141], e.g., **157**.

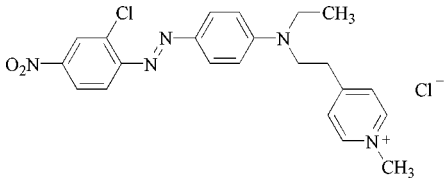

157

The place of the pyridinium residue may also be taken by the 1,2,4-triazolium [142], the benzimidazolium [143], or the imidazolium [144] moieties. An example is the dye **158** [*47660-05-9*].

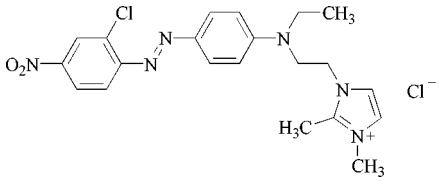

158

The cyclic ammonium residue may also be linked to the coupling component via a sulfonamide [145] or an ester [146] group.

A yellow cationic substantive paper dye is obtained by coupling diazotized 2-(4′-aminophenyl)-5-methylbenzothiazole onto an ace-

toacetic arylide, e.g., *N*-aryl-acetoacetamide, substituted by a pyridinium residue [147], or by coupling tetraazotized 4,4′-diaminostilbene with the coupling component of dye **155** [148].

5.1.6. Coupling Components with Condensed Cyclic Ammonium Residues

Heterocyclic compounds wherein the condensed benzene ring is substituted by a hydroxyl or an amino function may be coupled with diazonium compounds and may also be quaternized, either prior or subsequent to the coupling reaction, to yield cationic azo dyes. 1,2-Dialkyl-6-nitroindazolium salts are reduced to the 6-amino compounds and then coupled with diazonium salts of aromatic amines. These dyes (e.g., **159**) color polyacrylonitrile in bright yellow to orange shades [149].

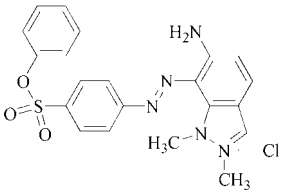

159

6-Amino-1,3-dimethylbenzotriazolium chloride may also be used as the coupling component [150]. By hydrolysis of 6-amino-1,2-dialkylindazolium chloride in dilute sulfuric acid at 180 – 190 °C, a hydroxyl group is introduced into the 6-position. The 6-hydroxy-1,2-dialkylindazolium salt formed is suitable for the generation of mono- [151] (**160**) and disazo dyes [152] (**161**) [*92888-19-2*].

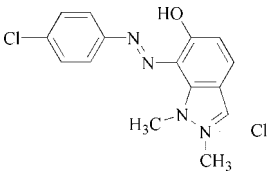

160

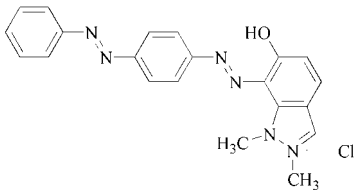

161

In a similar manner 6-amino-1,3-dimethylbenzotriazolium chloride may be converted into the 6-hydroxy compound for use as a coupling component [153].

5.1.7. Cyclic Ammonium Residues at the Azo Group

If azo dyes from aromatic diazonium compounds and 1-phenyl-3-methyl-5-pyrazolone or 1-phenyl-3-methyl-5-aminopyrazole are quaternized at the 2-nitrogen of the pyrazole, cationic azo dyes are obtained. A resonance charge exchange is prevented through lack of a donor function at the aromatic moiety [154], [155]. An example is dye **162**.

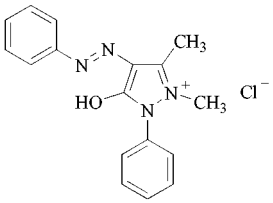

162

Azo dyes from aromatic diazonium compounds and indolizine derivatives may also be converted into cationic azo dyes by quaternization [156].

5.1.8. Coupling Components with two Different Cationic Residues

Mono- and polyazo dyes have been prepared, in which the coupling component carries at least two different basic and/or cationic groups. Of these, mostly yellow to orange dyes are derived from pyridones with a cyclic ammonium group and an amino- or trialkylammoniumalkyl group. To obtain a technically sufficient affinity to cellulose fibers disdiazo components are preferred either with the same coupling component [157–161] (e.g., **163**) or with different ones [162], [163] (e.g., **164**).

A step further is the connection of two suitable molecules with the help of a dihaloalkane [164], [165] (e.g., **165**).

Dyes of this kind can be used to color anodically generated oxide layers on aluminum [166].

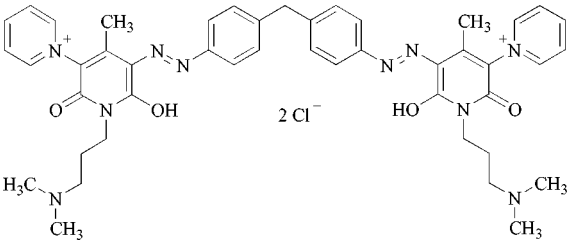

163

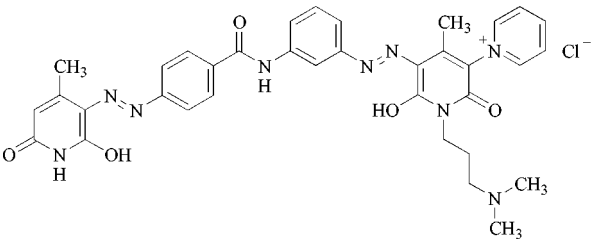

164

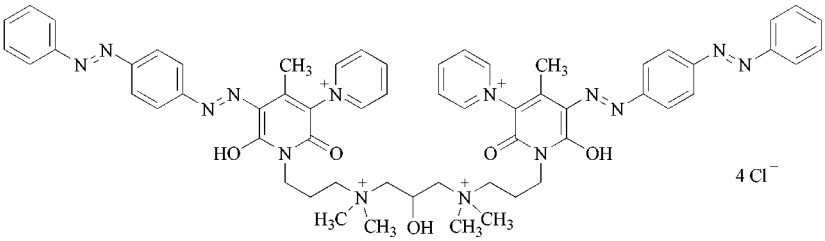

165

A *commercial product* is Cartasol Yellow K-3GL liq.

Orange, red and violet dyes can be obtained by using as coupling components aminonaphtholsulfonic acids that are connected via the N atom to a 1,3,5-triazine with different aminoalkylamino, trialkylammonium-alkylamino or cyclic ammonium substituents in positions 4 and 6 [167] (e.g., **166**).

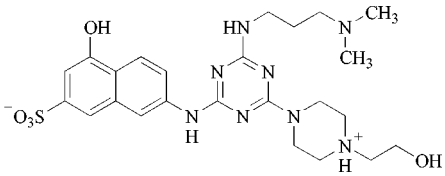

166

For example, **167** dyes paper in a scarlet shade [168].

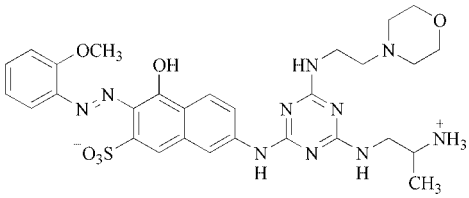

167

Between the N atom of the aminonaphtholsulfonic acid and the triazine ring, spacer groups can be introduced, e.g., aminoaroyl or aminophenylcarbamoyl groups [169], [170].

A different way to synthesize a coupling component with two different basic groups is to react *N*-dibromopropionylaminonaphtholsulfonic acids with dialkylaminoalkylamines to give aziridine derivatives; red dyes are obtained by coupling these with a diazonium salt of *p*-aminoazo-benzene [171] (e.g., **168**).

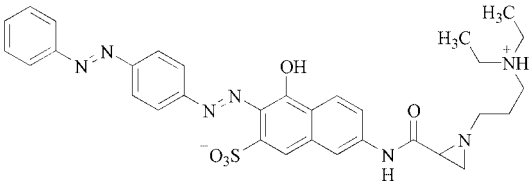

168

5.2. Cationic Charge in the Diazo Component

Diazo components may contain the cationic charge at amino, dialkylamino, trialkylammonium, or cycloammonium groups.

5.2.1. Diazo Components with Aminoalkyl Moieties

Yellow dyes are prepared by aminomethylation of 2-(4′-aminophenyl)-5-methylbenzothiazole followed by diazotization and coupling with acetoacetarylides. Red cationic dyes are obtained with naphthols or hydroxynaphthoic acid arylides. All are suitable for dyeing paper. The bleachability of these dyes is important for the recycling of waste paper [172]. The aminomethylation of aromatic amines is carried out by reaction with formaldehyde and phthalimide, followed by hydrolytic scission of the phthalic acid residue. By diazotization of 4-methoxy-3,5-bis(aminomethyl)aniline trihydrochloride and coupling with 3-hydroxynaphthoic acid arylides, red cationic substantive paper dyes are obtained [173]. Of importance in the manufacture of cationic substantive paper dyes are reaction products of cyanuric chloride with 2 mol dialkylaminoalkylamine and 1 mol 1,3- or 1,4-diaminobenzene.

Yellow paper dyes (e.g., **169**) are obtained by diazotization of these aromatic amines and coupling with acetoacetarylides [174], [175].

By coupling to derivatives of J acid, substantive dyes (e.g., **170**) are obtained wherein the charge on the sulfonic acid group is more than compensated for by the multiple cationic charges. These dye paper in shades of red [176].

Dyes wherein the dialkylaminoalkyl group is connected to the diazo component via the ester group of anthranilic acid [177–180], a carboxamide [181], a sulfonamide [184], [182] or an ether [183] residue have also been described.

5.2.2. Diazo Components with Trialkylammonium Residues

The trialkylammonium group may be linked to the diazo component either via aromatic substitution or by way of an aliphatic residue.

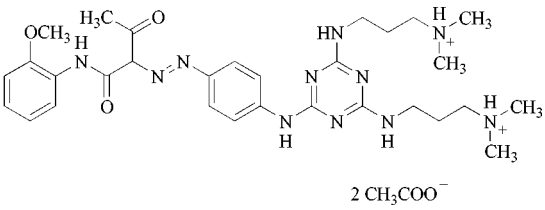

169

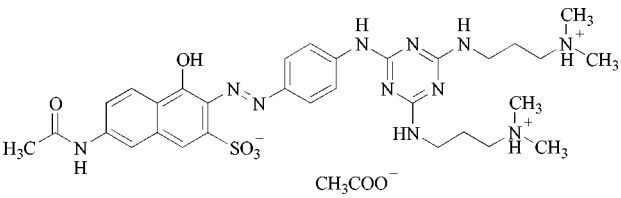

170

5.2.2.1. Aromatically Linked Trialkylammonium Residues

Aminophenyltrimethylammonium chloride was already an important diazo component of the Janus line of dyes. The fastness of these dyes, e.g., **171**, on polyacrylonitrile is enhanced by introduction of halogens. Mono- and dialkylanilines are suitable coupling components [185].

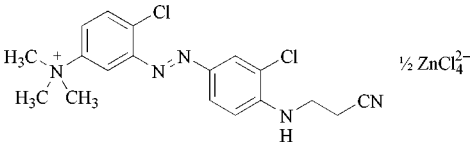

171

Disazo dyes of this series (e.g., **172**) are suitable for dyeing polyacrylonitrile and paper [186].

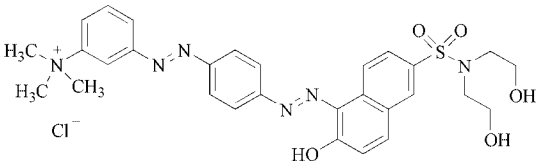

172

A *commercial product* is Anilan Golden Yellow RL (Ciech).

4-Hydroxynaphthalimide as a coupling component yields cationic dyes that confer an orange-red shade to polyacrylonitrile [187].

Heterocyclic coupling components that have been coupled with diazotized aminophenyltrimethylammonium chloride are 5-hydroxy-1-phenylpyrazole-3-carbonamide [188], 1-alkyl-6-hydroxy-2-pyridone [189], 1-amino-3-hydroxy-isoquinoline [190], and 2,4-diamino-6-hydroxypyrimidine [191]. The trialkylammoniumaryl residue may also be connected to the aromatic diazo component via a sulfone or a sulfonamido function [192]. Disazo dyes in this series (e.g., **173**) [*77901-21-4*] may also be generated from monoazo dyes that still contain a primary amino group by dimerization using phosgene [193], [194] or cyanuric chloride [195].

Coupling with β-naphthol results in a yellow dye suitable for polyacrylonitrile [196], coupling with 2,5-dimethoxyaniline gives an orange dye that is recommended for hair coloration [197].

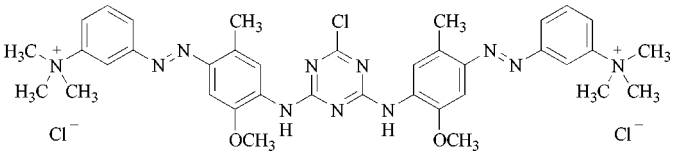

173

5.2.2.2. Aliphatically Linked Trialkylammonium Residues

The Friedel – Crafts acylation of acetanilide with chloroacetyl chloride yields 1-acetamido-4-chloroacetylbenzene. The trimethylammonium group is introduced by reaction with trimethylamine, followed by hydrolysis of the acetamide group. This diazo component is a constituent of numerous yellow, orange, and red cationic azo dyes. Using diethyl-*m*-toluidine as the coupling component, the lightfast red dye (**174**) [*67905-12-8*] is obtained [198].

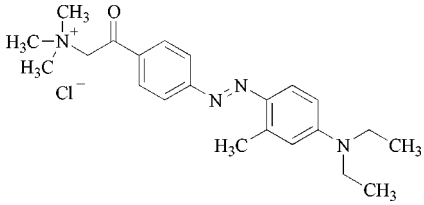

174

Introduction of an unsaturated alkyl residue improves the solubility of the dye [199]. *N*-Methyl-*N*-cyanoethylaniline yields an orange dye. The shade is shifted to yellow with a red cast using *N*-cyanoethyl-2-chloroaniline (**175**) [*72208-25-4*].

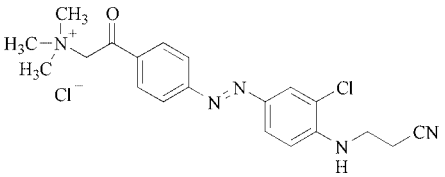

175

Commercial products are Sevron Yellow 3RL (CKC), and Sumiacryl Yellow E-3RD (Sumitomo).

Use of 2-methylindole as the heterocyclic coupling component has led to commercially important dyes (e.g., **176**) [*25784-16-1*] [200].

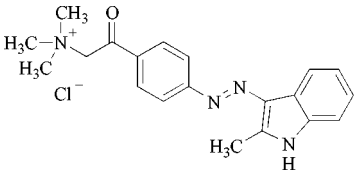

176

However, dyes have also been generated from 1-phenyl-3-methyl-5-aminopyrazole [200], 2,4-dihydroxyquinoline [201], 2,6-dihydroxypyridine [202], 2,6-diaminopyridine [203], 2,4,6-triaminopyridine [204], and 3-phenyl-5-hydroxyisoxazole [205]. The solubility of such dyes may be enhanced by introduction of polyether moieties [206].

Cationic dyes with reactive functional groups, e.g., dichlorotriazine or vinylsulfone, suitable for dyeing wool and silk, have also been generated in this series [207]. Bis-cationic dyes, for instance **177**, are suitable for dyeing mixed-fiber fabrics containing acid-modified and unmodified polyamide, whereas the latter does not take up the dye [208].

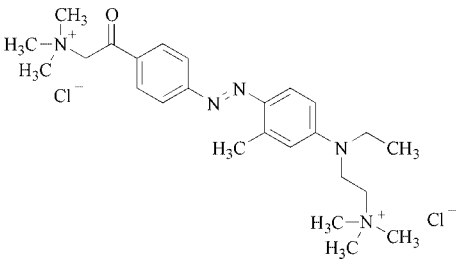

177

A *commercial product* is Sevron Red YCN (CKC).

The trialkylammonium alkyl residue may also be attached directly to the aryl residue of the diazo component [209]. Dyes that derive from aminomethylated 2-aminonaphthalene and contain acetoacetarylides, naphthols, or pyrazolones as the coupling components are suitable for dyeing paper. An example is the yellow monoazo dye **178**.

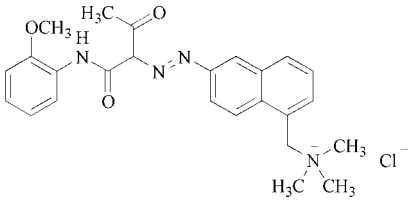

178

Disazo dyes are obtained by using coupling components with a primary amino group and dimerization with phosgene [210]. The diazo components are prepared by exhaustive alkylation of aminomethylated 2-aminonaphthalene-1-sulfonic acid, followed by scission of the sulfonic acid group in the presence of mineral acids [211].

The trialkylammonium residue may also be connected to the diazo component by way of an ether [212], a carbonamide [213], [214], a sulfone [215], or a sulfonamide group [216], [217].

Bis-cationic metal-complex dyes are suitable for dyeing leather. Cationic monoazo dyes wherein a trialkylammonium group is linked via a carboxamide group to the diazo component [218] and disazo dyes are known. The latter are prepared from dialkylamino compounds by quaternization using alkylene dibromides. They are suitable for dyeing paper [219].

Azo dyes with 2-(4′-aminophenyl)-5-methylbenzothiazole as the diazo component and carrying trialkylammoniumalkyl residues are suitable for dyeing paper in the same manner as the dyes described in Section 5.2.1 [172], [220]. Also described as carriers of the cationic charge were dialkylhydrazinium residues linked to the diazo component via a sulfonamide [221] or a carboxamide group [222].

5.2.3. Diazo Components with Cyclic Ammonium Residues

The cyclic ammonium residue may either be linked directly to the aromatic ring or attached to the diazo component via an aliphatic residue.

5.2.3.1. Aromatically Linked Cyclic Ammonium Residues

For the preparation of a diazo component that contains a 1,2,3-triazolium residue in the *para* position, *p*-nitrophenylazide in acetone solution is first reacted with acetylene in an autoclave to give 1-(4-nitrophenyl)-1,2,3-triazole. The nitro group is then reduced, the resultant amine acetylated, and the triazole moiety is quaternized with dimethyl sulfate. After hydrolysis of the acetyl group, the resultant amine may be diazotized and treated with ethylcyanoethylaniline to give dye **179** [*47553-94-6*] [223]. The dye colors polyacrylonitrile to a lightfast orange.

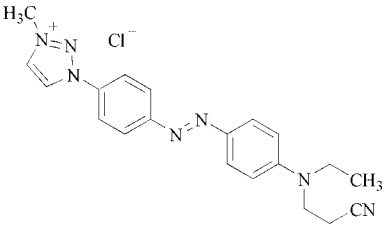

179

The triazole moiety may also be linked in the 2-position with respect to the azo linkage [224]. Among cyclic ammonium residues linked via carbon bonds to the aromatic ring of the diazo component have been cited: 1,3-dialkylbenzimidazolium [225], benzothiazolium [226], methylpyridiniumoxazole [227], and pyridinium [228] moieties, linked via nitrogen, as well as the benzotriazolium moiety [229].

5.2.3.2. Aliphatically Linked Cyclic Ammonium Residues

Upon reaction with pyridine azo dyes that carry a chloroethylsulfonamido group in the diazo component form cationic azo dyes, e.g., **180** [*33869-97-5*], which color polyacrylonitrile in brilliant shades [230].

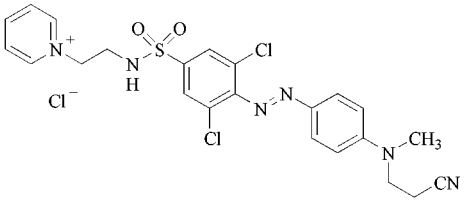

180

The chloroethylsulfonamide may also be reacted with 1,2,4-triazole [231].

Dyes containing the pyridinium group linked to the diazo component via a carboxylic ester function (e.g., **181**) [*32017-47-3*] are obtained by condensation of 4-nitrobenzoyl chloride with chloroethanol, reaction of the ester with pyridine, reduction of the nitro to an amino group, diazotization, and coupling with aromatic amines [232].

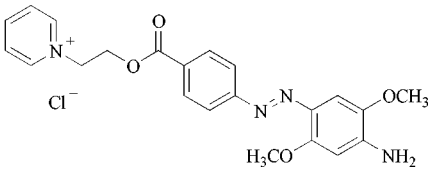

181

This dye imparts an orange shade to polyacrylonitrile. The cyclic ammonium residues may also be linked to the diazo component via a carboxamide [233] or an ether group [234] or directly by means of an alkyl residue [235].

5.2.4. Diazo Components with two Different Cationic Residues

Mono- and disazo dyes have been prepared in which the diazo compound carries at least two different basic and/or cationic groups.

Reaction of 2,4- or 2,5-diaminobenzenesulfonic acids with 2-chloro-1,3,5-triazines that are substituted in the 4- and 6-positions by aminoalkylamines and cyclic ammoniumalkylamines, respectively, leads to diazo components which, when diazotized and coupled with derivatives of aminonaphtholsulfonic acids, give rise to dyes for cotton, leather, and especially paper [236]. An example is **182**, which colors paper in a violet hue.

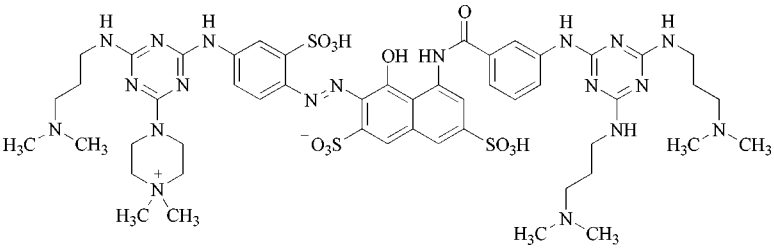

182

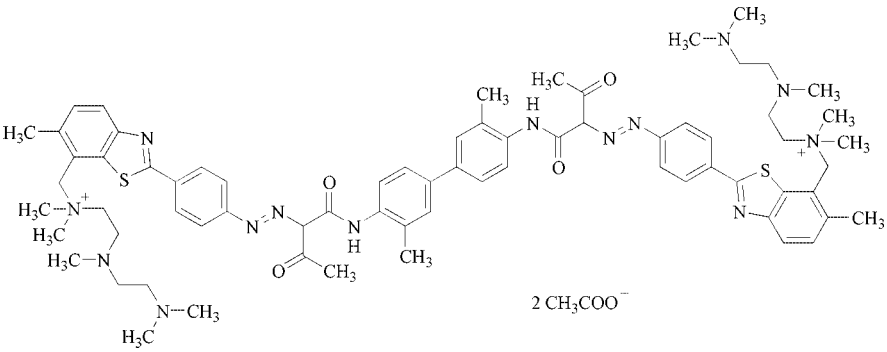

183

Reaction of chloromethyl derivatives of 4′-aminophenyl-2-benzothiazoles with e.g. *N′,N″,N‴*-permethyldiethylene-triamine leads to compounds with alkylaminoalkylammonium residues that can be diazotized and coupled with, e.g., acetoacetarylides [237]. An example is **183**, which colors paper in a greenish shade of yellow.

5.2.5. Diazo Components with Aromatic Condensed Cyclic Ammonium Residues

Polynuclear N heterocycles that carry an amino group in the carbocyclic aromatic ring may be diazotized and then joined to azo dyes using aromatic or heterocyclic coupling components. Quaternization at the heterocyclic nitrogen atom may occur before or after coupling. The azo dye **184**, from 2-methyl-5-aminobenzimidazole and 1-phenyl-3-methyl-5-pyrazolone, used in the form of its hydrochloride, dyes paper and leather in clear yellow shades [238].

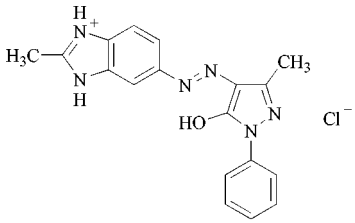

184

By exhaustive alkylation, dyes are obtained that are suitable for polyacrylonitrile [239].

Introduction of halogen yields dyes with especially valuable properties. A suitable diazo component is 4-amino-6-chlorobenzimidazole. The sequence of diazotization, coupling with aromatic amines, and quaternization gives colorants (e.g., **185**) [*36116-31-1*] that confer fast orange and red shades to polyacrylonitrile [240].

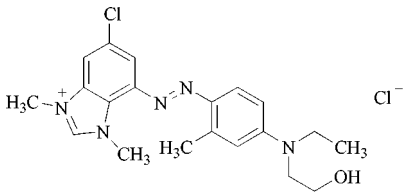

185

Trifluoromethyl-substituted aminobenzimidazole compounds have also been described as diazo components [241].

A bathochromic shift is obtained by using 4-amino-6-chlorobenzotriazole in place of the benzimidazole. The quaternized coupling products with dialkylanilines (e.g., **186**) [*36116-25-3*] dye polyacrylonitrile in red shades with a blue cast [242].

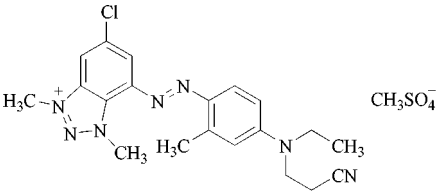

186

Red dyes for polyacrylonitrile are obtained with 6- or 7-aminoindazole as the diazo component [243], (e.g., **187**).

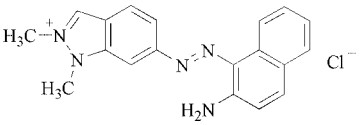

187

Other heterocyclic diazo components that have been used to generate cationic dyes are 4-amino-7-nitrobenzisothiazole [244], 2-methyl-6-aminoquinoline [245], and 5-aminoimidazo[1,2-*a*]pyridine compounds [246].

5.2.6. Cyclic Ammonium Compounds Linked in Meta Position to the Azo Group

Coupling of diazotized N-heterocyclic amino compounds that carry the amino group in the 2- or 4-position to the heterocyclic nitrogen to aromatic or heterocyclic coupling components and subsequent quaternization of the heterocyclic nitrogen leads to diazacyanine dyes (→ Methine Dyes and Pigments).

If the starting materials are heterocyclic amino compounds with the amino group in a meta position, a mesomeric charge exchange is not possible. Therefore, dyes of this type are grouped with the cationic azo dyes.

By coupling of diazotized 3-aminopyridine to 2-naphthol, followed by quaternization, dyes (e.g., **188**) [*41313-61-5*] are obtained that color polyacrylonitrile in fast yellow shades and exhibit excellent leveling characteristics [247].

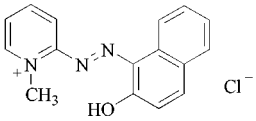

188

This dye is also recommended for dyeing of human hair [248], as well as the dye containing 2-methylindole as coupling component [249], [250].

5.3. Different Cationic Charges in Both the Coupling and the Diazo Component

A further means to introduce different basic and/or cationic residues (cf. Sections 5.1.8 and 5.2.4) into a dye has been developed by combining diazo and coupling components that each carry different basic or cationic groups. They are synthesized according to the procedures that have been described in Sections 5.1 and 5.2.

Dyes have been prepared that color paper in hues ranging from orange [251] (e.g., **189**, see next page)

to red [252–254] (e.g., **190**)

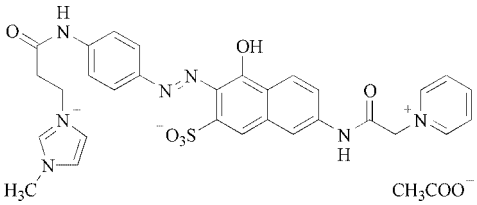

190

to black [255] (e.g., **191**)

The yellow dye **192**

can be used in ink-jet systems to flocculate anionic pigment dispersions, thus improving the bleedfastness [256], [257].

5.4. Introduction of Cationic Substituents into Preformed Azo Dyes

Cationic substituents may also be introduced into preformed azo dyes. In this manner, mixtures of isomers are obtained which may carry

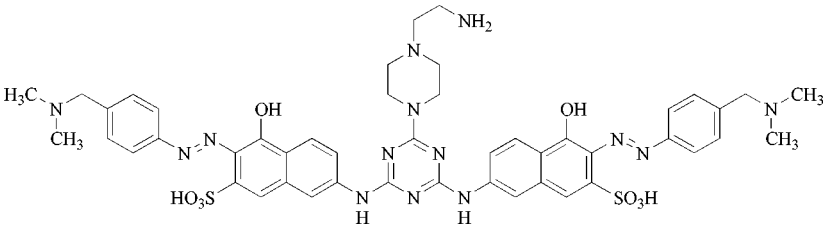

189

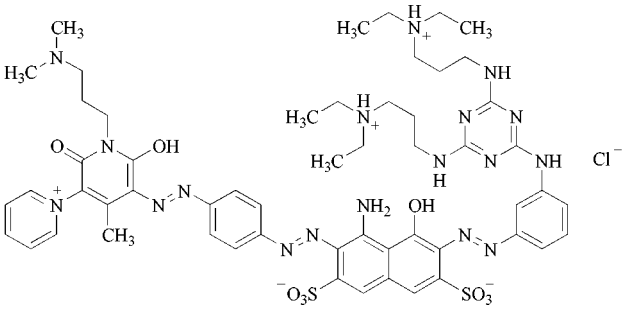

191

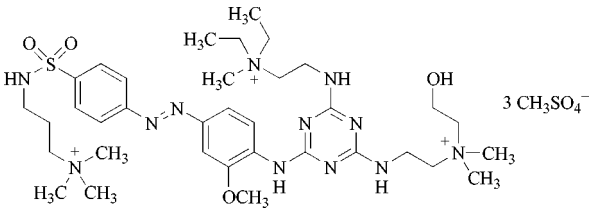

192

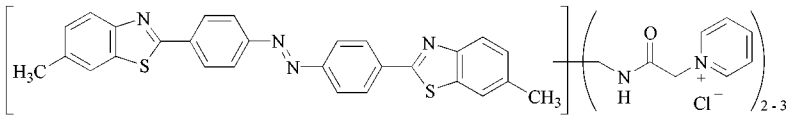

193

one to four cationic groups in various positions. The multiplicity of isomers is a cause for good water solubility. These dyes are predominantly used for coloring paper.

Reaction of chlorosulfonic acid with mono- [258] or disazo dyes [259] yields sulfonyl chlorides, which can then be transformed into cationically substituted sulfonamides by reaction with dialkylaminoalkylamines. The coupling product of diazotized 2-anisidine with 2-hydroxynaphthoic acid arylide upon such treatment dyes paper red and the azo dye from tetraazotized dianisidine and 1-phenyl-3-methyl-5-pyrazolone gives yellowish orange shades.

Numerous cationic azo dyes are prepared by the action of *N*-hydroxymethylchloroacetamide on azo dyes in sulfuric acid medium, followed by displacement of the reactive chloro substituent by pyridine or trialkylamine. Of special significance for dyeing paper are dyes that are prepared by coupling of diazotized 2-(4′-aminophenyl)−5-methylbenzothiazole to acetoacetarylides, pyrazolones, naphthols [260], or barbituric acid derivatives [261], followed by reaction with *N*-hydroxymethylchloroacetamide and pyridine or *N*-methyl-1,3-diaminopropane [262] . The azo dye obtained by oxidative dimerization of 2-(4′-aminophenyl)−5-methylbenzothiazole may also be subjected to this conversion [263]. Dye **193** colors paper yellow.

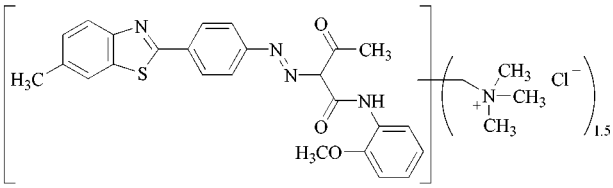

194

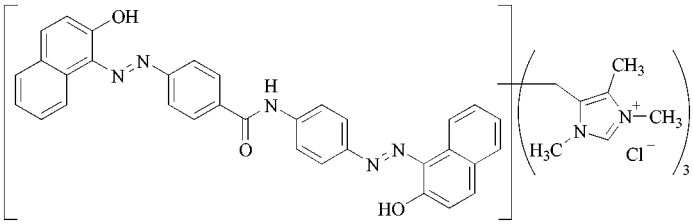

195

Disazo dyes with β-naphthol as the coupling component give red shades [264].

By removal of the chloroacetyl residue aminomethylated dyes are obtained which are used for coloring paper either as such or upon further alkylation [265]. An example is the yellow dye **194**.

Reaction of azo dyes with formaldehyde and 4-methylimidazole, followed by alkylation, produces cationic azo dyes that dye paper with excellent wetfastness [266]. An example is the disazo dye **195**, which confers a brilliant red shade to paper.

A *commercial product* is Cartasol Brilliant Yellow K-5GL (Clariant).

5.5. Cationic Dyes with Sulfur or Phosphorus as Charge-Carrying Atoms

When dyes containing anionically displaceable groups are reacted with thioethers, cationic dyes (e.g., **196**) are obtained in which the sulfonium moiety carries the charge [267]. The sulfonium group may also be linked directly to the aromatic coupling component [268].

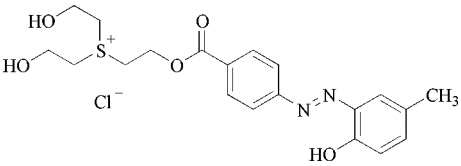

196

By reacting dyes that contain a chloroethyl group with thiourea, the cationic charge may be introduced in the form of an isothiouronium residue (e.g., **197**) [269].

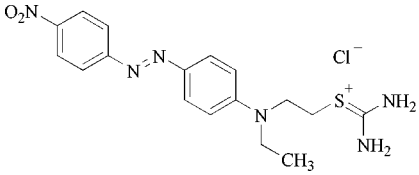

197

To introduce phosphorus as the charge-carrying moiety, dyes that contain haloalkyl groups are reacted with phosphines, such as dimethylphenylphosphine (e.g., **198**) [270].

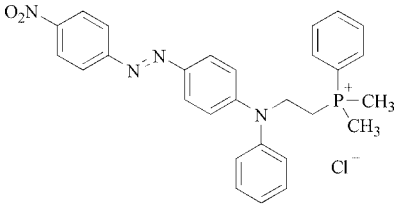

198

5.6. Dyes with Releasable Cationic Groups

Dyes with releasable cationic groups are converted to water-insoluble dyes when heated in the dye bath. They can then enter textile mate-

rials that are amenable to dyeing with disperse dyes. Using this method, for instance, it is possible to dye polyester materials without having first to generate a finely dispersed form of the colorant with the aid of surfactants. An example of such a releasable group is the isothiouronium moiety [271]. Dyes with isothiouronium groups are also suitable for wool [272].

Reaction of dyes containing a primary amino group with dimethylformamide and an inorganic acid chloride, e.g., phosphoryl chloride, permits introduction of the formamidinium group, which is also scissioned off upon heating in the dye bath (e.g., **199**) [273].

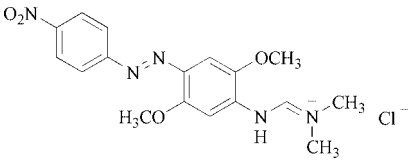

199

A similar reaction occurs with the trialkylhydrazinium moiety obtained by reacting formyl-substituted azo dyes with dialkylhydrazines and subsequent quaternization [274].

6. Developing Dyes

General Survey. There is no standard definition of the term developing dyes. They can include all dyes that are synthesized *on the fiber* or converted there into their final form. In this respect it is possible to distinguish between two main groups:

Developing dyes whose final chromophoric system is synthesized on the fiber with formation of an *insoluble* dye. In the azo dye series, these mainly include the *β-naphthol* and *Naphtol AS dyes*, which are directly synthesized on the fiber; the substrate impregnated with the coupling component is "developed" in a suitable diazonium salt solution. When dyeing or printing the fiber, it is also possible for coupling component and diazo component to be applied simultaneously; not until an aftertreatment is carried out is the coupling process then initiated.

The *diazotization dyes* are also frequently numbered among this group (see Section 4.2.3). As a result of the diazotization of direct dyes with free amino groups and subsequent reaction with suitable coupling components, an azo dye of higher molecular mass with especially good adhesion is synthesized on the fiber.

Developing dyes from another dye class are the *phthalocyanine developing dyes* (e.g., *Phtalogen* dyes (Bayer) → Phthalocyanines).

Developing dyes whose color-imparting structure is synthesized before they are applied to the substrate in the form of *soluble* derivatives; they are converted into an insoluble form by aftertreatment on the fiber.

These include the *polycondensation dyes,* in particular in the azo range, which contain thiosulfate groups (–S–SO_3–) as water-solubilizing groups. In the presence of condensing agents (e.g., sodium sulfide), these compounds undergo polycondensation on the fiber to form insoluble dyes, S–S bridges being formed at the same time (e.g., *Inthion* dyes from Hoechst) [275], [276].

A further example is that of *esterified Naphtol AS* dyes. If azo dyes formed from Naphtol AS components and diazotized color bases are reacted with suitable acylation agents containing sulfo groups, soluble esters of azo compounds, which can be applied to the fiber, are obtained. By saponification in an alkaline medium the esters are then converted on the fiber into insoluble dyes, e.g., *Neocoton* dyes (Ciba); *Tinogenal* dyes (Geigy) and *Neogenol* dyes (Sandoz) [277].

Developing dyes in the second group have never been able to maintain their market position over a lengthy period: they are only mentioned here because their mode of action is of interest from the chemical aspect.

The first group of developing dyes, on the other hand, includes products of major industrial interest. Among all the developing dyes, by far the greatest importance attaches to the insoluble dyes known as Naphtol AS dyes, which are produced on the fiber from 2-hydroxy-3-naphthoic acid amides and other carboxylic acid amides capable of coupling (Naphtol AS components) and suitable diazo components. Only members of this class of compounds are dealt with in detail

in this chapter. In accordance with dyeing and performance aspects they can be divided into

developing dyes for cotton
developing dyes for animal fibers
developing dyes for hydrophobic fibers

The large number of dyeing and printing methods used will be discussed under the keywords → Textile Dyeing and → Textile Printing.

6.1. Developing Dyes for Cotton [5, vol. III, p. 290], [42, vol. 1, p. 502], [278–280]

The principal fields of application are the dyeing and printing of textiles made from cellulosic fibers (natural or regenerated), of which cotton is the main material. Cotton dyeing with developing dyes was being practiced by the end of the 19th century (ice colors); it received a boost with the introduction of Naphtol AS components in 1911. Ten years later, the *Rapid Fast* range (see page 82) made possible an especially simple method of application in the printing of cotton. Cotton dyeing and printing on the basis of developing dyes have now reached such an advanced stage of development, and the dyeings obtained possess such a high quality level, that this class of dye is likely to continue to maintain its market position in the future.

6.1.1. Developing Dyes for Dyeing

6.1.1.1. β-Naphthol Dyes

β-Naphthol dyes are only of historical interest. The initial results achieved in the dyeing of cotton by producing an azo dye on the fiber were patented by TH. and R. HOLLIDAY (1880). They impregnated cotton with an alkaline solution of β-naphthol and, after drying, "developed" the impregnated material by immersing it in an acetate-buffered solution of diazotized α- or β-naphthylamine. In particular the attractive red (para red) obtained by GALLOIS and ULLRICH (1889) using 4-nitroaniline as diazo component was for a long time of importance in the dyeing and textile printing industries. The introduction of further suitable diazo components led to a whole range of shades in these "ice colors" (so-called because of the ice used during diazotization); see Table 12.

Table 12. Amines for the development of cotton impregnated with β-naphthol and shades obtained

5-Nitro-2-aminotoluene	orange
5-Nitro-2-aminoanisole	pink
4-Chloro-2-aminoanisole	scarlet
4-Nitroaniline	red
1-Naphthylamine	bordeaux
3,3′-Dimethoxy-4,4′-diaminodiphenyl (with Cu salts)	blue
3,3′-Dimethyl-4,4′-diaminodiphenyl	brown
Mixture of 4,4′-diaminodiphenylamine and 4-dimethylamino-4′-aminoazobenzene	black

Numerous disadvantages of β-naphthol dyeing (difficulty in obtaining uniform fixation of the coupling component when drying material impregnated with β-naphthol; very limited fastness of the dyeings to light, chlorine, washing, and rubbing) pushed subsequent developments in the direction of the so-called Naphtol AS dyes.

6.1.1.2. Naphtol AS Dyes

In Offenbach (now Hoechst) in 1911, A. WINTHER, L. LASKA, and A. ZITSCHER discovered that instead of β-naphthol it is more advantageous to use the anilide of 2-hydroxy-3-naphthoic acid, listed in the technical literature under its trade name *Naphtol AS;* AS is short for "**A**mid einer **S**äure" (amide of an acid). Compared with that of β-naphthol, the sodium salt of this coupling component possesses the great advantage of having a certain degree of substantivity, i.e., it is absorbed onto the cotton fiber like a direct dye. Uniform fixation on the substrate is thus guaranteed. Moreover, in comparison with β-naphthol, the alkaline solutions of Naphtol AS possess higher air resistance, with the result that the moist cotton fibers impregnated with it can be exposed to air for a prolonged period without undergoing a change. The dyeings obtained on the fiber after development with suitable diazo components are distinguished by a high fastness level.

With the development of a large number of substituted anilides of 2-hydroxy-3-naphthoic acid (known as "β-oxynaphthoic acid arylides") and related compounds as Naphtol AS coupling components (the various products are denoted by additional letters following the AS) and a selection of suitable aromatic amines ("Fast

Color Bases") and their stabilized diazonium salts ("Fast Color Salts") as diazo components, the Naphtol AS dyeing sector underwent rapid expansion.

World production of Naphtol AS components, Fast Color Bases, and Fast Color Salts for dyeing and printing of cellulosic fibers amounts in 1982 to approx. 24000 t (around 6000 t are in addition used for production of azo pigments). This is equivalent to just under 5 % of colorant consumption in the textile sector as a whole, although as a result of a decrease in the proportion of world fibers demand accounted for by cotton, a declining trend can be observed in the percentage rate.

The percentage of Naphtol AS developing dyes in the dyestuffs requirement for textiles made from cellulosic fibers has for years remained constant at around 8 %, although the use of reactive dyes, particularly since the start of the 1970s, has gained increasingly in importance. Around 60 % of Naphtol AS dyes are employed for dyeing, around 40 % in the printing sector.

Table 13 provides a survey of the most important ranges of coupling and diazo components for Naphtol AS dyeing (and printing). The manufacturing countries are combined into three large groups: Western Europe and USA, India and Japan, and the state-trading countries. The first group accounts for about 50 % of production volume, and the other two have approximately equal importance.

The trade names indicate that the spelling Naph*t*ol AS is reserved for the Hoechst range. Naph*th*ol AS components belong to other ranges. The Hoechst spelling is the one mainly used below.

Coupling Components. The following factors were of major importance to the rapid development of numerous new coupling components and their use in cotton dyeing: a simple synthesis process for substituted 2-hydroxy-3-naphthoic acid aryl amides, improved dyeing properties (substantivity, fastness, shades for combinations), and reduction in the sensitivity of the moist impregnated material to air.

2-Hydroxy-3-naphthoic acid can be produced under pressure by reaction of carbon dioxide with the sodium salt of β-naphthol (Kolbe-Schmitt synthesis). The aryl amides are obtained by one-step reaction, 0.4 – 0.5 equivalents of phosphorus trichloride being added slowly at 70 – 80 °C to a mixture of one equivalent each of carboxylic acid and amine in an inert solvent (usually toluene and polychloroalkanes). The reaction mixture is then boiled for some hours and, after cooling, treated with soda to remove the acid. The Naphtol AS compound is filtered off, the solvent recovered. The yield is almost quantitative. The addition of formaldehyde to alkaline solutions of 2-hydroxy-3-naphthoic acid amides improves the air stability of the damp impregnations. The stabilizing action of the formaldehyde can be explained by the reversible formation of hydroxymethyl compounds, the sodium salts of which possess improved hydrolysis resistance.

By varying the substituents in the aryl amide radical and in the hydroxynaphthoic acid moiety, a wide range of coupling components was synthesized. The substituents naturally influence the shade and fastness properties of the azo dyes obtained after development. For example, the introduction of alkoxy groups in the aryl amide radical has a favorable influence on light fastness.

With the commercial products compiled in Tables 14 and 15, the shades obtained with diazotized color bases, such as chloro- or nitroanilines, are almost always orange, red, or bordeaux. Therefore, these Naphtol AS derivatives can also be referred to as "Red Naphtols."

The classification principle used in the Tables is largely that adopted in the Colour Index [7], which is based on the chemical constitution. The trade names are mainly those of the Hoechst range, since as a result of the historical development, the field of Naphtol AS developing dyes is an undisputed domain of this manufacturer. A number of compounds are listed under the different designations of other manufacturers (e.g., *Naphthol, Naphtanilid*). Only in some cases do the letter combinations in the trade names of different firms agree with regard to one and the same compound.

Whereas with 2-hydroxy-3-naphthoic acid amides, only red shades (apart from a few reddish brown shades) can be obtained by varying the aryl amide radical in combination with simple diazo compounds, coupling components derived from *o*-hydroxycarboxylic acids of other ring systems yield brown and black shades. Here, derivatives of carbazolecarboxylic acid

Table 13. Trade names of important coupling components and diazo components for Naphtol AS dyeing and their manufacturers

Country	Manufacturer	C. I. Azoic Coupling Component	C. I. Azoic Diazo Component	
			Bases	Stabilized salts
Germany, Fed. Rep.	Hoechst	Naphtol AS-	Echt-Base	Echt-Salz Azanil-Salz Variaminblau
Switzerland	Rohner	Naphtanilid	Fast Base	Diazo Fast
USA	Blackman Uhler Pfister Chem.	Naphtol AS- Naphtol AS-	Fast Base Azoene Fast Base	Fast Salt Azoene Fast Salt
India	Amar Dye Atul Prod.	Amarthol AS- Tulathol AS-	Amarthol Fast Base Tulabase Fast	Amarthol Fast Salt –
Japan	Daito Chem. Ind. Sanyo Colour Works Sumitomo Showa Chem. Co.	Daito Grounder Sanatol Naphthoide Kako Grounder	Daito Base Sanyo Fast Base (Fast) Base Kako Base	Daito Salt Sanyo Fast Salt Salt Kako Salt
China, P.R.		Youhaothol AS-	Fast Base	Fast Salt Youhaoriamin Blue
Czech Rep.		Ultrazol	Fast Base	Fast Salt
Poland	Ciech	Naphtoelan	Naphtoelan Fast Base	Naphtoelan (Fast) Salt
fromer Soviet Union		Azotol	Azoamin	Diazol

and dibenzofurancarboxylic acid are of particular importance (see Table 16).

Yellow dyeings are not obtainable by combination of *o*-hydroxyarylcarboxylic acid amides with color bases. The coupling components used for yellow shades are derived from acetyl- and benzoylacetic acid amides. Although in the chemical sense they are therefore not naphthol derivatives, they are classified as Naphtol AS components and in accordance with their coloristic properties are often called "Yellow Naphtols" (Table 17).

In contrast to the red Naphtols (Tables 14 and 15) and heterocyclic coupling components (Table 16), the yellow Naphtols are stable in alkaline solution; therefore, no addition of formaldehyde is necessary, and because of side reactions it is even inadmissible.

Apart from the coupling components listed in the tables, mention should also be made of those that as a special structural characteristic contain azo groups, e.g., *Naphthol AS-BN* (Blackman Uhler) (*C. I. Azoic Coupling Component 43,* exact constitution not as yet published).

An interesting development during the sixties was *Naphtol AS-FGGR* (Bayer) (*C. I. Azoic Coupling Component 108*): the condensation product of a blue dye (nickel phthalocyanine sulfochloride) and a coupling component in the yellow range (aminophenyl pyrazolone derivative) yields in combination with suitable diazo components brilliant yellowish greens with outstanding light fastness [281]. Since the coupling component possesses virtually no affinity for the cellulosic fiber, the product was mainly used for textile printing by the Rapidogen process (see page 83). Green dyeings are now obtained almost exclusively by combination of phthalocyanine developing dyes, e.g., *Phtalogen* (Bayer), with yellow Naphtol AS dyes.

Diazo Components. A large number of diazotizable aromatic amines are available for azoic coupling with the coupling component (see Table 18). The diazo components are sold either as nondiazotizable bases, so-called "Fast Color Bases" (both as free amines, if they are stable and nontoxic, and as hydrochlorides or sulfates), or as stabilized diazonium salts, so-called "Fast Color Salts" (for trade names see Table 13). The diazonium compounds are isolated ("precipitated") in the form of chlorides, sulfates, hydrogen sulfates, double salts of zinc chloride, fluoroborates, oxalates, and neutral or acid 1,5-

Table 14. Naphthol AS components: 2-hydroxy-3-naphthoic acid amides

Trade name	Constitution 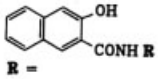	C. I. Azoic Coupling Component, generic/constitution no.	CAS reg. no.
Naphtol AS		2/37505	[*92-77-3*]
Naphtol AS-BS		17/37515	[*135-65-9*]
Naphthol AS-AN	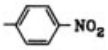	27/37516	[*2208-20-0*]
Naphthol AS-MCA		22/37511	[*5442-40-0*]
Naphtol AS-E		10/37510	[*92-78-4*]
Naphtol AS-D		18/37520	[*135-61-5*]
Naphthol AS-RT Naphtanilid PTO	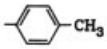	31/37521	[*3651-62-5*]
Naphtol AS-OL		20/37530	[*135-62-6*]
Naphtol AS-RL	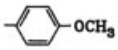	11/37535	[*92-79-5*]
Naphtol AS-PH		14/37558	[*92-74-0*]
Naphtol AS-VL Naphthol AS-BP	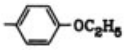	46/37559 30/37559	[*4711-68-6*]
Naphtol AS-TR		8/37525	[*92-76-2*]
Naphtol AS-KB	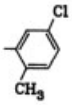	21/37526	[*135-63-7*]
Naphtol AS-CA Naphtanilid EL	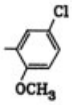	34/37531	[*137-52-0*]
Naphthol AS-MX Naphtanilid MX	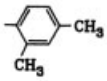	29/37527	[*92-75-1*]
Naphtol AS-AG*	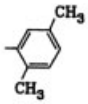	42/37611	[*3785-25-9*]

Table 14. (Continued)

Trade name	Constitution 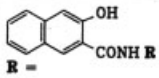	C. I. Azoic Coupling Component, generic/constitution no.	CAS reg. no.
Naphtol AS-LT	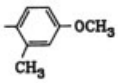	24/37540	[*3689-20-1*]
Naphtol AS-BG	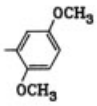	19/37545	[92-73-9]
Naphtol AS-RS	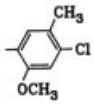	28/37541	[*5165-81-1*]
Naphtol AS-LC	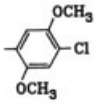	23/37555	[*4273-92-1*]
Naphtol AS-ITR	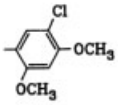	12/37550	[*92-72-8*]
Naphtol AS-BO		4/37560	[*132-68-3*]
Naphtanilid S	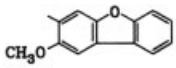	32/37580	[*2672-81-3*]

∗ A former Hoechst product, no longer manufactured by any company.

naphthalenedisulfonates and other aromatic sulfonates, the optimum conditions depending on the type of base. Final stabilization is achieved by mixing ("adjusting") with inert salts (usually aluminum sulfate). For further details → Diazo Compounds.

With the diazo components, too, the chemical constitution naturally has a great influence on fastness properties and shade of the dyeing obtained with a coupling component. It has been found that negative substituents (halogens, nitro, nitrile, RO_2S, and, in particular, CF_3 groups) in the aromatic nucleus exert a favorable influence on light fastness. In addition to suitable substituted aniline derivatives, which in combination with 2-hydroxy-3-naphthoic acid amides almost without exception yield red shades (orange, red, bordeaux), color bases have been developed that can be combined with the same Naphtol AS components to give blue and violet shades. Here, chief mention can be made of: *N*-aroyl-*p*-phenylenediamine derivatives, which contain no negative substituents, and 4-aminodiphenylamine derivatives. With compounds of the latter type (*Variamin Blue* bases and salts; Hoechst) attractive blue shades are obtained and in combination with *Naphtol AS-GR* (*C. I. Azoic Coupling Component 36,* 37585; see Table 16) green dyeings as well. 4-Aminoazobenzene derivatives produce black dyeings in combination with Red Naphtols.

The diazo components compiled in Table 18 have been classified first by coloristic aspects (shades) and second by constitutional characteristics (in accordance with the Colour Index [7]). The shades specified in the trade names relate,

Table 15. Naphthol AS components: derivatives of 2-hydroxy-3-naphthoic acid amides

Trade name	Constitution 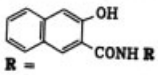	C. I. Azoic Coupling Component, generic/constitution no.	CAS reg. no.
Naphtanilid CF	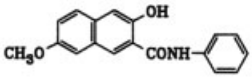	113/37567	
Naphtanilid CB	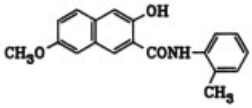	111/37568	
Naphtanilid HS	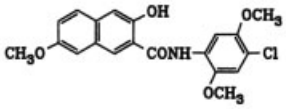	112/37569	
Naphtanilid M	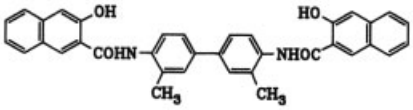	40/37570	[*3692-05-5*]
Naphtol AS-BR	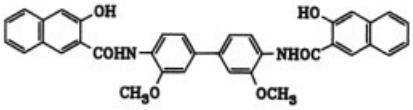	3/37575	[*91-92-9*]

with only a few exceptions to the dyeings obtained with Red Naphtols.

Basically speaking it is possible for the dyer to combine any aromatic base with any Naphtol AS component. However, only a limited selection of all conceivable combinations fulfil the preconditions for a coloristically high-grade dyeing. The dyeings generally possess high fastness to rubbing, washing, chlorine, and light. Many of the Naphtol AS dyes obtained on the fiber attain fastness levels of Indanthren dyes, without being chemically related to these compounds, which are usually polycyclic (examples see Table 19).

Other Developments. A new development in Naphtol AS dyeing, which commenced at the start of the 1960s, utilized the tendency of certain azo dyes to form metal complexes. Heteroaromatic amines that fulfil certain structural pre-conditions (see Section 2.3.4) are used as diazo components and the dyeings obtained on the fiber with Naphtol AS components by the usual method are converted into especially lightfast metal complex dyeings by aftertreatment with heavy-metal salts (copper, cobalt):

C. I. Azoic Diazo Components 135 – 137 (bases I, II, III from Hoechst, now withdrawn by the manufacturer on price grounds) [282–284], [15, p. 676].

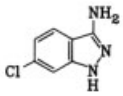

Co or Cu complex
C.I. 135/37310

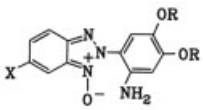

Cu complex
X = H, R = CH_3: C.I. 136/37325
X = OCH_3, R = C_2H_5: C.I. 137/37330

A further development concerns improvement in the dyeing methods. The conventional two-bath technique (impregnation of the material with the alkaline solution of the substantive coupling component, removal of the excess, and subsequent treatment with the diazonium salt solution) can often be considerably simplified and shortened by using the so-called *one-bath method:* a highly substantive Naphtol AS impregnating component and the developing component are prepared together in one bath. Not until the acid has been added is formation of the dye initiated. Diazo components suitable for this purpose consist of a diazoamino compound not

Table 16. Naphthol AS components: o-hydroxy compounds of polycondensed and heterocyclic arylcarboxylic acid amides

Trade name	Constitution	C. I. Azoic Coupling Component, generic/constitution no.	CAS reg. no.
	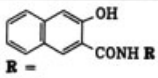		
Naphtol AS-GR Naphtanilid GR	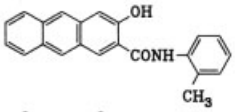	36/37585	[*1830-77-9*]
Naphtol AS-SG	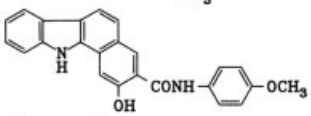	13/37595	[*86-19-1*]
Naphtol AS-SR	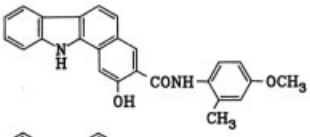	25/37590	[*5840-22-2*]
Naphtol AS-LB	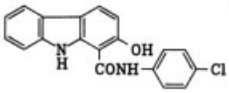	15/37600	
Naphtol AS-BT	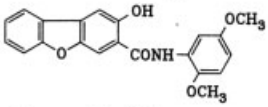	16/37605	[*132-62-7*]
Naphtol AS-KN	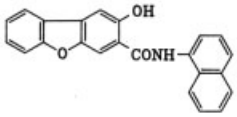	37/37608	[*4444-48-8*]

capable of coupling in the neutral and alkaline range, which in the presence of acid is regenerated into the active diazonium salt (see Section 6.1.2). The addition of dispersing agents to the developing component ensures that the insoluble dye formed in the liquor remains in colloidal dispersion and can be removed during the subsequent washing process (*Azanil* Color Salts from Hoechst) [285–288].

The introduction of *liquid brands* of azoic coupling components and azoic diazo components results in a considerable simplification of important steps in the dyeing and printing with Naphtol AS developing dyes. The former are 25 – 45 % solutions of monoalkali salts of Naphtol AS components in high-boiling alcohols (usually ethylene or diethylene glycol) miscible with water. These highly concentrated molecular solutions with good storage stability permit dust-free handling of the coupling component and enable ready-to-use impregnating baths to be prepared especially quickly by simply pouring in the relevant liquid brand into aqueous diluted sodium hydroxide solution. The liquid brands of the diazo components are either molecular solutions or dispersions: aromatic amines with low melting points (down to approx. 80 °C) are dissolved as free bases in monoalkyl ethers of diethylene glycol (occasionally with the addition of dimethylsulfoxide or *N*-methylpyrrolidone); higher melting amines (*mp* above 100 °C) can in the presence of suitable surfactants and other auxiliaries (ethylene glycol, glycerol, or sorbitol) be processed in more highly concentrated dispersions. Apart from dust-free handling, all liquid brands of diazo components are distinguished by particularly rapid and complete diazotizability without troublesome side reactions [289] (see also page 82).

Table 17. Naphthol AS components: acylacetic acid amides

Trade name	Constitution 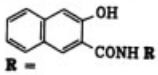	C. I. Azoic Coupling Component, generic/constitution no.	CAS reg. no.
Naphtol AS-IRG	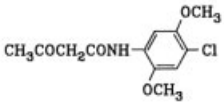	44/37613	[*4433-79-8*]
Naphtanilid L4G	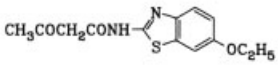	9/37625	[*4273-88-5*]
Naphtol AS-G	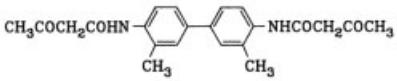	5/37610	[*91-96-3*]
Naphtanilid LG	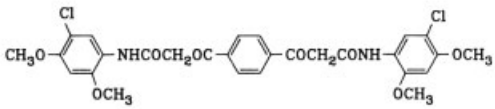	35/37615	[*92-21-7*]
Naphtanilid L3G	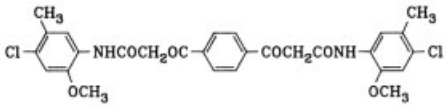	33/37620	[*6492-81-5*]

6.1.2. Developing Dyes for Printing [5, vol. III, p. 290 – 301], [278–280], [290]

For details of the printing method → Textile Printing.

Two-Stage Printing Processes. As with the dyeing of cotton fibers, two-stage methods can also be employed when printing. Depending on whether the coupling component (as sodium salt) or the diazo component (as diazonium salt) is applied during the printing process, a distinction is made between naphtholate printing processes and base printing processes.

In the *naphtholate printing process,* the sodium salts of the Naphtol AS compounds in the form of aqueous, thick pastes (additions of high-swelling, nonionic or anionic polysaccharides) are printed on the textile material and then developed as azo dyes on the fiber by treatment with the solution of a diazotized base.

Naphtholate printing has always been of only secondary importance.

In the so-called *base printing process,* the cellulosic material is first impregnated with the solution of a sodium naphtholate and, after drying, the thickened solution of the diazotized Fast Color Base is printed onto the material.

In contrast to cotton dyeing, with the wide diversity of Naphtol AS derivatives that are used, price, coloristic, and substantivity reasons dictate that only two components capable of coupling are in the main used for impregnation of the textile to be printed: *Naphtol AS* (*C. I. Coupling Component 2,* 37505, see Table 14) as red component and *Naphtol AS-G* (*C. I. Coupling Component 5,* 37610, see Table 17) as yellow component. The cotton impregnated with the sodium salts of these Naphtols by padding and subsequent drying has good, if only limited, stability when stored in a cool, dark place. Because of their low substantivity for the cellulosic fiber, Naphtol AS and Naphtol AS-G additionally satisfy the important requirement that after the base printing process they can be completely washed off again from the unprinted parts.

By combining the two Naphtols with numerous azoic diazo components, a relatively wide range of printed shades is possible. The addition of Naphtol AS enables shades in the or-

Table 18. Color bases and color salts

Trade name* Fast Base or Salt	Constitution 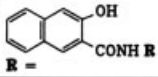	C. I. Azoic Diazo Component, generic/constitution no.	CAS reg. no.
Yellow GC		44/37000	[*137-04-2*]
Orange G Orange GC		2/37005	[*17333-84-5*]
Orange GR		6/37025	[*88-74-4*]
Orange R		7/37030	[*99-09-2*]
Orange GGD	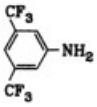	16/37045	[*121-50-6*]
Orange RD	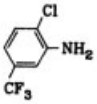	49/37050	
Orange LG	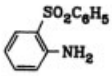	18/37060	[*4273-98-7*]
Golden orange GR	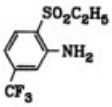	19/37065	
Scarlet GG Scarlet GGS	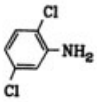	3/37010	[*95-82-9*]
Scarlet VD	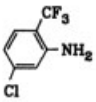	17/37055	[*445-03-4*]
Scarlet TR		46/37080	[*6259-40-1*]
Scarlet G	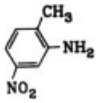	12/37105	[*99-55-8*]
Scarlet R Scarlet RC	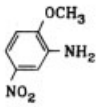	13/37130	

Table 18. (Continued)

Trade name* Fast Base or Salt	Constitution 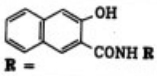	C. I. Azoic Diazo Component, generic/constitution no.	CAS reg. no.
Scarlet LG	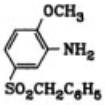	31/37145	[*6368-86-1*]
Red GG		37/37035	[*100-01-6*]
Red 3GL	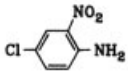	9/37040	[*89-63-4*]
Red FG	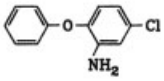	26/37070	[*6259-38-7*]
Red FR Red FRN	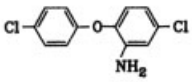	33/37075	
Red TR	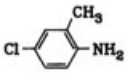	11/37085	[*3165-93-3*]
Red KB		32/37090	[*6259-42-3*]
Red RL	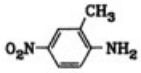	34/37100	[*99-52-5*]
Red GL	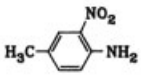	8/37110	[*89-62-3*]
Red RC	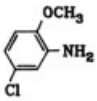	10/37120	[*4274-03-7*]
Red B	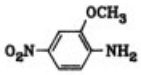	5/37125	[*97-52-9*]
Red SL	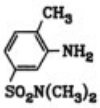	132/37111	[*12216-87-4*]
Red ITR	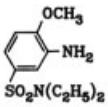	42/37150	[*97-35-8*]

Table 18. (Continued)

Trade name* Fast Base or Salt	Constitution 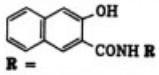	C. I. Azoic Diazo Component, generic/constitution no.	CAS reg. no.
Red SW	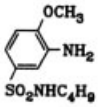	14/37151	[*80-22-8*]
Red AL	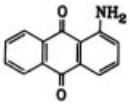	36/37275	[*16048-40-1*]
Red KL	not yet published	121/–	
Garnet GBC	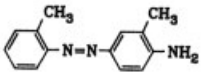	4/37210	[*2298-13-7*]
Garnet GC	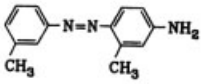	27/37215	[*4335-70-0*]
Garnet RH	not yet published	123/–	
Bordeaux GP	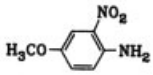	1/37135	[*96-96-8*]
Bordeaux BD	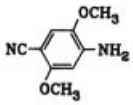	40/37170	[*6368-91-8*]
Corinth LB	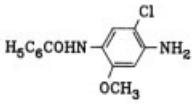	43/37160	[*6368-90-7*]
Corinth V	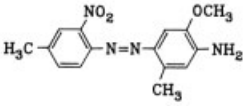	39/37220	
Violet B	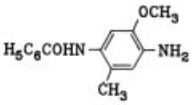	41/37165	[*27761-27-9*]
Blue RR	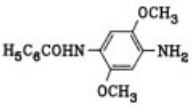	24/37155	[*6268-05-9*]
Blue BB	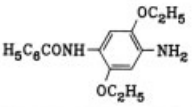	20/37175	[*120-00-3*]

Table 18. (Continued)

Trade name* Fast Base or Salt	Constitution 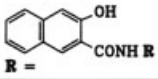	C. I. Azoic Diazo Component, generic/constitution no.	CAS reg. no.
Navy blue RA Navy blue RAS	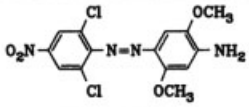	51/37195	
Blue B	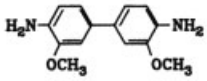	48/37235	
Variaminblue RT Variaminblue RTN	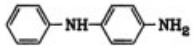	22/37240	[*101-54-2*]
Variaminblue FG Variaminblue FGC	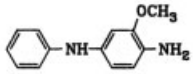	47/37250	
Variaminblue B Variaminblue BN	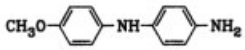	35/37255	[*6254-98-4*]
Bluegreen B	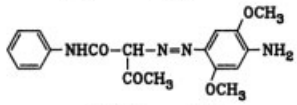	118/37222	
Olive BR	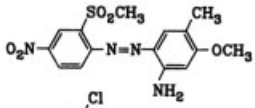	125/37197	
Brown RR	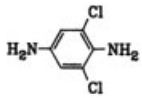	117/37020	[*609-20-1*]
Brown V Brown VA	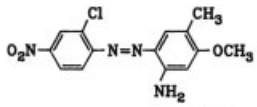	21/37200	[*4274-06-0*]
Black K	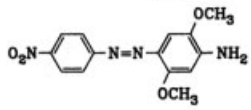	38/37190	[*27766-47-8*]
Black B Black BN	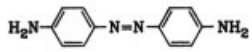	109/37245	[*6369-04-6*]
Black G	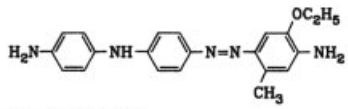	45/37260	[*6369-07-9*]
Black M	not yet published	131/–	

* Mainly those of the Hoechst range

ange, scarlet, red, bordeaux, violet, navy blue, and black range to be obtained, Naphtol AS-G is used in the yellow to brown range. If higher requirements must be met with regard to light fastness, Naphtol AS-G can be replaced by *Naphtol AS-IRG* (*C. I. Coupling Component 41,* 37613, see Table 17).

Right into the 1940s, the base printing process enjoyed considerable industrial importance, but in comparison with the one-stage process it then receded into the background. The development of reactive dyes resulted in a worldwide revival of the two-stage base printing process, because vinyl sulfone dyes, in particular, can in many cases be applied together with color salts or diazotized color bases. The possibility of printing a broader range of shades on cotton, which had long been desired, can thus be achieved [291].

The additional combination of the yellow to golden yellow azoic developing dyes with the blue phthalocyanine developing dyes, in particular with a cobalt-phthalocyanine derivative of clear, dark blue color (*C. I. Ingrain Blue 5,* Phtalogen Blue IBN [*37370-33-5*] from Bayer) makes it possible to match deep green shades and thus close a gap in the color range.

These subsequent developments have resulted in the base printing method acquiring considerable importance, in particular for so-called African and Java prints.

Other Developments. In the conventional diazotizing process, the differences in the reactivity of aromatic amines frequently lead to the uncontrolled release of nitrous gases and undesired side reactions, in spite of the fact that ice is used. With the use of liquid brands (see page 75) instead of the powder of the free base, it was possible to increase the rate of diazotization considerably and to eliminate the formation of byproducts. A further refinement of this technique led to an especially elegant variation of the base printing process, the so-called *DIP process* ("**D**iazotization **i**n the **p**aste"): the aqueous – alcoholic dispersion or molecular solution of the base is stirred into a paste of sodium nitrite which without being cooled is then mixed with the aqueous paste of a medium-strong acid (it is best to use phosphoric acid). Diazotation is completed quickly without any formation of nitrous gases. The diazo salt paste obtained no longer needs to be buffered; during the subsequent printing process it couples rapidly and completely with the Naphtol in the impregnated fabric [289], [292].

The possibility of diazotizing mixtures of different azoic diazo components at the same time in a single paste has proved to be a particular advantage of the diazotization process described. The use of mixed coupling products thus opened up a further possible method of shade matching.

One-Stage Printing Processes. Especially simple methods of printing cotton fibers were developed using the one-stage direct printing process. In this technique, the diazo component in the form of a stabilized diazo compound, i.e., not initially capable of coupling, is applied at the same time as the coupling component and coupling is then initiated by a damp aftertreatment in an acid medium or by acid, or frequently also neutral steaming.

Three different diazo derivatives are employed in textile printing as stabilized diazo compounds (a more appropriate term is concealed diazo compounds): antidiazotates, diazo sulfonates, and diazoamino compounds (for properties and production → Diazo Compounds).

Combinations with diazoamino compounds were of particular industrial importance, but with the appearance of the reactive dyes, they, too, receded into the background.

Because of its relatively complicated chemical background, the one-stage printing process must be dealt with in somewhat more detail.

Combinations with Antidiazotates. Unlike the unstable syndiazotates, the stable antidiazotates do not couple with alkaline β-naphthol solution, or if so only very slowly.

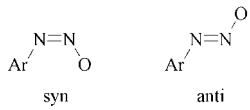

Not until the acid treatment is carried out is the diazonium ion formed with subsequent rapid azo coupling. This fact acquired importance in textile printing, diazotates of weakly basic, mainly chlorinated amines being used in

particular, because of their good stability and development ability.

Antidiazotates are obtained in a simple manner by introducing a diazonium salt solution into concentrated aqueous alkali and heating to temperatures above 100 °C, whereby the syndiazotate formed initially is rearranged into the anti compound.

Mixtures of sodium salts of 2-hydroxy-3-naphthoic acid amides and other coupling components in the Naphtol AS range with antidiazotates in the form of pastes and powder brands have been marketed since 1921 as *Rapid Fast* dyes. If the fibers impregnated or printed with them undergo an acid aftertreatment, regeneration of the active diazonium salt and coupling of the azo dye take place on the substrate. A number of combinations and the shades obtained with them are compiled in Table 20, considerable importance attaching in particular to *Rapid Fast Blue BN*.

Table 19. Combinations for developing dyes in the Naphtol AS range with high fastness properties

Naphtol	Color base	Dyeing obtained
AS-L4G	Fast Orange RD	yellow
AS-OL	Fast Orange GGD	orange
AS-OL	Fast Scarlet GGS	scarlet
AS-LT	Fast Scarlet LG	red
AS-ITR	Fast Red ITR	red
AS-ITR	Fast Bordeaux BD	bordeaux
AS-E	Variamin Blue FGC	blue
AS-FGGR	Fast Yellow GC	green
AS-BT	Fast Scarlet TR	brown
AS-SG	Fast Red B	black

Table 20. Examples of printing combinations with antidiazotates

Dye	Antidiazotate of	Coupling component
Rapid Fast Yellow GH	Fast Scarlet GG Base	Naphtol AS-G
Rapid Fast Orange RH	Fast Scarlet GG Base	Naphtol AS-PH
Rapid Fast Red RH	Fast Red RC Base	Naphtol AS-OL
Rapid Fast Bordeaux RH	Fast Red R Base	Naphtol AS-BS
Rapid Fast Blue BN	Fast Blue B Base	Naphtol AS
Rapid Fast Brown IBH	Fast Red R Base	Naphtol AS-LB

The generally limited storage stability of the mixtures has proved to be a disadvantage of the process described, because slow splitting of the antidiazotate takes place in the alkaline printing paste.

Combinations with Diazosulfonates. Diazosulfonates $Ar{-}N{=}N{-}SO_3^-$, obtained by reaction of diazonium salts with sulfites, do not couple in their stable anti form with 2-hydroxy-3-naphthoic acid amides, or if so only slowly. The diazonium compound capable of coupling is reformed when exposed to light or when the temperature increases.

The *Rapidazol* dye range introduced in 1932 utilizes this property: ready-to-use mixtures of diazosulfonates, in particular from the 4-aminodiphenylamine range (*Variamin Blue* bases, see Table 18), with coupling components of the Naphtol AS type are printed on the fiber and then developed into azo dyes in neutral steam, especially in the presence of oxidizing agents.

In spite of its limited color scale, the range, with its high-quality blue and black brands, was a welcome addition to the diazotate mixtures. It was especially important at a time when no diazoamino compounds capable of being developed in a neutral medium were available.

Combinations with Diazoamino Compounds [6, p. 177]. Among the Naphtol AS combinations for printing, it is the stable diazoamino compounds that are mainly used. Since the release (1930) of the first *Rapidogen* dyes up to the worldwide introduction of the reactive dyes, standardized mixtures of these combinations had enjoyed major industrial importance. The most important commercial ranges based on diazoamino compounds are compiled in Table 21.

Table 21. Designations of important ranges of Naphtol AS printing combinations with diazoamino compounds

Country	Manufacturer	Ranges
Germany, Fed. Rep.	Bayer	Rapidogen
France	ICI-Francolor	Neutrogen
Switzerland	Rohner	Ronagen, Sinagen
USA	Blackman Uhler	Printing
	Pfister Chem.	Azogen
India	Amar Dye	Stabagene
	Atul Prod.	Tulagene
Japan	Kiwa Chem. Ind.	Polydogen
	Sanyo Col. Works	Thiugen
	Sumitomo	Sumika Fast
Poland		Pologen

Diazoamino compounds (further details → Diazo Compounds) are obtained by reacting

a diazonium salt (**200**) with a primary or secondary aliphatic or aromatic amine:

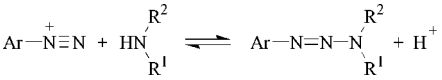

In the case of aromatic amines, the diazoamino compound only forms if blocking or deactivation of the coupling positions in the aromatic nucleus prevents the formation of an aminoazo compound. The reaction (13) is reversible, i.e. the addition of acid causes the diazoamino compounds to be split again into their initial components. If a primary amine R–NH_2 (**201**) is used, the resulting triazene (**202**) is in balance with its tautomer (**202a**):

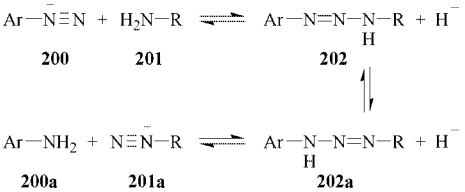

The splitting process can thus lead to the formation of a new diazonium salt (**201a**) ("rediazotization").

The tautomer equilibrium **202** ⇌ **202a** is dependent on the nature of the two radicals Ar and R: as the basicity of amine Ar−NH_2 (**200a**) increases and that of amine R−NH_2 (**201**) decreases, the equilibrium shifts in favor of the diazoamino compound (**202**), which is split in the desired manner to form the original diazonium salt.

For practical usage as a stabilized diazo compound in printing combinations, only those diazoamino compounds are suitable that satisfy the following conditions:

Well-defined splitting without rediazotization.
No possibility for the diazonium salt obtained after splitting to couple with the amine component.
Solubility of the diazoamino compound in water.

As amine component, the so-called stabilizer, use is thus generally made of non-diazotizable secondary amines, or the equilibrium is influenced by 2nd order substituents in the stabilizing amine (decreasing basicity). When aromatic stabilizers are used, the *o*- and *p*-position in relation to the amino group is usually blocked by substituents, with the result that after splitting of the diazoamino compound no coupling to aminoazo products can occur. If a possible coupling position is free, the azo coupling is handicapped by the deactivating influence of carboxylic acid or sulfonic acid groups. These substituents additionally impart the desired water solubility to the diazoamino compound used in the printing combinations.

Stabilizers frequently used are sarcosine, *N*-methyltaurine and anthranilic acid derivatives:

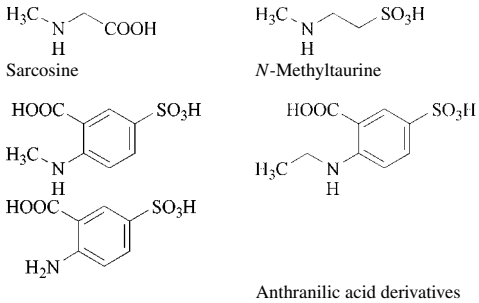

The diazoamino compound (for further details → Diazo Compounds) is generally obtained by adding a diazonium salt solution to the neutral or alkaline solution of the stabilizer in the form of its sodium salt. The reactivity of the diazonium salt determines the speed of the reaction. The diazoamino compound precipitated or salted out is filtered, dried, ground, and adjusted with the sodium salt of a Naphtol AS component.

After the textile fabric has been printed with the printing combination, the dyeing is developed by acid steaming (in the presence of acetic and formic acid vapors). The reaction of the two soluble components on the fiber to form the insoluble azo dye is illustrated by the example of *Rapidogen Red GS* (**205**).

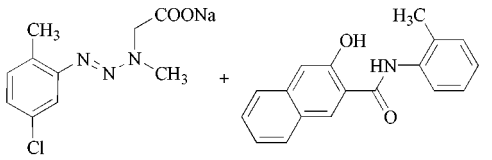

203 **204**

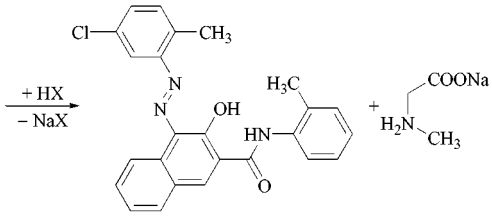

205

203: Diazoamino compound of Fast Red KB base and sarcosine.
204: Naphtol AS-D

The diversity of combinations on the market makes it possible to obtain dyeings of any shade from greenish yellow through orange, scarlet, red, bordeaux, violet, blue, green to black.

The need for acid development has occasionally proved to be a certain disadvantage. This step can be avoided by using a steam-volatile organic base, such as diethylaminoethanol, instead of the alkali employed in the Naphtol AS diazoamino combinations. During neutral steaming, the alkali component in the printing mixture then escapes and azo coupling takes place, especially in the presence of salts which in the steam undergo hydrolysis to form compounds with an acid reaction (e.g., ammonium or pyridinium salts). Selection of suitable modified stabilizers, such as **206** or **207**, also enables the splitting tendency of the diazoamino compounds to be so carefully "counterbalanced" that the relevant printing combinations can be developed in neutral steam (*Rapidogen N, Neutrogen dyes*, etc., see Table 21).

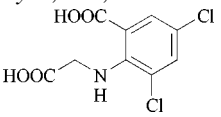

206

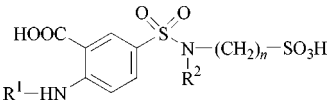

207

For use of diazoamino compounds for the dyeing of cotton, see page 77, and of synthetic polyamide fibers, see Section 6.3.

6.2. Developing Dyes for Animal Fibers

Compared with the dyeing and printing of hydrophilic cellulosic fibers, the dyeing of animal fibers (protein fibers) with products in the Naphtol AS range has never attained industrial importance. In this connection, mention is made here only of the Hoechst ranges of *Ofna-Lan* and *Rauna* dyes, which are no longer on the market and were used for the dyeing of wool – cellulosic blends and furs. The problem with this dyeing method proved to be the alkalinity of the impregnating baths (alkaline solutions of the Naphtol AS coupling components), because of the sensitivity to hydrolysis of the peptide compounds of animal fibers [293].

Pure silk can be dyed in the same way as cellulosic fibers, since, unlike wool, silk is largely insensitive to alkali.

6.3. Developing Dyes for Hydrophobic Fibers

Naphtol AS dyes have also been used commercially for dyeing hydrophobic synthetic fibers, particularly polyester and polyamide. Two methods were developed [294]:

In the *modified azo process*, the coupling component and the azoic diazo component are applied to the fiber consecutively or from an aqueous bath at 80 – 130 °C, if necessary in the presence of a carrier, and the fiber impregnated by this means is then treated with nitrous acid. Development of the dyeing takes place with diazotization and simultaneous coupling.

Products are being introduced that in the optimum mixture ratio contain selected combinations of coupling and diazo components capable of development (*Intramin* dyes, Hoechst; *Ronasyn* dyes, Rohner). The process only gained importance in the inexpensive black dyeing of polyester fibers (*C. I. Azoic Black 16,* e.g., *Intramin Black G*, Hoechst). Multicolored shades are more advantageously obtained with disperse dyes.

When *using diazoamino compounds* (so-called *Ofnaperl process*, Hoechst) a suitable diazoamino compound is added to the alkaline bath, in which the Naphtol AS component is dissolved. After the fiber is impregnated in this bath at 60 – 80 °C, the development process is initiated in the warm acid bath by liberating the concealed diazonium salt. Unlike the components of the printing combination, the diazoamino compounds used contain no carboxylic acid or sulfonic acid groups; they have only limited solubility in water but must be readily dispersible (*C. I. Azoic Diazo Components 127, 128*, and *130*, *Ofnaperl* Salts, Hoechst). The process has proved especially effective for the dyeing of polyamide fibers (perlon, nylon).

7. Disperse Azo Dyes (→ Disperse Dyes)

Disperse dyes are dyes sparingly soluble in water which are used for dyeing synthetic fibers. The dyes are added to the aqueous dye bath in finely dispersed form and are uptaken out of this dispersion by the hydrophobic fiber.

The fiber hereby represents the organic phase, in which the disperse dyes are more soluble than in the aqueous dyebath.

About half of all disperse dyes belong to the azo range. Originally developed for cellulose acetate fibers, disperse dyes are nowadays used in large quantities for dyeing polyester fibers.

8. Azo Pigments (→ Pigments, Organic)

Pigments are coloring matters that are practically insoluble in the application media (DIN 55944). Coloring matters are all color-spending materials according to DIN 55945.

Azo pigments are the most important class within the group of organic pigments. Mono- and disazo pigments are the two main ranges.

Whereas in English a strict differentiation exists between pigments (insoluble coloring matters) and dyes (soluble coloring matters) in German some confusing terms are still occasionally used, e.g., "Pigmentfarbstoffe", "Körperfarben", "Farbkörper", "Lackfarbstoffe". These terms should be discontinued and only "pigments" and "dyes" used.

Azo pigments, like all other organic pigments, serve for the coloration of polymeric materials such as plastics, elastomers, rubber, paints and other coating materials, and for printing pastes. The relatively inexpensive azo pigments have gained noticeable importance due to the rapid economic and industrial development in the past three decades.

9. Alcohol- and Ester-Soluble Dyes

With the exception of the blue copper phthalocyanine derivatives, these products are azo dyes soluble in polar solvents, such as alcohols, glycols, esters, glycol ethers, and ketones. Dyes soluble in alcohol and esters are used in protective lacquers for the transparent coating of metal (aluminum) foils and other materials, such as wood (greening lacquers); in flexographic inks for the printing of metal foils, cellophane and paper; as well as for the coloration of cellulose esters, celluloid, and poly(vinyl acetates), and, in the office supplies sector, for stamping inks and pastes for pressure recorders. See Table 22 for important product ranges.

Table 22. Designation of important alcohol- and ester-soluble dye ranges

Manufacturer	Range
BASF	Neozapon, Zapon Fast
Bayer	Irisol Fast
Ciba-Geigy	Grasol Fast, Irgacet, Oracet, Orasol
GAF	Azosol Fast
ICI	Methasol
Sandoz	Savinyl

In modern formulations for foil lacquers, dyes with better solubility in esters are given preference over the older products, which are mainly soluble in alcohols. As regards chemical constitution, there is no precise delineation between alcohol- and ester-soluble dyes, nor between these two groups and the fat- and oil-soluble dyes described in Chapter 10.

In accordance with chemical aspects, the most important alcohol- and ester-soluble azo dyes can be divided into three groups:

1 : 2 Metal complexes of (mainly mono-) azo dyes, without sulfonic or carboxylic acid groups, and trivalent metals (→ Metal-Complex Dyes).

The metals chiefly suitable are chromium and cobalt, as well as nickel, manganese, iron, or aluminum. Diazo components are mainly chloro- and nitroaminophenols or aminophenol sulfonamides, coupling components are β-naphthol, resorcinol, and 1-phenyl-3-methylpyrazolone (5). Formation of a complex from an azo dye and a metallic salt generally takes place in the presence of organic solvents, such as alcohols, pyridine, or formamide. An example is *C. I. Solvent Red,* 12715 (**208**) [*33270-70-1*].

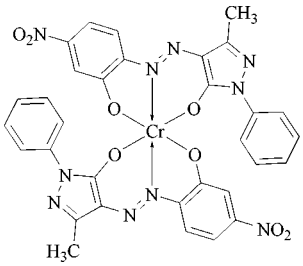

208

1 : 1 Metal-complex azo dyes that contain sulfonic acid or carboxylic acid groups and are present in the form of internal salts. Here, azamethin – metal complexes are also of importance. An examples is *C.* I. Solvent Yellow 32, 48045 (**209**) [*61931-84-8*].

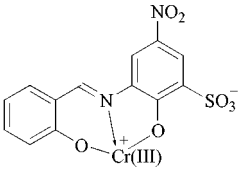

209

In the case of *C. I. Solvent Orange 56,* similar to 18745 : 1 [*13463-42-8*], also a metal complex, diazotized 2-hydroxy-3-amino-5-nitrobenzene sulfonic acid is coupled to 1-phenyl-3-methylpyrazolone (5).
C. I. Solvent Yellow 82, similar to 18690, is a 1 : 1 metal complex of the azo dye obtained from anthranilic acid and 1-phenyl-3-methylpyrazolone (5).
Reaction products of acid azo dyes, acid 1 : 1 metal-complex azo dyes, or 1 : 2 metal complex azo dyes without acid groups, with organic bases or cationic dyes.

Cyclohexylamine, dodecylamine, and sulfonium or phosphonium compounds serve as bases, whereas derivatives of the xanthene range (rhodamines) are mainly used as cationic dyes. Example: *C. I. Solvent Red 109* [*53802-03-2*] is composed of *Solvent Yellow 19*, 13 900 : 1, (**210**) [*10343-55-2*] and *Solvent Red 49*, 45 170 : 1, (**211**) [*81-88-9*]:

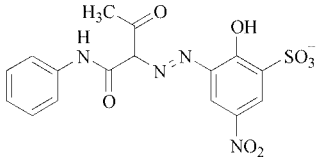

210

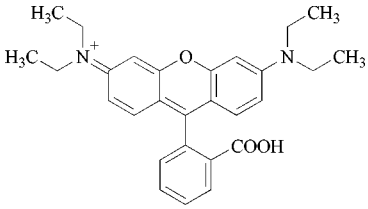

211

These dyes are saltlike compounds of a metal-complex azo dye acid and a base.

10. Fat- and Oil-Soluble Dyes

Fat- and oil-soluble dyes are also soluble in waxes, resins, lacquers, hydrocarbons, halogenated hydrocarbons, ethers, and alcohols, but not in water. It is not possible to differentiate clearly between them and the alcohol- and ester-soluble dyes (see Chap. 9).

Table 23. Designations of important fat- and oil-soluble dye ranges

Manufacturer	Range
Amer. Cyanamid	Calco
BASF	Sudan
Bayer	Ceres
Bitterfeld	Sudan
Du Pont	Oil
Ethyl Corp.	Oil Soluble
Hoechst	Fat
ICI	Waxoline
Siegle	Sico Fat
Williams	Oil

With the exception of the blue anthraquinone derivatives, all fat- and oil-soluble dyes belong

to the azo range. These azo dyes are generally based on simple components. According to their degree of solubility they usually contain hydroxy and/or amino groups, but not sulfonic acid and carboxylic acid groups. See Table 23 for important product ranges.

As examples, mention can be made of some of the most important fat- and oil-soluble azo dyes produced on an industrial scale: *C. I. Solvent Yellow 56,* 11021 (**212**) [*2481-94-9*]

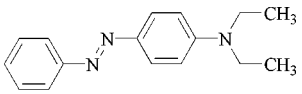

212

C. I. Solvent Yellow 14, 12055 (Orange) (**213**) [*842-07-9*]

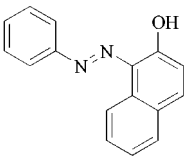

213

C. I. Solvent Yellow 16, 12700 (**214**) [*4314-14-1*]

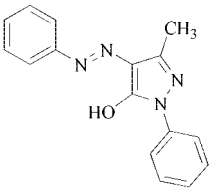

214

C. I. Solvent Red 23, 26100 (**215**; R = H) [*85-86-9*]; *C. I. Solvent Red 24,* 26105 (**215**; R = CH_3) [*85-83-6*]

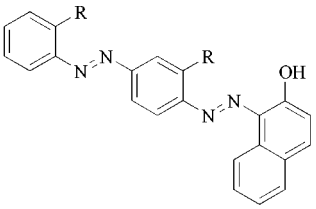

215

C. I. Solvent Black 3, 26150 (**216**) [*4197-25-5*]

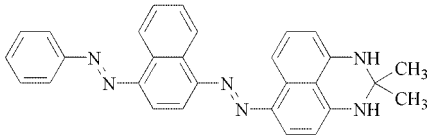

216

The products are sold in powder form, as granules, or as flakes, and some dyes also as liquid brands. The liquid brands are highly concentrated solutions of fat-soluble dyes in aromatic hydrocarbons, in some cases also solvent-free 100 % liquid products.

Fat- and oil-soluble dyes are used on a large scale in a wide variety of industrial sectors. Main fields of application are the coloration of products in the mineral oil and plastics industries, as well as of wax products (e.g., candles, shoe polish, floor polishes).

Mineral oil products (fuels, fuel oil, lubricating oils, and greases) are colored as a means of distinguishing between different grades (supergrade petrol) or for compulsory identification, e.g., of diesel and fuel oils, for duty purposes. The plastics industry values the fat- and oil-soluble dyes for the highly transparent colorings, usually with very good light fastness, that are obtainable with them. Products in this dye class are most frequently used for the coloration of polystyrene, as well as for incorporation in polymethacrylate and unsaturated polyester casting resins.

An interesting factor is the high degree to which the light fastness of fat- and oil-soluble dyes is dependent on the medium colored. Whereas, for example, 0.05 % colorations of the dye *Solvent Yellow 56,* 11021, in candle materials possess only moderate light fastness (step 3 on the eight-step Blue Scale), transparent colorations in polystyrene are distinguished by outstanding light fastness (step 8).

Other fields of application are the pyrotechnics industry (high sublimation tendency of the dyes), the office supplies industry (inks for felt-tip pens, oil-based stamping inks), the lacquers industry (especially coloration of transparent lacquers on aluminum foil) and the cosmetics industry (approval restricted to a few products).

11. References

1. P. Rys, H. Zollinger: *Farbstoffchemie,* 3rd ed., Verlag Chemie, Weinheim 1982.
2. *Houben-Weyl,* **10/3,** 213 – 465, 467 – 485.
3. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin 1964.
4. W. Seidenfaden: *Künstliche organische Farbstoffe und ihre Anwendungen,* Enke Verlag, Stuttgart 1957.
5. K. Venkataraman: *The Chemistry of Synthetic Dyes,* Academic Press, New York, vol. I (1952), vol. II (1952), vol. III (1970), vol. IV (1971), vol. VI (1972), vol. VII (1974).
6. H. Zollinger: *Azo and Diazo Chemistry,* Interscience, New York – London 1961.
7. Colour Index, Society of Dyers and Colourists and American Association of Textile Chemists and Colorists, 3rd ed., Bradford 1971, First Revision 1975, Second Revision 1982.
8. K. Holzach: *Die aromatischen Diazoverbindungen,* Enke Verlag, Stuttgart 1947.
9. *Houben-Weyl,* **10/3,** 1 – 212.
10. N. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* 4th ed., Akademie Verlag, Berlin (DDR) 1966.
11. H. E. Fierz-David, L. Blangey: *Grundlegende Operationen der Farbenchemie,* 8th ed., Springer Verlag, Wien 1952.
12. P. F. Gordon, P. Gregory: *Organic Chemistry in Colour,* Springer Verlag, Berlin – Heidelberg – New York 1983.
13. A. Eibner, *Justus Liebigs Ann. Chem.* **316** (1901) 126; H. Bucherer, A. Schwalbe, *Ber. Dtsch. Chem. Ges.* **39** (1906) 2798.
14. M. Martynoff, *Bull. Soc. Chim. Fr.* **18** (1951) 214.
15. H. Baumann, H. R. Hensel: "Neue Metallkomplexfarbstoffe. Struktur und färberische Eigenschaften", *Fortschr. Chem. Forsch.* **7** (1967) no. 4.
16. H. Pfitzner, H. Baumann, *Angew. Chem.* **70** (1958) 232.
17. D. Vorländer, F. Meyer, *Justus Liebigs Ann. Chem.* **320** (1902) 122.
18. D. C. Freemann, C. E. White, *J. Org. Chem.* **21** (1956) 379.
19. H. H. Hodgson, F. Leigh, G. Turner, *J. Chem. Soc.* 1942, 744.
20. F. Suckfüll, H. Dittmer, *Chimia* **15** (1961) 137.
21. S. Hünig et al., *Angew. Chem.* **70** (1958) 215; **74** (1962) 818.
22. M. Regitz, *Angew. Chem.* **79** (1967) 786.
23. T. Vickerstaff: *The Physical Chemistry of Dyeing,* Oliver and Boyd, London – Edinburgh 1954.
24. E. M. Diskant, *Anal. Chem.* **24** (1952) 1856.
25. G. Schetty, *J. Soc. Dyers Colour.* **71** (1955) 705.
26. Sandoz, BE 717458, 1967 (M. Studer).
27. Bayer, DE-OS 1923680, 1969 (H. Nickel, F. Suckfüll).
28. Toms River Chem. Corp., DE-OS 1931691, 1969 (H. A. Stingl).
29. Crompton & Knowles, DE-OS 1957115, 1969 (J. F. Feeman).
30. Bayer, DE-OS 1960816, 1969 (K. L. Moritz).
31. J. Boulton, *J. Soc. Dyers Colour.* **67** (1951) 522.
32. I. D. Rattee, M. M. Breuer: *The Physical Chemistry of Dye Absorption,* Academic Press, London – New York 1974.
33. H. Bach, E. Pfeil, W. Phillippar, M. Reich, *Angew. Chem.* **75** (1963) 407.
34. E. Schirm, *J. Prakt. Chem.* **144** (1935) no. 2, 69.
35. H. Zollinger, *Helv. Chim. Acta* **38** (1955) 1597, 1623; R. Pütter, *Angew. Chem.* **63** (1951) 186.
36. Cassella, DE 204212, 1907 (A. Greßly).
37. BASF, DE 46737, 1888; DE 47902, 1889 (C. L. Müller).
38. Ciba, DE 436179, 1923.
39. Ciba, DE 335809, 1915.
40. BASF, DE 807289, 1948 (H. Pfitzner, H. Merkel); H. Pfitzner, H. Baumann, *Angew. Chem.* **70** (1958) 232.
41. J. Offenbach: *Deutscher Färbekalender,* Eder Verlag, Stuttgart 1972, pp. 185 – 194.
42. L. Diserens: *Die neuesten Fortschritte in der Anwendung der Farbstoffe,* vol. 1, 3rd ed., 1951; vol. 2, 2nd ed., 1949, Birkhäuser, Basel.
43. Geigy, DE 114634, 1899.
44. GAF, DE 901419, 1950; DE 901540, 1950.
45. BASF, DE-OS 2557561, 1975 (H. Kast, G. Riedel).
46. Tekhnolog Constr. Techn. Bur., RU 21 02 415, 1994.
47. BASF, DE-OS 3011235, 1980 (K. Schmeidl).
48. BASF, DE-OS 37 13 618, 1987 (K. Schmeidl).
49. BASF, EP 341 325, 1988 (K. Schmeidl).
50. Bayer, DE-OS 3025557, 1980 (K. Linhart, H. Gleinig, R. Raue, H.-P. Kühlthau).
51. Bayer, DE-OS 33 03 512, 1983 (R. Raue).
52. Amer. Cyanamid, DE-AS 1072222, 1958.
53. Bayer, DE-AS 1044310, 1956.
54. VEB Bitterfeld, DE-OS 2038411, 1970.
55. Bayer, DE-OS 3048998, 1980 (K. Kunde).

56. Sandoz, DE-OS 2915323, 1979; 3030197, 1980 (R. Pedrazzi).
57. Sandoz, DE-OS 33 11 091, 1982 (J. Troesch, W. Portmann).
58. Sandoz, DE-OS 33 29 817, 1982 (J. Troesch, W. Portmann).
59. Sandoz, DE-OS 34 12 762, 1983 (J. Troesch).
60. Sandoz, DE-OS 35 03 844, 1984 (R. Pedrazzi).
61. Sandoz, EP 341 214, 1988 (H. A. Moser, R. Wald).
62. Bayer, DE-OS 3133360, 1981 (K. Kunde).
63. Nippon Chem. Works, JA 04 93 365, 1990.
64. BASF, EP 281 920, 1987 (J. Dehnert).
65. Bayer, EP 632 104, 1993 (K. Hassenrueck, P. Wild).
66. Bayer, EP 312 838, 1987 (A. Engel).
67. Intreprenderea "Sintofarm", RO 88 681, 1984.
68. BASF, DE-OS 42 02 566, 1992 (E. Hahn, H. Hengelsberg, U. Mayer).
69. Orient Chem. Works, JA 08 048 924, 1994.
70. Nippon Chem. Works, JA 04 183 754.
71. BASF, DE-OS 42 35 154, 1992 (U. Schloesser, U. Mayer).
72. BASF, EP 666 287, 1994 (U. Schloesser, U. Mayer).
73. Wella, EP 852 942, 1996 (U. Lenz, E. Nowald).
74. L'Oreal, WO 97/20 545, 1997 (H. Samain).
75. Sandoz, DE-OS 1644138, 1966.
76. Ciba, FR 1322766, 1961.
77. Bayer, DE-OS 32 31 398, 1982 (P. Wild, F.-M. Stoehr).
78. Bayer, DE-OS 3138182, 1981 (K. Kunde, P. Wild).
79. ICI, GB 951667, 1961.
80. Bayer, DE-AS 1011396, 1955 (W. Kruckenberg).
81. VEB Chem. Bitterfeld, DDR 227 148, 1984 (E. Guhl, G. Knochel, H. Noack, J. Schlykow).
82. VEB Chem. Bitterfeld, DDR 227 149, 1984 (R. Freitag, E. Guhl, G. Knochel, H. Noack, J. Schlykow).
83. Bayer, DE-OS 2850706, 1978 (K. Leverenz, J. Reppert).
84. Chemie Bitterfeld-Wolfen, DE-OS 41 28 490, 1991 (H. Hartenstein et al.).
85. Mitsubishi Chem., JA 59 170 147, 1983.
86. Sumitomo, JP 7313753, 1969.
87. Bayer, DE-OS 2024184, 1970; DE-OS 2036952, 1970; DE-OS 2036997, 1970; DE-OS 2041690, 1970; DE-OS 2065685, 1970 (W. Kruckenberg). ICI, DE-OS 2646967, 1975 (P. Gregory).
88. Sandoz, DE-OS 2338729, 1972 (B. Henzi).
89. Bayer, DE-OS 2059947, 1970.
90. Bayer, DE-OS 2309528, 1973; DE-OS 3030918, 1980 (K. Linhart, H. Gleinig, G. Boehmke, K. Breig); DE-OS 3101140, 1981 (K. Linhart, H. Gleinig, G. Boehmke).
91. Bayer, DE-OS 2508884, 1975; DE-OS 2631030, 1976 (W. Kruckenberg).
92. Ciba, DE-OS 2011428, 1970 (G. Hegar). Sandoz, CH 546269, 1964; CH 552032, 1964 (R. Entschel). Du Pont, US 3890257, 1973; US 3987022, 1973 (D. S. James).
93. Pentel K.K., JP 77101125, 1976.
94. Mitsubishi Kasei, JA 01 284 570, 1988.
95. Bayer, DE-OS 2533428, 1975 (G. Boehmke, U. Hendricks).
96. Sumitomo, JP 7237675, 1969.
97. Ciba-Geigy, EP 696 619, 1994 (A. Kaeser).
98. Sandoz, DE-OS 2118536, 1970; DE-OS 2712265, 1976; DE-OS 2741010, 1976; DE-OS 2805264, 1978 (M. Greve); BE 756820, 1970; CH 549629, 1971. Montedison, DE-OS 2851373, 1977 (F. Merlo).
99. ICI, DE-OS 2200270, 1971; DE-OS 2364592, 1973; DE-OS 2364593, 1972.
100. Sandoz, DE-OS 2313526, 1972 (W. Steinemann).
101. Ciba-Geigy, DE-OS 2429927, 1973 (V. Ramerathan).
102. Sandoz, DE-OS 31 17 127, 1980 (P. Doswald, E. Moriconi, H. Moser, H. Schmid).
103. Mitsubishi, JP 4957169, 1972; JP 4990327, 1972; JP 49109422, 1973; JP 5000179, 1973; JP 5003123, 1973; JP 5031181, 1973; JP 5246184 1973.
104. Ciba, BE 587048, 1959. GAF, FR 1295862, 1960.
105. I. Gmai et al., SU 287217, 1966.
106. Inst. f. Org. Ind., PL 64734, 1968.
107. Du Pont, DE-OS 2003540, 1969.
108. Ciba-Geigy, DE-OS 2509596, 1974.
109. Sandoz, DE-OS 20 61 964, 1969 (H. Moser, H. von Tobel).
110. Ciba-Geigy, EP 15 233, 1979 (P. Galafassi, J.-M. Adam, P. Loew, H. Scheidegger).
111. Ciba-Geigy, DE-OS 38 21 627, 1987 (W. Stingelin).
112. Ciba-Geigy, EP 294 330, 1987 (W. Stingelin).
113. Ciba, BE 672152, 1964, Amer. Aniline, US 3881866, 1972 (E. E. Renfrew). ACC, US 4204994, 1979; US 4211697, 1979 (K. A. Desai).
114. Mitsubishi Chem., JA 59 140 265, 1983.
115. Mitsubishi Chem., JA 61 151 269, 1984.
116. Mitsubishi Chem., JA 62 263 262, 1986.
117. Sandoz, FR 1325176, 1961.
118. Sandoz, CH 508709, 1969. Bayer, DE-OS 2042662, 1970.

119. Sandoz, DE-OS 2054697, 1969 (W. Steinemann); DE-OS 2752282, 1976 (M. Greve). ICI, DE-OS 2638051, 1975; DE-OS 2656879, 1975; DE-OS 2657147, 1975; DE-OS 2657149, 1975; DE-OS 2702777, 1976; DE-OS 2702778, 1976 (B. Parton).
120. TU Dresden, DDR 289 043, 1989 (P. Bellmann, K. Gewald, H. Schafer).
121. Sandoz, DE-OS 2627680, 1975 (M. Greve, H. Moser); CH 601433, 1976; FR 2360632, 1976 (H. Moser). Ciba-Geigy, DE-OS 3223436, 1981 (V. Ramanathan, W. Stingelin).
122. Hodogaya Chem. Ind., JA 03 192 158, 1989.
123. Nippon Kayaku, JA 01 283 269, 1988.
124. BASF, DE-OS 38 01 945, 1988 (H. Bruder, R. Dyllick-Brenzinger, U. Mayer, S. J. Bares).
125. ICI, DE-OS 2920487, 1978 (B. Parton).
126. ICI, EP 23770, 1979 (B. Parton, P. F. Gordon). BASF, DE-OS 2500064, 1975 (J. Dehnert, P. Miederer).
127. BASF, EP 482 508, 1990 (M. Ruske, U. Mayer).
128. BASF, DE-OS 19 809 994, 1997 (M. Ruske, P. Erk, H. Hengelsberg).
129. Ciba, BE 698730, 1966; BE 698917, 1966; EP 55221, 1980 (P. Loew). Yorckshire Dyeware Corp., GB 1047293, 1964; GB 1157867, 1967; GB 1349511, 1970 (J. F. Dawson, J. Schofield). ICI, GB 2009208, 1977 (M. G. Hutchings).
130. ICI, GB 1211078, 1967.
131. Yorckshire Dyeware Corp., GB 1233729, 1968.
132. GAF, DE-OS 2244459, 1971.
133. Yorckshire Dyeware Corp., GB 1380571, 1971.
134. ICI, GB 1247683, 1968.
135. ICI, DE-OS 2331953, 1972.
136. Yorckshire Dyeware Corp., GB 1361542, 1972.
137. Ciba-Geigy, DE-OS 2125911, 1970. ICI, DE-OS 3008899, 1979 (P. Gregory, D. Thorp).
138. BASF, DE-OS 36 11 228, 1986 (K. H. Etzbach, G. Hansen, H. Reichelt, E. Schefczik).
139. ICI, DE-OS 2548879, 1974; DE-OS 2703383, 1976; DE-OS 2817638, 1977 (P. Gregory). Sandoz, CH 627487, 1977 (U. Blass).
140. Ciba-Geigy, EP 514 336, 1991 (A. Nicopoulos, H. Birri, G. Hanika).
141. Sandoz, BE 668688, 1964. ICI, DE-OS 2817638, 1977; DE-OS 3008899, 1979; GB 2074598, 1980 (P. Gregory).
142. ICI, GB 1374801, 1971.
143. Ciba-Geigy, DE-OS 2154477, 1970.
144. BASF, DE-OS 2046785, 1970.
145. ICI, GB 1282281, 1967; GB 1290321, 1967.
146. Ciba-Geigy, CH 745871, 1974; CH 745872, 1974.
147. Ciba-Geigy, EP 16726, 1979 (J. M. Adam, P. Galafassi).
148. Ciba-Geigy, EP 306 452, 1987 (W. Stingelin).
149. Kuhlmann, BE 626882, 1962; DE-OS 2314406, 1972 (J. P. Stiot, C. Brouard).
150. Kuhlmann, FR 1364560, 1962.
151. Kuhlmann, FR 1364647, 1963.
152. Kuhlmann, FR 1389432, 1963; FR 1414876, 1964; DE-OS 2140511, 1970; DE-OS 2314406, 1972 (J. P. Stiot, C. Brouard).
153. Kuhlmann, FR 1467822, 1965.
154. Kuhlmann, FR 1403092, 1964. Geigy, FR 1322747, 1961.
155. Chemie Bitterfeld-Wolfen, DDR 293 604, 1990 (E. Dehmel et al.).
156. ICI, GB 1528801, 1975; GB 1531752, 1975 (B. Parton, F. L. Rose).
157. Sandoz, DE-OS 36 09 590, 1985 (H. A. Moser, R. Wald).
158. Sandoz, DE-OS 35 16 809, 1984 (J. Doré).
159. Sandoz, DE-OS 37 15 066, 1986 (H. A. Moser, R. Wald).
160. Sandoz, DE OS 33 40 483, 1982 (M. Greve, H. Moser).
161. Sandoz, DE-OS 33 11 294, 1983 (H. Moser, W. Samhaber, H. Schmid).
162. Sandoz, DE-OS 35 38 517, 1984 (H. A. Moser, R. Wald).
163. Sandoz, DE-OS 19 629 238, 1995 (J. Geiwiz, H. Moser, R. Pedrazzi, H. A. Moser).
164. Sandoz, DE-OS 39 32 566, 1988 (H. A. Moser, R. Wald).
165. Sandoz, DE-OS 19 500 203, 1994 (H. A. Moser).
166. Sandoz, DE-OS 44 39 004, 1994 (J.-P-Chavannes, G. Schöfberger, G. Scheulin).
167. Bayer, DE-OS 31 14 088, 1981 (F.-M. Stöhr, H. Nickel).
168. Sandoz, DE-OS 36 25 576, 1985 (J. Doré, R. Pedrazzi).
169. Bayer, DE-OS 31 33 568, 1981 (F.-M. Stöhr, H. Nickel).
170. Clariant, WO 97/35925, 1996 (R. Pedrazzi).
171. Ciba-Geigy, EP 122 458, 1983 (J.-M. Adam, H. Schwander).
172. Sandoz, DE-OS 2250676, 1971 (H. Moser).
173. Sterling Drug, DE-OS 2604699, 1975 (N. C. Crounse).
174. Bayer, EP 54616, 1980; EP 74589, 1981; DE-OS 3222965, 1982 (K. Kunde).
175. Bayer, DE-OS 35 02 927, 1985 (K. Kunde, N. Petzeit).

176. Sandoz, DE-OS 2915323, 1978 (R. Pedrazzi).
177. BASF, DE-OS 33 26 812, 1983 (H. Loeffler).
178. BASF, DE-OS 34 13 022, 1984 (H. Colberg).
179. BASF, DE OS 34 18 672, 1984 (H. Colberg, H. J. Degen, J. P. Dix).
180. BASF, DE-OS 35 02 693, 1985 (H. Colberg, E. Hahn).
181. Sandoz, DE-OS 3030196, 1980 (R. Pedrazzi).
182. Nippon Chem. Works, JA 04 183 755, 1985.
183. Ciba-Geigy, DE-OS 34 20 777, 1983 (V. Ramanathan).
184. Sterling Drug, US 4376729, 1980 (N. C. Crounse).
185. Hoechst, DE-OS 2028217, 1970; DE-OS 2057977, 1970.
186. Mitsubishi, JP 2181/66, 1963. ICI, GB 1516978, 1974 (P. Gregory). Bayer, DE-OS 3110223, 1981 (P. Wild, H. Nickel).
187. Mitsubishi, JP 4986424, 1972; JP 49116384, 1973.
188. Chemie Bitterfelden-Wolfen, DDR 296 501, 1990 (E. Forst, E. A. Jauer, R. Muehlberg, W. Mueller, H. Noack).
189. ICI, NL 6918341, 1968; DE-OS 2216206, 1971.
190. Hoechst, DE-OS 2347756, 1973.
191. Hoechst, FR 1601787, 1967.
192. Sandoz, CH 555876, 1971 (R. Entschel).
193. Bayer, EP 24322, 1979 (H. Nickel, F. Müller, P. Mummenhoff).
194. BASF, EP 324 415, 1988 (H. Bruder, U. Mayer).
195. Bayer, EP 24321, 1979 (H. Nickel, F. Müller, P. Mummenhoff).
196. Nippon Chem. Works, JA 05 132 629, 1991.
197. Warner-Jenkinson, DE-OS 19 752 665, 1996 (R. W. Pengilly, F. C. P. McLoughlin, R. G. J. C. Twigg).
198. Du Pont, DE-AS 1054616, 1954.
199. Bayer, DE-OS 2135152, 1971.
200. Ciba, DE-AS 1156188, 1959.
201. Du Pont, US 2965631, 1958.
202. Ciba, DE-OS 1956142, 1968. Sandoz, CH 553241, 1969 (R. Entschel, C. Müller, W. Steinemann).
203. Hoechst, DE-OS 2222099, 1972 (E. Fleckenstein, R. Mohr, E. Heinrich).
204. Ciba-Geigy, DE-OS 2263109, 1971.
205. ICI, GB 1414503, 1972 (D. B. Baird, J. L. Lang, D. F. Newton).
206. Bayer, DE-OS 2559736, 1975 (W. Kruckenberg).
207. Hodogaya, JP 7215473, 1969; JP 7215476, 1969.
208. Du Pont, DE-OS 2137548, 1970.
209. Sterling Drug, GB 1299080, 1968; JP 4702127,1970; US 3935182, 1973; US 3996282, 1974. Scott Paper, GB 2020703, 1978.
210. Bayer, DE-OS 3114075, 1981; DE-OS 3114087, 1981 (H. Nickel, P. Wild, F. M. Stöhr).
211. Bayer, DE-OS 3048694, 1980 (H. Nickel, P. Wild).
212. Mitsubishi, JP 49118721, 1973.
213. Ciba-Geigy, EP 263 073, 1986 (W. Stingelin).
214. Ciba-Geigy, EP 434 623, 1989 (W. Stingelin).
215. Eastman Kodak, US 3804823, 1972. Hoechst, DE-OS 2339713, 1973 (E. Fleckenstein).
216. Toyo Ink, JA 05 125 290, 1991.
217. Bayer, DE-OS 2315637, 1973 (K. L. Moritz, K.-H. Schündehütte). Sandoz, DE-OS 2120878, 1970; EP 41040, 1980; EP 93828, 1980; EP 93829, 1980 (P. Doswald, E. Moriconi, H. Moser, H. Schmid). Kuhlmann, DE-OS 2119745, 1970. Eastman Kodak, US 3836518, 1972 (G. T. Clark).
218. Bayer, DE-OS 2101999, 1971 (M. Wiesel, G. Wolfrum). Mitsubishi, JP 5043284, 1973.
219. Sandoz, DE-OS 3313965, 1982.
220. Sandoz, DE-OS 2251041, 1971 (H. Moser).
221. Sandoz, DE-OS 2120590, 1970; DE-OS 2120878, 1970.
222. Sandoz, CH 516622, 1968.
223. Bayer, DE-OS 1544529, 1965.
224. Hoechst, FR 1484099, 1965.
225. Hoechst, FR 1484503, 1965. Amer. Cyanamid, DE-AS 1155547, 1957; DE-AS 1162014, 1957.
226. IG Farbenind., DE 477913, 1927. ICI, DE-OS 2319090, 1972.
227. Ciba-Geigy, EP 284 567, 1987 (J. M. Adam).
228. ICI, DE-OS 2421822, 1973 (J. L. Leng, D. F. Newton); GB 2011454, 1977 (M. G. Hutchings). Ciba-Geigy, GB 1257346, 1968.
229. Kuhlmann, FR 1391676, 1963.
230. ICI, GB 1246632, 1967; GB 1365625, 1972 (J. S. Hunter, J. L. Leng, C. Morris).
231. Bayer, DE-OS 2141987, 1971 (M. Wiesel, G. Wolfrum).
232. GAF, DE-OS 2050246, 1969 (N. A. Doss).
233. Sandoz, CH 554917, 1969 (R. Entschel, C. Müller, W. Steinemann). Hodogaya, JP 7130110, 1968.
234. Ciba, DE-OS 1644110, 1966.
235. MLB Hoechst, DE 105202, 1898. Bayer, DE 272975, 1911. Geigy, DE 473 526, 1927.
236. Sandoz, DE-OS 37 17 869, 1986 (R. Pedrazzi).
237. Sandoz, DE-OS 33 40 486, 1982 (M. Greve, H. Moser).

238. Bayer, DE 181783, 1906.
239. Kuhlmann, FR 1380628, 1963. VEB Bitterfeld, DE-OS 2031561, 1970.
240. Hoechst, DE-AS 1242773, 1963; FR 2019355, 1968; DE-OS 2516687, 1975; DE-OS 2522174, 1975 (R. Mohr, E. Mundlos, K. Hohmann).
241. Hoechst, DE-OS 1816985, 1968.
242. Hoechst, BE 677795, 1965; DE-OS 2221989, 1972; DE-OS 2222042, 1972 (E. Mundlos, R. Mohr, K. Hohmann); DE-OS 2357448, 1973 (K. Hohmann); FR 2016982, 1968.
243. Kuhlmann, DE-OS 2154662, 1970.
244. BASF, DE-OS 2020479, 1970.
245. Leningrad Lensovet, Tech., RA 595309, 1975.
246. Eastman Kodak, US 3928309, 1974; US 3936435, 1974 (G. T. Clark).
247. Amer. Cyanamid, FR 1340397, 1961; DE-OS 2741395, 1976 (R. B. Balsley). Hodogaya, JP 56134280, 1980.
248. Ciba-Geigy, WO 95/15 144, 1993 (P. Moeckli).
249. L'Oreal, WO 99/20 235, 1997 (Ch. Rondeau).
250. L'Oreal, EP 920 856, 1997 (C. Deneulenaere).
251. Ciba-Geigy, EP 116 513, 1983 (V. Ramanathan, P. Moeckli).
252. BASF, DE-OS 35 29 968, 1985 (M. Ruske, H.-J. Degen).
253. Sandoz, DE-OS 33 13 965, 1982 (J. Doré, R. Pedrazzi).
254. Sandoz, DE-OS 34 40 777, 1983 (J. Doré, R. Pedrazzi).
255. Sandoz, DE-OS 34 16 117, 1983 (M. Greve, H. Moser, R. Pedrazzi, R. Wald).
256. Lexmark, GB 23 15 493, 1996 (J. F. Feeman, J. X. Sun).
257. Lexmark, GB 23 15 759, 1996 (B. L. Beach, J. F. Feeman, T. E. Franey, A. P. Holloway, J. M. Mrvos, J. X. Sun, A. K. Zimmer).
258. Sterling Drug, US 4379088, 1980.
259. Sterling Drug, US 4379089, 1980.
260. Sandoz, DE-OS 2403736, 1973 (H. Moser).
261. Ciba-Geigy, EP 84372, 1982 (J. M. Adam).
262. Bayer, EP 321 778, 1987 (G. Franke, P. Wild).
263. Ciba-Geigy, EP 14677, 1979 (J. M. Adam, V. Ramanathan, P. Galafassi).
264. Ciba-Geigy, EP 57159, 1981 (J. M. Adam).
265. Sandoz, DE-OS 1903058, 1969; DE-OS 1965994, 1968. Sterling Drug, DE-OS 2810246, 1977.
266. BASF, EP 34725, 1980 (M. Patsch, M. Ruske).
267. ICI, BE 601144, 1960.
268. CZ 182 658, 1976 (J. Kroupa, V. Chmatal, J. Dvorak).
269. ICI, BE 601145, 1960; BE 601146, 1960; DE-OS 2436138, 1973 (C. W. Greenhalgh, D. S. Phillips).
270. ICI, BE 603028, 1960. V. V. Kormachev, G. P. Pavlov, V. A. Kukhtin, R. S. Tsekhawski. *Zh. Obshch. Khim.* **49** (1979) no. 111, 112479.
271. Bayer, BE 661025, 1964.
272. G. B. Guise, I. W. Stapleton, *J. Soc. Dyers Colour.* **91** (1975) 259.
273. Bayer, BE 657179, 1963; BE 657302, 1963.
274. Bayer, FR 1459140, 1964.
275. K. Schimmelschmidt, H. Hoffmann, E. Baier, *Angew. Chem.* **74** (1962) 975.
276. G. Kaufmann, *Melliand Textilber.* **44** (1963) 1245.
277. Ch. Gränacker, H. Brüngger, F. Ackermann, *Helv. Chim. Acta* **24** (1941) 40 E.
278. D. A. W. Adams, *J. Soc. Dyers Colour.* **67** (1951) 223.
279. W. Kirst, W. Neumann, *Angew. Chem.* **66** (1954) 429.
280. W. Staab, *Melliand Textilber.* **42** (1961) 1373.
281. F. Gund, *J. Soc. Dyers Colour.* **76** (1960) 151.
282. Hoechst, DE 1086832, 1958 (U. Dreyer, W. Kirst); DE 1088060, 1959 (R. Mohr, H. Hertel); DE 1126545, 1959 (R. Gross, H. Hertel, W. Kirst, R. Mohr).
283. A. Rückert, *Melliand Textilber.* **41** (1960) 1541.
284. R. Gross, *Text. Prax.* **15** (1960) 1046.
285. W. Kirst, *Z. Gesamte Textilind.* **68** (1966) 772.
286. P. Heinisch, *Text. Prax.* **21** (1966) 342.
287. R. Löwenfeld, *Textilveredlung* **2** (1967) 677.
288. R. Löwenfeld, *Z. Gesamte Textilind.* **70** (1968) 618.
289. P. Frey, H. Hertel, *J. Soc. Dyers Colour.* **99** (1983) 286.
290. W. Seidenfaden, *Melliand Textilber.* **28** (1947) 432.
291. E. Fees, *Melliand Textilber.* **47** (1966) 1162, 1300.
292. Hoechst, DE 2850901, 1978 (E. Feess, P. Karacsonyi, H. Curtius).
293. Naphtol Chemie Offenbach, DE 848639, 1948 (W. Kirst).
294. R. Löwenfeld, *Melliand Textilber.* **40** (1959) 295.

Barium and Barium Compounds

Robert Kresse, Chemetall GmbH, Frankfurt am Main, Germany (Chap. 1)

Ulrich Baudis, Houghton Durferrit GmbH, Hanau, Germany (Chap. 1)

Paul Jäger, Kali-Chemie AG, Hannover, Germany (Chap. 2)

H. Hermann Riechers, Kali-Chemie AG, Hannover, Germany (Chap. 2)

Heinz Wagner, Kali-Chemie AG, Hannover, Germany (Chap. 2)

Jochen Winkler, Sachtleben Chemie, Duisburg, Germany (Chap. 2, 5, 3 and 4)

Hans Uwe Wolf, Universität Ulm, Ulm, Germany (Chap. 5)

1. Barium Metal 1
1.1. Physical Properties 2
1.2. Chemical Properties 2
1.3. Production 3
1.4. Chemical Analysis 3
1.5. Transportation, Storage, Safety . . . 4
1.6. Uses 4
1.7. Economic Aspects 4
2. Barium Compounds 4
2.1. Barium Sulfate 6
2.1.1. Properties 6
2.1.2. Barite 6
2.1.3. Blanc Fixe 8
2.2. Barium Sulfide and Polysulfides . . 9
2.3. Barium Carbonate 11
2.4. Barium Chlorate 13
2.5. Barium Chloride 13
2.6. Barium Ferrite 14
2.7. Barium Oxide 14
2.8. Barium Peroxide 15
2.9. Barium Hydroxide 15
2.10. Barium Nitrate 16
2.11. Barium Titanate 16
2.12. Barium Salts of Organic Acids . . . 16
3. Barium in the Environment 17
4. Dietary Intake of Barium 17
5. Toxicology 17
6. References 18

The history, production, analysis, and uses of barium and barium compounds are described. Barite (barium sulfate) is the most important barium compound. Most of its use is in drilling fluids for crude oil and natural gas exploration. Use of pure barium sulfate is, e.g., as a filler and extender and as a contrast agent in X-ray diagnosis. Barium carbonates are mainly used to remove sulfate ion impurities and to introduce barium oxide into special fine glassware such as cathode tube television screens. The production pathway of barium and barium compounds typically starts from barite. Volatile barium compounds impart a pale green color to flames, and this is an effective, simple qualitative test for barium.

1. Barium Metal

Barium belongs to group 2 of the periodic table and is one of the alkaline-earth metals. Of the elements barium is 19th in atomic abundance. Its average content in the earth's crust is estimated to be 250 g/t [10].

The soft, silver-colored metal is very reactive and does not occur free in nature. However, its compounds are widely distributed. Its major minerals are barium sulfate (barite, heavy spar) and barium carbonate (witherite). Barium sulfate occurs in large deposits throughout the world, and it is the most important raw material for all other barium chemicals. The total annual world production of barium compounds is estimated around 6×10^6 tons. The total world reserves are thought to be around 450×10^6 tons [11].

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_325.pub2

DAVY first prepared barium metal by electrolyzing mixtures of barium oxide and mercuric oxide and subsequently evaporating the mercury in the year 1808 [12]. He proposed the name barium. The principles of the reduction of the oxides of alkaline-earth metals with aluminum were described by MALLET [13] as early as 1878, but GUNTZ [14] is considered to be the first who obtained a considerable amount of barium of good purity by aluminothermic reduction of the oxide.

1.1. Physical Properties

Barium has a body-centered cubic crystal structure at standard temperature and pressure. It is a soft, silvery white metal of medium specific weight and good electrical conductivity (see Table 1). Ultrahigh-purity barium is described as slightly golden yellow [15]. The color changes into silvery white as soon as the surface is contaminated. It is not easy to obtain samples of ultrahigh purity, and therefore accurate measurements of some physical properties of barium metal are difficult to carry out. In fact, the values for some physical properties are still the subject of controversy.

1.2. Chemical Properties

Barium is clearly metallic and is strongly electropositive. In all of its compounds it has an oxidation state of +2. Barium is very reactive, and its compounds have high energies of formation. Barium forms stable salts with nearly all oxoacids. Table 2 summarizes the physicochemical properties that are associated with the chemical properties of the element.

Generally speaking, barium is more reactive than magnesium, calcium, or strontium. On heating, it readily combines with the halogens in strongly exothermic reactions. It also reacts with the chalcogens to form BaO, BaS, BaSe, and BaTe [16]. These compounds are also formed by reduction of the carbonate, sulfate, selenate, or tellurate [17, pp. 917–934]. Barium combines with oxygen even at room temperature. Therefore, the metal must be protected from oxidation by storage under paraffin oil or an inert gas. A series of phosphides can be isolated [18]. Reaction of barium with nitrogen starts at about 250 °C. The compound Ba_3N_2 is formed [17, pp. 917–934], [19].

Table 1. Physical properties of barium

Property	Value
Atomic number Z	56
Relative atomic mass A_r	137.3274
Mass number (natural abundance, %) of stable isotopes	130 (0.106), 132 (0.101), 134 (2.417), 135 (6.592), 136 (7.854), 137 (11.232), 138 (71.698)
Density ϱ at 20 °C	3.74 g/cm^3
Melting point *mp*	726.2 °C
Boiling point *bp* at 101.3 kPa	1637 °C
Hardness (Mohs scale)	1.25
Crystal structure	body-centered cubic
Lattice constant α_0 at 20 °C	0.5025 nm
Coefficient of linear thermal expansion α_1 (mean, 0–100 °C)	1.8×10^{-5} K^{-1}
Modulus of elasticity E	1.265×10^{10} N/m^2
Heat of fusion ΔH_m	7.98 kJ/mol
Heat of vaporization ΔH_v	140.3 kJ/mol
Specific heat capacity c	
at 20 °C	192 J kg^{-1} K^{-1}
at 900 °C	230 J kg^{-1} K^{-1}
Vapor pressure P at	
630 °C	1.33 Pa
730	1.33×10^1 Pa
860	1.33×10^2 Pa
1050	1.33×10^3 Pa
1300	1.33×10^4 Pa
1520	5.33×10^4 Pa
1637	1.013×10^5 Pa
Electrical resistivity ϱ	
Commercial purity	40×10^{-6} Ω cm
Extra high purity	30×10^{-6} Ω cm
Liquid barium at *mp*	314×10^{-6} Ω cm
Thermal coefficient of electrical resistivity $d\varrho/dT$ (mean, 0–100 °C)	6.5×10^{-3} K^{-1}
Thermal conductivity at 273.2 K	0.185 W/cm K

Table 2. Physicochemical properties of barium

Property	Value
Work function Φ	3.38×10^{-19} J or 2.11 eV
Ionization potential	
First	8.33×10^{-19} J or 5.212 eV
Second	1.60×10^{-18} J or 10.00 eV
Standard electrode potential	
$Ba^{2+} + 2e \rightarrow Ba$	− 2.91 V
Electronegativity (Pauling)	0.9
Ionic radius (Ba^{2+}, Pauling)	0.135 nm
Atomic radius	0.217 nm
Enthalpy of hydration (Ba^{2+})	1290 kJ/mol
Standard heats of formation (25 °C)	
BaO	− 558 kJ/mol
$BaCl_2$	− 858
BaH_2	− 191
Ba_3N_2	− 377

Silicon reacts with barium at 1150 °C to form the silicides BaSi, $BaSi_2$, and Ba_3Si_4 [20]. Barium combines with carbon to form the carbide BaC_2 [21]. At moderately elevated temperatures barium reacts with hydrogen to form the hydride BaH_2 [17, pp. 921–922]. Barium reacts vigorously with water, liberating hydrogen, to form the strong base barium hydroxide. The metal reacts similarly with alcohols to form the alkoxides and hydrogen. Barium metal is used for synthesis of barium dipivaloylmethanate as the basic compound for the preparation of high temperature superconducting films by the chemical vapor deposition technique [22]. Barium dissolves readily in liquid ammonia. The blue or black solutions contain amine complexes such as $Ba(NH_3)_6$. Evaporation of the solvent leaves residues of barium amide, $Ba(NH_2)_2$, imide, Ba(NH), and nitride, Ba_3N_2 [23, 24].

There is an increasing number of organobarium compounds. Alkyl- and arylbarium halogenides, RBaX, as well as dialkyl- and diarylbarium, R_2Ba, have been synthesized [25, 26]. For a comprehensive revue see [27].

Barium forms many alloys, and many binary phases have been described [28]. Nevertheless, only the system barium – aluminum has commercial importance.

1.3. Production

On an industrial scale barium metal is produced in a vacuum by reduction of its oxide with aluminum:

$$4\,BaO + 2\,Al \rightarrow 3\,Ba + BaAl_2O_4$$

The two-stage process is similar to that for calcium. The raw material is barium oxide (see Section 2.7).

The barium oxide is mixed with aluminum granules, and the mixture briquetted and charged into long tubular retorts of heat-resistant steel. These are evacuated and heated to ≈ 1100 °C at the end containing the charge, while the other end is kept cool. Molten aluminum and aluminum vapor react with the solid barium oxide, releasing barium vapor, which condenses in the cooler part of the apparatus [29], where it is collected and cast into chill molds under argon.

$BaAl_4$ is an intermediate [30], and the net reaction is actually the sum of three reactions:

$$3\,BaO + 14\,Al \rightarrow 3\,BaAl_4 + Al_2O_3$$

$$8\,BaO + BaAl_4 \rightarrow 7\,Ba + 2\,BaAl_2O_4$$

$$BaO + Al_2O_3 \rightarrow BaAl_2O_4$$

In the laboratory the aluminothermic reduction of barium oxide is also the most useful method of preparation [7, pp. 920–921], [23, 31 – 34]. It is also the method of choice for preparing of high-purity samples [35].

Silicon powder can be used to reduce barium oxide, but somewhat higher temperatures are required, 1200 °C in a vacuum [36]:

$$3\,BaO + Si \rightarrow 2\,Ba + BaSiO_3$$

Either no reduction or only partial reduction combined with alloying occurs when barium oxide is treated with magnesium, calcium, sodium, or potassium. Barium can be isolated by reaction of barium iodide with sodium, whereas the chloride does not react. Electrolytic methods are not useful for industrial production and are only rarely useful for laboratory synthesis because of the solubility of barium in its molten halides [37]. It is unlikely that barium has ever been isolated in sufficient purity by electrolysis, even though this method has been mentioned numerous times [2].

1.4. Chemical Analysis

Volatile barium compounds impart a pale green color to flames, and this is an effective, simple qualitative test for barium (455.4, 493.4, 553.6, and 611.1 nm).

Barium is separated from magnesium, strontium, and calcium by precipitation from a dilute solution in nitric or hydrochloric acid with a solution of potassium dichromate in aqueous acetic acid. Barium is determined gravimetrically by precipitation of the sulfate [38]; small quantities are determined spectrometrically [39].

The metallic impurities in commercial barium (Table 3) are determined by atomic absorption and flame emission spectroscopy. Trace impurities are best determined by inductively coupled emission spectroscopy (ICP) [41]. The carbon content in barium is determined by combustion; nitrogen, by the Kjeldahl method; and hydrogen, by vacuum hot extraction. Vacuum hot

extraction is not useful for analysis of oxygen. Neutron activation analysis based on the reaction

$$^{16}\mathrm{O}\,(\mathrm{n,p}) \rightarrow {}^{16}\mathrm{N}$$

is the recommended method for determination of oxygen in alkaline-earth metals [42].

Table 3. Chemical analysis of commercial barium [40]

Element		Percentage
Ba	(incl. Sr)	99.2 ± 0.30 %
Sr	max.	0.80
Ca	max.	0.25
Al	max.	0.06
C	max.	0.06
Mg	max.	0.02
N	max.	0.02
Fe	max.	0.02
Cl	max.	0.01
Li, Na, K	max.	0.01

1.5. Transportation, Storage, Safety

Barium metal is commercially available in bars up to 20 kg or in rods 22 mm in diameter and 400 mm in length. The rods can be cut into small pieces or extruded into wires.

Barium is packaged in airtight steel drums containing up to 100 kg of the metal under argon or paraffin oil. Smaller amounts (1–10 kg) are packaged in tin cans, and even smaller samples are packaged in hermetically sealed glass bottles [43, 44].

Barium is a flammable solid and cannot be mailed. If it comes into contact with water, there is always the danger of explosion on account of the liberated hydrogen. Therefore barium should always be stored in a dry, well ventilated place and every contact with moisture and air avoided. Protective glasses and safety gloves should be worn while handling barium. Burning barium can be extinguished with sand, aluminum oxide, etc.

Transport classification [40]:

- GGVE, GGVS, RID, ADR: class 4.3, fig. 11b
- IMDG-Code: class 4.3 UN-No. 1400 PG.II
- ICAO: class 4.3 UN-No. 1400 PG.II/Drill-Code 4W

1.6. Uses

The main use of barium and barium – aluminum alloys is as *getter* to remove the last traces of unwanted gases from television picture tubes, transmitter valves, and other vacuum tubes [45 – 51]. Barium and its aluminum alloys also are used in incandescent lamps [52, 53]. Barium is particularly well suited as a getter because it has a low vapor pressure at the working temperatures of the tubes and is very reactive toward undesired gases, such as oxygen, nitrogen, hydrogen, carbon dioxide, water vapor, even removing residues of inert gases by inclusion.

Many other uses of barium have been described in the literature, but they are minor. For example, barium has been used in bearing alloys [54]; in lead – tin soldering alloys, where it increases creep resistance [55]; in nickel alloys for spark plugs [56]; in alloys with calcium, silicon, aluminum, and manganese that are used as deoxidizers for high-grade steel [20, 57]; and as an inoculant for steel and cast iron [58, 59]. Barium has also been used as a modifying agent for silumin, instead of strontium or sodium, because barium also refines the structure of the eutectic aluminum – silicon alloy [60]. Barium is also a good reducing agent; however, it is not normally used for this purpose because of its high atomic mass and its relatively high price, one kilogram of the metal costing about € 130 ($ 150) in 1998.

1.7. Economic Aspects

The leading producer, Chemetall GmbH, Germany, covers the global demand of Ba and $BaAl_4$. Nothing is known about production in the GUS and China, although facilities doubtlessly exist.

It is difficult to predict the future demand for barium. However, the replacement of television picture tubes by liquid crystal displays, or similar systems, is likely to decrease demand and production over the long term.

2. Barium Compounds

History. The first mention of barium compounds dates from the year 1602 when V. Cas-

CIOROLUS, in Bologna, realized that barite heated in the presence of organic substances phosphoresces in the dark. Much later, after 1774, C. W. SCHEELE discovered barium oxide and described its properties. He noted that a precipitate formed when sulfuric acid was added to barium oxide dissolved in water.

J. G. GAHN recognized that this was a reaction of the oxide to form barite, i.e., that the oxides were a new class of compounds. At that time the oxide was given the name baryta ($\beta\alpha\varrho\acute{\upsilon}\varsigma$ = heavy), which in some countries is still the name for technical grades: caustic baryta for the oxide, hydrate of baryta for the hydroxide. Today, however, the mineral heavy spar, $BaSO_4$, is meant by baryta or barite. Barium carbonate also occurs in nature. Native barium carbonate was named witherite, for W. WITHERING discovered the mineral in Cumberland, Great Britain, in 1783.

Raw Materials and Processing. The most important barium mineral is barite (barytes, heavy spar, barium sulfate, $BaSO_4$), today almost exclusively used as the raw material for the production of barium compounds. Barite occurs all over the earth in hydrothermal vein and sedimentary deposits. The main deposits of the mineral witherite ($BaCO_3$) are situated in England, Romania, and the USSR, but witherite is no longer of economic importance. No other mineral containing barium is used as a raw material for barium or its compounds at present.

The maximum world output of barite was $\approx 8.3 \times 10^6$ t (1981); however, only 7–8 % of this was used for the production of other barium compounds. Figure 1 shows how dressed barite is converted into other barium compounds. The barite is first reduced with carbon to give water-soluble barium sulfide, which is used mainly as an intermediate for the production of all other barium compounds and lithopone [61 – 65].

Chemical Analysis. All barium compounds may be identified by the green color imparted to flames. For quantitative analysis, atomic absorption spectroscopy (AAS), flame emission spectroscopy (OES-ICP), X-ray spectroscopy, complexometry, ion chromatography (IC), and gravimetry are suitable. (For chemical analysis of barium metal see Section 1.4.)

Economic Aspects [66, 67]**.** Production data is available only for barium sulfate, the most important barium compound and the starting material for the production of most other barium compounds. The worldwide barite production has continued to increase from about 5.3 $\times 10^6$ t in 1996 to 5.8×10^6 t in 2000, 6.5×10^6 t in 2003, 7.3×10^6 t in 2004 up to about 7.6×10^6 t in 2005. Since the 1990s, China has been the main producer of barites and is continuing to strengthen its position. Its share of the market has increased from 30 % in 1990 to more than 50 % in 2005. Other significant producers are India (14 %) and the USA (7 %), followed by Marocco (4 %), Mexico (4 %), and Iran (3 %). Barite production primarily reflects worldwide crude oil and natural gas exploration, since at least about 80 % of its production is used in drilling fluids for these industries. A large portion of this type of barites is used in the oil industry of the Golf of Mexico. Consumption and the source of barites is mostly dominated by freight costs. In the Golf of Mexico, for example, most of the barites is imported from China in the form of lumps that are then processed in grinding mills at various locations on the Golf coast. American sources, mostly from Nevada, are primarily used in Canada and the western and north-western parts of the USA. Indian barite is most competitive in the Middle East and in local markets. Due to the increasing oil demand and the phasing out of Chinese VAT rebates, the prices for barites have continued to rise. Increasing freight rates due to globalization of the markets have assisted in this development. Depending upon the price of the barites, haematite may be a possible replacement, depending upon freight rates. As of today, awareness for environmental issues is growing, such as the mercury contents of the barites used.

The other important uses of barite are as a starting material for the production of barium compounds (see Fig. 1) and its direct use in bleached or unbleached form as a filler in paints and plastics. Consumption of barite in drilling muds varies considerably, while the other two markets are less subject to change. Therefore, it is difficult to obtain reliable relative percentage data. An estimated balance for 1995 is 76–80 % for oil development, 9–13 % for industrial use, and 5–9 % for direct use as a filler.

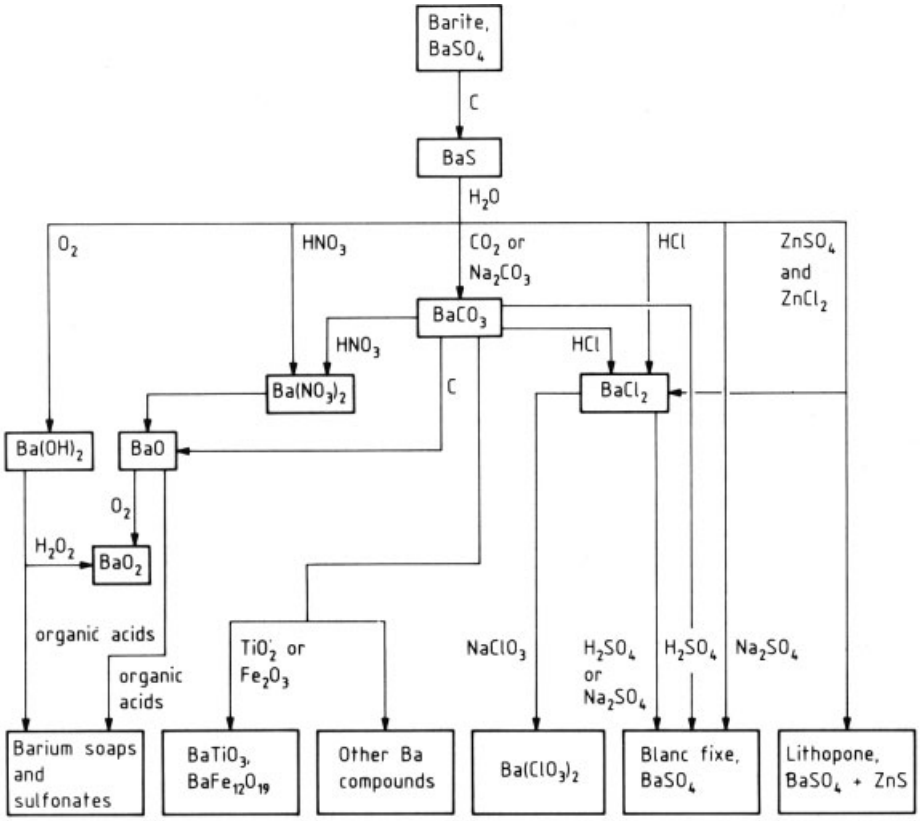

Figure 1. Production pathways to various barium compounds starting from barite

Identified barite reserves amount to 170×10^6 t with prospects for a total of 500×10^6 t worldwide.

2.1. Barium Sulfate

Barium sulfate is the most important barium compound, and the mineral barite, $BaSO_4$, also called heavy spar or barytes, is by far the most important barium mineral.

2.1.1. Properties

Barium sulfate [*13462-86-7*], $BaSO_4$, M_r 233.39. The industrially precipitated product is called blanc fixe [*7727-43-7*], which is French for "permanent white". Pure $BaSO_4$ forms colorless rhombic crystals of refractive index $\approx$ 1.64, hardness 3–3.5 (Mohs scale), and density ϱ 4.48 g/cm^3. Heating $BaSO_4$ decomposes it, the rate increasing with temperature over the range 1100–1500 °C. The main products are BaO, SO_2, and O_2. For this reason melting points given in the older literature — 1345 °C or 1580 °C — differ. Barium sulfate is almost insoluble in water, the solubility being 2.5×10^{-4} wt % at 20 °C and 3.9×10^{-4} wt % at 100 °C. The solubility of barium sulfate is enhanced by the presence of chelating compounds. For example, it is dissolved by aqueous ethylenediamine tetraacetic acid (EDTA). Barium sulfate is more soluble in hot concentrated sulfuric acid and in melts of alkali-metal salts. The most important reaction of barite is its reduction with carbon to produce barium sulfide.

2.1.2. Barite

In nature, barite [*13462-86-7*] rarely occurs as a pure white mineral; generally, its color is yellowish, pink, reddish brown, dark gray, or black.

The main impurities are quartz, calcium carbonate, iron and manganese oxides, fluorite, strontium sulfate, pyrite, lead glance, zinc blende, and sometimes bituminous substances, the impurities and their amounts depending on the particular deposit. To remove these impurities before the barium sulfate is used industrially, in most cases barite is given a preliminary treatment to increase the $BaSO_4$ content to 90–97 wt %.

Deposits. Barite deposits are found in all regions of the world. The most important sources in Europe are in Germany, France, Italy, the UK, Belgium, and Spain, with minor amounts coming from Portugal and Greece. Kasakhstan, Bulgaria, Romania, and, to a lesser extent, Georgia, Slovakia, and Poland supply the Central and Eastern European markets.

In the Americas, significant deposits are found in the United States, Canada, Mexico, Brazil, and Argentina.

Morocco is the most important producer in Africa, followed by Algeria and Tunesia.

The largest producer of barite is China (appr. 3.8×10^6 t/a). Other major barite producers in the middle and Far East are India, Iran, and Turkey, with further amounts coming from Indonesia, North Korea, and Thailand, and lesser amounts from Pakistan, Malaysia, Japan, and Australia.

Mining and Dressing of the Ore. Barite is usually mined in open cuts, only sometimes in open pits or underground galleries. The ore is washed in trommels to remove adherent argillaceous impurities, then crushed in breakers, and classified. The various grain sizes are liberated from the mostly quartz gangue in jigs, thus enriching the barium sulfate. This wet mechanical dressing is adequate for many uses. If there is a high degree of intergrowth of quartz and barite and if the ore contains much iron, lead, or zinc, flotation is required. This produces 98 wt % $BaSO_4$ [69]. Sometimes barite is a by-product of the flotation of pyritiferous ores. The flotation agents and the pH depend, among other things, on the impurities present in the crude ore [70]. Generally the barite is wet ground to a grain size of 0.1 mm before flotation. Sodium oleate, tall oil, oleic acid, and alkyl sulfonates and sulfates have been used as collectors. Water glass is the preferred depressant, especially for quartz and iron oxides. Flotation processes to depress the $BaSO_4$ and float the impurities are confined to a few special cases, e.g., ore rich in fluorite [71].

Another method applicable only to special barite types is based on the tendency of these barites to decrepitate. The broken ore is heated to ≈ 700 °C in rotary kilns, the barite crystals bursting while the gangue is unaffected. The decrepitated fines rich in barite are screened off from the gangue. The yield depends above all else on the suitability of the barite for this treatment.

Barite grades with special properties are produced for the various industrial uses.

Drilling Muds. About 80 % of the world output of barite is ground and used to increase the density of drilling muds, especially for drilling oil and gas wells. The muds consist of an aqueous clay suspension and additives. The large amount of barite that can be added to this suspension allows adjustment of the density to a high value without the viscosity being affected significantly. The increased density allows faster discharge of cuttings and prevents blowouts. The standards for drilling-mud-grade barite vary from country to country; however, the OCMA or API standards widely accepted [72]. These regulate the viscosities of aqueous suspensions, the density of the barite (related to the barium sulfate content), its particle size distribution, and the content of water-soluble materials. Higher contents of water-soluble materials, especially divalent cations such as calcium, can cause flocculation of the clay components of drilling muds [73].

For the production of barite for drilling muds, generally it is sufficient to jig the ore to increase the density by removing the quartz. In most cases the other impurities do not cause trouble. On the other hand, barite that has been floated must be heated or treated in some other way to destroy its hydrophobic character.

Raw Material for Barium Chemicals. The 7–8 % of the total barite production used as a raw material for other barium compounds is always reduced to barium sulfide as the first stage. The $BaSO_4$ content must be $\geqq 95$ wt %, and the $BaSO_4$ should contain as little silica and iron as possible, as these impurities reduce the

yield of water-soluble barium sulfide [*21109-95-5*]. Fluorspar (fluorite) and lead glance interfere with the reduction. Such barite is generally shipped in bulk as coarsely broken material, either granular (jigged) or ground (floated).

Mined Barites [66]**.** Mined barites may be separated into "white" and "off-white" grades. The whiteness of a barite is a measure for its purity. Off-white grades are used as a source for barium chemicals and directly as off-white fillers.

Off-white Barite Filler. The use of off-white barites is more or less limited to sound-deadening applications and to dark formulations, such as automobile underbody sealants. The sound-deadening action is related to the high density of barium sulfate.

White Ground Barite Filler. Ground barite is used as a filler owing to its resistance to chemicals and weathering, its wettability and dispersibility, and its low adsorption of binders. The properties of barite flour vary, depending on the commercial grade:

Grain size	0.1–40 µm
Brightness (reference, $BaSO_4$ white standard = 100)	85–95
Refractive index n_D	$\approx$ 1.64
Oil requirement	10–12 g of oil/100 g
Bulk density, tamped	1.4–2.2 g/cm^3
Density	4.0–4.25 g/cm^3

Some white barite grades may be ground directly without further treatment. More usual, however, is to subject the spar to coarse crushing and gravity dressing. Then the granular material is treated in bleaching vats with hot dilute acid, preferably sulfuric or hydrochloric acid, to remove the colored iron or manganese oxides, washed, and dried. The bleached spar is dry-milled in edge or ball mills to obtain the desired grain size, then air classified.

There is a more economical procedure: the coarsely crushed, gravity dressed barite is ground to the desired grain size by continuous wet milling in lined ceramic tubes. The screened suspension of ground material is bleached in agitator vessels, with dilute sulfuric or hydrochloric acid along with reducing agents (e.g., sodium sulfite) or oxidizing agents (e.g., nitric acid), depending on the kind of impurities, then washed, decanted, and dried. If floated barite is used in this process, then the barite must be floated again after bleaching to remove the decomposed organic material [74].

A third bleaching procedure lacks economic importance: here the barite is heated in a rotary kiln by adding alkali-metal salts, e.g., sodium chloride or sodium hydrogensulfate. This procedure is confined to special cases due to its high cost.

The specifications for ground barite are given in DIN 55911 for the Federal Republic of Germany and ISO 3262 internationally. The requirements of the buyers differ in regard to grain size and whiteness. Some products, such as EWO-Albaryt grades of the Deutsche Baryt-Industrie, Bad Lauterberg/Harz, are equivalent to synthetic blanc fixe in many respects. White ground barite is generally shipped in 25 or 50 kg paper bags, 1000 kg big bags, or in silo trucks.

Barite flours are mainly used as fillers and extenders in paints, varnishes, plastics, rubber, linoleum, sealing compounds, stoppers, plasters, and adhesives (5–9 % of the total barite production). High-quality grades can be used as an X-ray-opaque medium in medicine, provided they meet the requirements of the individual countries as given in USP, EP, JP, DAB, etc.

2.1.3. Blanc Fixe

Precipitated barium sulfate is usually called blanc fixe [*7727-43-7*]. It is mainly used as a filler and extender, in some fields also as a white pigment. The most important properties for the consumer are its insolubility in water and organic binders, its high degree of whiteness, and its homogeneous granulation:

Grain size	0.1–20 µm
Average grain size	0.2–4 µm
Brightness (reference, $BaSO_4$ white standard = 100)	95–99
Refractive index n_D	$\approx$ 1.65
Oil requirement	12–22 g of oil/100 g
Bulk density, tamped	0.9–2.2 g/cm^3
Density	4.0–4.4 g/cm^3

Certified ultrahigh-purity blanc fixe, which must also fulfill certain handling requirements, is used as a white standard in UV – Vis spectroscopy instead of magnesium oxide, which is no longer user for this purpose.

Production. Blanc fixe is usually precipitated in stirred tanks from solutions of barium chloride or barium sulfide by adding dilute sulfuric acid or sodium sulfate solutions. The sulfuric acid or the sodium sulfate solution must be free from heavy metals that can impart color to the product.

The method of manufacture and the reaction conditions depend on the ultimate use of the blanc fixe, especially the desired grain size. For example, fine-grained blanc fixe is produced by rapid precipitation from highly concentrated solutions at high pH and low temperatures. Coarse blanc fixe is produced from dilute solutions at high temperatures, say 80–100 °C. The use of seed crystals and prolonged residence time, e.g., by stirring for an extended period at low pH, produces a larger average grain size in the finished product. The precipitated blanc fixe is filtered or centrifuged and washed. It is drawn off and packed as paste, slurry, or powder, the last involving an additional process step of drying and perhaps one of regrinding.

A special problem in preparing blanc fixe from barium sulfide and sodium sulfate solutions is washing out the sodium sulfide. However, this problem can be overcome by enlarging the particle size of the extremely fine material by sintering (at 300–1000 °C) with fluxes like sodium sulfate and extracting the clinker [75].

Another procedure for the manufacture of blanc fixe is based on the conversion of an aqueous suspension of barium carbonate with dilute sulfuric acid, sometimes in the presence of small amounts of hydrochloric acid. The resulting product is characterized by its whiteness [76].

Formerly blanc fixe was obtained as a byproduct of hydrogen peroxide production by reaction of barium peroxide and sulfuric acid, but today hydrogen peroxide is prepared by the anthraquinone process (see → Hydrogen Peroxide).

Quality Specifications. In the Federal Republic of Germany, the standards for blanc fixe, the methods of analysis, and tests are defined in DIN 55 911. The various grades differ mainly in grain size. The smaller the grain diameter the higher the luster. In fact, grades are still sometimes characterized as mat, semimat, or lustrous.

The standards for blanc fixe to be used as X-ray-opaque medium in medicine are described and defined for the various countries in USP, EP, JP, DAB, etc. The standards deviate slightly from one to another.

Shipping and Packing. Blanc fixe is shipped as powder (usually in 25 kg paper bags, big bags, or in silo trucks); slurry (30–40 % water in tank trucks); and agglomerate-free paste (25–35 % water in 50 kg polyethylene-lined sacks). Because blanc fixe is nontoxic, it is not classified as a dangerous good.

Uses. About 70 % of the blanc fixe consumed in Western Europe is used in paint and varnish as well as in printing ink as a filler. It is used as a coating pigment to upgrade the surface of paper and is sometimes used as a pulp filler. Photographic paper requires grades free from sulfur and heavy metals, but this market has lost significance. Today, anatase-grade TiO_2 is used instead. Blanc fixe is also used as a filler in plastics, rubber, and adhesives and as an additive in the negative plates of lead storage batteries. Special grades are used as a medium opaque to X-rays for medical purposes (stomach X-rays).

Nanoscaled barium sulfates have proven useful in improving mechanical properties of certain polymers such as epoxy systems. Similar to nanoscaled titanium dioxides, they are able to improve the hardness as well as the fracture toughness of these polymers [77]. Because of the low refractive index of barium sulfate, it is possible to formulate clear coats with a high transparency even with large amounts of the nanoparticles. These films have an enhanced resistance to chemicals and improved scratch resistance.

2.2. Barium Sulfide and Polysulfides

The principal intermediate between barium sulfate and other barium compounds is barium sulfide.

Properties. Pure barium sulfide [*21109-95-5*], BaS, M_r 169.40, ϱ 4.3 g/cm^3, forms colorless cubic crystals. It melts above 1200 °C. Its heat capacity c_p is 287 J kg^{-1} K^{-1}. At room temperature barium sulfide is unaffected by dry air, but over the range 500–1000 °C it is oxidized to barium sulfate, at an ever-increasing rate. In humid

air containing carbon dioxide, some barium carbonate forms, and the odor of hydrogen sulfide is noticeable. The solubility of barium sulfide in water depends strongly on the temperature, the maximum being near 90 °C:

°C	0	20	40	60	80	90	100
wt % BaS	2.8	7.3	13.0	21.7	33.3	40.2	37.6

Owing to complete hydrolysis, a solution of sulfide is actually an equimolar solution of barium hydroxide [*17194-00-2*], $Ba(OH)_2$, and barium hydrosulfide [*25417-81-6*], $Ba(SH)_2$. Accordingly, barium sulfide does not precipitate when the solution is cooled; instead, a double salt of composition $Ba(OH)_2 \cdot Ba(SH)_2 \cdot 10\ H_2O$ [*42821-46-5*] precipitates. Excess hydrogen sulfide converts the barium hydroxide in solution quantitatively to barium hydrosulfide, which can be precipitated as $Ba(SH)_2 \cdot 4\ H_2O$ [*12230-74-9*] by addition of ethyl alcohol.

Aqueous barium sulfide reacts with carbon dioxide to form barium carbonate [*513-77-9*], $BaCO_3$, and the solution is also susceptible to atmospheric oxygen, which easily oxidizes the hydrosulfide to polysulfide, thiosulfate, and sulfite.

Production. Barium sulfide is produced by reducing barite ($BaSO_4$) with carbon. The net reaction is

$$BaSO_4 + 2\ C \rightarrow BaS + 2\ CO_2$$

However, the thermodynamic data suggest that the actual reducing agent is carbon monoxide [78]:

$$BaSO_4 + 4\ CO \rightarrow BaS + 4\ CO_2$$

Barium sulfite, $BaSO_3$, is an intermediate [79]. The CO forms in the reaction between carbon and CO_2, the Boudouard equilibrium:

$$C + CO_2 \rightleftharpoons 2\ CO$$

The barite should have a $BaSO_4$ content $\geqq 95$ wt %. Large amounts of impurities, especially silica and iron oxide, reduce the yield of water-soluble barium sulfide by forming the insoluble barium silicates and ferrates. The spar is ground to 0.1–2 mm and mixed with low-ash reducing carbon or coke, one part of carbon to 4–6 parts of barite.

Today the reduction is usually carried out in a continuous process centered around rotary kilns up to 50 m long. A burner at the outlet of the kiln fired with fuel oil, natural gas, or coal dust provides countercurrent heating. Reduction takes place at 900–1200 °C; average residence time is 90–120 min. The operating conditions must be carefully controlled to prevent secondary reactions with the gangue and to impede caking. The barium sulfide melt should be obtained in a fine-grained condition. The sulfide melt leaves the kiln continuously through a cooling drum. Generally further grinding is not necessary.

Various processes using a fluidized-bed kiln have been reported. Here, the pulverized spar, e.g., from flotation, is granulated together with carbon and reacted in a fluidized bed, a part of the carbon being used for heat generation and a part for reduction [79, 80]. Other processes use a reducing gas, e.g., carbon monoxide or methane, instead of carbon. The gas serves three functions: it provides heat upon combustion, it supports the fluidization, and it is the reducing agent. The advantages are a reduction temperature below 1000 °C, a residence time less than 1 h, and a more favorable energy balance [81, 82]. However, large-scale plants of this kind have not been reported.

Crude barium sulfide (black ash) is *leached* countercurrently with 60–80 °C water to dissolve the barium sulfide. This leaching is carried out batchwise in batteries of Shank leaching vats or continuously in closed agitated vessels followed by several thickeners. The suspended matter is filtered off, and the lye, containing 13–18 % BaS, is immediately treated further to prevent oxidation or crystallization. At this point the barium sulfide solution is very pure.

The residue still contains 50 % barium, in the form of compounds insoluble in water, such as silicates and ferrates of barium and unreacted barite. The silicates and ferrates can be dissolved with hydrochloric acid to form barium chloride; however, this process is often uneconomical because of the cost of purifying the barium chloride solution. To avoid these problems it has been proposed that the leaching residue be resuspended in water and, by adding hydrogen sulfide, the portion of barium compounds that are insoluble in water be converted into readily soluble barium hydrosulfide that can be led back into the sulfide lye [83].

Quality Specifications. Barium sulfide is available only as a technical-grade, granular, dark gray crude product called black ash. It contains 75–85 % water-soluble BaS, barium compounds soluble in hydrochloric acid (8–10 % as BaO), 2–4 % SiO_2, 2–3 % Fe_2O_3 and Al_2O_3, 2–4 % $BaSO_4$, and 1–2 % carbon. The economic value depends directly on the content of water-soluble BaS, which is determined analytically by acidimetry and iodometry.

Packing and Shipping. Barium sulfide is packed in 25 kg paper bags with an inner polyethylene bag. Shipment is governed by IMDG Class 6.1, UN 1564, and RID/ADR Class 6.1, 60c.

Uses. Barium sulfide is mainly used for the production of other barium compounds, especially blanc fixe and lithopone. Some BaS is used for the production of hydrogen sulfide; in this case, barium chloride is the byproduct. Other uses (depilatories, purification of mineral acids) are far less important.

Barium Polysulfides. The polysulfides [*50864-67-0*] have the general composition BaS_x (x = 2–5). They include barium disulfide [*12230-99-8*], BaS_2; barium trisulfide [*12231-01-5*], BaS_3; barium tetrasulfide [*12009-34-6*], BaS_4; barium tetrasulfide monohydrate [*12448-67-8*], $BaS_4 \cdot H_2O$; barium tetrasulfide dihydrate [*12231-03-7*], $BaS_4 \cdot 2\,H_2O$; and barium pentasulfide [*12448-68-9*], BaS_5. Polysulfides form when barium sulfide solutions are oxidized by air, when sulfur is dissolved in barium sulfide solutions, when barium sulfide is fused with sulfur in a vacuum, etc.

The isolation of individual polysulfides from solution is difficult because of their tendency to decompose. Technical barium polysulfide can also be produced by grinding barium sulfide with sulfur (Ba : S = 1 : 4–5). It is marketed as a grayish-yellow powder. The formation of polysulfide does not take place until the powder is dissolved.

One process, however, produces a mixture of true barium polysulfides, BaS_{4-5}, by sudden evaporation of a solution in a spray dryer [84]. The resulting light yellow powder is stable and dissolves in water without leaving a residue.

As a 1–3 % solution, barium polysulfide is a fungicidal and acaricidal agent used in fruit and grape growing.

Shipment of barium polysulfides is governed by the IMDG Class 6.1, UN 1564, or RID/ADR Class 6.1, 60c.

2.3. Barium Carbonate

Barium carbonate is the most important manufactured barium compound after blanc fixe. It occurs in nature as witherite, which has little economic significance due to rareness, inevitable impurities, and almost fully depleted deposits.

Properties. Barium carbonate [*513-77-9*], $BaCO_3$, M_r 197.35, *mp* 1360 °C accompanied by loss of CO_2. Pure barium carbonate is a white fine crystalline powder having several modifications:

$$\gamma\text{, rhombic} \xrightleftharpoons{803\,^\circ\mathrm{C}} \beta\text{, hexagonal} \xrightleftharpoons{980\,^\circ\mathrm{C}} \alpha\text{, cubic}$$

The rhombic modification is isomorphous with the carbonates of calcium (aragonite), strontium (strontianite), and lead (cerussite).

Barium carbonate has a density of 4.29 g/cm^3. Its solubility in water at 20 °C is only 2 $\times 10^{-3}$ wt %, although it is a little more soluble in water containing CO_2. Its refractive index is $\approx$ 1.6. Its heat capacity c_p at room temperature is 433 J kg^{-1} K^{-1}.

When barium carbonate is heated, it loses carbon dioxide and forms barium oxide. The dissociation is promoted by traces of water. Barium carbonate and its decomposition product barium oxide are partially soluble in each other at elevated temperatures, at 1000 °C $\approx$ 10 mol % oxide in the carbonate and $\approx$ 5 mol % of carbonate in the oxide — this partial solubility was not adequately appreciated in the older literature. Moreover, there is a eutectic at 1060 °C with 34 mol % barium oxide. The equilibrium CO_2 pressure above barium carbonate saturated with oxide increases rapidly above 1000 °C [85, 86]:

°C	900	1000	1100	1200	1250	1300	1360
kPa	0	0.21	1.6	10.1	21.6	46.4	101.3
Torr	0	1.6	12	76	162	348	760

Production. Barium carbonate is produced almost exclusively by precipitation from barium sulfide solutions. There are two processes:

1) precipitation with carbon dioxide
2) precipitation with soda

The first is based on the reactions

$$2\,BaS + CO_2 + H_2O \rightarrow BaCO_3 + Ba(HS)_2$$

$$Ba(HS)_2 + CO_2 + H_2O \rightarrow BaCO_3 + 2\,H_2S$$

The carbon dioxide is usually available from combustion gases or native carbonic acid. The precipitation is carried out batchwise in gastight vessels, several of which may be arranged in series to absorb the carbon dioxide that breaks through with the hydrogen sulfide in fresh lye; a utilizable hydrogen sulfide poor in carbon dioxide is produced in this way. Also there are continuous processes, e.g., using precipitation columns [87, 88].

The second process is based on the reaction

$$BaS + Na_2CO_3 \rightarrow BaCO_3 + Na_2S$$

The barium sulfide lye containing 15–18 wt % BaS is run into ≈ 30 wt % soda solution in agitator vessels.

The decision to favor one or the other of the two processes does not depend only on the availability and relative cost of carbon dioxide or soda. Above all else the value or sales prospects of the byproducts, H_2S or Na_2S, is decisive. The hydrogen sulfide must be worked up to sulfur, sulfuric acid, sodium hydrosulfide, or sodium sulfide. The dilute sodium sulfide solution produced in the soda process in most cases is upgraded to crystalline sodium sulfide hydrate (60 wt % Na_2S).

The main impurity in $BaCO_3$, sulfur as sulfide, elemental sulfur, or sulfate, is the major problem. Patents offer a number of solutions: adding aluminum powder to the barium sulfide lye to reduce the polysulfide sulfur [89]; continuously maintaining a residual barium sulfide content in the solution of ≈ 0.5 wt % during the precipitation with carbon dioxide to keep the sulfur dissolved [90]; or a careful step-by-step precipitation, perhaps first to $BaCO_3$ and $Ba(HS)_2$ and then to $BaCO_3$ and H_2S [87, 88].

Quality Specifications. Purity, bulk density, and grain-size distribution depend on the particular use. Usually, low sulfur and heavy-metal content is demanded. The manufacturers of electroceramics require low strontium content. The numbers below refer to products of various origin, quality, and process and should only be regarded as reference values:

Barium carbonate	≧ 98.5 %
Sulfur	0.08 (0.02) %
Strontium	0.5–1.4 (0.3) %
Iron	$6–20 \times 10^{-4}$ %
Insoluble in $HCl_{(soln)}$	0.1–0.6 %
Density	4.29 g/cm^3
Bulk density, tamped	0.4–1.8 (2.2) g/cm^3
Grain size, < 5 µm	80–95 %

An important standard for barium carbonate as a sulfate-binding agent is its reactivity. This is determined by conversion of barium carbonate with known quantities of sulfate, as gypsum water or sodium sulfate solution.

Packing and Shipping. Barium carbonate is stored in silos and shipped in multilayer paper bags, big bags, or silo wagons. Shipment is governed by the IMDG Class 6.1, UN 1564, or RID/ADR Class 6.1, 60c.

Uses. Barium carbonate is used to remove the detrimental sulfate ion from the input solutions of the chlor-alkali electrolysis, e.g., the mercury process, as well as from industrial waters of various kinds. In brick production, barium carbonate is a binder for the soluble sulfates in the raw mixture. The formation of insoluble barium sulfate prevents discoloration (efflorescence) during production. Moreover, barium carbonate reduces the tendency of baked bodies to effloresce on exposure to moisture. Barium carbonate is used to incorporate barium oxide into special fine glassware, apparatus glass, and many optical glasses. The barium oxide increases the workability of the glass melt, improves the mechanical and chemical resistance, and increases the density and refractive index of the glass.

Barium in television screens prevents the leakage of X rays. In the western world, the demand for cathode ray tube (CRT) televisions is declining since LCD (liquid crystal displays) and plasma screens are increasingly being used. As CRT televisions are, however, much cheaper,

they are still favored in India and China, where most CRT production is still based.

Barium carbonate is also used in enamels and for the production of the electroceramic barium titanate [*12047-27-7*] and the magnetoceramic barium ferrites [*12047-11-9*]. These last two uses require high-purity $BaCO_3$. Barium carbonate is also used in the synthesis of high-temperature superconductors.

2.4. Barium Chlorate

Barium chlorate monohydrate [*10294-38-9*], $Ba(ClO_3)_2 \cdot H_2O$, M_r 322.26, forms monoclinic short prismatic crystals of density 3.18 g/cm^3. Its solubility in water is 21.2 wt % at 10 °C and 51.3 wt % at 100 °C. Near 120 °C barium chlorate begins to lose its water of crystallization, and above 250 °C oxygen is lost. The anhydrous salt melts at 415 °C; however, explosions can occur on fusion or if a mixture with sulfur or organic substances is heated even at 110 °C. Barium chlorate is a good oxidant.

Barium chlorate is produced on an industrial scale by mixing concentrated barium chloride solution into hot concentrated sodium chlorate solution. On cooling, barium chlorate monohydrate crystallizes. Although it crystallizes in a rather pure form, barium chlorate can be purified even further by recrystallization. This double conversion is more economical than electrolysis of a barium chloride solution, a process that is no longer used.

Barium chlorate is almost exclusively used in pyrotechnics, where it produces a green color. In addition, it has been used as an oxidizing agent, e.g., in textile printing. Shipment is governed by IMDG Class 5.1, UN 1445, or RID/ADR, Class 5.1, 29b.

2.5. Barium Chloride

Anhydrous barium chloride [*10361-37-2*], $BaCl_2$, M_r 208.25, *mp* 962 °C, ϱ_{20} 3.91 g/cm^3, crystallizes in two modifications: α-$BaCl_2$ (rhombic, mostly lamelliform) and β-$BaCl_2$ (cubic). The conversion temperature is 925 °C. Barium chloride is soluble in water:

°C	0	20	40	60	80	100
wt % $BaCl_2$	23.6	26.3	29.0	31.6	35.0	36.6

The enthalpy of fusion is 193 kJ/kg, and the entropy of fusion is 87.7 J kg^{-1} K^{-1}. The heat capacity c_p is 361.8 J kg^{-1} K^{-1} near room temperature.

Barium chloride crystallizes from aqueous solutions as the dihydrate [*10326-27-9*], M_r 244.31, in monoclinic colorless plates, ϱ 3.097 g/cm^3. Water of crystallization is not lost much before 110 °C, this process being finished around 160 °C.

Production. Today the production of barium chloride on an industrial scale is carried out by the reaction of 15–18 wt % barium sulfide solution with 31–32 wt % hydrochloric acid. In addition, the reaction of barium sulfide solution with chlorine is of commercial interest.

In both methods, reaction is continued until the solution becomes acidic to assure complete conversion of the barium sulfide and to remove from solution the hydrogen sulfide formed. The reaction with hydrochloric acid is often followed by oxidation with air or chlorine to convert any dissolved sulfur compounds into barium sulfate, which removes the sulfur from the solution. In both methods, the solution is neutralized afterwards with barium carbonate or sodium hydroxide to precipitate heavy metals.

The solubility of $BaCl_2$ in water does not depend strongly on temperature. If the barium chloride is to be crystallized, this is carried out by evaporative crystallization. In most cases, vacuum evaporation is used, the salt crystallizing as the dihydrate. Normally it is only dried but it can be dehydrated further if necessary. The mother lye, which still contains barium chloride, can be processed to produce insoluble blanc fixe ($BaSO_4$) or barium carbonate. The byproduct hydrogen sulfide from the hydrochloric acid process can be converted into sodium sulfide or sodium bisulfide or oxidized to produce sulfuric acid.

In the chlorine process, the byproduct is sulfur, usually containing barium carbonate and heavy metals.

Byproduct barium chloride is obtained in the production of lithopone grades containing more than 30 wt % zinc sulfide, for in these cases barium sulfide is reacted with zinc sulfate and zinc chloride. Another source of barium chloride is the black ash leaching residues. These are reacted with hydrochloric acid by some manufac-

turers. However, the resulting barium chloride solutions contain such a concentration of impurities that they must be purified extensively before further use.

The literature reports various procedures to convert barite directly into barium chloride. These procedures, e.g., the reaction of barite with carbon and calcium chloride or the chloridizing reduction of barite, no longer have significance in Western Europe or America, for they are now uneconomical. Direct conversion may still be of interest in Eastern Europe, where numerous scientific papers and patents on direct conversion have been issued, e.g., [91]. The industrial production of barium chloride from barium sulfate, carbon, and calcium chloride is reported only from the People's Republic of China.

Quality Specifications, Packing, Shipping. Both the dihydrate and anhydrous barium chloride are marketed as technical grades of 99–99.5 %, the dihydrate also being available as a high-purity grade (≈ 100 %). They are packed in 50 kg multilayer paper bags with polyethylene covers or liners to protect against moisture. Shipment is governed by IMDG Class 6.1, UN 1564 (harmful, keep separate from foodstuffs), or RID/ADR Class 6.1, 60c.

Uses. Along with other chlorides, barium chloride is used as a component of melt quenching baths for steel. It is used as a starting material for the production of sulfide-free blanc fixe grades and, instead of barium carbonate, to remove sulfate from solutions. Finally, it has a limited importance as a stabilizer for plastics.

2.6. Barium Ferrite

Barium hexaferrite [*12047-11-9*], $BaFe_{12}O_{19}$, M_r 1111.49, forms hexagonal needle-shaped ferrimagnetic crystals. It is usually manufactured in a two-stage solid-state reaction, first stage at 650–750 °C, second stage > 850 °C. An intermediate is barium monoferrite [*12009-00-6*], $BaFe_2O_4$ (→ Magnetic Materials).

2.7. Barium Oxide

Barium oxide [*1304-28-5*], caustic baryta (an older term), BaO, M_r 153.33, *mp* ≈ 1920 °C, sublimes ≈ 2000 °C. Pure commercial barium oxide is a white fine crystalline powder, whereas technical grades are often gray due to carbon impurities. Barium oxide forms face-centered cubic crystals. The densities reported in the literature vary considerably, perhaps depending on preparation. The most likely value is ≈ 5.7 g/cm^3.

Usually barium oxide is very reactive. However, its reactivity depends on the preparative conditions, the reaction temperature, and the coreactant. In the absence of water, BaO often reacts very slowly with carbon dioxide. In the presence of small amounts of water, however, it is converted by carbon dioxide into barium carbonate in a very exothermic reaction. The conversion of barium oxide by water into barium hydroxide is also very exothermic. Self-ignition may occur on contact with wet organic substances. Barium oxide is soluble in methanol: 20 wt % dissolves at 15 °C, and therefore barium oxide can be purified by recrystallization before it is used as a starting material for organic barium compounds or for incorporation, as a methylate, into organic systems.

Production. Barium oxide is mainly produced by decomposition of barium carbonate in the presence of carbonaceous material at high temperature. The use of carbon significantly lowers the temperature at which the barium carbonate decomposes, which to a large extent avoids unwanted corrosion of the reactor when barium carbonate is decomposed by heat alone at 1400 °C. Also important is the decomposition of barium carbonate mixed with channel black (6.2 wt % C) in retorts of silicon carbide – corundum externally heated to ≈ 1250 °C.

The modern procedures tend to use continuous fluidized beds. Barium carbonate – soot pellets are used [92, 93], or the carbon is added to the fluidized bed as methane [94]. The advantages are, among others, that the low reaction temperatures give less-sintered products, i.e., more reactive BaO.

There are also procedures using electrical arcs in a core zone [95]. In such processes a steep temperature drop is maintained between the hot core zone and the reactor shell, to provide a protective layer for the shell material. Addition of carbon is not absolutely necessary.

The crude barium oxide can be reacted [93, 94] in a second stage with carbon for more complete conversion.

Quality Specifications. Barium oxide is marketed in grades of a purity of 97–99 wt %. The main impurities are 1–3 wt % barium carbonate, $\leqq$ 1 wt % barium peroxide, and a few tenths of a percent of nonconverted carbon.

Storage and Shipping. Barium oxide is shipped mainly in iron drums or, for smaller quantities, in tin cans. Shipment is governed by the regulations on the transport of harmful goods of the IMDG Class 6.1, UN 1884, or RID/ADR Class 6.1, 60c.

Uses. Barium oxide, like barium hydroxide, is mainly used for the preparation of additives for oils and greases; however, this is losing significance. In addition, barium oxide is used as a starting material for organic barium salts used in plastics. Barium oxide is frequently used in the production of luminescent materials.

2.8. Barium Peroxide

Barium peroxide [*1304-29-6*], BaO_2, M_r 169.33, → Peroxo Compounds, Inorganic, Section 2.1.2.

2.9. Barium Hydroxide

The anhydrous hydroxide has only a secondary industrial importance; the monohydrate and octahydrate are used in industry on a far larger scale.

Properties. Barium hydroxide octahydrate [*12230-71-6*], hydrate of baryta, caustic baryta, $Ba(OH)_2 \cdot 8\ H_2O$, M_r 315.47, forms monoclinic pseudotetragonal crystals, often plates when rapidly crystallized. It dissolves at 78 °C in its own water of crystallization, releasing the water when heated above 108 °C to yield the monohydrate. The enthalpy of melting is high at a value of 280 J / g. The density is ≈ 2.18 g/cm^3. The specific heat capacity c_p of the solid is 1250 J kg^{-1} K^{-1}, of the melt 1675 J kg^{-1} K^{-1} at 80 °C.

The solubility in water depends dramatically on the temperature:

°C	0	20	40	60	78
wt % $Ba(OH)_2 \cdot 8\ H_2O$	3.0	6.9	14.1	32.6	100.0
g BaO/100 g H_2O	1.5	3.48	7.35	18.8	94.7

The solutions are strongly alkaline and tend to supersaturate.

Like all soluble hydroxides, barium hydroxide, especially in aqueous solution, reacts readily with acidic gases. The most important example in practice is carbon dioxide, to form barium carbonate.

Barium hydroxide monohydrate [*22326-55-2*], $Ba(OH)_2 \cdot H_2O$, M_r 189.37, crystallizes as often radially intergrown needles. Its density ϱ_{20} is ≈ 3.74 g/cm^3. The monohydrate does not react with dry carbon dioxide.

Anhydrous barium hydroxide [*17194-00-2*], $Ba(OH)_2$, M_r 171.35, *mp* ≈ 407 °C. The course of the dehydration, i.e., the temperatures and times, from monohydrate to anhydrous $Ba(OH)_2$ depends on the reactivity of the monohydrate. Molten anhydrous hydroxide is so corrosive, attacking steel and most ceramics, that the conversion of barium hydroxide into barium oxide, which takes place at 800 to 850 °C, is not carried out on an industrial scale.

Production. Barium hydroxide octahydrate can be produced by hydration of barium oxide. The barium oxide is suspended in recycled mother liquor and heated for several hours to destroy peroxides. The insolubles, e.g., $BaCO_3$, are separated, and the hydroxide is precipitated by crystallization at low temperatures in stainless steel vessels. The crystals are removed from the mother lye by centrifugation.

Another industrial method is based on the oxidation of a barium sulfide solution [96 – 98]. Air is passed into the stirred solution containing barium hydroxide and barium hydrosulfide, the hydrolysis products of barium sulfide, until the hydrosulfide is oxidized to form polysulfide. Then the barium hydroxide octahydrate is crystallized at low temperatures and separated with a centrifuge. The polysulfidic mother liquor is partly recycled; the rest is usually processed into other barium compounds. The advantage of this procedure over the direct reaction of BaS solu-

tions with NaOH is the lower strontium content of the barium hydroxide product.

The monohydrate is generally produced on a commercial scale by dehydration of the octahydrate in heated vacuum driers. A monohydrate of especially high reactivity is obtained with drum driers [99]. This highly reactive monohydrate is particularly suitable for the production of oil additives, PVC stabilizers, and catalysts.

Storage, Packing, Shipping. Both the monohydrate and the octahydrate are packed and stored in 25 or 50 kg polyethylene-lined multilayer paper bags or in big bags. Shipping is governed by IMDG Class 6.1, UN 1564, or RID/ADR Class 6.1, 60c.

Uses. Barium hydroxide, especially the monohydrate, is used to produce organic barium compounds such as additives for oil and stabilizers for plastics. In addition, barium hydroxide is used for dehydration and deacidification, especially for removing sulfuric acid from fats, oils, waxes, and glycerol. In some applications, e.g., removal of sulfate from water and as starting material for other barium compounds, barium hydroxide to some extent replaces barium carbonate. The earlier use of the hydroxide in some countries for extracting the residual sucrose from molasses is no longer of much significance. The patent literature describes a procedure in which barium hydroxide is reacted with hydrogen peroxide to produce barium peroxide monohydrate [100].

2.10. Barium Nitrate

Barium nitrate [*10022-31-8*], $Ba(NO_3)_2$, M_r 261.34, *mp* 593 °C, colorless cubic crystals, often octahedra, ϱ_{20} 3.24 g/cm^3. The solubility of barium nitrate in water increases with temperature:

°C	0	10	20	50	100
wt % $Ba(NO_3)_2$	4.8	6.5	8.4	14.6	25.4

On heating, barium nitrate decomposes into barium oxide, barium peroxide, nitric oxide, nitrogen, and oxygen. The gaseous products formed and their proportions depend largely on the temperature and rate of heating [101]. The presence of sulfur or organic substances may cause an explosion.

For the *production* of barium nitrate, various procedures are possible, starting from $BaCO_3$, $Ba(OH)_2 \cdot 8\,H_2O$, BaS, or $BaCl_2$. However, for industrial production there are two main procedures. In the first, barium carbonate is suspended in recycled mother lye and reacted with nitric acid. If necessary, heavy metals are precipitated, and the solution is filtered hot. The nitrate is then crystallized, if necessary with vacuum evaporation. The second procedure consists of reacting a barium sulfide solution, directly or after the removal of barium hydroxide, with nitric acid. After carefully removing any sulfur formed and the heavy metals, the barium nitrate is crystallized at low temperatures.

The preparation of $Ba(NO_3)_2$ by double conversion of barium chloride and sodium nitrate or calcium nitrate is described in the Eastern European patent literature.

Barium nitrate is packed and stored in 25 kg polyethylene-lined multilayer paper bags. Shipping is governed by IMDG Class 5.1, UN 1446, or RID/ADR Class 5.1, 29b.

One of the main uses, as a component in pyrotechnics to produce a green color, requires very pure barium nitrate, 99.5–99.8 %. The content of other flame colorants — sodium, strontium, calcium — must be extremely low. Barium nitrate is used as a plaining agent in the manufacture of special glasses and optical glasses in place of the combination alkali-metal nitrate – arsenic. In addition it is used to incorporate barium oxide into catalysts, in luminescent cathode-ray screens, and as a component of metallurgical hardening agents.

2.11. Barium Titanate

Barium titanate [*12047-27-7*], $BaTiO_3$, M_r 233.24, *mp* 1625 °C, ϱ_{20} 6.02 g/cm^3, has phase transitions and both ferroelectric and piezoelectric properties (→ Ferroelectrics; → Insulation, Electric).

2.12. Barium Salts of Organic Acids

Barium acetate is used as a mordant in the printing industry and as a catalyst in organic chemistry.

Barium stearate is used as a slip agent in plastics and metal working, as a stabilizer against heat and light in PVC, and also in the production of grease.

The compound known as barium octoate is normally the barium salt of 2-ethylhexanoic acid. It is used as a siccative (drying agent) for the curing of alkyd paints. Siccatives catalyze the reaction between double bonds of the resin and atmospheric oxygen.

Certain dyes and pigments, also known as lakes, contain barium either as the salt of a carboxylic acid group or of a sulfonic acid group within the molecule. Well-known examples are the β-naphthol azo pigments [102].

3. Barium in the Environment

The terrestrial abundance of barium is approximately 250 g/t [10]. The estimated average barium concentration in the soil is 500 g/t [103]. Measured concentrations range between 100 and 3000 g/t [104].

Barium occurs in seawater in a concentration of 6 µg/L [105]. This level is due to reaction between barium and sulfate ions also present in the oceans. The concentration of barium is higher in deep ocean waters than in surface ocean waters [11]. The precipitated barium sulfate forms a permanent part of the sediment on the ocean floor [106]. In fresh water the barium content depends upon the occurrence of barium and the concentration of anions which form barium salts of low solubility such as sulfate and carbonate ions. Values between 7 and 15 000 µg/L (average: 50 µg/L) have been reported [107].

Studies of drinking water gave a wide range of values between only traces and 10 000 µg/L in the USA [119], 5–600 µg/L in Canada [120], and 1–20 µg/L in municipal drinking waters in Sweden [107].

Barium levels in the air are not well documented. An estimate of a mean value for the USA is 0.05 µg/L [121]. There is no correlation between the degree of industrialization and the barium concentration in the air. Higher levels are found in areas with high natural dust levels. Anthropogenic emissions are primarily industrial. Reports about the release of barium by way of exhaust gases, mainly from diesel engines [122, 123], are no longer valid since in the last few years the primary cause, namely, the sulfur content of diesel fuels, has been reduced due to legislation. Therefore, there is no longer a need to use barium compounds as sulfur-scavenging fuel additives. Other atmospheric emissions result from the handling of barium compounds or materials containing barium compounds, such welding wires [124].

4. Dietary Intake of Barium

Figures for the mean dietary intake of barium in the USA ranged from 300 to 1700 µg/d [125] and 650 to 1770 µg/d in two other studies [126]. These results match well with barium intakes in other parts of the world [127] with reported levels between 80 and 1330 µg/d. Apart from drinking water, milk, potatoes, and flour have been suggested to be the main sources for barium in the diet [128], even though absolute barium concentrations are higher in, for example, eggs and certain types of cheese. The normal dietary intake of barium within these limits is apparently harmless to humans. Acute symptoms will, however occur in case of an overdose of bioavailable barium (see Chap. 5). There is no evidence that barium is carcinogenic [121].

Typical levels of barium in human tissues and blood [11]: 3–70 ppm in the bone; 0.04–1.2 ppm in the liver; 0.09 ppm in the muscle, and 0.06 mg/dm^3 in the blood. The total mass average for a 70-kg human is 22 mg.

5. Toxicology

Water-soluble barium compounds are toxic, as is shown in Table 4. Most barium poisonings are caused by mistaking soluble barium compounds, such as barium carbonate, for the insoluble — and therefore nontoxic — barium sulfate, which is used as a contrast agent in X-ray diagnosis.

The symptoms of poisoning are convulsions of both striated and smooth muscles, including the heart, followed by paralysis of the peripheral nerve system. These effects on the nerve system are accompanied by severe inflammation of the gastrointestinal tract.

Table 4. Acute lethal doses of soluble barium compounds

Compound	Toxicity	Ref.
Barium carbonate [*513-77-9*]	TDLo 29 mg/kg (human, oral)	[108]
Barium nitrate [*10022-31-8*]	LD_{50} 8.5 mg/kg (mouse, i.v.)	[109]
	LD_{50} 355 mg/kg (rat, oral)	[110]
Barium chloride [*10361-37-2*]	LDLo 11.4 mg/kg (human, oral)	[111]
Barium oxide [*1304-28-5*]	LD_{50} 50 mg/kg (mouse, subcutaneous)	[112]
Barium fluoride [*7787-32-8*]	LD_{50} 250 mg/kg (rat, oral)	[113]

Barium carbonate is often used in the ceramics industry. Detailed examination of workers exposed to $BaCO_3$ dust for 7–27 years did not reveal any specific chronic poisoning [114]. Damage to the lungs has been reported to have been caused by contaminants in the barium compounds, such as quartz [115] or zinc sulfide [116]. However, $BaSO_4$ may cause a benign pneumoconiosis (barytosis) because, unlike $BaCO_3$, $BaSO_4$ is not absorbed by the organism [117, 118]. The legally binding "Permissible Exposure Limit" (PEL) for $BaSO_4$ dust in the USA is 15 mg/m^3, whereas the voluntary "Threshold Limit Value" (TLV) was set to 10 mg/m^3. The German MAK value for $BaSO_4$ dust is 3 mg/m^3.

For soluble barium compounds, an exposure limit of 0.5 mg/m^3 (as Ba) has been established by both TLV (1989) and MAK (1996) commissions. According to the MAK, short-term exposure limits may be as high as 1 mg/m^3 (average over 30 min, four times within 8 h).

6. References

General References

1. *Ullmann*, 4th ed., vol. **8**, p. 301.
2. *Gmelin*, Barium (system no. 30), main volume, 1932; supplement volume, 1960.
3. R. Fromageau, *Monogr. Mét. Haute Pureté* 1977, no. 2, 153–183.
4. *Kirk-Othmer*, 3rd ed., vol. **3**, pp. 457–463.
5. R. D. Goodenough, V. A. Stenger: "Magnesium, Calcium, Strontium, Barium and Radium," in J. C. Bailar (ed.): *Comprehensive Inorganic Chemistry*, vol. 1, Pergamon Press, Oxford 1973, pp. 591–664.
6. J. Alexander, M. Steffel: "Calcium, Strontium and Barium," in F. Korte (ed.): *Methodicum Chimicum,* vol. 7, Georg Thieme Verlag, Stuttgart 1976, pp. 97–106.
7. *Winnacker-Küchler*, vol. **6**: Metallurgie, pp. 81–107.
8. C. L. Mantell: "Barium," in C. A. Hampel (ed.): *Rare Metals Handbook*, 2nd ed., Reinhold Publ. Co., London 1961, pp. 25–31.
9. *Römpps Chemie-Lexikon*, 8th ed., vol. **1**, Franckh'sche Verlagshandlung, Stuttgart 1979, p. 365.

Specific References

10. B. Mason: *Principles of Geochemistry*, J. Wiley & Sons, , New York 1952, p. 41.
11. Kenneth Barbalace. Chemical Database - Barium. EnvironmentalChemistry.com. 1995 - 2006. Accessed online: 11/15/2006, http://EnvironmentalChemistry.com//yogi /chemicals/cn/Barium.html.
12. H. Davy, *Nicholsons J.* **21** (Suppl. 1808) 366; **22** (1809) 54.
13. J. W. Mallet, *Justus Liebigs Ann. Chem.* **190** (1878) 62.
14. A. Guntz, *C. R. Hebd. Seances Acad. Sci.* **143** (1906) 339; *Ann. Chim. Phys.* **10** (1907) 441.
15. J. Evers, A. Weiss, *J. Less Common Met.* **30** (1973) 83.
16. E. Miller et al., *Trans. Metall. Soc. AIME* **218** (1960) 978.
17. P. Ehrlich, H. J. Seifert in G. Brauer (ed.): *Handbuch der Präparativen Anorg. Chemie*, vol. **2**, Enke Verlag, Stuttgart 1978, pp. 917–934.
18. K. E. Maas, *Z. Anorg. Allg. Chem.* **374** (1970) 1.
19. R. Schmid: "Nitride der Hauptgruppenmetalle," in F. Korte (ed.): *Methodicum Chimicum,* vol. 7, Georg Thieme Verlag, Stuttgart 1976, pp. 59–62.
20. I. Obinata et al., *Metall (Berlin)* **19** (1965) 21; also see [6] and the literature cited there.
21. J. J. Alexander: "Carbide der Elemente der 1.–4. Hauptgruppe," in F. Korte (ed.): *Methodicum Chimicum,* vol. 7, Georg Thieme Verlag, Stuttgart 1976, p. 67.
22. A. Drozdov et al., *Material Science and Engineering* **B18** (1993) 139–140.
23. W. Biltz, G. Hüttig, *Z. Anorg. Allg. Chem.* **114** (1920) 241.
24. R. Juza, H. Schuhmacher, *Z. Anorg. Allg. Chem.* **324** (1963) 278.
25. *Houben-Weyl*, **13** (2 a), 542–543.
26. W. Lindsell in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry*, vol. **1**, Pergamon Press, Oxford 1982, pp. 223–252.

27. W. E. Lindsell in: *Comprehensive Organometallic Chemistry II*, Pergamon Press, New York 1995, pp. 57–128.
28. M. Hansen: *Constitution of Binary Alloys*, McGraw-Hill, New York 1958;
1st Supplement, R. P. Elliott (ed.), New York 1965;
2nd Supplement, F. A. Shunk (ed.), New York 1969.
29. H. Seliger, *Freiberg. Forschungsh. B* **34** (1959) 80.
30. J. Krüger in O. Winkler, R. Bakish (eds.): *Vacuum Metallurgy*, Elsevier, Amsterdam 1971, pp. 207–212.
31. I. S. Bystrova et al., *Proizvod. Ferrosplavov* 1977, no. 5, 31–37;
Chem. Abstr. **87** (1977) 120951.
32. O. Orman, E. Zembala, *Pr. Inst. Hutn.* **4** (1952) 437.
33. E. Fujita, H. Yokomizo, *Tokyo Kogyo Shikensho Hokoku* **47** (1952) 291.
34. P. S. Danner, *J. Am. Chem. Soc.* **46** (1924) 2382.
35. J. Evers, G. Oehlinger, A. Weiss, *J. Less Common Met.* **81** (1981) 15.
36. C. Matignon, *C. R. Hebd. Seances Acad. Sci.* **156** (1913) 1378.
37. W. Sundermeyer, *Angew. Chem.* **77** (1965) 245.
38. Ges. Deutscher Metallhütten- und Bergleute: *Analyse der Metalle*, 2nd ed., vol. **II/1**, Springer Verlag, Berlin 1961, pp. 139–145.
39. L. Edelbeck, P. W. West, *Anal. Chim. Acta* **52** (1970) 447.
40. Chemetall GmbH, Lieferprogramm Sondermetalle, Frankfurt, FRG, 1998.
41. G. Tölg, *Fresenius' Z. Anal. Chem.* **294** (1979) 1;
also see [27], pp. 22–23.
42. L. Melnick, L. Lewis, B. Holt (eds.): "Determination of Gaseous Elements in Metals," *Monographs on Analytical Chemistry*, vol. 40, J. Wiley & Sons, New York 1974.
43. IMDG, Class 4.3, p. 4147; UN 1400; 4.3/1A RID, ADR, GGVS, GGVE.
44. I. Sax (ed.): *Dangerous Properties of Industrial Materials*, 4th ed., Van Nostrand Reinhold Co., New York 1975, p. 430.
45. S. Wagener, *Z. Angew. Phys.* **6** (1954) 433.
46. J. Maley, J. Moscony, *J. Vac. Sci. Technol.* **6** (1969) 51.
47. K. Marquard, *Residual Gases Electron Tubes Proc. Int. Conf. 4th* 1971, 315–334;
Chem. Abstr. **77** (1972) 106735.
48. M. Malev, *Vacuum* **23** (1973) 359.
49. J. Turnbull, *J. Vac. Sci. Technol.* **14** (1977) 636.
50. U.S.-Philips Corp., US 4077899, 1978; *Chem. Abstr.* **88** (1978) 181336.
51. A. Misumi et al., *Shinku* **23** (1980) 183;
Chem. Abstr. **93** (1980) 178254.
52. Phillips Electronic, GB 1182883, 1970; *Chem. Abstr.* **72** (1970) 115999.
53. P. Ward, *J. Illum. Eng. Soc.* **9** (1980) 194;
Chem. Abstr. **93** (1980) 190370.
54. W. Cowan et al., *Trans. Am. Electrochem. Soc.* **40** (1921) 27.
55. P. Paschen et al., *Metall (Berlin)* **23** (1969) 786.
56. D. Randolph, *Trans. Electrochem. Soc.* **66** (1934) 85.
57. Union Carbide Corp., US 3734714, 1973.
58. V. Nikolaev, V. Vasilev, *Liteinoe Proizvod.* 1980, no. 4, 6;
Chem. Abstr. **93** (1980) 135988.
59. V. D. Povolotskii et al., *Izv. Vyssh. Uchebn. Zaved. Chern. Metall.* **10** (1982) 69;
Met. Abstr. **16** (1983) 45-0527.
60. Z. Zhou et al., *Giesserei Prax.* 1983, no. 4, 49–52.
61. E. Sibbing: "Schwerspat: Rohstoffbasis für Füllstoffe und Bariumchemikalien," part 1, *Kunstst. J.* **8** (1974) no. 10, 30–36.
62. E. Sibbing: "Schwerspat: Rohstoffbasis für Füllstoffe und Bariumchemikalien," part 2, *Kunstst. J.* **8** (1974) no. 11, 26–31.
63. Barytes — Its Uses Outside Drilling Muds, *Ind. Miner. (London)* 1978, no. 133, 33–43.
64. J. Massonne: "Technology and Uses of Barium and Strontium Compounds," *Ind. Miner.* (London) 1982, no. 177, 65–69.
65. D. Rohe: "Nicht nur Masse, sondern auch Qualität durch mineralische Füllstoffe, part III: Sulfate," *Chem. Ind.* (Düsseldorf) 35 (1983) 369–373.
66. P. Moore, *Industrial Minerals*, February 2004, 26–33.
67. P. Moore, J. Newcaster, *Industrial Minerals*, December 2003, 36–39.
68. *Industrial Minerals Prices and Data 1997*, Industrial Minerals Information Ltd., London 1998.
69. *Gmelin*, Barium (system no. 30), main volume, 1932;
supplement volume, 1960.
70. M. Clement, H. Surmatz, H. Hüttenhain, *Z. Erzbergbau Metallhuettenwes.* **20** (1967) 512–522.
71. Klöckner Humboldt-Deutz, DE-OS 1783014, 1968.

72. Oil Companies Material Association (OCMA): *Specification No. DFCP 3*, Cecil Chambers, London 1973.
American Petroleum Institute (API): *API RP 13 B*, 6th ed., Washington, D.C., 1976.
73. K.-H. Grodde: *Bohrspülungen und Zementschlämme in der Tiefbohrtechnik*, Verlag O. Vieth, Hamburg 1963.
74. Sachtleben, DE-OS 1955881, 1969.
75. Sachtleben, DE 567348, 1930.
76. Titan Co., GB 427220, 1933.
77. B. Bittmann, B. Wetzel, F. Haupert, 10. Merseburger Problemseminar, June 15-17, 2006, Merseburg, Germany, Conference Proceedings.
78. R. Maiwald, Dissertation 1977, Bergakademie Freiberg.
79. A. A. Rawdel, N. A. Nowikowa, *Zh. Prikl. Khim. (Leningrad)* **36** (1963) 1433–1442.
80. Forschungsinstitut für NE-Metalle, DE-OS 1567512, 1965.
81. G. Salvi, A. Fiumara, *Riv. Combust.* **12** (1958) 525–561.
82. N. S. Safiullin, S. F. Gawrilowa, *Khim. Promst. (Kiev)* 1965, no. 3, 32–35.
83. Kali-Chemie, DE-OS 2034065, 1970.
84. Kali-Chemie, DE 1800866, 1968.
85. E. H. Baker, *J. Chem. Soc.* 1964, 699–704.
86. M. M. Evstigneeva, A. A. Bundel, B. V. Kondakov, *Tr. Mosk. Khim. Tekhnol. Inst. im. D. I. Mendeleeva* **62** (1969) 291–294; *Zh. Fiz. Khim.* **43** (1969) 2613–2614; **44** (1970) 2607–2609.
87. A. Schulze, DD 58501, 1966.
88. Kali-Chemie, DE 1907440, 1969.
89. A. Schulze, DD 45070, 1964.
90. PPG Industries, US 3467494, 1966.
91. A. F. Lozhkin, SU 245047, 1967.
92. Columbia Southern Chem., US 2772948, 1952.
93. Columbia Southern Chem., US 2772949, 1952.
94. Columbia Southern Chem., US 2915366, 1956.
95. FMC, US 3031266, 1959.
96. IG Farbenind., DE 519891, 1929.
97. IG Farbenind., DE 526796, 1930.
98. Kali-Chemie, DE 1792505, 1968.
99. FMC, DE 1186841, 1962.
100. Peroxid-Chemie, DE-OS 203063, 1983.
101. F. Lazarini, B. S. Brčié, *Monatsh. Chem.* **97** (1966) 1318.
102. *Ullmann*, 5th ed., **A 20**, 1992 371–412.
103. R. R. Brooks, "Pollution Through Trace Elements," in J. O. M. Bockris (ed.): *Environmental Chemistry*, Plenum Press, New York 1978, pp. 429–476.
104. W. O. Robinson, R. R. Whetstone. G. Edington, "The Occurrence Of Barium In Soils And Plants," *U.S. Dept. Agric. Tech.* **1013** (1950) pp. 429–476.
H. A. Schroeder, "Barium," *American Petroleum Institute, Air quality Monograph* **70** (1970) 12.
105. H. A. Schroeder. I. H. Tipton, A. P. Nason, H. A. Schroeder, "Trace Metals in Man: Strontium and Barium," *J. Chron. Dis.* **25** (1972) 491–517.
106. K. Wolgemuth, W. S. Broecker, "Barium in Sea Water," *Earth planet. Sci. Lett.* **8** (1970) 372–378.
107. A. L. Reeves, "Barium," in L. Friberg, G. F. Nordberg, B Velimir (eds.): *Handbook on the Toxicology of Metals – Volume II: Specific Metals* Elsevier Science Publishers, Amsterdam, 1986, pp. 84–93.
108. *Israel J. Med. Sci.* **3** (1967) 565.
109. *Toxicol. Appl. Pharmacol.* **22** (1982) 150.
110. J. V. Marhold: *Sbornik Vysledku Toxikologickeho Vysetreni Latek a Pripravku*, Institut Pro Vychovu Veducicn Pracovniku Chemickeho Prumyclu, Praha 1972, p. 10.
111. T. Sollmann (ed.): *Manual of Pharmacology and its Applications to Therapeutics and Toxicology*, 7th ed., Saunders, Philadelphia 1942, p. 478.
112. *Zh. Vses. Khim. Ova.* **19** (1974) 186.
113. *Vestn. Akad. Med. Nauk SSSR* **32** (1977) 28.
114. H.-G. Essing, G. Bühlmeyer, G. Kemmerer, R. Procharka, *Arbeitsmed. Sozialmed. Praeventivmed.* **11** (1976) no. 12, 299–302.
115. E. Browning: *Toxicity of Industrial Metals*, 2nd ed., Butterworths, London 1965.
116. E. Wende, *Arch. Gewerbepathol. Gewerbehyg.* **15** (1956) 171.
117. Z. Maretič, *Med. Klin. (München)* **52** (1957) 1950–1953.
118. E. P. Pendergrass, R. R. Gruning, *Arch. Ind. Hyg. Occup. Med.* **7** (1959) 44–48.
119. C. M. Durfor; E. Becker, "Public Water Supplies of the 100 Largest Cities in the United States," Water Supply Paper No. 1812, US Department of the Interior, Government Printing Office, Washington D.C. 1962.
P. R. Barnett, M. W. Skougstadt, K. J. Miller, "Chemical Characterization of a Public Water Supply," *J. Am. Water Works Assoc.* **2** (1969) 60–67.

L. McCabe in A. Saponzik (ed.): "Problems of Trace Metals in Water Supplies — an Overview," *Proceedings of the 16th Water Quality Conference on Trace Metals in Water Supplies: Occurence, Significance and Control* , University of Illinois, Urbana, IL, 12–13 Feb. 1974, p. 1–9.
L. J. McCabe, et al., *J. Am. Water Works Assoc.* **62** (1970) 670–687.
E. J. Calabrese, "Excessive Barium and Radium-226 in Illinois Drinking Water," *J. environ. Health* **39** (1977) no. 5, 366–369.
"An AWWA Survey of Inorganic Contaminants in Water Supplies," *J. Am. Water Works Assoc.* (1985) 67–72.

120. K. S. Subramanian, J. C. Meranger, "A Survey for Sodium, Potassium, Barium, Arsenic, and Selenium in Canadian Drinking Water Supplies," *At. Spectrosc.* **5** (1984) 34–37.
121. ICPS Environmental Health Criteria 107, Barium, World Health Organization, Geneva 1990, p. 14.
122. W. R. Pierson et al., "Effects of Barium Fuel Additive and Fuel Sulfur Level on Diesel Exhaust," *Environ. Sci. Technol.* **14** (1981) no. 9, 1121–1124.
123. S. Miner, *Air Pollution Aspects of Barium and its Compounds*, Litton Systems Inc, Bethesda, Maryland 1969, p. 63 (NTIS PB-188083).
124. *Health Hazard Evaluation Determination Report: Mark Steel Corporation, Salt Lake City, Utha*, NIOSH Report No. 78–93–536, National Institute for Occupational Health and Safety, Cincinnati, Ohio 1978.
125. E. J. Underwood: *Trace Elements in Human and Animal Nutrition*, Academic Press, New York 1977.
126. I. H. Tipton, P. L. Stewart, P. G. Martin, "Trace Elements in Diets and Excreta," *Health Phys.* **12** (1966) 1683–1689.
I. H. Tipton, P. L. Stewart, J. Dickson, "Patterns of Elemental Excretion in Long Term Balance Studies" *Health Phys.* **16** (1969) 455–462.
127. H.A. Schroeder, I. H. Tipton, A. P. Nason, "Trace Metals in Man: Strontium and Barium," *J. Chron. Dis.* **25** (1972) 491–517.
E. W. Murphy, L. Page, B. K. Watt, "Trace Minerals in Type A School Lunches," *J. Am. Diet. Assoc.* **59** (1971) 115–122.
E. I. Hamilton, M. J. Minski, "Abundance of the Chemical Element in Man's Diet and the Possible Relations with Environmental Factors," *Sci. Environ.* **1** (1972) 315.
International Commision on Radiological Protection (ICRP): *Recommendations of the International Commision on Radiological Protection, Report of Committee II*, ICRP Publication No. 2, Pergamon Press, New York 1959.
International Commision on Radiological Protection (ICRP): *Report of the Task Group on Reference Man, Report of Committee II*, ICRP Report No. 23, Pergamon Press, New York 1974.
128. E. J. Calabrese, A. T. Canada, C. Sacco, "Trace Elements and Public Health," *Annu. Rev. public Health* **6** (1985) 131–146.

Batteries

DIETRICH BERNDT, (formerly VARTA Batterie AG), Kronberg, Federal Republic of Germany (Chaps. 1 – 5, Chaps. 7 and 9)

DIETER SPAHRBIER, (formerly VARTA Batterie AG), Kelkheim, Federal Republic of Germany (Chaps. 1 – 6, Chap. 8)

1. Introduction 4
2. History 5
3. Fundamental Laws 6
3.1. Equilibrium or Thermodynamic Parameters 6
3.2. Current Flow, Kinetic Parameters 8
3.2.1. Electron Transfer 9
3.2.2. Tafel Lines 10
3.2.3. Influence of Temperature 11
3.3. Heat Effects 11
3.3.1. Reversible Heat Effect 11
3.3.2. Current Related Heat Effects (Joule Effect) 11
3.3.3. Heat Generation in Total 12
3.4. Heating of the Battery 12
4. Battery Properties and Characteristics 12
4.1. Cell and Battery 13
4.1.1. Active Material 13
4.1.1.1. Change of Volume 13
4.1.1.2. Change of Electrode Structure . 14
4.1.2. Nonactive Components 14
4.1.2.1. Conducting Components 14
4.1.2.2. Nonconducting Components . . 14
4.2. Battery Parameters 16
4.2.1. Voltage 16
4.2.2. Capacity 16
4.2.2.1. Secondary Batteries 16
4.2.2.2. Primary Batteries 17
4.2.3. Energy Content 18
4.2.4. Internal Resistance and Power Output 19
4.2.5. Charge Parameters 21
4.2.6. Self-Discharge 21
4.2.6.1. Mixed Potential 21
4.2.6.2. Further Self-Discharge Mechanisms 22
4.2.6.3. Apparent Self-Discharge 22
4.2.6.4. Capacity Loss During Storage . 22
5. Classification of Battery Systems 23
5.1. Classification by Electrolyte . 23
5.2. Classification by Construction 23
5.2.1. External Construction 23
5.2.2. Internal Construction 24
6. Primary Batteries, Systems, and Application 25
6.1. Systems Employing an Aqueous Electrolyte 25
6.1.1. Batteries with Weak Acid Electrolyte 25
6.1.1.1. Carbon – Zinc Systems 25
6.1.1.1.1. General 25
6.1.1.1.2. Electrochemistry 25
6.1.1.1.3. System-Immanent Properties . . 27
6.1.1.1.4. Batteries and Their Applications 28
6.1.1.1.5. Prospects 31
6.1.1.2. Air – Zinc System 31
6.1.1.2.1. General 31
6.1.1.2.2. Electrochemistry 32
6.1.1.2.3. System-Immanent Properties . . 32
6.1.1.2.4. Batteries and Their Applications 33
6.1.1.2.5. Prospects 34
6.1.2. Batteries with "Neutral Electrolyte" 34
6.1.3. Batteries with Alkaline Electrolyte 35
6.1.3.1. Manganese Dioxide – Zinc System 36
6.1.3.1.1. General 36
6.1.3.1.2. Electrochemistry 36
6.1.3.1.3. System-Immanent Properties . . 37
6.1.3.1.4. Batteries and Their Applications 39
6.1.3.1.5. Rechargeable Alkaline Manganese Dioxide – Zinc System . 40
6.1.3.1.6. Prospects 42
6.1.3.2. Silver Oxide – Zinc System . . . 42
6.1.3.2.1. General 42
6.1.3.2.2. Electrochemistry 42
6.1.3.2.3. System-Immanent Properties . . 43
6.1.3.2.4. Batteries and Their Application 44
6.1.3.2.5. Prospects 45
6.1.3.3. Mercuric Oxide – Zinc System . 46
6.1.3.3.1. General 46
6.1.3.3.2. Electrochemistry 46

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_343

6.1.3.3.3. System-Immanent Properties . . 46
6.1.3.3.4. Batteries and Application 47
6.1.3.3.5. Prospects 47
6.1.3.4. Alkaline Air – Zinc System . . . 47
6.1.3.4.1. General 47
6.1.3.4.2. Electrochemistry 48
6.1.3.4.3. System-Immanent Properties . . 48
6.1.3.4.4. Air Access Opening: Its Seal and Diffusion Resistance 50
6.1.3.4.5. Batteries and Their Application 51
6.1.3.4.6. Prospects 51
6.2. Batteries with Nonaqueous Electrolyte 52
6.2.1. Batteries with Organic Electrolyte 53
6.2.1.1. Manganese Dioxide – Lithium System 54
6.2.1.1.1. General 54
6.2.1.1.2. Electrochemistry 54
6.2.1.1.3. System-Immanent Properties . . 54
6.2.1.1.4. Batteries and Their Applications 54
6.2.1.1.5. Prospects 57
6.2.1.2. Carbon Poly-monofluoride–Lithium System 57
6.2.1.2.1. General 57
6.2.1.2.2. Electrochemistry 57
6.2.1.2.3. System-Immanent Properties . . 57
6.2.1.2.4. Batteries and Their Applications 57
6.2.1.2.5. Prospects 58
6.2.1.3. Iron Disulfide – Lithium Systems 59
6.2.1.3.1. "1.5 V" Lithium Systems 59
6.2.1.3.2. Electrochemistry 59
6.2.1.3.3. System-Immanent Properties . . 60
6.2.1.3.4. Batteries and Their Applications 60
6.2.1.3.5. Prospects 60
6.2.2. Batteries with Soluble Cathode 60
6.2.2.1. Sulfur Dioxide – Lithium System 60
6.2.2.1.1. General 60
6.2.2.1.2. Electrochemistry 61
6.2.2.1.3. System-Immanent Properties . . 61
6.2.2.1.4. Batteries and Their Applications 62
6.2.2.1.5. Prospects 62
6.2.2.2. Thionyl Chloride – Lithium System 62
6.2.2.2.1. General 62
6.2.2.2.2. Electrochemistry 63
6.2.2.2.3. System-Immanent Properties . . 64
6.2.2.2.4. Batteries and Their Applications 64
6.2.2.2.5. Prospects 65
6.2.3. Batteries with a Nonliquid Electrolyte 66
6.2.3.1. Polymer and Solid Electrolyte Systems 66
6.2.3.2. The $LiI(Al_2O_3)$ Solid Electrolyte System 67
6.2.3.3. The Lithium – Iodine System . . 67
6.2.3.3.1. General 67
6.2.3.3.2. Electrochemistry 67
6.2.3.3.3. System-Immanent Properties . . 68
6.2.3.3.4. Batteries and Their Applications 68
6.2.3.3.5. Prospects 69
6.3. Reserve Batteries 69
6.3.1. Principles 69
6.3.1.1. System Activation with Liquid Electrolyte 69
6.3.1.1.1. Water-Activated Batteries 69
6.3.1.1.2. KOH Activated Batteries 70
6.3.1.1.3. Ampoule-Type Batteries 70
6.3.1.1.4. Battery Activation by Ammonia 71
6.3.1.2. Thermal Batteries 71
6.3.1.2.1. General 71
6.3.1.2.2. Systems 72
6.3.1.2.3. Pyrotechnic Heat Sources 73
6.3.1.2.4. Battery Construction 73
6.3.1.2.5. Batteries 73
6.3.2. Prospects 74
7. Secondary Batteries 74
7.1. Batteries with Aqueous Electrolyte 74
7.1.1. Lead – Acid Batteries 76
7.1.1.1. Charge/Discharge Reactions . . 76
7.1.1.2. Secondary Reactions 77
7.1.1.3. Self-Discharge 79
7.1.1.4. Heat Effects 79
7.1.1.5. Valve-Regulated Lead – Acid Batteries 81
7.1.1.6. Discharge Performance 82
7.1.1.7. Charge Performance 83
7.1.1.8. Technical Design and Manufacturing 88
7.1.1.9. Lead Alloys 92
7.1.1.10. Separators 92
7.1.1.11. Container Material, Post Seals . 93
7.1.1.12. Vents, Valves 94
7.1.1.13. Water Refilling Systems, Forced Acid Agitation 95
7.1.1.14. Catalytic Gas Recombination . . 95
7.1.2. Batteries with Alkaline Electrolyte 95
7.1.2.1. Nickel – Cadmium Batteries . . 95
7.1.2.1.1. Thermodynamic Situation 96
7.1.2.1.2. Charge/Discharge Reactions . . 96
7.1.2.1.3. Equilibrium or Open-Circuit Voltage 98

7.1.2.1.4. Reversible Heat Effect, Calorific Voltage 98
7.1.2.1.5. Heat Capacity (Specific Heat Content) 99
7.1.2.1.6. Heat Effects 99
7.1.2.1.7. Self-Discharge 100
7.1.2.1.8. Types of Electrodes 102
7.1.2.1.9. Cell Design 105
7.1.2.1.10. Cell Assemblies, Power Packs . 107
7.1.2.1.11. Valve Design 107
7.1.2.1.12. Controlled Charging 108
7.1.2.2. Nickel – Hydrogen Batteries . . 110
7.1.2.3. Nickel – Metal Hydride Batteries 112
7.1.2.3.1. Thermodynamic Parameters . . 112
7.1.2.3.2. Hydrogen-Absorbing Alloys . . 113
7.1.2.3.3. Design 115
7.1.2.4. Nickel – Iron Batteries 115
7.1.2.5. Nickel – Zinc Batteries 117
7.1.2.6. Silver Oxide – Zinc Batteries . . 118
7.1.3. Rechargeable Batteries of Particular Design 119
7.1.3.1. The Chlorine – Zinc Battery . . . 119
7.1.3.2. The Bromine – Zinc Battery . . . 120
7.1.3.3. The Vanadium Redox Battery . 121
7.1.3.4. Rechargeable Air – Zinc Batteries 122
7.2. Batteries with Nonaqueous Electrolyte 122
7.2.1. Rechargeable Lithium Batteries 122
7.2.1.1. Button or Coin Cells 123
7.2.1.2. Lithium Batteries with Polymer Electrolytes 124
7.2.1.3. Batteries with Glassy Electrolytes 125
7.2.1.4. Lithium-Ion Batteries 125
7.2.1.4.1. Lithium – Carbon Electrodes (Lithiated Carbons) 126
7.2.1.4.2. Negative Electrodes Based on Alloys 128
7.2.1.4.3. Active Materials in the Positive Electrode (Cathode) 129
7.2.1.4.4. Electrolytes 130
7.2.1.4.5. Separators 131
7.2.1.4.6. Lithium-Ion Polymer Batteries . 131
7.2.1.4.7. Practical Batteries 131
7.2.2. High-Temperature Batteries . . . 132
7.2.2.1. Iron Sulfide – Lithium Batteries 133
7.2.2.2. Sodium – Sulfur Batteries 134
7.2.2.3. Nickel Chloride – Sodium (ZEBRA) Batteries 135
8. Battery Standards 137
8.1. General 137
8.2. National Standards 137
8.3. International Standards 138
8.3.1. Technical Committees and how they operate 138
8.3.2. TC21, SC21A , TC35, and Their Standards 139
8.3.3. TC21, SC21A, and TC35: Harmonization of Standards 139
8.3.3.1. IEC Vocabulary on Batteries . . 139
8.3.3.2. Harmonization of the Nomenclature Systems 140
8.3.3.3. Designation of Batteries, Present Status 140
8.3.3.4. IEC Nomenclature Systems for Batteries 140
8.3.4. Closing Remark 143
8.4. Annex 1: IEC Standards on Batteries 143
8.5. Annex 2: Designations by Manufacturers 144
8.6. Annex 3, Examples of IEC Nomenclature Systems for Batteries 144
8.7. Annex 4, Denoting Electrochemical Systems 145
8.8. Example 1: 145
8.9. Example 2: 145
9. Battery Disposal, Environmental Protection 146
9.1. Programs to Collect Disposed Batteries 146
9.1.1. The ISO Recycling Sign 147
9.1.2. Sorting of the Collected Batteries 147
9.2. Recycling of Batteries 148
9.2.1. Recycling of Lead – Acid Batteries 149
9.2.2. Recycling of Nickel – Cadmium Batteries 150
9.2.3. Recycling of Nickel – Metal Hydride Batteries 151
9.2.4. Recycling of Lithium-Ion Batteries 151
9.2.5. Recycling of Primary Batteries . 151
9.2.5.1. Recycling of Mercury-Containing Batteries 151
9.2.5.2. Recycling of Mercury-Free Batteries 152
10. References 152

1. Introduction

Batteries and fuel cells are electrochemical energy converters that directly convert chemical into electrical energy. In batteries the energy is generally stored within the electrodes. But there are exceptions, e.g., the air – zinc system, where one of the reacting substances is continuously supplied from outside, or the zinc – bromine system, where the active material is stored in separate tanks. *Fuel cells* receive their chemical energy from outside sources (→ Fuel Cells).

Batteries can be recharged, if they represent reversible systems. This, however, often is restricted, since not only the system's electrochemical reactions must be reversible, but also the structure of its electrodes. As a consequence, two basically different battery systems exist, i.e., primary and secondary batteries:

1) *Primary batteries* are designed to convert their chemical energy into electrical energy only once.
2) *Secondary batteries* are reversible energy converters and designed for many cycles, i.e., repeated discharges and charges.

The border between both types is blurred, since some primary battery systems permit charging under certain, but usually limited conditions. In general, the efficiency of their recharging process is poor and often results in early failure. The IEC therefore does not recommend the charging of any primary system [1], [2].

Batteries are of increasing importance in all areas of our daily life. They are fabricated in very large ranges of system, volume, weight, and construction. Batteries may be purchased as small button cells, that may have volumes of well below 1 cm^3 to operate a wrist watch, a pocket calculator, a hearing aid, or another small electronic device. Other types of batteries power large electric vehicles including trains and ships, and may have a volume of the order of cubic meters and a weight of several tons. In between a huge variety of batteries exists, designed for all kinds of applications. To meet all needs of the market, batteries offer a large range of optimized properties regarding electric energy, voltage, power output, self-discharge, environmental compatibility, and price.

The term "direct energy conversion" indicates that also nondirect energy conversion systems exist. A famous example is the combustion engine, which converts chemical into mechanical energy via a thermal process (Carnot process) and which may serve for a simplified comparison. Both systems are characterized by a reducing agent and an oxidizing agent. In the case of the combustion engine the reducing agent is the fuel and the oxidizing agent usually atmospheric oxygen. A battery is also characterized by a reducing agent ("fuel") and an oxidizing agent, that serve as energy storage. A well known fuel of this kind is metallic zinc (representing the negative electrode of, e.g., a flashlight battery), and the corresponding oxidant is manganese dioxide, constituting the positive electrode of that battery.

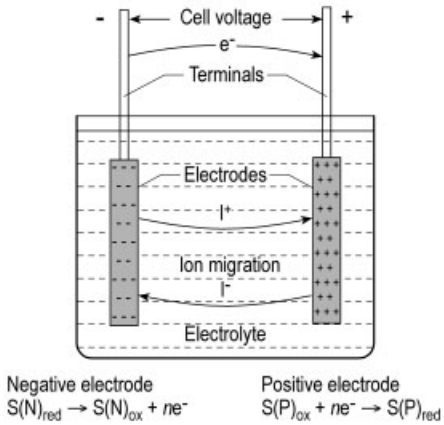

Figure 1. Electrochemical cell and the split-up of the cell reaction
$S(N)_{red}$ and $S(P)_{ox}$ are the reacting components of the negative and the positive electrode, respectively. The shown reactions apply when the battery is discharged.

In principle, direct and nondirect chemical energy conversion are comparable: In both cases the reductant is oxidized under consumption of the oxidant, the latter being stored inside or outside the system (e.g., air). However, the direct energy conversion is not restricted by the limited efficiency of the Carnot process of the combustion engine. In the battery the reaction is split up into one that releases electrons (which means oxidation), and another one that absorbs electrons (which means reduction) (see Fig. 1). The electron exchange that is connected with these reactions, can so be transformed directly into a

current that flows through the electrical appliance (or the charging device).

In principle, a large range of "fuels" and "oxidants" exists for electrochemical energy storage and conversion, each combination representing a new battery system. However, not all combinations are meaningful. A careful selection has to be made, also with respect to the type of electrolyte being employed.

2. History

The history of batteries is usually traced back to Luigi Galvani (1737 – 1798), who detected electrical phenomena between different metals in his famous experiments with frog legs. Allessandro Volta (1745 – 1827) built the first primary battery, the so-called Volta pile. The scientific foundation of electrochemical energy storage was derived by Michael Faraday (1791 – 1867).

The early development of *secondary batteries* is connected to names like Gaston Planté (1834 – 1899), Camille Alfons Faure (1840 – 1898), and Henry Owen Tudor (1859 – 1928) in the field of the lead – acid battery, while Thomas Alva Edison (1847 – 1931) and Valdemar Jungner (1869 – 1924) are the two important names in connection with nickel – iron and nickel – cadmium batteries.

In the decades following these discoveries the lead – acid and the nickel – cadmium system were constantly improved. Today's motor cars would be unthinkable without their reliable batteries. Sealed nickel – cadmium batteries supply the energy for many of our portable electronic devices, and valve-regulated lead – acid batteries have gained a large share of the market for standby batteries.

In the 1990s new rechargeable systems appeared on the market. Of special importance are the nickel – metal hydride and the lithium-ion system.

Primary batteries for practical use emerged in the middle of the 19th century. Special reference is made to George Leclanché, who pioneered the 1.5 V manganese dioxide – zinc system with aqueous ammonium chloride as electrolyte. This system was described first by him in 1866 [3]. It is still produced, technically much more advanced by now, of course, but still known as Leclanché or zinc-carbon battery.

Over the years another 1.5 V system became most successful. It is the alkaline version of the manganese dioxide – zinc system, first proposed in 1882. Its development was taken up again by W. S. Herbert [4], [5] about 70 years later. It took, however, another 20 years before the system became commercially available.

Further to be mentioned is the alkaline silver oxide – zinc system. First patents date back to 1883 – 1910 [6]. This 1.5 V battery is an expensive system, which, however, exhibits very favorable electrochemical properties. To become successful it needed a market for small batteries which opened up in 1975 with the introduction of microelectronics (e.g., wrist watches).

Another primary system that deserves mentioning is the air – zinc system. H. Cohen [7] was the first to develop a practical carbon – zinc element with aqueous ammonium chloride as electrolyte and air as oxidant in 1891. The system is still in use. Its main application is in the area of low power output over a long period of time (e.g., electric fence controllers).

Special future promise has the alkaline air – zinc version. C. Féry was obviously the first [8] to realize this system in 1925. Owing to further development the alkaline zinc – air element offers the highest energy density of all currently available battery systems. Since the alkaline air – zinc battery is also an environmentally benign system, it has replaced the mercuric oxide – zinc system in important application areas, e.g., hearing aids. After activation, however, its storage behavior is limited, depending on construction (e.g., air access) and environment (e.g., humidity).

Finally there are primary battery systems using nonaqueous electrolytes, i.e., lithium batteries. Research activities to develop high energy lithium systems started in about 1960, especially in the USA and soon spread worldwide. Of the many lithium systems developed since the 1960s, about ten systems have attained market relevance (see Section 6.2).

For a more detailed description of the history of batteries, see, e.g., [9–11].

3. Fundamental Laws

Direct conversion of chemical into electrical energy requires the split-up of the electrochemical reactions into reactions of reductants that deliver electrons and reactions of oxidants that accept electrons. Such a split-up is achieved in the electrochemical cell, shown in Figure 1. A positive and a negative electrode are immersed in the electrolyte and the reacting substances (the active materials) are usually stored in the electrodes and the electrolyte, if it participates in the overall reaction.

During discharge, oxidation of $S(N)_{red}$ occurs at the negative electrode according to

$$S(N)_{red} \longrightarrow S(N)_{ox} + n \cdot e^-$$

while $S(P)_{ox}$ is reduced at the positive electrode:

$$S(P)_{ox} + n \cdot e^- \longrightarrow S(P)_{red}$$

Both together form the cell reaction

$$S(N)_{red} + S(P)_{ox} \longrightarrow S(N)_{ox} + S(P)_{red} + \text{energy}$$

When the battery is charged, this reaction is reversed and energy has to be supplied to the cell.

The arrangement shown in Figure 1 resembles an electrolytic capacitor where also two electrodes are separated by the electrolyte. However, charging and discharging of such a capacitor involves only charge shifting within the double layer at the electrode – electrolyte interface. Chemical reactions do not occur and the physical structure of the electrodes is not affected. For this reason, a nearly unlimited number of charge – discharge cycles are possible. Since mass transport does not occur, charge and discharge of a capacitor occur extremely fast, but the amount of stored energy per weight or volume is comparatively small.

In batteries, such a double layer also exists, and the large surface area of the active material gives rise to a high double-layer capacitance that is observed when impedance measurements are made. The double-layer capacitance is caused by charge shifting and is given in the unit Farad, i.e., As/V. The real battery capacity, however, is based on chemical reactions and measured as As or Ah. (Actually, the double-layer capacity is always involved when a battery is charged or discharged). Each charge – discharge cycle changes the physical structure of the electrodes, and these changes inevitably cause an aging processes. Thus, with a battery the number of possible charge – discharge cycles is inevitably limited, and performance changes during service live are unavoidable.

The *cell reaction* characterizes each battery system and its components that represent the charged and discharged states determine the amount of energy that can be stored. As a consequence, the parameters of its cell reaction are of paramount importance for the battery system. While for primary batteries only reactions are relevant that describe the open circuit voltage and the discharge behavior, for rechargeable batteries these reactions must be reversible.

In the following, a brief survey is given of the most important rules. For details and derivations, see textbooks of electrochemistry or fundamental books on batteries, e.g., [12–17].

3.1. Equilibrium or Thermodynamic Parameters

The laws of thermodynamics generally apply to the state of equilibrium, which means that all reactions are balanced. In the electrochemical cell, equilibrium data can only be measured when no current flows through the cell or its electrodes. On account of this balance, the thermodynamic parameters do not depend on the reaction path but only on the difference in energy levels between the final and initial components of the system (the products and the reactants of the electrochemical reaction). The thermodynamic parameters describe the maximally achievable performance data. As soon as current flows through the cell, these values are lowered, owing to kinetic parameters.

The main thermodynamic parameters of an electrochemical reaction are:

1) The *enthalpy of reaction* ΔH which represents the amount of energy released or absorbed. ΔH describes the heat that is generated, provided that 100 % of the chemical energy is converted into heat.
2) The *free enthalpy of reaction* ΔG, also called *change of Gibbs free energy* ΔG, which describes the (maximum) amount of chemical energy that, under equilibrium conditions,

can be converted into electrical or mechanical energy and vice versa.

3) *The entropy of reaction* ΔS which characterizes the reversible heat effect $T \cdot \Delta S$ (where T = temperature in kelvins) which describes the energy loss or gain connected with any chemical or electrochemical process under equilibrium conditions.

Important relations between the three parameters are:

$$\Delta G = \Delta \mathrm{H} - T \cdot \Delta S \text{ or } \Delta H - \Delta \mathrm{G} = T \cdot \Delta S \quad (3.1)$$

The difference between ΔH and ΔG, the product $T \cdot \Delta S$, is the aforementioned reversible heat effect. $T \cdot \Delta S$ can be positive or negative. In the first case, additional energy is generated by cooling of the environment (Peltier or heat pump effect). Otherwise, $T \cdot \Delta S$ constitutes an additional heat increment.

The equilibrium voltage U^{o} is given by

$$U^{\mathrm{o}} = -\frac{\Delta G}{n \cdot F} \quad (3.2)$$

where n is the number of exchanged electronic charges; F the Faraday constant (96 485 As/equivalent; i.e., 26.802 Ah/equivalent); $n \cdot F$ denotes the amount of electrical charge connected with the reaction; and $n \cdot F \cdot U^{\mathrm{o}}$ describes the generated electrical energy (kJ). Thermodynamic quantities like ΔH and ΔG depend on the concentrations (or, more accurately, activities) of the reacting components, as far as these components are in solution. The corresponding relation is:

$$\Delta G = \Delta G^{\mathrm{o,s}} + R \cdot T \cdot \left(\sum \ln \left[(a_i)^{j_i} \right]_{\mathrm{prod}} - \sum \ln \left[(a_i)^{j_i} \right]_{\mathrm{react}} \right) \quad (3.3)$$

where a_i is the activity of the reacting component i (approximately its concentration) in mol/cm^3); j_i is the number of equivalents of this component that take part in the reaction; R is the molar gas constant for an ideal gas (R = 8.3145 JK^{-1} mol^{-1}); $\Delta G^{\mathrm{o,s}}$ is the standard free enthalpy value when all activities are unity. The terms "reac" and "prod" designate in the formula reactants and products for the forward (spontaneous) reaction. Combination of Equation (3.2) and Equation (3.3) results in the Nernst Equation:

$$U^{\mathrm{o}} = U^{\mathrm{o,s}} - \frac{R \cdot T}{n \cdot F} \cdot \ln \frac{\Pi \, (a_i)^j_{\mathrm{react}}}{\Pi \, (a_i)^j_{\mathrm{prod}}} \quad (3.4)$$

which is simplified for 25 °C (298.2 K) to

$$U^{\mathrm{o}} = U^{\mathrm{o,s}} - \frac{0.0592}{n} \cdot \log \frac{\Pi \, (a_i)^j_{\mathrm{react}}}{\Pi \, (a_i)^j_{\mathrm{prod}}} \quad (3.5)$$

under consideration that

$$\ln(..) = 2.303 \log(..),$$

$$R = 8.3145 JK^{-1} mol^{-1} = 8.3145 WsK^{-1} mol^{-1},$$

and

$$F = 96485 As/equiv.$$

thus

$R \cdot T/F$ = 0.02569 V · equiv. · mol^{-1} (T = 298.16 K)

or

2.303 · $R \cdot T/F$ = 0.0592 V · equiv. · mol^{-1}.

The thermodynamic data also determine the temperature coefficient of the equilibrium cell voltage according to:

$$\frac{\mathrm{d}U^{\mathrm{o}}}{\mathrm{d}T} = -\frac{\Delta S}{n \cdot F} \quad (3.6)$$

Single Electrode Potentials. Thermodynamic calculations are always based on an electrochemical cell reaction and the derived voltage and its temperature coefficient in fact concern the voltage difference between two electrodes. The voltage difference between the electrode and the electrolyte, the *absolute potential*, cannot be determined exactly, since potential differences can only be measured between two electronic conductors [18]. Thus, single electrode potential always means the cell voltage between the electrode in question and a reference electrode. To provide a basis for the electrode-potential scale, the zero point was arbitrarily equated with the potential of the *standard hydrogen electrode* (SHE), a hydrogen electrode under specified conditions at 25 °C [19].

Hydrogen reference electrodes are not used in battery practice. They are not only impracticable to handle but also involve some risk

of contamination of the battery's electrodes by the noble metals like platinum or palladium used as electrode materials in the hydrogen electrode [20]. Therefore, a number of reference electrodes are used instead, e.g., the *mercury/mercury(I) sulfate reference electrode* (Hg/Hg_2SO_4) in lead – acid batteries, and the *mercury/mercury(II) oxide reference electrode* (Hg/HgO) in alkaline solutions (cf., e.g. [21]). In lithium-ion batteries with organic electrolytes, the electrode potential is mostly referred to that of the lithium electrode.

3.2. Current Flow, Kinetic Parameters

When current flows, the cell reaction must occur at a corresponding rate. This means that electron transfer has to be forced into the desired direction, and mass transport is required to bring the reacting ions to the electrode surface or carry them away. To achieve this current flow, additional energy is required. This energy finds its expression in *overvoltages*, i.e., deviations from the equilibrium voltage (sometimes denoted as "irreversible entropy loss" $T \cdot \Delta S_{\mathrm{irr}}$). Furthermore, current flow through conducting elements causes *ohmic voltage drops*. Both mean irreversible energy loss and a corresponding heat generation, caused by current flow.

As a result, the voltage U under current flow i is reduced (on discharge) or increased (secondary cellon charge) compared to the equilibrium value U°. The difference $U - U^{\circ}$, when measured as deviation from the cell voltage comprises:

1) The overvoltage, caused by electrochemical reactions and concentration deviations on account of transport phenomena.
2) The ohmic voltage drops, caused by the electronic as well as the ionic currents in the conducting parts including the electrolyte.

The sum of all the voltage drops caused by the current flow is called *polarization*, i.e.,

$$\text{polarization} = \text{overvoltage} + \text{ohmic voltage drops} \tag{3.7}$$

The quantity determined in practice is always polarization. Overvoltage can be determined separately only by special electrochemical methods.

Usually, the reaction path consists of a number of reaction steps that precede or follow the actual charge transfer step as indicated in Figure 2.

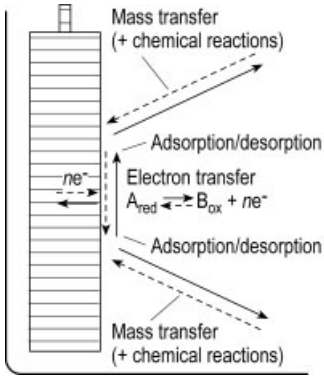

Figure 2. Course of an electrochemical reaction. Charge transfer often can only occur with adsorbed species, then adsorption/desorption steps are included. Furthermore, chemical reactions may precede or follow the electron transfer step.

The rate of each of these reaction steps is determined by kinetic parameters, such as exchange current density, diffusion coefficients, or transport numbers. The slowest partial step of this chain is decisive for the rate of the overall reaction. As a consequence, overvoltages, or even limitations of the reaction rate, often are not caused by the electron-transfer step itself, but by preceding or following steps.

Transport of the reacting species is achieved by two mechanisms, diffusion and migration, and, when only one-dimensional transport is assumed, is given by:

$$N_j = \frac{i_j}{n \cdot F} = -D_j \frac{\partial c_j}{\partial \zeta} + \frac{i \cdot t_j}{z_j \cdot F} \tag{3.8}$$

where N_j is the flux of species j in mol/cm^2; i_j/nF the current equivalent; c_j the concentration of species j in mol/cm^3; $\delta c_j/\delta\zeta$ the concentration gradient in mol/cm^4; D the diffusion coefficient in cm^2/s; t the transference number; z_j the valence number (charges per ion i); and ζ the diffusion direction (cm). Addend 1 describes mass transport by diffusion, which is proportional to the (negative) gradient of the concentration $\mathrm{d}c_j/\mathrm{d}\zeta$ that often can be approximated as a linear slope $\Delta c = c_j - c_{\mathrm{j,o}}$ where c_j is the concentration of the concerned species at the electrode surface, while $c_{j,\mathrm{o}}$ denotes the corresponding concentration in the bulk of the electrolyte.

Diffusion can limit the reaction rate when the mass transport by diffusion precedes the electron transfer step. If the concentration c_j attains the value zero at the electrode surface, all the arriving ions or molecules are charged or discharged immediately, and further increase of the current by increased overvoltage is not possible. Thus the current cannot exceed a certain value, the *limiting current* (i_{lim}) that can be derived from the first term in Equation (3.8) with $c_j = 0$.

Addend 2 in Equation (3.8) describes ion transport by migration by the ionic current. The transference number t denotes the share of the total current that is carried by the corresponding ionic species. In a binary electrolyte, dissociated into A^+ and B^-, the transference numbers are related by

$$t_+ + t_- = 1 \tag{3.9}$$

Transference numbers depend on the concentration of the ions and on temperature. In binary-salt solutions they are fairly close to 0.5, which means that both ion species more or less equally share in ion conductivity. Larger deviations are observed in acids and bases on account of the much higher ion mobility of H^+ and OH^- ions. The values for the battery electrolytes sulfuric acid (dissociated into H^+ and HSO_4^-) and potassium hydroxide are:

Sulfuric acid	$t_+ = t_{\mathrm{H}^+} = 0.9$	$t_- = t_{\mathrm{HSO}_4^-} = 0.1$
Potassium hydroxide	$t_+ = t_{\mathrm{K}^+} = 0.22$	$t_- = t_{\mathrm{OH}^-} = 0.78$

These transference numbers for sulfuric acid, given in [19, p. 255] for diluted solutions, are also true for concentrations used in batteries. The numbers for potassium hydroxide are valid for a wide concentration range [22].

The transference number indicates how much the concentration of the concerned ion is changed by migration due to the current flow. The small value for the HSO_4^- ion means that its concentration is only slightly influenced by migration. In lithium ion batteries, where lithium ions swing between the negative and the positive electrode, a transference number $t_{\mathrm{Li}} = 1$ would be desirable, since then a constant concentration profile would be maintained during discharging and charging. This is one reason to aim at utilization of conducting salts with large anions in such systems as mentioned in Section 7.2.1.4.4.

3.2.1. Electron Transfer

The electron transfer reaction denotes the central reaction step where electric charge is exchanged (cf. Fig. 2). Current flow affords additional forces because of an energy barrier that has to be surmounted by electrons. The required additional energy is the activation energy and the dependence of reaction rates on the activation energy is expressed by the Arrhenius equation

$$k = k_{\mathrm{o}} \cdot \exp\left(-\frac{E_{\mathrm{A}}}{R \cdot T}\right) \tag{3.10}$$

where k is the reaction constant and E_{A} the activation energy (J/mol). In electrode reactions, $n \cdot U \cdot F$ is the driving force, and the corresponding relation is:

$$i = k\prime \cdot c_j \cdot \exp\left(\frac{n \cdot F}{R \cdot T} U\right) \tag{3.11}$$

k' includes the "equivalence factor" $n \cdot F$ between mass transport and current i, U is the electrode potential, and c_j the concentration of the reacting substance that releases or absorbs electrons. Electron transfer, however, does not occur only in one direction: the reverse reaction is possible as well, and the balance between both depends on the electrode potential. Thus, Equation (3.11) has to be completed.

$$i = k_+ \cdot c_{\mathrm{red}} \cdot \exp\left(\frac{\alpha \cdot n \cdot f}{R \cdot T} U\right) - k_- \cdot c_{\mathrm{ox}} \cdot \exp\left(-\frac{(1-\alpha) \cdot n \cdot F}{R \cdot T} U\right) \tag{3.12}$$

where addend 1 describes the anodic reaction (e.g., $Pb \rightarrow Pb^{2+}$), addend 2 its reversal; a denotes the transference factor (usually ≈ 0.5) that indicates how symmetrically the reaction and its reversal depend on electrode potential (difference in activation energies); n is the number of charges, and c_{red}, c_{ox} are the concentrations of the reduced and oxidized states of the reactants. Such an electron transfer occurs also during equilibrium at equal rates in both directions as a dynamic equilibrium, although no current flow is observed through the electrodes. The equilibrium voltage (U^{o}) is determined by the point at which the forward and reverse reaction rates equal each other.

At the equilibrium potential $U = U^{o}$, the current flow in both directions is balanced, which means

$$i_{+}(0) = -i_{-}(0) = i_{o}$$

This balancing current is called *exchange-current density* (it is necessarily related to the surface area, therefore it is a current density given, e.g., in mA/cm^2. In battery practice, such current densities often are referred to the nominal capacity (Ah or 100 Ah, because the true surface area of the porous material is not known, but there is a certain relation between surface area and nominal capacity). Often the current-voltage curves are related to the deviation from the equilibrium potential, the overvoltage $\eta = U - U^{o}$. This leads to the usual form of Equation (3.12):

$$i = i_{o} \left[\exp\left(\frac{\alpha \cdot n \cdot F}{R \cdot T}\eta\right) - \exp\left(-\frac{(1-\alpha)\cdot n \cdot F}{R \cdot T}\eta\right)\right] \quad (3.13)$$

where i_{o} is the exchange current density that characterizes the dynamic equilibrium as shown in Figure 3. The resulting current is represented in Figure 3 by the solid curve as the combination of anodic and cathodic current-voltage curves.

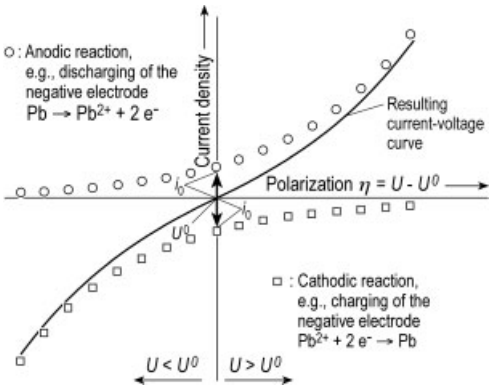

Figure 3. The current voltage curve. The horizontal axis (abscissa) represents the overvoltage $\eta = U - U^{o}$; The vertical axis (ordinate) the current density i, which is synonymous to the reaction rate; i_0 is the exchange current density that characterizes the dynamic equilibrium. The reaction of the lead electrode is inserted as an example.

Polarization. Polarization is an important parameter given by the relations

$$\eta_{+} = U_{+} - U_{+}^{o} \text{ or } \eta_{-} = U_{-} - U_{-}^{o} \quad (3.14)$$

where η_{+} and η_{-} are the polarization of the positive and negative electrodes, respectively; U_{+} and U_{-} are the actual potentials; and U_{+}^{o} and U_{-}^{o} the equilibrium potentials of the positive and negative electrodes. The cell voltage $U_{+} - U_{-}$ results in

$$U = U^{o} + \eta_{+} - \eta_{-} \quad (3.15)$$

According to the definition in Equation (3.14), the polarization of the negative electrode has a negative sign when the electrode potential is below its equilibrium value. If the cell voltage is considered, η_{+} and η_{-} are summed up to η.

Polarization is a very important parameter. The negative electrode is only kept fully charged when its polarization is negative or zero ($\eta_{-} \leq 0$) while for a charged positive electrode a positive polarization is required ($\eta_{+} \geq 0$).

3.2.2. Tafel Lines

If the potential is shifted far enough from the equilibrium value, the reverse reaction can be neglected in Equation (3.13). Then the resulting current – voltage curve in Figure 3 becomes a simple exponential function

$$i = i_{o} \cdot \exp\left[\frac{\alpha \cdot n \cdot F}{R \cdot T} \cdot \eta\right] \quad (3.16)$$

This equation can be rearranged into

$$\eta = \frac{R \cdot T}{\alpha \cdot n \cdot F} \cdot \ln(|i|) - \frac{R \cdot T}{\alpha \cdot n \cdot F} \cdot \ln(|i_{o}|) \quad (3.17)$$

that can be written in a form known as the Tafel equation (J. Tafel was the first to describe this relation in connection with hydrogen-overvoltage measurements on noble metals [23])

$$\eta = a + b \cdot \log(|i|) \quad (3.18)$$

The curves represented by Equation (3.18) are linearized when plotted semi-logarithmically and are called *Tafel lines*. The constant b represents the slope of the Tafel line and denotes the potential difference that causes a current increase of one decade. Tafel lines are important tools when reactions are considered that occur at high overvoltages, since such a linearization allows quantitative considerations (cf. Fig. 65).

3.2.3. Influence of Temperature

Temperature is an important parameter for all electrochemical reaction rates. In general, the approximation can be used that the rates of electrochemical reactions like hydrogen evolution, oxygen evolution, and corrosion are doubled per temperature increase of 10 °C. Then the corresponding Tafel lines are shifted upwards in parallel.

3.3. Heat Effects

Like chemical reactions electrochemical reactions are always connected with heat effects, determined by the reversible heat effect, mentioned in Equation (3.1). When current flows through the cell, additional heat is generated by overvoltages due to kinetic parameters and by ohmic resistances in the electrodes. The sum of both is the polarization that causes the Joule effect (Section 3.3.2).

3.3.1. Reversible Heat Effect

The reversible heat effect

$$Q_{\mathrm{rev}} = T \cdot \Delta S \tag{3.19}$$

represents an unavoidable heat absorption or heat emission that is connected with electrochemical reactions. It is related to the thermodynamic (equilibrium) parameters of the concerned reaction, and is strictly connected with the amount of material (in electrochemical equivalents) that reacts. Thus, the reversible heat effect does not depend on discharge or recharge rates. When the cell reaction is reversed, the reversible heat effect is reversed too, i.e., it acquires the opposite sign. Thus, energy loss in one direction means energy gain when the reaction is reversed. The reversible heat effect per unit time can be related to current flow, because each multiple of the cell reaction requires the current amount $n \cdot F$:

$$\frac{\mathrm{d}Q_{\mathrm{rev}}}{\mathrm{d}t} = \frac{Q_{\mathrm{rev}}}{n \cdot F} \cdot i \tag{3.20}$$

Q_{rev}/nF has the unit volts. So it is equivalent to a voltage, although it is not a voltage that can be measured, but for caloric evaluations it is convenient to combine U^{o} with Q_{rev}/nF. The difference

$$U_{\mathrm{cal}} = U^{\mathrm{o}} - \frac{Q_{\mathrm{rev}}}{n \cdot F} \tag{3.21}$$

is called calorific voltage (or *thermoneutral* potential E_{tp} [24]). U_{cal} is a hypothetical voltage that includes the reversible heat effect, and is used instead of the equilibrium voltage for caloric calculations. The combination with Equations (3.1) and (3.2) shows that U_{cal} also can be written as

$$U_{\mathrm{cal}} = -\frac{\Delta H}{n \cdot F}$$

3.3.2. Current Related Heat Effects (Joule Effect)

Current flow through any conducting object generates heat proportional to the voltage drop caused by the current itself according to

$$\mathrm{d}Q_{\mathrm{Joule}}/\mathrm{d}t = \Delta U \cdot i \tag{3.22}$$

This heat is called the Joule effect; it always means loss of energy.

Strictly speaking, the negative absolute value should be used in Equation (3.22) for consistency with the arithmetical sign of the thermodynamic parameters (lost energy always has a negative sign).

In an electrochemical cell, the voltage drop caused by the current is represented by the difference between the cell voltage under current flow U and the open-circuit cell voltage U^{o}. Then the Joule effect reads according to Equation (3.22):

$$\mathrm{d}Q_{\mathrm{Joule}}/\mathrm{d}t = (U - U^{\mathrm{o}}) \cdot i \tag{3.23}$$

or its integrated form for a period t (in hours)

$$Q_{\mathrm{Joule}} = \int_0^t \{(U - U^{\mathrm{o}}) \cdot i\} \, \mathrm{d}t \tag{3.24}$$

Strictly speaking Equation (3.23) describes a composite parameter of

1) The "genuine" Joule effect $dQ/dt = i^2 \cdot R$, which describes the heat generation by current flow (i) through the ohmic resistance (R) of the conducting elements and the electrolyte, and

2) The irreversible heat effect $T \cdot \Delta S_{irr}$, which causes voltage drops between the electrode surface and the electrolyte on account of reaction hindrance (overvoltage according to Eq. 3.7). This additional heat is caused by an additional entropy of reaction (ΔS_{irr}) which cannot be recouped when the reaction is reversed.

For heat effects this is not relevant, since heat generation is proportional to polarization.

$U - U^{\mathrm{o}}$ does not remain constant during charging or discharging, because it is related to the internal resistance, that usually increases with proceeding discharge, because of the lower conductivity of the discharged products. In lead – acid batteries the conductivity of the electrolyte also decreases with discharge, and acid dilution also influences U^{o}. In quite a number of battery systems the U^{o}-value is also influenced by the state of charge. An example is the manganese dioxide – zinc system [25].

3.3.3. Heat Generation in Total

Summation of the Joule effect and the reversible heat effect gives the total heat generated in the cell or the battery, i.e.,

$$Q_{\mathrm{total}} = Q_{\mathrm{Joule}} + Q_{\mathrm{rev}} \tag{3.26}$$

as energy, e.g., Wh, or as energy per time unit, i.e., power, e.g., W,

$$\frac{\mathrm{d}Q_{\mathrm{total}}}{\mathrm{d}t} = \frac{\mathrm{d}Q_{\mathrm{Joule}}}{\mathrm{d}t} + \frac{\mathrm{d}Q_{\mathrm{rev}}}{\mathrm{d}t} \tag{3.27}$$

Depending on the sign of $\mathrm{d}Q_{\mathrm{rev}}/\mathrm{d}t$, the total energy generation may be larger or smaller than the Joule effect.

According to Equation (3.23), $\mathrm{d}Q_{\mathrm{Joule}}/\mathrm{d}t$ can be substituted by $(U - U^{\mathrm{o}}) \cdot i$, which gives:

$$\frac{\mathrm{d}Q_{\mathrm{total}}}{\mathrm{d}t} = (U - U^{\mathrm{o}}) \cdot i + \frac{\mathrm{d}Q_{\mathrm{rev}}}{\mathrm{d}t} \tag{3.28}$$

Substitution of $\mathrm{d}Q_{\mathrm{rev}}/\mathrm{d}t$ by Equation (3.20) and application of Equation (3.21) results in the simple relation that is convenient for heat calculations:

$$\frac{\mathrm{d}Q_{\mathrm{total}}}{\mathrm{d}t} = (U - U_{\mathrm{cal}}) \cdot i \tag{3.29}$$

Strictly speaking, Equation (3.29) should have a negative sign according to thermodynamic parameters, since the Joule effect is always lost energy, as mentioned in connection with Equation (3.22). To overcome this difficulty, $Q_{\mathrm{gen}} = -Q_{\mathrm{total}}$ is introduced below.

3.4. Heating of the Battery

While a battery is being charged or discharged, the heat generation caused by the flowing current raises the temperature until balance is achieved between heat generation in the cell and heat dissipation to the environment. This balance depends on both: the rate of heat generation and the possibility of heat dissipation.

Heating Effects, Heat Capacity. The change of battery temperature during discharge ($\mathrm{d}T/\mathrm{d}t$) can be estimated by heat generation values:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{1}{C_{\mathrm{Batt}}} \cdot \left(\frac{\mathrm{d}Q_{\mathrm{gen}}}{\mathrm{d}t} - \frac{\mathrm{d}Q_{\mathrm{diss}}}{\mathrm{d}t} \right) \tag{3.30}$$

where $\mathrm{d}Q_{\mathrm{gen}}/\mathrm{d}t$ is the generated energy per unit time and $\mathrm{d}Q_{\mathrm{diss}}/\mathrm{d}t$ the dissipated energy per unit time; Q_{gen} corresponds to Q_{total} in Equation (3.29). C_{Batt} is the heat capacity of the battery (in kJ/K) defined by

$$\sum [m(j) \cdot C_p(j)] = C_{\mathrm{Batt}} \tag{3.31}$$

where i denotes one of the various components of the battery; $m(i)$ is the mass of component i in kg; and $C_p(i)$ is the specific heat of component i in kJ kg^{-1} K^{-1}.

4. Battery Properties and Characteristics

Figure 1 shows the basic design of a single cell. When the battery is discharged, the active material in the positive electrode is reduced

$$\mathrm{S(P)_{ox}} + n \cdot \mathrm{e}^- \longrightarrow \mathrm{S(P)_{red}}$$

Such a reducing current flow is called "cathodic".

In the negative electrode, the active material is oxidized

$$\mathrm{S(N)_{red}} \longrightarrow \mathrm{S(N)_{ox}} + n \cdot \mathrm{e}^-$$

by an "anodic" current. The terms cathodic and anodic are strictly connected to the direction of current flow. *Cathodic* means that a negative current (i.e., negatively charged ions) flows from the cathode to the anode via the electrolyte, *anodic* means a positive current (positively charged ions) flowing from the anode to the cathode via the electrolyte, i.e., in opposite direction. Thus, during charging, these currents are reversed. In accordance with these terms, in the field of *primary batteries* the positive electrode is usually called "cathode" since the positive electrode is discharged by a cathodic current, while the negative electrode is called "anode", and this is unambiguous, since only discharge occurs.

In *secondary batteries*, it depends on charge or discharge which of the two electrodes is anode or cathode. Thus, in lead – acid batteries, the negative electrode is anode during discharging ($Pb \longrightarrow Pb^{2+}$), but cathode during charging. The opposite applies to the positive electrode: $Pb^{4+} + 2 \cdot e^- \longrightarrow Pb^{2+}$ is a cathodic reaction, and the PbO_2 electrode is cathode during discharging but anode during charging. Because of this ambiguity, the terms positive electrode and negative electrode are preferred for secondary batteries.

For secondary lithium-ion batteries, however, that partially have been developed commonly with primary lithium batteries, the terms cathode and anode have become customary. Then the terms anode and cathode correspond to the discharging situation.

4.1. Cell and Battery

The basic element of each battery is the cell, corresponding to Figure 1. The term *battery* refers usually to several cells (at least two) being connected in series or in parallel. The International Electrotechnical Commission IEC, however, has meanwhile extended the term "battery" to also include single cells, if they have terminal arrangements, such that they can be placed into a battery compartment while providing adequate contact. Furthermore the single cell must carry markings as required by the IEC Standard in order to be called a battery [26], [27].

According to IEC the term *cell* still applies when referring to the component cells inside a multicell battery, like, e.g., a 9 V battery. Terms like smart batteries, battery packs, etc., are discussed in Sections 5.2.1 and 7.2.1.4.7.

4.1.1. Active Material

The term active material refers to the components of the cell reaction being used up or regenerated in consequence of a discharge or charge. The term active material includes in the first place materials in the positive and negative electrode, but may also include certain components of the electrolyte.

4.1.1.1. Change of Volume

The active material changes its volume on charge and discharge. When the volume of the active material is increased, mechanical forces may be generated inside the battery that even may disrupt the battery casing, if no volume for expansion is provided. Reduction in volume, on the other hand, can cause contact problems between the electrodes and the separator that contains the immobilized electrolyte unless there is a mechanical component like a spring that provides pressing forces.

For example, during discharge of the alkaline air – zinc system, its zinc anode reacts with the oxygen of the air according to the overall reaction

$$Zn + 1/2\,O_2 \longrightarrow ZnO$$

The molar volume of ZnO is ca. 65 % larger than that of metallic zinc, and the battery construction must provide adequate space to allow for this change in volume, which is actually caused by the intake of atmospheric oxygen. Otherwise the battery would swell during discharge.

Other electrochemical systems exhibit a reduction in partial molar volume in consequence of discharge, like the monovalent silver oxide-zinc system. According to the overall reaction

$$Ag_2O + Zn \longrightarrow 2\,Ag + ZnO$$

its partial molar volume decreases by ca. 15 %. The CuO – Li system exhibits even a reduction in partial molar volume of ca. 42 %.

4.1.1.2. Change of Electrode Structure

The solubility of reaction products is a very important parameter for electrode reactions that occur via dissolution of the reactants, as in the case of lead, cadmium, zinc, or lithium electrodes. If the product or the discharge is highly soluble the electrode will – to a large extent – be dissolved during discharge and will lose its former structure. This leads to problems during recharge because the redeposition of the metal is favored where the concentration of the solution has its highest value. As a consequence, the structure of the electrode will be changed as demonstrated in the upper line of Figure 4.

Connected to the shape change is a further drawback of the high solubility, namely the tendency that during recharging the precipitated material forms dendrites that may penetrate the separator and reach the opposite electrode, thus gradually establishing a short circuit.

A typical example of this situation is the zinc electrode, which allows only limited discharge-charge cycles before it starts to establish electron conducting bridges between the electrodes. Zinc electrodes are therefore not used for commercial secondary batteries with the exception of the rechargeable alkaline zinc manganese dioxide battery (RAM) [28], [29], which is a within certain limits rechargeable battery of low initial cost, but limited cycle life.

The metallic lithium electrode is another case where cycling causes problems, in this case, however, not only due to shape change but also due to unwanted reactions with the electrolyte (see Sections 6.2 and 7.2.1).

Extremely low solubility of the reaction products leads to immediate precipitation and thus to the formation of an impermeable covering layer (lower line in Fig. 4). As a consequence, the discharge reaction comes to a halt as soon as the passivating layer is completed. To counteract such a passivation, the active material in corresponding technical electrodes, e.g., lead and cadmium electrodes, is used as a spongy structure that has a large surface area of the order of square meters per gram. A further advantage of such a porous structure is that the reaction products are precipitated immediately after their formation within the pores of the active material, close to the place of their origin, and the structure of the electrode remains nearly stable.

4.1.2. Nonactive Components

The split-up of the cell reaction into two electrode reactions as indicated in Figure 1 requires a number of nonactive components. They can be classified into conducting and nonconducting components.

4.1.2.1. Conducting Components

The current has to be collected from the active material and conducted to the terminals. Often the current conductor simultaneously acts as a support for the active material. In some systems the container of the cell is made of metal and often simultaneously acts as terminal. In Leclanché cells the container represents simultaneously the active material of the negative electrode.

When a number of electrodes are connected in parallel within the cell, corresponding connecting parts like pole bridges are required.

Additives, like carbon or metal powder, sometimes are required to improve the conductivity of the active material, especially in thick electrodes, and in electrode materials with low conductivity.

4.1.2.2. Nonconducting Components

Separators. The electronic separation of the two electrodes is another important condition. Any electronic contact between positive and negative electrodes constitutes a short circuit that discharges the cell concerned. On the other hand, the ionic current through the separator should be impeded as little as possible. In the early days of battery development, the widely spaced electrodes were only separated by rods of glass or rubber. In modern batteries, thin sheets of plastic materials, e.g., polyethylene, are used with a fine pore system in the micrometer range that provides a void volume of more than 80 %. Fiber felts of plastic or glass are also used as separators.

In batteries with an internal oxygen cycle, like valve-regulated lead – acid, sealed nickel – cadmium, or nickel – metal hydride batteries, the separator is important also for the

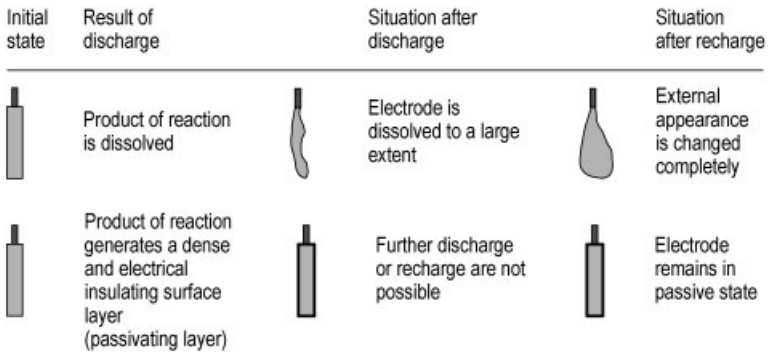

Figure 4. Effect of the solubility of the reaction products on electrode structure, when the discharging-charging mechanism occurs via the dissolved state

functioning of the oxygen cycle: the fine pores of the separator hold the electrolyte while the large pores stay open for fast oxygen transport through the gaseous phase.

In some alkaline batteries ion-conducting films like cellophane are used to prevent gas flow between the cells.

Polymer electrolytes can also be regarded as ion-conducting separators.

A semipermeable membrane that only allows the permeation of sodium ions is the β-alumina in sulfur – sodium or nickel chloride – sodium batteries (Sections 7.2 and 7.2.2.3).

Containers for rechargeable batteries are often made of nonconducting material. In lead – acid batteries it is a must to use glass, rubber, or plastics on account of the high cell voltage that would destroy all metals. Various types of plastic materials are used in the different systems. The advantage of a plastic container is that no insulation is required between adjacent cells. A general drawback of plastic materials is their permeability for gases, water vapor, and volatile substances. For this reason in sealed nickel – cadmium and nickel – metal hydride batteries metal has to be used instead of plastic material.

Seals. The seal of the terminals is a critical element. In *vented batteries* with liquid electrolyte it has to prevent creeping of the electrolyte which is especially observed for the negative terminal of batteries with alkaline electrolyte. With *sealed batteries*, the post seal has to prevent escape of electrolyte or its vapor, and the intake of oxygen from the surroundings. Special techniques have been developed for the different battery systems. Premium cells for space-craft applications but also lithium batteries for long service life often are equipped with metal – glass – metal seals that prevent any transport phenomena because of their glass body.

Vent plugs and valves are required, since a number of battery systems involve secondary reactions that generate gases which must escape. *Vents* as used for some rechargeable batteries are simple openings that allow gas flow in both directions, i.e., from outside into the battery and vice versa. The openings in such vent plugs are small to minimize water loss by evaporation. In modern batteries, such vents mostly are equipped with porous disks to prevent ignition sparks or flames from entering the cell and also the escape of acid or alkaline fumes from the cell.

Valves allow only the escape of gas and are used in most sealed batteries to prevent damage of the cell in the case of a too high internal pressure when the battery is abused, e.g., overcharged at a too high current rate (see Sections 7.1.1.12 and 7.1.2.1.11).

Rechargeable button cells in general have a *rupture vent* (breaking point) embossed in their metallic cell container that opens on a preset overpressure before the cell explodes. Safety features of this type are also employed in primary batteries but may be incorporated in their plastic seal instead.

4.2. Battery Parameters

The discharge/charge behavior of a battery depends on a number of parameters, like current, voltage, and temperature. These parameters have to be specified when such data are compared.

4.2.1. Voltage

In the case of reversible systems the *cell voltage* may be derived from the thermodynamic data of the cell reactions according to Equation (3.2). But often, this equilibrium voltage cannot exactly be measured (even not at open circuit), since secondary reactions cause a slight deviation (see Fig. 9). Then the *open cell voltage* (OCV) is actually measured, which, depending on the system, is often a function of the state of charge (SOC). Other systems, like some primary lithium systems, offer a rather high initial open circuit voltage, obviously due to the presence of higher valence oxides. This voltage collapses by several hundred millivolts on the slightest load and never recovers to its initial value.

The cell voltage under load, the *closed circuit voltage* (CCV) depends on the current, the state of charge, and on the cell's history, like, e.g., its lifetime or storage period.

A further term is the *nominal voltage* of a cell or battery that approximates the voltage of the concerned battery system. It is generally an open circuit voltage, rounded off to achieve a simple voltage designation, e.g., 3 V instead of 3.14 V or 2.86 V.

The *coup de fouet* or *initial voltage minimum* is a voltage minimum as indicated in Figure 5 at the beginning discharge of fully charged lead-acid batteries and is in the range of 20 mV per cell. It is caused by the positive electrode and can be ascribed to crystallization overvoltage [30]. Initial voltage minima are also known for lithium batteries. They are referred to as *initial voltage delay* and are caused by protecting layers being formed on the surface of the negative electrode during storage and being dissolved on discharge.

4.2.2. Capacity

The capacity of a battery is defined by international convention as the electrical charge in units of ampere hours that can be drawn from the battery. Thus, when the battery is discharged with a constant current, its capacity is given by the relation

$$C_{\mathrm{AH}} = i \cdot \Delta t \tag{4.1}$$

A more general definition would be

$$C_{\mathrm{AH}} = \int_0^t i\,(t) \cdot \mathrm{d}t \tag{4.2}$$

The discharge parameters that, besides the design of a battery, mainly influence the capacity are:

- Discharge current,
- Voltage limit that has to be specified and is known under various names: the final, the end point, the cut-off, or end-of-discharge voltage (EOD) and
- Temperature.

Further parameters that also influence the capacity, are the state of charge and the history of the battery, e.g., the preceding storage period. Any comparison of capacity data must always consider these parameters.

4.2.2.1. Secondary Batteries

Figure 5 shows as an example two typical discharge curves of a lead – acid battery. The upper curve represents a low-rate discharge, the lower curve a high-rate discharge.

The broken curve in Figure 5 represents the equilibrium voltage that for a lead – acid battery gradually decreases with progressing discharge on account of acid dilution. In other systems, where the electrolyte is not involved in the cell reaction, e.g., nickel – cadmium or nickel – metal hydride batteries, U^{o} remains approximately constant. The difference between the actual discharge curves and U^{o} is due to polarization according to Equation (3.7). This difference usually increases with progressing discharge because of a gradual increase of the internal resistance.

The *nominal* or *rated capacity* of a battery is specified by its manufacturer as the standard value that characterizes this battery. Usually it is specified for a constant-current discharge at

20 °C or at room temperature. For the various applications, nominal or rated capacities are often referred to different discharge duration, termed as C_{20}, C_{10}, or C_5, respectively, where the subscript denotes the discharge period in hours.

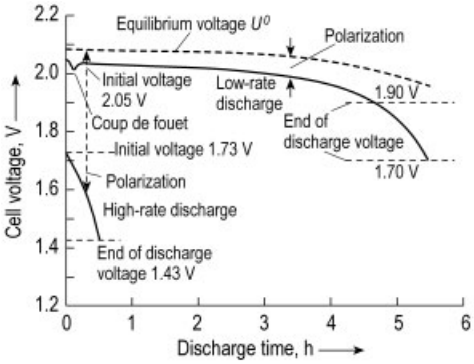

Figure 5. Typical low-rate discharge and high-rate discharge curves of a lead-acid battery and the main discharge parameters. The broken curve denotes the equilibrium or open circuit voltage.

Capacity ratings expressed as multiples of the nominal capacity were initially introduced for nickel – cadmium batteries. But it became common practice to address the current that characterizes charging or discharging of small lead – acid batteries and other rechargeable battery systems also in terms of capacity ratings. In this case the discharge current is written

$$i = m \cdot C_r \tag{4.3}$$

where i is the current in amperes; m is a number that indicates the multiple or fraction of C_r; C_r is the rated or nominal capacity in ampere hours, but expressed in amperes; and r is the discharge period in hours, the capacity is rated to. To indicate that the current is expressed as a multiple of the capacity, it is often written as "CA" or "C(A)". In this case "C" does not mean a capacity in ampere hours but a corresponding current in amperes. In regard to the dimensions, Equation (4.3) is not correct because it mixes quantities with units (i in A, C_r in Ah, m without unit). But in the battery field, this description of discharging and charging currents is well established and understood. It is even used in the standards of IEC/SC21A [31].

The factor m has no relation to discharge time: it is a pure number. For this reason, discharge rates in terms of multiples of the capacity cannot be converted into discharge durations. To evaluate the discharge duration, the discharge curves must be known.

The IEC Standard for lead – acid batteries [36], however, continues to use the corresponding currents, e.g., I_5, I_{10}, I_{20}, that describe the actual discharge duration.

The *depth of discharge* (DOD) is an important parameter in regard to the number of cycles that can be reached with rechargeable batteries. For lead – acid batteries, deep discharges that are continued beyond the recommended maximum DOD, can reduce the service life dramatically [14].

4.2.2.2. Primary Batteries

For primary batteries the term "capacity" as defined by Equations (4.1) or (4.2) is of course also applicable. Two discharge methods are in use:

1) Constant current discharge
2) Constant resistance discharge

Some test houses prefer method (1). By far predominant is, however, method (2) as is evident from manufacturer's technical handbooks. Method (2) prevails also in the IEC Standard for Primary Batteries [26], [33].

The discharge capacity of a battery decreases with increasing load as indicated in Figure 5 for a lead – acid battery. The same applies to primary batteries. The results of constant resistance discharges can be represented in a generalized form, shown in Figure 6: For a given cut-off voltage U_c, the discharge capacity $C(R)$ is plotted versus the discharge resistance R_d, measured in ohms.

Typical for this $C-R$ curve is the sharp increase in capacity C on increase of R_d in the range of low resistance values (high rate range), whereas in the low rate range (i.e., large R_d values) the discharge capacity C becomes finally independent of R_d. This is the battery's capacity plateau for which $dC/dR \approx 0$ is valid [32], i.e., $C \approx$ const. The $C-R$ plot is typical for all batteries. However, depending on the system's self-discharge which acts in parallel to the external discharge current, the plateau value of C_p may decrease again on further increasing R values. The plateau value $C_p \approx$ const is the battery's maximum capacity. $C_p \approx$ const implies

that within the plateau region also the discharge voltage U(plateau) is approximately constant. U(plateau) is independent of size and construction of the battery, and typical for its electrochemical system, because with such a low load, transport reactions are fast enough, and the battery's polarization is only determined by the charge-transfer resistance of its electrodes.

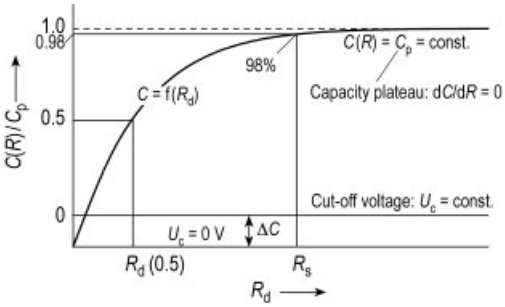

Figure 6. Schematic diagram of discharge capacity C versus discharge resistance R for a given end-point voltage and temperature. For the purpose of normalization $C(R)$ was divided by the plateau capacity C_{p}, that is reached at a sufficient high resistance. For the plateau $C(R)/C_{\mathrm{p}} = 1$ is valid.

Depending on manufacturer, the capacity of primary batteries may be given in terms of a service output duration, when simulating a certain application (e.g., transistor radio test, portable lighting test, toy test, etc.). The service output duration (in hours), as given for the individual application test, is proportional to the battery's capacity. The IEC Standard for Primary Batteries, which so far has proposed most of the existing application tests, quotes only "Minimum Average DurationValues" [26], [33] instead of capacities to describe battery performance. The "Minimum Average Duration Values" are exceeded significantly by at least 90 % of battery products available on the market.

4.2.3. Energy Content

In general form, the energy E that can be drawn from the battery is represented by the product

$$E = \int_0^t U(t) \cdot i(t) \cdot \mathrm{d}t \tag{4.4}$$

where E is the energy content in watt hours; U the voltage in volts; i the discharge current in amperes; and t the discharge period in hours. Direct measurements of the drawn energy are made with constant load (W), which means that the voltage decrease with proceeding discharge has to be compensated by an increase in current.

Usually, the battery is discharged with a constant current and the energy output is calculated by multiplying the measured Ah capacity by the discharge voltage. Characteristic voltages that are used for such an approximation are:

- *Initial discharge voltage*: voltage at the moment the load is applied.
- *Average discharge voltage* or *mean discharge voltage*: approximate average of the voltage during the whole discharge period (average discharge voltage × current = discharged energy).
- *Mid-point discharge voltage*: voltage after 50 % of the capacity has been discharged (≈ average voltage).

For constant resistance discharge as normally carried out for primary batteries, with R = const as resistive load, Equation (4.5) applies

$$E = \frac{1}{R} \int_0^t U(t)^2 \, \mathrm{d}t \tag{4.5}$$

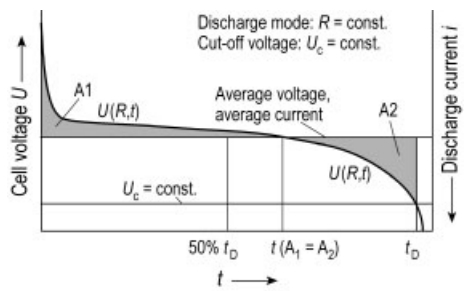

Figure 7. Schematic discharge curve $U(R,t)$ versus discharge time t for a given load resistance R and end-point voltage U_{c}

Figure 7 is a schematic battery discharge curve with load resistance R for a given end-point voltage U_{c} at constant temperature ϑ. The average discharge voltage (and average discharge current) is given by the horizontal line that is defined by equal areas for t(A1 = A2). For a symmetrical curvature this average voltage equals the mid point discharge voltage for t_{D} = 50 %.

Specific Energy and Energy Density. For system comparison it has become common practice to relate the energy content of a given battery either to its volume or to its weight. The *weight-related energy* in watt hours per kilogram is the specific energy. (In the negligent use of such terms it is often also called energy density or gravimetric density, although a density should always be referred to a volume.) The specific energy is of special interest in the field of electric vehicles, since the weight of the battery determines the cruising range of the car. A typical value for a lead – acid traction battery is, e.g., 25 Wh/kg. For comparison: A modern lithium-ion battery offers ca. 125 Wh/kg.

The *volume-related energy density* is given in units of watt hours per liter or milliwatt hours per cubic centimeter. The energy density is of special interest for batteries designed to power portable equipment. In this application the size of the battery is in general more important than its weight. Present systems cover a range of 150 mWh/cm^3 (carbon – zinc) up to 1.2 Wh/cm^3 (alkaline air – zinc).

4.2.4. Internal Resistance and Power Output

The internal resistance of a battery is an important parameter since the power output is proportional to the reciprocal value of its internal resistance. Therefore it permits to differentiate between, e.g., high rate, medium rate, and low rate batteries, which again is of importance in judging the possibilities and limits of a battery's application. A general requirement is that the internal d.c.-resistance of the battery must be significantly below that of the appliance (1/10 and less), otherwise the voltage drop caused by the current demand of the appliance would limit the battery's service output duration too early.

The meaning of the term "internal resistance" has to be considered with some caution, because it is not a simple ohmic resistance and depends on the method that is used for its determination.

The *direct-current method* is frequently applied in battery practice. With this method, the terminal voltage is compared at various loads. The battery is loaded with the current i_1 for a few seconds and the cell or battery voltage U_1 results. Then the current is increased to the value i_2 and the battery voltage is reduced to the value U_2.

The internal resistance R_i can be calculated according to:

$$R_i = \frac{U_1 - U_2}{i_2 - i_1} = \frac{\Delta U}{\Delta i} \tag{4.6}$$

Equation (4.6) represents the usual method to determine an ohmic resistance. Batteries, however, do not represent an ohmic resistance. Their characteristic is nonlinear. This is illustrated in Figure 8, representing a schematic diagram of a quasi stationary current – voltage curve of a battery. According to Equation (4.6) the internal resistance R_i is given by the slope determined by

$$\Delta U/\Delta i = (U_1 - U_2)\,/\,(i_2 - i_1)$$

As illustrated in Figure 8, the value depends on the difference Δi. For infinitesimal values of Δi the internal resistance becomes $R_\mathrm{i} = \mathrm{d}U/\mathrm{d}i$, which is the tangent for a certain current. As is obvious from Figure 8, the slope, i.e., the internal resistance, changes with current i. The higher the current i, the flatter is the slope $\mathrm{d}U/\mathrm{d}i$, i.e., the smaller is the differential internal resistance, until finally the current approaches the limiting current i_lim. Then $\mathrm{d}U/\mathrm{d}i$ increases again. For $i = i_\mathrm{lim}$, the voltage U collapses due to a preceding rate determining process, and the internal resistance becomes infinitely high, $R \to \infty$. An example for such a behavior is the air – zinc system: With increasing current i the transport of oxygen, depending on the diffusion resistance for oxygen, becomes finally rate determining, i.e., the voltage under load collapses.

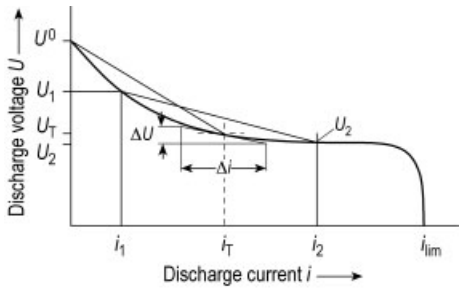

Figure 8. Quasi stationary current – voltage curve of a battery. Its internal resistance R_i is given by $\mathrm{d}U/\mathrm{d}i$ that often can be approximated by $\Delta U/\Delta i$. Such a curve naturally depends on the state of discharge

To achieve values for R_i that are relevant to practical applications, the currents i_1 and i_2 usually are specified and adapted to different battery types (examples, e.g., in [14, p. 90]).

The internal resistance R_i, determined according to Equation (4.6), comprises ohmic resistance within the electrodes and the electrolyte as well as overvoltage at the phase boundaries between the electrodes and their electrolyte. Equation (4.6) implies that the overvoltage is rather small compared to the ohmic voltage drop. For batteries with aqueous electrolyte this is true only for the discharge, not during charging because of the high overvoltage of the gassing reactions.

The *short-circuit current* is of interest especially for larger stationary batteries, since it stands for the maximum current that could be supplied by the battery for a short period of time. Its value helps in estimating the size of fuses that might operate with the battery. The short-circuit current is determined according to Equation (4.6) by extrapolating the resulting line to the voltage zero. The short-circuit current represents a dynamic parameter that decreases quickly with proceeding discharge. Published values are always referred to the fully charged battery.

The *a.c. internal resistance* of a battery is a complex parameter and difficult to interpret, since the a.c. behavior of a battery can only be approximated by an equivalent circuit of many components. In the 1990s, the ohmic part of the impedance gained importance as a possibility to check the situation of valve-regulated lead – acid batteries that otherwise could be determined only by discharging. The impedance is mostly measured at a frequency of 1000 Hz. At such a high frequency it does not give any information on the battery's electrochemistry but only on its ohmic components, including the electrolyte. The resulting values are used for comparison, to detect faulty cells within a string.

Specific Power Output. A performance feature which is of special interest in the field of primary batteries is the battery's specific power output $P*$ (see Eq. 4.7). It depends basically on the following five parameters:

1) Battery size
2) System
3) Internal battery construction
4) State of charge/discharge
5) Temperature

It is evident that the power output of a battery for given parameters 2 – 5 increases with size and vice versa. The change in size, however, does not mean that the battery may finally become a high rate or low rate battery, because of its increased or reduced power output. The terms *high rate* and *low rate* are special cases of the specific power output, which is obtained when the battery's power output in watts is divided by its envelope volume V as defined by the Equation (4.7).

$$P^* = \frac{U^2\,(C/2)}{R\,(C/2) \times V} \tag{4.7}$$

The value for $P*$ depends on the experimental conditions under which it is determined. The range of $P*$ values as quoted below for primary batteries was determined by the following schedule [32]:

1) Discharge fresh battery by 50 % through R_s that would yield 98 % (or above) of plateau capacity $C_p : C(R_1) \geq 0.98 \times C_p$
2) Apply discharge resistor $R(C/2) = R(0.5)$ that yields 50 % of C_p (cf. Fig. 6) and read voltage U after 3 s
3) Calculate $P*$-value according to Equation (4.7)

The approximate range as covered by present primary battery systems for consumer application is 3 mW/cm^3 up to 400 mW/cm^3. The highest values were verified for cylindrical batteries of the MnO_2 – Li system with an organic electrolyte. The large range of these values is partly caused by the different electrochemical systems, but mainly by differences in their internal construction. The first value (3 mW/cm^3) is a low rate value, being typical for a simple "bobbin" type electrode construction (cf. Section 6). The second value (400 mW/cm^3) is typical for a high rate battery construction, using spirally wound electrodes.

Present specific power outputs may be either determined individually or they may be grouped into ranges [34] like

1. Low rate: $P* < 5\,\mathrm{mW/cm^3}$
2. Medium low rate: $5\,\mathrm{mW/cm^3} \leq P* < 20\,\mathrm{mW/cm^3}$
3. Medium rate: $20\,\mathrm{mW/cm^3} \leq P* < 80\,\mathrm{mW/cm^3}$
4. Medium high rate: $80\,\mathrm{mW/cm^3} \leq P* < 320\,\mathrm{mW/cm^3}$
5. High rate: $P* \geq 320\,\mathrm{mW/cm^3}$

Under the conditions quoted above, cylindrical batteries, employing spirally wound electrodes, belong exclusively to the high rate range, i.e., group 5. Groups 1 to 4 serve only to differentiate between lower rate systems, which include most of the primary batteries and some of the rechargeable systems.

4.2.5. Charge Parameters

The charging process basically represents the reversal of the discharge. Theoretically it occurs as soon as the voltage of the cell attains a value just above the equilibrium voltage U^{o}. In practice, however, higher voltages have to be applied to achieve full charge within a desired period. With increased cell voltage, however, parasitic reactions, such as water decomposition in the case of aqueous electrolytes, may gain more significance. Charging methods therefore always have to compromise between charging at a rather low rate with high efficiency and faster but less effective charging. Furthermore, fast charging methods must be controlled, since they may cause thermal problems, and for a number of battery systems, like lithium-ion cells, overcharging must strictly be avoided since it would cause irreversible reactions.

As a consequence, a great number of different charging methods have been developed, adapted to the type of battery and to the requirements of the application.

The charging process is mainly influenced by three parameters:

Charging current,
Charging voltage, and
Temperature of the battery

Charge acceptance or *charge current efficiency* describes the share of the current, that actually is "accepted" by the battery and can be retrieved by subsequent discharging. It is determined by the balance between the kinetic parameters of the charging reaction and the secondary processes that simultaneously may occur.

4.2.6. Self-Discharge

Self-discharge means gradual loss of charge in the positive and/or negative electrode, when the battery is stored at open-circuit. One reason can be the gradual decrease of the oxidation state in the positive electrode, e.g., the loss of oxygen in nickel hydroxide electrodes.

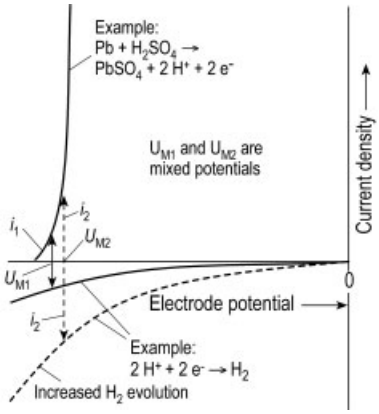

Figure 9. Mixed potential between discharge of the lead electrode and hydrogen evolution with the result of self-discharge. The dashed curve represents reduced hydrogen overvoltage
U_{M} = mixed potential; i_1, i_2 = rate of self-discharge

4.2.6.1. Mixed Potential

Secondary reactions can also cause self-discharge when these reactions are thermodynamically possible at the equilibrium potential and cause discharge of the concerned electrode. Then the secondary reaction and the discharge reaction form a mixed potential as shown in Fig. 9. The two electrochemical reactions compensate each other, and gradual discharge is the consequence (cf. Section 7.1.1.3).

The fairly steep current – voltage curve pointing upwards represents the discharge of the lead electrode. The fast rise of this curve indicates the low overvoltage that characterizes this reaction. It means that high discharge rates can be achieved at low overvoltage or polarization values. The high overvoltage, characteristic of the hydrogen-evolution reaction, is expressed by

[2]
the gradual rise of the corresponding flat current curve. At open-circuit voltage, discharge reaction and hydrogen evolution have to balance each other, because no external current flows through the electrode. The result of this balance is a mixed potential, U_M in Figure 9. As for an equilibrium potential, no external current appears. But the mixed potential is not an equilibrium potential, because two different reactions occur, and gradual discharge of the negative electrode is the result.

The position of the mixed potential is largely determined by the faster one of the two reactions, while the rate of the reaction, the self-discharge in this example, is determined by the slower reaction. This is illustrated by comparison of the continuous and broken curves in Figure 9. The continuous curve represents a fairly high hydrogen overvoltage; the broken curve depicts the case when hydrogen can be evolved more easily. The position of U_M is only slightly changed between the two examples, but the rate of the self-discharge, the current i_2, grows to a multiple of i_1 in the second case.

4.2.6.2. Further Self-Discharge Mechanisms

Self-discharge can also be caused by oxidizable or reducible substances in the electrolyte, when they reach the positive or negative electrode, respectively. Oxidation of the critical substance means a corresponding reduction of the active material within the electrode and vice versa.

Such an effect is called a "shuttle" when the oxidized and reduced form of the critical substance remain dissolved and alternatively are oxidized and reduced at the positive and the negative electrode, respectively. In nickel – cadmium batteries, for example, this can be caused by iron ions which are reduced at the negative electrode according to

$$Fe^{3+} + e^- \longrightarrow Fe^{2+}$$

The Fe^{2+} ions migrate to the positive electrode, where they are oxidized by the reverse reaction. The process is repeated, when the so formed Fe^{3+} ions again reach the negative electrode. In this way both electrodes are gradually discharged (cf. the "nitrate shuttle"in Section 7.1.2.1.7).

The formation of electron conducting bridges is another mechanism that may establish a self-discharge. An example is the formation of ZnO dendrites penetrating the separator system and getting in touch with the cathode, consisting, e.g., of Ag_2O. The cathode material itself, if soluble to a certain extent like Ag_2O, may also lead to the formation of electron-conducting Ag bridges if reduced by cell components, like, e.g., an unsuitable separator [35].

4.2.6.3. Apparent Self-Discharge

After prolonged storage of primary cells, their increased internal resistance is often misinterpreted as self-discharge. The reason for the increase in internal resistance has nothing to do with chemical or electrochemical self-discharge, but only with mass transfer. If, e.g., the cathode has void pores it will soak-up electrolyt during storage from the region between anode and cathode, thus reducing the battery's load capability. Lithium batteries with a plastic seal will pick up humidity during storage and loose volatile electrolyte components by permeation. The air – zinc battery will loose water by evaporation after activation, if stored in an environment with lower humidity, which may also lead to a considerable increase in internal resistance.

The increased internal resistance reduces the voltage under load, and the end-of-discharge voltage is reached earlier compared to a fresh battery. Thus the delivered capacity is reduced although the electrodes are still fully charged, i.e., the battery's full capacity may still be obtained at a reduced load.

4.2.6.4. Capacity Loss During Storage

For all the reasons given above, a battery loses capacity during storage. The loss of capacity depends on system, construction, and storage conditions, such as temperature. Therefore a wide region of self-discharge ranges is observed. In primary batteries with lithium electrodes, a protecting layer is formed at the electrode surface and extremely low self-discharge results. As a rule of thumb, for a given battery system the self-discharge rate correlates roughly with its

specific power output, i.e., the higher its specific power output, the higher the self-discharge. The reason is that both, self-discharge and power output, correlate with the electrode surface, i.e., an increased electrode surface means both, an increased power output and an increased self-discharge, especially at reduced electrode distances.

Practical values for *primary batteries* at ambient temperature, that also include the apparent self-discharge (see Section 4.2.6.3), are in the range of 0.5 % (and less) to 20 % per year. The low value applies, e.g., to lithium batteries with a high quality seal. The 20 % value is considered to be a maximum value for low cost carbon – zinc batteries, which is, however, mostly determined by the apparent self-discharge (see Section 4.2.6.3). The true chemical and/or electrochemical self-discharge is much smaller, unless low quality active materials are employed.

Also *secondary battery systems* exhibit a broad range of different rates of self-discharge. Their values, however are based on a one-month period in contrast to primary systems (one-year period). Depending on system and construction, typical values vary between 2 and 30 % at ambient temperature. For the lead – acid system the values vary between 2 and 20 % depending on antimony content and age. The lithium-ion system offers about 5 % per month. Values in the range of 20 to 30 % per month are observed for the nickel – cadmium and the nickel – metal hydride system.

Instead of the term capacity loss ΔC_s the term capacity retention ΔC_R is also in use. The capacity retention is defined by

$$\Delta C_R / C = 1 - \Delta C_S / C$$

C being the initial capacity of the fresh (and charged) battery.

5. Classification of Battery Systems

Both primary and secondary systems may be further classified by characteristics that allow for a systematic subclassification of the various battery systems. One guiding principle is the classification by electrolyte. Another possibility of battery classification is to base it on battery construction.

Choice of Battery Classification. For the classification of battery systems, preference was given to the classification by electrolyte according to chapter 5.1 with 5.2 as sub-system whenever appropriate.

5.1. Classification by Electrolyte

There are two groups of electrolyte, i.e., aqueous and nonaqueous electrolytes, any battery system may be allocated to:

1) Group 1: Battery systems employing an *aqueous electrolyte* may be divided up into following subgroups
   1) Strong acidic electrolyte (*lead – acid batteries*)
   2) Weak acidic electrolyte (e.g., *Leclanché batteries*)
   3) Neutral electrolyte (e.g., *water activated systems*)
   4) Alkaline electrolyte (e.g., *nickel–cadmium*, *alkaline manganese dioxide – zinc*). The prevailing electrolyte is potassium hydroxide (KOH).
2) Group 2: Battery systems employing a *nonaqueous electrolyte* may be divided up into following subgroups
   1) Organic electrolyte (e.g., *manganese dioxide – lithium, lithium-ion*)
   2) Electrolytes with soluble cathode (e.g., *thionyl chloride*)
   3) Polymer electrolyte (e.g., *vanadium pentoxide – lithium*)
   4) Solid electrolyte (e.g., *iodine – lithium, sulfur – sodium*)

5.2. Classification by Construction

Several constructional principles exist that are common for a great number of battery systems. This concerns the outward appearance as well as internal design features.

5.2.1. External Construction

The external constructional principles are represented by three groups:

Prismatic cells (e.g., Figs. 69, 77, and 86), are characterized by the dimensions length, width, and height. This design is mainly used for larger secondary cells with electrodes in form of plates. But also small prismatic cells exist and even miniaturized ones like those used in modern credit cards. Very flat cell constructions (e.g., below 1 mm in thickness) may be even flexible. One advantage of the prismatic design is the possibility to efficiently use the rectangular volume (cf., e.g., Figs. 77 and 87).

Cylindrical cells (e.g., Figs. 70, 88) are characterized by dimensions diameter d and height h for $h > d$. They are the domain of primary batteries, sealed nickel – cadmium, nickel – metal hydride, and lithium-ion batteries. To some extent they are also used for valve-regulated lead – acid batteries.

Button cells (e.g., Fig. 85) are characterized by dimensions diameter d and height h for $h \leq d$. They are in widespread use as primary batteries, sealed nickel – cadmium, nickel – metal hydride, and lithium batteries.

Battery Packs and Smart Batteries. For many applications, like tools, camcorders, mobile and cordless telephones, rechargeable batteries are assembled to *battery packs*, also called *power packs* that combine a number of cells within a common housing. The cells in such a pack are often selected to ensure their uniform capacity (see Section 7.1.2.1.10).

Smart batteries or "intelligent" batteries have an incorporated controlling system, based on a processor that provides information to the host device in regard to capacity and aging of the battery.

Battery packs and smart batteries are often characterized by different multi-contact and special mechanical fit-in arrangements, i.e., they often fit only one specific device they are made for.

5.2.2. Internal Construction

The internal constructional features are manifold, at least when they are considered in detail. Some of these features will be described in Chapters 6 and 7. The internal construction is mainly determined by the design of the battery's electrodes. Basically there are three different electrode designs:

1) *Pressed electrodes with direct contact to the container wall*. Examples are the disc electrodes in button cells (Fig. 85) and the manganese dioxide cathode rings of the alkaline manganese dioxide-zinc cell. (Fig. 21).
2) *Electrode plates* that usually are connected in parallel to enlarge the surface area and form plate groups (Figs. 68, 86). The thickness of the plates reaches from more than 1 cm to fractions of a millimeter.
3) *Spirally wound electrodes* are often used in cylindrical cells (Figs. 70, 88). Usually one positive and one negative electrode with the separator in between are wound together and form the "electrode group".

Safe-Guarding Devices. For the safe operation of batteries safe-guarding devices are integrated. Valves, vent plugs, and rupture vents as already discussed in Section 4.1.2.2, are *mechanical devices*. They are described in detail in Sections 7.1.1.12 and 7.1.2.1.11.

Furthermore *nonmechanical* safe guarding devices are employed for battery packs, and to some extent also for single cells. These may be thermal fuses, or devices that increase the internal resistance when a specified temperature is exceeded, e.g., PTC devices as in Figure 32. A safety switch, controlled by the current, is shown in Figure 10. The cell temperature is often used for charge control (see Section 7.1.2.1.12). Corresponding temperature sensors are installed in the charging device or already in the battery pack. The processor of smart batteries also controls proper charging, and prevents overdischarge.

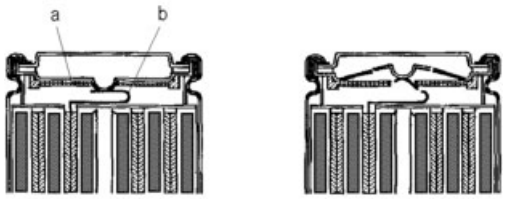

Figure 10. Schematic diagram of the safety device of the Sony 18650 cell
a) Al burst disk; b) Al connection between cell cover and positive electrode

6. Primary Batteries, Systems, and Application

The system classification of primary batteries is carried out according to Section 5.1 (classification by electrolyte), with the exception of Section 6.3, dealing with reserve batteries, which form a group of their own.

The first group of systems to be presented employ aqueous electrolytes. This group is governed by the carbon – zinc systems. Systems with aqueous electrolyte may be roughly subdivided by the pH value of their electrolyte. The first subgroup employs weak acid electrolytes having an pH value in the range of $2 < pH < 4$. This group is also dominated by the "carbon – zinc" system. The second subgroup employs a nearly neutral electrolyte like seawater and is governed by magnesium systems. The third subgroup, finally, uses strong alkaline electrolytes, predominantly KOH ($pH > 14$) and is again dominated by zinc systems.

The second group of systems employ nonaqueous electrolytes. They are governed by lithium systems using either an organic, inorganic, or solid electrolyte and are subdivided accordingly.

In this article only systems and batteries will be addressed, that have received commercial significance. Systems and batteries for very specific applications will not be covered, apart from a few exceptions. Reference is made instead to the existing literature, e.g., [37].

6.1. Systems Employing an Aqueous Electrolyte

6.1.1. Batteries with Weak Acid Electrolyte

6.1.1.1. Carbon – Zinc Systems

6.1.1.1.1. General

The system's popular name is carbon – zinc, its correct name would be the "manganese dioxide – zinc system with weak acid electrolyte".

Normally a battery system is named after its active materials. In the case of carbon – zinc, however, reference is made actually to its current collectors, which is carbon for its positive and zinc for its negative terminal. Zinc is simultaneously the active material of the negative electrode. For cylindrical batteries, zinc is also part of the battery construction, i.e., it is the can of the battery. The multifunction of the zinc is one of the reasons why the system can be produced at rather low costs and which in turn is the reason for its widespread use. The system has been standardized by the IEC [26], [35] but did not receive a system letter of its own in contrast to all other standardized systems.

There are two variants of carbon – zinc systems. Both variants exist in different grades, depending on the quality of the active material and the type of construction employed. The first variant uses predominantly ammonium chloride as electrolyte and is therefore named the "ammonium chloride" system. Since it was pioneered by George Lechlanché it is also known as "Leclanché" system. The second variant uses predominantly zinc chloride as electrolyte and is therefore called the ″zinc chloride″ system. The literature on this subject is abundant [38–40].

The alkaline version of the manganese dioxide – zinc battery surpasses the carbon – zinc systems in performance by far, however, at higher costs. This system will be dealt with in Section 6.1.3.1.

6.1.1.1.2. Electrochemistry

Both variants of the carbon – zinc system employ manganese dioxide as cathode material (positive electrode) and zinc as anode material (negative electrode). The reaction mechanisms of these systems are rather complex. Under light discharge conditions the following reactions occur:

"Ammonium Chloride" or Leclanché System. The electrochemical reaction of the negative electrode is given by Equation (6.1)

$$Zn + 2\,Cl^- \longrightarrow ZnCl_2 + 2\,e^- \tag{6.1}$$

The electrochemical reaction of the positive electrode is described by Equation (6.2)

$$2\,MnO_2 + 2\,NH_4Cl + 2\,e^- \longrightarrow 2\,MnOOH + 2\,NH_3 + 2\,Cl^- \tag{6.2}$$

Reactions (6.1) and (6.2) are followed by a precipitation of zinc diammine chloride (Eq. 6.3)

$$ZnCl_2 + 2\,NH_3 \longrightarrow [Zn(NH_3)_2]Cl_2 \tag{6.3}$$

The overall reaction is

$$2\,MnO_2 + Zn + 2\,NH_4Cl \longrightarrow 2\,MnOOH + [Zn(NH_3)_2]Cl_2 \tag{6.4}$$

and will be discussed below.

Zinc Chloride Battery. The electrochemical reaction of the negative electrode is given by Equation (6.5)

$$Zn + 2\,Cl^- \longrightarrow ZnCl_2 + 2\,e^- \tag{6.5}$$

The reaction of the positive electrode is described by Equation (6.6):

$$8\,MnO_2 + 4\,ZnCl_2 + 4\,H_2O + 8\,e^- \longrightarrow 8\,MnOOH + 4\,ZnO + 8\,Cl^- \tag{6.6}$$

Reactions (6.5) and (6.6) are followed by a precipitation of zinc oxychloride. At high current densities it is preferably generated in the anode compartment.

$$4\,Zn^{2+} + ZnCl_2 + 9\,H_2O \longrightarrow ZnCl_2 \cdot 4\,ZnO \cdot 5\,H_2O + 8\,H^+ \tag{6.7}$$

The overall reaction is

$$8\,MnO_2 + 4\,Zn + ZnCl_2 + 9\,H_2O \longrightarrow 8\,MnOOH + ZnCl_2 \cdot 4\,ZnO \cdot 5\,H_2O \tag{6.8}$$

System Comparison. One of the outstanding features of the zinc chloride system in comparison to the ammonium chloride system is its superior resistance to leakage, which follows from Equation (6.4) in comparison to Equation (6.8): In contrast to the zinc chloride system the ammonium chloride system does neither use up water in the course of a reaction, nor does it generate water, whereas the zinc chloride system reacts with water molecules by producing zinc oxychloride.

Under light discharge conditions with endpoint voltages of $U \geq 0.75\,V$ the ammonium chloride system is also unlikely to leak.

The difference in leakage behavior becomes, however, evident under the condition of prolonged discharge, i.e., discharge voltages of 0.7 V and below. Under this condition the following reaction occurs in the ammonium chloride system:

$$Zn + 6\,MnOOH \longrightarrow 2\,Mn_3O_4 + ZnO + 3\,H_2O \tag{6.9}$$

This reaction involves generation of water and is inevitably followed by electrolyte leakage. The leakage is also favored by an internal pressure built-up and the perforation of the zinc can that most likely occurs under these conditions. This is in contrast to the zinc chloride system, which continues to use up water by producing complex compounds similar to zinc oxychloride (see Eq. 6.8) and which also leads to a reduction in electrolyte volume, thus effectively reducing the problem of electrolyte leakage.

Another outstanding feature of the zinc chloride system is the property of its precipitate (Eq. 6.7) in comparison to the precipitate formed by the ammonium chloride system (Eq. 6.3). The precipitate of zinc diammine chloride is much more dense than the microporous zinc oxychloride. This difference becomes more pronounced at higher current densities. This means that the zinc diammine chloride precipitate not only represents a severe diffusion barrier leading to marked diffusion polarization for all diffusing species but also to an increase of the system's ohmic resistance. For the zinc chloride system these negative effects are much less pronounced, i.e., it offers a higher load capability.

Another advantage of the zinc chloride system is the larger volume of electrolyte being employed (approx. 15 % more). Due to a special mixing process for the positive mass (MnO_2, acetylene black, and $ZnCl_2$ electrolyte) pores are established, being filled with electrolyte. These pores permit approx. 36 % higher transport rates of species according to Equation (6.6), than those of the ammonium chloride system according to Equation (6.2). At higher discharge currents this also means lesser concentration polarization within the cathode of the zinc chloride system in comparison to the ammonium chloride system. These advantages of the zinc chloride system over the ammonium chloride system disappear for low and/or intermittent current densities.

A disadvantage of the zinc chloride system in comparison to the ammonium chloride system is its storage behavior. When employing the same active materials, swelling agents and seal, etc., the zinc chloride system tends to dry out faster than the ammonium chloride system. This, however, can be overcome by a proper choice of swelling agents, nonporous carbon as current collector, an increased amount of carbon black, high quality active materials, and a hermetic seal. These measures, however, mean higher costs in comparison to the ammonium chloride system.

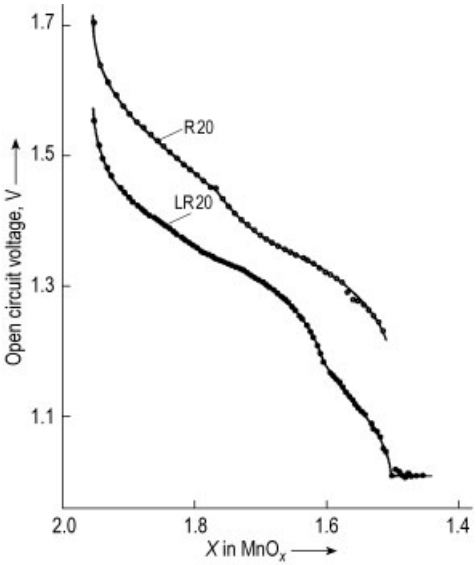

Figure 11. Open circuit voltage of R20 sized batteries containing electrodeposited MnO_2 in the IEC radio test [45] o R20 ("carbon – zinc") • LR20 (alkaline electrolyte)

6.1.1.1.3. System-Immanent Properties

System-immanent properties are directly correlated to the electrochemical reaction occurring in the system and to transport phenomena. According to Chapter 3, the open circuit voltage U^o may be calculated from the thermodynamic data of the cell reaction. This is also true for the carbon – zinc system. The MnO_2 used as cathode material, however, has the peculiarity that its open circuit potential U^o depends not only on the MnO_2 modification employed [The more important modifications are: natural manganese dioxide (NMD), lowest quality; activated manganese dioxide (AMD); chemically synthesized manganese dioxide (CMD), and electrodeposited manganese dioxide (EMD), highest quality)] but also on the degree of electrochemical reduction [41–43].

In a buffered ammonium chloride solution the open circuit voltage U^o of α-MnO_x covers a range of $1.6\,V > U^o \geq 1.05\,V$, for $1.98 > x \geq 1.87$ [44], indicating a homogeneous phase being followed by a heterogeneous phase for $1.87 > x > 1.39$ while exhibiting a constant open circuit value U^o of approx. 1.05 V.

A relevant dependence of the U^o-value on discharge is also observed for EMD in an alkaline electrolyte [45]. Experimental values obtained when applying an IEC radio test to R20 and LR20 batteries (IEC application test: 10 Ω; 4 h/d) are shown in Figure 11, indicating a homogeneous reaction between MnO_2 and $MnO_{1.5}$.

Another system-immanent property has to do with the zinc and the fact that its potential is − 0.42 V versus the hydrogen potential in the same electrolyte, i.e., Reaction (6.10) proceeds, if not inhibited.

$$Zn + H_2O \longrightarrow ZnO + H_2 \tag{6.10}$$

Indeed, this reaction, characterized by a dissolution of zinc and the generation of hydrogen, or variants of it have caused considerable problems in the past. These problems finally were solved by the addition of mercurous chloride, which forms an amalgam with zinc, leading thus to an inhibition of Reaction (6.10). This treatment led to a considerable improvement of the system's storage behavior (reduction of internal self-discharge, elimination of problems connected with internal H_2 pressure and can corrosion). However, because of the environmental problems, mercury had to be eliminated or replaced. At present nearly all battery manufacturers succeeded to replace zinc amalgam by other environmentally benign zinc alloys.

The system's specific energy (mWh/g) and the system's energy density (mWh/cm^3) are further system-immanent properties. When assuming an average discharge voltage of 1.3 V the following theoretical values may be derived under simplified conditions (one electron step reaction according to Reaction 6.8):

Specific energy: 222 mWh/g
Energy density: 727 mWh/cm^3

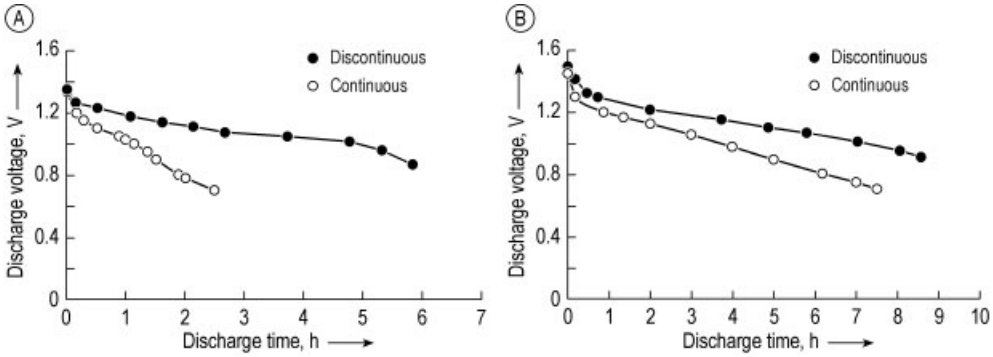

Figure 12. Continuous and discontinuous discharge through 3.9 Ω
A) Ammonium chloride R14S battery, endpoint voltage 0.9 V; B) Zinc chloride R14P battery, endpoint voltage 0.9 V

Values obtained under experimental conditions, however, are significantly below. When discharging a modern premium grade R20 zinc chloride battery at low current (e.g., at 0.15 mA/cm^3, endpoint voltage 0.9 V) the following experimental discharge values may be verified:

Specific energy: 120 mWh/g
Energy density: 200 mWh/cm^3

Low-cost carbon – zinc batteries, using NMD, exhibit an inferior performance. Their values may even be below 70 mWh/g and 120 mWh/cm^3.

The difference between theoretical and experimental values not only depends on the quality of the manganese dioxide and the porosity of the cathode used in the battery, but also on weight and volume of the excess zinc necessary to produce the can, and inactive materials such as separator, carbon rod, graphite, carbon black, labeled metal jacket plus insulation and contact arrangements.

For increased discharge currents (2 mA/cm^3 and above) further system-immanent properties become noticeable, which are governed by electrode kinetics and transport phenomena. Under increased discharge currents a diffusion barrier builds up (see Section 6.1.1.1.2). The formation of a diffusion barrier is a time dependent process that leads to an increasing diffusion polarization, which is accompanied by an increasing voltage drop under load (CCV). On current interruption the voltage recovers because the diffusion barrier regresses. Figure 12 illustrates the discharge of two ammonium chloride batteries type R14S. One discharge was carried out discontinuously according to an IEC portable lighting test, i.e., 3.9 Ω, 4 min/h, 8 h/d. The other R14S battery was discharged continuously through a 3.9 Ω resistance. The continuous test yields only 27 % of the capacity which the discontinuous test provides.

This is the fundamental reason why carbon – zinc batteries perform better under intermittent load conditions. This is especially true for the ammonium chloride system. The zinc chloride system offers a higher load capability and is less time dependent in its performance (see Fig. 12 B). Figure 12 B illustrates the discharge results of two zinc chloride batteries type R14P, which were discharged under identical conditions. The result for the zinc chloride system is that under continuous discharge conditions 58% of the capacity is available, instead of only 27 % for the ammonium chloride system (Fig. 12 A).

For low current demands the advantage of the zinc chloride system over the ammonium chloride system diminishes, i.e., both provide the same capacity. In those cases the cheaper ammonium chloride system may be the better choice.

6.1.1.1.4. Batteries and Their Applications

The final property of an electrochemical system is given by both, its system-immanent properties (including the choice of materials) and its battery design (the design includes battery construction and size).

According to Section 5.2 three constructions have been established by time: Cylindrical, button, and prismatic cells.

Carbon – zinc batteries are made only in cylindrical and prismatic shapes with a variety of terminal arrangements. They are manufactured with different internal constructions. The more important ones are described below.

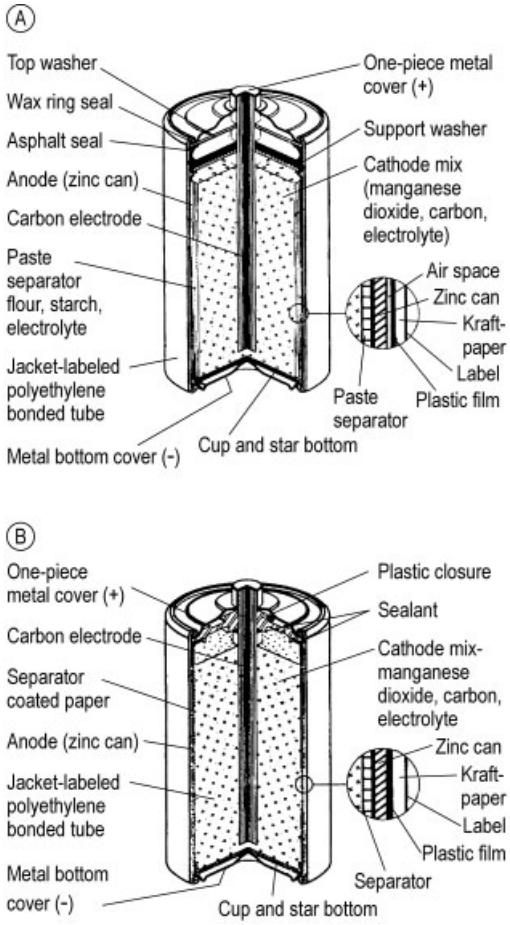

Figure 13. Cross-section of an ammonium chloride (A) and a zinc chloride cylindrical battery (B)

Cylindrical batteries are characterized by diameter d and height h, with $d < h$. The cylindrical carbon – zinc battery is made in two basically different internal constructions. The construction according to Figure 13 A is predominantly used for low cost ammonium chloride batteries, whereas the construction according to Figure 13 B is predominantly used for the more expensive zinc chloride battery, which needs a better seal (see Section 6.1.1.1.2).

For both carbon – zinc systems the following IEC battery sizes are available on the market (in sequence of increasing volume, cf. Chap. 9):

R03 (AAA), R1 (N), R6 (AA), R14 (C) and R20 (D) [The designations AAA, N, AA, C and D are no longer valid ANSI (American National Standard Institute) designations for cylindrical ammonium chloride batteries with natural manganese dioxide. They are unofficially still in use, however, only to designate battery sizes].

For batteries R6, R14 and R20 three performance variants are standardized, characterized by suffices S (standard), C (high capacity), and P (high power) [58], [33].

In general suffix S, e.g., R6S, R14S, R20S designates ammonium chloride batteries whereas suffix P designates in general zinc chloride batteries, e.g., R6P, R14P and R20P. Thus suffix "S" batteries are of a lower performance than suffix "P" batteries, whereas suffix "C" batteries offer a medium performance but at a higher capacity. Since the use of IEC designations is voluntary, not all manufacturers follow the IEC proposal to designate their performance variants by suffices S, C, and P.

The most important battery size by production volume is presently the R6 (AA) battery. The discharge behavior of both carbon – zinc systems is illustrated in Figure 14 A and B. Figure 14 A shows the voltage-time transient of an IEC transistor radio test, which is a discontinuous test with a 43 Ω load, 4 h/d. Figure 14 B is the voltage – time transient of an IEC pulse test (simulating a photo flash), which is a discontinuous test with a 1.8 Ω load, 15 s on and 45 s off, 24 h/d and which represents a much higher discharge rate.

Figure 14 A demonstrates a discontinuous low rate discharge. Under this condition the ammonium chloride system offers practically the same performance as the zinc chloride system. Under pulse test conditions, however, the R6P battery (zinc chloride) is superior to the R6S battery (ammonium chloride) by about 42 %, see Figure 14 B. If compared, however, to a LR6 battery (see Section 6.1.3.1), the R6S and the R6P battery are by far inferior. This is illustrated by the discharge result of a LR6 battery, as incorporated into Figure 14 B. With an endpoint voltage

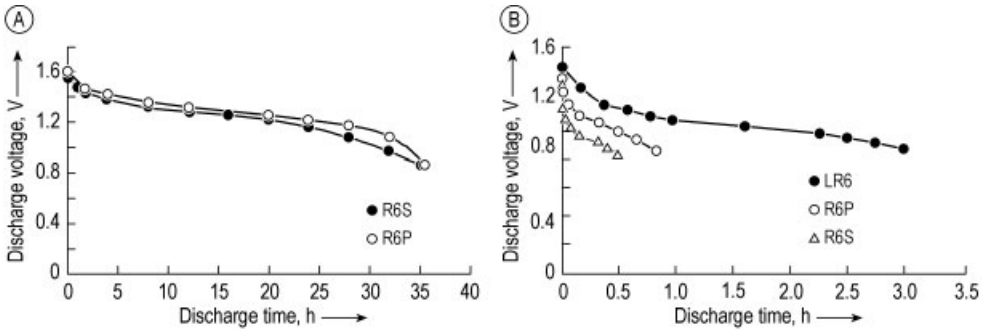

Figure 14. R6P and R6S batteries
A) Transistor radio test; B) 1.8 Ω pulse test

of 0.9 V it provides nearly four times more capacity than the R6P battery.

For an interpretation of the results of Figure 14 see the explanations given in Sections 6.1.1.1.2 and 6.1.1.1.3. The statements are valid independent of battery size. The applications for the carbon – zinc cylindrical batteries are manifold and change with changing technology. Typical applications are listed in Table 1.

Table 1. Applications of 1.5 V cylindrical carbon – zinc batteries according to IEC

Applications	R03	R1	R6	R14	R20
Portable lighting	x	x		x	x
Personal cassette players	x		x		
Tape recorder				x	x
Transistor radios	x		x	x	x
Pulse test (photo flash)	x		x		
Hearing aids		x			
Paging		x			
Motor/toy			x	x	x

Prismatic batteries are characterized by length, width, and height. Carbon – zinc prismatic batteries are mostly multicell batteries, i.e., they have normally at least two cells connected in series and therefore open circuit voltages of 3 V and above, e.g., 30 V. Batteries of this type, however, are more and more disappearing from the market with a few exceptions, like the 9 V battery according to IEC designation 6F22. It has six flat cells (IEC size F22) connected in series. The external and internal construction of the battery is shown in Figure 15.

The main applications of this battery are:

- Remote control,
- Smoke detectors,
- Electrical multi-meters (small transportable, multifunction instruments, e.g., for measurement of current, voltage, or resistance),
- Toys, and
- Transistor radios.

The 6F22 9 V carbon – zinc battery exists also as alkaline version, having the IEC designation 6LR61, (see Section 6.1.3.1). It surpasses all carbon – zinc versions in performance, but is also more expensive.

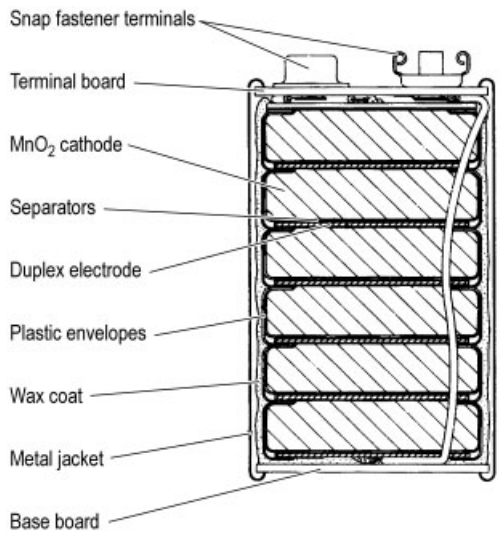

Figure 15. Construction of a 9 V IEC type 6F22 carbon – zinc battery

Another well known prismatic battery construction is the 4.5 V battery according to IEC designation 3R12. It contains three cylindrical cells in series (IEC size R12). Its external construction is illustrated in Figure 16. The main ap-

plications of this battery are in portable lighting and transistor radios.

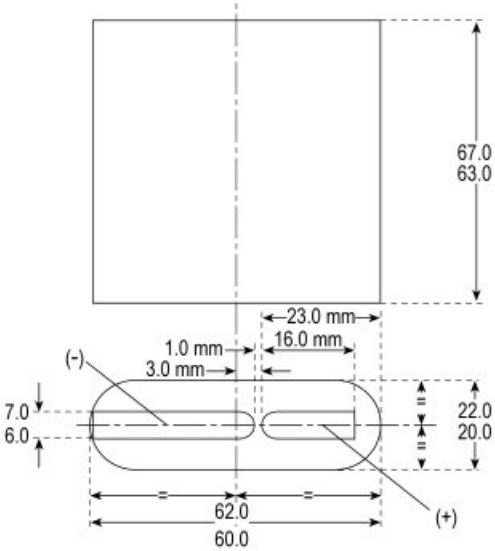

Figure 16. External construction of the 4.5 V IEC type 3R12 carbon – zinc battery

According to IEC this battery exists in three performance variants designated by suffices S (standard), C (high capacity) and P (high power). The 3R12 4.5 V carbon – zinc battery exists also as alkaline version, having the IEC designation 3LR12, (see Section 6.1.3.1). It surpasses all carbon – zinc versions in performance, but again it is offered at a higher price.

The specific energy and the energy density of multicell batteries are below corresponding values of single cell batteries like, e.g., the R20 cylindrical battery. The reason for this is the higher percentage of inactive materials and volume being unavoidable for multicell batteries. For a 6F22 battery, e.g., the values are approx. 30 % below those quoted in Section 6.1.1.1.3.

Battery Packs. A unique battery is the four cell flatpack 6 V battery, manufactured by the Polaroid Corporation [46] and being integrated into a 10 pictures film-pack. This concept ensures that each film-pack has always a fresh battery (film transport, flashlight, shutter operation, etc.). The pack basically consists of a carbon – zinc multicell battery with zinc chloride as electrolyte, having a geometric electrode area of approx. 5.1 by 5.1 cm. At a discharge current of 1 A it offers a capacity of approx. 100 mAh, at a 1 mA discharge rate the capacity rises to about 250 mAh. The exploded view in Figure 17 represents a single cell [47].

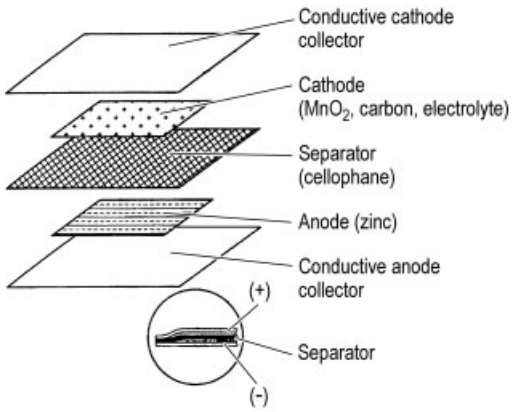

Figure 17. Exploded view of a single cell from a Polaroid P-80 battery pack

More information on carbon – zinc batteries (performance and applications) is given in manufacturer's data sheets, brochures and technical handbooks and in the IEC-Standard Public. 86 [26], [33].

6.1.1.1.5. Prospects

By now the carbon – zinc system has been on the market for about 100 years and will continue to exist. There is no reason to assume that this system will disappear in the foreseeable future. The worldwide annual production volume is estimated to be 18×10^9 carbon – zinc batteries in 2001. It is not to be expected that any other battery system will offer portable electrical energy at a price being competitive to the carbon – zinc system. It will continue to be produced in large volumes probably with an increasing share of zinc chloride batteries. It is, however, to be expected, that the total percentage market share of carbon – zinc will decrease in favor of the alkaline manganese dioxide – zinc system.

6.1.1.2. Air – Zinc System

6.1.1.2.1. General

The air – zinc system with a weak acid electrolyte, like ammonium chloride, is a system of

the past. It belongs in principle to the group of carbon – zinc systems. H. COHEN [48] was probably the first to develop this system, which still is available on the market, although by now technically surpassed by other air – zinc systems. The IEC allocated system letter A to this system and addresses the positive electrode as oxygen electrode, which, however, has to be corrected to air electrode (cf. Section 6.1.3.4.1).

An interesting variant was developed in the 1960s. In this variant the ammonium chloride electrolyte was replaced by manganese(II) chloride in mixtures with magnesium chloride. This electrolyte was used in combination with a flat plate-construction for the air cathode (B. SILLER [49]) to improve the system's storage behavior, energy density and load capability. Although technically advanced it never attained market significance.

The alkaline version suggested by C. FÉRY [50] is much more promising in this respect. FÉRY obviously built the first successfully operated alkaline air – zinc cell (see Section 6.1.3.4).

6.1.1.2.2. Electrochemistry

The electrochemical active materials are

- Zinc, which represents the negative electrode,
- atmospheric oxygen in combination with a suitable porous carbon electrode, which represents the positive electrode, and
- the electrolyte, because it also participates in the reaction like in all carbon – zinc systems

The electrochemical reactions may be described by

Negative electrode

$$2\,Zn \longrightarrow 2\,Zn^{2+} + 4\,e^- \tag{6.11}$$

Positive electrode:

$$O_2 + 4\,e^- + 4NH_4^+ \longrightarrow 2\,H_2O + 4\,NH_3 \tag{6.12}$$

Electrolyte:

$$2\,Zn^{2+} + 4\,NH_3 + 4\,Cl^- \longrightarrow 2\,[Zn(NH_3)_2]Cl_2 \tag{6.13}$$

Overall reaction:

$$2\,Zn + O_2 + 4\,NH_4Cl \longrightarrow 2\,H_2O + 2\,[Zn(NH_3)_2]Cl_2 \tag{6.14}$$

Like all gas electrodes, also this air cathode needs to fulfil one important requirement: it must create a three-phase boundary, which is the reaction zone, otherwise it would not be operable. At the three-phase boundary the following phases are in contact with each other:

1) The oxygen of the air,
2) The electrolyte, and
3) The porous carbon electrode with sufficient catalytic activity for the electrochemical oxygen reduction according to Equation (6.1)

For a three-phase boundary to be established, the porous carbon structure must be partially hydrophilic – to allow for pores being filled with electrolyte – and partially hydrophobic to stay open for gas transport.

Like the ammonium chloride manganese dioxide – zinc system, also this air – zinc system produces zinc diammine chloride (Eq. 6.13), which precipitates in the electrolyte-filled pores of the air cathode. This process finally leads to a total blockage of the porous carbon structure, which means the battery's end of discharge. In other words: the porous volume of the carbon electrode, that is filled with electrolyte, determines the amount of electrical energy, that can be withdrawn from this system. This is the reason why this system needs rather thick porous air cathodes.

On the other hand a thick air cathode gives rise to a relatively high diffusion polarization (problems of mass transport), especially at higher currents. Thus the thickness of the air cathode must constitute a balance between the average power output in terms of, e.g., mW and the electrical energy in terms of Wh that may be obtained from the system. A typical thickness of an air cathode is of the order of several centimeters.

6.1.1.2.3. System-Immanent Properties

Again, the immanent properties of this system are directly correlated to its electrochemistry. This also includes the open circuit voltage. According to calculation the standard cell voltage is 1.73 V [45]. Since, however, the oxygen electrode is not a reversible one, the observed open circuit voltage is below this value, i.e., 1.45 V.

According to IEC, the nominal voltage is chosen to be 1.4 V.

The system's specific energy and energy density depends very much on the discharge conditions, including the environment (see below). At a current density of about 0.2 mA/cm^3 and ambient temperature typical values are of the order of

Specific Energy: 140 mWh/g

Energy density: 150 mWh/cm^3

Another system-immanent property is due to the fact that the system is open to the atmosphere, which is of course a requirement to permit oxygen to enter the system. However, simultaneously an exchange of water vapor proceeds between the system and its environment. Since the equilibrium humidity of ammonium chloride corresponds to about 80 % relative humidity, while the environmental humidity is normally below this value, the system tends to dry out after opening the air access. Due to water loss it may exhibit a rather low initial voltage on load that, however, will recover during operation, because water is produced according to Equation (6.12), thus compensating for the water loss during storage. In order to avoid an initial low voltage under load, the air access should be closed during storage or prolonged idle hours. From the above it follows that under extreme high humidity conditions, i.e., a relative humidity of the atmosphere > 80 %, the system would absorb water, which also would deteriorate the system's performance by time.

Another system-immanent property is the fact that the electrolyte of this air – zinc system does not absorb CO_2 from the atmosphere, in contrast to the alkaline air – zinc system. This is a property that makes the system suitable for long periods of operation, provided its humidity household is under control.

(Reaction 6.10 Section 6.1.1.1.3 also applies to this air – zinc system).

6.1.1.2.4. Batteries and Their Applications

In the air – zinc system the variety in battery sizes is limited in comparison to the batteries of the carbon – zinc system.

Cylindrical Batteries. The AR40 battery is presently the only cylindrical air – zinc battery listed in IEC Standard Publication 60086-2. It has a diameter of 67 mm and an overall height of 172 mm. Its internal construction is given in Figure 18.

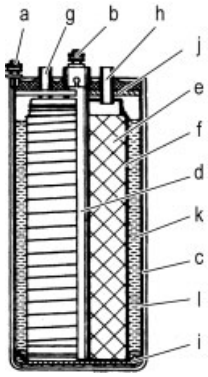

Figure 18. Cross-section of an air – zinc battery IEC type AR40 (Ø = 67 mm; h = 172 mm)

The electrical energy is taken from the battery's screw terminals a and b. Terminal a is the negative pole being connected to the zinc can c, which is also part of the battery construction.

Screw terminal b is connected to carbon rod d, which is in contact with the active material e of the air cathode. The active material consist of a mixture of active carbon, graphite and electrolyte and forms a porous body that permits internal air diffusion. Around the porous body a separator paper f is wrapped. Within this porous body the oxygen of the air, the electrolyte and the active carbon are in contact with each other, forming the reaction zone, i.e., the three-phase boundary. The air enters the system via the air access openings g and h. The bottom of the cathode mix is separated from the can by a paper cup i. The gap between separator f and zinc can c is filled up with gelled electrolyte k. A card board washer j and a bitumen seal on top closes the battery. A card board cup l protects the zinc can c from outside.

The AR40 is offered also as multicell battery, having the IEC designation 5AR40. It comprises five AR40 cells connected in series to give a 7 V/85 Ah battery. The cells are arranged in a circle of 184 mm diameter at a height of 190 mm. When discharged, e.g., through a 500 Ω resis-

tor it delivers at ambient temperature on average 11.5 mA at 5.7 V for 365 d, i.e., 574 Wh.

Prismatic Batteries. The IEC standard lists two 8.4 V prismatic multicell batteries, carrying the IEC designation 6AS4 and 6AS6. They comprise six AS4 or six AS6 prismatic single cell batteries, connected in series. The characteristics of these multicell batteries are given in Table 2 for ambient temperature.

The design of the prismatic single cell batteries AS4 and AS6 is identical to the AR40 single cell battery shown in Figure 18.

The application of the air – zinc batteries with ammonium chloride electrolyte is in the area of low power output (e.g., 50 to 100 mW) for a long time (e.g., 5 to 12 month). Typical applications are industrial applications (e.g., barricade flashlight) and fence control. Another interesting application area is still the operation of emergency phones in areas where no power line is available.

Advantageous for this system is its portability and low price compared to other batteries and systems.

6.1.1.2.5. Prospects

In the mid-term future the air – zinc system should continue to be competitive in the low power/long term application field and in locations where no alternative power supply is available. In the long term it will probably disappear from the market.

6.1.2. Batteries with "Neutral Electrolyte"

Systems working with a true neutral electrolyte do not exist, because this kind of "electrolyte" would not have sufficient ionic conductivity [at 20 °C the ionic conductivity of pure water is ca. $5 \times 10^{-8}\,\Omega^{-1}\,\mathrm{cm}^{-1}$]. The term "neutral electrolyte" denotes natural water compositions like ocean water and fresh water of lakes and rivers. The term "activation" as used in the following refers to the method by which water is brought into the space between the battery's positive and negative electrode(s), which initiates the electrochemical reaction.

Magnesium Systems.

General. Several magnesium systems are known. Attention is drawn to the literature, e.g., to [51].

Magnesium systems are normally listed under "reserve batteries" (cf. Section 6.3). This is due to their specific properties like their excellent storage behavior (in a dry environment) before activation and their immediate readiness for use after activation (see also Section 6.1.3.4).

Electrochemistry. The more prominent magnesium system is the silver chloride – magnesium system. Because of the silver, it is also the most expensive magnesium system. Many attempts have been made in the past to replace the expensive AgCl with less expensive cathode materials such as Cu_2I_2, CuCl, CuSCN, $PbCl_2$, and others.

The technical performances of these systems are inferior to that of the silver chloride system.

The electrochemical reaction of the magnesium electrode (anode) is described by Equation (6.15)

$$\mathrm{Mg} \longrightarrow \mathrm{Mg}^{2+} + 2\,\mathrm{e}^- \tag{6.15}$$

For the silver chloride electrode Equation (6.16) applies

$$2\,\mathrm{AgCl} + 2\,\mathrm{e}^- \longrightarrow 2\,\mathrm{Ag} + 2\,\mathrm{Cl}^- \tag{6.16}$$

The overall reaction follows from Equations (6.15) and (6.16)

$$\mathrm{Mg} + 2\,\mathrm{AgCl} \longrightarrow \mathrm{MgCl_2} + 2\,\mathrm{Ag} \tag{6.17}$$

The electrochemical reaction of the magnesium – cuprous iodide system may serve as a second example. The chemical reaction at its negative electrode is described by Equation (6.15). For its positive electrode Equation (6.18) applies:

$$\mathrm{Cu_2I_2} + 2\,\mathrm{e}^- \longrightarrow 2\,\mathrm{Cu} + 2\,\mathrm{I}^- \tag{6.18}$$

The overall reaction of the magnesium – cuprous iodide system is

$$\mathrm{Mg} + \mathrm{Cu_2I_2} \longrightarrow \mathrm{MgI_2} + 2\,\mathrm{Cu} \tag{6.19}$$

The electrochemical reactions of the other magnesium systems proceed in an analogue manner. A unique feature of all magnesium systems is that they produce water soluble conductive salt,

Table 2. Air – zinc batteries (8.4 V), constant resistance discharge

Type, IEC designation	Length/width/height, mm	Discharge resistance, Ω	Capacity, Ah	Endpoint voltage, V
6AS4	168/113/114	300	53	5.1
6AS6	192/128/162	150	130	5.1

in the above examples magnesium chloride or magnesium iodine).

Because of the negative potential of the $Mg/Mg(OH)_2$-couple (approx. -1.7 V vs. SHE) a parasitic side reaction of magnesium with water is inevitable, i.e., the decomposition of water while generating hydrogen gas, which has the beneficial effect of stirring the "electrolyte" between the electrodes.

$$Mg + 2\,H_2O \longrightarrow Mg(OH)_2 + H_2 \tag{6.20}$$

Heat, as also generated by Equation (6.20), favors the battery performance in addition.

System-Immanent Properties. Values of some of the reported electrochemical system properties are listed in Table 3 [51].

According to Table 3 the energy density (mWh/cm^3) and the specific energy (mWh/g) of the magnesium – silver chloride system is approx. three times, respectively, two times as high as that of other magnesium systems. The magnesium – silver chloride system also has the highest discharge voltage, The magnesium lead chloride system has the lowest. The other magnesium systems are similar to the $Mg - Cu_2I_2$ system.

The operating temperatures of the above systems cover a range of $-60\,°C$ to $+65\,°C$. If stored in a dry atmosphere at ambient temperature the magnesium systems may be kept for an indefinite time without noticeable deterioration.

Batteries and Their Applications. Magnesium batteries are normally not available on the commercial market. They are designed for special applications.

Magnesium batteries exist in three different designs [51], i.e., immersion-type batteries, forced flow-type batteries, and dunk-type batteries.

1) *Immersion-type* batteries are activated by immersion into water after opening their water access wholes. Life vests, e.g., are equipped with this type of battery that uses the magnesium – cuprous iodide system to operate a 1.5 V × 0.2 A light bulb for 8 h in the case of emergency.
2) *Forced flow-type batteries:* The torpedo battery belongs to this type of battery. While the torpedo shoots through the sea water, water is driven through the electrode system of the battery, which is in this case a magnesium – silver chloride system. In combination with a water recirculation pump especially flat discharge curves of 1.1 V/cell at high current densities are obtained until the end of discharge.
3) *Dunk-type batteries:* Batteries of this type are activated by pouring water/electrolyte into the system. An absorbent separator is placed between its electrodes, which transports the water/electrolyte to the electrodes initiating thus their electrochemical reaction, while delivering electrical power. Batteries have been designed from 1.5 V up to 130 V at current outputs of up to 10 A [51]. Batteries of this type are used in radiosondes but also to power emergency transmitters. A battery of this kind is a 10.6 V × 10 Ah silver chloride – magnesium reserve battery for aircrafts.

Prospects. No change in market demand or application is expected: Magnesium batteries will continue to be specialists in the area of reserve batteries especially for rescue systems.

6.1.3. Batteries with Alkaline Electrolyte

Alkaline electrolytes offer more possibilities to realize battery systems than acid or neutral electrolytes. This is due to the fact that more metals and metal compounds are stable in alkaline electrolytes than in acid and weak acid electrolytes. The latter only permit zinc, carbon, and manganese dioxide to be used as electrochemical active substances or as electronic conductor (noble metals, titanium, etc., are stable but too expensive for primary batteries).

Table 3. Properties of different magnesium systems

Property	System		
	Mg – AgCl	Mg – Cu_2I_2	Mg – $PbCl_2$
Energy density, mWh/cm^3	180 – 300	50 – 120	50 – 120
Specific energy, mWh/g	100 – 150	50 – 80	50 – 80
Voltage under load∗, V	1.42 – 1.52	1.33 – 1.49	0.90 – 1.06

∗ at 5 mA/cm^2.

For alkaline electrolytes the situation is different. Quite a few metals and metal compounds are stable and have become electrochemical active materials and components of alkaline primary battery systems. They have enabled the development of several commercial battery systems. Metals and metal compounds of this kind are: Nickel, stainless steel, iron, cobalt, manganese, carbon, silver, copper, zinc, brass, mercury, and cadmium, including their oxides and hydroxides.

Another advantage is the high conductivity of alkaline electrolytes (e.g., 30 % KOH: $6 \times 10^{-1}\,\Omega^{-1}\,\mathrm{cm}^{-1}$) and the rather high rates at which electrochemical reactions occur without any impeding precipitates as in the case of carbon – zinc systems (cf. Section 6.1.1.1). This in turn means that diffusion polarization does not play a major role in alkaline battery systems with the exception of very high discharge rates.

Typical for all alkaline batteries is that after prolonged storage or use their negative terminal tends to present signs of leakage, which is of eminent importance, e.g., in the case of button batteries as used in high value devices like wrist watches. Electrolyte leakage is not always a phenomenon that is solely determined by the quality of seal. It is also influenced by a phenomenon that has to do with under-potential deposition [81] of the electrolyte's cations in alkaline solutions (cf. Section 6.1.3.2.4)

6.1.3.1. Manganese Dioxide – Zinc System

6.1.3.1.1. General

The alkaline manganese dioxide – zinc system has become the most important primary battery system. Its market share is constantly growing. The worldwide market for manganese dioxide-zinc batteries, including carbon – zinc, was reported in 1993 to be above 12×10^9 batteries [52] and will further have grown since then.

6.1.3.1.2. Electrochemistry

The electrochemical active materials are zinc (granulated powder) for the negative and electrodeposited γ-MnO_2 (EMD) for the positive electrode. KOH at a concentration of 40 % ($\pm$ 10 %) serves as electrolyte.

The electrochemical reaction of the negative electrode is given by Equation (6.21), which is the overall reaction of Equations (6.21a) and (6.21b).

$$\mathrm{Zn} + 2\,\mathrm{OH}^- \longrightarrow \mathrm{ZnO} + \mathrm{H_2O} + 2\,\mathrm{e}^- \tag{6.21}$$

$$\mathrm{Zn} + 2\mathrm{OH}^- \longrightarrow \mathrm{Zn(OH)_2} + 2\,\mathrm{e}^- \tag{6.21a}$$

$$\mathrm{ZN(OH)_2} \longrightarrow \mathrm{ZnO} + \mathrm{H_2O} \tag{6.21b}$$

The electrochemical reaction of the positive electrode is described by Equation (6.22) which is generally assumed to be a homogenous reaction for MnO_x in the range $2 > x > 1.5$ [53]:

$$2\,\mathrm{MnO_2} + 2\,\mathrm{e}^- + 2\,\mathrm{H_2O} \longrightarrow 2\,\mathrm{MnOOH} + 2\,\mathrm{OH}^- \tag{6.22}$$

The overall reaction is a combination of Equations (6.21) and (6.22)

$$\mathrm{Zn} + 2\,\mathrm{MnO_2} + \mathrm{H_2O} \longrightarrow 2\,\mathrm{MnOOH} + \mathrm{ZnO} \tag{6.23}$$

more likely to happen is the following reaction

$$\mathrm{Zn} + 2\,\mathrm{MnO_2} + 2\,\mathrm{H_2O} \longrightarrow 2\,\mathrm{MnOOH} + \mathrm{Zn(OH)_2} \tag{6.23a}$$

since Reaction (6.21b) is a rather slow process. Under high rate conditions the availability of water molecules in the cathode compartment may become crucial, as may become the availability of hydroxyl ions in the anode compartment, depending on the relative diffusion resistance

and the rate at which they are required. Reaction (6.23) or (6.23a) proceeds in the voltage range $1.55\ V > U > 1.0\ V$. A voltage plateau just above 1 V [54] is determined by Equation (6.24)

$$2\,MnOOH + H_2O + Zn \longleftrightarrow 2\,Mn(OH)_2 + ZnO \qquad (6.24)$$

The standard potential of Reaction (6.24), calculated from Gibbs free energies is 1.06 V [55]. When the reduction is continued at a discharge voltage $U < 1.06$ V Reaction (6.25) applies:

$$MnOOH + H_2O + e^- \longrightarrow Mn(OH)_2 + OH^- \qquad (6.25)$$

This is a heterogeneous reaction, being accompanied by an increase in cathode volume [56].

6.1.3.1.3. System-Immanent Properties

System-immanent properties of electrochemical systems have always two aspects. One is of thermodynamic the other one of kinetic origin. During discharge both influences superimpose each other.

The thermodynamic properties of the alkaline manganese dioxide – zinc system are similar in pattern to what was discussed in Section 6.1.1.1.3) in respect to the correlation between the state of discharge of manganese dioxide, the type of reduction reaction (homogeneous or heterogeneous) and its open circuit potential (cf. Fig. 11).

Calculated values of the open circuit voltage on basis of thermodynamic data for the manganese dioxide – zinc system deviate from experimentally observed values [45]. The observed open circuit voltage is 1.55 V and thus within the reported calculated value of carbon – zinc batteries. The open circuit voltage depends, however, also on the MnO_2 content. According to IEC Publication 86-1 the maximum open circuit voltage of the alkaline manganese dioxide – zinc system may attain up to 1.65 V [26].

Much work has been devoted to the kinetics of the manganese dioxide electrode [57], [58]. Special attention is drawn to the analysis of H. Laig-Hörstebrock [59], stating that during galvanostatic discharge the potential of the manganese dioxide electrode is basically controlled by two time-dependent processes, which are experimentally determinable. The experiments were carried out on a manganese dioxide disk-shaped sample, made of a mass mix being identical to what is used in primary batteries, including porosity and electrolyte. The thickness of the MnO_2 disk was chosen to be $d = 2$ mm. Potential measurements were carried out simultaneously on both sides of the disk. The galvanostatic load, however, was applied only to one of its two sides.

The difference in potential $\Delta\varphi$ of both sides was determined. During the initial phase the potential difference changed according to Equation (6.26)

$$\Delta\varphi = S_I t^{1/4} \qquad (6.26)$$

i.e., it is proportional to the speed at which the current penetrates the porous system (S_I = const). In parallel to the current penetration the reduction reaction by proton diffusion into the MnO_2 lattice comes into play, a model being well established [60–62]. The overall reduction reaction is described by

$$MnO_2 + x\,H_2O + x\,e^- \longrightarrow MnOOH_x + x\,OH^- \qquad (6.27)$$

with x being the fraction of diffusing protons. The process comes to an end as soon as the current has penetrated the full electrode thickness (for $d = 2$ mm after approx. 1 min). At this point the electrode potential is only controlled by proton diffusion, which is described by Equation (6.28)

$$\varphi = S_{II} t^{1/2} \qquad (6.28)$$

Constants S_I and S_{II} of Equations (6.26) and (6.27) are derived on basis of theoretical considerations [59] being related to the MnO_2 disk and the proton diffusion [63]. The results are that the potential of the MnO_2 electrode is only controlled by the resistance of the electrolyte diaphragm of the MnO_2 disk and the proton diffusion and not by resistances such as the electronic resistance of the MnO_2 particles and the charge transfer.

It is reasonable to assume that the same factors are rate determining in the reduction of MnO_2 cathodes in carbon – zinc batteries. Their experimental determination, however, is difficult because of the marked influence of the diffusion polarization in carbon – zinc cells (cf. Section 6.1.1.1).

Another system-immanent property has to do with the zinc and the fact that its potential is ca. – 0.42 V versus the hydrogen electrode in the same electrolyte, thus that Reaction (6.29) occurs, if not inhibited.

$$Zn + H_2O \longrightarrow ZnO + H_2 \quad (6.29)$$

This problem is identical in nature to that discussed in Section 6.1.1.1.3. It is, however, more pronounced in the case of the alkaline system, because this system uses zinc powder, having a relatively large surface area and thus producing a relatively high amount of hydrogen gas if not properly inhibited. For safety reasons all alkaline manganese dioxide – zinc batteries have a rupture vent included, that opens when the internal gas pressure exceeds a predetermined value, thus avoiding a rupture of the battery.

In the past the amalgamation of zinc powder, i.e., treatment with mercury to form an alloy, has proven best in avoiding any critical rates of hydrogen evolution. However, the use of mercury was banned for environmental reasons. Costly research programs resulted finally in the development of highly purified zinc powder, which in combination with additives like indium and organic inhibitors, permitted to produce an anode material providing gassing rates being equivalent to that of amalgamated zinc powder.

When assuming an average discharge voltage of 1.3 V the following values for the specific energy and the energy density are obtained:

Specific energy: 272 mWh/g
Energy density: 1132 mWh/cm^3

These values are purely theoretical and exceed significantly the experimental values. When discharging a modern premium grade LR20 battery at a low current density [i.e., 0.3 mA/cm^3, endpoint voltage 0.9 V (endpoint voltage 0.9 V was chosen because practically all IEC discharge regimes for the manganese dioxide zinc systems are based on this voltage)], the following experimental discharge values may be verified:

Specific energy: 170 mWh/g
Energy density: 400 mWh/cm^3

In comparison to the carbon – zinc system (cf. Section 6.1.1.1.3) these experimental values are higher by 40 % in respect to the specific energy and by 100 % as regards the energy density, although these values are obtained at twice the discharge rate (0.3 mA/cm^3 compared to 0.15 mA/cm^3). These values demonstrate already in part the superior performance of the alkaline manganese dioxide system over the carbon – zinc system.

The experimental values are typically well below the theoretical values (cf. Section 6.1.1.1.3), i.e., the theoretical values consider only the active materials according to Equation (6.23), while assuming 100 % Faraday efficiency. The theoretical values do neither consider weight and volume of the inactive materials like the steel can, the separator, the current collector; nor additives like graphite and carbon black; nor the grommet or contact arrangements.

The difference between theoretical and practical values of the energy density is especially high. This also results from the porosity of the cathode and the low density of the zinc powder (approx. 3 g/cm^3). Since the endpoint voltage was chosen to be 0.9 V, the experimental values will also include a contribution from Equation (6.25), i.e., they are not based just on Equation (6.23) or (6.23a), respectively.

For increasing discharge currents the voltage under load continues to drop, depending on the current and the degree of discharge. However, in comparison to the carbon – zinc system the diffusion polarization is much smaller (see Fig. 19). In comparison to Figure 12 the performance of the LR14 battery is significantly better. Under continuous load conditions the battery provides still 89 % of the discontinuous discharge result, whereas the zinc – carbon systems provide only 27 %, respectively 58 %. The alkaline manganese dioxide – zinc system is therefore suitable for both: continuous and discontinuous discharge.

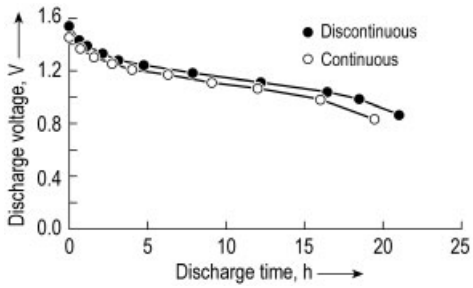

Figure 19. IEC type LR14 continuous (24 h/d) and discontinuous (4 h/d) discharge through a 3.9 Ω resistor

6.1.3.1.4. Batteries and Their Applications

Alkaline manganese dioxide – zinc batteries exist in three shapes: prismatic, cylindrical, and button. The more important batteries are listed in the IEC Standard [33]. To differentiate carbon – zinc batteries from alkaline manganese – dioxide zinc batteries by nomenclature, the alkaline system was given letter L by the IEC system (see Table 4).

Prismatic Batteries. All standardized prismatic batteries are multicell batteries constructed from 3 to 6 cylindrical cells, and connected in series. Relevant nominal voltages vary from 4.5 V (three cells) to 9 V (six cells).

The 6LR61 is a rather well known battery. It is the alkaline version of the 6F22 battery (cf. Section 6.1.1.1.4). The 6LR61 battery offers 500 mAh, i.e., twice the capacity of the 6F22. Like the 6F22 it has a nominal voltage of 9 V, the same external dimensions (26.5 mm × 17.5 mm × 48.5 mm) and identical terminal arrangements. Its internal construction, however, is different from the 6F22 battery (cf. Fig. 15). The 6LR61 is constructed of six alkaline manganese – zinc cylindrical cells, size R61 connected internally in series. Figure 20 displays a battery of this kind.

Figure 20. IEC type 6LR61 9 V battery
The top plate of this prismatic battery carries the terminals (spring loaded snap contacts). The battery contains six cylindrical 1.5 V round cells (LR61). Top and bottom plate carry inside connectors that connect the cells in series to give a 9 V battery. The six round cells are encapsulated in a metal jacket.

The applications for this 9 V battery includes those of the 6F22, i.e.,

- Remote control,
- Smoke detectors,
- Electrical multi-meters,
- Toys, and
- Transistor radios

In addition it permits higher rate applications, e.g., in personal tape recorder. Another well known prismatic battery is the 4.5 V 3LR12 battery which is the alkaline version of the 3R12 battery, however, with more than twice the capacity, i.e., 5900 mAh. Construction and typical applications are identical to the 3R12 battery (cf. Section 6.1.1.1.4), i.e., it is used in portable lighting and transistor radios. The 3LR12 battery also permits higher rate applications.

Cylindrical Batteries. A cross-section of a cylindrical battery is given in Figure 21. The construction is identical to the so-called rechargeable alkaline manganese – zinc battery (see Section 6.1.3.1.4).

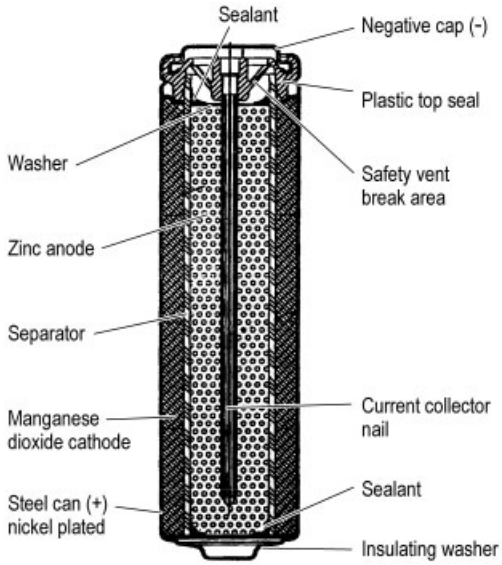

Figure 21. Schematic cross-section of a cylindrical alkaline manganese dioxide – zinc battery

The construction shown in Figure 21 is known as *inside-out* construction, which refers to the battery's polarity. In comparison to the carbon – zinc battery (cf. Fig. 13) the MnO_2 cathode is no longer inside but outside the internal battery construction. It is in direct contact with the nickel-plated steel can, while the anode (zinc powder) is arranged in the center

of the construction. In order to be interchangeable with "carbon – zinc" batteries the terminal arrangements had be changed, such that the negative electrode offers a flat external contact area, while the positive terminal is represented by the steel can which imitates the physical envelope of a carbon rod at its end. Furthermore the cylindrical part of the steel can requires an electrical insulation.

Table 4 lists five standard cylindrical batteries of system L, and zinc – chloride batteries of the same size, including their dimensions and typical applications. For comparison typical capacity values are listed for both systems. These values again demonstrate the superior capacity of the alkaline battery system.

This is especially true for high continuous discharge currents as illustrated in Figure 22. Under the conditions chosen, the R14S provides only 8.6 % and the R14P just 27 % of the capacity provided by the LR14 battery. For low rate and/or discontinuous applications, however, carbon – zinc batteries are still a good choice (cf. Section 6.1.1.1.4). Because of their present high market demand R6-sized batteries are produced worldwide at very high production volumes.

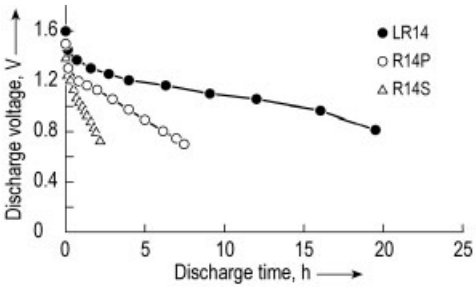

Figure 22. 3.9 Ω continuous discharge of R14S, R14P, and LR14 batteries

Besides single cell cylindrical batteries, also multicell cylindrical batteries are under production, which are constructed from in series connected button cells.

Button batteries are usually referred to as *button cells* although they comply with the criteria for single cell batteries. Single cell button batteries are characterized by their physical shape that resembles a button. Primary button batteries, using an alkaline electrolyte have a maximum diameter of $d = 11.6$ mm. The ratio between their diameter d and their height h is typically $1.5 < d/h < 6$. A schematic drawing of a LR-button battery is given in Figure 23.

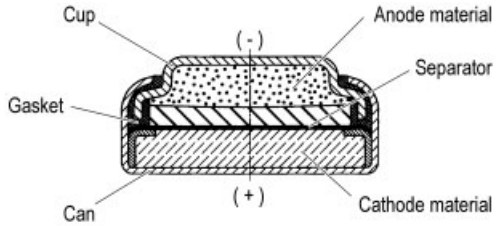

Figure 23. Schematic drawing of a (single cell) button battery

Presently five popular sizes for the manganese dioxide-zinc system are available on the market (see Table 5).

Multicell button batteries are constructed from several single cells as a pile, being connected in series, while having a common cylindrical envelope. They are available with different nominal voltages like 6 V, 12 V, 15 V and 22.5 V. For this type of batteries diameter ≤ height is valid. They therefore may be called *multicell cylindrical batteries*.

The *application* of multicell button batteries is manifold. Their predominant application is in the area of photo- and remote control equipment. Also single cell button batteries find widespread application. Specific application areas are electronic equipment like small hand held calculators with LCD display and low cost wrist watches.

6.1.3.1.5. Rechargeable Alkaline Manganese Dioxide – Zinc System

Rechargeable batteries are subject of Chapter 7. Since, however, some of the background of the rechargeable alkaline manganese dioxide – zinc system is already given in Section 6.1.3.1.2, the system's rechargeability will be discussed here.

The rechargeable alkaline manganese dioxide – zinc system was pioneered by K. V. Kordesch and co-workers ([64] and references cited therein, [56], [65]). One of the more attractive properties of this system is its low cost in comparison to other rechargeable systems.

True rechargeable systems are reversible systems in both respects: electrode structure and

Table 4. Features and applications of standard 1.5 V cylindrical zinc chloride and alkaline manganese – zinc batteries

IEC max. dimensions, mm (Ø: diameter, *h*: height)	Ø = 12.0 *h* = 30.2	Ø = 10.5 *h* = 44.5	Ø = 14.5 *h* = 50.5	Ø = 50.0 *h* = 26.2	Ø = 34.2 *h* = 61.5
Carbon – zinc batteries, IEC designation	R1	R03	R6	R14	R20
Typical capacity, mAh	350	500	1200	3200	8000
Alkaline MnO_2 – zinc batteries, IEC designation	LR1	LR03	LR6	LR14	LR20
Typical capacity, mAh	800	1100	2600	7800	16 500
Applications					
Portable lighting	x	x		x	x
Personal cassette players		x	x		
Tape recorder				x	x
Transistor radios		x	x	x	x
Pulse test (photo flash)		x	x		
Hearing aids	x				
Paging					
Motor/toy			x	x	x

Table 5. Alkaline manganese dioxide – zinc button batteries: IEC designation, dimensions, voltage, and capacities

IEC type	Diameter, mm	Height, mm	Nominal voltage, V	Typical capacity, mAh
LR41	7.9	3.6	1.5	30
LR55	11.6	2.1	1.5	25
LR54	11.6	3.05	1.5	50
LR43	11.6	4.2	1.5	80
LR44	11.6	5.4	1.5	120

electrochemistry. In the manganese dioxide-zinc system of the reactions occurring at the manganese dioxide electrode only Reaction (6.22) complies with this requirement (or Reaction 6.23, respectively, which also includes the zinc electrode). This reaction takes place in the voltage range 1.55 V $> U >$ 1.0 V. Below 1.06 V Reaction (6.25) proceeds, which is accompanied by an increase in volume. This reaction is irreversible and is therefore to be avoided in order to maintain the system's cyclability.

The reaction occurring at the zinc electrode, (Eq. 6.21), is electrochemically reversible. Not reversible, however, is its structure on recharge (shape change, see also Section 4.1.1.2). Furthermore, after the first charge the deposited zinc is much finer than the originally granulated zinc powder, which may lead (in the discharged state) to zinc particles being electronically isolated thus lowering the zinc electrode's capacity. This can be counteracted by the addition of graphite. Hydrogen gassing, being typical for all zinc systems, is inhibited by the addition of organic inhibitors. Nevertheless, the rechargeability of the manganese dioxide – zinc system is limited.

Furthermore, when charging the system it must be ensured that the charge voltage is limited to 1.65 – 1.68 V. Charging to higher voltages produces soluble manganese species and oxygen which oxidizes the zinc and/or increases the internal pressure, both processes resulting in a loss of cycle capacity. In no case a constant-current charger should be employed. This could lead to an uncontrolled built-up of internal gas pressure, such that the safety vent of the battery may open, permitting the electrolyte to escape.

The dimensioning of the zinc electrode determines the capacity and thus roughly the endpoint voltage of the battery. A better choice would be a voltage control by, e.g., electronic means to avoid the two-electron step (Eqs. 6.22 and 6.25), which occurs below 1.06 V. If not, the endpoint voltage of discharge will approach 0.8 V. Under this condition it takes about 20 discharge cycles until the capacity drops to about 50 % of the initial capacity, which is 70 % of that of the primary battery. The number of useful cycles and cycle capacity increases, when the cells are not deep discharged to below 0.9 V but only partially

discharged, while avoiding the irreversible two-electron discharge step.

An endpoint voltage of 0.9 V leads to much better results (see Fig. 24). In Figure 24 the cycle capacity and cumulative capacity for a R6(AA)-sized battery are plotted versus the number of cycles. An interesting result is that the cycle capacity decays linearly with the logarithm of cycle number.

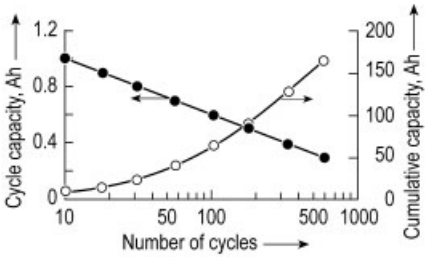

Figure 24. Cycle capacity and cumulative capacity versus the logarithm of cycle numbers of rechargeable alkaline manganese – zinc batteries

The construction of the rechargeable alkaline manganese dioxide – zinc battery is practical identical to that of the primary version (cf. Fig. 21). Essential differences exist, however, in the composition of the positive and the negative electrodes, i.e., to impede hydrogen gassing, to improve oxygen evolution during charge and to enhance cycle life, while limiting the discharge preferably to the one-electron step of Equation (6.22).

The rechargeable alkaline manganese dioxide – zinc battery is offered in the standard sizes R03(AAA), R6(AA), R14 (C) and R20 (D), all of which are cylindrical batteries. The highest specific load capability is offered by the small sized batteries because of their relatively thin cathodes.

The application of the single cell battery is practically identical to its primary version, which is also true for its storage behavior (in a charged state). If connected in parallel to so-called bundle-cells the discharge current may be increased to the needs of the application.

6.1.3.1.6. Prospects

The prospects of the primary alkaline manganese dioxide – zinc system are very promising as already pointed out (cf. Section 6.1.3.1.1). The future success of the rechargeable version of the alkaline manganese dioxide – zinc battery will not only depend on the system's characteristics but also on the consumer. Which system he will give preference will determine whether or not a substantial market will develop. This market will probably be identical to the market of present primary standard cylindrical batteries. New markets for the rechargeable version may open up, especially in the area of applications that only need low-rate shallow cycles in the voltage range 1.5 to 1.0 V per cell. In principle also high-rate applications are not out of range. The rechargeable alkaline manganese dioxide – zinc battery is, however, not expected to replace systems used in present automotive starting applications [66] like, e.g., the lead – acid system.

6.1.3.2. Silver Oxide – Zinc System

6.1.3.2.1. General

The silver oxide – zinc system exists in two variants. Variant 1 uses monovalent silver oxide Ag_2O while variant 2 is based on divalent silver oxide AgO. Variant 1 permits the construction of silver oxide batteries offering a relatively high energy density at a high load capability, including a flat discharge characteristic.

Batteries made by use of the divalent silver oxide system offer ca. 30 % more capacity in comparison to the monovalent silver oxide – zinc system. Silver is a precious metal and as such expensive. Much work has therefore been dedicated in the past to commercially employ the divalent silver oxide system, unfortunately so far with limited success, apart from a few years where they were in production.

6.1.3.2.2. Electrochemistry

Divalent Silver Oxide – Zinc System. The first step of electrochemical reduction of the divalent silver oxide is described by Equation (6.30)

$$2\,AgO + H_2O + 2\,e^- \longrightarrow Ag_2O + 2\,OH^- \qquad (6.30)$$

In combination with the electrochemical zinc oxidation according to

$$Zn + 2\,OH^- \longrightarrow ZnO + H_2O + 2\,e^- \quad (6.31)$$

the overall reaction of the first reduction step is derived

$$2\,AgO + Zn \longrightarrow Ag_2O + ZnO \quad (6.32)$$

Using the values of the Gibbs free energies the standard voltage of Equation (6.32) is calculated to be $U = 1.837$ V.

Since the electronic conductivity of AgO is poor, the AgO powder needs to be mixed with some conductive material like graphite. Under this condition the system's d.c. resistance is still relatively high and permits only a low rate discharge according to Equation (6.32), resulting in a discharge plateau at approx. 1.8 V.

Monovalent Silver Oxide – Zinc System. As soon as all of the divalent silver oxide is reduced to monovalent silver oxide according to Equation (6.32) the continuation of the reduction process follows Equation (6.33). Since the electronic conductivity of Ag_2O is even worse than that of AgO, the availability of conductive material in the Ag_2O powder is crucial, at least initially. As soon as metallic silver Ag is produced according to Equation (6.33) the system's internal resistance drops by about a factor of 30 to rather low values and the system becomes finally a high-rate system permitting discharge loads corresponding to current densities of up to 100 mA/cm^2 and even above [67].

On discharge the monovalent silver oxide – zinc system reacts in a heterogeneous reaction, which is reflected by a flat discharge characteristic. It is based on cathode reaction (6.33) and anode reaction (6.34)

Cathode reaction

$$Ag_2O + H_2O + 2\,e^- \longrightarrow 2\,Ag + 2OH^- \quad (6.33)$$

Anode reaction

$$Zn + 2\,OH^- \longrightarrow Zn(OH)_2 + 2\,e^- \quad (6.34)$$

Reaction (6.34) is followed by a decomposition of zinc hydroxide according to Reaction (6.35)

$$Zn(OH)_2 \longrightarrow ZnO + H_2O \quad (6.35)$$

The overall reaction is

$$Ag_2O + Zn \longrightarrow 2\,Ag + ZnO \quad (6.36)$$

The standard voltage of Reaction (6.36) is $U = 1.593$ V. The discharge voltage plateau at a medium discharge rate stabilizes at ca. 1.55 V.

The overall reaction that includes both discharge steps is given by

$$2\,AgO + 2\,Zn \longrightarrow 2\,Ag + 2\,ZnO \quad (6.37)$$

6.1.3.2.3. System-Immanent Properties

The theoretical values of the two discharge plateaus are:

First voltage plateau ($U = 1.83$ V)
Specific Energy: 259 mWh/g
Energy density: 1899 mWh/cm^3
Second voltage plateau ($U = 1.59$ V)
Specific Energy: 225 mWh/g
Energy density: 1651 mWh/cm^3

The total of both discharge plateaus is obtained by their addition:

Specific Energy: 484 mWh/g
Energy density: 3550 mWh/cm^3.

The applicability of the silver oxide system would undoubtedly increase, if it could be modified such that it offered the capacity of the divalent silver oxide at the discharge voltage and load capability of the monovalent silver oxide – zinc system.

Much work has been devoted in the past to this task [68–72]. The scientific background of this work was primarily built on the findings of Wales, Burbank, and Dirkse [73].

The basic idea of the development work was to (electro-)chemically reduce the surface of the AgO powder particles according to Equation (6.30) such that the AgO particles are covered by a layer of Ag_2O. In the second step these Ag_2O/AgO particles are contacted with metallic silver, either by adding a small amount of metallic silver powder, or by chemical and/or electrochemical reduction. Under this condition the potential of the Ag_2O/Ag electrode is realized, although most of the silver oxide is still in its divalent state. In combination with a zinc anode a voltage of 1.59 V is established while the

system offers the desired high load capability of the monovalent and the extended capacity of the divalent silver oxide zinc system.

At a first glance the above ideas appear simple and easy to realize. This is indeed true for batteries being put into service shortly after the reduction treatment or if stored only for a short period of time. For long term storage (a couple of months) problems are most likely to arise for the following reasons:

It is difficult to really produce a tight Ag_2O layer around all of the individual AgO particles, such that the AgO phase does not get into contact with the alkaline electrolyte.
If it gets into contact with the electrolyte the divalent silver oxide will decompose and/or oxidize the metallic silver.

At ambient temperature the decomposition reaction of divalent silver oxide occurs relatively slowly

$$2\,AgO \longrightarrow Ag_2O + \tfrac{1}{2}\,O_2 \tag{6.38}$$

However, in a hermetically sealed battery this will lead to a noticeable pressure built-up during storage, especially at elevated temperatures. Reaction (6.38) needs therefore to be inhibited. This can be achieved by the addition of, e.g., gold oxide [74], lead and/or cadmium compounds, which means either a relatively expensive treatment or the addition of environmentally harmful compounds.

The other obstacle of the "electrolyte-contacted AgO" is, that it oxidizes the surface of the metallic silver

$$AgO + Ag \longrightarrow Ag_2O \tag{6.39}$$

The voltage will only be stable at 1.59 V if the Ag surface is large enough in relation to the AgO surface being in contact with the electrolyte. If the Ag surface is oxidized before the AgO is consumed the potential of the system will rise again to that of the AgO/Ag_2O system, which – in combination with the zinc electrode – means a voltage of 1.83 V (cf. Eq. 6.32). Simultaneously the internal resistance rises again to its initial high value. However, on discharge the voltage plateau of the monovalent silver oxide – zinc system is reached again rather quickly.

Another important component is the *separator*, which determines the storage life of the silver oxide battery. In contrast to other cathode materials like MnO_2 and HgO, silver oxide is quite soluble in alkaline solutions (approx. 50 mg/L). The ions present in the alkaline solution are reported in the literature as $Ag_3O(OH)_2^-$, $Ag(OH)_2^-$ or AgO^- [75], [76]. Cellophane was the first separator that made it possible to manufacture silver oxide – zinc cells with a reasonable shelf life [77]. Since 1970 highly improved separator systems have become available [78], [79]. They are based on stable materials like grafted polyethylene (PE) or polypropylene (PP) in combination with laminates of cellophane. The separator system has been studied in detail [80] with the result that the diffusion of dissolved silver oxide species through the separator, an important factor for self-discharge, can be excluded. More important is the influence of self-discharge resulting from the oxidation of organic cell components and the cellophane. Decisive is the separator's tendency to allow for the formation of metallic silver bridges, their resistance being in the range of 0.5 to 12 kΩ. The formation of silver bridges is governed by statistical influences, i.e., after a long term storage some cells out of many may still exhibit nearly their full capacity while others are partially or even fully discharged.

The system employs a zinc (powder) electrode as anode, which is responsible for another system-immanent property, i.e., hydrogen evolution. The problems involved are identical to what has already been discussed (see, e.g., Section 6.1.3.1.3, Eq. 6.29).

Typical for the monovalent silver oxide – zinc system is its rather flat discharge curve and its rather low internal resistance. At a current density of 0.5 mA/cm^2 its discharge voltage is in the range $1.5\,V > U > 1.6\,V$ during 90 % of discharge time at less than $30\,\Omega \times cm^2$ (d.c. resistance).

6.1.3.2.4. Batteries and Their Application

Because of its high material costs, the commercial application of this system is limited to small sized batteries like button batteries, having an energy content of less than 0.5 Wh. For military application, however, battery units of up to 20 kWh have been realized (cf. Section 6.3.1).

The IEC system letter for the monovalent silver oxide zinc system is S. In combination with letter R, i.e., SR, a round battery is addressed, which includes both, cylindrical and button batteries. The divalent silver oxide system was designated with letter T as long a it was an IEC standardized system.

Batteries made from the Monovalent Silver Oxide – Zinc System. The single cell button battery is the most common construction for the alkaline silver oxide – zinc system in commercial application. The construction is practically identical to that shown in Figure 23.

Because of the problem of electrolyte leakage caused by under-potential deposition (see Section 6.3.1) there are two versions of silver oxide – zinc button batteries in use. The version that employs KOH is especially used for high rate applications, e.g., to power light emitting diodes (LEDs). The battery's lifetime in high-rate applications is relatively short, such that leakage problems do not show up during its active service life.

For long-term applications (i.e., low rate discharge) preference is given to the NaOH electrolyte. Its resistance is higher than that of the KOH electrolyte but the tendency to cause leakage at the negative electrode is – for above reasons – significantly reduced [82].

More than 25 different sizes of silver oxide – zinc button batteries are available on the market. Present batteries cover following ranges of dimensions and capacities:

Diameter: $5.8\,\text{mm} \leq \phi \leq 11.6\,\text{mm}$
Height: $1.25\,\text{mm} \leq h \leq 5.4\,\text{mm}$
Capacity: $5\,\text{mAh} \leq \text{c} \leq 170\,\text{mAh}$

Their specific energy and their energy density are far below the theoretical values and depend very much on battery size: The smaller the battery, the smaller its corresponding energy value due to the increasing percentage of inactive weight and volume of battery piece parts. Typical values are in the range of:

Specific energy: $52\,\text{mWh/g} < E*(\text{w}) < 113\,\text{mWh/g}$

Energy density: $235\,\text{mWh/cm}^3 < \text{E}*(\text{v}) < 462\,\text{mWh/cm3}$

If the electrochemical active materials would determine weight and volume of the battery, its specific values would increase to

Specific energy: 350 mWh/g (extrapolated value)

Energy density: 700 mWh/cm^3 (extrapolated value)

Since their main application is in the area of electronic wrist watches they are primarily addressed as *watch batteries*; but they are also in use in some photo applications.

Batteries made from the Divalent Silver Oxide – Zinc System. Button batteries constructed from the divalent silver oxide – zinc system to enter the market in 1974 – 1975. For larger sized button batteries they offered a capacity being 35 % above the capacity of the monovalent silver oxide zinc system. This was of considerable interest in those days when LED wrist watches were largely in use, because these wrist watches have a high current demand, especially in combination with additional functions. In the following decades, however, the technology changed and electronic wrist watches became smaller and smaller, while their current demand dropped. Consequently the batteries became smaller too. Smaller sized divalent silver oxide button batteries, however, offered less than an 35 % capacity advantage over the version with monovalent silver (see above).

This change in technology led finally to the disappearance of the divalent silver oxide system from the market, which in part was also due to the processing costs and difficulties to have its voltage stabilized.

6.1.3.2.5. Prospects

The divalent silver oxide system may come back if the market should demand again larger-sized high drain silver oxide – zinc button batteries. The monovalent silver oxide – zinc system is expected to remain in the commercial market as a button battery. It will continue to be the ideal energy source to power multifunction high quality wrist watches, photo devices, and appliances that need a constant voltage source, while offering a good rate capability.

6.1.3.3. Mercuric Oxide – Zinc System

6.1.3.3.1. General

The mercuric oxide – zinc system was pioneered by SAMUEL RUBEN [83] and developed during World War II. The system is known for its high energy density, constant voltage, good load capability, good temperature and excellent storage behavior. An important requirement for an excellent storage behavior is, among others, the use of high purity mercuric oxide, which renders the system relatively expensive.

In spite of its favorable performance characteristics the use of mercuric oxide – zinc battery has increasingly declined within the last decade because of the environmental problems involved. Whenever possible it is replaced meanwhile by other systems like lithium-based batteries, alkaline manganese dioxide – zinc, silver oxide – zinc, and air – zinc systems. A detailed description of the mercuric oxide – zinc system including battery sizes and their application etc. is given in [84].

6.1.3.3.2. Electrochemistry

Since mercuric oxide is a poor electronic conductor it has to be mixed with graphite to allow the cathode reaction (Eq. 6.40) in an alkaline electrolyte at reasonable rates

$$HgO + H_2O + 2e^- \longrightarrow Hg + 2OH^- \tag{6.40}$$

in combination with the anode reaction (Eq. 6.41)

$$Zn + 2OH^- \longrightarrow ZnO + H_2O + 2\,e^- \tag{6.41}$$

the overall reaction (Eq. 6.42) is derived

$$HgO + Zn \longrightarrow Hg + ZnO \tag{6.42}$$

6.1.3.3.3. System-Immanent Properties

Based on Gibbs free energies a standard voltage of $U^o = 1.35$ V is calculated from Equation (6.42), which is in good agreement with experimental observations. Theoretical values on specific energy and energy density are calculated from Equation (6.42) as well. They are solely based on active materials:

Specific energy: 256 mWh/g
Energy density: 2526 mWh/cm^3

Corresponding values of commercial products are far below. Larger sized products (R20 size and above) may reach values (extrapolated) up to:

Specific energy: 160 mWh/g
Energy density: 800 mWh/cm^3

Smaller sized button batteries at a capacity of 50 to 450 mAh cover a range of

Specific energy: 80 – 130 mWh/g
Energy density: 380 – 550 mWh/cm^3

Mercuric oxide is not readily soluble in the alkaline electrolyte, i.e., any self-discharge due to dissolved species does not occur. In comparison to silver oxide, mercuric oxide has a rather low oxidation potential, such that it does not attack the separator system.

A less favorable system-immanent property is the formation of microdroplets of mercury, generated during discharge according to Reaction (6.40), which may agglomerate. If so, larger agglomerations may finally cover up areas of active material and impede its reaction. Fortunately graphite powder being added (3 % to 10 %) to improve the cathode's electronic conductivity assists in the prevention of mass agglomeration of mercury droplets as do other additives like manganese dioxide. This addition increases the cell voltage to approx. 1.4 V.

A careful selection is also required for the separator system. Its porosity must be such that it by no means permits the penetration of microdroplets of mercury, which otherwise could establish electron conducting bridges with subsequent self-discharge. Another requirement is the mechanical stability of the separator in alkaline electrolyte (e.g., 30 % KOH) over prolonged periods of time at elevated temperatures (e.g., 40 to 60 °C). The identical requirement exists for the battery's plastic seal.

Typical for the mercuric oxide – zinc system is its nearly constant voltage discharge. At a current density of 0.5 mA/cm^2 its discharge voltage is in the range $1.35\,V > U > 1.30\,V$ during approx. 90 % of discharge.

Another system-immanent property are unfortunately the ecological problems involved. These problems have made the mercuric oxide – zinc system practically disappear from the commercial market. As long as the battery is hermetically sealed no problem exists. There would also be no problem if approx. 100 % of sold batteries would be returned after use for recycling. Since this cannot be enforced, batteries may be disposed off uncontrolled and open up one day by, e.g., corrosion or mechanical force, such that its contents may get into contact with the environment and form soluble compounds. The worst case would be the formation of soluble organic compounds, since they exhibit the highest toxicity.

6.1.3.3.4. Batteries and Application

System letter of the IEC for the mercuric oxide – zinc system was letter M. In combination with letter R, i.e., MR, a round battery was addressed. Meanwhile the mercuric oxide – zinc system is no longer included in the IEC Standard Public. 86 [26], [33].

The prevailing application of this miniature battery has been in the area of hearing aids, electronic wrist watches, photo and all kinds of electrical instruments. In this application area the prevailing construction is/was the button battery. The constructional details are identical to those shown in Figure 23.

Besides button batteries also cylindrical batteries exist(ed) on the market at nominal voltages between 2.7 and 7 V and capacities of 150 up to 1000 mAh.

6.1.3.3.5. Prospects

There is still a large number of electronic devices which only work properly on mercuric oxide – zinc batteries. As long as these devices are in use, the demand for mercuric oxide batteries will continue. But in the long term these devices will be replaced, the production of mercuric oxide batteries will probably fade away.

It is safe to assume that – because of the system's ecological problems – it will not be employed again in any new commercial application.

6.1.3.4. Alkaline Air – Zinc System

6.1.3.4.1. General

The IEC has allocated system letter P to the alkaline air – zinc system. The IEC, however, addresses this system as the "oxygen alkali metal hydroxide zinc" system, which actually should be changed to read the "air alkali metal hydroxide zinc" system, since there are distinct differences between the two systems. In contrast to the air electrode the oxygen electrode of the alkaline oxygen – zinc battery does practically not have any gas diffusion problems and therefore would permit a different construction. The reason for the diffusion problems of the air electrode (especially at higher discharge rates) is the high nitrogen content of the air (ca. 80 vol %). In consequence of this the alkaline oxygen – zinc system has a much higher rate capability than the alkaline air – zinc system and in addition a slightly higher open circuit voltage due to the higher partial oxygen pressure. It furthermore needs no fan for gas circulation as needed for larger air – zinc battery systems. Another principal advantage of the oxygen – zinc system over the air – zinc system is the fact that it is a closed system, i.e., there is no exchange of gaseous species between the system and its environment.

In comparison to the air – zinc system, however, the oxygen – zinc system has one big disadvantage: It needs a special container to store the oxygen, which increases the systems volume, weight, and costs. This is probably the reason why no commercial application for the oxygen – zinc system has emerged so far.

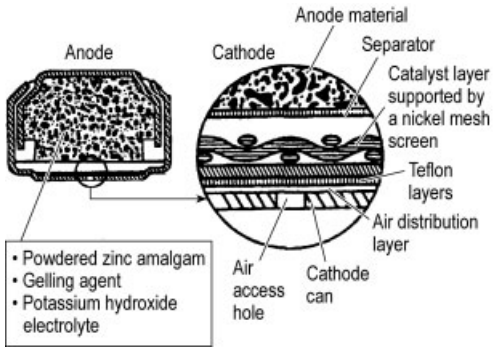

Figure 25. Schematic drawing of the cross-section of an air – zinc button battery including a close-up drawing of its air cathode arrangement

The alkaline air – zinc system is known since 1925 at the latest (cf. Chap. 2). The system's key component is its air cathode, which may be considered as a spin-off of fuel cell gas-electrode technology. A fuel cell gas-electrode needs – to be operable – also a three-phase boundary [85] being composed of the gas phase (air), the catalytic phase (electronic conductor) and the ion conducting phase (electrolyte). In this context the development of hydrophobic compounds like polytetrafluoroethylene was crucial. This is the reason why the system's commercialization did not start before the early 1970s [86].

Figure 25 is a schematic cross-section of an air zinc button battery with a close-up of its air cathode. The thickness of this multilayer air cathode is approximately 0.5 mm. In comparison to the air cathode of the weakly acidic ammonium chloride – zinc system (see Section 6.1.1.2), which has a thickness of ca. 3 cm, this air cathode is extremely thin. This is one of the reasons for its high load capability, i.e., low diffusion polarization. Depending on battery size up to four air access holes are placed into the bottom of the cathode can.

6.1.3.4.2. Electrochemistry

The electrochemical oxygen reduction can be described by Reaction (6.43)

$$\frac{1}{2}\,O_2 + H_2O + 2e^- \longrightarrow 2\,OH^- \qquad (6.43)$$

This reaction proceeds at the three-phase boundary of the porous air cathode, as already described above.

The reaction of the zinc powder electrode is identical to what has already been described in Section 6.1.3.1.3.

$$Zn + 2\,OH^- \longrightarrow ZnO + H_2O + 2\,e^- \qquad (6.44)$$

Equation (6.44) is simplified. The potential-determining reaction is given by [87]

$$Zn + 4\,OH^- \longrightarrow Zn(OH)_4^{2-} + 2\,e^- \qquad (6.44a)$$

it is succeeded by the decomposition reaction

$$Zn(OH)_4^{2-} \longrightarrow ZnO + H_2O + 2\,OH^- \qquad (6.44b)$$

The overall reaction is derived from Equations (6.43) and (6.44)

$$\frac{1}{2}\,O_2 + Zn \longrightarrow ZnO \qquad (6.45)$$

6.1.3.4.3. System-Immanent Properties

The alkaline air – zinc system is remarkable in various aspects:

The standard potential of Equation (6.45), calculated under equilibrium conditions from Gibbs free energies, is $U^0 = 1.647$ V which may be realized under ultrapure conditions (i.e., absence even of traces of reducing substances [88]). Under practical conditions, however, the reversible potential is not established. The OCV of an practical air – zinc cell tends to be 200 mV below the standard value, i.e., approx. 1.44 V. This is reportedly influenced by a hydrogen peroxide mechanism acting in parallel [89].

Furthermore, the partial oxygen pressure in the case of air is approx. only 1/5 of the standard value as used for the calculation of the oxygen standard potential, which, however, accounts only for a few millivolts less (a decrease in oxygen pressure by a factor 10 decreases the equilibrium potential only by 15 mV and vice versa).

As a single cell battery the alkaline air – zinc battery has the highest energy density of all electrochemical systems. A button battery offers up to 1200 mWh/cm^3 at 365 mWh/g. These specific values drop for larger multicell battery units to 250 mWh/cm^3 and 170 mWh/g, if auxiliary equipment like fan, heat exchanger, humidity control, etc., are necessary for system operation [90].

At ambient temperature its load capability is equal if not higher than that of the mercuric oxide – zinc system, i.e., 30 mA/cm^2 and above, depending on the potential of the uncovered cathode, when exposed to air.

The system is environmentally benign and therefore a very good replacement for the mercuric oxide – zinc system, especially in the hearing aid application.

A comparison of the design of the air – zinc battery and other zinc systems, explaining the high energy density of the air – zinc system, is given in Figure 26.

Both button batteries use the same number of components. However, in contrast to the mercuric oxide – zinc battery the air – zinc battery must not store its oxidant (atmospheric oxygen). Therefore the air-zinc battery can store more zinc than the mercuric oxide – zinc battery, which explains its higher energy content. However, the anode compartment of the air – zinc battery must in addition provide some extra space, because of the volume increase by 65 % due to Reaction (6.45). Otherwise the air – zinc battery would swell and leak on discharge. The same arguments are valid in comparison to other metal oxide batteries like the silver oxide – zinc and the manganese dioxide – zinc system.

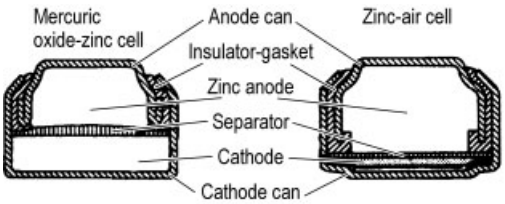

Figure 26. Cross-section of an air – zinc button battery in comparison to a mercuric oxide – zinc button battery

Besides these advantages the air – zinc system has one significant disadvantage in comparison to other systems. This is due to the fact that the system is open to air. Under this condition two deteriorating processes are acting in parallel. The carbonation of the alkaline electrolyte and exchange of water with the environment.

$$2\,\mathrm{KOH}\ (\text{electrolyte of battery} + \mathrm{CO_2}\ (\text{from outside}) \longrightarrow \mathrm{K_2CO_3} + \mathrm{H_2O} \qquad (6.46)$$

Equation (6.46) describes the carbonation of the KOH electrolyte by carbon dioxide that diffuses through the air access opening into the battery's interior. Although the air contains only 0.03 % carbon dioxide it will by and by use up the alkaline electrolyte, and thus increase the electrolyte's resistance. The final result is given by Reaction (6.46).

In parallel to Reaction (6.46) water is exchanged in the form of vapor either leaving the battery and diffusing away into the environment or coming from the environment and being absorbed as described by Reaction (6.47). The direction of diffusion is determined by the difference in relative humidity ΔRH between the inside of the battery RH_i and the environment RH_e.

$$\mathrm{H_2O}\ (\text{battery}) \longleftrightarrow \mathrm{H_2O}\ (\text{environment}) \qquad (6.47)$$

Typical values for the relative humidity inside the battery are in the range $50\ \% < RH_\mathrm{i} < 60\ \%$, depending on the concentration of the KOH-electrolyte. The environmental humidity RH_e varies in the range $10\ \% < RH_\mathrm{e} < 95\ \%$, the low value is represented, e.g., by a desert climate or an electrically heated room in wintertime without humidifier. The high value being represented by a wet environment, e.g., a tropical climate, a rainy day or an in-house sauna.

To illustrate the consequences of the environmental humidity and related properties, three different cases are discussed below:

1) Low environmental humidity: $\Delta RH > 0$; $RH_\mathrm{i} > RH_\mathrm{e}$.
   The air – zinc battery will dry out, a process additionally favored by Reaction (6.46). As a consequence the battery's internal resistance will increase, which finally leads to a bad battery performance and a short battery life. This effect is sometimes referred to as self-discharge, which is not correct since this effect has nothing to do with electrochemistry. The above effect may not be noticed if the battery is discharged quickly enough before it has lost too much water and/or if the condition $\Delta RH > 0$ lasts only for a short time and/or if its value is very small.
2) High environmental humidity: $\Delta RH < 0$; $RH_\mathrm{i} > RH_\mathrm{e}$
   The air – zinc battery will pick up additional water from outside. Under this condition the battery will swell and may finally leak. This situation becomes worse when the battery is additionally discharged, because when zinc is oxidized to zinc oxide according to Reaction (6.45) its partial molar volume increases by ca. 65 %.
   The above effect may not be noticed if the battery is discharged quickly enough before substantial amounts of water have been absorbed and/or if the condition $\Delta RH < 0$ lasts only for a short time and/or if it is very small.

3) Equilibrium humidity: $\Delta RH = 0; RH_i = RH_e$
This is the ideal case since Equation (6.47) is in equilibrium and would thus permit an unlimited storage life. This, however, is only true in a CO_2 free atmosphere, which unfortunately does not exist. Because of Equation (6.46), the relative humidity RH_i inside the cell would constantly rise to higher values thus initiating the process as already described in case 1), although at a much lower rate.

6.1.3.4.4. Air Access Opening: Its Seal and Diffusion Resistance

To ensure a good storage behavior of the air – zinc system its air access opening(s) must be absolutely sealed during storage to avoid deteriorating transport processes according to Equations (6.46) and (6.47).

When closing the air access opening(s), the system is deactivated, i.e., no current can be drawn anymore. Small batteries, like button cells, have typically one to four air access holes (depending on size, load capability and storage behavior), which have to be sealed during storage. This is usually done by an adhesive foil, preferably a metal foil. Plastic tape, e.g., is not suitable for very long storage periods, since it would permit the diffusion of oxygen, water vapor, and carbon dioxide, although at a highly reduced rate. Of significant importance is furthermore the adhesive that glues the foil to the battery surface. The adhesive must last at least the time of storage. It must further be ensured that the hydrogen evolution at the anode is minimized to an acceptable degree (cf. Section 6.1.3.1.3, Eq. 6.29). Otherwise the foil would be lifted, resulting in a bad storage behavior. Before use the foil must of course be removed from the battery (*activation*). In this respect the air – zinc battery may also be considered as a reserve battery (see Section 6.3).

Larger battery systems, using, e.g., a fan for air supply and circulation have other constructional means to suppress Reactions (6.46) and (6.47). They employ, e.g., flaps for the air in- and outlet and a CO_2 scrubber to eliminate the carbon dioxide from the air.

An interesting characteristic of the air – zinc battery is its limiting current i_L given by Equation (6.48)

$$i_L = G D_{O_2} \frac{4F}{RT} \Delta p^o_{O_2} \tag{6.48}$$

The limiting current i_L is defined for the case that the partial pressure of oxygen at the three-phase boundary (i.e., reaction zone) of the oxygen electrode is zero, i.e.

$$\Delta p_{O_2} = p^o_{O_2} - p^b_{O_2} = p^o_{O_2}; p^b_{O_2} = 0$$

$p^o_{O_2}$: partial oxygen pressure outside the battery
$p^b_{O_2}$: partial oxygen pressure at three-phase bondary
D_{O_2}: oxygen diffusion coefficient
G: geometric electrode factor

The limiting current i_L is experimentally determined by means of a voltage – current diagram. When constantly increasing the current the voltage finally drops more than proportionally. The limiting current is the current that causes in the end an abrupt drop in electrode potential. A typical diagram for a PR44 button battery (see Table 6) is given in Figure 27. The PR44 battery has two air access holes symmetrically arranged within the bottom area of the cathode can, each one having an diameter of ca. 0.5 mm and being separated from each other by ca. 5 mm.

Figure 27 presents two current – voltage diagrams with a sharp voltage drop at their end, which is related to a limiting current. With two air access holes the limiting current is ca. 22 mA. With one air access hole the limiting current is about half of it, i.e., 12 mA. The i_L value for the unprotected open air cathode is 100 mA, i.e., 150 mA/cm^2.

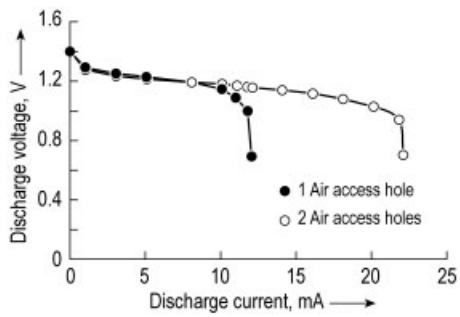

Figure 27. Current – voltage diagram and limiting current i_L of a PR44 air – zinc battery

Table 6. Air – zinc button batteries for hearing aids (presently standardized single cell button batteries, IEC code P)

IEC type	Diameter, mm	Height, mm	Nominal voltage, V	Typical capacity, mAh
PR	5.8	2.15	1.4	33
PR70	5.8	5.8	1.4	60
PR41	7.9	3.6	1.4	130
PR48	7.9	5.4	1.4	230
PR44	11.6	5.4	1.4	540

For the air – zinc system under given conditions the limiting current is proportional to the mass transport according to Equations (6.46) and (6.47), i.e., a high limiting current means a high rate of water loss (depending on $\Delta RH > 0$) and vice versa. The same is true for the rate of carbon dioxide pick-up, i.e. the higher the limiting current, the higher the absorption rate of carbon dioxide (depending on the partial pressure of CO_2).

Normal current loads $i < i_L$, should not exceed 1/2 of the limiting current i_L in order to realize an acceptable discharge voltage of the air – zinc battery. Therefore, if a given battery is able to supply 100 mA constant current while the application demands only 10 mA, the diffusion resistance should be increased to provide only 10 mA at an acceptable discharge voltage. This avoids an unnecessary high rate of water loss and CO_2 pick up during battery operation, resulting thus in a prolonged service life of the battery.

This optimization principle allows to realize, with otherwise identical battery parts, either a high-rate battery, having a relatively short service life after activation, or a low-rate battery having a relatively long service life, just by adapting the diffusion resistance to the current needs of the application.

6.1.3.4.5. Batteries and Their Application

Because of their attractive high energy density and high specific energy content, there have been many attempts to construct special air – zinc batteries, even as mechanically rechargeable batteries in the area of vehicle propulsion [91], [92] with an energy content of up to 110 kWh at approx. 50 kW power output. So far, however, all these activities have failed to enter the commercial market for various reasons, with the exception of primary button batteries. Button batteries are mainly applied in the area of hearing aids and to a smaller extent in electronic pagers.

Besides its electrical/electrochemical attractiveness the air – zinc system is, in comparison to the mercuric oxide – zinc system, also an environmentally benign system. Since the discharge voltage of the air – zinc system is close to that of the mercuric oxide – zinc system, and this at an improved energy density and load capability, it became soon its ideal replacement. The capacity of identical sized button batteries of the mercuric oxide – zinc system is by a factor 2.5 less than that of the air – zinc system which also has favored the replacement of the mercuric oxide – zinc system. Best results with air – zinc batteries are obtained if the hearing aid is used 12 to 14 h per day. Dissatisfaction will arise if the air – zinc hearing aid battery would be used only occasionally (e.g., once a week), because it probably will dry out before its stored electric energy is used up.

Presently five different button battery sizes are available on the market, designed to power different types of hearing aids (see Table 6).

The first column lists the international designation according to the IEC (manufacturer's designations differ in general from IEC designations). As can be seen from Table 6 the smallest type of air – zinc button batteries has not been standardized yet.

6.1.3.4.6. Prospects

In order to be successful in a given application, a new system must offer a technical, environmental and/or economical advantage over the existing system.

In the area of hearing aid batteries the air – zinc system has been very successful. It is therefore expected that this system will continue to stay in the market.

In general the alkaline air – zinc system should have good prospects in all those applica-

tions that discharge the battery to its end-point before it starts to deteriorate noticeably due to mass transfer according to Equations (6.46) and (6.47). The air – zinc system should also be very useful as a reserve system to be activated via electrolyte and/or air in application areas where a low weight and small volume are decisive.

Since the air cathode is not used up during discharge it should generally be possible to use it in a rechargeable system. Recharging could occur either mechanically (i.e., replacing the discharged zinc by fresh zinc and fresh electrolyte) or electrically by reducing the zinc(II) to metallic zinc. Both methods have been studied extensively in the past [91] and later by AER Energy Resources, Inc., Smyrna, Georgia, USA, however, so far without commercial success. Among other problems it is obviously difficult to develop an air cathode with a three-phase boundary, that is locally stable within the electrode's porous structure while remaining active over a sufficient length of time on repetitive use.

If present problems should be overcome it would be easy to predict a number of new commercial applications for the alkaline air – zinc system.

6.2. Batteries with Nonaqueous Electrolyte

Electrochemical battery systems employing a nonaqueous electrolyte are governed by lithium metal as anode material, in combination with a suitable cathode material. These batteries are characterized by some outstanding properties like a high energy density, high specific energy, low self-discharge, and a wide temperature range, within which they can be operated.

Several cathode materials (solid and soluble) and several nonaqueous electrolytes (organic and inorganic), were developed in the past. Certain combinations of both allow to compose different lithium battery systems with different characteristics regarding voltage, load capability, capacity, storage behavior, etc.

Table 7. Three crucial properties of lithium in comparison to zinc

Property	Lithium	Zinc
Standard potential∗	− 3.045 V∗	− 0.76 V∗
Density	0.534 g/cm^3	7.13 g/cm^3
Specific capacity	3.86 Ah/g	0.82 Ah/g

∗ Standard potentials of metal ion electrodes at 25 °C in aqueous solution according to Latimer [93]

In this section some typical properties of these lithium battery systems will be presented, including their construction. Presentation will be limited to those that have attained market significance, apart from some exceptions.

The reason to devote research and development activities to batteries employing lithium as anode material is based on its extraordinary properties. Three of them are listed in Table 7 in comparison to zinc.

Lithium's affinity to react with water has, however, initially retarded technical battery development work (until the 1960s). Lithium reacts with water according to Equation (6.49) rather vigorously under practical conditions, producing lithium hydroxide, hydrogen, and heat.

$$2\,Li + 2\,H_2O \longrightarrow 2\,Li^+ + 2\,OH^- + H_2 \qquad (6.49)$$

Reaction (6.49) is valid for all aqueous electrolytes.

The first task was therefore to develop suitable non-aqueous electrolyte(s).

Nonaqueous Electrolytes. A nonaqueous electrolyte is suitable for lithium batteries if it offers the following properties:

1) Stability in contact with lithium
2) Sufficiently high ionic conductivity
3) Compatibility with the cathode material
4) Safety on handling
5) Environmental friendliness
6) Low cost

Nonaqueous electrolytes may be grouped into organic electrolytes, inorganic electrolytes, electrolytes with soluble cathode, and solid electrolytes.

Liquid electrolytes consist in general of a mixture of solvents with the addition of lithium-based conductive salt(s). Today it is generally accepted that none of the known solvents is thermodynamically stable in contact with lithium [94]. The commercial success of primary lithium

batteries is actually based on a passivating protective layer between the electrolyte and the surface of the lithium electrode [95]. The composition and thickness of this protective layer depends on the composition of the electrolyte and its purity. This protective interphase is called *solid electrolyte interphase* (SEI). It is, among others, responsible for the excellent storage behavior of lithium battery systems.

The SEI represents a layer of relatively low conductivity and causes initially a voltage drop upon application of a discharge load. On continuation of the load, however, the layer is reduced in thickness, which leads to a significant increase of its conductivity and consequently to a decrease of the initial voltage drop [96]. This phenomenon is called *voltage delay* (see Fig. 28). For more information on the SEI see [97].

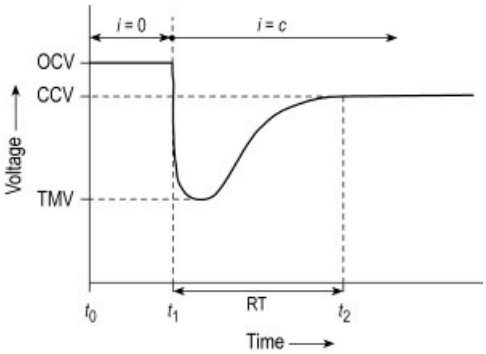

Figure 28. Schematic drawing of the initial voltage – time transient (voltage delay) in response to discharge current $i = c$
OCV = open circuit voltage ($i = 0$); CCV = closed circuit voltage ($i = c$); TMV = transition minimum voltage; RT = recovery time of the CCV until it becomes constant.

The parameters *transition minimum voltage* (TMV) and *recovery time* (RT), depicted in Figure 28, depend very much on the type of lithium system, the discharge load, and preceding storage conditions. TMV and RT generally increase with increasing discharge load, decreasing discharge temperature, and long preceding storage periods at elevated temperatures. The voltage delay is less pronounced for systems with organic electrolyte and solid cathode than in systems with soluble cathode materials, like thionyl chloride ($SOCl_2$) and sulfur dioxide (SO_2). Values of RT may reach up to 10 s.

In comparison to aqueous electrolytes, liquid nonaqueous electrolytes offer larger temperature ranges of operation, down to below − 150 °C [98] and up to above 300 °C [99] and voltage windows up to several volts. Their drawbacks are lower conductivity in comparison to aqueous systems (factor 10 to 100 and above), higher costs, inflammability, and environmental problems. For more information on nonaqueous electrolytes, see, e.g., [100].

Certain nonaqueous electrolytes fit best only certain cathode materials, they will be dealt with in the relevant sections.

Processing Techniques. Besides suitable nonaqueous electrolytes, also new processing techniques had to be developed to handle lithium metal in battery production, i.e., to protect it from any reaction with the ambient atmosphere, especially from air humidity. To be acceptable for sensitive operations, humidity must be reduced to the ppm range. Thus, relevant *dry rooms* had to be developed. Dry rooms also solved problems during handling of the nonaqueous electrolyte, i.e., its addition to the electrode system before battery closure. Under normal humidity conditions (50 % to 60 % RH) the electrolyte would absorb to much humidity from the atmosphere during addition to the system and thus become unusable.

For special operations (research, small production volumes, storage, etc.) *glove boxes* are still in use, normally with an argon atmosphere inside. The glove box system permits to control not only the humidity but also the oxygen content of the inside atmosphere to very low levels if required.

6.2.1. Batteries with Organic Electrolyte

Systems using an organic electrolyte may be subdivided into systems having an operating voltage of approx. 3 V and systems with approx. 1.5 V. They all employ solid cathode materials, such as V_2O_5, Ag_2CrO_4, MoO_3, MnO_2 (3 V systems) and FeS_2, CuO, and Bi_2O_3 (1.5 V systems). The more important ones will be described below.

6.2.1.1. Manganese Dioxide – Lithium System

6.2.1.1.1. General

The manganese dioxide – lithium system was one of the first systems with solid cathode. It was pioneered by Sanyo Electric Co. The company started the development work in 1975 and has finally granted their manufacturing technology to more than 12 international operating battery manufacturers. Today the manganese dioxide – lithium system is the most widely spread commercial lithium primary battery system in the world. It is manufactured in various shapes and sizes such as cylindrical, button, and prismatic batteries.

The IEC has allocated system letter C to the manganese dioxide – lithium system. Example: CR2032 designates a round battery (R) of the MnO_2 – Li system (C), having a diameter of 20 mm and a height of 3.2 mm. Since $d > h$ applies, the round battery is a button battery by definition (for battery designation see also Chap. 8).

6.2.1.1.2. Electrochemistry

The simplified reaction for this system is described by:

Anode reaction

$$Li \longrightarrow Li^+ + e^- \tag{6.50}$$

Cathode reaction

$$MnO_2 + Li^+ + e^- \longrightarrow MnO_2^- \ (Li^+) \tag{6.51}$$

Overall reaction

$$MnO_2 + Li \longrightarrow MnO_2^- \ (Li^+) \tag{6.52}$$

where MnO_2^- (Li^+) symbolizes that the lithium ion is introduced into the MnO_2 crystal lattice [101], i.e., an intercalation compound is formed. The cathode material is a specially prepared and heat-treated form of MnO_2. The electrolyte consists generally of a solvent-mix of propylene carbonate (PC) and 1,2-dimethoxyethane (DME) with lithium perchlorate ($LiClO_4$) as conducting salt. At high discharge rates, the limited diffusion rate of Li^+ ions into the pore system of the MnO_2 cathode may result in a depletion of the conducting salt in the cathode pores, thus causing a premature end of discharge [102]. For this reason, cathodes for high rate discharge generally do not exceed 1 mm in thickness. They need a metal grid for mechanical support, which serves simultaneously as current collector.

6.2.1.1.3. System-Immanent Properties

Voltage characteristics:

Maximum off-load voltage: 3.7 V [26]
Theoretical off-load voltage: 3.5 V [103]
Nominal voltage: 3 V [26]

Under standard discharge conditions the voltage is 2.9 V [26], which is experimentally verifiable.

If based on a theoretical voltage of 3.5 V the following theoretical energy values may be calculated:

Energy density: 3100 mWh/cm^3
Specific energy: 1000 mWh/g

Practical values depend on battery size, internal construction, discharge load, and temperature.

Under low-rate conditions and at ambient temperature the following maximum values may be obtained from bobbin-type cylindrical batteries:

Energy density: 720 mWh/cm^3
Specific Energy: 260 mWh/g

6.2.1.1.4. Batteries and Their Applications

The applications of a manganese dioxide – lithium battery depend on its shape, size, and internal construction.

Button batteries are defined by diameter > height. Because of their flatness, lithium button batteries are addressed also as *coin cells*. Their ratio d/h is typically in the range: $6 < d/h < 12$ (see Fig. 29). For normal button batteries the ratio is typically $d/h < 6$.

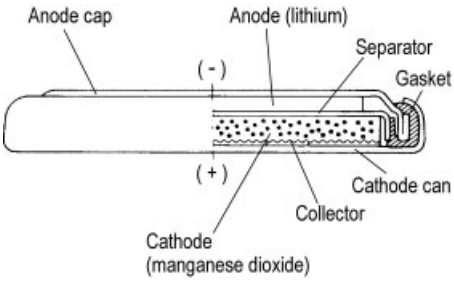

Figure 29. Cross-section of a manganese dioxide – lithium button battery

So far 21 manganese dioxide – lithium button batteries have been standardized [26], [33]. Their dimensions cover the following range

$$10.0\,\text{mm} \leq d \leq 30.0\,\text{mm}$$
$$1.2\,\text{mm} < h \leq 5.0\,\text{mm}$$

Their capacities range from 25 to 500 mAh. A typical discharge curve of a button battery is given in Figure 30.

From the voltage difference $U_b - U_t$ and the relevant change in discharge current the internal resistance R_i is calculated. Typical – as for most batteries – is the increase of the internal resistance with increasing degree of discharge. The capacity loss of these batteries during storage is of the order of 1 % per year and below. It is influenced also by the quality of the seal, the storage conditions and the purity of the materials used.

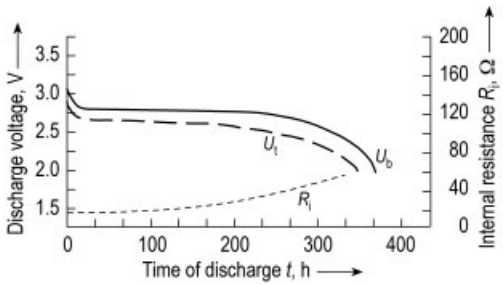

Figure 30. Discharge curve (U_b) of a CR2032 battery discharged through a 5.6 kΩ resistance with a superimposed pulse of 390 Ω for 2 s every 2 h, designated (U_t). R_i is the internal resistance of the battery.

Applications. Low-rate applications are typical for this type of battery (watches, small calculators, memory back-up, and cameras).

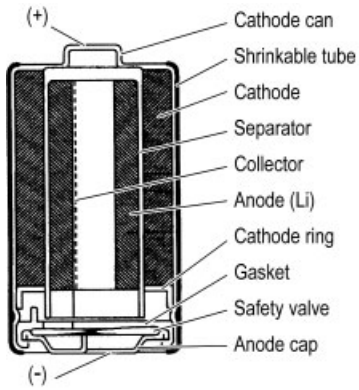

Figure 31. Bobbin-type construction of a cylindrical MnO_2 – Li battery [100]

Cylindrical Batteries. Of approx. 20 cylindrical MnO_2 – Li batteries available on the commercial market, so far six batteries have been standardized. Some of them are bobbin-type constructions for low-rate applications. Others employ spirally wound electrodes to comply with high rate requirements.

The *bobbin-type construction* is similar in electrode arrangement to the alkaline manganese dioxide – zinc battery (see Fig 21). A cross-section of a MnO_2 – Li bobbin-type cylindrical battery is presented in Figure 31.

Table 8 lists four cylindrical batteries with bobbin-type construction for low-rate application. A typical low-rate application is memory back-up in various kinds of electronic devices including computers.

Cylindrical *batteries with spirally wound electrodes* are designed for high-rate applications. Some of the more important batteries are listed in Table 9. The main application for these batteries is in the field of multifunction automatic cameras and heat meters (CR11108). A cross-section of a spirally wound cylindrical battery is presented in Figure 32.

Table 8. Cylindrical MnO_2 – Li batteries with bobbin-type electrodes for low-rate application

IEC Battery designation	Diameter, mm	Height, mm	Voltage, V	Capacity, mAh
CR12600	12.0	60.0	3	1500
CR14250	14.5	25	3	850
(CR17335)	17.0	33.5	3	1800
CR17450	17.0	45.0	3	2500

( ) = not standardized yet.

Table 9. Cylindrical MnO_2 – Li batteries with spirally wound electrodes for high rate application

IEC Battery designation	Diameter, mm	Height, mm	Voltage, V	Capacity, mAh
CR11108	11.6	10.8	3	180
2CR11108	13	25.2	6	170
(CR15H270)	17.0	27.0	3	750
(CR17335)	17.0	33.5	3	1300
CR17345	17.0	34.5	3	1300
CR17450	17.0	45.0	3	2000

( ) = not standardized yet.

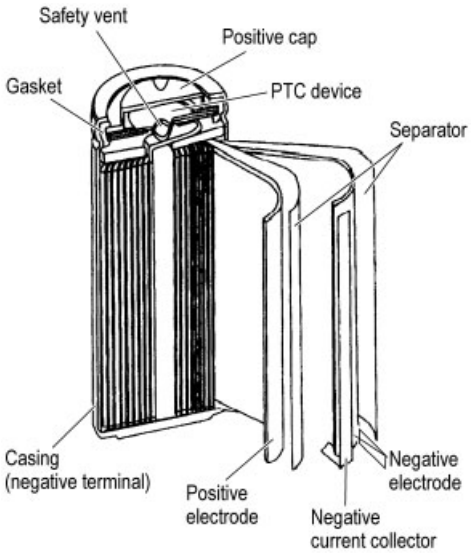

Figure 32. Cross-section of a high-rate lithium battery with spirally wound electrodes [125]

The high-rate capability of this battery is achieved by its large electrode area, its thin electrodes and its low-resistance separator. Figure 33 illustrates the discharge behavior of a cylindrical battery with spirally wound electrodes. If standardized it would carry the designation CR17335 (diameter 17.0 mm, height 33.5 mm), capacity 1300 mAh. Battery 17335 exists also as low-rate battery, with a nominal capacity of 1800 mAh (see Table 8). If standardized, the present IEC nomenclature system would designate both of them as CR17335 since the present designation system does not differentiate between a low-rate and a high-rate lithium battery. This ambiguity needs to be eliminated.

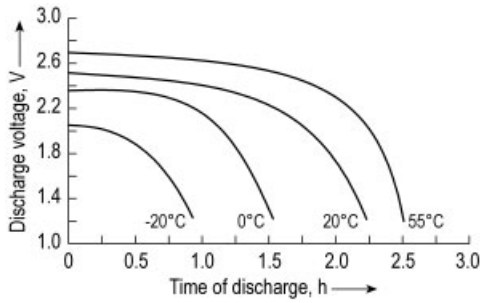

Figure 33. Discharge behavior of a high-rate battery CR17335 at different temperatures, discharge current 500 mA [125]

The construction according to Figure 32 offers short-period power outputs of up to 400 mW/cm^3(cf. Section 4.2.4). Constructions of this type are equipped with at least one safeguarding device (c.f. Section 5.2.2) to avoid battery rupture or explosion in case of an internal or external short circuit that would heat up the battery, thus causing an excessive internal pressure built up. The positive temperature coefficient (PTC) device reduces an external short circuit to an acceptable current level.

This kind of battery is relatively expensive because of its construction, safety features, ac-

tive materials, and production costs. It therefore will hardly be used, e.g., to power toys or torches. In this case batteries with aqueous electrolyte offer the better choice. Their specific energy cost is lower by a factor 6 to 8.

Prismatic Batteries. So far only two prismatic batteries of system C have been standardized by the IEC. They are not true prismatic batteries by construction, since both contain two cylindrical cells with spirally wound electrodes connected in series. Their nominal voltage is 6 V. Further details are given in Table 10.

Both batteries are designed to operate automatic cameras (film transport, flash light, shutter, light meter, etc). They could in principle be used also for medium- to low-rate applications, which, however, could be prohibitive by their price, because cheaper alternatives exist.

6.2.1.1.5. Prospects

The manganese dioxide – lithium system is currently the most successful primary lithium system on the commercial market. Presently only rough estimates exist on the worldwide production volume. The production volume of lithium button batteries was estimated in 1998 to be 10 % of that of alkaline MnO_2 batteries, i.e., 1.8×10^9 batteries, tendency increasing [120].

6.2.1.2. Carbon Poly-monofluoride–Lithium System

6.2.1.2.1. General

The carbon poly-monofluoride – lithium system $(CF)_n$ – Li was developed by Matsushita Battery Industrial Co. in the early 1970s. Since then batteries were developed in various shapes and sizes like button, cylindrical, prismatic, and pin-type batteries.

Their characteristics are similar to those of the MnO_2 – lithium system. The $(CF)_n$ – Li system is used primarily at low to medium discharge rates [104]. Advantageous is the large temperature range (− 40 °C to + 85 °C) within which it can be operated [105]. Carbon poly-monofluoride is an intercalation compound, formed by reaction between carbon powder and fluorine gas at elevated temperatures [106].

The IEC has allocated system letter B to the $(CF)_n$ – Li system [26]. Example: BR2325 designates a round battery (R) of the carbon poly-monofluoride – lithium system (B), having a diameter of 23.0 mm and a height of 2.5 mm. Since $d > h$ applies, this round battery is a button battery.

6.2.1.2.2. Electrochemistry

The discharge reaction is described by

Anode reaction

$$n\mathrm{Li} \longrightarrow n\mathrm{Li}^+ + n\mathrm{e}^- \tag{6.53}$$

Cathode reaction

$$(\mathrm{CF})_n + n\mathrm{e}^- \longrightarrow n\mathrm{C} + n\mathrm{F}^- \tag{6.54}$$

Overall reaction

$$n\mathrm{Li} + (\mathrm{CF})_n \longrightarrow n\mathrm{C} + n\mathrm{LiF} \tag{6.55}$$

Typical electrolytes are [104]: Lithium hexafluorarsenate ($LiAsF_6$) in dimethylsulfite (DMSI) Lithium tetrafluoroborate ($LiBF_4$) in a mixture of butyrolactone (BL) and tetrahydrofuran (THF) or of propylencarbonate (PC) and dimethoxyethane (DME).

6.2.1.2.3. System-Immanent Properties

According to IEC [26] the system's maximum off-load voltage is 3.7 V and its nominal voltage 3 V. Its theoretical specific energy is 2260 mWh/g [104]. The system's practical energy density and specific energy range up to

Energy density: 600 mWh/cm^3
Specific energy: 250 mWh/g

The system offers a low self-discharge of 0.5 % per year.

6.2.1.2.4. Batteries and Their Applications

Button Batteries. Of the 12 button battery sizes being produced, six have presently been standardized by the IEC. Their range is given by

$$12\,\mathrm{mm} \leq \text{diameter} \leq 30.0\,\mathrm{mm}$$

Table 10. 6 V high-rate prismatic batteries for automatic cameras

Designation accepted by IEC	Length, mm	Width, mm	Height, mm	Capacity, mAh
2CR5	34	17	45	1520
CR-P2	35	19.5	36	1300

$1.6\,\text{mm} \leq \text{height} \leq 3.2\,\text{mm}$
$48\,\text{mAh} \leq \text{capacity} \leq 190\,\text{mAh}$

The cross section of a $(CF)_n$ – Li system button battery is identical to that of Figure 29. The discharge behavior of battery type BR2032 (d = 20 mm, h = 3.2 mm; capacity 190 mAh) is given in Figure 34 A and compared to the discharge behavior of battery type CR2032 (210 mAh) under identical conditions (see Fig. 34 B). The comparison according to Figure 34 exhibits slight advantages for the MnO_2 – Li system over the $(CF)_n$ – Li system regarding both, voltage and capacity.

The commercial applications of the $(CF)_n$ – Li button batteries are identical to those of system MnO_2 – Li (watches, small calculators, memory back-up, etc.).

Cylindrical and Prismatic Batteries. The following cylindrical and prismatic batteries are standardized by the IEC:

1) Low-rate type BR17335:
   Nominal voltage 3 V; capacity 1.2 Ah; $d = 17.0$ mm; $h = 33.5$ mm
2) High-rate type BR17345:
   Nominal voltage 3 V; capacity 1.2 Ah; $d = 17.0$ mm; $h = 34.5$ mm

Their applications are identical to batteries CR17335 and BR17345 of the MnO_2 – Li system (e.g., memory back-up and automatic cameras).

An interesting comparison between a BR17335 battery and an R6P carbon – zinc battery (cf. Section 6.1.1.1.4) is given in Figure 35 for three different temperatures (20 °C, 0 °C and – 10 °C). The B-system battery uses spirally wound electrodes, the carbon – zinc battery is a bobbin-type construction. The discharge resistances are selected such that comparable discharge currents are drawn from the batteries. Although smaller by approx. 9 % in volume, the B-system battery offers at 20 °C approx. 3.5 times more energy than the carbon – zinc battery.

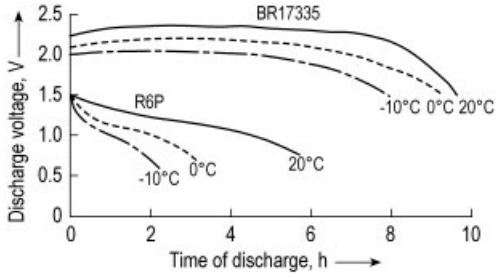

Figure 35. Comparison between a BR17335 battery and a R6P carbon – zinc battery

A specialty of the $(CF)_n$ – Li system is the so-called *pin-type battery* (BR425), used for fishing rods (illumination of the "fishing controller" during night time). Type BR425 is characterized by: Nominal voltage 3 V, capacity 25 mAh, diameter 4.2 mm, height 25.9 mm. Since $d < h$ applies, this battery is a cylindrical battery by definition, probably the smallest one in the world. A larger version of the BR425 is the BR435 (50 mAh).

Battery type BR-P2 is a 6 V high-rate prismatic battery for automatic cameras, see Table 11. It has the same application as the 6 V CR-P2 battery and is also constructed from two cylindrical high-rate batteries, connected in series. Larger prismatic batteries of the $(CF)_n$ – Li system with capacities of 25 Ah and above [104] are designed for applications in the noncommercial market.

6.2.1.2.5. Prospects

The carbon poly-monofluoride – lithium system is produced by one battery manufacturer [107]. Commercial markets being exclusive for this system are unknown with the exception of the pin-type battery. It appears that in regional and noncommercial markets as well as in special-purpose applications the $(CF)_n$ – Li system will

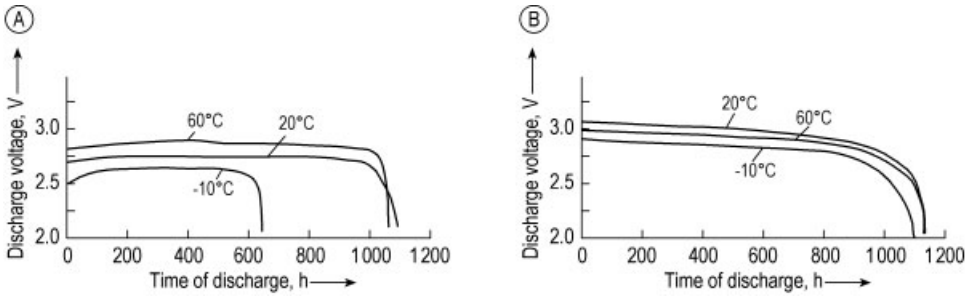

Figure 34. Constant resistance discharge (15 kΩ) of button batteries (diameter 20 mm, height 3.2 mm) at 60 °C, 20°C, and − 10 °C
A) BR2023 (190 mAh); B) CR2023 (210 mAh)

Table 11. BR and CR 6 V-high rate prismatic batteries for automatic cameras

Designation accepted by IEC	Length, mm	Width, mm	Height, mm	Capacity, mAh
BR-P2	33.9	19.5	36	1200
CR-P2	35	19.5	36	1300

continue to be an interesting alternative to other lithium systems.

6.2.1.3. Iron Disulfide – Lithium Systems

6.2.1.3.1. "1.5 V" Lithium Systems

The iron disulfide – lithium system (FeS_2 – Li) belongs to the group of "1.5 V" lithium systems like iron sulfide – lithium (FeS – Li), copper oxide – lithium (CuO – Li), bismuth plumbate – lithium ($Bi_2Pb_2O_5$ – Li, bismuth trioxide – lithium (Bi_2O_3 – Li) and others [104]. Much research and development work has been put into these systems in the 1970s. It was hoped that these systems could replace the expensive silver oxide – zinc or the environmentally harmful mercuric oxide – zinc button batteries to operate watches and/or hearing aids [108].

The above expectations could, however, not be realized. The main reason for this is the high internal resistance of these batteries, which is partly caused by the small diameter of the button battery: To be interchangeable with existing 1.5 V button batteries only standard diameters of 11.6 mm, 9.5 mm, 7.8 mm, or 5.8 mm are allowed. Under this condition existing 1.5 V button battery systems with alkaline electrolyte have a much lower resistance (factor 100 and above).

Another hope was that 1.5 V lithium systems could replace existing 1.5 V cylindrical batteries in low-rate applications. This was realized to a certain extent by the CuO – Li system [104]. Its excellent storage capability, high energy density, and wide temperature range has made it an interesting special-purpose battery, that, however, never will become a serious competitor for user-replaceable commercial 1.5 V cylindrical batteries.

More success on the commercial market is expected for the iron disulfide – lithium system, developed by the Eveready Battery Co., USA. The system is acknowledged by the IEC, and has received system letter F for identification. So far, however, only one battery size is available on the market. It is a R6-sized battery (diameter 14.5 mm, height 50.5 mm, capacity 2.6 Ah, nominal voltage 1.5 V) containing spirally wound electrodes.

6.2.1.3.2. Electrochemistry

The reaction (simplified) for the iron disulfide – lithium system is described by:

Anode reaction:

$$4\,Li \longrightarrow 4\,Li^+ + 4e^- \tag{6.56}$$

The cathode reaction proceeds in two steps

Step 1:

$$FeS_2 + 2\,Li^+ + 2e^- \longrightarrow FeS + Li_2S \tag{6.57a}$$

Step 2:

$$FeS + 2\,Li^+ + 2\,e^- \longrightarrow Fe + Li_2S \qquad (6.57b)$$

Overall reaction:

$$4\,Li + FeS_2 \longrightarrow Fe + 2\,Li_2S \qquad (6.58)$$

A mixture of propylene carbonate–dimethoxyethane with dissolved lithium perchlorate as conducting salt serves as electrolyte.

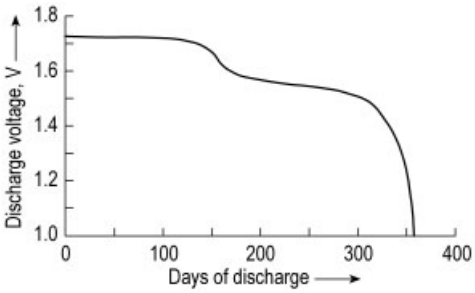

Figure 36. Low-rate discharge of a R6-sized FeS_2 – Li battery, (discharge resistance 5 kΩ) [125]

6.2.1.3.3. System-Immanent Properties

According to IEC the system's maximum off-load voltage is 1.83 V and its nominal voltage 1.5 V [26]. The energy density of an R6 sized battery with spirally wound electrodes is 450 mWh/cm^3 at a discharge rate of 8 mA.

At very low currents (0.4 mA and below) the discharge curve of the R6 sized battery exhibits two voltage plateaus according to Figure 36. The reason for this are the two discharge steps according to Reactions (6.57a) (first voltage plateau) and (6.57b) (second voltage plateau).

6.2.1.3.4. Batteries and Their Applications

At higher discharge rates (e.g., 60 mA and above), the discharge characteristics of the iron disulfide – lithium system change (see Figs. 37 A and B). Figures 37 A and B include a comparison with the LR6 battery (alkaline version of the MnO_2 – zinc system). Figure 37 A employs a 60 mA discharge current, Figure 37 B refers to a 800 mA discharge.

When comparing the two systems, it must be taken into account that the FeS_2 – Li battery uses spirally wound electrodes, i.e., a construction similar to Figure 32, which includes a safety vent and a PTC device. The LR6 battery, however, employs only a bobbin type construction (see Fig. 21). Because of this difference, the FeS_2 – Li battery offers a superior load capability, but at a significantly higher price.

Figure 37 A and B demonstrate that the FeS_2 – Li battery is especially attractive for high discharge currents: At 800 mA (21 °C, end-point voltage 0.9 V) the FeS_2 – Li battery delivers approx. 5 times more energy than the LR6. At 60 mA this energy factor drops to about 1.5 times.

The preferred application for this battery is therefore in the field of devices needing a high-rate power supply on a discontinuous basis like automatic cameras (film transport, flash light, etc.). Devices requiring a high-rate power on a continuous basis could technically also be operated. In this case, however, the use of a rechargeable battery system would certainly be more economic.

6.2.1.3.5. Prospects

The prospects of this battery are difficult to judge. They will depend on the number of applications that need this type of battery, and which cannot be operated satisfactorily by a cheaper 1.5 V battery system.

6.2.2. Batteries with Soluble Cathode

Systems with soluble cathode include the sulfur dioxide – lithium system (SO_2 – Li) with an organic electrolyte, the thionyl chloride – lithium system ($SOCl_2$ – Li) with an inorganic electrolyte, and the sulfuryl chloride – lithium system (SO_2Cl_2 – Li) with an inorganic electrolyte [104], [109].

6.2.2.1. Sulfur Dioxide – Lithium System

6.2.2.1.1. General

The sulfur dioxide – lithium system is a high-rate, high energy system. It was developed in

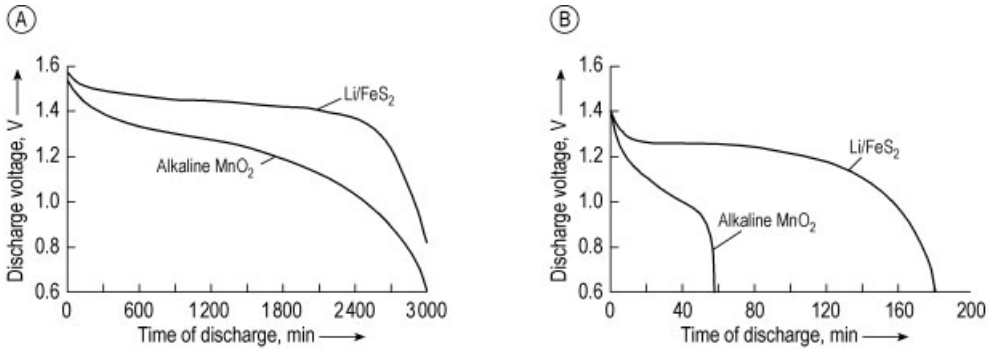

Figure 37. A) 60 mA-discharge performance of system FeS_2 – Li in comparison with the alkaline MnO_2 – zinc system during discharge of R6-sized batteries; B) 800 mA-discharge performance of system FeS_2 – Li in comparison with the alkaline MnO_2 – zinc system during discharge of R6-sized batteries [125]

the early 1970s by US companies essentially for space and military applications. It still is a special purpose battery that probably will never enter the commercial market. It also has not been proposed yet for International Standardization.

6.2.2.1.2. Electrochemistry

The discharge reaction (simplified) may be described by

Anode reaction

$$2\,Li \longrightarrow 2\,Li^{+} + 2\,e^{-} \tag{6.59}$$

Cathode reaction

$$2\,SO_2 + 2\,e^{-} \longrightarrow S_2O_4^{2-} \tag{6.60}$$

Overall reaction

$$2\,Li + 2\,SO_2 \longrightarrow Li_2S_2O_4 \tag{6.61}$$

A mixture of acetonitrile, lithium bromide, and sulfur dioxide serves as electrolyte.

A porous Teflon bonded carbon black layer serves as catalytic surface to promote Reaction (6.60) and in combination with a metal grid it also serves as current collector.

6.2.2.1.3. System-Immanent Properties

The open circuit potential of the sulfur dioxide – lithium system is approx. 3 V. Typical discharge voltages at a 10 to 20 h discharge rate are in the range 2.6 to 2.8 V.

Specific values for its energy content are

Energy density: 440 mWh/cm^3
Specific energy: 280 mWh/g

At ambient temperature the conductance of its electrolyte is $5 \times 10^{-2}\,\Omega^{-1}\,cm^{-1}$. It decreases only by 60% when the temperature is changed from + 50 °C to − 50 °C.

Reactions (6.59) and (6.60) have low reaction resistances such that the SO_2 – Li system exhibits an excellent low-temperature performance.

The good storage behavior results from the protective lithium dithionite layer on the surface of the lithium electrode formed by an initial reaction. Formation of this layer is also the reason for the initial voltage delay on discharge (see Figs. 28 and 39), which becomes more pronounced with increasing discharge current and decreasing temperature. The voltage delay can be eliminated by preconditioning with a short discharge at a higher rate [110].

Under normal conditions the SO_2 cathode material is gaseous. In order to have it in a liquid state it must be kept under sufficient pressure. At ambient temperature this pressure is 3 to 4 bar for an undischarged battery, at 100 °C it is about 20 bar. The system therefore needs an adequate battery design.

Because of its hazardous contents this battery must be used and handled extremely carefully [111]. Reference is also made to restrictions in transportation, shipment, and disposal of SO_2 – Li batteries and lithium batteries in general [1].

6.2.2.1.4. Batteries and Their Applications

The SO_2 – Li system is manufactured only in a cylindrical shape and assembled to multicell battery packs (including safety features like fuses etc.). It is predominantly used in military and noncommercial devices. A typical cylindrical battery construction is shown in Figure 38.

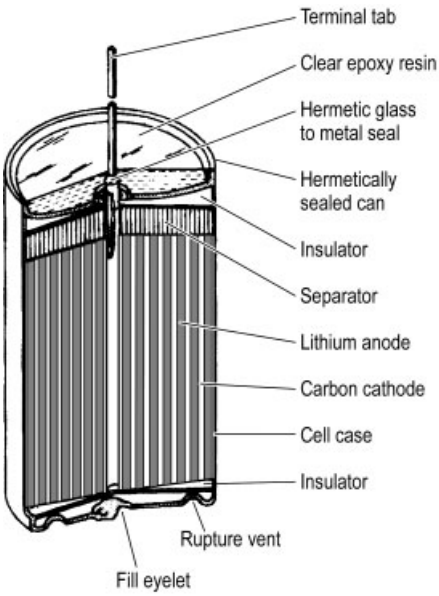

Figure 38. Cut-away view of a cylindrical SO_2 – Li battery [125]

Typical for this construction is the use of a hermetic glass-to-metal seal. It carries in its center the positive terminal, which is connected to the SO_2-electrode. The cylindrical cell case is connected to the negative pole, i.e., the lithium electrode. Integrated into the battery can is a rupture vent for safety reasons (opening pressure 20 to 35 bar, depending on product).

The electrode-pack consists of four rectangular strips, i.e.:

1) Lithium metal layer
2) Separator layer
3) Porous Teflon-carbon layer
4) Separator layer

These four layers are spirally wound to a cylindrical electrode-pack and inserted into the cell can. All layers are porous with the exception of the lithium electrode. Before cell closure they are soaked with electrolyte plus SO_2.

The discharge behavior of an 8 Ah SO_2 – Li battery at 21 °C is plotted in Figure 39 for different discharge currents. Typical is its flat discharge plateau and its initial voltage delay.

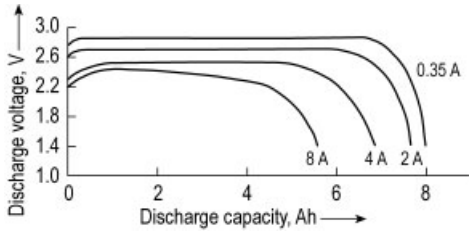

Figure 39. Discharge voltage vs. discharged capacity of an 8 Ah SO_2 – Li cylindrical battery at 21 °C for different discharge currents [110]

The discharge performance of the 8 Ah SO_2 – Li battery at different temperatures is illustrated in Figure 40. Selected temperatures are: Room temperature (RT), – 30 °C, – 40 °C and – 50 °C. In Figure 40 A the capacity is plotted as function of discharge current. In Figure 40 B the discharge voltage is plotted as function of discharge current. Figures 40 A and B demonstrate the outstanding low temperature behavior of the SO_2 – Li system.

6.2.2.1.5. Prospects

Since this system is predominantly used in military and professional applications, its prospects are difficult to judge. In any case, it will most likely not become a commercial consumer replaceable product but will continue to be the specialist for low-temperature/high-rate applications.

6.2.2.2. Thionyl Chloride – Lithium System

6.2.2.2.1. General

The 3.6 V thionyl chloride – lithium system ($SOCl_2$ – Li) is a high energy system. It is manufactured in two battery versions, i.e., low-rate and high-rate. Applications in the military and professional market are prevailing. The $SOCl_2$ – Li system has started to become also a consumer replaceable product. The IEC has meanwhile allocated system letter E to this system.

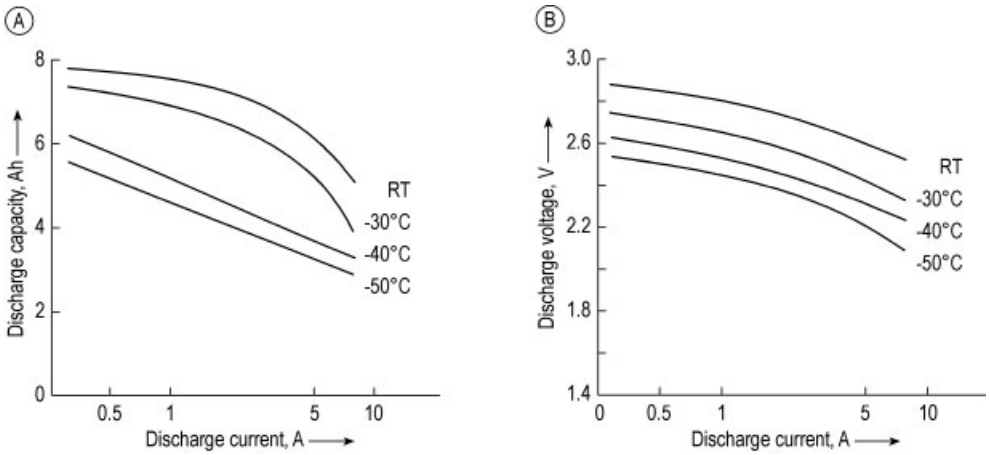

Figure 40. Discharge performance of an 8 Ah SO_2 – Li at different temperatures [110]; A) Discharge capacity versus current; B) Discharge voltage versus current.

Many types of this battery are available on the market. The preferred shapes are cylindrical and prismatic.

Because of their hazardous contents these batteries are not considered to be environmentally benign. Restrictions in handling, transportation, shipment and disposal have to be observed also for this system [1].

6.2.2.2.2. Electrochemistry

The reactions of the SO_2 – Li system (simplified) are described by:

Anode reaction

$$4\,Li \longrightarrow 4\,Li^{+} + 4\,e^{-} \tag{6.62}$$

Cathode reaction

$$2\,SOCl_2 + 4e^{-} \longrightarrow SO_2 + S + 4\,Cl^{-} \tag{6.63}$$

Overall reaction

$$4\,Li + 2\,SOCl_2 \longrightarrow S + SO_2 + 4\,LiCl \tag{6.64}$$

The electrolyte is composed of thionyl chloride, serving not only as cathode material but also as solvent for the conducting salt, which is usually lithium tetrachloroaluminate ($LiAlCl_4$).

Cathode reaction (6.63) is promoted by an active surface of a highly porous, electron conducting Teflon-bonded carbon-black structure, which represents the positive electrode. In addition, the carbon-black structure serves as current collector, and is connected to the positive terminal of the battery. It further functions as a repository for insoluble products of the discharge reaction like lithium chloride. At the end of discharge LiCl-precipitates may cover up the active carbon surface, finally terminating the discharge of a cathode-limited battery [112]. Since SO_2 dissolves very well in the electrolyte, its formation does not present any problems. It only leads to a slight increase of the battery's internal gas pressure. The formation of sulfur, however, may lead to a thermal runaway if substantial amounts should get into contact with the lithium electrode [104].

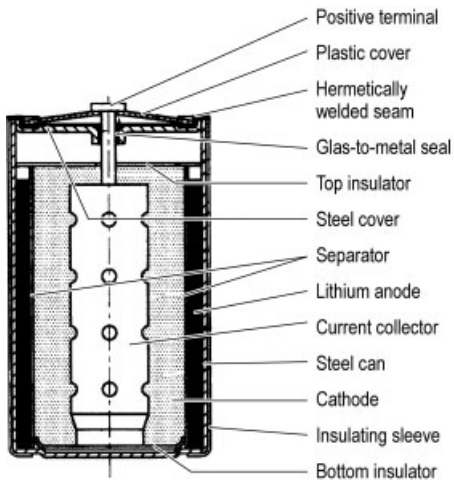

Figure 41. Cross-section of a bobbin-type $SOCl_2$ – Li cylindrical battery [110]

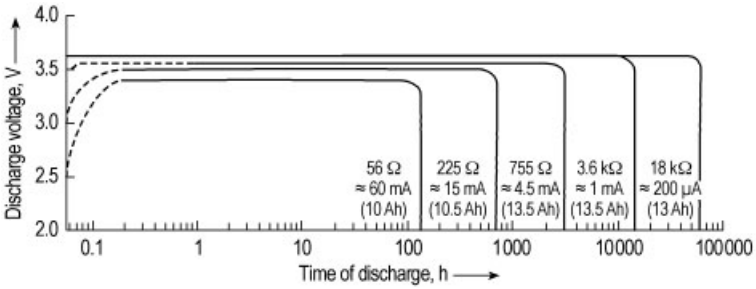

Figure 42. Discharge behavior of a R20-sized bobbin-type $SOCl_2$ – Li battery for different discharge resistances [110]

In contact with the thionyl chloride the lithium anode forms a protective LiCl film (SEI), which contributes to the system's excellent shelf life. Due to its low conductivity this LiCl film can cause a voltage delay at the beginning of a discharge (see Fig. 42). This voltage delay becomes more pronounced in the case of a preceding storage at elevated temperatures. Decreasing temperatures and increasing discharge currents increase the voltage delay. When continuing the discharge, the voltage recovers (cf. Fig. 28), due to a reduction of the thickness of the LiCl layer and thus in resistance.

6.2.2.2.3. System-Immanent Properties

According to the IEC [26] the thionyl chloride – lithium system has a maximum off-load voltage of 3.9 V and a nominal voltage of 3.6 V. Its standard discharge voltage is 3.4 V [26].

With large batteries, where the active material determines the battery volume, the following energy values may be obtained at ambient temperature and low-rate discharge:

Energy density: 900 mWh/cm^3
Specific energy: 460 mWh/g

The $SOCl_2$ – Li system has the highest specific energy of all primary battery systems and surpasses even that of the alkaline air – zinc system, which is about 360 mWh/g.

Its energy density is the highest of all lithium batteries.

Another advantage of this lithium system is that even a bobbin type construction permits moderate discharge rates up to a 50 h-discharge (example: a 5 Ah battery would carry a discharge current of up to 100 mA). Furthermore it has a wide temperature range within which it can be operated. The temperature range is determined by the electrolyte's freezing (– 110 °C) and boiling point (79 °C).

6.2.2.2.4. Batteries and Their Applications

All thionyl chloride – lithium batteries are characterized by a hermetic glass-to-metal seal of their positive terminal. *Button batteries* in 3 V standard dimensions [26], [33] do not exist.

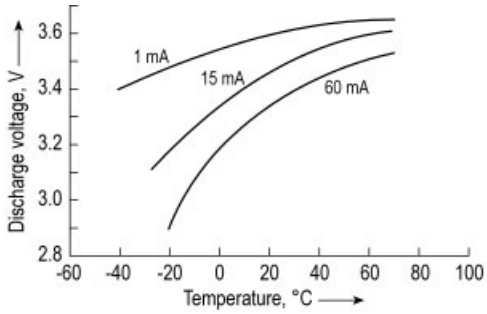

Figure 43. Discharge voltage of a R20-sized (13 Ah) bobbin type $SOCl_2$ – Li battery vs. temperature for three different currents [110]

Cylindrical batteries are manufactured in two versions: with bobbin-type electrodes for low- and medium-rate applications, and with spirally wound electrodes for high-rate applications.

The *bobbin-type construction* is illustrated in Figure 41. The cathode is a highly porous body, made of Teflon-bonded carbon black. The cathode is in electronic contact with the current collector, which in turn is connected to the battery's

positive terminal. The positive terminal is part of the glass-to-metal seal. The separator between anode and cathode is made of nonwoven glass. The lithium foil is contacted to the inner surface of a nickel-plated steel or stainless steel can, and carries the negative terminal. Before battery closure the porous cathode and the separator are soaked with electrolyte, which includes the thionyl chloride.

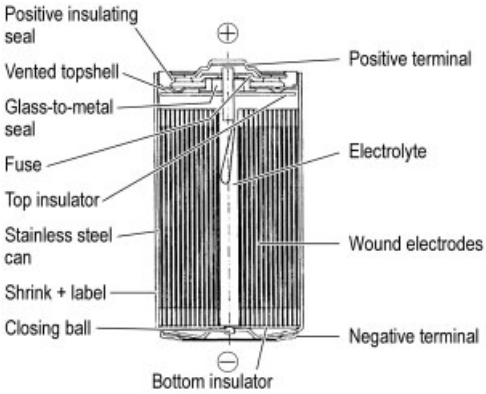

Figure 44. Construction of a high rate $SOCl_2$ – Li battery, using spirally wound electrodes [125]

The discharge behavior of an R20-sized (13 Ah) cylindrical battery is presented in Figure 42. Characteristic is the initial voltage delay and the succeeding flat discharge plateau. The discharge voltage as function of temperature and three different discharge currents is shown in Figure 43.

The construction of an $SOCl_2$ – Li battery, using *spirally wound* electrodes is shown in Figure 44. This design includes several safeguarding devices like a vented top-shell, a fuse and a closing ball. These devices are designed to prevent an explosion of the battery even under extreme conditions including cell reversal. The discharge behavior of a high-rate R20-sized $SOCl_2$ – Li cell is plotted in Figure 45.

The capacity of a R20-sized $SOCl_2$ – Li high-rate battery as function of discharge current for three different temperatures is illustrated in Figure 46. Even at − 40 °C it still provides 4 Ah at a 200 mA discharge.

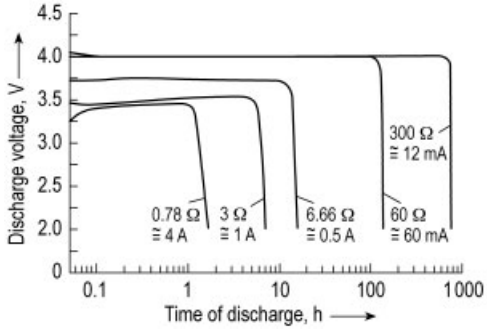

Figure 45. Discharge behavior of a high-rate R20 sized $SOCl_2$ – Li battery [110]

These high-rate cylindrical batteries are developed mainly for military applications and are assembled to multicell batteries. Military applications prevail also for the large high-capacity $SOCl_2$ – Li batteries (10 000 Ah and above), e.g., as standby power source [104] and for undersea applications [113].

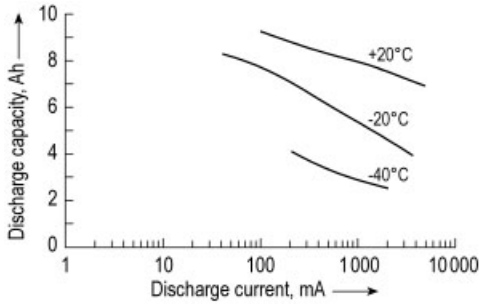

Figure 46. Discharge capacity of an R20-sized $SOCl_2$ – Li high-rate battery for three different temperatures, endpoint-voltage 2.0 V [110]

6.2.2.2.5. Prospects

For many years $SOCl_2$ – Li batteries have been produced by manufacturers in the USA, Japan, and Europe. It is safe to predict that this system will continue to remain in the market. Cylindrical bobbin-type constructions with capacities in the range of 0.3 Ah to 3 Ah are preferred. Prevailing application is memory back-up in all kinds of electronic devices, including computers. In this application the $SOCl_2$ – Li battery has started to become a consumer replaceable product. The first E-system battery being included in the IEC Standard is a 6 V consumer replaceable memory back-up prismatic battery for personal

computers (battery designation 2EP3863). With the increasing electronic market the production volume of this kind of battery system will most likely increase as well. The prospects of batteries with spirally wound electrodes and/or extremely large capacities are difficult to judge.

6.2.3. Batteries with a Nonliquid Electrolyte

The attractive aspect of a nonliquid electrolyte in combination with a solid cathode and a solid anode is the option to construct batteries with a high energy content, however, without any gassing and leakage problems and relevant safety issues. This kind of battery system should furthermore offer the ability to be operable over a wide range of environmental conditions (temperature, pressure, and acceleration). These options have been and still are the impetus for extensive research and development work.

6.2.3.1. Polymer and Solid Electrolyte Systems

The term *polymer electrolyte* can be applied to a broad family of ion-conducting materials [114]. The ionic conductivities of polymer electrolytes at room temperature are typically by a factor 100 to 1000 lower than those of liquid electrolytes. This high specific resistance, however, can largely be compensated by thin-electrolyte-film techniques. Another unique aspect of the polymer electrolyte is that it represents a flexible film of a polymer matrix and an ionic salt complexed into this matrix.

The focus of polymer research and development has largely been on polyether-based solvent-free systems for lithium *rechargeable* batteries [114]. For *primary* battery systems preference has been given to solid electrolytes.

A large variety of *solid electrolyte systems* were studied in the past [115], while new ones are subject of research and development [116]. Solid electrolytes for battery application generally need to have a high ionic conductivity. High ionic conductivities, however, are only provided by solid electrolytes with monovalent ions as charge carrier, as for example, lithium, sodium, silver, potassium, or rubidium. Multivalent ions offer the advantage of a larger charge transfer, however, they exhibit a much lower mobility and thus a lower conductivity because of their strong electrostatic interactions with the crystal lattice. Optimal conductivity is achieved in a lattice with an ion that is neither too big nor too small to fit the available pathways in the lattice [117].

At room temperature the conductivity σ of solid electrolytes covers a range of

$$10^{-0.5}\,\Omega^{-1}\,\mathrm{cm}^{-1} \leq \sigma \leq 10^{-7}\,\Omega^{-1}\,\mathrm{cm}^{-1}$$

The temperature dependence of the conductivity σ of a given ion conductor follows mostly the Arrhenius equation

$$\log\sigma = H - A/T \tag{6.65}$$

where T is the absolute temperature, H and A being specific constants for a given solid electrolyte. When changing the temperature from 25 °C to, e.g., 200 °C the conductivity σ may increase by several orders of magnitude [116], depending on the type of ion conductor. A sufficiently high ionic conductivity of a solid electrolyte is a requirement for a practical battery. It is, however, not the only aspect that must be observed. Of practical interest are electrolyte systems offering in addition the option to compose a battery system of high voltage and a high energy density. This requirement is unfortunately not fulfilled for silver- and copper-ion conductors such as Ag_4RbI_5 and $Cu_{16}Rb_4I_7Cl_{13}$, with extraordinarily high ionic conductivities of the of order $10^{-0.5}\,\Omega^{-1}\,\mathrm{cm}^{-1}$.

The prospects of lithium, sodium, or potassium conductors are much better in terms of energy densities, as shown by the theoretical values given in Table 12. These values are based on a metal – I_2 system [115].

From Table 12 it is evident that Li-, Na-, or K-based ionic conductors are the preferable candidates for practical solid electrolyte batteries. Consequently research and development has been and still is focussed on relevant ion conductors. In the attempt to realize a fast sodium conductor GOODENOUGH was successful in developing a structure with the formula $Na_3Zr_2PSi_2O_{12}$, which is named Nasicon [118]. At 25 °C its conductivity is approx. $10^{-2.5}\,\Omega^{-1}\,\mathrm{cm}^{-1}$. With Nasicon GOODENOUGH contributed to the understanding that high ionic mobility requires the existence of interconnected spaces of partially occupied lattice sites.

Table 12. Theoretical values for open circuit potential U^0 and energy density for metal – I_2 systems based on some monovalent anode metals

Anode metal	U^0, V	Energy density, Wh/cm^3
Li	2.8	1.9
Na	3.0	1.6
K	3.4	1.3
Cu	0.7	0.59
Ag	0.7	0.51

Present lithium ion conductors have *total conductivities* in the range $10^{-3}\,\Omega^{-1}\,cm^{-1}$ to $10^{-7}\,\Omega^{-1}\,cm^{-1}$ [116]. Their *bulk conductivities* may be higher by two orders of magnitude. Of concern is the fact that only a few compounds of presently known lithium-ion conductors show thermodynamic stability in contact with elemental lithium, i.e., stability of the interface electrode – electrolyte. Because only thin-film solid electrolytes can be employed, a high specific electronic resistance is required to avoid self-discharge. For a practical system the specific electronic resistance must be in the range 10^{12} to $10^{14}\,\Omega\,cm$, which is an additional requirement a solid lithium-ion conductor must comply with.

For above reasons fast lithium-ion conductors may not always be the best choice, unless they also provide thermodynamic stability and a high specific electronic resistance.

6.2.3.2. The LiI(Al_2O_3) Solid Electrolyte System

A solid electrolyte system that constitutes a good compromise on the above requirements is the LiI(Al_2O_3) electrolyte. In combination with certain cathode materials such as $PbI_2 + S$, $PbI_2 + PbS$, and $TiS_2 + S$ different solid-state battery systems were developed in the 1980s by DURACELL and made available for commercial use. Later on, however, the production of the LiI(Al_2O_3) electrolyte system was discontinued.

The battery was designed in a circular disk configuration as shown in Figure 47, having a diameter of about 29 mm and a height of 2.5 mm (version 1), or 5.8 mm, respectively (version 2), which was a 2-cell-in-series arrangement [119]. The thickness of the electrolyte (d) is about 0.2 mm. During discharge the partial molar volume of the active materials decreases (valid for all lithium systems). To avoid any loss of contact between anode, electrolyte and cathode during discharge they must be kept under pressure. This is done by a wave spring (k).

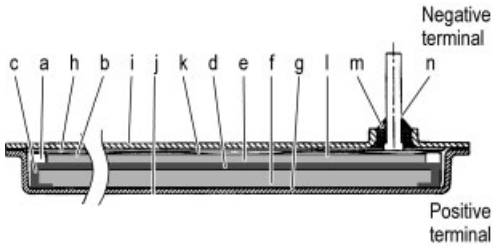

Figure 47. Cross section of a Li – LiI(Al_2O_3) – PbI_2,PbS solid state battery [125]
a) Anode insulating ring; b) Anode collector; c) Cathode ring; d) Electrolyte; e) Anode; f) Cathode; g) Cathode collector; h) Insulating disk; i) Top feed-through assembly; j) Can; k) Wave spring; l) Tab; m) Loctite; n) Glass-to-metal seal

Properties of the Single-Cell Battery.

U^0: 1.9 V
Capacity: 350 mAh
Energy density: 445 mWh/cm^3
Specific energy: 920 mWh/g

There is no loss of capacity after storage for one year, even at a storage temperature of 60 °C. The battery can withstand temperatures up to 200 °C [119].

6.2.3.3. The Lithium – Iodine System

6.2.3.3.1. General

The lithium – iodine system is only commercially available solid electrolyte battery system. Its development dates back to the early 1970s. Batteries of this kind are mainly produced in the USA (Catalyst Research Corporation, Wilson Greatbatch, Medtronic). Because of its high reliability in providing low power for a long time, the system has become a power source for heart pacemakers.

6.2.3.3.2. Electrochemistry

The basic reaction of the Li – I_2 system is described by Equation (6.66)

$$Li + 1/2I_2 \longrightarrow LiI \quad (6.66)$$

The reaction product is lithium iodide. It is formed as soon as the lithium surface is contacted by iodine. Lithium iodide constitutes the system's SEI.

In order to discharge the nonconducting iodine it must be mixed with an electronic conductor. In present commercial batteries poly-2-vinylpyridine (P2VP) is used as electron conducting medium. Thus the system may be designated by $Li - LiI - I_2(P2VP)$.

Its electrochemical reactions are:

Anode reaction

$$2\,Li \longrightarrow 2\,Li^+ + 2\,e^- \quad (6.67)$$

Cathode reaction

$$2\,Li^+ + 2\,e^- + P2VPnI_2 \longrightarrow P2VP(n-1)\,I_2 + 2LiI \quad (6.68)$$

Overall reaction

$$2\,Li + P2VPnI_2 \longrightarrow P2VP(n-1)\,I_2 + 2\,LiI \quad (6.69)$$

Its reaction product LiI is identical to that of Reaction (6.66).

6.2.3.3.3. System-Immanent Properties

A unique feature of the system is that its reaction product LiI serves simultaneously as separator and as lithium-ion conductor. Lithium iodide is a solid electrolyte, having a conductivity at the order of $10^{-7}\,\Omega^{-1}\,cm^{-1}$. On discharge the LiI-layer grows in thickness proportional to the capacity discharged from the battery, i.e., its internal resistance increases constantly during discharge.

The system's open circuit voltage is approx. 2.8 V. Depending on battery type and capacity (1 to 2.5 Ah) its practical energy density ranges from 0.7 to 0.99 Wh/cm^3, and may be obtained at a discharge current density of $1-2\,\mu A/cm^2$.

As typical for all lithium systems the partial molar volume of the $Li - LiI - I_2(P2VP)$ system decreases during discharge (in theory by 15 % for a complete discharge). Being a kind of viscous liquid, $I_2(P2VP)$ can adapt itself better to the system's volume changes than a solid.

6.2.3.3.4. Batteries and Their Applications

Pacemaker batteries are available from several companies. Figure 48 illustrates the construction of a pacemaker battery made by Wilson Greatbatch. Depending on type, its volume ranges from 3 to 6 cm^3. The battery thickness ranges from 5 mm to 10 mm.

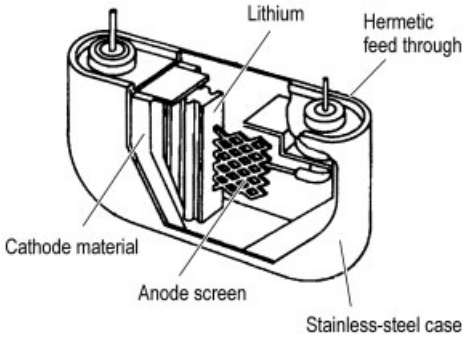

Figure 48. Cutaway view of a pacemaker battery [110]

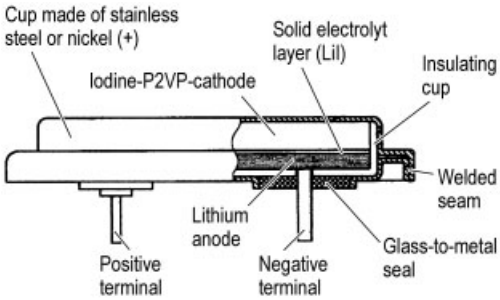

Figure 49. Cutaway view of a lithium – iodine button battery [110]

Typical discharge currents in application are of the order of 20 μA. Depending on battery capacity, the discharge time may be as long as 10 years. For discharge currents of 10 μA and below, the voltage under load drops about linearly with time down to approx. 2.5 V, mainly because the LiI layer increases in thickness. At the end of discharge nonohmic resistances come into play until the voltage drops abruptly. At high discharge rates nonohmic resistances influence the discharge curve from the beginning.

When looking for new applications the possibility was studied to operate wrist watches continuously for a ten years period by a suitable lithium – iodine battery. Figure 49 illustrates such a battery construction.

The discharge characteristic of a 650 mAh button battery is shown in Figure 50. Although

technically feasible, a commercial market for wrist watches with lithium-iodine batteries could not be realized.

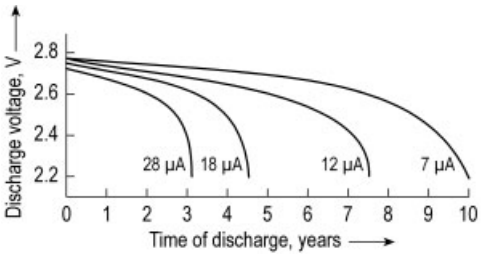

Figure 50. Discharge voltage vs. time for a 650 mAh lithium iodide button battery for different loads [110]

6.2.3.3.5. Prospects

The commercial market for present solid-state battery systems is limited because of their high internal resistance. If reduced by three to four orders of magnitude, new applications may evolve. The only commercial market for the lithium – iodine system is presently the heart pacemaker market, which also accepts the high price for the battery.

6.3. Reserve Batteries

6.3.1. Principles

A reserve battery is a primary battery system that only after activation provides electrical energy to operate a specific electric device for a predetermined period of time, which may range from seconds to weeks. The time at which the electrical power is needed is usually unknown (like a case of emergency). The waiting period until the reserve system is activated, may be ten years and more.

The battery employed is usually a high-rate system. Its self-discharge would not be negligible if kept active over a long period of time. This is especially true when the system is exposed to elevated temperatures during standby. But also at normal to low temperatures the overall reliability of the system may decline if kept active over several years. Therefore for reserve batteries preference is given to systems, that are kept inactive during standby.

Each electrochemical system consists of three key components, i.e., cathode, electrolyte, and anode. The reserve battery is characterized by the fact that one of these key components is separated from the other two. A preferred key component is the electrolyte (cf. Section 6.3.1.1).

Shortly before the electrical energy is needed, the activation is carried out by removing the separation between the electrolyte and the electrodes. The activation of the system occurs by a physical process, in consequence of which the electrolyte enters the space between anode and cathode while intimately contacting the total surface of both electrodes. In the simplest case the addition of the electrolyte is done by hand. Depending on the specific conditions, however, the activation mechanism may be more complex (cf. Section 6.3.1.1.2).

Another kind of electrolyte activation employs a solid-state electrolyte that only becomes an ion conductor if heated up to its melting point. During standby the solid state electrolyte is a nonconductor. This group of systems is known as thermal batteries (cf. Section 6.3.1.2).

In addition, a battery can be kept inactive during standby if its electrochemically active materials are stored separately. These materials should preferably be in the liquid or gaseous state to facilitate distribution and transport to the battery's electrodes in the case of activation. Reserve batteries that store gaseous or liquid cathode and/or anode materials during standby in separate containers are in principle similar in function to fuel cells [121]: The system starts to work as soon as the active material(s) are released to the relevant electrodes. In this sense the air – zinc system could also be counted to the group of reserve systems, since it complies with the afore mentioned criteria: It starts working as soon as its air access is opened (cf. Section 6.1.3.4).

6.3.1.1. System Activation with Liquid Electrolyte

6.3.1.1.1. Water-Activated Batteries

An example of this group of systems are the *water-activated metal chloride – magnesium batteries* (cf. Section 6.1.2). Their activation is

carried out by manually opening their water access holes. Properties and typical applications for this type of battery are described in Section 6.1.2.

6.3.1.1.2. KOH Activated Batteries

Because of its high energy density and high rate capability the silver oxide – zinc system is well suited for reserve battery systems (cf. Section 6.1.3.2). It may be based on monovalent or divalent silver oxide or mixtures thereof. In the simplest case the activation of this system is carried out manually by filling the electrolyte (KOH) into the battery. Alternatively, dry KOH pellets may be stored inside the battery. In this case only water must be added, having the beneficial effect that the heat, generated in consequence of KOH-dissolution, improves the battery performance.

Manually-Activated Batteries. A broad range of 1.5 V silver oxide – zinc battery modules for manual activation exists, predominantly used in military applications. The batteries cover a capacity range from 1 to 775 Ah at a weight from 25 g to 4.4 kg.

Gas Activation. A more sophisticated activation system for a multicell battery arrangement is shown in Figure 51. As can be realized from the drawing, the activation is carried out by activating the gas generator. The generated gas forces the piston to move, while pressing the electrolyte into the multicell battery arrangement, which activates the battery.

Systems of this kind are used for automatically activated silver oxide – zinc reserve batteries, for military (e.g., missiles) and space applications. Their characteristics cover following ranges:

Voltage: 26 – 50 V
Capacity: 0.8 – 20 Ah
Current: 2 – 80 A
Power: 260 W – 2.24 kW
Energy: 20 – 600 Wh

Torpedo batteries based on the silver oxide – zinc system have an energy content of 20 to 25 kWh at a power output of approx. 20 kW.

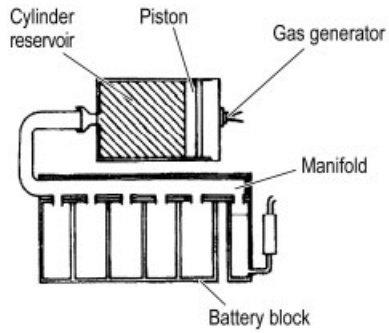

Figure 51. Schematic drawing of an activation system with piston activator [12]

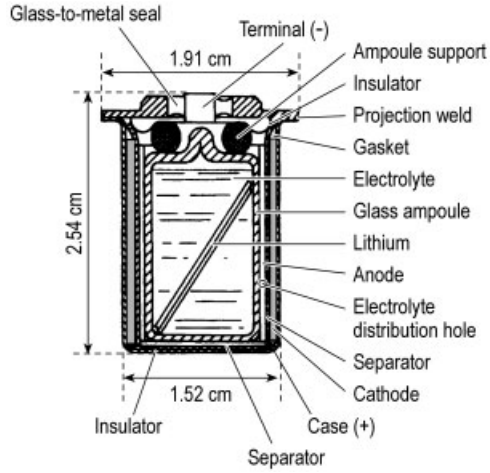

Figure 52. Vanadium pentoxide – lithium reserve battery ([125], Fig. 21.3)

6.3.1.1.3. Ampoule-Type Batteries

The unique feature of this type of battery construction is a glass ampoule, serving as electrolyte reservoir, which is arranged within the battery. The electrolyte is released (activation) as soon as the ampoule brakes due to mechanical force acting from outside.

This construction is preferred for small single-cell reserve batteries. The principle, however, is also used in multicell battery arrangements. The electrolyte used depends on the electrochemical system employed.

Vanadium Pentoxide – Lithium System (Fig. 52). In the vanadium pentoxide – lithium system (V_2O_5 – Li) the electrolyte is separated from the electrodes by a glass ampoule. Activation is carried out by an external force

that hits the battery's bottom, thus braking the glass ampoule inside the can. Under this condition the electrolyte is released and soaked up by the electrode – separator arrangement, and the battery is thus activated (voltage range during discharge: 3.4 – 2.3 V; capacity range: 100 – 500 mAh, depending on battery). Depending on manufacturer, the design and size of the battery varies.

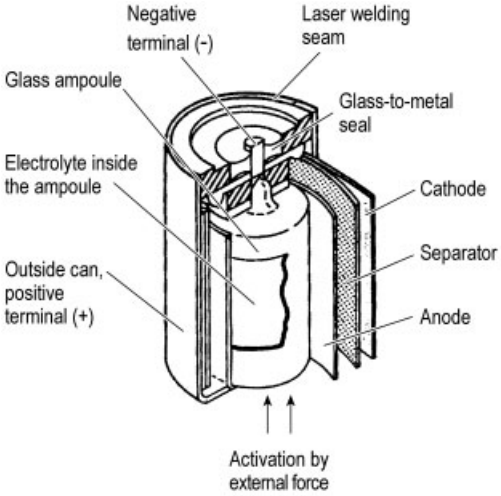

Figure 53. Cutaway view of a thionyl chloride – lithium single-cell reserve battery [110]

Thionyl Chloride – Lithium and Sulfur Dioxide – Lithium Systems. Figure 53 presents a cutaway view of a thionyl chloride – lithium single-cell reserve battery. Its construction is practically identical to that of Figure 52. Identical is also the battery's mode of activation being carried out by braking the ampoule by an external force.

However, there is a distinct difference between both, because thionyl chloride is not only the solvent of the electrolyte, which includes the conducting salt, but is simultaneously the cathode material. Thus, both electrolyte and cathode material are encapsulated in the ampoule. Similar arrangements exist for SO_2 – Li single-cell reserve batteries, which also encapsulate SO_2, the cathode material and the system's electrolyte, in an ampoule before activation.

Spin-Activated Battery. A variant of the above mechanism is used in the spin-dependent reserve battery. In this case the external force is represented by high spin forces as provided by artillery projectiles. The electrode stack of this battery consists of electrodes of an annular configuration with an electrolyte-filled ampoule in its center. Upon firing of the gun, the ampoule is torn apart and the electrolyte is released and activates the battery.

6.3.1.1.4. Battery Activation by Ammonia

This battery construction contains an electrolyte compartment, which is filled with ammonia, instead of an ampoule. It is separated from the cell stack by a seal.

A multicell ammonia reserve battery of this type is shown in Figure 54. When activating the gas generator (not shown) the membrane of the NH_3-reservoir is forced down, including the lance, that punctures the seal of the battery compartment (see right-hand figure). Under this condition the ammonia is injected into the battery compartment and distributed via its manifold system to the cell stack. The cell stack contains NH_4SCN and KSCN in a dry form. They are dissolved in ammonia, such that the electrolyte obtains its final composition within the electrode stack. About 200 ms after activation the battery is ready for use.

The liquid ammonia reserve battery employs ammonia in combination with ammonium thiocyanate (NH_4SCN) and potassium thiocyanate (KSCN) as electrolyte [122]. Ammonia offers several advantages like a low viscosity and in combination with above thiocyanates a high conductivity in a wide temperature range (– 55 – + 74 °C). The low viscosity of ammonia facilitates its distribution within a multicell battery arrangement. Different electrochemical systems are used like the 3 V – (PbO_2 – NH_4SCN – Mg)- or the 2.3 V – (S – KSCN – Mg)-system.

6.3.1.2. Thermal Batteries

6.3.1.2.1. General

Thermal batteries also belong to the group of electrolyte-activated primary reserve batteries. Their activation, as well as their cell and battery components differ, however, significantly

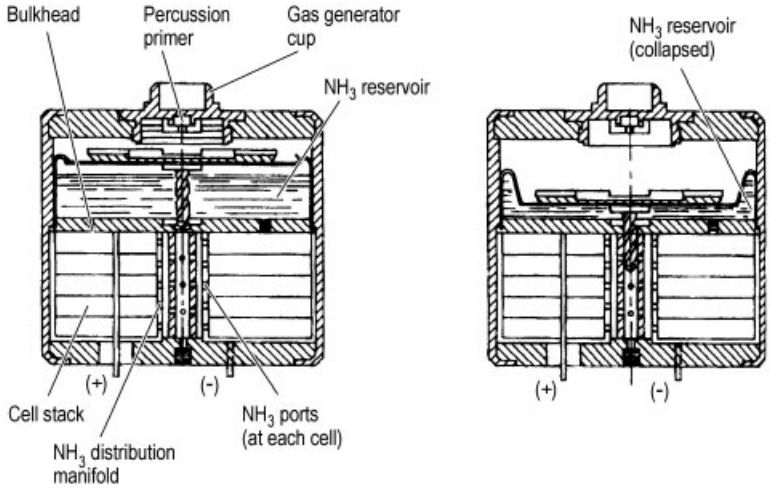

Figure 54. Liquid ammonia multicell reserve battery, activation by a pyrotechnic gas generator (capacity approx. 430 mAh, voltage 9 V) Alliant Techsystems, USA ([125], Fig. 20.4)

from ambient-temperature systems, such that they form a group of their own [123].

The unique feature of a thermal battery is the electrolyte which is an inorganic salt and a nonconducting solid under ambient temperature conditions. It is in direct contact with the system's anode and cathode. When the solid electrolyte is heated up to its melting point, i.e., activated, it becomes a very fast ion conductor and the electrochemical system is active and ready to provide electric energy.

The heat source to melt the electrolyte is a pyrotechnic material being integrated into the battery and activated by application of an energy impulse from an external source to a built-in initiator, which ignites the pyrotechnic. After activation, the temperature rises above the melting point of the electrolyte of 352 °C (eutectic mixture of LiCl + KCl) but remains below the melting point of the active materials (anode and cathode). Thermal insulation of the battery is important. The external temperature of the battery is at its peak at least 230 °C. If the temperature of the electrolyte falls below 352 °C the battery stops working. A schematic drawing of a thermal battery is shown in Figure 55.

General characteristics of thermal batteries are:

1) Long shelf life in a nonactivated state (> 10 a)
2) Provision of electric power in less than a second after activation
3) High reliability
4) Insensitivity to mechanical shock, acceleration, and vibration
5) Active life up to 1 h but typically less than 10 min

6.3.1.2.2. Systems

A wide variety of systems have so far been subject of research and development. The following *cathode materials* have been tested: Calcium chromate ($CaCrO_4$), potassium dichromate ($K_2Cr_2O_7$), potassium chromate (K_2CrO_4), lead chromate ($PbCrO_4$), metal oxides (V_3O_5, WO_3) and sulfides (CuS, FeS_2, CoS_2). Calcium, magnesium, lithium, and alloys thereof are in use as *anode materials*. Most thermal battery systems employ an eutectic mixture of lithium chloride and potassium chloride (*mp* 352 °C) as electrolyte.

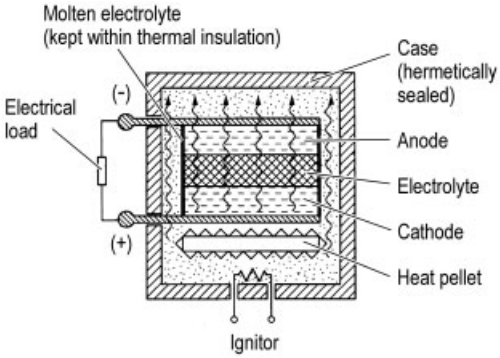

Figure 55. Schematic drawing of a thermal battery [110]

Of all systems investigated so far preference has been given in recent years to the iron disulfide – lithium system because of its stable and predictable chemistry at high temperatures. The calcium chromate – calcium system has been attractive for a long time because of its higher average operating cell voltage (2.4 V compared to 1.9 V). New promise for more high temperature stability and active battery lives of above 1 h is offered by the CoS cathode [124].

6.3.1.2.3. Pyrotechnic Heat Sources

Two different types of heat sources are in use, *heat paper* and *heat pellets*.

Heat paper has the composition Zr – $BaCrO_4$ and is manufactured from zirconium powder and $BaCrO_4$. The mix together with water is formed into a paper by a paper-making process, cut into parts and dried. In a dry state it must be handled very carefully to avoid its ignition.

Heat pellets are made by cold pressing a dry blend of iron powder (Fe in excess) and potassium perchlorate ($KClO_4$). Heat pellets provide a heat content of approx. 1 kJ/g. After combustion the heat pellet remains an electronic conductor, serving as intercell connector.

6.3.1.2.4. Battery Construction

Figure 56 illustrates an exploded view of a thermal cell including heat pellets on the cell's top and bottom. The top-cover contains four metal-to-glass terminals: Two for the battery (+ and −) and two for the ignition of the heat pellets.

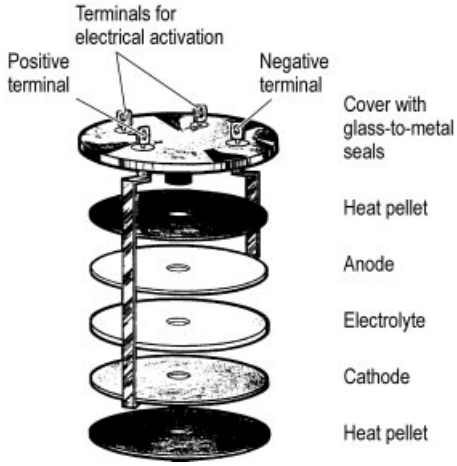

Figure 56. Exploded view of a thermal battery (schematic) [110]

Different cell constructions are in use. The *cup cell* uses two heat pads on its outside, two cathodes, two electrolyte pads and in the middle a "two-sided" anode (Fig. 57 A). The *pellet cell* construction consists of the anode (e.g., Ca – Fe bimetal) and a two-layer pellet, i.e., cathode – electrolyte plus a heat pellet on the cathode's side (Fig. 57 B).

6.3.1.2.5. Batteries

The *iron disulfide – lithium batteries* have the following characteristics and cover the following ranges:

Nominal voltage: 25 – 315 V
Current: 1 – 17 A
Peak power: 138 W – 3.6 kW
Average life: 35 – 900 s
Weight: 270 g – 6.62 kg
Volume: 92 cm^3 – 3.12 L

The *calcium chromate – calcium batteries* have the following characteristics and cover the following ranges:

Nominal voltage: 28 – 42 V
Current: 1.2 – 3 A
Peak power: 34 – 125 W
Average life: 25 – 150 s
Weight: approx. 310 g
Volume: approx. 120 cm^3

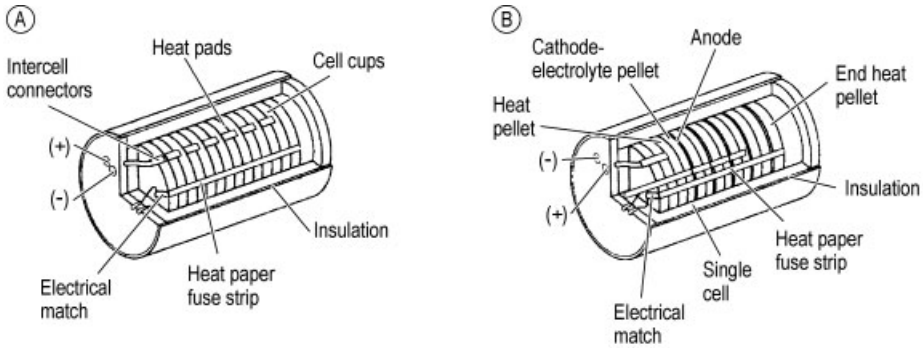

Figure 57. Thermal battery, multicell arrangements
A) Cup cell construction; B) Pellet cell construction ([125], Fig. 22.7)

6.3.2. Prospects

Reserve batteries will remain to be special-purpose batteries. They are not available on the commercial market. Their main use will continue to be in the field of military and space appliances, apart from rescue systems. Further developments are difficult to judge.

More information on reserve batteries is given in [125].

7. Secondary Batteries

Secondary batteries are genuine storage systems, since their discharging reaction can be reversed and the battery recharged. The efficiency of the storing process is limited for two reasons:

1) Kinetic parameters cause polarization that decreases the voltage during discharging, and increases it during charging. The gap between charging and discharging voltage characterizes the energy loss.
2) Secondary reactions that occur in parallel may consume a portion of the current applied for charging.

7.1. Batteries with Aqueous Electrolyte

Most of the commercially used batteries are based on aqueous electrolytes such as dilute solutions of sulfuric acid (H_2SO_4) and potassium or sodium hydroxide (KOH or NaOH, respectively). The main advantage of these electrolytes is their high conductivity.

Table 13 shows characteristic data of aqueous battery systems that are of technical interest. The values in the last two columns are only approximate data of comparison, since they largely depend on the design of the individual battery.

Redox and air – metal batteries that also use aqueous electrolyte, are treated in Section 7.1.3.3 and in Sections 6.1.1.2, 6.1.3.4, 7.1.3.4, respectively. The manganese dioxide – zinc system which has a certain rechargeability is treated in Section 6.1.3.1. A number of further electrode combinations that operate in aqueous electrolyte are known, but they have no technical importance.

Reactions of the Electrolyte. The main disadvantage of aqueous electrolytes is the comparatively low decomposition voltage of water of 1.23 V. As soon as the cell voltage exceeds this value, the water in the electrolyte is decomposed into hydrogen and oxygen according to:

$$H_2O \longrightarrow H_2 + 1/2 \cdot O_2$$

[Water decomposition]Water decomposition is the combination of two electrochemical reactions:

The evolution of oxygen at the positive electrode by oxidation of O^{2-} ions according to

$$H_2O \longrightarrow 1/2\,O_2 + 2\,H^+ + 2\,e^- \text{ for acidic media}$$

or

$$2\,OH^- \longrightarrow 1/2\,O_2 + H_2O + 2\,e^- \text{ for alkaline media}$$

and the evolution of hydrogen, by reduction of H^+ ions at the negative electrode:

Table 13. Rechargeable battery systems with aqueous electrolyte. Lead – acid systems use dilute sulfuric acid as electrolyte, the others dilute KOH or occasionally NaOH. Performance data refer to a single cell.

	System	Open circuit voltage, V/cell	Stored energy		
			Theory,	Practice,	
			Wh/kg	Wh/kg	Wh/kg
1	Lead – acid	2 [a]	161	20 – 50	60 – 100
2	Nickel hydroxide – cadmium	1.32 [b]	240	20 – 50	40 – 60
3	Nickel hydroxide – hydrogen	1.34 [b]	387	50 – 60	50 – 70
4	Nickel hydroxide – metal hydride	1.34 [b]	≈ 300 [c]	50 – 80	100 – 150
5	Nickel hydroxide – iron	1.3 [b]	267	40 – 60	80 – 120
6	Nickel hydroxide – zinc	1.73 [d]	341	80 – 100	140 – 180
7	Silver oxide – zinc	1.8/1.6 [e]		70 – 120	130 – 220
8	Silver oxide – cadmium	1.4 [f]	310	50 – 75	90 – 140

[a] Open circuit voltage depends on acid concentration.
[b] Uncertain value on account of peculiarities of the nickel hydroxide electrode; nominal voltage 1.2 V/cell.
[c] Depends on the alloy, used for hydrogen storage.
[d] Nominal voltage 1.6 V/cell.
[e] Two plateau discharge (cf. Fig. 102).
[f] Nominal voltage 1.2 V/cell.

$$2\,H^+ + 2\,e^- \longrightarrow H_2 \text{ for acidic media}$$

or

$$2\,H_2O + 2\,e^- \longrightarrow 2\,OH^- + H_2 \text{ for alkaline media}$$

The equilibrium voltage of water decomposition is $U^o = 1.229$ V. When the cell voltage exceeds this value, water is decomposed according to the above equations.

The reversible heat effect of water decomposition is (cf. Eq. 3.1)

$$Q_{\mathrm{rev}} = T \cdot \Delta S = T \cdot \Delta H - \Delta G = 48.6\mathrm{kJ}$$

This is 20.5 % of the free reaction enthalpy, and, according to Equation (3.21), the calorific voltage of water decomposition is

$$U_{\mathrm{cal}} = 1.229\,\mathrm{V} + 0.252\,\mathrm{V} = 1.481\,\mathrm{V}$$

In regard to battery practice, the reversible heat effect lowers heat generation during overcharging since the difference $U - U_{\mathrm{cal}}$ is smaller than $U - U_{\mathrm{o}}$ (see Eq. 3.29).

Internal Oxygen Cycle. Valve-regulated lead – acid batteries (VRLA batteries), sealed nickel – cadmium, and nickel – metal hydride batteries are characterized by the internal oxygen cycle: Oxygen that is generated during charging or overcharging at the positive electrode is subsequently reduced at the negative electrode by reversal of the above O_2 producing equations. Oxygen generation at the positive electrode is thus counterbalanced at the negative electrode, and the created oxygen cycle absorbs the overcharging current that otherwise would decompose water into hydrogen and oxygen.

To achieve the required fast transport of oxygen, the electrolyte must be immobilized, i. e., soaked in a mat of fine fibers of polyamide or polyethylene in alkaline batteries, or in an absorbent-glass mat (AGM) in lead – acid batteries (cf. Section 7.1.1.10). The amount of electrolyte must be limited so that larger pores of the matrix remain open and provide space for fast diffusion of oxygen in the gaseous phase. In lead – acid batteries, the electrolyte can alternatively be fixed as a gel by addition of SiO_2. In the gel cracks are formed which also provide open space.

In both cases, the hindrance of the oxygen transport to the negative electrode is nearly nullified, since oxygen transport occurs about 10^6 times faster through the gaseous phase than via dissolution through the electrolyte, and only a thin wetting layer on the surface of the negative electrode has to be penetrated by dissolved oxygen. Thus, nearly all oxygen that is evolved at the positive electrode, reaches the surface of the negative electrode, and is there reduced by reversal of the above oxygen evolution reaction.

Such an ideal situation can only be achieved in nickel – cadmium and nickel – metal hydride batteries that can virtually be sealed. In valve-regulated lead – acid batteries, hydrogen evolution and grid corrosion occur as unavoidable secondary reactions as shown in Fig. 70.

7.1.1. Lead – Acid Batteries

The lead – acid battery is the most important electrochemical energy storage system. Its storage capacity is moderate compared to other electrochemical systems, but the battery has several features that make it especially suited for standby applications. In addition, its comparatively low price is one reason for its widespread use.

7.1.1.1. Charge/Discharge Reactions

The electrode reactions of the lead – acid system are both based on lead, but in different states of oxidation, namely Pb^{2+} and Pb^{4+}.

During discharge, Pb is oxidized to Pb^{2+} at the negative electrode

$$Pb \leftrightarrow Pb^{2+} + 2\,e^-$$

and Pb^{4+} is reduced to Pb^{2+} at the positive electrode

$$Pb^{4+} + 2\,e^- \leftrightarrow Pb^{2+}$$

The double arrow in both equations indicates that these reactions can be reversed (under addition of energy). This reversion describes the charging process.

These equations have to be completed, because the solubility of Pb^{2+} ions in sulfuric acid is extremely low ($\approx 10^{-6}$ mol/L), and any Pb^{2+} formed beyond this concentration is precipitated as $PbSO_4$. Furthermore, in the positive electrode, Pb^{4+} ions do not occur in the free form, but as lead dioxide (PbO_2). When both facts are observed, the completed reactions are:

at the negative electrode:

$$Pb + H_2SO_4 \leftrightarrow PbSO_4 + 2\,H^+ + 2\,e^-$$

and at the positive electrode:

$$PbO_2 + H_2SO_4 + 2\,H^+ + 2\,e^- \leftrightarrow PbSO_4 + 2 \cdot H_2O$$

Summation of the two electrode reactions results in the cell reaction:

$$Pb + PbO_2 + 2\,H_2SO_4 \leftrightarrow 2\,PbSO_4 + 2\,H_2O$$

When the lead – acid battery is discharged, lead (Pb) and PbO_2 are converted into $PbSO_4$. In other words, during discharge sulfate ions from the electrolyte are absorbed in the electrodes. When the battery is charged, sulfate is released from the electrodes. The reactions at both electrodes include the dissolved state, i.e., the reacting species are dissolved in the course of the reaction. The new chemical compounds, formed by the reactions, are precipitated again as solid matter. About 8 g of substance have to be dissolved per drawn or charged ampere hour.

Detailed examinations revealed that lead dioxide does not always correspond exactly with the stoichiometric formula PbO_2, but some (small) deviations are observed that indicate that the true formula reads $PbO_{1.99}$.

In the concentration range used in batteries, dilute H_2SO_4 is dissociated mainly into H^+ and HSO_4^- ions. Only about 1 % of the H_2SO_4 molecules dissociate into $2\,H^+$ and SO_4^{2-}. In consideration of this dissociation the cell reaction can be written as

$$Pb + PbO_2 + 2\,H^+ + 2\,HSO_4^- \leftrightarrow 2\,PbSO_4 + 2\,H_2O$$

which describes the actual situation more adequately. The free enthalpy of this reaction is $\Delta G = -372.6$ kJ. When this value is inserted into Equation (3.2) the standard value of the equilibrium voltage results: $U^{o,s} = 1.931$ V.

This standard value applies when the activities (effective concentrations) of the soluble components of the reaction, namely H^+, HSO_4^-, and H_2O amount to 1 mol/L. This is approximated by an acid concentration of 1.083 mol/L (≈ 10 wt % or a density of 1.066 g/cm^3). The dependence of the equilibrium voltage on concentration is given by the Nernst equation, Equation (3.5):

$$U^o = 1.931 + 0.0592 \cdot \log \frac{a_{H^+} \cdot a_{HSO_4^-}}{a_{H_2O}}\,\mathrm{V}$$

The equilibrium cell voltage depends only on the acid concentration. It is independent of the present amount of lead, lead dioxide, or lead sulfate, as long as all three substances are available in the electrode. The result of this equation is plotted in Figure 58.

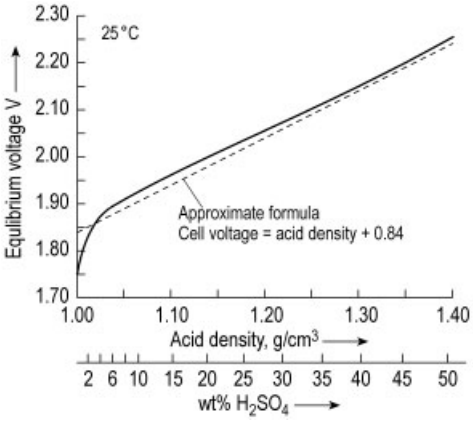

Figure 58. Equilibrium cell voltage of the lead – acid battery referred to acid density in g/cm^3 and acid concentration in wt % H_2SO_4

In battery practice, usually the following approximation is used:

Equilibrium cell voltage = acid density (in g/cm^3 or kg/L) + 0.84

Figure 58 shows that the calculated curve and the approximate formula coincide quite well.

Actually not the true equilibrium voltage but only the open circuit voltage can be measured with lead – acid batteries. Due to the unavoidable secondary reactions of hydrogen and oxygen evolution and grid corrosion, mixed potentials are established at both electrodes, which are somewhat different from the true equilibrium potentials (cf. Fig. 9). But the differences are small and can be neglected.

Since H_2SO_4 participates in the electrode reaction, the electrolyte becomes more diluted during discharge and is reconcentrated during charge. As a consequence, the discharge voltage does not remain constant even at low discharge rates. Furthermore, the conductivity of the acid varies when the battery is charged or discharged, and in cold climates the freezing point of the electrolyte may be reached during discharge. Then the resistance of the electrolyte is significantly increased and further discharging as well as subsequent charging are hindered. On the other hand, the changing acid concentration provides a tool to determine the state of charge.

The temperature coefficient of the equilibrium voltage is according to Equation (3.6):

$$\frac{\mathrm{d}U^{\circ}}{\mathrm{d}T} = 0.23\,\mathrm{mV/K}$$

This temperature coefficient is strictly valid only for standard condition, because the thermodynamic parameters depend on acid concentration as well as on temperature [126]. On the whole, the temperature coefficient of the equilibrium voltage is so small that it can be neglected in practice.

7.1.1.2. Secondary Reactions

The lead – acid battery is an unstable system in two respects:

1) Its cell voltage of 2 V is far above 1.23 V, the decomposition voltage of water.
2) The high potential of the positive electrode oxidizes all metals, usually applied as current conductors. Lead can be used, since on its surface a layer of lead dioxide is formed by corrosion that protects the underlying metal. But this layer is not totally stable and a certain current is continually required to reestablish the protecting layer. Corrosion is so gradually progressing into the metal that forms the grid and the conducting elements.

As a consequence, hydrogen evolution, oxygen evolution, and grid corrosion are unavoidable secondary reactions in the lead – acid system.

Figure 59 illustrates the situation. The electrode potential is plotted on the *x*-axis, the range of the negative electrode on the left-, the range of the positive electrode on the right-hand side. The two hatched columns represent the equilibrium potentials of the negative and positive electrodes. Their dependence on acid concentration is indicated by the width of these columns. The potential difference of 1.23 V between the beginning of hydrogen and oxygen evolution marks the water decomposition voltage.

The *charging and discharging reactions* are represented by the broken curves. They are very steep, since these reactions are fast, even at a small deviation from the equilibrium potential.

Hydrogen evolution begins at an electrode potential of 0 V which is about 0.3 V above the open-circuit potential of the negative electrode. *Oxygen evolution* starts at 1.23 V which is about 0.5 V below the positive electrode potential. Thus, hydrogen and oxygen evolution already occur at the open circuit potential of the

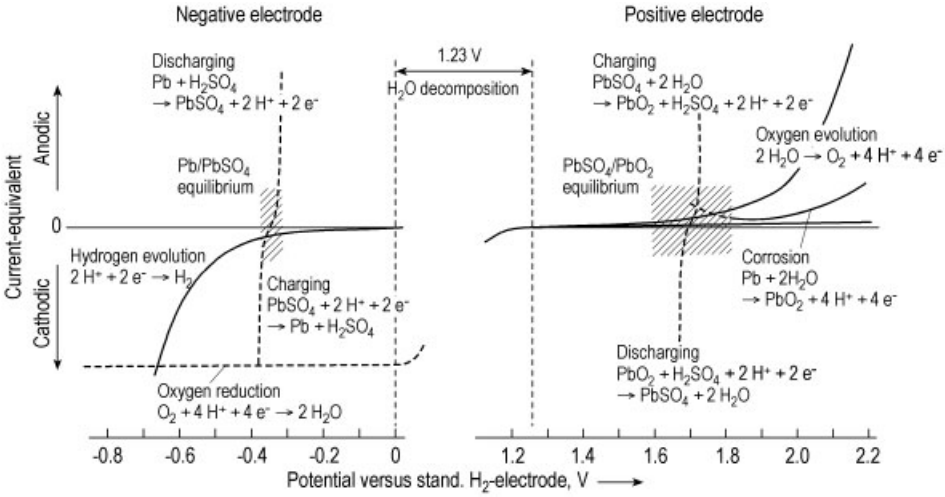

Figure 59. Reactions occurring in lead – acid batteries vs. electrode potential. The rates of these reactions are indicated by current-potential curves

negative and positive electrodes, respectively, and at open circuit, gradual self-discharge of both electrodes is the result. In Figure 59, both reactions are represented by continuous curves. The gradual increase at their beginning indicates that these reactions occur slowly as long as the potential remains fairly close to the origin of the corresponding curve. When, however, this deviation from the equilibrium potential of the concerned gas evolution reaction exceeds a certain value, the two curves show a steep increase. This means that hydrogen as well as oxygen generation gain in volume enormously at a correspondingly low and high overvoltage or polarization referred to their equilibrium potential. For this reason, small gassing rates characterize the lead – acid battery at the usual standby voltage of about 2.25 V/cell, but heavy gassing is observed at a voltage that exceeds 2.5 V/cell.

Corrosion of the positive grid also is an electrochemical reaction. The corrosion rate is at its minimum at about 40 to 80 mV above the $PbSO_4/PbO_2$ potential. At an electrode potential below this minimum, corrosion increases due to destabilization of the protecting PbO_2 layer. Above this minimum, the corrosion rate follows the usual exponential increase with increasing electrode potential.

Corrosion gradually destroys the current conducting elements, but is also a water consuming reaction, since oxygen that is required for the reaction is taken from the electrolyte

$$Pb + 2\,H_2O \longrightarrow PbO_2 + 4\,H^+ + 4\,e^-$$

Corrosion is a complex reaction, influenced by a number of parameters. Not only "external parameters" like temperature, electrode potential and acid concentration are important, but also "internal parameters" like composition of the corroding alloy, metallographic structure and local distribution of the alloying additives (→ Corrosion).

Oxygen reduction, the reversal of oxygen generation, is possible in the acid electrolyte at an electrode potential below 1.23 V vs. SHE. It is a quick reaction at the potential of the negative electrode that is far below + 1.23 V, and oxygen is immediately reduced when it reaches the surface of this electrode. For this reason, the rate of this reaction is largely determined by the rate of oxygen transport and therefore does not depend on the potential of the negative electrode. Oxygen reduction forms a limiting current that is characterized by the horizontal curve in Figure 59.

In flooded batteries, the transport of dissolved oxygen through the electrolyte limits this reaction rate to very low values. In VRLA batteries, however, the immobilization of the electrolyte allows fast transport of oxygen by diffusion in the gas-filled space. As a consequence, the inter-

nal oxygen cycle (cf. Section 7.1) characterizes this battery type.

7.1.1.3. Self-Discharge

At open circuit, the secondary reactions form mixed potentials with the discharge reactions, according to Figure 9.

At the positive electrode, oxygen evolution and grid corrosion are compensated by discharge, i.e., the reduction of PbO_2 to $PbSO_4$.

At the negative electrode, hydrogen evolution and oxygen reduction are compensated by the oxidation of lead to $PbSO_4$. Since the supply of oxygen from the positive electrode is negligible at open circuit, in flooded lead – acid batteries the self-discharge rate of the negative electrode is largely caused by hydrogen evolution.

The self-discharge rates of positive and negative electrodes occur independently from each other. They depend markedly on the state of charge of the concerned electrode and are reduced with progressing discharge. Furthermore, an increase in temperature of 10 °C approximately doubles the self-discharge rate.

7.1.1.4. Heat Effects

The reversible heat effect, defined by Equation (3.19), amounts to $Q_{\mathrm{rev}} = 13.2$ kJ per formula conversion of the cell reaction.

Comparison with the value for ΔG (− 372.6 kJ) shows that the reversible heat effect is only 3.5 % of the amount of energy converted during charge or discharge. The positive sign means cooling during discharge, and additional heat generation when the lead – acid battery is charged. But only when the lead – acid battery is discharged at very low rates (duration of discharge > 10 h) slight cooling of the battery can be observed. Usually, the reversible heat effect is concealed by the Joule effect.

According to Equation (3.20), Q_{rev} can be expressed as a (fictive) voltage. Since $Q_{\mathrm{rev}}/2 \cdot F = 0.068$ V, the calorific voltage in lead – acid batteries is according to Eq. (3.21)

$$U\mathrm{cal} = U^{\mathrm{o}} - 0.068\,\mathrm{V}$$

For heat calculations, this reduced value has to be used instead of the equilibrium voltage.

The above calculated values for the reversible heat effect, the temperature coefficient, and the calorific voltage are strictly valid only for standard conditions; the values actually depend on acid concentration and temperature, but the deviations are rather small and their contributions so insignificant that the conclusions can be extended to the usual range of acid concentration and temperature.

During charging, heat generation due to decomposition of water is much higher than that of the charging reaction itself, since the calorific voltage U_{cal} of water decomposition is 1.48 V and thus much lower than the cell voltage. For this reason, the current share that is consumed by water decomposition generates much more heat than it would when it were used for charging the battery. Figure 60 is an example of the heat generation in a charging/discharging cycle of vented lead – acid batteries, usually applied in traction applications.

The voltage curve is shown at the top of the figure:

- A current-limited initial step of charging is followed by a constant-voltage period at 2.4 V/cell. Equalizing charging at constant current up to about 2.65 V/cell is the final step of this charging schedule.
- Discharge is assumed at constant current ($I_5 = 20$ A/100 Ah). The broken curve shows the calorific voltage, while the full line represents the actual discharge voltage.

The middle of the Figure shows the current and its distribution with respect to charging, water decomposition, and discharging.

- During the initial stage, practically only charging occurs. Only when the voltage approaches the 2.4 V level, the onset of water decomposition becomes noticeable.
- As long as the cell voltage remains at 2.4 V, gas evolution is maintained at a roughly constant rate (assuming that the potentials of the positive and negative electrodes do not change too much).
- During the equalizing step, nearly all the current is used for water decomposition on account of the progressively reduced charge acceptance.
- During discharge, water decomposition is neglected.

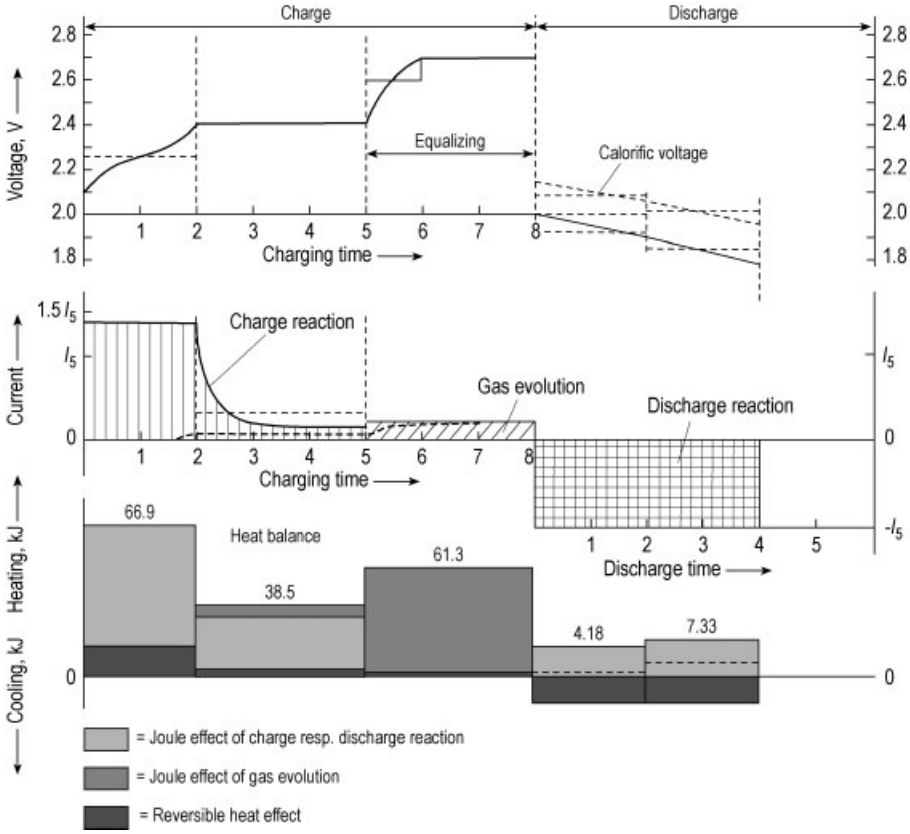

Figure 60. Charging/discharging cycle of a vented traction battery, a lead – acid battery with tubular positive plates (VARTA PzS), capacity 500 Ah.
Heat-generation values refer to 100 Ah of nominal current. The figures in the bottom part represent heat generation in total. The sum of the whole charging period amounts to 166 kJ/100 Ah.

At the bottom of Figure 60, the heat generation is drawn as blocks that represent average values for the corresponding sections of the charging/discharging processes. The distribution between reversible heat effect, charging, and water decomposition is marked by different patterns. The value above each block is the total heat generation in kilojoules.

- During the first stage of the charging process, gas evolution can be neglected. The main portion of heat is generated by the Joule effect, on account of the fairly high current. But the reversible heat effect also contributes noticeably to heat generation, since a large amount of active material is converted.
- When 2.4 V is reached, the current is reduced and, as a consequence, the Joule effect and the reversible heat effect are reduced, too. The now approximately constant gas evolution causes a corresponding heat generation.
- During the equalizing step, gas evolution (mainly required for mixing of the electrolyte, see Section 7.1.1.7) dominates. On account of the large difference between the actual charging voltage and the calorific voltage of water decomposition, heat generation is considerable, although the current is rather small.
- During discharge, heat generation is also small, owing to the small overvoltage, and further reduced by the reversible heat effect that now means cooling.

Heat Capacity. The temperature response of a battery to heat generation or heat loss is determined by its heat capacity according to Equa-

Table 14. Typical values for the heat capacity of various lead – acid batteries, and the share caused by the electrolyte

Type of battery	Heat capacity		Percentage caused by the
	$kJ\,K^{-1}\,kg^{-1}$	kJ/(K · 100 Ah)	electrolyte
Starter (flooded SLI ∗∗)	0.94 – 1.2	30 – 37∗	75 – 85 %
Traction (600 Ah, tubular)	1.05	7	76 %
Stationary (600 Ah, past.)	1.15	8	82 %
Stationary (600 Ah, tubular)	1.18	10	80 %
Valve-regulated	0.75 – 1.0	6 – 8	64 – 75 %

∗ per 12 V battery. ∗∗ SLI = starting, lightning, and ignition (starter battery for motor cars).

tion (3.31). In Table 14 heat capacities of some lead – acid batteries are compiled.

Table 14 shows that the heat capacity does not vary too much for the different types and is close to $1\,kJ\,K^{-1}\,kg^{-1}$. For batteries of similar design, it is mainly determined by the acid percentage of the battery weight, due to the high heat capacity of water.

Maximum Heat Generation During Discharging. To estimate possible heating effects at extremely high-rate discharges or even short-circuits, a specific energy of 35 Wh/kg can be assumed that is typical for a starter battery. With the specific heat capacity of $1\,kJ\,kg^{-1}\,K^{-1}$ the temperature increase would be:

$$\Delta T = \frac{35 \cdot 3.6}{1} = 126\,K$$

if the whole energy content were converted into heat and heat dissipation were excluded (adiabatic heating). This, however, is a theoretical value. Actually only a small fraction of the energy content can be drawn at such a high rate, partly because of the slow diffusion processes, partly because lead sulfate, precipitated at the surface of the active material, forms a barrier that hinders further discharge, and the residual energy content can only be drawn at a low rate. For this reason, lead – acid batteries cannot be heated up to detrimental temperatures by their own discharge. Even when discharged by short-circuiting and heat cannot dissipate, a temperature increase of $< 30\,°C$ is to be expected.

This, however, does not mean that loads above specified limits may not cause damage. Internal connectors or terminals may be interrupted by melting and the sparks may ignite explosive gas mixtures in the head space of the battery.

Heat Problems During Charging, Thermal Runaway. During charging, the energy that flows into the battery is limited by the charger that permanently supplies energy, different from the discharge of a battery that rather soon is exhausted due to its limited storage capability. Furthermore, when the battery is scheduled for a short bridging time, the charging device will be designed to supply a comparatively high current. If, for example, the standby battery is scheduled to bridge a current interruption of 10 min, the 10-minute rate will be the normal requirement of the concerned application and the charging device will supply the corresponding current. Thus, under unfavorable and uncontrolled conditions the battery might heavily be overcharged and a corresponding heat generation will result. In vented cells, the water loss can be so high that cells are dried out, and fire hazards might be the worst consequence. In valve-regulated lead – acid batteries (VRLA batteries), the intense heat generation of the internal oxygen cycle can cause a thermal runaway situation, where heat generation increases faster then heat can dissipate.

Such a situation can, however, only occur when the battery is not regularly controlled and short circuits in individual cells or failure of the charging device are not noticed (cf., e.g., [127]).

7.1.1.5. Valve-Regulated Lead – Acid Batteries

Valve-regulated lead – acid batteries (VRLA batteries) are based on the internal oxygen cycle, described in Section 7.1 but the unavoidable secondary reactions hydrogen evolution and grid corrosion disturb this cycle as indicated in Figure 61.

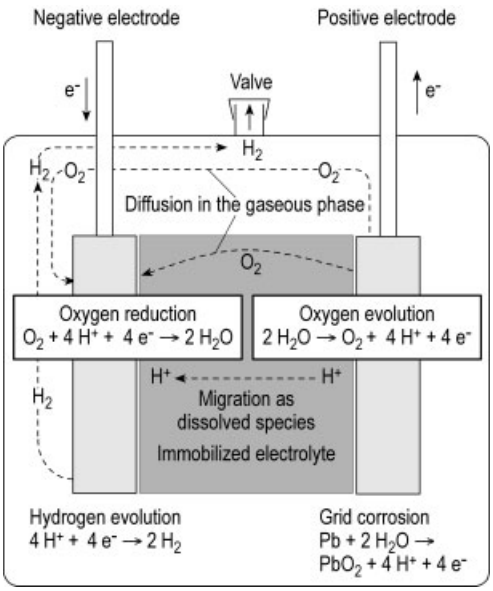

Figure 61. The internal oxygen cycle in valve-regulated lead – acid batteries that unavoidably is accompanied by evolution and escape of hydrogen, and by corrosion of the positive grid

As a consequence, the valve-regulated lead – acid battery cannot be made as a virtually sealed system. The evolved hydrogen has to leave the cell, since the rate of hydrogen oxidation that would form an internal hydrogen cycle is negligible. At the usual float voltage of about 2.25 V/cell, hydrogen evolution rates at minimum between 0.5 and 2 mL per 100 Ah and per hour are to be expected that escape periodically through the valve, if hydrogen is not lost otherwise, e.g., by permeation through the container wall.

Oxygen is consumed by corrosion. Both reactions mean water loss, although only to a very small extent that normally just causes a slight reduction of the performance data during service life. However, under unfavorable conditions (elevated temperature), water loss can be a serious aging factor, since it cannot be compensated by refilling.

Of great importance is that at a high efficiency of the internal oxygen cycle, hydrogen evolution must balance the corrosion rate to ensure proper float charging, as is described in detail in Section 7.1.1.7 and illustrated in Fig. 66.

The container of a VRLA battery must properly be welded, the terminals be sealed excellently, and the valve must close precisely. Otherwise oxygen, that permeates into the VRLA battery from the surroundings, quickly reaches the negative electrode and acts as an additional anodic current that causes discharge of the negative electrode.

The valve-regulated design has a number of advantages:

- Water loss is so widely reduced that refilling is not required for the whole service life.
- Hydrogen evolution is reduced to a fraction of that of vented batteries.
- Acid cannot be spilled, and acid fumes do not escape from the battery under normal operational conditions. For both reasons, the battery can be installed close to electrical equipment.
- VRLA batteries can be operated in horizontal positions which often saves volume required for the battery (cf. Fig. 72).

Due to their advantages, VRLA batteries came into widespread use within less than twenty years. In the field of stationary applications they substituted most of the vented batteries. However, they are more sensitive to unfavorable operational conditions than their vented sisters. They should always be carefully kept under control, especially when they are operated at elevated temperatures (cf., e.g., [128], [129]).

7.1.1.6. Discharge Performance

The discharge performance is largely determined by design parameters of the individual type of battery. Electrode thickness and size are such parameters, but also electrode spacing, kind of separation, and concentration of the electrolyte influence discharge performance.

A typical family of discharge curves is shown in Figure 62. The dotted line is the open-circuit voltage and represents the potential decrease caused by acid dilution (cf. Fig. 58). The slope of this curve is mainly determined by the electrolyte content in the cell.

As the load increases, the cell voltage does not only drop to a lower level due to the internal resistance, but also the actual capacity is reduced even at the lower end-of-discharge voltage. The reduction in capacity is caused mainly by decreased utilization of the active material at in-

creased current load. This again is a result of the rate-limiting diffusion steps.

The end-of-discharge voltage is determined by national or international standards.

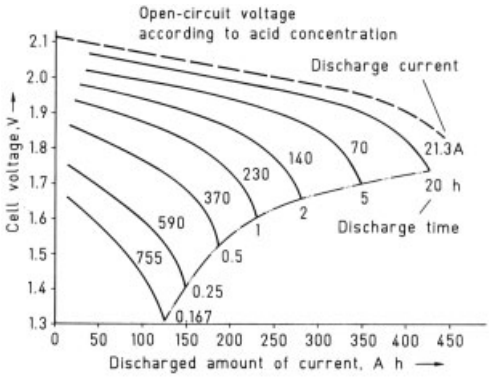

Figure 62. Discharge performance of a traction battery with tubular plates, nominal capacity (5 h rate) 350 A h

Initial Voltage Drop. At the beginning of discharge of a completely charged lead – acid battery, a small minimum is observed in the voltage curve (coup de fouet in Fig. 5). This voltage drop, on the order of 20 – 30 mV, is interpreted as nucleation overvoltage, i.e., the completely charged surface of the positive plate does not contain sufficient $PbSO_4$ nuclei for further crystal growth and the precipitation of new crystals affords oversaturation [130].

The initial voltage drop is the reason that in discharge curves published by manufacturers (Fig. 62) the first 10 % of discharge time is omitted because this voltage drop depends on the pretreatment of the battery.

7.1.1.7. Charge Performance

The charge performance of the lead – acid battery depends largely on the limiting influence of kinetic parameters. The charge acceptance is high as long as the state-of-charge is low, but charge acceptance is reduced as the battery approaches full charge. When the battery is charged at constant voltage, this results in the typical current patterns shown in Figure 63.

As charging begins, charge acceptance is high, and the current is limited by the charger. When the charging process is nearly completed, the charging current approaches zero and the current is reduced to the oxygen/hydrogen-evolution rate (float current). Three parameters determine the charging process: the initial charging current, the charging voltage, and the charging time. Increased charging voltage is effective only as long as charge acceptance is high; towards the end-of-charge it only increases the rate of water decomposition. For this reason, charging at increased voltage is recommended only for a limited charging period. The most effective way to reduce the charging time is to increase the initial charging current (Fig. 63).

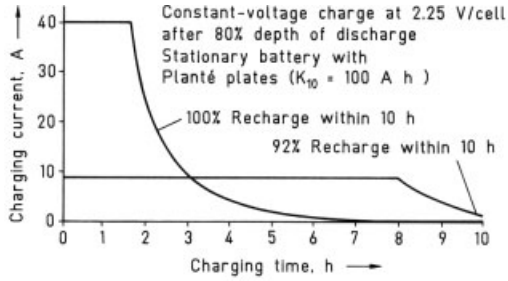

Figure 63. Charge acceptance at 2.25 V/cell; effect of initial charge current

Figure 64 shows the characteristic charging pattern for positive and negative electrodes, and for the cell. At the negative electrode, charge acceptance is close to 100 % for most of the charging time, as indicated by the horizontal curve III. The sudden switch to the hydrogen evolution potential corresponds to the cell-voltage step towards the end of the charge. The voltage curve of the positive electrode (curve II in Fig. 64) shows a slight increase from the beginning of the charging process, indicating a corresponding increase of oxygen evolution. This means that during the first stages of the charging process, the efficiency of the positive electrode is lower than that of the negative one. Because of different efficiencies, the positive plate is not charged completely when the voltage of the negative electrode suddenly increases, indicating complete charge of this electrode.

Actual charging schedules are always a combination of constant voltage and constant current steps as shown in Figure 60. When the battery is operated in charging – discharging cycles, a certain overcharge is required at the end of the cycle ("equalizing" in Fig. 60).

In flooded batteries, this final overcharge is mainly required to avoid acid stratification. During discharge, SO_4^{2-} is consumed at both electrodes. The resulting dilution of the acid means

reduction of its specific weight, and causes natural convection that mixes the electrolyte between, beside, and above the electrodes. Only the acid below the electrodes remains unchanged. During charging, the acid concentration is increased between the electrodes. Now the gravity causes mixing with the exception of the acid above the plates. At the end of the cycle, the acid above the plates is still diluted, while the concentration of the rest is slightly increased compared to the situation of a completely uniform concentration. If such cycles are repeated, a significant stratification of the acid is soon observed.

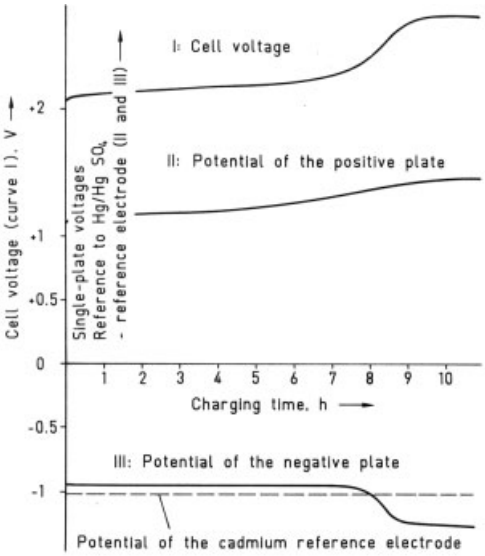

Figure 64. Constant current charging of a laboratory cell with excess electrolyte

To avoid stratification, conventional lead – acid batteries in traction applications are overcharged by 10 to 20 % per cycle to produce gas bubbles for mixing the electrolyte (the equalizing step in Fig. 60). This method is, however, not very effective and causes not only increased water loss, but also increased corrosion attack and heat generation, besides the fact that energy is wasted.

A superior method is to mix the acid by air-lift pumps (cf. [127, p. 258]). In such pumps gas bubbles are generated in a vertical tube by feeding a small volume of air from a central supply. These bubbles reduce the specific weight of the liquid inside the tube, which results in an upward movement of the acid column within the tube, and so acid is lifted from the bottom and poured above the top of the plates. Thus, complete mixing of the electrolyte can be achieved with a small amount of air and within a short period of time. Overcharging of traction batteries can so be reduced to about 4 % instead of 20 %. This method has been common for many years in submarine batteries [131] and is also used for solar applications. Since the beginning of the 1990s, it has been adopted for specially designed batteries for load-leveling applications [132] for electric vehicles [133], and also for normal traction applications [134].

In valve-regulated lead – acid batteries, acid stratification is no problem with gelled electrolyte due to the strong immobilization of the electrolyte. In cells with glass mat separators (AGM), a small influence of gravity is still observed, and the general recommendation is to avoid a separator height above about 30 cm by operating taller cells in a horizontal position (cf. Section 7.1.1.10). But VRLA batteries also require equalization charges when cycled. Here the reason is that the internal oxygen cycle consumes most of the current when the battery approaches the state of full charge. Small variations in the efficiency of this cycle would cause corresponding variations in the state of charge of individual cells within a string already after a fairly small number of cycles. For this reason, equalizing steps at increased voltage are usually applied in the charging schedule of VRLA batteries when cycled, although they increase the water loss to a certain extent (cf. [127, p. 422]).

Fast recharge is often desired, especially in electric vehicle applications. This can be achieved by high initial currents. In addition, such a high charging current is beneficial for the cycle stability of the active material, as more recent experiments with electric vehicles showed. However, controlled charging methods are a premise to avoid high-rate overcharging of individual cells in the battery.

Float Charging. One advantage of lead – acid batteries is that they can be recharged and kept fully charged at a voltage slightly above the equilibrium value. This is applied in many stationary applications where the battery is used as a standby battery and is only discharged for emergency energy supply if the network fails. Usually, the permanent overcharge, the float charge, is accomplished by connecting the bat-

tery to the d.c. power supply bus. Then the battery is continuously slightly overcharged, and in the case of a mains failure security is provided for uninterrupted power supply (although at a slightly reduced voltage).

Quasi float charging is applied to automotive batteries. Here the battery is also directly connected to the d.c. bus bar of the car, and is charged at a slightly increased voltage as long as the car runs (cf., e.g., [135]). But this charge is limited to the driving periods of the car, most of the time the battery stands at open circuit. It depends on the balance between charging periods and energy consumption, whether the battery really reaches the float situation. In many private cars, especially in winter time, the battery is not fully charged but remains between 75 and 90 % of its full charge.

Polarization of the positive and negative electrodes is an important parameter for float charging, since the positive electrode only remains fully charged when its potential is kept above its equilibrium value, while the potential of the negative electrode must be kept below its equilibrium value. At a given float voltage, polarization of the electrodes is determined by the general rule:

> Equal currents must flow through positive and negative electrodes

Figure 65 illustrates the situation for a flooded battery on float charge or overcharging at 2.25 and 2.35 V/cell, respectively. It corresponds in principle to Figure 59, but is drawn on a semilogarithmic scale and the value of the various parameters are quantified. The horizontal axis shows the electrode potential versus the polarization of the concerned electrode. The $PbSO_4/PbO_2$ and the $Pb/PbSO_4$ equilibrium potential in Figure 65 correspond to a definite acid density. Since negative logarithms do not exist, cathodic currents, like those for hydrogen evolution and oxygen reduction, are also drawn in positive direction.

In Figure 65, hydrogen and oxygen evolution form Tafel lines that are drawn with their typical slopes of 0.12 and 0.08 V per decade of current, respectively. These slopes and the vertical position of the Tafel lines depend of course on the individual parameters of the considered type of battery, such as cell design, used materials, and manufacturing methods. The values used in Figure 65 represent data that are typical for modern batteries with antimony free grid alloys (low hydrogen evolution). Increased hydrogen evolution, e.g., due to contamination by nickel or antimony, would cause a marked shift of the concerned curve in parallel to higher current values.

Corrosion is approximated by two Tafel lines, a falling one at small polarization values and an increasing one at higher polarization, which form a minimum at 60 mV above the $PbSO_4/PbO_2$ equilibrium potential. This is, of course, an arbitrary approximation, but reflects the general experience fairly well.

For oxygen reduction a limiting current density of 2 mA/100 Ah is assumed which applies for flooded batteries [136], where most of the generated oxygen escapes by ascending gas bubbles from the cell, and only a small portion reaches the negative electrode. At very low cell voltages, the generation of oxygen at the positive electrode can be the limiting factor as indicated by "lack of oxygen supply".

The rule that equal currents must flow though both electrodes is only fulfilled at a certain polarization of the positive and negative electrodes as indicated by the horizontal arrows. The broken curves in Figure 65 represent the total current

- hydrogen evolution + oxygen reduction at the negative electrode
- oxygen evolution + grid corrosion at the positive electrode.

In this example, at 2.25 V/cell a float current of 13 mA/100 Ah results. At the negative electrode, hydrogen evolution and oxygen reduction contribute 11 mA and 2 mA, respectively, while at the positive electrode it is the sum of 2.4 mA by corrosion and 10.6 mA by oxygen evolution. With the data used in this example, the result is a critically low polarization of the positive electrode due to the low hydrogen evolution rate that requires a comparatively large polarization of the negative electrode. Such a situation usually is experienced, when low float voltages are applied to flooded antimony free batteries, and it is a well known practice to preserve full capacity of the positive electrode by addition of traces of platinum that depolarize the negative electrode (increase the hydrogen evolution rate) [137].

Hydrogen evolution equals hydrogen escape, since hydrogen oxidation at the positive elec-

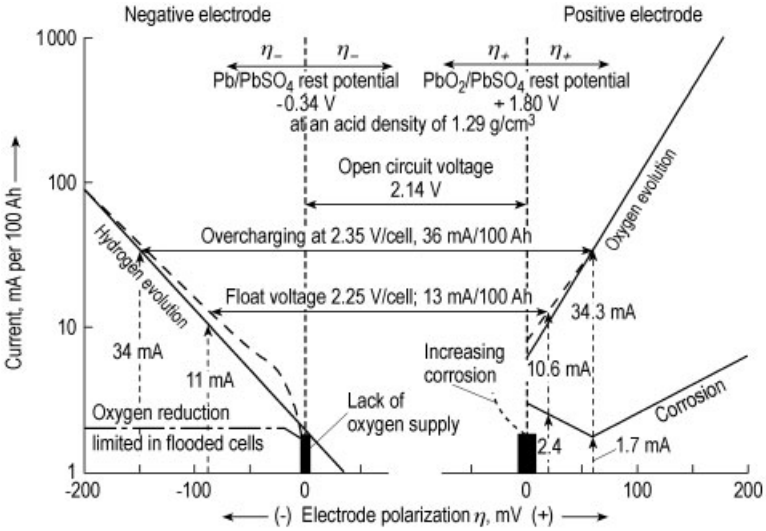

Figure 65. Float charging and overcharging of a flooded lead – acid battery

trode can be neglected. In this example, hydrogen evolution corresponds to an equivalent of 11 mA.

The total loss of oxygen is also equivalent to 11 mA: an equivalent of 8.6 mA escapes, an equivalent to 2.4 mA is lost by corrosion, while an equivalent of 2 mA is recombined at the negative electrode and so remains within the cell.

Summed up for a year's period, these figures mean a coulomb equivalent of 96 Ah (11 mA × 24 h × 365/1000) which corresponds to a water loss of 32 g per year and 100 Ah (1 Ah decomposes 0.336 g water).

Water loss is proportional to hydrogen escape. Thus water loss can be derived from determination of the amount of gas evolved combined with gas analysis. Determination of the weight loss is misleading, since the oxygen lost by corrosion remains in the cell. This applies also for VRLA batteries.

When this battery is overcharged at 2.35 V/cell, the polarization of both electrodes is automatically increased to fulfill the above rule that equal currents flow through positive and negative electrodes. Water loss is then increased to an equivalent to 34 mA or 297 Ah per year, i.e., 100 g water is lost per year and 100 Ah. This is an increase by approximately a factor of 3 compared to the voltage of 2.25 V/cell.

Figure 66 illustrates the situation for the valve-regulated design of the same battery, under the assumption that the efficiency of the internal oxygen cycle is 100 %, i.e., all the oxygen, that is generated at the positive electrode, is subsequently reduced at the negative electrode. (In practice, values very close to 100 % are usually reached). Since oxygen evolution is completely compensated by oxygen reduction, the equivalents of hydrogen evolution and corrosion must equal each other. Otherwise, the fundamental rule would be violated that equal currents have to flow through both electrodes. Consequently, polarization of both electrodes is now determined by the balance between hydrogen evolution and grid corrosion, as indicated by the broken horizontal double arrows.

At the float voltage of 2.25 V/cell, the required balance is reached at a polarization $\eta_- = -11$ mV and $\eta_+ = 99$ mV of negative and positive electrodes, respectively, and the resulting float current is 106 mA/100 Ah (103 mA O_2 evolution + 2.6 mA corrosion at the positive electrode and 103 mA H_2 evolution + 2.6 mA O_2 reduction at the negative electrode).

Compared to Figure 65, the situation in Figure 66 is quite different, although both are based on the same cell data. In the vented design (Fig. 65), the polarization of the electrodes is mainly determined by the balance between hydrogen and oxygen evolution. In the valve-regulated design, the polarization of the electrodes depends on the balance between hydro-

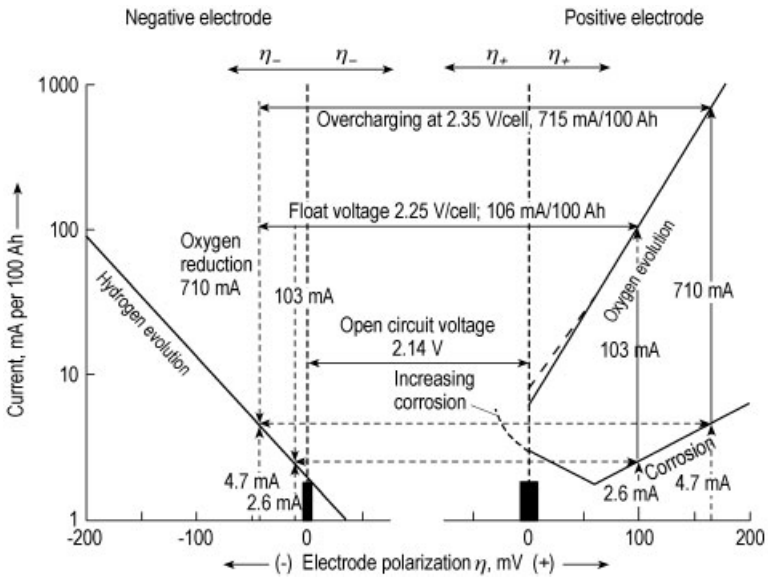

Figure 66. Float charging of a valve-regulated battery, under the assumption that the internal oxygen cycle has an efficiency of 100 %

gen evolution and grid corrosion, although both reactions contribute only to a small portion of the current.

Figure 66 indicates the typical situation in the VRLA battery: The polarization of the negative electrode is small compared to that of the positive one. The advantage of this situation is the correspondingly reduced hydrogen evolution. Only an amount equivalent to 2.6 mA is evolved, i.e., only 24 % compared to the vented design in Figure 65, although the float current is much higher. Water loss is correspondingly reduced. Oxygen is not lost from the cell. Only an oxygen equivalent of 2.6 mA is consumed by corrosion. 2.6 mA correspond to 23 Ah per year, that is equivalent to 8 g of water. The initial water content of a VRLA battery is about 1000 g per 100 Ah. So, in this example 0.8 % of the water content would be lost per year.

Figure 66 however, also shows a problem that may be encountered with valve-regulated batteries: If in this example, hydrogen evolution would slightly be increased, or the cell voltage be decreased, the required balance to corrosion could only be achieved at a positive polarization of the negative electrode, but this means discharge. Under such conditions, after a certain period of float charging, the full capacity of the battery could no longer be drawn, due to partly discharged negative electrodes, although the battery had properly been floated at the specified voltage.

At a higher charging voltage, as shown by "overcharging at 2.35 V/cell", the situation at the negative electrode is less critical, since the positive electrode is polarized to values with an increased corrosion rate. The so increased deficit between float current and oxygen evolution requires a correspondingly increased hydrogen evolution. This again causes a more negative polarization of the negative electrode, and ensures so that it remains fully charged. In practice, however, float charging at an increased voltage is usually not possible. But terminated boost charging or equalizing charging is occasionally included in monitoring routines for valve-regulated lead – acid batteries to recoup the full capacity of the negative electrodes [138].

As a consequence, the manufacturer of valve-regulated lead – acid batteries has to consider the balance between hydrogen evolution and grid corrosion. If highly corrosion resistant alloys are applied for the positive grid, a corresponding low gassing rate of the negative electrodes is required. The use of highly purified lead for the active material in the negative electrode would be the consequence. Battery practice, however, shows, that the required balance not always can

be reached and that "discharged negatives" occasionally are observed.

Developments during the 1990s indicate that the application of a small catalyst within the cell can be used as a remedy against such an imbalance caused by a too high hydrogen-evolution rate [139–141]. The catalyst in the top space of the cell recombines a certain portion of the oxygen directly with the abundant hydrogen, reduces so the efficiency of the internal oxygen cycle and in this way effects the required polarization of the negative electrode. Water loss is not increased, since the gas that is recombined by the action of the catalyst remains as water within the cell and water loss is even slightly reduced to the equivalent of corrosion.

7.1.1.8. Technical Design and Manufacturing

Batteries are mostly designed for specific applications. This leads to an indeterminable number of different designs. Only a few of the most important features can briefly be treated in this article.

Plate Construction, Cell Design. The design of the electrode must provide both mechanical support for the active material and electronic conductivity for the collected current. The most common plate design is the *grid plate* or pasted plate. A network of lead wires is filled with pellets of active material. The grid itself may be produced by casting of lead or lead alloys or by expanding or punching sheet metal. Two examples of grids are shown in Figure 67.

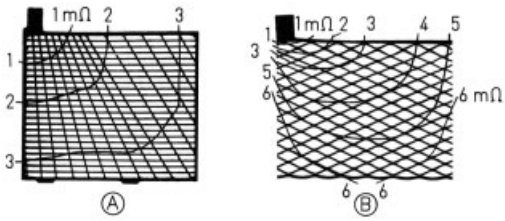

Figure 67. Two examples of SLI-battery grids with ohmic resistance map [142]
A) Cast grid; B) Expanded lead strip

The basic material for positive and negative active material usually is gray oxide, a fine powder consisting of about 70 – 80 % lead oxide (PbO) and the rest lead. This powder is mixed with water and sulfuric acid to form a paste, which is applied to the grid by pasting machines. To improve the mechanical strength of the active material, short fibers (0.1 – 0.2 %) are added. Furthermore, the paste applied in the negative electrode contains expanders, e.g., organic substances and barium sulfate, which maintain the activity (surface area) of the negative electrode material.

After the plates are pasted they are dried gradually over several days under controlled conditions. During this *curing process* 95 – 98 % of the lead in the paste is oxidized and the water, which originally accounts for 50 vol % of the paste, is evaporated. The volume occupied by the water forms the pore volume of the active material.

In most cell designs several plates are connected in parallel to form a plate group (Fig. 68). The positive (b) and negative (a) plates face each other. Separators between plates of different polarity prevent short circuits. The metallic extensions of the grids, called lugs (c), are welded to a top bar or strap (d) and the terminal post (e).

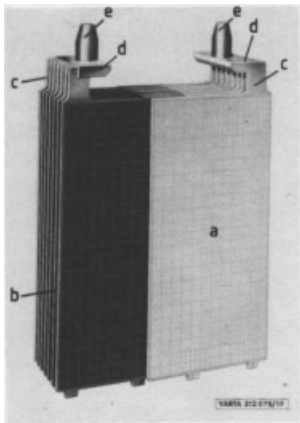

Figure 68. Pasted or flat plate group, without separators
a) Negative plate; b) Positive plate; c) Lug; d) Connecting bar; e) Terminal post

Often, a number of cells are combined into a common container to save volume and weight and form a mechanically stable unit. Such *block batteries* or monoblocks are mainly used with three cells (6 V block) or six cells (12 V block).

The usual starting, lighting, and ignition (SLI) battery in the motor car is a 12 V block.

Block batteries, as in Figure 69, are used in all kinds of electric vehicles like passenger cars, but also battery-powered handcarts. A further field of application is in wheel-chairs, where VRLA batteries are preferred due to the fact that acid cannot be spilled. Furthermore, the cycle performance of such valve-regulated batteries is superior to corresponding batteries of the flooded design [143] [$>$ 700 cycles (80 % depth of discharge) compared to $>$ 400 for similar batteries of vented design]. The main reasons for this superiority are that acid stratification does not occur and shedding of active material is prevented.

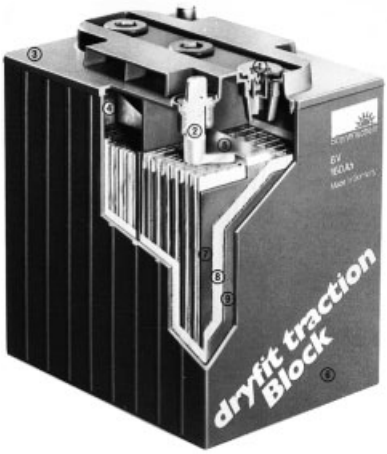

Figure 69. Valve-regulated lead – acid battery with gelled electrolyte for traction; three cells (6 volt), 160 Ah (Sonnenschein)
1) Valve (opening pressure $\approx$ 10 kPa); 2) Positive terminal; 3) Cell cover (polypropylene); 4) Intercell connection; 5) Pole bridge; 6) Container (polypropylene); 7) Positive plate; 8) Microporous separator with glass felt adjacent to the positive electrode (not a microporous glass mat); 9) Negative Plate

A different design principle is shown in Figure 70. It is used for batteries in the capacity range between 2.5 and 25 Ah. On the left, a cross-section of a monocell battery, on the right a 12 volt-block composed of six of these cells are shown. Each cell contains one positive and one negative electrode, which are separated by an absorbent-glass mat and spirally wound to form the cell element. The grid is punched from a lead sheet (no alloying additives, except a little amount of tin), soft enough to be wound as shown in the cutaway cell of the block battery. The cross-section shows how the grid is surrounded by the active material.

The cell is sealed in a polypropylene container, welded to a cover of the same material. In the monocell the container is mechanically supported by a metallic casing. A number of lugs on the top edge of the electrodes are welded to the terminal. The valve consists of a synthetic rubber cap which in the monocell opens at 3 to 4×10^5 Pa and in the blockbattery between 4.1×10^4 and 5.5×10^4 Pa because of the reduced mechanical support of the cells compared to the monocell. (In most VRLA batteries, the opening pressures are in the range of 10 kPa to prevent bulging of the container. Since such batteries are not overcharged at high current rates, the partial pressure of oxygen in the gaseous phase remains so low that it cannot open the valve, and that only insignificant amounts of oxygen escape when hydrogen exceeds the opening pressure of the valve).

A battery also based on wound electrodes, but designed for extremely high loads uses thin foils (0.05 mm) of a lead tin alloy instead of grids. The active material is coated on both sides in 0.075 mm layers. The separator (absorbent-glass mat) is 0.18 mm thick) [146].

The *tubular plate* design, sometimes called ironclad plate, is common in many countries for the positive electrode (Fig. 71). The principle of this design is to separate the conducting elements from the components which contribute mechanical support. The grid consists of vertical lead rods, which are the centers of tubes formed by woven, braided, or nonwoven fabrics. The tubes are filled with gray oxide or minium powder; some manufacturers use a paste-filling process. Improved utilization of the active material is the advantage of the tubular electrode. The central current collecting spine produces homogeneous current flow across the active material, and the mechanical support by the tube allows use of rather light, active material, which is equivalent to high porosity.

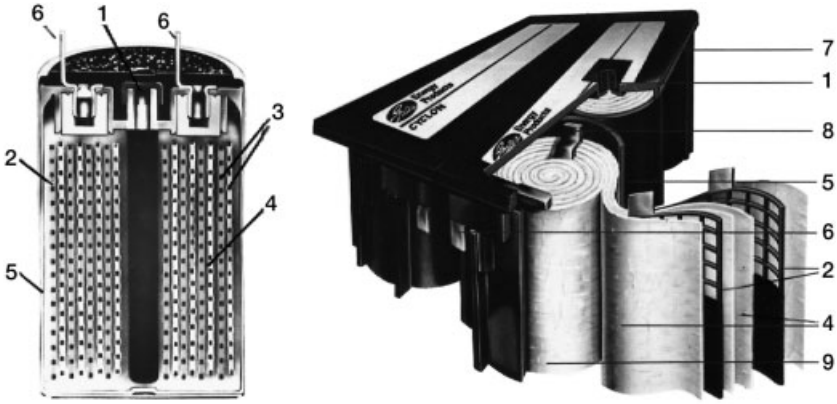

Figure 70. Battery with spirally wound electrodes (Left: Monocell [145]; right: 12-volt block)
1) Valve; 2) Grid, punched from lead; 3) Positive and negative electrodes; 4) Glass-mat separator; 5) Polypropylene container (as a monocell encased by a metallic housing); 6) Terminal (bobbin contacts or thread (for 25 Ah only); 7) Monoblock cover; 8) Intercell connector; 9) Spirally wound element of the two electrodes and the separator
(Gates Energy Products [144], now Hawker Energy Products Inc.)

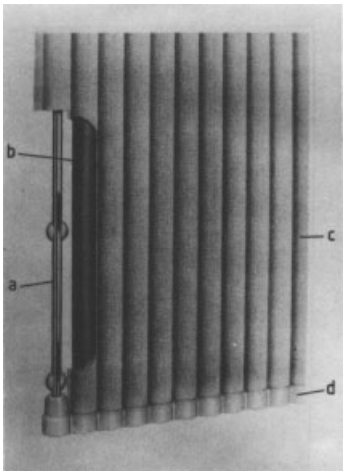

Figure 71. Tubular plate cross section
a) Lead alloy spine; b) Active material, PbO_2; c) Tube, in this example mixed fabric of glass and polyester fibers; d) Bottom seal of plastic caps

Tubular plates are in widespread use for flooded traction and stationary batteries. But they are also applied for VRLA batteries (see Fig. 72).

Figure 72. Stationary valve-regulated lead – acid battery with gelled electrolyte (Sonnenschein dryfit A 600)
48 volts, 1000 Ah (2 strings in parallel)

Figure 72 clearly demonstrates an advantage of valve-regulated lead – acid batteries, namely

that the blocks or cells can be operated in horizontal position. Despite the compact installation, easy access is given to each terminal for voltage measurements. But attention has to be paid to the high weight of lead – acid batteries. It may exceed the load capacity of the floor. Furthermore, heat problems may arise if the current flow is too high to be compensated by the cooling effects. A spacing of at least 1 cm has to be provided around each block or cell if heat dissipation by air convection is to be effective.

The *Planté plate* is an old plate design, based on castings of pure lead.

Figure 73 is a cross section of a Planté plate common in Europe. The lamellas of the casting enlarge the surface area by a factor of 8 – 12. The layer of active material is thin (ca. 200 μm). It is produced by anodic corrosion using perchloric acid to increase the corrosion rate. The thin layer of active material in combination with the large surface area of the current collecting lead body results in an excellent high-rate performance. The corrosion resistance of pure lead guarantees long service life. Disadvantages of Planté plates are the poor relation of capacity – lead content, and a poor capability for charge – discharge cycles because of capacity loss.

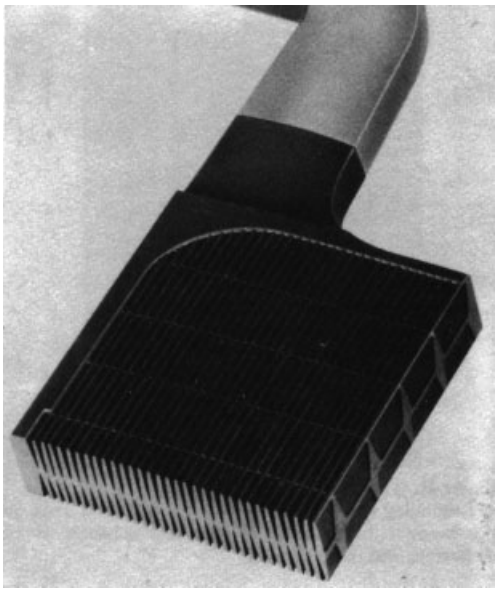

Figure 73. Planté plate cross section

Planté plates are still in use for flooded stationary batteries. Like tubular plates, they are always combined with negative grid plates.

In a battery especially designed by Bell for telephone stations, saucer formed pastes plates form horizontal stacked electrodes [147].

Formation, "Dry" Delivered Batteries. The conversion of lead oxide or minium to the final active material, i.e., lead and lead dioxide, is achieved by an electrochemical process called formation. *Tank formation* is formation of the plates before they are installed in the battery container. Usually positive and negative plates are assembled together in special jars with excess acid and, when the current flows, the material in the positive electrode is converted to lead dioxide, while that in the negative electrode is reduced to lead. After this process the plates are washed and dried. The spongy lead of the negative plate can be preserved by special treatments, otherwise the moist plate would react violently with oxygen to re-form lead oxide. A tedious recharge process would be required to recover full capacity. Charge preserving processes are: drying in vacuum, in vapor, or in hot mineral oil, or immersion in boric acid solution prior to the drying stage. Preserved negative plates can be stored for years without a significant loss of capacity. Cells equipped with precharged dry plates can be filled and put into service without any recharge even after long storage time. The service life starts with filling.

The lead dioxide in the positive electrode does not need preservation; the only requirement is to remove the sulfuric acid by thorough washing; otherwise the plate could not be dried.

Often raw plates are assembled in plate groups and inserted into the cell container prior to "jar formation," "cell formation," or "battery formation." The advantage of jar formation is that washing and drying after formation is avoided. On the other hand, these batteries can only be delivered filled with electrolyte, which has the consequence of limited storage time.

Valve-regulated lead – acid batteries are always delivered in the filled state.

7.1.1.9. Lead Alloys

The unavoidable corrosion of the positive grid and other parts of the conducting material has far-reaching consequences for the lead – acid battery:

1) Grid corrosion finally limits service life of every lead – acid battery, unless other failures occur earlier. A battery designed for a long service life must be equipped with a corresponding "corrosion reserve" of the positive grid.
2) Grid corrosion is a water consuming reaction which is of great importance in valve-regulated lead – acid batteries.

Table 15. Grid alloys for lead – acid batteries, their characteristic additives and fields of application

Kind of alloy	Additives	Field of applications
Standard antimony	4 – 11 % Sb, As, Sn, Cu (Ag)	traction
Low antimony	0.5 – 3.5 % Sb, Se, Te, S, Cu[a], As, Sn (Ag)	low maintenance (stationary, traction) SLI batteries
Standard calcium	0.06 – 0.12 % Ca, 0 – 3 % Sn, (Al)	stationary, vented and VRLA batteries SLI batteries
Low calcium	0.02 – 0.05 % Ca, 0.3 – 3 % Sn; Ag 0.008 – 0.012 % Al	valve-regulated, SLI batteries (continuous grid manufacturing)
Lead – tin	0.2 – 2 % Sn	valve-regulated batteries, conducting elements
Pure lead		Planté plates, Bell System round cell, (Bell Linage 2000)
ASTAG, ASTATIN[b]	0.009 % As, 0.065 Te, 0.08 % Ag (+ Sn = ASTATIN)	submarine batteries stationary batteries (Scandinavia, Italy)
Antimony – cadmium[c]	1.5 % Sb, 1 % Cd	valve-regulated

[a] Se, Te, S, and Cu are added as "grain refiners". [b] Alloys used by the Tudor Group in Scandinavia. [c] Only used by GNB.

The selection of appropriate alloying additives is of immense importance, since the rate of grid corrosion is influenced by the composition of the grid alloy. Furthermore, alloying constituents are released when the grid material is transformed into lead dioxide. These substances can be leached out of the positive plate and reach the negative electrode by diffusion, and cause increased hydrogen evolution by contamination (cf. Section 7.1.1.2).

Thus, the selection of grid alloys must not only consider fabrication requirements (to produce flawless grids), mechanical strength (for handling during production), creep strength (to resist growth during service), and corrosion resistance, but also pay attention to the fact that the alloy and the alloying constituents may cause electrochemical problems. As it is not possible to fulfill all these requirements at their optimum simultaneously, a number of different grid alloys are in use. They are described in detail in → Lead Alloys. (A survey on metallurgical properties of lead alloys for batteries is also given in [148]).

Table 15 shows the most frequently used alloy and their field of application.

7.1.1.10. Separators

In lead – acid batteries, the separator must not only prevent direct contact between the electrodes, but it also has to provide space for the required volume of acid. This is achieved by ribs or additional spacers or glass mats. Furthermore, the separator must suppress the growth of lead dendrites from the negative to the positive electrode.

PVC, rubber, polyethylene, and phenolic resin supported by polyester fibers are the most frequently used basic materials for separators in flooded batteries. An inexpensive separator that has been mainly used in SLI batteries is based on cellulose soaked up in phenolic resin that protects the cellulose fibers. Nowadays it is widely substituted by polyethylene and PVC based separators.

Sintered PVC separators are easy to manufacture. Their pore system is formed by voids between the sintered particles. Disadvantages are that they contain a certain amount of larger pores and that a certain thickness is required for their mechanical stability. For another kind of PVC separators, the molten plastic is extruded together with a high portion of silica (SiO_2) that forms the rather uniform and fine pore system (≈ 0.1 µm). Rather uniform pores of about 1 µm can also be obtained by addition of starch that later is extracted [149].

Polyethylene separators also contain silica, but also oil that later is extracted. In the *phenolic resin – polyester* separator a rather uniform sys-

tem of fine pores (1 μm) is formed by a special curing process of the resin.

Rubber separators are the oldest types. They also contain silica but also expanding additives to form the pore system. Rubber separators are still used in larger types of industrial batteries.

In many applications, mainly in SLI batteries, separators are used as pockets that envelope the plates and prevent dendritic growth of lead around the edges. In this regard, it is an advantage of PVC and polyethylene that both can easily be heat sealed.

AGM separators that are used in VRLA batteries, simultaneously act as the separator and the spacer that contains the electrolyte (cf. Section 7.1.1.5). Since glass is wetted by sulfuric acid, the glass fibers are covered with liquid and so the small voids between them are filled with electrolyte. Larger voids are not filled and form the pores for gas transport. To achieve the desired ratio of filled and open pores, glass fibers of different diameter are mixed. Sometimes, layers of different fiber types are combined to achieve uniform gas transport across the separator [149]. When the cell is assembled, a certain compression of the glass mat has to be observed. Otherwise, the contact between the separator and the electrode surface might be lost and cause capacity decay after a certain service life, since water loss causes a slight shrinkage of the glass mat.

Batteries with jellified electrolyte require conventional separators between the positive and negative electrodes, since the gel itself would allow the penetration of lead dendrites. Due to this additional separator, the internal resistance of a gel battery is slightly higher than that of a corresponding type with AGM separation. On the other hand, the influence of gravity is smaller in batteries with gelled electrolyte. Gelled sulfuric acid behaves in the same way as in a felt with a pore structure one order of magnitude finer than that of the usual microglass-mat separator. So the acid is more strongly fixed and gravity effects that are observed in a 30 cm high AGM battery would be observed to a similar extent in a 3 m high battery with gelled electrolyte. As a consequence, tall cells with gelled electrolyte can be used upright, while for corresponding AGM batteries the operation in horizontal position is mostly recommended to avoid stratification effects. For the same reason, the gel prevents electrolyte stratification even during extensive cycle operations.

7.1.1.11. Container Material, Post Seals

Containers nowadays are nearly exclusively made of plastics; seldom is hard rubber used. Because of its high impact resistance, relatively soft polypropylene is preferred for starter batteries or for small block batteries which attain mechanical strength by the multicell structure. It is also the material of choice for vehicle power batteries where the cell containers are supported by a tray (cf. Fig. 77).

Stiffer materials, such as polystyrene, or copolymers, like (styrene–acrylonitrile, SAN, shock-resistant polystyrene), are preferred for stationary batteries where the single cells usually are installed without any support (cf. Fig. 74). In special cases polycarbonate (extremely shock-resistant) or PVC (flame-resistant) is used.

Figure 74. Portion of a large, stationary battery

In most cases the lids are heat-sealed to the container; adhesive is only used for big stationary batteries. The post seal to the container cover or to the container itself is an important construction detail. Sulfuric acid wets the lead post, creeps along the surface, and tends to penetrate the seal. At the positive post this penetration may be accelerated by corrosion (Pb oxidizing to PbO_2), which, under unfavorable conditions, may result in detrimental crevice corrosion. Many battery manufacturers have made efforts to achieve post seals which definitely prevent acid leakage in order to achieve the maximum service life of the battery.

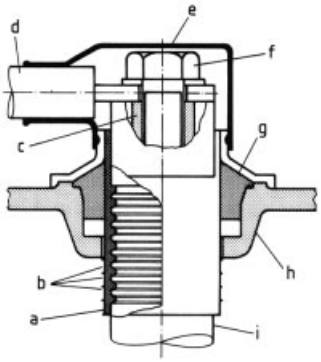

Figure 75. Post seal for stationary batteries (courtesy of Varta)
a) Injection-molded plastic sleeve; b) Grooves; c) Copper or brass insert to improve conductivity and contact to connector; d) Cable connector; e) Plastic cover; f) Bolt; g) Seal sleeve and cover; h) Cell cover; i) Lead post

Figure 75 is an example of a post seal. Grooves (b) are cut in the post and covered by an injection-molded sleeve (a). The shrinkage of this sleeve forms a tight seal around the lead post. Even after several years of service the acid penetration usually does not occur beyond the second groove.

In VRLA batteries, corrosion problems of the terminal post are of minor importance, since the distance between the electrolyte level and the post is larger, and acid fumes are hardly present. Nevertheless the post seal has to be made carefully, since any leakage would allow the penetration of oxygen into the cell and cause self-discharge of the negative electrode.

7.1.1.12. Vents, Valves

In *flooded lead–acid batteries*, the vent plug permits the exchange of gas between the interior of the cell and the surrounding atmosphere. Normal vent plugs have only a few narrow holes to keep the evaporation rate of water low. Furthermore, vent plugs are frequently fitted with porous disks or cylinders that only allow the passage of gas or vapor and retain fumes. Simultaneously, these disks prevent ignition sparks or flames from entering the cell from the outside.

In *VRLA batteries*, the valve allows only escape of gas from the cell to avoid overpressure that leads to deformation of the container (bulging of the cell) and may cause problems in regard to the contact pressure within the plate block. The valve must, however, totally prevent any flux or diffusion of air into the cell. If air penetrates, the atmospheric oxygen content of 21 vol % is reduced at the negative electrode and causes chemical self-discharge.

Two valves that have been developed for mass production under the aspect of high reliability are shown in Figure 76. Their parts have to be made with high precision.

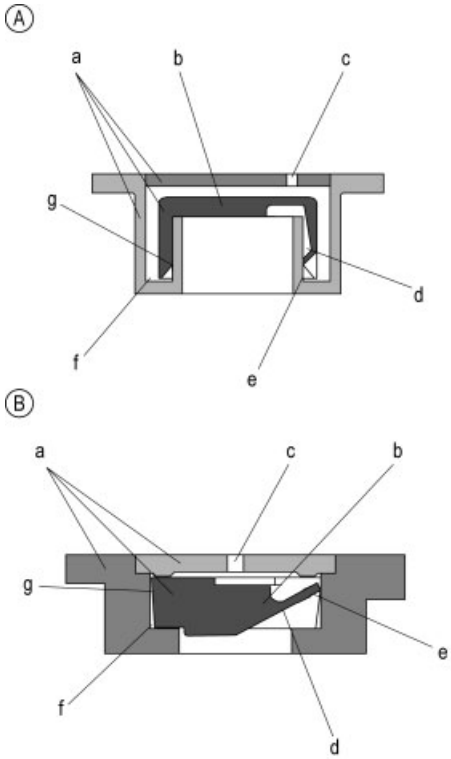

Figure 76. Valves for batteries, based on the deformation of elastomers
a) Both valves are made up of only three parts; b) Elastic cap in (A) or base of EPDM or FPM in (B); c) Outlet; d) Enlarged area for pressure response; e) Dynamic seal; f) Casing; g) Permanent seal (large area in (A) or sealing bead in (B)) [150]

The valve in Figure 76 A corresponds to a Bunsen valve. Such a valve is based on a rubber ring or cap that closes the opening of the cell in normal conditions, but is blown open when the internal pressure exceeds a specific limit. This design is widely used in small VRLA batteries. In the example of Figure 76 B, a cup of an elastomer forms the seal. Three grooves in the cap enlarge the surface area and so enhance the pressure response (the cut through one is shown in

the Figure). Furthermore, the seal is reduced in this section to a small edge (in Fig. 76) while the nongrooved section of the cap provides a safe permanent seal of a correspondingly large area (in Fig. 76) between the casing and the cap. This small edge is beneficial to achieve a definite opening pressure, since it reduces the danger that the cap sticks to the plastic tube.

The elastic base in Figure 76 B forms a butterfly valve. A bead seals the rest of this base so that air intake is absolutely prevented.

Oil can be used to improve the situation in both valves. Furthermore, both can be placed into a vent cup or into the lid of the container. Often such valves are covered by a sintered disk to prevent ignition from outside.

7.1.1.13. Water Refilling Systems, Forced Acid Agitation

Automatic water refilling systems are used for traction batteries.

In Figure 77, for example, each cell is equipped with a valve plug which automatically cuts water inflow when a float is lifted by the electrolyte level.

Figure 77. Traction battery (24 V, 500 Ah) equipped with automatic water refilling system (Varta Aquamatic)

The water pipe is simultaneously connected to its supply when the battery is connected to the charging device, but, water supply starts only at the end of the charging process to prevent overfilling.

7.1.1.14. Catalytic Gas Recombination

In flooded batteries, catalytic recombination of hydrogen and oxygen, with the produced water returning to the cell, is one alternative to reduce water loss. Appropriate devices are on the market and are installed instead of the usual vent caps [151]. These devices contain traces of palladium as the catalyst. Catalytic efficiency reaches 90 % or more, but some of the produced water is lost as water vapor. The efficiency (referred to water loss) therefore is reduced about 60 – 80 %. Nevertheless, recombination devices can extend the maintenance interval by a factor of three to five.

The possibility to improve the float charge behavior of VRLA batteries by catalysts has been described at the end of Section 7.1.1.7.

7.1.2. Batteries with Alkaline Electrolyte

A number of secondary battery systems is based on alkaline electrolyte, nearly almost diluted potassium hydroxide (KOH), because of its higher conductivity compared to that of sodium hydroxide (NaOH). Technically most important are nickel – cadmium and nickel – metal hydride batteries, especially as small sealed power sources for portable applications. Nickel – hydride batteries are employed as storage batteries in most satellites, for terristic applications they are too expensive. Nickel – iron batteries cannot be sealed because of the unavoidable hydrogen evolution at the iron electrode. They were mainly used as inexpensive traction batteries, but gradually lost their market share. The nickel – zinc system is attractive because of its high storage capability and its inexpensive negative electrode material. Despite enormous efforts that contiguously have been spent, however, an acceptable solution has not yet been found, because of the problems to recharge the zinc electrode. The silver oxide – zinc system is the most powerful alkaline storage system. Its high price and the limited number of possible recharges, however, limits use to special applications.

7.1.2.1. Nickel – Cadmium Batteries

The nickel – cadmium battery is one example of a larger number of batteries, based on an alkaline electrolyte (cf. rows 2 to 8 in Table 13).

Nickel – cadmium batteries have been in industrial production nearly as long as lead – acid batteries. In their sealed version they have gained enormous technical importance as maintenance-free batteries.

Characteristic features of nickel – cadmium batteries are:

- The electrolyte is (almost) not involved in the cell reaction, its concentration changes during charging or discharging can be neglected, and its volume can be minimized, which results in a battery of low weight and low internal resistance. Furthermore, the conductivity of the electrolyte remains fairly constant.
  As further consequences, problems of electrolyte stratification (see Section 7.1.1.7) are of little significance, and the fairly constant electrolyte concentration means a constant low freezing point of the electrolyte (between – 30 and – 60 °C).
  These features are the reason for the excellent high-rate and low-temperature performance of nickel – cadmium batteries and their capability of intermediate charging.
- Corrosion reactions of the current-conducting parts can usually be neglected during normal operation. A certain exception are foam electrodes (see Section 7.1.2.1.8), but even there the corrosion problem is not comparable with that in lead – acid batteries.
- Hydrogen is not evolved at the equilibrium potential of the cadmium electrode. Thus, hydrogen evolution can be avoided under certain preconditions (oversizing the capacity of the negative electrode so that always a rest is left uncharged as "charge reserve" and avoidance of too high charging currents).
  Since neither hydrogen evolution nor corrosion occur, nickel – cadmium batteries can be sealed completely, and the only reaction during overcharging is the internal oxygen cycle, described in Section 7.1. Such batteries are absolutely independent of their operational position, electrolyte cannot be spilled, and gases or fumes do not escape from the battery. For all these reasons, sealed nickel – cadmium batteries are ideal for the cordless power supply of electric or electronic devices and are in widespread use.

7.1.2.1.1. Thermodynamic Situation

Figure 78 illustrates the general situation for nickel – cadmium batteries. The sketch corresponds to Figure 59, but for the x-axis two scales are employed: The upper one represents the potential referred to the hydrogen electrode in the same solution (HESS). Since the equilibrium potentials of all the reactions in Figure 78 are shifted in the same manner with changing pH value (– 0.059 V/pH), this scale is independent of the electrolyte concentration.

The lower scale represents the electrode potential referred to the SHE. The electrolyte density has to be considered in this case, because, referred to this scale, all potentials in the Figure, are shifted with KOH concentration by – 0.059 V/pH.

The equilibrium potential of the cadmium electrode is 20 mV more positive than that of the hydrogen electrode. As a consequence self-discharge of the cadmium electrode caused by hydrogen evolution does not occur. Furthermore, *hydrogen evolution* can completely be avoided, if the (negative) polarization of this electrode during charging is kept so small that the hydrogen equilibrium potential is not reached. In practice, overcharging at the 5-hour rate is possible for many of the sealed designs without hydrogen evolution.

Oxygen evolution cannot be avoided at the positive electrode, because already the equilibrium potential of this electrode exceeds the equilibrium value of oxygen evolution.

Oxygen reduction occurs at the negative electrode. In flooded nickel – cadmium batteries this limiting current is as low as that in vented lead – acid batteries. In sealed nickel – cadmium batteries with immobilized electrolyte, oxygen transport rates can be quite high, and the internal oxygen cycle (cf. Fig. 61) can be correspondingly fast due to void volume within the separator and narrow spacing of the electrodes.

Corrosion of the nickel substrate in the positive electrode usually is negligible, and thus not shown in Figure 78.

7.1.2.1.2. Charge/Discharge Reactions

The discharge reaction (the oxidation) of the negative electrode is

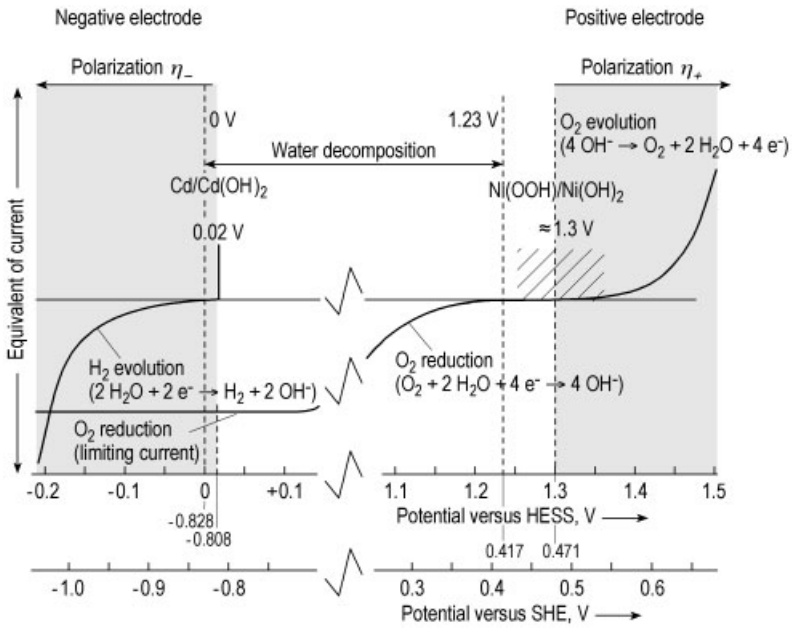

Figure 78. Equilibrium potentials of the negative and positive electrodes in a nickel – cadmium battery and current – voltage curves of hydrogen evolution, oxygen evolution, and oxygen reduction. (The latter is limited by the oxygen transport rate, as indicated by the horizontal curve). The SHE scale corresponds to a KOH solution of 1.250 g/cm^3 or 26 wt % at 25 °C.

$$Cd + 2\,OH^- \longrightarrow Cd(OH)_2 + 2\,e^-$$

When the positive electrode is discharged, Ni^{3+} ions are reduced to Ni^{2+} according to the (simplified) equation

$$NiOOH + H_2O + e^- \longrightarrow Ni(OH)_2 + OH^-$$

When both are summed up, the cell reaction during discharge results, that can formally be written as:

$$2\,Ni(OOH) + 2\,H_2O + Cd \longrightarrow 2\,Ni(OH)_2 + Cd(OH)_2$$

For the positive electrode, the reactions in fact are more complicated. Nickel hydroxide is not a stoichiometric chemical compound, and not only trivalent nickel ions are present in the charged material, but it can be a mixture of Ni^{2+}, Ni^{3+}, and Ni^{4+} (cf. Fig. 79). The charged state should be written more precisely as $u\,NiO_2 \cdot v\,NiOOH \cdot w\,H_2O$, where u, v, and w are factors that describe the share of the three components [152].

Furthermore, the reactions at the nickel hydroxide electrode are not quite reversible. Thus, the positive electrode in nickel – cadmium batteries does not attain a true equilibrium potential, and thermodynamic data and values are only approximate.

While the discharging/charging reaction of the cadmium electrode occurs via the dissolved state, the nickel ions are charged and discharged in the crystal lattice; as illustrated in Figure 79. To preserve electrical neutrality, a corresponding number of protons must migrate into the crystal lattice during the reduction (the discharge) of the Ni^{3+} or Ni^{4+} ions. Otherwise, the increasing space charge within the crystal would soon stop the process. When the nickel electrode is charged (oxidized) as shown in Figure 79, these protons have to leave the crystal lattice. This exchange of protons is accompanied by the formation of a reduced or an oxidized crystallographic phase. Figure 79 shows the oxidation and reduction of nickel hydroxide in an idealized form.

Actually, the charging/discharging reactions are complex. Two different couples of charged/discharged species have been specified: the β/β couple and the α/γ couple which are not independent of each other [153]. Both couples are characterized by a different maximum oxidation

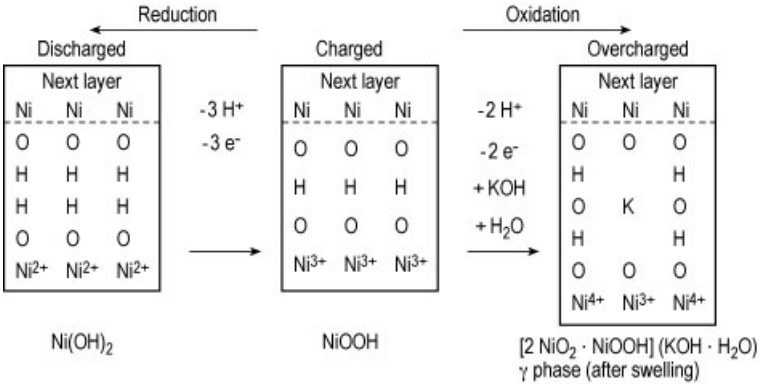

Figure 79. Charge and discharge of the nickel hydroxide electrode with simultaneous release and absorption of protons and incorporation of a small amount of potassium ions

state and different contents of KOH. The α/γ-couple absorbs about four times as much KOH as the β/β-couple [154]. Beyond that, the active material in actual battery electrodes shows poor crystallinity, and the reactions do not lead to definite chemical compounds, but cause progressively increased or decreased oxidation of the nickel ions (Ni^{2+}, Ni^{3+}, and Ni^{4+}). The consequence in battery practice is that the attained state of charge of a nickel hydroxide electrode depends on the charging period and on the electrode potential [155].

A remarkable feature of the above given simplified charge/discharge reaction is that neither KOH nor potassium appear in it or in one of the electrode reactions. Only H_2O and its dissociation product OH^- ions take part. Water, however, is the main constituent of the electrolyte, because the usually applied concentrations of KOH are in the range of 20 – 32 wt % (densities at 20 °C between 1.19 and 1.30 g/cm^3). From the discharge equation it can be deduced that 0.67 g water are required per Ah of discharge. Since in vented nickel – cadmium batteries, a minimum of 10 g of electrolyte is contained per Ah of nominal capacity, the change in density is only 0.015 g/cm^3, which will not be noticed in practice. In sealed batteries, despite their lower content of electrolyte, the density change caused by discharge does not exceed 0.02 g/cm^3, and can thus also be neglected.

7.1.2.1.3. Equilibrium or Open-Circuit Voltage

The free enthalpy of the discharge reaction is $\Delta G = -256$ kJ.

Inserting this value into Eq. (3.2) results in the equilibrium or open-circuit voltage: $U^o = 1.32$ V, in practice $U^o \approx 1.3$ V. Actually, after charging values between 1.3 and 1.4 V are observed, depending on the previous treatment of the battery. But on open circuit the cell voltage decreases to less than 1.3 V within about ten days [156].

The temperature coefficient of the equilibrium voltage is according to Equation (3.6)

$$\frac{dU^o}{dT} \approx -0.45\,\mathrm{mV/K}$$

This decrease of the equilibrium voltage of about 0.5 mV per kelvin can usually be neglected in battery practice.

7.1.2.1.4. Reversible Heat Effect, Calorific Voltage

The reversible heat effect of the discharge reaction is $Q_{rev} = T\Delta S \approx -26$ kJ per formula conversion of the cell reaction. Comparison with the value for ΔG shows that the reversible heat effect is 10.5 % of the converted energy when the nickel – cadmium battery is discharged or charged.

The negative sign means that additional heat is generated during discharge; and attention has to be paid to the heat balance when the battery is discharged with extremely high currents.

During the charging process, the reversible heat effect causes remarkable cooling as long as the heat generating secondary reaction of water decomposition does not occur. As a heat source, however, water decomposition can only be critical when the cell voltage considerably exceeds 1.48 V (= U_{cal} for water decomposition). As a result, vented nickel – cadmium batteries can be charged at quite high rates without suffering heat problems. But in the sealed version, the much higher heat generation by the internal oxygen cycle can cause serious thermal problems (see Fig. 81).

7.1.2.1.5. Heat Capacity (Specific Heat Content)

The temperature response of the battery to heat generation or heat loss is determined by the heat capacity of the battery, which can be calculated according to Equation (3.31). Table 16 gives characteristic heat capacities of some nickel – cadmium batteries.

Table 16 shows that the heat capacities of the nickel – cadmium batteries are of the same order of magnitude as those of lead – acid batteries, compiled in Table 14. Also here, the share of the electrolyte is a decisive factor. An exception is the heat capacity of the small button cell that is mainly determined by the metallic housing.

7.1.2.1.6. Heat Effects

Figure 80 shows heat generation in a vented nickel – cadmium battery when charged and discharged with a constant current (5-hour rate) and the charging voltage is limited to 1.65 V/cell. The calculation is based on the equilibrium voltage $U^{\mathrm{o}} = 1.3$ V and the calorific voltage $U_{\mathrm{cal}} = 1.44$ V. Due to the uncertain thermodynamic data, these calculations are only rough approximations, but correspond with practical experience.

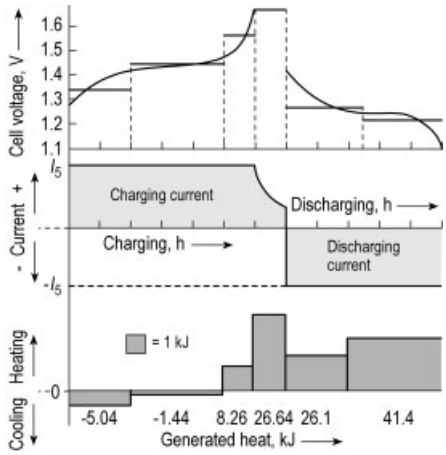

Figure 80. Heat generation during charge and discharge of a vented nickel – cadmium battery (VARTA TS-type, values referred to 100 Ah of nominal capacity). Charging is performed with constant current I_5 (5-hour rate) until 1.65 V/cell is reached. Discharge also occurs with I_5. In the top part, sections are marked that were used to calculate the average heat generation, shown in the bottom part. (Calculation based on $U^{\mathrm{o}} = 1.3$ V; $U_{\mathrm{cal}} = 1.44$ V)

Figure 80 shows that during the initial two sections of the charging period, slight cooling is observed on account of the reversible heat effect that consumes heat at a constant rate proportional to the current. With increasing cell voltage, the Joule effect is increased, and more and more conceals the cooling effect. But the main reason of the increased heat generation at the end of the charging period is that a growing portion of the current is consumed by water decomposition. The contribution of the charging reaction with its cooling reversible heat effect is so reduced, and when the charging voltage exceeds 1.48 V/cell, water decomposition contributes a corresponding amount of heat.

During discharging, heat generation increases with decreasing cell voltage. In total 67.5 kJ of heat are produced during discharging, while heat generation during charging only amounts to 28.4 kJ. This is mainly caused by the reversible heat effect that generates additional heat during discharge, while it compensates for heat generation during charging.

The heat evolution of a sealed nickel – cadmium battery during constant-current charging is illustrated in Figure 81. A charge factor of 1.4 is applied (such an amount of overcharge is usual for conventional charging

Table 16. Typical values of the heat capacity and its share assigned to the electrolyte (1 J = 1 Ws; 1 kJ = 0.2778 Wh)

Type of battery	Heat capacity,		Percentage share of electrolyte
	$kJ\,K^{-1}\,kg^{-1}$	Wh/(K · 100 Ah)	
Stationary, 315 Ah, pocket plates	1.3	1.9	78 %
Aircraft battery, 40 Ah (vented)	0.98	0.97	48 %
Sealed round cell, 4 Ah	0.81	0.72	47 %
Sealed button cell, 0.06 Ah	0.52	0.72	41 %

methods but can only be applied to comparably small batteries < 10 Ah).

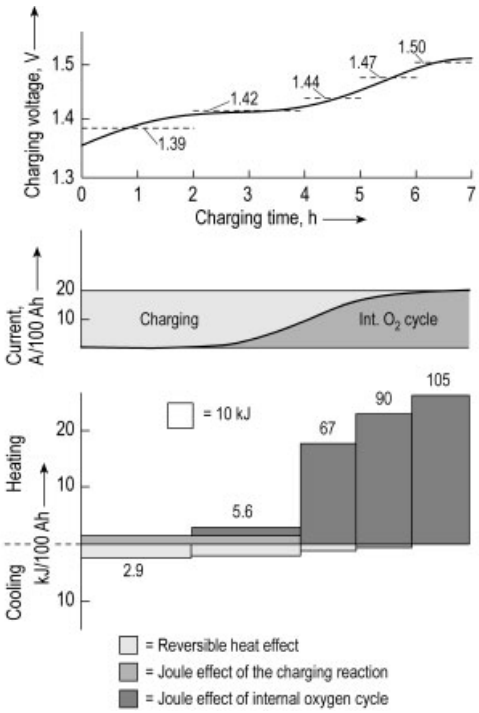

Figure 81. Charging of a sealed nickel – cadmium battery with constant current 0.2 C(A). During 7 h 140 % of the nominal capacity are recharged which corresponds with a charge factor 1.4.
For comparison, all values are converted to 100 Ah of nominal capacity. Actually, batteries of this type are in the size < 10 Ah. Larger sealed nickel – cadmium batteries can only be charged by controlled charging processes (Section 7.1.2.1.12)
Middle: current distribution between charging and internal oxygen cycle; Bottom: heat generation as an average of the different sections (slight cooling during the first 2 h)

- The voltage curve at the top shows the gradual increase of charging voltage with charging time. The generated heat is calculated as an average value for different sections of this curve. The numbers at the charging curve are the average voltages (V per cell) for the corresponding section.
- The figure in the middle depicts the (constant) current and its distribution between charging process and internal oxygen cycle.
- The bottom figure shows the heat generation as average value for the different sections. The numbers are the heat in kilojoules, converted to 100 Ah of nominal capacity for comparison.
  During the first two hours, the reversible heat effect exceeds the Joule effect and cooling is observed. Therefore the number for this section is written below the zero line.
  When the charging process approaches completion, nearly all the current is used for the internal oxygen cycle, which generates a large amount of heat.

Altogether 264.7 kJ = 73.53 Wh of heat are generated, referred to a nominal capacity of 100 Ah. Figure 81 shows that this heat is generated practically during the last 3 h of the charging process. This means an average heat generation of 24.51 W/100 Ah for these three hours. Thus, sealed nickel – cadmium batteries can be charged at a high rate as long as the current is actually used for charging and not for the internal oxygen cycle. Rapid charging methods, are always based on this principle.

Battery manufacturers usually strongly advise their customers not to operate sealed nickel – cadmium batteries at constant-voltage float charge without monitoring, because of the heat generation by the internal oxygen cycle. The fact that this cycle can attain extremely fast rates makes the situation really dangerous with respect to thermal runaway.

7.1.2.1.7. Self-Discharge

The factors that cause self-discharge of nickel – cadmium batteries are:

1) Decomposition of the charged nickel hydroxide by separation of oxygen according to

$$Ni_6O_6(OH)_6 + H_2O \longrightarrow Ni_6O_4(OH)_8 + 1/2\,O_2$$

$$[6NiOOH \approx Ni_6O_6(OH)_6]$$

With progressing reduction of the oxide, this disintegration process decreases and comes to a standstill when a stable state of oxidation is reached. Theoretically, a final oxidation state of 1.33 (Ni_3O_4) can be expected instead of 1.5 (NiOOH), which means that, at most, about half the charge would be lost. In practice however, only 10 to 20 % of capacity loss are caused by oxide decomposition.

The process is markedly increased at elevated temperature. At 40 °C, a steep decrease in capacity by nearly 20 % is observed within two days (see Fig. 82). But even at that temperature further self-discharge does not proceed to more than 22 % during 30 days.

In sealed nickel – cadmium batteries, oxygen release at the positive electrode also causes self-discharge of the cadmium electrode, since oxygen can easily reach the negative electrode where it is consumed.

2) The shuttle of a redox system always discharges both electrodes (see Section 4.2.6.2). In nickel – cadmium batteries, such a shuttle is usually caused by nitrate ions (*nitrate shuttle*), that are dissolved in the electrolyte. When they reach the negative electrode by diffusion, they are reduced to NH_4^+:

$$NO_3^- + 10\,H^+ + 8\,e^- \longrightarrow NH_4^+ + 3\,H_2O$$

The NH_4^+ ions migrate back into the electrolyte and are oxidized to nitrite when they again reach the positive electrode (the NO_3^- oxidation potential is not reached in open-circuit conditions). The cycle is repeated when the NO_2^- ions again reach the negative electrode and are reduced to NH_4^+ ions. Thus, the alternating oxidation to NO_2^- and reduction to NH_4^+ ions forms a shuttle as described in Section 4.2.6.2.

Current transport by the nitrate shuttle is effective, since six electrons are exchanged in each cycle when nitrogen is reduced from the oxidation state + 3 to − 3:

$$NO_2^- + 6\,H_2O + 6\,e^- \longrightarrow NH_4^+ + 8\,OH^-$$

Self-discharge caused by the nitrate shuttle is mainly observed in batteries with sintered electrodes, because nickel and cadmium nitrates are used in the manufacturing process to precipitate the active material into the sintered substrate. Another source of this shuttle is degradation of the polyamide separator.

3) The occurrence of short-circuits also affects both electrodes. Short-circuits can be caused by penetration of the separator by small particles washed out of the active material, or by the growth of crystal bridges between the electrodes.

4) Self-discharge of the positive electrode can furthermore be caused by hydrogen that has been evolved during overcharging of nickel – cadmium batteries at too high rates. But the effect is small, since the available amount of hydrogen is limited.

5) Self-discharge of the negative electrode occurs when oxygen penetrates into a sealed cell through a leakage (or is released at the positive electrode according to item 1). On account of the immobilized electrolyte, oxygen has access to the negative electrode surface and oxidizes cadmium according to

$$Cd + 1/2\,O_2 + H_2O \longrightarrow Cd(OH)_2$$

Quantities of Self-Discharge. Self-discharge in nickel – cadmium batteries depends largely on cell design. In general, large surface areas of the active material favor self-discharge, as do several other parameters (e.g., narrow electrode spacing) that reduce the internal resistance. So, heavy-duty cells are more susceptible to self-discharge (see, e.g., Fig. 26.11 in [157]).

In Figure 82 typical self-discharge curves are shown for vented nickel – cadmium batteries with pocket plates. Figure 83 shows curves for button cells (with pressed powder electrodes and comparatively high internal resistance) and cylindrical cells with sintered electrodes. Both figures illustrate the enormous influence of battery design on the self-discharge rate.

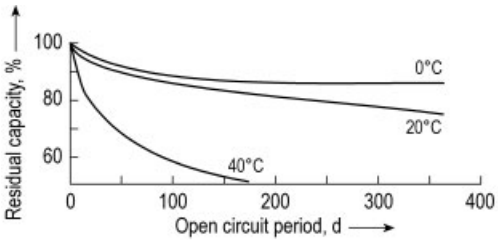

Figure 82. Charge retention of a vented nickel – cadmium battery with pocket plates at different temperatures (Vantage Battery of ALCAD Ltd. [158])

Figure 82 also shows the strong increase of self-discharge at 40 °C. This initial steep part of the curves is mainly caused by decomposition of the nickel hydroxide.

The strong influence of temperature on the self-discharge rate in both examples of Figure 83 is mainly caused by the usual acceleration of electrochemical reactions, but may also have been increased by separator hydrolysis.

The self-discharge rate can also be influenced by additives. Cobalt addition to the nickel hydroxide, for example, reduces self-discharge at elevated temperatures.

7.1.2.1.8. Types of Electrodes

Compared to the grid of lead – acid batteries, much more variety is possible in the construction of nickel – cadmium batteries, since the corrosion of nickel usually is negligible. Thus, thin layers, sintered bodies of high porosity, or foam-like structures of nickel can be used as stable substrates in the positive electrode. Nickel can also be used as conducting or supporting material in negative electrodes, despite the comparatively low hydrogen overvoltage of nickel, since the potential of the electrode is determined by the slightly positive $Cd/Cd(OH)_2$ potential, and hydrogen evolution is only possible when the negative electrode is correspondingly polarized.

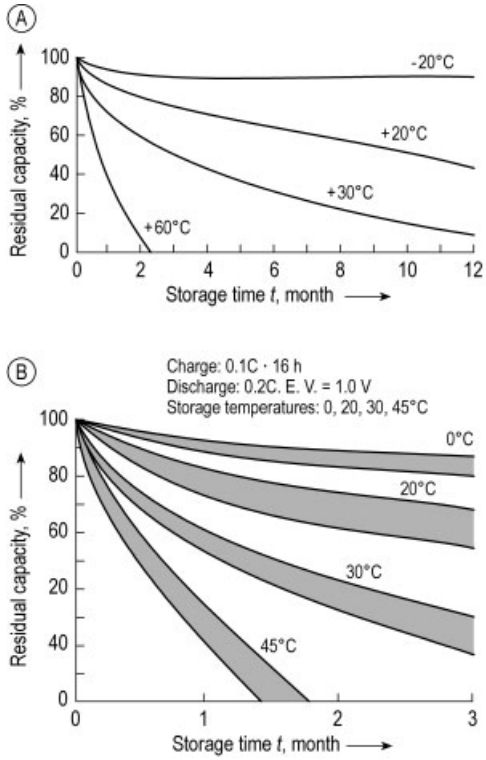

Figure 83. Capacity loss during storage of sealed nickel – cadmium batteries
A) Button cells, Varta DKZ [159]; B) Cylindrical cells, Sanyo N-Type (Fig. 4.1 in [160])

Pocket electrodes are in widespread use in vented nickel – cadmium batteries (see Fig. 84). They consist of strips of perforated nickel-plated iron that form "pockets", filled with the active material ([153], p. 59 f). A number of such strips is assembled to a flat plate. Positive and negative electrodes usually contain graphite as a conducting aid. In "genuine" sealed nickel – cadmium batteries, pocket plates are not used, but in so called "ultra-low maintenance batteries" as shown in Figure 84. The latter are flooded nickel – cadmium batteries. But they take advantage of the internal oxygen cycle, although limited to small rates, to reduce water loss. The vent cap that allows gas exchange in both directions is replaced by a valve through which gas only can escape.

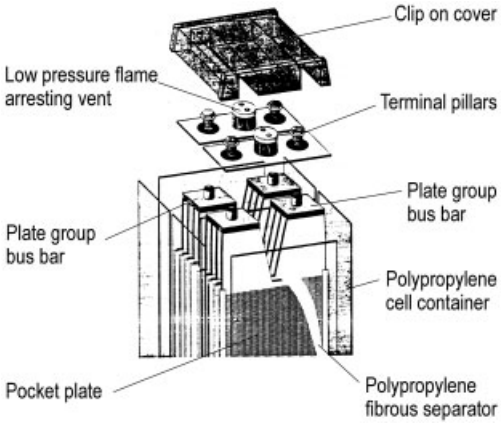

Figure 84. Schematic construction diagram of an "ultra-low-maintenance battery" (1.2 V single cell). Such batteries with limited internal recombination are available in nominal capacities between 8 and 476 Ah.
They are offered as monoblocks up to 12 V per block (AL-CAD Vantage [161])

Figure 84 shows the arrangement of the pocket plates that corresponds to the plate group in Figure 68. The valve has an opening pressure of about 2×10^4 Pa. The battery is filled with liquid electrolyte. The separator is a polypropylene fibrous material that is claimed to support the oxygen diffusion from the positive to the negative electrodes. An efficiency of the internal oxygen cycle between 85 and 95 % is observed at low float voltages and correspondingly low currents. Otherwise, the limited oxygen transport rate reduces this efficiency. Presumably a considerable share of the oxygen reduction occurs at the conducting parts that are not immersed into the electrolyte, but are in contact with the fibrous separator which is wet with electrolyte.

Pressed electrodes are mainly used in button cells (Fig. 85). Thin tablets of the active material are pressed and enclosed in a nickel-wire gauze that simultaneously provides mechanical support and acts as a current collector (in Fig. 85 sections of the wire gauze are removed from both electrodes to show the active material). The conductivity of the active material itself is improved by the addition of nickel powder. In this example, one pair of pressed electrodes forms the plate group. Stacks of several parallel electrodes are occasionally used to increase the electrode surface.

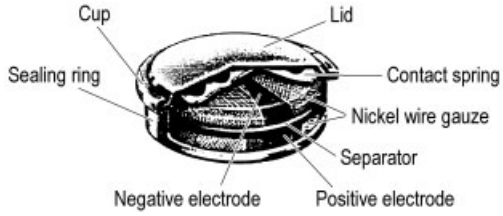

Figure 85. Construction of a button cell (VARTA)

The upper electrode in Figure 85 is the negative one; the positive electrode can be seen at the bottom. The separator is a felt disk between the electrodes. It consists of a nonwoven fiber felt of polyamide or polypropylene.

The cup of the nickel-plated steel casing acts as the positive terminal. The steel spring between the negative electrode and the cup provides the desired pressure for the separator and compensates for swelling, especially of the positive electrode. Furthermore, the spring forms void space for fast oxygen transport to the surface of the negative electrode.

Sintered electrodes are based on carbonyl nickel, an extremely fine powder produced by thermal decomposition of nickel tetracarbonyl, $Ni(CO)_4$. The powder is sintered at about 1000 °C. The result is a spongy sintered body with fine pores and a porosity of about 80 vol %.

Sintered plates can be manufactured with a thickness of up to several millimeters, but nowadays fairly thin plates are generally used with a thickness in the range of 0.5 to 1 mm. The core of the plates is a perforated, nickel or nickel-plated iron strip (0.05 to 0.1 mm thick) coated on both sides with the sintered nickel. Such electrodes normally are manufactured by a continuous sintering process [162].

The sintered plates are filled with the active material by different methods. Always, the sintered substrate is first soaked by the dissolved active material that subsequently is precipitated within the pore system. Otherwise, it would not be possible to fill the fine pores. Furthermore, a close contact between the substrate and the active material is achieved by such a precipitation. High utilization of the active material is so attained between the thin layer of active material on the large surface of the current-conducting substrate and also by the fact that the surface of

the electrode is not partly covered by the conducting material, as in pocket plates.

Chemical impregnation is used by most manufacturers. The sintered substrate is inserted into a tank, which is evacuated and then filled with nickel nitrate solution when positive active material is impregnated, or with cadmium nitrate solution when negative electrodes are processed. Temperatures between 50 and 80 °C are applied, and the sintered material is soaked completely with the solution of the nitrate, i.e., $Ni(NO_3)_2$ or $Cd(NO_3)_2$

After drying with steam, the soaked material is transferred to a tank which is filled with aqueous sodium hydroxide. When positive electrodes are processed, the impregnation of the nickel hydroxide occurs according to:

$$Ni(NO_3)_2 + 2\,NaOH \longrightarrow Ni(OH)_2 + 2\,NaNO_3$$

The negative electrode is impregnated by the following reaction:

$$Cd(NO_3)_2 + 2\,NaOH \longrightarrow Cd(OH)_2 + 2\,NaNO_3$$

After completion of these reactions the impregnated substrate is rinsed thoroughly with water and dried.

The described sequence is repeated four to eight times, depending on the type of electrode and the amount of active material required in the plate. Additives like cobalt or cadmium are added to one or several of these impregnation steps. Sintered plaques produced as single plates can be impregnated similarly.

Electrochemical impregnation is another possibility for filling the sintered substrate with active material. Electrochemical impregnation is also based on $Ni(NO_3)_2$ or $Cd(NO_3)_2$ solutions, but the precipitation is achieved by a cathodic current that changes the pH value of the solution [163].

Electrochemical impregnation has been revised and improved repeatedly [164]. It has not gained the importance of chemical impregnation, but it is used occasionally for electrodes based on fiber supports.

Thermal decomposition is a third way to fill the positive active material into the sintered substrate (cf. [153, p. 129]). The substrate is soaked with $Ni(NO_3)_2 \cdot 6\,H_2O$, which then is subjected to thermal decomposition at 180 to 250 °C. More recent variations of this method are described in [165].

Sintered electrodes are used in prismatic cells where they form plate groups as shown for a lead – acid battery in Figure 68 (see also Fig. 86). Mostly they are used in cylindrical cells as shown in Figure 88.

Foam Electrodes. Foam substrate is used for positive electrodes of high specific capacity. The basis is foamed polyurethane resin that is plated with nickel. After plating, the polyurethane is removed by burning and a nickel skeleton is left with about 95 % of void space (porosity) [166].

The active material is usually filled as a paste into this framework of nickel. The final thickness of the plate is achieved by pressing after pasting and drying.

Fiber Electrodes. Repeated efforts have been made to replace the fairly expensive sintered substrate. The *composite nickel electrode* is one result. Felts of graphitic-fiber material are coated with a layer of nickel of the thickness 0.6 to 0.8 µm. These coated fibers are sintered together and then welded to a supporting thin nickel grid (usually of expanded metal). The completed plate is impregnated by the electrochemical process described above [167]. The advantage of this electrode is that it is lighter than the sintered type.

Fiber or *felted support structures* are manufactured by precipitation of a nickel layer onto specially prepared fibers. The organic or carbon fibers are left underneath the nickel coating or are removed by thermal decomposition [168]. The porosity of fiber supports is 85 – 90 vol %.

The advantage of electrodes based on fiber substrates is that they can be manufactured with thicknesses between 0.6 and 10 mm with approximately constant performance features, because the conducting substrate increases with electrode thickness. The substrate retains also some flexibility even when the fibers are coated with nickel, and this flexibility helps when the active material swells.

The fiber substrate itself can be used as an auxiliary electrode to increase the oxygen-reduction rate [169], [170].

Figure 86 shows such a cell. The arrangement of the electrodes can be seen in the opened part: a negative electrode in the foreground is accom-

panied by an auxiliary electrode made of nonimpregnated fiber felt to improve the rate of oxygen reduction. The terminal on the right is connected with the positive electrode. The plate lugs are bolted, as usual in larger nickel – cadmium batteries of the vented type.

Figure 86. Sealed nickel – cadmium battery (1.2 V, 19 Ah nominal capacity) with fiber structure electrodes and auxiliary fiber mats for oxygen reduction (Daug/Hoppecke)

The increased oxygen-reduction rate allows fast recharge of the battery. But for thermal reasons this can only be done under controlled charging conditions.

Plastic-bonded negative plates were also developed to replace the expensive sintered electrodes. The active material is mixed with 5 to 10 wt % of polytetrafluoroethylene (PTFE) and pressed or rolled on a substrate, usually a nickel-wire net [171]. Plastic-bonded electrodes are mainly used for larger batteries in traction applications.

To achieve sufficient porosity of these active materials, graphite has to be added, which forms a fine pore system and simultaneously increases the conductivity [172]. Plastic-bonded electrodes are applied only as negative plates, since graphite would be oxidized in the positive electrode and form carbonate.

Electrodeposited (Negative) Electrodes. The cadmium electrode can also be made by galvanic precipitation of cadmium on a nickel or nickel-plated steel strip, similar to that used as substrate for the sintering process [173]. The metal strip is passed through a bath containing Cd^{2+} ions while it is connected to the negative pole of a rectifier. The positive electrode in this bath consists of metallic cadmium, which is dissolved when the current flows.

The current density has to be adjusted so that the precipitation of cadmium and evolution of hydrogen occur simultaneously and form a spongy cadmium layer, which by rolling is compressed to the desired thickness.

7.1.2.1.9. Cell Design

Prismatic cells are available in the vented and the sealed design. Examples of larger cells are shown in Figures 84 and 86. But small types are also offered, comparable to the prismatic cell in Figure 98.

Vented cells types (available in the range 7.5 to 1250 Ah) are usually equipped with pocket plates like the example shown in Figure 84. Smaller types are provided with containers of polystyrene or steel, while the larger types mostly have containers of stainless steel. The plates are spaced by nets or a number of rods, both of plastic material. Separators, as in the sealed version, are not required, since the active material is stabilized by the perforated steel stripes that form the pocket.

Batteries of this type are mainly used for stationary applications, as standby batteries or for diesel cranking. They are preferred in cold climates where lead – acid batteries are endangered to freeze, and in hot climates where lead – acid batteries suffer from increased corrosion.

For high load applications, vented nickel – cadmium batteries are also offered with sintered plates. An example are air-craft batteries as shown in Figure 87.

Air-craft batteries are designed for high load at minimum weight and volume. Figure 87 shows a typical nickel – cadmium aircraft battery. The plates are spaced by a felt of polyamide, and the distance between the plates is kept as small as possible. A membrane of cellophane prevents direct access of oxygen to the negative electrode. The limited amount of electrolyte affords frequent refilling of water, but this draw-

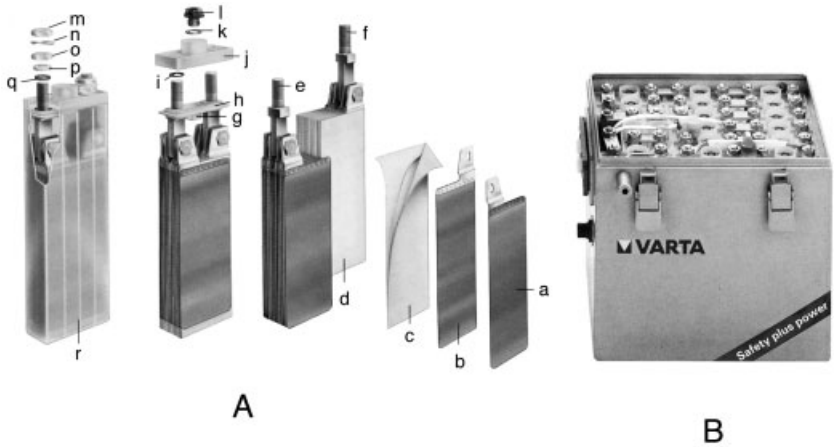

Figure 87. Nickel – cadmium air-craft battery with sintered plates
A) Exploded view of a cell
a) Negative electrode; b) Positive electrode; c) Separator; d) Plate block; e) Negative pole; f) Positive pole; g) Electrolyte level indicator; h) Baffle; i) Post sealing ring; j) Lid; k) Sealing ring; l) Vent; m) Nut; n) Spring washer; o) Spacing cap; p) Sealing ring; q) Plastic container.
B) Complete battery 24 V/40 Ah in metallic tray (20 cells)

back is outweighed by the comparatively low weight and volume.

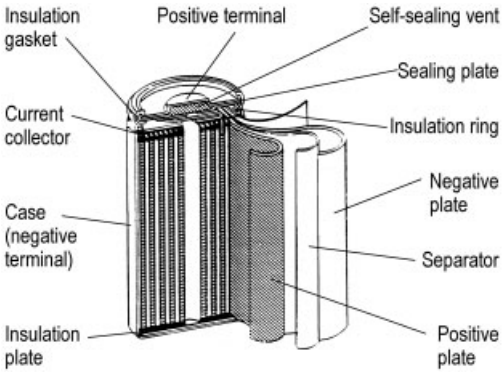

Figure 88. Cross-section of a sealed cylindrical nickel – cadmium battery (Panasonic) [174]

Cylindrical cells are widely used as sealed cells for power supply, mainly for portable devices, when capacities are required between 150 mAh and about 5 Ah. Nearly all types of electrodes, i.e., sintered, fibers, foam, and plastics electrodes are employed in this cell design.

The characteristic constructional features of cylindrical cells are shown in Figure 88.

A positive and a negative electrode are spirally wound together with the separator between them. A nickel-plated steel case encloses the electrode group and withstands the fairly high internal pressures that may occur during overcharge. The case acts as the negative terminal. A flat button in the cover is the positive terminal. The lateral face of each cell is usually covered with a shrink sleeve of plastic to prevent short-circuits.

Inside the cell, both electrodes are equipped with a number of metal strips that act as current conductors. The strips of the positive electrode are welded to the cover. The negative electrode is connected in a similar way with the bottom of the cell case.

The cell shown in Figure 88 is equipped with a resealing valve based on the compression of a rubber pad (see Section 7.1.2.1.11, especially Fig. 91).

The performance data required by the varying applications are very different and cannot be achieved by one "universal" battery design. So, most manufacturers offer several series of cylindrical sealed nickel – cadmium batteries. Such basic types are:

- *Standard types*, which represent the most economic solution.
- *High-current types*, characterized by low internal resistance, mainly achieved by a special design of the conducting elements within the cell. They are used for high-load applications like motor drives.

– *High-capacity batteries*, specially designed for high loads. They combine a maximum amount of active material with low internal resistance.
– *High-temperature batteries*, characterized by stabilized active material in the positive electrode, achieved by an increased cobalt content and by heat resistant separators. They may comprise an electrolyte that contains sodium hydroxide to reduce leakage. Emergency lighting is one field of application.
– *Fast-charge types*, suitable for recharging within one hour or less. This feature is mainly demanded for cordless tools. Fast charging is achieved with layers of active carbon which are rolled onto the surface of the negative electrode and increase the rate of oxygen reduction. But this "auxiliary electrode" requires volume that cannot be used for active material. So, maximum capacity cannot be achieved with this design.
– *Memory-backup batteries* are used for standby operation (memory backup) frequently at elevated ambient temperatures. Often they are directly soldered to printed circuit boards and therefore have to be leakproof for a long service life. Such batteries are equipped with temperature-resistant separators, special seals that prevent leakage, and often they contain a sodium hydroxide electrolyte. The advantage of the sodium hydroxide electrolyte is its reduced tendency to seep out of the seal. The main field of application is the protection of electronic memories. As a consequence of the reduced conductivity of the electrolyte, the internal resistance is somewhat higher than that of the aforementioned types.

Button cells (see Fig. 85) are in widespread use as small energy sources at moderate loads in pocket computers and similar applications.

7.1.2.1.10. Cell Assemblies, Power Packs

In general, sealed nickel – cadmium batteries are not used as single cells, because of the higher voltage usually required. For this reason, several cells are usually assembled for a battery. There are many ways of arranging the cells in a battery. Some examples of possible combinations of cylindrical cells are shown in Figure 89.

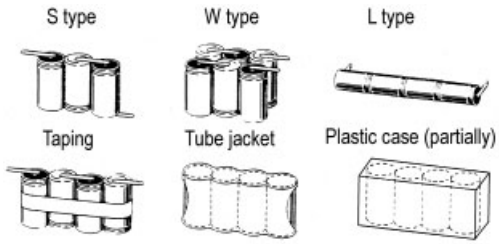

Figure 89. Examples for the connection and fixing of cylindrical cells that are assembled to multicell batteries, also called "power packs" (Sanyo Cadnica [175])

The upper line in Figure 89 shows various ways to connect cells. Contact between the individual cells is achieved by spot welding with corresponding tags. The battery on the right-hand side, that is formed by stacking the cells, is mainly used for smaller cells. Sometimes two or more strings form a slim battery that can be placed in the handle of tools. To avoid short circuits, the battery usually is covered completely by insulating plastic material, except for the soldering tags.

Mostly, power packs are completely encapsulated by plastic material, as shown on the right-hand side in the lower line of Figure 89. Such packs are offered in a wide variety since a battery often is tailored to its application. The fast growing market for cordless telephones, tools, camcorders, and toys caused a nearly indeterminable variety of batteries.

The cells that are combined in a battery pack usually are matched for capacity. Otherwise, individual cells may suffer deep discharge or inadmissible overcharge. This is true especially for batteries operating in rigorous conditions, e.g., in tools. In regard to such a matching, the customer should not combine single cells in batteries himself, especially in the case of heavy-duty batteries.

Button cells are also combined to batteries. Often they are stacked and enclosed by a shrink sleeve, but the are also available encapsulated by plastic containers as batteries of different shape.

7.1.2.1.11. Valve Design

Nickel – cadmium batteries can be sealed completely. In theory, this battery system should

need no valve. Nevertheless, for safety reasons, also nickel – cadmium batteries usually are equipped with valves to prevent explosion from unexpected increase of internal pressure, usually caused by abuse of the battery, e.g., during overcharge with a too high current.

Two kinds of valve are used:

- Genuine "safety vents" destroyed in an emergency,
- Resealing valves, which open at a certain overpressure but close again when the internal pressure drops below the critical value.

An example of the first kind is shown in Figure 90. The sealing element is a metallic membrane which bulges upwards with increasing internal pressure (diaphragm in the figure). A pointed piece of metal is bent down from the outer cell cover, and its point pierces the membrane when its bulge is excessive.

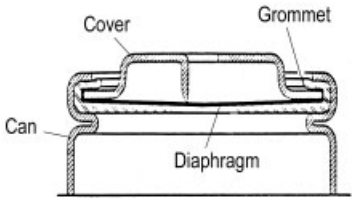

Figure 90. Safety vent in a cylindrical sealed nickel – cadmium battery ([153, Fig. 2.59, p. 212], reprinted by permission of The Electrochemical Society, Inc.)

Safety vents as in Figure 90 were used mainly during the first decades of the manufacture of sealed nickel – cadmium batteries. Absolutely tight sealing can be achieved with this vent. But as soon as the membrane is pierced, the sealing is completely destroyed. Then the cell operates like a vented all, which means that water loss ends the service life within a short time. Thus, an occasional overcharge might spoil the whole battery.

For this reason, safety vents as in Figure 90 have been largely replaced by resealing valves. Today, the safety vent, as described in Figure 90, is used only with batteries that have to be sealed absolutely, e.g., for space application or with button cells where a point of fracture in the cup acts as a safety vent.

Improvement of elastomers (synthetic rubber), enabled the development of simple and also reliable resealing valves. One example of various similar designs is shown in Figure 91.

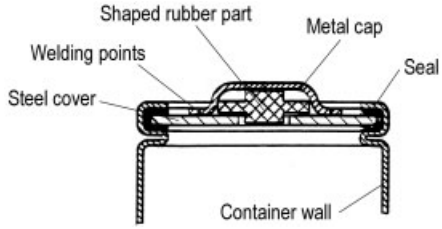

Figure 91. Resealing valve for nickel – cadmium batteries based on a compressed rubber part

The shaped rubber part is pressed on a seat in a metallic cap that is welded on top of the cell cover. This cap is welded only at certain points, and slots are left for the gas to escape between the cell cover and the metallic cap. When the internal pressure exceeds the critical value, the rubber part is compressed so much that a slot is formed between the steel cover and the rubber, and the gas can escape.

The advantage of the resealing valve is obvious: occasional abuse of the battery causes some water loss; but, if this loss is limited, it will not influence the performance significantly. The battery will operate as before and can reach the usual service time. Hydrogen that has been evolved will even be gradually removed by oxidation at the nickel surface in the positive electrode.

A disadvantage of the resealing valve is that it cannot be sealed as perfectly as the safety vent in Figure 90. However, for usual battery applications the advantages outweigh the disadvantages. Especially in batteries composed of a larger number of cells, occasional overcharging or overdischarging (reversal) of single cells often cannot be excluded. In such cases, the battery as a whole is better protected by resealing valves, when the damage of the concerned cells does not exceed certain limits.

7.1.2.1.12. Controlled Charging

A simple charging method for sealed nickel – cadmium batteries is shown in Figure 81. But it takes about 14 h for complete recharge, and for many applications, e.g., in tools, a short charging period is desired. Acceleration of the charging process can be achieved by an increased charging current. But inadmissible overcharging has to be avoided, which means that the charging current has to be reduced as soon as the cell reaches a certain stage

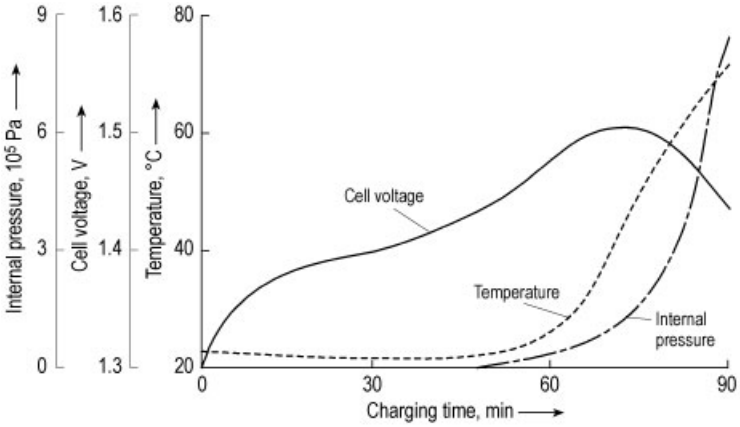

Figure 92. Cell voltage, cell temperature, and internal pressure of a sealed nickel – cadmium battery versus charging time during recharge at 1 C(A) (1-hour rate) [176]

of recharge. So additional parameters are required to determine this critical stage of the charging process. Such parameters are:

- *Amount of charge*, measured directly or calculated as the product of charging current and charging time. One simple method that uses this parameter, is to discharge the battery completely before the charging process starts. Subsequently the amount of current is charged that is required for a complete charge.
- *Cell temperature* increases significantly when the internal oxygen cycle has started. Temperature sensors can be used to terminate the charging process.
- *Cell voltage* is an often used parameter. Cell voltage decay at a constant current is caused by reduced polarization of the positive electrode due to increased heat generation when the internal oxygen cycle starts. The voltage decay thus indicates that the positive electrode is fully charged. If the charging current is not constant, the varying voltage drop has to be considered, e.g., by measuring the cell voltage during periods of current interruption.
- *Internal pressure* also reflects oxygen evolution at the positive electrode. Its measurement is, however, rather expensive and hardly applied in practice.

The behavior of the three parameters: cell voltage, cell temperature, and internal pressure during constant-current charge is shown in Figure 92. All three parameters are plotted versus charging time during recharge at 1 C(A) ($\approx$ 1-hour rate).

The cell voltage increases with charging time and reaches a maximum shortly after about 100 % of the nominal capacity has been restored. But the battery is not recharged completely at this point, because a portion of the current has been used for the internal oxygen cycle. The subsequent decline of the cell voltage is caused by the rising cell temperature, mainly on account of the increasing current share that is consumed by the internal oxygen cycle. This increasing temperature accelerates the oxygen-evolution reaction at the positive electrode and thereby reduces its polarization. The resulting voltage decay ΔU is often used to terminate the charging process as indicated in Figure 93.

This voltage decay would be dangerous if the battery were charged at constant voltage. The more and more reduced polarization would continuously increase the current and so result in a further increasing temperature. If no other parameter terminates the charging process, this situation would finally lead to thermal runaway.

The temperature curve in Figure 92 shows a slight decrease during the first stage of the charging process, caused by the negative reversible heat effect. With the onset of the internal oxygen cycle, heat production is increased corresponding to Figure 81 and the cell temperature rises.

The internal pressure is related to the balance between the oxygen-evolution rate at the positive electrode and the oxygen-reduction rate at the negative electrode and mainly determined by the transport rate of oxygen.

Fig. 93 shows a number of parameters that are used in controlled charging methods.

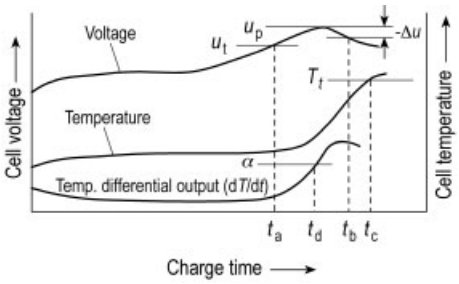

Figure 93. Charging characteristics that are used in controlled charging methods for nickel – cadmium and nickel – metal hydride batteries [177]

Cell temperature: at the time t_a the voltage U_t is reached that is below the maximum voltage U_p and even below the voltage that finally would be reached. By that time, about 70 % of recharge is achieved, and the thermal situation of the cell is still stable. Ut is often used to reduce the high initial charging current so far that thermal risks are excluded during the final charging period.

Cell voltage decay: the cell voltage decay ΔV that is reached at the time t_b is frequently used to terminate the charging process.

Turning point α of the differential of the temperature curve dT/dt: this turning point, reached at the time t_d, is caused by the increased heat dissipation due to the increase of the cell temperature. It is used to reduce the charging current or switch to a short additional charging period.

Cell temperature: The cell temperature Tt that is observed at the time t_c is mainly used as an additional parameter to cut the charging current when other controlling mechanism have failed. This parameter is not suited for direct control of the charging, since it is observed too late.

7.1.2.2. Nickel – Hydrogen Batteries

Nickel – hydrogen batteries, known also as hydrogen – nickel oxide batteries, have been developed for aerospace applications and are still the number one energy storage system in many satellite projects [178]. This system employs the same positive electrode and also the same electrolyte as the nickel – cadmium battery. But for the negative electrode hydrogen is used as active material instead of cadmium. Furthermore, the negative electrode is different from a conventional battery electrode and corresponds to a fuel-cell electrode. The electrode material is not transformed, but acts only as a catalytic surface that provides or absorbs the electrons.

The catalytic effect of the negative electrode is not a pure surface effect, it rather includes absorption and the formation of metal hydrides. But, the amount of platinum or palladium, used in such batteries, is so small that the storage capacity of the electrode has no relevance.

Hydrogen is not stored in the electrode. In the charged state of the battery, it remains as gaseous hydrogen within the cell. The container therefore has to withstand pressures up to about 50 bar. On account of the hydrogen storage, nickel – hydrogen batteries must be hermetically sealed.

The discharge reaction of the negative electrode is:

$$H_2 + 2\,OH^- \longrightarrow 2\,H_2O + 2\,e^-$$

Its reversal (from right to left) represents the charge.

The generation of gaseous hydrogen during charging and its consumption during discharging means that the internal gas pressure is proportional to the state of charge. This offers the possibility to check the state of charge by measurement of the internal cell pressure.

Usually platinum is applied as electrode material on account of its outstanding catalytic capabilities and its chemical stability. Platinum is usually bonded by polytetrafluoroethylene (Teflon) or covered with microporous Teflon membranes to make the surface hydrophobic and so prevent the formation of thicker layers of water that would hinder gas access. The electrolyte is immobilized by soaking up in fiber felts or knitted textiles of mineral fibers (fuel-cell grade asbestos paper) or zirconia fibers stabilized with yttria (Zircar cloth). They provide high temperature resistance and chemical resistance for long-life batteries.

Combination with the reaction of the nickel electrode (Section 7.1.2.1.2) leads to the cell reaction (discharge from left to right)

$$2\,NiOOH + H_2 \longleftrightarrow 2\,Ni(OH)_2$$

During discharge, NiOOH is reduced to $Ni(OH)_2$ and hydrogen is consumed. During charge, this equation is reversed. Hydrogen is stored in each cell as gas. Figure 94 shows a sectional view.

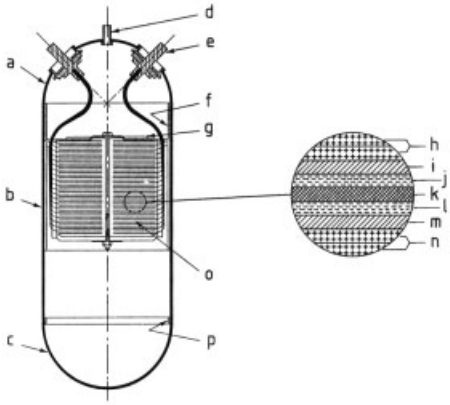

Figure 94. Construction of a nickel oxide – hydrogen cell, showing details of the cell construction
a) Container top hemisphere; b) Container cylindrical section; c) Container bottom hemisphere; d) Filling tube; e) Terminal; f) Welding ring; g) Pressure plate; h) Positive electrode; i) Separator; j) Negative electrode; k) Gas diffusion screen; l) Negative electrode; m) Separator; n) Positive electrode; o) Electrode stack; p) Welding ring

Thermodynamic Situation, Secondary Reactions. The thermodynamic situation of the nickel – hydrogen battery is shown in Figure 95. It is similar to that in the nickel – cadmium battery shown in Figure 78. The positive electrode is identical, while at the negative electrode the charging reaction is hydrogen evolution (H^+ oxidation):

$$2\,H_2O + 2\,e^- \longrightarrow H_2 + 2\,OH^-$$

and the charging reaction its reversal:

$$H_2 + 2\,OH^- \longrightarrow 2\,H_2O + 2\,e^-$$

The equilibrium potential of the hydrogen electrode is only 0.02 V lower than that of the cadmium electrode. Thus, the cell voltage is $U^o = 1.34$ V, in practice $U^o \approx 1.3$ V.

Oxygen evolution, oxygen reduction, and the internal oxygen cycle occur as in nickel – cadmium batteries. If no current flows through the cell, any oxygen that had been generated within the cell is removed by direct catalytic recombination of oxygen and hydrogen, because at the negative electrode a mixed potential is formed between hydrogen oxidation ($H_2 \longrightarrow 2\,H^+ + 2\,e^-$) and oxygen reduction.

Hydrogen oxidation is the discharge reaction at the negative electrode. But due to its high pressure, hydrogen is also oxidized at a perceivable rate at the positive electrode as an unwanted secondary reaction, that reduces the efficiency of the charging reaction and acts as a parasitic discharging reaction during discharge. At open circuit, reduction of nickel hydroxide is the counter reaction; and self discharge according to

$$2\,NiOOH + H_2 \longrightarrow 2\,Ni(OH)_2$$

is the result. This direct reaction between H_2 and Ni(OOH) causes the considerable self-discharge rate of nickel – hydrogen batteries. An *internal hydrogen cycle* is established when the battery is discharged so deeply that hydrogen is evolved at the reversed nickel hydroxide electrode. This hydrogen subsequently is oxidized at the negative electrode according to the normal discharge reaction: So, the negative electrode is not reversed, but remains at its normal discharge potential, while H_2 is now supplied by the reversed positive electrode. This internal hydrogen cycle allows deep discharges that include the reversal of individual cells.

Advantages of the nickel – hydrogen battery system are:

- A fairly high specific energy of 50 to 60 Wh/kg depending on design and size.
- The ability to withstand a great number of deep discharges (40 000 cycles at 40 % DOD).
- A service life of 15 years.
- The battery can be deeply discharged and single cells reversed without damage.
- The state of charge can easily be controlled by measurement of the H_2 pressure.
- Each cell is hermetically sealed. No maintenance is required.

Disadvantages are:

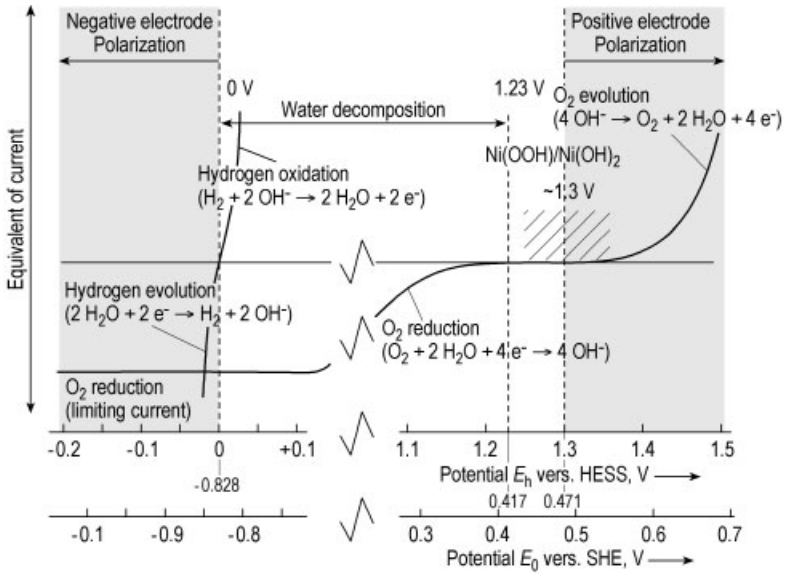

Figure 95. Equilibrium potentials of the negative and positive electrodes in nickel – hydrogen and nickel – metal hydride batteries, and current – voltage curves for O_2 and H_2 evolution, O_2 reduction, and H_2 oxidation.
For the x-axis, the electrode potential, two scales are employed: The upper one represents the potential E_h referred to the hydrogen electrode in the same solution (HESS), the lower scale E_o referred to the standard hydrogen electrode (SHE) for 1.250 g/cm^3 density of the KOH solution at 25 °C. (cf. Fig. 72)

– Its high price which limits the application more or less to aerospace projects. (More recently, efforts have been made [179] to reduce the price and develop a nickel – hydrogen battery also for remote terrestrial applications like photo-voltaic stations).
– High self-discharge rate of about 5 % per day at room temperature, due to direct reaction between hydrogen and nickel.
– The energy density amounts to only 20 to 40 Wh/L, because of the volume required for gas storage and the shape of the pressure vessel.

The above listed advantages explain the widespread use in the aerospace field.

Low-Pressure Nickel – Hydrogen Battery. A modified version of this system is the low-pressure nickel – hydrogen Battery. It has in principle the same design as the normal nickel – hydrogen battery, but hydrogen is stored as a metal hydride in alloys as they are used in nickel – metal hydride batteries. The advantages are that the container must not withstand high pressure and the rate of self-discharge is lower. On the other hand, the alloy contributes additional weight.

7.1.2.3. Nickel – Metal Hydride Batteries

The nickel – metal hydride battery is based on the same reactions as the nickel – hydrogen battery. But in nickel – metal hydride batteries, special alloys act as the catalytic electrode surface and simultaneously absorb H_2 by forming metal hydrides under low pressure. Thus the negative electrode material catalyzes the reaction and simultaneously stores the hydrogen that is formed during charging.

7.1.2.3.1. Thermodynamic Parameters

For the cell reaction

$$2\,(\beta\text{-NiOOH} \cdot H_2O) + H_2 \longleftrightarrow 2\,Ni(OH)_2 + 2\,H_2O$$

the free enthalpy of reaction ΔG is -259 kJ and the reversible heat effect $Q_{rev} = T\Delta S = -36\,kJ$. Q_{rev} is very close to that of nickel – cadmium batteries (Section 7.1.2.1.4).

But here, heat effects have also to be considered that are caused by hydrogen absorption and desorption in the material of the negative electrode. Hydrogen absorption is an exothermic reaction and supplies additional heat during charging. The quantity of heat depends on the alloy applied for hydrogen storage. It might compensate or even over-compensate the reversible heat effect. For this reason, the characteristic feature of nickel – cadmium and nickel – hydrogen batteries that the charging reaction causes cooling of the battery, may not apply to nickel – metal hydride batteries, (see, e.g., [180]).

The equilibrium voltage U^{o} is the same as that of the nickel – hydrogen battery, $U^{o} \approx 1.32$ V, and also its temperature coefficient $U^{o}/dT \approx 0.6$ mV/K. Both values almost correspond to those of the nickel – cadmium battery system.

The fairly high value for the specific energy of 387 Wh/kg effected by the cell reaction, is markedly reduced by the required weight of the storing alloy. So this quantity is lowered to about 300 Wh/kg as shown in Table 13.

7.1.2.3.2. Hydrogen-Absorbing Alloys

(→ Hydrogen, Chap. 9.3.1.2.; → Hydrides, Chap. 5.)

Hydrogen can be stored at a fairly low pressure when it is absorbed by metals which form hydrides. Palladium, for example, absorbs between 0.4 and 2 wt % of hydrogen. This means that 1 cm^3 palladium is able to store up to 1200 cm^3 of gaseous H_2 at 25 °C and 1 bar. This demonstrates the enormous hydrogen storage capability of such metal hydrides.

Less noble metals are also able to form hydrides and to store considerable amounts of hydrogen, and a great variety of such elemental hydrides exists. But most of them are of little practical use, since only a few hydrides offer a reasonable hydrogen equilibrium pressure at room temperature. This situation changed when intermetallic compounds were developed that combine elements that form strong hydrides with those that form weak hydrides. Such intermetallic compounds are alloys of two or more metallic elements, having narrow integer stoichiometries. The number of possible intermetallic compounds is nearly unlimited, and a great number of alloys have been discovered or developed that form metal hydrides [181], providing hydrogen storage in a suitable range of pressure and temperature (surveys are given in [182]). The intermetallic compounds are divided into groups $A_{\mu}B_{\nu}$, based on composition and crystal structure. A represents a strong hydride forming element and B the weak hydride forming component. For practical purposes A and B often are combinations of various elements as shown in Table 17.

Combinations that proved to be suited for hydrogen storage in batteries are known as AB, AB_2, A_2B and AB_5 (see, e.g., [184]). Table 17 shows their characteristic composition and some features.

AB_5 alloys are nowadays mainly used for batteries. Mischmetals are mostly substituted for lanthanum, mainly because they are less expensive than pure La. But such substitutions also are required to stabilize the alloy and prevent its premature degradation.

The AB_2 type in row 2 of Table 17 is based on vanadium, titanium, zirconium, and nickel. Especially for these intermetallic compounds the stoichiometry is not well defined. So, some alloys ascribed to this group may be named A_2B or AB as well. They often have the crystallographic structure of Laves phases and are called Zr – Ni – V – Cr multicomponent alloys (e.g., the *Ovonic alloys* [187]).

Titanium – nickel intermetallic phases (A_2B and AB type) were developed in the 1970s with great assiduity for hydrogen storage in road vehicles with hydrogen combustion engines [191].

Figure 96 shows pressure – composition isotherms of an AB_5 alloy ($LaNi_{4.7}Al_{0.3}$). For each temperature the upper and lower curves denote absorption and desorption, respectively. Initially hydrogen is absorbed with increasing pressure, until the formation of the MH_n phase starts. Then the pressure remains approximately constant, until the metallic phase is converted into hydride. Desorption of hydrogen occurs at lower pressure and the hysteresis between absorption and desorption pressure indicates the energy loss between absorption and desorption. The horizontal axis is given in moles hydrogen per mole metal. So the ratio 1 means 6 H per alloy formula, or, with the molecular masses of La, Ni, and Al = 138, 59, and 27 respectively, 6 g hydrogen per 423 g of alloy. This gives a storage

Table 17. Classes of intermetallic compounds that are used for negative electrodes of nickel – hydride batteries; alloys corresponding to rows 1 and 2 are nowadays mainly applied

	A_uB_v Class (basis)	Actual components	Storage capability, Ah/kg	Remarks
1	AB_5 ($LaNi_5$)	A: mischmetal ∗, La, Ce, Ti B: Ni, Co, Mn, Al	≈ 300 at max. (e.g., [185])	at present, mostly used alloy group
2	AB_2 ($TiNi_2$)	A: V, Ti B: Zr, Ni (+ Cr, Co, Fe, Mn)	≈ 400 at max. (e.g., [186])	basis of "multi-component alloys" [187]
3	AB (ZrNi)	A: Zr, Ti B: Ni, Fe, Cr, V		early used for hydrogen storage in cars and as electrodes [188], [189]
4	A_2B (Ti_2Ni)	A: Mg, Ti B: Ni		

∗ Mischmetal (Mm) is an unrefined rare earth mixture (mainly Ce, La, Nd, and Pr) (→ Cerium Mischmetal, Cerium Alloys, and Cerium Compounds). Its composition depends on the ore [184], but "synthetic mischmetals" are also known [190].

capability of about 1.4 wt %. In electrochemical equivalents 6 g of H_2 mean 3 times 53.6 Ah or 160.8 Ah per 423 g which equals 380 Ah/kg that are stored in this case.

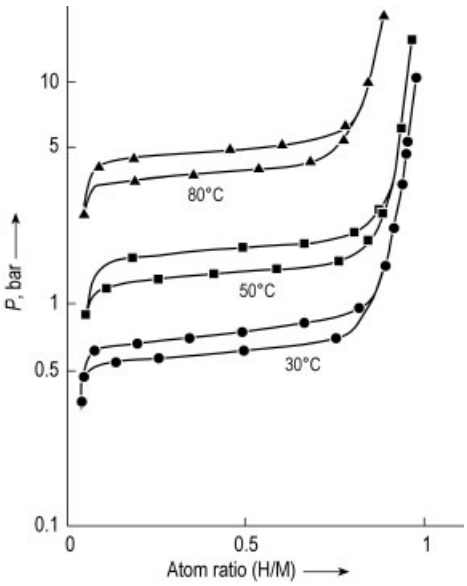

Figure 96. Pressure – composition isotherms (PCI curves) of an AB_5 alloy with nickel partly substituted by aluminum ($LaNi_{4.7}Al_{0.3}$) (Fig. 2.18 in [192])

The hydrogen storing alloys are usually not manufactured in battery plants, but are provided by specialized suppliers (some references are given in Section 2.5 of [192]). Selection of the raw materials and preparation of the hydrogen absorbing alloy are most important, because not only the composition of the alloy but also its structure and the distribution of the alloying elements are decisive for capacity and service life [193].

In the battery, the hydrogen storing alloy is in direct contact with the electrolyte and also acts as the electrode. So the interface layer has to resist corrosion and oxidation attacks and may undergo significant structural and compositional changes [194]. This layer plays an important role: While the bulk of the electrode material determines the capability for hydrogen absorption and so the capacity, the interface layer is decisive for most performance data, e.g., high rate discharge capability and cycle life.

The performance characteristics of nickel – metal hydride batteries are very similar to those of sealed nickel – cadmium batteries. The cell voltage differs by only 20 mV, and charging as well as discharging performance are so alike that both battery systems can replace each other in all normal applications. The discharge curves in Figure 97 confirm this: they differ from each other only by the discharge duration.

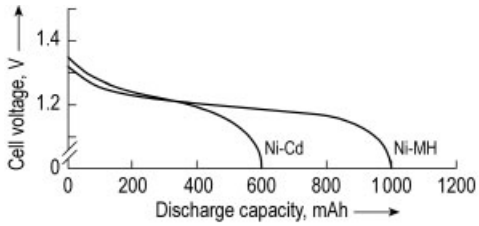

Figure 97. Comparison of discharge voltage and capacity between nickel – cadmium and nickel – metal hydride batteries of the same size [196]
Cylindrical cells, AA type (Mignon cell), discharging current 0.2 CA (≈ 5 h)

In general, metal-hydride electrodes are inferior to cadmium electrodes at extremely high rates and low temperatures, because of the reduced rate of hydrogen desorption. In prac-

tice, however, discharge performance under these conditions is mainly a question of electrode thickness. Thus, such a comparison depends largely on the special battery design, and nickel – metal hydride batteries with thin electrodes can give a better performance at low temperatures than nickel – cadmium batteries with thicker electrodes.

Self-discharge in nickel – metal hydride batteries is based on hydrogen oxidation as has been described for nickel – hydrogen batteries (Section 7.1.2.2). But, due to the low hydrogen equilibrium pressure, the rate of self-discharge is reduced so far that it only attains rates known for nickel – cadmium batteries.

7.1.2.3.3. Design

Due to similar cell voltage and discharge/charge behavior, nickel – metal hydride batteries can replace nickel – cadmium batteries. For this reason, they are manufactured in the same dimensions. Button, cylindrical, and prismatic cells are offered, but also battery or power packs and "smart batteries".

For the positive electrode, usually foam electrodes (see Section 7.1.2.1.8) are applied, to balance the comparatively large capacity of the hydrogen storage electrode.

Constructional details of cylindrical and small prismatic nickel – metal hydride batteries are shown in Figure 98. The arrangement of the electrodes in cylindrical cells is similar to that of cylindrical lead – acid and nickel – cadmium batteries (Figs. 70, and 88 respectively). In the prismatic cells, a number of positive and negative electrodes are paralleled and form the plate group as in Fig. 68.

The electrode assemblies are inserted into a nickel-plated steel can. The can acts as the negative pole, while the lid is the positive terminal, both are insulated by a plastic gasket. The cell is sealed by crimping the lid to the can. Perfect sealing is of great importance because of the volatility of hydrogen. The lid contains a resealable valve similar to that described in Figure 91. As in the case of nickel – cadmium batteries, this valve opens only in an emergency to prevent inadmissible pressure increase. In normal operational conditions the nickel – metal hydride battery remains completely sealed, and hydrogen does not escape.

Due to their high specific energy content and their capability of high discharge rates and rapid charging, nickel – metal hydride batteries are promising candidates as storage systems for hybrid electric vehicles that combine an ignition engine with electric drive. Longer distances are covered by the gasoline engine, in the cities the vehicle is driven by battery power. This requires a comparatively small battery that can be heavily loaded. Figure 99 shows such a module. The fan belongs to the air cooling system.

The expected cycle life of these batteries is $\geq$ 2000 cycles (100 % DOD).

Nickel – metal hydride batteries designed for extremely high load may also be used in future motor cars to supply peak power for cranking and electrically driven accessories [195].

7.1.2.4. Nickel – Iron Batteries

The nickel – iron battery was developed in parallel to the nickel – cadmium battery at the beginning of the 20th century (see, e.g., [153, pp. 22]). It is based on the same electrolyte and the same (pocket-type) positive electrode. However, the active material of the negative electrode is iron.

Iron ions exist in two oxidation states as Fe^{2+} and Fe^{3+} ions. As a consequence two steps of electrode reaction are possible. The oxidation (discharge) of Fe to Fe^{2+} is described by:

$$Fe + 2\,OH^- \longrightarrow Fe(OH)_2 + 2\,e^-$$

with the standard electrode potential (in alkaline solution at pH 14) $U^{o,s} = -0.89$ V and the further oxidation of Fe^{2+} to Fe^{3+} occurs according to

$$Fe(OH)_2 + OH^- \longrightarrow Fe(OH)_3 + e^-$$

with the standard potential (in alkaline solution at pH 14) $U^{o,s} = -0.56$ V. The corresponding standard value of the hydrogen electrode

$$2\,H_2O + 2\,e^- \longleftrightarrow 2\,OH^- + H_2$$

is for comparison $U^{o,s}(H_2/H^+) = -0.828$ V. Thus the Fe/Fe^{2+} potential is 60 mV below the equilibrium potential of the hydrogen electrode,

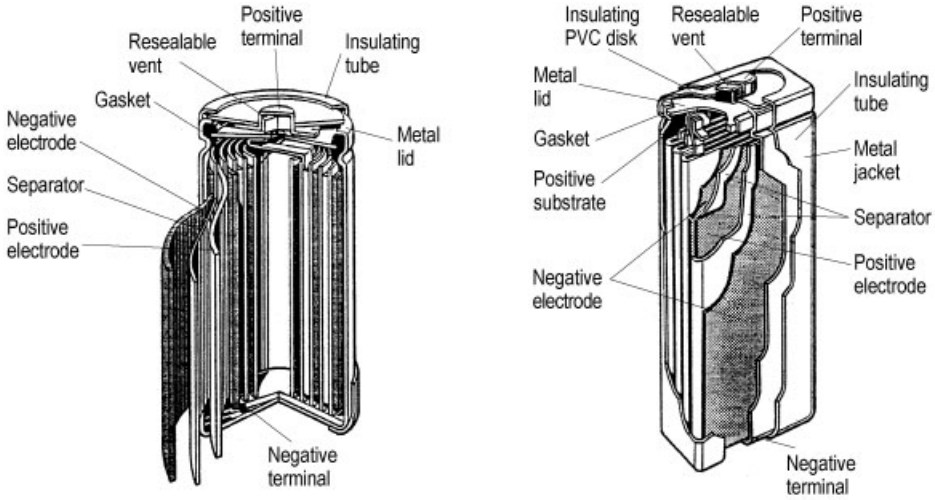

Figure 98. Cross sections of cylindrical and prismatic nickel – metal hydride batteries (VARTA) [197]

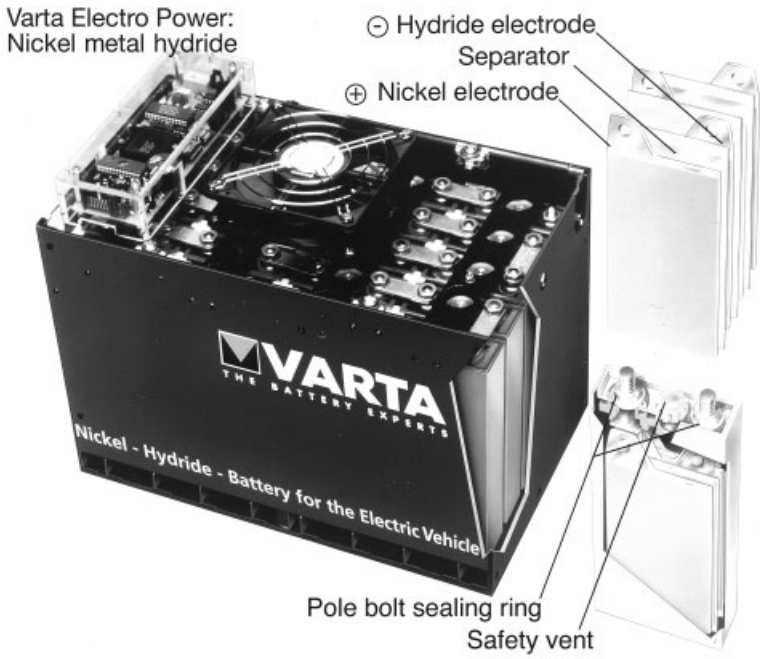

Figure 99. Nickel – metal hydride battery; module for electric vehicle applications with battery management system and air cooling [198]

while the Fe^{2+}/Fe^{3+} potential is 268 mV positive compared to the H_2/H^+ equilibrium potential.

The normal charge/discharge cycle uses only the Fe/Fe^{2+} step. Fe^{3+} ions are only formed at extremely deep discharges or cell reversal. Thus under normal cycling condition the (simplified) cell reaction (see Section 7.1.2.1.2) is (left to right = discharge)

$$2\,NiOOH + Fe + H_2O \longleftrightarrow 2\,Ni(OH)_2 + Fe(OH)_2$$

and the open-circuit cell voltage $U^{o,s} = 1.38$ V, since the potential of the nickel hydroxide elec-

trode in alkaline solution is 0.49 V. The iron hydroxide forms together with OH^- ions complex compounds ($Fe(OH)_n^{2-n}$) of extremely low solubility. Thus, precipitation of the reaction product occurs within the pore system of the active material and the physical structure of the electrode remains almost stable (see Section 4.1.1.2). This situation is similar to that of the lead – acid battery and it is the main reason why this reaction can be reversed very often and that nickel – iron batteries can reach 4000 discharge/charge cycles.

At deep discharges, the final product is Fe_3O_4, that contains two Fe^{3+} ions and one Fe^{2+} ion:

$$8\,NiOOH + 3\,Fe + 4\,H_2O \longleftrightarrow 8\,Ni(OH)_2 + Fe_3O_4$$

The general situation of the nickel – iron battery is similar to that in Figure 78, the only difference is that the equilibrium potential of the negative electrode is 60 mV below the H_2/H^+-potential, and so hydrogen evolution occurs already at the open-circuit potential. This is a decisive disadvantage of the nickel – iron battery compared to the nickel – cadmium battery, and this all the more since the hydrogen overvoltage is low at the iron electrode. This means high gassing rates, a corresponding water loss, and high self-discharge rates that, at open circuit, discharge the battery completely within about three months.

Charging at too high rates has to be avoided, otherwise the increased polarization of the iron electrode causes such a high hydrogen evolution rate that the electrolyte is sprayed though the vent caps. To keep the battery fully charged in a standby application; a float current of about 0.5 A/100 Ah is required.

The discharge/charge characteristic shown in Figure 100 is similar to that of nickel – cadmium batteries. The slope of the discharge curve is steeper than that of nickel – cadmium batteries, since the iron electrode has a less stable discharge potential. At low temperatures a considerable decrease of the available capacity is observed, caused by the lowered reaction rates at the iron electrode.

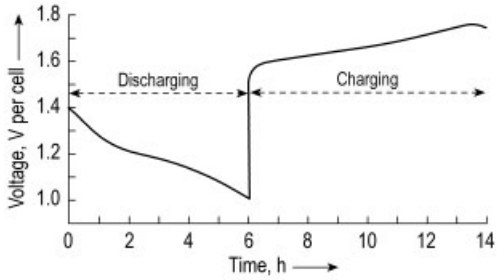

Figure 100. Typical discharging – charging characteristic of a nickel – iron battery when charged at constant current. The positive electrode is capacity limiting so that the negative electrode is not deeply discharged.

Nickel – iron batteries were mainly used for traction applications due to their mechanical stability and the high cycle numbers that can be obtained. They were also employed for signal system since they can withstand abuse. In the western hemisphere, however, they completely lost their market in the 1970s to lead – acid or nickel – cadmium batteries. They are still in use in Eastern Europe, especially in the countries of the former Soviet Union.

The search for batteries, suited for electric vehicles caused a certain renaissance of this system, since it would be an inexpensive battery with high cycle capability. So, quite a number of projects were started to improve this battery system by a new design (cf., e.g., [175, p. 243 f]).

7.1.2.5. Nickel – Zinc Batteries

The nickel – zinc battery is a combination of the nickel electrode and a zinc electrode in an alkaline electrolyte. It is also closely related to the nickel – cadmium battery system. The oxidation reaction (discharge) of the zinc electrode is

$$Zn + 2\,OH^- \longrightarrow Zn(OH)_2 + 2\,e^-$$

with the standard electrode potential (in alkaline solution at pH 14) $U^{o,s} = -1.245$ V. Combination with the nickel hydroxide electrode gives the cell reaction

$$Zn + 2\,NiOOH + 2\,H_2O \longleftrightarrow 2\,Ni(OH)_2 + Zn(OH)_2$$

(left to right = discharge) with the cell voltage $U^o = -1.73$ V, since the potential of the nickel hydroxide electrode is 0.49 V (at pH 14). The

zinc electrode usually is made by pressing or rolling zinc oxide together with a binder and graphite as conducting aid. The result is the uncharged electrode that then is converted into a porous zinc material by charging (electrochemical reduction).

Zinc is an interesting electrode material. Its atomic mass of 63.4 g/mol is about half that of cadmium, and its more negative potential means a higher cell voltage and a correspondingly increased storage capability. The hydrogen overvoltage of zinc is high, and although the potential of a zinc electrode is 0.42 V negative compared to the hydrogen electrode, the hydrogen evolution rate is low as long as the zinc is not contaminated by foreign metals. Furthermore, for a technical product it is important that zinc is an inexpensive metal.

The problem of the zinc electrode is its solubility. This causes shape changing when discharging/charging cycles are repeated as indicated in Figure 4. Furthermore, zinc dendrites can be formed by precipitation of the dissolved material. They can penetrate the separator and form short circuits. To reduce this problem, various separators or combinations of separators have been tried (cf. Table 29.3 in [157]).

The design of a sealed nickel – zinc battery is possible. As in nickel – cadmium batteries, the negative electrode must have a "charge reserve" of uncharged material to avoid polarization that causes increased hydrogen evolution at the end of the charging process. The possible overcharging rate is limited in nickel – zinc batteries, on account of the required more dense separation between the electrodes to avoid the mentioned penetration by zinc dendrites. Figure 101 shows a typical charge/discharge curve.

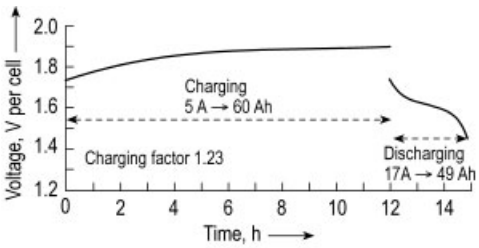

Figure 101. Typical constant current charge – discharge curve of a sealed nickel – zinc battery

The general problem of nickel – zinc batteries is the limited number of possible cycles. At 100 % DOD, the capacity is reduced to 60 % within less than 300 cycles, but even at 30 % DOD only about 400 cycles are reached before the capacity drops below 60 %.

7.1.2.6. Silver Oxide – Zinc Batteries (cf. Section 6.1.3.2)

The silver oxide electrode as the positive one represents in combination with the zinc electrode the highest possible storage capability per weight and volume of the aqueous systems. Due to the existence of two positive oxidation states of silver (Ag^+ and Ag^{2+}) two discharge/charge reactions are possible. The discharge of divalent Ag^{2+} (its reduction) is described by

$$2\,AgO + H_2O + 2\,e^- \longrightarrow Ag_2O + 2\,OH^-$$

with the standard electrode potential (in alkaline solution at pH 14) $U^{o,s} = 0.599$ V; while the lower discharge plateau is given by the conversion of Ag^+ into Ag and the following reaction:

$$Ag_2O + H_2O + 2\,e^- \longrightarrow 2\,Ag + 2\,OH^-$$

with the standard electrode potential (in alkaline solution at pH 14) $U^{o,s} = 0.344$ V. In the battery, however, the discharge also of Ag^{2+} mainly occurs at the lower plateau, especially at increased load (cf. Fig. 102). When the electrode is recharged, complete conversion of Ag to the Ag^{2+} state is not reached.

The most common fabrication technique for silver oxide electrodes is by sintering silver powder at about 700 °C onto a substrate, usually expanded silver foil or silver-plated nickel. Subsequently, the sintered material is electrochemically converted to Ag_2O and AgO. Other methods start with the silver oxides that are pasted as a slurry onto the substrate. Subsequently, the slurry is sintered and simultaneously reduced to silver. Occasionally but mainly for primary batteries, chemically prepared AgO is pressed together with a binder onto the substrate and directly used as electrode.

Besides the high energy content, an advantage of silver-oxide electrodes is that the internal resistance of the electrode is reduced with progressing discharge, since metallic silver is formed by the discharge reaction. The result is a very flat discharge curve as shown in Figure 102.

Disadvantages of silver oxide electrodes are their high price and a certain solubility of silver oxides. Cellophane membranes are used as additional separators to hinder the migration of silver ions to the negative electrode. This affects the discharge performance at low temperatures. But even with such a multilayer separator, cycle live is limited on account of silver migration.

Combination with the zinc electrode gives the reaction equation

$$AgO + Zn + H_2O \longleftrightarrow Ag + Zn(OH)_2$$

with the open circuit voltage (upper plateau) $U^{o,s} = 1.84$ V. Combination of the silver-oxide electrode with the cadmium electrode results in

$$AgO + Cd + H_2O \longleftrightarrow Ag + Cd(OH)_2$$

with the open circuit voltage (upper plateau) $U^{o,s} = 1.41$ V. On account of the smaller cell voltage, the energy content of the silver – cadmium system is inferior to that of the silver – zinc system. But cycle performance is slightly better. As in nickel – cadmium batteries, the electrolyte does not participate in the cell reaction in both systems. Figure 102 shows typical discharge curves of silver-oxide batteries. Both systems are restricted to a niche market, mainly in medical or military applications.

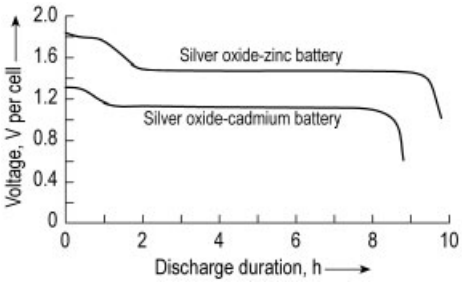

Figure 102. Typical constant current discharge curves of silver – zinc and silver – cadmium batteries of comparable size

7.1.3. Rechargeable Batteries of Particular Design

The batteries treated in this section differ in their design from the batteries discussed in connection with Figure 1. In this "classical" design, the reacting substances are stored as "active material" in the electrodes. An already mentioned exception is the nickel – hydrogen battery (Section 7.1.2.2). Here the active material of the negative electrode, hydrogen, in its charged state is stored as a gas under high pressure not in the electrodes but in the container of each cell.

The battery systems, described in this section, store the active material outside the cells in separate tanks, or take it in the form of oxygen from the surrounding atmosphere. To some extent these batteries are hybrids of batteries and fuel cells. Their technical development was mainly initiated by the desire to overcome the limited storage capability of "classical" batteries. One mainspring of these efforts was the development of electric vehicles. Only batteries with increased storage capability would solve the problem of limited cruising range that opposes the acceptance of such vehicles.

In *chlorine – zinc* and *bromine – zinc batteries*, the active material is partially stored in exterior tanks, that in parallel supply all cells of the battery.

The *vanadium redox battery* stands for a whole family of batteries in which the active material is exclusively stored in the electrolyte. The electrode material does not participate in the electrochemical reactions but only acts as acceptor or donor of electrons. The capacity of such a battery is determined by the amount of electrolyte that is stored in the tank.

Air – zinc batteries take the active material of the positive electrode, oxygen, from the surrounding atmosphere. So the positive electrode does not contain active material, but only delivers electrons for the reduction of oxygen ($O_2 + 4\,e^- \longrightarrow 2\,O^{2-}$) (cf. Section 6.1.3.4).

Table 18 shows characteristic data of the mentioned battery systems.

7.1.3.1. The Chlorine – Zinc Battery

The chlorine – zinc battery was a predecessor of the bromine – zinc battery. It had been developed in the 1960s and 1970s for vehicle application (cf., e.g., [200], p. 302). It is based on the cell reaction

$$\text{Charging} \longleftrightarrow \text{Discharging}$$

$$Zn + Cl_2 \longleftrightarrow ZnCl_2$$

Table 18. Rechargeable battery systems of special design

	System	Equilibrium voltage, V/cell	Stored energy		
			Theory, Wh/kg	Practice, Wh/kg	Practice, Wh/L
1a	Zinc – chlorine	2.12	$\approx 500^a$	$\approx 110^b$	$\approx 130^b$
1b	Zinc – bromine	1.8	438	65 to 70	max. 70
3	Vanadium redox	1.6^c	$\approx 40^d$	20 – 30	20 – 40
2	Zinc – air	1.65^e	1320^f	$\approx 170^g$	≈ 200

[a] Assuming that chlorine is stored as $[Cl_2 \cdot 7\,H_2O]$. [b] Whole battery, including the thermos container. [c] When fully charged. At standard concentrations 1.26 V/cell. [d] When the vanadium solution alone is considered. [e] Open-circuit voltage in practice ≈ 1.4 V/cell. [f] Referred to pure zinc as negative electrode and to the charged state: Oxygen intake during discharge reduces the figure to 1211 Wh/kg. [g] Referred to the weight of the charged state.

The design uses a cell stack similar to that shown in Figure 103 for the bromine – zinc battery and two separate tanks to store the electrolyte, an aqueous solution of zinc chlorine. The positive electrodes are supplied in parallel by one tank, and the negative electrodes by the other tank. The electrolyte is pumped through the cell stack. Chlorine, that is generated during charging, is dissolved in water and forms the solid compound chlorine hydrate ($Cl_2 \cdot xH_2O$; $6 \leq x \leq 8$) at temperatures below 10 °C. This hydrate is stable at temperatures below 10 °C, accumulates at the bottom of the tank, and can so be stored. Therefore, adequate cooling of the tank is required during charging and storage in the charged state. When the battery is discharged, the electrolyte has to be warmed up and then releases the required chlorine.

Although the data shown in Table 18 are promising, the development was not continued due to technological problems.

7.1.3.2. The Bromine – Zinc Battery

The bromine – zinc system was pioneered by Exxon and Gould in the USA. The development was mainly continued by SEA in Austria. Prototypes of this battery system have been built. A great number of tests has been carried out in electric vehicles, and also stationary applications, like load leveling, have been included in the tests [201].

This battery uses zinc and bromine as active materials. The reactions are

	Charging	$\longleftrightarrow$	Discharging
Positive electrode	$Br_2 + 2\,e^-$	$\longleftrightarrow$	$2\,Br^-$
Negative electrode	Zn	$\longleftrightarrow$	$Zn^{2+} + 2\,e^-$
Cell reaction	$Br_2 + Zn$	$\longleftrightarrow$	$ZnBr_2$

During charging, zinc is precipitated as a solid metal, that is dissolved during discharge. At the positive electrode, the components of the reactions remain dissolved. Figure 103 shows the principle design of such a battery.

The cell stack shown in Figure 103 consists of only three cells. Actual batteries contain stacks of 50 or more cells. Except from the two end-plates, all electrodes are bipolar. They consist of an electron-conducting plate of carbonized plastic enframed by insulating plastic. On the positive electrode, a porous layer of carbon increases the surface to enhance the reaction rate of bromine. In the center of each cell, a microporous separator is positioned as used in lead – acid batteries (e.g., DARAMIC, a polyester nonwoven impregnated with phenol – resorcinol – formaldehyde resin) to suppress the direct contact between bromine and zinc as far as possible.

The electrolyte forms two circuits: one that flows in parallel through the compartments formed between negative electrodes and separators and through the tank on the left-hand side of Figure 103, and a second circuit along the positive electrodes and through the tank on the right-hand side of the figure.

On charging, zinc is deposited at the negative electrodes, and shape changing problems (c.f. Fig. 4) or the formation of zinc dendrites are not relevant, since the precipitation occurs from the fast streaming electrolyte. When the battery is discharged, the $ZnBr_2$ concentration is correspondingly increased.

At the positive electrode, bromine is formed during charging. The tendency of bromine to form complex compounds is supported by the addition of two complex-forming substances to the electrolyte: pyrrolidinium bromide and mor-

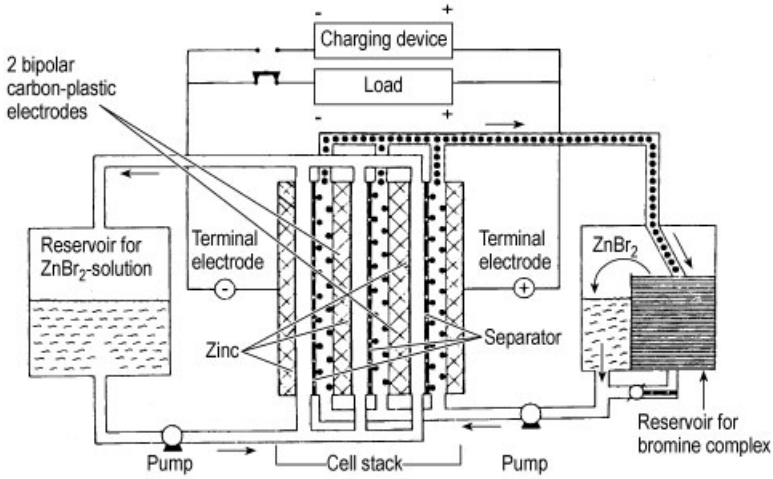

Figure 103. Bromine – zinc battery

pholinium bromide [202]. They form nonaqueous complexes with the generated Br_2 that is parted from the electrolyte as indicated in the tank on the right-hand side of Figure 103. When the battery is discharged, the bromine complex is added to the streaming electrolyte and bromine released.

7.1.3.3. The Vanadium Redox Battery

The vanadium redox battery uses also a cell stack of bipolar plates and two separate tanks that supply the compartments of the positive and the negative electrodes with electrolyte, similar to the design of Figure 103 (cf., e.g., Fig. 7.9 in [200]). An ion exchange membrane is used as separator in each cell.

This system makes use of the different states of oxidation that are characteristic for vanadium. The V^{5+}/V^{4+} equilibrium and the V^{3+}/V^{2+} equilibrium are used at the negative and positive electrode, respectively. The corresponding reactions are:

	Charging	$\longleftrightarrow$	Discharging
Positive electrode	$V^{5+} + e^-$	$\longleftrightarrow$	V^{4+}
Negative electrode	V^{2+}	$\longleftrightarrow$	$V^{3+} + e^-$
Cell reaction	$V^{5+} + V^{2+}$	$\longleftrightarrow$	$V^{4+} + V^{3+}$

The electrolyte is a mixture of 2 mol/L vanadium sulfate and 20 wt % sulfuric acid. Figure 104 shows charge discharge curves of a cell stack.

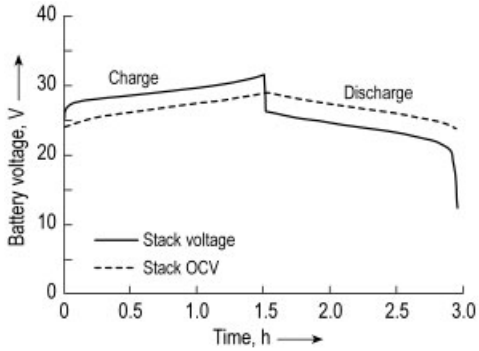

Figure 104. Charge – discharge curves of a vanadium redox battery. The cell stack consists of 17 cells; constant current of 100 Ah was applied for charging and discharging

A main advantage of such a redox system is that the electrode reactions do not involve solid phase changes or precipitation mechanisms. As a consequence, a large number of cycles can be reached, and deep discharges are not critical.

The most striking disadvantages of these systems are the comparatively low values of specific energy (Wh/kg) and energy density (Wh/L), shown in Table 18. They are caused by the low concentration of the reacting species on account of limited solubility. The specific power (W/kg) is limited for the same reason.

A further disadvantage is a certain transition of the positive vanadium ions from the compart-

ment of the positive electrode to that of the negative electrode, caused by migration (cf. Section 3.2). It results in gradual concentration increase in the compartments of the negative electrodes and requires occasional remix of the two electrolytes.

Shunt currents are unavoidable, caused by the required electrolytic connection between the individual cells of a stack. To keep these parasitic currents low, the connecting tubes must be small, but this simultaneously limits the high rate performance.

7.1.3.4. Rechargeable Air – Zinc Batteries

The air – zinc system is in widespread use as primary battery system, e.g., in the form of button cells for hearing aids (cf. Section 6.1.3.4).

The main advantages of the air – zinc system are the high values of specific energy and energy density shown in Table 18. This is partly due to the fact that the active material of the positive electrode is not contained in the battery when charged. Furthermore zinc is an element of low specific weight and offers a very negative equilibrium potential when used as an electrode.

The (simplified) electrode reactions are

	Charging	$\longleftrightarrow$	Discharging
Positive electrode	$1/2\,O_2 + H_2O$	$\longleftrightarrow$	$2\,OH^-$
Negative electrode	$Zn + 2\,OH^-$	$\longleftrightarrow$	$ZnO + H_2O$
Cell reaction	$Zn + 1/2\,O_2$	$\longleftrightarrow$	ZnO

Two problems arise in respect to recharging:

- The high solubility of zinc causes shape change as indicated in Figure 4.
- Intensive oxygen evolution that occurs at the positive electrode during recharge of the zinc electrode may damage the positive electrode.

A result of the first item is that in general zinc electrodes can stand only a very limited number of cycles.

One possibility to overcome this problem is to replace the used zinc electrodes after each discharge by new ones ("mechanical recharge"). The zinc oxide of the removed negative electrodes is regenerated in a central electrolysis unit that operates for a whole fleet of cars, equipped with such zinc – air batteries [204–206]. This procedure also solves the second problem, since the electrode is not recharged within the cell. Results of a larger fleet test in Bremen were disappointing [206], mainly for economical reasons connected to the regeneration of the zinc electrodes.

Another approach applies a small number (< 10) of recharges within the cell before the zinc electrode is replaced [207]. When the negative electrodes are recharged within the battery, auxiliary electrodes are used as positives. Extended fleet tests are planned.

7.2. Batteries with Nonaqueous Electrolyte

Alkaline metals, like lithium and sodium, are promising substances as negative electrodes for batteries of high storage capacity due to their negative electrode potential and low specific weight. They can, however, not be used together with aqueous electrolytes, because of their high reactivity to water. Alternatives are aprotic solvents that allow a certain conductivity, salt metals, or solid state electrolytes. The disadvantage of all these electrolytes is their comparatively low conductivity. In organic solvents this can be compensated by extremely thin and narrow spaced electrodes. Salt melts and solid state electrolytes require elevated or even high temperatures to achieve a sufficient high conductivity.

7.2.1. Rechargeable Lithium Batteries

The fundamental properties of lithium are shown in Section 6.2. As in primary batteries the layer that is caused by reactions between lithium and the electrolyte, called SEI (solid electrolyte interface) is formed at the electrode, prevents further attack at the metal, but allows the penetration of Li^+ ions. When the electrode is discharged, the protecting layer is destroyed, at least partially, and during a subsequent recharge some of the lithium is not precipitated as metal but contributes to further growth of the surface layer and is lost as capacity. Compared to primary batteries, the surface layer is less stable and normally a voltage delay as shown in Figure 28 is not observed with secondary batteries.

To compensate for the loss of capacity, rechargeable lithium batteries have to be equipped with a large excess of lithium to reach

a satisfactory number of cycles. But this implies the potential of hazards. If under anomalous operational conditions the temperature of the battery would reach the melting point of lithium (180 °C), the comparatively large amount of residual metal would react directly with the electrolyte and the battery might explode. For this reason, larger rechargeable batteries (cylindrical or prismatic) with metallic lithium electrodes and liquid electrolytes are only used for military applications, and are not available on the market.

7.2.1.1. Button or Coin Cells

Lithium as Negative Electrode. A great variety of small rechargeable button or coin cells exist, based on lithium as the negative electrode. These cells are mainly designed for memory backup, solar watches, pocket calculators, or similar applications. Most of them are 3 V systems, based on lithium alloys with aluminum as alloying additive as the negative and manganese dioxide or vanadium pentoxide as the positive electrode. But also polymers that can insert lithium, like polyaniline or polypyrrol, are used in the positive electrode.

The electrochemical reactions (when MnO_2 is used) correspond to those described in Section 6.2.1.1.2 for primary batteries, but here the subscript $(1-x)$ indicates that in such oxides the intercalation/de-intercalation mechanism can be repeated only within certain limits of the lithium content. Usually, in such layered oxides x is not allowed to exceed 0.5, otherwise the crystalline structure is changed (cf. Section 7.2.1.4.3).

	Charging	↔	Discharging
Positive electrode	$x\mathrm{Li}^+ + \mathrm{Li}_{(1-x)}\mathrm{MnO}_2 + x \cdot \mathrm{e}^-$	↔	LiMnO_2
Negative electrode	$x\mathrm{Li}$	↔	$x\mathrm{Li}^+ + x\mathrm{e}^-$
Cell reaction	$x\mathrm{Li}^+ + \mathrm{Li}_{(1-x)}\mathrm{MnO}_2$	↔	LiMnO_2

During discharge, in the positive electrode Mn^{4+} ions are reduced to Mn^{3+} and lithium ions are intercalated into the layered structure to compensate for the added negative charge. At the negative electrode, lithium is oxidized and dissolved.

During charging, the Li^+ ions are de-intercalated at the positive electrode, and a corresponding number of them are reduced and precipitated at the negative electrode.

The design of such button or coin cells corresponds to that shown in Figure 85. In general, the height of lithium button cells is small, particularly flat types occasionally are called coin cells. Figure 105 shows the cross-section of a typical example of the 3 V systems.

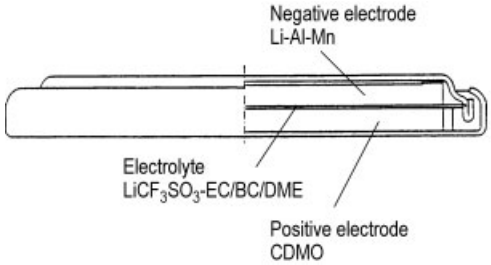

Figure 105. Cross-section of a button cell (Sanyo ML2430). CDMO means "composite dimensional manganese oxide" i.e., lithiated manganese dioxide Li_xMnO_2 $(0 \leq x \leq 1)$, produced by heat treatment (375 – 420 °C) of a mixture of LiOH and γ-MnO_2. Dimensions of the cell: diameter 24.5 mm; height 3.0 mm; weight 4.1 g

In this battery, the electrolyte is a mixture of ethylene carbonate (EC), butylene carbonate (BC), and dimethoxyethane (DME) to achieve desired features, like melting point and vapor pressure. The melting point must be low enough to provide sufficient conductance also at low temperatures, and the vapor pressure should not be too high to ensure proper sealing of the cell. Furthermore, the composition of the electrolyte greatly influences the stability of the protecting layer at the lithium surface. In the example of Figure 105, the conducting salt is $LiCF_3SO_3$. Often $LiAsF_6$ is applied for button cells, since As stabilizes the protecting layer at the negative electrode. A typical discharge curve is shown in Figure 106.

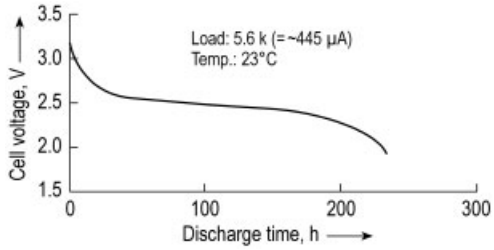

Figure 106. Discharge of the cell shown in Figure 105 by a resistor of 5.6 kΩ

7.2.1.2. Lithium Batteries with Polymer Electrolytes

Polymer electrolytes react less violently with metallic lithium electrodes and allow the direct combination without the danger of hazards. Such electrolytes are made of a polymer, like poly(ethylene oxide), with an incorporated lithium salt (e.g., $LiClO_4$, $LiPF_6$, $LiCF_3SO_3$) to achieve ionic conductivity. The disadvantage of these electrolytes is their low conductance that requires extremely thin electrolyte layers and mostly increased temperature to reach sufficient conductivity. Furthermore, the chemical stability of polyethylene oxide does only allow voltages up to about 3 V. For this reason, vanadium oxides (VO_x, e.g., V_2O_5 or V_6O_{13}) usually are employed in the positive electrode because the resulting cell voltage is in the range of 3 V. The electrode is formed as a thin layer mixed with carbon, rolled or pressed on a metal foil as current conductor. Figure 107 shows a corresponding arrangement of the electrodes and the polymer electrolyte. This combination reaches an overall thickness of only 100 µm:

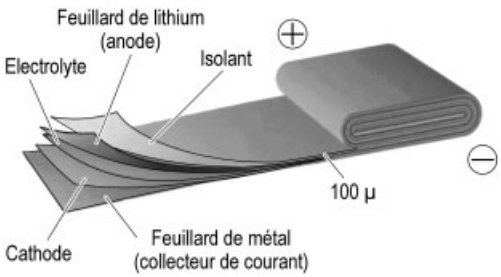

Figure 107. Arrangement of the electrodes in a lithium polymer cell. The positive electrode (cathode) is a mixture of the polymer electrolyte with about 50 % of vanadium oxide (mainly the phases V_2O_3 or V_6O_{13}) and a small amount of carbon black (≈ 5 %) (3M/Hydro Quebec, [208])

The electrochemical reactions are (discharge from left to right):

Positive electrode	$y \cdot Li^+ + Li_{(1-y)}VO_x + y \cdot e^-$	$\leftrightarrow$	$LiVO_x$
Negative electrode	$y\,Li$	$\leftrightarrow$	$y \cdot Li^+ + y \cdot e^-$
Cell reaction	$y \cdot Li^+ + Li_{(1-y)}VO_x$	$\leftrightarrow$	$LiVO_x$

During discharge, lithium is dissolved at the negative electrode and absorbed by the positive one. During charging, lithium leaves the positive and is precipitated at the negative electrode. The electrode arrangement shown in Figure 107 can be used for various battery designs: especially flat and thin designs, spirally wound for cylindrical cells, but also for prismatic cells. A larger monoblock battery is shown as an example in Figure 108.

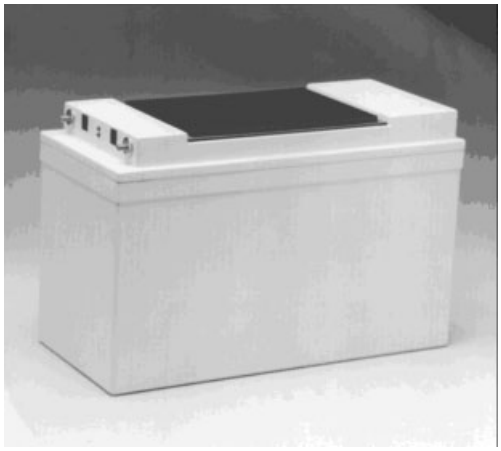

Figure 108. Prototype of a 24 V/80 Ah block battery (9 cells) designed for stationary applications, based on polymer electrolyte. (Dimensions: length 394 mm, width 176 mm, height 251 mm, weight 20 kg) [208]

This battery was developed to replace lead – acid batteries in stationary applications where the battery is continuously operated at elevated temperatures and the service life of a lead – acid battery is dramatically reduced. In such an application, positive utilization could be made of the disadvantage that sufficient conductivity of the polymer electrolyte is only reached at temperatures above 40 °C.

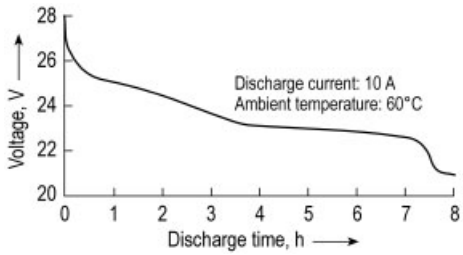

Figure 109. Discharge curve of the battery shown in Fig. 108 [208]. The voltage drop from about 3 to 2.4 V/cell is mainly caused by the positive electrode

The discharge curve of this battery is shown in Figure 109. It is typical for the VO_x cathode, and means that the voltage of a single cell drops from 3 V to 2.3 V during discharge. During the

discharge, the vanadium oxide passes through various structural phases that can be recognized by the different steps of the curve (cf. e.g., [209], p. 58).

The cycle number of such batteries is limited, mainly on account of the above mentioned growing layer at the lithium surface that consumes lithium and increases the internal resistance of the battery.

7.2.1.3. Batteries with Glassy Electrolytes

The possibility to design rechargeable all-solid state batteries, based on glassy electrolytes is described in [209], p. 199. Lithium ion conductive glasses reach conductivities in the range of 10^{-3} S/cm. Their general composition is $x\mathrm{LiI} \cdot y\mathrm{Li_2S} \cdot z\mathrm{SiS_2}$ where $\mathrm{SiS_2}$ can be replaced by $\mathrm{P_2S_5}$ or $\mathrm{B_2S_3}$ and instead of the lithium-iodine compound, LiBr or LiCl may be used.

Lithium is used as the negative electrode. Combination with $\mathrm{Li}_x\mathrm{TiS_2}$ and $\mathrm{Li}_x\mathrm{CoO_2}$ in the positive electrode leads to 3 V and 4 V batteries, respectively. Small coin cells with capacities in the range of 1 to 20 mAh reached more than 100 cycles at constant capacity and 100 % of coulombic efficiency.

7.2.1.4. Lithium-Ion Batteries

In lithium-ion batteries, intercalation and de-intercalation of lithium ions characterizes both electrodes. During discharge, the negative electrode releases $\mathrm{Li^+}$ ions, that are absorbed by the positive electrode, while these ions swing back, when the battery is charged.

In the negative electrode, nowadays mostly carbon intercalation compounds ($\mathrm{Li}_x\mathrm{C}_n$) are applied, only 1.5 V button cells employ certain oxides or sulfides as mentioned at the end of this section. Promising tin-based alloys as negative electrodes are still (2000) under development.

Specially pretreated carbons, based on graphite or coke, can absorb lithium ions or release these ions while a corresponding amount of electrical charges is drawn from or added to the carbon, respectively, according to

$$\mathrm{Li}_x\mathrm{C}_n \longleftrightarrow x\mathrm{Li^+} + \mathrm{C}_n + x \cdot \mathrm{e^-}$$

From left to right, this equation means discharging of the electrode (oxidation of carbon = de-intercalation of $\mathrm{Li^+}$ ions). The reversal, the insertion or intercalation of $\mathrm{Li^+}$ ions, means charging, i.e., reduction of the carbon by absorption of electrons. The negative charging simultaneously is compensated by the intercalation of a corresponding number of $\mathrm{Li^+}$ ions. Lithium does not form a real chemical compound with carbon, rather "intercalation" or "insertion" is largely based on electrostatic forces. (The reaction resembles the intake of $\mathrm{H^+}$ ions in the nickel hydroxide electrode, illustrated in Fig. 79).

Intercalation and de-intercalation at the negative electrode correspond to the precipitation and dissolution reactions of lithium electrodes, and the potential of such carbon electrodes is close to that of metallic lithium. Thus the combination with corresponding positive electrodes results in batteries of similar cell voltage. But the decisive difference is that the structure of the carbon host remains nearly unchanged when the electrode is charged or discharged. Only a slight volume increase is observed when the lithium ions are intercalated.

In the battery, such a carbon electrode is combined with a positive electrode of a metal oxide that also intercalates $\mathrm{Li^+}$ ions (mainly $\mathrm{Li}_x\mathrm{CoO_2}$, $\mathrm{Li}_x\mathrm{NiO_2}$, or $\mathrm{Li}_x\mathrm{Mn_2O_4}$). These positive electrodes intercalate the lithium when discharged, i.e., the lithium-filled material characterizes the discharged state. Then the $\mathrm{Li^+}$ ions compensate for a corresponding reduction of the metal ions ($\mathrm{Me^{4+}} + x \cdot \mathrm{e^-} \longrightarrow \mathrm{Me}^{(4-x)+}$). During charging, the $\mathrm{Li^+}$ ions are de-intercalated.

For $\mathrm{Mn_2O_4}$ (spinel) the reaction is

$$\mathrm{Li}_x\mathrm{Mn_2O_4} + x \cdot \mathrm{e^-} \longleftrightarrow \mathrm{Mn_2O_4} + x \cdot \mathrm{Li^+}$$

As in the carbon electrode, also in the positive electrode the host material remains (nearly) unchanged, and in the whole cell, only the $\mathrm{Li^+}$ ions swing between the positive and negative electrode, according to the cell reaction

$$|n\mathrm{C} + \mathrm{Li}_x\mathrm{Mn_2O_4} \longleftrightarrow \mathrm{Mn_2O_4} + \mathrm{Li}_x\mathrm{C}_n$$

This gave the battery the name "rocking chair battery" sometimes used. Figure 110 illustrates the system as a simple model.

The electrolyte is an organic (aprotic) solvent with a dissolved lithium salt that provides conductivity. On account of the comparatively low

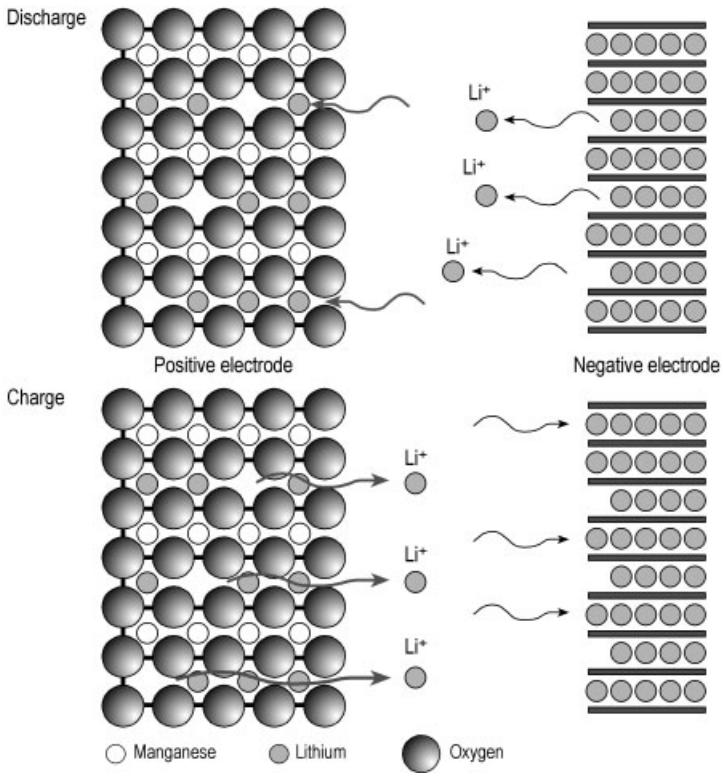

Figure 110. Electrode reactions in rechargeable Li-C – Li-metal oxide batteries. In the positive electrode a layered structure of an oxide (here Mn_2O_4) absorbs Li^+ ions when discharged (left hand side), while in the negative electrode (on the right hand), carbon (here the layered structure of graphite, Li_xC_6) absorbs Li^+ ions when charged (from [210])

conductivity of organic electrolytes, the spacing between the electrodes has to be as small as possible. For this reason, in most lithium-ion batteries the thickness of the separator is about 25 µm and the thickness of the electrodes in the same order of magnitude. Only in low-rate button cells, these dimensions are larger (cf. Section 7.2.1.1). The arrangement of the electrodes is similar to that shown in Figure 107. Instead of the lithium electrode, a copper foil as current conductor with a layer of the carbon material forms the negative electrode, and the liquid organic electrolyte is soaked by a microporous separator. In most cases, the electrodes are wound up for cylindrical cells, corresponding to Figure 88.

Molybdenum Oxide, Niobium Oxide, or Titanium Oxide as Negative Electrode. Button cells, with MoO_2 $Nb_2O_5TiS_2$ as negative electrode materials that intercalate lithium, represent 1.5V systems. They were mainly designed for the direct substitution of nickel – cadmium or nickel – metal hydride batteries. The reaction equation for the negative electrode is, e.g., for MoO_2:

$$\begin{array}{ccc} \text{Charging} & \longleftrightarrow & \text{Discharging} \\ \text{LiMoO}_2 & \longleftrightarrow & \text{Li}_{(1-x)}\text{MoO}_2 + x \cdot \text{Li}^+ + \text{x} \cdot \text{e}^- \end{array}$$

At this negative electrode, the lithium-rich phase represents the charged state (the reduced state). Li^+ ions are de-intercalated during discharge, while the are intercalated during charging.

The potential of these electrodes is about 1 V positive compared to that of the metallic lithium electrode, and in combination with positive intercalation electrodes based on $LiMn_2O_4$, $LiCoO_2$, or $LiNiO_2$ a cell voltage of about 1.5 V results.

7.2.1.4.1. Lithium – Carbon Electrodes (Lithiated Carbons)

Carbon is available in a variety of crystallographic forms. For lithium intercalation,

graphite, but also other special modifications of carbon, are important (→ Carbon, Chap. 4.3.2.).

In such substances the carbon atoms form planes of hexagonal rings (C_6) that are arranged in layers with each second layer slightly shifted against the one below or above. The layers are only bound by van der Waals forces, and when the Li^+ ions are intercalated between these layers, the vertical distance between the layers is slightly increased (from 0.67 nm to about 0.75 nm).

Nongraphite carbons also consist of C_6 rings, but only small units of these rings exists, and these units are rotated with respect to each other. Such modifications are called "turbostratic carbons". They are denoted as *soft carbons* when they can be converted to a certain extent into graphite by annealing at high temperatures (up to 3000 °C). *Hard carbons* or *nongraphitizable carbons* are amorphous or glassy carbons that withstand such a conversion even at extremely high temperatures. In these materials the hexagonal fragments are fixed by cross-linking bonds that hinder the ordering into a graphitic structure. Hard carbons are mainly produced by pyrolysis of resins.

For graphite, the above general equation reads

$$Li_xC_6 \longleftrightarrow xLi^+ + C_6 + x \cdot e^-$$

where x can reach the value of 1, i.e., in maximum one Li^+ ion can be intercalated per C_6. The molar mass of LiC_6 is 79 g/mol, i.e., the storage capability of LiC_6 is $1 \cdot F = 28.6$ Ah per 79 g or 340 mAh/g. [Often this value is referred to the discharged electrode (C_6) with the molar mass 72.1 g/mol. Then the value is 372 mAh/g.] This is in the same order as the corresponding figure for the cadmium/cadmium hydroxide electrode. Charging and discharging of such a graphite electrode occurs at a potential that is only 0.1 to 0.2 V more positive than that of an electrode of metallic lithium. An example is shown in Figure 111. Thus, in regard to the charging – discharging characteristic, lithium – graphite electrodes can replace metallic lithium, but they cannot reach the specific energy of the metal electrode.

Fig. 111 shows a symptom that characterizes cells with lithiated carbon electrodes: the initial loss of lithium. The lithiated carbon is not stable, and the cell has to be assembled in the discharged state, i.e., with a lithium-free carbon electrode and the lithiated metal oxide (Li_xMO_2) in the positive electrode. During the first charge, lithium is intercalated in the carbon. Simultaneously, the surface layer (SEI), mentioned at the beginning of Section 7.2.1, is formed at the negative electrode by direct reaction between lithium and components of the electrolyte. During the subsequent discharge (and the following cycles), this amount of lithium cannot be recouped as capacity but remains in the layer. This share of the original lithium content is called "irreversible capacity" or "lost capacity" (C_{irr} in the upper part of Fig. 111).

The formation of such a protective film is of great importance. The protective layer must prevent co-intercalation of the solvent, but Li^+ ions must be able to pass. Especially in graphite, due to the comparatively weak van der Waals forces, solvated Li^+ ions [$Li_x(solvent)_y$] can be intercalated. On account of their large size, such solvated ions bend apart the graphite layers with causes a volume increase of more than 100 % and disintegration of the carbon electrode by exfoliation. The possibility to avoid such co-intercalation of the solvent depends to a large extent on the composition of the electrolyte, but also on the structure of the graphite.

Hard carbon or nongraphitizable carbon contains short layers of hexagonal carbon rings that are quite disordered, as indicated in Figure 112. As a result, both sides of the carbon layers are available for the adsorption of Li^+ ions. Furthermore, micropores may exist as an additional reservoir for Li^+ ions.

The increased number of positions for Li^+ ion adsorption results in a larger capacity than in the case of graphite-lithium intercalation compounds. In test cells, up to 700 mAh/g have been achieved, corresponding to the formula Li_2C_6. Fig. 113 shows the constant current-open circuit charge-discharge curve of the initial cycle. The irreversible capacity (C_{irr} in Fig. 111) is here denoted "loss". During the following cycles, the hysteresis between charge and discharge will be reduced, corresponding to the second cycle in Figure 111.

On account of nonequivalent sites of intercalation, the curves in Figure 113 do not show such a plateau slightly above the lithium potential, as it is observed in the case of graphite in Fig-

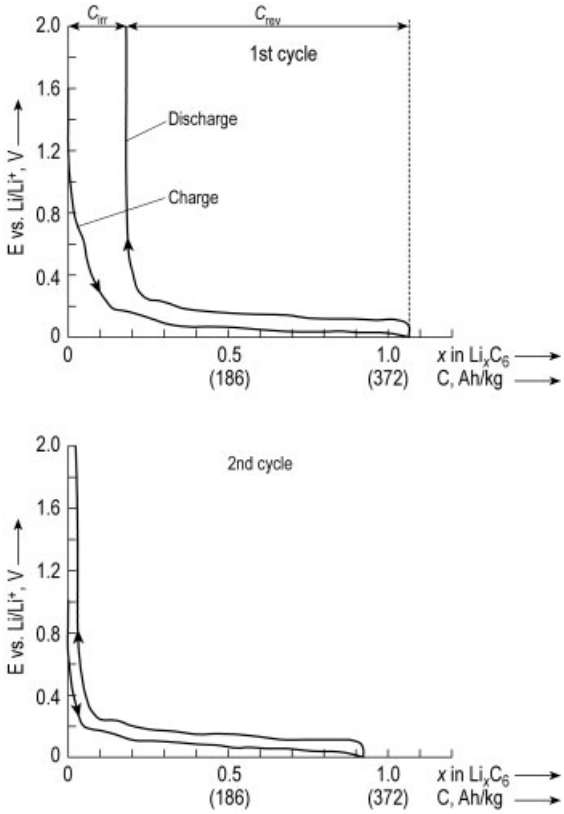

Figure 111. Constant-current charge/discharge curves of a graphite electrode referred to a metallic lithium electrode (Li/Li^+). Electrolyte: $LiN(SO_2CF_3)_2$/ethylene carbonate (EC)/dimethyl carbonate (DMC) (Fig. 7 on p. 392 of [213])

ure 111. Instead, charge and discharge occurs in a broad range of the potential. This means that a certain portion of the capacity can only be discharged at a lower cell voltage. Furthermore, at the end of the charging process, the electrode potential comes down very close to the potential of the metallic lithium electrode. This implies the danger of precipitation of metallic lithium which might form dendrites and so cause short circuits. In any case, such precipitated metallic lithium will react with the electrolyte and so get lost. On account of these limitations, in practical batteries, the aforementioned high capacity can only partly be utilized.

An advantage of hard carbons is that the cross-linking bonds do not allow the co-intercalation of solvents, and in general, volume expansion connected with Li^+ intake can be neglected.

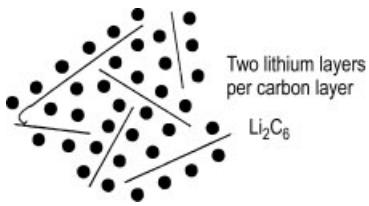

Figure 112. Schematic model of the possibilities for lithium intercalation in hard carbon. The lines symbolize the carbon layers, the discs the intercalated Li^+ ions (part of Fig. 5.14 in [215], p. 113)

7.2.1.4.2. Negative Electrodes Based on Alloys

The application of lithium alloys in the negative electrode has been studied since the 1970s, because a great number of metals form lithium alloys that contain more lithium than can be inter-

calated in carbon negatives (a review is given in [211]). The main disadvantage of such alloys are large volume changes when lithium is absorbed (up to 100 %) that destroy the electrode. Another drawback is the low diffusion rate of the lithium ions at room temperature that limits the penetration depth of lithium. As a consequence, the size of the particles must be in order of manometers and they should be embedded in a host material in a way that allows volume increase.

To solve these problems, tin-based oxides (amorphous tin-based oxide, ATCO, or tin composite oxides, TCO) have been introduced by Fuji in 1994 [212]. They are mixtures of SnO, B_2O_3, $Sn_2P_2O_7$, Al_2O_3, and graphite and offer the possibility of negative electrodes with high capacity up to 650 mAh/g and also a certain cycle stability (cf., e.g., p. 406 in [213]). During charging the SnO obviously is reduced to fine Sn particles, that alloy up to 4.4 Li ions per Sn [214] and are embedded in an amorphous ductile surrounding.

A further method to create alloys of a sufficient small particle size is electroplating of corresponding mixtures. The cycle stability of such mixtures is much better than that of the pure substance. Thus, mixtures of Sn – SnSb or Sn – $SnAg_3$ – $SnAg_4$ are superior to Sn alone. The explanation is that the various phases react at different potential while the rest remains unchanged and so acts as a stable host ([211], p. 41).

On the whole, the described alloys are not yet (2000) stable enough to be used in commercial batteries but they are object of a worldwide intensive research.

7.2.1.4.3. Active Materials in the Positive Electrode (Cathode)

In the positive electrode, mainly three compounds and their variations are applied that have a high oxidation potential and a layered structure suitable for lithium intercalation: cobalt dioxide ($LiCoO_2$), nickel dioxide ($LiNiO_2$), and manganese dioxide (spinel) ($LiMn_2O_4$). Figure 114 shows a comparison of corresponding charge – discharge curves.

Figure 114 shows that also the positive electrode suffers a certain capacity loss between the first and the tenth cycle, the so-called "capacity fade". This parameter is strongly influenced by the composition of the compound but also by its structure that is determined by its preparation.

The extraction of lithium is limited, and the maximum of x usually should not exceed 0.5, when the oxide is described by $Li_{(1-x)}MO_2$ (M stands for Co, Ni, or Mn). When de-intercalation is continued beyond this limit by heavy overcharging, the structure of the material is irreversibly changed, and loss of capacity results.

Cobalt oxide was the first one used in commercial batteries. It has a stable layered structure. The mean discharge voltage against lithium is 3.8 V (Fig. 114 A). The main disadvantage of cobalt oxide is its high price. Furthermore, there are concerns about the toxicity of cobalt when run-down batteries are disposed of.

$LiNiO_2$ has a slightly lower discharge voltage than $LiCoO_2$ (mean value 3.52 V in Fig. 114 C) but a high capacity. It is more difficult to prepare and to maintain at an ideal layered structure, since nickel ions tend to occupy sites in the lithium layer and so disturb the structure. As Figure 114 indicates, a marked reduction of the capacity is observed between the first and the tenth discharge due to irreversible phase transitions during the first recharge. The main disadvantage of nickel dioxide is that Ni^{4+} oxides, formed by overcharging, are unstable and can cause hazardous reactions within the cell. For this reason, larger cells with nickel hydroxide in the positive electrode will be critical for safety reasons.

Substitution of cobalt for nickel, as in example B of Figure 114 ($LiCo_{0.5}Ni_{0.5}O_2$), can improve the situation.

Lithium manganese dioxide ($LiMnO_2$) is applied for primary batteries with lithium electrodes. However, the stability of the layered structure is not sufficient for a high cycle stability. But the lithium manganese spinel ($LiMn_2O_4$) is of great interest. Its structure corresponds to that of the mineral spinel $MgAl_2O_4$. $LiMn_2O_4$ is a relatively inexpensive material, offers a high electrode potential (mean discharge voltage 3.97 V in Fig. 114 D) and causes no safety risks as compared with nickel oxide and no environmental problems when the battery is disposed of. Disadvantages of the spinel are that at deep discharges (high Li^+ content) Mn^{2+} ions are formed that are dissolved and so lost from the electrode. Overcharging, on the other

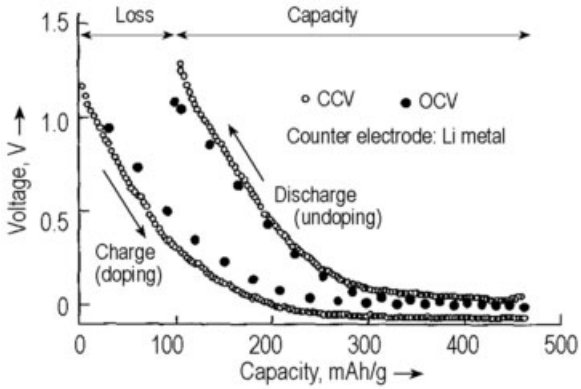

Figure 113. Typical intermittent charge and discharge of a nongraphitizable carbon electrode at the initial cycle (doping = intercalation of Li^+ ions, undoping = de-intercalation). CCV: closed circuit voltage during constant-current charging and discharging. OCV: open circuit voltage.

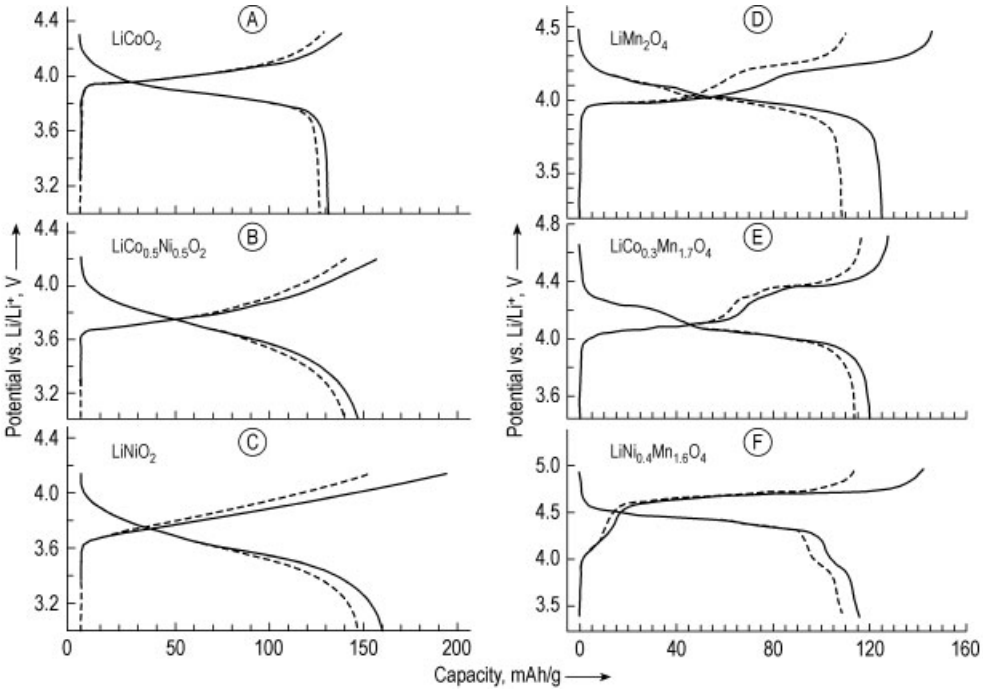

Figure 114. Charge – discharge curves of various layered oxides versus lithium (button cells with lithium – aluminum alloys as negative electrodes). The solid lines represent the first cycle, the dashed ones the tenth cycle [215]

hand, causes the already mentioned undesired changes of the phase structure. Substitution of manganese, as shown in Figures 114 E and F ($LiCo_{0.3}Mn_{1.7}O_4$ and $LiNi_{0.4}Mn_{1.6}O_4$ respectively), provides the possibility to tailor the active material to the desired performance data and to overcome the mentioned difficulties.

7.2.1.4.4. Electrolytes

The liquid nonaqueous electrolytes are always a combination of an organic solvent and a lithium salt that effects the conductivity. One important condition is stability of the solvent in the wide voltage range of the 4 V system (the so-called

"voltage window"). Furthermore, the choice of the electrolyte is of great importance, since the protecting layer on the surface of the negative electrode results from reactions between lithium and components of the electrolyte.

Solvents, used in practical batteries, are, e.g., propylene carbonate (PC), ethylene carbonate (EC), or dimethyloxyethane (DME) (cf., e.g., Table 36.8 in [216] or Table 1, p. 460 in [213]). Often mixtures are applied to reach required features. The mostly used salt is $LiPF_6$. $LiAsF_6$is also applied, but has the disadvantage of environmental problems. Both salts are highly resistant to oxidation (large voltage window) and impart a comparatively high conductivity. But they tend to cause unwanted side reactions at higher temperatures. Especially developed for lithium-ion batteries were a number of salts with large molecular anions, like $Li[N(CF_3SO_2)_3]$ ("lithium imide") or $Li[C(SO_2CF_3)_3]$ (further examples and details cf. [213], p. 462). But the large molecular mass requires corresponding amounts of these salts to reach the ionic concentration, needed for sufficient conductivity, and this might not be possible due to limited solubility.

The electrolyte must be thoroughly freed from water, since even traces of H_2O would cause corrosion reactions with lithium and result in unwanted surface layers.

7.2.1.4.5. Separators

The separators in lithium-ion batteries are usually made of polyethylene or polypropylene films. Nonwoven material is only used in button cells designed for low load rates. Such a separator would not be suited in cells for higher loads, since there the thickness of the separator has to be in the range of 25 μm on account of the comparatively low conductivity of the organic electrolyte. Special films are required, characterized by an extremely fine pore system (≤ 0.1 μm) and able to withstand the mechanical forces of the cell manufacturing process, especially the winding for the production of cylindrical cells.

Such separators are known under various trade names, e.g., Cellgard and Setela. (cf. e.g., Table 1 on page 555 in [213]). The pore system is created either by filling fluids that later are extracted (wet process) or by overstretching the film (dry process).

Combination of layers of polypropylene and polyethylene are used to increase the safety of the battery. When the melting point of polyethylene (135 °C) is exceeded, the pores in this layer are closed, and the current flow is interrupted, while the polypropylene layers (*mp* 165 °C) are still stable. Thermal runaway is so avoided.

7.2.1.4.6. Lithium-Ion Polymer Batteries

Particular interest exists for lithium-ion batteries with polymer electrolyte, since such batteries could be manufactured in any shape, they would not have leakage problems and could easily be miniaturized. Credit cards or similar applications afford such a battery design. Disadvantages of polymer electrolytes are their low conductivity, and that an additional resistance may be caused by contact problems between the surface of the electrodes and the polymer (wetting problems).

A solution of these problems is offered by "gel electrolytes" or "gelionic electrolytes". They can be produced in two ways: immobilization of the liquid electrolyte by adding a soluble polymer or by filling a porous matrix with the liquid electrolyte. An example for the latter is the "Bellcore" process that has been licensed to a number of manufacturers (cf., e.g., Fig. 10.19 on p. 235 in [209]). It employs carbon in the negative and $LiMn_2O_4$ in the positive electrode. Both electrode films are rolled with a plastic layer on top of each, which later is filled with the electrolyte. Then both are combined with the separator between them. Subsequently the plasticizer is extracted from the plastic layers, and so a pore system created within them. Then the electrode group is filled with electrolyte. Thus, the electrolyte is not only soaked into the plastic layers and the separator, but it also gets in intimate contact with the electrode surface.

7.2.1.4.7. Practical Batteries

Since Sony introduced the first lithium-ion battery into the market in 1991, an exciting growth has occurred of the market share of this battery system. The first one was a cylindrical cell

(Sony 18650), based on nongraphitizable carbon in the negative and $LiCoO_2$ in the positive electrode. The battery reached the specific energy of 120 Wh/kg and the energy density of 300 Wh/L. A typical discharge curve is shown in Figure 115.

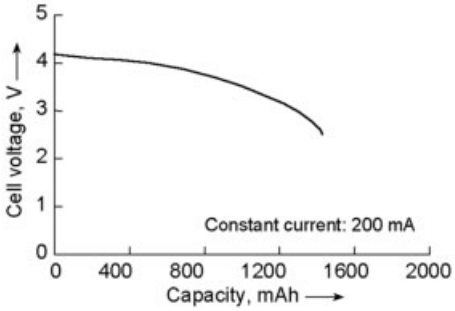

Figure 115. Discharge of a rechargeable lithium-ion cell (Sony 18650)

The cycle behavior of this cell is shown in Figure 116. Depending on the charging/discharging conditions, more than 2000 cycles have been attained [217].

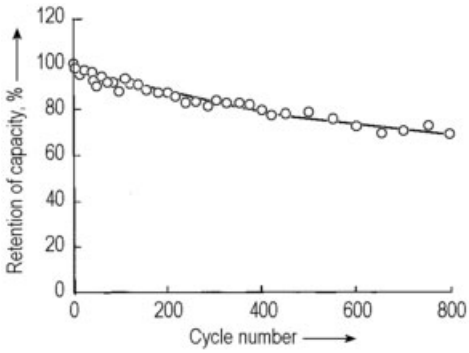

Figure 116. Cycle performance of a cylindrical lithium-ion cell (Sony 18650). Charge: constant current 1 A; constant voltage 4.2 V; Discharge: constant current 1 A, end of discharge voltage 2.7 V

The $LiCoO_2$ used in the Sony 18650 cell contains a small amount of Li_2CO_3 which is added when $LiCoO_2$ is synthesized. When the cell is charged to a voltage higher than 4.8 V, the Li_2CO_3 is decomposed and the resulting CO_2 gas increases the internal pressure. Then, a concave aluminum disk at the top of the cell is pushed upwards, as shown in Figure 10, which interrupts the current flow between the cover and the electrodes. Hazards by overheating are so avoided.

Today (2000) a great variety of lithium-ion batteries is offered by different manufacturers. Usually they are smaller sizes cylindrical cells or button cells, since most applications are portable devices, such as cellular phones, notebooks, and camcorders.

The strong efforts to develop advanced batteries for electric vehicles, sponsored, e.g., by United Stated Advanced Battery Systems Consortium (USABC), initiated the development of large batteries for traction applications. Figure 117 shows such a battery.

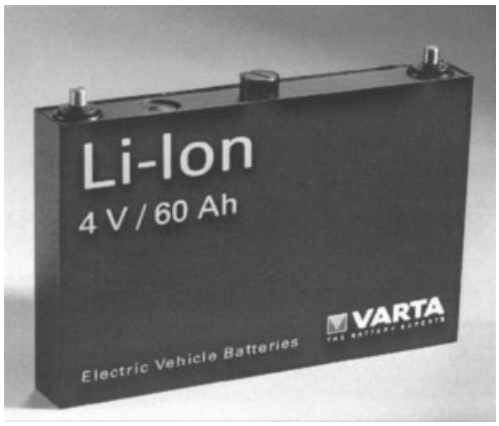

Figure 117. Prototype of a lithium-ion cell for electric vehicle application
Dimensions: width 30 mm; height 153 mm; length 210 mm; specific energy (at 25 W/kg continuously): 115 Wh/kg; energy density (C_5) 270 Wh/L (VARTA R&D Center, Kelkheim [218])

On account of the high energy content, batteries of such cells can only be operated in connection with battery-management systems that control each cell.

7.2.2. High-Temperature Batteries

The high-temperature batteries make use of a salt melt or of β-alumina, a solid electrolyte that gains its conductance by the mobility of sodium ions. In both cases, temperatures between 300 and 400 °C are required to reach sufficient conductivity of the electrolyte.

The high operating temperature of these batteries is a drawback, since it requires expensive thermal insulation and temperature management. But the performance data are attractive as shown in Table 19.

Table 19. Rechargeable battery systems that operate at high temperatures

	System	Equil. voltage, V/cell	Stored energy		
			Theory, Wh/kg	Practice, Wh/kg	Wh/L
1a	Iron sulfide – lithium [a,e]	1.35	459	90 – 110	max. 200
1b	Iron disulfide – lithium [b,e]	1.73	490	150 – 170 [c]	max. 350 [c]
2	Sulfur – sodium [f]	2.1	795	90 – 120 [d]	max. 180 [d]
3	Nickel chloride – sodium (ZEBRA) [f]	2.6	719	80 – 100 [d]	max. 160 [d]

[a] Li(Al) – FeS. [b] Li(Al) – FeS_2 when the upper plateau is only used (Fig. 118). [c] Bipolar version. [d] Figures concern the whole battery including thermal insulation. [e] Operating temperature ≈ 450 °C. [f] Operating temperature temperature ≈ 300 °C.

The development of these three systems has greatly been favored by the search for advanced batteries to power electric vehicles. The low specific energy of lead-acid batteries limits the range of such vehicles to about 80 km, and this range can be more than doubled by a high temperature battery of the same weight, as a comparison between Tables 19 and 13 shows.

7.2.2.1. Iron Sulfide – Lithium Batteries

The development of this system started in 1968 and subsequently has been continued to an essential part at the Argonne National Laboratory. Originally, lithium and sulfur were tested as electrode materials. Later, lithium – aluminum alloys and metal sulfides have been used as active material in the negative and positive electrodes, respectively. A molten eutectic mixture of lithium chloride and potassium chloride is used as the electrolyte in the battery that is operated at about 450 °C.

Iron and sulfur form two compounds: iron sulfide (FeS) and iron disulfide (FeS_2, pyrite). In both iron is in the divalent state (Fe^{2+}). Iron disulfide belongs to the group of polysulfides (MeS_x), containing S_x^{2-} anions.

If iron sulfide is used as the positive electrode and a lithium – aluminum alloy as negative electrode, reduction of iron (II) and oxidation of lithium results in the overall reaction

$$FeS + 2\,Li(Al) \longleftrightarrow Li_2S + Fe + 2\,Al$$

Proceeding from left to right means discharge, and the electrode reactions are:

$$\text{Positive electrode } FeS + 2\,e^- \longrightarrow Fe + S^{2-}$$

$$\text{Negative electrode } 2\,Li + S^{2-} \longrightarrow Li_2S + 2\,e^-$$

When the battery is charged, both reactions are reversed. The open-circuit voltage of the cell reaction is $U^{o,s} = 1.64$ V.

Compared to FeS, the energy content of pyrite is slightly higher. In combination with a lithium – aluminum alloy, the (simplified) cell reaction results:

$$FeS_2 + 2\,Li(Al) \longleftrightarrow +2\,Li_2FeS_2 + 2\,Al$$

with the open circuit cell voltage $U^{o,s} = 1.73$ V. This reaction causes the "upper plateau" discharge curve in Figure 118. The formed intermediate phase Li_2FeS_2 can further be discharged according to

$$Li_2FeS_2 + 2\,Li(Al) \longleftrightarrow 2\,Li_2S + Fe + 2\,Al$$

This second step causes the lower plateau in Figure 118. But when this reaction occurs, FeS_2 can only be recouped by charging voltages above 2.5 V/cell. For this reason, the discharge of FeS_2 is usually restricted to the upper plateau (see [220, p. 464]).

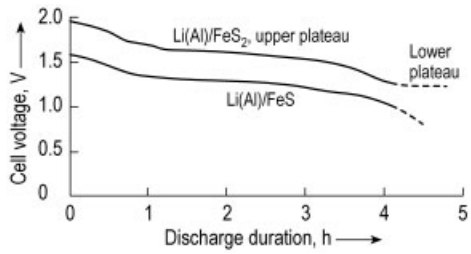

Figure 118. Constant-current discharge curves of iron sulfide – lithium batteries

Iron sulfide – lithium batteries have been designed as prismatic cells with an appearance similar to the nickel – cadmium battery shown

in Figure 86 (c.f. Fig. 39.1 in [199]). Felts of boron nitride or pressed powder sheets of magnesium oxide have been used as separators. The latter contain a limited amount of electrolyte, so that free liquid electrolyte is not present in the cell. The current collectors are made of steel for the negative electrode and also for the positive one, when FeS is used. In the FeS_2 battery molybdenum usually is applied as current conductor in the negative and the positive electrode. A ceramic seal is required for the post, usually the positive one, while the negative electrode is connected to the container.

The bipolar version represents the most compact design due to the minimized path length for the current. The principle is shown in Figure 119. Lithium – iron sulfide batteries in the bipolar design (cf. Fig. 39.11 in [199]) gave the excellent performance data shown in row 1 b in Table 19.

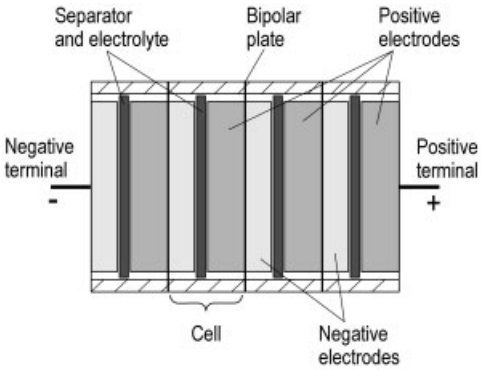

Figure 119. Bipolar battery design consisting of four cells. In the case of the iron sulfide – lithium battery, the separator is pressed MgO filled with the (starved) electrolyte, and the bipolar plates are molybdenum sheets. Special ceramics have been developed to achieve safe sealing of the bipolar plates and so prevent electronic or ionic contacts between the cells.

Despite enormous efforts by Argonne National Laboratories and a number of battery manufacturers for more than thirty years, and despite its excellent performance data, the iron sulfide – lithium battery did not yet reach a position in the market. Main problems are corrosion and high manufacturing and material costs.

7.2.2.2. Sodium – Sulfur Batteries

The start of an intense technical development of the sulfur – sodium battery system began at the Ford Motor Company's laboratory [219] and was strengthened as a reaction to the first oil crises in 1973. The idea was to make electric road vehicles more competitive by an increased driving range (for a detailed description see [220]). Research and development were continued by a number of groups, so on a large scale by Chloride in England and BBC (later ABB) in Germany [221]. The batteries consisted of a great number of identical cells in parallel strings (similar to the battery shown in Fig. 125). Practical experiments with a larger fleet of various vehicles were conducted for a number of years. However, Chloride and ABB discontinued their development in 1996. At present (2000), the system is further developed and tested for stationary applications in Japan by NGK Insulators Ltd. [222].

Sulfur – sodium and nickel chloride – sodium batteries are closely related to each other. Both employ β-alumina that simultaneously acts as the solid electrolyte and the separator. β-Alumina is a compound composed of aluminum oxide and sodium oxide. Their molar ratios ($Na_2O : Al_2O_3$) can vary between 1 : 1 and 1 : 11 (→ Aluminum Oxide, Chap. 1.6.). β-Alumina has the appearance of unglazed ceramic. At temperatures of about 300 °C, a sufficient high mobility of the sodium ions is achieved. Figure 120 shows the construction of a cell of the sodium – sulfur battery. Such cells manufactured by ABB had a capacity of 40 Ah per cell. Chloride used small cells of only 10 Ah per cell and a correspondingly higher number of cells per battery.

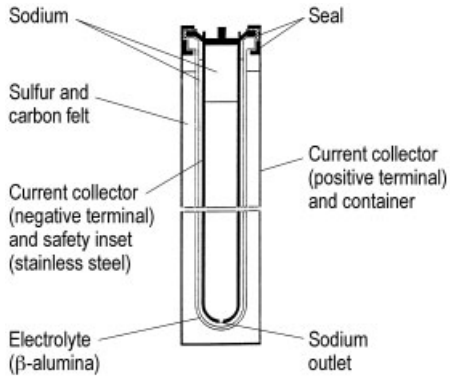

Figure 120. Cell of a sulfur – sodium battery (ABB)

The separator of β-alumina that simultaneously is the electrolyte, forms a cylindrical cup. Within this cup, the sodium is located. It is the active material of the negative electrode, that is molten at the operating temperature. The space between the β-alumina cup and the outer cell case of aluminum or mild steel is filled with sulfur, which is also liquid at this temperature. Sulfur is mixed with graphite felt, since sulfur itself does not conduct the electric current. In the β-alumina cup, a cup of stainless steel is inserted with a small hole in the bottom. This "safety inset" prevents an uncontrollable thermal situation in the case that cracks were formed in the ceramic separator and sodium and sulfur would get into direct contact (internal short-circuit). The small opening in the bottom of the safety inset limits the flow of sodium. Thus the rate of the reaction and also the heat generation are limited, and the diameter of the opening in the safety inset is designed so small that safe heat dissipation is possible also in the case of the internal short-circuit. The safety inset simultaneously acts as current collector.

When the cell is discharged, sodium ions migrate through the β-alumina wall and form NaS_x compounds with increasing sulfur content according to the sequence of reactions given in Table 20.

Table 20. Sequence of discharge – charge reactions of the sodium – sulfur cell

Charge $\longleftrightarrow$ Discharge	$U^{\circ}*$
$2\,Na + 5\,S \longleftrightarrow Na_2S_5$	2.08 V
$2\,Na + 4\,Na_2S_5 \longleftrightarrow 5\,Na_2S_4$	1.97 V
$2\,Na + 3\,Na_2S_4 \longleftrightarrow 4\,Na_2S_3$	1.81 V

$*$ U° values at 300 °C.

According to this sequence, the equilibrium cell voltage is slightly reduced with progressing discharge. Typical charge – discharge curves are shown in Figure 121. Since there are no secondary reactions, a steep increase of the charging voltage indicates that the state of complete recharge is reached. Thus, in a series-connected string of cells, the voltage across the cell that first reaches full charge may attain very high values, and damage of the β-alumina may be the consequence. As a remedy, a sulfur – sodium battery always contains a number of paralleled strings of cells with cross connections.

With rising temperature, the voltage difference between charging and discharging decreases, i.e., the efficiency of the battery is improved. Since there are no secondary reactions, the coulomb efficiency (charged/discharged Ah) is 100 %, and only the voltage gap between charging and discharging causes that the energy-storage efficiency (charged/discharged Wh) is lower than 100 %. Since this gap is reduced with rising temperature, the Wh efficiency is increased with temperature and reaches values up to 90 %.

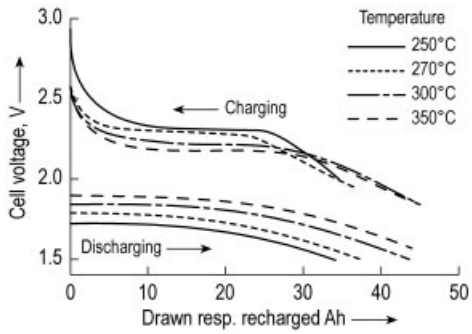

Figure 121. Charge/discharge curves of a sulfur – sodium battery at various temperatures [223]

The design of a complete battery including the thermal insulation is similar to that shown in Figure 125 for nickel chloride – sodium batteries.

The main problems of sulfur – sodium batteries are:

- Internal short circuits, caused by overcharging or hot spots (sodium dendrites). The failed cell attains a high internal resistance so that the whole string is blocked.
- Mechanical defects (separator or cell seal) caused by thermal cycles (too rapid heating or cooling).

7.2.2.3. Nickel Chloride – Sodium (ZEBRA) Batteries

The nickel chloride – sodium battery also uses sodium for the negative electrode and β-alumina as separator and solid electrolyte. Nickel chloride ($NiCl_2$) is used as active material for the positive electrode that is reduced to the metal when discharged. Both nickel chloride and nickel metal remain in the solid state despite the high

operating temperatures of the battery. For this reason, these batteries must use an intermediate electrolyte which is a molten salt, namely sodium aluminum chloride ($NaAlCl_4$). This intermediate electrolyte involves some benefits, described below.

The development of this battery system was started under the name "ZEBRA battery" by the Anglo African Oil Company in South Africa (code for the project was the abbreviation **ZE**olites applied to **B**attery **R**esearch **A**frica [220, p. 380]) and continued in cooperation with AEG Anglo Batteries GmbH, Ltd. in Germany and England. Although a large number of such batteries were being tested in vehicles and proved quite successful [224], it was decided to discontinue the development at the end of 1998.

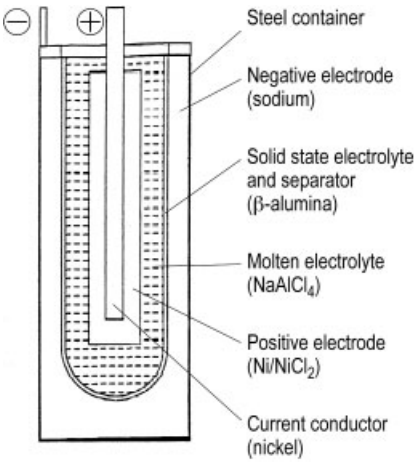

Figure 122. Vertical cross section of a nickel chloride – sodium cell

In principle, an iron chloride – sodium electrode can also be used as the positive electrode and has also been tested as well as mixtures of iron and nickel. Iron would be attractive as a less expensive material offering an only slightly reduced cell voltage ($U^o = 2.35$ V/cell). But iron proved to have a number of drawbacks, and the later development concentrated on positive nickel electrodes (cf. [220, p. 387]).

The construction of a nickel chloride – sodium cell is shown in Figure 122. It looks similar to a sodium – sulfur cell, but here sodium fills the volume between the β-alumina cup and the cell container, while the positive electrode, the $Ni/NiCl_2$ electrode, is located in the center. Molten $NaAlCl_4$ forms the ionic contact between the solid positive electrode and the solid electrolyte.

The nickel electrode is produced by sintering a fine powder of nickel tetracarbonyl together with sodium chloride and aluminum. The result is a structure of high porosity.

The intermediate electrolyte ($NaAlCl_4$) involves some benefits. When the cell is overcharged, the intermediate electrolyte reacts with the active material of the positive electrode according to the reaction:

$$2\,NaAlCl_4 + Ni \longrightarrow 2\,Na + 2\,AlCl_3 + NiCl_2$$

which has the equilibrium cell voltage $U^o = 3.05$ V. On account of this high equilibrium voltage, this reaction does not occur during the normal charging process. If the battery has been overcharged, this reaction is reversed during the subsequent discharge.

When the battery is overdischarged, the intermediate electrolyte reacts with sodium according to

$$NaAlCl_4 + 3\,Na \longrightarrow Al + 4\,NaCl$$

with the equilibrium cell voltage $U^o = 1.58$ V. This reaction is reversed when the battery is recharged. Thus, the intermediate electrolyte allows overcharging and overdischarging to a certain extent. Furthermore, the molten $NaAlCl_4$ bridges the cell in the case of failure. A string of cells is so kept workable despite of failed cells, although at a reduced string voltage.

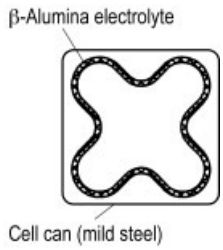

Figure 123. Horizontal cross section of a nickel chloride – sodium cell [225]

In the latest design, a cloverleaf shaped β-alumina cup was used instead of the simple cylinder, and a prismatic cell container instead a cylindrical one, as the horizontal cross-section in Figure 123 shows. (In Fig. 125 the older cylindrical cell containers are shown). The surface of the separator was significantly increased by this

improved design, and the prismatic container makes better use of the battery volume.

The electrochemical reactions are:

		Charge $\longleftrightarrow$ Discharge
Positive electrode	$NiCl_2 + 2\,Na^+ + 2\,e^-$	$\longleftrightarrow Ni + 2\,NaCl$
Negative electrode	$2\,Na$	$\longleftrightarrow 2\,Na^+ + 2\,e^-$
Cell reaction	$NiCl_2 + 2\,Na$	$\longleftrightarrow Ni + 2\,NaCl$

with the equilibrium cell voltage $U^{o,s} = 2.58$ V. Figure 124 shows typical discharging and charging curves. The decreasing voltage near the end of discharge indicate the increasing internal resistance.

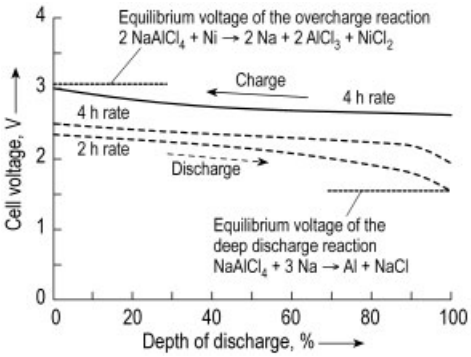

Figure 124. Typical voltage curves for the discharge of a ZEBRA battery at the 4-h and 2-h rate, and recharge at the 4-h rate. Charging and discharging occurs with constant current.

The equilibrium potentials of the overcharge and overdischarge reactions are inserted in Figure 124. They show that the cell voltage will not significantly exceed the normal range, when the cell is overcharged or deep discharged.

Figure 125 shows a complete battery and its construction. The battery consists of a great number of cells, which stand upright side by side. The cells have metallic containers and are insulated from each other by foils of mica. The cell assembly is encapsulated by a specially developed container of excellent thermal insulation, including vacuum. For temperature control, cool air can be blown through the interior. Electrical heating elements are inserted for heating up the battery to the minimum value of 200 °C where the charging process can be started. These heating elements are also used to keep the battery hot, if it is not used for a certain period of time.

8. Battery Standards

8.1. General

A standard is an item considered by general consent as a basis of comparison, if it is in compliance with the requirements of relevant directives and guidelines. A technical standard may relate, e.g., to shape and dimensions. In the case of a battery, the standard will also refer to the electrochemical system employed, its electrical performance characteristics, its terminal arrangements, and to its unambiguous identification, derived by a unique nomenclature system.

A technical standard is expected to

- foster international trade, and/or
- protect the consumer, and/or
- assure interfacing, interchangeability compatibility or interworking, and/or
- improve safety and health, and/or
- protect the environment.

8.2. National Standards

Before International Standards became available, National Standards for battery systems were established by National Committees, working independently from each other. National Standards are for instance ANSI (American National Standards Institute), BSI (British Standards Institute), DIN (German Industrial Standard), JIS (Japanese Industrial Standard), etc.

With the growth of international trade there was a need to coordinate the activities of these National Committees. Within the European Community this coordination is done by CENELEC [227]. CENELEC works closely together with the IEC and is expected to become one day the only European Standards Organization, superseding all National Committees of the European Community.

During the last two decades quite a few National Committees have started to harmonize their Standards with IEC Standards, especially for primary batteries. In the meantime the majority of National Committees have been fully harmonized. Nevertheless, some national and/or regional standards are still in existence, and will continue to be part of their country's legislation.

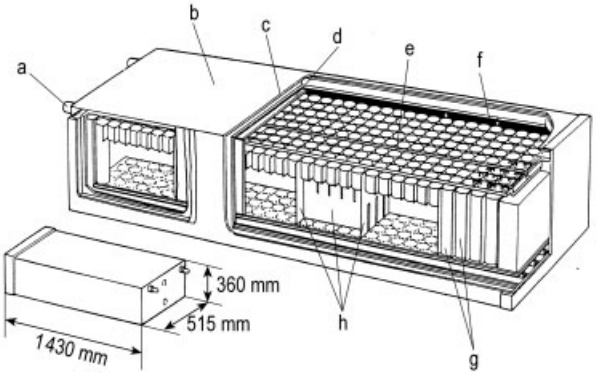

Figure 125. Construction of a nickel chloride – sodium battery (from [226]) In the latest design, the cylindrical cell containers are substituted by prismatic ones (see Fig. 123). The energy content of such a battery is about 30 kWh.
a) Terminal; b) External container wall; c) Thermal insulation (vacuum); d) Internal container wall; e) Heating element; f) Electrical insulation; g) Cell; h) Electrical insulation between cells

8.3. International Standards

8.3.1. Technical Committees and how they operate

In consequence of the growing needs to coordinate the activities of National Committees, two Technical Committees on batteries were set up under the auspices of the IEC [228]. The first IEC Technical Committee on batteries that was established in 1931 was Technical Committee No. 21 (TC21). It focusses solely on lead – acid batteries. In 1965 a Sub-Committee SC21A was added, dealing with "*Secondary Cells and Batteries, containing alkaline (e.g. Nickel-Cadmium batteries) or other non acid electrolytes (e.g. Lithium-Ion batteries)*". In 1949 a second IEC Technical Committee was set up, dealing solely with primary batteries, i.e., Technical Committee No. 35 (TC35).

Meanwhile more than 100 IEC Technical Committees and Sub-Committees are in existence, covering the whole field of electrotechnical components, devices, and machinery. Each Technical Committee and Sub-Committee is guided by a Chairman and a Secretary, being elected according to the "Procedures" of the IEC Directives Part 1 [229]. These procedures also clarify the rights and duties of National Committees, Chairman, and Secretary when carrying out IEC work. National Committees, Chairman, and Secretary are in close contact with each other, and with the Central Office of the IEC, which provides technical and administrative services to the Technical Committees.

The IEC Technical Committees are operated by consultation and discussion; the participants being experts appointed by their National Standards Organization. When a point of agreement is confirmed by a formal postal vote, it is incorporated into the IEC Standard.

A new standard starts always with a "New Work Proposal" that may be submitted by any National Committee or by the Secretary himself. It depends on National Committees whether or not this New Work Proposal is approved and whether or not it is finally processed to an International Standard. The standardization procedure takes in general three to four years.

At present, well above 25 countries are participating in the work of Technical Committees No. 21 and No. 35. IEC Standards are voluntary standards, unless they are part of a country's legislation.

In order to progress the IEC work, the experts of a Technical Committee have to meet regularly. The meeting may be a Working Group meeting or a meeting of the Technical Committee, which has the authority to take decisions on new standards and future work (strategic policy). The meeting place is determined by the inviting National Committee and may be at any place of the world. This means travel expenses for the experts, which are not covered by their National Committee but by the company the expert is associated with. Experts of TC21, SC21A

and TC35, are prevailingly associated with battery companies.

Part 2 of the IEC Directives cover the "Methodology for the development of International Standards" [230], its principles and requirements to be observed by Technical Committees and their subsidiary bodies. Part 2 is considered to be the most important part of the IEC Directives, since it has a direct impact on the quality of the IEC Standard. In contrast to Part 1 of the Directives, the contents of Part 2 and its application to standards is not enforced by an official IEC body. This is left to the individual Technical Committee, which is required to comply with the principles and rulings of the Directives, Part 2. In practice, however, international consensus is given the first priority, which is not always based on the requirements of the IEC Directives, Part 2.

Part 3 of the IEC Directives contains the "Rules for the structure and drafting of International Standards" [231].

When voting is required, each National Committee has one vote only, independent of the size of the population it represents.

8.3.2. TC21, SC21A , TC35, and Their Standards

The IEC Standards on Batteries are listed in Section 8.4 ((Annex 1)). Since the Standards of TC21, SC21A, and TC35 were set up at different times and by different experts, their contents differ in structure. The contents may, however, in general be allocated to following subjects

1) Scope
2) Normative references
3) Definitions
4) Requirements
5) Designation and marking of cells and batteries
6) Dimensions
7) Performance testing
8) Test conditions
9) Specification tables and sheets
10) Product relevant annexes (normative and informative)

The contents of Standards on Battery Safety may be allocated to following subjects

1) Scope
2) Normative references
3) Definitions
4) Requirements for safety
5) Instruction for use
6) Testing and requirements
   Intended use
   Reasonably foreseeable misuse
7) Safety precautions during handling
8) Safety precautions during packaging, storage, transportation, and disposal
9) Marking
10) Annexes (normative and informative)

IEC standards are under constant review, thus ensuring that their contents reflect current technology. When using the IEC Standard it must be ensured that its latest edition is consulted.

8.3.3. TC21, SC21A, and TC35: Harmonization of Standards

A comparison of the standards of TC21, SC21A, and TC35 has revealed that, due to the independent work of the committees, their standards have arrived at different results on some similar subjects. In this respect two areas were identified:

1) Definitions as presented in the relevant IEV vocabularies
2) Nomenclature system for batteries as presented in the relevant standards

Consequently two joint working groups (experts from both Technical Committees) were set up in recent years with the task to harmonize the standards and to resolve unnecessary differences.

8.3.3.1. IEC Vocabulary on Batteries

The harmonizing work was based on existing IEC vocabularies, i.e.,

1) IEV 60050-481: Primary Cells and Batteries [228]
2) IEV 60050-486: Secondary Cells and Batteries [228]

The task was to amalgamate both vocabularies. Since the Joint Working Group could not resolve the discrepancies the project was transferred to IEC Technical Committee No. 1, which has specialized on IEC definitions.

8.3.3.2. Harmonization of the Nomenclature Systems

Three IEC nomenclature systems plus variants are presently in use. They were developed by TC21, SC21A, and TC35 at a time when no IEC Directives were available. Consequently none of them was developed in compliance with the IEC Directives, Part 2, section 7, annex E: Designation of Standardized Items [230].

The task of the joint working group was therefore to harmonize existing nomenclature systems and/or to derive a completely new unique and logical nomenclature system for all batteries, being in compliance with the IEC Directives, Part 2 and its principles, and being suitable for data processing. A proposal of this kind [232] that initially found favorable support finally failed. An agreement on a different proposal [233] could not be reached either.

As a result the task was abandoned. This outcome is regrettable, not only for the consumers for reasons given below (cf. Section 8.5 and 8.6) but also in regard to the technical professionalism of this International Standard.

8.3.3.3. Designation of Batteries, Present Status

Since batteries have been available on the market, consumers have been confronted with battery designations of different origin, i.e.:

1) Manufacturer's designation
2) Designations according to National Standards
3) Designations according to IEC Standards

and mixtures thereof.

The present status is illustrated in Table 21. Table 21 lists 15 batteries arbitrarily picked from a range of button batteries, cylindrical batteries, and prismatic batteries of different primary and secondary electrochemical systems. Each row contains designations of batteries being identical in dimensions and electrochemical system. All batteries belong to the group of portable batteries.

The left part of Table 21 lists the battery type, its dimensions, followed by IEC codes, i.e., nominal voltage/system letter and the battery's IEC designation.

The right part of Table 21, as given by columns 1 – 6 contain manufacturers designations. The designations listed are used by well-known international battery companies. Some of them have started to use IEC designations instead, especially in Japan.

The confusing number of different designations for batteries, being identical in size and system, speaks for itself. A more detailed discussion of the contents of Table 21 is given in Section 8.5 ((Annex 2)).

Former ANSI designations are still in use, although having been no more valid for a long time. The more popular ones of these designations are listed in Table 22. Table 22 also contains equivalent IEC size designations. Originally, the former ANSI designations addressed carbon-zinc batteries with different dimensions, as given in Table 22. Today former ANSI codes are still in use as size codes, e.g., in combination with IEC system- and shape-letters like CR1/2AA.

8.3.3.4. IEC Nomenclature Systems for Batteries

In general, presently valid IEC nomenclature systems address two aspects of a battery:

1) Physical aspects such as shape, dimensions, and terminal arrangements
2) Electrical aspects such as voltage, electrochemical system, capacity, load capability and type of electrode (pocket plates, sintered plates etc.)

The information given in the designation, has in general the following sequence:

1) Voltage: Either in clear or by code, i.e., number of cells connected in series in combination with a system letter, to which the system's nominal voltage is allocated.
2) The system letter(s) denotes the electrochemical system in the sequence (see Section 8.7 ((Annex 4)) for details): positive electrode – electrolyte – negative electrode

Table 21. Designations of Batteries being identical in Dimensions∗ and Electrochemical System

No	Battery type	Dimensions, mm			IEC	Battery designations as used by different international manufacturers					
		Length	Width or Height diameter	Norm. voltage/syst. letter(s)	Designation	1	2	3	4	5	6
	Button, primary										
1	Silver oxide-zinc		7.9	2.6	1.55 V/S	SR59	397	SR726SW	D397	BSR59L	V397	26
2	Alkaline air-zinc		11.6	5.4	1.4 V/P	PR44	V675AT	PR44	AC675E	675HPX	PR675	L675ZA
3	Manganese dioxide-lithium		20	1.6	3 V/C	CR2016	CR2016	CR2016	DL2016	ECR2016	CR2016	CR2016
	Button, secondary										
4	Nickel-metal hydride		11.6	5.5	1.2 V/H	HB116/055	V40H				
5	Manganese dioxide-lithium alloy		20	3.2	3.6 V/LM	LM2032	ML2032	(MT2032)	ML2032	(VG2032)	
	Cylindrical, primary										
6	Carbon-zinc		14.5	50.5	1.5 V/none	R6S	M15P	7AA	R6KG	1015	R6P(NR)	2006
7	Manganese dioxide-lithium		14.5	25	3 V/C	CR14250	CR1/2 6L	CR14250SE	CR1/2AA			
8	Manganese dioxide-lithium		17	34.5	3 V/C	CR17345	CR123A	K123LA	DL123A	EL123AP	CR123A	CR123A
	Cylindrical, secondary										
9	Nickel-cadmium		14.5	50.5	1.2 V/K	KR15/51	VRE AA600R	NC15	N-600AA	CH15	P-60AA	700 RSE
10	Lithium-ion		18.8	65.2	3.6 V/IC	ICR18650	UR18650	CGR18650	US18650	ICR18650	LGR18650	ICR18650
	Prismatic, primary										
11	Alk. manganese dioxide-zinc	26.5	17.5	48.5	9 V/L	6LR61	AL-9V	PC1604	6LF22G	EN22	GP910A	4022

∗ Within given tolerances

Table 21. (Continued)

No	Battery type	Dimensions, mm				IEC	Battery designations as used by different international manufacturers					
		Length	Width or diameter	Height	Norm. voltage/syst. letter(s)	Designation	1	2	3	4	5	6
12	Manganese dioxide-lithium	34	17	45	6 V/C	2CR5	2CR5	2CR5	DL245	EL2CR5	2CR5M	2CR5
12a						(CP3845)						
	Prismatic, secondary											
13	Lead acid, valve regulated	70	48	102	6 V/none	6P40	PE6V4	240 703	NP4-6W			
14	Nickel-metal hydride	17.3	6.1	48.2	1.2 V/H	HF18/07/49	HF-B1	HP6	VH600F6	HHF60S		
15	Lithium-ion	34	8	48	3.6 V/IC	ICP340848	US083448	LP9	LGQ863448C	CGP34488	LGR083448	MP9

* Within given tolerances

3) Shape letter: B for button batteries, R four round batteries (including button and cylindrical shape), P for prismatic and non-round batteries.
4) Size: Either by code or by dimensions. Since the battery volume correlates with its capacity, reference is sometimes made to the capacity instead.
5) Load capability: L, M, H and X for low, medium, high and very high rate for nickel-cadmium batteries. For primary batteries no load-classification is available.
6) Contact and performance variants.

It would be beyond the scope of this article to present all three IEC nomenclature systems including their variants in full length. For further information please refer to the IEC Standards, listed in Section 8.4 ((Annex 1)). A short characterization of the different designation systems is given in Section 8.6 ((Annex 3)).

The problem of battery designation was already highlighted in 1990 by an article in a consumer journal entitled "Big Confusion in Battery Nomenclature" [234].

To the benefit of the consumers, the manufacturers, and the trade, it is hoped that future experts of the IEC will be able one day to take on the work again, i.e., to derive a unique and logical nomenclature system for all batteries, being in compliance with the IEC Directives, Part 2 and its principles and being suitable for data processing, as anyway required by today's information technology.

Furthermore it would be important especially for primary batteries to include into the nomenclature systems a well defined code, denoting the system's load capability. This matter is not only related to fit and function of the battery but also to safety. A relevant proposal being based on extensive experimental testing [235] was not supported. Presently the category of rate capability is not included in the designation, i.e., a difference of a factor 100 in load capability is not indicated on the primary battery. All that is indicated are system, voltage, and dimensional characteristics. This is in contrast to the IEC designation system for sealed nickel – cadmium batteries.

Table 22. Cylindrical batteries, popular size designations of ANSI and IEC

	Dimensions, mm		Size designations				
			Primary batteries			Secondary batteries	
No	1 Diameter	2 Heigth	3 Present ANSI	4 Former ANSI	5 IEC	6 IEC	7 IEC
1	10.5	44.5	24	AAA	R03	R03	R11/45
2	12	30.2	910	N	R1		R12/30
3	14.5	50.5	15	AA	R6	R6	R15/51
4	26.2	50	14	C	R14	R14	R26/50
5	34.2	61.5	13	D	R20	R20	R33/62

8.3.4. Closing Remark

Nevertheless, in spite of the deficiencies considerable progress has been made especially in areas as the standardization of new electrochemical systems and batteries and the issue of new IEC Standards on Battery Safety. Since IEC Standards are under constant review, it is hoped that existing discrepancies and nonconformities with the Directives, Part 2 will be eliminated in the foreseeable future.

8.4. Annex 1: IEC Standards on Batteries

Standards of TC21: Lead-Acid Batteries.

1) IEC 60095: Lead – Acid Starter Batteries
2) IEC 60254: Lead – Acid Traction Batteries
3) IEC 60896: Stationary Lead – Acid Batteries
4) IEC 61056: Portable Lead – Acid Cells and Batteries

Standards of SC21A: Secondary Cells and Batteries containing Alkaline or other Non-acid Electrolytes.

1) IEC 60285: Sealed Nickel – Cadmium Cylindrical Rechargeable Single Cells
2) IEC 60509: Sealed Nickel – Cadmium Button Rechargeable Single Cells
3) IEC 60622: Sealed Nickel – Cadmium Prismatic Rechargeable Single Cells
4) IEC 60623: Vented Nickel – Cadmium Prismatic Rechargeable Single Cells
5) IEC 60952: Aircraft Batteries
6) IEC 61150: Sealed Nickel – Cadmium Rechargeable Monobloc Batteries in Button Cell Design
7) IEC 61434: Secondary Cells and Batteries with Alkaline or other Non-acid Electrolyte: Rules for the Determination of the Current according to the Standards for Alkaline Secondary Cells and Batteries
8) IEC 61436: Secondary Cells and Batteries Containing Alkaline or other Non-acid Electrolytes: Sealed Rechargeable Nickel – Metal Hydride Single Cells
9) IEC 61440: Secondary Cells and Batteries Containing Alkaline or other Non-acid Electrolytes: Sealed Small Prismatic Rechargeable Nickel – Cadmium Single Cells
10) IEC 61808: Secondary Cells and Batteries Containing Alkaline or other Non-acid Electrolytes: Sealed Rechargeable Nickel – Metal Hydride Single Button Cells
11) IEC 61809: Secondary Cells and Batteries Containing Alkaline or other Non-acid Electrolytes: Safety Requirements for Portable Sealed Alkaline Secondary Cells and Batteries
12) IEC 61951: Secondary Cells and Batteries Containing Alkaline or other Non-acid Electrolytes: Portable Sealed Rechargeable Single Cells
13) IEC 61960: Secondary Lithium Cells and Batteries (This Product Standard includes also Safety Requirements).

Standards of TC35: Primary Batteries.

1) IEC 60086-1: Primary Batteries, General
2) IEC 60086-2: Primary Batteries, Physical and Electrical Specifications
3) IEC 60086-3: Watch Batteries
4) IEC 60086-4: Safety of Lithium Batteries
5) IEC 60086-5: Safety of Batteries with Aqueous Electrolyte

8.5. Annex 2: Designations by Manufacturers (Table 21)

Primary Batteries.

Rows 6 and 11. When looking at the designations given in rows 6 and 11 of columns 1 to 6, they all look different. None of them complies with the IEC designation. It is difficult even for the expert to recognize that all of them refer to the same type of carbon-zinc battery (row 6) and the same type of alkaline 9 V battery (row 11).

Row 3. Much easier is the case given in row 3. In most cases the designation CR2016 is used, which complies with the IEC designation for primary batteries, being standardized after October 1990. It actually was taken over from JIS just by agreement. For explanation: C means the electrochemical system (manganese dioxide – lithium), letter R is the shape letter (round), 20 means the diameter in millimeters and 16 denotes the height of the battery in tenth of millimeters. The designation CR2450 of the same series (however, not shown in Table 21) is interpretable on the same basis. In this case it must be known that 24 does not mean 24 mm but 24.5 mm.

Row 12 follows essentially also IEC recommendations, however, with exceptions. The correct designation following from the designation system, is given in row 12a. By way of exception the well-known designation 2CR5 was taken over.

Row 2 refers to an air – zinc button battery. This IEC designation is rarely reflected at all. The number 675 is used more frequently, which has to do with the tradition to designate a hearing aid button battery of this size (11.6 × 5.4 mm) with code no. 675 with preceding and/or succeeding letters, depending on manufacturer.

Row 8 is similar to row 2 in that the IEC designation for the high-rate cylindrical battery, i.e., CR17345 is rarely used. In some cases the first two letters CR are applied. As size code the code 123A is used, with preceding and/or succeeding letters, depending on manufacturer.

Row 7 refers to a cylindrical low-rate manganese dioxide – lithium battery, having a diameter of 14.5 mm and a height of 25 mm. The designation method is identical to the IEC method according to row 3. Some manufacturers, however, prefer a different size designation like, e.g., 1/2AA or 1/2 6L. The code AA is taken from the former ANSI standard (see Table 22) which refers to a battery size with 14.5 mm diameter and a height of 50.5 mm. The code 1/2AA refers to a height of 1/2 × 50.5 mm = 25 mm. The same is true for the size code 1/2 6, which refers to a R6 sized battery with letter L as suffix.

Secondary Batteries (Table 21). Like in the case of primary batteries, a logical background for the designations of secondary batteries can not be identified, apart from a few exceptions.

Rows 5 and 10. Row 5 refers to lithium-alloy button batteries. With the exception of the two-letter codes, their size code 2032 is identical to the IEC standard (see row 3 for explanation). Letter codes MT and VG may refer to different electrochemical systems.

Row 10 is an exception since the IEC size code 18650 is used by all manufacturers, however, not the IEC system and shape code ICR.

Row 9 is an example where none of the designations comply with the IEC designation. Some of the designations use size code AA (see row 7 for explanation).

Row 15 offers another case, where even the expert has difficulties to recognize that the manufacturers designation refer to the same type of lithium-ion battery.

8.6. Annex 3, Examples of IEC Nomenclature Systems for Batteries

IEC Nomenclature System 1. Example 1:The designation 6P40 (see IEC 1056-2, 1st ed. 1994) refers to a 6 V (6) prismatic lead acid battery (P), rated capacity $\underline{4}$ Ah ($\underline{40}$/10 Ah). A system letter is not used. Instead of battery dimensions or volume, the capacity is given.

Variant of IEC Nomenclature System 1. Example 2:The designation KCSH 15 (see IEC 60622, 2nd ed., 1988) refers to a sealed nickel – cadmium (K) prismatic single cell (C), employing sintered electrodes (S) for high rate (H) discharge, and a rated capacity of 15 Ah (15). The designation does not contain a shape letter. The rated capacity is given instead of battery dimensions or volume. The system letter K for nickel – cadmium stands for 1.2 V.

Attention is drawn to the deviations in structure and information, in comparison to example 1.

IEC Nomenclature System 2. Example 3:The designation KB116/055 (see IEC 509 2nd ed., 1988) refers to a nickel – cadmium (K) button (B) battery, having a diameter of 11.6 mm and a height of 5.5 mm. System letter K with no preceding figure means 1.2 V.

IEC Nomenclature System 3. Example 4:If the designation system according to IEC 6086-1, 1997 would be applied (nomenclature valid before October 1990) the designation of the nickel-cadmium battery in Example 3 would read KR44, letter K means nickel – cadmium, letter R means round and size code 44 denotes a battery having a diameter of 11.6 mm and a height of 5.4 mm.

Example 5:If the nomenclature system according to IEC 6086-1, 1997 (nomenclature valid after October 1990) would be applied, the designation for the same battery would change to KR1154. This size code is a variant of example 3: Figure 11 refers to the rounded down value of 11.6 mm and Figure 54 refers to the height of the battery in tenth of mm.

Example 6:The designation ICR18650 (see IEC 61960-1, 1st ed.) refers to a cylindrical (R), cobalt based (C), lithium-ion (I) secondary cell, having an approx. diameter of 18 mm and an approx. height of 65 mm. This example is a variant of example 5 in that two system letters (IC) are employed. The size code is derived on identical grounds.

Example 7:The designation ICR20/1050 (see IEC 61960-2) refers to a cylindrical (R), cobalt based (C), lithium-ion (I) secondary cell, having an approx. diameter of 20 mm and an approx. height of 105 mm. The slash (/) shall indicate, that the cell height is above 100 mm.

Example 8:The designation HF18/09/49 (IEC 61436, 1998) refers to a sealed nickel – metal hydride (H) prismatic single cell battery, designated as flat (F), having a width of 17.3 mm, a thickness of 8.3 mm and a height of 48.2 mm.

Example 9:When applying the dictum of IEC 61960-2, 1st ed., the designation of the battery in example 8 would read HP081748 (note: the size code of example 9 uses the sequence thickness, width, and height, whereas in example 8 the sequence width, thickness, height is used. Furthermore approximate dimensions are used in example 9 in contrast to example 8, which uses rounded up values with shape letter P instead of F.

Example 10:When applying the teaching of IEC 6086-1 (section A2) to the designation of example 9, it would read HP1948. The last two characters denote the height of 48.2 mm. The first two digits addresses the diagonal of the prismatic battery, as given by thickness and width and calculated via the Pythagorean theorem (integer value of 19.19 mm = 19 mm).

In conclusion: The different IEC nomenclature systems and their variants are actually not justified by technical reasons. They are non-uniform and can hardly be used for modern data processing. They are in general too long and therefore not consumer-friendly. Example: The designation LR61 of a small cylindrical alkaline battery has changed by now to LR8D425.

8.7. Annex 4, Denoting Electrochemical Systems

When addressing an electrochemical system according to the International Standards (see Section 8.4 ((Annex 1)), the positive electrode is addressed first, followed by the electrolyte and concluded with the negative electrode:

8.8. Example 1:

The manganese dioxide – ammonium chloride – zinc system

8.9. Example 2:

The nickel hydroxide–potassium hydroxide–metal hydride system

Since the correct designation is in general too long, popular abbreviations have come into use like the carbon – zinc system, the nickel – metal hydride system, the lead – acid or the lithium-ion system, etc.

A more correct abbreviation would be the designation by the system's formula:

	Designation by formula	Abbreviated formula
Example 1:	$MnO_2 - NH_4Cl - Zn$	$MnO_2 - Zn$
Example 2:	$NiOOH - KOH - MeH$	$Ni - MeH$

When naming only the positive and the negative electrode of the system, the positive should be named first and the negative second. For reasons of consistency this style was applied throughout this article.

In order to avoid ambiguities the type of electrolyte needs sometimes to be added like in the case of the $MnO_2 - Zn$ system, which also exists as alkaline version.

9. Battery Disposal, Environmental Protection

The growing attention that nowadays is paid to environmental protection also raises the question how to handle batteries after the end of their service life, since they often contain heavy metals and toxic substances. Run-down larger stationary and motive power batteries are usually replaced by a manufacturer and directed to a recycling plant to regain the valuable raw materials. This also applies to most automotive lead-acid batteries that were recycled to more than 96.5 % in the USA already in 1996 [236]. The situation, however, is different for the extremely fast growing market of small batteries which often are disposed of as municipal waste. Then their toxic constituents can be released into the environment from landfills or incinerators. This concerns more or less all battery systems.

A tremendous amount of primary batteries is used mainly in private homes, e.g., in flash lights, cameras, alarm clocks, and similar applications. This group includes button cells for primary use, e.g., in hearing aids, watches, and small calculators. Most of these batteries, however, do not contain heavy metals or other critical substances and therefore are not subject to official regulations in many countries.

Nickel – cadmium, Nickel – metal hydride, and lithium-ion batteries are increasingly connected to our daily life, for example in portable tools, such as drilling machines or electrical screw drivers, or other equipment, like mobile phones and notebooks. They exist as multicell "power packs", but also as single cells of cylindrical and prismatic shape or as button cells.

Small valve-regulated lead – acid batteries are used as standby batteries in household-fire alarm or similar security systems.

The problem with these small batteries is not the recycling itself, since suitable reprocessing methods and plants are available for most of them. The problem is the collection of these batteries, because they are indeterminably distributed in almost every home. Thus it would not have been sufficient to establish legislative regulations for recycling, but simultaneously effective collecting and sorting systems had to be installed to direct the various types of batteries to their corresponding recycling plant.

Parallel to these attempts, the industry made great efforts to reduce the toxic constituents in the batteries and to reduce the potential of environmental hazards. Thus today almost all primary batteries do no longer contain mercury, and an important incentive for the development of nickel-metal hydride batteries was the substitution of cadmium in the nickel-cadmium battery by a less toxic negative electrode.

9.1. Programs to Collect Disposed Batteries

The programs to handle disposed batteries are different in the various countries. The programs always include two aspects:

1) Legal regulations that concern the handling of hazardous waste
2) Programs to collect and sort the different batteries

Often the regulations have to be modified, when they do not allow the accumulation of hazardous waste, what necessarily is connected to the collection of batteries, containing substances that are defined as hazardous. In the USA, for example, this problem has been solved by the Resource Conservation and Recovery Act (RCRA) of 1976 (40 CFR) that regulates the handling of hazardous waste such as cadmium, but excludes residential waste. Thus batteries, discarded by households, are no longer subject to this regulation. This regulation was modified in 1995 (40 CFR, Part 273), and in 1996 President Clinton signed into law the so-called Battery Act

(*Mercury Containing and Rechargeable Battery Management Act*) that facilitated the collection of such batteries. The rechargeable types at that time were nickel – cadmium and valve-regulated lead – acid (VRLA) batteries. In these regulations, the latter are always designated as SSLA batteries (sealed small lead – acid batteries) [237]. Besides these general rules, details are handled different by the various states of the USA.

To solve the second problem, the collection of the batteries, in 1991 five major rechargeable battery manufacturers funded the Portable Rechargeable Battery Association (PRBA) to establish recycling programs and educational programs in the USA [238]. For the collection of the batteries the Rechargeable Battery Recycling Corporation (RBRC) was funded by the industry as an independent, nonprofit public service organization.

In Europe, the European Portable Battery Association (EPBA) established rules that act as a guide for all battery manufacturers in the EC member countries. In the meantime legislative regulations have been passed for the handling of disposed batteries everywhere, although these regulations vary in the different countries. As an official document, the EC directive 91/157 was adopted by the Environmental Counsel in 1991. It prescribed the collection of all batteries that contain lead, cadmium, or more than 0.025 % of mercury [239]. In Germany 1998 a more rigid decree came into force demanding that all spent batteries have to be collected, sorted, and recycled or disposed. In the same year a common collection system was established and the corresponding foundation (*Stiftung gemeinsames Rücknahmesystem, GRS*) founded in Hamburg as a nonprofit organization (details in [239]). Members are the main battery manufacturers and the Zentralverband der Elektrotechnik und Elektroindustrie e.V. (ZVEI) as the representative of the producers of battery-powered appliances. [A collecting system under the name ARGE – BAT (Arbeitsgemeinschaft Batterien) had been founded already in 1988, but failed since the importers refused to participate.]

All batteries are collected in special containers that are positioned in shops and public areas. The batteries are sorted and transported to recyclers or they are properly disposed. The total costs are estimated to 40 to 30 $\times$ 10^6 € per year which corresponds to about 0.026 € for a typical cell of the AA size [239].

Similar collecting systems have been established in a number of European countries and also in Japan.

9.1.1. The ISO Recycling Sign

To support the efforts to collect batteries, a special sign has been introduced worldwide, characterized by three chasing arrows, as shown in Figure 126. All lead – acid, nickel – cadmium, and nickel – metal hydride batteries are marked directly or indirectly (on the package) with this recycling sign. The sign means that the batteries can be reprocessed after the end of their service life. Batteries with this sign should never be disposed of with the normal household rubbish, but returned to the dealer or to special collecting points.

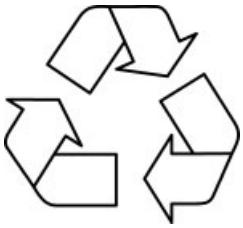

Figure 126. The ISO recycling sign (ISO Symbol 7000-1135) [240]

9.1.2. Sorting of the Collected Batteries

A general problem is caused by the fact that the collected small batteries form a mixture of all the systems and types that are on the market. Table 23 shows the sales of such batteries in the US between 1985 and 1992. A corresponding composition of the collected batteries has to be expected a few years later when these batteries are disposed.

A similar mixture of the collected batteries is to be expected today, however, the percentage of silver oxide and mercuric oxide batteries will be smaller, since those batteries were mainly used for special hearing aids and are now replaced by air – zinc batteries. On the other hand, nickel – metal hydride and lithium-ion batteries

Table 23. Percentage sales of domestic household batteries in the U.S, 1985 – 1992 [241]

Alkaline	Carbon – zinc	Nickel – cadmium	Air – zinc	Silver oxide	Mercuric oxide	Lithium
63.5 %	19.7 %	9.4 %	3.4 %	2.6 %	1.2 %	0.2 %

appeared on the market in the meantime and increasingly will contribute to the mix.

All these batteries must be sorted, since special recycling processes have to be used for the various battery systems. Separate collection would not make much sense, since the user often does not know which battery system he/she uses. Furthermore, it will happen rather often that batteries are thrown into the wrong bin.

The consequence is that sorting stations have to be installed to ensure effective separation of the various systems. Such stations are costly and always involve combinations of hand sorting and a number of automated steps [242]. In the latter, vibrating tables may separate different sizes of batteries. Magnetic separation and X-ray analysis can be used to discern the different battery systems. The battery can be transported from the conveyor belt into a collection device by mechanical forces or an air stream.

As an example, Table 24 shows results that were obtained in 1999 in Germany. The table lists the number of collected batteries compared to the delivery of new batteries in that year, both published by GRS [243], [244].

The figures of collected and delivered batteries in Table 24 cannot be compared directly, since most of the batteries collected in 1999 had been delivered in previous years. But the table shows that the ratio of the various systems within the collected mix reflects fairly well the delivered mix. This indicates that the collecting systems includes all battery systems, and that the market situation did not change too much during the previous years.

Manganese – zinc and carbon – zinc batteries constitute by far the largest number of collected batteries, according to their large share in the consumer market. The collected amounts of nickel – metal hydride and lithium-ion batteries were comparatively small in 1999, since both systems are rather young. The numbers of mercury oxide batteries, on the other hand, will diminish within the next years, since they are no longer manufactured in Europe.

Comparison of the total amount of collected batteries to that of delivered ones indicates that about 30 % of the delivered batteries were collected in 1999, although both figured cannot be compared, as already has been mentioned. According to [244] about 20 % of these batteries could be recycled while the residual 80 % had to be deposited, mainly on account of their mercury content (see Section 9.2).

9.2. Recycling of Batteries

The re-use of the raw materials must be the goal for the recycling processes to keep the expenses as low as possible. This means:

- The recycling process must result in materials that can compete with the primary materials or at least can be used instead of them. In lead – acid batteries, for example, primary lead is still preferred by many manufacturers as active material. Increasingly, however, recycled lead is refined so far that it can also be used for this purpose.
- The recycling method must be economical, i.e., the price of the secondary raw material must compete, at least approximately, with that of the primary raw material. This is mainly a question of the energy required, but labor cost may also be important, for example when the battery scrap has to be sorted laboriously.
- The method must be ecologically suitable, i.e., the recycling process itself must not cause environmental problems or generate residues that are difficult to dispose of.

The first item mainly concerns the purity of the recovered material. The purity achieved is a question of expenditure and thus is closely connected to the second item. But also the price of the primary metal is decisive and must be compared to that of the recycled material. It furthermore depends on the specification of the required material which kind of recycling process may be suitable and how far purification has to be carried.

Table 24. In Germany 1999 delivered and sorted batteries, data derived from [243], [244]

System	New batteries delivered in 1999		Disposed batteries collected in 1999	
	t	%	t	% of collection
Primary batteries				
Alkaline manganese + carbon – zinc	21 393	85.48	6012	79.56
Lithium	356	1.42	106	1.4
Air – zinc	75	0.30	75	1
Silver oxide + mercuric oxide	61	0.25	46	0.61
Secondary batteries				
Nickel – cadmium	1842	7.35	970	12.84
Nickel – metal hydride	676	2.70	7	0.09
Lithium-ion	324	1.29	47	0.2
Small lead – acid	290	1.16	232	3.08
Total	**25 017**	**100**	**7557**	**98.7**∗

∗1.3 % of the collected batteries could not be sorted.

Overall, the economic aspect of recycling is favored by the dramatic increase in dumping costs in most countries, and the dwindling possibilities of battery disposal in landfills greatly stimulate recycling efforts.

A fairly detailed review of the various recycling methods for batteries is given by a brochure edited 1991 by the Association of European Accumulator Manufacturers (EUROBAT) and the Association of European Dry Battery Manufacturers (EUROPILE) [245].

9.2.1. Recycling of Lead – Acid Batteries

Recycling of lead – acid batteries has been well established for decades, and most lead – acid batteries are actually recycled. Two main processes are used [246]:

1) The *blast-furnace* or *shaft-furnace process* which does not require previous separation of the battery components. The more or less smashed batteries, together with coke and smelting additives, are directly filled into the furnace. The organic components such as separators and containers are used as additional fuel, but they are lost. Lead sulfate and residual sulfuric acid are reduced and, together with added iron, form a slag of iron sulfide. Chlorine, which mainly originates from separators or spacers made of PVC, forms lead chloride, which is volatile at temperatures above 900 °C and escapes from the furnace together with the furnace gas. Lead chloride has to be collected by effective filter systems after cooling.
2) The *rotary furnace process* that requires preceding separation of the metallic parts of the battery from the other inorganic components. This means intensive pretreatment, including flotation processes to separate the different base materials according to their different specific weights. Sulfate, which forms a part of the active material as lead sulfate is removed with the aid of soda which converts the lead sulfate into lead carbonate which is further reduced to the metal. Only the metal-containing components are smelted in the rotary furnace.

In both cases, the crude lead obtained has to be refined by subsequent metallurgical processes.

The advantages and disadvantages of the processes are:

– The blast-furnace process requires practically no pretreatment and generates no wastewater that has to be neutralized and thoroughly decontaminated of residual lead. So it is economical. However, all the organic components of the battery are not recycled but lost (used as fuel).
– In the rotary-furnace process, all the organic components survive and may be recycled. The main problem is that a great variety of organic substances is used in batteries which can hardly be separated, and that these materials are contaminated with lead and other metals. But the increasing necessity for recycling everything may stimulate further development of this process.

The diluted acid is a special problem for economic reasons. It contains 10 to 20 wt % H_2SO_4; the rest of the sulfuric acid remains in the partly discharged plates of the scrapped batteries. But the diluted acid is contaminated, mainly with lead but also with other metals such as antimony, traces of copper, and occasionally nickel. So it cannot be reconcentrated, but has to be distilled to be reused as battery electrolyte. This distillation has to be carried out in a vacuum; otherwise sulfuric acid is decomposed before its boiling point of 338 °C is reached. Nowadays, this process is too expensive and the scrap acid is used for other chemical or metallurgical processes, like pickling, and is finally deposited as (heavy-metal-contaminated) gypsum ($CaSO_4$).

9.2.2. Recycling of Nickel – Cadmium Batteries

Recycling of larger nickel – cadmium batteries, used in stationary or traction applications, mainly concerns recovery of the valuable cadmium that forms about 15 % of the scrap weight. Cadmium can easily be regained by distillation. Recycling of small and especially sealed nickel – cadmium batteries on a larger scale was started at the end of the 1970s. Reasons for this rather late start were the above described problem of collecting these batteries and the more expensive treatment that is required. For this reason only about 15 % of the discarded nickel – cadmium batteries have been recycled in the U.S. in 1995 [241]. In Europe corresponding figures differ for the various countries between 5 and 50 % [247].

Problems connected with the recycling of nickel – cadmium batteries are:

- The collection of sealed nickel – cadmium batteries usually leads to a mixture with primary batteries. Therefore, sorting plants had to be established.
- The batteries first have to be dismantled before they can be treated.
- Iron and nickel, both of which are contained in these batteries, cannot be separated by the usual metallurgical processes. Both form an alloy that can only be used for production of stainless steel. The condition that the recycling process must result in materials that can be used again for battery production cannot be fulfilled for these components.

In the meantime the situation with small cells and batteries has changed, since collecting systems have been installed in many countries. Recycling capabilities are available as Table 25 shows.

According to the figures in Table 25, more than 30 000 t of nickel – cadmium batteries could be recycled. This figure is far above the amount of battery nowadays collected, and it surely could be increased. Thus, recycling itself is not the bottle-neck.

Predominantly pyrometallurgical processes are applied, especially in the plants that are dedicated NiCd recyclers (rows 4, 7, and 8 in Table 25).

In the NIFE process (row 7 in Table 25), larger cells are dismantled and positive and negative electrodes separated. The negative electrodes are rinsed to remove residual electrolyte and organic substances. Then the cadmium is distilled at 850 °C. The condensed metal is collected as a granulate that can directly be used as raw material in the production of new batteries. The positive plates are sold to steelworks where their content of nickel is used as an alloying additive. Plastic containers are cleaned and the material is granulated to be re-used. Small consumer batteries are shredded, than heated up to about 500 °C where the plastic components and the carbon are pyrolyzed. Then the rest is heated to 850 °C and the cadmium distilled as described above. The residuals that contain about 50 % of nickel are also sold the steelworks (more detailed description in [245]).

The SNAM/SAVAM process (row 8 in Table 25) is similar in principle. But here after dismantling the cells the material is pressed in cylindrical packages (900 mm diameter; 900 mm height) that are treated in electrical furnaces. Cadmium again is distilled at 900 °C (cf. also [245]).

INMETCO also applies a pyrometallurgical process that is similar to the above described ones. But this process forms only a small part of a much larger recycling operation to recover iron, chromium, and nickel for the stainless steel industry.

TNO/Esdex uses a hydrometallurgical process where cadmium is leached out of the

Table 25. Facilities that can recycle nickel – cadmium batteries and their capacity in t/a (1996)

Company	Location	Type	Capacity, t
1Esdex Corporation	Canada	hydrometallurgical	2000∗
2Hydrometal S.A.	Belgium	hydrometallurgical	1300
3INMETCO	USA	stainless steel	10 000
4Japan Recycle Centre	Japan	NiCd recycler	3000
5Kansay Catalyst	Japan	zinc refinery	500
6Mitsui Mining	Japan	zinc refinery	750
7SAFT/NIFE	Sweden	NiCd recycler	1500
8SNAM/SAVAM	France	NiCd recycler	5400
9TNO/Esdex	Netherlands	hydrometallurgical	200
10Toho Zinc Co., Ltd.	Japan	zinc refinery	1700
11Uniniquel S.A.	Spain	hydrometallurgical	2400∗
12U.S. Filter Recovery	USA	hydrometallurgical	3000∗

∗ in 1996 planned figures (Table 8 in [247])

shredded batteries with the aid of hydrochloric acid. Cadmium is separated by solvent extraction (with tributylphosphate) and subsequently precipitated electrochemically as a rather pure metal. Nickel is recovered also by electrochemical deposition.

9.2.3. Recycling of Nickel – Metal Hydride Batteries

The recycling of nickel – metal hydride batteries is less attractive as a commercial process, since the raw materials of the negative electrode can hardly be regained and the hydrogen content of this alloys makes the handling of scrapped batteries more risky. But the fast increasing use of such batteries in many applications requires recycling capabilities.

In Germany NIREC developed a process (according to [242]) which is based on the recovery of the nickel content. In this process the batteries are shredded under vacuum, while the released hydrogen is continuously removed.

9.2.4. Recycling of Lithium-Ion Batteries

Nearly all lithium-ion batteries that are on the market contain cobalt in the form of cobalt oxide (CoO_2). Cobalt is a heavy metal and therefore should not be dumped in landfills. Furthermore, it is an expensive material that should be recovered. At the moment (2000), recycling processes are tested as pilot projects.

9.2.5. Recycling of Primary Batteries

In the recycling of primary batteries, a fundamental difference exists between batteries that contain mercury and others that are mercury-free. Formerly carbon – zinc (Leclanché) batteries and alkaline manganese dioxide batteries contained mercury on the order of 1 %, mainly employed to reduce self-discharge by hydrogen evolution at the zinc electrode. Today these batteries no longer contain any mercury, and only remainders of older or imported types may be found within the collected manifold. To allow an easy identification of the modern mercury-free batteries by the sorting machine, the European manufacturers cover them with a special varnish that can be detected by UV light.

9.2.5.1. Recycling of Mercury-Containing Batteries

Mercury gradually will diminish in waste batteries, because, as already mentioned, the addition of mercury has been abandoned in carbon – zinc and alkaline manganese batteries, which represent by far the largest share of primary batteries.

Batteries with mercury oxide as the active material of the positive electrode, which mainly had been used as button cells in hearing aids, will also disappear soon, because they are no longer produced.

Anyway it will take a number of years until mercury-containing batteries will have totally disappeared from in collected batteries. Until then they have to be sorted out and the mercury has to be recovered by vacuum distillation

(ALD Process). The removal of small mercury contents, however, is quite expensive.

9.2.5.2. Recycling of Mercury-Free Batteries

Primary lithium batteries contain organic electrolytes but no heavy metals. They do not contain substances that could be recovered and re-used as raw material for batteries.

10. References

1. IEC Public. 60086 part 4, 2nd ed., Safety standard for lithium batteries, 2001, Central Office of the IEC, 3 rue de Varembé, Genève, Switzerland.
2. IEC Public. 60086 part 5, 1st ed., 2001 *Safety standard for batteries with aqueous electrolyte* Central Office of the IEC, 3 rue de Varembé, Genève, Switzerland.
3. FR 71865, 1866 (G. Leclanché).
4. US 2 650 945, 1953 (W. S. Herbert).
5. W. S. Herbert, *J. Electrochem. Soc.* **99** (1952) 190 C.
6. GB 1932, 1883 (C. L. Clarke).
7. GB 15407, 1891 (H. Cohen).
8. DE 440 806, 1927 (C. Féry).
9. A. J. Salkind (ed): Proc. Symp. History of Battery Technology, The Electrochem. Soc. Proc., vol. 87 – 14, Pennington, NJ 1987.
10. R. H. Schallenberg: *Bottled Energy: Electrical Engineering and the Evolution of Chemical Energy Storage,* **vol. 148,** by the American Philosophical Society (Memoirs Series), 1982.
11. K. Jäger (ed.): *Geschichte der Elektrotechnik 13, Gespeicherte Energie,* VDE-Verlag, Berlin-Offenbach, 1994.
12. D. Linden: *Handbook of Batteries,* 2nd ed., McGraw-Hill, New York 1995.
13. H. Bode: *Lead – Acid Batteries,* John Wiley & Sons, New York 1977.
14. D. Berndt: *Maintenance-Free Batteries,* 2nd ed., John Wiley & Sons, New York 1997.
15. S. U. Falk, H. J. Salkind: *Alkaline Storage Batteries,* John Wiley & Sons, New York 1969.
16. M. Wakihara, O. Yamymoto: *Lithium Ion Batteries,* Wiley-VCH, Weinheim 1998.
17. K. Kordesch, G. Simader: *Fuel Cells,* Wiley-VCH, Weinheim 1996.
18. A. J. Bard, L. R. Faulkner: *Electrochemical Methods – Fundamentals and Applications,* John Wiley & Sons, New York 1980, p. 49.
19. J. S. Newman: *Electrochemical Systems,* 2nd ed., Prentice-Hall, Inc., Englewood Cliffs 1993, p. 120 f.
20. D. J. G. Ives, J. J. Jantz: *Reference Electrodes,* Academic Press, New York – London 1961, p. 71.
21. D. Berndt: *Maintenance-Free Batteries,* 2nd ed., Research Studies Press, John Wiley & Sons, New York 1997, p. 20.
22. S. U. Falk, A. J. Salkind: *Alkaline Storage Batteries,* John Wiley & Sons, New York 1969, p. 598.
23. J. Tafel, *Z Phys. Chemie* **50** (1905) 641.
24. P. Reasbeck, J. G. Smith: *Batteries for Automotive Use,* RSP Ltd, Taunton, and John Wiley & Sons Inc, New York 1997, p. 29.
25. K. J. Vetter, *J. Electrochem. Soc.* **110** (1963) 597.
26. IEC 60086-1, 9th ed., 2001, EC, 3 rue de Varembé, Genève, Switzerland.
27. IEC Public. 60050(481), 1st ed., 1996-07, IEC, 3 rue de Varembé, Genève, Switzerland.
28. K. Kordesch et al., 36th Power Sources Conference, Palisades Institute for Research Services, Inc., New York.
29. K. Kordesch et al., Proc. 182 Meet. Of the Electrochem. Soc., Toronto, Ont. Canada, Oct. 11 – 16, 1992, Abstract 10.
30. D. Berndt, E. Voss: "Batteries 2," *Proc. 4th Intern. Symp. Brighton 1964,* Pergamon Press, Oxford 1965.
31. IEC Public. 60285, 60622, 60509, 61808, 61436, 61440 (1999) Central Office of the IEC, 3 rue de Varembé, Genève, Switzerland.
32. D. H. Spahrbier, IEC/TC35, document 35/1109/CDV, 1999.
33. IEC 60086-2, 10th ed., 2001, IEC, 3 rue de Varembé, Genève, Switzerland.
34. Proposal (1999) by D. H. Spahrbier, formerly VARTA Batterie AG, R & D Centre, Kelkheim.
35. D. H. Spahrbier et al.: *Proc. 13th Int. Pow. Sourc. Sympos., 333,* Academic Press, New York 1983.
36. IEC 1056-1, 1991, Central Office of the IEC, 3 rue de Varembé, Genève, Switzerland.
37. D. Linden (ed.): *Handbook of Batteries,* 2nd ed., part 5, McGraw-Hill, New York 1994.
38. M. Barak (ed.): *Electrochemical Power Sources,* The Institution of Electrical Engineers, Stevenage 1980.
39. D. W. McComsey, W. B. Felegyhazi in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 8, McGraw-Hill, New York 1994.

40. R. Huber in K. V. Kordesch (ed.): *Batteries,* **vol. 1,** chap. 1, Marcel Dekker, New York 1974.
41. H. Bode, A. Schmier, D. Berndt, *Z. Elektrochem. Ber. Bunsenges.* **66** (1962) 586.
42. K. J. Vetter, *Z. Elektrochem. Ber. Bunsenges.* **66** (1962) 577.
43. G. S. Bell, R. Huber, *J. Electrochem. Soc.* **111** (1964) 1.
44. H. Bode, A. Schmier, *Ber. Bunsenges.* **68** (1964) 954.
45. F. L. Tye: "Primary batteries for civilian use," in M. Barak (ed.): *Electrochemical Power Sources,* The Institution of Electrical Engineers, Stevenage 1980.
46. Polaroid Corporation, Commercial Battery Division, 868 Winter Street, Waltham, MA 02154, USA.
47. D. W. McComsey, W. B. Felegyhazi in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 8, McGraw-Hill, New York 1994.
48. H. Cohen, GB 15407, 1891.
49. B. Siller, *Elektrotechn. Ausg. B* **17** (1965) 328.
50. C. Féry, DE 440 806, 1927.
51. R. F. Koontz in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 17, McGraw-Hill, New York 1994.
52. K. Kordesch et al., 18th Int. Power Sources Symp., Stratford-upon-Avon, April 1993.
53. K. Kordesch (ed.): *Batteries,* **vol.1, Manganese Dioxide,** Dekker, New York 1974.
54. G. S. Bell, R. Huber, *J. Electrochem. Soc.* **111** (1964) 1 – 6.
55. F. L. Tye, *Electrochim. Acta* **21** (1976) 415 – 420.
56. K. Kordesch et al. in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 34, McGraw-Hill, New York 1994.
57. A. Kozawa, R. J. Brodd, Proceedings of the Manganese Dioxide Symposium, vol. 1, Cleveland, 1975, I.C. Sample Office, Cleveland.
58. B. Schumm, H. M. Joseph, A. Kozawa, Proceedings of the Manganese Dioxide Symposium, vol. 2, Tokyo, Japan 1980, I.C. Sample Office, Cleveland.
59. H. Laig-Hörstebrock, *J. Electroanal. Chem.* **180** (1984) 599 – 607.
60. J. J. Coleman, *Trans. Electrochem. Soc.* **90** (1946) 545.
61. A. B. Scott, *J. Electrochem. Soc.* **107** (1960) 941.
62. J. P. Gabano, J. Seguret, J. F. Laurent, *J. Electrochem. Soc.* **117** (1970) 147.
63. E. Warburg, *Wied. Ann.* **67** (1899) 493.
64. K. V. Kordesch et. al., 18th Int. Power Sources Symp., Stratford-upon-Avon, England Apr. 19 – 21, 1993.
65. K. Kordesch, J. D. Ivad in J. O. Besenhard (ed.): *Handbook of Battery Materials,* Wiley-VCH, Weinheim 1999.
66. W. J. Wruck et al., *J. Electrochem. Soc.* **138** (1991) no. 12, 3560 – 3567.
67. D. Spahrbier, H. M. Lippold, *Progr. in Batteries & Sol. Cells* **1** (1978) 48 – 50.
68. U.S. 4 078 127, 1978 (E. A. Megahed, C. R. Buelow).
69. U.S. 4 009 056, 1977 (E. A. Megahed, C. R. Buelow, P. Spellmann).
70. E. A. Megahed, D. V. Davig, *Progr. in Batteries & Sol. Cells* **4** (1982) 83 – 86.
71. E. A. Megahed, D. V. Davig, *Power Sources,* **vol. 8,** Academic Press, London 1981.
72. C. P. Wales, J. Burbank, *J. Electrochem. Soc.* **106** (1959) 885.
73. T. P. Dirske, *J. Electrochem. Soc.* **109** (1962) no. 3, 173 – 177.
74. S. M. Davis, U.S. Patent 3, 853, 623, 1974.
75. Y. V. Pleskov, B. N. Kabanow, *Zh. Neorg. Khim* **2** (1957) 1807 – 1811.
76. R. F. Amlie, P. Ruetschi, *J. Electrochem. Soc.* **108** (1961) 813 – 819.
77. H. Andre, *Bull. Soc. Fr. Electr.* **1** (1941) 132 – 146.
78. V. D'Agostino in A. Fleischer, J. Lander (eds.): *Zinc Silver Oxide Batteries,* John Wiley & Sons, New York 1971, p. 271.
79. L. B. Adams et al. in J. Thompson (ed.): *Power Sources 7,* Academic Press, London 1979.
80. H. M. Lippold, H.-W. Nientiedt, D. H. Spahrbier in J. Thompson (ed.): *Power Sources,* Academic Press, London 1979.
81. H.-W. Nientiedt, *J. Power Sources* **8** (1982) 257 – 265.
82. H.-W. Nientiedt, H. Laig-Hoerstebrock, *J. Power Sources* **8** (1982) 267 – 272.
83. S. Ruben, *Proc. Electrochem. Soc., General Meeting, Boston,* October 1947.
84. D. Naylor in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 11. McGraw-Hill, New York 1994.
85. K. Kordesch, G. Simader, *Fuel Cells,* Wiley-VCH, Weinheim 1996.
86. R. G. Biddick, U.S. 4 129 633.
87. T. P. Dirske, *J. Electrochem. Soc.* **101** (1954) 328 – 331.
88. G. N. Lewis, M. Randall, *J. Am. Soc.* **36** (1914) 1969.
89. K. J. Vetter: *Elektrochemische Kinetik,* chap. 154, Springer Verlag, Heidelberg 1961.

90. J. Whartman, B. Koretz: "The Electric Fuel Battery System", 2nd Annual Intern. Zinc Conference, in Scottsdale Arizona, February 21 – 24, 1993.
91. Electric Fuel Ltd., Har Hotzvim Science Park P.O. Box 23073, Jerusalem 91230, Israel.
92. W. Buckentin, Presentation at the DGS-Meeting in Karlsruhe, Germany, April 29/30, 1999.
93. W. M. Latimer (ed.): "Oxidation Potentials", 2nd ed., Prentice-Hall, Inc. Englewood Cliffs 1952.
94. S. B. Brummer et al. in D. W. Murphy, J. Broadhead, B. C. H. Steele (eds.): *Materials for Advanced Batteries,* Plenum Press, New York 1980, p. 123.
95. A. N. Dey, *Thin Solid Films* **43** (1977) 131.
96. A. N. Dey, C. R. Schlaikjer, Proceedings of the 26th Power Sources Symposium (1974).
97. E. Peled, D. Golodnitzky, J. Penciner: "The Anode/Electrolyte Interphase" in J. O. Besenhardt (ed.): *Handbook of Battery Materials,* Wiley-VCH, Weinheim 1999.
98. S. Ching et al., *J. Electrochem. Soc.* **138** (1991) 2308.
99. G. Geblewicz et al., *Trans. Inst. Met. Finish* **64** (1986) 134.
100. J. O. Besenhardt (ed.): *Handbook of Battery Materials,* chaps. 7, 8, 9, Wiley-VCH, Weinheim 1999.
101. B. B. Owens, P. M. Skarstadt: "Ambient Temperature Solid State Batteries," *Solid State Ionics* **43** (1992) 665.
102. H. Laig-Hoerstebrock, *Electrochim. Acta* **35** (1990) no. 11/12, 1889.
103. D. Linden (ed.): *Handbook of Batteries,* 2nd ed., part 5, McGraw-Hill, New York 1994, pp. 14 – 48.
104. D. Linden in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., part 5, chap. 14, McGraw-Hill, New York 1994.
105. *Catalogue of lithium-carbon monofluoride batteries,* Matsushita Battery Industrial Co., Ltd., Osaka 1999.
106. K. Nishio, N. Furukawa in J. O. Besenhardt (ed.): *Handbook of Battery Materials,* chap. 2, Wiley-VCH, Weinheim 1999.
107. Matsushita Battery Industrial Co., Ltd., Japan.
108. G. Lander, D. Spahrbier, *Deutsche Gesellschaft für Chronometrie DGC* **30** (1979) 191 – 203.
109. H. D. Jaksch: *Batterie Lexikon,* Pflaum Verlag München 1993.
110. J. R. Rea et al., Proc. Symp. on Biomedical Implantable Applications and Ambient Temperature Lithium Batteries, vol 80-4, Electrochem. Soc., Princeton, N.J., 1980, pp. 245 – 253.
111. E. H. Reiss, Proc. 29th Power Sources Conf., Electrochem. Soc., Pennington, N.J., June 1980.
112. M. Fukuda, T. Lijima in J. P. Gabano (ed.): *Lithium Batteries,* Academic Press, New York 1983, p. 287.
113. J. F. MacCartney et al., "Development of a 200 kWh Li/SO_2 Battery for Undersea Applications," Proc. 29th Power Sources Conf., Electrochemical Society, Pennington, N.J., June 1980.
114. Fiona Gray, Michel Armand in J. O. Besenhardt (ed.): *Handbook of Battery Materials,* Wiley-VCH, Weinheim 1999.
115. B. B. Owens et al. in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 15, McGraw-Hill, New York 1994.
116. P. Birke, W. Weppner: "Solid Electrolytes" in J. O. Besenhardt (ed.): *Handbook of Battery Materials,* Chap. 9, Wiley-VCH, Weinheim 1999.
117. L. W. Strock, *Z. Phys. Chem.* **B25** (1934) 441.
118. J. G. Goodenough et al., *Mater. Res. Bull.* 1976, 11, 203.
119. B. B. Owens et al. in D. Linden (ed.): *Handbook of Batteries,* 2nd ed., chap. 15.2.3, McGraw-Hill, New York 1994.
120. K. Kordesch, J. D. Ivad: "Global competition of Primary and Secondary Batteries", in J. O. Besenhardt (ed.): *Handbook of Battery Materials,* chap. 3, Wiley-VCH, Weinheim 1999.
121. K. Kordesch, G. Simader, *Fuel Cells and their application,* Wiley-VCH Weinheim 1996.
122. H. R. Smith, B. C. Tierney, "Ammonia Batteries", Proc. 19th Power Sources Conf., Electrochemical Soc., Pennington, N.J., 1965.
123. V. Klasons: "Thermal Batteries" in D. Linden (ed.): *Handbook of Batteries,* second edition, chapter 22, McGraw-Hill, Inc., New York 1994.
124. R. A. Guidotti, F. W. Reinhardt, "The Relative Performance of FeS and CoS in Long-Life Thermal-Battery Applications," Proc. 9th Int. Symp. Molten Salts, 1994.
125. D. Linden (ed.): *Handbook of Batteries,* second edition, chapt. 14 – 22, McGraw-Hill, Inc., New York 1994.
126. H. Bode: *Lead – Acid Batteries,* John Wiley & Sons, New York 1977, Table. 2.36, p. 98.

127. D. Berndt: *Maintenance-Free Batteries,* 2nd ed., Research Studies Press, John Wiley & Sons, New York 1997, p. 136.
128. D. Berndt, R. Bräutigam, U. Teutsch, Proc. 17th Intelec Meeting, The Hague 1995, IEEE Communications Soc., 1.
129. D. O. Feder, W. E. M. Jones, Proc. 18th Intelec Meeting, Boston 1996, IEEE Communications Soc., 184.
130. D. Berndt, E. Voss: "Batteries 2", *Proc. 4th Symp. Brighton 1964,* Pergamon Press, Oxford 1965.
131. N. E. Bagshaw: *Batteries on Ships,* Research Studies Press (John Wiley & Sons), New York 1982, p. 106.
132. G. D. Rodriguez, W. C. Spindler, D. S. Carr, *J. Power Sources* **31** (1990) 311.
133. K. R. Bullock, B. K. Mahato, G. H. Brilmyer, G. L. Wierschem in: "Advances in Lead – Acid Batteries," The Electrochem. Soc. Proceedings, vol. 84-14, p. 451.
134. W. Huber, Batteries International, Issue 4, July 1990, p. 48.
135. P. Reasbeck, J. G. Smith: *Batteries for Automotive Use,* RSP Ltd and John Wiley & Sons Inc., New York 1997, p. 7.
136. D. Berndt: *Maintenance-Free Batteries,* RSP Ltd and John Wiley & Sons Inc., New York 1997, p. 53.
137. S. S. Misra, T. M. Noveske, A. J. Williamson, Proc. 16th Intelec Meeting, Vancouver 1994, IEEE Communications Soc., p. 250.
138. K. E. White, Proc. 18th Intelec Meeting, Boston 1996, IEEE Communications Soc., p. 624.
139. W. E. M. Jones, Batteries International, Issue 33, October 1997, p. 77.
140. D. Berndt, W. E. M. Jones, Proc. 20th Intelec Meeting, San Francisco 1998, IEEE Communications Soc., p. 443.
141. W. E. M. Jones, Proc. 20th Intelec Meeting, San Francisco 1998, IEEE Communications Soc., p. 461.
142. J. R. Pierson, C. E. Weinlein, in B. Culpin, K. Peters, N. R. Young: *Power Sources 9* Academic Press, New York1983, p. 49.
143. J. Strebe, B. Reichman, B. Mahato, K. R. Bullock, *J. Power Sources* **31** (1990) 43.
144. "Prospects of Gates Energy Products", Gates Battery Application Manual, Gates Energy Products Inc., 1 050 S Broadway, Denver, Colorado 80217, U.S.A.
145. M. R. Harrison, Proc. Intelec Meeting, Washington, D.C. 1982, IEEE Publishing Inc., p. 424.
146. Batteries International, Issue 19, April 1994, p. 36.
147. D. E. Koontz, D. O. Feder, L. D. Barucci, H. J. Luer, *Bell Syst. Tech. J.* **49** (1970) 1253. R. V. Biagetti, H. J. Luer, *J. Power Sources* **4** (1979) 309.
148. N. E. Bagshaw: A Battery Technologists Guide to Metallurgical Structures and Alloy Properties, Advanced Lead – Acid Battery Consortium (ALABC), 1994, ILZRO, Inc, 2525 Meridian Parkway, Suite 100, P.O Box 12036, Research Triangle Park, NC 27709-2036, USA.
149. T. Ferreira, Batteries International, Issue 36, July 1998, p. 83.
150. "Ventseal" published by Busak + Shamban GmbH & Co, Handwerkstraße 5 – 7, D-70565 Stuttgart, Germany.
151. K. Ledjeff, A. Winsel, *J. Power Sources* **8** (1982) 211.
152. H. Bode, K. Dehmelt, J. Witte, *Z. Anorg. Allg. Chemie* **366** (1969) 8.
153. S. U. Falk, A. J. Salkind: *Alkaline Storage Batteries,* John Wiley & Sons, New York 1969, Fig. 2.2, p. 50.
154. R. Barnard, C. F. Randell, F. L. Tye, *J. Appl. Electrochem.* **10** (1980) 109.
155. R. Barnard, C. F. Randell, *J. Appl. Electrochem.* **13** (1983) 97.
156. VARTA Batterie AG (ed.): *Sealed Nickel Cadmium Batteries,* VDI Verlag GmbH, Düsseldorf 1982, Fig. 5/1, p. 59.
157. D. Linden (ed.): *Handbook of Batteries,* 2nd ed., McGraw-Hill Inc., New York 1995.
158. J. N. Taylor (ALCAD Ltd), Proc. 17th Intelec Meeting, The Hague 1995, IEEE Communications Soc., p. 778.
159. VARTA Batterie AG (ed.): Sealed Ni-Cd Cells – Sales Program and Technical Handbook, Ellwangen/Jagst, Figs. 2.14 and 3.10.
160. SANYO Electric Co. Ltd.: Cadnica, Sealed Type Nickel-Cadmium Batteries, Engineering Handbook, Osaka 1995.
161. Vantage Technical Manual, Alcad Limited, Hampton, Middlesex.
162. G. Halpert in D. A. Corrigan (ed.): *Nickel Hydroxide Electrodes,* **Proc. vol. 90-4,** The Electrochem. Soc. Inc., Pennington, N.J. 1990, p. 3.
163. DE 1 133 442, 1958 (L. Kandler).
164. H. N. Seiger in H. C. Maru (ed.): *Porous Electrodes,* **Proc. vol. 84-8,** The Electrochem. Soc. Inc., Pennington, N.J. 1984, p. 50.

165. M. Paszkiewicz, G. Mlynarek in D. A. Corrigan (ed.).: *Nickel Hydroxide Electrodes,* **Proc. vol. 90-4,** The Electrochem. Soc. Inc., Pennington, N.J. 1990, p. 222.
166. I. Matsumoto, H. Ogawa, T. Iwaki, M. Ikeyama in T. Keily, B. W. Baxter (eds.): Power Sources 12, Intern. Power Sources Symp. Committee, Brighton 1988, p. 203.
167. W. A. Ferrando, R. A. Sutula in R. G. Gunther, S. Gross (eds.): *The Nickel Electrode,* **Proc. vol. 82-2,** The Electrochem. Soc. Inc., Pennington, NJ 1982, p. 27.1.
168. B. Bugnet, D. Doniat in L. J. Pearce (ed.): Power Sources 11, Intern. Power Sources Symp. Committee, Brighton 1986, p. 179.
169. "HOPPECKE SPECIAL III/June 1986", Akkumulatorenwerke HOPPECKE Carl Zoellner & Sohn GmbH, 1986.
170. G. Benczúr-Ürmössy, *Dechema Monographs* **109** (1987) 379.
171. J. Lindra et al., *J. Power Sources* **4** (1979) 227.
172. J. Mrha et al., *J. Power Sources* **4** (1979) 239.
173. Société Piles Wonder, DE 2822-821-C, 1977 (B. Sticker, J. Jolas).
174. Panasonic Nickel Cadmium Batteries Technical Handbook 89/, Matsushita Battery Industrial Co, Ltd., Chigasaki, Kanagawa.
175. "Sanyo Rechargeable CADNICA", Sanyo Electric Co., Ltd., Sumoto-city, Hyogo.
176. VARTA Sealed Ni-Cd Cells; Sales Program and Technical Handbook, VARTA Batterie AG, Ellwangen/Jagst.
177. VARTA, Ni-MH Cylindrical and Prismatic Cells; Sales Program and Technical Handbook, VARTA Batterie AG, Ellwangen/Jagst.
178. J. M. Dunlop: "Nickel Hydrogen Batteries," in D. Linden (ed.): *Handbook of Batteries and Fuel Cells,* McGraw-Hill, New York 1995, Chap. 32.
179. D. M. Bush, J. F. Sindorf in: "Stationary Energy Storage: Load Leveling and Remote Applications," **Proc. vol. 88-11,** The Electrochem. Soc. Pennington, N.J. 1988, p. 318.
180. D. Berndt: *Maintenance-Free Batteries,* RSP Ltd and John Wiley & Sons New York 1997, p. 227.
181. W. M. Mueller, J. P. Blackledge, G. G. Libowitz: *Metal Hybrides,* Academic Press, New York 1988.
182. L. Schlapbach (ed.): "Hydrogen in Intermetallic Compounds I and II," in*Topics in Applied Physics,* **vols. 63 and 67,** Springer Verlag, Berlin – Heidelberg 1988 and 1992.
183. G. G. Libowitz in: *Proc. Symp. Hydrogen Storage Materials, Batteries and Electrochemistry,* **Proc. vol. 92-5,** The Electrochemical Soc., Inc. 1992, p. 3.
184. G. Sandrock in D. Bennett, T. Sakai (eds.): *Symp. Hydrogen and Metal Hydride Batteries,* **Proc. vol. 94-27,** The Electrochemical Society, Inc., Pennington, N.J. 1994, p. 1.
185. J. S. Eager, V. J. Puglisi: "Nickel-Metal Hydride Performance and Environmental Advantages," in The Fourth International Rechargeable Battery Seminar, Deerfield Beach, FL, March 1992, Shawmco, Inc. P.O. Box 9495, Tulsa, OK 74137 USA.
186. K. Kajita, Y. Uetani: "Ni/MH Batteries using Muliphase-Multicomponent Alloys," in The Tenth International Seminar on Primary and Secondary Battery Technology an Application, Florida Educational Seminars, Inc., Deerfield Beach, FL 1993.
187. M. A. Fetcenco, S. Venkatesan, S. R. Ovshinsky in: *Proc. Symp. Hydrogen Storage Materials, Batteries and Electrochemistry 1992,* **Proc. vol. 92-5,** The Electrochemical Soc., Inc., 141.
188. E. W. Justi et al., *Energy Conversion* **10** (1970) 183.
189. M. A. Gutjahr, H. Buchner, K. D. Beccu, H. Säufferer, in D. H. Collins (ed.): *Power Sources,* **vol. IV,** Oriel Press, Newcastle upon Tyne 1973, p. 79.
190. S. Mukerjee et al. in D. Bennet, T. Sakai (eds.): *Proc. Symp. Hydrogen and Metal Hydride Batteries,* **Proc. vol. 94-27,** The Electrochemical Society, Inc., Pennington, N.J. 1994, p. 33.
191. H. Buchner, M. A. Gutjahr, K. D. Beccu, H. Säufferer, *Z. Metallkde.* **63** (1972) 497.
192. A. Percheron-Guégan, J.-M. Welter: "Preparation of Intermetallics and Hydrides," in L. Schlapbach (ed.): *Hydrogen in Intermetallic Compounds I, in Topics in Applied Physics,* **vol. 63,** Springer Verlag, Berlin – Heidelberg 1988, p. 11.
193. M. Tadokoro et al.: *Proc. Symp. Hydrogen Storage Materials, Batteries and Electrochemistry 1992,* **Proc. vol. 92-5,** The Electrochemical Soc., Pennington, N.J. 1992, p. 92.
194. L. Schlapbach, F. Meli, A. Zuettel, in D. Bennet, T. Sakai (eds.): *Proc. Symp. Hydrogen and Metal Hydride Batteries,* **Proc. vol. 94-27,** The Electrochemical Society, Inc., Pennington, N.J. 1994, p. 102.

195. *Batteries International,* Issue 42, January 2000, p. 71.
196. VARTA, "Ni/MH Cylindrical and Prismatic Cells, Sales Program and Technical Handbook," 43221e, VARTA Batterie AG, Ellwangen/Jagst.
197. VARTA, "Sealed Ni/MH Cells, Sales Program and Technical Handbook," VARTA Batterie AG, Toshiba Battery Co., Ltd and DURACEll International Inc. 1992, mailed by VARTA Batterie AG, Ellwangen/Jagst.
198. VARTA (ed.): Electric vehicles – current situation and perspectives, VARTA Batterie AG, Hannover.
199. D. Linden (ed.): *Handbook of Batteries,* 2nd ed., McGraw-Hill Inc., New York 1995.
200. D. R. J. Rand, R. Woods, R. M. Dell (eds.): *Batteries for Electric Vehicles,* Research Studies Press, Taunton, England and John Wiley & Sons, New York1997.
201. Ch. Fabjan, K. Kordesch, *Dechema Monographs;* **vol. 109,** DECHEMA, Frankfurt 1987, p. 355.
202. J. O. Besenhard (ed.): *Handbook of Battery Materials,* WILEY-VCH Verlag, Weinheim 1999, p. 181.
203. M. Skyllas-Kazacos, C. Menictas, Proc. 19th Intelec Melbourne 1997, IEEE Communications Soc., p. 463.
204. Electric Fuel Limited, Paper in the course of 'The Second Annual International Zinc Conference', Scottsdale, Arizona, February 1993.
205. Batteries International, Issue 21, October 1994, p. 10 – 13.
206. W. Bucketin "Zink-Luft-Batterie-Einsatz in Bremen", DGES Fachtagung, April 1999 at Karlsruhe, DGES Geschäftsstelle: c/o BEWAG, Motzstr.89 D-10779 Berlin.
207. S. Ray, D. Kurzeja, H.-J. Pauling, "ZOXY, ein neues Zink/Luft Batterie System", GDCh Monographien, Vol. 12, GDCh, Frankfurt 1998, p. 407.
208. C. Létourneau, J. Rouillard, Y. Choquette, D. C. Worboys, Proc. 20th Intelec San Francisco 1998, IEEE Communications Soc., p. 656.
209. M. Wakihara, O. Yamamoto (eds.): *Lithium Ion Batteries,* Kodansha Ltd., Tokyo; Wiley-VCH, Weinheim 1998.
210. R. Gereth in 'New Concepts', VARTA Batterie AG, Am Leineufer 51, D-30419 Hannover, 1990, p. 21 cf. R. Gereth VARTA special report 1/1991, VARTA-special-report, P.O. Box 540, D-30405 Hannover.
211. M. Winter, J. O. Besenhard, *Electrochim. Acta* **45** (1999) 31.
212. FUJI Photo Film Co., EP 651,450 A1, 1994 (Y. Idota et al.).
213. J. O. Besenhard (ed.): *Handbook of Battery Materials,* WILEY-VCH, Weinheim, 1999.
214. Jim Y. Lee, Ruifen Zhang, Zhaolin Liu, *Electrochemical and Solid-State Lett.* **3** (2000) 167.
215. D. Huang, *Advanced Battery Technology,* **34** Nov. 1998, no. 11, 21.
216. D. Linden (ed.): *Handbook of Batteries,* 2nd ed., McGraw-Hill Inc., New York 1995.
217. M. Nagamin, H. Katoh, Y. Nishi, Electrochem. Soc. of Japan, 33rd Battery Symposium Abstract, 1992, p. 83.
218. U. Koehler et al., Proc. 32nd IECEC, Honolulu, 1997, p. 93.
219. Y. F. Y. Yao, T. J. Kummer, *J. Inorg. Nucl. Chem.* **29** (1967) 2453.
220. D. R. J. Rand, R. Woods, R. M. Dell: *Batteries for Electric Vehicles,* Research Studies Press, Taunton, England and John Wiley & Sons, New York1997, Chap. 8.
221. C. H. Dustman: "Entwicklungsstand und Einsatzfelder der ABB Hochenergiebatterie," inH. Kahlen (ed.): *Batterien,* Haus der Technik Fachbuchreihe, Vulkan-Verlag, Essen 1992, p. 161.
222. *Batteries International,* October 1998, Issue 37, 16.
223. W. Fischer, BBC, BMFT report FB-T 86-059 (1986).
224. *Batteries International,* October 1998, Issue 37, 93.
225. H. Böhm, J. Croetzer, J. L. Sudworth, in F. J. Kruger (ed.): *Batterien, GDCh-Monographie,* **vol. 12,** Gesellschaft Deutscher Chemiker, Frankfurt 1997, p. 345.
226. H. Böhm, in H. Kahlen (ed.): *Batterien,* Vulkan Verlag, Essen 1992, p. 153.
227. CENELEC (European Committee for Electrotechnical Standardization), Central Secretariat, 35, rue de Stassart, B-1050 Brussels, Belgium.
228. International Electrotechnical Commission, 3, rue de Varembé Genève, Switzerland.
229. IEC Directives, Part 1, Procedures for the technical work, 3rd edition, 1995, 3, rue de Varembé Genève, Switzerland.
230. IEC Directives, Part 2, Methodology for the development of International Standards, 2nd edition, 1992, 3, rue de Varembé Genève, Switzerland.

231. IEC Directives, Part 3, Rules for the structure and drafting of International Standards, 3rd edition, 1997, 3, rue de Varembé Genève, Switzerland.
232. IEC Document 35/1108/CDV, issued 1999-05-21.
233. IEC Document 35/1124A/CDV, issued 2000-05-15.
234. "Großer Wirrwarr bei der Batteriekennzeichnung," Markt und Technik (1990) no. 35, p. 24.
235. IEC Document 35/1109/CDV, issued 1999-05-21.
236. Battery Council International, http://www.batterycouncil.org/recycling.html.
237. Implementation of the Mercury-Containing and Rechargeable Battery Management Act, published by the US Environmental Protection Agency (EPA) as EPA530-K-97-009; http//www.epa.gov.
238. Battery Reclamation Manual, 1998 PRBA; Portable Rechargeable Battery Association, 1000 Parkwood Circle, Suite 430, Atlanta, Georgia 30339, USA. http//www.prba.org.
239. J. F. Fricke, H.-A. Kiehne, Proc. 3rd Telescon Meeting, Dresden 2000 VDE Verlag, Berlin 2000, p. 115.
240. H. A. Kiehne, Proc. 11th Intelec Meeting, Florence 1989, IEEE Communications Society 1989, Paper 17.3, Fig. 8
241. B. Fishbein, Industrial Program to Collect Nickel-Cadmium (Ni-Cd) Batteries, http//www.informinc.org
242. J. L. Fricke, N. Knudsen, Presentation at the Technische Akademie Essen, 2000.
243. www.grs-batterien.de/facts/erfolg.htm
244. www.grs-batterien.de/facts/sheets.htm
245. Eurobat - Europile, 'Realized and Projected Recycling Processes for Used Batteries', Edited by the Association of European Accumulator Manufacturers, Secretarial Office, PO Box 5032, CH-3001 Bern, Switzerland, 1991.
246. H. Niklas in F. Hiller (ed.): *Batterie und Umwelt,* 2nd ed., Chap. 7, Expert-Verlag, Ehningen 1969. Plants using both processes are presented in 'The Battery Man', Independent Battery Manufacturers Association Inc. (IBMA), Largo, Florida, Issue February 1992, pp. 12 – 18.
247. H. Morrow, Batteries International, Issue 29, October 1996, p. 51.

Bearing Materials

Harald Pfestorf, Kolbenschmidt AG, Neckarsulm, Federal Republic of Germany

Franz Weiss, Kolbenschmidt AG, Neckarsulm, Federal Republic of Germany

Karl Heinz Matucha, Metallgesellschaft AG, Frankfurt/Main, Federal Republic of Germany

Peter Wincierz, Metallgesellschaft AG, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Rolling-Element Bearings – Materials 1
3. Sliding Bearings 3
3.1. Tribology of Sliding Bearings 3
3.2. Microstructure of Important Bearing Materials 5
3.3. Metallic Materials for Sliding Bearings 7
3.3.1. Metals for Solid Bearings 8
3.3.2. Materials for Metallic Multilayer Thin-Walled Bearings 11
3.3.3. Sintered Materials 15
3.4. Plastics for Sliding Bearings 17
3.4.1. Plastics for Solid Bearings 17
3.4.2. Materials for Plastic Multilayer Thin-Walled Bearings 18
3.5. Artificial Carbon and Graphite for Solid Sliding Bearings 19
3.6. Ceramic Materials for Solid Sliding Bearings 20
3.7. Soft Rubber for Sliding Bearings . . 21
3.8. Practical Selection of Materials . . . 22
4. References 22

1. Introduction

A bearing is a machine element that transmits loads or reaction forces to a shaft that rotates or oscillates relative to the bearing. *Radial* bearings are those where the forces act at right angles to the axis of rotation. *Thrust* bearings take up axial forces acting parallel to rotating shafts. Depending on how the load is carried and what kind of friction occurs, a distinction is made between *rolling-element bearings* and *sliding* or *plain bearings.*

The alloy designations and chemical compositions given in the tables are standardized by national and international organizations. Such organizations include the American Iron and Steel Institute (AISI, Washington, D.C., United States), American Society for Testing and Materials (ASTM, Philadelphia, Pa., United States), Society of Automotive Engineers (SAE, Warrendale, Pa., United States), British Standards Institution (BS, London, Great Britain), Deutsches Institut für Normung (DIN, Berlin, Federal Republic of Germany), and International Organization for Standardization (ISO, Genève, Switzerland, international).

2. Rolling-Element Bearings – Materials

In rolling-element bearings the load is carried by balls, needles, or rollers that are in intimate contact with the raceways. The starting friction of ball or roller bearings is only slightly higher than the operating friction, which itself is usually as low as that of hydrodynamic plain bearings. Rolling-element bearings are used for low to average speeds, the permissible speed decreasing as shaft diameter increases. The space occupied by a ball or roller bearing is relatively large. As a result of the comparatively long time before relubrication, maintenance expense is low. Bearings that support loads are subject to surface-pressure wear. Alternating tension and pressure stresses occur at the surface of the bearings. *Cracks* appear and propagate if the fatigue resistance of the bearing material is exceeded. In bearings subjected to high surface pressures, there may be Hertz-type stresses up to 3000 N/mm^2 [11]. Because of the great contact stresses between the rolling elements and the raceways, the bearing material must offer adequate strength at the operating temperature.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_399

The most common material for ball and roller bearings is steel (→ Steel). Important examples are listed in Table 1. Usually throughhardening chromium steels having a carbon content of ≈ 1 % and a chromium content of 0.5 – 1.5 % are used up to 160 °C. The chromium content must be increased as the cross section of the rolling elements increases, the hardness ranging from 58 to 65 HRC (Rockwell hardness). Toughness and dimensional stability, which depend on the kind of thermal treatment and on the material selected, are important properties that must be taken into account.

Hardening is mainly the result of the transformation of austenite into metastable martensite, a transition that requires sufficiently rapid cooling rates. Tempering relieves stress and causes carbides to precipitate, especially from the martensite. The service life of bearing steel tends to increase with hardness, up to a maximum. Although great hardness does inhibit formation of cracks, once formed cracks propagate more rapidly because of the concentration of stress. The difference in hardness between the rolling element and the race also determines the life of a rolling-element bearing. For steel bearings, differences of 1 – 2 Rockwell (70 – 80 Vickers) give good results.

Generally, the service life of ball and roller bearings can be prolonged by reducing the content of nonmetallic oxide and sulfide impurities and by preventing the segregation of carbides. This all can be accomplished by remelting the bearing steel, e.g., by electro-slag remelting or vacuum remelting in an arc furnace or an electron-beam furnace.

Case-hardened steels, for example, SAE 4720, are used for bearing components of large dimensions, especially when conventional steels are difficult to harden through. Whereas in the case of normal hardening the steel must contain sufficient carbon from the very beginning to ensure that the desired hardness is reached, case-hardened steels have low carbon contents, for example, 0.2 %. The surface is carburized by annealing in carbonaceous media. Often the graded microstructure, with the hardness decreasing from the surface to the core, can be subjected to the same stresses as completely hardened structures. The higher the nickel content of the case-hardened steels the more uniform the hardness profile from surface to core, even in cases of mild oil tempering or martempering. The composition of hardened steels also determines the maximum operating temperature, as is shown in Table 1.

Steels with high chromium contents, such as X90CrMoV18, are resistant to corrosion and tempering up to ≈ 450 °C due to the fact that precipitation of carbon from martensite is prevented. These martensitic stainless steels can be used in rolling-element bearings operating up to 300 °C. For operating temperatures of 450 °C or even 540 °C, steels that retain their hardness are used: those containing molybdenum (4 or 8 %, respectively) along with chromium, vanadium, and tungsten. These elements are all carbide formers. Essentially, such steels are high-speed tool steels austenitized around 1100 °C. Tempering is carried out at ≈ 560 °C to precipitate finely dispersed carbides and ensure good creep resistance.

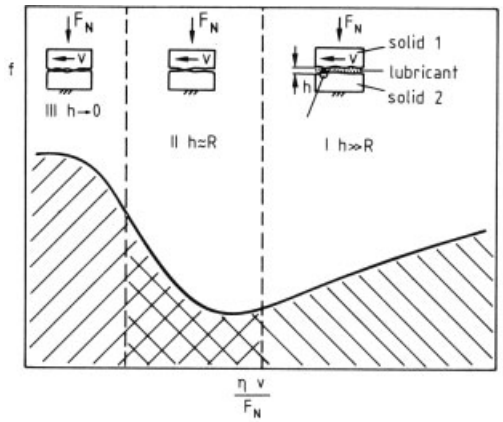

Figure 1. Schematic Stribeck curve and lubrication regions [7]
Region I. Hydrodynamic lubrication
Region II. Mixed lubrication
Region III. Boundary lubrication

For even higher operating temperatures, up to 800 °C, and oxidizing atmospheres, cobalt-base materials alloyed with chromium and tungsten, such as Stellite 3 or 100, are preferred. Basically they have an austenitic structure with a high carbide content. Even higher carbide contents are found in TiC cermet, where nickel acts as the metallic binder. Rolling-element bearings made of silicon nitride have been subjected to extraordinary high stress, e.g., in space flights. Plastic rolling bearings are used for lower specific loads at normal ambient temperatures, for example,

Table 1. Steels for rolling-element bearings [6]

Alloy designation	Specification standard*	Nominal composition, %		Maximum operating temperature, °C	Rockwell hardness, HRC	Characteristics; uses
105Cr2 50100	1 2	C 1.0 Si 0.25 Mn 0.35	Cr 0.5	120	58–65	small cross-section parts, e.g., needle rollers
100Cr6 52100 534A99	1 2 3	C 1.0 Si 0.25 Mn 0.35	Cr 1.5	160	58–65	most popular through hardening steel for ball, roller, and needle bearings
17CrNiMo6 4720 9310 665M17	1 2 2 3	C 0.1–0.2 Si 0.2–0.3 Mn 0.5–0.6	Cr 0.5–1.7 Ni 1.1–3.3 Mo 0.1–0.3	180		most used carburizing rolling-element steel, shock resistance and core hardness increase with nickel content
X90CrMoV18 X105CrMo17	1 1	C 0.9–1.1 Si 1.0 Mn 1.0	Cr 17–18 Mo 0.6–1.1 V 0–0.1	300	55–62	martensitic stainless steels for corrosive environment
440CM	2	C 1.0 Si 1.0 Mn 1.0	Cr 15 Mo 4.0 V 0.15	430	55–62	stainless steel, high retention of hardness
80MoCrV4216 M50	1 2	C 0.8 Si 0.25 Mn 0.35	Cr 4.0 Mo 4.0–4.25 V 1.0	310	58–64	intermediate high-temperature steels
BM-1	4	C 0.8 Si 0.3 Mn 0.3	Cr 4.0 Mo 8.0 V 1.0 W 1.5	450	58–64	
X82WMoCrV654	1	C 0.8 Si 0.4 Mn 0.4	Cr 4.0 Mo 4.25 V 1.9 W 6.5	540	58–64	high-speed high-temperature tool steels

* Standards: 1) DIN 17230 (FRG); 2) AISI-SAE Steel Classification (USA); 3) BS 970 (GB); 4) BS 4659.

in chemical plants. Conventional bearing cages made of plastic are becoming increasingly important.

3. Sliding Bearings

3.1. Tribology of Sliding Bearings

Since its definition in 1966, *tribology* has become widely recognized as a general concept embracing all aspects of the transmission and dissipation of energy and the dissipation of materials in mechanical equipment, including the topics friction, wear, and lubrication. The purpose of lubrication is to separate surfaces moving relative to one another with a film of material that can be sheared with low resistance without causing damage to the surfaces.

The mode of lubrication depends on the thickness of the lubricant. The *Stribeck curve* (Fig. 1) shows the general characteristic of lubricated moving surfaces and the relationship between friction coefficient f and a function of the lubricant viscosity η, the velocity v, and the normal load F_N (or pressure p), namely, $\eta\, v\, F_N^{-1}$. Surface roughness R, the operating conditions, and the separation of the surfaces h are also important factors. There are three main lubrication regions.

Fluid-film lubricated bearings are usually operated in region I, where the rigid surfaces are separated by a continuous lubricant film of thickness much larger than the combined surface roughness measure R. There is no direct physical contact between the surfaces, and wear processes cannot take place, except surface fatigue wear, cavitation wear, erosion, or corrosion. This is referred to as a *hydrodynamic lubrication.*

If at a constant load and lubricant viscosity the velocity is decreased, the lubricant film becomes thinner and the surfaces come closer together, producing the *mixed lubrication* associated with region II. The first asperity contacts occur: the load is carried partly by the fluid film and partly by the contacting asperities. In principle, all wear mechanisms, including abrasive wear, can take place. An even lower velocity ultimately results in considerably increased asperity interactions, the film thickness decreases down to a few molecules. In this region of *boundary lubrication,* the load is carried almost entirely by the deformation of the asperities, and the physical and chemical interactions at the solid – lubricant – solid interfaces determine the friction and wear. Although fluid-film lubricated bearings are designed to operate in region I, mixed and boundary lubrication do occur on starting and stopping machines. For some specific applications sliding bearings must be operated without lubricant – they must be run dry. The materials used have to be wearproof and scuffproof.

Sliding bearings do not operate reliably

1) if they run hot
2) if they show excessive wear

These two conditions are avoided by appropriate calculation, design, manufacture, and operation.

For hydrodynamically lubricated bearings, this involves:

1) selecting a lubricant of sufficiently high viscosity at operating temperature to reach fluid-film pressures sufficiently high to carry the load
2) avoiding excessive lubricant viscosity to reduce friction losses
3) assuring an adequate supply of lubricant and sufficient heat dissipation
4) designing and manufacturing bearings to produce a clearance between bearing and shaft appropriate for the formation of the required minimum fluid-film thickness.

Bearings operating in mixed and boundary lubrication regions are provided with reservoirs for the lubricant, e.g., impressed calottes or grooves in the sliding surface or pores of sintered material.

The demands on a sliding bearing are determined by the magnitude of forces to be transmitted and the space available for the bearing. The loads to be transmitted range from $\approx 10^2$ to 10^7 N for shaft diameters from 5 to 500 mm. The combination of load and available space for shaft and bearing produce specific loads that have to be taken up, statically and dynamically, by the bearing material as alternating and impact stresses. These stresses range over several orders of magnitude, reaching maximum values $> 100\,\mathrm{N/mm^2}$. Such stresses must always be considered in connection with the sliding velocities, which vary over a wide range [12]. The surrounding medium and the temperature must also be considered.

Even under usual, hydrodynamic conditions, the bearing materials are subjected to additional stresses as a result of cavitation, erosion, e.g., by dirt particles, and corrosion caused by additives or old lubricants.

Table 2. Definition of tribological terms (DIN 50282)

Term	Definition
Mechanical load limit	maximum possible load (force/projection area of bearing). If this limit is exceeded, the material will be subject to impermissible permanent deformation, forced rupture, or fatigue fracture
Load carrying capacity	load that a material can take with a specific kind of stress without exceeding the mechanical load limit and a defined amount of wear
Conformability	ability of a material to adapt to stresses by elastic or plastic deformations without permanently affecting the sliding behavior
Beddability	ability of a material to adapt to stresses by conformability and wear without permanently affecting the sliding behavior
Compatibility between bearing and shaft material	resistance to welding and scoring
Embeddability	ability of the bearing material to embed hard particles
Running-in characteristic	ability of a material to reduce the high initial friction and the high initial wear within a short time by adapting the sliding surfaces
Emergency-run characteristic	ability of a material to ensure sliding during a limited period of time if unforeseen unfavorable lubrication occurs

Deviations from normal operating conditions can result in solid contact between shaft and bearing in the mixed or boundary lubrication area, causing both elements to wear. Therefore, material properties and surface quality of shaft and bearing have to be selected with a view to the stress requirements. Specifications to aid in the

selection of compatible bearing and shaft materials are available [2]. The wide-ranging requirements for bearing materials have been defined for metallic antifriction materials in DIN 50282. The terms needed are defined in Table 2. The terms primarily describe properties that cannot be expressed numerically; the evaluations are qualitative – very good, good, etc.

3.2. Microstructure of Important Bearing Materials

Important properties of the base metals, including the modulus of elasticity, coefficient of thermal expansion, hardness, yield point, and fatigue resistance, both at room and elevated temperatures, correlate with the melting points of the metals. Therefore the melting point of the base bearing metal indicates application limits. Eespecially the mechanical properties depend also on the microstructure, which is affected by alloy constitution and production process.

The common metallic bearing materials have heterogeneous structures. Exceptions are the *copper – tin alloys* with less than 12 wt% tin, materials usually occurring as homogeneous supersaturated solid solutions (Fig. 2). The dissolved tin increases the load-carrying capacity because the solid solution is harder and more corrosion resistant to aged lubricants and additives. Grain-boundary hardening associated with the small grain size and work hardening increase the strength.

Figure 2. Microstructure of KS 934 (CuSn8P). The alloy is hardened by solid-solution hardening, grain-boundary hardening, and cold working, which is indicated by slip lines
100X

Heterogeneous bearing materials can be classified as follows:

1) materials where the microscopically distinguishable particles are softer than the matrix
2) materials where the particles are harder than the matrix
3) materials with both hard and soft particles in the matrix

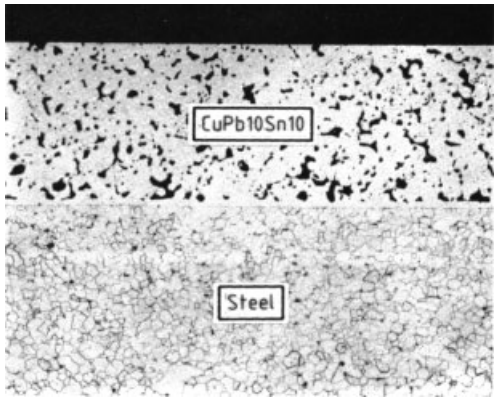

Figure 3. Microstructure of KS 940 S (CuPb10Sn10) with steel backing. Soft particles of lead (dark) in the solid-solution hardened matrix (bright)
100X

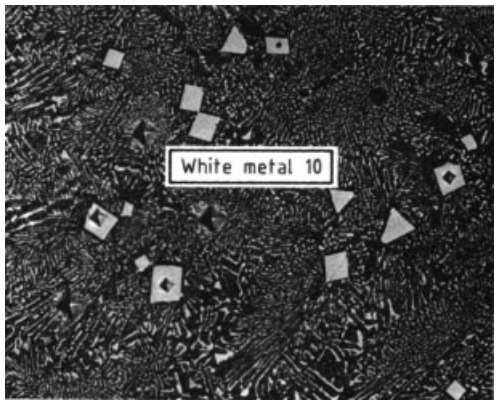

Figure 4. Microstructure of KS 904.1 (white metal 10). Hard cuboids of SbSn in a soft eutectic matrix. The differing hardness of the two phases is shown by the micro-hardness impressions
100X

Case 1 includes the soft lead-containing tin bronzes, the tin- and zinc-containing multicomponent bronzes, and the tin-containing aluminum alloys. The lead that is insoluble in copper is distributed as soft particles in the lead-containing tin bronze matrix (Fig. 3). The dissolved tin increases the matrix hardness. While

the mechanical load-carrying capacity decreases as the content of soft particles increases, the sliding properties improve.

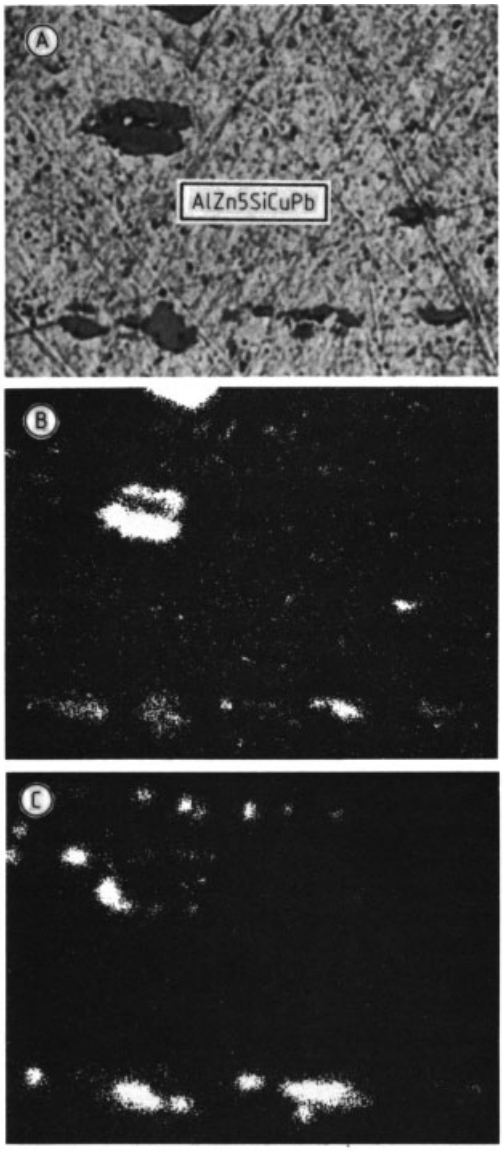

Figure 5. Microstructure of KS 960 (AlZn5SiCuPb) (Fig. 5 A). The distribution of Si (Fig. 5 B) and Pb (Fig. 5 C), revealed by microprobe analysis, indicate hard and soft phases
1500X

Case 2 includes the lead and tin alloys, the special brasses, and the aluminum alloy AlSi12CuMgNi. Figure 4 shows for the alloy PbSb15Sn10 the cubic particles of hard intermetallic SnSb in the soft eutectic matrix. Whereas hard iron – manganese silicides form in the special brass CuZn40Al2 during solidification, nearly all silicon in the high-silicon alloy AlSi12CuMgNi is distributed as hard particles. The matrix is precipitation-hardened by a fine distribution of magnesium- and silicon-containing particles that cannot be seen with an ordinary microscope.

Case 3 includes the aluminum alloy AlZn5SiCuPb. This alloy contains soft lead particles and hard silicon and aluminum silicide constituents, both in a matrix of medium hardness (Fig. 5). Also included in this group are the aluminum alloys AlSi4Cd and AlCd3CuNi, which are used in multilayer bearings.

The effects of the microstructure on some important properties are not yet understood quantitatively, especially because the structure may change considerably near the sliding surface during operation [13–15]. Qualitatively it is known that the usual hardening methods serve to improve the mechanical load carrying capacity, whereas the beddability essentially depends on the elastic and plastic properties of the base metal.

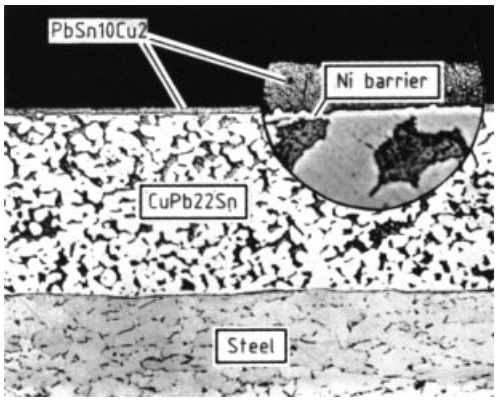

Figure 6. Microstructure of the multilayer bearing KS 941.1 S (CuPb22Sn) with overlay. The overlay of PbSn10Cu2 is deposited on steel-backed lead bronze
100X

Above all else the system properties are determined by the hard and soft phases. *Soft* components improve embeddability, the run-in characteristic, and the emergency-run characteristic. *Hard* phases increase wear resistance. The higher the content of hard-phase particles, the worse the embeddability and the emergency-run characteristic. Both hard and soft phases are able to limit scuffing. Under conditions of mixed and

boundary lubrication, these phases show little tendency to seize up [13].

Examples of the microstructure of *multilayer materials* are given in Figures 3 and 6. In both cases the steel supports the actual bearing material. For especially heavy loads, three-layer bearings are used: a layer, approximately 0.02 mm, of PbSn10Cu2 is electrodeposited onto a steel – lead bronze base (Fig. 6). This very soft, very thin overlay is then able to withstand the severe mechanical stresses. The overlay is beddable, embeddable, and serves as the sliding surface. The intermediate lead bronze ensures that the bearing remains operative even if the overlay is subjected to intensive wear.

Sintered metallic materials based on iron or copper contain a connected network of pores. The lubricant distributed in the pores serves as a reservoir for long service life, especially if the sliding velocities are low. These bearings are often referred to as maintenance-free bearings. Chemical compositions and microstructures that are not possible with melting techniques can be realized. An additional advantage is that solid lubrication can be employed.

The properties of *thermoplastics* can be determined by the selection of the monomers, the degree of cross-linkage, the degree of crystallization, and the polymerization process. The tribological behavior is determined by the micromolecular composition. The adhesion, which is undesirable for metallic materials, may be an advantage with polymers. For instance, PTFE firmly adheres to the steel shaft at low sliding velocities and high surface pressures, the individual lamellae or molecules of the PTFE sliding off each other, and the coefficient of friction is low [16], [17]. The principal disadvantages of polymers are their low thermal conductivity, their low wear resistance, and in many cases their high friction. Hard components (e.g., glass, metals, metal oxides) are used to increase the wear resistance; fillers (e.g., MoS_2, lead), to improve the sliding properties; and metallic particles (e.g., bronze), to increase the thermal conductivity. Since the load carrying capacity of thermoplastics is low, multilayer materials are often used. Figure 7 shows a plastic – metal multilayer bearing having a complex structure: A porous bronze structure (CuPb10Sn10) is sintered onto a steel strip. The pores of the bronze are filled with a PTFE – lead mixture by rolling.

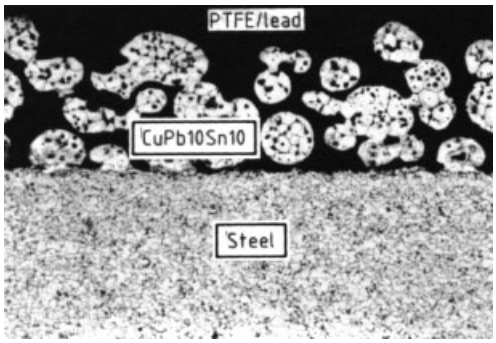

Figure 7. Microstructure of the plastic – metal multilayer bearing Permaglide P 10. The porous CuPb10Sn10 sintered on a steel strip is filled with a PTFE – lead mixture by rolling 100X

Unlike thermoplastics, *thermosetting plastics* are tightly cross-linked by covalent or ionic bonds. Solid lubricants – graphite, MoS_2, PTFE – are used to improve sliding behavior.

The lubrication by *graphite* or *artificial carbon* running against steel shafts arises from their layer structure and the H_2O molecules present in the graphite lattice. These reduce the bonding forces between the lattice planes, and the planes easily slide against one another.

The strong covalent bond gives *ceramic* materials high compression strength, resistance to chemicals, and high melting temperatures. These properties make possible employing them as bearing materials, especially in aggressive media and at high temperatures.

3.3. Metallic Materials for Sliding Bearings

Three groups of metallic materials are used in bearings:

1) Metallic material for solid sliding bearings. These bearings consist of only one material. They are machined from castings or cold drawn tubes or roll-formed from strips. Copper, aluminum, tin, and lead alloys are used.
2) Multilayer metallic materials for thin-walled sliding bearings. These bearings are made from two or more layers of different composition. The preferred manufacturing processes include composite casting, composite sintering, cladding by rollbonding, and

galvanic deposition. Most common is steel-backed lead bronze. Also used is steel-backed aluminum alloy. Both are sometimes overlaid by lead – tin or lead – tin – copper alloys.

3) Sintered materials. Generally, sintered bearings consist of only one alloy, which is often based on iron or copper. They are manufactured by sintering of powdered metal to form shapes of defined porosity.

3.3.1. Metals for Solid Bearings

Lead and Tin Alloys. Combined with shafts usually made of unhardened steel, the lead and tin alloys offer good sliding properties. Because of their low strength, they require backings, mostly cast iron or steel, sometimes bronze, red brass, or brass. The composition and the most important properties of the lead and tin alloys known as *babbitt metals* are given in Table 3. Babbitt metals are used for bearings of big engines and heavy machinery. The specific load can be kept low by use of a large bearing, especially in the case of stationary machinery. For example, bearings for steam-turbine rotors carry specific loads of $\approx 1\ \mathrm{N/mm^2}$ at sliding velocities $\leqq 150$ m/s. The group of babbitt metals have the lowest melting point and the lowest hardness of the metallic bearing materials. The low hardness gives good embeddability for any dirt in the oil. The ductility allows good conformability to the shaft. The excellent sliding properties and emergency-run characteristic result from the microstructure. Babbitt metals offer good corrosion resistance to lube oils [18], [19].

The disadvantage of babbitt metals is that their mechanical load limit decreases sharply at higher temperatures, the maximum operating temperature being $\leqq 140\,^{\circ}\mathrm{C}$. Their low fatigue strength can be increased to some extent by design, manufacturing methods, and the use of modified alloys. Up to 0.15 % arsenic increases the strength of tin alloys. Arsenic concentrations up to and exceeding 1 % provide higher fatigue strength for lead alloys at elevated temperatures.

Reaction with the lube oil forms an oxide layer on lead alloys. This film gives lead alloys a higher resistance to scuffing at elevated temperatures than the tin alloys have. The stationary casting process or the more economic centrifugal casting process is used to manufacture bearing linings up to 6 mm thick from lead and tin alloys. Rapid dissipation of heat during solidification creates a fine-grained microstructure with good sliding properties and sufficient load carrying capacity.

Copper Alloys. The standard copper alloys for solid bearings are given for cast materials in Tables 4, 5, and 6 and for wrought materials in Table 7. The important cast materials are listed in order of increasing load carrying capacity, the result of increasing hardness, wear resistance, and fatigue strength.

Soft *lead – tin casting bronzes* (Table 4) have relatively low hardness but excellent emergency-run characteristic. In these respects they are similar to babbitt metals. They can withstand loads of $\approx 40\ \mathrm{N/mm^2}$ at average sliding velocities of ≈ 3 m/s. The shafts must be made from hardened steel. The manufacture of solid bearings from this and other copper alloys is usually by centrifugal or continuous casting. Lead contents > 20 % in alloys containing > 5 % tin are avoided because of the enlarge liquid miscibility gap. A fine dispersion of soft lead-rich eutectic particles is distributed throughout the cavities of the skeleton-type matrix. Too many such soft particles reduce the strength. Multilayer bearings are used when high mechanical load limits and good emergency-run characteristic are required.

Partial replacement of expensive tin by zinc combined with reduced lead content produces *copper – tin – zinc – lead casting alloys* (Table 5). The zinc improves the castability. These bronzes are usually produced for general mechanical engineering, where they are used for a variety of unproblematic uses involving small loads, low sliding velocities, and low operating temperatures. However, good lubrication is needed.

Tin bronzes, such as CuSn10P, or tin bronzes with low lead contents, such as CuSn12Pb2, are suitable for higher loads. Although their emergency-run characteristic is only average, these hard bearing materials have high load carrying capacity, offer good wear resistance and fatigue strength, and are insensitive to impact. Phosphorus prevents the oxidation of tin. Particles of the hard tin oxide distributed in the

Table 3. Lead and tin casting alloys (babbitt metals)

Alloy designation	Specification standard*	Nominal composition, %	Melting range, °C	Yield strength, N/mm²	Modulus of elasticity, 10^3 N/mm²	Brinell hardness, HB	Characteristics; uses
PbSb10Sn6 No. 13 Grade 19	1, 5 2 3	Sb 9–11 Sn 5–7 rest Pb	240–260	39	30	16	corrosion resistant, low loads at typical sliding speed
PbSb15Sn10 No. 14 Grade 7	1, 5 2 3	Sb 14–16 Sn 9–11 rest Pb	240–270	43	31	21	NH_3 resistant, average loads; crosshead blocks, crossheads
PbSb15SnAs No. 15 Grade 15	1, 5 2 3	Sb 13.5–15.5 Sn 0.9–1.7 As 0.8–1.2 rest Pb	240–350	39	27	18	low loads at low sliding speeds; camshaft bushes, main and connecting rod bearings in compressors
PbSb14Sn9CuAs	1	Sb 13–15 Sn 8–10 Cu 0.7–1.5 As 0.3–1.0 rest Pb	240–420	46	54	22	good sliding behavior even for mixed friction region, high thermal loads; bearings in gears and roll mills
SnSb8Cu4 No. 12 Alloy No. 2 No. 1	1, 5 2 3 4	Sb 7–8 Cu 3–4 Sn 88–90	233–360	47	54	22	average loads at high sliding speeds, good compatibility; roll mill bearings
SnSb8Cu4Cd	1	Sb 7–8 Cu 3–4 Cd 0.8–1.2 Sn 88–90	233–360	62	54	28	corrosion resistant, high loads at high sliding speeds; main and connecting rod bearings for large piston machinery
SnSb12Cu6Pb	1	Sb 11–13 Cu 5–7 Pb 1–3 Sn 79–81	183–400	61	55	25	average loads at high sliding speeds, high wear resistance; turbine bearings

* Standard: 1) ISO 4381 (international), DIN/ISO 4381 (FRG); 2) SAE-J 460e/J 459c (USA); 3) ASTM B 23 (USA); 4) BS 3332 (GB); 5) ISO 4383, DIN/ISO 4383.

bearing material may cause shaft scoring and increase the tendency to seize.

Special casting brasses and *multicomponent aluminum casting bronzes* (Table 6) belong to the group of high load carrying copper casting materials. The latter can carry even higher loads than the former. Because of their poor emergency-run characteristic, they can only be used if lubrication is good. The wear resistance of special brasses [20] is limited, and their fatigue strength diminishes when they are subjected to corrosion. For extremely high bearing loads at low sliding velocities, the wear- and corrosion-resisting aluminum multicomponent bronzes are effective even when there are impact and shock stresses. Their limited beddability and compatibility between bearing and shaft material require precision machining.

Copper wrought alloys are available as tubes or strips, and they are especially economical in the production of large quantities of solid bearings in the form of bushings or bushes and thrust washers. Tubes or strips for wrapped bushes are manufactured with as close a tolerance as possible from the materials listed in 7 by extruding and cold drawing or rolling of billets made by continuous or stationary casting. In this way machining times and material wastage in working the tin bronze CuSn8P and the special brass CuZn31Si1 are minimized. Slightly greater tolerances must be expected for materials such as CuZn40Al2 and most multicomponent aluminum bronzes. While casting materials cannot be work hardened, the mechanical properties of the copper wrought alloys can be improved by cold deformation.

Table 4. Lead – tin casting bronzes for solid and multilayer bearings*

Alloy designation	Specification standard **	Nominal composition, %	Yield strength Rp 0.2, N/mm²	Tensile strength Rm, N/mm²	Elongation A_5, %	Modulus of elasticity, 10^3 N/mm²	Brinell hardness, HB	Characteristics; uses
CuPb20Sn5 LB 5-C	1, 5, 6 2	Pb 18–23 Sn 4–6 rest Cu	80	180	7	75	50	difficult to cast, resistant to sulfuric acid, excellent sliding properties; cold rolling and foil rolling mills, water pumps
CuPb15Sn8 LB 1-C CA 938 B 144-3D	1, 5, 6 2 3 4	Pb 13–17 Sn 7–9 rest Cu	100	220	8	85	65	resistant to sulfuric acid, moderate pressure, and high speeds, insensitive to impact, shock, and compression across the edges; valves
CuPb10Sn10 LB 2-C CA 937 B 144-3A	1, 6 2 3 4	Pb 8–11 Sn 9–11 rest Cu	110	220	6	90	70	heavily loaded plain bearings, insensitive to compression across the edges; vehicles, rolling mills
CuPb9Sn5 LB 4-C CA 935 B 144-3C	1, 6 2 3 4	Pb 8–10 Sn 4–6 rest Cu	130	230	9	85	60	corrosion and acid resistant, high speeds, high loads

* The mechanical properties are minimum values for continuous-cast material.
** Standards: 1) ISO 4382 (international), DIN/ISO 4382 (FRG); 2) BS 1400 (GB); 3) SAE-J 462 (USA); 4) ASTM (USA); 5) DIN 1716 (FRG); 6) ISO 1338 (international).

The load carrying capacities of the various wrought and casting alloys are similar. Bearings made of tin bronzes such as CuSn8P, combined with hardened steel shafts, can be used to transmit high loads. Because of the hardness of this material, special attention should be paid to surface quality, the prevention of machining and alignment errors, and oil contamination. The load carrying capacity and the wear resistance of the special brass CuZn31Si1 are inferior to those of CuSn8P in spite of their similar mechanical properties. However, this low-priced special brass has proved its effectiveness for average loads and is widely used for wrapped bushes.

Aluminum Alloys. Solid bearings made from aluminum alloys are used to a limited extent only. Because the coefficient of thermal expansion of aluminum alloys is higher than that of the bearing housing material, e.g., cast iron and steel, problems develop at high temperatures regarding the press fit in the housing and the bearing clearance. However, the corrosion resistance of aluminum alloys is superior to that of copper materials, especially in sulfur-containing lube oils.

Aluminum alloys are listed in Table 8 in order of increasing load carrying capacity. The alloy AlSn6Cu, which contains soft tin particles, belongs to the bearing materials having good sliding properties. The harder alloy AlZn5SiCuPb can take slightly higher loads at average sliding velocities. Both materials are used for the main bearings of combustion engines with specific loads of $\approx 40\,\mathrm{N/mm^2}$ at sliding velocities of $\approx 15\,\mathrm{m/s}$, although such a high sliding velocity requires an additional overlay. The alloy AlSi12CuMgNi is a high wear-resisting material that permits high loads. However, the emergency-run characteristic is less good, and only low sliding velocities are permissible at high loads. This alloy is then preferred for high, even impact-type load conditions involving oscillating sliding motion, e.g., piston pin bushes with specific loads $\leqq 80\,\mathrm{N/mm^2}$ and a sliding velocity of $\approx 0.2\,\mathrm{m/s}$. This alloy is also used for floating bushes, e.g., in waste gas turbochargers. In this case the specific load of about $0.5\,\mathrm{N/mm^2}$

Table 5. Tin bronze and leaded tin bronze casting alloys*

Alloy designation	Specification standard **	Nominal composition, %	Yield strength Rp 0.2, N/mm²	Tensile strength Rm, N/mm²	Elongation A_5, %	Modulus of elasticity, 10^3 N/mm²	Brinell hardness, HB	Characteristics; uses
CuSn10P PB 1-C CA 907	1, 6 2 3	Sn 10–11.5 P 0.5–1 Cu ≥ 89	170	330	4	95	95	high loads, shock and impact resistant, seawater resistant
CuSn8Pb2	1, 6	Sn 6–9 Pb 0.5–4 Cu ≥ 82	130	230	4	75	85	moderate loads, sufficient lubrication required
CuSn12Pb2 (CA 927)	1, 5, 6 3	Sn 11–13 Pb 1–2.5 Cu ≥ 84	150	280	5	95	90	high load, good emergency-run characteristic, resistant to impact and shock, resistant to corrosion and seawater
CuSn10Zn2 G 1-C CA 905 B 143-1A	5, 6 2 3 4	Sn 9–11 Zn 1–3 Cu ≥ 86	140	270	7	85	75	corrosion-resistant to seawater; bushes for heavy loads and low speeds
G 2-C CA 923 B 143-2B	2 3 4	Sn 7.5–9 Zn 3–5 Pb 0.8–1.5 Cu ≥ 85	115	240	15	97	60	better sliding behavior than CuSn10Zn2, otherwise similar
CuSn5Zn5Pb5 LG 2-C CA 836 B 145-4A	1, 5, 6 2 3 4	Sn 4–6 Zn 4–6 Pb 4–6 Cu ≥ 84	100	250	13	90	65	bearings and components under normal load, water pump fittings
CuSn7Pb7Zn3 CA 932 B 144-3B	1, 5, 6 3 4	Sn 6–8 Pb 5–8 Zn 2–5 Cu ≥ 81	120	260	12	85	70	normal load bushes for general mechanical engineering, electric motors, and ship building

* The mechanical properties are minimum values for centrifugal cast material.
** Standards: 1) ISO 4382 (international), DIN/ISO 4382 (FRG); 2) BS 1400 (GB); 3) SAE-J 462 (USA); 4) ASTM (USA); 5) DIN 1705 (FRG); 6) ISO 1338 (international).

is so low that a maximum sliding velocity of 25 m/s can be permitted.

3.3.2. Materials for Metallic Multilayer Thin-Walled Bearings

The plain bearings in combustion engines have to fulfill demands that cannot be met by one material: high load carrying capacity and fatigue strength at temperatures up to ≈ 180 °C must be combined with high sliding velocities. The emergency-run characteristic must be adequate, and beddability and embeddability should be good. Despite small dimensions and wall thicknesses, the press fit in the housing bore must be tight even at high operating temperatures.

Such requirements are best met by multilayer materials. The steel backing ensures perfect press fit in the housing bore and gives the composite material the necessary fatigue strength by backing the bearing alloy, thus increasing its mechanical load limit. Then it is possible to use soft bearing alloys based on lead, tin, copper, and aluminum, all alloys characterized by good emergency-run characteristic but low fatigue strength, to produce high load carrying sliding bearings.

The steel backing usually consists of a low-carbon steel whose composition depends on the manufacturing process chosen for the production of the multilayer plain bearing. Since multilayer plain bearings, especially for small and medium-sized engines, are produced in large quantities, the production starts from steel strips coated with bearing alloy. The coating is carried out in a continuous process where the ac-

Table 6. Special copper base casting alloys for solid bearings*

Alloy designation	Specification standard **	Nominal composition, %	Yield strength Rp 0.2, N/mm²	Tensile strength Rm, N/mm²	Elongation A_5, %	Modulus of elasticity, 10^3 N/mm²	Brinell hardness, HB	Characteristics; uses
CuZn35Al1 B 147-7 A HTB 1-C (CA 865)	1 3 4 6	Cu 56–65 Al 0.5–2 Fe 0.5–2 Mn 0.3–3 Zn ≥ 34	200	500	18	100	110	moderate sliding properties
CuZn25Al5 B 147-8C HTB 3-C CA 863	1 3 4 6	Cu 60–67 Al 3–7 Fe 1.5–4 Mn 2.5–5 Zn 22–28	480	750	5	110	190	moderate sliding properties; heavily loaded, low-speed bearings
CuAl10Fe B 148-9A AB 1-C CA 954	2 3 4 6	Cu ≥ 83 Al 8–11 Fe 2–4	200	550	15	115	115	high temperature, very high specific pressure at small sliding speeds, wear and corrosion resistant
CuAl11Ni AB 2-C	2 4	Cu ≥ 73 Al 9–12.3 Ni 5–7.5 Fe 4–7	400	750	5	120	185	very highly stressed bearings especially for momentary high loads, high wear, corrosion, cavitation, and erosion resistant
CuAl10Fe5Ni5 B 148-9D CA 955	5 3 6	Cu ≥ 78 Al 8–11 Fe 3.5–5.5 Ni 3.5–5.5	280	680	12	120	140	tends to not embed dirt, wear and corrosion resistant; structurals parts

* The mechanical properties are minimum values for centrifugal cast material.
** Standards: 1) DIN 1709 (FRG); 2) DIN 1714; 3) ASTM (USA); 4) BS 1400 (GB); 5) ISO 4382 (international), DIN/ISO 4382 (FRG); 6) SAE-J 462 (USA).

tual bearing material is cast on, sintered on, or rollbonded to the steel strip.

The thickness of the bearing alloy is usually 0.2 – 0.6 mm when the bearings are finished. Often an additional, ≈ 0.02-mm-thick electroplated overlay of tin-containing lead alloy is provided to further improve the sliding properties, thus making a three-layer metal bearing.

Steel-Backed Lead and Tin Alloys. Lead and tin alloys for thin-walled steel composite bearings, such as PbSb15SnAs and SnSb8Cu4 (Table 3), owe their use to their excellent emergency-run characteristic as well as their good compatibility and embeddability. Hardened steel shafts are not required. However, the use of these multilayer bearings is restricted because of their low fatigue strength and permissible temperature, these limits being slightly higher for lead alloys than for tin alloys. Bearing metal thicknesses of 0.1 – 0.5 mm can carry specific loads up to ≈ 18 N/mm². The lead or tin alloys for the manufacture of steel composite materials are cast onto a tinned steel strip to get good bonding. The rapid heat dissipation during manufacture produces a fine-grained microstructure with good sliding properties.

Steel-Backed Copper Alloys. Lead bronzes containing tin (Table 9) have proved especially successful as bearing materials when combined with steel backing. Materials containing a large proportion of the soft structure component, e.g., those with high lead contents, are characterized by good emergency-run characteristic and embeddability. This permits high sliding velocities. The higher the tin content the higher the hardness, the load carrying capacity, and the corrosion resistance.

There are two different processes for the industrial production of steel-backed composite materials: (1) the strip casting process, where the copper alloy is cast onto a steel strip, and (2) the sintering – rolling – sintering process, where

Table 7. Wrought copper alloys for solid bearings*

Alloy designation	Specification standard **	Nominal composition, %	Yield strength Rp 0.2, N/mm²	Tensile strength Rm, N/mm²	Elongation A_5, %	Modulus of elasticity, 10^3 N/mm²	Brinell hardness, HB	Characteristics; uses
CuSn8P CA 521 B 139-Alloy C PB 104	1, 5 2 6 7	Sn 7.5–9 P 0.1–0.4 Cu 90–92.5	200–480	400–580	10–55	115	80–160	high loads at average sliding speeds, wear and corrosion resistant; automotive engineering
CuSn4Pb4Zn4 CA 544 B 139-Alloy B 2	5 2 6	Sn 3–5 Pb 3–5 Zn 1–5 Cu 88	275	415	20	105	120	high loads at low-to-average sliding speeds, wear resistant; general engineering
CuZn31Si1	1, 3	Zn 28–32 Si 0.7–1.3 Cu 66–70	240–450	440–560	10–30	105	100–160	high loads at average sliding speeds; bushes for vehicle construction, e.g., spring suspension
CuZn37Mn2-Al2Si CA 674	1, 5 2	Zn 32–40 Mn 1.5–3.5 Al 1–2.5 Si 0.3–1.3 Cu 57–60	300	600	15	100	150	highly wear resistant at high loads, corrosion resistant; tool and agricultural machines
CuAl9Fe4Ni4 CA 630 B 150-Alloy 2 CA 105	1, 5 2 6 7	Al 8–11 Fe 2.5–4.5 Ni 2.5–5 Cu 78–87	400	700	15	118	160	heavy load, good wear resistance, excellent corrosion resistance; aircraft and marine parts
CuZn40Al2	3, 4	Zn 56–59 Al 1.3–2.3 Mn 1.4–2.6 Si 0.3–1.0 Cu 56.5–59	240–310	540–640	10–18	105	150–170	lower wear resistance than CuZn37Mn2Al2Si, otherwise similar
CuAl10-Fe3Mn2 B 150-Alloy 1 CA 103 CA 623	4, 5 6 7 2	Al 8–11 Fe 2–4 Mn 1.5–3.5 Cu ≥ 80	250–340	590–690	7–12	110	150–180	high strength at elevated temperatures, excellent wear resistance, acid resistant; forging presses, apparatus
CuAl11Ni6Fe5 CA 104	4 7	Al 10.5–11.5 Ni 5–7 Fe 4.8–7.3 Cu ≥ 70	270–390	640–740	10–15	120	190–240	extreme loads, resistant to scaling, cavitation, erosion, and corrosion

* Some of the mechanical properties are taken from DIN 17671 and DIN 17672.
** Standards: 1) ISO 4382 (international), DIN/ISO 4382 (FRG); 2) SAE-J 461d/I (USA); 3) DIN 17660 (FRG); 4) DIN 17665; 5) ISO 427/ISO 428 (international); 6) ASTM (USA); 7) BS 2872 (GB).

atomized powdered alloy is sintered on the steel strip in a three-step process. Specifically, the copper alloy particles are spread on the steel strip and sintered in an inert-gas atmosphere. The porous alloy layer is compacted by rolling and finally sintered again in an inert-gas atmosphere.

The alloy CuPb30 has an excellent emergency-run characteristic because of its high lead content; however, load carrying capacity and corrosion resistance are limited. This alloy is employed for wrapped bushes for hydraulic pumps and automatic gears. A lead bronze that finds widespread use in overlaid multilayer bearings is CuPb24Sn. It has good emergency-run characteristic and good fatigue strength owing to the steel backing. Main bearings and connecting rod bearings in combustion engines are its most important use. Specific loads $\leqq 70\,\mathrm{N/mm^2}$ and sliding velocities $\leqq 20\,\mathrm{m/s}$ are encountered.

Table 8. Aluminum alloys for solid bearings*

Alloy designation	Specification standard **	Nominal composition, %	Yield strength Rp 0.2, N/mm²	Tensile strength Rm, N/mm²	Elongation A_5, %	Brinell hardness, HB	Characteristics; uses
AlSn6Cu No. 770	1, 2 3	Sn 5.5–6.5 Cu 1.3–1.7 Ti 0.05–0.2 rest Al	60	130	20	40	for average loads and sliding speeds, high speeds possible with overlay; main bearings for piston engines
AlZn5SiCuPb	–	Zn 5.0–5.5 Si 1.2–2.0 Cu 0.9–1.2 Pb 0.7–1.3 Mg 0.4–0.6 Ti 0.02–0.15 rest Al	100	180	19	50	higher load carrying capacity than AlSn6Cu, otherwise similar
AlSi12CuMgNi	–	Si 11.0–13.0 Cu 0.8–1.5 Mg 0.8–1.3 Ni ≤ 1.3 rest Al	190	200	1	90	for high loads at low sliding speeds; piston pin bushes for compressors

* The mechanical properties are minimum values for extruded material.
** Standards: 1) ISO 6279 (international); 2) DIN/ISO 6279 (FRG); 3) SAE-J 459/J 460 (USA).

Table 9. Copper multilayer materials *

Alloy designation	Specification standard **	Nominal composition, %	Method of manufacture	Brinell hardness, HB	Characteristics; uses
CuPb10Sn10 No. 792 No. 797	1, 2 3 3	Pb 9–11 Sn 9–11 rest Cu	cast, sintered	70–130 60–90	high fatigue and impact strength, wear resistant, limited emergency-run characteristic, hard shafts required; wrapped bushes, thrust washers, piston pin bushes, gear boxes
CuPb17Sn5	1, 2	Pb 14–20 Sn 4–6 rest Cu	cast	60–95	slightly lower fatigue strength than CuPb10Sn10, otherwise similar, positive emergency-run characteristic; rocker valve bushes
CuPb24Sn4 No. 794 No. 799	1, 2 3	Pb 19–27 Sn 3–4.5 rest Cu	cast, sintered	60–90 45–70	moderate fatigue strength, suitable for high speeds, good emergency-run characteristic; crankshaft bushes, washers in gear boxes
CuPb24Sn No. 49	1, 2 3	Pb 19–27 Sn 0.6–2 rest Cu	cast, sintered	55–80 40–60	moderate fatigue strength, good emergency-run characteristic; main and connecting rod bearings in combustion engines, preferably with overlay
CuPb30 No. 48	1, 2 3	Pb 26–33 rest Cu	sintered	30–45	moderate fatigue strength, excellent emergency-run characteristic; wrapped bushes for hydraulic pumps

* The steel backing usually consists of low-carbon steels.
** Standards: 1) ISO 4383 (international); 2) DIN/ISO 4383 (FRG); 3) SAE-J 459/J 460 (USA).

The alloy CuPb24Sn4 is slightly harder due to its greater tin content. It is suitable for average to high load carrying capacity combined with a good emergency-run characteristic. The corrosion resistance is good. This alloy is used in wrapped bushes and thrust washers under loads $\leqq 50\,N/mm^2$ and sliding velocities $\leqq 10\,m/s$, for example, in gears, piston compressors, and pumps. The alloy CuPb10Sn10 is characterized by a high mechanical load limit combined with a high wear resistance. The high tin content is the reason for the corrosion resistance to old

lube oils. The emergency-run characteristic is, however, low, so only low sliding velocities are permissible. Typical applications are piston pin bushes for combustion engines (specific loads of 140 N/mm^2 at a sliding velocity of 0.2 m/s) as well as wrapped bushes and thrust washers for gears.

Steel-Backed Aluminum Materials. The positive effects of a steel backing have permitted aluminum alloys, with their normally restricting properties at elevated temperatures, to gain in significance. The most important components are tin, copper, silicon, zinc, and cadmium. Generally corrosion resistance to sulfur-containing lube oils is good. Tin forms soft structure constituents and improves the emergency-run characteristic; silicon, like copper, increases the load carrying capacity. Industrial production of these multilayer composite materials involves cladding by roll-bonding. The aluminum alloy and steel are bonded at various temperatures in the roll gap while subject to great deformation. The aluminum alloys for multilayer materials are compiled in Table 10.

In the case of the alloy AlSn20Cu finely distributed tin particles are embedded in an aluminum – copper matrix. This high tin content ensures an excellent emergency-run characteristic even at high sliding velocities, but the fatigue life is limited, and therefore only average loads are possible. The most important uses are main bearings and thrust washers in combustion engines. For specific loads $\leqq 45$ N/mm^2 and sliding velocities $\leqq 20$ m/s, no additional overlay is necessary. Owing to the good emergencyrun characteristic as well as the excellent wear resistance, AlSn20Cu is also employed for all kinds of pumps for liquid. The material AlSn6Cu is suitable for high loads and has good emergency-run characteristic. It is used in main and connecting rod bearings of combustion engines. An additional overlay is necessary, and hardened steel shafts are recommendable. The load limit is ≈ 60 N/mm^2 for sliding velocities $\leqq 20$ m/s.

In respect to properties and applications the alloys AlSi4Cd and AlCd3CuNi are similar to AlSn6Cu. A material of high strength but limited emergency-run characteristic is the alloy AlSi11Cu, which is used in heavy load carrying connecting rod bearings in turbo-charged diesel engines. In this case, an additional overlay is necessary. Bearing loads $\leqq 70$ N/mm^2 and sliding velocities $\leqq 20$ m/s are permissible. The advantages of the alloy AlZn4,5SiCuPb are high load carrying capacity and good wear resistance. The emergency-run characteristic is adequate. It is also used for the heavy load carrying connecting rod bearings in turbo-charged diesel engines. This alloy requires an additional overlay as well as hardened steel shafts. The permissible bearing loads and sliding velocities are similar to those of the alloy AlSi11Cu.

Overlay Alloys. Steel-backed copper or aluminum alloys for main and connecting rod bearings in internal combustion engines almost always require additional electrodeposited sliding layers about 0.02 mm thick. Thicker layers are possible, though the fatigue life is shortened. Between the bearing alloys and the overlay there is a nickel barrier electrodeposited to a thickness of ≈ 0.002 mm. This layer (1) prevents diffusion of tin, which would lead to the formation of brittle phases in the copper alloy, or (2) ensures perfect adhesion to the aluminum alloy. The electroplated overlay alloys are compiled in Table 11. The alloy PbSn10Cu2 is probably the most widely used. Its tin content makes it corrosion resistant to old lube oils. The copper content increases the load carrying capacity. Greater copper contents, up to ≈ 6 %, have been used, and even higher contents, up to ≈ 15 %, are used in special cases. Such overlays give good operating results with heavy load carrying connecting rod bearings in turbo-charged diesel engines. The properties of PbSn10 are similar to those of PbSn10Cu2, but it has a slightly lower load carrying capacity. The PbIn7 overlay is comparable to PbSn10. The nickel barrier is not needed with PbIn7.

3.3.3. Sintered Materials

An important characteristic of sintered material is the pore volume, which ranges from ≈ 15 to ≈ 25 %. The materials listed in Table 12, which have a pore volume of ≈ 20 %, are the ones most widely used. The powder-metallurgy method of production results in a coherent pore network that has a high storage volume for lubricants owing to the capillary effect. The base metals for the sintered materials are powdered copper and

Table 10. Aluminum multilayer materials *

Alloy designation	Specification standard **	Nominal composition, %	Brinell hardness, HB	Characteristics; uses
AlSn20Cu No. 783	1, 2 3	Cu 0.7–1.3 Sn 17.5–22.5 rest Al	30–40	average fatigue strength with excellent emergency-run characteristic, good corrosion resistance; soft shafts, nodular cast-iron shafts, main bearings and thrust washers for combustion engines
AlSn6Cu No. 780	1, 2 3	Cu 0.7–1.3 Sn 5.5–7 rest Al	35–45	high to average fatigue strength, good emergency-run characteristic, good corrosion resistance; hard shafts required; overlaid for main and connecting rod bearings of internal combustion engines
AlSi4Cd No. 781	1, 2 3	Cu 0.05–0.15 Cd 0.8–1.4 Si 3.5–4.5 rest Al	30–40	slightly lower compatibility than AlSn6Cu, but similar properties and application
AlCd3CuNi No. 782	1, 2 3	Cu 0.7–1.3 Ni 0.7–1.3 Cd 2.7–3.5 rest Al	35–55	properties and application of AlSi4Cd
AlSi11Cu	1, 2	Cu 0.7–1.3 Si 10.0–12.0 rest Al	45–60	high fatigue strength, limited emergency-run characteristic, good corrosion resistance, hard shafts required; overlaid for main and connecting rod bearings of internal combustion engines
AlZn4,5SiCuPb	–	Cu 0.9–1.2 Si 1.0–2.0 Zn 4.4–5.0 Pb 0.7–1.3 rest Al	45–60	better emergency-run characteristic than AlSi11Cu, but similar properties and application

* The steel backing usually consists of low-carbon steel.
** Standards: 1) ISO 4381 (international); 2) DIN/ISO 4381 (FRG); 3) SAE-J 460e (USA).

Table 11. Electroplated overlays for multilayer materials

Alloy designation	Specification standard *	Nominal composition, %	Characteristics; uses
PbSn10Cu2 No. 19	1, 2 3	Sn 8.0–12.0 Cu 1.0–3.0 rest Pb	good fatigue strength because of thin overlay, good compatibility, very good emergency-run characteristic, and good corrosion resistance; third layer for main and connecting rod bearings for internal combustion engines
PbSn10 No. 19	1, 2 3	Sn 8.0–12.0 rest Pb	
PbIn7	1, 2	In 5.0–10.0 rest Pb	

* Standards: 1) ISO 4383 (international); 2) DIN/ISO 4383 (FRG); 3) SAE-J 459c/J 460e (USA).

iron. The sintered bronze SINT B 50 has a composition comparable to that of conventional tin bronze and is characterized by a relatively good dynamic load carrying capacity of $\approx 5\,N/mm^2$ at 1 m/s. Graphite in the SINT B 51 material improves the low noise properties at the expense of the load carrying capacity. Compared with SINT B 50 sintered bronze, SINT B 00 sintered iron is characterized by a lower load carrying capacity, of $\approx 3\,N/mm^2$ at 1 m/s. Higher loads are permissible if copper, $\leqq 5$ %, is present. More copper changes the properties and makes them similar to those of sintered bronze. Depending on the lube oil, the maximum service life temperature of sintered bearing materials ranges between 90 and 180 °C [22]. Because of their low

noise level, sintered bearings are employed in almost all fields of technology, especially for fine mechanics and in automobiles. Solid lubricants other than graphite, e.g., sulfides, have been tried [23]. Bearings of sintered materials are produced by compacting mixtures of powdered metals in press molds and then sintering below the melting point of the main alloy components. Post-calibration ensures that bearings of high accuracy are obtained [24].

3.4. Plastics for Sliding Bearings

Plastics are used for plain bearings if their special properties, e.g., chemical resistance, insulating capacity, or attenuation capacity, are required. This occurs where plain bearings are operated with mixed or boundary lubrication or have to run dry, or where their use offers economic advantages. The main applications for dry bearings are the following:

1) vacuum operations or aggressive media, where there are no fluid lubricants
2) food industry, etc., where contamination by lubricants must be avoided
3) maintenance-free operations, where there is no checking or changing of lubricant

Plastics can be used for both solid and multilayer bearings.

3.4.1. Plastics for Solid Bearings

If compared with metals, *thermoplastics* (Table 13) are characterized by low density, low modulus of elasticity, good resistance to corrosion, high attenuation, and low lubricant requirements. Their disadvantages are low mechanical load capacity, inferior thermal conductivities, and high coefficients of thermal expansion. These three drawbacks and the considerable swelling in moist atmosphere have to be taken into account when designing the bearing clearance.

Fillers may be used to modify the properties of thermoplastics (Table 14). If the fillers are harder than the matrix, the mechanical properties are usually improved. For example, glass fibers improve the properties of polyamides. Softer fillers lower mechanical strength, for example, POM + PTFE. For dry bearings, fillers serve above all to increase wear resistance (graphite, glass, metal), to decrease the coefficient of friction (PTFE, lead, MoS_2), or to increase the thermal conductivity (bronze).

Unlike those of most metallic bearing materials, the mechanical properties of thermoplastics change considerably between $-40\,°C$ and $+100\,°C$. Since the macromolecular structure of the plastics can, moreover, be destroyed gradually as a result of extended exposure to elevated temperatures, the permanent service life temperature of the plastics with low melting temperatures (listed in Table 13) is $\approx 100\,°C$. PTFE and polyimide can be used up to about $260\,°C$.

Usually steel shafts are used in combination with plastic plain bearings. Independent of the type of plastic, the most favorable sliding conditions are reached if the Rockwell hardness of the steel surface is more than 50 HRC. A major factor affecting the sliding behavior is the roughness of the mating material. In general, the smoother the mating metal the lower the frictional wear of the plastic. In the case of sensitive plastics, e.g., PTFE and HDPE, wear can be minimized by using a metallic mating surface of slight roughness, which, however, may increase the coefficient of sliding friction.

The limits of load carrying capacity are determined by the resistance to compression and the temperature. With plain bearings of PA 6 having wall thicknesses 3 – 4 mm, average specific loads of $14\,N/mm^2$ are permissible for very low sliding velocities at $20\,°C$. Fiberglass-filled PA 6 can withstand average specific bearing loads of $24\,N/mm^2$. For higher sliding velocities, the $(p \times v)$ value is generally the load limit, however, only as a guide, since the limit is defined by the sliding surface temperature. For PA 6 bearings of 1 mm wall thickness, the $(p \times v)$ value for dry running operation is $0.07\,N/mm^2 \times m/s$ for $v = 0.1 - 1.5\,m/s$. For thick-walled bearings, 3 mm, however, the $(p \times v)$ value is only $0.04\,N/mm^2 \times m/s$. Since thermoplastics tend to become viscous at slightly elevated temperatures – a reversible process – accurate bearing shapes can often be manufactured by injection or compression molding. This is especially economical for long-run production.

Table 12. Sintered materials for bearings [21]

Standard grade	Material	Nominal composition, %	Open porosity, %	Density, g/cm³	Radial crushing strength, N/mm² min.	Elongation, % min.	Brinell hardness, HB min.	Characteristics [22]; uses
SINT-B-00 P 012*	iron	C 0.3 Cu 0–1 rest Fe	20 ± 2.5	6.0–6.4	80	3	30	standard material, average requirements for load and noise
SINT-B-10 P 2012	iron-copper	C 0.3 Cu 1–5 rest Fe	20 ± 2.5	6.0–6.4	170	2	40	wearproof, shock loads possible; agitators
SINT-B-11	iron-copper-graphite	C 1.5 Cu 2 rest Fe	20 ± 2.5	5.8–6.2	250		70	low noise, "similar to sintered bronze"
SINT-B-20	iron-copper	C 0.3 Cu 20 rest Fe	20 ± 2.5	6.0–6.4	200	2	45	good noise conditions, (p × v) value similar to sintered bronze, static load carrying capacity lower than sintered iron
SINT-B-22	iron-copper	C 1.5 Cu 20 rest Fe	20 ± 2.5	6.0–6.4	270		25	lower load carrying capacity than sintered bronze, cheaper, poor emergency-run characteristics
SINT-B-50 P 4013	bronze	Sn 9–11 rest Cu	20 ± 2.5	6.8–7.0	100	4	30	standard material, low noise, good dynamic load carrying capacity and corrosion resistance
SINT-B-51 P 4022	bronze with graphite	Sn 9–11 C 0.2–2 rest Cu	20 ± 2.5	6.6–7.0	90	3	25	suitable for special requirements such as sliding behavior, load carrying capacity, and low noise
SINT-C-51 P 4023			15 ± 2.5	7.0–7.5	124	2.5		

* All "P" designations are from ISO 5755 (1980).

Thermosetting plastics based on epoxy or phenolic resin with solid lubricants such as graphite, MoS_2, or PTFE are primarily used as dry bearing materials. Then such bearings can be subjected to loads $\leqq 1$ N/mm² at sliding velocities of 1 m/s. They are manufactured by molding of powdery granulates followed by a heat treatment to increase strength. This hardening involves cross-linking.

Greater surface pressures with very low sliding velocities can be taken by phenolic resins reinforced by asbestos or cotton yarns. Bushes and bearings are manufactured from tubes produced by a winding process. Such bearings are especially suited for water lubrication (water lubricated roll-neck bearings, marine stern tubes). Synthetic resin bearings based on phenol – formaldehyde containing chip- or striplike fillers of paper or textiles are often used where exposure to dust and dirt has to be expected, for example, in agricultural machinery, construction machinery, cement mills, and roll mills.

3.4.2. Materials for Plastic Multilayer Thin-Walled Bearings

The metal – plastic composite structure combines the good properties of plastics with the good properties of metals. Metal – plastic composites are able to take higher mechanical loads than plastic bearings. The greater thermal conductivity of the metal component allows higher (p × v) values. These two advantages are most important for dry running and boundary lubrication. Plastic multilayer bearings are used for hydrodynamic lubrication where occasional mixed-friction operation has to be expected. The usual lubricants are employed.

Maintenance-free plastic – metal multilayer bearings of the type shown in Figure 7 allow

Table 13. Thermoplastics for sliding bearings [25] [a]

	Symbol (DIN 7728)	Density[b], g/cm³	Modulus of elasticity[c], N/mm²	Tensile strength[d], N/mm²	Coeff. of thermal expansion, $10^{-6}\,K^{-1}$	Thermal conductivity, $W\,m^{-1}\,K^{-1}$	Melting temperature, °C	Characteristics; uses
Polyamide 66	PA66	1.14	3000	85	85	0.29	250–255	universal plain bearing material; general mechanical engineering, fabrication by injection molding up to 10 mm thick
Polyamide 6	PA6	1.14	2700	80	85	0.29	215–220	
Cast polyamide 6	GPA6	1.15	3300	85	75	0.29	215–220	universal plain bearing material; general mechanical engineering, for thick-walled and very large bearings
Cast polyamide 66	GPA66							
Polyamide 11	PA11	1.04	1500	55	150	0.29	185	
Polyamide 12	PA12	1.01	1800	55	110	0.29	175–180	
Polyoxymethylene homo- and copolymeride (polyacetal)	POM	1.42	3000	70	120	0.29	165–175	plain bearing materials for precision mechanics, bearings with dimensional accuracy
Poly(ethylene terephthalate)	PETP	1.36	3000	70	80	0.23	255	
Poly(butylene terephthalate)	PBTP	1.29	2800	60	60	0.21	255	
High-density high molecular mass polyethylene	HDPE	0.95	950	22	190	0.36	135	linings, sliding skirts, joint prothesis
Polytetrafluoroethylene	PTFE	2.2	400–700	25–30	160	0.23	327[e]	bridge bearings
Polyimide		1.43	3400	100	51	0.30		turbine construction, high thermal loads, radiation resistant
Poly(vinylidene fluoride)	PVDF [26]	1.78	1800–2200	50–60	90	0.17	170	pump construction in chemical industry

[a] Properties of dry specimens. [b] DIN 53479. [c] DIN 53457, 4 point bend. [d] DIN 53455. [e] Transformation temperature.

(p×v) values of $1.8\,N/mm^2 \times m/s$ for continuous dry running. Experience shows that the sliding velocity should be kept $\leqq 2$ m/s for dry running and $\leqq 3$ m/s with lubrication. Table 15 lists metal – plastic composite materials.

Because of its chemical resistance and high service life temperature, PTFE is used most often. Plastics such as POM and PVDF are also used. Fillers may be incorporated.

To fabricate a composite material, for example, that illustrated in Figure 7, a 0.2 – 0.3 mm thick lead-bronze layer with a pore volume of ≈ 35 % is sintered onto a steel backing. The pores of this bronze are filled with a pasty blend of PTFE and lead in a rolling process that also applies a layer 0.01 – 0.03 mm thick. The impregnated strip is hardened at ≈ 320 °C.

3.5. Artificial Carbon and Graphite for Solid Sliding Bearings

Graphite has good sliding properties and is used as a solid lubricant in sintered materials (see Table 12) and thermoplastics (Table 14). In addition, artificial carbon and graphite are used for solid plain bearings, which are characterized by good dry-running behavior at temperatures up to 350 or 500 °C in oxidizing atmospheres. Both are resistant to chemicals in wet operations.

Artificial carbon bearings are used in drying units and kilns and in operations involving aggressive liquids or media that may affect the operational reliability of the bearing. Other important properties of these materials are a low coefficient of thermal expansion, ranging between 2.5×10^{-6} and $4.0 \times 10^{-6}\,K^{-1}$, and a thermal conductivity that can be compared to that of metals. In dry operation, the friction coeffi-

Table 14. Thermoplastics with fillers for sliding bearings [25]

	Density, g/cm³	Modulus of elasticity, N/mm²	Tensile strength, N/mm²	Coeff. of thermal expansion, $10^{-6}\,K^{-1}$	Thermal conductivity, $W\,m^{-1}\,K^{-1}$	Melting temperature, °C	Characteristics; uses
PA66 + glass fibers	1.35	9500	190	25	0.23	250–255	high loads along with short total sliding distance, suitable for water lubrication
PA6 + glass fibers	1.35	9000	160	25	0.2	215–220	
PA11 + glass fibers	1.26	3200	105	30	0.2	185	
PA12 + glass fibers	1.23	3500	65	30	0.21	175–180	
PA12 + graphite (40–50%)	1.35	5000	50	30		175–180	high thermal conductivity; semiconductors
PA66 + MoS_2	1.16	4200	90	65	0.25		
POM + glass fibers	1.56	9000	140	70	0.35	165–175	specific high load carrying capacity
POM + PTFE	1.52	2600	50	110	0.3	165–175	low stick-slip tendency
POM + PE	1.36	2400	45	140	0.3	165–175	
Polyimide + 15% graphite	1.51	4500	81	41	0.73	no melting or softening point	high thermal load carrying capacity; vacuum applications, spacecraft
+ 40% graphite	1.65	5500	68	29	1.43		
+ 15% graphite + 10% PTFE	1.55	3500	58	45	0.65		
PTFE + artificial carbon	2.08		>12	95	0.65	327*	especially suited for water lubrication; compressors, pumps
PTFE + graphite	2.14		>16	125	0.83	327*	foil bearings, chemical industry
PTFE + bronze	3.72		>15	95	0.75	327*	high specific loads; foil bearings, hydraulics

* Transformation temperature.

cient of graphite is 0.1 – 0.2 if the air is humid or other vapors surround the bearing (see Section 3.2). In a high vacuum, the coefficients of friction increase by a factor of 10. In normal dry operation, a dynamic load of 0.2 N/mm² can be permitted for bearings made from heat-hardened carbon at sliding velocities of 0.5 m/s, whereas 0.8 N/mm² must not be exceeded for bearings made from electrographite at the same velocity. In wet operation, higher sliding velocities are permissible, e.g., up to 4 m/s for a specific load of about 5 N/mm², this in the case of geared pumps.

Sliding bearings of artificial carbon require shafts having a high surface quality. A shaft roughness of $R_z = 0.5 - 1.0\,\mu m$ is desirable. Hardened shafts, for example, hardened chromium steels, are advantageous. A porosity of artificial carbon between 10 and 15 % enables additional impregnation, both with metal and synthetic resins. Impregnation increases the load carrying capacity for a dry bearing up to ≈ 3.7 N/mm² if the electrographite is impregnated with a special salt. However, the temperature limit is lower. For example, artificial carbon impregnated with lead or synthetic resin can only be employed ≦ 180 °C.

Graphite and artificial carbon are brittle. For high-temperature use, artificial carbon bearings shrunk onto a steel ring give the best press fit in the housing bore. Presently there is no standard available for materials, and therefore manufacturers' information has to be used. Design features are standardized in DIN 1850.

Artificial carbon bearings are fabricated from compacted blanks and annealed. Heat-hardened carbon is obtained at 1200 – 1300 °C. Firing temperatures of 2800 – 3000 °C produce electrographite quality especially suitable for plain bearings. The burnt blanks are finished by cutting.

3.6. Ceramic Materials for Solid Sliding Bearings

Of the ceramic materials, only *alumina* is widely used for speciality solid plain bearings. Like graphite and artificial carbon, alumina is resistant to aggressive media and high temperatures.

Table 15. Thermoplastic composite bearing materials [27]

Material	Additions	Operating temperature, °C	Thermal conductivity, $W m^{-1} K^{-1}$	Characteristics
Steel backing with porous sintered tin or lead–tin bronze layer, filling and finish of polytetrafluoroethylene and additives	Pb, MoS_2, graphite, in special cases no additives	− 200 to + 280, continuous operation ≤ 260	40	dry running, boundary lubrication, hydrodynamic lubrication
Bronze backing with porous sintered tin or lead–tin bronze layer, filling and finish of polytetrafluoroethylene and additives	Pb, MoS_2, graphite, in special cases no additives		50	
Steel backing with porous sintered tin or lead–tin bronze, coated with thermosetting plastics, e.g., POM, PVDF	PTFE, Pb	− 40 to + 130, continuous operation ≤ 100	2–10	initial lubrication, boundary lubrication, hydrodynamic lubrication
Easily shapable re-sintered tin bronze fabric, embedded in polytetrafluoroethylene sheet with additives	graphite, glass fiber powdered tin bronze, MoS_2, in special cases no additives	− 200 to + 260, higher for short intervals	0.8	dry running, initial lubrication, hydrodynamic lubrication
Easily shapable fabric of stainless steel, embedded in polytetrafluoroethylene sheet with additives			0.8	
Steel backing with glued on polytetrafluoroethylene mixed fabric, filling of phenolic resin	polyester fabric	− 200 to + 180, higher for short intervals	0.8 without backing	dry running, initial lubrication, hydrodynamic lubrication
Backing of steel, stainless steel, or nonferrous metals with glued-on polytetrafluoroethylene mixed fabric, filling of phenolic resin	fabric of glass fiber, bronze, cotton	− 200 to + 180, higher for short intervals	0.8 without backing	dry running, initial lubrication, hydrodynamic lubrication

It is very hard and is resistant to abrasive wear. However, it is brittle and has a tendency to break. Its very high wear resistance is useful for circulation pumps for hot-water heating systems. Solid particles carried in the water circuit are not embedded in the bearing, but crushed between the aluminum oxide bearing and shaft.

Wear resistance and insensitivity make this material useful in supply pumps for polluted or aggressive liquids. Alumina is useful for valves.

Aluminum oxide plain bearings are fabricated from molded parts and sintered. The blanks are cut with diamond tools.

3.7. Soft Rubber for Sliding Bearings

Important criteria for the use of soft rubber as a bearing material are its low coefficient of friction and its high elastic deformability, shock absorption capacity, and wear resistance. Soft rubber, having a Shore Hardness A between 33 and 81, is used in water-lubricated plain bearings, especially for water pumps, since this material is completely resistant to water, even seawater, and has a high wear resistance. In addition, soft rubber is resistant to almost all nonoxidizing acids and alkaline solutions below 65 °C. On the other hand, it is very sensitive to oil and petroleum products, which cause the rubber to swell and lose its elasticity and strength. However, some special grades, such as butyl rubber, are resistant.

Unlike other plain bearings water-lubricated soft-rubber radial plain bearings have an elastic gap size as a consequence of the great deformability of rubber. The elastic deformability partially compensates for the low viscosity of the lubricant water [28]. As a result less precision is required to fabricate rubber bearings than for other water-lubricated bearings, such as bronze or artificial carbon.

Generally soft rubber plain bearings with backings are usually made from metals, e.g., CuSn5Zn5Pb5, but can be made from thermosetting plastics where the rubber layer is vulcanized into the backing.

Soft rubber is very sensitive to temperature fluctuations because of its low thermal conductivity, and the highest operating temperature ranges from 70 to 100 °C depending on the type of rubber. Lubrication and cooling has to be sufficient. The rubber sliding layer often forms a polygon with longitudinal grooves. In the case of larger bearings, especially with marine shaft bearings, individual rubber-lined metal rods with interspaces are mounted around the periphery of the backing brass. The hard particles in the water are temporarily embedded in the elastic rubber before they pass into the next groove and are flushed away. However, the grooves increase the coefficient of friction and lead to mixed lubrication operation. Despite their excellent sliding properties, smooth circular cylindrical rubber bearings without grooves are not used. If the bearing runs dry, the rubber is soon destroyed as a result of the quickly rising temperature [29].

Soft rubber can be used with water lubrication at $v = 2$ m/s and $p = 1$ N/mm^2. Rubber containing ≈ 50 % graphite had an increased limit load [30].

3.8. Practical Selection of Materials

In addition to knowing the tribological fundamentals, the following factors are important in selecting the bearing material best suited for a specific application:

1) load, specified by kind, intensity, and direction
2) sliding speed, specified by range and kind – either oscillating, rotating, or linear
3) space available for installation and bearing dimensions
4) temperature
5) environmental conditions, such as wetness, humidity, corrosive media, dirt, dust, and external vibration
6) lubrication (or dry), specified by type, rate, and supply

Other aspects are reliability in operation, exchangeability in case of damage, and the availability of specific bearing materials and designs. Last, but not least, is the cost, not only production costs but also downtime costs if the bearing has to be replaced.

The wide variety of proven bearing materials on the market is the result of the variety of situations in which plain bearings serve. The practical know-how of the bearing manufacturers can be useful in selecting a specific material.

4. References

General References

1. E. Schmid, R. Weber: *Gleitlager,* Springer Verlag, Berlin 1952.
2. *Metals Handbook,* 8th ed., **vol. 1: "Sleeve Bearing Materials,"** American Society for Metals, Noveity 1961, p. 843.
3. F. P. Bowden, D. Tabor: *The Friction and Lubrication of Solids,* part II, Clarendon Press, Oxford 1964.
4. G. Vogelpohl: *Betriebssichere Gleitlager,* Springer Verlag, Berlin 1967.
5. G. C. Pratt: "Materials for Plain Bearings," *Int. Metall. Rev.* **18** (1973) 62.
6. M. J. Neale (ed.): *Tribology Handbook,* Newness-Butterworth, London 1975.
7. H. Czichos: *Tribology Series 1,* Elsevier, Amsterdam – Oxford – New York 1978.
8. K. H. Habig: *Verschleiß und Härte von Werkstoffen,* Hanser Verlag, München – Wien 1980.
9. W. J. Bartz et al.: *Gleitlagertechnik,* part 1, Expert, Grafenau 1981.
10. *SAE Handbook 1982,* Society of Automotive Engineers, Warrendale, Pa., 1982.

Specific References

11. H. Schlicht, *Wälzlagertechnik* **9** (1970) 2.
12. E. Roemer, *VDI Ber.* **410** (1981) 133.
13. K. Anderko, R. Weber, E. Grau, *VDI Ber.* **156** (1970).
14. R. Weber, *Metall (Berlin)* **29** (1975) 447.
15. W. A. Glaeser, *J. Met.* **35** (1983) 50.
16. C. M. Pooley, D. Tabor, *Proc. R. Soc. London Ser. A* **329** (1972) 251.
17. H. U. Mittmann, H. Czichos, *Materialprüfung* **17** (1975) 366.
18. E. Roemer, *Metall (Berlin)* **16** (1962) 766.
19. W. Hilgers, *VDI Ber.* **248** (1979) 149.
20. R. S. Montgomery, *Wear* **87** (1983) 339.

21. Werkstoff-Leistungsblätter des Fachverbandes Pulvermetallurgie, Schwelm 1974.
22. F. Rübenach, *Powder Metall. Int.* **15** (1983) 20.
23. I. M. Fedorchenko, *Powder Metall. Int.* **15** (1983) 126.
24. A. Cheng et al., *Powder Metall. Int.* **15** (1983) 178.
25. VDI-Richtlinien 2541, VDI-Verlag, Düsseldorf 1975.
26. P. Gebauer, *Kunststoffe* **65** (1975) 3.
27. VDI-Richtlinien 2543, VDI-Verlag, Düsseldorf 1977. KS-Catalogue Permaglide, Kolbenschmidt AG, Neckarsulm.
28. H. Piper, *Konstruktion* **22** (1970) 179.
29. A. Wilda, *Pumpen und Verdichter Informationen* 1980, 11.
30. L. Pietrasik, K. Ziemianski, *Schmierungstechnik* **10** (1979) 142.

Beer

Hans Michael Esslinger, Freiberger Brauhaus Aktiengesellschaft, Freiberg/Sachsen, Federal Republic of Germany

Ludwig Narziss, Freising, Federal Republic of Germany

1. Introduction . . . 1
2. Raw Materials . . . 3
2.1. Starch-Containing Raw Materials . 3
2.1.1. Malting Barley . . . 3
2.1.2. Malting Wheat . . . 4
2.1.3. Unmalted Grains . . . 4
2.1.4. Other Sources of Extract . . . 5
2.2. Hops and Hop Products . . . 6
2.2.1. Resins . . . 6
2.2.2. Hop Oil . . . 7
2.2.3. Polyphenols . . . 7
2.2.4. Processing of Hops and Hop Products 7
2.3. Brewing Water . . . 9
2.3.1. Salts . . . 9
2.3.2. Brewing Water Treatment . . . 10
2.4. Beer Yeasts . . . 12
2.5. Auxiliary Materials and Brewing Aids . . . 12
3. Production Technology . . . 14
3.1. Malting . . . 14
3.1.1. Steeping . . . 15
3.1.2. Germination . . . 16
3.1.3. Kilning . . . 18
3.2. Technology of Wort Production . . . 20
3.2.1. Grinding of the Malt . . . 21
3.2.2. Mashing . . . 22
3.2.3. Separation of Wort . . . 23
3.2.4. Wort Boiling and Hopping . . . 24
3.2.5. Wort Treatment . . . 26
3.3. Bottom Fermentation . . . 27
3.3.1. Fermentation . . . 28
3.3.2. Maturation . . . 29
3.3.3. Cold Storage . . . 31
3.3.4. Filtration . . . 31
3.3.5. Stabilization . . . 31
3.3.6. Types of Bottom-Fermented Beers . . 32
3.4. Top Fermentation . . . 32
3.5. Special Production Methods . . . 34
3.5.1. Dietetic Beer . . . 34
3.5.2. Nutrient Beer . . . 34
3.5.3. Low-Alcohol Beer and Alcohol-Free Beer . . . 34
3.5.4. High-Gravity Brewing . . . 34
3.6. Filling . . . 35
3.7. Beer Dispensing . . . 36
4. Properties and Quality . . . 36
5. Analysis . . . 37
5.1. Analysis of Raw Materials . . . 37
5.1.1. Water . . . 37
5.1.2. Malt . . . 37
5.1.3. Hops and Hop Products . . . 37
5.2. Brewhouse Control . . . 37
5.3. Wort . . . 37
5.4. Fermentation . . . 38
5.5. Microbiological Process Monitoring 38
5.6. Beer . . . 38
5.7. Legally Required Controls . . . 39
6. Economic Importance . . . 40
7. Physiology and Toxicology . . . 41
8. References . . . 42

1. Introduction

Beer is described as a beverage containing alcohol, extract, and carbon dioxide. Beer is prepared from barley malt, raw hops or other hop products, brewing water, and top- or bottom-fermenting yeast. The alcohol must be produced exclusively from these ingredients, which are converted to fermentable products during the brewing process. Barley malt may be combined with wheat malt, unmalted cereal adjuncts (raw grain), and other extract-containing materials. Legal regulations concerning raw materials and additives as well as rules for listing these vary from one country to another.

To obtain barley malt, the grain is germinated under controlled conditions; the controls comprise moisture content, germination temperature, the ratio of oxygen to carbon dioxide in the germinating grain, and germinating time. *Green malt* is formed once a certain increase in enzyme activity and a partial degradation of the starchy endosperm have taken place; this latter process is called by the brewer *modification*. Green malt is then processed into *kiln* or *brew-*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_421

ing malt by drying. This malt is subsequently milled (ground) and mixed with brewing water. During the subsequent procedure, called *mashing*, high molecular mass components of malt are degraded by enzymes at specific well-defined temperatures (*rest periods*). The suspension is filtered to separate the liquid, called *wort*, from the spent grains. This process is called *lautering*. The wort is subsequently boiled with the addition of hops; this causes a coagulation of constituents, which are then called *hot trub* or *break*. They are separated together with the solids from the hops (*hot trub removal*). The clarified wort is subsequently cooled to the *pitching temperature* required by the fermentation method and the yeast strain used. The fermentation process is initiated by *pitching*: this consists of saturating the cold wort with air and adding cultured yeast. The fermentable low molecular mass components of the wort are converted to ethanol and numerous aroma compounds (fermentation byproducts) according to the metabolism of the yeast strain. After maturing to taste and enriching with carbon dioxide produced by fermentation, the beer is filtered to clearness and bottled.

The *extract of original wort* (original extract; original gravity) is defined as the mass fraction of nonvolatile, dissolved extract substances in the unfermented, cold pitching wort. During fermentation most of these substances are metabolized by the yeast, and the content of extract continually decreases. The extract of original wort can, however, be determined in samples taken during fermentation and in the finished beer, provided that the alcohol concentration and the real extract (extract contained in the dealcoholized beer) are both known.

The *degree of attenuation* is defined as the ratio of fermented extract to original extract, expressed as a percentage. The *attenuation limit* indicates the maximum amount of extract that the yeast will ferment.

Furthermore a distinction has to be made between top- and bottom-fermenting types of beer. The *top-fermenting yeast* rises to the surface at the end of the main fermentation process, whereas the *bottom-fermenting yeast* settles at the bottom. Both yeast types differ morphologically, in their enzyme composition, and in their physiological behavior.

The nomenclature of beer types is determined by their general production method (see Chap. 3). Typical beer designations have developed from the natural chemical composition of the water in various regions, e.g., Pilsener, Munich, Dortmunder.

In the Federal Republic of Germany, teaching and research in the area of brewing is concentrated at the technical universities of Berlin and of Munich-Weihenstephan. The universities of other European countries and of Japan also have formidable research capacities. Furthermore, in numerous countries there are research facilities operated jointly with the brewing industry. In addition, there are the laboratories of the large brewing companies, which have contributed greatly to the level of knowledge currently available. Results of research are discussed at scientific meetings in Europe (European Brewery Convention, EBC) and at international meetings in the United States, South America, Australia, and Japan.

History. The word beer comes from the Latin word "bibere" (to drink), which is the origin of the Old English word "bēor" (the brewed), akin to the Old High German word "bior", from which also the French word "bière", the Italian "birra", the East European "pivo", and the Spanish "cerveza" developed.

The roots of beer production, however, go back much farther to the first agrarian societies, the Sumerers. They used a variety of grain called emmer (*Triticum dicoccum*) which was dehusked and baked to give flat bread. The flat breads were soaked in water and then allowed to ferment spontaneously through the action of wild yeasts. The Babylonians developed the art of brewing further and distinguished between twenty different types of beer, in which the emmer and barley content as well as the strength of the beer were closely regulated. The "Codex Hammurabi" contained regulations regarding the quality of beer and described strict punishment for beer adulterators.

The Egyptians refined further the art of beer brewing and the legal requirements. They made the grain germinate and eliminated the soaked pieces of bread by sieving.

The Jews, Greeks, Romans, and Germanic peoples all knew beer, but partly preferred wine as a drink. From the seventh century, malting and brewing processes were researched with great experimental zeal, mainly in German monas-

teries. In the following centuries, brewing and dipensing rights were loaned out to several monasteries. It was also the monks who first used hops as a flavoring agent. From the fourteenth century, hops were generally accepted as an additive; earlier, tree bark, bitter herbs, and berries were added to the brew.

Until the sixteenth century, only the spontaneous top fermentation, which occurs at higher temperature, was known; later, bottom fermentation also was discovered.

War changed drinking habits because of the destruction of cultural values. Bavaria, for instance, became a beer country only after the Thirty Years' War, when its vineyards were destroyed.

Beer as it is known today was only made possible through numerous inventions in technology. Hot air kilns, steam engines, refrigerating machines, and filtration equipment enabled brewers to work throughout the year. Increasingly, more precise control of the brewing process also became available. The most significant improvement in quality was finally accomplished after the invention of the microscope: yeasts were found to cause the alcoholic fermentation. At the end of the nineteenth century, yeasts were cultivated and introduced into the brewery as pure strains.

Purity Law. Laws that regulate the production of beer have always been regarded as consumer protecting regulations. One of these statutory regulations was laid down by the Bavarian dukes Wilhelm IV and Ludwig X at the State Parliament at Ingolstadt on April 23, 1516; the law was accepted and is known today as the *Purity Law* (Reinheitsgebot). It requires that only barley (barley malt), hops, and water are to be used for the production of beer; yeast, as the fourth raw material, was only mentioned for the first time in 1551. Germany, Greece, and Switzerland brew according to this very strict law, the oldest in the world that pertains to food processing.

Other sources of extract and numerous additives and brewing aids are permitted in countries elsewhere. Because brewing technology does not require additives and chemicals for the production of beer of a consistently high quality, the Purity Law brewers insist on the maintenance of the Purity Law.

2. Raw Materials

2.1. Starch-Containing Raw Materials

For further information, see also the articles on → Cereals and Cereals Products, and on → Starch.

2.1.1. Malting Barley

Barley belongs to the family of grasses (*Gramineae*) and is found in various forms. First, a distinction is made between summer barley (spring sowing) and winter barley (late fall sowing), and second, between two-row and multiple-row barleys, according to the number of blossoms on the stalk. Multiple-row barleys produce malts richer in husks, protein, and enzymes, which may prove advantageous when using unmalted grain adjuncts. Two-row barley is divided into two main groups: the straight barley, and the "nodding" barley, whose ear hangs down during maturation. Two-row spring barley (*Hordeum distichum nutans*) has the best malting and brewing properties; the most important varieties of spring barley in Europe are Barke (Germany, Denmark), Chariot (UK), Optic (France, Denmark, UK), Pasadena (Germany, Denmark), Reggae (Netherlands), and Scarlett (France, Germany, Denmark, Netherlands). Two-row winter barley varieties are Plaisant (France), Pipkin (UK), Tiffany (Germany), and Vanessa (Germany). The advantages of barley over other grains that could be used for malting are the easier regulation and control of the germination process, the superior taste of beer made with barley, and the brewing technology available. Barley grows best in a humid climate and on soils with moderate contents of nutritients. The husks produce a filter bed for lautering, and the enzyme complement is advantageous for bringing about the desired modification. A similar uniform development of kernels as in the two-row spring barley can only be achieved in the newly-bred two-row winter barley varieties. Mixtures of spring and winter barleys cause technological difficulties in the malting process. Because the developmental rhythms of the two types of barley are different, the germination process occurs unevenly when processing mixed charges. These problems may be

overcome by using separate malting processes, but processing difficulties can be expected in the brewery in connection with lautering, fermentation, and the filterability of the beer; an impaired beer stability can also be expected.

Quality Requirements. Minimum specifications are laid down in the EC quality standard. Moreover, the maltster puts further requirements on his raw material:

Germinative ability > 98 %
Screening: > 90 % > 2.5 mm (Vollgerste)
Raw protein content < 11.5 %
Extract > 80 %
Attenuation limit > 80 %

Smell, color, gloss, and husk fineness are checked by manual appraisement. Further mechanical, chemical, and physiological inspection methods are listed in [29], [30].

The most important malting attribute of barley lies in its *germinative vitality*, which is defined as the percentage of kernels that initiate germination in the very beginning of the malting process. Freshly harvested barley must pass through a post-maturation period. This dormancy may be shortened by physical or chemical methods.

The germinative energy is a measure of germination maturity: it indicates how many grains have actually germinated after three days under conditions that are similar to those used in practice. A further indicator of maltability is the sensitivity of the barley to an overexposure to water during steeping (water sensitivity); the importance of this criterion has declined following the application of extended air rest periods during steeping.

The standardization of pilot-scale malting methods has become an indispensable aid for the assessment of barley and optimization of malting technology parameters; furthermore these methods provide reliable criteria in the selection of newly-bred barley varieties for planting.

Directions for the cultivation, storage, and physiological preservation of brewing barley, as well as detailed descriptions of its morphology and breeding, may be found in the literature [8], [10], [12], [13].

Chemical Composition. The chemical composition of brewing barley versus a pale malt is shown in Table 1. The moisture content, which is especially relevant to the storage quality of freshly harvested barley, may range from 12 to 20 %; 10 to 12 % are optimal for storage. As shown in Table 2, the moisture content is the limiting factor for storage. The α-glucans (amylose and amylopectin of the starch) are the most important carbohydrates in barley. Other carbohydrate components include β-glucans (cellulose, hemicellulose, gums), pentosans, as well as minute portions of low molecular mass sugars. Proteins are especially important for maltability, yeast nutrition, foam, taste, and the stability of the beer. Lipids are only partially used up during malting, the remainder staying mainly in the spent grains. Other important components are phosphates (about 0.3 %), minerals (2.5 – 3.5 %), vitamins (about 0.5×10^{-3}%), and phenolic substances (about 0.2 %). The enzymes of barley and of malt, as well as their effects during mashing, are shown in Table 4.

2.1.2. Malting Wheat

For some top-fermenting beers (see Section 3.4) wheat malt is added in order to achieve a special flavor quality. Brewing wheat should contain 11.5 – 12.5 % protein. Its extract yield is in the range of 83 to 87 %.

2.1.3. Unmalted Grains

Economic reasons or insufficient supplies of brewing barley or brewing malt have resulted in obtaining part of the starch by the addition of other, unmalted grain types. These *adjuncts* may account for up to 30 % of the grist in Europe, and up to 50 % in the United States. However, the success of this practice requires the use of very high enzyme-containing and protein-rich barley malts. In many countries these adjuncts are defined as replacement material for brewing malt, which add mostly carbohydrates to the wort; their use must be legally permitted.

Unmalted barley does not present any economic advantage over barley malt. Lower yields are achieved and frequently insufficient conversion in the brewhouse has to be corrected by adding enzymes produced microbially. Beers produced in this manner contain less nitrogen,

Table 1. The chemical composition of brewing barley and malt (mass fractions in %)

	Malting barley		Malt	
	As is	Dry matter	As is	Dry matter
Moisture	12 – 14		4 – 5	
Starch	55 – 57	64 – 66	56 – 58	58 – 60
Other non-nitrogen extract compounds	12 – 14	14 – 16	16 – 18	17 – 19
Protein	9 – 10.5	10 – 12	8.5 – 10	9 – 11
Fiber	4	4.5	5	5.2
Minerals	2.5	2.8	2.4	2.5
Fat	2	2.3	2	2.1

Table 2. Maximum storage time for maintaining malting barley quality under different storage conditions

Storage temperature	Seed moisture content			
	10 %	12 %	14 %	16 %
8 °C	7.5 years	2.5 years	1 year	170 days
10 °C	6 years	2 years	300 days	140 days
12 °C	5 years	1.6 years	240 days	110 days
14 °C	3.8 years	1.3 years	190 days	85 days
16 °C	3 years	1 year	150 days	65 days

have a lower attenuation limit, and display better head retention, but filtration is more difficult.

Unmalted wheat. Partially unmalted wheat is sometimes added to the mash. Its composition is similar to that of barley. With a moisture content of 15 %, wheat contains 65 % starch and other carbohydrates, 12 – 14 % protein, and 1.7 % fat.

Rice is processed as broken rice. This by-product of table rice production must be pure white. The moisture content is 12 – 13 %; other concentrations (expressed on a dry basis) are: extract 93 – 95 %, fat 0.5 – 0.7 %, protein 8 – 9 %. Rice starch gelatinizes at 65 – 70 °C, sometimes only at about 80 °C. The co-processing of rice generally results in very light and dry beers.

Corn (maize) is processed as corn starch, flakes, or larger corn grits, and is popular because extract yields range from 87 to 91 % after the removal of the oil-rich embryo. Moisture content should not exceed 12 – 13 %; protein content (dry basis) is in the range of 8.5 – 9 %; a residual fat content of less than 1 % in the grits is not harmful to head retention. Addition of corn results in sweetish, full-bodied beers. Commercial corn starch is practically free from protein and fat. The final yield is about 103 %, because of the addition of water during enzymatic hydrolysis

Sorghum, as a malt additive, has only regional importance in Africa.

2.1.4. Other Sources of Extract

Other sources of extract are processed *starch preparations* and *carbohydrates* in fermentable low molecular mass form. Besides starch flour itself, syrups frequently are used, which are manufactured from grain or starch flour by enzymatic or acidic hydrolysis. All syrups have an extract content of about 80 %, while their fermentability is in the range of 40 – 78 %. They are added to the wort kettle. The concentration of the wort may be increased in this manner to 15 – 18 % without affecting the further process development adversely (see Section 3.5.4).

Sugar is added to the wort kettle shortly before the end of boiling in order to raise the proportion of fermentable extract. Amounts up to 15 % yield very soft beers with a wine-like flavor. For malt beers and nutrient beers, sugar is added to the filtered beer in order to achieve the desired character and extract of original wort. Sugar is added as sucrose, as invert sugar, or as glucose. For nutrient beers caramelized brewing sugar may be added. The extract content of the sugar solution is between 65 and 85 %, depending on consistency and quality.

2.2. Hops and Hop Products

Hops added to the wort are an indispensable ingredient for many reasons [7], [9]. They impart a bitter taste to the beer, a specific aroma, and promote clarification. Furthermore, they are considered to have foam-improving qualities. Hops also are believed to act as an antiseptic in beer. The hop plant, *Humulus lupulus*, is a hardy, dioecious, climbing plant, and belongs to the hemp family. Only female vines are planted in hop farms; propagation is carried out by perpetuating vegetative clones. Hop cultivation requires special climatic conditions and a soil varying in texture from sandy to muddy. The hop cones are of primary interest to the brewer. They represent clusters of blossoms on the female plant and grow despite an absence of pollination. Hops are picked in August or September. Yellowish-green, sticky, cup-like glands (lupulin glands) are located on the inner sides of the inner and outer bracteoles; they contain the aromatic and bitter substances.

Hops are classified according to their origin and type; for similar varieties, the influence of origin dominates. The cultivated types have been obtained by separation according to shape. Classification by type in the EC is made on the basis of the content of compounds responsible for the bitter taste and flavor:

1) *EC List A aroma hops*: Hallertauer, Hersbrucker, Perle, Saazer, Spalter, Tettnanger
2) *EC List B bitter hops*: Brewers Gold, Magnum, Northern Brewer, Nugget, Target, Taurus

According to the certification policy within the EC hop market law, since 1978 producers of hops are required to designate varieties and to indicate origin. The most important hop-producing countries, together with their quantities harvested, are listed in Chapter 6. Hops can be judged on the basis of manual classification (appearance, color, pest infestation), whereas brewing values can only be determined by chemical analysis. Table 3 lists the most important components found in hops. The bitter acids (α- and β-acids) and the aroma substances (hopoils) are of great importance both from the viewpoint of brewing technology and for the taste of the beer.

Table 3. The chemical composition of hops (mass fraction in %)

	As is	Dry matter
Water	9 – 12	
α-Acids	2 – 17	2.2 – 13.5
β-Acids	2 – 10	2.2 – 11.2
Hop oils	0.5 – 2.5	0.6 – 2.8
Non-nitrogen extract compounds	4 – 9	4.5 – 10
Protein	15 – 21	13 – 22
Fiber	10 – 17	11 – 19
Polyphenols	3 – 8	4.5 – 16
Minerals	7 – 11	8 – 12
Lipids and waxes	up to 3	up to 3.4
Fatty acids	0.05 – 0.2	0.06 – 0.22

2.2.1. Resins

The bitter constituents of hops consist of α-acids (humulones), β-acids (lupulones), soft resins, and the oxidation products of bitter acids, called hard resins. The brewing value of the individual fractions varies and depends on their solubility in beer and wort and on their bitterness potential.

The α-acids are the most important of these bitter components because of their high bitterness potential.

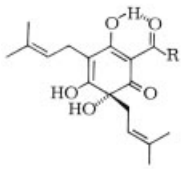

R = $CH_2CH(CH_3)_2$: Humulone [*23510-81-8*]

R = $CH(CH_3)_2$: Cohumulone [*511-25-1*]

R = $CH(CH_3)CH_2CH_3$: Adhumulone [*475-36-5*]

R = $CH_2CH_2CH(CH_3)_2$: Prehumulone [*59122-94-0*]

R = CH_2CH_3: Post humulone

The soft resins produce a lower bitterness level, about 10 – 33 % of the intensity produced by humulones. Hard resins are even less bitter. The β-acids are not bitter at all. The solubility of the α-acids is very dependent on the pH value. At the pH of the wort, these substances will dissolve to only a limited degree. In beer itself, α-acids are found only in very limited quantities, because they are isomerized during the boiling process whereas the remaining acids are removed during fermentation as part of the foam cover. The α-acids isomerize to iso-α-acids, e.g., *cis*- and *trans*-isohumulone. All α-acids and their isomerization products have approximately the same bitterness.

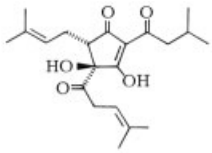

cis-Isohumulone

[1534-03-8]

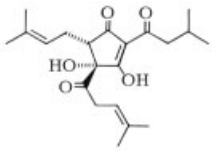

trans-Isohumulone

[467-72-1]

The isomerization products are significantly more soluble than the original α-acids, and are stable at the pH value of beer. During the isomerization process the following α-acid derivatives also are formed: *cis*- and *trans*-alloisohumulones, abeoisohumulones (oxidized isohumulones), antiisohumulones, spiroisohumulones, humulinic acids, and their various homologues; for details, see [2].

The lupulones (β-acids) are insoluble at the normal pH value of the wort. They do not isomerize during boiling, and are largely removed with the spent hops and the trub without being utilized. Lupulone differs from humulone by a side chain on the C_3 of the six-membered ring, and gives rise to the same homologues as humulone. During storage, lupulones oxidize to β-soft resins, which are soluble in the wort and in beer and also impart bitterness to the latter.

2.2.2. Hop Oil

The volatile and nonvolatile aromatic components of hops also determine hop quality. The aromatic content of hops is influenced not only by the variety but also very much by the drying and storing conditions and by processing.

The relative amounts of the almost insoluble sesquiterpenes, such as humulenes, farnesenes, and β-caryophyllenes, can be used to distinguish between the different varieties.

The mono- and sesquiterpenes become more soluble by oxidation, i.e., they are converted to epoxides; during fermentation they are partly transformed by the yeast to the corresponding sesquiterpene alcohols and can therefore contribute to the aroma of the beer.

Some volatile compounds (aldehydes, ketones, alcohols, and esters) are also produced during aging of the hops by side-chain breakdown of the bitter acids. Most of these, however, evaporate during wort boiling.

2.2.3. Polyphenols

Hops contain a number of phenolic components and derivatives of low molecular mass, such as phenolic acids (e.g., *p*-hydroxycoumaric, gallic, ferulic, protocatechinic, and caffeic acid), coloring components (catechins, flavones, and anthocyanidines), and polycondensated substances with strong tanning properties, which can originate from procyanidines. Hop polyphenols play a role in the establishment of the intrinsic colloidal stability of beer [21], [22]. The net result of boiling is a dramatic change in the already complex polyphenol composition of wort. Part of the complexity can undoubtedly be ascribed to the ready oxidation and the ease of polymerization of many polyphenols. Most of the hop polyphenols are removed during wort boiling by precipitation with proteins. The addition of hop polyphenols does not alter the foam consistency and color of beer, but very high dosages of hop polyphenols and long boiling times may adversely affect the colloidal stability. Low molecular mass polyphenols act as antioxidants and hence have a beneficial effect on beer flavor stability. The positive effects of low molecular mass polyphenols on human health are reported in Chapter 7.

2.2.4. Processing of Hops and Hop Products

After harvesting, the hop cones must be dried from a moisture content of 75 – 80 % to a level of 10 – 12 % using low temperature and strong air circulation. The hops are sulfurized before being packed into bales or compressed into ballots. Hops should be stored in cool, dry, dark areas at about 0 °C to preserve their quality. Vacuum packing followed by impregnation with an inert gas is necessary if prolonged storage of the hop product is anticipated. High temperature, oxygen, moisture, and light will cause changes in the oil fraction of hops; bitter acids become resins and lose part of their bitterness potential, whereas the polyphenols transform into higher condensed products [32–34].

Various hop products find acceptance nowadays in the brewing industry because of their improved storage capacity, longer preservation

of quality, and simpler handling in the brewery. Regulations control the use of these products in various countries. Hops products can be classified into four general types:

Non-Isomerized Hop Products. In this type of product the α-acids remain unchanged during hop processing. To achieve sufficient isomerization during wort production they must be added to the kettle at the start of or during wort boiling. Products included in this category are:

Double compressed whole hops

Hop pellets type 90. By drying hops down to a moisture content of 6 – 9 % and grinding, a powder is obtained. The powder is usually pelletized and packed under inert conditions. Because these pellets possess a larger surface area, their components are more readily soluble than those from raw hops. Therefore, the use of hop pellets type 90 offers savings of 10 – 15 % in terms of α-acids. Pellets also occupy a considerable smaller volume than raw hops.

Hop pellets type 45. After drying, hops are deep-frozen to about – 30 °C and ground. Lupulin is separated from the cones and from other nonessential parts of the leaves by sieving at the same low temperature. This treatment also reduces the polyphenol concentration in the resulting hops, depending on the extent of removal of nonessential components. Furthermore, the content of undesired components such as nitrate or pesticides is reduced. The powder, which usually is pelleted, offers savings in terms of α-acids of about 15 % when compared with natural hops.

Stabilized hop pellets are produced by blending magnesium oxide or calcium oxide with the hop powder prior to pelletization. During the formation of the pellet, the α-acids react with magnesium ions to form salts, which isomerize to *iso*-α-acids during storage.

Hop Extracts. Solvents used for the extraction of hops include ethanol and carbon dioxide (supercritical or liquid). For the production of the extract the hops are ground, the valuable components are dissolved, and the solution is separated from the solid substances. In all production methods, the solvent is finally evaporated at 40 – 60 °C and reclaimed. The extracts are packed in cans.

Carbon dioxide and ethanol dissolve the α-acids quantitatively. During extraction with 90 % ethanol, isomerization of 10 % of the α-acids takes place. Carbon dioxide dissolves only traces of the hard resins under the prevailing extraction conditions for the removal of bitter substances.

The aroma-contributing components are quantitatively dissolved; losses caused by the evaporation of the solvent are minimal in the case of carbon dioxide and substantial with ethanol. In the process employing supercritical carbon dioxide, the extraction conditions are controlled by pressure and temperature. The composition of the resulting extract of aroma-contributing substances and resins can be controlled by fractionation.

Carbon dioxide does not dissolve tannins and hop proteins. These may be obtained during a secondary extraction step by the use of hot water. During one-step procedures using ethanol as solvent, a certtain proportion of tannins also is extracted, which can easily be removed by centrifugation.

All extract can be standardized to a defined α-acid content with hot water extracts of hops or glucose syrup. The savings in bitter substances when using extracts are aprroximately 10 % in terms of the amount of the α-acid.

Isomerized Hop Extracts. In these products, α-acids have been isomerized to iso-α-acids during processing. Unlike α-acid based products, iso-α-acid products can be added either during wort boiling or after fermentation, depending on the composition of the product. Isomerized hop products include:

Isomerized Hop Pellets. The α-acids of stabilized hop pellets are converted easily and quantitatively into the corresponding iso-α-acids by heat treatment at temperatures below 60 °C.

Isomerized Kettle Extracts. There are two basic methods for the production of isomerized kettle extracts. In the first method pure resin extract is heated in contact with aqueous potassium carbonate-hydroxide solution. This gives an isomerized kettle extract in which the iso-α-acids are present as

potassium salts. In the second method the pure resin extract is mixed and heated with magnesium oxide. No water is added during the isomerization process. In the magnesium salt form the extract is very viscous and not easy to handle. By treatment with a strong acid, the magnesium iso-α-acid salts are converted to the free acids. This extract can be handled similarly to a conventional pure resin extract.

Isomerized hop extract is used for dosing into cold, fermented beer and must therefore be of high purity and essentially free of insoluble hop resins.

Reduced isomerized hop extracts comprise three different reduced forms of iso-α-acids: extracts containing rho-iso-α-acids (dihydroiso-α-acids), those containing tetrahydroiso-α-acids and those containing hexahydroiso-α-acids.

Hop oil products are specifically used to impart aroma to beer. They include the following preparations:

Hop pellets type 100 are produced by simple compression of whole hops. Waste materials are removed, but the hops are not dried or milled prior to compression. The pellets are vacuum packed in order to preserve the quality of the oil during storage. Whole hop pellets are generally added to cask conditioned ales in order to impart the beer 'dry hop' aroma.

Oil-Rich Hop Extract. An extract, that has a higher than normal oil content, can be obtained by partial extraction with carbon dioxide or fractionation during production of pure resin extract with supercritical carbon dioxide or by addition of pure hop oil to pure resin extract.

Pure Hop Oil. A common method for isolating pure hop oil is a steam distillation of an extract under vacuum or atmospheric pressure.

Hop oil emulsions

Fractionated hop oils

Miscellaneous Hop Products. Specialty hop products are available, which do not contribute significantly to beer bitterness, but are used instead to prevent overfoaming in the kettle and ensure normal fermentation characteristics. These products include:

Base hop extracts

Purified β-acid fractions

2.3. Brewing Water

Natural water always comes as a highly diluted mineral salt solution; its chemical composition and the concentrations of various constituents usually are determined by the geology of the specific region of origin. In addition, impurities, organic substances, and organisms may subsequently enter the water. Because of their low concentration, the salts are almost completely dissociated into ions, which indirectly influence the quality of the beer.

Today almost any water can be made suitable for brewing but at a corresponding cost. Expensive water treatment is unavoidable if certain ions are present in such high concentrations as to be detrimental to the beer, e.g., an excessive amount of sodium chloride or more than 30 mg/L of nitrate ions. Nitrate is reduced during fermentation to nitrite, which is toxic to yeasts.

2.3.1. Salts

The total solids content of natural water usually is 50 – 2000 mg/L, with an average of ca. 500 mg/L. The following ions are predominantly found in the water:

Cations: H^+, Na^+, K^+, NH_4^+, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Al^{3+}

Anions: OH^-, Cl^-, HCO_3^-, NO_3^-, NO_2^-, CO_3^{2-}, SO_4^{2-}, SiO_3^{2-}, PO_4^{3-}

Calcium and magnesium salts are most commonly found in natural waters. The sum of these two cations determines the total hardness of the water, which is expressed in milliequivalents per liter (mval/L). Carbonate, hydrogencarbonate (HCO_3^-), and free carbon dioxide have to be controlled carefully because of the corrosive nature of free carbon dioxide. The non-carbonate hardness is accounted for by salts of acids other than carbonic, e.g., sulfates. These salts remain dissolved during boiling [35].

Effects During Mashing. At the concentrations of mineral salts normally found in water, interactions with soluble components of the malt as well as various effects on the enzymes of the malt and on the ingredients derived from the hops are of practical importance, especially with regard to the net effect on the solubility of the α-acids. It is essential to observe how different ions present in the water, together with the malt ingredients, influence the pH values of the mash, the wort, and the beer. The phosphates of sodium, calcium, and magnesium are most effective in this regard. Further reactants include the potassium and calcium salts of such organic acids as lactic, malic, and succinic. The various hydrogencarbonates influence the acidity to varying extents, because the solubilities of the corresponding secondary or tertiary phosphates are different. Accordingly, the effect of magnesium hydrogencarbonate is greater than that of calcium hydrogencarbonate, because the secondary magnesium phosphate will remain in solution. The effect of sodium hydrogencarbonate is even more marked, because strongly alkaline tertiary sodium phosphate is formed, which remains in solution together with the secondary sodium phosphate. In addition, the precipitation of phosphates by calcium and magnesium ions results in a decreased buffering capacity of the mash, so that the pH may drop too rapidly during fermentation.

The alkalinity of the wort caused by hydrogen carbonates also affects the solubility behavior of the bitter substances present in hops. On the one hand, when the pH is increased, the solubility is increased, but an unpleasant, harsh bitterness is created in the beer. On the other hand, the ions of the alkaline-earth metals are able to compensate for the increase in pH caused by hydrogencarbonate. The following equation demonstrates this effect:

$$3\,Ca^{2+} + 4\,HPO_4^{2-} \rightarrow Ca_3(PO_4)_2 + 2\,H_2PO_4^-$$

Calcium ions convert the secondary phosphate to the acidic primary phosphate. Magnesium ions are only half as effective by comparison with calcium ions.

Alkalinity. The term alkalinity is used to designate the concentration of the hydrogencarbonate ions in water (→ Water, Chap. 6.1.). The "residual alkalinity" is calculated by subtracting the effect of the acidity-improving alkaline-earth ions from the effect of the acidity-destroying hydrogencarbonate ions:

$$\text{residual alkalinity} = \text{total alkalinity} - \left(\frac{\text{Ca hardness}}{3.5} + \frac{\text{Mg hardness}}{7}\right)$$

Hardness is given in degrees German hardness (dH). The total alkalinity is equal to the hardness linked to the presence of hydrogencarbonate ions.

The residual alkalinity indicates the suitability of the brewing water for the various varieties of beer. The residual alkalinity for pilsener should be below 35 ppm $CaCO_3$, and 90 – 110 ppm $CaCO_3$ for pale beer. Decreasing the residual alkalinity by 90 ppm $CaCO_3$ will lower the mash pH by 0.15 units. Should the residual alkalinity increase substantially, considerable disadvantages can be expected. The brewhouse yield is reduced by 2 % when the residual alkalinity is in the range of 180 – 210 ppm $CaCO_3$. Similarly, high values for residual alkalinity resulting in a mash pH of over 5.8 will reduce the effectiveness of most of the hydrolytic enzymes. As already mentioned, a high pH value causes a molecular solution of the α-acids, which will impair the bitter flavor of the beer. The release of polyphenols from the husks of barley malt will occur with greater ease in this pH range, and is acceptable only in the case of dark beer.

2.3.2. Brewing Water Treatment

Water used for brewing should correspond in quality to drinking water; it should be clear, colorless, and neutral in taste and smell. It should not contain heavy metals, especially iron and manganese, and should not be corrosive. Natural waters originate from many different geological formations and therefore very rarely conform to the quality required for brewing water. For this reason treatment is required before water from such sources can be used for brewing. The actual extent of water conditioning is governed by the concentrations of anions, the dissolved organic substances, and the presence of aggressive gases in the untreated water. Processing of brewing water also depends on the specific needs of

the brewer and on the particular beer required. The selection of a treatment method also is influenced by the prevailing laws.

Attention is directed to both the carbonate hardness and to the ratio of carbonate to non-carbonate hardness, which should be around 1 to 2 – 2.5. Bearing profitability in mind, the treatment methods mentioned in the following can be applied:

Heat Precipitation. Carbonates can be precipitated by boiling. However, this method is not economical.

Decarbonation. The simplest procedure for the precipitation of carbonates is to add milk of lime. The following components can be removed by $Ca(OH)_2$: free CO_2, $Ca(HCO_3)_2$, $Mg(HCO_3)_2$, and $MgCO_3$. For the latter, however, a pH of 10.5 – 11 is required. If the non-carbonate hardness in the raw water is higher than the magnesium hardness, quick or pressure decarbonation may be used. In this case, water and milk of lime are intensively mixed in a reaction vessel, where calcium carbonate precipitates as a coarse product that sediments well.

Variations in magnesium hardness are more problematic. If the magnesium hardness exceeds the non-carbonate hardness, slimy magnesium hydroxide will partly precipitate and clog the gravel filter bed; it can also contaminate the brewing water. To overcome this problem, a two-step *precipitation – decarbonation process* is used. In the first step, the calculated amount of milk of lime that is needed for the decarbonation and for the removal of magnesium is added to a partial stream of the raw water (two thirds). This is done in a superliming reactor, where both calcium carbonate and magnesium hydroxide are precipitated. This alkaline water subsequently reacts with the remaining one third of the raw water stream in a refining reactor. In this reactor, the excess lime is used to precipitate calcium carbonate whereas the magnesium carbonate of this 30 – 40 % stream remains in solution. The total water is finally collected in a storage tank after filtration through a gravel bed.

In order to control variability in the composition of raw water hydrogen ion exchange facilities can be installed at the end. This equipment will automatically be activated whenever the pH rises above a predetermined value. The procedure ensures extensive removal of the magnesium hardness. The brewing water thus obtained is very poor in carbonates, and does not contain an excess of free carbon dioxide after irrigation.

Ion Exchange (→ Ion Exchange). Brewing water can be readily deionized by using ion exchange resins, which require little space and facilitate a quicker throughput. The availability of food-grade ion exchange resins enables raw water to be processed to give water of any desired composition.

Cation exchangers can be either weakly or strongly acidic, depending on the anticipated load. Weakly acidic exchangers remove only the hydrogencarbonates (of Ca^{2+} and Mg^{2+}) from the raw water. During this process a large amount of free, corrosive CO_2 is liberated; this needs to be removed by blowing a countercurrent of air through the water. The remaining CO_2 reacts with milk of lime or marble to form $Ca(HCO_3)_2$; in this way slight hardening is achieved. Alternatively, the deionized water can be blended with raw water.

Strongly acidic ion exchangers operate in a similar manner. All cations are removed from the raw water, even those of the strong acids, and are replaced by hydrogen ions. These exchangers are especially effective with hardness caused by magnesium. The free mineral acids that are thus formed are neutralized with saturated milk of lime.

Anion exchange: SO_4^{2-}, Cl^-, and NO_3^- ions may be removed with anion exchangers; this will decrease non-carbonate hardness. Even though total deionization is not desirable in the production of brewing water, complete ion exchange enables the brewer to prepare brewing water from sources with widely differing compositions.

Electroosmosis. For total deionization by electroosmosis, investment and operating costs are much higher than in the treatment procedures described above. In addition, this procedure may be applied only if combined with other deionization methods. In a continuous-current field the salts in the water migrate towards the electrodes, which are separated by membranes. The water trapped between the membranes is appreciably lower in its salt content and can be run off.

Reverse Osmosis. Ions, molecules, and other small particles can also be removed from water by reverse osmosis. Additionally, table waters may be obtained from the resulting concentrate by proper sterilizing and carbonating procedures, provided that the ionic composition is appropriate.

Acid Addition. The addition of mineral acids, such as hydrochloric, sulfuric, or phosphoric acid, changes the carbonate hardness into non-carbonate hardness, but the total hardness remains unaltered.

Other Processes. Another important treatment may become necessary if bad-smelling or bad-tasting substances have to be removed. This can be achieved by installing an activated carbon filter upstream of the deionization equipment. Suspended matter is removed with flocculants and gravel bed filters, so as to avoid clogging of the downstream equipment.

Furthermore, such mineral salts as are found in natural water may be added to brewing water. Acceptable food-grade salts are: $CaCO_3$, $MgCO_3$, $CaCl_2$, $CaSO_4$, and NaCl.

2.4. Beer Yeasts

Brewery yeasts belong to the family of Saccharomycetaceae and the genus *Saccharomyces*. They have the main advantage that their cells propagate by budding. For the production of bottom-fermenting beers, *Saccharomyces carlsbergensis* are used, and for top-fermenting beers, *Saccharomyces cerevisiae* [36].

After the malt enzymes have been destroyed during wort boiling, the yeast provides the original wort with its enzyme system. Wort is not an ideal nutrient medium for yeast. In order to metabolize actively, the yeast must synthesize those substances it needs according to the wort composition. Depending on the syntheses required, the amounts of the metabolic products that are found in the fermenting substrate will differ. References to taxonomic and technological structure, morphological characteristics, chemical composition, propagation, metabolism, and yeast enzymes can be found under the appropriate keywords: → Ethanol, Chap. 5.1.; → Yeasts.

The pitching yeast should be selected as carefully as the other raw materials. Furthermore, its introduction into production, its crop after fermentation, and its further handling should be managed with great care. Short-term storage of yeast and aeration before pitching retains fermenting power and keeps the yeast in good physiological condition. If an infection with organisms detrimental to beer occurs, the yeast must be removed quickly from the production process

2.5. Auxiliary Materials and Brewing Aids

This group comprises all those chemicals that come into contact with the raw materials, the wort, and the beer; they are not essential for the production of beer, and are used for the correction of deficiencies. No general directions exist in the various countries where beer is brewed regarding the use of these materials.

Malting Auxiliaries. Various chemicals may be added to the steeping water. Hydrogen peroxide decreases the sensitivity of barley to water, and reduces the danger of mold formation. Alkaline additives, such as $Ca(OH)_2$ or NaOH, will increase the leaching out of husk tannins.

Gibberellic acid is an effective growth agent; it stimulates enzyme formation and accelerates germination. The amount of growth factor that is to be added to the final steeping water should not exceed 0.05 – 0.1 mg/kg of barley because of the danger of overmodification and additional coloration.

In order to avoid excessive losses of extract during germination, it is possible to add growth inhibitors during the later stage of germination. For this purpose, dilute nitric acid, formaldehyde, or, most commonly, potassium bromate are used.

Sprinkling the green malt towards the end of the germination phase with an aqueous glucose solution will increase the amount of extract to a greater extent than would be expected from the amount of sugar used.

The addition of sulfur dioxide to the air flow during kilning will result in a lighter color and a higher extract yield by lowering the pH value.

Other aids to water treatment are described in Section 2.3.

Brewhouse Auxiliaries. The first concern of the brewer is to eliminate differences in quality of the raw materials which have remained through insufficient modification in the malthouse and which could not be compensated for during mashing. These cytolytic, proteolytic, and amylolytic deficiencies, or the use of an excessively high proportion of adjuncts, will be compensated in the mash by adding enzyme preparations that originate from sources other than malt. The hard, raw aftertaste that frequently occurs in such beers can be reduced by adding mineral acid (HCl, H_2SO_4, H_3PO_4) or concentrated lactic acid to the mash.

During wort boiling, protein-precipitating substances such as formaldehyde, tannin, carrageenan (Irish moss), or protein-stabilizing substances may be added in order to improve the clarification process and stability of the end product. The addition of cross-linked polyvinylpyrrolidone (polyvinylpolypyrrolidone; PVPP) reduces the polyphenol concentration. Synthetic additives based on polyester result in a compact separation of the unpleasant, coarse, and bitter-tasting hot trub.

Numerous advantages can be gained by using a lactic acid culture propagated in the brewery (biologic acidification). The lactic acid bacteria obtained from malt ferment at 45 – 47 °C in 10 – 12 % unhopped first wort, and produce lactic acid (concentration 0.8 – 1.2 %) within 12 – 20 h. The culture is added in the brewhouse during mashing or wort boiling to obtain the desired pH value.

Fermentation and Maturation Aids. Deficiencies in yeast nutrients can be counteracted by the addition of yeast food, which contains amino acids, minerals, vitamins, and zinc salts, to the substrate in order to achieve a vigorous fermentation. Because all enzymes of the malt are denatured during wort boiling, one may wish to add enzymes with a wide activity spectrum, which may additionally enhance the clarification and filtration of the beer in the fermenting room. This will become necessary if the naturally-occurring enzymes achieve an insufficient hydrolysis of malt components with a high molecular mass.

If foam production is impaired by this enzymatic action, which is rather difficult to control, foam stabilizers (alginates, poly(propylene glycol), or polysiloxane) can be added.

After fermentation, bacterial 2-acetolactatedecarboxylase may be added to the fermentation substrate, which will transform diacetyl directly to acetoin (see Section 3.3.2). In this way the maturation rate can be increased considerably.

Antifoaming agents based on silicones have come into use in order to reduce the headspace needed during warm fermentation in tall, cylindrical vessels. A small amount (4 – 8 g/hL) of these agents is sufficient to prevent the formation of a fermentation cover. The foam inhibitor is removed completely by kieselguhr filtration, so that the final foam quality of the beer will not be impaired.

Clarification Aids. Before bottling, the beer is filtered through a filter cake. Proven materials for this purpose are kieselguhr (mud-free, calcined and screened diatomaceous earth) of various particle size distribution, or perlite (ground and calcined glassy rock of volcanic origin.

Activated carbon may be used to correct a mild off-taste; it is usually used in the treatment of rest beers. Shortly before filtration, hydro- and xero-silica gels may be added, which selectively affect the high molecular mass nitrogen fraction, and which also contribute to the build-up of the filter cake.

Stability Improvers. The stability of beer is defined as the ability to preserve its characteristics from bottling to consumption. To ensure an adequate biological stability in zones of moderate climate it is sufficient to clarify the beer through kieselguhr and sheet filters, provided that the beer is delivered to the consumer quickly.

Higher demands can only be met either by the use of insoluble substances which act mechanically or adsorptively, or by pasteurizing the bottled beer (with 12 – 14 pasteurization units), or by hot bottling at 68 – 75 °C. However, heat treatment results in early aging of the beer. Oxygen trapped during filling reacts quickly with the beer. As a result a nonbiological haze may form

sooner or later. This is derived either from tannins or from proteins. Tannins may be removed by adsorption onto polyvinylpolypyrrolidone, or by precipitation with formaldehyde. Haze-forming protein particles may be removed with silica preparations or bentonite. Proteolytic enzymes, such as papain, bromelain, or ficin, decompose high molecular mass proteins to non-hazing components.

During the breakdown of proteins, two characteristics, stability and foam, again run counter to each other. Metal ions accelerate haze formation; they can be complexed with ethylenediaminotetraacetic acid.

During long-term storage beer becomes vulnerable to other dangers including flavor changes, particularly through oxidation, different climatic and mechanical conditions, and effects due to light. The most effective remedy is low-oxygen bottling; if this is not possible, the beer needs to be protected against oxidation by the addition of such antioxidants as ascorbic acid, sulfites, sugar reductones, or by enzymes such as glucose oxidase or catalase. The influence of oxygen during bottling can also be overcome by evacuating the bottles, by pressurizing the bottles with fermentation carbon dioxide, and by allowing fobbing. For additional safety the cleaned beer barrels, kegs, or bottles can be treated with peracetic acid or other products based on hydrogen peroxide, iodine-releasing chemicals, or sulfur dioxide.

3. Production Technology

3.1. Malting

Malting is defined as allowing grain to germinate under well-controlled conditions. The main purposes of malting are the development of enzymes in the grain with simultaneous degradation of high molecular substances in the cell walls (modification), the achievement of a distinctive character by color and aroma compounds, and removal of undesired aroma compounds (i. e. *S*-methylmethionine and dimethyl sulfoxide). High extract yield and low malting loss are economic goals. The schematic procedure of malt production is shown in Figure 1.

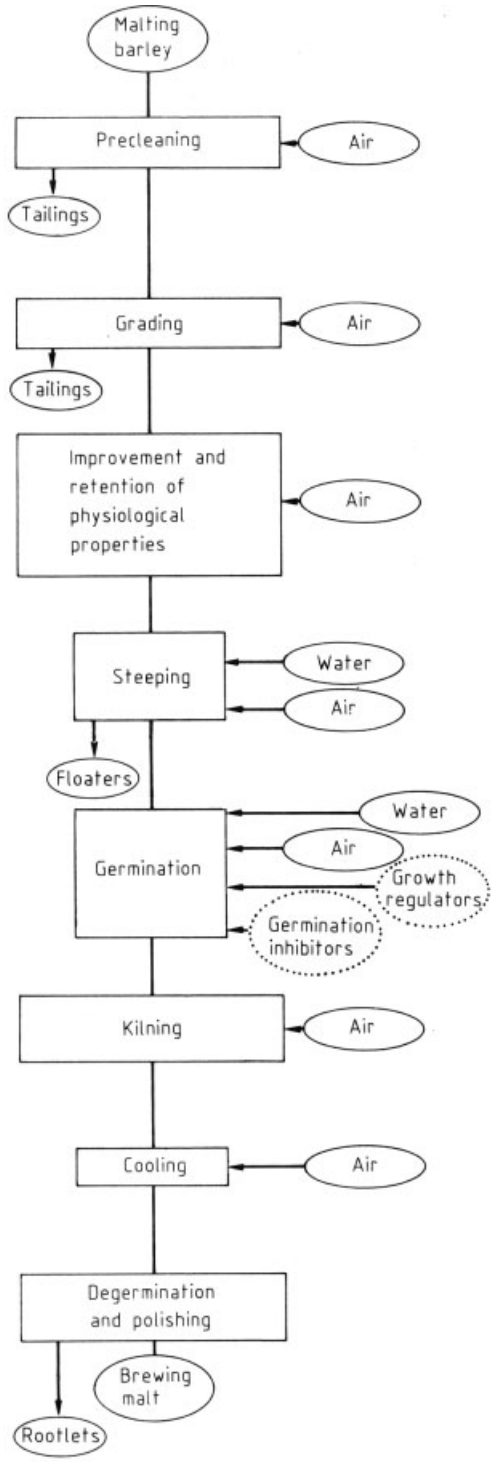

Figure 1. Flowsheet of malt production
Additives that are shown in dotted ellipses are not necessary

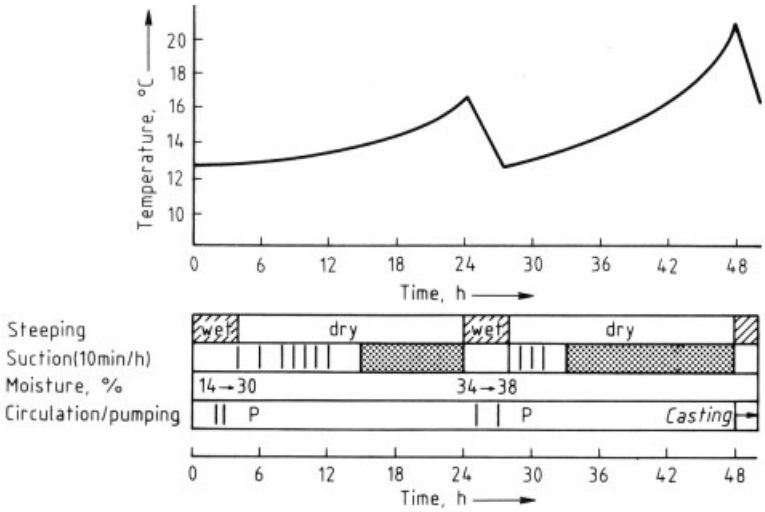

Figure 2. Steeping

After thorough cleaning with various separators and grading machines, the brewing barley is dried to a moisture content of ca. 12 %, which allows storage without damage to the embryo. If proper drying equipment is not available, cooling of the barley can help prevent its spoilage to a certain extent. Before malting is actually started periodic aeration is necessary to remove the carbon dioxide emanating from the grains, so as to keep the germinating facilities healthy and to maintain germinative ability.

3.1.1. Steeping

Upon addition of water germination of the grain begins. No particular requirements are placed on the steeping water during malting, but it should be of potable quality. In the beginning the water uptake is an osmotic process and depends on water quality and temperature. The grains begin to respire. Respiration is continued during the entire steeping period. At 30 % moisture content the water is taken up via the micropyle (physiological process) and living tissue in the embryo is formed.

Steeping Equipment. The round or rectangular steeping drums with conical bottoms (capacity 2.2 – 2.4 m^3 per ton of barley) must be equipped with a water inlet and outlet, discharge valves, overflow for floaters, a carbon dioxide exhaust vent, and an air inlet for pressure aeration and recirculation. Recirculation pumps and sprayers are optional. The first cleaning is often performed in washing drums.

Steeping Technology. Rapid water uptake and enhanced germination is possible in *pneumatic steeping*, in which wet steeping periods alternate with extended dry steeping periods (the latter during 50 – 80 % of the total period). First of all, steeping provides a definite moisture content appropriate to the physiological characteristics of the barley. During subsequent dry steeping (16 – 24 h) at a moisture content of 30 % the water sensitivity of the barley declines. After further increasing the moisture content to 38 %, a uniform germination of the kernels can be expected within a period of 14 – 20 h. The moisture content is then raised to the final level in the germination box by adequate spraying. An example of the steeping procedure and the necessary handling involved is shown in Figure 2.

When the temperature is raised from one wet steeping period to the next, the temperature of the steeping materials is taken into account. The casting of steeped barley at 18 °C permits the use of a germination procedure with decreasing temperature as shown in Figure 3. Other steeping methods exist based on similar principles: the flushing procedure, the resteeping procedure, and the water-saving spray steeping.

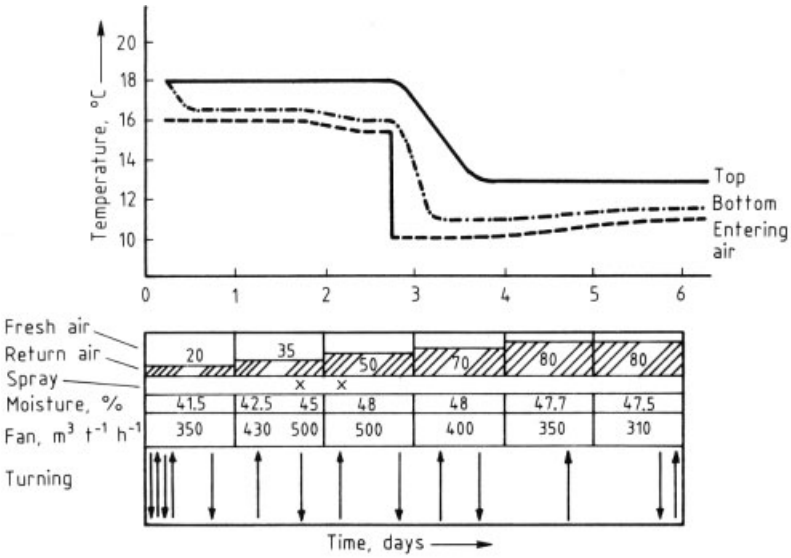

Figure 3. Germination

3.1.2. Germination

Germination is a physiological process during which the embryo develops rootlets and acrospire. During this process, the nutrients stored in the endosperm are partly consumed. The aim of controlled germination is to produce a green malt of a definite composition, but not to allow the development of a new plant.

Germination Conditions. Germination takes place only under appropriate conditions in order to achieve the desired metabolic changes during the time necessary for germination. Parameters are: moisture, temperature, ratio of air to carbon dioxide, and time. Steeping and germination conditions have to be adapted to barley variety, harvest year, water sensitivity, vitality, and furthermore to the protein content and the expected structure of the endosperm matrix. The moisture content must not decrease during the entire germination period. Temperatures favorable to uniform germination range from 14 to 18 °C.

A sufficient supply of air must ensure both normal respiration and removal of carbon dioxide. Germination will manifest itself first in noticeable changes in the appearance of the kernel. After "chitting" (breakthrough of the rootlet), the rootlet divides into a main root and secondary roots (forking). The acrospire also breaks through the aleurone layer and pericarp, and begins to grow under the husk towards the tip of the kernel. During malting the development of the germ buds is desirable only to a certain degree.

Besides these manifestations of growth, chemical transformations occur in the endosperm; stored materials are broken down by enzymes and changed into soluble matter. These are either used as an energy supply, or are built into new tissue in the germ buds. The growth factors which develop during germination cause the formation of a number of enzymes in the scutellum and the aleurone layer.

Enzyme Formation. In addition to certain modifications to the substances of the barley, the main purpose of malting is the induction and increase in hydrolytic enzymes. The most important groups of these are: hemicellulases, proteolytic enzymes, amylases, and phosphatases (see Table 4). Of the cytolytic enzymes, which break down the hemicelluloses to low molecular mass materials, the β-glucanases are considerably more effective than the pentosanases. The degree of cytolysis may be determined empirically by the continuous increase in friability of the endosperm; analytical measurements of the difference between coarse and fine grind extract and the viscosity of the congress wort can be made. Excellent cytolysis is achieved under con-

Table 4. Malt enzymes and their behavior during mashing

Enzyme	CAS registry number	E.C. number	Optimum conditions in mash		Inactivation temperature, °C
			pH	t, °C	
Oxidoreductases					
Peroxidase	[*9003-99-0*]	1.11.1.7		40 – 50	65
Lipoxygenase	[*9029-60-1*]	1.13.11.12	6.5	40	70
Polyphenoloxidase		1.14.18.1		60 – 65	80
Hydrolases					
Lipase	[*9001-62-1*]	3.1.1.3	6.8	35 – 40	60
Acid phosphatase	[*9001-77-8*]	3.1.3.2	4.5 – 5.0	50 – 53	70
α-Amylase	[*9000-90-2*]	3.2.1.1	5.6 – 5.8	70 – 75	80
β-Amylase	[*9000-91-3*]	3.2.1.2	5.4 – 5.6	60 – 65	70
endo-β (1 $\rightarrow$ 4)-Glucanase	[*9074-99-1*]	3.2.1.8	4.7 – 5.0	40 – 50	55
Cellulase	[*9012-54-8*]	3.2.1.4	4.5 – 5.0	20	20
Laminarinase	[*9025-37-0*]	3.2.1.6	5.0	37	50
Limit dextrinase	[*9025-70-1*]	3.2.1.11	5.1	55 – 60	65
Maltase	[*9001-42-7*]	3.2.1.20	6.0	35 – 40	40
β-Mannosidase	[*9025-43-8*]	3.2.1.25	3 – 6	55	70
Invertase	[*9001-57-4*]	3.2.1.26	5.5	50	55
exo- and *endo*-Xylanases		3.2.1.37	5.0	45	
endo-β-(1 $\rightarrow$ 3)-Glucanase	[*9044-93-3*]	3.2.1.39	4.7 – 5.0	40 – 45	55
exo-β-Glucanases		3.2.1	4.5	40	40
Pullulanase	[*9012-47-9*]	3.2.1.41	5.0 – 5.2	40	70
Arabinosidase	[*9067-74-7*]	3.2.1.55	4.6 – 4.7	40	60
β-Glucan-solubilase					
with esterase activity		3.2	6.6 – 7.0	62	73
with carboxypeptidase activity		3.4	4.6 – 4.9	62	73
Aminopeptidases		3.4.1	7.2	40 – 45	55
Carboxypeptidases		3.4.2	5.2	50 – 60	70
Dipeptidases		3.4.3	7.2 – 8.2	40 – 45	55
Endopeptidases		3.4.4	5.0 – 5.2	50 – 60	70

ditions of high moisture, average temperatures of up to 18 °C, plenty of oxygen, and long germination periods. Before the cell wall is broken down, proteins must be hydrolyzed to a certain extent by proteolytic enzymes. The proteolysis is favored by high moisture content, low temperature, and an optimum germination period; if germination periods are too long, low-molecular protein substances will be consumed for the synthesis of the rootlets and the acrospire. The modification of the proteins can be quantified by the degree of protein hydrolysis (ratio of soluble nitrogen to total nitrogen) and by the total amount of soluble nitrogen.

The protein content of barley decreases somewhat during germination, because the protein-rich rootlets are removed after kilning. The modification of starch by the action of α- and β-amylases occurs to only a moderate degree. The liberation, activation, and formation of the β-amylase, as well as the de-novo formation of α-amylase during the germination process, are both important. α-Amylase can be formed only in the presence of oxygen; its production is favored by a high moisture content during germination, low germination temperatures, and long germination periods.

Germination Technology Practice. The control of the batch is governed by the malting system, and is also determined by the steeping method and by casting. *Germination with increasing temperature* proceeds between 12 and 16 °C; for barley grown in hot, dry climates, the temperature may rise as high as 20 °C towards the end of germination. Slow but even growth, slow enzyme formation, and low enzyme activity characterize the germination process. Green malts are sometimes kept warm during the last 24 h of the germination period in order to attain the desired extent of cytolysis.

During *germination at constant temperature*, respiration and temperature increase are facilitated; this results in a more active metabolism and in strong growth. Maintaining the grain bed at 14 – 15 °C results in most even modification.

With modern steeping and germination equipment it has become common practice to *germinate with decreasing temperature* (Fig. 3).

The relatively warm initial germination phase at a still low moisture content favors quick hydrolysis and high enzyme activity. Enzyme formation is further stimulated by quick cooling of the grain bed and a simultaneous increase in moisture content. Because of the high moisture level, modification occurs satisfactorily in spite of the low temperature. The embryo tries to maintain its original growth rate, and compensates for the impaired living conditions by increased enzyme formation.

In the steep tank and during the early "biological" phase of germination, sufficient oxygenation of the grain bed is necessary for the formation of endo enzymes. In the subsequent "modification phase", a moderate increase of carbon dioxide in the grain bed (up to 4 %) will inhibit too vigorous growth and improve cytolysis. Germination usually takes six days. The different modification properties of the various types of malt can be compensated by varying the germination parameters, especially the moisture content.

Malting Equipment. Floor malting is the oldest and simplest malting method, but it is found only very rarely today. The metabolic activity of the steeped grain, which has been spread on the floor, is controlled by the temperature of the room, by the height of the grain bed, and by turning. The temperature is allowed to rise steadily from the "young pile" through the "growth" and "matting pile" up to the "old pile".

All methods that use aeration (pneumatic systems) are characterized by malting in deep beds. The most important and most difficult task in this type of malting is the constant maintenance of effective cooling of the grain bed by means of an air stream saturated with moisture. Besides supplying oxygen and removing carbon dioxide resulting from respiration, the air stream must also inhibit excessive loss of water during germination. This is not easy to accomplish because the air in the grain bed warms up, and therefore removes moisture. Each pneumatic germination facility consists of aeration installations equipped with temperature control and moisturizing capabilities, ducts for fresh air, exhaust air, and recycled air, and fans. In addition the germination box is fitted with perforated floors, turning and spraying devices, and with discharging equipment.

The various pneumatic malting installations originate from drum or box malting. Drum malting is practically only to be found in the various forms of the Galland drum, or malting box drum. Box malting, with the further development of tower malting or moving grain bed (Wanderhaufen), is almost the only method that has survived. The specific load of a box is 350 – 500 kg/m^2; this corresponds to a depth of the green malt bed of 0.7 – 1.25 m. Among the special malting systems, either the combined germination – kilning system or malting systems where steeping, germination, and kilning occur in one combined installation are used in practice.

3.1.3. Kilning

In order to stop the chemical and biological transformations that take place during germination, the green malt is dried by kilning. This procedure yields a storable product. Another function of kilning is to remove the vegetable-like flavor of the green malt and to impart to the kilned malt a specific aroma and a defined color characteristic for the type of malt required. Finally, the rootlets, which are very much valued as nutritional feed for cattle, are also removed.

The control of drying processes (withering and kilning), together with the quality and modification of the green malt, determine the character and color of the product.

Drying and Kilning. Moist green malt is very sensitive to high temperature. In order to make allowance for this sensitivity, whithering is carried out at high aeration and low temperature. This step decreases the moisture content from the green malt stage to the hygroscopic point (18 – 20 %).

Further drying is accomplished somewhat more slowly, but is still relatively easy to handle up to the point of "breakthrough", at which the temperature above the kiln floor is higher than that of the wet bulb temperature. Breakthrough occurs when the moisture content is ca. 10 %. The subsequent drying procedure at higher temperature and low aeration becomes progressively more difficult, because the internal moisture must be conveyed from the interior of the grain to the surface. The progress of whithering

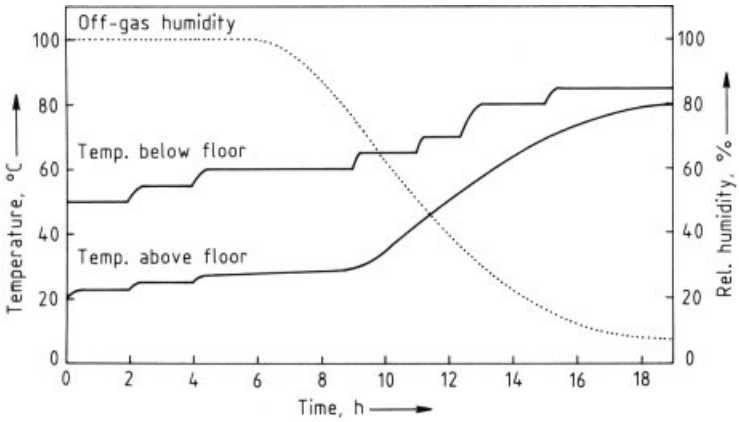

Figure 4. Drying and kilning of pale malt (single-floor kiln, no air recycling)

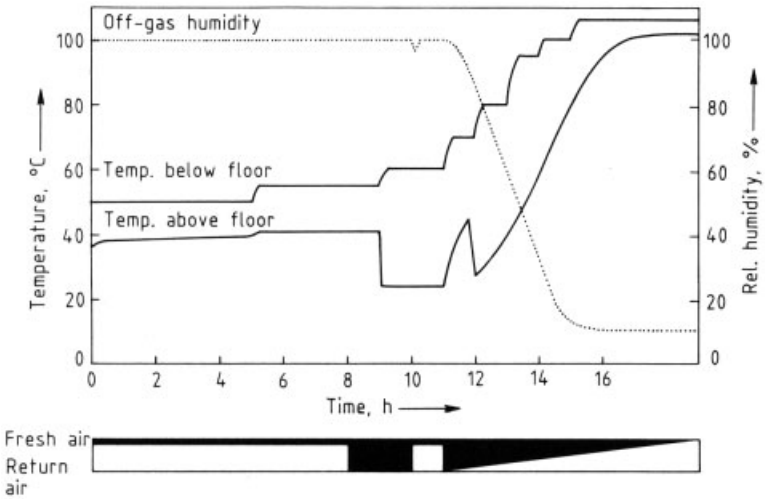

Figure 5. Drying and kilning of dark malt (single-floor kiln, air recycling)

and kilning for pale and dark malts in single-floor, high-capacity kilning equipment is shown schematically in Figures 4 and 5.

During the drying of malt the moisture content, volume, weight, and color of the grain change. The dehydration of the green malt lowers the moisture content, which ranges according to the malting technique from 41 to 46 % (pale green malt) and 48 to 50 % (dark green malt) initially, to 3.5 – 4 % for pale malt and 1.5 – 2 % for dark malt. Careful dehydration ensures that the original volume of the green malt remains unaltered.

The chemical and biological changes encompass three stages with relatively different reactions: natural growth, breakdown, and build-up. The formation of aromatic and colored compounds is most important (Maillard reaction). During initial drying most enzyme levels increase and then decrease during kilning, depending on the temperature [8].

Malt Kilns. During malt kilning the grain bed is aerated with drying air. The most primitive, directly heated, single-floor kilns were further developed to the indirectly heated double-floor and triple-floor kilns. Later single-floor, high-capacity kilns were built.

The combined germination – kiln boxes have a specific loading capacity of 420 – 500 kg/m^2. Indirect heating systems, which keep the formation of *N*-nitrosodimethylamines (NDMA) on a low level, require big heating ovens with large heat-exchange surfaces. The use of recircu-

lated air offers various technological and energy-saving advantages. New methods use a thin vertical or horizontal layer and accomplish drying in 8.5 – 9.5 h. The technique exposes practically every grain to the desired kilning temperature without causing any shrinking of the kernels.

After cooling the kilned malt, the rootlets are removed, and the malt is polished. The malt should then be stored before use in the brewhouse for at least four weeks.

Malt Loss. The changes in weight and volume occurring during steeping, germination, and kilning are shown in Table 5.

Table 5. Volume and mass changes during malting

	Moisture content, wt %	Volume, hL	Mass, kg
Malting barley	14	100	100
Steeped barley	41	145	145
Green malt	48	220	147
Kilned malt	3.5	118	79
Stored malt	4.7	120	80

The maltster is mainly concerned with the quality of the malt and the costs of an adequate barley supply [37]. The production costs are also influenced by malt loss. Loss arises during steeping, respiration, and germination, and amounts to 16 – 25 % (average 20 %) and 5 – 12 % (average 8 %) on a dry-weight basis. For quality control standards, see [8] and [23].

Special Malts. *Wheat malt* is produced according to the processes followed for barley malt. *Chit malt* and *short-grown malt* are exposed to short germination periods and have a low degree of modification. *Caramel malt* is made from kiln malt by resteeping and roasting; kiln malt is converted to *black malt* by roasting at a temperature as high as 220 °C and moisturizing. *Scalding malt* is heated to 50 °C at the end of germination and is subsequently kilned. *Acid malt* is enriched with lactic acid bacteria.

3.2. Technology of Wort Production

The high molecular mass substances of the malt and other adjuncts must be solubilized by grinding and mashing. The extract solution is subsequently separated from the solids by lautering. The lautered wort is then boiled with hops, clarified, and cooled. The basic procedures are shown in Figure 6.

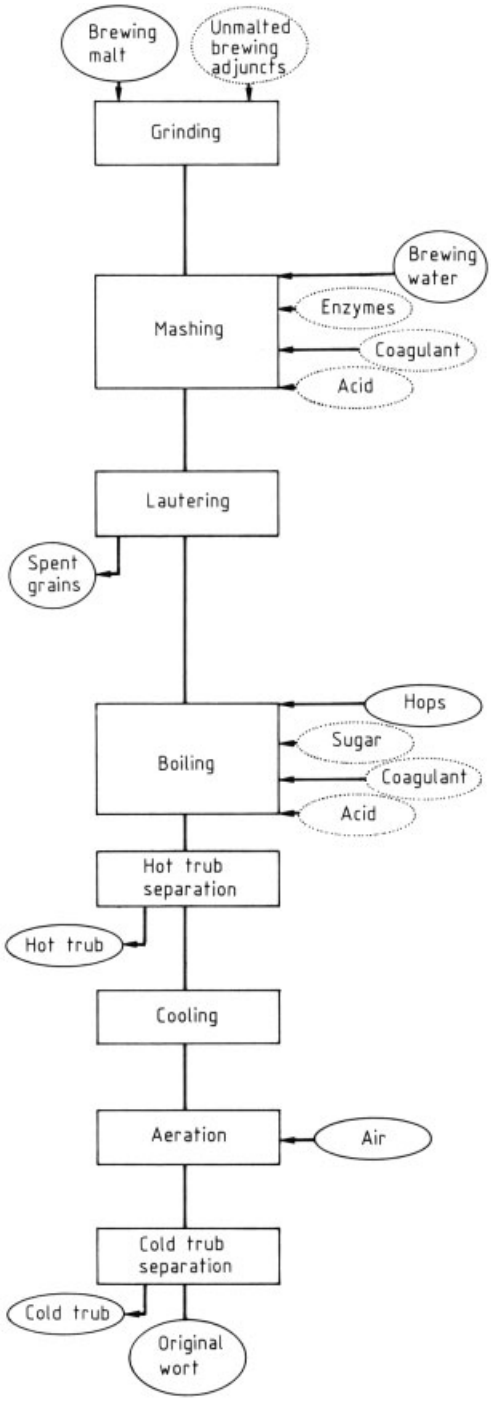

Figure 6. Flowsheet of wort preparation
Additives that are shown in dotted ellipses are not necessary

3.2.1. Grinding of the Malt

The manner in which this purely mechanical procedure is carried out is of critical importance both to the chemical and biological transformations during the mashing process and to the composition and yield of the wort. The husks should be crushed as little as possible in order to prevent the undesirable solution of tannins, bitter compounds, and coloring substances, which could have an adverse effect on the taste of the beer. Furthermore, the husks must serve the specific purpose of forming a filter layer during lautering when a lauter tun is used. In contrast, the endosperm requires fine grinding, because it contains the extract components that are to be dissolved.

Nonuniform modification and the resulting differences in hardness of the individual parts of the malt grains cause the products of grinding to differ in size, extract yield, and ease of breakdown. Particles located at the tip of the kernel are undermodified; they are therefore tough and hard, and result in a coarse grind. The portion near the embryo is more friable, and yields a finely ground flour. Grinding also is the basis for wort preparation, because the fineness of the grist determines the grist volume, which in turn determines the volume of the spent grains. During lautering of the wort, the technique of sparging and raking depends on the characteristics of the spent grains (see Section 3.2.3 and Figure 9). The extent of modification and the moisture content of the malt, the mashing procedure, and the lautering equipment determine the degree of grinding. Proper sampling is most important when particle size is assessed by sieving.

Gristing Mills. The grinding of malt is accomplished by using smooth or fluted cast-iron rollers which rotate either at the same or at different speeds relative to each other. The grinding process may be accomplished in one step, or in such a fashion that certain particles are subjected to repeated grinding. Accordingly, the number of rollers ranges from two to six.

The simplest grinding equipment consists of the two-roller mill, which, however, can be used satisfactorily only on malt that is well modified. When four rollers are used, the two upper rollers function as crushers, whereas the lower rollers accomplish the final grinding. Modern types of mills use shaking screens after the first pair of rollers; smaller particles and husks are thus removed, and only the harder coarse particles are ground twice.

Mills with three passages adjust best to the variations in malt quality. To obtain higher capacities, five- and six-roller mills are exclusively employed (Fig. 7). After crushing, the grist is separated into three main fractions: husks, grits, and flour. The husks are directed to the two-husk rollers, where they are freed of the rough particles clinging to them. The coarse grits from the first and second grinding are finally milled in the grits roller pair. Mills with five rollers work similarly. Mills with several milling stages also permit removal of the husks, which are added to the mash at a later stage of mashing. Conditioning of the malt with low-pressure steam or warm water (increasing the water content by 1 – 1.5 %) during multiple-roller milling proved advantageous. In this way, the husks remain elastic and are not destroyed even with thorough grinding.

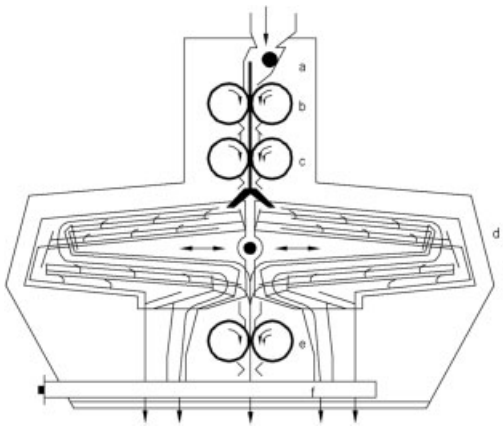

Figure 7. Six-roller malt mill
a) Feed roll; b) Pair of precrushing rolls; c) Pair of husk rolls; d) Sieve box; e) Pair of grits rolls; f) Sampler

In *wet milling* the malt is steeped and then ground in simple double or quadruple roller mills; in the older systems this is accomplished in one vessel, whereby a moisture level of about 30 % is reached within 10 or 20 min at a water temperature of 50 °C or 30 °C, respectively. In modern systems, moisturizing is accomplished by the use of a worm drive, giving a moisture content of 20 – 22 %.

3.2.2. Mashing

During mashing the solid particles are solubilized by the action of the brewing water and the enzymes formed during malting. The optimum conditions at which these enzymes act on the individual compounds are shown in Table 4.

Starch breakdown is most important during mashing. The solubilization of the starch granules during mixing with water and heating proceeds in various steps. After swelling of the starch kernels, gelatinization of the starch occurs as the enzymatic hydrolysis starts. Starch hydrolysis is allowed to continue until no more α-glucans (dextrins), which give a color with iodine, are present, and the desired attenuation limit is achieved.

Protein hydrolysis is as important as starch hydrolysis, even though smaller amounts are transformed. During mashing, numerous endo- and certain exopeptidases (carboxypeptidases) attack the proteins that were already extensively modified during germination. Depending on the degree of protein modification, the enzyme content of the malt, and the mashing conditions, peptides and amino acids are formed, as well as proteinaceous substances of high molecular mass; the latter are responsible for foam, palate fullness, and nonbiological haze.

The hemicelluloses and gums are hydrolyzed during mashing only after they have been released at temperatures up to 50 °C. The β-glucanases are no longer active above this temperature, so that substances of high molecular mass that are released later during wort preparation by the action of another enzyme cannot be broken down any further. For this reason, uniformity and extent of cytolytic modification is of greatest importance.

The acid phosphatases which occur in the malt hydrolyze the organic phosphates; during this process phosphoric acid is released, which decreases the pH value and increases the buffering capacity of the malt, wort, and beer. Excessive buffering, however, attenuates the drop in pH during fermentation.

Increasing the temperature of the mash and the duration of mashing increases the release of polyphenols and anthocyanidines.

Mashing Parameters. The amount of brewing water used for dissolving the grist and for the chemical and biological transformations constitutes the mash liquor; *spargings* are used to yield the residual extract compounds, which remain in the spent grains after lautering the first wort. The amount of brewing water employed is basically different for all-malt pale and dark beers. For pale beers, a greater volume of mashing-in water is chosen in order to obtain thinner mashes and to speed up the enzymatic processes. The volume of brewing water is about 4 – 5 hL per 100 kg of malt for pale beers, and 3 – 3.5 hL per 100 kg malt for dark beers (mash rates 1 : 4 – 5 and 1 : 3 – 3.5, respectively). The spargings (4 – 5 hL per 100 kg malt) must be added in such a manner that the remaining extract can be yielded as completely and quickly as possible. The last runnings (end of spargings) should always be analyzed for economic usefulness and beer quality (tannins, boiling time).

The intensity of the mashing process is determined by the mashing-in temperature. Temperatures of 35 – 40 °C facilitate solution of the substrate and the enzymes, so that these can be more effective at their optimum temperature. At around 50 °C the breakdown of proteins, gums, and phosphates is increased; thoroughly modified malts can be mashed in even at 62 °C, the optimum temperature for β-amylase.

The dissolution of the ingredients of the brewing materials is already initiated by grinding. Further parameters include the rest periods at different temperatures, the pH value of the mash, and the choice between decoction mashing (boiling parts of the mash) and infusion mashing (no boiling of mash portions).

Mashing Vessels. For mashing-in and for the storage of partial mashes a heatable mash tun, equipped with an agitator, is used. In a similar mash kettle, which is usually smaller, the decoction mashes are heated and boiled. In breweries with a two-vessel brewhouse, the lauter tun is also used as mash tun, and the brewing kettle is also used as mash kettle.

Mashing Procedure. The brewer determines the composition of the various specific beer types by the choice of the mashing procedure. All mashing methods can be derived from the three-mash decoction procedure shown in Figure 8. In two-mash procedures, one of the decoction mashes is eliminated. Single-mash

procedures can be accomplished with infusion before drawing off the decoction mash, with infusion after drawing off the decoction mash, or as kettle-mash procedure (boiling of the total mash, cooling, and saccharification with an enzyme extract that was drawn off at the beginning from the top liquid layer). When the malt is overmodified and enzyme-rich, then a high-temperature two-mash method can be used to advantage. The *infusion methods* require good malt modification. Infusion-mash methods with decreasing temperature have become popular for extremely modified malts and certain beer types such as ale. Mashing-in at 75 °C enhances the activity of α-amylase. During mashing-in the temperature drops to 65 °C. *Grist mashing methods* are intended to treat the different grist fractions on the dual basis of their enzyme content and their convertibility. In the process involving husk separation, the boiling of the husks, and therefore their extensive leaching, is avoided.

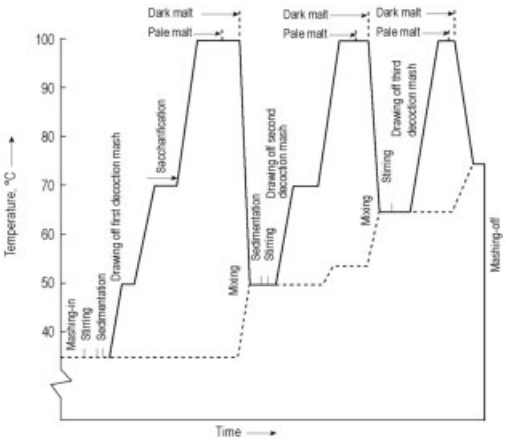

Figure 8. Three-mash method
—- Temperature of decoction mash
- - - - Temperature of rest mash

Processing of Adjuncts. Unmalted cereals must be treated thermally before enzymatic breakdown takes place, thus ensuring the conversion of extract-forming components. Malt amylases facilitate the gelatinization of the finely ground adjuncts at low temperature. The grinding of the malt grist must correspond in fineness to the grist part of the adjuncts, and the concentration of the malt mash must be kept very high because a large amount of water is required for optimum adjunct gelatinization (1 : 4 – 5).

Corn is easier to process than rice, because it gelatinizes at lower temperature. Up to 15 % of adjunct can be added to the first decoction mash without any special treatment. If higher proportions of corn are added, a special adjunct mash will be required, in which the ratio of corn to malt is 2 : 1. When almost complete gelatinization and liquefaction have been achieved, the adjunct is boiled and subsequently added to the malt mash.

Some varieties of *rice* will gelatinize completely only at 88 – 90 °C, even if malt is added. After gelatinization, sufficient malt mash held at temperatures of 30 – 40 °C is added; liquefaction then takes place within 10 min at 78 °C. The liquefied material is subsequently boiled and added to the actual malt mash. After that, the mashing procedures described above can take place. The processing of 10 – 15 % unmalted barley will be possible only if the brewing malts are rich in enzymes. Higher proportions of unmalted barley (30 – 40 %) require the addition of enzyme products.

3.2.3. Separation of Wort

After mashing wort is obtained in two steps: (1) lautering of the first wort by filtration, (2) leaching of the extract that remains in the spent grains with hot water (sparging).

Lauter Tun. Commonly used lauter tuns have a total capacity of 8 hL/100 kg and spent grain depths of 30 – 65 cm. They have a perforated false bottom with an open surface representing about 6 – 30 % of the plate area. Rotating, height-adjustable raking machines loosen up the spent grain bed and a spray device delivers the sparging water. During lautering aeration of the wort should be avoided. The mash should remain homogeneous during pumping in, so that a loose, even filter cake can form. Finally, the wort that runs off should be as clear as possible, so that no particles which could disintegrate further during wort boiling (filterability), and only small amounts of the long-chain fatty acids, which destroy foam, can get into the kettle.

Because of the danger of washing out iodine-reactive α-glucans, the sparge water should have

a maximum temperature of 78 °C. In order to avoid channelling, it should be delivered evenly. The ideal performance of a lautering process in a modern lauter tun is shown in Figure 9.

Mash Filter. In place of the lauter tun, several brewers employ mash filters to separate the mash into solids and liquid. The total mash is transferred into a vertically arranged filter press. The frames are covered on both sides with filter cloth made of synthetic material. At the same time, the air must escape quickly when the homogeneous mash is pumped into the chambers. After yielding the first wort, water is pumped in and the filter compartments of the filter press are pressed together with corrugated steel plates. Because very low volumes of sparging water (0.5 hL/100 kg) are necessary, this facility is very suitable for high-gravity brewing.

Characteristic differences between well-automated mash filters and lauter tuns are the following: (1) independence from the quality of the malt and from the proportion of adjunct, (2) quicker lautering of the more highly concentrated first wort, (3) higher yields, and (4) mostly a hazier filtrate.

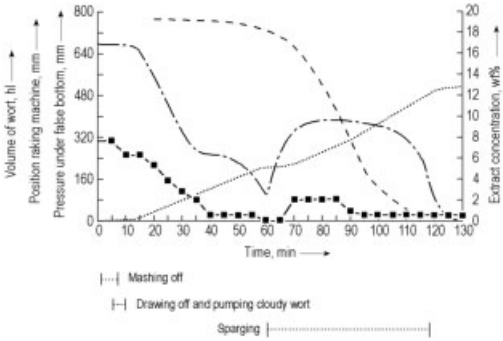

Figure 9. Idealized lautering diagram using a lauter tun
- - - - - - Extract concentration, wt %
–·–·–·– Pressure under false bottom, mm
– – – – Volume of wort, hL
–■–■– Position raking mashine, mm

Strainmaster. The strainmaster consists of a rectangular vessel, the bottom half of which is conically shaped. In the lower part of the vessel perforated sieve pipes of triangular cross section, which have an open surface area of 10 %, are arranged on top of each other. After pumping off the first wort the sparging water can be pumped in from the top and/or the bottom. A very high first wort concentration of 20 – 23 % is necessary in order to achieve a homogeneous mash. Even with large amounts of sparging water it is not possible to achieve yields as high as those obtained with the lautering systems previously discussed, primarily because the spent grains must be withdrawn while very wet. The major advantages of the strainmaster are its large capacity and rapidity in action – lautering is finished in 70 – 80 min – and its simple, automatic operation.

Continuous Lautering. Equipment that allows continuous wort lautering has not been successfully introduced into breweries because the wort that runs off is very hazy. The most important features of such equipment include a rotary mash filter, a screen-conveyor centrifuge, a belt filter, a vacuum drum filter, the Pablo system with screen centrifuges and separators, and hydrocyclones.

Spent Grain Removal. Spent grain with a moisture content of 75 – 80 % is usually sold to farmers and provides a protein-rich nutrient feed for cattle. Spent grains obtained from a strainmaster must first be demoisturized in a screw press. Spent grains obtained from mash filters have a moisture content of only 60 %. An amount of 100 kg of barley malt yields 120 – 130 kg wet spent grains, whereas wheat malt yields 10 – 15 % less. The dry matter consists of 23 – 28 % protein, 5 – 9.5 % fat, 40 – 47 % nitrogen-free extract materials, 16 – 21 % crude fiber, and 4 – 6 % minerals.

3.2.4. Wort Boiling and Hopping

The wort obtained by lautering (full kettle wort) is boiled; during this time hops are added. The process has several functions, namely: (1) evaporation of excessive water in order to achieve the desired wort concentration (original extract), (2) destruction of enzymes, (3) sterilization of the wort, (4) coagulation of the proteins (the flaky precipitate is called hot break or hot trub), and (5) dissolution and isomerization of the bitter substances of the hops in the wort. In addition, undesirable aroma components are removed, reducing substances are formed, color increases, and the pH value falls slightly.

Wort Boiling. The kettles must have a capacity of 9 hL per 100 kg of malt grist in order to achieve the desired effect through boiling. The most important part of the kettle is its heating mechanism. At first, direct firing was used; this was later changed to oil-heating, and then to steam-jacketed kettles with two-zone heating and a cone-shaped central core.

Today hot water or saturated steam systems (internal or external heating) are frequently used. In conventional kettles the evaporation rate, relative to the final volume, is 8 – 10 % per hour. For technological reasons, the boiling time should not be less than 90 min. The denaturation and subsequent coagulation of proteins will lead to coarse, flaky precipitation products (break) and nonopalescent worts only if the finely divided protein complexes have sufficient chance to come into contact with each other for agglutinization; this is achieved by using small steam bubbles and boiling movement. Boiling by means of various external heating systems provides proper rolling and control of the wort; it results in a high evaporation rate of 10 – 15 % per hour as well as a quick and extensive separation of coagulable nitrogen, so that the boiling time can be shortened to 75 min. Further decrease of the boiling time is not possible because of the length of time required for the isomerization of the bitter substances. However, heating at 108 – 112 °C enables the boiling time to be further reduced to about 60 min.

Continuous Wort Boiling. Completely new possibilities are offered by continuous high-temperature wort boiling systems. At a maximum temperature of 135 °C, the hot holding time required is claimed to be only 150 – 160 s. At higher temperature, heterocyclic nitrogen compounds will increasingly form; these have a very low flavor threshold, and impart a bread-like or cracker-type aroma to beer [30]. After the hot holding period, evaporation proceeds in tandem flash evaporators, and should amount to at least 6 % of the original volume in order to drive off such undesirable aromatic substances as aldehydes, ketones, hydrogen sulfide, and dimethyl sulfide. When these parameters are carefully observed, beers can be produced with high-temperature wort boiling which are of the same quality as those produced under normal boiling conditions [38].

Hopping. The dissolution of the bitter substances is very dependent on pH. At a pH of 5.9, the α-acids are molecularly dispersed, partly as salts (humulates); their solubility is 480 mg/L. However, at a pH of 5.2 only 84 mg/L can be dissolved in a colloidal state. At the normal wort pH of 5.4 – 5.6, the colloidal solution prevails. During wort boiling, up to 40 – 60 % of humulone, cohumulone, and adhumulone isomerize; 5 – 1 % remain unisomerized. A proportion of the bitter substances is oxidized, which also contributes to the bitterness of the beer. The major part of the loss must be attributed to incomplete extraction from the hop cone or hop products, and to precipitation. The remaining α-acids are rendered insoluble as a result of the fall in pH during fermentation, and are expelled in the fermentation cover. The β-acids are not transferred to the wort, but the soft and hard resins are dissolved during boiling, and also impart a certain bitter taste.

The isohumulones are more soluble at low pH values, e.g., about 2000 mg/L at pH 5. Boiling and hot stand times are essential for isomerization. In addition, extraction speed, the age of the hops, the amount of α-acids added, and the wort pH all affect isomerization. If additional contact surfaces are available (as in hop extract powders), the isomerization will occur faster. Separate addition of tannin extracts at the beginning of boiling and of extracts of bitter substances later will result in a higher bitter substance yield, because the bitter substances otherwise adsorb onto the hot trub particles and are lost. In the brewhouse, about 60 % of the bitter substances can be transferred to the wort; only 30 % of the amount originally added remains in the finished beer because of further precipitation during fermentation.

The amount of bitter substance added is expressed in mg α-acid per liter of finished wort, and varies according to the hop product used, and the beer type and variety. Pale beer contains about 65 mg/L, export 80 mg/L, and pilsener 80 – 160 mg/L.

The addition of hops according to variety and timing also influences the bitterness of the beer, but most of all the yield of aroma substances.

Color. Coloring and reducing substances are formed during kilning and mashing. Others develop during boiling of the wort. The numer-

ous amino acids react with reducing sugars to form intermediary products which undergo further transformation to brown melanoidines. The products thus formed possess reducing properties and give acid reactions. Polyphenols also add to the color formation by nonenzymatic oxidation and polymerization. Strong formation of coloring compounds leads to broad and harsh tasting beers which age quickly.

The Finished Wort. The hopped beer wort obtained after boiling is designated as finished wort (cast wort, hot wort). Its extract content and volume needs to be analyzed. The brewhouse yield can be determined from these values; the extract yield is expressed as a percentage of the amount of malt used (range between 78 and 81 %). Table 6 shows the composition of a pale 12 % finished wort.

Table 6. The composition of pale lager wort (12 %) made from barley malt

Carbohydrates (100 %):	
Hexoses	7 – 9 %
Sucrose	3 %
Maltose	43 – 47 %
Maltotriose	11 – 13 %
Lower dextrins	6 – 12 %
Higher dextrins	19 – 24 %
Pentosans	3 – 4 %
Gums	0.2 %
Nitrogen compounds:	
Total nitrogen	950 – 1150 mg/L
High molecular mass nitrogen	200 – 300 mg/L
Low molecular mass nitrogen	550 – 700 mg/L
Free amino nitrogen	200 – 250 mg/L
Bitter substances	25 – 35 EBC bitter units
Polyphenols:	
Total polyphenols	180 – 250 mg/L
Anthocyanidines	70 – 110 mg/L
Minerals	15 – 20 mg/L (80 % inorganic, 20 % organic)
Zinc	0.1 – 0.25 mg/L
pH	5.0 – 5.7
Viscosity (20 °C)	1.7 – 2.0 mPa · s

Wort Concentrates. In some countries it is permitted to manufacture hopped wort concentrates and unhopped malt extracts from wort made by the usual brewing method. Concentration to 70 – 80 % dry matter is achieved either by vacuum evaporation or by lyophilization. These concentrates can be diluted back to the desired extract content in the kettle; they are processed subsequently in the normal manner.

3.2.5. Wort Treatment

Hot Trub Separation. The hot trub contains the nitrogen compounds that coagulate during boiling. These must be removed completely before fermentation, or else the beer will taste wort-like, bitter, and even harsh. The hot trub consists of 40 – 70 % protein, 7 – 15 % bitter substances, 20 – 30 % of other organic compounds, such as polyphenols, and mineral substances. Hot trub yield is between 400 and 800 mg of extract-free dry matter per liter of finished wort. When whole hops are used, the cast wort must first be cleared of spent hops by passing it through a hop back. In the past, the wort was pumped into the flat coolship, where it would cool, and the hot trub and part of the cold trub would sediment. Because of the danger of infection on the large surface of the coolship, the heat loss, and the great expenditure of work, other methods of hot trub separation have been developed.

The *settling tank* (hot wort receiver) permits a good separation of the trub and spent hops, but the amount of sludge collecting on the bottom of the vessel creates problems because 2 – 5 % of the wort is trapped in this sludge. It can, however, be recovered by centrifuging the mixture of trub, spent hops, and wort employing a chamber or plate centrifuge. The wort recovered in this way can be added to the same batch of wort or to the next brew.

The total hot wort is frequently clarified by means of efficient, self-cleaning centrifuges; these, however, will only work reliably if the wort – trub mixture is added homogeneously. This can usually be achieved by means of an intermediate hot wort tank with a stirring mechanism. The most thorough removal of trub and spent hops can be achieved by filtering the hot wort through kieselguhr.

An inexpensive solution is the *whirlpool tank*, where the wort is pumped tangentially into a cylindrical vessel. This creates an even, rotating stream. The solid particles suspended in the rotating liquid will separate due to friction (tea cup effect), migrate to the bottom center, and coalesce to form a cake. It is also possible to achieve this effect in whirlpool kettles, where both boiling and trub separation take place, provided that the bottom of the kettle is appropriately shaped.

Wort Cooling. The wort, which has been freed of hot trub, is cooled to 4 – 7 °C for cold bottom fermentation, to 10 – 15 °C for accelerated bottom fermentation, and to 12 – 18 °C for top fermentation.

Cold Trub Removal. The very fine flaked cold trub appears at temperatures of 70 – 55 °C. It consists of around 50 % protein, combined with 15 – 25 % polyphenols and 20 – 30 % carbohydrates of high molecular mass. At 0 °C about 150 – 300 mg/L of cold trub is formed. Opinions are divided regarding the necessity of cold trub separation. The presence of cold trub can, under certain circumstances, accelerate fermentation because of the presence of long-chain unsaturated fatty acids. The yeast will contain a higher level of impurities and the filterability of beer brewed in this manner may be poor.

A fairly good separation of the cold trub can be obtained in the starting vessel by sedimentation after pitching. Even more effective are: cold sedimentation with or without the addition of kieselguhr, cold centrifugation, and, very elegantly, flotation not only with aeration but the simultaneous addition of pitching yeast. Filtration of the wort removes the cold trub quantitatively; however, a deficiency of minerals and fatty acids could be created by this method. Blending of unfiltered wort can be helpful in this case.

Aeration. The oxygen necessary for yeast propagation (7 – 8 mg/L, corresponding to 80 % saturation of the wort with O_2) is usually introduced in the form of air at the pitching temperature. If pure oxygen is used, it must be added carefully, so that the oxygen level does not exceed 15 mg/L; a higher level is detrimental to the yeast. Optimum aeration can compensate for the disadvantages of an extensive wort clarification. All closed cooling systems require separate aeration of the wort. During normal wort aeration 5 – 15 L of air per hectoliter of wort is needed; flotation requires 40 – 60 L/hL.

Sterile air may be introduced on the cold wort side of the plate cooler with even dispersion by air nozzles or venturi jets, or in the hot wort centrifuge, provided the wort is cooled immediately. A combined hot and cold aeration in the ratio of 1 : 5 also is possible.

3.3. Bottom Fermentation

The flow sheet of Figure 10 describes the basic operations for popular beer varieties from fermentation to finished beer.

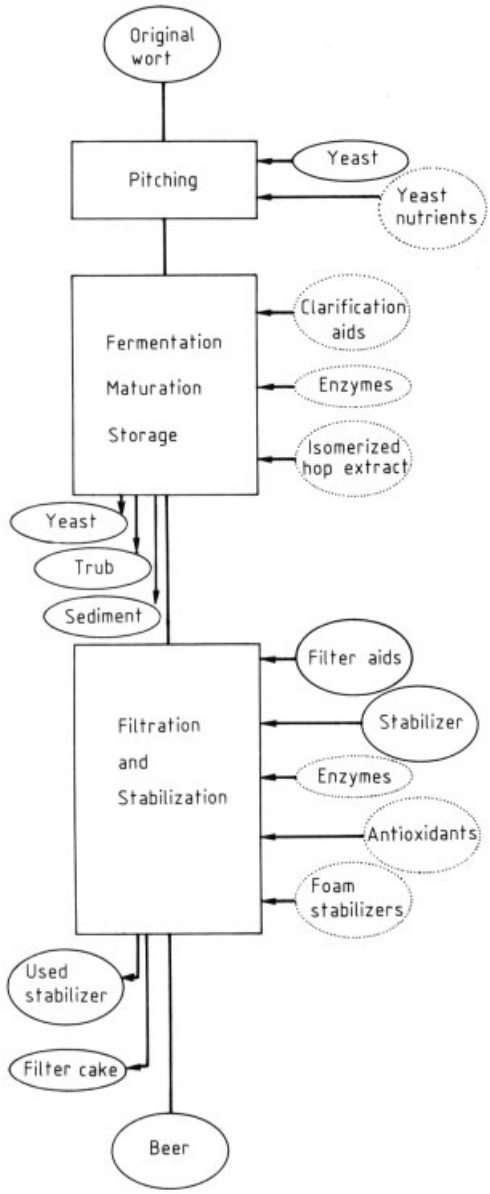

Figure 10. Flowsheet of beer fermentation from pitching to final product
Substances that are shown in dotted ellipses are not necessary

3.3.1. Fermentation

The fermentation process is initiated by the addition of 0.5 – 0.7 L of a heavy yeast slurry per hectoliter of wort, corresponding to $15 - 20 \times 10^6$ yeast cells per milliliter of cooled and aerated wort. This procedure is called *pitching*. The batchwise addition of original wort to fermenting green beer is called *doubling*. The main products resulting from the fermentation are ethanol and carbon dioxide. Other reaction products include: higher aliphatic and aromatic alcohols, esters, organic acids, carbonyl compounds, sulfur-containing compounds, and polyhydric alcohols, all of which are important for the properties and quality of the resulting beer (for details on alcoholic fermentation, see → Ethanol). All the compounds formed have different taste and odor thresholds. Their combined contributions make up the flavor or off-flavor of the beer; the amounts produced can be influenced to some degree by brewing technology. Further changes in the wort are caused by the fall in the pH value (see below). The pH drop leads to the precipitation of nitrogen compounds of high molecular mass, of polyphenols, and of bitter agents. The result is a decrease in color and a debittering effect.

Yeast has the ability to adjust its metabolism to aerobic as well as to anaerobic conditions. The yeast doubles or triples its mass during fermentation. For the build-up of cell substance (proteins and enzymes) the yeast needs mostly amino acids, which it either takes from the fermentation substrate or must synthesize by itself. Besides proteins, lipids are also synthesized for yeast propagation because they are important components of the cell wall, and are needed for the uptake of nutrients. For the synthesis of these lipids from acetyl coenzyme A, molecular oxygen is needed; after lautering, wort itself contains only few lipids. Finally, the yeast also requires minerals for the stabilization of its enzyme systems.

Fermentation Byproducts. During the main fermentation, the pH value decreases by one unit because volatile (acetic, formic) and nonvolatile *organic acids* (pyruvic, malic, citric, lactic) are formed. The pH of beer ranges from 4.3 – 4.6. The intensity and speed of acid formation is determined by the buffering action of the wort, the amount of easily assimilated nitrogen, the yeast strain, and the fermentation schedule used. If the pH drops quickly, gums that retard filtration will precipitate, but valuable colloids also are lost. The pH has a direct effect on the flavor and the liveliness (sparkle) of the beer. Short-chain fatty acids are formed during the fatty acid synthesis at the beginning of the main fermentation process: butyric, isovaleric, hexanoic, octanoic, and decanoic acid. Their amounts can be controlled by the wort composition, aeration, yeast strain, and general fermentation conditions. During pressure fermentation, increased levels of these compounds can be expected during maturation. Even in very low concentrations, they cause a yeasty odor and impair head retention.

The higher *aliphatic alcohols* (1-propanol, 2-methyl-1-propanol, 2-methyl-1-butanol, and 3-methyl-1-butanol) and *aromatic alcohols* (especially 2-phenyl-1-ethanol) represent the largest fraction of the compounds responsible for the aroma of the beer; at concentrations which are too high, they will adversely affect taste and quality. Their levels can be controlled to some extent by manipulating the content of free amino nitrogen, wort concentration, pitching rate, yeast strain, pitching temperature, and fermentation temperature.

Because of their low taste threshold values, *esters* strongly influence the organoleptic properties of the beer. Esters are products of enzymatic catalysis and their formation is very closely related to yeast propagation and lipid metabolism. Wort aeration, pitching rate, fermentation temperature, the yeast strain, and the counterpressure during fermentation all have great influence on ester formation. Measures intended to intensify yeast propagation will lower ester concentration.

Sulfur compounds (hydrogen sulfide, sulfur dioxide, dimethyl sulfide, 3-methylthio-1-propanol, and thiols) are not desirable in beer because of their specific odor and taste. Besides efficient trub removal the most important factor in the formation and reduction of these flavor-active substances is the yeast strain.

Glycerol (1300 – 2000 mg/L) is formed as a byproduct during glycolysis; its concentration depends on the amount of fermented sugars.

Aldehydes and ketones are responsible for the aroma of green beer and for the stale flavor. Acetaldehyde formed in the green beer does

not present any technological difficulties. Off-flavors in beer are usually caused by a high level of diacetyl and 2,3-pentanedione; these compounds are responsible for unfavorable, buttery flavor in the beer. The taste threshold of diacetyl depends on beer type, and ranges from 0.10 to 0.12 mg/L; that of 2,3-pentanedione is 0.5 – 0.6 mg/L.

3.3.2. Maturation

The total diacetyl concentration is used to judge the maturity of purged beer, and must be decreased below the flavor threshold by means of brewing technology. The diacetyl precursor 2-acetolactate is called "potential diacetyl", because it transforms into free diacetyl only in the filtered, yeast-free beer, and can then not be broken down any further. In calculating the total diacetyl concentration, 2-acetolactate must be added to the amount of free diacetyl.

Diacetyl Metabolism. During the propagation phase the yeast cells need numerous nitrogen compounds for the formation of yeast protein. If there is not sufficient fixed nitrogen present in the wort in the form of compounds which can be assimilated, the yeast will use a combination of carbohydrate and protein metabolism. During the valine synthesis diacetyl can be formed via 2-acetolactate by oxidative decarboxylation as shown in Figure 11. This step, which is catalyzed by yeast enzymes, is highly temperature-dependent, and occurs very slowly below 10 °C. Diacetyl itself is present in very small quantities in fermentation samples and in green beer, because its reduction to acetoin is much faster than its formation. The final product 2,3-butanediol is, as far as taste is concerned, unobjectionable; its concentration in beer never exceeds its flavor threshold. Bacteria, which may occur in the brewery as infections, also are likely to promote the formation of diacetyl by the route shown in Figure 11. Some types of bacteria contain an enzyme which can decarboxylate 2-acetolactate directly to acetoin while avoiding the limiting maturation step.

Fermentation Conditions. The formation and reduction of 1,2-diketones is dependent on (1) a sufficient supply of free amino nitrogen and other yeast nutrients, (2) proper pitching and doubling conditions (sufficient aeration, low pitching temperature, optimization of pitching rate regarding yeast quantity and timing), (3) a careful selection of the yeast (yeast strain, physiological condition of the yeast, no infection), and (4) control of the fermentation and maturation conditions favorable for the degradation of diacetyl. The most effective parameter in this respect is temperature control during fermentation and maturation, which is the basis for the fermentation procedure shown in Figure 12.

Using the combination of *cold fermentation – cold maturation*, a tasty beer can be produced by simple means. It is advantageous to remove most of the yeast at the end of fermentation and to achieve secondary fermentation by the addition of "krausen" (green beer in its initial fermenting stage).

If the combination of *warm fermentation – warm maturation* is considered, the formation of fermentation byproducts is increased at the prevailing high fermentation temperatures. The rapid pH drop results in losses of bitter substances, decreased foam stability, and an unsatisfactory yeasty flavor. Pressure fermentation using carbon dioxide is a remedy. This will slightly decrease the fermentation rate and control yeast propagation, thus curbing the formation of fermentation byproducts.

The combination *cold fermentation – warm maturation* avoids the formation of undesirable flavors and decreases the level of diacetyl safely; it leads to beer of constant quality. In this case, temperature control is optimally adjusted to the metabolism of the yeast. In *programmed maturation*, heat exchangers are used in order to raise the temperature to 20 °C. The addition of 10 % "krausen" with an apparent degree of attenuation of 20 – 30 % is practiced at the beginning of this maturation phase. Accelerated fermentation and maturation are also achieved by stirring fermentations and then maintaining a maturation step and subsequently purging green beer with carbon dioxide. However, an excess of fermentation byproducts is formed by this method.

Fermenters. Fermentation and maturation are carried out in open or closed fermenters, horizontal tanks, or vertical fermentation tanks with conical bottoms [39]. In two-tank processes, fermentation, maturation, and storage occur sepa-

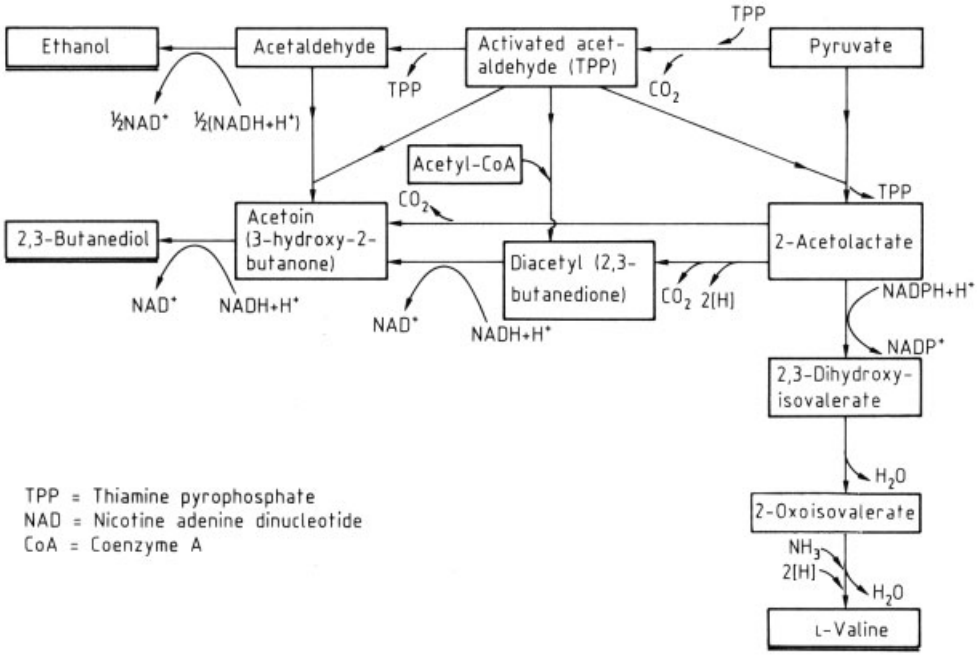

Figure 11. Synthesis of diacetyl by yeast and bacteria

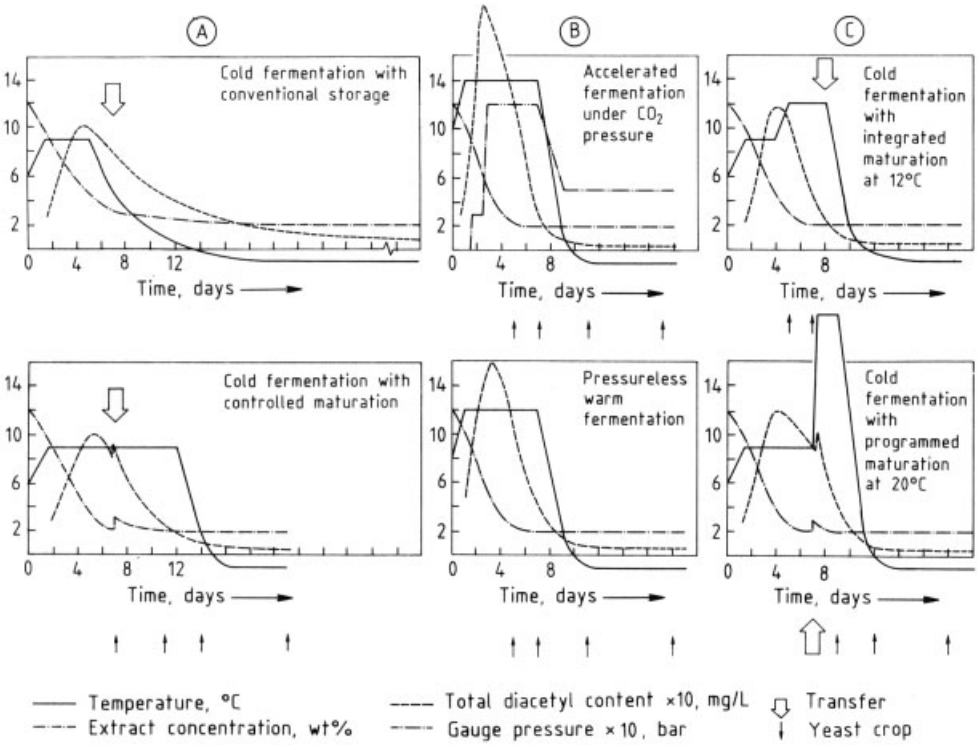

Figure 12. Fermentation in practice
A) Cold fermentation – cold maturation. B) Warm fermentation – warm maturation. C) Cold fermentation – warm maturation

rately; vertical tanks with conical bottoms are used in this case.

The capacity of the cooling equipment and heat exchange surface of the tanks should be designed for maximum heat development or, in the case of single-tank processes, for a maximum cooling rate.

Fermentation Stages. The brewer identifies the changes occurring in the green beer by carefully observing the individual fermentation stages: the creaming, the head formation, the rocky head, and the decreasing of head at the end of fermentation. Extract decrease and temperature level must be checked constantly. Attenuation also is indicated by a fall in pH value, by a phase of cell propagation and cell sedimentation (turbidity), by a decrease in coloration, and by a lowering in redox potential.

The carbon dioxide formed during fermentation amounts to 2 – 2.5 kg per hectoliter of beer. It is collected and can be used for carbonation of soft drinks or for low-oxygen bottling and racking.

3.3.3. Cold Storage

During cold storage the beer must be carbonated to the desired CO_2 level (0.48 % for draft beer, 0.50 % for canned beer, 0.55 % for bottled beer). This can be achieved in the conventional procedure by using a definite bunging overpressure of 0.2 – 0.6 bar, depending on hydrostatic pressure and temperature. Beers that underwent warm maturation require either higher pressure, or carbon dioxide to be added during transfer from warm to cold storage tanks. During storage, the beer must clarify by allowing the yeast and other haze-causing materials to settle, and its taste must refine and round off. In order to achieve these requirements, the beer must be stored at 0 to $-2\,^{\circ}$C during the last week or two. When warm maturation is practiced, the storage period cannot be as easily defined. Frequently fermentation, maturation, and storage take place in the same vessel (one-tank process).

By using separate tanks for the sedimentation of the yeast after fermentation (flocculation tanks), yeast content and degree of attenuation are further balanced. During transfer to the cold storage tanks krausen is added. Separate storage tanks also are used when the beer is stabilized with bentonite during the second half of the storage period.

Continuous fermentation may be accomplished by through-flow or by tributary-flow systems or in a bioreactor. These methods are rarely found in large production facilities.

3.3.4. Filtration

For filtration theory and beer filtration, see [40].

Besides an impeccable taste, perfect clarity is expected of a stored and matured beer. Solid and hazy particles still present in the beer (yeast, protein – tannin particles, and hop resins) are removed by filtration. Filtration also improves biological and chemical-physical stability. Filtration is carried out at low temperature (possibly at 0 to $-2\,^{\circ}$C) under a counterpressure of carbon dioxide above its saturation level, and with minimum uptake of oxygen [42].

Filtration systems used for pre-clarification were formerly pulp filters; today brewers use mostly plate and frame filters, pressure leaf filters, or candle filters for cake filtration with a filter aid (diatomaceous earth [41], perlite), or centrifuges. For subsequent final clarification and sterile filtration, sheet filters made from cellulose and diatomaceous earth are used. For sterile filtration filter membranes made from cellulose esters of definite pore size may be used.

The choice of the clarification method depends on capacity, technical considerations, and economic conditions. Another factor is the filterability of the beer, which is not always proportional to the viscosity, but also depends upon the nature and amount of filtration-retardant materials present, such as α- and β-glucans, proteins, and a high yeast content.

3.3.5. Stabilization

In bright beer high molecular proteins and tannins tend to aggregate and form haze. This process can be delayed by removing one of these fractions which improves the chemical-physical shelf life of the product. Most common is the removal of part of the tannins by adding polyvinylyrrolidone (10 – 50 g/hL) as an adsorbent. Polyvinylpyrrolidone is dosed in the filtered

beer and retained in a pressure leaf filter. The loaded polyvinylpyrrolidone is subsequently reprocessed and reused.

3.3.6. Types of Bottom-Fermented Beers

In many countries, bottom-fermented beer is designated as *lager beer* [43]. Its extract of original wort varies according to local laws (tax classification) from 7 to 14 %. Lager beers are the most popular, with an average bitter substance content of 20 EBC bitter units (see Table 7 and Section 5.3).

Table 7. The composition of bottom-fermented pale lager beer (12 % extract of original wort)

Extract of original wort	12.0 wt % (125.6 g/L)
Attenuation limit (apparent)	78 – 85 %
Real degree of attenuation	65 – 80 %
Apparent residual extract	1.7 – 3.0 wt %
Real residual extract	2.0 – 3.5 wt %
Alcohol concentration	3.5 – 4.5 wt %
Total nitrogen	700 – 900 mg/L
Coagulable nitrogen	8 – 28 mg/L
High molecular mass nitrogen	150 – 250 mg/L
Low molecular mass nitrogen	300 – 600 mg/L
Free amino nitrogen	80 – 160 mg/L
Bitter substances	16 – 25 EBC bitter units
Total polyphenols	130 – 180 mg/L
Anthocyanidines	40 – 80 mg/L
pH	4.3 – 4.6
Viscosity (20 °C)	1.4 – 1.7 mPa · s

Beer with an Extract of Original Wort of 10 – 14 %. This range comprises an extraordinarily large variety of beer types, including pale and dark beers, export beers (more than 12 % extract of original wort), Märzen beers, special beers, and festival beers (13 – 14 % extract of original wort). Within these limits there are such different beer types as Pilsener, Dortmunder, Munich, as well as smoky-flavor beers and cellar beers; these are however restricted to certain localities. The upper gravity limits for special beers differ from country to country: for instance, In Germany 14 %, in Austria 13 %, in Italy 14 %, and in France 15 %.

Strong Beer. In Italy, pale and dark beers of more than 15 % extract of original wort are classified as strong beers; in Austria and Germany beers from 16 % up to a maximum of 28 % extract of original wort also belong to this class.

3.4. Top Fermentation

Top-fermented beers differ from bottom-fermented beers by their special aroma which is primarily induced by the top-fermenting yeast strains of *Saccharomyces cerevisiae*. The particular yeast strain employed has a higher optimum fermentation temperature, and therefore the fermentation proceeds between 12 and 25 °C. During fermentation, the yeast rises and can be skimmed off the top. In large vessels, especially the cylindrical fermentation tanks with conical bottoms, the yeast is, however, cropped in the same manner as in bottom fermentation. The number of yeast generations is considerably greater. At the higher fermentation temperature, the amount of diacetyl is usually easily decreased. Because of the fast rate at which fermentation proceeds, a relatively low pH value of 4.1 – 4.3 results.

Types and Production of Top-Fermented Beers.

Wheat Beer [44]. The extract of original wort in wheat beer is 11 – 14 %; the wheat malt portion can range from at least 50 to 100 %. An intensive two-mash decoction procedure is needed in the brewhouse in order to ensure a satisfactory protein modification, because an increase in wheat malt proportionally decreases the concentration of assimilable nitrogen compounds in the wort. Because of the higher pitching temperature (12 – 18 °C), the pitching rate required will be lower (0.3 – 0.5 L of yeast per hectoliter of original wort, corresponding to $7 - 15 \times 10^6$ yeast cells per mL). The aeration should ensure an oxygen concentration of 6 – 8 mg/L. The initial fermentation stage is characterized by the rise of trub particles and hop resins to the surface. After their removal, yeast rises to the top and can be cropped; this continues until the end of fermentation. The special wheat beer yeast can be repitched as often as 200 – 500 times. Wheat beer is characterized by a typical spectrum of fermentation byproducts, such as 4-vinylphenol and 4-vinylguaiacol; these two compounds are responsible for the typical aroma of wheat beer. A more rapid and extensive pH drop, an increased formation of higher alcohols and esters, together with a greater decrease in nitrogen and bitter compounds, mark the course of wheat beer

fermentation as compared with lager beer procedures. A higher bunging pressure during storage ensures a carbon dioxide concentration in wheat beer of 0.7 – 0.9 %. If wheat beer is marketed as "naturally hazy", the secondary fermentation can be accomplished in the bottle by adding unfermented wort and bottom-fermenting yeast or bottom-fermented krausen. Crystal-clear wheat beer remains in the tank until mature and is subsequently filtered and bottled.

Alt beer also has an extract of original wort of 11 – 14 %. Methods vary widely for the production of the dark Alt beer which gives readings of 25 – 40 EBC coloring units: the wort may be produced from pale malt with the addition of caramel coloring or colored beer, 100 % dark malt, or 90 % pale malt and 10 % dark caramel malt. A proportion of 10 – 15 % pale wheat malt sometimes is used to round off the taste. The black malt which is used as a substitute for caramel may also be produced from wheat. The bitter substance concentration of Alt beer amounts to 28 – 40 EBC bitter units. The fermentation temperature is 12 – 22 °C.

Kölsch beer (extract of original wort 11 – 14 %) may be produced only in the town of Cologne. It is brewed with pale barley malt by adding up to 10 – 20 % wheat malt, and fermented at 12 – 22 °C, but sometimes up to 28 °C. The character of Kölsch and Alt, formerly designated as top-fermented bitter beers, is strongly determined by the properties of the specific yeasts; a large variety of flavors results.

Berliner Weisse (White) also is named after its place of origin; it has an extract of original wort of 7 – 8 %. Its very low pH value of 3.2 – 3.4 originates from the combined yeast and lactic acid fermentation. A degree of attenuation sometimes exceeding 100 % results in an alcohol concentration of 2.5 – 3 % and a lactic acid concentration of 0.25 – 0.8 %. The bitter substance content amounts to 4 – 6 EBC units, whereas the carbon dioxide concentration lies between 0.6 and 0.8 %. The beer has an acidic taste, which is marked by a pleasant estery – flowery quality, depending on the production method and the particular type of beer; it is frequently served with raspberry or woodcruff syrup.

Malt Beer. Sugar and sugar syrup may be used in the production of top-fermented nutrient beers. The beers are brewed with 7 – 8 % extract of original wort, and are enriched after filtration with sugar until an extract of original wort of 12 % results. The alcohol concentration of malt beer must be under 0.5 %. The pH lies between 4.5 and 4.9 depending upon the method of wort production and the degree of attenuation; the carbon dioxide concentration is 0.4 – 0.5 %, the color rating ranges from 50 to 80 ECB units (65 – 80 % dark malt, 3 – 5 % dark caramel malt, 3 – 5 % acid malt, and the remainder pale malt) and the bitter substance content is 6 – 10 EBC units. Malt beers are pasteurized on account of their high content of fermentable sugar.

British Ales. British ales have an extract of original wort of 7 – 13 %, according to the tax category. Depending on the type of malt charge and adjunct (mostly corn), on the choice of top-fermenting yeast strains, as well as on the "primings" added in the cold storage tank (caramel coloring), various types are obtained such as ale, brown ale, and bitter ale. The so-called "best bitter" is remarkably bitter due to special hopping methods; it also possesses a strong hop aroma. Bottled and canned beers are filtered, carbonated, and stabilized, according to the demands of the market. The original draft ale is clarified with gelatine or isinglass, and dry-hopped in the barrel or keg.

Stout is a dark, strongly hopped beer, with an extract of original wort between 11 and 18 %. The weaker stouts also are called *porter*. Some of these are treated with a special post-fermentation yeast (*Saccharomyces brettanomyces*), which imparts a typical aroma to the product.

Geuze and *Lambic* are Belgian spontaneously fermenting beers with an extract of original wort of 11 – 12 % [45]. They are produced from certain admixtures of raw materials other than malt. The hop content varies widely, as does the method of wort production. Fermentation is not caused by a definite yeast strain, but by the airborne organisms of the fermentation rooms and vessels. Because of this spontaneous fermentation, the beers vary not only in their alcohol level, but also in their lactic acid concentration. According to the amount of acidity, the beers are either left in a natural state and consumed with added sweeteners, or a specific amount of sweet mash is added.

Top-fermented strong beers with an extract of original wort of more than 16 % include the

German pale and dark *Weizenbock* beers as well as the different varieties of stout.

3.5. Special Production Methods

A number of special production methods have been designed to produce beers with very specific properties. The legal regulations covering such products differ in various countries.

3.5.1. Dietetic Beer

Dietetic Beers are pale beers of Pilsener brewing type; in Germany they may contain only 0.75 g bioavailable carbohydrates and 0.5 g protein per 100 g of beer. In order to achieve these figures, wort of 11 – 12 % extract must be fermented until the apparent degree of attenuation is more than 100 %. Brewhouse procedures are adapted towards achieving this aim by employing mashing with extended rest periods at 60 – 66 °C. During cold fermentation (7 – 12 °C), a small proportion of malt extract that has been drawn from the mash at 50 °C is added. This causes breakdown of the remaining high molecular mass carbohydrates and proteins. The addition of malt flour also has proved successful in obtaining the properties of dietetic beer. The beer must be carefully stabilized because the addition of unboiled extracts increases considerably the amount of coagulable nitrogen compounds. Almost complete breakdown and fermentation of the extract results in the alcohol concentration rising to 4.8 – 5 %. This is partly viewed as a disadvantage. Thus, the alcohol concentration is subsequently lowered by either distillation in film evaporators or reverse osmosis. It is much easier to produce dietary beer with a lower extract of original wort because the control of both alcohol concentration and fermentation is simpler. The intensive fermentation leads to pH values in the range of 4.1 – 4.5; the bitter substance content varies between 22 and 40 EBC units.

3.5.2. Nutrient Beer

Nutrient beers are bottom-fermented beers which are brewed with 100 % malt. They are classified either as "low-alcohol" (alcohol concentration under 1.5 %) or as "alcohol-free" (alcohol concentration under 0.5 %). The extract of original wort of these very dark beers (60 – 80 EBC coloring units) is between 11.5 and 12.7 %; the apparent degree of attenuation is 25 – 30 % in the low-alcohol beer and 8 – 10 % in the alcohol-free beer. The pH is 4.7 – 4.9, according to the degree of attenuation and other technological steps (acid malt); the bitter substance content is low (6 – 10 EBC units). The flavor of alcohol-free nutrient beers may be improved by increasing the alcohol concentration to 0.7 % and then blending with first-wort extract before filtration. It is also possible to remove the alcohol partly. Complete pasteurization of such beers is mandatory.

3.5.3. Low-Alcohol Beer and Alcohol-Free Beer

The many varieties of low-alcohol beers may be divided into three groups:

1) Beers where the fermentation is stopped by filtration and pasteurization. An alternative to heat treatment is a *cold shock process*, during which a beer that has just started to ferment is rapidly cooled in a thin film. This produces a spectrum of fermentation byproducts similar to that of normal beer.
2) Beers that are fermented with a special yeast strain, e.g., *Saccharomyces ludwigii*; these yeasts metabolize only hexoses and sucrose. The major fraction of the malt sugars, such as maltose and maltotriose, remains unfermented.
3) Beers that are fermented almost normally, but where part of the alcohol is later withdrawn by thin-film evaporation, vacuum distillation, reverse osmosis, or dialysis.

They latter two methods are currently the most common ones, because the taste of the dealcolized beer is more similar to that of normal beer.

3.5.4. High-Gravity Brewing

This procedure is popular because existing installations can be better utilized and because a higher capacity can be achieved without new investment [43]. In general, the wort is brewed

as strongly as the brewhouse equipment will permit, e.g., with 13 – 18 % extract of original wort instead of 11 – 12 %. Concentrations of 16 – 18 % are only attainable without loss of yield if syrups are added at the end of wort boiling. After fermentation and maturation, these stronger beers are adjusted to the desired extract of original wort with carefully processed water (deaerated, carbonated, and sterilized). The wort concentrations are usually no higher than 14 – 1 5 % so as to maintain the ratio and level of byproducts and, thereby, the normal beer taste. High-gravity beers always have a poor head retention.

3.6. Filling

After filtration, all operations must be directed to maintain the quality of the beer. Mistakes made at this point are very hard to rectify and will only become noticeable much later (re-infection and aged taste caused by a high oxygen concentration). Furthermore, the beer must be bottled under an appropriate pressure to prevent the escape of carbon dioxide. Filling and packaging are subject to various legal regulations just as the production of beer itself. These regulations pertain to labeling, container capacity, and volume tolerance.

The amount of beer lost depends on the production facilities, and to a certain extent on capacity, and ranges between 3 and 10 %. Examination of *beer loss* provides an insight into loss of volume, which occurs from casting to the actual process of bottling, and into loss of extract.

Kegging. Barrels are manufactured of oak wood, aluminum alloys, stainless steel, or synthetic materials lined with stainless steel; their size ranges from 10 to 250 L. After the barrel has been thoroughly cleaned, it is purged with carbon dioxide or, less advantageously, with air, and filled under counterpressure (isobarometric). Bowless filling units not only avoid the danger of infection, but they also protect the beer from extensive contamination with oxygen.

Barrels with a cylindrical edge are called *kegs* or *system barrels*; they have a permanently installed fitting for cleaning, sterilization, filling, and tapping. With such a device, the barrel remains sealed and under a pressure of carbon dioxide even after it has been emptied. In this fashion the drying-up of beer remnants is avoided, and cleaning is facilitated. Cleaning and filling under sterile conditions are easily automated. Kegs and fittings are standardized.

Cellar Tanks. For larger distributors stationary draft beer tanks which hold 10 to 30 hL can be installed. The beer is delivered from the brewery in large tank trucks. The cellar tanks are equipped with cooling devices. They are lined with disposable polyethylene bags, which makes it possible to store the beer under impeccable sanitary conditions without the need to clean the cellar tank. The excess pressure necessary to dispense the beer is created by gas which is admitted between the inner wall of the tank and the polyethylene bag.

Bottling. Bottles may be made of glass or plastics [poly(ethylene terephthalate), PET, or poly(ethylene naphthalate, PEN)]. Whereas glass is diffusion-proof, not only can carbon dioxide escape from plastic bottles, but also oxygen can diffuse into the beverage, and cause an undesirable oxidation taste. A metal crown with synthetic liner is widely preferred over other means of sealing the bottles. Returnable bottles must be cleaned to a microbiologically impeccable standard before refilling. The step can be preceded by a steam treatment. Disposable bottles are relatively expensive because their manufacture requires more energy and raw material than multiple-use bottles.

Isobarometric bottling of glass bottles is best carried out by pre-evacuation of the bottles, intermediate puging with carbon dioxide, second evacuation, counterpressurizing and then filling. After filling fobbing is induced by various means; this provides for a maximum displacement of air. The crown closure machine presses the crimped edges of the crown around the mouth of the bottle, which is then labeled automatically. Modern bottling machines fill up to 120000 bottles per hour. Brown glass bottles are preferred to those made of green glass, because brown bottles protect the beer better against light of short wavelengths, which causes photodegradation of the bitter acids of hops, and hence gives rise to a “light stroke”.

Canning. Cylindrical cans of aluminum or tinplate are closed with a flat lid equipped with a ring pull-tab. A nonporous synthetic resin coating is applied to the inside of the can so as to prevent a chemical reaction occurring between the metal and the liquid. Pre-evacuation is only possible with stable tinplate cans. Otherwise cans are purged with an inert gas and carbon dioxide is blown under the lid of the can before sealing.

3.7. Beer Dispensing

After filling and transportation the beer should be stored at a temperature of 6 – 9 °C without movement. Bottled beer should be stored in the dark. For draft beer it is important to maintain the carbon dioxide level established in the brewery right up to the end of dispensing, and moreover to transfer the carbonation into the glass without loss. This is achieved by dispensing the beer with a carbon dioxide pressure higher than that of the saturation pressure in combination with a pressure compensator [46]. The clean, thin-walled glasses used exclusively for beer drinking should be rinsed with fresh, cold water just before the beer is tapped so as to equalize the temperature. Special care should be taken in the cleaning of beer glasses in order to remove traces of fat which might otherwise cause premature collapse of the sensitive foam, and also to maintain the good flavor of the beer.

4. Properties and Quality

Foam. More than any other beverage, beer is marked by its natural ability to form foam [46]. The release of carbon dioxide during tapping is responsible for foam stability whereas the sparkling of carbon dioxide bubbles in the tapped beer is responsible for head retention. Both foam stability and head retention, as well as the liveliness of the beer, depend largely on the carbon dioxide concentration. A foam-stabilizing effect is ascribed to colloids, such as glycoproteins, tannins, β-glucans, and isohumulone complexes; on the other hand, fatty acids, glycerides, and ethanol in an amount in excess of 7 – 8 wt % destabilize foam.

Color. The color of beer is first of all determined by the malt type. During the brewing process an increase in color is caused by temperature-dependent, nonenzymatic color reactions of the Maillard type; these occur especially during kilning of overmodified green malts treated with gibberellic acid. The beer is further darkened during boiling of mash portions and of wort, during the thermal stand of the wort, and also by oxidation reactions during wort preparation and bottling. A paling of the beer color occurs during fermentation, as well as later during filtration.

Buffering and Reduction Potential. The buffering substances present in the beer are chiefly weak acids and their salts, primary and secondary phosphates, and proteins. Depending on their chemical identities and concentrations, they can counteract the pH changes that occur during malting and during wort preparation, and they can attenuate the pH drop during fermentation. The reductones that are formed during malt, wort, and beer production are of great importance for the stability of the beer. By adding antioxidants, the reduction potential can be increased further.

Sensory Qualities. The odor and the taste of the beer, as well as the mouthfeel, are evaluated with regard to quality and intensity by taste experts. The temperature of the beer during tasting should be 6 – 10 °C. In a taste evaluation one differentiates between the *initial impression*, for which the aroma and palatefullness are mainly responsible, the *tingle* or *sparkle*, where the impression of freshness emerges (a function of carbon dioxide release and the organic acids present), and the *aftertaste*, where the quality and intensity of the bitterness are put to the test. Taste experts must be selected carefully and their sensory skills constantly practiced [22]. An international nomenclature with guidelines for the description of the impressions of beer flavor has been developed [47], [48]. Five basic types of flavor evaluation methods exist [47]:

> *Differentiation tests* are used to find out (a) if there is a difference between two or more samples, or (b) if two or more samples are identical to each other.

Descriptive tests are applied to "measure" the sensory quality of one or more samples using a vocabulary of flavor terms.
Scaling tests rank the intensity of one or more beer flavor attributes
Preference tests are performed by consumers to generate a simplified determination of beer flavor or to assess a preference.
Drinkablity tests indicate the relation between the brewing process and the consumer's judgement of drinkability.

Flavor Stability. After bottling the beer, numerous changes in its original properties occur. The changes in palatefullness and in liveliness are caused by an agglomeration of colloidal particles, which in turn is facilitated by oxygen and by changes in temperature. Palatefullness declines, and the bitterness becomes coarser and broader. Beer aroma undergoes a change due to numerous reactions contributing to an "aged" quality, which is also described as oxidation or bready flavor [49], [50].

5. Analysis

5.1. Analysis of Raw Materials

5.1.1. Water

In the food-processing industry, drinking and process water must comply with certain quality standards from the bacteriological, hygienic, and chemical point of view. The hardness of water can be determined by complexometric titration, which gives an idea of the concentrations of carbonate, hydrogen carbonate, and hydroxide in the water. The calcium and magnesium hardnesses must be determined to estimate the residual alkalinity.

5.1.2. Malt

The congress mashing method is used to study the parameters that determine the modification properties of the malt. In this method, ground malt is subjected to a simple infusion mash process whereby the temperature must first be maintained at 45 °C for half an hour, and is then gradually increased at the rate of 1 K/min to 70 °C; this temperature is finally maintained for one hour.

The fine – coarse difference is calculated from the dry matter extract of finely and coarsely ground malt. The degree of protein modification is the ratio of the amount of dissolved nitrogen compounds to the total nitrogen content of the malt (as determined by the Kjeldahl method).

Microscopic staining in combination with statistical methods is being increasingly used to estimate the uniformity of modification. Other qualitative characteristics are the enzyme potential, friability, and purity of the barley variety.

5.1.3. Hops and Hop Products

The bitter ingredients can be isolated by the fractionation of resins or by using HPLC. During the fractionation of resins, hop ingredients are distributed between an acidic aqueous methanol phase and diethyl ether. The bitter ingredients, which are extracted into the ether phase, are classified according to their solubility in methanol and hexane into total resins, soft resins, and hard resins. Depending on their ability to form lead salts the soft resins are further classified into α- and β-fractions. Using the HPLC method each of the homologues of the α- and β-acids and their oxidation products can be separated [3].

5.2. Brewhouse Control

The yield of the as-is extract (air-dried) from the congress mashing method is a measure of the yield achieved in the brewhouse during the large-scale preparation of the wort. For this the extract of original wort and the quantity of the cast wort must be determined.

5.3. Wort

The chemical composition of the wort is also of interest; amylolytic degradation and saccharification of starch components should be advanced in such a manner that only very few compounds that are stainable with iodine remain in the wort . When turbid lautering occurs, further stainable starch particles are solubilized during wort boiling. Furthermore, turbid lautering also leads to

an increase in the long-chain fatty acids content in the wort.

The gum content of the wort (determined by precipitation with salts, or precipitating agents or after isolation, using gel permeation chromatography) can only slightly be influenced by the choice of the mashing procedure. It depends mainly on the initial gum content of the barley and on the modification of the malt.

The degree of protein degradation during mashing can be measured from the contents of amino acids and dipeptides in the wort and from a fractionation of the nitrogen compounds. The amount of coagulable nitrogen and the quantity of precipitated trub give information on the intensity of wort boiling. An idea of the various Maillard reactions, which take place during boiling, is given by the concentrations of the aroma compounds in the wort (carbonyl compounds, alcohols, Maillard products), which are quantitatively determined using GC.

The isohumulones are the most important bitter compounds in the finished wort and in the beer. They are extracted from the acidified sample with isooctane, and their concentration is determined by spectrophotometry. The extinction of the isooctane extract at 275 nm is multiplied by 50; the value thus obtained is taken as the bitterness (in EBC bitter units) according to the standards of the European Brewery Convention (EBC) and the American Society of Brewing Chemists (ASBC).

The attenuation limit is determined by fermenting a wort sample with an excess of yeast at room temperature.

5.4. Fermentation

The most important parameters during fermentation are the decrease in extract, the pH, and the temperature. The practice of keeping a record of the yeast cell concentration from pitching to the end of fermentation using the electronic Coulter counter or microscopic count methods (Thoma chamber) is gaining increasing importance.

Gas chromatographic methods have proved to be useful in estimating the aroma components. Whereas the more concentrated fermentation byproducts in the ppm-range can be used to judge the fermentation and eventually occurring organoleptic deviations, certain trace components in the ppb-range can serve as indicator substances for the identification of the process technology. The acetoin concentration is used as an indicator for the vitality of the yeast. After a relatively simple work-up method (steam distillation of the sample followed by extraction of the distillate with dichloromethane) about 60 individual components can be detected in a single run using modern GC and high-resolution capillary columns. Along with selective detectors (sulfur detector, nitrogen detector), this figure can be increased considerably.

5.5. Microbiological Process Monitoring

In a brewery the main contaminants are lactobacilli, *Pediococcus cerevisiae*, and "wild yeasts" (*Saccharomyces* species). Gram-negative bacteria (genus *Megasphaera* or *Pectinatus*) may also be dangerous [24].

Direct viewing under the microscope, biological shelf life tests, and enrichment methods are used. Shelf life tests show microbiological stability of the beer during a storage at 27 °C. The yeast sediments of samples taken from the fermentation and the storage cellars are observed under the microscope after a period of about 20 days. In the case of filtered samples, trace infections are detected by enrichment on a solid or liquid nutrient medium. By using membrane filtration techniques, the microorganisms on the filter are incubated anaerobically under optimal growth conditions on solid agar. The samples can be evaluated after an incubation period of two days with aerobic incubation (wort agar) and five days (NBB agar; NBB = culture medium for bacteria harmful to beer) with anaerobic incubation. Color changes in liquid NBB broth or solid NBB agar indicate the presence of harmful bacteria.

The number of biological controls and the choice of culture media depend on such factors as selectivity, specificity, and output of the brewery. As a routine procedure, the pitching yeast as well as bottled beer stored for a specific period should be analyzed.

5.6. Beer

The extract of original wort of the finished product can be calculated from the density and re-

Table 8. Beer production, exports, imports, consumption and per capita consumption in 2000 (countries are ranked in the order of decreasing overall consumption

	Production, 10^6 hL/a	Exports, 10^6 hL/a	Imports, 10^6 hL/a	Consumption, 10^6 hL/a	Per capita consumption, L/a
1. United States	213.1	2.7	**23.3**	**235.9**	86.1
2. China	**223.1**	0.8	0.8	223.1	17.2
3. Germany	113.3	11.1	3.2	105.4	128.0
4. Brazil	86.6	0.5	0.1	86.2	51.2
5. Japan	71.1	0.5	0.9	71.5	56.2
6. United Kingdom	55.2	3.7	5.5	57.0	95.9
7. Mexico	57.8	7.9	0.5	50.4	48.9
8. Russia	48.8	0.3	0.4	48.9	33.6
9. Spain	26.4	0.6	2.7	28.5	72.0
10. South Africa	24.2	0.3	0.5	24.4	55.5
16. Czech Republic	18.0	1.7	0.2	16.5	**160.6**
21. Netherlands	25.1	**12.9**	0.9	13.1	82.8
27. Austria	8.8	0.5	0.5	8.9	110.5
34. Denmark	8.1	2.6	0.1	5.6	106.1
35. Ireland	8.0	3.1	0.7	5.6	150.5
Americas				473.8	57.3
Europe				463.6	53.2
Asia/Pacific				384.9	11.2
Africa				64.1	6.4
World total				**1386.3**	**22.6**

fractive index of the decarbonated beer or from the densities of beer distillate and the distillation residue.

The methods described in Section 5.3 are also used for beer. The carbon dioxide content is determined by measuring the total pressure of beer after vigorous shaking at a specific temperature.

The physicochemical stability of the beer is estimated from an accelerated aging process. In this method filled beer bottles are kept alternately at 40 °C (or at 60 °C for stabilized beer) and at 0 °C until an increase of turbidity by 2 EBC formazine units occurs after cooling. By multiplying the obtained stability period – expressed as 40 °C- (or 60 °C-)warm days – with a factor that is specific to each brewery, one can calculate the period of time during which the beer would not become turbid under normal storage conditions.

The carbon dioxide content of the beer determines its foaming capacity. The head retention is dependent on the composition of the beer. The methods of foam measurement are classified according to the way the foam is generated: (1) free fall, (2) shaking, (3) bubbling of a gas, and (4) catalytic release of carbon dioxide.

Color can either be measured by visual comparison under defined conditions or spectrophotometrically by measuring the extinction of the sample at 430 nm and that at 700 nm; this difference, multiplied by 25.5, is defined as the color in EBC color units.

5.7. Legally Required Controls

In order to comply with the requirements of the food laws, the extract of original wort or the alcohol content must be within certain limits. Further, the filling volume in the bottles must be within required tolerances. Nevertheless, it is still indispensable that the brewmaster is responsible for determining by taste test whether the beer is ready for filling and before it leaves the brewery.

In the last years breweries introduced integrated management systems and had them validated by a external accredited and recognized certification body. The following systems have been introduced [51], [52]:

Quality Management Systems in accordance with DIN EN ISO 9001 (design, development, production, installation, and maintenance),
Environmental Management System in accordance with DIN EN ISO14001,
Working Security System, and

Table 9. Top world breweries

	Production, 10^6 hL/a	Regional ranking	Leading brands	
			Names	Output, 10^6 hL/a
1. Anheuser – Busch (USA)	124.1	Americas no. 1	Budweiser	51.9
			Bud	39.9
			Busch	17.1
2. Heineken (Netherlands)	74.8	Europe no. 1	Heineken	21.6
		Africa no. 2	Amstel	10.8
3. Interbrew (Belgium)	64.1	Europe no. 3	Stella Artois	7.0
4. Brahma/AmBev (Brazil)	62.4	Americas no. 2	Skol	25.4
			Brahma Chopp	20.9
			Antarctica	20.9
5. South African	61.3	Africa no. 1	Castle Lager	14.4
		Europe no. 5		
6. Carlsberg (Denmark)	60.9	Europe no. 2	Carlsberg	10.0
			Baltika	8.0
7. Miller (USA)	50.6	Americas no. 3	Miller LITE	19.6
			Miller Genuine Draft	9.1
8. Kirin (Japan)	38.5	Asia/Pacific no. 1	Kirin Lager	11.1
			Tanrei – Nama	8.4
9. S & N-Kronenbourg (France)	36.8	Europe no. 4	Kronenbourg	6.6
10. Modelo (Mexico)	36.6	Americas no. 4	Corona	24.1

Hazard Analysis Critical Control Points (hygiene in food producing companies)

6. Economic Importance

During the period of 1991 – 2000 beer output increased by more than 20 % worldwide. Despite stagnation in several large beer consuming nations the total market size of 1386×10^6 hL reached a new peak in 2000. The per capita consumption increased only marginally, to reach 22.6 L in 2000 [27].

The most popular product is the pale, bottom-fermented lager of typically 4.5 – 5.5 vol % alcohol. Worldwide 65 % of the beer is marketed in bottles, 22 % in cans and 13 % in kegs or barrels (draught beer).

Because the cereals produced in the individual countries are not exclusively used for malting and brewing, it is not possible to list which and how much of the different cerials are used. The world crop of hops in 2001 was 97 730 t (62 % bitter hops, 38 % aroma hops). The content of α-acids was about 8 % on average, so 7928 t of α-acids were harvested. Production and consumption figures are given in Table 8, Table 9 lists the top brewers of the world and their leading brands.

A survey of the productivity and quantities of energy and water consumed in breweries of different sizes is given in Table 10.

Table 10. Productivity and specific energy and water comsumption as an function of brewery size

	Breweries with an annual output of		
	$> 10^6$ hL	$> 10^5$ hL	$> 10^4$ hL
Productivity, hL per employee	7000	3000	1000
Current consumption, kWh/hL	8	10	13
Heat consumption, kWh/hL	30	40	50
Water consumption, hL/hL	4	6	8

Taxation. The varieties of beer are classified into various tax groups on the basis of their extract of original wort or alcohol content. The groups differ from country to country. There are different means of tax collection: (1) raw-material taxation, which is based on weight or volume of the brewing materials used; (2) intermediate product taxation, which is assessed according to the volume of the wort, and (3) finished product taxation, which is levied on sales beer.

7. Physiology and Toxicology

According to standard definitions, beer is classified as a food. In addition, it acts as a stimulant. The calorific content of beer as a nutrient may be calculated from its concentrations of protein, carbohydrate, and alcohol, as well as organic acids. A rough estimate in kJ/L can be obtained by multiplying the extract of original wort (in wt %) by a factor of 150. Hence beers with 10 – 14 % extract of original wort have an approximate calorific value of 1500 – 2100 kJ/L. From a nutritional point of view, the amino acids (on an average 140 mg/L) and vitamins (20 – 25 mg/L) are especially valuable. One liter of beer contains about 40 µg thiamine, 400 µg riboflavin, 7500 µg niacin, 650 µg pyridoxine, 1500 µg pantothenic acid, and 800 µg folic acid.

The consumption value of beer may be derived from its ingredients. Thirst quenching is accomplished by its high water content and its mineral concentration (total around 1000 mg/L, consisting of: sodium 20 – 30 mg/L, potassium 500 mg/L, calcium 30 mg/L, phosphorus 300 mg/L, and magnesium 100 mg/L). Carbon dioxide (4 – 8 g/L) and organic acids (up to 600 mg/L) have a relaxing, calming and, at the same time, stimulating effect. The dietetic effect of the beer is based mainly on its low sodium concentration. Because it also is free of fats, it will promote urination.

Pathogenic and toxic bacteria cannot survive in beer because of the presence of alcohol, carbon dioxide, bitter substances, and the low pH. A proper balance between alcohol and assimilable carbohydrates, proteins, phosphates, and vitamins results in a lower increase in blood alcohol levels by comparison with other alcoholic beverages. The toxicology of ethanol is treated elsewhere (→ Ethanol, Chap. 11.).

Benefits of Moderate Beer Consumption. Beer as a wholesome beverage has formed a staple part of human diet for thousands of years. When consumed moderately and regularly beer is not only enjoyable to drink, but also beneficial to health. A daily consumption of 60 – 80 g alcohol (corresponding to 1.5 to 2 L of beer) by a healthy, adult male, or 40 – 60 g alcohol by an adult woman is not harmful. These guidelines vary according to body weight, age and nutritional habits as well as physiological condition. Rules for reasonable beer consumption are given in [53], [54].

It has been shown in many studies throughout the world that moderate consumption of beer – in contrast to heavy drinking or abstention – is protective against cardiovascular diseases such as heart attack and some forms of stroke [55]. One explanation for this effect is that beer increases the level of "good HDL cholesterol" in the blood [56], another explanation is that alcohol reduces blood coagulation [57]. Beer is reported to protect against gallstone formation [58], diabetes [59], and even osteoporosis [60]. Hops contain active compounds which prevent calcium depletion of bones. Beer drinkers are protected against *Heliobacter pylori* bacteria [61], which cause stomach ulcers and may increase the risk of stomach cancer. Beer is also a source of soluble fiber (5 – 10 mg/L), originating from the cell walls of malted barley. A liter of beer contains on average 20 % of the recommended daily intake of fiber. Fibers do not only support a healthy bowel function, they also retard the digestion and adsorption of food, and lower cholesterol levels, which may help to reduce the risk of heart diseases [62].

Beer is also a source of antioxidants (200 – 600 mg phenolic carboxylic acids, flavonoids and tannoids per liter of beer), which show a cancerostatic activity [63–65]. Additional research has shown that the antioxidants present in beer are more readily available to the body than those from solid foods [66]. Furthermore, beer contains a large amount of dietetically valuable silicon as orthosilicate, its biologically available form. The silicon content of the human body has a direct effect on the mineralization of the bones and the density of the bone marrow. Beer contains 10 – 40 mg SiO_2 per liter, which is absorbed to more than 50 % by the human body. Thus beer is one of the most important dietetic silicon sources.

Harmful Substances. Numerous nutritional regulations require the amount of additives in beer to be kept to a minimum.

Heavy metals derived from raw materials are mostly removed during the malting and brewing process; beer is one of the beverages lowest in heavy metals (lead, 1 – 3 µg/kg; cadmium, 0.1 – 0.5 µg/kg; mercury, always less than 0.1 µg/kg). *Mycotoxins* and *insecticides* have

not been found in beer until now. *Fungicide* sprays, especially dithiocarbamates, are used in hop growing. The residues contained in the hops are metabolized during wort boiling to ethylenethiourea and propylenethiourea, which are only partially removed in the further brewing process and remain in the beer. The necessary sprayings could be reduced considerably by breeding spore-resistant types as well as by an early warning system. The concentrations of residue can also be cut to less than 5 mg/kg by allowing a time lapse of several weeks between the last spraying and the actual harvesting. Various hop extraction procedures also can lower the concentration of dithiocarbamates.

The World Health Organization (WHO) recommends a limit of *nitrates* of 50 mg per liter of drinking water. The nitrate in the wort originates mainly from water and hops. Nitrate is reduced to nitrite, which is poisonous to yeast. However, its concentration is much below the threshold considered critical by the WHO. The concentration of nitrate in beer corresponds to that of the brewing water. Nitrate can be most effectively removed from the water by anion exchange.

The concentration of highly carcinogenic *N*-nitrosodimethylamines is virtually negligible, because kilns are fired indirectly nowadays. With average values of less than 0.5 µg/kg, beer produced today is no longer significant as a potential source of nitrosoamines.

The byproducts formed during fermentation, such as higher alcohols, esters, and aldehydes, are of more importance for the aroma of the beer than for its wholesomeness. No connection can be found between the level of higher alcohols, fats, and aldehydes and the incidence of hangover.

Sulfur dioxide is a true fermentation byproduct. Bottom-fermented lager or pilsener contain 0 – 10 mg/kg, top-fermented beers 0 – 5 mg/kg, whereas strong beers may contain more than 10 mg/kg of the gas. The source cannot be traced back to the sulfur content of the malt, the hops, or the brewing water.

Histamine, a metabolite of histidine, occurs in beer in amounts of under 0.5 mg/kg, and has no toxic effect at this level. The *nucleic acid components* of beer originate from the raw materials. The yeast needs adenine and guanine for its growth and removes part of these substances during fermentation. The purines of the finished beer are broken down to uric acid in the human body, which can cause gout if the amount exceeds the solubility in blood. Bottom-fermented lager beers contain 70 – 130 mg purine per kilogram; in wheat beers the figure is around 80 mg/kg [67].

8. References

General References

1. B. E. Briggs: Malts and Malting, Chapman & Hall, Andover 1997.
2. H. J. Barth, C. Klinke, C. Schmidt: *Der große Hopfenatlas/The great hopatlas,* Verlag Hans Carl, Nürnberg 1994.
3. M. Verzele, D. De Keukeleire: *Chemistry and Analysis of Hop an Beer Bitter Acids,* Elsevier Science Publishers B.V., Amsterdam 1991.
4. J. L. Benitez et al.: *Hops and Hop Products,* Verlag Carl Hanser, Nürnberg 1997.
5. G. Reed, T. Nagodawithana: *Yeast Technology,* 2nd ed., Van Nostrand Reinhold, New York 1991.
6. C. P. Kurtzman, J. W. Fell: *The Yeasts,* 4th ed., Elsevier Science Publishers B.V., Amsterdam 1998.
7. L. Narziß: *Abriß der Bierbrauerei,* 6th ed., F. Enke Verlag, Stuttgart 1995.
8. L. Narziß: *Technologie der Malzbereitung,* 7th ed., F. Enke Verlag, Stuttgart 1999.
9. L. Narziß: *Technologie der Würzebereitung,* 7th ed., F. Enke Verlag, Stuttgart 1992.
10. K. U. Heyse (ed.): *Handbuch der Brauereipraxis,* 3rd ed., Verlag Hans Carl, Nürnberg 1994.
11. K. U. Heyse (ed.): *Praxishandbuch der Brauerei,* B. Behr's Verlag GmbH & Co., Hamburg 2000 and supplements.
12. W. Kunze: *Technologie Brauer und Mälzer,* 8th ed., Verlag der VLB Berlin, Berlin 1998.
13. W. Kunze: *Technology Brewing and Malting,* 2nd ed., Verlag der VLB Berlin 1999.
14. T. Wainright: *Basic Brewing Science,* Wainright, Reigate 1998.
15. European Brewery Convention Monographs: Separations Process 1990, Packaging 1990, Wort Boiling and Clarification 1992, Waste Reduction in Brewery Operations 1992, Instrumentation and Measurement 1992, Process Hygiene 1994, Hops 1994, Malting Technology 1994, Immobilized Yeast Applications in the Brewing Industry 1995, Draught Beer ∗Packaging∗ Dispense 1996,

Quality Issues & HACCP 1997, Beer Foam Quality 1998, Yeast Physiology – a new Era of Opportunity 1999, Assuring Product Safety in the Brewing Industry 2000. Verlag Hans Carl, Nürnberg.
16. European Brewery Convention Proceedings: 23rd Congress, Lisbon 1991, 24th Congress, Oslo 1993, 25th Congress, Brussels 1995, 26th Congress, Maastricht 1997, 27th Congress, Cannes 1999, 28th Congress, Budapest 2001, University Press, Oxford.
17. N. Buchner: *Verpackung von Lebensmitteln – Lebensmitteltechnologische, verpackungstechnische und mikrobiologische Grundlagen,* Springer Verlag, Berlin, Heidelberg 1999.
18. O. G. Piringer, A. L. Baner: *Plastic Packaging Materials for Food,* Wiley-VCH, Weinheim 2000.
19. Deutscher Brauerbund: Leitfaden "Schankanlagen" (1997).
20. P. Dilly: *Qualitätssicherung in der Brau- und Getränkewirtschaft,* B. Behr's Verlag, Hamburg 1990.
21. P. Dilly: *Handbuch Umweltaudit,* B. Behr's Verlag, Hamburg 1996.
22. *Analytica EBC,* 5th ed. and supplements, Verlag Hans Carl, Nürnberg 1998.
23. H. Pfenninger (ed.): *Brautechnische Analysenmethoden, Methodensammlung der Mitteleuropäischen Brautechnischen Analysenkommission (MEBAK),* 3rd. ed, **vol. 1:** Rohstoffe, 1997, 3rd ed., **vol. 2:** Würze und Bier, 1993, 2nd ed., **vol. 3:** Instrumentelle Analytik und Mikrobiologie, 1996, 2nd ed., **vol. 4:** Technische Hilfsstoffe, Bedarfsgegenstände und Ausschank, 1998, **vol. 5: Gebinde und Produktausstattungsmittel, MEBAK,** Freising-Weihenstephan 1999.
24. W. Back: *Handbuch und Farbatlas der Getränkebiologie Teil 1: Kultivierung und Methoden Brauerei/Mälzerei,* Verlag Hans Carl, Nürnberg 1994.
25. M. Knab, J. Schulters: *Bierologie – Einführung in die Geschmackswelt des Bieres,* Verlag Hans Carl, Nürnberg 2000.
26. A. Piendl, K.-U. Heyse, H. Kracun, M. Unkel: *Bier Erleben,* Verlag Hans Carl, Nürnberg 2001.
27. World Beer Report, Plato Logic Ltd., October 2001 edition.
28. Brauerei-Datenbanken. Die Infothek für die Getränkewirtschaft Website http://www.braudatenbanken.de

Specific References
29. W. Back, L. Narziß: "Malt parameters and beer quality", *Brauwelt International* **15** (1997) 29 – 35.
30. K. Wackerbauer, C. Zufall: "From malthouse to brewery – malt specifications and analysis", *Brauwelt International* **15** (1997) 22 – 29.
31. A. Forster, B. Beck, R. Schmidt: "Hopfenpolyphenole – mehr als nur Trübungsbilder in Bier", *Hopfenrundschau International* (1999) 68 – 74.
32. A. Forster, U. Balzer, W. Mitter: "Commercial brewing tests with different hop extracts", *Brauwelt International* **14** (1996) 324 – 326.
33. D. Kaltner, B. Thum, C. Forster, W. Back: "Hops – Investigations into technological and flavour effects in beer", *Brauwelt International* **19** (2001) 40 – 45.
34. A. Forster: "Significance of crop year in qualitative assessment of hop products", *Brauwelt International* **19** (2001) 32 – 37.
35. S. Schildbach: "Evaluation a drinking water analysis", *Brauwelt International* **19** (2001) 401 – 407.
36. S. Donhauser: "Characterisation of yeast species and strains", *Brauwelt International* **15** (1997) 52 – 58.
37. R. Schildbach: "Raw materials supply for maltings in the European Union", *Brauwelt International* **17** (1999) 41 – 48.
38. A. Schwill-Miedaner, H. Miedaner: "Würzekochung – heutiger Stand der Technologie und Technik", *Brauwelt* **141** (2001) 670 – 673.
39. H. M. Eßlinger: "Fermenters and storage tanks", *Brauwelt International* **13** (1995) 50 – 154.
40. F. Hebmüller: Einflußfaktoren auf die Kieselgurfiltration von Bier, PhD Thesis TU Freiberg 2002.
41. R. Schildbach, W. Ritter: *Brauerei-Kieselgur, Wissenschaftsförderung der Deutschen Brauwirtschaft e.V.,* Verlag der VLB, Berlin 1995, pp. 1 – 43.
42. H. M. Anger: "Assuring nonbiological stability of beer as an important factor for guaranteeing minimum shelf-life", *Brauwelt International* **14** (1996) 142 – 150.
43. L. Narziß: "Global brewing technology – a look over the fence", *Brauwelt International* **15** (1997) 16 – 21.
44. W. Back, C. Diener, B. Sacher: "Hefeweizenbier – taste spectrum and technology", *Brauwelt International* **18** (2000) 112 – 119.

45. A. Perrier-Robert, C. Fontaine: *Belgium by Beer – Beer by Belgium,* Edition Schortgen, Esch/Alzette 1996.
46. G. Hauser: "Beverage Quality with Keg Dispensing", *Brewing and Beverage Industry International* **2** (1995) 17 – 19.
47. W. J. Simpson, E. Canteranne: "Sensory analysis in modern brewery operations", *Brauwelt International* **19** (2001) 280 – 286.
48. L. Narziß: "Beer taste and how it is influenced by raw materials and technological factors", *Brauwelt International* **16** (1998) 14 – 25, 46.
49. W. Back, C. Forster, M. Krottenthaler, J. Lehmann, B. Sacher, B. Thum: "New research findings on improving taste stability", *Brauwelt International* **17** (1999) 394 – 405.
50. K. Wackerbauer, R. Hardt: "Radical reactions and flavour stability of beer", *Brauwelt International* **15** (1997) 320 – 327.
51. F. Jacob, S. Donhauser, E. Geiger: Quality assurance aimed at certification", *Brauwelt International* **13** (1995) 56 – 64.
52. F. Jakob: "Quality assurance in breweries as a function of plant size", *Brauwelt International* **18** (2000) 486 – 488.
53. A. Piendl: "Definitions of moderate alcohol consumption", *Brauwelt International* **17** (1999) 49 – 55.
54. A. Piendl: "Physiological effects of alcohol consumption", *Brauwelt International* **15** (1997) 310 – 319.
55. B. Fagrell et al.: "The effects of light to moderate drinking oncardiovascular diseases", *J. Intern. Med.* **246** (1999) 331 – 340.
56. M. V. McConnel et al.: "Effects of a single daily alcoholic beverage on lipid and haemostatic markers of cardiovascular risk", *Am. J. Cardiol.* **80** (1997) 1226 – 1228.
57. S. Gorinsten et al.: "Moderate beer consumption and the blood coagulation in patients with coronary heart disease", *J. Intern. Med.* **241** (1997) 47 – 51.
58. M. F. Leitzmann et al.: "Prospective study of alcohol consumption patterns in relation to symptomatic gallstone disease in men", *J. Clin. Exp. Res.* **23** (1999) 835 – 841.
59. E. B. Rimm, M. J. Stampfer, G. A. Colditz, W. C. Willett: "Prospective study of cigarette smoking, alcohol use and the risk of diabetes in men", *Br. Med. J.* **310** (1995) 555 – 559.
60. P. Charles, K. Laitinen, A. Kardinaal: "Alcohol and Bone" in *Health Issues Related to Alcohol Consumption,* 2nd ed., Ian Macdonald, Blackwell Science Ltd., 1999.
61. H. Brenner, D. Rothenbacher, G. Bode, G. Adler: "Relation of smoking and alcohol and coffee consumption to active heliobacter pylori infection: Cross sectional study", *Br. Med. J.* **315** (1997) 1489 – 1492.
62. S. Bell et al.: "Effect of β-Glucan from oats and yeast on serum lipids", *Crit. Rev. Food Sci. Nutr.* **39** (1999) 189 – 202.
63. T. Tagashira, M. Watanabe, N. Uematsu: "Antioxidative activity of hop bitter acids and their analogues", *Biosci. Biotech. Biochem.* **59** (1995) 740.
64. C. I. Miranda et al.: "Antiproliferative and cytotoxic effects of xanthohumol in human prostate cancer cells", *Toxicologist* **48** (1999) 348.
65. S. Arimoto-Kobayashi et al.: "Inhibitory effects of beer and other alcoholic beverages on mutagenesis and DNA adduct formation inducted by several carcinogens", *J. Agric. Food Chem.* **47** (1999) 221 – 230.
66. L. Bourne, G. Paganga, D. Baxter, P. Hughes, C. Rice-Evans "Absorption of ferulic acid from low-alcohol beer", *Free Radical Research* **32** (2000) 273 – 280.
67. E. Geiger, D. Wagner, F. Briem, S. Englert: Biogene amines in malt and beer preparation", *Brauwelt International* **14** (1996) 328 – 330.

Bentonite → Clays
Benzal Chloride → Chlorinated Hydrocarbons

Benzaldehyde

FRIEDRICH BRÜHNE, Bayer AG, Krefeld-Uerdingen, Federal Republic of Germany (Chaps. 1 – 9)

ELAINE WRIGHT, General Motors Research Laboratories, Warren, Michigan 48090, United States (Chap. 10)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Hydrolysis of Benzal Chloride 3
4.2. Oxidation of Toluene 5
4.3. Other Production Processes 6
5. Quality Specifications and Test Methods 6
6. Storage, Transportation, and Safety Regulations 7
7. Uses 7
8. Derivatives 7
8.1. Chlorobenzaldehydes 7
8.2. Nitrobenzaldehydes 8
8.3. Hydroxybenzaldehydes 9
8.4. Aminobenzaldehydes 10
8.5. Methylbenzaldehydes 10
9. Economic Aspects 11
10. Toxicology 11
11. References 11

1. Introduction

Benzaldehyde is the simplest and industrially the most important aromatic aldehyde. It exists in nature, occurring in combined and uncombined forms in many plants. The best known natural source of benzaldehyde is amygdalin, in which it exists in a combined form as a glycoside and which is present in bitter almonds. The odor of bitter almonds arises from small amounts of free benzaldehyde formed by hydrolysis of the amygdalin. Owing to its occurrence in bitter almonds, the aldehyde was formerly referred to as "bitter almond oil." Benzaldehyde is also the main constituent of the essential oils obtained by pressing the kernels of peaches, cherries, apricots, and other fruits.

Benzaldehyde
[100–52–7]

In 1818 – 1819 VOGEL and MATRÈS reported for the first time, and independently of one another, that, in addition to hydrocyanic acid, a volatile oil could be obtained from bitter almonds. Later, in 1832, this oil was investigated by WÖHLER and LIEBIG; they determined its composition and recognized its relationship to benzoic acid and benzoyl chloride.

2. Physical Properties

Benzaldehyde, C_7H_6O, M_r 106.13, is a colorless and highly refractive liquid with an odor of bitter almonds. Like many other aromatic aldehydes, it is volatile in steam, *bp* 179 °C at 101.3 kPa, *mp* −56 °C, refractive index n_D^{20} 1.5450, density ϱ at 0 °C 1.063 g/cm^3, at 20 °C 1.046 g/cm^3, at 50 °C 1.018 g/cm^3, specific heat capacity c_p 1.676 J g^{-1} K^{-1} at 25 °C, heat of evaporation 371.0 J/g at 179 °C, standard heat of combustion 33.19 kJ/g, flash point 64.5 °C, autoignition temperature 190 °C, lower explosive limit 1.4 vol %. Vapor pressure versus temperature data for benzaldehyde are given in Table 1.

Benzaldehyde is miscible with many organic solvents; at 25 °C it is miscible with concentrated sulfuric acid, liquid carbon dioxide, liquid ammonia, methylamine, and diethylamine. The compositions of several binary azeotropic mixtures are given in Table 2. For further azeotropes, see [3]. Vapor – liquid equilibria data for binary mixtures of benzaldehyde with water or organic solvents will be found in [2]. One hundred milliliters of water dissolves 0.40 mL of benzaldehyde at 25 °C or 0.96 mL at 60 °C. Dynamic viscosity η 1.40×10^{-3} Pa · s at 25 °C, 1.11×10^{-3} Pa · s at 40 °C, surface tension σ 40.04×10^{-3} N/m at 20 °C, dipole moment μ (measured in liquid benzene) 2.92 D (9.74×10^{-30} C m), dielectric constant ε_r 17.7 at 25 °C.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_463

Table 1. Vapor pressure vs. temperature [2]

t, °C	40	60	80	100	120	140	160	180
p, kPa	0.3	1.1	3.3	8.2	17.8	34.7	62.3	104.3

Table 2. Azeotropic mixtures of benzaldehyde [3]

Second component	*bp* at 101.3 kPa, °C	Benzaldehyde, wt %
3-Bromotoluene	179.0	92.0
Ethyl phenyl ether	169.8	12.0
Benzyl ethyl ether	177.5	92.0
2-Octanol	174.0	25.0
α-Terpinene	170.0	38.0
Diisoamyl ether	168.6	37.5

3. Chemical Properties

The chemical behavior of benzaldehyde and of its nuclear substitution derivatives corresponds largely to that of the aliphatic aldehydes (→ Aldehydes, Aliphatic and Araliphatic). But the reactivity of the carbonyl group of the aromatic aldehydes is somewhat less because the π electrons of this group are included in the resonance system of the aromatic ring. Thus benzaldehyde forms Schiff bases with amines, an oxime with hydroxylamine, a hydrazone with phenylhydrazine, and acetals with alcohols. It adds hydrogen cyanide, sodium bisulfite, and Grignard compounds. If benzaldehyde is reacted with hydrogen cyanide in the presence of ammonia and the aminonitrile formed as an intermediate is then saponified, DL-2-phenylglycine is formed (Strecker synthesis).

Benzaldehyde undergoes autoxidation in air. This process, the final product of which is benzoic acid, is accelerated by light and by the addition of peroxides or heavy-metal salts. It is retarded by antioxidants, such as phenolic compounds and diphenylamine. Nitric acid, chromium(VI) oxide, and other oxidizing agents also oxidize benzaldehyde to benzoic acid.

The reduction or hydrogenation of benzaldehyde leads to benzyl alcohol, dibenzyl ether, benzoin, 1,2-diphenylethane-1,2-diol, stilbene, toluene, methylcyclohexane, and other products, depending on the reaction conditions. Catalytic hydrogenation of benzaldehyde is used industrially to produce benzyl alcohol. Reduction with aluminum alcoholates (Meerwein-Ponndorf-Verley reduction) also gives benzyl alcohol. This reaction also permits the reduction of unsaturated aldehydes, such as cinnamaldehyde, to the corresponding alcohols, in which the olefinic double bonds are retained.

The reaction of benzaldehyde with ammonia and hydrogen in the presence of hydrogenation catalysts leads to benzylamine and is an important industrial process. The side-chain chlorination of benzaldehyde gives benzoyl chloride [4].

Like the aliphatic aldehydes, benzaldehyde participates in condensation reactions with a number of organic compounds possessing active hydrogen atoms. Some of these reactions are used for industrial purposes. The Claisen-Schmidt condensation of benzaldehyde with acetaldehyde in the presence of aqueous alkali leads to cinnamaldehyde.

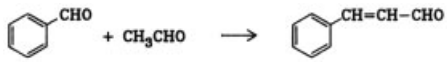

The Perkin condensation of benzaldehyde with acetic anhydride in the presence of sodium acetate or potassium acetate as a condensing agent is carried out industrially in the production of cinnamic acid (→ Acetic Anhydride).

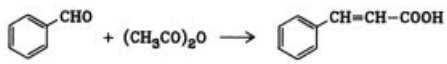

Cinnamic acid is also produced by Knoevenagel condensation of benzaldehyde with malonic acid in the presence of weakly basic catalysts, such as ammonia and amines.

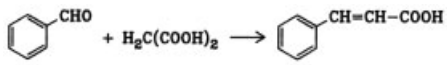

In the Reformatsky reaction, ethyl cinnamate is obtained from benzaldehyde and ethyl bromoacetate in the presence of activated zinc.

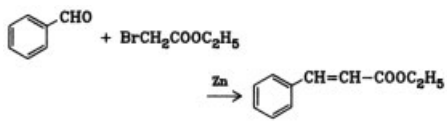

The Cannizzaro reaction is typical of benzaldehyde and other aldehydes that have no α-hydrogen atoms. If concentrated sodium or potassium hydroxide is present, this reaction results in disproportionation to benzoic acid and benzyl alcohol.

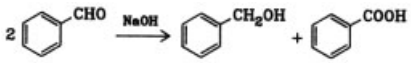

Sodium benzylate and aluminum benzylate catalyze the Claisen-Tishchenko condensation of benzaldehyde, which leads to benzyl benzoate.

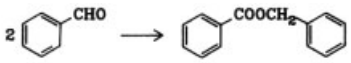

In comparison with their aliphatic analogues, benzaldehyde and other aromatic aldehydes have certain peculiar aspects in their reaction behavior. In the benzoin condensation, two molecules of benzaldehyde condense in the presence of cyanide to form the hydroxyketone benzoin. Thiazolium salts can be used as catalysts in place of cyanide [5].

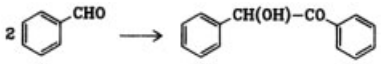

The reactivity of benzaldehyde towards ammonia also differs from that of aliphatic aldehydes. Instead of stopping at the aldehyde – ammonia adduct, the reaction continues until 1-phenyl-*N,N′*-bis(phenylmethylene)-methanedi-amine (hydrobenzamide) is obtained.

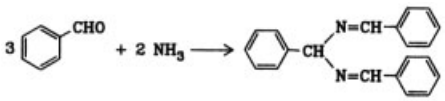

Benzyl alcohol and benzoic acid, but no copper(I) oxide, are formed when benzaldehyde is treated with Fehling's solution. Aromatic aldehydes do not show the polymerization or formation of cyclic compounds that is observed with aliphatic aldehydes. The condensation of benzaldehyde with phenols, aromatic amines, and benzene leads to triphenylmethane derivatives. This reaction is exploited industrially in the production of Malachite Green dyes.

Electrophilic substitution of the aromatic nucleus (chlorination, nitration, sulfonation) occurs preferentially in the meta position.

4. Production

Benzaldehyde is produced principally by the hydrolysis of benzal chloride or the partial oxidation of toluene. There are various other manufacturing processes, but at present they have no industrial importance.

4.1. Hydrolysis of Benzal Chloride

The hydrolysis of benzal chloride, which is readily obtainable by side-chain chlorination of toluene, is among the oldest industrial processes for the production of benzaldehyde. It can be carried out either in an alkaline or in an acidic medium.

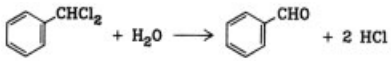

Hydrolysis under basic conditions can be carried out with calcium hydroxide, calcium carbonate, sodium hydrogencarbonate, or sodium carbonate. Excessive alkalinity increases the probability of side reactions. The preferred alkaline saponifying agent is sodium carbonate.

According to a fairly old process [6], benzal chloride is saponified with a small excess of 15 % sodium carbonate solution at 138 °C. The reactor is a carbon steel agitator vessel lined internally with Heresite (a phenolic resin) and having an agitator made of a copper – silicon alloy. The cooler, separator, and pumps are made of Karbate (a graphitic material). On completion of the saponification, the benzaldehyde is isolated from the reaction mixture by steam distillation and distilled in a stainless steel column at 2 – 3 kPa (20 – 30 mbar). The chlorine content of the distilled product is less than 0.01 %.

Instead of pure benzal chloride, a mixture of benzal chloride and benzotrichloride, as obtained industrially in the side-chain chlorination of toluene (→ Chlorinated Hydrocarbons), can be hydrolyzed. Under the reaction conditions, benzotrichloride is saponified to sodium benzoate, which remains in the aqueous alkaline

solution. It is converted separately to benzoic acid.

In a new continuous process [7], benzal chloride and the alkaline hydrolyzing agent are reacted as countercurrents in a flow reactor in an unreactive organic solvent (e.g., toluene, xylene). A temperature of 125 – 145 °C and a pressure of 1200 – 1800 kPa (12 – 18 bar) are maintained. In the extraction zone the dissolved aldehyde is extracted from the aqueous alkaline phase with an unreactive organic solvent flowing in the opposite direction. Water, likewise as a countercurrent, washes the solution of the crude benzaldehyde in the washing zone. The flow reactor, extraction zone, and washing zone are integrated in a special apparatus. The wash water and extract are returned to the hydrolysis zone in order to recover bezaldehyde (Fig. 1).

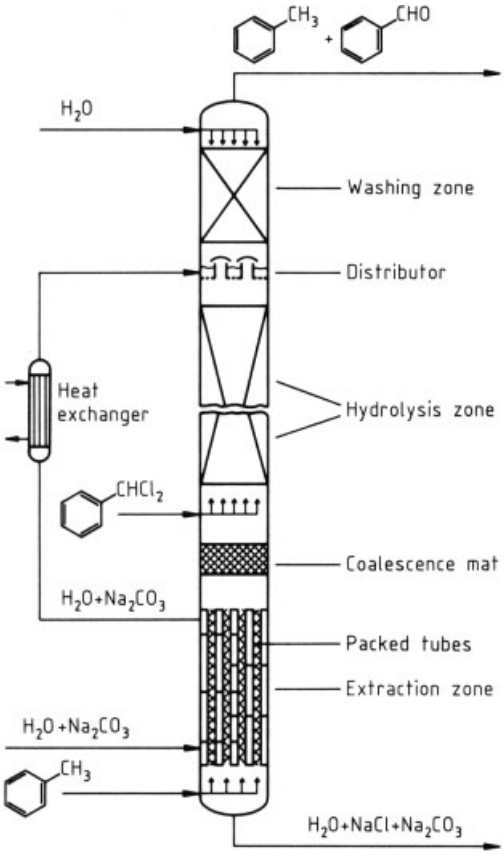

Figure 1. Hydrolysis of benzal chloride to benzaldehyde in a flow-through reactor [7]

Benzal chloride can also be converted into benzaldehyde by boiling with aqueous solutionsof hexamethylenetetramine [8]. Because benzyl chloride also reacts with hexamethylenetetramine to give benzaldehyde (Sommelet reaction), industrial mixtures of benzyl chloride and benzal chloride can be used to form benzaldehyde in this way.

The acid hydrolysis of benzal chloride is carried out in the presence of acids and with metal salts as catalysts. It gives very high yields of benzaldehyde (more than 90 %). Hydrogen chloride is formed simultaneously and is usually passed through an absorption unit to be recovered as concentrated hydrochloric acid.

The hydrolysis of benzal chloride was at one time commonly carried out in the presence of concentrated sulfuric acid. This process has the disadvantage that large amounts of dilute sulfuric acid are formed as a waste product. It can be used for the hydrolysis of certain substituted benzaldehydes that are difficult to saponify in other ways [9], [10]. Benzal chloride is also saponified in the presence of phosphoric acid or sulfonic acids [11], hydrochloric acid [12], and formic acid [13]. The reaction of benzal chloride with organic acids, such as acetic acid, in the presence of a tin(II) or tin(IV) chloride catalyst [14], gives benzaldehyde and the corresponding acid chloride.

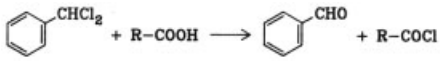

R = alkyl, aryl, or cycloalkyl

The hydrolysis of benzal chloride is catalyzed by metal salts, preferably those of iron or zinc. Water should not accumulate in the reaction mixture during the reaction because it would reduce the activity of the catalyst [15], [16].

Two parts by weight of anhydrous zinc chloride are added to 700 parts by weight of distilled benzal chloride, stirred, and heated to 105 – 110 °C. The mixture is maintained at 100 – 120 °C, while 85 parts by weight of water are added at the rate at which water is consumed. After all the water has been added, the reaction temperature is maintained for one hour to ensure removal of the liberated hydrogen chloride [16]. The yield of crude benzaldehyde is 95 %, and the purity of the undistilled product is greater than 96.5 %.

The catalyst is not deactivated by a momentary excess of water if insoluble zinc catalysts, such as zinc phosphate or zinc laurate, are used [17]. Tin chlorides [18] and copper(II) chloride [19] are also catalytic. The hydrolysis of benzal chloride can be carried out batchwise or continuously, e.g., in a cascade consisting of several reactors.

A new continuous process [20] hydrolyzes benzal chloride in the vapor phase at temperatures of 100 – 300 °C. The reaction is catalyzed by activated carbon which has been treated with an acid, such as sulfuric acid, or impregnated with a metal chloride, such as iron(III) chloride, or with a sulfate, such as copper(II) sulfate. The benzaldehyde yield is 97 %. This process is particularly suitable for the hydrolysis of trifluoromethyl-substituted benzal chlorides, which are otherwise difficult to convert into the corresponding benzaldehydes. In a similar process, benzal chloride in the vapor phase at 300 °C is hydrolyzed to benzaldehyde using a silicon dioxide or aluminum oxide catalyst [21].

4.2. Oxidation of Toluene

The partial oxidation of toluene with oxygen to give benzaldehyde can be carried out in either the gas phase or the liquid phase. Benzaldehyde itself is easily further oxidized to benzoic acid and other products. Conditions must therefore be carefully chosen to favor only partial oxidation.

In the gas phase, the oxidation is carried out by passing toluene vapor, together with oxygen in a gaseous mixture such as air, through a catalyst bed in a tube-bundle or fluidized-bed reactor at a temperature of 250 – 650 °C. The reaction is highly exothermic, and effective cooling is necessary. It is advisable to dilute the mixture of toluene vapor and oxygen-containing gas with an unreactive gas, such as water vapor, nitrogen, or carbon dioxide. The benzaldehyde yield is favored by a low conversion rate (10 – 20 %) per pass, a short residence time (0.1 – 1.0 s), and a precise adjustment of the amount of oxygen. Even then, it is only 40 – 60 % of the theoretical yield based on toluene. The oxidation of toluene in the gas phase produces maleic anhydride, citraconic anhydride, phthalic anhydride, anthraquinone, cresol, acetic acid, and other compounds, in addition to substantial quantities of benzoic acid, carbon monoxide, and carbon dioxide [22]. The complete combustion of the toluene is retarded by applying high pressure [23] or by adding potassium sulfate or sodium fluoride to the catalyst [24].

Oxides of the elements of Groups V and VI of the periodic table are frequently used as catalysts, sometimes with other oxides. These oxide catalysts contain molybdenum and at least one additional element from among iron, nickel, cobalt, antimony, bismuth, vanadium, phosphorus, samarium, tantalum, tin, and chromium [25]. Some have palladium and phosphoric acid on activated carbon [26]. Others are mixed oxides of silver and transition metals, such as cobalt nickel, vanadium, platinum, or palladium [27], and one is a mixture of silver vanadate and iron vanadate [28].

Another process uses a mixed-oxide catalyst that has silicon dioxide as a carrier and contains uranium, copper, iron, phosphorus, tellurium, and lead, in addition to molybdenum [29]. At a toluene consumption of 35 – 50 % per pass, the selectivity of the benzaldehyde formation at reaction temperatures of 475 – 550 °C is 40 – 70 %. Catalysts based only on mixed oxides of molybdenum and uranium [30] require reaction temperatures of around 600 °C.

More important than gas-phase oxidation for the production of benzaldehyde is the oxidation of toluene in the liquid phase by oxygen in the form of air or other gaseous mixtures. This is carried out at 80 – 250 °C, preferably in the presence of cobalt, nickel, manganese, iron, or chromium compounds (alone or in combination) as catalysts. Lead compounds [31], ruthenium compounds [32], and thallium salts of organic acids [33] are also catalytic. A pressure at which the reaction medium remains liquid is chosen. Alkali, alkaline earth, and tertiary alkyl hypochlorites have been used as oxidation promoters [34] in place of the usual bromides. These promoters may cause corrosion, however.

Toluene can also be oxidized in an alcohol, e.g., methanol [35], an aliphatic aldehyde, e.g., acetaldehyde [36], or an acid, e.g., benzoic acid [37] or acetic acid [38]. Water can also be added to the reaction [39], but one process im-

proves the benzaldehyde selectivity by continuously removing the water formed in the oxidation reaction [40]. When a cobalt catalyst is used and conversion rates of up to about 25 % are achieved, this process gives aldehyde selectivities of 40 – 80 %.

In another process, the selectivity of the benzaldehyde formation is improved by liquid-phase oxidation of the toluene with oxygen in the presence of a phosphoric acid – palladium catalyst that also contains certain phosphorus, sulfur, or nitrogen modifiers [41]. For an oxygen consumption of 63 %, the yield of benzaldehyde is 41 %.

In the large-scale production of phenol by the Dow process and of ε-caprolactam by the Snia Viscosa process via the toluene – benzoic acid synthesis route (→ Benzoic Acid), considerable amounts of benzaldehyde are formed as a by-product. Because a large proportion of this byproduct is worked up into pure benzaldehyde, oxidation processes intended exclusively to produce benzaldehyde are seldom used industrially at the present time.

Even when catalytic liquid-phase oxidation of toluene is carried out for the express purpose of manufacturing benzaldehyde, substantial quantities of benzoic acid are formed. Depending on the reaction conditions, carbon dioxide, carbon monoxide, formaldehyde, formic acid, benzyl alcohol, benzyl hydroperoxide, biphenyl, methylbiphenyls, and other compounds are formed also as byproducts. The crude benzaldehyde is normally refined by distillation at reduced pressure in a stainless steel column.

Benzaldehyde produced by oxidation of toluene often contains small amounts of impurities that are difficult to remove by distillation and which discolor the product. To remove these impurities the benzaldehyde can be treated with aqueous alkali [42] or with water and a base metal such as zinc powder [43]. A color-stable product can be obtained by treatment with oxidizing agents, e.g., hydrogen peroxide, followed by distillation [44].

Those processes in which toluene is oxidized with other agents, such as manganese dioxide in sulfuric acid, sodium persulfate, chromium(VI) oxide in acetic anhydride, or chromyl chloride (Etard reaction), cause waste water disposal problems and are without industrial importance.

4.3. Other Production Processes

Other processes for the manufacture of benzaldehyde include the reaction of benzene with carbon monoxide [45], the oxidation or dehydrogenation of benzyl alcohol [46], and the ruthenium-catalyzed oxidation of styrene with periodate or hypochlorite [47]. Benzaldehyde can also be made by hydrolysis of mixtures of benzyl chloride and benzal chloride in dilute nitric acid, using vanadium pentoxide as catalyst [48], or in the presence of an aqueous solution of hexamethylenetetramine [8]. The reduction of benzoyl chloride [49] or methyl benzoate [50] can also be used. These processes have no industrial importance in the production of benzaldehyde and are more likely to be used to make certain nuclear-substituted derivatives.

5. Quality Specifications and Test Methods

Benzaldehyde is commercially available in two grades (Table 3). *Pure* benzaldehyde, which is suitable for most uses, accounts for more than 95 % of the amount sold. *Double-distilled* benzaldehyde is used mainly in the pharmaceutical, perfume, and flavor industries and must therefore be particularly pure and free from foreign odors.

Table 3. Purity specifications of benzaldehyde grades

Property	Industrial grade	
	Pure	Double-distilled
Assay, wt %	> 99.0	> 99.5
Benzoic acid, wt %	< 0.3	< 0.1
Water, wt %	< 0.3	< 0.1
Chlorine, wt %	< 0.02	< 0.007
Color (APHA)	< 20	< 20

Traces of chlorine compounds may be detectable in benzaldehyde obtained by the hydrolysis of benzal chloride, whereas benzaldehyde obtained by oxidation may contain byproducts formed in the oxidation of toluene. Benzaldehyde can now be produced, by hydrolysis of benzal chloride and by partial oxidation of toluene, to such a high standard of purity that it is suitable for all known applications.

A suitable quantitative determination method consists of reacting benzaldehyde with hydroxylamine hydrochloride and titrating the hydrochloric acid liberated. Gas chromatography in packed or capillary columns is becoming increasingly important, particularly in the determination of impurities. The pharmacopeias of the various countries should be consulted regarding the testing of benzaldehyde that is to be used for pharmaceutical purposes. United States Pharmacopeia XX includes tests for halogen compounds, hydrocyanic acid, and nitrobenzene.

6. Storage, Transportation, and Safety Regulations

Benzaldehyde is readily oxidized to benzoic acid in the presence of air and should therefore be stored under nitrogen. Tanks for benzaldehyde should be made of aluminum or stainless steel. Benzaldehyde can be transported in drums with baked enamel finishes or in tanks made of aluminum or stainless steel. The air content of containers used for shipping benzaldehyde must be minimized.

The transportation of benzaldehyde by rail and road is regulated by the provisions of Class 3, Figure 32c, of GGVE/RID and GGVS/ADR. The transportation of benzaldehyde by sea is not restricted by the IMDG Code. The labeling of benzaldehyde must comply with national regulations for dangerous substances or with Supplement I of the EEC Directive on the Packaging, Classification, and Labelling of Dangerous Substances, 6th Amendment (see Journal of the European Community no. L 360 of December 30, 1976).

Benzaldehyde belongs to Temperature Class T 4 (autoignition temperature 190 °C) and has a flash point of 64.5 °C. Electrical installations must be labeled accordingly. If it is finely divided in air, e.g., on cleaning cloths, sawdust, kieselguhr, or other materials with large surface areas, benzaldehyde shows a very strong self-heating tendency. Because of the relatively low autoignition temperature, there is a risk of spontaneous combustion under such circumstances.

7. Uses

Benzaldehyde is an important starting material for the manufacture of odorants and flavors. The compound is responsible for the odor of natural bitter almond oil and is incorporated directly in perfumes, soaps, foods, drinks, and other products. Substantial amounts are used in the production of derivatives that are also employed in the perfume and flavor industries. Examples of these derivatives are cinnamaldehyde, cinnamaldehyde dimethyl acetal, α-amylcinnamaldehyde, α-hexylcinnamaldehyde, dihydrocinnamaldehyde, cinnamyl alcohol, cinnamic acid, 4-phenyl-2-butanone (benzylacetone), β-bromostyrene, 2,2,2-trichloro-1-phenylethyl acetate, and benzyl benzoate.

Another application of benzaldehyde is the production of triphenylmethane dyes. For example, the leuco base of Malachite Green is obtained by condensation of benzaldehyde with dimethylaniline. The acridine dye benzoflavin is obtained from benzaldehyde and *m*-toluenediamine. In the pharmaceutical industry benzaldehyde is used as an intermediate in the manufacture of chloramphenicol, ephedrin, ampicillin, diphenylhydantoin, and other products.

Other important chemical intermediates obtained from benzaldehyde are benzoin, benzylamine, benzyl alcohol, mandelic acid, and 4-phenyl-3-buten-2-one (benzylideneacetone). Benzaldehyde is used in photochemistry, as a corrosion inhibitor and dyeing auxiliary, in the electroplating industry, and in the production of agricultural chemicals. The use of benzaldehyde in the extractive separation of isomeric amines has been proposed [51].

8. Derivatives

8.1. Chlorobenzaldehydes

2-Chlorobenzaldehyde [*89-98-5*], C_7H_5ClO, M_r 140.57, is a colorless to yellowish liquid with a penetrating odor and low solubility in water, *mp* 11.6 °C, *bp* 212 °C at 101.3 kPa, *bp* 91 – 92 °C at 2.0 kPa, density ϱ 1.248 g/cm^3 at 20 °C, refractive index n_D^{20} 1.5660, flash point 90 °C.

2-Chlorobenzaldehyde is produced mainly by chlorination of 2-chlorotoluene to form

2-chlorobenzal chloride, which is then subjected to acid hydrolysis. Metal salts, such as iron(III) chloride, are used as catalysts [52]. The hydrolysis can also be accomplished using formic acid without a catalyst [13]. 2-Chlorobenzaldehyde can also be produced by oxidation of 2-chlorobenzyl chloride with N-oxides of tertiary amines [53] or with dilute nitric acid [48]. 2-Chlorobenzaldehyde is considerably more resistant to oxidation than benzaldehyde. When it is heated with sodium sulfite solution under pressure, benzaldehyde-2-sulfonic acid forms.

3-Chlorobenzaldehyde [*587-04-2*], C_7H_5ClO, M_r 140.57, *mp* 17 °C, *bp* 213 – 214 °C at 101.3 kPa, *bp* 93 – 96 °C at 2.0 kPa, density ϱ 1.236 g/cm^3 at 20 °C, refractive index n_D^{20} 1.5641.

3-Chlorobenzaldehyde is formed in the nuclear chlorination of the benzaldehyde – aluminum chloride complex compound in the presence of chlorinated hydrocarbon solvents [54]. It can also be obtained from the diazonium salt of 3-aminobenzaldehyde by the Sandmeyer reaction.

4-Chlorobenzaldehyde [*104-88-1*], C_7H_5ClO M_r 140.57, is a colorless to yellowish crystalline mass with a penetrating odor and low solubility in water, *mp* 48 – 49 °C, *bp* 214.5 – 216.5 °C at 101.3 kPa, *bp* 98 °C at 1.87 kPa, flash point 88 °C.

4-Chlorobenzaldehyde is produced by the same methods as 2-chlorobenzaldehyde. Chlorobenzonitrile can be hydrogenated to form 4-chlorobenzaldehyde in the presence of an acid and a nickel catalyst that has been treated with a copper salt [55].

2,4-Dichlorobenzaldehyde [*874-42-0*], $C_7H_4Cl_2O$, M_r 175.02, *mp* 72 °C, *bp* 233 °C at 101.3 kPa, *bp* 105 – 106 °C at 2.0 kPa, flash point 135 °C.

2,4-Dichlorobenzaldehyde is produced by hydrolysis of 2,4-dichlorobenzal chloride with sulfuric acid.

2,6-Dichlorobenzaldehyde [*83-38-5*], $C_7H_4Cl_2O$, M_r 175.02, *mp* 71 – 72.5 °C.

2,6-Dichlorobenzaldehyde is formed by hydrolysis of 2,6-dichlorobenzal chloride with concentrated sulfuric acid. In aromatic sulfonic acids, hydrolysis can be catalyzed by zinc chloride, iron(III) chloride, or aluminum chloride [56]. 2,6-Dichlorobenzaldehyde has also been obtained by the oxidation of 2,6-dichlorotoluene with manganese dioxide and sulfuric acid. To separate the 2,6-dichlorobenzaldehyde from isomeric chlorine-containing benzaldehydes, the solution of the aldehyde mixture is extracted in an organic solvent with aqueous sodium bisulfite solution [57]. The 2,6-dichlorobenzaldehyde accumulates in the organic phase.

The chlorobenzaldehydes are important intermediates in the production of dyes, optical brighteners, agricultural chemicals, and pharmaceuticals.

8.2. Nitrobenzaldehydes

2-Nitrobenzaldehyde [*552-89-6*], $C_7H_5NO_3$, M_r 151.12, *mp* 42 – 44 °C, *bp* 153 °C at 3.07 kPa, highly volatile in steam.

The classical method for the preparation of 2-nitrobenzaldehyde begins with cinnamic acid, which is nitrated to form a mixture of 2-nitro- and 4-nitrocinnamic acid. The two isomers are separated by fractional crystallization after which the 2-nitrocinnamic acid is oxidized with potassium permanganate to form 2-nitrobenzaldehyde. Nitrobenzaldehyde can also be prepared by the oxidation of 2-nitrostyrene with ozone [58] or molecular oxygen [59]. Other processes begin with 2-nitrobenzyl bromide, which is oxidized to the aldehyde with dimethyl sulfoxide [60] or N-oxides of tertiary amines [53]. According to a similar process [61], 2-nitrobenzaldehyde can be obtained by the alkaline hydrolysis of 2-nitrobenzyl bromide followed by the oxidation of the intermediate 2-nitrobenzyl alcohol using nitric acid.

3-Nitrobenzaldehyde [*99-61-6*], $C_7H_5NO_3$, M_r 151.12, *mp* 56 – 58 °C, *bp* 164 °C at 3.07 kPa.

3-Nitrobenzaldehyde is obtained by the nitration of benzaldehyde with a mixture of concentrated sulfuric acid and nitric acid at 5 – 10 °C. In this process, up to 20 % of the 2-isomer is formed.

Warning: *A risk of explosion exists while the crude product is being distilled under reduced pressure* [62]. The two isomers can be separated by exploiting the difference in their reactions with bisulfite.

4-Nitrobenzaldehyde [*555-16-8*], $C_7H_5NO_3$, M_r 151.12, *mp* 107 °C, relatively involatile in steam.

4-Nitrobenzaldehyde can be obtained by the oxidation of 4-nitrotoluene with chromium (VI)oxide in glacial acetic acid. It can also be obtained by hydrolyzing 4-nitrobenzal bromide in 93 % sulfuric acid [63].

The nitrobenzaldehydes serve as intermediates in the production of dyes and pharmaceuticals.

8.3. Hydroxybenzaldehydes

2-Hydroxybenzaldehyde [*90-02-8*], salicylaldehyde, $C_7H_6O_2$, M_r 122.12, is a yellowish, oily liquid with an odor resembling that of bitter almonds; it is volatile in steam. It occurs naturally in cassia oil and in the essential oils of several plants of the genus *Spirea*, *mp* 1.6 °C, *bp* 196 °C at 101.3 kPa, *bp* 92 °C at 2.93 kPa, density ϱ 1.167 g/cm^3 at 20 °C, refractive index n_D^{20} 1.5718, dynamic viscosity η 2.50×10^{-3} Pa · s at 25 °C, surface tension σ 42×10^{-3} N/m at 25 °C, dipole moment μ (measured in liquid benzene) 9.54×10^{-30} C m (2.86 D) heat of combustion 27.29 kJ/g at 20 °C and constant pressure.

2-Hydroxybenzaldehyde is soluble in many organic solvents, in 20 % aqueous sodium carbonate, and in 10 % aqueous sodium hydroxide. At 86 °C, 1.7 g can be dissolved in 100 g of water. With iron(III) chloride, 2-hydroxybenzaldehyde gives an intense violet color. It forms yellow alkali salts through inner complexing.

The two most important processes for the manufacture of 2-hydroxybenzaldehyde begin with phenol. The best known process is based on the Reimer-Tiemann reaction, in which phenol is reacted with chloroform in the presence of an aqueous alkali. Moderate yields of 2-hydroxy- and 4-hydroxybenzaldehyde are obtained in a ratio of about 85 : 15. The use of aqueous methanol as a reaction medium [64] improves the yield, especially that of 4-hydroxybenzaldehyde. According to a modern variation [65] of the Reimer-Tiemann reaction, the process is carried out in the absence of water. An alkali-metal phenolate reacts with chloroform and a suspension of an alkali-metal hydroxide in an inert organic solvent.

One industrial process starts with phenol and forms 2-hydroxybenzyl alcohol (saligenin) by reacting triphenyl metaborate with formaldehyde [66]. The alcohol is catalytically oxidized in air to give 2-hydroxybenzaldehyde [67]. It is more economical to react phenol with formaldehyde in the presence of alkaline catalysts and in the absence of boric acid [68]. The yield is approximately 85 %. The ratio of 2-hydroxybenzyl alcohol to 4-hydroxybenzyl alcohol can be altered by varying the catalyst. The oxidation of the hydroxybenzyl alcohols is carried out in an aqueous solution with a platinum – lead – carbon catalyst and gives the corresponding hydroxybenzaldehydes in yields of more than 98 %.

Several other processes are based on side-chain chlorination of 2-cresol and saponification of the resulting dichloromethyl group to form the aldehyde group. The phenolic hydroxyl group is protected prior to chlorination by esterification with an inorganic or organic acid chloride, such as phosphorus oxychloride [69], phosgene [70], tetrachlorosilane [71], or acetyl chloride [72]. Hydrolysis splits the ester. 2-Hydroxybenzaldehyde can also be produced by electrolytic reduction of salicylic acid on a rotating amalgam cathode [73] and by catalytic reduction of a salicylic acid halide [74].

2-Hydroxybenzaldehyde is one of the most important derivatives of benzaldehyde. Estimated annual production worldwide is 4 – 6 kt. Because of its pleasant aromatic odor, 2- hydroxybenzaldehyde is used in perfumes and, on a large scale, as a starting material in the production of cumarin.

Its condensation products with amines have complex-forming properties and are used as additives, e.g., in petrochemistry. 2-Hydroxybenzaldehyde is also an intermediate for dyes, pharmaceuticals, plastics, photographic chemicals, agricultural chemicals, and electroplating chemicals.

3-Hydroxybenzaldehyde [*100-83-4*], $C_7H_6O_2$, M_r 122.12, *mp* 108 °C, *bp* 240 °C

at 101.3 kPa, heat of combustion 27.04 kJ/g at 20 °C and constant pressure, involatile in steam.

3-Hydroxybenzaldehyde can be produced by heating 3-chlorobenzaldehyde with aqueous potassium hydroxide under pressure at 250 °C [75] or by the reduction of 3-nitrobenzaldehyde to 3-aminobenzaldehyde and hydrolysis of the diazonium salt. It is used as an intermediate in producing dyes.

4-Hydroxybenzaldehyde [*123-08-0*], $C_7H_6O_2$, M_r 122.12, *mp* 117 °C, sublimes at 110 – 120 °C and 4 Pa, heat of combustion 27.18 kJ/g at 20 °C and constant pressure involatile in steam. At 18 °C, 0.81 g dissolves in 100 g of water.

4-Hydroxybenzaldehyde can be produced similarly to 2-hydroxy- and 3-hydroxybenzaldehyde. It can also be produced by treating benzaldehydes or acetalized benzaldehydes that have a benzyloxy residue in the 4-position with aqueous hydrogen halides [76]. The benzaldehyde derivatives used as starting materials are obtained by catalytic or electrochemical oxidation of *p*-benzyloxytoluene.

4-Hydroxybenzaldehyde is used in electroplating and is an important intermediate for making dyes, pharmaceuticals, textile auxiliaries, odorants, and agricultural chemicals.

8.4. Aminobenzaldehydes

2-Aminobenzaldehyde [*529-23-7*], C_7H_7NO, M_r 121.14, *mp* 39 – 40 °C, *bp* 80 – 85 °C at 270 Pa, low solubility in water.

2-Aminobenzaldehyde is obtained by the reduction of 2-nitrobenzaldehyde with iron(II) sulfate and ammonia or by catalytic hydrogenation. It is also formed in the hydrolysis of 2-dichloromethylphenyl isocyanate [77].

4-Aminobenzaldehyde [*556-18-3*], C_7H_7NO, M_r 121.14, *mp* 71 °C, soluble in water.

The preferred method for the production of 4-aminobenzaldehyde is the conversion of 4-nitrotoluene with sodium polysulfide. The reaction is carried out in an aprotic polar solvent, such as dimethylformamide [78].

The aminobenzaldehydes have a strong tendency to undergo self-condensation, in which they expel water. The aldehydes should therefore be isolated in the form of their stable N-acetyl compounds. They are used as intermediates in the production of dyes. 2-Aminobenzaldehyde is a starting material in the manufacture of quinoline derivatives [77].

8.5. Methylbenzaldehydes

2-Methylbenzaldehyde [*529-20-4*], C_8H_8O, M_r 120.15, liquid with the odor of bitter almonds, *bp* 199 – 200 °C at 101.3 kPa, *bp* 102 – 104 °C at 4.5 kPa, density ϱ 1.0328 at 20 °C, refractive index n_D^{20} 1.5462.

2-Methylbenzaldehyde can be produced by partial oxidation of *o*-xylene, e.g., in acetic acid with cobalt salts as catalysts and bromine compounds as promoters [79]. It can also be obtained from 2-methylbenzyl chloride by gas-phase oxidation at 350 – 450 °C in the presence of $Al_2O_3-V_2O_5$ catalysts [80]. In the Sommelet reaction, 2-methylbenzyl chloride reacts with hexamethylenetetramine to form 2-methylbenzaldehyde. This reaction can also use industrial chlorination mixtures containing small amounts of highly chlorinated products in addition to 2-methylbenzyl chloride and unreacted *o*-xylene [81].

3-Methylbenzaldehyde [*620-23-5*], C_8H_8O, M_r 120.15, liquid with the odor of bitter almonds, *bp* 200 °C at 101.3 kPa, *bp* 83 °C at 2.13 kPa, density ϱ 1.019 at 20 °C, refractive index n_D^{20} 1.5410.

3-Methylbenzaldehyde can be prepared similarly to 2-methylbenzaldehyde.

4-Methylbenzaldehyde [*104-87-0*], C_8H_8O, M_r 120.15, liquid with a pepper-like odor, *bp* 204 °C at 101.3 kPa, *bp* 85 – 86 °C at 2.0 kPa, density ϱ 1.019 at 20 °C, refractive index n_D^{20} 1.5454, dipole moment μ (in liquid benzene) 10.81×10^{-30} C m (3.24 D).

4-Methylbenzaldehyde is produced industrially by a modified Gattermann-Koch reaction. Toluene is reacted with carbon monoxide in the presence of hydrogen fluoride and tantalum(V) fluoride [82] or hydrogen fluoride and boron trifluoride [83]. It can also be produced similarly to 2-methylbenzaldehyde.

The modified Gattermann-Koch synthesis for the production of 4-methylbenzaldehyde [83] is the first of two stages in a process for the production of terephthalic acid from the inexpensive raw materials toluene and carbon monoxide [84].

The methylbenzaldehydes serve as intermediates (e.g., in the production of terephthalic acid) and are also used in the flavors industry.

9. Economic Aspects

The benzaldehyde production capacities of Western Europe, the United States, and Japan in 1983 were approximately 14 kt, 7 kt, and 3.5 kt, respectively. The corresponding production figures were approximately 12 kt, 4.5 kt, and 2 kt.

10. Toxicology

Benzaldehyde is used as a flavoring and fragrance in food, cosmetics, pharmaceuticals, and soap and is "generally regarded as safe" (GRAS) by the U.S. Food and Drug Administration. Benzaldehyde has industrial and agricultural applications, but no environmental or workplace exposure limits have been proposed.

Acute toxicity of benzaldehyde is moderate, with oral LD_{50} values reported to be 1.3 g/kg in rats and 1 g/kg in guinea pigs [85]. The probable lethal dose by mouth for a 70-kg human is estimated to be 50 mL. Subchronic oral administration to rodents produced no effects at daily doses of 400 mg/kg in rats and 300 – 600 mg/kg in mice. Higher doses caused necrotic and degenerative lesions in brain, kidney, and forestomach [86]. Tests for mutagenicity were negative in the Ames TA 100 *Salmonella* strain at concentrations of $0.1 - 10^3$ μg per plate [87]. No studies of carcinogenic, teratogenic, or reproductive effects have been reported.

Effects of toxic doses in acute and subchronic studies included depression, inactivity, tremors, seizures, and coma. Death may result from respiratory depression. Benzaldehyde has a weak local anesthetic effect and is mildly irritating to the eye and upper respiratory tract. Skin irritation is moderate, and allergic sensitivity may occur in some individuals [88].

Contact with eyes and skin should be avoided by use of gloves and protective clothing. Self-contained breathing apparatus should be used in poorly ventilated areas to prevent inhalation exposure.

11. References

General References

1. *Beilstein,* **7,** 174; **7 (1),** 113; **7 (2),** 145; **7 (3),** 805; **7 (4),** 505. *Houben-Weyl,* **vol. E 3** (1983). *Ullmann,* 4th ed., **vol. 8,** pp. 343 – 351. T. Laird in: *Comprehensive Organic Chemistry,* **vol. 1,** Pergamon Press, Oxford 1979, pp. 1105 – 1160.

Specific References

2. J. Gmehling, U. Onken, W. Arlt: "Vapor-Liquid Equilibrium Data Collection," *Dechema Chemistry Data Series,* vol. I, part 2 d (1982) 474 – 475, 568 – 571, 606, 620; part 3 + 4 (1979) 62 – 63.
3. L. H. Horsley: "Azeotropic Data-III," *Adv. Chem. Ser.* **116** (1973) 360 – 362.
4. Stamicarbon B.V., DE-OS 2421268, 1973 (J. H. Gregoire).
5. H. Stetter, R. Y. Rämsch, H. Kuhlmann, *Synthesis* 1976, 733.
6. W. H. Shearon, H. E. Hall, J. E. Stevens, *Ind. Eng. Chem.* **41** (1949) 1812 – 1820.
7. EKA AB, EP-A 64486, 1981 (R. K. Rantala, G. L. F. Hag).
8. B. Ya. Libman et al., *Zh. Prikl. Khim. (Leningrad)* **39** (1966) 1669 – 1670; *Chem. Abstr.* **65** (1966) 18518 f.
9. Nippon Kayaku Co., JP-Kokai 7589337, 1973 (T. Ishikura); *Chem. Abstr.* **83** (1975) 205940 s.
10. Diamond Shamrock Corp., US 3499934, 1967 (W. J. Pyne).
11. Albright & Wildon, DE-OS 2261616, 1971 (H. Coates, W. E. Billingham).
12. Bayer, DE-OS 2752612, 1977 (F. Brühne, K. A. Lipper).
13. Bayer, EP-A 41672, 1980 (H. U. Blank, E. Wolters).
14. Argus Chem. Corp., US 3691217, 1970 (T. J. McCann).
15. Ciba-Geigy, DE-AS 2044832, 1969 (P. Liechti, F. Blattner).
16. General Aniline & Film Corp., US 3087967, 1960 (D. E. Graham, W. C. Craig).
17. Tenneco Chem. Inc., US 3524885, 1967 (A. J. Deinet).

18. Instytut Przemyslu Organicznego, PL 93511, 1975 (D. Narkiewicz, J. Legocki); *Chem. Abstr.* **90** (1979) 87043 w.
19. Hodogaya Chem. Co., JP 7107927, 1967 (Y. Murakami, T. Koizumi); *Chem. Abstr.* **75** (1971) 19945 g.
20. Central Glass Co., DE-OS 3226490, 1981 (T. Kondow, K. Okazaki, Y. Katsuhara, K. Matsuoka).
21. Ihara Chem. Ind. Co., JP 8202699, 1971; *Chem. Abstr.* **97** (1982) 5980 g.
22. H. Pichler, F. Obenaus, *Brennst. Chem.* **45** (1964) 97 – 103.
23. A. Guyer, DE-AS 1236493, 1963 (A. Guyer, P. Guyer, G. Gut).
24. Mitsubishi Chem. Ind. Co., JP-Kokai 79151937, 1978 (T. Onoda, K. Wada, S. Fujii); *Chem. Abstr.* **92** (1980) 180826 e.
25. BASF, DE-OS 2136779, 1971 (R. Krabetz, Ch. Dudeck, W. Fuchs).
26. National Distillers & Chem. Corp., US 3946067, 1972 (J. Kwiatek, J. H. Murib, C. K. Brush).
27. Standard Oil Co. (Indiana), US 4005049, 1975 (E. K. Fields).
28. Stamicarbon B.V., DE-OS 2730761, 1976 (P. C. van Geem, A. J. Teunissen).
29. Ashland Oil Inc., US 4390728, 1981 (C. Daniel).
30. Institut Français du Pétrole, des Carburants et Lubrifiants, FR 1568763, 1967 (B. Delmon).
31. Mitsubishi Chem. Ind. Co., JP 7528946, 1970 (Y. Suzuki, T. Maki, H. Nakajima, K. Mineda); *Chem. Abstr.* **84** (1976) 89823 a.
32. Universal Oil Products Co., US 3775472, 1970 (S. N. Massie).
33. Universal Oil Products Co., US 3723517, 1970 (S. N. Massie).
34. Universal Oil Products Co., US 3790624, 1972 (S. N. Massie, H. S. Bloch).
35. Mitsubishi Gas Chem. Co., Inc., JP-Kokai 8062029, 1978; *Chem. Abstr.* **93** (1980) 149992 c.
36. Union Carbide Corp., US 3931330, 1972 (N. S. Aprahamian).
37. Mitsubishi Chem. Ind. Co., JP-Kokai 7805132, 1976 (Y. Murao, T. Nakanome, T. Yamaura); *Chem. Abstr.* **88** (1978) 169778 n.
38. Phillips Petroleum Co., US 4088823, 1975 (H. D. Holtz, L. E. Gardner).
39. Mitsubishi Chem. Ind. Co., JP-Kokai 75108231, 1974 (A. Wada, N. Nakajima, T. Hironaka); *Chem. Abstr.* **84** (1976) 43580 r.
40. Hoechst, EP-A 71166, 1981 (H. Kuckertz, G. Schaeffer).
41. National Distillers & Chem. Corp., DE-OS 2745511, 1976 (J. A. Scheben).
42. Mitsubishi Chem. Ind. Co., JP 7424467, 1969 (K. Okuno, K. Itagaki, T. Hironaka); *Chem. Abstr.* **82** (1975) 72638 e.
43. Stamicarbon B.V., EP-A 15616, 1979 (C. Jongsma).
44. Stamicarbon B.V., EP-A 1660, 1977 (J. Elmendorp).
45. Mobil Oil Corp., US 3369048, 1963 (L. A. Hamilton, P. S. Landis).
46. Stamicarbon B.V., EP-A 37149, 1980 (T. F. De Graaf, H. J. Delahaye).
47. Mitsubishi Petrochemical Co., JP-Kokai 8087739, 1978; *Chem. Abstr.* **94** (1981) 103012 a.
48. Mitsubishi Gas Chem. Co., JP-Kokai 8218644, 1980; *Chem. Abstr.* **97** (1982) 72084 f.
49. Fuso Kagaku Kogyo Co., JP-Kokai 7823913, 1976 (K. Kobayashi, N. Ishino, Y. Ota, S. Wakita); *Chem. Abstr.* **89** (1978) 23964 u.
50. Dow Chem. Co., US 4328373, 1980 (E. J. Strojny).
51. UOP Inc., US 4174351, 1978 (J. P. Shoffner).
52. Bayer, DE-OS 2026817, 1970 (J. Schneider).
53. Dynamit Nobel, DE-OS 2948058, 1979 (G. Bernhardt, E. N. Petersen, G. Daum).
54. Fujisawa Pharmaceutical Co., JP-Kokai 7777021, 1975 (K. Kariyone, H. Yagi); *Chem. Abstr.* **87** (1977) 201095 e.
55. Sumitomo Chem. Co., EP-A 87298, 1982 (H. Yamachika, H. Nakanishi).
56. S. Kobayashi, JP-Kokai 7966638, 1977; *Chem. Abstr.* **91** (1979) 175009 r.
57. Hooker Chem. & Plastics Corp., US 4136122, 1978 (B. R. Cotter).
58. Ciba-Geigy, DE-OS 2829346, 1977 (J. Gosteli).
59. Bayer, DE-OS 2805402, 1978 (F. Hagedorn, L. Imre, K. Wedemeyer).
60. Bayer, DE-OS 2808930, 1978 (W. Ertel).
61. VEB Arzneimittelwerk Dresden, DE-OS 2708115, 1976 (W. Sauer, H. Goldner, H. J. Heidrich, G. Faust, W. Fiedler, E. Carstens, G. Heine).
62. J. Lange, T. Urbański, *Chem. Ind. (London)* 1967, 1424.
63. Nippon Kayaku Co., JP-Kokai 7589337, 1973 (T. Ishikura); *Chem. Abstr.* **83** (1975) 205940 s.
64. Dow Chem. Co., US 3365500, 1964 (D. F. Pontz).
65. Sumitomo Chem. Co., EP-A 68725, 1981 (K. Hamada, G. Suzukamo).

66. Rhône-Poulenc S. A., FR 1328945, 1962 (P. Marchand, J. B. Grenet).
67. Rhône-Poulenc S. A., FR 2305420, 1975 (J. Le Ludec).
68. H. Fiege, K. Wedemeyer, K. A. Bauer, A. Krempel, R. G. Mölleken: "Fragrance and Flavor Substances," in R. Croteau (ed.): *Fragrance Flavor Subst. Proc. Int. Haarman & Reimer Symp. 2nd* **1979,** 63 – 73; *Chem. Abstr.* **95** (1981) 150058 a.
69. Tenneco Chem., Inc., US 3641158, 1969 (A. J. Deinet, D. X. Klein).
70. BASF, DE-OS 1925195, 1969 (H. Hoffmann, J. Datow, G. Wenner).
71. Tokuyama Soda Co., JP-Kokai 83124729, 1982; *Chem. Abstr.* **99** (1983) 158016 p.
72. Yoshitomi Pharmaceutical Ind., JP 7303831, 1968 (K. Saruwatari, T. Gono, K. Tsubone); *Chem. Abstr.* **79** (1973) 18387 x.
73. K. S. Udupa, G. S. Subramanian, H. V. K. Udupa, *Ind. Chem.* **39** (1963) 238 – 241; *Chem. Abstr.* **59** (1963) 10986 b.
74. Seimi Kagakagu Co., JP 6813204, 1965 (S. Abe, K. Sato, T. Asami, T. Amakasu, T. Itakura); *Chem. Abstr.* **70** (1969) 28646 j.
75. Bayer, DE 942808, 1953 (W. Müller, W. Möllering, W. Schommer).
76. BASF, DE-OS 2904315, 1979 (M. Barl, D. Degner, H. Siegel).
77. Bayer, DE-OS 2730061, 1977 (V. Ehrig, H. S. Bien, E. Klauke, D. I. Schütze).
78. K. Kato, M. Kawamura, T. Nishi, H. Hata, *Nippon Kagaku Kaishi* 1981, no. 2, 255 – 258; *Chem. Abstr.* **95** (1981) 6705 s.
79. Agency of Industrial Sciences and Technology; Sanko Chemical Co., JP-Kokai 79157534, 1978 (J. Imamura, K. Kizawa); *Chem. Abstr.* **92** (1980) 215066 m.
80. Rhône-Poulenc S. A., DE-OS 2062522, 1969 (Y. Colleuille, R. Perron).
81. BASF, DE-OS 2942894, 1979 (W. Schoch, M. Kröner, R. Widder).
82. Texaco Development Corp., US 4218403, 1979 (S. H. Vanderpool).
83. Mitsubishi Gas Chemical Co., Inc., DE-OS 2460673, 1973 (S. Fujiyama, T. Takahashi, S. Kozao, T. Kasahara).
84. Mitsubishi Gas Chemical Co., Inc., GB 2025956, 1978 (M. Komatsu, T. Ohta, T. Tanaka, R. Oda, Y. Takamizawa). S. Fujiyama, T. Kasahara, *Hydrocarbon Process.* **57** (1978) no. 11, 147 – 149.
85. P. M. Jenner, E. C. Hagan, J. M. Taylor, E. L. Cook, O. G. Fitzhugh, *Food Cosmet. Toxicol.* **2** (1964) 327 – 343.
86. W. M. Kluwe, C. A. Montgomery, *Food Cosmet. Toxicol.* **21** (1983) 245 – 250.
87. W. H. Rapson, M. A. Nazar, V. V. Butsky, *Bull. Environ. Contam. Toxicol.* **24** (1980) 590 – 596.
88. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **14 Suppl.** (1976) 693 – 698.

Benzanthrone → Anthraquinone Dyes and Dye Intermediates

Benzene

HILLIS O. FOLKINS, (Union Oil Company of California), Claremont, California 91711, United States

1.	**Introduction**	1
2.	**Physical Properties**	1
3.	**Chemical Properties**	3
4.	**Raw Materials**	4
5.	**Production**	6
5.1.	**Production from Coal**	6
5.2.	**Production from Petroleum**	8
5.2.1.	Benzene from Reformate	8
5.2.2.	Aromatics from Pyrolysis Gasoline	9
5.3.	**Methods of Increasing Production**	11
5.3.1.	Hydrodealkylation	11
5.3.2.	Disproportionation	13
5.3.3.	Combination Dealkylation Processes	14
5.3.4.	Benzene from Other Sources	15
6.	**Separation and Purification**	16
6.1.	**Removal of Dienes, Olefins, and Sulfur Compounds**	16
6.2.	**Close Fractionation**	17
6.2.1.	Azeotropic Distillation	17
6.2.2.	Extractive Distillation	18
6.2.3.	Liquid – Liquid Extraction	18
6.3.	**Processes for Aromatic Separation**	18
7.	**Quality Specifications**	25
8.	**Storage and Transportation**	25
9.	**Economic Aspects**	26
10.	**Uses**	29
11.	**Toxicology and Occupational Health**	29
12.	**References**	32

1. Introduction

Benzene [*71-43-2*] is a single-ring, aromatic compound, C_6H_6, M_r 78.11; colorless, flammable liquid, *bp* 80.1 °C, *fp* 5.5 °C. The term benzene denotes the pure compound; benzol is still used to a small degree in some countries to represent the compound or a material having benzene as its main component. Benzine, on the other hand, is a low-boiling hydrocarbon mixture or naphtha, often nonaromatic in composition.

Benzene

Benzene is thermally stable but is chemically reactive and thus serves as a source for the production of many petrochemicals and hydrocarbon materials. Other chemicals derived from benzene are styrene, phenol, and cyclohexane; manufactured products include plastics, resins, and other intermediates used in insecticides, drugs, dyes, and detergents.

Benzene is an excellent solvent, but because of its high toxicity its use has decreased greatly, with less toxic materials taking its place. Benzene has a high octane rating and is an important component of gasolines.

Benzene was first isolated by M. FARADAY in 1825 [1]; he isolated "bicarburet of hydrogen" from experiments on the pyrolysis of whale oil and other materials. A. W. HOFMANN and C. MANSFIELD of the Royal College of Chemistry also were working on liquids obtained from pyrolysis of coal. They developed commercial processes for obtaining benzene and other aromatics from coal tar between 1840 and 1850.

Prior to World War II, benzene was obtained mostly from coal. Catalytic processes developed in the petroleum industry have made petroleum the main source for benzene and related aromatics.

2. Physical Properties

The structure of benzene was postulated by A. KEKULÉ in 1865 as a planar, six-carbon ring with alternating single and double bonds and with one hydrogen atom attached to each carbon atom. Thermodynamic and spectral studies indicate that this simple picture cannot explain the true nature of benzene or related molecules. For instance, when benzene is formed from cyclohexane, less energy (151 kJ, 36 kcal) is required than is needed for the formation of three double bonds. The NMR spectrum of benzene shows the protons at lower field than would be expected for simple vinylogous ones. Spacing

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_475

Table 1. Physical data of $C_6 - C_8$ aromatics

	M_r	Density d_{20}	*mp*, °C	*bp*, °C	n_D^{20}	Critical density, g/cm^3	Critical pressure, bar	Critical temp., °C	Flash point, °C
Benzene [*71-43-2*]	78.11	0.87901	+ 5.533	80.099	1.50112	0.309	48.9	288.9	−11
Toluene [*108-88-3*]	92.13	0.86694	− 94.991	110.626	1.49693	0.291	40.7	319.9	4.0
Ethylbenzene [*100-41-4*]	106.16	0.8669	− 94.975	136.186	1.49588	0.286	37.4	346.4	15
o-Xylene [*56004-61-6*]	106.16	0.88020	− 25.182	144.411	1.50545	0.288	36.6	357.9	17
m-Xylene [*108-38-3*]	106.16	0.86417	− 47.872	139.104	1.49722	0.286	37.0	353	23.2
p-Xylene [*41051-88-1*]	106.16	0.86105	+ 13.263	138.351	1.49582	0.286	36.5	343	25

	Ignition temp., °C	Explosion limits in air, vol%		Heat of fusion, J/g	Heat of combustion, kJ/mol	Gross heating value, J/g	Specific electrical conductivity, cm$^{-1}\Omega^{-1}$
		lower	upper				
Benzene	595	1.4	6.7	125.9	− 3275.3	41932	3.8×10^{-14}
Toluene	552	1.4	6.7	71.8	− 3911.3	42454	$2-7 \times 10^{-13}$
Ethylbenzene	460	1.0	–	86.3	− 4568.6	43060	–
o-Xylene	502	1.0	6.0	128.2	− 4570.7	43079	$< 10^{-15}$
m-Xylene	562	1.1	7.0	109.0	− 4556.9	42949	–
p-Xylene	500	1.1	7.0	161.2	− 4559.8	42977	–

Table 2. Physical properties of benzene at different temperatures

Temp., K	Vapor pressure, kPa	Specific heat, liquid, J/g	Specific heat, vapor, J/g	Heat of vaporization, J/g	Thermal conductivity, W m^{-1}K^{-1}	Viscosity, mPa · s	Dielectric constant	Miscibility of water and benzene: % water in benzene	Miscibility of water and benzene: % benzene in water
223	0.103	1.097							
243	0.479	1.223							
263	1.950	1.491							
273	3.466	1.612	1.114	448.0		0.906			
283	5.965	1.683						0.034	
292							2.26		
293	9.97	1.708		435.0	14.58	0.654			
298							2.282	0.066	0.113
313	24.19	1.763		421.6	14.06			0.095	
323	35.82					0.499			
328								0.113	
330									0.248
333	38.4			408.2	13.54	0.398			
343	72.2	1.900							
353	99.7			394.4	12.02	0.336	2.17		
363	135	1.943							
373	178		1.260	379.3					
380									0.507
413	469.2			345.8					
433	765			329.1					
453	1015.3			310.2					
473	1419.6		1.394	286.8					
493	1935.6								
533	2579.4								
553	4369.8								
573			1.528						
673			1.617						

between the neighboring carbon atoms is constant (0.139 nm), which is less than single-bond spacing (0.154 nm), and more than, but close to, double-bond spacing (0.134 nm); C – H bond spacing is 0.108 nm. Benzene is thermally stable. It does not react as a compound having typical double bonds; however, it is more reactive than simple hydrocarbons.

Table 3. Binary azeotrope mixtures of benzene

Component B	Benzene, wt%	Azeotrope *bp*, °C
Water	91.17	69.25
Formic acid	31.0	71.05
Nitromethane	85.7	79.2
Methanol	60.4	58.34
Acetonitrile	60.0	73.7
Acetic acid	98.0	80.5
Ethyl nitrate	85.0	80.08
Ethyl alcohol	67.6	68.24
Acrylonitrile	53.0	73.3
Allyl alcohol	82.6	76.75
1,3-Dioxolane	15.0	74.0
Dimethyl carbonate	99.0	80.17
Isopropyl alcohol	66.7	71.92
Propyl alcohol	83.1	77.12
Diacetyl	45.0	79.3
2-Butanone	37.5	78.4
Dioxane	88.0	82.4
Ethyl acetate	6.0	76.95
Methyl propionate	48.0	79.45
Propyl formate	53.0	78.5
Isobutyl alcohol	90.7	79.84
sec-Butyl alcohol	84.0	78.8
tert-Butyl alcohol	63.4	73.95
tert-Amyl alcohol	85.0	80.0
Cyclohexene	85.0	79.45
Cyclohexane	55.0	77.5
Methylcyclopentane	10.0	71.4
n-Hexane	5.0	68.87
2,4-Dimethylpentane	48.4	75.2
2,3-Dimethylpentane	79.5	79.2
n-Heptane	99.3	80.1
2,2-Dimethylpentane	46.3	75.85
Trimethylbutane	50.5	76.6
2,2,4-Trimethylpentane	97.9	80.01

Tables 1,2,3,4,5, and 6 give important physical data for benzene and for related aromatic compounds.

3. Chemical Properties

Benzene is the basic unit of the aromatic class of compounds. It is one of the largest volume organic chemicals, with the United States being the largest producer. Benzene is the source of a variety of organic chemicals, many of which are intermediates for the production of a host of commercial products. Benzene is thermally stable and its formation is kinetically and thermodynamically favored at temperatures $\geqq 500\,°C$. Therefore elevated temperatures are required for its thermal decomposition or for condensation or dehydrogenation reactions to occur. As an example, at ca. 650 °C in contact with iron, lead, or over other catalytic materials, such as vanadium, *condensation* reactions take place to form diphenyl and other polyaromatic compounds.

Benzene is quite stable to *oxidation*, but under severe conditions it is oxidized to water and carbon dioxide. With a deficiency of air or oxygen under oxidizing conditions, partial decomposition and deposition of soot occur. Oxidation with air or oxygen in the vapor phase at 350 – 450 °C over a V–Mo catalyst produces maleic anhydride in yields of 65 – 70 % [2]. The use of pure oxygen offers no advantage over air [3]. Phenol can be obtained in low yield from the high-temperature oxidation of benzene with air [4].

Table 4. Ternary azeotrope mixtures of benzene

Component		Benzene,	B,	Azeotrope
B	C	wt%	wt%	*bp*, °C
Water	ethyl alcohol	74.1	7.4	64.86
Water	allyl alcohol	82.1	8.6	68.3
Water	*n*-propyl alcohol	82.4	8.6	68.48
Water	isopropyl alcohol	73.8	7.5	66.51
Water	2-butanone	73.6	8.9	68.9
Water	*sec*-butyl alcohol	85.55	8.63	69.0
Water	*tert*-butyl alcohol	70.5	8.1	67.3
Chloroform	1,2-dichloroethane	66.4	5.0	79.2
Water	acetonitrile	68.5	8.2	66.0

Table 5. Volume (or cubic) expansion coefficient of benzene

Temperature, K	$\gamma \times 10^6$
90 – 203	350
293	1237
273 – 303	1229
278 – 279	1160
279 – 283	1140
283 – 293	1180
203 – 313	1280
313 – 333	1380
333 – 353	1460
288 – 303	1060

Table 6. Boiling points and octane ratings of benzene and related compounds

Compound	*bp*, °C	Octane rating, RON∗
Benzene	80.1	98
Cyclohexane	81.4	83
Toluene	110.6	120
Ethylbenzene	136.1	107
o-Xylene	144.4	120
m-Xylene	139.1	115
p-Xylene	138.4	116
Cumene	152.4	113

∗ Research octane number

Substitution reactions of benzene are of primary importance. Depending on reaction conditions one or more of the hydrogen atoms in the benzene ring may be exchanged for nitro or sulfonic acid radicals, for amine or hydroxyl groups, or for atoms, such as chlorine or bromine. Products include phenol, nitrobenzene, chlorobenzene, benzenesulfonic acid, and others. For disubstituted benzene three isomers are possible:

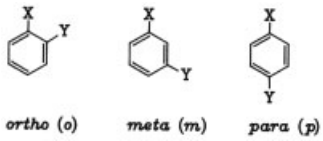

Other important reactions of benzene involve *addition* and include alkylation and hydrogenation. These reactions generally take place at elevated temperature and pressure, sometimes requiring active catalysts. Ethylbenzene results from the alkylation of benzene with ethylene in the presence of a catalyst, such as aluminum chloride [5] (→ Ethylbenzene). The reaction is carried out at 40 – 100 °C, gauge pressure < 0.7 MPa. With less active catalysts higher temperature and pressure are employed [6].

The production of cumene by the vapor-phase catalytic alkylation of benzene with propene is another important primary addition reaction of benzene. The reaction takes place at 200 – 250 °C and gauge pressure of 2.7 – 4.2 MPa over active catalysts, such as phosphoric acid or kieselguhr, with yields of ca. 95 % [7].

Hydrogenation is also an addition reaction. One of the most important is the hydrogenation of benzene at elevated temperature and pressure to produce cyclohexane. The reaction may be carried out in the liquid phase [8] or in the vapor phase at higher temperatures.

The dehydrogenation of ethylbenzene to form styrene is an important commercial reaction (→ Styrene). High temperature is necessary both kinetically and thermodynamically for this endothermic reaction to proceed.

Most of the reactions discussed in this chapter are primary reactions of benzene. A myriad of reactions and products rely on benzene, either directly or indirectly, as their source. Many of these are shown schematically in Figure 1.

4. Raw Materials

Benzene and its homologues, such as toluene and xylenes, are found in crude oils but in such small amounts that physical separation and recovery generally has not been economically feasible. Table 7 shows the aromatic, naphthenic, and paraffinic contents of the petroleum and naphtha fractions of Saudi Arabian crudes and those of a similar fraction of a United Kingdom crude oil [9], [10]. Because the naphtha fractions represent < 15 % of the total respective crudes, the percentages of benzene, toluene, and xylene (BTX) are quite low. The benzene contents in these fractions are ≦ 50 % of the total aromatic contents shown. However, the naphthenes, which are generally potential benzene precursors, are usually present in larger amounts than the aromatics.

With heavier crudes the aromatic contents of the light naphtha fractions may be somewhat greater. In synthetic crudes, such as those derived from tar sands, an increased concentration of aromatics in the light naphtha fraction can

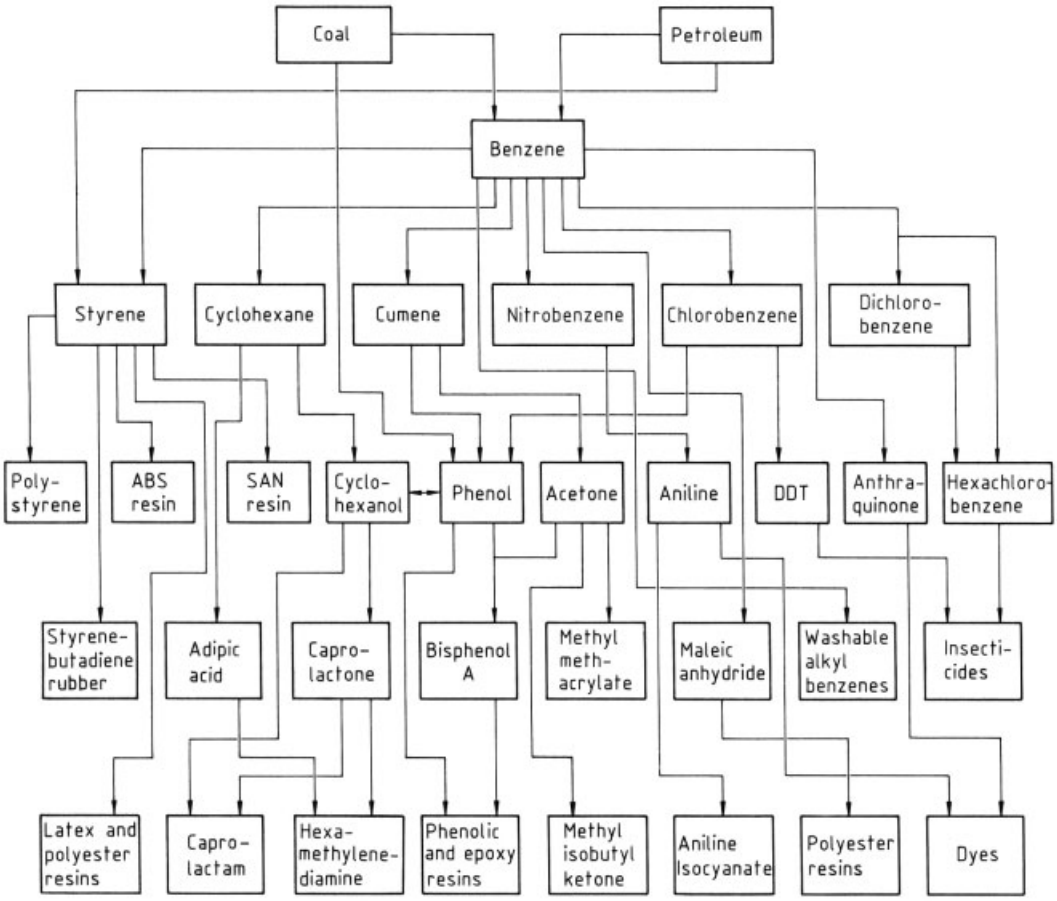

Figure 1. Products from benzene

Table 7. Analysis of crude oil fractions

Property	Saudi Arabian mixed	Thistle, U.K.	Saudi Arabian light
Crude gravity, °API	34 – 34.9	37.4	34.1
Sulfur, wt%	1.63	0.31	1.72
C_5 and lighter, vol %	1.6	3.3	1.5
Light naphtha, yield, vol% (debutanized)	13.7	13.23	9.3
T.b.p. ∗ range, °C	C_5 – 126 ∗∗	31 – 105	C_5 – 93 ∗∗
End point, °C	132		
Gravity, °API	72.5	71.3	75.0
Sulfur, wt%	0.02	0.04	0.03
Aromatics, vol%	5	3.1	2.4
Naphthenes, vol%	5	30.5	12.3
Paraffins, vol%	90	66.4	85.3

∗ T.b.p. is true boiling point distillation.
∗∗ $C_5 - t$ refers to C_5 plus content of the light naphtha.

be expected. Such may not be the case, however. Table 8 contains an analysis of the C_5 – 150 °C naphtha fraction from a conventional crude and that of a synthetic naphtha derived from Athabasca tar sands [11]. The naphtha from the tar sands has lower contents of aromatics and naphthenes than does the fraction from the conventional crude. However, when the higher naphtha fractions are compared the synthetic crude has a considerably higher aromatic content (32 versus 19 %), although the naphthene content is lower in the synthetic crude. The higher aromatic content of the heavier fraction can be explained on the basis of dealkylation or of ring opening of the higher molecular mass aromatics in the bitumen.

Table 8. Properties of naphthas from synthetic and conventional crudes

	Synthetic	Conventional
Boiling range	C_5 – 150 °C	
Density, g/cm^3	0.697	0.718
Sulfur, ppm	1	30
Hydrocarbons, vol%∗		
Aliphatics	69	43
Naphthenes	25	46
Aromatics	6	11
Benzene	1.1	1.4
Toluene	2.7	4.3
Xylenes	2.0	0.5

∗ By GC analysis

The commercial sources of benzene, as well as toluene and xylenes in primary products, are coal and petroleum. Historically benzene production has been a byproduct of the carbonization of coal to produce coke for the steel industry. More recently petroleum has become the primary source as catalytic and thermal methods have been developed for producing aromatic or related compounds. These materials are used to improve gasoline quality or as feedstocks to produce petrochemicals. For example, catalytic reformates, pyrolysis gasoline, and alkylated aromatics, such as toluene and xylenes, provide sources for the commercial production of benzene. Production methods are discussed in Chapter 5.

5. Production

Until World War II, the coal industry supplied most of the commercial benzene in the United States and other countries. The pyrolysis of coal to yield metallurgical coke produced byproduct gas and aromatic liquid from which benzene and related aromatics were recovered by extraction and distillation. The small concentrations of benzene, toluene, xylene, etc., in crude oil fractions made this potential source noneconomic.

New processes discovered and commercialized in the petroleum industry during the 1930s and 1940s and the greatly increased demand for aromatics over this period changed the situation. Petroleum became more important as a raw material for the production of benzene and its homologues in the United States. Today, petroleum is the main source of the BTX aromatics – benzene, toluene, and xylenes.

During the 1930s catalytic cracking made its debut, first as fixed-bed processing and later as moving-bed processes. Cracked light naphthas contained a considerable aromatic content in contrast to the low concentrations found in virgin light naphthas. During this period catalytic reforming of virgin naphthas also became a reality. Aromatics, such as benzene, were made by dehydrogenation of cyclohexane or by isomerization and dehydrogenation of methylcyclopentane. *n*-Hexane was converted to benzene in smaller yields and under modified reaction conditions.

5.1. Production from Coal

Light oils from the carbonization or coking of coal have long been sources of BTX (→ Coal Pyrolysis). The carbonization process also produces a C_5 and lighter fraction and a heavier coal tar fraction. In the coking of coal, the yields and composition of the light oils depend on the type of coal being processed and on the carbonization temperature. *Low-temperature coking* was formerly employed to produce gas for heating and lighting. *Low-temperature* (< 700 °C) *carbonization* was important, especially in industrialized Europe before the advent of natural gas, petroleum, and other forms of energy. The benzene content of the BTX is lower in this case than when coking is carried out at a higher temperature. Low carbonization temperatures produce a rather low-density, clean burning coke.

High-temperature carbonization was developed to produce a hard coke for use in blast furnaces of the steel industry. *High-temperature*

coking is done at $\geqq 900\,°C$. A typical composition of light oils produced in coke-oven operation is (vol%): benzene (63), toluene (13), xylenes (8), higher aromatics (10), nonaromatics (6) [8].

Typical yields of liquid products obtained from slot-type coke-oven operations are 11 – 15 L of light oil and 30 – 38 L of coal tar pitch per ton of coal charged [12]. About 1 % of ammonium sulfate and around 2800 L of coke-oven gas per ton of coke are produced. Coke and coke breeze yields are about 75 %.

Whereas the benzene content of light oils derived from coke-oven operation is $\geqq 65$ wt%, the concentration is much lower (ca. 45 %) in light oils derived from vertical retort operations. The typical, crude BTX material from coke oven operation mentioned previously contains considerable amounts of higher boiling aromatic and nonaromatic materials. These include around 5 % of material boiling below benzene, which may be designated C_5 and lighter. These compounds may include aliphatics, simple olefins, and unsaturates, such as cyclopentadiene. Carbon disulfide, another low-boiling impurity, may be present at ca. 0.5 %. It as well as thiophenes must be removed (see Chap. 6).

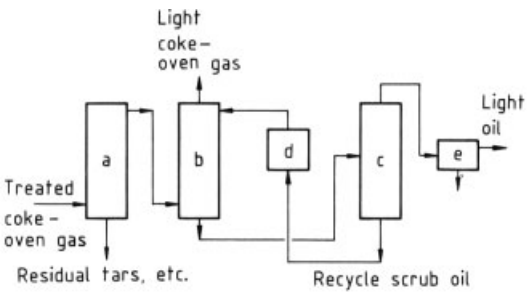

Figure 2. Recovery of light oil from coke-oven gas
a) Cooler and removal of residual tars and impurities; b) Light oil scrubber; c) Light oil stripper; d) Separation and cooling; e) Separator

The overhead gas and the liquids from the coke-oven operation pass through a separation system. The gas and BTX fractions are separated from the coal tars and pass overhead to a recovery system where the C_5 and lighter fraction compounds are separated overhead from the BTX. The BTX fraction is recovered from the C_5 and lighter gas by countercurrent absorption using a high-boiling (generally 300 – 400 °C) petroleum fraction. Subsequent recovery of the BTX from the absorbent oil is accomplished by steam stripping. The BTX overhead products are condensed and separated from the water. One version of the process is covered in a U.S. patent issued to Bethlehem Steel [13]. A simplified flow sheet common to several processes is shown in Figure 2.

Separation and Refining of BTX Fractions. The crude BTX as obtained from the coke-oven recovery system must be refined to remove impurities so that the different BTX components meet purity specifications. Distillation removes C_5 and lighter hydrocarbons, carbon disulfide, and cyclopentadiene. Treatment with sulfuric acid removes thiophene and other sulfur compounds. Further purification by hydrotreating removes unsaturates, sulfur compounds, and nitrogen impurities. Hydrodealkylation reactions may accompany the hydrotreating process. Finally, recovery of the desired aromatics may be accomplished by solvent extraction and close fractionation to separate the individual aromatic components.

Individual processes for these different operations are being used. These include several combination processes that have been developed where the different chemical reactions are included in the overall process. The patent literature contains many references to combination processes designed to process coke-oven light oils into high-purity benzene or into high-quality BTX components. Modifications of these processes are used for the recovery of naphthalene from the crude BTX [14], [15].

One combination process, the *Houdry-Litol process,* has been in commercial operation since 1964. In 1981 there were six commercial installations operating in various countries [16], [17]. In this process a prefractionation removes the C_5 and lighter fraction, and the bottoms containing C_9 and heavier hydrocarbons. The BTX cut is passed to a pretreat catalytic reactor. Reactive unsaturates, e.g., cyclopentadienes and styrene, are selectively hydrogenated over a catalyst, such as cobalt molybdate on alumina. Hydrogenation of some sulfur compounds also occurs. Temperatures of ca. 250 – 350 °C and pressures of 5 – 7.5 MPa are used. More active catalysts allow lower reaction temperatures.

A second reaction system operates at $\geqq 600\,^{\circ}C$ using a chromia – alumina catalyst. The reactions that take place include:

1) Desulfurization

$$CS_2 + 4\,H_2 \longrightarrow CH_4 + 2\,H_2S$$

2) Hydrocracking

$$C_6H_{12} + 3\,H_2 \longrightarrow 3\,C_2H_6$$

3) Aromatization

$$C_6H_{12} \longrightarrow C_6H_6 + 3\,H_2$$

or

$$C_6H_{14} \longrightarrow C_6H_6 + 4\,H_2$$

4) Hydrodealkylation

$$C_6H_5CH_3 + H_2 \longrightarrow C_6H_6 + CH_4$$

5) Dehydrogenation of naphthenes

$$C_6H_{12} \longrightarrow C_6H_6 + 3\,H_2$$

A modified Litol process is more adaptable for complicated coke-oven aromatic materials [17]. In the Litol and related processes the readily polymerizable unsaturates should not exceed 2 – 3 % and nonaromatics should be limited to a similar amount. In the hydrodealkylation step about 50 % of the toluene and 75 % of the xylenes are dealkylated to benzene. After treatment of the Litol reactor effluent with clay, benzene is recovered and the remaining toluene separated by fractionation.

A small amount of benzene is available from the distillation of coke oven tars or tars from other coal carbonization processes. The low-boiling fractions containing BTX materials are treated with base to remove tar acids. Purification by hydrodealkylation produces benzene. Generally, however, the light oil fraction is 1.0 % or less of the coal tar [18]. Hence, this source is not of major importance in the production of benzene. Coal tars are more important as sources of pitches and of naphthalene, anthracene, or other condensed aromatic materials.

5.2. Production from Petroleum

Catalytic cracking and catalytic reforming of petroleum, adventing during the late 1930s and the 1940s, became a new source for the production of benzene, toluene, xylenes, and higher aromatics. Petroleum rather than coal began to emerge as the source of BTX materials, especially in the United States. In the United Kingdom and Europe the transition from coal to petroleum came around 1960.

Toluene was in great demand during World War II for the production of the explosive, trinitrotoluene (TNT). However, production of toluene from virgin naphthas and later from catalytically cracked and catalytically reformed naphthas was expensive. One of the early catalytic cracking processes was developed by Houdry Corp. [19], [20]. The silica – alumina catalyst used in this process was deposited in several fixed catalyst beds. The catalyst was regenerated frequently with air to burn deposited coke. Another process developed by Mobil Oil used a moving-bed silica – alumina catalyst [21]. The third type of process, developed by Exxon [22] and others, is known as the fluid process (FCC), and is now used exclusively.

Although the catalytic cracking processes were developed primarily for increasing the yield and octane rating of the motor fuels produced, the aromatic content of the gasolines was quite high [23] and so these processes provided a source for petrochemical benzene.

5.2.1. Benzene from Reformate

Catalytic reforming provided another petroleum source for BTX aromatics. Installation of these processes took place in the 1940s and early 1950s, at a time when the demand for benzene was expanding rapidly because of development of processes requiring it to manufacture resins, detergents, synthetic fibers, and other chemicals. Chemical reactions involved in reforming include dehydrogenation of naphthenes to aromatics, or isomerization of alkylnaphthenes followed by dehydrogenation. Likewise, paraffins may be dehydrocyclized to aromatics. The reaction is quite slow and process conditions must be adjusted to make it economic. A small amount of dealkylation of alkylated aromatics may occur.

Early catalytic reforming units used base-metal catalysts, generally molybdena on activated alumina. These processes operated on a short process cycle with frequent catalyst regeneration. The hydroforming process [24] operated at elevated temperatures of about 520 °C and at pressures of 2.5 – 7 MPa with hydrogen recycle. This process converted naphtha fractions into high-octane gasolines. The 10 % molybdena on alumina catalyst required regeneration every 8 – 16 h. The process consisted of two parallel systems of two reactors in series. One pair of reactors could be in use while the catalyst in the other pair was being regenerated. Later catalytic reforming processes using base-metal catalysts employed either moving catalyst beds with the catalyst in granular form or used fluidized beds with a finely divided catalyst of small particle size.

Various catalytic processes using platinum on alumina and platinum with modifiers on alumina have been developed [25]. These methods may have process cycles as long as six months or much shorter cycles, depending on the feedstock, the process conditions, and the severity of operation. Pretreatment of the naphtha feedstock by catalytic hydrotreating is required to remove sulfur compounds. Commercial catalytic reforming processes include: catalytic reforming (Institut Français du Pétrole), magnaforming (Engelhard Industries), platforming (UOP), powerforming (Exxon Research & Engineering), rheniforming (Chevron Research Co.), and ultraforming (Standard Oil of Indiana).

Operating conditions cover a considerable range: temperatures vary from 425 to 525 °C and gauge pressures from 0.7 to 3.5 MPa. Hydrogen recycle rates depend on whether long process cycles or whether frequent or continuous catalyst regeneration is employed. Naphtha feedstocks vary but generally are in the gasoline boiling range. Narrower boiling range stocks may be used if aromatics are the desired end products. Regenerative-type products generally allow lower pressures, less recycle hydrogen, and higher severity operation. Naphthene conversions approach 100 %. Cyclization of paraffins is much lower. For example, for benzene production from C_6 and higher stocks, methylcyclopentane is isomerized to cyclohexane and dehydrogenated to benzene. Conversion of *n*-hexane to benzene is limited.

Catalytic reformer units may be operated at relatively high pressures with a series of fixed-bed reactors, the catalyst being regenerated at defined intervals. Units designed for higher conversion operate at lower pressure with more frequent catalyst regeneration. Shorter cycles, swing reactors, or continuous regeneration are employed. The version of platforming [26] shown in Figure 3 employs continuous catalyst regeneration. Another design developed by Standard Oil of Indiana uses a swing reactor in which the fixed-bed catalyst in one reactor is regenerated and brought back onstream, with the catalysts in each of the remaining reactors being regenerated in rotation. Gauge pressures of ca. 0.7 – 2.4 MPa are employed in the different processes. For gasoline production, naphthas used have boiling ranges of around 80 – 210 °C. If reforming is carried out for the production of aromatic charge stocks, the composition and boiling range of the naphtha may be altered. The reformate may have RON clear octane rating as high as 100 and may contain up to 70 vol% aromatics. Typical compositions of reformate at different levels of reforming severity are shown in Table 9.

Over the years, the process has been made more effective. Reforming catalysts have been modified. Water in small amounts was found to enhance benzene yields with certain catalysts when producing BTX [27]. Higher xylene yields have resulted from operating at increased space velocities.

5.2.2. Aromatics from Pyrolysis Gasoline

Pyrolysis gasoline (dripolene), produced as a byproduct from the high-temperature, short-residence time cracking of paraffin gases, naphthas, gas oils, or other hydrocarbons used to produce ethylene, is another excellent source for BTX. The quantity of byproduct pyrolysis gasoline produced is a function of feedstock and operating conditions [28], [29]. Although the amount may be small with ethane and other gaseous paraffins as the ethylene charge stocks, it increases when heavier charge stocks are used. On a once-through basis, amounts may be 20 % in the high-severity cracking of a medium range naphtha. As the boiling point of the feed increases, a larger yield of toluene and xylene, rel-

Table 9. Reformer cuts and reformate composition (amounts in wt %)

	RON ∗ clear	Benzene	Toluene	Xylene	C_9 aromatics	Non-aromatics
Total reformate	95	5.5	19.5	27.1	13.1	34.8
C_6						
Aromatic fraction		8.4	29.9	41.6	20.1	
BTX fraction		10.6	37.4	52.0		
Total reformate	98	6.6	22.1	29.5	15.3	26.5
C_6						
Aromatic fraction		9.0	30.0	40.2	20.8	
BTX fraction		11.3	38.0	50.7		
Total reformate	101	7.8	25.9	32.0	16.3	18.0
C_6						
Aromatic fraction		9.5	31.6	39.0	19.9	
BTX fraction		11.8	39.4	48.8		

∗ Research octane number: indicates relative degree of reforming severity

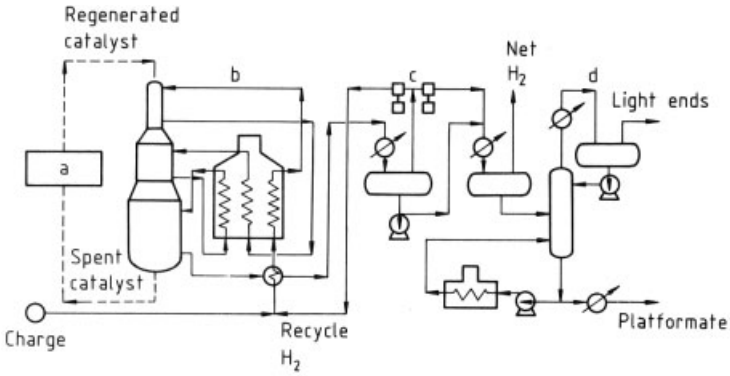

Figure 3. Continuous platforming process
a) Catalyst regeneration; b) Reactors; c) Product separation; d) Stabilizer
Reproduced with permission from [26]

Table 10. Cracking severity and product composition from pyrolysis gasoline (mass fractions in %)

Composition	Moderate severity			High severity		
based on	Pyrolysis gasoline	Aro-matics ∗	BTX∗	Pyrolysis gasoline	Aro-matics ∗	BTX∗
Nonaromatics	35.3			28.2		
Benzene	23.7	36.6	39.4	43.2	60.2	62.2
Toluene	20.2	31.2	33.6	17.4	24.2	25.0
Xylenes including ethylbenzene	12.6			3.5		
Styrene	3.6			5.4		
Total C_8 aromatics	16.2	25.1	27.0	8.9	12.4	12.8
Higher aromatics	4.6	7.1		2.3	3.2	

∗ Including styrene

ative to that of benzene, is produced. Table 10 shows a typical composition of pyrolysis gasoline [30]. The BTX content of the pyrolysis gasoline is > 60 %. Because of the high temperature necessary for ethylene production, the pyrolysis gasoline contains other unsaturates and diolefins. Diolefin content is typically $\geqq 5$ % as is the total content of cyclic olefins and cyclic diolefins [31].

Table 11 gives an analysis of pyrolysis gasolines by carbon groups [30], [32]. Pyrolysis gasoline is quite unstable because of the appreciable amount of diolefinic material. Lower boiling diolefins may be removed by distillation although their tendencies to polymerize and depolymerize may cause difficulties. Therefore a two-stage hydrotreating is generally employed. In the first stage a mild selective hydrotreating converts diolefins to olefins. If this gasoline is to be used as a fuel, further treatment may not be necessary. However, if the gasoline is a source of aromatics, a second stage is employed to saturate olefins and to remove residual sulfur.

5.3. Methods of Increasing Production

Depending on demand, benzene production from other aromatic compounds can be increased, particularly from the BTX fractions obtained from catalytic reformate, pyrolysis gasoline, or light oils from coal carbonization. Among the methods used are hydrodealkylation, disproportionation, or combination processes.

5.3.1. Hydrodealkylation

Hydrodealkylation of alkyl aromatics is a source of benzene. Toluene is the usual charge stock, although higher alkylated aromatics may be used:

$$C_6H_5CH_3 + H_2 \xrightarrow{\text{Heat}} C_6H_6 + CH_4$$

With more highly alkylated benzenes, the reaction proceeds stepwise:

$$C_6H_4(CH_3)_2 + H_2 \xrightarrow{-CH_4} C_6H_5CH_3 \xrightarrow{H_2} C_6H_6 + CH_4$$

Other alkylbenzenes, such as ethyl- and propylbenzene, dealkylate in a single step and form the corresponding alkanes. Both catalytic and thermal processes are used. The catalytic processes, e.g., Detol (Air Products & Chemicals) [33] and Hydeal (UOP) [34], operate at 575 – 650 °C and 2.5 – 6 MPa. The temperature of this exothermic reaction is controlled by recycling cold hydrogen.

Thermal dealkylation processes, such as HDA (Arco and Hydrocarbon Research) and THD (Gulf Oil), operate at higher temperatures than the catalytic, but yields and reaction systems are similar in both types. For a comparison of the different processes see [35].

The amount of toluene in BTX from catalytic reformate is greater than that of benzene. Whether or not hydrodealkylation is used is related directly to the demand and price of benzene relative to toluene. If the relative demand for the two products changes, the dealkylation units may be placed on standby.

Detol Process [33], [36]. This catalytic dealkylation process is illustrated in Figure 4. Benzene is produced from a feed consisting of toluene or mixtures of toluene and other alkylated benzenes. The alkylated aromatics and hydrogen pass at elevated temperature and pressure over a fixed-bed dealkylation catalyst, generally dispersed in more than one reactor in series. Heat exchangers cool the reactor products and condense benzene, unreacted toluene, and the heavier unreacted alkylated benzenes. A high-pressure flash drum separates recycle hydrogen and product gas. The gas is split into streams for fuel gas, for hydrogen purification if needed, and for recycle hydrogen.

The product gas from the flash drum is condensed and the liquid pumped to a stabilizer, where remaining gas and low-boiling hydrocarbons are driven off and utilized as fuel gas. The bottoms from the stabilizer are clay treated and passed to a column where the benzene is distilled overhead. Unreacted toluene and heavier aromatics are recycled. The benzene produced in yields of around 99 mol% is highly pure, typically 99.95 %, with a freezing point of 5.45 °C. Minor side reactions, such as hydrocracking and hydrogenation, occur and produce small amounts of light hydrocarbons and also eliminate sulfur compounds (e.g., thiophene). Catalyst cycle life in these dealkylation processes is long, as it is in other high-pressure reactions carried out in the presence of hydrogen. As of 1981 twelve commercial Detol plants had been licensed.

Hydeal Process [34], [37]. This process was developed jointly by Ashland Oil and UOP. The process is similar to the Detol one described previously. Toluene alone or in mixtures containing xylenes and other alkylated benzenes is

Table 11. Composition of pyrolysis gasoline according to carbon groups

Substances by type (mass fractions in %)	Groups					
	C_5	C_6	C_7	C_8	C_9+	Total
Diolefin	8	4	1	1	0	14
Monoolefin	2	2	1	3	0	8
Saturates	1	1	1	2	0	5
Aromatics	0	34	12	12	12	73
Aromatics by type	0	81.9	83.3	66.7	100	73

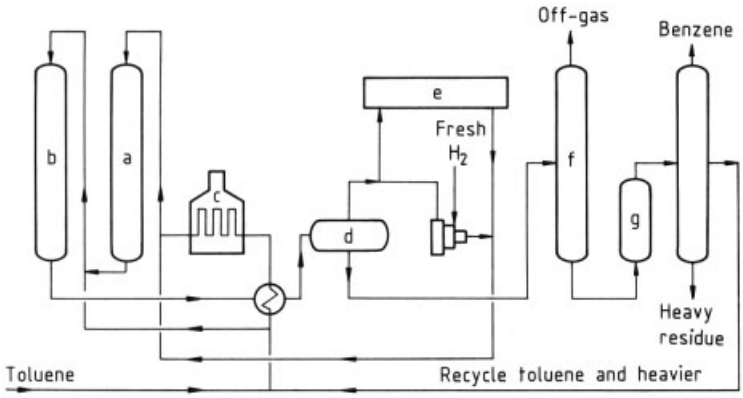

Figure 4. Detol process
a) Reactor I; b) Reactor II; c) Furnace; d) Separator; e) Hydrogen purification; f) Stabilizer; g) Clay treater

charged to the catalytic reactor, which contains a chromia – alumina catalyst. Hydrogen is added along with the alkyl aromatics. Nitration-grade benzene is obtained in yields of ca. 98 mol%. Unreacted aromatic charge material from the once-through operation is recycled to obtain practically complete conversion. The process is also used to hydrodealkylate alkylnaphthalenes.

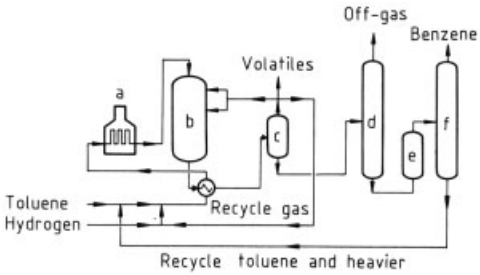

Figure 5. HDA process
a) Heater; b) Reactor; c) Separator; d) Stabilizer; e) Clay treater; f) Distillation

HDA Process [38], [39]. This process, developed by Arco and Hydrocarbon Research, operates at 600 – 660 °C and at gauge pressures of 3.45 – 6.9 MPa (see Fig. 5). Operation is similar to that of the catalytic processes, with benzene yields of 99 mol%. The small amounts of diphenyls formed are recycled and a low-equilibrium concentration is maintained, permitting high benzene yields.

THD Process [40]. This Gulf process is similar to the other thermal processes.

MHC Process [41], [42]. Mitsubishi Petrochemical Co. has described a thermal dealkylation process known as MHC. Methods are emphasized for controlling condensation reactions and coke formation. The following advantages are claimed for the MHC process. Feedstocks containing up to 30 % of nonaromatics can be handled without resorting to aromatic extraction and fractional distillation steps. The process will operate on low-purity hydrogen, thus reducing

the amount of makeup hydrogen required. Benzene of 99.95 % purity is readily achieved. Figure 6 is a flow sheet of the MHC process.

5.3.2. Disproportionation

Processes have been developed to disproportionate or transalkylate alkylated aromatics to produce benzene and alkylated benzenes. For example, the Tatoray process (UOP and Toray Industries, see Fig. 7) [43–45] produces benzene and equilibrium xylenes from toluene or from a mixture of toluene and C_9 aromatics:

$$2\,C_6H_5CH_3 \longrightarrow C_6H_6 + C_6H_4(CH_3)_2$$

$$C_6H_5CH_3 + C_6H_3(CH_3)_3 \longrightarrow 2\,C_6H_4(CH_3)_2$$

The reaction proceeds with a high H_2 recycle rate over a noble-metal or rare-earth catalyst. Operating conditions are 350 – 525 °C and 1 – 5 MPa.

Benzene of high purity (*fp* 5.45 °C) and xylenes containing $< 10^{-9}$ parts of saturated hydrocarbons are produced. When the feedstock is limited to C_9 and C_{10} aromatics, a mixture of benzene, toluene, and xylenes results. Typical yields with toluene as the charge are about 97 %. Yields for both a pure toluene feed and a mixed feed of toluene and C_9 + aromatics are shown in Table 12.

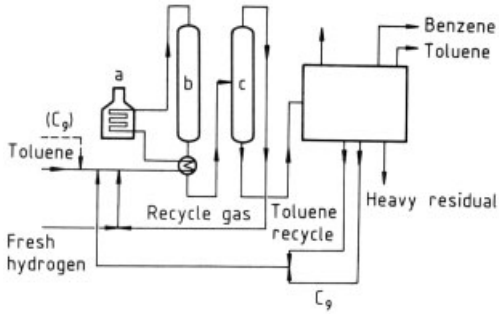

Figure 7. Tatoray process for combination toluene disproportionation and C_9-aromatic transalkylation
a) Preheater; b) Reactor; c) Separator

Table 12. Tatoray process yields, wt % [43]

Products ∗	Aromatic Feed	
	70 % Toluene 30 % C_9+	Pure toluene
Benzene	29.6	41.2
Toluene		
Ethylbenzene	2.3	
Xylenes	61.9	55.9
C_{10}+		1.0

∗Balance of products are low-boiling hydrocarbons and C_{10}+aromatics

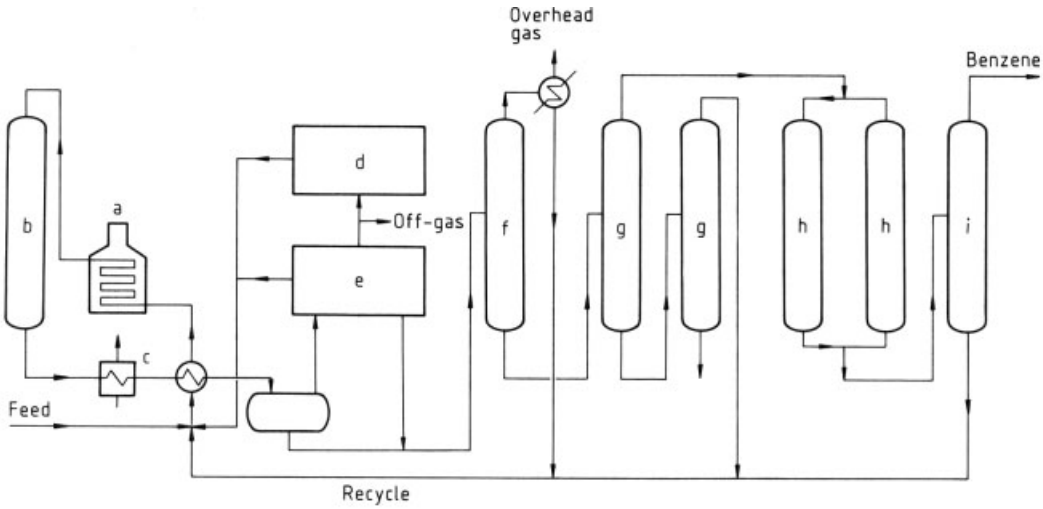

Figure 6. MHC process
a) Heater; b) Reactor; c) Waste-heat boiler; d) Hydrogen regeneration; e) Aromatic recovery; f) Fractionation; g) Vacuum columns; h) Clay treaters; i) Benzene column

Four commercial units were in operation in 1981 and others were being designed or built.

5.3.3. Combination Dealkylation Processes

Several combination dealkylation processes have been developed for producing benzene from impure BTX feedstocks. The Houdry Division of Air Products developed the *Pyrotol process* [46], [47] by a combination of hydrogenation and dealkylation (Fig. 8).

Essentially the process involves fractionation of the feed to remove C_5 minus and C_9 plus material from the crude BTX. The $C_6 - C_8$ is vaporized and passes to a catalytic pretreat reactor for selective hydrogenation of diolefins, cyclic diolefins, and styrene. The effluent from the pretreat reactor is subsequently charged to the Pyrotol reactors, where aromatics are dealkylated to benzene. Other reactions include desulfurization and hydrocracking of nonaromatics. Unreacted toluene and heavier aromatics from the first pass are recycled. The benzene product contains less than 0.5 ppm of thiophene and has a freezing point of at least 5.47 °C. Table 13 contains typical yield data. As of 1982 thirteen plants had been licensed.

A somewhat similar process was developed by the Houdry Division of Air Products and by Bethlehem Steel. The *Litol process* has been in operation in the United States and other countries. It produces pure benzene from aromatic light oil obtained from coal carbonization [48]. This process, as the Pyrotol process, utilizes two catalytic reaction stages. After prefractionation of the raw light oil to remove C_5 and lighter fractions overhead and C_9 and heavier as bottoms, the $C_6 - C_8$ fraction is vaporized.

At the first stage diolefins and styrene are hydrogenated. The effluent then contacts the Litol chromia – alumina catalyst at 500 – 600 °C, where desulfurization, hydrocracking, and dealkylation occur. Benzene purity of 99.97 % with freezing point of 5.5 °C and a thiophene content of less than 0.5 ppm are typical.

In 1968 a modified Litol process was developed [49]. This modified process is designed to include a heavier fraction of the light oil, mostly C_9s and heavier, along with the secondary light oil ($C_6 - C_8$) normally used in the process. The lighter fraction typically contains around 70 % benzene, 20 % toluene, and 6 % xylenes and ethylbenzene. The primary, or heavier fraction of the light oil, has a typical boiling range of 180 – 260 °C and is composed mainly of C_9 and heavier aromatics. Pilot plant operation indicates that a 50 – 50 mixture of the lower and higher boiling range light oils can be processed.

Mitsubishi Petrochemical Co. described a process [50] for producing benzene from pyrolysis gasoline by a combination method similar to those discussed previously. After a two-stage hydrogenation a solvent was used to extract the aromatics. Alternatively, the hydrogenated material was reacted in Mitsubishi's MHC process [42], [51], [52]. Advantages of the process are low-temperature liquid-phase operation of the first stage with no need for hydrogenated product recycling, and a reactor design incorporating an improved method for controlling temperature of the exothermic reactions.

Mobil Chemical Co. has developed an aromatics processing catalyst and a process for the liquid-phase disproportionation of toluene to form benzene and xylenes (Fig. 9) [53], [54]. This *low-temperature – disproportionation process* (LTD) employs a zeolite-based catalyst of high activity which permits the reaction to proceed at temperatures as low as 260 °C; gauge pressure is 4.5 MPa. No hydrogen recycle is required in this liquid-phase process. An hourly mass-space velocity is standard. A catalyst life of 1.5 years is claimed with infrequent regenerations. With fresh or regenerated catalyst the initial reactor temperature is 260 °C. As the catalyst slowly deactivates, the temperature is raised in increments to 315 °C; then the catalyst is regenerated by burning accumulated coke.

After cooling the reactor effluent passes to three distillation columns arranged in series. In the first column non-aromatics plus a small amount of benzene are taken overhead. This benzene may be recovered by extraction if deemed necessary. Benzene is taken overhead as product in the second column and toluene is distilled overhead for recycle in the third column. Xylenes and heavier aromatics are withdrawn as bottoms from this last column. The C_9 + aromatics may be recycled. Toluene conversion on a once-through basis is around 45 % and the balance is recycled to extinction. Table 14 contains

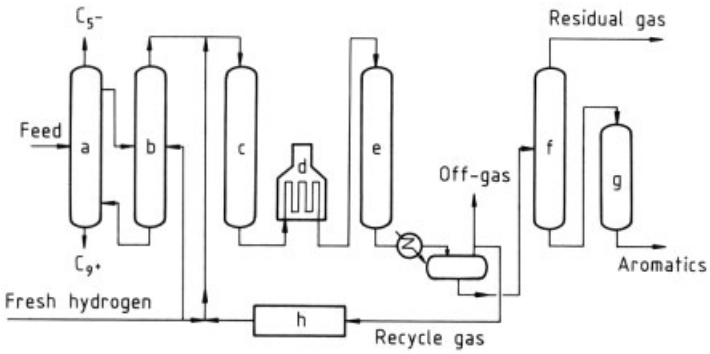

Figure 8. Pyrotol process
a) Distillation; b) Vaporizer; c) Prereactor; d) Heater; e) Pyrotol reactor; f) Stabilizer; g) Clay treater; h) Hydrogen purification

Table 13. Pyrotol yield data and material balance, wt %

Component	Dripolene raw charge	Fresh feed to pyrotol reactors	Benzene product
C_5s	7.1	0.2	
Benzene	48.3	54.4	99.92
Cyclohexane	2.4	2.7	
Toluene	16.2	18.3	0.08
Other C_7s	8.7	9.8	
Ethylbenzene	1.8	2.0	
Xylenes	9.2	10.2	
Other C_8s and heavier	6.3	2.4	
Total	100.0	100.0	100.0
wt% of charge	100.0	88.63	68.88

a typical wt% composition of the toluene fresh feed and the end product.

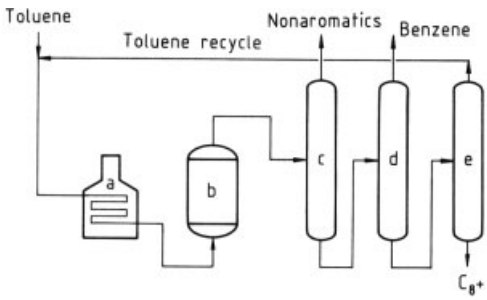

Figure 9. Mobil LTD process
a) Furnace; b) Reactor, c) Nonaromatics column; d) Benzene column; e) Toluene column

5.3.4. Benzene from Other Sources

The development of energy from solid fuels, such as coal, oil shale, and tar sands, may provide other potential sources for the production of benzene [55]. For example, if coal is developed as a source of liquid fuels, it has potential for benzene production [56–59]. Some of the processes being investigated include noncatalytic liquefaction, SRC-I and the related SRC-II. Others are COED, Exxon Donar Solvent, indirect liquefaction via the Fischer-Tropsch method, and the LCFFC catalytic process by Lummus. Coal gasification also may yield naphthas that can serve as charge material for benzene production.

Production of liquid fuels from coal has been known since 1911 when Bergius hydrogenated coal. Hydrogenation and Fischer-Tropsch plants were a source of liquid fuels in Germany during the 1940s. Two commercial processes for the production of liquid fuels from coal are the Sasol plants in South Africa [55] and the Mobil methanol-to-gasoline plant now under construction in New Zealand [60–62]. A third commercial process, using tar sands as raw material for the production of liquid fuels, has been described

Table 14. LTD fresh feed and product composition, wt % [54]

Component	Feed	Product
Nonaromatics		0.2
Benzene		43.8
Toluene	99.7	0.1
Ethylbenzene	0.1	0.2
p-Xylene	0.1	12.2
m-Xylene	0.1	27.3
o-Xylene		11.5
Trimethylbenzenes		4.0
Other C_9 + aromatics		0.7
Total	100.0	100.0

[55]. Other potential sources of liquid fuels include shale oil developments, biomass conversion, and the production of hydrocarbon liquid fuels from ethanol.

The yields of chemical feedstocks and gasolines from coal conversion depend on the process or combination of processes employed. For example, a BTX yield of 2.5 %, based on coal charged, was projected using a combination of processing including the Fischer-Tropsch route [63]. Hydroliquefaction followed by processing of the naphtha produced can increase BTX. Another study indicates a liquid yield of ca. 50 %. Distillate yields from other liquefaction processes, such as SRC-II and H-coal, are ca. 50 %, with an aromatics content of ca. 25 %. Fischer-Tropsch gasolines generally have a lower aromatic content than those from liquefaction processes. Unless secondary aromatization is carried out, the benzene content of distillates from Fischer Tropsch processes is generally lower than that found in petroleum derived gasolines.

Thus carbon-containing materials, such as coal, tar sands, and oil shale, are potential raw materials for benzene production. However, their commercial use is not imminent as long as petroleum is in ample supply at a reasonable price. Exceptions are light oils from coal carbonization and possibly coal conversion processes, such as Sasol and others in operation or under construction.

6. Separation and Purification

The separation and purification of benzene from crude BTX fractions or the separation of the different BTX components is accomplished by different methods and processes, such as those outlined in this chapter. The BTX fractions from catalytic reforming are the easiest to handle in that most of the objectionable impurities have been removed during reforming. Light oils from coal carbonization and pyrolysis gasoline from high-temperature manufacture of ethylene need additional treatment to remove impurities, such as dienes, olefins, and sulfur compounds, as well as small amounts of nitrogen and oxygen before proceeding with the recovery and separation of the aromatic components.

6.1. Removal of Dienes, Olefins, and Sulfur Compounds

Pyrolysis gasolines need pretreatment because of appreciable diene content. A two-stage hydrotreating process is generally employed. The dienes are hydrogenated to olefins in a mild first stage. In a second stage olefins are saturated and sulfur compounds are removed under more severe conditions. However, conditions are mild enough to avoid saturation of aromatics. Both base-metal and noble-metal catalysts are employed in diolefin removal. Nickel molybdate and cobalt molybdate generally are used in olefin and sulfur removal. Temperatures of 30 – 175 °C and pressures of 2 – 6 MPa are typical for diolefin operation. Noble-metal catalysts generally require lower temperatures than base-metal catalysts. Space velocities range from 2 to 10 vol of feed per volume of catalyst per hour, with higher values permitted with noble-metal catalysts.

Shell Development Co. has described a process for diene removal or, alternately, a two-

Table 15. Hydrogenation of pyrolysis gasoline (Shell Development)

Conditions	First stage	Second stage
Temperature, °C	80 – 130	230 – 380
Pressure, MPa	6	4.5 – 6.5
Recycle ratio	3 – 5	0.2
Space velocity, t fresh feed per m^3 cat./h	0.8 – 1.5	1.3
H_2 makeup, m^3/t (STP) fresh feed	50 – 100	50 – 125
Recycle gas, m^3/t (STP) fresh feed	200 – 500	450 – 1200

Properties	Feed	Product	Feed	Product
Maleic anhydride number mg/g	100 – 250	5 – 15	5 – 15	0
Bromine number g/100 g	60 – 100	25 – 80	25 – 80	1
Sulfur, ppm	100 – 400	100 – 400	100 – 400	0.5

stage process for diene removal followed by hydrogenation of sulfur and olefin compounds [64]. Diolefin hydrogenation is carried out in the first-stage reactor in a liquid flow with a special nickel-containing catalyst. Temperature control of the exothermic reaction is maintained by liquid product recycle. No desulfurization takes place in the first stage. If desired, olefins and sulfur may be removed in a second-stage hydrogenation in the presence of a nickel molybdate catalyst. Liquid product recycle may be used if necessary. Desulfurization and olefin saturation are nearly complete. Typical operating conditions when processing a C_5 – 200 °C pyrolysis gasoline charge are shown in Table 15.

A somewhat similar process has been described by the Institut Français du Pétrole. Reference [65] gives data for the hydrotreating of a C_5 – 205 °C pyrolysis gasoline. This is a one-stage process for the selective hydrogenation of diolefins or can be done as a two step process if complete hydrogenation of diolefins and olefins plus desulfurization is desired. Table 16 shows a partial listing of results.

In 1982, 65 units were in operation and others were under construction [65]. Several other processes of this type have been developed, including ones by British Petroleum [66] and Bayer.

Clay treating was formerly used for removal of diolefinic material and has been described in several publications. The process is outmoded, however, and is rarely used because of poor efficiency and disposal problems.

6.2. Close Fractionation

Precise fractionation may be used in some cases to separate nonaromatics from benzene or from other aromatics if the concentration of nonaromatics is not too great. However, some of the nonaromatics, such as cyclohexane, form azeotropes with benzene and thus cannot be completely separated by close fractionation at atmospheric pressure. In some cases distillation under elevated pressure will change the respective vapor – liquid equilibria so that separation may be improved.

6.2.1. Azeotropic Distillation

Naphthenes and paraffins are separated from aromatics by azeotropic distillation. Olefins in the charge mixture are generally detrimental to efficient separation. Usually the light ends from the crude BTX material are removed, followed by distillation of the benzene fraction from the toluene-xylene mixture. Examples of azeotrope formers include acetone and methanol with added water. Acetone is particularly suitable for the benzene fraction.

The azeotropic mixtures of nonaromatics and promoting agent are distilled. The nonaromatics are recovered by breaking the azeotropic mixture with water. In the benzene application, benzene is removed as a bottom product, washed, and distilled. Methanol is effective as an azeotrope forming agent in the treatment of the toluene – xylene fractions. By the azeotropic method, benzene with *fp* 5.5 °C is recovered in 98 % yield. Pure toluene is separated in 95 %

Table 16. Hydrogenation of C_5 – 205 °C pyrolysis gasoline (Institut Français du Pétrole)

	Feed-stock	After first stage	After second stage
Specific gravity	0.815	0.813	0.860
Diene value	30	1	≈0
Bromine number g/100 g	75	56	0.1
Sulfur, ppm	400	400	0.5

yield. Generally the process is limited to materials containing < 12 % nonaromatics. Azeotropic distillation has found limited application but is not as effective as other methods, such as extractive distillation and liquid-liquid extraction, in the isolation of high-purity aromatic compounds.

6.2.2. Extractive Distillation

Extraction and extractive distillation are commonly used to separate aromatics from crude BTX feeds. Extractive distillation is distinguished from normal distillation in that it is carried out in the presence of a high-boiling solvent having great solubility for aromatics and little or no solubility for paraffins and naphthenes. The solvent enters the distillation column near the top and above the point at which the BTX feed is introduced. The nonaromatics are distilled from the top of the column and the aromatic-rich solvent is withdrawn as a bottoms product. The aromatics are then distilled or stripped from the solvent in a second column and the lean solvent is recycled. Separation of the different aromatic components is carried out in subsequent distillation steps. Solvent to BTX feed ratios are generally in the order of 2 : 1 to 6 : 1 and conditions vary according to the nature of the solvent, column efficiency, and properties of the feedstock.

Extraction and extractive distillation methods as well as liquid – liquid extraction methods depend upon the use of very selective polar substances. Most of these compounds or solvents are high-boiling but some, such as SO_2, are not.

6.2.3. Liquid – Liquid Extraction

Liquid – liquid extraction in itself or in conjunction with other separation processes is the most widely used method for the isolation of aromatic fractions. All these processes involve the extraction of aromatics from nonaromatic material by the use of a polar solvent having a high-selective affinity for the former compounds.

The solvent, which is generally a high-boiling polar liquid, enters the top of the extractor vessel. Various designs of extractors have been used, including packed columns, sieve tray types, and others. A rotating disk contactor (RDC) has been used in some processes. The crude BTX or other aromatic-containing material enters near the middle of the extractors and thus contacts the solvent countercurrently in the column, or intimate contact is effected by other means. The nonaromatic raffinate leaves at the top of the extractor, whereas the aromatic-rich solvent is withdrawn at the bottom and passes to a distillation column, where the aromatics are taken overhead. The denuded solvent is then recycled to the extractor. The overhead aromatic stream is distilled to separate benzene and other aromatic components. The fundamentals of liquid – liquid extraction, extractive distillation, and various methods of aromatic recovery by extraction are discussed in [67], [68].

6.3. Processes for Aromatic Separation

Numerous methods and processes have been proposed and developed for the separation of pure aromatics from BTX feedstocks [69], [70]. The following are discussed in this article:

Process	Solvent
Edeleanu	sulfur dioxide
Udex	diethylene glycol and others
Sulfolane	sulfolane
Arosolvan	1-methyl-2-pyrrolidone
IFP	dimethyl sulfoxide
Formex	*N*-formylmorpholine
Morphylane	*N*-formylmorpholine

Edeleanu Process. This process uses as solvent liquid sulfur dioxide [*7446-09-5*], SO_2, M_r 68.06. It was developed in 1908 primarily for the

removal of aromatics from kerosine lamp oil and is of historical interest only. Liquid sulfur dioxide is effective in extracting of aromatics but is now used only for special purposes. This process is not competitive for aromatic removal in stocks with a high-aromatic content.

Udex Process (UOP-Dow). One of the early liquid – liquid extraction processes for the recovery of BTX aromatics from catalytic reformate was the Udex process [71], [72]. The process is also effective for recovering aromatics from coal tar, light oil, and pyrolysis gasoline. The first commercial installation was in conjunction with the UOP platforming unit at Cosden Petroleum Corp. [73].

The Udex process (see Fig. 10) uses diethylene glycol [*111-46-6*], 2-hydroxyethyl ether, $(HOCH_2CH_2)_2O$, M_r 106.12. The glycol contains about 8 % water to increase aromatic selectivity. This combination is effective as a solvent for separating aromatics from paraffins in the same boiling range. For aromatics production, a reformate fraction of 65 – 150 °C is charged to the liquid – liquid extractor where a multistage, countercurrent extraction takes place. The solvent enters at the top of the extractor column and passes countercurrent to the hydrocarbon feed, which is charged at an intermediate point. The raffinate passes overhead from the extractor. The aromatic-rich solvent is withdrawn from the bottom and passes to the solvent stripper, where the BTX material passes overhead and the denuded solvent is recycled to the extractor. The BTX mixture is washed with water and treated with clay before being separated into the benzene, toluene, and xylene by precise fractionation. Compounds of nitration grade specifications are obtained with only a small amount of xylene from the xylene distillation being lost from the bottoms. Typical results in the processing of catalytic reformate are shown in Table 17.

Improvements have been made in the Udex process in terms of solvent and of operating conditions. Aromatics recovery was improved and capacity was increased by using a solvent mixture of triethylene glycol and diethylene glycol or tetraethylene glycol. Pilot-plant studies carried out at Union Carbide showed that, where a solvent to feed ratio > 10 was required for high-aromatics recovery with diethylene glycol as the solvent, the ratio could be reduced to around 6 with the ethylene glycol – propylene glycol mixture. With tetraethylene glycol as the solvent, the ratio could be reduced to as low as 4 [74]. Operation at Sun Oil's Corpus Christi, Texas, plant [75] was improved when tetraethylene glycol was substituted for propylene glycol – ethylene glycol. Solvent/feed ratio was decreased with a resulting increase in capacity. Although the original process was based on the use of diethylene glycol as the solvent, most plants now use other more effective solvents, such as a tetraethylene glycol – water mixture.

Sulfolane Process. Shell Development Co. and Royal Dutch Shell have developed an aromatic extraction process based on the use of sulfolane as a selective solvent [76–79].

Sulfolane [*126-33-0*], tetramethylene sulfone, $C_4H_8O_2S$, M_r 120.17, is made by reacting sulfur dioxide and butadiene to form sulfolene, which is then hydrogenated to give sulfolane (→ Sulfones and Sulfoxides):

Some of the advantages of the sulfolane process are high selectivities and high yields. Also sulfolane has suitable physical properties, such as high density, low-heat capacity, and a boiling point high enough to permit easy separation for recycling. One disadvantage is thermal instability at the boiling point. However, this shortcoming is met by adding small amounts of water, which depresses the boiling point.

The sulfolane process (see Fig. 11) is similar to other aromatic extraction processes but differs in some operational procedures. The BTX containing feed is charged to a rotating disk contactor or extractor. Aromatics are extracted with a countercurrent flow of the sulfolane solvent. The aromatic-rich solvent from the bottom of the extractor passes to an extractive stripper. The overhead from this stripper is returned to the extractor as a backwash stream, for recovery of any residual aromatics. The aromatic – solvent mixture is charged to a recovery column where the aromatic product passes overhead. The lean solvent from the bottom of the column is recycled to the extractor.

A comparison of sulfolane with diethylene glycol [*111-46-6*] in the treating of light cat-

Table 17. Results of Udex extraction (U-2)

Properties of products	Udex charge	Benzene	Toluene	Xylene mixture
Distillation				
initial *bp*, °C	66	79.7	110	
50 %	99	80.1	110.8	
end point	150	80.3	111.0	
fp, °C		5.49		
% Aromatics	51.3	99.9	99.8	>99.8

Yields	Vol% of charge	Vol% of extract	% Recovery in extract
Benzene	7.6	15.3	>99
Toluene	21.2	42.5	95
Xylenes and ethylbenzene	20.0	40.0	94
C_9 + Aromatics	0.9	1.8	75

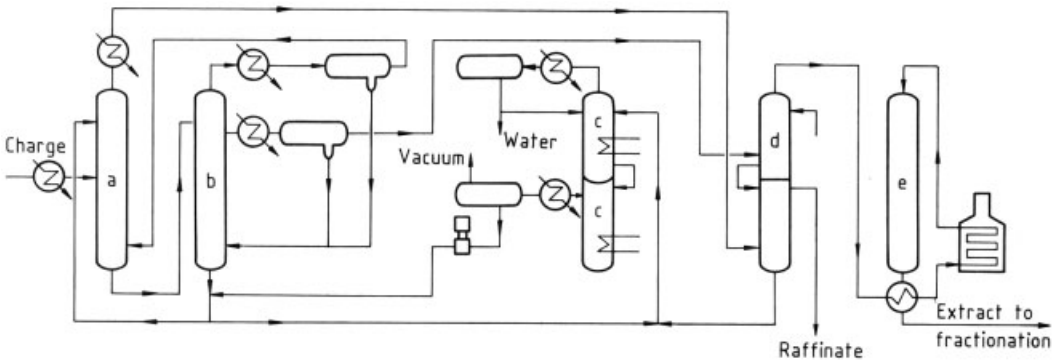

Figure 10. Udex extraction process
a) Extractor; b) Stripper; c) Water and solvent distillation; d) Water; e) Clay treater

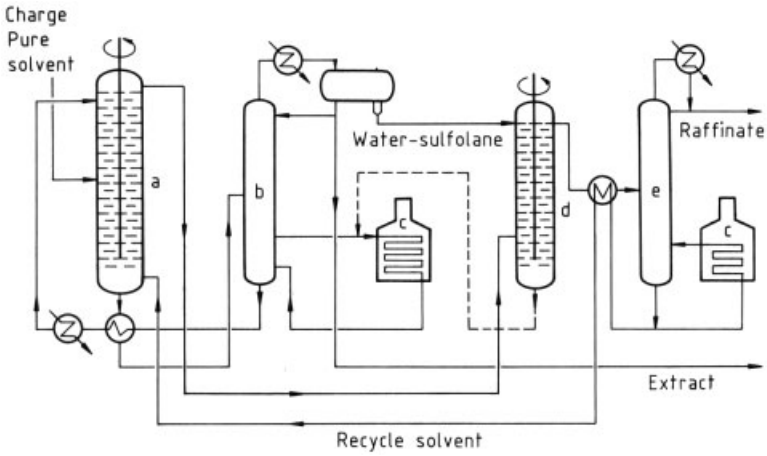

Figure 11. Sulfolane extraction process
a) Extractor; b) Extract stripper; c) Preheater; d) RDC washer; e) Rectifying column

alytic reformate to separate and produce nitration grades of benzene, toluene, and xylenes is summarized in Table 18.

Table 18. Comparison of sulfolane and diethylene glycol extraction [77], [80]

Operating conditions	DEG∗	Sulfolane
Solvent-to-feed, wt ratio	20	6.8
Backwash-to-feed, wt ratio	1.1	0.5
Stripping steam-to-feed, wt ratio	0.6	0.13
Stripper bottom temp., °C	143	191
Extractor top temp., °C	143	100
Feed temp., °C	116	116
Extractor gauge pressure, MPa	0.76	0.10
Water content of lean solvent	7.5	1.3

∗ Diethylene glycol

Yields and amounts of nonaromatic impurities (in parentheses) are: benzene, 99.9 % (0.01 %); toluene, 99.5 % (0.02 %); xylenes, 98.0 % (0.10 %).

As of 1980, over 55 sulfolane units were in operation with capacities ranging from a few hundred to over 45000 barrels/d [78].

Arosolvan Process. The Arosolvan process [81–83], licensed by Lurgi, uses an aqueous 1-methyl-2-pyrrolidone solvent. The water content ranges from 10 to 20 %, but is generally 12 to 14 %. Aromatics are extracted at 30 – 35 °C and at about atmospheric pressure. A comparison of properties of 1-methyl-2-pyrrolidone [*872-50-4*] and other solvents is shown in Table 19. Advantages of the Arosolvan process include operation of the extraction tower at atmospheric pressure and at room temperature, and low consumption of solvent (< 0.09 mg/L) [82]. Feedstocks may contain considerable unsaturates, and depentanization of the feedstock is not required because of the low-temperature operation of the extractor. Other advantages are clay treating of recovered aromatics is not necessary and waste disposal is not a problem because wash water is recycled.

Figure 12 is a simplified flow sheet of the process. The BTX feedstock is fed to the middle of the extractor where the aromatics are removed by countercurrent extraction with the solvent. The raffinate, containing pentane, other nonaromatics, and a trace of solvent, passes overhead from the extractor. The aromatic-rich solvent is passed to the first stripper, where pentane and some benzene passes overhead and, after joining with the overhead from the depentanizer, serves as a recycle stream to the bottom of the extractor. In the second stripper, high-purity aromatics and water are taken overhead while aromatics-free solvent is withdrawn from the bottom for recycle. The raffinate from the bottom of the depentanizing tower is separated from the solvent by water washing, and is removed from the process for fuel blending or other uses. The aromatic fraction from the overhead of the second stripper is separated by precise distillation.

Table 20 shows data obtained in the operation of two early commercial plants using the Arosolvan process. The Japanese plant operation involved the recovery of benzene, toluene, and xylene fractions from a feedstock composed of hydrogenated coal-gas and oil-gas benzene. The German operation used a hydrogenated pyrolysis gasoline as the feedstock.

DMSO Process. In 1965 the Institut Français du Pétrole described a process for the extraction of aromatics using dimethyl sulfoxide (DMSO) [*67-68-5*], $(CH_3)_2SO$, M_r 78.13, admixed with small amounts of water as the solvent (→ Sulfones and Sulfoxides) [84–86].

The DMSO process is shown in Figure 13. In the first extractor the BTX or other aromatic mixture is contacted at ambient temperature and atmospheric pressure in countercurrent fashion with the dimethyl sulfoxide solvent (containing < 9 % water). To maximize aromatic purity a mixture of aromatics and paraffins is fed to the bottom of the first extractor, countercurrent to the solvent flow, so as to combine reflux and backwash effects. The aromatic – solvent mixture passes from the first contactor to the second extractor where the DMSO solvent is recovered by re-extraction of the aromatics with a paraffin solvent. Extraction and re-extraction are carried out at room temperature and atmospheric pressure in rotating disk contactors of 10 – 12 theoretical stages. The paraffin – aromatic mixture from the second stage is separated by distillation and the paraffin is recycled to the second contactor. The small amount of DMSO that remains in the paraffin phase may be recovered by water washing. The DMSO is separated from the wash water by vacuum distillation.

Some of the advantages of this process are: DMSO has high selectivity for aromatic extraction; this selectivity is increased by the addition

Table 19. Comparison of properties of different aromatic solvents

Property	1-Methyl-2-pyrrolidone	Sulfolane	Diethylene glycol
bp, °C	206	287	245
fp, °C	−24	27.9	−8
Refractive index			
n_D^{20}	1.4703		1.447
n_D^{30}		1.481	
Density, g/mL			
20 °C	1.033		
30 °C	1.024	1.261	1.11
200 °C		1.116	
Viscosity (mPa s)			
30 °C	1.017	10.2	2.7
100 °C	0.97	2.5	2.5
200 °C		0.97	
Interfacial surface tension, mN/m			
30 °C	8.2		
120 °C		2.0	8.5

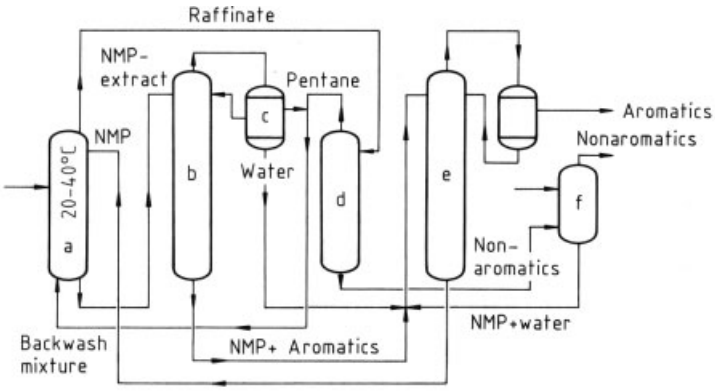

Figure 12. Arosolvan process
a) Extraction column; b) Stripper I; c) Separator; d) Pentane column; e) Stripper II; f) Recovery of 1-methyl-2-pyrrolidone (NMP)

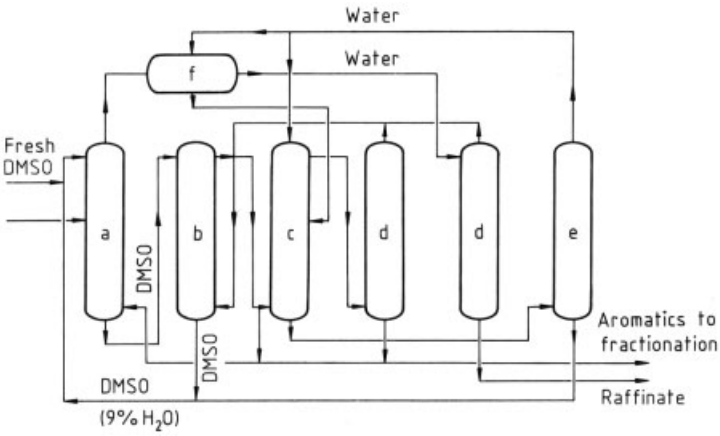

Figure 13. DMSO process
a) Extractor I; b) Extractor II; c) Washer for DMSO residue; d) Column for removal of pentane; e) Vacuum column for dehydration of DMSO; f) Raffinate container

Table 20. Plant data from Arosolvan process [80]

	Plant A Japan			Plant B Germany	
Feedstock	hydrogenated coal-gas and oil-gas benzene			hyrogenated pyrolysis gasoline	
Feedstock composition, wt%					
benzene	56			35	
toluene	16			25	
C_8 aromatics	8			0	
nonaromatics	20			40	
Yield, wt% in each fraction					
benzene	99.9			99.9	
toluene	99.7			99.8	
xylene + ethylbenzene	96.8			–	
Properties of fractions	B	T	X	B	T
fp, °C	5.50	–	–	5.50	
density, g/mL	0.884	0.871	0.870	0.884	0.872
benzene, wt%	99.99	<0.1	–	99.99	<0.01
toluene, wt%	0.0	>99.9	0.1	0.01	99.99
C_8 aromatics, wt%	–	<0.1	99.8	–	<0.01
nonaromatics, wt %	0.01	0.01	<0.1	0.01	<0.01

of small amounts of water; extraction and re-extraction are carried out at room temperature and atmospheric pressure; and DMSO is non-corrosive and is chemically stable under process conditions.

Yield data for benzene, toluene, and xylene in the treatment of a catalytic reformate and a pyrolysis gasoline, after hydrogenation of olefins and diolefins, are presented in Table 21.

Formex Process. The Formex aromatic extraction process (SNAM Progetti) uses *N*-formylmorpholine (FM) in admixture with water as the extraction solvent for aromatics [87], [88]. The process is designed for the extraction of benzene, toluene, and xylenes from BTX feedstocks, such as catalytic reformates or hydrogenated pyrolysis gasoline. The process can be used also for the recovery of C_9 and C_{10} aromatics or for the removal of small amounts of aromatics from paraffinic hydrocarbon mixtures.

N-Formylmorpholine [*4394-85-8*], 4-formylmorpholine, 4-morpholinecarboxaldehyde, $C_5H_9NO_2$, M_r 115.13, has the following structure (→ Amines, Aliphatic):

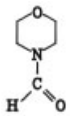

The Formex process applied to the recovery of BTX components is shown in Figure 14 [83]. The BTX feedstock is charged at the middle of the extractor column.

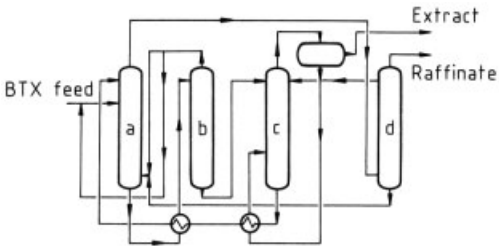

Figure 14. Formex process, BTX extraction
a) Extractor; b) Stripper; c) Extraction column; d) Raffinate washer

Solvent enters at the top. Extractor conditions (atmospheric pressure, ca. 40 °C) are controlled so that the raffinate, containing a minimum amount of aromatics, passes overhead. The aromatic-rich solvent from the extractor is fed to the extractive stripper, where a stream of residual saturated hydrocarbons is taken overhead and returned to the extractor as reflux. The bottoms from this extractive stripper pass to the extract recovery column. Here the aromatics are taken overhead and the lean solvent is recycled to the extractor.

Typical yields and quality of products for a BTX charge containing 5 – 40 % benzene,

Table 21. DMSO process for aromatic extraction (yield data for different feedstocks)

	Catalytic reformate		Hydrotreated pyrolysis gasoline	
	Feed components, wt%	Aromatic recovery, wt%	Feed components, wt%	Aromatic recovery, wt%
Benzene	15	>99.9	53.5	99.7
Toluene	40	99	17.5	99.7
Xylenes	10	75	6.0	90.0
C_9 + aromatics	2.5	20		
Total aromatics	67.5	93	77.0	99.0
Paraffins	32.5		23.0	

15 – 30 % toluene, and 15 – 30 % xylenes are [89]:

	Benzene	Toluene	Xylenes
Yield, %	100	99	95 – 97
Nonaromatics, ppm	–	<200	<200
fp, °C	5.5		

The basic process may be modified. For benzene-rich feedstocks the liquid – liquid extraction column may be omitted and the feed charged in the vapor phase to a column where the raffinate, containing a minor amount of benzene, passes overhead and the aromatic-rich solvent from the extractor bottom is handled in the conventional manner. Other modifications include removal of minor amounts of aromatics from primarily saturated hydrocarbon stocks.

Commercial operations include a 500 $\times 10^3$ t/a BTX extraction unit and a 24×10^3 t/a unit producing aromatic-free aliphatic solvents from a $C_6 - C_7$ feedstock.

Morphylane Process. Krupp-Koppers has described processes for separation and recovery of BTX components from aromatic feedstocks, such as catalytic reformate and hydrogenated pyrolysis gasoline [90–92]. The Morphylane process is essentially an extractive distillation. The solvents used are derivatives of morpholine, principally *N*-formylmorpholine.

Generally, extractive distillation processes are used for the recovery of one aromatic component. The addition of *N*-formylmorpholine to the BTX mixture makes nonaromatics that boil above benzene more volatile than benzene. Therefore, these nonaromatics can be extractively distilled from a binary mixture of aromatics (e.g., benzene and toluene).

The *N*-formyl-morpholine, without added components, such as water, is fed at ambient temperatures to the top of an extractive distillation column (Fig. 15). It contacts a countercurrent flow of the preheated BTX feedstock, which is charged at a middle point in the column. The nonaromatics are distilled and leave at the top of the column. The aromatic-rich solvent is withdrawn from the bottom and passes to the stripper, where the aromatics are discharged overhead. The cooled, lean solvent is recycled to the top of the extractive distillation column. The recovery of benzene from a hydrogenated fraction of pyrolysis gasoline is 97 %, purity is 99.95 wt%, and $fp \geqq 5.5$ °C.

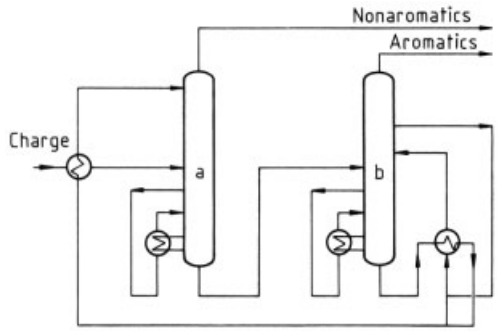

Figure 15. Morphylane extractive distillation process
a) Extractive distillation; b) Stripper

A toluene – xylene fraction from a feed consisting of catalytic reformate and hydrogenated pyrolysis gasoline, with morphylane processing under high-yield conditions, gives toluene in 99 wt% yield, purity > 99.9 %; and xylenes in 99.3 wt% yield, purity of 98 % [93].

A modified process for the combined and simultaneous recovery of BTX components from aromatic feedstocks, such as catalytic reformate or hydrogenated pyrolysis gasoline, has been described by Krupp-Koppers [94]. This process, called Aromex or Morphylex, combines liquid – liquid extraction with extractive distillation. In the liquid – liquid extraction stage, high-boiling nonaromatics are sep arated. In the extractive distillation stage, lower boiling nonaromatics are separated by extractive distillation. The overhead, after water removal, is fed back to the liquid – liquid extractor. The bottoms from the extractive distillation are passed to a stripper where the aromatics are distilled overhead and the solvent is recycled to the two extraction columns.

Typical data for the processing of a BTX – containing feedstock are:

	Benzene	Toluene	Xylenes
Yield, wt%	>99.9	>99.5	>97.0
Purity after fractionation	99.99	99.98	99.90

In 1982 six commercial units of the Morphylane process and one Morphylex unit were operating.

7. Quality Specifications

The ASTM has specified standards and test methods for benzene and related aromatic compounds. Likewise, specifications have been developed and used in other countries. As examples, the American National Standards have been used. The DIN Standards are used in Germany, and those of the Institute of Petroleum in England. The Technical Committee of ISO has standardized many of the specifications and test methods. Many of the ASTM methods have an Institute of Petroleum designation also.

The ASTM specifications for different grades of benzene are shown in Table 22. Test methods developed by ASTM are used in determining the specific properties of benzene:

1) Sampling: method D-270 for petroleum and petroleum products
2) ASTM D-891: specific gravity
3) ASTM D-1209: color
4) ASTM D-850: distillation
5) ASTM D-852: freezing point
6) ASTM D-848: wash color
7) ASTM D-847: acidity
8) ASTM D-849: copper corrosion
9) ASTM D-853: sulfur compounds
10) ASTM D-1685: thiophene
11) ASTM D-2360: nonaromatics

Similar specifications and tests have been developed for related aromatic compounds as well as for reagent grades of other materials; e.g., ASTM methods D-362–75 and D-841–71 describe standard specifications for industrial and for nitration-grade toluene.

The concentration of nonaromatics permitted in different grades of benzene is limited and in most cases must not exceed 0.15 %. For the 5.5 °C freezing point grade the maximum allowable is 0.01 %. The amount of paraffinic and naphthenic materials in the finished benzene is reflected in the boiling range and freezing point of the benzene.

The tests outlined by the ASTM methods and similar tests in different countries characterize benzene purity and whether or not the desired specifications are met. Gas chromatography is an important and efficient method for determining the presence of impurities in benzene and also the amount of benzene in different potential charge stocks available for the production of benzene.

8. Storage and Transportation

Benzene is a flammable, volatile, and toxic material, and its storage and transportation is closely regulated. Labeling is particularly important for the shipment of benzene. United States agencies involved in these regulations include OSHA, EPA, DOT, NIOSH, and others. The regulations dealing with handling and shipping are updated and published each year in the CFR. Other countries have regulations and safety practices quite similar to those in the United States. Although specific regulations must be met for manufacture, laboratory testing, storage, handling, and transporting of benzene, the safety rules are similar to those for other flammable and toxic liquids or gases.

Benzene is stored and transported in steel drums or tanks. Adequate ventilation is neces-

Table 22. Specifications for different grades of benzene

Specification	Industrial grade (ASTM D836−71)	Refined benzene-535 (ASTM D2359−69)	Refined benzene-485 (nitration grade) (ASTM D835−71)
Appearance		clear, water white	
Specific gravity (15.56/15.56 °C)	0.875 – 0.886	0.882 – 0.886	0.882 – 0.886
Color	not darker than 20 max.	20 max.	20 max.
Distillation range at 101.3 kPa	not more than 2.0 °C	not more than 1.0 °C	not more than 1.0 °C
Solidification point		5.35 °C min. (dry basis)	4.85 °C min. (dry basis)
Acid wash color	not darker than no. 3	no. 1 max.	no. 2 max.
Acidity	no free acid	nil	no evidence of acidity
H_2S, SO_2		free	
Thiophene		1 ppm	
Sulfur compounds	free of H_2S and SO_2		free of H_2S and SO_2
Copper corrosion	copper strip strip shall not show irridescence nor a gray or black deposit or discoloration		
Nonaromatics		0.15 % max.	
Carbon disulfide		if needed use ASTM D2324	

sary. Workers must be protected from skin contact or inhalation of fumes. Periodic physical examinations are recommended for those who work in areas where there is probable exposure to benzene. The handling of gasoline has caused concern because of its toxic components, which include benzene. Either CO_2 or dry chemical extinguishers are recommended for combatting benzene fires.

9. Economic Aspects

The demand for benzene as a raw material for the manufacture of petrochemicals has increased dramatically since the 1950s. Before World War II, light oil from coal carbonization was the main source for the production of benzene and the other BTX compounds. The development of new processes in the petroleum industry, such as catalytic reforming, provided a new raw material source for benzene. The rapid growth in the demand and production of benzene during 1950 – 1970, using United States production figures as an example, is shown in Table 23. The same increase in benzene production over the period before 1970 occurred in many other countries, especially those of Western Europe. For example, 1965 benzene production in the Federal Republic of Germany was ca. 330×10^3 t, but this had increased to ca. 820×10^3 t by 1970. Since the early 1970s the rate of growth in the quantities of benzene produced has leveled off somewhat, with downtrends apparent during periods of industrial recession. Because of decreased use of tetraethyllead in gasolines over the past several years, the demand for BTX aromatics in gasoline has increased. This may affect the overall demand and hence price level for benzene. Generally about 50 % of the benzene produced by petroleum refiners is retained for use as a gasoline component.

Benzene production by country over the past several years is shown in Table 24. Benzene production in 1980 for the non-communist world was over 13×10^6 t, with the United States producing about 40 % of the total [101]. An estimated annual growth of 2.0 – 2.5 % is predicted over the next years. The amount of benzene available from light oils produced in coal carbonization has not increased for some years, especially in the United States, for several reasons. Depression of the steel industry and the adoption of more modern steel processing methods have resulted in decreased coke production and hence decreased amounts of light oil. In addition, much of the light oil formed is not segregated but is burned as fuel in the steel industry. Petroleum refiners, however, purchase and refine this material because of its high content of benzene and appreciable amounts of the other BTX components. Table 25 shows the decline in coal-derived benzene in the United States since the 1950s. In 1953, 77 % of the benzene produced was derived from coal, but in 1981 this had decreased to 7 %. A further decline is projected. As mentioned earlier the change from coal-derived benzene to that produced by petroleum refiners came at a later date in Europe because of the greater dependency on coal. Thus in the Federal Republic of Germany in 1970, 38 % of the benzene produced was coal derived, as compared to

Table 23. Benzene production, United States

Year	Amount, 1000 t	Reference	Year	Amount, 1000 t	Reference
1953	911	[95]	1974	4970	[97]
1955	1031	[95]	1975	3420	[97]
1960	1539	[95]	1976	4760	[97]
1965	2770	[95]	1977	4866	[97]
1968	3340	[96]	1978	4970	[97]
1970	3788	[96]	1979	5588	[97]
1970	3812	[95]	1980	5294	[97]
1971	3594	[96]	1981	4472	[97]
1972	4182	[97]	1982	3583	[97]
1973	4853	[97]	1983	4175∗	[98]

∗ Projected

Table 24. Benzene production, 1000 t

Country	Year							References
	1977	1978	1979	1980	1981	1982	1983	
United States	5102	5048	5769	4980	4494	3585	4175	[95], [96], [98], [100]
United Kingdom	1020	883	1144	897	816	625	–	
Federal Republic of Germany	824	892	1004	917	919	1024	1333	[94], [98], [100]
Japan	1951	2015	2178	2060	1899	1814	1750	[94], [98], [100]
Canada	383	555	493	604	571	504	580	[94], [96], [97], [100]
Mexico	–	79	71	79	77	96	139	[97], [100]
Italy	–	–	528	449	423	392	518	[95], [100]
Brazil	145	172	260	309	336 ∗	336 ∗	–	[98], [99]
France	–	–	631	546	499	514	623	[97], [100]

∗ Estimates based on six-month values

about 10 % in the United States. About 50 % of the light oils produced in the United States is processed into BTX materials for use as petrochemical charge stocks or as a motor fuel component.

Table 26 gives benzene production and sales by coke-oven operators and by petroleum refiners. The petroleum refinery figures include benzene derived from light oils purchased from coke-oven operators and processed by the petroleum refiners. The differences between the production and sales figures are indicative of the benzene retained by the petroleum refiners for use as fuel components.

Except for short intervals, production of benzene approximates demand. After a rapid growth in production facilities, the United States and other countries have production capacity far greater than demand. For the years shown, the ratio of benzene production to existing capacity was approximately [101]:

United States	1980	0.62
	1981	0.56
Western Europe	1980	0.59
Japan	1980	0.71

Demand is expected to increase moderately during the 1980s and the ratio may improve. On the other hand, with projected increase in demand, production capabilities may increase, especially in industrial developing countries and in countries with excess hydrocarbon supplies. Illustrative of this projected increase in capacity are the benzene projects planned or under construction in late 1981 [102]. The survey showed that 32 projects were being considered or under construction. In North America there were five units and in Western Europe four units. Eastern Europe with eight units and the Asian area with seven units indicate a rather sizable expansion. Eight units were planned for other areas.

Because toluene and xylenes are related components of BTX feedstocks and because the alkylated benzenes may serve as feedstocks for benzene production, the amounts of toluene and

Table 25. United States benzene production from petroleum and from coal, 1000 t

Year	Production from		Total	% From	Reference
	petroleum	coal		coal	
1953	211	702	913	76.9	[95]
1955	332	699	1031	67.8	[95]
1960	1034	495	1529	38.4	[95]
1965	2335	435	2770	15.7	[95]
1970	3415	402	3817	10.5	[95]
1978	4573	398	4971	8	[101]
1981	4170	314	4484	7	[101]
1986 ∗	4930	315	5245 ∗	6	[101]

∗ Projected

Table 26. Benzene production and sales of synthetic organic chemicals (USITC)

Year	Production, 1000 t		Sales, 1000 t	
	Petroleum refiners	Coke-oven operators	Petroleum refiners	Coke-oven operators
1977	4580	216	1983	217
1978	4793	179	2347	181
1979	5383	204	2817	200
1980	6535	170	3661	169
1981	4368	105	2190	107

Table 27. Toluene and xylene production in different countries, 1000 t [96–99]

Country	1976	1978	1979	1980	1981	1982
Toluene						
United States	3279	4195	5379	5111	4680	3478
United Kingdom			293	190	179	169
Federal Republic of Germany	233	225	298	260	262	268
Japan	858	861	925	889	821	803
Canada			493	560	569	504
Mexico		124	108	125	132	138
Italy			366	286	278	268
Brazil		72	82	104	60	69
France			86	64	56	31
Xylenes						
United States	2363	2794	3125	2902	2916	2399
Federal Republic of Germany	477	414	492	355	486	459
Japan	1169	1248	1317	1195	1202	1225
Mexico		93	82	93	103	115
Italy			323	283	256	268
Brazil			117	134		

Table 28. Aromatics prices, US $/L

Year	Benzene∗	Toluene∗	Xylenes∗	Reference
1977	0.20	0.15	0.13	[103]
1979	0.08	0.28	0.28	[103]
1980	0.25	0.30	0.25	[103]
1981	0.45	0.33	0.39	[103]
1982	0.41	0.39	–	[101]
Aug. 1983	≈0.41	≈0.29	≈0.30	[104]

∗ All grades, average

xylenes produced are of direct significance. Table 27 contains production figures for toluene and xylenes.

The prices of the BTX chemicals in the United States are documented annually by the United States International Trade Commission (USITC) [103]. Prices have fluctuated but in general have increased over the years (see Table 28). The values given in this table are for all grades and include values assigned to the aromatics retained and used as components in fuel blends. Some small variations in price are evident depending on the source and quality of the aromatic. For benzene in 1981, the all-grades average price was $ 0.45/L. Benzene from coke-oven sources commanded $ 0.44/L and 1 °C and 2 °C benzene from petroleum refiners averaged $ 0.45/L. Spot prices for benzene, toluene, and xylenes are published regularly in technical and trade journals. In the August 17, 1983 issue of Chemical Week [104], the price of benzene was $ 0.41/L or less, with toluene and xylenes endeavoring to command $ 0.29/L and $ 0.30/L, respectively. A comparison of benzene prices in July, 1982 for United States, Japan, and the Federal Republic of Germany as published by SRI International [101] showed roughly equivalent prices, with the United States price range of $ 0.40 – 0.41/L; Japan, $ 0.40 – 0.42/L; and the Federal Republic of Germany $ 0.38 – 0.40/L.

10. Uses

With the exception of its use as a motor fuel component and with the restrictions on its use as a solvent, the three main applications for benzene are production of ethylbenzene, cumene, and cyclohexane. These three products account for 75 – 80 % of the benzene consumed as a chemical feedstock. Table 29 gives 1980 benzene consumption by product [101]. A distribution of products and end uses of benzene in the United States is shown in Table 30.

Most of the primary products of benzene may be classified as intermediates. For example, benzene is alkylated with ethylene to form ethylbenzene, which in turn is dehydrogenated to styrene (→ Ethylbenzene, → Styrene). Styrene is the monomer employed for the production of several important polymeric or copolymeric products. These include acrylic – butadiene – styrene resins, styrene – butadiene rubber, and others [105].

Cumene, another large-volume product made from benzene, is used as a charge stock for the production of phenol and acetone. Cumene is made by the vapor-phase alkylation of benzene with propene. Most of the phenol produced in the United States and non-communist countries is derived from the hydroperoxidation of cumene. Phenol is used mainly for the production of phenolic resins.

Cyclohexane is the third largest volume intermediate product derived from benzene. It is produced by the direct hydrogenation of benzene [106]. Cyclohexane is oxidized to form intermediates, such as adipic acid, and thus serves as the starting point for the production of nylons.

About 5 % of the benzene produced is nitrated to form nitrobenzene, which in turn is hydrogenated to form aniline.

The oxidation of benzene is an established process for the production of maleic anhydride but it is receiving stiff competition from other processes, notably those involving the oxidation of butane and butenes. Maleic anhydride serves as a precursor for polyester resins.

Other products of benzene include the halogenated benzenes and the linear alkylbenzenes. The latter are a source for the production of detergents.

11. Toxicology and Occupational Health

Benzene has long been recognized as a toxic compound with both acute and chronic effects. Poisoning occurs through inhalation, ingestion, and by rapid absorption through the skin. Manifestations of acute benzene poisoning include headaches, confusion, loss of muscular control, and irritation of the respiratory and gastrointestinal tract. Greater concentrations may result in unconsciousness and even in death. Exposure to long-term chronic poisoning or shorter term exposure at higher concentrations may have drastic aftereffects, such as anemia and probably leukemia. Brain damage and damage to the urinary tract, mucous membranes, and other body parts were found in postmortem studies of acute and fatal poisoning [107].

Table 29. Benzene consumption in various countries (1980), 1000 t [101]

	United States	Western Europe	Canada	Mexico	Japan
Production	5293	4059	604	79	2046
Capacity	8578	6880	–	–	2882
Demand	5166	4247	–	–	1825
End use, %					
Ethylbenzene	51.0	48.6	67	22	50.4
Cumene	20.7	19.3	7	–	12.1
Cyclohexane	13.8	13.4	23	37	25.6
Nitrobenzene/aniline	5.3	6.7	–	2	–
Detergent alkylate	3.1	2.0	–	22	–
Chlorobenzenes	2.6	5.2	–	–	3.7
Maleic anhydride	2.8	3.3	1	6	2.5
Other	0.8	1.5	2	11	5.7

Table 30. Benzene consumption in the United States

Year	1975	1977		1980	1986∗
Demand, 1000 t	3596		4886	5170	5610
Consumption, wt%					
Ethylbenzene	52	55	48	51	51
Cumene	17	17	20	21	24
Cyclohexane	17	16	17	14	14
Nitrobenzene/aniline	4.7	5	5	5	5
Detergent alkylate	3.4	2.8	–	3	2.7
Chlorobenzenes	4.5	2.9	–	2.6	2.1
Maleic anhydride	3.2	2.9	4	2.8	–
Other	–	–	6	0.8	0.8
Reference	[101]	[101]	[2]	[101]	[101]

∗ Projected

Inhalation of benzene at different concentrations in air is the most common source of benzene poisoning because benzene is rapidly absorbed by the lungs. Ingestion by mouth causes irritation of the mouth, esophagus, and stomach. Absorption of benzene by the blood may follow, with resultant chronic or acute poisoning. Absorption of benzene through the skin is not a major source of benzene poisoning [108].

The potential toxicity of benzene has been the subject of numerous studies [109], [110]. Benzene is selectively absorbed by the lungs [111]. About 80 % of the benzene inhaled by breathing air containing 3000 – 5000 ppm of benzene is retained. The benzene is absorbed by the blood and can become fixed in bone marrow, fatty tissues, and the liver. It is slowly metabolized to phenols and related compounds before being eliminated. Removal of benzene from the body by metabolism is determined by the presence of phenols in the urine. The phenols may act as toxic intermediates. The relationship between metabolism and benzene toxicity has been the subject of many studies [112], [113]. If benzene metabolism is inhibited by the addition of toluene, the destruction of red cell production is reduced [114]. Some inhaled benzene is slowly removed by exhalation. People breathing air containing 100 ppm benzene exhaled benzene 24 h after they ceased inhaling the benzene-contaminated air [115].

Susceptibility to benzene poisoning varies with the individual. Women, especially pregnant women, reportedly are more susceptible than men, but this has not been substantiated in other studies [116]. Individual susceptibility probably results from the natural resistance of the person or from the state of health at the time of exposure. Cases have been cited where workers, generally unconscious as a result of exposure to benzene fumes in enclosed areas, have recovered, whereas members of the rescue teams have died [117].

The literature is replete with references concerning the potential toxicity of benzene or that of its metabolized products and the var-

ious effects on the body. NTIS Report PB-289189 [118] reviews research findings of the effects of benzene poisoning at environmental exposure levels. Potential effects considered include leukemia, especially acute myelogenous leukemia, and pancytopenia (including aplastic anemia and chromosomal effects) and may result from long-term exposure. Again indications suggest that benzene toxicity and damage occur because of a toxic metabolite of benzene.

A group of chemical workers from seven industrial plants who had been exposed for 30 years to benzene for extended periods of time, but not necessarily continuously, were studied [119]. The incidence of various forms of cancer and the mortality rate of these workers was higher than, but not significantly above, those of the national norm.

Although conflicting reports have been issued over the years, it is generally believed that inhalation of benzene in occupational environments is potentially a cause of leukemia. An early indication of this relation was contained in a 1928 report [120]. In 1939, several cases of leukemia were reported in patients who had been chronically exposed to benzene [121]. In 1973, chronic benzene poisoning was thought to be responsible for four cases of acute leukemia. These patients had worked in an atmosphere containing 150 – 200 ppm benzene. According to other studies the connection between benzene poisoning and leukemia and other malignant conditions is much less certain. A 1974 survey included 38000 petroleum workers from eight refineries in Europe [122]. The purpose of the study was to determine if the incidence of leukemia could be related to workers chronically exposed to benzene at low-concentration levels. The conclusion was that the incidence of leukemia in the chronically exposed workers was not significantly different from that of the general population as a whole. However, this work has been criticized because of the different variables and assumptions made in the study. The results of other investigations indicate that benzene is suspected as a potential cause of leukemia.

Benzene toxicity has been related also to pancytopenia or destruction of elements in the blood. A related effect of benzene exposure is aplastic anemia in which bone marrow is damaged or destroyed. One early investigation studied workers in the printing industry who had been exposed for a long period to benzene vapors [123]. More recently the long-term effect of exposure to benzene has been studied. Partial recovery of blood and bone marrow damage may result on cessation of exposure if the damage has not been too great.

An EPA report [118] discusses the health effects caused by benzene at relatively low levels of concentration in terms of leukemia and blood disorders. Studies were made on humans and on animals.

Some investigations suggest that chromosome rearrangement may be caused by chronic exposure to benzene [110], [124]. Chromosomal damage of both stable and unstable nature can occur after benzene exposure. The stable effect was found to remain for several years. Alkylated benzenes did not cause this effect. These chromosome aberrations might result in the development of leukemia.

Although benzene has been universally recognized as a very toxic chemical compound with chronic and acute poisoning effects, questions still remain regarding its carcinogenicity. For instance, much research continues to determine whether benzene exposure results in leukemia. Studies have been done on animals and on groups of workers exposed to benzene vapors under varying operational conditions. Two examples of studies relating to leukemia indicate that much additional work is needed to sort out the complexities associated with the problem and to determine the quantitative level of exposure [125], [126].

Exposure to benzene emissions may occur from many sources. A major source of evaporative benzene emissions is from motor gasoline, both in exhaust gases from automobiles and from handling of gasoline during transport to service stations and from the filling of automobile tanks. Gasolines generally contain from 3 – 5 vol% benzene. Depending on control of emissions, a concentration of up to 3.2 ppm of benzene in the air may exist in the vicinity of the operation [127].

Another source of benzene emissions is in the manufacture of benzene and gasolines. Efficient process control in refineries, however, maintains evaporative losses at a minimum. Historically, a major source of benzene emissions was in the use of benzene or benzene-containing solvents in the formulation of paints, thinners, and ad-

hesives. Because of the high-toxic characteristics of benzene its use in these applications has been practically eliminated. Regulations state that substitute solvents should be used wherever possible. At one time alkylated mononuclear aromatics, such as toluene and xylene, were considered to have toxic effects equal to that of benzene. However, benzene is unique in most of its toxic effects. Earlier reports on the toxic effects of toluene and xylene may have overlooked the fact that benzene was present in these compounds [128]. The lower relative volatility of these alkylated aromatics is another factor in their favor.

Coal carbonization, for the production of coal coke for steel mill use, produces large amounts of aromatic compounds with benzene present in high concentrations.

Treatment for acute benzene poisoning is similar to that used for many cases of poisoning from toxic volatile chemicals. These steps include removing the person from the source, artificial respiration, if necessary, and other standard first-aid procedures.

Good laboratory and plant practices are paramount in preventing benzene poisoning. These involve minimum concentrations of benzene, either in liquid or vapor form, in work areas. Nonpermeable gloves have been suggested and tested as a means of preventing skin contact. Viton rubber or poly(vinyl alcohol) materials were found to be the best of those tested. However, these gloves are expensive and gloves are not always a guarantee of protection [129], [130].

The TLVs for benzene adopted by the ACGIH for 1983 – 1984 are: TWA 10 ppm, 30 mg/m^3; and STEL 25 ppm, 75 mg/m^3. Benzene is listed as a class A 2 industrial substance suspected of carcinogenic potential for humans (10 ppm) [131].

The DFG does not give a MAK for benzene but lists it in Section III A 1 under working materials which have been unequivocably proved carcinogenic. It is a compound capable of inducing malignant tumors as shown by experience with humans [132].

With the spotlight on the toxicity of benzene, tighter restrictions will probably be imposed on the use, the handling, and the exposure limits of benzene. The new regulations may be in the 1-ppm range. Recently the EPA has issued standards mostly aimed at benzene, projected to reduce emissions of volatile organic compounds by 70 %. This would apply to petroleum refineries and chemical plants [133].

Indicative of the activity and studies related to the toxicity of benzene and other poisonous materials, the following may be cited. A report on different carcinogens and methods for defining their risk factors has been issued by the Interagency Regulatory Liaison Group (IRLG) listing 26 substances, including benzene, and relating these to occupations that show high cancer rates [134].

The International Register of Potentially Toxic Chemicals (IRPTC) represents an international agency that is studying and registering the various effects of toxic chemicals under different conditions for each material. These conditions include factors related to production sampling, carcinogenicity, treatment of poisoning, and others [135].

The IRPTC was initiated in 1971 and is planning to publish its findings. The initial list of chemicals and products number 300 and more are to be added each year. The register is hoped to bring an international uniformity of information and control.

12. References

Specific References

1. M. Faraday, *Trans. R. Soc. (London)* **115** (1825) 440.
2. L. H. Hatch, S. Matar, *Hydrocarbon Process.* **57** (1978) no. 11, 291 – 301.
3. R. Maux, *Hydrocarbon Process.* **55** (1976) no. 3, 90.
4. *Hydrocarbon Process.* **60** (1981) no. 11, 196.
5. A. C. MacFarlane, *Oil Gas J.* **74** (1976) no. 6, 99 – 102.
6. *Hydrocarbon Process.* **56** (1977) no. 11, 144.
7. P. R. Pujado, *Hydrocarbon Process* . 55 (1976) no. 3, 91 – 96.
8. F. Treeny, *Erdöl Kohle* **20** (1967) no. 9, 629 – 634.
9. *Oil Gas J.* **61** (1963) no. 15, 125.
10. L. Aalund, *Oil Gas J.* **74** (1976) no. 27, 98 – 108.
11. D. E. Steeve, W. A. MacDonald, D. H. Stone, *Hydrocarbon Process.* **61** (1981) no. 9, 263 – 268.

12. M. Sittig: *Aromatic Hydrocarbons, Manufacture and Technology,* Noyes Data Corp., London 1976, p. 73.
13. Bethlehem Steel Co., US 3312749, 1967 (C. J. Hess, W. M. Perry).
14. Air Products, US 3564067, 1971 (W. C. Brenner, L. C. Doelph).
15. Bethlehem Steel Co., US 3623973, 1971 (M. O. Tarhan).
16. M. O. Tarhan, L. H. Windsor, *Chem. Eng.* 1966 (March 28) 96 – 98.
17. J. M. Duffalo, D. C. Spence, W. A. Schwartz, *Chem. Eng. Prog.,* **77** (January 2, 1981) 56 – 62.
18. D. McNeil: *Coal Carbonization Products,* Pergamon Press, 1966, p. 54.
19. E. J. Houdry, W. F. Burt, A. E. Pew, W. A. Peters, *Pet. Refiner* **17** (November 1938) 574 – 582.
20. Houdry Chemical, US 2078945, 1937 (E. J. Houdry).
21. Socony, US 2384942, 1945 (M. M. Marisic).
22. E. V. Murphree et al., *Pet. Refiner* **24** (1945) 97 – 100.
23. R. T. Sebulsky et al., *Oil Gas J.* **66** (1968) no. 21, 103 – 109.
24. A. N. Sachanen: *Conversion of Petroleum,* Reinhold Publ. Co., New York 1948, pp. 345 – 350.
25. *Hydrocarbon Process.* **61** (1982) no. 9, 164 – 169.
26. E. A. Sutton, A. R. Greenwood, F. H. Adams, *Oil Gas J.* **70** (1972) no. 21, 52 – 56.
27. *Oil Gas J.* **65** (1967) no. 8, 70 – 75.
28. S. B. Zdonik, E. J. Green, L. P. Hallee, *Oil Gas J.* **64** (1966) no. 51, 75 – 80.
29. J. G. Freiling, A. A. Simone, *Oil Gas J.* **71** (1973) no. 1, 25 – 31.
30. W. Kronig, *7th World Petroleum Congress,* **4** (1967) 235.
31. D. J. Griffiths, D. M. Luntz, J. L. James, *Oil Gas J.* **66** (1968) no. 8, 107 – 108.
32. S. Field, *Am. Pet. Inst. Proc. Refining Division* **50** (1970) 340 – 365.
33. A. H. Weiss, J. A. Marker, R. Newwirth, *Oil Gas J.* **60** (1962) no. 4, 64 – 71.
34. G. F. Asselin, R. A. Erickson, *Chem. Eng. Prog.* **58** (1962) no. 4, 47 – 52.
35. *Ind. Eng. Chem.* **54** (1962) no. 2, 28 – 33.
36. *Hydrocarbon Process.* **60** (1981) no. 11, 138.
37. *Hydrocarbon Process.* **46** (1967) no. 11, 184.
38. [37], p. 187.
39. S. Feigelman, C. B. O'Connor, *Hydrocarbon Process.* **45** (1966) no. 5, 140 – 144.
40. *Hydrocarbon Process.* **44** (1965) no. 11, 277.
41. S. Masamune et al., *Hydrocarbon Process.* **46** (1967) no. 2, 155 – 158.
42. S. Masamune, T. Kawatani, *Hydrocarbon Process.* **47** (1968) no. 12, 111 – 117.
43. *Hydrocarbon Process.* **61** (1981) no. 11, 139; **50** (1971) no. 11, 133.
44. S. Otani, *Jpn. Chem. Q.* **4** (1968) no. 6, 16.
45. E. G. Hancock (ed.): *Benzene and Its Industrial Derivatives,* J. Wiley & Sons, New York 1975, pp. 66, 67.
46. R. G. Craig, C. E. Fowler, M. L. Raczynski, *ACS Meeting,* Minneapolis, April 13 – 18, 1969.
47. *Hydrocarbon Process.* **61** (1982) no. 9, 153.
48. M. O. Tarhan, L. H. Windsor, *Chem. Eng.* **73** (March, 28, 1966) 96 – 98.
49. W. Lorz et al., *Erdöl Kohle* **21** (1968) 610.
50. H. Kubo, S. Masamune, R. Sako, *Hydrocarbon Process.* **49** (1970) no. 7, 111 – 113.
51. S. Komatsu, *Ind. Eng. Chem.* **60** (1968) no. 2, 36.
52. Mitsubishi Petrochemical Co., US 3400168 1968 (F. Fukudaetal).
53. P. Grandio: *162nd ACS Annual Meeting,* Washington, D.C., September 12 – 17, 1971.
54. P. Grandio, F. H. Schneider, A. B. Schwartz, J. J. Wise, *Oil Gas J.* **69** (1971) no. 48, 62 – 69.
55. T. Wett, *Oil Gas J.* **78** (June 16, 1980) 55 – 61.
56. R. Serrurier, *Hydrocarbon Process.* **55** (1976) no. 9, 253 – 257.
57. R. G. Nene, *Hydrocarbon Process.* **60** (1981) no. 11, 287 – 290.
58. R. C. Robinson, D. H. Broderick, H. A. Frumkin, *Hydrocarbon Process.* **61** (1982) no. 11, 102 – 108.
59. H. Hiller, O. L. Garkisch, *Hydrocarbon Process.* **59** (1980) no. 9, 238 – 242.
60. J. J. Wise, A. J. Silvesti, *Oil Gas J.* **74** (1976) no. 47, 140 – 142.
61. M. Steinberg, P. T. Fallon, *Hydrocarbon Process.* **61** (1982) no. 11, 92 – 96.
62. *Oil Gas J.* **82** (May 14, 1984) 76 – 80.
63. J. B. O'Hara et al., *Hydrocarbon Process.* **57** (1978) no. 11, 117 – 121.
64. *Hydrocarbon Process.* **61** (1982) no. 9, 126.
65. [64], p. 125.
66. R. Lester, *Hydrocarbon Process.* **40** (1961) 175.
67. K. H. Eisenlohr, *Erdöl Kohle* **16** (1963) no. 6, 523 – 533.
68. J. A. Gerster, *Chem. Eng. Prog.* **65** (1969) no. 9, 43 – 46.
69. M. Sittig: *Aromatic Hydrocarbons,* Manufacture and Technology, Noyes Data Corp. N.J., London 1976, pp. 104 – 141.

70. E. G. Hancock: *Benzene and its Industrial Derivatives,* J. Wiley & Sons, 1975, Chaps. 2, 3.
71. H. W. Grote, *Chem. Eng. Progress.* **54** (1958) no. 8, 43 – 48.
72. *Hydrocarbon Process.* **49** (1970) no. 9, 248.
73. D. P. Thornton, Jr., *Pet. Process.* **8** (March 1953) 384 – 387.
74. G. S. Somekh, B. I. Friedlander, *Hydrocarbon Process.* **48** (1969) no. 12, 127 – 130.
75. T. S. Hoover, *Hydrocarbon Process.* **48** (1969) no. 12, 131 – 132.
76. H. G. Staaterman, R. C. Morris, R. M. Stager, G. J. Pierotti, *Chem. Eng. Prog.* **43** (1947) no. 4, 148 – 151.
77. G. H. Deal, Jr., H. D. Evans, E. D. Oliver, M. N. Papaopoulos, *Pet. Refiner* **38** (1959) no. 9, 185 – 192.
78. *Hydrocarbon Process.* **59** (1980) no. 9, 203.
79. D. B. Broughton, G. F. Asselin: *7th World Petroleum Congress,* Mexico City, April 2 – 8, 1967, Section 4, pp. 65 – 73.
80. *Ullmann,* 4th ed., **8**: 395, 398.
81. [67], p. 52.
82. *Oil Gas J.* **64** (1966) no. 29, 83 – 84.
83. K. H. Eisenlohr, W. Grosshans, *Erdöl Kohle* **18** (1965) no. 8, 614 – 618.
84. J. Lautier, J. Durandet, C. Raumbault, M. Viguier, *Rev. Inst. Fr. Pet.* **20** (1965) 181 – 190.
85. B. Choffe, C. Raumbault, F. P. Navarre, M. Lucas, *Hydrocarbon Process.* **45** (1966) no. 5, 188 – 192.
86. *Hydrocarbon Process.* **51** (1972) no. 9, 185.
87. E. Cinelli, S. Noe, G. Paret, *Hydrocarbon Process* . 51 (1972) no. 4, 141 – 144.
88. K. H. Eisenlohr: *6th World Petroleum Congress,* Frankfurt/Main, June 19 – 26, 1963, Section 4, Paper 8.
89. *Hydrocarbon Process.* **61** (1982) no. 9, 182.
90. J. Morrison, *Oil Gas J.* **68** (1970) no. 47, 49 – 64.
91. F. Trefny, *Erdöl Kohle* **23** (1970) no. 6, 337 – 340.
92. M. Stein, *Hydrocarbon Process.* **52** (1973) no. 4, 139 – 141.
93. *Hydrocarbon Process.* **61** (1982) no. 9, 186.
94. *Hydrocarbon Process.* **61** (1982) no. 9, 187.
95. *Ullmann,* 4th ed., **8**:384.
96. "Facts and Figures," *Chem. Eng. News* **57** (June 11, 1979) 32.
97. "Facts and Figures," *Chem. Eng. News* **61** (June 13, 1983) 26.
98. *Chem. Eng. News* **60** (Dec. 20, 1982) 48.
99. "Facts and Figures", *Chem. Eng. News* **60**(June 14, 1982) 81.
100. "Facts and Figures", *Chem. Eng. News* **62**(June 11, 1984) 32.
101. C. S. Hughes, A. B. Abshire: "Benzene," *Chemical Economics Handbook,* SRI International, February, 1983, 300.7000 I, J.
102. T. Wett, *Oil Gas J.* **80** (1982) no. 13, 81 – 84.
103. "Synthetic Organic Chemicals (U.S. Production and Sales)" U.S. International Trade Commission *Publications 920* (1977), *1099* (1979), *1183* (1980), *1292* (1981).
104. *Chem. Week* **133** (1983) no. 7, 29.
105. T. C. Ponder, *Hydrocarbon Process.* **56** (1977) no. 7, 135 – 138.
106. *Hydrocarbon Process.* **60** (1981) no. 11, 147, 148.
107. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons,* Elsevier, New York 1960, pp. 97 – 98.
108. *H. E. W. Publication,* NIOSH, Washington, D.C., 1974, pp. 74 – 137.
109. "Assessment of Benzene as a Potential Air Pollution Problem," *Report P. B. 258-356* National Technical Information Service, vol. 4, 1976.
110. "Health Effects of Benzene," *Report P. B. 254-388,* National Technical Information Service, June, 1976.
111. "Metabolism versus Toxicity of Benzene," *Report P. B. 289-789,* National Technical Information Service, Sept. 1978, p. 8.
112. A. Stulmann, C. P. Stewart: *Toxicology of Volatile Liquids,* **vol. 1,** Academic Press, New York 1960, pp. 61 – 63.
113. P. C. Parke, R. T. Williams, *Biochem. J.* **54** (1954) 231 – 238.
114. L. S. Andrews, E. W. Lee, C. M. Witmer, J. J. Kocsis, *Biochem. J.* 1977 243.
115. C. G. Hunter, *Proc. R. Soc. Med.* **61** (1968) 913 – 915.
116. F. T. Hunter, *J. Ind. Hyg. Toxicol.* **21** (1939) 331.
117. G. L. Winek, W. D. Collom, *J. Occup. Med.* **13** (1971) 250 – 261.
118. *Report PB 289-789,* National Technical Information Service, Sept., 1978, p. 47.
119. O. Wong: "An Industry-Wide Mortality Study of Chemical Workers Occupationally Exposed to Benzene," *Report to Chemical Manufacturers Assoc.,* Environmental Health Assoc., Berkely, Calif., Dec. 8, 1983.
120. P. Delore, Borganamo, *J. Med. Lyon* **9** (1928) 227 – 233.
121. T. B. Mallory, E. A. Gall, Brickley, *J. Ind. Hyg. Toxicol.* **21** (1939) 355 – 393.

122. J. J. Thorpe, *J. Occup. Med.* **16** (1974) 375 – 382.
123. L. J. Goldwater, Tewksbury, *J. Ind. Hyg.* **26** (1941) 957-973.
124. E. Vigiliani, A. Forni, *J. Occup. Med.* **11** (1969) 148 – 149.
125. *Chem. Eng. News* **59** (September 7, 1981) 68.
126. *Chem. Eng. News* **54** (April 26, 1976) 7.
127. G. S. Parkinson, *Ann. Occup. Hyg.* **14** (1971) 145 – 153.
128. J. Aryanpur: *Tenth World Petroleum Congress,* Bucharest, 1979, vol. 5, p. 237.
129. E. B. Sansone, V. B. Tewari, *Am. Ind. Hyg. Assoc. J.* **39** (1978) no. 2, 169 – 174.
130. *Chem. Week* **125** (October 31, 1979) no. 18, 20.
131. ACGIH (ed.): *Threshold Limit Values (TLV)1983 – 1984,* ACGIH Cincinnati, Ohio 1984.
132. DFG (ed.): *Maximum Concentrations at the Workplace (MAK),* Verlag Chemie, Weinheim 1982.
133. *Chem. Eng. News* **62** (June 4, 1984) 5.
134. *Chem. Eng. News* **60** (Feb. 12, 1982) 4.
135. *Chem. Eng. News* **60** (May 10, 1982)25.

Benzene Hexachloride → Chlorinated Hydrocarbons

Benzenesulfonic Acids and Their Derivatives

For nitrobenzenesulfonic acids, see → Nitro Compounds, Aromatic; sulfobenzoic acids are treated under → Benzoic Acid; alkylbenzenesulfonic acids with aliphatic side chains longer than ethyl are treated under → Surfactants

Otto Lindner, Bayer AG, Leverkusen, Federal Republic of Germany

Lars Rodefeld, Bayer AG, Leverkusen, Federal Republic of Germany

1. General Aspects 1
1.1. Physical Properties 1
1.2. Chemical Properties 2
1.3. Production 3
1.3.1. Direct Sulfonation 3
1.3.2. Oxidation of Sulfur Compounds . . . 5
1.3.3. Diazo Reaction 5
1.3.4. Sulfite Reaction 6
1.4. Uses 6
1.5. Analysis 6
1.6. Environmental Aspects 7
1.7. Economic Aspects 8
1.8. Toxicology and Ecotoxicology 8
1.9. General Characteristics of Benzenesulfonic Acid Derivatives 8
1.9.1. Benzenesulfonyl Chlorides 8
1.9.2. Benzenesulfonamides 10
1.9.3. Benzenesulfonohydrazides 11
1.9.4. Benzenesulfonic Esters 11
2. Individual Benzenesulfonic Acids and Derivatives 11
2.1. Benzenesulfonic Acid 11
2.2. 1,3-Benzenedisulfonic Acid 14
2.3. Alkylbenzenesulfonic Acids 15
2.4. Formylbenzenesulfonic Acids 20
2.5. Chlorobenzenesulfonic Acids 20
2.6. Aminobenzenesulfonic Acids 22
2.7. Hydroxybenzenesulfonic Acids . . . 25
2.8. Chloro-alkylbenzenesulfonic Acids 27
2.9. Amino-alkylbenzenesulfonic Acids . 29
2.10. Amino- and Hydroxycarboxybenzenesulfonic Acids 29
2.11. Chloro-hydroxybenzenesulfonic Acids 30
2.12. Aminohydroxybenzenesulfonic Acids and Aminochlorohydroxybenzenesulfonic Acids 31
2.13. Amino-chlorobenzenesulfonic Acids 33
2.14. Amino-chloro-alkylbenzenesulfonic Acids 34
3. References 35

1. General Aspects

1.1. Physical Properties

The benzenesulfonic acids, with the exception of the aminobenzenesulfonic acids, are strong acids, which are almost entirely dissociated in aqueous solution. Some of them, including benzenesulfonic acid, can be distilled in a high vacuum without decomposition. They have a tendency to form hydrates, which are very stable in some cases and can be isolated from aqueous hydrochloric acid and dried at temperatures up to 100 °C without the water being lost. However, azeotropic distillation with benzene or chlorobenzene (depending on the heat stability of the hydrate) gives the anhydrous sulfonic acid.

Aminobenzenesulfonic acids are weak acids; they form zwitterions, which are poorly soluble in water, but their alkali salts are water-soluble. Unlike the free acids, the metal salts of the benzenesulfonic acids, particularly the alkali salts, are stable at temperatures up to and exceeding 300 °C. Most of them can be precipitated by adding an excess of salt to their aqueous solution. The calcium salts are more freely soluble in water than calcium sulfate. This is exploited industrially in the "liming" of the sulfonation mixture (see Section 2.2). In contrast to these metal salts certain organic salts are poorly soluble in water; for example, the salts with arylamines and *S*-benzylisothiourea have definite melting points and are therefore suitable derivatives for the identification of benzenesulfonic acids. For the physical properties of individual

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_507

compounds, see Section 1.9 and Chap. 2 of this article.

1.2. Chemical Properties

Desulfonation. The sulfonation of aromatic compounds to form arylsulfonic acids is a reversible reaction. It proceeds through various activated complexes and intermediates; Figure 1 shows a schematic plot of the enthalpy along the reaction coordinate. The ease with which sulfonation and desulfonation occur (kinetic control) and the position of the equilibrium (thermodynamic control) are dependent on the relative height of these enthalpies [8]. They are largely affected by other substituents present in the benzene nucleus. This is reflected by the various temperatures at which acidic desulfonation occurs [13]:

benzenesulfonic acid	227 °C
2-methylbenzenesulfonic acid	188 °C
4-methylbenzenesulfonic acid	186 °C
3,4-dimethylbenzenesulfonic acid	176 °C
3-methylbenzenesulfonic acid	155 °C
2,4-dimethylbenzenesulfonic acid	137 °C
2,4,6-trimethylbenzenesulfonic acid	100 °C

Figure 1 also explains the general rule according to which acidic desulfonation of benzenesulfonic acids occurs the more easily the easier it is to introduce the sulfonic acid group (higher sulfonation rate). This is often exploited industrially in the separation and purification of aromatic compounds as described in the next paragraph.

Pure *p*-xylene can be obtained from a mixture of *o*- and *p*-xylene by sulfonation and subsequent removal of water-soluble *o*-xylenesulfonic acid [14]. The same principle can be used to separate *m*-xylene from *p*-xylene [15], [16] and from ethylbenzene [17], or 1-ethyl-3,5-dimethylbenzene from other ethylxylenes [18], or 4-chloro-1,3-xylene from its isomers [19]. In the manufacture of DDT excess chlorobenzene was recovered by sulfonation, removal of the sulfonic acid, and subsequent desulfonation at 190 °C [20]. *m*-Dichlorobenzene can be separated as a sulfonic acid derivative from a mixture of isomeric dichlorobenzenes and obtained by subsequent desulfonation [21], [22]. The sulfonic acid group also serves as a protective group and permits the selective introduction of further substituents, e.g., in the manufacture of *o*-chlorotoluene:

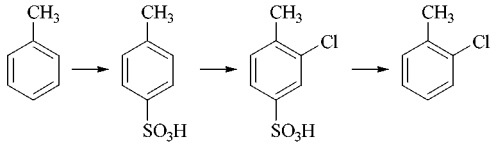

2,6-Dichlorophenol and 6-halogeno-2-methylphenol can be produced analogously. In a similar manner 2,6- and 2,5-dichlorotoluene and 2,3,6-trichlorotoluene are obtained from toluenesulfonic anhydride by chlorination and subsequent desulfonation [24]. Sulfonation – desulfonation is also important in the production of 1,4-benzenedisulfonic acid (Section 2.2).

Alkali Fusion. Alkaline cleavage of benzenesulfonic acids leads to phenols according to the following equation:

$$ArSO_3Na + 2\,NaOH \longrightarrow ArONa + Na_2SO_3 + H_2O$$

In former times this was the method of choice to produce phenol on an industrial scale; however, this process has mostly been replaced by the Hock process (→ Acetone, Chap. 4.1.).

Sulfone Formation. Benzenesulfonic acids react with aromatic compounds to form sulfones. This reaction is an undesired side reaction in industrial sulfonation (see Section 2.1); it is promoted by Friedel – Crafts reagents and inhibited by alkali metal ions [25].

Reaction with Halogenides. The reaction with phosphorus halogenides (PCl_5, PBr_5), chlorosulfuric acid, thionyl chloride, or phosgene leads to sulfonyl halogenides:

$$C_6H_5SO_2OH + PCl_5 \longrightarrow C_6H_5SO_2Cl + POCl_3 + HCl$$

If benzenesulfonic acid reacts with an excess of PCl_5, chlorobenzene is formed according to

$$C_6H_5SO_2Cl + PCl_5 \longrightarrow C_6H_5Cl + SOCl_2 + POCl_3$$

Sulfonyl fluorides are obtained by the reaction of sulfonyl chlorides with KF in the presence of phase-transfer catalysts in sulfolane.

Anhydride Formation. Sulfonic anhydrides are made by the reaction of P_2O_5, $SOCl_2$, or SO_3 with benzenesulfonic acids [26]. They are

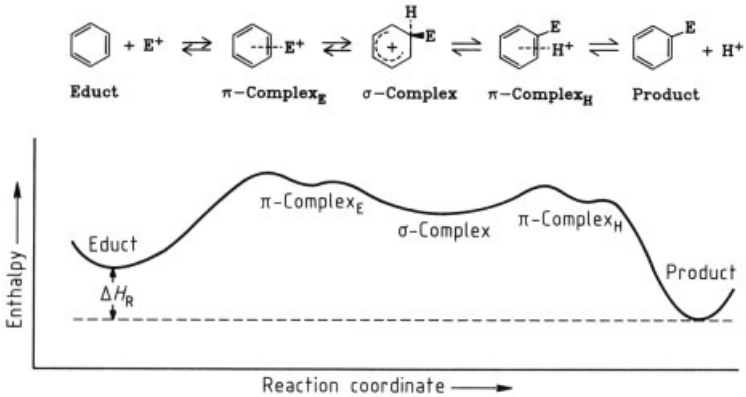

Figure 1. Sulfonation and desulfonation of benzene ($E^+ = SO_3$ or HSO_3^+), ΔH_R = reaction enthalpy

also formed through side reactions in the preparation of benzenesulfonic acids or benzenesulfonyl chlorides; in this case, however, they are frequently not observed because they are rapidly hydrolyzed in aqueous acid.

Esterification. Benzenesulfonic esters are formed in the reaction of the sulfonic acids or sulfonyl chlorides with alcohols or ethers (see Section 1.9.4).

1.3. Production

1.3.1. Direct Sulfonation

The direct introduction of the sulfonic acid group is one of the most important reactions in industrial organic chemistry. Together with nitration and chlorination, it belongs to the important group of electrophilic aromatic substitution reactions. It gives high yields under relatively mild conditions and usually results in well-defined benzene derivatives.

Aqueous sulfuric acid, at various concentrations from about 76 % up to 100 %, is frequently used as the sulfonating agent. In industry 100 % sulfuric acid is often called *monohydrate*. *Oleum* is a solution of sulfur trioxide in sulfuric acid; for sulfonation it is generally used at SO_3 concentrations of 20 or 65 wt % because the solidification points at these two concentrations are minimal (0 and 2 °C, respectively). Other concentrations are less favorable with regard to storage and transportation because the solidification points are higher (for example, 45 % oleum solidifies already at 35 °C) [27].

Sulfonation with sulfuric acid is a reversible reaction (see also Figure 1):

$$ArH + H_2SO_4 \rightleftharpoons ArSO_3H + H_2O$$

It can be shifted optimally to the right if the water of the reaction is bound or removed by distillation. Higher temperatures usually shift the equilibrium towards the right [28]. However, both methods increase sulfone formation.

The water of the reaction can also be bound by adding thionyl chloride:

$$C_6H_6 + H_2SO_4 + SOCl_2 \longrightarrow C_6H_5SO_3H + SO_2 + 2\,HCl$$

For example, the sulfonation of benzene with sulfuric acid and thionyl chloride gives a mixture consisting of 96.3 % benzenesulfonic acid, 2.7 % sulfuric acid, and 1.6 % diphenyl sulfone [29]. Similar yields are obtained when sulfonating toluene or chlorobenzene in this way.

No reaction water is formed in sulfonations using chlorosulfuric acid or sulfur trioxide. However, increased sulfone formation occurs. As a sulfonating agent pure *sulfur trioxide* generally reacts too violently and leads to extensive side reactions including oxidation and sulfone formation. Hence it is used in complexed form, e.g., with pyridine, dioxane, trimethylamine, or dimethylformamide [30]. Many industrial processes use sulfur trioxide gas as the sulfonating agent, normally mixed with an inert gas such as air, nitrogen, or carbon dioxide [31–33]. Thus, the sulfonating agent is diluted or evenly distributed so that the heat of the reaction is quickly dissipated [34] (see also → Surfactants). In the

liquid phase, sulfur dioxide or dichloromethane is suitable as diluent. However, dichloromethane reacts with sulfur trioxide at temperatures above 0 °C, and highly poisonous decomposition products are formed, such as bis(chloromethyl) ether. 1,1-Dichloroethane also reacts with sulfur trioxide at 80 °C; the decomposition products include phosgene.

In stoichiometric amounts, *chlorosulfuric acid* serves as a sulfonating agent according to

$$ArH + ClSO_3H \longrightarrow ArSO_3H + HCl$$

When using an excess of chlorosulfuric acid, sulfonyl chlorides are formed, see Section 1.9.1. The chlorosulfonation of toluene, for example, with chlorosulfuric acid at low temperature yields predominantly *o*-toluenesulfonyl chloride. This is the first step in the industrially important synthesis of saccharin (→ Sweeteners).

Reaction Mechanism. The basic reaction mechanism of aromatic sulfonation is described in Figure 1. The nature of the electrophilic agent E^+, however, varies according to the sulfonating agent used and the solvent.

In monohydrate (100 %) or only slightly diluted (92 – 99 %) sulfuric acid, sulfur trioxide is the electrophilic agent [35], [36]. The rate-determining step in this reaction is the elimination of the proton.

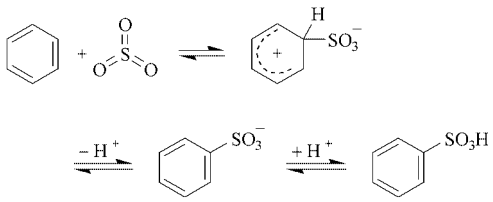

In sulfonation with disulfuric acid and with chlorosulfuric acid the SO_3H^+ cation (sulfate cation) is assumed to be the sulfonating agent [37]:

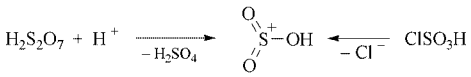

The SO_3H^+ cation is also formed as an intermediate when sulfur trioxide is used as the sulfonating agent [38–40]

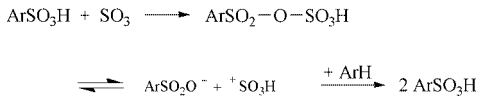

Orientation Rules. The sulfonation of benzene derivatives C_6H_5Y is facilitated over that of benzene if the substituent Y supplies electron density to the aromatic ring. Examples are:

Substituent Y	Electronic effect
Alkyl	+I
OR, NR_2	+M > − I
O^-	+M; +I

where I = inductive effect and M = mesomeric effect

Benzene derivatives are sulfonated with greater difficulty than benzene if the substituent Y withdraws electron density from the aromatic ring. Examples are:

Substituent Y	Electronic effect
COR, COOR, CN, NO_2, N_2^+, SO_3H	−M; −I
Cl, Br	−I > +M
NR_3^+	−I

where I = inductive effect and M = mesomeric effect

The electron-releasing substituents and the halogens direct the SO_3H group predominantly into the *o*- and *p*-positions (*first-order substituents*). On the other hand, the electron-withdrawing substituents (except for the halogens) generally direct the SO_3H group into the *m*-position (*second-order substituents*). The seemingly exceptional role of the halogens is explained by the electronic stabilization of the various reaction intermediates (π complexes, σ complex); this is discussed extensively in textbooks on organic chemistry.

This fundamental rule of aromatic substitution is valid particularly at low temperature. Raising the temperature may change the isomer distribution in the product mixture due to the transition from kinetic to thermodynamic control. For example, sulfonation of benzene with sulfuric acid at about 80 °C yields predominantly 1,3-benzenedisulfonic acid (see Section 2.2); however, by raising the temperature to 250 °C an isomer mixture consisting of 67 % 1,3-disulfonic acid and 33 % 1,4-disulfonic acid is formed. When sulfonating first-order-

substituted benzene derivatives, the ratio of the *ortho* and *para* isomer formed is also shifted in favor of the *para* isomer if the temperature is raised.

Technology of Sulfonation. Sulfonation is carried out in cast-steel or enameled steel vessels; in continuous processes, the vessels are arranged in cascades. In the batchwise mode the sulfonating agent is introduced into the vessel, the aromatic compound is subsequently added, and the reaction is controlled by means of temperature profiles or metering. In continuously run cascades the sulfonating agent is often introduced into two or three vessels simultaneously.

On completion of the reaction the batch is forced into water, which causes unconverted aromatic compounds to be expelled. The dilute sulfonation mass is then cooled, and the free acid separated by filtration. If the free acid is too soluble, sodium sulfate or sodium chloride is added, after which the acid is separated as the sodium salt (salting out). Often it is also possible to neutralize a concentrated solution with sodium hydroxide and then remove the precipitated sodium sulfate at 70 – 90 °C. The filtrate then contains the sodium salt of the arylsulfonic acid almost free from inorganic salt. It can be dried in drum or spray dryers.

If isolation is not possible in this way, the excess sulfuric acid can be neutralized by adding calcium carbonate. This leads to a large amount of gypsum, which is removed in the hot state ("liming" or "chalking"). The dissolved calcium arylsulfonate is then treated with soda and the precipitated calcium carbonate is removed by filtration. The filtrate contains the sodium arylsulfonate. Today, in large-scale manufacture the economical removal of byproduct gypsum is, especially in Europe, very difficult, and other isolation techniques such as temperature-controlled crystallization or reactive extraction have become more favorable.

In the reactive extraction process the unconverted sulfuric acid is recovered by converting the sulfonic acid into its ammonium salt with a long-chain aliphatic amine. This salt is separated from the sulfuric acid as a liquid phase and then converted with sodium hydroxide solution into sodium sulfonate solution and the amine; the latter can be separated as a liquid phase and is re-used. Sulfonates practically free from inorganic salts are obtained in this way [41], [42].

Special sulfonation and processing methods have been developed for compounds that withstand high temperatures and therefore allow the excess sulfuric acid to be distilled off in vacuo (see Hooker process for the production of benzenedisulfonic acid, see page 15). This gives sulfonic acids almost free from sulfuric acid, which, with skillful mixing, can be neutralized and subjected to alkali fusion in a single step [43].

Aromatic amines may be sulfonated by a special solid-state process, called baking: the ammonium hydrogensulfate that is first formed in the reaction between the amine and sulfuric acid is dewatered and simultaneously rearranged to aminobenzenesulfonic acid at 150 – 300 °C. The reaction product is solid, which requires the reaction to be carried out on trays in an oven.

The solids can also be suspended in an organic liquid such as diphenyl sulfone [44] or *o*-dichlorobenzene [45]; in this way the mass can be stirred during the entire reaction. This gives final products of higher quality because local overheating cannot occur (see also manufacture of *p*-sulfanilic acid, see page 23).

1.3.2. Oxidation of Sulfur Compounds

Such sulfur compounds as thiophenols and diaryl disulfides can be oxidized with chlorine solution, permanganate, or nitric acid to form sulfonic acids. The process is industrially important wherever it is difficult or impossible to introduce the sulfonic acid group directly and the starting compounds are easily produced. An example of the use of halogens as oxidizing agents is given in [46].

1.3.3. Diazo Reaction

Aryldiazonium halides are converted to aromatic sulfonic acids in glacial acetic acid containing SO_2 in the presence of copper(I) chloride [47].

This reaction, which was first observed by L. Landsberg, permits the production of, for example, 1,2-benzenedisulfonic acid from orthoanilic acid with a yield of 68 % [48].

$$HO_3S{-}Ar{-}N{\equiv}N^+\ Cl^- + \tfrac{1}{2}\,Cu_2Cl_2 \longrightarrow$$

$$HO_3S{-}Ar{-}N{\equiv}N^+\ [CuCl_2]^- \longrightarrow$$

$$HO_3S{-}Ar\bullet + N_2 + CuCl_2 \xrightarrow[-\,N_2]{+\,SO_2}$$

$$HO_3S{-}Ar{-}SO_2\bullet + CuCl_2 \xrightarrow{H_2O}$$

$$HO_3S{-}Ar{-}SO_2OH + [CuCl_2]^- + H^+$$

1.3.4. Sulfite Reaction

Aromatic halogen compounds can react with sulfite to form sulfonic acids if the halogen substituent is activated by nitro groups. For example, 1-chloro-2,4-dinitrobenzene reacts with sulfite to give 1,3-dinitro-4-benzenesulfonic acid. The reaction is catalyzed by copper ions. 2-Formylbenzenesulfonic acid and 2-sulfobenzoic acid are obtained similarly from the corresponding chlorine compounds.

The addition of hydrogensulfite to aromatic systems succeeds when their aromatic character is disturbed by the presence of certain substituents. The formation of 3-hydroxybenzenesulfonic acid from resorcinol [49] and the sulfitation of *m*-dinitrobenzene with simultaneous reduction of a nitro group [50] (Piria reaction) are examples:

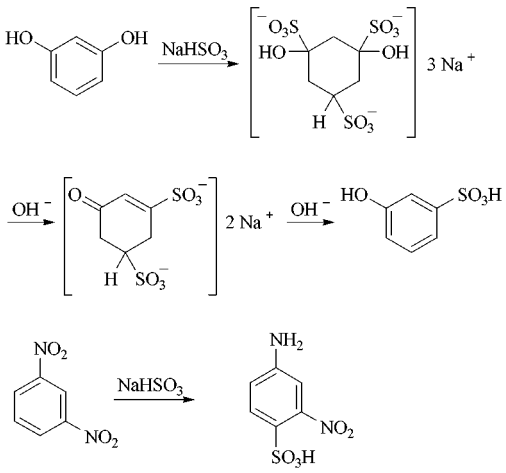

Hydrogensulfite addition, followed by the formation of a benzenesulfonic acid, gives satisfactory yields in special cases only and has therefore acquired little industrial importance so far.

1.4. Uses

Benzenesulfonic acids are used chiefly as intermediates. They are employed in the manufacture of sulfonic acid amides, hydrazides, and esters; of sulfinic acids, sulfones, phenols, and thiophenols; and of other compounds. Sulfonic acids that are substituted with OH and/or NH_2 groups serve as intermediates in the manufacture of finishing agents, optical brighteners, pickling agents, dyes, tanning agents, water-soluble resins, insecticides, ion-exchange resins, wetting agents, pharmaceuticals, polymeric thickeners, plasticizers, etc. Benzenesulfonic acids are also used as such as acidic catalysts and standardizing agents in dye manufacture.

1.5. Analysis

Today benzenesulfonic acids and most of their derivatives are analyzed by HPLC with high accuracy and efficiency [52]. By this method it is possible, both in the laboratory and in industry, to analyze the exact composition of sulfonation mixtures, which frequently consist of many components, and to observe how the composition is affected by the reaction parameters. Gas chromatography (GC) is used less often, but can be as effective, for example, after derivatization of reactive functional groups. GC methods have been reported for the analysis of *o*-aminobenzenesulfonic acids via their sulfonyl chlorides [51]. Benzenemonosulfonic and benzenedisulfonic acid mixtures can also be separated and analyzed quantitatively by GC analysis of their sulfofluorides [53].

In particular cases, benzenesulfonic acids are determined quantitatively in the presence of sulfuric acid and sulfones by titrating the total acid with alkali potentiometrically and then determining the sulfuric acid content gravimetrically with barium chloride. The difference gives the content of sulfonic acid. In general, in the case of the analysis of isolated salts of sulfonic acids,

the content should usually be referred to the free acid.

Older laboratory methods are the following: Aminobenzenesulfonic acids can be rendered distinctly visible on a paper chromatogram by diazotization and subsequent azo dye formation. Hydroxybenzenesulfonic acids can be stained well on the paper by coupling with a diazonium salt. Most other sulfonic acids can be recognized clearly from their fluorescence-extinguishing effects on chromatography paper previously sprayed with Pinakryptol Yellow and thus rendered fluorescent in UV light [54].

1.6. Environmental Aspects

Organic Compounds in Wastewater. Benzenesulfonic acid is biologically degradable. However, the direct biological degradation of many substituted benzenesulfonic acids is difficult or impossible; these compounds are therefore not suitable for treatment in biological wastewater purification plants without oxidative pretreatment.

If the desired benzenesulfonic acid can be precipitated from sulfuric acid solution without salt being added, the spent acid can be upgraded in a sulfuric acid recovery plant. If the wastewater contains salts and nondegradable sulfonic acids, low-pressure wet oxidation is a possible method of treatment that subsequently allows effective degradation in wastewater treatment works.

Inorganic Compounds in Wastewater. In the operation of large plants the treatment of such inorganic constituents as acids and salts is becoming an increasingly important cost factor in addition to the treatment of the organic constituents.

The following methods are available to decrease the acid or salt content of the wastewater.

1) Azeotropic distillation of the reaction water with a solvent. This improves the utilization of the sulfuric acid.
2) Use of oleum, sulfur trioxide, or chlorosulfuric acid in place of sulfuric acid. The resulting increase in sulfone formation can be counteracted, though not eliminated, with the aid of additives.
3) Precipitation of the final product from the sulfuric acid solution without adding salt, e.g., by cooling (see isolation of *p*-toluenesulfonic acid, see page 17).
4) Distillative separation of the sulfuric acid (see manufacture of benzenedisulfonic acid, see page 15).
5) Formation of a separate liquid phase consisting of an ammonium sulfonate. After phase separation, the sulfuric acid phase is returned to the sulfonation reactor.
6) Precipitation of the excess sulfuric acid as sodium sulfate if sodium sulfate is less soluble than the sodium sulfonate.
7) Precipitation of the excess sulfuric acid as calcium sulfate. The gypsum thus obtained is either dumped or used, depending on its quality.
8) Removal of the excess sulfuric acid by treating it with additional aromatic compounds that are more reactive than that used for the desired reaction. The resulting mixture could be subjected to alkali fusion to form phenols, which can be separated by distillation.

Waste Gas. The main organic constituents of waste air from benzenesulfonic acid production plants are unconverted portions of such starting materials as benzene, toluene, or halogenated benzene derivatives. They are usually removed in a sulfuric acid scrubber with a downstream thermal waste-gas remover. The waste sulfuric acid obtained in the scrubber can be recycled as a sulfonating agent. Because benzene is carcinogenic (Group C 1, MAK commission) the benzene emissions of newly built plants for the production of benzenesulfonic acid must be in accordance with the accepted lowest technically feasible levels.

Sulfur trioxide in waste gas forms aerosols that can only be removed effectively with specially designed scrubbers loaded with concentrated sulfuric acid. In most sulfonations the oxidative effects of sulfuric acid or sulfur trioxide lead to the formation of sulfur dioxide, which can suitably be removed from the waste gas by alkali scrubbing.

In the production of benzenesulfonyl chlorides the main impurity of the waste gas is hydrogen chloride. It can be used to produce hydrochloric acid or chlorosulfuric acid, provided that organic matter has been removed from it by

sulfuric acid scrubbing. Phosgene in the waste gas is hydrolyzed with water on activated carbon to form carbon dioxide and hydrogen chloride.

1.7. Economic Aspects

Most benzenesulfonic acids, chlorides, and amides are used as intermediates in the manufacture of dyes, crop protection agents, and pharmaceuticals. Because of the large number of producers and uses, little is known about exact market volumes. Especially in Asia and India the number of producers has increased, making the business for European suppliers increasingly difficult.

Table 1 lists some estimated production data for some benzenesulfonic acids, chlorides, and amides.

Table 1. Estimated production data for some benzenesulfonic acids, chlorides, and amides

	Production, 10^3 t/a	Year
p-Toluenesulfonic acid		
W. Europe	65.5	1990
World	115	1990
4-Aminobenzenesulfonic acid		
W Europe	11.5	1994
World	25	1994
m-Aminosulfonic acid		
W. Europe	2.3	1994
China	5.0	1994
India	9.4	1994
Japan	1.0	1994
World	17.7	1994
2-Amino-1,4-benzenedisulfonic acid		
World	2.0	1995
2-Amino-5-chloro-4-methylbenzene-sulfonic acid		
World	9.9	1990
2-Formylbenzenesulfonic acid		
World	4.0	1996
2-Methylbenzenesulfonyl chloride		
World	11.2	1995
4-Methylsulfonamide		
World	10 – 15	1994

Some use of selected benzenesulfonic acids, chlorides, and amides follow:

p-Toluenesulfonic acid	resins, *p*-cresol, C-Acid, acid catalyst
4-Aminobenzenesulfonic acid	dyes, optical brightener
m-Aminosulfonic acid	dyes
2-Amino-1,4-benzenedisulfonic acid	dyes, optical brightener
2-Amino-5-chloro-4-methylbenzene-sulfonic acid	Lake Red C
2-Formylbenzenesulfonic acid	triphenylmethane dyes, optical brightener
2-Methylbenzenesulfonyl chloride	saccharin
4-Methylsulfonamide	plasticizer

1.8. Toxicology and Ecotoxicology

In general, benzenesulfonic acids, chlorides, and amides are not very toxic. Nevertheless, benzenesulfonic acids and chlorides are strong skin irritants. Amino substitution does increase their pK_a values, but aminobenzenesulfonic acids and -chlorides are still moderately irritant. In addition, benzenesulfonyl chloride can act as a skin sensitizer [55]. Benzenesulfonyl hydrazide is suspected to be mutagenic in vitro [56]. Concerning ecotoxicology, except for benzenesulfonyl chloride, none of the compounds listed in Table 2 represent a potential major hazard for aquatic ecosystems.

1.9. General Characteristics of Benzenesulfonic Acid Derivatives

1.9.1. Benzenesulfonyl Chlorides

Properties. The chlorides of benzenesulfonic acids are crystalline substances with definite melting points. They decompose with formation of sulfur dioxide at temperatures between 100 and 200 °C, depending on their structure. They undergo hydrolysis, which in some cases is very slow. Hydrolysis is accelerated by hydrogen ions; this explains why the decomposition of moist pastes may not occur until a considerable time has expired after their production, and why it may then take place exothermally and very rapidly, with the result that dangerous local overheating occurs. The chlorides of the benzenesulfonic acids react with amines to form sulfonamides and with alcohols to form esters.

Production. The most convenient method for the production of benzenesulfonyl chlorides is the *chlorosulfonation reaction* of benzene or substituted benzene with chlorosulfuric acid. Sulfonation and formation of the sulfonyl chloride take place in one reaction sequence:

Table 2. Toxicological and ecotoxicological data for benzenesulfonic acids and their derivatives

Compound	LD_{50} (rat, oral), mg/kg	Aquatic toxicity			Degradation	
		species	LC_0, mg/L	LC_{50}, mg/L	Test method	%
Benzenesulfonic acid	1170				Zahn – Wellens [57]	90
4-Methylbenzenesulfonic acid	2480	*Leuciscus idus*		>500	Zahn – Wellens	100
2-Aminobenzenesulfonic acid	2700	*Leuciscus idus*	1000		closed bottle	90
4-Aminobenzenesulfonic acid	12 300	*Leuciscus idus*	1000		OECD screening [58]	90
4,6-Diamino-1,3-benzenesulfonic acids	2990	*Brachydanio rerio*	100		closed bottle	0
3-Amino-4-methylbenzenesulfonic acid	8480	*Brachydanio rerio*	500			0
4-Amino-3-methylbenzenesulfonic acid		*Salmo gairdneri*		133	dissolved organic carbon	24
2-Amino-5-methylbenzenesulfonic acid	11 700					
3-Amino-4-methoxybenzenesulfonic acid	> 5000	*Brachydanio rerio*	100		closed bottle	0
3-Amino-4-chlorobenzenesulfonic acid	> 5000	*Brachydanio rerio*	100		closed bottle	0
2-Amino-4,5-chlorobenzenesulfonic acid	> 5000					
3-Amino-5-chloro-2-hydroxybenzenesulfonic acid	> 5000	*Brachydanio rerio*	100		closed bottle	3
2,5-Dichlorobenzenesulfonic acid	1170					
Benzenesulfonyl chloride	1960	*Salmo trutta*		3	closed bottle	65
4-Methylbenzenesulfonyl chloride	4680	*Brachydanio rerio*		>1000		
Benzenesulfonamide	960	*Leuciscus idus*	200		closed bottle	0
N-(Butyl)benzenesulfonamide	1725 – 2050					
4-Methylbenzenesulfonamide	2330				$(BSB_5:CSB) \times 100$	13
4-Aminobenzenesulfonamide	>5000	*Leuciscus idus*	1000		Closed bottle	0
4-Chlorobenzenesulfonamide	515	*Brachydanio rerio*	100	198	Zahn – Wellens	6
3-Amino-4-hydroxybenzenesulfonamide	2160				$(BSB_5:CSB) \times 100$	1 – 10

$$Ar{-}H + Cl{-}SO_3H \longrightarrow ArSO_3H + HCl$$

$$ArSO_3H + Cl{-}SO_3H \rightleftharpoons ArSO_2Cl + H_2SO_4$$

Sulfone is formed in a side reaction of the chlorosulfonation (up to 7 % of sulfone is formed when benzene is used as raw material and up to 35 % when chlorobenzene is used). It is attributed to the reaction of the initially formed sulfonic acid with unconverted aromatic compound:

$$ArSO_3H + ArH \longrightarrow ArSO_2Ar + H_2O$$

$$H_2O + ClSO_3H \longrightarrow H_2SO_4 + HCl$$

In this reaction the chlorosulfuric acid serves as a condensation agent.

The amount of sulfone formation can be reduced by diluting with a solvent, by using a large excess of chlorosulfuric acid, or by adding sulfone-inhibiting substances, e.g., alkali metal and ammonium salts, acetic acid, phosphoric acid, dimethylformamide, or amidosulfuric acid [59].

For *industrial chlorosulfonation* the chlorosulfuric acid is introduced into a cast-steel or enameled steel vessel and 10 – 25 mol % of the aromatic compound is stirred in at 25 – 30 °C, whereupon sulfonation of the aromatic compound and HCl formation occur. The formation of sulfonyl chloride is initiated by heating the reactants to 50 – 80 °C. The reaction is exothermic. The temperature must be controlled accurately to ensure uniform release of the HCl gas. Restarting of the agitator after an interruption of the electricity supply is hazardous and may cause the contents of the vessel to foam over. The sulfonyl chloride is isolated by draining the reaction mass onto water and simultaneous cooling. Excess chlorosulfuric acid is decomposed, and the sulfonyl chloride either precipitates or separates as an organic liquid phase. The quality of the chlorosulfuric acid affects the yield.

In the case of aromatic compounds that easily take up two sulfochloride groups, e.g., anisole, monochlorosulfonation is carried out with only a little more than the calculated amount of chlorosulfuric acid at a low temperature (0 °C) and in the presence of a diluent such as dichloromethane. The reaction of benzenesulfonic acids with chlorosulfuric acid is an equilibrium reaction. Therefore, the yield can be in-

creased by using an excess of chlorosulfuric acid [60].

Other chlorinating agents, such as phosgene, thionyl chloride, sulfuryl chloride, or phosphorus pentachloride, can be used instead of chlorosulfuric acid. When thionyl chloride is used the sulfonyl chloride is obtained from the sulfonic acid in high yield and without formation of sulfuric acid:

$$ArSO_3H + SOCl_2 \longrightarrow ArSO_2Cl + SO_2 + HCl$$

A one-pot procedure is described in [61]. According to [62] the highest degree of conversion is obtained in the presence of at least 5 % of a sulfonating agent, e.g., sulfur trioxide, or an excess of chlorosulfuric acid. If phosgene is used as a chlorinating agent, the reaction can be accelerated by adding dimethylformamide [62] or the less problematic dibutylacetamide for dimethylcarbamoyl chloride as byproduct is avoided. The crude sulfonyl chlorides can be purified either by fractional distillation or by crystallization from an anhydrous solvent.

In a *diazo reaction,* an NH_2 group can be ultimately replaced by the SO_2Cl group; this reaction is suitable for the preparation of special sulfonyl chlorides whose SO_2Cl group must be in a specific position [63]. A hydrochloric acid solution of the diazonium salt, whose concentration must be as high as possible, is allowed to flow into a $CuCl_2$-containing 30 % solution of sulfur dioxide in glacial acetic acid [64]. A similar procedure using thionyl chloride was also described [65]. Diazonium chlorides with substituents having a strong negative inductive effect (NO_2, SO_3H) give the best results. Poor yields can be improved by adding magnesium chloride or a water-immiscible solvent having a low dielectric constant. Disulfides can be converted into sulfonyl chlorides with chlorine in aqueous suspension. The reaction is also applicable to thiophenols.

$$(C_6H_5)_2S_2 + 5\,Cl_2 + 4\,H_2O \longrightarrow 2\,C_6H_5SO_2Cl + 8\,HCl$$

The dechlorosulfonation of benzenesulfonyl chlorides offers the possibility to selectively introduce chlorine into aromatics. 1,2,4-Trichlorobenzene can be produced from 1,2-dichlorobenzene without formation of undesirable tetra- or higher chlorinated products via chlorosulfonation and subsequent dechlorosulfonation at 200 °C and passing chlorine gas through the reaction mixture [66].

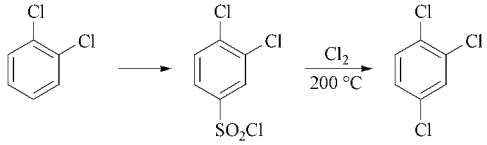

Due to their poor storage properties, many benzenesulfonic acids are reduced to the corresponding benzenesulfinic acids at 60 °C and pH 6 – 9.

$$ArSO_2Cl + NaHSO_3 + 3\,NaOH \longrightarrow ArSO_2Na + NaCl + Na_2SO_4 + 2\,H_2O$$

Sulfinates are often converted with ethylene oxide into β-hydroxyethylsulfones.

Uses. Benzenesulfonyl chlorides are important as intermediates. They are used to produce sulfonic acids, sulfonyl fluorides, sulfonamides, sulfonohydrazides, sulfonic esters, sulfinic acids, sulfones, and thiophenols.

1.9.2. Benzenesulfonamides

Properties. Benzenesulfonamides are readily crystallizing, colorless compounds with defined melting points and poor solubility in water. They are therefore suitable for the characterization of sulfonic acids (via the sulfonyl chlorides) and of primary and secondary amines and for the separation of amine mixtures (*Hinsberg method*) [67], [68].

Benzenesulfonamides are weak acids and form salts with bases. They are thermally stable and very difficult to hydrolyze with alkali; however, they are more easily hydrolyzed with mineral acids. In concentrated sulfuric acid sodium nitrite splits them into sulfonic acids and nitrogen. The hydrogen atoms bound to the nitrogen can be substituted.

Dibenzenesulfonylamines are acids similar in strength to the mineral acids [69]; they react with amines to form salts with definite melting points [70].

Production. Sulfonamides are produced by treating the sulfochlorides with ammonia, primary, or secondary amines in water. Alternatively they are produced in the presence of a base such as sodium hydroxide solution, sodium hydrogencarbonate, calcium carbonate, sodium sulfite, a second equivalent of amine or pyridine, or sodium acetate in glacial acetic acid; in this case, the reaction proceeds in an inert solvent, e.g., toluene or acetone.

Dibenzenesulfonylamines are obtained analogously according to the following reaction [71]:

$$2\,ArSO_2Cl + NH_3 \longrightarrow ArSO_2NHSO_2Ar + 2\,HCl$$

Unsymmetrical disulfonylamines are produced from sulfonamides and sulfochlorides in the absence of water [72] or in sodium hydroxide solution [73], [70].

Uses. Benzenesulfonamides serve as intermediates in the production of polysulfonamides, which are used as tanning agents and plastics. *N*-Alkylamides of the benzenesulfonic and toluenesulfonic acids can be used as plasticizers. Aminobenzenesulfonamides and diarylsulfonylamines with amino groups serve as intermediates in the production of azo dyes.

Sulfonamides, especially those derived from *p*-sulfanilic acid, sulfonyl ureas and sulfonyl guanidines are important in medicine as antibiotics, antidiabetics, diuretics, and anthelmintics [74], [75].

1.9.3. Benzenesulfonohydrazides

Properties. Benzenesulfonohydrazides, unlike the benzenesulfonamides, are thermally unstable. They decompose, when heated in water in some cases, to form sulfides, thiosulfonates, and nitrogen; for this reason they are used as blowing agents in the production of foams.

Production. Benzenesulfonohydrazides are produced by treating benzenesulfonyl chlorides with hydrazine in dioxane, alcohol, or water and in the presence of a second equivalent of hydrazine or another base. They react with nitrous acid to form benzenesulfonyl azides according to the following equation:

$$ArSO_2NHNH_2 + HNO_2 \longrightarrow RSO_2N_3 + 2\,H_2O$$

1.9.4. Benzenesulfonic Esters

Esters of benzenesulfonic acid are liquids or low-melting crystalline substances. They have relatively high thermal stability and can be purified by vacuum distillation. In general, they can be hydrolyzed with alkali at elevated temperatures only. The esters of benzenesulfonic acid are produced by treating benzenesulfonic acid halides with alcohols in the presence of a base [76], [77] or by treating benzenesulfonic acids with ethers [78]. These compounds are used as alkylating agents and plasticizers.

2. Individual Benzenesulfonic Acids and Derivatives

2.1. Benzenesulfonic Acid

Benzenesulfonic acid [*98-11-3*], $C_6H_5SO_3H$, M_r 158.17, was first obtained, together with diphenyl sulfone, by E. Mitscherlich in 1834 by heating benzene with fuming sulfuric acid. The industrially important reaction of benzenesulfonic acid with alkali hydroxide to form phenol (alkali fusion) was developed by A. Wurtz and A. Kekulé in 1867 and by P. O. Degener in 1878. Until the early 1960s benzenesulfonic acid was used chiefly in the manufacture of phenol. Other phenol syntheses are preferred now (→ Phenol).

Properties. Benzenesulfonic acid crystallizes from aqueous solution as a hydrate with $1.5\,H_2O$ in the form of deliquescent needles; its melting point is 43 – 44 °C. The monohydrate melts at 45 – 46 °C. The anhydrous acid, which can be distilled without decomposition at 171 – 172 °C (0.13 mbar), melts at 65 – 66 °C. It is very easily soluble in water and alcohol, poorly soluble in benzene, and insoluble in diethyl ether and carbon disulfide. Its dielectric constant in aqueous solution is 0.2. For the density and refractive indices of aqueous solutions, see [79].

Sodium benzenesulfonate decomposes at about 450 °C. It is soluble in 1.75 parts of water at 30 °C and in 0.8 part of boiling water, and crystallizes from aqueous solutions with 1 mol of water of crystallization. The calcium and barium salts are also soluble in water.

Benzenesulfonic acid has the characteristic reactions of a strong aromatic sulfonic acid. Acid hydrolysis at 175 °C splits it into benzene and sulfuric acid. Additional sulfonation with fuming sulfuric acid gives 1,3-benzenedisulfonic acid, which reacts further to 1,3,5-benzenetrisulfonic acid, and also diphenyl sulfone disulfonic acid (see Section 2.2).

Benzenesulfonic acid reacts with benzene to form diphenyl sulfone according to a Friedel – Crafts-type reaction:

$$C_6H_5SO_3H + C_6H_6 \xrightarrow{SO_3} C_6H_5SO_2C_6H_5 + H_2SO_4$$

Benzenesulfonic acid reacts with alkali metal hydroxide at 320 – 350 °C to form sodium phenolate according to

$$C_6H_5SO_3Na + 2\,NaOH \longrightarrow C_6H_5ONa + Na_2SO_3 + H_2O$$

This reaction was used in the first industrial synthesis of phenol.

Production. Benzenesulfonic acid is formed from benzene and sulfuric acid in an exothermic reaction according to

$$C_6H_6 + H_2SO_4 \longrightarrow C_6H_5SO_3H + H_2O$$

Crooks and White investigated the effects of temperature and of the composition of the sulfonation mixture on the reaction rate and developed a formula for calculating this rate [80]. The reaction stops at a particular sulfuric acid concentration between 74 and 78 %, which depends on the temperature and the amount of water formed. A proportion of the sulfuric acid (45 % in the classical process) is not utilized, but serves as a solvent and diluent and ensures that byproducts, particularly diphenyl sulfone, are formed in low yield only. In industry the excess acid is removed by adding calcium carbonate.

Any method that serves to reduce the excess of sulfuric acid increases the sulfone formation to a greater or lesser extent. The following methods have been used industrially:

1) Use of oleum or sulfur trioxide instead of sulfuric acid
2) Extraction of the benzenesulfonic acid from the reaction mixture with benzene
3) Azeotropic removal of the reaction water with benzene

Classical Process. The sulfonation by the classical process [81] is now of historical interest only, because it is coupled with the synthesis of phenol (use of sulfite from phenol production to neutralize the benzenesulfonic acid). Details of the process can be found in [82].

Continuous Sulfonation with Oleum (Monsanto process) [83]. Benzene and oleum are pumped simultaneously into the first of six sulfonation vessels forming a cascade. The vessels, which have propeller stirrers, are connected to one another by overflow pipes so that the reaction mixture can flow from one vessel to the next.

The first two vessels are cooled, whereas the others are heated. If optimal temperature conditions are provided, the reaction is complete when the sulfonation mixture leaves the last vessel. The excess sulfuric acid is removed by neutralizing the reaction mixture with sodium sulfite or sodium hydroxide solution. Sodium sulfate precipitates and is separated in a number of centrifuges. The solution of sodium benzenesulfonate is concentrated in an evaporator, which causes more sodium sulfate to precipitate. After centrifuging, the benzenesulfonic acid can be converted directly to phenol or dried. According to Carswell oleum containing 35.6 % of sulfur trioxide is used and the temperatures of the first and last sulfonation vessels are held at 70 – 80 °C and 110 °C, respectively. About 1 % of the benzene is converted to diphenyl sulfone.

Continuous Extraction Process. Under vigorous stirring excess benzene and sulfur trioxide are introduced into a vessel through separate pipes leading into the bottom of the vessel [84]. A benzene layer saturated with benzenesulfonic acid collects in the upper part of the vessel and passes through an overflow pipe into a second agitator vessel. In this vessel the benzenesulfonic acid is removed from the benzene extract by continuous washing with water or sodium hydroxide solution. The benzene is separated from

the aqueous benzenesulfonic acid solution in a separator. After being dried it is returned to the reaction vessel. The sulfuric acid consumption is 1260 kg for 1000 kg of converted benzene. Diphenyl sulfone formation is below 2 % [85].

Azeotropic Removal of Reaction Water [86], [87]. Sulfuric acid with a concentration of, for example, 79 % is heated to 170 °C in a cast-iron or enamelled steel sulfonation vessel and finely divided benzene vapor is introduced through a perforated plate situated in the lower part of the vessel. A portion of the benzene vapor, of which an excess is used, is sulfonated, whereas the unconverted portion continuously removes the water from the reaction mixture as it is formed. The mixture of benzene and water vapor leaving the vessel is condensed; after phase separation the benzene is returned to the evaporator.

Continuous operation of this process gives a final product containing 80.2 % of benzenesulfonic acid and 14.3 % of sulfuric acid; in batch operation the final product contains 93.1 % of sulfonic acid and 4.8 % of sulfuric acid. When adding sodium benzenesulfonate to the sulfonation mixture the sulfone formation can be kept below 2 % [88].

Analysis. A sample is taken from the sulfonation reactor and is diluted with water. Unconverted benzene is determined by gas chromatography, benzenesulfonic acid and diphenyl sulfone by high performance liquid chromatography. The sulfate concentration is determined at pH 2 – 3 by photometric titration with lead nitrate solution and dithizone (1,5-diphenylthiocarbazone) indicator in aqueous acetone solution [89].

Quality. Benzenesulfonic acid generally is sold as the sodium salt. This is almost colorless and without substantial impurities except for a small percentage of sodium sulfate.

Uses. Benzenesulfonic acid is used as an acid catalyst. The sodium salt is used to standardize dyes.

Benzenesulfonyl Chloride [*98-09-9*], $C_6H_5SO_2Cl$, M_r 176.62, $bp_{1013\,mbar}$ 251.5 °C, $bp_{13.3\,mbar}$ 120 °C, *mp* 14.5 °C, d_{15}^{15} 1.3842. Benzenesulfonyl chloride is made by treating benzene with chlorosulfuric acid at 30 – 35 °C [90]:

$$C_6H_6 + 2\,ClSO_3H \longrightarrow C_6H_5SO_2Cl + HCl + H_2SO_4$$

The sulfuric acid formed simultaneously inhibits the reaction. Therefore the yield of sulfonyl chloride is about 75 %, based on benzene. Benzenesulfonic acid and diphenyl sulfone are formed as byproducts in yields of approximately 18 % and 5 – 7 %, respectively.

The yield can be raised by partly replacing the chlorosulfuric acid with thionyl chloride. If this is done, the raw material and wastewater treatment costs are reduced, but the flue gas contains sulfur dioxide in addition to hydrogen chloride. Absorption of the gases in water recovers impure hydrochloric acid. The sulfur dioxide leaves the top of the hydrochloric acid recovery plant and can be used further in a sulfuric acid plant.

At the end of the reaction the products are cooled and simultaneously forced into a vessel containing water. The benzenesulfonyl chloride separates from the aqueous phase as a liquid organic phase. It is then washed twice with water and distilled; before that, the residual water is removed in vacuo. The main bottom product of the distillation column is diphenyl sulfone. The reaction can be carried out in vessels of cast iron or enameled steel.

Uses. Benzenesulfonyl chloride is produced on an industrial scale in Europe, the United States, and Japan. It is used as an intermediate for dyes and in the manufacture of benzenesulfinic acid, sulfonamides, and sulfonic esters. Benzenesulfonohydrazide serves as a blowing agent in the production of foams (trade name: Porofor BSH, Bayer).

Methyl Benzenesulfonate [*80-18-2*], $C_7H_8O_3S$, M_r 172.20, $bp_{20\,mbar}$ 150 °C, d_4^{17} 1.2730, is produced by treating benzenesulfonyl chloride with methanol in the presence of sodium carbonate [91] and is used as an alkylation agent. Yield: 87 %.

Ethyl Benzenesulfonate [*515-46-8*], $C_8H_{10}O_3S$, M_r 186.23, $bp_{20\,mbar}$ 156 °C, d_4^{17} 1.2192. Its manufacture and use are analogous to those of the methyl ester.

2-Chloroethyl Benzenesulfonate [*27887-43-0*], $C_8H_9ClO_3S$, M_r 220.67, $bp_{12\,mbar}$ 184 °C, d_4^{15} 1.353, is produced by treating 2-chloroethanol and benzenesulfonyl chloride at 12 – 15 °C (at most 25 °C) and adding sodium carbonate.

The reaction product is diluted with water and rendered alkaline by adding sodium carbonate. The organic lower layer is removed and distilled in vacuo [92]. Yield: 87.8 %, calculated on benzenesulfonyl chloride.

***p*-Chlorophenyl Benzenesulfonate** and **2,4-Dichlorophenyl Benzenesulfonate** are important as acaricides.

Benzenesulfonamide [*98-10-2*], $C_6H_5SO_2NH_2$, M_r 157.19, needles from water, flakes from alcohol, *mp* 156 °C. Obtainable from benzenesulfonyl chloride and ammonia. Used to manufacture chloramine B and dichloramine B (→ Chloroamines).

***N*-(Methyl)benzenesulfonamide** [*5183-78-8*], $C_7H_9NO_2S$, M_r 171.22, $bp_{22.5\,mbar}$ 202 °C, *mp* 30 °C, freely soluble in dilute alkaline solutions. Produced in the same way as the amide. Used as a plasticizer.

***N*-(Ethyl)benzenesulfonamide** [*5339-67-3*], $C_8H_{11}NO_2S$, M_r 185.25, *mp* 57 – 58 °C. Obtained as crystals from alcohol. Soluble in excess alkaline solution. Produced from benzenesulfonyl chloride with ethylamine and sodium carbonate. Yield: 98.7 % [93].

***N*-(Butyl)benzenesulfonamide** [*3622-84-2*], $C_{10}H_{15}NO_2S$, M_r 213.23, oil. Used as a plasticizer.

Benzenesulfonohydrazide [*80-17-1*], $C_6H_8N_2O_2S$, M_r 172.21, *mp* 101 – 103 °C, white powder, ϱ 1.48 g/cm^3, is used as a blowing agent in foam manufacture (Porofor BSH, Bayer).

4-[*N*-(Phenylsulfonyl)amino]acetanilide [*565-20-8*], $C_{14}H_{14}N_2O_3S$, M_r 290.34, *mp* 157 °C, crystallized from alcohol. Manufactured by treating 4-(amino)acetanilide with benzenesulfonyl chloride in water at 65 °C to which lime is added [94]. The compound is used to manufacture *N*-(4-aminophenyl)benzenesulfonamide.

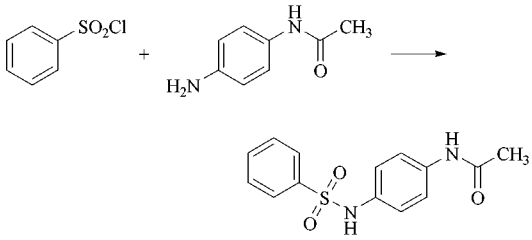

1

***N*-(4-Aminophenyl)benzenesulfonamide** [*5466-91-1*], $C_{12}H_{12}N_2O_2S$, M_r 248.3, *mp* 171 – 172 °C. Made from the 4′-acetylaminoanilide by splitting off the acetyl group in an alkaline medium. Used as an intermediate for dyes.

***N*-(Phenylsulfonyl)benzenesulfonamide** [*2618-96-4*], $C_{12}H_{11}NO_4S_2$, M_r 297.35, *mp* 157 – 158 °C, poorly soluble in water and alcohol, insoluble in ether, slightly soluble in acetone. Sodium salt: *mp* 314 – 316 °C; anilinium salt: *mp* 176 °C. Manufactured by treating benzenesulfonamide in 5 % sodium hydroxide solution with benzenesulfonyl chloride at 50 – 55 °C (1 – 2 h). During the reaction the pH value must be kept at 7.2 by adding 5 % sodium hydroxide solution. At the end of the reaction the disulfonylamine is precipitated from the clear solution by adding hydrochloric acid. Yield: 70 – 98 % [95]. Additive for electroplating baths [96].

2.2. 1,3-Benzenedisulfonic Acid

1,3-Benzenedisulfonic acid [*98-48-6*] (**1**), $C_6H_6O_6S_2$, M_r 238.24, is a hygroscopic substance with a solidification point of 137 °C.

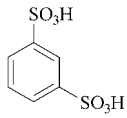

1

It is formed as the main product in the sulfonation of benzene or benzenesulfonic acid with excess oleum at temperatures of 80 – 250 °C. In proportion to the oleum concentration and reaction temperature

1 – 29 % of diphenyl sulfone disulfonic acid is formed as an undesired byproduct [97]. 1,4-Benzenedisulfonic acid is formed also, especially at high temperature, through rearrangement; mercury accelerates this process. At temperatures of 240 – 250 °C an equilibrium exists at a ratio of the *p*-acid to the *m*-acid of 1 : 2.

The disodium salt of 1,3-benzenedisulfonic acid is soluble in water at 27.5 °C to the extent of 41 %. With potassium or sodium hydroxide in aqueous solution under pressure it forms 3-hydroxybenzenesulfonic acid at 180 – 200 °C. At higher temperature resorcinol is formed.

Industrial Production [98]. Benzene is introduced into a cast-steel agitator vessel containing 100 % sulfuric acid. The contents of the reactor are heated to 100 °C and kept at this temperature for one hour. The products of the reaction (monosulfonation) are transferred to an agitator vessel containing 65 % oleum.

The temperature rises from 30 to 80 °C and is maintained at this level for another 1 – 2 h. The sulfonation mass is freed from sulfate ions in the normal way by adding calcium hydroxide or calcium carbonate and then removing the gypsum. It is then converted with sodium carbonate into the disodium salt, which is dried on a drum dryer. The yield is 93 % and the product has a purity of 96 %.

In a process patented by Farbwerke Hoechst [25] benzenesulfonic acid is mixed simultaneously with benzene and sulfur trioxide at temperatures above 120 – 140 °C. Sulfone formation is prevented by adding 3 – 15 % of the disodium salt of benzenedisulfonic acid. The excess of sulfur trioxide is about 0.2 mol per mole of benzene. The process can be operated continuously or discontinuously. After being neutralized with sodium hydroxide solution the final product can be melted to resorcinol directly. There is no necessity to remove sulfate ions. Excess sulfonating agent can be removed by subsequent reaction with benzene, toluene, or xylene. A mixture of resorcinol and phenol, cresol, or xylenol, is then formed by alkali fusion; the components of this mixture can be separated by distillation [99–101].

If benzene is sulfonated with 30 % oleum at 120 °C in the presence of sodium sulfate and the excess sulfuric acid is then converted into benzenemonosulfonic acid by introducing benzene vapor, a mixture of monosulfonic and disulfonic acid is obtained which can be used directly to produce phenol and resorcinol in the ratio 8 : 1 [99].

Particularly pure 1,3-benzenedisulfonic acid is obtained according to the process of Hooker Chemical Corp. [102]. Benzene is sulfonated with 96 % sulfuric acid at a molar ratio of the reactants of 1 : 3.4. The temperature is initially 65 °C, but is then raised to 80 °C. The water of the reaction is then removed by evaporation at 33 mbar and 200 °C. Finally, during disulfonation at 1.3 mbar and 210 – 235 °C, sulfuric acid and reaction water are distilled. The reaction product has a purity of 98.7 % and a solidification point of 136 °C. The process also can be operated continuously.

In a patent Koppers Co. [103] proposes that the disulfonation and simultaneous removal of water and excess sulfuric acid be carried out in a thin-film evaporator at a film thickness of 0.5 – 3 mm and a residence time of one minute. A vacuum of 13 mbar is claimed to be sufficient.

Uses. 1,3-Benzenedisulfonic acid is used to produce resorcinol and 3-hydroxybenzenesulfonic acid.

1,3-Benzenedisulfonyl Chloride [*585-47-7*], $C_6H_4Cl_2O_4S_2$, M_r 275.13, prisms from ether, *mp* 63 °C, $bp_{1.3\,mbar}$ 145 °C, $bp_{27\,mbar}$ 210.7 °C. Manufactured from *m*-benzenedisulfonic acid and phosphorus pentachloride or from benzenesulfochloride and chlorosulfuric acid. The compound is used to produce 1,3-benzenedisulfonohydrazide.

1,3-Benzenedisulfonamide [*3701-01-7*], $C_6H_8N_2O_4S$, M_r 236.27, obtained as needles from water, *mp* 229 °C. Manufactured from *m*-benzenedisulfochloride and aqueous ammonia.

1,3-Benzenedisulfonohydrazide [*4547-70-0*], $C_6H_{10}N_4O_4S_2$, M_r 266.31, white powder, *mp* 163 °C (decomp.), ϱ 1.24 g/cm^3, is suitable as a blowing agent for foams (Porofor B 13 CP 50, Bayer).

2.3. Alkylbenzenesulfonic Acids

This group of compounds comprises methylbenzenesulfonic acids (toluenesulfonic acids),

[*25231-46-3*], $C_7H_8O_3S$, M_r 172.20, methylbenzenedisulfonic acids (toluenedisulfonic acids) $C_7H_8O_6S_2$, M_r 252.26, dimethylbenzenesulfonic acids (xylenesulfonic acids) [*58723-02-7*], $C_8H_{10}O_3S$, M_r 186.23, and their derivatives.

Production. The sulfonation of toluene is accomplished more easily than that of benzene and is accompanied by less sulfone formation. Low temperatures favor *o*-substitution; *m*-substitution occurs only to the extent of about 2 – 3 %. At temperatures around 100 °C the ratio of *o*- to *p*-substitution is about 3 : 7. The sulfuric acid concentration has a strong influence on the isomer ratio [104]. Surprisingly, almost 100 % *p*-substitution is obtained at −10 °C if sulfur dioxide is used as the solvent [105].

At a sulfonation temperature of 100 °C and a sulfuric acid concentration of 97.4 %, 31 % *m*-substitution is achieved [106]. Through rearrangement at 140 – 200 °C the concentration of *m*-toluenesulfonic acid rises to about 46 % [107], [108]. If this is followed by hydrolysis with steam at 170 – 200 °C, the concentration of *m*-isomer rises to more than 90 % [109].

Isolation from the Mixture. *p*-Toluenesulfonic acid can be crystallized from 66 – 71 % sulfuric acid, or from concentrated hydrochloric acid at temperatures below 0 °C [110]. Then, according to the reaction conditions of the preceding sulfonation, the *o*- or *m*-toluenesulfonic acid can be obtained from the mother liquor. If the compounds are required to be very pure, the route via the toluidines (diazotization and reaction with sulfur dioxide) is preferred [47].

2,6-Toluenedisulfonic acid can be obtained by adding 65 % oleum to the sulfonation reaction mixture while the temperature is below 100 °C and then raising the temperature to 125 °C. In 90 % sulfuric acid at 190 °C, 64 % of the 2,4-toluenedisulfonic acid rearranges to 3,5-toluenedisulfonic acid. The 3,5-isomer is isolated from hydrochloric acid solution at −15 °C [111].

Uses. Toluenesulfonic acids are used as isomer mixtures, or as mixtures with xylenesulfonic acids; they serve as solubilizers, as acidic catalysts, and as additives for detergents [112]. They are used as intermediates in the production of cresols and sulfobenzoic acids.

2-Methylbenzenesulfonic Acid [*88-20-0*] (**2**), *o*-toluenesulfonic acid, $C_7H_8O_3S$, M_r 172.20, flaky crystals (dihydrate), *mp* 62.1 °C.

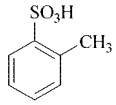

(**2**)

Produced together with 4-methylbenzenesulfonic acid by sulfonation of toluene with 96 % sulfuric acid at 40 °C, or preferably with 1 mol of chlorosulfuric acid at 0 to −10 °C. Used to produce *o*-sulfobenzoic acid.

2-Methylbenzenesulfonyl Chloride [*133-59-5*], $C_7H_7ClO_2S$, M_r 190.65, $bp_{13\,mbar}$ 126 °C, *mp* 15.5 °C. Manufactured by treating toluene with chlorosulfuric acid at temperatures not exceeding 3 °C, forcing the reaction mixture into ice, siphoning off the water, and centrifuging 4-methylbenzenesulfonyl chloride. Used in the production of toluenesulfonamide.

2-Methylbenzenesulfonamide [*88-19-7*], $C_7H_9NO_2S$, M_r 171.22, *mp* 156.5 °C, is obtained by treating 2-methylbenzenesulfonyl chloride with aqueous ammonia (*d* 0.910). Purified by fractional precipitation: hydrochloric acid is added to a solution which has been rendered alkaline with sodium hydroxide; the impurities are precipitated first. Used in the manufacture of saccharin (→ Sweeteners, Chap. 6.4.).

3-Methylbenzenesulfonic Acid [*617-97-0*] (**3**), *m*-toluenesulfonic acid, $C_7H_8O_3S$, M_r 172.20. The pure acid can be obtained from *m*-toluidine. Isomer mixtures with *m*-isomer concentrations of more than 50 % are obtained from toluene sulfonation mixtures by bubbling in toluene vapor at 180 – 205 °C [113], see also [109]. Used to manufacture *m*-cresol (→ Cresols and Xylenols).

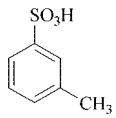

(**3**)

3-Methylbenzenesulfonyl Chloride [*1899-93-0*], $C_7H_7ClO_2S$, M_r 190.65, $bp_{29\,mbar}$ 146 °C, *mp* 11.7 °C. Obtained from *m*-toluidine according to [33]. Yield: 71.2 %.

3-Methylbenzenesulfonamide [*1899-94-1*], $C_7H_9NO_2S$, M_r 171.22, *mp* 103 °C.

4-Methylbenzenesulfonic Acid [*104-15-4*] (**4**), *p*-toluenesulfonic acid, $C_7H_8O_3S$, M_r 172.20, *mp* 38 °C, monohydrate *mp* 106 °C, trihydrate *mp* 93 °C, $bp_{26.3\,mbar}$ 140 °C.

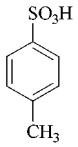

(**4**)

Production. According to the process of Allied Chemical and Dye Corp. [114], boiling toluene is sulfonated with 90 – 95 % sulfuric acid. Evaporating toluene is condensed, separated from water, dried, and recycled. A mixture with a *p*-isomer content of 75 – 85 %, an *o*-isomer content of 10 – 20 %, a *m*-isomer content of 2 – 5 %, and less than 1 % of sulfuric acid is present at the end of the reaction. There is very little sulfone formation. The process can also be performed continuously in a vertical reactor [116].

The reaction of toluene in a continuous process with sulfur trioxide as sulfonating agent allows toluenesulfonic acid to be generated with high *para* selectivity and with the advantage of few byproducts and low waste-gas emissions [115].

For some purposes, for example, for use as an acidic catalyst or in cresol manufacture, the crude toluenesulfonic acid obtained according to the above processes does not have to be purified. Purification is possible by crystallization from 66 % sulfuric acid or via the barium salt.

Uses. The applications of 4-methylbenzenesulfonic acid include the manufacture of 4-formylbenzenesulfonic acid, *p*-sulfobenzoic acid, 2-chlorotoluene-4-sulfonic acid, and 4-(chloromethyl)phenylmethanesulfonic acid.

4-Methylbenzenesulfonyl Chloride [*98-59-9*], tosyl chloride, $C_7H_7ClO_2S$, M_r 190.65, $bp_{25\,mbar}$ 138 – 139 °C, *mp* 69 °C. Formed as a byproduct in the manufacture of 2-methylbenzenesulfonyl chloride, from which it can be separated by remelting several times under water.

This compound is used in the production of sulfonamides, arylides, nuclear-substituted toluenesulfonyl chlorides, 4-(chloromethyl)benzenesulfonyl chloride, and *p*-toluenesulfinic acid. The tosylation of hydroxyl groups, i.e., their esterification with *p*-toluenesulfonyl chloride, serves to protect these groups; this method is employed particularly often in the chemistry of natural substances [117].

4-Methylbenzenesulfonamide [*70-55-3*], $C_7H_9NO_2S$, M_r 171.22, *mp* 137 °C, dihydrate *mp* 105 °C, weak acid, soluble in aqueous alkali. Produced from the sulfonyl chloride with aqueous ammonia. Used in the production of chloroamines, e.g., chloramine T and dichloramine T (→ Chloroamines).

Methyl 4-Methylbenzenesulfonate [*80-48-8*], $C_8H_{10}O_3S$, M_r 186.23, $bp_{17\,mbar}$ 168 – 170 °C, *mp* 28 °C, is obtained from the chloride of the acid by reaction with methanol; important alkylating agent.

Ethyl 4-Methylbenzenesulfonate [*80-40-0*], $C_9H_{12}O_3S$, M_r 200.26, $bp_{0.4\,mbar}$ 137 – 139 °C, $bp_{12\,mbar}$ 165 – 166 °C, *mp* 34 °C, is obtained from the sulfonyl chloride and ethyl alcohol at −5 to 0 °C after addition of 45 – 50 % sodium hydroxide solution [118]. Used as an alkylating agent and as a plasticizer.

4-(Chloromethyl)benzenesulfonic Acid [*46062-27-5*], $C_7H_7ClO_3S$, M_r 206.65, hygroscopic crystals, is obtained at a yield of 90 % through the action of chlorine on sodium 4-methylbenzenesulfonate in the presence of dichlorobenzene [119]. Used in the manufacture of surfactants.

4-(Chloromethyl)benzenesulfonyl Chloride [*2389-73-3*], $C_7H_6Cl_2O_2S$, M_r 225.09, *mp* 64 – 65 °C, $bp_{20\,mbar}$ 183 – 195 °C, is obtained by treating 4-methylbenzenesulfonyl chloride with chlorine in the presence of PCl_5 at 120 – 140 °C.

Methyl 4-(Chloromethyl)benzenesulfonate [*89981-68-0*], $C_8H_9ClO_3S$, M_r 219.67, is used as a quaternizing agent. Reactions with tertiary amines, e.g., trimethylamine, give sulfobetaines with fungistatic properties [120].

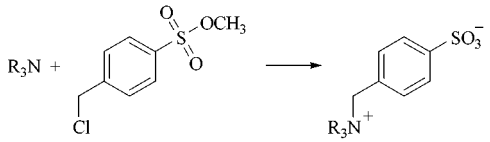

4-(Acetylaminomethyl)benzenesulfonyl Chloride [*39169-92-1*] (**5**), $C_9H_{10}ClNO_3S$, M_r 247.70, *mp* 95–97 °C, is obtained from *N*-acetylbenzylamine and chlorosulfuric acid [121].

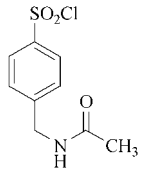

(**5**)

4-(Acetylaminomethyl)benzenesulfonamide [*2015-14-7*], $C_9H_{12}N_2O_3S$, M_r 228.27, *mp* 177 °C (from water or aqueous alcohol), is produced from the sulfonyl chloride with aqueous ammonia at 15 and 70 °C; yield 80 %. It is an important intermediate in the manufacture of 4-(aminomethyl)benzenesulfonamide.

4-Methyl-1,3-benzenedisulfonic Acid [*121-04-0*] (**6**), $C_7H_8O_6S_2$, M_r 252.26, is a viscous oil. Obtained by additional sulfonation of the *o*- and *p*-toluenesulfonic acid mixture with 66 % oleum at 125 °C. Used in the manufacture of 4-formyl-1,3-benzenedisulfonic acid, 4-carboxy-1,3-benzenedisulfonic acid, and 4-methylresorcinol.

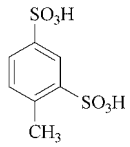

(**6**)

4-Methyl-1,3-benzenedisulfonyl Chloride [*2767-77-3*], $C_7H_6Cl_2O_4S_2$, M_r 289.16, *mp* 56 °C, is produced from toluene and chlorosulfuric acid at 120 °C [122]. Addition of phosphorus pentoxide raises the yield [123].

4-Methyl-1,3-benzenedisulfonamide [*717-44-2*], $C_7H_{10}N_2O_4S_2$, M_r 250.30, *mp* 190–191 °C.

2,4-Dimethylbenzenesulfonic Acid [*88-61-9*] (**7**), *m*-xylenesulfonic acid, $C_8H_{10}O_3S$, M_r 186.23, dihydrate: *mp* 57 °C, prisms or flakes. Produced by sulfonating *m*-xylene with concentrated sulfuric acid; the acid is precipitated by dilution with water. Of the three isomeric xylenes, *m*-xylene is the one most easily sulfonated. The ease of sulfonation decreases in the order $m>p>o$, whereas the ease of hydrolysis increases in the order $o<m<p$. *m*-Xylenesulfonic acid is hydrolyzed at 180 °C. Used in the production of nitroxylenesulfonic acid, chloroxylenes, and chloroxylenesulfonic acids.

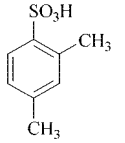

(**7**)

2,4-Dimethylbenzenesulfonyl Chloride [*609-60-9*], $C_8H_9ClO_2S$, M_r 204.68, *mp* 34 °C.

2,4-Dimethylbenzenesulfonamide [*7467-12-1*], $C_8H_{11}NO_2S$, M_r 185.25, *mp* 138 °C.

2,5-Dimethylbenzenesulfonic Acid [*609-54-1*] (**8**), $C_8H_{10}O_3S$, M_r 186.23, *mp* 48 °C, *bp* 149 °C (cathode vacuum), is soluble in chloroform. The dihydrate, *mp* 95 °C, is obtained by sulfonating *p*-xylene with 93 % sulfuric acid and removing the water by distillation [124]. Used in the production of 2,5-xylenol.

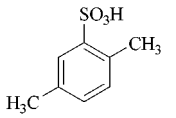

(**8**)

2,5-Dimethylbenzenesulfonyl Chloride [*19040-62-1*], $C_8H_9ClO_2S$, M_r 204.68, *mp* 25.5 °C, $bp_{29\,mbar}$ 152 – 153 °C.

2,5-Dimethylbenzenesulfonamide [*6292-58-6*], $C_8H_{11}NO_2S$, M_r 185.25, *mp* 148 °C.

4-Ethylbenzenesulfonic Acid [*98-69-1*], $C_8H_{10}O_3S$, M_r 186.21, is produced by sulfonating ethylbenzene with sulfuric acid. It is separated from the isomers via the aniline salt [125]. Used in the production of *p*-ethylphenol.

4-(2-Bromoethyl)benzenesulfonic Acid [*54322-31-5*] (**9**), $C_8H_9BrO_3S$, M_r 265.13; *S*-benzylisothiuronium salt, *mp* 149 – 150 °C. The acid is obtained by sulfonating 1 mol of (bromoethyl)benzene with 1.1 mol of SO_3 in methylene chloride [126], see also [127], [128]. The potassium salt is obtained by treating the sulfonyl chloride with potassium carbonate in water [129].

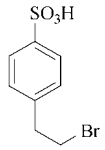

(**9**)

4-(2-Bromoethyl)benzenesulfonyl Chloride [*64062-91-5*], $C_8H_8BrClO_2S$, M_r 283.58, is produced from 2-(bromoethyl)benzene and chlorosulfuric acid at temperatures not exceeding 25 °C [130]. Used in the production of 4-vinylbenzenesulfonic acid.

4-(2-Bromoethyl)benzenesulfonamide [*5378-84-7*], $C_8H_{10}BrNO_2S$, M_r 264.15, *mp* 185.5 – 186 °C.

4-Vinylbenzenesulfonic Acid [*98-70-4*] (**10**), $C_8H_8O_3S$, M_r 184.21; *p*-toluidine salt, *mp* 182 – 183 °C, is obtained in the form of the potassium salt from 4-(2-bromoethyl)benzenesulfonic acid with methanolic KOH [131] or from 4-(2-bromoethyl)benzenesulfonyl chloride with alcoholic KOH [132]. The monomer can be stabilized by adding 0.5 – 5 % of sodium nitrite [133].

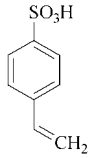

(**10**)

4-Vinylbenzenesulfonic acid has been proposed for copolymerization with acrylonitrile to improve the dyeing properties of the fibers [134] and as a starting product in the manufacture of polymeric styrenesulfonic acid [131], [132].

4-Vinylbenzenesulfonamide [*2633-64-9*], $C_8H_9NO_2S$, M_r 183.23, *mp* 138 – 139 °C [135].

4-*tert*-Butyl-2,6-dimethylbenzenesulfonic Acid [*28188-48-9*] (**11**), $C_{12}H_{18}O_3S$, M_r 242.31, is prepared from 1-*tert*-butyl-3,5-dimethylbenzene with 5 % oleum at 20 – 45 °C. Used in the manufacture of 4-*tert*-butyl-2,6-dimethylphenol and 2,6-xylenol [136].

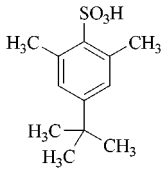

(**11**)

2,3-Dihydro-1*H*-indene-5-sulfonic acid [*40117-41-7*] (**12**), $C_9H_{10}O_3S$, M_r 198.24, is obtained by precipitation as a sodium salt, after 2,3-dihydro-1*H*-indene and concentrated sulfuric acid have been reacted at 150 °C [137]. At low temperatures mainly 1*H*-indene-4-sulfonic acid is formed. 1*H*-Indene-5-sulfonic acid has been proposed as a starting material for the manufacture of sulfonylureas with hypoglycemic effects [138].

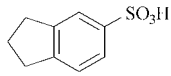

(**12**)

2,3-Dihydro-1*H*-indene-5-sulfonyl Chloride [*52205-85-3*], $C_9H_9ClO_2S$, M_r 216.69, *mp* 49 °C, $bp_{21\,mbar}$ 180 °C, is obtained by treating

the sodium salt of 1*H*-indene-5-sulfonic acid with PCl_5, or by the reaction of 2,3-dihydro-1*H*-indene and chlorosulfuric acid at 0 – 15 °C, pouring onto ice, and extracting with chloroform [139].

2,3-Dihydro-1*H*-indene-5-sulfonamide [*35203-93-1*], $C_9H_{11}NO_2S$, M_r 197.26, *mp* 135 °C.

2.4. Formylbenzenesulfonic Acids

2-Formylbenzenesulfonic Acid [*91-25-8*] (**13**), $C_7H_6O_4S$, M_r 186.18, is obtained as a thick syrup by treatment of its barium salt with sulfuric acid. Acquires a deep reddish-violet color if Schiff's reagent is added. In contrast to the poorly soluble barium salt, the sodium salt is freely soluble in water; it is also freely soluble in hot, but not in cold, ethyl alcohol.

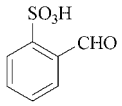

(**13**)

Production. The starting material is *o*-chlorobenzaldehyde, which is heated with sodium hydrogensulfite solution in an autoclave to 190 – 200 °C, with the result that its chloro substituent is replaced by the sulfonic acid group. The batch is allowed to cool, sulfuric acid is added, and excess sulfur dioxide and unconverted chlorobenzaldehyde are driven off by boiling. The resulting solution is used directly in the manufacture of dyes. The acid can also be produced by oxidation of *o*-toluenesulfonic acid with oxygen in the presence of bromine and cobalt as catalysts [140].

Uses. 2-Formylbenzenesulfonic acid is an intermediate in the manufacture of optical brighteners [141], trade names including Uvitex NFW (for textiles) and Tinopal CBS (for detergents). Triphenylmethane dyes are produced by condensation of the acid with various *N,N*-dialkylanilines, *N,N*-dialkyl-*m*-aminophenols, and their sulfonic acids.

4-Formylbenzenesulfonic Acid [*5363-54-2*] (**14**), $C_7H_6O_4S$, M_r 186.18. The thiosemicarbazone decomposes at 230 °C. 4-Formylbenzenesulfonic acid is produced by oxidation of *p*-toluenesulfonic acid with manganese dioxide in 25 % oleum or by oxidation of *p*-toluenesulfonyl chloride in glacial acetic acid – acetic anhydride with chromium trioxide and concentrated sulfuric acid at 0 °C and hydrolysis of the sulfonyl chloride of the aldehydehydrate diacetate with dilute hydrochloric acid [142].

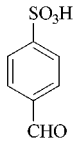

(**14**)

4-Formyl-1,3-benzenedisulfonic Acid [*88-39-1*] (**15**), $C_7H_6O_7S_2$, M_r 266.25, is freely soluble in water. The sodium salt crystallizes with 2 mol of H_2O and is freely soluble in water and poorly soluble in ethyl alcohol. It is made by reaction of 2,4-dichlorobenzaldehyde with sodium hydrogensulfite liquor or oxidation of 2,4-toluenedisulfonic acid with manganese dioxide in sulfuric acid. The acid is used in the production of triphenylmethane dyes and wetting agents.

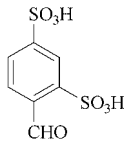

(**15**)

2.5. Chlorobenzenesulfonic Acids

4-Chlorobenzenesulfonic Acid [*98-66-8*] (**16**), $C_6H_5ClO_3S$, M_r 192.62, *mp* 68 °C, deliquescent needles, is soluble in water and ethanol. Prepared by sulfonation of chlorobenzene with sulfuric acid; the water formed in the reaction is removed continuously. It can be hydrolyzed to chlorobenzene with steam at 190 °C [143] and is used in the manufacture of 4-chloro-3-nitrobenzenesulfonic acid.

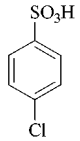

(**16**)

4-Chlorobenzenesulfonyl Chloride [*98-60-2*], $C_6H_4Cl_2O_2S$, M_r 211.01, $bp_{16\,mbar}$ 140 °C, *mp* 55 °C, is manufactured by mixing 4-chlorobenzene and chlorosulfuric acid at 35 °C, heating for two hours at 80 °C, and discharging the contents of the reactor into ice water; yield 81 % of theory, together with a small amount of bis(4-chlorophenyl) sulfone [145], [146]. Used in the manufacture of chlorobenzenesulfinic acid and sulfimides; three-step reaction to the sulfone in a single reactor is possible without intermediate isolation [147].

4-Chlorobenzenesulfonamide [*98-64-6*], $C_6H_6ClNO_2S$, M_r 191.63, *mp* 146 °C, is freely soluble in ethyl alcohol, ether, and hot water.

4-Chloro-*N*-[(4-chlorophenyl)sulfonyl]-benzenesulfonamide [*2725-55-5*], $C_{12}H_9Cl_2NO_4S_2$, M_r 366.24, *mp* 298 °C (from water), is prepared by treating the corresponding sulfonyl chloride with 0.45 mol of ammonium chloride at 0 – 3 °C, with dropwise addition of sodium hydroxide solution at pH 8; yield 91 % [148]. Used as a reagent for identifying amines.

2,5-Dichlorobenzenesulfonic Acid [*88-42-6*] (**17**), $C_6H_4Cl_2O_3S$, M_r 227.06, crystals, freely soluble in water. Obtained by sulfonating 1,4-dichlorobenzene with oleum.

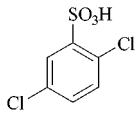

(**17**)

2,5-Dichlorobenzenesulfonyl Chloride [*5402-73-3*], $C_6H_3Cl_3O_2S$, M_r 245.51, *mp* 39 °C, needles (from ethyl alcohol), is made by treating 1,4-dichlorobenzene with chlorosulfuric acid at 150 °C and stirring into ice water; yield 85 % [149]. It has been described as a modifier for Nylon 6 [150].

3,4-Dichlorobenzenesulfonic Acid [*939-95-7*] (**18**), $C_6H_4Cl_2O_3S$, M_r 227.06, is poorly soluble in water; the calcium salt is freely soluble, barium and lead salts are slightly soluble. Manufactured by sulfonating *o*-dichlorobenzene with oleum [151].

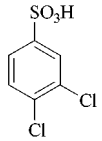

(**18**)

3,4-Dichlorobenzenesulfonyl Chloride [*98-31-7*], $C_6H_3Cl_3O_2S$, M_r 245.51, *fp* 22.4 °C, is manufactured from *o*-dichlorobenzene and chlorosulfuric acid; used as an inexpensive acylating agent.

3,4-Dichlorobenzenesulfonamide [*23815-28-3*], $C_6H_5Cl_2NO_2S$, M_r 226.08, *mp* 140 °C.

2,4,5-Trichlorobenzenesulfonic Acid [*6378-25-2*] (**19**), $C_6H_3Cl_3O_3S$, M_r 261.51, is obtained in the form of the sodium salt by sulfonating 1,2,4-trichlorobenzene with 25 % oleum and pouring the clear sulfonation mixture into a solution of sodium chloride in water; yield 94 – 95 %.

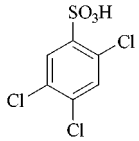

(**19**)

2,4,5-Trichlorobenzenesulfonyl Chloride [*15945-07-0*], $C_6H_2Cl_4O_2S$, M_r 279.96, *mp* 65 – 67 °C (from petroleum ether), $bp_{0.6\,mbar}$ 138 °C, is obtained from trichlorobenzene by chlorosulfonation with chlorosulfuric acid at 90 °C and pouring the reaction mass into water [152–154].

Esters and amides of several chlorobenzenesulfonic acids (particularly 4-chlorobenzenesulfonic acid, 2,5-dichlorobenzenesulfonic acid,

and 2,4,5-trichlorobenzenesulfonic acid) have acquired some importance as crop protection products and agents for the control of fiber pests.

2.6. Aminobenzenesulfonic Acids

The rearrangement of the *N*-phenylsulfamic acid initially formed in the sulfonation of aniline is dependent on temperature. At elevated temperatures it gives 4-aminobenzenesulfonic acid almost quantitatively.

Even at low temperatures, the two other isomers are formed in only relatively small amounts. They are therefore manufactured industrially in other ways (see below).

For identification, aromatic aminosulfonic acids cannot be converted into amides via the sulfonyl chlorides; they must first be converted to the corresponding chlorobenzenesulfonic acids by Sandmeyer's reaction [155], and then derivatized. Today, identification is by HPLC.

2-Aminobenzenesulfonic Acid [*88-21-1*], $C_6H_7NO_3S$, M_r 173.19, is obtained as a byproduct in the manufacture of 3-aminobenzenesulfonic acid by nitration of benzenesulfonic acid at 100 °C, separation of the resulting nitrobenzenesulfonic acids via the magnesium salts, and reduction with iron turnings.

In the reaction between amidosulfuric (sulfamic) acid and excess aniline at temperatures up to 160 °C, 2-aminobenzenesulfonic acid and 4- aminobenzenesulfonic acid are obtained in a ratio of 3 : 2 [156]. The two isomers are separated by means of their different solubilities in aqueous KOH [157].

2-Aminobenzenesulfonic acid is obtained from 2-nitrothiophenol by boiling in a mixture of dioxane and water (ratio 20 : 1) for seven hours; the yield is 86.7 % of theory [158]. The acid can also be produced as follows: 2-chloronitrobenzene is reacted with sodium disulfide in aqueous alcoholic solution to form 2,2′-dinitrodiphenyldisulfide. After having been dissolved in a mixture of hydrochloric acid and nitric acid this is oxidized with chlorine to form 2-nitrobenzenesulfonyl chloride [159], which is then hydrolyzed with soda solution and finally reduced with iron turnings [160]. Alternatively 2,2′-dinitrodiphenyl disulfide can first be reduced to 2,2′-diaminodiphenyl disulfide and then oxidized with hydrogen peroxide in aqueous sulfuric acid to 2-aminobenzenesulfonic acid [161], [162].

2,2′-Diamino-diphenyldisulfide can also be arrived at by alkaline hydrolysis of benzothiazole or mercaptobenzothiazole, followed by oxidation with hydrogen peroxide. The additional oxidation leading to 2-aminobenzenesulfonic acid can be performed in alkaline solution. Thus the synthesis starting from benzothiazole can be carried out exclusively in alkaline solution in a single reactor [163]. Yield: 70 %, based on benzothiazole.

2-Aminobenzenesulfonamide [*3306-62-5*], $C_6H_8N_2O_2S$, M_r 172.21, *mp* 153 °C, crystals (from water), is freely soluble in water, glacial acetic acid, ethyl alcohol, acetone, and methanol; insoluble in benzene.

2-Amino-*N*-methylbenzenesulfonamide [*16288-77-0*], $C_7H_{10}N_2O_2S$, M_r 186.23, is obtained by treating 2-nitrobenzenesulfonyl chloride with methylamine and subsequent reduction. Used as a dye component [164].

The amides of 2-aminobenzenesulfonic acid can be treated with aldehydes or orthoformic acid esters to give 1,2,4-benzothiadiazine 1,1-dioxides or their dihydro compounds; these substances have been proposed as diuretics [165].

2-Amino-*N*-(phenylsulfonyl) benzenesulfonamide, $C_{12}H_{12}N_2O_4S_2$, M_r 312.37, *mp* 193 – 194 °C, is obtained by treating 2-nitrobenzenesulfonyl chloride with benzenesulfonamide at 70 °C in the presence of aqueous sodium hydroxide solution and subsequent reduction of the nitro group [87]. Used as a component of azo dyes.

Phenyl 2-Aminobenzenesulfonate [*68227-69-0*], $C_{12}H_{11}NO_3S$, M_r 249.29, *mp* 70 °C, is obtained by treating 2-nitrobenzenesulfonyl chloride with sodium phenolate and subsequent reduction of the nitro group. Used as an azo dye component.

3-Aminobenzenesulfonic Acid (metanilic acid) [*121-47-1*], $C_6H_7NO_3S$, M_r 173.19, can be obtained by sulfonation of aniline sulfate with oleum and boric acid at 20 – 50 °C. The yield is 98 % [166].

Production. For industrial production, however, nitrobenzene is sulfonated to *m*-nitrobenzenesulfonic acid, which is then reduced either with iron [167], [168] or catalytically with hydrogen [169].

Crude 3-aminobenzenesulfonic acid can be used directly, either as a solution or dried, to obtain 3-aminophenol by alkali fusion. It is not sufficiently pure for use in dye manufacture, however. Therefore the crude solution is acidified; after filtration the product thus obtained is of satisfactory purity.

3-Aminobenzenesulfonic acid is used in the manufacture of dyes and optical brighteners [170]. It is also used to manufacture 3-hydroxybenzenesulfonic acid and 3-aminophenol. The latter is an important intermediate in the manufacture of 4-amino-2-hydroxybenzoic acid (*p*-aminosalicylic acid) used to fight tuberculosis.

3-Aminobenzenesulfonamide [*98-18-0*], $C_6H_8N_2O_2S$, M_r 172.21, *mp* 142 °C, is poorly soluble in cold water.

3-Amino-*N*-(phenylsulfonyl)benzenesulfonamide, $C_{12}H_{12}N_2O_4S_2$, M_r 312.37, is obtained by treating 3-nitrobenzenesulfonyl chloride with benzenesulfonamide and subsequent reduction of the nitro group. It is used as a dye component [171]. The use of the *N*-acrylsulfonamide in the polymerization of acrylonitrile has been proposed to improve the dyeing properties of the fibers [172].

4-Aminobenzenesulfonic Acid (sulfanilic acid) [*121-57-3*], $C_6H_7NO_3S$, M_r 173.19, is made by the reaction of aniline with sulfuric acid at temperatures above 190 °C:

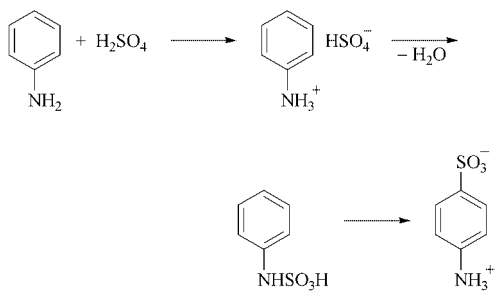

Production. Equimolar amounts of aniline and sulfuric acid are reacted in a lead-lined reactor. The reaction mass is spread on trays and baked in an oven at 190 – 220 °C, the water of the reaction being removed continuously. The process is complete if a sample dissolves clearly in a dilute alkaline solution and has practically no smell of aniline. The crude product is dissolved as a sodium salt in water that has been rendered alkaline with sodium carbonate solution; traces of aniline are removed by blowing with steam. The solution is clarified by filtration and can then be processed further immediately. The baking operation can be carried out in vacuo, which facilitates the removal of the water of the reaction.

In industrial manufacture the aniline hydrogensulfate is placed on trays in trolleys and these are passed at constant speed through a tunnel-shaped oven against an air current at 260 – 280 °C. The aniline hydrogensulfate is thus brought to the baking temperature very gradually. It leaves the oven after 12.5 h [173]. According to [174] the reaction can be carried out in a fluidized bed consisting of sulfanilic acid.

The baking process can also be carried out in a multiphase reactor, with a resultant shortening of reaction time and improvement in quality of the end product [175]. The baking process can be simplified if the formation of the aniline salt and subsequent rearrangement are carried out successively or simultaneously in a solvent, e.g., diphenyl sulfone [67] or 1,2-dichlorobenzene [176]; this enables the reaction mass to be stirred throughout the production process [177]. Particularly high yield and end product purity can be achieved if the reactants are used in a molar ratio of 1 : 1; the rearrangement is carried out under a slight positive pressure (1 – 3 bar) and at 200 – 240 °C; water vapor is expelled from the reactor simultaneously [178]. The reaction time is three hours.

Analysis. The assay is carried out by diazotization. Impurities (aniline, 2,4-disulfanilic acid, 2- and 3-aminobenzenesulfonic acid) are determined by HPLC.

Uses. 4-Aminobenzenesulfonic acid is an important intermediate in the production of dyes, pesticides, pharmaceuticals, etc.

4-Aminobenzenesulfonamide [*63-74-1*], $C_6H_8N_2O_2S$, M_r 172.21, forms the basis for chemotherapeutical sulfonamides.

4-(Acetylamino)benzenesulfonic Acid [*121-62-0*], $C_8H_9NO_4S$, M_r 215.23, is produced by dissolving aminobenzenesulfonic acid in water, adding sodium carbonate, acetylating with acetic anhydride at 45 – 50 °C, and salting out with sodium chloride; the yield is 82 % [179].

4-(Acetylamino)benzenesulfonyl Chloride [*121-60-8*], $C_8H_8ClNO_3S$, M_r 233.67, *mp* 149 °C (from benzene), is relatively stable to water; therefore it can be precipitated from alcoholic solutions with water. It is freely soluble in diethyl ether and ethyl acetate; obtainable as an anhydrous product from the aqueous paste if this is subjected to azeotropic distillation with a water-immiscible solvent at a low temperature.

Production (see also [180]). Acetanilide is introduced into stirred chlorosulfuric acid solution at 20 °C. The solution is then heated to 55 °C and kept at this temperature for 1.5 h. The solution is cooled to 20 °C and introduced as a thin jet into water, the temperature of which is kept at 0 – 5 °C by external cooling. The precipitated sulfochloride is filtered and washed with cold water until neutral to Congo Red. Yield is 80 – 82 %. The crude product contains about 60 % water; the dried product melts at 144 – 147 °C. About 1 – 2 % of bis(acetylaminophenyl) sulfone is obtained as a byproduct.

If thionyl chloride is added, less chlorosulfuric acid is needed because it can be replaced partly by oleum. The product is stable for several days if kept in a cool place. It is generally used directly in the moist state. 4-(Acetylamino)benzenesulfonyl chloride is an important intermediate in the manufacture of sulfonamides and in the synthesis of parabase ester [181].

4-(Acetylamino)benzenesulfonamide [*121–61–9*], $C_8H_{10}N_2O_3S$, M_r 214.24, *mp* 219 °C, is obtained from 4-acetylaminobenzenesulfonyl chloride and aqueous ammonia at temperatures not exceeding 30 °C. Uses include the manufacture of sulfanilamide.

5-Amino-1,3-benzenedisulfonic Acid [*5294-05-3*] (**20**), $C_6H_7NO_6S_2$, M_r 253.25, is obtained by reduction of 5-nitro-1,3-benzenedisulfonic acid.

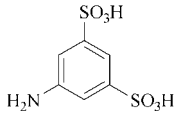

(**20**)

5-Amino-1,3-benzenedisulfonyl Chloride, $C_6H_5Cl_2NO_4S_2$, M_r 290.14, is obtained from the sodium salt or free acid with chlorosulfuric acid at 100 – 160 °C [182].

5-Amino-*N,N*-bis(2-chloroethyl)-1,3-benzene-disulfonamide [*22480-69-9*], $C_{10}H_{15}Cl_2N_3O_4S_2$, M_r 376.28, *mp* 126 – 128 °C, is formed by treating the dichloride with 2-chloroethylamine [183], [184].

4-Amino-1,3-benzenedisulfonic Acid [*137-51-9*] (**21**), $C_6H_7NO_6S_2$, M_r 253.25, crystallizes with two moles of H_2O. The compound is freely soluble in water and ethyl alcohol. It is prepared by sulfonating 4-aminobenzenesulfonic acid with oleum at 20 °C and 130 °C (4 – 6.5 h), draining the sulfonation mixture into water under pressure, adding lime at 70 – 75 °C, separating the resulting gypsum, precipitating the calcium ions with soda, removing the calcium carbonate formed, and concentrating the sodium salt solution by evaporation. Yield 90 – 91 % [185].

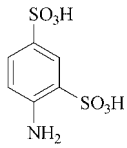

(**21**)

4-Amino-1,3-benzenedisulfonamide [*40642–90–8*], $C_6H_9N_3O_4S_2$, M_r 251.28, *mp* 235 °C, crystals (from water). The compound is used as a dye intermediate.

2-Amino-1,4-benzenedisulfonic Acid [*98-44-2*] (**22**), $C_6H_7NO_6S_2$, M_r 253.25, is obtained by sulfonating metanilic acid in 100 % sulfuric acid with 50 % oleum at 160 °C (5 – 6 h); the sulfonation mixture is diluted with water, lime is added, and the resulting calcium sulfate

is removed; the product is then precipitated with excess concentrated hydrochloric acid in hot solution; yield 90 – 91 % [186].

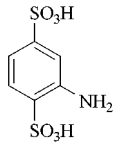

(**22**)

The trisulfonic and tetrasulfonic acids formed in this sulfonation process are unstable in hot aqueous solution. When water is added to the sulfonation mass they eliminate sulfonic acid groups, thus forming the end product; see also [187]. It is also possible to start from 4-chloro-3-nitrobenzenesulfonic acid, in which case sulfite is used to replace the chlorine by the sulfonic acid group and the resulting 2-nitro-1,4-benzenedisulfonic acid is then reduced [188].

Compound **22** is used as an intermediate for dyes; for example, see [189]. 2-Amino-1,4-benzenedisulfonic acid is converted to triazine derivatives, which are intermediates in the manufacture of dyes and optical brighteners [190], [191].

2,4-Diaminobenzenesulfonic Acid [*88-63-1*] (**23**), $C_6H_8N_2O_3S$, M_r 188.2, monoclinic plates or triclinic prisms, poorly soluble in cold water, is obtained by sulfonating a solution of *m*-phenylenediamine in 100 % sulfuric acid with oleum at 155 °C, adding water to the sulfonation mixture, desulfonating the resulting disulfonic acids at 140 °C, and isolating the free acid at 10 °C; yield 93 % [192]. Another process starts from 1-chloro-2,4-dinitrobenzene, which, as an alcoholic solution, is reacted with sodium sulfite solution to form 2,4-dinitrobenzenesulfonic acid [*89-02-1*], from which the end product is obtained by reduction with iron according to BÉCHAMP [188, pp. 99, 100].

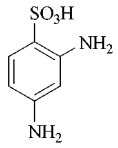

(**23**)

4,6-Diamino-1,3-benzenedisulfonic Acid [*137-50-8*] (**24**), $C_6H_8N_2O_6S_2$, M_r 268.27, is obtained from *m*-phenylenediamine by sulfonation with excess oleum at 70 – 130 °C. Used as an intermediate for dyes [193], [194].

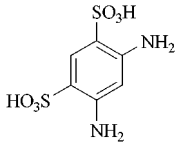

(**24**)

2,5-Diamino-1,4-benzenedisulfonic Acid [*7139-89-1*] (**25**), $C_6H_8N_2O_6S_2$, M_r 268.27, is obtained from *p*-phenylenediamine and 25 % oleum at 140 °C [195] or by heating *p*-phenylenediamine sulfate and 98 % sodium hydrogen sulfate in a mixture of 1,2,4-trichlorobenzene and 1-chloronaphthalene to the boiling point of the solvent mixture (215 °C), removing the solid product, dissolving it in water, neutralizing it with sodium hydroxide solution to pH 7.5 – 8, and clarifying the solution with kieselguhr. The product contains 5 % of monosulfonic acid, but only a small amount of the isomeric 2,6-disulfonic acid; yield 90 % [196]. The compound is used as an intermediate for dyes.

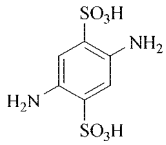

(**25**)

2.7. Hydroxybenzenesulfonic Acids

The chemical properties of the phenolsulfonic acids are characterized by the reactive aromatic hydroxyl group. The reactions of these acids therefore include azo coupling reactions with diazonium salts to form azo dyes. Hydroxybenzenesulfonic acids with more than one hydroxyl group on the benzene nucleus form complexes with metal ions.

2-Hydroxybenzenesulfonic Acid [*609-46-1*] (**26**), $C_6H_6O_4S$, M_r 174.17, *mp* 145 °C (mono-

hydrate); aniline salt *mp* 165 °C; potassium salt (dihydrate) *mp* 235 – 240 °C. If phenol is sulfonated with an equal amount of 100 % sulfuric acid under mild conditions, a mixture consisting of 2-hydroxy- and 4-hydroxybenzenesulfonic acid in a ratio of roughly 2 : 3 is obtained. The 2-hydroxybenzenesulfonic acid can be isolated from this mixture via the monobarium salt [197]. The pure compound is obtained by diazotizing 2-aminobenzenesulfonic acid and boiling it under acid conditions.

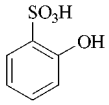

(**26**)

3-Hydroxybenzenesulfonic Acid [*585-38-6*] (**27**), $C_6H_6O_4S$, M_r 174.17; potassium salt (monohydrate) *mp* 200 – 210 °C; sodium salt (from ethyl alcohol) *mp* 314 °C, is synthesized by the reaction of 1,3-benzenedisulfonic acid and sodium hydroxide in water in an autoclave for 30 h at 250 °C. The reaction mass is neutralized and evaporated to dryness. The sodium salt of the acid is extracted with 78 % ethyl alcohol; yield 78 % [198]. The crude solution obtained by acidifying the reaction mass with hydrochloric acid and freeing it from sulfur dioxide by boiling can be used in the manufacture of dyes [199].

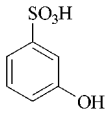

(**27**)

4-Hydroxybenzenesulfonic Acid [*98-67-9*] (**28**), $C_6H_6O_4S$, M_r 174.17, aniline salt *mp* 170 °C, *S*-benzylisothiuronium salt (monohydrate) *mp* 168.7 °C.

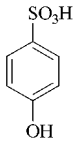

(**28**)

Production: phenol (25 kg) and 96 % sulfuric acid (28 kg) are heated to 50 °C; then 1.25 kg additional sulfuric acid is added and the reaction mixture is heated at 110 °C for 5 – 6 h. Reaction water, together with about 5 % of the phenol, is distilled off during this period. Yield 95 % [200].

Crude 4-hydroxybenzenesulfonic acid obtained by hot sulfonation, i.e., mixed with 2-hydroxybenzenesulfonic acid, is used as such in the manufacture of synthetic tanning agents. Condensation products of phenolsulfonic acid mixtures with formaldehyde improve the dyeing properties of polyamide fibers [201].

4-Hydroxybenzenesulfonic acid and 2-halogenomalonic esters form the corresponding ethers, which are used in the manufacture of polyesters with improved affinity for basic dyes [202].

If a mixture of hydroxybenzenesulfonic acids and bis(hydroxyphenyl) sulfone is condensed with formaldehyde, and hydrogensulfite is subsequently added, substances with good dispersing effects on dyes are obtained. They also are used as thinning agents for cement [203], [204].

Most of the 4-hydroxybenzenesulfonic acid is used as an additive for electroplating baths. By comparison, the acid is relatively unimportant as a dye intermediate. With ethylene oxide it forms poly(ethoxy)oxybenzenesulfonic acid, which can be used as a plasticizer and emulsifier.

4-Methoxybenzenesulfonic Acid

[*5857-42-1*] (**29**), $C_7H_8O_4S$, M_r 188.20, ammonium salt *mp* 285 °C. The acid is obtained by sulfonation of anisole. The ammonium salt of (**29**) can be obtained in high yield by treating anisole with an equal mass of sulfamic acid at 140 – 150 °C [205]. The compound is used in the manufacture of 4-methoxybenzenesulfonylureas, which have hypoglycemic effects.

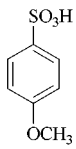

(**29**)

4-Hydroxy-1,3-benzenedisulfonic Acid [*96-77-5*] (**30**), $C_6H_6O_7S_2$, M_r 254.24, hy-

groscopic crystals, is obtained by hydrolysis of the dichloride with ethyl alcohol and/or water. Nitration and reduction of this acid gives 5-amino-4-hydroxy-1,3-benzenedisulfonic acid. 4-Hydroxy-1,3-benzenedisulfonic acid accelerates the rearrangement of cumene hydroperoxide in the Hock – Lang phenol process [206].

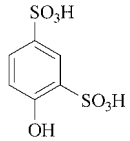

(**30**)

4-Hydroxy-1,3-benzenedisulfonyl Chloride [*1892-33-7*], $C_6H_4Cl_2O_5S_2$, M_r 291.13, *mp* 89 °C, needles (from gasoline), is obtained by treating phenol with chlorosulfuric acid and introducing the sulfonation mixture into concentrated hydrochloric acid [207].

2,4-Dihydroxybenzenesulfonic Acid [*6409-58-1*] (**31**), $C_6H_6O_5S$, M_r 190.17, ammonium salt (from ethyl alcohol) *mp* 190 °C [208]. The acid is prepared by sulfonating resorcinol with 96 % sulfuric acid in a molar ratio of 1 : 1. The compound is used as a catalyst in the curing of resins [209].

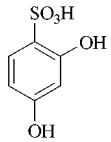

(**31**)

3,4-Dihydroxybenzenesulfonic Acid [*7134-09-0*] (**32**), $C_6H_6O_5S$, M_r 190.17, deliquescent crystals; the ammonium salt (flakes, from aqueous ethanol) melts at 260 °C (decomp.). Synthesis: 0.3 mol of disodium 4-hydroxy-1,3-benzenedisulfonate is reacted with 1.2 mol of sodium hydroxide for 36 min at 340 °C; the reaction mass is diluted with water, acidified with sulfuric acid, and boiled to remove sulfur dioxide; the product is extracted with ether. Yield 80 %. Pyrocatechol is obtained by hydrolyzing the sulfonic acid group at 190 °C [210].

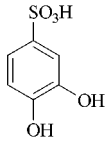

(**32**)

4,5-Dihydroxy-1,3-benzenedisulfonic Acid [*149-46-2*] (**33**), $C_6H_6O_8S_2$, M_r 270.23, is produced by sulfonation of pyrocatechol with oleum [211] or by alkali fusion of 2-hydroxy-1,3,5-benzenetrisulfonic acid: 10 kg of the tetrasodium salt of phenoltrisulfonic acid is introduced into 16 kg of sodium hydroxide at about 20 °C. After adding some water, the temperature is gradually raised to 160 °C; at the end of the reaction the melt is diluted with water and acidified with sulfuric acid, and sulfur dioxide is driven off. After cooling, precipitated sodium sulfate is removed by filtration and excess sulfate ions are precipitated in fractions. Finally the barium salt, which has poor solubility in cold water, is precipitated.

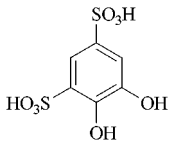

(**33**)

The disodium salt of **33**, which is referred to in the literature as Tiron [212], serves to detect traces of iron in blood and is used as a reagent for colorimetric determination of Ti^{4+} and Mo^{4+}. The ability to form water-soluble complex salts, e.g., with antimony, has been exploited in pharmacology [213]. Tiron is claimed to improve the storability of silver halogenide emulsions [214].

2.8. Chloro-alkylbenzenesulfonic Acids

3-Chloro-4-methylbenzenesulfonic Acid [*98-34-0*] (**34**), $C_7H_7ClO_3S$, M_r 206.65, decomposes above 110 °C. Production is from toluene by sulfonation with anhydrous sulfuric acid, followed by chlorination with gaseous chlorine [215].

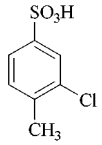

(**34**)

3-Chloro-4-methylbenzenesulfonic acid is an intermediate in the manufacture of 2-chlorotoluene, 5-chloro-4-methyl-2-nitrobenzenesulfonic acid, 2,4-dichlorobenzoic acid, and 2-chloro-4-sulfobenzoic acid.

3-Chloro-4-methylbenzenesulfonyl Chloride [*42413-03-6*], $C_7H_6Cl_2O_2S$, M_r 225.09, $bp_{32\,mbar}$ 166 °C, *mp* 38 °C, is obtained by chlorination of 4-methylbenzenesulfonyl chloride in the presence of iron trichloride or antimony trichloride.

A high yield of sulfochloride with only a small percentage of isomers is obtained if toluene is sulfonated with sulfur trioxide in 1,2-dichloroethane. The resulting sulfonic acid is chlorinated in the presence of iodine, and the resulting chlorobenzenesulfonic acid is reacted with chlorosulfuric acid. The crude product can be used immediately in the production of 2,4-dichloro-α,α,α-trifluorotoluene, which is a precursor of several herbicides [216].

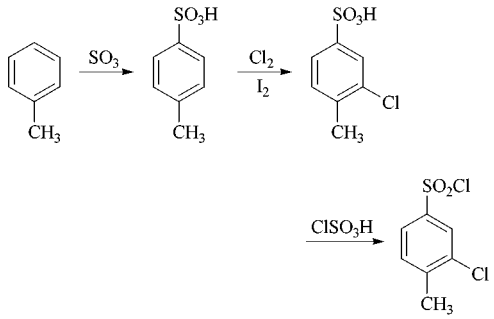

3-Chloro-4-methylbenzenesulfonamide [*51896-27-6*], $C_7H_8ClNO_2S$, M_r 205.66, melts at 137 °C.

5-Chloro-2-methylbenzenesulfonic Acid [*133-73-3*] (**35**), $C_7H_7ClO_3S$, M_r 206.65; the sodium salt crystallizes from water with 1/2 mol of H_2O. The acid is formed as the main product in the sulfonation of 4-chlorotoluene with 100 % sulfuric acid at 100 °C. It is purified by recrystallizing the barium salt. The compound is used in the production of 5-chloro-2-formylbenzenesulfonic acid.

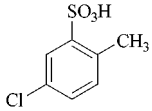

(**35**)

4-Chloro-3-(trifluoromethyl)benzenesulfonyl Chloride [*32333-53-2*] (**36**), $C_7H_3Cl_2F_3O_2S$, M_r 279.06, $bp_{16\,mbar}$ 128 – 131 °C, n_D^{20} 1.5158, is obtained from 2-chloro-α,α,α-trifluorotoluene and chlorosulfuric acid in 65 % oleum [217]. The compound is used as an intermediate in the production of pesticides and dyes [217].

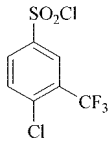

(**36**)

3,5-Dichloro-4-methylbenzenesulfonic Acid [*2225-18-5*], $C_7H_6Cl_2O_3S$, M_r 241.09, is obtained (in addition to tri- and tetrachlorotoluenesulfonic acids) from 4-toluenesulfonic acid by chlorination with chlorine or sulfuryl chloride in sulfuric acid [218], or by hydrolysis of the sulfonyl chloride (see next paragraph). The compound is used as an intermediate in the manufacture of 2,6-dichlorotoluene.

3,5-Dichloro-4-methylbenzenesulfonyl Chloride [*24653-79-0*], $C_7H_5Cl_3O_2S$, M_r 259.54, is formed from 4-toluenesulfonyl chloride by chlorination with chlorine, antimony trichloride being added as a catalyst [219]. The compound is used as an intermediate in the manufacture of 2,6-dichlorotoluene [218].

2.9. Amino-alkylbenzenesulfonic Acids

2-Amino-5-methylbenzenesulfonic Acid [*88-44-8*] (**37**), $C_7H_9NO_3S$, M_r 187.22, crystallizes with 1 mol of H_2O. It is produced by baking *p*-toluidine sulfate at 180 °C and 97 – 101 mbar, dissolving the product in 3 % sodium hydroxide solution, and precipitating the free acid by adding concentrated hydrochloric acid [220, 222].

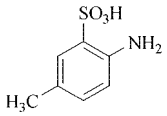

(**37**)

2-Amino-3,5-dimethylbenzenesulfonic Acid [*88-22-2*] (**38**), $C_8H_{11}NO_3S$, M_r 201.24; the barium salt crystallizes from water with 2 mol of H_2O; the lead salt is very poorly soluble in water. Production is from 4-amino-1,3-dimethylbenzene (*m*-xylidine) and 96 % sulfuric acid by baking the sulfate paste in vacuo, subsequently purifying the product with boiling milk of lime, and finally precipitating the sodium salt with sodium chloride [223]. 2-Amino-3,5-dimethylbenzenesulfonic acid can also be produced by sulfonation of *m*-xylidine in 1,1,2,2-tetrachloroethane with chlorosulfuric acid and subsequent boiling until no more hydrogen chloride is formed.

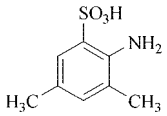

(**38**)

3-Amino-4-methylbenzenesulfonic Acid [*618-03-1*] (**39**), and **5-Amino-2-methylbenzenesulfonic Acid** [*118-88-7*] (**40**), $C_7H_9NO_3S$, M_r 187.22; both compounds are produced analogously to the 3-aminobenzenesulfonic acids from the corresponding nitrotoluenesulfonic acids by Béchamp reduction.

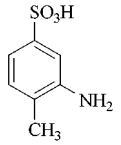

(**39**)

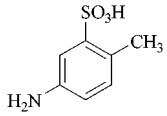

(**40**)

4-Amino-3-methylbenzenesulfonic Acid [*98-33-9*] (**41**), $C_7H_9NO_3S$, M_r 187.22, crystallizes with 1 mol of H_2O. It is prepared from *o*-toluidine and 98 % H_2SO_4 via the sulfate, which is then baked. The sodium salt can be isolated after liming and precipitation of calcium sulfate [224]. The solution can be used directly in the manufacture of dyes [225].

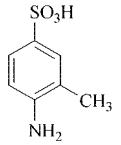

(**41**)

4-Amino-2-methylbenzenesulfonic Acid [*133-78-8*] (**42**), $C_7H_9NO_3S$, M_r 187.22, crystallizes from water with 3 mol of H_2O. It is produced analogously to 4-amino-3-methylbenzenesulfonic acid by baking of *m*-toluidine sulfate [226].

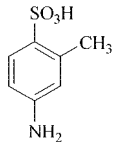

(**42**)

2.10. Amino- and Hydroxycarboxybenzenesulfonic Acids

2-Amino-4-sulfobenzoic Acid [*98-43-1*] (**43**), $C_7H_7NO_5S$, M_r 217.2, 2,4-sulfanthranilic acid, crystallizes with 1 mol of H_2O; it is freely soluble in hot water, less soluble in cold water. The compound is obtained at a moderate yield by treating 4-methyl-3-nitrobenzenesulfonic acid with sodium hydroxide solution:

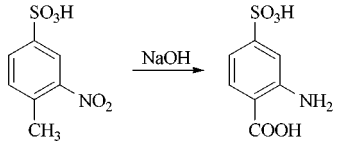

43

This acid is used mainly in the manufacture of diazoamino compounds of the R–N=N–NH–R′ type. As mixtures with naphthols these compounds have acquired importance as printing inks (Rapidogen dyes).

2-Amino-5-sulfobenzoic Acid [*3577-63-7*], (**44**) $C_7H_7NO_5S$, M_r 217.2, is produced from anthranilic acid by sulfonation with chlorosulfuric acid in anhydrous sulfuric acid or nitrobenzene or by baking the sulfate of anthranilic acid at 200 °C [227], [228]. This compound is used, like its *N*-alkyl derivatives, as a stabilizer for diazo compounds (see next paragraph).

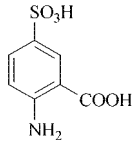

(**44**)

2-Isobutylamino-5-sulfobenzoic Acid, $C_{11}H_{15}NO_5S$, M_r 273.31, is obtained in crystalline form by sulfonating *N*-isobutylanthranilic acid in 98 % sulfuric acid with 65 % oleum at 80 °C and pumping the sulfonation mixture into water; yield 93 – 94 % [229].

2-Hydroxy-5-sulfobenzoic Acid [*97-05-2*] (**45**), $C_7H_6O_6S$, M_r 218.18, needles, *mp* 120 °C, soluble in water, ethyl alcohol, and ether, is obtained by heating one part of salicylic acid with 5 parts of concentrated sulfuric acid [230]. It can also be prepared from the sulfochloride.

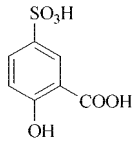

(**45**)

2-Hydroxy-5-sulfobenzoic acid forms a dihydrate and is used in analytical chemistry to detect albumins, iron, and mercury [231]. Applications described in patents include the use as a heat stabilizer for polyacrylonitrile [232], as an additive used in the anodizing of metals [233–235], and – as a tributyltin compound – as an agent for the antibacterial protection of organic materials [236]. The acid is also claimed to be suitable for a pregnancy test for horses [237].

5-(Chlorosulfonyl)-2-hydroxybenzoic Acid [*17243-13-9*], $C_7H_5ClO_5S$, M_r 236.63, *mp* 171 °C (decomp.), is obtained from salicylic acid and chlorosulfuric acid [238], [239]. The compound is used to manufacture the corresponding sulfinic acid and sulfones.

2.11. Chloro-hydroxybenzenesulfonic Acids

3-Chloro-4-hydroxybenzenesulfonic Acid [*46060-27-9*] (**46**), $C_6H_5ClO_4S$, M_r 208.62, is prepared by chlorination of 4-hydroxybenzenesulfonic acid or by sulfonation of 2-chlorophenol with 98 % sulfuric acid [240]. The compound can be converted by nitration into 3-chloro-4-hydroxy-5-nitrobenzenesulfonic acid, from which 3-amino-5-chloro-4-hydroxybenzenesulfonic acid is obtained by reduction. The sulfonic acid group is eliminated easily at 180 – 190 °C.

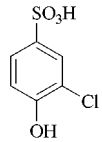

(**46**)

5-Chloro-2-hydroxybenzenesulfonic Acid [*2051-65-2*] (**47**), $C_6H_5ClO_4S$, M_r 208.62, is obtained from 2,5-dichlorobenzenesulfonic acid by partial hydrolysis with sodium hydroxide solution in water at 170 – 190 °C and precipitation with concentrated hydrochloric acid. Used to prepare 5-chloro-2-hydroxy-3-nitrobenzenesulfonic acid, which is then reduced to 3-amino-5-chloro-2-hydroxybenzenesulfonic acid.

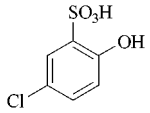

(**47**)

3,5-Dichloro-2-hydroxybenzenesulfonic Acid [*26281-43-6*] (**48**), $C_6H_4Cl_2O_4S$, M_r 243, forms a poorly water-soluble barium salt.

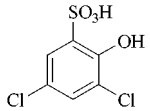

(**48**)

3,5-Dichloro-2-hydroxybenzenesulfonamide [*35337-99-6*], $C_6H_5Cl_2NO_3S$, M_r 242.08, melts at 230 °C. It is obtained from the chlorosulfonation of 2,4-dichlorophenol and conversion with ammonia or NH_4Cl. It can be used as a precursor to synthesize some 2-hydroxybenzenesulfonamides that are poorly accessible by other routes, by dehalogenation with H_2 and Pd/C catalyst [241].

2.12. Aminohydroxybenzenesulfonic Acids and Aminochlorohydroxybenzenesulfonic Acids

All the compounds described in this section are used in the manufacture of dyes.

3-Amino-4-hydroxybenzenesulfonic Acid [*98-37-3*] (**49**), $C_6H_7NO_4S$, M_r 189.19, is freely soluble in hot water, slightly soluble in cold, dilute hydrochloric acid. Obtained by boiling 4-chloro-3-nitrobenzenesulfonic acid with sodium hydroxide solution, followed by Béchamp reduction [242]:

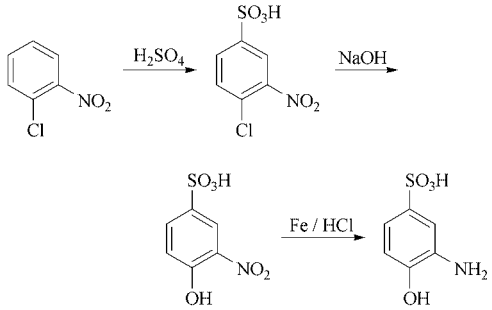

49

3-Amino-4-hydroxybenzenesulfonic acid is an important intermediate in the production of dyes and is used to manufacture Diamond Black PV and 3-amino-4-hydroxy-5-nitrobenzenesulfonic acid.

3-Amino-4-hydroxybenzenesulfonamide [*98-32-8*], $C_6H_8N_2O_3S$, M_r 188.2. Manufacture: 2-nitrochlorobenzene is sulfochlorinated with chlorosulfuric acid; the resulting sulfochloride is treated with ammonia to give 4-chloro-3-nitrobenzenesulfonamide; chlorine is replaced by hydroxyl by heating the product at 100 °C in aqueous sodium hydroxide solution. Finally, the hydroxybenzenesulfonamide is reduced with iron turnings [243].

3-Amino-4-methoxybenzenesulfonic Acid [*98-42-0*] (**50**), $C_7H_9NO_4S$, M_r 203.22, needles, is freely soluble in water. The compound is obtained from 2-anisidine and 98 % sulfuric acid at 50 – 55 °C [244].

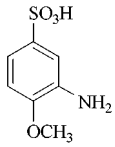

(**50**)

3-Acetylamino-4-methoxybenzenesulfonyl Chloride [*3746-67-6*] (**51**), $C_9H_{10}ClNO_4S$, M_r 263.7, *mp* 152 – 163 °C, is obtained from 2-acetylaminoanisole and chlorosulfuric acid at 35 – 40 °C [245–247]. The compound is used to produce the corresponding sulfinic acid and Fast Red ITR base.

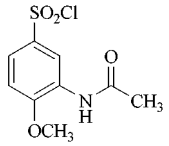

(**51**)

3-Acetylamino-4-methoxybenzenesulfonamide [*85605-29-4*], $C_9H_{12}N_2O_4S$, M_r 244.27, melts at 226 °C.

3-Acetylamino-4-methoxy-*N*-methylbenzenesulfonamide, $C_{10}H_{14}N_2O_4S$, M_r 258.3, is obtained by treating the chloride with 2 mol of ethylamine at 25 °C. Other *N*-alkylamides of this sulfonic acid are produced in the same way [248].

5-Amino-4-hydroxy-1,3-benzenedisulfonic Acid [*120-98-9*] (**52**), $C_6H_7NO_7S_2$, M_r 269.25, is obtained by nitration of 4-hydroxy-1,3-benzenedisulfonic acid and reduction of the hydroxynitrobenzenedisulfonic acid [84, p. 148].

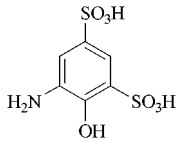

(**52**)

5-Acetylamino-2-hydroxybenzenesulfonic Acid [*55034-25-8*] (**53**), $C_8H_9NO_5S$, M_r 231.23, needles, freely soluble in water, decomposition on melting. Manufactured by sulfonating *N*-(4-hydroxyphenyl)acetamide with 98 % sulfuric acid at 80 °C [249] or by acetylating 5-amino-2-hydroxybenzenesulfonic acid with acetic anhydride and sodium acetate in water.

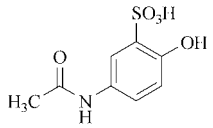

(**53**)

3-Amino-6-ethoxybenzenesulfonic Acid [*6375-02-6*] (**54**), $C_8H_{11}NO_4S$, M_r 217.24, needles, poorly soluble in water and ethyl alcohol. Manufactured by treating *o*-dichlorobenzene with 96 % sulfuric acid and baking the resulting sulfate or from phenetidine by reaction with oleum [250].

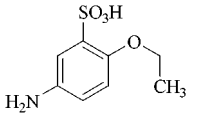

(**54**)

4-Amino-6-methoxy-1,3-benzenedisulfonic Acid (**55**), $C_7H_9NO_7S_2$, M_r 283.28.

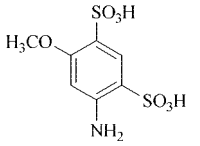

(**55**)

4-Amino-6-methoxy-1,3-benzenedisulfonyl Chloride [*670-02-0*], $C_7H_7Cl_2NO_5S_2$, M_r 320.17, *mp* 129 – 130 °C, is obtained by adding sodium chloride to *m*-anisidine and chlorosulfuric acid while cooling with ice and then heating [251].

3-Amino-5-chloro-2-hydroxybenzenesulfonic Acid [*88-23-3*] (**56**), $C_6H_6ClNO_4S$, M_r 223.63, colorless crystals, freely soluble in hot water, less soluble in cold water, very easily soluble in sodium hydroxide solution. The industrial product has a weak brownish-pink color and is produced by sulfonating 2-amino-4-chlorophenol or from *p*-chlorophenol:

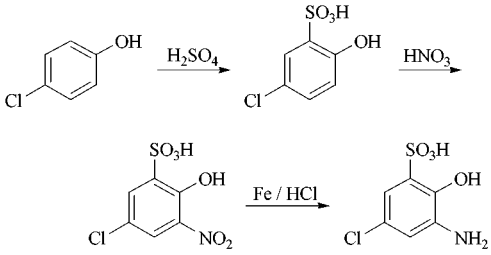

56

The 5-chloro-2-hydroxy-3-nitrobenzenesulfonic acid obtained from *p*-chlorophenol by sulfonation and nitration is reduced with iron and hydrochloric acid without being precipitated. After filtering from the iron sludge, the highly alkaline reduction liquor is concentrated by evaporation and decolorized with activated carbon. Upon acidification 3-amino-5-chloro-2-hydroxybenzenesulfonic acid precipitates. Yield 72 %, based on *p*-chlorophenol. The acid is an important intermediate for dyes. It is used mainly in the manufacture of Diamond Black P 2 B and of acid chrome blue dyes.

3-Amino-5-chloro-4-hydroxybenzenesulfonic Acid (**57**) is similar in properties and manufacture to 3-amino-5-chloro-2-hydroxybenzenesulfonic acid but is less important industrially [252].

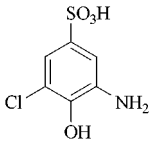

(**57**)

2.13. Amino-chlorobenzenesulfonic Acids

2-Amino-5-chlorobenzenesulfonic Acid [*133-74-4*] (**58**), $C_6H_6ClNO_3S$, M_r 207.64, decomposes at 280 °C and crystallizes from water with 1 mol of H_2O. Manufactured by baking 4-chloroaniline hydrogensulfate in vacuo; after the mass has been pumped into water, the acid is neutralized with calcium oxide; the solution is suitable for use in the manufacture of azo dyes [253].

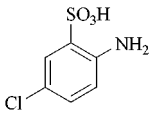

(**58**)

3-Amino-4-chlorobenzenesulfonic Acid [*98-36-2*] (**59**), $C_6H_6ClNO_3S$, M_r 207.64, needles (from water), crystallizes with 1 mol of H_2O. It is obtained from 4-chloro-3-nitrobenzenesulfonic acid by reduction; the compound is used for the production of azo dyes.

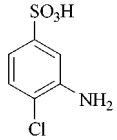

(**59**)

2-Amino-4,5-dichlorobenzenesulfonic Acid [*6331-96-0*] (**60**), $C_6H_5Cl_2NO_3S$, M_r 242.08, crystallizes as needles from water. Manufactured by baking 3,4-dichloroaniline sulfate. The compound is used as a component of azo dyes.

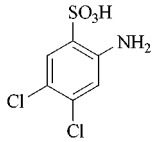

(**60**)

2-Amino-4,5-dichlorobenzenesulfonyl Chloride [*36110-12-0*], $C_6H_4Cl_3NO_2S$, M_r 260.53, is obtained by treating 3,4-dichloroaniline with chlorosulfuric acid in the presence of sodium chloride.

2-Amino-4,5-dichlorobenzenesulfonamide [*16948-63-3*], $C_6H_6Cl_2N_2O_2S$, M_r 241.1, *mp* 183 – 185 °C, crystals from 50 % alcohol, is obtained by treating the chloride with aqueous ammonia [254].

4-Amino-2,5-dichlorobenzenesulfonic Acid [*88-50-6*] (**61**), $C_6H_5Cl_2NO_3S$, M_r 242.08, is obtained by sulfonating 2,5-dichloroaniline with 65 % oleum in anhydrous sulfuric acid; the compound is purified via the sodium salt [255].

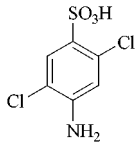

(**61**)

4-Amino-6-chloro-1,3-benzenedisulfonic Acid (**62**), $C_6H_6ClNO_6S_2$, M_r 287.7.

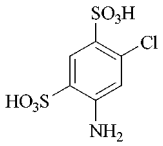

(**62**)

4-Amino-6-chloro-1,3-benzenedisulfonyl Chloride [*671-89-6*], $C_6H_4Cl_3NO_4S_2$, M_r 324.59, *mp* 142 – 144 °C, is obtained by chlorosulfonation of 3-chloroaniline with 11 mol of chlorosulfuric acid at 130 °C [256–258]; after the chlorosulfonation mixture is treated with 4 mol of thionyl chloride at 80 °C, the yield is 90 % [259].

4-Amino-6-chloro-1,3-benzenedisulfonamide [*121-30-2*], $C_6H_8ClN_3O_4S_2$, M_r 285.73, *mp* 258 – 260 °C, is obtained by treating the chloride with ammonia in *tert*-butanol and concentrating the solution by evaporation; the yield is 80 % [259]. Condensation with formic acid gives 6-chloro-7-sulfamoyl-1,2,4-benzothiadiazine 1,1-dioxide, an important diuretic [260], [261]. The 1-monoamide **63** is an intermediate for the production of azo dyes for polyamide fibers. For manufacture, see [262].

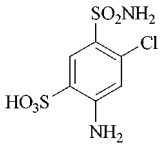

(**63**)

4-Amino-6-chloro-N^1,N^2-dimethyl-1,3-benzene-disulfonamide [*60385-32-2*], $C_8H_{12}ClN_3O_4S_2$, M_r 313.78, *mp* 183 °C (crystals from aqueous alcohol), is prepared by boiling the chloride with aqueous methylamine in dioxane at reflux temperature for 3 h and precipitating with water [256].

2.14. Amino-chloro-alkylbenzenesulfonic Acids

2-Amino-5-chloro-4-methylbenzenesulfonic Acid [*88-53-9*] (**64**), $C_7H_8ClNO_3S$, M_r 221.66, is a colorless to slightly reddish crystalline powder. The compound is obtained by chlorination of *p*-toluenesulfonic acid to form 3-chloro-4-methylbenzenesulfonic acid, subsequent nitration of this acid to a mixture of nitro isomers, precipitation of the 2-chloro-4-methyl-6-nitrobenzenesulfonic acid with sodium chloride solution, and, finally, Béchamp reduction of the nitro group with iron and acetic acid [263].

A different synthesis route in which no isomers are formed is as follows [264]:

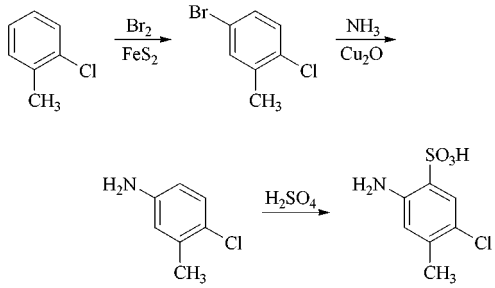

64

The compound is used as Lake Red C and CLT Acid as a diazo component for azo dyes and azo pigments.

3-Amino-5-chloro-4-methylbenzenesulfonic Acid [*6387-27-5*] (**65**), $C_7H_8ClNO_3S$, M_r 221.66, is a moist, slightly reddish paste in raw form. It is manufactured by sulfonation of 1-chloro-2- methyl-3-nitrobenzene [*83-42-1*] with oleum, followed by Béchamp reduction [265]. The compound is used as a diazo component for azo dyes in the dyeing of leather.

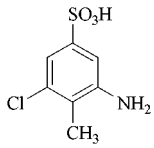

(**65**)

4-Amino-5-chloro-3-methylbenzenesulfonic Acid [*6387-14-0*] (**66**), $C_7H_8ClNO_3S$, M_r 221.66, needles from water, splits off the sulfonic acid group in 75 % sulfuric acid at

150 – 160 °C. It is made by sulfonation of 2-chloro-6-methylaniline in *o*-dichlorobenzene with anhydrous sulfuric acid in vacuo, during which the water of the reaction distills off [266]. Alternative: chlorination of 4-acetylamino-3-methylbenzenesulfonic acid in water and deacetylation. The compound is used as a dye component.

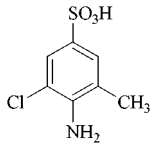

(**66**)

4-Amino-5-chloro-2-methylbenzenesulfonic Acid (**67**), $C_7H_8ClNO_3S$, M_r 227.66, is obtained from *m*-toluidine sulfate by chlorination, followed by sulfonation in an inert solvent [267].

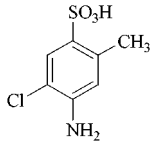

(**67**)

2-Amino-4-chloro-5-methylbenzenesulfonic Acid [*88-51-7*] (**68**), $C_7H_8ClNO_3S$, M_r 221.66, is insoluble in cold water. The compound is made by baking 3-chloro-4-methylbenzeneamine hydrogensulfate in vacuo at 195 °C [268], and it is used for dyes; see, for example, [269].

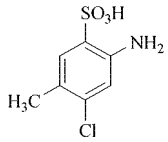

(**68**)

6-Amino-2,4-dichloro-3-methylbenzenesulfonic Acid (**69**), $C_7H_7Cl_2NO_3S$, M_r 256.11, is obtained by sulfonation of 3,5-dichloro-4-methylaniline in *o*-dichlorobenzene with chlorosulfuric acid [270].

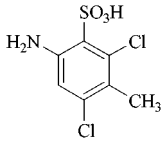

(**69**)

3. References

General References

1. B. T. Brooks, C. E. Brooks, S. S. Kurtz, L. Schmerling: "Sulfonation of Aromatic Hydrocarbons," in: *The Chemistry of Petroleum Hydrocarbons,* **vol. III,** Reinhold, New York 1955.
2. H. Cerfontain: *Mechanistic Aspects in Aromatic Sulfonation and Desulfonation,* J. Wiley & Sons, New York 1968.
3. E. E. Gilbert: "The Reaction of Sulfur Trioxide and of its Adducts with Organic Compounds," *Chem. Rev.* **62** (1962) 549 – 589.
4. C. W. Suter: *The Organic Chemistry of Sulfur. Tetravalent Sulfur Compounds,* J. Wiley & Sons, New York 1945.
5. C. M. Suter, A. W. Weston: "Direct Sulfonation of Aromatic Hydrocarbons and their Halogen Derivatives," in: *Organic Reactions,* **vol. III,** J. Wiley & Sons, New York 1946, pp. 141 – 197.
6. N. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* Akademie-Verlag, Berlin 1966.
7. E. E. Gilbert: *Sulfonation and Related Reactions,* Interscience Publ., New York 1964.
8. K. LeRoi Nelson in G. A. Olah (ed.): *Friedel-Crafts and Related Reactions,* **vol. 3,** Interscience Publ., New York 1964, pp. 1355 – 1392.
9. F. Muth: "Methoden zur Herstellung und Umwandlung aromatischer Sulfonsäuren," in: *Houben-Weyl,* **vol. 9,** 1955, pp. 435 – 556.
10. W. J. Hickingbottom, in *Rodd's Chemistry of Carbon Compounds* **, vol. III A,** Elsevier, Amsterdam 1954, pp. 230 – 248.
11. C. Grundmann: "Desulfonation," in *Houben-Weyl,* **vol. 5/2 b,** 1981, pp. 354 – 357.
12. H. Cerfontain et al.: "Aromatic Sulfonation" (a series of articles), *J. Chem. Soc. Perkin Trans. 2* 1978, 719; *ibid.* **1979,** 224; *J. Am. Chem. Soc.* **100** (1978) 8244; *J. Chem. Soc. Perkin Trans 2* **1979,** 673 – 679; *Tetrahedron Lett.* **1978,** 3263; *J. Chem. Soc. Perkin Trans 2* **1979,** 844 – 850; *ibid.* **1979,** 851 – 852; *ibid.*

1980, 13 – 38; *ibid.* **1980,** 358 – 362; *ibid.* **1980,** 358 – 62; *ibid.* **1980,** 904 – 914; *Can. J. Chem.* **71** (1993) 210 – 215.

Specific References

13. V. Vesely, T. Stojanova, *Collect. Czech. Chem. Commun.* **9** (1937) 465.
14. G. Bourjol, *Chem. Ind.* **78** (1957) 214.
15. Texaco, US 3311670, 1963.
16. California Res. Corp., US 2511711, 1946.
17. Rütgerswerke, DE 1468913, 1965.
18. California Res. Corp., US 2541959, 1947.
19. Hoechst, DE 950464, 1954.
20. I. T. Rodeanu, RO 60761, 1972.
21. Y. and G. Erykalov, V. G. Chirtulov, A. A. Spryskov, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **14** (1971) no. 1, 79 – 82.
22. Mitsui Toatsu Chem., JP 53044528, 1976.
23. V. D. Simonov, M. A. Ikrina, N. F. Popova, L. A. Kozlova, *Zh. Prikl. Khim. (Leningrad)* **45** (1972) no. 12, 2765 – 2766.
24. Nitto Chem. Ind., JP-KK 20541/65, 1962.
25. Hoechst, DE 1063151, 1957.
26. L. Field, P. H. Settlage, *J. Am. Chem. Soc.* **76** (1954) 1222.
27. C. M. Gable, H. F. Betz, S. H. Maron, *J. Am. Chem. Soc.* **72** (1950) 1445 – 1448.
28. Allied Chem. & Dye Corp., US 2697117, 1951.
29. J. A. Bradley, H. H. Harkins, *Abstr. Pap. Am. Chem. Soc.* **127th,** 1955.
30. E. E. Gilbert, *Chem. Rev.* **62** (1962) no. 6, 549.
31. L. Leiserson et al., *Ind. Eng. Chem.* **40** (1948) 508.
32. Monsanto, GB 679827, 1949, US-Prior. 1948.
33. Matsuyama Sekiyuka, JP 52133945, 1976.
34. *Chem. Eng.* **69** (1962) no. 5, 70 – 72.
35. V. Gold, D. P. N. Satchell, *J. Chem. Soc.* 1956, 1635.
36. M. Kilpatrick et al., *J. Phys. Chem.* **64** (1960) 1433.
37. J. C. D. Braml, *J. Chem. Soc.* 1950, 1004, 1952, 3922, 3927.
38. J. E. Woodbridge, *J. Am. Oil Chem. Soc.* **35** (1958) 528.
39. Vulcan Chem., GB 747659, 1952.
40. A. W. Kaandorp, H. Cerfontain, F. L. J. Soxma, *Recl. Trav. Chim. Pays-Bas* **81** (1962) 969 – 992.
41. BASF, EP 41134, 1981 (F. Brunnmueller, W. Boehm, V. Weberndoerfer).
42. BASF, US 5 892 105, 1997.
43. Hoechst, DE 1493663, 1964.
44. Z. Skrowaczewska, *Pr. Wroclaw, Tow. Nauk. Ser. B.* **61** (1953) 5; *Chem. Abstr.* **48** (1954) 7568.
45. Bayer, EP 63271, 1981 (H. Emde, H. U. Blank, P. Schnegg).
46. Sumitomo, EP 770 599, 1995.
47. H. Meerwein et al., *Ber. Dtsch. Chem. Ges.* **90** (1957) 846.
48. H. Meerwein et al., *Ber. Dtsch. Chem. Ges.* **90** (1957) 851.
49. Z. J. Allan, J. Podstata, *Collect. Czech. Chem. Commun.* **31** (1966) 3573.
50. W. H. Hunter, M. M. Sprung, *J. Am. Chem. Soc.* **53** (1931) 1432.
51. A. Amer et al., *J. Chromatogr. 3* **62** (1986) 413 – 418.
52. E. Toinlinson, T. M. Jefferies, C. M. Riley, *J. Chromatogr.* **159** (1978) 315 – 358.
53. V. V. Kharitonov, A. A. Spryskov, V. P. Leshchev, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **14** (1971) no. 2, 238.
54. J. Borecky, *J. Chromatogr.* **2** (1959) 612.
55. M. J. Stasik *Arch. Toxicol.* **33** (1975) 123 – 127.
56. D. H. Blakey et al., *Toxicology and Industrial Health* **8** (1992) 369 – 376.
57. R. Zahn, H. Wellens, *Z. Wasser Abwasser Forsch.* **13** (1980) 1 – 7.
58. P. Gerike, W. K. Fischer, *Ecotoxicol. Environ. Saf.* **5** (1981) 45 – 55.
59. Hoechst, EP 403 947, 1989.
60. L. I. Levina, S. N. Patrakova, L. A. Patrusev, *Zh. Obshch. Khim.* **28** (1958) no. 9, 2427.
61. Hoechst, DE 3 302 647, 1983.
62. Bayer, EP 1276, 1978 (H. U. Blank, T. Pfister).
63. H. Meerwein et al., *Ber. Dtsch. Chem. Ges.* **90** (1957) 841.
64. Bayer, DE 859461, 1942.
65. Zeneca, WO 98/40 332, 1997.
66. Occidental Chemical Cooperation, US 5 710 355, 1996.
67. C. R. Gambill, T. D. Roberts, H. Shechter, *J. Chem. Educ.* **49** (1972) no. 4, 287.
68. O. Hinsberg, J. Kessler, *Ber. Dtsch. Chem. Ges.* **38** (1905) 906.
69. F. Runge et al., *Fresenius Z. Anal. Chem.* **158** (1957) 266.
70. F. Runge et al., *Ber. Dtsch. Chem. Ges.* **88** (1955) 533.
71. M. L. Crossley et al., *J. Am. Chem. Soc.* **60** (1938) 2222.
72. VEB Fahlberg-List, DD 9132, 1942.
73. Bayer, DE 694974, 1955.
74. E. Schlittler, G. de Stevens, L. Werner, *Angew. Chem.* **74** (1962) no. 9, 317 – 326.
75. J. A. Bogan, *Med. Actual.* 1977, 393 – 395.
76. Chem. Fabrik Griesheim, DE 840240, 1943.

77. H. R. Slagh, E. C. Britton, *J. Am. Chem. Soc.* **72** (1950) 2808.
78. Esso, DE 1227009, 1962.
79. H. Kohner, M. L. Gressmann, *Z. Phys. Chem. Abt. A* **144** (1929) 144.
80. R. C. Crooks, R. R. White, *Chem. Eng. Progr.* **46** (1950) 249.
81. BIOS Final Report no. 664, 8; no. 1149, 128.
82. *Ullmann,* 3rd ed., **vol. 4,** p. 305.
83. R. L. Kenyon, N. Boehmer, *Ind. Eng. Chem.* **42** (1950) 1446.
84. A. N. Planowski, S. Z. Kagan, *Chem. Zentralbl.* 1941, 1075.
85. G. F. Lisk, *Ind. Eng. Chem.* **40** (1948) 1678.
86. A. Guyot, *Chim. Ind. Paris* **2** (1919) 879 – 891, 1167.
87. D. H. Killeffer, *Ind. Eng. Chem.* **16** (1924) 1066.
88. E. Sobczak, M. Badzynski, S. Chojnocky, K. Blanowicz, *Pr. Wydz. Nauk Tech., Bydgoskie Tow. Nauk., Ser. A* **13** (1979) 109 – 112; *Chem. Abstr.* **91,** 211020 y.
89. H. Soep, P. Demoen, *Microchem. J.* **4** (1960) 82 – 87.
90. BIOS 986, I, 56.
91. BIOS 986, II, 415.
92. FIAT 1313, I, 101.
93. BIOS 986, II, 414.
94. BIOS 1153, 150.
95. N. N. Dykhanov, *J. Gen. Chem. (USSR Engl. Transl.)* **29** (1959) 3563.
96. Harshaw Chem., US 2757133, 1952.
97. A. A. Spryskov, S. P. Starkov, *Zh. Obshch. Khim.* **27** (1957) 2780.
98. BIOS 986, II, 385.
99. Ube Ind., JP 71/41529, 1971.
100. Koppers Co. Inc., US 4302403, 1976.
101. Taoka Chemical, JP-KK 80034776, 1969.
102. Hooker Chem., DE 1468040, 1964.
103. Koppers, BE 698821, 1967.
104. H. Cerfontain, F. L. J. Sixma, L. Vollbracht, *Recl. Trav. Chim. Pays-Bas* **82** (1963) no. 7, 659 – 670.
105. Tennessee Corp., US 2841612, 1956.
106. V. S. Patwardhan, R. E. Eckert, *Ind. Eng. Chem. Process Des. Dev.* **20** (1981) no. 1, 82 – 85.
107. A. A. Spryskov, V. A. Kozlov, *Khim. Khim. Tekhnol. Alma Ata 1962* **12** (1969) no. 2, 166 – 169 (1969).
108. A. A. Spryskov, V. A. Kozlov, *Khim. Khim. Tekhnol. Alma Ata 1962* **12** (1969) no. 7, 900 – 902 (1969).
109. Koppers, BE 844860, 1975 (US Appl. 602768, 1975).
110. A. A. Spryskov, *Khim. Khim. Tekhnol.* **4** (1961) no. 6, 981 – 984.
111. T. I. Potapova, A. A. Spryskov, *Khim. Khim. Tekhnol.* **10** (1967) no. 8, 885 – 887.
112. Witco Chem., DE 1812708, 1968.
113. Taoka Dye MFG., JP 68/22861, 1964.
114. Allied Chem. & Dye Corp., US 2362612, 1942.
115. Hoechst, DE 2 837 549, 1987.
116. Standard Oil Develop. Co., US 2540519, 1946.
117. *Org. Synth.,* **Coll. Vol. III** (1955) 366. G. Storck, R. Borch, *J. Am. Chem. Soc.* **86** (1964) 937.
118. M. J. Morgan, L. H. Cretcher, *J. Am. Chem. Soc.* **70** (1948) 375.
119. Phillips Petroleum Co., US 2678947, 1949.
120. BASF, DE-AS 1157629, 1962.
121. F. H. Gereim, W. Braker, *J. Am. Chem. Soc.* **66** (1944) 1459.
122. W. Herzog, *Angew. Chem.* **39** (1926) 728.
123. Sankyo Kasei Co., JP 67/13937, 1965.
124. H. Meyer, *Justus Liebigs Ann. Chem.* **433** (1923) 333.
125. Sugai Chemical, JP-KK 52136144, 1976.
126. Dow, US 2821549, 1954.
127. Toyo Soda MFG, JP-KK 55031059, 1973.
128. Toyo Soda MFG, JP-KK 55064565, 1978.
129. Du Pont, US 2837500, 1953.
130. G. E. Inskeep, *J. Am. Chem. Soc.* **69** (1947) 2237.
131. R. H. Wiley, *J. Am. Chem. Soc.* **76** (1954) 720.
132. R. H. Wiley, S. F. Reed, Jr., *J. Am. Chem. Soc.* **78** (1956) 2171.
133. Du Pont, US 2822385/6, 1955.
134. Dow, US 2913438, 1955.
135. R. H. Wiley, C. C. Ketterer, *J. Am. Chem. Soc.* **75** (1953) 4520.
136. Mitsubishi Petro-Chem., JP 70/36498, 1967.
137. W. Borsche, M. Pommer, *Ber. Dtsch. Chem. Ges.* **54** (1921) 105.
138. Chem. Fabrik von Heyden, DE-AS 1159937, 1960.
139. P. Cagniant, *Bull. Soc. Chim. Fr.* 1950, 29.
140. Agency of Ind. Sci. Techn., JP 54100343, 1978.
141. Ciba-Geigy, DE-OS 2209223, 1971; DE-OS 2241304, 1972.
142. T. P. Sycheva, M. N. Shchukina, *Sb. Statei Obshch. Khim.* 1953, 527.
143. J. N. Tay, M. L. Dey, *J. Chem. Soc.* **117** (1920) 1407.
144. Y. and E. Briskin et al., SU 192192, 1966.
145. FIAT 949, 23.
146. Weiler – Ter Meer, DE 385049, 1921; *Friedländer* **14** (1926) 386.

147. Union Carbide, DE-OS 2704972, 1977 (U. A. Steiner; US Appl. 657209, 1976).
148. F. Runge, F. Pfeiffer, *Ber. Dtsch. Chem. Ges.* **90** (1957) 1757.
149. J. Stewart, *J. Chem. Soc.* **121** (1922) 2557.
150. S. A. El-Gerf, *Textile Res. J.* **67** (1997) 13 – 17.
151. G. M. Kraay, *Recl. Trav. Chim. Pays-Bas* **49** (1930) 1083.
152. Phillips, FR 1119189, 1954.
153. W. V. Farrar, *J. Chem. Soc.* 1960, 3063.
154. SU 166678, 1962 (E. V. Sergeev, V. I. Zetkin).
155. C. F. H. Allen, G. F. Frame, *J. Org. Chem.* **7** (1942) 15.
156. Sugai Chem. Ind., JP 70/19892, 1967.
157. M. Saito, M. Dehara, O. Manabe, *Nippon Kagaku Kaishi* 1972, 2, 380 – 382.
158. Imperial Chemical Industries, EP 25274, 1979.
159. H. E. Fierz et al., *Helv. Chim. Acta* **12** (1929) 665.
160. E. Wertheim, *Org. Synth.* **15** (1935) 55 – 58.
161. H. E. Fierz et al., *Helv. Chim. Acta* **12** (1929) 663.
162. Amer. Cyanamid Co., US 3038932, 1958.
163. Bayer, DE-OS 3224155, 1982 (M. Michna, H. Henk).
164. Eastmann Kodak, US 2529924, 1948.
165. J. H. Freemann, E. C. Wagner, *J. Org. Chem.* **16** (1951) 815.
166. Chelevin RN, SU 740758, 1978.
167. FIAT Final Report no. 1313, **vol. 1** (1948) 187 – 191.
168. H. E. Fierz-David: *Grundlegende Operationen der Farbenchemie,* Springer Verlag, Wien 1952, p. 117.
169. BASF, FR 1336648, 1962.
170. Bayer, US 3532692, 1967.
171. BASF, DE 1115384, 1957.
172. Bayer, DE 1089548, 1959.
173. FIAT Final Report no. 1313, **vol. 1** (1948) 255 – 256.
174. BASF, DE-OS 2049639, 1970.
175. Hoechst, DE-OS 2439297, 1974.
176. J. Casper, W. Petzold, PB Rept. 73911, FIAT Microfilm no. 87, 4648-60.
177. SU 667550,1977 (B. I. Kissin, A. V. Tarasevich).
178. Bayer, EP 63271, 1981 (H. Emde, H. U. Blank, P. Schnegg).
179. BIOS 1149, 125.
180. *Org. Synth* ., Coll. Vol. 1 (1932) 8.
181. Hoechst, DE 19 741 455, 1997.
182. O. Lustig, E. Katscher, *Monatsh. Chem.* **48** (1927) 94.
183. I.G. Farbenindustrie, DE 743766, 1940.
184. Bayer, DE 899536, 1951.
185. BIOS 1153, 171.
186. FIAT 1313, I, 54.
187. P. K. Maarsen, R. Bregman, H. Cerfontain, *J. Chem. Soc. Perkin Trans 2* 1977, 14, 1863 – 1868.
188. H. E. Fierz-David: *Grundlegende Operationen der Farbenchemie,* Springer Verlag, Wien 1952, 102.
189. Imperial Chemical Industries, DE 2921309, 1978.
190. Imperial Chemical Industries, DE 2045086, 1969.
191. Geigy, US 3589921, 1969.
192. Sumitomo Chemical, JP-KK 57048961, 1980.
193. Imperial Chemical Industries, GB 1205017, 1966.
194. Imperial Chemical Industries, JP 4706532, 1970.
195. Ciba-Geigy, CH 523865, 1965.
196. Hoechst, DE 2162963, 1971.
197. J. Obermiller, *Ber. Dtsch. Chem. Ges.* **40** (1907) 3637.
198. F. Willson, K. H. Meyer, *Ber. Dtsch. Chem. Ges.* **47** (1914) 3162.
199. Bayer, DE-OS 1921046, 1969.
200. M. Hazard-Flammand, DE 141751, 1897; *Friedländer* **6,** 1295.
201. E. N. Anischuk, SU 678109, 1977.
202. FMC Corp., DE 2064944, 1971.
203. C. Stopanski, DE 3004543, 1979.
204. Bayer, DE 2934980, 1979.
205. K. A. Hofmann, E. Biesalski, *Ber. Dtsch. Chem. Ges.* **45** (1912) 1396.
206. V. I. Burmiskov et al., SU 213892, 1969.
207. J. Pollak, *Monatsh. Chem.* **46** (1925) 395.
208. A. Quilico, *Gazz. Chim. Ital.* **57** (1927) 799.
209. Bakelite, FR 1012550, 1949.
210. Ube Ind., JP 7021501, 1966.
211. Eastman Kodak, US 3772379, 1971.
212. *Römpp,* 7th ed., p. 3604.
213. V. Fischl, H. Schlossberger: *Handbuch der Chemotherapie,* Thieme Verlag, Leipzig 1934, p. 583.
214. Eastman Kodak, DE 1302776, 1971.
215. FIAT 1313, I, 124.
216. Dow, US 4131619, 1978.
217. Bayer, BE 758231, 1969.
218. Ishihara Sangyo Kaisha, JP 66/10810, 1962; GB 1017976, 1963.
219. Shell, BE 625993, 1962; GB-Prior. 1961.
220. BIOS 1153, 176.
221. FIAT 1313, I, 51.
222. D. W. Hein, E. S. Pierce, *J. Am. Chem. Soc.* **76** (1954) 2729.
223. BIOS 1153, 177.

224. *Org. Synth* **27** (1947) 88 – 91
225. BIOS 1153, 173; 986, 441.
226. BIOS 986, 399.
227. Ciba, US 2353351, 1943.
228. W. F. Harris, T. R. Sweet, *J. Am. Chem. Soc.* **77** (1955) 2893.
229. FIAT 1313, I, 184.
230. BIOS 1153, II, 255.
231. SU 251911, 1967 (L. N. Lapin, I. V. Reis).
232. SU 318595, 1969 (Z. I. Mironova, L. N. Smirnov, A. G. Temnova).
233. Kaiser Aluminium and Chem. Corp., DE 1446002, 1959.
234. Sumitomo Chem., DE 2243178, 1971.
235. G. Scholze, DD 96261, 1972.
236. P. B. Hutson, US 3534077, 1968.
237. Teikoku Hormone MFG., JP 70/24372, 1966.
238. FIAT 1313, I, 77.
239. BIOS 1153, 255.
240. BIOS 1153, 223.
241. Du Pont de Nemours, US 45 56 733, 1984.
242. BIOS 1153, 214.
243. BIOS 1153, 217.
244. BIOS 1153, 179.
245. BIOS 1149, 24.
246. BIOS 1153, 113.
247. FIAT 1313, I, 18.
248. BIOS 1153, 114, 117.
249. BIOS 1153, 234.
250. Hoffmann-La Roche, DE 98 839, 1898.
251. Merck, DE 1153761, 1957.
252. BIOS 1153, 222.
253. BIOS 1153, 182.
254. A/S Ferrosan, DE 1157628, 1960.
255. BIOS 1153, 185.
256. W. Logemann et al., *Justus Liebigs Ann. Chem.* **623** (1959) 162.
257. F. C. Novello et al., *J. Org. Chem.* **25** (1960) 965.
258. Merck, DE 1088504, 1958; DE 1153761, 1957.
259. Merck, DE 1161910, 1958.
260. F. C. Novello, J. M. Sprague, *J. Am. Chem. Soc.* **79** (1957) 2028.
261. Merck, US 2809194, 1956; DE 1119282, 1957.
262. Ciba-Geigy, EP 39306, 1980.
263. BIOS 986, 130 – 134.
264. Hodogaya Chem. Ind., JP-KK 4151945, 1978.
265. BIOS 986, 129 – 130.
266. BIOS 986, 128.
267. Dainichisaika Color Chem., JP 4125636, 1978.
268. BIOS 1153, 188.
269. Du Pont, US 2744027, 1952.
270. Amer. Cyanamid Co., US 2754294, 1953.

Benzidine and Benzidine Derivatives

HANS SCHWENECKE, Hoechst Aktiengesellschaft, Frankfurt / Main, Federal Republic of Germany (Chaps. 1 – 7)

DIETER MAYER, Hoechst Aktiengesellschaft, Frankfurt / Main, Federal Republic of Germany (Chap. 8)

1. Introduction 1
2. Production 1
2.1. Reduction 1
2.2. Benzidine Rearrangement 4
2.3. Isolation 4
3. Storage and Transportation 5
4. Occupational Health 5
4.1. General Safety Precautions 5
4.2. Special Safety Regulations 5
5. Environmental Protection 7
6. Characteristics of the Diphenyl Bases 7
6.1. Benzidines 7
6.2. Tolidines 9
6.3. Halogenated Benzidines 10
6.4. Alkoxybenzidines 11
6.5. Benzidinecarboxylic Acids 12
6.6. Benzidinesulfonic Acids and Benzidinesulfones 12
7. Economic Aspects 13
8. Toxicology 13
8.1. Benzidine 13
8.2. 3,3′-Dichlorobenzidine 14
8.3. *o*-Tolidine (3,3′-Dimethylbenzidine) . 15
8.4. *o*-Dianisidine (3,3′-Dimethoxybenzidine) 15
9. References 15

1. Introduction

Benzidine [*92-87-5*], 4,4′-biphenyl-diyldiamine, and its substitution products represent the group called the diphenyl bases. They are used mainly as intermediates in the production of azo dyes and azo pigments. Symmetrically or asymmetrically coupled products can be produced by simultaneous or successive diazotization (coupling), respectively.

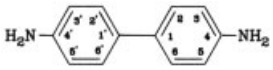

The diphenyl bases Benzidine have aroused interest as cross-linking agents, e.g., in polyurethane plastics [10], in which they can noticeably increase the temperature stability [11]. The diphenyl radical has a chain-stiffening effect in polyamides [12]. The ability of the diphenyl bases to react with numerous cations, anions, and organic substances, such as oxidizing agents and blood, is utilized for analytical and diagnostic purposes.

The economic importance of the diphenyl bases is apparent from the fact that world production in 1983 was about 15 kt. The major producing countries were the United States, Japan, and the Federal Republic of Germany.

The most important of the diphenyl bases are *o*-tolidine, 3,3′-dichlorobenzidine, and *o*-dianisidine. The physical and chemical properties of benzidine and its derivatives are treated in detail in Chapter 6. The following sections describe the general methods for the production and handling of benzidine and benzidine derivatives.

2. Production

Benzidine and the other diphenyl bases are produced in three separate processing stages:

1) Reduction of nitro groups to form hydrazo compounds
2) Benzidine rearrangement
3) Isolation of the bases

2.1. Reduction

The alkaline reduction of nitrobenzene to hydrazobenzene can be represented by the simple equation

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_539

$$2\,C_6H_5NO_2 + 10\,H \longrightarrow C_6H_5NH\text{–}NHC_6H_5 + 4\,H_2O$$

It passes, according to the Haber process, through several stages, including the intermediate condensation of a nitroso with a hydroxylamine compound to form the azoxy compound.

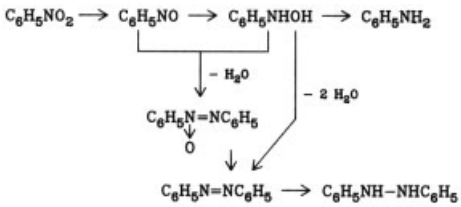

This method always forms symmetrical hydrazo compounds.

The main byproduct is the monocyclic primary amine (e.g., aniline) corresponding to the original nitro compound. This is formed not only by further reduction of the phenylhydroxyamine, but also by disproportionation of the unstable hydrazo compound.

Various reduction methods are used in industry.

Reduction with Zinc Dust. The reduction of aromatic nitro compounds with zinc dust in alkaline medium takes place according to the equation:

$$2\,C_6H_5NO_2 + 5\,Zn + 10\,NaOH \longrightarrow C_6H_5NH\text{–}NHC_5H_5 + 5\,Na_2ZnO_2 + 4\,H_2O$$

In industry, the nitro compound is reduced in batches of 500 kg or more. The nitro compound is dissolved in a high-boiling solvent, e.g., *o*-dichlorobenzene or solvent naphtha, and alcohols (methanol, ethanol) are often added to keep the resultant hydrazo compound in solution. Zinc dust is then suspended in this solution and sodium hydroxide solution is emulsified with it. A quantity of zinc dust as much as 50 % above the theoretical requirement is usually needed. The necessary alkali is introduced in the form of a concentrated (40 – 50 %) sodium hydroxide solution. Only a small percentage (ca. 2 – 10 %) of the theoretical quantity of sodium hydroxide is required because the sodium zincate quickly regenerates alkali by hydrolysis.

$$Na_2ZnO_2 + H_2O \longrightarrow 2\,NaOH + ZnO$$

The reduction is highly exothermic. It can be regulated by controlling the addition of one or more of reactants. The optimum reduction temperature is between 80 and 120 °C, depending on the aromatic compound. Higher temperatures can be used with the unsubstituted nitrobenzene without causing overreduction than are possible with *o*-alkyl and *o*-alkoxy nitrobenzenes. In many cases the addition of small amounts of zinc dust and water is necessary to complete reduction and convert zinc oxide into zinc hydroxide.

The reduction is terminated when the red color of the azo stage has disappeared. The hydrazobenzene solution is then separated, either directly or after dilution with hot water, from the zinc hydroxide, which may still contain unconverted zinc dust. The zinc hydroxide is extracted with fresh solvent to increase the yield. The hydrazobenzene solutions are usually subjected to a subsequent benzidine rearrangement without intermediate isolation of the hydrazo compound. The zinc hydroxide is freed of solvent in a vacuum or with steam and either further processed into zinc salts or dried and returned to the zinc works [13], [14]. The reduction process using zinc and alkali is technically the most important and can be used for all hydrazo compounds.

However, because of the high price of zinc dust there have been numerous attempts to replace it with cheaper reducing agents.

Reduction with Iron. Reduction with iron and sodium hydroxide solution, only of minor importance in preparative chemistry, has been investigated extensively for commercial production ever since 1900.

$$2\,C_6H_5NO_2 + 4\,Fe + 6\,H_2O \longrightarrow C_6H_5NH\text{–}NHC_6H_5 + 2\,Fe(OH)_2 + 2\,Fe(OH)_3$$

Finely ground cast iron turnings, iron powder obtained by decomposition of iron carbonyl, carbide-containing iron, or ferrosilicon are used. The iron turnings are first slightly corroded with 60 % sodium hydroxide solution. To prevent overreduction the concentration of sodium hydroxide is maintained at a high level and sometimes even increased during reduction. Reduction is carried out using nitrobenzene, iron, and

sodium hydroxide in the molar ratio 1 : 2 : 4 at a temperature of 100 – 120 °C.

The process can be used for the reduction of nitrobenzene to hydrazobenzene but less often for its substitution products, especially because it is difficult to separate the hydrazo compound from the ferruginous mud. A yield of 88 – 92 % has been reported for the reduction of nitrobenzene to hydrazobenzene using ferrosilicon (15 % Si) [14].

Reduction with Sodium Amalgam. Sodium amalgam at 70 – 100 °C in an upright, cylindrical nickel reactor is brought into contact with nitrobenzene emulsified in water or sodium hydroxide. The process is carried out in batches to produce azobenzene. The reduction conditions can be varied by adding solvent (solvent naphtha), changing the concentration of the sodium hydroxide solution or the amalgam, or by adding decomposition catalysts (graphite, activated charcoal).

Nevertheless, in the further reduction from the azo to the hydrazo stage, some overreduction to the monocyclic amine cannot be avoided. The product formed contains up to 90 % azobenzene in addition to hydrazobenzene, aniline, and nitrobenzene. It is therefore isolated and further reduced to the hydrazo stage with zinc dust and sodium hydroxide solution. The amalgam reduction is used industrially only with unsubstituted nitrobenzene.

Electrolytic Reduction. The electrochemical reduction of nitro to hydrazo compounds is also carried out on a commercial scale.

$$2\,C_6H_5NO_2 + 10\,H^+ + 10\,e^- \longrightarrow C_6H_5NH\text{–}NHC_6H_5 + 4\,H_2O$$

Electrolytic cells, with cathodes, diaphragms, and anodes positioned either side by side or above each other, are used.

The satisfactory functioning of electrolytic reduction depends largely on the condition of the surface of the cathode. Sheet iron cathodes with a lead sponge are used. Overreduction is less likely to occur with a lead sponge than with shiny lead surfaces. Even lead cathodes with spongy zinc have been used, but the spongy zinc was partially used up by chemical reduction. Nickel cathodes have also been described [14], [15].

The catholyte consists of a vigorously stirred emulsion of the nitro compound in 8 – 10 % sodium hydroxide solution. Ethanol and toluene can be used as diluents. Depending on the geometry of the cell, solvents with high specific gravities (chlorinated benzene, chlorinated toluene, or chlorinated naphthalene) are used to keep the cathode liquid away from the anode compartment above it. Cement or special ceramic compounds are used to make the diaphragms. The anolyte, consisting of 10 % sodium hydroxide solution, contains the lead or iron anode, which has a PbO_2 coating or, for preparative purposes, may also be nickel-plated.

Electrochemical reduction is carried out at temperatures of 65 – 90 °C, current densities of about 20 A/dm^2 and voltages of 4 – 10 V. Electrochemical reduction is technically satisfactory and may also be cheaper than zinc dust reduction if it is interrupted when only about 50 % of the starting material has reached the hydrazo stage. After rearrangement and separation of the diphenyl base salts, the azo components and the solvent are recovered for the next batch. Nitrobenzene, 2-chloronitrobenzene, 2,5-dichloronitrobenzene, 2-nitrotoluene, and 2-nitroanisol can be electrochemically reduced to hydrazo compounds in yields of 70 – 90 % [14], [15].

Catalytic Reduction. Nitrobenzene can be reduced to hydrazobenzene with hydrogen in the presence of a palladium–carbon catalyst. Industrially, reduction is carried out in dilute alcohol (ethyl or isopropyl alcohol) in the presence of a base, such as 50 % sodium hydroxide solution. The reaction takes place in an autoclave at 80 – 85 °C and with a slightly elevated pressure. Overreduction is suppressed by the high concentration of alkali. The yield can also be influenced by buffering the reaction medium, e.g., by the addition of acetates, borates, or alkanolamines. This reduction method also applies to 2-nitrotoluene and 2-nitroanisol.

For preparative purposes, Raney nickel can be used at elevated temperatures and with mixed catalysts. Instead of hydrogen, hydrazine hydrate or sodium borohydride can be used with palladium – calcium carbonate, Raney nickel, ruthenium – carbon, or copper ions as the catalyst [14].

Other Reduction Methods. In preparative chemistry, numerous other reducing agents are used to produce hydrazo compounds. Lithium, sodium, potassium, magnesium, and aluminum are used for reduction, either as metals or as alcoholates. Silicon, lithium, and aluminum are used as amalgams. Of the nonmetal reducing agents, special mention should be made of hydrogen sulfide and its salts (sulfide reduction). Methanol, formaldehyde, and glucose, sometimes combined with quinonoid catalysts, are organic reducing agents.

2.2. Benzidine Rearrangement

Aromatic hydrazo compounds, formed by the alkaline reduction of aromatic nitro compounds, are easily rearranged by mineral acids into diaminodiphenyl compounds (**1**) – (**3**) and aminodiphenylamine compounds (**4**), (**5**). This is the second stage in the production of diphenyl bases.

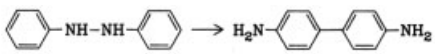

Hydrazobenzene Benzidine (1)

Traces of *o*-benzidine (2,2′-diaminodiphenyl) (**2**) are formed as a byproduct by ortho rearrangement, and diphenyline (2,4′-diaminodiphenyl) (**3**) is also formed in appreciable quantities by ortho-para rearrangement.

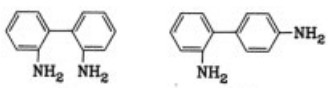

2 3

Partial rearrangements yield *o*-semidine (2-aminodiphenylamine) (**4**) and *p*-semidine (4-aminodiphenylamine) (**5**).

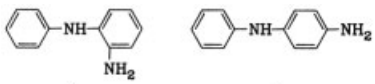

4 5

The type and quantity of the rearrangement products are largely dependent on the chemical constitution of the reactant and can only be influenced to a limited extent by the reaction conditions. The byproducts, mainly diphenyline (**3**), are formed in quantities of up to 15 % and have no commercial value.

The benzidine rearrangement is a true intramolecular reaction because no mixed benzidines are formed from mixtures of different hydrazo compounds and because only the corresponding unsymmetrical benzidine is obtained from unsymmetrically substituted hydrazobenzenes [16].

The industrial benzidine rearrangement usually starts with the hot solution that is obtained from the reduction of the nitro compound to the hydrazo compound. Intermediate isolation of the hydrazo compound is not necessary, but it is advisable in some cases, such as amalgam reductions. Because hydrazo compounds disproportionate to form azo compounds and the corresponding monocyclic amines at higher temperatures, some reducing agent, such as sodium hydrosulfite (sodium dithionite) or zinc dust, can be added to reduce the azo content of the hydrazo compound.

$$2\,C_6H_5NH\text{–}NHC_6H_5 \longrightarrow C_6H_5N{=}NC_6H_5 + 2\,C_6H_5NH_2$$

The hydrazo solution is mixed with a mineral acid in an acid-resistant mixer. Depending on the substituents, suitable mineral acids are 10 – 30 % hydrochloric acid, 20 – 80 % sulfuric acid, or a mixture of these two. The optimum rearrangement temperature varies from 0 – 5 °C for dianisidine to 100 °C for the less heat-sensitive diphenyl bases. Temperatures higher than necessary diminish the yield, which should be 70 – 95 %.

2.3. Isolation

The most important stage of the benzidine rearrangement is the formation of a hydrochloric or sulfuric acid salt of the diphenyl base. This is either isolated directly, e.g., by salting out with sodium chloride or sodium sulfate, or first converted into the free base. For this a dilute alkali, such as sodium hydroxide solution or ammonia solution, should be used. The undesirable byproducts, especially the monocyclic amine and diphenyline, can be separated out relatively easily because of the greater solubility of these bases and their salts. Finally, only the azo compound remains in the inert solvent after acid extraction. It is returned to the reduction process along with the solvent.

3. Storage and Transportation

The diphenyl bases and their salts are usually introduced onto the market in the form of moist solids to prevent dust from forming when they are handled. They are packed in corrugated-iron drums with plastic liners; in addition, plastic-lined fiber drums, which can be burnt after they have been emptied and rinsed out, are often used as disposable packs. Wooden drums are avoided. The packs should be kept closed if possible so that dust does not form as a result of drying out of the diphenyl bases.

Benzidine is often isolated by collecting it on suction filters as a slurry and then transporting it in this form. The earlier procedure of drying, distilling, and grinding diphenyl bases is no longer used.

If the diphenyl bases are to be further processed by the producer in an integrated system, they are transported through pipe systems as aqueous suspensions.

4. Occupational Health

4.1. General Safety Precautions

Benzidine has properties which are dangerous to health and the health effects of the substituted diphenyl bases are unknown. The diphenyl bases should therefore be produced and processed in special, separated areas. Walls and floors must not be capable of absorbing chemicals. The floors should have adequate drainage and wood should not be used either as a building material or for equipment. Good ventilation must be ensured and provisions must be made for air circulation. Ventilation of the workrooms must not be restricted. The exhaust air must be effectively cleansed using filters or exhaust-gas scrubbers. Wherever possible, equipment should operate as a closed system. Filter presses and open suction filters must be avoided. A vacuum in the system is a precaution against leakage or the uncontrolled escape of vapors. When evacuated machines are opened, attention must be paid to the formation of sublimates. Tasks that involve an increased risk of human contact include sampling during manufacturing and processing, isolation, packing, and filling. If possible, the work is carried out with automated machines or in separately ventilated areas. Before undertaking repairs, special care must be taken to ensure that the equipment has been decontaminated.

Scalding out should only be allowed if the vapors can be carefully condensed and do not enter the atmosphere. If necessary and if permitted by the construction material, purification can be carried out by the addition of chemicals. Reaction with hypochlorite or even diazotization is suitable for the chemical destruction of amines. When using dilute hypochlorite solution, care must be taken to ensure that the solution is always kept considerably basic by the appropriate addition of alkali. This avoids formation of explosive nitrogen trichloride. Access to contaminated containers should only be permitted if complete protective clothing is worn and breathing apparatus is used. Diphenyl bases penetrate the surface of protective articles made of rubber or plastic. The workplaces must be monitored for emissions of benzidine and its derivatives [17].

4.2. Special Safety Regulations

In the United States and most other industrialized countries, rules and regulations have been issued for benzidine, which is regarded as carcinogenic in humans, and for 3,3′-dichlorobenzidine, which has proved to be carcinogenic in animals [18]. The aim is to protect the worker from contamination. Mixtures containing 0.1 wt% or more of benzidine are subjected to detailed handling regulations. The production of benzidine and its salts is prohibited in Japan and has been stopped in the Soviet Union.

Nowadays the most commonly produced and used diphenyl base is 3,3′-dichlorobenzidine (as the free base or one of its salts). The appropriate U.S. regulations are therefore presented. They apply to substances that contain at least 1 wt% 3,3′-dichlorobenzidine or its salts [19].

1) Areas in which these substances are produced, processed, used, repackaged, released, handled, or stored are designated *controlled areas.*
2) The number of people entering and leaving these areas must be restricted and monitored. In these areas a distinction is made between
   a) isolated systems, e.g., glove boxes

b) closed systems, e.g., sealed containers or piping systems
Special permission must be obtained for access to closed systems; employees shall be required to wash themselves upon each exit from the regulated areas.
c) open vessel system operations are prohibited.

3) Upon completion of their work in isolated or closed systems, employees must observe certain minimum regulations in terms of personal hygiene.
4) When opening a system employees shall be required to wear clean, full-body protective clothing, shoe covers, and gloves. Employees engaged in 3,3′-dichlorobenzidine (or its salts) handling operations shall be required to use a half-face filter-type respirator for dusts, mists, and fumes.
5) For the cleanup of leaks or spills and maintenance or repair operations on contaminated systems, impervious garments, gloves, boots, and a hood with a continuous supply of air should be worn.
6) Prior to each exit from a regulated area, employees shall be required to remove and leave behind protective clothing and to place used clothing and equipment in impervious containers for purposes of decontamination or disposal.
7) Employees shall be required to wash on each exit from the regulated area and to shower after the last exit of the day.
8) A daily roster of employees entering regulated areas shall be established and maintained. Before taking up employment they must be informed of the carcinogenic hazard of 3,3′-dichlorobenzidine, of the possibilities of contact with the product in their work, and of the medical surveillance program. They are subjected to an annual medical examination.
9) Signs bearing the following inscription should be posted at the entry points of regulated areas:
CANCER-SUSPECT AGENT
AUTHORIZED PERSONNEL ONLY
10) More detailed warnings apply to contaminated areas. Additional signs should inform employees of the procedures that must be followed upon entering or leaving a regulated area.
11) The storage, consumption, and use of foodstuffs, drinks, tobacco, cosmetics, and the like are prohibited in regulated areas.
12) Decontamination procedures shall be established. Dry sweeping and dry mopping are prohibited.
13) The contents of the containers must be identified by suitable labels. The words *Cancer-Suspect Agent* must be added, the minimum size print being stipulated.
14) Any incidents must be reported to the relevant authorities.
15) Deposits on machinery or parts of the building are detected by a swipe test. Sodium hypochlorite solution must be used for decontamination.

In Great Britain the use (for research purposes) of *Prohibited Substances*, such as benzidine and its salts, is only possible if special permission is obtained for work in completely closed systems or if the benzidine, in the form of a hydrochloride, is always kept moist with at least one part water to two parts benzidine hydrochloride.

Dichlorobenzidine, *o*-tolidine, *o*-dianisidine, and their salts are designated *Controlled substances*, and employees should be protected from contamination by these substances. Workers are medically monitored and records kept [20].

In the Federal Republic of Germany the important diphenyl bases are included among the regulated dangerous working substances.

1) Benzidine and its salts are regarded as carcinogenic hazards in concentrations above 0.01 % by weight in the working material.
2) 3,3′-Dichlorobenzidine is regarded as a carcinogenic hazard in concentrations above 0.1 % by weight in the working material.
3) *o*-Tolidine is considered harmful to health.
4) *o*-Dianisidine and its salts are considered toxic.

When working with these substances, all the necessary measures relating to safety, industrial medicine, and occupational hygiene should be taken to protect employees from the effects of these substances and to keep their health under constant medical check. The air at the workplace should be examined to monitor the benzidine

and 3,3′-dichlorobenzidine contents. If benzidine or 3,3′-dichlorobenzidine is marketed, the package must bear the words *May cause cancer* [21].

Equipment in which benzidine or its salts may be present must be operated in such a way that breakdowns are prevented and the harmful effects are kept to a minimum. All the necessary precautions are summarized in [22].

In accordance with the technical regulations for dangerous working substances, the presence of an effect must always be assumed when handling benzidine. Here "effect" means the risk of damage to workers' health. If the concentration of 3,3′-dichlorobenzidine in the air is below 0.1 mg/m^3 at the workplace, it can be assumed to have no effect on the worker [23].

In Belgium [24], Japan [25], and numerous other countries there are similar regulations for handling benzidine and its derivatives [26]. Before starting a job in which diphenyl bases can be formed or are used, it is high advisable to obtain exact details of the local regulations.

5. Environmental Protection

The acute fish toxicity LC_{50} is about 5 – 50 ppm, depending on the type of fish and the diphenyl base. In terms of acute toxicity, diphenyl bases can be regarded as only slightly toxic to mammals. The oral LD_{50} for rats is 3820 mg/kg for 3,3′-dichlorobenzidine, 404 mg/kg for *o*-tolidine, and 1920 mg/kg for *o*-dianisidine.

Benzidine achieves only a slight, and 3,3′-dichlorobenzidine a moderate bioaccumulation in bluegill sunfish [27]. Thus benzidine is unlikely reach a significant level in the food chain [28].

Because of the very low vapor pressures of the diphenyl bases, emissions into the ambient air from production or processing works are only possible in the form of dust. Such emissions can certainly be prevented by filtering or by using wet separators. Diphenyl base concentrations in the ambient air are monitored using TLC or HPLC [29].

In aqueous solution 3,3′-dichlorobenzidine base is very quickly degraded by photolysis, with monochlorobenzidine and benzidine being formed as intermediates [30], [31]. 3,3′-Dichlorobenzidine, *o*-tolidine, and *o*-dianisidine can be slowly biodegraded. More than 90 % of these two compounds can be eliminated in biological waste water treatment plants. Elimination takes place mainly by adsorption on sediments followed by slow biodegradation. Only a portion can be extracted again. This part declines as the adsorption process continues, because the 3,3′-dichlorobenzidine reacts with the sediments [30].

3,3′-Dichlorobenzidine cannot be removed quantitatively from waste water by hypochlorite solution because higher chlorinated benzidines are formed. Tetrazotization and conversion into 3,3′-dichloro-4,4′-dihydroxybiphenyl are more suitable procedures; they reduce the 3,3′- dichlorobenzidine content to less than 0.1 mg/kg [32]. Other appropriate waste water purification processes are oxidation with peroxide, adsorption on suspended adsorption agents, precipitation with iron(III) hydroxide, and combinations of these methods.

For decontamination of workrooms, buildings, and machines, a hot solution of 5 wt% tetrapotassium pyrophosphate and 10 wt% sodium ethyl hexyl sulfate is sprayed on with a hydraulic jet cleaner in the ratio 1 : 10 with water under pressure, so that 90 – 99 % of the adherent 3,3′-dichlorobenzidine is removed [32].

Wastes, such as production residues, contaminated disposable protective equipment, and empty packaging, which may contain diphenyl bases, are best incinerated. Suitable incinerators must have a secondary incineration compartment in which a dwell time of at least 0.3 s and a minimum temperature of 900 °C are maintained [33].

6. Characteristics of the Diphenyl Bases

6.1. Benzidines

Benzidine [*92-87-5*], 4,4′-biphenyl-diyldiamine, $C_{12}H_{12}N_2$, M_r 184.24, *mp* (coarse rods) 128 °C, crystallizes out of the melt usually in a metastable modification with a melting point of 122 – 125 °C, *bp* 400 – 401 °C at 98.7 kPa.

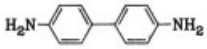

Benzidine crystallizes out of water above 80 °C in the anhydrous form; below 60 °C, it retains one molecule of water of crystallization; the monohydrate melts at 104 – 105 °C. Pure benzidine is colorless. One part by weight of benzidine dissolves in 2447 parts of water at 12 °C, in 106.5 parts of water at 100 °C, and in 45 parts of ether or 13 parts of absolute ethyl alcohol at 20 °C. Benzidine is diacidic; at 30 °C, $K_1 = 9.3 \times 10^{-10}$ and $K_2 = 5.6 \times 10^{-11}$. The heat of neutralization is 106.5 kJ/mol.

Benzidine slowly becomes discolored on exposure to air. It is chemically resistant to water. Blue, green, or red colorations and precipitates, many of which are based on quinonoid structures, form depending on the nature of the oxidizing agent and the reaction conditions used. These color reactions can be used to detect numerous oxidizing agents.

When chlorine is passed into a suspension of benzidine hydrochloride in concentrated hydrochloric acid, 3,3′,5,5′-tetrachlorobenzidine [*41687-08-5*] is formed. Bromine forms the corresponding tetrabromobenzidine [*62477-23-0*]. 2-Nitrobenzidine [*2243-78-9*], 2,2′-dinitrobenzidine, and a little 2,3′-dinitrobenzene are formed from benzidine sulfate in concentrated sulfuric acid upon the addition of potassium nitrate [34]. Sulfonic acids or sulfones of benzidine are formed under sulfonating conditions. The N-acetylation products, which also occur as metabolites in animal digestion, are formed with acetic anhydride. They are *N*-acetyl-benzidine [*3366–61-8*], $C_{14}H_{14}N_2O$, and *N,N′*-diacetylbenzidine [*613–35–4*], $C_{16}H_{16}N_2O_2$.

The most important commercial reaction is the diazotization of the two amino groups. Reaction with nitrous acid converts benzidine into the tetrazonium compound [13], in which the first diazonium group is coupled very vigorously whereas the second reacts more slowly. As a result it is possible to produce asymmetrical diazo dyes. Gradual diazotization is also possible (→ Azo Dyes).

In the presence of acids, benzidine forms salts that are difficult to dissolve. Among these are benzidine monohydrochloride [*14414–68–7*], $C_{12}H_{12}N_2 \cdot HCl$, needles, sparingly soluble in water, readily soluble in dilute hydrochloric acid; benzidine dihydrochloride [*531-85-1*], $C_{12}H_{12}N_2 \cdot 2\,HCl$, flakes, readily soluble in water and in alcohol. The dihydrochloride hydrolyzes to form the monohydrochloride and hydrochloric acid.

Benzidine sulfate [*531-86-2*], $C_{12}H_{12}N_2 \cdot H_2SO_4$, forms microscopic flakes and is virtually insoluble in water and alcohol. Because of its poor solubility, benzidine sulfate can be used for the quantitative determination of sulfuric acid. Benzidine disulfate, $C_{12}H_{12}N_2 \cdot 2\,H_2SO_4$, is formed from benzidine sulfate and sulfuric acid.

Benzidine has been produced from nitrobenzene on an industrial scale since about 1880. Commercial production methods include alkaline iron reduction [14], amalgam reduction, and electrochemical reduction. The resultant hydrazobenzene is rearranged with hydrochloric acid or sulfuric acid during cooling. The base is then isolated in the form of benzidine hydrochloride or benzidine sulfate. The conversion of these salts to the free base is avoided as much as possible because of the chronic toxicity of benzidine.

An aqueous benzidine solution gives a blue coloration in the presence of aqueous potassium ferricyanide. The sensitivity is about 10 mg/kg.

In very dilute colorless bromine water, benzidine yields a blue coloration. The solution turns green when more reagent is added and finally becomes colorless again upon the precipitation of reddish flakes. Quantitative determination is by titration with 0.1 N sodium nitrite solution in hydrochloric acid, using potassium iodide and starch paper as the indicator. Quantitative determination of benzidine in dyestuffs involves petroleum ether extraction, column chromatographic separation, and colorimetric determination [35]. Benzidine can be detected in urine or air by extraction with *m*-cresol methyl ether and reaction with potassium 1,2-naphthoquinone-4-sulfonate [*5908-27-0*]; the detection limit is 0.33 mg/kg [36], [37]. A simple and rapid method for the detection of benzidine in the urine is based on its the color-forming reaction with cyanogen bromide; the detection limit is 0.05 mg/kg [38]. Benzidine in the urine can be separated from its metabolites by means of paper chromatography [39].

Benzidine is used as the reagent base for the production of a large number of dyes, particularly azo dyes for wool, cotton, and leather. However, because benzidine has been found to be car-

cinogenic in humans, there has been a marked decline in the use of the benzidine dyes.

Benzidine is used for the quantitative determination of sulfuric acid and for the detection and determination of numerous anions and metal ions. The reaction of benzidine with pyridine in the presence of elemental chlorine is suitable for detecting traces of free chlorine or pyridine in drinking water [38]. The green to blue coloration which occurs when benzidine reacts with hydrogen peroxide in the presence of peroxidases can be used to detect blood [40]. Benzidine still plays a role in many chemical syntheses.

3,3′,4,4′-Tetraaminodiphenyl [*91-95-2*], diphenyl tetramine, 3,3′,4,4′-biphenyltetrayltetramine, $C_{12}H_{14}N_4$, M_r 214.27, *mp* 178 °C, forms fused blocks and flakes which turn a dark color upon exposure to air; it is soluble in hot water.

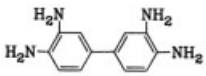

The hydrochloride [*7411-49-6*], $C_{12}H_{14}N_4 \cdot 4\,HCl \cdot 2\,H_2O$, forms needles and is readily soluble in water. The sulfate, $C_{12}H_{14}N_4 \cdot H_2SO_4$, forms small needles, is sparingly soluble in cold water, alcohol, and ether, and is soluble in hot water.

3,3′,4,4′-Tetraaminodiphenyl can be produced from benzidine by acetylation followed by nitration, saponification, and reduction with iron powder and hydrochloric acid [41].

The reduction of 3,3′-dinitrobenzidine can be carried out with hydrogen using a catalyst [42], with sodium dithionite in aqueous methanol, or (for preparative purposes) with tin(II) chloride and hydrochloric acid. The reaction of 3,3′- dichlorobenzidine with ammonia at high temperature and under pressure can also be used [43]. The crude product can be purified by recrystallization from aqueous sulfuric acid [44].

3,3′,4,4′-Tetraaminodiphenyl is suitable for the production of fibers based on polybenzimidazole. These fibers are resistant to high temperature and to chemicals. Fabrics made from them do not burn, do not melt in contact with a flame, and do not become hard when carbonized. They are suitable for the production of protective clothing and flue–gas filters and as substitutes for asbestos [45].

6.2. Tolidines

***o*-Tolidine** [*119-93-7*], 3,3′-dimethyl-4,4′-biphenyl-diyldiamine, $C_{14}H_{16}N_2$, M_r 212.30, *mp* 130 °C, colorless (in technical grades also brownish flakes), sparingly soluble in water, readily soluble in alcohol and ether. *o*-Tolidine is diacidic; at 30 °C, $K_1 = 6.2 \times 10^{-10}$, and $K_2 = 2.9 \times 10^{-10}$.

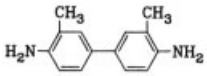

o-Tolidine barely becomes discolored upon exposure to air. It is chemically resistant to water. Numerous color reactions occur as a result of oxidation, e.g., an intense blue coloration forms on exposure to bromine or chlorine vapors. As with benzidine, these color reactions form the basis of a number of methods of detection and determination, such as those used for chlorine in air or water, for nitrates, and for hydrogen peroxide.

Mono- and disulfonic acids and *o*-tolidine-6,6-sulfone can be formed from *o*-tolidine sulfate.

o-Tolidine forms acid salts, such as *o*-tolidine hydrochloride, $C_{14}H_{16}N_2 \cdot HCl$, flakes. At 12 °C, one part dissolves in 112.4 parts of water.*o*-Tolidine dihydrochloride [*612-82-8*], $C_{14}H_{16}N_2 \cdot 2\,HCl$, forms crystals. At 12 °C, one part dissolves in 17.3 parts of water. *o*-Tolidine sulfate, $C_{14}H_{16}N_2 \cdot H_2SO_4$, forms needles or flakes. Water at 20 °C dissolves 0.12 g per 100 mL and more in the presence of hydrochloric acid. *o*-Tolidine disulfate [*64969-36-41*], $C_{14}H_{16}N_2 \cdot 2\,H_2SO_4$ forms crystals and is sparingly soluble in water.

o-Nitrotoluene undergoes alkaline reduction with zinc dust [13], electrolytic reduction [46], and catalytic reduction to form 2,2′-dimethylhydrazobenzene. This is rearranged using dilute hydrochloric acid or 20 % sulfuric acid at 5 – 50 °C. The free base or the dihydrochloride can be isolated.

The color reactions previously described can also be used for the detection of *o*-tolidine. Quantitatively, it is determined by titration with nitrite, as was described for benzidine.

o-Tolidine can be detected in the urine by extraction with 2,4-dichloroisopropylbenzene and conversion with potassium 1,2-naphthoquinone-4-sulfonate [47]. A rapid and simple procedure is based on the color reaction with cyanogen bromide; the detection limit in both cases is 0.05 mg/kg [38].

o-Tolidine is marketed as the free base and as the dihydrochloride. Both are supplied as moist products with a water content of 10–30 %.

o-Tolidine is a starting material in the production of a large number of azo dyes and pigments. *o*-Tolidine is used in the determination of oxygen and chlorine present in water, and for the colorimetric determination of cations of gold, cerium, and manganese.

An important derivative of *o*-tolidine is its diacetoacetyl compound, 4,4′-bisacetoacetylamino-3,3′-dimethyl diphenyl, which is marketed under the trade name Naphtol AS-G. It is a coupling agent which is frequently used and in combination with chloroanilines gives yellow shades. *o*-Tolidinediisocyanate is employed as a cross-linking agent to make polymers.

***m*-Tolidine** [*84-67-3*], 2,2′-dimethyl-4,4′-bi-phenyl-diyl-diamine, $C_{14}H_{16}N_2$, M_r 212.30, *mp* 109 °C, forms prisms from water, soluble in cold water, alcohol, and ether.

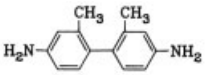

m-Tolidine gives no color in the presence of iron(III) chloride. It is tetrazotized with nitrite and acid and couples to form dyes.

m-Tolidine sulfate, $C_{14}H_{16}N_2 \cdot H_2SO_4$, is insoluble in cold alcohol, sparingly soluble in cold water, and soluble in hot water.

m-Tolidine is made by the reduction of *m*-nitrotoluene with zinc dust and sodium hydroxide solution [13]. The subsequent rearrangement is carried out with hydrochloric acid or sulfuric acid and gives rise to *m*-tolidine dihydrochloride or *m*-tolidine sulfate, which are isolated as salts. The free base is difficult to isolate and is therefore not a commercial product.

m-Tolidine was used in the past for the production of azo dyes.

6.3. Halogenated Benzidines

2,2′-Dichlorobenzidine [*84-68-4*], 2, 2′-dichloro-4, 4′-biphenyl-diyldiamine, $C_{12}H_{10}Cl_2N_2$, M_r 253.13, *mp* 166.8 °C, forms prisms from alcohol. It is almost insoluble in water but readily soluble in alcohol.

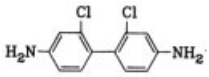

2,2′-Dichlorobenzidine dihydrochloride, $C_{12}H_{10}Cl_2N_2 \cdot 2\,HCl$, crystallizes in the form of flakes from water and is relatively soluble in water.

2,2′-Dichlorobenzidine is made from *m*-nitrochlorobenzene, preferably using zinc dust and sodium hydroxide solution. It is subsequently rearranged and isolated in the form of a dihydrochloride.

3,3′-Dichlorobenzidine [*91-94-1*], 3,3′-dichloro-4,4′-biphenyl-diyldiamine, $C_{12}H_{10}Cl_2N_2$, M_r 253.13, *mp* 132 – 133 °C, forms needles from alcohol or benzene. The commercial product is light brown to violet in color. It is virtually insoluble in water but readily soluble in alcohol, benzene, and glacial acetic acid.

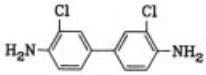

3,3′-Dichlorobenzidine is barely discolored on exposure to air and is chemically resistant to water. Oxidizing agents, such as bromine water, potassium dichromate, and iron(III) chloride, cause the formation of a green color. With gold, the green color is still detectable at a dilution of 1 : 5000000.

N-Acetyl-3,3′-dichlorobenzidine and *N,N*′-diacetyl-3,3′-dichlorobenzidine form when 3,3′- dichlorobenzidine is treated with acetic anhydride in dilute alcohol. The tetrazotizing behavior is similar to that of benzidine and *o*-tolidine.

3,3′-Dichlorobenzidine dihydrochloride [*612-83-9*], $C_{12}H_{10}Cl_2N_2 \cdot 2\,HCl$, is sparingly soluble in water and readily soluble in alcohol. It is marketed in the form of a colorless to pale grey crystalline powder.

3,3′-Dichlorobenzidine is made from *o*-nitrochlorobenzene by reduction with zinc dust and sodium hydroxide solution and subsequent rearrangement with dilute hydrochloric acid or sulfuric acid. It is assayed by titration with sodium nitrite solution in dilute hydrochloric acid and detected in human urine by colorimetry using chloramine. The detection limit is 0.1 mg/kg [38].

Like *o*-tolidine, 3,3′-dichlorobenzidine is marketed as a free base and as the dihydrochloride. Both are supplied as moist products with water contents of 5 – 30 %. The so-called *urethane quality* is taken to mean an anhydrous base, free of hydrochloric acid.

3,3′-Dichlorobenzidine was introduced about 1932 and is now the most important diphenyl base. It is used as the starting material for pigments with yellow and red shades. These are used for coloring printing inks, paints, plastics, and rubbers. The important diarylide yellow pigments, which are incorrectly known as benzidine yellows, are formed by the combination of 3,3′-dichlorobenzidine with acetic acid arylides. 3,3′-Dichlorobenzidine is also used in the production of polyurethane rubbers.

2,2′,5,5′-Tetrachlorobenzidine [*15721-02-5*], 2,2′,5,5′-tetrachloro-4,4′-biphenyldiyldiamine, $C_{12}H_8Cl_4N_2$, M_r 322.02, *mp* 134 – 137 °C.

2,2′,5,5′-Tetrachlorobenzidine is made by the reduction of 2,5-dichloronitrobenzene with zinc dust and sodium hydroxide solution. The subsequent rearrangement is carried out using sulfuric acid or hydrochloric acid. The dihydrochloride melts at 230 °C. Tetrachlorobenzidine is marketed as the free base and is used to produce yellow pigments.

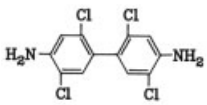

6.4. Alkoxybenzidines

***o*-Dianisidine** [*119-90-4*], 3,3′-dimethoxy-4,4′- biphenyl-diyldiamine, $C_{14}H_{16}N_2O_2$, M_r 244.30, crystallizes dimorphically, rarely in needles, with *mp* 133 °C, often in flakes with *mp* 137 – 138 °C. It forms colorless crystals, but commercial products have a tinge of violet. It is sparingly soluble in water but soluble in alcohol, ether, and benzene. One gram of ethyl acetate dissolves 0.285 g of *o*-dianisidine at 73 °C.

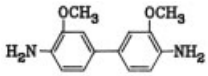

The pure compound is stable upon exposure to air, but commercial products turn violet. *o*-Dianisidine is resistant to water but sensitive to oxidizing agents.

o-Dianisidine dihydrochloride [*20325-40-0*], $C_{14}H_{16}N_2O_2 \cdot 2\,HCl$, forms prisms with *mp* 272 °C (decomp.). It is readily soluble in hot water and sparingly soluble in cold water and alcohol.

The hydrazo compound (2,2′-dimethoxyhydrazobenzene [*787-77-9*]) is obtained by the reduction of *o*-nitroanisole [*91-23-6*] with zinc dust and sodium hydroxide solution, often in a medium containing ethyl alcohol.

The reduction can also be performed electrochemically or with a palladium catalyst in an aqueous solution of isopropyl alcohol and sodium hydroxide. Rearrangement is carried out at a maximum of 20 °C with 20 % sulfuric acid [13], [14].

Like benzidine and *o*-tolidine, *o*-dianisidine forms colors with numerous oxidizing agents, e.g., copper, cobalt, and gold ions. *o*-Dianisidine is quantitatively determined by titration with nitrite using potassium iodide and starch paper as the indicator. *o*-Dianisidine can be detected in the urine using potassium 1,2-naphthoquinone-4-sulfonate after extraction [47]. A rapid and easy method is based on the color formed with cyanogen bromide. Its detection limit is 0.05 mg/kg [38].

The free *o*-dianisidine base and the dihydrochloride are marketed in moist forms with a 10 – 30 % water content.

o-Dianisidine is a starting material for the production of disazo dyes and pigments.

***o*-Diphenetidine** [*6264-77-3*], 3, 3′-diethoxy- 4, 4′-biphenyl-diyldiamine, $C_{16}H_{20}N_2O_2$, M_r 272.35, *mp* 119 °C, forms needles or flakes. It assumes a grey or violet shade in commercial products. It is insoluble in cold water, slightly

soluble in boiling water, and very soluble in alcohol, ether, and chloroform.

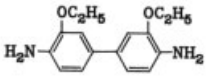

The aqueous solution turns red on contact with oxidizing agents, such as bromine, iron(III) chloride, or potassium permanganate. The dihydrochloride melts at 254 °C with decomposition.

o-Diphenetidine is made from 2-nitro-1-ethoxybenzene [*610-67-3*] by reduction using zinc dust, followed by rearrangement using hydrogen chloride gas in ethanol or dilute sulfuric acid.

Dichlorodianisidine [*5855-70-9*], 2,2′-dichloro-5,5′-dimethoxy-4,4′-biphenyl-diyldiamine, $C_{14}H_{14}Cl_2N_2O_2$, M_r 313.19, *mp* 175 °C, is a violet crystalline powder.

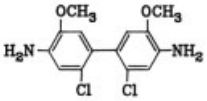

Dichlorodianisidine is made from 4-chloro-2-nitro-1-methoxybenzene [*89-21-4*] by zinc dust reduction. Dichlorodianisidine is used in special pigments.

6.5. Benzidinecarboxylic Acids

Benzidine-2,2′-dicarboxylic acid [*17557-76-5*], 4,4′-diaminobiphenyl-2,2′-dicarboxylic acid, $C_{14}H_{12}N_2O_4$, M_r 272.26, *mp* 265 °C, crystallizes from water in needles. It is sparingly soluble in water, alcohol, and ether. The dihydrochloride, $C_{14}H_{12}N_2O_4 \cdot 2\,HCl$, crystallizes in columns or needles.

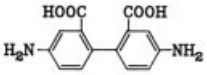

Benzidine-2,2′-dicarboxylic acid is made from 3-nitrobenzoic acid [*121-92-6*] by reduction with zinc dust and sodium hydroxide solution to form the hydrazo compound. This is followed by rearrangement using hydrochloric acid to form the dihydrochloride.

Benzidine-3,3′-dicarboxylic acid [*2130-56-5*], 4,4′-diaminobiphenyl-3,3′-dicarboxylic acid, $C_{14}H_{12}N_2O_4$, M_r 272.26, crystallizes into needles which are converted into benzidine by decarboxylation at temperatures above 250 °C. It is slightly soluble in alcohol and ether.

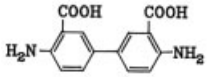

Benzidine-3,3′-dicarboxylic acid is made by reducing 2-nitrobenzoic acid [*552-16-9*] with zinc dust and sodium hydroxide solution. The resultant hydrazobenzene-2,2′-dicarboxylic acid [*612-44-2*] is rearranged by heating with hydrochloric acid. The free benzidine-3,3′-dicarboxylic acid is precipitated from the hydrochloric acid solution by adding ammonia or sodium acetate [13]. Electrolytic reduction followed by rearrangement with dilute sulfuric acid is also possible.

Benzidine-3,3′-dicarboxylic acid is used for the manufacture of azo dyes containing copper.

6.6. Benzidinesulfonic Acids and Benzidinesulfones

Benzidine-2,2′-disulfonic acid [*117-61-3*], 4,4′-diaminobiphenyl-2,2′-disulfonic acid, $C_{12}H_{12}N_2O_6S_2$, M_r 344.37, crystallizes in monoclinic prisms, chars without melting. One hundred grams of the aqueous solution at 22 °C contains 0.0791 g of the anhydrous acid. The compound is virtually insoluble in alcohol and ether.

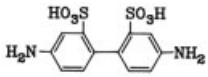

Benzidine-2,2′-disulfonic acid is made from the potassium or sodium salt of 3-nitrobenzenesulfonic acid by a gradual reduction using zinc dust and sodium hydroxide solution in a dilute aqueous solution. The product is subsequently rearranged using hydrochloric acid or dilute sulfuric acid [13]. Electrochemical or catalytic reductions are also possible [48].

The sulfonic acid must be diazotized indirectly because of its poor solubility. This is achieved by dissolving it in the necessary amount of sodium hydroxide solution, mixing the neutral solution with sodium nitrite and allowing the mixture to trickle into hydrochloric or sulfuric acid in a thin stream [13].

Benzidine-2,2′-disulfonic acid is used in the production of yellow and red azo dyes.

Benzidine-3,3′-disulfonic acid [*3365-90-0*], 4,4′-diaminobiphenyl-3,3′-disulfonic acid, $C_{12}H_{12}N_2O_6S_2$, M_r 344.37, forms tetrahedral flakes, virtually insoluble in alcohol and ether, very slightly soluble in boiling water.

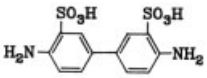

Benzidine-3,3′-disulfonic acid is made by reacting one part of benzidine sulfate with two parts of sulfuric acid monohydrate at 210 °C.

***o*-Tolidine-6,6′-disulfonic acid** [*83-83-0*], 4,4′-diamino-5,5′-dimethylbiphenyl-2,2′-disulfonic acid, $C_{14}H_{16}N_2O_6S_2$, M_r 372.42, crystallizes from water in needles. One hundred grams of the aqueous solution at 18 °C contains 0.226 g of the anhydrous acid. The substance is insoluble in alcohol, ether, and glacial acetic acid.

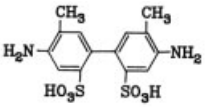

o-Tolidine-6,6′-disulfonic acid is produced from 2-nitrotoluene-4-sulfonic acid [*97-06-3*] by reduction with zinc and sodium hydroxide solution and subsequent rearrangement.

o-Tolidine-6,6′-disulfonic acid is used to make yellow dyes for wool.

Benzidinesulfone [*6259-19-4*], 3,7-diaminodibenzothiophene-5,5-dioxide, $C_{12}H_{10}N_2O_2S$, M_r 246.29, forms yellow needles or flakes, *mp* 330 – 332 °C, is virtually insoluble in boiling water, boiling alcohol, ether, and benzene.

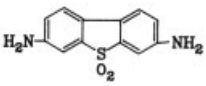

The dihydrochloride, $C_{12}H_{10}N_2O_2S \cdot 2\,HCl$, crystallizes as flakes from dilute hydrochloric acid and is decomposed by water. The sulfate, $C_{12}H_{10}N_2O_2S \cdot H_2SO_4 \cdot 1.5\,H_2O$, crystallizes as needles or flakes, is slightly soluble in water containing sulfuric acid, and is decomposed by water.

Benzidinesulfone is produced by heating benzidine sulfate with an excess of 20 – 40 % oleum in a water bath.

Benzidinesulfone-disulfonic Acid, 3,7-diaminodibenzothiophene-2,8-disulfonic acid 5,5- dioxide, $C_{12}H_{10}N_2O_8S_3$, M_r 406.41, crystallizes in pale yellow needles, soluble in hot water, slightly soluble in alcohol, virtually insoluble in cold hydrochloric acid and in dilute sulfuric acid.

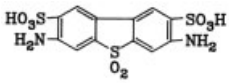

Benzidinesulfone-disulfonic acid is made by heating benzidine or benzidine sulfate with 40 % oleum at 150 °C.

7. Economic Aspects

Because of the carcinogenic properties of benzidine, a number of older benzidine plants have been replaced by new ones built to meet the industrial hygiene requirements. At the same time, every effort is being made to replace benzidine with less dangerous bases. Thus the demand for benzidine, which once amounted to several thousand tons per year, has been falling for some time.

There are no statistics available on the world production of diphenyl bases, but in 1983 about 15 kt was produced. Dichlorobenzidine and its salts accounted for about 7 – 10 kt, followed by *o*-tolidine with 2 – 3 kt. In contrast, *o*-dianisidine production was only about 1 – 2 kt and the production of tetrachlorobenzidine hardly more than 100 t.

8. Toxicology

8.1. Benzidine

The *acute* oral LD_{50} of benzidine administered to male rats is 1.57 g/kg [49]. *Subacute* dietary exposure of mice to concentrations of

100 – 800 ppm resulted in cloudy swelling of the liver, vacuolar degeneration of the renal tubules, and hyperplasia of myeloid elements of the bone marrow and lymphoid cells in the thymus and spleen [50]. Little is known about the *dermal* or *pulmonary absorption,* but the toxic systemic manifestations following these modes of exposure indicate that proportional amounts are absorbed [51].

If benzidine is injected intravenously, its half-life in blood exhibits several phases. The fourth and final phase has a half-life of 65 h in rats and 88 h in dogs [52]. The routes of elimination differed in the various animal species investigated. In rats, approximately 80 % of an i.v. dose was excreted via feces. In dogs and monkeys 67 % and in monkeys 50 % was eliminated via the urine [52].

Benzidine is metabolized to *N*-acetylbenzidine and *N,N′*-diacetylbenzidine, which is converted by an NADPH-dependent reaction to *N*-hydroxy-*N,N′*-diacetylbenzidine and 3-hydroxy-*N,N′*-diacetylbenzidine; the latter binds to nucleic acids [53].

Benzidine is positive in the Ames test following metabolic activation with liver homogenate from rats [54], [55] or humans [56]. The Ames test is also positive for the urine of rats that have been exposed to benzidine by oral administration [57]. *N*-Acetylbenzidine [57] and *N*-hydroxy-*N,N′*-diacetylbenzidine [53] also induce point mutations in the Ames system. Benzidine induces unscheduled DNA synthesis in HeLa cells [58] and in rat hepatocytes [59]. After metabolic activation, DNA strand breaks [60], [61] and cell transformation [62], [63] occur in V-79 cells.

Benzidine hydrochloride produces hepatocellular carcinoma in mice at a dose level of 150 mg/kg [64–67]. Evidence of the carcinogenic properties of benzidine has been found in rats [68] and hamsters [69].

The first information about the carcinogenic properties of benzidine in humans was published by Oppenheimer in 1927 [70]. He observed tumors of the bladder among workers in the dyestuffs industry. This observation has been confirmed by many other researchers [71–80]. The incidence of bladder cancer in workers decreased after reduction of the industrial exposure [80].

In summary, there is sufficient evidence that benzidine is carcinogenic in mice, rats, and hamsters. Ample evidence exists that benzidine also is carcinogenic in humans.

The MAK commission has classified benzidine in group A 1 (clearly carcinogenic). The American Conference of Governmental Industrial Hygienists (ACGIH) lists benzidine in group A 1 b: "no exposure or contact by any route – respiratory, skin, or oral, as detected by the most sensitive methods – shall be permitted."

8.2. 3,3′-Dichlorobenzidine

The oral LD_{50} (rat) of 3,3′-dichlorobenzidine is 7.07 g/kg [81].

In chronic studies 100 mg was administered orally to beagle dogs three times weekly for six weeks and five times weekly thereafter for seven years. Serum-glutamic-oxalacetic transaminase levels were increased [82], a sign of a slightly toxic effect upon the liver. Carcinoma of the bladder, hepatocellular carcinoma, and mammary gland tumors were also observed.

3,3′-Dichlorobenzidine is rapidly metabolized. After an oral dose of one gram, dogs excrete only 2 % unchanged in the feces and urine [83]. Rats and dogs treated with 3,3′-dichlorobenzidine showed multiphasic blood clearances. The half-life of the final phase was 68 h in rats and 86 h in dogs. The majority of the administered dose was recovered from the bile, thereby demonstrating the importance of the hepato-biliary excretion [52].

In the Ames test, 3,3′-dichlorobenzidine was positive with and without adding a metabolic activation system [55], [84], [85]. Positive results were also obtained in the unscheduled DNA synthesis test in HeLa cells [58] and in the cell transformation assay in baby hamster kidney cells [86].

When mice were treated with 3,3′-dichlorobenzidine in a dietary concentration of 1000 ppm for one year, all of the animals developed hepatomas [87]. In another experiment, rats developed tumors in various organs after an exposure for one year followed by an unlimited observation period [88]. The results are difficult to evaluate because of the absence of an adequate control group. When rats were treated with 1000 ppm of 3,3′-dichlorobenzidine in the

diet, significant statistical increases were found for granulocytic leukemias, mammary adenocarcinomas, and zymbal gland carcinomas [89]. Hamsters did not develop tumors at the same dose level, but following doses of 3000 ppm, tumors of the bladder and the liver were observed [69], [90].

In epidemiological studies of workers exposed to 3,3′-dichlorobenzidine, no bladder tumors were found [81], [91], [92]. In summary, 3,3′-dichlorobenzidine is carcinogenic in laboratory animals, but proof of its oncogenic properties in humans does not exist. Both commissions (MAK and ACGIH) have classified this compound in category 2, respectively, group A II. Neither MAK nor TLV values have been established. 3,3′-Dichlorobenzidine is absorbed by the skin.

8.3. *o*-Tolidine (3,3′-Dimethylbenzidine)

o-Tolidine can enter the body by percutaneous absorption, ingestion, or inhalation [93]. It is metabolized in exposed workers to diacetyl-*o*-tolidine and to a hydroxyamino metabolite (probably 5-hydroxy-*o*-tolidine), which are excreted via the urine [94]. Dogs are unable to acetylate aromatic amines such as *o*-tolidine [95].

In the Ames test, *o*-tolidine is a weak mutagen [85], [96]. *o*-Tolidine acts as a systemic carcinogen in Sherman rats when administered subcutaneously at a dose level of 60 mg per rat per week. It induces mainly tumors of the zymbal glands [97]. No carcinogenic potential could be observed in hamsters after feeding them a diet containing 1000 ppm of *o*-tolidine [69], [90]. No epidemiological studies in humans are available.

Exposure limits have been established. The MAK value is 5 mg/kg (22 mg/m^3, group III B), and the TLV is 2 mg/kg (9 mg/m^3, group A 2).

8.4. *o*-Dianisidine (3,3′-Dimethoxybenzidine)

Information about the biotransformation of *o*-dianisidine is rather limited. In dogs, 3,3′- dihydroxybenzidine is formed as a major metabolite [83].

In the presence of rat liver homogenate, *o*-dianisidine is positive in the Ames test [85]. It is capable of inducing unscheduled DNA synthesis in HeLa cells [58] and in primary rat hepatocytes [98], and it provokes cell transformation in baby hamster kidney cells in vitro [86]. *o*-Dianisidine is carcinogenic in rats. An oral dose of 30 mg administered three times per week for 13 months produced tumors in the zymbal glands [98], [99] and in other tissues [100], [101].

No conclusive epidemiological studies have been reported on the carcinogenicity of *o*-dianisidine in humans. No MAK value has been established. *o*-Dianisidine is classified in group III B.

9. References

General References

1. *Beilstein,* **13,** 214, 227, 234, 255, 256, 340, 807, 808; **14,** 567, 568, 770, 794, 795, 796; **18,** 591, 636; **13 (1),** 58, 66, 67, 79, 331, 332; **14 (1),** 647, 737, 743; **18 (1),** 590; **13 (2),** 90, 102, 106, 116, 152, 502, 504; **14 (2),** 344, 345, 479; **18 (2),** 490; **13 (3),** 425, 437, 447, 484, 2310; **14 (3),** 2264, 2265; **18 (3/4),** 7289, 8286.
2. *Houben-Weyl,* **vol. X/2,** pp. 693 – 743; **vol. XI/1,** pp. 839 – 848.
3. T. S. Scott: *Carcinogenic and Chronic Toxic Hazards of Aromatic Amines,* Elsevier, Amsterdam – New York 1962.
4. R. Powell, M. Murray, C. Chen, A. Lee: *Survey of the manufacture, import and uses for benzidine, related substances and related dyes and pigments,* Report 1979 EPA/560/13 – 79/005; Order No. PB-296544. Avail. NTIS. *Gov. Rep. Announce. Index (U.S.),* **79** (1979) no. 20, 120; *Chem. Abstr.* **92** (1980) 7831 e.
5. T. J. Haley: " Review of the literature and problems associated with the use of benzidine and its congeners," *Clin. Toxicol.* **8** (1975) no. 1, 13 – 42 (106 refs.).
6. W. Seidenfaden: *Künstliche organische Farbstoffe und ihre Anwendungen,* Enke Verlag, Stuttgart 1957.
7. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. I. (1952); vol. III (1970); vol. V (1971),** Academic Press, New York – London.
8. W. Herbst, K. Hunger: "Azo-Pigmente, Eigenschaften, Anforderungen, Entwicklungstendenzen," *Progr. Org. Coat.* **6** (1978) 105 – 210.

9. *Ullmann,* 4th ed., **vol. 18,** pp. 661 – 695.

Specific References

10. W. A. Rye, P. F. Woolrich, R. P. Zanes: "Facts and Myths Concerning Aromatic Diamine Curing Agents," *JOM J. Occup. Med.* **12** (1970) no. 6, 211 – 215.
11. R. Merten: "Die Synthese heterocyclischer Ringsysteme für wärmebeständige Kunststoffe aus Polyisocyanaten," *Angew. Chem.* **83** (1971) 339 – 347; *Angew. Chem. Int. Ed. Engl.* **10** (1974) 294.
12. A. Horvath, B. Vollmert: "Über den Einfluß von Triäthylamin auf die Viskosität von Polyamidcarbonsäuren aus Pyromellitsäuredianhydrid und Benzidin in Dimethylacetamid," *Angew. Chem.* **83** (1971) 375; *Angew. Chem. Int. Ed. Engl.* **10** (1971) 348.
13. H. E. Fierz-David, L. Blangey: *Grundlegende Operationen der Farbenchemie,* 8th ed., Springer Verlag, Wien 1982.
14. P. H. Groggins: *Unit Processes in Organic Synthesis,* 5th ed., McGraw-Hill, New York – Toronto – London 1958.
15. *Houben-Weyl,* **vol. IV/2,** p. 494.
16. D. V. Banthorpe et al., *J. Chem. Soc. Perkin Trans. 2* 1973, 551 – 56. H. J. Shine et al., *J. Am. Chem. Soc.* **99** (1977) 3719 – 23. Z. J. Allan, *Justus Liebigs Ann. Chem.* 1978, no. 5, 705 – 09.
17. International Labour Office: "Occupational Safety and Health Series No. 39," *Occupational Cancer-Prevention and Control,* Geneva 1977. M. R. Zavon, U. Hoegg, E. Bingham: "Benzidine exposure as a cause of bladder tumors," *Arch. Environ. Health* **27** (1973) no. 1, 1 – 7.
18. *IARC Monogr.* **29** (1982) 149 – 181 (Benzidine and its Sulfate, Hydrochloride and Dihydrochloride); 239 – 256 (3,3′-Dichlorobenzidine and its Dihydrochloride).
19. Department of Labor. Occupational Safety and Health Administration: *Carcinogens,* Occupational Health and Safety Standards, Federal Register, vol. 39, no. 20, Part III, Washington, D. C., 1974, pp. 3757, 3771 – 3773.
20. Statutory Instruments 1967, no. 879: Factories. The Carcinogenic Substances Regulations 1967. Publ. by Her Majesty's Stationery Office 1967.
21. Verordnung über gefährliche Arbeitstoffe vom 11. 2. 1982, Bundesgesetzblatt 1982, part I, p. 144; Berufsgenossenschaft der chemischen Industrie, Unfallverhütungsvorschrift Nr. 47; Schutzmaßnahmen bei Umgang mit krebserzeugenden Arbeitsstoffen vom 1. 10. 1982.
22. Zwölfte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes – Störfall-Verordnung – vom 27. 6. 1980. Bundesgesetzblatt 1980, part I, p. 772.
23. Hauptverband der gewerblichen Berufsgenossenschaften, Spezifische Einwirkungsdefinitionen, ZH 1/600.16, proposal of 10.1979, Carl Heymanns Verlag, Köln.
24. Arrêté Royal du 3. 10. 1973, Moniteur Belge 23. Nov. 1973.
25. Labour Safety and Hygienic laws 2. 6. 1972.
26. R. Montesano, L. Tomatis: "Legislation Concerning Chemical Carcinogens in Several Industrialized Countries," *Cancer Res.* **37** (1977) 310 – 315.
27. H. T. Appleton, H. C. Sikka: "Accumulation, Elimination and Metabolism of Dichlorobenzidine in the Bluegill Sunfish," *Environ. Sci. Technol.* **14** (1980) no. 1, 50 – 54.
28. P. Y. Lu, R. L. Metcalf, N. Pummer, D. Mandel: "The environmental fate of three carcinogens," *Arch. Environ. Contam. Toxicol.* **6** (1977) no. 2 – 3, 129 – 42; *Chem. Abstr.* **88** (1978) 27196.
29. R. Morales, S. M. Rappaport, R. E. Hermes: "Air sampling and analytical procedures for benzidine, 3,3′-dichlorobenzidine and their salts," *Am. Ind. Hyg. Assoc. J.* **40** (1979) 970.
30. H. C. Sikka, H. T. Appleton, S. Banerjee; *Fate of 3,3′-Dichlorobenzidine in Aquatic Environments,* U.S.-E.P.A.-600/3-78-068, National Technical Information Service, Springfield, Virginia 22161, USA, 1978.
31. S. Banerjee, H. C. Sikka, R. Gray, C. M. Kelly: "Photodegradation of 3,3′-Dichlorobenzidine," *Environ. Sci. Technol.* **12** (1978) no. 13, 1425 – 27.
32. R. J. Hackman, T. Rust: "Removal and decontamination of residual 3,3′-Dichlorobenzidine," *Am. Ind. Hyg. Assoc. J.* **42** (1981) 341 – 47.
33. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft) vom 28. 8. 1974. Gemeinsames Ministerialblatt Ausgabe A, **25**, 4. 9. 1974, p. 442.
34. *Ullmann,* 4th ed., **vol. 8,** p. 359.

35. F. Jones, D. Patterson, D. Srinivasan: "The determination of free benzidine in dyes," *J. Soc. Dyers Color.* **96** (1980) no. 12, 628 – 31.
36. P. Engelbertz, E. Babel: "Nachweis von Benzidin und seinen Umwandlungsprodukten im Harn und in Organteilen," *Zentralbl. Arbeitsmed. Arbeitsschutz* **3** (1953) 161 – 68. P. Engelbertz, E. Babel: "Über die Anwendung der Benzidin-Testmethode mit 1,2-naphthochinon-4-sulfonsaurem Kalium zur Bestimmung der gesundheitlichen Gefährdung in Benzidinbetrieben," *Zentralbl. Arbeitsmed. Arbeitsschutz* **4** (1954) 40 – 42.
37. P. Engelbertz, E. Babel: "Nachweis und Bestimmung von 3,3′-Dichlorbenzidin und 3-Oxybenzidin," *Zentralbl. Arbeitsmed. Arbeitsschutz* **6** (1956) 58 – 60.
38. O. Winkler: "Beitrag zum Nachweis von Diphenylbasen im Harn," *Zentralbl. Arbeitsmed. Arbeitsschutz* **9** (1959) 140 – 42.
39. S. Laham, J. P. Farant, M. Potvin, *Ind. Med. Surg.* **39** (1970) 142 – 47.
40. H. Steinberg: "The hazard of benzidine to criminal justice personnel," *NBS Spec. Publ. U.S.* 1977, 480 – 421; *Chem. Abstr.* **87** (1977) 178486.
41. W. Zönnchen, DD 31935, 1963.
42. Bayer, DE 2232095, 1972.
43. Celanese Corp., US 3865876, 1969.
44. Celanese Corp., US 3943175, 1969.
45. D. R. Coffin, G. A. Serad, H. L. Hikcs, R. T. Montgomery: "Properties and Applications of Celanese PBI-Polybenzimidazole Fiber," *Text. Res. J.* **52** (1982) no. 7, 466 – 72.
46. Borma BV, DE 3020846, 1979.
47. P. Engelbertz, E. Babel: "Nachweis von o-Tolidin und o-Dianisidin sowie ihrer Umwandlungsprodukte im Harn und in Organteilen," *Zentralbl. Arbeitsmed. Arbeitsschutz* **4** (1954) 179 – 83.
48. Clayton Aniline Co., EP 32784, 1980.
49. J. Marhold, M. Matrka, M. Hub, F. Ruffer, *Neoplasma* **15** (1968) 3 – 10.
50. K. V. N. Rao, J. H. Rust, N. Mihailovich, S. D. Vesselinovitch, J. M. Rice, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **30** (1971) 444.
51. G. Ghetti, *Med. Lav.* **51** (1960) 102 – 114.
52. H.-M. Kellner, O. E. Christ, K. Lötzsch, *Arch. Toxicol.* **31** (1973) 61 – 79.
53. K. C. Morton, F. A. Belland, F. E. Felland, N. F. Fulland, F. F. Kandlubarr, *Cancer Res.* **40** (1980) 751 – 757.
54. B. N. Ames, W. E. Durston, E. Yamasaki, F. D. Lee, *Proc. Natl. Acad. Sci. U.S.A.* **70** (1973) 2281 – 2285.
55. D. Anderson, J. A. Styles, *Br. J. Cancer* **37** (1978) 924 – 930.
56. S. Haworth, T. Lawlor, in: 9th Annual Meeting of the Environmental Mutation Society, San Francisco 1978, Abstract No. 1 a – 8.
57. K. J. Tanaka, S. Marui, T. Mii, *Mutat. Res.* **79** (1980) 173 – 176.
58. C. N. Martin, A. C. McDermid, R. C. Garner, *Cancer Res.* **38** (1978) 2621 – 2627.
59. G. M. Williams, *Cancer Let.* **4** (1978) 69 – 75.
60. J. A. Swenberg, G. L. Petzold, P. R. Harbach, *Biochem. Biophys. Res. Commun.* **72** (1976) 732 – 738.
61. G. L. Petzold, J. A. Swenberg, *Cancer Res.* **38** (1978) 1589 – 1594.
62. J. Ashby, J. A. Styles, D. Paton, *Br. J. Cancer* **38** (1978) 34 – 50.
63. R. J. Pienta, in F. J. de Serres, E. Hollaender (eds.): *Chemical Mutagens. Principles and Methods for their Detection,* **vol. 6,** Plenum, New York 1980, pp. 175 – 202.
64. S. D. Vesselinovitch, K. V. N. Rao, N. Mihailovich, *Cancer Res.* **35** (1975) 2814 – 2819.
65. S. D. Vesselinovitch, K. V. N. Rao, N. Mihailovich, *Natl. Cancer Inst. Monogr.* **51** (1979) 239 – 250.
66. C. H. Frith, K. P. Baetcke, C. J. Nelson, G. Schieferstein, *Toxicol. Lett.* **4** (1979) 507 – 518.
67. C. H. Frith, K. P. Baetke, C. J. Nelson, G. Schieferstein, *Eur. J. Cancer* **16** (1980) 1205 – 1216.
68. E. Boyland, J. Harris, E. S. Horning, *Br. J. Cancer* **8** (1954) 647 – 654.
69. U. Saffiotti, F. Cefis, R. Montesano, A. R. Sellakumar, in W. B. Deichmann, K. F. Lampe (eds.): Bladder Cancer: A Symposium, Birmingham, Aesculapius Publ. Co., Birmingham 1967, pp. 129 – 135.
70. R. Oppenheimer, *Z. Urol. Chir. (Gynaekol.)* **21** (1927) 336 – 350.
71. A. Muller, *Z. Urol. Chir. (Gynaekol.)* **36** (1933) 202 – 219.
72. G. Di Maio, *Arch. Ital. Urol.* **14** (1937) 283 – 385.
73. M. W. Goldblatt, *Br. J. Ind. Med.* **6** (1949) 65 – 81.
74. M. Barsotti, E. C. Vigliani, *Med. Lav.* **40** (1949) 129 – 138.
75. T. S. Scott, *Br. J. Ind. Med.* **9** (1952) 127 – 132.
76. P. Aboulker, G. Smagghe, *Arch. Mal. Prof.* **14** (1953) 380 – 386.
77. F. Uebelin, A. Pletscher, *Schweiz. Med. Wochenschr.* **84** (1954) 917 – 920.

78. M. Douilett, J. Bourret, A. Convert, *Arch. Mal. Prof.* **20** (1959) 713 – 733.
79. J. L. Billard-Duchesne, *Acta Unio Int. Cancrum* **16** (1960) 284 – 288.
80. K. M. Ferber, W. J. Hill, D. A. Cobb, *Am. Ind. Hyg. Assoc. J.* **37** (1976) 61 – 68.
81. H. W. Gerarde, D. F. Gerarde, *J. Occup. Med.* **16** (1974) 322 – 344.
82. E. F. Stula, H. Sherman, W. R. Clayton, *Toxicol. Appl. Pharmacol.* **31** (1975) 159 – 176.
83. L. J. Sciarmy, J. W. Meigs, *Arch. Environ. Health* **2** (1961) 584 – 588.
84. R. C. Garner, A. L. Walpole, F. L. Rose, *Cancer Lett.* **1** (1975) 39 – 42.
85. E. J. Lazear, S. C. Louie, *Cancer Lett.* **4** (1977) 21 – 25.
86. I. A. Styles, *Br. J. Cancer* **37** (1978) 931 – 936.
87. H. Osanai, *J. Sci. Labour* **52** (1976) 179 – 201.
88. G. B. Pliss, *Vopr. Onkol.* **5** (1959) 524 – 533.
89. E. F. Stula, J. R. Barnes, H. Sherman, C. F. Reinhardt, J. A. Zapp, *J. Environ. Pathol. Toxicol.* **1** (1978) 475 – 490.
90. A. R. Sellakumar, R. Montesano, U. Saffiotti, *Proc. Am. Assoc. Cancer Res.* **10** (1969) Abstr. 309.
91. T. Gadian, *Chem. Ind.* 1975, 4th Oct., 821 – 831.
92. I. McIntyre, *J. Occup. Med.* **17** (1975) 23 – 26.
93. J. W. Meigs, L. J. Sciarini, W. A. Van Sandt, *Arch. Ind. Hyg.* **9** (1954) 122.
94. H. M. L. Dieteren, *Arch. Environ. Health* **12** (1966) 30.
95. IARC, *Monographs on the evaluation of carcinogenic risk of chemicals to man,* **vol. I,** International Agency for Research on Cancer, Lyon 1972, pp. 88 – 91.
96. J. I. Ferretti, W. Lu, M. B. Liu, *Am. J. Clin. Pathol.* **67** (1977) 526 – 527.
97. G. B. Pliss, M. A. Zabezhinsky, *J. Natl. Cancer Inst.* **45** (1970) 283.
98. G. B. Pliss, *Gig. Tr. Prof. Zabol.* **9** (1965) 18.
99. G. B. Pliss, *Acta Unio Int. Cancer* **19** (1963) 499.
100. ACGIH, Documentation of the Threshold Limit Values, 3rd ed., American Conference of Governmental Industrial Hygienists, Cincinatti, Ohio, 1971.
101. Z. Hadidian, T. N. Fredrickson, E. K. Weisburger, J. H. Weisburger, R. M. Glass, N. Mantel, *J. Natl. Cancer Inst.* **41** (1968) 985.

Benzoic Acid and Derivatives

TAKAO MAKI, Dia Research Martech Inc., Yokohama, Japan (Chaps. 1 – 8)

KAZUO TAKEDA, Mitsubishi Chemical Safety Institute Ltd., Yokohama, Japan (Chap. 9)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 2
5. Quality Specifications 3
6. Storage, Transportation and Legal Aspects 3
7. Uses and Economic Aspects 4
8. Derivatives of Benzoic Acid 5
8.1. Salts of Benzoic Acid 5
8.2. Esters of Benzoic Acid 5
8.3. Benzoyl Chloride 6
8.4. Benzonitrile 7
8.5. Alkyl and Acyl Analogues 7
8.6. Chlorobenzoic Acids 9
8.7. Aminobenzoic Acids 10
8.8. Nitrobenzoic Acids 12
8.9. 3-Sulfobenzoic Acid 13
8.10. Hexahydrobenzoic Acid 13
9. Toxicology 13
10. References 13

1. Introduction

Benzoic acid [*65-85-0*], $C_7H_6O_2$. The name benzoic acid originates from gum benzoin, a balsamic resin obtained from a South Asian plant called styrax. The extraction of benzoic acid was carried out by SCHEELE in 1775. Its structure was determined by LIEBIG and WÖHLER in 1832. The initial production methods were developed in the late 1800s. They were based on the hydrolysis of benzotrichloride or the decarboxylation of phthalic anhydride. Today, benzoic acid is produced by oxidation of toluene with air, which has displaced dichromate and nitric acid oxidation processes.

Occurrence. Benzoic acid and its derivatives are widely distributed in nature. Gum benzoin contains from 12 – 18 % benzoic acid in free and esterified forms. Other natural products containing benzoic acid are the bark, foliage, fruits, and seeds of various plants, including cherries and prunes. Hippuric acid, found in the urine of herbivores, is a glycine derivative of benzoic acid.

2. Physical Properties

Benzoic acid, benzenecarboxlic acid, M_r 122.12, *mp* 122 °C, *bp* 250 °C, forms white monoclinic crystals that begin to sublime at about 100 °C. It is volatile in steam and forms azeotropes with biphenyl, butyl benzoate, catechol, diphenyl ether, diphenylmethane, and naphthalene. At 89.7 °C, a mixture of excess benzoic acid and water forms two stable liquid phases. The water content of the benzoic acid phase is 26.5 wt %. The two phases become homogeneous at 117.2 °C. The mixture then contains 32.34 % benzoic acid and 67.66 % water.

Some physical properties of benzoic acid are listed in the following:

Solubility in water, g/100 g	
at 0 °C	0.17
at 20 °C	0.29
at 40 °C	0.60
at 60 °C	1.20
at 80 °C	2.75
at 95 °C	6.85
Solubility in organic solvents at 25 °C, g/100 g	
Acetone	55.60
Benzene	12.17
Tetrachloromethane	4.14
Ethanol (absolute)	58.40
Hexane	0.94 (at 17 °C)
Methanol	71.50 (at 23 °C)
Toluene	10.60
Vapor pressure, kPa	
at 96 °C	0.133
at 132 °C	1.33
at 186 °C	13.3
Vapor density relative to air	4.2
Refractive index	
n_D^{15}	1.53974
$n_D^{131.9}$	1.504
Relative density	
d_4^{20}	1.321
$d_4^{122.375}$	1.0819
d_4^{180}	1.02942

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_555

Specific heat	
solid (20 – 122 °C)	1.204 J g^{-1} K^{-1}
liquid (122 – 322 °C)	1.775 J g^{-1} K^{-1}
Heat of fusion	18 kJ/mol
Heat of sublimation at 110 °C	86.2 kJ/mol
Heat of vaporization at 150 °C	68.2 kJ/mol
Heat of formation	− 384.9 kJ/mol
Heat of combustion at 25 °C	3227 kJ/mol
Free energy	
solid	− 251 kJ/mol
liquid	− 247 kJ/mol
aqueous solution	− 242 kJ/mol
Flash point	121.1 °C
Ignition temperature	
in air	573 °C
in oxygen	556 °C
Dissociation constant at 25 °C, K_a	6.335×10^{-5}
pH of the saturated solution at 25 °C	2.8
Change in volume upon freezing	− 0.138 cm^3/g
log POW *	1.870

* Bioaccumulation factor (octanol/water partition factor)

3. Chemical Properties

Benzoic acid is stable towards common oxidizing agents. Air, permanganate, chromic acid, hypochlorite, and dilute nitric acid do not affect it. However, above 220 °C it reacts with copper(II) salts to form phenol and its derivatives [1]. This reaction, followed by regeneration of the copper(II) with molecular oxygen, is the basis for the commercial production of phenol from benzoic acid [2] (→ Phenol). Benzoic acid reacts with ammonia under similar conditions to form aniline [3]. Upon heating to 150 °C, some dehydration takes place to form benzoic anhydride. Decarboxylation occurs when benzoic acid is heated above 370 °C or as low as 245 °C in the presence of catalysts [4]. Benzene and a small amount of phenol are formed. Copper and cadmium powder accelerate the decarboxylation. When potassium benzoate is heated in carbon dioxide, it disproportionates to terephthalate and benzene. Cadmium and zinc salts catalyze this reaction. Benzoic acid is converted into hydroxybenzoic acids by fused potassium hydroxide [5].

Benzoic acid can be directly hydrogenated with molecular hydrogen to afford benzaldehyde in quantitative yield. Zirconium and other metal oxides catalyze the reaction [31]. Hydrogenation on noble metal catalysts gives cyclohexanecarboxylic acid (hexahydrobenzoic acid). Chlorination yields chiefly 3-chlorobenzoic acid [6]. Nitration forms 3-nitrobenzoic acid [7], and sulfonation forms 3-sulfobenzoic acid.

4. Production

Liquid-Phase Oxidation of Toluene with Molecular Oxygen. This process was developed by I.G. Farbenindustrie in Germany during World War II and was operated until the end of the war [8]. The plant consisted of a bubble-column oxidation reactor made of aluminum-lined steel and equipped with an air inlet tube, a separator, and an absorber to recover toluene. The reactor contained heat-exchange coils to remove the great amount of heat generated in the reaction. The purity of the toluene was critical because sulfur compounds, nitrogen compounds, phenols, and olefins inhibit the oxidation reaction. The reaction was carried out at 140 °C and ca. 0.2 MPa with a cobalt naphthenate catalyst (0.1 %) for 30 h. About 50 % conversion of toluene was achieved, 80 % of which formed benzoic acid. Other oil-soluble cobalt salts were also used as catalysts. The crude oxidation product was neutralized with sodium carbonate to produce sodium benzoate. Unreacted toluene was recovered from the spent oil phase by distillation.

The elemental mechanism of this oxidation is a free-radical chain process [9]. The cobalt catalyst shortens the induction period of the reaction and retards the accumulation of inhibitors. Peroxides are reaction intermediates. Manganese behaves similarly to cobalt as a catalyst, but its performance is weaker. Copper adversely affects the oxidation. High pressures retard the reaction, especially at lower catalyst concentrations. Above 130 °C, the reaction rate is determined by the diffusion of oxygen. Phenolic compounds gradually accumulate during the reaction and eventually inhibit it.

In a typical modern process the oxidation is conducted at 165 °C and 0.9 MPa. The reaction heat is removed by external circulation of the contents of the reactor. The pressure of the liquid discharged form the reactor is reduced to atmospheric, and unreacted toluene is recovered. Benzoic acid is purified by rectification. The bottom residue is extracted to recover the cobalt

catalyst. The exhaust gas is cooled to recover most of the toluene, purified, and vented. The oxidation reactor can be a bubble column or a stirred tank. Stainless steel is generally used as the construction material. Heat removal is still a serious problem, and air distribution within the reactor is critical to prevent a runaway reaction. The oxygen content of the exhaust gas should be strictly monitored to avoid an explosion.

The oxidation produces several byproducts. Benzaldehyde is formed in fairly large amounts, especially at lower conversions. It can be recovered by distillation. A limited amount of benzyl alcohol is formed, but its recovery is impractical because of the unavoidable esterification with benzoic acid. Benzyl benzoate can be recovered from the bottoms of the benzoic acid column [10]. Other esters, including benzyl formate and benzyl acetate, are also present. Biphenyl and methylbiphenyls are formed in smaller amounts. Because their vapor pressures are close to that of benzoic acid, they play a role in the rectification of benzoic acid. The product is contaminated by small amounts of phthalic acid. Carbon monoxide and carbon dioxide are detected in the exhaust gas. The yield of benzoic acid is about 90 %, but it varies with toluene conversion. At lower conversions, the selectivity for benzaldehyde and benzyl alcohol exceeds 10 %.

Conventional technical-grade benzoic acid is usually of as-rectified quality. Further processing is necessary for higher grade products. Sublimation, recrystallization, and neutralization processes have been proposed. Thermal treatment in an inert atmosphere is attempted [11]. To remove phthalic acid, whose presence is not allowed for food or pharmaceutical uses, treatment with amines [32] and rinsing [33] methods have been reported. Methylbiphenyls are responsible for the odor of the product. For their removal, extraction [34] and treatment in an inert gas stream [35] have been proposed.

The oxidation reactions of alkylaromatic hydrocarbons are dramatically enhanced by adding bromine compounds to the cobalt catalyst in acetic acid [12]. This method is now widely used in the production of terephthalic acid and is also applicable to the benzoic acid process. Under certain conditions, benzoic acid itself can be used as a solvent. Manganese has synergistic effect on the cobalt catalyst. This process can realize high reaction rates and better toluene conversion, as well as eliminating reaction inhibition. However, to avoid corrosion, the equipment must be constructed of expensive titanium.

Another type of cobalt catalyst activator is a readily oxidizable carbonyl compound such as methyl ethyl ketone or acetaldehyde. These additives permit a considerable reduction in reaction temperature, but because of the high consumption of these activators, this technology does not seem to be practical.

The catalytic oxidation of toluene in aqueous sodium hydroxide directly affords sodium benzoate. This process has been tried since 1940s, but satisfactory yields have not been achieved.

Vapor-phase oxidation of toluene at higher temperatures is another potential process. Various catalysts (vanadium, tungsten, uranium, molybdenum, silver, etc.) have been investigated, but yields were not comparable to those of the liquid-phase process. However, a patent claims a significant improvement in yield by using complex vanadia catalysts [36].

5. Quality Specifications

Quality specifications for benzoic acid are listed in Tables 1 and 2. An EU guideline for the use of benzoic acid and its derivatives has been issued [37].

6. Storage, Transportation and Legal Aspects

Benzoic acid is commercially handled as flakes packed in 25 kg paper bags or in 500 kg flexible containers. Bulk quantities are transported as molten liquid in stainless steel vessels. For transport, benzoic acid is not regulated as a hazardous material (IMO, IATA, ADR/SDR, and RID/SDR). There is no IMDG code. The reportable quantity in USA is 2270 kg (40 CFR 355.40). The European hazard-labeling symbol is Xn (harmful). The MSDN is available on the Internet [38], [39].

Benzoic acid can be stored without deterioration, but the low-grade product tends to solidify on storage. Precautions should be taken against dust during handling since it irritates the

Table 1. Quality specifications for technical-grade benzoic acid

Items	Specifications	Specifications for commercial products			
	Japan	Japan	USA		Europe
	(JIS K 4127, 1995)	Commercial grade	Industrial grade	Technical grade	Commercial grade
mp, °C	120 – 123	121 – 123	121 – 123	121.5 – 123	122 – 122.5
Assay, % min.	99.0 (by glc)	99.6	97.5	99.0	99.5
Moisture, % max.	0.3 (by titration)				
Low-boiling substances, % max.	0.5 (by glc)				
High-boiling substances, % max.	0.5 (by glc)				
Heavy metals, ppm max.				10	10
Phthalic acid, % max.	0.5 (by glc)				
Ignition residue, % max.	0.1	0.1	0.1	0.1	0.1
Appearance	white	off-white	brownish	white	white

Table 2. Quality specifications of pharmaceutical- and food-grade benzoic acid

	Pharmaceutical			Food additive	
	Japan (JP12 1993)	USA (USP23 1995)	Germany	Japan (Official)	USA (FCC 1996)
Identification	pass test	pass test	pass test		
mp, °C	121 – 124	121 – 123 *(fp)*	121 – 124	121 – 123	121 – 123 *(fp)*
Assay, % min. (by titration)	99.5	99.5 – 100.5	99.0 – 100.5	99.5	99.5 – 100.5
Moisture, % max.	0.5 (by weight loss)	0.7		0.5	0.7 (by titration)
Heavy metals, ppm max.	20	10	10	10	10
Arsenic, ppm max.		3		4	
Readily oxidizable substances (by permanganate)	pass test	pass test	pass test		pass test
Readily carbonizable substances (by sulfuric acid)	pass test	pass test	pass test		pass test
Phthalic acid (by fluorescence)	pass test			pass test	
Halogen compounds	pass test		pass test	0.014 % max.	
Ignition residue, % max.	0.05	0.05	0.1		0.05
Solubility			pass test		

eyes and skin. Dust/air mixtures can be explosive above the flash point (121°C). The explosion limit of the vapor/air mixture is 0.085-0.99g/m^3.

Benzoic acid has the following registry number for the legal control: TSCA 65-85-0, EINECS (European Inventory of Existing Chemical substances) 200-618-2, ECNS (Japan) (3)-1397, RTECS DG0875000.

7. Uses and Economic Aspects

The majority of the benzoic acid is utilized captively to produce phenol. The rest comes onto the market. The benzoic acid market in the USA and Japan is summarized in Table 3. Major producers are listed in Table 4.

Glycol esters of benzoic acid are used as plasticizers for vinyl resins. Demand has increased in the USA as a replacement for butyl benzyl

phthalate in poly(vinyl acetate) emulsion adhesives [40]. Another major market is sodium benzoate for preservatives and anticorrosives. Benzoic acid is used in modified alkyd resins for automotive refinishing enamels to improve surface properties. Benzoic acid is also used as an intermediate for perfumes, pharmaceuticals, and cosmetics. The market in USA has experienced growth rate of 4 % per annum [42].

Table 3. Benzoic acid consumption (in %)

Item	USA 1991	Japan 1991
Plasticizers	50	0
Sodium benzoate	25	55
Benzoyl chloride	10	15
Alkyd resins	5	10
Others	10	20

Table 4. Major benzoic acid producers

Company	Capacity, 10^3 t/a
Kalama Chemical (USA)	95
Velsicol Chemical (USA)	30
DSM (NL)	120
Chimica del Fruita (IT)	30
Velsicol Eesti (Estonia)	30
Liquid Quimica (E)	5
Mitsubishi Chemical (JP)	10
Nippon Steel Chemical (JP)	captive

8. Derivatives of Benzoic Acid

8.1. Salts of Benzoic Acid

Sodium benzoate [*532-32-1*], $C_7H_5O_2Na$, M_r 144.11, forms white granules, flakes, or a crystalline powder with a sweetish taste. One gram dissolves in 2 mL of water. Between 0 °C and 50 °C the solubility in water is nearly constant, but at higher temperature it increases sharply. At −13.5 °C, the eutectic mixture with water consists of 44.9 g of the salt and 100 g of water. The aqueous solution is slightly alkaline (pH ca. 8). Sodium benzoate is soluble to the extent of 0.8 g in 100 mL of ethanol, 8.2 g in 100 mL of methanol, and over 20 g in 100 mL of ethylene glycol. Aqueous and alcoholic solutions of sodium benzoate passivate the surfaces of metals and alloys.

Dry sodium benzoate is electrically charged by friction and forms an explosive mixture when its dust is dispersed in air. It absorbs moisture from air to some extent. Upon heating in an inert atmosphere, sodium benzoate carbonizes and forms sodium carbonate. In the presence of air, it burns to give carbon dioxide and sodium carbonate. Sodium benzoate reacts with benzoyl chloride to form benzoic anhydride [*93-97-0*]. Other salts of benzoic acid can be prepared by metathesis reactions with the salts of particular cations. Sodium benzoate is produced by the neutralization of benzoic acid with sodium hydroxide. Direct oxidation of toluene in sodium hydroxide solution is not currently in use.

The principal use of sodium benzoate is an anticorrosive. The expansion of the motor industry has increased demand. It is added mainly to antifreeze coolants. It is also used as a preservative for foods and cosmetics. Specifications for food-grade sodium benzoate are listed in Table 5. The CFR GRAS level is 0.1 % max., and its use below pH 4.5 is recommended [41].

Table 5. Specifications for food-grade sodium benzoate

	Japan	USA (FCC, 1996)
Appearance	pass test	pass test
Solubility	pass test	
Assay, % min. (by titration)	99.0	99.0 – 100.5
Moisture, % max. (by weight loss)	1.5	1.5
Free alkali, % max. (as NaOH)	pass test	0.04
Free acid	pass test	
Heavy metals, ppm max. (as Pb)	10	10
Arsenic, ppm max.	4	
Readily oxidizable substances (by permanganate)	pass test	
Phthalic acid (by fluorescence)	pass test	
Halogen compounds, % max. (by Ag)	0.014	
Sulfate, % max.	0.3	

Potassium benzoate [*582-25-2*], $C_7H_5O_2K$, M_r 160.22. At 17.5 °C, a saturated aqueous solution contains 41.1 wt % of the salt. It has similar properties to the sodium salt. It is used in the food industries where the presence of sodium is undesirable.

8.2. Esters of Benzoic Acid

Methyl benzoate [*93-58-3*], $C_8H_8O_2$, M_r 136.15, *mp* −12.21 °C, *bp* 199.5 °C, colorless liquid with a pleasant odor, n_D^{20} 1.516, d_4^{15}

1.09334, solubility in water at 20 °C 0.15 wt %, miscible with common organic solvents, vapor pressure 0.133 kPa at 39 °C, 13.3 kPa at 130.8 °C, flash point 83 °C, ignition point 505 °C, heat of combustion 3950 kJ/mol. Methyl benzoate forms an azeotrope (20.8 wt % ester) with water, *bp* 99.08 °C. Stable against oxidation and easily saponified by strong base.

Methyl benzoate is prepared by reacting methanol with benzoic acid. It is used as a perfume and as an intermediate in the production of other benzoic esters.

***n*-Butyl benzoate** [*136-60-7*], $C_{11}H_{14}O_2$, M_r 178.23, *mp* −22.4 °C, *bp* 250.3 °C, d_4^{20} 1.0061, flash point 111 °C; autoignition temperature 440 ± 5 °C, heat of combustion 5828 kJ/mol. Butyl benzoate forms an azeotrope (6 wt % ester) with water, *bp* 99.8 °C. Butyl benzoate is widely used as a dye carrier in polyester fibers.

Glycol Benzoates. *Diethylene glycol dibenzoate* [*120-55-8*], $C_{18}H_{18}O_5$, M_r 314.34, *mp* −28 °C, *bp* 240 °C at 0.7 kPa, d_4^{20} 1.178, n_D^{25} 1.5424, flash point 232 °C.

Propylene glycol dibenzoate [*19224-26-1*], $C_{17}H_{16}O_4$, M_r 284.31, *fp* -3 °C, *bp* 232 °C at 1.6 kPa.

Dipropylene glycol dibenzoate [*27138-31-4*], $C_{20}H_{22}O_5$, M_r 342.40, *mp* −40 °C, *bp* 232 °C at 0.7 kPa, d_4^{20} 1.129, n_D^{25} 1.5282, flash point 212 °C.

Glycerine tribenzoate [*614-33-5*].

The glycol benzoates are used as plasticizers in poly(vinyl chloride), poly(vinyl acetate), and polacrylate resins.

Phenyl benzoate [*93-99-2*], $C_{13}H_{10}O_2$, M_r 198.21; monoclinic crystals, *mp* 70 °C; *bp* 314 °C, d_4^{20} 1.235, slightly soluble in water, soluble in hot ethanol. Phenyl benzoate is an intermediate in the production of phenol from benzoic acid.

Benzyl benzoate [*120-51-4*], $C_{14}H_{12}O_2$, M_r 212.25, leaflets or oily liquid, aromatic odor, *mp* 21 °C, *bp* 323 – 324 °C, *bp* 189 – 191 °C at 2.13 kPa, *bp* 156 °C at 0.6 kPa. It decomposes above 200 °C and is slightly volatile in steam and insoluble in water.

Benzyl benzoate is produced by esterifying dry sodium benzoate with benzyl chloride [13] or by the dimerization of benzaldehyde [14]. It is also obtained from the bottoms of benzoic acid rectification plants (see page 2).

Benzyl benzoate is used as a dye carrier. Other uses include solvents for cellulose derivatives, plasticizers, and fixatives in the perfume industry.

8.3. Benzoyl Chloride

Physical Properties. Benzoyl chloride [*98-88-4*], C_7H_5ClO, M_r 140.57, *mp* −0.6 °C, *bp* 198.3 °C, vapor pressure 0.133 kPa at 32.1 °C, 1.33 kPa at 73.8 °C, 13.3 kPa at 128 °C, colorless liquid with a penetrating odor, lachrymatory, fumes on exposure to air, n_D^{20} 1.55369, d_4^{20} 1.2113, decomposes in water and alcohols, miscible with benzene, flash point 102 °C, autoignition temperature 600 °C, heat of combustion 3282 kJ/mol.

Chemical Properties. Benzoyl chloride reacts violently with water to give benzoic acid and hydrochloric acid. If the amount of water is limited, benzoic anhydride is formed. Reaction with ammonia forms benzamide [*55-21-0*]. Benzoyl chloride is an excellent benzoylation reagent for compounds with active hydrogen, such as alcohols, phenols, or amines, under Friedel – Crafts and Schotten – Baumann reaction conditions.

Production. Benzoyl chloride is produced commercially from benzotrichloride [*98-07-7*], which is available by the chlorination of toluene. The chlorination reaction proceeds stepwise via benzyl chloride [*100-44-7*] and benzal chloride [*98-87-3*].

$$C_6H_5CH_3 + Cl_2 \longrightarrow C_6H_5CH_2Cl + HCl$$

$$C_6H_5CH_2Cl + Cl_2 \longrightarrow C_6H_5CHCl_2 + HCl$$

$$C_6H_5CHCl_2 + Cl_2 \longrightarrow C_6H_5CCl_3 + HCl$$

The benzotrichloride is then partially hydrolyzed with a limited amount of water.

$$C_6H_5CCl_3 + H_2O \rightarrow C_6H_5COCl + 2\,HCl$$

This series of reactions consumes a large quantity of chlorine. An alternative process uses benzoic acid.

$$C_6H_5CCl_3 + C_6H_5COOH \longrightarrow 2\,C_6H_5COCl + HCl$$

Phosphorous chlorides or iron chlorides are used as catalysts at the solvolytic stage of both processes. Benzoyl chloride can also be made by the reaction of benzoic acid with phosgene, thionyl chloride, or phosphorous pentachloride [15]; chlorination of benzaldehyde, benzyl alcohol, or benzyl benzoate [16]; and direct reaction of benzene with phosgene or the combination of carbon monoxide and chlorine [17].

The benzoyl chloride prepared by the conventional processes may contain trace amounts of chlorobenzoyl chloride, but the product obtained by the reaction of molten phthalic anhydride with hydrogen chloride at 200 °C is free of such impurities.

Uses. Benzoyl chloride is used in making organic peroxides. Benzoyl peroxide [*94-36-0*] and *tert*-butyl peroxybenzoate [*614-45-9*] are widely used as polymerization initiators and as curing agents in the polymer industries. Other uses include intermediates for rubber additives, UV absorbers, dyes, insecticides, and pharmaceuticals.

Specifications. Assay 99.5 % min., *sp* above − 0.5 °C, benzal chloride 0.01 % max., trichlorobenzal chloride 0.01 % max., color (APHA) 10 max.

8.4. Benzonitrile

Physical Properties. Benzonitrile [*100-47-0*], C_7H_5N, M_r 103.12, *mp* −13.8 °C, *bp* 191.1 °C, vapor pressure 1.37 kPa at 70 °C, 2.27 kPa at 79 °C, 4.53 kPa at 94 °C, colorless liquid, almond oil odor, n_D^{15} 1.53056, d_4^{20} 1.0052, solubility in water 1 wt % at 100 °C, miscible with common organic solvents, heat capacity c_p 1.840 J g^{-1} K^{-1}, heat of vaporization 46 kJ/mol, heat of formation −155 kJ/mol, heat of combustion 3063 kJ/mol, flash point 70 °C.

Chemical Properties. Benzonitrile is stable in air and light. In the presence of a strong base or acid, benzonitrile is hydrolyzed to benzoic acid and ammonia (via benzamide). Benzylamine [*100-46-9*] and dibenzylamine can be prepared from benzonitrile by hydrogenation.

Benzonitrile is a powerful solvent. At room temperature it dissolves poly(vinyl chloride), poly(vinyl acetate) resins, polystyrene, polymethacrylate, and nitrocellulose. It does not dissolve polyethylene, polyamide, poly(vinyl alcohol), or fluoropolymers.

Production. Benzonitrile is produced by the high-temperature vapor-phase oxidation of toluene in the presence of ammonia. The traditional catalysts are vanadium and molybdenum, but they suffers from low selectivity and serious decomposition of ammonia [18]. A tungsten – manganese complex catalyst shows better performance [19]. With the latter catalyst, the reaction is carried out in a fixed-bed reactor at 450 °C. The concentration of toluene in the feed stream is 1.5 vol %. The ratio of ammonia to toluene is 4 : 1 with a contact time of 2.4 s. The reaction converts 97 % of the toluene and 30 % of the ammonia. The selectivity is 87.4 %. The crude product, which contains ammonia, hydrogen cyanide, toluene, and high-boilers, is purified by distillation.

Benzonitrile has also been produced from benzoic acid and ammonia in the vapor phase at 400 – 410 °C over alumina [20] or in the liquid phase at 225 – 245 °C [21]. Other methods include liquid-phase ammoxidation of toluene in the presence of a cobalt or manganese bromide catalyst [22], dehydrogenative condensation of benzyl alcohol or benzaldehyde with ammonia, the high-temperature reaction of toluene with nitrous oxide, and cyanation of benzene with cyanogen chloride or dicyanogen.

Uses. The major derivative of benzonitrile is benzoguanamine [*91-76-9*], which is made by the reaction of benzonitrile with dicyanodiamide in the presence of a strong base [23]. Benzoguanamine reacts with formaldehyde and/or alcohols to afford a heat-resistant thermosetting resin.

8.5. Alkyl and Acyl Analogues

Methylbenzoic acids, $C_8H_8O_2$, M_r 136.15.

The physical properties of the methylbenzoic acids are presented in Table 6.

Table 6. Physical properties of the methylbenzoic acids

	mp, °C	*bp*, °C	Density, g/cm^3	Solubility in water 25 °C, wt %	Solubility in water 100 °C, wt %	Solubility in chlorobenzene, 31.8 °C, mol %	Solubility in corresponding xylene, 14 °C, wt %	Heat capacity, J g^{-1} K^{-1}	Heat of fusion, kJ/mol	Heat of combustion, kJ/mol	Heat of formation, kJ/mol	Flash point, °C	Ignition point, °C
2-Methylbenzoic acid	105	259	1.062 at 115 °C	0.118		12.73	6.64	1.285	20.2	3887	−416.5	148 ± 3	495 ± 5
3-Methylbenzoic acid	110	263	1.0543 at 112 °C	0.098		14.07	7.89	1.219	15.7	3886	−426.1	146	
4-Methylbenzoic acid	180	274	1.23 at 20 °C	0.034	1.16	1.76	1.45	1.204	22.7	3880	−429.2	181 ± 1	570 ± 5

The chemical properties of the methylbenzoic acids are similar to those of benzoic acid. They are better preservatives than benzoic acid, but their higher toxicity limits their application. Their alkali metal salts have better anticorrosion effects than benzoates.

2-Methylbenzoic acid [*118-90-1*], *o*-toluic acid, can be produced by the oxidation of *o*-xylene. It is an intermediate for agrochemicals, such as the fungicide mepronil and the herbicide londax.

3-Methylbenzoic acid [*99-04-7*], *m*-toluic acid, is predominantly used for the production of an insect repellant, *N*,*N'*-diethyl-*m*-toluamide [*134-62-3*]. Other uses are as an intermediate for poly(vinyl chloride) stabilizers and agrochemicals.

4-Methylbenzoic acid [*99-94-5*], *p*-toluic acid, is an intermediate in the production of terephthalic acid by oxidation of *p*-xylene. The methyl group of 4-methylbenzoic acid, unlike that of toluene or *p*-xylene, is resistant to oxidation by molecular oxygen with a conventional cobalt catalyst. The addition of bromide overcomes this difficulty. 4-Methylbenzoic acid can be prepared from *p*-xylene by vapor-phase oxidation or nitric acid oxidation. 4-Methylbenzoic acid is used for producing intermediates for antibiotic pharmaceuticals, and organic pigments.

4-*tert*-Butylbenzoic acid [*98-73-7*], $C_{11}H_{14}O_2$, M_r 178.23, crystals, *mp* 168.5–169 °C, dissociation constant 6.99×10^{-7} at 25 °C.

4-*tert*-Butylbenzoic acid is obtained by the liquid-phase air oxidation of 4-*tert*-butyltoluene (prepared from toluene and isobutylene) [24]. It is a modifier for alkyd resins and a polymerization regulator for polyesters. It is also used as an additive in cutting oils and as a corrosion inhibitor. It is hydrogenated to afford the fragrance 4-*tert*-butylbenzaldehyde [*939-97-9*].

2-Benzoylbenzoic acid [*85-52-9*], benzophenone-*o*-carboxylic acid, $C_{14}H_{10}O_3$, M_r 226.23, *mp* 127 °C.

The Friedel–Crafts acylation of benzene with phthalic anhydride is generally used to produce 2-benzoylbenzoic acid. 2-Benzoylbenzoic

Table 7. Physical properties of the monochlorobenzoic acids

	2-Chlorobenzoic acid	3-Chlorobenzoic acid	4-Chlorobenzoic acid
mp, °C	141	154	241.5
bp, °C	279	275	276
Density, g/cm^3	1.544 at 20 °C	1.517 at 25 °C	1.571 at 24 °C
Solubility, wt %			
in water at 25 °C	0.21	0.045	0.0077
at 100 °C	4.02	0.5	0.11
in benzene at 14 – 16 °C	0.92	0.66	0.017
Heat capacity, J g^{-1}K^{-1}	1.042	1.042	1.073
Heat of fusion, kJ/mol	25.68	23.85	32.25
Heat of combustion, kJ/mol	3084	3069	3066
Dissociation constant at 25 °C, K_a	1.2×10^{-3}	1.5×10^{-4}	1.05×10^{-4}
Flash point, °C	173 ± 6	150	238 ± 1

Table 8. Physical properties of the aminobenzoic acids

	2-Aminobenzoic acid	3-Aminobenzoic acid	4-Aminobenzoic acid
mp, °C	146.1	177.9	188.0
Crystal structure	rhombic	monocline	monoclinic
Density			
at 20 °C, g/cm^3	1.367	1.5105	
Solubility, wt %			
in water	0.35 at 14 °C	0.59 at 15 °C	0.59 at 25 °C, 1.1 at 100 °C
in 90 % ethanol	10.7 at 9.6 °C	2.2 at 9.6 °C	11.3 at 9.6 °C
in benzene	0.18 at 11.4 °C		0.06 at 11 °C
Dissociation constant at 25 °C			
K_{acid}	1.07×10^{-5}	1.67×10^{-5}	1.2×10^{-5}
K_{base}	1.1×10^{-12}	7.9×10^{-12}	1.7×10^{-12}

Table 9. Physical properties of the mononitrobenzoic acids

	2-Nitrobenzoic acid	3-Nitrobenzoic acid	4-Nitrobenzoic acid
mp, °C	148	142	240
Density, g/cm^3	1.575	1.494	1.610
Solubility, g per 100 mL of solvent at 10 °C			
in water	0.75 (at 25 °C)	0.24 (at 15 °C)	0.02 (at 15 °C)
in methanol	42.72	47.34	9.6
in benzene	0.294	0.795	0.017 (at 12.5 °C)
Heat capacity, J g^{-1}K^{-1}	1.149	1.035	1.077
Heat of fusion, kJ/mol	28.02	29.33	37.02
Heat of combustion, kJ/mol	3078	3053	3881
Dissociation constant, K_a	6.1×10^{-3} at 18 °C	3.48×10^{-4} at 25 °C	3.93×10^{-4} at 25 °C
Decarboxylation temperature, °C	180	238	> 240

acid is principally used for the production of photosensitizers.

8.6. Chlorobenzoic Acids

The physical properties of the monochlorobenzoic acids are listed in Table 7.

2-Chlorobenzoic acid [*118-91-2*], $C_7H_5ClO_2$, M_r 156.57.

2-Chlorobenzoic acid is the strongest acid among the three isomers. Its chlorine atom is easily exchangeable. It is decarboxylated in aqueous solution at 300 °C. With sodium sulfite it gives 2-sulfobenzoic acid. 2-Aminobenzoic acid is formed with ammonia. 2-Chlorobenzoic acid is produced by oxidation of 2-chlorotoluene [*95-49-8*] with molecular oxygen [25] or by the hydrolysis of α,α,α,2-tetrachlorotoluene [*2136-89-2*]. 2-Chlorobenzoic acid is used

for the manufacture of sweeteners, anti-inflammatory pharmaceuticals, and fungicides.

3-Chlorobenzoic acid [*535-80-8*], $C_7H_5ClO_2$. The reactivity of the chlorine atom of 3-chlorobenzoic acid is sluggish compared with the other isomers. Decarboxylation at higher temperatures is insignificant, and it resists oxidation. 3-Chlorobenzoic acid is made by the oxidation of 3-chlorotoluene [*108-41-8*] or by the hydrolysis of 3-chlorobenzoyl chloride [*618-46-2*], which is derived from benzoyl chloride. 3-Chlorobenzoic acid is used for making diphenyl ether type herbicides.

4-Chlorobenzoic acid [*74-11-3*], $C_7H_5ClO_2$ is produced by the oxidation of 4-chlorotoluene [*106-43-4*].

2,4-Dichlorobenzoic acid [*50-84-0*], $C_7H_4Cl_2O_2$, M_r 191.01, *mp* 164 °C, needles, slightly soluble in water.

2,4-Dichlorobenzoic acid is an important pharmaceutical chemical used in diuretics, antirheumatics, and antimalarials. It is produced by the oxidation of 2,4-dichlorotoluene [*95-73-8*], which is obtained by the direct chlorination of toluene.

2,5-Dichlorobenzoic acid [*50-79-3*], $C_7H_4Cl_2O_2$, *mp* 154.4 °C, *bp* 301 °C.

2,5-Dichlorobenzoic acid is made from benzoyl chloride by chlorination and hydrolysis. It is used as a starting material in the production of herbicides.

8.7. Aminobenzoic Acids

Aminobenzoic acids, $C_7H_7NO_2$, M_r 137.14.

The physical properties of aminobenzoic acids are presented in Table 8.

2-Aminobenzoic acid [*118-92-3*], anthranilic acid, was first obtained from indigo by alkali fusion. The methyl derivative is found in essential oils. It is an off-white crystalline powder of sweetish taste, sublimable, and freely soluble in hot water. It has an amphoteric character as a weak acid and a weak base. Therefore, it is soluble in strong acids and bases. It is easily decarboxylated to form aniline.

2-Aminobenzoic acid is usually made from phthalic anhydride via sodium phthalamate.

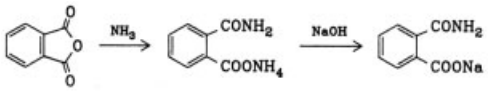

Oxidative decarboxylation (the Hofmann reaction) is carried out by the addition of sodium hypochlorite solution at 60 – 100 °C followed by acidification.

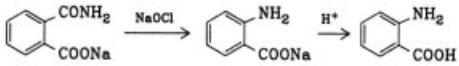

The product is purified with activated carbon and bisulfite. The overall yield is 88 %. Other synthetic routes include the amination of 2-chlorobenzoic acid and the reduction of 2-nitrobenzoic acid.

2-Aminobenzoic acid is called vitamin L_1; it enhances the milk production of cows. The major use of 2-aminobenzoic acid is as an intermediate for dyes. As its occurrence suggests, it is an intermediate for indigo synthesis. Mordant Brown 40 and Vat Violet 13 are other derivatives. 4-Hydroxy-1-methyl-carbostyril [*1677-46-9*] is an important derivative used in making dyes and pigments. Its methyl and ethyl esters are used as fragrances for toiletries. In the pharmaceutical industry, it is used as an intermediate for tranquilizers and antiphlogistics.

3-Aminobenzoic acid [*99-05-8*] is prepared by the reduction of 3-nitrobenzoic acid and used in dye syntheses. Its derivative *N,N*-dimethyl-3-aminobenzoic acid [*99-64-9*] is also used in the production of dyes.

4-Aminobenzoic acid [*150-13-0*] is widely found in nature as a factor in vitamin B complex. Baker's yeast contains 5 or 6 mg/kg of 4-aminobenzoic acid, and brewer's yeast 10 – 100 mg/kg. It enhances the growth of various microorganisms, and it is essential to the anaerobic metabolism of some bacteria. It is known as bacterial vitamin H and is antagonistic to sulfonamide drugs. It darkens slightly on exposure to air or light. It is slightly soluble in cold water. The pH of a 0.5 wt % aqueous solution is 3.5. It is soluble in ethyl acetate and acetic acid but practically insoluble in petroleum ether.

Table 10. Acute toxicity data for benzoic acid and its derivatives (RTECS)

Compound	Species	Route		Acute toxicity
Benzoic acid	rat	oral		LD_{50} 1700 mg/kg
[*65-85-0*]	mouse	oral		LD_{50} 1940 mg/kg
	rat	inhalation		$LC_{50} > 26$ mg/m^3
	rat	i.p.		LD_{50} 1600 mg/kg
	mouse	i.p.		LD_{50} 1460 mg/kg
	human	skin		TDL_0 6 mg/kg
	rabbit	s.c.		LDL_0 2 g/kg
	rabbit		irritation (skin)	mild
	rabbit		irritation (eye)	severe
Sodium benzoate	rat	oral		LD_{50} 4070 mg/kg
[*532-32-1*]	mouse	oral		LD_{50} 1600 mg/kg
	dog	oral		LD_{50} 2 g/kg
	mouse	i.v.		LD_{50} 1440 mg/kg
	rat	s.c.		LD_{50} 2 g/kg
	mouse	i.m.		LD_{50} 2306 mg/kg
Methyl benzoate	rat	oral		LD_{50} 1177 mg/kg
[*93-58-3*]	mouse	oral		LD_{50} 3330 mg/kg
	rabbit		irritation (skin)	mild
	rabbit		irritation (eye)	mild
n-Butyl benzoate	rat	oral		LD_{50} 735 mg/kg
[*136-60-7*]	rabbit	skin		LD_{50} 4 g/kg
	rabbit		irritation (skin)	severe
	rabbit		irritation (eye)	mild
Diethylene glycol dibenzoate	rat	oral		LD_{50} 2830 mg/kg
[*120-55-8*]	rabbit	skin		LD_{50} 20 g/kg
	rabbit		irritation (skin)	mild
	rabbit		irritation (eye)	mild
Phenyl benzoate	mouse	oral		LD_{50} 1225 mg/kg
[*93-99-2*]				
Benzyl benzoate	rat	oral		LD_{50} 1700 µL/kg
[*120-51-4*]	mouse	oral		LD_{50} 1400 µL/kg
	dog	oral		$LD_{50} > 22.44$ g/kg
	mouse	i.p.		$LD_{50} > 500$ mg/kg
	rabbit	skin		LD_{50} 4 g/kg
Benzoyl chloride	rat	oral		LD_{50} 1900 mg/kg
[*98-88-4*]	rat	inhalation		LC_{50} 1870 mg/m^3
Benzonitrile	rat	oral		LDL_0 720 mg/kg
[*100-47-0*]	mouse	oral		LD_{50} 971 mg/kg
	rat	inhalation		LCL_0 950 ppm
	mouse	inhalation		LC_{50} 1800 mg/m^3
	mouse	i.p.		LD_{50} 400 mg/kg
	rabbit	s.c.		LDL_0 200 mg/kg
	rabbit	skin		LD_{50} 1250 mg/kg
	rabbit		irritation (skin)	moderate
Benzaldehyde	rat	oral		LD_{50} 1300 mg/kg
[*100-52-7*]	mouse	oral		LD_{50} 28 mg/kg
	rat	inhalation		$LC_{50} > 500$ mg/m^3
	rat	s.c.		LDL_0 5 g/kg
	mouse	i.p.		LD_{50} 9 mg/kg
	rabbit		irritation (skin)	moderate
2-Methylbenzoic acid	mouse	i.p.		LD_{50} 422 mg/kg
[*118-90-1*]				
3-Methylbenzoic acid	rat	oral		$LD_{50} > 5$ g/kg
[*99-04-7*]	mouse	oral		LD_{50} 1630 mg/kg

The conventional method of producing 4-aminobenzoic acid is the reduction of 4-nitrobenzoic acid. The Hofmann degradation reaction of terephthalic acid is another source of 4-aminobenzoic acid. Dimethyl terephthalate reacts with one equivalent of potassium hydroxide in methanol. The resultant salt is then amidated with liquid ammonia in dimethylformamide at 130 – 140 °C to give potassium terephthalamate. This then reacts with sodium hypochlorite solution at 50 – 92 °C. The overall yield is approximately 70 % [26].

Table 10. (Continued)

Compound	Species	Route		Acute toxicity
	mouse	i.p.		LD_{50} 562 mg/kg
4-Methylbenzoic acid	mouse	oral		LD_{50} 2340 mg/kg
[*99-94-5*]	rat	i.p.		LD_{50} 874 mg/kg
	mouse	i.p.		LD_{50} 916 mg/kg
4-*tert*-butylbenzoic acid	rat	oral		LD_{50} 700 mg/kg
[*98-73-7*]	rat	inhalation		$LC_{50} > 1900$ mg/m^3
2-Chlorobenzoic acid	rat	oral		$LD_{50} > 500$ mg/m^3
[*118-91-2*]	rat	i.p.		LD_{50} 2350 mg/kg
	rabbit		irritation (skin)	mild
	rabbit		irritation (eye)	moderate
3-Chlorobenzoic acid	rat	i.p.		LD_{50} 750 mg/kg
[*535-80-8*]				
4-Chlorobenzoic acid	rat	oral		LD_{50} 1170 mg/kg
[*74-11-3*]	mouse	oral		LD_{50} 1170 mg/kg
	rat	i.p.		LD_{50} 1 g/kg
2,4-Dichlorbenzoic acid	mouse	oral		LD_{50} 830 mg/kg
[*50-84-0*]	mouse	s.c.		LD_{50} 1200 mg/kg
2,5-Dichlorobenzoic acid	mouse	i.p.		LD_{50} 237 mg/kg
[*50-79-3*]	mouse	s.c.		LD_{50} 1200 mg/kg
2-Aminobenzoic acid	rat	oral		LD_{50} 5410 mg/kg
[*118-92-3*]	mouse	oral		LD_{50} 1400 mg/kg
	mouse	i.v.		$LD_{50} > 500$ mg/kg
	mouse	i.p.		LD_{50} 2500 mg/kg
3-Aminobenzoic acid	mouse	oral		LD_{50} 6300 mg/kg
[*99-05-8*]	mouse	i.p.		LD_{50} 500 mg/kg
4-Aminobenzoic acid	rat	oral		$LD_{50} > 6$ g/kg
[*150-13-0*]	mouse	oral		LD_{50} 2850 mg/kg
	dog	oral		LD_{50} 1 g/kg
	rabbit	i.v.		LD_{50} 2 g/kg
	rat	i.p.		$LD_{50} > 3450$ mg/kg
3-Nitrobenzoic acid	mouse	i.v.		LD_{50} 640 mg/kg
[*121-92-6*]	mouse	i.p.		LD_{50} 610 mg/kg
4-Nitrobenzoic acid	rat	oral		LD_{50} 1960 mg/kg
[*62-23-7*]	mouse	i.v.		LD_{50} 770 mg/kg
	mouse	i.p.		LD_{50} 880 mg/kg
	rabbit		irritation (eye)	moderate
Hexahydrobenzoic acid	rat	oral		LD_{50} 3265 mg/kg
[*98-89-5*]				

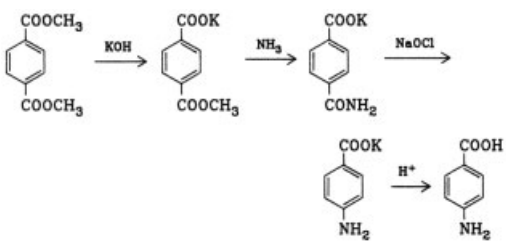

A two-stage Hofmann reaction has been developed. The yield is 93 % [27].

4-Aminobenzoic acid is used principally in the pharmaceuticals industry. Other uses are as a cross-linking agent for polyurethane resins, dyes, and feedstock additives. Ethyl 4-aminobenzoate [*94-09-7*], anesthesin, *mp* 90 – 91 °C is a local anesthetic of low toxicity, irritation, and persistency. It is used especially by dentists.

Folic acid [*59-30-3*], another derivative of 4-aminobenzoic acid, is a hematopoietic vitamin.

8.8. Nitrobenzoic Acids

The physical properties of the mononitrobenzoic acids are given in Table 9.

2-Nitrobenzoic acid [*27178-83-2*], $C_7H_5NO_4$, M_r 167.12, can be prepared by the oxidation of 2-nitrotoluene [*88-72-2*] with nitric acid.

3-Nitrobenzoic acid [*121-92-6*] is prepared by nitrating benzoic acid at low temperature. Approximately 20 % of the 2-nitro isomer and 1.5 % of the 4-nitro isomer are co-produced. The 3-Nitrobenzoic acid can be purified by recrystallization of the sodium salt. The oxidation of 3-nitrobenzaldehyde [*99-61-6*], prepared by

the controlled oxidation of benzaldehyde, gives a higher yield. 3-Nitrobenzoic acid is an intermediate in the preparation of 3-aminobenzoic acid and azo dyes. Its derivative 4-chloro-3-nitrobenzoic acid [*96-99-1*] is an intermediate in the production of dyes.

4-Nitrobenzoic acid [*62-23-7*] is produced commercially by the oxidation of 4-nitrotoluene [*99-99-0*] with molecular oxygen. Oxidation with 15 % nitric acid at 175 °C produces the acid in 88.5 % yield. An interesting method involves the nitration and subsequent oxidation of polystyrene [28]. This method uses the steric hindrance of the polymer chain to improve the *para* to *ortho* ratio of the product.

4-Nitrobenzoic acid is an intermediate for 4-aminobenzoic acid. 4-Nitrobenzoyl chloride [*122-04-3*], mp 75 °C, is a starting material for procaine hydrochloride and folic acid.

8.9. 3-Sulfobenzoic Acid

3-Sulfobenzoic acid [*121-53-9*], *mp* 141 °C, *mp* of dihydrate 98 °C, hygroscopic, soluble in water and ethanol. It is produced by the direct sulfonation of benzoic acid with oleum in ca. 85 % yield. 3-Sulfobenzoic acid is chiefly used for the production of 3-hydroxybenzoic acid [*99-06-9*] by alkali fusion.

8.10. Hexahydrobenzoic Acid

Hexahydrobenzoic acid [*98-89-5*], cyclohexanecarboxylic acid, $C_7H_{12}O_2$, M_r 128.17, *mp* 30 – 31 °C, *bp* 232 – 234 °C, *bp* 150.7 °C at 6.53 kPa, *bp* 120 – 121 °C at 1.87 kPa, d_4^{20} 1.0274, slightly soluble in water, soluble in ethanol and petroleum ether.

Hexahydrobenzoic acid reacts with nitrosylsulfuric acid to produce caprolactam. This is the basic reaction of the toluene-based caprolactam process (→ Caprolactam). Hexahydrobenzoic acid can be oxidized by molecular oxygen at 220 °C in the presence of a copper catalyst [29]. The main products are cyclohexene and cyclohexanone. It can be dehydrocarbonylated in the vapor phase over a heteropoly acid catalyst to yield cyclohexene [30]. Hexahydrobenzoic acid is produced in quantitative yield by the hydrogenation of benzoic acid over noble metal catalysts at 150 – 170 °C, at elevated pressure.

9. Toxicology

Acute toxicity data for benzoic acid and its derivatives are given in Table 10.

Benzoic acid is not accumulated in the human body. After administration it reacts with glycine to form hippuric acid, which is excreted in the urine. The FAO/WHO guideline for the upper limit of daily intake of benzoic acid is 5 – 10 mg/kg. The reproductive toxicity test for rats showed no abnormality.

Benzoic acid has an irritating effect on human mucous membranes. When dust formation is anticipated, protections should be arranged.

Benzoyl chloride is significantly toxic, like other acid chlorides. It irritates mucous membranes strongly (for further data on toxicology, see → Chlorinated Hydrocarbons, Chap. 10.2.6.).

Benzonitrile is less toxic than aliphatic nitriles. It is, nevertheless, absorbed through the skin and may cause convulsions of tissue and paralysis of nerves.

The esters of 4-aminobenzoic acid have a local anesthetic effect. After absorption they stimulate the central nervous system. At the workplace, some of these derivatives may have an allergenic and skin hypersensitizing effect.

4-Nitrobenzoic acid has a significant bactericidal action against *Staphylococci* and *Streptococci*. It is reported to be slightly mutagenic.

10. References

1. W. W. Kaeding, *J. Org. Chem.* **26** (1961) 3144.
2. W. W. Kaeding, R. O. Lindblom, R. G. Temple, *Ind. Eng. Chem.* **53** (1961) no. 10, 805.
3. American Cyanamid, DE 2 258 227, 1973.
4. C. R. Kinney, D. P. Langlois, *J. Am. Chem. Soc.* **53** (1931) 2189.
5. Mitsubishi Chem. Ind., JP 47-44215, 1972.
6. J. T. Bornwater, A. F. Holleman, *Recl. Trav. Chim. Pays-Bas* **31** (1922) 221.
7. W. J. LeNoble, G. W. Wheland, *J. Am. Chem. Soc.* **80** (1958) 5397.
8. BIOS 1786; BIOS (Misc.) 112.

9. N. Ohta, T. Tezuka, *Kogyo Kagaku Zasshi* **59** (1956) 71.
10. Mitsubishi Chem. Ind., JP-Kokai 56-39045, 1981.
11. Bayer, JP-Kokai 51-88934, 1976.
12. Y. Kamiya, *Adv. Chem. Ser.* **76** (1968) 193.
13. I. P. Tharp, H. A. Nottorf, *Ind. Eng. Chem.* **39** (1947) 1300.
14. *Org. Synth.* **2** (1922) 5.
15. SU 56 693, 1936 (Uwarov, Stepanov).
16. Chemische Werke Witten, DE 1 070 616, 1956.
17. ICI, GB 987 516, 1965.
18. Allied Chemical, US 2 486 934, 1949; US 2 499 055, 1950; US 2 828 325, 1958.
19. T. Ohara, Z. Iwao, M. Ninomiya, Y. Nakagawa, *Yuki Gosei Kagaku Kyokaishi* **26** (1968) 213.
20. BIOS **986**[2]**,** no. 209, 417.
21. Rohm & Haas, US 2 770 641, 1954.
22. G. D. Shik et al., *Neftekhimiya* **21** (1981) 133.
23. *Org. Synth.* **33** (1953) 13.
24. G. W. Hearne et al., *Ind. Eng. Chem.* **47** (1955) 2311.
25. Toray Industries, JP-Kokai 57-175143, 1982.
26. C. S. Rondestvedt Jr., *J. Org. Chem.* **42** (1977) 3118.
27. Akzo, DE 2313580, 1973.
28. C. S. Rondestvedt, Jr., J. R. Jeffrey, J. E. Miller, *Ind. Eng. Chem. Prod. Res. Dev.* **16** (1977) no. 4, 309.
29. J. A. Bigot, P. L. Kerkhoffs, *Recl. Trav. Chim. Pays-Bas* **82** (1963) 677.
30. Mitsubishi Chem. Ind., JP 55-37532, 1980; JP 55-50928, 1980.
31. T. Yokoyama, T. Setoyama, N. Fujita, M. Nakajima, T. Maki, K. Fujii, *Appl. Catal. A* **88** (1992) 149.
32. Bayer, DE 3 420 111, 1984 (W. Schulte-Huermann). Stamicarbon, EP-A 453 022, 1990 (U. F. Kragten, M. K. Frohn-Schloesser).
33. Mitsubishi Chem. Ind., JP-Kokai 4-69357, 1990 (K. Fujii, M. Nakamura).
34. Stamicarbon, EP-A 183 318, 183 319, 1984 (J. J. P. M. Goorden, A. J. F. Simons, L. A. L. Kleintjens).
35. Stamicarbon, EP-A 188 298, 1985 (S. M. P. Mutsers, M. H. Willems, W. P. Wolvers).
36. Nippon Kokan, JP-Kokai 5-255 181, 1991 (I. Osada, A. Imai, A. Miki).
37. European Parliament Council, Official Journal, no. L061 (1995) 0001 – 0040.
38. Iowa State University, material safety data sheets, see http://avogadro.chem.iastate.edu/msds/msds.html
39. Safety Information on the Internet, see http://siri.org/index.html
40. C. Barron: ″Benzoic acid″ in *SRI International Chemical Economics Handbook*, 1995, **618**, 8000A.
41. 21 CFR 184. 1733 (Report no. 88, US Dept. of Agriculture).
42. *Chemical Market Reporter* Sept. 29, 1997.

Benzoin → Ketones
Benzophenone → Ketones

Benzoquinone

K. Thomas Finley, State University of New York, Brockport, New York, United States

1. **Introduction** 1
2. **Physical Properties** 2
3. **Chemical Properties** 2
4. **Production of 1,4-Benzoquinone** . . . 3
5. **Uses** 3
6. **Toxicology** 4
7. **References** 4

1. Introduction

In 1838 the parent compound 1,4-benzoquinone (**1**) [*106-51-4*] was prepared by oxidation of quinic acid [*36413-60-2*] with maganese dioxide [1]. More than half a century later its much less stable isomer, 1,2-benzoquinone (**2**) [*583-63-1*] was obtained by silver ion oxidation of catechol [*120-80-9*] [2]. Since then these colorful and reactive compounds have occupied a central place in practical and theoretical studies, which are the subject of two extensive reviews [3, 4]. Other, more specialized reviews are given in [5 – 8]. The chemistry of the quinones includes a vast array of specific reactions and has been explored in great detail, including addition and cycloaddition reactions, photochemistry, biochemistry, oxidative chemistry, and electrochemistry [9 – 19].

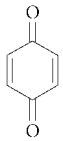

1

2

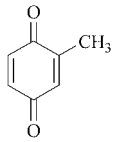

3

While there are many different quinonoid structures by far the most common are **1** and **2**. This article will be restricted to the physical, chemical, industrial, and toxicological characteristics of these compounds and their simple derivatives. The general references provide coverage of many other structural types.

1,4-Benzoquinones can be substituted by a wide variety of alkyl, aryl, halo, oxygen, nitrogen, and sulfur groups. They are usually named as derivatives of **1**, e.g., 2-methyl-1,4-benzoquinone (**3**) [*553-97-9*]. The CAS nomenclature of the quinones, e.g., 2,5-cyclohexadiene-1,4-dione for **1**, is seldom used. Trivial names like *p*-benzoquinone and *p*-tolyquinone (**3**) exist and are used occasionally. Two very useful oxidizing agents are universally known by a trivial name and an acronym: *p*-chloranil [*118-75-2*] (**4**) (2,3,5,6-tetrachloro-1,4- benzoquinone) and DDQ [*84-58-2*] (**5**) (2,3-dichloro-5,6-dicyano-1,4-benzoquinone).

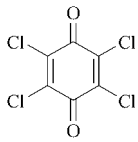

4

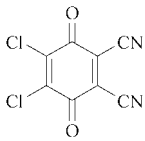

5

This lack of widespread concensus concerning nomenclature for the quinones is well illustrated by the latest *Aldrich Catalog Handbook of Fine Chemicals* [20] where one finds: "DDQ, see 2,3-dichloro-5,6-dicyano-1,4-benzoquinone", but "2,5-dichloro-3,6-dihydroxy-1,4-benzoquinone, see chloranilic acid" (**6**), [*87-88-7*]. Thymoquinone (**7**) is still the principal entry for 2-isopropyl-5-methyl-1,4-benzoquinone [*490-91-5*].

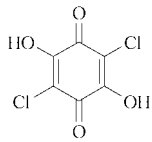

6

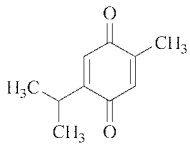

7

A few compounds with no apparent relationship to quinones continue to be misnamed, e.g., camphorquinone [*10373-78-1*] (2,3-camphandione).

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_571.pub2

2. Physical Properties

1,4-Benzoquinone (**1**), $C_6H_4O_2$, M_r 108.10, *mp* 113 °C, d_4^{20} 1.318, p_v^{20} 12 Pa, is soluble in most oxygenated organic solvents (ether, alcohol, acetone), and hot ligroin, slightly soluble in pentane, and insoluble in water. Crystallization from alcohol or sublimation produces yellow monoclinic prisms.

1,2-Benzoquinone (**2**), $C_6H_4O_2$, M_r 108.10, *mp* 60 – 70 °C (decomp.), crystallizes in red plates or prisms. It is much less stable than **1**, e.g., it decomposes in water. Compound **2** is soluble in ether, acetone, and benzene, but insoluble in pentane.

2,3,5,6-Tetrachloro-1,4-benzoquinone (**4**), *p*-chloranil, $C_6Cl_4O_2$, M_r 245.88, *mp* 290 °C (sealed tube); and 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (**5**), DDQ, M_r 227.01, *mp* 213 – 216 °C are yellow to bright yellow in color and crystallize as monoclinic prisms and plates.

These and most other low molecular mass benzoquinones have high vapor pressures and pungent, irritating odors.

3. Chemical Properties

Many of the quinones found in nature are important as oxidants, and this property is well illustrated in metabolic processes [14, 15] and has found widespread synthetic applications [16, 17]. The cross-conjugated system of two α,β-unsaturated carbonyl groups creates versatile reactivity.

Michael addition has a long and productive history; similar radical and electrophilic reactions offer important alternative and often complementary synthetic routes [9 – 11]. The synthesis of natural products [21] and polycyclic aromatic compounds [22] shows quinones to be important dienophiles for Diels – Alder chemistry. When additions fail to achieve the desired result, substitution reactions are often the route of choice [9 – 11]. There is also a growing literature on the role of quinones in photochemical processes [12].

The addition of arylsulfinic acids to 1,4-benzoquinones has been studied in detail as a model for the more complicated cases of thiol and sulfite anion addition [23]. These other oxidation states of sulfur are interesting in biochemistry [24] and photography [25]. The protonated quinone is the reactive species, a fact that is supported by the product distribution observed with 2-methyl-1,4-benzoquinone (Eq. 1). While 2,5-addition (**8**) is favored by both resonance and molecular orbital arguments, the 2,6-product (**9**) is obtained in a 3:1 ratio.

It has also been shown that the product distribution in this reaction can be strongly influenced by changes in acidity and solvent mixture [26]. Furthermore, the use of ionic liquids shows great promise as "green" solvents [27] in this chemistry. When the acidity or solvent is varied, the relative yields of **8** and **9** can change dramatically [26]. Over the pH range 1 to 4.5 in aqueous buffers one observes the **9/8** ratio changing from 4:1 to 2:3. In a similar fashion this same ratio is 9:1 with absolute ethanol as the solvent and 2:3 when carried out in 50% aqueous acetic acid. This same reaction (Eq. 1) has been studied in ionic liquids [27]. The **9/8** ratio almost exactly reverses to 1:8.5. These interesting media are being promoted as much more environmentally friendly green solvents.

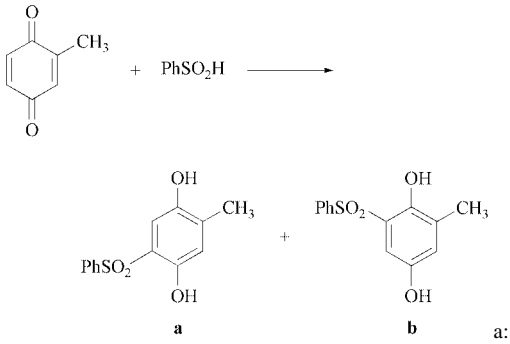

(**8**); b: (**9**)(Eq. 1)

With a single electron-withdrawing substituent 2,3-addition provides the major product [28]. This observation has been exploited as an important route to heterocyclic molecules as illustrated in Equation 2.

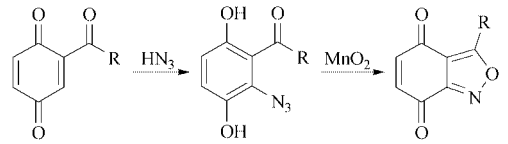

(Eq. 2)

The tripeptide glutathione (**10**) adds to 1,4-benzoquinone to form the thioether **11**. Subsequent cross-oxidation and further addition (Eq. 3) give a mixture of the three possible isomeric products [29]. This situation has only been reported once before, and then it also involved the formation of a bisthioether addition product [30].

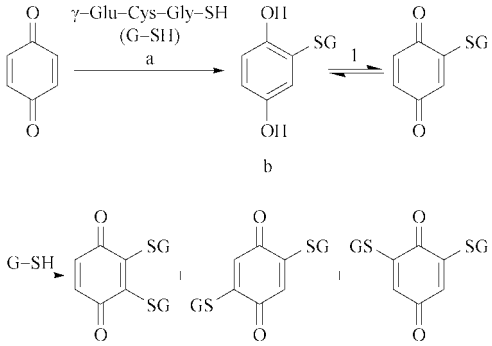

a: (**10**); b: (**11**)(Eq. 3)

A widely used example of electrophilic quinone chemistry may be described as a tandem reaction of acetic anhydride [31]. The result is the reductive addition of acetate together with esterification. This Thiele – Winter acetoxylation has been used productively, for example, in the synthesis of the antitumor, antibiotic fredericamycin A (Eq. 4) [32].

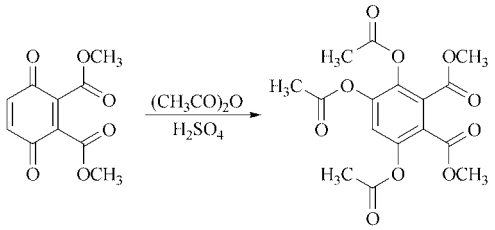

(Eq. 4)

The Meerwein arylation reaction has been reviewed [33] and support obtained for the importance of the semiquinone radical ion as an intermediate [34]. The synthesis of natural products has involved such radical arylations (Eq. 5) [35].

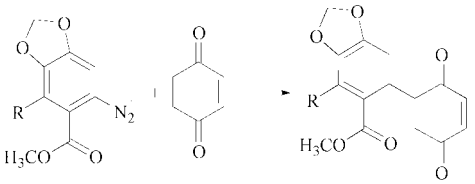

(Eq. 5)

4. Production of 1,4-Benzoquinone

A variety of quinones are available as fine organic chemicals and most of these are prepared by standard oxidative procedures. Quinones of higher oxidation potential such as chororanil (**4**) (742 mV) or 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (**5**) (ca. 1000 mV) are often the reagents of choice. There are many proposed syntheses of 1,4-benzoquinone, but the oxidation of aniline or phenol remains the practical route. Steam distillation produces a high yield and excellent purity [36]. Glucose is a possible raw material for the preparation of quinones (Eq. 6) [37].

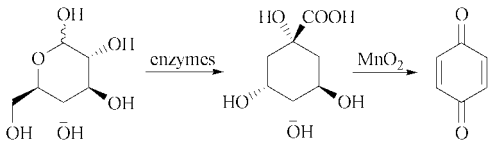

(Eq. 6)

The original preparation of the simplest quinones by oxidation continues to be the principal method for the commercial and laboratory synthesis of quinones [16, 17, 38]. The use of ceric ammonium nitrate for the oxidation of 1,4-dimethoxybenzene derivatives has been used frequently and with great success in the preparation of sensitive quinones. The synthesis of *blattellaquinone* (gentisyl quinone isovalerate) [*849762-24-9*], the sex pheromone of the German cockroach is an important example [39].

5. Uses

Important uses of 1,4-benzoquinone include applications in the inhibition of polymerization, determination of amino acids, and an additive in adhesive mixtures. Additional uses include

research laboratory applications, especially oxidative syntheses, analytical methods, and bactericidal substances. Hydroquinone is important in the photographic, dye, and leather industries and is often prepared from 1,4-benzoquinone. Naphthoquinones are important as vitamin K substitutes, fungicides, and antihemorragic agents.

6. Toxicology

Benzoquinone and many of its low molecular mass derivatives have a high vapor pressure, a penetrating odor, and present a health hazard [40 – 42]. Quinone vapor and the quinone resulting from hydroquinone dust in moist air produce eye injuries that can lead to appreciable loss of vision. The odor of 1,4-benzoquinone is definite at concentrations of 0.1 – 0.15 ppm, and at 0.5 ppm it becomes irritating. Both the MAK and TLV (TWA) are 0.1 ppm and 0.4 mg/m^3; the STEL is 0.3 ppm and 7 mg/m^3 [43, 44]. However, Russia has fixed the threshold limit in air at 0.01 ppm [45]. Both solid quinones and their solutions must be handled with care because they can cause severe local damage to the skin and mucous membranes [46, 47].

The limited carcinogenicity studies on 1,4-benzoquinone conducted with mice and rats do not permit definite classification, and it is listed as a questionable carcinogen. Results for *p*-chloranil are suggestive but inconclusive [48, 49].

7. References

Specific References

1. A. Woskressenski, *Justus Liebigs Ann. Chem.* **27** (1838) 268.
2. R. Willstätter, A. Pfannenstiel, *Chem. Ber.* **37** (1904) 4744.
3. S. Patai (ed.): *The Chemistry of Quinonoid Compounds*, Wiley-Interscience, New York 1974.
4. S. Patai, Z. Rappoport (eds.): *The Chemistry of Quinonoid Compounds*, vol. **2**, Wiley-Interscience, New York 1988.
5. S. M. Bruce: "Benzoquinones and Related Compounds," in S. Coffey (ed.): *Rodd's Chemistry of carbon Compounds*, 2nd ed., vol. **33**, Chap. 8, Elsevier, Amsterdam 1974.
6. T. Laird: "Quinones," in J. F. Stoddart (ed.): *Comprehensive Organic Chemistry, the Synthesis and Reactions of Organic Compounds*, vol. **1**, Chap. 5.5, Pergamon Press, Oxford 1979.
7. R. H. Thomson: *Naturally Occurring Quinones*, 3rd ed., Chapman & Hall, London 1987.
8. R. H. Thomson, in J. ApSimon (ed): *The Total Synthesis of Natural Products*, Vol. 8, John Wiley & Sons, Inc., New York 1992.
9. K. T. Finley: "The Addition and Substitution Chemistry of Quinones," Chap. 17, in [3].
10. K. T. Finley: "Quinones as Synthones", Chap. 11, in [4].
11. K. T. Finley, in S. Patai (ed.): *Supplement E: The Chemistry of Hydroxyl, Ether and Peroxide Groups*, vol. **2**, John Wiley & Sons, Inc., New York 1993, pp. 1027 – 1134.
12. J. M. Bruce: "Photochemistry of Quinones," Chap. 9, in [3].
13. K. Maruyama, A. Osuka, "Recent Advances of the Photochemistry of Quinones," Chap. 13, in [4].
14. R. Bentley, I. M. Campbell: "Biological Reactions of Quinones," Chap. 13, in [3].
15. H. Inouye, E. Leistner: "Biochemistry of Quinones," Chap. 22, in [4].
16. H.-D. Becker: "Quinones as Oxidants and Dehydrogenating Agents," Chap. 7, in [3].
17. H.-D. Becker, A. B. Turner: "Quinones as Oxidants and Dehydrogenating Agents," Chap. 23, in [4].
18. J. Q. Chambers: "Electrochemistry of Quinones," Chap. 14, in [3].
19. J. Q. Chambers: "Electrochemistry of Quinones," Chap. 12, in [4].
20. *Aldrich Catalog/Handbook of fine Chemicals*. Aldrich Chemical Co., Milwaukee WI, 2005 – 2006.
21. W. B. Manning, D. J. Wilbur, *J. Org. Chem.* **45** (1980) 733.
22. B. Simoneau, P. Brassard, *Tetrahedron* **42** (1986) 3767.
23. J. M. Bruce, P. Lloyd-Williams, *J. Chem. Soc. Perkin. Trans. 1* (1992) 2877.
24. K. G. Eckert, P. Eyer, J. Sonnenlichler, I. Zetl, *Xenobiotica* **20** (1990) 351.
25. M. P. Youngblood, *J. Org. Chem.* **51** (1986) 1981.
26. D. E. Allgeier, S. A. Herbrert, R. Nee, K. D. Schlecht, K. T. Finley, *J. Org. Chem.* **68** (2003) 4988.
27. J. S. Yadav, B. V. S. Reddy, T. Swamy, N. Ramireddy, *Synthesis* (2004) 1849.

28. R. Cassis, M. Fernández, R. Tapin, J. A. Valderrama, *Synth. Commun.* **17** (1987) 1077.
29. S. S. Lau, B. A. Hill, R. J. Highet, T. J. Monks, *Mol. Pharmacol.* **34** (1988) 829.
30. H. G. Wilgus, III, E. Frauenglass, E. T. Jones, R. F. Porter, J. W. Gates, Jr., *J. Org. Chem.* **29** (1964) 594.
31. J. F. W. McOmie, J. M. Blatchy in: *Organic Reactions*, vol. **19**, John Wiley & Sons, New York 1972, pp. 199 – 277.
32. K. A. Parker, D. M. Spero, K. A. Koziski, *J. Org. Chem.* **52** (1987) 183.
33. C. S. Rondestvedt, Jr. in: *Organic Reactions*, vol. **24**, John Wiley & Sons, New York 1976, pp. 225 – 259.
34. Y. A. Ilan, G. Czapski, D. Meisel, *Biochem. Biophys. Acta* **430** (1976) 209.
35. K. Krohn, A. Mondon, *Chem. Ber.* **109** (1976) 855.
36. W. H. Shearon, L. G. Davy, H. Von Bramer, *Ind. Eng. Chem.* **44** (1952) 1730.
37. K. M. Draths, T. L. Ward, J. W. Frost, *J. Amer. Chem. Soc.* **114** (1992) 9725.
38. J. Cason, "Synthesis of Benzoquinones by Oxidation," Chap. 6, in *Organic Reactions*, Vol. 4, John Wiley & Sons, Inc., New York 1948.
39. S. Nojima, C. Schal, F. X. Webster, R. G. Santangelo, W. L. Roelofs, *Science* **307** (2005) 114.
40. *OSHA Regulated Substances*, Noyes Data Corp., Park Ridge NJ, 1990, 1693.
41. *Material Safety Data Sheet Collection*, No. 637, Genium Publishing, Schenectady, NY 1997.
42. R. E. Lenga, K. L. Votoupal, *The Sigma-Aldrich Library of Regulatory and Safety Data*, vol. **I**, Aldrich Chemical Co., 1993, 491M. (Note typo in listing not 1491M).
43. DFG Deutsche Forschungsgemeinschaft (ed.): *Maximum Concentrations at the Workplace (MAK)*, Wiley-VCH Verlag, Weinheim 1997.
44. ACGIH (ed.): *Threshold Limit Values (TLV)*, ACGIH, Cincinnati, OH 1983.
45. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals*, 2nd ed., Van Nostrand, New York 1977, pp. 280, 350.
46. R. J. Lewis, Sr.: *Sax's Dangerous Properties of Industrial Materials*, 11th ed., vol. **III**, Wiley-Interscience, New York 2004, p. 2855.
47. R. J. Lewis, Sr.: *Carcinogenically Active Chemicals. A Reference Guide*, Van Nostrand-Reinhold, New York 1991, pp. 829 – 830.
48. *Toxic and Hazardous Industrial Chemicals Safety Manual*, The International Information Institute, Tokyo, 1989.
49. R. J. Lewis, Sr. (ed.). *Hazardous Chemical Desk Reference*, 5th ed., Wiley-Interscience, New York 2002, p. 1013.

Benzoquinone and Naphthoquinone Dyes

JOHN GRIFFITHS, Department of Colour Chemistry, The University, Leeds LS2 9JT, United Kingdom

1. **Introduction** 1
2. **Benzoquinone Dyes** 1
3. **1,4-Naphthoquinone Dyes** 2
3.1. **Simple 1,4-Naphthoquinones** 2
3.2. **Heteroannelated 1,4-Naphthoquinones** 3
4. **1,5-Naphthoquinones** 5
5. **References** 6

1. Introduction

A large number of naphthoquinone and benzoquinone dyes occur naturally as plant constituents, and in the past many of these have found use as colorants. The chemistry of synthetic analogues began with the discovery of *naphthazarin* (**1**) [*475-38-7*] by ROUSSIN in 1861, which was followed ten years later by the synthesis of *naphthoxidine* (the "naphthazarin intermediate") (**2**) [*27823-83-2*] by DE AGUIR and BAYER. Naphthazarin and its derivatives first found use as black chrome mordant dyes for wool and silk. In contrast to the naphthoquinones, synthetic benzoquinone dyes are few in number and are generally of the 2,5-diarylamino-1,4-benzoquinone type, suitable for application as vat dyes, e.g., *Helindon Yellow CG* (**3**) [*4370-55-2*], C.I. 56 005.

From about 1930 onwards, developments in the field of naphthoquinone dyes concentrated on the use of naphthazarin and naphthoxidine as intermediates for the preparation of violet, blue, and green acid and disperse dyes. Perhaps the most successful of these was *Artisil Blue GFL* (**4**), a blue dye for acetate fibers, introduced by Sandoz in 1946. The many developments in this area were reviewed by MERIAN in 1959 [1]. More recently there has been interest in the synthesis and color and constitution properties of simple colored naphthoquinones, stimulated by the fact that such dyes have similar tinctorial properties to the anthraquinones but a smaller molecular size. The naphthoquinones provide a useful alternative to the anthraquinones for certain specialized applications, e.g., as pleochroic dyes with improved solubility for liquid crystal display devices. As a result, research interest in these chromogens remains unabated, even though they have failed to make any major impact as textile dyes.

Details of naphthoquinone and benzoquinone dyes that have been marketed at one time or another may be found in the Colour Index [2], and natural coloring matters of this type have been discussed by THOMSON [3]. Several reviews of naphthoquinone and benzoquinone chemistry have appeared [4 – 8].

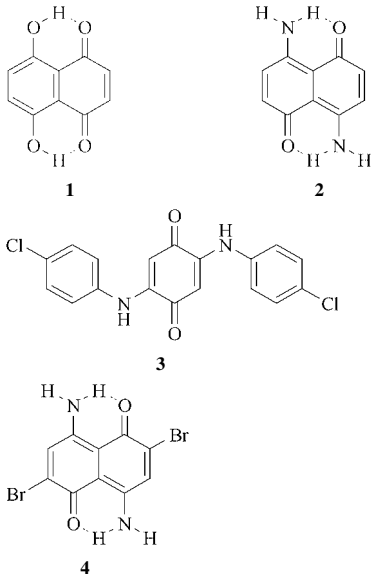

2. Benzoquinone Dyes

From the commercial point of view, only the 2,5-diarylamino-1,4-benzoquinones have been of

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a03_575.pub2

any significance, e.g., *Helindon Yellow CG* (**3**). These were used as vat dyes for wool and cotton, but are no longer in commercial production. Other examples of early dyes of this type are listed in the Colour Index under numbers 56 000 – 56 050. The inadequate color properties of the amino-substituted benzoquinone chromogen are exemplified by 2,5-bis(dimethylamino)-1,4-benzoquinone [*1521-02-4*], which absorbs at 530 nm in ethanol with an absorption coefficient of only ca. 500 L mol^{-1} cm^{-1}. This low intensity is characteristic of the 2,5-disubstitution pattern. Although other substitution patterns have higher intensities (e.g., the 2-amino derivatives have ε_{max} ca. 5000 L mol^{-1} cm^{-1}), dyes based on these have not been exploited.

More recently the 2,5-diamino-1,4-benzoquinone dyes have found applications other than as colorants. For example, they form stable complexes with *trans*-bis-(triphenylphosphine)phenylnickel(II) chloride, and these have proved to be useful catalysts for the polymerization of ethylene [9]. If 1,4-benzoquinone is condensed with a diamine under oxidizing conditions, polymeric 2,5-diaminobenzoquinones are formed, and these afford excellent anticorrosive coatings for steel [10].

One or both of the carbonyl groups of the 1,4-benzoquinone system have the potential to be replaced by imino groups, but the corresponding dyes have received little attention until recently. 1,4-Benzoquinone monoimines with electron-donating groups (e.g., alkylamino, alkoxy) in the 2- or 2,5-positions provide useful red permanent hair dyes [11]. The color properties of the 2,5-dialkylamino-1,4-benzoquinone diimines are more complex, and show an interesting dependence of color on pH [12]. Thus, the neutral species is yellow, the monoprotonated form red, and the diprotonated form blue.

3. 1,4-Naphthoquinone Dyes

3.1. Simple 1,4-Naphthoquinones

Color can be produced in the 1,4-naphthoquinone chromogen by introduction of amino and hydroxyl groups into the quinonoid ring (i.e., positions 2 and 3), into the benzenoid ring (more particularly positions 5 and 8), or into both rings. The spectroscopic effects of such substitutions have been investigated in detail [13 – 16].

2-Amino-1,4-naphthoquinones are generally yellow to red, with ε_{max} values of ca. 2000 – 5000 L mol^{-1} cm^{-1}, depending on the nature of the substituents attached to the amino group and to the 3-position of the ring. In general, substituents in the 3-position cause a decrease in intensity due to steric crowding, whereas *N*-aryl groups induce a higher intensity. Typically, 2-phenylamino-1,4-naphthoquinone [*6628-97-3*] has λ_{max} 472 nm, ε_{max} 4800 L mol^{-1} cm^{-1}. A now obsolete vat dye of this type is *Helindon Red CR* (**5**) [*3144-89-6*], C.I. 56 050. Fluoro- and trifluoromethylphenylamino analogues have been described for coloring wool and polyester [17].

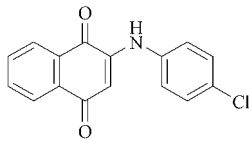

5

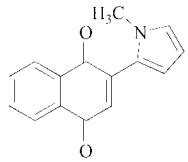

A

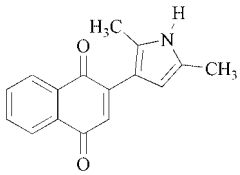

B

The 2-pyrrolyl derivatives **A** and **B**, formed respectively by condensation of 1,4-naphthoquinone with N-methylpyrrole [18, 19] and 2,5-dimethylpyrrole [20], can be regarded as vinylogues of the 2-amino-1,4-naphthoquinones. Such compounds have proved useful as direct dyes for hair [19, 20].

The more intensely colored 1,4-naphthoquinones with amino and/or hydroxyl groups in the 5- and 8-positions are analogous to the commercial anthraquinone dyes, but are

generally more bathochromic [14]. For example, 5-methylamino-1,4-naphthoquinone has λ_{max} 529 nm in cyclohexane, whereas 1-methylaminoanthraquinone has λ_{max} 495 nm. This is also true for the 5,8-disubstituted naphthoquinones, but this potential advantage has not proved of commercial significance and few such compounds have been considered as textile dyes. An exception is 5-amino-2,3-dichloro-8-hydroxy-1,4-naphthoquinone (**6**) [*68217-33-4*], which has been claimed as a violet disperse dye for acetate fibers [21]. Naphthazarin (**1**) is an example of a 5,8-disubstituted naphthoquinone of greater value as an intermediate than as a dyestuff.

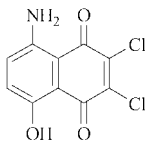

6

When auxochromes such as −NHR or −OH are present in both rings of the naphthoquinone system, the absorption spectrum consists of two overlapping bands, and this can result in dull colors, typically blue-greys. A 3-cyano group appears to counteract the dulling effect, and **7** [*79469-30-0*] has been described as a blue dye for polyester [22].

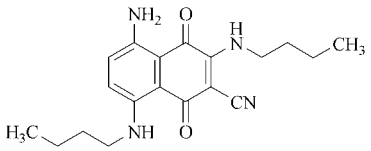

7

There is considerable interest in 1,4-naphthoquinones that absorb in the near-infrared, particularly for application in optical recording materials and related high-technology areas. The first simple near-infrared 1,4-naphthoquinone to be described was the 5-amino-2,3-dicyano-8-phenylaminoderivative (**8**, X = H) [*68217-41-4*], which despite its relatively small molecular size has λ_{max} at 770 nm in dichloromethane [14]. The derivative **8** (X = $C_{12}H_{25}$) has been produced as Langmuir – Blodgett films which can be written on with 830 nm laser light with high sensitivity [23]. Other derivatives related to **8** have been patented extensively for optical recording discs [24] and have found commercial use.

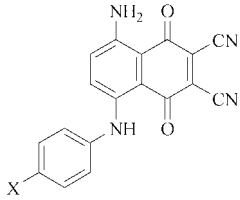

8

3.2. Heteroannelated 1,4-Naphthoquinones

A very large number of quinones of the general structure **9** are known, where X corresponds to a heterocyclic residue containing one or more rings, and the chemistry of these has been reviewed extensively [6 – 8]. If X consists of a benzenoid or azabenzenoid ring system, then such compounds are best regarded as anthraquinone analogues. In the case of other heterocyclic residues, they may be considered as 1,4-naphthoquinone derivatives.

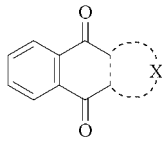

9

The simplest members of this group are those where X is a monocyclic residue, such as furan, thiophene, and pyrrole ring systems. Color may then be produced in the usual way by attaching auxochromes to the 5- and/or 8-positions. Few dyes of this type have been suggested for textile applications, presumably because their chemistry is generally more complex than that of comparable anthraquinones. It appears that the color properties of such dyes differ only slightly from those of related anthraquinone structures, as exemplified by studies of imidazole dyes of general structure **10** [25]. However, triazole dyes of formula **11**, where R^1 to R^3 are auxochromes, have been patented as disperse dyes for polyester [26].

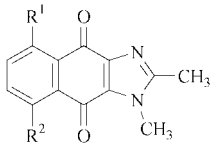

10

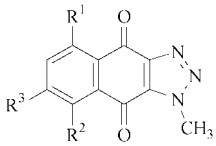

11

Of greater potential are dyes of general structure **9** in which X corresponds to a polycyclic heterocyclic system, and a large number of these have been examined as possible dyes and pigments. Perhaps the most closely investigated are the *brazanquinones* (**12**), which have been known for more than eighty years. They can be prepared readily by condensing 2,3-dichloro-1,4-naphthoquinone [*117-80-6*] with substituted phenols and naphthols. For example, arylamides of 2-hydroxy-3-naphthoic acid, which are commercially available as coupling components for azoic dye combinations, condense to give yellow and orange vat dyes (**13**) with good substantivity for cotton [27].

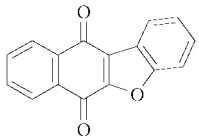

12

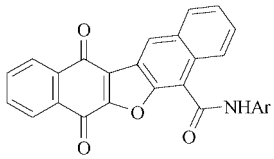

13

The brazanquinone system has also been used to provide disperse dyes by attaching auxochromes to the benzene ring of the naphthoquinone moiety. Thiophene equivalents of the brazanquinones have received much less attention. Disperse dyes of the type **14** proved to be rather more bathochromic than corresponding anthraquinones, but of lower intensity [28]. Pyrrole analogues have also been neglected, but it is interesting to note that the old dye *Helindon Yellow R* (**15**) [*6451-05-4*], C.I. 56 070, a vat dye for cotton, is the parent compound of this type.

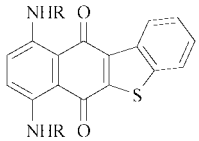

14

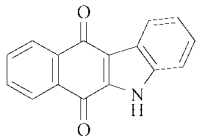

15

A different class of polycyclic naphthoquinone dyes is that based on the naphth[2,3-*b*]indolizine-6,11-dione system (**16**), which may be accessed readily by condensation of 2,3-dichloro-1,4-naphthoquinone with active methylene compounds in the presence of pyridine [29], or by reaction of 2-methoxy-3-pyridinio-1,4-naphthoquinone with an active methylene compound [30]. In **16** the bridgehead nitrogen atom acts as an effective auxochrome and thus orange to red colors are observed without further substitution. Derivatives **16** (R = amide group) are of particular value as vat dyes and pigments. Related isomeric heterocyclic structures have also attracted interest, e.g., **17**, a yellow disperse dye for polyester [31].

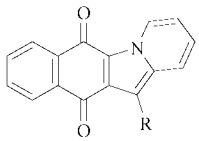

16

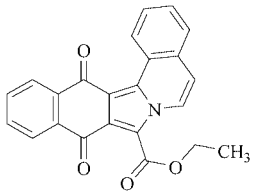

17

As with the simple 1,4-naphthoquinones, research activity in the heteroannelated compounds has been directed towards the development of dyes that absorb in the near-infrared. Noteworthy is the dithiadinaphtho-1,4-naphthoquinone system **18** (X = S), which can be prepared by condensing 2,3-dichloro-5,8-dihydroxy-1,4-naphthoquinone with 2-aminothiophenol [32]. The parent compound **18** (X = S) [*96692-25-0*] shows multiple absorption bands between 400 and 800 nm, with the most intense peak at 725 nm in chloroform (ε_{max} 15 200 L mol^{-1} cm^{-1}) [32]. Oxidation of the sulfur atoms to SO groups produces a large bathochromic shift, and **18** (X = SO) shows a peak at 827 nm (ε_{max} 17 600 L mol^{-1} cm^{-1}) in chloroform [32].

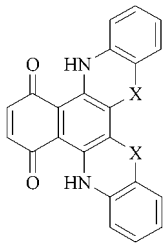

18

4. 1,5-Naphthoquinones

Dyes of this class have the general formula **19**, in which R^1 and R^2 may be hydrogen, alkyl, or aryl groups. Historically and structurally, *naphthoxidine* (**2**) may be regarded as the parent compound; it is a violet-blue compound possessing strong intramolecular hydrogen bonding and exhibiting solubility in organic solvents and in water. This substance and its *N*-substituted derivatives have been represented variously as 1,5-naphthoquinones (**19**) and as the tautomeric 1,4-naphthoquinone imines (**20**). Naphthoxidine itself, for example, is listed under two CAS registry numbers, i.e., [*27823-83-2*] and [*6259-68-3*], for the 1,5-quinone and the 1,4-quinone imine forms, respectively. Although definitive structural evidence has yet to be provided, the generally preferred formulation is the former, and the exceptionally low IR carbonyl frequencies of these compounds (i.e., below 1600 cm^{-1}) support this suggestion. Such a low frequency would be characteristic of a 1,5-naphthoquinone [33], particularly if intramolecular hydrogen bonding were present. Hydrogen bonding in a 1,4-quinone structure such as **20** would not account for such low frequencies, because, for example, 5-amino-1,4-naphthoquinones show absorptions for hydrogen-bonded carbonyl groups near 1630 cm^{-1}. In the following account, the 1,5-quinone structure (**19**) will be assumed.

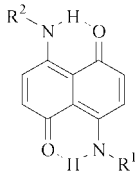

19

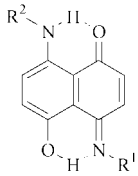

20

Naphthoxidine is a very useful intermediate for dyes of this group, and it can be prepared conveniently by reacting 1,5-dinitronaphthalene with sulfur in oleum.

Procedure [34, 35]. A solution of sulfur (1.25 g) in 20 % oleum (15 cm^3) was added slowly over 35 min to a slurry of 1,6-dinitronaphthalene (2.5 g) in concentrated sulfuric acid (10 cm^3), maintaining the temperature at 50 – 55 °C. The dark brown mixture was stirred for a further 10 min at this temperature and then for 30 min at 20 °C. The mixture was then poured onto ice (200 g) and the suspension filtered. The blue filtrate was made alkaline with sodium hydroxide solution, employing efficient external cooling in order to keep the temperature below 40 °C. The resultant blue precipitate of the sodium salt of naphthoxidine was filtered off and washed with a little water. The solid was extracted with boiling acetic acid, and concentration of the extracts gave lustrous red needles of naphthoxidine (1.44 g, 68 %).

The most successful naphthoxidine-derived dye was *Artisil Blue GFL* (**4**) [*26846-51-5*], which is prepared by bromination of naphthoxidine in acetic acid in the presence of a catalytic amount of iodine or iron. The commercial dye is in fact a mixture of various brominated products, of which **4** is the predominant component. The product is a clear blue disperse dye suitable for acetate and polyester fibers.

Condensation of naphthoxidine with arylamines results in replacement of one or both primary amino groups by the amine, and the resultant *N*-aryl derivatives are greener than **2** and can be used as disperse dyes. An example is *Setacyl Blue Green BS* (**19**; R^1 = 4-ethoxyphenyl, R^2 = H) [1]. Arylamines react directly with **2** in acetic acid, but the reaction is unsuccessful with alkylamines. The latter can be introduced by first reducing **2** to a stable leuco compound with hydrosulfite and then condensing this with the alkylamine. The product is then oxidized to give the quinone dye. It appears that monocondensed dyes have greater affinity for cellulose acetate fibers than the bis-condensation products. Interest in the latter has been stimulated by the discovery that such compounds can provide useful dyes for liquid crystal display devices and exhibit high-order parameters (i.e., contrast ratios). For example, the compound **19** with R^1 = R^2 = 4-methoxyphenyl is a blue-green dye providing an excellent contrast ratio in liquid crystal media [36].

Brominated condensation products can be obtained by using appropriately halogenated naphthoxidine intermediates, e.g., **4**. For example, **4** may be condensed with aniline in acetic acid at 90 °C to give **21** [1]. Naphthoxidine and its brominated derivatives may also be condensed with aminoazo compounds, giving mixed azo – quinone chromogens, e.g., **22**. These are of interest as homogeneous green dyes [1].

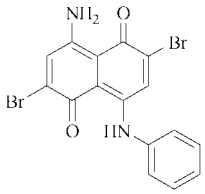

21

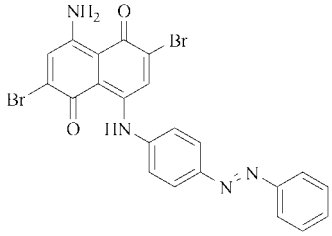

22

It is possible by modification of reaction conditions to effect the introduction of amino groups into the 2-position of naphthoxidine, but this generally produces dyes with reduced purity of shade and of lower light fastness [1].

As in the case of the 1,4-naphthoquinones, there has been considerable interest in developing near-infrared dyes of this class. The dye **23** [*100012-51-9*] is essentially the 1,5-quinone analogue of dye **18** (X = S) and shows a long-wavelength peak at 750 nm in trichloromethane (ε_{max} 32 000 L mol^{-1} cm^{-1}) [37]. Thus although **23** has the same molecular mass as **18** (X = S), it is both more bathochromic and more intensely absorbing.

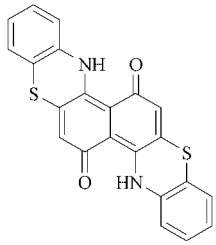

23

5. References

General References

1. E. Merian, *Chimia* **13** (1959) no. 6, 181; *Am. Dyest. Rep.* **48** (1959) 31.
2. *Colour Index*, 1st ed. (1924), 2nd ed., vol. **1 – 6** (1956), 3rd ed., vol. **1 – 6** (1971), The Society of Dyers and Colourists, Bradford/England.
3. R. H. Thomson: *Naturally Occurring Quinones*, Butterworths, London 1957.
4. *Houben-Weyl*, **7 (3 a).**
5. S. Patai (ed.): *The Chemistry of the Quinoid Compounds*, parts 1 and 2, J. Wiley & Sons, London–New York 1974.

6. M. F. Sartori, *Chem. Rev.* **63** (1963) 279.
7. I. Baxter, B. A. Davis, *Q. Rev. Chem. Soc. (London)* **25** (1971) 239.
8. B. D. Tilak in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. V, Academic Press, New York–London 1971.

Specific References

9. D. Zhang, G.-X. Jin, *Organometallics* **22** (2003) 2851.
10. S. Muralidharan, S. Ravichandran, S. S. Pitchumani, K. L. N. Phani, *J. Mater. Sci. Lett.* **19** (2000) 1299.
11. Fr. Demande, FR 2002-9144, 2002 (P. Braunstein, O. Siri, J. P. Taquet).
12. O. Siri, P. Braunstein, M.-M. Rohmer, M. Benard, R. Welter, *J. Am. Chem. Soc.* **125** (2003) 13793.
13. K. Y. Chu, J. Griffiths, *J. Chem. Soc. Perkin Trans. 1* 1978, 1083.
14. K. Y. Chu, J. Griffiths, *J. Chem. Res. Synop.* 1978, 180;
J. Chem. Res. Miniprint 1978, 2319.
15. K. Y. Chu, J. Griffiths, *J. Chem. Soc. Perkin Trans. 1* 1979, 696.
16. C. Blackburn, J. Griffiths, *J. Chem. Res. Synop.* 1983, 168;
J. Chem. Res. Miniprint 1983, 1556.
17. L. K. Vinograd, S. M. Shein, A. P. Cherepivskaya, G. V. Shalimova, *Zh. Prikl. Khim. (Leningrad)* **38** (1965) 208;
Chem. Abstr. **62** (1965) 11941 e.
18. C. Lion, R. Baudry, M. Hedayatullah, L. Da Conceicao, M. Hocquaux, S. Genard, J. Maignan, *J. Heterocycl. Chem.* **37** (2000) 1635.
19. C. Lion, R. Baudry, M. Hedayatullah, L. Conceicao, S. Genard, J. Maignan, *J. Heterocycl. Chem.* **39** (2002) 125.
20. L'Oreal, EP 98-400298 1998 (S. Genard, H. Andrean, M. Hocquaux).
21. Sandoz, US 2 764 600, 1956 (E. Merian).
22. Mitsubishi Chemical Industries, JP 57 145 157, 1982.
23. D. Heard, G. G. Roberts, J. Griffiths, *Thin Solid Films* **180** (1989) 305.
24. NEC, JP 60 149 490A, 1983.
NEC, JP 60 150 241A, 1983.
NEC, JP 60 150 242A, 1983.
NEC, JP 60 161 192A, 1984.
NEC, JP 60 190 388A, 1984.
25. G. Green-Buckley, J. Griffiths, *J. Chem. Soc. Perkin Trans. 1* 1979, 702.
26. American Cyanamid, US 2 967 863, 1961 (M. Scalera, A. S. Tomcufcik, W. B. Hardy).
27. B. D. Tilak, V. K. Dikshit, B. Suryanarayana, *Proc. Indian Acad. Sci. Sect. A* **37** (1953) 92.
28. A. T. Peters, D. Walker, *J. Chem. Soc.* 1956, 1429.
29. E. F. Pratt, R. G. Rice, R. W. Luckenbaugh, *J. Am. Chem. Soc.* **79** (1957) 1212.
30. J. A. VanAllan, R. G. Reynolds, R. E. Adel, *J. Org. Chem.* **28** (1963) 3502.
31. Sandoz, DE 2425662, 1974 (B. L. Kaul).
32. K. Takagi, M. Kawabe, M. Matsuoka, T. Kitao, *Dyes Pigments* **6** (1985) 177.
33. L. K. Horst, H. Kratzin, P. Boldt, *Liebigs Ann. Chem.* 1976, 1560.
34. Du Pont, US 2238959, 1941 (M. S. Whelen).
35. K. Y. Chu: Ph. D. Thesis, University of Leeds, 1977.
36. Merck, DE 3126108, 1983 (G. Haas, G. Weber).
37. K. Takagi, S. Kanamoto, K. Itoh, M. Matsuoka, S. H. Kim, T. Kitao, *Dyes Pigments* **8** (1987) 71.

Benzyl Alcohol

Friedrich Brühne, Bayer AG, Krefeld-Uerdingen, Federal Republic of Germany (Chaps. 1 – 9)

Elaine Wright, General Motors Research Laboratories, Warren, Michigan 48090, United States (Chap. 10)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Hydrolysis of Benzyl Chloride 3
4.2. Hydrogenation and Reduction of Benzaldehyde 4
4.3. Oxidation of Toluene 4
4.4. Other Manufacturing Processes 5
5. Quality Standards and Test Methods 5
6. Storage, Transportation, and Safety Regulations 6
7. Uses 6
8. Derivatives 6
8.1. Dibenzyl Ether 6
8.2. Benzyl Acetate 7
8.3. Benzyl Benzoate 7
8.4. Benzyl Salicylate 7
9. Economic Aspects 7
10. Toxicology 7
11. References 8

1. Introduction

Benzyl alcohol [*100-51-6*] is the simplest and also the industrially most important aromatic alcohol. Liebig and Wöhler first prepared benzyl alcohol from bitter almond oil (benzaldehyde) in 1832 [3]. The structure of benzyl alcohol was determined in 1853 by Cannizzaro [4]. Cannizzaro used the reaction named after him, in which benzaldehyde is disproportionated into benzoic acid and benzyl alcohol through the action of an alkali.

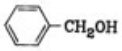

Benzyl alcohol occurs in nature both free and in combined forms. In the latter case it exists as esters of acetic, benzoic, salicylic, and cinnamic acids, among others. It is known to occur in the balsams of Peru and Tolu, in the flower oils of hyacinths and wallflowers, in ylang-ylang oil, and in other essential oils. It also occurs as a glucoside in maize.

2. Physical Properties

Benzyl alcohol, C_7H_8O, is a colorless liquid with a faint aromatic odor and a mildly irritating effect on the mucous membranes.

M_r 108.14, *bp* 205.4 °C at 101.3 kPa, *mp* −15.4 °C, refractive index n_D^{20} 1.5400, density ϱ 1.061 g/cm^3 at 0 °C, 1.045 g/cm^3 at 20 °C, 1.030 g/cm^3 at 40 °C, 1.015 g/cm^3 at 60 °C.

Specific heat 1972 J kg^{-1} K^{-1} at 20 °C, 2135 J kg^{-1} K^{-1} at 50 °C, heat of fusion 82.9 J/g, heat of evaporation 467.0 J/g at 205.4 °C, standard combustion enthalpy 34.58 kJ/g.

Flash point 101 °C, autoignition temperature 435 °C, lower explosive limit (at 170 °C and 101.3 kPa) 1.3 vol %, upper explosive limit 13.0 vol %.

Table 1 shows the vapor pressure in relation to temperature.

Table 1. Vapor pressure of benzyl alcohol vs. temperature

Temperature, °C	Vapor pressure, kPa
60	0.18
80	0.66
100	2.02
120	5.28
140	12.20
160	25.47
180	48.88
200	87.43

Benzyl alcohol is miscible with many organic solvents. At 20 °C, 4.0 g of benzyl alcohol is soluble in 100 g of water and 5.1 g of water is soluble in 100 g of benzyl alcohol. The compositions of several binary azeotropic mixtures are

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_001

given in Table 2. For further azeotropes, see [5]. Liquid – vapor equilibrium data for binary mixtures of benzyl alcohol with water or organic solvents are given in [6].

Table 2. Azeotropes of benzyl alcohol

Component	*bp* (101.3 kPa), °C	Benzyl alcohol, wt %
Water	99.9	9.0
Hexachloroethane	182.0	12.0
Ethylene glycol	193.4	46.5
Nitrobenzene	204.2	62.0
Dimethylaniline	193.9	6.5
Diethylaniline	204.2	72.0
Naphthalene	204.1	60.0
p-Cresol	206.8	62.0
p-Bromotoluene	184.5	8.0

Dynamic viscosity η 5.584×10^{-3} Pa · s at 20 °C, surface tension σ 39.96×10^{-3} N/m at 20 °C. Dipole moment μ (measured in liquid benzene) 5.571×10^{-30} C · m (1.67 D), relative dielectric constant ε_r 11.92 at 30 °C, 9.81 at 60 °C.

3. Chemical Properties

The chemical properties of benzyl alcohol are determined mainly by the hydroxyl group. This group shows the typical reactions of aliphatic alcohols (→ Alcohols, Aliphatic). The reactivity of the hydroxyl group is increased by the proximity of the aromatic nucleus. Benzyl alcohol is less acidic than the isomeric cresols and therefore not completely soluble in aqueous alkalis.

If benzyl alcohol is heated in the presence of dehydrating compounds (e.g., aluminum oxide), dibenzyl ether, toluene, and benzaldehyde are formed. Dibenzyl ether is also formed when benzyl alcohol is heated with strong acids or bases. Mixed ethers can be prepared simply through the action of alkyl halides on benzyl alcohol in the presence of bis(acetylacetonato)nickel as a catalyst [7].

Oxidation results in benzaldehyde or benzoic acid, depending on the nature of the oxidizing agent and the reaction conditions. Thus nitric acid as an oxidizing agent gives benzaldehyde [8], whereas solid sodium permanganate monohydrate gives benzoic acid [9]. Under the conditions of the Oppenauer oxidation, benzyl alcohol is oxidized to benzaldehyde in the presence of furfural as a hydrogen acceptor [10]. Benzaldehyde is also formed slowly if benzyl alcohol is exposed to air. Benzyl alcohol can be dehydrogenated in the gas phase using catalysts containing copper or noble metals. The main product of the dehydrogenation is benzaldehyde.

Depending on the reaction conditions, the hydrogenation of benzyl alcohol gives toluene, benzene, methylcyclohexane, cyclohexane, and hydroxymethylcyclohexane.

The reactions between benzyl alcohol and hydrogen halides give the corresponding benzyl halides [11]. Under the conditions for chlorinating the side chains of aromatic compounds, benzyl alcohol forms benzoyl chloride [12].

With ammonia or amines, benzyl alcohol can be converted into *N*-alkyl-, *N,N*-dialkyl-, *N,N,N*-trialkylamines. In the presence of Friedel-Crafts catalysts, aromatic hydrocarbons can be alkylated on the aromatic nucleus. Thus benzyl alcohol and benzene form diphenylmethane, whereas benzyl alcohol and phenol give a mixture of 2- and 4-benzylphenols.

With organic and inorganic acids, acid halides, or acid anhydrides, benzyl alcohol reacts in the normal way to give the corresponding esters. The acid-catalyzed reactions of benzyl alcohol with aldehydes give acetals. The addition of carbon monoxide to benzyl alcohol in the presence of carbonyl catalysts, e.g., rhodium trichloride with benzyl iodide, leads to phenylacetic acid [13].

The treatment of benzyl alcohol with Friedel-Crafts catalysts (e.g., anhydrous aluminum chloride, zinc chloride, concentrated sulfuric acid, or perchloric acid) results in the elimination of water and the formation of resinous products having the composition $H(C_7H_6)_n$ OH and corresponding to the polybenzyl made from benzyl chloride. At temperatures above 150 °C, as little as 1 % hydrogen bromide and 0.04 % iron (II) can cause the exothermic polycondensation of benzyl alcohol to proceed very rapidly. In this reaction the temperature rises to 240 °C, and the sudden pressure increase may be sufficient to break a closed vessel [14]. Thus, for safety reasons, benzyl alcohol should not be heated to temperatures above 100 °C unless it is free from acidic matter and dissolved iron.

4. Production

Benzyl alcohol can be produced in many ways. At present there are only two processes of substantial industrial importance.

1) The hydrolysis of benzyl chloride

2) The hydrogenation of benzaldehyde

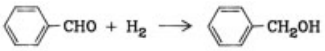

4.1. Hydrolysis of Benzyl Chloride

The hydrolysis of benzyl chloride is a reversible reaction which leads to the almost quantitative formation of benzyl alcohol only in the presence of alkaline saponifying agents that combine with the hydrogen chloride as it is formed. It is therefore carried out by heating benzyl chloride with stoichiometric excesses of aqueous solutions of oxides, hydroxides, or carbonates of the alkali or alkaline earth metals. As in the following example, soda is the preferred saponifying agent.

To 610 parts of boiling 10 % soda solution 126.5 parts of benzyl chloride is added with stirring. The reaction mixture is refluxed and stirred until carbon dioxide no longer escapes; this takes an average of five to six hours. An alkali-resistant steel reactor with brick walls is used. The stirrer and heating coil are made of an alloy with a high nickel content. After the reaction mixture has cooled, the upper layer, consisting of crude benzyl alcohol, is removed. The sodium chloride solution below still contains some soda. Dissolved benzyl alcohol can be obtained from it by adding salt or by extracting with benzene or toluene. The crude benzyl alcohol is carefully fractionated at a reduced pressure in an efficient column. The yield is 85 parts of benzyl alcohol and 10 parts of dibenzyl ether.

In the discontinuous hydrolysis of benzyl chloride the reaction time can be reduced by hydrolyzing the final residue of benzyl chloride with sodium hydroxide solution after the saponification with soda solution has been in progress for three hours [15]. The total saponification time required is then only four hours. Normally 8 – 12 % of dibenzyl ether is formed as an unavoidable byproduct of the alkaline hydrolysis of benzyl chloride. Pure dibenzyl ether is recovered by distillation from the residue in the distillation column. The use of alkali hydroxides instead of carbonates reduces the reaction time but favors the formation of dibenzyl ether, which may then exceed 20 % in the crude benzyl alcohol. The percentage of dibenzyl ether can be reduced if the hydrolysis is performed in the presence of an inert solvent, such as benzene, toluene, or xylene.

Benzyl alcohol can be produced continuously if benzyl chloride and the aqueous alkaline solution of saponifying agent react in a flow reactor at an elevated temperature (150 – 350 °C) and sufficient pressure to maintain the reactants in the liquid phase. Thorough mixing of the reactants is important; it can be achieved by the use of suitable baffles in the flow reactor [16] or by maintaining an adequate flow rate [17]. The residence time need then be no more than a few minutes. The mixture emerging from the hydrolysis zone is processed in the normal way. Only a small amount of dibenzyl ether is formed and this can be further reduced by using an inert solvent.

In a new continuous process, benzyl chloride, flowing in one direction, and the alkaline saponifying agent, flowing in the other, are reacted as counter-currents in an inert organic solvent (e.g., toluene, xylene) in a flow reactor. In the extraction zone, the alcohol dissolved in the aqueous alkaline phase is extracted by an inert organic solvent flowing in the opposite direction. Water, likewise as a countercurrent, washes the solution of the crude benzyl alcohol in a washing zone. The flow reactor, extracting zone, and washing zone are integrated in a special apparatus (Fig. 1). The wash water and extract are returned to the hydrolysis zone. This process can be operated at low reaction temperatures (120 – 150 °C) and with a small stoichiometric excess of saponifying agent (5 – 25 %); a small reactor can be used [18].

In a two-step process, benzyl chloride reacts with sodium acetate to form benzyl acetate, which is then saponified to form benzyl alcohol [19]. In a similar process, benzyl chloride reacts with an alkali or alkaline earth formate in the presence of a catalyst to form benzyl formate. The benzyl alcohol is liberated by subsequent

transesterification using an alcohol and a catalyst [20]. Neither of these processes has acquired much importance in the manufacture of benzyl alcohol. However, they could have advantages in the hydrolysis of certain substituted benzyl chlorides that react with soda solution to give dibenzyl ethers and not the desired alcohol.

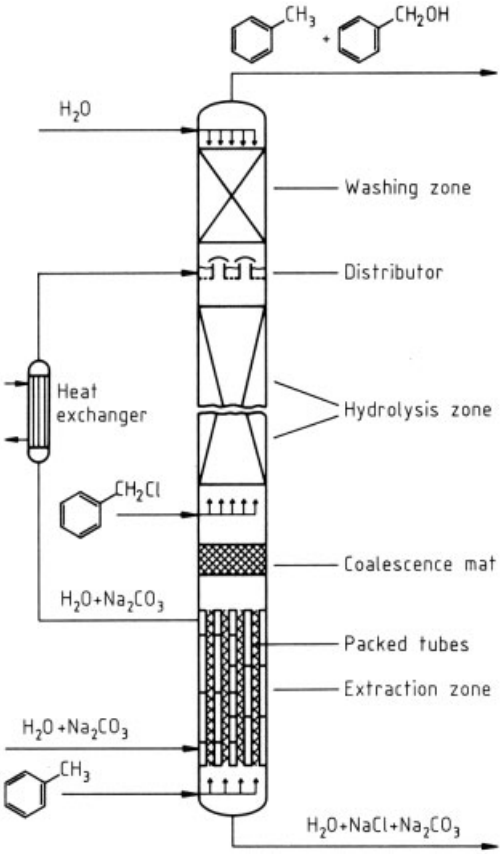

Figure 1. Hydrolysis of benzyl chloride to form benzyl alcohol in a continuous-flow reactor [18]

4.2. Hydrogenation and Reduction of Benzaldehyde

The industrial production of benzyl alcohol by the hydrogenation of benzaldehyde gained importance because substantial quantities of benzaldehyde were available as a byproduct of the production of phenol (Dow process) and ε-caprolactam (Snia Viscosa process) by the oxidation of toluene (→ Benzaldehyde).

Depending on the reaction conditions, benzaldehyde can be hydrogenated to form benzyl alcohol, toluene, hydroxymethylcyclohexane, methylcyclohexane, and other products. However, very high yields of benzyl alcohol are obtainable if suitable reaction conditions and catalysts are chosen. Raney nickel doped with transition metals [21], nickel or platinum metals with phosphines or phosphine oxides [22], palladium combined with an organic nitrogen [23] or alkali base [24], water [25], or another transition metal [26] can serve as the catalyst. If benzaldehyde is hydrogenated at temperatures of 70 – 200 °C and a hydrogen pressure of 1 – 4 MPa, high yields of benzyl alcohol are obtained in short reaction times. In a well-known continuous process [27], benzaldehyde is hydrogenated to form benzyl alcohol with a high degree of efficiency and a high selectivity in the presence of a platinum–aluminum oxide–lithium oxide catalyst. Stannane [28], sodium hydride [29], zinc [30], and microorganisms [31] can also be used to reduce benzaldehyde to benzyl alcohol, but these processes have no industrial importance. Benzyl alcohol is no longer produced industrially using the Cannizzaro reaction.

4.3. Oxidation of Toluene

The catalytic oxidation of toluene [32] gives only low yields of benzyl alcohol because the reaction conditions support further oxidation to benzaldehyde and benzoic acid. Therefore intermediates are isolated in most benzyl alcohol production processes that start with toluene. The liquid-phase oxidation of toluene with oxygen-depleted air at 170 – 220 °C and pressure sufficient to keep the reaction medium liquid is carried out in the presence of stabilizers for organic hydroperoxides (e.g., sodium pyrophosphate or sodium fluoride). It leads mainly to benzyl hydroperoxide [33] if not more than 10 % of the toluene is permitted to react. The subsequent decomposition of the oxidation product at 165 °C in the presence of soluble cobalt salts yields benzyl alcohol. Benzaldehyde and benzoic acid are formed in fairly large quantities as byproducts. Benzyl hydroperoxide can also be reduced to benzyl alcohol using alkali metal sulfites [34].

The oxidation of toluene as a liquid phase in air is conducted in the presence of acids, acid chlorides, or acid anhydrides; it forms the benzyl esters of these acids. These are then saponified to form benzyl alcohol. For example, the oxidation of toluene in the presence of acetic anhydride at temperatures of 140 – 240 °C and pressures of 1 – 3 MPa yields benzyl acetate [35]. The oxidation is discontinued after 10 % of the toluene has reacted. After the saponification of the oxidation product, 350 g of benzyl alcohol, 55 g of benzaldehyde, and 67 g of benzoic acid are obtained from 500 g of reacted toluene. Phenolic impurities, e.g., cresols, are unavoidably present in the benzyl alcohol obtained by the oxidation of toluene. These impurities can be removed by washing the benzyl alcohol vapor with a countercurrent of alkali benzylate solution in a plate column or a packed column [36].

4.4. Other Manufacturing Processes

The hydrogenation of benzoic acid esters to form benzyl alcohol is of particular interest in cases where benzoic acid esters are formed in substantial amounts. For example, methyl benzoate is formed in the production of dimethyl terephthalate by the Witten process. Catalysts containing copper are preferred. Benzoic acid esters can be hydrogenated to form benzyl alcohol with the aid of a copper catalyst and a carrier consisting of alkaline earth oxides or carbonates [37]. The selectivity can be increased if the catalyst also contains chromium. Temperatures of 100 – 300 °C and pressures above 6 MPa are used.

The hydrogenation of methyl benzoate in the presence of a copper–chromium catalyst is a feature of a three-step process for the production of benzyl alcohol from toluene [38]. The first step is the oxidation of toluene to form benzoic acid. The acid is esterified with methanol in the second step. The hydrogenation of methyl benzoate to form benzyl alcohol is performed in the final step. This process produces benzyl alcohol at an exceptionally low cost.

Ruthenium, rhodium, platinum, and palladium, activated by alkali metal arenes, ketyls, or alkoxides, are catalysts for the hydrogenation of the esters [39]. Using these catalysts, benzoic acid esters can be hydrogenated to form benzyl alcohol under mild conditions and with high selectivity.

Other processes for the production of benzyl alcohol include the hydrogenation of benzoic acid [40], the electrochemical reduction of benzoic acid [41], the hydrolysis of benzylsulfonic acid [42], and the decarboxylation of benzyl formate [43]. These processes have no importance in the industrial production of benzyl alcohol, but they may be used to produce derivatives substituted on the aromatic nucleus. Benzyl alcohol can also be obtained from the benzyl benzoate which is formed in the manufacture of benzoic acid [44].

5. Quality Standards and Test Methods

Several grades of benzyl alcohol are commercially available (Table 3). *Technical* grade is obtained as the first and last fractions in the distilla tion of the crude alcohol, it has little importance. *Pure* benzyl alcohol is the most important grade in terms of volume and is required for most applications. *Double-distilled* benzyl alcohol is used mainly in the perfume and pharmaceutical industries and must therefore be exceptionally pure and free of foreign odors. At 25 °C it forms a clear 4 % aqueous solution. The *photographic* grade normally contains 0.01 – 0.02 % of hydroquinone monomethyl ether to inhibit oxidation.

Gas chromatography using packed or capillary columns is suitable for the quantitative determination of benzyl alcohol. It is more accurate than the method involving acetylation of the hydroxyl group with pyridine and acetic anhydride. Traces of chlorine compounds are unavoidably present in the benzyl alcohol produced by hydrolysis of benzyl chloride. Byproducts, e.g., cresols, may be present in the benzyl alcohol obtained by the oxidation method or via benzaldehyde. Chlorine is quantitatively determined by Wickbold combustion [45], and cresols can be detected by color reactions with aminoantipyrine [36]. Benzaldehyde can be quantitatively determined using hydroxylamine hydrochloride.

The pharmacopoeias of the various countries should be consulted regarding the testing of benzyl alcohol which is to be used in the pharmaceutical industry.

Table 3. Specifications for certain grades of benzyl alcohol (wt %)

Compound	Grades			
	Technical	Pure	Double distilled	Photographic
Benzyl alcohol	> 97.0	> 99.0	> 99.8	> 99.5
Benzaldehyde	< 0.5	< 0.2	< 0.04	< 0.03
Chlorine	< 0.2	< 0.02	< 0.01	< 0.005
Dibenzyl ether	< 2.0	< 0.05	< 0.02	< 0.02

6. Storage, Transportation, and Safety Regulations

Benzyl alcohol is stored in stainless steel tanks. Because benzyl alcohol oxidizes readily, it is advisable to cover the surface of the liquid with nitrogen. Benzyl alcohol can be transported in drums protected by stoving finishes and in tank wagons of aluminum or stainless steel.

Benzyl alcohol belongs to Temperature Class T 2 (autoignition temperature 435 °C). The transportation of benzyl alcohol by rail, road, or water is not subject to special regulations. The labeling of benzyl alcohol must comply with the national and international regulations on dangerous substances.

7. Uses

By virtue of its good solvent properties, benzyl alcohol is an important solvent for surface-coating materials and resins. It dissolves cellulose esters and ethers, alkyd resins, acrylic resins, and fats; it is also used as an ingredient in the inks for ball-point pens. It is added in small amounts to surface-coating materials to improve their flow and gloss. In the textile industry, it is used as an auxiliary in the dyeing of wool, polyamides, and polyesters. Because it has only a relatively faint odor, it is used as a solvent and diluting agent in the manufacture of perfumes and flavors. In color photography it is increasingly important as a development accelerator. In pharmacy it is used as a local anesthetic and, because of its microbicidal effect, as an ingredient of ointments and other preparations.

Benzyl alcohol is a starting material for the preparation of numerous benzyl esters that are used as odorants, flavors, stabilizers for volatile perfumes, and plasticizers. Benzyl alcohol is also used in the extractive distillation of *m*- and *p*-xylenes and *m*- and *p*-cresols.

8. Derivatives

The most important derivatives of benzyl alcohol are dibenzyl ether, benzyl acetate, benzyl benzoate, and benzyl salicylate.

8.1. Dibenzyl Ether

Dibenzyl ether [*103-50-4*], $C_{14}H_{14}O$, M_r 198.27, is a colorless liquid with a faint odor; *mp* 3.6 °C, *bp* 296–298 °C at 101.3 kPa, *bp* 157–158 °C at 1.2 kPa, density ϱ 1.043 g/cm^3 at 20 °C, refractive index n_D^{20} 1.5622, flash point 135 °C. The compound is very slightly soluble in water.

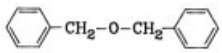

Production. Dibenzyl ether is formed when benzyl alcohol is heated with strong acids or bases. It is produced almost exclusively by the alkaline hydrolysis of benzyl chloride. Heat treatment decomposes dibenzyl ether into benzaldehyde and toluene.

Use. Dibenzyl ether is used as a plasticizer in the surface coatings industry, for special purposes in the rubber and textile industries, and as a solvent for musk and other odorants.

Excess dibenzyl ether can be converted into a mixture of benzotrichloride and benzoyl chloride under the conditions of side chain chlorination with chlorine [46]. Treatment of dibenzyl ether with hydrochloric acid in the presence of trialkyl- or tetraalkylammonium salts gives benzyl chloride [47].

8.2. Benzyl Acetate

Benzyl acetate [*140-11-4*], $C_9H_{10}O_2$, M_r 150.18, is a colorless liquid with an odor of pears and jasmine. It is a constituent of jasmine, gardenia, and other flower oils; *mp* −51.5 °C, *bp* 215 °C at 101.3 kPa, density ϱ 1.056 g/cm^3 at 20 °C, refractive index n_D^{20} 1.5020, flash point 102 °C, very slightly soluble in water.

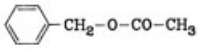

Production and Use. Benzyl acetate is produced by reacting benzyl chloride with an alkali acetate or by esterifying benzyl alcohol with acetic anhydride. It is also formed in the oxidation of toluene in the presence of acetic acid or acetic anhydride. Benzyl acetate is one of the most used odorants. It is also used as a flavor and, to a small extent, as a high-boiling solvent.

8.3. Benzyl Benzoate

Benzyl benzoate [*120-51-4*], $C_{14}H_{12}O_2$, M_r 212.25, is a colorless, viscous liquid with a pleasant odor. It is the main constituent of Peru balsam, Tolu balsam, and certain flower oils; *mp* 19–21 °C, *bp* 323–324 °C at 101.3 kPa, density ϱ 1.112 g/cm^3 at 25 °C, refractive index n_D^{24} 1.5672, flash point 148 °C, insoluble in water and glycerol.

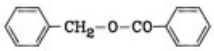

Production. Benzyl benzoate is produced from benzyl alcohol and sodium benzoate in the presence of triethylamine or by transesterification of methyl benzoate with benzyl alcohol in the presence of an alkali benzyl oxide. In another manufacturing process benzaldehyde is condensed to form benzyl benzoate in the presence of sodium (Claisen-Tishchenko condensation). The presence of a small amount of an aliphatic ether improves this reaction [48]. Benzyl benzoate is a byproduct in the manufacture of benzoic acid by the oxidation of toluene; it is present in the benzoic acid distillation residue.

Use. Benzyl benzoate is used in perfume manufacture, mainly to stabilize volatile odorants. It has vasodilating and spasmolytic effects and is present in many asthma and whooping cough drugs. It is used as an insect repellent and to treat scabies. Benzyl benzoate is used as a plasticizer in the surface coatings and plastics industries.

8.4. Benzyl Salicylate

Benzyl salicylate [*118-58-1*], $C_{14}H_{12}O_3$, M_r 228.25, is a colorless, viscous liquid with a pleasant odor. It occurs naturally in several essential oils; *mp* 24 °C, *bp* 211 °C at 2.67 kPa, density ϱ 1.175 g/cm^3 at 25 °C, refractive index n_D^{25} 1.5787, slightly soluble in water.

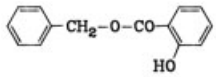

Production and Use. Benzyl salicylate can be manufactured by reacting benzyl chloride with an alkali salicylate or by the transesterification of methyl salicylate with benzyl alcohol. It is used as a perfume stabilizer and, because it absorbs ultraviolet light, as an ingredient in sunscreening preparations.

9. Economic Aspects

The benzyl alcohol production capacities of Western Europe and the United States in 1983 were approximately 10 kt and 6 kt, respectively. About 7 kt of benzyl alcohol was actually produced in Western Europe in 1983. Approximately 3.4 kt of benzyl alcohol was produced in the United States in 1979.

10. Toxicology

Benzyl alcohol is used in pharmaceuticals, soaps, detergents, and cosmetics as a fragrance, preservative, and antimicrobial, and is "generally regarded as safe" (GRAS) for use in food by the U.S. Food and Drug Administration. It has industrial applications as a solvent and dye

assistant, but no environmental or workplace exposure limits have been established.

Benzyl alcohol is moderately toxic with acute oral LD_{50} values of 1.2 g/kg in rats and 1.6 g/kg in mice [49]. The LD_{50} for inhalation exposure is approximately 1.0 g/kg (eight hours) in rats [50]. Benzyl alcohol can be absorbed through the skin in sufficient quantities to cause toxicity; the dermal LD_{50} in guinea pigs is < 5 mL/kg [51]. In vitro screening tests for chemical carcinogens in a repair-deficient bacterial strain (P 3478 *E. coli*) were negative [52]. Benzyl alcohol is not considered to be a carcinogen and no data are available regarding teratogenic or reproductive effects.

Benzyl alcohol is widely used as a preservative in solutions of drugs for intravenous administration. The minimum lethal dose in dogs (i.v.) is between 88 and 113 mL/kg of a 0.9 vol % solution of benzyl alcohol in water or in physiological saline solution [53]. Similar doses have been fatal for newborn humans receiving benzyl alcohol in the course of intravenous drug treatment [54]. Benzyl alcohol is estimated to be more than twenty times more toxic than ethyl alcohol (i.v.) [53].

Benzyl alcohol is severely toxic and highly irritating to the eye. Contact with benzyl alcohol causes mild to moderate skin irritation [55]. Hypersensitivity reactions may occur in some individuals with both topical and parenteral administration [56].

Signs of intoxication include vomiting and diarrhea, collapse, central nervous system depression, excitability, muscle paralysis, convulsions, dyspnea, and respiratory arrest [55]. Exposure should be avoided by the use of adequate ventilation systems, self-contained respiration equipment, protective goggles, gloves, and clothing.

11. References

General References

1. *Beilstein* **6,** 428; **6 (1),** 217; **6 (2),** 403; **6 (3),** 1445; **6 (4)** 2222.
2. M. Watanabe, K. Namikawa, *CEER Chem. Econ. Eng. Rev.* **2** (1970) 50 – 52.

Specific References

3. F. Wöhler, *Ann. Pharm. (Lemgo, Ger.)* **3** (1832) 249, 254.
4. S. Cannizzaro, *Justus Liebigs Ann. Chem.* **88** (1853) 129.
5. L. H. Horsley: "Azeotropic Data-III," *Adv. Chem. Ser.* 1973, no. 116, 374 – 376.
6. J. Gmehling, U. Onken, W. Arlt: "Vapor-Liquid Equilibrium Data Collection," *Dechema Chem. Data Ser.,* vol. I, part 1 a, 447 – 450, 1981, part 2 b, 434, 1978, part 2 d, 563 – 577, 1982.
7. M. Yamashita, Y. Takegami, *Synthesis* 1977, 803.
8. A. McKillop, M. E. Ford, *Synth. Commun.* **2** (1972) 307 – 313.
9. F. M. Menger, C. Lee, *Tetrahedron Lett.* **22** (1981) 1655 – 1656.
10. SCM Corp., DE 2556161, 1975 (W. J. Ehmann, W. E. Johnson Jr.).
11. J. F. Norris, *Am. Chem. J.* **38** (1907) 631.
12. ICI, GB 310909, 1928 (B. W. Henderson).
13. A. Masuda, H. Mitani, K. Oku, Y. Yamazaki, *Nippon Kagaku Kaishi* 1982, 249 – 256; *Chem. Abstr.* **96** (1982) 141989 y.
14. E. Iselin, F. Barfuss, *Chem. Ing. Tech.* **44** (1972) 462.
15. National Oil Products Co., US 2221882, 1938 (L. T. Rosenberg).
16. All-Union Scientific Research Institute of Synthetic and Natural Perfumes, SU 225183, 1967 (V. V. Kashnikov, N. I. Gel'perin, S. A. Voitkevich, O. N. Zhuchkova, T. P. Khrenkova); *Chem. Abstr.* **70** (1969) 28614 x.
17. Velsicol Chem. Corp., US 3557222, 1968 (H. W. Withers, J. L. Rose).
18. EKA, EP 64486, 1981 (R. K. Rantala, G. L. F. Hag).
19. Chemische Werke Witten, DE 1108677, 1959 (G. Renckhoff, H. L. Hülsmann).
20. Dynamit Nobel, DE 2825364, 1978 (G. Bernhardt, G. Daum).
21. Kazakh Chemical-Technological Institute, SU 235 002, 1966 (F. B. Bizhanov, D. V. Sokol'skii, N. I. Popov, A. M. Khisametdinov, N. Ya. Malkina); *Chem. Abstr.* **70** (1969) 114813 t.
22. Hoechst, DE 2837022, 1978 (K. Kümmerle, H. Heise).
23. Mitsui Petrochem. Industries, JP-Kokai 77 27738, 1975 (K. Teranishi, T. Shimizu, K. Koga); *Chem. Abstr.* **87** (1977) 134502 v.
24. Mitsui Petrochem. Industries, JP-Kokai 77 27737, 1975 (K. Teranishi, T. Shimizu, K. Koga); *Chem. Abstr.* **87** (1977) 134501 u.
25. Sumitomo Chem. Co., JP-Kokai 82 26634, 1980; *Chem. Abstr.* **96** (1982) 199279 h.

26. Mitsui Petrochem. Industries, JP-Kokai 76 86432, 1975 (K. Teranishi, T. Ito); *Chem. Abstr.* **85** (1976) 192342 n.
27. Universal Oil Products Co., US 3663626, 1969 (J. T. Arrigo, N. J. Christensen).
28. M & T Chemicals Inc., US 3708549, 1970 (G. H. Reifenberg, W. J. Considine).
29. Sagami Chem. Res. Center, JP-Kokai 76 39609, 1974 (T. Fujisawa, K. Sugimoto, H. Ota); *Chem. Abstr.* **85** (1976) 122719 s.
30. Sumitomo Chem. Co., JP-Kokai 74 135933, 1973 (T. Mizutani, Y. Ume, T. Matsuo); *Chem. Abstr.* **82** (1975) 155745 p.
31. Politechnika Wroclawska, PL 114065, 1979 (J. Wisniewski, A. Poranek, T. Winnicki); *Chem. Abstr.* **97** (1982) 214282 v.
32. Universal Oil Products Co., US 3775472, 1970 (S. N. Massie).
33. Rhône-Poulenc, FR 1366078, 1963 (J. Bonnard, G. Poilane).
34. Nippon Oils & Fats Co., JP-Kokai 78 21133, 1976 (S. Suyama, H. Ishigaki); *Chem. Abstr.* **88** (1978) 190371 j.
35. Institute of Petrochemical Synthesis, DE 1912078, 1969 (A. N. Baskirov, M. M. Grozhan, V. V. Kamzolkin, Yu. A. Lapitskii).
36. Rhône-Poulenc, DE 1668646, 1967 (J. Bonnart, P. Rey).
37. Nippon Mining Co., DE 1768262, 1967 (M. Watanabe, K. Mamikawa, K. Yasuda, M. Takagi).
38. K. Kato, K. Namikawa, M. Watanabe, *Bull. Jpn. Pet. Inst.* **14** (1972) 206 – 216; *Chem. Abstr.* **78** (1973) 60190 g.
39. Allied Chem. Corp., EP 36939, 1980 (R. A. Grey, G. P. Pez).
40. Sumitomo Chem. Co., EP 28422, 1979 (T. Takano, G. Suzukamo, M. Ishino, K. Ikimi).
41. Union Rheinische Braunkohlen Kraftstoff AG, DE 2237612, 1972 (G. Schwarzlose).
42. Universal Oil Products Co., US 3641165, 1967 (L. Schmerling, R. A. Dombro).
43. Dynamit Nobel, DE 2825362, 1978 (G. Bernhardt).
44. Stamicarbon B. V., EP 19341, 1979 (C. Jongsma, L. H. B. Frijns, P. A. M. Raven-Donners).
45. F. Ehrenberger, S. Gorbach: *Methoden der organischen Elementar- und Spurenanalyse,* Verlag Chemie, Weinheim 1973, p. 238 ff.
46. Bayer, DE 1909523, 1969 (W. Böckmann).
47. Odessa State University, SU 872525, 1980 (G. L. Kamalov, D. G. Chikhichin, L. Ya. Glinskaya, G. P. Fedorchenko, F. P. Filek); *Chem. Abstr.* **96** (1982) 85217 f.
48. Dow Chem. Co., US 3387020, 1964 (C. E. Handlovits, J. B. Louch).
49. P. M. Jenner, E. C. Hagan, J. M. Taylor, E. L. Cook, O. G. Fitzhugh, *Food Cosmet. Toxicol.* **2** (1964) 327 – 343.
50. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, *AMA Arch. Ind. Hyg. Occup. Med.* **4** (1951) 120.
51. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **11** (1973) 1011 – 1081.
52. E. R. Fluck, L. A. Poirier, H. W. Ruelius, *Chem. Biol. Interact.* **15** (1976) 219 – 231.
53. E. T. Kimura, T. D. Darby, R. A. Krause, H. D. Brondyk, *Toxicol. Appl. Pharmacol.* **18** (1971) 60 – 68.
54. J. Gershanik, B. Boecler, H. Ensley, S. McCloskey, W. George, *N. Engl. J. Med.* **307** (1982) 1384 – 1388.
55. V. K. Rowe, S. B. McCollister in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* **vol. 2 C,** John Wiley & Sons, New York 1982, pp. 4636 – 4703.
56. J. A. Grant, O. A. Bilodeau, B. G. Guernsey, F. H. Gardner, *N. Engl. J. Med.* **307** (1982) 108.

Benzyl Chloride → Chlorinated Hydrocarbons

Benzylamine

LUTZ HEUER, Lanxess Deutschland GmbH, Federal Republic of Germany

1. **Physical Properties** 1
2. **Chemical Properties** 1
3. **Production** 2
4. **Quality Specifications and Analysis** . . . 2
5. **Uses** 2
6. **Storage and Transportation** 2
7. **Ecology and Toxicology** 3
8. **References** 3

Benzylamine (α-aminotoluene) [*100-46-9*] is an important industrially used amine, which is readily available by hydrogenation of benzonitrile or amination of benzaldehyde or benzyl chloride. It also occurs in nature and can be isolated from the leaves of *Reseda media* in small amounts. Benzylamine is a colorless liquid of ammonia-like odor, which is miscible with water in all proportion. It is used in the chemical industry as a starting material for other products or as a corrosion inhibitor. Benzylamine is listed in EINECS 202-854-1, TSCA, and MITI/MO and traded world-wide. It can be transported in steel drums and in high-density polyethylene (HDPE) containers.Benzylamine is harmful in contact with skin and if swallowed. It causes burns, especially when skin is exposed to benzylamine for prolonged periods. Because of the fast reaction of benzylamine and carbon dioxide, slightly leaking flanges or other seals are easily identified by the white solid formed.Benzylamine is produced by Lanxess and other companies and is available in drums, IBC's (intermediate bulk containers), and in bulk.

1. Physical Properties

Benzylamine [*100-46-9*], EINECS 202-854-1, TSCA and MITI/MO listed, α-aminotoluene, phenylmethylamine, moringine, C_7H_9N, M_r 107.16, is a colorless liquid with a weak amine-like odor.

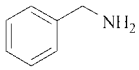

Benzylamine is miscible with water, alcohol, and diethyl ether in all proportions; *bp* 184.5 °C at 101.3 kPa, *bp* 90 °C at 1.6 kPa, solidification point < -30 °C, vapor pressure ca. 60 Pa at 20 °C, ca. 130 Pa at 30 °C, and 520 Pa at 50 °C.

Examples of azeotropic mixtures are given in Table 1.

Heat of vaporization 49 kJ/mol, heat of combustion 4058.7 kJ/mol at 101.3 kPa and 20 °C, refractive index n_D^{20} 1.5401, density d_4^{20} 0.9813, $d_4^{86.6}$ 0.9272, dielectric constant 4.6 at 20.6 °C, electrolytic dissociation constant in water 2.35×10^{-5} at 20 °C, viscosity 1.78 mPa · s at 21.2 °C, 0.295 mPa · s at 178.2 °C, surface tension 38.82×10^{-5} N/cm at 21.1 °C, 31.70×10^{-5} N/cm at 88 °C, autoignition temperature 390 °C, flash point 65 °C, explosion limit 0.9 % (lower) to 14 % (upper); dipole moment (20 °C) 1.18 D (neat), 1.25 D in toluene, 1.29 D in hexane, 1.30 D in cyclohexane, and 1.33 D in benzene [7]; dynamic viscosity 0.01596 g cm^{-1} s^{-1} at 25 °C [8]; absorption maximum λ_{max} in water 255.9 nm (ε 249 L mol^{-1} cm^{-1}) [9], magnetic susceptibility -75.3×10^{-6} cm^3/mol [10].

2. Chemical Properties

Benzylamine is a stronger base than its toluidine isomers and reacts strongly alkaline even if diluted with some water (pH 11.6 in water at 100 g/L) [11]. It forms addition compounds with, e.g., phenol (1 : 3, *mp* 15.3 °C), *p*-cresol (1 : 1, *mp* − 6 °C), and formic acid (1 : 1, *mp* 81 °C). Examples of the salts of benzylamine are the hydrochloride, $C_6H_5CH_2NH_3^+Cl^-$, *mp* 260 °C (decomp.), and the picrate, *mp* 194 °C.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_009.pub2

Table 1. Azeotropic mixtures containing benzylamine

Other component, A	*bp* of A, °C	*bp* of azeotropic mixture, °C	Component A in the azeotropic mixture, wt %
Aniline	184.35	186.2	42
Phenol	182.2	196.8	45
m-Cresol	202.2	207.2	56

Benzylamine absorbs carbon dioxide from the air to form the solid carbamic acid salt. If benzylamine is boiled with glacial acetic acid, *N*-acetylbenzylamine is formed. Benzylamine reacts with isocyanates to give substituted ureas, $C_6H_5CH_2NHCONHR$. Benzylamine is nitrated on the aromatic nucleus (8 % *ortho*-, 49 % *meta*-, 43 % *para*-nitrobenzylamine) by nitric acid. Catalytic hydrogenation gives hexahydrobenzylamine, $C_6H_{11}CH_2NH_2$.

3. Production

Benzylamine is produced by the reaction of benzyl chloride with ammonia [12] in aqueous solution. The catalytic hydrogenation of benzonitrile [13 – 15] and the reaction of benzaldehyde with ammonia in the presence of hydrogen and catalysts [16, 17] are alternatives. Both processes use some organic solvent in the reaction mixture.

In the benzaldehyde process, 8 kg of Raney nickel that has been kept under water (corresponding to about 4 kg of 100 % Ni), 10 g of glacial acetic acid, and 110 kg of ammonia are added to a mixture of 250 kg of methanol and 500 kg of benzaldehyde. The resulting batch is hydrogenated in a 1200-L steel autoclave at 100 °C and 15 MPa. The process takes 3 – 5 h. The hydrogen consumed in the reaction is made up at intervals of about 10 min. The contents of the reaction vessel are maintained at the same temperature and pressure for 30 min after hydrogen consumption ends. The catalyst is then separated from the reaction products on a pressure filter [16]. The filtrate contains 470 kg of benzylamine, about 1 kg of benzyl alcohol, less than 1 kg of dibenzylamine, and about 2 kg of Schiff base (*N*-benzalbenzylamine), in addition to methanol, reaction water, and ammonia. The yield is 93 %. Vacuum distillation yields benzylamine of the required technical purity.

Benzylamine has been isolated as a natural compound in the leaves and flowers of *Reseda media* [18].

4. Quality Specifications and Analysis

Commercial benzylamine has an assay of 99 % or more and contains less than 1.0 % water. Depending on the production process, some minor impurities are specified. Color number is typically less than 20 Hazen (APHA; DIN-ISO 6271), and the density at 20 °C 0.98 g/cm^3. For use in the pharmaceuticals industry, some customers require the absence of organic solvents such as benzene, xylene, and chlorinated solvents.

Benzylamine can be analyzed by gas chromatography on commercial columns such as Chrompack CP-SIL 8 for amines or Chrompack WACHS 51 for amines. The results should be corrected for water content.

5. Uses

Benzylamine is used in synthetic textiles, in paints, and as corrosion inhibitor [19]. It is also used as an intermediate in the production of pharmaceutical active substances [21] and compounds for plant and material protection. In this cases benzylamine is often used as “protected nitrogen” [24].

6. Storage and Transportation

For storage and transportation of benzylamine, most types of steel may be used (unalloyed steel, V2A, V4A types; e.g. 1.4301, 1.4306, 1.4571, 1.4401 [DIN 6601]). If chloride or bromide is present, some steel types should be avoided [DIN 6601]. One-tonne containers made from HDPE (Lupolen 4261 A) are also suitable (e.g. Schütz Container, intermediate bulk containers),

as are drums lined with this material. Benzylamine is slowly oxidized on contact with air and forms benzalbenzylamine. It also reacts with carbon dioxide to form its solid carbamic acid salt. It therefore should be stored in a tight closed container for not more than one year. For longer storage, a nitrogen blanket is recommended. In Germany the conditions for storage are classified as 8L (VCI-Lagerklasse) [25]. The handling of benzylamine is shown in a video sequence.

Benzylamine (EC No. 612-047-00-x) is included in the list of substances contained in EC Guideline 67/548, Appendix I and is therefore to be considered a dangerous industrial material. Containers must be marked as follows (EEC Label No. 202-854-1):

C	Corrosive; Benzylamine
R 21/22	Harmful in contact with skin and if swallowed.
R 34	Causes burns.
S 26	In case of contact with eyes, rinse immediately with plenty of water and seek medical advice.
S 36/37/39	Wear suitable protective clothing, gloves, and eye/face protection.
S 45	In case of an accident or if you feel unwell, seek medical advice immediately (show the label where possible).

The following symbols and indications are required on labels and/or transport papers [25]:

Sea (GGVSee/IMDG Code)	Class 8, UN no. 2735, MFAG 320, EmS 8-05, Packing Group (PG) III, Marine Pollutant (MPO): No
Rail/road, Germany (GGVE/GGVS)	Class 8, Item 53C
Rail/road, international (RID/ADR)	Class 8, Item 53C
Hazard symbol	Hazard 080, Compound 2735
Inland vessels (ADNR)	Class 8, Item 53C
Air (ICAO/IATA-DGR)	Class 8, UN no. 2735, PG III

Benzylamine is not allowed to be sent by mail within Germany but can be sent by express mail within Germany if 12 L per package is not exceeded. Declarations are as follows:

Land	UN no. 2735,amines, liquid, corrosive, n.o.s. (Benzylamine), Class 8, Item 53c, ADR (RID)
Sea	UN no. 2735, amines, liquid, corrosive, n.o.s. (Benzylamine)
Air	UN 2735, amines, liquid, corrosive, n.o.s. (Benzylamine)

7. Ecology and Toxicology

Results of laboratory investigations into the ecological behavior of benzylamine are summarized in Table 2 [25].

Table 2. Biodegradability of benzylamine

Method [26]	Degree of elimination	Result
OECD screening test	$> 95\,\%$	readily biodegradable
Sturm test	90 %	readily biodegradable
Zahn – Wellens test	95 %	well eliminated

Benzylamine is readily biodegradable. It is harmful in aquatic environment [German class for hazards in aquatic environment: 1 (weak effect) (WGK 1)]. Acute toxicity: *Pseudomonas fluorescens* (bacteria) 500 mg (EC_0); *Scenedesmus quadricauda* (algae) 6 mg (96-h EC_{10}); *Daphnia magna* 60 mg (48-h EC_{50}); *Leuciscus idus* (fish) 20 mg (48-h EC_0); *Pimephals promelas* (fish) 102 mg (96-h EC_{50}) [25].

Benzylamine causes burns to skin and eyes and may cause sensitization. Rinse immediately with plenty of water. Wash skin also with soap. When used regularly a 5 % aqueous solution of acetic acid should be available for neutralization of contaminated skin. If swallowed dilute by drinking water, rinse mouth. If risk of inhalation is given, use ABEK filters in breathing masks (DIN 3181).

Benzylamine is Ames test negative (no mutagenic effect). Acute oral toxicity: LD_{50} 1130 mg/kg (rat). Acute percutaneous toxicity: LD_{50} 1340 mg/kg (rat) [25]. Toxicity class 3 (Switzerland). All rats inhaling benzylamine for 3 h survived a two-week period (full body exposure). After full body exposure for 5 h, 17 % died.

8. References

General References

1. *Beilstein* **12**, 1013; **12 (1)**, 445; **12 (2)**, 540; **12 (3)**, 2194; **12 (4)**, 2155.
2. *Houben-Weyl*, **11/1**, Stuttgart 1957.
3. *Merck Index*, 12th ed., 1160.
4. *Merck FT-IR*, 1532.
5. *Fieser*, **10** 26.
6. *Kühn-Birett*, **B050**.

Specific References

7. *Beilstein* **12 (3)**, 2194; **12 (4)**, 2155.
8. A. G. Mussell, F. B. Thole, A. E. Dunstan, *J. Chem. Soc.* **101** (1912) 1014.
9. P. R. Haddad, R. C. Foley, *Anal. Chem.* **61** (1989) 1435 – 1441.
10. French, *Trans. Faraday Soc.* **1954**, 1320 – 1321.
11. Merck, Data Sheet for Benzylamine for Synthesis, 801812, Catalogue 1999.
12. A. T. Mason, *J. Chem. Soc.* **63** (1893) 1311. California Research Corp., US 2987548, 1961.
13. W. H. Carothers, G. A. Jones, *J. Am. Chem. Soc.* **47** (1925) 3051
14. BASF, FR 2165515, 1939.
15. M. Grunfeld, US 2449036, 1941.
16. C. F. Winans, *J. Am. Chem. Soc.* **61** (1939) 3566.
17. BASF, FR 146354, 1966.
18. O. Olsen, H. Soerensen, *Phytochemistry*, **19** (1980) 1783 – 1788.
19. B. Wanderott, *Z. Metallkd.* **56** (1965) no. 1, 63 – 64.
20. *Chem. Abstr.* **47** (1950) 7149 h.
21. H. Ippen: *Index Pharmacorum*, Thieme Verlag, Stuttgart 1968.
22. *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlamm-Untersuchung*, 3rd. ed., Verlag Chemie, Weinheim 1960 – 1971.
23. Bayer AG, Toxicological Institute, Elberfeld 1977.
24. T. W. Green (Ed.): *Protective Groups in Organic Synthesis*, Wiley, New York, 1981, p. 272.
25. Lanxess Safety Data Sheet Benzylamine 012193 (Sept. 26, 2005).
26. P. Gerike, W. K. Fischer, *Ecotoxicol. Saf.* **3** (1979) 159 – 173.

Beryllium and Beryllium Compounds

GÜNTER PETZOW, Max-Planck-Institut für Metallforschung, Institut für Werkstoffwissenschaften, Stuttgart, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14)

FRITZ ALDINGER, W. C. Heraeus GmbH, Hanau, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14)

SIGURD JÖNSSON, Leybold-Heraeus GmbH, Hanau, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14)

PETER WELGE, Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin, Bochum, Federal Republic of Germany (Chap. 15)

VERA VAN KAMPEN, Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin, Bochum, Federal Republic of Germany (Chap. 15)

THOMAS MENSING, Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin, Bochum, Federal Republic of Germany (Chap. 15)

THOMAS BRÜNING, Berufsgenossenschaftliches Forschungsinstitut für Arbeitsmedizin, Bochum, Federal Republic of Germany (Chap. 15)

1. Introduction 1
2. Physical Properties 2
3. Mechanical Properties 3
4. Chemical Properties 6
5. Resources, Raw Material 7
6. Production 7
6.1. Processing of Beryl 7
6.2. Processing of Bertrandite 9
6.3. Production of Reducible Compounds 10
6.4. Metal Production 11
6.5. Refining and Further Processing . . . 12
7. Environmental Protection 14
8. Quality Specifications 15
9. Beryllium Alloys 17
10. Beryllium Compounds 17
11. Chemical Analysis 20
12. Storage, Transportation, Legal Aspects 20
13. Uses of the Metal 20
14. Economic Aspects 20
15. Toxicology and Occupational Health . 20
16. Acknowledgement 24
17. References 24

1. Introduction

Beryllium [*7440-41-7*], Be, is the first element in the second main group of the periodic table. It is a light metal with hexagonal-closest-packed (hcp) structure. Both finely divided metal and the compounds have serious toxic effects on the lungs. In addition to the name beryllium, derived from the mineral beryl, this element with atomic number 4 was formerly called glucinium, chiefly in France, because of the pronounced sweetish taste of its salts.

In 1797 VAUQUELIN [1] discovered beryllium oxide while analyzing varieties of beryl including emerald and aquamarine. WÖHLER [2] and BUSSY [3], working independently, prepared elemental beryllium in 1828 by reduction of beryllium chloride with potassium. In 1899 LEBEAU [4] obtained beryllium crystals by electrolysis of molten sodium beryllium fluoride. Industrial production of beryllium began in 1916 in several countries. The publication in 1916 by OESTERHELD [5] of the phase diagrams of beryllium with iron, nickel, and copper led to the first important use of beryllium in 1931, namely, as an additive, about 2 wt%, in age-hardenable copper alloys [6].

The demand for beryllium increased around 1950 with the development of nuclear reactors because the nuclear properties and low density of beryllium seemed to indicate that the metal would be especially suitable for nonstationary

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_011.pub2

nuclear reactors of the types used in airplanes, space craft, ships, and submarines [7, 8]. However, beryllium was a disappointment because it swells and becomes brittle when exposed to the simultaneous action of radioactivity and heat [9]. However, the increased research activity in beryllium technology associated with this period produced significant advances in the chemistry, physics, and processing technology of beryllium [7, 10 – 14].

Since the end of the 1950s, there has been increased interest in beryllium as a structural material, especially in aeronautical and astronautical uses because its mechanical and thermal properties relative to weight are superior to those of all other materials [15 – 19]. Most of the beryllium produced today is used for military purposes. Greater use in the civilian sector is hindered mainly by the high price, the toxicity, and the room-temperature susceptibility to brittle fracture of the metal. On the other hand, beryllium oxide is finding increasing use in electrical engineering as an electrical insulator [20]. Beryllium properties, production methods, and applications have been comprehensively reviewed [21].

2. Physical Properties

The properties listed here were measured, for the most part, on commercial grades of beryllium. These contain various impurities and have various grain structures. The most reliable values of the important physical properties of beryllium [10, p. 320], [18, p. 201], [22 – 36] are compiled below. On first glance the most outstanding characteristics of the metal are low density and high melting point.

Atomic number		4
Atomic weight		9.0122
Electronic structure		$1s^2 2s^2$
Atomic radius		112.50 pm
Atomic volume at 298 K		4.877 cm^3/mol
Crystal structures		
hexagonal closest packed at 293 K	*a*	228.56 pm
	c	358.32 pm
	c/a	1.5677
body-centered cubic at 1523 K	*a*	255.0 pm
Melting point		1560 K or 1287 °C
Boiling point		2745 K or 2472 °C
Transformation point		1527 K or 1254 °C

Density at	
298 K	1.8477 g/cm^3
1773 K	1.42
Standard enthalpy	216 J/g
Standard entropy	1.054 J g^{-1} K^{-1}
Heat of fusion	1357 J/g
Entropy of fusion	0.87 J g^{-1} K^{-1}
Heat of transformation	837 J/g
Entropy of transformation	0.186 J g^{-1} K^{-1}
Vapor pressure at	
500 K	5.7×10^{-29} MPa
1000 K	4.73×10^{-12}
1560 K	4.84×10^{-6}
Specific heat at	
298 K	1.830 J g^{-1} K^{-1}
700 K	2.740
Thermal conductivity at 298 K	165 ± 15 W m^{-1} K^{-1}
Linear coefficient of thermal expansion, 298 – 373 K	11.5×10^{-6} K^{-1}
Electrical resistivity at 298 K	4.31×10^{-8} Ω m
Magnetic susceptibility at	
93 K	-0.75×10^{-9} m^3/kg
293 K	-1.00×10^{-9}
573 K	-1.20×10^{-9}
Hall coefficient	2.4×10^{-9} $m^3A^{-1}s^{-1}$
Thermoelectric voltage at 298 K	$\approx 5\times10^{-6}$ V/K
Photoelectric work function	3.92 eV
Compressibility	
isothermal	0.0883×10^{-10} m^2/N
adiabatic	0.103×10^{-10}
Speed of sound	12600 m/s
Surface tension at	
1557 K	1.960 J/m^2
1773 K	1.100
Volume contraction on solidification	3%

The *c*-to-*a* ratio of hexagonal-closest-packed beryllium is very small and decreases with temperature [26]. The existence of the body-centered-cubic high-temperature modification was disputed for many years but has now been demonstrated conclusively [24, 27]. The transformation point is slightly below the melting point [24 – 28]. The sum of the heats of fusion and transformation is given as 1641 J/g, which allows a heat of transformation of 284 J/g to be estimated [28]. However, a much higher value, ≈ 837 J/g, was determined directly [18, p. 201].

The *vapor pressure* values in the literature show considerable variation, probably due to the variable formation of oxide and nitride coatings [37]. They were compiled by Hultgren [28]. The high *specific heat* [28] and *thermal conductivity* [7], [10, p. 320], [31] are noteworthy. Low-temperature measurements show that the thermal conductivity perpendicular to the basal plane is greater than that parallel to the basal plane [38].

The coefficient of *linear expansion* is comparatively low. The values given in the literature vary widely, a fact that can be attributed to differences in grain structure of the samples. This is shown by measurements on single crystals parallel and perpendicular to the basal plane [39] and on highly textured material perpendicular and parallel to the pressing direction [40].

The *electrical resistivity* between 40 and 200 K is significantly lower than that of copper or aluminum [41]. Of course, the differences among the reported values are particularly great for electrical properties [11, 22, 31, 36, 42 – 45] because impurities are especially important here, as has been clearly shown by measurements of residual resistance in materials of different degrees of purity [46, 47]. The information on the superconductivity of beryllium is contradictory. Some investigators [22, 48] have found *superconductivity* at temperatures below 11 K, others [49] have found superconductivity only in thin films below transition temperatures between 7 and 9 K, and still others [45, 50, 51] have been unable to find superconductivity at all.

Beryllium is *diamagnetic;* little work [33, 34] has been reported on the magnetic properties. The available measurements on the thermoelectric force show considerable variation, probably a result of the experimental conditions and sample characteristics [7, 38, 52].

The reflexivity of beryllium is high, particularly in the infrared region [53]. Beryllium transmits x rays well because of its low atomic number. Even relatively long-wave gamma rays cause beryllium to emit neutrons, on account of a (γ,n) reaction. Other nuclear reactions of beryllium and its most important nuclear properties [10, pp. 14, 328], [22, 53, 54] are given below and in Table 1:

Magnetic dipole moment	5.9468 A m^2
Electric quadrupole moment	0.02×10^{-24} cm^2
Binding energy of the last neutron	1.664 MeV
Absorption cross section for thermal neutrons (0.025 eV)	
microscopic (σ_a)	0.9 fm^2
macroscopic (Σ_a)	0.0011 cm^{-1}
Scattering cross section for epithermal neutrons	
microscopic (σ_s)	610 fm^2
macroscopic (Σ_s)	0.76 cm^{-1}

The nuclear reactions of beryllium with fast neutrons and hard γ rays do not result in long-lived radioactivity. However, additional neutrons and gaseous reaction products may be produced. For example, bombardment with 10^{22} fast neutrons per cm^2 generated a total of 23 cm^3 of gas per cm^3 of beryllium: 5% of this gas was tritium, the remainder was helium [56]. The gas was initially located at interstitial sites or at dislocations. Heating at 1170 K caused the formation of gas-filled pores at the grain boundaries and within the grains, which resulted in a decrease in density of 20%. The volume increase due to pore formation begins at about 870 K. In addition to the formation of noble gases and tritium, normal radiation damage also occurs as a result of elastic collisions of fast neutrons with beryllium atoms [7], but the effect of this radiation damage can hardly be separated from the effect of the gas formation.

The moderator and reflector properties of beryllium for thermal and epithermal neutrons are also favorable [8, 9].

3. Mechanical Properties

As in the case of other metals and alloys, e.g., aluminum or brass, the term beryllium is understood to mean a large group of different beryllium grades, whose properties differ to a greater or lesser extent as a result of differences in recovery and processing. This is also the reason for the more or less large ranges for many properties. Nevertheless, the principles of the deformation modes are known, which is the bridge to an understanding of the mechanical behavior of polycrystalline beryllium.

Beryllium *deforms* primarily on basal planes and secondarily on first-order prismatic planes. Slip systems with a nonbasal Burgers vector, e.g., second-order pyramidal slip, are difficult to activate, especially at room temperature. The strong anisotropy of flow stresses on the active slip systems is illustrated in Figure 1. The basal plane has a very low surface energy and hence acts as a cleavage plane. One main reason for the *brittleness* of beryllium is the fact that the basal plane acts both as the cleavage plane and the slip plane [21, vol. I, pp. 7 – 114]. Twinning, which is a very important deformation mechanism in several hcp metals, e.g., zinc, titanium, zirco-

Table 1. Properties of beryllium isotopes

Isotope	Half-life	Radiation, MeV	Source
^{6}Be	0.4 s	–	–
^{7}Be	53.5 d	γ 0.453 – 0.485	^{6}Li(d,n) ^{7}Li(p,n) ^{10}B(p,d) ^{10}B(p,α) ^{10}B(d,α,n)
^{8}Be	1.4×10^{-16} s	α 0.055	^{9}Be(γ,n) ^{9}Be(n,2n)
^{9}Be	stable	–	–
^{10}Be	2.7×10^{6} a	β 0.56 – 0.65	^{9}Be(α,p) ^{9}Be(d,p) ^{9}Be(n,γ) ^{10}Be(n,p) ^{13}C(n,α)
^{11}Be	–	–	–
Nuclear reactions of the stable isotope ^{9}Be:			
Gamma-ray bombardment		^{9}Be(γ,n)^{8}Be	
Neutron bombardment		^{9}Be(n,α)^{6}He ^{9}Be(n,γ)^{10}Be ^{9}Be(n,2n)^{8}Be	

nium, does not play a similar role in beryllium. In beryllium, twinning acts more as a crack inducer or enhances crack propagation, rather than to increase the deformability.

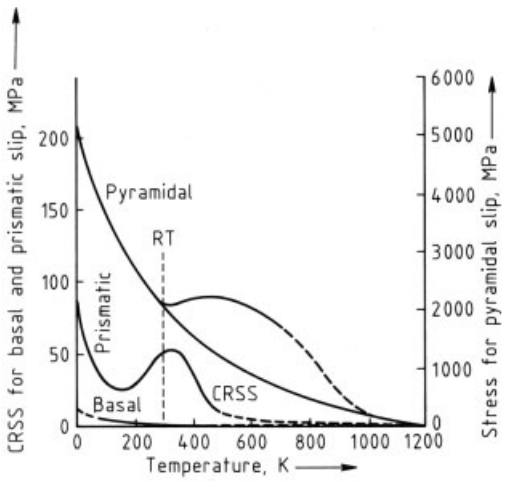

Figure 1. Flow stress for basal, prismatic, and pyramidal *c* + *a* slip versus temperature [21, vol. I, pp. 7 – 114]
CRSS stands for the critical resolved shear stress.

In polycrystalline beryllium brittle fracture occurs even after comparatively low strain at room temperature [11, 15, 57 – 59].

The reason for this tendency toward brittle fracture is the preferred occurrence of basal slip due to the described strong deformational anisotropy, which induces microcracks [10, p. 372], [60 – 62]. The brittleness can be reduced by purification of the material (e.g., to reduce oxide content) and special thermomechanical treatment (upset forging), which results in a microstructure with a very small grain size [63].

The average *modulus of elasticity* of beryllium is unusually high, 320000 MPa, and is strongly anisotropic [64, 65]: 269000 MPa in the basal plane and 365000 MPa perpendicular to the basal plane [28]. As a result the modulus of elasticity of individual polycrystalline specimens shows considerable variation depending on the type of working, and thus on differences in microstructure (Fig. 2 A). The shear modulus decreases somewhat more rapidly than does the modulus of elasticity as the temperature increases. The Poisson ratio shows a corresponding increase with temperature (Fig. 2 B).

The deformational anisotropy and susceptibility to brittle fracture become lower as the temperature is increased [10, p. 372] so that beryllium is readily workable above 450 K [57]. However, the *elongation* of some commercial grades of beryllium decreases sharply in the temperature range 650 – 900 K, then increas-

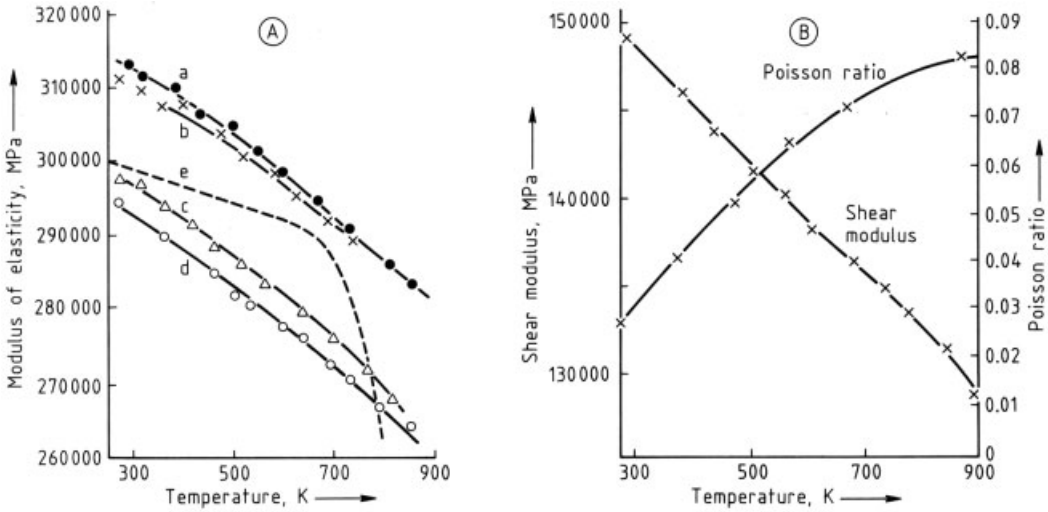

Figure 2. Mechanical properties of beryllium as a function of temperature
A) Modulus of elasticity of various grades of beryllium: a) Pressure sintered [66]; b) Cross-rolled cast material [66]; c) Extruded cast material [66]; d) Extruded P/M (powder-metallurgy) material [66]; e) Pressure sintered, 1 wt% BeO [67]
B) Shear modulus and Poisson ratio of cross-rolled cast beryllium [66]

ing again [15, p. 3], [57, 58] (Fig. 3, Curves a, b, and c). This decrease in elongation at intermediate temperatures is caused by impurities. However, when the concentrations of aluminum, iron, and beryllium oxide are adjusted in a well-defined manner, the hot-working characteristics can be greatly improved by effective heat treatment [15, p. 112], [68, 69] (Fig. 3, Curve e). The preferred occurrence of basal slip during warm working, which is usually performed in the temperature range 670 – 1360 K, also results in a pronounced microstructure. This consists of a preferential orientation of the basal planes, e.g., parallel to the direction of extrusion or parallel to the rolling plane [12, p. 240], [15]. The microstructure strongly limits the basal slip at room temperature in the case of deformation parallel to this preferential direction so that the susceptibility to brittle fracture in this direction of deformation is reduced. Considerable elongation can be attained because of the preferred occurrence of prismatic slip. Perpendicular to the preferred direction of the basal plane the potential elongation is almost equal to zero because the direction of prismatic slip lies in the basal plane.

Improvement of the *forming characteristics* by grain refining during melting and casting is possible only to a limited extent [10, p. 136], [12, p. 55], [15, pp. 193, 237, 246], [58, 70 – 72]. The mechanical properties after powder-metallurgical processing are significantly better than those after casting, and material produced by powder metallurgy is now used almost exclusively.

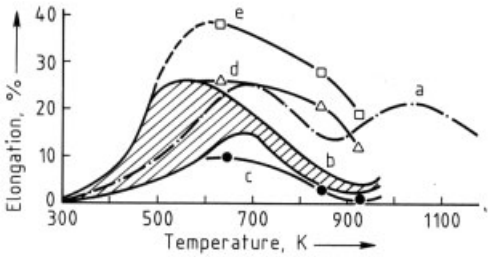

Figure 3. Elongation of various beryllium grades as a function of temperature
a) Cast, extruded [57]; b) Pressure sintered [15, p. 3]; c) Pressure sintered, impurities uncontrolled; d) Pressure sintered, Fe and Al content controlled; e) Pressure sintered, BeO, Al, and Fe content controlled [69]

The susceptibility of beryllium to cleavage fracture is significantly reduced under compressive load so that the deformation values of cast or pressure-sintered material are usually greater than 20%. The buckling strength is comparable to the yield strength; however, the compressive

strength is significantly higher than the tensile strength (Table 2).

Table 2. Strength values, MPa, of extruded beryllium measured in compression test [73]

	Cast		Pressure sintered	
	Room temperature	670 K	Room temperature	670 K
Two-tenths buckling strength	210	179	240	178
Compressive strength	657	460	635	460

4. Chemical Properties

The two 2s electrons of beryllium are valence electrons. The first ionization potential ($Be^0 \rightarrow Be^+ + e$) is 9.32 eV, and the second ($Be^+ \rightarrow Be^{2+} + e$) is 18.21 eV [74]. Although beryllium is almost always divalent in its compounds, it is far more similar to aluminum than to magnesium or calcium [7, 10, 11, 37, 74, 75]. For example, like aluminum, it is amphoteric, and in humid air and water vapor forms a strongly adhering surface layer of oxide that prevents further oxidation up to about 600 °C [37, 76]. Above 600 °C the oxide layer thickens, but oxidation that destroys the metal by break away does not occur below 700 °C.

The affinity of beryllium for oxygen is very great

$$\mathrm{Be\,(s)} + 1/2\mathrm{O_2\,(g)} \rightarrow \mathrm{BeO\,(s)}$$

$$\Delta G^0_T(\mathrm{J/g}) = -66469 + 10.5T \pm 4.7$$

and beryllium is an excellent reducing agent. At temperatures $> 900\,°C$ it reacts violently with nitrogen or ammonia to form beryllium nitride, Be_3N_2. However, beryllium does not react with hydrogen, even at high temperatures, because the formations of the hydrides BeH_2 and BeH, which are difficult to prepare, are endothermic, and the hydrides rapidly decompose at only slightly elevated temperatures. Below 500 – 600 °C beryllium is not attacked by dry carbon dioxide and is attacked only very slowly by moist carbon dioxide. Beryllium powder reacts with fluorine at room temperature, and at elevated temperatures it reacts with chlorine, bromine, and iodine and with sulfur, selenium, and tellurium vapor; in each case, beryllium burns with a flame.

The adhering oxide film protects beryllium from attack by both cold and hot water. It also protects cold beryllium from attack by oxidizing acids [76]. The stability of the film can be increased by addition of dichromate to the water to form a protective chromate layer as in the case of aluminum. Beryllium dissolves in dilute nonoxidizing mineral acids, accompanied by hydrogen evolution and salt formation. In accord with its amphoteric nature it is also attacked by aqueous hydroxide accompanied by hydrogen evolution and beryllate formation. Since most beryllium compounds have a highly exothermic heat of formation, beryllium reduces the salts and the borates and silicates of many metals. For example, the only halides that are stable towards beryllium are those of the alkali metals and magnesium; all others are reduced by beryllium. Molten alkali-metal hydroxides react explosively with beryllium.

Beryllium forms several intermetallic phases with carbon; all transition metals; the alkaline-earth metals magnesium, calcium, and strontium; and the metalloids boron, arsenic, tellurium, etc. [25]. Some of these phases have extensive homogeneity ranges and high melting points [77 – 79]. With very few exceptions, the solubility of metals in solid beryllium and vice versa is very low, even at high temperatures. The binary and ternary beryllium systems were reviewed [21, vol. I, pp. 235 – 305].

Beryllium is very stable towards liquid lithium, sodium, potassium, zinc, magnesium, cadmium, mercury, gallium, indium, tin, lead, antimony, and bismuth. Some of these metals do not dissolve beryllium at all, even at high temperatures, so long as oxygen is not present. If the melts contain metal oxides, the metal oxides are reduced by beryllium with the formation of beryllium oxide. The layer of BeO formed in this way is not adherent and is removed by the melt, so that a fresh surface of beryllium is always available for reduction.

Beryllium is very reactive in the liquid state, reacting with most oxides, nitrides, sulfides, and carbides, including those of magnesium, calcium, aluminum, titanium, and zirconium.

5. Resources, Raw Material

Beryllium is a rare element; its abundance in the earth's crust is about 6 ppm. However, it occurs in concentrated form in various minerals. The total exploitable world reserves of beryllium are estimated at 200000 tons.

The first beryllium mineral commercially exploited was *beryl* [*1302-52-9*], 3 BeO · Al_2O_3 · 6 SiO_2, *mp* 1650 °C, ϱ 2.67 – 2.76 g/cm^3, hardness of 7.5 – 8 on the Mohs' scale, optically negative with weak double refraction, n = 1.57 – 1.60. It crystallizes in hexagonal prisms, which are often very large in size; indeed, crystals weighing several tons have been discovered. It is found in mica schist, granite, pegmatite, and argillite. The crystals often show signs of weathering, and only rarely do they have the ideal composition. Beryl contains about 11% beryllium oxide (4% Be) and is often obtained as a byproduct of feldspar quarrying. The largest deposits are found in Brazil, Argentina, India, Mozambique, Madagascar, Uganda, Zimbabwe, South Africa, and the Former States of USSR. In addition to the aluminum oxide and silicon dioxide, principal impurities in the ores include alkali metals, alkaline-earth metals, iron, manganese, and phosphorus.

Until the late 1960s, beryl was the only beryllium mineral having industrial importance. In 1969 Brush Wellmann Inc. began the processing of *bertrandite* [*12161-82-9*] ores in Utah. This is presently the most important commercial beryllium mineral. Its formula is 4 BeO · SiO_2 · H_2O, but ores contain only 0.2 – 0.35% Be [19, p. 21]. Chrysoberyl, BeO · Al_2O_3; phenakite, 2 BeO · SiO_2; euclase, $BeAlSiO_4(OH)$; and the precious varieties of beryl, emerald and aquamarine, are highly prized as gemstones. About 40 other beryllium-bearing minerals are known.

6. Production

Since the density of beryl is close to that of quartz, concentration by density is not possible. Hand sorting has been the chief means of separation. Flotation is successful only for a few ores [80]. An automatic machine, the Beryl Picker, has been used [81]: the ore is first γ-irradiated to render the beryllium atoms radioactive, $^9Be(\gamma,n)^8Be$. The beryl crystals are then selected out by a sorter guided by a neutron counter [82].

The recovery of beryllium from both beryl and bertrandite includes several stages. The ores are first converted to an acid-soluble form by fusion. Complex chemical processes are then used to obtain comparatively pure beryllium hydroxide or oxide and then beryllium chloride or fluoride. These halogenides are reduced to metallic beryllium with other metals or by melt electrolysis. The beryllium metal obtained is subjected to one or more refining processes and then to further treatment by powder metallurgy or in some cases fusion metallurgy [7, 10 – 12].

6.1. Processing of Beryl

Beryl is treated mainly by the sulfate and fluoride processes. The chloride process is a direct, inexpensive process for recovering beryllium chloride for electrolysis [10, p. 102]. Other processes are not used because of problems with reactions and apparatus [11, 83].

Sulfate Process. First the beryl, which is only slightly soluble in sulfuric acid even under such extreme conditions as heating for several hours in an autoclave at 400 °C, is subjected to either an alkali or a heat treatment.

The *alkali treatment* corresponds to the usual silicate treatment of analytical chemistry: finely ground beryl is heated until fusion or sintered below the melting point with a sufficient quantity of alkaline flux. Suitable alkalies include hydroxides or carbonates of sodium, potassium, and calcium; mixtures of these carbonates; calcium oxide; borax; lead chloride; and sodium sulfate and charcoal. The mechanism of the fusion is still not fully understood [84]. The ratio of flux to beryl depends on the operating conditions, especially the temperature: the higher the temperature the less flux. For alkali treatments involving molten material, gas-heated or oil-heated rotary furnaces or blast furnaces are suitable. Rotary kilns, muffle furnaces, or tunnel kilns are used for sintering.

In the *heat treatment* the beryl is melted (*mp* 1650 °C) without additives and then quenched in water. After this treatment about 50 – 60% of the beryl has a greater solubility in sulfuric acid.

The rest of the beryllium oxide forms a solid solution with silicon dioxide that is not attacked by sulfuric acid. At 900 °C this solid solution separates into its components, and a heat treatment at this temperature produces a free beryllium oxide that is soluble in sulfuric acid. Such a two-stage heat treatment renders a total of 90 – 95% of the beryl soluble.

The heat-treated beryl is extracted with hot concentrated sulfuric acid, whereas alkali-treated beryl is extracted with cold sulfuric acid. In the second case there is no heating cost, but the total consumption of acid is greater because of the alkali added. The liquor is diluted with water, and the insoluble silicon dioxide separated by filtration. At this point the filtrate contains not only beryllium and aluminum sulfate but also considerable quantities of iron sulfate and smaller amounts of other impurities, all of which must be removed before precipitation of beryllium hydroxide. Examples of the many separation methods [11] include the alum separation, the ammonium carbonate separation, and the chelate separation.

In the *alum separation* excess ammonium or potassium sulfate is added to the hot sulfate solution. Upon cooling, all the aluminum precipitates as alum and can be separated from the beryllium sulfate solution in a centrifuge. The filtrate is diluted and then oxidized with potassium permanganate. It is then heated to the boiling point, and calcium carbonate slurry is added until a pH of 3.8 – 4.2 is reached. This causes iron and manganese to precipitate as the hydroxides and calcium to precipitate as calcium sulfate dihydrate. The beryllium hydroxide is then precipitated by addition of ammonia.

In the *ammonium carbonate separation* the liquor obtained from the fusion treatment is treated with saturated ammonium carbonate solution to precipitate aluminum, iron, and other heavy metals as hydroxides, which are then separated by filtration. The filtrate, which contains the soluble beryllium hydroxide, is concentrated by boiling to precipitate readily filterable sparingly soluble basic beryllium carbonate, which is converted to low-density beryllium oxide by calcining at low temperatures.

In the *chelate separation* iron and aluminum are chelated, e.g., with ethylenediaminetetraacetic acid. The beryllium can be precipitated as hydroxide, while these complexes remain in solution.

If the separation process does not lead directly to fine-grained, readily filterable beryllium hydroxide or oxide, the fusion liquor is adjusted to pH 7.5 with ammonia or sodium hydroxide and then heated to 95 – 98 °C. A crystalline, readily filterable beryllium hydroxide is obtained under these conditions. At lower pH gelatinous hydroxide is obtained; at higher pH, amorphous hydroxide is obtained [11]. If the temperature is too low, the crystalline hydroxide forms too slowly.

Brush Wellman Inc., the world's largest beryllium producer, has fully automated its process, called the *caustic process*. This process uses a heat fusion treatment and is characterized by low consumption of sulfuric acid, reliable separation of iron and other heavy metal impurities by combined alum and chelate separation, and recovery of an easily filterable beryllium hydroxide by hydrolysis of the sodium beryllate [10, p. 71].

Fluoride Process. In the fluoride process beryl is melted or sintered at ≈ 700 °C with sodium hexafluorosilicate. Only the beryllium oxide is converted to a water-soluble salt; aluminum oxide, silicon dioxide, and other impurities are essentially unattacked [83, 85]:

$$2(3BeO \cdot Al_2O_3 \cdot 6SiO_2) + 6Na_2(SiF_6) \rightarrow 6Na_2BeF_4 + 2Al_2O_3 + 15SiO_2 + 3SiF_4$$

The formation of other beryllium salts ($NaBeF_3$, BeF_2) is not observed. Water-soluble Na_3AlF_6 may also be formed in small amounts.

Other suitable extracting agents are hydrogen fluoride at 630 °C, molten fluorides, other fluorosilicates, or silicon tetrafluoride [86]. Rotary kilns are used for reaction with the gases; the solid mixtures are briquetted and sintered in muffle furnaces or tunnel kilns.

The sodium hexafluorosilicate needed for the fluoride process is relatively expensive. However, consumption can be greatly reduced by adding sodium fluoride [87] or sodium carbonate [88], which traps and thus "recycles" the liberated silicon tetrafluoride.

The sodium fluoride remaining in solution after precipitation of the beryllium hydroxide can

also be recovered. Sodium hexafluoroferrate(III) is obtained by precipitation with iron(III) sulfate [89] and can be used instead of sodium hexafluorosilicate.

The reaction product must be leached with water at room temperature because silicon dioxide is attacked by fluoride solutions at elevated temperatures. On the other hand, the leaching must be performed as rapidly as possible because otherwise the beryllium salt that is already dissolved will precipitate [11] and in this form it is sparingly soluble, unlike the form which is present in the original fusion product. In order to obtain beryllium hydroxide the solution is made alkaline with NaOH. Beryllium hydroxide precipitates when the solution is diluted with water and then boiled. The fluoride process is being tested in a pilot plant by Bhabba Atomic Research Centre, India [90].

Chloride Process. At temperatures above 630 °C the components of beryl can be extracted in a stream of chlorine; reducing conditions greatly enhance the yield:

$$3BeO \cdot Al_2O_3 \cdot 6SiO_2 + 18Cl_2 + 18C \rightarrow 3BeCl_2 + 2AlCl_3 + 6SiCl_4 + 18CO$$

In practice a mixture of pulverized beryl and carbon is either heated near 800 °C in a stream of chlorine [91, 92] or melted in an electric-arc furnace. In the second case the carbides that form are chlorinated [93, 94]. The reaction can also be carried out with gases that contain chlorine, such as hydrogen chloride [83], carbon tetrachloride [95], sulfur chloride [83], or phosgene [96, 97]. The volatile chlorides are separated by fractional condensation or liquid – liquid extraction with sulfur dichloride, sulfuryl chloride, phosphorus trichloride, phosphorus oxychloride, boron chloride, carbon tetrachloride, or phosgene [83].

Reducing chlorination of the beryllium hydroxide or oxide obtained in the sulfate or fluoride process gives beryllium chloride, the starting material for the electrolytic production of beryllium.

6.2. Processing of Bertrandite

The beryllium-poor bertrandite ores ($\leqq$ 0.5 – 0.8% BeO) discovered in the United States in 1960 cannot be economically smelted by conventional methods. A new process was therefore developed by Brush Wellman Inc., called the *SX-Carbonate process* (Fig. 4) [98, 99]. This process produces a very pure beryllium hydroxide (Table 3) by liquid – liquid extraction with organic phases [74].

Table 3. Typical analyses of beryllium hydroxide obtained by two processes, ppm, with respect to BeO [98, 99]

Element	SX-Carbonate process	Caustic process
Al	200	3 000
Fe	200	1 300
Cr	50	125
Li	70	300
Mg	100	6 000
Mn	10	900
Ni	10	30
Na	500	10 000
Ca	200	800
Cu	5	25
Si	100	1 500
Zn	20	700

Unlike beryl, bertrandite is soluble in sulfuric acid without pretreatment. The insoluble residue (SiO_2, etc.) and the aluminum content, which precipitates as an aluminum alum, are removed. The sulfate solution, pH = 0.5, then contains the beryllium and also aluminum, calcium, iron, sodium, and traces of other elements.

The sulfate solution is extracted with a 0.3 m solution of ammonium di-2-ethylhexyl phosphate/di-2-ethylhexyl phosphoric acid/kerosine (DAP) in a five-stage mixer at ≈ 50 °C. In the first mixer the pH is about 2, and after the last stage it is about 5.5. The acid strength is adjusted in a well-defined manner in each stage by means of the ratio of the ammoniacal to the acid component of the organic phase. Beryllium, aluminum, and iron enter the organic phase almost quantitatively, while the other impurities remain in the sulfate solution. Following separation of the two liquid phases, the metal ions are removed from the organic compounds by a 3.0 – 3.5 m ammonium carbonate solution in a two-stage mixer. This results in the formation of a basic beryllium carbonate solution, which is actually a slurry since it contains precipitated aluminum hydroxide (the ABC slurry). The organic phase, which is now free of metal ions, is returned to the process after purification. Iron is

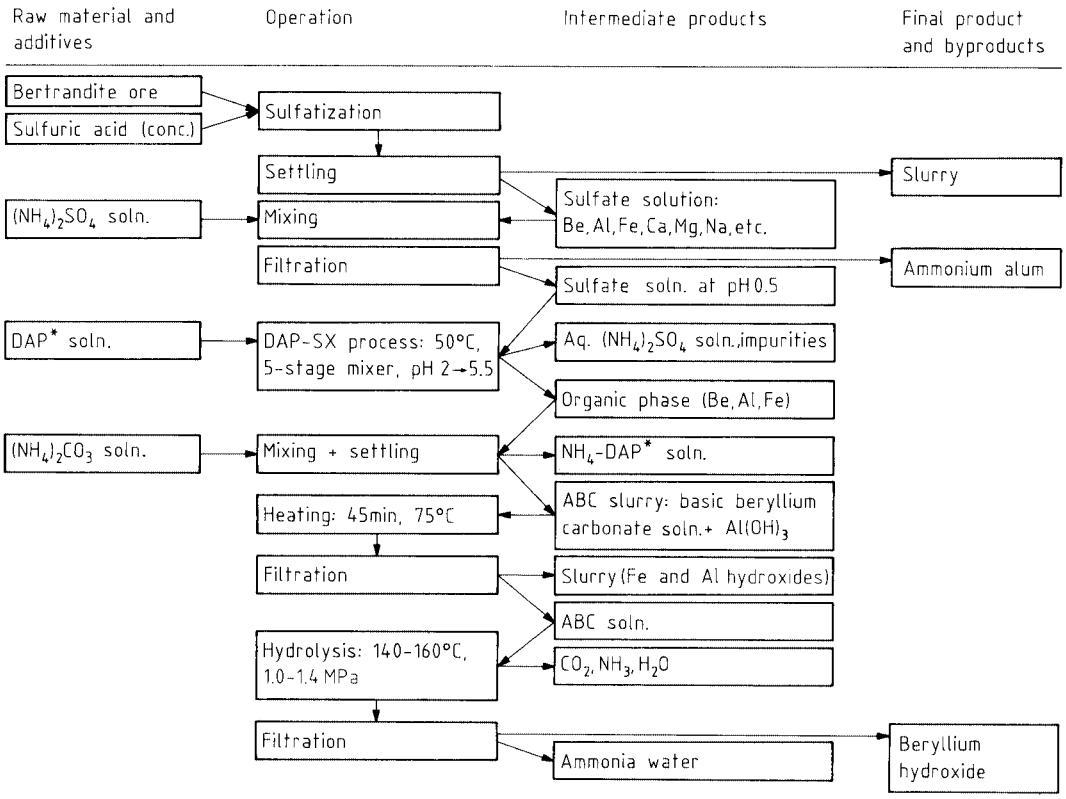

Figure 4. Flow diagram of the SX-Carbonate process (Brush Wellman Inc., USA) [98, 99]
DAP is the abbreviation for the solution ammonium di-2-ethylhexyl phosphate – di-2-ethylhexyl phosphoric acid – kerosine.

precipitated from the ABC slurry by heat treatment, 45 min at 85 °C, and filtered off with the aluminum hydroxide. The beryllium remains in solution as basic carbonate. The specified time must be observed as closely as possible: if the heat treatment is too short, iron is only partially precipitated. If it is too long, the beryllium loss will be too great. Even so, about 2% of the beryllium is coprecipitated and about 4 – 6% is contained in the hydroxide filter cake. Half of the beryllium can be separated by careful washing. The remainder is recovered by dissolving the precipitate in sulfuric acid, extracting the iron with an organic phase, and recycling the beryllium- and aluminum-containing sulfate solution to the process before the alum precipitation.

The basic beryllium carbonate solution separated from the ABC slurry by filtration is decomposed in an autoclave (140 – 160 °C, 1.0 – 1.4 MPa) into ammonia, carbon dioxide, and a beryllium hydroxide that is easily filtered off and that is quite pure (Table 3). The ammonia and carbon dioxide are used to produce ammonium carbonate, which is returned to the process. The ABC slurry can be treated additionally to further reduce the concentration of impurities, e.g., to reduce the iron content to less than 50 ppm. For this reason beryl ores are today also treated by the SX-Carbonate process, following the usual heat treatment.

6.3. Production of Reducible Compounds

The most important metal production processes are based on reduction of beryllium fluoride by magnesium and beryllium chloride by melt electrolysis. Therefore, the beryllium hydroxide or oxide obtained from the fusion process must be converted into one of these halides.

Beryllium Fluoride. First beryllium hydroxide is dissolved in an ammonium hydrogen fluoride solution. The resulting ammonium tetrafluoroberyllate is dissociated to such a small extend in the solution that the pH can be varied over a wide range without beryllium hydroxide precipitating. Impurities can thus be precipitated relatively easily as hydroxides. Aluminum is precipitated as hydroxide by increasing the pH to 8.3. Chromium and manganese are precipitated by oxidation with lead dioxide, and manganese is precipitated by oxidation with lead dioxide and manganese dioxide. Copper, nickel, and lead are precipitated as sulfides.

Ammonium tetrafluoroberyllate is freely soluble in water. When concentrated by evaporation, it crystallizes without water of hydration. Above 130 °C it dissociates into ammonium fluoride and beryllium fluoride. Between 900 and 1100 °C ammonium tetrafluoroberyllate dissociates into gaseous beryllium fluoride and gaseous ammonium fluoride. The latter can be recycled to the process as ammonium hydrogen fluoride by dissolution in aqueous hydrogen fluoride. Cooling produces beryllium fluoride as vitreous grains [10, p. 71], [100].

Beryllium Chloride. Beryllium chloride can be prepared either directly from beryl by the chloride process or by chlorination of beryllium oxide under reducing conditions [101].

Beryllium chloride is especially well suited for purification by distillation in a stream of hydrogen [10, p. 102] and fractional condensation [102]. The significantly lower-boiling chlorides of aluminum, silicon, and iron(III) can be separated by careful temperature control. Iron(II) chloride, which is reduced by hydrogen, stays in the residue.

6.4. Metal Production

The chief difficulties involved in the production of beryllium metal are the reactivity and high melting point of the metal and the extreme stability of the oxide. Of the many possible methods of producing beryllium [10, 11, 84] two are used in industry presently: fusion electrolysis and reduction of halides by metals.

Only calcium is able to reduce the extremely stable beryllium oxide directly under standard conditions. However, calcium forms the very stable intermetallic compound CaBe [103]. Reduction of BeO with titanium, zirconium [104], or carbon [105] at high temperatures is theoretically possible: the volatile beryllium would have to be distilled off from the reaction zone immediately. However, this method is not used.

Reduction with Metals. The most important industrial process is the reduction of beryllium fluoride with magnesium [10, p. 71], [106 – 109], the process presently used by Brush Wellman Inc.

The reaction is started by heating a mixture of relatively coarse-grained beryllium fluoride and magnesium in a graphite crucible. At a temperature of about 1300 °C the reaction produces a mixture of beryllium pebbles and magnesium fluoride. The reaction cannot be performed without external heating because the heat of reaction is not great and because the beryllium fluoride must be processed in the form of relatively large particles due to its toxicity and hygroscopicity. Normally about 70% of the stoichiometric quantity of magnesium is used. In this way a readily water-soluble beryllium magnesium fluoride slag, easily separated from the coarse-grained beryllium, is produced. The unreacted beryllium fluoride is separated from the magnesium fluoride and returned to the process.

The reaction of beryllium chloride with magnesium or sodium or the reaction of beryllium fluoride with sodium cannot be carried out in the same way. Sodium is more difficult to work with because of its low boiling point and its reaction with graphite. Beryllium chloride (*bp* 520 °C) is too volatile. The reduction of beryllium chloride with sodium in the vapor phase at 250 °C is the only process developed to the production stage [110 – 112]. However, this process has no economic importance.

Fusion Electrolysis. All practical electrolytic methods of production are based on the decomposition of beryllium fluoride, beryllium oxide, or beryllium chloride mixed with halides of the alkali metals or alkaline-earth metals [83]. These main group I and II halides are necessary to form a stable melt electrolyte with good conducting properties — the molten beryllium halides are covalently bonded and thus poor electrical conductors. In addition, only beryllium fluoride has a sufficiently large

interval between melting point and boiling point at atmospheric pressure.

The electrolysis of the fluoride was developed by A. Stock, H. Goldschmidt [113], Siemens & Halske [114], and the Beryllium Research Society. The electrolysis was carried out above the melting point of beryllium, at 1290 – 1400 °C. These methods and the similar methods of J. Dickinson [115] and A. C. Vivian [116] are now obsolete.

The electrolysis of beryllium chloride [117] can be carried out at temperatures so low that the metal neither melts nor oxidizes. The beryllium is obtained as solid flakes, which are separated by washing out the electrolyte.

Mixtures of approximately equal parts of beryllium chloride and sodium chloride are electrolyzed at ≈ 350 –400 °C. The composition of the electrolyte is held as constant as possible. The apparatus consists of two identical nickel vessels hanging in electric furnaces. The crucible covers are fitted with the connections. One of the covers contains the anode and can be placed alternately on each crucible; the crucibles themselves serve as cathodes. During electrolysis (4 – 6 V, 0.08 A per cm^2 of cathode surface area) the metal is deposited as flakes on the cathode wall. When the beryllium chloride concentration in the melt has fallen to 45 wt%, the melt is siphoned into the second crucible, in which beryllium chloride and sodium chloride have previously been melted down in quantities corresponding to the quantities consumed by the electrolysis in the first crucible. In this way the bath is again given its original composition. The crucible covers are then exchanged, and electrolysis is started in the second crucible. The chlorine generated at the anode during electrolysis is collected and used in the production of beryllium chloride.

The equipment for this process was developed by Degussa. Since the Second World War this process has been used by Pechiney Cie. Chloride electrolysis was also used in France by Trefimetaux G.P., a joint subsidiary of Pechiney Cie. and Kawecki Berylco Ind. Inc., and by NGK Insulators in Japan. It was used in slightly modified form by Murex Ltd., England [10]. It was also used in the Soviet Union [118].

6.5. Refining and Further Processing

The as-reduced beryllium pebbles or flakes still contain many impurities and have to be refined before they are used to fabricate structural pieces. The main impurities in electrolytic beryllium are sodium and chlorine. The main impurities in beryllium from the magnesium reduction process are magnesium and magnesium fluoride. Other impurities include beryllium oxide, carbon, and metals, the most important being aluminum, iron, and silicon. Typical analyses of magnesium-reduced beryllium and electrolytic beryllium (CR flakes, Pechiney Cie.) are given in Table 4. Some of the impurities have adverse effects on the mechanical properties and corrosion of beryllium or are reactor poisons. Commercial grades of beryllium are refined exclusively by vacuum melting in beryllium oxide or magnesium oxide crucibles and casting in graphite ingot molds. The ingots weigh ≈ 200 kg. Their surfaces are machined (skinned) to remove the high carbon content material adjacent to the mold wall.

Table 4. Composition of beryllium produced by various methods, ppm [12]

Element	Mg-reduced, coarse-grained	Electrolysis	
		CR flakes	SR flakes
O	–	2300	500 – 800
Al	400	<200	60
Fe	500	300	50
Ni	100	150	30
Si	50	75	30
B	–	2	2
Cr	100	20	5
Mn	100	60	30
Ca	–	200	30
Mg	15000	50	30
Na	–	300	–
C	300	300	150
F	20000	–	–
Cl	–	1000	700
Be, wt%	96.0	96.5	99.87

The melting of magnesium-reduced beryllium in a high vacuum produces a degree of purity comparable to that of electrolytic beryllium. Melting the electrolytic flakes in a vacuum further reduces the content of halides and low-boiling metals. A very pure grade of beryllium, particularly with respect to the content of oxide, aluminum, iron, silicon, carbon, and halides, can be produced by electrolytic purification (SR

flakes, Pechiney Cie.). Other methods, such as vacuum distillation [10, 11, 18, 119 – 123], zone melting [18, 47, 124, 125], or liquid-metal extraction [18] have been used in the laboratory only.

Melting and Casting. The main purpose of melting and casting of beryllium is purification (see above); the production of semifinished goods is of secondary importance [12, p. 55]. The reason for this is that parts produced from castings have significantly poorer properties, even after working, than parts made by powder metallurgy. Cast flaws due to gas evolution during solidification and transcrystalline microcracks [10, p. 136], [15, p. 677] caused by cooling stresses cannot be prevented, even though they can be greatly reduced by mold design and construction [126]. Another significant disadvantage is the coarse-grained texture that develops during solidification, resulting in poor mechanical properties not improvable by thermomechanical treatment [10, p. 136], [12, p. 55], [15, pp. 193, 237, 246], [71, 72]. Fairly often crystals several centimeters in length are observed after induction melting [12, p. 55], and often crystals several millimeters in length are observed after electric-arc melting [15, pp. 677, 687], [126]. No effective grain-reducing agent is known [126].

Powder Manufacture. Beryllium was first worked by powder metallurgy in 1946. Since then many methods of powder production and compaction capable of producing ductile beryllium parts with good high-temperature strength have been studied and tested [12], [21, vol. 2, pp. 13 – 29], [127]. However, the concentration of impurities is increased during powder-metallurgical working. Atmospheric oxygen and nitrogen, as well as iron and carbon from the mechanical working of the material into powders, are especially important. Above all, there is an increase in the oxide content as the grain size is decreased, i.e., as the specific surface increases. On the other hand, the oxide is desirable in certain circumstances because it increases strength and inhibits grain growth.

Powder manufacturing starts with cast ingots of inherently coarse grain size characterized by low ductility and resistance to fracture. These ingots are chipped with a lathe. The equipment is heavy duty: a multipoint tool can chip a 65-kg ingot in a single pass. These relatively coarse chips must be ground, or otherwise comminuted to finer particle sizes.

The first milling method commercially employed was *rotary grinding* between one stationary water-cooled and one rotating beryllium disk. The coarse chips were fed through a central hole in the stationary disk, and the fine powder emerged peripherally. The mill was enclosed into a sealed system to minimize contamination of the powder and to protect the environment from the toxic beryllium dust. This method was the standard for many years. Almost everywhere it has now been replaced by the impact attrition mill.

Within the *impact attrition mill* the coarse powder is fed into a gas stream, accelerated through a nozzle and impacted against a beryllium target. The morphology of the powder particles comminuted by this process is more spherical and less flat than disk-milled powder. Therefore, the mechanical properties of impact-ground powder compacts are more uniform. Because of the improved isotropy and a lower amount of inclusions and impurities the elongations of beryllium grades produced from impact-attrition material are higher.

Another commercial production process is *ball-milling*. It is used for high-oxide ultrafine powder ($\approx 5\ \mu m$), which is made especially for instrumentation uses. The high microyield strength specified for such qualities is provided by the fine particle size and the high oxide content that strengthen the grain boundaries within the unpacked material. Because ball-milled powders are flat, like the particles produced by rotary grinding, they are difficult to handle and cause anisotropic physical and mechanical properties in the compacted beryllium billets.

Other powder production methods, like fluid-energy milling, disintegration of beryllium melts by gas atomization, rotating electrode process, or centrifuged shot-casting, have been used only in the laboratory [21, vol. 2, pp. 13 – 29]. Nevertheless, some of these processes may be used to produce commercial beryllium powder within the foreseeable future.

Powder Consolidation. *Vacuum hot pressing* of beryllium powder in graphite or steel dies

is by far the most important commercial consolidation process. The powder is either dynamically compacted in the die or prepressed under cold isostatic conditions. After the die has been heated to 1000 – 1200 °C, a pressure of 1.4 –10 MPa is applied. In this way billets up to 1.80 m in diameter and 6 t in weight, with relative density of nearly 100%, have been manufactured.

Hot isostatic pressing is used to produce more complicated parts with improved isotropic properties. The powders are either directly consolidated in a one-step process or preconsolidated. In preconsolidation the powder is prepressed at 210 – 700 MPa under cold isostatic conditions in rubber containers, the cold compact degassed at 650 °C, and the compact pressed isostatically for 2 – 3 h at 70 MPa and ≈ 760 °C in evacuated stainless steel containers [128].

The combination of *cold pressing and sintering* has been applied in the production of special parts, e.g., aircraft brake segments. Preforms may be sintered using mandrels to limit distortion and give better shape control. Sintering is usually carried out in a vacuum, which ensures better consolidation and higher densities than those achieved in argon.

The hot-pressed ingots are worked into semifinished goods or finished parts by machining, rolling (760 – 870 °C), extrusion (900 – 1090 °C), drawing (400 – 430 °C), deep-drawing (710 –760 °C), forging, and compression (400 – 430 °C) [13, 14].

During hot deformation the beryllium is usually enclosed in a steel container. Working with such "canned" material greatly simplifies the operation because the workpieces need not be in a vacuum and need to be heated only slightly or not at all. The steel cover reduces oxidation and surface damage, improves lubrication, and produces a uniform temperature distribution in the workpiece. Although hot-deformed parts are usually less expensive than machined parts, they have highly anisotropic properties, with the favorable values in the direction of material flow during deformation and the less favorable values perpendicular to this direction. Cross-rolling produces uniform properties, e.g., in the sheet plane. Other processes used only in special cases are slip casting [21, vol. 2, pp. 13 – 29], [127, 129, 130], spark sintering [131], sinter forging [12, p. 102], [132], plasma injection [133], and isostatic hot-pressing [128].

Production of Single Crystals. The first comprehensive work on beryllium single crystals was done in the early 1950s [10, p. 372], [134 – 137]. These crystals were used mainly for investigations of the mechanical and cleavage fracture behavior of beryllium and were produced only in the laboratory. They were grown by slow solidification from the melt in a crucible [10, p. 372], [137], direct crystallization from vapor [138], and crucibleless zone melting [134, 139 – 145]. Efforts to produce beryllium single crystals using the Bridgeman method were unsuccessful [146].

During the late 1970s it was recognized that beryllium would be the most efficient material for neutron monochromators, both because of its excellent nuclear properties and its excellent crystallographic properties [147, 148]. Many attempts to obtain beryllium single crystals with monochromator qualities failed [146, 149 – 152].

Crystals with a diameter of 15 mm having a proper substructure can be grown by crucibleless zone melting at a rate of 1 mm/min. The application of the necking technique is inevitable. A modified double-ellipsoid mirror furnace, originally developed for crystal growth experiments during spacelab missions, is used [153]. The mosaic spread of the as-grown beryllium crystals can be increased from a few minutes to any desired value by plastic deformation on prismatic planes [154].

7. Environmental Protection

Because of the toxicity of beryllium vapor and dust, all operations should be carried out in properly ventilated rooms, and with vented equipment. Source exhaust has been found to be most effective [155 – 157]. In the presence of poorly controlled high workplace concentrations, fine-dust masks with filters of the specific safety level must be worn. The following limits have been recommended by the U.S. Atomic Energy Commission and the American Conference of Governmental Industrial Hygienists as a guideline for controlling beryllium hazards:

1) In-plant atmospheric concentration should not exceed 2 $\mu g/m^3$ throughout an 8-h working day.

2) A brief exposure should not exceed 25 μg/m^3.
3) In the neighborhood of a plant handling beryllium the average monthly concentration should not exceed 0.01 μg/m^3.

In order to meet these safety regulations the plant exhaust air has to be efficiently filtered before it is discharged into the atmosphere. All persons employed in processing areas with high risk potential for beryllium dust exposure are advised to wear shop-laundered work clothes to prevent neighborhood contamination. Care must also be exercised in disposal of wastes from beryllium facilities: securely packed solid wastes are disposed by ground burial in concrete trenches. Liquid wastes are chemically treated to remove the toxic substances and to bring down the beryllium concentration to a safe level of approximately 1 ppm before discharge into public sewers.

In order to monitor the workplace concentration of 2 μg/m^3, dust collectors must have suction speeds between 0.02 and 2 m^3/min, depending on the type of sampling [155]. Presently a sufficiently sensitive analytical method for detecting the comparatively low beryllium concentrations is spectral analysis. Metallic beryllium in powder form can burn when ignited by sparks or an external heat source but does not self-ignite or explode. Beryllium oxide and alloys are not combustible. The naturally occurring isotope ^{9}Be is not radioactive.

8. Quality Specifications

To date there is only one company — Brush-Wellman Inc. — in the Western world that produces beryllium, beryllium-rich alloys, and beryllium compounds commercially. Brush-Wellman Inc. produces beryllium from bertrandite by the SX-Carbonate process, reduces beryllium fluoride with magnesium, and refines it by vacuum melting and casting. Different grades are produced by different powder-metallurgy processes and metallurgical treatments, e.g., forging, extruding, rolling, and annealing.

The composition, yield strength $\sigma_{0.2}$, tensile strength σ_F, and failure strain δ of several typical, commercial types of beryllium are compiled in Tables 5 and 6. However, these values guaranteed by the manufacturers are generally exceeded. The vacuum hot-pressed materials of Table 5 are used in nuclear technology, aeronautics, astronautics, and wherever high strength is required. The table shows that the yield strength and tensile strength increase with oxide content, while the elongation is not significantly affected. The brake grade (BG-170) is characterized by greatly reduced hot brittleness (Curve e in Fig. 3) [68, 69]. In the case of the instrument grades I-220A and I-400, the concentration of impurities is used to give a high precision elastic limit (PEL), below which the plastic strain is smaller than 10^{-6} or one microstrain.

Table 5. Manufacturer-guaranteed composition and mechanical properties of various grades of beryllium [158]

	S-65B	S-100E	S-200E	I-70A	I-220A	I-400	BG-170
Chemical composition							
Be, min. %	99.0	98.5	98.0	99.0	98.0	94.0	98.0
BeO, max. %	1.0	1.2	2.0	0.7	2.2	4.2 *	1.2
Al, max. ppm	600	1400	1600	700	1000	1600	500
C, max. ppm	1000	1500	1500	700	1500	2500	1500
Fe, max. ppm	800	1500	1800	1000	1500	2500	1500
Mg, max. ppm	600	800	800	700	800	800	800
Si, max. ppm	600	800	800	700	800	800	500
Others, each, max. ppm	400	400	400	400	400	1000	–
Tensile properties							
σ_F, min. MPa	290	241	276	241	310	344	345
$\sigma_{0.2}$, min. MPa (0.2% offset)	207	186	207	172	241	–	207
Elongation, min. %	3	1	1	2	2	–	1
Microyield, min. MPa	–	–	–	–	34	55	–

* BeO specified is minimum in this instance.

Table 6. Mechanical properties of various semifinished beryllium products [158, 159]

	Plates *	Extruded parts *	Forged parts *	Wire, Ø <0.25 mm
BeO content, wt% max.	2.0	2.0	2.0	0.02
$\sigma_{0.2}$, MPa	350	280	280	760
σ_F, MPa	480	480	450	900
δ, %	10	5	3	1

* Properties parallel to the direction of extrusion or to the rolling plane.

The development of texture in extruded parts and plates results in failure strain values of 5 and 10% in the direction of extrusion and in the rolling plane, respectively (Table 6); perpendicular to this direction the failure strain is < 1%. Forged parts have elongations of 3%, and high-strength wire (Ø < 0.25 mm) has an elongation of 1%.

As an example of the effect of temperature on tensile strength, yield strength, elongation, and necking at failure, the values measured for plate material parallel to the plane of the plate are plotted in Figure 5. The plate material was produced by cross-rolling high-purity isostatically hot-pressed ingot material.

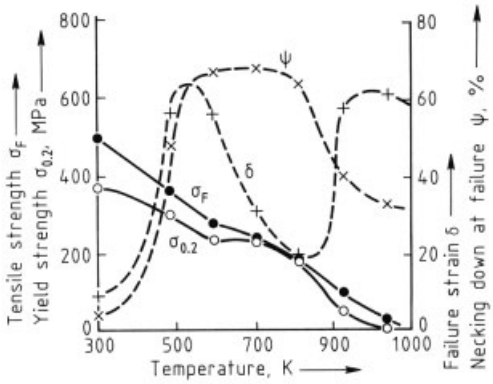

Figure 5. Mechanical properties of high-purity beryllium [160]

The creep rupture strength is usually high [67, 161, 162]. In the case of cross-rolled plates the ten-hour creep rupture strength is equal to the tensile strength up to 800 K and can be significantly increased by carbon or beryllium oxide [37]. The little information that is available on the fatigue strength [161 – 165] (Fig. 6) indicates that beryllium has high fatigue strength and that the dynamic crack propagation is much less critical under alternating stress than under normal stress.

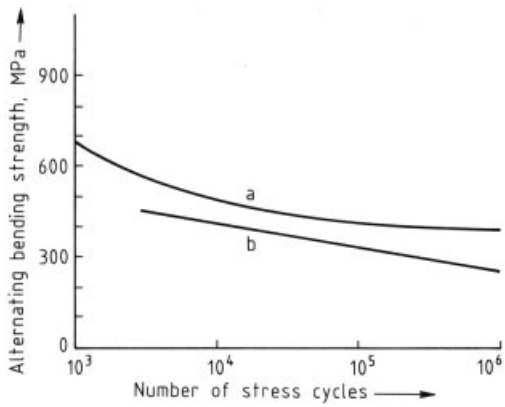

Figure 6. Fatigue strength of pressure-sintered, hot-worked beryllium [161]
a) Extruded and forged; b) Cross rolled

A compilation [163, 164] of the K_{Ic} (a measure of the toughness) values published so far shows that there is a great deal of variation. Most values lie between 10 and 20 MPa/m$^{1/2}$, but values up to 40 MPa/m$^{1/2}$ have been reported. However, the impact strength of notched and unnotched specimens is very low and depends on the type of working (Table 7).

Table 7. Impact strength (Charpy test), 10^{-2} MPa, of beryllium produced in various ways [58]

	Un-notched	Notched
Cast and extruded	0.16	–
Pressure sintered	1.1 – 3.4	0.68 – 1.3
Pressure sintered and extruded	5.6 – 6.1	–
Pressure sintered and rolled	2.7 – 4.1	–

The hardness of high-purity beryllium (Pechiney SR) is plotted in Figure 7 as a function of the mean grain size. The Brinell hardness of commercial pressure-sintered beryllium varies between 100 and 200, depending on the composition of the tested material [52].

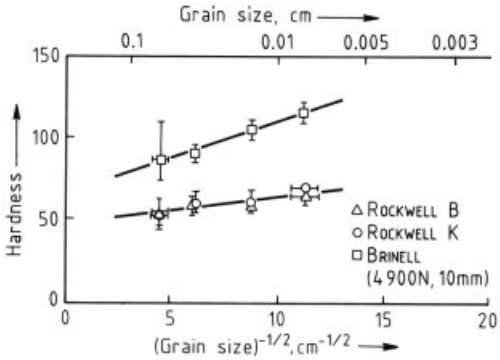

Figure 7. Brinell and Rockwell hardnesses of high-purity beryllium (Pechiney SR) as a function of the mean grain size [74]

9. Beryllium Alloys

Beryllium-rich alloys have little industrial importance. The motivation for industrial research on beryllium alloys [10, p. 555], [12, p. 179], [15, pp. 601, 634], [17, p. 103], [132, 166 – 174] is not great because the solubility of most elements in solid beryllium is low ($\ll$ 1 mol%), and only copper, nickel, cobalt, iron, silver, and platinum have solubilities between 1 and 10 mol %, even at elevated temperatures. In addition, most metals form intermetallic compounds with beryllium [25, 77 – 79], and these intermetallics are usually quite brittle.

The only alloy containing large amounts of beryllium is Lockalloy [169], which is produced by powder metallurgy and contains 38 wt% Al. Aluminum does not form any beryllides. Therefore, this alloy contains beryllium grains embedded in the ductile aluminum phase with a more or less uniform distribution. This combination has a comparatively high modulus of elasticity (E), low density, and comparatively useful deformation behavior at room temperature. In addition, the notch sensitivity is lower than that of commercial beryllium (Table 8) [17, p. 103], [173]. However, Lockalloy has found only limited commercial use.

Composite materials made of beryllium and titanium have been developed for use, e.g., as compressor blades in gas turbines [174]. At present, these materials are the only composite materials that allow definite plastic deformation, are impact resistant, and offer the choice of secondary cold or hot working (Table 8). Furthermore they can be used to a limited extent even at 400 – 500 °C.

Beryllium is used as an alloying component in a number of age-hardening alloys [6] based on copper, nickel, cobalt, and/or iron [10, pp. 24, 49], [175 – 179]. Small amounts of beryllium improve the castability and oxidation resistance of aluminum and magnesium alloys [180 – 187]. These alloys are described under the principal component, → Copper Alloys, etc.

Table 8. Mechanical properties of powder-metallurgy Lockalloy [169] and beryllium – titanium composite materials (extruded powder mixtures) [174]

	ϱ, g/cm^3	σ_F, MPa	E, MPa	δ, %	$\sigma_{0.2}$, MPa
Lockalloy	2.09	386	200000	≈10	255
60 vol % Ti	3.4	1100	175000	5	≈770
40 vol % Ti	2.85	829	215000	2.5	≈580

10. Beryllium Compounds

Most beryllium compounds are important only in the production of beryllium metal, and these aspects are described earlier in this article.

Beryllium oxide [*1304-56-9*], BeO, M_r 25.0122, is obtained directly from beryllium hydroxide by calcination or from the basic carbonate, acetate, or sulfate by ignition [11]. The white, reactive powders differ in grain size, morphology, and impurity content [188]. They are used in the production of beryllium oxide ceramics, some of which have most unusual properties (Table 9). These ceramics are being used to an ever increasing extent [20].

Beryllium oxide is extremely stable [37]. However, its stability depends on its structure and is determined by the content of bound water. The higher the calcination temperature the more inert the beryllium oxide.

Beryl, see Chapter 5 and 6.

Bertrandite, see Chapter 5.

Beryllium nitride [*1304-54-7*], Be_3N_2, M_r 55.0440, is formed by heating beryllium in dry nitrogen in the presence of 2 – 6% hydrogen (700 – 1400 °C) or ammonia (1000 °C) [10, p. 570]. Moldings are produced by nitration of beryllium powder in porous molds by hot pressing the nitride at 1800 – 1900 °C. Relative densities up to almost 100% are achieved in this

way. Beryllium nitride is stable in dry oxygen up to 500 °C, but in the presence of water vapor it reacts at lower temperatures. Beryllium nitride is cubic at room temperature, converting to hexagonal at 1400 °C. Two other nitrides, $Be(N_3)_2$ and BeN, and two beryllium silicon nitrides, Be_4SiN_4 and $BeSiN_4$, are known. These ternary nitrides are very hard, extremely inert, and stable up to 1850 °C [191].

Table 9. Physical properties of beryllium oxide and carbide [7, 8], [10, p. 599], [11], [12, p. 267], [37, 189, 190]

	BeO	Be_2C
Structural type	ZnS	CaF_2
Melting point, °C	2300 – 2570	≈ 2400
Density, g/cm^3	3.01	≈ 2.3
Modulus of elasticity, MPa	370000	320000
Linear coefficient of thermal expansion, K^{-1}	5×10^{-6}	5.8×10^{-6}
Specific heat, J g^{-1} K^{-1}		
60 K	0.029	
292 K	1.00	1.4
1173 K	2.04	
Thermal conductivity, W m^{-1} K^{-1}		
room temp.	264	24
673 K	92	
1073 K	29	
Electrical resistance, Ω m		
room temp.	10^{16}	1.1×10^{-2}
1273 K	8×10^{11}	0.45×10^{-2}
2373 K	8×10^{6}	
Compressive strength, MPa	800	720
Bending strength, MPa	285 – 380	91 – 98
σ_a*, fm^2	0.9	2.3
Σ_a*, cm^{-1}	0.00066	0.0011
σ_s*, fm^2	990	1690
Σ_s*, cm^{-1}	0.72	0.81

* Defined in Chapter 2.

Beryllium carbide [*506-66-1*], Be_2C, M_r 30.0355, ϱ 2.3 g/cm^3, *mp* ≈ 2400 °C, cubic, is obtained by direct reaction of solid or liquid beryllium with carbon or by reaction of beryllium oxide with finely divided carbon at 2080 °C [7], [10, p. 570]. Moldings are produced by hot pressing the carbide powder in graphite molds at 1800 – 1950 °C and 7 – 35 MPa. Beryllium carbide hydrolyzes very readily, even at room temperature, to form beryllium oxide and methane. At high temperatures it is not stable in air. Beryllium carbide vaporizes and decomposes in a vacuum above 2100 °C. The properties given in Table 10 are reference values because the beryllium carbide grades which were tested had indefinite structure. The nuclear properties of beryllium carbide are favorable. Beryllium borocarbide is extremely hard and is formed by melting mixtures of beryllium oxide, boron, and carbon in an electric arc.

Beryllium hydride [*7787-52-2*], BeH_2, M_r 11.028, is a very light (ϱ 0.58 g/cm^3), white powder at room temperature that is difficult to prepare. It dissociates into beryllium and hydrogen at ≈ 130 °C [37]. Beryllium hydride is of interest as a high-energy solid fuel.

Beryllium hydroxide [*13327-32-7*], $Be(OH)_2$, M_r 43.028, is obtained as a white voluminous precipitate on addition of bases to beryllium salt solutions. The fresh precipitate dissolves in sodium and potassium hydroxide and ammonium carbonate solution, but it ages quickly and becomes less and less soluble. The hydroxide also dissolves in aqueous $BeSO_4$ to form basic sulfates.

Basic beryllium carbonate, $BeCO_3 \cdot n\ Be(OH)_2$, $n = 2 – 5$, is formed in the reaction of beryllium salt solutions with alkali-metal or ammonium carbonate solutions. If excess ammonium carbonate is used, then a readily filtered precipitate of variable composition is formed on boiling. This salt is a suitable starting material for the preparation of beryllium salts of all types. Gentle calcining causes ammonia to escape, leaving beryllium oxide.

Beryllium nitrate, $Be(NO_3)_2 \cdot 4\ H_2O$, M_r 205.08, is obtained by dissolving pure hydroxide or gently calcined, pure basic carbonate in nitric acid and concentrating the solution until crystallization takes place. Concentrated solutions are also commercially available. The nitrate is sometimes used in the incandescent gas mantle industry for stiffening mantles.

Beryllium sulfate [*7787-56-6*], $BeSO_4 \cdot 4\ H_2O$, M_r 177.14, is obtained by heating beryllium oxide with sulfuric acid. The tetrahydrate crystallizes from the aqueous solution in well-developed crystals. It can be obtained in very pure form by this method. The tetrahydrate is stable in air, can be dehydrated at about 400 °C, and decomposes above 600 °C to sulfur trioxide. The aqueous solution can dissolve considerable quantities of BeO to form basic sulfate. The sulfate is important chiefly for the recovery of beryllium.

Table 10. Properties of some beryllides [193]

Beryllide	CAS Registry Number	*mp*, K	Oxidation in air, 10^{-3} mm in 100 h		Coeff. of thermal exp., 10^{-6} K^{-1}	Thermal conductivity, W m^{-1} K^{-1}	Bending strength, MPa			Modulus of elasticity, 10^4 MPa		Strain at failure, %			Compressive strength, MPa	
			1640 K	1750 K	RT – 1790 K	1750 K	RT	1640 K	1790 K	RT	1640 K	RT	1640 K	1790 K	RT	1640 K
$NbBe_{12}$	[*12010-12-7*]	1960	22	50	16.8	107.5	152	276	152	31.4	17.2	0.1	1.0	2.4	1380	551
Nb_2Be_{17}	[*12010-34-3*]	1980	15	47.5	15.9	112	214	434	248	29.6	17.2	0.1	2.0	~7	–	–
Nb_2Be_{19}		~1980	–	–	–	–	206	448	186	–	–	0.1	2.0	3.0	–	–
$TaBe_{12}$	[*12010-13-8*]	2130	12.5	22	15.1	123	214	296	179	31.0	13.8	0.1	1.1	2.6	1030	–
Ta_2Be_{17}	[*12010-36-5*]	2265	7.5	22	15.7	–	206	386	241	31.0	13.8	0.1	2.6	~5	–	–
$ZrBe_{13}$	[*12010-33-2*]	2200	12.5	32.5	17.7	119	172	255	172	32.4	13.8	0.05	0.25	0.6	1310	482
Zr_2Be_{17}	[*12010-37-6*]	2260	12.5	30	15.1	–	172	276	241	31.0	13.8	0.1	0.3	0.1	–	–

RT = room temperature.

Beryllium fluoride [*7787-49-7*], BeF_2, M_r 47.01, is a white, freely soluble, low-melting salt. Molten beryllium fluoride is a poor electrical conductor due to the its covalent character. (See Section 6.3 for production of BeF_2.) In addition to its use as an intermediate in the preparation of beryllium, it is also sometimes used as an additive to welding and soldering fluxes because it dissolves metal oxides readily.

Beryllium chloride [*7787-47-5*], $BeCl_2$, M_r 79.92, crystallizes in asbestos-like matted needles. It is extremely hygroscopic, forming a tetrahydrate. Its reaction with water is intensely exothermic and is accompanied by the release of HCl vapor. Beryllium chloride melts at 416 °C, boils at 520 °C, and, like the fluoride, is a poor electrical conductor in the molten state. (See Section 6.3 for production of $BeCl_2$.) Its principal uses are as the raw material for the electrolytic production of beryllium and as the starting material for syntheses of organoberyllium compounds.

Beryllium bromide, [*7787-46-4*], $BeBr_2$, M_r 168.85, melts at 490 °C. It is obtained as white needles by the reaction of bromine with a mixture of beryllium oxide and carbon. It is used in organic syntheses.

Beryllium orthophosphate, $Be_3(PO_4)_2 \cdot 6\ H_2O$, M_r 325.08, is obtained by reaction of beryllium salts with disodium phosphate or diammonium phosphate in acetic acid solution. Its solubility is very low at room temperature but increases sharply with temperature.

Beryllium ammonium phosphate, NH_4BePO_4, whose composition varies greatly due to partial separation of beryllium hydrogen phosphate, is sparingly soluble. Freshly precipitated NH_4BePO_4 is amorphous.

Aqueous *beryllium perchlorate* [*13597-95-0*], $Be(ClO_4)_2$, is a solvent for cellulose.

Basic beryllium acetate [*19049-40-2*], a hexaacetato complex, $Be_4O(CH_3COO)_6$, like the acetylacetonate and a few other organic beryllium compounds can be distilled without decomposition. Insoluble in water, it melts at 280 °C and boils at 331 °C. The organic beryllium compounds are being used to an increasing extent as intermediates in the purification of beryllium.

Beryllides of some transition metals are promising structural materials or protective coating materials for high-temperature applica-

tions because of their low density, high melting point, good mechanical properties, and resistance to oxidation, even up to high temperatures (Table 10) [192 – 197].

11. Chemical Analysis

The quantitative determination of beryllium, even in very small amounts, is especially important because of the toxicity of beryllium [155, 198]. The best method is atomic absorption spectrometry. When an acetylene/nitrous oxide flame is used, even beryllium oxide is excited, and it is possible to analyze any solution or dispersion without preliminary treatment [155, 199, 200]. Other methods [201 – 205] have also been described.

12. Storage, Transportation, Legal Aspects

Solid parts and structures made from beryllium metal, alloys, and ceramic beryllium can be stored and transported safely without special precautions as long as neither vapor nor dust is generated. Beryllium and beryllium oxide powder have to be handled with care, as is described in Chapter 7. For storage and transportation of the powders, sealed, unbreakable containers and proper instructions are a necessity.

Because beryllium is used mainly in military, nuclear, aeronautical, and rocket technologies, its trade is restricted. A special license is necessary for export from the United States.

13. Uses of the Metal

The properties of beryllium, especially the mechanical properties, are strongly affected by even small amounts of additives, e.g., 1000 ppm or less, and in this sense even commercial grades of beryllium (Table 5) represent alloys. The crucial advantage of beryllium over other materials is its combinations of favorable properties. Detailed discussions of the relationship between properties and uses have been published [19, 41].

The low density of beryllium combined with its high strength, high melting point, and resistance to oxidation is the basis for its use in structural parts that must be light and are exposed to inertial or centrifugal forces. In addition, it has a high modulus of elasticity, which gives light-weight components a high degree of rigidity [19]. Another area of use is in so-called dimensionally stable parts, in which high microyield strength and good thermal conductivity and reflexivity are exploited. It is an excellent heat sink, and therefore used especially for brake systems of airplanes. Applications in civilian nuclear reactors as reflectors and moderators have not met expectations because of radiation damage. However, its favorable nuclear properties have been exploited in weapons systems and in various physical instruments. Beryllium single crystals are the most efficient material for neutron monochromators. Beryllium foil is widely used as windows in energy-dispersive X-ray analyzers.

14. Economic Aspects

Data on production and consumption of beryllium is scanty and incomplete, in most cases because beryllium is used mainly for military and similar purposes. The most complete summary is published annually by the U.S. Bureau of Mines (Table 11), but the data mainly concerns mined beryl.

15. Toxicology and Occupational Health

Occupational Exposure. Exposure to beryllium and inorganic beryllium compounds (e.g., beryllium oxide, beryllium sulfate, beryllium fluoride) can lead to impairment of human health. The highest levels of human exposure occur through occupational inhalation of beryllium dusts or fumes. But also dermal contact with beryllium-containing products is a possible exposure pathway. The general population can be exposed to trace amounts of beryllium through inhalation of air, consumption of food and water, and skin contact with air, water, or soil that contains beryllium. The inhalation route is of greatest concern for systemic effects. Beryllium and its compounds are poorly absorbed after oral (<

Table 11. World production of beryl, as compiled by the U.S. Bureau of Mines, in short tons (2000 pounds)

	1975	1976	1977	1978	1979 r	1980 r	1981 p	1982 p
Argentina	303	123	182	24	13	34	33	29
Brazil	770	406	547	815	498	606	894	882
Madagascar	0	0	17 e	12	11	11	10	11
Mozambique	9	0	NA	NA	31	22	20	17
Portugal	–	–	–	–	6	21	20	21
Rwanda	20	51	60 e	64	51	119	100	110
South Africa	3	3	3	4	1	–	134	66
Uganda e	0	60	50	–	–	–	–	–
Former States of USSR e	1760	1820	1870	1930	2000	2000	2000	2000
Zimbabwe e	70	70	114	39	31	10	46	22
Other b	135	20	1	–	6	21	20	NA
Total	3070	2553	2844	2888	2642	2823	3257	3158

b U.S. bertrandite included;
e estimated;
p preliminary;
r revised; NA, not available.

1%) and dermal exposure. The respiratory tract is the primary target of beryllium toxicity [206]. Workers with the highest potential for exposure include beryllium miners, beryllium alloy makers and fabricators, phosphorus manufacturers, ceramics workers, missile technicians, nuclear reactor workers, electric and electronic equipment workers, and jewelers [207]. About 134 000 workers are potentially exposed to beryllium and its compounds in the USA [208].

Toxicokinetics and Metabolism. After inhalative exposure beryllium and its inorganic compounds are absorbed primarily in the respiratory tract. The absorption of inhaled beryllium depends on physical and chemical properties like particle size and solubility of the beryllium compounds. The higher the solubility, the faster relatively uniform distribution in the organism takes place, with moderate retention in the lungs [209]. In the case of poorly soluble beryllium compounds, 70 – 99% remains in the lungs [210]. Clearance from the lung is biphasic; the initial phase is rapid and involves mucociliary clearance, while the second, slower phase involves translocation to tracheobronchial lymph nodes, alveolar macrophage clearance, and solubilization. In humans beryllium often is detectable in the lungs for several years and can also serve as evidence of previous exposure. Beryllium is poorly absorbed from the gastrointestinal tract. Dermal absorption from intact skin is also low. Absorbed beryllium is distributed throughout the body via blood. Beryllium was detected in lungs, brain, kidney, spleen, liver, muscle, heart, and skeleton. There is suggestive evidence that beryllium is transferred across the placenta and excreted via breast milk. The primary routes of elimination are urine and feces. Beryllium and its inorganic compounds are not metabolized in the body. Only soluble beryllium salts are partially converted in the lungs into less soluble forms [206].

Mechanism of Action. Beryllium can elicit a strong cellular immune response in vivo. Chronic beryllium disease is an allergic reaction of the delayed type, in which beryllium appears to act as a hapten that binds to peptides on mucosal surfaces. In susceptible individuals the beryllium-peptide complex initiates an immune response which may progress ultimately to granuloma formation in the pulmonary interstitium. The genotoxic effects of beryllium are based on the induction of DNA-protein complexes [211]. Beryllium may inhibit the repair of DNA damage and act in a cooperative manner to enhance the genotoxicity of other agents. The effects of beryllium on cellular metabolism and DNA replication combined with its ability to promote an immune response with concomitant induction of reactive oxygen species (and oxidative DNA damage to the target tissue) may be sufficient to promote carcinogenesis in vivo [212]. Beryllium is the least studied carcinogenic metal with respect to gene modulation.

The mechanisms by which beryllium induces tumors are still obscure [213].

Effects in Humans. The main adverse health effects of beryllium are respiratory effects and lung cancer following inhalative exposure, gastrointestinal effects following oral exposure, and skin effects following dermal exposure.

Respiratory Effects. The inhalation of beryllium can cause, depending on duration and level of exposure, acute or chronic beryllium disease (berylliosis). Acute beryllium disease is usually observed after inhalative exposure to high beryllium concentrations ($> 100\ \mu g/m^3$). It is the result of the direct toxicity of water-soluble beryllium compounds. Acute beryllium disease has a short induction period and is usually resolved within a couple of months after termination of exposure. It is believed to be an inflammatory response to beryllium, and most regions of the respiratory tract are affected. Individuals who have previously suffered from acute beryllium disease have an elevated risk of lung cancer. Long-term exposure to relatively low beryllium concentrations ($\geq 0.5\ \mu g/m^3$) of soluble or insoluble beryllium compounds can result in chronic beryllium disease. The duration of this disease is more than one year, and the latency period may range from a few weeks to 20 years after exposure. Here immunological mechanisms in the sense of an allergic reaction of delayed type play a role.

Dermal Effects. Edematous papulovesicular dermatitis was observed in workers exposed to airborne beryllium compounds. Beryllium exposure may also cause a delayed hypersensitivity reaction in the skin. Skin granulomas from beryllium workers contained the same mononuclear infiltrates as detected in the lungs [206].

Cardiovascular Effects. Severe cases of chronic beryllium disease can result in cor pulmonale, that is, hypertrophy of the right heart ventricle. It is not likely that the cardiac effects are due to direct toxicity to the heart, but rather are a response to impaired lung function.

Renal Effects. Kidney stones were observed in association with chronic beryllium disease [206].

Immunological and Lymphoreticular Effects. While acute beryllium disease is a chemical pneumonitis, chronic beryllium disease appears to be an immunological disease. Beryllium can induce cell-mediated immune responses. Beryllium-sensitized cells accumulate at sites of chronic beryllium disease, resulting in granulomas in the lungs. Nonspecific immunological findings in chronic beryllium disease include an increase in serum gamma globulin levels. The existence of specific antibodies to beryllium was reported.

Genotoxicity. Beryllium compounds yielded contradictory results in mutation or chromosomal aberration assays. Bacterial tests were largely negative; mammalian test systems showed evidence of beryllium-induced mutations, chromosomal aberrations, and cell transformation. No genotoxic effects (micronuclei and sister chromatid exchange) were observed in a group of beryl cutters. Thus, the results of genotoxicity assays of soluble beryllium compounds are inconsistent. Differences in the results depend on assay conditions, type of assay, and chemical form of beryllium. Soluble beryllium compounds appear to be weakly genotoxic. The carcinogenicity of beryllium is supported by the mutagenic potential in some of these studies [206, 214, 215].

Reproductive and Developmental Effects. The available data on the potential of beryllium to induce reproductive and/or developmental effects are inconclusive [206].

Carcinogenicity. Increased mortality from lung cancer was found in epidemiological studies of workers in beryllium plants [216 – 221] and in an evaluation of the American Beryllium Case Register [222, 223]. A relationship between the duration of employment and lung cancer mortality was not determined. However, there was an association between death from lung cancer and an increasing time span since the beginning of employment or previous acute beryllium disease. Exposure to high levels of beryllium, which is regarded as the trigger for acute beryllium disease, is associated with a significantly increased risk of lung cancer [224, 225]. The data indicate that the carcinogenic effects occur mainly at high beryllium concentrations, which are no longer to be expected in workplaces today. The epidemiological data for the pulmonary carcinogenicity of beryllium and its inorganic compounds is also supported by the results of animal experiments [210].

Occupational Health (Clinical Symptoms, Diagnosis, Treatment and Prevention).

Acute Beryllium Disease. Acute beryllium disease develops after a short and heavy exposure. Because such high beryllium concentrations are no longer to be expected in workplaces today, this form of the disease is rarely encountered today, except for accidental or unexpected exposure. Acute beryllium disease usually lasts for less than one year, has a latency period of usually around three days, and may take several forms. Symptoms are coughing, breathing difficulties, tightness of the chest, appetite and weight loss, and general weakness and tiredness. Pneumonitis is the most serious of the acute effects from beryllium exposure. It varies in severity and can result in death as a result of the formation of severe pulmonary edema. But fatal cases are rare and recovery is usually complete in about six months. In isolated cases transition to the chronic form is possible without further exposure and after several years [211]. Diagnosis and evaluation of acute beryllium disease requires the identification of the exposure conditions and the analysis of the employment history. Treatment for acute beryllium disease includes removal from exposure, bed rest, and administration of oxygen and corticosteroids, which prevent inflammation.

Chronic Beryllium Disease (Berylliosis). Unlike the acute form, chronic beryllium disease is still frequent in exposed and susceptible individuals. This disease is caused by an immune reaction to inhaled beryllium that is deposited in lung airspaces and retained for a prolonged period. This disease primarily affects the lungs, but it may also affect other organs because the blood transports beryllium throughout the body. Observed symptoms were weight loss, dyspnea, cough, fever, chest pain, cor pulmonale, fatigue, and appetite loss.

On physical examination, signs of chronic beryllium disease are rales (changes in lung sounds), cyanosis (lack of oxygen), digital clubbing, and lymphadenopathy (enlarged lymph nodes). An X-ray of the lungs may show many small scars. Patients may also have an abnormal lung function. Histological examination of the lung tissue after biopsy may show granulomas, which are signs of damage due to the body's reaction to beryllium. Chronic beryllium disease may be confused with other lung diseases, especially sarcoidosis. A blood test, the beryllium-induced lymphocyte proliferation test (Be-LPT), is a valid method for detecting beryllium sensitization, but it is not sufficient for diagnosis of chronic beryllium disease. The presence of granulomas in the lung of an individual with a positive lung Be-LPT confirms the presence of chronic beryllium disease. In the absence of granulomas or other clinical evidence, individuals with a positive Be-LPT are classified as sensitized to beryllium. In beryllium workers involved in the production of beryllia ceramics and nuclear weapons, sensitization rates of 1 – 15% were found [206]. In more than half of the beryllium-sensitized workers chronic beryllium disease was diagnosed.

The only available therapy for chronic beryllium disease is the administration of corticosteroids. Individuals with impaired respiratory gas exchange may require continuous oxygen administration. Asymptomatic individuals who are only sensitized do not need medical treatment. But once sensitization has occurred, it is medically necessary to prevent additional exposure to beryllium.

Because high beryllium concentrations trigger the occurrence of acute beryllium disease, which is a risk factor for lung cancer, the aim of prevention must be to minimize and control exposure at the workplace. Enclosed processes separate workers from harmful substances and prevent toxic substances from escaping into the workroom. Workers should use wet processes or pellets instead of powders. Workplaces need warning devices in areas with a potential of massive contamination. Air from the local exhaust system should be conducted to the atmosphere through high-efficiency filtration equipment. Good housekeeping includes proper storage of substances, frequent disposal of waste, prompt spill cleanup, periodic maintenance of equipment, and no dry sweeping. Respiratory protective devices and protective clothing can help reduce exposure. The workplace should also provide training and educational programs to inform workers about the nature of substances they are exposed to and how to work safely.

Medical surveillance should include chest X-rays and spirometry prior to employment, annually, and on termination. For minimum exposure risks, less frequent intervals may be sat-

isfactory. A possible sensitization to beryllium may be tested by Be-LPT.

Regulations. The OSHA PEL and the ACGIH TLV-TWA for beryllium are 2 μg/m^3. The OSHA ceiling concentration is 5 μg/m^3, and the ACGIH TLV-STEL is 10 μg/m^3. The National Toxicology Program (NTP) lists beryllium and certain beryllium compounds as human carcinogens. The IARC has classified beryllium and beryllium compounds in Group 1, carcinogenic to humans [226]. In Germany, beryllium and its inorganic compounds are considered as human carcinogens and classified in carcinogenicity category 1 by DFG Commission for the Investigation of the Health Hazards of Chemical Compounds in the Work Area [210]. In contrast, the EPA concluded that the human data only provide limited evidence and classified inhaled beryllium in Group B1, a probable human carcinogen [206].

16. Acknowledgement

The authors would like to thank James Marder, James Stonehouse, and W. Raymond Cribb of Brush Wellman Inc. for arranging for the translation of our article and after translation for their careful review and helpful suggestions.

17. References

1. L. N. Vauquelin, *Ann. Chim. (Paris)* **26** (1798) 155, 170, 259; **30** (1799) 82.
2. F. Wöhler, *Pogg. Ann.* **13** (1828) 577.
3. A. A. B. Bussy, *J. Chim. Med. Pharm. Toxicol.* **4** (1828) 453.
4. P. Lebeau, *Ann. Chim. Phys.* s7 **16** (1899) 457.
5. G. Oesterheld, *Z. Anorg. Allg. Chem.* **97** (1916) 14.
6. G. Masing, O. Dahl, *Wiss. Veröff. Siemens Werke* **8** (1929) 126.
7. W. W. Beaver, D. W. Lillie in C. R. Tipton ,Jr. (ed.): *Reactor Handbook*, Interscience, New York 1960, p. 897.
8. H. H. Hausner: "Beryllium as a Moderator and a Reflector for Nuclear Reactor," *At. Energy Rev.* **1** (1963) no. 2, 99.
9. J. B. Rich, G. B. Redding, R. S. Barnes, *J. Nucl. Mater.* **1** (1959) 96.
10. D. W. White, J. E. Burke: *The Metal Beryllium*, The American Society for Metals, Cleveland 1955.
11. G. E. Darwin, J. H. Buddery: *Beryllium*, Butterworths, London 1960.
12. H. H. Hausner: *Beryllium, Its Metallurgy and Properties*, University of California Press, Berkeley – Los Angeles 1965.
13. R. F. Williams, S. E. Ingels, *NASA Tech. Rep.* TM-X-53453, vol. **1 – 6** (1966).
14. G. E. Meyer, J. H. Henning, *DMIC Rep.* **S-29** (1970).
15. *The Metallurgy of Beryllium*, Proc. Int. Conf., London 1961, Institute of Metals, Monograph and Report Series no. 28, Chapman & Hall, London 1963.
16. *Proc. Int. Conf. Beryllium 1st*, AIME, Gatlinburg, Pa., 1963.
17. *Proc. Int. Conf. Beryllium 2nd*, AIME, Philadelphia 1964.
18. *Conf.* Int. Metall. Beryllium Commun. 3rd 1965.
19. *Proc. Beryllium Conf.,* Arlington, Va., 1970, NMAB Rep. **272** (1970).
20. F. E. Buresch: *Berylliumoxidkeramik unter besonderer Berücksichtigung moderner Anwendungen; Appl. Mineralogy – Technische Mineralogie*, Springer Verlag, Berlin 1974 /75.
21. D. Webster, G. J. London (eds.): *Beryllium Science and Technology,* vol. I D. R. Floyd, J. N. Lowe (eds.): *Beryllium Science and Technology,* vol. II, Plenum Press, New York 1979.
22. M. C. Udy, H. L. Shaw, F. W. Bougler, *Nucleonics* **11** (1953) 52.
23. C. J. Smithells: *Metals Reference Book*, Interscience, New York 1955.
24. A. J. Martin, A. Moore in: *Symposium on the Study of Metals and Alloys above 1200 °C*, Oxford Univ. Press, England 1958.
25. W. B. Pearson: *Lattice Spacings and Structure of Metals and Alloys,* vol. 2, Pergamon Press, London 1967.
26. P. Gordon, *J. Appl. Phys.* **31** (1960) 1221.
27. A. J. Martin, A. Moore, *J. Less-Common. Metals* **1** (1959) 85.
28. R. Hultgren et al.: *Selected Values of Thermodynamic Properties of Metals and Alloys*, J. Wiley & Sons, New York 1963.
29. R. W. Hill, P. L. Smith, *Philos. Mag.* **44** (1953) 636.
30. V. N. Eremenka, V. I. Nizhenko, Shou Wei Tai, *Izv. Akad. Nauk SSSR Otd. Tekh. Nauk Metall. Topl.* **3** (1960) 116.

31. R. L. Powell, *Philos. Mag.* **44** (1953) 645.
32. R. M. Treco, *Trans.* **188** (1950) 1274.
33. M. Owens, *Ann. Phys.* **37** (1912) 657.
34. A. Ciccione, *Nature (London)* **130** (1932) 315.
35. M. M. Mann, Jr., L. A. du Bridge, *Phys. Rev.* **51** (1937) 120.
36. C. B. Sawyer, B. R. F. Kjellgren, *Met. Alloys* **11** (1940) 163.
37. F. Aldinger, *Metall (Berlin)* **26** (1972) 711.
38. R. L. Powell, *Philos. Mag.* **31** (1960) 1221.
39. R. W. Meyerhoff, J. F. Smith, *J. Appl. Phys.* **33** (1962) 223.
40. A. R. Kaufmann, *USAEC* **MIT-1113** (1953).
41. F. Aldinger, G. Petzow, *Radex Rundsch.* **3/4** (1972) 275.
42. E. J. Lewis, *Phys. Rev.* **34** (1929) 1575.
43. L. Losana, *Aluminio* **8** (1939) 67.
44. P. Gordon, *USAEC* **MDDC-1370, CT-3315** (1945).
45. F. K. Lampson, *USAEC* **NEPA-186** (1951).
46. S. H. Gelles, *NMI Q*. Prog. Rep. (Jan. – March 1962).
47. M. Wilhelm, *Dissertation*, University of Stuttgart, 1974.
48. *Extracts on Soviet Beryllium and Beryllium Alloys*, DMIC Tech. Note (June 1969).
49. B. G. Lazarev, A. I. Sudovtsev, E. E. Semenenko, *Zh. Eksp. Teor. Fiz.* **37** (1959) 1461.
50. A. E. Cruzon, A. J. Mascall, *J. Phys. C* **2** (1969) 383.
51. G. K. White, S. B. Woods, *Can. J. Phys.* **33** (1955) 58.
52. M. C. Udy, H. L. Shaw, F. W. Boulger, *USAEC* **AECD-3382** (1949).
53. J. E. Janssen et al., *Honeywell Center Rep.* **ASD-TR-61-147** (1961).
54. D. E. Gray: *Am. Inst. of Phys. Handbook*, McGraw-Hill, New York 1957.
55. D. Strominger, J. M. Hollander, G. T. Seaborg, *Rev. Mod. Phys.* **30** (1958) 585.
56. C. E. Ells, E. C. Perryman, *J. Nucl. Mater.* **1** (1959) 96.
57. A. R. Kaufmann, P. Gordon, D. W. Lillie, *Trans. Am. Soc. Met.* **42** (1950) 785.
58. W. W. Beaver, K. G. Wikle, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **200** (1954) 559.
59. H. D. Hanes, S. W. Porembka, J. B. Melehan, P. J. Gripshover, *DMIC Rep*. (1965).
60. H. Conrad, G. London, V. V. Damiano in F. W. Vahldiek, S. A. Mersol (eds.): *Anisotropy in Single Crystal Refractory Compounds,* vol. 2, Plenum Press, New York 1968.
61. A. N. Stroh, *Philos. Mag.* **3** (1958) 397.
62. J. J. Gilman, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **200** (1954) 621.
63. D. Webster, *Conf. Strength Met. Alloys Conf. Proc. 4th* **2** (1976) 669.
64. J. F. Smith, C. L. Arbogast, *J. Appl. Phys.* **31** (1960) 99.
65. W. D. Rowland, J. S. White, *J. Phys. F* **2** (1972) 231.
66. M. P. Baldwin, *AWRE Rep.* **O-49/70** (1970).
67. D. R. G. O'Rourke et al., *USAEC* **COO-312** (1956).
68. R. A. Foos, A. J. Stonehouse, K. A. Walsh, *BBC Rep.* **TR-456** (1970).
69. K. W. Walsh, A. J. Stonehouse, A. J. Sandor, *BBC Rep.* **TR-465** (1971).
70. J. Greenspan, *USAEC* **NMI-1174** (1957).
71. N. Maropis, J. B. Jones, *USAEC* **NYO-7788** (1957).
72. V. V. Damiano et al., *FIRL Rep.* **F-B2373** (1966), **F-C1820** (1967), **F-C2031** (1968), **F-C2233** (1969), **F-C2521** (1970).
73. R. Syre, *Met. Corros. Ind.* **33** (1958) 406.
74. A. V. Novoselova, L. R. Batsanowa: *Analytical Chemistry of Beryllium*, Ann Arbor-Humphrey Science Publ., Ann Arbor – London 1969.
75. *Gmelin*, system no. 26, Beryllium (1930).
76. P. D. Miller, W. K. Boyd, *DMIC Rep.* **242** (1967) .
77. M. Hansen, K. Anderko: *Constitution of Binary Alloys*, McGraw-Hill, New York 1958.
78. R. P. Elliott: *Constitution of Binary Alloys*, **1st Suppl**., McGraw-Hill, New York 1965.
79. F. A. Shunk: *Constitution of Binary Alloys*, **2nd Suppl**., McGraw-Hill, New York 1969.
80. B. H. Clemmons, J. S. Browning, *Min. Eng. (Littleton, Colo.)* **5** (1953) 786.
81. A. M. Gaudin et al., *Min. Eng. (Littleton, Colo.)* **2** (1950) 495.
82. R. G. Bellamy, N. A. Hill: *Extraction and Metallurgy of Uranium*, Thorium and Beryllium, Pergamon Press, Oxford 1963.
83. G. Jäger, *Metall (Berlin)* **4** (1950) 183.
84. K. C. Chen, *Dissertation*, London 1950.
85. W. L. W. Ludekens, *Dissertation*, London 1950.
86. *Ullmann*, 3rd ed., vol. **4**, p. 322.
87. H. Claflin, US 1929014, 1933.
88. R. A. Opatowski, US 2209131, 1941.
89. H. C. Kawecki, US 2312297, 1943; *Trans. Electrochem. Soc.* **89** (1946) 229.
90. C. V. Sundaram, C. M. Paul, B. P. Sharma, *Trans. Indian Inst. Met.* **35** (1982) no. 2, 171.
91. R. W. Winters, L. F. Yntema, *Trans. Electrochem. Soc.* **55** (1929) 205.

92. F. K. McTaggart, *J. Counc. Sci. Ind. Res. (Aust.)* **20** (1947) 564.
93. L. Burgess, *Am. Electrochem. Soc.* **47** (1925) 317.
94. J. Kielland, L. Tronstad, *K. Nor. Vidensk. Selsk. Arsberetn.* **8** (1935) 147.
95. J. Besson, *C. R. Hebd. Séances Acad. Sci.* **214** (1942) 861.
96. C. Matignon, J. Cathala, *C. R. Hebd. Séances Acad. Sci.* **181** (1925) 1066.
97. J. M. Schmidt, *Sci. Ind. (Paris)* **13** (1929) 110.
98. R. A. Foos: *Applications of Solvent Extraction to Beryllium Oxide Manufacturing*, AIME Conference, New York 1968.
99. Brush Beryllium Co., US 3259456, 1964.
100. C. W. Schwenzfeier et al., US 2660515, 1953.
101. G. Jeager, DE 752729, 1938.
102. *Industrial Challenge of Nuclear Energy III*, part 3, Stresa Conference, 1959, OEEC 1960.
103. R. W. Buddery, R. W. Thackray, *J. Inorg. Nucl. Chem.* **3** (1956) 1901.
104. W. Kroll, *Z. Anorg. Allg. Chem.* **240** (1939) 332.
105. O. Kruh, AT 147634, 1934.
106. H. v. Zeppelin, DE 705645, 1937.
107. B. Wempe, US 2091087, 1933.
108. B. R. F. Kjellgren, *Trans. Electrochem. Soc.* **93** (1948) 122.
109. L. J. Derham, D. A. Temple: *Extraction and Refining of the Rarer Metals*, Inst. Min. Metall, London 1957.
110. J. M. Tien, *Trans. Electrochem. Soc.* **89** (1946) 237.
111. Brush Beryllium Co., *Progr. Rep.* **12 – 17** (1947).
112. T. T. Campbell, R. E. Mussler, F. E. Block, *Metall. Trans.* **1** (1970) 2881.
113. A. Stock, H. Goldschmidt, DE 375824, 1924.
114. K. Illig et al., *Wiss. Veröff. Siemens Werke* **8** (1929) no. 1, 47.
115. J. Dickinson, US 1511829, 1921.
116. A. C. Vivian, *Trans. Faraday Soc.* **22** (1926) 211.
117. Degussa, DE 646088, 1937.
118. G. A. Meyerson, *Proc. U.N. Int. Conf. Peaceful Uses At. Energy, 1st* 1955, 633.
119. R. B. Holden et al., *J. Am. Chem. Soc.* **70** (1948) 3897.
120. E. A. Gulbranssen, K. F. Andrew, *J. Electrochem. Soc.* **97** (1950) 383.
121. N. D. Erway, R. I. Seifert, *J. Electrochem. Soc.* **93** (1951) 83.
122. J. P. Pemsler et al., *NMI Rep.* **TJ-38** (1961).
123. S. H. Gelles et al., *ASD Rep.* **TDR-62-509** (1962).
124. G. W. Pfann, *J. Met.* **4** (1952) 747.
125. C. S. Pearshall, *USAEC* **MIT-1103** (1952).
126. J. P. Denny, B. H. Hessler, *ASD Rep.* **TDR-62-390** (1962).
127. S. W. Porembka, H. D. Hanes, P. J. Gripshover, *DMIC Rep.* **239** (1967) .
128. F. T. Zurey, H. D. Hanes, P. J. Gripshover: *Hot Isostatic Pressing of Beryllium,* vol. I, Batelle Mem. Inst., Columbus, Ohio, 1968.
129. G. W. Fischer et al., *GEC Rep.* **IR-8-12** (1965).
130. C. A. Bielowski, J. G. Theodore, W. W. Beaver, *NASA Tech. Rep.* **334-241** (1963).
131. *Spark Sintering of Parts and Preforms*, Lockheed Missiles & Space Comp., Sunnyvale, Calif., 1969.
132. M. Hirschvogel, F. Aldinger, *Proc. Int. Powder Met. Conf. 4th.*, Toronto 1973.
133. M. L. Headmann, T. J. Roseberry, F. L. Parkinson, *West. Gear Corp. Rep.* **649-229** (1964).
134. H. T. Lee, R. M. Brick: "Slip and Twinning in Single Crystals of Beryllium," *J. Met.* **4** (1952) 147 – 148.
135. H. T. Lee, R. M. Brick: "Deformation of Beryllium Single Crystals at 25 to 500 °C," *Trans. Am. Soc. Met.* **48** (1956) 103 – 107.
136. C. S. Pearsall, *Nucl. Sci. Abstr.* **7** (1953) 1434.
137. L. Gold, *USAEC* **AECD-2643** (1949).
138. I. I. Papirov, G. F. Tikhinskij, *Kristallografiya* **9** (1964) 310.
139. M. Wilhelm, F. Aldinger, *Z. Metallkd.* **66** (1975) 323.
140. G. J. London, J. D. Meakins, *FIRL Rep*. **P-C1870-3-1**, Philadelphia 1970.
141. G. E. Spangler, M. W. Herman, E. J. Arndt, *FIRL Rep*. **F-A2476**, Philadelphia 1961.
142. M. Herman, G. E. Spangler, in: *The Metallurgy of Beryllium*, Institute of Metals, Monograph and Report Series no. 28, Chapman & Hall, London 1963, p. 75.
143. L. M. McDonald-Schetky, H. A. Johnson (eds.): *Beryllium Technology,* vol. 1, Gordon & Breach, New York 1964.
144. D. F. Kaufman, E. D. Levine, J. J. Pickett, L. R. Aronin in: *Proc. Conf. Phys. Metall. Beryllium*, **AEC-CONF-170** (1963) 69.
145. D. Beasley, *Conf. Int. Metall. Beryllium Commun. 3rd 1965*, 129.
146. J. D. Meakin, *FIRL Rep*. **F-C1870**, Philadelphia 1971.

147. A. Freund, J. O. Forsyth in: *Treatise on Materials Science and Technology*, vol. **15**, Neutron Scattering, Academic Press, New York 1979, chap. 10.
148. S. Jönsson, *Naturwissenschaften* **69** (1982) 483.
149. F. Aldinger, A. Freund in: "Beryllium 1977," *Proc. 4th Int. Conf. Beryllium*, Royal Society, London 1977, p. 55/1.
150. J. Faure-Deloche, *Thesis*, University of Grenoble 1973.
151. A. S. Kapcherin, I. I. Papirov, G. F. Tikhinskij, A. S. Avotin in: "Beryllium 1977," *Proc. 4th Int. Conf. Beryllium*, Royal Society, London 1977, p. 5/1.
152. S. Jönsson, A. Freund, F. Aldinger, *Metall (Berlin)* **33** (1979) 1257.
153. S. Stilz, S. Jönsson, *Metall (Berlin)* **38** (1984) 748.
154. A. K. Freund, S. Jönsson, S. Stilz, G. Petzow, *J. Nucl. Mat.* **124** (1984) 215.
155. S. Mönch, *Metall (Berlin)* **23** (1969) 238.
156. H. Zorn, H. Diem, *Zentralbl. Arbeitsmed. Arbeitsschutz* **24** (1974) 3.
157. G. Petzow, H. Zorn, *Chem. Ztg.* **98** (1974) no. 5, 236.
158. Brush Wellman Inc., Cleveland, Ohio, *Specification Sheets* (1972).
159. Kawecki Berylco Ind., Hazleton, Pa., *Product Specifications* (1971).
160. Kawecki Berylco Ind., Hazleton, Pa., *Product Specification*, File no. 302-PD10 (1972).
161. G. G. Winkle et al., *AMC Rep.* **60-7-631** (1960).
162. E. Crawford et al., *ASD Rep.* **TR 61-692** (1962).
163. P. C. Paris, D. O. Harris, *UCRL Rep.* **72442** (1970) .
164. R. E. Cooper, *AWRE Rep.* **O-17/72** (1972).
165. J. G. Klein, L. M. Perelman, W. W. Beaver, *WADC Rep.* **TR-58-478** (1958).
166. R. E. Evans, D. Beasley, J. N. Lowe, "Alloy Strengthened Beryllium Sheet," *AWRE Rep.* **O-12/67** (1967).
167. J. G. Klein, L. M. Perelman, W. W. Beaver, *WADC Rep.* **58-478**, I (1959), II (1960).
168. W. W. Beaver, *BBC Rep.* **TR-242** (1968).
169. Lockheed Missiles & Space Comp., US 3337334, 1963.
170. C. Carson, J. P. Denny: *Solid Solution Strengthened Beryllium Alloy Ingot Sheet*, Beryllium Corp., Reading, Pa. 1966 /67.
171. W. Taylor, C. Carson: *Solid Solution Strengthened Beryllium Alloy Ingot Sheet*, Beryllium Corp., Reading, Pa. 1967 /68.
172. W. Fischer, F. Aldinger, *Tagung Verbundwerkstoffe*, Konstanz 1974.
173. R. W. Fenn et al., *SAE Rep.* **660652** (1966).
174. D. B. King, A. J. Stonehouse, *Ordnance*, in press, 1985.
175. D. Böhme, *Z. Metallkd.* **56** (1965) 487.
176. Vakuumschmelze Hanau, *Tech. Informationsblätter F 3* (1967).
177. Vakuumschmelze Hanau, *F 004* (1967).
178. Deutsche Beryllium GmbH, Tech. Inf. no. 1006 G (1982).
179. I. Pfeiffer, *Z. Metallkd.* **57** (1966) 635.
180. J. R. Burns, *Trans. Am. Soc. Met.* **40** (1948) 143.
181. S. Balicki, *Pr. Inst. Hutn.* **10** (1958) 208.
182. *Tech. Rundsch.* **51** (1959) no. 35, 5.
183. K. E. Mann, *Z. Metallkd.* **46** (1955) 17.
184. E. Nachtigall, H. Landerl, *Alum. Ranshofen Mitt.* **4** (1956) 15.
185. L. David, *Light Met. (London)* **18** (1955) 15.
186. V. de Pierre, H. Bernstein, *Trans. Am. Soc. Met.* **43** (1951) 635.
187. A. L. Mincher, *Met. Ind. (London)* **76** (1950) 435.
188. F. E. Buresch, *Jahresber. Kernforschungsanlage Jülich* **JÜL-552-RW** (1968).
189. E. Ryshkewitch: *Oxide Ceramics*, Academic Press, New York – London 1960, p. 318.
190. A. Boltax, J. H. Handwerk, *Proc. Conf. Nucl. Appl. Nonfissionable Ceram.* 1966, 169.
191. A. Rabenau, P. Eckerlin in P. Popper (ed.): *Special Ceramics*, Heywood & Co., London 1960, p. 136.
192. J. R. Lewis, *J. Met.* **13** (1961) 357.
193. A. J. Stonehouse, *Mater. Des. Eng.* **55** (1962) no. 2, 84.
194. G. Ervin Jr., M. M. Nakata, *J. Electrochem. Soc.* **110** (1963) 1103.
195. R. M. Paine, A. J. Stonehouse, W. W. Beaver, *Corrosion (Houston)* **20** (1964) 307.
196. W. W. Beaver, A. J. Stonehouse, R. M. Paine, *Met. Space Age: Plansee Proc. Pap. Plansee Semin. "De Re Met." 5th* 1964, 682.
197. G. Petzow, M. Stümke, *Keram. Z.* **20** (1968) no. 12, 795.
198. H. E. Stockinger: *Beryllium, Its Industrial Hygiene Aspects*, Academic Press, New York 1964.
199. D. L. Bokowski, *RFP Rep.* **787** (1966) .
200. J. D. Taylor, T. W. Steele, *Rep. Nat. Inst. Metall. (S. Afr.)* **173** (1967) .
201. G. Kimmerle: *Handbuch der experimentellen Pharmakologie XXI: Beryllium*, Springer Verlag, Berlin 1966.

202. D. A. Everest: *The Chemistry of Beryllium*, Elsevier, Amsterdam 1964.
203. J. S. Pollok, I. S. Jones, *AERE Rep.* **R-5106** (1966).
204. P. Iredale, *AERE Rep.* **EL/M 108** (1960).
205. W. D. Ross, R. E. Sievers, *Talanta* **15** (1968) 87.
206. ATSDR: Toxicological Profile for Beryllium, U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta, GA, 2002.
207. *Report on Carcinogens*, 11th ed., U.S. Department of Health and Human Services, Public Health Services, National Toxicology Program, 2005.
208. P. K. Henneberger, S. K. Goe, W. E. Miller, B. Doney, D. W. Groc: "Industries in the United States with Airborne Beryllium Exposure and Estimates of the Number of Current Workers Potentially Exposed", *J. Occup. Environ. Hyg.* **1** (2004) 648 – 659.
209. U.S. EPA: *Toxicological Review of Beryllium and Compounds*, Washington, D.C., 1998.
210. DFG: "Beryllium und seine anorganischen Verbindungen" in H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten (Maximale Arbeitsplatzkonzentrationen)*, Lfg. 37, Wiley-VCH, Weinheim 2003.
211. WHO: *EHC (Environmental Health Criteria) 106, Beryllium*, International Programme on Chemical Safety, Geneva 1990.
212. E. T. Snow: "Metal Carcinogenesis: Mechanistic Implications", *Pharmacol Ther.* **53** (1992) 31 – 65.
213. D. Beyersmann: "Effects of Carcinogenic Metals on Gene Expression", *Toxicol. Lett.* **127** (2002) 63 – 68.
214. T. Gordon, D. Bowser: "Beryllium: Genotoxicity and Carcinogenicity", *Mutat. Res.* **533** (2003) 99 – 105.
215. R. Wegner et al.: "Lung Function, Biological Monitoring, and Biological Effect Monitoring of Gemstone Cutters Exposed to Beryls", *Occup. Environ. Med.* **57** (2000) 133 – 139.
216. T. F. Mancuso: "Relation of Duration of Employment and Prior Respiratory Illness to Respiratory Cancer Among Beryllium Workers", *Environ. Res.* **3** (1970) 251 – 275.
217. T. F. Mancuso: "Occupational Lung Cancer among Beryllium Workers" in Proceedings of the Conference on Occupational Exposure to Fibrous and Particulate Dust and their Extensions into the Environment, R. Lemen, J. Dement (eds.): *Dust and Disease*, Pathotox Publishers, Park Forest South 1979, pp. 463 – 482.
218. T. F. Mancuso: "Mortality Study of Beryllium Industry Workers' Occupational Lung Cancer", *Environ. Res.* **21** (1980) 48 – 55.
219. T. F. Mancuso, A. A. El-Attar: "Epidemiological Study of the Beryllium Industry. Cohort Methodology and Mortality Studies", *J. Occup. Med.* **11** (1969) 422 – 434.
220. J. K. Wagoner, P. F. Infante, D. L. Bayliss: "Beryllium: an Etiologic Agent in the Induction of Lung Cancer, Nonneoplastic Respiratory Disease, and Heart Disease among Industrially Exposed Workers", *Environ. Res.* **21** (1980) 15 – 34.
221. E. Ward, A. Okun, A. Ruder, M. Fingerhut, K. Steenland: "A Mortality Study of Workers at Seven Beryllium Processing Plants", *Am. J. Ind. Med.* **22** (1992) 885 – 904.
222. P. F. Infante, J. K. Wagoner, N. L. Sprince: "Mortality Patterns from Lung Cancer and Nonneoplastic Respiratory Disease among White Males in the Beryllium Case Registry", *Environ. Res.* **21** (1980) 35 – 43.
223. K. Steenland, E. Ward: "Cancer Incidence among Patients with Beryllium Disease: A Cohort Mortality Study", *J. Natl. Cancer Inst.* **83** (1991) 1380 – 1385.
224. W. T. Sanderson, M. R. Petersen, E. M. Ward: "Estimating Historical Exposures of Workers in a Beryllium Manufacturing Plant", *Am. J. Ind. Med.* **39** (2001) 145 – 157.
225. W. T. Sanderson, E. M. Ward, K. Steenland, M. R. Petersen: "Lung Cancer Case-Control Study of Beryllium Workers", *Am. J. Ind. Med.* **39** (2001) 133 – 144.
226. IARC: IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 58, Beryllium, Cadmium, Mercury and Exposures in the Glass Manufacturing Industry, Lyon 1994, 41 – 117.

Beverages, Nonalcoholic

→ Beer, → Coffee, → Milk and Dairy Products, → Spirits, → Tea, → Water, → Wine are separate keywords

Philip Crandall, University of Florida

Chin Shu Chen, University of Florida

Steven Nagy, Florida Dept. of Citrus, Lake Alfred, Florida 33850, United States (Chaps. 1, 3, 5)

Georges Perras, Perrier Group – 75382 Paris – 30310 Vergeze, France (Chap. 2)

Johannes A. Buchel, Pepsi Co, Inc., Valhalla, New York 10595, United States (Chaps. 4, 5)

William Riha, Pepsi Co, Inc., Valhalla, New York 10595, United States (Chaps. 4, 5)

1. Introduction . . . 1
1.1. Drinking Water . . . 2
1.2. Fruit and Vegetable Juices and Derived Juice Products . . . 2
1.3. Carbonated and Non-carbonated Beverages . . . 3
2. Mineral Water . . . 4
2.1. Definitions of Mineral Water . . . 4
2.1.1. European . . . 4
2.1.2. French . . . 5
2.1.3. Codex Alimentarius . . . 6
2.2. Development of Mineral Water . . . 6
2.3. Legislation . . . 7
2.4. Packaging of Mineral Water . . . 7
2.5. Regulations for Labeling . . . 8
2.6. Bottling . . . 8
2.7. Conclusions . . . 8
3. Fruit and Vegetable Juice Products 8
3.1. Temperate Fruits . . . 8
3.1.1. Production . . . 9
3.1.2. Harvesting and Transportation . . . 9
3.1.3. General Fruit Quality . . . 9
3.1.4. Apple Processing . . . 12
3.1.5. Grape Processing . . . 14
3.1.6. Nutritive and Chemical Composition 15
3.1.7. Regulations . . . 15
3.2. Subtropical Fruits . . . 17
3.2.1. Quality of Fruit for Processing . . . 17
3.2.2. Processing . . . 18
3.2.3. Processed Citrus Products . . . 23
3.2.4. Chemical and Nutritive Data . . . 24
3.3. Tropical Fruits . . . 25
3.3.1. Introduction . . . 25
3.3.2. Quality of Fruit for Processing . . . 25
3.3.3. Processing . . . 25
3.4. Vegetable Juices . . . 28
4. Sweet Carbonated and Non-carbonated Beverages . . . 29
4.1. Types of Beverages . . . 30
4.1.1. Still Drinks . . . 30
4.1.2. Carbonated Beverages . . . 31
4.2. Flavoring Materials and Concentrates . . . 31
4.2.1. Natural Flavoring Materials . . . 31
4.2.2. Synthetic Flavoring Materials . . . 32
4.2.3. Flavoring Concentrates . . . 32
4.3. Beverage Production . . . 33
4.3.1. Water Treatment . . . 33
4.3.2. Syrup Preparation . . . 34
4.3.3. Carbonation and Filling . . . 34
4.3.4. Wastewater Treatment . . . 35
4.3.5. Quality Control . . . 35
5. Economic Aspects . . . 35
5.1. Fruit Juices . . . 35
5.2. Carbonated Soft Drinks . . . 36
5.2.1. United States Market . . . 37
5.2.2. International Markets . . . 39
6. References . . . 40

1. Introduction

A *beverage* is defined as a liquid for drinking, and includes the following general categories:

1) water (natural, mineral, carbonated, non-carbonated);
2) fruit and vegetable juices (single-strength, concentrated), and their modifications (drinks, comminuted, nectars, fortified, carbonated);
3) soft drinks (carbonated, non-carbonated, packaged, diet);
4) tea;

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_035

5) coffee;
6) milk;
7) beer;
8) wine (still, carbonated, dry, sweet);
9) distilled spirits.

Beverages can be further differentiated into nonalcoholic and alcoholic types; in general, a *nonalcoholic beverage* may be fermented or unfermented but must contain less than 0.5 vol% ethanol. Differentiation is best exemplified by the malt beverages. "Beer" (alcoholic beverage) is defined as beer, ale, porter, stout, or other similar fermented beverages containing 0.5 vol% or more ethanol. The product must be brewed or produced wholly or in part from malt. On the other hand, a "cereal beverage" (nonalcoholic beverage) is defined as a fermented or unfermented malt beverage containing less than 0.5 vol% ethanol [1]. Details on regulatory control of standards and of the various phases of production, labeling, and advertisement of alcoholic beverages are found in Title 27 of the United States Code of Federal Regulations [1]. This article covers only nonalcoholic beverages and, specifically, those beverages confined to categories (1), (2), and (3).

1.1. Drinking Water

Water [*7732-18-5*] for drinking contains no pathogenic organisms, has a pH between 6.5 and 8.0, contains minor amounts of organic matter, and may or may not contain mineral salts. The maximum levels of allowable constituents in drinking water are: 1000 ppm total solids, 250 ppm Cl, 250 ppm SO_4, 100 ppm Mg, 0.5 ppm Zn, 0.3 ppm Fe, 0.2 ppm Cu, and 0.1 ppm Pb [2]. Drinking water containing no mineral salts is classified as:

1) *Demineralized:* manufactured by passage of natural water through an ion-exchange medium.
2) *Distilled:* natural water is vaporized and subsequently condensed leaving behind nonvolatile minerals.

1.2. Fruit and Vegetable Juices and Derived Juice Products

The expression or pressing of a fruit or vegetable yields a product which contains not only a liquid component (water, soluble solids) but, most often, an insoluble component (finely suspended colloidal particles). Generally, insoluble components are removed commercially by a finishing process, but they may be deliberately added to certain juices (apple, citrus) to mimic fresh juice and/or as an economic matter to increase juice yield. United States Standards of Identity explicitly define processing and compositional variable limits for most juice products.

Fruit. A *fruit juice* is the unfermented, undiluted, extractable fluid content resulting from expression of fruit. *Fruit drink* or *diluted fruit beverage* are descriptive phrases used to define a non-carbonated beverage containing > 0 % but < 100 % fruit juice. A diluted fruit juice beverage may contain added water, sugar, other sweeteners, gum arabic, artificial colors, flavors, vitamins, and acids (citric, malic, fumaric). The percent of fruit juice(s) in a diluted juice beverage is calculated on the basis of the soluble-solids content of the undiluted (single-strength) juice and is declared on a vol% basis. In diluted juice beverages containing only one type of fruit juice, the U.S. Food and Drug Administration [3] has proposed a rule that would require a label statement declaring the percent of juice in 5 % increments. Beverages containing > 0 % but < 5 % would be declared as "less than 5 %". Beverages containing multiple juices (e.g., fruit punches) may declare the percentage of the total juice content or a statement of the percentage of each juice represented.

Some fruit juices are too acidic, too strongly flavored, or too bland to yield a natural, pleasant-tasting beverage. However, delicious drinks may be prepared from these fruits by addition of flavor-modifying ingredients and by dilution with a thin syrup. Also, the delicate flavor of many fruit beverages is enhanced if fruit solids are added. The solids are converted to a soft, smooth pulp which enhances flavor and imparts a desirable mouth-feel. Therefore, a *fruit nectar* is a pulpy fruit juice to which may be added sugar syrup, acids, vitamins, pectin, essences, and water. Sweeteners include sugar, invert sugar syrup,

dried corn syrup, glucose syrup, and dried glucose syrup; optional ingredients include lemon juice; citric, malic, and fumaric acid as acidifiers; and ascorbic acid as an antioxidant and as a vitamin. A typical nectar, e.g., pear, contains water, pear pulp and juice, corn sweetener, sugar, citric acid, and ascorbic acid (vitamin C). Most nectars contain $\geqq 50$ % pure juice, but for purposes of identity, they cannot be classified as a fruit juice because of added ingredients.

Vegetable. *Vegetables* may be grouped into 5 general types based on the part which is consumed and, hence, on the part that is most often used to derive a juice product:

1) Leaf or stem vegetables, e.g., celery, cabbage, rhubarb
2) Root vegetables, e.g., carrot, beet, turnip
3) Fruiting vegetables, e.g., tomato, cucumber, eggplant, pepper
4) Immature flower bud vegetables, e.g., cauliflower, broccoli, artichoke
5) Legumes, e.g., pea, bean, soybean

The major constituent of all vegetables is water; the percentage ranges from about 70 % (legumes) to about 95 % (roots, leaves, stems). Vegetable juices cannot satisfy all of the daily nutrient requirements of human diets, but they are an important contributor of vitamins and minerals to that diet.

The *juices* derived from vegetables are generally grouped into 6 classes [4]:

1) Juices prepared from normally acid vegetables, e.g., tomato and rhubarb; may be processed at low temperature.
2) Vegetable juices or their blends acidified with highly acid products, such as tomato, sauerkraut, or rhubarb juice; tomato juice is often a primary ingredient.
3) Vegetable juices acidified with organic or mineral acids can be processed at relatively low temperature.
4) Excess juices derived from fermented vegetables are either used in their natural state or heat processed at low temperatures, e.g., sauerkraut juice.
5) Vegetable juices or their blends which are not acidified must be processed in a retort at high temperature.
6) Juices freshly extracted from nonacid vegetables must be consumed quickly (specialty or health food stores); these juices are neither heat-stabilized nor acidified.

1.3. Carbonated and Non-carbonated Beverages

Carbonated and non-carbonated beverages, more commonly known as soft drinks, represent a very broad category that overlaps, to some extent, the category of fruit and vegetable juices and their modified products. A fine distinction cannot be made because the two beverage types contain natural product ingredients (in whole or part). A *soft drink* is defined as a nonalcoholic beverage of many flavors [2], and includes: (1) any fruit drink (synthetic or natural), squash, crush, or cordial; (2) soda water, Indian or quinine tonic water, and any sweetened, artificially carbonated water; (3) ginger beer and any herbal or botanical beverage. Soft drinks are also called carbonated beverages, soda water, soda, cola, pop, or drinks.

Carbonated. Carbonation is a term used to identify the dissolution of carbon dioxide gas in water utilizing pressure and temperature. The dissolved gas not only imparts a distinctive taste, zest, and sparkle to the beverage, but the dissolved gas also plays a major part in inhibiting or destroying harmful bacteria [5]. Soft drink beverages contain carbonation ranging from 1 to 5 volumes of gas (1.05 to 5.27 kg/cm^2 gas pressure) per volume of liquid; there is no standard volume of carbonation for the various type soft drinks. Soft drink classification based on the degree of carbonation has been defined as [6]: 3.5 or more CO_2 volumes (ginger ale, colas, tonics, club soda), 2.5 – 3.5 CO_2 volumes (root beer, lemon, lime, cream soda, grapefruit), and 1.0 – 2.5 CO_2 volumes (strawberry, cherry, grape, orange, pineapple, punch).

There are virtually hundreds of carbonated soft drink beverages on the market with scores of natural and synthetic flavors. The ingredients of a typical carbonated soft drink include: purified water that has been saturated with CO_2 gas, sweetening agents (dry or liquid sugars, non-nutritive sweeteners), acids (acetic, cit-

ric, fumaric, gluconic, tartaric, phosphoric), flavors (derived from fruit, vegetables, bark, roots, leaves, or artificial flavors), color (natural or artificial), preservatives, and other optional ingredients (e.g., vitamin C).

Non-carbonated. A non-carbonated, nonalcoholic beverage cannot be sharply defined. Simply stated, commercially-made and natural (coffee, tea, bottled water) non-carbonated beverages are those which do not contain pressurized carbon dioxide. In contrast to soft drinks, non-carbonated beverages contain components derived from natural fruit, vegetable, meat, or dairy products. The most popular of the non-carbonated types are the fruit beverages; these are found under a variety of names – fruit drinks, juice drinks, breakfast drinks, punches, ades, treats, etc. Modifications of non-carbonated fruit beverages include the electrolyte drinks – Gatorade, Sunshine Orange Thirst Quencher [7], fortified drinks (mostly with vitamin C), and protein-enriched drinks (protein-enriched citrus juices, fruit-milk beverages). In addition to the fruit beverages, there are non-carbonated beverages of the broth type (clam, meat), vegetable (vegetable juice cocktails, vitamin fortified), and dairy (chocolate drinks).

2. Mineral Water

History [8–11]. Water is an essential element of life and health. Although humans could survive for more than 30 d without eating, they would die of thirst after 8 – 10 d without drinking.

The first civilizations grouped around sources of water, and considerable engineering efforts were spent on the construction of aqueducts, e.g., by the Romans. Since ancient times, famous mineral springs were praised, e.g., by Homer or Pindar. The healing power of the sulfur springs of Thessalia was mentioned by Aristotle. Hannibal sent his soldiers to the Pyrenees, where they recovered at the mineral spring of "Perrier", which rises at Vergèze in Southern France. The legionaries of Caesar cured themselves at the springs of Plombières and Vichy, where the first medical facilities were installed.

Pliny the Elder was the first to classify mineral waters according to their *nature* (hot, warm, cold, sulfurous, aluminous, nitrous, bituminous, ferrous, and carbonated), and according to their *therapeutic effect* (purgative, purifying, stomachic, digestive, fortifying, recalcifying, reconstituent), or their efficiency against gout, sciatica, fractures, dislocations, injuries, fever, gallstones, or infections of the throat, ears, skin, or eyes.

Mineral waters were administered in the same way then as today, i.e., by drink, inhalation, bath, shower, sauna, etc.

The fall of the Roman Empire was accompanied by a decline in the use of mineral springs. This trend was not reversed until the 16th century. In 1605, Henry IV installed the "Superintendent General of Spas and Mineral Springs of the Kingdom." In the same year, one of the first treatises on modern hydrology appeared: *"De l'usage et l'employ des eaux naturelles contre les maladies"* (On the use and application of natural waters against diseases) by Jean Ban.

2.1. Definitions of Mineral Water

2.1.1. European

This definition appears in the European Economic Community (EEC) guideline 80/777 of July 15, 1980 [12].

1) *Natural mineral water* is water free of harmful bacteria, originating from a subterraneous layer, and emanating from a source exploited by one or more natural or drilled outlets. Natural mineral water is clearly distinguished from ordinary drinking water by (a) its contents of minerals, trace elements, or other constituents, and sometimes by certain therapeutic effects; (b) its original purity. These characteristics are preserved because the water is underground and therefore sheltered from pollution.

2) These characteristics, which by nature give the natural mineral water properties beneficial to health, must be evaluated on the following levels: (a) geological and hydrogeological, (b) physical, chemical, and physicochemical, (c) microbiological, and (d) if necessary, pharmacological, physiological, and clinical.

(a) Geological and hydrogeological regulations require:

- exact location of recovery, i.e., the altitude and the topographical level, on a scale of 1 : 1000 or greater;
- detailed geological report on the origin and nature of the ground;
- description of the hydrogeological layer;
- description of the recovery operations;
- determination of the zone or of other measures for protecting the source from pollution.

(b) For physical-chemical and physicochemical regulations, the following must be determined:

- output of the source;
- temperature of the water at the outlet and the ambient temperature;
- relations existing between the nature of the ground and the type of mineralization;
- dry residues at 180 – 260 °C;
- conductivity at the given temperature;
- pH value;
- anions and cations;
- nonionized elements;
- trace elements;
- radioactivity at the outlet;
- if applicable, the relative proportions of isotopes of the constituent elements of water: oxygen (^{16}O, ^{18}O) and hydrogen (^{1}H protium, ^{2}H deuterium, ^{3}H tritium);
- toxicity of certain constituents of the water, taking account of limits fixed for each constituent.

(c) The microbiological regulations for the outline concern particularly:

- demonstration that parasites and pathogenic microorganisms are absent;
- the quantitative determination of revivable microorganisms as indicated by fecal contamination. The absence of the following microorganisms must be demonstrated: *Escherichia coli* and other coli strains in 250 mL at 37 °C and 44.5 °C; fecal streptococci in 250 mL; sulfite-reducing aerobic spores in 50 mL; *Pseudomonas aeruginosa* in 250 mL;
- the determination of the total content of revivable microorganisms per milliliter of water.

(d) The water must be subjected to clinical and pharmacological tests, following recognized scientific methods. According to the relevant regulations, the nature of the tests must be adapted to the natural mineral waters' own characteristics and to their effects on the human organism, such as diuresis, effects on gastric or intestinal functions, and compensation for mineral deficiency.

The characteristics of the water must be determined according to scientific methods approved by the authority responsible. The tests specified under (b) – (d) can be voluntary when the water shows characteristics of composition conforming to those of a water considered as a natural mineral water in the member state of origin before application of the present directive started. This is particularly the case when the water contains, at the origin and after bottling, more than 1000 mg of total solids or more than 250 mg of free carbon dioxide.

2.1.2. French

The term mineral water is reserved exclusively for spring water from deep, protected sources or from drilled sources subject to authorization and control by the mining office (Ministry of Industry) and whose exploitation is authorized by the Minister of Health after its therapeutic properties have been recognized by the Academy of Medicine (see also [13]).

As a result of the particular regulation applied, mineral waters are not required to conform to the norms defined for drinking water with respect to content of mineral salts or total hardness.

Mineral waters are compared neither to food (as drinking water), nor to dietetic products, nor to medicaments.

Normally, the harmlessness of mineral water is guaranteed by compulsory authorization for bottling (decree of 1964), which is only granted for waters whose sale poses no danger for the consumer.

All possible precautions must be taken to preserve all physiological, bacteriological, and pharmacological qualities of the water after recovery.

Nevertheless, mineral waters are very diverse in composition and effects. They can be either still or carbonated. In the latter case, the quantity of gas in the bottle is the same as the orig-

inal content from the source. These waters are termed "natural carbonated mineral waters" and are symbolized by "Perrier".

Two classifications for still waters have been adopted:

1) Classification according to *composition,* which is based on chemical content rather than on therapeutic criteria. Thus distinction is made between waters weakly or very weakly mineralized and those rich in mineral salts, those that are bicarbonated, sulfated, chlorinated, fluorinated, or rich in calcium, magnesium, or sodium.
2) Classification according to *therapeutic effects,* based on therapeutic properties: diuretics, laxatives; treatment of rheumatism, arthritis, lithiasis, diabetes, obesity, gout, high cholesterol levels, and ailments of the kidney, cardiovascular system, liver, stomach, and intestines.

2.1.3. Codex Alimentarius

Control of mineral waters by the Codex Alimentarius is under development. Initial studies retain certain elements of the European definition, but they do not appear to mention the properties beneficial to health, which appears to be a fundamental omission.

Of the three definitions discussed in this section, the French legislation appears to be the best because of its precise wording (see Section 2.3), and because its has benefited from long experience. Authorization for production of mineral water in France can only be granted after complete hydrogeological, chemical, and bacteriological studies lasting several years, carried out under the supervision of the ministries of agriculture, industry, and health, together with active participation of the Academy of Medicine and of the Superior Council of Hygiene.

The European legislation is precise in its definition but does not mention the authority which grants the authorization, nor the methods accepted. It systematically classifies highly mineralized water (more than 1000 mg/L) as mineral water. As for the Codex, this includes more or less everything from mineral water to simple source waters.

In the United States water is only designated as natural mineral water if it contains more than 500 mg/L of dissolved minerals.

2.2. Development of Mineral Water

Thermal Springs and Bottling. Over the centuries, consumption of mineral water was primarily the privilege of inhabitants and clientele of thermal springs. In the last decade, this empiricism has given way to methodological application of thermal therapy, which has become the subject of regular instruction in French medical faculties. Many of the doctors specialize in natural medication. They have succeeded in developing appropriate thermal cures for various illnesses. In France more than 500000 persons are treated each year at the thermal springs. Details of treatment and different techniques used can be found in [8–11], [14–19].

After 1918 mineral waters became increasingly available to the average consumer. Improved packaging methods coupled with increasing demand permitted large-scale commercialization. Patients at the thermal springs could benefit from these waters all year and could prolong the effects of their cure. It also resulted in new clientele concerned about health.

The need for a pure product resulted from "quality defects" of public water, which is initially impure and requires numerous treatment stages. Even if such treatment eliminates bacterial pollution, it still leaves traces of chlorine, which on one hand is considered by some to be carcinogenic and on the other hand gives the water a particularly unpleasant taste.

Also, the market for bottled mineral waters, which was only some 0.35×10^9 L in 1948, represented more than 10×10^9 L in Europe and the West in 1983. France sold more than 3.25×10^9 L, the Federal Republic of Germany ca. 2.6×10^9 L, and Italy 1.7×10^9 L. In terms of per capita consumption, France leads with more than 56 L/a, followed by Belgium (54 L/a) and the Federal Republic of Germany (46 L/a).

As a consequence of this considerable commercial expansion in Europe, the demand for local or imported mineral water increased all over the world. The most striking illustration of such an evolution is provided by the worldwide distribution of Perrier, a naturally carbonated mineral

water, covering the markets not only of Europe, but also of the United States, Japan, Australia, etc.

2.3. Legislation [12], [20], [21]

As a consequence of the growing popularity of bottled water, the French legislature formulated the decrees of April 30, 1930, March 28, 1957, and August 1, 1961; the decisions of August 1, 1961, and December 21, 1964; and the circular of July 23, 1957. These general texts, briefly summarized herein, effectively control French mineral water. Authorization for mineral water is granted to natural sources by the Academy of Medicine after examination of the dossiers from the Mining Office, the Superior Council of Hygiene, etc. The dossiers consist of: (1) geological studies and recovery work; (2) physiological and bacteriological analysis of the water; and (3) pharmacological, physicological, and clinical tests.

The water is subject to numerous controls at the outlet, in the thermal installations, during bottling, and during commercialization. These controls are carried out by the official laboratories, approved by the decrees of August 10, 1961, and December 21, 1964, and repeated by the directive of the EEC [12].

Bottling is subject to authorization; the criteria defining the norms of bottling are defined by all the existing regulations, including the draft regulations of the Codex Alimentarius and the French regulations [20] (Regulations 57404 of March 28, 1957 and 641255 of December 11, 1964; decisions of August 10, 1961 and December 21, 1964; circulars of July 23, 1957, October 24, 1960, and January 27, 1965), and the German draft regulations (DBMJFG 416/415 6411-I of February, 1983, draft regulations on natural mineral, spring water, and table water; Mineral- and Tafelwasser Verordnung), which retain the main points of the European directive 60/777 [12].

Certain regulations, such as the Codex, the German regulations, and even the European directive, tend to limit various minerals present in mineral water. This is contrary to the French concept of mineral water, which is that certain ions, in quantities much above those recommended for public water, are responsible for the beneficial effects.

2.4. Packaging of Mineral Water [22]

Mineral water must be bottled at the outlet, and can in no way be processed. Some companies have been granted special authorization to remove iron from their water to eliminate precipitation of brown ferric oxide. This can only be done if other mineral components are not affected. Some companies are allowed to reincorporate carbon dioxide if technical procedures require separation of this gas from the water before bottling. The percentage of carbon dioxide in the bottled water must be the same as in the original source water. In fact, if this water is transported long distances, a constant gas percentage identical to the original cannot be maintained, especially when the amount of gas is high.

Reincorporation of carbon dioxide generally involves collecting the gas at the fountainhead, storing it at high pressure in a tank, and eventually mixing the gas with the water during bottling.

Transportation of the water through a pipeline to a bottling plant close to the fountainhead requires a ministerial license with approval from the Academy of Medicine.

Containers must be transparent (apart from those intended for special markets, such as the airline companies, which demand cans). The maximum capacity of containers for retail sale must not exceed 2 L. Use of barrels and tanks is prohibited. Until 1968, glass bottles were the only authorized containers. Now packaging has been revolutionized with development mainly of poly(vinyl chloride) bottles. The plastic bottles must be manufactured in the bottling factory. Those used for plain and low carbonated mineral water are subject to government approval [20]. The plastic bottle must be tested separately with each mineral water to make certain that all constituent materials are inert and will not affect the properties and features of the water.

The tests, lasting at least three months, are effected by the National Health Laboratory in Paris. One of the main points is detection of vinyl chloride monomer (VCM) (European Directive of January 30, 1978). Up to now, no VCM has ever been detected in mineral water, the minimum detectable quantity being 0.002 ppm [23].

2.5. Regulations for Labeling

Labeling regulations are very precise. However, because there is no significant difference between the various sets of regulations, only the European directive of December 18, 1978 will be mentioned. This Directive specifies that any label, package, or advertisement must in no way be likely to delude the consumer. It also mentions the type of information that must be available to the consumer: (1) name of the product on sale, i.e., natural mineral water (a brand name, a trademark, or fictitious name cannot serve as a substitute), together with the name of the source and of the bottling site; (2) ingredients; (3) net volume.

2.6. Bottling

During the last few years, bottling methods have become increasingly more efficient because of the ever growing use of disposable (nonrefillable) bottles. In modern factories, these bottles represent 90 % of the production. The production rate on highly automated bottling lines is about 30000 to 50000 bottles an hour. Bottling lines using plastic bottles are called "integrated" lines, because these bottles are manufactured on site. Mineral water companies working with glass bottles are dependent on supply from external bottle manufacturers. The Perrier Co. is an exception to this rule. It has integrated a glass bottle factory with its bottling unit. This company now controls all stages of manufacturing.

2.7. Conclusions

Mineral waters are of considerable interest and are often misunderstood.

Health Interest. Millions of patients are received annually at thermal springs; vast quantities ($> 10^9$ L) of mineral waters are consumed annually.

Economic Aspects. An industry, directly or indirectly employing millions of people, has been created to satisfy the demand for mineral water by consumers, increasingly aware of their beneficial effects.

3. Fruit and Vegetable Juice Products

3.1. Temperate Fruits

Temperate fruits are grown on almost every continent, from the cool northern and southern temperate zones to the warm equatorial tropics where sufficient altitude permits growth. These fruits enhance the diets of millions upon millions of people by providing nutrients, diversification of diet, good taste, and eating enjoyment. The major temperate fruits covered in this section are apples, apricots, cherries, grapes, peaches, pears, and plums. World production figures of these temperate fruits in selected countries are shown in Table 1. Apples and grapes are the two principal temperate fruits.

The Federal Republic of Germany, Italy, and France are all major European *apple* producers. Even countries that are not major contributors to world apple production, such as Switzerland, find apples to be the most important fruit for fruit juice processing. The United States is the largest apple producer. Washington is the leading state with New York second. European apple crops show large variation in production so that in years of excess, apples normally used for the fresh market are processed into juice. In years of short supply, apples are imported from the United States and elsewhere.

Grapes have been cultivated for more than 5000 years. They are the largest in volume of any temperate fruit. Of the grapes grown in Europe, about 85 % are for wine, 15 % for fresh market, and only a negligible amount for grape juice or other products [27]. The United States produced 5.9×10^6 t in 1982. Of this California produced almost 93 %. The California grape crop was utilized as 12.7 % fresh, 28.5 % dried, and 58.2 % wine. Grape juice is processed from grapes grown in Washington, New York, and Michigan, with 44.5 %, 24.7 %, and 15.3 %, respectively, of the 316×10^3 t utilized for United States grape juice production [24], [25].

Apples are used as the example temperate fruit in this section because of their economic prominence and the wealth of scientific information on all phases of apple production and processing. Although grapes are the largest fruit

Table 1. Percentage of 1982 world production by select countries of major temperate fruits – 1982 [24–26]

	Apples	Apricots	Cherries	Grapes	Pears	Peaches and nectarines	Prunes and plums
Austria	1.7	2.5	2.5	1.0	1.1	0.3	0.2
Belgium and Luxembourg	1.4	–	1.0	0.02	1.9	–	–
Canada	2.5	0.4	1.2	0.1	0.8	0.7	0.4
United Kingdom	2.0	–	0.7	–	0.9	–	1.1
France	10.5	10.0	9.7	15.9	8.4	8.1	6.4
Federal Republic of Germany	15.3	0.4	25.4	2.7	11.6	0.7	22.6
Italy	12.3	20.5	12.0	15.8	22.0	29.5	5.7
Netherlands	2.2	–	–	0.004	2.1	–	–
Switzerland	0.7	–	–	0.3	0.4	–	–
United States	20.6	14.4	24.6	8.3	14.9	22.5	17.6
Other	30.7	51.6	22.5	55.9	35.8	38.1	42.4
Total, kt/a	18 110	707	1158	70605	4901	5324	2974

crop produced in the world (70×10^6 t), the majority are consumed fresh or go into wine production (→ Wine). Only those aspects of grape production and products related to nonalcoholic beverages, such as grape juice, are covered here.

Before discussing the quality of the fresh fruit going into processed products, it is important to know some of the major factors influencing the production of temperate fruits. This is necessary because processing can only preserve, not create, good quality. Quality begins with production.

3.1.1. Production

From earliest times, humans have selected and husbanded the plants in their surroundings. They took plants to and from the places they visited and constantly sought ways of increasing their variety and yield. For example, apples, taken from their origin in China, are grown around the world. Through selection, grafting, and hybridization, today's apple is of much better quality than its ancestors.

Some of the major factors to consider in producing high-quality fruit which leads to high-quality beverages are: *Temperature.* Extremely high temperature during growing, or freezes before harvest will decrease quality. *Rainfall.* Heavy torrential rains or long dry summers make it difficult to produce high-quality fruit. *Cultivar.* Selection of scion and rootstock are responsible for genetic control of quality. Cultural practices include every step from selection of resistant cultivars and site selection to the timing of spraying. Growers are responsible for controlling pesticide applications to be sure the fruit contains less than the maximum pesticide tolerance at harvest. Additional information can be found in [28–30]. Further details on apple harvesting can be found in [31], and on grape harvesting in [31], [32].

3.1.2. Harvesting and Transportation

Careful harvesting and rapid transportation to processing facilities ensure that a high-quality starting material will be available for beverage manufacture. Depending on the type of fruit and location in the world, between 30 and 60 % of the fruit's selling price is due to harvesting and transportation costs.

3.1.3. General Fruit Quality

Fruit quality is a multifaceted property which depends partly on the person judging the quality and on the intended use of the fruit. There are cultural aspects of fruit quality like annual bearing, pest and disease resistance, hardiness, and good yield of the trees and a high return per hectare [29]. To be of the highest quality, fruit must be harvested at the correct stage of maturity which may last from as short as 1 or 2 days, but rarely longer than 1 to 2 weeks. There are fresh fruit aspects of quality, such as good handling and storage characteristics, freedom from internal (physiological) defects, and minimal external defects like mechanical injury and dis-

eases. High-quality fruit must have all pesticide residues below prescribed limits.

Postharvest quality deterioration includes: *Aging.* Most fruits are highly perishable, living commodities which respire, oxidizing their internal energy reserves. *Desiccation.* A 5 % loss of water will not only detract from a fruit's appearance but will result in a monetary loss when fruit is sold on a mass basis. A loss of water and turgidity can be caused by bruising during handling. *Decay.* The attack of microorganisms can affect a single fruit or the entire lot.

There are economic aspects to fruit quality. How much a buyer is willing to pay for fruit at a given time depends on the availability of supply, current and projected demand, and any premiums paid for deep rich color, desirable size, high sugar or vitamin content, small seed cavity, etc.

Fruits also have processing quality aspects. Processors need to remember the grower is ultimately responsible for fruit quality. While the processor can preserve the good quality of fruit he receives, he can do little with poor quality raw material. Processing quality includes those subjective attributes already discussed as well as objective measurements like percentage of juice, yield, ^{0}Brix (measurement of percent dissolved sugar or percent soluble solids), percent of acid, pH, color, flavor, and more. Before discussing the quality characteristics of individual fruit, it is important to have up-to-date, accurate, and specific information on government policies, grade, and product standards. Current information on quality standards can be obtained from governmental agricultural officials. Standards among countries in Europe and within the United States can be obtained from addresses given in Section 3.1.7.

Apple Quality. The desirable quality of apples to be processed depends to a certain extent on what product is being made and the properties of the individual apple cultivar. In many countries, the highest quality apples go to the fresh market because growers usually receive the most money there. Highest quality culls go into whole or sliced apples, with the remaining culls going into juice or apple sauce. The apples rejected from these operations are used to make vinegar.

Apples should be left to mature on the trees for optimum quality. Undermature apples are poor in flavor, are hard, and have higher starch and acid content. Overmature apples have a bland flavor, are soft, give low yields of juice, and are difficult to press and clarify [33]. Mature, firm apples should be harvested and handled carefully to minimize bruises and cuts. Bruises usually do not worsen during storage and do not permit increased decay but may require trimming. Cuts in the skin can permit entry of microorganisms that cause spoilage and decay. This in turn can affect loss during storage, product yield, and quality of the finished product [34]. Losses due to improper handling and storage can run as high as 20 %.

Only a few of the differences among apple cultivars are meaningful in utilizing a specific cultivar for a specific processed product. Soluble solids, mostly sugars, are measured by a refractometer and range from 10 to 15 ^{0}Brix. Organic acids, measured by titration, range from 0.25 to 1 %. The sugar-acid ratio (Brix-acid) has been found to be a good indicator of maturity, harvest timing, and product quality [35]. For European cider apples, an average sugar-acid ratio of 12 and total acid of 8 g/L (expressed as tartaric acid) is recommended [27].

Typical apple flavor is important in some products and is formed by a combination of acids, sugars, tannins, and aromatic substances. The quality factors of aroma and taste are affected by differences in cultivars, maturity, starting fruit quality, processing parameters, and storage conditions [36]. In other products like juice concentrate most of the volatile contributions to flavor are lost during evaporation, but can be replaced using commercially available flavors and essences. Maximum juice yield and pressability are important quality factors for apple juice. For additional information on processed apple products see [37], [38]. A list of recommended apple juice varieties for Germany can be found in [27]. The mean and standard deviation for ten parameters of quality have been measured for 15 commercial apple cultivars, grown in 8 seasons over a 3-year period [39].

Apricot Quality. Apricots harvested at optimum ripeness for processing are firm, deep yellow in color, and have good flavor. The firmness measured by a pressure tester ranged from 0.19 to 2.3 kg for apricots ready to process [40]. Fruit

left to mature after the optimum for processing reached maximum flavor and soluble solids but decreased in firmness and acidity [41]. Apricots for puree or nectar should be so ripe they are soft [42]. One of the major quality constraints of apricots is their variable maturity both within an individual tree and across the orchard. Three to four harvests may be necessary to get the best flavored, tree-ripened fruit. Minor to moderate bruising is not a large problem for processed apricot products so mechanical harvesting is satisfactory [43]. Small sized fresh fruit, with a count of more than 30/kg, are often graded out for canning as whole packed apricots or for production of nectar. Details on processing of apricot puree, nectar, and concentrate are found in [42].

Cherry Quality. There are two principal types of cherries: the first has several names, tart or red tart is a preferred name to sour or pie cherries; the second type is the sweet cherry. In the United States, 85 % of the tart and 15 % of the sweet cherries are mechanically harvested [44]. Growers strive for removal of more than 90 % of the cherries in a marketable condition, but late in the season, harvesting machines may greatly reduce the quality.

Tart cherries used for making juice or nectar need to be fully mature and dark in color to give a strongly colored juice. They also need to have a high acid content, good juice yield, and a typical cherry flavor [27].

Sweet cherries for brining and canning also need to be fully mature and dark in color. There is a pronounced increase in the sugars and soluble solids during maturation. In Oregon (USA) Napoleon cherries increase from 12 to 22 % soluble solids from first to last picking. There is a gradual decline in acid content with time. Napoleon cherries go from 0.64 to 0.35 % acid. Some juice processors feel sweet cherries are too low in acid and flavor to make good juice. There is a gradual increase in mass of 0.5 to 2 %/d. Napoleon cherries increase in size 36 % from beginning to full maturity. Size, however, is not one of the main guides to maturity. Firmness is an important quality consideration, but the pressure test is not a good indicator of maturity. Napoleon cherries decrease in resistance 17 % from immature to overripe. During the harvestable period, the decrease is only 6 %. Mature fruit is larger, firmer, more aromatic, and has a more distinctive cherry flavor. Overmature or cull fruit can give cherry juice an off-flavor, while underripe fruit produces a sour, poorly colored juice. Details on manufacture and composition of cherry juice, carbonated cherry juice, and cherry syrup are in [45]. An extensive review on cherry production with less emphasis on processing is available [46].

Grape Quality. There are three grape cultivars of commercial importance, each with its own quality characteristics and products. *Vitis vinifera* cultivars make up about 75 % of the grapes grown in Europe and more than 90 % of those in the United States. They are thick skinned, the skin adheres firmly to the pulp, and the pulp is firm and thoroughly sweet. This cultivar is grown in a relatively mild climate and an example is the Tokay. The *Vitis labrusca* has a relatively thin skin, with the skin slipping easily from the pulp (slip-skin). The pulp is soft and more acidic near the seeds. An example of this cultivar is the Concord, which requires a temperate climate. The third cultivar is *Vitis rotundifolia.* They grow singly or in small bunches, have large seeds, and ripen unevenly. They are Muscadines and have a 'musky' odor. This cultivar grows under a hot, humid climate. The vine is long lived and disease resistant. An example would be the Scuppernong cultivar [29].

Preferred grape maturity depends somewhat on the intended product and cultivar. Grapes should not be picked too early because there is no increase in sugar or color after separation from the vine. Grapes for fresh market are picked somewhat earlier than for juice, before berries begin to shatter from the bunch. Grapes intended for juice production are almost all of the *V. labrusca* type and are picked fully mature. Ripe Concord grapes can have total soluble solids ranging from 15 to 20 % measured by refractometer; for acid, measured by titration, less than 0.85 % tartaric acid is desirable. The minimum acceptable sugar-acid ratio is different for each cultivar [47]. Grapes for juice must possess a good fresh grape flavor. Although the Concord is widely used for juice, it has a strong astringent flavor sometimes described as "foxy." Additional information on grape quality is available [47], [48].

Peach Quality. There are two principal types of peaches, the cling and freestone. Clingstone, where the pit is firmly attached to the flesh, is preferred for canned half peaches and slices. Freestone can be used for nectars. Peaches for processing should have these qualities: large size, symmetrical shape, and a small pit. Yellow-fleshed cultivars with a firm texture to withstand harvesting and hauling are preferred. They should have a characteristic peach flavor and aroma. Peaches are harvested when they are firm and ripe. A significant amount of mechanical harvesting is used. Peaches in Europe are processed into pulpy nectars, which preserve their carotene content. Several peach beverages, freestone peach nectar, pressed peach juice, peach puree, and mixed fruit drinks are made in the United States. Details on these processes are found in [42], [49].

Pear Quality. Pears are grown principally for fresh market with a small amount being processed. Formerly, high-acid cultivars were made into an alcoholic pear cider [27]. Harvesting of pears for use in canning begins when they have reached their maximum size but are still hard and green. Harvesting of pears according to a plunger pressure tester, 8.2 – 9.7 kg, storage at 1 °C for a week or more, ripening at 21 °C at 85 % natural humidity to a 0.9 – 1.4 kg pressure reading gives the best canned pear products [42], [49]. Pears handled in this manner give better quality product than those without 1 °C storage. Preparation of pulpy pear juice, nectar, clear juice, and aseptic canned puree are given in [42], [50].

Plum and Prune Quality. Plums and prunes are terms often used interchangeably for different cultivars. Most plums are canned whole. Most prune cultivars cannot be juiced by pressing and are mostly dried. Selected cultivars of fresh prunes with high acid, and distinctive flavor and aroma can be processed into a pulpy nectar [27], [42]. Dried prunes can be water extracted to produce a prune juice which is rich in minerals and has a mild laxative effect. Details on processing these fruit are found in [42], [49].

3.1.4. Apple Processing

For more than 100 years, apples have been processed by drying and by conversion to cider. This was done to utilize windfall or drops which were low in quality and may have been decayed. Today, high-quality apples are processed into juice, concentrate, fermented cider, wine, apple sauce, and vinegar. Apples are also dried or dehydrofrozen, canned, and baked.

Apple Juice. There are several traditional methods of producing apple juice or unclarified apple cider. These have evolved into modern processing techniques in Europe and have been adapted to the situation in the United States. A generalized flow diagram for apple juice processing is shown in Figure 1 (for additional details, see [38], [51], [52]). Apples from storage, fresh fruit eliminations, or apples directly from the orchard go to the processing plant. First, leaves, grass, and rotten fruit are eliminated. The fruit is thoroughly washed, sorted, and may be trimmed if this can be done profitably. Removal of partly or wholly decayed fruit may be the single most important factor in producing a high-quality apple juice. If the decayed fruit is not eliminated, it produces off-flavors, increases the probability of microbial spoilage, and introduces possible microbial toxins into the finished product [50]. Leaves and grass must be separated as they also can cause off-flavors in the juice. As a general rule, at least three cultivars of 'eating ripe' apples are blended prior to grinding. This gives a desirable blend of the sugar-acid ratio and a preferable aroma. Processing a single cultivar and blending the juice after it is concentrated is not economical.

Sound fruit are sent to a grater or hammer mill to reduce the apples to pulp. If the pulp is too coarse, the juice will be low in yield and soluble solids; if the pulp is too fine, it is difficult to press and the juice has an excess of fine particles.

Enzymatic browning of ground apples takes place in a few minutes and needs to be controlled by inhibiting the polyphenolase enzymes with heat, ascorbic acid, or sulfur dioxide. Ascorbic acid at 0.05 % can be added directly into the grinder but the reaction is difficult to control. The tannins may be oxidized, destroying the desirable amber color. Next the juice is ready to be extracted from the pulp.

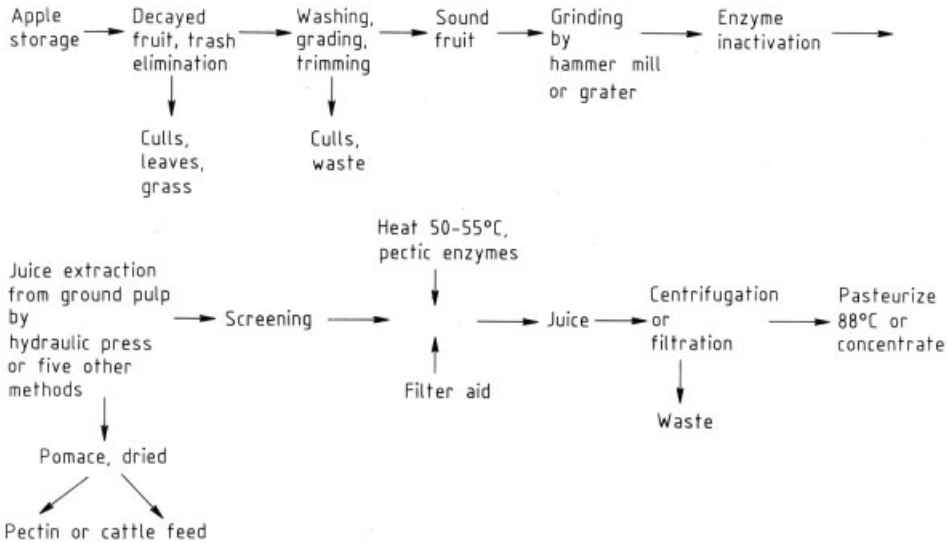

Figure 1. Production of apple juice [38], [51], [52]

Juice extraction is very labor intensive, so six methods have been developed and used to reduce labor and increase yield from this operation [51], [52]. Apple juice manufacturers select the method that best meets their criteria.

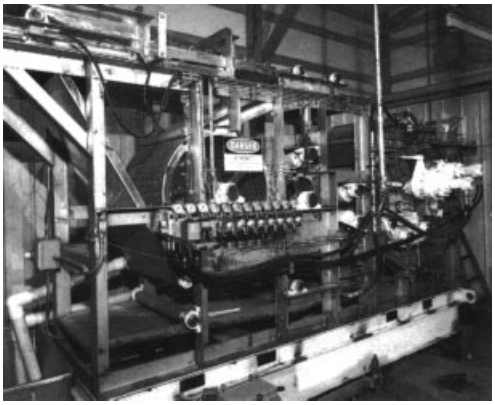

Figure 2. Newly designed belt press. Apple mash blended with a press aid enters upper left. It is lightly pressed to remove the "free run juice." The press cake is then fed into a high-pressure section which removes the remaining juice. ∗ Courtesy Eimco Process Equipment Co.

1) The *hydraulic press* uses a press cloth placed inside a square wooden rack. The cloth is filled with ground apple pulp then the cloth is folded to completely surround the pulp. A series of racks and cloths are stacked together in preparation for pressing. Pressing usually takes place in two stages. The first, at about 4000 kPa, is applied rapidly to remove the "free run" juice. The second pressing, up to 19000 kPa, is applied slowly and maintained to remove the remaining juice.
2) A *pneumatic press* is composed of a large air inflated rubber bladder and a horizontal screen lined with a press cloth. Pulp is pumped into the space between the bladder and press cloth. Air inflates the bladder and the juice is pressed through the cloth and screen.
3) In the *continuous screw press,* wood pulp as a "press aid" is mixed with apple pulp. This mixture is pressed by the screw past a cone, which partially blocks the exit from the press. The restriction of the cone creates backpressure in the press, forcing the juice from the pulp through the surrounding screen.
4) The *continuous belt press* has apple pulp together with a press aid placed between two continuous belts, then pressed between rollers with increasing pressure. The juice flows out, the pomace is discharged, and the belts are fed back for more pulp (Fig. 2).
5) The *horizontal basket press* is composed of a horizontal cylinder with screen. At one end is a pressing plate and at the other end a stationary press plate. Nylon hoses connect between the two plates. The cylinder is filled with pulp and the pressing plate alternately compresses

so the juice can flow through the cylinder and retracts, allowing the nylon hoses to break up the pomace.

6) The *screen type centrifuge* is a cone-shaped, self-cleaning centrifuge. The pulp is fed in the top and the juice pressed through the screen by centrifugal force.

The pulp exiting the bottom is normally given a second pressing to increase yield.

Pomace from the presses is dried and used to make pectin or cattle feed. Apple juice yield from the presses ranges from 58 to 70 % of the mass of the fruit. The juice is passed over fine screens to remove coarse particles. Pectic enzymes are added to hydrolyze the water soluble pectin, which in turn reduces the viscosity making the juice easier to filter and concentrate [53], [54]. Ultrafiltration can be used to clarify the juice (Fig. 3), or filter aid, usually diatomaceous earth, is added and the juice filtered to give a clear liquid. Screened juice can also be centrifuged to remove most of the small particles, leaving an opaque juice. The juice can be evaporated to give juice concentrate, about 72 °Brix, or pasteurized at 88 – 90 °C for 25 – 30 s to make single-strength juice. Apple essence and flavor compounds are used to add bouquet and increase consumer acceptance. Apple juice is stored cold (4 – 6 °C) to reduce development of browning and off flavors [37]. One report found no difference in the sensory and chemical evaluations after 12 months storage at 3 °C. There were significant aroma changes after 12 months storage at 20 °C [36].

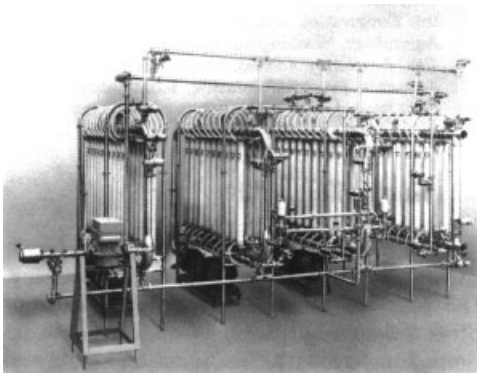

Figure 3. Four-stage, 80 cartridge, sanitary, ultrafiltration system (197 m^2) – used to treat depectinized apple juice (2 – 5 % solids) to produce a "sparkling" clear juice with reduced microorganism counts
∗ Courtesy Romicon, Inc.

Hard Cider and Wine. Alcoholic hard cider, sometimes referred to as cider, and apple wine are covered in more detail under → Wine. European hard cider apples are usually small in size, astringent in taste, and distinctive in flavor. Higher soluble solids means less sugar has to be added before fermentation. American preference for less strongly flavored cider means fresh market apple cultivars may be used. The process is similar to that just described with an additional alcohol fermentation production step.

3.1.5. Grape Processing

World grape production is about 70×10^6 t/a, about 60 % of which is sold fresh, about 40 % goes into wine and 1 – 2 % into raisins and other grape products. Within the EEC, grape hybrids for juice are just being developed. For the United States in 1982, the amount of grapes utilized for raisins was only about half that going into wine; but there were four to five times as many raisins as all other processed grape products combined. Processed grape products besides juice and raisins are juice concentrates, drinks and blended fruit drinks, fruit cocktail, and jelly [24].

Grape Juice. The principal grape cultivar for juice is the Concord. It gives the dark purple anthocyanin color, the astringent-sweet flavor, odor, and other characteristics necessary for quality juice [55].

Processing

is similar to apple juice extraction except that a hot extraction is necessary to obtain the rich purple color, derived from a complex mixture of anthocyanins. Grapes are gently washed with water sprays. Crushing takes place inside a perforated drum where rotating blades and beating bars break the grape's skin but not the seeds and stems. The rotating blades produce a centrifugal action which forces juice and pulp through the slotted drum while the stems and some seeds are discharged out the opening. The juice and pulp are extracted by heating at 60 – 63 °C while being stirred with pectic enzymes in a steam-jacketed kettle. Heating time and enzyme concentration is adjusted to extract the desired color and solubles. Traditionally juice has been

pressed from the pulp by a plate and frame press. Hot pulp and juice are piled on a press cloth which is folded to surround them. A second press cloth is placed on the first and the process repeated. About 50 % of the juice freely flows from the pulp before pressing. Hydraulic pressures of 1380 – 2050 kPa are used to complete the separation of juice, which may yield up to 90 %. Plate and frame presses require considerable time and labor to load and unload plus the juice can pick up considerable microbial contamination from the press cloths. Several other types of presses (discussed in the apple juice section) have been developed to try and overcome these problems. The juice is then pasteurized at about 82 °C, cooled to about 0 °C, and held in a tank for a month to allow the tartrate salts (argols) to crystallize before final clarification and bottling. More information can be found in [49], [56].

3.1.6. Nutritive and Chemical Composition

Temperate fruits are composed of 75 – 85 % water unless they have been dried; then they are only about 5 – 30 %. Water provides the turgidity and crispness when these fruit are eaten fresh. It is the major component of fruit juice and gives shape and mass to other processed products.

The two main dietary contributions of temperate fruits are carbohydrates and vitamins. *Carbohydrates* supply the body's need for food energy. Fruit carbohydrates are mainly in the form of sugars like fructose, glucose, sucrose, and to a lesser extent starch. The mature, 150 g apple shown in Table 2 contained 126.6 g of water; 20 g of the remaining 23.4 g, or more than 85 % of the dry mass, was carbohydrates. Apple juice contains about three times more fructose (4.4 – 8.2 %) than glucose (1.3 – 2.0 %) or sucrose (1.7 – 4.2 %) [52]. The ratio of these sugars is one of the important parameters in determining the authenticity of fruit juice [58].

Vitamin C (ascorbic acid) and vitamin A are present in temperate fruit (Table 2). Some of the richer sources of *vitamin C* are apricots, peaches, and cherries. Vitamin C is essential for healthy teeth, bones, and gums, and it aids in wound healing. Those factors affecting vitamin C content of the fresh fruit have been reviewed [59]. Apple juice may be fortified with vitamin C because apple juice contains so little compared to citrus juices, and also it helps reduce settling in cloudy juice.

Vitamin A occurs as carotene in apricots and peaches. It is important for vision and maintaining healthy linings of the mouth, nose, throat, and digestive tract.

Temperate fruits contain small amounts of several *minerals.* Dried fruits contain iron. From 150 g of raisins, 9.5 mg of iron are supplied, 150 g of dried apricots provide 8 mg, and 150 g dried peaches 5.3 mg. Reviews of the nutrition and composition of raw and processed products from temperate fruits have been published [60], [61]. Processing involving heat and storage has a complex effect on nutrient retention [62].

3.1.7. Regulations

Regulations from an international to a local level are critical to temperate fruit processing and marketing. In order to get the most relevant and up-to-date information, the following organizations should be contacted. Background information for regulations within the United States can be found in [63].

Regulatory agency addresses:

– For standards within the United States:

1) *USDA,* Chief, Processed Products Branch, Fruit and Vegetable Division, Agricultural Marketing Service, U.S. Dept. of Agriculture, Room 0709, South Building, Washington, DC 20250, USA
2) *FDA,* Director, Division of Food Technology, Food and Drug Administration, 200 C Street S.W., Washington, DC 20204, USA

– For standards within the European Economic Community:

1) *CODEX Regulations,* Executive Director, International Organization Affairs, Room 6552, Office of Int. Cooperation and Development, U.S. Dept. of Agriculture, Washington, DC 20230 USA
2) *EEC Regulatory Address,* Mr. A. Korth, Verband der deutschen Fruchtsaftindustrie, Konstantinstr. 3, D-5300 Bonn 2, Federal Republic of Germany

Table 2. Composition and nutritive values for edible portion of major temperate fruits [57]

	Mass, g	Refuse Peel, Core, Stem, Seeds, %	H_2O, %	kJ	Protein, g	Fat, g	Carbohydrate, g	Ca, mg	P, mg	Fe, mg	Na, mg	K, mg	Vitamin A value, IU [a]	Thiamin, mg	Riboflavin, mg	Niacin, mg	Vitamin C, mg
Apples (1) 7 cm [d] raw, commercial with skin,	150	8	84.4	335	0.3	0.8	20	10	14	0.4	1	152	120	0.04	0.03	0.1	6
Apricots (3) raw	114	6	85.3	230	1.1	0.2	13.7	18	25	0.5	1	301	2890	0.03	0.04	0.6	11
Cherries (237 cm^3) raw, sour, red	155	10	83.7	251	1.2	0.3	14.7	23	19	0.4	2	196	1030	0.05	0.06	0.4	10
Grapes (237 cm^3) slipskin	153	34	81.6	293	1.3	1.0	15.9	16	12	0.4	3	160	100	(0.05) [b]	(0.03) [b]	(0.3) [b]	4
adherent, seedless	160	–	81.4	448	1.0	0.5	27.7	19	32	0.6	5	277	(160) [b]	0.08	0.05	0.5	6
Peaches (1) 7 cm [d] raw, whole	175	13	89.1	243	0.9	0.2	14.8	14	29	0.8	2	308	2030	0.03	0.08	1.5	11
Pears (1) 8.9 cm [e]																	
Bartlett 6.4 cm [d]	180	9	83.2	418	1.1	0.7	25.1	13	18	0.5	3	213	30	0.03	0.07	0.2	7
Boscs 6.4 cm [d]	155	9	83.2	360	1.0	0.6	21.6	11	16	0.4	3	183	30	0.03	0.06	0.1	6
D'Anjous 7.6 cm [d]	220	9	83.2	510	1.4	0.8	30.6	16	22	0.6	4	260	40	0.04	0.08	0.2	8
Plums, Raw Damson (10) 2.5 cm [d]	110	9	81.1	276	0.5	trace	17.8	18	17	0.5	2	299	(300) [b]	0.08	0.03	0.5	– [c]
Japanese and hybrid (1) 5.4 cm [d]	70	6	86.6	134	0.3	0.1	8.1	8	12	0.3	1	112	160	0.02	0.02	0.3	4
prune type (3) 3.8 cm [d]	90	6	78.7	264	0.6	0.3	16.8	9	15	0.3	trace	144	240	0.03	0.03	0.3	3
Prunes-dried (10) without pits	102	– [c]	28.0	1088	2.1	0.6	68.7	52	81	4.0	8	708	1630	0.09	0.17	1.6	3

[a] International Units.
[b] Imputed values usually from another form of the food or from a similar food.
[c] Lack of reliable data for a constituent believed to be present in a measurable amount.
[d] Diameter.
[e] Height.

3.2. Subtropical Fruits

Citrus fruits are one of the largest and most important groups of fruits for juice processing. The place of origin of nearly all species of citrus fruits was probably in warmer Asian regions: Southeastern China, India, and Malaysia. Nowadays citrus fruits are grown in all tropical and subtropical regions (i.e., between the latitudes 40 °N and 40 °S); the warmer areas of North America, South and Central America, the Mediterranean, the Far East, and a few other Southern Hemisphere countries, such as South Africa and Australia.

Important commercial citrus fruits can be classified into four main groups: (1) oranges, (2) tangerines, (3) grapefruits, (4) lemons and limes. World citrus production in 1983 was estimated at 55×10^6 t; oranges (69 %), tangerines (14 %), grapefruit (8 %), lemons and limes (9 %). More than 125 countries and territories in the world produce some type of citrus fruit. The following top ten producing countries account for nearly two-thirds of the total: United States (22.2 %), Brazil (18.6 %), Japan (6.6 %), Spain (5.5 %), Italy (4.8 %), Mexico (4.0 %), Argentina (2.8 %), Israel (2.6 %), Egypt (2.6 %), Morocco (1.8 %) [64], [65].

In 1946, a method for producing a high-quality frozen concentrated orange juice (FCOJ) was developed [66]. Juice was concentrated under vacuum at a low temperature, fresh cut-back juice freshly squeezed (nonpasteurized) was added to the over-concentrated product to bring it back to the desired concentration and to restore fresh flavor and aroma, and the final material was packed as a frozen product. This momentous discovery was the major stimulus for the development of more efficient juice extractors, finishers, evaporators, and other important juice processing equipment, and for the dramatic increase in worldwide orange production especially in the United States and Brazil.

Approximately 40 % of world citrus production is processed into juices and concentrates. Processed oranges and tangerines represented over 84 % of total processed citrus fruit in 1982. United States and Brazil account for nearly 80 % of the total juice production [64].

3.2.1. Quality of Fruit for Processing

From exterior to interior, the major parts of the citrus fruit are:

1) peel or rind (epidermis, flavedo, oil glands, albedo, and vascular bundles)
2) segments or sections (segment wall, juice vesicles, and seeds)
3) core (vascular bundles and parenchymous tissue).

*Flavedo*is the outer, colored skin which is covered with a thin, wax layer and contains essential oils as well as carotenoids (coloring matter). *Albedo* is the thick, white porous inner skin which is closely attached to the flavedo. The thickness of this albedo layer in orange and grapefruit is a predominant factor in determining the proper adjustment of extractor parameters for optimum quality and juice yield. *Segments* are composed of juice vesicles which are enclosed by the segment wall. *Seeds* are contained in most fruit segments but some hybrids are wholly or partially seedless. The *core* is the fruit axis. Bitter substances (polyphenols) are contained in skins, vesicle membranes, albedo, and core; consequently, extraction treatments minimize introduction of undesirable substances from these tissues into juice for processing [67].

In commercial operations, average juice yield from oranges varies from 43 to 60 %, with residues, i.e., peel, pulp, and seeds, left after juice extraction, 40 – 57 wt % [68]. The proximate compositions of various citrus juices are shown in Table 3 [69], [70].

The quality of citrus juice is judged by its color, flavor, °Brix/acid ratio, and defects. A good quality juice should possess good color (the appearance of fresh juice), should be free from defects (e.g., seed fragments, insect fragments), and should have good flavor, proper °Brix value and °Brix/acid ratio, and proper pulp and oil contents. Therefore, every step of processing must preserve good quality and minimize introduction of detrimental components.

The fruit used should be whole, mature, sound, and recently harvested. Before harvest, fruit for processing must meet certain minimum maturity requirements established by regulatory agencies. Those requirements may vary from one citrus-producing area to another, but standards are designated for nearly all the citrus

Table 3. Proximate composition of citrus juices, g/100 g [69], [70]

Juice	Water	Protein	Fat	Sugar	Acid	Vitamin C
Orange	88.3	0.6	0.2	9.3	0.6 – 1.6	0.05
Tangerine	88.9	0.5	0.2	9.2	0.6 – 1.6	0.04
Grapefruit	90.2	0.5	0.1	6.0	1.2 – 2.2	0.04
Lemon	87.6	0.4	0.2	3.2	4.0 – 6.0	0.05

processing areas throughout the world. Such requirements are usually based on: color break, minimum juice content, minimum acid content, minimum percentage of total soluble solids, and °Brix/acid ratio.

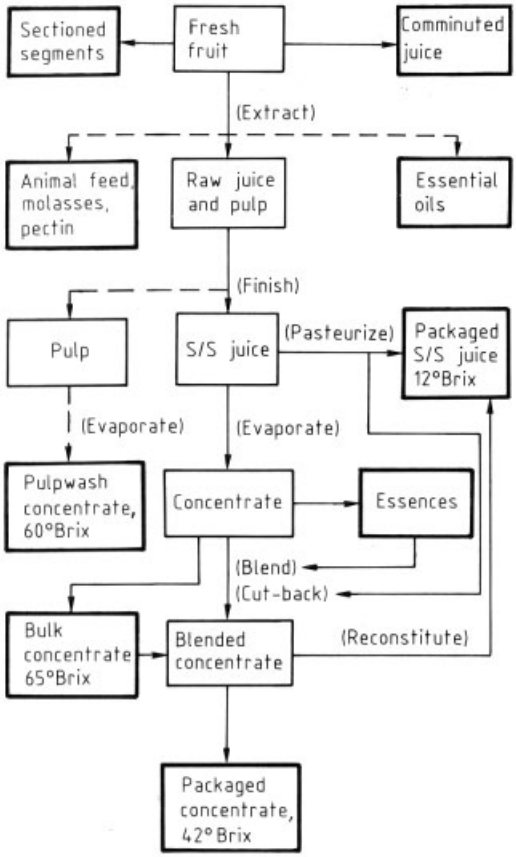

Figure 4. Basic elements of the citrus processing industry
▭ Intermediary product; ▭ End product; —- Juice line; - - - - - Byproduct line

3.2.2. Processing

Processing of citrus fruit can yield many different products (see Fig. 4). In the 1950s the citrus industry considered nonjuice components of fruit as having little or no commercial value. Juice remains the main product of citrus fruit, but now commercial processing of other components also is important economically. Processed citrus *juice products* include: single strength, concentrate, comminuted, orange pulpwash, bases, crystals, and segments. Production of pure juice, either single strength (S/S) or concentrate, probably accounts for over 90 vol%. Byproducts and other *nonjuice containing products* of citrus fruit include: oils, essences, D-limonene, peel and pulp for cattle feeds, molasses, alcohols, pectin, and flavonoids [71].

A basic citrus processing plant generally includes the following operations:

- bulk concentrate juice processing
- pulpwash recovery and addback
- cold pressed oil recovery combined with D-limonene recovery
- essence recovery
- animal feed processing from peel and rag wastes.

The variations to those basic operations include:

- concentrate packed in consumer sized containers
- single strength packed in consumer sized containers for the local market only both directly from fruit and reconstituted from concentrate
- bulk concentrate packed for remanufacture.

Major unit operations involved for citrus juice processing are: extraction, finishing and centrifuging, pasteurization, evaporation, freezing, packing, and storage.

Extraction. Three major types of juice extractors are commonly used in the citrus juice industry. They are fast, rugged, and easy to clean and to adjust to fruit of different sizes and characteristics.

Rotary Juice Press (Fig. 5). This was the first machine developed for commercial juice extraction. The process involves cutting the fruit in half, then passing the fruit halves between rotating cones and the cups which press out the juice. In another version the cut fruit is pressed between two rotating cylinders. In the latter case, oil has to be extracted prior to juice extraction. The major disadvantage of the method is the relatively low juice yield. The rotary press has been generally superseded by more modern and sophisticated equipment in most processing plants in the United States, but it is still used commercially in Italy, Spain, and Central and South America.

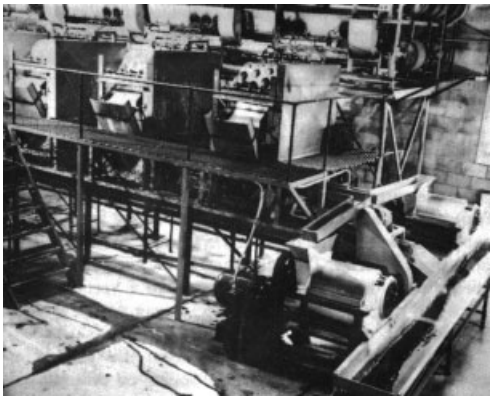

Figure 5. Rotary juice press
∗ Courtesy S. A. Bertuzzi

In-line Extractor, FMC (Fig. 6). This machine is used worldwide. The unique aspect of the FMC machine is that juices are extracted without first halving the fruit. The main steps involved are:

1) Fruit is placed in lower of two cups formed by a series of metal fingers.
2) Similar designed upper cup descends on fruit and as it closes the steel tube opens a hole in bottom of the fruit.
3) Fingers of two cups interlock crushing fruit and expelling juice, pulp, and some rag into a perforated steel tube.
4) Piston in tube compresses fruit flesh forcing juice and some juice sacs through perforated inner cylinder walls into juice collector tube. Remaining pulp and peel plug then eject from bottom of steel tube.
5) Peel and rag forced through cup fingers release oil, which is washed away and collected separately as a water emulsion. Peel is ejected from top of cups.

The major advantages of the FMC machine are its elimination of the halving step, its minimization of contact between juice and other fruit components, especially flavedo and albedo, its simultaneous oil separation, and its prefinishing ability. Both quality and juice yields are excellent depending on proper adjustment of the clearance of the inner pistons. Disadvantages are its cost and complexity. The fact it lacerates the peel can be either an advantage or a disadvantage depending on subsequent peel processing.

Extractors, Brown (Automatic Machinery Corp.). These machines are also used worldwide. There are several models of different sizes and capacities. However, the basic principles are similar. As in the case of the rotary press, fruit are initially cut in half. The halves are then held by rubber cups, while plastic reamers enter and express the juice from the fruit. This all occurs as the cups and reamers are rotating on a large turntable, or in a vertical plane. Juice and pulp collects in a pan, while rag and peel halves are ejected.

Both quality and juice yields from this process are excellent. Oil content is low because few oil cells (flavedo oil glands) rupture. Oil must be extracted separately. However, the company has developed an excellent machine for this purpose.

The disadvantages, like the FMC in-line extractors, are relatively high cost and complexity.

Finishing and Centrifuging. These two operations complete the basic separation process initiated and are largely accomplished by extraction. Both operations separate pulp from raw juice.

Primarily two machines are available for *finishing* raw juice. These are manufactured by the two major extractor manufacturers, i.e., Brown and FMC. Both machines operate on a rotating screw principle (Fig. 7). Raw juice enters at one end. Pulp is squeezed out at the other end, while finished juice passes radially through sieving cylindrical shaped screens. Finishers can also be adjusted to influence the amount of juice recovered. A paddle finisher (Fig. 8) provides more gentle finishing.

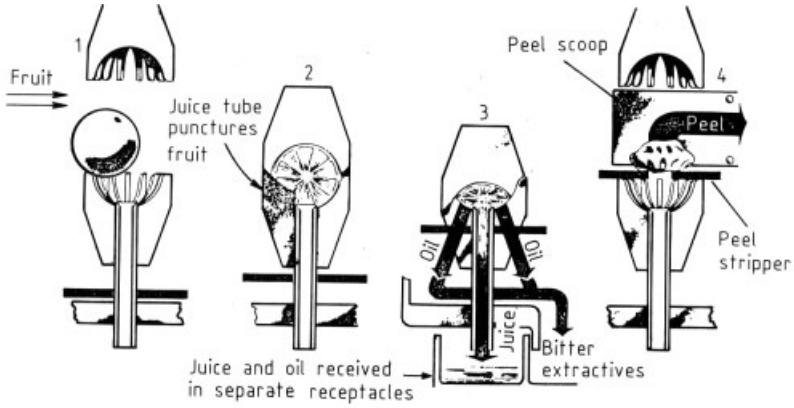

Figure 6. Principles of FMC in-line extractor
∗ Courtesy FMC Corp.

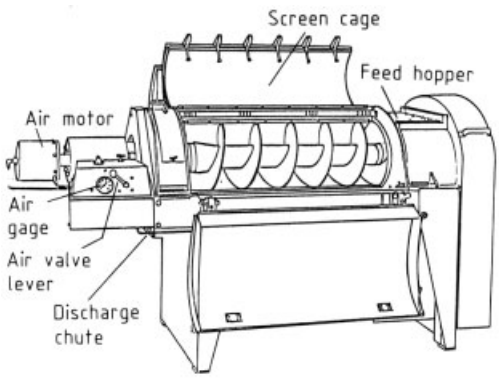

Figure 7. Brown model 2503 screw finisher
∗ Courtesy Automatic Machinery Corp.

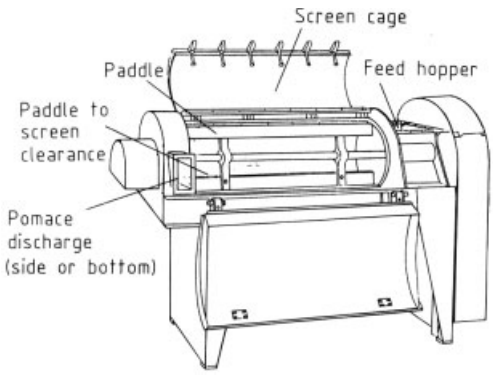

Figure 8. Brown model 200 paddle finisher
∗ Courtesy Automatic Machinery Corp.

The *centrifuge* completes the separation of pulp particles from juice prior to evaporation. Standard finishing is generally capable of reducing pulp content of juice to about 12 %. Florida juice specifications state this level is acceptable. As a result some Florida plants do not centrifuge their juice. However, European standards generally require lower pulp content (as low as 3 – 5 %). Only centrifuging can ensure these low levels.

Centrifuging offers other advantages. It eliminates undesirable particles, broken seeds, etc., which finishing will not remove. Thus, the resulting juice is of a higher quality. It also permits the juice to be concentrated to a higher °Brix with lower viscosity. However, centrifuging does reduce overall yield.

Evaporation. Vacuum evaporation of water is the key concentration activity. In theory, water can be removed by freezing as well as by evaporation. In practice, evaporation is the only economically feasible method in present commercial use for citrus.

Freeze concentration has several advantages over evaporation:

– Low process temperatures, normally between 0 °C and – 15 °C depending on the type of product, result in better product quality.
– The absence of liquid-vapor interfaces, combined with a totally closed system, results in no losses of volatile aroma components.

The disadvantages are higher solids loss and higher energy costs than by evaporation. An

improved freeze concentration process, developed by Grenco Special Applications B.V. in the Netherlands, may overcome the drawbacks of freeze concentration. Freeze concentration was tested in the citrus processing industry [72].

The purpose of concentration operations is to convert single-strength juice into a product form that will be more economical to store, distribute, and sell. Concentration by *evaporation* involves enzyme stabilization, pasteurization, and deaeration in addition to concentration. Traditionally the former operations were performed separately from evaporation. Now all three occur simultaneously during evaporation.

The evaporation process is critical for two reasons:

1) *Economics* – the fuel costs necessary for evaporation represent by far the largest operating cost of all the juice processing within the concentrate plant, excluding the extractor rental fee.
2) *Technology* – evaporation is a fairly complex process calling for considerable technical skill and experience. It subjects the juice to high temperatures. As a result evaporation greatly influences the quality of concentrate it produces.

There are three major types of evaporators each associated with a specific, largely specialized manufacturer.

TASTE Evaporator. The TASTE (thermally accelerated short time evaporator) is produced by Gulf Machinery of Clearwater, Florida (USA), and is the most common evaporator used worldwide. It accounts for around 80 % of the world market for citrus evaporators. The TASTE evaporator was developed in the late 1950s and early 1960s.

A typical TASTE evaporator, 5-effect, 8-stage, is shown in Figure 9. The word "stage" indicates the sequence of product flow as it travels through the evaporator. The term "effect" indicates the sequence of vapor (or steam) flow through the evaporator. The roman numerals as in Figure 9 indicate the effect while the arabic numerals indicate the stage. Thus, the number of stages is merely the number of bodies in series in terms of juice flow. The number of effects represents the number of different pressure drops being used for concentration.

TASTE evaporators use long vertical tubes, with inside diameters of 22 – 46 mm and lengths of 3 – 18 m. Total system heights are 18 – 35 m. Juice enters the evaporator by a countercurrent flow through interstage heaters which preheat the juice using vapor from the previous effect. The juice then enters the top of the first effect tube bundle through a nozzle type distributor, and travels downward through the tubes, where it absorbs heat from the steam on the outside of the tubes. At some stages, the juice is reheated in the reheater to provide sufficient superheat so that the liquid will flash and distribute itself uniformly upon entering the subsequent stage. Evaporation takes place on passage down the tubes, not as a liquid on the tube wall, but as a mixture of juice and vapor; heat is transferred to the turbulent mixture. The exit velocity from the tubes is ca. 20 – 100 m/s [73].

On leaving the tubes, the liquid and vapor mixture travels into the centrifugal-type vapor separator connected to the bottom of each effect. There is no juice level in the separator. The juice from the first effect separator is pumped to the top of the second effect where it enters through another nozzle, flashes, and evaporates in a similar manner. The vapor from the first effect supplies the heat for the second effect and so on for subsequent effects [74], [75].

Most TASTE evaporators are forward flow and single pass (no recirculation). There are normally 4 – 7 effects with the last composed of 1 – 4 stages. They are capable of concentrating the juice from 12 to 72 °Brix. For orange juice, the concentrate from the last stage is ca. 60 – 65 °Brix, and is at a temperature of 35 – 38 °C. The residence time from the first stage to the last stage is ca. 6 – 8 min. From the last stage, the concentrate is pumped into a flash chamber maintained at a high vacuum to chill the juice to 4 – 10 °C before exit. Another 1 – 3 °Brix increase in concentration results from this method of cooling. Practically instant cooling is achieved in the flash chamber. No heat transfer surfaces are required. This method of cooling minimizes heat damage which might result from high-temperature evaporation and can be used with viscous materials. The concentrate is further cooled to – 5 to – 10 °C prior to pumping into bulk storage tanks.

The evaporation to steam ratio is roughly $0.8\,N : 1$ where N is the number of effects, i.e.,

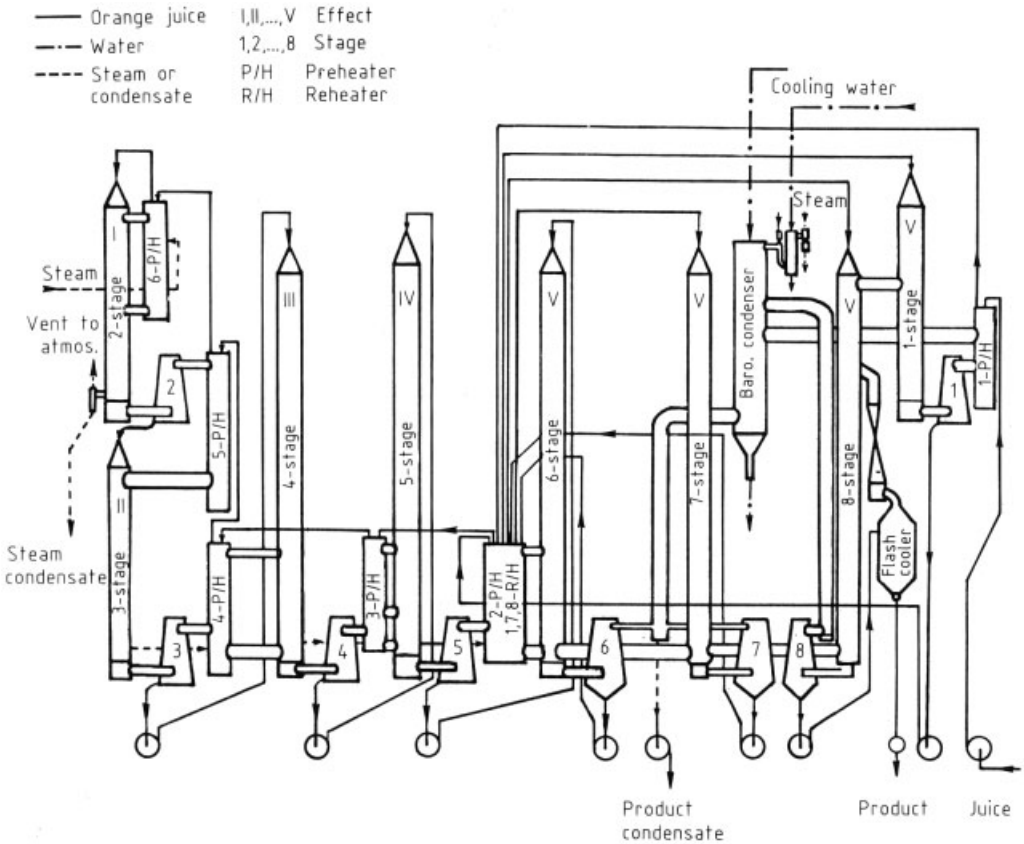

Figure 9. Schematic of typical TASTE evaporator (5 effect – 8 stage)
∗ Courtesy Gulf Machinery Co.

for a 4-effect evaporator the evaporation to steam ratio is 3.2 kg evaporation to 1 kg steam.

Plate Evaporator, APV. This type of evaporator is produced by APV with headquarters in the United Kingdom and manufacturing operations around the world.

The APV company developed their first plate evaporator in 1957. These evaporators are widely used in many countries in the dairy, food, fruit juice, pharmaceutical, and chemical industries.

The origins and basic concepts of the plate evaporator are very similar to those of the tubular, i.e., TASTE type evaporator just described. The major difference is that banks of gasketed plates held together in a frame replace the tube bundles and tubular bodies. Steam and juice pass between alternate plates.

The evaporation to steam ratio is about the same as the TASTE evaporator or slightly better. Higher efficiency vapor recompression is also available.

Centrifugal Evaporator, Alfa-Laval. This type of evaporator, produced by Alfa-Laval with headquarters in Sweden and worldwide manufacturing operations, differs significantly from the previous two. The main variations of the Alfa-Laval evaporator are: evaporation and centrifugal separation occur simultaneously (and continuously); residence times are short (1 – 2 s compared to several min) and temperatures low (45 °C compared to 80 °C); technology and economics are comparable to single-effect evaporation only.

Centritherm type evaporators were specially developed as single-effect, short time, low-temperature units to be used where conven-

tional evaporation methods did not give good results. The evaporation to steam ratio is no better than for a single-effect evaporator, i.e., about 0.9 : 1. Because of low steam economy, these evaporators have limited application in the citrus industry.

For further information on concentration and citrus evaporator technology see [73], [76].

The final operations after evaporation involve blending, packaging, storing, and shipping of the concentrated juice product. Concentrate to be used for reprocessing can be stored in drums at about − 20 °C. Tank or bulk storage is used for large operations at storage temperatures of − 5 to − 10 °C.

3.2.3. Processed Citrus Products

Citrus juice production can be described by the form in which the extracted juice is initially produced and stored. Market volumes can be measured by the retail form of the juice. Reprocessing activities which link the two could include reconstitution (i.e., dilution from concentrate to single strength) and/or consumer packaging (i.e., repackaging from bulk to consumer sized containers). Processing characteristics for different types of juice have been described [77].

The three major types of juice are:

1) Nonreconstituted single strength, i.e., produced direct from fresh fruit without evaporation;
2) Reconstituted single strength, i.e., produced from concentrate;
3) Concentrate, i.e., juice that has passed through evaporation.

There are three variable aspects of consumer packaging. These are:

1) Process used for filling – hot pack, refrigerated (chilled), or aseptic;
2) Type of container – can, glass, cardboard, or plastic;
3) Size of container – 165 mL through 19 L.

The estimated citrus juice production patterns and market volume are shown in Table 4. Citrus juice production patterns are illustrated in Figure 10.

Table 4. Estimated citrus juice production patterns and market volumes [78–80]

Patterns	Market volume, %∗			
	United States		World	
	Orange	Grapefruit	Orange	Grapefruit
Nonreconstituted S/S ∗∗	19	59	14	55
Reconstituted S/S ∗∗	14.5	20	29	24
Concentrate	66.5	21	57	21

∗ Expressed as % of all juice
∗∗ Single strength

Several factors affect the quality and stability of the final processed products. The two most important factors are dissolved oxygen and heat treatments. Oxygen must be removed because of its adverse effect on color, vitamin C, and citrus oils. Oxygen can be removed by deaeration; proper packaging design prevents it from re-entering the product.

Citrus juices are perishable at room temperature and can deteriorate in quality even in aseptic packages. Their delicate fresh aroma and flavor may be lost or damaged at room temperature. The extracted juices are usually pasteurized and chilled as rapidly as possible to minimize the loss of flavor.

The purpose of pasteurization is to destroy spoilage organisms, inactivate enzymes, or both. Heating to 65 °C will destroy most spoilage organisms, but temperatures as high as 92 °C for about 30 s may be required to inactivate the natural enzyme, pectinesterase, and stabilize the juice. The time required can be lowered by using higher temperatures. The temperature necessary to stabilize the juice is also pH dependent. Juices at higher pH require higher temperatures for stabilization. Pasteurization requirements for different products have been reviewed [81].

Concentrate Blending and Reprocessing. The purpose of concentrate blending is to achieve a uniform, high-quality final product prior to cooling, storage, distribution, and sale.

A process known as *cut-back* involves addition of fresh juice to the concentrate. This process was the traditional method of restoring volatile aromas and essences to the concentrate and is still used by some processors in Florida. However, with the development of essence recovery systems, the use of essence and oil in-

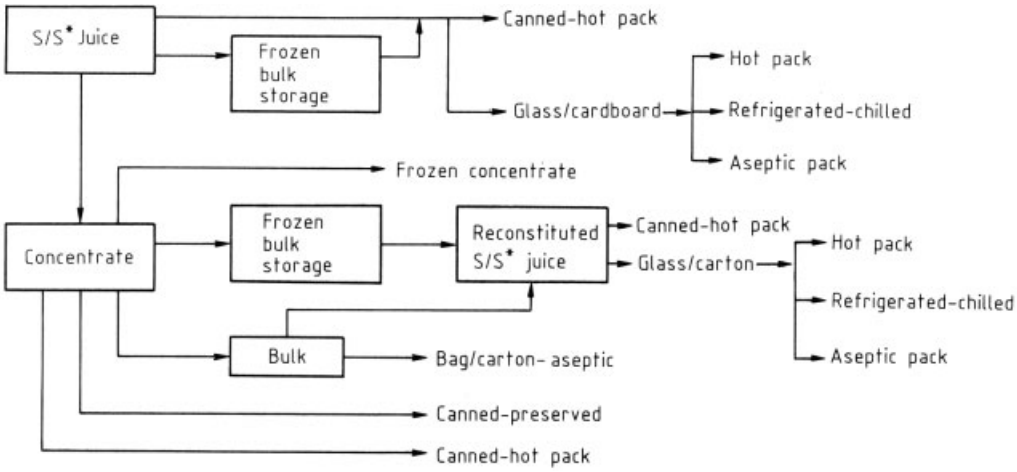

Figure 10. Citrus juice production patterns
∗ Single strength

stead of cut-back allow the concentrate to be maintained at high °Brix, reducing distribution costs. For this reason, virtually no plants use cut-back in Brazil.

Essences are stored as two separate phases: water phase essence and oil phase essence. These are generally added back to concentrate during blending along with cold-pressed oil, obtained from peel rather than juice. However, various combinations of essence and cold-pressed oil can be added back depending on the type of concentrate, time of year, customer requirements, etc. Blending is probably the most important operation for concentrate reprocessing. Technical aspects of reconstitution and blending of concentrate for consumption as a single strength juice have been described [82].

3.2.4. Chemical and Nutritive Data

Citrus juice is a complex mixture of hundreds of chemical constituents. Many of these relate to the quality and nutritive value of these products [83].

Citrus fruits have long been recognized as excellent sources of ascorbic acid (vitamin C), which is the most abundant vitamin in citrus fruits. Citrus fruits are also quite rich in the mineral element, potassium. Other nutrients in orange juice that are of dietary significance, according to standards set by the U.S. Food and Drug Administration [84], i.e., they are present at a level of 10 % or more of the U.S. RDA (Recommended Daily Allowance) per serving, are folic acid and thiamine (vitamin B_1). A serving size for orange juice is generally 177 mL. Average nutrient delivery per serving of reconstituted FCOJ (12.8 °Brix) in relation to U.S. RDA is shown in Table 5 [84], [85]. For more information on nutritive data, see [86], [69], [87].

Table 5. Average nutrient delivery per serving of reconstituted FCOJ (12.8°Brix) in relation to United States RDA

Nutrient	U.S. RDA [84]	Average % U.S. RDA/ 177 mL FCOJ [85]
Vitamin A	5000 IU∗	1.4
Vitamin C	60 mg	131
Thiamine	1.5 mg	9.8
Riboflavin	1.7 mg	2.4
Niacin	20 mg	2.0
Calcium	1.0 g	1.8
Iron	18 mg	1.1
Vitamin B_6	2 mg	4.9
Folic acid	0.4 mg	20.3
Phosphorus	1.0 g	3.3
Magnesium	400 mg	4.9
Zinc	15 mg	0.7
Copper	2 mg	4.4
Pantothenic acid	10 mg	3.3
Vitamin D	400 IU∗	–
Vitamin E	30 IU∗	–
Vitamin B_{12}	6 µg	–
Iodine	150 µg	–
Biotin	0.3 mg	–

∗ International Units

3.3. Tropical Fruits

There is no fine dividing line between a tropical and subtropical fruit, but a tropical fruit is generally defined as one which cannot be grown at temperatures which fall much below 5 °C [88]. The vast majority of tropical fruits thrive best in a region lying between the Tropic of Cancer and the Tropic of Capricorn; examples of tropical fruits would include banana and plantain, pineapple, mango, guava, Annona (soursop, cherimoya, custard apple, sugar apple), sapodilla, and papaya. Although many subtropical fruits, for example citrus, passion fruit, and longan grow also in tropical regions, they are classified as subtropical because they can tolerate temperatures around − 4 °C for the short term.

The entire topic of tropical fruit juices cannot be covered in this article. Therefore, the pineapple has been selected to illustrate the various steps in the processing of tropical fruit into beverages.

3.3.1. Introduction

The pineapple, *Ananas comosus* (L.) Merrill, is a native of tropical America, and appears to have originated in Brazil and/or Paraguay. *Ananas* is derived from the name used by the Guarani Indians of Paraguay [89]. Columbus found it cultivated in the West Indies in 1493, and by 1600, Portuguese sailors had carried it to Java, Malaysia, and India.

The pineapple is a multiple fruit composed of a large numer of fruitlets growing from a central stem called the core. Each fruitlet is a portion of a small flower which forms off the stem before the fruit matures. All pineapple cultivars examined produced functional pollen and egg cells but all are self-incompatible; therefore, they cannot set seed from self-pollination. The five groups of cultivated pineapples are Spanish, Queen, Abacaxi, Cayenne, and Maipure. The pineapple is best suited to a mild tropical climate with temperatures between 16 and 32 °C.

World production of pineapples in 1982 exceeded 5×10^6 t. The major producers with production in 10^3 t were [90]: Thailand (1824), Philippines (890), United States (560), Mexico (550), Ivory Coast (320), South Africa (232), Malaysia (170), Australia (125), and Taiwan (99). Total production of pineapple products ($\times10^3$ t) were: canned pineapple (804), pineapple juice concentrate (38), single-strength pineapple juice (195), and fruit salad, frozen pineapple, and other products (6).

Pineapples produced commercially in the United States are grown almost entirely in Hawaii (Smooth Cayenne variety) and Puerto Rico (Red Spanish and Smooth Cayenne varieties). United States pineapple products are destined primarily for the domestic market, but some are exported to Canada, Europe, and the Middle East.

3.3.2. Quality of Fruit for Processing

Nutrition, climate, variety, geographical growing area, season, and ripeness are important factors affecting pineapple fruit quality. The ultimate quality of pineapple juice is primarily dependent upon the relative ripeness of the fruit. Selection of individual fruit for harvest and processing requires considerable human judgment because fruit in the field does not ripen simultaneously.

Additionally, pineapples do not ripen after harvest; therefore, fruits harvested prior to optimum ripeness remain inferior and those harvested at an over-mature stage exhibit a poor flavor. Fruits destined for the fresh market are generally harvested at a less ripe stage than those going to the cannery.

Flavor and composition of pineapple juice mimic the flavor and composition of the fresh fruit from which it is prepared [91]. Pineapple fruits which have a °Brix/acid ratio of about 20 produce high-quality juice. In Hawaii the pineapple reaches peak fruit quality in June and July; during winter the fruit tends to be pale, lower in °Brix, higher in acid, and lacking in volatile flavor components.

3.3.3. Processing

Pineapple juice may be regarded as a byproduct resulting from the preparation of primary pineapple products such as slices, chunks, and crushed meat. Juice is primarily derived from cores, juice trimmings, eradicator meat, small

fruit, juice drained from crushed pineapple, and juice drained from eradicator meat. The FDA Standard of Identity [92] for pineapple juice is defined as the "juice intended for direct consumption, obtained by mechanical process, which may include centrifuging but not filtering, from the flesh or parts thereof, with or without core material, of sound, ripe pineapple. Canned pineapple juice contains finely divided insoluble solids, but it does not contain pieces of shell, seeds, or other coarse or hard substances." Added sweeteners, vitamin C, and dimethyl polysiloxane (defoaming agent) may be added to canned pineapple juice.

A simplified schematic showing the processing steps in the preparation of commercial pineapple juice is shown in Figure 11. Harvested fruit is unloaded from bulk bins at the cannery, mechanically graded for size, washed, and sent by conveyors to Ginaca machines (FMC, Madera, Calif.) (Fig. 12). These machines are adjusted to accommodate fruit of varying diameter sizes; that is, specific machines will process only fruit within a specified diameter range. The Ginaca machine is fully automatic and removes the inedible component of the fruit from the edible parts; it processes the fruit by cutting off both ends, coring the cylinder, and trimming the cylinder from the shell. One part of the machine, called an "eradicator," shaves the juicy meat portion (eradicator meat) from the shell for use in crushed pineapple or juice. The meat adjacent to the shell contains a large amount of valuable aromatics used in making a quality juice [93]. The core is used for the preparation of crushed pineapple or juice. The major product from the Ginaca is the pineapple cylinder, which is conveyed to trimming tables where each cylinder is hand-trimmed to remove the last traces of shells and blemishes. The cylinders are subsequently washed, sliced, and packed in cans. Slices that are broken, too thick, or too thin are unsuitable for the sliced product, but can be used for crushed pineapple or juice products. The fruit production from a single Ginaca machine is up to 150 t/d; about 75000 L of juice are produced [91].

The next primary stage in the preparation of the pineapple juice beverage is the disintegration of all solid constituents remaining after preparation of the basic pineapple products (slices, chunks, and crushed meat). Prior to disintegration, a first stage screw press may be used to remove part of the extractable juice; the screw press reduces the amount of fine solids that would have to be removed later to produce an acceptable juice. The disintegrating device commonly used is the Reitz disintegrator; this machine is comprised of a rotating shaft with fixed or pivoted disintegrating arms to which the fruit is introduced axially. The arms are surrounded by a screen (0.64 cm opening) through which material passes after having been appropriately reduced in size. The material emerging from the disintegrator is a free-flowing juice containing small sharp free-draining fibers [94]. Although many pine-apple canneries use the Reitz disintegrator, some prefer a Schwarz extractor which produces a liquid that goes directly to a juice tank and a coarse pulp that is then fed to a press.

The slurry material emerging from the disintegrator passes through a heat exchanger (about 60 °C) which increases cell wall breakdown, releases more juice, and inactivates the protein-digesting enzyme, bromelain. The heated juice then goes into a finisher (generally a screw type), which separates the liquid from the comminuted solids. The separated liquid flows into centrifuges (Sharples, Merco, or Westphalia continuous types) to remove excess fiber and other material.

Juice separated from insoluble tissue components by screw extractor and centrifuge machines is fed into a central holding or blending tank. This tank allows for the mixing of constituents of several lots of fruit. Additionally, liquid discharges from various process lines (crushed meat, slices, cores) may be collected separately and characterized for taste, solids content, °Brix/acid ratio, flavor quality, etc. The liquid discharge from each line may then be judiciously blended to create a juice product with characteristics best suited for that product, i.e., having the desired levels of sugar, acid, insoluble solids, flavor, and aroma.

In the preparation of single-strength juice, cans are filled at about 60 °C with the juice from the blending tank and closed. The juice is pasteurized by passing the closed cans through a standard reel cooker and cooler combination; processing time is regulated to bring the temperature of the juice to no less than 87.7 °C [91]. Pineapple concentrate is produced from single-strength juice with equipment similar to

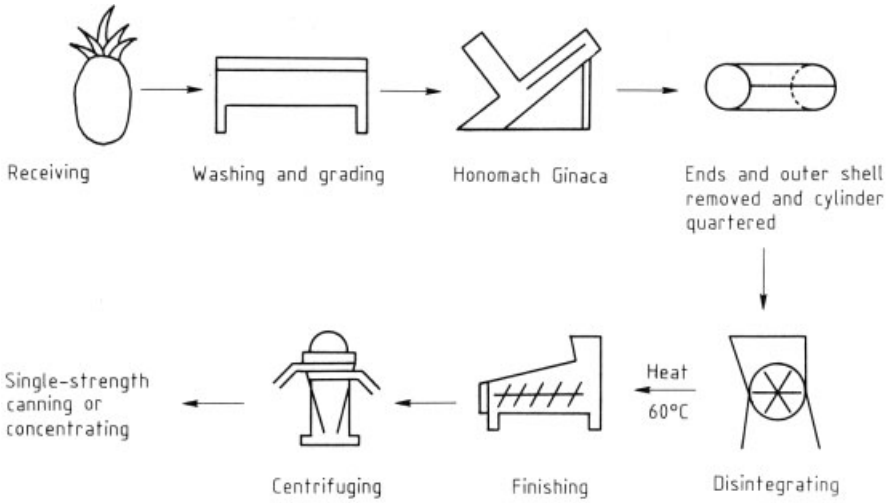

Figure 11. General processing steps in the preparation of pineapple juice

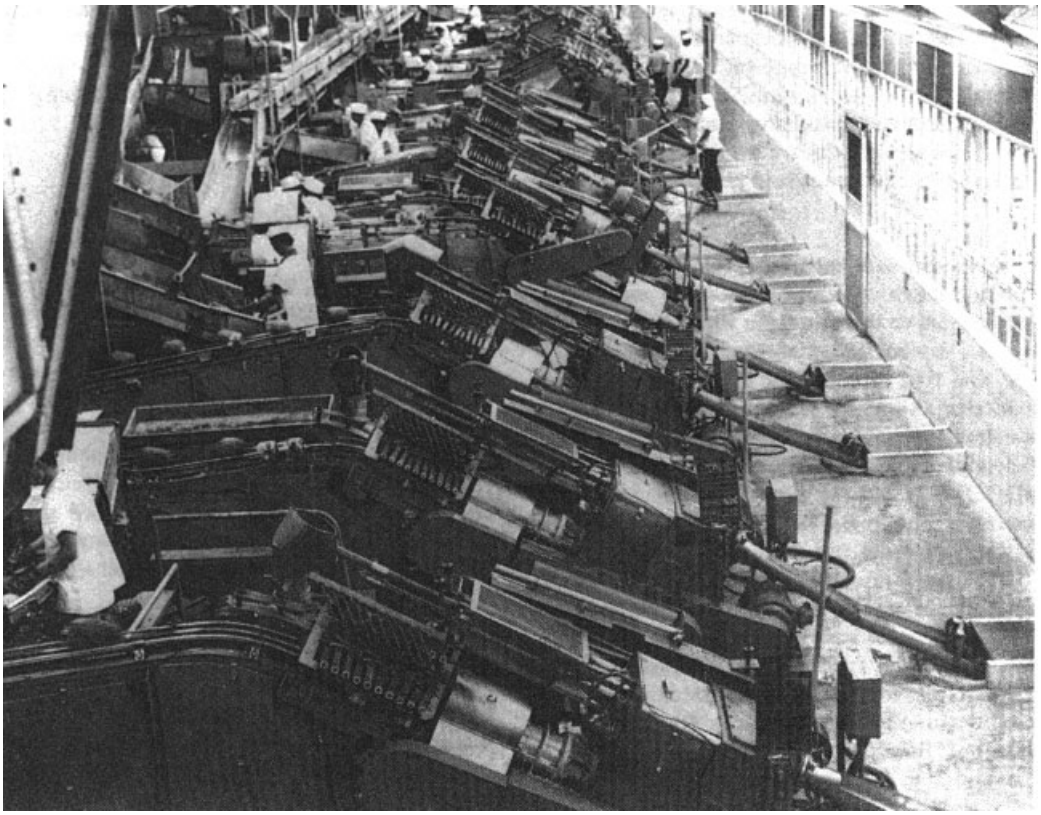

Figure 12. Ginaca machines
∗ Courtesy AVI Publishing Co.

that used in the preparation of citrus concentrate (see Section 3.2.2.).

Types of Products. The major pineapple juice products are canned single-strength, concentrate, and nectar. The standard of quality for canned single-strength pineapple juice [92] is: soluble solids content without added water shall not be less than 10.5 °Brix; the acidity is not more than 1.35 g of anhydrous citric acid per 100 mL juice; °Brix/acid ratio is not less than 12; and finely divided insoluble solids is not less than 5 % nor more than 30 %. Pineapple concentrates are produced at °Brix values of 46 ° (3 to 1 for reconstitute), 61 ° (4.5 to 1), and 72 ° (6 to 1). Mixed nectar products containing pineapple pulp as one of the ingredients are available, but a pure pineapple nectar is not sold [95].

Table 6. Mineral and vitamin composition of frozen pineapple juice [96]

	Proximate composition		
	Maximum	Minimum	Average
Solids, %	15.1	13.4	13.8
Ash, %	0.39	0.31	0.35
Ether extract, %	0.04	0.03	0.03
Protein, %	0.50	0.36	0.41
Crude fiber, %	0.19	0.08	0.11
Total carbohydrate, %	14.3	12.6	13.0
Energy in 100 gm, kJ	222	197	205
Minerals in 100 gm, mg			
Calcium	12.8	9.0	10.8
Total iron	0.52	0.24	0.32
Phosphorus	9.6	7.6	8.3
Magnesium	10.7	4.6	8.9
Potassium	170	135	143
Sodium	1.0	0.6	0.8
Vitamins in 100 gm, mg			
Ascorbic acid	16.6	11.0	13.0
β-Carotene	0.011	0.006	0.009
Folic acid	0.002	0.001	0.001
Niacin	0.29	0.20	0.25
Pantothenic acid	0.157	0.094	0.125
Riboflavin	0.019	0.012	0.016
Thiamine	0.069	0.058	0.066
Vitamin B_6	0.098	0.053	0.074

Chemical and Nutritive Data. The composition and nutritive value of pineapple juice is not only related to the fruit from which it is derived, but also is related to processing conditions, storage, and type of container in which the juice is marketed. An approximate composition of frozen pineapple juice is shown in Table 6 [96]. Pineapple juice is well buffered (range of about pH 3.3 – 3.7) and the average free acid content averages about 0.8 %. The major acids in descending order are citric (0.7 – 0.9 %), malic (0.2 %), and ascorbic (0.008 – 0.02 %); small amounts of oxalic, malonic, succinic, galacturonic, tartaric, and glycolic are also present [97]. The soluble solids content of pineapple juice is mainly derived from carbohydrates, in particular, the mono and disaccharides. Representative values of °Brix (10.8 – 17.5), glucose (1.0 – 3.2 %), fructose (0.6 – 2.3 %), and sucrose (5.9 – 12.0 %) have been reported [98]. Pineapple juice is an attractive food because of its good taste and pleasing aroma. Numerous studies on the aroma constituents and other aspects are available [91], [98], [99].

3.4. Vegetable Juices

The processing of vegetables into juice products has many features in common with fruit juice processing. Juice is derived from a fresh vegetable by comminution, followed by separation of the large colloidal particles, or by pressing the comminuted vegetable to obtain a clear juice. Comminuting machines express about 60 –90 % of the vegetable into a product of juice-like consistency. Quite often, the suspended solids are allowed to remain in the juice because much of the flavor, aroma, and mouth-feel is due to these solid constituents.

The maceration of a vegetable liberates enzymes. Unless heat treatment is used to deactivate them, either before or immediately after the cellular disruption process, detrimental juice quality changes will result. The processing of low-acid vegetables (tomato, rhubarb) or the addition of acid (commonly HCl) to a vegetable will allow for lower processing temperatures. However, for most vegetables, high-temperature heat processes are required to inactivate enzymes and to kill spores of spore-forming bacteria that may germinate in neutral or slightly acidic vegetable products [100].

Tomato juices represent the largest and best known segment of the vegetable juice market. Tomatoes are used in making tomato juice, cocktails, and tomato juice blends. Because of the prominence of tomato juice, a brief discussion of the processes involved in its preparation follows. Tomatoes used for preparation of tomato

juice possess high color, rich flavor, and high total acidity. These *quality attributes* are primarily influenced by variety, maturity, climate, geographical growing area, and cultural practice. In the *juicing process,* tomatoes are first chopped into fine pieces or coarsely ground, and then extracted by a pressing action, or they are sieved under great agitation as is normally encountered with a pulper or finisher. The pressing action of a screw-type juice extractor yields an average of about 79 % juice, whereas a paddle pulper and a paddle finisher give an average yield of about 82 % [101]. Either type of extractor can be preset to yield either a high or low percentage of juice.

Consistency is an important quality factor; tomato juice with proper consistency has a smooth mouth-feel and its solids will not separate readily from the serum [102], [103]. Tomato juice consistency can be increased by addition of acid to the tomato tissue immediately prior to or during comminution and heating [104], [105]. The method of manufacture, i.e., by use of the cold break method or the hot break method, also affects consistency of the tomato juice product. In the *cold break method,* tomatoes are macerated at ambient temperature and the resulting pulp is pressed through wire mesh screens to separate juice from seeds, skins, and cores. This process yields a juice of low consistency because of pectin degradation by naturally occurring pectic enzymes. In the *hot break method,* the same mechanical steps are used but heat (71 – 88 °C) is applied during maceration or immediately after to inactivate the pectic enzymes. The juice obtained by the hot break method has a higher consistency and is generally preferred by the industry.

During the maceration, finishing, and filling steps, deaerators are often used to remove dissolved or occluded air. Deaeration minimizes vitamin C loss and off-flavor development. In addition deaeration may be used after pasteurization or high-temperature presterilization to remove air incorporated by pumping, to cool the juice to filling temperatures, and to prevent foaming in the filler bowl.

Tomato juice may also be homogenized to prevent settling of solids and separation of serum. In this process, juice is routinely forced through narrow orifices at pressures between 70 and 98 kg/cm^2 to finely break up the suspended solids. Viscosity of juice is increased and a smoother consistency is attained.

Prior to packaging, juice is sterilized at a recommended temperature of 121 °C for 0.7 min and cooled rapidly to about 93 °C. The juice is conveyed to filler bowls where salt is ordinarily added or the salt may be added to individual cans (0.5 – 1.25 wt %). Finally the juice is packed in plain or enameled tin cans or in glass.

A more extensive coverage of other vegetable juice products, such as fermented vegetable juices, nonacid vegetable juices, and soybean milk, may be found elsewhere [4], [100].

4. Sweet Carbonated and Non-carbonated Beverages [106–108]

The most important requirements for sweet nonalcoholic beverages (soft drinks) are that they must be refreshing as well as pleasing in taste and odor. The taste factors perceived by the tongue are sweet, sour, bitter, and salty. These factors are influenced by beverage ingredients, such as sugar (sweet), phosphoric and citric acids (sour), quinine and caffeine (bitter), and various salts.

The odor perceived by the nose is determined predominantly by the flavoring ingredients, as will be discussed in Section 4.2, and to a lesser extent, also by fruit juices, caramel, etc.

Other food additives play an important role in soft drinks as well, and according to their functional properties include preservatives, antioxidants, emulsifiers, food colors, sequestrants, stabilizers, and buffers.

The soft drink industry is based on two different technologies, namely for carbonated products and for still drinks. This industry is the major user for most of the ingredients mentioned previously.

History. The oldest reference to soft drinks is probably in the *Old Testament* where unfermented "wines" are mentioned. Lemon juice is also one of the oldest bases for soft drinks, such as the "lemonata" from Sicily. Still lemonade was well known in France and Italy at the end of the 16th century, and in Britain by the end of the 17th. One of the first legal regulations on record for such products is from Paris in 1676 regarding the sale of "limonade". The history of sweetened carbonated beverages is not quite clear;

it started around 1800 after the skills involved in making artificial mineral water were developed. LEONHARD THURNEYSSER is usually credited with the first attempts to make carbonated water in Switzerland in the 16th century. However, the potential of artificial carbonation was not realized until experiments, such as JOSEPH PRIESTLEY's in 1767, brought it to public notice. In 1781, the world's first artificially produced mineral water was sold by THOMAS HENRY in Manchester, but on a very small scale. This may, in fact, have been the first carbonated soft drink, because he recommended to take it with lemon juice.

Commercialization on a somewhat larger scale was realized in Geneva in 1783, by PAUL and GOSSE, financed by SCHWEPPE. This was soon followed by JOSEPH HAWKINS, who started a new business in Philadelphia in 1808, producing "soda water" from artificial mineral water, sugar, and fruit juices. The next step took place in 1820 in Dresden, where F. A. A. STRUVE began the manufacture of soft drinks based on artificial mineral water and introduced the use of dyes and fruit acids. In 1839 J. W. WRIGHT produced a carbonated drink in Hamburg containing citric acid and lemon oil. The first Cola drink was produced in Atlanta by J. S. PEMBERTON in 1886, who experimented with a tonic wine in order to remove the alcohol and make it taste better. The result was Coca-Cola. He was soon followed by druggist C. BRADHAM who compounded Pepsi-Cola in 1893, based on extracts from kola nuts and vanilla, together with essential oils. It took until after World War II to gain worldwide acceptance of soft drinks, and they now constitute a significant part of our daily fluid intake. The cola beverages have developed into truly international drinks, and can be found virtually anywhere in the world. High-quality requirements and strict regulations have made possible reaching the mass production and public confidence that is today's standard.

4.1. Types of Beverages

Standards and regulations concerning soft drinks differ significantly from country to country. *Standards* specify those ingredients that must be present, usually at a minimum level, in order to be classified in certain categories, such as fruit juice drinks, sugar content, etc. *Regulations* are more concerned with additives and contaminants, and prescribe maximum levels. Many countries, e.g., United States, European Economic Community (EEC), use a "positive" list for those additives that may be used and none other. Other countries use a "negative" list wherein only those materials are listed that may not be used. Attempts are being made to simplify the complex system of food laws, e.g., Codex Alimentarius, and "harmonization" in the EEC. However, as long as there is no agreement on what constitutes "safe", it is unlikely that these differences in regulatory philosophy will disappear soon.

4.1.1. Still Drinks

Non-carbonated beverages were the first in recorded history, and they continue to represent the majority of drinks consumed today. In the United States about 25 % of the non-alcoholic beverages are carbonated, and in most other countries this is considerably lower (see Chap. 5). The non-carbonated drinks include water, coffee, tea, dairy beverages, and fruit juice products. Most non-carbonated beverages are produced in a ready-to-drink form, but in some markets, such as the EEC, fruit syrups are popular.

Production technology of non-carbonated drinks concerns mainly beverage stability, i.e., maintaining appearance and preventing microbiological contamination. Processing may include homogenizing dairy products or centrifuging fruit juices, but virtually always some form of thermal processing is required. The classical processes were retorting dairy products, and pasteurizing (tunnel or flash) juice products. Under these processing conditions, the containers could only be cans or bottles.

Aseptic packaging, originally developed for milk products in 1952 but also applied in the 1970s to fruit juices, has led to a tremendous number of better products and different containers, such as paper laminates. In this process, the product is pasteurized at high temperature for a short time (HTST), and aseptically transferred to a container that has been chemically sterilized with hydrogen peroxide. The worldwide growth in this type of product is expected to be between

8 – 10 %/a, mainly based on economic advantage. However, these products also often have a quality advantage because the thermal processing in this case does not affect flavor and color as much as retorting or hot packing does [109].

Most still drinks are *fruit flavored* and, in fact, contain fruit juice. The juice content varies from 100 % for pure juices, to 30 – 50 % for nectars, and to 10 % for fruit juice drinks. Other ingredients are sweeteners, acidulants (mainly citric acid), food color, flavoring, and ascorbic acid for vitamin C enrichment or as an antioxidant. *Nonfruit drinks* include protein-based beverages, such as drinking yoghurt, flavored milk, and soy beverages. Finally, there are coffee- and tea-based beverages, predominantly found in the Far East.

4.1.2. Carbonated Beverages

Standards for carbonated beverages differ significantly from country to country, but in many instances at least three categories can be identified:

1) *Fruit Juice Beverages.* This type of soft drink generally derives most of the flavor and appearance from fruit juice. The minimum juice requirement for this classification is regulated, and depends on the country and the type of fruit. For orange drinks it is, in most cases, at least 10 %; for lemon, typically about 5 %. These products usually can only be flavored with natural essential oils or extracts (see Section 4.2). Use of coloring is often prohibited or limited to natural products. In many countries use of preservatives is also prohibited in this category, and therefore tunnel pasteurization is usually applied to the bottled or canned beverage.
2) *Beverages with Natural Flavor.* This category can be separated into groups of *cloudy beverages* and *clear beverages.* In Europe the names "limonade" and "soda" are used, but are not always defined exactly the same from country to country. On a worldwide basis, this group represents the largest volume of carbonated beverages because it includes colas and lemon-lime beverages. The *colas* are the only truly international beverages because the major brands are marketed with virtually the same quality in about 150 countries; minor variations are the result of regulatory differences. Several countries have standards for cola drinks, requiring the presence of extracts (e.g., from cola nut), natural flavor, caramel color, and exclusively allowing caffeine and phosphoric acid. Tonic is also in the clear category, with quinine or other bitter substances added according to local regulations. The *cloudy group* contains orange flavored beverages, which in many countries is the second or third largest flavor category. The cloudy appearance may be due to relatively low levels of juice or to emulsions based on orange oil and emulsifiers as permitted by local regulations.
3) *Beverages with Artificial Flavor.* This is a catch-all category where generally more additives are allowed than in the above two groups. This is often the group where *diet beverages*, based on intense sweeteners, are found. The situation in the United States is somewhat different because the soft drink standards cover only carbonated beverages based on nutritive sweeteners and, therefore, diet beverages are nonstandardized, thus requiring declaration of ingredients and other information.

4.2. Flavoring Materials and Concentrates

Flavor is one of the main properties of beverages and is often the distinguishing characteristic between brands. The industry uses natural as well as synthetic flavoring materials and these are usually incorporated in some convenient form of "concentrate" which makes it possible to add flavoring to the beverage in a controlled fashion.

4.2.1. Natural Flavoring Materials

Aromatic vegetable materials, such as fruits, flowers, roots, seeds, and bark, are the most common resource for natural flavors. In beverages the most popular flavors are derived from essential oils. Odoriferous oils have been known throughout history, but only early in the sixteenth century was distillation used to isolate

such oils. In fact, in the search for the "Quinta essentia" or the active principle of drugs, the volatile oil was discovered. Distillation did not yield the most effective part of the drug, but instead gave the volatile oil that represented the odor of the drug. The name *essential oil* was maintained, however.

Essential oils are produced by cold-pressing, distillation, or extraction. Citrus oils are the most popular in the beverage industry, and these are generally produced by cold-pressing or by scraping the peel, e.g., orange, lemon, and grapefruit. Lime oil is typically produced by steam distillation, but some cold-pressed is also done. The yield of citrus oils is about 3 – 7 kg/t. The citrus oils consist of > 90 % terpenes, empirical formula $C_{10}H_{16}$, and low levels of oxygenated compounds, such as octanal, citral, terpineol, etc. The terpenes contribute little flavor and are very slightly soluble in water. For these reasons, and especially in clear beverages, the terpenes are often partially or completely removed. The most popular method to remove them is distillation. In addition extraction is used and, to a small extent, chromatography.

Most other essential oils used in beverages are produced by steam distillation, e.g., ginger, cinnamon, anise, etc. Yield varies widely, such as 0.2 % for cinnamon and 17 % for clove.

Oils produced by extraction are not essential oils according to definition because they also contain nonvolatile ingredients. Examples are some of the floral oils, popular in fragrances, and used at low levels in some beverage flavors. Spice extracts, such as ginger root for ginger ale, are more popular, and in this case the extract contains not only the essential oil but also the pungent character of the root. Citrus extracts are also widely used, particularly for clear beverages. The citrus peel or the oil is extracted with alcohol and water in order to obtain the "soluble essence". In the case of peel extraction, this is usually followed by distillation to separate the terpenes, which are poorly soluble in the distillate [110–112].

Many other extracts are used for flavoring of beverages, and some of the best-known examples are vanilla bean extract and cola nut extract, used in cola beverages worldwide. Another example is the extract of guarana seed, which contains caffeine, and is used in Brazil. The most common solvents for such flavor extracts are ethanol and propylene glycol.

4.2.2. Synthetic Flavoring Materials

The beverage industry applies predominantly natural flavoring ingredients, but some synthetic materials are also used, alone or in combination with natural ones. Synthetic material is found mainly in noncitrus flavors, such as strawberry, cherry, etc.

Most of the synthetic flavoring compounds occur also in nature, and only a few are strictly unique. An example of a natural flavor is vanillin [*121-33-5*] (4-hydroxy-3-methoxybenzaldehyde), which occurs in the vanilla bean. Synthetic vanillin can be produced from lignin, a byproduct from paper manufacture, or from eugenol, a natural constituent of clove oil. An example of a manufactured or artificial flavoring compound is ethyl vanillin [*121-32-4*] (3-ethoxy-4-hydroxybenzaldehyde).

The largest volume flavoring materials are simple esters, such as ethyl butyrate and amyl acetate, as well as aldehydes, such as benzaldehyde and vanillin.

4.2.3. Flavoring Concentrates

Most flavoring ingredients, with or without other additives, are used in a conveniently diluted form in order to ensure homogeneous mixing with the syrup or beverage. In Europe, where fruit juice-based beverages are quite common, a one component "Grundstoff" based on concentrated fruit juices together with flavoring extracts or oils is often used. Preparation of the beverage only requires addition of sugar and water to such a concentrate. The juices are often specially treated in terms of pulp content, particle size, and pectin content in order to ensure a stable appearance of the beverage with the least amount of sediment. Such concentrates are usually homogenized at high pressure, particularly if essential oils are used, to stabilize the emulsion.

In the case of franchised beverages, the concentrates usually contain several components, some of them in dry form, but always containing all the ingredients except for water and sugar to

prepare a fixed amount of bottling syrup. Well-known examples are cola and lemon-lime beverages.

The most popular flavor concentrates, such as cola and orange, are in *emulsion* form because they are based on essential oils which need to be dispersed. The cola concentrates are to a large extent caramel-based emulsions [113], whereas most other flavors use gum acacia or modified starch, etc., as emulsifiers. In cloudy beverages without juice, other emulsifiers and stabilizers, e.g., glyceryl ester of wood rosin, brominated vegetable oil, sucrose acetate isobutyrate, are required in order to reduce and maintain the oil phase at a particle size of about 1 μm. With clear beverages, the concentrates are either a soluble essence, such as lemon-lime, or a solution of flavoring ingredients in suitable solvents, such as ethanol, propylene glycol, 2-propanol, etc.

Other important concentrate ingredients can be found in the list of additives that are allowed in soft drinks for each country. The most popular *acidulants* include orthophosphoric acid, used strictly for cola beverages, citric acid and malic acid for fruit drinks, and tartaric acid, used almost exclusively for grape beverages. Among *food colors,* the most important one for beverages is undoubtedly caramel, a polymer prepared by heating dextrose in the presence of ammonium sulfite. The most important artificial color is Sunset Yellow (C.I. 15985), followed by Tartrazine (C.I. 19140). The most controversial colors are found among the reds where regulatory opinions differ significantly. The well-known red colors are: Amaranth (C.I. 16185), Carmoisine (C.I. 14720), and Allura Red (C.I. 16035).

Preservatives are needed to retard growth of microorganisms. Benzoic acid is the most widely used one, followed by sorbic acid, and to a much lesser extent sulfites. *Alkaloids* are another important class of additives, and these are used to modify taste and to add refreshment. Caffeine is used in cola beverages at about 100 mg/L, and also in some other beverages. Quinine is only used in tonics. Additives also include antioxidants, such as ascorbic acid, buffers, such as sodium citrate, and several others for special applications [114].

4.3. Beverage Production

The technology involved in preparing beverages is mainly found in equipment engineering, such as high-speed filling lines, aseptic fillers, etc., and is beyond the scope of this review.

Most *non-carbonated beverages* are prepared by simple mixing of the ingredients followed by pasteurization either as is or within the container. Dairy products usually require more processing, such as homogenizing, and are often sterilized, in a retort for instance.

In the production of *carbonated drinks* five separate activities can be distinguished: 1) water treatment, 2) syrup preparation, 3) carbonating and filling, 4) wastewater treatment, 5) quality control.

4.3.1. Water Treatment [115]

The type and amount of impurities in water vary considerably and depend largely on the source. Water treatment removes impurities and makes it suitable for beverage use. Common *impurities* include:

1) turbidity and sediment due to suspended insoluble matter
2) color caused by organic matter
3) microbiological organisms
4) salts which raise the alkalinity and may affect taste
5) mineral acids
6) hydrogen sulfide.

Most water treatment systems (see Fig. 13) involve four basic *processes:* alkalinity reduction, chlorination, coagulation, and clarification. The *chemicals* required are: lime, chlorine, and ferrous sulfate. These are added in a reaction tank where most of the sludge settles. The clear overflow from this tank is filtered through sand or diatomite, and a carbon purifier removes chlorine and trihalomethanes. For water containing high-organic matter, ozone is used instead of chlorine, in order to prevent formation of additional trihalomethanes. In cases where dissolved solids are extremely high and sodium content must be reduced, three alternatives to chemical treatment are: reverse osmosis, electrodialysis, and ion exchange.

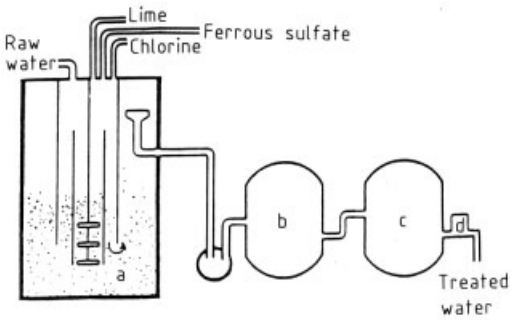

Figure 13. Water treatment plant
a) Reaction tank; b) Sand filter; c) Carbon purifier; d) Polisher∗ Courtesy PepsiCo, Inc.

4.3.2. Syrup Preparation

Sucrose [*57-50-1*] , mainly in granulated form, is the most common sweetener in carbonated beverages (→ Sugar). The beverage industry uses cane as well as beet sugar of high purity. Sucrose is also sold as syrup, the so-called liquid sugar, which is popular in the United States and Europe. Sugar syrup is also supplied partially *inverted*, i.e., the sucrose is converted to dextrose and fructose in order to improve stability and allow a higher solids content. Inversion can be accomplished by heating with mineral acid, by ion exchange, or by use of invertase. A relatively new sweetener is *high-fructose corn syrup* which is prepared by enzymatic conversion of corn starch. In this syrup, about 55 % of the solids is fructose, about 42 % dextrose, and the remainder higher saccharides. In the United States this sweetener has already replaced more than 50 % of the sucrose that is used in soft drinks.

When granulated sugar is used, the bottler prepares a syrup (melting), which usually requires at least filtration, and often treatment with activated carbon to remove color or odor. The syrup room contains various stainless steel mixing tanks where the syrup is prepared and held until it can be bottled. The sugar syrup is measured into a tank along with the flavor concentrate and other additives. After the required quality control steps, the bottling syrup is ready for use.

In *low-calorie drinks,* the nutritive sweetener is replaced by intense sweeteners, such as saccharin, cyclamate, or aspartame (→ Sweeteners). The use of these sweeteners is restricted through food regulations and differs from country to country. Blends of intense sweeteners are preferred because of synergism and also undesirable taste can be minimized. The sweeteners are used in a soluble form, such as sodium or calcium salts, and a "syrup" is prepared similar to a regular bottling syrup. Sugar sweetened drinks normally contain about 10 % sucrose, but in low-calorie drinks, the amount of sweetener is very low, e.g., 0.04 – 0.05 % for saccharin and aspartame, or 0.2 % for cyclamate.

4.3.3. Carbonation and Filling

The final stage of soft drink preparation consists of mixing bottling syrup and treated water in a certain ratio, carbonating under pressure, and filling into bottles or cans.

The traditional method was to chill syrup and water separately, then "throw" an accurately measured quantity of syrup into the containers, and fill it with water that had been passed through the carbonator. After closing the container, the mixture was agitated to blend syrup and water. This technique (post-mix system) is still applied in low-speed lines and in product with high-pulp content.

In high-speed lines, water and syrup are blended in the correct proportions, usually 5 : 1, either through an orifice (United States) or through metering pumps (Europe). The premix beverage is then cooled, carbonated, and filled. The carbonator is a pressurized vessel in which liquid and carbon dioxide are brought in intimate contact. Degree of carbonation is a function of temperature and pressure (solubility of carbon dioxide increases as the temperature of the liquid decreases). The solubility of carbon dioxide is adversely affected by the presence of dissolved air in water, and therefore deaerators are often employed, certainly in canning lines. Variations of the two basic carbonation methods, such as injection of liquid carbon dioxide, are also being used [116].

A typical bottling line layout is shown in Figure 14; however, this is a basic line, and not the most sophisticated, such as syrup batching by computer, continuous blending, etc. A high-speed line typically can fill 1200 – 1500 con-

tainers per minute depending on size and type of container. For further information see [116–120].

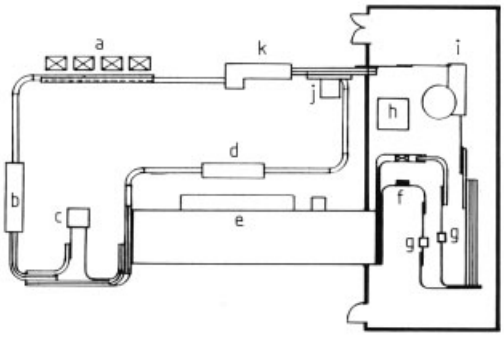

Figure 14. Typical bottling line layout
a) Pallet unloading, loading; b) Uncaser; c) Decapper; d) Case washer; e) Bottle washer; f) Visual inspection; g) Electronic bottle inspection; h) Carbo cooler; i) Filler; j) Accumulation table; k) Case packer∗ Courtesy PepsiCo, Inc.

4.3.4. Wastewater Treatment

In bottling operations with returnable bottles, as much as 60 % of the incoming water may go to waste; with proper reclamation, this can be reduced to about 30 %. In operations with nonreturnable containers, the wastewater load is considerably smaller, and is generally less than 10 % of incoming water.

Wastewater treatment consists at least of neutralization and biological purification. The latter method is widely used to reduce dissolved organic matter, such as sugar, through breakdown by bacteria. Two broad types of equipment are trickling filters and activated sludge [121].

4.3.5. Quality Control

Manufacturing of soft drinks begins and ends with quality control. First there is quality control on all concentrate ingredients, such as essential oils, and then on the concentrate itself. Secondly, the beverage ingredients, such as sugar and water, are checked on chemical and physical parameters and on microbiological condition. Finally, beverage samples are removed from the line periodically and checked to ascertain whether they are within specification. Tests may include carbonation, solids, titratable acidity, pH, color, and taste. In addition microbiological checks are performed. Electronic as well as visual inspections are made during filling on containers and fill height as well as on net content. In-line quality monitoring is also possible electronically, e.g., densitometer for solids, infrared absorption for carbon dioxide [118], [119].

5. Economic Aspects

From 1979 to 1983, nonalcoholic beverages recorded continuous growth in both consumption and monetary value in markets of the United States and Europe. The estimated retail volume of nonalcoholic beverages consumed in the United States over this 5-year period is noted in Table 7. In 1983 about 40.4×10^9 L of drinks and juices, and about 2.6×10^9 L of bottled waters were consumed (Table 7) [122]. The retail dollar value of the United States nonalcoholic beverage market in 1983 amounted to: soft drinks ($\$26.1\times10^9$), fruit juices and drinks ($\$7.8\times10^9$), drink mixes ($\$1.7\times10^9$), and bottled waters ($\$0.7\times10^9$).

The ten major companies and their market share of the soft drink, bottled water, and juice and fruit drink segments of the nonalcoholic beverage market are noted in Table 8 for the United States. The fastest growing areas of the soft drink industry are diet and caffeine-free beverages. Diet soft drinks account for about 18 % of the market and, in terms of flavors, cola leads with 63 % of the sales followed by lemon-lime with 14 %. Caffeine-free has become a major advertising issue, and some soft drink manufacturers enthusiastically proclaim that their products contain “no caffeine”.

5.1. Fruit Juices

The fruit juice market is depicted in Table 9. As noted orange juice is the overwhelming favorite of the American consumer with about 62 % market share. Frozen concentrated orange juice is the predominant product, but chilled orange juice has shown spectacular gains – consumption of 772×10^6 L in 1978 to over 1306×10^6 L in 1983 (represents sales of $\$542.5\times10^6$ to $\$1.1\times10^9$). Grapefruit juice is preferred after

Table 7. Estimated retail sales of single-strength nonalcoholic drinks and juices, $\times 10^6$ L [122]

Year	Packaged soft drinks (carbonated & noncarbonated)	Powdered soft drink mixes	Fruit flavored drinks	Juices (100 %)		Total all nonalcoholic drinks and juices
				Citrus juices	Other juices	
1 979	23 942	4 197	1 849	3 077	1 284	34 350
1 980	24 224	4 248	2 257	3 424	1 312	35 465
1 981	25 193	4 205	2 061	3 404	1 578	36 441
1 982	26 453	4 331	1 975	3 386	1 658	37 803
1 983	28 040	4 768	2 252	3 604	1 769	40 432

Table 8. The ten major companies in the United States nonalcoholic beverage market in 1983 [123]

Soft drinks		Juice and fruit drinks		Bottled water	
Company	Sales, $\$\times 10^9$	Company	Sales, $\$\times 10^9$	Company	Sales, $\$\times 10^9$
Coke USA	1.737	Coca-Cola Foods	1.995	Beatrice	0.127
Pepsi USA	1.164	General Foods	0.595	McKesson	0.120
Seven-Up	0.411	Del Monte	0.528	The Perrier Group	0.092
Dr. Pepper USA	0.345	Tropicana	0.383	Hinckley & Schmitt	0.028
Royal Crown Cola	0.291	Ocean Spray	0.367	Beverage Mgt.	0.025
Shasta	0.245	Tree Top	0.185	Kentwood	0.013
Canada Dry USA	0.181	Welch	0.183	Mountain Valley	0.012
Sunkist Soft Drinks	0.176	Borden	0.122	Nestle	0.011
P & G Beverage Div.	0.176	TreeSweet	0.108	Absopure Water	0.010
Schweppes USA	0.070	Gatorade	0.097	Black Mountain Spring	0.010

orange juice (7.2 % market share). It has not shown the spectacular growth demonstrated by orange juice because 70 % of consumers surveyed found the taste of grapefruit juice too tart for their liking.

Table 9. 1983 Fruit juice market for the United States [124]

Commodity	$\$\times 10^6$	% Market share
Orange juice	3842.3	62.1
Apple juice and cider	1094.0	17.7
Grapefruit juice	442.4	7.2
Grape juice	292.5	4.7
Pineapple juice	142.5	2.3
Prune juice	119.4	1.9
Blended fruit juice	94.0	1.5
Lemon and lime juice	92.8	1.5
Fruit nectars	65.9	1.1

The apple juice and cider segment (about 18 % market share) of the juice industry has been growing at a rate of about 20 % annually in recent years, with a jump of about 96 % in liters sold between 1978 and 1983, and a 169 % gain in dollars. The other fruit juices (grape, pineapple, blended, prune, lemon-lime, and nectar) collectively comprise a modest 11 % of the juice market. For the 1978 – 83 period, grape juice and pineapple juice showed retail gains of 13.5 and 11.9 %, respectively. Blended 100 % fruit juice ranks about seventh in the total fruit juice market but is increasing in market share; from 1978 – 1983 blended juice showed a 140 % volume increase and a 248 % dollar increase. Traditional blends are orange-grapefruit, pineapple-grapefruit, apple-grape, and applebery, but many other diverse blends are entering the market. The lemon-lime juice category remained stable, whereas prune juice and fruit nectars dropped more than 12 % and 18 %, respectively, during 1978 – 1983.

5.2. Carbonated Soft Drinks

Figure 15 compares the history of soft drink production in the United States with the rest of the world. In 1982, an estimated 16.8 $\times 10^9$ cases were manufactured throughout the world [125]. Of this, almost 37 % was produced and consumed in the United States.

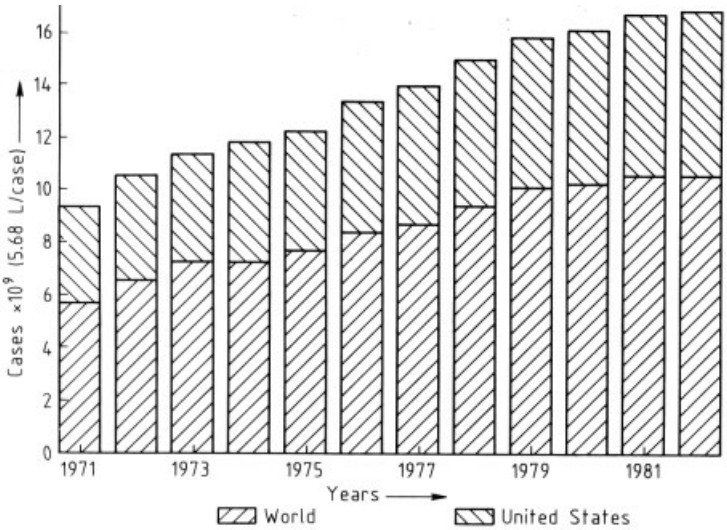

Figure 15. Estimated annual production of sweetened car-bonated soft drinks throughout the world [125]

Table 10. History of colas in the United States

Major brand introductions	Brand
May, 1886	Coca-Cola
April, 1898	Pepsi-Cola
April, 1934	Royal Crown (RC)
March, 1962	Diet Rite
May, 1963	Tab
March, 1964	Diet Pepsi
February, 1977	Pepsi Light (sugar-free)
April, 1980	Sugar-Free RC 100∗
March, 1982	Like∗
	RC 100∗
August, 1982	Pepsi Free∗
	Diet Pepsi Free∗
	Diet Coke∗
October, 1982	Sugar Free Like∗
May, 1983	Caffeine-Free Coke∗
	Caffeine-Free Diet Coke∗
	Caffeine-Free Tab∗
August, 1983	Diet Rite (salt/sodium free)∗

∗ Caffeine-free

5.2.1. United States Market

The United States continues to be a unique, evolutionary, and often revolutionary soft drink market. It is here where cola beverages were first discovered (Table 10), and where diet beverages became a major market segment. In 1949, the United States manufactured just over 1×10^6 cases (8.52 L) valued at \$760000. In 1982, 4.07×10^9 cases were manufactured valued at $\$21.2\times10^9$ (wholesale). This growth is illustrated in Figure 16.

The extraordinary growth of carbonated soft drink consumption has taken place at the expense of other beverages (Fig. 17). In 1962, carbonated soft drinks accounted for 9.2 % of per capita beverage consumption. In 1982, almost 22 % of all liquid consumption in the United States was soft drinks. Over this period (1962 – 1982), the percentage of coffee, milk, and water consumed per capita fell dramatically because of the increase in soft drink consumption as well as increases in beer, wine, bottled water, and powdered beverage consumption.

A major factor in the extraordinary growth of soft drink consumption has been cola beverages, first introduced over 100 years ago. From 1966 to 1982, colas consistently accounted for 62 – 63 % of total soft drink production, lemon – lime beverages 11 – 12 %, and root beers, another unique United States beverage, about 5 – 5.5 % (Fig. 18) [125–127]. During this same period, the acceptance of pepper beverages has steadily increased to about 7.5 % of total production. Orange consumption fell slightly, dropping from 9.5 % in 1966 to 7.6 % in 1982, as did the production of other flavored soft drinks, which accounted for 5.1 % of 1982 production.

In order to maintain growth, the United States marketplace has become much more segmented, particularly in the cola category, as evidenced

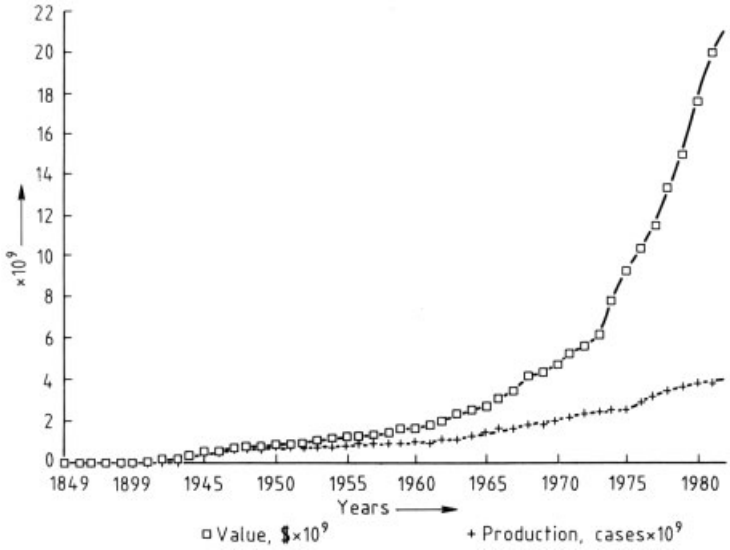

Figure 16. Growth of United States soft drink industry [125]

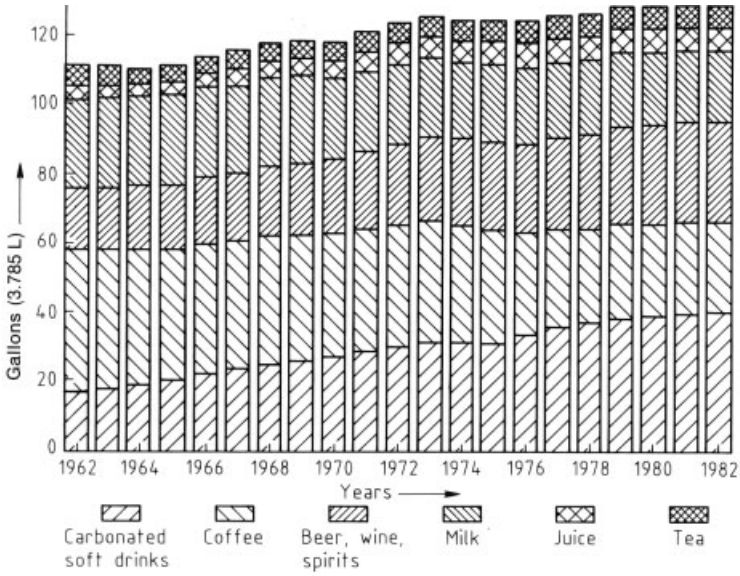

Figure 17. Per capita consumption of various beverages by United States consumers [125–127]

in Table 10. Figure 19 shows the extraordinary growth and acceptance of diet beverages. The growth of diet beverages can be correlated especially to the entry of major manufacturers into the diet cola market. In 1983, diet colas accounted for approximately 67 % of the total diet market. The cola market has been further segmented with the introduction of caffeine-free colas, and their impact has been significant. In 1983, regular caffeine-free colas accounted for 6.1 % of the cola market, and diet caffeine-free colas for 5.4 %.

The prime movers and benefactors of the United States consumers' love affair with colas

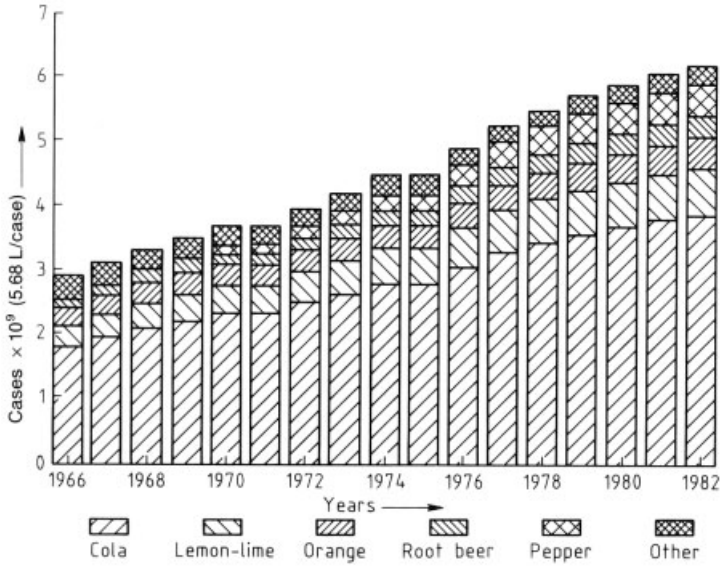

Figure 18. United States soft drink production by flavor type [125–127]

Figure 19. Share of United States diet beverages sold per annum – based on sales of bottled and canned beverages, and including all flavor types

has been the major cola manufacturers, as shown in Figure 20. From 1966 to 1982, the market was dominated by Coca-Cola and Pepsi-Cola companies, accounting for 34.5 and 25.0 % respectively, of 1982 production, followed by Seven-Up, Royal Crown, and Dr. Pepper, which accounted for 6.9, 4.3, and 6.4 %, respectively.

5.2.2. International Markets

In past years, international sales of carbonated soft drinks has grown in a manner similar to the United States, although volume has shown some instability because of tensions from worldwide economic conditions. In 1982, 10.6×10^9 cases of carbonated soft drinks were sold throughout

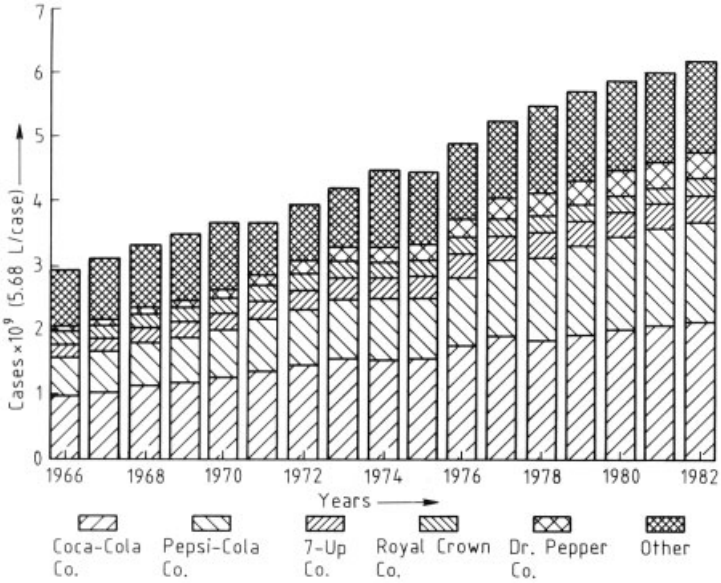

Figure 20. United States production of carbonated soft drinks by major manufacturers [125–127].

the world (Fig. 21). Over half this production was in those countries listed in Table 11 [128].

Table 11. Estimated market size of selected major international markets [128]

Country	1982 $\times 10^6$ cases *
Argentina	173.2
Australia	174.5
Belgium	103.9
Brazil	500.2
Canada	293.9
Federal Republic of Germany	851.2
France	212.4
Italy	284.5
Japan	480.8
Mexico	971.4
Netherlands	156.4
South Africa	176.1
Spain	322.1
United Kingdom	352.2
Venezuela	195.7
Total	5248.5

* 1 Case = 5.68 L

Like the United States, the international soft drink market also supports much cola beverage consumption, but to a much lesser extent (Fig. 21). In 1982, cola beverage production accounted for 37.1 % of production, up from 31.3 % in 1971. Diet colas, as well as other diet soft drinks, played an insignificant role in worldwide soft drink production in 1982, accounting for only 0.5 % of total production. Further segmentation into caffeine-free colas is essentially nonexistent in international markets, with the primary exception of Canada.

Although cola beverages are not dominant internationally, the primary United States cola manufacturers accounted for nearly 45 % of all international sales in 1982 (Fig. 22).

6. References

1. Y. H. Hui: *United States Food Laws, Regulations, and Standards,* J. Wiley & Sons, New York 1979.
2. H. W. Ockerman: *Source Book for Food Scientists,* AVI Publishing Co., Westport, Conn. 1978.
3. U.S. Food and Drug Administration, *Fed. Reg.* **49** (1984) 22831.
4. C. S. Pederson in B. S. Luh, J. G. Woodroof (ed.): *Commercial Vegetable Processing,* AVI Publishing Co., Westport, Conn. 1975, pp. 290 – 316.
5. J. G. Woodroof, G. F. Phillips: *Beverages: Carbonated and Non-carbonated,* AVI Publishing Co., Westport, Conn. 1981.

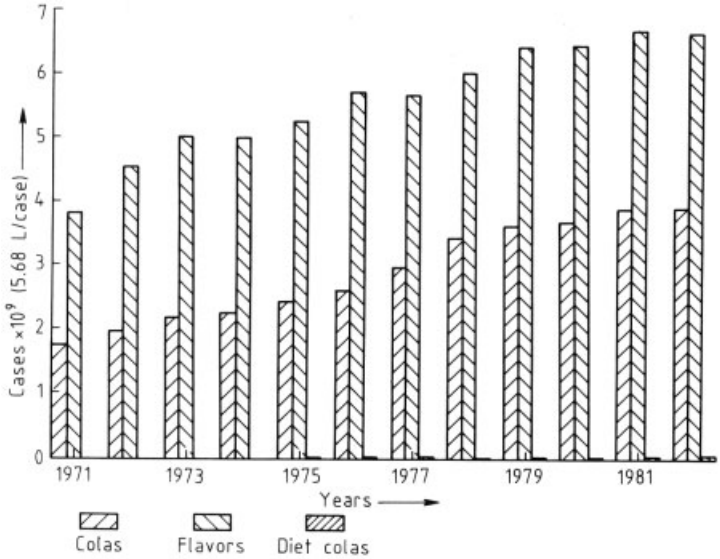

Figure 21. International production of cola, flavored, and diet cola soft drinks

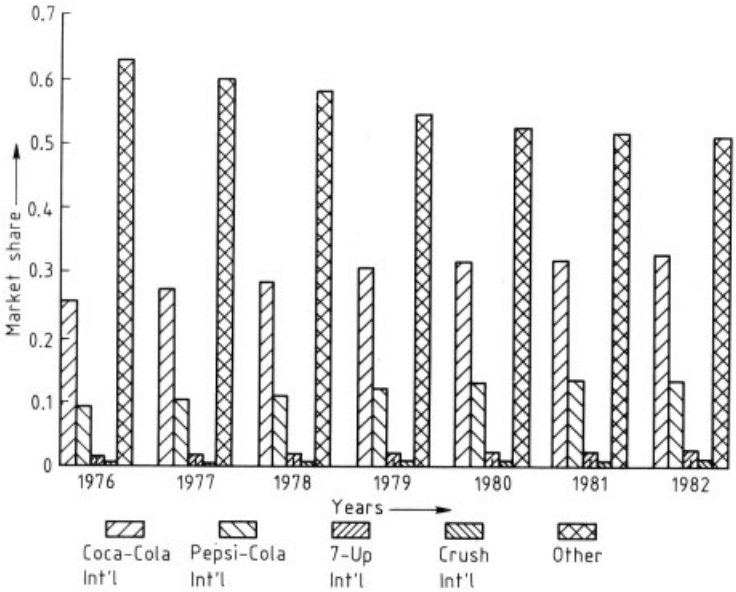

Figure 22. Share of international markets by major United States beverage manufacturers

6. American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE) (ed.): *ASHRAE Data Book-Applications,* ASHRAE, New York 1971.
7. J. A. Attaway, C. D. Atkins, M. D. Maraulja in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 2,** AVI Publishing Co., Westport, Conn. 1977, pp. 382 – 400.
8. R. Pimienta: *L'Antiquité du Thermalisme,* Agrégé de l'Université.
9. C. James: *Guide pratique du Médecin et du Malade aux Eaux Minérales,* Librairie Victor Masson, Paris, 1857.
10. Mace: *Guide Villes d'Eaux, Bains de Mer et Stations,* 1881.
11. M. Jaltel: *La Santé par les Eaux – 2000 ans de Thermalisme,* 1983.

12. "Relative au rapprochement des législations des Etats Membres concernant l'exploitation et la mise dans le commerce des eaux minérales naturelles," *Journal Officiel des Communautés Européennes,* **1 – 10** (Aug. 30, 1980) no. 2229, Directive du Conseil July 15.
13. B. Ninard: "Considérations sur le conditionnement moderne des eaux de boisson," *Bordeaux Médical* **12** (1979) no. 15, 1085.
14. G. Roumilhac: *Le Thermalisme français,* Syndicat National des Etablissements Thermaux de France.
15. L. Merklen: *Richesses hydro-minérales françaises,* Institut d'Hydrologie et de Climatologie à la Faculté de Médecine de Nancy, Nancy.
16. *Thérapeutique thermale et climatique,* Masson & Cie, Paris 1972. (Précis d'hydrologie et de climatologie pour la préparation du diplôme français de Docteur en Médecine.)
17. C. Dedet: *Guide des cures thermales,* Editions du Jour 1975.
18. *Les travaux du Centre de Recherche d'Hydrologie de Vichy* .
19. J. C. Gérardin, B. Hacq: *Eau sulfatée calcique et magnésienne de Contrexéville et Diurèse* .
20. *Thermoclimatisme et eaux minérales,* Journal Officiel de la République française, no. 1409.
21. Projet de législation allemande DBMJFG. 416/415, 6411, I STAND February 25, 1983, Entwurf einer Verordnung über natürliches Mineralwasser, Quellwasser und Tafelwasser (Mineral und Tafelwasser Verordnung).
22. J. Brod: "Considérations Générales sur les procédés modernes d'embouteillage des eaux minérales," *Bulletin de la Société Scientifique d'Hygiène alimentaire "L'alimentation et la vie,"* **52** (1964) no. 1, 2, 2.
23. *Latest Developments in P.V.C. Packaging of Foodstuffs,* GECOM PVC, Paris 1980.
24. USDA, *Agricultural Statistics 1983,* U.S. Gov't. Printing Office, Washington, D.C., 1983, pp. 188 –237.
25. USDA, *Fruit Outlook and Situation,* U.S. Gov't. Printing Office, Washington, D.C., March 1982, pp. 1 – 53.
26. F.A.O., *Production Year 1982,* vol. 36, F.A.O. of United Nations, Rome, Italy, 1982.
27. H. Thonges, H. Rinder, *Confructa* **28** (1984) no. 1, 8 – 54.
28. J. L. Apple, R. F. Smith (ed.): *Integrated Pest Management,* Plenum Publishing, New York 1976.
29. N. F. Childers: *Modern Fruit Science: Orchard and Small Fruit Culture,* 9th ed., Horticultural Publications, Gainesville, Fla., 1983.
30. J. Janick, J. N. Moore (ed.): *Advances in Fruit Breeding,* Purdue University Press, West Lafayette, Ind., 1975.
31. B. F. Cargill, D. E. Booster: "Vibrational Removal Techniques, High Density, Applications," in M. O'Brien, B. F. Cargill, R. B. Fridley (ed.): *Principles and Practices for Harvesting and Handling Fruits and Nuts,* AVI Publishing Co., Westport, Conn., 1983, pp. 189 – 243.
32. V. E. Petrucci, C. D. Clary, M. O'Brien: "Grape Harvesting Systems" in M. O'Brien, B. F. Cargill, R. B. Fridley (ed.): *Principles and Practices for Harvesting and Handling Fruits and Nuts,* AVI Publishing Co., Westport, Conn., 1983, pp. 525 – 574.
33. J. G. Woodroof, B. S. Luh (ed.): *Commercial Fruit Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1984.
34. R. L. LaBelle, R. L. Lind, R. W. Kime, *Proc. N.Y. State Hortic. Soc.* **123** (1978) 180 – 183.
35. R. L. LaBelle, R. S. Shallenberger, R. D. Way, L. R. Mattick et al.: "The Relationship of Apple Maturity to Applesauce Quality," *Food Technol.* **9** (1960) 463 – 468.
36. L. Poll, J. M. Flink: "Influence of Storage Time and Temperature on the Sensory Quality of Single-Variety Apple Juice," *Lebensm.-Wiss. Technol.* **16** (1983) 215 – 219.
37. R. L. LaBelle: "Apple Quality Characteristics as Related to Various Processed Products," in R. Teranishi, H. Barrera-Benitez (ed.): *Quality of Selected Fruits and Vegetables of North America,* ACS Symposium Series 170, Amer. Chem. Soc., Washington, D.C., 1981, pp. 61 – 76.
38. R. M. Smock, A. M. Neubert: "Apples and Apple Products," in Z. I. Kertesz (ed.): *Economic Crops,* vol. 2, Interscience, New York 1950.
39. L. R. Mattick, J. C. Moyer: "Fruit and Fruit Products: Composition of Apple Juice," *J. Assoc. Off. Anal. Chem.* **66** (1983) no. 5, 1251 – 1255.
40. A. El-Sayed, B. S. Luh: "Amino Acids and Quality of Canned Apricots as Affected by Nitrogen Fertilization," *Food Technol.* **21** (1967) no. 3 A, 98 A – 102 A.
41. R. J. Romani, W. G. Jennings: "Stone Fruits," in A. C. Hulme (ed.): *The Biochemistry of Fruits and Their Products,* vol. 2, Academic Press, London and New York 1971.

42. B. S. Luh: "Nectars, Pulpy Juices and Fruit Juice Blends," in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* 3rd ed., AVI Publishing Co., Westport, Conn., 1980, pp. 436 – 505.
43. B. J. E. Teskey, J. S. Shoemaker: "Apricots and Nectarines," in B. J. E. Teskey, J. S. Shoemaker (ed.): *Tree Fruit Production,* 3rd ed., AVI Publishing Co., Westport, Conn., 1978, pp. 387 – 401.
44. G. K. Brown, D. L. Peterson, J. H. Levin: "Tree Fruit Harvesting Systems: I. Individual Trees," in M. O'Brien, B. F. Cargill, R. B. Fridley (ed.): *Principles and Practices for Harvesting and Handling Fruits and Nuts,* AVI Publishing Co., Westport, Conn. 1983, pp. 575 – 587.
45. D. K. Tressler, V. L. S. Charley, B. S. Luh: "Cherry, Berry and Other Miscellaneous Fruit Juices," in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* 3rd ed., AVI Publishing Co., Westport, Conn., 1980, pp. 310 –343.
46. B. J. E. Teskey, J. S. Shoemaker: "Cherries," in B. J. E. Teskey, J. S. Shoemaker: *Tree Fruit Production,* 3rd ed., AVI Publishing Co., Westport, Conn., 1978, pp. 292 – 357.
47. J. S. Shoemaker: "Grapes," in J. S. Shoemaker: *Small Fruit Culture,* 5th ed., AVI Publishing Co., Westport, Conn., 1978, pp. 1 – 102.
48. A. D. Webb: "Quality Factors in California Grapes," in R. Teranishi, H. Barrera-Benitez (ed.): *Quality of Selected Fruits and Vegetables of North America,* ACS Symposium Series 170, Amer. Chem. Soc., Washington, D.C., 1981, pp. 61 – 76.
49. B. S. Luh, C. E. Kean: "Canning of Fruits," in J. G. Woodroof, B. S. Luh (ed.): *Commercial Fruit Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1984, pp. 1 – 164.
50. D. M. Wilson, G. J. Nuovo, *J. Appl. Microbiol.* **26** (1973) 124 – 125.
51. B. S. Luh, C. E. Kean: "Canning of Fruits," in J. G. Woodroof, B. S. Luh (ed.): *Commercial Fruit Processing,* AVI Publishing Co., Westport, Conn., 1975, pp. 141 – 265.
52. J. C. Moyer, H. C. Aitken: "Apple Juice," in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* 3rd ed., AVI Publishing Co., Westport, Conn., 1980, pp. 212 – 267.
53. F. M. Rombouts, W. Pilnik: "Enzymes in Fruit and Vegetable Juice Technology," *Process Biochem.* **13** (1978) no. 8, 9 – 13.
54. A. Kilara: "Enzymes and Their Uses in the Processed Apple Industry, A Review," *Process Biochem.* 17 (1982 – July/August) no. 4, 35 – 41.
55. T. E. Acree: "The Odor Quality of Labrusca Grapes," in R. Teranishi, H. Barrera-Benitez (ed.): *Quality of Selected Fruits and Vegetables of North America,* ACS Symposium Series 170, Amer. Chem. Soc., Washington, D.C., 1981, pp. 11 – 19.
56. C. S. Pederson: "Grape Juice," in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* 3rd ed., AVI Publishing Co., Westport, Conn., 1980, pp. 268 – 309.
57. C. F. Adams: "Nutritive Value of American Foods In Common Units," *Agriculture Handbook No. 456,* USDA, U.S. Gov't. Printing Office, Washington, D.C., 1975.
58. R. E. Wrolstad, C. J. Cornwell, J. D. Culbertson, F. G. R. Reyes: "Establishing Criteria for Determining the Authenticity of Fruit Juice Concentrates," in R. Teranishi, H. Barrera-Benitez (ed.): *Quality of Selected Fruits and Vegetables of North America,* ACS Symposium Series 170, Amer. Chem. Soc., Washington, D.C., 1981, pp. 77 – 93.
59. J. M. Krochta, B. Feinberg: "Effects of Harvesting and Handling on Fruits and Vegetables," in R. S. Harris, E. Karmas (ed.): *Nutritional Evaluation of Food Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1975, pp. 98 – 118.
60. C. T. Young, J. G. Woodroof: "Composition and Nutritive Value of Raw and Processed Fruits," in J. G. Woodroof, B. S. Luh (ed.): *Commercial Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1984, pp. 539 – 579.
61. R. S. Harris, E. Karmas: *Nutritional Evaluation of Food Processing,* 2nd ed., AVI Publishing Co., Westport, Conn. 1975.
62. D. B. Lund: "Effects of Blanching, Pasteurization and Sterilization on Nutrients," in R. S. Harris, E. Karmas (ed.): *Nutritional Evaluation of Food Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1975, pp. 205 – 240.
63. J. G. Woodroof: "Grades and Standards for Raw and Processed Fruits," in J. G. Woodroof, B. S. Luh (ed.): *Commercial Fruit Processing,* 2nd ed., AVI Publishing Co., Westport, Conn., 1984, pp. 580 – 594.
64. C. Davis in R. F. Matthews (ed.): *Future Trends in the Citrus Processing Industry,* IFAS, University of Florida, Gainesville, Fla., 1983, pp. 43 – 68.

65. USDA, *Foreign Agricultural Circular FHORT 1 – 84,* USDA, Washington, D.C., 1984.
66. United States Government, US 2543109, 1948 (L. G. MacDowell, E. L. Moore, C. D. Atkins).
67. L. G. Albrigo, R. D. Carter in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 1,** AVI Publishing Co., Westport, Conn., 1977, pp. 37 – 73.
68. Florida Department of Agriculture, Division of Fruit and Vegetable Inspection, *Annual Report* Winter Haven, Fla., 1980.
69. S. V. Ting in S. Nagy, J. A. Attaway (ed.): *Citrus Nutrition and Quality,* ACS Symposium Series 143, Amer. Chem. Soc., Washington, D.C., 1980, pp. 3 – 24.
70. H. E. Swisher, L. H. Swisher in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 2,** AVI Publishing Co., Westport, Conn., 1977, pp. 253 – 289.
71. J. W. Kesterson, R. J. Braddock, *University of Florida Bulletin 784,* University of Florida, Gainesville, Fla., 1976.
72. W. H. J. M. Pelt, W. J. M. Swinkels: *Trans. 1984 Citrus Eng. Conf.,* ASME, Lakeland, Fla., 1984.
73. C. S. Chen, *Trans. ASAE* **25** (1982) no. 5, 1457 –1463.
74. R. W. Cook: *Trans. 1963 Citrus Eng. Conf.,* ASME, Lakeland, Fla., 1963.
75. R. W. Cook: *Proc. Third International Congress Food Science and Technology,* Washington, D.C., 1970, pp. 452 – 454.
76. C. S. Chen: *AICHE Winter Meeting,* Orlando, Fla., 1982.
77. R. E. Berry, M. K. Veldhuis in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 2,** AVI Publishing Co., Westport, Conn., 1977, pp. 177 – 252.
78. USDA, *Foreign Agricultural Circular FCF 4 – 81,* U.S. Gov't. Printing Office, Washington, D.C., 1981.
79. Florida Citrus Processors Association, *Statistical Summary, 1981 – 1982,* Winter Haven, Fla., 1982.
80. D. L. Gunter: *Trans. 1983 Citrus Eng. Conf. ASME,* Lakeland, Fla., 1983.
81. C. Varsel in S. Nagy, J. A. Attaway (ed.): "Citrus Nutrition and Quality," *ACS Symposium Series 143,* Amer. Chem. Soc., Washington, D.C., 1980, pp. 225 – 272.
82. R. D. Carter, *Technical Manual,* Florida Department of Citrus, Lakeland, Fla., 1981.
83. USDA, *Agricultural Handbook 98,* U.S. Gov't. Printing Office, Washington, D.C., 1962.
84. "Food and Drugs", CFR Title 21, U.S. Food and Drug Administration, U.S. Gov't. Printing Office, Washington, D.C., 1979, parts 100 – 199.
85. S. V. Ting, E. L. Moore, J. W. McAllister, R. R. Streiff et al., *Proc. Fla. State Hortic. Soc.* **89** (1974) 206 – 209.
86. S. V. Ting in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 2,** AVI Publishing Co., Westport, Conn., 1977, pp. 401 –444.
87. P. E. Araujo in S. Nagy, P. E. Shaw, M. K. Veldhuis (ed.): *Citrus Science and Technology,* **vol. 1,** AVI Publishing Co., Westport, Conn., 1977, pp. 1 – 32.
88. W. Popenoe: *Manual of Tropical and Subtropical Fruits,* Hafner Press, New York 1974.
89. R. Knight in S. Nagy, P. E. (ed.): *Tropical and Subtropical Fruits,* AVI Publishing Co., Westport, Conn., 1980, pp. 1 – 120.
90. USDA, *Foreign Agricultural Circular FHORT 2 – 82,* USDA Washington, D.C., 1982.
91. F. P. Mehrlich, G. E. Felton in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* AVI Publishing Co., Westport, Conn., 1980, pp. 180 – 211.
92. *The Almanac,* Edward E. Judge & Sons, Inc., Westminster, Md., 1983.
93. Pennwalt Corp., US 3318709, 1967 (F. J. Cygan, J. Farmer).
94. J. K. Paul: *Fruit and Vegetable Juice Processing,* Noyes Data Corp., Park Ridge, N.J., 1975.
95. B. S. Luh in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* AVI Publishing Co., Westport, Conn., 1980, pp. 436 –505.
96. M. Burger, L. W. Hein, L. Tepley, P. H. Derse et al., *J. Agric. Food Chem.* **4** (1956) 418 – 425.
97. H. T. Chan, E. Chenchin, P. J. Vonnahme, *J. Agric. Food Chem.* **21** (1973) 208 – 211.
98. G. G. Dull in A. C. Hulme (ed.): *The Biochemistry of Fruits and Their Products,* Academic Press, London 1971, pp. 303 – 324.
99. R. Flath in S. Nagy, P. E. Shaw (ed.): *Tropical and Subtropical Fruits,* AVI Publishing Co., Westport, Conn., 1980, pp. 157 – 183.
100. C. S. Pederson in P. E. Nelson, D. K. Tressler (ed.): *Fruit and Vegetable Juice Processing Technology,* AVI Publishing Co., Westport, Conn., 1980, pp. 573 – 596.
101. J. C. Moyer, W. B. Robinson, J. R. Ransford, R. L. LaBelle et al., *Food Technol.* **13** (1959) 270 – 272.

102. D. B. Hand, J. C. Moyer, J. R. Ransford, J. C. Hening et al., *Food Technol.* **9** (1955) 228 – 235.
103. W. B. Robinson, L. B. Kimball, J. R. Ransford, J. C. Moyer et al., *Food Technol.* **10** (1956) 109 – 112.
104. J. R. Wagner, J. C. Miers, D. W. Sanshuck, R. Becker, *Food Technol.* **22** (1968) 1484 – 1488.
105. J. R. Wagner, J. C. Miers, D. W. Sanschuck, R. Becker, *Food Technol.* **23** (1969) 247 – 250.
106. G. B. Beattie: *The Soft Drinks Bottler,* The Mineral Water Trade Review Co., Ltd., London 1950.
107. J. F. Riley: *A History of the American Soft Drink Industry,* American Bottlers of Carbonated Beverages, Washington, D.C., 1958.
108. J. C. Louis, H. Z. Yasijian: *The Cola Wars,* Everest House Publishers, New York 1980.
109. J. E. Tillotson, *Food Technol.* **38** (1984) 3, 63.
110. E. Guenther: *The Essential Oils,* Robert E. Krieger Publishing Co., Huntington, New York 1974.
111. S. Arctander: *Perfume and Flavor Materials of Natural Origin,* Privately published, Elizabeth, N.J., 1960.
112. H. B. Heath: *Source Book of Flavors,* AVI Publishing Co., Westport, Conn., 1981.
113. PepsiCo Inc., US 3622343, 1971 (M. H. Anwar).
114. T. E. Furia: *Handbook of Food Additives,* Chemical Rubber Co., Cleveland, Ohio, 1968.
115. The American Water Works Association: *Water Quality and Treatment,* McGraw-Hill, New York 1971.
116. L. F. Green: *Developments in Soft Drinks Technology I,* Applied Science Publishers, Ltd., London 1978.
117. M. T. Gillies: *Soft Drink Manufacture,* Noyes Data Corp., Park Ridge, N.J., 1973.
118. M. H. Gutcho: *Fortified and Soft Drinks,* Noyes Data Corp., Park Ridge, N.J., 1977.
119. M. B. Jacobs: *Manufacture and Analysis of Carbonated Beverages,* Chemical Publishing Co., New York 1959.
120. J. G. Woodroof, G. F. Philipps: *Beverages Carbonated and Non-carbonated,* AVI Publishing Co., Westport, Conn., 1974.
121. C. E. Hamilton: *Manual on Water,* ASTM, Philadelphia 1978.
122. *Beverage World* **103** (April, 1984) 30 – 45.
123. *Beverage World* **103** (July, 1984) 36 – 44.
124. J. D. Stacey, *Beverage World* **103** (July, 1984) 54 – 60.
125. *Beverage Industry Annual Manual,* Harcourt Brace Jovanovich Publications, Cleveland, Ohio, 1984, no. 17.
126. *Beverage Industry Annual Manual,* Harcourt Brace Jovanovich Publications, Cleveland, Ohio, 1974 –1975, no. 8.
127. *Beverage Industry Annual Manual,* Harcourt Brace Jovanovich Publications, Cleveland, Ohio, 1979 –1980, no. 13.
128. *Beverage Digest,* vol. 4, no. 1, J. Meyers, Publ., Greenwich, Conn., 1984.

Biacetyl → Ketones
Biguanide → Guanidine
Biliary Tract Stimulants → Gallbladder and Liver Therapy

Biological Control

Leslie C. Lewis, Agricultural Research Service, USDA, Ankeny, Iowa 50021, and Iowa State University, Ames, Iowa 50011, United States

1. Introduction 1
1.1. Entomophagous Insects 2
2. Biological Control of Insects 3
2.0.1. Parasitoids 3
2.0.2. Insect Predators 4
2.1. Other Predators 5
2.2. Entomopathogens 5
2.2.1. Bacteria 5
2.2.2. Pathogens with Protozoal Characteristics 7
2.2.3. Insect Viruses 8
2.2.4. Fungi 9
2.3. Nematodes 10
3. Biological Control of Weeds 11
3.1. Phytophagous Insects 11
3.2. Plant Pathogens 11
3.3. Higher Animals 12
4. Biological Control of Plant Pathogens 12
4.1. Macroorganisms 12
4.2. Microorganisms 12
5. Other Biological Methods of Control 12
5.1. Biochemicals 12
5.1.1. Growth Regulators 12
5.1.2. Semiochemicals 14
5.1.3. Bioderived Toxicants 15
5.2. Autocidal 15
5.2.1. Sterile Male 15
5.2.2. Hybrid Sterility 16
6. Formulation, Application, and Efficacy 16
7. References 19

1. Introduction

Biological control is the action of natural enemies in maintaining another organism's population density at a lower level than would otherwise occur. In this article, however, the traditional definition has been expanded to include chemicals that duplicate or mimic natural ones. The major effort in biological control has been focused on the suppression of populations of pest insects. However, biological control of weeds and plant pathogens is a very active area of research. Biological control techniques also have been used successfully to suppress populations of animals other than insects. Biological control is a global endeavor; specialists in this discipline have developed a worldwide organization, *International Organization for Biological Control of Noxious Animals and Plants* (IOBC). More information on IOBC may be obtained by contacting: Secretary-General, G. Mathys, 1, rue Le Nôtre, F-75016 Paris.

Classical biological control involves: (1) discovery, (2) importation, (3) release, and (4) establishment of exotic species of natural enemies. Most pest insects, especially in the United States, are pests because they were introduced accidentally, or on purpose, leaving their complex of natural enemies behind. For example, several thousand species of insects have been introduced into the United States in attempts to control a few species of insect pests.

Applied biological control is sometimes implemented when a natural enemy does not become established. This involves: (1) augmentation and (2) conservation. Augmentation is the manipulation of a natural enemy, and conservation is the manipulation of the environment, both to enhance the impact of the natural enemy on the pest population. Examples of augmentation are colonization and periodic release of natural enemies and improvement of the efficiency of a natural enemy through selection. Conservation involves, e.g., using different cultural practices that serve to protect the natural enemy in the environment.

Natural enemies are: (1) parasitoids, (2) predators, and (3) pathogens. A parasitoid (parasite) is an insect that lays an egg on, or in, a single host insect. The parasitoid develops internally, killing the host. Predators are free-living organisms that kill and consume their prey im-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_077

mediately or within a short period of time. Predatory insects kill and consume more than one host or prey during their development. Predators are found in all major orders of insects. Parasitoids and predators are referred to as entomophagous insects. Pathogens are microorganisms that cause diseases in insects. Most insect diseases are caused by bacteria, fungi, protozoa, microspora, and viruses. These organisms are collectively referred to as entomopathogens. The phylum Nematoda consists of a group of round worms called nematodes. Some nematodes are natural enemies of insects; some nematodes behave like parasitoids, whereas others behave like insect pathogens. These nematodes are referred to as entomophagous nematodes.

In the biological control of weeds, natural enemies usually are phytophagous insects and microorganisms that cause diseases of plants. Natural enemies of plant pathogens include predacious insects that feed on the fruiting bodies of pathogenic fungi and microorganisms that are parasitic on other microorganisms.

History. Several references are made to early examples of natural enemies being used to control pests. For example, in the year 1200, Chinese citrus growers purchased nests of the predacious ant, *Oecophylla smaragdina*, and placed them in citrus trees to control a caterpillar, *Tessarotoma papillosa.* Many other examples of successful biological control are recorded, including the introduction to California of the vedalia beetle, *Rodolia cardinalis,* in 1888 to control the cottony cushion scale, *Icerya purchasi* [24]. This project saved the citrus industry of California and was the impetus for additional biological control projects against insect pests.

Widespread adoption of host-specific biological control techniques was slowed by the development of relatively inexpensive synthetic chemical insecticides that had a broad spectrum of insect targets. These compounds were envisioned to be the permanent solution for problems with insect pests.

The public concern for environmental conservation that has surfaced over the past 25 years has again stimulated interest in, and financial support for, biological control programs. During this time, intensive research has been conducted in the area of insect biochemistry; specifically, insect hormones and semiochemicals. *Hormones* are chemicals that are produced in certain body tissues, are transported within the body by the circulatory system, and elicit a response in another part of the body (→ Hormones). *Semiochemicals* also are produced within the body but are emitted externally to elicit a response. The hormones of interest in biological control are those that regulate growth and molting of the insect; the semiochemicals are those that elicit a sexual response or attract an insect to a point source. The use of biochemicals, such as hormones and semiochemicals, is not a method of biological control in the strict sense. These chemicals, plus male-sterile techniques, host plant resistance, and genetic manipulation, are merely other biological methods used to control insects (see Chap. 5).

Entomologists have determined the economic threshold or economic injury levels for many important insect pests on many agricultural crops. These data allow a grower to determine how many insects can be tolerated by a crop before control is economically feasible. In developing this information, researchers have become aware that complete control or elimination of a pest is not necessary and that measures less drastic than a chemical insecticide, such as biological control, can be effective in protecting crops. This has brought about the present-day approach to pest control known as *integrated pest management* where a holistic approach is taken in suppressing pest populations. Briefly, this involves the assessment of a pest problem and the judicious use of two or more methods of control (parasitoids, chemical insecticides, predators, pathogens, hormones, semiochemicals, plant resistance, and cultural practices) in a compatible system that will reduce the pest population to an acceptable level. At all times, attention must be paid to using sound ecological principles so as not to disturb the beneficial regulatory factors present in nature.

1.1. Entomophagous Insects

An ecosystem has a relationship between a host (plant, animal, or commodity), insects feeding on the host, and organisms feeding on the insects. An example of this relationship is illustrated in Figure 1.

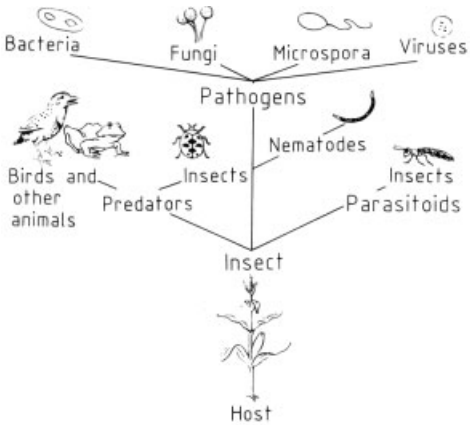

Figure 1. Example of an ecosystem

2. Biological Control of Insects

2.0.1. Parasitoids

The great majority of parasitoids are found in the insect orders of Strepsiptera, Coleoptera, Lepidoptera, Hymenoptera, and Diptera, with by far the most being classified as Hymenoptera and Diptera. To be successful, a parasitoid should have the following attributes: (1) good searching ability, (2) high biotic potential, and (3) ability to operate effectively throughout the entire geographic range of its host. A typical female parasitoid selects a suitable host, deposits a single egg either on or within it, and leaves in search of other suitable hosts. The parasitoid egg hatches and develops within the host insect. During the early stages of development, the parasitoid does not greatly disturb the host insect. Later, as the parasitoid reaches maturity, the host insect becomes immobile, at which time the parasitoid will leave the host's body and pupate. Shortly afterward, the host usually dies.

There are many variations in the parasitoids as to developmental stage of host most desirable for parasitization. Some parasitoids prefer young larvae, some older larvae, and some even the pupal stage. Other parasitoids parasitize only insect eggs.

An example of a "generalized" hymenopteran parasitoid is *Bathyplectes curculionis*, imported from Italy to assist in controlling the alfalfa weevil, *Hypera postica.* Some parasitoids deposit several eggs in a host insect. These eggs hatch, and then at maturity several hymenopterous pupae are seen attached to the surface of the host insect. *Trichogramma* spp. are well-known as egg parasitoids. The females of these minute insects deposit one or more eggs within the eggs of different species of insects (Fig. 2). There is sufficient nutritive material present within the egg to support the development of this small parasitoid.

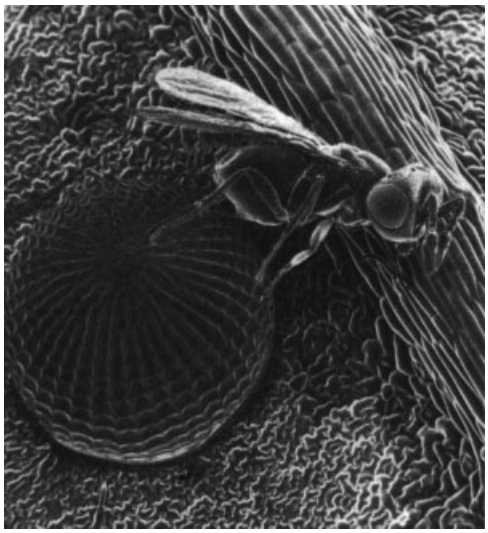

Figure 2. *Trichogramma* wasp preparing to oviposit in an insect egg
∗ Courtesy of Gerald R. Carner, Clemson University

Some Hymenoptera possess a biological modification known as polyembryony, the development of several individuals from a single egg. In this situation, an adult female, e.g., *Macrocentrus grandii* (Fig. 3), a parasitoid of the European corn borer, *Ostrinia nubilalis*, deposits a single egg in a second-stage corn borer larva. The primary embryonic germ divides by fission to form two secondary germs which split into a variable number of morulae, each of which develops into an embryo.

Dipterous parasitoids exhibiting a general developmental cycle are represented by *Winthemia quadripustulata*, a parasitoid of the armyworm, *Psuedaletia punctata.* The adult female deposits an egg directly behind the head capsule of a suitable host. The parasitoid larva hatches from the egg and chews its way into the body tissue of the host, develops to a mature larva, chews its way out of the larva, and pupates. A dipterous parasitoid with a different biology is *Blepharipa*

prantensis, a parasitoid of the gypsy moth, *Lymantria dispar.* The female parasitoid deposits its egg on the foliage where gypsy moth larvae will feed. The parasitoid is fully developed within the egg, and the slight pressure exerted by the mandibles during feeding causes the parasitoid to break out of the egg shell.

Figure 3. Adult female *Macrocentrus grandii,* parasitoid of the European corn borer, *Ostrinia nubilalis*

Bonnetia comta (Fig. 4), a parasitoid of several cutworms, deposits eggs on host fecal material. These eggs immediately hatch, and the neonates or planidia rest on their freshly cast egg chorion waiting for their host to move.

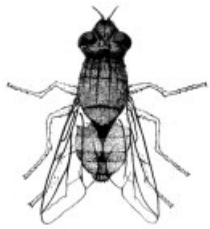

Figure 4. Adult female, *Bonnetia comta,* parasitoid of several cutworm larvae

Still another unique modification is the larvipositing phenomenon exhibited by *Lydella thompsoni*, a parasitoid of *O. nubilalis.* The adult female searches for entrance holes that tunneling larvae of the European corn borer have made and deposits live larvae in the fecal material voided by the feeding corn borer larva. The parasitoid larvae actively search for and enter the host by chewing through the insect cuticle.

These few examples of parasitoids give only a brief overview of the fascinating and sometimes complex biology of insect parasitoids.

2.0.2. Insect Predators

Predacious insects differ from parasitoids in that a predator consumes several insect forms during its life cycle. Also, in many instances, both the immature and adult stages of the insect are active predators. Most orders of insects contain at least one family of predators.

The scope of this article does not allow a treatment of examples from all orders of insects; only a few predators that a home gardener or general naturalist would most likely recognize are described. The order Coleoptera (beetles) contains by far the greatest number of predacious species. The ubiquitous lady beetle (Fig. 5) has been observed by many. The female lady beetle lays its eggs in clusters in a habitat near a food source. The newly hatched larva is a voracious feeder, consuming insect eggs, aphids, and small immature insects. The adult lady beetle also is predacious, devouring a general variety of prey. However, some of these coccinelid beetles are very specific in their prey, i.e., the vedalia beetle, *Rodolia cardinalis*, feeds only on the cottony cushion scale, *Icerya purchasi.* The ground beetles are another group of ubiquitous Coleoptera that are general predators, feeding on insect eggs, caterpillars, and even the adult moths of some species. There also are many predacious aquatic beetles that prey on almost any aquatic form of life that can be captured.

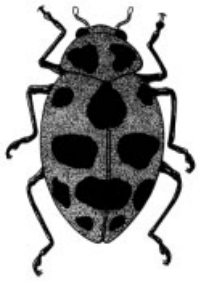

Figure 5. Adult lady beetle, *Coleomegilla maculata*

Dragonflies and damselflies (order Odonata) frequent country ponds in the summer. The immature forms of these insects are aquatic and are predators of pond life. The adults capture their prey in flight. Depending on the size of the predators, the prey ranges from mosquitoes to adults of the orders Lepidoptera and Hymenoptera.

Another common predator found near water is the robber fly (order Diptera). The larvae feed on pest insects in the soil, whereas the adults capture flying insects in flight. Their prey consists mainly of mosquitoes and midges that are readily found in an aquatic environment. Syrphid larvae, also in the order Diptera, feed on aphids.

Most members of the order Neuroptera are predators. A representative is the green lacewing. The larvae of this insect feed on aphids, mealybugs, thrips, and other small insects. These larvae also are known to prey on the eggs of several lepidopterans. The adult lacewings prey on aphids and mealybugs if they feed on insects at all.

The praying mantid is predacious in both the immature and adult stages. These insects prey on almost any insect that they can overcome. They frequent flowers and capture insects visiting the flowers for nectar. The Chinese mantid, *Tenodera aridifolia sinensis,* is frequently made available commercially for the home gardener.

2.1. Other Predators

Insects are a main ingredient in diets of several birds, rodents, fish, amphibians, reptiles, and arachnids. For example, several birds, including the downy woodpecker, crow, robin, red-winged blackbird, purple grackle, chickadee, starling, and ring-necked pheasant, are known predators of the European corn borer [25].

2.2. Entomopathogens

Insect pathogens (entomopathogens) are microorganisms that cause diseases in insects. These pathogens can cause diseases in beneficial insects as well as pest insects. In this article, only diseases of pest insects will be addressed.

2.2.1. Bacteria

Entomopathogenic bacteria can be divided into three groups – those that fit the description of classical biological control (see Chap. 2), those that fit the description of applied biological control, and those that fit into either group. Regardless of the category in which these bacteria are classified, they have one thing in common, they can be grown in an insect and/or in a bacteriological medium. The most prominent bacteria in biological control are in the genus *Bacillus*, the group of gram-positive, mobile, spore-forming rods.

Bacillus thuringiensis is the most widely known and researched bacterium within this group and is differentiated from other spore-forming bacilli by the presence of a parasporal body that is formed within the sporangium during sporogenesis. The parasporal body is a high-molecular-mass protein crystal that is referred to as crystalline protein, δ-endotoxin, as well as a parasporal body. This protein moiety possesses some of the insecticidal properties of the bacterium [26]. Certain subspecies of *B. thuringiensis* under specific growing conditions produce several other toxins with insecticidal activity: α-exotoxin, β-exotoxin, and γ-exotoxin. The δ-endotoxin, however, is the most important toxin relative to insecticidal activity and thus is the one most studied.

The δ-endotoxin is produced commercially in submerged culture under the conditions described in the patent [27]. Fermentation medium is crucial to successful production. It primarily contains an energy source (carbohydrate) and a nitrogen source. Common sources of carbohydrate are hydrolyzed corn products, starch, dextrose, and molasses. Sources of nitrogen are fishmeals, cottonseed flour, corn steep liquor, soybeans, autolyzed yeast, and casein. Trace minerals, namely Mg^{2+}, Mn^{2+}, Fe^{2+}, Zn^{2+}, and Ca^{2+}, are usually added to the medium.

Bacillus thuringiensis is primarily a pathogen of lepidopterans. A general infection cycle is illustrated in Figure 6. To elicit its effect, the crystalline protein must be consumed by a susceptible insect larva. The protein crystal is actually a protoxin that is hydrolyzed by enzymes in the gut of susceptible insects, releasing the pure toxin. The toxin causes paralysis of the gut. The insect either starves to death, or the midgut epithelial cells are damaged, allowing the gut contents, including *B. thuringiensis* spores, to enter the hemocoel. At this time, a general infection occurs.

Most Lepidoptera are susceptible to the δ-endotoxin, but there are basic differences in the

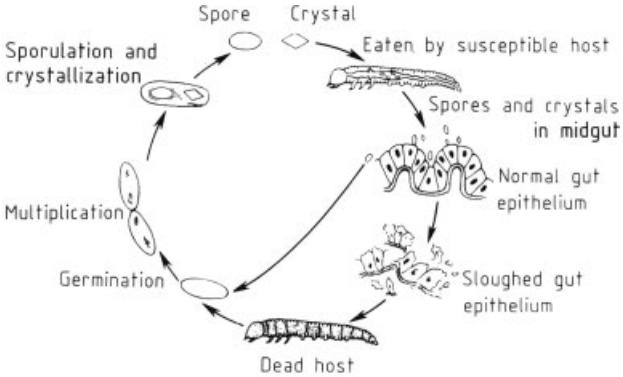

Figure 6. Infection cycle of the bacterium *Bacillus thuringiensis*

pathological responses of insects to this toxin. Researchers have divided lepidopterous larvae into four groups according to response [28]. The first group, designed Type I, shows a rapid general gut paralysis resulting in death 1 – 7 h after ingestion of the crystal toxin. Type II larvae are characterized by midgut paralysis a few minutes after ingestion of crystals and by cessation of feeding, but not by general paralysis. Type III larvae have to ingest both spores and crystals, do not exhibit either general or gut paralysis, but do stop feeding. Affected larvae enter a morbid state of varying duration with a fairly slow bacterial multiplication and finally, with a lethal infection. Type IV larvae are not susceptible to the crystal toxin or spores but are susceptible to the thermostable exotoxin(s).

Within these categories, the definitive response is dependent on the age of the insect. For example, the European corn borer is a Type III larva responding to a combination of spores and crystals. However, pure crystals are highly toxic to neonate larvae and moderately toxic to 6-d-old larvae. A combination of spores and crystals is necessary to cause maximum larval mortality [29].

Not only are the insects categorized relative to their response, but also isolates are divided into subspecies based on the antigenic properties of the flagella that are present during the vegetative stage of growth [30], [31]. Some subspecies are further divided by the serology of the crystalline protein [32]. With use of one or both of these techniques, > 20 subspecies of *B. thuringiensis* have been identified.

Insects susceptible to *B. thuringiensis* do not necessarily respond the same to all subspecies of *B. thuringiensis.* Likewise, a subspecies of *B. thuringiensis* does not have the same virulency to all susceptible insects. A bioassay technique is used to quantify the toxicity or potency of the many isolates of the several subspecies.

Bacillus thuringiensis subspecies *israelensis* was isolated from a mosquito breeding site in Israel [33]. The discovery of this subspecies provided researchers with an isolate with consistently high toxicity against larvae of mosquitoes and blackflies (order Diptera). All other subspecies are predominantly toxic for larvae of Lepidoptera. This subspecies also contains a parasporal crystal protein. However, this crystal is irregular in size and instead of a single crystal, there are 3 – 5 crystals. The pathology caused by subspecies *israelensis* is similar to that of other subspecies of *B. thuringiensis* in susceptible Lepidoptera. The pathology in black flies is most likely very similar. *Bacillus thuringiensis* subspecies *israelensis* is also produced by submerged culture.

β-Exotoxin [*23526-02-5*], the other toxin produced by *B. thuringiensis,* is being emphasized as an insecticide.

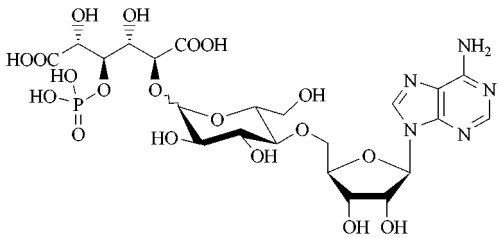

β-Exotoxin

β-Exotoxin is a nucleotide produced during the vegetative stage of bacterial growth [34]. It is composed of adenine, ribose, glucose, and allaric acid with a phosphate group. β-Exotoxin is heat stable and is produced by subspecies *thuringiensis, kenyae, morrisoni, tolworthi,* and *darmstadiensis* [35]. These subspecies produce different quantities of exotoxin, depending on the culture medium. Also, indications are that some subspecies produce more than one exotoxin. Commercial production of β-exotoxin is carried out under patent US 3758383 [36].

Bacillus popilliae causes a "milky disease" in some beetles in the family Scarabaeidae, but most noted in the Japanese beetle, *Popillia japonica. Bacillus popilliae* is eaten by the grub; the spore germinates and penetrates the gut wall. Once in the hemocoel, the bacterium readily multiplies (septicemia), giving the grub a whitish appearance visible through the cuticle, thus the name milky disease.

Bacillus popilliae readily develops in larvae of the beetle, but microbiological techniques have not been developed to induce *B. popilliae* to sporulate in vitro; thus, all production is carried out by inoculating Japanese beetle larvae and harvesting the infected larvae. This technique has a drawback because the Japanese beetle is not easily grown in the laboratory.

Bacillus sphaericus is a spore-forming rod that is readily found in nature. It has been isolated from mosquito larvae, and much research has been conducted on several strains of this bacterium that are insecticidal to several species of mosquitoes. *Bacillus sphaericus* does not produce a parasporal body or β-exotoxin as does *B. thuringiensis.* However, there is a proteinaceous toxin located in the spore coat. Susceptible mosquito larvae consume the spore and partly digest the spore wall, releasing a toxin. This toxin causes feeding inhibition, swelling of the midgut cells, body tremors, and then death. Death can occur as soon as 0.5 h after consumption of the spores. Like *B. thuringiensis, B. sphaericus* is produced in submerged culture.

2.2.2. Pathogens with Protozoal Characteristics

Microspora. The phylum Protozoa has been split into seven distinct phyla [37]. This phylum contains the order Microsporida, intercellular parasites found essentially in all orders of insects, with new species being described and host records being published frequently. Pathologically, microsporidia differ greatly from other disease-causing organisms in that they generally do not cause immediate death of the host. A microsporidian infection usually is chronic. There are no specific symptoms to describe insects infected with a microsporidium. The infected insect usually is lethargic, has reduced feeding, is small, has morphological deformities, and fails to molt or pupate.

Although microsporidia are widely distributed in the insect world, most microsporidia infect a single, or at most, a very few hosts. *Nosema pyrausta* is a microsporidium of this type and is a classical biological control agent. *Nosema pyrausta* predominantly infects larvae of the European corn borer, *Ostrinia nubilalis.* A schematic illustration of the relationship between *N. pyrausta* and *O. nubilalis* is presented in Figure 7. *Nosema pyrausta* has a life cycle typical of the order Microsporida. A corn borer larva eats a spore (the resting stage *N. pyrausta*); the spore extrudes its polar filament, injecting a sporoplasm into a midgut cell. Some development occurs in the midgut, whereas other sporoplasms migrate through the midgut into the hemocoel and eventually infect other tissues. In the European corn borer, infections predominantly take place in the Malpighian tubules and in the reproductive tissues of the female insect. At this point, two routes are possible for the microsporidium (both routes occur in all cases of infection). The infection can develop in the midgut, causing these cells and spores to slough into the lumen and pass from the body in the fecal material. The fecal material then becomes a reservoir of *N. pyrausta* spores available to infect other corn borer larvae feeding in the same plant. This is known as *horizontal transmission*, i.e., transmission within a generation. This route readily occurs in nature and is an effective means of disease transmission [38]. If the insect in which the reproductive tissues became infected was a female, the spores either are dis-

seminated within the developing insect egg and infect the embryo during development (transovarial transmission), or they are passed on the egg shell and are consumed by the larvae at eclosion (transovum transmission). Transovarial and transovum transmission are collectively known as *vertical transmission*, i.e., transmission to a subsequent generation. Horizontal and vertical transmission are two very effective means of transmitting and maintaining an organism within an ecosystem.

The impact of *N. pyrausta* occurs in several ways. If the infection is intense enough (a tremendous number of spores per insect larva), the infected insect will die. If the insect develops to the adult with this infection, the impact of *N. pyrausta* is elicited by a shortened life span, reduction in the number of eggs laid, and a reduction in the number of eggs hatching and developing to maturity.

Vairimorpha necatrix is a microsporidium that infects a great number of insects, all being phytophagous Lepidoptera. This microsporidium also can cause an acute pathology, resulting in death of the susceptible larvae within several hours, and is an example of an applied biological control agent. If a newly hatched or relatively early stage larva consumes an excessive number of spores, the midgut cells are damaged, allowing entry of the midgut contents into the hemocoel, and the larva dies from bacterial septicemia. If the larva does not die from septicemia, the spore enters the hemocoel, and an acute infection of the fat body occurs. This infection becomes very intense, and the host usually dies before pupation or during the pupal period. The European corn borer is not as susceptible to *V. necatrix* as are several larvae in the family Noctuidae (cutworms, armyworms, etc.), and a few insects develop to adults. However, because of the intense infections at this time, the production of offspring is unlikely. The acute pathology caused by *V. necatrix* makes it improbable that any horizontal or vertical transmission will occur. If transmission does occur, it is likely by an insect larva feeding on infected cadavers, by cannibalism, or by insects feeding on foliage contaminated by infected insects that ruptured after death (horizontal transmission) [39].

Apicomplexa. The gregarines and some coccidia with potential as biological control agents are representatives of this phylum. An infection usually occurs by consumption of spores. The development cycle involves schizogeny, which increases the number of spores tremendously, destroying tissues and depleting energy reserves. This cycle takes 1 – 6 weeks, depending on the specific organism. *Mattesia grandis,* a pathogen of the boll weevil, *Anthonomus grandis,* is an example of this group and contributes to overwintering mortality of this insect. *Mattesia trogodermae,* a pathogen of a stored-grain beetle, *Trogoderma glabrum,* has great potential for controlling these beetles in warehouses. Adult males can be contaminated with spores and can transmit these spores to females during copulation [40].

Commercial application of pathogens with protozoal characteristics is limited because these organisms must be produced in vivo, in their respective hosts.

2.2.3. Insect Viruses

There are two general types of insect viruses, the occluded and nonoccluded viruses. The *occluded*-type viruses are characterized by having the infective units, the virions, enclosed by a protein capsule (occlusion body). This is the most emphasized group relative to biological control and includes the nuclear polyhedrosis viruses (NPV), the cytoplasmic polyhedrosis viruses (CPV), the granulosis viruses (GV), and the entomopox viruses or insect pox viruses. The iridescent viruses, parvoviruses and picornaviruses are representatives of the *nonoccluded* viruses.

The occlusion body of an NPV is many-sided and encloses several virions. The virions are rod-shaped, and replication occurs within the cell nucleus. A generalized cycle, using a nuclear polyhedrosis virus, is illustrated in Figure 8. The inclusion bodies (occlusion bodies) are consumed by the insect and the proteinaceous capsule is hydrolyzed in the alkaline gut, releasing the virions. The virions enter the midgut cell nucleus, at which point they replicate and destroy the cell, or else the virions pass through the cell into the hemocoel and infect cell nuclei of other tissues, replicate, encapsulate, and rupture the

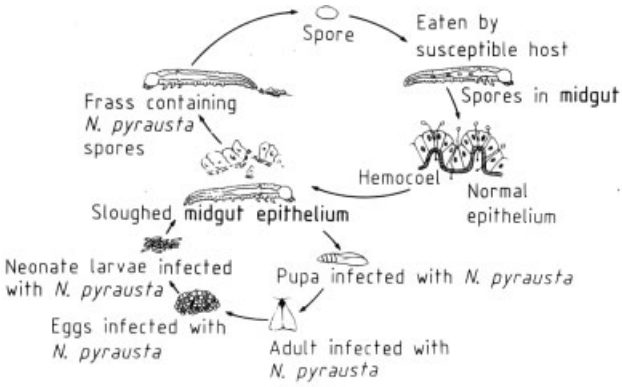

Figure 7. Infection cycles of the microsporidium *Nosema pyrausta*

cell. Insects infected with NPV exhibit sluggish retarded growth and a behavioral change characterized by moving to the top of a host plant where they die, and the integument usually ruptures. The primary tissues infected with a nuclear polyhedrosis virus are the midgut, tracheal matrix, fat body, and hypodermis. A viral infection in these tissues with eventual rupturing of the cells gives rise to a classical viral death where the internal organs have liquefied, and the cuticle of the insect is a mere "bag" of viral occlusion bodies. Once this "bag" breaks, the occlusion bodies contaminate the surface to which the insect was attached and become available for consumption by other susceptible insects.

Insects infected with a CPV display discoloration of the midgut (visible through the integument), experience retarded growth, and eventually die. The integument does not rupture. The shape of the occlusion body varies, and several virions, icosahedral in shape, are within each occlusion body. The infection occurs within the cytoplasm of the cell.

A GV infection is relatively nondescript but is generally characterized by retarded growth of the insect and a paling of color of the integument. In some infections, the occlusion bodies are passed into the hemolymph, giving a milky color. The occlusion bodies of this virus are formed within the cell nucleus and usually contain a single rod-shaped virion. Insects infected with a pox virus are sluggish and have an extended developmental stage. The occlusion bodies develop predominantly in the fat body, both in the nucleus and in the cytoplasm of the cell. The occlusion body contains many ovoid infective units.

Insect viruses were believed to be host specific (only infecting a single species of insect), but research has revealed that some nuclear polyhedrosis viruses will infect more than one species. For example, the nuclear polyhedrosis virus from the alfalfa looper, *Autographa californica,* infects several alternate hosts [41], including some from which a virus has never been isolated, i.e., the European corn borer and the black cutworm [42], [43]. This aspect of a wider host range for a virus increases the feasibility of using viruses in applied biological control. Insect viruses can be produced only in respective hosts.

2.2.4. Fungi

Most entomopathogenic fungi belong to the class Deuteromycetes (Fungi Imperfecti), with some very important entomopathogenic fungi in other classes. Insects infected with a fungus exhibit general lethargy, slowed growth, cessation of feeding, and changes in coloration of the integument. The difference between fungi and other pathogens is that the former do not have to be eaten by the insect to cause disease, but instead grow through the insect's skin.

A general infection cycle for fungi is illustrated with *Beauveria bassiana* as an example (Fig. 9). The resting stage or conidium comes in contact with the insect cuticle. The conidium germinates and enzymatic action partly digests the cuticle, allowing the hyphae to penetrate the

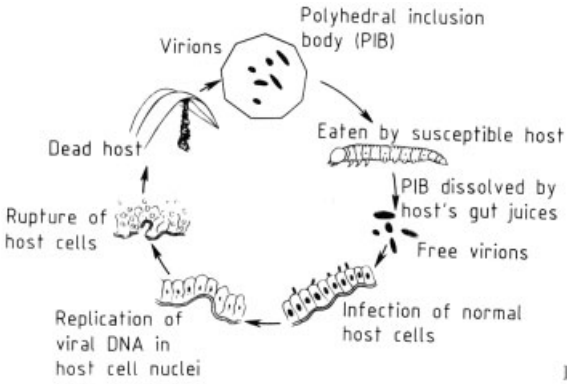

Figure 8. Infection cycle of a nuclear polyhedrosis virus

cuticle. The hyphae develop, forming a network of mycelia within the body cavity. Death usually occurs after this mycelial growth has spread throughout the body cavity. At this time, the mycelia grow out of the body, forming conidiophores and conidia. These formations are unique and definitive in some fungi. For example, *B. bassiana* conidia cover the infected insect, transforming it into a white covered mummy. This fungus is referred to as the white muscardine. *Metarrhizium anisopliae* turns an insect into a green-covered mummy and is called the green muscardine fungus. Both these fungi infect a wide range of insects. *Nomuraea rileyi* also coats insects with a green covering but is not as "fuzzy" as *M. anisopliae.* The conidia of these fungi become airborne and eventually infect other insects.

Some insects exhibit a typical behavior pattern once infected. Diptera (flies) infected with *Entomophthora musca* climb to the top of vegetation, wrap their legs around a grass stem, for example, and die; then mycelia grow on the external surface. A similar response occurs in grasshoppers infected with *Entomophaga grylli.* These fungi forcibly discharge the conidia from the conidiophores. Both these fungi are members of the class Zygomycetes. The class Oomycetes contains the species *Lagenidium giganteum,* a pathogen of several species of mosquitoes.

Entomopathogenic fungi are omnipresent; however, their effectiveness as a biological control agent is dependent on temperature, moisture, and wind. Each species has certain requirements relative to these environmental variables that must be met for the fungus to be effective. Fungi can be produced on submerged culture or on agar plates.

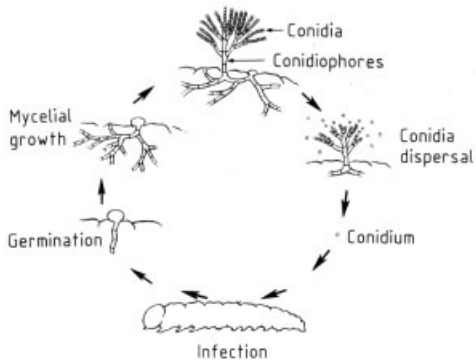

Figure 9. Infection cycle of the fungus, *Beauveria bassiana*

2.3. Nematodes

Nematodes (phylum Nematoda) are elongated cylindrical worms that are parasitic in animals or plants, or are free-living in soil or water. Nematodes are of importance in the biological control of insects. Several families of nematodes have members parasitic to insects. Nine families show potential as biological control agents [14]. In this article the families Mermithidae and Steinernematidae are be mentioned as examples of insect parasitic forms.

The **mermithids** have both aquatic and terrestrial species. *Romanomermis culicivorax* is an aquatic species that is parasitic to several

species of mosquitoes. *Romanomermis culicivorax* hatches from an egg and actively seeks a mosquito larva and penetrates the larval integument by use of a stylet and enzymatic secretions. The nematode completes its development within the insect hemocoel and then leaves the host by boring through the integument. At this time, the host usually dies. Once in the aquatic environment, the nematode develops to adulthood and mates. The female deposits its eggs to continue the cycle.

Mermis nigrescens, a representative of a terrestrial mermithid, is mainly a parasite of grasshoppers, but will occasionally parasitize other insects, including beetles. Adult *M. nigrescens* live in the soil. The female climbs vegetation, lays her eggs, and returns to the soil. An insect eats an egg. Once in the gut, the egg hatches, and the nematode penetrates the midgut, enters the hemocoel, develops within for several weeks, and then leaves the host (the host dies at this time); the nematode enters the soil and develops to adulthood. Mating may or may not occur. Reproduction can take place parthenogenetically.

The **steinernematids** are represented by *Neoaplectana carpocapsae,* which parasitizes terrestrial insects within several orders. Infective nematode larvae actively enter a host through the mouth, the anus, or through a spiracle, and proceed to the hemocoel, where they begin development. The steinernematids have a symbiotic relationship with a bacterium. As the *N. carpocapsae* larvae develop, a bacterium, *Achromobacter nematophilus,* is released through the anus. This bacterium creates a desirable environment in which *N. carpocapsae* develops. The nematodes develop to maturity and mate, and the female deposits her eggs. These eggs hatch, and the larvae develop to infective stage and then leave the host, ready to start the cycle again.

Nematodes exist in nature. However, some nematodes, including both the aforementioned ones, are produced commercially and, therefore, are available as applied biological control agents. Methods have been developed for in vitro production of nematodes. A high-protein source free from bacterial contamination is a necessity [14].

3. Biological Control of Weeds

The biological control of weeds theoretically involves the use of any organism that uses the weed as a food source or as a host. In general, most successful programs in biological control of weeds involve the use of phytophagous insects or insects that feed on these weeds. There are several programs throughout the world that have been successful in controlling weeds. For illustrative purposes, a few of the more prominent ones will be emphasized.

3.1. Phytophagous Insects

Control of the **prickly pear cactus** in Australia in the 1920s was one of the earliest successes. Larvae of the moth *Cactoblastis cactorum* were imported from Argentina and destroyed populations of this group of cacti. In the United States, these cacti have been substantially controlled by a cochineal insect, *Dactylopius opuntiae,* imported from Hawaii.

The **Klamath weed, Hypericum perforatum,** was introduced to northwestern North America in the early 1900s and has been successfully suppressed by two beetles, *Chrysolina hyperici* and *C. quadrigemina,* which were imported from Europe via Australia.

Musk thistle, a weed in range and pasture land in many areas, is host to a weevil, *Rhinocyllus conicus,* that lays its eggs in the seed head of the thistle. *Rhinocyllus conicus* was introduced into western Canada and into the United States (Montana, Nebraska, and more recently Iowa) and has a definite negative impact on this weed.

These three plants are examples of control of terrestrial weeds. Successful control of an aquatic plant, the alligator weed *Alternanthera philoxeroides,* has been demonstrated in inland waterways in Florida. Control has been obtained with a beetle, *Agasicles hygrophila,* and a moth larva, *Vogtia malloi.*

3.2. Plant Pathogens

The use of plant pathogens for control of weeds is a relatively new approach. Most of the emphasis has been with plant pathogenic fungi. Basic

research is performed under carefully controlled conditions to assure the specificity of pathogens. *Phytopthora palmivora* is a fungus that is specific for the **milkweed vine.** Some of these plant pathogenic fungi maintain themselves in the soil and thus are excellent biological control agents. *Alternaria cassiae* is a fungus that shows potential for control of **sicklepod.** Several other fungi are being extensively researched for biological control of weeds. This discipline will have a tremendous impact on weed control in the future.

3.3. Higher Animals

Tansy ragwort, *Senecio jacobaea,* is a biennial weed containing pyrrolizidine alkaloids. These alkaloids cause liver damage in livestock that consume these weeds. However, sheep are seemingly immune to this alkaloid. Therefore, they can graze this plant and keep it from producing seeds, thus reducing the plant population [44].

Biological control of weeds is a very active area of research. More and more phytophagous insects and plant pathogens are being evaluated as potential biological control agents. Once the basic research has been completed, several of these organisms will most likely be used to aid in suppression of weeds.

When selecting weedy plants as potential targets for biological control, one must be certain that the weed is classified as a weed in all areas of the country where it occurs. In some parts of a country, a certain plant might be a weed. However, in other locations, the plant might be beneficial as food source, habitat for wildlife, or for ground cover. The host range of phytophagous insects and plant pathogens also must be extensively researched to be absolutely sure that desirable plants that are closely related and might become an alternative host, do not exist in the country of introduction.

4. Biological Control of Plant Pathogens

Plant pathogens are a group of pests that cause tremendous loss of plant life every year and are primarily suppressed or controlled by chemical pesticides. However, there are biological control agents that use these microorganisms as sources of food or as hosts.

4.1. Macroorganisms

Fungi (smuts, rusts, and mildews) are food sources for several insects, but applied biocontrol of plant pathogens by insects has not been encouraging.

4.2. Microorganisms

The complex interrelationships between plant pathogens and microorganisms that have the potential of controlling them have been studied. Cucumber powdery mildew has been experimentally controlled by application of the fungus *Ampelomyces quisqualis* [45]. *Ampelomyces quisqualis* is a hyperparasite that obtains nutrients from the mildew. The fungus *Trichoderma harzianum* has been shown to degrade plant pathogens, such as *Schlerotium rolfsii, Rhizoctonia solani,* and *Pythium aphanidermatum.* Plant nematodes, *Criconemella xenoplax,* have been suppressed by the fungus *Hirsutella rhossiliensis.* Here again, researchers are faced with a very complicated biological system in which a tremendous amount of basic research must be conducted before biological control of plant pathogens can be realized to its full potential.

5. Other Biological Methods of Control

5.1. Biochemicals

Interfering with an insect's growth to make the insect vulnerable to natural control measures and maintaining an insect in a "nonpest" developmental stage have long been intriguing ideas for insect control. During normal insect development, immature stages are regulated by hormones (growth regulators). During the adult stage, insect behavior is influenced by pheromones and kairomones (semiochemicals).

5.1.1. Growth Regulators

Ecdysone is secreted by the prothoracic glands and initiates cell growth. Once released into

the hemolymph, ecdysone causes the insect to molt. α-Ecdysone [*3604-87-3*], $C_{27}H_{44}O_6$, M_r 464, *mp* 237 – 239 °C, is a water-soluble crystalline compound; IR (KBr, cm^{-1}) 3333, 1657; UV (MeOH) 242 nm (ε 12400); NMR (C_5D_5N) [46]. It was originally isolated from silkworm

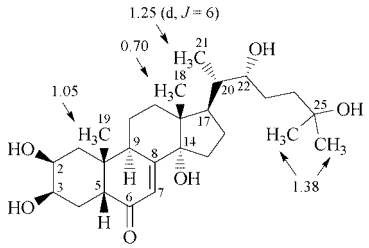

α-Ecdysone

pupae. β-Ecdysone [*5289-74-7*], 20-hydroxy-α-ecdysone, is a slightlymore biologically active compound. Both α- and β-ecdysone are biosynthesized from cholesterol [47]. Phytoecdysones, steroids with molting hormone activity, have been isolated from several plants. The isolation and synthesis of phytoecdysones and α- and β-ecdysone have been described in detail [47]. Exogenous ecdysones will initiate molting in insects when applied to the surface of the cuticle.

Juvenile hormones (JHs) are synthesized in and secreted by the corpus allatum. These hormones stimulate larval development while preventing development of adult characteristics. They also promote synthesis of yolk proteins and accumulation of these proteins by the developing oocytes. Juvenile hormones control the development of imaginal discs. Application of JH to a young embryo will block its development to a larva; however, JH may be important in normal embryonic development of some insects.

Five juvenile hormones, JH 0 [*55333-14-7*], iso-JH 0, JHsI [*13804-51-8*], JH II [*34218-61-6*], and JH III [*22963-93-5*], have been isolated from insects. The JH I, II, and III are liquids, JH 0 and iso-JH 0 are presumed to be liquids; JH I *bp* 130 – 145 °C; IR (cm^{-1}) 1715, 1640, 1230, 1150, 850, 1730, 1655, 1722, 1650.

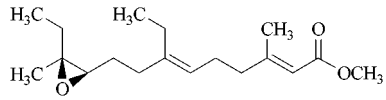

Juvenile hormone I

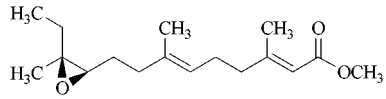

Juvenile hormone II

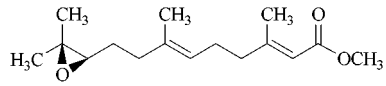

Juvenile hormone III

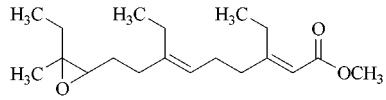

Juvenile hormone 0

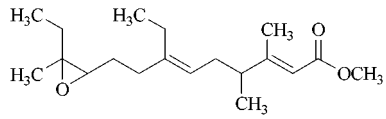

iso-Juvenile hormone 0

Juvabione [*17904-27-7*], a juvenoid (compound with JH activity), was isolated from the balsam fir tree [48]; MS 266.188 (M^+ $C_6H_{26}O_3$); UV 223 nm; IR (cm^{-1}) 1722 (α, β-unsaturated ester), 1645 (saturated ketone).

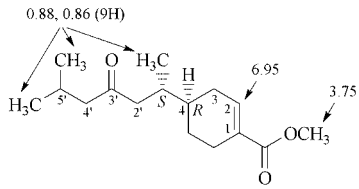

Juvabione

Several other juvenoids have been isolated from plants [49]. The determination of the structure of the naturally occurring JHs led to the synthesis of JH-analogs. Syntheses of JHs, juvenoids, and JH-analogs have been reviewed [50], [51]. Commonly used synthetic JH-analogs are: methoprene [*40596-69-8*], Altosid, ZR 515, *mp* 135 – 136 °C, d^{20} 0.926; hydroprene

[*41096-46-2*], Altozar, ZR 512; epofenonane [*57342-02-6*], RO 10-3108. The applied use and chemistry of these and other JH analogs have been thoroughly discussed [52].

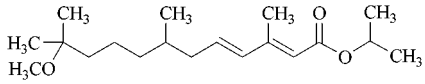

Methoprene

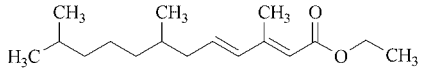

Hydroprene

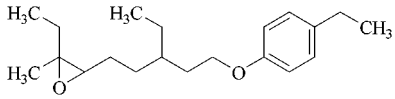

Epofenonane

Two compounds, precocene 1 [*17598-02-6*] and precocene 2 [*644-06-4*], isolated from the common bedding plant, *Ageratum houstonianum,* have anti-JH activity against several hemipteran species [53]. Precocenes destroy the corpora allata of insects in the orders Hemiptera, Homoptera, and Orthoptera. There is no anti-JH activity against holometabulous insects; however, certain ones have been forced into diapause, have been sterilized, or both. Synthesis of precocenes has been described [54].

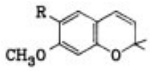

Precocene 1, R = H
Precocene 2, R = OCH_3

Leaves of the neem tree, *Azadirachta indica,* yield chemicals that are strong antifeedants [55], [56]; however, these chemicals have not led to any significant use for pest control. The main component of the neem fractions is azadirachtin [*70357-01-6*], $C_{35}H_{44}O_{16}$, a compound that acts as a phagorepellent and as a growth regulator; UV(MeOH) 217 nm (ε 9100); CD (MeOH), no extrema above 220 nm; IR (KBr, cm^{-1}) 3410, (OH) 1745, 1720, 1700 (esters), 1645, 1620, 843 (double bonds); IR (chf. concentration variation) 3580 cm^{-1} (free OH), 3465 (intramolecular bonded OH), 3380 (intermolecular H bonded OH).

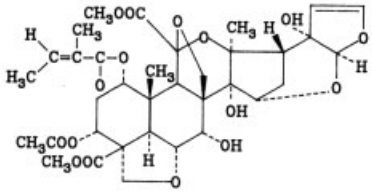

Azadirachtin

5.1.2. Semiochemicals

Pheromones. A pheromone is a substance that is secreted by an organism to the outside and mediates a response in an organism of the same species. Sex pheromones are specific chemicals used in the mating process. Generally, they are emitted by the female insect and attract the male insect. Chemical structures of pheromones are scattered among many types of volatile compounds ranging from alkanes to nitrogen heterocycles. Earlier research usually reported one chemical as the pheromone of an insect. Synthesis of these compounds and field testing for attractancy has resulted in various degrees of success. However, although a single compound is the major component, there are trace amounts of additional chemicals that make up the full complement of the pheromone [57–59].

Using open-tubular capillary chromatography analysis, the following composition was determined for the pheromone of the lesser cornstalk borer, *Elasmopalpus lignosellus*: 16.5 % (*Z*)-7-tetradecen-1-ol acetate, 8.8 % (*Z*)-9-tetradecen-1-ol acetate, 3.8 % 1-tetradecanol acetate (C_{14}), 28.2 % (*Z*)-9-hexadecen-1-ol acetate, 17.4 % (*Z*)-11-hexadecen-1-ol acetate, 9.6 % 1-hexadecanol acetate (C_{16}), 2.1 % (*Z*)-7-tetradecen-1-ol, 2.1 % (*Z*)-9-tetradecen-1-ol, 7.0 % (*Z*)-9-hexadecen-1-ol, and 4.4 % (*Z*)-11-hexadecen-1-ol [59]. The organic synthesis of sex pheromones has been reviewed [50], [60], [61]. Biosynthesis of pheromones takes place in a glandular area of the abdominal tip of the adult female and is regulated by a hormonal factor produced in the brain [62].

Several pheromones have been synthesized and these synthetic pheromones are available for certain pests of agricultural crops [63].

Kairomones. Kairomones are compounds given off by one species that induce a greater behavioral response in another species than in the species exuding the compound. Kairomones include host-seeking chemicals, e.g., compounds that aid parasitoids in search of suitable hosts. Host-seeking compounds, extracted from the wings (scales) of *Heliothis zea* and *Cadra cautella,* increase parasitization of eggs by *Trichogramma evanescens.* These compounds stimulate the parasitoid to search for suitable host eggs.

Some plants produce substances that attract insects, e.g., cucurbitacins. Cucurbitacins are oxygenated tetracyclic triterpenes produced by plants of the family Cucurbitaceae, i.e., Cucurbitacin B [*6199-67-3*]. A complete discussion of the chemistry of these compoundsis available [64].

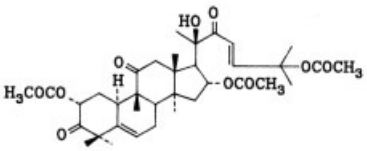

Cucurbitacin B acetate

The fruits of wild species of *Cucurbita* contain amounts of these chemicals that are toxic to animals. Cucurbitacins are also present in extremely small quantities, < 0.02 mg/g, in domesticated fruits of *C. ficifolia, C. maxima, C. mixtra, C. moschata,* and *C. pepo* [65]. This small amount, however, acts as an arrestant (kairomone) and feeding stimulant to several beetles in the genus *Diabrotica* [66].

5.1.3. Bioderived Toxicants

Some organisms produce chemicals that are mere poisons without any novel mode of action, such as an attractant and/or growth regulator. The avermectins are a class of compounds described as pentacyclic lactones isolated from fermentation of the actinomycete *Streptomyces avermitilis.* These natural products contain an α-L-oleandrosyl-α-L-oleandrosyl disaccharide attached to the macrocyclic lactone ring through the allylic C_{13}-hydroxy group [67]. These compounds have demonstrated toxicity in laboratory and field tests against several insect pests, phytophagous mites, and a plant parasitic nematode.

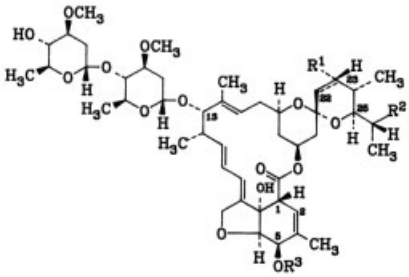

Avermectin	R^1	R^2	R^3	CAS-Reg. No.
A_1a		C_2H_5	CH_3	[*65195-51-9*]
A_1b		CH_3	CH_3	[*65195-52-0*]
A_2a	OH	C_2H_5	CH_3	[*65195-53-1*]
A_2b	OH	CH_3	CH_3	[*65195-54-2*]
B_1a		C_2H_5	H	[*65195-55-3*]
B_1b		CH_3	H	[*65195-56-4*]
B_2a	OH	C_2H_5	H	[*65195-57-5*]
B_2b	OH	CH_3	H	[*65195-58-6*]

Where R^1 is absent, the double bond (=) is present. Both sugars are α-L-olendrose.

5.2. Autocidal

Autocidal methods involve alteration of genetic material of insects, thereby reducing the insect's ability to reproduce. Briefly, after the genetic material of an insect population is altered, this population is released into nature. Mating of the genetically abnormal population with the wild population causes reduction in the natural population.

5.2.1. Sterile Male

The primary autocidal method is the sterile male technique in which tremendous numbers of male insects are sterilized either by radiation or by chemical means, and released into natural populations in such numbers that the sterilized males greatly outnumber the naturally occurring ones. This technique has been employed with success against the screwworm, *Cochliomyia hominivorax,* Mediterranean fruit fly, *Ceratitis capitata,* melon fly, *Dacus cucurbitae,* Mexican fruit fly, *Anastrepha ludens,* and pink bollworm, *Pectinophora gossypiella.*

5.2.2. Hybrid Sterility

Hybrid sterility involves the crossing of two closely related but different species of insects with production of sterile progeny and subsequent release in natural populations. These hybrids must be fully viable and able to compete with feral adults. Other genetic alterations are: (1) cytoplasmic incompatibility – certain strains within a species when crossed produce either sterile offspring, or sterile and fertile offspring, or sterile males; (2) conditional lethals – certain alleles of genes controlling adaptability produce normal effects at certain temperature ranges but produce lethal effects at higher or lower temperatures; (3) growth alterations – for example, introducing a gene for "nondiapause" into a strain of insects that must diapause in order to survive the winter.

6. Formulation, Application, and Efficacy

Insect Control. Formulation and application of biological control agents are a "hand-in-glove" type of relationship. The proposed use of an agent dictates to a great extent the way in which that agent must be formulated to elicit its greatest possible impact. Formulation must be done with the recognition that biological control agents are living organisms which vary in their susceptibilities to adverse environmental conditions.

The formulation techniques for parasitoids and predators are limited. Basically, these insects are colonized, produced in large numbers in an insectary, and then released into the environment at a time predetermined to be most advantagous to the beneficial insect.

A special case of formulation and application of parasitoids is employed when *Trichogramma* are used in applied biological control. Briefly, *Sitotroga cereallella* or *Anagasta kuehniella* adults are constrained to lay many eggs on sheets of cardboard; these eggs are parasitized by *Trichogramma* adults, and the cards of parasitized eggs are placed in an ecosystem containing recently oviposited eggs of a pest insect. This procedure is timed so that the *Trichogramma* will emerge within a few hours after being placed in the environment. *Trichogramma maidis* is utilized in eastern France to control populations of the European corn borer in corn. Researchers obtained 90 – 95 % control by placing 300000 parasitized *Anagasta kuehniella* eggs/ha [68] in corn fields infested with the European corn borer. Techniques also have been developed to apply *Trichogramma* with a crop-dusting plane. The host eggs are sometimes oviposited on a carrier, such as wheat bran, and this admixture is air – dropped over an insect-infested crop area.

Pediobius foveolatus, a parasitoid of the Mexican bean beetle, was released annually in Maryland. Approximately 490000 parasitoids were released over 94000 ha of soybeans, resulting in up to 84 % parasitism of the Mexican bean beetle [69].

Excellent control of aphids on potatoes was obtained by releasing a lacewing, *Chrysopa* spp., and a coccinellid, *Coccinella septempunctata,* at the same time. These researchers released 210300 and 74600/ha of *Chrysopa* and *C. septempunctata,* respectively [70].

Pathogens are formulated in the same way that chemical insecticides are, i.e., as dusts, wettable powders, flowable concentrates, granules, or baits. An insect-pathogen dust or granule is made by mixing the pathogen with an inert material called a diluent. Diluents most often used are talcs, prophyllite clays, clays, calcium carbonate, precipitated hydrated silicates, silicon dioxides, synthetic calcium silicate, and diatomaceous earth. Potato starch, cornstarch, wheat flour, wheat bran, ground corn cobs, and ground corn have been used as experimental diluents. The basic differences between a dust and a granule is the mere size of the diluent.

Wettable powders are formulations of pathogens containing sufficient wetting agent that the pathogen will be wetted and suspendible in water. Powders can be any of the named diluents, only finely ground, so that a sprayable end product is obtained. A flowable concentrate is essentially a wettable powder suspended in an oil or liquid base of some type that can readily be mixed with water to form a stable suspension.

Baits usually are botanical products, such as apple pomace, grape pomace, citrus pulp, wheat bran, etc., that are food sources and attract the target insect.

In all these formulations, great care is taken to assure shelf life, stability in the field, and mainte-

nance of viability of the pathogen. Commercial additives are available to protect the pathogen from severe drying or wetting and from ultraviolet light.

Wettable powder and flowable concentrate formulations are employed when complete coverage of an area or vegetation is necessary for maximum efficiency. The use of *Bacillus thuringiensis* subspecies *israelensis* for mosquito and black fly control is a situation in which these formulations are used.

Two wettable powder formulations of *B. thuringiensis* subspecies *israelensis,* Bactimos (Biochem Products, USA) and Vetobac (Abbott Laboratories, USA), and a flowable concentrate, Teknar (Sandoz, USA), were tested against mosquito populations in central Italy [71]. Bactimos and Teknar were applied at 0.5 kg/ha and 2.5 L/ha, respectively, to control the mosquito *Aedes caspius* in a salt marsh. These applications gave complete control of the larvae. Vectobac at 0.5 kg/ha gave a range of 67 – 91 % control. A flowable concentrate formulation of Teknar was prepared to give a concentration of 10 ppm and applied to the width of a fresh water stream for 1 min with complete elimination of the black fly larvae [72].

Aqueous suspensions of insect viruses are prepared from wettable powder formulations. Viruses are quantified as polyhedral inclusion bodies (PIB) per unit mass. The European pine sawfly, *Neodiprion sertifer,* can be controlled on *Pinus resinosa* with an application of 2.5×10^9 PIB/ha [73]. Application rates of this magnitude have been used to control other forest pests [74] and pests of other agricultural crops, i.e., cole crops, cotton, and soybeans [75]. Granular formulations are used in an ecosystem for placement of a pathogen on a plant or within a certain area, or for incorporation into soil. For example, larvae of the European corn borer feed on a corn plant at the interface where the developing leaves are unfolding. Because of the morphology of the plant, there is a natural funnel that will direct granules of a pathogen to this area and thus make them immediately available for consumption by the insect. Not only is the pathogen more readily available to the insect, but also it is protected from ultraviolet radiation.

An 11.25 wt % formulation of Thuricide HPC on clay granules was applied at 1.12 kg of *B. thuringiensis* HPC/ha to control the European corn borer on corn [76]. This application caused a significant reduction in damage by the borer compared to the untreated check. Several mosquito habitats are in areas of heavy vegetation. A granular formulation, when applied in this type of environment, will pass through the vegetation, enter the water, and become available to the mosquito larvae. Granular formulations of *Bacillus popilliae* are incorporated into soil for control of Japanese beetles in Russia [77]. A specialized formulation and application is the impregnation of foundation wax with *Bacillus thuringiensis* subspecies *galleriae* for control of greater wax moth larvae in beehives [78]. *Nosema locustae* is formulated as a bait on wheat bran and applied to rangeland to control grasshoppers at 2.5×10^9 *N. locustae* spores on 2 kg of wheat bran/ha [79]. This application reduced the population of immature grasshoppers by 50 %. A similar treatment effectively reduced populations of the Mormon cricket, *Anabrus simplex* [80]. Nematodes, *Neoaplectana carpocapsae,* applied as an aqueous suspension ($\geqq$ 1000 nematodes/mL) were highly effective against larvae of the artichoke plum moth, *Platytilia cardiudactyla* [81]. The neoaplectanid nematodes show potential for control of several insect pests [82]. *Romanomermis culicivorax,* a nematode parasite of mosquitoes, has been tested worldwide for mosquito control. Different stages of the nematode, egg, pre-, and postparasitic, are applied to the aqueous environment at rates of a few hundred to 200000 units/m^2 with varying degrees of success. A wettable powder or dust formulation of *Hirsutella thompsoni,* a myoacaricide, was effective in stimulating premature fungal epizootics in citrus rust mite populations [83].

The environs of a greenhouse create special pest problems but also present a controllable situation for the use of biological control agents. The successes, problems, and failures of biological control in such ecosystems have been reviewed [84], [85]. In a greenhouse-like situation, the spotted spider mite, *Tetranychus urticae,* was controlled in a plastic walk-in tunnel by introducing one predatory mite per strawberry plant [86]. Control of the greenhouse whitefly, *Trialeuroes vaporariorum*, is difficult. A combination of attracting the whitefly to yellow sticky traps and using the parasitoid *Encarsia formosa* has resulted in 90 % parasitization of this pest on

spring tomatoes in the greenhouse [87]. In other research [88], several species of parasitoids with the potential of controlling the chrysanthemum leaf miner have been identified.

The use of fishes to control insects is primarily focused on controlling the larvae of mosquitoes that transmit human disease. *Gambusia affinis,* a larvivorous fish, is a native of the southeastern United States but has spread to most parts of the world where it can survive and give excellent mosquito control. *Gambusia affinis* seeded at 2500 – 5000 fish/ha will reproduce quickly and provide complete mosquito control in 6 – 8 weeks [89], [90].

Spiders definitely have an impact on control of insects; however, quantification is difficult [91].

Weed and Plant Pathogen Control. Mycoherbicides are also formulated as an aqueous suspension and sprayed on foliage for control of weeds. Spores of *Colletotrichum coccodes,* a fungus, were sprayed on velvetleaf, *Abutilon theophrasti,* at 10^7 and 10^9 spores/m^2 in 467.5, 935, and 1870 L of H_2O/ha. These concentrations of spores reduced the velvetleaf biomass by 40 and 50 % at the 10^7 and 10^9 treatment levels, respectively [92]. Progress has been made using a rust, *Puccinia chondrilla,* to suppress rush skeletonweed, *Chondrilla juncea* [93]. Plant size, number of flowers, and seed viability have been reduced, but there has been no substantial reduction in plant populations.

Damage caused by the root-knot nematode *Meliodogyne javanica* has been reduced using biological control. Tomato roots containing the nematode infected with *Bacillus penetrans* were air-dried, finely ground, and incorporated into the soil at 212 – 600 mg/kg, reducing the number of nematodes and the damage caused by the nematodes [94]. Organic extracts from the fungus *Aphanocladium album* induce teliospore formation on rust fungi. *Puccinia graminis, P. dispersa, P. striiformis,* and *P. sorghi* formed premature telia on plants treated with the extract [95]. The biology of introduced predators and parasitoids of arthropod and weed pests is available [96].

Other Methods of Biological Control. Growth regulators have been tested for insect suppression. Lethal effects were induced in a population of western spruce budworms at dosages comparable to several conventional insecticides [97]. Treating male or female adults of the western spruce budworm with ZR575, ZR512, or RO 10-3108 reduced the reproductive potential of these insects [98]. Precocene 1 and 2 deterred *Heliothis zea* larvae from feeding in laboratory studies, resulting in starvation [99].

Azadirachtin from the neem tree interfered with molting of the face fly, *Musca autumnalis,* when larvae were exposed to concentrations as low as 0.00001 µg/mL [100]. The application of 2 µg of azadirachtin/larva killed 92 – 100 % of Japanese beetle, *Popillia japonica* [101].

Insect pheromones have potential for several uses in insect management, e.g., monitoring populations, trapping insects to a point source, and interfering with mating. The science of insect pheromones involves several steps, from isolation, identification, and synthesis of the pheromone to learning the behavioral change that these chemicals manifest in a target insect. The state of the art of pheromones has been reviewed [102]. Minute amounts of the pheromone of a given species of insect will attract the male. In some insects, several compounds constitute the pheromone. In *H. zea,* maximum male response is elicited with compounds (*Z*)-7-hexadecenal, (*Z*)-9-hexadecenal, (*Z*)-11-hexadecenal, and hexadecanal, whereas *H. virescens* males respond to a mixture of the above chemicals plus (*Z*)-9-tetradecenal, tetradecanal, and (*Z*)- 11-hexadecen-1-ol [57], [58]. The female sex pheromone of the lesser cornstalk borer, *Elasmopalpus lignosellus,* is composed of 10 compounds (Section 5.1.2). A mixture of four of these compounds, 91.6 µg of (*Z*)-11-hexadecen-1-ol acetate, 11.2 µg of (*Z*)-9-tetradecen-1-ol, 46.4 µg of (*Z*)-9-tetradecen-1-ol acetate, and 86.8 µg of (*Z*)-7-tetradecen-1-ol acetate applied to a rubber septum produced maximal capture of lesser cornstalk borer females [59].

Cucurbitacins act as kairomones to beetles in the subtribe Diabroticites, which includes several species that feed on Cucurbitaceae. This tribe also includes the corn rootworm beetles. Amounts as small as 0.003 µg cause complete arrest of these beetles. If they pass by a point source, they stop and begin compulsive feeding until the food source is depleted [66]. Re-

searchers have demonstrated a strong correlation between total Cucurbitacin content and the extent of aggregation and feeding by two species of *Diabrotica.* Other researchers have proposed using fruits of Cucurbitaceae coated with a chemical insecticide to bait insects to a point source and kill them [103].

Bioderived Toxicants. Ivermectin [*70288-86-7*], 22,23-dihydro derivative of avermectin, is active at extremelylow dosages against a wide variety of arthropod parasites and nematodes [104]; i.e., 0.001 mg/kg is effective against immature stages of *Dirofilaria immitis,* when administered orally to dogs. Also, 0.02 mg/kg is efficacious against *Haematopenus suis* when injected subcutaneously in hogs. A subcutaneous dosage of 200 µg/kg killed blood sucking flies on steers [105]. Pheromone production was significantly reduced when adult boll weevils, *Anthonomus grandis* were topically treated with 7 µg of avermectin B_1 [106].

The sterile autocidal male technique has been used successfully but depends on the ability to rear large numbers of insects for sterilization and release. The screwworm, *Cochliomyia hominivorax,* has been eliminated from Curaçao and southeastern United States. A program is ongoing to suppress the screwworm in Mexico and inhibit its spread to the southwestern United States. To accomplish this, ca. 350×10^6 sterile flies are released weekly. The established population of flies has been pushed back in excess of 600 km from the United States – Mexican border.

The sterile insect technique is a major component of the Mediterranean fruit fly, *Ceratitis capitata,* program in California. Up to 40×10^6 flies are released weekly, enabling success in keeping this pest out of California.

An extensive program has been undertaken to suppress populations of the codling moth, *Laspeyresia pomonella,* in the state of Washington, United States, and British Columbia, Canada. The codling moth is widely distributed and feral populations continually migrate into suppression areas rending the program ineffective.

Other autocidal techniques are generally on an experimental basis. The hybrid sterility technique crossing *Heliothis virescens* males with *H. subflexa* females has been tried experimentally in Puerto Rico, indicating a good potential for infusion of the sterility factor into the native population [107].

Whenever a component is used to manage a pest population, the question of **safety** arises. Usually, pest control is undertaken to protect a commodity that is in the food chain which eventually is consumed by humans. Thus, human safety is extremely important. Recently, safety to the environment and to the ecological balance has become recognized as important. The organisms and chemicals discussed in this article are entities that are indigenous to a certain ecological niche or they are synthesized chemical compounds that are an instrumental part of nature. Thus, if these components are used within the environmental framework from which they evolved, safety is relatively assured.

Exhaustive lists are available for some organisms, stating which target and nontarget organisms are susceptible [108]. For others, guidelines for safety tests and registration have been established [109], [110]. In depth studies have been conducted examining the effect of insect nuclear polyhedrosis viruses in mammalian cells [111], with no consistent deleterious effects directly attributable to the viruses being observed.

Prior to inoculative releases, phytophagous insects and plant pathogens that are candidate biological control agents are subject to exhaustive host range studies to be certain that the introduced species will not become pests in a different ecosystem.

In general, the biological control agents discussed here will not have a detrimental impact when used sensibly in an integrated pest management program.

7. References

1. C. B. Huffaker, P. S. Messenger (ed.): *Theory and Practice of Biological Control,* Academic Press, New York 1976.
2. P. Debach: *Biological Control by Natural Enemies,* Cambridge University Press, New York 1974.
3. C. P. Clausen: *Entomophagous Insects,* McGraw-Hill, New York 1940.
4. E. Kurstak (ed.): *Microbial and Viral Pesticides,* Marcel Dekker, New York 1982.
5. G. E. Cantwell (ed.): *Insect Diseases,* vol. 1 and 2, Marcel Dekker, New York 1974.

6. E. A. Steinhaus (ed.): *Insect Pathology: An Advanced Treatise,* vol. 1 and 2, Academic Press, New York 1963.
7. E. W. Davidson (ed.): *Pathogenesis of Invertebrate Microbial Diseases,* Allanhead, Osmun Publ., Totawa, N.J., 1982.
8. *Ullmann,* 4th ed., **13** : 238.
9. H. D. Burges, N. W. Hussey (ed.): *Microbial Control of Insects and Mites,* Academic Press, New York 1971.
10. H. D. Burges (ed.): *Microbial Control of Pests and Plant Diseases 1970 – 1980,* Academic Press, New York 1981.
11. L. A. Bulla, Jr., T. C. Cheng (ed.): *Comparative Pathobiology, Biology of the Microsporidia,* vol. 1, Plenum Publishing, New York 1976.
12. K. M. Smith: *Virus-Insect Relationships,* Longman Group Ltd., London 1976.
13. D. C. Kelly, *J. Gen. Virol.* **63** (1982) 13.
14. G. O. Poinar, Jr.: *Nematodes for Biological Control of Insects,* CRC Press, Boca Raton, Fla., 1979.
15. J. J. Menn, M. Beroza (ed.): *Insect Juvenile Hormones: Chemistry and Action,* Academic Press, New York 1972.
16. D. A. Nordlund, R. L. Jones, W. J. Lewis (ed.): *Semiochemicals: Their Role in Pest Control,* J. Wiley & Sons, New York 1981.
17. M. Beroza (ed.): *Pest Management with Insect Sex Attractants and Other Behavior-Controlling Chemicals,* Amer. Chem. Soc., Washington, D.C., 1976.
18. V. B. Wigglesworth: *Insect Hormones,* W. H. Freeman & Co., San Francisco 1970.
19. *Biological Control of Pests in China,* USDA, Office of International Cooperation and Development, Scientific and Technical Exchange Division, China Program, Washington, D.C., 1982.
20. R. J. Cook, K. F. Baker: *The Nature and Practice of Biological Control of Plant Pathogens,* The American Phytopathology Society, St. Paul 1983.
21. R. Van den Bosch, P. S. Messenger, A. P. Gutierrez: *An Introduction to Biological Control,* Plenum Publishing, New York 1982.
22. R. M. Anderson, E. U. Canning, A. E. R. Taylor, R. Muller (ed.): "Parasites as Biological Control Agents," *Symposia of the British Society for Parasitology,* vol. 19, Cambridge University Press, Cambridge 1982.
23. H. G. Miltenburger (ed.): *Safety Aspects of Baculoviruses as Biological Insecticides,* Bundesministerium für Forschung und Technologie, Bonn 1980.
24. Reference 1, F. J. Simmonds, J. M. Franz, R. I. Sailer: "History of Biological Control," Chap. 2.
25. W. A. Baker, W. G. Bradley, C. A. Clark: *Biological Control of the European Corn Borer in the United States,* USDA Technical Bulletin No. 983, 1948, pp. 180 – 181.
26. T. A. Angus, *Nature (London)* **173** (1954) 545 – 546.
27. Bioferm Corp., US 3073749, 1963 (J. C. Megna).
28. Reference 6, A. M. Heimpel, T. A. Angus: "Diseases Caused by Certain Sporeforming Bacteria," vol. 2, Chap. 2.
29. M. B. Mohd-Salleh, L. C. Lewis, *J. Invertebr. Pathol.* **39** (1982) 290 – 297.
30. H. deBarjac, A. Bonnefoi, *Entomophaga* **7** (1962) 5 – 31.
31. H. deBarjac, A. Bonnefoi, *Entomophaga* **18** (1973) 5 – 17.
32. J. Krywienczyk, H. T. Dulmage, P. G. Fast, *J. Invertebr. Pathol.* **31** (1978) 372 – 375.
33. L. J. Goldberg, J. Margalit, *Mosq. News* **37** (1977) 355 – 358.
34. J. Farkas, K. Sebesta, K. Horska, Z. Samek, et al., *Collect. Czech. Chem. Commun.* **34** (1969) 1118 – 1119.
35. M. B. Mohd-Salleh, C. C. Beegle, L. C. Lewis, *J. Invertebr. Pathol.* **35** (1980) 75 – 83.
36. International Mineral & Chemicals Corp., US 3758383, 1973 (T. R. Shieh, M. H. Rogoff).
37. N. D. Levine, J. O. Corliss, F. E. G. Cox, G. Deroux et al., *J. Protozool.* **27** (1980) 37 – 58.
38. L. C. Lewis, *Can. Entomol.* **110** (1978) 897 – 900.
39. L. C. Lewis, R. D. Gunnarson, J. E. Cossentine, *Can. Entomol.* **114** (1982) 599 – 603.
40. T. J. Shapas, W. E. Burkholder, G. M. Boush, *J. Econ. Entomol.* **70** (1977) 469 – 474.
41. P. V. Vail, D. L. Jay, *J. Invertebr. Pathol.* **21** (1973) 198 – 204.
42. L. C. Lewis, R. E. Lynch, J. J. Jackson, *Environ. Entomol.* **6** (1977) 535 – 538.
43. L. C. Lewis, J. R. Adams, *J. Invertebr. Pathol.* **33** (1979) 253 – 256.
44. S. H. Sharrow, W. D. Mosher, *J. Range Manage.* **35** (1982) 480 – 482.
45. L. Sundheim, *Plant Pathol.* **31** (1982) 209 – 214.
46. J. B. Siddall, A. D. Cross, J. H. Fried, *J. Am. Chem. Soc.* **88** (1966) 862 – 863.
47. K. Nakanishi: "Steroids," in K. Nakanishi, T. Goto, S. Ito, S. Natori et al.: *Natural Products, Chemistry,* vol. 1, Academic Press, New York-London 1974, p. 535.

48. W. S. Bowers, H. M. Fales, M. J. Thompson, E. C. Uebel, *Science* **154** (1966) 1020 – 1021.
49. N. Wakabayashi, R. M. Waters: "Juvenile Hormones and Analogs," in E. D. Morgan, B. Mandava (ed.): *Handbook on Natural Pesticides: Insects*, vol. 1, CRC Press, Boca Raton, Fla., (in press).
50. K. Mori: "Synthetic Chemistry of Insect Pheromones and Juvenile Hormones," in *Recent Developments in the Chemistry of Natural Carbon Compounds*, vol. 9, Publishing House of the Hungarian Academy of Sciences, Budapest 1979, pp. 11 – 209.
51. Reference 47, S. Nozoe: "Mono- and Sesquiterpenes," p. 39.
52. C. A. Hendrick: "Juvenile Hormone Analogs: Structure-Activity Relationships," in J. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, pp. 315 – 402.
53. W. S. Bowers, T. Ohta, J. S. Cleere, P. A. Marsella, *Science* **193** (1976) 542 – 547.
54. W. S. Bowers: "Toxicology of Precocene", in J. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, 403 – 427.
55. A. Juss: "Natural Pesticides from the Neem Tree (*Azadirachta indica*)," in H. Schmutterer, K. R. S. Ascher, H. Rembold (ed.): *Proceedings of the First International Neem Conference,* German Agency for Technical Cooperation, Eschborn, Federal Republic of Germany, 1981.
56. J. D. Warthen, Jr., USDA, Sci. and Educ. Admin., *Agricultural Research Results,* 1979, no. 1.
57. J. A. Klun, J. R. Plimmer, B. A. Bierl-Leonhardt, A. N. Sparks, et al., *Science* **204** (1979) 1328 – 1330.
58. J. A. Klun, J. R. Plimmer, B. A. Bierl-Leonhardt, A. N. Sparks, *J. Chem. Ecol.* **6** (1980) 165 – 175, 177 – 183.
59. R. E. Lynch, J. A. Klun, B. A. Leonhardt, M. Schwarz, *Environ. Entomol.* **13** (1984) 121 – 126.
60. J. M. Brand, J. C. Young, R. M. Silverstein: "Insect Pheromones: A Critical Review of Recent Advances in Their Chemistry, Biology, and Application," *Prog. Chem. Org. Nat'l. Prod.* **37** (1979) 1 – 190.
61. H. J. Bestman, "Stereoselective Synthesis of Pheromones via Phosphonium Ylides," in P. Doyle, T. Fujita (ed.): *Pesticide Chemistry: Human Welfare and the Environment, Synthesis and Structure Activity Relationships,* vol. 1, Pergamon Press, Oxford 1982, pp. 77 – 82.
62. A. K. Raina, J. A. Klun, *Science* **225** (1984) 531 –532.
63. W. Klassen, R. L. Ridgway, M. Inscoe: "Chemical Attractants in Integrated Pest Management Programs," in A. F. Kydonieus, M. Beroza (ed.): *Insect Suppression with Controlled Release Pheromone Systems,* vol. 1, CRC Press, Boca Raton, Fla., 1982.
64. D. Lavie, E. Glotter: "The Cucurbitanes, a Group of Tetracyclic Triterpenes," in *Progress in the Chemistry of Organic Products,* vol. 29, Springer-Verlag, Wien, New York 1971, pp. 308 – 362.
65. R. L. Metcalf, A. M. Rhodes, R. A. Metcalf, J. Ferguson et al., *Environ. Entomol.* **11** (1982) 931 –937.
66. R. L. Metcalf, *Bull. Entomol. Soc. Am.* **25** (1979) 30 – 35.
67. R. A. Dybas: "Avermectins: Their Chemistry and Pesticidal Activities," in P. Doyle, T. Fujita (ed.): *Pesticide Chemistry: Human Welfare and the Environment, Synthesis and Structure Activity Relationships,* vol. 1, Pergamon Press, Oxford 1982, pp.83 – 90.
68. M. Stengel, *Entomophaga* **27** (no° H. S.) (1982) 105 – 114.
69. L. M. Stevens, A. I. Steinhauer, J. R. Coulson, *Environ. Entomol.* **4** (1975) 947 – 952.
70. W. A. Shands, G. W. Simpson, R. H. Storch, *J. Econ. Entomol.* **65** (1972) 799 – 805.
71. G. Majori, A. Ali, *J. Invertebr. Pathol.* **43** (1984) 316 – 323.
72. L. A. Pistrang, J. F. Burger, *Can. Entomol.* **116** (1984) 975 – 981.
73. J. D. Podgwaite, P. Rush, D. Hall, G. S. Walton, *J. Econ. Entomol.* **77** (1984) 525 – 528.
74. Reference 4, J. C. Cunningham, pp. 335 – 386.
75. Reference 4, W. C. Yearian, S. Y. Young, pp. 387 – 423.
76. R. E. Lynch, L. C. Lewis, E. C. Berry, *J. Econ. Entomol.* **73** (1980) 4 – 7.
77. C. Bajan, T. Bilewicz-Pawinska, A. Fedorka, K. Kmitova: *Report 8th International Plant Protection Congress,* Sec. 5, Moscow 1975, pp. 33 – 40.
78. H. D. Burges, *J. Invertebr. Pathol.* **28** (1976) 217 – 222.
79. J. E. Henry, *Ann. Rev. Entomol.* **26** (1981) 49 – 73.
80. J. E. Henry, J. A. Onsager, *Entomophaga* **27** (1982) 197 – 201.
81. M. A. Bari, H. K. Kaya, *J. Econ. Entomol.* **77** (1984) 225 – 229.
82. R. Gaugler, *J. Nematol.* **13** (1981) 241 – 249.

83. C. W. McCoy, T. L. Couch, *Fla. Entomol.* **65** (1982) 116 – 126.
84. N. W. Hussey, L. Bravenboer: "Control of Pests in Glasshouse Culture by the Introduction of Natural Enemies," in C. B. Huffaker (ed.): *Biological Control,* Plenum Publishing, New York 1971, pp. 195 – 216.
85. I. J. Wyatt: "Progress Towards Biological Control Under Glass," in D. P. Jones, M. E. Solomon (ed.): *Biology in Pest and Disease Control,* Blackwell Scientific Publ., London 1974, pp. 293 – 301.
86. J. V. Cross, *Plant Pathol.* **33** (1984) 417 – 423.
87. M. Van De Veire, V. Vacante, *Entomophaga* **29** (1984) 303 – 310.
88. S. J. Cornelius, H. C. J. Godfray, *Entomophaga* **29** (1984) 341 – 345.
89. R. Haas, R. Pal, *Bull. Entomol. Soc. Am.* **30** (1984) 17 – 25.
90. E. C. Bay, C. O. Berg, H. C. Chapman, E. F. Legner in C. B. Huffaker, P. S. Messenger (ed.): *Theory and Practice of Biological Control,* Academic Press, New York 1976, Chapter 18.
91. S. E. Riechert: "Spiders as Biological Control Agents," *Ann. Rev. Entomol.* **29** (1984) 299 – 320.
92. A. R. Gotlieb, M. H. Brosseau, A. K. Watson: *Abstracts of 1984 Weed Science of America Meeting,* Miami, Fla., 1984, p. 68.
93. E. B. Adams, R. F. Line, *Phytopathology* **74** (1984) 745 – 748.
94. G. R. Stirling, *Phytopathology* **74** (1984) 55 – 60.
95. H. R. Forrer, J. Werder: "The Influence of Antagonistic Fungi on the Spore-Formation of Rust Fungi," in H. Geissbühler (ed.): *Advances in Pesticide Science,* Pergamon Press, Oxford 1978, Part 2, pp. 383 – 388.
96. C. P. Clausen (ed.): "Introduced Parasites and Predators of Arthropod Pests and Weeds: A World Review," *Agriculture Handbook No. 480,* U.S. Agriculture Research Service, USDA, Washington, D.C., 1978.
97. J. L. Robertson, R. A. Kimball, *Can. Entomol.* **111** (1979) 1361 – 1368.
98. J. L. Robertson, R. A. Kimball, *Can. Entomol.* **111** (1979) 1369 – 1380.
99. C. S. Wisdon, J. T. Smiley, E. Rodriguez, *J. Econ. Entomol.* **76** (1983) 993 – 998.
100. I. A. Gaaboub, D. K. Hayes, *Environ. Entomol.* **13** (1984) 803 – 812.
101. T. L. Ladd, Jr., J. D. Warthen, Jr., M. G. Klein, *J. Econ. Entomol.* **77** (1984) 903 – 905.
102. J. A. Klun: "Insect Sex Pheromones," in J. L. Hilton (ed.): *Agricultural Chemicals of the Future, Proc. Beltsville Agric. Symp. 8*, Rowman & Allanheld, Totowa, N.J., 1985, pp. 381 – 386.
103. A. M. Rhodes, R. L. Metcalf, E. R. Metcalf, *J. Am. Soc. Hortic. Sci.* **105** (1980) 838 – 842.
104. W. C. Campbell, M. H. Fisher, E. O. Stapley, G. Alberts-Schönberg et al., *Science* **221** (1983) 823 – 828.
105. H. A. Standfast, M. J. Muller, D. D. Wilson, *J. Econ. Entomol.* **77** (1984) 419 – 421.
106. J. E. Wright, *J. Econ. Entomol.* **77** (1984) 1029 – 1032.
107. D. F. Martin, M. L. Laster, F. I. Proshold, P. D. Lindgren et al., *Environ. Entomol.* **13** (1984) 701 – 707.
108. C. C. Beegle, L. C. Lewis, T. Yamamoto: "Bacteria," in C. M. Ignoffo (ed.): *Handbook of Naturally Occurring Pesticides: Microbial Insecticides,* CRC Press, Boca Raton, Fla., 1985.
109. H. D. Burges, J. Huber, G. Croizier, *Entomophaga* **25** (1980) 341 – 348.
110. H. D. Burges, A. Krieg, P. Luthy, H. deBarjac, *Entomophaga* **27** (1982) 225 – 236.
111. R. Reimann, H. G. Miltenburger, *Entomophaga* **27** (1982) 267 – 276.

Biomass Chemicals

Barbara A. Tokay, Englewood, New Jersey 07631, United States

1. Introduction 1
2. Raw Materials 1
3. Process Technology 2
4. Biomass Chemicals 3
4.1. Ethanol and Its Derivatives 3
4.2. Other Fermentation Chemicals . . . 3
4.3. Sugar Chemicals 4
4.4. Starch Chemicals 4
4.5. Oilseed Chemicals 4
4.6. Furfural and Its Derivatives 4
4.7. Cellulose Polymers 5
4.8. Lignin Chemicals 5
4.9. Methanol 5
4.10. Methane 6
5. Future Contributions 6
6. References 6

1. Introduction

Biomass chemicals are those chemicals made from renewable, organic materials. Most biomass materials are plant products or wastes, but animal products, such as fats or manures, also are included.

Although this category enjoyed renewed interest because of the oil crises during the 1970s, biomass chemicals are not new inventions. In fact, at one time, the chemical industry was largely based on chemical products made from wood, sugars, starches, and animal fats (→ Wood, → Sugar, → Starch, → Polysaccharides, → Fats and Fatty Oils) [1].

Some biomass chemicals are still staples of chemical industry. Cellulose fibers, for example, command large markets even after years of competition from petroleum-based materials. Many other biomass chemical processes were in operation until the 1960s, and some are still operating in developing nations or in countries without ready access to petroleum [2]. Among the biomass chemicals that are, or were, important are ethanol, acetone/butanol, itaconic acid, lactic acid, xanthan gum, sorbitol, industrial starches, fatty acids and alcohols from vegetable and animal sources, glycerol, soap, furfural and its derivatives, rayon, cellophane, carboxymethylcellulose, lignosulfonates, tall oil, vanillin, and methanol.

During the oil shocks of the 1970s, a number of chemical companies decided to take another look at biomass chemicals [3]. In the United States, the biomass chemical receiving the most attention was ethanol made by fermentation of grain (→ Ethanol). Ethanol was, at that time, vigorously supported by the federal government as a gasoline extender and octane enhancer. Other biomass-based commodity chemicals receiving attention at that time were methanol, acetic acid, butanol/acetone, and ethanol derivatives.

Petroleum prices have stabilized and most United States' chemicals producers have abandoned their interests in biomass commodity chemicals [3]. However, several manufacturers are developing new lines of biomass specialty chemicals, such as novel starch polymers [1], [2, p. 147], [4], new lignin chemicals [4–6], and polysaccharide gums [3].

The rapid rise and fall of interest in biomass chemicals parallels, in part, the curve of petroleum prices, but other characteristics of biomass chemicals have contributed to the cautious attitude toward their redevelopment. These characteristics and the potential for development and manufacture of new and old biomass chemicals are detailed in the following sections.

2. Raw Materials

One major problem in organizing production of biomass chemicals is the diversity of materials from which these chemicals may be made. The biomass *feedstock* category is vast; it includes grains, sugar crops, oilseed crops, agri-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_099

cultural wastes, food processing wastes (liquid and solid), wood, wood chips, bark, mill residues, forest residues, pulping liquors, animal fats, manures, algae, and exotic crops, such as guayule, jojoba, and euphorbia.

All biomass materials contain carbon, hydrogen, and oxygen, but the chemical *composition* of each type of material is unique. Because most biomass materials contain considerable amounts of oxygen (as much as 40 %), they are generally considered to be most useful as sources of oxygenated chemicals [7], [8].

The diversity of biomass materials means that each biomass chemical *process* must be designed to use a specific type of material. Unlike petroleum, which is a much simpler, more homogeneous material, biomass from different sources is not interchangeable.

Production and *collection* of biomass feedstocks also present problems. If a process uses a well-known biomass feedstock, for example, starch, then production and collection is simple because a starch-processing industry already exists. However, use of starch for production of new chemicals or for increased production of conventional starch-based chemicals may create competition with existing starch markets. The same line of reasoning applies to the use of timber or pulp-grade wood as a feedstock for chemicals production.

If a nonconventional biomass feedstock is selected, agricultural residues, for example, then collection of this dispersed material poses logistical problems. In general, biomass wastes are bulky, dispersed, and difficult to collect and transport economically for any appreciable distance.

The biomass chemicals processes in operation today use conventional biomass feedstocks (sugar, starch, wood) or biomass wastes (forest residues, pulping liquors, agricultural residues). Investigations are under way on new crops, such as guayule, jojoba, euphorbia, or crambe, that can be grown on currently uncultivated land and used exclusively for chemicals production [9]. Some of these projects, as well as creation of short-rotation forestry projects, still need further development of production and collection infrastructures, processing industries, and markets.

3. Process Technology

Because biomass materials are complex in composition, *conversion* of these substances to chemicals is often inefficient [1], [10]. Plant materials, for example, contain three major components: cellulose, hemicellulose, and lignin (→ Cellulose, → Lignin). If a chemical process uses only one component and the others become wastes that must be disposed, the process often becomes uneconomic. In traditional wood pulping, for example, cellulose for cellulose fiber and chemical production is separated from lignin and hemicellulose fractions, which are usually burned for their energy content or occasionally made into specialty chemicals. New wood chemicals processes would almost certainly have to include higher-value uses for the lignin and hemicellulose. Some biomass experts have suggested that a "biomass refinery" concept is necessary to any successful biomass chemicals plan in the future [4], [11].

Separation of the components of lignocellulose materials without intensive use of chemicals and energy, as is done in traditional wood pulping, is no easy matter. Development of new process technology that would convert the six-carbon sugars of wood cellulose to ethanol has been slowed considerably by the separation problem, although it was the subject of intensive research [10].

Biomass chemicals made from a purified component of lignocellulose are often produced via *fermentation.* Feedstock for these chemicals, which include ethanol, butanol/acetone, citric acid, and xanthan gum, is usually sugar or starch [2, pp. 5, 71]. Fermentation is a well-established process, but there are some problems in its wider application to chemical production. A typical fermentation produces a dilute product stream containing 10 % or less of the desired product, which must then be concentrated. Ethanol is usually separated from the fermentation beer by energy-intensive distillation. There has been much research and development on ethanol separation and progress has been made, but the problem is by no means eliminated [8].

Biomass chemicals may also be made by *chemical conversion* routes. Chemically modified industrial starches are produced by this route, as are the starch-based polymers [1], [2, pp. 137 – 154]. Vegetable-oil-based fatty acids

and alcohols are other products manufactured via chemical process routes [2, pp. 155 – 223].

Because separation of biomass components is often difficult and costly, some biomass chemicals are made by *thermochemical routes* that use the entire material. The destructive distillation of wood to methanol [*67-56-1*] is the oldest and best known example of this type of process(→ Methanol). The newer processes gasify biomass to synthesis gas, then reform it to methanol (→ Gas Production).

Biomass may also be *gasified* to methane or pyrolyzed to tars and liquid hydrocarbon products. Most work on gasification and pyrolysis for chemical production is in the laboratory or in early pilot plant development [12].

Another whole material chemical conversion route involves *anaerobic digestion* of biomass solids or liquids. Developed as a method for treating liquid sewage, methane-producing anaerobic digestion was initially conceived as a less energy-intensive method of reducing chemical oxygen demand of wastes. The process was also explored as a means of producing pipeline-quality methane.

4. Biomass Chemicals

4.1. Ethanol and Its Derivatives

The biomass chemical that has attracted the most interest in the early 1980s was ethanol [*64-17-5*] (→ Ethanol) [1], [4], [13]. In the United States, production of fermentation ethanol for use as a fuel extender was supported by a fuel tax exemption that extended to the end of the 1980s. The effect of this tax incentive has been a large growth in production of fuel and industrial-grade fermentation ethanol. Industrial grade fermentation ethanol is displacing synthetic ethanol in the chemicals market and some analysts predicted that the industrial ethanol market would be dominated by fermentation ethanol by 1990 [13].

Fuel ethanol is made by traditional batch fermentation of starches or sugars with the yeast *Saccharomyces cerevisiae.* New continuous fermentation processes are approaching commercialization [14].

Biomass-based ethanol may be converted to other large-volume commodity chemicals. Ethanol may be converted to ethylene by passing ethanol vapor over a heated fixed-bed alumina catalyst. The ethylene – water mixture is quenched and the two components separated [2, pp. 40 –47]. Although this route would be prohibitively expensive in countries with a well-developed petrochemical industry [1], [7], the ethanol-to-ethylene method is used in Brazil, which has a large government-supported fermentation ethanol program.

Acetaldehyde, now made via a petrochemical route, can also be produced from ethanol (→ Acetaldehyde, Chap. 4.1.). Ethanol is vaporized and passed over a heated, fixed-bed chromium-activated copper catalyst. Hydrogen is separated from the acetaldehyde and untreated ethanol is recycled. Acetaldehyde may also be converted to *n*-butanol via acetaldol and crotonaldehyde, although this process is expensive [2, pp. 48 – 54].

Butadiene was once produced from ethanol in a single-step process in which acetaldehyde and ethanol were passed over a tantalum oxide on silica gel catalyst. Energy intensive distillation operations necessary for product purification make this process expensive [2, pp. 62 – 70].

4.2. Other Fermentation Chemicals

Although ethanol production is the best known chemical fermentation process, there are many others. The butanol/acetone fermentation, still operated in South Africa, was once of great industrial importance (→ Acetone, Chap. 4.6.; → Butanols). Raw material for the process is grain or molasses, and the anaerobic fermentation uses the bacterium *Clostridium acetobutylicum.* The fermentation process produces a 60 : 30 : 10 mixture of butanol/acetone/ethanol, which must be separated by distillation [2, pp. 76 – 83].

Submerged aerobic fermentations for antibiotic production, initially developed during World War II, opened up new chemical production routes. Citric, gluconic, and itaconic acids are made by this route, as is xanthan gum, a polysaccharide with applications as a thickener, emulsifier, and stabilizer of food products (→ Citric Acid, → Gluconic Acid, → Polysaccharides) [15]. Xanthan gum and other similar polysaccharides may have poten-

tial as oil recovery agents or in oil clean-up applications [16].

4.3. Sugar Chemicals

Pure dextrose, usually obtained from starch, is the raw material for manufacture of sorbitol, a well-established chemical used in foods, dentifrices, and cosmetics, and as a chemical intermediate (→ Sugar Alcohols) [2, pp. 129 – 136]. Sucrose polyols also have a role in urethane foam manufacture.

4.4. Starch Chemicals

More than half the starch isolated from grains and potato crops is converted into sugars. The remainder has many industrial applications (→ Starch). Industrial starch may be used as is or it may be modified by partial depolymerization to dextrins or by chemical reactions that replace the hydroxyl groups of the dextrose unit with other substituents. Among the many modified starches are oxidized starch, acetates, phosphates, amino alkyl ethers, and other ethers. These large-volume, inexpensive materials are used as paper and textile sizes and as binders [2, pp. 137 – 146].

Some more drastically modified starch chemicals have also been introduced. A starch-acrylonitrile graft copolymer, developed and licensed by the USDA, has specialty uses in personal care products and in medical applications. A starch slurry is treated with acrylonitrile in the presence of a ceric ammonium nitrate catalyst in nitric acid. The product, called "Super Slurper," absorbs hundreds of times its mass in water [2, pp. 147 – 154].

Another starch-based chemical is α –methyl glucoside, a starch polyol used in production of rigid urethane foam (→ Polyurethanes) [7].

4.5. Oilseed Chemicals

Linseed, soybean, cottonseed, coconut, palm, and other plant oils and animal fats have always been important raw materials for chemical manufacture. These oils contain a mixture of triglycerides of fatty acids in the 8 to 20 carbon range.

Although biomass fatty acids have suffered from competition from petrochemical equivalents, they still retain sizable market shares in surfactants, paint binders, and plasticizers.

Natural fatty acids are made by hydrolyzing fats or plant oils with high-pressure steam, which requires investment in high-pressure equipment plus extensive vacuum distillation capacity to fractionate the acid mixture. The products find big markets, however, in surfactants, plasticizers, greases, synthetic lubricants, cosmetics, toiletries, textile chemicals, emulsifiers, secondary petroleum recovery products, and mineral beneficiation [2, pp. 155 – 223]. Processes and markets are described in more detail under → Fatty Acids.

Triglycerides also yield long-chain fatty alcohols, which have lubricant and surfactant applications. Soap, another natural oil product, has lost ground to synthetic detergents, but glycerol, which is a byproduct of soap or fatty acid manufacture, holds its own against synthetic glycerol.

Fatty acid derivatives, such as azelaic and pelargonic acids, formed by ozonolysis from oleic acid, find niches in specialty markets. Azelaic and pelargonic acids are used in modified polyester fibers, elastomers, adhesives, plasticizers, synthetic lubricants, and surface coatings. Nylon 11, made from ricinoleic acid derived from castor oil, is marketed as a plastic for molded parts in automobiles, conveying equipment, and other machinery. Other specialized nylons may also be made from natural fatty acids. Nylon 9 is derived from soybean oil; nylon 1313 is made from erucic acid from crambe oil.

Jojoba oil, extracted from the seed of the jojoba plant, an exotic desert shrub, is used in cosmetics. Its structure is very close to that of sperm oil, now on the proscribed list in many nations because of efforts to save the near extinct sperm whale. Jojoba oil producers are trying to introduce the oil, an ester of a long-chain fatty acid and a long-chain alcohol, into the high-pressure lubricant market and perhaps to the automobile lubricant market [9].

4.6. Furfural and Its Derivatives

Furfural (2-furaldehyde) [*98-01-1*] and its-derivatives, furfuryl alcohol [*98-00-0*], furan

resins, and tetrahydrofuran [*109-99-9*], are produced in many countries from corn cobs, wheat and oat hulls, and many other biomass materials (→ Furan and Derivatives, → Tetrahydrofuran). *Furfural,* which is derived from the pentoses (five-carbon sugars) present in hemicellulose, one of the major components of lignocellulosic materials, is produced by steam treatment followed by dehydration. Furfural was often mentioned in the early 1980s as a possible new product for developing countries with plentiful biomass resources, and several nations have successfully established plants [13].

One-third of the furfural produced is used as a solvent; the other two-thirds is used as an intermediate in the manufacture of furfural derivatives. The most important *furfuryl alcohol,* used to produce furan resins for foundry sand binders. Another derivative, *tetrahydrofuran,* can also be produced via a petrochemical route. Other uses for furfural, such as production of adiponitrile, might be found if furfural prices were reduced by expanded production [13].

4.7. Cellulose Polymers

Cellulose polymers, such as rayon, cellulose acetate, cellulose ester, cellophane, and modified cellulose, are still important in spite of heavy competition from petrochemical alternatives. Markets and process technology for these chemicals are described more fully under → Cellulose, → Cellulose Esters.

Interest in cellulose stems from alternative fuel research growing out of the oil crises of the 1970s. Cellulose is viewed as a possible source of hexose sugars, which can be fermented to ethanol. New lignocellulose fractionation schemes resulting from the cellulose-to-ethanol research may contribute novel pulping technologies that will preserve the value of the hemicellulose and lignin fractions of the lignocellulosic materials [4], [10].

4.8. Lignin Chemicals

Paper pulp mills have long produced vast waste streams containing the lignin (→ Lignin) and hemicellulose fractions of wood. These waste streams, which contain chemicals from the pulping process, are usually burned for their energy value, but a small portion is converted to a variety of chemicals.

Neutral sulfite semichemical pulping yields a black liquor from which acetic and formic acids may be extracted. Kraft black pulping liquor may be a source of dimethyl sulfoxide, which is made by heating the liquor to convert the lignin methyl groups to methyl sulfide. Oxidation then produces dimethyl sulfide [2, pp. 295 – 302].

Sulfonated lignins, a major component of Kraft black liquor, can be recovered as lignosulfonates, which are used as dispersants, sequestrants, emulsion stabilizers, and humectants. Crude calcium lignosulfonate can be oxidized to vanillin, which competes with the extract from the vanilla bean [2, pp. 311 – 319].

Tall oil is a byproduct of the Kraft process for pulping softwoods. Tall oil is a mixture of rosin and a group of fatty acids (→ Tall Oil) [2, pp. 320 – 328].

New lignocellulose separation processes, such as steam explosion, have led to production of an unmodified, low-molecular-mass lignin [5]. Although this material has yet to be widely tested, it may be useful as a partial substitute for phenolic resoles in phenol – formaldehyde adhesives. Other uses for this type of lignin have also been suggested: binder for animal feeds, encapsulating agent, feed grain coating, road dust treatment chemicals, reinforcing fillers in rubbers, and setting agent in Portland cement [5]. In addition use of lignin copolymers in drilling muds has been discussed [17].

Some proponents of "biomass refining" suggest that lignin be hydrocracked and hydroalkylated to phenol and benzene [6], [11].

4.9. Methanol

The old route to biomass methanol, often called wood alcohol, was destructive distillation. In this method, 1 t of wood heated in steel ovens produces 0.25 t charcoal, 30 kg acetic acid, 10 kg methanol, and 60 L tar and oil from low-energy gas. This process is not economically feasible in industrialized countries although it might be in less developed nations [2, pp. 339 –347].

New methods of methanol production require gasification of biomass, often in an oxygen-

blown gasifier, to produce synthesis gas, which is then reformed to give methanol. Several projects are now in the pilot plant or demonstration stage (→ Methanol) [18].

4.10. Methane

There are several ways to produce methane [*74-82-8*] from biomass. The most widely used is anaerobic digestion, a conversion process originally developed for clean up of low-strength organic liquid wastes. Anaerobic digestion is a slow fermentation process in which a variety of bacteria convert organic matter first to organic acids and then to methane. Much experimentation is now going on in this field, with most applications being found in waste treatment [19]. Methane production in these applications is secondary, with gas produced being used to fuel the process. However, the process can be used for its methane-producing capacity alone. New reactor designs and process research make this possibility more likely (→ Methane).

5. Future Contributions

Biomass-based chemicals always have been important to the chemical industry and they are likely to remain so in the future. If their contribution is to grow, however, several conditions must be met.

First, *new markets* for biomass-based products must be identified. Although some have proposed constructing a chemicals industry based on four fermentation products [20], most authorities believe biomass chemicals must be introduced to specific target markets where they may replace petrochemical equivalents by indirect substitution [1], [4], [8]. Specialty chemicals from biomass are probably more likely to meet with success than commodity chemicals.

Second, *economic conditions* must be realistically assessed. Although production of many oxygenated chemicals could be based on starch, for example [7], [8], this is not likely to happen as long as petroleum and natural gas prices are stable. However, if these prices rise significantly, such commodity chemicals options might become more promising. Some government support of research and development is likely to be necessary in development of new commodity chemicals process technology [3].

Third, production of new biomass chemicals will likely involve *new collaborations* between producers of raw materials and chemical manufacturers. Such collaboration is now beginning between starch processing companies and a few chemical manufacturers. Forest products companies, which are likely suppliers of wood and wood wastes for chemical production, are much less interested at present in such collaboration [21].

6. References

1. C. E. Dunlap: "Opportunities for Chemicals from Carbohydrates," *American Institute of Chemical Engineers Annual Meeting,* Washington, D.C., October 30, 1983.
2. *Biomass Process Handbook: A Production/Economic Guide to 42 Chemical Processes that Use Biomass as a Raw Material,* revised ed., Technical Insights, Inc., Fort Lee, N.J., 1983, pp. 1 – 4.
3. B. A. Tokay, "New Perspectives on Biomass Chemicals," *Chemical Business* 1983 (August 22) 8 – 14.
4. E. S. Lipinski: "Chemicals from Biomass: Petrochemical Substitution Options," *Science* **212** (1981) 1465 – 1471.
5. D. W. Sundstrom, H. E. Klei: "Use of By-Product Lignins from Alcohol Fuel Processes," *Fourth Symposium on Biotechnology in Energy Production and Conservation,* J. Wiley & Sons, New York 1982, pp. 45 – 56.
6. T. L. Donaldson, O. L. Culberson: "Potential for Chemicals from Biomass," *Symposium Papers Energy from Biomass and Wastes VI,* Institute of Gas Technology, Chicago, Ill., 1982.
7. R. M. Busche: "Will Biomass Supply Feedstocks of the Future?", *Ind. Chem. News* **1983** (May) 34 – 35.
8. T. K. Ng, R. M. Busche, C. C. McDonald, R. W. F. Hardy: "Production of Feedstock Chemicals," *Science* **219** (1983) 733 – 740.
9. Soil and Land Use Technology, *Feasibility of Introducing New Crops: Production/Marketing/Consumption (PMC)* Systems, Rodale Press, Emmaus, Pa., 1981.

10. J. M. Nystrom, C. G. Greenwald, F. G. Harrison: "The Production of Ethanol from Cellulosic Feedstocks," *American Institute of Chemical Engineers Annual Meeting,* Washington, D.C., October 30, 1983.
11. W. H. Klausmeier: "Biomass Chemicals Production by Thermochemical Conversion," in C. Scott (ed.): *Fifth Symposium on Biotechnology for Fuels and Chemicals,* J. Wiley & Sons, New York 1984.
12. *Biomass Digest* **5** (1983) no. 11, 2, 3, 6.
13. R. C. Johnson, C. G. Miller, M. B. Neuworth, A. Talib: *Chemicals from Wood: The Policy Implications of Federal Study,* **vol. 2,** Mitre Corp., McLean, Va., 1983, Chap. 4.
14. L. K. Garlick: "Continuous Fermentation Using the Biostil System," *American Institute of Chemical Engineers Summer National Meeting,* Denver, Colo., August 31, 1983.
15. E. C. Donaldson: *Proceedings of 1982 International Conference on Microbial Enhancement of Oil Recovery,* Bartlesville, Okla., 1982.
16. "Emulsan Biopolymer Aimed at Oil Clean-Up Market," *Biomass Digest* **5** (1983) no. 7, 1 – 2.
17. "Lignin Copolymer Gains Water Solubility," *Biomass Digest* **5** (1983) no. 10, 3 – 4.
18. T. B. Reed, M. Graboski (ed.): *Proceedings: Biomass-to-Methanol Specialists Workshop, National Technical Information Service,* Springfield, Va., 1982.
19. "Anaerobic Digestion Development Moves Across a Broad Front: Use Spreads, Reactor Designs Command Interest," *Biomass Digest* **5** (1983) no. 9, 4 – 5.
20. B. O. Palsson, S. Fathi-Afshar, D. F. Rudd, E. N. Lightfoot: "Biomass as a Source of Chemical Feedstocks: An Economic Evaluation," *Science* **213** (1981) 513 – 517.
21. Reference 13, vol. 1, pp. 26 – 31.

Biopolymers → Cellulose
Biopolymers → Gelatin
Biopolymers → Lignin
Biopolymers → Nucleic Acids
Biopolymers → Proteins
Biopolymers → Starch
Biopolymers → Resins, Natural

Biotechnology

THOMAS BECKER, Institute of Process Analytics, University of Hohenheim, Germany

DIETMAR BREITHAUPT, Institute of Food Chemistry, University of Hohenheim, Germany

HORST WERNER DOELLE, Department of Microbiology, University of Queensland, St. Lucia, Queensland 4067, Australia

ARMIN FIECHTER, Institute of Biotechnology, Eidgenössische Technische Hochschule, Zürich, Switzerland

MARTIJN GRIENSVEN, Ludwig Boltzmann Institute, Wien, Austria

CORNELIA KASPER, Institute of Technical Chemistry, University of Hannover, Germany

STEPHAN LÜTZ, Institute of Biotechnology 2, Research Centre Jülich, Germany

RALF PÖRTNER, Institute for Bioprocess Engineering, Technical University of Hamburg-Harburg, Germany

HANS-GÜNTHER SCHLEGEL, Institute of Microbiology, University of Göttingen, Göttingen, Germany

DIETER SELL, DECHEMA e. V., Frankfurt, Germany

SAKAYU SHIMIZU, Department of Agricultural Chemistry, Kyoto University, Kyoto, Japan

FRANK STAHL, Institute of Technical Chemistry, University of Hannover, Germany

KIRSTIN SUCK, Institute of Technical Chemistry, University of Hannover, Germany

ROLAND ULBER, Institute of Bioprocess Engineering, University of Kaiserslautern, Germany

JOACHIM WEGENER, Institute of Biochemistry, University of Münster, Germany

KERSTIN WÜRGES, Institute of Biotechnology 2, Research Centre Jülich, Germany

HIDEAKI YAMADA, Department of Agricultural Chemistry, Kyoto University, Kyoto, Japan

HOLGER ZORN, Institute of Food Chemistry, University of Hannover, Germany

1. Introduction 3
2. Basics in Microbiology 5
2.1. Microbiology – the Science of Microscopic Life Forms 5
2.2. Phylogeny and Taxonomy of Microorganisms 7
2.2.1. Definition and Survey 7
2.2.2. Bacteria 7
2.2.3. Archaea 8
2.2.4. Morphological and Physiological Properties for the Identification of Prokaryotic Species 9
2.2.5. Eukaryotic Microorganisms 11
2.2.5.1. Definition 11
2.2.5.2. Fungi 11
3. Metabolism 13
3.1. Microbial Systems Biology 14
3.2. Energy Production 14
3.3. Substrate Transport 15
3.4. Catabolism 16
3.4.1. Photosynthesis 16
3.4.2. Chemosynthesis 16
3.4.3. Carbohydrate Metabolism 17
3.4.4. Aerobic Processes 18
3.4.5. Fats and Fatty Acid Metabolism . . 20
3.4.6. Hydrocarbon Metabolism 20
3.4.7. Amino Acid Metabolism 21
3.4.8. Anaerobic Metabolic Processes . . . 21
3.4.9. Single-Carbon-Compound Metabolism 21
3.4.10. Inorganic Metabolism 22
3.5. Biosynthesis 23
3.5.1. Amino Acids 23
3.5.2. Lipids 23
3.5.3. RNA and DNA 24
3.6. Regulation 24
4. Metabolic Engineering 24
4.1. Analysis of the Transcriptome, Proteome, and Metabolome 25
4.1.1. Gene Expression Analysis using DNA Microarrays 26
4.1.2. Fabrication of DNA microarrays . . 27
4.1.3. Proteome Analysis using Protein Microarrays 27

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_107.pub2

4.1.4. Metabolome Analysis and Metabolite flux analysis 28
4.2. Design and Production of Genetically Optimized Strains for Production – In Vitro Mutagenesis 30
4.3. Random Mutagenesis, Isolation and Selection of Mutants 31
4.4. Types of Mutants and Selection Principles 32
4.4.1. Auxotrophic Mutants 32
4.4.2. Regulatory Mutants 32
4.4.3. Other Selection Methods 32
4.4.4. Targeted or Site-Directed Mutagenesis 34
5. Cultivation and Bioprocesses . . . 35
5.1. Isolation of Microorganisms . . . 35
5.2. Requirements for Growth 36
5.2.1. Chemical Composition of Bacterial Cells 37
5.2.2. Carbon and Energy Sources 37
5.2.3. Accessory Nutrients 38
5.2.4. Sulfur and Nitrogen 38
5.2.5. Oxygen 38
5.2.6. Complex Media 38
5.2.7. Solid Media 38
5.2.8. Hydrogen Ion Concentration 38
5.2.9. Carbon Dioxide 38
5.2.10. Aeration 38
5.2.11. Anaerobic Techniques 39
5.2.12. Media Preparation 39
5.3. Sterilization 39
5.3.1. Moist Heat 39
5.3.2. Dry Heat 39
5.3.3. Filtration 39
5.3.4. Irradiation 40
5.3.5. Chemical Means 40
5.4. Types of Bioprocesses 40
5.4.1. Surface Culture 40
5.4.2. Submerged Culture 40
5.5. Process Layout 41
5.5.1. Reactors 41
5.5.2. Containments for Anaerobic Processes 41
5.5.3. Reactors for Aerobic Processes . . . 41
5.5.4. Inoculation 41
5.5.5. Operation Modes 42
5.6. Process and Product Overview . . 45
6. Biocatalysis and Biotransformation 45
6.1. Introduction 45
6.2. Classification of Biocatalysts . . . 47
6.3. History 47
6.4. Characteristics of Enzyme Reactions Used in Biotransformations 50
6.5. Types of Biocatalysts and Reaction Systems 51
6.5.1. Biotransformation with Growing Cultures 51
6.5.2. Biotransformation Conversion with Previously Grown Cells 52
6.5.2.1. Vegetative or Washed Cells 52
6.5.2.2. Permeabilized Cells 52
6.5.2.3. Dried Cells 52
6.5.3. Biotransformation with Spores . . . 52
6.5.4. Biotransformation with Immobilized Cells 53
6.5.5. Biotransformation with Cell-free Enzymes or Purified Enzymes . . . 54
6.5.6. Multistep Reactions Using Different Biocatalysts 54
6.5.7. Multiphase Reaction Systems 54
6.6. Process Design 55
6.6.1. General Considerations 55
6.6.1.1. Evaluating Enzyme Potential 55
6.6.1.2. Finding Suitable Enzymes 57
6.6.1.3. Substrates 57
6.6.1.4. Media 58
6.6.2. Selection of Biocatalysts 58
6.6.2.1. Screening 58
6.6.2.2. Enrichment 58
6.6.2.3. Molecular Engineering 59
6.7. Improvement of Conversion Processes 59
6.8. Conclusion and Outlook 60
7. Downstream Processing 61
7.1. Sample Disruption 62
7.2. Solid–Liquid Separations 63
7.3. Product Recovery 64
7.4. Solvent Extraction 65
8. Monitoring and Modeling of Bioprocesses 65
8.1. Characteristics of Bioprocesses . 66
8.1.1. System Definition 66
8.1.2. System Description 69
8.1.3. Dynamics of Biosystems and Real-Time Considerations 71
8.2. Biotechnological Measurement Systems 73
8.2.1. Process Requirements Concerning Measuring Quantities 74
8.2.2. Online Sensing Devices 75
8.2.3. Further Aspects Concerning Measuring Systems 79
8.3. Cognitive Computing 80
8.3.1. Fuzzy Logic Systems 82

8.3.2. Artificial Neural Networks (ANN) . 85
8.4. Modeling Aspects of Biological Systems 88
8.4.1. Steps in Creating a Model 88
8.4.2. Reasons for Making a Model 91
8.4.3. Different Types and Basic Approaches for Building a Model . 93
9. Special Applications in Biotechnology 96
9.1. Mammalian Cell Culture Technology 96
9.1.1. Introduction 96
9.1.2. Products from Mammalian Cells . . 98
9.1.3. Cell Types 99
9.1.4. Growth Medium for Cell Culture . 101
9.1.5. Small-Scale Culture Systems for Routine Use 102
9.1.6. Types of Bioreactors 103
9.1.7. Process Strategies 107
9.1.8. Downstream Processes 108
9.1.9. Regulatory and Safety Issues 108
9.2. Tissue Engineering 109
9.2.1. Application of Tissue Engineering . 109
9.2.2. Principle of Tissue Engineering . . 110
9.2.3. Strategies 111
9.2.4. The Essentials 111
9.2.5. Cells 111
9.2.6. Biomatrices 111
9.2.7. Bioreactors for Tissue Engineering 112
9.2.8. Growing New from Old 112
9.3. Biotechnology and Food 113
9.3.1. Production of Food Additives by Cell Culture Systems 113
9.3.1.1. Amino Acids 113
9.3.1.2. Organic Acids 114
9.3.1.3. Vitamins 114
9.3.1.4. Sweet Compounds 115
9.3.1.5. Sugar Alcohols 115
9.3.1.6. Microbial Saccharides 116
9.3.1.7. Conjugated Linoleic Acids (CLA) . 116
9.3.1.8. Lactulose 116
9.3.2. Enzyme-Catalyzed Processes 117
9.3.2.1. Starch-Modifying Enzymes 117
9.3.2.2. Lipases 118
9.3.2.3. Pectin-Degrading Enzymes 118
9.3.2.4. Chymosin (Aspartic Protease) . . . 119
9.4. Biotechnology and Health 120
9.4.1. Individualized Medicine 120
9.4.2. Clinical Diagnosis as Indicated in Genetic Anomalies in Cancer 120
9.4.3. Pharmaceutical Development 120
9.4.4. Define Molecular Mechanisms of Toxicity 122
9.4.5. Detection of Genetically Modified Organisms 122
10. Concluding Remarks 122
11. Acknowledgement 123
12. References 123

Biotechnology can be regarded as one of the key technologies of the 21st century. It is the commercial application of living organisms such as bacteria, fungi, yeasts, plant cells, viruses, and mammalian cells or their products, which involves the deliberate manipulation of their DNA molecules. This article gives an introduction into the basics in microbiology and provides an exhaustive description of the relevant microbial species and metabolic pathways. Identification, analysis, and manipulation of the genome, proteome, and metabolome is described, and cultivation requirements as well as process parameters discussed. Biotransformation and enzyme technology plays a central role in industrial biotechnology, and a focus is given on the development of molecular engineering techniques and new screening methods. Computational Biochemistry comprises the definition, monitoring, and modeling of bioprocesses. Examples of biotechnology applications include mammalian cell technology, tissue engineering, and the production of relevant food additives as well as of various medical and pharmaceutical products. In conclusion, biotechnology offers manifold possibilities in industrial and medical applications.

1. Introduction

Biotechnology can be described as "the commercial application of living organisms or their products, which involves the deliberate manipulation of their DNA molecules". This includes the use of bacteria, fungi, yeasts, plant cells, viruses, and mammalian cells. This definition implies a set of laboratory techniques developed within the last 20 years that have been responsible for the tremendous scientific and commercial interest in biotechnology, the founding of many new companies, and the redirection of research efforts and financial resources

among established companies and universities. Thus, biotechnology can be regarded as one of the key technologies of the 21st century. However, the use of microorganisms and their products is as old as mankind. Fermented products with yeast such as beer, wine, or sake are known since several thousands of years (Table 1). The roots of modern biotechnology are coming from the era of microbiology, which was developed in the late 19th century. Cagniard-Latour, Pasteur, Koch and Kühne were important scientists who are decoding the principles of fermentation. The use of microorganisms for the production of bulk chemicals such as butanol or acetone and for pharmaceuticals (antibiotics) was strongly developed during World War I and II. With the better understanding of the gene function and its relation to the metabolisms of cells, the modern era of biotechnology started in the 70th of the last century.

Nowadays, biotechnology is a cross-sectoral technology that has been successfully applied in many industrial branches. One distinguishes between several areas or "colors" of biotechnology (the so-called red, white, green, and blue biotechnology).

Red Biotechology. Named as red biotechnology are applications in medicine or in the pharmaceutical industry. Red biotechnology has already made an impact on healthcare and will continue to contribute to improving human health and life expectancy. Today, 20 % of marketed medicines, 50 % of those in clinical trials, and 80 % in early development are biotech-based products. Forty percent of these candidate medicines are for the treatment of cancer. Typical products of red biotechnology are recombinant vaccines, antibodies, blood clotting agents, and hormones. In addition, tissue engineering is a part of the area of biotechnology. Tissue engineering aims at the functional regeneration of tissues through implantation of tissue cultured in vitro (see Section 9.2).

White Biotechology. The potential and applications of biotechnology in other sectors largely pass unnoticed especially in the chemical industry. The chemical industry has produced coal, gas, and petroleum-based goods and products for over 150 years. This era is likely to end within the next 100 years at the latest. An extrapolation of oil consumption from 2002 (3.3 billion tonnes), for example, yields a supply period of 50 years with estimated stocks of 165 billion tonnes. The underlying reasons for a future era of new raw materials are the finite nature of fossil resources on the one hand, and the connection with many environmental problems on the other hand. In the past few years conditions for the application of biotechnological processes in industrial production have improved

Table 1. Timetable of biotechnology [1, 2]

Year	Biotechnological progress
3000 b.c.	Fermentation of sugar containing juices to various alcoholic beverages
2800 b.c.	brewing rooms in Mesopotamia and leaven in Egypt
1500 b.c.	Use of microorganisms for the production of copper and production of soya sauce
300 b.c.	Use of vinegar
1300	Micro algae (*Spirulina*) as food additives (Aztecs)
1400	Production of saltpetre (potassium nitrate) with *Nitrosomas sp.* and *Nitrobacter sp.* in Germany
1676	Anthony van Leuwenhoek observes bacteria through a microscope
1837	Charles Cagnaird-Latour identifies yeast as causer of fermentation (one year later confirmed by Theodor Schwann and Friedrich Kützing)
1849	Industrial production of bakers yeast
1856	Louis Pasteur separates brewer yeasts from lactic bacteria
1876	Kühne creates the term "enzymes"
1881	Industrial microbial production of lactic acid (Boehringer)
1890	Development of first vaccines by Pasteur and Koch
1893	Industrial production of citric acid using *Aspergillus niger*
1897	Starting point of enzyme technology (Eduard Buchner)
1915	Patent for enzymes in washing powder
1916	Industrial fermentation process for acetone and n-butanol (Chaim Weizmann)
1928	Alexander Fleming discovers penicillin
1937	Microbial production of vitamin C
1941/42	Industrial production of antibiotics
1957	Use of *Corynebacterium glutamicum* for the production of amino acids
1972/1973	Stanley Cohen and Francis Boyer develop a procedure for in vitro recombination of DNA, using plasmid vectors
1977	First recombinant proteins from bacteria
1982	First transgenic plants and animals
1985	Development of polymerase chain reaction (PCR) by Kary Mullis
1990	Start of human genome project (HUGO)
1996	Yeast genome is completely sequenced
2000	Human genome is completely sequenced (Craig Venter)

(so-called *white biotechnology*). New tools such as screening methods and metabolic engineering, and also global analysis methods such as genomics, proteomics, metabolomeics, and bioinformatics are gradually becoming more widely available. These new instruments make it possible to reduce the time needed to develop and establish new industrial biotechnological products and processes; this was hitherto one of the major drawbacks of biotechnological, as opposed to chemical, processes. In addition, they help to develop biocatalysts (enzymes) and microorganisms which render manufacturing processes more economical and facilitate new manufacturing processes. For the first time since the beginning of the oil age in the early fifties, one is able to apply processes with economic potential in the production of basic chemicals and biopolymers based on biotechnological processes. In many cases, bioprocesses operate under milder conditions (lower temperatures and pressures, etc.) and more selectively than their competitors (chemical processes) by these means bioprocesses conserve resources and improve production processes economically and ecologically. Public debate has often emphasized the advantages for the environment; however, companies face hard economic competition so that when a biotechnological process is weighed against a classical chemical process, only the potential economic advantages can affect a shift in favour of biotechnology.

Green Biotechnology. In addition, to overcome the problems in the production both of bulk and fine chemicals but also in the production of vaccines and pharmaceuticals, the so-called green biotechnology can offer new solutions. This area of biotechnology involves the introduction of foreign genes into economically important plant species, resulting in crop improvement and the production of novel products in plants. Green biotechnology might also produce more environmentally friendly solutions than traditional industrial agriculture. An example of this is the engineering of a plant to express a pesticide, thereby eliminating the need for external application of pesticides.

Blue Biotechnology. In a very broad sense, marine or blue biotechnology can be understood as the various means of techniques of managing marine living systems for mankind's profit. The domain covered by marine biotechnology is vast and range over various overlapping disciplines, from developmental biology to chemistry of natural substances and bioprocess engineering. Not all these fields, however, are ready for practical and industrial applications. Biomass from fishing or aquaculture industry is, in fact, complex, geographically and seasonally dependent. Furthermore, many "natural substances", for which we do not know of any terrestrial counterparts, and, therefore, present an up most interest, exists only in tiny amounts in rare biological species and their exploitation is likely to call for costly synthesis procedures. Marine natural products not only display novelty but also complexity in terms of chemical structures. Isolation and structure elucidation represents only the emerged part of an iceberg. The question of the functionality of new isolated molecules within the perspective of challenging major public health and environmental problems is crucial. Likely, in this domain, ecological and evolutionary approaches should help the classical screening systems for determining the right target systems. In addition, a better understanding of the complex interactions between macro- and microorganisms is necessary to be able to use these resources for industrial purposes. Thus, more and more groups are focussing on the part of bioprocess engineering and downstream processing in blue biotechnology [3].

2. Basics in Microbiology

2.1. Microbiology – the Science of Microscopic Life Forms

Microbiology deals with microscopically tiny life forms which the resolution of the human eye (approx. 20 µm) cannot detect. Many of the single-cell microorganisms, such as yeasts, algae, or protozoa, do not reach this size. Not until the invention of the microscope (Antonie van Leeuwenhoek, 1684) was the detection, and thus also the characterization, of microorganisms possible [4].

If microorganisms are defined by their size, many varied life forms come into this category: single-cell life forms with a real cell nucleus

(e.g., algae, fungi, protozoa) and also some without a real cell nucleus (bacteria and archaea) which are of particular interest in general microbiology. Organisms with a genuine cell nucleus are termed eukaryotic organisms, those without prokaryotic organisms [5]. Today some 6500 species of prokaryotic bacteria and archaea are known, a relatively small number compared with the number of the known species of eukaryotic organisms. However, it is estimated that the number of actually living species of prokaryotic organisms is much higher, although most of them are regarded as not cultivable under laboratory conditions. Analyses of the metagenomes of the most varied habitats (total DNA assays of samples from various environments, such as forest soil, compost, pond water, etc.) have yielded estimated values of up to one billion potential species of prokaryotic organisms worldwide.

For most people, microorganisms have a bad image. The reason is not hard to find: some microorganisms can cause disease in man and animals (pathogenic microorganisms). Thanks to the achievements of *medical microbiology*, the often devastating effects of pathogenic microorganisms in earlier times have been confined due to the development of vaccines, antibiotics, etc. [6]. Nevertheless, microorganisms still represent a threat to health (for instance to people with a weakened immune system). The following are a few examples of bacteria that are pathogenic to humans: *Bacillus anthracis* (anthrax), *Bordetella pertussis* (whooping cough), *Clostridium botulium* (botulism), *Clostridium tetani* (tetanus), *Corynebacterium diphtheriae* (diphtheria), *Shigella dysenteriae* (shigellosis), *Vibro-cholerae* (cholera). The development of resistance to known antibiotics is a phenomenon that documents the variability of pathogenic microorganisms [7]. This also explains why it is necessary to promote progress in the field of combating infectious diseases with microbiological methods.

Biotechnology is closely linked to microbiology. Biotechnology uses microbiological metabolism performance to manufacture industrially relevant products. The sector with the oldest tradition of applying microorganisms for production purposes is the food industry. The manufacture of beer, wine, vinegar, bread, and cheese are early achievements of human history and pertain to the field of *food microbiology*. Microorganisms are used in the production of sausages (particularly salami), sauerkraut, and milk products (lactobacilli). In Asia, a variety of fermented food products are produced on the basis of rice and soya [8, 9].The typical aromatic properties and consistency of many food products originate from the activity of the microorganisms used. Nowadays, these organisms are generally used in the production process as starter cultures. The fermentation products are also often used as preservatives (acidification in lactic acid production by lactobacilli or alcohol production by yeasts).

The monitoring of food in the framework of quality control involves excluding undesired microbial infestation or providing evidence of it.

Microorganisms contribute towards stabilizing the global natural material cycles (e.g., carbon and nitrogen cycles). Furthermore, the waste flows produced by man are largely reintroduced into the natural material flows through the activity of microorganisms. The field of *environmental microbiology* addresses the question of how microorganisms can be utilized for the remediation of contaminated soil, water, and waste gases. Successful industrial processes have been developed and are used worldwide for wastewater and drinking water purification, brownfield clean-up, composting, and waste fermentation and also for the purification of contaminated gases [10 – 12].

Industrial microbiology, or white biotechnology, deals with the production of industrially interesting substances involving production volumes of thousands of tons. Effective production of substances requires on the one hand an efficient production strain to supply as much target product as possible in the shortest time possible. On the other hand, it needs a technical facility where the high-performance microorganisms find optimum living and, therefore, production conditions and where in addition monoseptic cultivation in large volumes is guaranteed. Bioreactors fulfill these requirements: they can be operated with volumes of up to several hundred m^3 and are equipped with automatic measurement and control systems adjusted to the growth needs of the microorganism [13, 14].

Examples of microbially produced industrial products are diverse. The list begins with microorganisms that can themselves be the prod-

uct of a biotechnological process and reach the market as starter cultures [15]. Currently the quantitatively dominant product is ethanol, of which over 24×10^6 (metric) tons are produced worldwide. Ethanol is produced with the yeast *Saccharomyces cerevisiae* and is mainly used as a substitute fuel or in admixture with other fuels. Other products include pharmacologically active agents (e.g., antibiotics, steroids, and alkaloids), bulk, special, and fine chemicals, food and feed additives, chiral intermediates, enzymes, antibodies, etc.

The total value of all biotechnologically produced substances worldwide (including biopharmaceuticals which are often manufactured with cultures of animal cells) must be in the realm of $> 100 \times 10^9$ euros annually.

2.2. Phylogeny and Taxonomy of Microorganisms

2.2.1. Definition and Survey

Taxonomy is the branch of biology which subsumes individual groups of species into a hierarchical system. For a long time, the kingdom was the highest category of creatures. Originally, a distinction was made between animals and plants. Later, the bacteria were added and then the fungi and plants were split up [16]. Finally the archaea were given their own kingdom. In recent years, genetic investigations have led to a new classification with the "*domain*" superimposed over the kingdom. Whereas in earlier times the aim was to understand and deduce relationships among microorganisms on the basis of morphological, physiological, and cytological similarities, in the last few decades genetic similarity has been the criterion used to identify relationships directly from the genetic material. Nowadays, sequence analysis of ribosomal RNA (rRNA), which was developed by Carl Woese, is one of the methods by which phylogenetic trees can be generated [17].

Probably the best-known gene in the world is the 16S rRNA (S, Svedberg unit for characterizing the behavior of sedimentation) of prokaryotic organisms. This gene is composed of around 1500 base pairs and in the course of evolution was only subjected to relatively minor changes through mutations. The reason is that ribosomes are the sites of protein biosynthesis and the molecular mechanism taking place is extremely sensitive to radical mutations. Thus, there are areas within 16S rRNA that are extremely well preserved, but also some where base exchanges have been more frequent. An evolutionary gap between two organisms can be determined from the relative similarity of different genes for the 16S rRNA of various prokaryotic organisms; this can be used as a measure of the degree of relationship [18].

Analogous to the investigations on prokaryotic organisms, the characterization of the degree of relationships of eukaryotic organisms is derived from their 18S rRNA, a molecule with approximately 2300 bases.

Besides comparing the base sequences of rRNA, the amino acid sequences of universal proteins may also be drawn on to establish relationships. This does not always involve the same branchings of the phylogenetic tree as with the comparison of rRNAs. This explains why taxonomy, which owes its motivation primarily to the potential inherent in the new molecular biology tools, is currently a highly dynamic area of science.

Today, all living organisms on earth are classified into three domains: archaea, bacteria, and eukaryota. The following representation of the manifold life forms of microorganisms adheres to this phylogenetic classification.

2.2.2. Bacteria

Bacterial cells are small, generally less than 2 μm in diameter and 2–5 μm long (there are only a few giants of 5×20 μm dimensions). The cells do not contain a membrane-enveloped nucleus; their DNA is rather a circularly closed double strand lying within the cytoplasm. The bacterial DNA contains the total genetic information of the cell. In many bacteria there are additional DNA circles, the plasmids; however, they are not required for growth under ordinary culture conditions. The ribosomes are small (sedimentation constant in the ultracentrifuge: 70 S), and their function is sensitive to various antibiotics that do not act upon the cytoplasmic ribosomes of eukaryotic organisms. In contrast to eukaryotic cells, bacterial cells do not contain organelles,

but often polyphosphates, poly(β-hydroxybutyric acid), or glycogen as storage substances (Fig. 1). Most bacteria are surrounded by a cell envelope that contains peptidoglycan (murein). Many bacteria are motile by either flagellar or gliding movements.

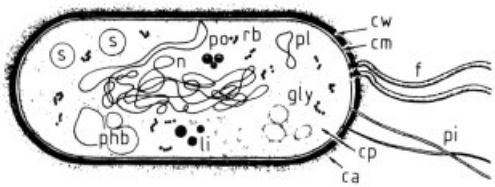

Figure 1. Longitudinal section through a bacterial cell
ca) capsule; cm) cytoplasmic membrane; cp) cytoplasm; cw) cell wall; f) flagellum; gly) glycogen; li) lipid droplets; n) nuclear material or bacterial chromosome; phb) poly(β-hydroxybutyric acid); pi) pili; pl) plasmid; po) polyphosphate; rb) ribosomes; s) sulfur granules

Classification of Bacteria According to the Phylogenetic System. Originally, the subdivision of bacteria was based on their morphology and physiology, in the last few years, however, their classification has been completely revised and established on a molecular biology level. Thus, the taxonomy of bacteria, especially the higher levels of classification, cannot yet be regarded as definitive. Some authorities believe that the degree of variance between different bacterial groups is sufficient to give them each "kingdom status" of their own. Thus, in some publications reference is made to 13 kingdoms of bacteria, in others to "phyla" instead of kingdoms.

The classification presented is derived from Bergey's Manual of Determinative Bacteriology [9], which the bacteria are subdivided into the following phyla:

- Proteobacteria
- Gram-positive bacteria
- Cyanobacteria
- Chlamydia
- Plantomyces
- Bacteroids/Flavobacteria
- Green sulfur bacteria
- Spirochetes
- Deinococci
- Green non-sulfur bacteria
- Hyperthermophiles (3 kingdoms)

Bearing in mind all the postulated, noncultivable species, the actual number of bacterial phyla (or kingdoms) is probably far higher. Estimates assume that there will ultimately be about 50 phyla (or kingdoms) of bacteria. The current phylogenetic tree, therefore, can only be regarded as a snapshot.

The bacterial phyla with the highest number of species known today are the proteobacteria, Gram-positive bacteria and cyanobacteria. Cyanobacteria are phototrophic organisms closely related to Gram-positive bacteria. Gram-positive bacteria form a group of chemo-organotrophic bacteria. The phylum of proteobacteria is the largest and physiologically most varied of all bacteria. Representatives of this phylum account for the majority of known Gram-negative bacteria of medical- and application-oriented interest [20].

Even the possibly best-known bacterium, *Escherichia coli*, belongs to this phylum. This example is used to demonstrate the taxonomic classification of an organism:

Domain:	Bacteria
Phylum:	Proteobacteria
Class:	Gamma protobacteria
Order:	Enterobacteriales
Family:	Enterobacteriaceae
Genera:	Escherichia
Species:	Escherichia coli

The binomial nomenclature is used throughout biology. Each organism has a genus name and a species name and is thus specified unambiguously.

2.2.3. Archaea

Several properties distinguish archaea from bacteria and eukaryota. Archaea are single-cell organisms mainly with a closed DNA molecule lying within the cytoplasm. They are of the same size as bacteria. Archaea have neither a cytoskeleton nor cell organelles, they are distinguished from bacteria by their lack of peptidoglycan and the different structure of their ribosomes. With over 200 species they are often found where extreme conditions prevail.

Archaea have cell walls of pseudopeptidoglycan and single-layer cell membranes formed from ether lipids with covalent bonded chains. Many species of archaea have adapted to their

extreme surroundings, hence some species prefer temperatures of over 80°C (thermophile), others live in salt waters (halophile) or in very acidic or alkaline environments (acidophile/alkaliphile) [21].

Enzymes from such extremophile organisms, which are not only found among archaea, are particularly interesting for industrial applications; indeed they seem predestined for use in industrial processes with high temperatures and salt concentrations [22]. In the area of amylases thermotolerant enzymes are already applied for starch saccharification. Moreover the enzymes used in detergents (proteases, lipases, amylases, cellulases) are active and fulfill their functions at temperatures of 60°C and above [23].

The phylogenetic system of the archaea that is valid at present subdivides the domains into the following phyla [19]:

- Euryarchaeota
- Crenarchaeota
- Korarchaeota
- (Nanoarchaeota is proposed as an additional phylum)
- Archaeota

The *Euryarchaeota* are a diverse group. This phylum comprises not only methanogenic species, which are strictly anaerobic organisms, but also strictly aerobic organisms. Furthermore a large group of hitherto noncultivable organisms belong to this phylum and similarly to that of the *Crenarchaeota*, which include both hyperthermophilic representatives and mesophilic marine species. It is only very recently that representatives of the *Korarchaeota* phylum have been cultivated successfully.

2.2.4. Morphological and Physiological Properties for the Identification of Prokaryotic Species

The properties of prokaryotic organisms as presented here permit the identification of species by conventional methods if molecular biology techniques are not available or cannot be applied. The shapes of most species are similar, and there are only a few morphological characteristics suitable for differentiation. In contrast, the diversity of metabolic features is tremendous. Prokaryotic organisms are extremely versatile. Therefore, a great number of physiological and biochemical properties have to be determined to describe and identify a species unambiguously.

Properties and description of prokaryotic organisms:

1) *Shape.*
   The majority of prokaryotic organisms are spheres, rods, or helices; these organisms are called cocci, rods, and spirilla, respectively (Fig. 2). Some organisms are encapsulated by proteins or polysaccharides, and some are combined to form packets, filaments, or consortia.
2) *Flagellation.*
   The kind of flagellation, length, type of undulation, and mode of attachment of flagellae are important characteristics of prokaryotic organisms (Fig. 3).
3) *Gram Stain.*
   The Gram stain, originally devised by CHR. GRAM (1884) to make bacteria visible in animal tissues, is a reliable and fundamental cellular characteristic by which the bacteria are divided into two parts: Gram-positive and Gram-negative bacteria. It is based on the retention or discoloration of bacteria by ethanol after a staining procedure involving an aniline dye and iodine. In Gram-positive bacteria, the dye is retained by a cell wall consisting of a multilayered network of peptidoglycan; in Gram-negative bacteria, the dye is readily washed out because the cell wall consists of only two peptidoglycan layers and a highly permeable outer membrane envelope. The basic structural differences are easily visible in ultrathin sections (Fig. 4).
4) *Endospores.*
   Endospores are highly refractile, thermoresistant stages in the life cycle of two groups of Gram-positive bacteria belonging to the genera *Bacillus* and *Clostridium*. Endospores are easily visible and tolerate pasteurization (treatment by moist heat for 10 min at 80 °C). Heating in an autoclave for 20 min at 120 °C is required to kill the spores.
5) *Need for Oxygen.*
   Many prokaryotic organisms grow only in the presence of atmospheric oxygen. They are *strictly aerobic* organisms. In contrast, *strictly anaerobic* organisms can only grow in the absence of oxygen. *Facultatively anaerobic* or-

ganisms can grow under both conditions. The last group includes aerotolerant species which tolerate but do not use oxygen. Others use oxygen if it is available. Some aerobic species are microaerophilic, i.e., they grow at low partial pressures of oxygen but not in equilibrium with air.

6) *Energy Generation.*
The production of energy, i.e., regeneration of adenosine triphosphate (ATP), occurs via three fundamental processes, namely, fermentation, respiration, and photosynthesis. The strictly anaerobic fermentative prokaryotic organisms rely only on fermentation. All strict aerobes regenerate ATP by oxidative phosphorylation with oxygen as the electron acceptor. Mechanistically, energy generation by denitrifying, sulfate-reducing, methanogenic, or acetogenic species is quite similar to aerobic respiration and is therefore designated as anaerobic respiration. All photosynthetic microorganisms contain light-absorbing pigments, such as chlorophyll derivatives and carotenoids. The processes by which ATP is regenerated are similar in respiration and in photosynthesis; both are described as electrophosphorylation.

7) *Effects of Temperature and pH.*
The majority of organisms are *mesophilic:* they grow at their maximum rate between 20 and 42 °C. *Thermophilic* bacteria and archaea reach their maximum growth rates between 40 and 80 °C. A few of them are able to grow above 80 °C and up to temperatures of over 100 °C; they are called *extremely thermophilic* organisms. The *psychrophilic* (or kryophilic) species prefer temperatures below 20 °C. Most organisms tolerate pH values between 5 and 8 but prefer neutrality; a few are acid-tolerant, acidophilic, or alkali-tolerant, alkaliphilic.

8) *Nutritional Types.*
All organisms that can use light energy for growth are called *phototrophic*. In contrast, the term *chemotrophic* denotes energy conversion by oxidation – reduction reactions that involve either fermentation or respiration. *Organotrophic* denotes the utilization of organic compounds, and *lithotrophic* the utilization of inorganic compounds (ammonia, nitrite, hydrogen sulfide, sulfur, hydrogen, carbon monoxide, iron(II)) as reductants. *Autotrophic* characterizes the ability to synthesize the majority of cellular constituents from carbon dioxide, and *heterotrophic*, the derivation of cellular constituents from organic compounds. The designations are usually combined, e.g., *Chromatium okenii* is photolithoautotrophic. Some organisms require accessory nutrients, such as vitamins or amino acids, for growth.

9) *Natural Habitats.*
The ecological habitat is the place where an organism or a population can usually be found. Examples are marine water, the sediment of a lake, fertile soil, or the intestinal tract.

10) *Symbiotic or Parasitic Relationships.*
The close relationship between two dissimilar organisms is called symbiosis. It may be favorable for both (mutualistic) or unfavorable for one (parasitic).

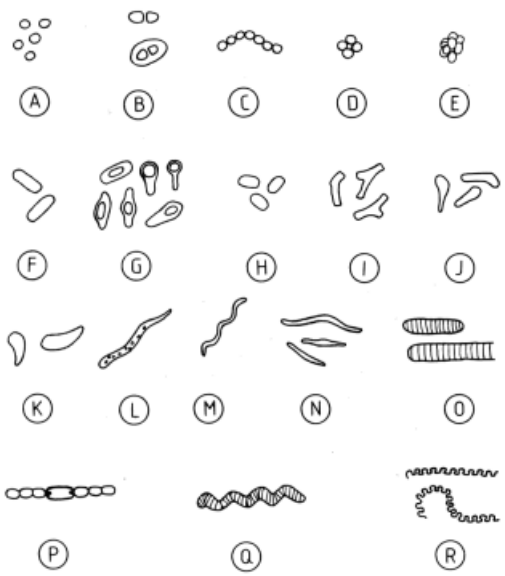

Figure 2. Shapes of bacteria
A) Cocci; B) Diplococci, without and with capsule; C) Streptococci; D) Tetrads; E) Package cocci (*Sarcina*); F) Rods; G) Spore-forming bacteria, differing with respect to shape and localization of endospores in the mother cell; H) Short rods, coccobacilli; I) Coryneform bacteria; J) Mycobacteria; K) Vibrio-like bacteria; L) Spirillum (*Thiospirillum*); M) Spirilla (*Aquaspirillum*); N) Spindle-shaped bacteria; O) Trichomes of filamentous cyanobacteria; P) Filamentous cyanobacteria with heterocyst; Q) Filamentous helically shaped cyanobacterium (*Spirulina*); R) Spirochaetes (*Spirochaeta plicatilis*)

11) *Composition of Cellular Macromolecules.*
The description of bacteria includes the Gram type of the cell wall, the type of peptidoglycan structure, the fraction of cytosine + guanine in the DNA, the similarity index of the 16 S rRNA, and data on DNA–DNA hybridization.
12) *Antibiotic Resistance.*
The characterization of an organism may be supplemented by its pattern of antibiotic resistance.

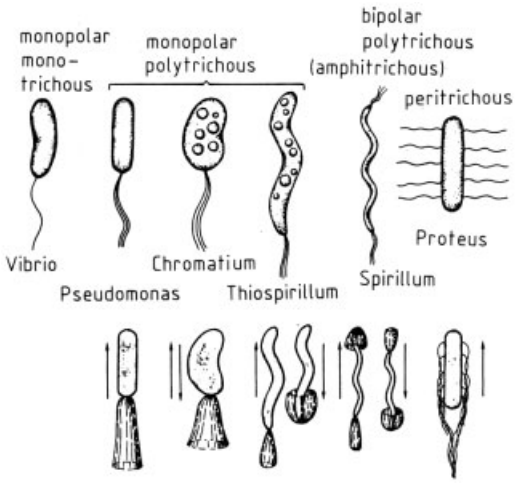

Figure 3. Flagellation of bacteria, flagellar movement

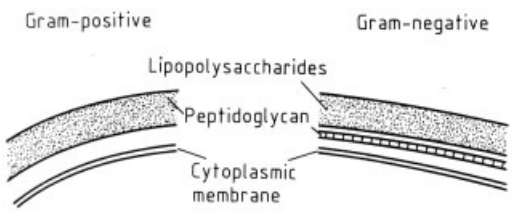

Figure 4. Gram-positive and Gram-negative bacterial cell walls

2.2.5. Eukaryotic Microorganisms

2.2.5.1. Definition

The cells of most eukaryotic organisms are usually much larger than prokaryotic cells and have diameters of 10 µm or more [24]. They contain several structural components, compartments, and organelles. Figure 5 shows a plant cell as an example of a eukaryotic cell. The nucleus, consisting of a set of chromosomes that divide by mitosis, is surrounded by a double membrane. The DNA is associated with histone protein. Mitochondria (for respiratory energy conversion) and chloroplasts (in plants for photosynthetic energy conversion) contain a prokaryotic type of DNA and ribosomes. Cytoplasmic ribosomes are large (80 S). Endoplasmic membranes and various kinds of membrane vesicles compartmentalize the cell; they are involved in the ingestion of nutrients (endocytosis) and the production and excretion of proteins and particles (exocytosis).

An internal skeleton, the cytoskeleton, consisting of contractile protein (actin) filaments and microtubules, gives the cell its shape and confers ability to move and to transport membrane vesicles within the cell. The cells of eukaryotic microorganisms can be motile by either ameboid or flagellar movement.

Most eukaryotic cells and organisms are aerobic. Energy for growth is derived from respiration or photosynthesis.

Eukaryotic microorganisms include the protozoans, algae, and fungi. Fungi are the most relevant representatives of eukaryotic microorganisms in biotechnology and are used to produce antibiotics, secondary metabolites, and also organic acids, vitamins, etc. For this reason, fungi will be considered here in more detail.

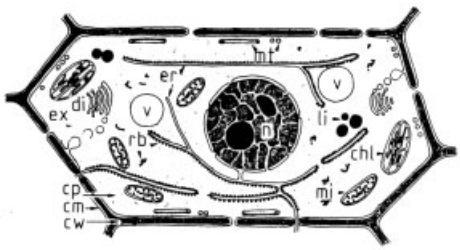

Figure 5. Longitudinal section of a eukaryotic plant cell
cw) Cell wall; chl) Chloroplast; cm) Cytoplasmic membrane; cp) Cytoplasm; di) Dictyosome; er) Endoplasmic reticulum; ex) Secretion vesicle for exocytosis; li) Lipid droplet; mi) Mitochondrion; mt) Microtubule; n) Nucleus or karyon; rb) Ribosomes; v) Vacuole

2.2.5.2. Fungi

Fungi are aerobic eukaryotic organisms; they form a separate kingdom within the domain of the eukaryota [25, 26]. There are unicellular fungi, but the majority form filaments (hyphae), (usually 5–10 µm in diameter) and grow as masses of hyphae (mycelia). Many fungi form

fruiting bodies. The shapes of fungi are diverse, and fungi are usually identified on a morphological basis. They are much less versatile metabolically than bacteria, and metabolic properties are almost useless for identification. Fungi have cell walls that often consist of chitin or cellulose. The group is rich in species, their number exceeding 100 000. Some estimates assume that 95 % of fungi have not yet been described.

Growth and Reproduction. Fungal hyphae grow at their tip (Fig. 6). Each part of the mycelium is potentially able to grow, and a small piece of a mycelium can serve as an inoculum for growth. There are two kinds of reproduction, asexual and sexual.

A sexual reproduction occurs by spore formation, bud formation, or fragmentation. Conidiospores are formed on the tips of hyphae (conidiophores), e.g., in the genera *Aspergillus* and *Penicillium*. If the spores are formed within special vessels, the sporangia, they are called sporangiospores, e.g., in the genera *Mucor* and *Rhizopus*. Bud formation is the mechanism of asexual reproduction among the yeasts.

Sexual reproduction involves mating of two nuclei (karyogamy), zygote formation, and meiosis. It usually results in the formation of spores. In the lower fungi the zygote is usually transformed to a durant organ which after meiosis forms a sporangium. In the higher fungi, the zygote nucleus divides meiotically, and spores are formed either within a sac (*ascus*) or by budding on top of the basidia; the spores formed are called ascospores and basidiospores, respectively.

The taxonomic classification of fungi is currently under debate and different suggestions can be found. The subdivision presented herein closely follows a traditional classification und contains the following classes: basidiomycetes, ascomycetes, zygomycetes, deuteromycetes, myxomycetes.

Basidiomycetes. In basidiomycetes, the zygote enlarges to form a club-shaped cell, the basidium. The basidiospores are formed from projections of the basidium usually resulting in four spores on the tip. Most species are terrestrial fungi and live as saprophytes or parasites; many form mycorrhizas on trees. Nutrient media and methods have been developed to grow several hundred species and strains of these higher fungi, which include many edible mushrooms, in submerged culture. Some could be of interest in biotechnology.

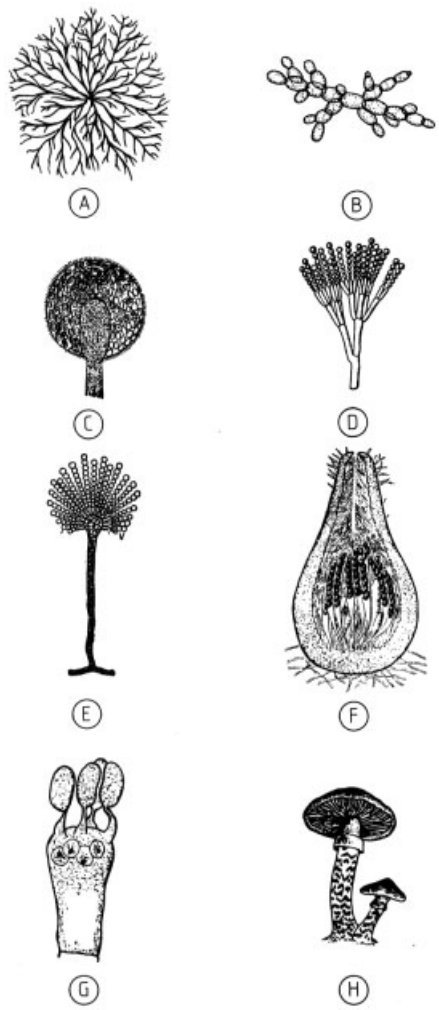

Figure 6. Mycelium, conidiophores, and fruiting bodies of fungi
A) Mycelium of fungal hyphae; B) Yeast dividing by budding; C) Sporangium of *Mucor mucedo* containing sporangiospores; D) Sporophore of *Penicillium* forming conidiospores on the tips of branched hyphae; E) Sporophore of *Aspergillus* forming conidiospores on the tips of hyphae (sterigmata) located on the spherical end of the sporophore cell; F) Fruiting body (perithecium) of a typical ascomycete containing asci and ascospores; G) Basidium with four basidiospores; H) fruiting body of a typical basidiomycete

Ascomycetes. The ascomycetes are named for their formation of spores within a sac-like zygote, the ascus. With the exception of a few unicellular forms they grow as branched mycelia with open cross walls; they are coenocytic. Many ascomycetes grow on excrements of animals, i.e., they are coprophilic, whereas others are parasites on plants or insects. Asexual reproduction occurs by various kinds of conidiospores.

The order **Saccharomycetales** within the Ascomycetes comprises all yeasts. Typical yeasts, such as *Saccharomyces cerevisiae*, are unicellular. They grow either as diplonts (with diploid nucleus) or as haplonts (with haploid nucleus), such as *Schizosaccharomyces pombe*. All yeasts are able to grow on sugar substrates, and a few can even grow on alkanes (*Saccharomycopsis, Candida lipolytica*) or on methanol (*Candida boidinii*, *Hansenula anomala*). Yeasts are organisms often used in industrial microbiology.

Zygomycetes. The zygomycetes are lower fungi and comprise unicellular and mycelial fungi; they have diverse nutritional habits. They grow in water, soil, or moist habitats. There are saprophytic and parasitic species. The most important group biotechnologically is that of the Zygomycetales, which include *Mucor mucedo*, *Rhizopus nigricans*, and *Phycomyces blakesleeanus*. Several species are used for the production of vitamins, carotenoids, organic acids, and enzymes.

Myxomycetes. The myxomycetes or slime moulds grow on decaying plant tissue, form multinuclear plasmodia, and produce fruiting bodies of yellow, red, or brown color. There is not yet a consensus on the exact evolutionary affinities of myxomycetes, but these organisms constitute a well-defined, homogeneous group of approximately 900 species.

Deuteromycetes. If a fungus is incapable of sexual reproduction, it cannot be classified with either the ascomycetes or basidiomycetes, hence a third group of higher fungi was created. There are many deuteromycetes of biotechnological importance: *Aspergillus*, *Penicillium*, and *Cephalosporium*. Investigations of the 18S rRNA revealed that the majority of deuteromycetes could be assigned to the basidiomycetes and ascomycetes so that the group of deuteromycetes will possibly be abandoned in the future [27, 28].

3. Metabolism

The decisive factor for the development of a successful biotechnological production is an efficient biocatalyst. If this biocatalyst is a microbial production strain, knowledge of its metabolic capabilities is very important. Such knowledge can lead to an increase in product concentration and product variety, because one catalyst can often be used to produce different products [29].

Metabolism is generally concerned with all the events occurring within a cell that not only keep the cell alive but are also manifested in growth [30]. All synthetic reactions leading to growth are energy-consuming (endergonic) and are commonly referred to as *anabolic reactions*. In order to carry out these reactions, each individual synthesis must be coupled with energy-producing reactions of the *catabolic sequence*. Metabolism therefore consists of an intricate coupling between anabolism and catabolism (Fig. 7), whereby energy transactions and transfer mechanisms are based upon the basic laws of thermodynamics [31, 32].

Furthermore, in order to obtain a functional biosynthesis, the microbial cell must also produce a reducing agent, because biosynthesis involves the reduction of small molecules of high oxidation state to larger molecules of lower oxidation states. Both the energy and the reducing agent can be obtained from any of a number of diverse reactions depending upon the growth conditions and the genetic system available [33, 34] within the metabolism. Energy is generally dependent on adenosine triphosphate (ATP) and the reductant on nicotinamide dinucleotide NADH or NADPH.

The cell has a very sophisticated regulatory control system that avoids oversynthesis of any chemical compound, thus securing its integrity. These regulatory control systems start at the cell membrane (substrate uptake) [35] and are present throughout the pathways of catabolism and anabolism.

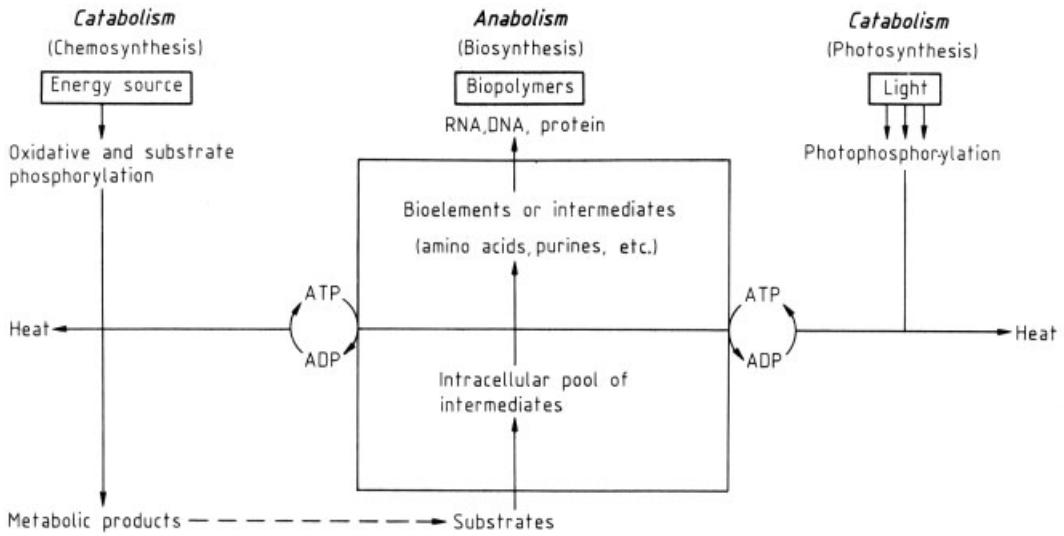

Figure 7. Anabolic and catabolic reactions in microbial metabolism

3.1. Microbial Systems Biology

In the future, systems biology will enable the metabolism of some microorganisms to be modeled and mapped in silico to such an extent that physiological performance and effects caused by environmental changes will be predictable. The essential prerequisites include sufficient knowledge of metabolic material flow and network analysis, reaction kinetic analysis of biochemical networks (enzyme kinetics, global metabolism regulation), global analysis of gene regulation networks (signal transduction), analysis of populationwide systems (quorum-sensing networks), and the abstraction of fundamental regulation patterns. The first such model that integrates signal transduction, gene expression and metabolism has already been introduced for the response of yeast to osmotic shock [36].

There is no doubt that such models signify great progress for microbiology and biotechnology: quantitative simulations and targeted interventions in the metabolic network ideally open up the way to highly specialized, optimized production organisms. In order to optimize specific types of metabolic performance or to accelerate desired biosyntheses, industrial biotechnology is already relying to a great extent on the metabolic engineering of microorganisms. Here, mention should be made of the synthesis of amino acids [37, 38], optimized riboflavin production in *B. subtilis* [39], and, in the case of terpenoids, yield increases by engineering mevalonate biosynthesis in *E. coli* [40]. It is likely that knowledge from systems biology will trigger tremendous progress in this application-oriented field. Moreover, strategies based on systems biology will also be important for biotechnological process development where an understanding of the complex regulatory networks of production organisms will be indispensable for process optimization [41].

3.2. Energy Production

Redox reactions are among the most important chemical reactions in living organisms. These oxidation – reduction reactions are of special importance for energy production by living organisms:

$$\Delta G^0 = n\,F\,\Delta E^0$$

where ΔG^0 is the standard free energy of the reaction, n is the number of electrons (or hydrogen atoms) involved, F is the Faraday constant, and ΔE^0 is the potential difference between the two redox systems. The simplest way to think about these oxidation – reduction reactions is in terms of electron donors and acceptors. Thus, the substrate oxidized has a specific redox potential that is at the lower end of the electrode

potential scale, whereas the final electron acceptor has a potential towards the more positive end of the scale. The difference between the two potentials corresponds to the energy obtainable in a chemical reaction between them. If the potentials of the substrate (electron donor) and of the final electron acceptor are known, one can easily determine the energy production in a microbial system (Fig. 8) [42, 43].

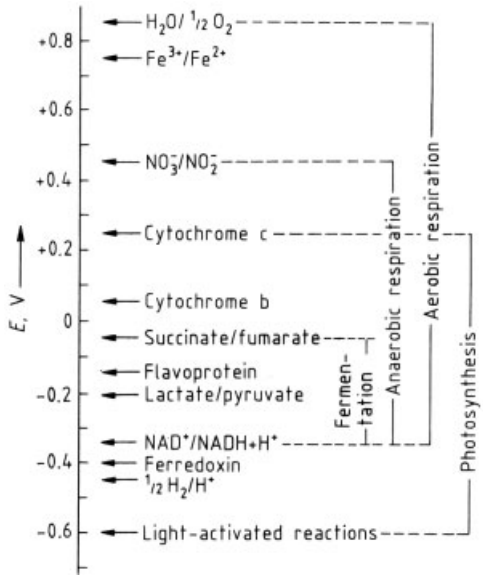

Figure 8. Redox potentials of various energy-producing microbial reactions

Because the potential difference between electron donor and acceptor can be significant, the energy released could produce so much heat that the cell would be damaged. In order to avoid such cell damage, the cell has an energy-release-controlling cascade system, which makes certain that energy is released only stepwise. This cascade system is referred to as the respiratory chain or the electron transport chain. It is located in the cellular membrane and is responsible for the production of energy by electron transfer involving an enzyme called ATPase (which forms or hydrolyzes ATP) [35, 44]. The ATPase also plays an important role for the transportation of substances across the membrane into the cell.

According to the chemiosmotic hypothesis for ATP production, the electron transfer chains are coupled to ATP synthesis by a proton electrochemical gradient ($\Delta\mu_H^+$) across the energy-transducing membrane. Proton electrochemical potential is a thermodynamic measure of the extent to which the proton gradient across the membrane deviates from equilibrium. This hypothesis is based on two separate proton pumps, whereby the $\Delta\mu_H^+$ is generated by electron transfer and is used to drive the second pump, involving the enzyme ATPase, backward in the direction of ATP synthesis. The final result of substrate oxidation is therefore the generation across the cell membrane of gradients of both pH (ΔpH) and electrical potential (ΔE). Both gradients exert a force on the protons extruded by the respiratory chain, tending to pull them back across the membrane. This proton motive force, $\Delta\mu_H^+$, is the key element:

$$\Delta\mu_H^+ = \Delta E - 2.303\, RT/F \cdot \Delta\mathrm{pH}$$

The numerical value of 2.303 *RT/F* is 59 mV at 25 °C.

3.3. Substrate Transport

Substrates undergoing catabolic reactions must enter the cell before catalysis can occur, because most catalytic enzymes are inside the cell. The cell membrane is a barrier for most ions and other molecules. For an ion to be transported across a membrane both a pathway and a driving force are required. Driving forces can be concentration gradients, electrical potentials, metabolic energy, or a combination of these. Four processes are known [45, 46]:

1) In *simple diffusion*, solutes move passively across the cell membrane, depending on the concentration gradient and thermal motion of the molecules (e.g., water, gases, low molecular hydrophilic compounds, organic acids in the protonated form).
2) In *facilitated diffusion*, solutes require a carrier for their transport in addition to the concentration gradient and thermal motion of the molecules. In both simple and facilitated diffusion, no metabolic energy or proton motive force is required. The energy for the transport stems from existing concentration gradients.
3) The term *active transport* is applied to a tight coupling of transport to metabolism in the ion pumps, which are central to chemiosmotic energy transduction. There are at least two

distinct classes of active transport systems: the membrane-bound and the binding protein transport systems.

4) The *group-translocation transport* system comprises processes in which the passage of the substrate across the membrane occurs simultaneously with, and as a consequence of, chemical transformation of the substrate (e.g., phosphoenolpyruvate–glucose–phosphotransferase system with most anaerobic and facultatively anaerobic bacteria). Active transport and group-translocation transport are often referred to as "primary transport", simple and facilitated diffusion as "secondary transport". The carrier systems involved can be of the "uniport", "symport", or "antiport" type.
   Many bacteria pump natrium ions out of the cell by taking in two protons per natrium ion in an antiport system. Sulfate ions, on the other hand, are only taken in by some transport systems if protons are simultaneously "symported".

3.4. Catabolism

3.4.1. Photosynthesis

Photosynthesis creates living matter out of inorganic material, replenishes the reservoir of oxygen in the atmosphere in the cases of algae and plants, stores the energy of sunlight to support the life activities of organisms, and removes carbon dioxide:

$$CO_2 + H_2A + \text{light} \rightarrow (CH_2O) + 2\,A + \text{chemical energy}$$

If A is taken to be oxygen the process is plant (or algae or cyanobacteria) photosynthesis; if it represents a sulfur atom the process is bacterial photosynthesis [43].

The process of photosynthetic energy conversion is initiated when photons are absorbed by specific molecules, such as chlorophyll. This absorption leads to an ejection of electrons, which are accepted by the compound having the lowest redox potential in animate nature, namely ferredoxin. The electrons then move through the electron transport system, and the energy is finally converted to ATP. Because the electrons return to the reaction center (Fig. 9), this system is called cyclic photophosphorylation. The reducing agent is produced separately. In plants and algae, a second photosystem exists that splits water into H^+ and oxygen, whereby the hydrogen ions reduce $NADP^+$ and oxygen is set free. In bacteria, however, such a second photosystem does not exist, and inorganic compounds such as thiosulfate are oxidized, with the electrons running through the electron transport system to the reaction center and the proton via ferredoxin to NAD^+ (Fig. 9). This is the reason why plants and algae possess mainly cyclic and bacteria noncyclic photophosphorylation.

3.4.2. Chemosynthesis

In chemosynthesis, all energy and reducing agents are obtained by catalytic reactions of organic or inorganic compounds. In the former case, one refers to *chemoorganotrophy* and in the latter to *chemolithotrophy*. In addition, three different energy modes, which depend upon the electron donors and acceptors, are distinguished: aerobic respiration, anaerobic respiration, and fermentation [47].

Organic substances are used as an energy source by the vast majority of microorganisms that live in natural environments (Fig. 10). Nature provides them with an abundance of large organic polymers. A polymeric substance consists of series of monomeric units linked together. These monomeric units can be of the same kind (e.g., cellulose, starch) or of different kinds (e.g., pectin, lignin, sucrose). Within the polymeric structure, the monomeric units can be linked in a straight chain with identical linkages or different linkages or in branched chains; moreover, they can be linked in a specific sequence or at random. The structure of each of the polymer substrates for microorganisms is very important, because the enzymes produced by the microorganisms are specific not only to the types of monomers, but also in most cases to the way in which the monomers are linked.

Since microorganisms can only take up monomers, all enzymes involved in the breakdown of polymeric substances must be extracellular. These enzymes are either membrane-bound or released into the medium. In general terms, Gram-negative bacteria accommodate these enzymes in the periplasmic space between the cell membrane and the cell wall and

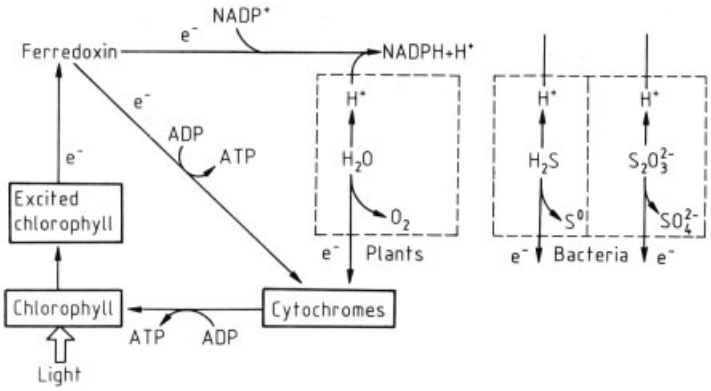

Figure 9. Schematic representation of cyclic and noncyclic photophosphorylation in nature

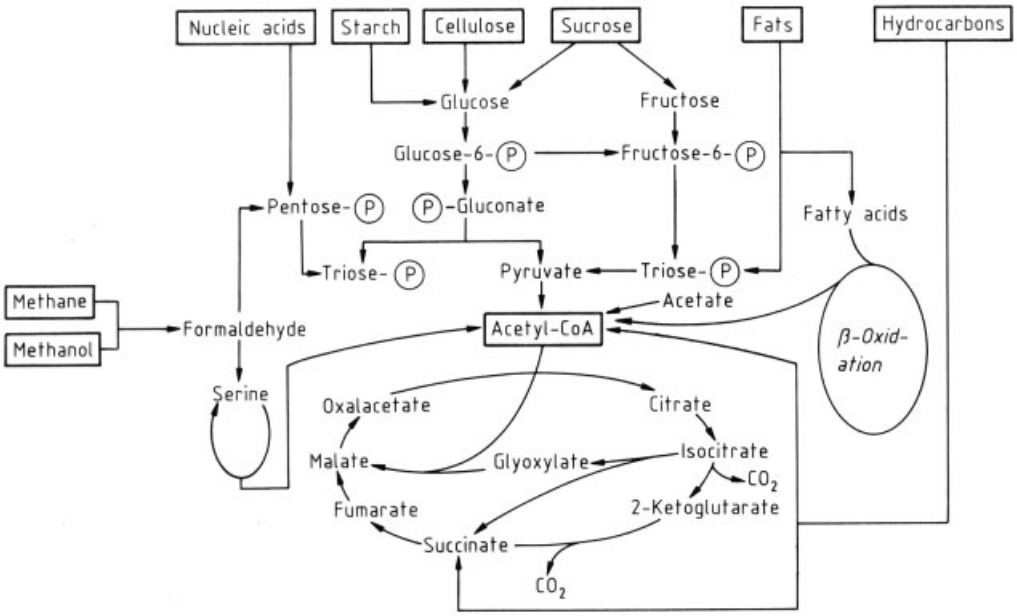

Figure 10. Schematic representation of aerobic catabolism in microorganisms

the enzymes are thus membrane-bound, whereas Gram-positive bacteria and fungi have a tendency to release these enzymes into the medium. This is one of the reasons why extracellular enzyme studies are carried out predominantly with Gram-positive microorganisms, as well as yeasts and other fungi.

3.4.3. Carbohydrate Metabolism

Most renewable resources are carbohydrates. The principal carbohydrate that serves as a carbon and energy source for microorganisms is glucose. However, not all microorganisms follow the same route of glucose utilization. There are at least four major pathways of glucose metabolism, of which three lead directly to pyruvate:

1) *Embden-Meyerhoff* (EM) pathway, often referred to as the glycolytic or fructose bisphosphate pathway
2) *Hexosemonophosphate* (HMP) pathway, often referred to as the pentose shunt or pentose phosphate pathway
3) *Entner-Doudoroff* (ED) pathway

4) *Phosphoketolase* (PK) pathway, almost entirely specific for heterofermentative lactic acid bacteria

All four pathways of glucose metabolism have a great number of intermediates and enzymes in common. The details of these pathways can be found in many textbooks [32, 34, 48].

The EM pathway, for example, provides the greatest amount of ATP, but it does not produce ribose-5-phosphate, the important precursor for RNA and DNA biosynthesis; nor does it produce erythrose-4-phosphate, which is important for amino acid biosynthesis. Microorganisms that are capable of using only the EM pathway for glucose utilization are therefore not able to grow on simple media with glucose as the sole carbon source. They require growth factors or organic compounds (e.g., yeast extract) for growth and are referred to as fastidious microorganisms.

In contrast, the HMP pathway produces all the precursors necessary for both RNA and DNA as well as for aromatic amino acid biosynthesis, but it produces only half the amount of ATP energy. This pathway does not produce pyruvate directly. The microorganisms must therefore possess part of the enzymes of the EM pathway. It is therefore not surprising that both pathways, EM and HMP, are common combinations, particularly in facultative anaerobic microorganisms.

The ED pathway is unique, however, and has so far only been found in bacteria. Although this pathway is linked partly to the HMP pathway in the reverse direction for precursor formation, pyruvate is formed directly because of the aldolase cleavage of 3-ketodeoxy-6-phosphogluconate. The ED pathway can exist on its own and is used by the majority of strictly aerobic microorganisms. The net result is similar to the HMP pathway, although one mole of ATP can be formed only if the carbon atoms go into pyruvate instead of precursor.

3.4.4. Aerobic Processes

All aerobic microorganisms use the tricarboxylic acid (TCA) cycle as a main metabolic pathway. In this cycle, appropriate groups of enzymes catalyze a series of consecutive transformations, including oxidations, which finally result in the complete oxidation of pyruvate to carbon dioxide. The electrons and hydrogen atoms removed from the individual organic compounds are accepted by oxygen and form the other two end products, ATP and water. The tricarboxylic acid cycle, however, is not only an energy-producing cycle, but it is also responsible for the production of precursors for amino acid and antibiotic biosyntheses [49, 50]: 2-ketoglutarate and oxalacetate ions. The tricarboxylic acid cycle serves, therefore, as the central pathway for the production of energy as well as for anabolic precursors.

Under ideal growth conditions, the intermediates of the tricarboxylic acid cycle would be continuously withdrawn for anabolic reactions. Then, the formation of oxalacetate at the end of the cycle became vulnerable. Because oxalacetate is required not only as a precursor for anabolic amino acid biosynthesis, but also for the condensation with acetyl-CoA to keep the tricarboxylic acid cycle alive, the organism must have some safety device to ensure that oxalacetate is always available; otherwise the tricarboxylic acid cycle would be interrupted. The organism possesses a very fine control system for energy production and biosynthetic requirements in form of replenishment sequence reactions. Altogether, the microbial world has five enzymes at its disposal for such sequences, of which phosphoenolpyruvate carboxylase is probably the dominant: it catalyzes the formation of oxalacetate from phosphoenolpyruvate and carbon dioxide. Phosphoenolpyruvate, in turn, is the intermediate from which pyruvate is formed in the EM pathway.

This double function of the tricarboxylic acid cycle as an energy-generating cycle as well as a biosynthetic precursor-supplying cycle is very often referred to as amphibolic. This dual function, of course, must be controlled. There exists one enzyme in the tricarboxylic acid cycle, isocitrate dehydrogenase (it catalyzes the reaction: isocitrate $\rightarrow$ oxalosuccinate), which performs this regulatory control. In the case of an overproduction of ATP this enzyme is inhibited by ATP, preventing any further oxidation and ATP synthesis, until the biosynthetic steps leading to RNA, DNA, aromatic amino acids, and fatty acids have caught up and utilized the excess ATP [33].

The general scheme in Figure 11 outlines the intricate interconnections between catabolism

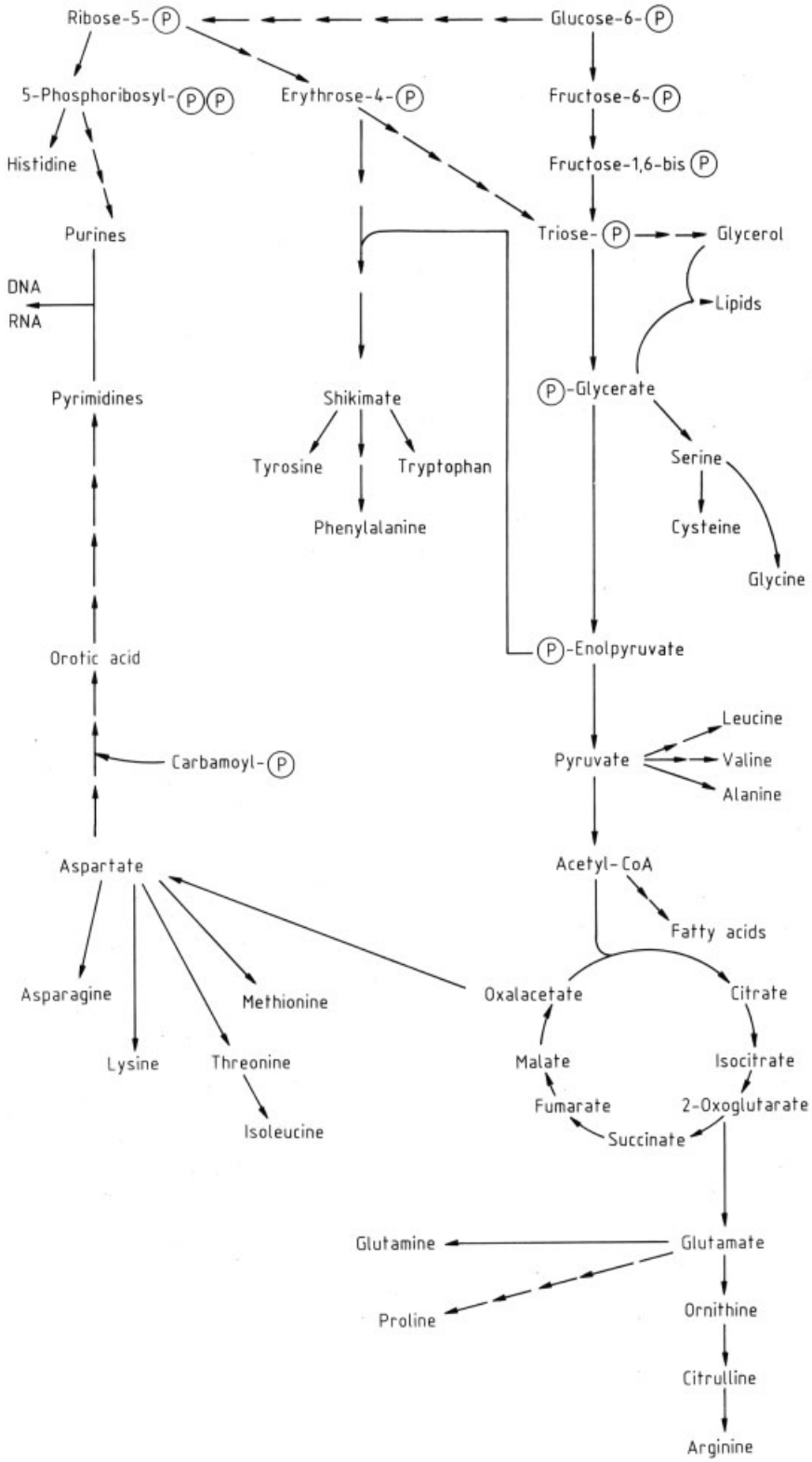

Figure 11. General scheme of interconnections between catabolism and anabolism (cell biosynthesis)
Ⓟ= Phosphate

and anabolism among all those precursors necessary for cellular biosynthesis. Aerobic processes in biotechnology are used to produce various products, for example, antibiotics, amino acids, and organic acids [29, 51, 52].

3.4.5. Fats and Fatty Acid Metabolism

Fats are another major group of naturally occurring compounds (see Fig. 10). The majority of fats are triglycerides, that is, fatty acids (R^1, R^2, R^3) esterified with glycerol:

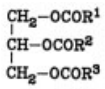

These fats have to be converted into the monomeric components, the fatty acids and glycerol. This hydrolysis is carried out by enzymes called lipases. The glycerol component is converted to dihydroxyacetone phosphate, which participates in the EM pathway. The fatty acids, which vary in their carbon chain length, have to be activated by the formation of their respective coenzyme A (CoA) esters; this reaction step requires energy. These fatty-acid-CoA esters subsequently undergo cyclic oxidation, eliminating an acetyl unit at each turn. This type of oxidation is referred to as *β-oxidation*. The acetyl-CoA formed after each cycle can then enter the tricarboxylic acid cycle for the formation of energy and biosynthetic precursors.

However, this type of metabolism leading to acetyl-CoA does not produce the important ribose-5-phosphate and erythrose-4-phosphate. In order to obtain these precursors, all organisms whose metabolism leads directly into the tricarboxylic acid cycle via acetyl-CoA and not via pyruvate have to build up the carbon chain mentioned above and have introduced two pathways for this biosynthetic purpose:

1) The *glyoxylate cycle*. As mentioned earlier, overproduction of ATP inhibits the enzyme isocitrate dehydrogenase. The organism is then capable of inducing two new enzymes, isocitrate lyase and malate synthase, to circumvent the carbon dioxide-producing steps of the tricarboxylic acid cycle. This new shortcut TCA cycle leads to succinate and via glyoxylate to malate, and oxalacetate is formed despite the inhibition of isocitrate dehydrogenase (see Fig. 10).
2) The *gluconeogenic pathways*. From the oxalacetate formed, the organism is using one of the earlier mentioned replenishment sequence reactions to produce phosphoenol-pyruvate and reverses the EM pathway to glucose-6-phosphate; from there ribose-5-phosphate and erythrose-4-phosphate can be formed via the forward HMP pathway.

A great number of enzymes are common to both carbohydrate and fatty acid metabolism, which shows the economy of bacterial metabolism. The principal differences are in the control systems, which are outlined in detail in the literature, for example, [32, 34].

3.4.6. Hydrocarbon Metabolism

The metabolism of hydrocarbons has been examined in detail in the framework of environmental biotechnology in connection with cleanup of contaminated sites. Aliphatic hydrocarbons are good substrates for a large number of microorganisms. The straight-chain alkanes are more readily attacked than substituted or branched-chain alkanes. The attack occurs at either one end (monoterminal) or both ends (diterminal). The alkane is converted via the primary alcohol into the corresponding fatty acid and then via β-oxidation into acetyl-CoA as outlined previously. Methyl group oxidation incorporates one atom of oxygen; this reaction is catalyzed by a mixed oxidase system consisting of three proteins: rubredoxin, NADH-rubredoxin reductase, and -hydroxylase. For the oxidation of methylene groups, rubredoxin is replaced by cytochrome P450 and -hydroxylase by methylene hydroxylase [53, 54].

Aromatic hydrocarbon utilization follows a uniform biochemical concept. The great majority of aromatic hydrocarbons, irrespective of the number of benzene rings, converge in their oxidative metabolism to three major intermediates: catechol, 3,4-dihydroxybenzoic acid (protocatechuic acid), and 2,5-dihydroxybenzoic acid (gentisic acid):

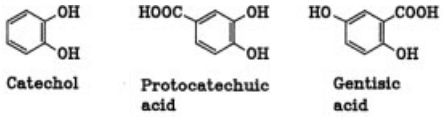

The benzene nucleus in these intermediates is cleaved either between adjacent hydroxyl groups (ortho cleavage) or between a hydroxyl and a carboxyl group (meta cleavage). In both cases, the active incorporation of an oxygen molecule is required. Ortho cleavage, via the 2-ketoadipate, and meta cleavage, via the keto acid pathway, lead into the tricarboxylic acid cycle at the level of acetyl-CoA, succinate, or malate and fumarate. As in fatty acid metabolism, the organisms use the tricarboxylic acid cycle for energy production and require the glyoxylate cycle together with gluconeogenesis to obtain pentose for RNA and DNA precursors.

3.4.7. Amino Acid Metabolism

The aerobic utilization of amino acids follows a pattern that is very similar to the metabolism of other carbon sources [29, 49]. The only difference is the -amino group. In most cases, the amino acid is first converted into the corresponding keto acid by either cytochrome-linked oxidases, transaminases, or NAD(P)-linked dehydrogenases. The final product of the metabolism always leads to the intermediates of the tricarboxylic acid cycle. This is one of the reasons why -amino acids are able to function as the sole source of carbon, nitrogen, and energy for many microorganisms. As in fatty acid and hydrocarbon metabolism, microorganisms growing on amino acids must possess the glyoxylate cycle and the gluconeogenesis pathway.

3.4.8. Anaerobic Metabolic Processes

The breakdown, or catabolism, of organic compounds in the absence of oxygen is generally referred to as *fermentation*. The microorganisms that carry out fermentations are either facultative or obligate anaerobes. The most characteristic difference between respiratory and fermentative metabolism is in ATP energy production. Because no electron transport occurs, redox reactions of organic compounds play a major role. The number of cells obtained per mole of substrate in simple defined media is much smaller than under aerobic conditions. In addition to the cell material, large amounts of organic end products are formed, mainly primary alcohols, e.g., ethanol or butanol.

Fermentations are normally classified according to their principal fermentation products. Figure 12 summarizes the end products from carbohydrates.

The distinction between aerobic and anaerobic carbohydrate metabolism is made at the pyruvate level, because in anaerobic fermentation the organism has to substitute for the tricarboxylic acid cycle. In general the organisms are not able to produce the intermediates that serve as precursors for amino acid and protein biosynthesis. Therefore, fermentation processes cannot be carried out on simple defined media, but always require complex media if the organism is to grow and maintain its metabolism. Most of these fermentation processes are rather complex and details may be obtained from the relevant literature [32, 34, 55].

3.4.9. Single-Carbon-Compound Metabolism

A few isolated genera of microorganisms are capable of oxidizing single-carbon compounds. Microorganisms are called *methylotrophs* if they have the ability to derive both carbon and energy from the metabolism of methanol, methanotrophs if they can do the same with methane [53, 56 – 58]. Methane and methanol were the most common substrates used in the production of single-cell protein. The oxidation of methane leads to methanol and from there to carbon dioxide.

The oxidation of methane or methanol to carbon dioxide, however, does not provide the organism with any precursors for biosynthesis. These organisms therefore possess an unusual carbon assimilatory pathway. Depending upon the enzymic assemblage, formaldehyde is incorporated into ribulose-5-phosphate or into the serine pathway in order to build up the carbon chain for the biosynthesis of RNA, DNA, and all the amino acids required for protein biosynthesis. The absence of a tricarboxylic acid cycle, the presence of anaplerotic sequence reactions, and

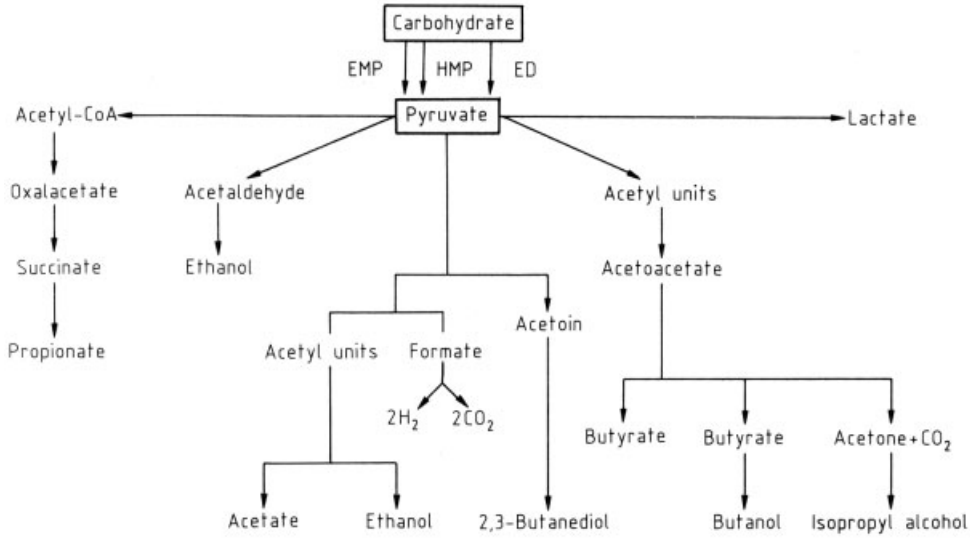

Figure 12. Fermentation end products of carbohydrate metabolism

the glyoxylate cycle demonstrates the biosynthetic capability of these pathways.

3.4.10. Inorganic Metabolism

Some microorganisms are capable of using inorganic compounds as an energy source in aerobic metabolism or as a final electron acceptor in anaerobic metabolism (anaerobic respiration). Both of these are combined in nature in the sulfur cycle and the nitrogen cycle, [32, 34].

Sulfur Cycle. Under anaerobic conditions, sulfate or thiosulfate can be reduced to sulfide by sulfate-reducing bacteria. Such processes occur mainly in water-logged soils or stationary estuaries rich in organic matter, and can cause serious corrosion of iron and steel pipes in the soil. Hydrogen sulfide is an extremely toxic product. If the soil or estuaries are supplied with oxygen, hydrogen sulfide can be reoxidized to sulfate by aerobic sulfur-oxidizing microorganisms, for example, thiobacilli. The reoxidation of sulfur or sulfide to sulfate is a microbial process of considerable biotechnological significance. It not only acidifies the soils and thus makes plant growth possible, but the sulfuric acid formed can also be used for ore leaching processes. Sulfuric acid can form soluble sulfates with a wide range of metals and by leaching old mining dumps it helps concentrate these for further utilization of the ore. Such ore leaching processes have led to a vastly improved mining industry and increased the economy of ore mining.

Nitrogen Cycle. Nitrogen is the most abundant gas in the atmosphere. Nitrogen-fixing organisms are capable of converting nitrogen into nitrogen compounds that are acceptable to plants. Nitrogen-fixing microorganisms can be divided into free-living (cyanobacteria, *Clostridium, Azotobacter*) and symbiotic (*Rhizobium*). The last group in particular may become of great biotechnological importance, because they live in association with legume plants and provide these with the necessary nitrogen source even in nitrogen-starved soils, thus eliminating the need for large amounts of nitrogen fertilizers. The nitrogen-fixing organisms convert nitrogen to ammonia, which then can be oxidized by chemolithotrophs (*Nitrosomonas* and *Nitrobacter*) via nitrite to nitrate. This conversion of nitrogen to nitrate is referred to as *nitrification*. The reverse reaction, called *denitrification*, is an anaerobic process in which the nitrate or nitrite serves as electron acceptor. The product of denitrification is nitrogen gas. Nitrification and denitrification are important process steps in biological wastewater cleaning.

The **chemolithotrophs**, however, face an energy problem. The redox potential of the inorganic compounds used as electron donors is well

above that of the NAD/NADH couple that is required as reducing agent for biosynthesis. Since the electrons can only enter the redox system at the cytochrome level, electron transport must be reversed to obtain NADH. In photosynthesis, this reversal can be carried out by the reaction center of the photosynthetic apparatus, but in chemosynthesis the reversal has to occur with energy input from the oxidation of the inorganic source. It is therefore not surprising that such organisms grow much better in the presence of organic material, which they are able to assimilate directly, thus obviating the reductive biosynthesis and consequently the need for reductant formation.

3.5. Biosynthesis

The microbial cell consists of five major types of macromolecules: proteins, polysaccharides, lipids, RNA, and DNA [32, 34]. These macromolecules are built up from relatively few monomers, which the cell has to provide or to synthesize: amino acids, sugar phosphates, fatty acids, ribonucleotides, and deoxyribonucleotides.

Microorganisms are divided into two groups: *autotrophs*, which are able to produce all their cellular requirements from carbon dioxide as the sole carbon source, and *heterotrophs* – the majority of microorganisms – which produce the cellular material from organic compounds. Photo- and chemolithotrophs use ATP and the reducing power produced by photosynthesis and/or oxidation of inorganic substrates to reduce carbon dioxide to cellular material. The pathway for carbon dioxide reduction is referred to as the Benson-Calvin cycle. The actual carbon dioxide fixation reaction is catalyzed by ribulose-1,5-bisphosphate carboxylase; the primary product of this reaction is 3-phosphoglycerate, which leads to glyceraldehyde-3-phosphate, a member of the EM pathway:

$$\begin{aligned} &3\,CO_2 + 9\,ATP + 6\,NADH_2 \\ &\quad \rightarrow \text{glyceraldehyde}-3-\text{phosphate} \\ &\quad \rightarrow + 9\,ADP + 8\,P_i + 6\,NAD^+ \end{aligned}$$

From glyceraldehyde-3-phosphate, these organisms are capable of producing the important pentoses, trioses, and the dicarboxylic acids required for macromolecular biosynthesis [46, 47].

3.5.1. Amino Acids

About twenty -amino acids are required for the biosynthesis of the numerous types of proteins that provide the catalytic capability of the microorganisms. These amino acids are produced from the following precursors:

Erythrose-4-phosphate	→ tyrosine, tryptophan
Phosphoenolpyruvate	→ phenylalanine
Ribose-5-phosphate	→ histidine
3-Phosphoglycerate	→ serine, glycine, cysteine
Pyruvate	→ alanine, valine, leucine
2-Oxoglutarate	→ glutamic acid, glutamine, arginine, proline
Oxalacetate	→ aspartic acid, asparagine, methionine, lysine, threonine, isoleucine

The most important amino acids are glutamic acid and glutamine, which figure in transamination reactions during biosynthesis. Therefore, the availability of these two amino acids is vital for protein biosynthesis. The detailed pathways can be found in the literature [32, 34, 46].

Theoretically, each of these -amino acids can be produced commercially by microorganisms, some are already produced, but some practical problems remain to be solved for some of them [49].

3.5.2. Lipids

Most of the fatty acids that occur in lipids contain 16 or 18 carbon atoms and are saturated or unsaturated with one or more double bonds. Acetyl-CoA is the only precursor for the biosynthesis of all fatty acids. In order to differentiate between fatty acid catabolism and fatty acid biosynthesis, the organisms employ CoA activation in the former and an *acyl carrier protein* (ACP) in the latter. The C_2 units are added to the acetyl-ACP in the form of malonyl-ACP. Once the required chain length is reached, ACP is hydrolyzed and the appropriate fatty acid is formed. Unsaturated fatty acids are generally formed by dehydration of a hydroxy fatty acid.

The fatty acids are then esterified with glycerol, which is readily available from dihydroxyacetone phosphate, a member of the EM path-

way. Finally, triglycerides are formed. The introduction of a phosphate group gives a phosphatidic acid and a whole range of phospholipids. If the phosphate group in the phosphatidic acid is replaced by a carbohydrate, glycolipids are produced [32, 34, 46].

3.5.3. RNA and DNA

Ribonucleotides consist of a purine or pyrimidine derivative, ribose, and phosphate groups. A purine or pyrimidine base attached to a ribose is referred to as a ribonucleoside; if a phosphate group is attached to this ribonucleoside it is called a ribonucleotide. The most important pyrimidine derivatives are uracil, cytosine, and thymine. The most important purine derivatives are adenine and guanine. The biosynthesis of pyrimidines and purines starts at the ribose-5-phosphate level, which also represents the ribose in the nucleotide. The RNA consists of ribonucleotides. Reduction of the ribonucleotides leads to deoxyribonucleotides, the building blocks for DNA biosynthesis [44, 46].

3.6. Regulation

The microbial cell possesses a complex of catabolic and anabolic capabilities with many interconnected pathways. The cell is required to maintain balance among the various parts of this extremely complex network of reactions [32, 34, 44, 46, 48]. The cell therefore has developed very intricate and refined devices to streamline its economy.

Regulation can either be manifested in the enzyme synthesis or in the enzyme activity. Induction and repression of enzyme synthesis occurs at the genetic code, whereas feedback inhibition simply regulates the activity of the enzyme. The importance for biotechnology is to realize that the first type of regulation is concerned with substrate, catabolite, and end product inhibition, whereas the second mainly involves a transitional end product inhibition.

More than one enzyme usually is required to channel a substrate into intermediary metabolism. The substrate can therefore be responsible for a coordinate or sequential induction of enzymes. This activity of the cell normally occurs during the lag phase of growth and can be reduced to almost zero by preculture conditions.

A much more complicated regulation is catabolite repression, which is mainly manifested in the so-called *diauxie phenomenon*. The latter is biphasic growth in the presence of more than one carbon source in the medium. In this reaction the organism takes up only one substrate at a time and the presence of this particular substrate represses the enzyme system that is responsible for the metabolism of the other substrate. This repression occurs at the operon level and occurs as soon as the concentration of the first substrate becomes very small.

Enzyme synthesis can also be inhibited by end products within a short period of time. For example, if the medium contains any α-amino acid, the organism would use this amino acid directly instead of producing it. The added amino acid therefore represses the synthesis of each enzyme in its synthesis pathway.

In addition to the regulation of enzyme synthesis, the cell must also adjust the *activity of enzymes* to its metabolic requirements. If a monomer is synthesized in larger amounts than needed for polymer synthesis, it is not necessary to stop the synthesis of the enzyme completely, but it suffices to reduce the activity of that particular enzyme. This fast control action is referred to as *feedback inhibition*. The target enzymes for feedback inhibition are called allosteric enzymes.

4. Metabolic Engineering

The term *metabolic engineering* refers to the genetic optimization of prokaryotic or eukaryotic cells that are used in industrial fermentation processes. In these processes, they are utilized for the economically production of proteins or other fine chemicals. Bailey [59] firstly introduced the definition of metabolic engineering in 1991 as "the improvement of cellular activities by manipulation of enzymatic, transport, and regulatory functions of the cell with the use of recombinant DNA technology." However, one of the early examples of successful genetic engineering of a living organism for production purposes, namely the production of human insulin in *E. coli* cells, had already been described

quite some time before a name was given to this new direction of science.

According to CAMERON and TONG [60], there are five categories of metabolic engineering that are grouped with respect to the objective of the genetic modification:

1) Improved production of chemicals already produced by the host organism.
2) Extended substrate range for growth and product formation.
3) Introduction of new catabolic activities for the degradation of toxic materials.
4) Production of chemicals that are new to the host organism.
5) Modification of cell properties, for instance, to make them more resistant to the production conditions.

The process of metabolic engineering is most often described by a circle of elementary steps that have to be performed once or several times. This circular process is demonstrated in Figure 13.

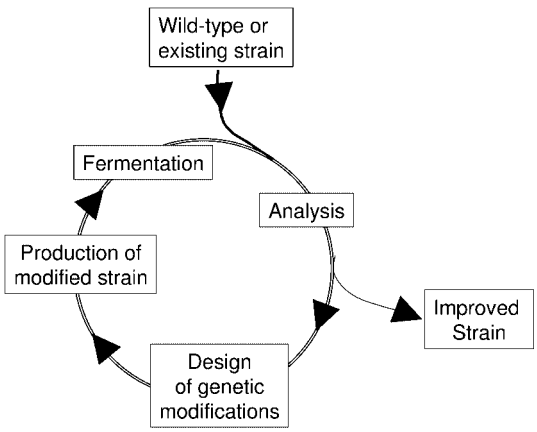

Figure 13. Schematic illustration of the elementary processes of metabolic engineering

The elementary processes are

– *Analysis* of the organisms with respect to the Proteome and/or metabolome
– *Design* of a genetic modification that improves/allows the synthesis of the compound of interest
– *Production* of a modified strain
– *Fermentation* and production

Depending on the task at hand, one enters the engineering cycle at different entry points. When, for instance, an organism should be modified to express a recombinant protein, the first step is to design and create the DNA sequence that codes for the protein in an appropriate vector system. Thus, one will enter the cycle at the *design* level. The same is true when an existing metabolic pathway should be extended in the way that the original end product of that pathway is metabolized to a compound of interest. However, when an existing metabolic pathway shall be used to produce a given compound in high yields, one will have to start by analyzing the metabolic situation and the metabolic fluxes in order to identify targets for genetic engineering. In this case, one will enter the cycle at the *analysis* level. In particular in the latter case, it is often necessary to go through the cycle of metabolic engineering more than once, since complex metabolic networks within living cells often require several genetic modifications to provide the compound of interest in improved quantities.

Metabolic engineering is to a high degree dependent on novel developments and improvements in chemical analytics as well as molecular biology. In both disciplines, the last decade has brought about powerful new technologies – such as protein and DNA chip technology to mention only two – that provided considerable progress in metabolic engineering. For further information, the interested reader is referred to the reviews by NIELSEN [61] and OSTERGAARD [62] and the website of a scientific journal that is exclusively devoted to this research area (www.apnet.com/mbe).

The following paragraphs will discuss in more detail the three elementary processes of metabolic engineering, analysis, design, and production of a modified strain. Fermentation will be addressed in other chapters of this series. (→ Ethanol, Chap. 5, → Antibiotics, Chap. 5)

4.1. Analysis of the Transcriptome, Proteome, and Metabolome

The bottleneck of all of the traditional methods of transcriptome and proteome characterization is the limitation to an analysis of just a few samples at a time. While conventional methods are confining themselves to the examination of single genes, a DNA or protein chip experiment

delivers a complete gene or protein expression pattern of the cell. The high degree of parallelization realized in biochips is the great advantage over classic molecular biological approaches.

4.1.1. Gene Expression Analysis using DNA Microarrays

In the past few years, the complete DNA sequences of a number of different microorganisms have been determined and can be exploited to optimize strains as well as recombinant protein production. Strain optimization involves measurement of genomewide mRNA levels in wild-type and mutant strains using DNA microarrays (Fig. 14) Furthermore, microarray analysis can help identifying previously unknown genes required for recombinant protein production.

Bioprocess optimization using microarrays entails metabolic control analysis, modeling, and molecular biology to create new mutants and strains with an, for example, optimized protein production rate. Recombinant protein expression exerts a metabolic burden on the host cell, whereby stress response mechanisms are triggered on different levels. Microarrays allow the investigation on a genome scale, which enables the qualitative and quantitative characterization of the burden on host cell metabolism. Thereby, a better understanding of the impact of recombinant protein production can be achieved by using the generated "snapshot" of the actual cellular composition and activity. Additionally, the knowledge of the interaction of host cell metabolism with recombinant protein production is improved and contributes to process optimization.

Chip Technology. Alongside metabolome analysis, analysis of various components of the proteome, the transcriptome or the genome will become increasingly valuable. New biosensors known as biochips will be required. DNA chip technology has already opened up new ways of studying disease in more depth and identifying far more possible targets. A DNA chip is an array of synthetic DNA sequences representing different genes. Up to now, membrane and immune biosensors are used to analyze substances of the metabolome. One important goal in chemical and biological sensing is the sensitive detection and selective identification of biochemical discrete compounds (carcinogens, metabolites, proteins, etc.) or living systems (bacteria, viruses, or related components) at low levels in complex biological matrices, tissues, and blood. In the future, it will be indispensable to analyze compounds of the genome and the proteome. Therefore, new formats of biosensors, so called biochips, have to be developed. Such biochips or microarrays can be used for identifying genes as well as changes in their activity. Furthermore, these biochips allow the simultaneous detection of several disease endpoints using different bioreceptors such as DNA, antibodies, enzymes, or cellular probes.

The high parallelization degree of biochips is a great advantage over classic molecular biological methods. While conventional methods are confining themselves to the examination of single genes, a DNA chip experiment delivers a complete gene expressions pattern of the cell. DNA microarrays can be used for the purpose of monitoring expression levels of thousands of genes simultaneously. Furthermore, biochips can solidly simplify and accelerate a number of long and expensive diagnostic methods. Additionally, gene expression analysis across various biological conditions, cell cycle states, tissues, and subjects may help identify differentially expressed genes. This type of information is a valuable pinpoint in the investigation of biological processes and functional disorders. Recapitulating, DNA chips provide a format to profile complex diseases and discover novel disease-related genes. Applications for this technology include gene expression monitoring, mutation detection, metabolic engineering, drug development, tailor made therapeutics, SNP research, GMO detection, and high-throughput screening, among others. They have a profound impact on biological research, industrial production, medicine and pharmacology and will be used as the biosensors of the future [63 – 67]. For the understanding of biological systems with up to 30 000 genes, the measurement of RNA levels for a complete set of transcripts of an organism will be necessary. The use of glass slides as medium enables high spot densities on microarrays (up to 10 000 spots per slide). A DNA chip experiment works by hybridizing all of the gene probes on the chip with the nucleic acid target to be tested. cDNA labeled with fluorescent tags is used, produced by

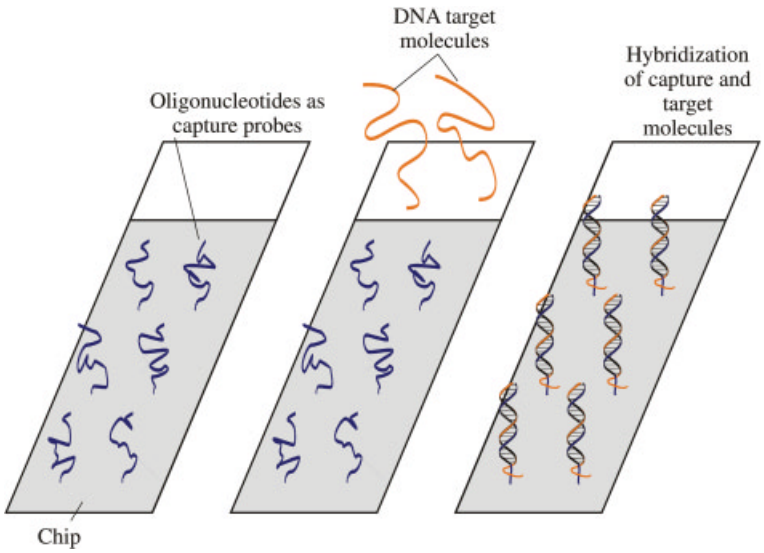

Figure 14. Interaction of labelled target molecules with probe molecules on a glass array

reverse transcription of RNA from the sample to be analyzed. Hybridizing a sample of nucleic acid with many complementary gene probes in parallel on a DNA chip produces a hybridization pattern with corresponding hybridization intensity. DNA chip technology therefore enables large numbers of genes to be screened simultaneously, giving a comprehensive, detailed picture of changes in gene expression, shedding light on complex regulatory interactions.

4.1.2. Fabrication of DNA microarrays

There are three primary technologies used presently in automated microarray fabrication including photolithography, ink-jetting, contact printing, and derivatives thereof. Each of these technologies has specific advantages and disadvantages in microarray manufacturing. The photolithographic approach relies on the in situ synthesis of 25mer oligonucleotides using photomasks. This means that each probe is individually synthesized on the chip surface. Photolithography was developed by Fodor et.al. [68] and commercialized by Affymetrix (Fig. 15). In contrast, the ink-jetting and contact printing methods attach presynthesized DNA probes to the chip surface. While the in situ probe synthesis necessitate expensive and sophisticated equipment the contact and noncontact spotting methods made DNA chips affordable for academic research laboratories. Since 1996, many DNA arrayers have become available and the self spotted glass slide DNA arrays are today the most popular format for gene expression profiling experiments.

Figure 15. Microarray fabrication using an Affymetrix 427 arrayer

4.1.3. Proteome Analysis using Protein Microarrays

The utilization of protein biochips for the Proteome analysis offers a number of alternatives and advantages in metabolic engineering. Protein microarrays make use of technological innovations and enable the proteome analysis in

miniaturized test formats, thereby promising significant advantages such as an improved analytical speed, a better separation efficiency, reduced sample/reagent consumption, contamination reduction and reduced costs. Protein microarrays consist of bound antigens or antibodies as capture molecules for detecting proteins in a complex mixture. They provide a huge potential for monitoring protein expression and protein profiling.

Although the basic construction of such protein microarrays is similar to DNA microarrays, the more delicate nature of protein structures has hindered the development of such devices for the analysis of proteins for a long time. This is due to their more complex coupling chemistry, the instability of the immobilized protein and the far weaker detection signals. In principal, microarray technology enables both, the probing of the genome and the proteome. Today, protein chip technology focuses basically on the understanding of molecular pathways which allows conclusions concerning the molecular diagnostic, the drug discovery, and metabolic engineering applications. These interferences can be done because of the capability of protein microarrays to analyze the protein function of a whole genome level [69 – 73]. The results of such high- throughput screening approaches can change our fundamental understanding of the cellular processes of life on the molecular level. However, gene expression analysis does not sufficient enable a reliable prediction of the function of a protein. Monitoring protein interactions is an extremely complex matter since the proteome is the quantitative representation of the complete protein expression pattern of a cell under accurately defined conditions. The proteome represents the protein equivalent of the genome. In contrast to the genome which is determined by the sequence of its nucleotides and therefore static, the proteome represents an extreme dynamic object which is influenced by many parameters. Not all genes will be switched on at the same time in a cell and the sensitive balance between protein synthesis and protein degradation can vary widely under different metabolic conditions. Consequently, a repeated analysis of the proteome will be successful only under exactly defined conditions, for example, cell culture conditions. In practice, this proves to be very difficult since conditions to which the cell severely reacts are often unknown. However, the sensitive dependence of most different parameters offers the possibility of applying specific small changes of the protein expression pattern in order to create sensitive biosensors.

Ideally, the analysis of the proteome delivers the currently available set of all proteins, if there is a way to maintain the quantitative relations of all proteins during the analysis. Such data cannot be obtained with classical molecular biology since no strict connection between the amount of mRNA and the amount of protein exists. Parameters such as mRNA stability, posttranslational modifications, protein degradation, and others consequently prevent a statement over the currently available amount of protein. However, this information is of utmost importance making a high throughput analysis necessary. One attractive method is the use of protein microarrays. Genomewide screens for protein function are of biological importance for many applications:

- Analyzing protein expression profiles
- Monitoring protein-protein interactions
- Identifying protein posttranslational modifications
- Screening the substrates of protein kinases
- Examining the protein targets of small molecules
- Proteomic analysis as a function of bioprocess cultivation conditions

4.1.4. Metabolome Analysis and Metabolite flux analysis

In order to improve a given strain of microorganisms or even more complicated a eukaryotic cell line that is used in a biotechnological production process, it is not only necessary to analyze gene expression patterns or the proteome. Analysis of metabolic turnover is central for recognizing possible targets for genetic engineering. Metabolic turnover comprises a large number of biochemical reactions. For instance, in the well-examined yeast *Saccharomyces cerevisiae*, about 1500 biochemical reactions have been estimated involving more than 850 metabolites and cofactors [74]. Since in most of these reactions there is more than two educts and more than one product, not to speak of the necessary cofactors, the metabolic turnover represents a complex network of biochemical reactions, also referred to

as the metabolic network. The concept of isolated metabolic pathways such as glycolysis, citric acid cycle, or urea cycle as presented in most biochemical textbooks is therefore somewhat misleading since it indicates that these pathways proceed independently. However, they are all interconnected and represent a huge interdependent network that can activate or slow down individual branches. Control over the activity of individual metabolic branches can be exerted on the levels of transcription, translation, protein-protein interaction, or enzyme activity regulation, which adds another level of complexity. According to Nielsen [77] it is therefore necessary to analyze the metabolic network in three instances (*metabolic pathway analysis*) in order to fully oversee the metabolic activities in a living cells, which is considered as a prerequisite for a rational optimization of the production strain:

- Network structure (pathway topology).
- Quantification of fluxes through the branches of the network.
- Identification of control structures.

Much work has been done in the past in biochemical laboratories around the globe to identify the metabolic network structure at least for the more well-known microorganisms – in particular when they are of industrial importance. Thus, very often the presence of a certain metabolic pathway can be deduced from the available literature. The most powerful approach to identify a certain pathway when no information is available is called metabolic labeling. Here, the organism is fed with ^{13}C-labeled glucose (or any other appropriate fundamental metabolite) and after a given time that is allowed for metabolization the labeling pattern of other intracellular components is identified and reports on the metabolic pathways involved. From the analytical perspective, analysis of the metabolite pattern is most often done by GC-MS (gas chromatography coupled to mass spectrometry) but also by NMR (nuclear magnetic resonance). Later on, biochemical assays addressing the activity of certain key enzymes can be used to confirm the presence of a certain pathway and dig out the cofactor requirements for the individual steps within the pathway. Pathway identification just by enzyme assays is possible, however, very tedious and time consuming.

Flux Analysis. Once the network topology has been identified, it is necessary to learn about the fluxes through each of the network branches. This kind of analysis requires experiments as well as mathematical modeling. The most important concept behind flux analysis is referred to as *metabolite balancing*. According to this, the concentration of each metabolite is stationary inside the cell or in other words: the generation rate of a given metabolite in one metabolic pathway (v_+) equals the rate of its onward reaction through the sum of all downstream metabolic pathways (v_-). Thus, for a given metabolite the corresponding material balance can be described as

$$v_{+1}+v_{+2}+v_{+3}+\ldots+v_{+i}=v_{-1}+v_{-2}+v_{-3}+\ldots v_{-i}$$

Such balances are established for all relevant metabolites, such that a set of algebraic equations results that describe the flux of molecules through the different branches of the metabolic network. When some of the fluxes are experimentally determined, the remaining fluxes can be deduced mathematically from this set of equations. The beauty of the metabolic balancing approach is its simplicity. But reliable predictions can only be obtained when the cofactor (NADH, NADPH, . . .) balances are also known which means that all the pathways that generate or consume any of these cofactors have to be considered in the formalism. Feeding the organisms with ^{13}C-glucose in combination with a molecular analysis of the resulting metabolite labeling downstream of glucose provides, however, an easier solution. From the labeling pattern of all metabolites it is possible to set up equations for the balances of individual carbon atoms. Thus, the number of equations describing the overall system increases the number of constraints for the solution and thus, makes balances of cofactors unimportant.

The final step in analyzing the metabolic network is to identify the regulatory mechanisms that control the fluxes through each of the branches [75, 76]. It is very obvious that detailed knowledge about flux control might immediately identify possible targets of metabolic engineering, for instance, to redirect the metabolic flux away from a pathway that does not lead to the desired product or consumes precursors. In order to understand flux control, it is necessary to study the regulation of the enzymes

that are active around the branching points of the metabolic network [77]. Here the regulatory instances can be found at several hierarchical levels as mentioned earlier in this chapter. The expression level of a given enzyme or any of its regulatory proteins may be controlled on the level of transcription, RNA processing, or translation. DNA and protein chip technology, as described above, provide powerful approaches to unravel whether any control occurs on the level enzyme expression. Finally, allosteric control over enzyme activity introduces another level of metabolic fine-tuning. The kinetic properties of an enzyme like, for instance, its affinity for the substrate molecules, the maximum substrate turnover rate and its sensitivity for product inhibition have to be characterized by biochemical assays. When these information and the current concentrations of the relevant metabolites are available, it should be possible to predict how the fluxes through individual branches of the metabolic network will change when the activity of one enzyme is altered by genetic engineering or when the feeding situation is changed.

However, it should be emphasized at this point, that due to the enormous complexity of the metabolic network and all its regulatory instances, it is very difficult to predict all metabolic consequences induced by just one genetic modification. Thus, it may be necessary to completely analyze the metabolome of the modified strain and – if the results are not satisfactory – go through one or several more rounds of metabolic engineering as sketched in Figure 13. This is particularly relevant when the organisms respond to alteration of their genetic make-up by opening up metabolic side pathways that are silent in the wild-type strain.

4.2. Design and Production of Genetically Optimized Strains for Production – In Vitro Mutagenesis

In order to generate genetically modified organism that might be superior to wild-type strains for production purposes, there are basically three major options:

A) Introduction and expression of a gene that is not originally expressed by the organism or only in small copy numbers
B) Silencing of a gene that codes for an enzyme or regulatory protein which should be switched off or reduced in expression level
C) Modifying an existing gene – and hence the coded protein – that is expressed by the organism of interest, for instance, to alter the substrate specificity of an enzyme or engineer its catalytic performance

In all three cases, the most suitable target for genetic manipulations is cloned DNA that can be handled and tailored in the well-defined environment of a test tube without the complexity of an entire organism around, in particular the cellular membranes that may hinder the accessibility of the various reagents. Cloned DNA is usually handled in form of plasmid DNA. A plasmid (cyto*plasm* + chromat*id* = *plasmid*) is a circular piece of double-stranded DNA that is autonomously replicated in bacteria and may even be exchanged between them. Plasmid DNA occurs also in wild-type strains since they are regular extrachromosomal genes that can be easily shared between certain bacteria. Many molecular tools have been developed that are specifically designed to work with plasmid DNA in vitro. A particular family of plasmids, the so-called *Expression plasmids* or *expression vectors* as the more general term, additionally contain the necessary nucleic sequences to initiate transcription (promotor) and translation of a gene that has been correctly inserted into the plasmid – instead of multiplication only. Expression plasmids for prokaryotic cells differ from those for eukaryotic cells [78]. The three different strategies to alter gene expression in more detail:

A) When a new gene should be introduced into an organism or the expression level of an existing one should be increased, this particular gene has to be cloned in a suitable expression vector controlled by an appropriate promotor that ensures a sufficiently high copy number of mRNAs and hence protein. The production of human insulin in *E. coli* is a famous example for the production of a recombinant eukaryotic protein in a prokaryotic organism.
B) When an existing gene should be silenced or reduced in activity, there are several experimental strategies. Today the most promising technique is RNA interference (RNAi), a process by which double-stranded RNA si-

lences gene expression. By transfecting cells with small interfering RNAs (siRNAs) with 100 % homology to a target mRNA sequence, a specific knock-down of the corresponding protein can be achieved. Bound by an RNA-induced silencing complex (RISC), the siRNAs induce the degradation of their homologous mRNA molecules in the cell, thus the protein expression level decreases. RNAi is a reverse-genetics approach to identify gene function by altering the phenotype of a cell [79 – 82].

C) Until the middle of the last century, traditional biotechnology concentrated on elucidating new metabolic pathways and improving the strains used for the production of antibiotics and other useful natural products, usually by exploiting random mutagenesis and selection. Later in the 1980ies, biologists began to improve strains through metabolic engineering, that is, through the modification of single metabolic pathways by means of *targeted* or *site-directed* genetic changes. The two different approaches – random or site-directed mutagenesis – to improve the genetic make-up of an organism are still applied today and shall be summarized here. However, throughout the last years progress in molecular biology has provided a pool of experimental approaches for targeted genetic changes that are way too many and multifaceted to be reviewed comprehensively. Thus, only a few principles can be discussed here, and are confined to the best-known models of genetic research, the bacteria. The interested reader is referred to other chapters of this series or the literature on molecular biology for more detailed and organism specific information [83, 84] (→ Genetic Engineering).

4.3. Random Mutagenesis, Isolation and Selection of Mutants

Spontaneous Mutants. In bacterial populations, mutational events occur at certain rates without experimental treatment. These spontaneous mutations, which result in mutants, are due to errors that occur during DNA replication. The frequency of spontaneous mutations in a bacterial population varies from gene to gene and species to species and may involve 10^{-2} to 10^{-11} of the total cell population. The spontaneous mutation rate (probability of a mutation per cell and generation) for a gene is about 10^{-5}, and for a pair of nucleotides it is 10^{-8}. The number of reversions (or functional back-mutations) is of the same order of magnitude.

Induction of Mutations. The mutant frequency can be considerably increased by treatment of the cells with chemical, physical, or biologic mutagenic agents. Three classes of mutants can be distinguished with respect to their genetic structure: 1) replacement of one base pair by another, e.g., AT by GC; 2) a base pair can be inserted or eliminated; 3) several base pairs, DNA fragments, or even genes can be lost (deletion), change their position within the chromosome (transposition), or be interrupted by insertion of foreign DNA (insertion). Class-1 mutants, called point mutants, revert readily to their original structure.

Various chemical agents are in use to induce mutations [85, 86]:

1) *Incorporation of Base Analogues.*
Base analogues are antimetabolites. Some are sufficiently similar to the natural purine and pyrimidine bases that they are taken up by the cells and incorporated into DNA during replication. They are able to fulfill their functions, but tend to bind a wrong counterpart during replication, thus introducing a wrong base pair and causing a point mutation.
2) *Chemical Change of Bases.*
Several mutagenic agents effect a chemical change in a base and thus cause an error in replication. For example, treatment of cells with nitrite results in the de-amination of adenine, guanine, and cytosine, and consequently in mispairing. Alkylating agents, such as ethyl or methyl methanesulfonate, ethyleneimine, nitrogen mustard, and *N*-methyl-*N*-nitro-*N*-nitrosoguanidine (MNG), belong to the most effective mutagenic agents. Acridine dyes (proflavin) function by intercalation and result in insertions or deletions of single base pairs.
3) *Irradiation.*
UV irradiation mainly affects pyrimidine bases and results in replication errors.
4) *Transposon Mutagenesis.*
By conjugation, transposon-containing plasmids (Tn elements) can be transferred from

an appropriate donor bacterium to the bacterium to be mutagenized. Transposons are DNA fragments that tend to "jump" to various positions in the DNA sequence of the chromosome or plasmid. Their insertion interrupts genes and results in insertion mutants.

The changes in the DNA described above involve an alteration of its base sequence, called a premutation. Subsequently, the mutant character has to be expressed in the phenotype of the cell. About 10–20 growth cycles may be required for nuclear segregation, dilution of enzymes, starting or stopping the synthesis of enzymes, and other functions.

4.4. Types of Mutants and Selection Principles

A mutation is a rare event, and the number of mutants in a population is small. Many thousands of Petri dishes would have to be inspected and many millions of colonies (clones of single cells) would have to be examined to recognize and isolate a mutant if there were not a selection (enrichment) procedure for increasing the number of the desired mutants within the population. Enrichment is achieved either by killing part of the parent cells or by growth under conditions that confer growth advantage to the mutant cells. There are many means for enriching mutants. In fact, designing strategies to select mutants is one of the most difficult and demanding arts of the microbiologist, and each type of mutant requires a special trick.

The recognition of mutants poses another problem. Mutants that have lost or gained pigment, colony size or characteristics, or that have become resistant to toxic compounds or bacteriophages can be easily detected in a lawn of parent cell colonies. Some mutants show up after addition of indicators. The detection of auxotrophic mutants requires comparison of growth on two different nutrient media. One tries to recognize mutants on an agar plate and to avoid laborious test tube assays of thousands of isolated cell clones. There are, however, many mutant types that require these efforts.

4.4.1. Auxotrophic Mutants

If the mutation has affected the ability to synthesize, for example, the amino acid leucine, the mutant will require leucine in the nutrient medium. A mutant that is auxotrophic for leucine (leu^-) can be recognized by comparison of its growth on two agar plates, one with and the other without leucine. While the phototrophic parent type (leu^+) will grow on both plates, the auxotrophic leu^- mutant will grow only on the supplemented agar.

The enrichment of auxotrophic mutants can be achieved by application of agents such as penicillin that kill only growing cells and leave non-growing cells unimpaired. If the population of parent cells and rare leu^- mutants is grown on a minimal medium (lacking leucine) only the parent cells will grow and after the addition of penicillin will be killed within a few hours. The mutants and some parent cells will survive and the mutant fraction is increased by a factor of 10^4 to 10^6. Another method involves the principle of "lethal synthesis." Growing cells will, for example incorporate radioactive phosphate or antimetabolites and will be killed while the non-growing mutants will survive.

4.4.2. Regulatory Mutants

Some mutants, which in contrast to the parent type form a biosynthetic enzyme constitutively, can be selected directly on the agar plate. For example, if an antimetabolite is added to the agar, the parent cells will incorporate it into their protein and stop growing. Mutants, however, that overproduce the normal metabolite because of overproduction of the biosynthetic enzyme will not incorporate the antimetabolite; they will grow, and sometimes even show their regulatory defect by excreting the metabolite, thus initiating secondary growth of the parent cells in the surroundings of the mutant colonies. Some procedures for the selection and recognition of mutants are listed in Table 2.

4.4.3. Other Selection Methods

Various other methods can be used to separate mutants from their parent cells. Because some

Table 2. Procedures for selecting and recognizing various mutants

Kinds of mutants	Selection or enrichment procedure	Recognition of mutants
Mutants resistant to inhibitors, antibiotics, toxic compounds, or bacteriophages	about 10^8 cells spread on a nutrient agar medium containing the inhibitory or killing agent.	only the desired resistant mutants able to grow.
Auxotrophic mutants that require accessory nutrients (vitamins, amino acids, or other metabolites) for growth	penicillin technique or analogous procedures: cells grown in a medium lacking the accessory nutrients but containing penicillin or another agent that kills only growing cells. Auxotrophic mutants not able to grow in minimal medium survive.	if the cell suspension is spread on a complete medium containing the accessory nutrients and if the colony pattern is then replicated on a minimal medium, colonies that do not grow on the minimal medium are auxotrophs.
Mutants that lack the ability to utilize a special substrate	penicillin technique and/or direct isolation of pinpoint colonies: the cell suspension is spread on a nutrient agar that contains the substrate utilizable by the wild-type cells at normal concentrations (0.5 vol%), and the substrate accessible by the desired mutants at a very low concentration (0.005 vol%). Pinpoint colonies are transferred and screened.	comparison of colony patterns on agar plates containing different substrates; mutants unable to grow on a substrate will need a different nutrient for growth and are therefore recognized. Excretory mutants can be recognized by staining reactions (pH indicators) or by dyes added to the agar or after colony growth.
Temperaturesensitive (conditional lethal) mutants	penicillin technique to kill the wild-type cells growing at their upper temperature limit.	comparison of colony pattern on plates that have been incubated at different growth temperatures. Only those colonies that grow, e.g., at 25 °C but not at 37 °C are isolated.
Mutants derepressed or constitutive for catabolic enzymes	1) continuous culture with the substrate as growth-limiting factor; 2) alternating growth on two different substrates; 3) growth in the presence of an agent suppressing induction (antiinducer).	If the cell suspension is spread on a nutrient agar containing a noninducing substrate and after incubation the colonies are sprayed with a solution containing the constitutively utilizable substrate + indicator + inhibitor of enzyme protein synthesis, only constitutive mutants immediately degrade the substrate and change the indicator.
Mutants derepressed or constitutive for anabolic enzymes	growth in the presence of an antimetabolite that inhibits the growth of the wild-type cells. Among the resistant cells are some mutants that are not subject to end product repression.	only the resistant mutants grow, colonies of a clone forming the enzymes of a biosynthetic pathway constitutively will be surrounded by a halo of satellite colonies due to the excretion of a metabolite.

cell constituents have lower or higher densities than the average cell mass, the cells can be incubated under conditions that promote synthesis only in the parent or the mutant cells, followed by centrifugation in a sucrose gradient. Cells rich in lipids (poly(β-hydroxybutyric acid), triglycerides), glycogen, calcium dipicolinate, magnetosomes, etc. can thus be separated. In other cases tactic responses, such as migration in light–dark fields (phototaxis), in gradients of nutrient (chemotaxis), or oxygen concentrations (aerotaxis), can be used to select mutants that are defective in transport systems or in perception mechanisms. Adsorption on particle surfaces (glycogen or starch granules, lipid droplets, lignocellulose particles) followed by fractional centrifugation can be used to select for mutants that have altered surface structures. The separation efficacy can be increased by the use of column chromatography or by suicide techniques. *Suicide techniques* employ compounds that are converted by the wild types to toxic compounds; this lethal synthesis will finally kill the parent cells. Examples are monofluoroacetate, bromopyruvate, or chlorate, which are converted into fluorocitrate, bromolactate, or chlorine, respectively. The mutants are unable to catalyze such conversions and will therefore survive. With this methods it is possible to select for cells that are deficient in nitrate reductase (chlorate, perchlorate), or for fermentative mutants that have lost the ability to form acids (bromide, bromate). Successful biotechnology in many cases depends on the background information and the ingenuity of the researcher in selecting, recognizing, and manipulating the desired organism.

In most cases, a single random mutagenesis decreases the characteristic of an organism that is to be changed, such as, for instance, the catalytic performance of an enzyme. By applying *high-throughput screening* technologies, it

is, however, possible to screen large numbers of strains that are produced by random mutagenesis and find the valuable ones with beneficial mutations. Thus, by going through several rounds of random mutagenesis on the one hand and screening for the desired phenotype on the other, this process will identify those strains with considerably improved properties for production [87].

4.4.4. Targeted or Site-Directed Mutagenesis

Site-directed mutagenesis allows modifying a DNA sequence and thus the protein that is coded by this sequence in a specific and well-controlled way. Thus, it is possible to exchange one or more amino acid within the primary structure of the protein and thereby alter its functionality, for instance the catalytic performance of an enzyme. Compared to random mutagenesis, this is the more straightforward and direct approach not relying on chances that a genetic change may occur at a site that is useful for strain improvement. However, it requires detailed information on the three-dimensional shape of the protein and its primary structure. Many different approaches for targeted genetic alterations have been described that cannot be fully reviewed here. We will summarize the most widely applied and flexible concept of oligonucleotide-based mutagenesis.

In *oligonucleotide-based mutagenesis*, the gene of interest is cloned in a plasmid. After denaturation of the double-stranded circular DNA, oligonulceotides are added to the system that share sufficient sequence homology with the coding strand of the gene to allow for hybridization. However, at the site where the mutations should be introduced the sequence of the olignucleotide is altered. Since this is only a minor portion of the entire olignucleotide, hybridization still occurs but the base exchange has been introduced. Figure 16 sketches the underlying principle.

Chain extension by DNA polymerase and ligation results in a double-stranded DNA with one mutant strand and one wild-type strand. When these plasmids are then introduced in bacteria by transformation, semiconservative replication produces homoduplexes with both strands being either of the wild-type or the mutant. When colonies are grown, one can now screen for those strains in which the mutant plasmid has prevailed.

Many more experimental options exist to introduce site-specific mutations. Many of these are based on polymerase chain reaction (PCR) and the application of primers that already hold the mutation, or insertion or deletion. The interested reader is referred to Bowen [88].

The aim of modern metabolic engineering is to study the cell as an integrated system of genetic, protein, metabolite, and pathway events that are continually changing. This approach should help to optimize the production of substances or to produce substances with improved properties. Historically, metabolic engineering is the targeted recombination of the DNA for proteins involved in the metabolism or in regulation. Today, the objective of metabolic engineering is to quantify the pathway alterations in response to environmental mediators taking into account the entire metabolism. Knowledge of in vivo flux distributions in cells at different physiological states is of increasing importance for the evaluation of biotechnological processes. Gene expression data provide information on pathways relevant to the metabolic models. Furthermore, the so-called *combinatorial biosynthesis* uses techniques from molecular biology to alter or combine genes from biosynthesis pathways of the secondary metabolism to generate new products with improved pharmacological profil. The importance of studying the entire metabolism rather than one gene or protein at a time has become increasingly relevant with the advent of high-throughput genomic and proteomic technologies, especially microarrays. The DNA microarray technology will gain great importance in the field of future biological research developing into a key technology of the 21st century hence revolutionizing modern biotechnology. DNA chips are used as biosensors in industrial analysis, biomedical diagnosis, and forensic science. Although this technology provides a powerful tool that is widely utilized for gene expression, it is just beginning to find applications outside of genomics. Further development of protein, cell, and tissue chips will make it possible in the future to carry out miniaturized, highly parallel analysis of metabolic pathways. The challenge is to transfer these principles into technically applicable and precise ana-

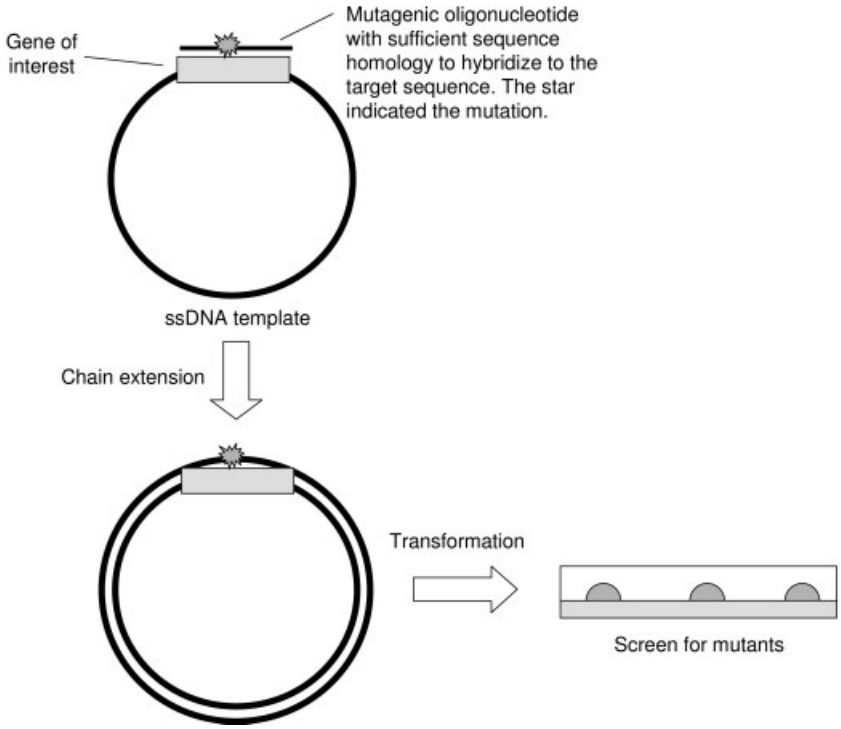

Figure 16. Site-directed mutagenesis of cloned DNA by oligonucleotide mutagenesis

lytical systems that can be used for many applications repeatedly.

5. Cultivation and Bioprocesses

In this chapter, only the basic requirements for bioprocesses are described very shortly. For any further details concerning different types of bioprocesses, bioreactors, kinetics, sterility, and cleaning, the reader is referred to (→ Biochemical Engineering). Very detailed information about the monitoring and control of bioprocesses can be found in this contribution in Chapter 8.

Bioprocesses [89, 90] are used for the transformation of organic matter with the help of biocatalysts, such as living cells, dead cells, or their components, e.g., enzymes [91, 92]. Bioprocesses are utilized for chemical syntheses and for the conversion of waste material to either useful products (recycling) or effluents harmless to the environment (waste treatment). The outstanding feature of bioprocesses is their high synthetic potential to carry out a series of very complicated chemical reactions in a one-step process. Classic biocatalysts are yeasts, fungi, and bacteria; animal and plant cell cultures have become important only in more recent times (see Chapter 9). Process development [93] [94] [95] involves the areas of biology (strain development and bioregulation), reactor design, process control [96], and medium design [97, 98] as well as downstream processing (product recovery, see Chapter 7).

5.1. Isolation of Microorganisms

Microorganisms can be purchased from culture collections. But "microbes are everywhere, the environment selects," and a desired type of microorganism can often be isolated from its most probable natural habitat. For example, methanogenic bacteria are present in the anoxic mud sediments of ponds and lakes, hemoglobin-degrading bacteria in abattoirs, and hydrocarbon-oxidizing bacteria around oil fields or leaky engines of motor cars. From samples of these materials, bacteria can be isolated by enrichment culture. The technique of enrichment culture is simple and straightforward [99, 100].

By establishing defined environmental conditions with respect to the energy, carbon, and nitrogen sources; hydrogen acceptor; gas atmosphere; temperature; pH value; light; etc., and by inoculating with a mixed population that contains the desired metabolic type, the best adapted organism will dominate and overgrow all accompanying organisms.

Liquid enrichment culture is recommended if the fastest growing organism is desired. In this case, the organism is repeatedly transferred from the original liquid culture through several subcultures. If, however, a great number of strains with only slightly differing traits are to be separated, the *direct plating method* is the preferable technique: the inoculum is diluted and spread on a solidified enrichment medium (agar dish). During incubation the cells form colonies and can be isolated separately.

Whereas formerly these kinds of naturally pure or well-balanced mixed cultures were used in industrial practice (e.g., yeast fermentations, vinegar and cheese production) now only pure cultures are used. Pure cultures are obtained by diluting the enrichment culture and spreading the suspended cells on an agar surface or distributing the cells in an agar medium. After incubation, single colonies are isolated and streaked on various media to test for commensalic bacteria. Finally, a clone of the desired organism is obtained as a pure culture.

Preliminary results indicate that previously noncultivable microorganisms could be the key to the biosynthesis of active substances. The genes responsible for the expression of the natural substance biosynthesis could be inserted into microorganisms which can be grown easily, a procedure known as the *metagenome approach*. The metagenome is a set of all genetic material from organisms which cannot be cultured, e.g., from the soil or from communities of organisms. The development of the metagenome techniques becomes possible with the development of high-throughput techniques. The development of high-throughput DNA sequencing began in the early 1990s as simplifications and automation made it possible to decode the sequence of entire genomes. In 1995 the first full genome sequence was published, that of the Gram-negative bacterium *Haemophilus influenzae*. Of particular interest for biotechnology is the possibility of sequencing microbial production strains developed by repeated random mutagenesis and screening. A comparison with precursor strains could identify the mutations responsible for increased production. That way, production strains could be designed with just a minimal number of defined mutations [101].

Designer Bugs. The term used for microorganisms tailored by genetic engineering methods so that they can carry out desired biochemical conversions efficiently is designer bugs. Contrary to the strategy used in the past to obtain microorganisms for the production of a desired product by coincidental mutagenesis and selection, genetic engineering has opened up the possibility of customized construction of production strains by *metabolic engineering*. The genome permits insight into the entire metabolic potential of the organism. This insight is currently limited because the function of many gene products is not yet known. Organisms, such as *E. coli, Bacillus subtilis, Corynebacterium glutamicum, Pichia pastoris*, or *S. cerevisiae*, are preferentially applied in the development of designer bugs. Further important aspects governing the choice of host organism for a designer bug are knowledge of metabolism, regulation mechanisms, growth properties, and hazard potential. Based on methods developed in the last few years for the analysis of the genome, transcriptome, proteome and metabolome, which thus permitted a holistic consideration of the organism ("systems biology"), it will be possible in future to develop designer bugs for biotechnological production of a desired product considerably more quickly and specifically than in the past.

A list with mostly all culture collection worldwide can be found in the internet [102] under the URL http://www.bacterio.cict.fr/collections.html and in ref. [103 – 105].

5.2. Requirements for Growth

The process solution must primarily supply carbon and energy. In order to maintain growth, nutrients, trace elements, and such growth factors as vitamins must be added to the process solution. Appropriate selection of the medium components is essential for proper functioning and activity of the biocatalysts involved in the reaction.

Table 3. Standard nutrient media

a) *Minimal medium for bacteria*	
K_2HPO_4	0.5 g/L
NH_4Cl	1.0 g/L
$MgSO_4 \cdot 7\,H_2O$	0.2 g/L
$CaCl_2 \cdot 2\,H_2O$	0.1 g/L
$FeSO_4 \cdot 7\,H_2O$	0.01 g/L
Glucose	5.0 g/L
Trace element solution	1.0 g/L
b) *Complete medium for bacteria*	
Peptone	10 g/L
Yeast extract	1.0 g/L
NaCl	2.0 g/L
$MgSO_4 \cdot 7\,H_2O$	0.2 g/L
c) *Complete medium for fungi* (pH 6.0)	
Malt extract	10 g/L
Yeast extract	4 g/L
Glucose	2 g/L
KH_2PO_4	0.5 g/L
NH_4Cl	1.0 g/L
d) *Vitamin solution for soil and water bacteria*	
Biotin	0.2 mg
Nicotinic acid	2.0 mg
Thiamin	1.0 mg
Sodium 4-aminobenzoate	1.0 mg
Sodium pantothenate	0.5 mg
Pyridoxamine	5.0 mg
Cyanocobalamine	2.0 mg
Distilled water	100.0 mL
Two to three mL of this solution is added to 1 L of nutrient solution.	
e) *Trace element solution*	
$MnCl_2 \cdot 4\,H_2O$	3 mg/L
$CoCl_2 \cdot 6\,H_2O$	5 mg/L
$CuCl_2 \cdot 2\,H_2O$	1 mg/L
$NiCl_2 \cdot 6\,H_2O$	2 mg/L
$Na_2MoO_4 \cdot 2\,H_2O$	3 mg/L
$ZnSO_4 \cdot 7\,H_2O$	5 mg/L
H_3BO_3	2 mg/L

The growth of microorganisms is dependent on the presence of water or moisture. Nutrients to be used as sources of energy and for synthesis of cell constituents are dissolved in water. The growth requirements for various microorganisms are different, and many recipes for the composition of nutrient media are known. Basically, all chemical elements that constitute the cell substance have to be present in utilizable forms. There are ten *macroelements* that are constituents of all organisms: carbon, oxygen, hydrogen, nitrogen, sulfur, phosphorus, potassium, calcium, magnesium, and iron. In addition, there are *microelements* (trace elements), such as manganese, nickel, cobalt, molybdenum, zinc, copper, vanadium, boron, chlorine, sodium, selenium, silicon, tungsten, and a few others, that are not required by all organisms. The compositions of a few simple synthetic nutrient media are presented in Table 3.

5.2.1. Chemical Composition of Bacterial Cells

Bacterial cells harvested by centrifugation from a culture growing in liquid medium contain about 70 wt% water. The elemental assay of the dry mass of *E. coli* is: approximately 50 % carbon, 20 % oxygen, 14 % nitrogen, 8 % hydrogen, 3 % phosphorus, 1 % sulfur, 2 % potassium, 0.05 % each of calcium, magnesium, and chlorine, 0.2 % iron, and a total of 0.3 % trace elements. The organic analysis of the dry mass of cells is presented in Table 4. It should be noted that the values refer to a specific bacterium grown in a specific environment and harvested in a specific phase of growth. The results may differ if, e.g., the cells encountered limitation of the nitrogen source and accumulated poly(β-hydroxybutyric acid) (up to 90 wt%) or glycogen (about 50 %) within their cells.

Table 4. Overall macromolecular composition of *Escherichia coli* (mass percent, dry matter basis) [106]

Protein	55.0
RNA	20.5
DNA	3.1
Lipid	9.1
Lipopolysaccharide	3.4
Peptidoglycan	2.5
Glycogen	2.5
Total macromolecules	96.1
Soluble pool of building blocks, vitamins	2.9
Inorganic ions	1.0

5.2.2. Carbon and Energy Sources

Only autotrophic organisms are able to synthesize all organic cell constituents from carbon dioxide as the main carbon source. All others derive the cell carbon from organic compounds, which usually serve as both carbon and energy sources; they are partially assimilated and partially oxidized (dissimilated). Glucose, the monomeric constituent of polysaccharides such as cellulose or starch, can be used by the majority of microorganisms. Many other natural

compounds can be utilized and degraded by one or another microorganism.

5.2.3. Accessory Nutrients

In addition to carbon and energy sources, many organisms require accessory nutrients (growth factors), such as vitamins, amino acids, purines, or pyrimidines (see Table 3).

5.2.4. Sulfur and Nitrogen

These elements can be used in their oxidized forms as sulfate and nitrate. Ammonium ion is the most common nitrogen source for microorganisms.

5.2.5. Oxygen

The oxygen atoms of the cellular constituents are mainly derived from water, substrates, or carbon dioxide. Atmospheric oxygen serves mainly as a terminal hydrogen acceptor of aerobic respiration, being reduced to water.

5.2.6. Complex Media

Many microorganisms can grow in a simple nutrient medium such as that listed in Table 3 a. For *Leuconostoc mesenteroides* such a synthetic medium contains 40 components. If the nutrient requirements of an organism are not exactly known, complex nutrient media can be used, such as yeast extract, yeast autolysate, peptone (Table 3), meat extract, wort, carrot juice, coconut milk, or horse manure extract. In other cases complex media are used for economical reasons, for example, whey permeate, corn steep liquor, soybean extract, or molasses.

5.2.7. Solid Media

For solidification of media, agar is added at a concentration of 1.5 to 2.0 wt%. Agar melts at 100 °C and solidifies on cooling to below 45 °C.

5.2.8. Hydrogen Ion Concentration

The majority of microorganisms prefer a pH of about 7.0. However, there are acidophilic and alkaliphilic microorganisms. Yeasts and other fungi prefer pH 5.0. Buffers, in most cases phosphates, are used to maintain the desired pH, although the CO_2/HCO_3^- system or organic substances may do as well. The pH value is very important: the drop from pH 7.0 to pH 6.0 means a tenfold increase in the concentration of H^+ ions, which is a significant change.

5.2.9. Carbon Dioxide

Many microorganisms require a higher concentration of CO_2 than exists in the atmosphere, such as 10 vol%. This requirement applies to those bacteria that in their natural habitats (intestinal tract, tissue, milk, blood, fermenting juices) are exposed to high CO_2 partial pressures.

5.2.10. Aeration

All obligately aerobic microorganisms require oxygen as an electron acceptor for energy generation. For growth in thin layers on liquid or solid media atmospheric oxygen may suffice. In liquid media aerobic bacteria grow only on the surface if the medium is not agitated and aerated. Only dissolved oxygen is utilized; its solubility in water is very low (6.2 mL/L at atmospheric pressure and 20 °C). Therefore oxygen has to be supplied continuously to the cell suspension growing in submerged culture in flasks or fermenters. To meet the increasing oxygen demand in a growing culture, the speed of agitation, the oxygen concentration in the gas phase, and the total gas pressure can be increased. Various kinds of fermenters have been designed to increase the gas transfer from the gaseous to the liquid phase. The final cell density and yield of microorganisms depends on the rate of oxygen transfer. For fermenter design and scale-up, both the respiration rate (oxygen uptake) of the microorganisms and the oxygen transfer rate from gaseous to liquid phase have to be known. Because the formation of desired products is highly dependent on oxygen supply – whether the cells

are able to take up oxygen at their maximum respiration rate or only much less – aeration poses one of the most important problems in fermenter technology.

5.2.11. Anaerobic Techniques

For the growth of strictly anaerobic bacteria the complete exclusion of oxygen is required. Anaerobic techniques involve deaerated nutrient solutions, tightly sealed flasks, incubation in anaerobic jars, the use of chemical oxygen absorbers (pyrogallol, dithionite or the "gas pak," which contains hydrogen- and carbon dioxide-evolving agents and a catalyst to promote the oxyhydrogen reaction), resazurin as a redox indicator, anaerobic hoods, and skilled hands to practice these anaerobic techniques.

5.2.12. Media Preparation

Substrates for industrial applications are usually included in complex media, which must be supplemented with special compounds, such as a nitrogen source, various nutrient salts, or certain trace elements. Organic precursors are also added in some cases for efficient product formation. Many of the feedstocks need appropriate pretreatment in order to provide for good assimilation by the microorganisms. Starch-containing materials are prepared by milling or steam treatment for softening and swelling. For many processes, especially when yeasts are used, starch must first be broken down to sugar by treatment with amylase derived from barley malt or with microbial amylases. Other feedstocks, such as wood, must be pretreated with acid or alkali. Sulfite liquor is stripped from sulfur dioxide by aeration or neutralization. Molasses is purified by acidification and centrifugation or filtration. In some cases heavy metals have to be removed from feedstocks prior to their use. For processes using solid substrates, see [107].

In general, the raw materials are dissolved or suspended in water and the resulting medium is heated, filtered, and sterilized. The complex composition of the media used in industry causes considerable problems. For downstream processing (harvest, concentration, and purification of product) or for analytical assays during the process, additional pretreatment of the raw material is needed to avoid unfavorable side effects.

5.3. Sterilization

The basis of microbiological laboratory methods and of the conservation of food and feed products is the killing of microorganisms. Sterilization means the removal of living microorganisms, and can be achieved by moist heat, dry heat, filtration, irradiation, or chemical means [108].

5.3.1. Moist Heat

Vegetative cells of bacteria and fungi suspended in water are killed at temperatures around 60 to 80 °C within 5 to 10 min; yeast and fungal spores are killed above 80 °C, and endospores of bacteria at 120 °C within 15 min. Because endospores of bacteria are highly tolerant to various environmental conditions, they are ubiquitous and are distributed through the air. Their presence requires that all nutrient media and canned foods be sterilized in an autoclave at about 121 °C for 20 min. If conditions do not allow the germination of spores and the growth of spore-forming bacteria, e.g., in acid fruit juices, jam, or desserts, heating to 80–100 °C for 10 min will suffice. This is called partial sterilization or pasteurization.

5.3.2. Dry Heat

For killing bacterial endospores by dry heat, longer exposure times and higher temperatures are required than with moist heat. Glass and other heat-resistant equipment can be sterilized for 2 h at 160–180 °C in a dry oven. In any case appropriate indicators or soil samples that contain bacterial spores are included in the oven to test whether adequate temperatures have been reached and even the extremely heat-resistant spores have been inactivated.

5.3.3. Filtration

Solutions containing thermolabile compounds can be sterilized by filtration through nitrocellulose membranes, kieselguhr, porcelain, asbestos, and others. Usually filters with a pore size $<0{,}2$ µm are used.

5.3.4. Irradiation

UV irradiation is used to keep rooms partially sterile. Bacteria and their spores are killed rather quickly, but fungal spores are only moderately sensitive to radiation. Ionizing radiation (X-ray, gamma radiation) is used to sterilize food and other compact materials.

5.3.5. Chemical Means

Ethylene oxide is used to sterilize food, plastics, glassware, and other equipment. β-Propionolactone and diethyl carbonate are added to nutrient solutions. The presently used soaps and detergents applied at moderate temperature kill all vegetative cells and even many spores. Cleaning by the methods practiced in the kitchen provides almost sterile glassware.

Remarkably, the killing of microbes is a process that follows first-order reaction kinetics. In a population there are allways some cells or spores that are more resistant than the majority. Accordingly, sterilization procedures are most likely to be effective on relatively clean substrates.

5.4. Types of Bioprocesses

In chemical technology, a choice is to be made between batch processes or continuously operated processes. Many parameters usually favor the second over the first choice. However, in biotechnology, the batch process still predominates. This specific problem is treated in more detail (also in regard to engineering aspects) in → Biochemical Engineering.

5.4.1. Surface Culture

Surface cultures are mostly used when fungal mycelia act as catalysts. Mycelium is grown in shallow pans containing a small volume of the medium. As the mycelium grows, it eventually forms a compact mat. This method was typically used in the production of acids, such as citric acid. In large-scale production, surface culture is being abandoned more and more and replaced by the submerged culture method [109 – 111].

Surface cultivation is usually applied in the lab scale in fungal research. Large Erlenmeyer flasks or specially designed vessels, such as the Fernbach, Roux, or Sakaguchi flasks, containing small volumes of liquid are used. The flasks and liquid are sterilized and inoculated with spore suspensions or small pellets of mycelium. Growth and product formation are dependent upon the depth of the medium layer and the amount of surface provided by the flasks, which in turn affects the necessary oxygen supply.

In large-scale production large pans are used. They are filled with substrate, inoculated, and then kept at controlled temperature and air humidity. Pans are stacked on suitable racks, and connected with a system of overflowing tubes. The liquid medium is poured in at the top. In this way, an entire stack can be loaded at once. If the stack is designed appropriately, it can be sterilized, inoculated, aerated, and maintained under controlled conditions. The same principle can be applied for solid substrates, which are first inoculated and then loaded into the pans. The material is spread at a thickness of several centimeters onto screens or perforated metal sheets.

Columns filled with carrier material onto which the microbes are immobilized are also considered a surface culture as long as the column is not completely flooded. The medium is trickled through the column and air is injected at the bottom. The product is then collected from the carrier, which is made of wood shavings, plastics, or ceramics. This type of surface culture has been used for vinegar production and sewage treatment. However, it was abandoned and replaced with submerged culture. Remarkably, the immobilization methods that were developed for such biocatalysts as living cells, resting cells, enzymes, or organelles are formally similar to these conventional processes.

5.4.2. Submerged Culture

In this process, the microbes are suspended in the liquid medium. The simplest example of a submerged culture consists of a vessel without any agitation device, in which microorganisms settle at the bottom and form a compact layer. This is the classic arrangement for all types of ethanol-producing processes. Some agitation will occur,

however, from the movement of rising carbon dioxide bubbles.

For submerged, *anaerobic* processes, simple agitation devices for gentle mixing will suffice. This configuration allows for the mixing of liquid and gas. Carbon dioxide or other gases, such as nitrogen, can be added during the process.

5.5. Process Layout

5.5.1. Reactors

Appropriate reactors in biotechnology have to meet the requirements for agitation, aeration, corrosion resistance, and aseptic operation. Various designs of stirred vessel are therefore used in the laboratory and in production plants. Container construction is based on principles similar to reactors used in heterogeneous catalysis.

In former times, bioprocesses were carried out under nonsterile conditions and took place in vats of wood. Beer today is brewed in large cubical containers coated with ceramic tiles. For the more conventional surface processes (citric acid, acetic acid), small pans were used.

A basic problem in aerobic processes is the efficient transport of oxygen into the liquid and the microbial cells. Very efficient devices have been developed to provide air for aerobic reactions. Some examples of more modern reactor configurations for agitation and aeration are shown in Figures 20 and 21.

5.5.2. Containments for Anaerobic Processes

Large units for anaerobic processes are made of wood, metal, or coated concrete. They have no covers and consequently cannot be operated aseptically. However, submerged anaerobic processes can be performed under conditions that favor the formation of product and inhibit the accumulation of unwanted organisms by proper selection of reaction parameters, such as pH and temperature. Lactic acid, for example, is produced at 50 °C and low pH, and the development of competing microbes is negligible.

In other cases, closed vessels are used. They are usually made of stainless steel, have large capacities of up to 1000 m^3 or more, and are equipped with agitators and, if necessary, with cooling and heating coils. Gases that form during the process (CO_2, H_2, or evaporated solvents) are collected at the top. These large tanks have openings for loading, inoculation, harvesting, and cleaning. Large units are mostly sterilized with steam.

Extremely large containments with capacities of thousands of cubic meters are used in sewage treatment. They are usually constructed of concrete and allow automatic feed and removal of material. Sludge treatment plants are often equipped with heating elements to allow thermophilic processes for methane production. The gas is collected and stripped of hydrogen sulfide if necessary.

5.5.3. Reactors for Aerobic Processes

Bioreactors used for aerobic processes are basically similar to those used for anaerobic processes. In addition, they contain devices for aeration and agitation. The capacities used in industry may vary considerably, from 10 to 200 m^3. Large tanks of 1000 m^3 and more are also used for aerated, sterile processes (e. g., glutamic acid). For aseptic processes the tank is sterilized with steam before loading; a slight overpressure is often maintained during cultivation. Agitation principles vary considerably and include mechanical as well as pneumatic and hydrodynamic systems.

The most commonly used type of bioreactor is the *stirred tank reactor* (STR) using a flat blade turbine (FBT) for agitation (Fig. 17). This configuration provides excellent mixing and mass transfer in media of low viscosity. Drawbacks are high power demand, poor performance with highly viscous liquids, and rather poor interchange of material between mixing regions around the blades. Air is normally injected by a sparger. Larger units are equipped for pneumatic agitation, in which movement is caused by injected air. For proper hydrodynamic control, various kinds of baffles and draft tubes can be inserted.

5.5.4. Inoculation

After inoculation with the microorganisms, the process should start immediately and the reaction should proceed fast. The amount of active

cell culture added is therefore critical and depends on the size of the batch. Scaling-up from the original starter culture to the inoculation broth is done in several steps in large-scale industrial processes (Fig. 19). The starter culture is kept deep frozen (−80 to −90 °C) in sealed vials or in vials that were previously subjected to lyophilization (a). They can be stored in this way for several months or even years without any damage. For short-term storage at 4 °C, agar slants are prepared (g) and purity is tested on Petri dishes (h) before inocula are scaled up for production.

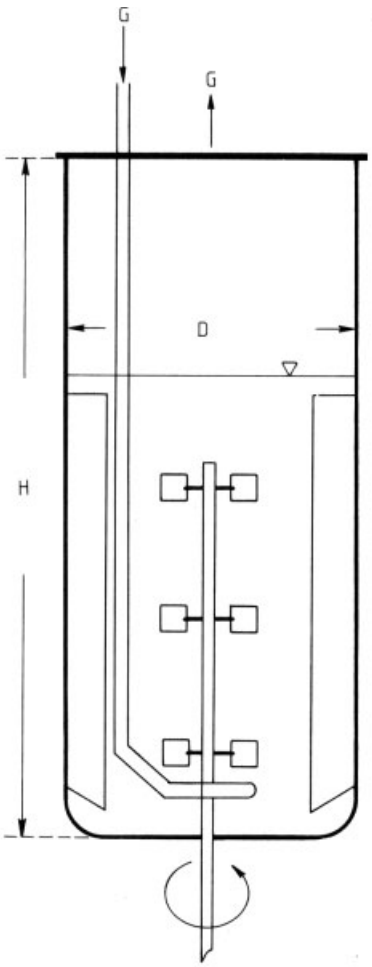

Figure 17. Stirred tank reactor (STR) equipped with flat blade turbine (FBT) and baffles for agitation and a sparger for gas distribution
The ratio of H: D is usually 3 (max. 5).

For scaling-up, the liquid cell culture is placed in liquid broth, first in test tubes (b), then in small (c), and subsequently in large Erlenmeyer flasks (d) containing approximately 200 mL of medium. Proliferation of the culture takes place by shaking the flasks overnight at constant temperature. Larger amounts of inocula are then prepared in agitated and aerated fermenters of various sizes depending on the volume of the production plant (e, f). In the early stages of this procedure (steps a through d) the inoculum is transferred under sterile conditions in glove boxes.

In *fungal inocula* proper wetting of the spores is achieved by adding small amounts of surfactants to the broth. If inoculation by spare suspensions is not optimal, mycelial pellets can be used for start-up. *Bacterial spores* must be activated by thermal treatment before they can be used for inoculation. During the exponential growth phase of the bioprocess cells can be harvested for following inoculations.

5.5.5. Operation Modes

Optimum production depends strongly on process layout. The standard process is still the *batch operation*, which consists of loading, inoculation, processing, harvest, and cleanup of the vessel. The advantages of this procedure are its adaptability to weekly work hours and small losses in case of contamination or decreases in productivity.

Greater specific production can be attained by using *semicontinuous* or *sequential-batch* operations. In this method only a partial stream of the broth is withdrawn after completion of the reaction (e.g., 90 %) and new medium is added directly into the remaining volume. This type of operation is also called cyclic continuous production. Modern vinegar production is operated in the cyclic manner and maintained by fully automated process control.

Fully *continuous operation*, that is, the constant flow of medium into a reaction vessel and the simultaneous removal of an identical amount of the process solution, has shown the highest production rate. However, practical and biological constraints, especially the high contamination risk, have so far prevented the general introduction of continuous processing into biotech-

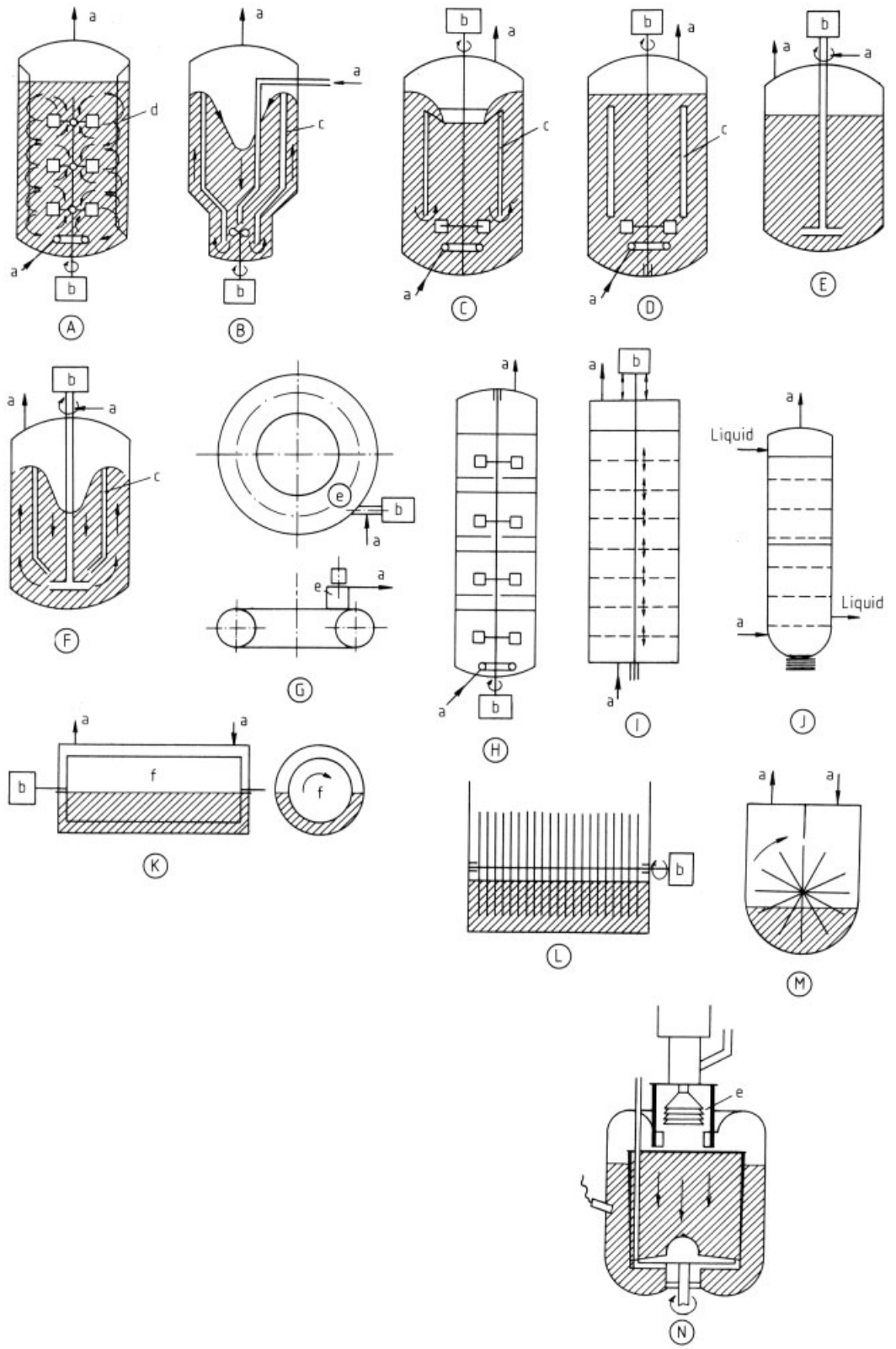
A
B
C
D
E
F
G
H
I
J
K
L
M
N
Liquid
Liquid

Figure 18. Various bioreactor configurations [112]
a) Gas; b) Motor; c) Draft tube; d) Baffles; e) Mechanical foam breaker; f) Cylinder
A) Multistage flat blade turbine reactor (FBT) and baffles (d) making up the classical stirred tank reactor (STR)
B) Forced agitation with draft tube (c) and ship propeller
C) Combination of flat blade turbine and draft tube, operated with an overflow pattern
D) Combination of flat blade turbine and draft tube, flodded
E) Simple vessel with hollow stirrer element for self-aeration and distribution
F) Same as E, but with draft tube resulting in vortex flow patterns
G) Torus reactor; annular form with ship propeller
H) Tower with multistage flat blade turbines
I) Vibrating elements in tower
J) Pulsated tower
K) Horizontal rotating cylinder
L) Rotating disks on horizontal axis
M) Rotating elements in a bath
N) Compact loop reactor with gas – liquid separating element. Operated as completely filled vessel. Forced agitation with ship propeller.

nology. Therefore, only processes that need not be carried out under aseptic conditions are operated continuously such as sewage treatment, methane formation, and feed yeast or ethanol production from sulfite liquor.

The potential advantages of continuous cultivation are substantial because of the efficiency of the process and the uniformity of the final product. As a result, much effort is being made to develop a sound basis of knowledge necessary for using this process strategy.

In contrast to large-scale industrial processes continuous cultivation methods are widely used in the laboratory [113 – 115]. The growth-controlling factor is either the concentration of nutrients in the medium (chemostat system) or the concentration of insoluble biomass sensed by optical measurement (turbidostat methods). The purpose of these control systems is to maintain a constant concentration of nutrients, cells, and metabolic products in the vessel (steady state). The kinetics of growth or product formation can

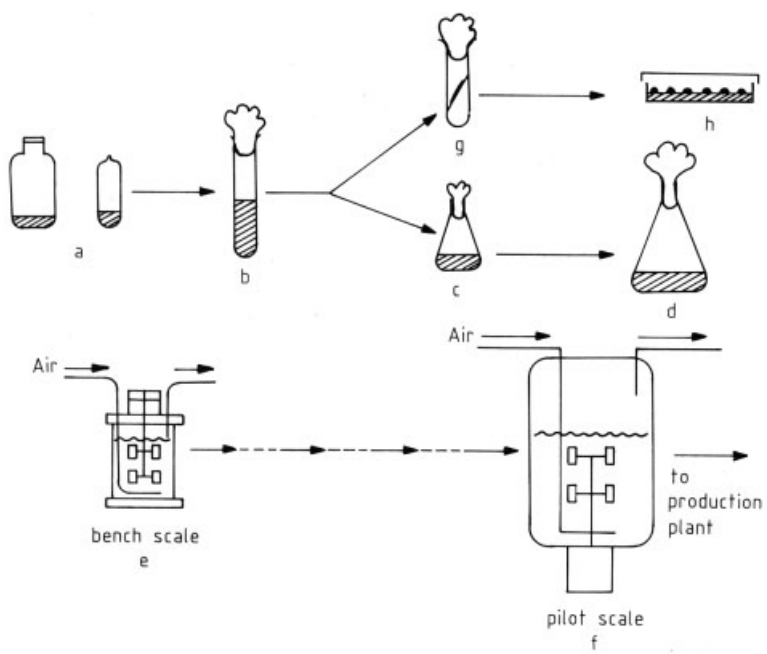

Figure 19. Preparation of inocula.
The individual steps a through h are explained in the text

be kept under precise control by selecting the appropriate dilution rate. All types of variations of the single-stage continuous cultivation are possible, such as multistage arrangements, with or without recycling.

5.6. Process and Product Overview

The potential of biotechnology consists in its ability to replace classical chemical production processes and to facilitate the production of new products. It is undisputed that, particularly in the area of basic and fine chemicals production, the use of biotechnological processes is meaningful since

- biotechnological processes are usually distinguished by their high specificity (relating to the conversion of substrates) and selectivity (relating to the product spectrum)
- biotechnological processes often use renewable resources as raw materials, thus contributing to the much discussed sustainability of products and processes
- biotechnological processes can be carried out under mild reaction conditions in terms of pressure, temperature, and pH.

In view of these facts, in the past hundred years a multitude of industrial biotechnological processes have been developed [116 – 120] whose efficiency exceeds that of chemical processes, and this has helped to establish them in the long run (Table 5 and 6). One interesting area is the production of fine chemicals. The term "fine chemicals products" refers to substances that are highly functional and for which world demand is typically quantities of less than 10 000 t/a. From a chemical point of view these products are usually distinguished by having several reaction centers and frequently by chirality. Classical syntheses of these substances include several reaction steps using stoichiometric quantities of reagents and often deploy extravagant protective group chemistry, expensive noble metal/heavy metal catalysts, and drastic reaction conditions, e.g., aggressive solvents. Here biocatalysis allows synthesis under considerably milder reaction conditions in terms of pressure, temperature, and acidity. At present, due to the excellent enantioselectivity of enzymes, white biotechnology methods are primarily used for the production of chiral compounds. Straathof et al. [121] list 134 industrial biotransformations. Just short of 90 % of the products described are chiral fine chemicals. In future, this field will undoubtedly acquire greater importance and new reaction sequences for the production of compounds with several stereo centres will be established.

Also bulk chemicals are produced by fermentation or biotransformation processes. Bulk products mean products exceeding 10 000 t annually. It is expected that, by the year 2010, 6–12 % of bulk products and polymers produced by chemical means will already be produced by biotechnological processes. High-volume, biotechnologically produced goods are to be found in the food, livestock feed, and the drinks and tobacco industries. A prime example of a bulk product derived by an enzymatic process is acrylamide. Global biotechnological production capacities have continually expanded in the last few years and are today probably in the region of 100 000 t/a. Monomers and polymers produced by biotechnological processes for the plastics and polymer industry are becoming increasingly interesting. Biotechnologically produced polymers, such as polylactide (PLA), 1,3-propandiol (PDO) and poly-3-hydroxybutyrate-co-3-hydroxyhexanoate (PHBH) could provide the basis for innovations.

As for the energy industry, besides bioethanol two further products should be mentioned: biogas and hydrogen. Whereas the technology for biogas recovery is state-of-the-art, hydrogen production is far more complex and requires long-term research efforts. Biogas recovery is a recycling method for residues. The product can be separated comparatively easily, which is crucial to the cost-effectiveness of the process.

6. Biocatalysis and Biotransformation

6.1. Introduction

Biocatalysis (also called biotransformation) is the conversion of one substance (substrate) to another (product) by a microorganism or an isolated enzyme. It is a chemical reaction catalyzed by a particular cellular enzyme or by an enzyme originally produced within the cell. Most

Table 5. Products obtained by fermentative processes of biotechnology [122 – 124]

Product/process	Annual production t/a	Price EUR/kg	Market value 10^6 EUR	Main application
Amino acids				
L-Glutamate	1500 000	1.20	1800	flavour enhancer
L-Lysine	700 000	2	1400	animal feed additive
L-Threonine	30 000	6	180	animal feed additive
L-Phenylalanine	10 000	10	100	aspartame, medicine
L-Tryptophan	1200	20	24	animal feed, nutrition
L-Arginine	1000	20	20	medicine, cosmetics
L-Cysteine	500	20	30 (incl. extraction)	food, pharma
Other amino acids and derivates	3000			pharmaceuticals, cosmetics, nutrition
Acids				
Lactic acid	150 000	1.80	270	food, leather, textiles
Gluconic acid	100 000	1.50	150	food, textiles, metal, construction
Citric acid	1000 000	0.80	800	medicine, food, metal, detergents
Acetic acid	190 000	0.50	95	food
Itaconic acid	4000			co-monomer
Solvents				
Acetone*	3000 000	–		solvent
1-Butanol*	1200 000	–		solvent
Bioethanol	>18 500 000	0.4		solvent, basic chemical, energy source
Biomass				
Starter cultures			100	food
Baker's yeast	1800 000		2300	
Mineral yeast				
Enzymes				
Enzymes (total)			1830	
Detergents			580	detergents
Food industry			500	starch degradation, proteases
Textile, leather industry			250	tanning
Animal feed			170	phytases, proteases
Food				
Cheese	15 000 000		150 000	
Erythritol	30 000	2.25	67	sugar substitute
Drinks and tobacco				
Beer	138 000 000	2.5		drinks, tobacco
Wine	27 766 000			drinks, tobacco
Antibiotics				
Bacitracin A	4	3000	12	healing of wounds
Bacitracin A	>200	<120	24	animal feed additive
Virginiamycin	70	250	17.5	hog feeding
Cyclosporin	3	5200	15.6	organ transplantation
Monensin	>3000	8	24	animal feed additive
Penicillins	45 000	300*	13 500	medicine, animal feed additive
Cephalosporins	30 000			medicine, animal feed additive
Tetracyclins	5000	50	250	
Antibiotics (ca. 160 on market)			19 000	
Other active substances				
Acarbose			300	active substance
Biopolymers				
Polyhydroxyalkanoate				packaging
Polylactide	140 000	2.25	315	packaging
Xanthan	40 000	8.4	336	food, oil production
Scleroglucan				oil production
Pullulan				film former in foodstuff applications
Dextran (derivatives)	2600	200**	520	blood substitute
Hyaluronic acid	500			
Polyglutamic acid				

Table 5. (continued)

Product/process	Annual production t/a	Price EUR/kg	Market value 10^6 EUR	Main application
Vitamins				
Riboflavin (b_2)	30 000			active substance, animal feed additive
Cyanocobalamin (b_{12})	20	25 000	500	active substance, animal feed additive
Vitamin C	80 000	8	640	food, animal feed
L-Sorbose (vitamin C precursor)	50 000			
Amino sorbite				
Lipids				
Phytosphingosine				cosmetics
Animal feed				
Galactooligosaccharides	2500	3.50	9	prebiotics
Cosmetics				
Dihydroxyacetone				suntan preparations

of these enzymes are necessary for the normal metabolism and reproduction of the cell. In biocatalysis, however, these enzymes are simply used as catalysts for chemical reactions. In addition to their natural substrates, many enzymes can utilize other structurally related compounds as substrates and therefore may catalyze nonnatural reactions upon addition of foreign substrates to the reaction medium. Thus, biocatalysis or biotransformation is a specific category of chemical synthesis.

6.2. Classification of Biocatalysts

Enzymes and whole cells can be applied in a biotransformation in many different forms. Usually a whole cell biotransformation is a growth decoupled process, so resting cells (harvested and washed after fermentation) are employed. The cells may be intact or permeabilized (e.g., for better substrate uptake). The different types of biocatalysts are:

- Whole or treated cells (e.g., lyophilized, permeabilized)
- Organelles
- Cell-free multienzyme systems
- Combination of individual cell-free enzymes
- Single cell-free enzyme

Isolated Enzymes may also be applied in different forms. In most cases, it is not necessary to purify an enzyme to homogeneity, a crude cell extract or partially purified solution may also be employed as biocatalyst. In some well-known technical enzyme preparations, several isoenzymes are present (e.g., pig liver esterase). In both whole cells and enzymatic biotransformations the catalytically active species is the enzyme. Enzymes can be subdivided into six classes (classified by the Committee on Enzyme Nomenclature of the International Union of Biochemistry) depending on the type of reaction they catalyze (Table 7) and are listed numerically in a catalogue published by the committee [125] (→ Enzymes).

6.3. History

Biocatalysis includes some of the oldest human known chemical transformations such as the use of yeast for baking or brewing or the use of chymosin from the stomach of young cattle for cheese production. Records proving these ancient technologies date back to about 7000 to 6000 BC.

Another old and well-known example is vinegar production, which has been known since the dawn of recorded history. It is the oldest example of microbial oxidation (Fig. 20).

The process has evolved along traditional lines of fermentation without knowledge of the underlying biochemistry. In 1862, PASTEUR described the biochemical nature of vinegar production [126]. Subsequently BROWN confirmed in 1886 that the oxidation of alcohol to acetic acid proceeded through acetaldehyde [127]. *Bacterium xylinum*, which causes the reaction to proceed, was first described by BROWN and is the first microbial catalyst.

Today, the oxidation of ethanol to acetic acid by acetic acid bacteria is well understood; it proceeds via two successive dehydrogenations,

Table 6. Biotechnology products obtained by biocatalysis and biotransformation [122 – 124]

Product/process	Annual production t/a	Price EUR/kg	Market value 10^6 EUR	Main application
Basic chemicals				
Acrylamide	100 000	1.40	28	–
Amino acids				
L-Aspartic acid	13 000	–		aspartame production
L-Methionine	400	20		infusion solutions
L-Dopa	300	–		active substance
L-Alanine	500	–		infusion solutions
D- and L-Valine	50	–		–
L-*tert*-Leucine	10	500		–
L-Carnitine	200	–		–
β-Phenylalanine	>1			
Food				
Glucose	20 000 000	0.30	6000	liquid sugar,fermentation medium
Fructose				sugar from HFCS
Isoglucose, HFCS	8000 000	0.80	6400	liquid sugar
L-Hydroxybutanedioic acid	100	20		acidifying agent
Palatinite				sugar substitute
Isomalt	70 000			sugar substitute
Aspartame	10 000	–		sweetening agent
Fructooligosaccharide (inulin)	10 500	2–3		prebiotics
Antibiotics (derivatives)				
6-APA	10 000	–		–
7-ACA	4000	–		–
D-4-Hydroxyphenylglycine	7000	–		–
Intermediate products				
(*S*)-2-Chloropropionic acid	2000	–		herbicide synthesis
D-Pantolactone	2000	–		–
(*S*)-Methoxyisopropylamine	several thousands	–		herbicide synthesis (Outlook after chiral switch by Frontier)
(*R*)-2-(4′-Hydroxyphenoxy)propionic acid (R-HPOPS)	1000	–		–
(*S*)-Phenethylamine etc., optical active amines	500	–		–
D-Mandelic acid	>200	ca. 20		chiral auxiliary
Ethyl (*S*)-4-chloro-3-hydroxybutyrate	>150	–		cholesterol-reducing drugs, e.g., Atorvastatin (Lipitor)
m-Phenoxybenzaldehyde cyanohydrin	100	–		–
(*S*)-3-Acetyl thioisobutyrate	100	–		–
(-)-RAN	50	–		–
(*R*)-Glycidyl butyrate	50			–
Active substance (precursors)				
Dilthiazem precursor	50	–		–
Nicotinamide	3000	–		–
Progesterone	200	–		active substance
Ephedrine	1500	60–90		pseudoephedrine, chiral auxiliary
N-Acetylneuraminic acid				sialidase inhibitor (e.g., Relenza)
Special products				
Cyclodextrins	5000	10	50	domestic, food, stabilizers
Pantothenic acid	11 600	ca. 6	ca. 70	

which depend on the cytochrome system for electron transfer to atmospheric oxygen, the ultimate hydrogen acceptor:

Table 7. Classification of enzymes and their reactions

Class	Reaction
Oxidoreductases:	Enzymes catalyzing oxidation–reduction reactions involving oxygenation, such as C–H → C–OH, or overall removal or addition of hydrogen atom equivalents, for example CH(OH) → C=O and CH–CH→C=C. Dehydrogenases, oxidases, peroxidases, hydroxylases, and oxygenases are involved in this class.
Transferases:	Enzymes catalyzing the transfer of various groups (e.g., aldehyde, ketone, acyl, sugar, phosphoryl) from one molecule to another.
Hydrolases:	Enzymes catalyzing hydrolytic cleavage of C–O, C–N, C–C, etc. Esterases, amidases, peptidases, phosphatases, glycosidases, etc. are involved.
Lyases:	Enzymes catalyzing cleavage of C–C, C–O, C–N and other bonds by elimination, leaving double bonds, or by adding groups to double bonds.
Isomerases:	Enzymes catalyzing a change in the configuration of a molecule. Racemases and isomerases are involved.
Ligases:	Enzymes catalyzing the joining of two molecules with the accompanying hydrolysis of a high-energy bond. Such enzymes are often termed synthetases.

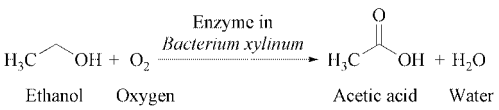

Figure 20. Biotransformation of ethanol to acetic acid

$$C_2H_5OH + \frac{1}{2}O_2 \rightarrow CH_3CHO + H_2O$$

$$CH_3CHO + \frac{1}{2}O_2 \rightarrow CH_3COOH$$

Other historical examples of biocatalytic oxidation are the oxidation of glucose to gluconic acid by *Acetobacter aceti* [128] and the oxidation of sorbitol to sorbose by *Acetobacter* species [129]. A variety of such early reports on biocatalytic conversions was collected by PLIMMER [130]. Since then many other books have been published on this topic.

In 1921, a microbial reaction for the stereospecific production of D-(−)ephedrine was described [131]. A yeast strain, cultivated to produce acetaldehyde from glucose, was found to condense benzaldehyde with acetaldehyde to form optically active L-phenyl-1-hydroxy-2-propanone. This product was in turn converted chemically into D-(−)ephedrine (Fig. 21). This is an example of a successful combination of chemical and biological reactions.

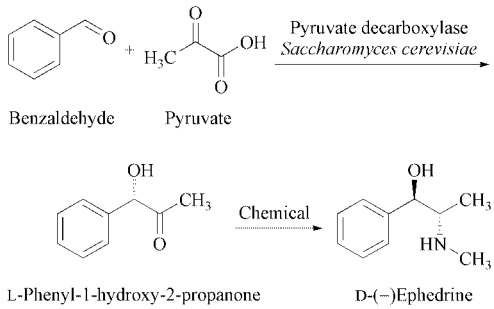

Figure 21. Chemoenzymatic synthesis of optically active l-phenyl-1-hydroxy-2-propanone and chemical conversion into D-(-)ephedrine

Biocatalyis reached its present significance much later, when specific modifications of steroids (or sterols) by microorganisms were discovered. Various steroid reactions, such as hydroxylation, epoxidation, dehydrogenation, isomerization, and hydrolysis, are performed by a wide variety of microbial enzymes. These include the following important reactions: reduction of androstenedione to testosterone by yeasts [132], hydroxylation of progesterone to 11-α-hydroxyprogesterone by *Rhizopus arrhizus* [133], 11-β-hydroxylation of C_{21}-steroids by *Curvelaria lunata* [134], introduction of a Δ^1-double bond into hydrocortisone by *Corynebacterium simplex* [135], 16-α-hydroxylation of 9α-fluorohydrocortisone by *Streptomyces reseochromogenes* [136], and elimination of the side chain of cholesterol by *Arthrobacter simplex* [137]. Many novel intermediates for the chemical synthesis of new steroids became available in this way.

In addition to the considerable practical value, the amazing versatility of microorganisms in the transformation of steroids provided an important theoretical background for the application of microbial enzymes to chemical synthesis.

Many of these reactions cannot be achieved by conventional chemical synthesis.

Table 8. Advances in biocatalysis

Biocatalytic synthesis	Example reactions
Transformation of steroids and sterols	hydroxylation (oxygenase), dehydrogenation (dehydrogenase), side-chain degradation (oxygenase), etc.
Transformation of terpenoids	hydroxylation of beta-ionone with monooxygenase from *Bacillus negaterium* [138].
Transformation of alkaloids	oxidation of morphine to 2,2′-bimorphin (pseudomorphine) by *Cylindrocarpon didymium* [139]; *N*-demethylation of codein to norcodein by *Mucus piriformis* [140].
Synthesis of semisynthetic antibiotics	synthesis of semisynthetic penicillins and cephalosporins (amidase).
Synthesis of organic acids	hydration of fumaric acid (fumarase), diterminal oxidation of alkanes (oxygenase), asymmetric hydrolysis of epoxysuccinic acid (hydrolase), etc.
Transformation of sugars	isomerization of glucose (glucose isomerase).
Protein synthesis	synthesis of plastein (protease), semisynthesis of human insulin (protease), etc.
Synthesis of nucleic acid related compounds	trans-*N*-ribosylation and trans-*N*-arabinosylation (phosphorylase), phosphorylation of nucleosides (phosphotransferase), pyrophosphorylation of nucleotides (pyrophosphotransferase), synthesis of sugar nucleotides (pyrophosphorylase), synthesis of coenzymes, etc.
Synthesis of amino acids	asymmetric hydrolysis of α-amino-ε-caprolactam, 2-amino-Δ^2-thiazoline-4-carboxylic acid, and 5-substituted hydantoins (hydrolase), amination of fumaric acid (aspartase), synthesis of L-tyrosine-related amino acids (β-tyrosinase), L-tryptophan-related amino acids (tryptophanase), etc.
Synthesis of amines	decarboxylation of amino acids (decarboxylase).
Synthesis of industrial chemicals	synthesis of alkene oxides (oxygenase), chiral alcohols and ketones (dehydrogenase), pyrogallol (decarboxylase), amides (nitrile hydratase), etc.

Innumerable biocatalytic conversions involving different types of reactions with organic compounds and natural products were found during the following years; some of them are very useful for chemical synthesis. Advances in biocatalysis are summarized in Table 8. Although biocatalysis has become an important technology for synthetic chemistry, it also has a high potential for many fields of biotechnology.

6.4. Characteristics of Enzyme Reactions Used in Biotransformations

Chemical reactions performed by biocatalysts (biotransformations) are essentially the same as those carried out in inorganic or organic chemistry.

Enzymes (or biocatalysts) used for microbial conversions increase the reaction rate by lowering the activation energy as normal catalysts do. The most striking difference between enzymes and chemical catalysts, however, lies in their substrate specificity. They catalyze specific reactions involving one or only a few structurally related compounds and they distinguish almost absolutely between stereoisomers or regioisomers. Therefore, only a very specific change in a functional group or bond of a compound is accelerated by the enzyme. As a result, one single product is expected as long as only one enzyme is involved in a conversion.

In contrast to chemical reactions that often require high activation energies (high temperature, pressure, etc.) biotransformations can be performed under considerable mild reaction conditions (Table 9). Due to the low energy input needed for biocatalytic processes and the high *atom utilization* (percentage of M_r of desired product to sum of M_r of all substrates), biotransformations are considered as an environmentally friendly alternative to classical organic chemistry processes. To compare biocatalytic and chemical processes on the basis of the amount of waste, Roger A. Sheldon [141] devised the "environmental quotient" (EQ), which is the quotient of the amount of waste product per kilogram of product (E) and an "unfriendliness quotient" (Q): $EQ = E \times Q$.

High catalytic efficacy is another characteristic property of enzymes. They increase the turnover rate without large energy requirements and only a small amount of enzyme is needed to catalyze the conversion of a large amount of substrate.

Table 9. General characteristics of enzymatic and chemical reactions

Parameter	Enzymatic reaction	Chemical reaction
Reaction conditions:		
Temperature	physiological	high
Pressure	atmospheric	high
pH	neutral	different
Reaction energy	enzyme conformation (van der Waals, hydrogen bonds, etc.)	thermal
Solvent	water, (organic solvents and ionic liquids under special conditions possible, e.g., low wateractivity)	water, organic solvents
Specificities (substrate, stereo-, regiospecifity)	high	low
Concentration of substrate and/or product	low	high

Table 10. Characteristics of biotransformation microbial conversion and fermentative production

Parameter	Biotransformation	Fermentative production
Microorganism	resting or treated cells	growing cells
Reaction	simple catalytic reaction (one or few steps)	life process (multistep reaction)
Reaction time	short	long
Starting materials	expensive substrates	cheap carbon and nitrogen sources
Product	natural and/or unnatural	natural
Product concentration	high	low
Product isolation	easy	tedious

As a result, enzymes exhibit their activity under mild reaction conditions, such as atmospheric pressure, temperatures around 20–40°C, and pH values near neutrality, where normal chemical catalysts are inactive. On the other hand, enzymes cannot function under drastic reaction conditions because of their delicate protein structures. The unique properties of enzymes are extremely useful when unstable molecules are to be converted without undesired side reactions.

There are two ways for using biocatalysts in the production of useful compounds: biotransformation and fermentation. *Fermentation* has been used since ancient times (e.g., alcoholic fermentation) and is still of industrial importance in the production of such compounds as organic acids, solvents, or antibiotics. In contrast to biotransformation, fermentation is considered to be a biologic method, because it results from the actual life process of the microorganisms. In other words, the reaction product results from the complex metabolism of the microorganism from cheap carbon and nitrogen sources. Therefore, living cells are required for fermentation, whereas this is not so important in biotransformation. Fermentation products are always natural products. Advantages and disadvantages of these two microbial processes are briefly summarized in Table 10. However, it is not always possible to draw a clear line between the two categories.

6.5. Types of Biocatalysts and Reaction Systems

Because biotransformations are essentially catalytic chemical reactions, a suitable catalyst for the desired conversion must be prepared carefully. Although conventional vegetative cell cultures are most commonly used, there are several alternative methods; some of these have already been operated commercially. These alternatives have been evaluated extensively with regard to, for example, increased yield, control of side reactions, simplified processes, and improved economy. They have great future potential.

Various types of biocatalysts are useful for microbial conversions (see Section 6.2. Any combination of these systems may also be possible. Several successful conversions are known in which multistep reactions are catalyzed by different catalysts (see Section 6.5.6).

6.5.1. Biotransformation with Growing Cultures

The substrate is added to the growth medium during inoculation or an appropriate phase of growth. The preparation and inoculation of the medium, addition of substrate, and incubation are successively carried out in one flask until the reaction is completed. Because of its extreme simplicity, this procedure is frequently used not only for screening (see Section 6.6), but also in large-scale production. Usually, the substrate is added directly to the medium without sterilization if the fully grown culture is the catalyst. Continuous chemostat methods may be used to maintain a steady state of growth, enzyme levels,

and substrate concentration. If high cell densities are required by the conversion or if the nature of the growth medium causes problems for the isolation of the product, the reaction should be performed after separating the cells from the medium, see next section.

6.5.2. Biotransformation Conversion with Previously Grown Cells

The microorganism is grown under suitable conditions and the cells are then collected by centrifugation or filtration. In this way, the cells retain most of their enzyme activity. For the conversion they are resuspended either in an incomplete medium (e.g., one without a nitrogen source; this prolongs cell viability and activity as well as possible cofactor regeneration) or in a simple reaction medium (e.g., a buffer solution or water containing the substrate). The reaction is then carried out directly or after suitable treatment of the cells. It is theoretically not important whether the cells are dead or alive. Vegetative, washed, and dried cells fall into this category. The advantages of this method are listed below:

1) Growth and conversion steps are controlled independently for optimization.
2) The grown cells can be stored for some period of time until initiation of the reaction.
3) The cell density for the conversion is easily controlled.
4) The reaction can be carried out under nonsterile conditions.
5) Product isolation is usually easier because of the simple composition of the reaction medium.

In some cases, however, the strict separation of microbial growth and conversion is considered to be unfavorable for economic reasons.

Depending on the ability of substrate permeability through the cell membrane these biocatalysts may have to be treated differently prior to their use.

6.5.2.1. Vegetative or Washed Cells

In cases where the cell membrane is permeable to substrate and product, no special technique is needed for the preparation of the cells and they can be used as vegetative or washed cells.

6.5.2.2. Permeabilized Cells

In cases where substrate or product molecules do not readily permeate the cell membrane, a modification of the cell becomes necessary. Commonly used techniques to change the permeability of the cell membrane to control the diffusion of substrate into the cell and of product out of the cell are: using surfactants, organic solvents, antibiotics or enzymes. Besides these chemical methods physical procedures, such as osmotic shock or freezing and thawing may also be useful [142]. The production of high-energy phosphate esters, such as adenosine 5-triphosphate, nicotinamide adenine dinucleotide, flavin adenine dinucleotide, or coenzyme A, is a typical example [143].

6.5.2.3. Dried Cells

Dried cells are an alternative to permeabilized cells. Drying causes structural changes in the cell wall or membrane. Contact between the cell enzyme and the substrate is more readily achieved. In some cases, drying also eliminates undesired side reactions, because some enzymes are damaged through drying. Dried cells are usually prepared by air-drying, lyophilization and acetone treatment. The powdered dried cells can retain their enzyme activities in the frozen state for years. They can be used as "instant catalysts" in the same way as normal chemical catalysts.

6.5.3. Biotransformation with Spores

Many fungal spores have as many useful enzymes as vegetative or grown cells. The fungi are cultivated under conditions that induce high spore yields. The spores are then separated from the mycelium and used as catalysts in a similar way as mentioned in the previous section. Processes using spores have the same advantages as those using grown cells. Spores are usually more stable than cells. Some types of spores can be stored for several years without loss of conversion activity. They usually do not germinate during the reaction, but in some cases germination is required for maximum conversion activity. The 11-α-hydroxylation of progesterone by

Aspergillus ochraceus spores represents a successful spore process [144]. Many other applications have been reported, e.g., fatty acids conversions, triglycerides, carbohydrates, or penicillin V [145].

6.5.4. Biotransformation with Immobilized Cells

Immobilized cells are cells that have either been fixed or bound to a surface, entrapped in a gel matrix, or retained in a membrane reactor. All of the above-mentioned cells can be immobilized. Numerous immobilization methods have been reported; they can be divided into the following four groups:

1) Entrapment or encapsulation in a porous polymer network, e.g., polyacrylamide, alginate, κ-carrageenan, gelatine, agar, collagen, cellulose, polyurethane, poly(vinyl alcohol).
2) Covalent, ionic, or physical attachment to an appropriate water-insoluble solid support, e.g., ion-exchange resins, silica gels, metal oxides.
3) Aggregation of cells by physical or chemical cross-linking with glutaraldehyde, polyethyleneimine, or other agents.
4) Retention of cells by membranes (microfiltration, cutoff 0.2–2 µm), e.g., in form of membrane reactors or hollow fiber modules [153].

In successful cases, the immobilized cells retain their activities over several months and show a higher operational stability than unsupported cells. Additional advantages are: 1) easy removal of the cells from the reaction mixture, 2) repeated use of the cells, 3) continuous operation of the reaction, 4) cell regeneration by immersion into an appropriate nutrient medium, and 5) easy product isolation. Some of these features are especially advantageous if the collection of a sufficient cell mass for the conversion is costly.

However, disadvantages must also be considered: 1) The catalytic activity of cells usually decreases during the immobilization procedure because the cells are partly damaged and permeability is further inhibited. This decreased activity can be compensated by increasing the cell density. 2) The immobilization itself sometimes requires special equipment. The conversion step, especially when operated continuously, also requires a special engineering design compared to conventional conversion processes. Examples of successful large-scale applications are listed in Table 11.

An advance in this field was the introduction of an immobilized system of growing cells

Table 11. A selection of commercial applications of immobilized biocatalysts: microbial cells or enzymes

Microorganism or immobilized enzyme	Methods of immobilization	Applications	Year of introduction	Ref.
L-Amino acylase from *Aspergillus oryzae*	adsorption to DEAE-Sephadex	optical resolution of racemic amino acids	1969	[146]
Escherichia coli (aspartase)	entrapment in χ-carrageenan gel	continuous production of L-aspartic acid from ammonium fumarate	1973	[147]
Brevibacterium ammoniagenes (fumarase)	entrapment in χ-carrageenan gel	continuous production of L-malic acid from fumaric acid	1974	[148]
Escherichia coli (aspartase) and *Pseudomonas dacunhae* (aspartate -decarboxylase)	entrapment in χ-carrageenan gel	continuous production of L-alanine from ammonium fumarate	1982	[149]
Penicillin acylase from *Escherichia coli*	entrapment in a cellulose triacetate matrix	production of 6-amino-penicillanic acid from penicillin G	1973	[150]
Glucose isomerase from *Bacillus coagulans* or *Streptomyces* species	many methods available	production of high-fructose syrup from glucose	1973	[151]
β-Galactosidase from *Lactobacillus bulgaricus, Aspergillus oryzae*, etc.	many methods available	production of low-lactose milk and production of sweetenings	1977	[152]
Whole cells *Arthrobacter sp.* (oxygenase)	retention by microfiltration in cross-flow membrane reactor	production of muconic acid (raw material for new resins, pharmaceuticals and agrochemicals)		[153]

into a conversion process. For example, cells entrapped in -carrageenan are viable and reproduce without loss of activity per unit cells [154]. The continuous production of ethanol from glucose by *S. cerevisiae* in κ-carrageenan [155] and the Δ^1-dehydrogenation of cortisol by *Arthrobacter simplex* in calcium alginate [156] are examples of such successful operations. Further improvement of this technique is now expanding into the application to complex multistep reactions, such as the production of antibiotics.

6.5.5. Biotransformation with Cell-free Enzymes or Purified Enzymes

In general, the use of cell-free enzymes or purified enzymes for conversions is expensive, because purification of enzymes is often tedious and time-consuming. Therefore, they are not so suitable for large-scale industrial production. However, there are several cases in which they are advantageous over systems using cells. Such cases are:

1) Poor substrate or product permeability through the cell wall
2) Problems caused by undesired side reactions
3) Commercial availability of the desired enzyme at an acceptable price

If an extracellular enzyme is used, cells are not involved at all.

Cell-free enzymes can also be used as immobilized catalysts [157]. The same techniques as those used for cell immobilization are available. Immobilization of penicillin acylase [150], L-amino acid acylase [146], lactase [152], and glucose isomerase [151] have been applied to commercial production. The entrapment of soluble enzymes in a membrane-type reactor with simultaneous coenzyme regeneration is an alternative to the use of cell-free enzymes. L-Amino acid dehydrogenases in combination with formate dehydrogenase as a coenzyme regenerator and nicotinamide adenine dinucleotide covalently bound to polyethylene glycol in a membrane reactor have been successfully applied to the continuous amination of keto acids yielding optically active amino acids [158]. L-Leucine is already produced commercially by this process.

Whitesides [159] has reviewed the potential applications of cell-free enzymes in organic synthesis.

6.5.6. Multistep Reactions Using Different Biocatalysts

In order to perform more than two sequential reactions, two or more different biocatalysts are sometimes required either simultaneously or sequentially. Theoretically, any combination of catalysts can be used for such multistep reactions. Examples of successful applications are: continuous production of L-alanine from ammonium fumarate via L-aspartate as the intermediate by immobilized *Escherichia coli* (aspartase) and *Pseudomonas dacunhae* (L-aspartate β-decarboxylase) cells [149], complete condensation of racemic homocysteine by washed or dried *Pseudomonas putida* (racemase) and *Alcaligenes faecalis* (*S*-adenosylhomocysteine hydrolase) cells [160], complete hydrolysis of racemic α-amino-ε-caprolactam (see Section 6.6.2.1), and continuous amination of α-keto acids to L-amino acids by two types of cell-free enzymes (see Section 6.5.5).

Another remarkable example is the synthesis of coenzyme A from pantothenic acid, L-cysteine, and ATP (or AMP), which proceeds through five sequential steps (see below) and is commercially performed solely with *Brevibacterium ammoniagenes* cells [161, 162].

Step 1:	pantothenic acid + ATP $\rightarrow$ phosphopantothenic acid + ADP
Step 2:	phosphopantothenic acid + L-cysteine + ATP $\rightarrow$ phosphopantothenoyl-L-cysteine + ADP + inorganic phosphate
Step 3:	phosphopantothenoyl-L-cysteine $\rightarrow$ phosphopantetheine + CO_2
Step 4:	phosphopantetheine + ATP $\rightarrow$ dephosphocoenzyme A + inorganic pyrophosphate
Step 5:	dephosphocoenzyme A + ATP $\rightarrow$ coenzyme A + ADP

6.5.7. Multiphase Reaction Systems

Since their discovery, enzymes have been thought to be inert catalysts in organic solvents. This is basically true, but enzymes have become available that are catalytically active in water-miscible organic solvents. Many reactions that are impossible in water because of kinetic or thermodynamic reasons can be performed in organic solvents [163, 164] (e.g., transesterification reactions [165]). Chemically modified peroxidase exhibits 21 % of its activity in benzene

as compared to that of the unmodified enzyme in aqueous solution [166]. Polyphenol oxidase converts nearly 100 % of phenol to catechol when the reaction is carried out in chloroform containing only 1–2 % water [167]. Such liquid–liquid two-phase systems with low water content can also be used to shift the hydrolytic equilibrium toward water elimination. The synthesis of esters, amides, and peptides by this method shows great practical potential.

Multiphasic conversion systems using microbial cells as catalysts have also been extensively studied, especially in the case of lipophilic compounds, which have a limited solubility in water. In some cases, increased reaction rates and/or product yields have been achieved.

6.6. Process Design

6.6.1. General Considerations

When designing a biocatalytic process many important aspects require careful consideration. Besides the essential question of the type of enzyme used for biocatalysis there is the selection of a compound to be synthesized and a survey on available substrates and routes or reactions that are of special importance in designing a conversion process (also see Chapter 5).

6.6.1.1. Evaluating Enzyme Potential

When a new conversion or a new enzyme useful to biocatalytic conversions is discovered, the question as to the potential of the new enzyme for practical purposes must be answered first. An example for a high-potential biocatalysis with practical application in the pharmaceutical industry is L-phenylalanine dehydrogenase (PheDH, EC 1.4.1.20) from *Thermoactinomyces intermedius* for the conversion of ketoacid acetals to acetal amino acids [168] (Fig. 22). The amino acid acetal is a precursor for Omapatrilat®, an antihypertensive drug, which inhibits the angiotensin-converting enzyme (ACE) and neutral endopeptidase (NEP). The synthesis is carried out in 16,000 L batch reactions with heat dried *E.coli* cells containing the cloned and overexpressed PheDH from *Thermoactinomyces intermedius* and the formate dehydrogenase from *C. boidinii* for cofactor regeneration. Yield (97 %), conversion (97 %), and optical purity of the product (ee > 99 %) are very high.

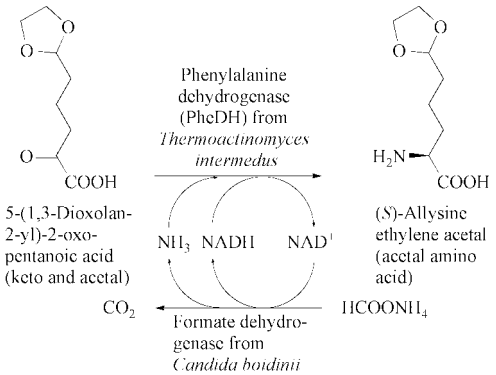

Figure 22. Synthesis of acetal amino acid (Bristol-Myers Squibb)

Another interesting example is the production of L-tyrosine and L-dopa (dihydroxyphenylalanine) by bacterial β-tyrosinase [169]. This enzyme has long been known to catalyze the α, β-elimination of L-tyrosine to phenol, pyruvate, and ammonia (Eq. 1).

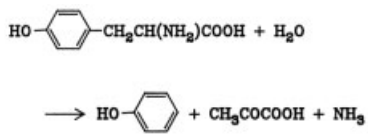

Because the enzyme also catalyzes the reverse reaction (Eq. 2) and the β-replacement reaction of L-tyrosine with phenol derivatives (Eq. 3), it did not take long for the application of these reactions in the synthesis of L-tyrosine and L-dopa.

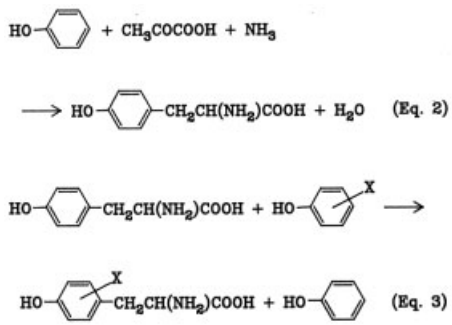

Subsequently, similar reactions were found to be catalyzed by several pyridoxal phosphate dependent enzymes, such as tryptophanase, cysteine desulfhydrase, and 3-chloro-D-alanine chloride lyase. These enzymes were then successfully used to synthesize L-tryptophan and 5-hydroxy-L-tryptophan, L-cysteine, and D-cysteine [169 – 171] (see Table 12). These reactions can be summarized as follows:

$$\mathrm{L\,(D)-RCH_2CH\,(NH_2)\,COOH+H_2O} \rightarrow \mathrm{RH+CH_3COCOOH+NH_3}$$

$$\mathrm{L\,(D)-RCH_2CH\,(NH_2)\,COOH+R'H} \rightarrow \mathrm{L\,(D)-R'CH_2CH\,(NH_2)\,COOH+RH}$$

$$\mathrm{R'H+CH_3COCOOH+NH_3} \rightarrow \mathrm{L\,(D)-R'CH_2CH\,(NH_2)\,COOH+H_2O}$$

where for β-tyrosinase R = hydroxyphenyl, –OH, –SH, –Cl and R′ = hydroxyphenyl; for tryptophanase R = indolyl, –OH, –SH, –Cl and R′ = indolyl; for cysteine desulfhydrase R = –SH, thiol radicals, –Cl and R′ = –SH, thiol radicals; and for 3-chloro-D-alanine chloride-lyase R = –Cl, – SH, thiol radicals and R′ = –SH.

These successful examples demonstrate the importance of reassessing the capabilities of well-known enzymes or reactions.

Table 12. Synthesis of L-tyrosine, L-tryptophan, L-cysteine, D-cysteine, and related amino acids by pyridoxal phosphate-dependent enzymes

Product	Yield,		Substrates	Reaction
	g/L	mol%		Enzyme Microorganism
L-Tyrosine	58	88 [a]	sodium pyruvate ammonium acetate phenol	reverse of α,β-elimination β-tyrosinase *Erwinia herbicola*
L-Tyrosine	53.5	78 [b]	D,L-serine ammonium acetate phenol	β-replacement β-tyrosinase *Erwinia herbicola*
L-Dopa	58.5	[f]	sodium pyruvate ammonium acetate pyrocatechol	reverse of α,β-elimination β-tyrosinase *Erwinia herbicola*
L-Dopa	53	71 [c]	D,L-serine ammonium acetate pyrocatechol	β-replacement β-tyrosinase *Erwinia herbicola*
L-Tryptophan1	00	100 [d]	sodium pyruvate ammonium acetate indole	reverse of α,β-elimination tryptophanase *Proteus rettgeri*
5-Hydroxy-L-tryptophan	23.3	57 [e]	sodium pyruvate ammonium acetate 5-hydroxyindole	reverse of α,β-elimination tryptophanase *Proteus rettgeri*
L-Cysteine	50	88[f]	3-chloro-L-alanine sodium sulfide	β-replacement cysteine desulfhydrase *Enterobacter cloacae*
D-Cysteine	22	88[g]	3-chloro-D-alanine sodium hydrogensulfide	β-replacement 3-chloro-D-alanine chloride lyase *Pseudomonas putida*

[a] Based on sodium pyruvate.
[b] Based on D,L-serine.
[c] Based on indole.
[d] Based on 5-hydroxyindole.
[e] Based on 3-chloro-L-alanine.
[f] Based on 3-chloro-D-alanine.
[g] Sodium pyruvate was added at 2-h intervals to maintain a concentration of 5 g/L for 48 h.

6.6.1.2. Finding Suitable Enzymes

Occasionally a synthetic scheme has been designed but suitable enzymes or conversion processes are yet unknown. In such a case, it becomes necessary to assay suitable microbial strains or enzymes (as described in Section 6.6.2) as to their suitability for catalyzing the desired reaction. The production of L-lysine from D,L- α- amino- ε-caprolactam (ACL) is another typical example [172]. This process was designed to utilize cyclohexene, a major byproduct of nylon production. In this process, the biocatalyst is required to hydrolyze only the L-form of α-amino- ε-caprolactam (L-ACL) to yield L-lysine (Fig. 23).

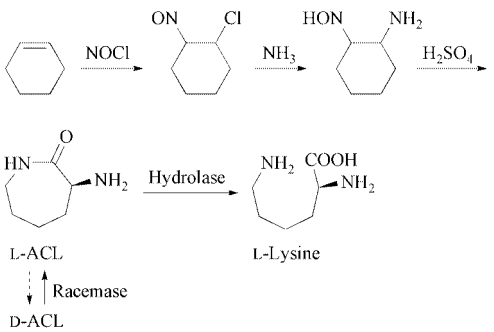

Figure 23. Hydrolysation of the L-form of -amino- -caprolactam (l-ACL) to yield L-lysine

Two microorganisms were found during the assay: one of them hydrolyzes the L-form of the substrate and the other racemizes the remaining D-form. As a result, all the α-amino- ε-caprolactam is converted into L-lysine without requiring any optical resolution.

6.6.1.3. Substrates

Any substrate is theoretically suitable for biotransformations, provided the substrate molecules come into contact with the enzyme. Even gases, such as methane, are suitable substrates; they are bubbled into the reaction medium. The substrate should be soluble in the medium and able to pass through the cell membrane unless the reaction is catalyzed by cell-free enzymes. The substrate is usually added to the reaction medium, neat or as a concentrated solution. Sterilization of the substrate is recommended prior to its use, if required.

Only dissolved substrate is converted. Therefore, slightly soluble substrates must first be dissolved and the products may then crystallize from the solution after conversion:

$$\text{Substrate (solid)} \rightleftharpoons \text{Substrate (solution)}$$
$$\rightleftharpoons \text{Product (solution)} \rightleftharpoons \text{Product (solid)}$$

For example, powdered cortisol in concentrations up to 500 g/L is converted to crystalline prednisolone with a good yield by *Arthrobacter simplex* [173]. Substrates may also be dissolved in water-miscible organic solvents. The lower alcohols, acetone, dimethylformamide, and dimethylsulfoxide are suitable solvents. Surfactants may also be used to disperse the substrate. The combination of water-miscible solvents and surfactants may offer some advantage. In some cases, chemical modification of the substrate may improve its solubility.

An example for a process using substrate with poor solubility is the regioselective oxidation of the *tert*-butyl group of terfenadine using *Cunninghamela blakesleana* in a whole-cell biotransformation [174]. Due to its low solubility in water, microcrystalline terfenadine is solved in the water-miscible organic solvent dimethylfomamide prior to its addition to the aqueous reaction medium. This regioselective oxidation was studied as a biocatalytic alternative for the chemical synthesis of the antihistaminic drug fexofenadine.

In cases where the cell membrane is impermeable to the substrate or a cell-free or purified biocatalyst is required, the permeability of the cell must be improved, respectively, the cell wall has to be disrupted. This is accomplished by air drying, acetone drying, lyophilization, autolysis, lysozyme digestion, surfactant treatment, osmotic shock, freezing and thawing, or ultrasonic treatment. For cell-free preparations, the cell debris can be removed by centrifugation or microfiltration.

If the substrate does not suit the desired conversion, its chemical modification should be considered. Addition, variation, or removal of a protecting group or modification of a functional group in the substrate molecule sometimes makes its interaction with the enzyme more efficient. In addition, these modifications may prevent undesirable side reactions or degra-

dation. This method is frequently used in steroid conversions. A systematic assay for a suitable chloroacetoacetate ester for enantioselective reduction yielding the L-form of 4-chloro-3-hydroxybutanoate, a promising precursor for the chemical synthesis of L-carnitine, is an example [175].

6.6.1.4. Media

Besides performing biocatalysis in aqueous media, there has developed the so-called "nonaqueous enzymology" which has become an important area of research and development during the last decades [176]. Enzymes exhibit a wide array of novel reactivities and selectivities in nonaqueous solvents. For example, many reactions that are impossible in water due to kinetic or thermodynamic reasons can be performed in organic solvents, [163, 164, 177] due to the suppression of water-induced side reactions. Improved and altered substrate specificities [178, 179] and selectivities can be observed. Examples of practical applications are enantioselective synthesis [180], chiral resolution [181], and combinatorial biocatalysis [182]. The possibility of the solubilization of hydrophobic substrates or products in organic solvents opens opportunities for the enzymatic production of poorly water-soluble fine chemicals and pharmaceuticals. The thermal and storage stability of enzymes can be significantly enhanced in nonaqueous media [164, 176, 178].

6.6.2. Selection of Biocatalysts

More than 8000 enzymes are known [125]. Almost all types of chemical reactions which are also catalyzed by normal chemical catalysts are involved. According to the desired type of conversion and the classification of the enzymes one can easily determine which enzyme would be suitable for the desired conversion by inspecting the enzyme catalogue published by the committee [125]. However, the properties of enzymes from different sources may vary widely, even if they catalyze the same reaction. Therefore, testing is indispensable. A wide variety of biocatalysts are tested for the purpose of determining their ability to carry out the desired reaction. Such a test, also called *screening*, may be one of the most important steps for a successful biocatalytic conversion.

6.6.2.1. Screening

Kitahara and co-workers [183] screened bacterial strains from cultures grown aerobically at 30–37°C in nutrient broth containing malt extract. Sodium fumarate and ammonium chloride were added 24 h after growth started. After incubation for another 24 h the L-aspartate in the culture broth was analyzed by chromatography. A strain of *E. coli* was selected by this screening as the most promising producer of aspartase. Under suitable reaction conditions, 56 g of L-aspartate was produced per 100 mL with a molar yield of 99 % by using 1 g of dried cells. Subsequently, this process was developed into large-scale technology by which L-aspartate was synthesized continuously with *E. coli* cells immobilized on κ-carrageenan [184].

The screening procedure described above may be rather simple, but it reflects the essence of screening. Normal screening procedures can be more complicated and deal with a great variety of microorganisms, including bacteria, actinomycetes, molds, yeasts, and basidiomycetes.

6.6.2.2. Enrichment

Another method to find a suitable biocatalyst can be *enrichment*. Sometimes it is desired or necessary to find biocatalysts capable of performing a specific conversion from natural sources. Soil samples containing mixed microbial populations are rich sources of test organisms. Because the enzymes in these microorganisms modify or degrade a great variety of complex organic compounds, at least one of the microorganisms is expected to offer one or several enzymes that perform the desired conversion. These microorganisms are selected using the *enrichment technique*. A typical enrichment procedure is the isolation of bacterial strains that perform the asymmetric hydrolysis of D,L-2-amino- Δ^2-thiazoline-4-carboxylic acid (D,L-ATC), an intermediate of D,L-cysteine synthesis, to L-cysteine [185]. Soil samples were incubated in a medium containing 0.3 % D,L-ATC as the sole

nitrogen source. After spreading each culture on a plate of the same medium, colonies that appeared were isolated and grown on nutrient agar containing 0.2 % D,L-ATC as the enzyme inducer. The resulting cells were then transferred to a solution containing 1 % D,L-ATC and the formation of L-cysteine was demonstrated by chromatographic or microbiological methods. Among 1975 colonies from 388 soil samples, 31 L-cysteine-producing bacteria were isolated. One of them, *Pseudomonas thiazolinophilum*, was mutagenized. Mutants with inhibited L-cysteine metabolism were screened from the mutagenized cultures. One of these mutants converted the added substrate almost stoichiometrically to L-cysteine with a yield of 31.4 mg/mL [186].

The enzyme system of *Pseudomonas thiazolinophilum* hydrolyzes D,L-ATC in two steps as shown in Figure 24.

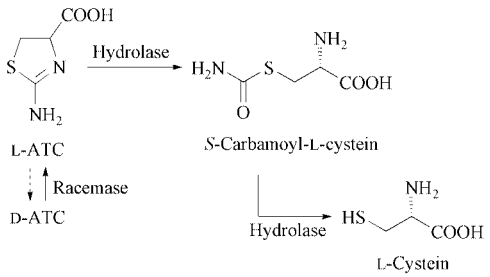

Figure 24. Asymmetric hydrolysis of D,L-2-amino- 2-thiazoline-4-carboxylic acid (D,L-ATC) to L-cysteine

Although screening sometimes is regarded as old-fashioned, time-consuming, and even illogical, some consider it to be one of the most promising methods for finding new genes in microorganisms.

6.6.2.3. Molecular Engineering

Sometimes enzyme yields in wild-type organisms are low or the desired enzyme is not found in the available microorganisms. In this case, molecular engineering techniques can lead to an enzyme catalyzing the desired reaction. There are two different strategies that represent the "state-of-the-art" technologies: *rational design* and *directed evolution* [153]. Both technologies require different knowledge of the target enzyme as well as laboratory equipment (Table 13).

Table 13. Requirements and molecular methods for rational design and directed evolution technique

	Rational design	Directed -evolution
Requirements:	3D structure of the target enzyme	high-throughput-screening/selection technology to analyze enzyme properties
	knowledge of reaction mechanism	
	modeling of enzyme-substrate interaction	
Molecular methods:	site-directed mutagenesis	random mutagenesis

6.7. Improvement of Conversion Processes

After an adequate process has been designed, it must be optimized. This includes improvement of the strain and the search for optimal growth (highest enzyme yield and maximum activity) and optimal reaction conditions.

If the enzyme to be used is constitutive, the amount of biomass containing that enzyme will have to be increased. This is usually sufficient because the constitutive enzyme is formed independently of the medium composition. However, if the enzyme is inducible, an effective inducer, which may be the substrate itself or another structurally related compound, will be required. If the enzyme is repressive, the repression factor will either have to be eliminated or the repressive compounds, which may be present in the growth medium or formed during the growth, must be removed.

Catabolites of an easily metabolizable carbon source, e.g., glucose or sucrose, and the end product of a metabolic pathway usually cause catabolite or product repression, respectively. Therefore, the composition of the growth medium, growth period, and various physical parameters of growth, such as temperature, aeration, agitation, and pH, must be carefully controlled. In successful cases, the concentration of the required enzyme reaches more than 10 % of the total soluble protein [187].

Optimal physical and chemical conditions for the biotransformation itself must be determined with the same care as for the optimization of growth conditions. This is especially important

if side reactions are observed, because they decrease product yield and make product isolation difficult. Physical or chemical treatment of the biocatalyst, such as heating, pH change, and the addition of detergents, organic solvents, or specific inhibitors may cause specific inactivation of undesired enzymes. For example, the *trans-N*-arabinosylation of adenine from synthetic uracil arabinoside (Fig. 25) yields adenine arabinoside, an antiviral agent.

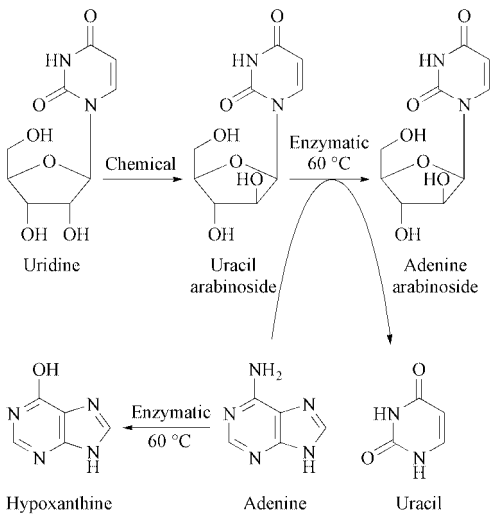

Figure 25. *trans-N*-arabinosylation of adenine from synthetic uracil arabinoside

This reaction is catalyzed by cells of *Enterobacter aerogenes* and takes place only at temperatures above 60°C [188]. The reaction does not proceed below 50°C because hypoxanthine is formed by adenine deaminase before the *trans-N*-arabinosylation by two kinds of nucleoside phosphorylases takes place. Fortunately, these phosphorylases are still active at 60°C, whereas the deaminase is completely inactivated at this temperature. Other examples are summarized in Table 14.

In the case of immobilized biocatalysts, concentration gradients of substrates and products may occur that could lead to decreased biotransformation rates. These gradients can be reduced by decreasing the *Thiele modulus*, e.g., by using smaller carrier particles or by reducing the diffusion distance between the free solution and the immobilized enzymes.

Microbial conversion can also be optimized by improving the selected strain. Standard methods of strain improvement include the isolation of single colonies, optionally after mutagenesis, and screening of the best suited isolates. Such conventional genetic manipulation techniques have been developed during the last decades for the improvement of useful strains. Recombinant DNA technology is providing new prospects for strain improvement. Plasmid or phage vectors can transfer DNA fragments of microbial, plant, animal, or even synthetic origin into a suitable recipient microorganism. These gene cloning techniques are still restricted to microbial strains such as *E. coli, B. subtilis*, and *S. cerevisiae*, in which the genetic background and the appropriate host–vector systems are well known, but their practical application is rare [191]. However, this technology will undoubtedly provide one of the most important keys for the future development of microbial conversions.

6.8. Conclusion and Outlook

Biocatalysts have become one of the most important tools in nearly all fields of industrial biotechnology. The main advantages of using biocatalysts are their unique properties such as high chemo-, regio-, and stereoselectivity and the ability to catalyze reactions under mild conditions.

During the last couple of years, many new methods and techniques have been developed to overcome common problems. Some examples here might be molecular engineering techniques such as rational design or directed evolution to increase low biocatalytic activities [153], biocatalysis in nonaqueaous reaction media to bypass solubility problems of substrate and/or product [179], or the applications of high-throughput screening methods for the search of new and suitable biocatalysts. Another field of interest that will become of great impact is "metagenomic research" that will enable scientists to find new biocatalysts with desired qualities [192].

To make new biocatalysts attractive for industrial applications they have to be available in sufficient amounts. High expression rates of enzymes can be achieved by genetical and physiological manipulations of the expressing mi-

Table 14. Elimination of side reactions by physical or chemical treatment

Desired reaction	Side reaction	Treatment	Reference
Ammonium fumarate → L-aspartic acid (*Escherichia coli*/aspartase)	fumaric acid → L-malic acid	incubation of culture broth at pH 5 and 45 °C for 1 h	[149]
Fumaric acid → L-malic acid (*Brevibacterium ammoniagenes*/fumarase)	fumaric acid → succinic acid	treatment of immobilized cells with 0.6 % bile extract	[148]
L-Aspartic acid → L-alanine (*Pseudomonas dacunhae*/aspartate β-decarboxylase)	L-alanine → D-alanine	incubation of culture broth at pH 4.75 and 30 °C for 1 h	[149]
Cholesterol → 1,4-androstadiene-3,17-dione (*Arthrobacter simplex*/multistep conversion)	further decomposition of the sterol molecule	addition of α,α'-dipyridyl (1 mM) to the culture broth	[137, 189]
β-Sitosterol → 17-ketosteroids (*Nocardia* species/multistep conversion)	further decomposition of steroid ring	addition of α,α'-dipyridyl (0.3 mM) to the conversion mixture	[190]
D,L-2-Amino-Δ^2-thiazoline-4-carboxylate → L-cysteine (*Pseudomonas thiazolinophilum*/multistep conversion)	further degradation of L-cysteine	addition of hydroxylamine or semicarbazide to the reaction mixture	[186]

croorganism such as optimizing the number of plasmids per cell or inactivation of "negative" cell functions (e.g., deletion of protease coding genes). Apart from deleting unwanted side activities also novel metabolic pathways can be implemented. These microorganisms, customized only to fulfill special needs for biocatalytic applications, are called "designer bugs" or "tailor-made microorganisms".

7. Downstream Processing

Downstream processing is one of the most underestimated steps in bioprocesses and it is well known especially in the pharmaceutical industry that downstreaming is the most expensive and unfortunately the most ineffective part of a bioprocess. Thus, one can assume that new developments are widely described in the literature. Unfortunately, this is not the case. Only a few working groups focus on new and more effective procedures to separate products from fermentation broths or biotransformations. A chief characteristic of biotechnology is the wide variety of products. Due to this variety, a broad spectrum of separation techniques must be applied. However, for nearly all products one starts with a dilute suspension and tries to produce a highly purified dry product. In the case of extracellular products, the solids in this suspension may include intact organisms, other insoluble fractions of the medium or natural sample, and perhaps insoluble products. Concerning intracellular products the solids include in addition fragmented mycelia caused by cell disruption, which is necessary to gain the products. According to this starting point, nearly each downstream process consists of the four following steps [193]:

- Removal of insoluble particles
- Isolation of the product
- Purification
- Polishing

With respect to downstreaming in laboratory scale, one is normally only limited by the available equipment. However, with regard to upscale processes one should have in mind that the developed process should also be applicable in industry. Thus, Belter et al. [193] have formulated the following questions, which are also crucial to biotechnology downstream processes:

- What is the value of the product?
- What is the acceptable product quality?
- Where is the product in each process stream?
- Where are the impurities in each process stream?
- What are the unusual physicochemical properties of the product and the principal impurities?
- What are the economics of various alternative separations?

In addition, it is important to use materials which are available for all upscale processes, since the downstream behavior of the product may change while changing, for example, the type of chromatographic resin or membrane material. Normally the recovery costs exceed the bioprocess costs; however, some upstream parameters influence the downstream processing:

– The characteristic properties of the producing microorganism or cell line
– The location of the product
– The stability of the product
– Byproducts and impurities
– Concentration of the product in the medium from which it is to be recovered

A general downstream scheme based on these parameters is given in Figure 26.

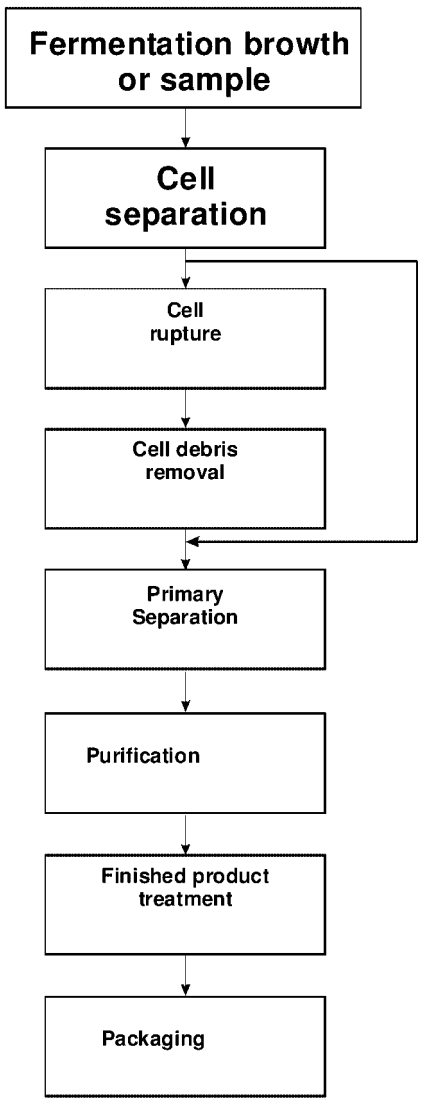

Figure 26. General downstream scheme in biotechnology

The main aims of the primary separation step are to achieve a volume reduction and to make a first stage purification of the product by removing dissimilar components from the broth. The most important techniques in use are:

– Membrane processing
– Ion exchange chromatography

In the membrane process, ultrafiltration and reverse osmosis are often used for separation, concentration, and desalting. In addition, polar membranes are used for ion exchange and desalting. Ion exchange chromatography is applied in order to remove either major contaminants from the broth or the desired product from the broth. Chromatographic procedures – based on columns or membranes – are also often used in the purification step. Other techniques in this part of the downstream process are precipitation and liquid–liquid extraction.

7.1. Sample Disruption

The influence of the sample disruption step in upstream and downstream processes cannot be ignored. The disruption is dependent on the properties of the sample. The main focus of the disruption is always to release as much as possible of the product. However, depending on the mechanism, parts of the product may be destroyed during the disruption process. Thus, very effective and short procedures are required. A distinction can be made between mechanical, chemical, and enzymatic procedures for product release. Mechanical methods are often preferred because of short residence time, lower operating costs, and contained operation [194]. The most common mechanical means of disruption are:

– Homogenizers [195, 196]
– Bead mills [196].

A *homogenizer* consists of a positive-displacement pump, which supplies the liquid sample at high pressure through a small nozzle or an orifice valve. The disruption results from the combination of shear force and impingement on the valve. Bead mills use horizontal grinding chambers filled with glass beads or other resistant materials such as zirconium oxide, zirconium silicate, titanium carbide, etc [197]. The dispensed sample is introduced into the grinding chamber on a continuous basis. The level of

disruption depends on the variable turning speed of the bead mill.

By using *bead mills*, cell disruption can be achieved in a single run with better temperature distribution and temperature control. Another possibility for cell disruption or at least product liberation is the use of microwaves or ultrasound which can be combined with sample extraction by organic solvents. Microwave techniques are widely used in acid digestion of solid samples. Their use in the extraction of organic analytes from environmental samples is less widespread, despite the availability of commercial devices for this purpose and their potential for reducing analysis time and solvent consumption.

A possibility for cell lysis under very mild conditions is the use of *hydrolyzing enzymes*. In addition, enzymes offer selectivity during product release. Enzymes hydrolyse the walls of cells, and when sufficient wall has been removed, the internal osmotic pressure bursts the periplasmic membrane allowing the intracellular components to be released [195]. The effect of lytic enzymes is specific to particular groups of cell types, which is attributed to the differences in cell wall composition. For example, the most efficient lytic enzyme for bacteria is lysozyme from hens' egg. This enzyme is also used in large-scale processes for enzyme production [198]. Even more highly specialized procedures for sample disruption can be applied in biotechnology. For example, high yields of intracellular enzymes from yeast can be obtained by applying a series of electric field pulses [199]. By using this technique, up to 90 % of the total activity can be released without any further or previous treatment of the cells. The method is based on electroinduced changes in the cell envelope leading to a leakage of part of the intracellular proteins without the formation of debris and permits the treatment of large volumes.

7.2. Solid–Liquid Separations

Solid–liquid separation is a very important procedure during downstream processing in biotechnology for cell separation, cell debris removal, and also for product recovery. The most important solid–liquid separation techniques in use for are filtration and centrifugation. Filtration separates solids from a liquid by forcing the liquid through a solid support or filter medium. Dead-end filtration and cross-flow filtration are two different designs of filtration which can be used. In *dead-end filtration* mode, the total process fluid stream flows through the membrane. The retained solids accumulate on the membrane and build up a filter cake. The membrane has to be changed when the membrane pores are clogged by the solids. When the feed flow is directed parallel to the membrane surface, the term *cross-flow filtration* is used. The tangential flow of liquid removes any retained molecules or particles from the membrane surface, which results in a stable flux for a longer time period. By using cross-flow filtration, relatively low shear stress is possible and filter aids are not required. In addition, scale-up is simple and cell washing is possible in a single process step.

Over the last 30 years, a number of further membrane processes have been developed for molecular separation. These filtration techniques can be divided into four major groups: reverse osmosis (hyperfiltration, RO), nanofiltration (NF), ultrafiltration (UF), and microfiltration (MF) [200]. Membrane processes are easy to scale up and the possibility of using the same materials and configurations in different sizes from laboratory to process scale reduces the validation effort enormously. However, filtration processes are limited with regard to selectivity. The fractionation of proteins can only be achieved with large differences in the molecular weight of the proteins and it is important to keep in mind that a certain difference in the molecular weight of two proteins does not mean the same degree of difference in molecular size. Proteins that differ in molecular weight by 10 times may differ in size by only 3 times when in globular or folded form. Another problem encountered when using membrane processes for product recovery can be the slow retentate flux, which can result in the formation of a thick secondary membrane. Another possibility is the strong interaction of the sample with the membrane material. This often depends on unspecific protein adsorption which is affected by several factors [201]. Thus, different membrane types have to be screened to minimize the unspecific binding of the sample when a new filtration procedure has to be developed.

7.3. Product Recovery

One of the most important procedures in product recovery are chromatographic processes. During chromatographic steps the sample is separated into several fractions according to interactions between the different molecules in the liquid phase and the stationary phase (chromatographic resin). The molecules can be separated according to their size and/or charge and based on hydrophobic/hydrophilic or affinity interaction with the stationary phase. Column chromatography utilizes a vertical column filled with the solid support with the sample to be separated placed on top of this support. The rest of the column is filled with a solvent (eluent) which, under the influence of gravity, moves the sample through the column. Differences in rates of movement through the solid medium are translated to different exit times from the bottom of the column for the various elements of the original sample. Based on the specific interactions between the sample in the mobile phase and the stationary phase the column chromatography can be performed as:

- Ion-exchange chromatography,
- Size-exclusion chromatography (gel-filtration chromatography),
- Reversed-phase chromatography, and
- Affinity chromatography.

Immobilized-metal-ion affinity chromatography (IMAC) as a special form of affinity chromatography is a popular and powerful way to purify proteins. It is based on the specific coordinate covalent binding between histidine or other unique amino acids and various immobilized metal ions (e.g., nickel).

Most of the chromatographic systems are run in a batch or quasicontinuous mode. Nowadays, new developments such as simulated moving bed chromatography [202] and anular chromatography [203] offers the possibility of continuous processes. One problem in column chromatography are the difficulties in upscaling. If the column radius is increased, unless special packing techniques are employed, the packing procedure becomes inefficient and the packing itself unstable. In addition, to maintain the optimum mobile phase velocity, the flow rate will need to be substantially increased and the consumption of mobile phase will eventually become economically impractical. Conversely, if the column length is increased, then the impedance to flow will become greater leading to high column pressures. If large column radii are employed, then the mechanical strength of the column system will limit the maximum permissible pressure. Consequently, lengthening the column will eventually require the particle diameter to be increased to provide adequate permeability. Increased particle diameter will, in turn, reduce the column efficiency, which may impair the resolution of the compounds of interest [204].

Membrane. Until today, the required process steps for the product recovery are carried out separately, which leads to many energy-intensive process steps such as filtration or chromatography and production of vast wastewater amounts, respectively. Membrane adsorption allows the integration of several process steps into one-unit operation, which means the downstreaming of biotechnologically produced proteins can be carried out by saving process time and resources [205, 206]. In addition to this, conventional chromatographic techniques which are based on packed columns often require lengthy procedures. This can lead to the degradation of sensitive proteins [207]. Furthermore, packed columns exhibit a high pressure drop across the column and a slow diffusion-controlled binding process of solutes within the matrix [208]. Whereas membrane systems show several advantages to packed columns, most important, mass transfer takes place through convection rather than through diffusion. Due to this fact, membrane adsorbers enable a time-effective performance with high flow rates without high back pressure [209]. Membranes can be converted into efficient adsorbers by attaching functional groups to the inner surface of synthetic microporous membranes. Affinity adsorption, ion exchange, or immobilized-metal affinity chromatography can be carried out by these membranes. Membrane ion exchangers of strong acidic (sulfonic acid), strongly basic (quarternary ammonium), weakly acid (carboxylic acid), and weakly basic (diethylamine) types are commercially available. A chelating membrane based on the iminodiacetate (IDA) group is applicable for IMAC. Membrane adsorber technology has several major advantages compared to classical separation methods. Due

to the membrane structure, the binding of proteins is not limited by diffusional processes, therefore loading and elution can be performed at very high fluxes resulting in very short cycle times. Compressibility of the membrane under normal operation conditions can be neglected, channeling cannot occur, and the pressure distribution inside the modules is designed to have plug flow through the module, all of which lead to sharp breakthrough curves. Scale-up is very easy, materials and systems allow cleaning in place (CIP), and the validation of the process is made easier due to of standard products and validation service of suppliers.

7.4. Solvent Extraction

Solvent extraction is the most common method for the recovery of hydrophilic substances and, therefore, a method for separating well-soluble metabolites from cultivation media. Classical extraction processes use organic solvents, which are often rarely suitable for effective recovery of the solute. Recently, new extractions have been developed which form specific adducts with the metabolite in question and allow its recovery with high efficiency and selectivity [210]. Solvent extraction in biotechnology focuses on the recovery both of primary metabolites (e.g., ethanol, acetic acid, citric acid, and amino acids) and of secondary metabolites (e.g., antibiotics or vitamins). The concentration of secondary metabolites is usually much lower than that of primary metabolites. Since most secondary metabolites are for use as therapeutics, the quality requirements of the products are high. Solvent extraction can help to fulfill these requirements. In addition, the set-up of integrated bioprocesses (production and downstreaming) can be performed by solvent extraction [211]. Supercritical fluid extraction (SFE) becomes also an important tool in biotechnological downstream processing since it offers important advantages compared with other solvent extraction methods. It is possible to work in an oxygen-free system, which prevents oxidation. The low temperatures applied minimize thermal degradation and microbes or their spores are not soluble. In addition, supercritical fluids for extractions are inexpensive. The successful implementation of this technique can lead to improved sample throughput, more efficient recovery of analytes, cleaner extracts, economic replacement of halogenated solvents, and a high level of automation, compared to conventional sample preparation procedures [212]. Supercritical fluid processes are being commercialized in the polymer, pharmaceutical, specialty lubricants, and fine-chemicals industries. Supercritical fluids are advantageously applied to increase product performance to levels that cannot be achieved by traditional processing techniques.

8. Monitoring and Modeling of Bioprocesses

The requirement to operate biotechnological processes in a cost-effective manner by simultaneously maintaining a high product quality and safety is becoming increasingly relevant. Historically, the most important means of achieving increased productivity from bioprocess plant has been through time- and resource-consuming experimental developments. More recently, however, significant advances have been made in the area of computer applications for bioprocess supervision, modeling and control, and their introduction in process industry.

Typically, little use is made of the large amounts of bioprocess online and laboratory assay data after an experiment is completed. Recently, attempts have been made using enhanced automation strategies to improve the quality of information presented to operators and increase the level of automatic process supervision, although few industrial applications have yet been reported. On a plantwide perspective, the scheduling of process operations is also traditionally manually undertaken by experienced personnel by using a trial and error approach with a planning board. This also appears to be a potential area for modern bioprocess management.

Knowledge-based systems, such as fuzzy logic systems, can be designed to cope with uncertainty and allow the coupling of quantitative information with the qualitative or symbolic expressions (in the form of heuristics), so as to reproduce the actions of an experienced process operator. They might therefore be used as an intelligent, online virtual assistant to the process

operator, or, with extra knowledge as a supervisor in running and maintaining bioprocess operations within optimal operating conditions. The main objective in this approach is to develop an expert supervisory system which uses bioprocess and control knowledge in the form of rules in conjunction with bioreactor state estimation (soft sensing), parametric identification and process control algorithms running in real time. In addition, provision needs to be made for process models and known bioprocess behavioral characteristics to be incorporated in order to enhance the overall supervisory control strategy. The complete system aims to perform process monitoring, process control, sensor validation, fault detection, and diagnostic tasks and, in addition, provide recovery advice so as to achieve continuous online optimization.

Collateral with the above approach is the online determination of the critical key variables which govern process production. The measurement and estimation problems are well known and represent the main drawback concerning an effective biosystem optimization. Three ways of reducing the difficulties encountered can be proposed:

- Better sensors
- Better sampling and automatic analysis systems
- Online estimation of difficult to measure, or immeasurable, variables.

Although all of these are attracting research effort, the techniques of online state and parameter estimation are receiving the greatest interest. Whilst developments in biosensors are in some cases helping to contribute to the online determination of bioprocess parameters, they are by no means at the stage where general applicability has been achieved. Until this is the case, the development of online estimation techniques will provide the major means by which online closed loop feedback control of critical bioprocess variables can be achieved.

Several different approaches can be adopted in the development of bioprocess estimation algorithms. The easiest but potentially least accurate way of obtaining an estimate of a "difficult to measure" primary process variable is to establish a correlation with a measurable secondary process variable, whilst ignoring measurement errors, noise, etc. A more robust approach might be to adopt a numerical estimation technique, either for estimating the parameters of a predefined model structure, usually of a generalized linear time series form, or directly to obtain an estimate of a difficult to measure primary process variable (state estimation or software sensor). An alternative approach is to exploit the nonlinear mechanistic structure of the bioprocess, if known, in the form of a state observer.

Neural Computing. A relatively new development in artificial intelligence is the area of neural computing. Neural networks are dynamic systems composed of highly interconnected layers of "simple" neuron-like processing elements. In the field of process engineering, it is their ability to capture process nonlinearities which offers potential benefits. Other useful characteristics are their ability to adjust dynamically to environmental and time-variant changes, to infer general rules from specific examples and to recognize invariances from complex, high-dimensional data. These properties provide neural networks with the potential to outperform other "learning" techniques. Given a series of experimental data the network is able to establish the governing relationships in these training data. This ability can be exploited to aid the nonlinear modeling, control, and optimization of complex processes, as well as being used in existing predictive control methods.

8.1. Characteristics of Bioprocesses

8.1.1. System Definition

A biotechnological process consists of a system of chemical, biochemical, and microbiological reactions, incorporating process-engineering aspects such as mass and energy conservation, thermodynamics, or heat and mass transfer correlations. (Fig. 27). The biological part of the system represents the peculiarity of the system characteristics and causes the great difficulties in considering it from the system theoretical point of view.

Common for all system models is their mapping of inputs to outputs, of stimuli to responses in some with the additional introduction of state variables. Therfore, the task is to define proper inputs and outputs, dependent on the problem

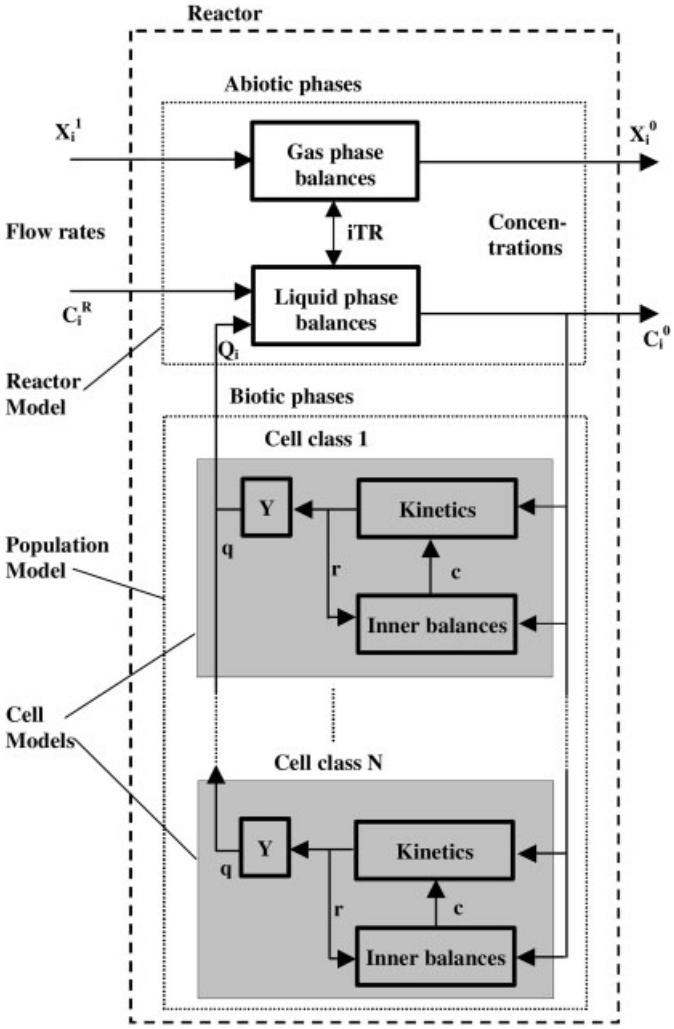

Figure 27. General model structure for biotechnological processes
x = external flow and transport rates; c = concentration of the component $(\cdot)^I$ inlet, $(\cdot)^O$ outlet; TR = transfer rates; Q = volumetric reaction rate; q = specific reaction rate; Y= yield coefficient; r = intrinsic reaction rate; $(\cdot)_i$ = index for quantity i; $(\cdot)_j$ = index for quantity j

available, to use appropriate signals (trajectories) to describe the system behavior in their important aspects and to build up the model.

A biotechnological system can be defined as a set of *i* biochemical reactions involving *m* input variables $\vec{x}(t)$, *o* output variables $\vec{y}(t)$, *m* state variables $\vec{z}(t)$ and *r* parameters $\vec{p}$ notated in operator transformation *T(·)*:

$$\vec{z}(t) = T\left(\vec{x}(t), \vec{p}\right) \tag{1}$$

$$\vec{y}(t) = T^*\left(\vec{x}(t), \vec{z}(t), \vec{p}\right) \tag{2}$$

with

$$\begin{aligned} \vec{z}(t) &= \left[z_1(t) \quad z_2(t) \quad \ldots \quad z_n(t) \right]^T \\ \vec{x}(t) &= \left[x_1(t) \quad x_2(t) \quad \ldots \quad x_m(t) \right]^T \\ \vec{y}(t) &= \left[y_1(t) \quad y_2(t) \quad \ldots \quad y_o(t) \right]^T \\ \vec{p} &= \left[p_1 \quad p_2 \quad \ldots \quad p_r \right]^T \end{aligned} \tag{3}$$

The model does not determine the input variables $\vec{x}(t)$. They represent the systems stimuli and can be divided up into control variables and

random forcing functions. Control variables are those which are chosen by the operator to manipulate the system in such a way the benefits are maximized. Typical examples in bioprocess engineering are: temperature trajectories, agitation rate, or feed rate of key components. Random forcing functions can be seen as outer disturbances of the system whose origin comes from the system environment and not from internal uncertainties, e.g., impurities of the inoculums or the feed medium. To evaluate the system dynamics, it is important to define meaningful trajectories of the control variables. Keeping the values constant will not result in a representative information retrieval of the biosystem. Only if characteristic curves – or specific combinations – are chosen to which the system reacts sensitive a representative process description is possible. Therefore, it is evident that the experimental as well as the theoretical design must be performed from deep problem awareness.

The state variables $\overrightarrow{z}(t)$ represent storage elements for mass quantities (e.g., substrate, products, biomass, inhibitors, interferences), information, or energy of the system. The state vector $\overrightarrow{z}(t)$ is composed of any set of quantities sufficient to completely describe the behavior of that system. Given a state vector at a particular point in time and a description of the system forcing functions, in an appropriate uncertainty formulation, and control functions from that point in time forward, the state at any other time could be computed. In biotechnological processes state variables are frequently defined, if the key quantity of interest can not measured directly and must therefore determined by state estimators from other process variables [213, 214]. The introduction of state variables is not strictly necessary. But in their absence inner relation can only be hardly understood and described. Also a direct mapping of an input vector to an output vector without consideration of state quantities in form of a signal model, as it is typical in conventional signal analysis in electrical engineering, can not be accomplished. Their application depends on the point of view. A signal model provides mainly information about the behavior of changing influencing parameters and how they affect the outputs based upon given inputs. No information can be provided about the intrinsic reactions or the basic mechanism, the focus is outside of any interpretation purposes [215, 216]. Mostly, they are used for software sensors [217, 218] or optimization purposes [219, 220]. Both represent black-box applications (see below), where only the outputs are of special interests, the inner relations or interpretation aspects are secondary.

As output variables $\overrightarrow{y}(t)$, very often quantities are specified which are accessible by online measuring systems or which can be estimated by software algorithms. They can be seen as the result of a measurement model which relates the state variables $\overrightarrow{z}(t)$ to the output variables $\overrightarrow{y}(t)$. These quantities represent the values with whose the system can be evaluated. The output vector $\overrightarrow{y}(t)$ is strongly related to the aspect of observability (see below).

The division into parameters, state variables, or output variables is not an immanent property of a system, but a specific view by the analyzer of the system. It can be changed in relation to the mode of process operation, to the process targets or the simplifying assumptions of the model. A state variable z_i can be exchanged to an output variable y_i and vice versa. A mathematical system should provide a coherent description of the entire bioprocess including all relevant system aspects. The degree of complexity of the bioprocess or the possibility for a simplified modeling of some parts of the system is determined by the intended application and by the skills of the system analyzer.

Parameters $\overrightarrow{p}$ can be seen as fixed quantities of the system. In the ideal case, they should be time invariant, as it is indicated by their original definition. However, sometimes a time dependency of parameters is introduced to gain a better system description. Normally, the parameters are used to interpret the biosystem characteristics. However, the meaning of the parameters can be quite different depending on the point of view from which the biosystem is described. In terms of a mechanistic approach using fundamental physico–chemical equations, the determined parameters, after they are fitted by experimental data, have a direct physical meaning and can be used for interpretation purposes and are comparable between different experiments. If the behavior is characterized phenomenological, by means of descriptive equations, the parameters determined are only valid under the specific boundary conditions of the experiments.

They can be quite different in different experiments and possess nearly no general validity. One must be very careful when comparing different approaches upon the quantities of these parameters.

8.1.2. System Description

A biotechnological system consists in particular of a set of chemical, biochemical, and microbiological reactions whose components, reactions rates, and other characteristics as temperature, pH-value, or internal energy, are mostly not exactly known. The metabolism of the biological unit is a very complicated process, not only because the intracellular reaction network may be quite complex, but also due to the large number of inner relations, control loops on reaction level, regulatory mechanism, and genetic level that are overlaid to coordinate the elementary reactions. So even when the microkinetics could be measured exactly, it would be impossible to establish a correct system map in every detail without the introduction of rigorous simplifications. Another reason for uncertainty is the inhomogenity of the units (e.g., cells) in the whole ensemble, which is normally not considered when characterizing biosystems. The relations are not constant and vary with time. Furthermore, there might be a morphological differentiation biological unit which is accompanied generally by changes in the dynamic behavior.

Therefore, the interesting question is why these complex systems can often be described for specific process targets by a few mathematical equations only. One reason is that the functional blocks of the biosystem operate together in a network of regulation, and exchange of mass, charge, and energy and a few bottleneck processes determine the behavior of the whole system. Another reason is the tremendous number of units in the ensemble of a biosystem which hides individual variations in their growth and leads to a smoothed average behavior. In most cases, the problem shifts. It is not the difficulty to build up the equations, but it is hardly to determine the characterizing parameters, which are generally dependent on time and other system state quantities. Furthermore, a few reactions can already determine the rates of many others. When the main inputs and the final overall outputs are known, a prediction is often even possible with incomplete information on the involved intracellular reaction network [221 – 223].

For analysis and design of those systems, two aspects have to be considered – the biological reactions catalyzed by the bioactive compound, and the numerous chemical and physical processes which precede, accompany, and follow them. Within the biosystem, the most important physical processes, which are intimately bound to the biological reactions, are associated with transport of material, energy, and information to and from the location of the bioactive unit. The dependency of the biological reactions on the microenvironment around the unit is called *microkinetics*. Due to the transport phenomena, the microenvironment will vary along different locations in the systems. The integral description of microkinetics in connection with the transport processes that may be included in a system model is called the *macrokinetics*. Ideally, one would always attempt to model the transport phenomena and the microkinetics of the model, because this model could be combined with proper reactor models to predict the macrokinetics of different biosystems. Unfortunately, at present, the effort for simulation of detailed reactor models is rather high and methods for analyzing the microenvironment of the cells are still rare. Therefore, what is usually observed is always a kind of macrokinetics. To find a compromise between these facts and our limited knowledge about the process, the description of the biological system is often characterized by means of so-called *formalkinetic* approaches. This means a formal application of system equations that represent actually microkinetics to a process, where only averaged macrokinetics can be measured. As an immediate consequence, the model parameters will change when the operating conditions of the reactor are changed, or even more, if a different reactor with changing boundary conditions is used. This necessity limits the predictive power of *formalkinetic* models for biotechnological processes significantly.

In principle, independent of the available system volume, the mass balance of a component i from the general point of view can be described using a finite volume approach solving the governing equations for fluid flow and heat and mass transfer:

$$\frac{\partial c_i}{\partial t}+div\left(c_i\ \vec{u}\right)-div\left(\Gamma\ gradc_i\right)-\frac{dc_i}{dt}=0 \qquad (4)$$

where the dependent variable is denoted by the concentration c_i and $\vec{u}$ represents the velocity. Both Γ, the diffusion coefficient, and dc_i/dt, the source term, are specific to a particular meaning of a component *i*. From this general description, a one-dimensional model can be derived. In the presented approach a dynamic process description is the goal rather than a rigorous three-dimensional simulation of a compartmented system. For describing biological systems a coupling with the involved chemical, biochemical, and microbiological reactions is necessary.

In the considered (finite) volume, homogeneous (bio-)chemical reactions system can be analyzed according to Gavalas [224]. He assumes a detailed knowledge of the chemical behavior of all the system components. The dynamic of those reaction systems is described by a set of ordinary differential equations, by using the idea that every isolated chemical kinetic system should be consistent with the conservation of atomic species and of internal energy. In principle, this approach can also consider regulatory and closed-loop effects at the component level. A general stoichiometric exact model as a closed homogeneous reaction system of constant (finite) volume can be described as follows.

$$\frac{dc_i\left(t\right)}{dt}=\sum_{j=1}^{M}k_{ij}\varphi_j\left(c_1\left(t\right),\ldots,c_N\left(t\right),T\left(t\right)\right) \qquad (5)$$

where

$(\cdot)_i$	Component *i* with *i* = 1, . . . , *N*
$(\cdot)_j$	Reaction *j* with *j* = 1, . . . , *M*
$c_i(\cdot)$	Concentrations of the different (bio-)chemical components *i*
$j_j(\cdot)$	Reaction rate of the *j*-th reaction
$T(\cdot)$	Temperature
k_{ij}	Exact stoichiometric coefficient from component *i* in reaction *j*

If we consider a reaction in a volume with possible in- and outflow, the above approach must be extended with the balances of the internal energies

$$\frac{\partial T}{\partial t}+div\left(T\ \vec{u}\right)-div\left(\Gamma\ gradT\right)-\frac{dT}{dt}=0 \qquad (6)$$

$$\frac{dT\left(t\right)}{dt}=\frac{1}{c_p}\sum_{j=1}^{M}H_j\varphi_j\left(c_1\left(t\right),\ldots,c_N\left(t\right),T\left(t\right)\right) \qquad (7)$$

where

H_j	Reaction enthalpy of *j*-th reaction
c_p	Averaged specific heat capacity of mass in the considered volume

In this model, all the components involved in the process reactions take part in the mathematical consideration. Furthermore, Gavalas assumes that the *j*-reactions are independent of each other. The mass conservation and more precisely the conservation of atomic species impose certain relations between the stoichiometric coefficients k_{ij}.

In this case, a so-called *forward problem* is considered. A forward problem is one in which the parameters and starting conditions of the system, and the kinetic or other equations which govern its behavior, are known and the equations can be solved by typical numerical methods. From a mathematical and physical point of view, it can be seen as the calculation of the "effect of a cause" and not to estimate the "cause of an effect", as it is accomplished in an inverse problem. In other words, we usually know how to use mathematics and physical reasoning to describe what would be measured if conditions were well known [225, 226]. Most mathematical models in fluid dynamics are of the "forward" type; the relevant properties of the aquifer or reservoirs are assumed to be known, as well as the initial and boundary conditions. A model then predicts the resultant flow. This is typically the approach used to map a scenario caused by lot of complex interacting partial processes or taken in sensitivity studies, which are quite useful, and can show what the most important features or processes are likely to be for a site [227 – 229]. Following the ideas of forward problems, structured or cybernetic approaches for the description of biological systems can also be mentioned. They structure the cell mass into several intracellular compounds and functional groups which are connected and regulated via an optimal criterion to each other and to the environment by fluxes of material and information. The functional groups may, in one extreme, consist of detailed reaction network considering as many as known reactions and regulatory loops [230, 231].

Mostly, however, in bioprocess engineering when microbiological and biochemical reactions are involved, the exact stoichiometry is unknown. In the engineering literature of the last 20 years a less precise mathematical model for

biotechnological processes is encountered [232, 233]. The so-called "general reactor model" – firstly introduced by Bastin and Dochain [232] – tries to unify chemical, biochemical and microbiological process using the same approach. The bioprocess, an n-dimensional system of ordinary nonlinear differential equations, is based on a simple representation of the biotechnological processes, where superfluous details are omitted. Only the dynamics of those components that play an important role in the process are considered, irrelevant byproducts or nonlimited substrates are neglected. The stoichiometry of the chemical reactions is only qualitatively taken into account by means of yield coefficient $Y_{\zeta(\cdot)/c(\cdot)}$ or stoichiometric coefficients k^*_{ij} (Eqs. 8 and 9).

$$\frac{d\varsigma_i(t)}{dt} = \sum_{j=1}^{M^*} k^*_{ij}\varphi_j\left(\varsigma_1(t), \ldots, \varsigma_N(t), T(t)\right) \tag{8}$$

$$Y_{\varsigma_j/c_i^*} = \frac{\sum_{J=1}^{M^*} \frac{d\varsigma_j}{dt}}{\frac{dc_{i^*}}{dt}} \tag{9}$$

where $c_i(\cdot)$ and $\zeta_i(\cdot)$ are concentrations of the different (bio-)chemical components, $c_i(\cdot)$ are those substances (i^*=1, ... ,N^{**}) whose reactions are calculated by the yield coefficients $Y_{\zeta(\cdot)/c(\cdot)}$, ζ_i are those components (j=1, . . ., N^*) whose turnover are determined by the dominating reaction rates ψ_j.

M^*	Number of the dominating reactions in the considered volume
$\psi_J(\cdot)$	Reaction rate of the j-th reaction
$T(\cdot)$	Temperature

In general, these coefficients do not have any mechanistic meaning in contrast to the above-mentioned approach, instead they represent formalkinetic quantities, which must be estimated, determined experimentally, or established by practical experiences. This allows the inclusion of chemical, biochemical, and microbiological processes in a unified approach. However, the reaction scheme may be inconsistent with the law of conservation of mass. Also cell-to-cell heterogeneity, e.g., due to different proliferation phases, are not considered and simplifies a segregated viewpoint to an unsegregated perspective.

The reactions rates can be often a very complex function of the operating conditions and of the state of the biosystem. In the case were ψ is proportional to the specific growth rate of the organism μ, there exist several possible models [234]. This approach is very common in literature (overviews in [235, 236]) and can be assigned – in contrast to the above-mentioned approach – to the class of inverse problems.

It is typical for *inverse problems* that in many situations the quantities that we wish to determine are different from the ones which we are able to measure. If the measured data depend, in some way, on the quantities we want, then the data at least contain some information about those quantities. Starting with the data that we have measured, the problem of trying to reconstruct the quantities that we really want to know is called an inverse problem; we measure an effect and want to determine the cause. Typical applications of inverse problems are model identification and estimation, image analysis, numerical analysis, or navigation [237, 238].

8.1.3. Dynamics of Biosystems and Real-Time Considerations

Real-time considerations play an important rule dealing with process control strategies. A notorious underestimation of the dynamic properties of microbial and cellular populations exists and results mainly from matching the duration of the respective batch cultivations to the relevant time constant of the biosystem under investigation. However, metabolic regulation of enzyme activities and fluxes often takes place on a time scale of seconds rather than days although the latter may also be true for some processes. It is therefore in the scope of promoting biotechnological research to adopt and develop appropriate experimental concepts, methodologies and equipment in order to consider all relevant time constants [239].

What is fast? What is slow? What is relevant? The last question is the most important when dealing with modeling. The relaxation time concept of Harder and Roels [240] (Fig. 28) maps typical time constants of microbial and cellular control on the level of modification of enzymes (activation, inhibition, dis-/association of subunits, covalent modification or digestion) to the range of milliseconds to seconds, on the level of regulation of gene expression (induction, repression, or derepression of transcription) to min-

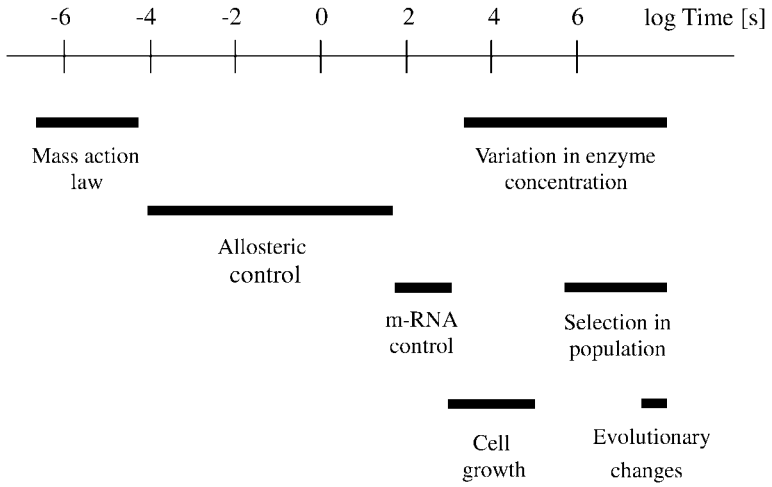

Figure 28. Concept of relevant relaxation times and time constants (according to [240]). Note that the time scale is logarithmic

utes, on the level of population selection and evolution to days and larger units. Considering mainly the growth of microorganisms, typical time scales of 0.3 to 20 h are relevant. The examples discussed below will illustrate how bioengineering is facing the individual time constants.

A typical bioprocess, if operated in batch mode, extends over several hours or a few days. If operated in continuous mode, it is not reasonable to accept operating times of less than a month (see, e.g., Heijnen et al. [241]). If an organism has special physiological features such as baker's yeast, a transition from one to the other domain (e.g., from low to high dilution rates or, in other words, from purely oxidative to oxidoreductive growth) may also result in considerable changes in the time required to approach a new steady state. Axelsson et al. [242] reported a rough estimate for this case: the time constant for experiments at low dilution rates (D<DR, where DR=D at which the regulatory switch between oxidative and oxidoreductive metabolism occurs) is, as expected, in the order of the mean residence time (τ=D$-$1) however, above DR, the time constant was predicted to be at least one order of magnitude greater. In experiments specifically designed to verify this, either even greater time constants or no unique stable steady states at all were found [243].

If excess carbon and energy source is pulsed to a carbon- and energy-limited culture, the intracellular ATP concentration initially drops because of the phosphorylation sink (glucokinase and/or hexokinases). Only later, during catabolism, can the energy provided by this extra substrate be liberated in terms of new ATP. The duration of the ATP sink was predicted to be in the order of a few tens of seconds by Nielsen et al. [244]. Obviously, such rapid reactions are under kinetic control and the necessary enzyme activities are present in sufficient quantities. It is not clear, at present, whether the fluxes attain their organism typical maximal values immediately or some fine tuning of enzymatic control precedes this event. It is likely that in the latter case a regulation of (intracellular) enzyme activity, if at all necessary, takes place on the level of enzyme modification (e.g., phosphorylation) rather than on the level of de novo production (i.e., control on the transcriptional level) because the latter would require much more time [245].

The *dynamics of microbial cultures* have an important impact on the characteristics of measurement and process control. The "typical time constant" in a bioprocess is often erroneously anticipated to be equivalent to the entire duration of cultivation. However, the quantitative investigation of substrate uptake requires a time resolution of a few 100 ms, otherwise artifacts must result [246, 247]. This statement was evidenced only after a suitable technique had been established: glucose metabolism was stopped within 100 ms by spraying the cell suspension from the overpressurized bioreactor into 60 % methanol which was pre-chilled to $-40°$C.

The relevant *relaxation times* of a culture system are determined by the actual cell density and the specific conversion rate (capacity) of the culture, that is, by one or more operational and state variables (for instance feed rate, the concentrations or activities of cell mass and of effectors, if relevant) and inherent characteristic properties of the biosystem which are represented by parameters. There are metabolites with a long lifetime and other (key) metabolites with very short lifetimes, e.g., molecules representing the energy currency of cells such as ATP and other nucleotides. Rizzi et al. [248] and Theobald et al. [249] have shown that the energy charge response to a pulse challenge (ATP) of a yeast culture is a matter of a few seconds only. However, the responses can differ considerably when pulses or shifts are applied to cultures of different recent history [245, 250]. Neubauer et al. [251] found that substrate oscillations greatly affected the growth performance of *E. coli*. Short term, that is, less than 2 min, glucose excess, and starvation were investigated in a looped system of a stirred tank and a plug flow reactor. The metabolism changed within some few minutes totally from anaerobic to aerobic and vice versa. Similar observations have also been made for yeast [252].

Two other paradigms demonstrate that the band width of relaxation times is extremely broad: 1) the time required to achieve a new steady state in a chemostat culture is approximately determined by $(\mu_{\max} - \mathrm{D})^{-1}$. 2) The time required by a culture to consume a considerable fraction of small amounts of residual substrate during sampling – and thus systematically falsify the analytical results if no appropriate inactivation takes place depends, among others, on the cell density and can also be in the order of a few seconds only.

Even simple static models are very valuable for compensation of systematic errors built into automated analytical procedures. One important example is the case when sampling requires a well-known but non-negligible time. A bypass behaves as a plug flow type reactor fraction where flow dependent spatial gradients develop and where no inactivation can take place because the bulk of the bypassed aliquots are returned to the reactor. The cells continue to consume substrate while they are being transported from the exit of the reactor to the filter. The permeate recovered is representative for the filter site but not for the reactor. Knowing the transport time and some basic kinetic parameters one can easily compensate online for such errors provided that a useful estimate of the actual biomass concentration is available. Even though a bypass can be tuned to operate at a mean residence time of 5 s or less, this can be enough for a significant decrease in substrate concentration in high-density cultures. Sample buses need a minimal (dead) tune for transportation of the sample and in situ filters tend to fail in high density cultures because of rapid fouling. Hence, the problem is real and can not be ignored.

8.2. Biotechnological Measurement Systems

All measurements have the ultimate goal of creating representative information from the biosystem involved. Cellular or biochemical quantities are the primary variables in bioprocess engineering. They mainly determine the performance of the bioprocess and are therefore of special interest [253, 254]. Measurement and acquisition of information are building the fundament for all process observations as well as for the development of new bioprocesses. They represent the presupposition for process optimization and control. However, they must be available online to exploit them in a control strategy. Measurements are not only key issues in modern process development, they also open the door to a detailed process supervision and understanding, and avoid restricted views when analyzing processes just through a keyhole [255, 256].

In comparison to other disciplines such as physics, mechanical or electrical engineering, sensors useful for online monitoring of biotechnological processes are comparatively few; they are mostly available for physical and chemical quantities rather than for biological ones. The reasons are manifold, but generally biologically relevant information is much more difficult and complex to access and interpret than those for typical physical or chemical quantities [257]. Other important reason derives from restricting requirements for the sensor equipment, namely

- Sterilization procedures,
- Stability and reliability over extended periods,

– Application over an extended dynamic range,
– No interferences with the sterile barrier,
– Insensitivity to protein adsorption, surface growth,
– Resistance to degradation or enzymatic break down, and
– No disturbances from matrix compounds (interferences, matrix effect).

These aspects are in an industrial environment covered by the fact that the operation of the analyzer and its service must be as simple as possible. The aim of applying such a measurement system is to get more information but not to increase the chances of malfunctions of the whole process. Due to the complex nature of many process analyzers this requirement can only be met in some exceptional cases. Finally, material problems can arise from the constraints dictated by sterile process conditions or by biocompatibility, which often make the construction of the sensor hardware rather difficult.

8.2.1. Process Requirements Concerning Measuring Quantities

In modern bioprocess engineering there are undoubtedly only a few variables that are generally regarded as essential. Among these are several physical (i.e., temperature), less chemical (pH-value) and even less biological quantities (cell concentration). Figure 29 gives a summary of what is nowadays believed to be a minimum set of required measurements in a bioprocess. Such equipment is typical for standard production of biomaterial [258]. However, the conclusion that these variables are sufficient to characterize the microenvironment and activity of the biocomponents, of course, is more than questionable. For the consideration and compliance of the process targets, as described above, besides some environmental and operational quantities, in particular the important biochemical state variables of the biosystems must be known, namely the amounts and composition of the active *biomass*, of the *starting material,* of the *products* (and byproducts) or other metabolites. Modern bioprocess observation systems try to overcome the lack in the instrumentations, by building up functional relationships between these quantities. The approach is to substitute the direct measurement by a function based value, calculated by other measured values. But this implies thorough problem awareness and is restricted to specific boundary conditions (see Section 8.2.2).

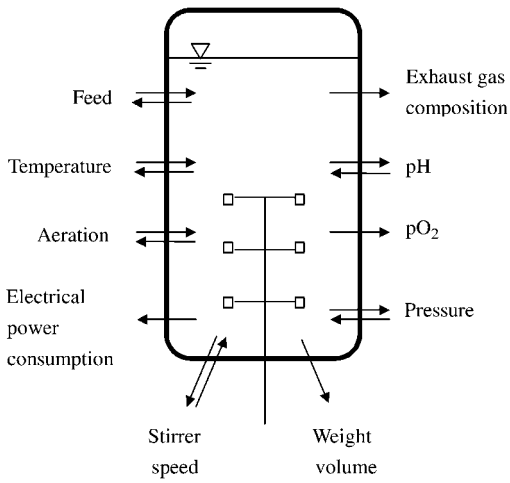

Figure 29. Common measurement instruments and control units of bioreactors as generally accepted as routine equipment (→ measurement only, ⇆ measurement and open- or closed-loop control)

Biomaterial. The active biomaterial is of paramount importance to scientists as well as to engineers. It is an indicator for the available quantity of biocatalysts. It is definitely an important key variable, because it determines – simplifying- the rates of growth and that for biotransformation. Almost all process descriptions and optimization tools contain the quantity of biomass as the most important state variable.

The biomaterial of interest depends on the system considered: microorganism, enzymes, nucleotides, or simple protein amount and constitution. Many control strategies involve the objective of optimizing the biomaterial concentration. For automation purposes, the biomass is mostly seen as a homogeneous entity, being aware that this represents an uncertainty because segregation into different individuals is more realistic. An ideal measurement of the bioactive compound would include their activity, physiological state, morphology, or other classifications rather than just their mass. But these quantities are normally ill-defined and, in general, inaccessible to online measuring devices.

A series of sensors and methods, that can be automated and have appeared in the recent decades for the measurement of the total biomass without discrimination of any segregation (e.g., active or inactive cells). Many of them rely on optical measuring principles [259, 260], others exploit filtration characteristics [261, 262], electrical properties of suspended cells [263, 264], or thermodynamic laws [265]. One has to decide according to the specific application, which method seems to be appropriate concerning the boundary conditions, because all methods reply on indirect measuring principles, and the access of the information carrying signal differs in non general ranges. To get online information about the actual state of the biomaterial techniques based on fluorescence principle, especially the 2-D fluorescence spectroscopy, should be mentioned. It represents a noninvasive method to analyze intracellular compounds. Technically, either intra- or extracellular fluorophores (e.g., NAD^+, Riboflavin, Tyrosine, ATP) are excited by visible or ultraviolet light, and the fluorescent light emitted by the fluorophores at a longer characteristic wavelength is collected and gives information about their concentration [266 – 268].

Substrate. The presence of sufficient and appropriate starting material (substrate) is the cause of any biotransformation and represents the supposition of the product formation. One can solve the inverse problem, namely conclude that biological activities cease whenever an essential substrate is exhausted, and thus omit the measurement of the substrate, provided the progress of the bioprocess and/or product formation is known [269]. But, this is not always a proper solution because there are many more plausible, and also probable, reasons for a decrease in bioactivities than just their limitation by depletion of a substrate. One must, then, solve the direct problem, namely analyze the relevant quantities of the biological unit. From an engineering point of view this measurement should be available instantaneously in order to be able to control the desired process characteristics. In environmental biotechnology, in particular, the objective of a bioprocess can be to remove a starting material as completely as possible rather than making a valuable product. So, the starting material is identical to the process product. In some other bio- or food technological processes, the conversion of a substrate from specific state of aggregation to another is of special interest. For example, in the case of fouling of heat exchangers where dissolved proteins are transformed to a solid protein layer, the starting material changes only its state not its chemical structure. Or, considering drying or thawing where water changes its state (or is removed) which strongly influence the quality and behavior of the biomaterial involved, but no (bio-) chemical conversion takes place. The classical methods to determine these quantities are offline laboratory methods so far. This implies that samples are taken, sometimes aseptically, pre-treated, and transported to a suitable location, where storage of these samples might be necessary before analyzing manually. These steps need a lot of human and instrumental resources.

Product. The product is almost the only reason why a process should run. The main concern is in maximizing the profit, which depends directly on the volumetric productivity and/or on the purity of the resulting product. It is therefore of essential interest to know the quantities, which require measurements. What was said above about substrate determination and the application of classical methods is valid here, too.

In summary, bioprocess science needs significantly more quantitative measurements. It is insufficient to know that something happens, we need to know almost instantaneously, online and automatically why, how, and to what extent a bioprocess behaves [270].

8.2.2. Online Sensing Devices

The development of increasingly sophisticated equipments for measuring of starting material and products represents one trend to satisfy the growing demand for high-quality measurements. In contrast to engineering disciplines such as mechanical or electrical engineering, in bioprocess engineering almost no measurement performed is directly, i.e., none is obtained by immediate comparison with a reference quantity. Most measurements are achieved by means of some specific physical or chemical property of the process value, often hidden in the very complex sample matrix com-

mon to bioprocesses. This very complex matrix, which consists of the numerous substances and parameters surrounding the analyte and their mutual influences, makes the task of measuring difficult and leads to the development of more and more complicated measurement techniques [271, 272]. The surrounding equipment, i.e., for the sample pretreatment, exceeds the pure analytical core [273 – 275]. However, the greater the measuring system complexity, the more difficult it is to ensure the accuracy and the reliability of the collected data. The need for accuracy and reliability increases even more if the data of the process analyzer should be used as input for a closed-loop controller. When unreliable measurement systems are used, the system operator can only hardly decide whether an occurring variation is due to an error of the analyzer or a result from the bioprocess itself.

The final goal of the measuring unit of a bioprocess is to build up simple online instruments in order to collect all necessary information from the biosystem, which is necessary to hold the defined process targets. Therefore, the process targets should represent the selection criterion for defining the measuring equipments. The information should be available online and no manual interaction should be necessary to obtain the desired measuring results. This aspect touches the term "system observability" (see Section 8.4.1), which is strictly defined for linear systems, but eludes a definition for nonlinear systems as they are common for typical biological systems [276, 277]. It is not obvious and definable whether all system information is available und known to reach the observability in respect to the process goals or which sensors are still necessary to reach the specified process goals. The consequence is that there are normally one or more state variables (or linear combinations of them) that are hidden from the view of the observer (i.e., measurement equipment).

Concerning the information treatment there can generally be made different classifications, which are important for industrial applications. From the communications engineering point of view the information can be distinguished between *online* and *offline*. If a continuous automatic correlation between the signal from the sensor and the product or process quantity is possible an online signal is available, otherwise it is offline. A measurement is called *inline*, when a direct interaction between a product property and sensor exist, otherwise *exline*; this is more a technological characterization. Depending on the site of installation, one discriminates further between in situ, which means built-in, and ex situ, synonym for a bypass configuration or for an exit line. In the latter case, the withdrawn volumes are lost for the process. A further classification can be made by the mode of operation of the sensing device. One can discriminate between continuous and discontinuous signal generation in a predefined time step. This is a very important fact for planning and developing an appropriate process controller considering the time scales of the bioprocess. In the latter casem, one has to guarantee that the process information is available in real time.

Only online sensors are of special interest concerning the process control of biotechnological systems. Indeed there exist a couple of different, very powerful systems for the offline analysis of nearly all relevant substances. For online sensors, the situation is quite different. In situ instruments exist only for a few quantities like the measurement of temperature, pH, pressure, oxygen, carbon dioxide, density, flow rates, electric power consumption, or redox potential. They use different measuring principles and are widely introduced in industrial process engineering.

However, for the most important biological quantities in biotechnological systems especially the measurement of active or inactive biomass or the specific determination of individual biochemical substances there exists a deep lack in online instruments. In the following, principal techniques should be presented, which possess a certain online capability. See also → Biochemical Separations; → Mass Spectrometry; → Surface and Thin-Film Analysis.

Chromatographic Methods. A review of chromatographic methods is beyond the scope of this contribution. Both (high-pressure) liquid chromatography and gas chromatography have been applied in numerous cases to offline analyses of biotechnological samples, but online applications, especially for industrial processes, have only recently been reported [278 – 282]. The scope of chromatographic methods is the separation of the individual constituents of mixtures as they pass through columns filled with a

suitable stationary phases. As disadvantages the consumption of almost expensive reagents must be mentioned. The samples injected are to be pretreated in such a way, that they possess nearly no impurities, like gas bubbles or solid compartments. Any existence of those interferences disturbs the analysis in a significant manner, and manual service procedures are mostly the consequence. Furthermore the measuring time last up to 30 or 40 min; a time period which is often too large for bioprocess supervision. But also economical aspects must be taken into consideration. Chromatographic systems possess a high cost price. Furthermore, high-specialized staff must service them frequently, an aspect, which represents the major drawback for an industrial application.

Mass spectrometry. A further measurement method, in principal suitable for online applications represents mass spectrometry (MS). Mass spectrometry has been mainly applied for the online detection and quantification of gases such as p_{O_2}, p_{CO_2}, p_{N_2}, p_{H_2}, p_{CH_4}, and even H_2S, or volatile substances such as alcohols, acetoin and butandiol [283]. Generally, the detection principle allows simultaneous monitoring and, consequently, control of metabolites. The principles, sampling systems, control of the measuring device and application of MS for bioprocesses have been summarized elsewhere [284]. Of course, mass spectrometry requires very expensive equipment and needs frequent and complex service procedures. But it should be taken into account that automatic multiplexing of different sample streams is possible and, in addition, a great variety of different substances can be determined simultaneously. Thus, one has to decide, whether the high economical and applicational expense is justifiable in comparison to the advantages due to the acquisition of information.

Biosensors. The most promising technique to overcome these problems, represented in the last decade, is the use of biosensors . Biosensors consist of a sensing biological module of either catalytic (e.g., enzymes or microorganism) or affinity reaction type (antibodies, cell receptors) in intimate contact with a physical transducer (electrochemical, optic, calorimetric or acoustic sensor). The latter finally converts the (bio-) chemical into an electrical signal (overviews in [285 – 287]). In spite the fact, that principally for nearly all biochemical substances an biological recognition unit exists, biosensor applications in an industrial surrounding often concentrate only on a small group of substrates, namely glucose, ethanol, or lactate [288 – 292]. Generally, biosensors are tricky to handle and must in principle be recalibrated in the matrix where the measurement should take place every time the process conditions change. Due to the sensible biological element – an enzyme or living microorganism – they principally cannot be sterilized. They also suffer from changes in the environment, for example, changes in pH or formation or existence of aggressive chemicals such as H_2O_2 [293]. Therefore, they must be used in a suitable environment. Furthermore, the surrounding must be constant, because in general an interrelation to the endogenous matrix exists. If the process media changes significantly, the biosensor alters its behavior. A compensation of the changed matrix interferences becomes necessary. This represents one of the most decisive drawbacks in inline biosensor applications for bioprocess engineering [294 – 296]. The long-term stability under working conditions is often poor; they must be serviced for several hours per week preferentially by high-specialized staff. In this way, only limited experience could have been gained under technical process monitoring conditions. They possess only in some exceptional cases a real industrial potential.

A reasonable way around the problem represents the measuring of the biotechnological media after removing a sample from the reactor in a bypass configuration. Consequently, the sensor is not installed in situ as it would be optimal but the information could be provided online and can be fed into an automatic process control system. Depending on the analyte of interest, i.e., whether it is soluble or (in) the dispersed phase, one needs to sample either the entire culture liquid or just a supernatant. The latter can be acquired using for example a membrane module. Concerning sampling, four aspects must be taken into account.

1) The system must be opened in such a way, that no infections can enter the reaction space either during sampling or between the sampling events.

2) A representative sample must be provided. So the overall volume of the sampling chamber as well as the position of device at the bioreactor must be determined properly. Additionally, the sample taken can still continue to react on the way to the sensor. In this case the sample would not be representative for the interior of the reactor, and appropriate additional measures or modeling approaches need to be taken in order to assure representativity.
3) When native samples are applied to the analytical system, problems can arise from *matrix interferences* and *matrix effects* [297]. Therefore, the sampling device should be chosen in that way that no interaction with the matrix occur which interferes the measurement (e.g., absorption of substrate at the membrane). On the other side, the sample device should withhold all the cross-sensitivities for disturbing co-components.
4) Due to the high number of individual system components and their complex interaction (i.e., sample pretreatment and analytical part), sampling devices provoke a small reliability in regard to the required process time. Furthermore, the hardware, necessary for the whole analysis system entails high original costs. The specific adaptation to one concrete process condition implies inflexibility. It excludes the determination of other metabolites in other process media. Its adaptation is only possible by time- and money-consuming alterations.

Flow Injection Analysis. The most important technique of sampling methods in biotechnology, behind some conventional devices like membrane modules, represents the flow injection analysis (FIA), firstly introduced by Ruzicka and Hansen [298]. It can be viewed as a general solution-handling technique or sampling device combined with a sensing unit. This combination causes a high flexibility with respect to the combined analytical procedure. It can be seen as a principle where a small injected volume (10-200 μL) is introduced into a continuous unsegmented stream of carrier. The sample disperses in a well-defined concentration gradient and is transported by the carrier stream to the reaction zone and a subsequent detector module. FIA cannot generate continuous signals. But, there are several important advantages like a high sampling frequency (up to $>100\ h^{-1}$), small sample and reagent consumption, high reproducibility and nearly total versatility of the sensing methods.

Software Sensors. The most promising method to get online access to important key components of bioprocess quantities are software sensors. In general, software sensors supply the estimation of the missing measurements by using an appropriate model that relates the corresponding variable with other physical or chemical measurements that are correlated to it in any way. The fact that the model is implemented by means of a software package (hence the expression soft(ware)-sensor) denotes, that software sensors provide a software backup for unavailable sensors, as an alternative to a hardware back-up using spare sensors [299 – 302].

Generally, software sensors are typical solutions of so-called inverse problems (see Section 8.1.2). In a complex biological system, in particular, the quantities which are normally easiest to measure are the variables, not the parameters. In the case of metabolism, the usual parameters of interest are the enzymatic rate and affinity constants which are difficult to measure accurately in vitro and virtually impossible in vivo [303 – 306]. Yet to describe, understand, and simulate the system of interest we need knowledge of the parameters. In other words, one must go backwards from variables such as fluxes and metabolite concentrations, which are relatively easy to measure, to the parameters.

In other words, when using software sensors there must always be a model available that reliably relates the measured variable with the target variable or parameter of interest. Normally, measured variables are easily measurable effects that are caused and influenced by the target. It is the special objective of the software sensor to reach a maximal degree of generalization. However, this is very difficult – even impossible – to achieve or furthermore to prove this claim. Thus, the basic question is: Is the available information representative enough to generate a model for the accurate estimation of the quantity of interest, or is the desired information inaccessible by the chosen sensor collection? Consequently, the development of software sensors is often restricted to a specific application; any transfer to

other measuring problems is almost combined with considerable alterations. This does not concern only the determination of the new parameter at an identical model structure. Instead a complete new system definition can be necessary, where new input variables must be added or dispensable one can be deleted.

Spectroscopic Methods. In this context the use of spectroscopic methods [307 – 310], after optical or sonic stimulation play a dominant rule. They must be seen as multisensor systems. Their common property is the detection of absorption degrees after stimulation at various frequencies. By this, a set of information is generated, which possesses more information than the measurement at one specific frequency, the measurement at one frequency can be seen as the output of one sensor. The multivariate evaluation in an appropriate model should subsequently reveal the accurate determination of the desired quantity. In such models, normally pure data are used, the integration of knowledge from well-known fundamental equations is only very hardly to realize. The most important advantage of spectroscopic methods is their non-invasive character. The sensing device is not in direct contact with the media or the process itself, so no interference is to be expected.

The situation is even more complicated since things develop in relation to time. The estimation of the physiological state of a culture involves more than one (measurable) variable at a time and the recent history of this set of individual signal trajectories involved. Consequently, physiological state estimation requires recognition of complex patterns. Various algorithms used for this purpose to build up the (data-)model have in common that it is not always the present values alone that are evaluated, there is always the recent history of signal trajectories involved. Concerning the models, some authors define software sensors only on the basis of neural networks but a broader point of view should be adopted since software-sensor models are also obtained by using regression or correlation techniques as well as fuzzy logic or first-principle models or combinations of all of them [306, 311 – 314].

In some cases, the data describing the actual state and their recent history are compared with so-called reference patterns: these are data from historical experiments or runs which an expert has associated with a typical physiological state. A physiological state is recognized either if the actual constellation matches any one of the reference sets best – in this case, there is always an identification made – or if the match exceeds a predefined degree of certainty, e.g., 60 % – then it can happen that no identification or association is possible when the pre-selected threshold value was not reached. The direct association with reference data needs normalization (amplitude scaling) and, probably, frequency analysis in order to eliminate dependencies on (time) shifts, biases or drifts.

In other cases, the data trajectories are translated into trend qualities via shape descriptors. These combinations of trends of the trajectories of various state variables or derived variables define a certain system state; the advantage of this definition is that the association is no longer dependent on time and on the actual numerical values of variables and rates [315].

8.2.3. Further Aspects Concerning Measuring Systems

Online measurements produced with in situ sensors are difficult to validate. The usual procedure for evaluating the quality of a measurement is restricted to calibration/checking prior to and after an experimental run. A few sensors such as the p_{CO_2}- or the Cranfield-glucose sensor [316] allow removal and, therefore, recalibration of the transducer during a run. External chromatographs and FIA systems can be regularly recalibrated but the sterile interface cannot. Other sensors such as a pH or a p_{O_2} probe can be mounted via a retractable housing, which allows either sterile exchange or withdrawal for external recalibration during a run. A further possibility to gain information about the reliability of a measurement is to mount a number of identical sensors in easy-to-compare positions and to check the individual signals for equality. This opportunity has been exploited in particular at applications where a high process safety must be guarantied. It is highly desirable to have alternative principles of measurement at hand, which operate simultaneously (heterogeneous redundancy) or to introduce a self-supervision op-

tion, like pattern recognition to generally identify malfunctioning sensors [276].

Elemental balancing permits the determination of (other) metabolic rates provided that the stoichiometry is known. Carbon balances are the most useful but the carbon lost via the exhaust gas (as CO_2) and culture liquid (as HCO_3^-) must be measured. HEINZLE et al. [317] determined that the state predictions based on experiments with a small mass spectrometer are not useful due to unacceptable error propagation; for instance, a 1 % relative offset calibration error could result in a prediction error for an intracellular storage material (PHB) of >50 %. Instead, using a highly accurate and precise instrument (absolute errors <0.02 % gas composition) together with automatic, repetitive recalibration resulted, however, in reasonable estimation of substrates, PHB, and biomass. Others have also experienced these findings, e.g., [318, 319].

The undelayed evaluation of the state of a culture by using software sensors and computers, based on the quantitative analytical information provided by hardware sensors and intelligent analytical subsystems, constitutes an excellent basis for targeted process control. Experts – either human or computer – have the data and the deterministic knowledge to trace observed behavior back to the physical, chemical, and physiological roots thereby gaining a qualitative improvement of bioprocess control, a quantum leap: process control can act on the causes of effects rather than just cure symptoms. A simple standard operating procedure [320] has proven useful in this concern, namely:

- Measure everything that can be measured at the very beginning of process development
- Decide whether or not a variable is relevant
- Determine the relevant variables to be measured, controlled, and/or documented
- Collect all raw data at any time and distinguish online between variable and parameter behavior
- Organize an archive of all these data accordingly and do not discard seemingly useless data since they contribute to the treasure of experience

If it is, as some people say, correct that today's bioengineering with all its tools and methodologies is too slow and not efficient enough, then it is all the more urgent to improve the performance of the methods, tools, and equipment currently available and to invent new and better ones. In essence, techniques of instrumentation, operation, and causal analytical interpretation of measurements need massive impulses.

8.3. Cognitive Computing

*Artificial neural network*s (ANNs) are dating back to 1943, firstly reported by MCCULLOCH and PITTS [321], then fallen into oblivion, but started a tremendous comeback in engineering science with the publications of RUMELHART in 1985, who introduced – again – the backpropagation algorithm and demonstrated its potential as a learning procedure in neural networks [322, 323]. *Fuzzy Logic* has been introduced by ZADEH 1965, firstly ignored consequently by technicians because of their undeterministic and uncertain nature, but provoked a second wave of interest in the 1980s predominantly in Japan in respect to their potential in treating highly complex, nonlinear multiple input multiple output (MIMO) systems in technical applications [324]. JOHN H. HOLLAND published 1975 his classic work "Adaption in Natural and Artificial Systems"; it can be seen as the pioneer work in *evolutionary programming* [325]. Compared to other mathematical principles, these three methods are fairly new approaches and did not gain not exclusive acceptance by scientists and engineers. The three methods mentioned above can be seen as the main representatives of a class of problem-solving methods taking nature as a model, especially how it has learned to solve problems. They can be summarized to the area of cognitive computing.

It is not within scope of this contribution to explain these methods in their basic principles and functionalities. Rather, their intrinsic principles are assumed to be well known; overviews and extensive explanations can be found for fuzzy in [326 – 328], for ANN in [328 – 330] (see also → Process Control Engineering, Chap. 13) and for evolutionary programming in [331, 332]. In the following, the focus lies on two methods, on fuzzy logic system and on *ANN*, because of their high applicational potential, especially in bioprocess engineering [333, 334]. The explanations should show their potential, typical aspects and also some critical

remarks, in particular in respect to bioprocess engineering applications.

Neural networks and fuzzy theory have been underground technologies for many years. They have had far more critics than supporters for most of their brief history. However, in particular due to their application potentials in different engineering areas where classical approaches failed, their acceptance is growing steadily. Neural networks and fuzzy systems estimate functions from sample data, they map inputs to outputs. Statistical and artificial intelligence approaches also estimate functions. On the other hand, for each problem statistical approaches require knowledge how outputs functionally depend on inputs, in other words they need a mathematical model. Thus, their fundamental ideas can be more compared to typical conventional mathematical principles. Instead, neural and fuzzy systems do not require such a mathematical model. They can be seen as model-free estimators [328].

From another point of view, artificial intelligence expert systems can also be seen as model-free estimators, they map conditions to actions. However, experts do not articulate a mathematical transfer function from the condition space to the action space and the AI framework is symbolic. Symbolic processing favors a prepositional and predicate calculus approach to machine intelligence. It does not favor numerical mathematical analysis or hardware implementation. In particular, symbols do not have derivatives; only sufficiently smooth functions have derivatives. Symbolic systems may change with time, but they are not properly dynamical systems, nor systems of first-order difference or differential equations. Therefore, they are unsuitable for modeling purposes in bioprocess engineering.

A small excurse to neural (pre-) attentive processing in human intelligence should illustrate which ideal result a cognitive system provides after its learning phase. Take a look at the *Kanizsa-square illusion* (Fig. 30).

What do we see when we look at the Kanizsa square [335]. We see a square with bright interior. We see illusory boundaries, or do we? We recognize a bright square. Indeed, technically we cannot see it, because it is not there. The only things which are shown are four three-quarter circles, nothing more. The square exists only in our brain, not "out there" in physical reality on this page. Immanuel Kant called these four ink stains 1783 noumena "things in themselves". Applied to technical problems, this means that we recognize information or generalize artifacts in that way, that we see not only the input–output mapping, but the real relations or laws the data stands for. The data can be seen as facilities or indications for the perception of the underlying intrinsic information.

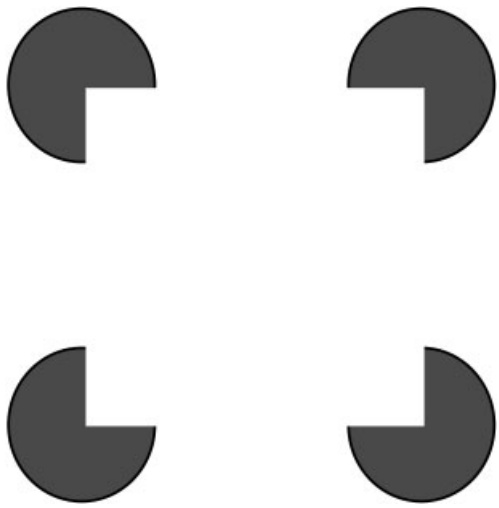

Figure 30. Kanizsa-square illusion

Today, many of the neural mechanisms that Kant could only guess are understood. We take for granted our high-speed, distributed, nonlinear, massively parallel pre-attentive processing. In our visual processing, we pay no attention to how we segment images, enhance contrasts, or discount background luminosity, even if they are the carrier of the real information. When we process sound, we pay no attention to how our cochleas filter out high-frequency signals even if we collect the data. We likewise ignore our real-time pre-attentive processing. We experience these pre-attentive phenomena, but we ignore them and cannot control or completely explain them. Natural selection has ensured only that we perform them, ceaselessly and fast. But what we really do, we recognize segmented image pieces and parsed speech units; we try to extract and introspect the intrinsic carried information based upon measured quantities.

Neural network studies both pre-attentive and attentive processing of stimuli. This leaves unaddressed the higher cognitive functions involved in reasoning, decision making, planning, or control. The nonlinear neurons and synapses in our brain perform these functions; they can deal with ill-posed problems or with a high degree of uncertainty. Natural selection evolved this capability in exemplary manner. Furthermore, it is the attempt of engineers to copy exactly this capability to technical problems by developing and using fuzzy logic systems or artificial neural networks.

Both methods try to imitate nature and both estimate input–output functions, in spite of totally different underlying principles. In contrast to statistical estimators, they estimate a function without a mathematical model of how outputs depend on inputs; this represents their main feature. They learn "from experience" with numerical, sometimes, linguistic sample data. Learning can be accomplished either by changing the system structure or by changing the system parameter; both principles are described for both algorithms [336 – 339]. Thus, the principle modus operandi is comparable. After the system problem is defined, the procedure can be divided into learning phase and prediction phase. The learning phase of ANN is quite obvious. Input data and corresponding output data are provided to the ANN; the system changes its structure or parameter, to minimize a predefined cost function. At first sight, the adaptive and learning character of fuzzy systems is not obvious. Indeed, there exist some techniques where the fuzzy system is built up autonomously based upon available data sets [339, 340]. On the other hand, in most heuristic applications a highly experienced technician does the "learning" step. He collects data, recognizes the intrinsic information, and formulates upon this information rules upon sets. Thus, the modeler performs the learning step and its result is introduced into the fuzzy system.

The most important features intelligent systems may show are *generalization* and *learning*. All intelligent systems should generalize. Their behavioral repertoires exceed their learned repertoire in that way, that intelligent systems associate similar responses with similar stimuli, small input changes produce small output changes. Furthermore, intelligent systems should learn or adapt. They learn new associations, new patterns, new functional dependencies. They sample flux of experiences and encode new information. They compress the sampled flux into a small but statistically representative set of prototypes or exemplars. Sample data, provided directly or transformed into linguistic rules, changes the system structure or parameters. In all cases, learning means changing. Neural networks learn patterns, functions, or probability distributions to recognize future patterns, filter future input streams of data, or solve future combinatorial optimization problems. Fuzzy systems can learn or formulate associative rules to estimate functions or control systems, either directly by means of data and an optimization add-on or guided by a highly experienced supervisor. It can be shown under some assumption that a neural net can approximate to any degree of accuracy using a fuzzy expert system and vice versa [341].

Neural and fuzzy systems ultimatively learn or estimate some unknown probability function $p(x)$. The probability density function $p(x)$ describes the distribution of vector patterns x, a few of which the neural or fuzzy system samples. When a neural or fuzzy system estimates a relation $r: X \rightarrow Y$, it in effect estimates the joint probability densities $p(x,y)$. Then the solution points $(x,r(y))$ should reside in high-probability regions of the input–output product space $X \times Y$. We do not need to learn if we know $p(x,y)$. We could proceed directly to our computational task with the techniques from numerical analysis, combinatorial optimization, calculus of variations, or other mathematical discipline. The need and the extent to learn varies inversely with the quantity of information available. *Supervised learning* uses or creates class–membership functions. It compares every input with the corresponding class D and proceeds learning based upon the difference. *Unsupervised learning* system processes each sample x but does not know that x belongs – or not – to class D. Neither supervised nor unsupervised learning systems assume knowledge of the underlying probability density function $p(x)$.

8.3.1. Fuzzy Logic Systems

Is uncertainty the same as randomness? Bayesian statistician Dennis Lindley [342] stated

that probability is the only sensible description of uncertainty and is adequate for all problems involving uncertainty. Lindley directs his challenge in large part at fuzzy theory, the theory that all things admit degrees, but admit them deterministically. Although, both expressions are used simultaneously in the same context, probability and fuzziness differ conceptually and theoretically. In this contribution some important differences are illustrated, more theoretical aspects and proves can be found elsewhere [343]. Fuzzy sets can be more compared with possibilities than probability. However, probability and fuzziness also share many similarities. Both systems describe uncertainty with numbers in the unit interval [0,1], consequently describe it numerically. Both systems combine sets and propositions associatively, commutatively, and distributively. The key distinction concerns how the systems jointly treat a set A and its opposite A^c. Classical set theory demands $A \cap A^c$, and probability theory conforms: $P(A \cap A)=0$. So $A \cap A^c$ represents a probabilistically impossible event. But fuzziness begins when $A \cap A^c=0$, hence $P(A \cap A) \neq 0$.

Fuzziness measures the degree to which an event occurs not whether it occurs. Probability describes the uncertainty of event occurrence. Whether an event occurs is "random"; to what degree it occurs is fuzzy. Consider an illustrating example. The probability this book will be published is one thing. The degree to which it will be published is another. From this book, only some special chapters may be edited or the whole entity gets public; this is typical for fuzziness. Fuzziness is a type of deterministic uncertainty; it handles with ambiguities. Unlike fuzziness, probability dissipates with increasing information. In fuzziness the uncertainty arises always form the simultaneous occurrence of two properties. More formally, does $m_A(x)$, the degree to which element x belongs to fuzzy set A, equal the probability that x belongs to A? This statement can be true or not, it depends on the problem existing, but in fact the point of views are quite different and so far hardly to compare. The following explanation should not explain the fuzzy steps in great detail, The focus lies upon two aspects. How are the sets defined and which rules can be determined. This is judged as the crucial aspects for fuzzy application in bioprocess engineering [344].

Following the introduction of fuzzy sets, Zadeh began steadily to develop the necessary inferencing mechanisms and modeling techniques to bring this concept to fruition [345]. In this period, it was recognized that fuzzy logic could aid in the development of control systems in which nonlinearities and time variance make the development of traditional control systems very difficult. In these cases, a human operator is often capable of controlling the plant, so a fuzzy controller can often be designed based on the operator's expert knowledge. Since that time, control applications based upon fuzzy logic models have been extensively investigated [346]. Fuzzy logic systems are similar to expert systems in their use of linguistic relationship, but the substantial difference is that the outputs are in general continuous by variation of the inputs [347]. It is exactly this property which causes the high interest in mathematical modeling, and hence in engineering science.

Mapping with fuzzy logic is analogous to classical modeling. The maps have variables which influence system behavior and relationships among the variables which describe the system. In classical models, variables have real number values, the relationships are defined in terms of mathematical functions. In fuzzy logic however, the values of variables are expressed by linguistic terms such as "large, medium, and small", the relationships are defined in terms of "if-then" rules. By using defuzzification techniques, exact numerical outputs are calculated from fuzzy subsets.

A fuzzy subset A is defined by a membership function $m_A(x_i)$, where x_i is the domain, of the variable on which A is defined.

$$m_A(x_i) \in [0,1] \, \forall \, x_i \in A \tag{10}$$

The value of $m_A(x_i)$ for each x determines the degree to which each element in the domain belongs to A. Although both classical and fuzzy subsets are defined by membership functions, the degree to which an element belongs to a classical subset is limited to being either zero or one. This means that $m(x)$ may only be a step function. In fuzzy logic, the degree to which an element belongs to a subset may be any value in the interval [0, 1]. Since $m_A(x_i)$ may be defined by any function, it is easy to see that a classical subset is a special case of a fuzzy subset [348]. The fuzzification process transforms exact values x

of fuzzy variable v_j to $m_{Ai}(x_i)$ for all subsets A_i, which are defined for the variable v_j. More rigorous definitions of membership functions may be found elsewhere [327, 349].

One technique for determining the shape of membership functions is seeking knowledge of the system from experts then constructing membership functions that represent expert opinion. Observations of many experts should be preferred. Based upon this observations, membership functions can be constructed using statistical methods accompanied by using the analysis of preferences [350, 351]. Many researchers have investigated more "rational" techniques for determining membership functions. One of these approaches represents one variant of *fuzzy clustering* [352, 353]. It requires a set of input and/or output data from the system to be collected. Patterns of data are found within the input or output space such that the variance within the sets is minimized and the variance between sets is maximized. The prototype, or center point, of a cluster X is determined and a distance metric is used to determine the degree of membership that a value has to X. Another technique for determining membership functions involves *neural networks*. Neural networks are networks of simple processors connected by adjustable weighting factors [328]. The weights are adjusted by using a backpropagation or other known algorithm to minimize the difference between the predicted output and the measured output. The parameters of membership functions can be "learned" by a neural net from a set of input–output data [354, 355]. Genetic algorithms have also been used to determine the optimal shape for membership functions [356, 357].

For fuzzy subsets to be useful in "modeling", there must be a way to define relationships between them. This is accomplished by means of their membership functions $m_{Ai}(x)$ with logical operators and inferences, to extract ambiguities. There are three basic operators in fuzzy logic, (1) OR, (2) AND, and (3) NOT, which link two or more membership functions $m_{Ai}(x)$ of different variables. The activation level of each rule is normally equal to that of the premise (activation). All rules are evaluated. If some sets of an output variable are activated more than once, an aggregation method must be chosen to determine the final activity. After defining the interferencing method, the output fuzzy sets are defuzzified to get a precise value, for further explanation see [327, 349].

Expert knowledge is the most common technique for determining rules. The expert is asked to summarize the knowledge about the system in the form of cause and effect relationships. From these the rules are formulated. When no experts are available or a more analytical approach is desired, other techniques of system identification must be used. One approach is to apply a set of *metarules* to adaptively acquire information about the system. Linguistic self-organizing controllers issue control actions and observe the environment [358]. The metarules evaluate the performance of the control actions from the effect of the control actions on the environment. The metarules then determine whether the control actions should be modified and perform the modification, if necessary. This technique has been applied especially to robotics applications [359]. Fuzzy classifier systems are another way to discover rules. Fuzzy classifier systems are generalizations of genetic classifier systems. A genetic classifier system uses some of the ideas from genetic algorithms to develop expert systems [360, 361]. Standard genetic algorithms have also been used to discover rules [362]. Sugeno and Yasukawa use the second variant of fuzzy clustering [fuzzy clustering methods can also be used to determine the membership functions (see above). Both approaches are addressed by the term "fuzzy clustering"] on data in the output space to determine fuzzy rules [363]. The clusters in the output space are used to induce clusters in the input space. Then, input space clusters are projected onto the axes to find the fuzzy subsets for the input variables. Since the input clusters are induced from the output clusters, rules can be constructed between the input subsets and the output subsets. Yoshinori et al. (1996) discuss another method of fuzzy model construction based upon clustering techniques [364]. This method produces a relational as opposed to a functional model, and consists of a series of local, linear models to approximate a global nonlinear one. Recently; neural networks have become an active field of interest and have been used to learn rules as well as membership functions [365, 366]. However, all methods, which automatically generate rules, have deficiencies in their interpreting ability. They are generated to fit available data best and it is not

obvious that the rules express reasoning relations. This can be seen as a strong disadvantage compared to the generation of rules by means of expert knowledge.

The application area of fuzzy systems is widespread and should only be presented here for applications in bioprocess engineering. The fermentation industry was one of the first to recognize the potential of fuzzy logic for biological processes [367, 368]. Konstantinov and Yoshida developed a fuzzy logic system for identifying physiological states in fermentation processes and used it to control distributed an *E. coli* fermentation [369]. Meanwhile the range of applications has enlarged tremendously. In principle, they can be divided in two fields. The first one are those where the knowledge bases is constructed such that it is based upon existing (qualitative) information, mainly from high-specialist staff. This represents the area of fuzzy modeling [370, 371], fuzzy controller [372, 373], or estimator for biosystem states or process faults [374, 375]. One of the intrinsic characteristics is that significant knowledge engineering effort is required for setting up a complete and consistent rule base. The other domain of fuzzy logic is relational model-based systems where the linguistic information is extracted from available data to build a system model. This second approach is also directed to the field of bioinformatics or experimental design [376, 377]. This approach does not need any preinformation about the intrinsic relation, however a knowledge base can arise which is hardly to evaluate or interpret. Additionally, no information can be given about the generality of the established rules. At all accounts, when a data driven construction is used, a subsequent mathematical or manual evaluation, especially of the rule bases, is necessary after the automatic construction of the fuzzy system. For the process industry, expert-knowledge-based approaches possess more interests because they follow a concrete process target. The second one can be seen as a special form of knowledge discovery which enables to accomplish some linguistic evaluation based upon data, if no expert knowledge exists. In some exceptional cases, these clustering techniques are exploited for control purposes [378]. An aggregation of both principles is also fruitful in the case where the existing information is too little to build up a representative "model". In this case, a data driven approach can help to complete the fuzzy structure.

8.3.2. Artificial Neural Networks (ANN)

Artificial neural networks are designed to mimic the actions of neurons in the human brain. For detailed information, see [328 – 330]. They represent massive collections of interconnected neurons (nodes), which individually perform a relative simple signal processing; they are interconnected via synaptic links. Like brains, neural networks recognize patterns or relations between inputs and outputs we cannot even define. They can be seen as pure function approximators whose abilities depend on the behavior of the individual nodes, the structure of the network, the learning procedure used, and on a very strong degree of the quantity of representative data. Not the total amount of data is crucial but the information content with small degree of redundancy. Almost the whole quantity space of the input and output values should be spanned by the database. Recently, e.g., Cybenko [379] and Hornik et al. [380], have proved that any continuous function can be approximated to an arbitrary degree of exactness on a compact set by a feedforward neural network comprising two or more hidden layers and a continuous nonlinear activation function, providing appropriate data.

Working with neural networks always consists of two phases, a *training* and a *prediction* phase. In the first one, the variable parameters $\vec{p}\ (t)$ of the net are changed in that way, the ANN fits best to the trainings database. In most cases, $\vec{p}\ (t)$ represents weight parameters of the intrinsic functions, e.g., the weight how the output of a neuron contributes to the activation of a subsequent node. On the other hand, there exist also some nets where the structure of the net (nodes and their synapses) can be seen as their parameters which are changed during learning. The indicator for changing $\vec{p}\ (t)$ represents a formulated performance index *J*. Based upon the absolute value *J*, the parameters are changed in order to minimize *J*.

$$\min\left(J\left(\vec{p}\right)\right) = \| \frac{\Delta J}{\Delta p_1}, \frac{\Delta J}{\Delta p_2}, \cdots, \frac{\Delta J}{\Delta p_n} \| \tag{11}$$

where p_i is a parameter *i* with $i = 1 \ldots n$ parameters in the ANN.

The definition of J depends on a couple of different aspects, but in most cases the Euclidian distance between the desired target and the calculated output integrated over all training data sets is used. Performing learning should result in the global optima concerning the specified problem. However, most training procedures are local methods, requiring a gradient calculation. And obviously, any algorithm that relentlessly crawls downward must have a very lucky starting position if it is to settle into the lowest local minimum. Avoiding false minima requires two separate procedures. First, we must elute in the initializing phase starting in their vicinity. If we settle into the neighborhood of a broad minimum, it will be very hard to escape later. Second, we require a procedure for determining whether or not we are in a local minimum, and escaping if we are. This procedure will be constructive in that we were in a local minimum. Otherwise, we assume that we have found the global optimum. In the second phase, the prediction phase, the net is conditioned to treat new input patterns, hopefully gaining the correct corresponding output values.

The nodes are connected in that way that the output of one node represents one input of the adjacent node. A node is stimulated by one or more inputs $\vec{x}\,(t)$ and it generates one output, a scalar $y(t)$ that is sent to other neurons. The output $y(t)$ is dependent on the weighted activities of each input, on the nature all inputs are transformed within the node and the parameters constellation $\vec{p}\,(t)$ in each neuron. The actual relationship between the inputs and outputs can be enormously complex, depending on the chosen of entry $E(\cdot)$, aggregation $S(\cdot)$, activation $A(\cdot)$ and output function $O(\cdot)$:

$$y\,(t) = O\,(A(\,S\,(E(\,\vec{x}\,(t)\,,\,\vec{p}\,)))) \tag{12}$$

Thus, the resulting behavior of individual neurons can be modeled by a simple weighted sum of inputs, a complex collection of interrelated or subsequent set of (differential) equations, or anything in between. There can be significant time delays in the steady-state output value after stepwise input stimuli. Neurons do not always respond in the same way to the same inputs. Even random events can be considered in the operation of neurons. Luckily, a large body of research indicates that simple models, which account for only the most basic neural processes, can provide excellent solutions to practical problems. The human brain contains roughly 10^{11} neurons. As many as 10^4 synaptic junctions may abut onto single neuron. That gives roughly 10^{15} synapses in the human brain [381]. Consequently, the brain represents an asynchronous, nonlinear, massively parallel, feedback dynamical system of nearly "cosmological" proportions.

The ability of ANN to emulate complex dynamical systems stems from the *interconnectivity of neurons*. In general an ANN has n neurons distributed in an input, an hidden, and an output section. Connections between neurons can be organized in layers such that information flows in one direction only, or it can circulate throughout the network in cyclic patterns. All neurons can be updated simultaneously or time delays can be introduced. All responses can be strictly deterministic, or random behavior can be allowed; the variations are nearly endless and should not be described in detail. Mathematically they can be formulated in the following vector equation.

$$\vec{y}\,(t) = \vec{h}\,(\vec{o}\,(\,\vec{z}\,(t)\,,\,\vec{x}\,(t)\,,\,\vec{p}\,)) \tag{13}$$

where

$\vec{y}\,(t)$	Vector of y_i with $i = 1 \ldots n_i$ outputs of the ANN
$\vec{x}\,(t)$	Vector of x_j with $j = 0 \ldots n_j$ inputs of the ANN
$\vec{p}$	Vector of pm with $m = 0 \ldots n_m$ variable trainable parameters of the ANN
$\vec{z}\,(t)$	Vector of z_j with $n = 0 \ldots n_n$ for all outputs of all individual nodes
$\vec{o}\,(t)$	Vector function of function o_m with $m = 0 \ldots n_m$ for the calculation of the individual contribution to the output y_i
$\vec{h}\,(t)$	Vector function of function h_i with $i = 0 \ldots n_i$ for the calculations of the real outputs of the ANN

There also exist some criteria to decide whether a structure is appropriate for a good modeling approach. Always, a compromise must be made between the desire to have a simple model with fewer parameters and more accurate predictions at the cost of a large number of parameters. The *Akaike's information criterion* (AIC) uses the number of data points, the overall error, and the number of parameters to determine an index, which can be analyzed to fix the optimal quantity of free parameters [382]. Another concept that may be of interest is the

use of a periodiograms, which gives an indication regarding the frequency content of the input signals, typical frequency can be characterized, which fixes the optimum degree of freedom [383]. Another way is the determination of principal components before fixing the structure of the net, thus, incorrect or redundant information can be rejected before entering in the learning phase [384]. However, care must be taken in reducing system complexity or rejecting data sets, because they can be meaningful and represent the behavior of the biological system even though, this may not be obvious from the data set.

When considering a particular application, it is to decide what type of network should be used. Principally, ANN structures are classified as either global or local. Both possess specific properties and hence applicational preferences. Global networks are the most popular und important ones. Among them, different types exist: feedforward backpropagation neural networks (FBNN); recurrent neural networks (RNNs), or cascade correlation networks (CCNs). Unlike the FBNN, RNNs are more general in the sense that connections are allowed both ways between a pair of neurons and even from a neuron back to itself on any way. The RNN allows the dynamics of the network to be considered intrinsically. In contrast, in FBNN for mapping a dynamic behavior the input variables must be supplemented by their corresponding deviations or time series. However, the number of weights (for a fully recurrent network) to be determined may easily become quite large. Therefore structural optimizers are recommended, e.g., the cascade correlation, where connections and nodes are added as required, [385]. Local networks, such as the radial basis function network (RBFN) process data in discrete areas not continuously over the whole space of the data. They are particularly suitable for applications in online control and optimization [386]. In the RBFN, it is necessary to locate centers among the input/output pairs such that the sum of the squares of the distance from the center to the training data set is minimized. These centers are equivalent in concept to the weights in the FBNN. However, the RBFN may fail in predicting values if the prediction space does not contain any center. This must be seen as a crucial disadvantage, especially for online applications. So, global networks should be preferred if high uncertainty within the data exists.

The "explosion" of ANN applications in nearly all areas of process engineering can be attributed to the following reasons:

- The tremendous hardware advances in digital technology over the past decade have enabled simulations of neural nets to be made both economically and with relative ease hand speed.
- Applications of neural networks for sensor pattern classification have been found to be superior to the traditional algorithmic techniques or the expert system approaches.
- Neural networks offer the promise of being able to extract information from a plant in an efficient manner with normal availability of data, especially if severe or unknown nonlinearities and time invariances, typically for biosystems, exists.
- Some practitioners claim that neural networks may be easier to use and apply in the real process plant, with difficult to handle nonlinearities, as compared with the modeling approach, which can be subject to various modeling errors.
- Finally, the versatility in structure and application of neural networks enables them to be utilized in the middle ground between conventional model-based approaches and black-box approaches for solving many classes of problems. These hybrid-type approaches have been another factor, which has further attracted their use in chemical process systems recently.

The applications utilizing neural network based strategies in bioprocess engineering are wide ranging. A detailed description of all application is outside the scope of this contribution. Neural networks can be used for *classification*, to specify typical classes in the sets of the data base. Any task that can be done by traditional classification analysis can be accomplished at least as well and almost always much better by networks. Very important applications in bioprocess engineering are the multivariate exploitation of sensor arrays [387, 388] or the fault diagnosis [218, 389]. An ANN can be trained for *pattern recognition*. These patterns can be a intervals of time, cultivation states, fermentation series, images et cetera. If a version

of one of these patterns, corrupted by noise, is presented to a properly trained network, the network can provide the original pattern on which it was trained [390, 391]. A very common problem and the most popular application field of ANN is that of *modeling, identification*, or *estimating* the value or the state of a variable, given historic values of itself and other variables; e.g., they were used to estimate the state of microbiological cultures [392], the concentration of unmeasurable mostly intracellular biological key components [393 – 395], or for the identification of bioprocesses [396, 397]. ANNs can also replace or act as *process controllers*. The applications utilizing these neural-network-based strategies are wide ranging and vary from linear time invariant to highly nonlinear time variant systems [398, 399]. Typical advanced representatives are the model-predictive control, the inverse model-based and the adaptive control techniques. These methods have all in common, that the ANN reacts as a process controller, modified or expanded in some way by a reference model or by an online training unit.

8.4. Modeling Aspects of Biological Systems

By definition, biochemical engineers are concerned with biochemical systems, with reactions and artifacts of biochemical substances like proteins or with systems employing growing cells. Even the simplest living cell is a system of such a forbidding complexity that any forward mathematical description of it is – in most cases – an extremely modest approximation. This situation prompts for bioprocess modeling the question from a formal logical viewpoint: "What kind of relation ship should exist between the underlying physical system and its mathematical description and which approaches should be exploited?".

Contemplation of this question leads, in one direction, into the labyrinths of philosophy science. Guidance into this territory from the learning guide Rutherford Aris led to John Casti's two-volume treatise "Reality Rules", which explores the general definitions of a mathematical model as well as numerous specific examples of models in different contexts [400, 401]. Gratefully for engineers, who are (or should be) engaged in discovery and development of useful modeling technology, Casti asserts an inextricable coupling between a model and its intended application: "Basically, the point of making models is to be able to bring a measure of order or probable system performance to our experience and observations, as well as to make specific predictions about certain aspects of the world by experience". Therefore, mathematical modeling does not make sense without formulating before making the model what is its use and what problem it is intended to help to solve. Thus, a definition of a model can be given as follows: *A model is an image of a real system that shows analogues behavior in the important properties, and that allows within a limited region in respect to the intended purposes a description of the behavior of the original systems*.

The emphasis represents the word *important*. A model should be always as simple as possible; the degree of simplicity depends only on the model purpose and on the modeler skills. Modeling the same real system with a different focus, the significant properties are mostly quite different. Single parameters can represent in one case the most specifying quantities and in the other case they are completely unimportant. The focus and the intention of the model determine not only the parameter, but also the proceeding in the formulation of the relations. The modeler must identify the important variables (inputs, outputs, states) and their separate effects, which in practice may have a very highly interactive combined effect on the overall process. This is only possible if the modeler hat a profound technological knowledge. Once the model is established, it can be used, with reasonable confidence, to predict performances under differing process conditions. But this is only allowed under specific boundary conditions, under those which were used for the set-up. If you leave them, the describing ability gets lost. Any extrapolation should be seen as critical.

8.4.1. Steps in Creating a Model

Model building is always a combination of theoretical studies and practical experiments in a very iterative sequence (Table 15). The described sequence should be observed in order

to get proper problem awareness and to minimize the efforts to reach the main targets. The objective of a scientific investigation is to improve understanding of a system by testing a hypothesis about that system. The formulation of the hypothesis is the most important action in the modeling process. In a scientific process, a problem or question is posed and stated as a hypothesis or theoretical statement of the problem.

Table 15. Step sequences for model building in bioreaction engineering. The steps must be seen as a loop whose termination is determined by the defined cost function

Step	*Action*
1	Proper definition of the problem (hypothesis), the goals, and the objectives of the study, fixing of presuppositions, boundary conditions and constraints, defining the evaluation and error criteria
2	Analysis of the system, determination of the structural elements, description of the key elements (variables, processes)
3	Running typical representative experiments, exploiting the parameter space of the control variables (experimental design)
4	Establishing the type of model by use of balances, physical–chemical–biological laws, available data, empirical equations, uncertainty treatment, problem formulation in mathematical or linguistic terms
5	Simplifying assumption (e.g. about mixing, process structure and dynamics, metabolism, kinetics, neglecting aspects)
6	Choice and definition of the important process variables: input and output variables, states, and parameters for the model
7	Simulation of the model, parameter identification, sensitivity analysis, determination of estimation properties
8	Evaluation of the model by using test data considering the evaluation criteria from step 1, if necessary starting again with step 1

For complex biological systems, however, the hypothesis may need to be translated into a mathematical tractable building and the model predictions are compared to the observed data as a basis for rejecting or accepting a hypothesis. This step also includes the definitions of the goals and the objectives. The purpose of the model usually dictates the form and the detail as well as the data that are required to develop and test it. Typical goals can be the identification of typical parameters or simulations, to predict responses to a perturbation. If the purpose is to get a general idea on how the system would behave under a new condition, the accuracy may not be important, however, if the purpose is to determine an "optimal" process trajectory, the model accuracy may be critical. Finally, the purpose may be to derive the mechanistic principles underlying the system behavior. So, it is useful to obtain as much data from the system as possible. It can be necessary to examine all different portions of the curves individually. Even, small deviations can hold large insight into the intrinsic physico-chemical effects of the biosystem.

In all real cases of *system modeling*, also boundary conditions or system constraints must be fulfilled. This requirement addresses two different aspects. The first one considers the fact that the created model is only valid inside typical boundary limits of the system variables and inside formulated constraints in respect to realized assumptions. The second one indicates that building up a model in bioprocess engineering is always accompanied with an identification step for optimizing the parameters. But the parameters are reasonable only within some limits, though they fulfill the mathematical optimal criteria. Otherwise, they can loose any physicochemical meaning or interpretational ability. Having more than one parameter to determine can lead, depending on the complexity of the system under consideration, to ambiguities. In this case, although available knowledge, e.g., in form of experimental data is represented well, the meaning of the parameters is outside of any interpretation scope. The extrapolation ability gets lost.

To accomplish a mathematical treatment of the model in order to reach the predefined goals, an evaluation criteria must be formulated. This can be seen as the mathematical description to the formulation of the modeling purpose and is realized by formulating a scalar performance index. Its quantity gives information about the model quality. In bioprocess engineering, it is normally calculated from the results obtained by the model compared with those of the corresponding experimental data.

There are two common approaches to fit models to data, the *least-square* approach where parameter are adjusted to minimize the sum (over all target values and existing data sets) of squares of the residuals between the calculated and experimental data and the *maximum-likelihood* approach, which maximizes the probability of the assumed parameterized probability density function [234]. Also, the above mentioned boundary conditions and system constraints must be considered in the performance

index. This aspect is from the logical point of view strictly necessary, but invokes in some cases crucial mathematical problems [402].

The system under study may consist of a single biomolecule, a cell, or a whole plant. Biological systems are normally open systems. They have inputs and losses, although they may also be reduced to closed systems as in vitro studies. Once the system to be studied is defined, the inputs and outputs are identified. Additionally, its processes, subsystems, and key compounds characterize the system. Processes are movements or changes in the system (e.g., absorption, metabolism, transport, chemical reactions), subsystems are components of the whole system (e.g., cells by considering the fermenter as the system), and compounds/states are the quantities under consideration (e.g., metabolites).

Almost all modeling approaches in bioprocess engineering involve an identification step of unknown parameters in the model by experimental data. The aim of the experiment is to obtain data to confirm or reject a hypothesis. It is essential to decide from which environmental and operating conditions measurable quantities should be gained, and to which precision and in which measuring range. The physical, chemical, and biological ones should be wide ranging, and should also include the quality of the whole environment and pre-cultures or pre-states. All measurable bioprocess data must be treated initially as variables. Afterwards, it could be decided, whether they remain constant or not. Automated measurement and control of bioprocesses, presently an art but – hopefully – routine in the near future, generates a tremendous amount of data. This requires judgment of the importance of these data for documentation to reduce data effectively without loss of valuable information. Experts – either human or computer – have the data and the deterministic knowledge to trace observed behavior back to the physical, chemical, and physiological roots.

Once the biosystem is structured and experiments are done, the model should be fixed. There exist different approaches to find the best appropriate model type for the underlying problem (see below). In principle, the model should represent the theories or hypothesis about how the system works. Normally, there exist different structures and parameter constellations which will fit a particular set of data. The crucial task for the modeler is to decide which one should be accomplished from the viewpoint of deep understanding of the biosystem. It is important and in most cases very effective that new approaches are built up upon information of earlier studies gained from literature. Differences between the models should be evaluated to choose the model most acceptable with current and new biological knowledge. Since the problem of parameter identification and model verification increases rapidly with model complexity, one should begin with as far as possible simplified assumptions and withdraw them step by step, if the model quality is judged to be not sufficient. In this way, a most simple initial model grows step by step in complexity and accuracy, without becoming too complicated. Modeling includes not only the selection of correct model structure, but also the determination of the undefined system parameters.

To give a final statement about the quality of model, its robustness has to be evaluated. This concerns *sensitivity*, *identifiability*, and *stability* aspects. Sensitivity refers to the relative influence of individual parameters on the solution. Identifiability refers to the uniqueness of the model and of its parameters. They should be determined in that way that a data-fitting process should return the same estimates of parameters more or less independently of the starting points for estimation and noise in the data. Stability is the behavior of a system with respect to a perturbation. Since nearly all real biological processes are time-variant and highly nonlinear in nature, the powerful collection of methods concerning the aspects of controllability, stability, and operability of linear systems cannot be applied [403].

Sensitivity analysis is one of the most important tools to optimized models. It refers to calculations and analyses performed to describe the relative dependency of the model parameters. Its analysis can be used for model validation as well as for fixing variables to constant values in order to decrease the degree of freedom. The sensitivities are also very important for interpretation purposes, e.g., to detect more or less important partial processes or quantities of the biosystem. From a practical perspective, it represents calculations which reveal the impact of assumptions made in conjunction with model development. A distinction can be made between point sensi-

tivity, relative sensitivity, and overall sensitivity. Considering the model

$$\vec{y}\ (t) = \vec{y}\ (t,\ \vec{p})) \tag{14}$$

with t representing the process time and $\vec{p}$ the parameter to identify; $\vec{y}\ (t)$ are the measurable output variables. The relationship between the state variables and the point sensitivity can be show using the first two terms in the Taylor expansion.

$$\vec{y}\ (t) = \vec{y}\ (t, \vec{p}_0) + G^T \partial\ \vec{p}\ + K \tag{15}$$

where G is the partial derivative matrix of $\vec{y}\ (t)$ with respect to $\vec{p}\ (t)$. An element of G is defined as the point sensitivity of y_i with respect to p_j. An alternative scale-free sensitivity is the relative sensitivity S_{ij}

$$S_{ij} = \frac{p_i}{y_i} \frac{\partial y_i}{\partial p_i} \tag{16}$$

To represent the overall sensitivity of a state variable as opposed to point sensitivity to parameter p_i over some time interval, S_{ij}^G can be defined

$$S_{ij}^G = \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} \frac{\partial y_i}{\partial p_j} dt \tag{17}$$

As applied to modeling, *identifiability* is a mathematical concept, which attempts to steer model development largely on the basis of an analysis reflecting whether features of a model for a presumed system can be extracted from a proposed experiment. Selecting a standardized method for an experiment directed for elucidating a model and its features involves selecting time points for observations, observation sites, sites for tracer input, as well as forms of tracer application. In bioprocess engineering, it is seldom possible to sample all units, and the number of units to which tracer can be applied is limited. Each choice we make here stands for a chance of mitigating against identifying aspects of the model, even in an error-free situation. If the lack of identifiability makes it impossible to estimate unique parameters or test relevant hypothesis, we should like to know this before conducting the experiment or interpreting the modeling results. If the experiment is adequate to identify the model, we are still left with the estimation problem, that of resolving the parameters amidst noisy data and to reach the observability. Consider the following general representation of a time-variant nonlinear system:

$$\vec{z}\ (t) = T^* \left(\vec{x} \left(t \right), \vec{y}\ (t), \vec{p} \right) \quad \text{with} \quad \vec{z}^0 = \vec{z}\ (t = t_0) \tag{18}$$

A system is defined as observable at the time t_0, if, admitting an arbitrary initial state $\overleftarrow{z}^0$, a finite point of time $t_1 > t_0$ exist, so that assuming known input $\vec{x}\ (t)$ and output $\vec{y}\ (t)$ trajectories the state $\vec{z}\ (t)$ can uniquely be determined in a time interval $t_0 < t < t_1$. In contrast to linear systems, this problem involving time-variant nonlinear equations as they arise in bioprocess engineering results in nonlinear algebraic equations for the solution of which similar powerful methods do not exist. Consequently, several assumptions and individual specifications and transformations have to be made to make the problem mathematically tractable with reasonable effort [404].

8.4.2. Reasons for Making a Model

There is a zoo of mathematical models in the biochemical engineering and mathematical biology literature. Many of these appear to naive as well as to sophisticated reader to have little more purpose than calculating numbers and writing scientific papers which conform reasonably to experimental data. This is, in itself, not a distinguished endeavor; it is not particularly difficult, and it teaches little. One reason that mathematical biology receives so little respect from biological scientists – and is generally not recognized as a credible research tool in biological science and a biotechnological discovery – is that it represents the failure to communicate clearly and persuasively the reasons and the motivation for constructing a model. Sometimes the modeler himself has not clearly asked this question while doing the work, sometimes biochemical scientists make one experiment after the other, not reflecting which mathematical approach exists to enhance the evaluation or interpretation of the experiments. However, modeling as a typical domain of engineering gets increasing interest in the study of biological systems. The challenge of

modeling biological systems is not to determine an arbitrary function to fit the data but to use the modeling process to understand the system.

Modeling can be used to determine the *structure of a system* where structure refers to the relationships between various parts or processes of a system. A model can be used to determine the type of relationship as well as the sequence of events that occur between the various substances of interest. In some systems, this information may be known, whereas in others the structure can only be inferred from the data. Modeling improves understanding, and it is through understanding that progress is made. In formulating a mathematical model, the modeler is forced to consider the complex cause-and-effect sequences of the process in detail, together with all the complex inter-relationships that may be involved in the process. The comparison of a model prediction with actual behavior usually leads also to an increased understanding of the process, simply by having to consider the ways in which the model might be in error. The results of a simulation can also often suggest reasons as to why certain observed and apparently inexplicable phenomena occur in practice.

Models can be used to determine parameters of interest, such as pool sizes clearance rates or transport rates. If the purpose is to accurately determine a particular parameter, it is important that the model can be uniquely determined. However, the model must also be consistent with known biological information. A model may be well-determined but incorrect.

The purpose of the model may be to determine the *interaction of parts of a system*. Models can integrate information on a number of subsystems into an integrated form to represent the process under consideration in the whole system. Because of the interactions, dependencies, and feedback of the processes, large complex systems can only be understood by using modeling. An example could be to link metabolism in one tissue with metabolism in another, or to link metabolism of a nutrient in one form with the metabolite in a second form. Large systems and systems with dynamic properties can only be understood with the use of models. Because many processes in large systems occur simultaneously and dynamic systems are accompanied by complex interactions between processes, it is unlikely that responses to perturbations could be predicted based solely on intuition and experience without a model.

Models can be used to simulate *inputs into a system* at various sites and to simulate short-term as well as long-term responses. Experiments can be simulated rapidly using a model to predict likely scenarios before undertaking expensive experimental data collection. While models will never replace experiments, they can be used to avoid experiments where insufficient or inappropriate data will be collected for testing a certain hypothesis. Models may be used predictive for design and control. Once the model has been established, it should be capable of predicting performance under differing sets of process conditions. Thus, mathematical models can be used for the design of relatively sophisticated control and optimization algorithms, and the model, itself, can often form an integral part of the control algorithm. Optimization usually involves considering the influence of two or more variables, often one directly related to profits and one related to costs. For example, the objective might be to run a reactor to produce a product at a maximum rate, while leaving a minimum amount of unreacted substrate.

An important use for models is to identify *sites of change* in a system when studied under different conditions. In some cases, the altered condition may result in large changes in the kinetic curves and several pathways or it may result in a subtle change in the data caused by a large change in only one parameter. A model helps to identify which parameters change between the conditions and the degree of change. The conditions may be an untreated versus treated state, a healthy subject versus a diseased subject, or a normal versus high intake of a nutrient.

Models form valuable tools for teaching the process of *scientific inquiry* and for challenging students to think creatively and quantitatively about a system. Models can be used to demonstrate properties of a system, teach principles (such as feed-back loops, saturation kinetics etc.), test theories, and design studies. With the increased speed and convenience of computers, the availability of modeling software and the access to it, the researchers who are developing and publishing models, need to be cognizant of this. They need to make their models understandable

and accessible to the biological community at large as well as to students.

Models help in *experimental design*. It is important that experiments be designed in such a way that the model can be properly tested. Often, the model itself will suggest the need for data for certain parameters, which might otherwise be neglected, and hence the need for a particular type of experiment to provide the required data. Conversely, sensitivity tests on the model may indicate that certain parameters may have a negligible effect, and hence these effects can be neglected both from the model and from the experimental program.

8.4.3. Different Types and Basic Approaches for Building a Model

Incomplete knowledge of the dominant biological pathways as well as low availability of sensor information about the current physiological state is the characteristics of biochemical process modeling. From system analysis, as already described above, the definitions of the describing variables (e.g., inputs, outputs, or states) can be given.

In principle, there can be postulated that these variables are intrinsically interconnected by means of a set of – indeed real existing – differential equations, as it is already described in previous chapters. But even being aware that these relations may exist in principal, for mathematical treatment they can only be specified in their structure and intercorrelation as well as in their parameter constellation in some exceptional cases with a high degree of accuracy. Thus, when dealing with modeling of real biological systems, uncertainty in various aspects can never be avoided [405]. In bioprocess engineering, uncertainty is an inseparable companion of any modeling approach.

Uncertainties in engineering systems can be mainly attributed to *ambiguity* and *vagueness* in defining the architecture, parameters, and governing prediction models for the systems. The ambiguity component is generally due to noncognitive sources. These sources include 1) physical randomness; 2) statistical uncertainty due to the use of sampled information to estimate the characteristics of these parameters; 3) lack of knowledge; and 4) modeling uncertainty which is due to simplifying assumptions in analytical and prediction models, simplified methods, and idealized representations of real performances.

The *vagueness*-related uncertainty is due to cognitive sources that include 1) the definition of certain parameters, e.g., structural performance (failure or survival), quality, deterioration, skill, and experience of construction workers and engineers, environmental impact of projects, conditions of existing structures; 2) other human factors; and 3) defining the interrelationships among the parameters of the problems, especially for complex systems. Other sources of uncertainty can include conflict in information and human and organizational errors [406].

Once the systems are defined, the relations between the outputs and states based upon the inputs and states have to be formulated to perform simulations. Today we have many types of models, which can be categorized in various ways such as deterministic versus nondeterministic, logical (linguistic) versus mathematical-equations-based or data- versus knowledge-driven. *Physical models* can be the realization of the original system in a different (usually smaller) scale or with structural modifications. A second type of physical models is obtained by turning to a different physical system, e.g., from the original biosystem to a corresponding electrical circuit [407]. Another group represents *verbal models*. They give a linguistic representation of our knowledge about the system, usually formulated as rules, e.g., "IF this or that happens THEN the system reacts by...". They are widely applied in the area of artificial intelligence, namely experts system [408, 409]. In the form of fuzzy-based systems, they gain increasing industrial interest in biotechnological control theory, which is due to their effective treatment of qualitative knowledge of experienced technologists in process control strategies (see Section 8.3.1), which could only hardly realized in terms of mathematical equations [410 – 412]. *Mathematical models* form the third class. They can be classified further depending on the mathematical formalism or the methods for model building. Theoretical models as mechanistic models are based on physical and chemical laws and our knowledge about the inner structure and function of the system. In contrast, experimental or nonmechanistic mathematical models try to give – without

looking into the interior of the system – only a description of the observed reaction of the system in response to a certain forcing signal. Many "mechanistic" models in biotechnology are actually due to their oversimplification quite closer to black-box models than to real mechanistic models, although mostly mechanistic interpretations are given [413 – 415].

White-box model. As described above (Section 8.3.1), the relation can strictly be mechanistic, considering the four differential equations for mass and heat balance and corresponding transformations. In this so-called *white-box modeling* strategy, the model development is mainly driven by the knowledge of the relevant mechanisms, macroscopic balances and predefined parameters. Indeed, if all the biosystem underlying mechanism are known and can be described by means of the above-specified equations and the accompanied parameters are known, the resultant white-box model is generally applicable and show good extrapolation properties. However, it is not an easy task to reveal all the relevant mechanisms and quantify these mechanisms correctly when dealing with biological processes. The *metabolic modeling* approach is one such approach and has been the focus of recent attention [416, 417]. Additionally, several attempts have recently been made in utilizing metabolic information for online identification and control as a consequence of the white-box strategy [418, 419].

Black-box model. The black-box model is, on the other hand, mainly driven by measured data obtained from the process, either online or offline. In terms of mathematical expressions, probability density functions for the important parameters are specified by means of available data. Based upon these distributions, the time-dependent system behavior can be forecasted [420 – 422]. The earliest stochastic biosystem model is perhaps that proposed by Yule, to mimic the evolution of a new species within a genus [423]. However, they are mainly used for systems only, where isolated parameters are treated. It is unrealistic that for a real biological MIMO system (multiple input multiple output), all joint probability densities, which are necessary to describe the key behavior, can be approximated by performed series of experiments. Even for systems without the existence of any driving-force function the expense for their determination is disproportional to the possible result. If inputs exist, which direct the system in a defined direction, reaction patterns must be formulated additionally in terms of probabilistic equations.

Another type of black-box models represents those where a specific functional structure is assumed, which is combined with a subsequent parameter estimation to adapt to the specific underlying problem (e.g., polynomial regression, ANN, chemometric models). However, on applying this, it has to be considered that the simplest and usually fairly accurate method for predicting values of the next data point in a time series from a "well-behaved" biosystem is to assume that it is the same as the present data point, a principle known as the first-order trivial predictor. Similarly, in the second-order trivial predictor the value of data point n+2 is equal to that for point n plus twice the signed difference between the values of point n+1 and n. If any complex nonlinear predictive model like ANN can do no better than the first- and second-order trivial predictors – which are indeed very often concurrently – then their development is a waste of time. Nevertheless, especially ANNs have been the focus of much attention for model development and have already been applied to various biochemical processes (Section 8.3.2). Although the main advantage of the development of black-box models is that a reasonably accurate model can be obtained without detailed mechanistic system knowledge, it should be noted that its accuracy depends only on the quality of the available input and output data. As black-box models are not believed to have any extrapolation properties, one has to obtain a large body of data for process identification by employing the relevant input variables in a wide range of fluctuations [218].

Markov chains. Another very attractive way to use probabilistic approaches are the formulation of the problem by means of so-called Markov chains [424, 425]. A Markov process allows the modeling of uncertainties in real-world systems that evolve dynamically with time without using mechanistic approaches. The basic concepts of a Markov process are those of states of a system and of state transitions. In specific applications, the modeling "art" is to find

an adequate state description such that the associated stochastic process has the Markovian property that the knowledge of the present state $X(t)$ is sufficient to predict the future stochastic behavior of the system $X(t+1)$. The moving of $i=t$ to $j=t+1$ is defined by means of the transition probability P_{ij} [426]. For bioprocess engineering, controlled Markov chains or Markov decision processes (MDP) possess a high potential, for simple modeling as well as for process optimization [427, 428]. Applied in bioprocess engineering, they involve the selection of an action U_t from a set of containing a finite number of A actions when observing the current state of the system. The problem is to define the probability that the system goes to a state j from a state i under the action of a:

$$P\{X_{n+1}=j|X_n=i,U_t=a\}=P_{ij}(a) \tag{19}$$

The controlled Markov chain is completely defined by the optimality criterion for calculation of the transition probability matrix, the state vector X, the action set A, and the decision epoch. For bioprocess modeling, concentration values grouped in clusters are often used as state variables; considering more than one state variable for each one a separate model is constructed. One great disadvantage of this approach represents the huge effort to determine the transition probability matrix P_{ij}(a) for each control action, especially if more than one manipulated variable should be considered [428].

Grey-box model. A grey-box model may be defined as a suitable combination of a black-box principle and a white-box model. The expectation is to obtain good interpolation and extrapolation properties. Especially, hybrid approaches, where portions from each principle are pieced together based upon the amount and the structure of available knowledge, gain increasing importance. From different points of view, such hybrid systems can be build up, i.e., analog or discrete consideration, stochastic or deterministic combinations [429 – 431]. Only some important concepts for bioprocess engineering should be presented here. In some cases, statistical methods are included into a mathematical discrete formulation, this concerns mainly *Kalman filtering*. They exploit statistical information about the error distribution of the state and a measurement values, and their propagation with time to predict online a future state based upon new measurements and already existing quantities. Their potential in bioprocess modeling and control is obvious [432 – 434].

From the structural point of view, hybrids are described in parallel and serial configurations. In the former case, the black box is placed in parallel with a white-box model and its role is to describe and to weight the difference between a white-box model and the real process. This model has been applied to several processes, but its performance might be limited to some extent to the viewpoint of extrapolation properties [435]. In the serial structure, the black box is placed in series with a white-box model. Different applications are described. Combined with a fuzzy expert system, the hybrid was applied to model a real-time fed-batch culture for baker's yeast production, a production scale beer-brewery fermentation, and mammalian cell cultures [431]. To show the better extrapolation properties of grey-box-model exemplarily, the performance of a serial-type grey-box model containing a neural network and a more knowledge-driven white-box model were compared with that of a more data-driven black-box model in the enzymatic hydrolysis of penicillin G to 6-amino-penicillanic acid and phenyl acetic acid by using the enzyme penicillin acylase. The grey-box model was shown to have significantly more reliable extrapolation properties than the black-box model [436].

Also linguistic models such as fuzzy models can be seen as a special type of grey-box models. Although the intrinsic behavior cannot be formulated by means of mathematical equations, a priori knowledge is included into the model by the linkage of various "IF. . .THEN . . . " conclusions, which is built up by the knowledge of highly specialized staff. Their fine-tuning is often accomplished by using data from experiments [437].

Being aware that different approaches exist how biological systems can be modeled and treated computationally, an important aspect for a successful application is the proper choice which principle fits best to the specific problem and its boundary conditions. In most cases, more than one specific approach can be applied in principal. The most important aspect for choos-

Table 16. Different types of models based upon the system knowledge

High	System knowledge				Low
Model	deterministic	stochastic, statistical	expert systems	approximation	classification
Examples	unstructured, structured, CFD[a]	Markov chains, maximum likelihood, Kalman,	fuzzy models, set theory	neural networks, PLS[b], PCR[c]	pattern recognition

[a] CFD, computional fluid dynamics.
[b] PLS, partial least square.
[c] PCR, principal component regression.

ing the right model should be the system knowledge about the underlying problem (Table 16).

Not only the amount of system knowledge is decisive, but also its structure, the quality and quantity of accompanied data, and the purpose for which the model should be used is of special importance. As a basic principle for building up a model, it could be postulated "use as much a priori information as possible". In mechanical engineering and fluid dynamics, strictly deterministic approaches are typical, where the underlying differential equations are "exactly" known. Thus, there exists no reason to deal with uncertainties and the problem is forwarded. However, this situation generally does not exist in bioprocess engineering. A biological system consists of a very complex, strongly regulated reaction network, where only some simple mechanisms can be formulated in terms of mathematical discrete forwarded formulations. Thus, one has to deal in a very strong degree with uncertainties. If one is able to formulate a set of differential equations, which represents not an exact image of the underlying mechanism but is descriptive enough for the existing problem and for the modeling purpose, one should formulate such a model and append a subsequent parameter estimation step where statistical properties or available data are exploited to consider unknown or not structural included information. In the case where a formal-kinetic approach cannot be formed but qualitative knowledge is available or can be extracted from available data, expert systems should be preferred. Especially fuzzy systems as a special representation of an expert system show advantages in modeling approaches because of their smoothness concerning the system output. In the case where a priori knowledge of the system is not available but pure data from a well-defined experimental design do exist, approximation or classification approach like ANN or chemometric models should be used. In those cases, they are the only alternative to realize appropriate prediction ability. They assume a system structure, which possesses no causal relation with the underlying real mechanism, but is able to incorporate and describe the system dynamics. They possess normally good approximation abilities, but their extrapolation property reveals a low reliability especially if outputs should be forecasted where the corresponding inputs are outside of the trained data space.

9. Special Applications in Biotechnology

The following chapters highlight some special aspects in biotechnology. However, there exist even more interesting applications in biotechnology. The reader should regard the areas chosen by the authors as examples for the wide distribution of biotechnology in the daily living.

9.1. Mammalian Cell Culture Technology

9.1.1. Introduction

Mammalian cell culture technology has become a major field in modern biotechnology, especially for the area of human health, and fascinating developments achieved in the last decades are an impressive example for an interdisciplinary interplay between medicine, biology, and engineering [438, 439]. Among the classical products from cells, we find viral vaccines, monoclonal antibodies, interferon, as well as recombinant therapeutic proteins. Tissue engineering or gene therapy open new, challenging areas [440].

Under "mammalian cell culture" or "animal cell culture" is to be understood the cells of a mammalian, isolated from specific tissues (i.e., skin, liver, glands, etc.) and further cultivated and reproduced in an artificial medium (Fig. 31) [441, 442]. During cultivation of mammalian cells in vitro, outside of a living organism, some distinct difficulties arise from the extraction of the cells from a "safe" tissue. Slow growth rates with doubling times between 18 and 28 hours, low productivity, a high sensitivity against shear stress due to the lack of a cell wall [443] and high demands in respect to the growth medium challenge the techniques required for mammalian cells. Furthermore, many cell lines grow adherent, and a suitable surface for attachment must be provided for these cells to proliferate. As of the origin from multicellular organisms, mammalian cells still hold the genetic program of inducing their own cell death, a process called "apoptosis" or "programmed cell death" [444 – 446]. This can limit culture productivity in biotechnological processes. Another major problem is the finite lifespan of primary cells, which die after several doublings in vitro. This problem was solved by transforming primary cells into immortal "established" or "continuous" cell lines. The engineering targets related to mammalian cell culture technology can be identified as:

– Production of a valuable product (viral vaccine, protein) from mammalian cells, development of processes for the efficient production of the desired product from a laboratory to a industrial scale under the restrictions given by the cell properties (*products from cells*).
– Development of processes for the cultivation of organic tissues, which can be used as substitute to original tissues (tissue engineering, gene therapy, *cells as products*).

Much effort has been put into the development of mammalian cell culture technology, and great progress was achieved in the last decades [447, 448]. Mammalian cells may now be cultured to very large volumes (up to 20 000 L) to provide the necessary quantities of a desired protein. Media that used to contain up to 10 % serum were continuously improved, and the cultivation in defined serum-free and even chemically defined, protein-free media is now common for most relevant industrial cell lines. Bioreactors were developed that provide the required low-shear-stress environment by, e.g., introducing gentle agitation and aeration with slow moving stirrers in stirred tanks, designing special aerators for air-lift reactors, or by separating the cells from the stressful conditions such as hollow-fiber, fluidized-bed, and fixed-bed reactors. In industrial scale, adherent cell lines can be cultivated on microcarriers (e.g., for vaccine production) or are adapted to grow in suspension, e.g., cell lines derived from baby hamster kidney cells (BHK) or chinese hamster ovary cells (CHO).

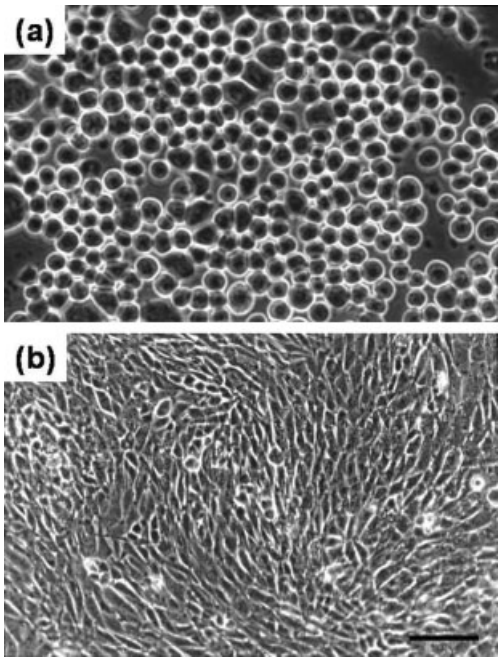

Figure 31. Morphology of (a) suspendable and (b) adherent mammalian cells (bar approx. 30 µm)

Genetic engineering contributes a great deal to recent progress [449 – 451]. With this technology, functional proteins can be produced by introducing recombinant DNA into cell lines, e.g., chimeric (humanized) antibodies are produced for in vivo applications in transfectoma or recombinant CHO cells. New promoters have been developed to enhance productivity, and product titers over one gram per liter have been reported for some industrial cell lines. Further, novel cell lines can now be constructed from

primary cells without loosing functionality by genetically induced proliferation [452, 453].

The following will provide a basic understanding of the specific requirements of mammalian cells, will describe state of the art process technology for cultivation of these cells and will give an outlook to future prospective. A comprehensive overview on cell culture technology is given by OZTURK and HU [454].

9.1.2. Products from Mammalian Cells

A detailed overview on products from mammalian cells is given in [454, 455]. Within "products from cells", viral vaccines [456] against polio, hepatitis B, measles, or mumps for human use, rubella, rabies, or food-and-mouth disease (FMD) for veterinary use play an important role. Viral vaccines are produced very efficiently by means of a cell-based vaccine technology. For this, mainly primary cells, diploid cells, or permanent cell lines (e.g., VERO) are applied, recently even recombinant cell lines. A breakthrough for the large-scale production of viral vaccines with anchorage dependent cells was the development of microcarriers in the late 1960s, allowing cultivation in stirred tanks in thousand-liter scale. New targets for cell-based vaccines are human immunodeficiency virus (HIV), herpes simplex virus, or influenza. Furthermore, new developments include genetically engineered or DNA vaccines (→ Immunotherapy and Vaccines).

Monoclonal antibodies [457] have become a valuable tool for diagnostic purposes as well as in therapy. Antibodies synthesized by B-lymphocytes play an important role in the immunosystem of mammalians. Traditionally, polyclonal antibodies were isolated from blood samples. In the 1970s MILSTEIN and KÖHLER developed a technology to generate hybridoma cells producing monoclonal antibodies [458]. As of the specific binding, monoclonal antibodies are widely used for diagnostics, where tens of thousands different monoclonal antibodies are available. The importance of monoclonal antibodies as therapeutic agents has evolved only recently, as immunogenic mouse antibodies were replaced by chimeric, humanized, or human antibodies. Fields of application are organ transplantation (OKT3), cancer diagnostic and treatment, rheumatoid arthritis, leukaemia, asthma, or multiple sclerosis. Nowadays several antibodies are produced in kilogram quantities [440]. Modern recombinant techniques focus on new antibody formats such as fragmented antibodies (FAbs) or bivalent antibodies with a broad spectrum of applications.

Glycoproteins are a further important group of products produced by mammalian cells. Starting with the production of α-interferon as an anti-infectious drug by (nonrecombinant) Namalwa cells in the late 1970s, nowadays a growing number of glycoproteins for treatment of a wide variety of diseases are produced by means of mostly recombinant mammalian cells. Prominent examples are cytokines (e.g., interferons and interleukins), hematopoetic growth factors (e.g., erythropoeitin for treatment of anemia), growth hormones, thrombolytic agents (e.g., tissue plasminogen activator [tPA]), coagulation factors (factor VII, factor VIII, factor IX etc.), and recombinant enzymes (DNAse) [440].

Recombinant proteins may be produced by either bacterial, yeast, or mammalian cells. From a technological point of view, hosts such as bacteria or yeast have an advantage with respect to growth rate, final cell density, and product concentration. Nevertheless, mammalian cells are preferred for those proteins requiring a specific, humanlike glycosylation pattern [459, 460], which is difficult to obtain in other host systems. Another problem for microorganisms is the maximal size of the produced protein which must be below a molecular mass of approx. 30.0 kDa. Further, in contrast to extracellular release of most proteins produced in mammalian cells, products from microorganisms are often accumulated intracellularly in "inclusion bodies". This demands a more complex downstreaming. Besides this, for mammalian cells important parameters such as product yield, medium requirements, and growth characteristics (suspendable, shear resistant) have been significantly improved.

Proteins produced in the milk of transgenic animals have started to compete against "classical" mammalian cell culture [461 – 464]. The proteins are usually expressed in enormous titers, over one order of magnitude higher than that obtained from cell cultures. The price for the production in transgenic animals will probably continue to decrease significantly due to

the development of more efficient reproduction technologies. These new approaches will significantly decrease the time for the development of a product. Among the disadvantages of transgenic animals, we find a more sophisticated downstreaming (high protein loads), the long development time, the inability to produce proteins that might impair the health of the animal (e.g., insulin), higher risk of viral contamination and the possibility of prion contamination (scrapie, BSE).

The novel field *cells as products* includes the development of artificial organs (tissue engineering of liver, kidney) and tissues (skin, cartilage, bone), or the expansion of hematopoietic cells for bone marrow transplantation or gene therapy. The exciting prospects of *tissue engineering* [465 – 467] are outlined in Section 9.2. *Somatic gene therapy* implies transfection of a specific gene to cells isolated previously from a patient suffering from a genetic disease [468 – 470]. The transfected cells are then reintroduced into the patient. Gene therapy involves transfer of genetic material, encoding therapeutic genes and the sequence necessary for expression to target cells to alter their genetic code for a desired therapeutic effect. Many diseases are originated by gene defects. Expression of the transferred genes can result in the synthesis of therapeutic proteins or correction of a gene defect. Gene transfer could also lead to desired apoptosis or inhibition of cell proliferation. Theoretically, gene therapy can be applied for repairing single entailed gene defects, acquired gene defects such as chronic infectious diseases, multifactor genetic diseases such as cardiovascular disease, and finally cancer. Since gene therapy was applied the first time in 1990, the number of clinical trials increased. Today gene therapy is widely used in ongoing clinical trials for the treatment of cancer and infectious diseases such as human immunodeficiency virus (HIV) infection.

Cell culture products are currently used mainly as medicines or in diagnostics. For medicines, some products have a demand of 500 kg/a and generate $ 1–2 billion dollars in revenue. Among these are tissue plasminogen activator (tPA), erythropeithin (EPO) products, Remicade® (Centocor) or MAbThera® [440]. The pipeline for new biopharmaceutical therapeutics targets a huge number of diseases (e.g., 425 for U.S.A in 2002 [471]), most of them for cancer therapy, infectious diseases, autoimmune diseases, or AIDS/HIV. It can be expected that at least some of these will find their way to the market. Furthermore, new products or medical protocols can be expected for tissue engineering and cell therapy. As for some compounds patents are already expired or will expire in the near future, there will be a market chance for biosimilar or generic products. On the other hand, the costs for target identification, clinical trials, and process development will increase (to date approx. $ 0.5–1 billion) and only block busters will make it finally to the market.

9.1.3. Cell Types

In general, mammalian cells relevant for industrial processes can be divided into the following groups [441, 472]:

- *Primary cells* have been isolated from a tissue and than taken into culture (primary culture).
- *Permanent or established cell lines* originate from a primary culture, but because of some transformation became able (at least theoretically speaking) to divide and proliferate indefinitely. Those cell lines are kept in a cell bank and are very often used as host cells for the expression of recombinant proteins.
- *Hybridoma cells* are cells which are obtained through the fusion of lymphocytes and tumor cells and which are able to express monoclonal antibodies.

The common procedure for the isolation of a mammalian cell from a tissue involves the following steps: first, a small piece of tissue is decomposed into single cells or cell clots mechanically, enzymatically (e.g., typsin [473], collagenase), or through a combined procedure. The cells are separated from the enzymes by centrifugation and resuspended in culture medium. After that, the cells are inoculated into a special glass or plastic vessel with flat bottom. The anchorage-dependent cells start to adhere to the bottom surface and, after a lag phase, start to divide by mitosis. Such a cell culture, in which the cells come from a differentiated tissue, is called a primary cell culture. The most critical point in the isolation of primary cells is to avoid external contamination by practicing aseptic techniques.

When the primary cells have covered the bottom of the culture vessel almost completely, they are enzymatically cleaved (i.e., trypsin) from their support and used to inoculate new cultures. This subculture gives origin to the so-called secondary cultures. Part of the cells is stored deeply frozen in liquid nitrogen, to remain as a safety stock, from which it is possible, at any time it might be necessary, to get enough "fresh" cells to start a new series of subcultures for mass production. With the primary cells, it is possible to repeat subcultures several times. However, the primary cultures have a finite lifespan and therefore, after a certain number of doublings (from 50 to 100 times), the cells cease to grow and die [441].

A finite growth capacity is a characteristic of all cells derived from normal mammalian tissue. A cell can be considered as normal when it shows a certain set of characteristics [474]:

- *A diploid number of chromosomes* (i.e., 46 chromosomes for human cells) with which it is shown that no gross chromosome damage has occurred.
- *Adherence*: the cells require a surface to grow attached to (anchorage dependent). The growth phase extends until the cells reach a stage of confluence (contact inhibition).
- *A finite life span* in culture.
- *Non malignant*: the cells are not cancerous, i.e., they do not cause tumor in mice.

In the 1960s, normal cells were required for the production of human vaccines in order to ensure the safety of these products [441].

Not all cell types produce exclusively primary cell cultures, which after a limited number of "passages" die. Some cells acquire an infinite lifespan and such a population is usually called a "permanent" or "established cell line" (Table 17). These cells have undergone a "transformation", i.e., they have lost their sensitivity to the growth control mechanisms. Transformed cells can also lose the characteristic to grow adhered to a surface and thus become able to grow in suspension. These transformations are also sometimes reflected in the chromosomes, changing the genotype of the cells. Transformed cells can be easily grown in relatively simple media without addition of expensive growth factors. At first, transformed cells were identified just by chance, but nowadays there are techniques to cause cell transformation and "immortalisation", as for example the treatment with mutagenic substances, with virus or with oncogenic substances [441]. Transformation of cells in vitro shows some similarities with carcinogenese in vivo, but is not identical to it; for instance, not all the transformed cells are malignant. On the other hand, all cells that are isolated from tumors (i.e, HeLa or Namalwa Cells) can be kept in a permanent culture.

Table 17. Examples of permanent cell lines important for research and production (data from [441], modified)

Cell Line	Origin	Application
Baby hamster kidney BKH-21	syrian hamster, 1963	adherent cells, but can be adapted to suspension, FMD vaccine, rabies vaccine, recombinant proteins (Factor VIII)
Chinese hamster ovary, CHO-K1	ovary of chinese hamster, 1957	adherent cells, but can be adapted to suspension, recombinant proteins (HBstg, tPA, Factor VIII)
COS	monkey kidney	transient protein expression [475]
NAMALWA	human lymphatic tissue	alpha-interferon
HeLa	tumor in a cervical vertebra	Fast-growing tumor cell line that was isolated in the beginning of the 1950s.
HEK-293	human embryonic kidney, 1977	transient protein expression
MDCK	rabbit kidney	adherent cell line with good growth characteristics, animal vaccines
MRC-5	human embryonic lung cells	"normal" cells with a finite life span, vaccine production
NS0 and SP2/0	mouse myeloma from B-lymphocytes	antibody production
3T3	mouse connective tissue	suspensible, used in the development of the cell culture technology
WI-38	human embryonic lung cells	"normal" cells with a finite life span vaccine production
Vero	long-tailed monkey kidney, 1962	established cell line, but with some characteristics of the normal diploid cells.

Hybridoma cells are "artificial" cells produced by the fusion of human or mammalian lymphocytes, which can secret specific antibodies, with a myeloma cell. Usually, lymphocytes do not survive outside the original organism, i.e.,

cannot be kept alive in an artificial medium. MILSTEIN and KÖHLER [458] were able to overcome this problem by fusing a mouse myelom cell similar to a tumor cell with a lymphocyte from a mouse spleen, which was able to produce antibodies. The cells generated in this way are hybrid cells (also called "hybridoma cells") and have the lymphocyte characteristic to produce antibodies against a certain antigen, as well as the ability from the myelom cells to survive in culture.

An alternative expression system for recombinant proteins offer insect cells, especially the insect-cell baculovirus expression system (ICBEVS) [476]. Mostly applied are Sf9- or Sf21-cells isolated from *Spodoptera frugiperda*. Production of heterologous proteins by the ICBEVS consists of two stages. Insect cells are first grown to a desired concentration and then infected with a recombinant baculovirus containing gene coding for the desired protein. Meanwhile, several recombinant proteins produced by this technique have reached the market.

The largest and most well known international animal cell culture collections are the American type culture collection (ATCC) [477], the European collection of animal cell culture (ECACC) [478] as well as the German resource centre for biological material (DSMZ) [479].

The application of permanent cell lines for the production of vaccines faced strong opposition, mainly in the 1960s, because there was the suspect that these cell lines, which behave similarly to tumor cells, could infect patients with unknown cancerous substances. Fortunately, no observation of this kind has been done until today. In the meanwhile, more and more permanent cell lines have been applied for the production of therapeutic and diagnosis substances. However, the application of permanent cell lines, and above all, of recombinant cell lines, is strictly controlled by supervisory boards (as the FDA [480] in the USA, and the EMEA [481] within the EU). During the whole production cycle of a pharmaceutics, not only the product has to be identical from charge to charge, but also the phenotypical characteristics of the cell line has to be maintained. To assure that, before the beginning of the production, a "master cell bank" has to be created, from which a "master working cell bank" is derived, providing the inoculation material for the production cultures.

9.1.4. Growth Medium for Cell Culture

Basically, a growth medium for mammalian cells have to supply all the necessary nutrients required for growth and product formation [482, 483]. Furthermore, it should have a certain buffer capacity to stabilize the pH (pH optimum 7.0–7.3) and should provide an appropriate osmolality in order to avoid damaging of the sensitive cell membrane [484 – 488]. A fundamental component of all media is a salt solution, which provides the ions necessary for life, keeps the osmotic pressure within the desired range, contains one or more buffer systems (sodium phosphate buffer, HEPES, and/or carbon dioxide–carbonic acid buffer) for pH regulation, and contains, in some cases, a pH indicator (phenol red). Furthermore the media contain glucose and glutamine as a source of carbon and nitrogen, other amino acids, vitamins, mineral salts, and trace elements.

A number of medium formulations have been developed, e.g., Eagle's minimal essential medium (MEM), Dulbecco's enriched (modified) Eagle's medium (DMEM), Ham's F12, and RPMI 1640, among others. Examples of medium compositions can be found in the literature [442, 454, 489]. Traditionally these basic media were supplied with approx. 5-10 % serum (e.g., foetal calf serum [FCS] or horse serum [HS]) in order to supply specific growth factors and to protect the cells against shear stress. The disadvantage of media containing serum are the indefinite composition of such media, the high serum costs, the difficulty of product purification, variations between the charges and the risk of virus contamination [490].

Serum can be partially substituted by the addition of transferin, insulin, ethanol amine, albumin, or eventually fibronectin as adherence factor in a serum-free but still protein-containing medium [491 – 493]. A further step to chemically defined, protein-free media became possible by replacing these animal-derived proteins by iron salts or iron complexes, IGF-1, chemically defined lipid concentrates, precursors or other stimulating agents such as fatty acids, biotin, cholin, glycerin, ethanolamin, thiole, hor-

mones, and vitamins [494 – 496]. Furthermore, addition of peptone or yeast extract can be helpful [497, 498]. However, the growth rate and productivity in a serum-free medium can decrease [499]. In addition, the sensitivity to shear stress increases, requiring additives with protecting characteristics, such as Pluronic F68 [500, 501]. Immobilization can improve the culture stability of nonanchorage-dependent animal cells grown in serum-free media [502, 503]

Different cell lines require different compositions. Adaptation of a cell line to grow without serum is quite time-consuming and not all cell lines have been adapted to serum-free or protein-free media [504, 505]. For cultivation of primary cells, for basic cell-culture research or for vaccine production still mostly complex, serum-containing media are common. In industrial production with established, optimized cell lines, serum-free, bovine-serum and protein-free as well as chemically defined media are state of the art [439].

9.1.5. Small-Scale Culture Systems for Routine Use

Techniques and culture systems for cultivation of mammalian cells differ significantly from those used for microbes. For basic research, a large number of culture systems have been developed (Fig. 32), mostly designed for use in an incubator aerated with 5 % CO_2 to maintain the pH within the desired range. Multiwell plates (6–96 wells), T-flasks (25–100 mL), Petri dishes or roller bottles (50 mL to 5 L) are usually used for cell maintenance and proliferation, especially for adherent cells [441, 506 – 509]. These systems allow for sterile handling procedures and are easy to use, disposable, and low-cost. On the other hand, they require individual handling, for example in medium exchange and cell seeding; their usefulness is limited when large quantities of cells or products are required. This can be overcome to some extent by using sophisticated robotics [510]. Furthermore, environmental parameters including pH, dissolved oxygen, and temperature can hardly be controlled within the medium. A further drawback

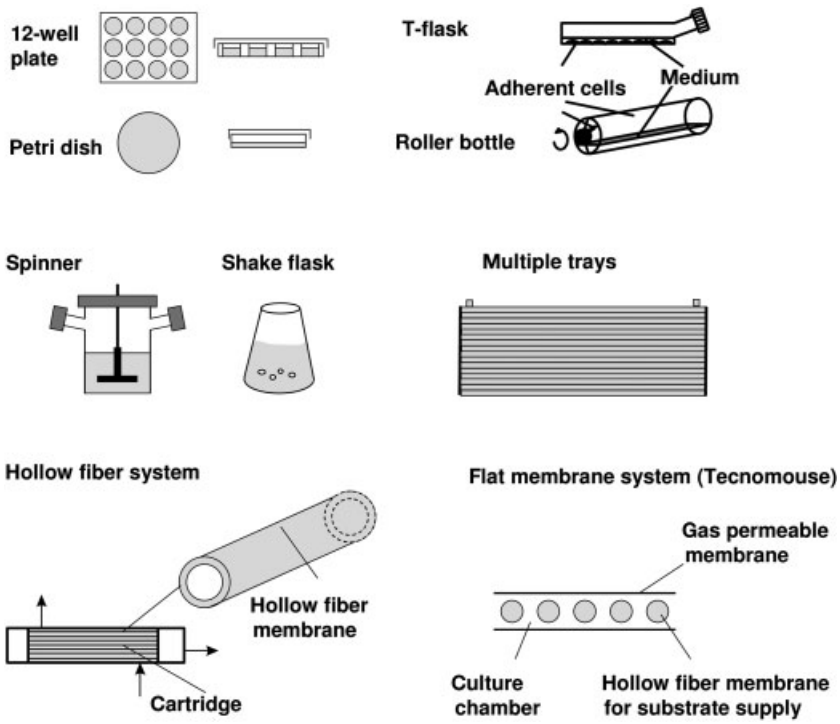

Figure 32. Routine cultivation systems for use in incubators

is the limited increase in cell number (approximately 10-20 times during cultivation). Alternatively, small well mixed bioreactors (for example, shake flasks, stirred vessels, and "super spinner" [511]) can be used, where adherent cells are grown on microcarriers.

Microcarriers are small beads, either solid or macroporous, having a diameter of approx. 100-300 μm and a density slightly higher than the growth medium. When first developed in the late 1960s [512 – 515], microcarrier culture introduced new possibilities and suspension culture of anchorage-dependent cells in high density. Nowadays, beads made of DEAE-Sephadex, DEAE-polyacrylamide, polyacrylamide, polystyrene, cellulose fibers, hollow glass, gelatin, or gelatin-coated dextran beads are in use [442]. In microcarrier culture, cells grow as monolayers on the surface of small spheres (Fig. 33) or as multilayers in the pores of macroporous structures that are usually suspended in culture medium by gentle stirring. By using microcarriers in simple suspension culture, fluidized or packed-bed systems, yields of up to 200×10^6 cells per milliliter are possible [516]. Microcarriers are extensively used for vaccine production on a larger scale (see below) [517 – 521].

Figure 33. Porcine chondrocytes grown on microcarriers (Cytodex 3, GE Healthcare)

Microencapsulation is another method for the immobilization of mammalian cells [441]. Basically, it can be used to protect cells against hazardous environmental conditions. The three basic encapsulation systems existing are the bead, the coated bead, and the membrane-coated hollow sphere [522 – 524]. Typical size for beads made of polysine alginate is 300 to 500 μm, and the molecular-mass cutoff of these capsule membranes is 60 to 70 kDa [442]. The capsules can be cultivated in suspension reactors similar to microcarriers. Detrimental with respect to scale-up is the fragile nature of the microcapsule.

In membrane bioreactors, including small hollow-fiber reactors [525], the miniPerm system [526] or the tecnomouse [527], cells are cultivated at tissuelike densities in a compartment which contains one or several types of membrane for nutrient and oxygen supply and removal of toxic metabolites. Hollow-fiber systems (compare Section 9.1.6) are widely used in the production of biopharmaceuticals including monoclonal antibodies. Several examples of modified membrane bioreactors exist for the three-dimensional culture of tissue cells [528 – 530] including hepatocytes [531 – 533], skin cells [534], or other human cells [535]. The specific characteristics of hollow-fiber reactors are discussed below.

9.1.6. Types of Bioreactors

Design and selection of cell culture bioreactors for larger scales have to meet special demands such as gentle agitation and aeration without cell damage, a well controlled environment with respect to pH, temp, dissolved oxygen, dissolved CO_2 concentration etc., low levels of toxic metabolites (ammonia, lactate), high cell and product concentrations, optimized medium utilization, surface for adherent cells, and scalability [536 – 545]. Bioreactors for mammalian cell culture can be divided into two categories (Fig. 34): *Low-density* or homogeneous systems (stirred-tank, bubble-column, air-lift reactor), where cells are cultivated either in suspension or, if required, on microcarriers; and *high-density* or heterogeneous systems (fixed-bed, fluidized-bed, hollow-fibre reactor), where cells are immobilized either in macroporous sup-

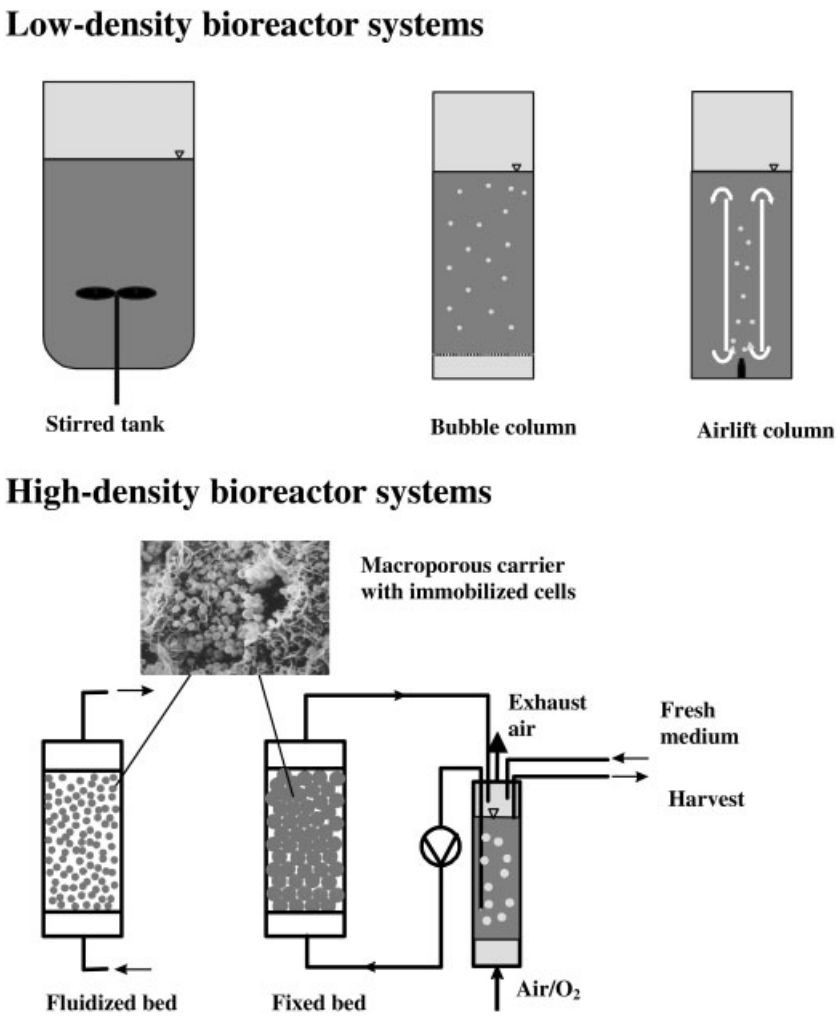

Figure 34. Bioreactor systems for cultivation of mammalian cells

port materials or within a compartment created by membranes. Selection of a reactor systems depends to a large extent on the specific purpose such as production of a certain amount of protein (small amounts for basic scientific studies up to kilogram quantities for medical application), three-dimensional tissue cultivation, or single-purpose or multipurpose facility.

Design of suspension reactors has to deal mainly with cell damage due to shear stress caused by agitation and/or aeration [546 – 548]. Stirred-tank reactors (Fig. 35), which are the most common type of reactor and are nowadays build up to 20 000 L scale, are especially suited for suspendable cells [549]. Adherent cells can be grown on microcarriers and by this handled similar to a suspension culture (see Section 9.1.5). For mixing, standard impellers (e.g., rushton turbine, pitched or marine impeller) as well as large paddle impellers are used depending on the shear sensitivity of the cells, sometimes combined to achieve a homogenous mixing throughout the bioreactor at lowest possible stirrer speed [441, 549]. For aeration, different methods are shown in Figure 36 [550 – 552]. Surface aeration can be applied only for low cell densities and volumes (below 1 L). Bubble aeration may cause cell damage mainly due to foam formation, especially in case of serum-containing medium. By applying special aerators (ring sparger, microsparger) in combination with serum- or protein-free medium,

the detrimental effects of bubble aeration can be overcome [553 – 556]. Alternatively, bubble-free membrane aeration systems were developed [557 – 560]. They are characterized by low shear rates, but require large membrane areas. On laboratory scale, a tumbling membrane basket can be used for mixing [511]. Rotating/vibrating sieves provide a gentle aeration, but are suitable mostly for laboratory scale [561]. *Bubble-columns* and *air-lift reactors* were build up to thousand liter scale, are easy to handle, but require serum-free media to prevent cell damage through bubbles and foam [549, 562]. For continuous operation of stirred tank reactors in perfusion mode (see Section 9.1.7), several techniques for cell and/or product retention are in use as shown in Figure 37 (e.g., microfiltration and ultrafiltration, sedimentation, centrifugation, spin filters, acoustic filter, dialysis for cell, and product enrichment [563 – 577]). *Suspension reactors* are characterized by advantages such as: conventional reactor systems, know-how on design and sterile operation, good mass transfer, homogeneous mixing, possibility of sampling and determination of the concentration of the cell, and high scale-up potential. Disadvantages are: difficult oxygen supply at high cell densities, cell damage by shear and/or foam (bubble aeration), relatively high demand for control (temperatur, oxygen, pH, flow rates), cell retention required for h igh cell densities.

Among high-density systems, we find *fixed-bed* and *fluidized-bed reactors* as well as *hollow-fiber reactors*. In fixed-bed and fluidized-bed bioreactors, cells are immobilized within macroporous carriers. The carriers are arranged in a column either packed (fixed-bed reactor [578 – 584]) or floating (fluidized-bed reactor [585 – 587]). The column is permanently perfused with a conditioned medium from a medium reservoir, mostly in a circulation loop. These types of reactor are very efficient for the long-term perfusion culture of mammalian cells for the production of biopharmaceuticals (e.g., monoclonal antibodies, recombinant drugs including tPA and EPO). With respect to tissue engineering, they have been investigated for several applications including the cultivation of "liver" cells as an extracorporal liver device [588 – 590] or proliferation of stem cells [591 – 593]. Furthermore they were successfully used for cultivation of cells producing viruslike particles for gene therapy [594 – 597]. Again, scale-up is critical for both fixed-bed and fluidized-bed reactors. For fixed-bed reactors, a radial-flow fixed-bed geometry was suggested [598 – 600] and successfully applied up to approximately 25 L fixed-bed volume. At perfusion rates of 10-20 L L_{FB}^{-1} d^{-1}, this would correspond to suspension reactors in the thousand liter scale. Alternatively, rotating-bed bioreactors are in use [601]. For fluidized-bed bioreactors, several concepts for scale-up have been suggested as well [514, 585 – 587, 602]. Compared to suspension reactors, fixed-bed and fluidized-bed reactors have the following advantages: high volume-specific cell density (approx. 5×10^7 cells/mL reactor) and productivity, low shear rates, simple medium exchange and cell/product separation, productivity on a high level during long-term cultures. Disadvantages are: nonhomogeneous cell distribution, difficult determination of cells and cell harvest.

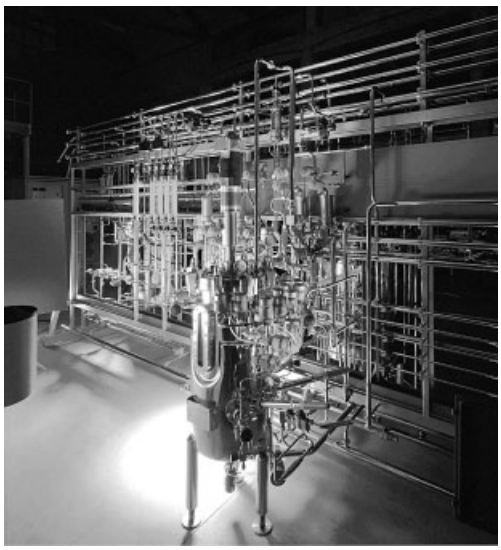

Figure 35. Multipurpose bioreactor for continuous perfusion culture of mammalian cells (Bioengineering AG, CH-Wald, with permission)

Within *hollow-fiber reactors* (compare Section 9.1.5), cells are immobilized in the extracapillar space of a hollow-fi4ber bundle. Nutrients as well as oxygen pass through the fibers, metabolites leave the culture chamber by this way. Cell concentrations comparable to those found in tissues are possible. The use of hollow-

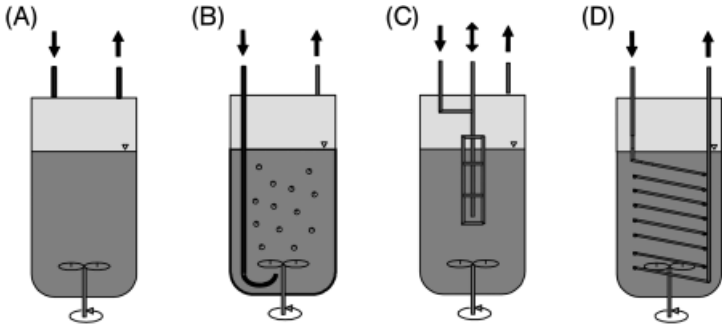

Figure 36. Aeration systems for stirred-tank reactors
A) Surface aeration; B) Sparging/bubble aeration; C) Rotating or vibrating sieve; D) Bubble-free membrane aeration

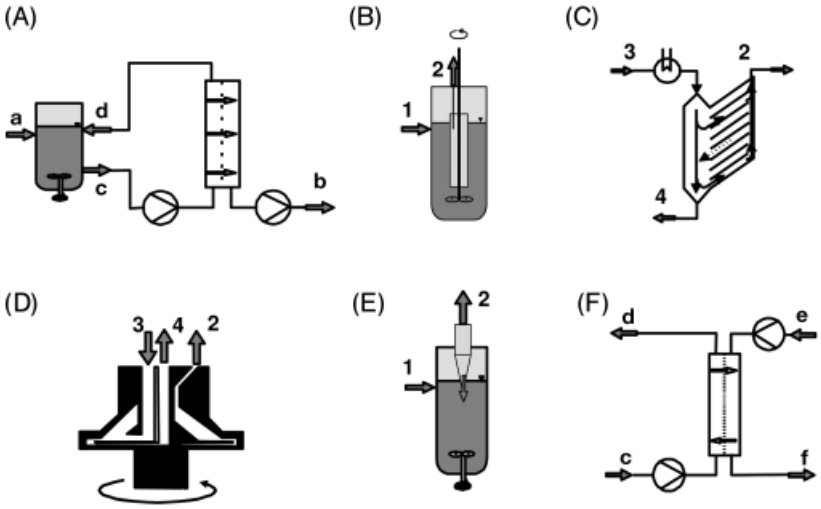

Figure 37. Devices for cell or/and product retention
A) Micro- or ultrafiltration, internal or external; B) Spin filter, internal or external; C) Settler, external or inclined; D) Centrifuge, external; E) Acoustic filter, inclined; F) Dialysis, internal or external
a) Feed; b) Cell-free harvest; c) Medium from bioreactor; d) Medium back to bioreactor with enriched cells; e) Dialysing fluid in; f) Dialysing fluid out

fiber culture is a convenient method for making moderately large quantities of high-molecular-weight products secreted by human and animal cells, at high concentrations and with a higher ratio of product-to-medium-derived impurities than is generally achieved using homogeneous (e.g, stirred-tank) culture [603 – 605]. Advantages are: very high cell densities, concentrated product, lower serum demands (if required), long-term stability, easy handling, and low cost. Disadvantages are: limited scale-up, mass-transfer problems/concentration gradients, difficult determination of the cell number, proteolytic activity on the product.

During the last years, a growing number of disposable or single-use bioreactor systems appeared on the market besides hollow-fiber reactors, addressing especially problems related to early process development such as flexibility, cost effectiveness, time-to-market as well as quality and regulatory issues [606 – 608]. These systems are mostly based on a bag technology [609, 610]. The bags with volumes up to 250 L are placed on tilting or rocking devices for mixing and oxygen supply. Advantages of single-use bioreactors are reduced cleaning procedures, lack of validation issues (cleaning, sterilization, etc.), lower investment costs, easier adaptation to changing process demands, less contamination risk [611 – 613].

The main characteristics of the culture systems discussed above are summarized in Table

18. All these systems support growth of mammalian cells in one or the other way. On laboratory scale, mostly disposable flask, membrane, or bag systems are the method of choice. Small suspension reactors as well as fixed bed and fluidized bed reactors are mostly used for research or process development. For larger amounts of products, only suspension reactors or, up to a certain scale, fixed-bed or fluidized-bed reactors have the required scalability. For a detailed discussion, compare [440, 555, 614 – 618]. Bioreactor design, in particular, is addressed in [619].

9.1.7. Process Strategies

As for all biotechnological processes, process strategies for operation of bioreactors can be classified in discontinuous (batch, repeated-batch or fed-batch) and continuous modes (chemostat, perfusion). Batch and fed-batch processes are mostly performed in small-scale culture systems (e.g., flasks, bags) or on larger scale in suspension reactors. Batch processes are usually starting with an initial cell density of approximately 1–2 10^5 cells mL^{-1} and last 1–2 weeks. After harvest, the bioreactor has to be cleaned and refurbished before the cycle is repeated [438, 440, 441]. Cell and product yield can be significantly improved by applying a fed-batch strategy, were nutrients are added after depletion according to an appropriate feeding strategy [[621 – 634]. Batch and fed-batch strategies are applied quite frequently on industrial scale [438, 440, 536].

The drawbacks of discontinuous modes such as large reactor volumes, high maintenance (cleaning, sterilization, etc.) can be overcome to some extend by using continuous mode, especially perfusion with cell and/or product retention [438, 563]. Chemostat cultures without cell retention are a valuable tool for research (e.g., kinetic studies) [635 – 639] but not for production scale because of low cell and product concentration. Continuous perfusion present several advantages over batch systems [563, 640 – 643]. The advantages include the ability to grow cells to a very high density, the ease of handling media exchanges for the purpose of fresh feed and product harvest, the easy removal of metabolites and other inhibitors, and the prospect of easy scale-up. When dealing with adherent cells, perfusion reactor design becomes slightly simpler (e.g., fixed-bed and fluidized-bed) because of the immobilization of cells within macroporous carriers (compare Section 9.1.6). There is a growing number of reports on stable perfusion cultures even on an industrial scale over periods of several months [563, 644 – 646]. The main advantage of perfusion cultures can be seen in the reduced bioreactor size (approx. 1/10 of a suspension reactor without cell retention). Nevertheless, there are some difficulties in the perfusion concept [438, 563]. Further equipment such as the retention device itself, pumps for feeding, harvest, and medium circulation, storage tanks for feed and harvest are required. The amount of media needed to complete a moderate to long-term run can be excessively large. A further possible drawback of continuous cultivation is the possibility for variability over the time of the run. Batch or fed-batch cultivations are much shorter in duration and have fewer chances for random occurrences to happen. This is important in dealing with current manufacturing practice (cGMP, see below) conditions that dictate precise reproducibility or at a minimum the evidence of non-effects of minor deviations. In the meantime, there is a growing number of pharmaceutical products in the market, produced successfully from perfusion systems. Among these are block busters such as Kogenate-FS® (Bayer Health Care) or Remicade® (Centocor) [563]. Previous concerns about getting such processes approved are not longer an issue and it can be expected that because of an increasing pressure on production costs more perfusion processes will be installed in the future.

A process that becomes more and more important for production of small quantities of recombinant proteins is *transient transfection* [647 – 649]. Usually, nontransfected cells (most commonly HEK-293, COS, and BHK cells, compare Section 9.1.2) are first cultivated in batch mode to a certain cell density and then transfected with DNA encoding for a certain protein [441]. Usually, the cells are genetically not stable and soon lose their expression ability but can produce a certain amount of protein within one batch. Originally, it was used just as a preliminary test of gene expression. New developments show that large scale production up to 100 L is possible [650, 651].

Table 18. Main characteristics of cell culture systems and bioreactors ([620], modified)[a]

	T-flask/roller bottles/disposable bags	Membrane reactors (hollow-fibre, miniPerm, etc.)	Suspension (stirred-tank, air-lift reactors)	Fixed-bed/fluidized-bed reactors
Cell density	1	3	2	2–3
Homogeneity	1	1	3	2
Shear stress	N	N	Y	N
Product concentration	1	3	2	2
Porductivity	1	3	2	2
Medium efficiency	3	1	3	2
Continuous process (perfusion)	N	Y	Y[b]	Y
Control	0	1	2	3
Downstream process	1	3	2	2
Steam sterilizable	N	N	Y	Y
Reusable	N	N	Y	Y

[a] 0, not possible; 1-3, increasing efficiency; Y, yes; N: no.
[b] Additional equipment for cell retention required.

9.1.8. Downstream Processes

The desired protein product, mostly extracellular, has to be purified after cell cultivation by a combination of unit processes. From an engineering point of view, the challenges are 1) low product concentrations in the cell-free supernatant, 2) the complex structures of the proteins, especially the recombinant, glycosylated proteins, resulting in a high sensitivity against temperature, shear forces, extreme pH, or proteolytic activity, 3) contamination by nucleic acid (DNA fragments from dead cells), 4) contamination by other proteins (e.g., from serum), 5) potential contamination by viruses or virus particles [652, 653]. The main process steps involved are [654]: concentration of supernatant (e.g., precipitation, centrifugation filtration, or aqueous two-phase partition), primary purification/capture (e.g., chromatographic techniques such as ion-exchange, gel, or immunoaffinity chromatography as well as precipitation), further purification (e.g., ion-exchange, hydrophobic, and gel chromatography), and finally polishing (gel filtration, diafiltration, ultrafiltration) and techniques for virus inactivation (acid/base treatment, nanofiltration or ultraviolet-c (UVc)) [653, 655, 656]. The early stages within a purification strategy will aim to achieve both concentration and purification. The next step will be the major purification step. The number of steps included in "further purification and "polishing" will depend on the intended use of the protein, as they are mainly designed to reduce contaminants such as DNA, host cell protein, endotoxins, process chemicals, etc. The importance of the downstream process is underlined by the fact, that an estimated 60–80 % of the total production costs are attributed to purification [438, 439, 654].

9.1.9. Regulatory and Safety Issues

Pharmaceutical production processes are subject to strict regulations and regular inspections to ensure the high quality demands for medicines [439, 440, 657]. Guidelines for aspects relevant for the production process are summarized under cGMP and corresponding lists were released in the United States by the FDA [480] as well as by the EU [481]. There are some special concern with the use of mammalian cells for production of biopharmaceuticals, e.g., 1) risk of contamination by microbes, mycoplasma, or viruses, 2) the genetic stability of the host cell line, or 3) the consistency of the product (glycosylation pattern, etc.). The documents to be submitted to the authorities include detailed information concerning the product, the production process, product characterization including analytical methods, validation procedures, methods, and results of clinical trials. The efforts to fulfill these requirements are quite high and bind significant resources within the companies [439].

9.2. Tissue Engineering

In this chapter, an overview in the field of "red biotechnology" is presented. Red biotechnology comprises three major fields:

- Gene therapy
- Production of proteins, antibodies, and vaccines for diagnostics and therapy
- Tissue engineering

Gene therapy is a therapeutic approach based on the correction of a gene defect causing a disease. Different techniques exist for achieving this aim. In most cases, the wild-type gene is inserted into the genome by special methods replacing the mutant gene, which leads to the development of the disease. The most common vectors for the transport of the correct gene into the genome are viruses.

In the mid-eighties, several treatments with the genetically modified cells were reported. The first gene therapy approach was reported in 1990 on a young girl suffering from severe combined immunodeficiency (SCID). But the therapy did not work out, after a few months the procedure had to be repeated regularly.

A major drawback within this field emerged when it was evident that a young patient died after being treated by gene therapy [658]. The patient was suffering from a rare liver disease and was a volunteer in a gene therapy program. Similar incidences occurred in the year of 2000 within a group of boys after gene therapeutic treatment of SCID. The majority of the boys in this group developed leukemia. Since then, gene therapy is only broadly applied within science-fiction movies. Researchers working in this field went back a few steps to elucidate the devastating incidences [659].

Biological active compounds produced from genetically modified cell cultures have been in use for several years in the treatment of a whole range of diseases. In most cases, the active ingredients are proteins such as blood clotting factors (e.g., tissue plasminogen activator, t-PA), growth factors (e.g., erythropoitin, EPO), or monoclonal antibodies [660]. All these substances can be produced in animal cell cultures. These proteins can also be applied in diagnostic assays (e.g., ELISA). For more "simple" molecules such as non-glycosylated proteins, peptides and hormones, bacterial or yeast cells can be utilized for production.

9.2.1. Application of Tissue Engineering

Tissue engineering is a rather young and rapidly growing interdisciplinary research field which combines the principles of engineering, life, and clinical sciences (Fig. 38). The aim of this research field is the development of biological substitutes that restore, maintain, or improve tissue functions [661, 662]. Therefore, the knowledge of cellular interactions, molecular transport phenomena, interactions of cells with material surface, mechanical loading, and biochemical interventions is applied.

An official definition was presented at a conference of the National Science Foundation in 1988: "*... the application of principles and methods of engineering and life sciences toward fundamental understanding of structure–function relationship in normal and pathological mammalian tissues and the development of biological substitutes to restore, remain, or improve tissue function*".

Over the past two decades, encouraging results have been achieved. Methods and techniques of tissue engineering have already entered therapeutic practice, e.g., for the treatment of sports injuries and damage to cartilage. The main approach adopted in these cases is known as autologous chondrocyte transplantation (ACT). Cells isolated from cartilage tissue of the patient (autologous) are seeded onto a collagen matrix and cultivated until a three-dimensional tissue structure is generated. Then the construct can be implanted into the defect [663 – 665]. Also tissue-engineered skin is already used, e.g., for therapy of severely burned patients [666 – 669]. Researchers are attempting to engineer almost every human tissue or organ because the demand of tissue-engineered cells and organs is still growing because of the fact of organ shortage for transplantation. The creation of three-dimensional tissue constructs which can mimic all functions and metabolic activity like organs (e.g., liver [670] or pancreas [671]) is still a crucial challenge. The first tissue-engineered heart valves have been implanted to a patient and show promising results in function and mechanical stability [672, 673].

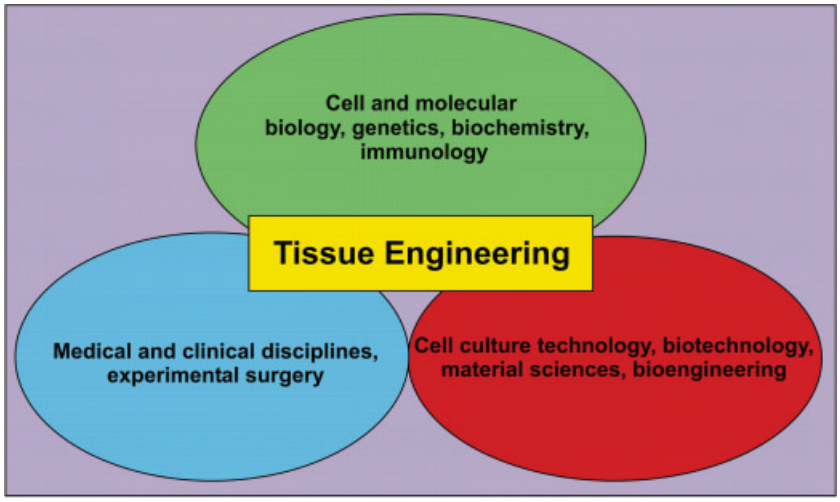

Figure 38. Disciplines involved in tissue engineering

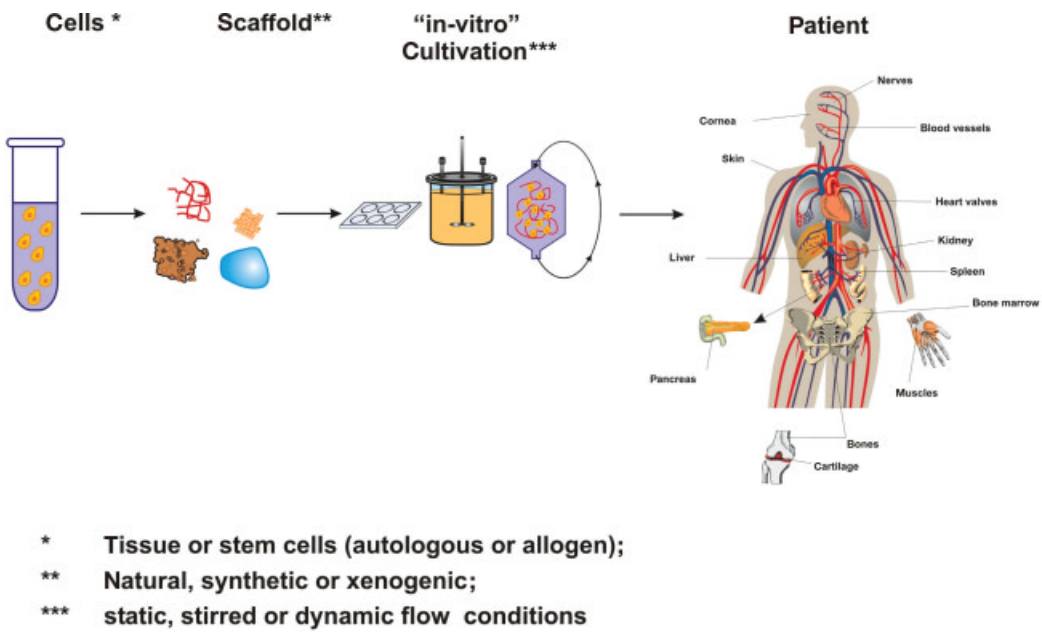

Figure 39. Principle of tissue engineering

9.2.2. Principle of Tissue Engineering

The principle of tissue engineering can be demonstrated in the following way: First, cells have to be isolated. The cells are seeded to a matrix and are extracorporal cultivated in a suitable culture system or bioreactor in medium containing nutrients and growth factors. After cultivation, the resulting tissue is reimplanted into the patient (Fig. 39).

Tissue engineering is driven by the hopes of the patients of increased availability of transplantable organs, to ease their suffering, forestall imminent death, and enabling a better quality of life. Alongside these medical objectives, there are also huge economic issues at stake because treatment of tissue and organ damage or dysfunction often involves long hospitalization and thus is very expensive. In contrast to alternative approaches of xenogenic (from an-

other species) and allogenic (same species but another individual) organ transplant, tissue engineering involves transplanting cells from the body of the patient itself. This is the reason for which such great attention has been focused in the last few years on this new area of research. By growing artificial autologous tissue and organs in vitro, the risk of rejection by the immune system of the body can be avoided. Tissue engineering techniques can be used to generate replacement tissue for internal organs (liver, pancreas, heart, kidneys, lungs), sensory organs (eyes, ears, nose), the skeleton (bones and cartilage), the brain and nerves (in particular for the treatment of Alzheimer's disease and Parkinson's disease), or the skin (e.g., treatment of severe burns).

A closely related concept to tissue engineering is *regenerative medicine*. Here, the emphasis is placed on supporting the body's natural healing processes. The potential of stem cells is crucial for both fields.

9.2.3. Strategies

Tissue engineering can be performed by using one of the three approaches:

In vitro cultivation of autologous cells on organic, synthetic, or natural matrices
In vitro cultivation of autologous cells on xenogenic matrices
Iin vitro tissue generation from embryonic stem cells

Depending on the tissue of interest, either a close system (e.g., extracorporal device, artificial liver) or an open system based on a biodegradable polymer scaffold is developed. The autologous cells can be isolated tissue-specific primary cells or stem cells.

9.2.4. The Essentials

The essentials for tissue engineering are shown in Figure 40. For a successful generation of three-dimensional tissues, suitable cells have to be isolated and cultivated on a matrix in an appropriate medium containing all necessary growth and differentiation factors by using an optimized bioreactor system. During cultivation, the application of a mechanical load (e.g., strain, compression) is beneficial for the development of three-dimensional cell–cell contacts resulting in functional tissue [674 – 682].

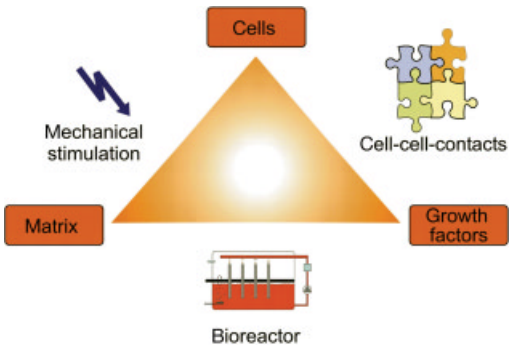

Figure 40. The "essentials" for tissue engineering

9.2.5. Cells

For tissue engineering, either tissue-specific cells can be isolated and cultured or stem cells can be used for the directed differentiation to the desired tissue. Due to their potential for differentiation into several different tissue types, nowadays the use of stem cells is intensively discussed [683]. In numerous publications adult stem cells isolated from bone marrow (bone marrow stromal cells BMSC [684 – 686], fat tissue (adipose-derived stem cells adSC [687]), and peripheral blood or umbilical cord blood (both haematopoietic cells [688, 689]) are applied. The potential for differentiation into several different types of tissues is already approved. Mesenchymal stem cells (MSC) play a major role in modern tissue engineering [690, 691]. These cells are easy to isolate from bone marrow or fat tissue and their availability is not so limited compared with that of the small number of stem cells present in, e.g., peripheral blood (0.1–0.2 %). Treatment concept based on the use of autologous stem cells would eliminate problems of donor site scarcity, immune rejection and pathogen transfer.

9.2.6. Biomatrices

There has been also huge progress in the development of biomaterials for use in tissue-engineering applications [692, 693]. Research

has concentrated on the development and production of biocompatible and biodegradable materials [694 – 697]. Such biomaterials take over the task of the extracellular matrix (ECM), which naturally provides cells with a supportive framework of structural proteins, carbohydrates, and signaling molecules. The ideal scaffold has to mimic ECM features and would be made of a biomaterial that provides all the necessary signals for the cells to grow, differentiate and interact, forming the desired structure. Regardless that general properties and design features of the biomaterials for discussed purposes have been published in a numerous reviews [698 – 702], the subject still represents intensive scientific and practical interest. Besides natural occurring compounds or scaffolds, also synthetic biomaterials are reviewed in the literature [703 – 705]. Along with polysaccharide- and protein-based (e.g., hyaluronic acid, collagen) materials, also many synthetic polymers are used for tissue-engineering application (e.g., polylactic acid, polyglycolic acid, polyurethane, polyesters).

9.2.7. Bioreactors for Tissue Engineering

The design, control, and operation of bioreactors in technical and engineering disciplines are well-established. Their use in medical sciences requires an interdisciplinary approach aiming at the development of functional tissues. Therefore, frequently the application of mechanical loading is needed [706 – 709]. The loading can be disposed in different ways; for example, a longitudinal straining leads to the formation of ligaments or tendons; compression is applied when aiming at the development of, e.g., cartilage; pressure (perfusion) can be applied for bone tissue formation and pulsed bioreactors are used for blood vessel generation. In Figure 41, a selection of bioreactor schemes for tissue-engineering applications is presented. Unlike in conventional mammalian cell culture, where only a limited number of bioreactor types are used (e.g., T-flaks, roller bottles, spinner systems, stirred-tank reactors), in tissue engineering tailor-made bioreactors have to be developed and constructed for each specific tissue type, considering in particular the fluid dynamics which are realized by these different systems.

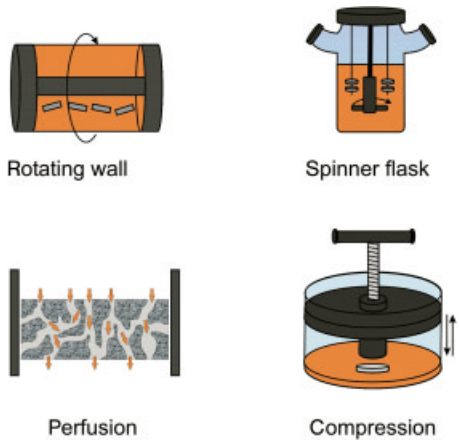

Figure 41. Selected bioreactor types for tissue engineering

Since human cells are anchorage-dependent, they must be attached to a matrix in the bioreactor during the cultivation. In a rotating-wall system, the cell-seeded matrices are cultured while the reactor wall is rotated, thus the cells float through the medium and are affected by microgravitylike conditions, and the mass-transfer rates are improved. Moreover, cell-seeded matrices can be cultured while hanging in spinner flasks, which results also in an improved mass transfer. Another possibility is the continuous pumping of the medium through a fixed bed with cells or the cells can be placed into a chamber where a mechanical stimulation (here compression) is applied.

9.2.8. Growing New from Old

The idea of organs one day being freely available "off the shelf" is still an aspiration today. The need is great, however, and patients are of course very eager to have "personalized" treatment from "organ designers" using tissue engineering. The most ambitious concepts involve applying a sort of "building-block" principle to the culture of each tissue or organ type: so-called precursor or stem cells would be used, whose potential for differentiation could be exploited through the use of appropriate growth or differentiation factors. Thanks to this method, one would "only" need the appropriate "ingredients": Further advantages of this approach are the fact that an organ or tissue replacement can be carried out at any time – in other words it can

be scheduled in advance. Regenerative medicine has great potential: according to a report in Time Magazine, over the next 10 to 20 years, tissue engineering will become one of the top careers and a new era in life sciences and medicine dawns. The potential applications are highly diverse but are to be found essentially in the treatment of cells and tissue which possess little or no capacity to repair themselves such as cartilage, heart muscle and nerve tissue.

The medicine of tomorrow will strive to develop personalized therapeutic procedures involving the patient in each treatment choice. There will be a move from a general "one-size-fits-all" approach to personalized treatment based on the principle of "one drug, one therapy, one patient". This will mean using a systems biology approach to fully restore the original healthy state of the patient. This dream could be realized thanks to the emerging technologies of regenerative medicine and tissue engineering.

Groups of cells and tissue generated by using tissue engineering can also serve as test systems which exhibit characteristics specific to a given organ. These can be used to test the effectiveness and duration of certain drugs or to study their mechanisms and metabolism, thus representing an alternative to animal experiments, and a rapid, broadly applicable and cost-effective way of screening pharmaceutically active substances.

Over the last ten years, there has been an explosion of research in tissue engineering, which shows the enormous importance attached to this technology. Within Europe, German researchers are playing a major role. Leading research is also being carried out by teams in France, Britain, Italy, and the Netherlands. Large numbers of small biotechnology companies are working in tissue engineering and more and more are being set up all the time.

In the 21st century, major scientific questions concern modern developments within the field of regenerative medicine, which includes tissue engineering, transplantation, prosthesis, as well as pharmacological interventions stimulating in situ tissue regeneration.

9.3. Biotechnology and Food

It is not presumptuous to claim that modern biotechnology has developed from traditional processes of food preparation. The production of alcoholic beverages, the coagulation of milk proteins, and the preservation of vegetable food by fermentation has already been familiar to most ancient civilizations. Thus, several microorganisms have a very long history of safe use in human food production. The main focus of this chapter is not on traditional fermentations but on recent developments in food biotechnology. Since this field is remarkably ambitious and complex, only some techniques of high industrial relevance are covered exemplarily. Genetic engineering of plants has yielded, for example, herbicide tolerant "RoundUp Ready" soy, *Bacillus thuringiensis* toxin expressing maize (Bt maize) and transgenic rice lines with an increased concentration of β-carotene in the endosperm ("golden rice"). Genetically engineered animals may produce milk with an altered protein profile, and modification of the hormone expression of fish imparts accelerated growth rates. A number of comprehensive reviews on food biotechnology are available [710].

9.3.1. Production of Food Additives by Cell Culture Systems

9.3.1.1. Amino Acids

L-Amino acids represent the largest group of food and feed additives produced by means of biotechnology [711]. Consequently, the literature published within the last decade is notably versatile. L-Glutamic acid ((*S*)-2-aminoglutaric acid, E 620) was the first amino acid produced biotechnologically on an industrial scale, and is still the leading amino acid by volume. More than 800 000 t/a of monosodium glutamate, which serves as a flavor enhancer ("umami taste"), are produced worldwide by cultivation of *Corynebacterium glutamicum* and related organisms. The production of L-glutamic acid may be enhanced by overexpression of the glutamate transporter gene (yhfK) [712], by use of streptomycin resistant *Brevibacterium lactofermentum* strains [713], and by optimization of the

culture conditions [714]. The second most important amino acid is L-lysine ((*S*)-2,6-diaminohexanoic acid), which is mainly used to supplement animal feed with this essential and often limiting amino acid. It is industrially produced by *C. glutamicum* or by *E. coli* species [711]. Efficient industrial producer strains have been developed by multiple rounds of mutagenesis. Defined L-lysine producing *C. glutamicum* mutants have been generated for example by modification of the enzyme aspartate kinase (lysC mutation) [715], attenuation of the transcription regulator tipA gene [716], and the introduction of a mutation (S361F) into the 6-phosphogluconate dehydrogenase gene (gnd) [717]. In the latter case, the modified strain was less sensitive to allosteric inhibition by intracellular metabolites than the wild-type strain. Product yields of up to 100 g/L are attainable within 24 h of cultivation with the recombinant production strains. A third amino acid, L-cysteine ((*R*)-2-amino-3-mercaptopropanoic acid), which is applied as a dough conditioner in the bakery industry, may be obtained from recombinant *E. coli* [718 – 720].

9.3.1.2. Organic Acids

Citric acid (2-hydroxy-1,2,3-propanetricarboxylic acid, E 330) is widely used to lower the pH of food (candy) and beverages (soft drinks, wine). Furthermore, it acts as an efficient chelator of prooxidative iron and copper ions. Various fungi and bacteria may be used for the biotechnological synthesis of citric acid starting from numerous substrates. The process has recently been comprehensively reviewed in [721]. *Aspergillus niger* grown as surface cultures on diluted molasses has been employed as early as in the 1920s, and still is the most popular production organism. Today, distillers' grains [722], palm oil mill effluents [723], or cassava starch factory wastes [724] may serve as cheap and abundantly available substrates. By optimization of the growth conditions (sufficient oxygen supply in submerged cultures; media deficient in manganese, copper, iron, and zinc ions; high glucose concentrations), and by selection of overproducing *A. niger* strains, yields of > 100 g/L are attainable [725]. The yeast *Yarrowia lipolytica* represents a future potential production organism. The strain *Y. lipolytica* 187/1 showed maximum synthesis rates when grown on raw agro-industrial fats as substrates [726], while Rymowicz and Cibis [727] selected the acetate-negative strain *Y. lipolytica* AWG-7, cultivated on glucose syrup, for citric acid production. As a second important organic acid, L-lactic acid ((*S*)-2-hydroxypropanoic acid, E 270), has been used from ancient times as a food preservative by lowering the pH during fermentation. More recently, lactic acid has gained increasing interest since the monomer is used for generation of the biodegradable food packing polymer "polylactic acid (PLA)". However, only few PLA-generating processes have been commercialized and the cost of PLA is still not able to compete with synthetic plastics [728, 729]. Apart from traditional lactic acid bacteria such as *Lactobacillus delbrueckii* [730] or *L. plantarum* [731], a wide range of further microorganisms, especially strains of *Rhizopus oryzae,* has been employed for the biosynthesis of L-lactic acid (for example [732, 733]). A current review on metabolic engineering approaches of lactic acid bacteria is available [734].

9.3.1.3. Vitamins

L-Ascorbic acid (vitamin C, E 300) is produced by a combination of chemical synthesis and enzymatic reaction steps. Known as "Reichstein process", the procedure was established already in the 1930s. The key biotransformation step, the oxidation of D-sorbitol to L-sorbose, is catalyzed by *Acetobacter suboxidans* [735]. Due to the worldwide use of L-ascorbic acid as vitamin in dietary supplements and as antioxidant in food and feed, the annual need sums up to 60 000 t [711]. Novel approaches use either recombinant microorganisms in one-step procedures [736] or *Azotobacter chroococcum* strains [737] for L-ascorbic acid production. For a review on biotechnological routes to vitamin C, see [738]. Further water-soluble vitamins, including riboflavin (vitamin B_2) and cyanocobalamin (vitamin B_{12}), are currently produced by microbial cell cultures. The chemical synthesis of the yellow-greenish-colored riboflavin has been largely replaced by biotechnological processes using either the overproducing fungus *Ashbya gossypii* [739] or mutant strains of *Bacillus subtilis*. The "generally recognized as safe" (GRAS)

organism *B. subtilis* is widely used in the food industry [725] and, for example, the mutant strain CJKB0002 produced 26.5 g/L riboflavin compared to 22.4 g/L obtained from the parent strain [740]. Detailed mechanistic studies on the riboflavin synthesis in *Ashbya gossypii* [739] and in the lactic acid bacterium *Lactococcus lactis ssp. cremoris* [741] have been published. Cyanocobalamin is derived from *Propionibacterium* or *Pseudomonas denitrificans* species [711]. Recently, a fermented milk product containing *Propionibacterium freudenreichii ssp. shermanii* B369 bacteria, capable of producing cyanocobalamin in situ, was patented [742]. Thus, a vitamin B_{12} enriched yogurt may be manufactured as a "nutraceutical" without external vitamin fortification. However, comparatively low world-wide annual production rates of 3000 t for vitamin B_2 and 10 t for vitamin B_{12} document the predominant use of theses vitamins in dietary supplements and in pharmaceuticals [711]. The main plant-derived carotenoid, β-carotene (E 160a), serves as provitamin A in the human diet. Since β-carotene additionally imparts an intense yellow-orange coloring, it is widely used as a natural food additive. The development of biotechnological processes for the production of β-carotene is mainly driven by the increasing demand of the food industry. Besides the extraction from plants (for example from lucerne), β-carotene is industrially obtained from the algae *Dunaliella salina* [743] or from the fungus *Blakeslea trispora* [744]. The carotenoid astaxanthin, which is used on a large scale in aquaculture for the coloring of salmonids and crustaceae, is formed by the yeast *Phaffia rhodozyma* [745].

9.3.1.4. Sweet Compounds

Palatinit (isomalt, E 953) is a non-cariogenic sugar substitute used in low-calory food, desserts, ice creams, and diabetic food. It is a white, crystalline, and non-hygroscopic substance with a sweetening power of roughly 0.45–0.60 compared to that of sucrose (1.0). In the first production step, sucrose is converted enzymatically (by action of a glycosyltransferase) into palatinose (isomaltulose, *O*-α-D-glucopyranosyl-(1–6)-D-fructofuranose) by immobilized cells of *Protaminobacter rubrum*. Subsequently, palatinose is reduced to palatinit, an equimolar mixture of the corresponding alcohols (1-*O*-α-D-glucopyranosyl-D-mannitol and 6-*O*-α-D-glucopyranosyl-D-sorbitol) by hydrogenation on a fixed-bed Ni catalyst. In the industrial process, it is advantageous to use immobilized non-viable cells of *P. rubrum* in a packed-bed reactor for the enzymatic step [746]. Apart from *P. rubrum*, immobilized cells of *Serratia plymuthica* [747], *Erwinia* sp. D12 [748], or *Klebsiella singaporensis* LX3T (isolated from soil) [749] have been used to convert sucrose into palatinose. In 2005, palatinose was approved according to the EU regulation (EC 258/97) concerning novel foods and novel food ingredients, and the GRAS registration was granted in 2006 by the US Food and Drug Administration (FDA). Recent inventions suggest a mixture of palatinose with cocoa powder for the preparation of instant beverages [750]. Further applications comprise the use of palatinose for the production of low-alcohol beer or beerlike soft drinks [751], and the preparation of fermented soybean milk [752].

In the industrial production of the sweetener aspartame (α-L-aspartyl-L-phenylalanine 1-methyl ester, E 951), the methyl ester of the "DF" dipeptide, both components are produced by biotechnological processes. L-Phenylalanine is usually manufactured by large-scale cultivation of the soil bacteria *Serratia marcescens* or *C. glutamicum*. Immobilized aspartase or alternatively intact cells of *E. coli* are used to synthesize L-aspartic acid starting from the precursors ammonia and fumarate [725]. Usually, amid formation is performed chemically although an enzymatic approach exists [753]. Thermolysine, immobilized on an agarose–aldehyd gel, is used as biocatalyst for the synthesis of enantiopure aspartame [754]. Thermolysine is a metallopeptidase with a zinc ion in its active site and four calcium ions, which stabilize the tertiary structure of the enzyme.

9.3.1.5. Sugar Alcohols

Sugar alcohols such as D-sorbitol (D-glucitol; E 420) and D-xylitol (E 967) have attracted the attention of the food and pharmaceutical industry because of their sweetening and non-cariogenic properties. Due to the fact that their metabolism

is insulin-independent, they are well suited for substituting sugar in diet schemes for diabetics. Since the biotechnological production of xylitol represents an economically attractive alternative to the industrialized chemical process, various process variants have been published. Somewhat less literature is available on sorbitol, which is mainly formed by cultures of *Zymomonas mobilis* [755, 756]. The synthesis of xylitol by *Candida guilliermondii* depends on several crucial process parameters, one of which is the concentration of phosphate buffer in the culture broth (optimum at 0.6 M) [757, 758]. In another approach, the D-xylitol production of the yeast *Debaryomyces hansenii* UFV-170 was monitored with respect to temperature and starting pH [759]. Various substrates, including for example hydrolysates of corn cobs [760] and sugarcane bagasse [761] have been used for the bioconversion of the pentose D-xylose to xylitol.

9.3.1.6. Microbial Saccharides

The most popular polymer currently produced by means of biotechnology is xanthan gum (E 415). Xanthan gum is a complex bacterial exopolysaccharide structurally closely related to cellulose. Today it is mainly produced by mutants of *Xanthomonas campestris* (40 000 t/a [711, 762]), a Gram-negative bacterium that is pathogenic to many crops. Due to its outstanding properties (high salt and pH tolerance, thermostability, and unusual rheological behavior), xanthan is predominantly used as a thickener in dressings and convenience food [763]. Sago starch has been suggested as an alternative carbon source instead of glucose [764]. Trehalose (α-D-glucopyranosyl-(1-1)-α-D-glucopyranose) is a nonreducing disaccharide that is thought to play a role as storage carbohydrate for many microorganisms. It may serve as an additive to improve the freeze stability of processed food. In 2000 and 2001, trehalose has gained the FDA GRAS and the EU novel food approval, respectively. Trehalose is obtainable inter alia from *Brevibacterium* [765], *Cellulosimicrobium cellulans* [766], or *Debaryomyces hansenii* strains. The latter two were cultivated under moderate saline stress for extracellular trehalose production [767]. Literature [768] reviews trehalose production (with a focus on the conversion of starch to trehalose) and potential applications in the food industry.

9.3.1.7. Conjugated Linoleic Acids (CLA)

Lactic acid bacteria play a prominent role in the traditional fermentation of food. One of the exceptional features of some lactobacilli is their ability to isomerize linoleic acid to different conjugated linoleic acid isomers (CLA). The *Z*9,*E*11- and *E*9,*E*11-octadecadienoic acid are assumed to possess beneficial physiological and anticarcinogenic properties [735]. *Lactobacillus plantarum* AKU 1009a formed up to 40 mg/mL CLA (33 % molar yield) in 108 h under optimized conditions from 12 % (w/v) linoleic acid as substrate [769]. Further lactobacilli capable of CLA production encompass for example *L. delbrueckii ssp. bulgaricus* [770] and *L. plantarum* JCM 1551 [771]. Castor oil proved to be an adequate substrate for CLA formation. Presumably, a two-step reaction mechanism, involving hydration of a double bond with subsequent elimination of water, is leading to different CLA isomers [772].

9.3.1.8. Lactulose

Lactulose acts as a prebiotic compound and thus may contribute to increased levels of "health-promoting" bacteria (bifidobacteria and/or lactobacilli) in the human intestinal tract [773]. The term "prebiotics" comprises a wide variety of oligosaccharides that are not digested in the stomach and in the small intestine. They naturally occur in garlic, asparagus, onion, and others. Apart from soybean oligosaccharides, a number of uncommon fructo-, galacto-, xylo-, and isomalto-oligomers are currently in use, especially in Japan [773]. The disaccharide lactulose (4-O-β-D-galactopyranosyl-D-fructose) is a non-naturally occurring compound with potential applications in infant milk formulations and foods [774]. Although lactulose is chiefly manufactured chemically using boric acid or aluminates as catalysts, a biotechnological route starting from lactose (as galactose donor) and fructose with permeabilized *Kluyveromyces lactis* yeast cells has been developed [775]. In

another approach, β-galactosidase from *Aspergillus oryzae* and a hyperthermostable β-glycosidase from *Pyrococcus furiosus* were used for the transgalactosylation of lactose in the presence of fructose [776]. Important food additives and food supplements produced by cell culture systems are summarized in Table 19.

9.3.2. Enzyme-Catalyzed Processes

The modification of starch and lipids as well as the production of cheese and fruit juices are prime examples of enzyme-catalyzed bioprocesses in the food industry. For economic reasons, the necessary enzymes are often employed as crude products or even as mixtures of enzymes with different properties. If required, the enzymes are purified by means of precipitation, ultrafiltration and/or fast protein liquid chromatography (FPLC). Common FPLC protocols comprise one ore more of the steps hydrophobic interaction chromatography (HIC), ion exchange chromatography (IEX), and size exclusion chromatography (SEC). Reusability and sometimes also stabilization may be achieved by immobilization of the biocatalyst on an inert carrier material.

9.3.2.1. Starch-Modifying Enzymes

Traditionally, the degradation of starch to maltodextrins and glucose/fructose syrups was performed by acid-catalyzed hydrolysis, using strong mineral acids such as hydrochloric or sulfuric acid. The harsh reaction conditions afforded a variable spectrum of unwanted by- and breakdown products. Therefore, the acid hydrolysis has been replaced more and more by enzymatic processes. Today, starch-converting enzymes comprise about 30 % of the enzyme market worldwide. Amongst the starch converting enzymes, α-amylase (EC 3.2.1.1, 1,4-α-D-glucan glucanohydrolase) holds the largest market share with major applications in the starch and bakery industries [777, 778]. α-Amylases may be derived from several plant materials, bacteria, yeasts, and fungi [779]. For commercial applications, mainly strains of *Bacillus amyloliquefaciens* and *Bacillus licheniformis* are employed as enzyme producers. α-Amylase attacks the α-(1-4)-links within the starch molecule (endoamylase), while the α-(1-4)-bonds near α-(1-6)-branches are resistant to hydrolysis. Thus, the end products of extensive α-amylase treatment are oligosaccharides of varying length with α-configuration and α-limit dextrins, which constitute branched oligosaccharides. Since α-amylase of *B. licheniformis* is remarkably thermostable (it remains active for several hours at temperatures of > 90 °C), its application in the industrial starch hydrolysis is highly efficient [780].

β-Amylases (EC 3.2.1.2; 1,4-α-D-glucan maltohydrolase) are present mainly in higher plants, but also in some microorganisms. Commercial products are available from malt, soy, and wheat. They hydrolyze α-(1-4)-glucosidic bonds, effecting successive removal of maltose units from the nonreducing end (exoamylase). In contrast to amylose, amylopectin is not completely hydrolyzed, as all reactions stop before branch points are reached (β-limit dextrin).

Glucoamylase (EC 3.2.1.3; 1,4-α-D-glucan glucohydrolase) is formed predominantly by fungi and yeasts (*Aspergillus*). It starts at the nonreducing end of α-(1-4)-D-glucans and successively liberates β-D-glucose units. The α-(1-6)-bonds in amylopectin are also cleaved, but around 30 times slower than the α-(1-4)-bonds. At higher concentrations, glucoamylase reforms all the glycosidic bonds that it hydrolyses. Thus, it condenses glucose, for example, to isomaltose (6-O-α-D-glucopyranosyl-D-glucose).

Debranching enzymes exclusively hydrolyze α-(1-6)-bonds in amylopectin, glycogen, and pullulan. Linear amylose fragments are formed from amylopectin. Pullulanase (EC 3.2.1.41; pullulan α-(1-6)-glucanohydrolase) hydrolyzes α-(1-6)-bonds in pullulan (a linear polymer of α-(1-6)-linked maltotriose units) and amylopectin, while isoamylase (EC 3.2.1.68) is only capable of the hydrolysis of the α-(1-6)-bonds in amylopectin. Pullulanases have been found in a variety of bacteria; industrially used is an enzyme derived from *Bacillus acidopullulyticus*.

Cyclodextrins are produced via an intramolecular transglycosylation reaction in which the enzyme cyclodextrin glycosyltransferase (EC 2.4.1.19; 1,4-α-D-glucan 4-α-D-(1,4-α-D-glucano)-transferase) cleaves the α-(1-4)-glycosidic bond and concomitantly links the reducing to the nonreducing end.

Table 19. Examples of food additives and food supplements produced by cell cultures

Product	Use	E number	Production organism
L-Glutamic acid	flavour enhancer	E 620	*Corynebacterium glutamicum*
L-Lysine	feed supplement	–	*C. glutamicum*
L-Cysteine	bread enhancer	E 920	*Escherichia coli*
Citric acid	acidity regulator	E 330	*Aspergillus niger*
L-Lactic acid	preservative	E 270	*Lactobacillus* ssp.
L-Ascorbic acid	vitamin C; antioxidant	E 300	*Acetobacter suboxidans*
Riboflavin	food coloring;vitamin B_2	E 101	*Bacillus subtilis*
Cyanocobalamin	vitamin B_{12}	–	*Propionibacterium* ssp.
β-Carotene	food coloring; provitamin A	E 160a	*Dunaliella salina, Blakeslea trispora*
Palatinit (Isomalt)	sugar substitute	E 953	*Protaminobacter rubrum*
Aspartame	sweetener	E 951	*Serratia marcescens, C. glutamicum*
D-Xylitol	sugar substitute	E 967	*Candida guilliermondii*
Xanthan	thickener; stabilizer	E 415	*Xanthomonas campestris*
Trehalose	cryoprotectant; stabilizer	–	*Brevibacterium* ssp.
Conjugated linoleic acids (CLA)	active ingredient of functional food	–	*Lactobacillus* ssp.
Lactulose	prebiotic	–	*Kluyveromyces lactis*

Syrups high in fructose level (42–44%, high-fructose corn syrups, HFCS) are obtained from glucose solutions by action of the enzyme xylose (glucose) isomerase (EC 5.3.1.5, D-xylose ketol isomerase) [725]. Technical enzymes are derived for example from *Streptomyces, Actinoplanes, Arthrobacter,* or *Bacillus* cultures. The large-scale isomerization of glucose is performed with immobilized enzymes, and product yields of up to 22 tons of isosugars per kilogram of catalyst are obtained. Fructose concentrations of about 90 % are accessible via a subsequent chromatographic purification step. An excellent survey of enzymes involved in the industrial processing of starch is presented in [781]; and the most important starch converting enzymes are summarized in Table 20.

9.3.2.2. Lipases

Lipases (EC 3.1.1.3, triacylglycerol acylhydrolase) are valued biocatalysts with numerous applications in food biotechnology. Many representatives of this enzyme class are highly stable in organic solvents, show broad substrate specificity, and high regio- and/or stereoselectivity [782]. Main fields of application include the rational design of lipids with specific structures ("functional oils and fats" [783]), and the production of volatile fatty acids and flavor esters ("natural food flavorings"). Volatile fatty acids are the so-called "character impact" flavor compounds of various types of cheese, while numerous flavor esters are responsible for fruity aroma impressions [784]. Industrial high-level production of lipases requires not only the efficient overexpression of the corresponding genes, but also a detailed understanding of the mechanisms governing the folding and secretion of the proteins [785]. Immobilized lipases from the yeast *Candida rugosa* are traditionally popular, as a wide core in the catalytic centre allows for acceptance of various fatty acids for *trans*-esterification or synthesis of mono- and diacylglycerides, for example, [786]. Especially *C. antarctica* lipase B features outstanding properties such as high thermostability, *sn*-2 recognition in hydrolysis of triacylglycerides, and selectivity towards *(E)*-fatty acids. A review on applications of the lipases A and B from *Candida antarctica* was published recently [787]. Several strains of another member of the *Candida* family, *Yarrowia (Candida) lipolytica*, were screened for lipase secretion. The highest lipase activity produced on rapeseed oil as sole carbon and energy source was found to be 2760 U/mL for *Y. lipolytica* 704 [726]. When cultured in a 2000 L bioreactor, the lipase activity reached 1100 U/mL after 53 h [788]. Lipases with novel catalytic properties are produced for example by the basidiomycetous fungus *Pleurotus sapidus*. An extracellular enzyme of the type B carboxylesterase/lipase family efficiently hydrolysed xanthophyll esters [789].

9.3.2.3. Pectin-Degrading Enzymes

Pectin is a complex polysaccharide (α-D-(1-4)-polygalacturonic acid, esterified with methanol in varying degree) that represents an essential

Table 20. Starch-converting enzymes

Enzyme	Enzyme number	Producer organism	Specificity
α-Amylase	EC 3.2.1.1	*Bacillus licheniformis*	*endo*-α-(1-4)
β-Amylase	EC 3.2.1.2	malt, soy, wheat	*exo*-α-(1-4)
Glucoamylase	EC 3.2.1.3	*Aspergillus* ssp.	*exo*-α-(1-4); exo-α-(1-6)
Pullulanase	EC 3.2.1.41	*Bacillus acidopullulyticus*	*endo*-α-(1-6)
Cyclodextrin glycosyltransferase	EC 2.4.1.19	*Bacillus macerans*	intramolecular transglycosylation
Xylose isomerase	EC 5.3.1.5	*Streptomyces* ssp.	isomerization of glucose to fructose

part of plant cell walls. The group of pectinolytic enzymes comprises pectinases (polygalacturonases), lyases, and pectin esterases. A total market share of 25 % of the global sales of food enzymes is currently attained by this enzyme family [790]. To improve the yields in fruit and vegetable juice production, and to obtain clear and concentrated juices, especially pectinases from the mold *Aspergillus niger* are employed [791, 792]. For example, the efficiency of kiwi juice production was increased $> 10\,\%$ by addition of 40 mg/L pectinase for 150 min [793], and the yield of loquat juice was 16 % higher when 160 mg/L pectinase were added for 4 h [794]. Under these conditions, the loss of vitamin C was low and the juice was clarified. Endo- and exopectinases (EC 3.2.1.15, endopolygalacturonase; EC 3.2.1.67, exopolygalacturonase) depolymerize the pectin chain by hydrolytic cleavage of the glycosidic α-(1-4)-bonds.

Usually, pectinases are combined with hemicellulases (mixture of cell-wall-degrading enzymes) and cellulases (EC 3.2.1.4) to improve fruit liquefication, leading to high-quality juices in terms of oxidation, aroma levels, and stability [795]. As an auxiliary enzyme, α-L-arabinofuranosidase (EC 3.2.1.55), was suggested for applications in the wine industry and for clarification of fruit juices [796]. Pectin lyase (EC 4.2.2.10) and pectate lyase (EC 4.2.2.2) catalyse the eliminative cleavage of pectin and pectate, respectively, to oligosaccharides with 4-deoxy-α-D-gluc-4-enuronosyl groups at their nonreducing ends. Potential production organisms include *Erwinia* species as well as a number of phytopathogenic fungi. Pectin esterases (EC 3.1.1.11) catalyze the hydrolytic cleavage of the methyl ester group. These enzymes generally exhibit a high specificity for pectin substrates.

9.3.2.4. Chymosin (Aspartic Protease)

Chymosin (EC 3.4.23.4) is a key enzyme with peptidolytic activity used in the production of cheese. The casein fraction of milk protein is coagulated by hydrolysis of κ-casein between Phe_{105} und Met_{106} to form para-κ-casein and a glycopeptide (κ-casein glycomacropeptide). From a mechanistic point of view, chymosin belongs to the class of aspartate peptidases, which are characterized by the presence of two aspartic acid residues in the catalytic center.

Since recombinant chymosin does not exhibit unwanted side activities and is not contaminated with microorganisms possibly causing quality problems, it replaces rennin from the stomachs of calves (abomasus) to an increasing degree. The gene encoding chymosin has been cloned into different microorganisms, and the properties of the recombinant enzymes expressed in the yeasts *Kluyveromyces lactis* and *Pichia pastoris* have been compared [797]. Since *K. lactis* is regarded as a safe organism traditionally used in food processing, it is of outstanding importance in the biotechnological production of chymosin on an industrial scale [798]. The technological characteristics of various enzymes from genetically modified organisms were nearly identical to those of natural rennin, and only slight differences in the coagulation time and curd firmness were observed [799]. Chymosin obtained from transgenic sheep milk and the recombinant protein expressed by *K. lactis* differed only in their action on low-molecular-weight substrates, but not in enzymatic activity on protein substrates [800]. Recently, extracellular acid peptidases from *Rhizopus oryzae* have been suggested as rennin substitutes in cheese manufacturing [801, 802].

9.4. Biotechnology and Health

9.4.1. Individualized Medicine

Single nucleotide polymorphisms (SNPs, a single base is exchanged by an alternate) are the most common form of genetic variant in the human genome, occurring on average 1 per 1000 base pairs. Since the influence of SNPs on medical therapy or diet is better understood, the SNP diagnostic is coming in the focus of interest. One important goal of genome research is studying genetic variance among individuals to improve the knowledge and treatment of disease and to identify the variations in our genes that influence the risk of an individual of becoming ill. Because inherited differences influence most common diseases and many drug responses, the understanding of existing genetic variations in the human population is a promising approach in the development of tailor-made therapeutics. Eighty percent of the individual variations are SNPs, which have a direct impact of gene regulation or of a protein product of a gene. In cases where one SNP is sufficient to cause disease it is possible to analyze the causal change and to improve our understanding of disease [803 – 806]. The hope is that a complete human SNP map will increase the efficiency of drug development and medical treatment of disease. Many genotyping platforms have been developed to detect many SNPs at once in a efficient, cost-effective manner including microarrays by the use of allel-specific oligonucleotides containing the variable base. Large-scale SNP genotyping should lead to a fully understanding of the genetic architecture of common traits underlying disease, drug or diet response. For example, microarray experiments could help to define the ideal dose of a drug and quantify the impact of this dose on human health. Since these are the most personally data we know, the access to these data must be protected so it is not misused by employers or insurers. The goal of this analysis is to relate SNP variation to disease or diet.

9.4.2. Clinical Diagnosis as Indicated in Genetic Anomalies in Cancer

Solid tumors often show alterations in the genome that lead to changes in DNA sequences and copy number. These changes can be detected and mapped by using comparative genomic hybridization (CGH). CGH is a molecular cytogenetic method of screening for genetic alterations. These changes are classified as DNA gains or losses and show a characteristic pattern that includes mutations at chromosomal level. In the past few years, microarray-based formats for CGH have been developed made from large genomic clones and cDNAs. These arrays provide many advantages over conventional CGH, including higher resolution, direct mapping of changes to the genome sequence, and high throughput. Together with the analysis of complex differences in gene expression between normal and cancer cells, these data provide insight into malignancy and reveal genes that may be useful as diagnostic or prognostic markers. One problem performing cancer diagnosis using gene expression profiles from microarray analysis considers the heterogeneity of a tumor. Samples taken from different locations in the tumors are dramatically different. Therefore, biological classification on the basis of random sample analysis leads to poor reproducibility and may not be adequate [807 – 809].

9.4.3. Pharmaceutical Development

Drug biotransformation in man and animals can lead to a modulation of pharmacodynamic effects, drug elimination, but also drug toxicity. Knowledge of drug biotransformation pathways at early stages of drug development could considerably speed up drug development [810 – 813] and increase the safety of drug development in the pharmaceutical industry. Each year, more than 668 000 animals are used in toxicological drug research in Europe. Animal models suffer from serious shortcomings regarding the prediction for a human situation as significant species differences in enzyme expression exist between man and animals. Specifically, enzyme induction experiments are frequently misleading. 10–20 % of all new drugs are required to return into preclinical development trials after first clinical tests because of unforeseen metabolite patterns, whose toxicological relevance is unclear or untested.

Classical animal-testing strategies are time-consuming and therefore rather expensive in

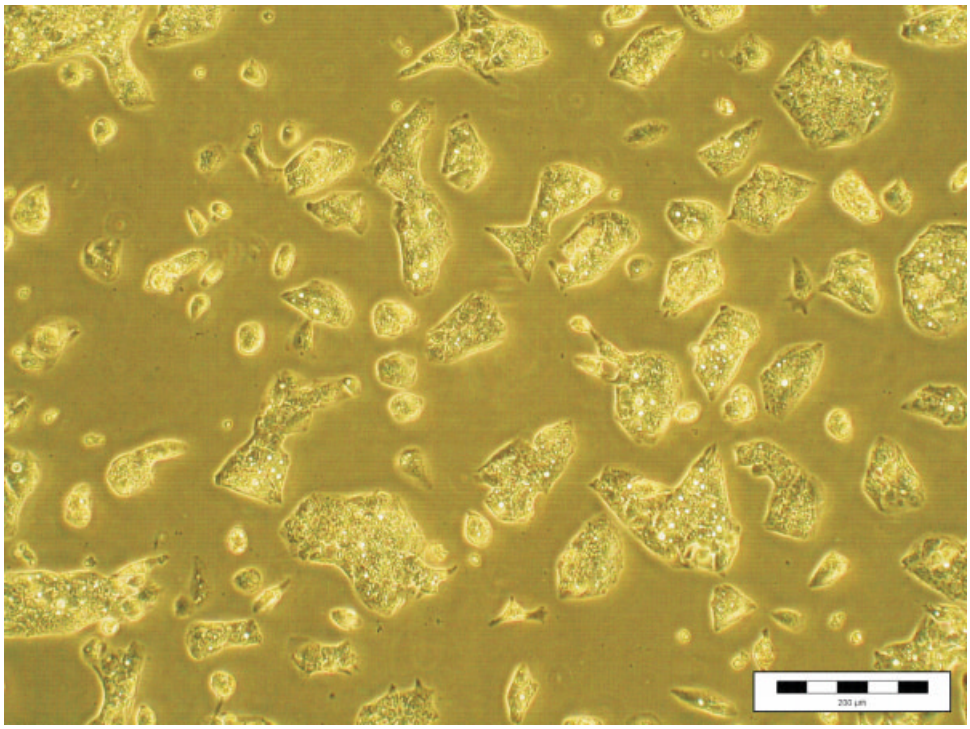

Figure 42. Human liver cells

view of a delayed market penetration and return of investment. The relevance of a strict control of development time is immediately explicable by the fact that preclinical development of a novel drug can total up to several hundred million Euros. In spite of this, modern cell-culture technology has progressed to the extent that it is now possible to culture primary human cells under conditions that maintain differentiated functions to perform drug biotransformation and enzyme induction studies with great reliability in individual laboratories. Human liver cells are at present the preferred target for drug studies as the use of such cells avoids species extrapolation problems (Figs. 42 and 43).

Another significant breakthrough in pharmaceutical drug metabolism research was that these models could be sufficiently automated to enter into the field of accelerated screening as a significant initiation step towards high-throughput screening. Industrial strategies are aiming at integrating pharmacokinetics and toxicology into the early phases of drug compound selection to save time, cost, and effort in drug development. Therefore, the cell cultures and the chip technology were brought together by the pharmaceutical industry for preclinical drug research. Induction phenomena could be correlated with microarray chips containing all known genes involved in drug biotransformation. This information is used to describe the culture models used in this study and to develop the tools to fingerprint interactions of inducing agents with drugs. The use of microarray technology in the field of drug discovery and development provide four significant advantages:

Figure 43. Bioreactor systems for in vitro testing

– Gene expression in response to drug treatments
– Gene expression in model systems
– Gene expression patterns in disease
– Gene expression patters in pathogens (host–pathogen interaction)

9.4.4. Define Molecular Mechanisms of Toxicity

DNA microarrays can be used to determine the relative safety of natural or synthetic chemicals to which humans are exposed. Lots of diseases are influenced by environmental factors, for example, chemicals, radiation, drugs, nutrition, viruses, carcinogens, reproductive toxins, neurotoxins, and immunotoxins [814 – 819]. Although most of these chemicals are harmless, it is important to identify potentially hazardous substances. In the past, toxicological assays have used rodent, required high doses, were expansive, and take years to complete. Unfortunately, the information gained from established life animal models such as rodent to investigate enzyme induction and toxicity cannot really lead to correct assumptions of hazards to humans with respect to species extrapolation. The shortcoming with respect to predictivity of animal models for human drug biotransformation and toxicity is based upon a well accepted species difference in enzyme expression between man and animals. Nevertheless, at present animal models in drug metabolism and toxicology are still legally required in all European countries for a variety of reasons and legislative bodies are still hesitative to recommend specific human liver cell based tests. In spite of this, modern cell-culture technology has progressed to the extent that it is now possible to culture primary human liver cells under conditions that maintain differentiated functions to perform drug biotransformation and enzyme induction studies with great reliability in individual laboratories. These developments have not yet found entry into legal recommendations as no prevalidation of such models on a European scale has been achieved so far. The widespread use of such alternative models to animal investigations is being hampered both by insufficient prevalidation and by further improvements still required in the field of automation. It would be a significant breakthrough in toxicology research if these models could be sufficiently automated to enter into the field of accelerated screening as a significant initiation step towards high-throughput screening. Microarrays are powerful tools for investigating the mechanism of toxicity (Fig. 44). A microarray chip for cytochrome P450s and Phase II enzymes can be used to evaluate the liver culture models [820].

9.4.5. Detection of Genetically Modified Organisms

The use of genetically modified organisms (GMOs) in food industry has been rising exponentially over the past years. In the space of time from 1996 to 2003, global area of transgenic crops increased 40-fold, from 1.7×10^6 hectares in 1996 to 67.7×10^6 hectares in 2003, with a dominant trait of herbicide tolerance, followed by insect resistance (www.isaaa.org, International Service for the Acquisition of Agribiotech Applications). A GMO is "an organism, with the exception of human beings, in which the genetic material has been altered in a way that does not occur naturally by mating and/or natural recombination" [821]. Microarrays are a useful tool for GMO detection and the application of array techniques to GMO detection implies the capability of sensitivity and standardization of the assay. Today, commercialized arrays often do not fulfill these requirements, in spite of a few promising attempts initialized by biotechnical companies and research projects. Nevertheless, microarrays are already being used in GMO analysis but well established quantitative real time PCR techniques are state of the art. GMO detection becomes increasingly important because by-products of these organisms could end up in the food chain and affect human health.

10. Concluding Remarks

As it has been demonstrated in the previous chapters, biotechnology offers manifold possibilities in the production of different products and in medical applications. It can be used for better food, better health, and a cleaner environment. However, in the area of production of fine and bulk chemicals, biotechnology has still to

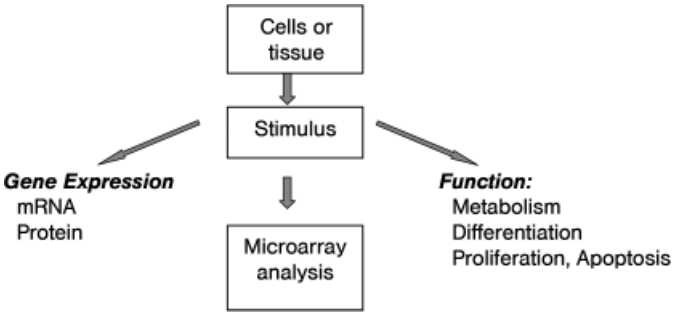

Figure 44. Investigation of toxicity through a combined cell-culture and microarray analysis

compete with the chemical industry. The chemical industry owes its success to the principle of unit construction: from simple basic substances, like ethylene, carbon monoxide or hydrogen, more complex precursors can be produced under controlled conditions by chemical reactions; because of the variety of combination options, the latter can in turn be converted into inconceivable quantities of derivatives and end products. Chemistry learned how to produce chemically pure basic substances from oil that are simple to handle and exactly defined; this is performed highly efficiently in refineries. This was the key to its success. Without exact knowledge of this functional principle, the triumphant success of plastics would have been just as impossible as the production of thousands of other chemical products that today make our lives safe and comfortable. This principle can also be applied to the so-called 'biorefinery' which has similar potential and could thus provide a further argument for the concept of material application of renewable resources and biotechnology. However, a technically viable separation procedure that would permit the separate use or further processing of all the primary products (such as cellulose, hemicellulose, and lignin) is still in its infancy. The main focus is on glucose, which can be derived by microbial or chemical methods from cellulose, since a wide range of biotechnological and chemical products can be obtained from glucose. The potential for the microbial conversion of glucose-based substances is huge and the reactions are advantageous from the viewpoint of energy. It is necessary to link the degradation processes to bulk chemicals via glucose as closely as possible to the synthesis processes of further derivatives.

Microorganisms still have a great potential for inducing new or novel enzyme systems capable of converting foreign substrates. It is possible to obtain and cultivate microorganisms that can survive or grow in extraordinary environments, e.g., psychrophilic, thermophilic, acidophilic, and alkaliphilic microorganisms. These microorganisms are capable of producing unique enzymes stable to heat, alkali, and acid. In addition, techniques for cultivation and enzyme purification, and gene and enzyme technology have made great advances and are still under development. Certainly biotechnological process development must be founded on a broader genomic basis. Ten years ago, only a handful of microorganisms with a completely sequenced genome existed, today there are several hundreds. Hitherto priority was given to sequencing pathogenic microorganisms, but it is now high time to sequence microorganisms that are relevant to industry. Taking all of these into consideration, there must be far more ways than one can imagine in using biotechnology for our daily life.

11. Acknowledgement

The current update was coordinated by Roland Ulber.

12. References

1. R. Ulber, K. Soyez, "5000 years of biotechnology", *Chem. Unserer Zeit* **38** (2004) 172.
2. R.D. Schmid: *Pocket Guide to Biotechnology and Genetic Engineering*, Wiley-VCH, Weinheim 2003.

3. Y. LeGal, R. Ulber: *Marine Biotechnology I and II; Advances in Biochemical Engineering/Biotechnology*, Springer, Berlin 2005.
4. H.G. Schlegel: *General Microbiology*, 7th ed., Cambridge University Press, Cambridge 1993.
5. M.T. Madigan, J.M. Martinko: *Brock – Biology of microorganisms*, 11th ed. (International Edition), Pearson, Prentice Hall, Upper Saddle River, NJ, 2006.
6. P.R. Murray, E.J. Baron, M.A. Pfaller: *Manual of clinical microbiology, American Society of Microbiology*, Washington D.C. 1999.
7. R. Wise et al., "Antimicrobial resistance – Is a major threat to public health", *BMJ* **317** (1998)no. 7159, 609.
8. J.M. Jay: *Modern Food Microbiology*, Springer, New York 2000.
9. M.P. Doyle, L.R: Beuchat, T.J. Montville: *Food Microbiology: Fundamentals and Frontiers*, 2nd ed., American Society of Microbiology, Washington, D.C. 2001.
10. R.M. Atlas, R. Bartha: "Hydrocarbon biodegradation and oil spill bioremediation" in K.C. Marshall (ed.): *Advances in Microbial Ecology*, vol. 12, Plenum, New York 1992.
11. F.H. Chapelle, "Bioremediation of Petroleum Hydrocarbon-Contaminated Ground Water: The Perspectives of History and Hydrology", *Ground Water* **37** (1999) no. 1, 122.
12. H.-J. Jördening, J. Winter: *Environmental Biotechnology*, Wiley-VCH, Weinheim 2005.
13. A.L. Demain, J.E. Davies: *Manual of Industrial Microbiology and Biotechnology*, ASM Press, Washington D.C. 1999.
14. M.C. Flickinger, S.W. Drew: *Encyclopedia of Bioprocess Technology*, John Wiley, New York 1999.
15. A. Liese, K. Selbach, C. Wandrey: *Industrial Biotransformations*, 2nd ed., Wiley-VCH, Weinheim 2006.
16. F.G. Priest, K. Tsubota, B. Austin: *Modern Bacterial Taxonomy*, Springer, New York 1993.
17. D.J. Woese, "Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya", *Microbiological Reviews* **51** (1987) 221.
18. G.E. Fox, A.K. Naik: "The Evolutionary History of the Ribosome" in L. Ribas de Poplana, P. Schimmel(eds.): *The Genetic Code and the Origin of Life*, Chapter 6, 92–105 Landes Bioscience, Austin, Texas 2004.
19. G.M. Garrity, J.A. Bell, T.G. Lilburn: "Taxonomic Outline of the Prokaryotes" in: *Bergey's Manual of Systematic Bacteriology*, 2nd ed., Release 5.0, Springer, New York 2004.
20. A. Balows et al. (eds.): *The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications*, 4 vols., Springer, New York, 1992.
21. K.O. Stetter: "The lesson of archaebacteria" in S. Bengtson(ed.): *Early Life on Earth. Nobel Symposium No. 84*, Columbia University Press, New York 1994, pp. 143–151.
22. R. Huber, H. Huber, K.O. Stetter, "Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties", *FEMS Microbiol. Rev.* **700** (2000) 1.
23. C. Bertoldo, C. Dock, G. Antranikian: "Thermoacidophilic Microorganisms and their Novel Biocatalysts", *Eng. Life Sci.* **4** (2004) no. 6, 521–532.
24. B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, P. Walter: *Molecular Biology of the Cell*, Garland Science, New York 1996.
25. C.J. Alexopoulos, W.W. Mims, M. Blackwell: *Introductory Mycology*, 4th ed., John Wiley & Sons, New York 1996.
26. K. Esser: *Cryptogams: Cyanobacteria, Algae, Fungi, Lichens*, Cambridge University Press, London 1982.
27. A.P. Rooney, T.J. Ward, "Evolution of a large ribosomal RNA multigene family in filamentous fungi: birth and death of a concerted evolution paradigm", *Proc. Natl. Acad. Sci. USA* **102** (2005) 5084.
28. Y. Van de Peer, R. De Wachter, "Evolutionary Relationships among the Eukaryotic Crown Taxa Taking into Account Site-to-Site Rate Variation in 18S rRNA", *J. Mol. Evol.* **45** (1997) no. 6, 619.
29. M. Roehr: "Products of Primary Metabolism", in H.-J. Rehm, H.-J.G. Reed(eds.): *Biotechnology*, 2nd ed., vol. 6, Wiley-VCH, Weinheim 1996.
30. A. Pühler, "Genetic Fundamentals and Genetic Engineering", in H.-J. Rehm, G. Reed(eds.): *Biotechnology*, 2nd ed., vol. 2, Wiley-VCH, Weinheim 1993.
31. G.M. Cooper: *The Cell: a Molecular Approach*, ASM Press, Washington D.C. 2000.
32. J.W. Lengeler, G. Drews, H. G. Schlegel: *Biology of the Prokaryotes*, Georg Thieme Verlag, Stuttgart 1999.
33. H. Sahm: "Biological Fundamentals", in H.-J. Rehm, G. Reed (eds.): *Biotechnology*, 2nd ed., vol. 1, Wiley-VCH, Weinheim 1993.

34. G. Gottschalk: *Bacterial Metabolism*, Springer, New York 1986.
35. D.G. Nichols: *Bioenergetics: An Introduction to the Chemiosmotic Theory*, Academic Press, London 1984.
36. E. Klipp, B. Nordlander, R. Kruger, P. Gennemark, S. Hohmann, "Integrative model of the response of yeast to osmotic shock", *Nat. Biotechnol*. **23** (2005) 975.
37. M. Koffas, G. Stephanopoulos, "Strain improvement by metabolic engineering: lysine production as a case study for systems biology", *Curr. Opin. Biotechnol.* **16** (2005) 361.
38. V.F. Wendisch, M. Bott, J. Kalinowski, M. Oldiges, W. Wiechert, "Emerging Corynebacterium glutamicum systems biology", *Biotechnol.* **124** (2006) 74.
39. N. Zamboni, N. Mouncey, H.P. Hohmann, U. Sauer, "Reducing maintenance metabolism by metabolic engineering of respiration improves riboflavin production by Bacillus subtilis", *Metab. Eng.* **5** (2003) 49.
40. V.J. Martin, D.J. Pitera, S.T. Withers, J.D. Newman, J.D. Keasling, "Engineering a mevalonate pathway in Escherichia coli for production of terpenoids", *Nat. Biotechnol.* **21** (2003) 796.
41. W.-D. Deckwer, D. Jahn, A.-P. Zeng, D.-C. Hempel, "Systembiotechnologische Ansätze zur Prozessentwicklung", *Chemie Ingenieur Technik* **78** (2006) 193.
42. D. Walz, H. Berg, G. Milazzo (eds.): *Bioelectrochemistry of Cells and Tissues*, Birkhäuser, Basel, Boston, Berlin, 1997.
43. P. Gräber, G. Milazzo (eds.): *Bioenergetics*, Birkhäuser, Basel, Boston, Berlin 1997.
44. D.L. Nelson, M.M. Cox: *Principles of Biochemistry*, Worth Publishers, New York 2002.
45. S.J. Singer, G.L. Nicolson, "The fluid mosaic model of the structure of cell membranes", *Science* **175** (1972) 720.
46. M.T. Madigan: J.M. Martinko: *Brock – Biology of Microorganisms*, 11th ed. (Internat.Ed.), Pearson, Prentice Hall, Upper Saddle River, NJ, 2006.
47. H.G. Schlegel: *General Microbiology*, Cambridge University Press, Cambridge 1993.
48. M. Gerhard: *Biochemical Pathways*, Wiley-VCH, Weinheim 1999.
49. R. Faurie: *Microbial Production of* L-*Amino Acids*, Springer, Berlin 2003 p. 20.
50. G. Lukacs, M. Ohno (eds.): *Recent Progress in the Chemical Synthesis of Antibiotics*, Springer, Berlin 1990.
51. A. Liese, K. Seelbach, C. Wandrey: *Industrial Biotransformations*, 2nd ed., Wiley-VCH, Weinheim 2006.
52. *OECD, The Application of Biotechnology to Industrial Sustainability*, Paris 2001.
53. O.P. Ward: *Fermentation Biotechnology: Principles, Processes and Products*, Open University Press, Milton Keynes 1989.
54. H.-J. Jördening, J. Winter: *Environmental Biotechnology*, Wiley-VCH, Weinheim 2005.
55. L.E. Erickson, D.Y.-C. Fung (eds.): *Handbook on anaerobic fermentations*, Marcel Dekker, New York 1988.
56. J. Helm: *Methanotrophe Bakterien als Produzenten von Poly-(β-hydroxybuttersäure) (PHB): Charakterisierung des Prozesses, des Polymers und einer stabilen Mischkultur*, UFZ-Umweltforschungszentrum Leipzig-Halle GmbH, Leipzig 2002.
57. I. Goldberg, J.S. Rokem (eds.): *Biology of methylotrophs*, Butterworth-Heinemann, Boston, Mass. 1991.
58. R.L.P. Adams, J.T. Knowler, D.P. Leader: *The Biochemistry of Nucleic Acids*, 11th ed., Chapman & Hall, London 1992.
59. J.E. Bailey, "Toward a science of metabolic engineering", *Science* **252** (1991) 1668.
60. D.C. Cameron, I.T. Tong, "Cellular and metabolic engineering", *Appl. Biochemistry and Biotechnology* **38** (1993) 105.
61. J. Nielsen, "Metabolic Engineering", *Appl. Microbiol. Biotechnol.* **55** (2001) 263.
62. S. Ostergaard, J. Nielsen, "Metabolic engineering of Saccharomyces cerevisiae", *Microbiol. Mol. Biol. Rev.* **64** (2000), 34.
63. E.M. Southern, U. Maskos, J.K. Elder, "Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models", *Genomics* **13**(1992) 1008.
64. M. Schena, D. Shalon, R.W. Davis, P.O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", *Science* **270** (1995) 467.
65. P.O. Brown, D. Botstein, "Exploring the new world of the genome with DNA microarrays", *Nat Genet.* **21** (1999) 33.
66. D.J. Duggan, M. Bittner, Y. Chen, P. Meltzer, J.M. Trent, "Expression profiling using cDNA microarrays", *Nat Genet.* **21** (1999) 10.

67. J. Khan, M.L. Bittner, Y. Chen, P.S. Meltzer, J.M. Trent, "DNA microarray technology: the anticipated impact on the study of human disease", *Biochim Biophys Acta.* **1423** (1999) M17.
68. S.P. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu, D. Solas, "Light-directed, spatially addressable parallel chemical synthesis", *Science* **251** (1991) 767–73.
69. G. MacBeath, "Protein microarrays and proteomics", *Nat Genet. 32 Suppl* (2002) 526.
70. M.F. Templin, D. Stoll, M. Schrenk, P.C. Traub, C.F. Vohringer, T.O. Joos, "Protein microarray technology", *Trends Biotechnol. 20* (2002) 160.
71. S.Y. Lee, S.J. Lee, H.-T. Jung, "Protein microarrays and chips", *J Ind. Eng. Chem.* **9** (2003) 9.
72. W. Kusnezow, J.D. Hoheisel, Antibody Microarrays, "Promises and Problems", *BioTechniques* **33** (2002) 14.
73. H. Zhu, M. Snyder, "Protein chip technology", *Current Opinion in Chemical Biology* **7** (2003) 55.
74. J. Nielsen: "Metabolic Turnover" in: *Encyclopedia of Life Sciences*, John Wiley & Sons, Inc., New York, USA 2002
75. H. Holms, "Flux analysis and control of the central metabolic pathways in Escherichia coli", *FEMS Microbiology Reviews* **19** (1996) 85.
76. G. Stephanopoulos, "Metabolic fluxes and metabolic engineering", *Metabolic Engineering* **1** (1999) 1.
77. J. Nielsen, "Metabolic Engineering", *Appl. Microbiol. Biotechnol.* **55** (2001) 263.
78. P. Novick: " Plasmids" in: *Encyclopedia of Life Sciences*, John Wiley & Sons, Inc., New York, USA 2002
79. A. Fire, S. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, C.C. Mello, "Potent and specific genetic interference by double stranded RNA in Caenorhabditis elegans", *Nature* **391** (1998) 806.
80. A. Grishok, et al., "Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C-elegans developmental timing", *Cell* **106** (2001) no. 1, 23.
81. A.J. Hamilton, D.C. Baulcombe, "A species of small antisense RNA in post transcriptional gene silencing in plants", *Science* **286** (1999) 950.
82. P. Nelson, M. Kiriakidou, A. Sharma, E. Maniataki, Z. Mourelatos, "The microRNA world: small is mighty", *Trends in Biochemical Sciences* **28** (2003) no. 10, 534.
83. P. Gerhardt (ed.): *Manual of Methods for General Microbiology*, Amer. Soc. for Microbiol., Washington, D.C. 1981.
84. J.R. Norris, D. W. Ribbons, (eds.): *Methods in Microbiology*, Academic Press, New York 1970.
85. P. Gerhardt (ed.): *Manual of Methods for General Microbiology*, Amer. Soc. for Microbiol., Washington, D.C. 1981.
86. H.W. Boyer, S. Nicosia: *Genetic Engineering*, Elsevier, Amsterdam-New York 1978
87. H. Chmiel: *Bioprozesstechnik*, Spektrum Verlag, Heidelberg 2006.
88. D. J. Bowen: "In vitro Mutagenesis" in: *Encyclopedia of Life Sciences*, John Wiley & Sons, Inc., New York 2005.
89. O. P. Ward: *Fermentation Biotechnology: Principles, Processes, and Products*, Prentice Hall, Englawood cliffs 1989.
90. R. D. Schmid: *Pocket Guide to Biotechnology and Genetic Engineering*; Wiley-VCH, Weinheim 2003.
91. A. Liese, K. Seelbach, C. Wandrey; *Industrial Biotransformations*, Wiley-VCH, Weinheim 2003.
92. K. Buchholz, V. Kasche, U. T. Bornscheuer: *Biocatalysts and Enzyme Technology*; Wiley-VCH, Weinheim 2005.
93. J.E. Bailey, D. F. Ollis: *Biochemical Engineering Fundamentals*, McGraw-Hill Book Company, New York 1977.
94. K. Schügerl, "Development of bioreaction engineering", *Adv. Biochem. Eng. Biotechnol.* **70** (2000) 41.
95. K.H. Lee, V. Hatzimanikati, "Biochemical Engineering", *Curr. Opin. Biotechnol.* **13** (2002) 85.
96. T. Scheper, B. Hitzmann, E. Stärk, R. Ulber, R. Faurie, P. Sosnitza, K. Reardon, "Bioanalytics: Detailed insight into bioprocesses", *Anal. Chim. Acta.* **400** (1999) 121–134.
97. A. Fiechter, "Physical and Chemical Parameters of Microbial Growth", *Adv. Biochem. Eng. Biotechnol.* **30** (1984) 7.
98. R. Greasham: "Media for Microbial Fermentations" in: *Biotechnology*, 2nd ed., vol. 3, VCH, Weinheim 1993.
99. K. Föbel, S. Metzger: "New Methods of Screening in Biotechnology" in: *Biotechnology*, 2nd ed., vol. 41, VCH-Wiley, Weinheim 2001.

100. M.P. Starr, H. Stolp, H. G. Trüper, A. Ballows, H. G. Schlegel, (eds.): *The Prokaryotes, A Handbook on Habitats, Isolation and Identification of Bacteria*, vol. I and II, Springer, Berlin 1981.
101. European Commission; Biotechnology 2020; Online in Internet: http://europa.eu.int/comm/research/rtdinfo/index_en.html, 2006.
102. http://www.bacterio.cict.fr/collections.html [2006].
103. ATCC, American Type Culture Collection, 12301 Parklawn Drive, Rockville, MD 20852, United States; http://www.lgcpromochem-atcc.com/ [2006].
104. DSMZ, German Collection of Microorganisms and Cell Cultures, Mascheroder Weg 1b, 38124 Braunschweig, Germany, http://www.dsmz.de/index.htm [2006].
105. CBS, Centraalbureau voor Schimmelcultures, Osterstraat 1, Baarn, Netherlands; http://www.cbs.knaw.nl/ [2006].
106. J.L. Ingraham, ø. Maaloe, F. C. Neidhardt: *Growth of the Bacterial Cell*, Sinauer Ass., Inc., Sunderland, Mass. 1983.
107. K. E. Aidoo, et al., *Adv. Appl. Microbiol.* **28** (1982) 201; I. Chibata, T. Tosa, *Adv. Appl. Microbiol.* **22** (1977) 1.
108. W.A. Rutala, D.J. Weber, *Emerg. Infect. Dis.* **7** (2001) 348.
109. J.R. Han, J. Xu, X.M. Zhou, *J. Basic Microbiol.* **42** (2002) no. 4, 254.
110. O.V. Singh, R.P. Singh, *J. Sci. Ind. Res.* **61** (2002) no. 5, 356–360.
111. A. Sakurai, H. Imai, M. Sakakibara, *J. Ferment. Bioeng.* **82** (1996) no. 5, 519.
112. K. Schuegerl, *Chem. Ing. Tech.* **52** (1980) 951.
113. S. Chamsartra, C.J. Hewitt, A.W. Nienow, *Biotechnol. Lett.* **27** (2005) no. 10, 693.
114. S. Mutafov, et al., *Biotechnol. Bioeng.* **84** (2003) no. 2, 160.
115. B. Jensen et al., *Biotechnol. Lett.* **24** (1) (2002) no. 1, 41.
116. H.-U. Blaser, E. Schmidt, (eds.): *Asymmetric Catalysis on Industrial Scale – Challenges, Approaches and Solutions*, Wiley-VCH, Weinheim 2003.
117. S.G. Burton, D.A. Cowan, J.M. Woodley, "The search for the ideal biocatalyst", *Nature Biotechnology* **20** (2002) 37.
118. D.E. Pszczola, "Products & Technologies", *Food Technology* **56** (2002) no. 2, 56.
119. A.M. Rouhi, "Chiral Chemistry," *Chem. Eng. News* 2004, 47.
120. A. Schmid, J.S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, *Nature* **409** (2001) 258.
121. A.J.J. Straathof, S. Panke, A. Schmid, *Curr. Opin. Biotech.* **13** (2002) no. 6, 548.
122. R.D. Schmid: *Taschenatlas der Biotechnologie*, Wiley-VCH, Weinheim 2002.
123. B. Biehl, K. Drauz, A. Karau, O. May: *Winnacker, Küchler, Chemische Technik*, Wiley-VCH, Weinheim 2004, p. 587–725.
124. H.-B. Rhein, S. Katzer, B. Hitzmann, D. Schnitzmeier, R. Ulber, *Umweltbundesamt: Ermittlung von Substitutionspotentialen von chemischen Verfahrenstechniken durch bio/gentechnische Verfahren zur Risikovorsorge, UBA-Text 29/02*, Berlin 2002.
125. *Enzyme Nomenclature 1984*, Academic Press, New York 1984.
126. L. Pasteur, *C. R. Hebd. Séances Acad. Sci.* **55** (1862) 28.
127. A.J. Brown, *J. Chem. Soc.* **49** (1886) 172.
128. L. Boutroux, *C. R. Hebd. Séances Acad. Sci.* **91** (1880) 236.
129. G. Bertrand, *C. R. Hebd. Séances Acad. Sci.* **122** (1896) 900.
130. R. H. A. Plimmer: *The Chemical Changes and Products Resulting from Fermentation*, Longmans, Green & Company, London 1903.
131. C. Neuberg, J. Hirsch, *Biochem. Z.* **115** (1921) 282.
132. C. Mamoli, A. Vercellone, *Ber. Dtsch. Chem. Ges.* **71** (1938) 1686.
133. D.H. Peterson, H.C. Murray, S.H. Eppstein, L.M. Reineke, A. Weintaub, P.D. Meister, H.M. Leigh, *J. Am. Chem. Soc.* **74** (1952) 5933.
134. G.M. Shull, D.A. Kita, *J. Am. Chem. Soc.* **77** (1955) 763.
135. A. Nobile, W. Charney, P.L. Perlman, H.L. Herzog, C.C. Payne, M.E. Tully, M.A. Jernik, E.B. Hershberg, *J. Am. Chem. Soc.* **77** (1955) 4184.
136. R.W. Thoma, J. Fried, S. Bonnano, P. Grabowich, *J. Am. Chem. Soc.* **79** (1957) 4818.
137. K. Arima, M. Nagasawa, M. Bae, G. Tamura, *Agric. Biol. Chem.* **33** (1969) 1636.
138. V. B. Urlacher; A. Makhsumkhanov, R.D. Schmid, *Appl Microbiol Biotechnol.* **70** (2006) 53.
139. P.J. Stabler, N. C. Bruce, *Appl Environ Microb* **64** (1998) 4106.
140. K.M. Madyastha, G. V.B. Reddy, *J Chem Soc Perkin Trans* **1** (1994) 911.
141. R.A. Sheldon, *Chemtech* **24** (1994) 38.

142. H. Felix, *Analytical Biochemistry* **120** (1982) 211.
143. S. Shimizu, H. Yamada, in H.J. Rehm, G. Reed, H. Pape (eds.): *Biotechnology*, vol. 4, VCH Verlagsgesellschaft, Weinheim 1985.
144. C. Schleg, S.G. Knight, *Mycologia* **54** (1962) 317.
145. C. Vezina, K. Singh in J.E. Smith, D. R. Berry (eds.): *The Filamentous Fungi*, vol. 1, Edward Arnold, London 1975 chap. 9.
146. T. Tosa, T. Mori, N. Fuse, I. Chibata, *Enzymologia* **31** (1966) 31.
147. T. Sato, Y. Nishida, T. Tosa, I. Chibata, *Biochim. Biophys. Acta* **570** (1979) 179.
148. I. Takata, K. Yamamoto, T. Tosa, I. Chibata, *Enzyme Microb. Technol.* **2** (1980) 30.
149. S. Takamatsu, I. Umemura, K. Yamamoto, T. Sato, T. Tosa, I. Chibata, *Eur. J. Appl. Microbiol. Biotechnol.* **15** (1982) 147.
150. D. Dinelli, *Process Biochem.* **7** (1972) no. 8, 9.
151. N.H. Aschengreen, *Process Biochem.* **10** (1975) no. 5, 17.
152. H.H. Weetall, N.B. Havewale, W.H. Pitcher, Jr., C.C. Deter, W.P. Vann, S. Yaverbaum, *Biotechnol. Bioeng.* **16** (1974) 295, 689.
153. A. Liese, K. Seelbach, C. Wandrey (eds.): *Industrial Biotransformations*, Wiley-VCH, Weinheim 2006.
154. M. Wada, J. Kato, I. Chibata, *Eur. J. Appl. Microbiol. Technol.* **8** (1979) 241.
155. M. Wada, J. Kato, I. Chibata, *Eur. J. Appl. Microbiol. Biotechnol.* **10** (1980) 275.
156. S. Ohlson, P.O. Larson, K. Mosbach, *Eur. J. Appl. Microbiol. Biotechnol.* **7** (1979) 103.
157. K. Buchholz, V. Kasche, U.T. Bornscheuer: *Biocatalysts and Enzyme Technology*, Wiley-VCH, Weinheim 2005.
158. C. Wandrey, R. Wichmann, W. Berke, M. Morr, M. R. Kula: *Third European Congress on Biotechnology*, vol. 1, Verlag Chemie, Weinheim 1984, p. 239.
159. G. M. Whitesides in R. Porter, S. Clark, (eds.): *Enzymes in Organic Synthesis*, Pitman, London 1985 76.
160. S. Shimizu, H. Yamada, *Trends Biotechnol.* **2** (1984) 137.
161. S. Shimizu, Y. Tani, K. Ogata in D. B. McCormick, L. D. Wright: *Methods in Enzymology*, vol. 62, Academic Press, New York 1979 p. 236.
162. S. Shimizu, A. Esumi, R. Komaki, H. Yamada, *Appl. Environ. Microbiol.* **48** (1984) 1118.
163. A. Zaks, and A.M. Klibanov, *Proc. Natl. Acad. Sci. U.S.A* **82** (1985) 3192.
164. A.M. Klibanov, *Chemtech* **16** (1986) 354.
165. M.T. Ru, K.C. Wu, J.P. Lindsay, J.S. Dordick, J.A. Reimer, D.S. Clark, *Biotechnol. Bioeng.* **75** (2001) 187.
166. K. Takahashi, H. Nishimura, T. Yoshimoto, Y. Saito, Y. Inada, *Biochem. Biophys. Res. Commun.* **121** (1984) 261.
167. *Bioprocessing Technology*, February 1985, p. 6.
168. A. Liese, S. Lütz: "Nonimmobilized Biocatalysts in Industrial Fine Chemical Synthesis", *Ullmann's Encyclopedia of Industrial Chemistry, Electronic Release*, 7th ed., Wiley-VCH, Weinheim 2004.
169. H. Yamada, H. Kumagai in D. Perlman (ed.): *Advances in Applied Microbiology*, vol. 19, Academic Press, New York 1975, p. 249.
170. H. Yamada, H. Kumagai, *Pure Appl. Chem.* **50** (1978) 1117.
171. H. Yamada, S. Shimizu, in J. Tramper (ed.): *Proc. Int. Symp. on Biocatalysts in Organic Synthesis*, Elsevier Science Publ., Amsterdam 1985.
172. T. Fukumura, *Agric. Biol. Chem.* **40** (1976) 1687; **41** (1977) 1327.
173. E. Kondo, E. Masuo, *J. Gen. Appl. Microbiol.* **7** (1961) 113.
174. G. Schmitz, D. Franke, S. Stevens, R. Takors, D. Weuster-Botz, C. Wandrey, *J. of Molecular Catalysis B-Enzymatic* **10** (2000) 313.
175. B. Zhou, A.S. Gopalan, F. vanMiddlesworth, W.-R. Shieh, C.J. Sih, *J. Am. Chem. Soc.* **105** (1983) 5925.
176. A.M. Klibanov, *Nature* **409** (2001) 241.
177. K. Wurges, P.H. Pfromm,M.E. Rezac, P. Czermak, *J. of Molecular Catalysis B-Enzymatic* **34** (2005) 18.
178. A. Zaks and A.M. Klibanov, *Science* **224** (1984) 1249.
179. A. Zaks and A.M. Klibanov, *J. Am. Chem. Soc.* **108** (1986) 2767.
180. M. Quiros, V.M. Sanchez, R. Brieva, F. Rebolledo, V. Gotor, *Tetrahedron: Asymm.* **4** (1993) 1105.
181. J.S. Dordick, *Enzyme Microb. Technol.* **11** (1989) 194.
182. D.H. Altreuter and D.S. Clark, *Curr. Opin. Biotechnol.* **10** (1999) 130
183. K. Kitahara, S. Fukui, M. Misawa, *Nippon Nogei Kagaku Kaishi* **34** (1960) 44.
184. T. Sato, Y. Nishida, T. Tosa, I. Chibata, *Biochim. Biophys. Acta* **570** (1979) 179.
185. K. Sano, K. Yokozeki, F. Tamura, N. Yoshida, I. Noda, K. Mitsugi, *Appl. Environ. Microbiol.* **34** (1977) 806.

186. K. Sano, K. Mitsugi, *Agric. Biol. Chem.* **42** (1978) 2315.
187. H. Yamada, H. Kumagai, H. Enei, H. Nakazawa, K. Mitsugi in G.B. Brown, G. Manecke, L.B. Wingard, Jr. (eds.): *Enzyme Engineering*, vol. 4, Plenum Publishing, New York 1978, p. 25.
188. T. Utagawa, H. Morisawa, T. Nakamatsu, A. Yamazaki, S. Yamanaka, *FEBS Lett.* **119** (1980) 101.
189. M. Nagasawa, M. Bae, G. Tamura, K. Arima, *Agric. Biol. Chem.* **33** (1969) 1644.
190. C.K.A. Martin, F. Wagner, *Eur. J. Appl. Microbiol.* **2** (1976) 243.
191. H. Mayer, J. Collins, F. Wagner in H.H. Weetall, G. P. Royer (eds.): *Enzyme Engineering*, vol. 5, Plenum Publishing, New York 1980, p. 61.
192. T. Drepper et al., *Chem. Ing. Technik* **78** (2006) 239.
193. P. A. Belter et al.: *Bioseparations*, John Wiley & Sons, New York 1988.
194. R. Hatti-Kaul, B. Mattiasson B: *Isolation and purifaction of proteins*, Marcel Dekker Inc., New York 2003.l
195. J.A. Asenjo: *Separation processes in Biotechnology*, Marcel Dekker, New York 1990.
196. E. Goldberg: *Downstream processing*, Chapman & Hall, London 1997.
197. A.P.J. Middelberg, *Biotechnol. Adv.* **13** (1995) 491.
198. G. Stephanopoulos: *Biotechnology*, VCH, New York 1993.
199. V. Ganeva, B. Galutzov, J. Teissie, *Anal. Biochem.* **315** (2003) 77.
200. M.J. Lewis: "Pressure-activated membrane processes", in A.S. Grandison, M.J. Lewis (eds.): *Separation Processes in the Food and Biotechnology Industries*, Woodhead Publishing, Cambridge 1996, p 65.
201. R. Ulber, K. Plate, O.W. Reif, D. Melzner: "Membranes for Protein Isolation and Purification", in R. Hatti-Kaul, B. Mattiasson, (eds.): *Isolation and purifaction of proteins*, Marcel Dekker Inc., New York 2003, p 191.
202. E. Huthmann, A. Juza, *J. Chromatogr. A* **1092** (2005) no. 1, 24.
203. F. Hilbrig, R. Freitag, *J. Chromatography B* **790** (2003) no. 1–2, 1.
204. R.P.W. Scott, Preparative Chromatography. Online in Internet: http://www.chromatography-online.org/Preparative/Introduction/rs1.html [2006].
205. R. Ulber et al., *Acta Biotechnol.* **21** (2001) 27.
206. W. Demmer, D. Nussbaumer, *J. Chromatogr. A* **852** (1999) 73.
207. D.K. Roper, E.N. Lightfoot, *J. Chromatogr. A* **702** (1995) 3.
208. R. Gosh, *J. Chromatogr. A* **952** (2002) 13.
209. H. Zou, Q. Luo, D. Zhou, *J. Biochem. Biophys. Methods* **49** (2001) 199.
210. K. Schügerl: *Solvent extraction in biotechnology*, Springer-Verlag, Berlin 1994.
211. K. Schügerl, *Biotechnology Advances* **18** (2001) 581.
212. E. D. Ramsey: *Analytical Supercritical Fluid Extraction Techniques*, Kluwer Academic Publishers, Dordrecht 1998.
213. L. Boillereax, J.M. Flaus, "A new approach for designing model-based indirect sensors" *IEEE Trans. Control Systems Tech.* **8** (2000) 601.
214. J.M. Flaus, L. Boillereaux, "Moving horizont state estimation for a bioprocess modeled by a neural network", *Trans. Inst. Measurement. Control.* **19** (1997) 263.
215. J. Giese, "Online sensors for food processing", *J. Food Technol.* **47** (1993) 88.
216. Y. H. Zhu et al., "Application of neural networks in lysine fermentation control", *Biochem. Eng. J.* **62** (1996) 207.
217. R. Simutis, "Fuzzy-Aided neural network for real-time state estimation and process prediction in alcohol formation step of production-scale beer brewing", *J. Biotechnol.* **27** (1993) 203.
218. M.N. Karim, "Global and local neural network models in Biotechnology: Application to different cultivation processes", *J. Ferment. Bioeng.* **33** (1997) 1.
219. J.F.M. Impe, "Power and limitations of model based bioprocess optimization", *Mathematics and Computers in Simulation* **42** (1996) 159.
220. P. Saha et al., "Multi-Input multi-output control a continuous fermenter using nonlinear model based controllers", *Bioprocess Engineering* **21** (1999) 533.
221. J. A. Roels: *Energetics and kinetics in biotechnology*, Elsevier Biomedical Press, Amsterdam 1983.
222. J.E. Bailey, "Mathematical Modeling and analysis in Biochemical Engineering: Past Accomplishments and Future Opportunities", *Biotechnol. Prog.* **14** (1998) 8.
223. J.J. Heijnen, "Thermodynamics ov microbial growth and ist implications for process design", *TibTech* **12** (1994), 483.

224. G. R. Gavalas: *Nonlinear Differential Equations of chemically Reaction Systems*, Springer, Berlin 1968.
225. R. Plonsey: "Volume Conductor Theory", in J.D. Bronzino (ed): *The Biomedical Engineering Handbook*, CRC Press, Boca Raton 1995, pp. 119–125.
226. C.C. Lin, L. A. Segel: *Mathematics applied to deterministic problems in the natural science*, SIAM, Philadelphia 1988.
227. D. A. Anderson, et al.: *Computional fluid mechanics and heat transfer*, Hemisphere Publishing Corporation, New York 1984.
228. H. Nirschl, et al., "Three Dimensional Calculations of the Simple Shear Flow around a Single Particle between two Moving Walls", *J. of Fluid Mech.* **283** (1995) 273.
229. H. Nirschl: *Partikelbewegung in Scherströmungen ohne und mit Berücksichtigung des Einflusses angrenzender Wände*, VDI-Verlag Düsseldorf 1997.
230. F.M. Williams, "A Model for cell growth dynamics", *J. Theor. Biol.* **15** (1967), 190.
231. D. Ramkrishna: "A cybernetic perspective of microbial growth", in E. Papoutsakis, G.N. Stephanopoulus, H.W. Blanch, (eds): *Foundations of biochemical engineering kinetics and thermodynamics in biological sytems*, American Chemical Society, Washington (1982).
232. G. Bastin,; D. Dochain, : *Online Estimation and adaptive control of bioreactors*, Elsevier Science Publisher, Amsterdam 1990.
233. J. Nielson,; J. Villadsen,: Bioreactors: "Description and modeling", in: *Biotechnology 3*. G. Stephanopolous; Eds: VCH-Verlag Weinheim 1993.
234. K. Schügerl K.H. Bellgardt: *Bioreaction Engineering: Modeling and Control*, Springer, Berlin 2000.
235. M. Gavrilescu, M. Macoveanu, "Process Engineering in Biological Aerobic Waste-Water Treatment", *Acta Biotechnol.* **19** (1999), 111.
236. I.J. Dunn; E. Heinzle; J. Ingham; J.E. Prenosil: *Biological Reaction engineering*, VCH Verlagsgesellschaft, Weinheim 1992.
237. D.M. Trujillo; H. R. Busby: *Practical Inverse Analysis in Engineering*, CRC Press, Boca Raton 1997.
238. E. Hensel: *Inverse Theory and Applications for Engineers*, Prentice-Hall, New Jersey 1991.
239. B. Sonnleitner: in U. Mortensen, H.J. Noorman, (eds): "Bioreactor performance", *Biotechnology Research foundation Lund* 1993,p. 143.
240. A. Harder, J.A. Roels, "Application of simple structured models in bioengineering", *Adv. Biochem. Eng.* **21** (1982), 56.
241. J.J. Heijnen, A. H. Terwisscha van Scheltinga, A.J. Straathof, "Fundamental bottlenecks in the application of continuous bioprocesses", *J. Biotechnol.* **22** (1992), 3.
242. J.P. Axelsson, T. Münch, B. Sonnleitner: in N. Karim (ed): *Computer applications in fermentation technology – Modeling and control of bioprocesses*, Pergamon (1992), 383–386.
243. B. Sonnleitner, U. Hahnemannn, "Dynamics of the respiratory bottleneck of Saccharomyces cerevisiae", *J. Biotechnol.* **38** (1994), 63.
244. J. Nielsen, K. Nikolajsen, J. Villadsen, "Structured modeling of a microbial system: I A theoretical study of lactic acid fermentation", *Biotech. Bioeng.* **38** (1991), 1.
245. B. Sonnleitner, S.A. Rothen, H. Kuriyama, "Dynamics of glucose consumption in yeast", *Prog. Biotechnol.* **13** (1997), 8.
246. W. de Koning, K. van Dam, "A method for the Determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH", *Anal. Biochem.* **204** (1992), 118.
247. K. van Dam, J. Vandervlag, B.N. Kholodenko, H.V. Westerhoff, "The sum of the control coefficients of all enzymes on the flux through a group-transfer pathway can be as high as two", *Eur. J. Biochem.* **212** (1993), 791.
248. M. Rizzi, U. Theobald, M. Baltes, M. Reuss: "Measurements and modeling of the dynamic glucose response of Saccharomyces cerevisiae in the time window of mixing times", in A.W. Nienow, (ed.): *3rd international conference on bioreactor and bioprocess fluid dynamics*. Mechanical Engineering Publishers London, 1993, pp.401–412.
249. U. Theobald, W. Mailinger, M. Reuss, M. Rizzi, "In vivo analysis of glucose-induced fast changes in yeast adenine nucleotide pool applying a rapid sampling technique", *Anal. Biochem.* **214** (1993), 31.
250. B. Sonnleitner, "Dynamic adaptation of microbes", *J. Biotech.* **65** (1998), 47.
251. P. Neubauer, L. Haggstrom, S.O. Enfors, "Influence of Substrate Oscillations on Acetate Formation and Growth Yield in Escherichia coli Glucose Limited Fed-Batch Cultivations", *Biotechnol. Bioeng.* **47** (1995) 139.

252. S. George, G. Larsson, S.O. Enfors: "Metabolic response of Saccharomyces cerevisiae in scale-down two- compartment reactor with controlled substrate oscillations", in L. Alberghina, L. Frontali, P. Sensi, (eds.): *ECB6: Proceedings of the 6th European Congress on Biotechnology*, Vol. 9 Part II, 1994, pp. 883–890:
253. A. Fiechter, B. Sonnleitner, "Non-invasive concepts in metabolic studies", *Adv. Microb. Physiol.* **36** (1994) 145.
254. S. Linko, P. Linko, "Developments in Monitoring and control of food processes", *Trans IChemE* **76** (1998) 127.
255. T.A. Haley, S. J. Mulvaney, "Advanced process control techniques for the food industry" *Trends Food Sci. Technol.* **6** (1994) 103.
256. S. Shioya, K. Shimizu, T. Yoshida, "Knowledge-based design and operation of bioprocess systems" *J. of Bioscience and Bioeng.* **87** (1999) 261.
257. J. R. Perez-Correa, C.A. Zaror, "Recent advances in process control and their potential applications to food processing", *J. Food Control.* **4** (1993) 202.
258. B. Sonnleitner: "Instrumentation of biotechnological processes", in T.Scheper (ed.): *Advances in biochemical engineering biotechnology*, Springer, Berlin 2000, pp. 3–54.
259. P. Ries: "Entwicklung einer Meßsystem-Konzeption zur dynamischen Messung der Trocken-Zelldichte von aufbereiteten Bakterien-Suspensionen auf der Grundlage der Streulichtmessung", Dissertation, ETH Zürich 1983.
260. P.C. Hauser, T.W.T. Rupansinghe, "A multi-wavelength photometer based on light-emitting diodes", *Talanta* **42** (1995) 605.
261. E. Nestaas, D.I.C. Wang, "A new sensor – the Filtration Probe" – for quantitative charakterization of penicillin fermentation. III. An automatically operating probe", *Biotechnol. Bioeng.* **25** (1983) 1981.
262. D.C. Thomas, V.K. Chittur, J.W. Cagney, H.C. Lim, "Online estimation of mycelial cell mass concentrations with a computer-interfaced filtration probe", *Biotechnol. Bioeng.* **27** (1985) 729.
263. G.H. Markx, D.B. Kell, "The use of dielectric permittivity for the control of the biomass level during biotransformations of toxic substrates in continuous culture", *Biotechnol. Prog.* **11** (1995) 64.
264. K. Asami, T. Yonezawa, "Dielectric analysis of yeast cell growth", *BBA Gen Subjects* **1245** (1995) 99.
265. L. C.M. Auberson, T. Kanbier, U. Stockar, "Monitoring synchronized oscillating yeast cultures by calorimetry", *J. Biotech.* **29** (1993) 205.
266. M. Rieseberg, C. Kasper, K.F. Reardon, T. Scheper, "Flow cytometry in biotechnology", *Appl. Microbiol. Biotechnol.* **56** (2001) 350.
267. J. Mukherjee, C. Lindemann, T. Scheper, "Fluorescence monitoring during cultivation of Enterobacter aerogenes at different oxygen levels", *Applied Microbiology and Biotechnology* **52** (1999) 489.
268. L.K. Ju, X. Yang, J.F. Lee, W.B. Armiger, "Monitoring of the Biological Nutrient Removal process by an online NADPH Fluorometer", *Biotechnol. Progress* **11** (1995) 545.
269. D. Kell, B. Sonnleitner, "GMP – Good modeling practise: an essential component of good manufacturing practise", TIBTECH **13** (1995) 481.
270. J. Maddox, "Towards more measurement in biology", *Nature* **368** (1994) 95.
271. K. Steube, U. Spohn, "Online monitoring of intracellular enzyme activities with flow injection analysis", *Anal. Chim. Acta* **284** (1994) 235.
272. I. Ogbomo, A. Steffl, W. Schuhmann, U. Prinzing, H.-L. Schmidt, "Online determination of ethanol in bioprocesses based on sample extraction by continuous pervaporation", *J. Biotech.* **31** (1993) 317.
273. N.C. van de Merbel, H. Lingemann, U. Brinkmann, "Sampling and analytical strategies in online bioprocess monitoring and control", *J. Chromatography* **725** (1996) 13.
274. J.M. Hilmer, T. Scheper, "A new version of an in-situ sampling system for bioprocess analysis", *Acta Biotechnol.* **16** (1996) 185.
275. L. Olsson, U. Schulz, J. Nielsen, "Online bioprocess monitoring – an academic discipline or an industrial tool", *Trends. Anal. Chem.* **17** (1998) 88.
276. G. Locher, B. Sonnleitner, A. Fiechter, "Online measurement in biotechnology: exploitation, objectives and benefits", *Journal of Biotechnology* **25** (1993) 55.
277. R.G. Brown; P.Y.C. Hwang: *Introduction to random signals and applied kalman filtering*, 3rd ed., John Wiley & Sons, New York 1997), pp. 260–265.

278. P.V. Zumbusch, T. Meyerjens, B. Brunner, H.Märkl, "Online monitoring of organic substances with high-pressure liquid chromatography (HPLC) during the anaerobic fermentation of waste-water", *Appl. Microbio. Biotechnol.* **42** (1994) 140.
279. C.S. Effenhauser, A. Manz, H.M. Widmer, "Glass chips for high-speed capillary electrophoresis separations with submicrometer plate heights", *Analytical Chemistry* **65** (1993) 2637.
280. N.C. van de Merbel, P. Zuur, M. Frijlink, J. J. M. Holthuis, H. Lingeman, U.A.T. Brinkman:, "Automated monitoring of biotechnological processes using online ultrafiltration and column liquid chromatography", *Anal. Chim. Acta* **303** (1993) 39.
281. H. Kurokawa, Y.S. Park, S. Iijima, T. Kobayashi, "Growth characteristics in fed-batch culture of hybridoma cells with control of glucose and glutamine concentrations", *Biotechnology and Bioengineering* **44** (1994) 95.
282. C. Zwiener, F.H. Frimmel, "Oxidative treatment of pharmaceuticals in water", *Water Research* **34** (2000) 1881.
283. T. Meyer-Jens, G. Matz, H. Märkl, "On-line measurement of dissolved and gaseous hydrogen sulphide in anaerobic biogas reactors", *App. Microbiol. Biotechnol.* **43** (1995) 341.
284. E. Heinzle, "Present and potential applications of mass spectrometry for bioprocess research and control" *J. Biotechnol.* **25** (1992) 81.
285. H.-L. Schmidt, W. Schuhmann, F.W. Scheller, F. Schubert,: "Specific features of Biosensors",in W. Göpel, et al. (ed.), *Sensors: a comprehensive survey, Chemical and Biochemical Sensors*,: VCH Verlagsgesellschaft, Weinheim 17 1992, pp. 717–818.
286. S. Zhang, G. Wright, Y. Yang, "Materials and techniques for electrochemical biosensor design and construction", *Biosensors and Bioelectronics* **15** (2000) 273.
287. D.G. Myszka, "Survey of the 1998 optical biosensor literature", *Journal of Molecular Recognition* **12** (1999) 390.
288. Q. Chen, B. Kersten, J. Li, C. Wu, P. Jagasia, A. Ewing, "A multichannel automated amperometric test system for glucose monitoring biosensor quality control testing", *Electroanalysis* **13** (2001) 61.
289. I. Eshkenazi, E. Maltz, B. Zion, J. Rishpon, "A three-cascaded-enzymes biosensor to determine lactose concentration in raw milk", *Journal of Dairy Science* **83** (2000) 939.
290. J. Perdomo, H. Hinkers, C. Sundermeier, W. Seifert, O.M. Morell, M. Knoll, "Miniaturized real-time monitoring system for L-lactate and glucose using microfabricated multi-enzyme sensors", *Biosensors and Bioelectronics* **15** (2000) Nov., 515.
291. B. Serra, A.J. Reviejo, C. Parrado, J.M. Pingarron, "Graphite-Teflon composite bienzyme electrodes for the determination of L-lactate: Application to food samples", *Biosensors and Bioelectronics* **14** (1999) 505.
292. M.D. Luque de Castro, G.J.L. Luque:, "Biosensors in wine production monitoring", *Analytical Letters* **33** (2000) 963.
293. M.J.F. Rebelo:, D. Compagnone, G.G. Guilbault, G.J. Lubrano, "Alcohol electrodes in Beverage Measurements", *Anal. Letters* **27** (1994) 3027.
294. C.S. Rui, K. Sonomoto, H.I. Ogawa, Y. Kato, "A flow-injection biosensor system for the amperometric determination of creatinine: Simultaneous compensation of endogeneous interferents", *Anal. Biochem.* **210** (1993) 163.
295. E. Mallat, C. Barzen, R. Abuknesha, G. Gauglitz, D. Barcelo, "Fast determination of paraquat residues in water by an optical immunosensor and validation using capillary electrophoresis-ultraviolet detection", *Analitica Chimica Acta* **427** (2001) 165.
296. J.J. Xu, H.Y. Chen, "Amperometric glucose sensor based on glucose oxidase immobilized in electrochemically generated poly(ethacridine)", *Analytica Chimica Acta* **423** (2000) 101.
297. D.L. Massart, B.G.M. Vandeginste:, S.N. Deming, Y. Michotte, L. Kaufman: "Chemometrics: a textbook", in B.G.M. Vandeginste:, L. Kaufman, (ed.), *Data Handling in science and technololgy*, vol. 2, Elsevier, Amsterdam 1988.
298. J. Ruzicka; E.H. Hansen: *Flow Injection Analysis*, 2nd edn., John Wiley & Sons, New York, 1988.
299. G. Bastin; D. Dochain: *Online Estimation and Adaptive Control of Bioreactors*, Elsevier, Amsterdam 1990.
300. A. Cheruy, "Software Sensors in bioprocess engineering" *J. Biotechnol.* **52** (1997) 193.
301. S. Vassileva, B. Tzvetkova, C. Katranoushkova, L. Losseva, "Neuro-fuzzy prediction of uricase procduction", *Bioprocess Eng.* **22** (2000) 363.

302. G. Liebsch, I. Klimant, B. Frank, G. Holst, O. S. Wolfbeis, "Luminescence lifetime imaging of oxygen, pH, and carbon dioxide distribution using optical sensors", *Applied Spectroscopy* **54** (2000) 548.
303. F. Shiraishi, "Apparent kinetic parameters of an immobilized enzyme reaction: What is expected from oversimplification?", *Enzyme Microb. Technol.* **16** (1994) 349.
304. Y.K. Jung, W. Hur, "A new method of online measurement of buffer capacity and alkali consumption rate of a fermentation process", *Journal of Bioscience and Bioengineering* **90** (2000) 580.
305. O. Bernard, S.Z. Hadj, D. Dochain, "Software sensors to monitor the dynamics of microbial communities: Application to anaerobic digestion", *Acta Biotheoretica* **48** (2000) 197.
306. V. Acha, M. Meurens, H. Naveau, D. Dochain, G. Bastin, S.N. Agathos, "Model-based estimation of an anaerobic reductive dechlorination process via an attenuated total reflection-Fourier transform infrared sensor", *Water-Science-and-Technology* **40** (1999) 33.
307. T. Yano, M. Harata, "Prediction of the concentration of several constituents in a mouse-mouse hybridoma by near infrared spectroscopy", *J. Ferment. Bioeng.* **77** (1994) 659.
308. R. Schindler, H.L. Thanh, B. Lendl, R. Kellner, "Determination of enzyme kinetics and chemometric evaluation of reaction products by FTIR spectroscopy on the example of beta-fructofuranosidase", *Vibrational Spectroscopy* **16** (1998) 127.
309. H. Hoyer, "NIR online analysis in the food industry", *Process Control and Quality* **9** (1997) 143.
310. G. Downey, "Authentication of food and food ingredients by near infrared spectroscopy", *J. Near Infrared Spectroscopy* **4** (1996) 47.
311. T.K. Hamrita, S. Wang, "Pattern Recognition for Modeling and online diagnosis of Bioprocesses", *IEEE Trans. Industr. Applications* **36** (2000) 1295.
312. T.T. Bachmann, B. Leca, F. Vilatte, J.L. Marty, D. Fournier, R. Schmid, "Improved multianalyte detection of organophosphates and carbamates with disposable multielectrode biosensors using recombinant mutants of Drosophila acetylcholinesterase and artificial neural networks", *Biosensors and Bioelectronics* **15** (2000) 193.
313. R.M.de Mol, W.E. Woldt, "Application of fuzzy logic in automated cow status monitoring", *Journal of Dairy Science* **84** (2001) 400.
314. Z. Shi, K. Shimizu, "Neuro-Fuzzy control of bioreactor systems with pattern recognition", *J. Ferment. Bioeng.* **74** (1992) 39.
315. G. Stephanopoulos, G. Locher, M. Duff: "Pattern recognition methods for fermentation database mining", *Reprints of the 6th Conference on Computer Applications in Biotechnology 1995*, 195–198.
316. J. Bradley, A.P.F. Turner:, R.D. Schmidt: "An in situ fermenter probe for bakers yeast propagation monitoring", in GBF Monographs 13, 1989, pp. 85–88.
317. E. Heinzle, A. Oeggerli, B. Dettwiler, "Online fermentation gas analysis: Error analysis and application of mass spectrometry" *Anal. Chim. Acta* **238** (1990) 101.
318. G. Locher, U. Hahnemann, B. Sonnleitner, A. Fiechter, "Automatic bioprocess control: 5. Biologically and technically caused effects during cultivation", *Journal-of-Biotechnology* **29** (1993) 75.
319. M. Dehghani, M. Bulmer, M.-E. Gregory, N.-F. Thornhill, "Measurement variability analysis for fermentation process assays", *Bioprocess-Engineering* **13** (1995) 239.
320. B. Sonnleitner, "New concepts for quantitative bioprocess research and development", *Advances in Biochemical Engineering, Biotechnology; Metabolic engineering* **54** (1996) 155.
321. W.McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Mathematical Biophysics* **5** (1943) 115.
322. D. Rumelhart, J.McClelland: *Parallel distributed Processing*, MIT Press, Massachussetts 1986.
323. D. Rumelhart, G.E. Hinton, R.J. Williams, "Learning representations by back-propagating errors", *Nature* **323** (1988), 533.
324. L.A. Zadeh, "Fuzzy Sets", *Information and Control* **8** (1965) 338.
325. J.H. Holland,: *Adaption in Natural and Artificial Systems*, 1. and 2. ed., MIT Press, Boston 1975, 1992.
326. A. Bardossy, L. Duckstein,: *Fuzzy rule-based modeling with applications to geophysical, biological and engineering systems*, CRC Press, Boca Raton 1995.
327. H. Kiendl,: *Fuzzy control – methodenorientiert*, Oldenbourg Verlag, München 1997.

328. B. Kosko: *Neural Networks and Fuzzy Systems – A dynamical systems approach to machine intelligence*, Prentice Hall International, London 1992.
329. A. Zell: *Simulation neuronaler Netze*, Addison-Wesley Publishing Company, Bonn 1994.
330. C.M. Bishop: *Neural Networks for pattern Recognition*, Clarendon Press, Oxford 1995.
331. D. Dasgupta, Z. Michalewicz,: *Evolutionary Algorithms in Engineering Applications*, Springer Verlag, Berlin 1997.
332. W. Banzhaf, P. Nordin, R.E. Keller, F.D. Francone,: *Genetic Programming: An Introduction: On the Automatic Evolution of Computer Programs and Its Applications*, Morgan Kaufmann Publisher, Los Angeles 1998.
333. T. Eerikäinen, P. Linko, S. Linko, T. Siimes, Y.-H. Zhu, "Fuzzy logic and neural network applications in food science and technology", *Trends in Food Science & Technology* **4** (1993) 237.
334. S. Shioya, K. Shimizu, T. Yoshida, "Knowledge-based design and operation of bioprocess systems", *J. Bioscience and Bioeng.* **87** (1999) 261.
335. B. Kanizsa, "Subjective Contours", *Scientific American* **234** (1976) 48.
336. H.J. Zimmermann: *Fuzzy Technologien – Prinzipien, Werkzeuge, Potentiale*, VDI-Verlag, Düsseldorf 1993.
337. A.E. Imrani, A. Bouroumi, Z.E.A. Abidime:, A. Essaid, "A fuzzy clustering-based niching approach to multimodal function optimiziation", *J. Cog. Systems Res.* **1** (2000) 119.
338. S.E. Fahlman, C. Lebiere: "The cascade-correlation learning architecture", in D.S. Touretzky, *Advances in Neural information processing systems*, Morgan Kaufmann Publisher, Los Angeles 1990, pp. 524–532.
339. R.O. Duda, P.E. Hart: *Pattern classification and scene analysis*, John Wiley, New York 1973.
340. L. Pickert, F. Kalwonn, E. Wingender, "Transcription regulatory region analysis using signal detection and fuzzy clustering", *Bioinformatics* **14** (1998) 244.
341. J.J. Buckley, Y. Hayashi, E. Czogala, "On the equivalenc of neural nets and fuzzy expert systems", *Fuzzy Sets and Systems* **53** (1993) 129.
342. D.V. Lindley, "The probability approach to the treatment of uncertainty in artificial intelligence and expert system", *Statistical science* **2** (1987) 17.
343. L.A. Zadeh, "Fuzzy Sets as a basis for a theory of possibilities", *Fuzzy sets and Systems* **1** (1978) 3.
344. D. Dubois, H. Prade, "Properties of measures of information in evidence and possibility theories", *Fuzzy sets and Systems* **24** (1987) 162.
345. L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning , part 1 to 3", *Information Science* **8** (1975) 199 199, 8 (1975) 301, 9 (1976) 43.
346. C. Lee, "Fuzzy logic in control system: Fuzzy logic controller-part i", *IEEE Transactions on Sytems, Man, and Cybernetics* **20** (1990) 404.
347. I. Turksen, H. Tanaka, J. Watade, "Special issue on expert decision-support systems – foreword", *Fuzzy Sets and Systems* **58** (1993) 1.
348. B. Kosko: *Fuzzy Thinking. The New Science of Fuzzy Logic*, Hyperion Inc., New York 1993.
349. R. Yager, D. Filev: *Essentials of Fuzzy Modeling and Control*, John Wiley & Sons, Inc., New York 1994.
350. R. Kruse, J. Cebhardt, F. Klawonn: *Foundations of Fuzzy Systems*, John Wiley & Sons, Inc., New York 1994.
351. I. Turksen, "Measurement of membership functions and their acquisition", *Fuzzy Sets and Systems* **40** (1991) 5.
352. Y. Yoshinari, W. Pedrycz, K. Hirota, "Construction of fuzzy models through clustering-techniques", *Fuzzy Sets and Systems* **78** (1996) 1.
353. M. Yang, "A survey of fuzzy clustering", *Mathematical and Computer Modeling* **18** (1993) 1.
354. H. Takagi, L. Hayashi, "NN-driven fuzzy reasoning", *International Journal of Approximate Reasoning* **5** (1991) 191.
355. T. Williams, "New tools make fuzzy/neural more, than an academic amusement", *Computer Design* **14** (1994) 69.
356. C. Karr, E. Gentry, "Fuzzy control of ph using genetic algorithms", *IEEE Transactions on Fuzzy Systems* **1** (1993) 46.
357. D. Park, A. Kandel, G. Langholz, "Genetic-based new fuzz reasoning models with application to fuzzy control", *IEEE Transactions on systems, man and cybernetics* **24** (1994) 39.

358. T. Procyk, E. Mamdani, "A linguistic self-organizing process controller", *Automatica* **15** (1979) 15.
359. R. Tansheit, E. Scharf, "Experiments with the use of a rule-based self-organizing controller for robotics applications", *Fuzzy Sets and Systems* **26** (1988) 195.
360. L. Journaler, D. Goldberg, J. Hollanid, "Classifier systems and genetic algorithms", *Artificial Intelligence* **40** (1989) 235.
361. T. Goldberg: *Genetic Algorithms in Search, Optimizations and Machine Learning*, Addison Wesley, Menlo Park 1989.
362. D. Park, A. Kandel, G. Langholz, "Genetic based new fuzzy reasoning models with application to fuzzy control", *IEEE Transactions on systems, man and cybernetics* **24** (1994) 39.
363. M. Sugeno, T. Yasukawa, "A fuzzy-logic based approach to qualitative modeling", *IEEE Transactions on Fuzzy Systems* **1** (1993) 7.
364. Y. Yoshinari, W. Pedrycz, K. Hirota, "Construction of fuzzy models through clustering-techniques", *Fuzzy Sets and Systems* **78** (1996) 1.
365. J. Keller, H. Tahani, "Implementation of conjunctive and disjunctive fuzzy logic rules with neural networks", *International Journal of Approximate Reasoning* **6** (1992) 221.
366. J. Jang, C. Sun: "Neuro-fuzzy modeling and control", *Proceedings of the IEEE 83*, 1995, 378–406.
367. K. Konstantinov, T. Yoshida, "Knowledge-based control of fermentation processes", *Biotechnology and Bioengineering* **39** (1992) 479.
368. M. Dohnal, "Fuzzy bioengineering models", *Biotechnology and Bioengineering* **27** (1985) 1146.
369. K. Konstantinov, T. Yoshida, "Physiological state control of fermentation processes", *Biotechnology and Bioengineering* **33** (1989) 1145.
370. G. Annadurai, S.R. Babu, V.R. Srinivasamoorthy, "Mathematical modeling of phenol degradation system using fuzzy comprehensive evaluation", *Bioprocess Engineering* **23** (2000) 599.
371. Ch.Venkateswarlu, K.V.-S. Naidu, "Dynamic fuzzy model based predictive controller for a biochemical reactor", *Bioprocess-Engineering* **23** (2000) 113.
372. B.Satishkumar; M. Chidambaram, "Control of unstable bioreactor using fuzzy tuned PI controller", *Bioprocess Engineering* **20** (1999) 127.
373. T. Miskiewicz, A. Kasperski, "A fuzzy logic controller to control nutrient dosage in a fed-batch baker's yeast process", *Biotechnology-Letters* **22** (2000) 1685.
374. H. Honda, T. Kobayashi, "Fuzzy control of bioprocess", *Journal of Bioscience and Bioengineering* **89** (2000) 401.
375. T.Tanaka. M. Taya, "Reduction in operator supervision time using a modified malfunction detection system with fuzzy inference in an industrial-scale fermentation of alkaline cellulase", *Journal of Bioscience and Bioengineering* **91** (2001) 106.
376. R. Berkholz, D. Roehlig, R. Guthke, "Data and knowledge based experimental design for fermentation process optimization", *Enzyme and Microbial Technology* **27** (2000) 784.
377. B. Yuan, X-Z Wang, T. Morris, "Software analyser design using data mining technology for toxicity prediction of aqueous effluents", *Waste-Management* **20** (2000) 677.
378. C.H. Sing, B. Postlethwaite, "Fuzzy relational model-based control applying stochastic and iterative methods for model identification", *Trans. IchemE.* **74** (1996) 70.
379. G. Cybenko, "Approximation by superposition of a sigmoidal function", *Math. Control Signals Systems* **2** (1989) 303.
380. K. Hornik, M. Stinchcombe, H. White, "Multilayer feedforward networks are universal approximators", *Neural Networks* **2** (1989) 359.
381. R.F. Thompson: *The brain: An Introduction to Neuroscience*, W.H. Freeman & Company, New York 1985.
382. V.M. Saucedo, B. Eikens, M.N. Karim: "Identification techniques for a recombinant fed-batch fermentation for ethanol production", in E. Galindo and O.T. Ramirez, (ed.): *Advances in bioprocess engineering*, Kluwer Academic Press, Amsterdam 1994.
383. P. Brockwell, R.A. Davis: *Time series: theory and methods*, 2nd ed., Springer, New York 1992.
384. U. Saner; S. Stephanopoulus: *Application of pattern recognition techniques to fermentation data analysisi*, 123–128. In M.N. Karim and G. Stephanopoulos (ed.), Modeling and control of biotechnical processes. Oxford: Pergamon Press (1992).
385. S.E. Fahlman, C. Lebeire: "The cascade correlation learning architecture", in D.S. Touretzky, (ed.): *Advances in neural*

information processing system, Morgan Kaufmann Publisher, Los Angeles 1990.
386. B. Eikens, M.N. Karim: "Real-time control of a waste neutralization process using radial basis functions", *Proc. ADCHEM'94 Conf.*, Kyoto, 1994 pp. 125–130.
387. A. K. Pavlou, N. Magan, D. Sharp, J. Brown, H. Barr, A. P. F. Turner, "An intelligent rapid odour recognition model in discrimination of Helicobacter pylori and other gastroesophageal isolates in vitro", *Biosensors and Bioelectronics* **15** (2000) 333.
388. R. M. Stuetz, S. George, R.A. Fenner, S.J. Hall, "Monitoring wastewater BOD using a non-specific sensor array", *Journal of Chemical Technology and Biotechnology* **74** (1999) 1069.
389. C.G. Raptis, C.I. Siettos, C.T. Kiranoudis, G.V. Bafas, "Classification of aged wine distillates using fuzzy and neural network systems", *Journal of Food Engineering* **46** (2000) 267.
390. C.H.Q. Ding:, I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks", *Bioinformatics* **17** (2001) 349.
391. T. Morimoto, T. Takeuchi, H. Miyata, Y. Hashimoto, "Pattern recognition of fruit shape based on the concept of chaos and neural networks", *Computers and Electronics in Agriculture* **26** (2000) 171.
392. A. Halme, A. Visala, J. Kankare: "Functional state concept in modeling biotechnological processes", in M.N. Karim, and G. Stephanopoulos, (ed.): *Modeling and control of biotechnical processes*, Pergamon Press, Oxford 1992, pp. 153–158
393. S. Linko, J. Juopa, Y.H. Zhu, "Neural networks as software sensors in enzyme procduction", *J. Biotechnology* **52** (1997) 257.
394. J. Glassey, M. Ignova, A.C. Ward, G.A. Montague, A.J. Morris, "Bioprocess supervision, neural networks and knowledge based systems", *J. Biotechnolgy* **52** (1997) 201.
395. D.S. Lee, J.M. Park, "Neural network modeling for online estimation of nutrient dynamics in a sequentially operated batch reactor", *J. Biotechnol.* **75** (1999) 229.
396. A. Tholudur, W.F. Ramirez, J.D. McMillan, "Mathematical modeling and optimization of cellulase protein production using Trichoderma reesei RL-P37", *Biotechnology and Bioengineering* **66** (1999) 1.
397. R.R. Leal, P. Butler, P. Lane, P.A. Payne, "Data fusion and artificial neural networks for biomass estimation", *IEEE Proc. Sci. Meas. Technol.* **144** (1997) 69.
398. M.A. Hussain, "Review of the applications of neural networks in chemical process control – simulation and online implementation", *Artificial Intelligence in Engineering* **13** (1999) 55.
399. M. Benne, P.B. Grondin, J.P. Chabriat, P. Herve, "Artificial neural networks for modeling and predictive control of an industrial evaporation process", *Journal of Food Engineering* **46** (2000) 227.
400. J.L. Casti: *Reality Rules: I. Picturing the world in mathematics – the fundamentals*, John Wiley & Sons, New York 1992.
401. J.L. Casti: *Reality Rules: I. Picturing the world in mathematics – the frontier*, John Wiley & Sons, New York 1992.
402. A.E. Brysonk: *Dynamic Optimization. Menlo Park: Addison Wesley Longman, Ins*. (1999).
403. R. Unbehauen: *Systemtheorie 1-mehrdimensionale, adaptive und nichtlineare Systeme*, 7th ed., Oldenbourg Verlag, München 1998.
404. K.S. Narenda, K. Parthasarathy, "Identification and Control of Dynamical Systems using neural networks", *IEEE Transaction of neural networks* **1** (1990) 4.
405. M. Smithson: *Ignorance and uncertainty: Emerging Paradigms*. In New York: Springer Verlag (1989).
406. B.M. Ayyub; M.M. Gupta: *Uncertainty Modeling and Analysis: Theory and Applications*. In: L. KanalN.; A. Rosenfeld (eds.) Machine Intelligence and Pattern Recognition. Amsterdam: Elsevier (1994).
407. M. Meyer: *Operation Research – Systemforschung: Eine Einführung in die praktische Bedeutung*. Stuttgart: Fischer (1990).
408. T. Siimes, P. Kinko, C. von Numers, M. Nakajjima, I. Endo, "Real-time fuzzy-knowledge-based control of baker's yeast production", *Biotechnol. Bioeng.* **45** (1995) 135.
409. K.B. Konstantinov, T. Yoshida, "Knowledge-based control of fermentation processes", *Biotechnol. Bioeng.* **39** (1992) 479.
410. Ch. Venkateswarlu, K.V.S. Naidu:, "Dynamic fuzzy model based predictive controller for a biochemical reactor", *Bioprocess-Engineering* **23** (2000) 113.
411. K. Nakano, R. Katsu, K. Tada, M. Matsumura, "Production of highly concentrated xylitol by

Candida magnoliae under a microaerobic condition maintained by simple fuzzy control", *Journal-of-Bioscience-and-Bioengineering* **89** (2000) 372.
412. T. Miskiewicz, A. Kasperski, "A fuzzy logic controller to control nutrient dosage in a fed-batch baker's yeast process" *Biotechnology-Letters* **22** (2000), 2000.
413. K. Kovarova-Kovar, S. Gehlen, A. Kunze, T. Keller, R. von Daniken, M. Kolb, P.G.M. van LoonAdolphus, "Application of model-predictive control based on artificial neural networks to optimize the fed-batch process for riboflavin production", *Journal of Biotechnology* **79** (2000) 39.
414. H.W. Chen, N.B. Chang, "Water pollution control in the river basin by fuzzy genetic algorithm-based multiobjective programming modeling", *Water Science and Technology* **37** (1998) 55.
415. C.C. Shih, K. Zuo, W.T. Wu: "Optimal fed-batch culture for penicillin G production via a hybrid neural model and a real-coded genetic algorithm", *Proceedings for the 7th international conference on computer applications in biotechnology, CAB 7*, Oxford 1998, pp. 51–54.
416. J.E. Bailey, "Toward a science of metabolic engineering", *Science* **252** (1991) 1668.
417. G. Stephanopoulos, J.J. Vallino, "Network rifidity and metabolic engineering in metabolite overproduction", *Science* **252** (1991) 1675.
418. N. Takiguchi, H. Shimizu, S. Shioya, "An online physiological state recognition system for the lysine fermentation process based on a metabolic reaction model", *Biotechnol & Bioeng.* **55** (1997) 170.
419. J. Nielsen, "Metabolic engineering: Techniques for analysis of targets for genetic manipulations", *Biotechnology & Bioengineering* **58** (1998) 125.
420. R. Brown G.; P. Hwang Y.C.: *Introduction to random signals and applied kalman filtering* 3rd ed. New York: John Wiley & Sons, New York 1997.
421. A. Kamina, R.W. Makuch, H. Zhao, "A stochastic modeling or early HIV-1 population dynamics", *Mathematical Bioscience* **170** (2001) 187.
422. Q. Zheng, "A stochastic two phase growth model", *Bulletin of Mathematical biology* **60** (1998) 151.
423. G.U. Yule: "A mathematical theory of evolution, based on the conclusions of Dr. J.C. Willis", in F.R.S. Phil. Trans. R. Soc. Lond. B213 1924, pp. 21–87.
424. A. Kolmogorov, "On the theory of markov chains", *Mathematische Annalen* **112** (1936) 155.
425. T. Rädler : "Modellierung and Simulation von Abfülllinien", VDI-Verlag Düsseldorf 1999.
426. H. Tijms: *Stochastic Modeling and Analysis: A computional Approach*. New York: John Wiley and Sons (1986).
427. L. Anderssson, L. Stradenberg, L. Haggstrom, S. Enfors, "Modeling of high cell density fed-batch cultivation", *FEMS Microbiol. Rev.* **14** (1994) 39.
428. V. Saucedo, M.N. Karim, "Online Optimization of stochastic processes using markov decision processes", *Comp. Chem. Eng.* **20** (1996) 701.
429. P. Mosterman, G. Biswas, "A comprehensive methodology for building hybrid models of physical systems", *Artificial Intelligence* **121** (2000) 171.
430. R. Tenno, P. Uronen, "State and parameter estimation for wastewater treatment processes using a stochastic model", *Control Engineering Practise* **3** (1995) 793.
431. T. Yoshida, S. Yoshida: "Horizon of bioprocess systems engineering", *Proceedings fo the 7th international conference on computer applications in biotechnology*, CAB 7, Oxford 1998.
432. A. Gelb: *Applied optimal estimation*. Cambridge: MIT Press (1996).
433. B. Hitzmann, O. Broxtermann, Y.L. Cha, O. Sobieh, E. Staerk, T. Scheper, "The control of glucose concentration during yeast fed-batch cultivation using a fast measurement complemented by an extended Kalman filter", *Bioprocess Engineering* **23** (2000) 337.
434. W.Y. Svrced, R.F. Elliott, J.E. Zajic, "The extended Kalman filter applied to a continuous culture model", *Biotechnol. Bioeng.* **16** (1974) 827.
435. M. Cote, B.P.A. Grandjean:, P. Lessard, J. Thibault, "Dynamic modeling of the activated sludge process: improving prediction using neural networks", *Water Research* **29** (1995) 995.
436. H.J.L. van Can, H.A.B. te Brakke, C. Hellinga, K.C.A.M. Leyben, "An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances", *Biotechnol. Bioeng.* **54** (1997) 549.

437. B. Center, B.P. Verma, "Fuzzy Logic for biological and agriculture systems", *Artificial intelligence Reviews* **12** (1998) 213.
438. B.D. Kelly, T-W. Chiou, M. Rosenberg, D.I.C. Wang:: "Industrial Animal Cell Culture", in G. Stephanopolous (ed.): *Biotechnol*, vol. 3 (Bioprocessing), VCH Verlagsgesellschaft, Weinheim 1993.
439. M. Howaldt, F. Walz, R. Kempken: "Kultur von Tierzellen", in H. Chmiel (ed.): *Bioprozesstechnik*, Elsevier – Spektrum Akademischer Verlag 2005.
440. S.S. Otzturk: "Cell Culture Technology – An Overview", in S.S. Ozturk and W.-S. Hu (eds.): *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*, Taylor & Francis Group, New York 2006.
441. M. Butler: *Animal Cell Culture and Technology – The Basics*. Oxford University Press (1996)
442. M.L. Shuler, F. Kargi: *Bioprocess Engineering – Basic Concepts*. Prentice Hall PTR (2002)
443. R.S. Cherry, "Animal cells in turbulent fluids: Details of the physical stimulus and the biological response", *Biotechnol. Adv.* **11** (1993) 279.
444. T.G. Cotter, M. Al-Rubeai, "Cell death (apoptosis) in cell culture systems", *Trends in Biotechnol.* **13** (1995) 150.
445. R.P. Singh, M. Al-Rubeai, C.D. Gregory, A.N. Emery, "Cell death in bioreactors: A role for apoptosis", *Biotechnol. Bioeng.* **44** (1994) 720.
446. M..Al-Rubeai, R.P. Singh, "Apoptosis in cell culture", *Curr. Opin. Biotechnol.* **9** (1998) no. 2, 152.
447. V. Glaser, "Current trends and innovations in cell culture", *Gen. Eng. N.* **21** (2001) 11.
448. G. Kretzmer, "Industrial processes with animal cell", *Appl. Microbiol. Biotechnol.* **59** (2002) 135.
449. R. Korke, A. Rink, T.K. Seow, M.C. Chung, C.W. Beattie, W.-S. Hu, "Genomic and proteomic perspectives in cell culture engineering", *J. Biotechnol.* **94** (2002) no. 1, 73.
450. F.M. Wurm, "Production of recombinant protein therapeutics in cultivated mammalian cells", *Nat. Biotechnol.* **11** (2004) 1393.
451. M. Butler, "Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals", *Appl. Microbiol. Biotechnol.* 68 (2005) no. 3, 283.
452. Y. H. Kim, T. Iida, T. Fujita, S. Terada, A. Kitayama, H. Ueda, E.V. Prochownik, E. Suzuki, "Establishment of an apoptosis-resistant and growth-controllable cell line by transfecting with inducible antisense c-jun gene", *Biotechnol. Bioeng.* **58** (1998) 65.
453. C.R. Bebbington, G. Renner, S. Thomson, D. King, G. Abrams, G.T. Yarrington, "High-level expression of a recombinant antibody from myeloma cells using a glutamine synthetase gene as an amplifiable selectable marker", *Biotechnol.* **10** (1992) 169.
454. S.S. Ozturk, W.-S. Hu (eds.): *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*. Taylor & Francis Group, New York (2006)
455. B. Griffiths: "Animal cell products, overview", in R.E. Spier (ed.): *Encyclopedia of Cell Technology*, vol. 1, John Wiley & Sons, New York, 2000, pp. 70–76.
456. S.t.J. Cryz, M. Granstro, B. Gottstein, L. Perrin, A. Cross, J. Larrick: "Immunotherapy and Vaccines". *Ullmann's Encyclopedia of Industrial Chemistry, Electronic Release*, 7th ed., Wiley-VCH, Weinheim 2005.
457. G.L. Galfre, D.S. Secher, P. Crawley: "Monoclonal Antibodies", *Ullmann's Encyclopedia of Industrial Chemistry, Electronic Release*, 7th ed., Wiley-VCH, Weinheim 2005.
458. G. Köhler, Milstein, "Continuous cultures of fused cells secreting monoclonal antibodies of predefined specifity", *Nature* **256** (1975) 495.
459. D.C. Andersen, C.F. Goochee, "The effect of cell-culture conditions on the oligosaccharide structures of secreted glycoproteins", *Curr. Opin. Biotechnol.* **5** (1994) 546.
460. S.W. Harcum: Protein Glycosylation: S.S. Ozturk, W.-S. Hu (eds.) *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*. Taylor & Francis Group, New York (2006)
461. R.G. Werner: "Transgenic technology – a challenge for mammalian cell culture production systems", in O.-W. Merten et al. (eds.): *New Developments and New Applications in Animal Cell Technology*, Kluwer Academic Publishers, 1998, 757–763.
462. M.W. Young, W.B. Okita, E.M. Brown, J.A. Curling JM, "Production of biopharmaceutical proteins in the milk of transgenic dairy animals", *BioPharm* **6** (1997) 34
463. I. Garner: "The production of proteins in the milk of transgenic livestock", A comparison of microinjection and nuclear transfer", in O.-W. Merten et al. (eds.): *New Developments and*

New Applications in Animal Cell Technology, Kluwer Academic Publishers, Amsterdam 1998, pp. 745–750.
464. I. Wilmut, A. Schnieke, J. McWhir, A. Kind, K. Campbell, "Viable offspring derived from fetal and adult mammalian cells", *Nature* **385** (1997) 810.
465. R. Langer, "Tissue engineering", *Mol. Ther.* **1** (2000) 12.
466. L.G. Griffith, G. Naughton, "Tissue engineering-current challenges and expanding opportunities", *Science* **295** (2002) 1009.
467. J.P. Petersen, A. Rücker, D. von Stechow, P. Adamietz, R. Pörtner, J.M. Rueger, N.M. Meenen, "Present and future therapies of articular cartilage defects", *Eur. J. of Trauma* **1** (2003) 1.
468. R.C. Mulligan, "The basic science of gene therapy", *Science* **60** (1993) 926.
469. J.Le Doux, M. Yarmush, "Engineering gene transfer technologies: Retroviral-mediated gene transfer", *Bioeng. Sci.* **20** (1996) 3.
470. A.J. Schlitz, K. Kühlcke, A.A. Fauser, H.G. Eckert, "Optimization of retroviral vector generation for clinical application", *J. Gene Med.* **3** (2001) 427.
471. PhRMA. *Biotechnol New Medicines in Development. Pharmaceutical Research and Manufacturers of America* 2002.
472. R. Freshney: *Culture of animal cells*. John Wiley, New York (1994)
473. V. Kasche, K. Probst, J. Maass, "The DNA and RNA content of crude and crystalline trypsin used to trypsinize animal cell cultures: kinetics of trypsinization", *Eur. J. Cell Biol* **22** (1980) 388.
474. L. Hayflick, P.S. Moorhead, "The serial cultivation of human diploid cell strains", *Exp. Cell Res.* **25** (1961) 585.
475. C.P. Edwards, A. Aruffo, "Current applications of COS cell based treatment expression system", *Curr. Opin. Biotechnol.* **4** (1993) 558.
476. L. Ikonomou, Y.J. Schneider, S.N. Agathos, "Insect cell culture for industrial production of recombinant proteins", *Appl. Microbiol. Biotechnol.* **62** (2003) no. 1, 1.
477. American Type Culture Collection (ATCC). http://www.atcc.org
478. European Collection of Animal Cell Culture (ECACC). http://www.ecacc.org.uk
479. German Resource Centre for Biological Material (DSMZ): http://www.dsmz.de
480. FDA:U.S. Food and Drug Administration. http://www.fda.gov
481. EMEA: European Medicine Agency. http://www.emea.eu.ing
482. M. Butler, H. Jenkins, "Nutritional aspects of the growth of animal cells in culture", *J. Biotechnol.* **12** (1989) 97.
483. H. Eagle, "Amino acid metabolism in mammalian cell cultures", *Science* **130** (1959) 432.
484. J. Zanghi, A. Schmelzer, R. Mendoza, R. Knop, W. Miller, "Bicarbonate Concentration and osmolality are key determinants in the inhibition of CHO cell polysialylation under elevated pCO2 or pH", *Biotechnol. Bioeng.* **65** (1999) 182.
485. S.K.W. Oh:, F.K.F. Chua:, A.B.H. Choo:, "Intracellular responses of productive hybridomas subjected to high osmotic pressure", *Biotechnol. Bioeng.* 46 (1995) 525.
486. J.S. Ryu, G.M. Lee, "Effect of hypoosmotic stress on hybridoma cell growth and antibody production", *Biotechnol. Bioeng.* **55** (1997) 565.
487. M.M. Zhu, A. Goyal, D.L. Rank, S.K. Gupta, T.Vanden Boom, S.S. Lee, "Effects of elevated pCO2 and osmolality on growth of CHO cells and production of antibody-fusion protein B1: a case study", *Biotechnol. Prog.* **21** (2005) no. 1, 70.
488. V.M. de Zengotita, A.E. Schmelzer, W.M. Miller, "Characterization of hybridoma cell responses to elevated pCO2 and osmolality: Intracellular pH, cell size, apoptosis, and metabolism", *Biotechnol. Bioeng.* **77** (2002) 369.
489. T. Fletscher, "Designing culture media for recombinant protein production", *BioProcess* **1** (2005) 30.
490. J.G. Griffiths, "Serum and growth factors in cell culture media: an introduction review", *Dev. Biol. Stand.* 66 (1987) 155.
491. D. Barnes, G. Sato, "Serum-free cell culture: an unifying approach", *Cell* **22** (1980) 649.
492. A. Gambhir, C. Zhang, A. Europa, W.-S. Hu, "Analysis of the use of fortified medium in continuous culture of mammalian cells", *Cytotechnol.* (1999) **31** 243.
493. H. Kallel, A. Jouini, S. Majoul, S. Rourou, "Evaluation of various serum and animal protein free media for the production of a veterinary rabies vaccine in BHK-21 cells", *J Biotechnol.* **95** (2002) no. 3, 195.
494. F. Frank, J. Dolníková, "Hybridoma growth and monoclonal antibody production in iron-rich protein-free medium: Effect of

nutrient concentration", *Cytotechnol.* **7** (1991) 33.
495. F. Frank, J.Dolníková, "Nucleosomes occurring in protein-free hybridoma cell cultures. Evidence for programmed cell death", *FEBS letters* **248** (1991) 285.
496. D. Fassnacht, S. Rössing, N. Ghaussy, R. Pörtner, "Influence of non-essential amino acids on apoptotic and necrotic death of mouse hybridoma cells in batch cultures", *Biotechnol. Letters* **19** (1997) 35.
497. P.L Pham, S. Perret, H.C. Doan, B. Cass, G. St-Laurent, A. Kamen, Y. Durocher, "Large-scale transient transfection of serum-free suspension-growing HEK293 EBNA1 cells: peptone additives improve cell growth and transfection efficiency", *Biotechnol Bioeng.* **84** (2003) Nov., 332.
498. J.S. Kim, B.C. Ahn, B.P. Lim, Y.D. Choi, E.C. Jo, "High-level scu-PA production by butyrate-treated serum-free culture of recombinant CHO cell line", *Biotechnol Prog.* **20** (2004), Nov.-Dec., 1788.
499. D. Fassnacht, S. Rössing, F. Fran (k. M. Al-Rubeai, R. Pörtner, "Effect of Bcl-2 expression on hybridoma cell growth in serum-supplemented, protein-free and diluted media", *Cytotechnology* **26** (1998) 119.
500. Z. Zhang, Y. Chisti, M. Moo-Young, "Effects of the hydrodynamic environment and shear protectants on survival of erythrocytes in suspension", *J Biotechnol.* **43** (1995) Nov., 33.
501. M. Al-Rubeai, A.N. Emery, S. Chalder, "The effect of Pluronic F-68 on hybridoma cells in continuous culture", *Applied Microbiology and Biotechnol* **37** (1992) 44.
502. G.M. Lee, B.O. Palsson, "Immobilization can improve the stability of hybridoma antibody productivity in serum-free media", *Biotechnol and Bioeng.* **36** (1990) 1049.
503. I. Lüdemann, R. Pörtner, C. Schaefer, K. Schick, K. Ŝrámková, K. Reher, M. Neumaier, F. Franék, H. Märkl, "Improvement of the culture stability of non-anchorage-dependent animal cells grown in serum-free media through immobilization", *Cytotechnology* **19** (1996) 111.
504. M.S. Sinacore, D. Drapeau, S. Radamson, "Adaptation of mammalian cells to growth in serum-free media", *Mol Biotechnol.* **15** (2000) July, 249.
505. T. Link, M. Backstrom, R. Graham, R. Essers, K. Zorner, J. Gatgens, J. Burchell, J. Taylor-Papadimitriou, G.C. Hansson, T. Noll, "Bioprocess development for the production of a recombinant MUC1 fusion protein expressed by CHO-K1 cells in protein-free medium", *J Biotechnol.* **110** (2004) May 13, 51.
506. R.R. Balcarcel, L.M. Clark, "Metabolic screening of mammalian cell cultures using well-plates", *Biotechnol Prog.* **19** (2003) Jan.Feb., 98.
507. J.T. Lu, Y.C. Chung, Z.R. Chan, Y.C. Hu, "A novel oscillating bioreactor BelloCell: Implications for insect cell culture and recombinant protein production", *Biotechnol Lett.* **27** (2005) Aug., 1059.
508. P. Girard, M. Jordan, M. Tsao, F.M. Wurm, "Small-scale bioreactor system for process development and optimization", *Biochem. Eng. J.* **7** (2001) no. 2, 117.
509. J.G. Auninŝ, B. Bader, A. Caola, J. Griffiths, M. Katz, P. Licari, K. Ram, C.S. Ranucci, W. Zhou, "Fluid mechanics, cell distribution, and environment in CellCube bioreactors", *Biotechnol. Prog.* **19** (2003) 2.
510. R. Kunitake, A. Suzuki, H. Ichihashi, S Matsuda, O. Hirai, K. Morimoto: *Fully automated roller bottle handling system for large scale culture of mammalian cells. Journal of Biotechnol 52*: 289 (1997)
511. R. Heidemann, U. Riese, D. Lutkemeyer, H. Buntemeyer, J. Lehmann, "The Super-Spinner: a low cost animal cell culture bioreactor for the CO2 incubator", *Cytotechnology* **14** (1994) no. 1, 1.
512. A.L. van Wezel "Growth of cell strains and primary cells on microcarriers in homogenous culture", *Nature* **216** (1967) 64.
513. K. Nilsson, F. Buzsaky, K. Mosbach, "Growth of anchorage dependent cells on macroporous microcarrier", *Bio/Technology* **4** (1986) 989.
514. B. Lundgren, G. Blüml: "Microcarriers in cell culture production", in G. Subramanian: *Bioseparation and Bioprocessing – A Handbook*, Wiley-VCH, Weinheim 1998.
515. C.A.M. van der Velden-de Groot, "Microcarrier technology, present status and perspective", *Cytotechnology* **18** (1995) 51.
516. G. Blüml: "Microcarrier Cell culture Technology", in R. Pörtner, (ed.): *Animal Cell Biotechnol – Methods and Protocols*, Humana Press 2006.
517. S.C. Wu, C.C. Liu, W.C. Lian, "Optimization of microcarrier cell culture process for the inactivated enterovirus type 71 vaccine development", *Vaccine* **22** (2004) Sept., 3858.
518. R.Z. Mendonca, L.M. Ioshimoto, R.M. Mendonca, M. De-Franco, E.J. Valentini, W.

Becak, I. Raw, C.A. Pereira ",Preparation of human rabies vaccine in VERO cell culture using a microcarrier system", *Braz J Med Biol Res.* **26** (1993) no. 12, 1305.
519. R.Z. Mendonca, L.R. Vaz-de-Lima, M.I. Oliveira, C.A. Pereira, S. Hoshino-Shimizu, "Studies on the efficiency of measles virus antigen production using VERO cell culture in a microcarrier system", *Braz J Med Biol Res.* **27** (1994) no. 7, 1575.
520. N.A. Bleckwenn, H. Golding, W.E. Bentley, J. Shiloach, "Production of recombinant proteins by vaccinia virus in a microcarrier based mammalian cell perfusion bioreactor", *Biotechnol. Bioeng.* **90** (2005) no. 6, 663.
521. R.M. Gallegos Gallegos, E.L. Espinosa L. Larios Ramos Ramirez, Kretschmer, R. Schmid, A. Aguilar Setien, "Rabies veterinary virus vaccine produced in BHK-21 cells grown on microcarriers in a bioreactor", *Arch Med Res.* **26** (1995) no. 1, 59.
522. H. Hübner: "Cell Encapsulation", in R. Pörtner (ed.): *Animal Cell Biotechnol – Methods and Protocols*, Humana Press 2006.
523. S. Murtas, G. Capuani, M. Dentini, C. Manetti, G. Masci, M. Massimi, A. Miccheli, V. Crescenzi, "Alginate beads as immobilization matrix for hepatocytes perfused in a bioreactor: a physico-chemical characterization", *J Biomater Sci Polym Ed.* **16** (2005) no. 7, 829.
524. T.D. Sielaff, M.Y. Hu, B. Amiot, M.D. Rollins, S. Rao, B.McGuire, J.R. Bloomer, W.S. Hu, F.B. Cerra, "Gel-entrapment bioartificial liver therapy in galactosamine hepatitis", *J Surg Res.* **59** (1995) 179.
525. J. Davis: "Hollow Fibre Cell Culture", in R. Pörtner, (ed.): *Animal Cell Biotechnol – Methods and Protocols*, Humana Press 2006.
526. F.W. Falkenberg, "Production of monoclonal antibodies in the miniPERM bioreactor: comparison with other hybridoma culture methods", *Res Immunol.* **149** (1998), July-Aug., 560.
527. U. Marx, H. Matthes, A. Nagel, R.V. Baehr, "Application of a hollow fiber membrane cell culture system in medicine", *Am. Biotechnol. Lab.*, **11** (1993) 26.
528. I. Martin, D. Wendt, M. Heberer, "The role of bioreactors in tissue engineering", *Trends Biotechnol.*, **22** (2004) 80.
529. R. Pörtner, St. Nagel-Heyer, Ch. Goepfert, P. Adamietz, N.M. Meenen, "Bioreactor design for tissue engineering", *Journal of Bioscience and Bioeng.* **100** (2005) no. 3, 235.
530. V. Barron, E. Lyons, C. Stenson-Cox, P.E. Mc Hugh, A. Pandit, "Bioreactors for cardiovascular cell and tissue growth: a review", *Ann. Biomed. Eng.* **31** (2003) 1017.
531. R.A. Chamuleau, "Artificial liver support in the third millennium. Artif. Cells Blood Substit. *Immobil. Biotechnol.* **31** (2003) 117.
532. L. De Bartolo, A. Bader, "Review of a flat membrane bioreactor as a bioartificial liver", *Ann. Transplant.* **6** (2001) 40.
533. J.C. Gerlach, "Development of a hybrid liver support system: A review", *International Journal of Artificial Organs* **19** (1996) 645.
534. J.E: Prenosil, P.E. Villeneuve, "Automated production of cultured epidermal autografts and sub-confluent epidermal autografts in a computer controlled bioreactor", *Biotechnol. Bioeng.* 59 (1998) 679.
535. A. Nagel, E. Effenberger, S. Koch, L.Lübbe, U. Marx: "Human cancer and primary cell culture in the new hybrid bioreactor system tecnomouse", in R.E. Spier, J.B. Griffiths and W. Berthold (eds.): *Animal cell technology: products of today, prospects for tomorrow*, Butterworth-Heinemann, 1994, pp. 296–298.
536. W.S. Hu, J.G. Aunins, "Large-scale mammalian cell culture", *Curr Opin Biotechnol.* **8** (1997) April, 148.
537. S. Ozturk, "Engineering challenges in high density cell culture systems", *Cytotechnology* **22** (1996) 3.
538. K.B. Konstantinov, "Monitoring and Control of the Physiological State of Cell Cultures", *Biotechnol and Bioeng.* **52** (1996) 271.
539. S.R. Fox, U.A. Patel, M.G. Yap, D.I. Wang, "Maximizing interferon-gamma production by Chinese hamster ovary cells through temperature shift optimization: experimental and modeling", *Biotechnol Bioeng.* **85** (2004) no. 2, 177.
540. S.M. Schatz, R.J. Kerschbaumer, G. Gerstenbauer, M. Kral, F. Dorner, F. Scheiflinger, "Higher expression of Fab antibody fragments in a CHO cell line at reduced temperature", *Biotechnol Bioeng.* **84** (2003) no. 4, 433.
541. S.S. Mostafa, X. Gu, "Strategies for improved dCO2 removal in large-scale fed-batch cultures", *Biotechnol Prog.* **19** (2003) Jan.-Feb., 45.
542. B. Frahm, H.-C. Blank, P. Cornand, W. Oelner, U. Guth, P. Lane, A. Munack, K. Johannsen, R. Pörtner, "Determination of dissolved CO2 concentration and CO2 production rate of mammalian cell suspension culture based on

off-gas measurement", *Journal of Biotechnol* **99** (2002) 133.
543. R. Kimura, W.M. Miller, "Glycosylation of CHO-Derived Recombinant tPA Produced under Elevated pCO2", *Biotechnol Progress* **13** (1997) 311.
544. T. Omase, K.-I. Higashiyama, S. Shioya, K.-I. Suga, "Effects of lactate concentration on hybridoma culture in lactate-controlled fed-batch operation", *Biotechnol and Bioeng.* **39** (1992) 556.
545. I. Lüdemann, R. Pörtner, H. Märkl, "Effect of NH3 on the cell growth of a hybridoma cell line", *Cytotechnology* **14** (1994), 11.
546. Y. Chisti "Hydrodynamic damage to animal cells", *Crit Rev Biotechnol.* 21 (2001) Oct., 67.
547. Y. Chisti, "Animal-cell damage in sparged bioreactors", *Trends Biotechnol.* (2000) **18** (10), 420.
548. J.J. Chalmers, "Cells and bubbles in sparged bioreactors", *Cytotechnology* **15** (1994) no.s 1–3, 311.
549. Ch. Fenge, E. Lüllau: "Cell Culture Bioreactors", in S. S:Ozturk, W.-S. Hu (eds.): *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*, Taylor & Francis Group, New York 2006.
550. A.N. Emery, D.C. Jan, M. Al-Rubeai, "Oxygenation of intensive cell-culture systems", *Applied Microbiology and Biotechnol* **43** (1995) 1028.
551. H.J. Henzler, O.J. Kauling, "Oxygenation of cell cultures", *Bioprocess Engineering* **9** (1993) 61.
552. N. Ma, M. Mollet, J.J. Chalmers: "Aeration, mixing and hydrodynamics in bioreactors", in S.S. Ozturk, and W.-S. Hu (eds.): *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*, Taylor & Francis Group, New York 2006.
553. S. Zhang, A. Handa-Corrigan, R.E. Spier, "A comparision of oxygenation methods for high-density perfusion cultures of animal cells", *Biotechnol and Bioeng.* **41** (1993) 685.
554. R. Pörtner, A. Schilling, I. Lüdemann, H. Märkl, "High density fed-batch cultures for hybridoma cells performed with the aid of a kinetic model", *Bioprocess Engineering* **15** (1996) 117.
555. Y. Christi, "Animal cell culture in stirred bioreactors: Observations on scale-up", *Bioprocess Engineering* **9** (1993) 191.
556. D. Nehring, P. Czermak, J. Vorlop, H. Lubben, "Experimental study of a ceramic microsparging aeration system in a pilot-scale animal cell culture", *Biotechnol Prog.* **20** (2004) Nov.-Dec., 1710.
557. J. Lehmann, G.W. Piehl, R. Schulz, "Bubble free cell culture aeration with porous moving membranes", *Development Biology Standard* **66** (1987) 227.
558. R. Johannsen, J. Lehmann: "Tierische Zellkulturen", in P. Präve, U. Faust, W. Sittig, D.A. Sukatsch: *Handbuch der Biotechnologie*. Oldenbourg Verlag, München 1994, pp. 179–206.
559. H.N. Qi, C.T. Goudar, J.D. Michaels, H.J. Henzler, G.N. Jovanovic, K.B. Konstantinov, "Experimental and theoretical analysis of tubular membrane aeration for Mammalian cell bioreactors", *Biotechnol Prog.* **19** (2003) Aug., 1183.
560. H. Märkl, "Bioreaktoren" *Chemie Ingenieur Technik* **63** (1991) no. 12, 1213.
561. W. Storhas, R. Gengenbach: *Sauerstoffeintrag in der Fermentation adhärenter Säugerzellkulturen*, GVC – VDI -Tagungen am 21. Mai 1990, Würzburg 1990.
562. J. Varley, J. Birch, "Reactor design for large scale suspension animal cell culture", *Cytotechnology* **29**() no. 3, 177.
563. D.S. Kompala, S.S. Ozturk: "Optimization of high cell density perfusion bioreactors", in S.S. Ozturk, and W.-S. Hu (eds.): *Cell Culture Technology For Pharmaceutical and Cell-Based Therapies*, Taylor & Francis Group, New York 2006.
564. L.R. Castilho, R.A. Medronho, "Cell retention devices for suspended-cell perfusion cultures", *Adv Biochem Eng Biotechnol.* **74** (2002) 129.
565. H. Dong, Y.J. Tang, R. Ohashi, J.F. Hamel, "A perfusion culture system using a stirred ceramic membrane reactor for hyperproduction of IgG2a monoclonal antibody by hybridoma cells", *Biotechnol Prog.* **21** (2005) Jan.-Feb., 140.
566. M.C. Dalm, S.M. Cuijten, W.M. van Grunsven, J. Tramper, D.E. Martens, "Effect of feed and bleed rate on hybridoma cells in an acoustic perfusion bioreactor: part I. Cell density, viability, and cell-cycle distribution", *Biotechnol Bioeng.* **88** (2004) Dec., 547.
567. R. Pörtner, M. Matsumura, T. Hatae, K. Murakami, H. Kataoka, "Perfusion-microcarrier cultivation of rCHO-cells in serum-free medium for production of human renin", *Bioprocess Engineering* **7** (1991) 63.

568. J. Stevens, S. Eickel, U. Onken: "Lamellar clarifier – A new device for animal cell retention in perfusion culture systems", in R.E. Spier, J.B. Griffiths and W. Berthold: *Animal Cell Technology: Products of today, prospects of tomorrow*, Butterworth-Heinemann, 1994, pp. 234–239.
569. T. Arai, S.Yokoyama M. Tokashiki: "50 l scale perfusion culture of hybridoma cells by gravitational settling for cell separation" in S. Kaminogawa et al. (eds.): *Animal Cell Technology: Basic & Applied Aspects*5 1993, pp. 341–346.
570. St. Apelman: "Separation of animal cells in continuous cell culture systems", in H. Murakami et al. (eds.): *Animal Cell Technology: Basic & Applied Aspects*, Kluwer Academic Publishers, Netherlands 1992, pp. 149–154
571. V. Jäger, "Kontinuierliche Durchflußzentrifuge für die Perfusionskultur tierischer Zellen", *BioTec* **2** (1992) 4.
572. P. Himmelfarb, P.S. Thayer, H.E. Martin, "Spin-filter culture: the propagation of mammalian cells in suspension", *Science* **164** (1969) 555.
573. D.C. Jan, A.N. Emery, M. Al-Rubeai, "Introducing spin filter eliminates hydrodynamic damage to hybridomas in bioreactor", *Biotechnol Techniques* **7** (1993) 351.
574. I.Z. Shirgaonkar, S. Lanthier, A. Kamen, "Acoustic cell filter: a proven cell retention technology for perfusion of animal cell cultures", *Biotechnol Adv.* **22** (2004) no. 6, 433.
575. V.M. Gorenflo, L. Smith, B. Dedinsky, B. Persson, J.M. Piret, "Scale-up and optimization of an acoustic filter for 200 L/day perfusion of a CHO cell culture", *Biotechnol Bioeng.* **80** (2002) Nov., 438.
576. M.J. Comer, M.J. Dearns, J. Wahl, M. Munster, T. Lorenz, B. Szperalski, S. Koch, U. Behrendt, H. Brunner, "Industrial production of monoclonal antibodies and therapeutic proteins by dialysis fermentation", *Cytotechnology* **3** (1990) 295.
577. R. Pörtner, H. Märkl, "Dialysis cultures", *Applied Microbiology and Biotechnol* **50** (1998) 403.
578. M.A. Applegate, G. Stephanopoulos, "Development of a single-pass ceramic matrix bioreactor for large scale mammalian cell culture", *Biotechnol and Bioeng.* **40** (1992) no. 9, 1056.
579. G. Wang, W. Zhang, C. Jacklin, L. Eppstein, D. Freedman, "High cell density perfusion culture of hybridoma cells for production of monoclonal antibodies in the celligen packed bed reactor", *Animal Cell Technology. Basic & Applied Aspects* **5** (1993) 463.
580. D.J. Oh, S.K. Choi, H.N. Chang, "High-density continuous cultures of hybridoma cells in a depth filter perfusion system", *Biotechnol and Bioeng.* **44** (1994) 895.
581. S.T. Yang, J. Luo, C. Chen, "A fibrous-bed bioreactor for continuous production of monoclonal antibody by hybridoma", *Adv Biochem Eng Biotechnol.* **87** (2004) 61.
582. R. Pörtner, H. Märkl, "Festbettreaktoren für die Kultur tierischer Zellen", *BIOforum* **18** (1995) 449.
583. D. Fassnacht, R. Pörtner, "Experimental and theoretical considerations on oxygen supply for animal cell growth in fixed bed reactors", *Journal of Biotechnol* **72** (1999) no. 3, 169.
584. N. Golmakany, M.J. Rasaee, M. Furouzandeh, S.A. Shojaosadati, S. Kashanian, K. Omidfar, "Continuous production of monoclonal antibody in a packed-bed bioreactor", *Biotechnol Appl Biochem.* **41** (2005) Jun, Pt 3, 273.
585. C. Born, M. Biselli C. Wandrey: "Production of monoclonal antibodies in a pilot scale fluidized bed bioreactor", in E.C. Beuvery et al. (eds.): *Animal Cell Technology: Developments Towards the 21st Century*, Kluwer Academic Publishers, Amsterdam 1995, pp. 683–686.
586. D. Müller, F. Unterluggauer, G. Kreismayr, C. Schmatz, A. Assadian, S. Wiederkum, S. Preis, K. Vorauer-Uhl, O. Doblhoff-Dier, H. Katinger: "Continuous perfused fluidized bed technology, increased productivities and product concentration", *Proceedings of the 16th ESACT Meeting: Animal Cell Technology: Products from Cells, Cells as Products*, 1999, pp. 319–321.
587. G. Blüml, M. Reiter, N. Zach, T. Gaida, C. Schmatz, A. Assadian, H. Katinger: "Fluidized bed technology: Influence of fluidization velocity on nutrient consumption and product expression", *Proceedings of the 15th ESACT Meeting: New developments and new applications in animal cell technology*, 1998, pp. 385—387.
588. H. Miyoshi, K. Yanagi, H. Fukuda, N. Ohshima, "Long-term continuous culture of

hepatocytes in a packed-bed reactor utilizing porous resin". *Biotechnol and Bioeng.* **43** (1994) 635.
589. T. Hongo, M. Kajikawa, S. Ishida, S. Ozawa, Y. Ohno, J. Sawada, A. Umezawa, Y. Ishikawa, T. Kobayashi, H. Honda, "Three-dimensional high-density culture of HepG2 cells in a 5-ml radial-flow bioreactor for construction of artificial liver" *J Biosci Bioeng.* **99** (2005) no. 3, 237.
590. D. Fassnacht, S. Rössing, J. Stange, R. Pörtner, "Long-term cultivation of immortalized mouse hepatocytes in a high density, fixed-bed reactor", *Biotechnol Techniques* **12** (1998) 25.
591. T. Noll, N. Jelinek, S. Schmid, M. Biselli, C. Wandrey, "Cultivation of hematopoietic stem and progenitor cells: biochemical engineering aspects", *Adv. Biochem. Eng. Biotechnol.* **74** (2002) 111.
592. G.J. Cabrita, B.S. Ferreira, C.L. da Silva, R. Goncalves, G. Almeida-Porada, J.M. Cabral, "Hematopoietic stem cells: from the bone to the bioreactor", *Trends Biotechnol.* **21** (2003) 233.
593. H. Schubert, I. Garrn, A. Berthold, W.U. Knauf, B. Reufi, T. Fietz, U.M. Gross, "Culture of haematopoietic cells in a 3-D bioreactor made of Al2O3 or apatite foam", *J. Mater. Sci. Mater. Med.* **15** (2004) 331.
594. D. Nehring, R. Gonzales R. Pörtner, P. Czermak, "Experimental and modeling study of different process modes for retroviral production in a fixed bed reactor", *Journal of Biotechnol* **122** (yyyy) no.2, 239.
595. S. McTaggart, M. Al-Rubeai, "Effects of culture parameters on the production of retroviral vectors by a human packaging cell line", *Biotechnol Prog.* **16** (2000) Sept.-Oct., 859.
596. O.W. Merten, "State-of-the-art of the production of retroviral vectors", *J. Gene Medicine* **6** (2004) 105.
597. O.W. Merten, P.E. Cruz, C. Rochette, C. Geny-Fiamma, C. Bouquet, D. Goncalves, O. Danos, M.J.T. Carrondo:, "Comparison of different bioreactor systems for the production of high titer retroviral vectors", *Biotechnol. Prog.* **17** (2001) 326.
598. D. Fassnacht, I. Reimann, R. Pörtner, H. Märkl, "Scale up von Festbettreaktoren zur Kultivierung tierischer Zellen", *Chemie Ingenieur Technik* **73** (2001) 1075.
599. R. Pörtner, H. Märkl D. Fassnacht: *Immobilization of Mammalian Cells in Fixed Bed Reactors. Bioforum*. Bioresearch plus Biotechnol. G.I.T. Verlag, Darmstadt 1999.
600. M. Motobu, S. Matsuo, P.-C. Wang, H. Kataoka, M. Matsumura, "High renin productivity of rCHO cells cultivated in radial-flow nonwoven fabric mat packed-bed reactor with increasing circulation flow rate", *Journal of Fermentation and Bioeng.* **83** (1997) 443.
601. H.P.A. Hoffmeister:, H.S. Conradt, R. Pörtner, A. Grebe, B.-U. Wilhelm and H.H.W. Hoffmeister:: "Rotating bed bioreactors: A novel way to natural proteins by tissue-like cultivation of mammalian cells", 19th ESACT Meeting, Harrogate, June 5–8, 2005
602. M.P. Durrschmid, K. Landauer, G. Simic, H. Klug, T. Keijzer, F. Trampler, A. Oudshoorn, M. Groschl, D. Muller, O. Doblhoff-Dier, "Comparison of fluidized bed and ultrasonic cell-retention systems for high cell density mammalian cell culture", *Biotechnol Prog.* **19** (2003) May-June, 1045.
603. J.M. Davis, J.A. Hanak, "Hollow-fiber cell culture", *Meth. Mol. Biol.* **75** (1997) 77.
604. N.S. Lipman, L.R. Jackson, "Hollow fibre bioreactors: an alternative to murine ascites for small scale (<1 gram) monoclonal antibody production", *Res Immunol.* **149** (1998) Jul.-Aug., 571.
605. D. Lowrey, S. Murphy, R.A. Goffe, "A comparison of monoclonal antibody productivity in different hollow fiber bioreactors", *J Biotechnol.* **36** (1994) July, 35.
606. D. Eibl, R. Eibl, "Einwegkultivierungstechnologie für biotechnische Pharma-Produktionen," *BioWorld* **3** (2005) 1. (Supplement BioteCHnet).
607. V. Glaser, "Disposable bioreactors become standard fare", *GEN* **25** (2005) no. 14, 80.
608. H. Blasey, C. Isch, C.R. Bernard, "Cellcube: a new system for large scale growth of adherent cells", *Biotechnol.Techn.* **9** (1995) no. 10, 725.
609. V. Singh, "Disposable bioreactor for cell culture using wave-induced agitation", *Cytotechnol.* **30** (1999) 149.
610. W. Weber, E. Weber, S. Geisse, K. Memmert, "Optimisation of protein expression and establishment of the wave bioreactor for Baculovirus/insect cell culture", *Cytotechnol.* **9** (2002) 77.
611. S. Fox, "The impact of disposable bioreactors on the CMO industry", *Contract Pharma* **6** (2005)

http://contractpharma.com/archive/2005/06/fature 2.php./21.8.05
612. S. Aldridge, "New biomanufacturing opportunities & challenges", *GEN* **25** (2005) no.12, 1.
613. L.N. Pierce P.W. Shabram, "Scalability of a disposable bioreactor from 12 L – 500 L run in perfusion mode with a CHO-based cell line: A tech review", *BioProcessing J* **3** (2004) no. 4, 1.
614. B. Griffiths, "Scale-up of suspension and anchorage-dependent animal cells", *Mol Biotechnol.* **17** (2001) no. 3, 225.
615. D.M. Marks, "Equipment design considerations for large scale cell culture", *Cytotechnology* **42** (2003) 21.
616. F. Hesse, M. Ebel, N. Konisch, R. Sterlinski, W. Kessler, R. Wagner, "Comparison of a production process in a membrane-aerated stirred tank and up to 1000-L airlift bioreactors using BHK-21 cells and chemically defined protein-free medium", *Biotechnol Prog.* **19** (2003) , May-June, 833.
617. C. Burger, M.J.T. Carrondo:, H. Cruz, M. Cuffe et al., "An integrated strategy for the process development of a recombinant antibody-cytokine fusion protein expressed in BHK cells", *Appl. Microbiol. Biotechnol.* **52** (1999) 345.
618. L. Chu, D.K. Robinson, "Industrial choices for protein production by large-scale cell culture", *Current Opinion in Biotechnol* **12** (2001) 180.
619. M. Krahe: *Biochemical Engineering. In: Ullmann's Encyclopedia of Industrial Chemistry*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany (2005)
620. J.B. Griffith, "Animal cell culture processes – batch or continuous?", *Journal of Biotechnol* **22** (1992) 21.
621. W. Zhou, J. Rehm, W.-S. Hu, "High viable cell concentration fed-batch cultures of hybridoma cells through online nutrient feeding", *Biotechnol and Bioeng.* **46** (1995) 579.
622. J.O. Schwabe, R. Pörtner, H. Märkl, "Improving an online feeding strategy for fed-batch cultures of hybridoma cells by dialysis and 'nutrient-split'-feeding", *Bioprocess Engineering* **20** (1999) 475.
623. P.W. Sauer, J.E. Burky, M.C. Wesson, H.D. Sternard, L. Qu, "High-yielding, generic fed-batch cell culture process for production of recombinant antibodies", *Biotechnol and Bioeng.* **67** (2000) 585.
624. P. Lenas, T. Kitade, H. Watanabe, H. Honda, T. Kobayashi, "Adaptive fuzzy control of nutrients concentration in fed-batch culture of mammalian cells", *Cytotechnology* **25** (1997) 9.
625. M. de Tremblay, M. Perrier, C. Chavarie, J. Archambault, "Fed-batch culture of hybridoma cells: comparison of optimal control approach and closed loop strategies", *Bioprocess Engineering* **9** (1993) 13.
626. S. Dhir, K.J. Jr. Morrow, R.R. Rhinehart, T. Wiesner, "Dynamic optimization of hybridoma growth in a fed-batch bioreactor", *Biotechnol and Bioeng.* **67** (2000) 197.
627. R. Pörtner, J.-O. Schwabe B. Frahm, "Evaluation of selected control strategies for fed-batch cultures of a hybridoma cell line", *Biotechnol and Applied Biochemistry* **40** (2004) 47
628. B. Frahm, P. Lane, H. Atzert, A. Munack, M. Hoffmann, V.C. Hass, R. Pörtner, "Automated, adaptive, model-based control by the Open-Loop-Feedback-Optimal (OLFO) controller for the effective fed-batch cultivation of hybridoma cells", *Biotechnol Progress* **18** (2002) 1095.
629. J.A. Roubos, C.D. de Gooijer, G. van Straten, A.J.B. van Boxtel, "Comparison of optimisation methods for fed-batch cultures of hybridoma cells", *Bioprocess Engineering* **17** (1997) 99.
630. B. Frahm, P. Lane, H. Märkl, R. Pörtner, "Improvement of a mammalian cell culture process by adaptive, model-based dialysis fed-batch cultivation and suppression of apoptosis", *Bioprocess and Biosystems Engineering* **26** (2003) 1.
631. A. Frison, K. Memmert, "Fed-batch process for MAb production" Genetic Engineering News **22** (2002) 11.
632. H. Kurokawa, Y.S. Park, S. Iijima, T. Kobayashi, "Growth characteristics in fed-batch culture of hybridoma cells with control of glucose and glutamine concentrations", *Biotechnol and Bioeng.* **44** (1994) 95.
633. L. Xie, D.I.C. Wang:, "Fed-batch cultivation of animal cells using different medium design concepts and feeding strategies", *Biotechnol and Bioeng.* **43** (1994) 1175.
634. D. Chee D. Furng Wong, K. Tin Kam Wong, L. Tang Goh, C. Kiat M. Heng Gek Sim Yap, "Impact of dynamic online fed-batch strategies on metabolism, productivity and N-glycosylation quality in CHO cell cultures", *Biotechnol Bioeng.* **89** (2005) Jan 20, 164.

635. W.M. Miller, H.W. Blanch, C.R. Wilke, "A kinetic analysis of hybridoma growth and metabolism in batch and continuous suspension culture: Effect of nutrient concentration, dilution rate and pH", *Biotechnol and Bioeng.* **32** (1987) 947.
636. M. Harigae, M. Matsumura, H. Kataoka, "Kinetic study on HBs-MAb production in continuous cultivation", *Journal of Biotechnol* **34** (1994) 227.
637. A.F. Europa, A. Gambhir, P.-C. Fu, W.S. Hu, "Multiple steady states with distinct cellular metabolism in continuous culture of mammalian cells", *Biotechnol and Bioeng.* **67** (2000) 25.
638. K.K. Frame, W.-S. Hu, "Kinetic study of hybridoma cell growth in continuous culture. I. A model for non-producing cells", *Biotechnol and Bioeng.* **37** (1991) 55.
639. R. Pörtner, and Th. Schäfer: *Modeling hybridoma cell growth and metabolism – A comparision of selected models and data; Journal of Biotechnol 49*: 119–135 (1996)
640. D. Voisard, F. Meuwly, P.A. Ruffieux, G. Baer, A. Kadouri, "Potential of cell retention techniques for large-scale high-density perfusion culture of suspended mammalian cells", *Biotechnol Bioeng.* **82** (2003) Jun 30, 751.
641. C. Geserick, H.P. Bonarius, L. Kongerslev, H. Hauser, P.P. Mueller, "Enhanced productivity during controlled proliferation of BHK cells in continuously perfused bioreactors", *Biotechnol Bioeng.* **69** (2000) Aug 5 266.
642. E. Lullau, A. Kanttinen, J. Hassel, M. Berg, A.Haag-Alvarsson, K. Cederbrant, B. Greenberg, C. Fenge, F. Schweikart, "Comparison of batch and perfusion culture in combination with pilot-scale expanded bed purification for the production of soluble recombinant beta-secretase", *Biotechnol Prog.* **19** (2003) Jan.-Feb. (1) 37.
643. B.G.D. Bödeker, R. Newcomb, P. Yuan, A. Braufman, W. Kelsey: "Production of recombinant factor VIII from perfusion cultures: I. Large-Scale fermentation" in R.E. Spier, J.B. Griffiths and W. Berthold (eds.): *Animal cell technology: products of today, prospects for tomorrow*, Butterworth-Heinemann, 1994, pp. 580–583.
644. A. Bohmann, R. Pörtner, H. Märkl, "Performance of a membrane dialysis bioreactor with radial-flow fixed bed for the cultivation of a hybridoma cell line", *Applied Microbiology and Biotechnol* **43** (1995) 772.
645. D. Fassnacht, S. Rössing, R.P. Singh, M. Al-Rubeai, R. Pörtner, "Influence of Bcl-2 on antibody productivity in high cell density perfusion cultures of hybridoma", *Cytotechnology* **30** (1999) 95.
646. M. Fussenegger, D. Fassnacht, R. Schwartz, J.A. Zanghi, M. Graf, J.E. Bailey, R. Pörtner,. "Regulated overexpression of the survival factor bcl-2 in CHO cells increases viable cell density in batch culture and decreases DNA release in extended fixed- bed cultivation", *Cytotechnology* **32** (2000) 45.
647. P. Meissner, H. Pick, A. Kulangara, P. Chatellard, K. Friedrich, F.M. Wurm, "Transient gene expression: recombinant protein production with suspension-adapted HEK293-EBNA cells", *Biotechnol Bioeng.* **75** (2001) Oct 20, 197.
648. M. Derouazi, P. Girard, F. Van Tilborgh, K. Iglesias, N. Muller, M. Bertschinger, F.M. Wurm, "Serum-free large-scale transient transfection of CHO cells", *Biotechnol Bioeng.* **87** (2004) Aug 20, 537.
649. L. Baldi, N. Muller, S. Picasso, R. Jacquet, P. Girard, H.P. Thanh, E. Derow, F.M. Wurm. "Transient gene expression in suspension HEK-293 cells: application to large-scale protein production", *Biotechnol Prog.* **21** (2005) Jan.-Feb., 148.
650. H.D. Blasey, J.P. Aubry, G.J. Mazzei, A.R. Bernard, "Large scale transient expression with COS cells", *Cytotechnology* **18** (1996) 183.
651. P. Girard, M. Derouazi, G. Baumgartner, M. Bourgeois, M. Jordan, B. Jacko, F. Wurm, "100-liter transient transfection", *Cytotechnology* **38** (2002) 15.
652. R.G. Werner, J. Walter, "Downstream Processing von aus Zellkulturen gewonnenen Produkten", *BioEng.* 1990, no. 5, 14.
653. R.G. Werner, W. Berthold, "Purification of proteins produced by biotechnological process", *Drug research* **38** (1988) no.3, 422.
654. P. Matejtschuk, R.M. Baker, G.E. Chapman: Purification and characterisation of monoclonal antibodies, in G. Subramanian: *Bioseparation and Bioprocessing – A Handbook*, Wiley-VCH, Weinheim 1998.
655. S. Liu, M. Carroll, R. Iverson, C. Valera, J. Vennari, KTurco, R. Piper, R. Kiss, H. Lutz, "Development and qualification of a novel virus removal filter for cell culture applications", *Biotechnol Prog.* **16** (2000) May-Jun., 425. Erratum in: Biotechnol Prog 16 (2000) Sep-Oct, 916.

656. J. Wang, A. Mauser, S.F. Chao, K. Remington, R. Treckmann, K. Kaiser, D. Pifat, J. Hotta, "Virus inactivation and protein recovery in a novel ultraviolet-C reactor", *Vox Sanguinis* **86** (2004) May, 230.
657. R.G. Werner, F. Walz, W. Noé A. Konrad, "Safety and economic aspects of continuous mammalian cell culture", *Journal of Biotechnol* **22** (1992) 51.
658. S.G. Stolberg, "The biotech death of Jesse Gelsinger", *N Y Times Mag* **1999**, p. 136–140, 149–150.
659. T. Ariga, "Gene therapy for primary immunodeficiency diseases: recent progress and misgivings", *Curr Pharm Des* **12** (2006) no. 5, 549.
660. R.G. Werner, et al., "Safety and economic aspects of continuous mammalian cell culture", *J Biotechnol* **22** (1992), no.s 1–2, 51.
661. R. Langer, "Tissue engineering", *Mol Ther* **1** (2000), no. 1, 12.
662. D.D. Jenkins, et al., "Tissue engineering and regenerative medicine", *Clin Plast Surg* **30** (2003), no. 4, 581.
663. B.D. Boyan, et al., "Bone and cartilage tissue engineering", *Clin Plast Surg* **26** (1999) no. 4, 629.
664. G. Vunjak-Novakovic: "The fundamentals of tissue engineering: scaffolds and bioreactors", *Novartis Found Symp*, 2003. 249: p. 34–46; discussion 46–51, 170–4, 239–41.
665. B.D. Boyan, et al., "Bone and cartilage tissue engineering", *Clin Plast Surg* 26 (1999) no. 4, 629.
666. T.B. Woodfield, et al., "Scaffolds for tissue engineering of cartilage", *Crit Rev Eukaryot Gene Expr* **12** (2002) no. 3, 209.
667. V. Zacchi, et al., "In vitro engineering of human skin-like tissue", *J Biomed Mater Res* **40** (1998) no. 2, 187.
668. B. Pomahac, et al., "Tissue engineering of skin", *Crit Rev Oral Biol Med* **9** (1998) no. 3, 333.
669. W.H. Eaglstein, V. Falanga, "Tissue engineering for skin: an update", *J Am Acad Dermatol* **39** (1998), no. 6, 1007.
670. A. Bader, et al., "Reconstruction of liver tissue in vitro: geometry of characteristic flat bed, hollow fiber, and spouted bed bioreactors with reference to the in vivo liver", *Artif Organs* **19** (1995) no. 9. 941.
671. E. Tziampazis, A. Sambanis, "Tissue engineering of a bioartificial pancreas: modeling the cell environment and device function", *Biotechnol Prog* **11** (1995) no. 2, 115.
672. G. Zund, et al., "Tissue engineering in cardiovascular surgery: MTT, a rapid and reliable quantitative method to assess the optimal human cell seeding on polymeric meshes", *Eur J Cardiothorac Surg* **15** (1999) no. 4, 519.
673. W.H. Zimmermann, T. Eschenhagen, "Tissue engineering of aortic heart valves", *Cardiovasc Res* **60** (2003) no. 3, 460.
674. G.H. Altman, "Cell differentiation by mechanical stress", *Faseb J* **16** (2002), no. 2, 270.
675. G.H. Altman, et al., "Advanced bioreactor with controlled application of multi-dimensional strain for tissue engineering", *J Biomech Eng* **124** (2002), no. 6, 742.
676. R. MauckL., et al., *Functional tissue engineering of articular cartilage through dynamic loading of chondrocyte-seeded agarose gels. J Biomech Eng*, 2000. 122(3): p. 252–60.
677. E.G. Lima, et al., "Functional tissue engineering of chondral and osteochondral constructs", *Biorheology* **41** (2004), no.s 3–4, 577.
678. A. Ignatius, et al., "Tissue engineering of bone: effects of mechanical strain on osteoblastic cells in type I collagen matrices", *Biomaterials* **26** (2005) no. 3, 311.
679. G. Grenier, et al., "Tissue reorganization in response to mechanical load increases functionality", *Tissue Eng* **11** (2005) no.s 1–2, 90.
680. J. Garvin, et al., "Novel system for engineering bioartificial tendons and application of mechanical load", *Tissue Eng* **9** (2003) no. 5, 967.
681. O. Demarteau, et al., "Development and validation of a bioreactor for physical stimulation of engineered cartilage", *Biorheology* **40** (2003) no.s 1–3, 331.
682. D.L. Butler, S.A. Goldstein, F. Guilak, "Functional tissue engineering: the role of biomechanics", *J Biomech Eng* **122** (2000), no. 6, 570.
683. H.D. Humes, "Stem cells: the next therapeutic frontier", *Trans Am Clin Climatol Assoc* **116** (2005) 167, discussion 183.
684. R.H. Lee, et al., "Characterization and expression analysis of mesenchymal stem cells from human bone marrow and adipose tissue", *Cell Physiol Biochem* **14** (2004), no.s 4–6, 311.

685. R. Cancedda, et al.: "Bone marrow stromal cells and their use in regenerating bone", *Novartis Found Symp*, 2003. 249: p. 133–43; discussion 143–7, 170–4, 239–41.
686. K. Anselme, et al., "In vitro control of human bone marrow stromal cells for bone tissue engineering", *Tissue Eng* **8** 2002. no. 6, 941.
687. P.A. Zuk, et al., "Multilineage cells from human adipose tissue: implications for cell-based therapies", *Tissue Eng* **7** (2001) no. 2, 211.
688. H.J. Kolb, E. Holler, "Hematopoietic transplantation: state of the art", *Stem Cells* **15** (1997) Suppl 1, 151; discussion 158.
689. M. Korbling, Z. Estrov, R. Champlin, "Adult stem cells and tissue repair", *Bone Marrow Transplant* **32** (2003) Suppl 1, S23.
690. D. Schmidt, et al., "Umbilical cord blood derived endothelial progenitor cells for tissue engineering of vascular grafts", *Ann Thorac Surg* **78** (2004) no. 6, 2094.
691. A.I. Caplan, S.P. Bruder, "Mesenchymal stem cells: building blocks for molecular medicine in the 21st century", *Trends Mol Med* **7** (2001) no. 6, 259.
692. D. Shi, *Biomaterials and tissue engineering. 1st ed. Biological and medical physics, biomedical engineering*. 2004, Berlin; New York: Springer-Verlag Berlin Heidelberg, 246 p.
693. M.P. Lutolf, J.A. Hubbell, "Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering", *Nat Biotechnol* **23** (2005) no. 1, 47.
694. A.P. Pego, et al., "Biodegradable elastomeric scaffolds for soft tissue engineering", *J Control Release* **87** (2003) no.s 1–3, 69.
695. H. Yoshimoto, et al., "A biodegradable nanofiber scaffold by electrospinning and its potential for bone tissue engineering", *Biomaterials* **24** (2003) no. 12, 2077.
696. K.A. Gross, L.M. Rodriguez-Lorenzo, "Biodegradable composite scaffolds with an interconnected spherical network for bone tissue engineering", *Biomaterials* **25** (2004) no. 20, 4955.
697. S.J. Lee, et al., "Macroporous biodegradable natural/synthetic hybrid scaffolds as small intestine submucosa impregnated poly(D,L-lactide-co-glycolide) for tissue-engineered bone", *J Biomater Sci Polym Ed* **15** (2004) no. 8, 1003.
698. D.W. Hutmacher, "Scaffold design and fabrication technologies for engineering tissues-state of the art and future perspectives", *J Biomater Sci Polym Ed* **12** (2001) no. 1, 107.
699. P.A. Gunatillake, R. Adhikari, "Biodegradable synthetic polymers for tissue engineering", *Eur Cell Mater* **5** (2003) 1, discussion 16.
700. D.W. Hutmacher, J.C. Goh, S.H. Teoh, "An introduction to biodegradable materials for tissue engineering applications", *Ann Acad Med Singapore* **30** (2001) no. 2, 183.
701. R.P. Lanza, R.S. Langer, W.L. Chick, "Principles of tissue engineering" 1997, San Diego Austin: Academic Press; R.G. Landes. 808 p.
702. D.L. Butler, S.A. Goldstein, F. Guilak. "Functional tissue engineering: the role of biomechanics." *J Biomech Eng* **122** (2000) no.6, 570.
703. L.G. Griffith, "Emerging design principles in biomaterials and scaffolds for tissue engineering," *Ann N Y Acad Sci* **961** (2002) 83.
704. D.W. Hutmacher, "Scaffolds in tissue engineering bone and cartilage", *Biomaterials* **21** (2000) no. 24, 2529.
705. J. Hodde, "Naturally occurring scaffolds for soft tissue repair and regeneration", *Tissue Eng* **8** (2002) no. 2, 295.
706. J. Chaudhuri, M. Al-Rubeai (2005). *Bioreactors for tissue engineeering: principles, design and operation. Dordrecht*; [London], Springer.
707. I. Martin, D. Wendt, M. Heberer, "The role of bioreactors in tissue engineering", *Trends Biotechnol* **22** (2004) no. 2, 80.
708. A. Ratcliffe, L.E. Niklason, "Bioreactors and bioprocessing for tissue engineering", *Ann N Y Acad Sci* **961** (2002), 210.
709. F. H. Silver (1994). *Biomaterials, medical devices, and tissue engineering: an integrated approach*. London; New York, Chapman & Hall.
710. K. Shetty, G. Paliyath, A. Pometto, R. E. Levin (Editors), *Food Biotechnology*, second edition, CRC Press, Boca Raton, FL (USA), 2006.
711. D. Wilke: "Bacterial food additives and dietary supplements", in J.R. Neeser, J.B. German (eds.): *Bioprocesses and biotechnology for functional foods and nutraceuticals*, Marcel Dekker, Inc., New York 2004.
712. Y. Hara, H. Izui, H. Ito, US 2005/196846 A1, 2005.
713. H.J. Lee, H.C. Seo, N.H. Park, KR 9701830 B1, 1997.
714. N. Prabhu, J.S.C. Babu:, S. Sundaram, "L-glutamic acid production in a continuous

stirred tank bioreactor using coimmobilized bio-catalyst using a fluorosensor", *Biomed. Sci. Instrum.* **38** (2002) 495.
715. J. Ohnishi, M. Ikeda, "Comparisons of potentials for L-lysine production among different Corynebacterium glutamicum strains", *Biosci. Biotechnol. Biochem.*, **70** (2006) 1017.
716. B. Bathe, EP 1574582 A1, 2005.
717. J. Ohnishi, R. Katahira, S. Mitsuhashi, S. Kakita, M. Ikeda, "A novel gnd mutation leading to increased L-lysine production in Corynebacterium glutamicum", *FEMS Microbiol. Lett.* **242** (2005) 265.
718. C. Winterhalter, W. Leinfelder, EP 885962/A1, 1998.
719. H. Takagi, S. Nakamori, A. Yamaguchi, K. Nishino, US 2005/221453 A1, 2005.
720. H. Takagi, S. Nakamori, M. Wada, H. Mori, EP 1571223 A2, 2005.
721. T. Roukas, "Biotechnology of citric acid production", *Food Sci. Technol.* **148** (2006) 349.
722. T. P. West, B. Reed, G. Xie, "Fungal citric acid production on distillers grains", *Curr. Trends Microbiol.* **1** (2004) 97.
723. P. Jamal, M. Z. Alam, M. Ramlan, M. Salleh, M.M. Nadzir, "Screening of Aspergillus for citric acid production from palm oil mill effluent", *Biotechnology* **4** (2005) 275.
724. V.R. Bindumole, K. Sasikiran, C. Balagopalan, B. Nambisan, "Cassava starch factory waste as an alternative substrate for citric acid production by Aspergillus niger MTCC 282", *Indian J. Microbiol.* **45** (2005) 235.
725. P. Johnson-Green, *Introduction to food biotechnology*, CRC Series in contemporary food science. CRC Press, Boca Raton, London, New York, D.C. Washington, 2002.
726. S.V. Kamzolova, I.G. Morgunov, A. Aurich, O.A. Perevoznikova, N.V. Shishkanova, U. Stottmeister, T.V. Finogenova, "Lipase secretion and citric acid production in Yarrowia lipolytica yeast grown on animal and vegetable fat", *Food Technol. Biotechnol.* **43** (2005) 113.
727. W. Rymowicz, E. Cibis, "Optimization of citric acid production from glucose syrup by Yarrowia lipolytica using response surface methodology", Electronic Journal of Polish Agricultural Universities 9 (2006) http://www.ejpau.media.pl/volume9/issue1/art-20.html
728. A. Vaidya, R. Pandey, S. Mudliar, M. Kumar, T. Chakrabarti, S. Devotta, "Production and recovery of lactic acid for polylactide – an overview", *Crit. Rev. Environ. Sci. Technol.* **35** (2005) 429.
729. Y. Tokiwa, T. Raku, "Polylactide (PLA) biodegradable plastics. The front line for the biological degradation mechanism and the fermentation production. Kankyo Gijutsu", *Environmental Conservation Engineering* **34** (2005) 399.
730. R.P. John, K.M. Nampoothiri, A. Pandey, "Solid-state fermentation for L-lactic acid production from agro wastes using Lactobacillus delbrueckii", *Process Biochem.* 41 (2006) 759.
731. X.M. Wang, Q.H. Wang, N.Q. Ren, X.Q. Wang, "Lactic acid production from kitchen waste with a newly characterized strain of Lactobacillus plantarum", *Chem. Biochem. Eng. Q* **19** (2005) 383.
732. E.N. Efremenko, O.V. Spiricheva, D.V. Veremeenko, A.V. Baibak, V.I. Lozinsky, "L-(+)-Lactic acid production using poly(vinyl alcohol)-cryogel-entrapped Rhizopus oryzae fungal cells", *J. Chem. Technol. Biotechnol.* **81** (2006) 519.
733. Y. Liu, W. Liao, C. Liu, S. Chen, "Optimization of L-(+)-lactic acid production using pelletized filamentous Rhizopus oryzae NRRL 395", *Appl. Biochem. Biotechnol.* (2006), 129–132, 844–853.
734. S.K. Singh, S.U. Ahmed, A. Pandey, "Metabolic engineering approaches for lactic acid production", *Process Biochem.* **41** (2006) 991.
735. S. Kardinahl, D. Rabelt, M. Reschke, "Biotransformation: from vision to technology", *Chem. Ing. Tech.* **78** (2006) 209.
736. C.I. Kado, C. J. Pujol, A.W. Chan, US 2004/086984 A1, 2004.
737. H.A. Emara, "L-ascorbic acid production by Azotobacter chroococcum", *New Egypt. J. Microbiol.* **9** (2004) 310.
738. R.D. Hancock, R. Viola, "Biotechnological approaches for L-ascorbic acid production", *Trends Biotechnol.* **20** (2002) 299.
739. A. Jimenez, M.A. Santos, M. Pompejus, J.L. Revuelta, "Metabolic engineering of the purine pathway for riboflavin production in Ashbya gossypii", *Appl. Environ. Microbiol.* **71** (2005) 5743.
740. K.H. Lee, J.K. Han, Y.H. Park, K.H. Lee, H. Choi, J.H. Park, EP 1426446 A1, 2004).
741. C. Burgess, M. O'Connell-Motherway, W. Sybesma, J. Hugenholtz, D. van Sinderen,

"Riboflavin production in Lactococcus lactis: Potential for in situ production of vitamin-enriched foods", *Appl. Environ. Microbiol.* **70** (2004) 5769.
742. P.G. Bruinenberg, E.J. Smid, EP 1625795 A1, 2006.
743. X. Duan, S. Zhang, "Dunaliella salina biomass harvest. Haihuyan Yu Huagong", *Sea-lake Salt & Chemical Industry* **27** (1998) 22.
744. A.A. Turenkov, Y.D. Turiyanskiy, E.A. Kunschikova, I.S. Kunschikova, A.S. Sten'ko, N.K. Bezkorovayna, M.S. Gorna, WO 2004/057012 A1, 2004.
745. T.J. Fang, Y.S. Cheng, "Improvement of astaxanthin production by Phaffia rhodozyma through mutation and optimization of culture conditions", *J. Ferment. Bioeng.* **75** (1993) 466.
746. T. Rose, M. Kunz, "Production of isomalt", *Landbauforschung Voelkenrode 241* **241** Sonderheft (2002), 75.
747. A. Krastanov, T. Yoshida, "Production of palatinose using Serratia plymuthica cells immobilized in chitosan", *J. Ind. Microbiol. Biotechnol.* **30** (2003) 593.
748. H.H. Sato, A.L.L. Moraes:, BR 2001002556 A, 2003.
749. X. Li, D. Zhang, F. Chen, J. Ma, Y. Dong, L. Zhang, "Klebsiella singaporensis sp. nov., a novel isomaltulose-producing bacterium" *Int. J. Syst. Evol. Microbiol.* **54** (2004) 2131.
750. T. Doerr, L. Guderjahn, J. Kowalczyk, WO 2006/007993 A1, 2006.
751. J. Kowalczyk, T. Doerr, L. Guderjahn, H. Evers, J.K. Nielebock, WO 2005/061690 A1, 2005.
752. Y. Tezuka, Y. Nagai, WO 2004/062385 A1, 2004.
753. P.S.J. Cheetham:, "Biotransformations; new routes to food ingredients", *Chem. Ind.* **7** (1995) 265.
754. J.M. Guisan, G. Alvaro, J. Aguado, M.D. Romero, M.J. Guerra, E. Polo, "Application of immobilized thermolysin derivatives for aspartame synthesis", *Rev. R. Acad. Cienc. Exactas Fis. Nat. Madrid* **88** (1994) 443.
755. J.A. Vignoli, M.A.P.C. Celligoi:, R.S.F. Silva:, "Development of a statistical model for sorbitol production by free and immobilized Zymomonas mobilis in loofa sponge Luffa cylindrica", *Process Biochem.* **41** (2006) 240.
756. M.L. Cazetta, M.A.P.C. Celligoi:, J.B. Buzato, I.S. Scarmino, R.S.F.da Silva, "Optimization study for sorbitol production by Zymomonas mobilis in sugar cane molasses", *Process Biochem.* **40** (2005) 747.
757. D.V. Cortez, I.C. Roberto, "Effect of phosphate buffer concentration on the batch xylitol production by Candida guilliermondii", *Lett. Appl. Microbiol.* **42** (2006) 321.
758. M.P. Vasquez, M.B. De Souza, Jr., N. Pereira, Jr., "RSM analysis of the effects of the oxygen transfer coefficient and inoculum size on the xylitol production by Candida guilliermondii", *Appl. Biochem. and Biotechnol.* (2006), 129–132, 256–264.
759. F.C. Sampaio, C. Alencar de Moraes, D. De Faveri, P. Perego, A. Converti, F.M.L. Passos:, "Influence of temperature and pH on xylitol production from xylose by Debaryomyces hansenii UFV-170", *Process Biochem.* **41** (2006) 675.
760. A.I. El-Batal, S.A. Khalaf, "Xylitol production from corn cobs hemicellulosic hydrolysate by Candida tropicalis immobilized cells in hydrogel copolymer carrier", *Inter. J. Agric. Biol.* **6** (2004) 1066.
761. J.M. Marton, M.G.A. Felipe:, J.B. Almeida e Silva, A. Pessoa, Jr., "Evaluation of the activated charcoals and adsorption conditions used in the treatment of sugarcane bagasse hydrolysate for xylitol production", *Braz. J. Chem. Eng.* **23** (2006) 9.
762. F. Kamal, H. Mehrgan, M.M. Assadi, S.A. Mortazavi, "Mutagenesis of Xanthomonas campestris and selection of strains with enhanced xanthan production", *Iran. Biomed. J.* **7** (2003) 91.
763. A. Becker, F. Katzen, A. Pühler, L. Ielpi, "Xanthan gum biosynthesis and application: a biochemical/genetic perspective", *Appl. Microbiol. Biotechnol.* **50** (1998) 145.
764. J. Janaun, K.W. Wong, C.M. Chu, A. Prabhakar: "Evaluation of xanthan production from Xanthomonas campestris using sago starch as carbon source", *7th World Congress of Chemical Engineering*, Glasgow, July 10–14, 2005.
765. E.J. Vandamme, W. Soetaert, "Biotechnical modification of carbohydrates", *FEMS Microbiol. Rev.* **16** (1995) 163.
766. A. Seto, H. Yoshijima, K. Toyomasu, H.O. Ogawa, H. Kakuta, K. Hosono, K. Ueda, T. Beppu, "Effective extracellular trehalose production by Cellulosimicrobium cellulans", *Appl. Microbiol. Biotechnol.* **64** (2004) 794.
767. J.C. Gonzalez-Hernandez, M. Jimenez-Estrada, A. Pena, "Comparative analysis of trehalose production by

Debaryomyces hansenii and Saccharomyces cerevisiae under saline stress", *Extremophiles* **9** (2005) 7.
768. M. Kubota, I. Sawatani, K. Oku, K. Takeuchi, S. Murai, "The development of (,(- trehalose production and its applications", *J. Appl. Glycosci.* **51** (2004) 63.
769. S. Kishino, J. Ogawa, Y. Omura, K. Matsumura, S. Shimizu, "Conjugated linoleic acid production from linoleic acid by lactic acid bacteria", *J. Am. Oil Chem. Soc.* **79** (2002) 159.
770. T.Y. Lin, "Conjugated linoleic acid production by cells and enzyme extract of Lactobacillus delbrueckii ssp. bulgaricus with additions of different fatty acids", *Food Chem.* **94** (2006) 437.
771. A. Ando, J. Ogawa, S. Kishino, S. Shimizu, "Conjugated linoleic acid production from castor oil by Lactobacillus plantarum JCM 1551", *Enzyme Microb. Technol.* **35** (2004) 40.
772. J. Ogawa, K. Matsumura, S. Kishino, Y. Omura, S. Shimizu, "Conjugated linoleic acid accumulation via 10-hydroxy-12-octadecaenoic acid during microaerobic transformation of linoleic acid by Lactobacillus acidophilus", *Appl. Environ. Microbiol.* **67** (2001) 1246.
773. R.A. Rastall, G.R. Gibson, "Prebiotic oligosaccharides: evaluation of biological activities and potential future developments", *Int. J. Probiotics Prebiotics*(2002) 107.
774. T. Mizoa, "Lactulose as a growth promoting factor for Bifidobacterium and its physiological aspects", *Bull. Int. Dairy Fed.* **313** (1996) 43.
775. Y.J. Lee, C.S. Kim, D.K. Oh, "Lactulose production by (-galactosidase in permeabilized cells of Kluyveromyces lactis", *Appl. Microbiol. Biotechnol.* **64** (2004) 787.
776. J. Mayer, J. Conrad, I. Klaiber, S. Lutz-Wahl, U. Beifuss, L. Fischer, "Enzymatic production and complete nuclear magnetic resonance assignment of the sugar lactulose", *J. Agric. Food Chem.* **23** (2004) 6983.
777. R. Gupta, P. Gigras, H. Mohapatra, V.K. Goswami, B. Chauhan, "Microbial (-amylases: a biotechnological perspective", *Process Biochem.* **38** (2003) 1599.
778. M.J.E.C. van der Maarel, B. van der Veen, J.C.M. Uitdehaag:, H. Leemhuis, L. Dijkhuizen, "Properties and applications of starch-converting enzymes of the (-amylase family", *J. Biotechnol.* **94** (2002) 137.
779. B. Balkan, F. Ertan, "Production and properties of (-amylase from Penicillium chrysogenum and its application in starch hydrolysis", *Prep. Biochem. Biotechnol.* **35** (2005) 169.
780. N. Declerck, M. Machius, P. Joyet, G. Wiegand, R. Huber, C. Gaillardin, "Engineering the thermostability of Bacillus licheniformis (-amylase", *Biologia* **57** (2002) 203.
781. J.R. Whitaker, A.C.J. Voragen, D. W. S. Wong(Eds.) *Handbook of food enzymology 2002*, Marcel Dekker, New York.
782. E.A. Snellman, R.R. Colwell, "Acinetobacter lipases: molecular biology, biochemical properties and biotechnological potential", *J. Ind. Microbiol. Biotechnol.* **31** (2004) 391.
783. Y. Shimada: "Modification of oils and fats by enzymatic process", *Mem. Fac. Eng. Osaka City Univ.* 2004, 45, 77–80.
784. W. Tungjaroenchai, C.H. White, W.E. Holmes, M.A. Drake, "Influence of adjunct cultures on volatile free fatty acids in reduced-fat Edam cheeses", *J. Dairy Sci.* **87** (2004) 3224.
785. K.E. Jaeger, T. Eggert, "Lipases for biotechnology", *Curr. Opin. Biotechnol.* **13** (2002) 390.
786. S. Ferreira-Dias, A.C. Correia, M.M.R. da Fonseca, "Response surface modeling of glycerolysis catalyzed by Candida rugosa lipase immobilized in different polyurethane foams for the production of partial glycerides", *J. Mol. Catal. B Enzym.* **21** (2003) 71.
787. P. Dominguez de Maria, C. Carboni-Oerlemans, B. Tuin, G. Bargeman, A. van der Meer, R. van Gemert, "Biotechnological applications of Candida antarctica lipase A: State-of-the-art", *J. Mol. Catal. B Enzym.* **37** (2005) 36.
788. P. Fickers, M. Ongena, J. Destain, F. Weekers, P. Thonart, "Production and down-stream processing of an extracellular lipase from the yeast Yarrowia lipolytica", *Enzyme Microb. Technol.* **38** (2006) 756.
789. H. Zorn, H. Bouws, M. Takenberg, M. Nimtz, R. Getzlaff, D.E. Breithaupt, R.G. Berger, "An extra-cellular carboxylesterase from the basidiomycete Pleurotus sapidus hydrolyses xanthophyll esters", *Biol. Chem.* **386** (2005) 435.
790. R.S. Jayani, S.S. Saxena, R. Gupta, "Microbial pectinolytic enzymes: a review", *Process Biochem.* **40** (2005) 2931.
791. L. Edens, A.A. van Dijk, S. Hopper, H. Ilgenfritz, W. Kemmner, S. Klugbauer, A. Fritz, D. Maier, P. Tan, F. Spreafico, A. Stock,

E. Kimpel, O. Heinrich, B. Gerhard, A. Ehrenreich, S. Losko, C. Wagner, U. Folkers, R. Albang, WO 2004/074468 A1, 2004.
792. Z. Yu, "Pectinase XXL – technical revolution in fruit and vegetable juice processing industry", *Shipin Gongye Keji* (Science and Technology of Food Industry) **24** (2003) 65.
793. B. Li, G. Zhao, G. Shuai, H. Hao, "Application of pectinase to the treatment of Kiwi fruit juice", *Niangjiu* (Liquer making) **29** (2002) 71.
794. Z. He, W. Li, X. Lin, "Study on pectinase processing techniques in loquat fruit juice", *Shipin Kexue* (Food Science) **25** (2004) 72.
795. C. Grassin, P. Fauquembergue, "Application of pectinases in beverages", *Prog. Biotechnol.* **14** (1996) 453.
796. M.T. Numan, N.B. Bhosle, "(-L-Arabinofuranosidases: the potential applications in biotechnology", *J. Ind. Microbiol. Biotechnol.* **33** (2006) 247.
797. A.E.S. Espinosa:, J.S. Pena, V.M. Santana, R.G. Fernandez, J.M. Grillo, "Comparison of bovine chymosin secreted from recombinant strains of Pichia pastoris and Kluyveromyces lactis", *Biotechnol. Apl.* **16** (1999) 160.
798. A.J.J. van Ooyen, P. Dekker, M. Huang, M.M.A. Olsthoorn:, D.L. Jacobs, P.A. Colussi, C.H. Taron, "Heterologous protein production in the yeast Kluyveromyces lactis", *FEMS Yeast Res.* **6** (2006) 381.
799. F. Dajnowiec, A. Reps, I. Jarmul, R. Wasilewski, "Rennet substitutes manufactured with the use microorganisms modified by genetic engineering methods", *Przem. Spozyw.* **51** (1997) 39.
800. V.V. Starovoitova, T.I. Velichko, L.A. Baratova, I.Y. Filippova, G.I. Lavrenova, "A comparative study of functional properties of calf chymosin and its recombinant forms", *Biochemistry Mosc.* **71** (2006) 320.
801. S. Kumar, N.S. Sharma, M.R. Saharan, R. Singh, "Extracellular acid protease from Rhizopus oryzae: purification and characterization", *Process Biochem.* **40** (2005) 1701.
802. S. Kumar, N.S. Sharma, M.R. Saharan, R. Singh, "Biochemical changes during ripening of Cheddar cheese prepared by milk clotting enzyme from Rhizopus oryzae", *J. Food Sci. Technol.* **41** (2004) 279.
803. J.E. Staunton, D.K. Slonim, H.A. Coller, P. Tamayo, M.J. Angelo, J. Park, U. Scherf, J.K. Lee, W.O. Reinhold, J.N. Weinstein, J.P. Mesirov, E.S. Lander, T.R. Golub, "Chemosensitivity prediction by transcriptional profiling", *Proc Natl Acad Sci U S A* **98** (2001) 10787.
804. F.S. Collins, L.D. Brooks, A. Chakravarti, "A DNA polymorphism discovery resource for research on human genetic variation", *Genome Res.* **8** (1998) 1229.
805. M.K. Halushka, I.B. Fan, K. Bently, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, A. Chakarvarti, "Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis", *Nat Genet.* **22** (1999) 239.
806. D.G. Wang, J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, E.S. Lander, et al. "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome", *Science* **280** (1998) 1077.
807. R.A. Heller, M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, R.W. Davis, "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays", *Proc. Natl. Acad. Sci. USA* **94** (1997) 2150.
808. K. Bussow, Z. Konthur, A. Lueking, H. Lehrach, G. Walter, "Protein array technology. Potential use in medical diagnostics", *Am J Pharmacogenomics*, **1** (2001) no.1, 37.
809. J. Khan, M. Bittner, Y. Chen, P.S. Meltzer, J.M. Trent, "DNA microarray technology: the anticipated impact on the study of human disease", *Biochim. Biophys. Acta* **1423** (1999) M17.
810. W.E. Evans, R.K. Guy "Gene expression as a drug discovery tool", *Nat Genet.* **36** (2004) no. 3, 214.
811. K. Ochi, Y. Daigo, T. Katagiri, S. Nagayama, T. Tsunoda, A. Myoui, N. Naka, N. Araki, I. Kudawara, M. Ieguchi, Y. Toyama, J. Toguchida, H. Yoshikawa, Y. Nakamura, "Prediction of response to neoadjuvant chemotherapy for osteosarcoma by gene-expression profiles", *Int J Oncol.* **24** (2004) no. 3 647.
812. P.Y. Lum, C.D. Armour, S.B. Stepaniants, G. Cavet, M.K. Wolf, J.S. Butler, J.C. Hinshaw, P. Garnier, G.D. Prestwich, A. Leonardson, P. Garrett-Engele, C.M. Rush, M. Bard, G. Schimmack, J.W. Phillips, C.J. Roberts, D.D.

Shoemaker, "Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes", *Cell* **116** (2004) Jan 9, 121.

813. K. Stegmaier, K.N. Ross, S.A. Colavito, S. O'Malley, B.R. Stockwell, T.R: Golub , "Gene expression-based high-throughput screening(GE-HTS) and application to leukemia differentiation", *Nat Genet.* **36** (2004) no. 3, 257.
814. D.R. Call, F.J. Brockman, D.P. Chandler, "Detecting and genotyping Escherichia coli O157:H7 using multiplexed PCR and nucleic acid microarrays", *Int J Food Microbiol.* **67** (2001) 71.
815. V. Chizhikov, A. Rasooly, K. Chumakov, D.D. Levy, "Microarray analysis of microbial virulence factors", *Appl Environ Microbiol.* **67** (2001) 3258.
816. I. Nagano, M. Kunishima, Y. Itoh, Z. Wu, Y. Takahashi, "Detection of verotoxin-producing Escherichia coli O157:H7 by multiplex polymerase chain reaction", *Microbiol Immunol.* **42** (1998) 371.
817. J. Yu, J.B. Kaper, "Cloning and characterization of the eae gene of enterohaemorrhagic Escherichia coli O157:H7", *Mol Microbiol.* **6** (1992) 411.
818. R. Prager, U. Strutz, A. Fruth, H. Tschape, "Subtyping of pathogenic Escherichia coli strains using flagellar (H)-antigens: serotyping versus fliC polymorphisms", *Int J Med Microbiol.* **292** (2003) 477.
819. Y. Liu-Stratton, S. Roy, C.K. Sen, "DNA microarray technology in nutraceutical and food safety" *Toxicol Lett.* **150** (2004) 29.
820. S. Wilkening, F. Stahl, A. Bader, "Comparison of primary human hepatocytes and hepatoma cell line HepG2 with regard to biotransformation properties", *Drug Metabolism and Disposition* **31** (2003) 1035.
821. P. Dabrowska: *The division of powers between the EU and the Member States with regard to deliberate release of GMOs (the new Directive 2001/18), GLJ*, 2002, p. 3.

Bismuth, Bismuth Alloys, and Bismuth Compounds

JOACHIM KRÜGER, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Federal Republic of Germany (Chaps. 1 – 9)

PETER WINKLER, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Federal Republic of Germany (Chaps. 1 – 9)

EBERHARD LÜDERITZ, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Federal Republic of Germany (Chaps. 1 – 9)

MANFRED LÜCK, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Federal Republic of Germany (Chaps. 1 – 9)

HANS UWE WOLF, Universität Ulm, Ulm, Federal Republic of Germany (Chap. 10)

1. Introduction 1
2. Properties 2
3. Occurrence 3
4. Production 3
4.1. Crude Bismuth from Bismuth-Rich Mixed Concentrates 3
4.2. Crude Bismuth from Lead Concentrates 5
4.2.1. Pyrometallurgical Separation of Bismuth from Metallic Lead 6
4.2.2. Processing of Bismuth-Containing Drosses 6
4.2.3. Electrolytic Separation of Lead and Bismuth. Bismuth Production 8
4.3. Crude Bismuth from Copper and Tin Concentrates 9
5. Refining 11
6. Bismuth Alloys 13
7. Bismuth Compounds 15
7.1. Inorganic Compounds 15
7.2. Organic Compounds 17
8. Chemical Analysis 17
9. Economic Aspects 18
10. Toxicology and Occupational Health 18
11. References 20

1. Introduction

Bismuth [*7440-69-9*], Bi, A_r 208.980, belongs to the fifth main group of the periodic table. Its chemical behavior is similar to that of the other group VB elements arsenic and antimony.

The electron configuration is [Xe] $4f^{14}$ $5d^{10}$ $6s^2$ $6p^3$. Only one stable isotope, ^{209}Bi, is known, but there are several unstable isotopes (^{199}Bi –^{215}Bi). Isotopes with mass number > 210 are found in the natural decay chains of radioactive elements. Isotopes with mass number $\leqq 208$ have been formed in nuclear transformations.

Bismuth is used as an alloying component in fusible alloys, but no commercial use of pure bismuth metal is known. Consumption seems to be declining; world production is only about 4000 t/a.

History. The name of the metal is to be found in all western languages. It is probably of Germanic origin and was mentioned first by ALBERTUS MAGNUS (1193 – 1280). Around 1450 BASILIUS VALENTINUS calls it *bismutum,* and PARACELSUS (1493 – 1541) named the metal *wismut.* GEORGUS AGRICOLA thought it to be a special kind of lead and describes the winning of *plumbum cinerum.* In 1739, POTT finally distinguished pure bismuth from other metals. Bismuth was known as a low-melting alloying element in the 16th and 17th centuries, and it was also used in the form of the subnitrate – in cosmetics and medicine. Industrial production be-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_171

gan around 1830 in Saxony and increased after the first bismuth ores were shipped to Europe in 1867. The discovery of fusible alloys caused production to increase in the 1930s [1–4].

2. Properties [1–6]

Physical Properties. Bismuth is a heavy, lustrous, silver-white metal with a slight pink cast. If solidification is slow, large brittle crystals form. For a metal, bismuth has some unusual properties:

1) Like germanium and gallium, its volume increases on solidification. The increase is 3.32 %.
2) Bismuth is the most diamagnetic metal. Its specific susceptibility is -16×10^{-9} m^3/kg at 293 K.
3) Only liquid mercury has a lower thermal conductivity.
4) The electrical resistance of solid bismuth is greater than that of liquid bismuth. Superconductivity has not been detected. The increase in resistivity in a magnetic field (the Hall effect) is the largest of all metals.
5) Bismuth alloys show large thermoelectric effects.
6) With the exception of beryllium, bismuth has the lowest absorption cross section for thermal neutrons.

Atomic radius	0.18 nm
Melting point	271.40 °C
Boiling point	1564 °C
Crystal system	rhombohedral
Lattice constant	$a = 0.47457$ nm
	$\alpha = 57.24°$
Latent heat of fusion	11280 J/mol
Latent heat of vaporization	178632 J/mol
Volume expansion on solidification	3.32 %
Coefficient of linear expansion	13.5×10^{-6} K^{-1}
Density at 20 °C	9790 kg/m^3
Electrical resistivity	
at 0 °C	106.8 μΩ cm
1000 °C	160.2
Specific heat at 25 °C	25.5 J $mol^{-1}K^{-1}$
Thermal conductivity	
at 0 °C	8.2 J $s^{-1}m^{-1}K^{-1}$
300 °C	11.3
400 °C	12.3
Vapor pressure	
at 893 °C	1.013×10^{-3} bar
1053 °C	1.013×10^{-2}
1266 °C	1.013×10^{-1}
Surface tension	
at 300 °C	376 mN/m
400 °C	370
500 °C	363
Absorption cross section for thermal neutrons	$(3.4\pm0.2)\times10^{-30}$ m^2/atom
Hardness, Brinell	184 N/mm^2
Hardness, Mohs	2.5
Poisson's ratio	0.33
Shear modulus	12400 MPa
Modulus of elasticity	338 GPa
Tensile strength at 20 °C	
soft	5 MPa
hard	20
Diffusion coefficient of Bi in dilute Bi–Sn alloys [7]	$(2.33\pm0.18\ cm^2/s)\times10^{-4}\exp-[Q/RT]$ where $Q = 10.6719\pm0.365$ kJ/mol
Viscosity [8]	
at 300 °C	1.65 mPa · s
350 °C	1.49
400 °C	1.37
500 °C	1.19
600 °C	1.06
Hall coefficient at 18 °C	-6.33×10^{-7} m^3 $A^{-1}s^{-1}$
Magnetic induction	0.393 T

Thermal expansion, compressibility, heat conduction, self-diffusion, electrical resistivity, and mass susceptibility depend on crystal orientation [8].

Chemical Properties. Bismuth is not very reactive; generally, it is less reactive than lead and more reactive than silver. It shows annealing colors when heated but otherwise does not oxidize in dry air. Liquid bismuth is covered by an oxide film of Bi_2O_3 that protects it from further oxidation. At 817 – 821 °C the oxide melts, and then the metal oxidizes rapidly. Therefore it is possible to separate bismuth from the noble metals by cupellation. Moisture is slightly oxidizing at high temperatures. Other possible oxides are Bi_2O_4 and Bi_2O_5. In most compounds, bismuth has a +3 oxidation state, although in some compounds, such as Bi_2O_5, bismuth pentahalides, and a number of organobismuth compounds, it has oxidation state +5. Stable compounds with selenium and tellurium have interesting semiconductivity and thermoelectric properties (→ Thermoelectricity). Bismuth has a standard electrode potential of +0.28 V, placing it between antimony and copper in the electrochemical series. It can be separated from lead by electrolysis [2].

3. Occurrence [2], [3]

Bismuth ranks 64th in abundance in the earth's crust; like antimony and cadmium, its abundance is estimated to be 0.17 to 0.2 ppm. Most bismuth sulfides occur associated with lead, copper, and silver. A few deposits in which bismuth is a major mineral are also known. The most important pure bismuth minerals are native bismuth and bismuth glance, Bi_2S_3. Other bismuth minerals are bismite, Bi_2O_3; bismuth ocher, $Bi_2O_3 \cdot 3\,H_2O$; and bismutite and bismutosphaerite, $Bi_2(O_2/CO_3)$.

Deposits from magmatic granite nearly always contain bismuth in low concentration. Bismuth is mainly found surrounding intrusive rocks, thus it is perimagmatic. Together with complex minerals of silver, copper, and lead, it can end up in deposits far from ingeneous bodies. If the other elements are carried further, the bismuth ocher left may be worth working.

Significant European deposits are situated in Spain and Saxony (Germany). In China bismuth occurs together with tungsten, molybdenum, and tin ores, which may be dressed up to 60 – 63 % Bi. Bismuth minerals are found together with tungsten ores in South Korea. Important USSR deposits are in Adrasman, Tajikistan SSR. The most important bismuth deposits are in South America. The Cu, Pb, and Zn concentrates from Peru (Junin and Oroya) contain bismuth; in Bolivia, bismuth is found together with tin and sometimes with tungsten. Mines are found in Tasua and Chorolque, Department Potosí, and Caracoles, Cimsa Cruz Cordillere, where oxide and sulfide concentrates with 35 – 45 % Bi are produced.

4. Production

Bismuth is produced in small amounts from so-called bismuth ores or concentrates. The largest amounts are recovered as byproducts from lead and copper production, mostly during processing complex copper concentrates. Bismuth is also won from molybdenum, tungsten, and tin concentrates.

4.1. Crude Bismuth from Bismuth-Rich Mixed Concentrates

Only in rare cases are as-mined ores rich enough to be treated directly to produce metal. Usually the ores have to be enriched before processing. Because of the large variety of ore compositions there is no standard procedure for upgrading. A special process has to be developed for each deposit: mechanical enrichment and flotation are usually the preferred methods. Often these methods do not produce the desired results, and the bismuth concentrates must be enriched further, either by leaching or by volatilization, or some combination of both.

For the selective separation of bismuth from complex nonferrous metal concentrates that contain up to several percent bismuth, leaching with hydrochloric acid has proved effective:

$$Bi_2S_3 + 6\,HCl \longrightarrow 2\,BiCl_3 + 3\,H_2S$$

If weak acid is used, the reaction must be supported with H_2O_2. The bismuth is precipitated by hydrolysis on dilution or partial neutralization,

$$BiCl_3 + H_2O \longrightarrow Bi(OH)Cl_2 + HCl$$

$$Bi(OH)Cl_2 \longrightarrow BiOCl + HCl$$

or by cementation with iron turnings,

$$2\,BiCl_3 + 3\,Fe \longrightarrow 3\,FeCl_2 + 2\,Bi$$

Neither procedure leads to a pure precipitate. The hydrolysis precipitates slightly soluble chlorides along with the bismuth compound. These chlorides can be precipitated only partially in advance. The cementation also precipitates the more noble metals. Such precipitates are usually smelted with soda or Na_2S to separate bismuth metal from copper and lead, which go into a matte low in bismuth.

Processing of Complex Molybdenum – Bismuth Preconcentrates. The bismuth in a preconcentrate containing (wt %) 4.8 % Bi, 14.5 % Mo, 17.9 % Fe, 8.0 % Pb, 2.5 % Cu, and 7.4 % SiO_2 can be dissolved with concentrated hydrochloric acid; molybdenum does not dissolve. The dissolved bismuth is then cemented with iron turnings.

	Composition, wt%		
	Mo	Bi	Cu
Concentrates before leaching	14.5	4.8	2.5
Concentrates after leaching	16.3	0.02	–*
Cementate	<0.1	86 – 95	<0.1

* Not usually given

The enriched product is refined. The bismuth-free molybdenum preconcentrates may be upgraded further by flotation [9]. The flow diagram of this process is shown in Figure 1. The Molybdenite Corporation of Canada floats bismuth-containing molybdenum ores: 70 – 80 % of the Bi and 93 – 95 % of the Mo are recovered. This process produces a highly enriched Mo – Bi concentrate and an Mo concentrate (Table 1). In the first stage the Mo – Bi concentrate is leached with hydrochloric acid at 90 – 100 °C and finally with nitric acid. BiOCl precipitates from the bismuth-containing solution upon dilution to pH 2 – 3. Silver and lead chlorides also precipitate. Copper may be cemented afterwards with iron scrap. The BiOCl can be smelted with coal, caustic soda, and sodium sulfate at 820 °C. The copper and iron go mostly into the slag. The crude bismuth has to be refined further [10].

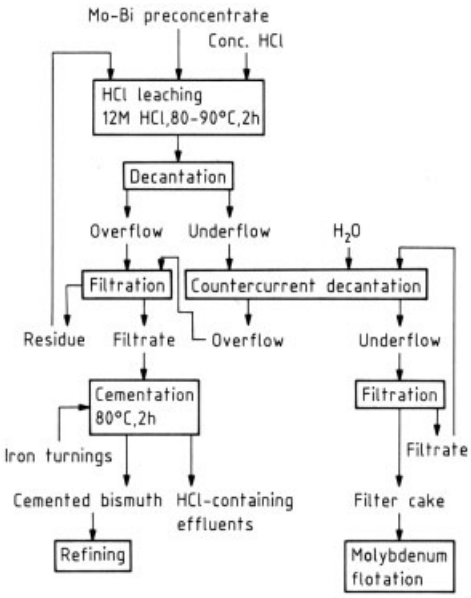

Figure 1. Hydrometallurgical separation of bismuth from Mo–Bi preconcentrate [9]

Processing of Sulfidic Copper – Bismuth Concentrates. The recovery of crude bismuth from sulfidic Cu – Bi concentrates is best illustrated by two examples.

The bismuth can be dissolved selectively with HCl in the presence of an oxidizing agent and precipitated as BiOCl by hydrolysis on neutralization of the solution. Figure 2 shows the flow diagram, and Table 2 shows the distribution of elements in the intermediate products. The wash water is recycled, which increases the recovery of Cu, Bi, and Pb. Altogether 95 % of the bismuth (and ≈ 1 % of the copper) is found in the BiOCl precipitate, and 98 % of the copper (and ≈ 4 % of the bismuth) is found with the copper concentrate and cement copper.

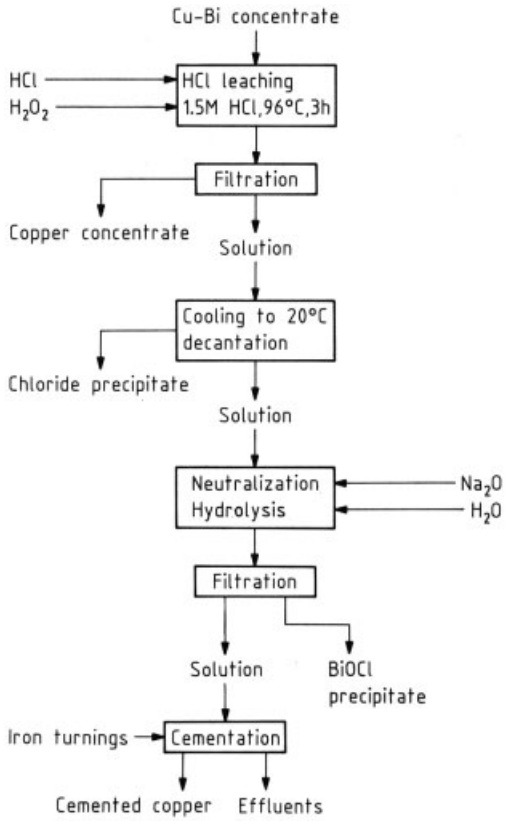

Figure 2. Hydrometallurgical separation of Bi from Cu–Bi concentrates [11]

The Cu – B concentrate is leached first with hydrochloric acid (90 – 100 °C) and then with nitric acid. BiOCl precipitates from the solution after dilution to a pH of 2 – 3. Silver and lead chlorides coprecipitate. Copper may be cemented afterwards with iron scrap [11].

Bismuth – copper concentrates are smelted in Telamayu, Bolivia, under reducing conditions after addition of soda to produce crude bismuth (> 92 % Bi), a lean alkaline copper matte (< 10 % Cu, < 0.3 % Bi, < 20 % Na), and a slag waste. The soda makes the slag low melting (< 1000 °C) and keeps the Bi content of the

Table 1. Bismuth separation from bismuth-containing molybdenum ores [10]

Substance	Concentration, wt %			
	MoS_2	Bi	Cu	Fe
Ore, as mined	0.32	0.042	$\leqq 0.1$	0.5 – 2.0
Mo – Bi concentrate	75 – 80	6 – 9	1 – 2	–
BiOCl	–	75	0.3 Ag	4 Pb
Crude bismuth	–	96	0.3 Ag	3 Pb

Table 2. Distribution of elements (wt %) in various intermediates of the oxidizing leaching process for Cu–Bi concentrates [11]∗

Substance	Bi	Cu	Pb	Ag	Au
Cu concentrate	4	76	80	82	100
Chloride precipitate	<1	–	9	8	–
BiOCl	74	<1	<2	<3	–
Cemented copper	–	14	<1	–	–
Wash water	21	8	3	–	–
Effluents	–	$\ll 1$	–	–	–

∗ Percentage of the initial concentrations in the Cu–Bi concentrate. These initial concentrations are 1.7 % Bi, 10.8 % Cu, 8.0 % Pb, and 0.34 % Ag, also present are 26.0 % Fe, 29.4 % S, 12.0 % SiO_2, and 2.8 % Zn

copper matte as low as possible. The flow sheet is shown in Figure 3. The crude bismuth is refined by conventional pyrometallurgical methods, the pre-refining in Bolivia, the final refining in Belgium [12].

4.2. Crude Bismuth from Lead Concentrates

Lead concentrates can contain significant quantities of bismuth. The average bismuth contents, g/t, are given below for several such concentrates [2]:

Kassandra, Greece	1800 – 2400
La Oroya, Peru	2300
Trepča, Yugoslavia	1600
Saxberget, Sweden	1100 – 1200
Milpo, Peru	1050
Aouli, Morocco	900
Atacocha, Peru	850
Raura, Peru	300 – 500
Bathurst, New Brunswick	450
Kimberley, British Columbia	350
Renström, Sweden	350
Garpenberg, Sweden	140 – 240
Rammelsberg, Germany	130 – 150

Lead production is usually carried out in two steps: (1) sintering and (2) reduction in a blast furnace (→ Lead). During this process the bismuth follows the lead without major losses, despite the volatility of bismuth. (See Table 3.) In any case, the flue dusts are usually recycled. The loss of bismuth in the slag is low, since bismuth is reduced before lead [14].

Table 3. Bismuth contents and distribution during conventional lead smelting [13]

Substance	Bi content, wt%	Bi distribution, %
Sinter roasting input		
Lead concentrates	0.03	66.3
Lead flue dusts	0.06	13.9
Copper flue dusts	0.07	15.1
Sinter roasting output		
Sinter	0.04	97.5
Flue dust	0.08	2.5
Blast furnace input		
Sinter	0.04	75.0
Drosses	0.07	14.4
Pb–Ag residue	0.02	3.1
Recycled slag	0.08	5.9
Blast furnace output		
Pig lead	0.08	77.3
Flue dust	0.04	12.8
Slag	0.001	1.1

Usually, bismuth is only a trace element throughout lead processing: only during lead refining is it enriched up to a few percent. The clean separation of bismuth from lead and the other elements present cannot be done by pyrometallurgical or electrolytic refining. Bismuth has less tendency than lead to form oxides, sulfides, or halides. Therefore, normal pyrometallurgical refining is not possible. Also, it is less electropositive:

$Sn^{2+} + 2e^- \longrightarrow Sn(s)$	-0.14 V
$Pb^{2+} + 2e^- \longrightarrow Pb(s)$	-0.13 V
$Sb^{3+} + 3e^- \longrightarrow Sb(s)$	$+0.2$ V
$Bi^{3+} + 3e^- \longrightarrow Bi(s)$	$+0.28$ V
$As^{3+} + 3e^- \longrightarrow As(s)$	$+0.3$ V
$Cu^{2+} + 2e^- \longrightarrow Cu(s)$	$+0.345$ V

Electrolytic refining can separate lead and bismuth, but Sb, As, Bi, and Cu remain together.

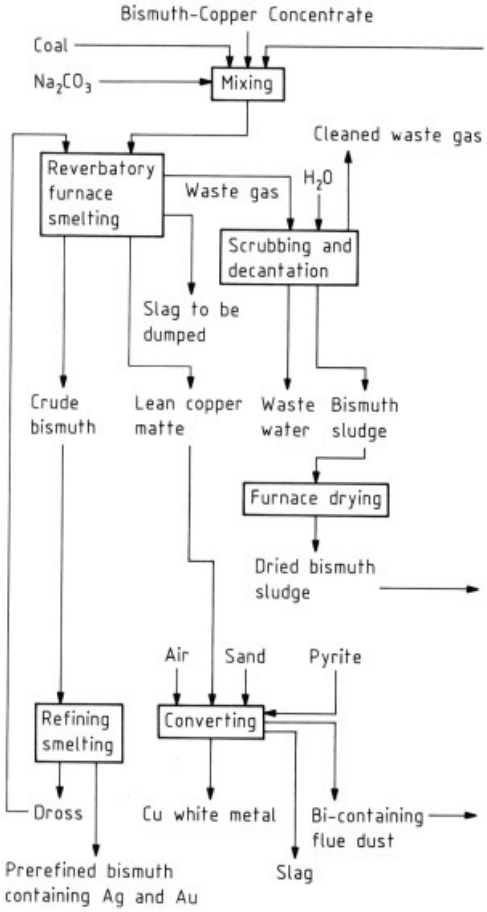

Figure 3. Processing of Bi–Cu concentrates in Telamayu, Bolivia [12]

4.2.1. Pyrometallurgical Separation of Bismuth from Metallic Lead

One method has been found to separate bismuth from lead: precipitation of the intermetallic compound $CaMg_2Bi_2$. The $CaMg_2Bi_2$ forms after addition of calcium–magnesium alloy to bismuth-containing lead melt (Kroll–Betterton process). The calcium and magnesium are added at 420 °C to the lead bath. The temperature is allowed to fall at a rate of 20 K/h until it is close to the melting point (330 °C). Then the intermetallic compounds rich in bismuth solidify, and, having a lower density than lead, float to the top, where they can be removed in the dross by skimming. The simultaneous solubilities of Ca, Bi, and Mg in lead close to its melting point can be calculated from the solubility product [15].

$$[\text{Mg,wt }\%]^2_{\text{Pb}} \cdot [\text{Ca, wt }\%]_{\text{Pb}} \cdot [\text{Bi,wt }\%]^2_{\text{Pb}} = 4.3 \times 10^{-8}$$

Figure 4 shows these equilibrium solubilities of bismuth in the lead corner of the Pb–Ca–Mg system. In practice, the bismuth concentrations in refined lead are slightly higher than the calculated values: usually ≈ 0.01 wt %. This is shown in Table 4 for two reported cases. The final bismuth concentration can be reduced below 0.01 % if antimony is added to a melt already treated with Ca and Mg [18].

The Ca–Mg consumption can be calculated. Because the residual concentrations of Ca and Mg are greater for lower bismuth contents, the specific consumption is much greater for low Bi contents, as can be seen in Table 5 [16], [17].

Table 5. Specific Ca and Mg consumptions for Bi removal from lead melts

$[Bi]_{Pb}$, wt %		Specific consumption, kg metal per kg Bi content	
before	after	Ca	Mg
0.05	0.009	1.06	3.5 calculated
0.09	0.009	0.62	2.05 calculated [16]
0.18	0.009	0.36	1.06 calculated [16]
0.2	0.020	0.25	0.60 ∗
0.68	0.014	0.15	0.39 [17]

∗ Calculated from data in *Ullmann,* 3rd ed., vol. 18, p. 635

Earlier, a mixture of K and Mg was used to separate the bismuth (Jollivet-Penarroya process), but the process is no longer in use. The reactions are analogous [15].

4.2.2. Processing of Bismuth-Containing Drosses

The dross resulting from the Kroll-Betterton process contains only a few percent bismuth. A typical composition is 6 % Bi, 0.8 % Ca, and 1.3 % Mg. The rest is Pb. If possible, the dross is enriched before bismuth recovery. Several methods are available: liquation, pressing with a Howard press, partial oxidation, or partial chlorination. Usually one-stage enrichment

Table 4. Concentrations, wt %, in molten lead after bismuth separation and calculated solubility product

Smelter	Bi	Mg	Ca	$[Mg]^2_{Pb} \cdot [Ca]_{Pb} \cdot [Bi]^2_{Pb}$	Ref.
Chimkent	0.009	0.145	0.066	1.1×10^{-7}	[16]
Norddeutsche Affinerie	0.014	0.1	0.03	0.6×10^{-7}	[17]

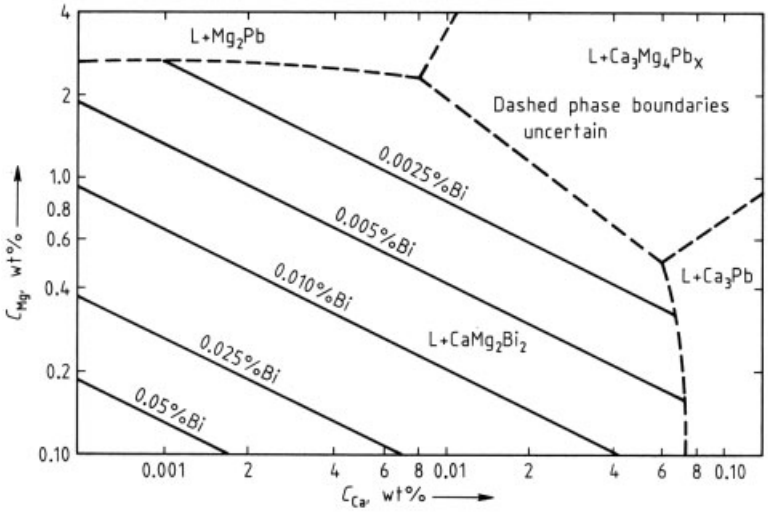

Figure 4. The Pb–Bi–Ca–Mg system, the liquidus surface in the lead corner [15]

is insufficient, so a multistage procedure is chosen. Often tar oil is mixed with the hot dross, which then burns immediately, oxidizing part of the Mg and Ca. At the same time the lead melts and separates from the dross. The flow sheet is shown in Figure 5. The solid dross remaining can be smelted to a lead – bismuth alloy. Calcium and magnesium can be removed from the dross by adding lead chloride.

$$PbCl_2 + [Ca, Mg] \longrightarrow Pb + [Ca, Mg]Cl_2$$

The resulting Pb–Bi alloy containing 20 – 30 wt % bismuth is chlorinated at 700 – 800 °C. The lead is converted to $PbCl_2$. The crude bismuth contains less than 0.5 % lead.

A process proposed by ASARCO (Fig. 6) minimizes the costs. Vacuum filtration at 450 °C separates most of the lead without oxidation. The filter cake is heated slightly higher and burns, mostly by self-sustaining combustion. Washing with $PbCl_2$ separates the oxide residue into a Pb–Bi alloy and an oxide chloride slag. The alloy is refined in the usual way [19].

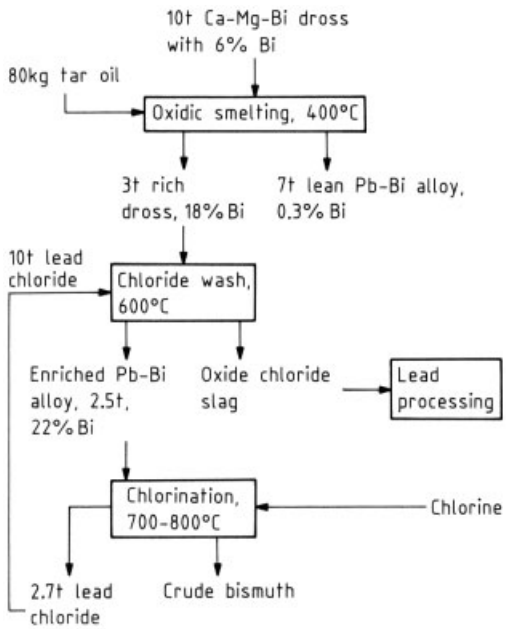

Figure 5. Conventional treatment of bismuth dross [19]

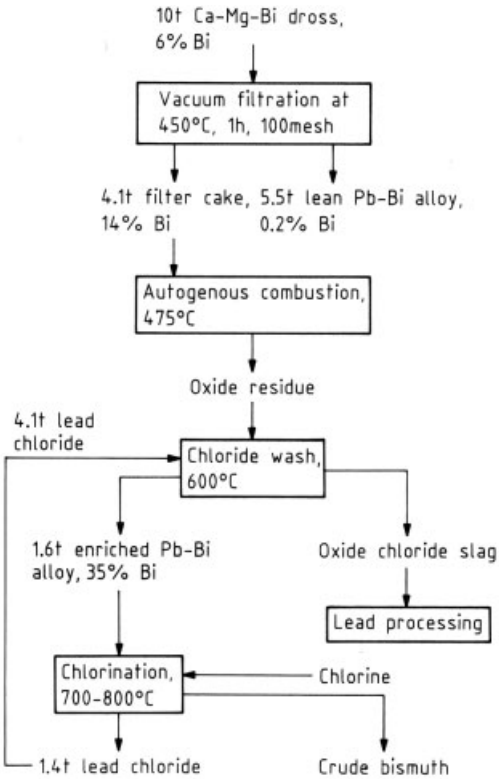

Figure 6. Bismuth treatment at ASARCO [19]

Although chlorine can be used to separate lead and bismuth, chlorine consumption is high and $PbCl_2$, not Pb, is produced. Both the Norddeutsche Affinerie, Hamburg, and the Vieille Montagne, Belgium, have chosen another method. The Bi–Cu–Mg dross is soaked in a Pb–Bi melt. The Ca and Mg are oxidized in the presence of small amounts of NaOH, < 20 kg per tonne dross. The bismuth-rich lead is cast into anodes and electrolyzed to pure lead and anode slime, which can be melted to give crude bismuth. Only the lead remaining in the crude bismuth is chlorinated. In this way the largest portion of the lead is obtained directly as metal [17], [20].

4.2.3. Electrolytic Separation of Lead and Bismuth. Bismuth Production

Some lead producers use electrolytic lead refining, usually with $PbSiF_6$–H_2SiF_6 electrolyte, which allows an almost complete lead – bismuth separation, independent of the bismuth content. Final bismuth contents $\leqq 10$ ppm can be achieved [17], [19], [21]. Almost all the bismuth goes into the anode slimes, which can contain between 2 and 25 % Bi, depending on the bismuth content of the anodes (Trail, 2 – 4 %; Oroya, 18 – 22 %; Kamioka, 22 % [2]).

At Centromin (formerly Cerro de Pasco Corp.) in La Oroya, Peru, dried anode slimes (35 % Sb, 22 % Bi, 14 % Pb, 9 % Ag, 0.03 % Au) are smelted in a reverberatory furnace to partially evaporate the antimony. The metal then is 22.9 % Sb, 30.8 % Bi, 21 % Pb, 19.6 % Ag, and 0.069 % Au. This metal is air-oxidized in a converter in two stages. In the first stage, at 900 – 1000 °C, a metal with ≈ 45 % Bi is produced. In the second stage, at 1100 – 1200 °C, a high-grade lead – silver – bismuth alloy, ≈ 50 – 55 % Ag and 30 % Bi, and a slag with 40 % Bi, 11 % Pb, 4 % Cu, and 2.5 % Sb are obtained. This slag can be reduced at 1000 – 1100 °C to crude bismuth (73 % Bi, 17 % Pb, 3.6 % Ag, 3 % Cu) [2], [22]. The flow sheet is shown in Figure 7.

The anode slimes of the electrolytic lead refinery of Consolidated Mining and Smelting Corp., Trail, Canada, contain only 2 – 4 % Bi. These anode slimes are smelted in a reverberatory furnace at 900 – 950 °C, and the resulting alloy is oxidized in two stages, at 800 – 850 °C for arsenic and antimony oxidation and at 1000 – 1100 °C for oxidation of bismuth, lead, and copper. The slag of the second stage is the raw material for bismuth recovery. It is reduced to give a Pb–Bi alloy with 25 – 30 % Bi, which is desilvered and again electrolyzed. The secondary anode slime consists mostly of bismuth. It contains 1 – 2 % Pb and only traces of other elements. The slime is reduced to crude bismuth. After treatment with chlorine to remove lead, final treatment with caustic soda yields bismuth of 99.994 % purity (20 ppm Pb, 14 ppm Cu, 33 ppm Ag). Figure 8 shows the flow sheet [2].

The lead anode slime in San Gavino Monreale is treated to produce antimonial lead and bismuth (→ Lead).

Norddeutsche Affinerie has a special status among bismuth producers. The Kroll-Betterton procedure gives a Bi–Ca–Mg dross, which is separated from lead (0.05 – 3.5 % Bi) and melted to a Pb–Bi alloy. This alloy is refined along with prerefined bismuth-rich lead in an electrolytic

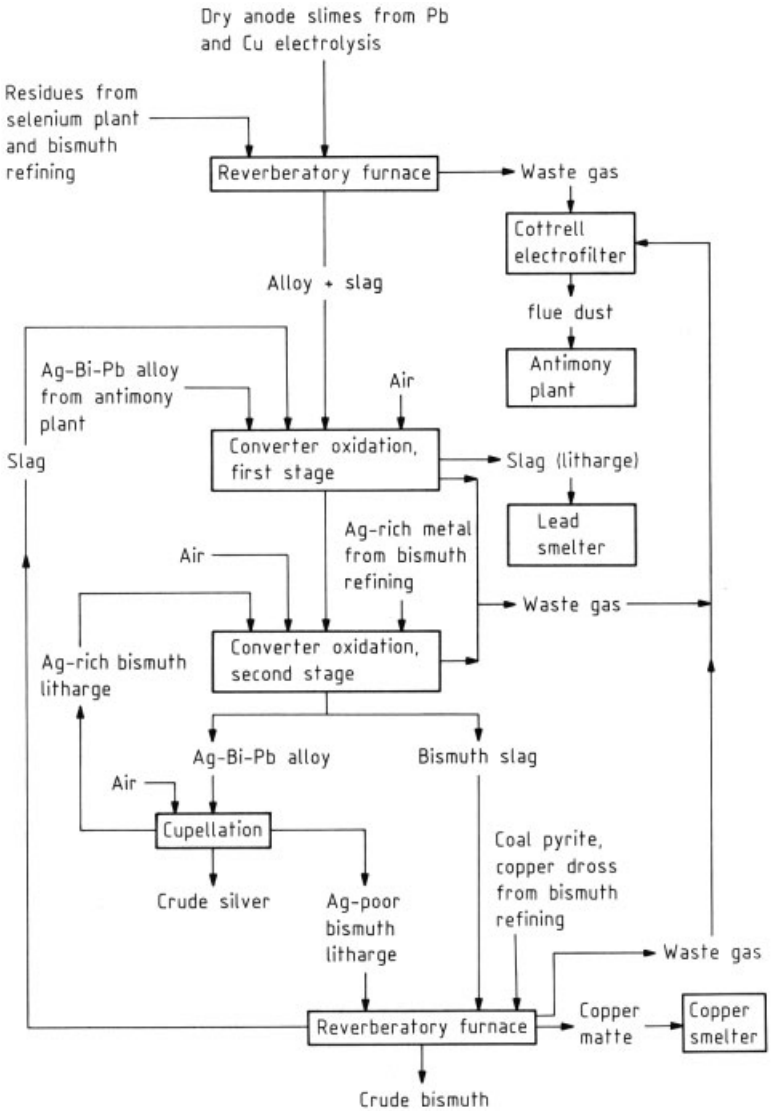

Figure 7. Recovery of crude bismuth from anode slimes at La Oroya, Peru [23]

lead refinery. The anode slimes, which are rich in bismuth, are smelted directly to crude bismuth [17]. The flow sheet is shown in Figure 9. The intermediate products have the following compositions (wt %):

	Pb	Cu	Bi
Anodes	90 – 95∗	<0.01	4 – 8
Cathodes	99.99	0.0003	0.001
Anode slimes	2 – 4	0.06	88 – 90

∗ Or more.

4.3. Crude Bismuth from Copper and Tin Concentrates

Bismuth Separation from Copper Concentrates. Most of the bismuth is volatilized during smelting of the concentrates, but the flue dusts are recycled to the smelting furnaces, at least in flash smelters, so that separation at this stage is not possible. In the following converter step the bismuth (more than 50 %), along with Pb, Sb,

and As, can be separated and enriched in the flue dust, which contains several percent bismuth. After decopperization by leaching with sulfuric acid, these flue dusts are sent to lead smelters for further metal recovery. The bismuth distribution in various steps is shown in Table 6.

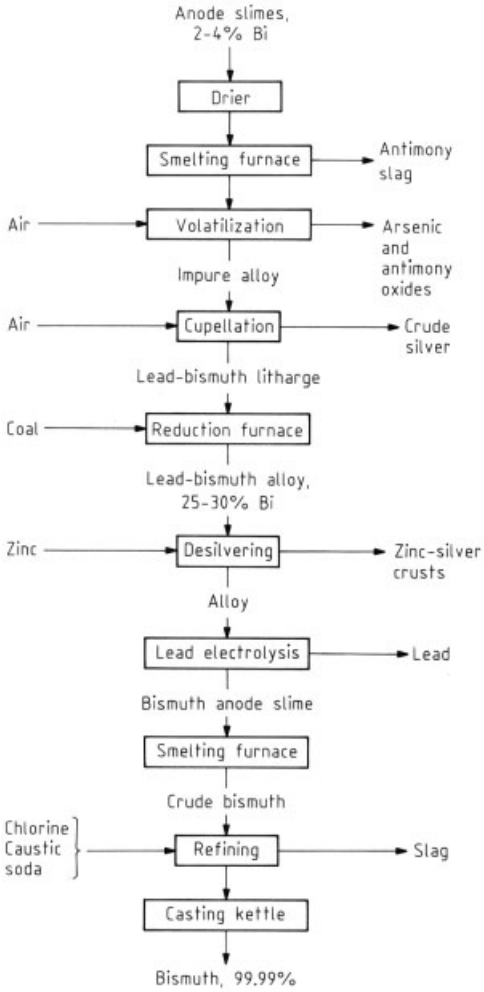

Figure 8. Bismuth recovery from lead anode slimes, Consolidated Mining and Smelting Co. of Canada [2]

Separate treatment of the converter flue dusts is also possible. After decopperization, bismuth can be leached out selectively with H_2SO_4–NaCl. It is precipitated as BiOCl [25].

The bismuth remaining in the blister copper is difficult to remove by distillation [26] or during copper electrolysis. The anode slimes contain considerable amounts of bismuth (Table 7), and treatment of these copper anode slimes enriches bismuth in the lead – bismuth litharge, which can be processed pyrometallurgically (see Figures 7 and 8).

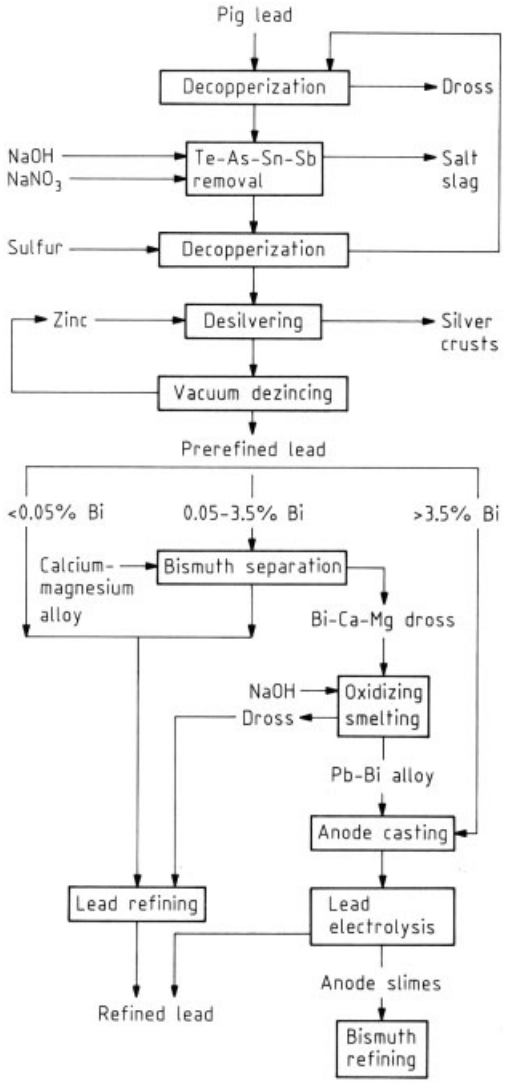

Figure 9. Simplified flow sheet of lead refining and bismuth separation at Norddeutsche Affinerie, Hamburg [17]

Bismuth Separation from Tin Concentrates. Tin concentrates often contain bismuth. The bismuth can be leached with HCl and then pre-cipitated as BiOCl, a process similar to the treatment of molybdenum concentrates [6].

In Vinto, Bolivia, the 0.5 – 5 % bismuth is separated from the rich tin concentrates (55 – 63 % Sn) by volatilization during a

Table 6. Bismuth distribution, %, during copper smelting and converting

Matte melting	matte	slag	flue dust	Ref.
	10	10	80	[24]
	35	15	50	[25]
Converting	blister copper	slag	flue dust	
	5	0	95	[24]
	10 – 15	0	90 – 85	[25]
Electrolysis	cathodes	anode slimes	elec-trolyte	
	≦ 5	30 – 60	65 – 35	[24]

Table 7. Average content of copper anode slimes, wt% [19]

	Element							
	Se	Te	Ag	Cu	Pb	Sb	As	Bi
Rönnskär, Sweden	20	1	10	40	5	1	2	0.5
Rokana NCCM, Sambia	13	1	4.5	44	1	0.1	0.3	1.1
Kennecott Copper Corp., Garfield, Utah	11 – 12	5	14	26	8.6	0.5	1.4	0.14
Mufulira RCM, Sambia	5	0.3	11 – 12	35 – 43	2 – 3	–	–	0.3
Port Kembla, New South Wales	3	2.5	6	18	10 – 25	8	7	0.4
Centromin, La Oroya, Peru	2.5	3	26	–	26	12	1.5	0.9
Anaconda Co., Great Falls, Montana	1.5	6	28	14 – 20	4 – 5	2 – 4	4 – 5	0.2

chloridizing roasting process in a multiple hearth furnace. Optimum volatilization occurs at 800 – 850 °C for an NaCl excess of only 25 – 50 %. A retention time of ≈ 1 h leads to a volatilization exceeding 97 %. The volatilized chlorides are absorbed in water and precipitated as BiOCl [27], [28].

If the bismuth is removed only partially or not at all from tin ores before reduction, it is reduced together with tin and can be separated with lead by vacuum distillation at ≈ 1150 °C and 10 Pa (0.1 mbar). The primary condensate is distilled again to give a secondary condensate. In this way the tin codistilled in the first distillation is recovered [29–32]. The rate-limiting step is evaporation at 1200 °C and diffusion at 1000 °C [29]. The extent of bismuth separation is shown in the following, all quantities in wt % [31], [32]:

	Sn	Pb	Bi
Crude tin	97	<2	<0.3
Primary condensate	50 – 60	10 – 40	3 – 30
Secondary condensate	2	80 – 90	<20
Refined tin	>99.97	<0.02	<0.01

5. Refining

Crude bismuth is always refined by pyrometallurgical methods, even though electrolytic refining is possible. The pyrometallurgical refining is almost identical with conventional lead refining (→ Lead), consisting of the following steps:

1) Decopperization by liquation and sulfuring
2) Te, As, and Sb removal by addition of $NaOH–NaNO_3$, the Harris process (→ Lead)
3) Removal of precious metals by Zn addition (Parkes process)
4) Lead and zinc removal by chlorination
5) Final oxidation with air in the presence of NaOH

With this procedure Norddeutsche Affinerie and Centromin produce 99.999 % Bi. The residual impurities are 2 ppm Cu, 3 ppm Zn, and 2 ppm Fe. Figure 10 shows a typical flow sheet.

Liquation slightly above the melting point of bismuth removes copper and other impurities. Copper can be removed more effectively if sulfur, soda, and charcoal are added. At 400 – 500 °C metal sulfides form, reducing the

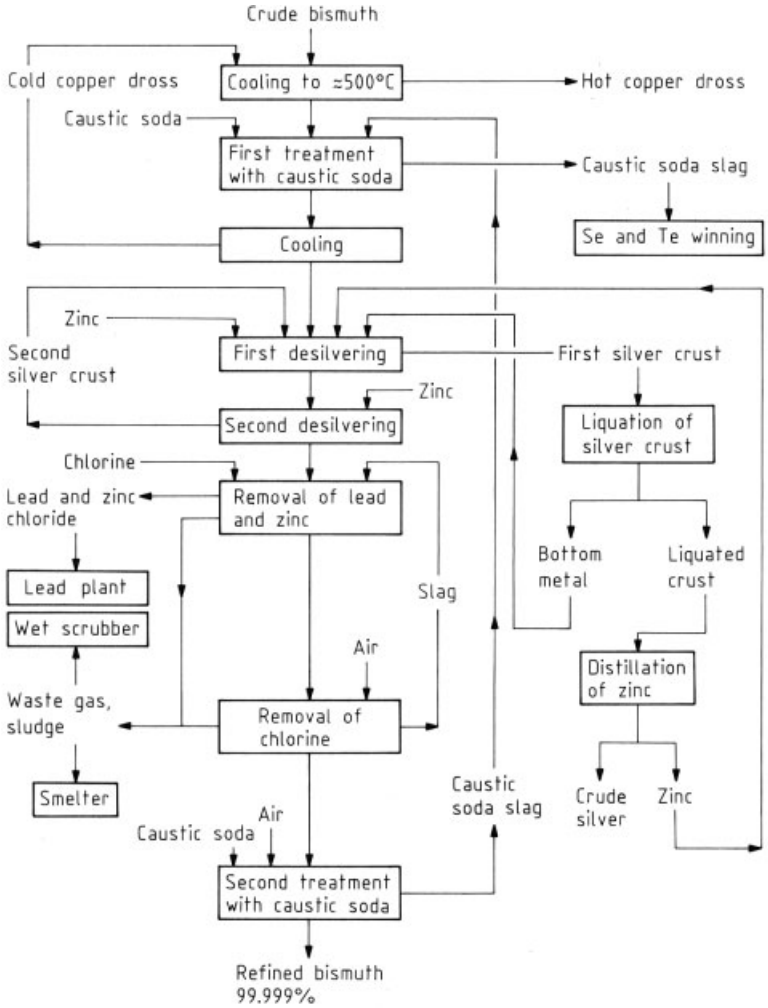

Figure 10. Refining of bismuth at Centromin, La Oroya, Peru [23]

copper concentration to 10 ppm. The elements Se, Te, As, Sb, Sn, and Zn can be oxidized with $NaNO_3$ at 400 °C and dissolved in NaOH melt. Precious metals are removed by the addition of zinc, which forms an insoluble silver – zinc alloy (→ Lead). Just as in lead refining, removal takes place in two stages: liquation of rich and poor silver crusts. Approximately 12.5 kg Zn per tonne of Bi plus 2 kg Zn per kg of Ag are required, much greater amounts than in lead refining. The residual contents of Cu, Ni, and Co are removed during desilvering, also by for-mation of insoluble alloys. After desilvering, bismuth contains more than 1 % Zn and residual contents of lead, both of which can be reduced to a few ppm by chlorination at 400 – 500 °C. Excess chlorine can be blown out with air, or the bismuth can be washed with NaOH. Also copper, nickel, and cobalt can be removed from bismuth with chlorine [33]. The chlorination can be supported electrochemically: the bismuth melt, which is covered with a chloride melt, is the anode. Generally $NaCl–CaCl_2$ or $KCl–ZnCl_2$ melt is used. Lead and copper are converted into chlorides. Current densities can reach 0.4 – 0.8 A/cm^2 [34].

This multiple-step refining can be shortened considerably if the starting material is crust and dross from bismuth separation or anode slime from electrolytic lead refining. The quality of refinery bismuth is shown in Table 8.

Bismuth of *very high purity* is produced by vacuum distillation or zone melting. The starting material is always prerefined metal. Highest purities can be achieved by a combination of vacuum distillation and multiple zone melting. Vacuum evaporation at 800 – 850 °C and condensation at 650 – 750 °C refine bismuth almost completely. Al, Au, Ba, Be, Ca, Co, Cr, Fe, Ga, In, Ir, Li, Mo, Ni, Pd, Pt, Rh, Sn, Sr, Ti, and V stay in the residue. Cd, K, Na, Se, and Zn evaporate but do not condense. Only a few elements cannot be removed: Ag, As, Mg, Pb, and Sb are insufficiently separated. Ge, Mn, and Tl follow the bismuth. At 800 °C the distillation rate is only 3.5 kg m^{-2} h^{-1} and at 850 °C it is only 14 kg m^{-2} h^{-1}; 75 – 80 % of the feed bismuth can be evaporated [35], [36].

Multiple zone melting of distilled bismuth produces very pure bismuth. Only 1 ppm of Ca, Pb, and Si and a total of 1.5 ppm of Al, Ca, Fe, and Mg are found.

Smelting bismuth with NaOH at 400 °C completely removes As, Sb, Sn, and Zn and almost completely slags Te and Pb. Careful chlorination before distillation and zone melting is recommended [37]. Treatment with chlorine removes Cu, Pb, Sn, and Zn completely. Zone melting then enriches Ag, Cu, Ni, Pb, Se, Sn, Tl, and Zn in the end piece of the rod.

6. Bismuth Alloys

The binary systems of alkali metals or alkaline-earth metals with bismuth show intermetallic compounds with high melting points. In the system Bi–Li, for example, the compound Li_3Bi melts at 1145 °C. Some intermetallic compounds, especially those with K, Ca, and Mg, are or were used to separate bismuth from lead (see Section 4.2.1). The systems with Zn, Al, Ga, Co, or Si show limited solubility in the liquid phase. Bismuth and iron are insoluble in one another; therefore, iron equipment can be used for melting and refining the metal. The rare case of complete solubility in both the liquid phase and the solid phase appears in the system Bi–Sb. Some binary alloys, particularly those with Cd, In, Pb, Sn, and Tl, as well as some multicomponent alloys, form low-melting eutectics (Table 9). The most common ones, called fusible alloys, are offered under the trade names Asarcolo, Cerrolow, Cerrobend, Belmont, and Ostalloy.

Some multicomponent and some non-eutectic compositions have both low melting points and the unusual characteristic of not contracting or even expanding on solidification. The alloy with 56 % Bi, 20 % Sn, and 24 % Pb is said to have the largest volume expansion of any alloy [6]. Some alloys show effectively no volume change: 50 % Bi – 50 % Pb and 70 % Bi – 30 % Pb, for example.

Bismuth forms with alkali metals and alkaline-earth metals alloys that superconduct at low temperatures. Examples are LiBi, NaBi, $CsBi_2$, and $BaBi_3$. Their transition temperatures are 2.47 K, 2.22 K, 4.75 K, and 4.69 K, respectively.

Bismuth alloys are usually prepared by melting the components together, but prealloyed metals are used for the exact adjustment of the composition.

Bismuth is used as an alloy addition in steel and aluminum. The fabrication of low-carbon steel and stainless steel is improved by addition of 0.15 % Bi and 2.2 % Pb. To aluminum alloys 0.2 – 0.7 % each of bismuth and lead are added [6].

Uses. The versatility of bismuth alloys results from their extremely low melting points and their volume changes on solidification combined with low viscosity and low surface tension. The composition can be chosen so that the alloy contracts, remains the same, or expands as a result of solidification. The last two cases, because they are unusual, are the most important.

The eutectic alloy Bi–Cd–In–Pb–Sn, melting at 47 °C, and some alloys melting higher are used for the automatic release of fire alarm systems, automatic sprinkler systems, electric fuses, and safety plugs for storage tanks.

Thin-walled tubes and parts for the automobile and aviation industries are covered or filled up with bismuth alloy melting below 100 °C before bending or deforming to prevent collapse. After forming, the alloy can be melted out by dipping the part into boiling water. Especially suitable is the eutectic alloy Bi – Cd – Pb – Sn, which melts at 70 °C. It is very ductile and expands on solidification and during the first hour thereafter; therefore, it keeps the worked part under tension during bending or forming.

Table 8. Composition of refinery bismuth [4]

	Bi, wt %	Contents, ppm				
		Cu	Ag	Zn	Pb	Fe
Centromin, La Oroya	99.999	2	–	3	–	2
Norddeutsche Affinerie, Hamburg	99.999	–	–	–	–	–
Cominco, Trail, Canada	99.999	–	–	–	–	–
Mitsui, Kamioka	99.99	6	9	–	10	–
San Gavino, Sardinìa	99.99	10	15	–	10	–

The noneutectic alloy with 48 % Bi, 28.5 % Pb, 9 % Sb, and 14.5 % Sn, which melts over the range 103 – 227 °C, expands greatly during the first hours after casting, reaching its final volume after ≈ 500 h. It can be used to grip tubes, tools, punches, and parts to be machined. Because it is diamagnetic, it is suited for embedding magnets in equipment or filling the space between magnets.

The noneutectic Bi–Sn alloy with 40 % Bi and 60 % Sn, which melts over the range 138 – 169 °C, contracts only 0.01 % during solidification and aging. Such an alloy exactly reproduces the shape and dimensions of the casting mold. In sprayed parts it reproduces the base in every detail. This alloy is used for test casts, copies of irregular parts, molds for thermoplastic material in the rubber and plastic industry, and for meltable casting cores. The eutectic Bi–Sn alloy with 57 % Bi, melting at 138 °C, expands 0.07 % on solidification but contracts during the next 5 h. The final net expansion is only 0.02 %. This alloy is used in dentistry for the production of models and molds [39].

Some bismuth alloys wet glass, mica, and enamelled ceramics. They are used to seal such substances to each other and to metals, as well as to produce high-vacuum seals for glass equipment. A so-called bismuth cement consists of 35 % Bi, 37 % Pb, and 25 % Sn. Such alloys are used in holding devices for grinding and polishing glass and plastic lenses. Low-melting bismuth solders are used for the soldering of temperature-sensitive electronic parts [40].

Cs–Bi alloys are used in photocathodes (→ Photoelectricity). The low conductivity of bismuth results in a good quantum yield. Bismuth tellurides and a bismuth – antimony alloy with 11 % Sb are used in photoelectric infrared detectors. The compound Bi_2Te_3, especially its solid solutions with Bi_2Se_3 and Sb_2Te_3, is a use-ful thermoelectric material (→ Thermoelectricity). The system $AgSbTe_2$ – $AgBiTe_2$ also has interesting thermoelectric properties [41]. Mn – Bi alloys are permanent magnets [34].

Table 9. Fusible alloys [6], [38]

System	Eutectic composition, wt %	Eutectic temperature, °C
Cd – Bi	Cd 40, Bi 60	144
In – Bi	In 66.3, Bi 33.7	72
	In 33, Bi 67	109
Pb – Bi	Pb 43.5, Bi 56.5	125
Sn – Bi	Sn 43, Bi 57	139
Tl – Bi	Tl 23.5, Bi 76.5	198
	Tl 52.5, Bi 47.5	188
Pb – Cd – Bi	Pb 40, Cd 8, Bi 52	92
Sn – Cd – Bi	Sn 26, Cd 20, Bi 54	102
In – Sn – Bi	In 16, Sn 26, Bi 58	79
Tl – Pb – Bi	Tl 11.5, Pb 33.3, Bi 55.2	91
	Tl 48, Pb 9.8, Bi 42.2	186
Tl – Sn – Bi	Tl 14.3, Sn 35.7, Bi 50	124
	Tl 25, Sn 31, Bi 44	167
Pb – Sn – Cd – Bi	Pb 26.7, Sn 13.3, Cd 10, Bi 50	70
In – Cd – Pb – Sn – Bi	In 19.1, Cd 5.3, Pb 22.6, Sn 8.3, Bi 44.7	47
	In 4.0, Cd 9.6, Pb 25.6, Sn 12.8, Bi 48.0	64

In nuclear technology, bismuth and its alloys may be useful for cooling and as heat-exchange mediums because bismuth has an extremely low absorption cross section for thermal neutrons. However, nothing specific has been reported about use in modern reactors.

Table 10. Physical properties of inorganic bismuth compounds

Compound	Formula	M_r	Gibbs free energy of formation, kJ/mol	Density, g/cm^3	*mp*, °C	*bp*, °C
Bismuth tribromide [*7787-58-8*]	$BiBr_3$	448.71	− 263.8	5.72	219	461
Bismuthyl carbonate [*5892-10-4*]	$(BiO)_2CO_3$	509.97	–	8.15	decomp.	–
Bismuth trichloride [*7787-60-2*]	$BiCl_3$	315.34	− 379.3	4.76	233.5	440.1
Bismuth chloride oxide [*7787-59-9*]	BiOCl	260.4	− 374.3	7.72	232.5	447
Bismuth trifluoride [*7787-61-3*]	BiF_3	266.0	− 883.4	7.90	227	405
Bismuth pentafluoride [*7787-62-4*]	BiF_5	304.0	–	5.40	151.4	230
Bismuth triiodide [*7787-64-6*]	BiI_3	589.7	− 333.5	5.80	408.5	542
Bismuth trinitrate pentahydrate [*10035-06-0*]	$Bi(NO_3)_3 \cdot 5\,H_2O$	484.1	–	2.80	75	decomp.
Bismuth trioxide [*1304-76-3*]	Bi_2O_3	465.96	− 575.7	9.32	824	1890
Bismuth trisulfide [*1345-07-9*]	Bi_2S_3	514.15	− 176.7	6.81	747	–

7. Bismuth Compounds

7.1. Inorganic Compounds

In its inorganic compounds bismuth usually has + 3 or + 5 oxidation state. The + 5 oxidation state is a strong oxidation agent, e.g., $NaBiO_6$ or BiF_5. Bismuth has an oxidation state of − 3 in a few intermetallic compounds, e.g., Li_3Bi and $CaMg_2Bi_2$. In gases, such as BiCl or BiO, in $Bi(AlCl_4)$, and in alloy-like compounds, such as BiS or BiSe, bismuth can have + 1 or + 2 oxidation state, and + 1, 0, and − 1 are present in polynuclear ionic species. However, the + 1 and + 2 states are rare, and compounds with + 4 probably do exist also. On dissolution, bismuth compounds hydrolyze easily, yielding nearly insoluble basic salts of the type BiOX. In the presence of strong acids or in complexing agents, bismuth remains in solution. Table 10 shows some physical properties of the most important inorganic bismuth compounds.

Only a few inorganic bismuth compounds are produced commercially. Bismuth oxychloride is used in cosmetics. It is the basic ingredient of a pearly pigment found in eye shadow, hair sprays, powders, nail polishes, and other cosmetic products (→ Pigments, Inorganic). Bismuth oxide can substitute for lead oxide in glass or porcelain.

The use of bismuth compounds in *medicine* is decreasing and the use of bismuth in pharmaceutical products is being viewed more and more as a questionable practice [42]. France, which in 1972 consumed more than 1000 t of bismuth, making it the largest bismuth consumer, has restricted the use of bismuth in pharmaceuticals, resulting in a sharp drop in bismuth consumption [43]. Bismuth compounds have been used because of their astringent, antiphlogistic, bacteriostatic, and disinfecting actions. In dermatology bismuth subgallate is still used in vulnerary salves and powders as well as in antimycotics. Medicines for depigmentation can no longer contain bismuth. Even bismuth subcarbonate, bismuth subnitrate, or complex bismuth citrates ought not be used for the treatment of gastric disturbances.

Bismuth Tribromide, $BiBr_3$. Golden-yellow crystals, cubic, hygroscopic. Prepared by reaction of the elements at 250 °C [44] or in methanol [45], by reaction of bismuth with molten iodine bromide and extraction with CCl_4 [46], or by dissolving Bi_2O_3 in concentrated hydrobromic acid, drying, and distillation in a stream of nitrogen. Bismuth tribromide is soluble in acetone, ether, acetic acetate, glacial acetic acid, and aqueous solutions of HCl, KCl, KBr, and KI. In water it hydrolyzes to BiOBr. It is insoluble in alcohol and carbon disulfide.

Bismuthyl Carbonate, $(BiO)_2CO_3$. Naturally occurring as bismutite. White powder, sensitive to light, decomposition during heating. Preparation by precipitation with sodium or ammonium carbonate from nitrate solutions [47], [48].

Bismuth Trichloride, $BiCl_3$. White powder, cubic, very hygroscopic. Prepared by chlorination of the molten metal [44] or by dissolution of

the metal in aqua regia, boiling down, and distillation. *Bismuth chloride oxide,* bismuth oxychloride, BiOCl, is formed by heating $BiCl_3$ in air or by hydrolysis of $BiCl_3$. The oxychloride is a white, lustrous, crystalline powder, ϱ 7.72 g/cm^3, that is practically insoluble in water, ethanol, acids, and bases. Hot concentrated alkali solutions convert it into the trioxide.

Bismuthine, BiH_3. Bismuthine [*18288-22-7*] is a colorless gas unstable at room temperature. Prepared by treating Mg_3Bi_2 with dilute acids [49], decomposition of methylbismuthine or dimethylbismuthine [50], or reaction of bismuth halides with $LiAlH_4$ in a vacuum [3].

Bismuth Trifluoride, BiF_3. White, fine pow-der, rhombic. Prepared by dissolving Bi_2O_3 in hydrofluoric acid [51] or by precipitation from an aqueous solution of bismuth trinitrate with aqueous KF or NaF [52]. Soluble only in acetone and liquid organic acids, insoluble in water. *Bismuth fluoride oxides* [*13520-72-4*], BiO_xF_{3-2x}, form when $BiF_3$is heated in air at 670 – 850 °C.

Bismuth Pentafluoride, BiF_5. White needles, formed by fluorination of bismuth metal or BiF_3 at 500 °C [53]; strong oxidant, violent reaction with water.

Bismuth Triiodide, BiI_3. Greenish-black crystalline powder. Prepared by heating stoichiometric quantities of the elements in a sealed vessel at 150 – 180 °C [54] or by reaction of concentrated hydriodic acid with a solution of $BiCl_3$ in hydrochloric acid [55]. Insoluble in cold water. Boiling BiI_3 with water produces BiOI. Bismuth triiodide is soluble in alcohol, natural gasoline, xylene, and toluene. Red triammine complexes are formed in aqueous ammonia. The yellowish-orange Dragendorff reagent, $KBiI_4$, for precipitating alkaloids, is formed by dissolving BiI_3 in potassium iodide solution.

Bismuth Trinitrate Pentahydrate, $Bi(NO_3)_3 \cdot 5\,H_2O$. Colorless, columnal crystals. Prepared by dissolving bismuth metal, Bi_2O_3, or $(BiO)_2CO_3$ in nitric acid [56]. Hydrolyzes in water to basic salts. The nitrate is soluble in strong inorganic acids, glacial acetic acid, and glycerol. It is the starting material for basic bismuth nitrates of composition $x\,Bi_2O_3 \cdot y\,N_2O_5 \cdot z\,H_2O$. Bismutum subnitricum, approximate composition $6\,Bi_2O_3 \cdot 5\,N_2O_5 \cdot 9\,H_2O$, was commonly used for medical purposes.

Bismuth Oxides. Bismuth trioxide, Bi_2O_3, is the only bismuth oxide which has been isolated in a pure state. There is also a Bi_2O_5, but its characteristics are not yet well defined. Information about other oxides and related compounds is scarce.

Bismuth trioxide, Bi_2O_3, is prepared by oxidation of bismuth metal at 750 – 800 °C, by thermal decomposition of carbonates, or by addition of alkali-metal hydroxides to a bismuth salt solution. In the last case hydrated bismuth trioxide precipitates. This gelatinous compound is usually represented by the formula $Bi(OH)_3$; calcining it yields Bi_2O_3.

Monoclinic α-Bi_2O_3 converts in an endothermic transition at 710 – 740 °C to the high-temperature cubic modification δ-Bi_2O_3, which is stable between 710 °C and the melting point. The transition is reversible, and δ-Bi_2O_3 is not stable at room temperature. Quenching thin layers of molten Bi_2O_3 forms metastable tetragonal β-Bi_2O_3, which converts back to α-Bi_2O_3 when heated. This metastable β-Bi_2O_3 dissolves oxygen in its lattice without a structure change; the composition can range from $BiO_{1.50}$ to $BiO_{1.75}$. Heating β-$BiO_{>1.50}$ to 750 °C and quenching at a moderate rate forms metastable cubic γ-Bi_2O_3, which contains more oxygen than the stoichiometric formula Bi_2O_3 expresses. The preparation of all these modifications is described in detail in [3].

Quenching molten mixtures of Bi_2O_3 and other metal oxides from 800 – 1000 °C to 0 °C produces stabilized δ- and β-modifications; tempering for 24 h at 600 °C yields a γ-modification contaminated with other oxides.

Bismuth trioxide is soluble in strong inorganic acids and concentrated alkaline solutions containing glycerol. Above 710 °C it attacks metal oxides; in fact, molten bismuth trioxide dissolves every metal oxide and even corrodes platinum.

Bismuth trioxide monohydrate [*1304-76-3*], $Bi_2O_3 \cdot H_2O$, is a white powder. Its composition and constitution are not completely elucidated.

$Bi(OH)_3$ and BiOOH probably do exist. A hydrated oxide precipitate is formed by reaction of an acidic bismuth nitrate solution with ammonia or caustic soda solution [57].

Bismuth pentaoxide [*35984-07-7*], Bi_2O_5, is prepared by oxidizing suspended Bi_2O_3 with Cl_2, Br_2, or H_2O_2 in an aqueous alkaline solution. Each of these reactions goes almost to completion. Molten mixtures of Bi_2O_3 and alkali-metal oxides can be oxidized by air or oxygen. The reaction is favored by excess alkali. The excess alkali can be stripped with methanol at 0 °C to form yellowish-orange sodium metabismuthate, $NaBiO_3$. Addition of nitric or perchloric acid leads to a compound whose composition varies from $Bi_2O_4 \cdot x\,H_2O$ to $Bi_2O_5 \cdot x\,H_2O$. Both compounds are strong oxidants.

Bismuth Trisulfide, Bi_2S_3. Naturally occurring as bismuth glance. Dark brown to grayish black, metallic luster, rhombic. Prepared by heating bismuth with sulfur or bismuth trioxide with sulfur compounds or by precipitation from aqueous bismuth(III) solutions with H_2S. It is insoluble in water and alkaline solutions but soluble in hot aqueous salt solution or concentrated nitric acid. Concentrated alkaline metal sulfide solutions or melts dissolve Bi_2S_3 to form compounds of the type $KBiS_2$. Similar compounds naturally occur with copper, silver, and lead.

7.2. Organic Compounds

There are numerous stable organobismuth compounds [3], [58–64]. The bismuth is either trivalent or pentavalent. The compounds are more subject to thermal decomposition than the corresponding phosphorus, arsenic, and antimony compounds, and they are more toxic. Both statements are especially true for halogeno- or pseudohalogenoorganobismuth compounds.

Trivalent Organobismuth Compounds (Bismuthines). Symmetrical tertiary bismuthines, R_3Bi where R is alkyl or aryl, are best prepared by reaction of organometallic reagents, usually Grignard compounds, with anhydrous $BiCl_3$ or $BiBr_3$ in a dry inert atmosphere. Although triarylbismuthines are stable, all trialkylbismuthines except trimethylbismuthine [*593-91-9*] are self-igniting in air. Tertiarybismuthines react exothermically with halogens, and they decompose in thepresence of water or alkaline solutions. Inorganic bismuth salts are formed in the presence of strong mineral acids.

Halo- and pseudohaloorganobismuthines are sensitive to oxygen and moisture and tend to self-ignition. They are produced by reaction of tertiary bismuthines with $BiCl_3$. Generally they are crystalline solids melting between 180 and 205 °C.

Pentavalent Organobismuth Compounds. After the tertiary bismuthines, the triarylbismuth dihalides are the most important organobismuth compounds. Only the chlorides and bromides can be prepared in high yields by oxidation of triarylbismuthines with the stoichiometric amount of chlorine or bromine. To prevent decomposition this has to be done at 0°C. Other triarylbismuth dihalides cannot be produced this way because of instant decomposition. Instead they are produced by reaction of triarylbismuth dichlorides or dibromides with alkali-metal or silver halides.

8. Chemical Analysis

For qualitative determination bismuth [65] is precipitated as Bi_2S_3 from slightly acidic solution with H_2S. The precipitate is brown black and soluble in strong acids and hot dilute nitric acid. Yellow-green Bi_2S_3 precipitates from alkaline solution on addition of Na_2S. Bismuth can be precipitated by hydrolysis from hot dilute hydrochloric acid as bismuth oxychloride, provided excess chloride ions are absent. Aqueous ammonia precipitates bismuth as white bismuth oxide hydrate, which can be reduced to bismuth metal with alkaline stannate(II) solution.

Bismuth phosphate, $BiPO_4$, can be precipitated quantitatively from very dilute nitric acid solution. Bismuth in medicinal preparations is determined by precipitation as basic bismuth carbonate. Calcining at 550 °C yields Bi_2O_3. A thiourea – bismuth mixture is of yellow color in nitric acid solution, and this is the basis for a photometric method. If ions such as Fe, Hg, Zr, etc., are absent, titration with EDTA is a selective, quantitative, and fast method of bismuth

analysis. This method has become very popular [66]. Emission, mass, or atomic absorption spectrometry allow detection of less than 1 ppm Bi.

9. Economic Aspects

Bismuth concentrates are a byproduct of nonferrous metal ore dressing. The metal available from concentrates has remained nearly the same, between 3600 and 4600 t/a since the 1960s. The world's one and only bismuth mine, Telamayu in Bolivia, which had produced about 15 % of the worldwide supply, stopped production in 1980 because of either declining demand or perhaps total exhaustion of the site. Also Australia and Peru are reducing their production. Only a part of bismuth refining is carried out in the mining countries. Table 11 shows the metal content of bismuth ores and concentrates, corresponding closely to the world production of 3795 t in 1980. The main producer of bismuth metal is the United States, producing $\approx$ 950 t/a [68]. Bismuth suppliers are listed in [70].

Bismuth (99.99 %) is quoted in London and New York in \$/lb (1 lb = 0.454 kg). After rising to 9 \$/lb in 1974 the price of bismuth fell steadily for years. The price was \$ 1.67 per pound in December, 1983. However, since then the price has increased, to \$ 6.60 in February, 1985. The prices are much higher for high-purity bismuth (99.999 %). Table 12 shows actual U. S. bismuth consumption. No consumption data for the European market have been published for several years. Because of substitution and regulatory restrictions bismuth consumption has decreased more in Europe than in the United States.

10. Toxicology and Occupational Health

Toxicology. Poisoning by bismuth [*7440-69-9*] and bismuth compounds has occurred more frequently during medical therapy using soluble compounds than by exposure at the workplace [71], [72].

Bismuth poisoning resembles in part poisoning caused by lead and mercury and their compounds. After oral administration, water-insoluble bismuth compounds such as bismuth nitrate oxide and bismuthyl nitrate [*1304-85-4*], $BiO(NO_3)$, are hardly absorbed, and acute poisoning is seldom. Increased absorption was observed, however, when large quantities of such compounds were applied to skin lesions, resulting in serious poisoning, sometimes fatal. Water-soluble bismuth compounds are absorbed quickly, and acute poisoning is likely to occur. A significant biochemical effect of bismuth incorporation is the increase in metallothionein concentration that occurs in rats after subcutaneous administration of bismuth chloride [*7787-60-2*] atdoses of 1 x 3 mg or 7 x 3 mg Bi/kg body weight, influencing the renal andhepatic levels of Cu and Zn [73]. There are indications that pretreatment with bismuth compounds protects mice and rats against the toxic effects of anticancer drugs [74].

The main hazard of bismuth is the chronic exposure that took place during long-term therapy of gastric ulcera (up to 30 years in some cases) practiced earlier in Australia and France. Such exposure can have serious consequences; even deaths can occur [75]. However, the relatively high toxicity of bismuth nitrate oxide and bismuth subnitrate [*1304-85-4*], which was observed especially in children, is most likely not caused by bismuth but by reduction of nitrate to nitrite by intestinal bacteria and subsequent methemoglobin formation [71], [72].

Initial symptoms of chronic intoxication by large amounts of bismuth compounds are hypersalivation, stomatitis, and a grayish blue seam surrounding the gums (symptoms similar to those caused by lead). According to many investigations [75–77], long-term oral administration caused encephalopathies accompanied by confusion, distractions, ataxia, myoclonic spasms, dysarthria, insomnia, damage to the kidneys (albuminuria, nephritis) and headache in many cases, and less frequently epileptic attacks. Some cases ended in death, the brains of the deceased containing large quantities of lipid-soluble bismuth compounds [78]. Symptoms were mostly reversible within several weeks or months after cessation of bismuth intake. The crossing of the blood – brain barrier by bismuth compounds was also demonstrated in other experiments, e.g., for trivinylbismuth in pigs [79]. In another epidemiological study, a group of patients with manifest bismuth poisoning was compared to a control group that had been exposed to the same amount of bismuth but that

Table 11. Bismuth in ores and concentrates, t [67–69]

	1970	1973	1977	1980
Australia	363	500	912	907
Bolivia	567	690	680	11
Canada	272	300	165	171
China, People's Republic of	200 ∗	250	249	259
France	59	65	66	59
Germany, Federal Republic of	18	20	10	9
Italy	23	25	15	10 ∗
Japan	68	75	80 ∗	80 ∗
Mexico	453	600	729	748
Peru	816	690	590	522
Romania	9	10	80	80 ∗
South Korea	136	150	134	90 ∗
Sweden	32	35	15	15 ∗
USSR	50∗	50	66	73
United States	363	400	400	400 ∗
Yugoslavia	108	120	74	73
Others	204	320	399 ∗∗	215∗∗
Total	3741	4300	4664	3722

∗ Estimated.
∗∗ Calculated.

Table 12. Bismuth consumption in the United States, t [68]

	1970	1973	1978	1980
Chemical industry∗	567	543	521	556
Fusible alloys	292	423	379	291
Metallurgical additions	236	376	210	211
Other alloys	6	7	10	11
Others	5	5	8	7
Total	1106	1354	1128	1076

∗ Including pharmaceutical products and cosmetics

did not show any symptoms of poisoning [80]. The investigators concluded that bismuth poisoning is caused by the transformation of low-toxicity bismuth compounds into highly toxic compounds by intestinal bacteria. Bismuth toxicity is enhanced significantly when applied intraperitoneally as a Bi – L-cysteine complex to mice [81].

Apparently, symptoms of neurotoxicity caused by bismuth compounds such as tripotassium dicitrato bismuthate (de-nol) [*57644-54-9*] [82], colloidal bismuth subcitrate [*57644-54-9*], and bismuth subnitrate [83] were causedby an excessive intake, since the normal clinical use of these two compounds in the treatment of gastric ulcera did not cause signs of neurotoxicity, as shown by clinical and neurophysiological tests that included magnetic resonance imaging, nerve-conduction studies, visual evoked responses, electroneurography, and electroencephalography. However, decreased kidney function may increase the risk of bismuth intoxication, even by therapeutic doses of bismuth compounds [84].

The occurrence of osteolytic and osteonecrotic arthropathies of the shoulder was pointed out in [85]. In over 59 cases of bismuth poisoning, bone lesions accompanied by encephalopathies were described [86]. As a result of the elimination of bismuth compounds in the urine, albuminuria and nephritis were also observed [72]. Bismuth also causes eye defects [87], which was confirmed by experiments with mice [88]. Inhalation of trimethylbismuth [*593-91-9*] by humans causes irritation of the respiratory tract and conjunctiva. Cats and dogs that inhaled trimethylbismuth for 10 – 20 min showed ataxia, restlessness, and convulsions; after 24 h severe encephalitis was observed [89]. Furthermore, the skin shows pigmentation, purulent pustules, and severe dermatitis.

For the treatment of bismuth poisoning, a series of heavy-metal chelators that enhance the excretion of bismuth is avail-

able. On the basis of experiments on rats, the chelators of choice for clinical practice are *meso*-2,3-dimercaptosuccinic acid (DMSA) and D,L-2,3-dimercaptopropane-1-sulfonic acid (DMPS) [90]. However, only in a small number of cases is the application of chelators such as DMPS or D-penicillamine actually necessary [72].

In a long-term carcinogenicity test, bismuth chloride oxide [*7787-59-9*], BiOCl, was fed to rats over a period of two years (1, 2, and 5 % in the feed) [91], and neither carcinogenic nor other toxic effects that were attributable to bismuth chloride oxide could be found. With the exception of rare single reports, there are no indications so far for carcinogenicity, mutagenicity, immunotoxicity, or reproductive toxicity, including teratogenicity, of bismuth compounds: Bismuth trioxide caused chromosomal aberrations in mice after the intake of 400 – 1000 mg/kg body weight [92]. The following toxicological data (in mg/kg) are available for bismuth and bismuth compounds: bismuth, LDLo (human, presumably oral): 221 [93]; bismuth chloride oxide, LD_{50} (rat, oral): 22 000; trimethylbismuth, LD_{50} (rabbit, oral): 484, LD_{50} (rabbit, s.c.): 182, LDLo (rabbit, i.v.): 11 [89].

Occupational Health. Bismuth and its compounds are not included the list of German MAK values or in the lists of carcinogenic or sensitizing workplace compounds [94]. Occupational safety and health guidelines for (undoped) bismuth telluride [*1304-82-1*], Bi_2Te_3:OSHA PEL: 15 mg/m^3(total dust) and 5 mg/m^3(respirable fraction) as an 8-h TWA [95]; NIOSH Recommended Exposure Limit (REL): 10 mg/m^3(total dust) and 5 mg/m^3(respirable fraction) as an 8-h TWA [96]; ACGIH TLV: 10 mg/m^{3}as a TWA for a normal 8-h working day and a 40-h working week [97].

11. References

1. V. Tafel: *Lehrbuch der Metallhüttenkunde,* Hirzel Verlag, Leipzig 1951.
2. J. Feiser in: *Die metallischen Rohstoffe,* vol. 17: *Nebenmetalle,* Enke Verlag, Stuttgart 1966.
3. *Gmelin,* System No. 19, Main Volume (1927, reprinted 1971), Supplement Volume (1964), Organobismuth Compounds (1977).
4. C. A. Hampel: *Rare Metals Handbook,* 2nd ed., Reinhold Publ. Co., London 1961.
5. *Winnacker-Küchler,* 3rd ed., **vol. 6,** pp. 350 – 354.
6. *Kirk-Othmer,* **vol. 3,** pp. 912 – 937.
7. J. D. Verhoeven, E. D. Gibson, M. B. Beardsley, *Metall. Trans. B* **6 B** (1975) no. 2, 349.
8. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* Springer Verlag, Berlin – Göttingen – Heidelberg – New York 1967.
9. I. I. Mekler, V. M. Yashchenkova, A. A. Egizarov, E. A. Simkin, M. I. Gorodi, N. F. Plaksa, *Sov. J. Non Ferrous Met. (Engl. Transl.)* **12** (1971) no. 7, 51 – 53.
10. E. Borbasch, *Can. Min. J.* **82** (1961) no. 8, 66.
11. L. A. Chazova, E. N. Glazkov, S. A. Ovintseva, S. M. Barakov, A. V. Levchenko, *Sov. J. Non Ferrous Met. (Engl. Transl.)* **7** (1966) no. 1, 36 – 38.
12. Z. Tapia, *Perfil. Minero Metalurgico* 1979, no. 14, 13 – 20.
13. K. Z. Kuanysheva, S. T. Takeshanov, G. L. Pashkov, *Sov. J. Non Ferrous Met. (Engl. Transl.)* **14** (1973) no. 2, 31 – 32.
14. D. O. Rausch, B. O. Mariacher: *Lead & Zinc, Mining and Concentrating of Lead and Zinc,* AIME, New York 1970.
15. J. M. Cigan, T. S. Mackey, T. J. O'Keefe: *Lead-Zinc-Tin '80,* AIME, Warrendale 1979.
16. R. Z. Khobdabergenov, M. P. Smirnov, U. B. Kolosov, *Sov. J. Non Ferrous Met. (Engl. Transl.)* **8** (1967) no. 2, 54 – 55.
17. K. Emicke, G. Holzapfel, E. Kniprath, *Erzmetall* **24** (1971) 205 – 215.
18. Am. Smelt. Ref. Co., US 2056164, 1935.
19. C. H. Cotterill, J. M. Cigan: *Lead & Zinc, Extractive Metallurgy of Lead and Zinc,* AIME, New York 1970.
20. L. Marck, *Z. Erzbergbau Metallhüttenwes.* **14** (1961) 113 – 123.
21. E. R. Freni: "Symposium Hydrometallurgie", *Erzmetall* **22** (1969) B 128 – B 132.
22. D. A. Pazour, *World Min.* **32** (1979) no. 8, 40 – 45.
23. I. L. Barker, *J. Met.* **8** (1956) 1058. W. C. Smith, P. J. Hickey, *Trans. Metall. Soc. AIME* **224** (1962) 224.
24. A. K. Biswas, W. G. Davenport: *Extractive Metallurgy of Copper,* Pergamon Press, Oxford 1976.

25. R. Hanks, J. van der Zel, P. Chesney, G. B. Harris, *Trans. Inst. Min. Metall. Sect. C* **88** (1979) C 99 – C 106.
26. K. G. Lombeck, J. Krüger, H. Winterhager, *Metall (Berlin)* **36** (1982) 1192 – 1196.
27. J. E. Joffre, E. R. Padilla, *Rev. Metall. (Bolivia)* 1975, no. 2, 10 – 13.
28. *Min. Mag.* 1976 (May) 369 – 374.
29. R. Kammel, H. Mirafzall, *Erzmetall* **30** (1977) 437.
30. J. A. Villaroel, *Rev. Metall. (Bolivia)* 1977 , no. 3, 35.
31. P. Paschen, *Metall (Berlin)* **33** (1979) no. 2, 137.
32. E. Müller, P. Paschen, *Erzmetall* **32** (1979) no. 6, 266 – 272.
33. M. Marone, R. Tarrico, *Metall. Ital.* **62** (1970) 482 – 484.
34. V. A. Shavga, V. N. Vigdorovich, J. U. Samson, *Izv. Akad. Nauk SSSR Met.* 1970, no. 5, 101 – 109. Yu. K. Delimarskii, A. A. Omel'schuk, O. G. Zarubitskii, *Ukr. Khim. Zh. (Russ. Ed.)* **42** (1976) no. 11, 1202 – 1204.
35. B. N. Aleksandrov, V. I. Udovikov, *Izv. Akad. Nauk. SSSR Met.* 1973, no. 1, 57 – 62.
36. B. N. Aleksandro, V. I. Udovikov, L. E. Usenko, *Russ. Metall. (Engl. Transl.)* 1969, no. 6, 57 – 63.
37. V. V. Marychev, A. P. Kolesnev, V. I. Panomarenko, V. N. Vigdorovich, *Tsvetn. Met.* 1973, no. 9, 9 – 11.
38. *Alloy Digest* Feb. 1978, Bi 3; Sept. 1978, Bi 7; Oct. 1978, Bi 8; Nov. 1980, Bi 20. *Metals Handbook,* Am. Soc. of Metals, Cleveland, Ohio, 1948, p. 744. M. Hansen: *Constitution of Binary Alloys,* McGraw Hill, New York 1958.
39. F. A. Eich, *J. Prosthet. Dent.* **1** (1951) 594.
40. D. A. Bancel, *Eng. Min. J.* **172** (1971) no. 3, 115 – 116.
41. W. C. Cooper: *Tellurium,* Van Nostrand-Reinhold, New York 1971, pp. 333 – 353. T. Durst, H. J. Goldsmid, L. B. Harns, *Sol. Energy Mater.* **5** (1981) no. 2, 181 – 186.
42. S. C. Harvey in L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* MacMillan, New York 1975, p. 930.
43. D. Fitzgerald, *Eng. Min. J.* **183** (1982) no. 3, 104.
44. H. A. Skinner, L. E. Sutton, *Trans. Faraday Soc.* **36** (1940) 681.
45. H. M. Haendler, F. A. Johnson, D. S. Crocket, *J. Am. Chem. Soc.* **80** (1958) 2662.
46. V. Gutmann, *Monatsh. Chem.* **82** (1951) 280.
47. C. Frondel, *Am. Mineral.* **28** (1943) 521.
48. H. Hecht, R. Reissner, *Z. Anal. Chem.* **103** (1935) 186.
49. F. Paneth, *Ber. Dtsch. Chem. Ges.* **51** (1918) 1704, 1728; **55** (1922) 769; **58** (1925) 1138.
50. E. Amberger, *Chem. Ber.* **94** (1961) 1447.
51. B. Aurivillius, *Acta Chem. Scand.* **9** (1955) 1206.
52. F. Hund, *Z. Anorg. Allg. Chem.* **258** (1949) 198.
53. J. Fischer, E. Rudzitis, *J. Am. Chem. Soc.* **81** (1959) 6375.
54. S. Y. Yosim et al., *J. Phys. Chem.* **66** (1962) 28.
55. E. Dönges, *Z. Anorg. Allg. Chem.* **263** (1950) 112.
56. G. Gattow, G. Kiel, *Z. Anorg. Allg. Chem.* **335** (1965) 61.
57. P. Kirkov, *Ber. Chem. Ges. Belgrad* **22** (1957) 431.
58. G. O. Doak, L. D. Freedman: *Organometallic Compounds of Arsenic, Antimony, and Bismuth,* Wiley, New York 1970.
59. A. E. Goddard in J. N. Friend (ed.): *A Textbook of Inorganic Chemistry,* vol. XI/III: *Derivatives of Phosphorus, Antimony and Bismuth,* Ch. Griffin, London 1936.
60. M. Dub: *Organometallic Compounds,* 2nd ed., **vol. III,** Springer Verlag, New York 1968; 1st Suppl., 1972.
61. F. G. Mann: *The Heterocyclic Derivatives of Phosphorus, Arsenic, Antimony, and Bismuth,* Wiley-Interscience, New York 1970.
62. H. Gilman, H. L. Yale, *Chem. Rev.* **30** (1942) 281.
63. P. G. Harrison, *Organomet. Chem. Rev. Sect. A* **5** (1970) 183.
64. *Houben-Weyl,* 4th ed., **vol. XIII/8,** 590 – 627.
65. *Analyse der Metalle,* vol. II/2, Springer Verlag, Berlin – Göttingen – Heidelberg 1961.
66. G. Schwarzenbach, H. Flaschka: *Die komplexometrische Titration,* Enke Verlag, Stuttgart 1965.
67. P. Zilveti, *Eng. Min. J.* **175** (1974) no. 2, 122.
68. *Metal Bull. Handbook,* 14th ed., Metal Bulletin Books, London 1981.
69. *Metallstatistik 69,* Metallgesellschaft AG, Frankfurt/Main 1982.
70. R. Serjeantson, R. Cordero (eds.): *Nonferrous Metal Works of the World,* 3rd ed., Metal Bulletin Books, London 1982.
71. M. J. Ellenhorn, D. G. Barceloux: *Medical Toxicology, Diagnosis and Treatment of Human Poisoning,* Elsevier, New York 1988, p. 1018.

72. W. Wirth, C. Gloxhuber: *Toxikologie,* Thieme Verlag, Stuttgart – New York 1981, p. 140.
73. J. A. Szymanska, J. Chmielnicka, A. Kaluzynski, W. Papierz, *Biomed. Environ. Sci.* **6** (1993) 134.
74. J. S. Lazo, "Metals and Anticancer Drug Resistance", in R. A. Goyer, M. G. Cherian, (eds.): *Toxicology of Metals – Biochemical Aspects* Springer, Berlin, Heidelberg, New York, London 1995, p. 267.
75. G. Martin-Bouyer, *Therapie* **31** (1976) 683.
76. A. Buge, G. Rancurel, H. Dechy, *Rev. Neurol.* **133** (1977) 401.
77. V. S. Viterbo, C. Sicard, H. P. Cathala, *Rev. Electroencephalogr. Neurophysiol. Clin.* **7** (1977) no. 2, 139.
78. J. L. Ribadeau Dumas, B. Lechevalier, M. Breteau, Y. Allain, *Nouv. Presse Med.* **7** (1978) 4021.
79. S. Pollet, S. Albouz, F. Le Saux, N. Baumann, R. Bourdon, *Toxicol. Eur. Res.* **2** (1979) 123.
80. G. Martin-Bouyer, G. Foulon, H. Guerbois, C. Barin, *Clin. Toxicol.* **18** (1981) 1277.
81. N. Krari, Y. Mauras, P. Allain, *Res. Commun. Mol. Pathol. Pharmacol.* **89** (1995) 357.
82. C. U. Nwokolo, J. D. Fitzpatrick, R. Paul, R. Dyal, B. J. D. E. Loft, *Aliment. Pharmacol. Ther.* **8** (1994) 45.
83. L. A. Noach, J. L. Eekhof, L. J. Bour, F. E. Posthumus-Meyes, G. N. Tytgat, B. W. Ongerboer de Visser, *Hum. Exp. Toxicol.* **14** (1995) 349.
84. R. J. Playford, C. H. Matthews, M. J. Campbell, H. T. Delves, K. K. Hla, H. J. F. Hodgson, J. Calam, *Gut* **31** (1990) 359.
85. A. Buge, A. Hubault, G. Rancurel, *Rev. Rhum. Mal. Osteo-Articulaires* **42** (1975) 721.
86. J. Emile, J. M. de Bray, M. Bernat, T. Morer, P. Allain, *Clin. Toxicol.* **18** (1981) 1285.
87. A. Neetens, J. Martin, Y. Hendrata, M. C. Rubbens, *Bull. Soc. Belge Ophtalmol.* **177** (1975) 51.
88. P. Lechat, J. Fontagne, N. Letteron, *Ann. Pharm. Fr.* **34** (1976) 173.
89. T. Sollmann, J. Seifter, *J. Pharmacol. Exp. Ther.* **67** (1939) 17.
90. A. Slikkerveer, H. B. Jong, R. B. Helmich, F. A. de Wolff, *J. Lab. Clin. Med.* **119** (1992) 529.
91. R. Preussmann, S. Ivankovic, *Food Cosmet. Toxicol.* **13** (1975) 543.
92. N. Gurnani, A. Sharma, G. Talukder, *Biol. Trace Elem. Res.* **37** (1993) 281.
93. J. M. Arena: *Poisoning; Toxicology, Symptoms, Treatments,* 2nd ed., C. C. Thomas, Springfield, Ill., 1970, p. 73.
94. Deutsche Forschungsgemeinschaft (ed.): *MAK- und BAT-Werte-Liste, Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte, Mitteilungen der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 36,* VCH, Weinheim, 1998.
95. OSHA: Computerized Information System. Washington, DC, U.S. Department of Labor, 1990.
96. NIOSH Recommendations for Occupational Safety and Health: Compendium of Policy Documents and Statements; Cincinnati, OH, U.S. Department of Health and Human Services, Public Health Service, DHHS (NIOSH) Publication No 92-100 (1992).
97. ACGIH: 1991 – 1992 Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices, Cincinnati, OH, 1991.

Bisphenols → Phenol Derivatives
Black Powder → Pyrotechnics
Blanc Fixe → Barium and Barium Compounds
Blast Furnace Process → Iron

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Bleaching

Standard Article

Hans Ulrich Süss[1]
[1]Degussa AG, Hanau-Wolfgang, Germany

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a04_191.pub2
Article Online Posting Date: December 15, 2006

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Bleaching Agents
2.1. Pretreatment
2.2. Oxidizing Agents
2.2.1. Oxygen
2.2.2. Chlorine Dioxide and Sodium Chlorite
2.2.3. Hydrogen Peroxide and Other Peroxo Compounds
2.2.4. Ozone
2.2.5. Chlorine and Hypochlorites
2.3. Reducing Agents
2.3.1. Sodium Dithionite
2.3.2. Sulfur Dioxide
2.3.3. Sulfinic Acid Derivatives
3. Enzymes
4. Bleaching Processes
4.1. Pulp
4.1.1. Chemical Pulp
4.1.2. Mechanical Pulp
4.1.3. Paper Recycling (Deinking)
4.1.4. Environmental Aspects of Pulp Bleaching
4.2. Textile Fibers
4.3. Other Products

Bleaching is the removal of color from various materials. It is predominantly achieved by an oxidative or reductive treatment, which destroys chromophores. In principle all oxidizing or reducing chemicals can act as bleaching agents, but on an industrial scale only a few compounds are used in bleaching processes. Chlorine dioxide, hydrogen peroxide, and sodium dithionite are predominantly applied. The alternative to a chemical reaction is the adsorption of colored impurities on pigment surfaces, like active carbon or bleaching clays. This method is applied mainly in purification or bleaching of liquids.

A bleaching process removes impurities; thus, in parallel the bleached good is cleaned. The “art of bleaching” is to reach a desired decrease in color with the least negative impact on quality and properties of the bleached material. Chemicals and processes are selected on the basis of their ability to achieve brightness and quality targets with a small amount of byproducts or residuals and low impact on the environment.

Bleaching of fibers for paper production is the most important industrial application. Highest effectiveness is guaranteed by a combination of several process steps in a bleach plant. Textile fabric bleaching is the second most important application. Modern bleaching technology aims for the most moderate use of resources, the lowest consumption of water, and best biodegradability of the compounds dissolved during the bleaching process.

[Top of Page]

1. Introduction

A bleaching effect is achieved by a removal of chromophores. This can be achieved by adsorption of the colored compounds on active surfaces like active carbon or bleaching clays. On an industrial scale this option is used to bleach and purify liquids like vegetable oil. Bleaching by chemical destruction of chromophores in organic or inorganic compounds is more common. Industrially, decolorization is often combined with purification or cleaning. Components affecting material quality, handling properties, or marketability of a product are removed. This is especially important for natural products, most of which contain impurities. For example, cotton is contaminated with seed-capsule residues, wax, fatty constituents, and inorganic compounds. Bleaching cotton increases brightness while simultaneously removing impurities. Bleaching also can continue or supplement an earlier processing step. For example, in continuation of chemical pulping bleaching removes lignin from wood fibers, thus brightening and further purifying the cellulose.

Very few compounds can be bleached nearly colorless without damage. The “art of bleaching” is to achieve a high degree of chromophore removal without damaging the bleached good. The aim of bleaching from the outset may also be to generate simply a more uniform hue, e.g., bleaching wood veneer.

In terms of the amount of material processed and the value of the product, bleaching of fibrous materials for paper and textile production is most important (Table 1). Detailed information about these processes is available under specific keywords (⟶ Paper and Pulp or ⟶ Textile Dyeing) and in [1], [2], [3]. Table 2 gives a summary of products bleached on industrial scale.

Table 1. World production of bleached fiber material (10^3 t)

Material	Year	Amount
Bleached chemical pulp	2004	80 000
Bleached mechanical pulp	2004	30 000
Cotton*	2004/2005	25 000
Wool (scoured)*	2003	1 260

*Dying is an important part of textiles manufacturing. A prerequisite of the dying process is a certain level of cleaning and bleaching in order to achieve uniform color distribution. It can be assumed that the predominant amount of fiber used for textiles production has been bleached during manufacturing.

Table 2. Important products requiring bleaching

Fibers for paper	
Chemical pulp,	
Mechanical pulp,	
Wastepaper, linters	
Fibers for textiles	
Natural products:	cotton, flax, wool, silk
Synthetics:	rayon, polyamide, polyester, polyacrylonitrile, polyurethane
Others	
Vegetable products:	fatty acids, resins, wax, starch, flour, vegetable oils, wood, wood veneer, peeled rattan
Animal products:	fatty acids, hair, bristle, feathers, horn, sponges, mother-of-pearl
Mineral and synthetic products:	paraffin, petroleum, glues, plastics, soap, sulfuric acid

Bleaching can be classified into oxidative and reductive bleaching. Oxidative bleaching destroys chromophores irreversibly, while reductive bleaching can be reversible, for example, by reoxidation by atmospheric oxygen.

Especially in the case of natural products, the chemical constitution of the colored compounds is not known in full detail. Generally these compounds contain conjugated carbon – carbon double bonds, e.g., polymethine chains, aromatic systems, or quinoid systems. Oxidative bleaching destroys the double bonds, removes the conjugation, and produces chemical compounds of lower molecular mass. Effective bleaching frequently is a multistep process. The repetition of acidic oxidation

and alkaline extraction allows moderate conditions and a low concentration of chemicals to be maintained.

The bleaching effect is measured by the increase in brightness. Remission photometers are used to determine brightness relative to a standard at several wavelengths. $BaSO_4$ is the most widely used standard. However, there are several similar methods that may be used to determine brightness [4], [5]. The quality and efficiency of bleaching is also reflected by brightness stability. Excessive oxidation can generate new chromophores or the potential for later yellowing. Yellowing takes place if the auxochrome groups in a chromophore can be reoxidized. Light and heat may change the color of bleached materials. Carbon – carbon bonds are readily cleaved by UV radiation to give radicals. Recombination of these radicals with oxygen or other parts of the molecule can form new chromophores.

In a bleaching processes the degree of polymerization of the polymers treated must not be affected negatively. A decrease in molecular mass of the polymers may affect fiber strength and reduce the economic value. Therefore, bleaching should destroy chromophores but not affect the molecular structure. A wide variety of bleaching methods have evolved because of the various complications resulting from the differences in the raw materials. An example is bleaching of cotton in comparison to wool. While strong alkaline conditions can be applied in bleaching the carbohydrate polymer cellulose (cotton), the keratin (protein) wool fiber is very sensitive towards alkali. Extraction of amino acids would destroy the polymer, so alkaline bleaching cannot be used.

[Top of Page]

2. Bleaching Agents

Detailed descriptions of the production and properties of bleaching chemicals are given under the specific keywords (for example, ⟶ Chlorine). The following sections are limited to a description of the applications and the advantages or disadvantages of the major bleaching agents.

2.1. Pretreatment

The removal of contaminants is part of the preparation for a bleaching process. A typical pretreatment might simply be washing. Sheared wool can contain up to 50 wt % of sand and fatty material (lanolin). Similarly, compounds consuming or decomposing the bleaching agent must be eliminated before the actual bleaching process. Washing ahead of bleaching in the presence of surfactants or chelating compounds is a common procedure. This decreases the demand for bleaching agent and lowers the potential for negative side reactions during bleaching. In chemical pulp bleaching this can be an important prerequisite of bleaching.

Acid washing below pH 3 solubilizes metal ions. The presence of reducing agents such as sulfur dioxide to lower the oxidation state of the metals can improve the effect. Chelating agents like the sodium salts of diethylenetriaminepentaacetic acid [140-01-2], DTPA, or ethylenediaminetetraacetic acid [64-02-8], EDTA, are used under mildly acidic conditions (pH 5–6) to accelerate the elimination of metals. This pH range allows the removal of transition metals with conservation of magnesium and calcium ions, which are of benefit as stabilizers in bleaching. For chemical pulp produced by alkaline pulping, such a metals-removal stage is necessary if totally chlorine free (TCF) bleaching is to be applied. In conventional or ECF (elemental chlorine free) bleaching metals are removed during the acidic chlorine dioxide stages. In acid sulfite pulping they are solubilized during the pulping process and removed in brown stock washing.

During alkaline pulping of hardwood xylan polymer double bonds are generated. These hexenuronic acids consume electrophilic bleaching chemicals. They are hydrolyzed in a hot acid pretreatment below pH 3 at about 90 °C to yield furancarboxylic acid [6]. This decreases the demand for bleaching chemicals.

Enzymes like xylanase may be applied as pretreatment. During alkaline pulping xylan dissolves but may precipitate again in the final steps of the cooking procedure. This can cause difficulties in bleaching. Xylanase removes the precipitate and facilitates bleaching. Preferred products are enzymes extracted from thermophilic bacteria with high temperature and pH tolerance.

In bleaching of textiles, treatment with amylase or peroxosulfate may be necessary to remove starch.

2.2. Oxidizing Agents

The ancient Egyptians and Phoenicians practiced oxidative bleaching of textiles in sunlight combined with a mild alkaline treatment using potassium carbonate leached from wood ashes. In the 1700s bleaching of textile fabrics by an alternating treatment with potassium carbonate, sunlight on grassy meadows, and sour milk was known as “Dutch bleaching”, because it reached its greatest perfection around Haarlem, in the Netherlands.

The increase in knowledge about chemistry in the late 1700s yielded a new bleaching agent, hypochlorite. For nearly 150 years the hypochlorites of potassium [7778-66-7] (Eau de Javelle), sodium [7681-52-9] (Eau de Labarraque), and calcium [7778-54-3] developed into the most important industrial bleaching agents. Chlorine was introduced in chemical pulp bleaching in the 1930s. Chlorine dioxide and hydrogen peroxide followed a decade later.

Industrial use of oxygen as a bleaching agent for chemical pulp started in the late 1960s. Ozone [10028-15-6] was later added to chemical pulp bleaching sequences. Other oxidizing compounds used for bleaching are derivatives of hydrogen peroxide, for example, peracetic acid, sodium perborate, peroxodisulfates (persulfates), and peroxomonosulfates (Caro's acid). These compounds are applied only under special conditions. This also applies for potassium permanganate [7722-64-7].

2.2.1. *Oxygen (⟶ Oxygen)*

The use of oxygen [7782-44-7] as a bleaching agent has increased sharply since the early 1970s. It is used in chemical pulp

bleaching to continue and intensify the lignin degradation process which normally starts with alkaline pulping. Oxygen delignification needs alkaline conditions (normally NaOH) for the oxidation process, a rather high temperature (75 – 100 °C), and an oxygen pressure of 0.3 – 0.5 MPa. Increasing the temperature, the pressure, or the caustic soda concentration increases the rate of oxidation. Oxygen delignification involves two simultaneous reactions: the desired delignification and undesired carbohydrate degradation. Therefore practical limits exist for temperature increase and the amount of alkali. The decrease in the degree of polymerization is inhibited to some extent by magnesium salts.

Although the effect of oxygen on chemical pulp is limited, treatment with oxygen decreases the demand for other bleaching chemicals by 40 – 50 %. Oxygen delignification allows the combined recovery of the effluents from pulping and the subsequent delignification step, effectively lowering the total amount of organic material discharged from the bleaching plant.

For magnesium-based pulping processes (magnesium sulfite pulping), the oxygen stage can be operated with magnesium hydroxide or oxide as the alkali source. Typically, hydrogen peroxide is added simultaneously (at about 90 °C) to increase the rate of delignification.

2.2.2. *Chlorine Dioxide and Sodium Chlorite (* ⟶ Chlorine Oxides and Chlorine Oxygen Acids*)*

Chlorine dioxide [10049-04-4] is a very selective bleaching agent. The first industrial use in bleaching chemical pulp was in 1946.

Chlorine dioxide has an unpaired electron; therefore, it is a free radical and has limited stability. It is a gas under standard conditions and explodes spontaneously at partial pressures above 40 kPa, decomposing into chlorine and oxygen. If heated or exposed to light, it can explode even at lower concentrations, but with less violence. For this reason, pure chlorine dioxide cannot be transported and stored. It is generated on site and normally handled as dilute aqueous solution.

The basic reaction for industrial chlorine dioxide generation is the reduction of chlorate [14866-68-3]. Historically, reduction with sulfur dioxide, the Mathieson process, is important:

$$2\,NaClO_3 + SO_2 \xrightarrow{H_2SO_4} 2\,ClO_2 + Na_2SO_4$$

Dilute sulfuric acid and sodium sulfate or sodium hydrogen sulfate are the byproducts. Other processes, which produce fewer byproducts, use hydrochloric acid, methanol, or hydrogen peroxide as the reducing agent. The advantage of hydrochloric acid is very low costs, but the amount of chlorine present as a byproduct in the chlorine dioxide can be rather high (about 20%). In bleaching this results in increased formation of undesired halogenated products. Nowadays the methanol and hydrogen peroxide processes are the most important. In these processes very pure chlorine dioxide is generated as a gas, highly diluted with air or water vapor to avoid explosions. It is absorbed in cold water to give a concentration of 6 – 12 g/L.

For reactions of ClO_2 with pulp a pH value between 2 and 5 is normal. Within this pH range chlorine dioxide does not affect the degree of polymerization of cellulose. Typical reaction temperature for chlorine dioxide is between 50 and 75 °C. In hardwood pulp bleaching it is favorable to start bleaching at very high temperature (> 90 °C), as this allows the simultaneous oxidation of lignin and hydrolysis of hexenuronic acids.

In bleaching textiles sodium chlorite [7775-09-9] is used as the source of chlorine dioxide. Hydrochloric acid, formic acid [64-18-6], or acetic acid [64-19-7] is used to liberate ClO_2.

$$5\,NaClO_2 + 4\,HCl \rightarrow 4\,ClO_2 + 2\,H_2O + 5\,NaCl$$

2.2.3. *Hydrogen Peroxide and Other Peroxo Compounds (* ⟶ Hydrogen Peroxide*;* ⟶ Peroxo Compounds, Inorganic*;* ⟶ Peroxy Compounds, Organic*)*

Hydrogen peroxide [7722-84-1] has been used as an industrial bleach since the early 1900s. At first the main use was bleaching textile fibers. Sodium perborate [7632-04-4], $NaBO_3 \cdot 4\,H_2O$, prepared from sodium peroxide and borax, was used in the laundry detergent Persil as a bleach as early as 1907. Even today, perborates and percarbonates are used in large quantities in laundry detergents. Their mild bleaching effect at elevated temperatures is gentler to fibers and therefore superior to that of the harsher hypochlorites, which are used as so-called cold bleaches. Because of the lower storage stability of pure percarbonate, specially coated products are applied.

Aqueous solutions of hydrogen peroxide are available in concentrations up to 90 wt %. Solutions with 35 – 70 wt % are generally applied in bleaching. The most important application of hydrogen peroxide is bleaching of pulp for paper production. Mechanical pulp and chemical pulp are bleached with alkaline peroxide solutions. Hydrogen peroxide is also used for bleaching textile fibers such as cotton, flax, wool, and synthetics.

Commercial processes use slightly to strongly alkaline conditions for hydrogen peroxide bleaching. Usually sodium hydroxide is the source of alkali. The initial reaction is the nucleophilic addition of hydroperoxide anions to carbonyl groups [7]. Therefore, the bleaching rate is increased by addition of alkali:

$$HOOH + OH^- \rightleftharpoons HOO^- + H_2O$$

However, alkaline conditions also favor side reactions like the disproportion of hydrogen peroxide into oxygen and water:

$$HOO^- + H_2O_2 \rightarrow H_2O + OH^- + O_2$$

The bleaching result depends on activation and stabilization. The reaction rate increases with temperature and alkali concentration. However, side reactions such as the thermal decomposition of hydrogen peroxide into radicals become more

likely or even dominant at high temperatures, which also lead to yield losses due to solubilization of the material being bleached. Consequently the normal temperature range for peroxide bleaching is between 60 and 90 °C.

Stabilization of the alkaline bleaching liquor may be required to minimize peroxide decomposition. Hydroxyl radicals, generated by the decomposition of peroxide, are highly unselective oxidizing agents that produce undesired side reactions. In chemical pulp bleaching the chelation of transition metals is only required in TCF bleaching. In ECF bleaching hydrogen peroxide typically is applied only following an acidic stage. This removes transition metals in the washing step. In mechanical pulp bleaching, detrimental metals are best removed by pretreatment with a chelating agent. Typical chelating agents are the sodium salts of diethylenetriaminepentaacetic acid (DTPA), ethylenediaminetetraacetic acid (EDTA), or the corresponding phosphonates, which have similar or even better chelating ability. In addition sodium silicate is applied as a stabilizer which buffers the bleaching liquor. Similarly, magnesium salts (mostly magnesium sulfate, Epsom salt) are used. The effect of the stabilizers is to precipitate or absorb detrimental compounds on colloidal silica or magnesium hydroxide.

In the bleaching process hydrogen peroxide cleaves side chains and destroys quinoid structures irreversibly. The oxidation and destruction of quinones is an important prerequisite for achieving good brightness stability. Peroxide-bleached products show lower reversion compared to other oxidants.

Peracetic acid [79-21-0], CH_3COOOH, is used in small amounts as an activation step prior to alkaline hydrogen peroxide treatment. Its application allows a higher brightness increase or a lower total consumption of bleaching chemicals. The reaction with pulp is conducted under mild acidic conditions; it is predominantly a hydroxylation reaction of the residual lignin. Peracetic acid is produced by mixing acetic acid and hydrogen peroxide. In the presence of a strong acid (e.g., sulfuric acid) an equilibrium between acetic acid, peracetic acid, hydrogen peroxide, and water is rapidly established. As the other compounds in the mixture do not contribute to the bleaching reaction of peracetic acid the application of equilibrium peracetic acid is expensive. An option to overcome this problem is the use of distilled peracetic acid. Vacuum distillation of equilibrium peracetic acid yields a rather pure peracetic acid solution in water. Distilled peracetic acid with a content of about 35 % in water should be stored only below 4 °C, because at higher temperature the equilibrium with hydrogen peroxide and acetic acid is again slowly established. An alternative is on-site mixing, which yields a peracetic acid/peroxide mixture which can be used for an initial peracetic reaction and a subsequent peroxide bleaching step. This is achieved by adding sufficient alkali after the acidic stage.

Peracetic acid is generated in situ from hydrogen peroxide and tetraacetylethylendiamine (TAED) [10543-57-4] under weakly alkaline conditions. This procedure is applied in laundering agents with percarbonate or perborate as peroxide source. It is also applicable in bleaching of specialties.

Potassium peroxosulfate [37222-66-5], $KHSO_5$, is used for bleaching polyamide and cellulose fibers. However, it is ordinarily used only to clean wool and to reduce its shrinkage. Sodium peroxodisulfate [7775-27-1], $Na_2S_2O_8$, and ammonium peroxodisulfate [7727-54-0], $(NH_4)_2S_2O_8$, are used to desize cotton and sometimes as an additive in hydrogen peroxide bleaching.

2.2.4. *Ozone (⟶ Ozone)*

Ozone [10028-15-6] is produced from oxygen by silent electrical discharge. For bleaching pulp, ozone is typically produced from pure oxygen. This allows a high ozone concentration in the gas (up to 12 wt %) and facilitates the mixing procedure. Ozone is used at a pH below 4. Temperature in ozonation must be kept below about 50 °C to avoid side reactions with the cellulose caused by the high oxidation potential. At too high a temperature brightness is less stable and yellowing takes place easily.

2.2.5. *Chlorine and Hypochlorites (⟶ Chlorine)*

In the past chlorine [7782-50-5] was widely used in chemical pulp bleaching for lignin oxidation (delignification). In chlorination substitution and oxidation of the aromatic rings in lignin takes place, producing dicarboxylic acid and chlorinated lignin fragments [7]. About 10 % of the applied chlorine is found as halogenated compounds in the effluent. Chlorinated phenols and polyhalogenated dibenzodioxins and dibenzofurans are among these products. Quantitative analysis of these compounds is performed by an extractive method with activated carbon, and the result is expressed as adsorbable halogenated organic compounds (AOX) [8]. Some of these compounds are mutagenic [9]. In most regions of the world the application of chlorine was phased out shortly after identification of these compounds in the effluent.

Today hypochlorite is used mainly as stain remover (“cold bleach” in laundering). Like chlorine, it is no longer applied on a large scale in chemical pulp bleaching. There application is limited to the use of very small amounts for controlled lowering of the degree of polymerization of dissolving pulp. The oxidation potential of hypochloric acid [7790-902-3], HOCl, is higher (+ 1.50 V) than that of chlorine (+ 1.35 V) or hypochlorite [14380-61-1] OCl^- (+ 0.94 V). To avoid side reactions (mainly depolymerization) involving the aggressive hypochloric acid, bleaching is carried out in alkaline solutions to minimize the concentration of undissociated hypochloric acid. Because the hypochlorite anion is only a weak nucleophile, unable to attack electron-rich aromatic systems, alkaline conditions cause fewer unfavorable side reactions. Bleaching with hypochlorite generates halogenated compounds in side reactions. However, in comparison to chlorine AOX formation is significantly lower. Chloroform is the main byproduct. Sodium hypochlorite [7681-52-9] solutions are commercially available with active chlorine concentrations of 150 – 160 g/L. Active chlorine is the amount of chlorine set free upon acidification of the hypochlorite solution.

2.3. Reducing Agents

The most important reducing bleaching agent is sodium dithionite. Sulfur dioxide has been used to bleach wool since ancient times, but it is of less importance today. Sulfinic acid derivatives are used mainly for the removal of colorants from textile fabrics.

2.3.1. *Sodium Dithionite (* ⟶ Sulfites, Thiosulfates, and Dithionites*)*

Sodium dithionite [7631-94-9], $Na_2S_2O_4$, is still known as hydrosulfite in North America, although this incorrect initial description was recognized and corrected as early as 1881. Sodium dithionite is today produced from sodium formate and sulfur dioxide. Zinc dithionite [13986-17-9], prepared from sulfur dioxide and zinc dust, was formerly more important than the sodium salt. However, the heavy metal zinc causes environmental problems, and today on-site production of zinc dithionite is no longer practiced. The same applies for production from sodium amalgam and sulfur dioxide. On-site generation of dithionite from alkaline sodium borohydride solution, sodium bisulfite, and sulfur dioxide is used by some paper mills.

Sodium dithionite is available as crystalline powder (90 % $Na_2S_2O_4$) or as a refrigerated 150 g/L solution stabilized with alkali. Solutions should be kept well below 10 °C because dithionite disproportionates to thiosulfate and bisulfite:

$$2\,S_2O_4^{2-} + H_2O \rightarrow S_2O_3^{2-} + 2\,HSO_3^{2-}$$

Dithionite destroys chromophores by reducing carbonyl groups to alcohols and quinones to phenols. It is used to bleach mechanical pulp, wool, silk, bristle, hair, straw, soap, glue, kaolin, and foods such as sugar, molasses, gelatin, and edible fats.

Dithionite is easily oxidized to sulfite and sulfate by oxygen. The solution decomposes at elevated temperatures to thiosulfate, sulfite, hydrogen sulfide, or sulfur, the exact products depending mostly on pH. The pH selected for bleaching with dithionite is a compromise between the most effective bleaching and the lowest loss of dithionite. Normally, a pH of 5 – 6 and temperatures of 50 – 80 °C give the best results. In mechanical pulp bleaching, hydrogen peroxide and dithionite bleaching are often combined in two sequential steps.

2.3.2. *Sulfur Dioxide (* ⟶ Sulfur Dioxide*)*

Sulfur dioxide [7446-09-5] is used mostly in the form of sodium hydrogensulfite, $NaHSO_3$. Because of the relatively poor bleaching effect, SO_2 is no longer widely used. It can be applied in mechanical pulp production as a pretreatment to dithionite bleaching. Another use is the treatment of wool.

2.3.3. *Sulfinic Acid Derivatives (* ⟶ Sulfinic Acids and Derivatives*)*

The important sulfinic acid derivatives are sodium hydroxymethanesulfinate [149-44-0]

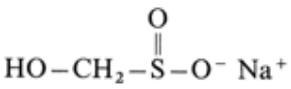

and formamidinesulfinic acid [4189-44-0]

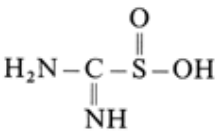

Hydroxymethanesulfinates are prepared by treating dithionites with formaldehyde; they are more stable than the dithionites. Higher temperatures are needed to produce the full reducing power. These sulfinic acid derivatives are therefore used in processes that operate above 75 °C, in which dithionite would be decomposed rapidly. They are used for bleaching soap, glue, gelatin, and fatty acids. However, their main use is in dyeing processes.

Formamidinesulfinic acid is prepared from thiourea and hydrogen peroxide. Therefore, it is sometimes called thiourea dioxide. It is a poorly water soluble crystalline product. It becomes more soluble under alkaline conditions, and it reaches its highest reducing power above 80 °C. It is used to increase brightness of textiles and to remove dyes from fabrics (like hydroxymethanesulfinic acid). Its strong reducing power at high temperature favors application in bleaching of secondary fibers in the dispersing step. It is very effective in the destruction of colorants used for colored paper. By volume this is the most important application.

[Top of Page]

3. Enzymes

The technical use of enzymes in bleaching is still the exception. The discovery of lignin-degrading enzymes in white rot fungi (*Phanerochaete chrysosporium*) [10] seemed to open a new option for industrial-scale delignification and bleaching. However, the reaction was very slow, not selective enough, and the enzyme required coenzymes which degraded cellulose. This is a general problem of enzyme application. Enzyme strains must be very pure to avoid negative side reactions by activity of other enzymes. In addition many industrial processes are operated at elevated temperature or extreme pH. Both require the use of adapted enzymes. Thermophilic organisms are a preferred source for the selection of an enzyme.

Xylanases are used to eliminate part of the precipitated xylan from unbleached kraft pulps [10]. This treatment lowers the demand for chlorine dioxide in ECF bleaching by approximately 10 %. It is suggested that lignin on the fibers is partly covered by precipitated xylan, and that removal of the xylan allows this lignin to be washed off more easily. The overall moderate effect has made xylanase application a niche option for pulp mills with limited availability of chlorine dioxide.

Laccase is the enzyme that polymerizes the phenylpropane units in plants and produces the lignin polymer. With the aid of a mediator, laccase can be used for the delignification of pulp [11]. Mediators are easily converted to radical cations by reaction

with laccase and oxygen. The intermediate product oxidizes the lignin. Theoretically, the mediator, for example, *N*-hydroxybenzotriazole [2592-95-2], can be reoxidized several times by the enzyme. However, repeated reaction is the exception. Consequently, demand for the mediator is high. Currently (in 2006), industrial application is not economical.

Cellulases are used to soften cotton fabric. They are applied in the stone washing of jeans as a substitute for the mechanical effect of pumice stones. This effect on the surface of the fabric causes a fading of the color, because the colorant is located only on the fiber surface.

Cellulases have also been recommended for improving the fiber strength of wastepaper pulp. The dissolution of fines and fibrils increases the average fiber length and increases the strength potential. However, this solubilization lowers yield and increases the effluent load. Therefore, large-scale application remains unlikely.

[Top of Page]

4. Bleaching Processes

The bleaching processes for the substances in Table 2 are described in detail in the articles on the substances. (For example, ⟶ Paper and Pulp.) The bleaching processes for pulp and textiles, the most important uses economically, are briefly summarized here.

4.1. Pulp (⟶ Paper and Pulp)

Fibers used for paper production are predominantly bleached [2]. A distinction can be made between the main types of paper fibers: chemical pulp and mechanical pulp. Wastepaper typically is a mixture of both types of fibers and contains in addition fillers such as china clay and calcium carbonate and printing ink.

4.1.1. *Chemical Pulp*

The most important method for the production of chemical pulp is the alkaline *kraft* or sulfate process, which is by far the dominant process worldwide. The acidic *sulfite* process plays a very minor role with a production share below 5 %. An oxidative process is required to destroy the residual lignin polymer. Because each bleaching agent has its specific advantages and disadvantages, generally several chemicals are applied in a sequence of steps. Bleaching is carried out continuously in towers or reactors. Between the steps, washing is carried out in drum, belt, displacement, press or diffusion washers. The bleachability of a pulp depends upon the pulping procedure. In general, kraft pulps require more effort and more bleaching steps to reach very high brightness. Typical market pulp brightness for softwood pulp is > 88 % ISO. The brightening process requires between four and six stages. Hardwood pulp is available with a brightness of 90 % ISO and higher. The abbreviations used to describe a bleaching sequence are listed in Table 3.

Table 3. Bleaching stage designations, conditions, and effects

Designation	Conditions and effects
O	oxygen stage for oxidation of lignin with molecular oxygen under alkaline conditions at 90 – 100°C.
A	acid stage to remove transition metals (at 40 – 60°C) or hexenuronic acid (at > 90° C) by hydrolysis; the acid typically is sulfuric acid.
Q	acidic stage (pH 5 – 6.5) with chelating agents (e.g., EDTA or DTPA) for transition metals removal.
D	chlorine dioxide stage using a solution of ClO_2 at pH < 5 in water for lignin oxidation.
E	extraction stage using caustic soda for solubilization of oxidized lignin (pH 9.5 – 11).
Eo	extraction with addition of oxygen gas for improved lignin removal by oxidation.
Eop	extraction reinforced with oxygen and hydrogen peroxide for improved lignin removal by oxidation and brightening.
P	alkaline bleaching stage with hydrogen peroxide, (pH >10 – 11, 60 – 90 °C).
OP	pressurized peroxide stage with addition of oxygen, potentially operated above 100 ° C with up to 0.3 MPa pressure.
Z	delignification stage with gaseous ozone.
Paa	weakly acidic stage (pH 5) with peracetic acid for lignin oxidation and activation for subsequent P stage.
X	enzyme treatment stage with xylanase or other hemicellulases to improve lignin accessibility by removal of precipitated carbohydrates.
Y	reductive treatment with dithionite.
N	neutralization.

Bleaching sequences are described by a sequence of these letters, e.g, ODED. In this combination the washing procedure between the stages is not explicitly mentioned. A combination without intermediate washing can be indicated by the use of brackets, e.g., (ZD). The way of combining the letters is not standardized. Sometimes in a combination of letters the capital letter is used to describe the main purpose of the stage; the combination “Eo” stands for an extraction stage with a support of the extraction by oxygen addition. Sometimes the description EO is used for the same purpose. A glossary of bleaching

terms is available in [12].

In the past in chemical-pulp bleaching chlorine was used as the main delignification chemical (abbreviated by C). Its selectivity in lignin oxidation and the low level of side reactions with cellulose made it the preferred delignification compound for some decades. A serious disadvantage is the generation of a large amount of various halogenated compounds with limited biodegradability. The detection of highly halogenated dioxins and furans initiated the elimination of chlorine from pulp bleaching. Its application today in pulp bleaching is limited to small, old pulp mills in exotic regions. Current bleaching processes use basically two alternative approaches: The replacement of chlorine by dominantly chlorine dioxide is labeled as ECF (elemental chlorine free) bleaching. Total chlorine free (TCF) bleaching requires a complete change in technology. TCF bleaching uses mainly oxygen and hydrogen peroxide for bleaching. Ozone or peracetic acid is in addition required to achieve high brightness. In ECF bleaching the simplest approach is to just exchange chlorine with chlorine dioxide. The optimization of an ECF sequence requires additional changes: higher temperature in the D stage and the reinforcement of the extraction stages with oxygen and hydrogen peroxide. ECF “light” is a label used to describe a sequence using less chlorine dioxide and more hydrogen peroxide.

Typical industrial bleaching sequences are listed in Table 4.

Table 4. Chemical pulp bleaching sequences

Pulp type	Bleaching process sequence	
Kraft pulp, softwood	ECF	ODEopDEpD, ODEoDnD
Kraft pulp, softwood	ECF “light”	OQ(OP)DP
Kraft pulp, softwood	TCF	OQ(OP)QPP, OOQ(OP)PaaP
Kraft pulp, hardwood	ECF	$OD_{(hot)}$EopDP, O(ZE)DP
Kraft pulp, hardwood	TCF	OQPPP
Sulfite pulp	TCF	OQP, OQ(PaaP), EopP

In bleaching kraft pulp, oxygen is used to decrease the amount of residual lignin without loss of viscosity and strength. Because kraft pulping and oxygen delignification both use sodium hydroxide as alkali source, combined evaporation and recovery of the effluents from pulping and oxygen delignification is possible. Thus, oxygen delignification not only decreases the amount of bleaching chemical required, it also lowers the overall effluent load. Oxygen delignification has become the standard measure for lowering the lignin level after pulping. Softwood kraft pulp can be further delignified by more than 50 %, and hardwood pulp by nearly 40 %.

The changes in bleaching technology have resulted in significantly lower levels of halogenated compounds in the effluent of pulp bleaching. These compounds are determined as AOX (adsorbable organic halogenated compounds). Historically the amount of AOX discharged per tonne of pulp could be higher than 5 kg/t. Most countries have set narrow limits to the discharge of AOX. ECF bleaching allows staying within these limits as it generates much less halogenated compounds, which show in addition improved biodegradability.

Combination of two bleaching chemicals in one stage is sometimes possible and advantageous. The effects of oxygen and hydrogen peroxide, for example, are to a certain extent additive. Chlorine dioxide and ozone can be applied in one combined step by using two mixers in line.

Water consumption in pulp production has decreased significantly. Volumes declined from more than 50 m^3/t to about 15 m^3/t in modern pulp mills. The limitation to further decreases is the presence of “nonprocess elements” in wood. The recovery of the process chemicals in parallel results in an increase of the concentration of these elements in the system. Therefore, discharge of effluent or solids will still be necessary. The cations are predominantly traces of metals such as Ca, Mg, Fe, Mn, and Cu. Chloride and sulfate are the main counterions. The target of process improvements is to minimize the impact not only of the discharge of liquid effluent but also to decrease atmospheric emissions and the amount of solid waste disposed of in landfills.

Bleaching chemicals are typically applied under medium-consistency conditions, that is, a fiber/water ratio of 1/9 to 1/11. This relatively thick mixture is fluidized by special pumps and pressed through upflow towers.

Typical market pulp brightness for ECF softwood pulp is 88 – 90 % ISO. Hardwood pulp is sold at a brightness of 90+ % ISO. Specialties such as photo paper can have a brightness > 92 % ISO. Lower brightness can be sufficient for a wide number of applications in papermaking and tissue production, and thus brightness varies over a wide range from unbleached to semibleached to fully bleached pulp. Typically TCF pulp is available with more brightness variation to balance brightness requirements and fiber strength properties. In TCF bleaching it is more difficult to simultaneously obtain top brightness and top fiber strength.

Special paper grades such as bank notes, cigarette papers, and tea bags are produced from special fibers. Esparto grass, hemp, flax, kenaf, ramie, and abaca are applied, as well as cotton linters. Although not all these fibers can be bleached to very high brightness, the requirement for bleaching chemical is low, and only short bleaching sequences are necessary. Another group of annual fiber plants includes the cereals, especially rice straw, bamboo, and depithed bagasse. They are used as a raw material for chemical or mechanical pulp.

4.1.2. *Mechanical Pulp*

Grinding tree trunks on rotating grinder stones is the traditional method of preparing mechanical pulp. The abbreviation describing this process is GW (groundwood) or PGW (pressurized groundwood) for processes operated at 2 – 3 bar. Today rotating metal disks, refiners, and wood chips as raw material are widely used; the conventional groundwood process has been phased out because of its high requirement for workforce. There are several slightly different processes: refiner mechanical pulp (RMP), thermomechanical pulp (TMP), chemothermomechanical pulp (CTMP), and alkaline peroxide mechanical pulp (APMP). In all cases, the amount of lignin in the pulp is reduced only slightly. At high temperature internal bonding within the wood chips decreases and longer fibers result from the mechanical action. Pretreatment of the wood chips with sodium sulfite, caustic soda (applied to softwood, e.g., spruce) or alkaline hydrogen peroxide (typically applied to hardwoods such as aspen or poplar) yields a defiberized product with even longer fibers.

The use of strong oxidizing agents such as oxygen or chlorine dioxide for bleaching mechanical pulp would result in an enormous consumption of chemicals and significantly lower the yield, because these chemicals would oxidize and dissolve the lignin present in the pulp. Hydrogen peroxide is used because it predominantly destroys chromophores without attacking the aromatic rings of lignin [7], so that yield losses are moderate. The amount of dissolved wood material depends on the intensity of bleaching; the more peroxide is applied, the more alkali is required for its activation, and in turn more organic material is dissolved. A simple possibility for a moderate increase in brightness is reductive treatment with sodium dithionite. This chemical can be added already during the refining process.

The maximum possible increase in brightness is lower for mechanical pulps than for chemical pulp. Softwood mechanical pulp can be bleached with dithionite to 70 – 75 % ISO. Hardwoods such as poplar and aspen bleach more easily, as their lignin contents are lower. Thermomechanical pulps give lower brightness increases, because the chromophores produced by high-temperature treatment are more stable towards reduction. Hydrogen peroxide bleaching produces higher brightness levels: softwood pulp can be bleached to 75 – 80 % ISO. Hardwoods mechanical pulp can reach brightness levels of up to 85 % ISO.

Hydrogen peroxide bleaching is best conducted at high pulp consistency, preferably above 30 %. At high consistency the relative concentration of the bleaching chemical is higher, and side reactions with organic compounds dissolved in the water are avoided. Bleaching processes are run continuously in downflow towers. Reductive dithionite bleaching is carried out conventionally at lower consistencies (4 – 6 %) in upflow towers to avoid consumption of the reducing agent by atmospheric oxygen. Deaeration in modern medium-consistency pumps allows reductive bleaching to be carried out at medium consistency (ca. 10 %) in upflow towers.

4.1.3. *Paper Recycling (Deinking)*

The amount of wastepaper recycled with a deinking step is continuously increasing. The aim of deinking is the removal of printing ink and other impurities from the fibers. Typically, printing inks cannot be bleached, as they consist mainly of pigments like carbon black and hydrophobic polymers. These inks must be removed mechanically. There are two major processes: washing and froth flotation [13]. For froth flotation, repulping is carried out under alkaline conditions with the addition of a sodium soap. The ink is detached and collected on precipitated calcium soap particles. The flotation of these precipitates removes not only ink particles but also other impurities. The addition of hydrogen peroxide prevents alkaline yellowing and improves brightness. As another additive sodium silicate is applied. It improves ink separation and suppresses the coflotation of fillers and fines. For best brightness and highest yield all four chemicals — caustic soda, soap, hydrogen peroxide, and sodium silicate — are required.

For deinking by washing, repulping of the wastepaper can be performed under alkaline or neutral conditions. To facilitate ink detachment surfactants or soaps are added. The washing process removes ink, fillers, and fines simultaneously. It is therefore the preferred method for the use of secondary fiber in tissue production.

Today the majority of recovered fibers subjected to a deinking process are used in newsprint production, where up to 100 % of the final product can be recycled paper. In smaller amounts deinked wastepaper is used as furnish in higher quality paper grades, such as SC (supercalendered) paper, copy paper, or printing and writing grades. Higher quality wastepaper (e.g., mixed office wastepaper) is used for the production of tissue paper grades. Brightening, especially the removal of dyes from colored paper grades, includes a combination of reductive and oxidative bleaching with formamidine sulfinic acid and hydrogen peroxide. A typical North American development is deinked market pulp of very high brightness. The intention of the recovery procedures is to achieve a fiber quality close to short-fiber chemical pulp. This requires a multistage process, including flotation, washing, and bleaching. The raw material generally is mixed office wastepaper. Bleaching stages include oxygen, hydrogen peroxide, and sodium dithionite or formamidine sulfinic acid. In some mills hypochlorite is still used. A higher content of mechanical fiber in such a process is difficult to handle, as oxygen delignification dissolves most of the fiber material and, in parallel to the loss of yield, a very high effluent load is generated.

4.1.4. *Environmental Aspects of Pulp Bleaching*

Bleaching removes part of the product being bleached. Washing is the most important method for removing the oxidized or reduced compounds and causes an effluent load which can be expressed as BOD or COD, or TOC (total organic carbon). Authorities normally set limits to the discharge of biodegradable compounds and oxidizable material. Treating the effluent starts with a “primary” step, the sedimentation of suspended solids (e.g., fibers). The standard method for removing organic material from the effluent in kraft pulp mills is biological treatment in aerated lagoons. This is denoted “secondary” treatment. The degradation of organic material with bacteria requires oxygen. Energy demand for aeration is about 1 kW·h/kg BOD_5 [14]. About half of the organic material is transformed into carbon dioxide, and the other half becomes biomass and must be separated as sludge. Sludge dewatering and disposal or combustion cause additional costs. An attractive alternative is an anaerobic fermentation step. This conversion leaves only about 5 % of the organic matter as residue for disposal. All other organic material is converted to biogas, mostly methane. Typically the BOD is decreased by more than 90 %. Anaerobic and aerobic treatment can be combined in two successive stages starting with the anaerobic treatment. The excess sludge from the aerobic stage can be recycled into the anaerobic stage. A “tertiary” treatment is the final precipitation of remaining suspended or dissolved compounds with chemical coagulation. Usually, compounds for such a treatment are aluminum salts, Fe^{III} salts, and lime slurries. Tertiary treatment is mainly applied to remove excess nutrients like phosphorous, but it can also

remove COD and AOX.

In ECF bleaching halogenated compounds are produced during the bleaching reaction. Because of the wide range of halogenated products, the use of a sum parameter is widely accepted to characterize the amount. A standard method is AOX analysis, which describes the amount of adsorbable (on activated carbon) halogenated compounds [8]. Conventional bleaching resulted in AOX amounts of 5 – 8 kg/t of organically bound chlorine. ECF conditions can lower the AOX discharge from the bleaching plant to less then 1 kg. A further decrease is achieved in biological treatment, which can cut the AOX level by more than 50 %.

The bleaching process leaves residual halogenated organic compounds in the pulp, which are measured as organic halogen (OX) in pulp. To a certain extent it is possible to identify the bleaching process applied by measuring the residual OX content of a pulp sample. ECF bleaching has residual OX contents between 150 and 30 g/t. This lower level is also the accepted level for TCF bleached pulp, because the pulp can have a natural background content of halogenated compounds and a higher level of chloride ions in pulp during bleaching can generate residual OX even with TCF bleaching.

In chemical pulp production specific water consumption is below about 30 m^3/t . A further decrease to less than 15 m^3/t is achieved with sophisticated washing equipment and countercurrent waterflow [15]. Complete bleach plant closure is very difficult to achieve. Most units labeled as “closed” in reality are still consuming water. One reason are the “nonprocess” elements present in wood. They become concentrated in the recovery loops and can cause equipment scaling or even upset the recovery procedure. The inclusion of bleach plant effluent in the recovery process increases scaling and corrosion potential. Consequently “closed” mills use “kidneys” to eliminate troublesome substances. Such purging consumes water or requires the disposal of solid waste after evaporation and combustion. (Solid waste would contain salts of the nonprocess elements and salts resulting from pulping and bleaching chemical application).

Totally effluent free bleaching only changes the kind of discharge: from water pollution to air pollution or to the need for solid waste disposal. The concentration process required (evaporation) will increase energy demand.

4.2. Textile Fibers (⟶ Textile Dyeing)

The quality of the raw material dictates the potential for a simple or a more complicated bleaching procedure. The higher the level of impurities in the material to be bleached, the more steps are required. Natural products like cotton, flax, and wool can be heavily contaminated with sand and other dirt. Washing in the presence of acid or chelating agents may be required to achieve good results in the bleaching stage.

Cotton, the most important natural fiber, is usually bleached with hydrogen peroxide because alkaline oxidation not only increases brightness but also removes extractives, saponifiable waxes, and pectins, and destroys the lignin of residual seed capsules. In the peroxide stage, chelating agents, magnesium salts, and sodium silicate are added to prevent transition metal catalyzed destruction of the cellulose.

Cotton linters are used as a specialty fiber for papermaking and as raw material for cellulose derivatives like cellulose acetate. Linters are normally alkaline pretreated and then bleached in several stages, for example, in a chlorine dioxide – hydrogen peroxide – chlorine dioxide sequence, analogous to chemical pulp bleaching.

Flax is the source plant for linen. Bleaching of flax is very similar to that of other plants containing bast fibers, for example, hemp or ramie. Because of the low lignin content a pulping process is not required. Most of the impurities can be removed by mechanical action. High brightness and cleanliness is achieved with multistage processes using a hot alkaline pretreatment followed by sodium chlorite and hydrogen peroxide.

Wool needs thorough washing after shearing. The scouring step removes sand and other dirt and separates lanolin (wool fat). Wool is a keratin (protein) fiber and thus sensitive to alkaline treatment. Hydrogen peroxide bleaching must be carried out at low temperature and below pH 8, often under neutral to even slightly acidic conditions. As a result the bleaching times are long and the potential for high brightness is limited. Combining oxidative peroxide treatment and reductive dithionite treatment gives the greatest increase in brightness.

Synthetic fibers like rayon are bleached with hydrogen peroxide or hypochlorite. Because polyurethane is sensitive to chlorinating chemicals, polyurethane fibers are bleached only with hydrogen peroxide. Polyamides, polyesters, and polyacrylonitriles are bleached with sodium chlorite. Hot alkaline hydrogen peroxide may cause polyamides to lose strength.

4.3. Other Products

Animal Products. Similar to wool hair and bristle are bleached with semineutral or slightly alkaline peroxide.

Mother-of-pearl, ivory, and bones are bleached with hydrogen peroxide solutions, which are stabilized with either sodium diphosphate in weak alkaline medium or sodium silicate, possibly in the presence of ammonia. Bones and ivory are bleached at room temperature for an extended time.

Vegetable Products. *Straw* is bleached with alkaline hydrogen peroxide. Straw-hat bleaching was the first important application for sodium peroxide at the beginning of the 1900s. Exotic straws such as pandan straw or bamboo grass require bleaching with sodium chlorite. The bleached products are typically used by florists.

Peeled rattan reeds, which are used for wicker baskets and furniture, are brightened with hydrogen peroxide (buffered with sodium carbonate and silicate) at ambient temperature and a dithionite stage.

Wood veneer and *wood* are bleached with hydrogen peroxide and ammonium carbamate. Offcolors can be corrected with this treatment, which is important in furniture production.

Palm oil and other *vegetable oils* for food and chemical production need to be purified before use. They are typically bleached not by a chemical reaction but by adsorption of colored impurities. High surface area clays (aluminum or magnesium silicates) are added at elevated temperature. Filtration is used for separation. The same procedure is applied for *paraffins* and *waxes*. Disposal of the clay contaminated with oil and colored material can become an environmental challenge.

Inorganic Products. *Clays* are bleached as aqueous slurries with sodium dithionite to remove the color resulting from metal ions such as iron. The reduced metal (Fe^{2+}) is washed off.

Technical-grade *sulfuric acid* is bleached with hydrogen peroxide. The oxidation of organic impurities takes place via the intermediate generation of Caro's acid (H_2SO_5).

[Top of Page]

References

1. C. W. Dence, D. W. Reeve (eds.): *Pulp Bleaching: Principles and Practice*, Tappi Press, Atlanta 1996.
2. H. Sixta: *Handbook of Pulp*, Wiley, New York 2006.
3. M. Peter, H. K. Rouette, *Grundlagen der Textilveredlung*, Deutscher Fachverlag, 1989.
4. *Prüfung von Papier Pappe Zellstoff und Holzstoff*, Band 1: Chemische und mikrobiologische Verfahren, Springer Verlag 1991; Band 2: Mikroskopische und photometrische Verfahren, Springer Verlag 1993; *TAPPI Test Methods*, TAPPI Press 1996, Atlanta.
5. ISO 2470-1973, Tappi T 646 om-86, DIN 53 145.
6. A. Maréchal, *J. Wood Chem. Technol.* **13** (1993) 6, 91.
7. J. Gierer, *Holzforschung* **36** (1982) 55. Links
8. DIN 38409 H 14.
9. K. Kringstad, K. Lindström, *Environ. Sci. Technol.* **18** (1984) no. 8, pp. 236A – 238A. Links
10. K.-E. L. Eriksson: *Biotechnology in the Pulp and Paper Industry*, Springer, Berlin 1997.
11. R. Bourbonnais, M. G. Paice, “Enzymatic Delignification of Kraft Pulp using Laccase and a Mediator”, *Tappi J.* **79** (1996) 199. Links
12. Glossary of Bleaching Terms, Bleaching Committee, Technical Section, CPPA, Montreal, QC Canada.
13. *Papermaking Science and Technology*, Book 7, “Recycled Fiber and Deinking”, Fapet Oy, Helsinki 2000.
14. C. H. Möbius: Abwasser der Papier- und Zellstoffindustrie, 3. Auflage, 2002, revised Dec. 2005, available at: www.cm-consult.de.
15. G. Annergren, P. Sandström, “A Bleach Plant Designed for Closure”, *Papier* **50** (1996) V12 – V16. C.-J. Alfthan, L. Ahlenius, G. Svensson, “Betriebserfahrung bei der Herstellung von Sulfatzellstoff im geschlossenen Bleichkreislauf”, *Papier* **50** (1996) V8 – V12.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Blood

Standard Article

Rainer Seitz[1], Herbert König[1], Johannes Dodt[1]

[1]Paul-Ehrlich-Institute, Dept. of Haematology and Transfusion Medicine, Langen, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a04_201.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Composition, Physiology and Chemical Properties of Blood
2.1. Blood Cells
2.1.1. Hematopoiesis, Stem Cells
2.1.2. Red Cells, Hemoglobin
2.1.3. Leukocytes, Cytokins
2.1.4. Thrombocytes
2.2. Plasma Proteins
2.2.1. Analysis of Plasma
2.2.2. Coagulation Factors
2.2.3. Fibrinolysis System
2.2.4. Plasma Proteinase Inhibitors
2.2.5. Immunoglobulins
2.2.6. Albumin
3. Production
3.1. Blood and Plasma Donation
3.2. Substances Derived from Blood Cells
3.3. Fractionation of Plasma
3.3.1. Quality Requirements
3.3.2. Preparation of Intermediate Products
3.3.3. Purification of Plasma Proteins
3.4. Recombinant Proteins
3.5. Adverse Drug Reactions of Blood Products
3.5.1. Immunogenicity
3.5.2. Thrombogenicity
3.5.3. Transmittable Pathogens
4. Therapeutic Uses
4.1. Blood Components
4.1.1. Blood and Stem Cell Preparations
4.1.2. Fresh Frozen Plasma
4.2. Hemoglobin-Derived Oxygen Carriers
4.3. Cytokins
4.4. Coagulation Factors
4.5. Plasminogen Activators
4.6. Inhibitors
4.7. Immunoglobulins
4.8. Albumin
5. Economic Aspects

Blood is a multifunctional transport medium of vital importance for the body of higher organisms. Its composition is very complex and includes various cell systems with peculiar functions, proteins, electrolytes, metabolites, and highly sophisticated mediator systems. Both cells and plasma, which can be separated by centrifugation, can be used as and further processed into medicines. Serum is the fluid remaining after coagulation of the blood. In this review, blood cell systems and plasma proteins are referred to and medicinal aspects including blood and plasma collection and processing are described. The medicines derived from blood and plasma, as well as basic research and clinical trials should keep an important place because of the medicine as a whole benefits extremely from transfusions and, in particular, hemophiliacs gain an enormous improvement of life quality and life expectancy.

[Top of Page]

1. Introduction

Blood is a multifunctional transport medium of vital importance for the body of higher organisms. The composition of blood is very complex and includes various cell systems with peculiar functions, proteins, electrolytes, metabolites, and highly sophisticated mediator systems. Its undisturbed circulation in an ubiquitous system of blood vessels with fine ramifications throughout the whole organism, connections to the lymph vessels and interfaces with the extravascular compartments is essential for the performance of all organs and enables numerous exchange processes. A sudden stop of blood circulation will cause the death of the organism within a short time. The functions of blood can be allocated to the following categories:

1. *Respiratory function*: the blood provides the capacity to transport oxygen needed for the energy metabolism from the lung to the organs, and of carbon dioxide back to the lung.
2. *Regulatory function*: besides by the nervous system, numerous specific body functions are controlled or influenced by messenger substances such as hormones or cytokins that are transported by the blood; the production of blood cells itself is a good example (see Section Hematopoiesis, Stem Cells). The blood has to provide also a stable internal milieu; the pH is maintained within a narrow range around 7.4 by buffer substances and by transport of acid or basic substances to metabolic or excretory organs. The blood is also involved in the regulation of further important parameters, e.g., water household or temperature.
3. *Nutritive and metabolic function*: nutrients absorbed from the gastrointestinal tract are collected into the splanchnic vessels, transported via the portal vein to the liver, the body's main synthetic and storage organ, and further transferred to the tissues in order to maintain energy supply and their specific metabolic functions. Waste and toxic products, as well as superfluous water and electrolytes are transported from the tissues to the sites of their elimination.
4. *Protective function*: the human body is permanently subject to external and internal attacks of agents disturbing its homeostasis or even threatening the individual's life. The blood contains highly specialized cellular (various types of phagocyte cells and lymphocytes) and humoral (e.g., immunoglobulins, complement, proteases, and inhibitors) protective systems providing effective surveillance, defense, and repair.

The blood volume in humans is equivalent to about 6 to 8 % of the total body mass, which again underlines the significance of this “fluid organ”.

[Top of Page]

2. Composition, Physiology and Chemical Properties of Blood

Blood can be obtained by donation into special containers that contain substances preventing coagulation and can be easily separated by centrifugation into the cellular fraction and the liquid supernatant (the *plasma*); *serum* is the fluid remaining after blood has undergone coagulation. Both cells and plasma can be used and further processed into medicines. Blood cell counts and other relevant values are shown in Table 1.

Table 1. Typical values of selected cellular parameters measured in normal adult blood*; (Units are given as usually used in clinical laboratories and text books; in some countries and publications SI-units are preferred for certain parameters. The actual reference ranges may vary depending on the assay methods)

Parameter	Value	Remarks
Red cells		
Erythrocyte (red cell) count	m: 4.4 – 6.0 × 10^6/µL f: 4.2 – 5.5 × 10^6/µL	
Hemoglobin	m: 14.0 – 18.0 g/dL f: 12.0 – 16.0 g/dL	
Hematocrit (ratio cell volume:/ total blood volume)	m: 0.40 – 0.54 f: 0.37 – 0.47	
MCH (mean corpuscular hemoglobin)	27 – 32 pg	mean hemoglobin mass per single red cell
MCHC (mean corpuscular hemoglobin concentration)	32 – 36 g/dL	mean hemoglobin concentration in a single red cell
MCV (mean corpuscular volume)	80 – 94 fL	

Reticulocytes	7 – 15 per 10^3 red cells	
White cells		
Total white cell count	4.3 – 10.0 × 10^3/µL	
Differential white cell count (relative to total number):		values obtained with microscopy of blood smears; values with automatic counters may be different
Neutrophils, band form	1 – 5 %	
Neutrophils, segmented	45 – 70 %	
Eosinophils	1 – 4 %	
Basophils	0 – 1 %	
Monocytes	2 – 8 %	
Lymphocytes	20 – 40 %	
Platelets		
Thrombocyte count	150 – 350 × 10^3/µL	

* m = male, f = female proband.

2.1. Blood Cells

Several highly differentiated cell systems with specialized characteristics and functions are circulating in the blood. The cells may be either nucleated such as leukocytes, in this case they are reactive to various stimuli, and actively mobile between blood stream and tissues. Otherwise they are nonnucleated, have a restricted metabolism, and stay within the blood their whole life-span. An example of the latter category are erythrocytes.

2.1.1. *Hematopoiesis, Stem Cells*

After the fetal period, the production of blood cells (hematopoiesis) takes place essentially in the red bone marrow within spongious bones (e.g., vertebrae, sternum, pelvis), and in case of lymphatic cells also in lymph nodes and other sites (e.g., spleen, tonsils, appendix, Peyer's plaques, or during childhood in the thymus). In adults, the hematopoietic marrow has a total mass of about 1000 to 1200 g and thus belongs to the largest organs. It possesses a special microenvironment, a “matrix” of cells, connective tissue and adhesion proteins, which supports homing, proliferation, and maturation of hematopoietic cells. There is a permanent turnover of mature blood cells, and the half lives of several types are short. Therefore, great quantities of new cells have to be produced: for a healthy adult, the estimate is about 10^{11} granulocytes, 2 × 10^{11} erythrocytes, and 2 × 10^{11} thrombocytes every day.

The microscopic examination of bone marrow is a very important feature in the diagnosis of many hematological disorders. Specimens can be obtained by needle biopsy from the pelvis (*spina iliaca posterior*) for histology, or by puncture and aspiration from the sternum, and subsequent preparation of smears for cytological examination. The specimen can be stained using conventional panoptic dyes or specific antibodies in order to visualize certain markers or structures. The bone marrow gives at first glance the impression of a colorful, somewhat “chaotic” conglomerate of various cells of different size and appearance (Fig. 1). However, the hematopoiesis is a complex and very thoroughly regulated process, which is effectively controlled and can be adapted within a wide range to the actual needs of the organism in case of stressing situations such as infections or blood loss.

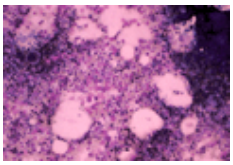

Figure 1. Bone marrow (overview)

Panoptic Pappenheim stain (same stain used also in Figs. 3-5) of a smear of a punctate obtained by sternal puncture and aspiration.

For the photographs in Figures 1 and 3-5, the authors are indebted to Mrs. Elke Eckstein, Philipps University Hospitals, Marburg, Germany.

The proliferating and differentiating bone marrow cells can be divided into several lineages, according to the resulting mature blood cells: (1) the “myeloid” lineages including the erythrocyte, thrombocyte, granulocyte of different types (neutrophil, eosinophil, basophil), and monocyte/macrophage precursors, and (2) the lymphatic cell lines. All of these cell lineages are ultimately derived from a common origin, the so-called *“pluripotent stem cells”* (Fig. 2). The latter are permanently proliferating in order to maintain a constant pool of pluripotent stem cells and to produce a sufficient number of progenitor cells for blood cell supply. Some stem and progenitor cells are permanently recirculated via the blood stream, which provides the opportunity to harvest them for therapeutic purposes by apheresis. The first step of hematopoiesis is the evolution into the myeloid or the lymphoid stem cells, and further into committed stem cells, which can be viewed as the start of the several lineages. By proliferation and progressive differentiation numerous cells at different stages are generated, until the cells reach the mature functional stage of blood cells and enter the blood via the ample bone marrow vessels (sinusoids). While pluripotent and committed stems cells can not be discriminated by morphology, but only by specific surface antigens (used as diagnostic markers), the more differentiated bone marrow cells can be identified by their typical microscopic appearance. There are, e.g., progenitors of erythrocytes (*erythroblasts*, Fig. 3), granulocytes (*myeloblasts*, Fig. 4), or thrombocytes (*megakaryocytes*, Fig. 5).

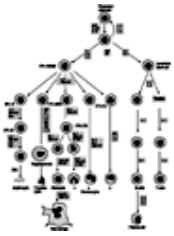

Figure 2. Scheme of cell line differentiation in hematopoiesis

BFU = burst-forming unit; E = erythrocyte; Epo = Erythropoetin; Eo = Eosinophil; GM = granulocyte-macrophage; MEG = Megakaryocyte; Tpo = Thrombopoetin.

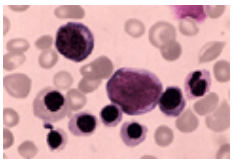

Figure 3. Erythropoiesis:

Large cell in the center is a proerythroblast; during further divisions and differentiation the cells become smaller, the nuclei increasingly dense, the cytoplasm contains more and more hemoglobin. Finally, the nucleus is extruded. At the left neutrophil granulocyte.

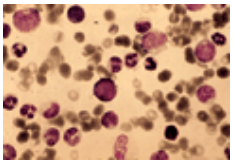

Figure 4. Myelopoiesis:

In the center myeloblast with sparse, basophilic cytoplasm, surrounded by more differentiated cells up to mature segmented granulocytes.

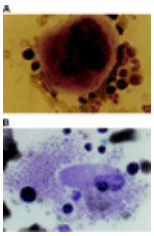

Figure 5. Megakaryocytes:

Largest cells in bone marrow (compare surrounding cells)

A) Mature megakaryocyte with giant nucleus and “granulation”; B) megakaryocyte after shedding of platelets, leaving nucleus behind.

The proliferation and differentiation of the lineages is controlled specifically by hematopoietic growth factors, which belong to an important group of substances referred to as cytokins or “biologic response modifiers”. Their nomenclature is somewhat confusing, due to historical reasons. A long-known substance is erythropoietin, a polypeptide produced by the kidneys in response to reduced oxygen concentration, which stimulates the production of oxygen carriers, i.e., erythrocytes. A rather “new” growth factor is thrombopoietin, which stimulates megakaryocytes [1], and shows about 50 % homology to erythropoietin. Other factors were discovered by their ability to promote in vitro the growth of progenitor cell colonies, which led to a nomenclature of the responsible cells as *colony forming units* (CFU) and of the growth factors as *colony stimulating factors* (CSF), for example the granulocyte-colony stimulating factor, G-CSF. It was also noted that several of a group of substances discovered as leukocyte-derived modifiers of inflammation, the interleukins, are also hematopoietic growth factors. But also “classical” hormones, such as thyroid or adrenal hormones, have significant influence on blood cell production. More detailed information about stem cell biology and their regulation can be obtained in [2], [3]. Much can be learned also from the pathophysiology of bone marrow failure in acquired aplastic anemia [4]. However, in face of the rapidly increasing knowledge with ongoing discovery of novel factors and interactions, it is impossible at present to provide a definite and comprehensive description of the regulation of hematopoiesis. A provisional, simplified overview is given in Figure 2, a selection of cytokins is listed in Table 2. A panel of mature blood cells is displayed in Figure 6 A – F.

Table 2. List of selected cytokins

Factor	Major source(s)	Main effect(s)
Granulocyte-colony stimulating factor (G-CSF)	monocytes, fibroblasts	stimulation of generation and differentiation of neutrophils
Granulocyte- macrophage-colony stimulating factor (GM-CSF)	T-cells, fibroblasts, endothelium	stimulation and differentiation of progenitor cells for granulocytes, monocytes, platelets, erythrocytes
Erythropoietin	kidney	stimulation of erythropoesis, some effect also on platelet production
Thrombopoietin	liver	stimulation of platelet production
Interleukin-1 (IL-1)	monocytes	enhancement of other growth factors, and of immune reactions
IL-2	T-cells (helper cells)	activation and proliferation of T-cells
IL-3 (multi-CSF)	T-cells	proliferation (self-renewal) and

		differentiation of early progenitor cells
IL-4	T-cells, macrophages	stimulation of B-cells
IL-6	macrophages, fibroblasts, T-cells	differentiation of B- and T-cells, acute-phase reaction (enhanced synthesis of fibrinogen and other proteins)
IL-16	T-cells	chemotaxis of CD4-cells, inhibits proliferation of HIV in infected cells
Stem cell factor (SCF)	fibroblasts and other cells	stimulates CD34-positive stem cells
Tumor necrosis factor (TNF	macrophages	enhancement of inflammatory reactions, pro-coagulatory stimulation of endothelium, antitumor effects
Interferons (α, β, γ)	leukocytes, fibroblasts, T-cells	various antiviral, antiproliferative (some types of malignancy) and immune-modulating effects
Platelet-derived growth factor (PDGF)	megakaryocytes and platelets	growth stimulation of vessel wall and inflammatory cells

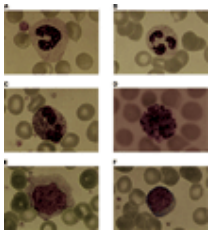

Figure 6. Mature leucocytes in the peripheral blood smear:

A) Band form; B) Segmented neutrophil; C) Eosinophil; D) Basophil; E) Monocyte; F) Lymphocyte

Compare size with surrounding erythrocytes and platelets.

2.1.2. *Red Cells, Hemoglobin*

The red cells, the erythrocytes, are produced by proliferation and differentiation of precursor cells (erythroblasts) in the bone marrow. The mature erythrocytes do not contain a nucleus and are no more capable of cell division. In the young erythrocytes, during the first days after they left the bone marrow, some residues of RNA and organelles can be found by a special stain. These cells are called reticulocytes. The normal life span of erythrocytes is about 110 d, thereafter they are phagocytized and degraded in macrophages of the spleen and the so called reticulo-endothelial system (RES). The iron and the protein of the erythrocytes is recycled for biosynthesis, the heme is degraded and metabolized into bilirubin, which is transported by the blood stream into the liver, where metabolites are excreted in the bile.

Mature erythrocytes have a characteristic shape of a biconcave disk. Normal red cells have a diameter of 7.5 to 8.7 µm; variations may be a diagnostic feature of disease states, e.g., smaller erythrocytes occur in iron deficiency, or larger cells in vitamin B_{12} deficiency. Erythrocytes have an average volume of 90 fL, and their shape provides a particularly large surface area of about 136 μm^2, which is important for their function in gas exchange. Red cells show a pronounced deformability which is needed for their passage through narrow capillaries. The erythrocytes are susceptible to alterations of the osmotic pressure and will shrink in hypertonic and swell in hypotonic solution until they are lysed at a critically low osmolality. Although the red cells do not contain a nucleus, they are quite differentiated cells which possess also an energy metabolism. Glucose is utilized as a source of energy via the glycolytic pathway and the hexose monophosphate shunt; red cells lack, however, the citric acid cycle. The energy is needed (1) to maintain the iron of hemoglobin in the divalent form; (2) to maintain the electrolyte gradient with high potassium level inside the erythrocytes, (3) to protect hemoglobin and enzymes from oxidation, and (4) to retain the biconcave shape of the cell.

The characteristic function of red blood cells, the transport of oxygen, is carried out by hemoglobin. Hemoglobin is a complex of a tetrameric protein (globin) component and the heme (which is made up by a characteristic porphyrin ring system similar to that found in chlorophyll, and divalent iron). The globin component consists of two alpha chains and two other chains. In adults, most of the hemoglobin contains beta chains (giving a formula $\alpha_2 \beta_2$). This hemoglobin is named hemoglobin A. This nomenclature must not be mixed up with the AB0 blood group system. In the fetal period, and in some hereditary disorders of hemoglobin metabolism, hemoglobin contains other molecules instead of the beta chains, e.g., gamma chains in the fetal hemoglobin ($\alpha_2 \gamma_2$, hemoglobin F). The binding of oxygen to hemoglobin is influenced by the temperature, pH value, and the red cell content of 2,3-diphosphoglycerate. The oxygen dissociation curve shows a typical S-shape.

2.1.3. *Leukocytes, Cytokins*

All cells circulating in the blood, which are neither erythrocytes nor thrombocytes, are summarized by the expression “leukocytes” or “white blood cells”. The leukocytes in the blood can be divided into granulocytes, monocytes, and lymphocytes.

Granulocytes. The name “granulocytes” refers to the characteristic granules making up the specific appearance of these blood cells under the light microscope. The granulocytes can be differentiated by the standard panoptic staining of blood smears (e.g., Pappenheim or Wright) as neutrophil, eosinophil, or basophil granulocytes.

Neutrophil granulocytes normally comprise more than 50 % of the total leukocyte count. They are derived from precursor

cells in the bone marrow (myeloblasts and myelocytes). During the proliferation and increasing differentiation of these precursor cells, the granules undergo a maturation process, and the nucleus changes from the big and round shape in myeloblasts to the beanlike appearance of the so called “band forms” of the young granulocytes. During their further live span, the nucleus of the neutrophil granulocytes undergoes progressive segmentation. While an increase of band forms in the blood smear is a sign of an acute challenge, in most cases bacterial infection, the appearance of granulocytes with an abnormally high number of segments may point to certain hematological diseases with impaired blood cell production. Normally, the neutrophil granulocytes circulate only several hours in the blood circulation, before they extravasate and enter the tissues. They are very reactive, nonspecific phagocytes, which can be viewed as the first line of combat against any invasion of pathogenic or toxic agents. The neutrophil granulocytes play a significant role in the surveillance against bacteria and fungi at outer surfaces of the body, e.g., the oral cavity. Therefore, a lack of functional granulocytes (e.g., in an acute leukemia or bone marrow failure) may lead to characteristic, severe local infection in the mouth (gingiva, tonsils) or in the perianal area. In individuals with normal bone marrow function, the production of neutrophil granulocytes can be increased over a wide range in case of an acute challenge such as an acute bacterial infection, intoxication or major trauma, with an increase of the total white cell count and the proportion of band forms in the blood. This increase is initiated by cytokins, particularly IL-3, GM-CSF and G-CSF; IL-8 is a strong activator of neutrophil granulocytes. The neutrophil granulocytes are activated by a variety of nonsoluble (e.g., pathogenic microorganisms) or soluble (endotoxins, leukotrienes, complement factors, fibrinopeptides) stimuli. They are capable of active locomotion, and they are attracted to the site of the lesion by chemotaxis. The phagocytosis is strongly enhanced by the so-called opsonization of particles, which means their coating by molecules such as complement factors or immunoglobulins. The neutrophil granulocytes possess a very efficient apparatus for intracellular killing of ingested microorganisms and degradation of foreign protein. They have the capacity to produce highly reactive oxygen radicals by means of specific enzymes, such as myeloperoxidase, in their azurophilic granules. The specific granules contain several enzymes such as acid hydrolases, alkaline phosphatases, and proteolytic enzymes, particularly elastase, with a high capacity to degrade foreign materials. During activation and phagocytosis, oxygen radicals and enzymes may leak out into the surrounding, where they may damage structures and proteins. Proteases from neutrophil granulocytes, particularly elastase, are thought to contribute to extravascular fibrinolysis [5].

Eosinophil Granulocytes. The eosinophil granulocytes contain specific granules which are larger than those of neutrophils and are stained red by eosin. Normally, they comprise 2 – 4 % of the total leukocyte count. The nucleus usually has two segments. Besides enzymes like in neutrophil granulocytes, the specific granules contain a so-called “major basic protein”, which makes up about 50 % of the granule content and plays a role in the killing of parasites and mammalian tissue cells. Besides their ability to nonspecific phagocytosis, eosinophil granulocytes exhibit some characteristic features: The cytotoxic destruction of parasites in cooperation with macrophages, cytotoxic damage to tissue cells (e.g., cells of the respiratory tract in patients with asthma bronchiale), and neutralization of mediators from mast cells such as histamine, serotonin, and bradykinin. During anaphylactic reactions, eosinophils contribute to the attenuation of the inflammatory process. Production and function of eosinophils is controlled by cytokins (e.g., IL-5), and is also influenced by steroid hormones (decreased by corticosteroids, increased by estrogens and androgens).

Basophil granulocytes have a very characteristic appearance due to their large and dark blue granules, which often hide the nucleus. The granules contain high amounts of heparin and biogenic amines (histamine, serotonin) and bradykinin. The effect of these mediators is generally an enhancement of inflammatory reactions by vasodilatation and prevention of coagulation in order to maintain free access for inflammatory cells and mediators to the area of the lesion. Basophils possess receptors for immunoglobulin E (IgE) on their surface and are specifically involved in type I allergic (anaphylactic) reactions, where they cause severe hypotension or even lethal shock by massive release of histamine and bradykinin.

Monocytes belong to the group of large, phagocytizing cells which are called macrophages. They share a common progenitor cell with the granulocytes in the bone marrow. The macrophages circulating in the blood stream are called monocytes due to their round, oval or kidney-shaped nucleus, which is not segmented in contrast to granulocytes. After a circulation time of a few days in the blood, the monocytes move to the tissues, where they may stay alive for months. The macrophages in the tissues are called histiocytes, or in the liver also Kupffer cells. Macrophages may be found in all tissues of the organism, but particularly rich of these cells are the liver, the spleen, lymphnodes, and bone marrow, as well as areas of chronic inflammation, particularly granulomas. Macrophages are able to kill bacteria and to phagocytize necrotic material, e.g., in tissue repair after a local infection has been overcome. Besides that macrophages have a very important function in the specific immune defense due to their cooperation with lymphocytes. After phagocytosis and proteolytic digestion, fragments of foreign material are transferred to the macrophage surface, where they are “presented” to lymphocytes.

Lymphocytes are produced in the bone marrow and in lymphatic organs such as the spleen, lymphnodes, tonsils, appendix, Peyer's plaques, and — during childhood — the thymus. Lymphocytes in the peripheral blood can be identified as small to mid-sized cells with a round nucleus and few cytoplasm. In contrast to other blood cells, the maturity stage and the specific function of the lymphocytes cannot be differentiated by morphologic examination. In order to identify and quantitate the different functional subsets of lymphocytes, the staining of characteristic antigens on the cell surface by specific antibodies has been introduced. The antigens have been described in the “cluster of differentiation” (CD-) nomenclature, leading to a large and still expanding list of antigens on blood and bone marrow cells (Table 3). The determination of the CD-antigens in cell samples is nowadays a very important tool for the differential diagnoses of disorders of the hematopoietic and immune system. A well known example is the disturbed balance between CD-4 and CD-8 lymphocytes in acquired immune deficiency syndrome (AIDS). The CD-antigens resemble in many cases important functional receptors on the cell surface.

Table 3. Cluster of differentiation (CD)-nomenclature; examples of CD entities and their expression by blood cells

CD designation	Remark	Main cell reactivity
CD3	T-cell receptor complex	T-cells
CD4	MHC class II and HIV receptor	T-cells
CD8	MHC class I receptor	T-cells

CD14	lipopolysaccharide (LPS) receptor	macrophages, granulocytes, dendritic cells
CD19	surface Ig family	B-cells
CD34	used for stem cell selection	hematopoietic stem cells
CD41	GP IIb-IIIa (fibrinogen receptor)	platelets
CD62P	P-selectin	activated platelets and endothelial cells

Lymphocytes can be divided into two main groups, the *T-lymphocytes* (which are primed in the thymus), and the *B-lymphocytes* (which are primed in the bone marrow, the equivalent to a specific organ of birds, called *Bursa Fabricii*).

T-Lymphocytes. The group of T-lymphocytes includes cells with important regulatory function (so called helper or suppressor cells), as well as cytotoxic cells. T-helper or inducer cells (T_H) can be divided into subpopulations: T_{h1}*cells* support the proliferation and the function of T-lymphocytes and thus the cell mediated immunity, T_{h2}*cells* support the differentiation of B-lymphocytes to the antibody producing plasma cells. There is evidence that T-suppressor cells are probably not a defined T-cell subpopulation of its own, “suppression” of the immune answer appears to be rather a special functional state of cytotoxic or helper cells. For instance, T_{h1}cells promote their own development, but inhibit at the same time the development of T_{h2}cells and vice versa.

A specific immune reaction is initiated by close contact and cooperation of an antigen-presenting cell (macrophages or dendritic cells) and T-cells. The interaction proceeds on several levels. A nonspecific cell adhesion is mediated by adhesion molecules, a specific interaction is brought about by molecules of the so-called major histocompatibility complex (MHC) which can be divided into class I and class II molecules. The pattern of MHC-antigens of an individual is the basis for the human leukocyte antigen (HLA) typing, which is useful for the assessment of compatibility in organ transplantation, as well as in platelet transfusion. The MHC-proteins are important for the antigen presentation. MHC-molecules bind short peptides derived from degraded antigens, class I peptides of viral proteins or endogenous (tissue derived) antigens, and class II those of exogenous, e.g., bacterial proteins. The MHC-molecules bearing the antigen peptides react with a T-cell receptor which recognizes specifically the presented antigen. By this way, T-cells with antigen specific T-cell receptors are activated and produce and release cytokins. Particularly by IL-1 and IL-2, the immune reaction is initiated and enhanced, and the antigen specific T-cell clones can proliferate and expand.

A subpopulation of T-cells possesses direct cytotoxic activity. These cells are able to attack specifically target cells presenting the antigen by direct cell contact and a complex reaction involving the release of proteins which induce a perforation of the membrane (perforine) and enzymes (granzyme A – H).

There are also *effector cells* which cannot be allocated to the T- or B-cell groups and are called natural killer cells.

B-Lymphocytes. The humoral immunity is a function of B-lymphocytes. They are stimulated by cytokins released from T $_{h2}$cells in cooperation with the antigen-presenting cells. On the surface of B-cells, their specific immunoglobulin is exposed in a way similar to a receptor. If the antigen, e.g., a virus protein on cell surface, is presented in a coordinated and repetitive arrangement, the B-cells may be activated directly without T-cell help [6].

A common principle of T- and B-cells is that a huge repertoire of preformed cells exists, each of which reacts specifically via its T-cell receptor or surface immunoglobulin to a certain antigen. If the body is confronted with such an antigen, there will be a clonal expansion of the lymphocytes which are able to react specifically with this antigen, and this expansion leads to a strong cytotoxic or antibody-mediated immune reaction.

2.1.4. *Thrombocytes*

Thrombocytes, also called blood platelets, are the smallest blood cells. They contain no nucleus, but several specialized organelles. Thrombocytes are derived from megakaryocytes in the bone marrow. In contrast to the other blood precursor cells megakaryocytes do not undergo cell divisions, but endomitosis leading to a polyploid huge cell nucleus, surrounded by a growing cytoplasm, which is roughly “granulated” at the end of maturation. The big “granules” are the preformed platelets, which develop by progressive spreading of a membrane system throughout the megakaryocyte cytoplasm and thus demarcation of the future platelets, and simultaneously development of the organelles. The maturation takes a few days, and one megakaryocyte can produce up to about 2000 platelets. At the end of maturation, the platelets fall apart from their mother cell and enter the blood stream, leaving the nude nucleus.

Thrombocytes have a life span of about one week in the circulation, where they are “a patrol”, the task of which is to detect any disturbances of the integrity of blood vessels and to seal them rapidly in order to prevent excessive blood losses. Due to this function the thrombocytes are the first line of hemostasis. Therefore, the time between a standardized cut into the skin and the termination of bleeding (“bleeding time”) is mainly a function of the thrombocytes. As long as the thrombocytes circulate through vessels with an intact endothelium, they are in a resting state; it has been calculated, that a platelet may pass more than 8000 times through the whole circulation during his life span without being activated. Aged platelets are cleared by the reticulo-endothelial system, mainly in the spleen. The spleen is also a kind of a storage pool which may host about a quarter to a third of the total platelet mass and may release them into the circulation by contraction in case of a challenge like a major injury. Spleenectomized individuals tend to have higher platelet counts in their blood.

Platelets have a discoid shape in the resting state and do not contain a nucleus. They possess characteristic and specific organelles: The α-granules contain proteins, mainly platelet factor 4 and β-thromboglobulin — which are anti-heparin factors — fibrinogen and coagulation factors, as well as thrombospondin, plasminogen activator inhibitor (PAI), and the platelet derived growth factor (PDGF). A very characteristic type of granules are the dense bodies, which contain serotonin

and adenosine diphosphate (ADP). The so-called dense tubular system is involved in the production of prostanoids. The open canalicular system is a system of channels spreading throughout the cell, enabling a rapid transport of the constituents to the outside of the cell, which is important for the release reaction upon activation of the thrombocytes. The outer plasma membrane of thrombocytes contains in the resting state only few negatively charged phospholipids, but a high concentration of the negatively charged phosphatidyl serine is located at the inner layer of the cell membrane. Of functional importance are also the cytoskeleton and in particular contractile proteins (actinin and myosin) which are important for the shape change (see below), and the clot retraction.

The thrombocytes are very reactive cells and are rapidly activated, whenever they come into contact with other surfaces than the intact endothelial cells. The are activated, in particular, by subendothelial structures such as collagen fibers or foreign surfaces such as glass. In contrast to granulocytes, thrombocytes do not have the capacity of locomotion. They are passively driven by the blood stream to the site of lesion. It has been estimated, that during the normal bleeding time of about five minutes about 10^{12} platelets may pass the defect of the vessel wall, but only about 10^5 are found in the clot. The platelets are particularly reactive in areas of fast and turbulent blood flow, where a considerable shear stress is imposed to them. This is one of the reasons, why thrombocytes play a much more prominent role in arterial than in venous vessels. The activation of thrombocytes leads to a rapid, microscopically visible reaction, the so-called shape change: The cell body changes from a discoid to a spherical shape, and a number of pseudopods is extruded. The electrical charge of the outer membrane changes by a fast transfer of phosphatidyl serine from the inner to the outer surface, and the platelets become “sticky”. The electrical charge is also of great significance for the procoagulant function of platelets, since the activated platelet membrane is a template for complexes of activated coagulation factors (see below). Besides the shape change, the thrombocytes undergo also the so-called release reaction, which means the rapid extrusion of mediators, which are important for the activation of surrounding thrombocytes and for other processes, such as the initiation of blood coagulation, and promotion of cell growth in the injured vessel wall by PDGF. In the activation of platelets, presently three pathways can be differentiated. The first pathway is mainly triggered by platelet adhesion to, e.g., collagen fibers, and involves the formation of prostanoids. By the action of a phospholipase, arachidonic acid is liberated from the platelet membrane and subsequently metabolized to prostanoids. An important step is the generation of cyclic peroxides by the enzyme cyclooxygenase, which is a target for the antiplatelet effect of aspirin. The cyclic peroxides are further metabolized to thromboxane A_2, which is one of the most potent mediators of platelet activation and vasoconstriction (see also ⟶ Cardiovascular Drugs; ⟶ Prostaglandins). In endothelial cells, there is also a cyclooxygenase pathway, which generates a counterpart to thromboxane A_2, prostacyclin, which is a potent inhibitor of platelet activation and a vasodilator (⟶ Prostaglandins). The second pathway of thrombocyte activation is mediated by ADP, which is stored in high amounts in the dense body granules of the platelet. ADP acts via a specific P_2T purine receptor. The ADP-pathway is also stimulated by thromboxane A_2, serotonin, and adrenaline (epinephrine). ADP may be liberated from platelets also in a turbulent blood stream with high shear stress leading to cell damage. The third pathway of platelet activation is mediated by thrombin, the key enzyme of blood coagulation (see below). Besides the three pathways outlined above, there are additional agonists and antagonists, such as platelet activating factor or nitric oxide. Generally, after activation by a primary agonist, the other pathways lead to a strong enhancement of the activation in a sense of a positive feed back reaction. The intracellular events associated with activation by the above agonists, include an increase of the intracellular calcium concentration, the formation of diacylglycerol and inositol triphosphate, leading to an activation of proteinkinase C, and an increase of cyclic AMP. An overshooting reaction should be prevented by natural antagonists, such as prostacyclin or nitric oxide.

The thrombocyte surface possesses a number of functionally important receptors, which have been mentioned above: A purine receptor, a thromboxane A_2 receptor, a thrombin receptor, an α_2-adrenergic receptor, and a platelet activating factor receptor. Besides that, there are a number of functionally important glycoproteins (GP), which act as receptors for specific structures. Important examples are (1) the GP IIb-IIIa, a receptor for fibrinogen, von Willebrand factor (vWF, a multimeric protein composed of up to 40 subunits each having an apparent molecular mass of 500 000), fibronectin, and vitronectin, which is important for platelet aggregation; (2) GP Ib-IX, a receptor for von Willebrand factor, which is important for platelet adhesion; and (3) GP Ia-IIa, a receptor for collagen. Platelet adhesion means the adherence of thrombocytes to subendothelium or foreign surfaces, which involves adhesive proteins and particularly von Willebrand factor. Adhesion has to be differentiated from aggregation, which means sticking of platelets to each other in the blood stream. Aggregation is effectively inhibited by aspirin, blocking the thromboxane A_2 pathway, while aspirin does not substantially influence adhesion. During aggregation, fibrinogen forms “bridges” between GP IIb-IIIa of aggregating platelets.

2.2. Plasma Proteins

2.2.1. *Analysis of Plasma*

Plasma, the cell-free part of blood, contains numerous different proteins, each having certain and essential function(s). Depending on the desired aim plasma proteins may be separated by different means. Generally, a distinction is made between analytical and preparative purposes. These purposes dictate which technique to apply. For analyzation of plasma electrophoretic techniques are preferred because they are both easy to perform and of remarkable resolving power. In this case plasma is subjected to an electric field whereby the proteins are separated into different groups. The resolution depends upon the electrophoretic technique applied. Using very common but relatively poor resolving techniques (e.g., zone electrophoresis) plasma is separated into six major fractions, constituting the most prominent proteins: Albumin, α_1-, α_2-, β-, γ-globulins, and fibrinogen (Fig. 7). With more elaborated separation (and detection) techniques (see below), however, plasma proteins can be separated into much more distinct fractions, hence allowing also the detection of minor constituents (examples are given in Table 4): From Table 4 it becomes obvious that plasma proteins may be classified according to several criteria, among others normal concentration, size, electrophoretic mobility, function.

Figure 7. Electrophoretic analysis of normal plasma. Typical separation of plasma proteins by electrophoresis

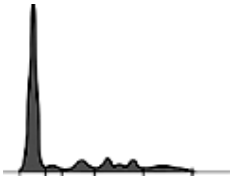

The anode is to the left. After electrophoresis the protein bands are stained and subsequently scanned by densitometry.

Table 4. Plasma proteins

Protein	Plasma concentration, mg/mL	Biological function	M_r
Albumin	40 – 50	transport, colloidosmotic	67 000
IgG Globulin	8 – 18	antibody	155 000
Fibrinogen	2 – 4.5	coagulation	340 000
Transferrin	2 – 4	iron transport	76 000
α_1-Proteinase inhibitor	2 – 3.5	inhibitor of serin proteases	54 000
α_2-Macroglobulin	1.5 – 3.5	general inhibitor of proteases	725 000
α_1-Lipoprotein	1.7 – 3	lipid, hormone transport	28 000
IgA Globulin	0.9 – 4.5	antibody	160 000
IgM Globulin	0.6 – 2.5	antibody	975 000
Haptoglobin 1-1, 2-1, 2-2	0.9 – 3	hemoglobin binding,	86 000
β-Lipoprotein	0.6 – 1.5	lipid, hormone transport	27 000
Hemopexin	0.5 – 1.1	hemin binding	57 000
C3-Component	0.5 – 1.2	complement factor	185 000
α_2-HS Glycoprotein	0.4 – 0.85	opsonin	49 000
Inter- α-trypsin inhibitor	0.4 – 0.7	trypsin inhibitor	160 000
α_1-Antichymotrypsin	0.3 – 0.6	chymotrypsin inhibitor	68 000
Ceruloplasmin	0.15 – 0.6	oxidase	132 000
Gc-Globulin	0.2 – 0.55	vitamin D3 binding	51 000
C1-Inhibitor	0.15 – 0.35	protease inhibitor	104 000
β_2-Glycoprotein I	0.15 – 0.3	heparin binding	40 000
C4 Component	0.2 – 0.5	complement factor	206 000
Prealbumin	0.1 – 0.4	thyroxin, retinol binding	55 000
Antithrombin III	0.15 – 0.2	inhibitor of activated coagulation factors	60 000
C3-Proactivator	0.1 – 0.4	proteinase	90 000
C1q-Component	0.1 – 0.25	complement factor	400 000
C1r-component	0.1	complement factor	83 000
C9-Component	0.08	complement factor	71 000
C8-Component	0.08	complement factor	153 000
C6-Component	0.075	complement factor	105 000
C7-Component	0.055	complement factor	92 000
Plasminogen	0.06 – 0.25	zymogen of plasmin	91 000
β_2-Glycoprotein III	0.05 – 0.15	unknown	35 000
Prothrombin	0.05 – 0.12	zymogen of thrombin	72 000
C5-Component	0.04 – 0.15	complement factor	180 000
9.5S α_1-Glycoprotein	0.04 – 0.08	constituent of amyloid deposits	308 000
Retinol binding protein	0.03 – 0.06	transport of retinol (vitamin A)	21 000
8S α_3-Glycoprotein	0.03 – 0.05	unknown	220 000
Transcortin	0.04	cortisol transport	56 000
C1s-Component	0.02 – 0.04	complement factor, esterase	85 000
Factor XIII	0.01 – 0.04	fibrin cross-linking, transglutaminase	340 000
Properdin	0.01 – 0.03	complement regulator	184 000
C2-Component	0.01 – 0.03	complement factor	110 000

Thyroxin binding glycoprotein	0.01 – 0.02	binding of thyroxin,	54 000
Factor X	0.01	zymogen of FXa	62 000
Zn α_2-Glycoprotein	0.02 – 0.15	unknown	41 000
Lysozyme	0.007 – 0.02	bacteriolysis	14 500
Serum cholinesterase	0.005 – 0.015	cleavage of cholin esters	348 000
Factor V	0.005	coagulation cofactor	330 000
Factor IX	0.004	zymogen of FIXa	62 000
Protein C	0.004	zymogen of inhibitor of coagulation	62 000
IgD Globulin	0.003 – 0.1	antibody	175 000
IgE Globulin	0.001 – 0.0014	antibody	190 000
Factor VII	0.0005	zymogen of FVIIa	50 000
Factor VIII	0.00015	coagulation cofactor	280 000

The *size* of a given protein is determined by the number of amino acids it is composed of and by posttranslational modifications. Most plasma proteins are modified by attachment of nonprotein side chains to the molecules (typical examples are sugar residues and phospholipids).

The *electrophoretic mobility* is determined by the relative amounts of positively, negatively, or uncharged amino acids, as well as by structure and size, depending upon which individual separation technique is applied (native gel electrophoresis, denaturing gel electrophoresis under reducing and nonreducing conditions, respectively, and isoelectric focusing, the latter being the system with the highest resolving power). Also, combinations are possible, e.g., isoelectric focusing followed by denaturing SDS gel electrophoresis.

The numerous proteins of plasma execute manifold functions. Some prominent examples are immunoglobulins as antidotes against foreign molecules, albumin for transport function, coagulation factors for sealing lesions, cytokins for the exchange of information between cells, numerous inhibitors for, e.g., the control of enzymatic reactions against overshooting.

To characterize single protein components of plasma these must normally be isolated from the other plasma constituents. For this purpose they are subjected to a series of separation techniques, each taking advantage of a different characteristic of the protein in question. Typical examples for fractionation sequences are salt (or alcohol) precipitation, hydrophobic interaction chromatography, ion exchange chromatography, affinity chromatography (if available), gel filtration, in certain cases reversed-phase chromatography.

2.2.2. *Coagulation Factors*

As mentioned above, coagulation factors have evolved as a response of organisms against injuries. It is now generally accepted that the system of coagulation may be divided into two branches, the extrinsic system and the intrinsic one (see Fig. 8). The extrinsic branch represents the immediate and direct response to lesions of the organism. This part of the coagulation system is composed of factor VII/VIIa and tissue factor, two coagulation components which come into contact with each other only in the case of disturbances of the normal tissue organization. The intrinsic branch (represented by factors XII, XI, and IX) may be interpreted as an additional amplification mechanism which is turned on, once the extrinsic system has become activated. As shown in Figure 8, both branches are interconnected at the level of coagulation factor X. Downstream of factor X both systems ultimately lead to the formation of fibrin by thrombin. From Figure 8 it can also be derived that normal coagulation is restricted to the vessel walls due to the dependence of the necessary events upon phospholipids. As the complete system is composed of a cascade of protease zymogens (i.e., enzyme precursors), this system is prone to overshooting and must be under tight and rapid control. This is fulfilled by specific and efficient inhibitors, such as antithrombin (AT) and tissue factor pathway inhibitor (TFPI). In addition, thrombin is transformed from a procoagulant protein into an anticoagulant protein by interaction with thrombomodulin.

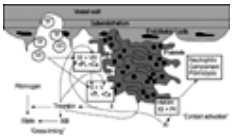

Figure 8. Scheme of blood coagulation

The events of the clotting cascade are illustrated schematically here. Only the unactivated coagulation factors are presented with the exception of factor IIa (thrombin) for simplicity. The organism's immediate response to a tissue lesion is shown on the left (FVII/TF to thrombin). The contact activation pathway (via FXII to FIX) is also illustrated. The whole process is dependent upon phospholipids which are contained in the cellular components of the system (vessel wall and platelets).

TF = Tissue factor; HMWK = High molecular weight kininogen: Pk = prekallikrein

Coagulation factors are listed according to their historical order of discovery by giving the term coagulation factor (often shortened to factor) followed by roman numbers. Most of the coagulation factors must be altered before they can fulfill their physiological function. These proteolytically modified coagulation factors are labeled by adding the letter “a” for activated.

Coagulation Factor I (Fibrinogen). Fibrinogen [9001-32-5], M_r 340 000, is the most prominent coagulation factor in plasma (3 mg/mL). Fibrinogen is the precursor of fibrin [9001-31-4], which is generated as the final event of clot formation. During normal coagulation fibrinogen is converted to fibrin by thrombin. Other enzymes of nonplasma origin, however, can induce clotting as well, e.g., papain and several proteolytic components of certain snake venoms. Several microorganisms, such as *Staphylococcus* also can cause coagulation by cleaving fibrinogen.

Fibrin, the cleavage product of fibrinogen is poorly soluble in aqueous salt solutions. This low solubility is, in the course of coagulation, further reduced by activated factor XIII (see below) to allow clot stabilization.

Coagulation Factor II (Prothrombin). Prothrombin [9001-26-7], M_r 72 000, is contained in normal plasma at a concentration of 100 µg/mL. It is a glycosylated protein containing about 5 % carbohydrate. Like several other coagulation factors it is produced in the parenchymal cells of the liver. For normal prothrombin synthesis vitamin K is required, as is for the synthesis of other zymogens of the coagulation pathways. During activation to the serine protease thrombin the protein's molecular mass is nearly halved to 37 000 whereas the other coagulation factors (see below) are shortened by few amino acids only when activated (the leaving peptide termed activation peptide). Although thrombin is the final protease of the coagulation cascade it plays an important role in regulating its own activity and production. This is accomplished by both, feedforward activity in activating cofactors V and VIII and negative feedback activity by activating inhibitors of coagulation including its own (protein C and protein S). Thus, thrombin is not only a procoagulant but simultaneously an important regulator of normal hemostasis. A rather common genetic variation in the prothrombin gene (G to A transition) at position 20210 leads to elevated prothrombin levels in plasma and is considered to increase the risk for venous thrombosis in people affected.

Coagulation Factor III (Thromboplastin, Tissue factor). Thromboplastin [9035-58-9] was initially discovered as a poorly defined substance that in coagulation experiments ultimately led to thrombin formation. It was then found to be composed of one part protein noncovalently bound to one part of heterogeneous phosphatides. The protein component is tissue factor (M_r 47 000), a glycoprotein with Ca^{2+}- dependent affinity to factors VII and VIIa. Unlike the other coagulation cofactors it does not circulate in the blood but is integrated into cellular membranes not normally in contact with the blood stream. A tissue injury results in exposure of tissue factor to blood, thus enabling its contact with FVII/ FVIIa. It is assumed to stabilize factor VIIa in a proteolytically active conformation to catalyze the conversion of FIX and FX to FIXa and FXa, respectively. The designation factor III meanwhile has been replaced by listing the respective components individually, namely tissue factor, phospholipids.

Coagulation Factor IV (Calcium). Calcium ions are required for most individual steps of coagulation as catalyst. The cation acts as a bridge between phospholipids in cellular membranes and carboxy glutamyl residues in certain coagulation factors. This results in a conformational change in the respective coagulation factor leading to formation of a complex which can be activated by the protease produced in the previous step of the coagulation cascade. Similar to the situation with FIII, calcium is the currently used denomination.

Coagulation Factor V (Plasma Accelerator Globulin). Factor V, M_r 350 000, is the inactive precursor to factor Va (see factor VI, below). Like factor VIII (see below) it is not a zymogen. Its concentration in plasma is about 10 µg/mL. After activation of factor V by thrombin, the resulting factor Va serves as a cofactor for the activation of prothrombin to thrombin by activated factor X, thus facilitating feed forward amplification of the clotting process. A genetically determined defect termed “factor Va Leiden” (Arg 506 is exchanged for Gln) results in activated protein C (APC) resistance, thus leading to an increased risk for thromboembolic complications. This genetic defect is almost exclusively restricted to humans of the western hemisphere. In this population, however, this defect is quite evident (one out of ten people is affected by this mutation).

Coagulation Factor VI (Activated Factor V). Unlike all other terms for coagulation factors, factor VI is the designation for (just) the activation product of another clotting factor (FV, see above), therefore it is somewhat confusing and its use is discontinued.

Coagulation Factor VII (Proconvertin). Proconvertin [9001-25-6], M_r 50 000, is the least prominent vitamin K dependent coagulation factor with a plasma concentration of 0.5 µg/mL. Together with tissue factor and phospholipids, activated factor VII (FVIIa) forms the initiator of coagulation as a response to tissue damage by activating factor X (extrinsic coagulation pathway) and by activating factor IX (intrinsic pathway). In contrast to the other coagulation factors some (about one in thousand molecules) activated factor VII is reported to circulate in the blood.

Coagulation Factor VIII (Antihemophilic Globulin A). Factor VIII [9001-27-8], M_r 280 000, circulates in plasma in tight association with von Willebrand factor. Due to its tight association with vWF the molecular mass of FVIII has long been overestimated. The physiological function of this tight association is assumed to lie in transport and protection of FVIII. In addition, vWF mediates the association of FVIII with platelets (see Section Thrombocytes). The normal plasma concentration of FVIII is 150 ng/mL, classifying it as a trace protein. Hereditary defects in FVIII synthesis are the cause for hemophilia A, leading to severe bleeding disorders. After activation by thrombin (or FXa), FVIIIa serves as a cofactor for the activation of FX by FIXa at a cellular phospholipid surface.

Coagulation Factor IX (Christmas Factor, Antihemophilic Factor B). Like the other factors of the prothrombin complex (FII, FVII, FX, and protein C and protein S) factor IX [9001-28-9], M_r 62 000, is a glycoprotein. It also contains several γ-carboxy glutamic acid residues that are essential for its function. FIX has a plasma concentration of 4 µg/mL. Similar to the situation with FVIII, genetically determined deficiency in this coagulation factor results in a bleeding disorder and a prolonged partial thromboplastin time, called hemophilia B in this case. Factor IX is the final zymogen of the intrinsic coagulation pathway. Once activated, the resulting enzyme together with activated FVIII, phospholipids, and calcium forms the tenase complex, that activates factor X.

Coagulation Factor X (Stuart Prower Factor). Factor X [9001-29-0] is the link between the intrinsic coagulation pathway and the extrinsic one, the former pathway being considered as an amplification mechanism of the latter. After activation by either FIXa (intrinsic) or FVIIa (extrinsic) to FXa this enzyme, with FVa as cofactor, phosholipids, and calcium ions forms the prothrombinase complex that converts prothrombin to thrombin. It can be interpreted as a feed forward amplification that activated factor X together with tissue factor can activate FVII, thus enabling the activation of additional FIX and FX molecules. Factor X has an apparent M_r of 67 000 and is contained in normal plasma at a concentration of 10 µg/mL. Unlike the other procoagulants of the prothrombin complex nonactivated FX is composed of two peptide chains held together by a

disulfide bridge.

Coagulation factor XI [9013-55-2], M_r 160 000, is a zymogen involved in the intrinsic coagulation pathway resulting by contact with foreign material (contact activation). It is composed of two identical peptide chains held together by a disulfide bond. After activation by factor XIIa (see below) it will activate FIX. The plasma concentration of FXI is 5 µg/mL.

Coagulation factor XII (Hagemann Factor), [9001-30-3], M_r 90 000, is the first zymogen in the contact activation of the coagulation system and the only one which is not dependent upon the presence of calcium ions for function. Its normal plasma concentration is 40 µg/mL. No bleeding disorders are reported as a consequence of a deficiency in this factor, people affected rather carry an increased risk for thromboembolic disorders. FXII plays an important (although not fully understood) role not only in the coagulation system but also due to its link to the kallikrein and plasmin systems.

Coagulation Factor XIII (Fibrin Stabilizing Factor). Unlike the aforementioned zymogens factor XIII [9013-56-3] is not the precursor of a protease, but of a transglutaminase. Once activated by thrombin, the enzyme stabilizes the fibrin clots by introducing covalent links between lysine and glutamine residues in individual fibrin molecules thus rendering the clot insoluble. It also introduces covalent links between other plasma proteins, e.g., fibrin, vitronectin, and α 2-plasmin inhibitor. Plasma FXIII is a heterodimer ($FXIIIA_2B_2$) composed of two subunits each, called FXIII-A (the proenzyme) and FXIII-B (formerly called FXIII-S). Activation of FXIII to FXIIIa is accomplished by splitting of an activation peptide from the A-chains. The plasma concentration of FXIII (M_r 320 000) is 20 µg/mL.

Protein C, M_r 62 000, is the zymogen of an inhibitor of coagulation. Like FX it is a two-chain molecule. Its plasma concentration is 5 µg/mL. After activation by thrombin bound to thrombomodulin (a “receptor”-like protein on the surface of endothelial cells), protein C together with its cofactor (activated protein S, see below) will inactivate activated factors V and VIII, respectively. Thus it acts as a modulator of coagulation by interference with FXa generation as well as thrombin formation. Due to its charge characteristics protein C is found in the prothrombin complex (factors II, VII, IX, X) during plasma fractionation. Genetically determined deficiency in protein C results in elevated risk for thromboembolic complications.

Protein S, M_r 69 000, has a plasma concentration of 10 µg/mL. Besides its cofactor function for activated protein C it has an anticoagulant effect via direct interaction with FVIIIa. Similar to the situation with protein C, genetically determined protein S deficiency results in an increased risk for thromboembolic complications.

2.2.3. *Fibrinolysis System*

The organisms needs an efficient and fast system to produce a clot in any case of violation of blood vessels. The generation of a fibrin matrix is also important for wound healing and inflammatory reactions. However, of equal importance is an efficient system for fibrin degradation. This fibrinolytic system is needed for the prevention of overshooting clotting, or at least, for recanalization and repair after thrombotic events. Intra- and extravascular fibrinolysis is also necessary for removal of the fibrin matrix along with the progressive “organization” during the healing of wounds and inflamed areas.

The main fibrinolytic system in the blood is the plasmin pathway. It is activated by the proteolytic cleavage of the proenzyme plasminogen [9001-91-6], M_r 93 000, a protein with a plasma concentration of about 2 µM (200 mg/L). In its N-terminal part plasminogen contains five characteristic domains, the so-called kringle domains, and the rest of the plasminogen molecule contains the active center of a serine protease. The kringle domains are important for the binding of plasminogen to fibrin and thus the interaction with the substrate of the activated enzyme. The activation takes place by cleavage between Arg_{560}-Val_{561}. The activation of plasminogen does not involve a complex cascade, such as the coagulation system. However, several physiologic pathways of activation exist. The most important pathway involves the tissue-type plasminogen activator (t-PA), which is synthesized predominantly in endothelial cells. The expression of t-PA can be enhanced by several mediators, the most important being thrombin. t-PA contains also two kringle domains, binds to fibrin and activates plasminogen. Further domains of t-PA, whose physiologic significance is not yet clear, are a so-called “finger” domain, which is also found in fibronectin, and a domain which is similar to epidermal growth factor. Another important physiologic plasminogen activator is the urokinase-type plasminogen activator (u-PA), which may be found as “pro-urokinase”, a single chain molecule (scu-PA), and the cleaved form urokinase, devoid of the N-terminal epidermal growth factor-like and kringle domain. The name pro-urokinase is somewhat misleading, since also this molecule has already some fibrinolytic activity. It has been speculated that pro-urokinase circulates bound to some “inhibitor”, which gets lost by binding to fibrin, where pro-urokinase can be readily further activated by proteolytic cleavage, e.g., by traces of plasmin. The fibrinolytic system is also activated by the so-called “contact activation” involving kallikrein and FXIIa, which is at least in part mediated by conversion of scu-PA to urokinase.

Plasmin is not only involved in fibrin cleavage, but has also a variety of other substrates [7]; of importance for hemostasis are the proteolytic cleavage of t-PA and scu-PA enhancing its own activation, and on the other hand the inactivation of coagulation factors V and VIII.

Plasmin can degrade both fibrinogen and fibrin at several cleavage sites. The result are several fibrin(ogen) degradation products of different size (X, Y, D, E). These fibrin(ogen) degradation products, particularly the large X-fragments, lead to a disturbance of fibrin polymerization and have thus an anticoagulatory effect. Fibrin degradation products (FDP) can be detected by several laboratory tests; specific tests are available for fibrinogen or fibrin degradation products. The demonstration of FDP in the plasma may be associated with an abnormal activation of fibrinolysis, e.g., during a disseminated intravascular coagulation. An isolated overstimulation, called hyperfibrinolysis can rarely be found in patients with certain tumors, e.g., prostate carcinoma. In diseases with a very strong inflammatory reaction, such as sepsis, pancreatitis, or polytrauma, a massive release of proteolytic enzymes from neutrophil granulocytes, particularly elastase, occurs which may contribute to intra- and extravascular fibrinolysis [5].

2.2.4. *Plasma Proteinase Inhibitors*

Most of the proteins involved in the coagulation and fibrinolytic system are proteolytic enzymes. These proteinases belong all to the class of serine proteinases. Their activity must be regulated because otherwise they would be a threat to the host that they are designed to protect. This problem has been solved, for instance, by the evolution of a series of proteins which decrease the activity of the proteolytic enzymes and which are called proteinase inhibitors (Table 4). These proteinase inhibitors are also class specific and therefore this chapter is dealing with proteinase inhibitors of serine proteinases. The inhibitors rapidly form tight stoichiometric complexes with proteinases in a substratelike manner which leads to inactivation of the proteolytic enzyme. Cleavage of the inhibitor may occur, if at all, at a much slower rate compared with a substrate. In the case of serine proteinase inhibitors, inhibition results from an interaction of the proteinase substrate binding site and a region of the inhibitor which is known as reactive site loop. In the case of canonical inhibitors like the kunins, the conformation of the reactive site loop is highly complementary to the surface of the enzyme and resembles a substrate bound to the specificity site [8], [9]. Unlike the Kunitz-type inhibitors, antithrombin has a reactive site loop which is not held rigidly by disulfide bonds but is able to swing in a hingelike manner. In the resting state, the reactive site loop projects away above the bottom of the serpin. When the reactive center docks with thrombin, the Arg393–Ser394 reactive site peptide bond is cleaved. At this stage, the complex is covalently linked. Normally, in the next step a substrate would dissociate after cleavage regenerating the active enzyme. In contrast, antithrombin remains linked to thrombin by one arm of the now cleaved reactive site loop. This arm becomes inserted into the body of the inhibitor as an additional β-pleated sheet. This leads to a stable complex in which the enzyme is crushed against the body of the inhibitor, distorting it, which results in irreversible inhibition [10]. It should be mentioned that α_2-macroglobulin is unique with respect to its completely different mechanism of inactivating proteinases and its ability to inhibit members of the four classes of proteinases [11].

Proteinases which initiate and amplify the coagulation and fibrinolytic systems are thought to have a restricted ability to cleave peptide bonds. Their role is to activate the zymogen of other components of the system usually by a single peptide bond cleavage. Proteinase inhibitors act to limit the activation by rapidly complexing key enzymes of the amplification system. However, they do not completely shut down the pathways, but rather inhibit systemic activation and thereby localise coagulation and fibrinolysis. The physiological importance of proteinase inhibitors may be seen from their relative plasma concentration. If one excludes albumin and immunoglobulins, about 20% of the remaining proteins are proteinase inhibitors. Although there are inhibitors of cysteine and metallo proteinases in plasma, most of the inhibitors regulate the serine proteinases. At this time, inhibitor concentrates of antithrombin III (AT III), α_1-proteinase inhibitor (α_1-PI), and C1-Inhibitor (C1-Inh) are commercially available. Therefore, these three inhibitors are described in the following.

Antithrombin III, AT III; M_r 58 000, [12], [13] is a single chain glycoprotein and a member of the serpin family. It is synthesized in the liver and circulates in plasma with a half-life of about 3 days. The primary structure exhibits 432 amino acid residues, six of which form 3 intramolecular disulfide bonds. The carbohydrate moiety of the protein contributes to the molecular mass by about 15%. The plasma concentration of AT III is about 2.6 μmol/L and the reference activity is given with 70-120%. Two isoforms exist in plasma, AT- α and AT- β. About 90% of the plasma AT III is present as the α-isoform. The two isoforms differ in their degree of glycosylation: the α-form exhibits four glycosylation sites, whereas the β-form lacks the carbohydrate at Asp135. AT- β has a greater affinity for heparin and for both the intact and injured vessel wall [14].

Antithrombin III is the major inhibitor of thrombin, factor Xa, and factor IXa but it also inactivtes factors XIa and XIIa as well as serine proteinases which do not belong to the coagulation system like plasmin, kallikrein, and complement C1.

Inhibition of proteinases by AT III is slow if compared other proteinase–proteinase inhibitor complex formation. The inhibitory action of AT III is greatly accelerated (about hundredfold) by heparin or related polysaccharides [15]. Among the most pertinent are heparan sulfate proteoglycans which are found on or underlying the vessel wall where anticoagulant activity occurs. While the order of magnitude of the rate enhancement for inhibition of thrombin, factor IXa, and factor Xa by antithrombin in the presence of heparin there is a fundamental difference in the mechanism of enhancement of inhibition. Thrombin inhibition is accelerated because the heparin chain provides a template to bring the reactive site of the inhibitor and the active site of thrombin in close proximity. This is observed with heparin oligosaccharides of size greater than 18 monosaccharide units but not for those of smaller size [16], Even though they are still inducing a conformational change in antithrombin which leads to the optimal inhibitory conformation of the reactive site loop [17]. The affinity of heparin to the thrombin-antithrombin (TAT) complex is much lower than to the free enzyme and heparin therefore dissociates from TAT complexes. The complexes are removed from circulation by receptors in the liver (half-life < 5 min). In contrast, accelerated inhibition of factor Xa is mainly achieved by the induced conformational change of AT III which is observed even with a pentasaccharide.

Antithrombin deficiency may be acquired or inherited. The pattern of inheritance is usually autosomal dominant (non-sex-linked) with levels of biologically active inhibitor around 40-70% of normal. Deficiencies of AT III but also of protein C and S are the most common known genetic causes for thrombosis. Venous thrombosis is a serious medical problem which affects annually one in thousand individuals causing morbidity and death. Commonly, antithrombin deficiencies are divided in two groups, type I and type II, based on the results from immunological and functional assays [18], [19]. Type I is characterized by a 50 % reduction of antigen (and activity) caused by heterozygous mutation of the gene. Homozygous AT III type I deficiency appears to be not compatible with life. Type II deficiencies are those cases where the antigen level is normal but functional activity is reduced (usually by a heterozygous mutation). Type II deficiencies are subdivided in mutations affecting the reactive site (RS), the heparin binding site (HBS). Acquired AT III deficiencies are due to either decreased liver synthesis, increased loss, increased consumption, or drug-induced.

α_1-Proteinase inhibitor, M_r 52 000, α_1-PI) is a single-chain glycoprotein and a member of the serpin family. It is synthesized primarily in the liver but also in mononuclear phagocytes, intestinal epithelium and kidney parenchyma. α_1-PI circulates in the blood with a half-life of 3-5 days and is cleared by the reticulo endothelial system. The primary structure exhibits 394 amino acid residues, containing one free cysteine residue and three carbohydrate side chains [20].

α_1-PI inhibits a large number of serine proteinases including trypsin, chymotrypsin, cathepsin G, plasmin, thrombin, tissue kallikrein, factor Xa, plasminogen, and neutrophil elstase. During acute phase response α_1-PI is upregulated in order to control neutrophil derived proteases at inflammatory sites. It is implicated that α_1-PI plays also a role in immune response which include inhibition of lymphocyte response to phytohemagglutinin, inhibition of chemotaxis, and enhancement of pokeweed mitogen driven IgG synthesis [21].

Protease inhibitor and individual alleles are named according to their isoelectric focusing characteristics (PI M, PI S, PI Z). In addition to the polymorphism at gene level, the protein exhibits microheterogeneity due to differences in glycosylation and N-terminal processing. Inheritance of α_1-PI alleles is autosomal codominant: Each allele contributes to the total serum concentration and overall isoelectric focusing pattern. Phenotypes are PI MM (two normal alleles) PI ZZ (homozygous inheritance of two defect Z alleles) and PI SS (homozygous inheritance of two defect S alleles) [22]. α_1-PI-deficiency affects about one in 6000 caucasian americans but only one in 40 000 to one in 100 000 black americans. 95 % of all α_1-PI-deficient patients belong to the PI ZZ which is characterized by a plasma level of 1.8 - 3.6 mg/mL (10 - 28 mg/mL for normal phenotypes, PI MM; threshold level for protection against lung damage due to proteolytic activity 10 - 12 mg/mL). Clinical manifestation of α_1-PI-deficiency is emphysema in adults and liver disease in children.

C1-Inhibitor, M_r 104 000, is a single-chain glycoprotein which belongs to the serpin family. It is mainly synthesized in the liver. Other sites of biosynthesis are skin fibroblast as well as monocytes. C1-Inhibitor circulates in the blood with a half-life of about 67 to 72 hours. Normal plasma levels are in the range of 0.2 mg/mL. The native molecule consists of 478 amino acid residues (including two disulfide bonds) and exhibits an unusually high carbohydrate content of about 50 % [23].

The inhibitor is involved in the regulation of several proteolytic systems including the complement (C1s, C1r), coagulation factor XIa, fibrinolytic (plasmin), and contact (factor XIIa, kallikrein) systems. Hereditary or aquired deficiency of C1-Inhibitor may result in accumulation of vasoactive substances. Sudden death due to acute larynx edema or nonitching, hard, painless swelling of soft tissues, and coliclike pains might result. In type I hereditary angioedema (HAE), an insufficient synthesis and elevated turnover of a normal C1-Inhibitor results levels of 5 - 30 % of normal. In type II, functionally inactive inhibitor is detected and the product of the normal allele is found in low concentrations. In aquired angioedema (AAE) type I, C1-inhibitor and C1 are diminished and in type II of AAE an anti-C1-Inhibitor autoantibody interferes with complex formation between enzyme and inhibitor [23].

2.2.5. *Immunoglobulins*

Immunoglobulins, also called antibodies (⟶ Immunotherapy and Vaccines – Antibodies) are produced by B-lymphocytes (see Section Leukocytes, Cytokins) (⟶ Immunotherapy and Vaccines – Immune Response). Their function is to “recognize” and bind specifically certain molecules, their corresponding antigens. The human organism possesses about 10^6 – 10^8 different antibody specificities and is thus able to react to a great variety of antigens. The immunoglobulins make up a specific humoral immunity and play an important role in the body's defense against pathogenic microorganisms and toxic substances.

As already outlined above, each B-cell carries its own specific immunoglobulin on the outer cell surface, which functions as an antigen receptor. If a certain antigen enters the organism, it binds selectively to corresponding antigen receptors (surface immunoglobulins) on B-cells. By this mechanism, B-cells, which are able to produce suitable antibodies, are selected and activated. Supported by cytokins produced by T-helper cells, a clonal proliferation and differentiation into plasma cells is initiated. The plasma cells produce and secrete high amounts of the specific antibody which mediates the humoral immune answer. The antibody production can last over different periods of time up to many years; in many cases, after the antibody production is down regulated, so called memory cells remain which may reside permanently in lymphatic organs and also may recirculate. By this mechanism, a fast and strong immune reaction can be initiated in case of repeated antigen exposure.

The basic structure of immunoglobulins is that of a Y-shaped four-chain polypeptide (Fig. 9). The polypeptide is a symmetric tetramer of two heavy (i.e., high molecular) chains and two light (i.e., low molecular) chains each. The C-terminal portion of these chains contains the so-called constant regions, while the N-terminal parts (the end of the two branches of the “Y”) have a very variable structure, which is responsible for the high diversity of binding properties corresponding to the numerous antigens, which can be recognized. The variability is brought about by gene rearrangements during the maturation of the B-cell precursors into mature B-cells (see [24]). While the two branches of the “Y”, also called Fab, contain the variable parts responsible for antigen recognition, the C-terminal part, called Fc, is important for effector functions of the immunoglobulins, e.g., complement binding and interactions with inflammatory cells. Immunoglobulins can be cleaved by enzymes like papain or pepsin, which can be utilized to produce fragments with particular characteristics (Fig. 10). Such a cleavage may be of interest in order to achieve a different distribution into body compartments of the Fab-fragments compared to intact immunoglobulins, but may also be utilized if only antigen binding, but not the effector functions mediated by Fc are required.

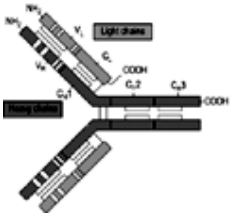

Figure 9. Schematic structure of immunoglobulin: V_L and V_H are variable regions, C_L and C_H constant parts of light and heavy chains

Figure 10. Products of enzymatic cleavage of IgG.

2.2.6. *Albumin*

Albumin is the most abundant plasma protein, with an average concentration of 36 – 50 g/L. It has a molecular mass of 66 000 and is the only plasma protein of this size without carbohydrate content. Albumin is synthesized by the liver. The half life of albumin is about 19 d, it is catabolized in many tissues and some loss occurs through the intestinal tract.

The most important function of albumin is to maintain the oncotic pressure, which is necessary to retain fluid within the vasculature against the hydrostatic blood pressure. While in the arterial part of capillaries the hydrostatic pressure is higher leading to extravasation of fluid, the hydrostatic pressure decreases towards the venous part of the capillaries below the oncotic pressure, which leads to an influx of fluid. This mechanism is of critical importance for the exchange of fluid and metabolites between the extra- and intravascular space and thus for the homeostasis of the body as a whole. Besides this function, albumin is a carrier protein for numerous substances, both physiological compounds as well as medications.

[Top of Page]

3. Production

As outlined in Chapter Composition, Physiology and Chemical Properties of Blood, there are numerous constituents of the blood that can be used as medicines. The majority of products on the market are derived from human blood. The use of animal material is for several reasons only of minor importance. Plasma proteins can be produced also by recombinant techniques, which is so far mainly used for coagulation factors for hemophilia treatment and for thrombolytic drugs.

3.1. Blood and Plasma Donation

The first attempt to transfer blood from one individual to another was in 1492, when blood from three young men was given to the Pope Innocent VII; unfortunately all four died [25]. Also trials with sheep blood were disappointing. A major step forward was finding a way to prevent clotting of the blood; the first nontoxic anticoagulant was sodium phosphate, which was found out by BRAXTON HICKS in 1869. Another breakthrough was the discovery of AB0 blood groups by KARL LANDSTEINER in 1901. Now the stage was set for further refinements and improvements in the collection, storage, and application of blood transfusions. This new medical discipline was strongly stimulated by World War II, when blood donation programs were developed, preservatives such as acid citrate-dextrose (ACD) were invented, and techniques for blood fractionation were developed. Of major importance for the development of a plasma industry was the work of COHN, who developed techniques for plasma fractionation (see Section Fractionation of Plasma).

The fundamental prerequisite for the manufacture of blood components and plasma derivatives remains the willingness of millions of people to donate their blood. One of the most important issues in the whole area is therefore the ongoing information, motivation, and protection of blood donors. Experience has shown that severe diseases, particularly infections with viruses such as human immune deficiency virus (HIV) can be transmitted by blood. In order to protect the recipients of blood products from such risks, careful selection of donors and testing of their donations is necessary. In order to ensure this, standards have been developed by organizations for blood collection (blood systems) and regulatory authorities, both on national and international levels. Although these standards are more or less comparable in developed countries, the organization of blood systems and the particulars of regulations and guidelines in this field may differ considerably not only from continent to continent, but also from country to country. Essential points of donor selection are the physical examination in order to verify a good health status, and an interview or questionnaire in order to find out whether there is any obstacle in the medical history or the behavior of the prospective donor. The latter point may be rather delicate, since, e.g., individuals with unsafe sexual behavior are not acceptable as blood donors.

If a donor is found suitable, the collection of his/her blood can start (in some countries the very first donation of an individual may be discarded as a measure of precaution). For the proper collection of blood, the rules of good manufacturing practice (GMP) have to be strictly followed. This includes suitable premises, well trained personal, and a quality assurance system. Crucial is the maintenance of hygiene from disinfection of the venipuncture site to a clean environment for processing and storage of blood components. The term blood components covers all blood-derived products for direct use as transfusion (e.g., erythrocyte or platelet concentrates, fresh or fresh frozen plasma). The blood is primarily collected into plastic blood bags. The collecting bags are usually centrifuged in order to produce different types of components. Numerous modifications exist, it is for instance possible to produce platelet concentrates from platelet rich plasma (supernatant of a “soft” spin of the blood) or from the so called buffy coat (white layer containing leukocytes and platelets on top of the red cells after a “hard” spin of the blood). In order to avoid a bacterial contamination of the products, the industry developed closed systems, i.e., a collection bag and additional bags in which the components can be transferred after separation, interconnected by appropriate tubes. For some special components it is necessary to open the system (an example is the addition of cryopreservatives to stem cells). For that purpose so called sterile connecting devices (SCD) have been developed which automatically cut and reconnect tubes with high temperature in order to avoid contamination. Of crucial importance for transfusion medicine is the careful documentation of the donation and all relevant data. The components must not be released before the negative results of safety relevant tests (serological markers for HIV and hepatitis B and C infections, additional parameters depending on the country) are available. Careful documentation and labeling is also necessary

because a confusion of components ("the wrong blood to the wrong patient") is a prominent cause of severe and potentially fatal complications.

Blood collection can also be carried out by apheresis techniques. This means, that a part of the blood is separated from the donation and the rest is given back to the donor. In the beginning, this was done by so called "manual" apheresis. An example was the manual plasmapheresis, where blood was taken from the donor and separated into plasma, which was used and the red cells which were retransfused into the donor. In the meantime techniques for "automatic" apheresis have been developed, which involve an extracorporal circulation of blood. In principle, blood flows from the donor via a tubing system into a machine, where a continuous flow differential centrifugation takes place with outlet tubes leading the desired component(s) into collection bag(s), and the rest of the blood back to the donor. By this way plasma can be obtained, as well as cellular components such as platelets or stem cells or erythrocytes.

The above described collection techniques may be used to obtain both blood components for transfusion and source plasma for the manufacture of plasma protein concentrates. The origin of the source plasma, i.e., donor population (geographical and epidemiological background), technique of collection (recovered from whole blood or plasmapheresis) and handling (freezing, storage, transport) may have significant impact on the quality of this important raw material. In order to enable the production on an industrial scale, source plasma is usually pooled; the size of these plasma pools varies considerably according to the facilities used and the product and can be up to several ten thousand donations. The fractionation of these plasma pools is described below.

3.2. Substances Derived from Blood Cells

Some attempts were made in the past to produce crude blood cell extracts or homogenates for medicinal purposes. These preparations are poorly characterized and have been part of alternative medicine in some countries.

However, there are also some examples for specific medicines derived from blood cells. From blood leukocytes β-interferon can be obtained (⟶ Interferons – Natural Leukocyte IFN). A preparation of hemin (derivative of the heme obtained from the hemoglobin of outdated erythrocytes), is used in severe hepatic porphyrias. An old desire in medicine, which had a renaissance in context with the experience of HIV transmission by blood products, is to find blood substitutes, which means nonerythrocyte oxygen carriers. Besides several synthetic substances such as perfluorocarbons, a current approach is to use modified or cross-linked hemoglobins. A limitation of this approach is that higher concentrations of free hemoglobin may be toxic, e.g., for the kidneys. Besides their capacity to transport oxygen, the hemoglobin-based blood substitutes appear to have also some impact on vascular tonus and blood pressure and other indirect effects; such effects may be of therapeutic use.

3.3. Fractionation of Plasma

3.3.1. *Quality Requirements*

In any kind of medical therapy, assessment of the risk-benefit as well as the development of therapeutic standards is an important consideration for the use of a medicinal product. Historically, in the case of hemophilia patients substitution therapy with plasma or plasma-derived products was of such a life-saving importance that other aspects were not considered adequately. However, with the HIV catastrophe this attitude has changed and virus safety of plasma and plasma derived products are issues of highest priority in health policy, the medical community, and the public interest. Although virus safety is obviously of major interest, besides the transmission of viruses and other hitherto unknown pathogens, possible other safety aspects must be considered. Examples are the development of antibodies (inhibitors) to plasma components and the risk of thrombogenicity due to the administration of activated coagulation factors.

In general, the whole manufacturing process contributes to the quality of a product. In the case of blood preparations, this process already starts with the collection of the source material, plasma for fractionation. Measures taken are the selection of donors, testing of single donations and subsequently pooling of single plasma donations to plasma pools, testing of the plasma pools, and the selection of suitable methods for the purification including virus inactivation/elimination steps. All manufacturing steps should be performed under validated conditions and according to cGMP yielding a product of consistent quality, safety, and efficacy. Regulatory approaches in different continents are still not entirely harmonized but their essential elements are similar. For the EU, the blood and plasma collection for both the manufacture of plasma for transfusion and plasma for fractionation is regulated by Directives [26], whereas the quality standard for plasma for fractionation, e.g., quality attributes, viral marker testing of plasma pools as well as freezing, storage, and transport conditions are provided by the monograph "Plasma for fractionation" of the European Pharmacopoeia (Ph. Eur.) [27]. In addition, there are monographs for plasma derived products available, e.g., human Factor VIII, Factor IX, Prothrombin Complex Concentrates, Fibrinogen, Albumin, Immunoglobulins, etc. setting quality standards for these products. Monographs are legally binding documents. Guidelines of the Committee for Medicinal Products for Human Use (CHMP), the scientific working arm of the European Medicines Agency (EMEA) in London [29] and its working groups, e.g., the Biological Working Party (BWP), provide further guidance on the manufacture of plasma derived medicinal products [28]. These guidance documents consider all steps in the manufacture of a plasma-derived product with a focus on infectious pathogens.

3.3.2. *Preparation of Intermediate Products*

Typical examples of plasma fractionation leading to low or intermediate purity are the Cohn fractionation procedure or derivatives thereof [30]. This technique is up to now in use for the production of major protein constituents of plasma (e.g., albumin and immunoglobulins). The Cohn fractionation takes advantage of the different physicochemical properties of proteins resulting in their different behavior at a given temperature in aqueous solutions of differing pH, salt, and alcohol concentration (see Fig. 11). For the purification of less abundant components it is now a widely common technique to separate plasma into the cryoprecipitate (obtained by thawing frozen plasma at or around 4 °C) and into so-called cryo-poor plasma (proteins which stay in solution after this procedure). Redissolved cryoprecipitate has long been in use for the treatment of hemophilia A, because coagulation factor VIII is enriched in this plasma fraction. Cryoprecipitate is also used as a source for fibrinogen which may then be marketed as tissue glue (in combination with thrombin). The cryo-poor plasma fraction is up to now in use for the preparation of intermediate-purity products, such as immunoglobulins, albumin, and

prothrombin complex concentrates (PCC). A derivative of the latter is activated prothrombin complex (aPCC) marketed as factor eight inhibitor bypassing activity (FEIBA) or Autoplex. PCC has long been used for the treatment of, e.g., hemophilia B while it is now occasionally in use for the treatment of acquired coagulation disturbances (e.g., reversal of anticoagulation by treatment with Marcumar during surgery). Activated PCC is given in certain cases of hemophilia A to patients with inhibiting antibodies against FVIII. The PCC is obtained from cryo-poor plasma by batch adsorption to anion-exchange media at relatively low pH, followed by washing and step elution of the coagulation factors. In some protocols a barium citrate precipitation step (or similar) is included in the process. Essentially the same procedure is performed for the preparation of Autoplex (FEIBA), followed by an activation step in order to obtain (among others) activated FVII.

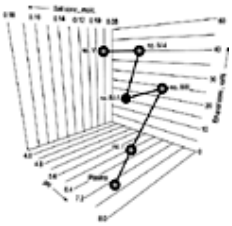

Figure 11. Cohn fractionation of plasma.

Plama fractionation according to Cohn and co-workers Plasma is fractionated by stepwise adding ethanol, adjusting pH, and/ or dilution. Although not illustrated here, the temperature is also carefully controlled. Each step (the corresponding designations in the figure are as in the original article) is followed by removing the resulting precipitate. The individual precipitates contain mainly fibrinogen (no. I), β-, γ- globulin, and prothrombin (no. II/III), α-globulin (no. IV-1), α-, β-globulin (no. IV-4), and albumin (no. V).

3.3.3. *Purification of Plasma Proteins*

The Cohn fractionation is up to now in use for the preparation of albumin and immunoglobulins, representing the major constituents of plasma proteins. Since the late 1980s, however, have not only analytical techniques been very much improved, but the same is true for preparative applications. Typical examples are FIX and FVIII high-purity concentrates which are now available. Both coagulation factors are characterized by their relative scarcity as compared to albumin and immunoglobulins. These novel developments became available through the now widespread use of affinity chromatography (see Section FVIII High Purity Concentrates.) as well as novel chromatography media and column designs (e.g., radial flow technology) allowing higher flow rates, thus shortening the purification process.

FIX High-Purity Concentrates. In contrast to PCC — the therapeutic for the treatment of hemophilia B in former years — FIX is purified by gradient elution of proteins from anion- exchange columns, followed by adsorption to at least group specific affinity columns, e.g., heparin or dextran sulfate modified media. Also in use are chromatography techniques employing immobilized monoclonal antibodies. These techniques lead to purities such that the target protein is at least the major component in these products.

FVIII High Purity Concentrates. Compared to high purity FIX concentrates the purity of so-called high purity FVIII concentrates is much less striking. This, however, is not surprising, considering that on a molar basis FIX is hundred-fold more concentrated in plasma than FVIII (see below). One major contaminant in several commercially available concentrates is serum albumin which is added after purification in order to stabilize FVIII. Beginning with cryoprecipitates, FVIII is purified using anion-exchange chromatography, in certain cases followed by affinity chromatography on monoclonal antibody media. Alternative purification strategies take advantage of the association of FVIII with von Willebrand factor. Being an unusually large protein complex, this association was successfully subjected to size exclusion chromatography, a technique which is normally characterized by poor resolution.

Although not a purification step, it should be mentioned here that for safety reasons all medium- or high-purity products on the market are subjected to two complemantary virus inactivation steps (e.g., heat treatment or solvent detergent treatment) prior to or after filling of the final containers. Before this procedure is performed the proteins are transferred into a buffer system capable of stabilizing the fragile coagulation factors. These stabilizing additives (mainly sugars and/ or amino acids, albumin) represent a contamination. Similarly, antithrombin III and/ or heparin are added to certain coagulation factor concentrates. This step is performed to prevent activation of coagulation factors or to inactivate those which were converted to proteases during production. These components may be considered a contamination, at least from a biochemical view. Generally, the purification of proteins for therapeutic use is considered to be beneficial by removing possibly hazardous contaminants. However, the procedures performed during purification may also be detrimental for the proteins to be purified. This may hold true especially for the purification of coagulation factors as many of these are progenitors of proteases and must not be activated, otherwise the therapeutics may lead to thromboembolic complications. Also, they are extremely sensitive against denaturation (due to their multi-subunit structure as in the case of, e.g., FVIII). The denaturation may lead to the formation of neoantigens which in turn will cause the development of inhibiting antibodies.

3.4. Recombinant Proteins

In order to circumvent the risk of contamination of coagulation factor preparations by human pathogens (Transmittable Pathogens) recombinant gene technology has been introduced for the production of several therapeutics instead of purifying these from human plasma. Recently, certain member states of the European Union (Ireland, UK, Sweden) have decided to use recombinant coagulation factors exclusively where available. Among others, FVIIa, FVIII, and FIX have been cloned and expressed in eucaryotic cells of nonhuman origin from which they can subsequently be purified. The purification strategies follow the same procedures as outlined above for the purification from human plasma to yield products free of human viruses. However, recently the problem of other pathogens has become apparent. At least during the initial phase of eucaryotic cell cultivation serum of animal origin is needed for feeding, thus raising the risk of contamination by nonhuman pathogens which eventually will adapt to the new host. Moreover, there is a theoretical risk that potentially novel pathogens may contaminate recombinant protein preparations. Here, prion proteins, the putative causative agent of bovine spongiforme encephalitis (BSE) or Creutzfeld Jacob disease in man (CJD) should be mentioned [31]. In addition, due to the expression in cells of nonhuman origin, posttranslational modifications may differ in proteins from recombinant sources which eventually lead to the recognition of these products as foreign and the development of antibodies against these agents. Apart from producing recombinant proteins in vitro in eucaryotic cells, expression systems have been developed which allow the production in living animals where the proteins of interest are delivered into the milk and can be purified from this source.

3.5. Adverse Drug Reactions of Blood Products

3.5.1. *Immunogenicity*

Proteins are increasingly used as therapeutic agents (see ⟶ Enzymes – Nonindustrial Enzyme Usage and ⟶ Enzymes – Therapeutic Enzymes). It has been recognized that these medicinal products frequently induce an immune reaction leading to antibody formation. The consequences of an immune reaction to a therapeutic protein range from transient appearance of antibodies without any clinical consequences to severe life threatening conditions. Potential clinical consequences are severe allergic reactions, reduction of efficacy, and induction of autoimmunity, including antibodies to the endogenous protein. Of further concern is the interference of antibody formation with clinical parameters, e.g., recovery, residence time, and the laboratory assays to measure the therapeutic protein [32], [33], [34].

Many factors may influence the immunogenicity of therapeutic proteins. They can be considered to be patient-, disease-, or product-related. Patient-related factors might predispose patients to an immune response, e.g., genetic background, ethnic origin, immune and infection status, and treatment with immune-suppressive medicines. Disease-related factors include various immune disorders. Product/treatment-related factors might also influence the likelihood of an immune response, e.g., intensity of treatment (dose frequency, duration of treatment), route of administration, source of protein (human or animal or recombinant origin), manufacturing process (contaminants, impurity profile), formulation and stability characteristics (degradation products, aggregates) of a given protein.

Development of antibodies against blood components is of special importance for the blood coagulation factors VIII and IX which are used for the treatment of hemophilia A or B patients [35]. These events are the most serious complications during the treatment of hemophiliacs besides the possible transmission of blood-borne pathogens. The antibodies, also called “inhibitors”, can interfere with the biological function of the coagulation factor [36] or can enhance for example the clearance of the coagulation factor [37]. Both events can lead to life threatening complications for the patients. Patients who develop inhibitors are classified as low or high responder according to the level of antibodies. Low responders are those who produce a low level of inhibitor (< 5 B.E.) and who do not react with an increase of inhibitor in case of new coagulation factor injections. High responders are those patients with a high titer of inhibitor (> 5 B.E.) and who react with an increase of inhibitor in case of new coagulation factor injections. Low responders can be treated with higher doses of the coagulation factor concentrate which overcome the inhibitor, whereas for the treatment of high responders other therapies like treatment with factor VIIa or activated prothrombin complex concentrates have to be considered.

Antibodies directed towards functional epitopes of the factor VIII molecule are currently evaluated in vitro by their capacity to inhibit the procoagulant activity of plasma factor VIII. Quantitation is obtained by mixing/incubating dilutions of the inhibitor containing plasma with a fixed amount of factor VIII. The active amount of inhibitor, expressed in units per milliliter plasma, is the reciprocal of the plasma dilution that neutralizes factor VIII activity. For routine analysis, the New Oxford assay (source of factor VIII is a concentrate) and the Bethesda method (source of factor VIII is a pooled normal plasma) have been used [38], [39]. It is now recommended to use a modification of the Bethesda method by which the progressive decrease of pH during incubation is prevented, which can lead to false positive results [40]. Inhibitors can also be detected independent of their functional activity using enzyme-linked immmunosorbent assay (ELISA) techniques.

Inhibitor development against the substitutive product is considered to be due to either the type of substitute or the severity of hemophilia. Patient-related factors for the development of antibodies are poorly understood and a relationship between gene defect and inhibitors is not manifest. Results for hemophilia A patients, however, point out that patients with severe gene defects are more prone to inhibitor development, perhaps due to the absence of circulating factor VIII and the lack of immune tolerance induction [41], [42]. Product-related factors could also be responsible for inhibitor development which was observed in hemophilia A patients treated for several years with factor VIII products without developing inhibitors. When these patients had been exposed to a certain factor VIII product, they developed inhibitors after a variable length of exposure days [43].

3.5.2. *Thrombogenicity*

Thromboembolic complications are the major adverse events during the treatment of patients with prothrombin complex concentrates (PCCs) [44-47]. Prothrombin complex concentrates contain the coagulation factors II, VII, IX and X, protein Z, as well as the coagulation inhibitors protein C and S. These products are indicated for the treatment of patients with decreased levels of factors of the prothrombin complex due to vitamin K deficiencies, overdosing of oral anticoagulants, e.g., vitamin K, or decreased synthesis caused by liver dysfunctions. The three component PCCs (factor II, IX and X) have previously been used for the treatment of hemophilia B but have been replaced by high-purity factor IX concentrates or recombinant factor IX. Different from PCCs are those products (FEIBA, Autoplex) which contain activated factors of the prothrombin complex and which have another indication.

Prothrombin complex concentrates are standardized with respect to their factor IX content. According to the Ph. Eur. monograph [27], PCCs should exhibit an equal composition of the prothrombin complex factors and their content of protein C and S should be indicated. In addition, most of the PCCs have heparin and antithrombin III added. PCCs have to comply with requirements of the nonactivated partial thromboplastin time and the thrombin fibrinogen clotting time. As these parameters, however, have been found to be of low predictive value with respect to in vivo correlation, additional measures should be followed by the treating physician in order to minimize the risk of thrombotic complications [46], [47].

During a 4-years survey from 1987 to 1990, 72 case reports on complications with PCCs were published [48]. Major risk factors of thrombogenicity have been reported to be (1) predisposing factors, i.e., the patient's underlying disease; (2) therapy factors, i.e., dosage (zymogen overload leading to a prothrombotic state) and concomitant therapy; and (3) quality of the PCCs. With respect to the quality aspect of PCCs, activated coagulation factors II, VII and IX as well as the absence or low levels of antithrombin III and protein C and S have been discussed to cause the clinical signs of thrombogenicity which manifest as phlebitis or venous thrombosis, pulmonary embolism, cerebral or myocardial infarction, disseminated intravascular coagulation with bleeding, and organ failure due to consumption of coagulation factors and fibrin formation. Coagulant active phospholipids, which have been detected in PCCs, were also discussed as potent thrombogenic triggers potentiating the effect of activated coagulation factors [49]. During the last few years, scientific data were accumulating which showed without a doubt that prothrombin overload is causing thromboembolic complications while treating patients with

PCCs [50-52].

3.5.3. *Transmittable Pathogens*

One immanent risk of blood products is the transmission of pathogens. This issue is therefore of particular concern and appropriate measures have to be taken by the manufacturer to prevent the transmission of infections by the use of plasma-derived medicinal products [53]. Those pathogens include HBV, HCV, HIV-1 and HIV-2, HAV, and Virus B19. During the recent years, emerging pathogens, e.g., variant CJD (vCJD) and West Nile virus (WNV) had to be added to this list of pathogens with importance for the safety of blood products. Therefore, the collection of the source material, plasma for fractionation, is of great importance in the assurance of the quality and safety of medicinal products derived from blood. Factors which should be considered for the collection of plasma (blood) include various factors like suitability of donors, screening of donations and epidemiology of the population from which plasma is collected. The risk assessment for medicinal products derived from plasma (blood) should consider the amount of virus which may contaminate the manufacturing plasma pool versus the capacity of the manufacturing process to remove or inactivate the pathogens. By considering the amount of starting material needed to manufacture a single dose of product, the probability of potential virus contamination can be estimated [54].

A recently emerged infection in North America is the WNV which belongs to the *Flaviviridae* family. Clinical symptoms in humans occur 2-14 days after infection and viraemia within 1-3 days after infection and lasts 1-11 days. Thus, an infected person could be viraemic before symptoms are seen. As mentioned above, the viral safety of plasma-derived medicinal products involves several approaches and re-evaluation those for the purpose of WNV lead to the assurance that the measures in place today are adequate to assure viral safety of plasma-derived products with respect to WNV [55].

Tracing recipients of blood transfusions from donors who subsequently developed vCJD has revealed possible cases of secondary transmission. This finding is consistent with the observation for sheep transfusion studies that infectivity in blood can be transmitted during preclinical phase of infection. To date many studies have been performed, which investigate the contribution of various manufacturing steps to reduction of infectivity. However, caution is needed in interpretation of the data since the effectiveness of a step is dependent on a number of variables, e.g., process conditions and state of the agent in the sample. Consequently, effectiveness of removal may vary from process to process. As vCJD is an emerging disease, uncertainties still exist about the number of cases of vCJD that will occur and the transmission of the infective agent by blood donations. Today, no screening test is available and therefore approaches are considered to identify donors who may present a higher risk. Therefore it is recommended to exclude donors who have spent a cumulative period of one year in the United Kingdom from donating blood/plasma for fractionation. In view of the lack of adequate information on vCJD, it is prudent to recall batches of plasma-derived medicinal products where a donor to a pool subsequently develops vCJD. Also medicinal products containing plasma-derived products as excipient should be recalled [56]. This issue is kept under review of the CHMP and its working groups.

[Top of Page]

4. Therapeutic Uses

Medicines derived from blood and plasma play an important role in the medical therapy in general. In most cases, they are used to replace the respective components or proteins in case of their loss or insufficient production. Particularly, the development of contemporary surgery, which is often associated with blood loss, would not have been possible without the possibility of transfusion. Also in hematology and oncology, and of course in the treatment of bleeding disorders, blood products are indispensable.

4.1. Blood Components

4.1.1. *Blood and Stem Cell Preparations*

Erythrocytes. The most important components in terms of the amounts used and the potentially immediate life-saving effect are those containing erythrocytes. Whereas initially whole blood was the mainstay of hemotherapy, nowadays erythrocyte concentrates are prevailing. These concentrates are conventionally produced from whole blood donations by centrifugation, and can be obtained also by modern apheresis techniques. With contemporary additives, erythrocyte concentrates may have a shelf life of up to 7 weeks. It is a well-documented experience [57] that immunization of blood recipients against donor antigens is mainly mediated by contaminating leukocytes. Immunization can cause severe allergic reactions or impair the therapeutic effect to subsequent transfusions. Also other unwanted effects of transfusions have been attributed to their leukocyte content. Thus, “leukodepletion”, the removal of leukocytes by filtration or modification of apheresis, is increasingly recommended, despite the increase of expenses for the concentrates.

There are several guidelines [59], but no rigid rule regarding the indication for erythrocyte transfusion. In every individual patient, the indication should be considered carefully. For instance, young, otherwise healthy patients tolerate lower Hb-levels than elderly patients, particularly those with cardiovascular disorders. Chronic anemia is far better compensated than acute blood loss. Therefore, the conventional recommendation of 10 g/dL Hb and below as a trigger for transfusion is at best a rough “rule of thumb”.

Platelets. A lack or functional disturbances of thrombocytes lead to a severe bleeding tendency. The typical clinical pattern is petechiae and bleeding from mucous membranes (e.g., nose, gingiva) but also life-threatening CNS bleedings. A bleeding tendency may be present at thrombocyte counts around 30 000/µL; a clear increase of the bleeding risk necessitating platelet transfusions occurs below 10 000/µL. Functional disturbances may be inherited such as defects of GP IIb-IIIa (thrombasthenia Glanzmann), or of GP Ib-IX (Bernhard Soulier Syndrome). More frequent are acquired functional defects in liver disease or renal insufficiency, or by medications such as analgesics or antibiotics. On the other hand, platelet inhibition may be a target of therapy in thromboembolic diseases, particularly in arterial events such as myocardial infarction or stroke. The classical substance for platelet inhibition is aspirin, but a number of new substances against specific receptor pathways are currently developed and evaluated.

Leukocytes. The use of leukocyte (e.g., granulocyte) concentrates is still restricted to very special circumstances, since considerable problems occur with unwanted effects, e.g., acute lung injury.

Stem Cells. An increasingly important field is the production and use of hematopoietic stem cells [61]. Stem cells can be obtained from aspirated bone marrow, from peripheral blood by apheresis, and from placenta via the umbilical cord; also fetal liver contains stem cells. Currently, the most important use is for "rescue" of hematopoiesis after otherwise lethal chemotherapy and/or radiation for treatment of several malignancies, or after damage or aplasia of the bone marrow (bone marrow or stem cell transplantation). Also "autologous" stem cells taken from the patient and reinfused after intensified antineoplastic treatment are in use; strategies are developed to purify the CD-34 positive stem cells, or to remove malignant cells ("purging") from the preparation. Still in the experimental stage are approaches to augment the stem cells in culture ("ex vivo expansion"), or to manipulate them by incubation with antigens and cytokins in order to obtain a specific immunological therapy. Of so far unforeseeable importance are stem cells as a vehicle for gene therapy.

4.1.2. *Fresh Frozen Plasma*

Most of the whole blood donations are separated into cells and plasma; plasma can also be obtained by plasmapheresis (see Section Blood and Plasma Donation). Plasma can be used as source for industrial production of protein concentrates, or as a component for direct transfusion. For the latter purpose, plasma is usually rapidly frozen as soon as possible (fresh frozen plasma) and stored in a freezer in order to preserve the sensitive proteins such as coagulation factors.

Plasma as a therapeutic component has the advantage that it is the physiologic fluid part of blood with a balanced content of proteins, particularly albumin, coagulation factors, inhibitors, and immunoglobulins, as well as electrolytes and other natural constituents. Therefore, plasma is in principle an ideal substitute in case of blood loss, in combination with red cells as clinically appropriate. However, it is no more recommended to use plasma just for volume replacement, since it is available in limited amounts only, and since alternatives, so-called "plasma expanders", e.g., dextrans or hydroxyethyl starch, are available. Moreover, plasma still carries a low residual infectious risk; there are strategies to keep plasma in quarantine until the donor is tested again after at least 6 months, to exclude the possibility of donations during the diagnostic window, i.e., the time between infection and the detectability of virus markers. There are also attempts to apply virus inactivation steps to single donations, or to pooled plasma, which is subsequently divided again into units of the usual size.

The main indication of plasma transfusion is the prevention and treatment of profound and complex disorders of blood coagulation, such as consumptive thrombohemorrhagic disorders, in surgical and intensive care patients. In the treatment of those disorders, e.g., disseminated intravascular coagulation, a multitude of factors and inhibitors are consumed, and plasma may be a suitable way to replace all of them in a balanced way. Large volumes of plasma may have to be transfused in order to obtain relevant replacement in full-blown consumption, potentially leading to fluid overload; some clinicians use therefore concentrates alone or in combination with plasma. The therapy of such complex disturbances, however, is far from being standardized, and should be directed by experienced specialists taking into account the individual clinical picture. In the treatment of hemophilia or other single-protein deficiencies, suitable concentrates should be used. In general, the indications for plasma transfusion are hard to define, and should be further evaluated in clinical studies.

4.2. Hemoglobin-Derived Oxygen Carriers

The driving forces behind the development of "artificial blood" are the avoidance of the risk of infection by allogeneic blood transfusion, and to overcome the problems posed by blood grouping and the limited shelf life of erythrocyte concentrates. In fact, hemoglobin isolated from blood, which may be modified e.g., by cross-linking, is capable of effective oxygen transport. These preparations may elicit also some increase of blood pressure, which might be interesting in particular circumstances.

The original hope to have a new, virtually unlimited source of oxygen carriers which would supersede conventional blood components has been darkened so far by the finding of substantial toxicity of larger doses of these carriers. Free hemoglobin outside the "parent" erythrocytes is bound to a specific transport protein, named haptoglobin, and has to be cleared from the circulation. If the capacity of haptoglobin and the clearance system is exhausted, free hemoglobin is excreted via the kidney, and may lead to kidney failure similarly as in the case of massive myoglobin release in severe trauma ("crush kidney"). Also the new hemoglobin-derived substances are still limited in their dose. It might turn out that they have a potential in the emergency ambulance, e.g., for immediate on-site resuscitation of accident victims. Hemoglobin-derived substances are also evaluated for other indications, e.g., oxygen delivery and pressor effect in septic shock, or sensitization to radiotherapy in cancer patients. However, the intriguing approaches appear to need substantiation by further research.

4.3. Cytokins

A growing number of cytokins have effects of potential therapeutic interest (see Section Leukocytes, Cytokins and Table 2). Here, only some examples already in clinical use are mentioned.

One of the first therapeutic cytokins was interferon (⟶ Interferons), due to its antiviral and potential antineoplastic effects. Interferon is used in chronic viral hepatitis and may lead to clinical remission in part of the patients [63]. It is not a universal antitumor drug, but in several malignancies, e.g., in chronic myeloid leukemia, interferon is a valuable part of the therapeutic repertoire.

Increasingly used in hematology and oncology are also the hematopoietic growth factors such as G-CSF or GM-CSF. The may be useful to stimulate granulopoiesis in patients during the aplastic phase after chemotherapy, and also for "mobilization" of stem cells from the bone marrow in order to harvest them by apheresis. A new, promising, substance is thrombopoietin, which stimulates platelet production. Another substance which is conventionally classified as a hormone should also be mentioned here: erythropoietin, which is physiologically produced by the kidneys. Recombinant erythropoietin has proved to be of great value for patients with chronic renal failure, which used to be dependent on transfusions before. Erythropoietin is also used in certain hematological disorders.

4.4. Coagulation Factors

Congenital coagulation factor deficiencies mostly affect the level of a single factor, while acquired deficiency (e.g., due to impaired synthesis in liver disease, loss or dilution of blood, or consumption in sepsis) includes usually many or all factors. All coagulation factors are contained in plasma, which may be used for the treatment of complex factor deficiencies (see Section Fresh Frozen Plasma), concentrates are recommended in cases where specific factors are to be administered. This concept of a targeted therapy helps to avoid fluid overload, and an unwanted and possibly overshooting increase of factors which are normal or even elevated at baseline. This chapter covers treatment of congenital, one-factor deficiencies. There is a trend to high purity concentrates, since the lacking factor can be substituted in a more controlled fashion, and impurities are kept low. Such impurities could be associated with unwanted effects, such as immunosuppression, or thrombogenicity [65], [44].

The classical indication for the use of concentrates are the hemophilias A (lack of factor VIII), and B (lack of factor IX). Both are congenital disorders mediated by X-chromosomal inheritance, and thus affect almost exclusively males. The disease can be expressed with different severity depending on the underlying gene defect; the severe hemophilias (levels of < 1 % of the normal plasma level) are usually manifested already in early childhood, when the babies start crawling and moving around and thereby may hit themselves mildly, which would be easily compensated by a healthy coagulation, but may lead to severe bleeding in hemophiliacs. The typical symptoms are large hematoma of the skin, and bleeding into muscles and joints; bleeding at other sites including the particularly life-threatening central nervous system bleeds also occur. The frequency of hemophilias is about 1 in 10 000 males. Being a congenital defect, hemophilia necessitates life-long treatment. The therapy may be given in case of symptoms (“on demand treatment”), many patients with severe hemophilia are substituted with a fixed continuous schedule of usually two to three injections per week (“prophylactic treatment”); a typical prophylactic single dose would be 30 iU per kilogram of body weight.

While during the early 1980s a disastrous epidemic of HIV infections, transmitted by coagulation factor preparations, hit the hemophiliac population, the safety of the concentrates has been increased fundamentally in the meantime by careful selection and testing of blood donors, and virus inactivation procedures. Currently, the most important problem of treatment is the occurrence of so-called “inhibitors”, which means neutralizing antibodies, during therapy. These inhibitors usually occur during the initial phase of treatment around the 10th to the 20th exposure to the factor, which is “strange” to the recipient's immune system. The inhibitors only rarely cause allergic symptoms (predominantly those against factor IX). Therefore they may develop unnoticed and may cause severe problems in case of bleeding episodes, since the effect of the transfused factor is impaired or even blocked. It is recommended to check the patients regularly for inhibitors, particularly during initial treatment. Inhibitors occur much more often in hemophilia A, which is also the far more frequent form of the hemophilias. The treatment of such inhibitor patients has two aims: (1) to stop and control acute bleeding and (2) to eliminate the inhibitor. For stopping and controlling of bleeding, high doses of the involved factor, PTC, “activated” PTC, or (in hemophilia A patients only) porcine factor VIII or, as a newer approach, recombinant factor VIIa are in use. For the elimination of inhibitors, also referred to as “immunetolerance treatment”, prolonged treatment with frequent and high doses of factor VIII (“Bonn protocol”), as well as a combination with extracorporal removal techniques and immunosuppressive treatment (“Malmö protocol”) have been developed [66]. The elimination of factor IX inhibitors is particularly problematic, since allergic and anaphylactic reactions and, in case of high doses of factor IX, immune complex disorders such as nephritis [67] may develop. The occurrence of inhibitors is a frequent problem (in hemophilia A in up to 50 % of patients). In two clusters, inhibitors with unusual characteristics were associated with certain products [68], [43], which had to be taken off the market.

Another congenital bleeding disorder, which is even more frequent than hemophilia, is von Willebrand disease. However, since there are a lot of different mutations and thus functional and clinical categories, the classification of severity and the therapeutic consequences are much less clear-cut than in the hemophilias. A lack or dysfunction of von Willebrand factor may lead to a reduction of factor VIII levels and/or disturbances of platelet adhesion; the clinical symptoms resemble in many cases more a platelet defect with prolonged bleeding time, petechiae, and mucous membrane bruising. Since the inheritance is autosomal (chromosome 12), both sexes are equally affected. It is strange that the clinical severity may differ from one affected family member to the other, and even during time course in an individual. By far the most patients have mild symptoms, and it is not unusual to detect the disease in an adult. Mild forms may be treated with a vasopressine analogue (DDAVP), which mobilizes residual von Willebrand factor from its production site, the endothelial cells; this therapy is effective only a limited time (several days). In any case, an accurate diagnosis should be established first, and the effect should be evaluated in a test treatment, before, e.g., surgery is to be covered by DDAVP; in some forms of dysfunctional von Willebrand factor DDAVP may even be contraindicated. In some countries, factor VIII concentrates containing von Willebrand factor are registered for substitution in severe forms of von Willebrand disease.

Other congenital coagulation factor deficiencies are less frequent or even rare; an example is factor XIII deficiency with an estimated prevalence of 1 in 1×10^6 people. Factor XIII is a fibrin-stabilizing factor, which is important to make clots stable and resistant against premature fibrinolysis. Characteristic symptoms are umbilical cord bleeding of newborns, repeated re-bleeding after initial hemostasis, wound healing disturbances and spontaneous abortions. Unfortunately, there is a quite high frequency of central nervous system bleeds. In the past, a concentrate had been manufactured from placenta; nowadays, for treatment, a plasma-derived concentrate is registered in some countries. The significance of an acquired factor XIII deficiency needs further elucidation; there are intriguing data pointing to a role of factor XIII in wound healing [69].

4.5. Plasminogen Activators

Plasminogen activators are used in the thrombolytic therapy of severe thromboembolic disorders such as myocardial infarction and lung embolism. The physiologic task of the fibrinolytic system is the prevention of an overshooting clotting. However, if clinically significant thrombosis occurs, the natural fibrinolysis may take considerable time (several months) in order to remove the fibrin in the thrombus. In many cases, the thrombus is perpetuated by “organization”, i.e., the conversion into a connective tissue “scar”. Particularly in emergencies like myocardial infarction, a very fast recanalization is needed.

The first, and still widely used substance was streptokinase, a plasminogen activator derived from streptococci. Streptokinase forms a complex with a plasminogen, and this complex activates free plasminogen to plasmin. The disadvantage of streptokinase is mainly its antigenicity. Also a number of physiologic plasminogen activators exist, and numerous attempts have been made to modify them by recombinant techniques. The most widely used are urokinase and t-

PA. While urokinase, similarly as streptokinase, produces free plasmin leading to a massive degradation of fibrinogen and coagulation factors and thus to a pronounced disturbance of coagulation, t-PA acts predominantly at the fibrin clot and has therefore also been named a “fibrin specific” thrombolytic agent. The initial hope that this would reduce bleeding complications was not fulfilled at the high t-PA doses needed for rapid thrombolysis. However, t-PA appears to act somewhat faster than the other thrombolytic agents which might be an advantage in certain emergencies. However, urokinase and particularly t-PA are very costly.

Compared to a thrombophilic state due to a congenital deficiency of coagulation inhibitors (see Section Inhibitors), inherited disorders of the fibrinolytic system are much less frequent and the clinical significance is less clear. Though there are some reports about an enhanced thrombotic risk for individuals suffering from disturbances of t-PA expression or release, and from a plasminogen deficiency [70], there was recently a report of a total plasminogen deficiency in several patients who did not experience serious thromboembolism [71].

4.6. Inhibitors

The congenital inhibitor deficiencies usually pertain to a single inhibitor — although evidence points to a higher prevalence of combined defects than in coagulation factors (Section Coagulation Factors — whereas an acquired reduction usually includes several inhibitors. The organism contains several proteolytic systems and respective protease inhibitors; this chapter is focused on the inhibitors relevant in the hemostatic system. Congenital deficiencies of *inhibitors of the fibrinolysis system* may cause severe, hemophilia-like bleeding, but are very rare; there are no specific concentrates available. Congenital deficiencies of *coagulation inhibitors* have attracted increasing interest in the past decade, and “thrombophilia” has been coined as a new term.

A total deficiency of antithrombin III is lethal, and total deficiency of protein C is associated with the life-threatening purpura fulminans of newborns. In contrast to, e.g., hemophilia (see Section Coagulation Factors), most symptomatic patients with deficiencies of the coagulation inhibitors antithrombin III, protein C, and protein S are heterozygotes with plasma levels of around 50 % of normal, which makes the diagnostic discretion from borderline normals sometimes difficult. The affected family members show an increased tendency to mostly venous thromboembolism at a relatively young age; in antithrombin III deficiency, up to 80 % of the affected individuals may experience a thrombotic event before the age of 40 years. In contrast to hemophilia, there is no indication for prolonged substitution, with the exception of particular risk situations like major surgery and in the peripartal period in pregnancy and after delivery. Usually, it is recommended that the patients are treated by prophylactic oral anticoagulation (cumarin treatment) after the first thrombotic manifestation. A problem in patients with protein C or protein S deficiency is that these proteins are also vitamin K-dependent and their level is further lowered during cumarin treatment. Particularly in the initial phase of anticoagulation, a critical dysbalance between coagulation factors and protein C may occur; a typical clinical complication is the so-called cumarin necrosis.

In some countries, antithrombin III concentrates are licensed, which are used predominantly in acquired deficiency, as in disseminated intravascular coagulation. However, a clinical study failed to show a reduction of mortality in severe sepsis [72]. Recently, recombinant activated protein C was licensed for treatment of severe sepsis [73]. Well-controlled clinical studies are extremely difficult to perform in those severe and complex disorders.

4.7. Immunoglobulins

The most common application of immunoglobulins is the replacement of endogenous immunoglobulins in patients with immune deficiency syndromes compromising the specific immune defense. In addition, high doses of immunoglobulins may be used to overcome certain autoimmune disorders, such as idiopathic thrombocytopenic purpura (ITP) [74]. A further therapeutic segment is the application of specific immunoglobulins, which can be gained from immunized donors. A typical example is immunoglobulin against the rhesus antigen D, which is used for prophylaxis of immunization of rhesus negative mothers by rhesus positive babies, which could lead to a severe hemolytic disorder in further rhesus positive offspring.

Immunoglobulins may also be produced by recombinant techniques in order to obtain antibodies with specific pharmacological effects. Examples are antibodies against the platelet GP IIb-IIIa (see Section Thrombocytes) as a potent antiplatelet drug in high risk situations such as thrombolysis for myocardial infarction, or antibodies with specific antitumor action, which may be conjugated with antineoplastic substances.

4.8. Albumin

The use of albumin has been scrutinized following a meta-analysis of Cochrane and co-workers, who suggested an increased mortality. Though this conclusion was not substantiated, it is recommended that the indication for human albumin should focus on the use of albumin for restoration and maintenance of circulating blood volume where volume deficiency has been demonstrated and use of a colloid is appropriate [75]. Haemodynamic parameters should be monitored in patients receiving albumin due to the risk of hypervolaemia and cardiovascular overload.

[Top of Page]

5. Economic Aspects

Blood is a very precious resource from the ethical standpoint, since it is a generous gift of voluntary donors. Blood and plasma and the derivatives thereof are also of high economical value, since there are considerable expenses for the infrastructure of donation centers, donor selection and testing, storage of labile components, handling and transportation of source plasma, and industrial production of concentrates, which includes plasma fractionation, purification of proteins and virus inactivation steps. Furthermore, the quality and safety of the products should be continuously refined by ongoing research, for instance on new pathogens or their potential to cause immunological or thrombotic side effects. The treatment of patients with blood and plasma products is therefore expensive, particularly for patients who need lifelong treatment for congenital diseases such as hemophilia or immune deficiencies. This is one of the reasons, why a functional blood system and particularly supply of plasma products is so far guaranteed mainly in the developed countries; it is estimated that currently about 80 % of all hemophiliacs worldwide do not receive medically adequate treatment.

Since the financial budgets of the health care systems are limited also in economically strong nations, there is an increasing pressure to use these resources as economically as possible. The decisions in every single country are strongly influenced by medical traditions and ethical views. However, a scientific basis for estimation of the cost vs. benefit ratio of certain expensive therapeutic interventions is increasingly advocated. A typical example is the competition between plasma-derived and recombinant coagulation factors. This necessitates well-designed basic research and clinical trials, as well as solid comprehensive concepts requiring the input of scientists, clinicians, health care suppliers, regulators, politicians and, of course, the patients and the citizens, who pay for the social systems.

Despite discussions about health care budgets, it should not be forgotten, how immense the benefit of the development of transfusion is for medicine as a whole, and how dramatically the life expectancy and the quality of life of hemophiliacs improved by the introduction of replacement therapy. Thus, also in the modern medicine of increasingly sophisticated invasive procedures and the upcoming paramount of gene therapy, the medicines derived from blood and plasma should keep an important place.

[Top of Page]

References

1. H. D. Bruhn *et al.* (eds.): “Megakaryocytopoiesis and Platelet Production in Health and Disease”, *Hämostaseologie*, **16** Schattauer-Verlag Stuttgart 1996, pp. 78 – 163.
2. R. S. Taichman: “Blood and bone: two tissues whose fates are intertwined to create the hematopoietic stem-cell niche”, *Blood* **105** (2005) 2631–2639. Links
3. N. S. Young, J. Maciejewski, “Mechanisms of Disease: The Pathophysiology of Acquired Aplastic Anemia,” *N. Eng. J. Med.* **336** (1997) 1365 – 1372. Links
4. M. Heiden, R. Seitz, R. Egbring, “The role of inflammatory cells and their proteases in extravascular fibrinolysis”, *Semin. Thromb. Hemost.* **22** (1996) 497 – 501. Links
5. E. Backström, K. Kristensson, H.-G. Ljunggren: “Activation of Natural Killer Cells: Underlying Molecular Mechanisms Revealed”, *Scand. J. Immunol.* **60** (2004) no. 1-2, 14–22. Links
6. M. F. Bachmann, R. M. Zinkernagel, “Neutralizing Antiviral B Cell Responses,” *Ann. Rev. Immunol.* **15** (1997) 235 – 270.
7. K. A. Hajjar, R. L. Nachmann: “The Human Endothelial Cell Plasmin-Generating System,” in: R. W. Colman, J. Hirsch, V. J. Marder, E. W. Salzmann (eds.): *Hemostasis and Thrombosis: Basic Principles and Clinical Practice*, 3rd ed., J. B. Lippincott Company, Philadelphia 1994, pp. 823 – 836.
8. M. Laskowski, Jr., I. Kato: “Protein inhibitors of proteinase”, *Ann. Rev. Biochem.* **49** (1980) 593 – 626.
9. W. Bode, R. Huber: “Natural protein proteinase inhibitors and their interaction with proteinases”, *Eur. J. Biochem.* **204** (1992) 433 – 451. Links
10. K. Gomez, J.H. McVey, E. Tuddenham: “Inhibition of coagulation by macromolecular complexes”, *Haematologica* **90** (2005) 1570–1576. Links
11. L. Sottrup-Jensen: “Alpha-macroglobulins: structure, shape, and mechanism of proteinase complex formation”, *J. Biol. Chem.* **264** (1989) 11539 – 11542. Links
12. U. Abildgaard: “Binding of thrombin to antithrombin III”, *Scand. J. Clin. Lab. Invest* **24** (1969) 23 – 27. Links
13. R. D. Rosenberg, P. S. Damus: “The purification and mechanism of action of human antithrombin-heparin cofactor”, *J. Biol. Chem.* **248** (1973) 6490 – 6505. Links
14. J. Swedenborg: “The mechanisms of action of alpha- beta-isoforms of Antithrombin”, *Blood Coagul. Fibrinolysis.* (1998) Nov. 9 Suppl 3, S7–10. Links
15. I. Bjork, S. T. Olson, J. D. Shore: “Molecular Mechanism of the Accelerating Effect of Heparin on the Reactions Between Antithrombin III and the Clotting Proteinases,” in D. A. Lane, U. Lindahl (eds.) *Heparin: Clinical Applications, Chemical and Biological Properties*, Edward Arnold, London 1989.
16. B. Bray, D. A. Lane, J. M. Freyssinet, G. Pejler, U. Lindahl: “Anti-thrombin activities of heparin, Effects of saccharide chain length on thrombin inhibition by heparin cofactor II and by antithrombin”, *Biochem. J* **262** (1989) 225 – 232. Links
17. R. Carell, R. Skinner, L. Jin, J. P. Abrahams: “Structural mobility of antithrombin and its modulation by heparin”, *Thromb. Haemost.* **78** (1997) 516 – 519. Links
18. D. A. Lane, R. J. Olds, M. Boisclair, V. Chowdhury, S. L. Thein, D. N. Cooper, M. Blajchman, D. Perry, J. Emmerich, M. Ajach: “Antithrombin III mutation database: first update. For the Thrombin and its inhibitors Subcommittee of the Scientific and Standardization Commitee of the International Society on Thrombosis and Haemostasis', *Thromb. Haemost.* **70** (1993) 361 – 369. Links
19. D. A. Lane, T. Bayston, R. J. Olds, A. C. Fiches, D. N. Cooper, D. S. Miller, K. Jochmans, D. J. Perry, K. Okajima, S. L. Thein, J. Emmerich: “Antithrombin mutation database: 2nd (1997) update. For the Plasma Coagulation Inhibitors Subcommittee of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis”, *Thromb. Haemost.* **77** (1997) 197 – 211. Links
20. G. Salvesen, S. V. Pizzo, J. D. Shore: “Proteinase Inhibitors: α-Macroglobulins, Serpins, and Kunins,” in R. W. Colman, J. Hirsch, V. J. Marder, E. W. Salzman (eds.): *Hemostasis and Thrombosis: Basic Principles and Clinical Practice*, 3rd ed., J. B. Lippincott Company, Philadelphia 1994, pp. 241 – 258.
21. C. A. Blank, M. Brantly: “Clinical features and molecular characteristics of alpha 1-antitrypsin deficiency”, *Ann. Allergy* **72** (1994) 105 – 122. Links
22. M. K. Fagerhol, *Haematologica* **1** (1968) 153 – 161.
23. A. E. Davis, 3rd: “C1 inhibitor and hereditary angioneurotic edema”, *Ann. Rev. Immunol.* **6** (1988) 595 – 628.
24. G. D. Yancopoulos, F. W. Alt, “Regulation of the assembly and expression of variable-region genes”, *Ann. Rev. Immunol.* **4** (1986) 339 – 368.
25. D. M. Harmening (ed.): *Modern Blood Banking and Transfusion Practices*, 3rd ed., F.A. Davis Company, Philadelphia

(1994).

26. Directive 2002/98/EC of the European Parliament and of the Council of 27 January 2003 setting standards of quality and safety for the collection, testing, processing, storage, and distribution of human blood and blood components and amending Directive 2001/83/EC; Commission Directive 2004/33/EC of 22 March 2004 implementing Directive 2002/98/EC of the European Parliament and of the Council as regards certain technical requirements for blood and blood components; Commission Directive 2005/61/EC of 30 September 2005 implementing Directive 2002/98/EC of the European Parliament and of the Council as regards traceability requirements and notification of serious adverse reactions and events"; Commission Directive 2005/62/EC of 30 September 2005 implementing Directive 2002/98/EC of the European Parliament and of the Council as regards Community standards and specifications relating to a quality system for blood establishments.
27. Monograph of human plasma for fractionation 01/2007:0853, European Pharmacopoeia, Strasbourg 2005.
28. www.emea.europa.eu
29. European Medicines Agency, "Note for guidance on plasma-drived medicinal products", CPMP/RWP/269/95
30. E. J. Cohn *et al.*, "Separation into Fractions of Protein and Lipoprotein Components," *J. Am. Chem. Soc.* **68** (1946) 459 – 475. Links
31. M. E. Bruce, R. J. Will, J. W. Ironside, I. McConnell, D. Drummond, A. Suttie, L. McCardle, A. Chree, J. Hope, C. Birkett, S. Cousens, H. Fraser, C. J. Bostock: "Transmissions to Mice Indicate that "New Variant" CJD is Caused by BSE Agent," *Nature* **389** (1997) 498 – 501. Links
32. H. Schellekens: "Immunogenicity of therapeutic proteins: clinical implications and future prospects", *Clin. Therap.* **24** (2002) 1720–1740.
33. S. Hermeling, D.J. Crommelin, H. Schelekens, W. Jiskoot: "Structure-immunogenicity relationship of therapeutic proteins", *Pharm. Res.* **21** (2004) 897–903.
34. E. Koren, L.A. Zuckerman, A.R. Mire-Sluis: "Immune response to therapeutic proteins in humans - Clinical significance, assessment and prediction", *Curr. Pharm. Biotechnol.* **3** (2002) 349–360.
35. J. G. Gilles, M. G. Jacquemin, J.-M. Saint-Remy: "Factor VIII inhibitors", *Thromb. Haemost.* **78** (1997) 641 – 646. Links
36. M. S. Gawryl, L. W. Hoyer: "Inactivation of factor VIII coagulant activity of two different types of human antibodies", *Blood* **60** (1982) 1103 – 1108. Links
37. I. Nilsson, E. Berntorp, O. Zettervall, B. Dahlbäck: "Nonecoagulation inhibitor factor VIII antibodies after induction of tolerence to factor VIII in hemophilia A patients", *Blood* **75** (1990) 378 – 383. Links
38. C. R. Rizza, R. Biggs: "The treatment of pateints who have factor-VIII antibodies", *Br. J. Haematol.* **24** (1973) 65 – 82. Links
39. C. K. Kasper, L. Aledort, D. Aronson, R. Counts, J. R. Edson, J. van Eys, J. Fratantoni, D. Green, J. Hampton, M. Hilgartner, P. Levine, J. Lazerson, C. McMillan, J. Penner, S. Shapiro, N. R. Shulman: "Proceedings: A more uniform measurement of factor VIII inhibitors", *Thromb. Diath. Haemorrh.* **34** (1975) 612. Links
40. B. Verbruggen, I. Novakova, H. Wessels, J. Boezeman, M. van den Berg, E. Mauser-Bunschoten: "The Nijmegen modification of the Bethesda assay for factor VIII:C inhibitors: improved specificity and reliability", *Thromb. Haemost.* **73** (1995) 247 – 251. Links
41. E. G. Tuddenham, R. Schwaab, J. Seehafer, D. S. Millar, J. Gitschier, M. Higuchi, S. Bidichandani, J. M. Connor, L. W. Hoyer, A. Yoshioka, et al.: "Haemophilia A: Database of Nucleotide Substitutions, Deletions, Insertions and Rearangements of the FactorVII Gene", *Nucl. Acids Res.* **22** (1994) 3511 – 3533.
42. S. E. Antonarakis *et al.*: "Factor VIII gene inversions in severe hemophilia A: results of an international consortium study", *Blood* **86** (1995) 2206 – 2212. Links
43. K. Peerlinck, J. Amout, M. Di Giambattista, J. G. Gilles, R. Laub, M. Jacquemin, J. M. Saint-Remy, J. Vermylen: "FactorVIII Inhibitors in Previously Treated Haemophilia A Patients with a Double Virus-inactivated Plasma Derived Factor VIII Concentrate," *Thromb. Haemost.* **77** (1997) 80 – 86. Links
44. I. Scharrer, "The Need for Highly Purified Products to Treat Hemophilia B," *Acta Haematol.* **94** (1995) Suppl 1: 2 – 7. Links
45. I. MacGregor, L. McLaughlin, O. Drummond, C. Prowse, J. Ferguson: "In vivo models of thrombogenic potential: usefulness and limitations", *Acta Haematol.* **94** (1995) Suppl 1: 18 – 24. Links
46. J. Oldenburg *et al.*, *Infusionsther. Infusionsmed.* **23** (1996) 271 – 280.
47. M. Köhler, P. Hellstern, G. Müller-Berghaus, E. Lechler, *Dt. Ärztebl.* **92** (1995) 1863 – 1865.
48. J. Lusher: "Thrombogenicity associated with factor IX complex concentrates", *Semin. Haematol.* **28** (Suppl. 6), (1991) 3 – 5.
49. A. R. Gilles, M. E. Nesheim, H. Hoogendoom, P. B. Tracy, K. G. Mann: "The coagulant-active phospholipid content is a major determinant of in vivo thrombogenicity of prothrombin complex (Factor IX) concentrates in rabbits", *Blood* **59** (1982) 401 – 407. Links
50. M. Kusch, R. Seitz, H. König: "High sensitivity detection of activated factor IX: application to the analysis of different therapeutical factor IX concentrates and prothrombin complexes", *Thromb. Haemost.* **79** (1998) 778–83. Links
51. C. Hanker Dusel, C. Grundmann, S. Eich, R. Seitz, H. König, "Identification of prothrombin as a major thrombogenic agent in prothrombin complex concentrates", *Blood Coagul Fibrinolysis* **15** (5) (2004) 405–411. Links
52. C. Grundmann, R. Piesker, M. Kusch, K. M. Hanschmann, S. Eich, R. Seitz, H. König: "Prothrombin overload causes thromboembolic complications in prothrombin complex concentrates: *In vitro* and *in vivo* evidence", *Thromb. Haemost.* **94** (2005) 1338–1339. Links
53. European Medicines Agency, "Note for guidance on plasma-derived medicinal products", CPMP/BWP/269/95 rev.3 (revision of [29]).
54. European Medicines Agency, "Guideline on assessing the risk for virus transmission" CPMP/BWP/5180/03
55. European Medicines Agency, "CPMP position statement on West Nile Virus and plasma-derived medicinal products",

EMEA/CPMP/BWP/3752/03.

56. European Medicines Agency, “CHMP position statement on Creutzfeld-Jakob disease and plasma-derived and urine-derived medicinal products”, EMEA/CPMP/BWP/2879/02/rev 2.
57. The Trial to Reduce Alloimmunization to Platelets Study Group: “Leukocyte Reduction and Ultraviolet B Irradiation of Platelets to Prevent Alloimmunization and Refractoriness to Platelet Transfusions,” *N. Engl. J. Med.* **337** (1997) 1861 – 1869. Links
58. Commission Directive 2004/33/EC of 22 March 2004 implementing Directive 2002/98/EC of the European Parliament and of the Council as regards certain technical requirements for blood and blood components.
59. J. Dzieczkowski, “Blood and Blood Product Transfusion, Indications, and Proper Use” *Infertility and Reproductive Medicine Clinics of North America* **8** (1997) 109 – 124.
60. W. Schramm, F. von Auer, F. Delaney, R. Seitz: *Blood Safety in the European Community: An initiative for optimal use*, European Commission. ISBN 3-00-005705-6 European Commission, Luxembourg, 1999.
61. N Schmitz, A. Gratwohl, J. M. Goldman: “Allogeneic and Autologous Transplantation for Haematological Diseases, Solid Tumors and Immune Disorders, Current Practice in Europe in 1996 and Proposals for an Operational Classification,” *Bone Marrow Transplant* **17** (1996) 471 – 477. Links
62. C. R. Seed, P. Kiely, A. J. Keller: “Residual risk of transfusion transmitted human immunodeficiency virus, hepatitis B virus, hepatitis C virus and human T lymphotropic virus”, *Int Med J* **35** (2005) 592–598;
63. J. H. Hoofnagle, D. Lau, “Chronic Viral Hepatitis – Benefits of Current Therapies,” *N. Engl. J. Med.* **334** (1996) 1470 – 1471. Links
64. E.M. Pogliani, P. Faccini, M. Fumagalli: “Haematopoietic growth factors: current status”, *Internal Medicine* **9** (2001) 1–9.
65. M. Wadhwa, P. Dilger, J. Tubbs, A. Mire-Sluis, T. Barrowcliffe, R. Thorpe: “Identification of Transforming Growth Factor-β as a Contaminant in Factor VIII Concentrates: A Possible Link With Immunosuppressive Effects in Hemophiliacs,” *Blood* **84** (1994) 2021 – 2030. Links
66. I. M. Nilsson, “Factor VIII inhibitor treatment. Immune Tolerance”, *Semin. Hematol.* **31** (1994) 44 – 48. Links
67. I. Warrier, B. M. Ewenstein, M. A. Koerper, A. Shapiro, N. Key, D. DiMichele, R. T. Miller, J. Pasi, G. E. Rivard, S. S. Sommer, J. Katz, F. Bergmann, R. Ljung, P. Petrini, J. M. Lusher: “FactorIX inhibitors and Anaphylaxis in Hemophilia B,” *J. Pediatr Hematol Oncol* **19** (1997) 23 – 27. Links
68. F. R. Rosendaal *et al.* and The Dutch Hemophilia Study Group, “A Sudden Increase in FactorVIII Inhibitor Development in Multitransfused Hemophilia A Patients in The Netherlands,” *Blood* **81** (1993) 2180 – 2186. Links
69. J. McDonagh, R. Seitz, R. Egbring (eds.): “FactorXIII,” *Second International Conference Marburg 1991*, Schattauer Verlag, Stuttgart – New York 1993.
70. K. Mayer *et al.*, “Congenital Herozygote Hypoplasminogenemia as a Risk Factor for Thrombosis,” *Fibrinolysis* **2** (1988) 37 – 38.
71. A. M. Mingers, N. Heimburger, E. Lutz, O. Schmidt, “Homocygotic and Heterocygotic Type I Plasminogen Deficiency in Two Families,” *Ann. Hematol.* **70** (1995) A75.
72. B.L. Warren *et al.*: “Caring for the critically ill patient. High-dose antithrombin III in severe sepsis: a randomized controlled trial” *JAMA* **286** (2001) 1869–1878. Links
73. G.R. Bernard *et al.*: “Efficacy and safety of recombinant human activated protein C for severe sepsis”, *N Engl J Med.* **344** (2001) 699–709. Links
74. C. Law, M. Marcaccio, P. Tam, N. Heddle, J. G. Kelton, “High-Dose Intravenous Immune Globulin and the Response to Splenectomy in Patients with Idiopathic Thrombocytopenic Purpura,” *N. Engl. J. Med.* **336** (1997) 1494 – 1498. Links
75. European Medicines Agency, “Guideline on the Core SPC for Human Albumin Solution” (CPMP/PhVWP/BPWG/2231/99 rev.2)

[Top of Page]

Abbreviations used in this article

AAE: acquired angioedema

ACD: acid citrate-dextrose

a-PTC: activated prothrombin complex

BFU-E: burst-forming unit-erythrocytes

BSE: bovine spongiforme encephalitis

CD: cluster of differentiation

CJD: Creutzfeld Jacob disease

CFU: colony forming unit

CPMP: Committee of Proprietary Medicinal Products

CSF: colony stimulating factors

Eo: eosinophil

EMEA: European Medicines Evaluation Agency

FDP: fibrin degradation products

FEIBA: factor eight inhibitor bypassing activity

GM-CSF: granulocyte-macrophage colony stimulating factor

GMP: good manufacturing practice

GP: glycoprotein

HAE: hereditary angioedema

HIV: human immune deficiency virus

HLA: human leukocyte antigen

HBS: heparin binding site

HMWK: high molecular weight kininogen

Ig: immunoglobulin

IL: interleukin

ITP: idiopathic thrombocytopenic purpura

LPS: lipopolysaccharide

MCH: mean corpuscular hemoglobin

MCHC: mean corpuscular hemoglobin concentration

MCV: mean corpuscular volume

MEG: megakarocyte

MHC: major histocompatibility complex

PA: plasminogen activator

PAI: plasminogen activator inhibitor

PDGF: platelet-derived growth factor

PTC: prothrombin complex

α_1-PI: α_1 -proteinase inhibitor

PK: prekallikrein

RES: reticulo-endothelial system

RS: reactive site

SCD: sterile connecting device

SCF: stem cell factor

TAT: thrombin-antithrombin

Tpo: thromopoetin

TNF: tumor necrosis factor

vWF: von Willebrand factor

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Antihypotensives

Hans Dieter Lehmann, Knoll AG, Ludwigshafen, Federal Republic of Germany

Marco Thyes, Knoll AG, Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **Sympathomimetics** 1
3. **Substances Having Other Modes of Action** 4
3.1. **Ergot Alkaloids** 4
3.2. **Mineralocorticoids** 5
4. **References** 5

1. Introduction

Blood pressure increasing agents are used to counteract acute drop of blood pressure, such as is observed in surgery, severe infectious diseases, and shock. A further field of use is persistent hypotension caused by circulatory dysregulation when standing (postural hypotension); this results in cerebral ischemia, which in turn produces clinical symptoms. Orthostatic dysregulation can also be induced by drugs, in particular by blood pressure lowering drugs (→ Antihypertensives), and by drugs acting on the central nervous system (psychotropic drugs). Hypotensive circulatory disorders may be volume-related or regulation-related. In volume-related hypotension there is an absolute deficit in blood volume, which must be compensated for by infusing blood or plasma expanders (→ Blood). Regulation-related hypotension originates either in the central nervous system or in the periphery, involving capacitance vessels (veins and venules) or resistance vessels (small arteries and arterioles) or both. In regulation-induced hypotension there is an absolute or relative deficiency in the release of norepinephrine from nerve terminals of the sympathetic nervous system in the vascular section. The release of norepinephrine is crucial to maintain vascular tone. Drug treatment, therefore, serves to supply vasoconstrictor substances, either acting like norepinephrine, i.e., sympathomimetics, or having other modes of action, i.e., ergot alkaloids, peptides, and mineralocorticoids. The mineralocorticoids also increase total blood volume. In acute hypotensive crises the drugs must be administered intravenously. For long-term hypotensive dysregulations orally effective substances are needed. The most important blood pressure increasing agents are described in this article [1].

2. Sympathomimetics

The physiological neurotransmitters of the sympathetic (adrenergic) nervous system, norepinephrine and epinephrine, and the other synthetic sympathomimetics stimulate the adrenergic α- and/or β-receptors. The stimulation of α-receptors causes the smooth muscle cells of resistance and capacitance vessels to contract; the stimulation of β_1-receptors leads to increased cardiac output, and the stimulation of β_2-receptors causes vascular smooth muscle cells to relax. Norepinephrine primarily activates α-receptors, whereas epinephrine stimulates both α- and β-receptors. Both neurotransmitters are only suitable for intravenous injection, especially for continuous i.v. infusion, because of their low bioavailability (≈ 3 %) and short half-life (≈ 2 min) [2].

Norepinephrine [*51-41-2*], (*R*)-(−)-4-(2-amino-1-hydroxyethyl)-1,2-benzenediol, (−)-norepinephrine, noradrenaline, (−)-noradrenaline, its hydrochloride [*329-56-6*], and its (*R,R*)-hydrogen tartrate [*51-40-1*] are described in more detail under → Hormones, Chap. 2.2.

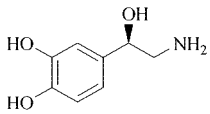

Trade names(hydrochloride): Arterenol (Hoechst Marion Roussel).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_229

R,R-Hydrogen tartrate: Levophed (Abbott, Breon), Levophed Bitartrate (Sanofi Winthrop), Levophed special (Abbott), Noradrenalin 1 : 1000 JENA-PHARM (Jenapharm), Noradrenalina tartrato (Monico, Salf, Sifra).

Epinephrine [*51-43-4*], (*R*)-(−)-4-[1-hydroxy-2-(methylamino)ethyl]-1,2-benzenediol, (−)-epinephrine, adrenaline, (−)-adrenaline, its hydrochloride [*55-31-2*] and its hydrogen tartrate [*51-42-3*] are described in more detail under → Hormones, Chap. 2.2.

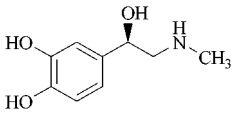

Trade names: Adrenalina ISM 1 : 1000 (Nuovo ISM), Anahelp (Stallergènes), Anakit (Dome-Hollister-Stier), Dyspne-Inhal (Augot), EPIFRIN (Allergan), EpiPen (Dey), Primatene (Whitehall-Robins), Sus-Phrine (Forest).

Hydrochloride: Adrenalin 1 : 10000 (B. BRAUN), Ana-Kit (Bayer Allergy), Anaphylaxie-Besteck (Bencard), Fastjekt (Allergopharma), Infectokrupp Inhal (Infectopharm), Suprarenin (Hoechst Marion Roussel).

Hydrogen tartrate: Adrenalin 1 : 1000 JENA-PHARM (Jenapharm).

Dopamine [*51-61-6*], 4-(2-aminoethyl)-1,2-benzenediol, 3-hydroxytyramine, $C_8H_{11}NO_2$, M_r 153.18, is very sensitive to oxygen; its hydrochloride [*62-31-7*] melts at 247 °C.

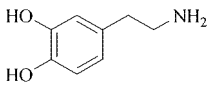

Dopamine is the direct endogenous precursor of norepinephrine. Its cardiovascular effects are different from those of norepinephrine and epinephrine. It dilates coronary, cerebral, renal, and mesenteric arteries by interacting with dopaminergic receptors. It also activates adrenergic α- and β-receptors, which in turn causes capacitance vessels to constrict and cardiac output to increase. With this combined effect dopamine is well suited for shock treatment: it increases blood pressure, and at the same time renal and mesenteric blood flow is augmented. In terms of pharmacokinetics, dopamine is similar to norepinephrine and epinephrine and must be administered by continuous intravenous infusion [3].

Dopamine is synthesized from veratrole by chloromethylation to form veratryl chloride, followed by cyanation, catalytic hydrogenation, and demethylation.

For further methods of synthesis, see [9].

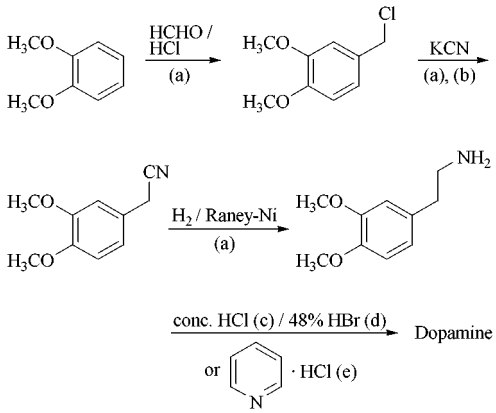

(see a: [4]; b: [5]; c: [6]; d: [7]; e: [8])

Trade names(hydrochloride): Dopamin Fresenius (Fresenius-Klinik), Dopamin-ratiopharm (ratiopharm), Dopamin Solvay (Solvay Arzneimittel), Dopamine Pierre Fabre (Pierre Fabre Cardio Vasculaire), Dopamine Lucien (Therabel Lucien Pharma), Dopamine Nativelle (Procter & Gamble), Revivan (Astra).

Norfenefrine [*536-21-0*], α-(aminomethyl)-3-hydroxybenzenemethanol, $C_8H_{11}NO_2$, M_r 153.18, is closely related to norepinephrine; its hydrochloride [*4779-94-6*] melts at 159 – 160 °C.

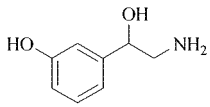

The bioavailability of norfenefrine is only slightly greater (≈ 5 %) than that of norepinephrine because the major portion is metabolized in the intestinal wall and the liver (extensive first-pass effect). When administered orally the compound is only effective at high dose levels [10].

Norfenefrine can be synthesized from *m*-hydroxyacetophenone as shown in the following scheme [11]:

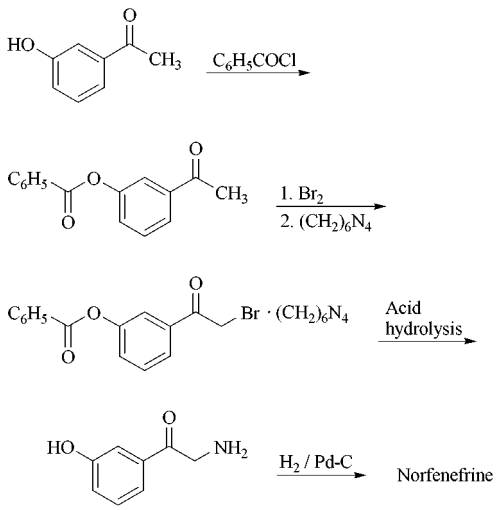

Norfenefrine is also obtained from *m*-hydroxyacetophenone by nitrosation to form *m*-hydroxy-(hydroxyimino)acetophenone and subsequent catalytic hydrogenation [12]:

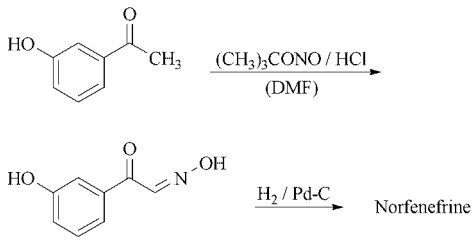

For a further method of synthesis, see [13]. Trade names(hydrochloride): Coritat (Green Cross), Energona (Maurer), Norfenefrin "Ziethen" (Ziethen), Norfenefrin retard forte-ratiopharm (ratiopharm), Novadral (Goedecke), Zondel (Grelan).

Midodrine [*42794-76-3*], 2-amino-*N*-[2-(2,5-dimethoxyphenyl)-2-hydroxyethyl]acetamide, $C_{12}H_{18}N_2O_4$, M_r 254.28, and its hydrochloride [*3092-17-9*], *mp* 202 – 204 °C, act on resistance and capacitance vessels by stimulating α-receptors. Midodrine is a prodrug from which the active major metabolite ST 1059, α-(aminomethyl)-2,5-dimethoxybenzenemethanol, is released by enzymatic cleavage of glycine. According to animal experiments the enteral efficacy of midodrine is better than that of the sympathomimetics already described. However, no data on the bioavailability and half-life of this relatively new compound have been published [14]. Synthesis [15]:

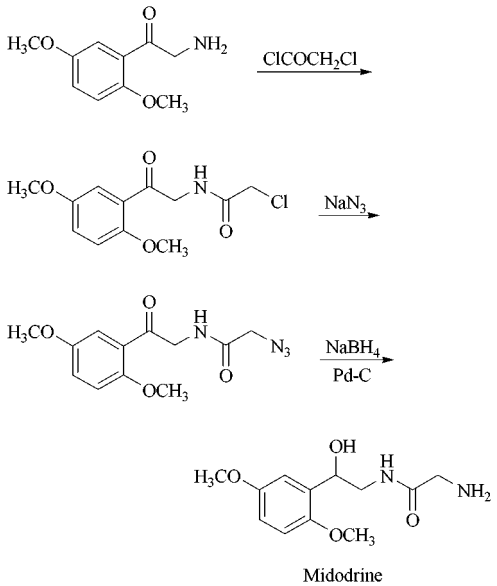

For further routes to midodrine, see [16]. Trade names(hydrochloride): Gutron (Guidotti, Nycomed), Metligine (Taisho), ProAmatine (Roberts).

Etilefrine [*709-55-7*], α-[(ethylamino)methyl]-3-hydroxybenzenemethanol, $C_{10}H_{15}NO_2$, M_r 181.23, *mp* 147 – 148 °C, and its hydrochloride [*943-17-9*], *mp* 121 °C, act both on α-receptors and on β-receptors. This not only increases the tone of resistance and capacitance vessels but also improves the cardiac output. The bioavailability of the compound is 55 % when administered in aqueous solution, 35 % when administered as an instant-release tablet, and 17 % as a slow-release tablet. Bioavailability increases to 61 % when 2 mg of dihydroergotamine (see page 4) is added to a slow-release tablet [17].

Etilefrine can be synthesized using *m*-acetoxy-ω-bromoacetophenone as the starting material:

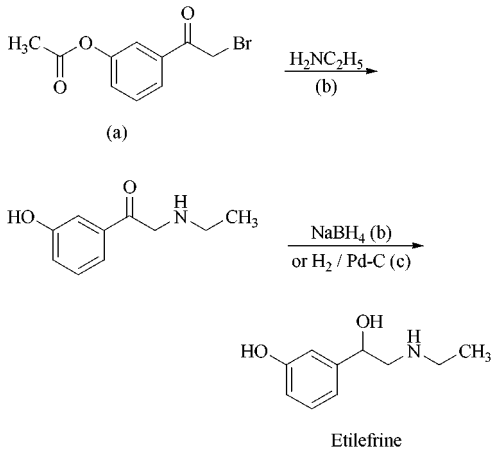

(see a: [18]; b: [19]; c: [20], [21])

For an alternate route to etilefrine, see [20].

Trade names(hydrochloride): Adrenam (NAM Neukönigsförder), Bioflutin-N (Südmedica), Cardanat (Temmler Pharma), Cardialgine (MIP Pharma), Circupon RR (gegepharm), Circuvit E (Pharma Wernigerode), Confidol (Medopharm), Effortil (Boehringer Ingelheim), etil 5 von ct (ct-Arzneimittel), ETI-PUREN (Isis-Puren), Kreislauf Katovit (Boehringer Ingelheim), Thomasin (Apogepha).

Hydrochloride in combination with dihydroergotamine methanesulfonate see page 4.

Ameziniummetilsulfate [*30578-37-1*], 4-amino-6-methoxy-1-phenylpyridazinium methyl sulfate, amezinium, $C_{12}H_{15}N_3O_5S$, M_r 313.33, *mp* 176 °C (decomp.), also acts via adrenergic α- and β-receptors. Its chemical structure differs from that of other sympathomimetics. At the adrenergic neurons the drug causes specific, reversible inhibition of monoamine oxidase, inhibition of norepinephrine re-uptake, and a slight increase in norepinephrine release. Therefore, the agent is only effective in vessels innervated by adrenergic nerves, thereby enhancing sympathetic activity. Unlike other indirectly acting sympathomimetic drugs, ameziniummetilsulfate does not deplete the norepinephrine stores in sympathetic nerve terminals. Its bioavailability is ca. 60 %, the terminal half-life is 9 – 17 h [22].

The substance is prepared by first treating 4,5-dichloro-2-phenyl-3(2*H*)-pyridazinone [23] with ammonia to give mainly 5-amino-4-chloro-2-phenyl-3(2*H*)-pyridazinone and some 4-amino-5-chloro-2-phenyl-3(2*H*)-pyridazinone [23].

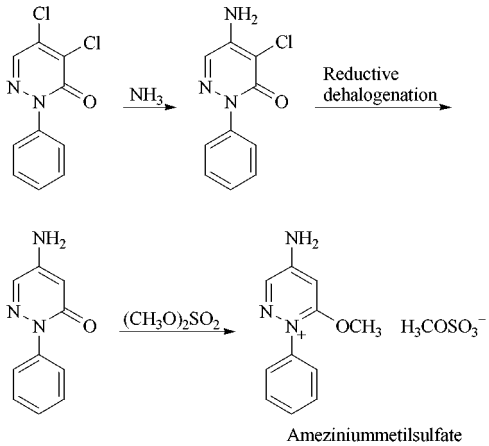

After removal of the 4-amino-5-chloro derivative by extraction with chloroform [24], [25], the 5-amino-4-chloro compound is dehalogenated with hydrogen over Raney nickel [23], [25] or with sodium borohydride in the presence of palladium on charcoal [26]. Subsequent reaction with dimethyl sulfate leads to ameziniummetilsulfate [25], [27].

Trade names: Regulton (Knoll Deutschland), Risumic (Dainippon), Supratonin (Grünenthal).

3. Substances Having Other Modes of Action

3.1. Ergot Alkaloids

Dihydroergotamine [*511-12-6*], (5′α,10α)-9,10-dihydro-12′-hydroxy-2′-methyl-5′-(phenylmethyl)ergotaman-3′,6′,18-trione, 9,10-dihydroergotamine, is treated in more detail under → Alkaloids, Chap. 11.3.2.; the methanesulfonate [*6190-39-2*] melts at 230 – 235 °C (decomp.).

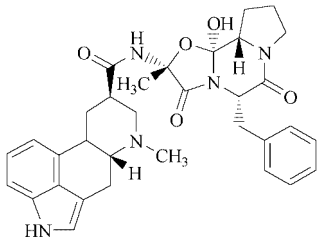

Dihydroergotamine acts on adrenergic α-receptors and serotonin receptors of capacitance vessels and increases their tone. In this way blood supply to the heart and stroke volume are improved. As a consequence, blood pressure is increased. To a great extent dihydroergotamine acts selectively on capacitance vessels, whereas an increase in the tone of resistance vessels has only been demonstrated in denervated vessel preparations. The bioavailability of the drug is only about 1 %. However, its effect on venous tone has been clearly demonstrated in animals and humans after oral administration and can be explained by an accumulation of the agent in the smooth muscle cells of the vessels. Therefore, its effect is largely independent of plasma levels [28].

Dihydroergotamine can be obtained by hydrogenating ergotamine (→ Alkaloids) in the presence of a palladium catalyst [29].

Trade names(methanesulfonate): Agit depot sanol (Sanol), Angionorm (Farmasan), clavigrenin (Hormosan), DET MS (FUISZ PHARMA), D.H.E. 45 (Novartis), DHE-PUREN (Isis-Puren), DHE-ratiopharm (ratiopharm), Dihydergot (Novartis Pharma), Dihytamin (Arzneimittelwerk Dresden, Temmler Pharma), Diidergot (Sandoz), Ergomimet (Klinge), Ergont (Desitin), ergotam von ct (ct-Arzneimittel), Ikaran (Pierre Fabre, Formenti), Migranal (Novartis), Seglor (Sanofi Winthrop, Synthelabo), Verladyn (Verla).

Methanesulfonate in combination with etilefrine hydrochloride: Effortil plus (Boehringer Ingelheim), Ergolefrin (gegepharm), Agit plus sanol (Sanol), Dihydergot plus (Novartis Pharma), Ergomimet plus (Klinge).

3.2. Mineralocorticoids

Therapy of postural hypotension by mineralocorticoids (renin – angiotensin – aldosterone mechanism) is based on sodium and water retention in the body. This increases the blood volume and makes the vessels more sensitive to vasoconstrictive agents, resulting in augmented tone of resistance and capacitance vessels. Fludrocortisone is normally used; it has a bioavailability of 100 %. On prolonged administration, fludrocortisone inhibits the production of renin, angiotensin, and aldosterone and causes atrophy of aldosterone-producing cells in the adrenal gland. Therefore, fludrocortisone is only used in the treatment of postural circulatory dysregulation where other drugs have failed. In any case treatment must be short [30].

Fludrocortisone [*127-31-1*], (11β)-9-fluoro-11,17,21-trihydroxypregn-4-ene-3,20-dione, 9α-fluorohydrocortisone, 9α-fluorocortisol, $C_{21}H_{29}FO_5$, M_r 380.46, *mp* 260 – 262 °C (decomp.), $[\alpha]_D^{23}$ + 139° (c = 0.55, 95 % ethanol); 21-acetate [*514-36-3*], $C_{23}H_{31}FO_6$, M_r 422.48, crystalline substance, several polymorphic forms, *mp* 233 – 234 °C (occasionally melting at 205 – 208 °C, resolidification and subsequent melting at 226 – 228 °C, $[\alpha]_D^{23}$ + 123° (c = 0.64, chloroform); for the synthesis of the 21-acetate, see [31–33] and for its conversion to fludrocortisone, see [31], [34]; see [35] for a broad description of the 21-acetate and [36] for a process giving dense granular prisms of the 21-acetate.

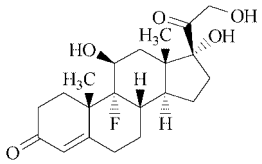

Trade names: Astonin H (Merck).

21-Acetate: Florinef (Bristol-Myers Squibb), Florinef Acetate (Apothecon), Fludrocortison (Bristol-Myers Squibb).

4. References

1. O. Thulesius, *Cardiology* **61, Suppl. 1** (1976) 180. M. G. Ziegler, C. R. Lake, I. J. Kopin, *New Engl. J. Med.* **296** (1977) 293. R. L. Klein, J. Mc C. Baggett, Å. Thureson-Klein, H. G. Langford, *J. Auton. Nerv. Syst.* **2** (1980)

205. H. Rieckert: *Hypotonie,* Springer, Berlin-Heidelberg-New York 1979.
2. I. J. Kopin in H. Blaschko, E. Muscholl (eds.): *Handbook of Experimental Pharmacology,* **vol. XXXIII,** Catecholamines, Springer, Berlin-Heidelberg-New York 1972, p. 270. M. Esler, G. Jackman, A. Bobik, P. Leonard, D. Kelleher, H. Skews, G. Jennings, P. Korner, *Hypertension (Dallas)* **3** (1981) 149.
3. L. Goldberg, J. D. Kohli, A. N. Kotake, P. H. Volkman, *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **37** (1978) 2396. J. R. Bianchine, G. M. Shaw, J. E. Greenwald, S. M. Demdalides, *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **37** (1978) 2434. J. W. Kebabian, D. B. Calne, *Nature (London)* **277** (1979) 93. R. J. Marino, A. Romagnoli, A. S. Keats, *Anesthesiology* **43** (1975) 570.
4. A. E. Bide, P. A. Wilkinson, *J. Soc. Chem. Ind., London* **64** (1945) 84.
5. Knoll, DE 896344, 1950 (F. Dengel). Knoll, DE 903931, 1951 (F. Dengel).
6. K. W. Rosenmund, C. Mannich, W. Jacobsohn, DE 247906, 1909. G. Hahn, K. Stiehl, *Chem. Ber.* **69** (1936) 2627. G. R. Clemo, F. K. Duxbury, G. A. Swan, *J. Chem. Soc.* **1952,** 3464.
7. C. Schöpf, H. Bayerle, *Liebigs Ann. Chem.* **513** (1934) 190.
8. Pierre Fabre, FR 2332748, 1975 (G. Mouzin, H. Cousse).
9. E. Waser, H. Sommer, *Helv. Chim. Acta* **6** (1923) 54. J. S. Buck, *J. Am. Chem. Soc.* **55** (1933) 3388.
10. J. H. Hengstmann, W. Konen, C. Konen, M. Eichel-baum, H. J. Dengler, *Eur. J. Clin. Pharmacol.* **8** (1975) 33.
11. A. D'Amico, L. Bertolini, C. Monreale, *Chim. Ind. (Milan)* **38** (1956) 93. R. Sachs, FR 851296, 1983. R. Sachs, FR 866569, 1939. Ciba, US 2312916, 1940 (H. Legerlotz).
12. Gödecke, DE-AS 2130710, 1971 (G. Satzinger, W. Herrmann).
13. H. Bretschneider, H. Hörmann, *Monatsh. Chem.* **84** (1953) 1021.
14. H. Pittner, H. Stormann, R. Enzenhofer, *Arzneim. Forsch./Drug Res.* **26** (1976) 2145. N. Kolassa, W. G. Schützenberger, H. Wiener, P. Krivanek, *Arch. Int. Pharmacodyn. Ther.* **238** (1979) 96.
15. Lentia, DE-AS 2523735, 1975 (G. Zölß, K.-A. Wismayr).
16. Österreichische Stickstoffwerke, AT 241435, 1963 (K. Wismayr, R. Kilches, O. Schmid, G. Zölß). Lentia, DE 2506110, 1975 (G. Zölß). Chemie Linz, AT 336582, 1975 (G. Zölß).
17. J. Tarnow, J. B. Brückner, H. J. Eberlein, D. Patschke, P. Schmicke, A. Weymar, *Acta Anaesthesiol Scand.* **17** (1973) 184. S. Mellander, *Angiologica* **3** (1966) 77. J. H. Hengstmann, U. Weyand, H. J. Dengler, *Eur. J. Clin. Pharmacol.* **9** (1975) 179. J. H. Hengstmann, R. Hengstmann, S. Schwonzen, H. J. Dengler, *Eur. J. Clin. Pharmacol.* **22** (1982) 463.
18. R. Sachs, FR 851296, 1938.
19. T. Goto, *J. Pharm. Soc. Japan* **74** (1954) 318; *Chem. Abstr.* **49** (1955) 3960 c.
20. Laboratorios Fher, ES 273595, 1962; *Chem. Abstr.* **60** (1964) 1649 c.
21. Boehringer Ingelheim, DE 566578, 1927 (H. Legerlotz).
22. H. D. Lehmann, H. Giertz, R. Kretzschmar, D. Lenke, G. v. Philipsborn, M. Raschack, J. Schuster, *Arzneim. Forsch./Drug Res.* **31** (1981) 1544. D. Lenke, J. Gries, R. Kretzschmar, *Arzneim. Forsch./Drug Res.* **31** (1981) 1558. M. Traut, E. Brode, H. D. Hoffmann, *Arzneim. Forsch./Drug Res.* **31** (1981) 1566. M. Traut, E. Brode, *Arzneim. Forsch./Drug Res.* **31** (1981) 1605. K. Starke, *Trends Pharmacol. Sci.* **4** (1983) 269. U. S. v. Euler, F. Lishajko, *Acta Physiol. Scand.* **73** (1968) 78.
23. K. Dury, *Angew. Chem.* **77** (1965) 282; *Angew. Chem. Int. Ed. Engl.* **4** (1965) 292.
24. BASF, DE 2100685, 1971 (F. Reicheneder, R. Kropp).
25. F. Reicheneder, T. F. Burger, H. König, R. Kropp, H. Lietz, M. Thyes, W.-W. Wiersdorff, *Arzneim. Forsch./Drug Res.* **31** (1981) 1529.
26. M. Beljean, M. Pays, *Bull. Soc. Chim. Fr.* (1973) 3324.
27. BASF, DE 1912941, 1969 (F. Reicheneder, R. Kropp). BASF, DE 3114496, 1981 (W.-W. Wiersdorff, R. Kropp).
28. S. Mellander, J. Nordenfelt, *Clin. Sci.* **39** (1970) 183. J. Nordenfelt, S. Mellander, *Acta Med. Scand.* **191** (1972) 115. E. Müller-Schweinitzer, *Eur. J. Pharmacol.* **27** (1974) 231. P. J. Little, G. L. Jennings, H. Skews, A. Bobitz, *Br. J. Clin. Pharmacol.* **13** (1982) 785. E. Müller-Schweinitzer, *Int. J. Clin. Pharmacol.* **22** (1984) 677.
29. Sandoz, DE 883153, 1941 (A. Stoll, A. Hofmann). A. Stoll, A. Hofmann, *Helv. Chim. Acta* **26** (1943) 2070.
30. H. Lettenbaur, H. Nowak, *Arzneim. Forsch./Drug Res.* **21** (1971) 1105. W. Vogt, I. Fischer, S. Ebenroth, S. Appel, M. Knedel,

P. W. Lücker, H. Rennekamp, *Arzneim. Forsch./ Drug Res.* **21** (1971) 1133. E. F. Gersmeyer, W. Mauer, W. W. Huep in H. J. Dengler (ed.): *Das Orthostasesyndrom,* Schattauer, Stuttgart-New York 1974, p. 153.
31. J. Fried, E. F. Sabo, *J. Am. Chem. Soc.* **76** (1954) 1455. Olin Mathieson, GB 792224, 1954.
32. R. F. Hirschmann, R. Miller, J. Wood, R. E. Jones, *J. Am. Chem. Soc.* **78** (1956) 4956.
33. Merck & Co., DE-AS 1 035 133, 1956 (R. F. Hirschmann, R. Miller). Merck & Co., US 2 894 007, 1959 (R. P. Graber, C. S. Snoddy, Jr.).
34. Schering, DE-AS 1 028 572, 1957 (E. Kaspar, W. Hiersemann, M. Schenck).
35. K. Florey, *Anal. Profiles Drug Subs.* **3** (1974) 281.
36. Merck & Co., US 2 957 013, 1960 (R. P. Graber, C. S. Snoddy, Jr.).

Antihypertensives

Alexander Scriabine, Miles Laboratories, Inc., Institute for Preclinical Pharmacology, New Haven, Connecticut 06511, United States

David G. Taylor, Miles Laboratories, Inc., Institute for Preclinical Pharmacology, New Haven, Connecticut 06511, United States

1. Introduction . . . 1
2. Diuretics . . . 2
2.1. Sulfonamide Diuretics . . . 2
2.1.1. Hydrochlorothiazide . . . 3
2.1.2. Chlorthalidone . . . 3
2.1.3. Furosemide . . . 3
2.1.4. Mefruside . . . 4
2.1.5. Indapamide . . . 4
2.2. K^+-Retaining Diuretics . . . 4
2.2.1. Amiloride . . . 4
2.2.2. Triamterene . . . 5
2.3. Other Diuretics . . . 5
2.3.1. Ethacrynic Acid . . . 5
2.3.2. Indacrinone . . . 6
2.3.3. Muzolimine . . . 6
3. Inhibitors of Angiotensin Converting Enzyme . . . 6
3.1. Captopril . . . 7
3.2. Enalapril . . . 7
4. Inhibitors of the Sympathetic Nervous System . . . 8
4.1. Drugs with Primarily Central Actions . . . 8
4.1.1. Methyldopa . . . 9
4.1.2. Clonidine . . . 9
4.1.3. Guanfacine . . . 9
4.1.4. Guanabenz . . . 10
4.2. Drugs with Primarily Peripheral Action . . . 10
4.2.1. Amine Depleting and Adrenergic Neuron Blocking Agents . . . 11
4.2.1.1. Reserpine . . . 11
4.2.1.2. Guanethidine . . . 12
4.2.2. α-Adrenoceptor Antagonists . . . 12
4.2.2.1. Prazosin . . . 12
4.2.2.2. Indoramin . . . 13
4.2.2.3. Urapidil . . . 13
4.2.3. β-Adrenoceptor Antagonists . . . 14
4.2.3.1. Propranolol . . . 14
4.2.3.2. Metoprolol . . . 15
4.2.3.3. Nadolol . . . 15
4.2.3.4. Pindolol . . . 15
4.2.3.5. Timolol . . . 16
4.2.3.6. Acebutolol . . . 16
4.2.3.7. Labetalol . . . 17
5. Smooth Muscle Relaxants . . . 17
5.1. Ca^{2+} Channel Antagonists . . . 17
5.1.1. Nifedipine . . . 17
5.1.2. Nitrendipine . . . 18
5.1.3. Verapamil . . . 19
5.1.4. Diltiazem . . . 19
5.2. Vasodilators with Other Mechanisms of Action . . . 20
5.2.1. Hydralazine . . . 20
5.2.2. Diazoxide . . . 21
5.2.3. Minoxidil . . . 21
5.2.4. Sodium Nitroprusside . . . 21
6. Perspectives in the Development of Antihypertensive Drugs . . . 22
7. Summary and Conclusions . . . 22
8. References . . . 23

1. Introduction

In spite of the fact that many drugs are now available for the treatment of hypertension, new drugs are being introduced every year. The existence of a lucrative and still expanding market as well as the medical need for effective antihypertensive drugs with minimal side effects stimulate continuous research activity in this field.

The available antihypertensive drugs can be classified on the basis of their mechanism of action as follows: (1) diuretics, (2) inhibitors of angiotensin converting enzyme (ACE), (3) drugs which interfere with the sympathetic nervous system, and (4) vascular smooth muscle relaxants. This article describes the major representatives of each of the above classes of antihypertensive drugs.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_235

2. Diuretics

The chemistry, pharmacology, and clinical use of diuretics in hypertension and edema are discussed in [1], also → Diuretics. The rapid expansion of the diuretic field occurred after the introduction of acetazolamide [2] and chlorothiazide [3], [4]. The discovery of the effectiveness of chlorothiazide in the therapy of hypertension [5] greatly expanded the market for diuretics and led to the synthesis of many chemically related compounds. Since hypokalemia is one of the major side effects of chlorothiazide and related thiazides, attempts were made to develop diuretics with K^+-retaining or isokalemic properties. Aldosterone antagonists, such as spironolactone or other K^+ retaining diuretics, e.g., amiloride [6] or triamterene [7] were shown to produce diuresis while reducing excretion of potassium. Combinations of thiazides with potassium retaining diuretics (e.g., Moduretic or Dyazide) have diuretic effects and lower arterial pressure without any change in the excretion of potassium. These combinations are widely used in general practice.

The *mechanism* of antihypertensive action of diuretics is still a subject of controversy, although numerous studies suggested that the natriuretic action of diuretics is primarily responsible for their antihypertensive effect. The kidneys of hypertensive animals or humans can excrete salt in normal amounts but only at elevated arterial pressure levels and diuretics increase the excretion of salt independent of blood pressure level [8]. Also the increase in sodium in the vascular smooth muscle cell was correlated with the increase in the intracellular calcium, and therefore, with the vasoconstrictor tone leading to the elevation of arterial pressure [9], [10]. The natriuretic effect of diuretics tends to prevent excessive accumulation of sodium.

2.1. Sulfonamide Diuretics

The first sulfonamide shown to have diuretic activity in humans was sulfanilamide [*63-74-1*] [11]. Subsequent attempts to improve the activity and to reduce the side effects of sulfanilamide led to the development of acetazolamide [*59-66-5*] [12], which can be considered the prototype of sulfonamide diuretics. Its continuous use led to metabolic acidosis and to the development of tolerance to its diuretic effect. The search for a better diuretic led to the development of chlorothiazide [*58-94-6*] [13]. The diuretic activity of chlorothiazide, unlike that of acetazolamide, is not determined by its carbonic anhydrase inhibitory activity. At therapeutic doses, chlorothiazide increases the excretion of chloride rather than that of hydrogen carbonate. By repeated administration chlorothiazide produces effective diuresis without metabolic acidosis.

Extensive structure – activity relationship studies led subsequently to the development of chemically and pharmacologically related thiazide diuretics which differed from chlorothiazide either in the relative potency or in the duration of action. The most commonly used thiazide diuretic is hydrochlorothiazide [14], which is ca. 10 times more potent than chlorothiazide. The most potent thiazide among marketed drugs is cyclopenthiazide [15], while methylclothiazide [16] and polythiazide [17] have exceptionally long durations of action. Their longer duration of action is due to higher lipid solubility (and therefore greater reabsorption of the drug in the proximal tubules) and higher protein binding. Other clinically important sulfonamide diuretics include mefruside [18], [19], indapamide [20], and chlorthalidone [21].

A significant advance in diuretic therapy was the introduction of furosemide [12], which has much higher ceiling natriuretic and chloruretic effects than thiazide diuretics. The high ceiling effect of furosemide is determined by the site of its tubular action: furosemide inhibits chloride and sodium reabsorption in the ascending limb of Henle's loop ("loop" diuretic). Furosemide is widely used clinically, although in hypertension it is not more effective than thiazide diuretics. An attempt to find a "loop" diuretic with a long duration of action led to the development of muzolimine [21], which is not yet available on the United States market.

2.1.1. Hydrochlorothiazide

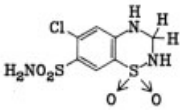

Hydrochlorothiazide [*58-93-5*], 6-chloro-3,4-dihydro-2*H*-1, 2, 4-benzothiadiazine-7-sulfonamide 1,1-dioxide, $C_7H_8ClN_3O_4S_2$, M_r 297.72, *mp* 273 – 275 °C, is insoluble in water, soluble in NaOH, ethanol, and acetone. Hydrochlorothiazide is widely used in the treatment of hypertension or edema.

Trade names: HydroDiuril, Esidrix.

Like other benzothiadiazine-type diuretics (or thiazides), hydrochlorothiazide increases the excretion of sodium and chloride independent of acid – base balance. Hydrochlorothiazide also increases the excretion of potassium. The resulting hypokalemia represents its major *side effect.* Other side effects include hyperuricemia, hyperglycemia, and hyperlipoproteinemia. The relative diuretic potency of hydrochlorothiazide is ca. 10 times higher than that of chlorothiazide [14]. The efficacy (ceiling effect) is similar to that of other thiazide diuretics and is considerably lower than that of furosemide. Daily clinical doses range from 25 to 200 mg and are given with potassium supplements or in combination with potassium-retaining diuretics to prevent hypokalemia.

2.1.2. Chlorthalidone

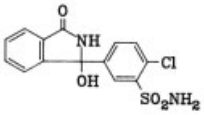

Chlorthalidone [*77-36-1*], 2-chloro-5-(1-hydroxy-3-oxo-1-isoindolinyl)benzenesulfonamide, $C_{14}H_{11}ClN_2O_4S$, M_r 338.78, *mp* 224 – 226 °C, is slightly soluble in water (12 mg/100 mL), more soluble in diluted NaOH or warm ethanol.

Trade name: Hygroton.

Chlorthalidone is pharmacologically similar to thiazide diuretics but is substantially longer acting than hydrochlorothiazide [21]. The duration of action is determined by its elimination half-life ($t_{1/2}$) of 49 h in humans. Because of longer duration of action, chlorthalidone can be given every 72 h. Chlorthalidone is widely used in the treatment of hypertension: it can be used alone or in combination with methyldopa or β-adrenoceptor antagonists. *Side effects* are also similar to those of thiazides and include hypokalemia, hyperuricemia, hyperglycemia, and hyperlipoproteinemia (elevation of low-density lipoproteins). Clinical doses range from 25 to 400 mg.

2.1.3. Furosemide

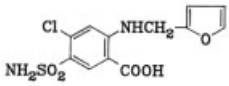

Furosemide [*54-31-9*], 5-(aminosulfonyl)-4-chloro-2-[(2-furanylmethyl)amino]benzoic acid, 4-chloro-*N*-furfuryl-5-sulfamoylanthranilic acid, $C_{12}H_{11}ClN_2O_5S$, M_r 330.77, *mp* 206 °C, is only slightly soluble in water, soluble in methanol, acetone, and dilute NaOH.

Trade name: Lasix.

The major site of action of furosemide is in the ascending limb of Henle's loop, proximal to the thiazide diuretics. Like thiazides, furosemide increases the excretion of sodium, chloride, and potassium, but its efficacy (ceiling effect) is much higher than that of thiazides and similar diuretics [12]. The duration of diuretic action of furosemide is short and is correlated with its elimination half-life (0.85 h). Independent of its diuretic action, furosemide, by i.v. administration, tends to increase venous capacitance and to reduce left ventricular end diastolic pressure. In spite of its greater diuretic efficacy, furosemide is not more effective than thiazide diuretics in the treatment of hypertension. *Side effects* of furosemide include those of thiazide diuretics: hypokalemia, hyperglycemia, hyperuricemia, and hyperlipoproteinemia. In addition, furosemide tends to produce metabolic alkalosis and ototoxicity, which is typical for "loop" diuretics. Daily clinical doses of furosemide range from 20 to 200 mg.

2.1.4. Mefruside

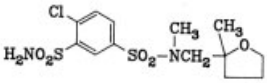

Mefruside [*7195-27-9*], 4-chloro-N^1-methyl-N^1-[(tetrahydro-2-methyl-2-furanyl)methyl]-1,3-benzenedisulfonamide, 4-chloro-N^1-methyl-N^1-(tetrahydro-2-methylfurfuryl)-*m*-benzenedisulfonamide, $C_{13}H_{19}ClN_2O_5S_2$, M_r 382.90, *mp* 149 – 150 °C (DL-crystals), 146 °C (D-form). The D-form is more active as a diuretic than the L-form.

Trade name: Baycaron.

The primary site of action of mefruside is the cortical diluting segment; duration of action is 20 to 24 h. Mefruside has carbonic anhydrase inhibitory activity but is less potent than acetazolamide [18], [19]. One of its metabolites, oxomefruside, also has diuretic activity. *Side effects* of mefruside are similar to those of thiazide diuretics, in that it produces hyperuricemia and hypokalemia. Daily clinical doses range from 25 to 200 mg. In hypertension, mefruside is effective at 25 – 50 mg/d or every other day.

2.1.5. Indapamide

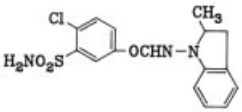

Indapamide [*26807-65-8*], 3-(aminosulfonyl)-4-chloro-*N*-(2,3-dihydro-2-methyl-1*H*-indol-1-yl)benzamide, 4-chloro-*N*-(2-methyl-1-indolinyl)-3-sulfamoylbenzamide, $C_{16}H_{16}ClN_3O_3S$, M_r 365.84, *mp* 160 – 162 °C, pK_a 8.3, is lipid soluble.

Trade names: Fludex, Ipamox, Natrilix.

The site of diuretic action of indapamide is similar to that of thiazide diuretics: the proximal part of distal tubules. Its duration of diuretic action can exceed 24 h. Indapamidediffers from thiazide diuretics in having additional direct smooth muscle relaxant effect and hypotensive effect at doses below those required to produce diuresis [20].

Side effects of indapamide include hypokalemia, hyperuricemia, and metabolic alkalosis, and are usually observed at doses higher than required for satisfactory control of arterial pressure in hypertensive patients (5 mg/d and higher). The safe and effective clinical dose is 2.5 mg/d.

2.2. K^+-Retaining Diuretics

The development of K^+-retaining (or sparing) diuretics was stimulated by the fact that K^+ loss and consequent hypokalemia represent the major side effect of sulfonamide diuretics. It was considered possible to develop K^+-retaining diuretics because aldosterone was known to retain sodium while increasing the excretion of potassium. The search for aldosterone inhibitors led to the development of spironolactone [*52-01-7*] [22], which is still used clinically. The activity of spironolactone is determined by its ability to compete with aldosterone at its receptor sites. It is inactive in the absence of aldosterone but has diuretic, natriuretic, and K^+ retaining effects in its presence. The pharmacological effects of spironolactone are to a great extent due to its active metabolite, canrenone, which is formed within 3 min after administration of spironolactone and has a half-life of ca. 12 h.

The clinical use of spironolactone is severely limited by *side effects,* including hyperkalemia, gynecomastia, impotence, and menstrual cycle abnormalities. In rats chronic spironolactone treatment was tumorigenic, and therefore, its use is restricted. Because of these side effects and the high cost of spironolactone, the search for nonsteroidal aldosterone inhibitors continued and led to the development of K^+-retaining diuretics that increase sodium excretion without inhibiting aldosterone.

Two major K^+-retaining diuretics which do not inhibit aldosterone are amiloride and triamterene.

2.2.1. Amiloride

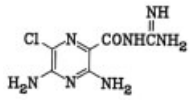

Amiloride [*2609-46-3*], 3,5-diamino-*N*-(aminoiminomethyl)-6-chloropyrazinecarboxamide, *N*-amidino-3,5-diamino-6-chloropyrazinecarboxamide, $C_6H_8ClN_7O$, M_r 229.65, *mp* 293.5 °C, is soluble in water. It is widely used in combination with hydrochlorothiazide (Moduretic).

Trade names: Colectril and Midamor.

The site of action of amiloride is in the distal portion of distal tubules and in the collecting duct, where amiloride inhibits sodium reabsorption and potassium secretion [6]. It is thought to block the Na^+ channel and the Na^+/H^+ exchange pump. In addition to the renal tubules, amiloride blocks Na^+ channels in various epithelial membranes, including the salivary ducts. In hypertension, if used alone, amiloride is not as effective as hydrochlorothiazide. The major *side effect* of amiloride is hyperkalemia. Other side effects include nausea, anorexia, and abdominal pain.

The clinical dose range of amiloride is 5 – 20 mg/d. Potassium supplements are contraindicated with amiloride; other K^+-retaining diuretics should not be combined with amiloride.

2.2.2. Triamterene

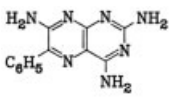

Triamterene [*396-01-0*], 6-phenyl-2,4,7-pteridinetriamine, 2,4,7-triamino-6-phenylpteridine, $C_{12}H_{11}N_7$, M_r 253.26, *mp* 316 °C, is insoluble in water or chloroform and only slightly soluble in alcohol.

Trade names: Dyrenium; Dyazide (combination with hydrochlorothiazide).

Triamterene increases urinary excretion of sodium and chloride and decreases excretion of potassium and hydrogen ions. The primary site of triamterene action is in the distal tubules [7]. The biochemical effects of triamterene are not well-known, although triamterene is capable of forming "false" nucleotides.

The principal use of triamterene is in combination with hydrochlorothiazide. The natriuretic effects of both diuretics are additive, but the kaliuretic effect of hydrochlorothiazide is antagonized by triamterene.

The major *side effect* of triamterene alone is hyperkalemia. Nausea, leg cramps, and dizziness have also been described. The initial clinical dose is 100 mg twice daily. Dyazide capsules contain 50 mg of triamterene and 25 mg of hydrochlorothiazide and are given 1 – 4 times daily.

2.3. Other Diuretics

In addition to sulfonamides and pyrazines, diuretic activity has been discovered among many other chemical classes of drugs. Among them are aryloxyacetic acids, 2-aminomethylphenols, and various heterocyclic compounds. Because all diuretics cannot be covered in this article, compounds that are important therapeutically as well as historically have been selected. The three diuretics discussed (ethacrynic acid, indacrinone, and muzolimine) differ from sulfonamides and pyrazines in their pharmacological profiles and represent three different chemical classes.

2.3.1. Ethacrynic Acid

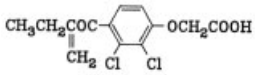

Ethacrynic acid [*58-54-8*], [2,3-dichloro-4-(2-methylene-1-oxybutyl)phenoxy]acetic acid, $C_{13}H_{12}Cl_2O_4$, M_r 303.14, *mp* 121 – 125 °C, is only slightly soluble in water but very soluble in alcohol (1 g/1.6 mL).

The efficacy (ceiling effect) of ethacrynic acid is similar to that of furosemide and considerably higher than that of hydrochlorothiazide. Ethacrynic acid increases the excretion of sodium and chloride. It also slightly increases the excretion of potassium. The major site of action is the ascending limb of Henle's loop [13]. The biochemical mechanism of action is not known precisely, although at high concentrations ethacrynic acid inhibits renal Na^+-K^+ dependent ATPase and oxygen consumption in medullary tubules. Ethacrynic acid reacts with thiols in the body fluids and forms adducts which are likely to represent the active forms of this

drug. The *side effects* of ethacrynic acid include abdominal discomfort, nausea, hyperuricemia, neutropenia, thrombocytopenia, ototoxicity, and metabolic alkalosis. The usual clinical doses are 50 mg twice a day; the maximal daily dose is 400 mg.

2.3.2. Indacrinone

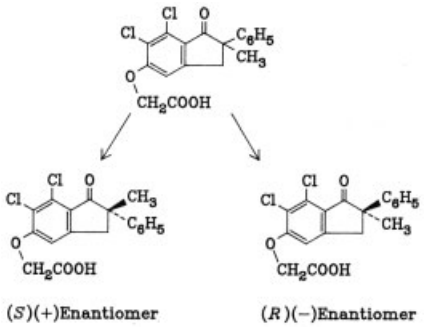

Indacrinone [*57296-63-6*], [(6,7-dichloro-2,3-dihydro-2-methyl-1-oxo-2-phenyl-1*H*-inden-5-yl)oxy]acetic acid, M_r 365.19. A 9 : 1 ratio of (+)- to (−)-indacrinone enantiomers was studied as MK-286 and found to be isouricemic during chronic use. The (+)-enantiomer is uricosuric, whereas the (−)-enantiomer retains uric acid. The racemic mixture was tested as MK-196 [23–25].

Indacrinone and its enantiomers lower arterial pressure in hypertensive animals and in humans, inhibit sodium and chloride reabsorption in the ascending limb of Henle's loop, and increase excretion of potassium. Hypokalemia can be antagonized by amiloride. The site of action of enantiomers in the renal tubules appears to differ. The (−)-enantiomer has its primary action in the medullary portion of the ascending limb, while the (+)-enantiomer has its primary action in the cortical diluting segment. The saluretic efficacy of indacrinone is similar to or greater than that of furosemide, while the duration of action is longer. *Side effects* of indacrinone in humans include headache, lethargy, weakness, dry mouth, and slight, but transient elevations in SGOT and/or SGPT. The clinical doses of the (−)-enantiomer ranged from 10 to 40 mg, while the (+)-enantiomer was given at doses from 10 to 400 mg.

2.3.3. Muzolimine

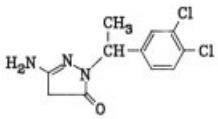

Muzolimine [*55294-15-0*], 5-amino-2-[1-(3,4-dichlorophenyl)-ethyl]-2,4-dihydro-3*H*-pyrazol-3-one, $C_{11}H_{11}Cl_2N_3O$, M_r 272.13, *mp* 127–129 °C, is insoluble in water, but is soluble in ethanol, polyethylene glycol, or dimethyl sulfoxide. It is highly lipophilic, with a pK_a value of 9.3 and a partition coefficient in octanol/water of 2.29.

Trade name: Edrul.

The site of diuretic action of muzolimine is the ascending limb of Henle's loop. Unlike other loop diuretics muzolimine does not act at the luminal site. It appears to inhibit sodium and chloride transport at the peritubular site of the cell membrane. Muzolimine exhibits a high-efficacy (ceiling) effect similar to that of furosemide, but has a considerably longer duration of action. The duration of antihypertensive action of muzolimine at a single dose exceeds 24 h [26], [27].

Clinical doses of muzolimine range from 10 to 480 mg; good therapeutic results were obtained with doses of 30 mg/d in most patients. Muzolimine is particularly useful in the treatment of edema associated with renal failure.

3. Inhibitors of Angiotensin Converting Enzyme

The renin–angiotensin system is involved in the control of blood pressure in normotensive and hypertensive states. The conversion of angiotensinogen to angiotensin II involves a two-step sequence:

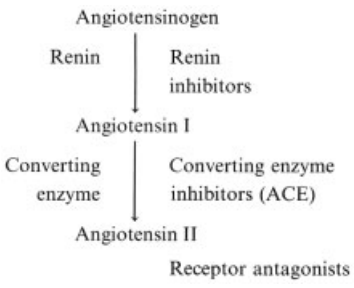

Pharmacological approaches for interference with the renin – angiotensin system have focused on renin inhibitors, angiotensin converting enzyme (ACE) inhibitors, and angiotensin II receptor antagonists [28]. Captopril was the first orally effective ACE inhibitor approved for use in mild-to-moderate hypertension. Enalapril was developed later as a nonsulfhydryl-containing inhibitor.

3.1. Captopril

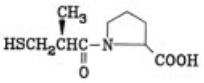

Captopril [*62571-86-2*], 1-(3-mercapto-2-methyl-1-oxopropyl)-L-proline, $C_9H_{15}NO_3S$, M_r 217.28, *mp* 87 – 88 °C, is soluble in water, alcohol, and chloroform;

Trade name: Capoten.

The rationale and design of the angiotensin converting enzyme (ACE) inhibitors have been described [29], [30]. The effectiveness of captopril on ACE was shown by its ability to inhibit the rise in blood pressure induced by angiotensin I, while leaving unchanged the responses to angiotensin II [31].

It has been postulated that the antihypertensive mechanism may include not only inhibition of plasma ACE activity, but also inhibition of local tissue ACE, especially in the vascular smooth muscle [32–35]. In addition, a number of alternative mechanisms could explain the efficacy of ACE inhibitors in nonrenin-dependent hypertension [36]. The alternatives include (1) interference with α-adrenergic vasoconstriction [37–42]; (2) a central inhibition of sympathetic nervous outflow [35], [43]; (3) elevation in bradykinin or prostaglandins, due to the similarity of kininase II and ACE [31], [32], [44]; and (4) alteration in the permeability of vascular smooth muscle to sodium ions [45].

Clinically, the antihypertensive actions of captopril are associated with a reduction in total peripheral resistance and a slight to moderate increase in cardiac output [46], [47]. In malignant hypertensive patients captopril was very effective during acute and long-term treatment [48].

The hypotensive responses are not usually accompanied by a reflex increase in heart rate. Captopril facilitates the reflex slowing of heart rate in some manner [38]. This could involve the removal of facilitatory effects of angiotensin II on sympathetic nervous transmission to the heart or a central inhibitory effect of angiotensin II on cardiac vagal tone [49]. Further studies have also indicated that captopril possesses cardioprotective properties [50].

Captopril decreases renal vascular resistance, and increases sodium and water excretion [51], [52].

Renin levels are increased and aldosterone levels are reduced by captopril. A certain number of patients with essential hypertension exhibit a subnormal aldosterone secretory response to angiotensin II challenge [53]. The abnormal responders also exhibit higher baseline plasma angiotensin II and lower levels of aldosterone. Captopril treatment for 72 h removed the difference in baseline values and partially corrected the abnormal aldosterone response to angiotensin II.

The absorption of captopril in humans is rapid and extensive; 75 % of the radioactivity after a single oral dose of ^{14}C-captopril was recovered in the urine within 48 h. The metabolites of captopril include disulfides and endogenous sulfhydryl compounds, e.g., cysteine.

The *side effects* of captopril include proteinuria, maculopapular rash, itch, disturbance of taste, and occasional leukopenia [54].

3.2. Enalapril

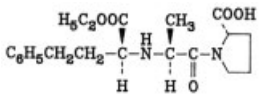

Enalapril [*75847-73-3*], 1-[*N*-[1-(ethoxycarbonyl)-3-phenylpropyl]-L-alanyl]-L-proline, $C_{20}H_{28}N_2O_5$, M_r 376.45, *mp* 143 – 144.5 °C, is a white to off-white crystalline powder, soluble in methanol and slightly soluble in water. Unlike captopril, enalapril does not contain the sulfhydryl moiety which has been implicated in some adverse side effects, such as rash and loss of taste.

Enalapril is readily absorbed by oral administration while the active diacid form is poorly absorbed. Deesterification of the prodrug takes place primarily in the liver [55]. Enalapril is about 17-fold more potent than captopril for inhibition of hog plasma ACE. In both rats and dogs, enalapril inhibits the pressor responses to angiotensin I for approximately 6 h. Although the onset of action is somewhat slower, the potency of enalapril in vivo is about 4 – 8 fold greater than that of captopril in the same species [56].

Similar to captopril, the antihypertensive action of enalapril is most pronounced in high-renin forms of hypertension and enalapril also exhibits antihypertensive activity in low-renin models [57].

In clinical studies, enalapril effectively lowers arterial pressure when given as sole therapy or in combination with diuretics [58–60]. Similar to preclinical studies, the onset of inhibition of ACE in humans is slower than with captopril. This finding has been attributed to the time needed for the deesterification reaction [61].

Reflex tachycardia is seldom observed with enalapril [62], [63]. Cardiac function is well-maintained [62], [64] and, in one group of hypertensive patients, the left ventricular mass was reduced during therapy [62].

With regard to renal function, enalapril increases sodium excretion, glomerular filtration rate, and renal blood flow in sodium-restricted dogs [55]. In spontaneously hypertensive rats (SHR), antihypertensive doses of enalapril do not cause sodium and water retention upon chronic administration [65]. Renal vasodilatation and natriuresis have been observed in hypertensive patients treated with enalapril [62], [63].

Hormonal responses during acute and chronic (3 months) enalapril treatment consist of decreased angiotensin II and aldosterone levels and a rise in plasma renin activity. In addition, bradykinin and PGE_2 were unchanged by a single dose of enalapril [63], [66]. Epinephrine levels were consistently decreased by enalapril, whereas, in hypertensive patients during chronic therapy with enalapril, plasma norepinephrine was reduced insignificantly [62]. In normotensive subjects, the direction of change in norepinephrine levels was dependent on the sodium intake [66], [67].

The mechanism of action of enalapril remains to be fully elucidated. The dominant mechanism most certainly involves the renin–angiotensin system; however, additional mechanisms must also be considered.

The major toxicological findings in dogs were renal functional changes, and renal tubular degeneration. Toxic actions of enalapril are exacerbated by coadministration of hydrochlorothiazide and attenuated by saline supplements [55]. In humans the major *side effects* include rash, angioedema, proteinuria, and renal insufficiency [54].

4. Inhibitors of the Sympathetic Nervous System

4.1. Drugs with Primarily Central Actions

Some of the widely used antihypertensive drugs are thought to lower arterial pressure by centrally induced blockade of sympathetic tone and consequent reduction of peripheral vascular resistance. The central control of the sympathetic nervous system is mediated by various neurotransmitters, including norepinephrine, and by drugs which enhance norepinephrine release or stimulate postsynaptic α-adrenoceptors involved in the regulation of sympathetic tone. The 5HT agonists, dopaminergic, or GABA ergic drugs, were also shown or are assumed to control hypertension by a central action. It is presently not clear whether their effects are mediated through central inhibition of the sympathetic nervous system or by other central mechanisms.

The major goal in the development of new centrally acting drugs is to achieve a high degree of selectivity for blood pressure control, which would lead to antihypertensive action without sedative, skeletal muscle relaxant, or emetic side effects. For further details on the central control of arterial pressure by drugs see reviews [68], [69].

4.1.1. Methyldopa

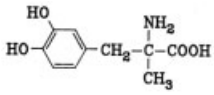

The antihypertensive activity of methyldopa [*555-30-6*], 3-hydroxy-α-methyl-L-tyrosine, $C_{10}H_{13}NO_4$, M_r 211.21, was first attributed to the inhibition of aromatic amino acid decarboxylase [70] and later to a formation of a "false transmitter" [71], [72]. It is postulated that α-methylnorepinephrine, a methyldopa metabolite, lowers arterial pressure by stimulating central adrenoceptors, which causes a reduction in sympathetic nervous outflow [73], [74].

The antihypertensive effect of methyldopa in humans is associated with a reduction in the peripheral vascular resistance and in the plasma renin activity [74–76]. Methyldopa increases coronary and cerebral blood flow [77], [78].

The oral absorption of methyldopa in humans was reported to vary from 26 to 74 % [79]. The peak plasma levels of methyldopa are observed 3 – 6 h after oral dose. The major urinary metabolites of methyldopa are methyldopa-*o*-sulfate and 3-methoxymethyldopa. About 80 to 90 % of the oral dose of methyldopa is eliminated within 48 h after treatment [80].

The initial dose of methyldopa in mild hypertension is 250 mg t.i.d. (3 times/d), whereas in severe hypertension methyldopa is used at doses up to 500 mg q.i.d. (4 times/d). The clinical advantages of methyldopa include its effectiveness in controlling hypertension, many years of clinical experience, and favorable hemodynamic effects, including maintenance of renal circulation and decrease in plasma renin activity. *Side effects* of methyldopa include sedation, positive direct Coomb's test, hepatotoxicity, and allergic reactions.

4.1.2. Clonidine

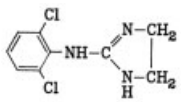

Clonidine [*4205-90-7*], 2-(2,6-dichloroanilino)-2-imidazoline, $C_9H_9Cl_2N_3$, M_r 230.10, *mp* ca. 300 °C, pK_a 8.2, is easily soluble in water (1 g/13 mL) and ethanol (1 g/25 mL).

Trade names: Catapres, Catapressan.

Clonidine is an α-adrenoceptor stimulant, which stimulates preferentially α_2-adrenoceptors [81], [82]. The stimulation of central α-adrenoceptors by clonidine leads to inhibition of central sympathetic tone, resulting in a lowering of arterial pressure and of heart rate [83–87].

Clonidine has many other pharmacological effects. The most prominent central effect of clonidine is sedation. Clonidine decreases spontaneous motor activity and produces ataxia and catatonia as well as signs of increased peripheral sympathetic activity in animals, i.e., piloerection and exophthalmus [88]. Clonidine has antinociceptive activity in mice [89] and is effective in the management of withdrawal reaction in opiate addicts [90]. Intraocular pressure is decreased by clonidine [91]. Clonidine reduces salivary secretion and has a biphasic effect on gastric secretion: inhibition at lower and stimulation at higher doses [88], [92]. In rats and dogs clonidine has a diuretic effect which appears to be mediated by inhibition of ADH release [93].

Clonidine is almost completely absorbed by oral administration; its bioavailability in humans is ca. 75 % and the drug has no extensive first-pass metabolism [94]. In humans ca. 50 % of clonidine is excreted unchanged. The major metabolites of clonidine are *p*-hydroxyclonidine and dichlorophenylguanidine [84].

The major *undesirable effect* of clonidine is the withdrawal reaction, which consists of a sudden rise in arterial pressure, nervousness, agitation, and an increase in heart rate [95], [96]. Other side effects include dry mouth and sedation.

The recommended starting dose of clonidine is 0.1 mg/d, but it should not exceed a total daily dose of 1 mg. The pharmacology of clonidine is reviewed extensively elsewhere [84].

4.1.3. Guanfacine

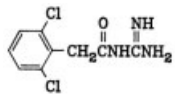

Guanfacine [*29110-47-2*], *N*-amidino-2-(2,6-dichlorophenyl)acetamide, $C_9H_9Cl_2N_3O$, M_r

246.08, *mp* 214 – 218 °C, is water soluble with a pK_a of 7.1. Guanfacine is one of the most potent members of a class of antihypertensive phenylacylguanidines [97].

Trade names: Estulic, Estuline.

Guanfacine is similar to clonidine in that the mechanism of antihypertensive action is attributed largely to stimulation of central adrenoceptors [98], [99]. Guanfacine is claimed to differ from clonidine in having less depressant effect on the central nervous system (CNS) and possibly may act at a different site within the CNS [100], [101]. Like clonidine, guanfacine exhibits a peripheral vasoconstrictor effect that is mediated most likely by stimulation of α_2-adrenoceptors. Guanfacine is rapidly absorbed following oral administration with a bioavailability of 100 %. The half-life is approximately 21 h after oral dosing. The major metabolite in humans is the 3-hydroxy derivative of guanfacine [100].

Dryness of mouth and sedation are the major *side effects* of guanfacine in the clinic. The incidence of side effects declines considerably with the duration of treatment. The withdrawal syndrome (rise in blood pressure and tachycardia) was seen after abrupt discontinuation of guanfacine in 2 – 4 % of the patients, but was less pronounced than after withdrawal of clonidine. The average dose of guanfacine in clinical trials was 3 mg/d. The optimal effect without side effects was obtained at a single dose of 2 mg [100], [102], [103].

4.1.4. Guanabenz

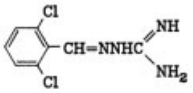

Guanabenz [*5051-62-7*], 2-[(2,6-dichlorophenyl)methylene]hydrazinecarboximidamide, [(2,6-dichlorobenzylidene)amino]guanidine, $C_8H_8Cl_2N_4$, M_r 231.07, *mp* 227 – 229 °C.

Trade names: Rexitene, Wytensin.

Like clonidine or guanfacine, guanabenz stimulates central α-adrenoceptors and lowers arterial pressure [104–108]. In hypertensive patients oral administration of guanabenz lowers arterial pressure and total peripheral vascular resistance with little or no effect on myocardial function.

In animals or humans guanabenz has no adverse effects on renal function except a slight and transient fall in filtration rate after the first dose [109], [110]. Under certain conditions guanabenz increased excretion of water, and this effect was attributed to the antagonism of vasopressin [111].

In humans guanabenz is well absorbed and exhibits a variable half-life of 7 – 10 h. Less than 1 % of the dose is excreted as unchanged drug. The major metabolites of guanabenz in rat, swine, Rhesus monkeys, and humans are *p*-hydroxyguanabenz and the glucuronide conjugate [110], [112]. Clinical studies demonstrated antihypertensive efficacy of guanabenz at doses from 16 to 65 mg/d [113], [114]. The most common *side effects* of guanabenz are dry mouth, sedation, weakness, and tiredness.

4.2. Drugs with Primarily Peripheral Action

Drugs can inhibit sympathetic transmission at various peripheral sites: (1) sympathetic ganglia, (2) sympathetic nerve endings, or (3) postsynaptic adrenoceptors. Ganglionic blocking drugs were used in the treatment of hypertension 20 to 30 years ago. Hexamethonium [*60-26-4*], a standard ganglionic blocking drug, is still used today as a tool in experimental pharmacology. Typical ganglionic blocking drugs (e.g., pentolinium, chlorisondamine, mecamylamine, and pempidine) block nicotinic receptors at the postsynaptic membrane of the sympathetic ganglia and prevent acetylcholine-induced depolarization. However, they are obsolete in the treatment of hypertension because of unacceptable side effects, which include atony of the gastrointestinal tract, inhibition of gastric and salivary secretion, incomplete emptying of the urinary bladder, cycloplegia, and impotence. These side effects are to a great extent due to the blockade of parasympathetic ganglia. Blockade of sympathetic ganglia can lead to additional side effects, e.g., postural hypotension, syncope, and inhibition of local vasomotor reflexes. Today the only indication remaining for ganglionic blockade is to produce so-called "controlled hypotension"

in surgery in order to reduce bleeding, facilitate small vessel surgery, and improve myocardial performance by reducing preload and afterload. For this indication, trimethaphan camsylate [*68-91-7*] (Arfonad) by infusion is the drug of choice. Its short duration of action permits precise control of blood pressure during infusion and blood pressure returns to control levels immediately after discontinuation of the infusion.

The search for new ganglionic blocking drugs was largely discontinued by 1960. It became clear that hypertension could be controlled by drugs relatively free from the disabling effects accompanying ganglionic blockade. The drugs which replaced the ganglionic blocking agents included diuretics and drugs which inhibited the sympathetic nervous system at loci peripheral to the ganglia; namely, at adrenergic nerve endings or at postsynaptic adrenoceptor sites.

4.2.1. Amine Depleting and Adrenergic Neuron Blocking Agents

At the sympathetic nerve endings drugs can decrease neurotransmitter (norepinephrine) release by depleting it from the storage sites in the nerve terminals or by decreasing the release by stimulation of inhibitory presynaptic α_2-adrenoceptors.

Antihypertensive drugs which lower arterial pressure primarily by an action at the sympathetic nerve endings include reserpine and guanethidine, which are discussed in this section. Other drugs which act by a mechanism similar to that of guanethidine and are marketed in some countries include: bethanidine, guanadrel, debrisoquin, guanacline, and guanochlor. They differ from guanethidine in potency, pharmacokinetics, and some aspects of their pharmacology. The mechanism of their antihypertensive action and the profile of their side effects are, however, not essentially different from those of guanethidine. Orthostatic hypotension is the major side effect which restricts their therapeutic usefulness.

4.2.1.1. Reserpine

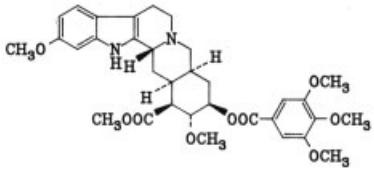

Reserpine [*50-55-5*], 11,17-dimethoxy-18-[(3,4,5-trimethoxybenzoyl)oxy]yohimban-16-carboxylic acid methyl ester, $C_{33}H_{40}N_2O_9$, M_r 608.70, *mp* 264 – 265 °C (decomp.), is poorly soluble in water, but is soluble in alcohol, glacial acetic acid, and chloroform. Reserpine was originally isolated from roots of *Rauwolfia serpentina* L. Benth [115].

Synthesis: [116].

Trade names: Alserin, Reserpex, Serpasil.

Reserpine depletes the transmitters from adrenergic, dopaminergic, and serotonergic neurons. Depletion of norepinephrine from adrenergic nerve endings is thought to be primarily responsible for the antihypertensive action of reserpine. The onset of the antihypertensive action is gradual. The maximal effect is seen 6 – 12 h after a single dose of reserpine. The antihypertensive effect is associated with a reduction in cardiac output and in peripheral vascular resistance. After repeated administration cardiac output returns to control values, while peripheral resistance remains at a reduced level [117–119].

Reserpine is well absorbed by oral administration. The major metabolites of reserpine are products of hydrolytic cleavage: methyl reserpate and trimethoxybenzoic acid. Urinary excretion of reserpine metabolites accounts for 20 – 40 % of the administered doses and fecal excretion for less than 10 % [120], [121].

The major *side effects* of reserpine are sedation, nasal stuffiness, and a depressive reaction. Reserpine is also known to increase prolactin release. Adrenocorticotrophic hormone (ACTH) secretion is initially increased and subsequently abolished by reserpine. Gastric secretion is increased and formation of gastric ulcers is promoted. The initial dose of reserpine utilized in hypertension is 0.12 – 1.0 mg/d and the maintenance dose is 0.25 mg or less. The use of reserpine alone has decreased because of the availability of drugs with less severe side effects. In

combination with thiazide diuretics reserpine is, however, still widely used.

4.2.1.2. Guanethidine

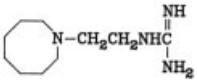

Guanethidine [*55-65-2*], [2-(hexahydro-1(-2*H*)-azocinyl)ethyl]guanidine, $C_{10}H_{22}N_4$, M_r 198.31, *mp* 276 – 281 °C.

Trade names: (sulfate salt) Ismelin, Guethine, Iporal, Isobarin.

The major pharmacological action of guanethidine is inhibition of release and of uptake of norepinephrine by sympathetic nerve endings [122–125]. After treatment with guanethidine, vasoconstrictor nerves do not provide reflex compensation to the erect position. Thus, the hypotension after guanethidine is largely orthostatically induced, which can lead to undesirable effects, such as dizziness and syncope. In addition to its adrenergic neuron blocking action, guanethidine has also a direct vasodilator activity, the mechanism of which may involve stimulation of vascular β-adrenoceptors [123], [126]. The hypotensive action of guanethidine can be antagonized by amphetamine, ephedrine, cocaine, or by tricyclic antidepressants [123], [127].

By acute intravenous administrations to animals or humans, guanethidine produces a transient increase in arterial pressure which is followed by hypotension. Peripheral vascular resistance is not changed, while cardiac output is decreased [128], [129]. In chronic oral therapy guanethidine does not change cardiac output [130].

Guanethidine is not completely absorbed from the gastrointestinal tract, with 20 – 25 % of oral dose being recovered from the feces [131]. 2-(6-Carboxyhexylamino)ethylguanidine and guanethidine *N*-oxide have been identified as metabolites of guanethidine [132], [133]. Guanethidine is given once a day at doses ranging from 12.5 to 150 up to 400 mg/d. The doses are adjusted by 10 or 25 mg at 2-week intervals. In addition to orthostatic hypotension, *side effects* of guanethidine include water retention, failure of ejaculation, diarrhea, tenderness in the region of parotid glands, and muscle weakness.

4.2.2. α-Adrenoceptor Antagonists

α-Adrenoceptor antagonists, e.g., phentolamine, tolazoline, or yohimbine, were known for many years [134], but were not used in the treatment of hypertension, because of their questionable effectiveness, rapid development of tolerance, and pronounced reflex tachycardia. The introduction of prazosin [135], [136] and demonstration of its usefulness in the treatment of hypertension resulted in the development of various new α-adrenoceptor antagonists. The selective α_1-adrenoceptor antagonists are more effective in hypertension than antagonists affecting both α_1- and α_2-adrenoceptors. The reduced effectiveness is most likely due to an inhibition of presynaptic α_2-adrenoceptor by nonselective antagonists. This leads to an increase in neurotransmitter release and consequent tachycardia and vasoconstriction, which oppose the hypotensive effects of the drugs as a result of blockade of postsynaptic α_1-adrenoceptors. The three drugs discussed here are relatively selective $\alpha_1$adrenoceptor antagonists.

4.2.2.1. Prazosin

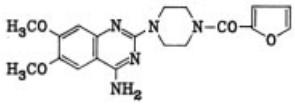

Prazosin [*19216-56-9*], 1-(4-amino-6,7-dimethoxy-2-quinazolinyl)-4-(2-furanylcarbonyl)piperazine, $C_{19}H_{21}N_5O_4$, M_r 382.42, *mp* 278 – 280 °C, is water soluble.

Trade names: Minipress, Hypovase, Sinetens.

The blockade of α_1-adrenoceptors represents the major mechanism of antihypertensive action of prazosin [135–139]. The α_1-adrenoceptor antagonist action of prazosin leads to arterial and venous dilatation, a reduction in preload and afterload, and, consequently, to an improvement in cardiac efficiency. Because of the indirect improvement in cardiac function, prazosin is also

useful in the therapy of heart failure [140]. In spite of pronounced vasodilatation, heart rate undergoes no change or only a slight increase after prazosin [138].

A consistent decrease in peripheral vascular resistance differentiates prazosin from methyldopa, clonidine or β-adrenoceptor antagonists. In spite of pronounced vasodilator action, prazosin in patients either decreases or has no effect on plasma renin activity [141], [142]. The peak plasma concentration of prazosin is reached 2 – 3 h after an oral dose and its half-life is 3.9 h [143]. The major metabolites of prazosin in rats, dogs, and humans are 6-*O*- and 7-*O*-demethylprazosin. Other minor metabolites have been identified [144], [145].

The antihypertensive activity of prazosin is comparable to, or better than, that of methyldopa [146], hydralazine [147], or clonidine [148]. Prazosin is successfully used in combination with thiazide diuretics [149] or β-adrenoceptor antagonists [150].

The major *side effect* of prazosin is the so-called "first dose phenomenon", a syncope in patients receiving a 2-mg tablet as the first dose [151]. The syncope can be avoided if the initial dose is reduced to 0.5 mg and given at bedtime. The dose can then be increased gradually over a 2-week period. Other reported side effects include an increase in the frequency of urination and in the frequency of anginal attacks. The usual dose of prazosin is 2 mg three times a day, but the drug was used at total daily doses up to 20 mg.

4.2.2.2. Indoramin

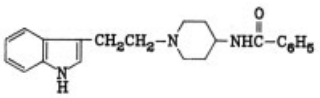

Indoramin [*26844-12-2*], *N*-[1-[2-(1*H*-indol-3-yl)ethyl]-4-piperidinyl]benzamide, $C_{22}H_{25}N_3O$, M_r 347.46, *mp* 208 – 210 °C. The synthesis and pharmacological activity of indoramin have been described [152–154].

Trade name: (hydrochloride salt) Baratol.

The mechanism of antihypertensive action of indoramin involves primarily the blockade of α_1-adrenoceptors. It is a competitive antagonist of norepinephrine and a noncompetitive antagonist of 5-HT and histamine. At doses higher than required for lowering of arterial pressure (5 vs. 1 mg/kg, i.v. in cats) indoramin reduces sympathetic nerve activity apparently by a central action [154–157]. Other pharmacological effects of indoramin include local anesthetic and antiarrhythmic effects [158], [159].

Indoramin is well absorbed orally. Its mean half-life in humans was estimated to be 5.5 h [160]. The metabolism of indoramin is extensive. The major human metabolite was identified as a conjugate of 6-hydroxyindoramin [161].

Extensive clinical studies have been conducted with indoramin [162–164]. Indoramin is effective in the treatment of hypertension at doses ranging from 20 to 150 mg/d. The major *side effect* is sedation. Indoramin does not produce the "rebound" phenomenon after discontinuation of therapy, nor syncope following the first dose. In addition to its antihypertensive action, indoramin was reported to produce therapeutic benefits in headache and bronchial asthma.

4.2.2.3. Urapidil

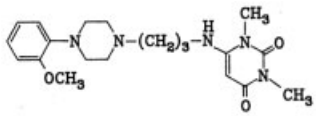

Urapidil [*34661-75-1*], 6-({3-[4-(2-methoxyphenyl)-1-piperazinyl]propyl}amino)-1,3-dimethyl-2,4(1*H*,3*H*)-pyrimidine-dione, $C_{20}H_{29}N_5O_3$, M_r 387.49, *mp* 156 – 158 °C, is poorly soluble in water, but readily soluble in dilute acids.

Synthesis: [165].

Trade name: Ebrantil.

Urapidil lowers arterial pressure in hypertensive and normotensive animals. Its antihypertensive activity is due to a combination of peripheral and central actions. The peripheral action involves blockade of α_1-adrenoceptors in vascular smooth muscle. By virtue of its central action urapidil diminishes sympathetic outflow, an effect which is probably mediated by central

α-adrenoceptors and is essentially similar to that of clonidine [166], [167].

Urapidil is well absorbed orally. The maximal blood levels are reached after 0.5 – 1 h after oral administration of tablets and 2.5 – 4 h after administration of slow-release capsules. The plasma half-life of urapidil in humans was estimated as 2.5 – 3.7 h. Urapidil is metabolized by *p*-hydroxylation, *O*-demethylation of the methoxyphenyl ring, and N^1-demethylation of the uracil residue. The same metabolites are found in humans, dogs, rats, and mice. The *p*-hydroxylation product has antihypertensive activity but is less potent than urapidil [167], [168].

In animals, urapidil exhibits sedative activity, while in humans, the major *side effects* include dizziness, headache, and nausea. In the treatment of hypertension, urapidil is effective at 30 – 90 mg/d given as slow-release capsules [167], [169].

4.2.3. β-Adrenoceptor Antagonists

β-Adrenoceptor antagonists were introduced in the treatment of hypertension on the basis of clinical observations that pronethalol, one of the first β-adrenoceptor antagonists, lowered arterial pressure in patients with angina pectoris [170]. The mechanism of antihypertensive action of β-adrenoceptor blocking drugs is still controversial. Various hypotheses have been critically reviewed [171]. Presently, resetting of baroreceptors and the inhibition of sympathetic transmission by blockade of presynaptic β-adrenoceptors are considered the most likely mechanisms of antihypertensive action, although lowering of plasma renin activity and a decrease in cardiac output by β-adrenoceptor antagonists may play a contributory role, particularly in the initial phase of therapy.

β-Adrenoceptor antagonists differ in potency, cardiac selectivity, intrinsic sympathomimetic activity (ISA), membrane stabilizing action (MSA), duration of action, tissue distribution, and metabolism.

A large number of β-adrenoceptor antagonists were developed and marketed during the last decade. The seven drugs discussed are more commonly used and represent different subgroups among β-adrenoceptor antagonists.

4.2.3.1. Propranolol

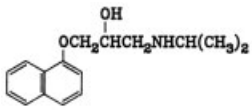

Propranolol [*525-66-6*], 1-(isopropylamino)-3-(1-naphthyloxy)-2-propanol, $C_{16}H_{21}NO_2$, M_r 259.34, pK_a 9.5, is moderately soluble in water and in ethanol. It is the most commonly used β-adrenoceptor antagonist. The β-adrenoceptor blocking and the antihypertensive activities reside in the (+)-isomer [172].

Trade names: Inderal, Inderex, Docitor, Tesnol, and many others.

Propranolol is not cardioselective; β_1- as well as β_2-adrenoceptors are blocked by the drug. Propranolol has MSA but no ISA. β_2-Adrenoceptor blocking activity can lead to bronchoconstriction and propranolol is, therefore, contraindicated in patients with bronchial asthma. The MSA of propranolol is correlated with local anesthetic as well as with the cardiac depressant actions. At therapeutic doses the MSA of propranolol is not prominent. The negative dromotropic activity of propranolol is evident following acute as well as chronic administration of the drug and is not opposed by the ISA [173], [174].

Following acute intravenous administration to animals or humans, propranolol reduces cardiac output, mean arterial pressure, heart work, left ventricular dp/dt, and blood flow in most vascular beds. Total peripheral vascular resistance is first increased but may be reduced during chronic therapy. Plasma renin activity is consistently reduced by the drug.

Propranolol is nearly completely absorbed from the gastrointestinal tract but is a subject of a large first-pass effect [175]. Peak plasma levels of propranolol are seen within 1 – 3 h after treatment [176]. Propranolol is extensively metabolized; 16 metabolites or conjugates have been identified. The major metabolites are naphthoxylactic acid and 4-hydroxypropranolol (an active metabolite) [177], [178].

In the treatment of hypertension propranolol is usually used at 80 – 320 mg/d in four divided doses, although it was used at doses as

low as 40 mg and as high as 2 g/d. The drug is commonly used in combination with a diuretic and/or a peripheral vasodilator drug [179], [180]. The *side effects* of propranolol include bradycardia, A-V heart block, cardiac insufficiency, bronchoconstriction, cold extremities, fatigue, vivid dreams, or even hallucinations. Rapid withdrawal of propranolol is not recommended because it may lead to coronary insufficiency or myocardial infarction.

4.2.3.2. Metoprolol

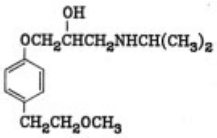

Metoprolol [*37350-58-6*], 1-[4-(2-methoxyethyl)phenoxy]-3-[(1-methylethyl)amino]-2-propanol, $C_{15}H_{25}NO_3$, M_r 267.38.

Trade names: Beloc, Betaloc, Lopressor, Lopresor.

Unlike propranolol, metoprolol is relatively specific for β_1-adrenoceptors and is, therefore, relatively free from bronchoconstrictor action [181], [182]. At therapeutic doses, metoprolol has no MSA, and therefore, has no local anesthetic or direct myocardial depressant effects. As a β_1-adrenoceptor antagonist, metoprolol is equipotent to propranolol [183]. Metoprolol is well absorbed by oral administration but has a substantial first-pass effect. Its plasma half-life is 3 – 4 h. Metoprolol is metabolized by *O*-demethylation, subsequent oxidation, oxidative deamination, and aliphatic hydroxylation. The two metabolites, *O*-demethylated product and α-OH-metoprolol, have β_1-adrenoceptor antagonist activity but are 5 times less potent than metoprolol [184], [185].

In clinical studies, metoprolol was shown to be equally effective as hydrochlorothiazide as an antihypertensive agent [186] and either equivalent or superior to other β-adrenoceptor antagonists [183]. In a double-blind study metoprolol was shown to reduce the mortality in patients with myocardial infarction [187].

The *side effects* of metoprolol include tiredness, headache, insomnia, nausea, heartburn, and depression. Withdrawal reactions consisting of headache, malaise, and palpitation were described for metoprolol. Thus, recommendations are to withdraw metoprolol slowly over a 10-d period. The initial dose of metoprolol in the treatment of hypertension is 50 mg twice daily. The maximal daily dose is 450 mg.

4.2.3.3. Nadolol

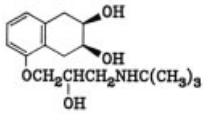

Nadolol [*42200-33-9*], 5-[3-[(1,1-dimethylethyl)amino]-2-hydroxypropoxyl]-1,2,3,4-tetrahydro-2,3-naphthalenediol, $C_{17}H_{27}NO_4$, M_r 309.42, *mp* 124 – 136 °C, pK_a 9.67.

Trade names: Corgard, Solgol.

Like propranolol, nadolol is not cardioselective and has no ISA. Unlike propranolol, nadolol has no MSA and, therefore, no direct cardiac depressant activity. The therapeutic advantages of nadolol include ability to increase renal blood flow, and a long duration of action (up to 24 h) [188–190]. The oral absorption of nadolol in humans ranges from 20 to 34 %. Serum half-life of nadolol in humans is significantly longer than that of other β-adrenoceptor antagonists and was estimated to equal 24 h [191].

In the treatment of hypertension nadolol is given once a day at 40 – 240 mg [192]. In addition to its antihypertensive effect nadolol was shown to be effective in the treatment of angina pectoris [193] and of certain types of ventricular arrhythmias [194]. Most of the *side effects* of nadolol are similar to those of propranolol.

4.2.3.4. Pindolol

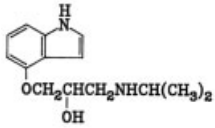

Pindolol [*13523-86-9*], 1-(1*H*-indol-4-yloxy)-3-[(1-methylethyl)amino]-2-propanol, $C_{14}H_{20}N_2O_2$, M_r 248.32, *mp* 171 – 173 °C.

Like propranolol, pindolol is not cardioselective, but unlike propranolol, pindolol has strong ISA, which is responsible for its acute vasodilator effect and for absence of bradycardia. The MSA of pindolol is considerably weaker than that of propranolol. In spite of the absence of cardioselectivity, pindolol is less likely to impair pulmonary function than propranolol, probably because of its strong ISA [195], [196].

Pindolol is well absorbed by oral administration and in comparison with other β-adrenoceptor antagonists, has a lower first-pass effect (13 %) [197]. The plasma half-life of pindolol is 3 – 4 h and the maximal plasma levels are reached 1.5 – 2 h after oral administration to humans [198]. It has a more rapid onset of action than propranolol and after withdrawal has a longer carryover effect [199–201]. The dose range of pindolol in the treatment of hypertension is 10 – 30 mg/d. The drug was initially administered t.i.d. but subsequently the majority of patients were found to respond equally well to once-a-day therapy [202].

The *side effects* of pindolol include insomnia and vivid dreams. Heart failure, heart block, bronchoconstriction, and circulatory disturbances are rare with pindolol.

4.2.3.5. Timolol

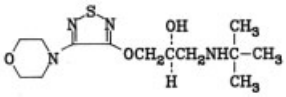

Timolol [*26839-75-8*], 1-[(1,1-dimethylethyl)amino]-3-{[4-morpholinyl-1,2,5-thiadiazol-3-yl]oxy}-2-propanol, $C_{13}H_{24}N_4O_3S$, M_r 316.42, is supplied as the hydrogen maleate salt, $C_{17}H_{28}N_4O_7S$, M_r 432.5, *mp* 201.5 – 202.5 °C, which is soluble in water and ethanol and is less lipophilic than propranolol. Unlike other β-adrenoceptor antagonists, timolol is available as the (−)-isomer and not as a racemic mixture.

Trade names: (hydrogen maleate salt) MK-950, Betim, Blocadren, Temserin, Timacor, Timolate, Timoptic, Timoptol.

Like propranolol, timolol is not cardioselective and has no ISA. At therapeutic doses timolol has little or no MSA and has practically no local anesthetic action. As a β-adrenoceptor antagonist timolol is ca. 8 times more potent than propranolol. As a negative inotropic agent timolol is 40 times less potent than propranolol [203–205].

Timolol is rapidly absorbed by oral administration with 72 % of the dose excreted within 84 h. The plasma half-life of timolol in humans was estimated as 5.5 h [206]. Five metabolites of timolol were identified in humans [205]. They have little or no β-adrenoceptor blocking activity.

The clinical study in essential hypertension indicated that timolol at 30 mg and propranolol at 120 mg had comparable antihypertensive activity [207]. Timolol is used in the treatment of hypertension at 10 – 30 mg/d. In patients surviving after acute myocardial infarction, timolol was shown to reduce mortality and reinfarction rates [208]. The *side effects* of timolol include tiredness, dizziness, increase in body mass, and an occasional increase in serum urea and creatinine.

4.2.3.6. Acebutolol

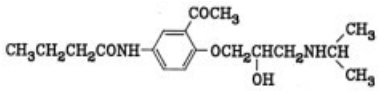

Acebutolol [*37517-30-9*], *N*-(3-acetyl-4-{2-hydroxy-3-[(1-methylethyl)amino]propoxy}-phenyl)butanamide, $C_{18}H_{28}N_2O_4$, M_r 336.43, *mp* 128 – 129 °C. Its *synthesis* and pharmacological activity were first described by Wooldridge [209].

Trade names: Sectral, Neptall, Prent.

Unlike propranolol, acebutolol is cardioselective and has weak ISA and MSA [210–214]. The major metabolite of acebutolol is diacetolol, *RS*-1-(2-acetyl-4-acetyl-amidophenoxy)-2-hydroxy-3-isopropylaminopropane), which has β-adrenoceptor blocking actions similar to those of acebutolol, but has no MSA [215]. The half-life of acebutolol and its metabolite is ca. 4 h. Acebutolol is metabolized in humans almost exclusively to diacetolol, 40 % of which is excreted unchanged [213].

The antihypertensive effect of acebutolol is more pronounced in severe than in mild hypertension [216] and was confirmed in numer-

ous controlled trials [217], [218]. In hypertension acebutolol is given once or twice daily at doses ranging from 200 to 1200 mg/d. *Side effects* of acebutolol include blurred vision, bronchospasm, and bradycardia.

4.2.3.7. Labetalol

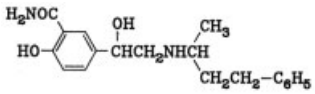

Labetalol [*36894-69-6*], 2-hydroxy-5-{1-hydroxy-2-[(1-methyl-3-phenylpropyl)amino]-ethyl}benzamide, $C_{19}H_{24}N_2O_3$, M_r 328.41; the hydrochloride salt, $C_{19}H_{25}ClN_2O_3$, is soluble in water and ethanol. Labetalol has two asymmetric centers and exists as a mixture of four isomers in equal proportions.

Trade names: (hydrochloride salt) Presdate, Trandate.

Labetalol differs from other β-adrenoceptor antagonists in its ability to block not only β- but also α-adrenoceptors, although it is more potent in blocking β-adrenoceptors. As a β-adrenoceptor antagonist, labetalol is not cardioselective, has weak ISA, but has clearly detectable MSA. As an α-adrenoceptor antagonist, labetalol is selective for α_1-adrenoceptors. Because of its α_1-adrenoceptor blocking activity labetalol decreases peripheral vascular resistance, particularly at higher doses. The antihypertensive activity of labetalol is likely to be due to α- as well as β-adrenoceptor blocking effects [219–221].

Following oral administration labetalol is well absorbed and excreted in the form of glucuronide conjugates. The major metabolite is the alcoholic glucuronide of labetalol [221]. Plasma labetalol levels are well correlated with its pharmacological actions [222]. Labetalol experiences high first-pass metabolism [223].

Extensive clinical studies with labetalol were published [224–226]. In humans labetalol lowers arterial pressure without any significant effect on resting heart rate or on cardiac output. It is used in hypertension at 200 – 2000 mg/d, usually in two divided doses. *Side effects* of labetalol include tiredness, dizziness, gastrointestinal disturbances, postural hypotension, impotence, and scalp tingling.

5. Smooth Muscle Relaxants

5.1. Ca^{2+} Channel Antagonists

The Ca^{2+} channel antagonists are a group of vasodilators used extensively in the treatment of angina pectoris (→ Calcium Antagonists, p. 519 ff.). However, their vasodilatory actions extend also to vascular beds other than the coronary and, consequently, their usefulness in hypertension has become apparent. The group classification emanates from the fact that these agents potently inhibit the entry of Ca^{2+} into vascular smooth muscle [227–229], although other mechanisms may contribute to their effectiveness as well [230], [231]. Considerable chemical diversity is found in the representatives of the Ca^{2+} channel antagonists [232], [233]. Nifedipine, nitrendipine, verapamil, and diltiazem have been studied most extensively in the treatment of hypertension; therefore, the emphasis here is placed on these agents.

5.1.1. Nifedipine

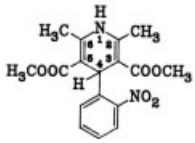

Nifedipine [*21829-25-4*], 1,4-dihydro-2,6-dimethyl-4-(2-nitrophenyl)-3,5-pyridinecarboxylic acid dimethyl ester, $C_{17}H_{18}N_2O_6$, M_r 346.34, *mp* 172 – 174 °C, is a yellow crystalline compound, practically insoluble in water, but soluble in organic solvents, such as ethanol, polyethylene glycol, and chloroform. In solution, it decomposes quite rapidly under normal laboratory light.

Trade names: BAY a 1040, Adalat, Procardia, Duranifin, Pidilat.

The structure – activity requirements are reported for several muscle systems, including vascular, cardiac, and nonvascular smooth muscle (see reviews [230], [234]). Specific requirements are the following: (1) the 1,4-dihydro-

pyridine (DHP) ring is essential; (2) an unsubstituted NH-group on the DHP-ring; (3) 2,6-substituents on the DHP-ring should be lower alkyl substituents, but an NH_2 group is acceptable; (4) ester groups in the 3- and 5-positions of the DHP-ring can be substituted to change the vascular selectivity and duration of action; (5) ester groups longer than COOMe will increase activity; (6) the aryl substituent (i.e., phenyl ring) is optimal in the 4-position of the DHP-ring; (7) position of the substituent on the phenyl ring is critical, i.e., *para* results in decreased activity, while *ortho* or *meta* result in increased activity; (8) stereoselectivity is obtained by different ester substitutions on C-3 and C-5 of DHP-ring [235–237].

Nifedipine lowers blood pressure in many experimental models of hypertension [238], [239]. It is efficacious in mild to moderate essentially hypertensive patients [240], [241] and in patients exhibiting severe hypertension [242], [243]. Following oral administration, the antihypertensive responses are rapid in onset (30 – 60 min), with a duration of about 3 h. Nifedipine exhibits diminished activity in normotensive subjects in comparison to hypertensives [240], [244–247].

The antihypertensive effect of nifedipine is due to arterial vasodilatation and a reduced total peripheral resistance [240], [244], [248–252]. Compensatory increases in heart rate, plasma norepinephrine, and renin levels are found after single doses of nifedipine; however, these elevations usually subside during long-term therapy [253–256].

Combined therapy of nifedipine with the β-adrenoceptor antagonists, propranolol or metoprolol, slightly enhanced the antihypertensive response in humans [240], [255]. The addition of methyldopa to nifedipine in severely hypertensive patients increased the antihypertensive response and reduced blood pressure fluctuations between dosing [242].

Sodium and water retention are not a problem during nifedipine treatment. Urinary volume and sodium excretion are elevated in hypertensive patients during nifedipine treatment [257], [258]. Should volume retention occur it can be controlled by diuretic therapy [241].

Nifedipine is bound extensively to plasma proteins (>90 %) [259], [260]. The plasma concentration of nifedipine correlates well to the hemodynamic responses [261], [262]. Metabolism occurs primarily in the liver, with the majority being excreted in the urine and feces as polar metabolites [260], [263].

The *adverse reactions* during therapy include headache, flushing, and nausea. In some patients, a localized ankle edema has been observed, which may be attributed to a change in capillary permeability or perfusion pressure [264]. The oral dosage form is presently approved for use in angina pectoris.

5.1.2. Nitrendipine

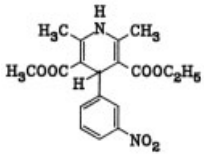

Nitrendipine [*39562-70-4*], 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)-3,5-pyridinedicarboxylic acid ethyl methyl ester], M_r 360.4, forms yellow crystals with solubility properties similar to nifedipine; however, nitrendipine is less light sensitive, with a decomposition half-life of about 7 h in laboratory light. It exhibits Ca^{2+} channel antagonism and produces antihypertensive effects in a number of animal models. It is more potent than nifedipine on vascular strips in vitro, and has a longer duration of antihypertensive action in the renal hypertensive dog model [265–275].

Trade names: Bayotensin, Baypress.

Synthesis: [268].

In clinical studies, single or repeated daily oral doses of nitrendipine (5 – 40 mg) effectively lower arterial pressure. The onset of action and peak effects are obtained at 15 min and at 60 – 90 min, respectively. The duration of action is approximately 8 h following a single oral dose, although, with long-term therapy, blood pressure can be controlled for up to 24 h with one dose [276–279]. In hemodynamic studies, nitrendipine lowered total peripheral resistance, but did not significantly affect cardiac output or heart rate [279], [280]. Plasma renin activity and catecholamines levels are increased by nitrendipine, but aldosterone levels are not significantly elevated [280].

In rats and humans, the metabolism of nitrendipine involves dehydrogenation, deesterification, and hydroxylation reactions. The identified metabolites exhibit essentially no antihypertensive activity [265]. *Adverse reactions* in humans include headache, flushing, and nausea.

5.1.3. Verapamil

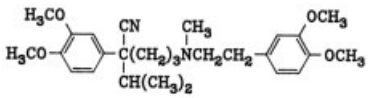

Verapamil [*52-53-9*], α-(3-{[2-(3,4-dimethoxyphenyl)ethyl]-methylamino}propyl)-3,4-dimethoxy-α-(1-methylethyl)benzeneacetonitrile, $C_{27}H_{38}N_2O_4$, M_r 454.59, is a white crystalline powder, soluble in water, chloroform, and methanol; the compound is not light sensitive. It is a Ca^{2+} channel antagonist which is unrelated chemically to nifedipine and nitrendipine.

Trade names: Isoptin, Calan, Cardibeltin, Cordilox, Vasolan.

Comprehensive reviews and studies on the pharmacology of verapamil have been published [281–287].

In clinical studies, verapamil lowered blood pressure in patients with essential hypertension [288–290]. The dosage ranged between 120 to 160 mg, given three times daily. Total peripheral resistance was decreased and cardiac output was elevated by verapamil. Heart rate was either unchanged or decreased during therapy. The antihypertensive efficacy was directly related to pretreatment levels of blood pressure and inversely related to plasma renin levels [290]. The vasodilator potency when given directly into the forearm circulation was greater in hypertensive patients in comparison with normotensive subjects. In addition, the increase in forearm blood flow correlated positively with basal concentrations of plasma epinephrine but not with plasma levels of renin or aldosterone [291]. Verapamil fails to elevate plasma catecholamine, renin, or aldosterone levels. The mechanism responsible for this effect is unknown; however, possibilities might involve catecholamine depletion, a decreased neuronal norepinephrine release, or depressed baroreceptor reflex sensitivity [292–296]. Constipation is the most prevalent *side effect* of verapamil therapy [283].

5.1.4. Diltiazem

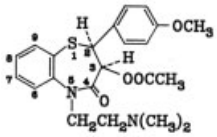

Diltiazem [*42399-41-7*], 3-(acetyloxy)-5-[2-(dimethylamino)ethyl]-2,3-dihydro-2-(4-methoxyphenyl)-1,5-benzothiazepin-4(5*H*)-one, $C_{22}H_{26}N_2O_4S$, M_r 414.52, is a white crystalline powder, freely soluble in water, methanol, or chloroform. The chemical synthesis of diltiazem and related substances has been reported [297], [298].

Trade names: Cardizem Herbesser, Dilzem.

The vasodilator structure – activity relationships have been reported and summarized [299–301]. Diltiazem is the D-*cis* isomer. The *trans* isomeric forms exhibit much lower vasodilator activity. Other substitutions which result in lower activity include replacement of the methoxy group with a hydroxyl group or additional alkoxy substituents on the aromatic ring. In addition, removal of the alkyl-aminoalkyl group at N_5 will significantly reduce the activity, as will dealkylation or quaternization of the terminal nitrogen. Other alterations, such as replacement of the methoxy group by a *p*-methyl and chlorine substitution at the 7-position of a condensed aromatic ring, do not reduce activity.

The antihypertensive activity of diltiazem has been demonstrated in DOCA, renal, and spontaneously hypertensive rats [302–304]. Diltiazem increased cardiac output, significantly elevated cardiac and skeletal muscle blood flow, and produced lesser flow increases to the mesenteric, cerebral, and renal vascular beds. Flow to the skin is usually unchanged or reduced by diltiazem [301], [305–307].

In humans, diltiazem is effective in mild-to moderate hypertension at oral doses ranging between 180 – 270 mg/d. The reduction in blood pressure (−10 to −15 %) was accompanied by an increase in cardiac output. Similar to preclinical studies, heart rate was not increased and plasma catecholamines were only moderately

increased [308]. In another study, intravenous diltiazem (0.2 mg/kg plus 0.1 mg/kg-min, i.v.) dilated both arterioles and larger conduit arteries in hypertensive patients. These results contrasted with dihydralazine which dilated arterioles, but constricted larger arteries [309].

Details on the mechanism of action and other pharmacological properties of diltiazem have been summarized [305]. Metabolism occurs primarily in the liver. Primary biotransformation entails demethylation of the terminal nitrogen, cleavage of the *p*-methoxy group, and oxidative hydroxylation of the aromatic rings. The 3-hydroxy compound is a major metabolite which exhibits some pharmacological activity [301], [310].

5.2. Vasodilators with Other Mechanisms of Action

Vasodilator drugs with a direct action on vascular smooth muscle were used in the therapy of hypertension long before the discovery of Ca^{2+} channel antagonists. Among vasodilator drugs hydralazine is the most commonly used drug, although its precise mechanism of action is still controversial. The common pharmacological properties of these drugs can be summarized as follows:

1) Their hypotensive activity is independent of the sympathetic tone.
2) They antagonize the vasoconstrictor effects of various endogenous substances including norepinephrine, serotonin, angiotensin II, and K^+.
3) Their site of action is at the membrane of vascular smooth muscle, at the contractile elements, or on the enzymatic processes involved in the delivery of energy for the contractile activity.
4) Their common side effects include tachycardia, salt and water retention, myocardial necrosis, and hyperglycemia.

5.2.1. Hydralazine

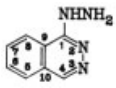

Hydralazine [*86-54-4*], 1-(2*H*)-phthalazinone hydrazone, $C_8H_8N_4$, M_r 160.18, *mp* 172–173 °C, is a yellow crystalline substance, soluble in acetic acid. Hydralazine has been available for the treatment of hypertension for about 30 years.

Trade names: Apresoline HCl, Hydralazine HCl; (combination with hydrochlorothiazide) Apresoline-Esidrix, Apresozide, Hydrap-ES; (with reserpine) Serpasil-Apresoline; (with hydrochlorothiazide plus reserpine) Ser-Ap-Es.

Hydralazine lowers arterial pressure in many experimental models of hypertension. Certain structure requirements are necessary for maximal activity. Among the more important requirements are the presence of a free NH_2 group and the placement of the hydrazino-group in the 1-position of the phthalazine moiety for optimal duration of action. The activity is affected minimally with the substitution of a carbon in the phthalazine or a pyridine replacement of the benzene ring [311].

Antihypertensive actions of hydralazine result primarily from arteriolar relaxation and a decreased peripheral resistance, with essentially no effects on venous tone. Heart rate and cardiac output are increased by hydralazine. The cardiotonic action results primarily from reflex adjustments to the decreased arterial pressure (baroreceptor reflex) [312] or quite possibly from elevation in central venous pressure (Bainbridge reflex) [313]. In addition, a direct cardiac stimulant effect has been observed in anesthetized dogs by some [314], but not by other investigators [313]. The direct and reflex actions of hydralazine are antagonized by propranolol. Cardiac hypertrophy occurring in SHR is not attenuated by hydralazine, whether given alone or in combination with propranolol [315]. Extensive pharmacological studies with hydralazine were published [316–328].

Clinically, the antihypertensive effect of hydralazine is enhanced by the addition of β-adrenoceptor antagonists, reserpine, guanethidine, methyldopa, clonidine, or diuretics [329], [330]. When given in combination with venodilators (e.g., isosorbide dinitrate or nitroglycerin) hydralazine has proved especially useful in the treatment of congestive heart failure [331].

5.2.2. Diazoxide

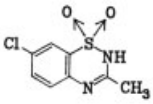

Diazoxide [*364-98-7*], 7-chloro-3-methyl-2*H*-1,2,4-benzothiadicazine-1,1-dioxide, $C_8H_7ClN_2O_2S$, M_r 230.70, *mp* 330 – 331 °C, is chemically related to chlorothiazide and other benzothiadiazine diuretics, but it exhibits no diuretic activity. It is an effective antihypertensive agent in spontaneous, DOCA, and renal hypertensive rat models [332], [333]. Antihypertensive doses of diazoxide, when given intravenously, range between 2 and 10 mg/kg in animal studies. The duration of action is dose-related and ranges between 0.5 – 4 h.

Trade name: Hyperstat I.V.

When given intravenously, diazoxide reduces total peripheral resistance and increases stroke volume and cardiac output [334].

In humans, heart rate and force of left ventricular contractions are increased by diazoxide, while left ventricular systolic and end-diastolic pressures are decreased. β-Adrenoreceptor blockade attenuates the elevations in heart rate and in left ventricular dp/dt, but augments the reduction in left ventricular systolic pressure [335]. The mechanism of action of diazoxide remains to be fully elucidated [336–340].

Hyperglycemia is observed during diazoxide therapy. This effect is attenuated by propranolol [341] and the ganglionic blocker chlorisondamine [342]. The propensity of diazoxide to cause hyperglycemia, and Na^+ and fluid retention has restricted its clinical use to acute life-threatening hypertensive emergencies. The dosage is titrated on the basis of individual responsiveness, but ranges between 100 and 300 mg, i.v. [343].

5.2.3. Minoxidil

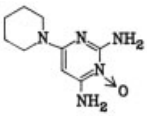

Minoxidil [*38304-91-5*], 6-(1-piperidinyl)-2,4-pyrimidinediamine-3-oxide, $C_9H_{15}N_5O$, M_r 209.25, *mp* 248 °C, 259 – 261 °C (decomp.), is a white crystalline solid. Minoxidil is primarily a dilator of arterial vasculature, thereby lowering total vascular resistance and blood pressure.

Trade name: Loniten.

Extensive literature on pharmacology of minoxidil exists [344–353].

Minoxidil is approved primarily for use in severely hypertensive patients who do not respond to maximal therapy with a diuretic or other antihypertensive agents, or in patients where side effects preclude other drug therapy. It is also useful in moderately hypertensive patients [354], those with impaired renal function [355], but is used rarely in hypertensive emergencies because of its availability in oral dosage form only and its long onset of action [345]. Minoxidil is combined with β-adrenoceptor blockade to reduce effects of reflex sympathetic activation. It is also effective when combined with other centrally acting sympatholytics, such as clonidine [356], or methyldopa, α-adrenoceptor antagonists, or neuronal blocking agents [357]. Captopril has been used as an adjunct to minoxidil in refractory hypertensive cases [358]. Because of sodium and water retention, a diuretic is given in almost all cases; however, this may not be necessary in patients with less severe hypertension [345].

The use of minoxidil is limited to a relatively small population of hypertensive patients because of the incidence of hypertrichosis, and potential for the occurrence of atrophic and ischemic lesions of the heart [359]. The hypertrichosis occurs in about 80 % of the patients treated. It occurs about 3 – 6 weeks after the onset of therapy and requires about 1 – 6 months for recovery after discontinuation of minoxidil. Cell culture experiments showed that minoxidil may produce the hypertrichosis by a direct action on epidermal cells [360].

5.2.4. Sodium Nitroprusside

Sodium nitroprusside [*14402-89-2*] is used as the dihydrate, $Na_2Fe(CN)_5NO \cdot 2\,H_2O$, M_r 216.91, readily soluble in water, but decomposes in light.

Trade names: Nipride, Nitropress.

Nitroprusside is not absorbed orally and thus must be given by systemic infusion. Vasodilatory effects are exerted on both the arterial and venous sides of the vasculature. Infusion of sodium nitroprusside at 1 to 100 $\mu g\, kg^{-1}\, min^{-1}$ produces a uniform fall in systolic and diastolic blood pressure. The onset of action is usually rapid (< 60 s) and the effect is well-maintained during the infusion. Recovery of blood pressure occurs rapidly following cessation of drug [361]. This most likely is the result of rapid cellular degradation of sodium nitroprusside. Because of this rapid recovery the magnitude of the hypotensive response can be controlled easily by varying the infusion rate. Tolerance to repeated sodium nitroprusside infusions occurs rarely [362]. Other pharmacological effects have been described [363–367].

The vasodilator activity of sodium nitroprusside is attributed to the NO-group (nitroso) [368–370]. The vasorelaxant effects do not directly involve α- or β-adrenoceptors and are quite fully manifested when vessels are contracted by angiotensin or vasopressin. Sodium nitroprusside may reduce the transsarcolemmal influx of calcium or enhance calcium efflux by stimulating $Na^+ - K^+$ ATPase, although quite disparate data have been found in both regards [371–373]. Elevations in cellular levels of cyclic GMP are intimately involved in the vasodilatory responses [369], [370].

A final metabolic product of sodium nitroprusside is thiocyanate. Toxicity during prolonged infusions is usually due to thiocyanate accumulation (i.e., hypoxia, nausea, tinnitus) and necessitates discontinuation of treatment.

6. Perspectives in the Development of Antihypertensive Drugs

A large number of new antihypertensive drugs have been introduced during the last 30 years. These drugs were usually justified by an improved side effect profile and/or by a more physiological mechanism of action. This trend is likely to continue during the next decade. Many new inhibitors of angiotensin converting enzyme have been introduced. The effectiveness of Ca^{2+} channel antagonists in the treatment of hypertension was recognized and novel inhibitors of Ca^{2+} channels are likely to be developed during the next decade. Drugs interfering with intracellular Ca^{2+} translocation, e.g., calmodulin inhibitors, also can be conceivably useful in the treatment of hypertension.

The central control of arterial pressure by drugs is receiving renewed attention. The recognition of the importance of various neurotransmitters, e.g., 5-HT, GABA, dopamine, in the central regulation of cardiovascular functions is likely to lead to novel antihypertensive drugs with central mechanism of action. Modification of the synthesis or release of various neuropeptides may also represent a future mechanism of antihypertensive action. The role of vasopressin in the cerebral as well as peripheral control of vascular tone has received renewed attention, and vasopressin antagonists can conceivably be useful in the treatment of hypertension.

The importance of the "natriuretic hormone" in the control of Na^+ levels in vascular smooth muscle cell and consequently of Ca^{2+} levels is now being recognized, and drugs controlling the release or the effects of "natriuretic hormone" may be useful in the treatment of hypertension.

7. Summary and Conclusions

In spite of the discovery of two new classes of antihypertensive drugs, inhibitors of the angiotensin converting enzyme and Ca^{2+} channel antagonists, the well-established antihypertensive drugs, such as diuretics and β-adrenoceptor antagonists, are still being widely used.

Diuretics which do not produce hypokalemia, hyperuricemia, hyperglycemia, or hyperlipoproteinemia, or diuretics with an additional direct vasodilator action are still in demand and may be introduced in the near future.

The novelty of the mechanism of action as well as the clinical efficacy and safety will help to expand the market for converting enzyme inhibitors. Their apparent usefulness in the treatment of heart failure will also contribute to the expansion of their market share.

Among drugs which interfere with the sympathetic nervous system, α_1-, β-, and α_1/β-adrenoceptor antagonists are widely used in the treatment of hypertension. The advantages of

α_1-adrenoceptor antagonists (e.g., prazosin) include increase in blood flow to vital organs and lowering of peripheral vascular resistance. βAdrenoceptor antagonists control arterial pressure with minimal side effects, and reduce mortality in patients with myocardial infarction. It remains to be seen whether new α_1/β-adrenoceptor antagonists (e.g., labetalol) will combine the benefits of both types of drugs.

With the discovery of the antihypertensive action of Ca^{2+} channel antagonists the interest in the direct acting smooth muscle relaxants has been renewed and drugs which interfere with vascular smooth muscle tone and affect intracellular calcium translocation are likely to be introduced in the near future.

8. References

1. E. J. Cragoe (ed.): *Diuretics: Chemistry, Pharmacology and Medicine,* J. Wiley & Sons, New York 1983.
2. T. H. Maren, *Trans. N.Y. Acad. Sci.* **15** (1952) 53.
3. F. C. Novello, J. M. Sprague, *J. Am. Chem. Soc.* **79** (1957) 2028 – 2029.
4. K. H. Beyer, *Ann. N.Y. Acad. Sci.* **71** (1958) 363 – 379.
5. E. D. Fries, A. Wanko, I. M. Wilson, A. E. Parrish, *JAMA* **166** (1958) 137 – 140.
6. J. E. Baer, C. B. Jones, S. A. Spitzer, H. F. Russo, *J. Pharmacol. Exp. Ther.* **157** (1967) 472 – 485.
7. V. D. Wiebelhaus, F. T. Brennan, G. Sosnowski et al., *Arch. Int. Pharmacodyn. Ther.* **169** (1967) 429 – 457.
8. A. C. Guyton, R. D. Manning, J. E. Hall, R. A. Norman et al., *J. Cardiovasc. Pharmacol.* **6** (1984) S 151 – S 161.
9. M. P. Blaustein, *Am. J. Physiol.* **232** (1977) C 165 – C 173.
10. L. Poston, H. H. Gray, A. Crowther, H. C. Dittrich et al., *J. Cardiovasc. Pharmacol.* **6** (1984) S 16 – S 20.
11. M. B. Strauss, H. Southworth, *Bull. John Hopkins Hosp.* **63** (1938) 41 – 45.
12. R. Muschaweck, P. Hajdú, *Arzneim.-Forsch.* **14** (1964) 44 – 47.
13. K. H. Beyer, J. E. Baer, J. K. Michaelson, H. F. Russo, *J. Pharmacol. Exp. Ther.* **147** (1965) 1 – 22.
14. G. de Stevens, L. H. Weiner, A. Halamandaris, S. Ricca, *Experientia* **14** (1958) 463.
15. W. E. Barrett, J. J. Chart, A. A. Renzi, *Arch. Int. Pharmacodyn. Ther.* **131** (1961) 325 – 338.
16. R. V. Ford, *Curr. Ther. Res.* **2** (1960) 422 – 429.
17. A. Scriabine, B. Korol, B. Kondratas, M. Yu et al., *Proc. Soc. Exp. Biol. Med.* **107** (1961) 864 – 872.
18. H. Horstmann, H. Wollweber, K. Meng, *Arzneim.-Forsch.* **17** (1967) 653 – 659.
19. K. Meng, G. Kroneberg, *Arzneim.-Forsch.* **17** (1967) 659 – 671.
20. R. Rutsaert, M. Fernandes in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 1,** Raven Press, New York 1983, pp. 49 – 67.
21. E. G. Stenger, H. Wirz, R. Pulver, *Schweiz. Med. Wochenschr.* **89** (1959) 1126 – 1130.
22. C. M. Kagawa, F. M. Sturtevant, C. G. Van Arman, *J. Pharmacol. Exp. Ther.* **126** (1959) 123 – 130.
23. L. S. Watson, G. M. Fanelli, H. F. Russo, C. S. Sweet et al. in A. Scriabine, C. S. Sweet (ed.): *New Antihypertensive Drugs,* Spectrum Publications, Inc., New York 1976, pp. 307 – 321.
24. G. M. Fanelli, Jr., D. L. Bohn, A. Scriabine, K. H. Beyer, *J. Pharmacol. Exp. Ther.* **200** (1977) 402 – 412.
25. W. B. Abrams, J. D. Irvin, J. A. Tobert, R. K. Ferguson et al. in A. Scriabine (ed.): *New Cardiovascular Drugs,* Raven Press, New York, 1985, pp. 1 – 20.
26. E. Moller, H. Horstmann, K. Meng, D. Loew, *Experientia* **33** (1977) 382 – 383.
27. B. Garthoff, M. Hanisch in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 2,** Raven Press, New York 1984, pp. 51 – 70.
28. C. S. Sweet, E. H. Blaine in M. J. Antonaccio (ed.): *Cardiovascular Pharmacology,* 2nd ed., Raven Press, New York 1984, pp. 119 – 154.
29. M. A. Ondetti, B. Rubin, D. W. Cushman, *Science* **196** (1977) 441 – 444.
30. D. W. Cushman, M. A. Ondetti, *Prog. Med. Chem.* **17** (1980) 41 – 104.
31. B. Rubin, R. J. Laffan, D. G. Kotler, E. H. O'Keefe et al., *J. Pharmacol. Exp. Ther.* **204** (1978) 271 – 280.
32. R. J. Laffan, M. E. Goldberg, J. P. High, T. R. Schaeffer et al., *J. Pharmacol. Exp. Ther.* **204** (1978) 281 – 288.
33. R. J. Bengis, T. G. Coleman, D. B. Young, R. E. McCaa, *Circ. Res.* **43** (1978) Suppl. 1, I 45 – I 53.

34. C. S. Sweet, P. T. Arbegast, S. L. Gaul, E. H. Blaine et al., *Eur. J. Pharmacol.* **76** (1981) 167 – 176.
35. M. L. Cohen, K. D. Kurz, *J. Pharmacol. Exp. Ther.* **220** (1982) 63 – 69.
36. T. Unger, D. Ganten, R. E. Lang, *Trends Pharmacol. Sci.* **4** (1983) 514 – 519.
37. M. J. Antonaccio, L. Kerwin, *Hypertension* **3** (1981) Suppl. 1, I 54 – I 62.
38. Y. Imai, K. Abe, M. Seino, T. Haruyama et al., *Hypertension* **4** (1982) 444 – 451.
39. R. Hatton, D. P. Clough, *J. Cardiovasc. Pharmacol.* **4** (1982) 116 – 123.
40. A. DeJonge, B. Wilffert, H. O. Kalkman, J. C. A. van Meel et al., *Eur. J. Pharmacol.* **74** (1981) 385 – 386.
41. C. Richer, M. P. Doussau, J. F. Giudicelli, *Eur. Heart J.* **4** (1983) Suppl. G, 55 – 60.
42. T. Unger, I. Kaufmann-Bühler, B. Schölkens, D. Ganten, *Eur. J. Pharmacol.* **70** (1981) 467 – 478.
43. R. E. McCaa, J. E. Hall, C. S. McCaa, *Circ. Res.* **43** (1978) Suppl. 1, I 32 – I 39.
44. V. S. Murthy, T. L. Waldron, M. E. Goldberg, *Circ. Res.* **43** (1978) Suppl. 1, I 40 – I 45.
45. K. Ito, H. Koike, M. Miyamoto, H. Ozaki et al., *J. Pharmacol. Exp. Ther.* **219** (1981) 520 – 525.
46. B. Rubin, M. J. Antonaccio in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 21 – 42.
47. E. E. Muirhead, R. L. Prewitt, B. Brooks, W. L. Brosius, *Circ. Res.* **43** (1978) Suppl. 1, I 53 – I 59.
48. D. B. Case, S. A. Atlas, P. A. Sullivan, J. H. Laragh, *Circulation* **64** (1981) 765 – 771.
49. W. B. Lee, M. J. Ismay, E. R. Lumbers, *Circ. Res.* **47** (1980) 286 – 292.
50. W. H. Van Gilst, P. A. DeGraeff, J. H. Kingma, H. Wesseling et al., *Eur. J. Pharmacol.* **100** (1984) 113 – 117.
51. B. G. Zimmerman, P. C. Wong, G. K. Kounenis, E. J. Kraft, *Am. J. Physiol.* **243** (1982) H 277 – H 283.
52. S. Chiba, C. P. Quilley, J. Quilley, J. C. McGiff, *Eur. J. Pharmacol.* **83** (1982) 243 – 252.
53. T. Taylor, T. J. Moore, N. K. Hollenberg, G. H. Williams, *Hypertension* **6** (1984) 92 – 99.
54. W. B. Abrams, R. O. Davies, R. K. Ferguson, *Fed. Proc.* **43** (1984) 1314 – 1321.
55. C. S. Sweet, E. H. Ulm in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 2,** Raven Press, New York 1984, pp. 1 – 17.
56. D. M. Gross, C. S. Sweet, E. H. Ulm, E. P. Backlund et al., *J. Pharmacol. Exp. Ther.* **216** (1981) 552 – 557.
57. C. S. Sweet, D. M. Gross, P. T. Arbegast, S. L. Gaul et al., *J. Pharmacol. Exp. Ther.* **216** (1981) 558 – 566.
58. J. Biollaz, M. Burnier, G. A. Turini, B. B. Brunner et al., *Clin. Pharmacol. Ther.* **29** (1981) 665 – 670.
59. R. K. Ferguson, P. H. Vlasses, J. D. Irvin, B. N. Swanson et al. *J. Clin. Pharmacol.* **22** (1982) 281 – 289.
60. H. Gavras, B. Waeber, I. Gavras, J. Biollaz et al., *Lancet* **2** (1981) 543 – 547.
61. H. Gavras, J. Biollaz, B. Waeber, H. R. Brunner et al., *Clin. Exp. Hypertens.* **4 (1), (2)** (1982) 303 – 314.
62. F. G. Dunn, W. Oigman, H. O. Ventura, F. H. Messerli et al., *Am. J. Cardiol.* **53** (1984) 105 – 108.
63. P. W. deLeeuw, R. P. L. M. Hoogma, G. A. W. van Soest, P. T. Tchang et al., *J. Cardiovasc. Pharmacol.* **5** (1983) 731 – 736.
64. S. Morioka, G. Simon, J. N. Cohn, *Clin. Pharmacol. Ther.* **34** (1983) 583 – 589.
65. C. Richer, M. P. Doussau, J. F. Giudicelli, *Eur. J. Pharmacol.* **79** (1982) 23 – 29.
66. D. M. Shoback, G. H. Williams, S. L. Swartz, R. O. Davies et al., *J. Cardiovasc. Pharmacol.* **5** (1983) 1010 – 1018.
67. O. Kohlmann, M. Bresnahan, H. Gavras, *Hypertension* **6** (1984) Suppl. 1, I 1 – I 6.
68. A. Scriabine, D. G. Taylor, E. Hong, *Prog. Drug Res.* **26** (1982) 353 – 371.
69. A. Scriabine, B. V. Clineschmidt, C. S. Sweet, *Annu. Rev. Pharmacol. Toxicol.* **16** (1976) 113 – 123.
70. T. L. Sourkes, *Arch. Biochem. Biophys.* **51** (1954) 444 – 456.
71. L. Gillespie, J. A. Oates, J. R. Crout, A. Sjoerdsma, *Circulation* **25** (1962) 281 – 291.
72. C. C. Porter, M. L. Torchiana, C. A. Stone in F. Gross (ed.): *Antihypertensive Agents,* Springer-Verlag, Berlin-Heidelberg-New York 1977, pp. 269 –297.
73. M. Henning, P. A. van Zwieten, *J. Pharm. Pharmacol.* **20** (1968) 409 – 417.
74. A. Scriabine in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 43 – 54.
75. G. Onesti, A. N. Brest, P. Novack, H. Kasparian et al., *Am. Heart J.* **67** (1964) 32 – 38.
76. S. Mohammed, T. E. Gaffney, A. C. Yard, H. Gomez, *J. Pharmacol. Exp. Ther.* **160** (1968) 300 – 307.

77. A. Cohen, J. S. Maxmen, M. Ragheb, H. Baleiron et al., *J. Clin. Pharmacol.* **7** (1967) 77 – 83.
78. J. S. Meyer, T. Sawada, A. Kitamura, M. Toyoda, *Neurology (N.Y.)* **18** (1968) 772 – 781.
79. A. Sjöerdsma, A. Vendsalu, K. Engelman, *Circulation* **28** (1963) 492 – 502.
80. L. F. Prescott, R. P. Buhs, J. O. Beattie, O. C. Speth et al., *Circulation* **34** (1966) 308 – 321.
81. S. Z. Langer, *Biochem. Pharmacol.* **23** (1974) 1793 – 1800.
82. Y. Pommier, M. Andréjak, P. Mouillé, H. Dabiré et al., *Naunyn-Schmiedebergs Arch. Pharmacol.* **318** (1982) 288 – 294.
83. J. R. Boissier, J. F. Giudicelli, J. Fichelle, H. Schmitt et al. *Eur. J. Pharmacol.* **2** (1968) 333 – 339.
84. W. Hoefke in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 55 – 78.
85. H. Schmitt in F. Gross (ed.): *Antihypertensive Agents,* Springer-Verlag, Berlin-Heidelberg-New York 1977, pp. 299 – 396.
86. T. Hukuhara, Jr., Y. Otsuka, R. Takeda, F. Sakai, *Arzneim.-Forsch.* **18** (1968) 1147 – 1153.
87. H. Schmitt, H. Schmitt, *Eur. J. Pharmacol.* **6** (1969) 8 – 12.
88. W. Hoefke, W. Kobinger, *Arzneim.-Forsch.* **16** (1966) 1038 – 1050.
89. G. A. Bentley, I. W. Copeland, J. Starr, *Clin. Exp. Pharmacol. Physiol.* **4** (1977) 405 – 419.
90. M. S. Gold, D. E. Redmond, Jr., H. D. Kleber, *Lancet* **2** (1978) 599 – 602.
91. R. C. Allen, M. E. Langham, *Invest. Ophthalmol.* **15** (1976) 815 – 823.
92. H. O. Karppanen, E. Westermann, *Naunyn-Schmiedebergs Arch. Pharmacol.* **279** (1973) 83 – 87.
93. W. Kobinger, *Naunyn-Schmiedebergs Arch. Pharmacol.* **260** (1968) 153 – 154.
94. D. S. Davis, E. Neill, J. L. Reid in T. A. Baillie (ed.): *Stable Isotopes,* MacMillan, London 1978, pp. 45 – 54.
95. M. A. Weber, *J. Cardiovasc. Pharmacol.* **2** (1980) Suppl. 1, S 73 – S 89.
96. M. J. M. C. Thoolen, P. B. M. W. M. Timmermans, P. A. van Zwieten, *Naunyn-Schmiedebergs Arch. Pharmacol.* **319** (1982) 82 – 86.
97. J. B. Bream, H. Lauener, C. W. Picard, G. Scholtysik et al., *Arzneim.-Forsch.* **25** (1975) 1477 – 1482.
98. G. Scholtysik, H. Lauener, E. Eichenberger, E. Bürki et al., *Arzneim.-Forsch.* **25** (1975) 1483 – 1491.
99. P. A. van Zwieten, *Prog. Pharmacol.* **1** (1975) 1 – 63.
100. G. Scholtysik, P. Jerie, C. W. Picard in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 79 – 98.
101. J. C. Doxey, *Eur. J. Pharmacol.* **54** (1979) 185 – 189.
102. A. Jaattela, *Eur. J. Clin. Pharmacol.* **10** (1976) 73 – 76.
103. A. J. Bune, J. P. Chalmers, J. R. Graham, P. R. C. Howe et al., *Eur. J. Clin. Pharmacol.* **19** (1981) 309 – 315.
104. I. L. Natoff, B. G. Katzung, F. Weir, J. K. Kodama, *Pharmacologist* **10** (1968) 157.
105. T. Baum, *Fed. Proc.* **28** (1969) 738.
106. P. Bolme, H. Corrodi, K. Fuxe, *Eur. J. Pharmacol.* **23** (1973) 175 – 182.
107. E. Marmo, R. K. Saini, A. P. Caputi, C. Vaca, *Res. Commun. Chem. Pathol. Pharmacol.* **6** (1973) 391 – 406.
108. T. Baum, G. Rowles, A. T. Shropshire, *Arch. Int. Pharmacodyn. Ther.* **183** (1970) 75 – 84.
109. P. Bosanac, J. Dubb, B. Walker, M. Goldberg et al., *J. Clin. Pharmacol.* **16** (1976) 631 – 636.
110. R. L. Wendt in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 99 – 111.
111. J. W. Strandhoy, B. D. Steg, V. M. Buckalew, Jr., *Life Sci.* **27** (1980) 2513 – 2518.
112. R. H. Meacham, M. Emmett, A. A. Kyriakopoulos, S. T. Chiang et al., *Clin. Pharmacol. Ther.* **27** (1980) 44 – 52.
113. F. G. McMahon, J. R. Ryan, A. K. Jain, R. Vargas et al., *Clin. Pharmacol. Ther.* **21** (1977) 272 – 277.
114. B. R. Walker, M. Geczy, J. A. Gold, *Clin. Pharmacol. Ther.* **25** (1979) 252.
115. L. Dorfman, A. Furlenmeier, C. F. Huebner, R. Lucas et al., *Hel. Chim. Acta* **37** (1954) 59 – 75.
116. R. B. Woodward, F. E. Bader, H. Bickel, A. J. Frey et al., *J. Am. Chem. Soc.* **78** (1956) 2023 – 2025.
117. R. Sannerstedt, J. Conway, *Am. Heart J.* **79** (1970) 122 – 127.
118. A. Scriabine in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 119 – 125.
119. M. J. Rand, M. Jurevics in F. Gross (ed.): *Antihypertensive Agents,* Springer-Verlag, Berlin-Heidelberg-New York 1977, pp. 77 – 159.

120. R. E. Stitzel, *Pharmacol. Rev.* **28** (1976) 179 – 208.
121. H. Sheppard, R. C. Lucas, W. H. Tsien, *Arch. Int. Pharmacodyn. Ther.* **103** (1955) 256 – 269.
122. R. A. Maxwell, A. J. Plummer, F. Schneider, H. Povalski et al., *J. Pharmacol. Exp. Ther.* **128** (1960) 22 – 29.
123. R. A. Maxwell in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 127 – 150.
124. R. A. Maxwell, W. B. Wastila in F. Gross (ed.): *Antihypertensive Agents,* Springer-Verlag, Berlin-Heidelberg-New York 1977, pp. 161 – 261.
125. B. B. Brodie, C. C. Chang, E. Costa, *Br. J. Pharmacol.* **25** (1965) 171 – 178.
126. F. M. Abboud, J. W. Eckstein, *Circ. Res.* **11** (1962) 788 – 796.
127. C. A. Stone, C. C. Porter, J. M. Stavorski, C. T. Ludden et al., *J. Pharmacol. Exp. Ther.* **144** (1964) 196 – 204.
128. A. N. Brest, *JAMA* **192** (1965) 41 – 44.
129. J. N. Cohn, T. E. Liptak, E. D. Freis, *Circ. Res.* **12** (1963) 298 – 307.
130. H. Villarreal, J. E. Exaire, V. Rubio, H. Dávila, *Am. J. Cardiol.* **14** (1964) 633 – 640.
131. C. T. Dollery, D. Emslie-Smith, M. D. Milne, *Lancet* **2** (1960) 381 – 387.
132. F. B. Abramson, C. I. Furst, C. McMartin, R. Wade, *Biochem. J.* **113** (1969) 143 – 156.
133. C. McMartin, *Biochem. Pharmacol.* **18** (1969) 238 – 243.
134. M. Nickerson, *Pharmacol. Rev.* **1** (1949) 27 – 101.
135. A. Scriabine, J. W. Constantine, H.-J. Hess, W. K. McShane, *Experientia* **24** (1968) 1150 – 1151.
136. J. W. Constantine, W. K. McShane, A. Scriabine, H.-J. Hess in G. Onesti, K. E. Kim, J. H. Moyer (ed.): *Hypertension, Mechanism and Management,* Grune & Stratton, Inc., Baltimore 1973, pp. 429 – 444.
137. A. Scriabine in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 151 – 160.
138. I. Cavero, A. G. Roach, *Life Sci.* **27** (1980) 1525 – 1540.
139. J. C. Schulz, D. P. Westfall, *Blood Vessels* **19** (1982) 79 – 87.
140. G. von der Lippe, O. J. Ohm, P. Lund-Johansen, *Acta Med. Scand.* **210** (1981) 213 – 216.
141. J. M. Hayes, R. M. Graham, B. P. O'Connell, E. Speers et al., *Med. J. Aust.* **1** (1976) 562 – 564.
142. P. Bolli, A. J. Wood, W. L. Phelan, D. R. Lee et al., *Clin. Sci. Mol. Med.* **48** (1975) 177 S – 179 S.
143. I. S. Collins, P. Pek, *Clin. Exp. Pharmacol. Physiol.* **2** (1975) 445 – 446.
144. T. H. Althuis, H.-J. Hess, *J. Med. Chem.* **20** (1977) 146 – 149.
145. J. A. Taylor, T. M. Twomey, M. S. von Wittenau, *Xenobiotica.* **7** (1977) 357 – 364.
146. D. S. Bloom, C. Rosendorff, R. Kramer, *Curr. Ther. Res.* **18** (1975) 144 – 150.
147. A. S. P. Hua, I. M. MacDonald, J. B. Meyers, P. Fang et al., *Med. J. Aust.* **2** (1977) 5 – 9.
148. W. M. Kirkendall, J. J. Hammond, J. C. Thomas, M. L. Overturf et al., *JAMA* **240** (1978) 2553 – 2556.
149. Y. K. Seedat, M. A. Seedat, R. Bhoola, *S. Afr. Med. J.* **51** (1977) 461 – 463.
150. A. J. Marshall, D. W. Barritt, J. Pocock, S. T. Heaton, *Lancet* **1** (1977) 271 – 274.
151. C. Rosendorff, *Br. Med. J.* **2** (1976) 508.
152. J. L. Archibald, *Chim. Ther. (Paris)* **3** (1968) 397.
153. J. L. Archibald, B. J. Alps, J. F. Cavalla, J. L. Jackson, *J. Med. Chem.* **14** (1971) 1054 – 1059.
154. J. L. Archibald in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 161 – 177.
155. B. J. Alps, E. T. Borrows, E. S. Johnson, M. W. Staniforth et al., *Cardiovasc. Res.* **6** (1972) 226 – 234.
156. T. Baum, A. T. Shropshire, D. K. Eckfeld, N. Metz et al., *Arch. Int. Pharmacodyn. Ther.* **204** (1973) 390 – 406.
157. T. Baum, A. T. Shropshire, *Eur. J. Pharmacol.* **32** (1975) 30 – 38.
158. B. J. Alps, M. Hill, K. Fidler, E. S. Johnson et al., *J. Pharm. Pharmacol.* **23** (1971) 678 – 686.
159. B. J. Alps, M. Hill, E. S. Johnson, A. B. Wilson, *Br. J. Pharmacol.* **40** (1970) 153 P.
160. G. H. Draffan, P. J. Lewis, J. L. Firmin, T. W. Jordan et al., *Br. J. Clin. Pharmacol.* **3** (1976) 489 – 495.
161. G. H. Draffan, *Pharmacol. Clin.* **1** (1973) 51 – 59.
162. R. B. Royds, D. J. Coltart, J. D. F. Lockhart, *Clin. Pharmacol. Ther.* **13** (1972) 380 – 392.
163. C. Rosendorff, *S. Afr. Med. J.* **50** (1976) 764 – 768.
164. C. T. Dollery, D. G. George, P. J. Lewis, *Br. J. Pharmacol.* **46** (1972) 542 P – 543 P.

165. K. Klemm, W. Prüsse, U. Krüger, *Arzneim.-Forsch.* **27** (1977) 1895 – 1897.
166. W. Schoetensack, P. Bischler, E. C. Dittmann, V. Steinijans, *Arzneim.-Forsch.* **27** (1977) 1908 – 1919.
167. W. Schoetensack, E. G. Bruckschen, K. Zech in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 1,** Raven Press, New York 1983, pp. 19 – 48.
168. K. Zech in W. Kaufman, E. G. Bruckschen (ed.): *Proc. 1st Symposium on Urapidil,* Excerpta Medica, Amsterdam-Oxford-Princeton 1982, pp. 50 – 64.
169. E. G. Bruckschen, F. Henze, G. Michael, *Arzneim.-Forsch.* **28 (2)** (1978) 1176 – 1184.
170. B. N. C. Prichard, *Br. Med. J.* **1** (1964) 1227 – 1228.
171. A. Scriabine, *Ann. Rev. Pharmacol. Toxicol.* **19** (1979) 269 – 284.
172. R. Howe, V. A. Cullum, *Nature (London)* **210** (1966) 1336 – 1338.
173. A. M. Barrett, R. G. Shanks, *Br. J. Pharmacol.* **34** (1968) 43 – 55.
174. J. D. Fitzgerald in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 195 – 208.
175. A. S. Nies, D. G. Shand, *Circulation* **52** (1975) 6 – 15.
176. D. G. Shand, E. M. Nuckolls, J. A. Oates, *Clin. Pharmacol. Ther.* **11** (1970) 112 – 120.
177. A. Hayes, R. G. Cooper, *J. Pharmacol. Exp. Ther.* **176** (1971) 302 – 311.
178. T. Walle, E. C. Conradi, U. K. Walle, T. C. Fagan, *Clin. Pharmacol. Ther.* **27** (1980) 22 – 31.
179. L. Hansson, R. Olander, H. Aberg, H. Malmcroma et al., *Acta Med. Scand.* **190** (1971) 531 – 534.
180. A. P. Douglas-Jones, *J. Int. Med. Res.* **7** (1979) 221 – 223.
181. B. Åblad, K.-O. Borg, E. Carlsson, L. Ek et al., *Acta Pharmacol. Toxicol.* **36** (1975) 7 – 24.
182. E. Carlsson, B. Åblad, A. Brändström, B. Carlsson, *Life Sci.* **11** (1972) 953 – 958.
183. B. Åblad, E. Carlsson, G. Johnsson, C.-G. Regårdh in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 247 –262.
184. K.-O. Borg, E. Carlsson, K.-J. Hoffmann, K.-J. Jönsson et al., *Acta Pharmacol. Toxicol.* **36** (1975) Suppl. 5, 125 – 135.
185. C.-G. Regårdh, G. Johnsson, L. Jordö, L. Solvell, *Acta Pharmacol. Toxicol.* **36** (1975) Suppl. 5, 45 – 58.
186. O. Lederballe Pedersen, *Eur. J. Clin. Pharmacol.* **10** (1976) 381 – 385.
187. Å. Hjalmarsson, D. Elinfeldt, J. Herlitz, S. Holmberg et al. *Lancet* **2** (1981) 823 – 827.
188. R. J. Lee, D. B. Evans, S. H. Baky, R. J. Laffan, *Eur. J. Pharmacol.* **33** (1975) 371 – 382.
189. D. B. Evans, M. T. Peschka, R. J. Lee, R. J. Laffan, *Eur. J. Pharmacol.* **35** (1976) 17 – 27.
190. M. J. Antonaccio, D. B. Evans in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 295 – 301.
191. R. A. Vukovich, J. E. Foley, B. Brown, D. A. Willard et al., *Br. J. Clin. Pharmacol.* **7** (2) (1979) 167 – 172.
192. G. Frithz, *Curr. Med. Res. Opin.* **5** (1978) 383 – 387.
193. B. Furberg, A. Dahlqvist, A. Raak, U. Wrege et al., *Curr. Med. Res. Opin.* **5** (1978) 388 – 393.
194. R. A. Vukovich, A. Sasahara, P. Zombrano, J. Belko et al., *Clin. Pharmacol. Ther.* **19** (1976) 118.
195. K. Saameli in H. J. Dengler (ed.): *Die Therapeutische Anwendung β* -sympathikolytischer Stoffe, F. K. Schattauer Verlag, Stuttgart-New York 1972, pp. 3 – 30.
196. C. Weil in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 237 – 246.
197. J. Meier, *Curr. Med. Res. Opin.* **4 Suppl. 5,** (1977) 31 – 38.
198. R. Gugler, W. Herold, H. J. Dengler, *Eur. J. Clin. Pharmacol.* **7** (1974) 17 – 24.
199. M. C. Laver, P. Fang, P. Kincaid-Smith, *Med. J. Aust.* **1** (1974) 174 – 176.
200. A. E. Doyle, *Drugs* **8** (1974) 422 – 431.
201. A. S. Turner, J. S. Peel, *Med. J. Aust.* **2** (1972) Spec. Suppl., 48 – 51.
202. J. Rosenthal, H. Kaiser, A. Raschig, D. Welzel, *Br. J. Clin. Pract.* **33** (1979) 165 – 174.
203. A. Scriabine, M. L. Torchiana, J. M. Stavorski, C. T. Ludden et al., *Arch Int. Pharmacodyn. Ther.* **205** (1973) 76 – 93.
204. R. A. Hall, R. O. Robson, N. N. Share, *Arch. Int. Pharmacodyn. Ther.* **213** (1975), pp. 251 – 263.
205. A. Scriabine in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 275 – 282.
206. D. J. Tocco, A. E. W. Duncan, F. A. De Lauwa et al., *Drug Metab. Dispos.* **3** (1975) 361 – 370.
207. G. Lohmoller, E. D. Frohlich, *Am. Heart J.* **89** (1975) 437 – 442.

208. The Norwegian Multicenter Study Group, *N. Engl. J. Med.* **304** (1981) 801 – 807.
209. K. R. H. Wooldridge, *Experienta* **28** (1972) 1404.
210. B. Basil, R. Jordan, A. H. Loveless, D. R. Maxwell, *Br. J. Pharmacol.* **50** (1974) 323 – 333.
211. R. H. Briant, C. T. Dollery, C. F. George, *Clin. Trials J.* **11** (1974) Suppl. 3, 25 – 28.
212. B. Levy, *J. Pharmacol. Exp. Ther.* **186** (1973) 134 – 144.
213. B. Basil, R. Jordan in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 283 – 294.
214. J. P. Amlie, S. Lessum, R. F. Collins, K. Landmark, *Acta Pharmacol. Toxicol.* **40** (1977) 378 – 388.
215. D. R. Maxwell, R. F. Collins, *Clin. Trials J.* **11** (1974) Suppl. 3, 9 – 18.
216. W. L. Ashton, *Curr. Med. Res. Opin.* **4** (1976) 442 – 454.
217. M. Lucsko, J. Guedon, *Clin. Trials J.* **11** (1974) Suppl. 3, 86 – 91.
218. L. Hansson, G. Berglund, O. Andersson, H. Holm, *Eur. J. Clin. Pharmacol.* **12** (1977) 89 – 92.
219. B. Carey, E. T. Whalley, *J. Pharm. Pharmacol.* **31** (1979) 791 – 792.
220. J. B. Farmer, I. Kennedy, G. P. Levy, R. J. Marshall, *Br. J. Pharmacol.* **45** (1972) 660 – 675.
221. G. P. Levy, D. A. Richards in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 325 – 347.
222. D. A. Richards, J. G. Maconochie, R. E. Bland, R. Hopkins et al., *Eur. J. Clin. Pharmacol.* **11** (1977) 85 – 90.
223. L. E. Martin, R. Hopkins, R. E. Bland, *Br. J. Clin. Pharmacol.* **3** (1976) Suppl. 1, 695 – 710.
224. B. N. C. Prichard, A. J. Boakes, *Br. J. Clin. Pharmacol.* **3** (1976) Suppl., 743 – 750.
225. P. Bolli, H. J. Waal-Manning, A. J. Wood, F. O. Simpson, *Br. J. Clin. Pharmacol.* **3 (Suppl.)** (1976) 765 – 771.
226. H. J. Dargie, C. T. Dollery, J. Daniel, *Br. J. Clin. Pharmacol.* **3** (1976) Suppl., 751 – 755.
227. A. Fleckenstein, *Ann. Rev. Pharmacol. Toxicol.* **17** (1977) 149 – 166.
228. T. B. Bolton, *Physiol. Rev.* **59** (1979) 606 – 718.
229. C. Cauvin, R. Loutzenhiser, C. Van Breemen, *Ann. Rev. Pharmacol. Toxicol.* **23** (1983) 373 – 396.
230. R. A. Janis, D. J. Triggle, *J. Med. Chem.* **26** (1983) 775 – 785.
231. R. A. Janis, A. Scriabine, *Biochem. Pharmacol.* **32** (1983) 3499 – 3507.
232. D. J. Triggle in S. F. Flaim, R. Zelis (ed.): *Calcium Blockers: Mechanisms of Action and Clinical Applications,* Urban & Schwarzenberg, Baltimore 1982, pp. 121 – 134.
233. R. G. Rahwan, D. T. Witiak, W. M. Muir, *Annu. Rep. Med. Chem.* **16** (1981) 257 – 268.
234. H. Meyer in L. H. Opie (ed.): *Calcium Antagonists and Cardiovascular Disease,* Raven Press, New York 1984, pp. 165 – 173.
235. H. Meyer, F. Bossert, E. Wehinger, K. Stoepel, W. Vater, *Arzneim.-Forsch.* **31** (1981) 407 – 409.
236. R. Towart, E. Wehinger, H. Meyer, *Naunyn-Schmiedebergs Arch. Pharmacol.* **317** (1981) 183 – 185.
237. R. Rodenkirchen, R. Bayer, R. Steiner, F. Bossert et al., *Naunyn-Schmiedebergs Arch. Pharmacol.* **310** (1979) 69 – 78.
238. M. Hiwatari, K. Satoh, N. Taira, *Arzneim.-Forsch.* **29** (1979) 256 – 260.
239. B. Garthoff, S. Kazda, *Eur. J. Pharmacol.* **74** (1981) 111 – 112.
240. K. Aoki, S. Kondo, A. Mochizuki, T. Yoshida et al., *Am. Heart J.* **96** (1978) 218 – 226.
241. M. T. Olivari, C. Bartorelli, A. Polese, C. Fiorentini et al., *Circulation* **59** (1979) 1056 – 1062.
242. M. D. Guazzi, C. Fiorentini, M. T. Olivari, A. Bartorelli et al., *Circulation* **61** (1980) 913 – 919.
243. M. D. Guazzi, A. Polese, C. Fiorentini, A. Bartorelli et al. *Hypertension* **5** (1983) Suppl. 2, I 85 – I 90.
244. O. Lederballe Pedersen, N. J. Christensen, K. D. Rämsch, *J. Cardiovasc. Pharmacol.* **2** (1980) 357 – 366.
245. M. Guazzi, M. T. Olivari, A. Polese, C. Fiorentini et al., *Clin. Pharmacol. Ther.* **22** (1977) 528 – 532.
246. O. Lederballe Pedersen, *Arch. Int. Pharmacodyn. Ther.* **239** (1979) 208 – 220.
247. M. J. Mulvany, N. Nyborg, *Br. J. Pharmacol.* **71** (1980) 585 – 596.
248. W. Vater, G. Kroneberg, F. Hoffmeister, H. Saller et al., *Arzneim.-Forsch.* **22** (1972) 1 – 14.
249. B. F. Robinson, R. J. Dobbs, C. R. Kelsey, *Br. J. Clin. Pharmacol.* **10** (1980) 433 – 438.
250. S. Kazda, B. Garthoff, G. Luckhaus, *Postgrad. Med. J.* **59** (1983) Suppl. 2, 78 – 83.

251. Y. C. Streifler, C. Wittenberg, J. Rosenfeld in M. Kaltenbach, H. N. Neufeld (ed.): *Fifth International Adalat Symposium,* Excerpta Medica, Amsterdam-Oxford-Princeton 1983, pp. 164 – 174.
252. S. Kazda, B. Garthoff, G. Thomas, *Drug Develop. Res.* **2** (1982) 313 – 323.
253. S. Kazda, A. Knorr, R. Towart, *Prog. Pharmacol.* **5** (1983) 83 – 116.
254. R. A. B. McLeay, T. J. Stallard, R. D. Watson, W. A. Littler, *Circulation* **67** (1983) 1084 – 1090.
255. L.-G. Ekelund, C. Ekelund, S. Rössner, *Acta. Med. Scand.* **212** (1982) 71 – 75.
256. W. Kiowski, O. Bertel, P. Erne, P. Bolli et al., *Hypertension* **5** (1983) Suppl. 1, I 70 – I 74.
257. A. Zanchetti, in A. Zanchetti, D. M. Krikler (ed.): *Calcium Antagonism in Cardiovascular Therapy: Experience with Verapamil,* Excerpta Medica, Amsterdam-Oxford-Princeton 1981, pp. 292 – 297.
258. G. Leonetti, C. Cuspidi, L. Sampieri, L. Terzoli et al. *J. Cardiovasc. Pharmacol.* **4** (1982) Suppl. 3, S 319 –S 324.
259. F. A. Horster in W. Lochner, W. Braasch, G. Kroneberg (ed.): *Second International Adalat Symposium,* Springer-Verlag, New York 1975, pp. 124 – 127.
260. K. Schlossmann, H. H. Medenwald, H. Rosenkranz in W. Lochner, W. Braasch, G. Kroneberg (ed.): *Second International Adalat Symposium,* Springer-Verlag, New York 1975, pp. 33 – 39.
261. S. R. Hamann, R. G. McAllister, *J. Cardiovasc. Pharmacol.* **5** (1983) 920 – 927.
262. C. H. Kleinbloesem, P. van Brummelen, J. A. van de Linde, P. J. Voogd et al., *Clin. Pharmacol. Ther.* **35** (1984) 742 – 749.
263. K. D. Rämsch, J. Sommer, *Hypertension* **5** (1983) Suppl. 2, I 18 – I 24.
264. R. Krebs, *Hypertension* **5** (1983) Suppl. 2, I 125 –I 129.
265. A. Scriabine, B. Garthoff, S. Kazda, K. Rämsch et al. in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 2,** Raven Press, New York 1984, pp. 37 – 39.
266. R. Towart, *J. Cardiovasc. Pharmacol.* **4** (1982) 895 – 902.
267. S. Kazda, B. Garthoff, A. Knorr, *Fed. Proc.* **42** (1983) 196 – 200.
268. H. Meyer, F. Bossert, E. Wehinger, K. Stoepel et al., *Arzneim.-Forsch.* **31** (1981) 407 – 409.
269. K. Stoepel, A. Heise, S. Kazda, *Arzneim.-Forsch.* **31** (1981) 2056 – 2061.
270. A. Knorr, K. Stoepel, *Arzneim.-Forsch.* **31** (1981) 2062 – 2064.
271. B. F. Johnson, L. Romero, R. Marwaha in A. Scriabine, S. Vanov, K. Deck (ed.): *Nitrendipine,* Urban & Schwarzenberg, Baltimore 1984, pp. 535 – 541.
272. K. Kobayashi, R. C. Tarazi, *Hypertension* **5** (1983) Suppl. 2, I 45 – I 51.
273. C. E. Hall, S. Hungerford, *Clin. Exp. Hypertens.* **5** (1983) 721 – 728.
274. D. G. Taylor, S. Shoemaker, T. E. Kowalski in A. Scriabine, S. Vanov, K. Deck (ed.): *Nitrendipine,* Urban & Schwarzenberg, Baltimore 1984, pp. 311 – 326.
275. D. C. Warltier, K. A. Lamping, M. G. Zyvoloski, G. J. Gross et al., *J. Cardiovasc. Pharmacol.* **5** (1983) 272 – 277.
276. H. R. Black, N. Vlachakis in A. Scriabine, S. Vanov, K. Deck (ed.): *Nitrendipine,* Urban & Schwarzenberg, Baltimore 1984, pp. 509 – 518.
277. J. F. Burris, A. V. Notargiacomo, V. Papademetriou, E. D. Freis, *Hypertension* **4** (1982) Suppl. 2, I 32 – I 35.
278. L. Hansson, L. Andrén, L. Oro, T. Ryman, *Hypertension* **5** (1983) Suppl. 2, I 25 – I 28.
279. H. O. Ventura, F. H. Messerli, W. Oigman, F. G. Dunn et al., *Am. J. Cardiol.* **51** (1983) 783 – 786.
280. F. M. Fouad, R. Pedrinelli, E. L. Bravo, F. Abi-Samra et al., *Clin. Pharmacol. Ther.* **35** (1984) 768 – 775.
281. H. Haas, E. Busch, *Arzneim.-Forsch.* **17** (1967) 257 – 272.
282. R. Mannhold, P. Zierden, R. Bayer, R. Rodenkirchen et al., *Arzneim.-Forsch.* **31** (1981) 773 – 780.
283. S. Baky, E. B. Kirsten in M. E. Goldberg (ed.): *Pharmacological and Biochemical Properties of Drug Substances,* Academy of Pharmaceutical Sciences, Washington, D.C. 1981, pp. 226 – 261.
284. S. Baky in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 2,** Raven Press, New York 1984, pp. 71 – 102.
285. B. N. Singh, J. T. Collett, C. Y. Chew, *Prog. Cardiovasc. Dis.* **22** (1980) 243 – 301.
286. R. A. Walsh, F. R. Badke, R. A. O'Rourke, *Am. Heart J.* **102** (1981) 341 – 350.
287. S. F. Flaim, K. Kanda in S. F. Flaim, R. Zelis (ed.): *Calcium Blockers: Mechanisms of Action and Clinical Applications,* Urban & Schwarzenberg, Baltimore 1982, pp. 155 – 178.

288. B. A. Gould, R. S. Hornung, S. Mann, V. B. Subramanian et al., *Hypertension* **5** (1983) Suppl. 2, I 91 – I 96.
289. K. Midtbo, O. Hals, J. van der Meer, *J. Cardiovasc. Pharmacol.* **4** (1982) Suppl. 3, S 363 – S 368.
290. P. Erne, P. Bolli, O. Bertel, U. L. Hulthén et al., *Hypertension* **5** (1983) Suppl. 2, I 97 – I 102.
291. U. L. Hulthen, P. Bolli, F. W. Amann, W. Kiowski, F. R. Bühler, *J. Cardiovasc. Pharmacol.* **4** (1982) Suppl. 3, S 313 – S 318.
292. G. Muiesan, E. Agabiti-Rosei, C. L. Alicandri in A. Zanchetti, D. M. Krikler (ed.): *Calcium Antagonism in Cardiovascular Therapy: Experience with Verapamil,* Excerpta Medica, Amsterdam 1981, pp. 238 – 249.
293. A. Chaudhry, M. M. Vohra, *Eur. J. Pharmacol.* **97** (1984) 156 – 158.
294. G. Haeusler, *J. Pharmacol. Exp. Ther.* **180** (1972) 672 – 682.
295. D. G. Taylor, T. E. Kowalski, *J. Pharmacol. Exp. Ther.* **228** (1984) 491 – 499.
296. L. H. Opie, *Lancet* **1** (1980) 806 – 810.
297. H. Kugita, H. Inoue, M. Ikezaki, M. Konda et al., *Chem. Pharm. Bull. (Tokyo)* **19** (1971) 595 – 602.
298. H. Kugita, H. Inoue, H. Ikezaki, S. Takeo, *Chem. Pharm. Bull. (Tokyo)* **18** (1970) 2028 – 2037.
299. T. Nagao, M. Sato, H. Nakajima, A. Kijomoto, *Jpn. J. Pharmacol.* **22** (1972) 1 – 10.
300. M. Sato, T. Nagao, I. Yamaguchi, H. Nakajima et al., *Arzneim.-Forsch.* **21** (1971) 1338 – 1343.
301. R. P. Hof, *Br. J. Pharmacol.* **78** (1983) 375 – 394.
302. H. Narita, T. Nagao, H. Yabana, I. Yamaguchi, *J. Pharmacol. Exp. Ther.* **227** (1983) 472 – 477.
303. M. Sato, S. Murata, H. Narita, M. Tomita et al., *Jpn. J. Pharmacol.* **75** (1979) 99 – 106.
304. I. Yamaguchi, K. Ikezawa, S. Murata, H. Narita et al., *Jpn. J. Pharmacol.* **24** (1974) 511 – 522.
305. S. F. Flaim in A. Scriabine (ed.): *New Drugs Annual: Cardiovascular Drugs,* **vol. 2,** Raven Press, New York 1984, pp. 123 – 156.
306. N. Nakaya, A. Schwartz, R. W. Millard, *Circ. Res.* **52** (1983) 302 – 311.
307. M. Matsuzaki, K. P. Gallagher, J. Patritti, T. Tajimi et al., *Circulation* **69** (1984) 801 – 814.
308. W. Klein, D. Brandt, K. Vrecko, M. Härringer, *Circ. Res.* **52** (1983) Suppl. 1, I 74 – I 173.
309. M. E. Safar, A. Ch. Simon, J. A. Levenson, J. L. Cazor, *Circ. Res.* **52** (1983) Suppl. 1, I 169 – I 173.
310. T. Meshi, J. Sugihara, Y. Sato, *Chem. Pharm. Bull. (Tokyo)* **19** (1971) 1546 – 1556.
311. J. E. Francis, *ACS Symp. Ser.* **27** (1976) 55 – 79.
312. H. A. J. Struyker-Boudier, H. Van Essen, J. F. M. Smits, *Eur. J. Pharmacol.* **95** (1983) 151 – 159.
313. E. G. Spokas, H.-H. Wang, *J. Pharmacol. Exp. Ther.* **212** (1980) 294 – 303.
314. I. Khatri, N. Uemura, A. Notargiacomo, E. D. Freis, *Am. J. Cardiol.* **40** (1977) 38 – 42.
315. S. Sen, R. C. Tarazi, *Am. J. Physiol.* **244** (1983) H 97 – H 101.
316. M. J. M. C. Thoolen, J. C. A. Van Meel, B. Wilffert, P. B. M. W. M. Timmermans et al., *Pharmacology* **27** (1983) 245 – 254.
317. C. Chevillard, M. N. Mathieu, B. Saiag, M. Worcel, *Br. J. Pharmacol.* **69** (1980) 415 – 420.
318. V. S. R. Krishnamurty, *Can. J. Physiol. Pharmacol.* **59** (1981) 1159 – 1169.
319. M. B. Murphy, A. J. Scriven, M. J. Brown, R. Causon et al., *Eur. J. Clin. Pharmacol.* **23** (1982) 479 – 482.
320. O. Kohlmann, M. Bresnahan, H. Gavras, *Hypertension* **6** (1984) Suppl. 1, I 1 – I 6.
321. E. G. Spokas, G. Folco, J. Quilley, P. Chander, J. C. McGiff, *Hypertension* **5** (1983) Suppl. 1, I 107 – I 111.
322. G. Haeusler, M. Gerold, *Naunyn-Schmiedebergs Arch. Pharmacol.* **310** (1979) 155 – 167.
323. K. Hermsmeyer, A. Trapani, P. W. Abel, M. Worcel, *J. Pharmacol. Exp. Ther.* **227** (1983) 322 – 326.
324. C. Chevillard, B. Saiag, M. Worcel, *Br. J. Pharmacol.* **73** (1981) 811 – 817.
325. A. J. McLean, P. DuSouish, K. W. Barron, A. H. Briggs, *J. Pharmacol. Exp. Ther.* **207** (1978) 40 – 48.
326. M. Khayyal, F. Gross, V. A. W. Kreye, *J. Pharmacol. Exp. Ther.* **216** (1981) 390 – 394.
327. J. L. Freslon, J. F. Guidicelli, *Br. J. Pharmacol.* **80** (1983) 533 – 543.
328. S. Greenberg, *J. Pharmacol. Exp. Ther.* **215** (1980) 279 – 286.
329. L. Hansson, R. Olander, H. Aberg, R. Malmcrona et al., *Acta Med. Scand.* **190** (1971) 531 – 534.
330. C. K. Newsome, *Southwest Med.* **42** (1961) 558 – 567.
331. J. A. Franciosa, G. Pierpont, J. N. Cohn, *Ann. Intern. Med.* **86** (1977) 388 – 393.

332. A. A. Rubin, F. E. Roth, R. M. Taylor, H. Rosenkilde, *J. Pharmacol. Exp. Ther.* **136** (1962) 344 – 352.
333. T. C. Hamilton, D. Robson, *Eur. J. Pharmacol.* **32** (1975) 273 – 278.
334. W. G. Nayler, I. McInnes, J. B. Swann, D. Race et al. *Am. Heart J.* **75** (1968) 223 – 232.
335. P. Limbourg, P. Figel, H. Justin, K. F. Lang, *Eur. J. Clin. Pharmacol.* **8** (1975) 387 – 392.
336. J. Taylor, R. D. Green, *Eur. J. Pharmacol.* **12** (1970) 385 – 387.
337. J. C. A. van Meel, P. B. M. W. M. Timmermans, P. A. van Zwieten, *Eur. J. Pharmacol.* **92** (1983) 27 – 34.
338. A. J. Wohl, L. M. Hausler, F. E. Roth, *J. Pharmacol. Exp. Ther.* **162** (1968) 109 – 114.
339. N. Winer, D. S. Chokshi, M. S. Yoon, A. D. Freedman, *J. Clin. Endocrinol.* **29** (1969) 1168 – 1175.
340. R. Vandongen, D. M. Greenwood, *Eur. J. Pharmacol.* **33** (1975) 197 – 200.
341. G. Sponer, P. Schelling, D. Ganten, F. Gross, *Naunyn-Schmiedebergs Arch. Pharmacol.* **303** (1978) 15 – 20.
342. I. I. A. Tabachnick, A. Gulbenkian, F. Seidman, *J. Pharmacol. Exp. Ther.* **150** (1965) 455 – 462.
343. A. McNair, F. Andreasen, P. E. Nielsen, *Eur. J. Clin. Pharmacol.* **24** (1983) 151 – 156.
344. D. W. DuCharme, W. A. Freyburger, B. E. Graham, R. G. Carlson, *J. Pharmacol. Exp. Ther.* **184** (1973) 662 – 670.
345. W. A. Pettinger, *N. Engl. J. Med.* **303** (1980) 922 – 926.
346. D. W. DuCharme, G. R. Zins in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 415 – 421.
347. M. Velasco, K. O'Malley, N. W. Robie, J. Wells et al., *Clin. Pharmacol. Ther.* **16** (1974) 1031 – 1038.
348. H. R. Brunner, P. Jaeger, R. K. Ferguson, E. Jequier et al., *Br. Med. J.* **2** (1978) 385 – 388.
349. H. N. Sapru, A. J. Krieger, *Am. J. Physiol.* **236** (1979) H 174 – H 182.
350. S. Sen, R. C. Tarazi, F. M. Bumpus, *Cardiovasc. Res.* **11** (1977) 427 – 433.
351. R. V. Jackson, M. M. Williamson, B. Seneviratne, R. D. Gordon, *Aust. N.Z. J. Med.* **13** (1983) 39 – 44.
352. J. M. McCall, J. W. Aiken, C. G. Chidester, D. W. DuCharme et al., *J. Med. Chem.* **26** (1983) 1791 – 1793.
353. C. J. Limas, J. N. Cohn, *Circ. Res.* **35** (1974) 609 – 617.
354. G. D. Nicholson, G. A. Alleyne, G. Valdes, R. L. Westerman, *J. Cardiovasc. Pharmacol.* **2** (1980) Suppl., 242 – 246.
355. F. C. Luft, R. Bloch, J. J. Szwed, C. M. Grim et al., *JAMA* **240** (1978) 1985 – 1987.
356. M. Valasco, A. Urbina-Quintana, E. Hernández, A. Ramirez et al., *Eur. J. Clin. Pharmacol.* **20** (1981) 259 – 262.
357. R. E. Mutterperl, F. B. Diamond, D. T. Lowenthal, *J. Clin. Pharmacol.* **16** (1976) 498 – 509.
358. Y. K. Seedat, R. Rawat, *Eur. J. Clin. Pharmacol.* **25** (1983) 9 – 11.
359. E. H. Herman, T. Balazs, R. Young, F. L. Earl et al., *Toxicol. Appl. Pharmacol.* **47** (1979) 493 – 503.
360. R. L. Cohen, M. E. Alves, V. C. Weiss, D. P. West et al., *J. Invest. Dermatol.* **82** (1984) 90 – 93.
361. V. A. W. Kreye, E. Marquard, *Naunyn-Schmiedebergs Arch. Pharmacol.* **306** (1979) 203 – 207.
362. I. H. Tuzel, *J. Clin. Pharmacol.* **14** (1974) 494 – 503.
363. M. Pagani, S. F. Vatner, E. Braunwald, *Circulation* **57** (1978) 144 – 151.
364. R. D. Bastron, G. J. Kaloyanides, *J. Pharmacol. Exp. Ther.* **181** (1972) 244 – 249.
365. G. Ross, P. V. Cole, *Anaesthesia* **28** (1973) 400 – 406.
366. A. P. Adams, T. N. S. Clarke, J. Edmonds-Seal, P. Föex et al., *Br. J. Anaesth.* **46** (1974) 807 – 817.
367. J. Kyncl, *Naunyn-Schmiedebergs Arch. Pharmacol.* **269** (1971) 390 – 391.
368. V. A. W. Kreye in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 373 – 396.
369. C. A. Gruetter, B. K. Barry, D. B. McNamara, D. Y. Gruetter, P. J. Kadowitz, L. J. Ignarro, *J. Cyclic Nucleotide Res.* **5** (1979) 211 – 224.
370. C. A. Gruetter, D. Y. Gruetter, J. E. Lyon, P. J. Kadowitz et al., *J. Pharmacol. Exp. Ther.* **219** (1981) 181 – 186.
371. V. A. W. Kreye, *Trends Pharmacol. Sci.* **1** (1980) 384 – 388.
372. V. A. W. Kreye, G. D. Baron, J. B. Lüth, H. Schmidt-Gayk, *Naunyn-Schmiedebergs Arch. Pharmacol.* **288** (1975) 381 – 402.
373. D. H. Foley, *Pharmacology* **28** (1984) 95 – 103.

Boric Oxide, Boric Acid, and Borates

Robert A. Smith, U.S. Borax Research Corp., Anaheim, California 92801, United States

1.	**Introduction**	1
2.	**Boric Oxide**	3
2.1.	**Physical Properties**	3
2.2.	**Chemical Properties**	4
2.3.	**Production**	4
3.	**Boric Acid**	4
3.1.	**Physical Properties**	4
3.2.	**Chemical Properties**	5
3.3.	**Production**	5
4.	**Aqueous Borate Solutions**	8
5.	**Sodium Borates**	8
5.1.	**Sodium Tetraborate Decahydrate**	9
5.2.	**Sodium Tetraborate Pentahydrate**	9
5.3.	**Anhydrous Sodium Tetraborate**	11
5.4.	**Sodium Metaborate Tetrahydrate**	11
5.5.	**Sodium Metaborate Dihydrate**	12
5.6.	**Sodium Pentaborate Pentahydrate**	12
5.7.	**Disodium Octaborate Tetrahydrate**	12
5.8.	**Disodium Tetraborate Tetrahydrate**	12
6.	**Other Metal Borates**	13
6.1.	**Calcium and Calcium Sodium Borates**	13
6.2.	**Lithium, Potassium, and Ammonium Borates**	13
6.2.1.	Lithium Borates	13
6.2.2.	Potassium Borates	13
6.2.3.	Ammonium Borates	14
6.3.	**Zinc Borates**	14
7.	**Borate Glasses**	15
8.	**Quality Specifications**	15
9.	**Uses**	15
10.	**Economic Aspects**	16
11.	**Toxicology and Occupational Health**	17
12.	**References**	17

1. Introduction

Boron, the fifth element in the periodic table, is the only electron-deficient nonmetallic element. It has a great affinity for oxygen. In nature boron always occurs in the oxygenated state, mainly as borates. The structure of these covalently bonded boron – oxygen compounds consists of either planar BO_3 units with 120° between bonds or tetrahedral BO_4 units. In addition, boron is the only light element having two abundant isotopes, ^{10}B (18.8 %) and ^{11}B (81.2 %).

Boron is found in the earth's crust at an average concentration of about 10 ppm. Large deposits of commercially valuable boron minerals are only found in a few localities, usually sites of earlier intense volcanic activity. There are about 150 known boron-containing minerals [1], and of these only tincal, colemanite, ulexite, kernite, probertite, and szaibelyite are of significant commercial importance (Table 1).

The Arabic word for borax, baurach, is found in old manuscripts from ancient Persia and Arabia dating 2000 years ago. Tincal, the mineral name for borax decahydrate, derives from tincana, the Sanskrit word for borax. Borates were known before Babylonian times to be useful as a flux for welding gold and have been found in enamels from China as early as 300 B.C. At the end of the 13th century Marco Polo brought borax from Mongolia to Europe. This became the primary European source for use as a soldering and enameling agent. In Turkey borates have been known since the 13th century. In 1772 boric acid or sassolite was discovered in hot springs located in Tuscany, Italy. About 1836, borates were discovered in Chile and Argentina, and these deposits quickly became the major source of world borates until the end of the century. By 1864 borates had been discovered in California and Nevada (USA).

The large sodium borates deposit of United States Borax & Chemical Corp. was discovered in California in 1913. Production began in 1927, and this has become the principal world source of refined sodium borates and boric acid. Reserves include tincal, kernite, and ulexite. American Borate Corp. near Death Valley in California produces colemanite and ulexite. Kerr McGee Chemical Corp. extracts borates from brines located at Searles Lake in California.

Reserves and production of borates in Turkey are operated by Etibank, a government con-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_263

Table 1. Important borate minerals

Mineral	CAS registry no.	Formula		%		Location
				B_2O_3	H_2O	
Tincal, borax	[*1303-96-4*]	$Na_2O \cdot 2B_2O_3 \cdot 10H_2O$	$Na_2B_4O_7 \cdot 10H_2O$	36.5	47.2	Boron, California, USA; Turkey; Argentina
Tincalconite	[*12045-88-4*]	$Na_2O \cdot 2B_2O_3 \cdot 5H_2O$	$Na_2B_4O_7 \cdot 5H_2O$	47.8	30.9	California, USA; Turkey; Argentina
Kernite	[*12045-87-3*]	$Na_2O \cdot 2B_2O_3 \cdot 4H_2O$	$Na_2B_4O_7 \cdot 4H_2O$	50.9	26.4	Boron, California, USA; Argentina
Ulexite	[*1319-33-1*]	$Na_2O \cdot 2CaO \cdot 5B_2O_3 \cdot 16H_2O$	$NaCaB_5O_9 \cdot 8H_2O$	43.0	35.6	California, USA; Nevada, USA; Argentina; Chile; Peru; Tibet; Turkey
Colemanite	[*12291-65-5*]	$2CaO \cdot 3B_2O_3 \cdot 5H_2O$	$Ca_2B_6O_{11} \cdot 5H_2O$	50.8	21.9	Turkey; California, USA; Nevada, USA; Mexico;
Pandermite, priceite	[*61583-61-7*]	$4CaO \cdot 5B_2O_3 \cdot 7H_2O$	$Ca_4B_{10}O_{19} \cdot 7H_2O$	49.8	18.1	Turkey
Hydroboracite	[*12046-12-7*]	$CaO \cdot MgO \cdot 3B_2O_3 \cdot 6H_2O$	$CaMgB_6O_{11} \cdot 6H_2O$	50.5	26.2	Caucasus, USSR; Argentina
Inyoite	[*12260-25-2*]	$2CaO \cdot 3B_2O_3 \cdot 13H_2O$	$Ca_2B_6O_{11} \cdot 13H_2O$	37.6	42.2	Kazakhstan, USSR; Argentina
Ascharite, szaibelyite	[*12447-04-0*]	$2MgO \cdot B_2O_3 \cdot H_2O$	$Mg_2B_2O_5 \cdot H_2O$	41.4	10.7	USSR; China
Datolite	[*1318-40-7*]	$2CaO \cdot B_2O_3 \cdot 2SiO_2 \cdot H_2O$	$Ca_2B_2Si_2O_9 \cdot H_2O$	21.8	5.6	Kazakhstan, USSR
Meyerhofferite	[*57572-66-4*]	$2CaO \cdot 3B_2O_3 \cdot 7H_2O$	$Ca_2B_6O_{11} \cdot 7H_2O$	46.7	28.2	Turkey
Inderite	[*12260-26-3*]	$2MgO \cdot 3B_2O_3 \cdot 15H_2O$	$Mg_2B_6O_{11} \cdot 15H_2O$	37.3	48.3	Argentina
Howlite	[*1318-68-9*]	$4CaO \cdot 5B_2O_3 \cdot 2SiO_2 \cdot 5H_2O$	$Ca_4B_{10}Si_2O_{21} \cdot 5H_2O$	44.4	11.5	Mexico; Turkey
Probertite	[*12229-14-0*]	$Na_2O \cdot 2CaO \cdot 5B_2O_3 \cdot 10H_2O$	$NaCaB_5O_9 \cdot 5H_2O$	49.6	25.6	California, USA
Sassolite	[*10043-35-3*]	$B_2O_3 \cdot 3H_2O$	H_3BO_3	56.3	43.7	Tuscany, Italy
Boracite	[*1303-91-9*]	$5MgO \cdot MgCl_2 \cdot 7B_2O_3$	$Mg_3B_7O_{13}Cl$	62.2	—	Turkey

trolled corporation [2]. A large sodium borates deposit found at Kirka has reserves of tincal and kernite, estimated at 500×10^6 t, which are very similar to the deposit in California. The Kirka reserve is by far the largest known borate deposit. Much smaller deposits of colemanite are found at Bigadic, Keatelek, and Hisarcik. Ulexite is also found at Bigadic. Colemanite and ulexite are the principal minerals mined.

Both tincal and ulexite are found in Argentina and ulexite is found in Peru. Borate reserves in the Soviet Union are found in the Inder district of Kazakhstan, and consist principally of szaibelyite and hydroboracite. Although production in the USSR is relatively small, reserves are believed to be extensive. Ascharite, a magnesium borate, is the source of most borates produced in the Peoples Republic of China.

2. Boric Oxide

Boric oxide (B_2O_3) [*1303-86-2*], M_r 69.62, exists as vitreous glass and as two crystallineforms. The *vitreous form* is a colorless, hard, tough, glasslike solid, commonly prepared by dehydration of boric acid. The most common crystalline B_2O_3 has a *hexagonal* or *α-form* and is crystallized at 200 – 250 °C at ambient pressure. A less common *monoclinic β-form* can be crystallized at 4000 MPa and 600 °C [3].

2.1. Physical Properties

Vitreous B_2O_3 is difficult to obtain in a completely anhydrous state even when heated at > 1000 °C [4]. Many physical properties of B_2O_3 are sensitive to small amounts of residual water. Very dry B_2O_3 can be prepared by heating at 200 – 400 °C in high vacuum [5]. Vitreous B_2O_3 softens between 325 and 450 °C and the melt is still quite viscous at > 1000 °C. The dynamic viscosity of the melt is:

t, °C	300	400	500	600	700
η, Pa · s	4.4×10^8	1.6×10^8	3900	480	85
t, °C	800	900	1000	1100	
η, Pa · s	26	12	7.4	4.3	

The density (ϱ) of the vitreous form is dependent on both its thermal history and moisture content. Commercial boric oxide has a density of 1.84 g/cm^3, whereas very dry B_2O_3 has a density of 1.82 g/cm^3 at 20 °C. The density at high temperature has been investigated [6]. Annealed B_2O_3 has a greater density compared to rapidly cooled melt (1.53 g/cm^3 when quenched from 1000 °C), implying a higher degree of molecular order in the annealed material [7]. The heat of formation of amorphous B_2O_3 glass is 1260 kJ/mol [8].

Crystalline hexagonal or α-form, *mp* 450 °C, ϱ 2.42 – 2.46 g/cm^3; monoclinic or β-form, *mp* ca. 510 °C [9], ϱ 2.95 g/cm^3; the β-form is much more resistant to attack by water and dilute HF than the α-form.

The heat of formation for crystalline B_2O_3 is 1274 kJ/mol. Its heat capacity increases with temperature from 1.05 J g^{-1}K^{-1} at 100 °C to 1.82 J g^{-1}K^{-1} at 275 °C [10]. The heat of vaporization is calculated as 391.6 kJ/mol at 1800 °C [11], and the heat of sublimation is 431.4 kJ/mol at 25 °C [12]. The electrolytic conductivity (κ) is only 5×10^{-4} S/cm at 300 °C [13]. The refractive index (n_D) varies between 1.4502 and 1.4633, depending on the thermal history [5]. The vapor pressure of B_2O_3 melt is [9], [14]:

t, °C	1163	1270	1673	1810	2000	2146
$p_{B_2O_3}$, kPa	0.0005	0.005	1.5	5.7	24	80

The vapor composition (by mass spectrometry) is predominantly $B_2O_3^+$ with about 7 % each of $B_2O_2^+$ and BO^+ and 2 % B^+. Boric oxide has been observed to be more volatile in steam than in a dry atmosphere. The green color of B_2O_3 in a flame has been attributed to BO_2 as the emitting species.

Historically there has been considerable controversy over the *molecular structure* in liquid and vitreous B_2O_3. The structure is believed to consist of a random network of trigonal BO_3 units and boroxol rings with short-range localized order. There is disagreement over the relative amounts of these two structural units. A summary of the structural data has been reported [15].

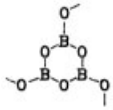

Boroxol

Trigonal BO_3

Apparently the structure interpretation can be influenced by the presence of residual moisture.

2.2. Chemical Properties

Molten B_2O_3 is corrosive to most metals and alloys at temperatures exceeding 1000 °C. This is because of the fluxing character of B_2O_3, which keeps the metal surface clean and vulnerable to attack by oxygen in the air. Molybdenum and nickel alloys are resistant to corrosion below 1000 °C, whereas silicon carbide is resistant above 1200 °C.

Reduction of B_2O_3 at high temperatures by aluminum, magnesium, or alkali metals produces $B_{12}O_3$ and other boron suboxides. Above 900 °C carbon reacts with B_2O_3 in a nitrogen atmosphere to form boron nitride, which can also be made by reacting B_2O_3 with ammonia at 600 – 900 °C.

Vitreous B_2O_3 reacts exothermically with 3 mol of water to form crystalline boric acid, with a heat of hydration of − 76.5 kJ/mol [16]; heat of hydration of crystalline B_2O_3 is − 58.2 kJ/mol. Above 135 – 140 °C boric oxide reacts with water to form metaboric acid.

Sodium hydride reacts with B_2O_3 at 330 – 350 °C to give sodium borohydride in about 60 % yield [17]. Reaction of either B_2O_3 or boric acid with anhydrous H_2SO_4 or oleum forms the acid $HB(HSO_4)_4$ [18].

2.3. Production

High-purity B_2O_3 (99 %) is made by fusing refined boric acid. A 96 % grade of B_2O_3 is made commercially by fusing a mixture of sodium tetraborate plus sulfuric acid at 800 °C. Two molten layers are formed, an upper layer of boric oxide and a lower layer of sodium sulfate. The molten B_2O_3 layer is separated and cooled on chill rollers. The resulting glass is ground, sized, and packaged in moisture-proof containers. The product B_2O_3 contains 3 – 4 % sodium sulfate.

Boric oxide can also be prepared by thermal decomposition of ammonium pentaborate, $NH_4B_5O_8 \cdot 4H_2O$ [*12229-12-8*], at 500 – 900 °C and through direct reaction of boron with oxygen [19].

3. Boric Acid

Boric acid can be viewed as a hydrate of boric oxide and exists both as a trihydrate, orthoboric acid ($B_2O_3 \cdot 3H_2O$ or H_3BO_3) [*10043-35-3*], and asa monohydrate, metaboric acid ($B_2O_3 \cdot H_2O$ or HBO_2) [*13460-50-9*]. Only the more stable orthoboric acid form is of commercial importance and is usually referred tosimply as boric acid. The terms "pyroboric acid" and "tetraboric acid" are sometimes encountered in the literature, but these acids do not actually exist as solid-phase compounds.

3.1. Physical Properties

Boric acid crystallizes from an aqueous solution as odorless, white, waxy platelets. Crystalline orthoboric acid (*mp* 170.9 °C, heated in a closed space) is monoclinic, having a sheet lattice structure. Sheets consist of coplanar $B(OH)_3$ molecules linked by hydrogen bonds. These sheetlike layers are held together only by van der Waals forces. This allows easy cleavage into flakes which feel slippery.

The three known crystalline modifications of metaboric acid are shown in the $B_2O_3-H_2O$ solubility phase diagram (Fig. 1) [20]. Orthoboric acid slowly dehydrates at 100 – 130 °C to form flaky crystals of orthorhombic HBO_2 (III) (α-form, *d* 1.784, *mp* 176 °C). After disappearance of the residual H_3BO_3 at temperatures up to 160 °C, a coarsely crystalline monoclinic HBO_2 (II) (β-form, *d* 2.044, *mp* 201 °C) is formed. At yet higher temperatures (200 – 250 °C) a highly viscous liquid is formed from which cubic HBO_2(I) (γ-form, *d* 2.486, *mp* 236 °C) slowly crystallizes. With about 1 % moisture remaining, HBO_2 (I) acts as seed for crystallization of hexagonal (α) boric oxide. The structures of these metaborate forms range from all trigonal borons in HBO_2 (III), to one-third tetrahedral borons in HBO_2 (II), to all tetrahedal borons in HBO_2 (I).

The solubility of boric acid in water is increased by addition of salts, such as KCl, KNO_3, RbCl, K_2SO_4, and Na_2SO_4, whereas addition of LiCl, NaCl, and $CaCl_2$ tends to lower the solubility.

Boric acid behaves as a weak acid in aqueous solutions, pK_a of 9.23 at 25 °C [21]. As

shown in Table 2, pH decreases as concentration increases. This is because of formation of trimeric and tetrameric species. The pH also increases with temperature. The solubility of boric acid in water is shown in Table 3.

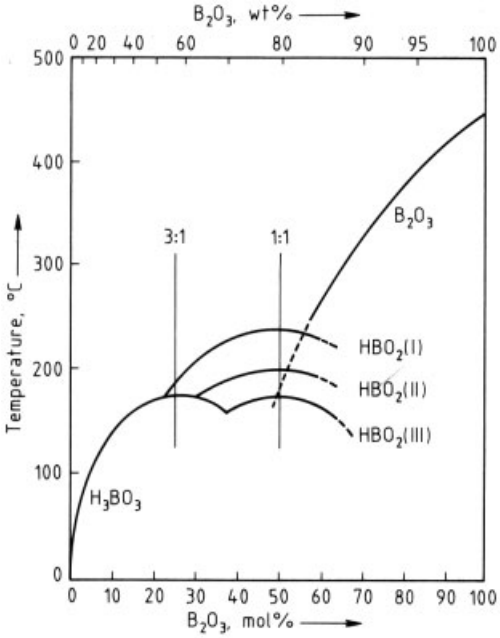

Figure 1. Solubility diagram for the system $B_2O_3-H_2O$.

The solubility of boric acid in many organic solvents is shown in Table 4. Boric acid is particularly soluble in polyhydric alcohols capable of forming complexes. Normally, boric acid is too weak an acid to titrate directly. Use of a polyhydric complexing agent, such as mannitol [*69-65-8*], enhances the acidity of boric acid sufficiently to allow accurate titration using phenolphthalein as the indicator.

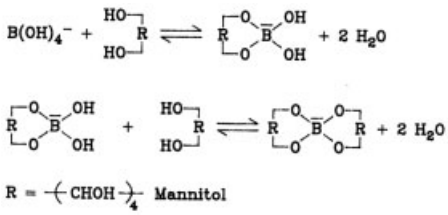

The vapor pressure of boric acid is primarily due to water from dehydration equilibria. Vapor below 160 °C is composed of water and boric acid molecules. At higher temperatures, HBO_2 is also found in the vapor phase, and above 940 °C $(HBO_2)_3$ becomes significant [22].

Boric acid is volatile with steam. Distillate from a boiling saturated solution contains about 0.18 % of boric acid [23]. Boric acid is particularly volatile from concentrated solutions made acidic by strong acids. The volatile loss from low molecular mass alcohol solutions can be attributed to ester formation.

3.2. Chemical Properties

Boric acid reacts with strong bases to form metaborate ion, $B(OH)_4^-$, and with alcohols to form borate esters. Boric acid reacts with fluoride ion to form tetrafluoroboric acid and with hydrofluoric acid to form trifluoroboric acid $H(F_3BOH)$, which can be distilled at reduced pressures (58.5 – 60.0 °C at 0.16 kPa) [24].

3.3. Production

Boric acid is manufactured industrially from borate minerals and brines [25]. Alkali and alkaline-earth metal borates, such as borax, kernite, colemanite, ascharite, ulexite, or hydroboracite, react with strong mineral acids to form boric acid. In the United States boric acid is made primarily from sodium borate minerals, whereas in Europe it is made from colemanite imported from Turkey. In the Soviet Union magnesium borates from the Inder region are used to make boric acid. Borates found in saline lake brine in California at Searles Lake are extracted and recovered as boric acid.

The world's largest and most modern boric acid plant was constructed by U.S. Borax in 1981 at Boron, California. Crushed kernite ($Na_2O \cdot 2\,B_2O_3 \cdot 4\,H_2O$) ore is reacted with sulfuric acid in recycled weak liquor (contains low concentration of borates) at 100 °C. Coarse gangue is separated by rake classifiers and fine particles are settled in a thickener. The boric acid strong liquor (high borate concentration) is nearly saturated with sodium sulfate. Complete solubility of sodium sulfate is maintained throughout the process by careful control of pH and temperature. The strong liquor is filtered at 98 °C and boric acid crystallized in two stages using continuous evaporative crystallizers. The temperature is dropped to 70 °C in the first stage and to 35 °C in the second. Crystals are filtered

Table 2. pH values of aqueous boric acid and borate solutions as a function of concentration

Compound	pH at wt % of compound							
	t, °C	0.1	0.5	1.0	2.0	4.0	10.0	15.0
$B(OH)_4$ [a]	20	6.1	5.6	5.1	4.5	3.9		
$Na_2B_4O_7 \cdot 10H_2O$	20	9.3	9.2	9.2	9.2	9.3		
$Na_2B_5O_8 \cdot 5H_2O$ [b]	23			8.5	8.4	8.1	7.6	7.3
$Na_2B_8O_{13} \cdot 4H_2O$	20			8.5	8.5	8.1	7.6	7.3
$K_2B_4O_7 \cdot 4H_2O$	25	9.2	9.1	9.1	9.2	9.3		
$KB_5O_8 \cdot 4H_2O$ [c]	25	8.4	8.4	8.4	8.1	7.9		
$(NH_4)_2B_4O_7 \cdot 4H_2O$	25	8.8	8.9	8.8	8.8	8.8	8.9	
$NH_4B_5O_8 \cdot 4H_2O$	25	8.5	8.4	8.3	8.2	7.9	7.3	
$NaBO_2 \cdot 4H_2O$	20	10.5	10.8	11.0	11.2	11.4	11.7	11.9

[a] Saturated solution at 20 °C, pH 3.7 at 4.7 wt %.
[b] At 40 °C, pH 6.9 at 20 wt %, pH 6.5 at 30 wt %.
[c] pH 7.6 at 5.0 wt %.

Table 3. Solubility of boric acid and borates in water

Compound	CAS registry no.	Solubility, wt % (anhydrous salt), at temp., °C										
		0	10	20	30	40	50	60	70	80	90	100
$LiBO_2 \cdot 8H_2O$ [a]	[*41851-38-1*]	0.9	1.4	2.5	4.6	9.4						
$LiBO_2 \cdot 2H_2O$	[*15293-74-0*]					7.4	7.8	8.4	9.5	10.6	11.9	13.4
$Li_2B_4O_7 \cdot 3H_2O$	[*37195-62-3*]	2.5		2.8	3.0	3.2	3.5	3.8	4.1	4.4	4.9	5.4
$LiB_5O_8 \cdot 5H_2O$ [b]	[*37190-10-6*]				17.9	18.7	20.9	24.3	28.0	32.0	36.6	41.42
$NaBO_2 \cdot 4H_2O$ [c]	[*10555-76-7*]	14.5	17.0	20.0	23.6	27.9	34.1					
$NaBO_2 \cdot 2H_2O$	[*16800-11-6*]							38.3	40.9	43.7	47.4	52.4
$Na_2B_4O_7 \cdot 10H_2O$ [d]	[*1303-96-4*]	1.0	1.6	2.5	3.8	5.9	9.5	16.0				
$Na_2B_4O_7 \cdot 5H_2O$	[*12045-88-4*]							16.4	19.5	23.4	28.1	34.6
$Na_2B_4O_7 \cdot 4H_2O$ [e]	[*12045-87-3*]						14.2	14.7	17.0	19.7	23.0	27.2
$NaB_5O_8 \cdot 5H_2O$	[*12046-75-2*]	5.8	7.5	10.6	13.7	17.5	21.7	26.9	35.0	38.1	44.3	51.0
$B(OH)_3$	[*10043-35-3*]	2.4	3.5	4.7	6.2	8.8	10.3	13.0	15.8	19.1	23.3	27.5
$KBO_2 \cdot 4H_2O$ [f]		34.2	38.2	43.8								
$KBO_2 \cdot 1.3H_2O$ [g]	[*27516-44-5*]	35.5			44.4	45.5	46.4	47.5	48.5	49.6	50.6	51.6
$K_2B_4O_7 \cdot 4H_2O$	[*12045-78-2*]	6.3	9.0	12.1	15.6	19.4	24.0	28.5	33.3	38.2	43.3	48.4
$K_4B_{10}O_{17} \cdot 5H_2O$	[*12429-89-9*]					23.2	24.9	26.9	28.6	31.0	32.5	33.8
$KB_5O_8 \cdot 4H_2O$ [h]	[*12229-13-9*]	1.6	2.1	2.8	3.8	5.1	6.9	9.0	11.7	14.7	18.3	23.3
$KB_5O_8 \cdot 2H_2O$ [i]	[*12513-21-2*]	2.0	2.6	3.2	3.8	5.0	6.4	7.8	9.2	10.6	13.9	15.3
$Na_2B_8O_{13} \cdot 4H_2O$	[*12280-03-4*]	1.9	3.7	7.8	18.1	22.9	26.4	28.9	39.3 (75 °C)		37.4 (94 °C)	
$(NH_4)_2B_4O_7 \cdot 4H_2O$	[*12228-87-4*]	3.8	5.3	7.6	10.8	15.8	21.2	27.2	34.4	43.1	52.7	
$NH_4B_5O_8 \cdot 4H_2O$	[*12229-12-8*]	4.0	5.4	7.1	9.1	11.4	14.4	18.2	22.4	26.4	30.3	

[a] Transition of the 8-hydrate to the 2-hydrate at 36.9 with 7.2 % $LiBO_2$.
[b] Incongruent solubility below 37.5 or 40.5 °C.
[c] Transition of the 4-hydrate to the 2-hydrate at 53.6 °C with 36.9 % $NaBO_2$.
[d] Transition of the 10-hydrate to the 5-hydrate at 60.7 °C with 16.6 % $Na_2B_4O_7$.
[e] Transition of the 10-hydrate to the 4-hydrate at 58.2 °C with 14.6 % $Na_2B_4O_7$.
[f] Transition of the 4-hydrate to the 1.3-hydrate at 24 °C with 43.8 wt % KBO_2.
[g] Metastable below 24 °C.
[h] More stable below 46 °C.
[i] More stable above 46 °C.

and washed with progressively weaker liquor in a countercurrent fashion. The final product is dried in rotary driers and screened for packaging. Weak liquor at 35 °C, nearly saturated in sodium sulfate, is heated to 100 °C and returned to the reactor. Sodium sulfate is removed from the process as solid in gangue and in effluent from a second-stage thickener, and is pumped to sealed solar ponds. The plant has a capacity of 113000 t/a (expressed as B_2O_3).

In Europe boric acid is produced from crushed colemanite ($2CaO \cdot 3B_2O_3 \cdot 5H_2O$) ore by reaction with sulfuric acid at 90 °C. Byproduct gypsum ($CaSO_4$) and gangue are fil-

Table 4. Solubility of boric acid and sodium tetraborate hydrates in organic solvents

Solvent	t, °C	Solubility, wt %			
		$B(OH)_3$	$Na_2B_4O_7 \cdot 10H_2O$	$Na_2B_4O_7 \cdot 5H_2O$	$Na_2B_4O_7$
Methanol	20	20.68			
	25		19.9	16.9	16.7 [b]
Ethanol	25	94.4 [a]			
Aqueous ethanol (46.5 %)	15.5		2.48		
n-Butanol	25	42.8 [a]			
n-Propanol	25	59.4 [a]			
2-Methylbutanol	25	35.3 [a]			
Isoamyl alcohol	25	2.39			
Ethylene glycol	25	18.5	41.6	31.2	15.5 [c]
Propylene glycol	25	15.06		21.9	
Diethylene glycol	25	13.6	18.6	10.0	
Glycerol (86.5 %)	20	21.1	47.2		
Glycerol (98.5 %)	20	19.9	52.6		
Glycerol	25	17.55			
Mannitol (10 %)	25	6.62			
Acetone	25	0.6	0.6		
Methyl ethyl ketone	20	0.7			
Methyl butyl ketone	20	0.23			
Ethyl acetate	25	1.5	0.14		
Acetic acid (glacial)	30	6.3			
Diethyl ether	20	0.008			
Dioxane	25	≈15 [a]			
Aniline	20	0.15			
Pyridine	25	≈70 [a]			
Ammonia (liquid)	25	1.88			40.6 [a]
Formamide	25				
Fuel oil	25	2.46			
Water	25	5.46			
	20	4.72			

[a] Solubility, g/L.
[b] Finely divided solid.
[c] Viscous solution.

tered and the hot mother liquor cooled to crystallize boric acid. The weak liquor is recycled to the reactor.

Magnesium borate ores, primarily of ascharite (szarbelyite, $2\,MgO \cdot B_2O_3 \cdot H_2O$), and hydroboracite ($CaO \cdot MgO \cdot 3\,B_2O_3 \cdot 6\,H_2O$), are finely ground and decomposed to boric acid with sulfuric acid at 95 °C. Insoluble gypsum and gangue are filtered and the mother liquor is cooled to 15 °C to crystallize boric acid. The weak liquor is nearly saturated in magnesium sulfate ($MgSO_4 \cdot 7\,H_2O$), which is crystallized by concentration of the liquor at elevated temperature; the concentrated liquor is then recycled to the reactor.

Datolite ($2\,CaO \cdot B_2O_3 \cdot 2\,SiO_2 \cdot H_2O$) is the most common silicoborate mineral. Ore containing 5 % B_2O_3 is finely ground and digested with sulfuric acid. Silica in the mother liquor is coagulated upon heating to 95 °C. The resulting boric acid solution is nearly free of silicic acid.

At Searles Lake, California, borax is found in brine at a concentration of about 1.5 % (expressed as anhydrous borax); a number of other salts are also present. A process using liquid–liquid extraction is used to separate borax selectively from brine. Borax is extracted into a water-insoluble solvent, such as kerosene, by use of an aromatic polyol that efficiently complexes the borate ion, e.g., 3-chloro-2-hydroxy-5-isooctylbenzenemethanol [*31567-74-5*]. The organic phase is isolated and acidified with sulfuric acid, giving an aqueous solution of boric acid and sodium sulfate. After concentrating by evaporation, the liquor is cooled to crystallize boric acid.

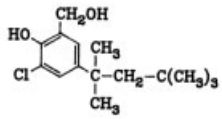

3-Chloro-2-hydroxy-5-isooctylbenzenemethanol

4. Aqueous Borate Solutions

Dilute borate solutions contain only monomeric species. Solutions of boric acid below 0.1 M contain monomeric boric acid molecules, $B(OH)_3$, and solutions of metaborate salts contain monomeric metaborate ions, $B(OH)_4^-$. Solutions made from mixtures of boric acid and metaborate salts contain a mixture of $B(OH)_3$ molecules and $B(OH)_4^-$ ions.

Polymeric borate anions form in more concentrated borate solutions. In addition to monomeric $B(OH)_3$ and $B(OH)_4^-$, concentrated borate solutions contain the polyions $B_3O_3(OH)_4^-$, $B_3O_3(OH)_5^{2-}$, $B_5O_6(OH)_4^-$, and $B_4O_5(OH)_4^{2-}$ (Fig. 2). These polyion species are all in rapid dynamic equilibrium and their distribution is dependent on the $Na_2O - B_2O_3$ molar ratio, temperature, and borate salt concentration, but is essentially independent of the cation [26].

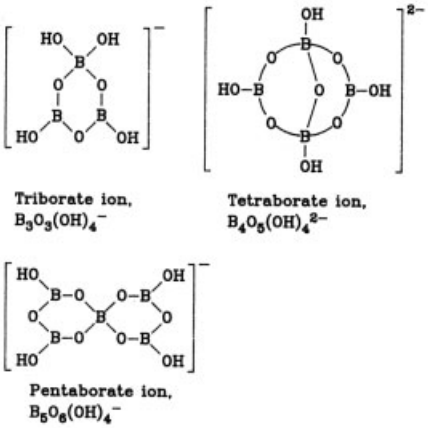

The pH of borate salt solutions varies with both concentration (see Table 2) and temperature as a consequence of these polyions. At the isohydric point the pH value is independent of borate concentration. This occurs at a pH of 8.9 and a $Na_2O - B_2O_3$ molar ratio of 0.41 for a sodium borate solution. Borate solutions with a $Na_2O - B_2O_3$ molar ratio less than 0.4 become more acidic at higher concentrations, whereas solutions with a $Na_2O - B_2O_3$ molar ratio greater than 0.4 become more alkaline with increasing concentration (Fig. 3). Borax, with a $Na_2O - B_2O_3$ molar ratio of 0.5 makes an excellent buffer solution.

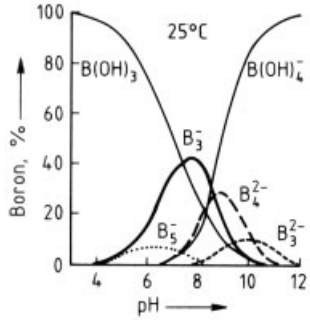

Figure 2. Distribution of borate polyanion species in aqueous solution:
B_3^- = $B_3O_3(OH)_4^-$,
B_3^{2-} = $B_3O_3(OH)_5^{2-}$,
B_5^- = $B_5O_6(OH)_4^-$,
B_4^{2-} = $B_4O_5(OH)_4^{2-}$ [20]

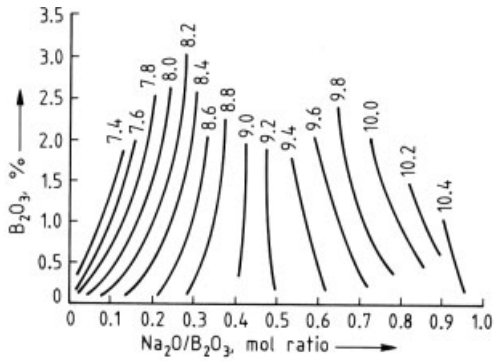

Figure 3. pH Values in the system $Na_2O-B_2O_3-H_2O$ at 25 °C [27]

5. Sodium Borates

The lack of a uniform system of nomenclature for naming metal borate salts has led to ambiguities in the past. Today metal borates are most commonly described by an empirical formula with the number of cations and boron atoms in the simplest stoichiometric units. The descriptive name is derived from this formula. In addition, an oxide formula is also in common use with the derived portions of metal oxide, boric oxide, and water in the simplest molar ratio. Many metal borates are also described using the mineral name, e.g., tincal for borax, $Na_2B_4O_7 \cdot 10\,H_2O$. Although useful for cataloging metal borates these formulas give no indication of the molecular structure of the borate anion found in the solid crystals. The solubilities of the borates described in this chapter are summarized in

Table 3. Of the sodium borates discussed, only borax decahydrate (Section 5.1) and borax pentahydrate (Section 5.2) are sold in large commercial volume.

5.1. Sodium Tetraborate Decahydrate

Sodium tetraborate decahydrate [*1303-96-4*], borax, tincal, $Na_2O \cdot 2\,B_2O_3 \cdot 10\,H_2O$ or $Na_2B_4O_7 \cdot 10\,H_2O$, crystals are odorless, white, monoclinic prisms with a relative density of 1.715. Based on the crystal structure, borax is best represented by the formula $Na_2\,[B_4O_5(OH)_4\,] \cdot 8\,H_2O$, with 2 mol of water existing as hydroxyl groups and 8 mol as water of crystallization [28]. The water in borax can be maintained at exactly 10 mol by storage in a closed container with a relative humidity produced by a solution saturated in both sucrose and salt [29].

Borax decahydrate slowly dehydrates at ambient conditions, depending on the relative humidity, to a water content of about 7 mol. The dehydration path is dependent on the thermal history of the sample [29]. Freshly crystallized borax previously heated to about 50 °C has a vapor pressure of 1.33 kPa at 19.8 °C and follows a stable dehydration path. That is, it dehydrates reversibly to borax pentahydrate at ambient temperature with a heat of dehydration of 54.17 kJ/mol. Further dehydration under vacuum produces an amorphous product containing 2 mol of water. Freshly crystallized borax decahydrate maintained below 50 °C has a vapor pressure of only 0.22 kPa at 20 °C and follows an unstable dehydration path. At ambient pressure and temperature over H_2SO_4 it undergoes irreversible dehydration leading directly to the amorphous hydrate (2 mol H_2O), bypassing borax pentahydrate. This hydrate can rehydrate first to borax pentahydrate and then to borax decahydrate.

Borax heated in a closed container melts in its own water of hydration at about 60 °C. Borax loses 5 mol of water by heating from 50 to 100 °C and loses an additional 3 mol up to 160 °C (determined by thermal gravimetric analysis). The remaining 2 mol of water are slowly lost during heating to about 400 °C. Anhydrous borax fuses at 742 °C forming a clear glass.

Rapid heating of borax decahydrate causes puffing [30]. Puffed borax consists of small glassy hollow spheres with a void volume as high as 90 % and a bulk density of about 0.07 g/cm^3.

Borax has a heat of formation of − 6264.3 kJ/mol [31], and a specific heat of 1.611 kJ kg^{-1}K^{-1} at 25 – 50 °C [32].

Borax decahydrate crystallizes from aqueous solution below 60.8 °C, the decahydrate – pentahydrate transition temperature (Fig. 4). This transition temperature is lowered by addition of either boric acid or sodium hydroxide [33]. In a solution saturated with sodium sulfate this transition temperature drops to 49.3 °C; with sodium chloride it drops to 39.6 °C; and when saturated with sodium chloride and potassium sulfate it drops to 35.5 °C. The solubility of borax in organic solvents is shown in Table 4.

The pH value of a borax solution increases slightly with increasing concentration (Table 2), and drops slightly with increasing temperature [34].

5.2. Sodium Tetraborate Pentahydrate

Sodium tetraborate pentahydrate[*12045-88-4*], tincalconite, $Na_2O \cdot 2\,B_2O_3 \cdot 5\,H_2O$ or $Na_2B_4O_7 \cdot 5\,H_2O$, crystals have rhombohedral symmetry and resemble octahedra; their relative density is 1.880. This pentahydrate has the same polyanion structure as borax decahydrate and is best represented by the formula $Na_2\,[B_4O_5\,(OH)_4] \cdot 3\,H_2O$, with 3 mol of water of crystallization and 2 mol of water existing as hydroxyl groups. Pure borax pentahydrate actually contains 4.75 mol rather than 5.0 mol of water because of apparent vacancies in the crystal structure [35].

Borax pentahydrate rapidly crystallizes from aqueous solution above 60.8 °C (Fig. 4). Solid pentahydrate in equilibrium with its saturated solution is metastable to kernite at all temperatures above 58.2 °C. When a solution saturated with borax pentahydrate is heated for one day or more, kernite crystals slowly form [36]. The solubility of tincalconite in organic solvents is shown in Table 4.

From thermogravimetric analysis 3 mol of water are lost on heating borax pentahydrate at 160 °C. The last 2 mol are slowly lost up to about

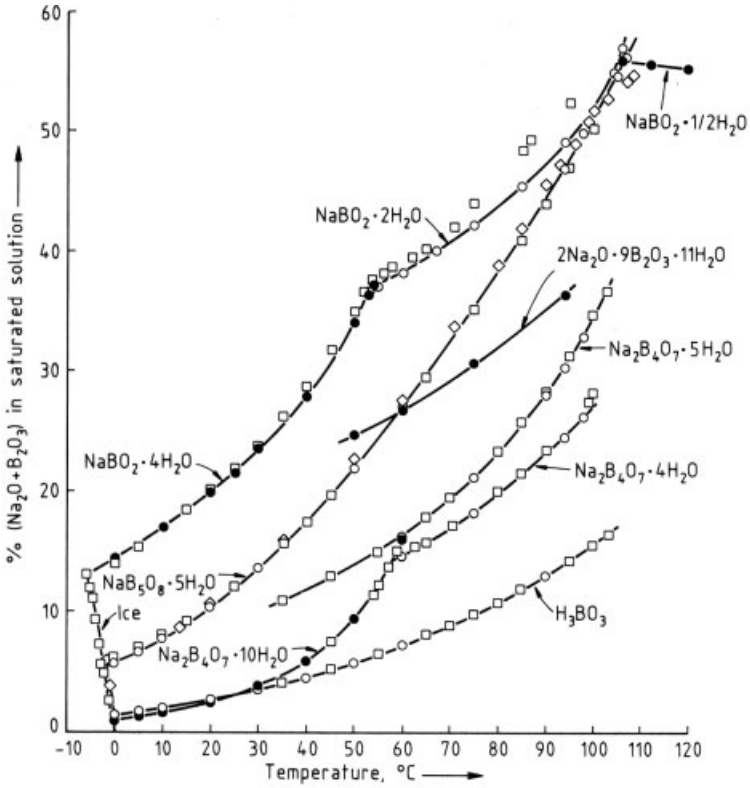

Figure 4. Solubility versus temperature curves for borate hydrates [33].

400 °C. Pentahydrate dehydrates reversibly to an amorphous dihydrate on heating at 80 °C and 0.3 kPa pressure or at 140 °C in air.

Like borax decahydrate, borax pentahydrate puffs when rapidly heated [30].

The heat of formation for borax pentahydrate is $-4\,784$ kJ/mol [31] and the specific heat is $1.318\,\mathrm{kJ\,kg^{-1}K^{-1}}$ [32].

Production. Refined borates are almost exclusively produced as a sodium salt, with a minor amount of potassium salt and even smaller amounts of lithium and ammonium salts.

Sodium borate ore consisting primarily of tincal and clay is crushed, mixed with Trona, and dissolved in weak liquor at 100 °C. The coarse gangue is separated on a screen and strong liquor and fine insolubles are transferred to a primary thickener. Clarified strong liquor is used to produce borax pentahydrate in continuous vacuum crystallizers at a temperature above the borax penta – decahydrate transition temperature of 60.8 °C. Borax decahydrate product crystallizes below this transition temperature. Crystals are filtered using a continuous centrifuge and dried in rotary driers. Borax crystals are washed with a weak boric acid solution leaving a very thin surface coat of pentaborate, which inhibits caking. Thickened underflow from the primary thickener is washed in a countercurrent fashion through four thickener stages with progressively weaker liquor in each stage. Fresh water is added to the last stage thickener. Liquid effluent is pumped to sealed solar ponds. The plant B_2O_3 capacity is 2030 t/d.

Calcium and sodium calcium borate ores were once a major source for borax production, but since the discovery of the high-grade borax deposit in California they are of less importance. Ulexite ($Na_2O \cdot 2\,CaO \cdot 5\,B_2O_3 \cdot 6\,H_2O$) is preferred over calcium borates and borosilicates because less alkali is required and less insoluble calcium salts are produced. Ulexite (ca. 38 % B_2O_3), which is finely ground and prefer-

ably calcined at 390 °C, is heated with aqueous sodium carbonate and sodium hydrogen carbonate to a near boil for several hours. The insoluble calcium carbonate is filtered and borax crystallized from the mother liquor in about 91 – 96 % yield.

Borax is recovered from upper lake brines by Kerr McGee at Searles Lake in California by heating and evaporating water to precipitate NaCl, Na_2CO_3, and Na_2SO_4. The hot liquor, which is nearly saturated in KCl, is then rapidly cooled in a vacuum crystallizer to precipitate KCl. The cool mother liquor, which is supersaturated with borax, is treated with $NaHCO_3$ to crystallize crude borax pentahydrate. This crude material is washed with brine and recrystallized as either borax deca- or pentahydrate. In a carbonation process borax is recovered from the lower brines by treatment with purified flue gas, causing $NaHCO_3$ to precipitate. The filtrate is neutralized with raw brine and cooled to crystallize crude borax.

A very pure grade of borax can be produced by reacting refined boric acid with hot NaOH solution and crystallizing the resulting borax.

5.3. Anhydrous Sodium Tetraborate

Anhydrous sodium tetraborate [*1330-43-4*], anhydrous borax, $Na_2O \cdot 2\,B_2O_3$ or $Na_2B_4O_7$. Fused borax forms an amorphous glass when rapidly cooled (density 2.36 g/cm^3). When slowly cooled orthorhombic crystals (α-form) are obtained. A β- and γ-form also exist [37]. The α-form is the most stable form with a congruent melting point at 742.5 °C. Its heat of formation has been calculated as − 3290 kJ/mol [38], and its vapor pressure is 0.73 kPa at 1200 °C, 1.73 kPa at 1250 °C, and 3.39 kPa at 1300 °C. Anhydrous borax dissolves more slowly in water than hydrated forms, giving off a large amount of heat; heat of hydration is 161 kJ/mol. The integral heat of solution in water to give a 2.5 wt % solution is − 192.6 J/g at 51.5 °C and 190.5 J/g at 60 °C. The solubility in several organic solvents is shown in Table 4.

The melt attacks most refractories at 800 °C. Many metal oxides are soluble in borax glass giving characteristic colors. Anhydrous borax reacts with carbon at 1200 °C in the absence of oxygen to form B_4C_3 and Na_2C_2, and reacts with metallic sodium to form elemental boron.

Production. Anhydrous borax is produced commercially by dehydration of hydrated sodium tetraborates in large fusion furnaces to form a borate glass. Anhydrous borax melts at 742 °C and this melt readily supercools to an amorphous glass. A crystalline form can be obtained by prolonged annealing or by seeding. The melt is very corrosive to furnace refractories and methods have been devised to keep borax away from the furnace walls [39].

5.4. Sodium Metaborate Tetrahydrate

Sodium metaborate tetrahydrate [*10555-76-7*], $Na_2O \cdot B_2O_3 \cdot 8\,H_2O$ or $NaBO_2 \cdot 4\,H_2O$, crystals are triclinic with a density of 1.743 g/cm^3. From X-ray studies the structure is best represented by the formula $NaB(OH)_4 \cdot 2\,H_2O$, containing $B(OH)_4^-$ ions [40]. Crystals begin to melt in their own water of hydration at 53.5 °C. At low humidity water is lost forming the dihydrate. By differential thermal analysis (DTA) 1 mol of water is lost at 130 °C, a second at 140 °C, and a third at 280 °C; the remaining water is slowly lost up to 800 °C. Heat of hydration is 52.51 kJ/mol.

Sodium metaborate tetrahydrate is the stable solid phase in contact with its saturated solution between 11.5 and 53.6 °C. Solutions are easily supersaturated. Both crystals and solutions absorb carbon dioxide from air forming borax and sodium carbonate. The pH of aqueous solutions increases with concentration (see Table 2).

Production. Sodium metaborate tetrahydrate is made commercially by crystallization (< 30 °C) from a solution containing borax and a slight excess of sodium hydroxide.

A solution of sodium metaborate, prepared from borax pentahydrate and sodium hydroxide, is reacted with hydrogen peroxide on a large commercial scale to produce sodium perborate tetrahydrate, $NaBO_3 \cdot 4\,H_2O$ or $Na_2O \cdot B_2O_3 \cdot 2\,H_2O_2 \cdot 6\,H_2O$, containing about 10 wt % available oxygen. Sodium perborate crystallizes in transparent monoclinic crystals, which are comparatively stable in air [41]. Its

aqueous solution slowly decomposes; decomposition is accelerated by catalysts or elevated temperature.

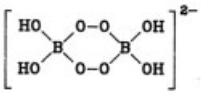

Perborate anion

5.5. Sodium Metaborate Dihydrate

Sodium metaborate dihydrate [*16800-11-6*], $Na_2O \cdot B_2O_3 \cdot 4\,H_2O$ or $NaBO_2 \cdot 2\,H_2O$, crystals are triclinic with a relative density of 1.909. The dihydrate is crystallized from aqueous solutions at 53.6 – 105 °C. When solutions are cooled rapidly very small needles, which are difficult to filter, form. Large platelike crystals form over several days during slow crystallization at 54 – 80 °C [42].

Sodium metaborate dihydrate slowly loses water at ambient temperatures and becomes sticky when heated. The heat of dehydration is 58.2 kJ/mol [43]. A monohydrate can be produced by dehydrating the dihydrate or by crystallization from highly basic solutions above 40 °C. Anhydrous sodium metaborate, $NaBO_2$, formed by heating to 350 °C, is very hygroscopic. At 966 °C it melts forming a clear glass.

5.6. Sodium Pentaborate Pentahydrate

Sodium pentaborate pentahydrate [*12046-75-2*], sborgite, $Na_2O \cdot 5\,B_2O_3 \cdot 10\,H_2O$ or $NaB_5O_8 \cdot 5\,H_2O$, crystals are triclinic with a density of 1.713 g/cm^3. The heat capacity, heat content, and entropy and other thermal functions have been reported from 0 to 345 K [44].

Sodium pentaborate is stable in contact with its saturated solution between 2 and 59.5 °C. Below 2 °C borax forms, and above 59.5 °C, $2\,Na_2O \cdot 9\,B_2O_3 \cdot 11\,H_2O$, Taylor's borate (*d* 1.903), slowly crystallizes if seeds are present. Supersaturated aqueous solutions of Taylor's borate easily form up to the boiling point. Still lower hydrates of 4, 2, 1, and 0 mol of water crystallize at temperatures above 100 °C under pressure [45].

Production. Sodium pentaborate can be made by gradually heating borax and boric acid (1 : 6 molar ratio) in a rotary drier from 30 to 66 °C [46]. It is also made by spray drying a borate solution of the appropriate composition.

5.7. Disodium Octaborate Tetrahydrate

Disodium octaborate tetrahydrate [*12280-03-4*], $Na_2O \cdot 4\,B_2O_3 \cdot 4\,H_2O$ or $Na_2B_8O_{13} \cdot 4\,H_2O$, is a spray-dry product. It is an amorphous solid, and is sold under the trade name Polybor; boron content is high (67.1 % B_2O_3), as is solubility in water. It is easily dissolved in cool water to give supersaturated solutions of considerable concentration. Above 60 °C, concentrated solutions become quite viscous. In aqueous solution the pH decreases with concentration (Table 2).

5.8. Disodium Tetraborate Tetrahydrate

Disodium tetraborate tetrahydrate [*12045-87-3*], kernite, $Na_2O \cdot 2\,B_2O_3 \cdot 4\,H_2O$ or $Na_2B_4O_7 \cdot 4\,H_2O$, forms monoclinic crystals (*d* 1.908) consisting of infinite linear chains of the polyanion $[B_4O_6(OH)_2]^{2-}{}_n$ made up of fused six-membered rings [47]. The heat of formation at 24 °C is − 4489.0 kJ/mol [31]. Crystal cleavage forms long fibrous needles.

Borax pentahydrate is metastable to kernite when in contact with its saturated aqueous solution above 58.2 °C (see Fig. 4). However, the rate of crystallization and of dissolution of kernite is very slow compared to borax pentahydrate. Calcined kernite dissolves at a considerably faster rate compared to the tetrahydrate.

Kernite hydrates irreversibly to borax decahydrate under ambient conditions. It dehydrates reversibly to the dihydrate, metakernite, over P_2O_5 in a vacuum or by heating at 100 – 120 °C. Further dehydration by heating to 190 °C causes irreversible dehydration forming an amorphous monohydrate. Rehydration of this monohydrate forms borax penta- and decahydrate.

Heating kernite to 180 °C results in the loss of 3 mol of water (DTA). The last mole of water is slowly lost up to 400 °C.

6. Other Metal Borates

6.1. Calcium and Calcium Sodium Borates

Calcium and calcium sodium borates are next to sodium borates in commercial importance. This includes the minerals colemanite, ulexite, and probertite.

Colemanite [*12291-65-5*] , $2\,CaO \cdot 3\,B_2O_3 \cdot 5\,H_2O$ or $Ca_2B_6O_{11} \cdot 5\,H_2O$, is the most widely occurring calcium borate mineral. Crystals of pure colemanite are monoclinic with a relative density of 2.42 and heat of formation of $-3\,469$ kJ/mol. Its crystal structure is represented by the formula $Ca_2B_6O_8(OH)_6 \cdot 2\,H_2O$ and is made up of polymeric chains of $[B_3O_4(OH)_3]^{2-}$ anions [48]. The solubility in water at 25 °C is about 0.18 wt % as B_2O_3 and 0.38 wt % at 100 °C. At about 350 °C, colemanite decrepitates violently losing all its water of hydration.

Ulexite [*1319-33-1*] , $Na_2O \cdot 2\,CaO \cdot 5\,B_2O_3 \cdot 16\,H_2O$ or $NaCaB_5O_9 \cdot 8\,H_2O$, forms triclinic needles with a density of 1.955 g/cm^3. The crystal structure consists of discrete pentaborate rings and is represented by the formula $NaCaB_5O_6\,(OH)_6 \cdot 5\,H_2O$ [49]. By DTA ulexite dehydrates losing 4 mol of water at 70 – 90 °C, and another 8 mol between 100 and 120 °C; the remaining water is gradually lost up to about 400 °C.

The solubility of ulexite in water is about 0.49 wt % as $NaCaB_5O_9$ at 25 °C. Calcination at 200 – 500 °C increases the solubility to 9 – 13 g/L [50].

Ulexite can be *prepared* by reacting borax, calcium metaborate, and water at 20 °C for 119 d or at 40 °C for 60 d [36]. Also moistened probertite heated at 80 – 100 °C forms ulexite in about four weeks [51].

Probertite [*12229-14-0*] , $Na_2O \cdot 2\,CaO \cdot 5\,B_2O_3 \cdot 10\,H_2O$ or $NaCaB_5O_9 \cdot 5\,H_2O$, exists as monoclinic crystals with a density of 2.14 g/cm^3. It has a crystal structure of $NaCaB_5O_7\,(OH)_4 \cdot 3\,H_2O$ and is made up of pentaborate rings linked in polymeric chains [49]. Probertite is a commonly found scale in borax refineries. By DTA 2 mol of water are lost at about 100 °C, 4 mol more between 100 and 180 °C, and the last 4 mol are slowly lost up to 400 °C. Probertite can be *prepared* by heating borax and calcium metaborate for 8 d at 105 °C [6].

6.2. Lithium, Potassium, and Ammonium Borates

Borates of lithium, potassium, and ammonium are of much smaller commercial volume than those described in Section 6.1.

6.2.1. Lithium Borates

Lithium metaborate octahydrate [*41851-38-1*] , $Li_2O \cdot B_2O_3 \cdot 16\,H_2O$ or $LiBO_2 \cdot 8\,H_2O$, with a trigonal crystal structure, is the stable solid in contact with saturated aqueous solutions below 36.9 °C [52]. Heating the crystals causes rapid loss of 6 mol of water below 70 °C, forming orthorhombic crystals of a dihydrate, $Li_2O \cdot B_2O_3 \cdot 4\,H_2O$ or $LiBO_2 \cdot 2\,H_2O$, with a density of 1.825 g/cm^3. This dihydrate is the stable solid in contact with its saturated solution above 36.9 °C.

Lithium pentaborate pentahydrate [*37190-10-6*] , $Li_2O \cdot 5\,B_2O_3 \cdot 10\,H_2O$ or $LiB_5O_8 \cdot 5\,H_2O$, has an incongruent solubility reported as both < 37.5 °C and 40.5 °C. By DTA the solid loses 3.7 mol of water when dehydrated between 110 and 200 °C and an additional 0.8 mol are lost up to 300 °C; it becomes anhydrous at 450 °C.

Dilithium tetraborate trihydrate [*37195-62-3*] , $Li_2O \cdot 2\,B_2O_3 \cdot 3\,H_2O$ or $Li_2B_4O_7 \cdot 3\,H_2O$, has a density of 1.88 g/cm^3. When crystallized from an aqueous solution, a variety of hydrates form; when boiled the trihydrate forms [52]. Dehydration occurs at 260 °C and fusion at about 890 °C.

6.2.2. Potassium Borates

Potassium tetraborate tetrahydrate [*12045-78-2*] , $K_2O \cdot 2\,B_2O_3 \cdot 4\,H_2O$ or $K_2B_4O_7 \cdot 4\,H_2O$, is the most common of this

class. Crystals are orthorhombic, having a relative density of 1.92 and a borate anion structure similar to borax [53]. The solid loses 2 mol of water between 112 and 180 °C, with a heat of dehydration of 86.6 kJ/mol; the last 2 mol are gradually lost between 180 and 420 °C. Potassium tetraborate is more soluble than the corresponding sodium tetraborate. In contact with its saturated solution above 56 °C, it is metastable to tetrapotassium decaborate pentahydrate, $2\,K_2O \cdot 5B_2O_3 \cdot 5\,H_2O$ or Auger's potassium borate. This Auger's form is very slow to crystallize in the absence of seed. The pH value of aqueous potassium tetraborate solutions increases slightly with concentration (Table 2).

Potassium tetraborate can be crystallized from aqueous solutions of potassium hydroxide and boric acid with a $B_2O_3 - K_2O$ molar ratio of 2.0. It can also be crystallized from borax solutions by addition of potassium chloride.

Potassium pentaborate [*12229-13-9*], $K_2O \cdot 5\,B_2O_3 \cdot 8\,H_2O$ or $KB_5O_8 \cdot 4\,H_2O$, crystals are orthorhombic with a relative density of 1.74, *mp* 780 °C, and heat capacity of 330.5 J mol^{-1} K^{-1} at 298 K [44]. The heat of dehydration is 110.8 kJ/mol between 106 and 134 °C. Potassium pentaborate is less soluble than the sodium analog. The pH value of its aqueous solution decreases with concentration (Table 2). Potassium pentaborate can be crystallized from a solution of potassium hydroxide and boric acid with a $B_2O_3 - K_2O$ molar ratio of about 5.0.

6.2.3. Ammonium Borates

Ammonium pentaborate octahydrate [*12229-12-8*], $(NH_4)_2O \cdot 5\,B_2O_3 \cdot 8\,H_2O$ or $NH_4B_5O_8 \cdot 4\,H_2O$, exists in two crystalline forms: an orthorhombic α-form with a relative density of 1.567 and heat capacity of 359.4 J mol^{-1} K^{-1} at 301.2 K, and a monoclinic β-form. The crystal structure of ammonium pentaborate has the same borate ion $[B_5O_6(OH)_4]^-$ found in sodium and potassium pentaborates. Unlike diammonium tetraborate the pentaborate is very stable to ammonia loss. At 150 °C all but 2 mol of water are lost, with less than 1 % ammonia loss [27]. At 200 °C and 0.1 kPa one more molecule of water is lost with only 2 % ammonia loss.

Ammonium pentaborate tetrahydrate can be crystallized from an aqueous solution of ammonia and boric acid with a $B_2O_3 - (NH_4)_2O$ molar ratio of 5.0 [54]. To determine its solubility in water, long equilibration times are required. For this reason solubility data published prior to 1966 are suspect [55]. The pH value decreases with solution concentration (Table 2).

Diammonium tetraborate tetrahydrate [*12228-87-4*], $(NH_4)_2O \cdot 2\,B_2O_3 \cdot 4\,H_2O$ or $(NH_4)_2B_4O_7 \cdot 4\,H_2O$, has a crystal density of 1.58 g/cm^3. The solid is somewhat unstable, with an appreciable vapor pressure of ammonia at ambient pressure and temperature. It is crystallized from a solution of ammonium hydroxide and boric acid with a $B_2O_3 - (NH_4)_2O$ molar ratio of about 2.0. Aqueous solutions up to 10 % concentration have an invariant pH of 8.8. The crystal structure can be described by the formula $(NH_4)_2B_4O_5(OH)_4 \cdot 2\,H_2O$ [56].

Other metal borates are of comparatively smaller commercial volume. They are generally prepared by reacting an aqueous solution of boric acid or alkali metal borate with a metal oxide or soluble metal salt.

6.3. Zinc Borates

Several hydrates of zinc borate are known. These include $2\,ZnO \cdot 3\,B_2O_3 \cdot 3.5\,H_2O$ or ZB 2335 [*12513-27-8*], which has a crystal density of 2.69 g/cm^3 and thermal stability to dehydration up to 290 – 300 °C [57]. It is made commercially from an aqueous boric acid solution with zinc oxide above 70 °C in the presence of product zinc borate seed. An induction period of 1 – 2 h is followed by a highly exothermic reaction.

A zinc borate of composition $2\,ZnO \cdot 3\,B_2O_3 \cdot 7$ to $7.5\,H_2O$ or ZB 237 [*12429-73-1*] and [*12280-01-2*], has a crystal density of 2.44 g/cm^3, and, from its X-ray structure, has the formula $Zn[B_3O_3(OH)_5] \cdot H_2O$ [58]. Six mol of water are lost when ZB 237 is heated from 120 to 170 °C.

The nine hydrate, $2\,ZnO \cdot 3\,B_2O_3 \cdot 9\,H_2O$ or ZB 239 is made using the same conditions as for ZB 2335 below 70 °C.

The zinc borate, $ZnO \cdot B_2O_3 \cdot 2H_2O$ or ZB 112 [*27043-84-1*], is thermally stable to dehydration up to 190 °C.

Several hydrated zinc borates are of commercial importance as fire retardant and smoke suppressant agents for polymers and coatings, most notably in PVC, halogenated polyesters, and nylon. For example, ZB 2335 is sold under the trade name Firebrake ZB or ZB 2335 (U.S. Borax & Chemical Corp., USA); it has a mean particle size of 5 – 10 µm. Its high-thermal stability is an important physical property because of the high temperatures required in some plastic processing. The zinc borate ZB 112 is sold by Humphrey Chemical Co. (USA) and Storey Brothers (U.K.). In addition Humphrey sells zinc borates described as ZB 235 and ZB 237, which are prepared by dry mixing boric acid with zinc oxide. The resulting product is actually a mixture of zinc borate $2ZnO \cdot 3B_2O_3 \cdot 9H_2O$ or ZB 239 and unreacted starting materials.

7. Borate Glasses

Various alkali-metal oxides, when fused with boric oxide, form a viscous melt, which on supercooling readily forms glasses (→ Glass). Mixtures of alkali-metal borates, boric acid, and alkali-metal carbonates or hydroxides can be used. Many physical properties of these glasses have maxima and minima with increasing alkali-metal oxide to borate ratio. This has been termed the *boron anomaly,* and has been observed in such properties, as viscosity, density, thermal expansion, and heat content [41]. This anomaly has been shown to be a result of structural changes in the glass, and of the variation in proportion of tetrahedral coordination of the boron in BO_4 groups and trigonal coordination in BO_3 groups.

The viscosity of $Na_2O-B_2O_3$ and $K_2O-B_2O_3$ glass has a maximum value at 20 – 25 mol % metal oxide below 900 °C [59], and the density increases from 1.85 g/cm^3 in pure B_2O_3 to a maximum of 2.36 in mixtures containing 27 – 33 mol % metal oxide at 900 °C [60]. Maximum hardness is achieved with a composition of 25 wt % Na_2O [61]. Thermal expansion reaches a minimum up to about 13 vol % Na_2O. The index of refraction (n_D) is 1.48 at 6 vol % Na_2O, and 1.51 at 24 vol % Na_2O.

Silica is quite soluble in molten sodium borate, giving a viscous liquid. The use of boric oxide in silica glass imparts many desirable properties. Borosilicate glass approaches the desirable properties of quartz glass, except the melting range and workability of the glass melt is significantly improved. Addition of boric oxide to alkali-metal silicate glass imparts increased hardness and decreased expansivity, lowers the melting range, and increases furnace production and durability of the glass. In ceramic glass, B_2O_3 improves resistance to weathering and chemicals.

8. Quality Specifications

Product specifications for boric oxide, boric acid, and borax are listed in Tables 5 and 6.

Table 5. Specifications for commercial B_2O_3

	% Maximum	% Typical
Na	0.1	0.015
SiO_2	0.21	0.15
Al_2O_3	0.14	0.08
Fe_2O_3	0.02	0.009
CaO	0.03	0.013
MgO	0.15	0.05
SO_4	0.41	<0.20

9. Uses

The glass and ceramic industry in the United States consumes about one-half of the total domestic B_2O_3. This includes borosilicate glasses, insulation fiberglass, textile fiberglass, porcelain enamels, and ceramic glazes. The largest market is for insulation fiberglass, with typically 5 – 7 % of contained B_2O_3 from borax pentahydrate, ulexite, or probertite; whereas textile fiberglass typically has 8 – 9 % of contained B_2O_3 from boric acid or colemanite. Borosilicate glass has 12 – 15 % of contained B_2O_3 largely from borax pentahydrate. Borates are used as a flux in the manufacture of porcelain enamels and ceramic glazes. Of the total United States domestic B_2O_3 consumption, about 15 % is used in the form of boric acid and metal borates as fire retardants in cellulosic insulation and plastics; 4 % as soluble borates in agriculture as a fertilizer to amend boron-deficient soils and as a herbicide at high dosages; 10 % from borax decahydrate

Table 6. Maximum specifications for grades of boric acid, borax, and borax pentahydrate, %

	Boric acid		Borax		Borax pentahydrate
	technical	special quality	technical	special quality	technical
$H_3BO_3 - Na_2B_4O_7 \cdot 10H_2O$	>99.5	>99.9	>99.5	>99.9	
SO_4	0.10	0.00015	0.06	0.0001	0.08
PO_4	–	0.001	0.001	0.001	–
Cl	0.01	0.00004	0.07	0.00004	0.05
Fe_2O_3	0.0007	0.0002	0.003	0.00028	0.004
Heavy metals as Pb	–	0.0002	–	0.001	–
Mn	–	0.0001	–	–	–
As	–	0.0005	–	0.0002	–
Ca	–	0.005	–	0.005	–
Na	–	0.001	–	–	–
Water insolubles	–	0.001	0.02	0.001	–

and sodium perborate for cleaning and bleaching of laundry products; and 2 % as corrosion inhibitors in aqueous systems, especially automotive antifreeze.

In Western Europe about one-half of the B_2O_3 is used in the manufacture of glass-related products. Sodium perborate accounts for about one-third of the total boron consumption as a bleach in high-temperature laundry products.

In Japan 70 % of B_2O_3 consumption is used in glass-related products.

Other minor uses of boric acid and metal borates are listed below.

Boric acid
- Flux in welding and brazing
- Bacteriostat and fungicide
- NF grade as a mild, nonirritating antiseptic
- Cockroach control
- Protection of wood products against insect damage
- Use in nuclear reactor cooling water
- Manufacture of refractories and abrasives
- Catalysts for air oxidation of hydrocarbons

Sodium borates, Calcium borates
- Manufacture of refractories and abrasives
- Metallurgical flux (smelting)
- Neutron absorber in nuclear reactors

Sodium metaborate
- Photographic chemical
- Herbicides
- Detergents and cleansers
- Textile-finishing compounds
- Adhesives

Potassium borates
- Non-sodium alkaline borate source
- Lubricant component
- Solvent for casein
- Welding and brazing flux for stainless steels and various nonferrous metals

Ammonium tetraborate
- Non-alkaline metal borate source

Ammonium pentaborate
- Component in electrolytes for electrolytic capacitors
- Flameproofing formulations
- Paper coating

Lithium meta- and tetraborates
- Glass-making

Barium borates
- Fire retardants
- Mildew inhibitor in latex, paints, plastics, textile, and paper products
- Preservative in protein-based glues

Copper metaborate
- Rot and mildew inhibitor
- Fungicide in lumber and other cellulosic materials
- Oil pigment

Manganese borate
- Printing ink drier

10. Economic Aspects

The estimated world reserves of borate minerals and world production of boron products for 1980 through 1982 are listed in Table 7. Of the total world borates the United States produces

62 %, Turkey 29 %, and all other sources 9 %, based on B_2O_3. World borate consumption in the United States is 30 %, Western Europe 37 %, Japan 5 %, and all other countries 28 %. The European and Japanese borate requirements are served through imports largely from the United States and Turkey. Almost all B_2O_3 production in Turkey and about one-half the production in the United States is exported.

Table 7. Estimated world borate reserves and mine production (as B_2O_3)

Country	Reserves, kt	Production, kt		
		1980	1981	1982
United States	70000	710	670	550
Turkey	70000	339	320	320
South America	35000	–	–	–
Argentina	–	–	30	27
Chile	–	–	1	4
Peru	–	–	3	3
Soviet Union	65000	–	40	40
China	30000	–	5	5
World total	270000	1049	1069	949

Nearly 99 % of the sales volume of borate products consists of borax pentahydrate, borax decahydrate, boric acid, boric oxide, anhydrous borax, colemanite, and ulexite. Borax pentahydrate is the highest volume commercial borate. Until mid-1984, U.S. Borax exported a crude borax pentahydrate, trade name Rasorite-46, containing less than 1 % clay. This product is now exported without clay as borax pentahydrate. Prices of the primary borate products are shown in Table 8.

Table 8. Borate prices [62]

Product	$/t
Borax, decahydrate, technical, granular, 99.5 %, bulk carlots	193 *
Borax, pentahydrate, technical, granular, 99.5 %, bulk carlots	222 *
Borax, anhydrous, technical, 99 %, bulk carlots	622 *
Boric acid, technical, granular, 99.9 %, bulk carlots	609 *
Tincal	148 – 185 **
Ulexite	120 – 170 **
Colemanite, glass quality	255 – 325 **
Colemanite, refining grade	160 – 190 **

* FOB (free on board) rail cars in California, United States.
** FOB ship Bandirma, Turkey

There has been a significant decline in production of anhydrous products because of the rapid increase in fuel costs. In addition there is a trend to use colemanite for the manufacture of textile fiberglass.

11. Toxicology and Occupational Health

Boron is present in animal tissue at about 1 ppm, but has no known essential biochemical function [63]. Humans absorb about 10 – 20 mg of boron in a normal daily diet, mainly from fruits and vegetables. Inhalation of boric acid and borate dust causes only irritation without permanent injury. It is not absorbed when contacted with healthy skin, but can be toxic when used on large areas of burned or abraded skin. The handling of boric acid and sodium borates is not generally considered hazardous [66].

Boric acid has an acute oral LD_{50} (rats) of 4 g/kg; for borax the oral LD_{50} (rats) is 6 g/kg. The TLV and MAK data are given in Table 9 [64], [65].

Table 9. Maximum concentrations in the work environment

Substance	TWA, mg/m^3	STEL, mg/m^3	MAK, mg/m^3
Sodium tetraborates [*1303-96-4*]			
Anhydrous	1	–	–
Decahydrate	5	–	–
Pentahydrate	1	–	–
Boron oxide [*1303-86-2*]	10	20	15

Boron is an essential trace element in higher plant nutrition [67]. At high levels boron becomes toxic, with damage manifested in deformation of leaves, which curl at the edge, often accompanied by necrosis. Fertilizers with calcium tend to reduce this damage [31].

12. References

1. R. C. Erd: *Supplement to Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* **vol. 5,** Longman, New York 1980, Part A, Section A 1, p. 7.
2. T. Dickson, *Ind. Miner. (London)* **12** (1983) 63.
3. F. Dachille, R. Roy, *J. Am. Ceram. Soc.* **42** (1959) 78. J. D. Mackenzie, W. F. Claussen, *J. Am. Ceram. Soc.* **44** (1961) 79.
4. D. White, D. E. Mann, P. N. Walsh, A. Sommer, *J. Chem. Phys.* **32** (1960) 481, 488. H. J. Emeleús, A. G. Sharpe: *Advances in*

Inorganic Chemistry and Radiochemistry, vol. 5, Academic Press, New York 1963.
5. P. Wulff, S. K. Majumdar, *Z. Phys. Chem.* **B 31** (1936) 319.
6. H. Menzel, H. Schultz, *Z. Anorg. Chem.* **245** (1940) 157.
7. A. Cousen, W. E. S. Turner, *J. Chem. Soc.* 1928, 2654.
8. E. J. Prosen, W. H. Johnson, F. Y. Pergiel, *J. Res. Nat. Bur. Stand.* **62** (1959) 43. *Chem. Abstr.* **53** (1959) 15746.
9. W. F. Claussen, J. D. Mackenzie, *J. Am. Chem. Soc.* **81** (1959) 1007.
10. S. T. Benson, G. S. Parks, *J. Phys. Chem.* **35** (1931) 2091.
11. A. Buechler, J. B. Berkovitz-Mattuk, *J. Chem. Phys.* **39** (1963) 286.
12. D. L. Hildebrand, W. F. Hall, N. D. Potter, *J. Chem. Phys.* **39** (1963) 296.
13. K. Yu. Kazema, A. K. Lindpere, L. A. Luts, *Tr. Inst. Fiz. Astron., Akad. Nauk Est., SSR* **7** (1958) 34. *Chem. Abstr.* **54** (1960) 19188.
14. F. T. Greene, J. L. Margrane, *J. Phys. Chem.* **70** (1966) 2112.
15. Reference 1, J. Dulat, Section A 4, p. 186.
16. J. C. Southard, *J. Am. Chem. Soc.* **63** (1941) 3147.
17. H. T. Schlesinger, H. C. Brown, A. E. Finholt, *J. Am. Chem. Soc.* **75** (1953) 205.
18. R. Flowers, R. J. Gillespie, J. V. Oubridge, *J. Chem. Soc.* 1956, 1925.
19. Olin Mathieson Chemical Corp., US 2867502, 1959 (H. Stange, S. L. Clark).
20. F. C. Kracek, G. W. Morey, H. E. Merwin, *Am. J. Sci.* **35 A** (1938) 143.
21. G. G. Manov, N. J. DeLollis, S. F. Acree, *J. Res. Nat. Bur. Stand.* **33** (1944) 287. *Chem. Abstr.* **39** (1945) 1348.
22. D. White, D. E. Mann, P. N. Walsh, A. Sommer, *J. Chem. Phys.* **32** (1960) 488. D. J. Meschi, W. A. Chupka, J. Berkowitz, *J. Chem. Phys.* **33** (1960) 530. S. P. Randall, J. L. Margrave, *J. Inorg. Nucl. Chem.* **16** (1960) 29.
23. M. V. Stackelberg, J. Dressel, *Z. Electrochem.* **43** (1937) 14.
24. J. S. McGarth, G. G. Stack, P. A. McCusker, *J. Am. Chem. Soc.* **66** (1944) 1263.
25. Reference 1, D. I. Prichard, E. V. Reid, Section A 10,p. 502.
26. C. G. Salentine, *Inorg. Chem.* **22** (1983) 3920.
27. United States Borax Research Corp., unpublished work.
28. N. Morimoto, *Mineral J. (Sapporo)* **2** (1956) 1. *Chem. Abstr.* **52** (1958) 12687.
29. H. Menzel, H. Schulz, L. Sieg, M. Voigt, *Z. Anorg. Chem.* **224** (1935) 1.
30. Stokely-Van Camp, Inc., US 4412978, 1983 (R. T. Ertle).
31. D. D. Wagman, W. H. Evans, E. J. Prosen, *Nat. Bur. Stand. Circ. (U.S.)* **500** (1952) 481.
32. S. Scholle, M. Szmigielska, *Chem. Prum.* **15** (1965) 530. *Chem. Abstr.* **63** (1965) 15629.
33. N. P. Nies, R. W. Hulbert, *Chem. Eng. Data* **12** (1967) 303.
34. "pH Values of Borate Solutions at 20 °C," U.S. Borax Product Catalog.
35. C. Giacorazzo, S. Menchetti, F. Scordari, *Am. Mineral.* **58** (1973) 523.
36. H. Menzel, H. Schulz, *Z. Anorg. Chem.* **245** (1940) 157.
37. G. W. Morey, H. E. Merwin, *J. Am. Chem. Soc.* **58** (1936) 2248.
38. G. Grenier, D. White, *J. Phys. Chem.* **61** (1957) 1681.
39. American Potash & Chemical Corp., US 2177766, 1940 (L. G. Black).
40. S. Block, A. Perloff, *Acta Crystallogr.* **16** (1963) 1233.
41. A. Hansson, *Acta Chem. Scand.* **15** (1961) 934.
42. United States Borax & Chemical Corp., FR 1379600, 1964 (N. P. Nies, P. F. Jacobs).
43. H. Menzel, H. Schulz, *Z. Anorg. Chem.* **251** (1943) 167.
44. G. T. Furukawa, M. L. Reilly, J. H. Piccirelli, *J. Res. Nat. Bur. Stand.* **68 A** (1964) 381. *Chem. Abstr.* **61** (1964) 7771.
45. A. Benhassaine, *C.R. Acad. Sci. Ser. C* **274** (1972) 1188. *Chem. Abstr.* **76** (1972) 158924.
46. United States Borax & Chemical Corp., US 3010786, 1961 (D. S. Taylor, N. P. Nies).
47. G. Cialdi, E. Corazza, C. Sabelli, *Atti. Accad. Naz. Lincei* **42** (1967) 236. *Chem. Abstr.* **67** (1967) 77039.
48. C. L. Christ, J. R. Clark, H. T. Evans, *Acta Crystallogr.* **11** (1958) 761.
49. J. R. Clark, C. L. Christ, *Am. Mineral.* **44** (1959) 712. J. R. Clark, D. E. Appleman, *Am. Mineral.* **50** (1965) 283.
50. H. Gode, A. Veveris, *Latv. PSR Zinat. Akad. Vestis, Kim. Ser.* 1970 722. *Chem. Abstr.* **74** (1971) 71020.
51. R. M. Adams (ed.): *Boron, Metallo-Boron Compounds and Boranes,* Interscience, New York 1964. *Chem. Abstr.* **61** (1964) 15652.
52. W. T. Reburn, W. A. Gale, *J. Phys. Chem.* **59** (1955) 19.
53. M. Marezio, H. A. Plettinger, W. H. Zachariasen, *Acta Crystallogr.* **16** (1963) 975.

54. H. A. Lehmann, W. Schmidt, *Z. Chem.* **5** (1965) 111.
55. H. A. Lehmann, W. Schmidt, *Z. Chem.* **5** (1965) 65.
56. G. B. Bokii, V. B. Kravchenko, *Zh. Strukt. Khim.* **7** (1966) 920. *Chem. Abstr.* **66** (1967) 49813.
57. United States Borax & Chemical Corp., US Reissue 27424, 1972 (N. P. Nies, R. W. Hulbert). United States Borax & Chemical Corp., US 3649172, 1972 (N. P. Nies, R. W. Hulbert).
58. J. Ozols, I. Tetere, A. Ievins, *Latv. PSR Zinat. Akad. Vestis Kim. Ser.* **1** (1973) 3. *Chem. Abstr.* **78** (1973) 165580.
59. B. E. Warren, *J. Appl. Phys.* **13** (1943) 602. *Tech. Rev.* **43** (1943) 253, 273. *Chem. Abstr.* **37** (1943) 239.
60. G.-J. Su, P.-C. Li, A. C. Ghose, Wright Air Development Division (WADD) Technical Report 59 – 21, Rochester University, April, 1960.
61. F. C. Eversteyn, J. M. Stevels, H. I. Waterman, *Phys. Chem. Glasses* **1** (1960) 134. *Chem. Abstr.* **55** (1961) 4905.
62. E. W. Fajans, *Mining Annual Review* 1984 (June) 102, published by Min. J. (London).
63. N. N. Greenwood: "Boron," in: *Comprehensive Inorganic Chemistry,* **vol. 1,** Pergamon Press, New York 1973, pp. 665 – 991.
64. ACGIH (ed.): *TLVs for Chemical Substances,* ACGIH, Cincinnati, Ohio, 1982.
65. DFG (ed.): *Maximale Arbeitsplatzkonzentrationen,* Verlag Chemie, Weinheim 1982.
66. R. S. Fischer, H. C. Freimuth, *J. Invest. Dermatol.* **30** (1978) 85. M. Pasdeloup, C. Brisson, *Org. Magn. Reson.* **16** (1981) 164.
67. J. A. Rajaratnam, J. B. Lowry, P. N. Avadhani, R. H. V. Corley, *Science* **172** (1971) 11421.

Borides → Boron Carbide, Boron Nitride, and Metal Borides

Boron and Boron Alloys

ULRICH BAUDIS, Degussa, Hanau, Federal Republic of Germany (Chap. 1)

RUDOLF FICHTE, Nürnberg, Federal Republic of Germany (Chap. 2)

1. Boron 1
1.1. Structure and Polymorphism 1
1.2. Physical Properties 2
1.3. Chemical Properties 2
1.4. Production 3
1.5. Chemical Analysis, Storage, Safety 4
1.6. Uses, Economic Aspects 5
2. Ferroboron and Other Boron Master Alloys 5
2.1. Physical Properties 5
2.2. Chemical Properties 6
2.3. Raw Materials 6
2.4. Production 7
2.4.1. Carbothermic Production 7
2.4.2. Aluminothermic Production 8
2.5. Quality Specifications for Commercial Grades 8
2.6. Storage and Shipment 10
2.7. Chemical Analysis 11
2.8. Pollution Control 11
2.9. Uses 11
2.10. Economic Aspects 12
2.11. Toxicology and Industrial Hygiene . 12
3. References 12

1. Boron

The nonmetallic element boron [*7440-42-8*] is situated in the top position of the third main group in the periodic table. Boron is not found free in nature but is always bound to oxygen. It occurs as orthoboric acid and as alkali-metal and alkaline earth metal borates. The most important sources of boron are the minerals rasorite and kernite, which are found in extensive deposits in California in the Mojave desert. There are also large ore deposits in Turkey and the USSR. Although widespread in nature, boron constitutes only about 3 ppm of the earth's crust [9].

Boron was discovered concurrently by the French chemist GAY-LUSSAC and the English chemist SIR HUMPHRY DAVY in 1808. HENRI MOISSAN [10] first obtained considerable quantities of boron (of 86 % purity) by reduction of boric oxide with magnesium in 1895. The Moissan process is still the basis for commercial production of amorphous low-purity boron.

In 1909 WEINTRAUB [11] succeeded in preparing 99 % pure boron by decomposition of BCl_3 in an electric arc. Since then numerous methods have been developed (see Section 1.4), but still research continues to find methods for producing commercial quantities of pure boron.

1.1. Structure and Polymorphism

There are several allotropic forms of boron. Well established are amorphous boron, α-rhombohedral boron, and β-rhombohedral boron. In addition, four tetragonal forms have been described. However, these are probably stabilized by small amounts of nitrogen or carbon [4].

The β-rhombohedral form is the thermodynamically stable modification at all temperatures. Amorphous boron slowly converts to the β-rhombohedral form at $\approx 1200\,^{\circ}C$ and to α-rhombohedral boron above $1500\,^{\circ}C$. Any type of boron recrystallizes in β-rhombohedral structure when heated above the melting point and cooled.

The unit cell of β-rhombohedral boron contains 105 boron atoms grouped $84\,B + 2\times 10\,B + 1\,B$ in a complex arrangement. The structure of α-rhombohedral boron can be described as a slightly deformed cubic close packing of B_{12} icosahedra. The unit cell contains 12 B atoms at the vertices of the icosahedron.

Amorphous boron exhibits broad reflections in the x-ray diffraction pattern. It is believed to be a frozen-in intermediate between the α- and β-modifications or a microcrystalline deposit of β-rhombohedral boron [4]. Amorphization can be achieved by intensive milling of β-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_281

rhombohedral samples [12]. Structures of the other polymorphs are too complex to be described in a few words; however, all boron structures have been discussed thoroughly elsewhere [4].

1.2. Physical Properties

Determination of precise physical properties of boron is made more complicated by the structural polymorphism and purification problems. Some values, rather arbitrarily collected from the literature, are listed below.

Atomic number	5	
Relative atomic mass	10.811	
Melting point	$2050 \pm 50\,^{\circ}C$	
Sublimation point	$2550\,^{\circ}C$	
Density ($20\,^{\circ}C$)	$2.3\,g/cm^3$	amorphous
	2.35	β-rhombohedral
	2.46	α-rhombohedral
	2.99	liquid at *mp*
	2.13	solid at *mp*
Crystal structure	amorphous, α-rhombohedral, β-rhombohedral, four tetragonal	
Hardness (Knoop)	$2390\,kg/mm^2$	crystallized from melt
	2690	vapor deposited
Electrical resistivity (300 K)	$7.5 \times 10^2\,\Omega\,cm$	amorphous
	7×10^5	β-rhomb., single crystal
	$10^6 - 10^7$	β-rhomb., polycrystalline
Heat capacity C_p	$12.054\,J\,K^{-1}\,mol^{-1}$	amorphous at 300 K
	11.166	β-rhomb. 300 K
	33.955	solid at *mp*
	39.063	liquid at *mp*
Entropy S (298 K)	$6.548\,J\,K^{-1}\,mol^{-1}$	amorphous
	5.875	β-rhomb.
Enthalpy of fusion ΔH_f	50.2 kJ/mol	
Enthalpy of sublimation ΔH_s	572.7 kJ/mol	

Boron is the second hardest element, the diamond allotrope of carbon being the hardest. α-Rhombohedral boron is red to brown; β-rhombohedral boron lustrous gray black; and amorphous material is brown to gray. The electrical resistivity changes drastically with temperature, varying from $10^{11}\,\Omega\,cm$ at $-160\,^{\circ}C$ to $10^6\,\Omega\,cm$ at $20\,^{\circ}C$ and to $0.1\,\Omega\,cm$ at $700\,^{\circ}C$ [4] for polycrystalline β-rhombohedral boron, be havior characteristic of a semiconductor. Detailed information about optical, magnetic, electrical, thermodynamic, and other properties of the various boron polymorphs are given in the literature [4], [5].

Boron is composed of two stable isotopes, ^{10}B [*14798-12-0*] and ^{11}B [*14798-13-1*]. Their nuclear properties are summarized in Table 1. The nuclear spin permits NMR spectroscopy. The high value of thermal neutron cross section of the isotope ^{10}B makes boron useful as a neutron absorber in nuclear power plants.

Table 1. Nuclear properties of boron

Property	Nuclide	
	^{10}B	^{11}B
Relative atomic mass	10.0129	11.0093
Natural abundance, %	19.8	80.2
Nuclear spin	+3	−3/2
Magnetic moment, nuclear magnetons*	+1.80063	+2.68857
Quadrupole moment, barn**	+0.074	+0.036
Cross section for (n, α), barn**	3837	0.005

* 1 nuclear magneton $= 5.0508 \times 10^{-27}\,A\,m^2$.
** 1 barn $= 10^{-28}\,m^2$.

1.3. Chemical Properties

The chemical properties of boron are determined to a large extent by its small atomic radius (0.25 nm) and high ionization potentials.

Ionization energy		
$B \longrightarrow B^+$	798 kJ/mol	(8.27 eV)
$B^+ \longrightarrow B^{2+}$	2426	(25.15)
$B^{2+} \longrightarrow B^{3+}$	3658	(37.92)
Standard electrode potential		
$B + 3\,H_2 \rightleftharpoons OH_3BO_3 + 3\,H^+ + 3\,e^-$	−0.73 V	
Electron affinity	32 kJ/mol (0.332 eV)	
Electronegativity	2.04 (Pauling)	
	2.01 (Mulliken)	
Ionic radius	0.25 nm	
Atomic radius	0.80 – 0.95 nm (depending on type of bonding)	
Standard enthalpy of formation		
BF_3	−1136 kJ/mol	
BCl_3	−402	
BBr_3	−239	
B_2O_3	−1269	
BN	−256	

The electronic configuration $2s^2 2p^1$ dictates predominant trivalency. However, simple B^{3+} ions do not exist.

Because boron has more orbitals available for bonding than electrons (4 orbitals, 3 electrons), it is an electron-pair acceptor, a Lewis acid, and it has a tendency to form multi-center bonds. Other dominant characteristics are the high affinity for oxygen and the tendency to combine with most metals to give refractory alloy-like metal borides.

The chemical behavior of elemental boron depends upon the morphology and particle size. Generally speaking, crystalline boron is rather unreactive, whereas amorphous boron does react more readily. At room temperature all modifications of boron are relatively resistant to chemical attack.

No direct reaction occurs with hydrogen, but indirect synthesis leads to a variety of boron hydrides (→ Boron Compounds).

The halogens react to form the trihalides:

$$2\,B + 3\,X_2 \longrightarrow 2\,BX_3$$

Fluorine reacts spontaneously at room temperature; chlorine, at 500 °C; and bromine, at 600 °C [13].

Reaction with oxygen starts at ≈ 600 °C but is restrained by the formation of a glassy liquid B_2O_3 film on the boron particles [14]. Boron reacts with sulfur to form glassy B_2S_3 and with Se to form B_2Se_3, both at ≈ 600 °C. There is no reaction with Te. With nitrogen, boron nitride, BN, is formed at 1100 °C. Phosphorus reacts completely at 1000 °C to give BP. Arsenic forms BAs at 800 °C under pressure.

Boron and carbon react to give $B_{12}C_3$ above 2000 °C; with silicon, B_4Si and B_6Si are formed. Heterocyclic compounds containing C, N, S, O, and boron have interesting structures and properties (see Carboranes, Borazine, Borthiine, and Boroxine under → Boron Compounds). There is also a large variety of organoboron compounds [15].

When it does react, elemental boron is an effective reducing agent. Water vapor is reduced at 800 °C to give H_2 and B_2O_3. Carbon monoxide is reduced at 1200 °C; boron burns in carbon dioxide at red heat. Nitrogen oxides are reduced to give BN and B_2O_3. Sulfur dioxide is also reduced. When transition metal oxides like Fe_2O_3, TiO_2, and Cr_2O_3 are reacted with elemental boron, B_2O_3 and metal borides form. Up to 1400 °C alkaline-earth oxides and aluminum oxide do not react with boron.

Boron is not attacked by aqueous NaOH but reacts completely with molten Na_2CO_3 or molten mixtures of sodium carbonate and sodium nitrate.

Boiling nonoxidizing acids, such as HF, HCl, HBr, or dilute sulfuric or phosphoric acid, do not attack boron. However, vigorous reaction occurs with concentrated HNO_3 or with HNO_3– H_2O_2 mixtures.

The chemical properties of boron are summarized in [17].

1.4. Production

A variety of preparative methods have been described. The following are the most important methods [6]:

1) Reduction of boric oxide with magnesium
2) Reduction of boron halides or fluoroborates with sodium or another metal
3) Reduction of boron halides with hydrogen
4) Thermal decomposition of boron compounds, especially boron halides and hydrides
5) Electrolysis of molten borates or fluoroborates

The modification of boron formed depends primarily on temperature and reaction time, only secondarily on the method [18]. As a rule of thumb, temperature below 900 °C and short reaction time produce amorphous boron; temperature above 1400 °C and long reaction time produce β-rhombohedral or the tetragonal modifications. The optimum conditions for formation of α-rhombohedral boron are in between.

The usual method for commercial production of large quantities of amorphous boron is the *reduction of boron trioxide with magnesium* in a thermite-like reaction [19–21]. The principles of this method date back to the work of Henri Moissan [10].

$$B_2O_3 + 3\,Mg \longrightarrow 2\,B + 3\,MgO$$

Simultaneously MgO reacts with excess B_2O_3:

$$MgO + B_2O_3 \longrightarrow Mg(BO_2)_2$$
$$MgO + 2\,B_2O_3 \longrightarrow MgB_4O_7$$

The reduction is very quick and highly exothermic; finely divided material may react explosively [7].

Reaction is smoother if there is an excess of B_2O_3 [7], [22]:

$$2\,B_2O_3 + 3\,Mg \longrightarrow 2\,B + Mg_3(BO_3)_2$$
$$4\,B_2O_3 + 3\,Mg \longrightarrow 2\,B + 3\,Mg(BO_2)_2$$

The optimum ratio B_2O_3 : Mg is about 1.8 : 3 [7]. Reaction is carried out in vertical steel retorts shielded from oxygen by a flow of argon. It is initiated by electric spark and igniter mixture or by external heating.

After cooling, the reaction mass is crushed to 1-mm particles and leached twice with hydrochloric acid. Crude amorphous boron of 86 – 88 % purity is obtained [21]. This material can be upgraded by treatment with B_2O_3 [22] or KHF_2 and KBF_4 [23], [24], subsequent leaching with acid, and final heating in a vacuum to remove boron suboxide and metals [25]. The boron content is then 90 – 92 or 95 – 97 %, adequate for use in pyrotechnics and fluxes.

Reduction of KBF_4 by sodium

$$KBF_4 + 3\,Na \longrightarrow 3\,NaF + KF + B$$

was used for the commercial production of boron in Germany until the end of the 1950s [26].

A common problem of metallothermic reduction is incomplete reaction and formation of nonremovable metal borides.

Samples of very pure boron, > 99 % B, can be obtained by *reduction of boron halides with hydrogen*, especially BBr_3 [27], [28] and BCl_3 [29–32]. This is also the method of choice for laboratory synthesis [27], [29]. The halides can be purified by distillation prior to reduction [4], [27]. However, the efficiency of the hydrogen reduction process is rather low [19], yields of 5 – 25 % being the usual. Unreacted boron halide must be recycled [30], washed out, or removed by other methods [33]. This complicates the process and makes it expensive.

Very pure boron is also obtained by *thermal decomposition* of BI_3 [34], [35], BBr_3 [36], or boron hydrides [37], [38] on tungsten wires or another type of incandescent filament. Boron of 99.9999 % (six nines) purity was obtained by decomposition of diborane and subsequent zone melting [39].

Electrolysis of molten borates or KBF_4 [40] is not an important method for commercial or laboratory preparation although it has been mentioned in numerous articles.

Boron can be refined either by zone melting [6], [35], [39], [41], [42] or volatilization of impurities in high vacuum or hydrogen at 2000°C [25], [35].

1.5. Chemical Analysis, Storage, Safety

Analysis. Boron content in elemental boron is generally *determined quantitatively* by decomposition to borate with fused carbonate in a platinum crucible, subsequent acid leaching, neutralization, addition of mannitol, and titration of boric acid. The mixture HNO_3–H_2O_2 also decomposes boron samples. Metal impurities in boron, such as Mg, Na, and Fe, are determined by atomic absorption or emission spectrometry after decomposition of the sample [43]. Chlorine and fluorine are determined by titration [44]. The elements O, N, H, P, Si, and C are difficult to determine. However, vacuum hot extraction and combustion methods [42], [43] or spark source mass spectrography [35], [45] can be used.

A sensitive *qualitative* test for boron is the green flame test. The sample is usually decomposed to methyl borate by heating a mixture of the sample, sulfuric acid, and methanol to ignition.

Storage. Dry boron is not very sensitive to oxidation. It is *packaged* in airtight metal drums, cardboard or plywood containers, or glass bottles [46]. Boron powder, like other combustible material, presents a dust explosion hazard. The lower explosive limit was determined to be 125 g/m^3 in air [47].

Boron ignites in air at 600 – 800 °C but does not burn violently because of the formation of a glassy boric oxide film that prevents the material from oxidizing rapidly [14]. Burning boron should be covered with sand, aluminum oxide, sodium chloride, or the like. Water must not be used, because of the formation of inflammable hydrogen. Boron is an explosion hazard when

mixed with strong oxidizing agents, such as nitrites, chlorates, lead dioxide, and silver fluoride. Such mixtures can react violently and even explode [13], [14], [16].

Safety. Elemental boron is considered to be nontoxic because of its chemical inertness and insolubility. In contrast, most boron compounds, especially the hydrides, are extremely toxic [48] (see Toxicology under → Boron Compounds). During the production process leading to boron the possible formation of any toxic boron hydrides must be either avoided or controlled.

1.6. Uses, Economic Aspects

Relatively large quantities of amorphous boron (90 and 95 % B) are used as additives in pyrotechnic mixtures, especially in flares, igniters, and delay compositions (→ Pyrotechnics); solid rocket propellant fuels [49]; and explosives [50]. Some boron is needed for the preparation of refractory metal boride additives for cemented carbides. Boron is also a useful reducing additive in fluxes for soldering stainless steel.

Use of boron in electronics is limited to high-purity boron (> 99.99 %). It is used as a ppm additive for germanium and silicon to make *p*-type semiconductors [51]. Crystalline high-purity boron is used in thermistors [52], [53] and delay lines [54]. Boron filaments have been developed as reinforcing material for light-weight, stiff composites for use in commercial and military aircraft [55]. Epoxy resins [56], polyamides, aluminum, magnesium, and other materials are the matrix [57], [58]. However, graphite filaments have replaced boron filaments in many applications.

In nuclear technology thin films of boron are used in neutron counters [52]. Boron powder dispersed in polyethylene castings is used for shielding against thermal neutrons [59]. In addition, composites of boron with Al or Fe, especially when enriched in ^{10}B, are used as neutron shields and absorbers in nuclear reactors [59].

Boron is an effective deoxidizing agent, e.g., for production of pure copper [60]. It also is important in the preparation of amorphous magnetic metals and alloys [61]. Boron, as ferroboron, is used in microalloyed steel (0.001 % B). This steel exhibits excellent resistance to stress cracks as well as improved tensile strength and hardness.

Amorphous boron of 90 – 92 % purity costs about DM 250 – 300 (≈ $ 100) per kg (1984). Material of higher purity is much higher in price: very pure boron (> 99.99 %, four nines) may cost up to DM 10000 (≈ $ 3500) per kg and more. Boron enriched in ^{10}B costs $ 9 – 10 per gram [59].

The leading *producers* are H. C. Stark, Goslar, Federal Republic of Germany, and Kerr McGee Chemical Corp., Oklahoma, United States. There are further production facilities at Degussa AG, Frankfurt am Main; Elektroschmelzwerk Kempten, München; Union Carbide Chemical Corp., New York; NL Industrial Corp., Wilmington, Delaware, United States; Murex Ltd., Rainham, United Kingdom; and Kawecki-Billiton Corp., Arnhem, the Netherlands [62]. Boron is also produced in the Soviet Union.

Total production is estimated to be 30 –50 t/a.

2. Ferroboron and Other Boron Master Alloys

Ferroboron is basically an iron – boron alloy containing 10 – 20 wt% B. It is used mainly in the steel industry. Ferroboron was first produced in 1893, by Henri Moissan [63], from boric acid, iron, and carbon in a single-phase electric-arc furnace lined with carbon. Ferroboron produced in this way contains carbon. The introduction of the thermite reaction by Goldschmidt in 1898 led to the aluminothermic reduction of boric oxide to ferroboron, for years the main commercial method for producing ferroboron. The carbothermic process has again found use for the production of ferroboron, now as a raw material for the manufacture of amorphous metals (→ Metallic Glasses). Aluminothermic ferroboron contains residual aluminum, which causes severe casting problems.

2.1. Physical Properties

In the system iron – boron, shown in Figure 1 [64], there are two intermetallic compounds, FeB [*12006-84-7*] and Fe_2B [*12006-86-9*] (ϱ

6.3 and 7.3 g/cm^3, respectively). The crystal structure of FeB is orthorombic ($a = 0.5502$, $b = 0.2948$, $c = 0.4057$ nm), whereas that of Fe_2B is tetragonal ($a = 0.5109$, $c = 0.4249$ nm). Additional physical properties of the two iron borides are summarized in Table 2. Also see [68].

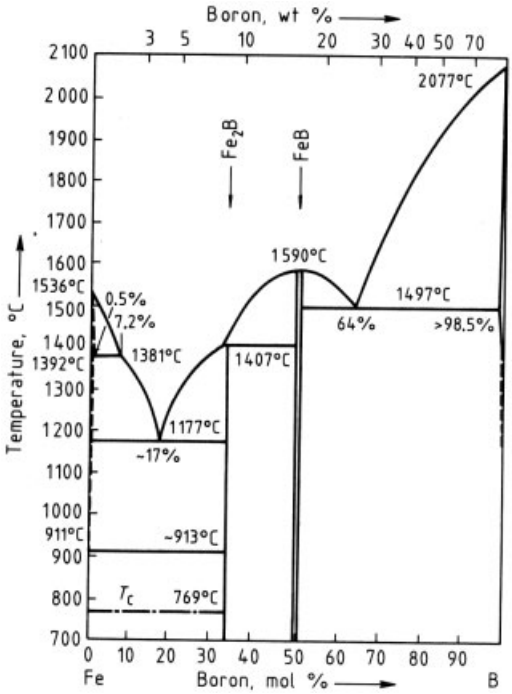

Figure 1. Phase diagram iron – boron [64]

Table 2. Properties of Fe_2B and FeB [65–67]

Property	Fe_2B	FeB
Microhardness, Vickers, 100 g, kN/mm^2	13.24	16.19
Thermal conductivity, W cm^{-1}K^{-1}	0.2 – 0.3	0.1 – 0.2
Electrical resistivity, at room temperature, μΩ cm	≈ 10	≈ 20
Temperature coefficient of resistance, μΩ/K	0.18	0.4
Curie temperature, K	1015	598
Coefficient of linear expansion (300 – 1000 K), K^{-1}	8×10^{-6}	10 – 16×10^{-6}

The enthalpies of formation at 1385 K have been determined by high-temperature solution calorimetry [69]: -67.87 ± 8.05 kJ/mol for Fe_2B and -64.63 ± 4.34 kJ/mol for FeB.

Commercial ferroboron alloys, either aluminothermic or carbothermic, are bright silver and brittle with a fine or coarse crystalline fracture. The density depends mainly on the boron content:

boron, wt %	20.2	17.9
aluminum, wt %	2.7	0.8
density, g/cm^3	6.02	6.41

2.2. Chemical Properties

Commercial ferroboron changes in appearance over prolonged exposure to humid air; its bright silver disappears, the surface turning dull gray, with almost red "rusty" stains. Ferroboron dissolves in mineral acids (Section 2.7). Distilled hot water may react slightly with ferroboron [70]. Some contradictory statements about the chemical behavior of ferroboron in the literature are probably due to other elements in the alloy. Fe_2B dissolves in HCl (1 : 2), hydrogen and boranes evolving, the latter hydrolyzing into suboxides [71], [72]. FeB is resistant to HCl ($\varrho = 1.19$ g/cm^3) and H_2SO_4 ($\varrho = 1.84$ g/cm^3), both at room temperature and at their boiling point. FeB dissolves completely in HNO_3. Both FeB and Fe_2B react with nitrogen at temperatures above 350 °C to give boron nitride [65].

The behavior of boronized steels, i.e., steel with thin layers of FeB on the surfaces, was discussed by A. Matuschka [66]. Such steels are oxidation resistant up to 700 – 900 °C. The boronized steel CK45 (AISI 1042) is resistant to 20 % HCl, 10 % H_3PO_4, and 10 % H_2SO_4 at 56 °C. Stainless steels, such as X10CrNiTi 18 9 (Werkstoff Nr. 1.4541, AISI 321, SAE 30321) can be made more resistant to attack by 20 % HCl or 10 % H_2SO_4 at 56 °C by boronizing the surface.

2.3. Raw Materials

The following boron minerals can be used for the manufacture of ferroboron: colemanite (51 % B_2O_3), pandermite (48 % B_2O_3), priceite (51 % B_2O_3), and boracite (62 % B_2O_3). However, the raw materials most commonly used are boric oxide (≈ 99 % B_2O_3) and boric acid (≈ 57 % B_2O_3).

2.4. Production

Boric oxide can be reduced by carbon, aluminum, or magnesium. Reduction by silicon is incomplete. Figure 2 shows the calculated free energies of the corresponding reactions as a function of temperature [67]. For commercial production, either reduction by carbon (carbothermic or endothermic) or reduction by aluminum, sometimes with some magnesium (aluminothermic or exothermic) is usual. The coreduction of iron and boric oxide proceeds more readily than the reduction of either alone because of the formation of the borides Fe_2B and FeB. Iron can be replaced by other boride-forming metals, manganese, nickel, or cobalt. In carbothermic reduction there is a tendency to form boron carbide, thus leading to a carbon-containing ferroboron.

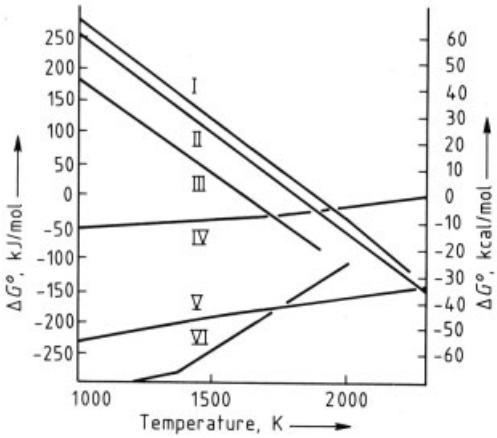

Figure 2. Free energies ΔG^0 of reactions reducing boric oxide as a function of temperature
I $2/3\,B_2O_3 + 2\,C \longrightarrow 4/3\,B + 2\,CO$
II $2/3\,B_2O_3 + 7/3\,C \longrightarrow 1/3\,B_4C + 2\,CO$
III $2/3\,B_2O_3 + 4/3\,Fe + 2\,C \longrightarrow 4/3\,FeB + 2\,CO$
IV $2/3\,B_2O_3 + Si \longrightarrow 4/3\,B + SiO_2$
V $2/3\,B_2O_3 + 4/3\,Al \longrightarrow 4/3\,B + 2/3\,Al_2O_3$
VI $2/3\,B_2O_3 + 2\,Mg \longrightarrow 4/3\,B + 2\,MgO$

Aluminum and boron also form borides (see Metal Borides under → Boron Carbide, Boron Nitride, and Metal Borides), and this results in residual aluminum in aluminothermic ferroboron.

2.4.1. Carbothermic Production

In order to produce a carbothermic ferroboron with a low carbon content the boron content must be high, i.e., the higher the boron content the lower the carbon solubility and thus carbon pick-up.

The reduction of boric oxide by carbon requires high temperature; therefore, the process is carried out in an electric-arc furnace (single- or three-phase type) of capacities up to 1500 kV A. These are relatively small production units, for the worldwide demand for ferroboron is low compared with that for other electrothermic ferro alloys.

The London & Scandinavian Metallurgical Co., United Kingdom, produces a ferroboron with 16 – 18 % B, 0.5 % C, and < 0.15 % Al by carbothermic reduction of boric oxide.

In a Japanese patent [73] the carbothermic production of ferroboron from boric acid, iron powder, and charcoal in a Héroult-type electric-arc furnace with carbon lining is claimed. One example describes the production of an alloy with 10.3 % B, 2 % Si, and 0.98 % C in a 350-kW three-phase electric-arc furnace. The mix is 100 parts boric acid, 135.8 parts iron powder (92.9 % Fe), and 57 parts charcoal powder. The boron recovery is 81.7 %, and the power consumption 4550 kWh per tonne of alloy. Another example describes the small-scale production of an alloy containing 15.3 % B. Another Japanese patent [74] claims the carbothermic production of a B- and Si-containing alloy, e.g., 3.3 wt% B, 2.9 wt% Si, and 3.0 wt% C, in a special vertical blast furnace. The ferroboron is intended for use in the manufacture of amorphous alloys.

The production of ferroboron from pig iron swarf and boric acid in an electric-arc furnace with a final oxygen blow leads to an alloy with 16.3 % B, 0.03 – 0.06 % Al, and 0.03 – 0.06 % C. The boron recovery in small-scale runs is said to be 60 – 65 %; the power consumption, ≈ 16500 kWh per tonne of alloy [75]. Silicon-containing boron alloy can be produced by carbothermic reduction. The process is described by K. D. Frank in [76]. A mixture of 100 parts colemanite, 100 parts quartzite, 40 parts iron turnings, and 60 parts low-temperature coke is charged to a 1000-kW electric-arc furnace. A metal containing 5.35 % B, 37.2 % Si, and 0.21 % C is tapped. Boron recovery is 70 %, and the power consumption is 6000 kWh per tonne of alloy.

Gesellschaft fuer Elektrometallurgie mbH, Duesseldorf, Federal Republic of Germany, has

developed a process for the carbothermic production of ferroboron containing 15 – 20 % B and < 0.1 % Al in a three-phase submerged arc furnace, whereby boric oxide and iron oxide are simultaneously reduced by charcoal and other low-density carbonaceous materials, e.g., wood chips [77]. Nickel boron can also be produced carbothermically. Boric oxide and/or boric acid and nickel are the raw materials. London and Scandinavian Metallurgical Co., London, United Kingdom, produces a nickel boron containing 15 – 18 % B; 0.5 and 0.15 % C (max.); and 0.20 % Al (max.) in the electric-arc furnace by carbon reduction.

2.4.2. Aluminothermic Production

Ferroboron can be made batchwise in a convenient way by the reduction of boric oxide and iron oxide with aluminum powder. Some magnesium in the aluminothermic mix is beneficial: Magnesium is the stronger reducing agent at temperatures below its boiling point, whereas aluminum is more effective above the boiling point of magnesium, where such metallothermic reactions generally take place.

The aluminothermic coreduction of boric oxide and iron oxide (Fe_2O_3) is highly exothermic, and only a little additional energy is necessary for a self-propagating reaction (→ Aluminothermic Processes). The thoroughly mixed compounds are charged into a refractory-lined pot and ignited, either the whole mix (top firing) or by igniting a starting mix, the rest then charged as the reaction proceeds over a few minutes. The liquid metal and slag separate on account of differing densities, and after cooling the metal button, up to 1500 kg, is removed. After mechanical cleaning, the metal button is broken and crushed to the desired size. The companies of the Metallurg Group [78], who produce ferroboron from boric oxide, iron oxide, and aluminum powder, obtain boron recoveries of 70 – 75 % for the 18 – 20 % B grade and 80 – 85 % for the 15 – 18 % B grade by optimizing process parameters. The aluminum consumption can be as low as 4.8 – 5 kg Al per kg of B.

N. A. Chirkov et al. [79] investigated heat balances for an electro-aluminothermic ferroboron process using either 1) boric oxide or 2) boric acid as the raw material. Three different mixes – igniting, main, and precipitation – are charged consecutively into a three-phase electric-arc furnace and reacted to produce a ferroboron melt with ≈ 20 % B. The composition of the charges and the results are summarized in Table 3. The heat balances for these two cases are shown in Table 4. The investigators are convinced that the process can be optimized to improve the boron recovery and the consumption figures.

When borate ores such as boracite or colemanite are used as raw materials, instead of boric oxide, the aluminothermic process gives boron recoveries as low as 40 – 50 % [80–82]. The specific aluminum consumption is normally higher because the process needs additional energy, in form of either a booster (peroxides or chlorates plus aluminum powder) or in the form of electricity by smelting the mix in the electric-arc furnace (electro-aluminothermic process).

Complex boron additives for steel, which contain boron in a dilute alloy together with protective elements, are produced from ferroboron and other alloying elements by a melting process (see Section 2.5, Table 7, and [83]). Other boron master alloys such as nickel boron and cobalt boron with ≈ 15 % B and 0.1 % or 1.0 % Al (max.) can be produced aluminothermically using boric oxide and nickel or cobalt (as metal and/or oxide) [84]. Copper boron with ca. 2 % B is commercially produced by Gesellschaft fuer Elektrometallurgie [85], using an electromagnesiothermic process with copper and boric oxide as raw materials. Aluminum master alloys with 2.5 – 4.5 % B and ≈ 5 % Ti + 1 % B are produced in an induction furnace by adding potassium boron fluoride and potassium titanium fluoride to an aluminum bath [86].

2.5. Quality Specifications for Commercial Grades

Quality standards are specified in Germany by DIN 17567 [87], which also describes sizing, sampling, analytical methods, shipment, and storage. Analyses of the standard grades are summarized in Table 5. The American ASTM recognizes six grades of ferroboron; analyses are given in Table 6. Ferroboron is shipped in various sizes. In Europe pieces of fist size and

Table 3. Charge composition during electro-aluminothermic production of ferroboron using boric oxide and boric acid raw materials [79]

	Mix 1:			Mix 2:		
	Igniting	Main	Precipitation	Igniting	Main	Precipitation
Raw materials						
Boric oxide, kg	–	1200	–	–	–	–
Boric acid, kg	–	–	–	–	1800	–
Iron ore, kg	200	200	1000	200	180	950
Al powder, kg	65	820	282	65	820	262
Lime, kg	70	100	170	70	450	170
Total, kg	335	2320	1452	335	3250	1382
Time, min	30	98	32	3	114	23
Power consumption, kWh		1130			1980	
Products						
Ferroboron, kg		1200 ∗			1100 ∗	
% B		20			18	
% Al		3			3.9	
% Si		1.13			1.7	
Slag, kg		2900∗			3080∗	
% B_2O_3		14 – 15∗			10 – 12∗	
Boron recovery, %		64			63	
Aluminum consumption, kg Al per tonne of ferroboron		973			1043	
Energy consumption, kWh per tonne of ferroboron		940			1800	

∗ Estimated.

Table 4. Heat balances for the electro-aluminothermic production of ferroboron using boric oxide and boric acid as raw materials [79]

Input and consumption	From B_2O_3		From H_3BO_3	
	ΔH, 10^6 J	, %	ΔH, 10^6 J	, %
1. Chemical reactions	13182	69.0	12345	58.3
2. Electric power	4072	21.3	7143	33.7
3. Side reactions	1833	9.7	1699	8.0
Total input	19087		21186	
1. Heat content of slag	9515	49.3	9237	42.9
2. Heat content of metal	3235	16.7	2907	13.5
3. Heat content of dust	63	0.3	125	0.6
4. Dehydration of acid	48	0.2	1084	5.0
5. Decomposition of carbonates	60	0.3	98	0.5
6. Heat content of gases	141	0.7	2805	13.1
7. Accumulation of heat by furnace brickwork	4678	24.2	4212	19.6
8. Radiation	1581	8.3	1045	4.8
Total consumption∗	*19319*		21514	

∗ The discrepancies between input and consumption are 232×10^6 J (1.2 %) and 328×10^6 J (1.5 %).

Table 5. German Standards DIN 17567 [87]

Designation	Composition							
	B wt %	Al max. wt %	Si	C	Mn	P	S	Co
FeB16	15 – 18	4.0	1.0					
				}0.10	}0.50	}0.005	}0.001	}0.005
FeB18	18 – 20	2.0	2.0					
FeB12C	10 – 14							
		}0.50	}4.0	}2.0	}0.50	}0.005	}0.1	}0.005
FeB17C	14 – 19							

Table 6. United States standards [88]

Grade	Composition, wt%				
	Boron min.	Boron max.	Carbon, max.	Silicon, max.	Aluminum, max.
A 1	12.0	14.0	1.5	4.0	0.5
A 2	12.0	14.0	1.5	4.0	8.0
B 1	17.5	19.0	1.5	4.0	0.5
B 2	17.5	19.0	1.5	4.0	8.0
C 1	19.0	24.0	1.5	4.0	0.5
C 2	19.0	24.0	1.5	4.0	8.0

in the United States sizes such as 2 in or 1 in x Down are commercial. Ferroboron can also be ground to powder of desired size (e.g., 20 mesh x D, i.e., less than 0.84 mm). In the United States, where the use of boron in steel is fast growing, and also in the United Kingdom, special boron alloys have been developed as hardenability intensifiers (Table 7). These complex alloys contain elements to protect boron against oxidation (Al, Si) and boron nitride formation (Ti, Zr) [89]. Boron-containing aluminum master alloys (Table 8) are standardized in Germany [90]. Nickel boron, cobalt boron, and copper boron are not standardized. Typical analyses for commercial alloys are given in Table 8.

In the United States and the United Kingdom special boron additives for the steel industry have been developed. The *trade names*, e.g., BATS (abbreviation for boron-aluminum-titanium-silicon), are shown in Table 7.

2.6. Storage and Shipment

Ferroboron and other boron master alloys are generally not toxic, and there is no special legislation on use and handling. Ferroboron is gener-

Table 7. Complex boron alloy "hardenability intensifiers" [83], [89]

Designation	Composition [a], wt%						
	B	Al	Ti	Si	Zr	Mn	C
BATS 50 [b]	0.50	13	20	7	5	8	–
BATS TT [b]	1.60	11	58	–	3.5	5	–
BATS 2Z [b]	2.50	11	36	8	6	–	–
BATS Wire	2.3	89.7	8	–	–	–	–
CARBOTAM [b]	2.0	3.0	15 – 20	4	–	–	0.50
BALCORE 101 [c]	2.5	25	–	–	–	–	–
BALCORE 102 [c]	3.4	34	–	–	–	–	–

[a] The remainder is iron.
[b] BATS 50, BATS TT, BATS 2Z, and CARBOTAM are registered trademarks of the Shieldalloy Corp., Newfield, N.J.
[c] BALCORE is a registered trademark of the LSM Co., London. The size of BALCORE 101 wire is 3.2 mm (34 g/m). The size of BALCORE 102 wire is 5 mm (75 g/m).

Table 8. Composition of boron master alloys [83–85]

Alloy	Analysis, wt%					
	B	Ti	Al	Si	C	Others
Aluminum boron V-Al B 3	2.5 – 3.4	–	rest	max. 0.2	–	*
V-Al B 4	3.5 – 4.5	–	rest	max. 0.2	–	*
Aluminum-titanium-boron V-Al Ti 5 B 1	0.9 – 1.4	5.0 – 6.2	rest	max. 0.2	–	*
Nickel boron	15 – 18	–	0.1 – 1.0	max. 0.8	0.1 – 0.5	rest Ni
Cobalt boron	15 – 18		max. 1.0	–	–	rest Co
Copper boron	≈2	–	–	–	–	rest Cu

* See DIN 1725 [90].

ally packed in steel drums marked with material grade, size, lot number, and producer. The shipment is accompanied by a data sheet on safe handling including physical and chemical properties, safe storage, and handling directions, e.g., use of protective gloves or dust respirators. Ferroboron and other boron master alloys and additives should be stored in a dry area in their original containers.

2.7. Chemical Analysis

For the determination of boron in ferroboron the sample is dissolved in HCl (1 : 1) or H_2SO_4 (1 : 1) containing a few drops of HNO_3 and refluxed. Residues rich in boron are fused with sodium peroxide. The solution is passed through an ion-exchange column, neutralized, diluted with polyalcohol, and titrated with sodium hydroxide [91–93]. Small amounts of boron in steel are determined by special methods [89]. One photometric determination of boron in steel is the curcumin method. A fully automated determination of boron in ferroboron, aluminum, and complex ferro alloys by neutron transmission is a convenient method for routine analyses [94]. Standard analytical methods in the United States and the United Kingdom [95], [96] have been described.

2.8. Pollution Control

During production the usual proscriptions on pollution have to be obeyed.

In both the carbothermic and the exothermic reduction processes the dust off-gases have to be cleaned, either dry in bag filters or electrostatic dust precipitators or wet in disintegrators or Venturi scrubbers (O. Renz in [76]). A special combined dry – wet cleaning plant for the off-gases of an aluminothermic plant is described by Seybold [97].

2.9. Uses

The main use of ferroboron is in steel (→ Steel) [89], [98]. Very small amounts of boron improve the hardenability of both plain carbon and alloyed steels. Boron acts as a hardenability intensifier of other hardening elements, such as carbon, manganese, chromium, molybdenum, etc. The boron should be present in the concentration range of 5 – 15 ppm (max., 30 ppm). Thus the use of boron saves the expense of more costly alloying elements. Because of the high affinity of boron for oxygen and nitrogen, also that dissolved in steel, boron has to be protected from these two elements. Therefore, good predeoxidation by calcium, aluminum, or silicon and denitrification by titanium or zirconium are necessary. In the United States and the United Kingdom complex boron master alloys that contain these protective elements have been developed and have proved very effective for controlled boron addition to steels. In case-hardened steels, boron exhibits better hardenability and strength; in nonaging steels, especially for the automotive industry, the sheet-forming properties are improved. Boron also offers benefits in high-strength low-alloy (HSLA) steels, and it improves the cutting properties in high-speed steels.

In stainless steels of the chrome – nickel type (ASTM 316), 0.004 – 0.009 % B significantly improves the ductility in the hot forming range (950 – 1250 °C). In very low concentration (0.002 – 0.005 %) boron is beneficial to creep properties in austenitic heat-resistant stainless steels, as well as in ferritic steels, then in combination with titanium, molybdenum, or niobium. Steels with boron contents up to 2.5 % are used for control rods in nuclear reactors because the thermal neutron cross section of boron is so high. Boron is also used in hardfacing and abrasion-resistant alloys; e.g., Colmonoy, a Ni-Cr-Si-B alloy containing up to 3 % B, and Stellite, an abrasion-resistant Co-Cr-W alloy containing up to 2.5 % B.

The copper – boron master alloy is the best agent for deoxidizing copper for the production of a highly conducting metal. Aluminum is treated with aluminum – boron master alloy to increase the electrical conductivity and with aluminum – titanium – boron master alloy to refine the grain [86].

A new field of boron usage is the production of amorphous alloys or metallic glasses (→ Metallic Glasses). Intensive development has been carried out by Allied Chemical Corp.

[99]. These alloys consist of a transition metal (Fe, Ni, or Co) and about 20 mol % of a metalloid (B, C, Si, or P). They exhibit extraordinary physical properties, especially the magnetic properties, and may be used for low-hysteresis-loss transformer cores [100].

2.10. Economic Aspects

The consumption of ferroboron is relatively low due to the small concentrations necessary in steel. In the Western World and Japan, ferroboron production was ≈ 1400 t/a in 1982 and ≈ 2000 t/a in 1983, showing an increasing tendency. In 1981 only 2 % of the total boron consumption in the United States was for the metallurgy and nuclear fields [101] (→ Boron Compounds). According to American Metal Market (28 Dec. 83), ferroboron costs $ 3.89 per pound (one pound = 454 g); nickel boron, $ 5.57 per pound of the regular grade, $ 5.67 per pound for the low-aluminum grade, and $ 6.30 per pound for the low-carbon grade.

2.11. Toxicology and Industrial Hygiene

Ferroboron and boron master alloys are generally not toxic, but in the handling of raw materials for their production protective measures must be taken, e.g., the use of protective gloves and glasses as well as dust respirators when handling boric acid. In the production facilities the usual protective measures (clothing, glasses, dust respirators) must also be taken.

3. References

General References

1. V. I. Matkovich (ed.): *Boron and Refractory Borides,* Springer Verlag, Berlin 1977.
2. R. M. Adams: *Boron, Metallo-Boron Compounds and Boranes,* J. Wiley & Sons, New York 1964.
3. R. F. Gould (ed.): "Borax to Boranes," *Adv. Chem. Ser.* 1961, no. 32, pp. 27 – 52.
4. *Gmelin,* Boron (system no. 13), main vol. (1926), supplement vol. (1954), supplement vol. 2 (1981).
5. N. N. Greenwood: "Boron," in J. C. Bailar (ed.): *Comprehensive Inorganic Chemistry,* vol. 1, Pergamon Press, Oxford-New York 1973, pp. 680 – 689.
6. R. Naslain in P. Hagenmuller (ed.): *Preparative Methods in Solid State Chemistry,* Academic Press, New York 1972, pp. 439 – 485.
7. A. F. Zhigach, D. C. Stasinevich: "Methods of Preparation of Amorphous Boron," in V. I. Matkovich (ed.): *Boron and Refractory Borides,* Springer Verlag, Berlin 1977, pp. 214 – 240.
8. H. S. Cooper: "Boron," in C. A. Hampel (ed.): *Rare Metals Handbook,* 2nd ed., Reinhold Publ. Co., London 1961, p. 77.

Specific References

9. B. Mason: *Principles of Geochemistry,* John Wiley & Sons, New York 1952, p. 41.
10. H. Moissan, *Ann. Chim. Phys.* **6** (1895) 296.
11. E. Weintraub, *Trans. Am. Electrochem. Soc.* **16** (1909) 165.
12. K. P. Tsomaya et al., *J. Less Common Met.* **47** (1976) 249.
13. J. Cueilleron, *Ann. Chim.* **19** (1944) 459.
14. Y. Hara et al., *Kenkyu Hokoku Asahi Garasu Kogyo Gijutsu Shoreikai* **26** (1975) 255; *Chem. Abstr.* **87** (1977) 44716 p.
15. G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* vol. 1, pp. 253 – 554, vol. 7, pp. 111 – 364, Pergamon Press, Oxford 1982.
16. L. Bretherick (ed.): *Hazards in the Chemical Laboratory,* 3rd ed., Royal Society of Chemistry, London 1981, p. 203.
17. J. Cueilleron, F. Thevenot: "Chemical Properties of Boron," in V. I. Matkovich (ed.): *Boron and Refractory Borides*, Springer Verlag, Berlin 1977, pp. 203 – 211.
18. L. Vandenbulcke, G. Vuillard, *J. Less Common Met.* **67** (1979) 65.
19. D. Z. Hobbs, T. T. Campbell, F. E. Block, *Rep. Invest. U.S. Bur. Mines* **6456** (1964) 1 – 14.
20. B. I. Silbergleit et al., SU 1004262, 1981.
21. J. C. Schuhmacher, US 3001855, 1961.
22. H. Mazza et al., US 2866688, 1958.
23. W. J. Kroll, N. P. Nies, E. W. Fajans, US 2893842, 1959.
24. J. Yannacakis, N. P. Nies, *Boron Synth. Struct. Prop. Proc. Conf. 1959* 1960, 38 – 41.
25. L. Markovskii, *Electron Technol.* **3** (1970) 95.
26. H. C. Starck, GB 747287, 1956 (H. Haag). F. Müller, H. Haag, DE 936745, 1955. H. Haag, DE 957299, 1957; US 2794708, 1957.
27. R. Naslain, *Inorg. Synth.* **12** (1970) 145.

28. T. Sugaya et al., *J. Less Common Met.* **47** (1976) 49.
29. H. J. Becher: "Boron," in G. Brauer (ed.): *Handbuch der Präp. Anorg. Chemie,* 3rd ed., vol. 2, Enke Verlag, Stuttgart 1978, pp. 787 – 788.
30. G. H. Fetterly, US 2542916, 1951.
31. D. Stern, L. Lynds, *J. Electrochem. Soc.* **105** (1958) 676.
32. A. J. Laubengayer et al., *J. Am. Chem. Soc.* **65** (1943) 1924.
33. Elektroschmelzwerk Kempten, DE 2115810, 1982 (G. Wiebke et al.).
34. J. Cueilleron, J. C. Viala, *J. Less Common Met.* **65** (1979) 167.
35. J. Cueilleron, J. Viala, *J. Less Common Met.* **67** (1979) 333.
36. A. van Arkel, US 1774410, 1930; *Metallwirtsch. Metallwiss. Metalltech.* **13** (1934) 405.
37. H. Schlesinger, G. W. Schaeffer, US 2528514, 1950.
38. J. Lagrenaudie, *J. Chem. Phys.* **50** (1953) 629.
39. W. H. Dietz, H. A. Herrmann, *Electron Technol.* **3** (1970) 195.
40. N. P. Nies, *J. Electrochem. Soc.* **107** (1960) 817.
41. W. Borchert et al., *Z. Angew. Phys.* **29** (1970) 277.
42. G. V. Tsagareishvili et al., *J. Less Common Met.* **67** (1979) 419.
43. F. Thevenot, J. Cueilleron, *Analusis* **5** (1977) 105; *Met. Abstr.* 1977, 23-0637.
44. F. Ehrenberger, *Fresenius Z. Anal. Chem.* **305** (1981) 181.
45. F. Thevenot, J. C. Viala, *Analusis* **3** (1975) 76.
46. Military Specification MIL-B-51092/ORD (1962).
47. Degussa, Research Dept. Phys. Chem., internal measurements (1983).
48. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* Thieme Verlag, Stuttgart 1980, pp. 206 – 208.
49. R. J. Markle, S. G. Stephenson, *Int. J. Powder Metall. Powder Technol.* **18** (1982) 243.
50. W. C. Goss, US 2415946, 1947.
51. *Winnacker-Küchler,* 4th ed., **vol. 3,** pp. 408 – 440.
52. G. K. Gaulé et al., *Boron Prep. Prop. Appl. Pap. Int. Symp. 1964* **2** (1965) 317 – 338.
53. G. F. Girri et al., *3rd Int. Conf. Thin Films,* Budapest 1975, p. 215; *Met. Abstr.* 1976, 63-0163.
54. J. T. Milek, S. J. Wellers, *Boron,* Electronic Prop. Inf. Center, Hughes Aircraft Co., Culfer City, Calif., Data Sheet DS 151 (1967).
55. J. L. Christian et al., *Met. Prog.* **97** (1970) 113.
56. G. Lubin, S. Dastin, *Mod. Plast.* **48** (1971) 62.
57. D. Kuehl, *Nat. SAMPE Symp. Exhib. Proc.* **20** (1975) 67; *Met. Abstr.* 1975, 61-0384.
58. M. Miller, A. Robertson, *Hybrid Sel. Met. Matrix Compos.* 1977, 99 – 157; *Met. Abstr.* 1979, 62-0211.
59. *Kirk-Othmer,* 3rd ed., **vol. 16,** p. 163.
60. *Met. Abstr.* 1981, 46-0040.
61. *Kirk-Othmer,* 3rd ed., **vol. 2,** pp. 537 – 569.
62. H. P. Münster, G. Kirchner: *Taschenbuch des Metallhandels,* Metall-Verlag, Berlin 1982, pp. 151 – 153.
63. J. Escard: *Les Fours Electriques Industriels et les Fabrications Electrothermiques,* 2nd ed., Dunod, Paris 1924, p. 458.
64. O. Kubaschewski: *Iron-Binary Phase Diagrams,* Springer Verlag, Berlin-Heidelberg-New York, Verlag Stahleisen, Düsseldorf 1982.
65. G. V. Samsonov: *Handbook of High Temperature Materials,* no. 2: Properties Index, Plenum Press, New York 1964.
66. A. Graf v. Matuschka: *Borieren,* Hanser Verlag, München-Wien 1977.
67. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1973, supplement (authors: I. Barin, O. Knacke, O. Kubaschwski) 1977.
68. H. Pastor, F. Thevenot, *Inf. Chim.* 1978, no. 178, 151 – 170.
69. S. Sato, O. J. Kleppa, *Metall. Trans. B* **13 B** (1982) 251 – 257.
70. *Gmelin,* Iron (system no. 59), main A 2 (1929) 1177 –1178.
71. L. Y. Markovskii, G. V. Kaputovskaja, *Zh. Prikl. Khim. (Leningrad)* **33** (1960) 569 – 577.
72. L. Y. Markovskii, E. T. Bezruk, *Zh. Prikl. Khim. (Leningrad)* **35** (1962) 491 – 498.
73. Mitsui Mining & Smelting Co., JP 60016/73, 1971 (T. Masuyama et al.).
74. Kawasaki Steel Corp., DE-OS 3228593, 1982, Prior. JP P 174960-81 (T. Hamada et al.).
75. Mitsui Mining & Smelting Co., JP 5826025, 1981.
76. G. Volkert, K. D. Frank (ed.): *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin 1972.
77. Gesellschaft fuer Elektrometallurgie, DE 3409311-7, 1984.

78. Metallurg Alloy Corp., New York, N.Y., USA; Shieldalloy Corp., Newfield, N.J., USA; London & Scandinavian Metallurgical Co., Ltd., London,England; Gesellschaft fuer Elektrometallurgie mbH, Duesseldorf, Germany (private communication).
79. N. A. Chirkov et al., *Stal'* 1967, 325 – 327, Transl. BISI 14535 (August 1976).
80. V. P. Elyutin, Y. A. Pavlov, B. E. Levin, E. M. Alekseev: *Production of Ferroalloys – Electrometallurgy,* 2nd ed., National Science Foundation, Washington, D.C., and the Department of the Interior; translated from Russian by the Israel Program for Scientific Translations (1961).
81. M. Riss, Y. Khodorovsky: *Production of Ferroalloys,* MIR Publishers, Moscow 1967 (transl. from the Russian by I. V. Savin).
82. G. V. Samsonov, *Vopr. Poroshk. Metall. Prochn. Mater.* 1960, no. 8, 8 – 23; *Chem. Abstr.* **57** (1962) 8274 c.
83. Shieldalloy Corp., Newfield, N.J.: *Product Specification Data Sheet* (1979).
84. London & Scandinavian Metallurgical Co., London: *The Metallurg Companies, Products etc.* (July 1975).
85. Gesellschaft fuer Elektrometallurgie, Duesseldorf: *Legierungen und Legierungsmetalle,* Prospectus 1980.
86. G. P. Jones, J. Pearson, *Metall. Trans. B* **7 B** (1976) 223 – 234.
87. DIN 17567: (Jan. 1970).
88. ANSI/ASTM A 323-76: *Standard Specification for Ferroboron* .
89. P. D. Deeley, K. J. A. Kundig: *Review of Metallurgical Applications of Boron in Steel,* Shieldalloy Corp., Newfield, N.J., Metallurg Alloy Corp., New York, 1982.
90. DIN 1725 Blatt 3: (June 1973).
91. Chemikerausschuß der GDMB: *Analyse der Metalle, I. Schiedsanalysen,* 3rd ed. (1966) ; II. Betriebsanalysen, 2nd ed. (1961), Springer Verlag, Berlin.
92. *Handbuch für das Eisenhütten-Laboratorium,* vol. 2 a, Verlag Stahleisen, Düsseldorf 1982.
93. G. Wünsch, F. Umland in W. Fresenius, G. Jander (ed.): *Handbuch der analytischen Chemie,* **3rd part: Quantitative Analyse, vol. III: "Elemente der dritten Hauptgruppe, Bor" 2nd ed.,** Springer Verlag, Berlin-Heidelberg-New York 1971.
94. H. L. Giles, G. M. Holmes, *J. Radioanal. Chem.* **48** (1979) 65 – 72.
95. British Standard 1121: *Methods for the Analysis of Iron and Steel,* part 49: *Boron in Ferro-Boron, Volumetric*, 1966.
96. ASTM (E 31): Chemical Analysis of Ferro-Alloys,
97. C. F. Seybold, *Electr. Furn. Proc.* **27** (1969) 99 – 108.
98. P. D. Deeley, K. J. A. Kundig, H. R. Spendelow Jr.: *Ferroalloys & Alloying Additives Handbook,* Shieldalloy Corp., Newfield, N.J., Metallurg Alloy Corp., New York, N.Y., 1981.
99. Allied, DE-OS 3011152, 1980.
100. W. Jaschinski, W. Wolf, *Metall (Berlin)* **36** (1982) 782 – 783.
101. E. W. Fajans: *Mining Annual Review – 1982,* Min. J., London, June, 1982, pp. 112 – 113.

Borides → Boron Carbide, Boron Nitride, and Metal Borides

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Boron Carbide, Boron Nitride, and Metal Borides

Standard Article

Jochen Greim[1] and Karl A. Schwetz[1]

[1]ESK Ceramics GmbH & Co. KG, A Ceradyne Company, Kempten, Germany

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a04_295.pub2
Article Online Posting Date: December 15, 2006

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Boron Carbide
1.1. Crystal Structure
1.2. The Boron - Carbon System
1.3. Physical Properties
1.4. Chemical Properties
1.5. Production
1.6. Uses
1.7. Economic Aspects
2. Boron Nitride
2.1. Properties
2.2. Production
2.3. Uses
3. Metal Borides
3.1. Properties
3.2. Production
3.3. Uses

[Top of Page]

1. Boron Carbide

Boron carbide [12069-32-8] belongs to the important group of nonmetallic hard materials, which includes alumina, silicon carbide, and diamond [1]. Although it was first prepared over a century ago, in 1883, by JOLY [2], the formula B_4C was assigned only in 1934 [3]. Today a homogeneity range from $B_{4.3}C$ to $B_{10.4}C$ has been established. The composition of commercial boron carbide is usually close to a boron : carbon stoichiometry of 4 : 1, the stoichiometric limit on the high-carbon side.

Boron carbide is an exceedingly hard material, inferior in hardness only to diamond and cubic boron nitride. In addition, it has a high melting point, high mechanical strength combined with low density, and high neutron cross section, and is a semiconductor. General information about boron carbide is available in the reviews of LIPP [4], THEVENOT et al. [5], [6] and TELLE et al. [7].

1.1. Crystal Structure

Structurally, boron carbide can be considered a prototype of the so-called interstitial compounds of α-rhombohedral boron [8], which include $B_{12}C_3$ [12075-36-4] (= B_4C), $B_{12}C_2Al$ [9], $B_{12}S$, $B_{12}O_2$, $B_{12}As_2$, $B_{12}P_2$, and $B_{3-4}Si$ [10]. The crystal structure for these materials consists of icosahedra that occupy the points of a rhombohedron, plus a chain of atoms that runs along the *c* axis of the rhombohedron. For the most part, boron atoms occupy the points of the icosahedra. The second constituent — C, Al, O — in general occurs on the chain (Fig. 1). (There is no chain at all in α-rhombohedral boron.) The rigid framework of polyatomic units of closely bonded atoms is reflected in the refractory nature and great hardness.

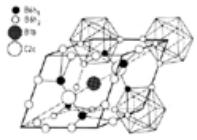

Figure 1. The structure of boron carbide for the stoichiometry $B_{13}C_2$.

For compositions close to $B_{12}C_3$, the lattice parameters are a_H = 0.55980 nm and c_H = 1.20707 nm [11]. However, the details of the boron carbide structure [12], [13] are not firmly established. According to investigations [14] the sequence C–B–C can be assumed for the chain. In addition to these two carbon atoms per unit cell, carbon likely replaces boron at the boron sites in the icosahedra. With 15 atoms per unit cell ($B_{12}C_3$) and 20 mol % C, boron carbide can be reformulated as $(B_{11}C)CBC$ to show this effect. However, SCHWETZ and KARDUCK [15] reported $B_{4.3}C$ as the carbon-rich limit of the homogeneity range of boron carbide. Optical absorption and reflectivity measurements on $B_{4.3}C$ indicated a structure with 81 % C – B – C chains, 19 % C – B – B chains, and exactly one carbon atom located in the $B_{11}C$ icosahedral site [16].

On the boron-rich side the extension of the homogeneity range to 8.7 mol % C, the ratio B : C = 10.4, has been reported by BOUCHACOURT, THEVENOT [17], who propose the structural formula $B_{12-x}C_x(C/B_xC_{1-x}/C)$ and n = 15.33 atoms per unit cell for $B_{10.4}C$. The density of boron carbide at the

boron-rich end is 2.465 g/cm^3; the volume of the unit cell, only 2.5 % larger than B_4C, corresponding to the lattice parameters a_H = 0.5651 nm and c_H = 1.2196 nm.

1.2. The Boron – Carbon System

Although numerous studies are available, not all parts of the B – C system have yet been fully elucidated.

SAMSONOV, SHURAVLOV, et al. [18], [19] reported the compounds $B_{13}C_2$ and $B_{12}C_3$, both with a large homogeneity range, in addition to the carbon-rich BC_2 and the boron-rich $B_{12}C$ phases. In the diagram of DOLLOFF [20], rhombohedral $B_{13}C_2$ and $B_{12}C_3$ phases are combined with a wide range of solid solubility of only one carbide phase, $B_{4+x}C$, which is stable only below 1800 °C. ELLIOT [21] reported the solid solubility of boron carbide from ca. 8 to 20 mol % C over the temperature range from room temperature to the melting point of 2450 °C. The B_4C – C eutectic temperature was reported to be 2375 °C at 29.0 mol % carbon. Investigations by KIEFFER et al. [22] support this broad range of solid solubility. In addition to the compounds given in these publications, $B_{25}C$ [23], B_8C [24], and $B_{13}C_3$ [25] have been reported. These are likely low-temperature phases, which are often observed in chemical vapor deposition. THEVENOT and BOUCHACOURT [5], [17] postulated a peritectic transformation of the β-rhombohedral boron – carbon mixed crystal, giving homogeneity range limits of $B_{10.4}C$ (8.8 mol % C) and $B_{4.0}C$ (20.0 mol % C). Their X-ray diffraction studies indicated a discontinuity in the lattice parameters at the composition $B_{6.5}C$. According to the phase diagram reported in 1983 by BEAUVY [26], the carbon-rich limiting composition corresponds to $B_{3.63}C$, i.e., "$B_{11}C_3$" with 21.6 mol % C. Investigations of the B – C system indicated [15], however, that the carbon content of fused boron carbide in equilibrium with graphite is only 19.2 mol%, corresponding to the formula $B_{4.3}C$. The phase diagram is presented in Figure 2. Recent careful measurements of the lattice constants on high-purity boron carbide showed the boundaries of the single-phase region to be at about 9 atom % C and near to, but likely some what less than, 20 atom % C [11]. Information on phase equilibria in the quaternary system B — C — Si — N is available in [27].

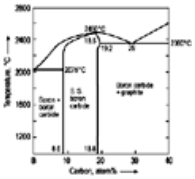

Figure 2. The B – C phase diagram [15]

The degenerated reaction at the B-rich corner is most probably an eutectic.

1.3. Physical Properties

The purest B_4C crystals yet obtained by the electrothermic production process are deep black and shiny, a contrast to SiC crystals, which in their purest form are transparent and colorless. The crystals melt congruently at ≈ 2450 °C; the liquid boils above 3500 °C. The density of B_4C having the natural isotopic composition is 2.52 g/cm^3; $^{10}B_4C$ is less dense, 2.37 g/cm^3.

The Knoop microhardness number HK-0.1, i.e., the hardness measured with a load of 0.1 kp (≅ 0.98 N), is ≈ 3000 [28] (2000 for α-Al_2O_3, 2600 for SiC and elemental boron, 4700 for cubic BN, and 7000 – 8000 for diamond). Therefore, except for diamond, the hardest known materials all contain boron. Table 1 presents some mechanical properties of hot-pressed B_4C and B_4C-based composites manufactured by ESK. The fracture mode of these materials is almost 100 % transgranular [29]. Boron carbide must be considered to be a refractory material, since hardness and strength do not degrade up to 1500 °C in an inert atmosphere [4].

Table 1. Mechanical properties of various B_4C ceramics

Property	Materials		
	HP B_4C[a]	B_4C (5 % C)[b]	BT 60/40[c]
Bulk density, g/cm^3	2.51	2.51	3.30
Flexural strength (4-point, 20 °C), MPa	480	560	730
Knoop hardness	2910	2860	2900
Fracture toughness K_{Ic}, MPa · m$^{1/2}$	2.2	2.1	6.5
Young's modulus, GPa	441	450	480
Mean grain size, µm	5	2	5

[a] Axially hot pressed B_4C with 1.7 wt % free carbon (C_{total} = 21.7 wt % C).

[b] HIP treated sintered B_4C with 4 wt % free carbon.

[c] HIP treated sintered B_4C – TiB_2 composite with 40 vol % TiB_2.

The average coefficient of thermal expansion is 4.6 × 10^{-6}K^{-1} over the range 25 – 800 °C, i.e., almost identical with that for pure SiC ceramic. The coefficient of thermal expansion as a function of temperature (t in °C) is

$$\alpha = 3.016 \times 10^{-6} + 4.30 \times 10^{-9}\, t - 9.18 \times 10^{-13}\, t^2$$

Thermal conductivity decreases with temperature:

t, °C	25	200	400	600	800
λ, W m^{-1} K^{-1}	35	28	23	19	16

However, B_4C has low resistance to thermal shock [30].

Boron carbide is a high-temperature *p*-type semiconductor. The width of the forbidden band is 0.8 eV. The electrical conductivity depends on the B : C ratio and impurity content. The specific electrical resistance of B_4C is 0.1 – 10 Ω · cm, i.e., of the same magnitude as that of SiC and graphite [4], [31]. Boron carbide has a negative temperature coefficient of resistivity, similar to other ceramic materials. Boron carbide is notable for its high thermoelectric power, which increases with temperature [5]. Electronic properties, i.e., electric conductivity, thermoelectric power, and the Hall effect, were measured as a function of B : C ratio by WERHEIT et al. [14]. A metal – insulator transition occurs at the composition B_4C.

The nuclear properties of boron carbide are such that it is a very effective neutron absorber in most types of reactors [4]. The effectiveness of boron as a neutron absorber is due to the large cross section of the ^{10}B isotope (≈ 4000 b). The natural abundance of ^{10}B is 19.9 ± 0.3 mol %. The (n, α) reaction

$$^{10}_{5}B + ^{1}_{0}n \rightarrow ^{4}_{2}He + ^{7}_{3}Li + 2.97\,keV$$

produces helium plus lithium — and no radioactive nuclei. (In contrast, cadmium or the lanthanides do produce radioisotopes.) In addition, the secondary gamma radiation is so small that it can be ignored. The cross section is a linear function of the neutron energy and follows a 1/v relation. The cross sections of boron over the entire neutron energy spectrum, 0.01 – 1000 eV, are sufficiently large to make boron effective in thermal, intermediate, and fast reactors.

1.4. Chemical Properties

Boron carbide is one of the most stable compounds. It is not dissolved by mineral acids or aqueous alkali; however, it is decomposed slowly by mixed hydrofluoric – sulfuric acid or hydrofluoric – nitric acid. One technique of refining B_4C from free carbon and boron is based on the differing oxidation rates of the parent elements and boron carbide. The boron carbide powder is heated in a mixture of H_2SO_4, HNO_3, $HClO_4$, and $K_2Cr_2O_7$ [32].

Molten alkali decomposes boron carbide to give borates. Boron carbide reacts with many metal oxides at elevated temperature to give carbon monoxide and metal borides. Boron carbide reacts with many metals that form carbides or borides at 1000 °C, i.e., iron, nickel, titanium, and zirconium. Aluminum [9] and silicon [33] form substitutional compounds with boron carbide. Boron carbide withstands metallic sodium fairly well at 500 °C [34] but is slowly etched by hydrogen at 1200 °C. At that temperature it does not interact with sulfur, phosphorus, or nitrogen. Above 1800 °C it reacts with nitrogen to give boron nitride. Elemental carbon can be dissolved in B_4C at the B_4C – C eutectic temperature of about 2400 °C, precipitating again on cooling [20].

Chlorine attacks B_4C at about 600 °C and bromine attacks it at 800 °C and above. These reactions can be used to prepare boron halides. Boron carbide can be melted without decomposition in an atmosphere of carbon monoxide, but it reacts in the temperature range 600 – 750 °C with CO_2 to form B_2O_3 and CO. Oxidation in air starts at 500 °C and is severe at 800 – 1000 °C. However, the oxidative damage depends to a great extent on the surface area of the object, powder being attacked more severely than bulk samples.

Detailed descriptions of chemical, mass spectrometric, and spectrochemical analysis of boron carbide have been issued by the American Society for Testing and Materials (ASTM) [35].

Etching methods for polished sections for microstructural inspections of dense boron carbide materials are described thoroughly by PREDEL et al. [36].

1.5. Production

Powder. The most common method for large-scale production of boron carbide is *carbothermic reduction* of boric oxide:

$$2\,B_2O_3 + 7\,C \rightarrow B_4C + 6\,CO$$

The process is so strongly endothermic, 1812 kJ/mol or 9.1 kWh/kg, that it is usually carried out at 1500 – 2500 °C in an electric furnace. Either arc furnaces or resistance furnaces, analogous to Acheson furnaces for SiC, are used [4]. The starting material is an intimate mixture of boric oxide and carbon, e.g., petroleum coke or graphite. Large quantities of carbon monoxide, 2.3 m^3/kg, are generated, and boron can be lost by evaporation of B_2O_3 at the high temperatures. In the electrothermic process used by ESK, a subsidiary of Wacker Chemie, the product is cooled, and the unconverted outer-zone material is removed to leave a fused coarse-grained boron carbide of high purity: B plus C ≥ 99 wt %, total metal impurity ≤ 0.2 wt %, N plus O ≤ 0.3 wt %, B : C mole ratio of 4.0 – 4.3. This boron carbide is obtained in the form of regular blocks, which are broken up and milled to produce the B_4C grain size appropriate for final use.

If the carbothermic reduction is carried out below the melting point of B_4C, e.g., at 1600 – 1800 °C, in a vented graphite tube furnace starting from a dehydrated mixture of H_3BO_3, acetylene black or sugar, and ethylene glycol, very fine-grained (0.5 – 5 µm), stoichiometric B_4C powders are produced [37], although productivity is low.

Uniformly sized submicrometer powders of boron carbide (specific surface area 20–35 m^2/g) can be synthesized continuously by rapid carbothermic reduction (RCR) of boron oxide in a graphite transport reactor at ca. 2000°C [38], [39].

Boron carbide powders are produced directly (no need for the expensive comminution) by magnesiothermic reduction of boric oxide in the presence of carbon at 1000 – 1800 °C:

$$2\,B_2O_3 + 6\,Mg + C \rightarrow B_4C + 6\,MgO$$

This reaction is strongly exothermic. It is carried out either directly by point ignition (thermite process) or in a carbon tube furnace under hydrogen [40]. The product is washed free of magnesium oxide and unconsumed magnesium metal with hydrochloric or sulfuric acid and can be further purified by heating under vacuum at 1800 °C. Since the MgO acts as a grain growth inhibitor, ultrafine boron carbide particles on the order of 0.1 – 1.5 µm are obtained.

There are other methods of powder preparation:

1. synthesis from the elements

   $$(4+x)B + C \rightarrow B_{4+x}C$$

2. reduction of boron trichloride by hydrogen in the presence of carbon

   $$4\,BCl_3 + 6\,H_2 + C \rightarrow B_4C + 12\,HCl$$

3. chemical vapor deposition [41]

   $$4\,BCl_3\,(BBr_3) + CH_4 + 4\,H_2 \rightarrow B_4C + 12\,HCl(Br)$$

However, these methods are generally used only on the laboratory scale for preparing high-purity boron carbide crystals or coatings [4].

Dense Shapes. Self-bonded [42] boron carbide shapes having 100 % theoretical density, small final grain size, and high mechanical strength can be produced by *hot pressing* [43]. This is accomplished by placing a loose powder or a compact preform in a graphite die and applying uniaxial pressure of 10 – 30 MPa at a temperature exceeding 2000 °C for 5 – 10 min. Inert or reducing atmospheres are necessary to protect B_4C and the graphite from oxidation. Increasing the boron content to the composition $B_{13}C_2$ has been found to facilitate sintering by reducing the necessary temperatures and to almost double the strength [44]. However, for boron-rich $B_{4+x}C$ compositions, expensive boron nitride-lined graphite molds must be used, to avoid carburization resulting in formation of carbon-rich B_4C [45].

In 1977, ESK developed a process for *pressureless sintering* of boron carbide, a process now used in everyday production [46]. The key observation, that carbon additives are extremely effective in inducing densification of B_4C, was also made by HENNEY, JONES [47] and SUZUKI et al. [48] almost simultaneously. There are two necessary conditions for obtaining a high sintered density:

1. The use of submicrometer B_4C powders, i.e., powders with a mean particle size below 1 µm
2. The use of small amounts of elemental carbon-containing additive (novolac resin), which pyrolyes to amorphous carbon during the heating cycle [29]

Up to the early 1990s, the carbon was added almost exclusively as novolac resin. However, after an improved method of dispersing carbon black (particle size 20 nm) was developed, the resin route was replaced by the carbon black method [49]. Using nanosized carbon black as sintering aid gives a better sintering behavior with very fine grained materials (mean grain size of B_4C = 1 – 4 µm) having an approximate composition of 96 wt % B_4C and 4 wt % C. The properties of sintered B_4C can be improved further by near net shape injection molding [50] and post-HIPing (HIP, hot isostatic pressing) after pressureless sintering [50], [51]. In this post-HIP stage, the sintered shapes are densified to > 99.5 % TD at 1950 – 2050 °C in a high-pressure autoclave (≥ 10 MPa).

The pressureless sintering/HIP route has also been used to prepare B_4C – SiC composites [52], [53] and B_4C – TiB_2 composites [54], [55]. The B_4C – SiC composites combine the good thermal shock and oxidation resistance of silicon carbide with the hardness and low density of boron carbide. In this way, a maximum strength of 550 MPa (four-point bend) was achieved for a composite of 59 wt % SiC, 37 wt % B_4C and 4 wt % C. This composite can be used in oxidizing atmospheres up to 1200 °C [52]. Excellent tribological and wear properties were obtained by KEVORKIJAN et al. [53] for B_4C – SiC composite materials, which suggest their use for heavily loaded mechanical seals and bearings.

The newly developed B_4C – TiB_2 composite materials [54], [55] are electroconductive and combine excellent fracture toughness (see Figure 3), with the high hardness of boron carbide (Table 2). They are used as nozzle materials for hard alumina grits in abrasive air-jet cleaning in dental laboratories. While the above-mentioned composite materials are formed by mixing the desired metal boride phases and subsequent sintering, processes have also been described in which the desired metal boride composition is only formed after a suitable reaction of the starting materials during the sintering step. SKOROKHOD and KRSTIC [56] fabricated a 85 vol % B_4C – 15 vol % TiB_2 composite with a flexural strength of 621 MPa and a fracture toughness of 6.1 MPa · $m^{1/2}$ (measured by the SENB method with 100 µm notch width) by reaction hot pressing of a boron carbide powder with submicrometer particle size using additions of submicrometer-sized TiO_2 and carbon at 2000°C. The formation of uniformly distributed TiB_2 particles (< 5 µm grain size) was in accordance with the reaction.

$$(1+x)B_4C + 2TiO_2 \rightarrow xB_4C + 2TiB_2 + 4CO$$

Another approach to produce tough and high-strength B_4C ceramics is liquid-phase sintering. LEE and KIM [57] showed that addition of alumina (Al_2O_3) promotes densification of boron carbide, and a maximum density of 96% of theoretical can be obtained with 3 wt % alumina-doped B_4C sintered at 2150°C, i.e., above the melting point of Al_2O_3. The microstructure showed equiaxial B_4C grains with a mean grain size of about 7 µm. However, when the addition of Al_2O_3 exceeded 3 wt %, exaggerated grain growth occurred, which was attributed to the liquid phase. It has been reported [58] that by hot pressing of B_4C with alumina additions of up to 5 vol % at 2000 °C the mechanical properties can be remarkably increased as compared to undoped, hot-pressed B_4C of 88 % relative density. Fracture toughness increased steadily with the addition of Al_2O_3 from ca. 3 MPa · $m^{1/2}$ (2.5 vol % Al_2O_3) up to 3.8 MPa · $m^{1/2}$ (10 vol % Al_2O_3). However, the achieved flexural strength was below 560 MPa.

The fracture toughness of boron carbide can be improved by addition of yttria [59] or of yttria in combination with other oxides [60]. The fracture toughness of B_4C can be increased to 3.9 MPa·$m^{1/2}$ by vacuum sintering at 1900 – 1975°C of powder compacts of composition 97.5 wt % B_4C and 2.5 wt % carbon packed in a yttria grit of 0.15 – 1.4 mm grain size [59]. The vacuum allows yttria vapor to penetrate the powder compact and promote reaction sintering of carbon-doped B_4C to full density (2.62 g/cm^3). X-ray diffraction showed that yttrium boride and yttrium borocarbide were coexistent with B_4C. The method disclosed in [60] uses liquid-phase sintering under vacuum or streaming argon of powder compacts comprising B_4C powder (average particle size 0.6 – 3.5 µm) and 1 – 28 wt % additions of Y_2O_3 in combination with Al_2O_3 or aluminum nitride (AlN) and La_2O_3 or CeO_2.

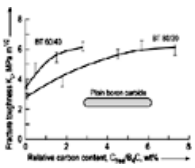

Figure 3. Fracture toughness of B_4C – TiB_2 composites and self-bonded B_4C as a function of free carbon content

60/40 and 80/20 indicate B_4C/TiB_2 vol % in the starting mixture [54].

1.6. Uses

The uses of boron carbide are described in [4-6], [30], [31], [61].

Uses Based on Hardness. The major industrial use of boron carbide is as abrasive grit or powder. Commercial grits are available in particle sizes from 1 µm to 10000 µm and predominantly used as polishing, lapping, and grinding media for hard materials, especially cemented tungsten carbide and fine ceramics. Water and commercial oils serve as lubricant and coolant. Available as a substitute for loose powders is B_4C paste, which can withstand temperatures up to 350 °C.

The second category is wear-resistant engineering components made of dense, sintered shapes. Boron carbide sand-blasting nozzles are now the most common form of such a component. These nozzles are characterized by minimum wear, even with corundum or silicon carbide grit. Sintered-B_4C

wheel dressing sticks are used to produce new cutting edges on surface grains of grinding wheels with minimum wear. So-called hand laps are used to render the cutting edges of cemented carbide tools keen and smooth. Mortars and pestles of boron carbide are suitable for grinding hard materials without contamination from the grinding equipment.

Ultrahard substances are nonductile. This factor, together with a density of only 2.5 g/cm^3, is fully exploited in the use of dense B_4C as ceramic plates against armor-piercing bullets of caliber up to the thickness of the plates [61]. Lightweight ceramic tiles have been used extensively for the protection of helicopters; monolithic breastplates, for the protection of personnel [62], [63]. Hot pressings made from a finer grade of boron carbide are used in air bearings of inertial navigation equipment.

Uses in Nuclear Technology [30], [64]. Boron carbide is both cheaper and easier to fabricate than the element boron itself. As a result it has found almost exclusive usage as a control rod material, neutron poison, and a neutron shielding material. The bulk density of vibro-packed B_4C powder, better the sintered density of self-bonded boron carbide pellets, or the ^{10}B : ^{11}B ratio can be varied to adapt absorber efficiency to the specific conditions in thermal (BWR, PWR), high-temperature (HTR), and fast breeder reactors (FBR).

A ^{10}B-enriched boron carbide, e.g., corresponding to ≤ 92 mol % ^{10}B, is commercially available; thus, the nuclear worth of an absorbing element can be increased nearly 4.5-fold by substituting enriched for natural material, without a change in dimensions. The helium formed during reactor exposure is partially retained within the solid B_4C as micropores, but the fraction released (≦ 50 %) must be accommodated by some aspect of control assembly design, such as a plenum within an individual pin or a reliable pressure relief vent. The He and Li atoms, because of their relatively large size, cause some radial swelling (≦ 10 %) of the B_4C pellets.

In a typical thermal reactor the operating temperature of the absorber is below 500 °C, and low-density (ca. 75 % of theoretical density) B_4C pellet columns clad in metal sheets have proved adequate for large-scale commercial use. In gas-cooled reactors, boronated graphite containing about 40 wt % B meets the high-worth control rod requirements. Ceramic-based dispersions, i.e., combinations of relatively small amounts of B_4C in an inert matrix, exhibit smaller macroscopic neutron absorption cross sections but swell less and release less gas. So-called burnable poison rods, which are consumed at the same rate as the fuel, are commonly designed in this fashion. Examples are Al_2O_3 – 3 vol % B_4C for water-cooled reactors and boronated graphite (ca. 6 wt % B_4C) for gas-cooled reactors.

^{10}B-enriched B_4C has been chosen as the absorber for fast-breeder reactor control material. A potential alternative for future fast breeder generations is ^{10}B-enriched europium hexaboride, which has better engineering properties and which loses its absorption reactivity because of ^{10}B burnup about 40 % slower than B_4C [65]. Carbon-bonded boron carbide plates (TETRABOR®) [66] provide for the closest possible arrangement of spent fuel elements in high density intermediate storage pools. Sizes range up to 500 mm in length and down to 2 mm in thickness. These shield plates with > 50 vol % B_4C usually contain more absorber material than aluminum-bonded B_4C (Boral) and silicone-bonded B_4C plates (ESK-Elastosorb). Doors, windows, or walls in laboratories where radioactive matter is investigated are screened with B_4C plates. Their low density of ca. 1.6 g/cm^3 makes them excellent for lining containers for nuclear fuel or any matter radiating neutrons.

Chemical Uses. Boron carbide powder is used to introduce boron into the surface layer of steels and other ferrous materials [31], [67]. This takes place by diffusion of boron into the surface layer, resulting in the formation of a thin (10 – 1000 µm) but hard, wear-resistant layer of iron boride, Fe_2B. Vickers hardness numbers of 2400 HV-0.1, corresponding to tungsten carbide or α-Al_2O_3, were measured on the surface of high-alloyed tool steels. The surfaces of cemented carbides (WC–Co, WC–TiC–TaC– Co/Ni/Fe, and Ferrotic) and stellite alloys can even be hardened further by boron [68]. According to the EKabor powder packing method B_4C – SiC powder mixtures with added sodium tetrafluoroborate activator are used.

Its high boron content makes B_4C useful as a raw material for the production of other boron compounds, including the boron halides and metal borides [4]. Much of the BCl_3 made commercially is used in the production of boron filament for reinforcement of metals and plastics. The exceptionally high heat of combustion of B_4C makes powders with a mean particle size of 2 µm useful energy carriers in solid rocket propellants [4].

Uses in Refractories. In carbon-bonded refractories, B_4C powder is used as an antioxidant. Small additions of B_4C to the refractory material reduce the burn out of the carbon and improve the corrosion resistance against slags and steel melts. A simultaneous addition of B_4C and a metallic antioxidant results in better oxidation resistance of the refractories than the addition of B_4C or the metallic antioxidant alone [69], [70]. The effect depends on the oxidation of the antioxidants and reaction with the oxides of the refractory material, which creates oxidic phases and gaseous components. Both prevent the infiltration of molten slag or steel, and the oxide phases protect the carbon against oxidation by forming a protective oxidic layer at the hot face of the refractory.

Uses in Metal Matrix Composites. Metal matrix composites (MMCs) made of B_4C and Al or Al-alloy have a higher stiffness, strength and hardness and lower weight than the metal itself. The content of B_4C in the MMCs is up to 50 wt %. This material could be used to produce magnetic disks for hard disk drives, automotive components, sports goods, etc. [71-73].

Other Uses. The thermoelectric properties of $B_{13}C_2$ [74] are such that $B_{13}C_2$ could be an interesting material for high-temperature thermoelectric conversion. Thermoelements made of the couple B_4C – C can be used for temperature measurements up to 2300 °C [75].

1.7. Economic Aspects

ESK is the world's largest producer of boron carbide. The granular and sintered B_4C products of this company are sold commercially under the trademark TETRABOR®. The price of B_4C is 10 – 150 $/kg, depending on grit size. Annual world production is about 3500 t.

[Top of Page]

2. Boron Nitride

The elements boron and nitrogen, both neighboring carbon in the periodic table, form the 1 : 1 compound BN [10043-11-5], which has allotropes isostructural to the allotropes of carbon:

1. Alpha-BN, hexagonal modification with a layered structure similar to graphite, sometimes called white graphite
2. Beta-BN, high-pressure diamond-like modification, cubic zinc blende structure
3. Gamma-BN, dense hexagonal modification, wurtzite structure

More detailed reviews of the properties, preparation, and uses of the various BN modifications are given in [76-78], [63].

2.1. Properties

α-BN. The most common form of boron nitride is the graphitelike hexagonal, used increasingly because of its unique combination of properties: low density, 2.27 g/cm^3; high-temperature stability, *mp* near 3000 °C; chemical inertness, especially its resistance to acids and molten metals and its stability in air up to 1000 °C and in inert gases like N_2, CO, and Ar up to 2700 °C; stability to thermal shock; workability of hot-pressed shapes; extremely low electrical conductivity, which makes it an excellent insulator; and high thermal conductivity.

In *thermal conductivity* BN ranks with stainless steel at cryogenic temperatures and with beryllium oxide, BeO, at higher temperature. Above 700 °C the thermal conductivity of α-BN exceeds that of BeO. Boron nitride has a dielectric constant of 4, one half that of α-Al_2O_3, which also has high dielectric strength. It is a lubricant over a wide range of temperatures; its low coefficient of friction is retained up to 900 °C, whereas other solid high-temperature lubricants, e.g., graphite and molybdenum disulfide, burn away well below 900 °C [80].

Because of its *stability* at high temperature and inertness towards carbon and carbon monoxide up to 1800 °C, it is used as a refractory ceramic. In this function it is superior to the nitride ceramics Si_3N_4 and AlN and the oxide ceramics MgO, CaO, and ZrO_2. It is not wet by and thus is stable against attack by molten glass, molten silicon, boron, nonoxidizing slags, molten salts such as borax and cryolite, and reactive molten metals, including aluminum, iron, copper, and zinc. Because it sinters poorly, dense shapes of α-BN are produced exclusively by hot pressing. Two material types are important for practical applications

1. Shaped bodies of pure hexagonal BN
2. Shaped bodies of hexagonal BN with addition of refractory oxides, carbides and nitrides with BN contents ≥ 20 wt % (composite ceramics)

Typical properties of hot-pressed BN and BN – ZrO_2 composite ceramics are summarized in Table 2.

Table 2. Property data on boron nitride ceramics

	Hot-pressed BN	Hot isostatically pressed BN	Hot-pressed composite ceramics BN/ZrO_2
Bulk density, g/cm^3	2.0	2.2	2.8 – 3.6[b]
Porosity, vol %	< 7	< 1	< 7
Modulus of rupture (4-point), [a] MPa	100/80	50	130/70
Young's modulus, [a] GPa	70/35	30	80/35
Thermal conductivity, [c] $Wm^{-1}K^{-1}$			
20 °C	65/45	50	35 – 25/20 – 18 [b]
400 °C	50/30	40	31 – 21/17 – 15
700 °C	30/20	30	28 – 18/15 – 13
1000 °C	15/10	20	25 – 15/13 – 18
Thermal expansion coefficient, [a] 10^{-6} K^{-1} (at 20 – 1000 °C)	1.2/8.0	4.0	4.5 – 9.0/8.5 – 11.0 [b]
Specific heat, J/gK (at 20 °C)	0.8	0.8	0.7
Electrical resistivity, Ωcm (at 20 °C)	> 10^{12}	> 10^{12}	> 10^{12}
Dielectric strength, kV/mm (at 20 °C)	> 6	> 6	> 6

[a] Test orientation vertical/parallel to pressing direction.
[b] Dependent on BN/ZrO_2 ratio.

Most chemical and physical properties of axial hot-pressed BN shapes depend on the nature and amount of densification additives (≤ 5 wt % B_2O_3 or metal borates) [77]. In addition, the coefficient of expansion and thermal conductivity as well as the flexural strength and Young's modulus are *anisotropic;* they vary with the direction of hot-pressing. In this respect, BN is similar to graphite. However, isostatic hot-pressing (HIP) of canned BN powder, produces pure isotropic α-BN shapes of theoretical density. Such shapes possess superior thermal and mechanical properties (see Table 2) [81].

β-BN. The cubic form of boron nitride, β-BN, the commercial Borazon, resembles diamond in crystal structure and other properties. Pure β-BN is colorless and a good electrical insulator. Doping with Li_3N or the elements Be, Si, C, or P turns β-BN yellow to black and changes other properties, such as conductivity or toughness.

The theoretical density of β-BN is 3.48 g/cm^3. The Knoop hardness number HK-0.1 is ca. 4700, somewhat less than the > 7000 for diamond. Because β-BN is stable in air up to 1400 °C — diamond burns away at only 800 °C — it is an effective high-temperature abrasive.

γ-BN. Wurtzite boron nitride, γ-BN, which is metastable at the pressures and temperatures where β-BN is fabricated, has a density close to cubic β-BN. Above 5.5 GPa and 1300 °C γ-BN transforms to cubic β-BN [82].

2.2. Production

Hexagonal α-BN was prepared for the first time by BALMAIN [83] in the middle of the nineteenth century. It remained a laboratory curiosity until the mid-1900s, when hot-pressed α-BN shapes were made for the first time [84]. Although there are a number of general methods of producing *powder* [76], principally two reactions are used on an industrial scale:

1. Reaction of boric oxide or boric acid with ammonia in the presence of a carrier, often tricalcium orthophosphate:

$$B_2O_3 + 2\,NH_3 \xrightarrow{900\,°C} 2\,BN + 3\,H_2O$$

   A second heat treatment for purification and stabilization is carried out > 1500 °C under N_2.
2. Reaction of boric acid or borax with organic nitrogen compounds, e.g., urea or melamine:

$$B_2O_3 + CO(NH_2)_2 \xrightarrow{>1000\,°C} 2\,BN + CO_2 + 2\,H_2O$$

Method 1 usually gives crystalline α-BN in the form of thin hexagonal platelets with a thickness of about 0.1 – 0.5 µm and a diameter of up to 5 µm. Method 2 can give a “turbostratic” boron nitride, a hexagonal structure characterized by partial or complete absence of three-dimensional order among its lamellae [85], [86].

Dense α-BN is produced in two ways:

1. Axial hot-pressing [84] or hot isostatic pressing [81] of fine powders to give billets that can be easily machined
2. Pyrolysis of a boron trichloride – ammonia mixture,

   $$BCl_3 + NH_3 \rightarrow BN + 3\,HCl$$

   on a graphite substrate on which the α-BN is deposited.

Pyrolytic BN, though pore-free and impervious to helium, is anisotropic because the α-direction layer planes of the hexagonal structure deposit parallel to the substrate surface [87]. α-BN is also available in the form of fibers [88], thin films [89], and multidimensional fiber infiltrated BN – BN composites [90].

The cubic zinc blende form of boron nitride was first prepared from the graphitic form by applying pressures of 4 – 6 GPa (40 – 60 kbar) at 1400 – 1700 °C in the presence of a lithium or magnesium nitride catalyst [91]. Later, the use of the ternary compound $Ca_3B_2N_4$ as the catalyst has been proposed to produce β-BN crystals of outstanding purity, quality, and size (0.1 – 0.3 mm) [92]. Details of the various high-pressure, high-temperature (HP-HT) syntheses and low-pressure syntheses for β-BN can be found in [63]. Deposition of β-BN coatings, mainly produced by PVD and plasma CVD methods, has been investigated. Due to the difficulties of growing large β-BN crystals and depositing thick layers, the low-pressure sintering method is not commercially used today.

Cubic BN particles can be sintered into strong masses at the high temperatures and high pressures at which this diamond-like form is stable [93]. Also starting from shock-synthesized γ-BN of 0.1 – 0.5 µm particle size and hot pressing at 6.7 GPa and 1600 °C produces self-bonded compacts of theoretical density and predominantly cubic structure, 80 % β – 20 % γ-BN [82].

Wurtzite-type boron nitride, γ-BN, can be obtained by static compression above 6.0 GPa or by dynamic compression at 12 – 13 GPa below 1700 °C [94], [95].

2.3. Uses

The many uses for α-BN powder and hot-pressed shapes are summarized in Table 3, which classifies specific properties. As can be seen from this table, α-BN has established itself in the various fields of chemistry, metallurgy, high-temperature technology, electrotechnics, and electronics. α-Boron nitride powder is also available in the form of coating suspensions and as a convenient aerosol spray. Annual world production of α-BN powder is ca. 1000 t (ca. 2005); the price of α-BN is 30 – 100 $/kg.

Table 3. Uses and potential uses for α-boron nitride

Use	Property desired					
	R	T	E	I	L	M
α-BN powder						
Solid lubricant for high-temperature bearings	+				+	
Mold release for die casting of glass and metals	+			+	+	
Active filler for rubber, resins, and plastics	+	+	+	+	+	
Additive to oils and high-temperature greases				+	+	
Ultrahigh-pressure transmitting agent	+			+	+	
Coatings for evaporation plants, as parting agent for deposited metal films	+			+		
Raw material for β-BN	+					
Component for cosmetic applications				+	+	
Coatings of tundishes for nonferrous metals	+				+	
Mold release for injection molding of plastics					+	
Coating for graphite hot-pressing molds	+			+	+	
Embedding medium for heating wires	+		+	+		
Boron source for preparation of other boron compounds	+					
α-BN hot-pressed shapes						
Crucibles for melting glass and metals	+	+		+		+
Break rings for horizontal continuous casting of high-alloy steels	+			+	+	+
Break rings for low-alloy steels	+			+	+	+
Spanker plates used in the production of TV tubes	+			+		+
End-dam material for thin-strip casting	+			+	+	+
Seal rings for gas sensors (lambda sensor)	+		+	+		+
Components for high-temperature electric furnaces	+	+	+	+		+
Structural parts for magnetohydrodynamic devices (MHD)	+		+	+		+
Dielectric for radar antennas and windows			+			+
	+		+			+

Insulators for low- and high-frequency equipment						
Insulators for plasma-jet furnace, arc pulse generators, and ion engines	+		+			+
Holders, mounting plates, substrates, and heat sinks in valve and transistor circuits		+	+			+
Pump components, pipes, and nozzles for handling liquid metals	+			+		+
Protective tubes and insulating sleeves for thermocouples	+	+	+	+		+
Protective sleeves for electrodes in automatic welding	+		+	+		+
Wafers for boron doping of semiconductors	+					+
Molds for hot-pressing of ceramics like B_6O or $B_{6.5}C$	+	+		+	+	+
Neutron absorbers and shields for nuclear reactors	+	+		+		+

R high-temperature refractoriness; T thermal conductivity; E electrical insulator; I chemical inertness, nonwetting; L lubricity; M machining ease.

Since cubic boron nitride is the second hardest material after diamond, its uses are similar to those of diamond — machining and drilling. Grinding wheels made from resin- or metal-bonded β-BN grits in many cases are superior and more economical than diamond wheels because of their better chemical resistance and greater toughness. Sintered dense compacts of β-BN are being used increasingly as cutting tools for hard steels and tough nickel-based superalloys, as wire drawing dies, and as cutting heads in rock drills [63]. The longer life of cutting tools containing wurtzite-type γ-BN in addition can be related to the lubrication provided by the α-BN that forms from the less stable γ-BN during cutting [96].

[Top of Page]

3. Metal Borides

Boron combines with a large number of metals and semimetals to form binary or higher solid compounds, the so-called borides. Most binary borides, especially those formed from transition metals having high melting points, are metallic high-temperature materials that can be grouped with the carbides, nitrides, and silicides as refractory hard materials [1], [97]. Although boride cermets [98], [99] that combine the extreme hardness of refractory borides with the high toughness and ductility of a metallic binder have been recently developed, they are far from attaining large-scale application.

Nevertheless, metal borides have some unique properties and are used in some specialized fields of modern industry, especially when the requirements cannot be met by other materials. More information about metal borides is available in [31], [61], [100-108], [109]], [7].

3.1. Properties

Metal borides are unique in the number of stoichiometries; compositions corresponding to at least 24 M : B ratios between 5 : 1 and 1 : 66 are known. However, the most common are the monoborides, MB; diborides, MB_2; tetraborides, MB_4; hexaborides, MB_6; dodecaborides, MB_{12}; and hectoborides, MB_{66}. The binary compounds of boron are given in the form of a periodic table in Table 4. As can be seen, many metal – boron systems contain three or more intermediate phases.

Table 4. Binary compounds of boron

Main group elements						
H_3B [a]						
Li_5B_4	Be_4B		$(B_{12}C)$ [b]	$(B_{12}N_2)$	BO	BF_3
Li_6B_{19}	Be_2B		$(B_{13}C_2)$	BN	B_2O_2	
Li_3B_{14}	BeB_2		$B_{4.3}C$		B_2O_3	
LiB_6	BeB_6				BO_2	
(LiB_{10})	(BeB_{12})				BO_3	
					$B_{12}O_2$	
NaB_6	MgB_2	AlB_2	$(B_{14}Si)$	$B_{12}P_2$	$B_{12}S_2$	B_2Cl_4
NaB_{15}	MgB_4	AlB_{12}	$B_{12}Si_2$	$B_{13}P_2$	BS	BCl_3
	(MgB_6)		B_3Si	BP	B_2S_2	
	MgB_7				B_2S_3	
	(MgB_{12})				BS_2	
					B_2S_5	
KB_6	CaB_4			$B_{12}As_2$	$B_{12}Se_2$	BBr_3
	CaB_6			$B_{13}As_2$	BSe	
				BAs	B_2Se_3	
					BSe_2	
(RbB_6)	SrB_6					BI_3
(CsB						

$_6$)	BaB_6							
Transition elements								
ScB_2	TiB	V_3B_2	Cr_2B	Mn_2B	Fe_2B	Co_3B	Ni_3B	
ScB_{12}	TiB_2	VB	Cr_5B_3	MnB	FeB	Co_2B	Ni_2B	
		V_5B_6	CrB	Mn_3B_4		CoB	Ni_4B_3	
		V_3B_4	Cr_3B_4	MnB_2			NiB	
			Cr_2B_3					
		V_2B_3	CrB_2	MnB_4				
		VB_2	CrB_4					
YB_2	ZrB_2	Nb_3B_2	Mo_2B	Tc_3B	Ru_7B_3	Rh_7B_3	Pd_4B	(AgB_2)
YB_4	ZrB_{12}	NbB	MoB	Tc_7B_3	$Ru_{11}B_8$	Rh_5B_4	Pd_3B	
YB_6		Nb_5B_6	MoB_2	TcB_2	RuB	RhB	Pd_5B_2	
YB_{12}		Nb_3B_4	$MoB_{2.3}$		Ru_2B_3		Pd_2B	
YB_{66}		Nb_2B_3	$Mo_{1-x}B_3$		RuB_2			
YB_{66}		NbB_2						
*	HfB	Ta_2B	W_2B	Re_3B	OsB	$IrB_{0.9}$	Pt_4B	(AuB_2)
	HfB_2	Ta_3B_2	WB	Re_7B_3	Os_2B_3	$IrB_{1.1}$	Pt_2B	
		TaB	WB_2	ReB_2	OsB_2	$IrB_{1.3}$	$PtB_{0.7}$	
		Ta_5B_6						
		Ta_3B_4	W_2B_5					
		TaB_2	$W_{1-x}B_3$					
**								

*	LaB_4	CeB_4	Pr_2B_5	Nd_2B_5	PmB_4	Sm_2B_5	EuB_6	Gd_2B_5	TbB_2	DyB_2	HoB_2	ErB_2	TmB_2	YbB_2	LuB_2
	LaB_6	CeB_6	PrB_4	NdB_4	PmB_6	SmB_4		GdB_2	TbB_4	DyB_4	HoB_4	ErB_4	TmB_4	YbB_4	LuB_4
	LaB_9		PrB_6	NdB_6		SmB_6		GdB_4	TbB_6	DyB_6	HoB_6	ErB_{12}	TmB_{12}	YbB_6	LuB_{12}
				NdB_{66}		SmB_{66}		GdB_6	TbB_{12}	DyB_{12}	HoB_{12}	ErB_{66}	TmB_{66}	YbB_{12}	LuB_{66}
								GdB_{66}	TbB_{66}	DyB_{66}	HoB_{66}			YbB_{66}	
**		ThB_4		UB_2	NpB_2	PuB_2	AmB_4	CmB_4							
		ThB_6		UB_4	NpB_4	PuB_4	AmB_6	CmB_6							
		ThB_{66}		UB_{12}	NpB_6	PuB_6	(AmB_{12})	(CmB_{12})							
					NpB_{12}	PuB_{12}	AmB_{66}	CmB_{66}							
					NpB_{66}	PuB_{66}									

[a] And higher boranes.
[b] Parentheses around a compound indicate that its existence has not been confirmed. Note: high pressure phases are not listed.

A classification of borides was first given by KIESSLING [110] on the basis of boron structural elements typical for various M : B ratios. With increasing M : B ratio the tendency to form boron – boron bonds increases, and pairs (M_3B_2), zigzag chains (MB), branched chains ($M_{11}B_8$), double chains (M_3B_4), plane nets (MB_2), and finally three-dimensional arrays of boron atoms, including cross-linked nets (MB_4), interconnected B_6 octahedra (MB_6), B_{12} cubic octahedra (MB_{12}), and interconnected B_{12} icosahedra (MB_{66}) are all found.

The most characteristic properties of the metal borides are their high melting points, extreme hardness, and in many cases high electrical and thermal conductivities, fair corrosion resistance, good wear resistance, and thermal shock resistances superior to those of oxide ceramics. Some physical properties of the most common and typical metal borides are given in Table 5. The most stable binary borides at high temperatures are the diborides of titanium, zirconium, and hafnium, each melting above 3000 °C. The melting points of the boron-rich borides MB_n, $n \geqq 2$, are generally higher than those of the parent transition metals. On the other hand, melting points of metal-rich borides are often as low as those of their parent metals. Most of these borides are paramagnetic. Knoop hardness numbers HK-0.1 usually fall in the range of 1100 – 2600 for diborides, 1650 – 2100 for hexaborides, and 2300 – 2600 for dodeca- and hectoborides; therefore, many borides are harder than WC or α-Al_2O_3, two cutting and grinding materials widely used. Hardness and high melting points reflect the properties of the rigid, three-dimensional boron frameworks.

Table 5. Physical properties of some refractory borides [61]

Boride	CAS Registry	Density,	Melting	Electrical	Knoop

	no.	10^3 kg/m^3	point K	resistivity, 10^{-8} Ω · m	hardness, 0.1 kp load
TiB_2	[12045-63-5]	4.52	3470	9 – 15	2600
ZrB_2	[12045-64-6]	6.09	3520	7 – 10	1830
ZrB_{12}	[12046-91-2]	3.61	2520	60 – 80	2580
HfB_2	[12007-23-7]	11.2	3650	10 – 12	2160
VB_2	[12007-37-3]	5.10	2670	16 – 38	2110
NbB_2	[12007-29-3]	7.21	3270	12 – 65	2130
TaB_2	[12007-35-1]	12.60	3370	14 – 68	2500
CrB_2	[12007-16-8]	5.20	2170	21 – 56	1100
Mo_2B_5	[12007-97-5]	7.48	2370	18 – 45	2180
W_2B_5	[12007-98-6]	13.1	2470	21 – 56	2500
Fe_2B	[12006-85-8]	7.32	1663		1800
FeB	[12006-84-7]	7.15	1820	30	1900
CoB	[12006-77-8]	7.32	1535	26	2350
NiB	[12007-00-0]	7.39	1325	23	
LaB_6	[12008-21-8]	5.76	2985	7 – 15	2010
EuB_6	[12008-05-8]	4.91	2890	80 – 170	1840
UB_4	[12007-84-0]	9.38	2768	30	1850
UB_{12}	[12008-80-9]	5.65	2500	22	2630
CaB_6	[12007-99-7]	2.46	2540	160	1650
SiB_6	[12008-29-6]	2.43	2140	2×10^5	2140
B_4C		2.52	2720	$10^5 – 10^7$	3000
β-B		2.35	2420	10^{12}	2600

The electrical characteristics, however, cover the entire spectrum: semiconductors, the MB_{66}, MB_6, and MB_{12} phases of Be, Mg, Ca, Eu, Al, and Si; metallic conductors, TiB_2, ZrB_2, and the majority of transition-metal borides; superconductors, NbB, YB_6, and ZrB_{12}.

In addition, LaB_6 and other lanthanide and actinide borides (YB_6, ThB_6, GdB_6, etc.) belong to the best-known high-temperature electron emitters. Most of these borides are intensively colored: for example, ZrB_{12} is pink, GdB_6 is blue, LaB_6 is purple, and ThB_6 is scarlet [61, 111].

The metal borides display resistance to oxidation in air at elevated temperatures, most of them up to 1000 °C, and to attack by molten metals, basic slags, and molten salts. Except for alkali metal and alkaline earth metal borides, which are saltlike and ionic, they are not readily attacked by nonoxidizing acids and aqueous alkalis. Resistance to oxidation is greatest for the transition metal diborides and is lower for greater boron contents. The controlling factor is the surface layer of parent boron and transition metal oxides, which prevents further oxidation up to temperatures at which boric oxide has appreciable volatility. Most metal borides are inert in hydrogen, nitrogen, and carbon, even at high temperatures. Chlorine and fluorine react vigorously with all borides, fluorine even at 400 °C. All types of borides are readily attacked by oxidizing molten salts such as nitrates and carbonates and hydroxides in the presence of air, and this is a usual method of decomposition for chemical analysis [112].

3.2. Production

Metal borides can be prepared by the following high-temperature reactions [61]:

1. synthesis from the elements by melting, sintering, or hot-pressing
2. borothermic reduction of metal oxides
3. aluminothermic, silicothermic, or magnesiothermic reduction of metal oxide – boric oxide mixture
4. carbothermic reduction of metal oxide – boric oxide mixtures, e.g.,

$$TiO_2 + B_2O_3 + 5\,C \rightarrow TiB_2 + 5\,CO$$

5. reduction of metal oxide with carbon and/or boron carbide, i.e., the boron carbide method, e.g.,

$$2\,ZrO_2 + B_4C + 3\,C \rightarrow 2\,ZrB_2 + 4\,CO$$

$$Eu_2O_3 + 3B_4C \rightarrow 2\,EuB_6 + 3\,CO$$

6. electrolysis of fused salts containing metal oxide and boric oxide
7. auxiliary-metal bath method, M + B being dissolved in molten Al, Cu, Sn, or Pb
8. deposition from the vapor phase involving coreduction of a metal halide – boron halide mixture by hydrogen, optionally as a plasma [113]:

$$TiB_2 + 2\,BCl_3 + 5\,H_2 \rightarrow TiB_2 + 10\,HCl$$

Reactions 1, 2, 6, 7, and 8 give pure products but are almost always restricted to the laboratory. For large-scale production, borides are prepared by reactions 4 and 5 followed by purification. These reactions are usually carried out in electric furnaces much like those for production of boron carbide. Molten metal borides often contain up to 3 wt % carbon and boron carbide, although these impurities can be scavenged with additional metal oxide or boric oxide on reheating the powdered mixture above 1400 °C under a partial vacuum. After cooling, the metal boride is crushed and milled to its final grain size. Monolithic boride shapes usually are fabricated from commercial boride powders by powder metallurgy techniques (⟶ Powder Metallurgy and Sintered Materials):

1. cold forming — isopressing, slip casting, etc. — followed by pressureless sintering of the powder compact [113-116], [117]

2. axial hot-pressing [118-121] and hot isostatic pressing [116], [122] (see Section Production, production of B_4C powder)

TiB_2 – Fe hard metals have been developed to sintered densities above 99 % TD by using fine-grained raw materials and vacuum sintering [98], [123-128]. The microstructure contains some Fe_2B and Ti_2O_3, which can be avoided by additions of Ti or TiH_2. The interesting strength, toughness and hardness properties of these materials (see Figure 4) suggest some future applications.

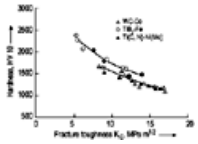

Figure 4. Hardness and fracture toughness of TiB_2 – Fe hard metals compared with commercial hard metals [129]

Cemented ternary transition metal borides have also been developed based on, e.g., $Mo_2(Fe, Ni)B_2$ as the hard phase sintered with iron metals (Fe, Ni, Co). Especially the Mo_2FeB_2-base cermet shows excellent wear behavior [130], [131]. Mo_2NiB_2 cermets were produced from a Mo_2NiB_2-type complex boride as a hard phase and a Ni-base binder by a novel sintering technique, named *reaction boronizing sintering*. The phase transformation of the Mo_2NiB_2-type complex boride from orthorhombic to tetragonal with optimal addition of Cr and V was essential for improvement of the mechanical properties (flexural strength, 2.5 GPa; Rockwell A hardness, ca. 90) and microstructural refinement of the cermet [99].

3.3. Uses

The principal use for metal diborides is as a crucible material for nonferrous metals, especially for aluminum, copper, magnesium, zinc, tin, and lead. Monolithic shapes of dense TiB_2 [132] are used for the electrolytic production of aluminum with the Hall – Héroult cell [133]. Hot-pressed TiB_2 armor plates are used in the protection of ground-based military vehicles. Their inertness to molten aluminum and cryolite and excellent electrical conductivity suggest titanium diboride and zirconium diboride as cathode leads, electrodes, and thermocouple sheaths in aluminum metallurgy [134]. Related uses are TiB_2 – BN or TiB_2 – BN – AlN hot-pressed composite crucibles, so-called evaporation boats, for the continuous evaporation of aluminum metal for vacuum metallizing [135]. The composites show high corrosion resistance against molten metals and good electrical conductivity (TiB_2 has high electrical conductivity). The crucibles/boats are heated in direct current, and by varying the BN content the electrical conductivity of the boats can be controlled. Coatings of BN are used in a wide range of applications due to the extraordinary properties of BN. They are used in the aluminum and steel industries to enhance corrosion resistance and reduce wear on sliding parts like crankshafts for compressors. Because of the nonwettability of BN against many metallic and nonmetallic melts various parts are often coated with BN, e.g., silicon in the semiconductor industry, or BN is used for linings to prevent reactions with Si or Al. Coatings can be produced by CVD techniques to obtain well-crystallized linings with relatively high strength. Also BN layers can be produced by spraying, brushing, dipping, and pouring. The coatings provide improved high-temperature electrical insulation for the semiconductor industry. In molding equipment BN coatings reduce sticking and reactions between powder and mold. High-temperature lubrication properties are important in aluminum and steel manufacturing processes such as casting. On BN-coated preforms, metal layers can be deposited and easily peeled off to give free-standing metal sheets. BN coatings on SiC or Al_2O_3 fibers reduce the thermal mismatch between fiber and matrix and also reduce fiber – matrix interfacial shear strength in ceramic matrix composites to give higher overall strength and toughness [63].

TiB_2 – Fe cermets are suitable for machining aluminum alloys [129], whereas ternary boride cermets based on Mo_2FeB_2, Mo_2NiB_2 and WCoB are suitable as machine parts for injection molding, bearings, wire-drawing cones, and cutting tools [99], [131]. The world production of TiB_2 is ca. 250 t/a, much of this for the manufacture of these resistance-heated evaporation boats. The price of TiB_2 powder is 30 – 60 $/kg.

Calcium hexaboride, CaB_6, is used as a deoxidizer for high-conductivity copper [136-138]. CaB_6 and ZrB_2 are used as antioxidants in carbon-bonded refractories (cf. Uses in Refractories.). For ZrO_2 – C submerged nozzles used in the continuous casting of aluminum a partially or total change of ZrO_2 to ZrB_2 at the powder can more than double corrosion resistance [139]. A major use of NiB, CrB, and CrB_2 is the production of welding and hard-facing alloys based on nickel – chromium – boron – silicon, called colmonoy [140]. An iron – boron alloy 85 wt % Fe/15 wt % B is used to introduce boron into steel (at about 0.002 wt %) in order to improve hardenability [141]. The ternary boride $Fe_{14}Nd_2B$ has been developed for use as a magnetic material (⟶ Magnetic Materials) [142].

The hardness and wear properties of borides are utilized economically as thin coatings on metal surfaces, prepared by the so-called boronizing process (cf. the chemical uses of B_4C in Chemical Uses.). Lanthanum hexaboride is useful as a high-current-density electrode for electron microscopes, electron-beam furnaces, and other devices of highest electronic emissivity. The lifetime of LaB_6 electrodes is reported to exceed that of tungsten cathodes by two orders of magnitude [143, 144].

The refractory hexaboride of europium was suggested as a neutron absorber to control the power of fast breeder reactors. As a neutron absorber EuB_6 is equivalent to 35 mol % ^{10}B-enriched boron carbide, and its nuclear worth is ca. 15 % higher than europium sesquioxide [30], [64], [65].

[Top of Page]

References

1. R. Kieffer, F. Benesovsky: *Hartstoffe*, Springer Verlag, Wien-New York 1963.
2. A. Joly, *C.R. Hebd. Séances Acad. Sci.* **97** (1883) 456 – 458.
3. R. R. Ridgway, *Trans. Electrochem. Soc.* **66** (1934) 117 – 133. Norton Co., US 1897214, 1933.
4. A. Lipp, *Tech. Rundsch.* **57** (1965) no. 14, 5; no. 28, 19; no. 33, 5; 58 (1966) no. 7, 3.
5. F. Thevenot, *J. Eur. Ceram. Soc.* **6** (1990) 205. Links
6. M. Beauvy, F. Thevenot, *Ind. Ceram.* **734** (1979) 811.
7. R. Telle, L. Sigl, K. Takagi: "Boride Based Hard Materials" in R. Riedl (ed.): *Handbook of Ceramic Hard Materials*, Vol 2, Wiley-VCH, Weinheim, Germany 2000, pp. 802–945.
8. J. Economy, V. I. Matkovich, R. F. Giese, *Z. Kristallogr.* **122** (1965) 248. Links
9. A. Lipp, M. Roeder, *Z. Anorg. Allg. Chem.* **343** (1966) 1. Links
10. P. Ettmayer, H. Ch. Horn, K. A. Schwetz, *Mikrochim. Acta, Suppl.* **4** (1970) 87 – 95.
11. T.L. Asalage, R.G. Tissot, *J. Am. Ceram. Soc.* **75** (1992) 2207–2212
12. A. Lipp, *Ber. Dtsch. Keram. Ges.* **43** (1966) 60 – 62.
13. G. Will, K. H. Kossobutzki, *J. Less-Common Met.* **47** (1976) 43 – 48.
14. H. Werheit, K. De Groot, W. Malkemper, *J. Less-Common Met.* **82** (1981) 153 – 162.
15. K. A. Schwetz, P. Karduck, *J. Less-Common Met.* **175** (1991) 1 – 11.

16. U. Kuhlmann, H. Werheit, K. A. Schwetz, *J. Alloys Compd.* **189** (1992) 249.
17. M. Bouchacourt, F. Thevenot, *J. Less-Common Met.* **82** (1981) 227 – 235.
18. G. V. Samsonov, N. N. Zhuravlev, I. G. Annuel, *Fiz. Met. Metalloved.* **3** (1956) 309.
19. N. N. Zhuravlev, G. Makarenko, G. V. Samsonov, *Izv. Akad. Nauk. SSSR Otd. Tekh. Nauk Metall. Topl.* **1** (1961) 133 – 141.
20. R. T. Dolloff, Air Force Report, WADD, Tech. Rep. **60-143** (1960) 22.
21. R. P. Elliott, U.S. AEC ARF-2200-12 (1961).
22. R. Kieffer, E. Gugel, G. Leimer, P. Ettmayer, *Ber. Dtsch. Keram. Ges.* **48** (1971) 385.
23. G. Will, K. H. Kossobutzki, *J. Less-Common Met.* **47** (1976) 33.
24. E. Amberger, M. Druminski, W. Dietze, *Angew. Chem.* **79** (1967) 278; Links *Angew. Chem. Int. Ed. Engl.* **6** (1967) 267.
25. K. Ploog, *J. Cryst. Growth* **24/25** (1974) 197 – 204.
26. M. Beauvy, *J. Less-Common Met.* **90** (1983) 169 – 175.
27. H. J. Seifert, F. Aldinger: "Phase Equilibria in the Si-B-C-N-System" in M. Jansen (ed.): *High Performance Non-Oxide Ceramics I*, Springer Verlag, Berlin, Heidelberg 2002, pp. 2–58.
28. A. Lipp, K. Schwetz, *Ber. Dtsch. Keram. Ges.* **52** (1975) 335 – 338.
29. K. A. Schwetz, W. Grellner, *J. Less-Common Met.* **82** (1981) 37 – 47.
30. R. A. Murgatroyd, B. T. Kelly, *At. Energy Rev.* **15** (1977) 3 – 74.
31. V. I. Matkovich: *Boron and Refractory Borides*, Springer Verlag, Berlin-Heidelberg-New York 1977.
32. G. N. Makarenko, in [31] p. 319.
33. A. Lipp, M. Roeder, *Z. Anorg. Allgem. Chem.* **344** (1966) 225.
34. *Handbook of Boron Carbide*, Norton Company, Worcester, Mass., 1955.
35. ASTM Standard C 791, *Annu. Book ASTM Stand*. 1981, part 45.
36. F. Predel, R. Telle, V. Carle, V. Schäfer, V. Täffner, G. Petzow: "Ceramography of High Strength Ceramics, Part VII, Boron Carbide", *Prakt. Metallographie* **31** (1994) 218–233.
37. G. Dufek, W. Wruss, A. Vendl, R. Kieffer, *Planseeber. Pulvermetall.* **24** (1976) 280.
38. A.W. Weimer, R.P. Roach, C.N. Haney, W.G. Moore, W. Raffaniello, *AIChE J.* **37** (1991) 759–768.
39. A.W. Weimer: *Carbide, Nitride and Boride Materials, Synthesis and Processing*, Chapman & Hall, 1992, pp. 3–113.
40. Quartz et Silice, GB 687946, 1950 (E. Gray); EP 11524, 1980.
41. L. Vandenbulcke, G. Vuillard, *J. Less-Common Met.* **82** (1981) 49 – 56.
42. A. Lipp, *Silikattechnik* **32** (1981) 373 – 375.
43. K. Adlassnig, *Planseeber. Pulvermetall.* **6** (1958) 92 – 103.
44. L. B. Ekbom, *Sci. Ceram.* **9** (1977) 183 – 189; **11** (1981) 237 – 243.
45. M. Bouchacourt, C. Brodhag, F. Thevenot, *Sci. Ceram.* **11** (1981) 231 – 236.
46. ESK GmbH, DE-OS 2751998, 1977; US 4195066, 1980.
47. United Kingdom Atomic Energy Authority, GB 2014193, 1978.
48. H. Suzuki, T. Hase, T. Maruyama, *Yogyo Kyokaishi* **87** (1979) 430 – 433.
49. L. S. Sigl, K. A. Schwetz, in G. Ziegler *et al.* (eds.): *EURO-CERAMICS II*, Vol **1**, Verlag DKG. e.V., 1993, 517.
50. K. A. Schwetz, L. S. Sigl, L. Pfau, *J. Solid State Chem.* **133** (1997) 68 – 76. Links
51. K. A. Schwetz, W. Grellner, A. Lipp, in "Science of Hard Materials, Conf. Proceedings", *Phys. Conf. Ser. No. 75,* Hilger Bristol, 1986, Chap. 5, p. 413.
52. K. A. Schwetz, K. Reinmuth, A. Lipp, *Sprechsaal* **116** (1983) no. 12, 1063 – 1070.
53. V. Kevorkijan, A. Bizjak, J. Vizintin, F. Thevenot in A. Bellosi (ed.): *FOURTH-EURO-CERAMICS*, Vol **4**, Faenza Editrice S.p.A, 1995, pp. 209 – 216.
54. L. S. Sigl, K. A. Schwetz, *Jpn. J. Appl. Phys. Series* **10** (1994) 224 – 225.
55. L. S. Sigl, H.- J. Kleebe, *J. Am. Ceram. Soc.* **78** (1995) 9, 2374 – 2380. Links
56. V. Skorokhod, V.D. Kristic, *J. Mat. Sci. Lett.* **19** (2002) 237–239.
57. C. H. Lee, C. H. Kim, *J. Mat. Sci.* **27** (1992) 6335–6340.
58. H.W. Kim, Y. H. Koh, H. E. Kim, *J. Am. Ceram. Soc.* **83** (2000) 2863–2865. Links
59. M. Marietta, US 5 330 942, 1993.
60. Harbin Songhuajiang Patent Agent Company, CN 1438201, 2003.
61. K. A. Schwetz, K. Reinmuth, A. Lipp, *Radex Rundsch.* 1981, no. 3, 568 – 585.
62. S. R. Skaggs: "A Brief History of Ceramic Armor Development", *Ceram. Sci. Eng. Proc.* **24** (2003) 337–349.
63. Technical Bulletin "Advanced Armor Products", Ceradyne Inc., Costa Mesa, CA, 2003, www.ceradyne.com.
64. K. Reinmuth, A. Lipp, H. Knoch, K. A. Schwetz, "Borhaltige, keramische Neutronenabsorberwerkstoffe," *J. Nucl. Mater.* **124** (1984) 175 – 184. Links
65. K. A. Schwetz, A. Lipp, *ATW Atomwirtsch. Atomtech.* **18** (1973) 531.
66. A. Lipp, Wacker-Zeitschrift *Werk + Wirken* **29** (1978) no. 9, 1.
67. A. Matuschka: *Borieren*, 1st ed., Delta Verlag, Vevey 1975. W. Fichtl, *HTM Härterei Tech. Mitt.* **33** (1978) 1, 13.
68. W. Fichtl, *Ber. Dtsch. Keram. Ges.* **52** (1975) 79.
69. F. Ye, M. Rigaud, *Proc. 5th UNITECR Conference 1997, New Orleans,* **Vol. II**, 807 – 815.
70. K. Hunold, *Int. Ceram.* **1** (1998) 41 – 43.
71. Alyn Corp., US 5 712 014, 1998.
72. D. Schuster, *Gorham Newsletter* (1995) no. 1, 1 – 3.
73. Alyn Corp., PCT US 97/09359, 1997; WO 98/00259, 1998.
74. M. Bouchacourt, F. Thevenot, *J. Mater. Sci.* **20** (1985) 1237 – 1247. Links
75. K. Hunold, "B_4C – C Thermoelements for High Temperatures," *Chem. Tech. (Heidelberg)* **14** (1985) no. 10.
76. A. Lipp, K. A. Schwetz, K. Hunold, *J. Eur. Ceram. Soc.* **5** (1989) 3 – 9. Links
77. E. Gugel: *Handbuch der Keramik*, Gruppe II K 2, Verlag Schmid GmbH, Freiburg 1975/1977, pp. 6 – 9.
78. T. A. Ingles, P. Popper in P. Popper (ed.): *Special Ceramics*, Heywood and Co., London 1960, pp. 144 – 167. *Gmelin*, Boron, system no. 13, main compounds 1 (1974) 1 – 87.
79. R. Haubner, M. Wilhelm, R. Weissenbacher, B. Lux: "Boron Nitrides – Properties, Sythesis and Applications" in M. Jansen (ed.): *High Performance Non-Oxide Ceramics II*, Springer Verlag, Berlin, Heidelberg 2002, pp. 1–45.
80. E. A. Smith, *Powder Metall.* **14** (1971) 110. Links

81. ESK GmbH, DE-OS 3201563, 1983; EP 84369, 1983.
82. A. Sawaoka, *Am. Ceram. Soc. Bull.* **62** (1983) 1379 – 1383. Links
83. W. H. Balmain, *J. Prakt. Chem.* **27** (1842) 422 – 430; **32** (1844) 494 – 495.
84. Carborundum Co., US 2808314, 1957.
85. J. Thomas, N. E. Weston, T. E. O'Connor, *J. Am. Chem. Soc.* **84** (1963) 4619 – 4622.
86. Moskovskij Chimiko-Technologitscheskij Institut Imeni D. I. Mendeleeva, DE-OS 2629960, 1978.
87. M. Basche, D. Schiff, *Mat. Des. Eng.* **59** (1964) 78 – 81.
88. R. Y. Lin, J. Economy, H. H. Murty, R. Ohnsorg, *Appl. Polym. Symp.* **29** (1976) 175 – 188.
89. S. Koide, K. Nakamura, K. Yoshimura, *Kenkyu Kiyo Nihon Daigaku Bunrigakubu Shizen Kagaku Kenkyusho* 1979, 14, 9 – 14; *Chem. Abstr.* **94** (1981) 19465 t.
90. R. Y. Lin, J. Economy, H. D. Batha, *Am. Ceram. Soc. Bull.* **55** (1976) 781. Links
91. R. H. Wentorf, Jr., *J. Chem. Phys.* **34** (1961) 809. Links
92. T. Endo, O. Fukunga, M. Iwata, *J. Mater. Sci.* **16** (1981) 2227 – 2232. Links
93. General Electric, US 3743489, 1973.
94. F. P. Bundy, R. H. Wentorf, *J. Chem. Phys.* **38** (1963) 1144. Links
95. T. Akashi, A. Sawaoka, S. Saito, *J. Am. Ceram. Soc.* **61** (1978) 245. Links
96. R. H. Wentorf, R. C. De Vries, F. P. Bundy, *Science (Washington, D.C.)* **208** (1980) 4446, 873 – 880.
97. H. Moissan: *Der elektrische Ofen*, Verlag M. Krayer, Berlin 1900.
98. L. S. Sigl, K. A. Schwetz, *Powder Metall Int.* **23** (1991) no. 4, 221 – 224.
99. K. Takagi, Y. Yamasaki, M. Komai, *J. Solid State Chem.* **133** (1997) no. 1, 243 – 248. Links
100. J. L. Hoard, R. E. Hughes in E. L. Mutterties (ed.): *The Chemistry of Boron and Its Compounds*, J. Wiley & Sons, New York 1967, pp. 22 – 154.
101. T. Lundström, *Ark. Kemi* **31** (1969) 227 – 266.
102. H. Nowotny in L. E. J. Roberts (ed.): *Solid State Chemistry*, vol. **10**, Butterworths, London 1972, pp. 151 – 188.
103. E. Rudy, *Tech. Rep. AFML-TR* **652, Part V.** (1969).
104. K. E. Spear: *Phase Diagrams*, vol. **IV**, Materials Science and Technology, Academic Press, New York 1976, pp. 91 – 159.
105. G. V. Samsonov, T. I. Serebryakova, V. A. Neronov, *Boridy (Borides)* Atomizdat Publ., Moscow, 1975.
106. R. Thompson in J. Brotherton, H. Steinberg (eds.): *Progress in Boron Chemistry*, vol. **2**, Pergamon Press, Oxford 1971, pp. 173 – 230.
107. F. Binder, *Radex Rundsch.* 1975, 4, 531 – 557; 1977, 1, 52 – 71.
108. H. Pastor, F. Thevenot, *Inf. Chim.* **178** (1978) 151 – 173.
109. R. Telle: "Boride and Carbide Ceramics", in M. V. Swain (ed.): *Materials Science and Technology: A Comprehensive Treatment*, Vol. **11**, VCH Verlagsgesellschaft, Weinheim, Germany 1994, pp. 175 – 265.
110. R. Kiessling, *Acta Chem. Scand.* **4** (1950) 209.
111. C. Mitterer *et al.*, *Vacuum* **46** (1995) no. 11, 1281 – 1294. Links K. A. Schwetz, Thesis, Technical University, Vienna 1971.
112. F. Thevenot, J. Cueilleron, *Analysis* **5** (1977) 105 – 121.
113. PPG, DE-OS 2523423, 1975.
114. PPG, DE-OS 2818418, 1978 (R. R. May, Jr.,
115. H. Pastor, in [31] pp. 457 – 493
116. ESK GmbH, DE-OS 3301841, 1984.
117. H. R. Baumgartner, R. A. Steiger, *J. Am. Ceram. Soc.* **67** (1984) no. 3, 207. Links
118. E. Fitzer, *Arch. Eisenhüttenwes.* **44** (1973) no. 9, 703 – 709.
119. R. D. Holliday, R. Mogstad, J. L. Henry, *Electrochem. Technol.* **1** (1963) 183 – 190.
120. S. K. Dutta, G. E. Gazza, *Bull. Am. Ceram. Soc.* **52** (1973) 552 – 554.
121. F. Rigby: "Development of Hot Pressing Techniques at Springfields Nuclear Laboratories," *Spec. Ceram.* **7** (1981) 249 – 258.
122. K. Hunold, *CFI Ceram. Forum Int.* **60** (1983) 182 – 189.
123. B. Y. Yuriditsky, *Refract. Mater. Hard. Mater.* **3** (1990) 32.
124. Th. Jüngling, R. Oberacker, F. Thümmler, L. S. Sigl, K. A. Schwetz, *Powder Metall Int.* **23** (1991) 5, 296 – 300.
125. I. Smid, E. Kny, *Refract Hard Mater* **8** (1988) 135 – 138.
126. V. Ghetta, N. Gayrand, N. Eustathopoulos, *Diffus. Defect Data Pt. B:* **25/26** (1992) 105.
127. L. Ottavi, C. Saint-Jours, N. Valignant, C. H. Allibert, *Z. Metallkunde* **83** (1992) 80 – 83.
128. Th. Jüngling, L. S. Sigl, R. Oberacker, F. Thümmler, K. A. Schwetz, *Int. J. of Ref. Hard Met.* **12** (1993 – 94) 71 – 88.
129. L. S. Sigl, K. A. Schwetz, U. Dworak, *Int. J. Refract. Hard Met.* **12** (1993 – 1994) 95 – 99.
130. K. Takagi, M. Komai, T. Watanabe, Y. Kondo, *Int. Powder Metallurgy Conference* (1990) no. 1, 374 – 384.
131. K. Takagi, *Jpn. Appl. Phys. Ser* **10** (1994) 200 – 203.
132. W. Grellner, K. A. Schwetz, A. Lipp, *Radex Rundsch.* 1983, 1/2, 146 – 151.
133. Aluminum Corporation of America, US 4071420, 1978.
134. C. E. Ransley, *J. Met.* **14** (1963) 129.
135. ESK GmbH, DE 2240655, 1972.
136. K. S. S. Murthy, *Br. Foundryman* **67** (1974) 335.
137. ESK GmbH, US 4118256, 1976.
138. K. G. Schmitt-Thomas, H. Meisel, S. Mirdamadi-Therani, *Metall (Berlin)* **32** (1978) 1103.
139. N. Kaji, H. Shikano, I. Tanaka, *Taikabutsu Overseas* **14** (1994) no. 2, 39 – 43.
140. O. Knotek, E. Lugscheider, H. Eschnauer: *Hartlegierungen zum Verschleißschutz*, Verlag Stahleisen, Düsseldorf 1975.
141. R. Thompson, in R. Freer (ed.): *The Physics and Chemistry of Carbides, Nitrides and Borides*, vol. **185**, NATO-ASI-Series, 1990 pp. 113 – 120.
142. M. Sagawa, S. Fugimura, N. Togawa, H. Yamamoto, Y. Matsuara, *J. Appl. Phys.* **55** (1984) 2083. Links
143. G. V. Samsonov, Yu. B. Paderno, V. S. Fomenko, *Sov. Powder Metall. Met. Ceram. (Engl. Transl.)* **6** (1963) 24.
144. M. Futamoto, M. Nakazawa, U. Kawabe, *Surf. Sci.* **100** (1980) 470 – 480. Links

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Boron Compounds

Robert J. Brotherton, U.S. Borax Research Corp., Anaheim, California 92801, United States (Chaps. 1 – 4, 6.4)

C. Joseph Weber, U.S. Borax Research Corp., Anaheim, California 92801, United States (Chaps. 1 – 4, 6.4)

Clarence R. Guibert, Callery Chemical Co., Callery, Pennsylvania 16024, United States (Chaps. 5, 6 – 6.3, 7)

John L. Little, Callery Chemical Co., Callery, Pennsylvania 16024, United States (Chaps. 5, 6 – 6.3, 7)

1. Boron Halides . . . 1
1.1. Physical Properties . . . 1
1.2. Chemical Properties . . . 1
1.2.1. Exchange Reactions . . . 2
1.2.2. Halide Displacement Reactions . . . 2
1.2.3. Boron Trihalides as Lewis Acids . . . 3
1.2.4. Reduction Reactions . . . 3
1.3. Production . . . 3
1.4. Quality Specifications, Chemical Analysis . . . 4
1.5. Sources, Handling, and Transportation . . . 4
1.6. Uses of Boron Trihalides . . . 4
1.7. Pollution Control, Toxicology, and Occupational Health . . . 5
2. Fluoroboric Acid and Fluoroborates 5
2.1. Physical and Chemical Properties . 5
2.2. Production . . . 6
2.3. Quality Specifications, Chemical Analysis . . . 6
2.4. Sources, Handling, and Transportation . . . 6
2.5. Uses . . . 6
2.6. Pollution Control, Toxicology . . . 7
3. Boron Subhalides . . . 7
4. Boron Sulfide . . . 8
4.1. Physical and Chemical Properties . 8
4.2. Production . . . 8
4.3. Uses . . . 8
5. Boranes . . . 8
5.1. Diborane and Higher Boranes . . . 8
5.1.1. Physical and Chemical Properties . . . 9
5.1.2. Production . . . 10
5.1.3. Uses . . . 10
5.2. Amine Boranes . . . 11
5.3. Borohydrides . . . 11
5.4. Carbaboranes . . . 12
6. Organoboranes . . . 14
6.1. Alkyl- and Arylboranes . . . 14
6.2. Borinic and Boronic Acids . . . 15
6.3. Borazines . . . 16
6.4. Boric Acid Esters . . . 16
7. Toxicology and Occupational Health 18
8. References . . . 19

1. Boron Halides

1.1. Physical Properties

Boron forms trihalides, sometimes called trihaloboranes, with fluorine, chlorine, bromine, and iodine. The trifluoride [*7637-07-2*], BF_3, M_r 67.80, and trichloride [*10294-34-5*], BCl_3, M_r 117.19, are gases at room temperature. The tribromide [*10294-33-4*], BBr_3, M_r 250.54, is a liquid, and the triiodide [*13517-10-7*], BI_3, M_r 391.52, is a solid. All four trihalides are colorless, but the tribromide and triiodide discolor on exposure to light. A summary of some physical properties of these compounds is given in Table 1 (see next page).

All four compounds have trigonal symmetry with X−B−X angles of 120°. A crystal structure has been reported for a metastable phase of BF_3 at 142 K in which boron atoms are trigonal prismatic with $B-X_{eq}$ bond lengths of 0.126 – 0.131 nm and $B-X_{ax}$ bond lengths of 0.268 – 0.271 nm [10].

1.2. Chemical Properties

Trivalent boron halides are coordinatively unsaturated and their chemistry is dominated by the unfilled boron p_z orbital. The compounds are Lewis acids and readily form adducts with a variety of electron donors. Overlap of the unfilled boron p_z orbital with nonbonding filled orbitals

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_309

Table 1. Physical properties of boron halides [5], [6]

	BF_3	BCl_3	BBr_3	BI_3
mp, °C	− 127.1	− 107	− 46	49.9
bp, °C	− 99.9	12.5	91.3	210
Density, g/mL	1.57 ∗	1.359	2.618	–
H° fusion, kJ/mol (at 298 K)	− 1136 [7]	− 403 [8]	− 240 [8]	− 47 [8]
ΔS°, J mol^{-1}K^{-1} (at 298 K)	254 [7]	290	324	–
$c_{p\,(v)}$, J mol^{-1}K^{-1} (at 298 K)	50.5	62.6	68.0	–
Electron affinity, J × 10^{-19}	< 0 [9]	0.53 [9]	1.31 [9]	1.81 [9]
B−X Bond energy, kJ/mol	154.3	106.1	88.0	63.7
B−X Bond length, nm	0.129	0.175	0.187	0.210

∗ At −100 °C.

on the halides decreases the electron acceptor ability of the molecule but increases the strength of the B−X bonds. The B−F bond in BF_3 is the strongest single bond known. The boron – halide p_z overlap decreases with increasing size of the halide. Thus, the acceptor strength increases as $BF_3 < BCl_3 < BBr_3 < BI_3$. However, the strength of the B−X bond decreases from BF_3 to BI_3. Boron trifluoride chemistry is dominated by Lewis acid – base reactions, whereas BBr_3 and BI_3 adducts are relatively unstable and halide displacement reactions predominate.

1.2.1. Exchange Reactions

All of the boron halides undergo halide exchange reactions, but of the mixed halide species only BBr_2I and $BBrI_2$ have been isolated. Exchange reactions are also known with trialkyl-, triaryl-, trialkoxy-, and triaryloxyboranes:

$$BX_3 + 2\,BR_3 \longrightarrow 3\,BXR_2$$

Exchange reactions between BF_3 and metal halides may be used to prepare other boron trihalides:

$$3\,KCl + 4\,BF_3 \longrightarrow 3\,KBF_4 + BCl_3$$
$$AlBr_3 + BF_3 \longrightarrow BBr_3 + AlF_3$$

1.2.2. Halide Displacement Reactions

Boron halides, with the exception of boron trifluoride, react readily with compounds that have active hydrogen atoms, including water, alcohols, thiols, amines, phosphines, and arsines. Because of the strength of the boron – fluorine bond, halide displacement reactions occur to a lesser extent with BF_3 than with other boron halides. Boron trifluoride reacts with water to give a number of partially hydrolyzed fluoroborate species including BF_3OH^-, $BF_2(OH)_2^-$, and $BF(OH)_3^-$, in addition to boric acid and the fluoroborate anion, BF_4^-. Two hydrated species, $BF_3 \cdot H_2O$ and $BF_3 \cdot (H_2O)_2$, can be isolated at low temperature. The dihydrate is formulated as $BF_3OH_2 \cdot H_2O$ [11].

In contrast to the slow substitution reactions of BF_3, other boron trihalides react vigorously with water:

$$BX_3 + 3\,H_2O \longrightarrow B(OH)_3 + 3\,HX$$

The reaction of BBr_3 or BI_3 with water may be explosive.

Most halide displacement reactions with BF_3 require either elevated temperature or an active nucleophile, such as a Grignard reagent:

$$2\,BF_3 + 6\,RMgBr \longrightarrow 2\,BR_3 + 3\,MgBr_2 + 3\,MgF_2$$

Halide displacement reactions between other boron halides and compounds with active hydrogen atoms are quite facile. The first step is often the formation of an adduct, which may be isolated at low temperature, followed by the elimination of the hydrogen halide:

$$BX_3 + 3\,ROH \longrightarrow B(OR)_3 + 3\,HX$$
$$CH_3SH + BCl_3 \xrightarrow{-20\,^\circ C} BCl_3 \cdot CH_3SH$$
$$\xrightarrow[-20\,^\circ C]{\text{above}} CH_3SBCl_2 + HCl$$

In general the reactivity of boron trihalides toward compounds with active hydrogen atoms

increases with decreasing B–X bond energy, in the order $BF_3 < BCl_3 < BBr_3$. Side reactions leading to alkyl halides and boric acid rather than simple halide displacement are more common with the triiodide and sterically hindered alcohols.

1.2.3. Boron Trihalides as Lewis Acids

Boron trifluoride forms adducts with a large number of electron donors including ethers, alcohols, ketones, amines, phosphines, arsines, thiols, and selenides. Those complexes with large electron donor atoms or sterically hindered bases tend to be less stable. Boron trichloride also forms many stable complexes with electron donors, but these are more prone to halide displacement reactions than are the BF_3 adducts. Few stable adducts are known with BBr_3 or BI_3.

1.2.4. Reduction Reactions

Reaction of boron halides with hydrogen, hydrides, or alkali metals often yields boron subhalides or elemental boron. If hydrogen or a hydride is used, boranes may result, e.g., in the following convenient preparation of diborane:

$$6\,LiH + 8\,BF_3 \cdot O(C_2H_5)_2 \longrightarrow B_2H_6 + 6\,LiBF_4 + 8\,(C_2H_5)_2O$$

1.3. Production

Boron Trifluoride. Boron trifluoride was first prepared by the reaction of boric acid and fluorspar (CaF_2) at red heat [12]. Commercially, BF_3 has been produced by the stepwise addition of hydrofluoric acid and sulfuric acid to anhydrous borax (sodium tetraborate):

$$Na_2B_4O_7 + 12\,HF \longrightarrow Na_2O \cdot (BF_3)_4 + 6\,H_2O$$

$$Na_2O \cdot (BF_3)_4 + 2\,H_2SO_4 \longrightarrow 4\,BF_3 + 2\,NaHSO_4 + H_2O$$

Another method involves addition of sulfuric acid and fluorspar and a source of boron, such as anhydrous borax or boric oxide (B_2O_3).

A process was developed in which BF_3 is obtained by the reaction of HF and boric acid, using oleum as a dehydrating agent [13].

$$H_3BO_3 + 3\,HF \longrightarrow BF_3 + 3\,H_2O$$

Boron trifluoride is made commercially in the United States by a proprietary process from boric acid and hydrofluoric acid.

A variety of synthetic routes are available for the production of small quantities of BF_3, including thermal decomposition of diazonium fluoroborate. The diethyl etherate adduct, $BF_3 \cdot O(C_2H_5)_2$ [*109-63-7*], produced by bubbling BF_3 into diethyl ether, provides a convenient source of BF_3 for many synthetic applications.

Boron Trichloride. This compound can be prepared by the reaction of boric oxide, a reducing agent, and chlorine in a melt at high temperatures, for example:

$$B_2O_3 + 3\,C + 3\,Cl_2 \xrightarrow[\Delta]{} 3\,CO + 2\,BCl_3$$

The product BCl_3 is recovered by fractional condensation. Sodium tetraborate $Na_2B_4O_7$ can be used in place of B_2O_3. Molten sodium chloride, formed as a byproduct, settles to the bottom of the reactor [14]. A molten combination of aluminum chloride and an alkali-metal chloride has also been used as a medium for reductive chlorination of borates [15].

Free-radical chlorination of borate esters also yields BCl_3:

$$B(OCH_3)_3 + 9\,Cl_2 \longrightarrow BCl_3 + 3\,COCl_2 + 9\,HCl$$

A modification of this process, in which a free-radical initiator is used, has been reported [16], [17]. A low reaction temperature, 40–90 °C, and the use of an initiator decrease the consumption of Cl_2 and prevent the formation of phosgene, $COCl_2$.

Commercially, boron trichloride is prepared by the reaction of boron carbide with chlorine in a borate melt. This reaction can also be done with dry B_4C in a fluidized-bed reactor using a transition-metal halide as a catalyst [18].

Contamination of BCl_3 with $COCl_2$ from the chlorination of carbon oxides is a serious problem in BCl_3 manufacturing. Because of similar vapor pressure–temperature curves, BCl_3 and $COCl_2$ cannot be separated by fractional distillation. Thermal or photochemical conversion of

$COCl_2$ to CO and Cl_2 followed by fractional distillation has been used to purify BCl_3.

Boron trichloride can be prepared in the laboratory by the reaction of boron trifluoride and aluminum chloride.

Boron Tribromide. This compound can be prepared by the reaction of boric oxide, carbon, and bromine, or from boron carbide and bromine analogous to the chlorine reactions.

Small quantities of BBr_3 can be made by the reaction of boron trifluoride and aluminum tribromide. A five-fold excess of $AlBr_3$ increases the yield of BBr_3.

Boron Triiodide. Boron triiodide can be prepared by the reaction of a borohydride and iodine:

$$3\,MBH_4 + 8\,I_2 \longrightarrow 3\,BI_3 + 3\,MI + 4\,HI + 4\,H_2$$

where M is lithium, sodium, or potassium.

The reaction between sodium borohydride and iodine in dimethylformamide (DMF) results in $BI_3 \cdot 4\,DMF$ [19]. Boron triiodide can also be made by exchange reactions between AlI_3 and BF_3, or between LiI and $LiBH_4$. Contrary to an earlier report, boron triiodide can be prepared from the elements by the reaction of amorphous boron with iodine at 870 °C [20].

1.4. Quality Specifications, Chemical Analysis

Boron trifluoride is available in quantity as a 99 % pure gas. Boron trichloride is typically sold as a 99.9 % pure liquid. Boron tribromide is offered in grades up to 99.9999 % for electronic applications.

Common impurities in BF_3 include air, silicon tetrafluoride, and sulfur dioxide. The major impurities in boron trichloride are phosgene and chlorine. Boron tribromide typically contains traces of bromine.

The boron halides may be analyzed by hydrolysis followed by analysis of B_2O_3 and the halide, or by base titration. The concentration of phosgene ($COCl_2$) in BCl_3 may be determined by infrared spectroscopy. The trifluoride, trichloride, and tribromide may be purified by fractional distillation.

1.5. Sources, Handling, and Transportation

Allied Corp. is the major producer of boron trifluoride. Boron trifluoride complexes with monoethylamine, phenol, and diethyl ether are also available in large quantities from Allied. Complexes of BF_3 with water, dimethylamine, monoethylamine, phosphoric acid, and piperidine are available from Harshaw/Filtrol. Kerr-McGee Corp. is the major producer of BCl_3 and BBr_3. Boron trifluoride and trichloride are available in steel cylinders. Boron tribromide is available in glass bottles. Boron tribromide for use in electronics is supplied by J. C. Schumacher Co. in specially sealed glass vessels equipped with teflon valves. Boron triiodide is available in small quantities from the Ventron Division of Morton-Thiokol Inc.

Boron trifluoride and boron trichloride sold for \$ 6.83/kg (\$ 3.10/lb) and \$ 8.38/kg (\$ 3.80/lb) in bulk in 1984. The 1982 price for BBr_3 was \$ 57/kg (\$ 26/lb) in large quantities from Kerr-McGee.

Boron trifluoride is the largest volume boron halide. The total annual production of BF_3 and its adducts is about 2300 – 4500 t/a. The annual production of BCl_3 is about 225 – 450 t/a. The annual production rate of BBr_3 is significantly less than that of BCl_3.

The boron trihalides are nonflammable. But because they hydrolyze easily, care should be taken to avoid exposure to moisture. The DOT classification for BCl_3 and BBr_3 is corrosive liquid; BF_3 is classified as a nonflammable gas. Equipment for conveying the compounds should be flushed with dry air prior to use. Stainless steel lines and valves are suitable for use with all dry boron halides.

1.6. Uses of Boron Trihalides

Boron Trifluoride. Boron trifluoride is the most widely used boron halide. Most applications take advantage of its strong Lewis acidity. The most important uses are in organic syntheses. Boron trifluoride is commonly used as a catalyst for Friedel-Crafts alkylation reactions. It also is used to catalyze the cleavage of ethers to alcohols, to catalyze esterification reactions,

and in the nitration and sulfonation of aromatic compounds. Many olefin polymerization reactions use BF_3 as an initiator, in conjunction with a proton donor, such as water. Also BF_3 is used to catalyze the isomerization of alkenes and alkanes and in petroleum cracking and desulfurization. Amine complexes of BF_3 are used as epoxy curing agents.

Boron Trichloride. One of the most important uses of BCl_3 is in the preparation of boron fibers (→ Fibers, 5. Synthetic Inorganic). Typically an electrically heated tungsten filament is passed through a chamber containing BCl_3 and hydrogen. The BCl_3 is reduced, and boron is deposited on the filament, producing a stiff, strong boron fiber.

Boron trichloride, like the trifluoride, has been used as a Lewis acid catalyst in organic synthesis in the polymerization of olefins and phosphazines, as well as in catalysis of other organic reactions. Boron trichloride is also used in plasma etching of aluminum and silicon, in semiconductor manufacturing, and as a source of boron for chemical vapor deposition. Steel is boronized by contacting it with a reactive mixture of hydrogen, hydrocarbons, and BCl_3 at high temperatures.

Boron Tribromide. The primary use of boron tribromide is as an initiator for the polymerization of olefins and as a catalyst in other organic reactions. It is also used in the electronics industry as a source of bromine for ion implantation in semiconductors, and for plasma etching in semiconductor device manufacturing.

Boron Triiodide. There are no important commercial uses for boron triiodide.

1.7. Pollution Control, Toxicology, and Occupational Health

Boron halides are not major pollution sources. Fluoride, from the hydrolysis of BF_3, can be precipitated from waste water as CaF_2 by the addition of a calcium salt [21–23]. Aqueous boron (boric acid or borates) can be precipitated by the addition of calcium salts at high pH.

Boron halides are quite toxic and contact with skin or eyes should be avoided. A TLV ceiling limit of 1.0 ppm, 3 mg/m^3 (TWA), has been set for BF_3 by the ACGIH. The MAK values are the same. For BBr_3, the adopted TLVs are 1 ppm, 10 mg/m^3 (TWA); 3 ppm, 30 mg/m^3 (STEL). The LCs in air for BF_3 and BCl_3 in male rats (1 h exposure) are 387 and 2.54 ppm, respectively [24]. These compounds are not known to be carcinogenic.

2. Fluoroboric Acid and Fluoroborates

2.1. Physical and Chemical Properties

In the older literature and in some commercial usage, HBF_4 and its salts are referred to as fluoboric acid and fluoborates. Nomenclature based on fluoroboric and fluoroborates has now become generally accepted in the literature and is used throughout this chapter [29].

The fluoroborate anion, BF_4^-, is essentially the adduct of BF_3 and F^-. The ion is tetrahedral, with B–F bond lengths of about 0.140–0.145 nm compared to a B–F bond length of 0.129 nm in BF_3.

The fluoroborate anion hydrolyzes in water in a stepwise manner to give first hydroxyfluoroborates, then boric acid:

$$BF_4^- + H_2O \longrightarrow (HO)BF_3^- + H^+ + F^-$$
$$(HO)BF_3^- + H_2O \longrightarrow (HO)_2BF_2^- + H^+ + F^-$$
$$(HO)_2BF_2^- + H_2O \longrightarrow (HO)_3BF^- + H^+ + F^-$$
$$(HO)_3BF^- \longrightarrow B(OH)_3 + F^-$$

Concentrated solutions of alkali-metal fluoroborates are stable toward hydrolysis; the anion is more easily hydrolyzed in solutions of transition-metal fluoroborates. Concentrated aqueous solutions of fluoroboric acid, tetrafluoroboric acid, HBF_4 [*16872-11-0*], can be made, but HBF_4 has not been isolated. Solvated species have been crystallized, including $H_3O \cdot BF_4$, $H_5O_2 \cdot BF_4$, and $[H(CH_3OH)_2] \cdot BF_4$ [30]. The term fluoroboric acid, as used commercially, always refers to an aqueous solution of HBF_4.

Many transition-metal fluoroborates are quite soluble in water. Alkali-metal salts are less soluble. Ammonium, sodium, and potassium fluoroborate crystallize as anhydrous salts.

Transition-metal salts all crystallize with various degrees of hydration. Many metal fluoroborates are also soluble in methanol and ethanol. Silver and copper(II) fluoroborate are soluble in diethyl ether and aromatic solvents.

Solid fluoroborate salts decompose upon heating, evolving BF_3. The heat of dissociation increases with cation size for the alkali-metal fluoroborates [31]. The reaction has been used as a convenient source of BF_3 for laboratory applications:

$$NaBF_4 \longrightarrow NaF + BF_3$$

2.2. Production

Fluoroboric acid is usually made by the reaction of hydrofluoric acid and boric acid in water:

$$4\,HF + B(OH)_3 \longrightarrow HBF_4 + 3\,H_2O$$

The reaction is exothermic.

Most of the commonly used fluoroborate salts can be prepared by reacting metal oxides, hydroxides, or carbonates with fluoroboric acid. Ammonium fluoroborate can be made from ammonia and fluoroboric acid. Fluoroborate salts can also be prepared by the reaction of BF_3 with a metal fluoride in a nonaqueous, inert solvent, such as HF, BrF_3, or SO_2.

2.3. Quality Specifications, Chemical Analysis

Fluoroboric acid is available from Harshaw/Filtrol in two grades, Electropure and Fluopure, a lower purity material. Specifications for Electropure fluoroboric acid are given in Table 2.

Table 2. Specifications for Harshaw/Filtrol Electropure fluoroboric acid

Component	Specification	Typical average
HBF_4	48 – 50 %	48.5 %
H_3BO_3	0.5 – 1.5 %	0.75 %
Fe	10 ppm max.	4 ppm
Cu	1 ppm max.	< 1 ppm
Pb	1 ppm max.	< 1 ppm
Zn	1 ppm max.	< 1 ppm
Cl	3 ppm max.	< 3 ppm
SO_4	50 ppm max.	< 50 ppm
Ni	2 ppm max.	< 1 ppm

Solid alkali-metal fluoroborates may contain some hydroxyfluoroborates and boric acid as contaminants. Metal fluoroborate solutions may contain free boric acid or fluoroboric acid as well as trace amounts of metallic impurities.

Fluoroboric acid can be analyzed by hydrolysis, followed by precipitation of fluoride with calcium and titration for the borate. Metal fluoroborates have been analyzed by electrolytic reduction followed by mannitol titration for boric acid [32]. A BF_4^- ion selective electrode is available from Orion. The tetrafluoroborate ion can also be analyzed by ion chromatography [33], [34].

2.4. Sources, Handling, and Transportation

Fluoroboric acid is available from Allied Corp., Harshaw/Filtrol, C.P. Chemicals, Fidelity Chemicals, Harstan, and Ozark-Mahoning. A variety of fluoroborate salts is produced by C.P. Chemicals, Harstan, Allied, Harshaw/Filtrol, and Fidelity Chemicals. Transition-metal fluoroborates are commonly sold as aqueous solutions. They may contain a slight excess of boric acid, added as a stabilizer. Information on commercial fluoroborates is given in Table 3.

Fluoroboric acid and solutions of fluoroborates are corrosive and are shipped in plastic drums or polyethylene pails. Glass vessels should not be used for containing these solutions.

Solutions of fluoroboric acid and fluoroborates are classified as corrosive materials for shipping.

2.5. Uses

The primary use of fluoroboric acid is in the preparation of other fluoroborate salts. Fluoroboric acid is also used in electroplating and in dipping solutions for surface treating of aluminum. It has been recommended as a pickling agent for hot rolled steel [35], and has been patented for use as an etchant for silicon and glass in the electronics industry (→ Electrochemical and Chemical Deposition, Chap. 4.) [36], [37].

Table 3. Commercial fluoroborates

Compound	CAS registry no.	Formula	Commercial form	Price (1984), $/kg
Fluoroboric acid	[*16872-11-0*]	HBF_4	49 % solution	2.01
Alkali-metal salts				
Ammonium	[*13826-83-0*]	NH_4BF_4	crystalline solid	3.73
Potassium	[*14075-53-7*]	KBF_4	crystalline solid	3.13
Sodium	[*13755-29-8*]	$NaBF_4$	solid	3.09
Transition-metal salts				
Antimony(III)	[*14486-20-5*]	$Sb(BF_4)_3$	solution	6.66
Cadmium(II)	>[*144-86-19-2*]	$Cd(BF_4)_2$	50 % solution	5.00
Cobalt(II)	[*26490-63-1*]	$Co(BF_4)_2$	solid	–
Copper(II)	[*38465-60-0*]	$Cu(BF_4)_2$	45 % solution	2.31
Indium(III)	[*27765-48-6*]	$In(BF_4)_3$	solution	–
Iron(II)	[*15283-51-9*]	$Fe(BF_4)_2$	41 % solution	1.06
Lead(II)	[*13814-96-5*]	$Pb(BF_4)_2$	51 % solution	1.87
Nickel(II)	[*14708-14-6*]	$Ni(BF_4)_2$	45 % solution	3.55
Silver(I)	[*14104-20-2*]	$AgBF_4$	powder	–
Tin(II)	[*13814-97-6*]	$Sn(BF_4)_2$	50 % solution	7.69
Zinc(II)	[*13826-88-5*]	$Zn(BF_4)_2$	40 % solution	–

Molten alkali-metal and ammonium fluoroborates are good solvents for metal oxides and are commonly used in fluxes for soldering and brazing.

A mixture of LiF and $NaBF_4$ has been studied as a nuclear reactor coolant [38] and as a solvent for fissionable materials. Mixtures of NH_4BF_4 with nitro compounds are useful as explosives [39].

Alkali metal fluoroborates and fluoroboric acid have also been used as catalysts in organic synthesis and in polymerization reactions.

Lithium fluoroborate is used as an electrolyte in lithium – sulfur batteries.

The fluoroborate anion is similar to perchlorate, ClO_4^-, in its size, shape, and lack of ability to form coordinate bonds with transition metals. Fluoroborates are often used in transition metal chemistry where a noncoordinating anion is required.

The primary use of transition metal fluoroborates is in electroplating. The fluoroborate ion is electrochemically inert and solutions of fluoroborates are highly conductive. Metals plated from fluoroborate solutions include cadmium, copper, indium, iron, lead, nickel, silver, tin, and zinc.

2.6. Pollution Control, Toxicology

Aqueous solutions of fluoroborates can be hydrolyzed and treated with calcium compounds to precipitate calcium fluoride for pollution control [40–42]. Reverse osmosis has also been used to purify waste streams containing potassium fluoroborate [43].

Protective equipment used with hydrofluoric acid should also be used when handling solutions of fluoroboric acid or fluoroborates.

Little information is available in the literature on the toxicity of fluoroborates. Fluoroborate given intravenously to rats depressed levels of L-thyroxine [44], and ammonium fluoroborate was shown to be nontoxic to fresh water minnows at concentrations less than 600 mg/L [45].

3. Boron Subhalides

All of the compounds, B_2X_4, where X is F [*13965-73-6*], Cl [*13701-67-2*], Br [*14355-29-4*], or I [*13703-80-5*], are known. The tetrafluoride is the most stable, but all four compounds decompose at room temperature to BX_3 and other boron halides. The compounds are Lewis diacids and form 1 : 2 complexes with amines and other Lewis bases. Boron tetrachloride, which is the most commonly used of these tetrahalides, can be prepared by reduction of BCl_3 in an electrical discharge apparatus. The most important use of B_2X_4 compounds is in diboration reactions, i.e., the addition of B_2X_4 to unsaturated organic molecules. These reactions are analogous to hydroboration reactions with diborane. Some of the decomposition products of B_2Cl_4 have been characterized. They include B_4Cl_4, B_8Cl_8, and B_9Cl_9.

4. Boron Sulfide

4.1. Physical and Chemical Properties

Boron sulfide [*12007-33-9*], B_2S_3, M_r 117.81, is a colorless compound normally obtained as an amorphous glass. It is extremely difficult to purify and often contains traces of silicon (from glass vessels), oxygen, or sulfur, which may impart a yellow color to the material [46]. The amorphous form does not have a well-defined melting point but begins to sublime at about 300 °C. Some of the physical properties are given in Table 4.

Table 4. Physical properties of B_2S_3

mp (crystalline form), °C	563 [47]
Flow point (amorphous form), °C	310 [46]
bp, °C	≈ 300, sublimes [46]
Density, ϱ, g/cm^3	1.92 [48]
B−S Bond energy, kJ/mol	360 [49]
Heat of fusion, kJ/mol	150 [46]
Heat of vaporization, kJ/mol	252 [50]
Average B−S bond length, nm	0.1808 [48]

An X-ray analysis of crystalline B_2S_3 indicates it is composed of $(BS)_3$ and $(BS)_2$ rings linked through bridging sulfur atoms in a two-dimensional lattice [48].

Boron sulfide reacts with a variety of protonic reagents [51]:

$$B_2S_3 + 6\,H_2O \longrightarrow 2\,B(OH)_3 + 3\,H_2S$$
$$B_2S_3 + 6\,ROH \longrightarrow 2\,B(OR)_3 + 3\,H_2S$$
$$B_2S_3 + 6\,R_2NH \longrightarrow 2\,B(NR_2)_3 + 3\,H_2S$$

The reaction with water is quite violent.

Boron sulfide inflames in chlorine, and is oxidized at high temperature by carbon dioxide or by oxygen [51]. At high temperature it also readily attacks quartz and silicate glass [46]. Boron sulfide reacts with sulfur to give B_8S_{16}, a planar molecule [52]. A variety of other higher molecular mass boron sulfides is also known [53].

Boron sulfide can also act as a sulfidizing agent [54] and can be used to replace a terminal oxygen atom with a terminal sulfur atom in both organic and inorganic systems [55], [56].

4.2. Production

Boron sulfide is difficult to prepare in a pure form. The reaction of boron hydrides or metal borides with sulfur or sulfur compounds yields impure B_2S_3 [46]. Higher purity product is obtained by the reaction of amorphous boron with H_2S or sulfur. Other preparative reactions that have been reported [47], [48] are:

$$3\,FeS + 2\,B \xrightarrow{700\,^\circ C} B_2S_3 + 3\,Fe \text{ [47]}$$
$$Ag_3B_5S_9 \xrightarrow[10\,d]{600-800\,^\circ C} B_2S_3 \text{ [48]}$$

The *recommended route* for the synthesis of pure B_2S_3 is the decomposition of $(XBS)_3$, where X is Cl, Br, or I [46], [51]:

$$BBr_3 \xrightarrow{H_2S} (SBSH)_2 \xrightarrow{BX_3} (XBS)_3 \longrightarrow B_2S_3$$

4.3. Uses

Boron sulfide has been used in organic synthesis to reduce sulfoxides to thiols [57], and to convert ketones to thioketones [55],[58]. The sulfide glasses, $M_2S-B_2S_3$, where M is Li or Na, have electrical conductivities 10^3 times greater than the corresponding oxide glasses [59]. Electrically conducting boron sulfide glasses have also been made incorporating LiI [60] and other glass-forming sulfides [61], [62]. These glasses have potential as electrolytes in batteries or fuel cells.

Boron sulfide is available in research quantities from the Alfa-Ventron Division of Morton-Thiokol (USA).

5. Boranes

5.1. Diborane and Higher Boranes

Introduction. Alfred Stock and co-workers began working on the hydrides of boron in 1912 [63]. From then until 1936, he discovered a series of boron hydrides of the composition B_nH_{n+4} and B_nH_{n+6}, which he named boranes.

Research on boron hydrides was started in the United States by H. I. Schlesinger, who with A. Burg published a new preparation of diborane in 1931 [64]. Impetus was added by

government sponsorship of various military programs to develop high-energy fuels for rockets and aircraft during World War II. Boron fuels (alkyl penta- and decaboranes) created great interest because they possessed much higher heats of combustion than conventional hydrocarbon fuels. Large quantities of boron fuels were manufactured under projects HEF (Olin Matheson Chemical Corp.) and ZIP (Callery Chemical Co.).

Attention has been expanded to other chemical properties of the boranes since 1956, when H. C. Brown discovered hydroboration procedures [65] (1979 Nobel Prize). His work has opened new avenues of research, one of which was to obtain chiral and optically specific compounds required in biological activity.

Progress in borane research was thwarted by inconclusive theories of structure and reactivity until the second half of the century, when W. Lipscomb and co-workers defined the structural and theoretical aspects of the boron hydrides through X-ray diffraction studies [66] (1976 Nobel Prize). This breakthrough in borane chemistry gave new impetus to boron research and to the discovery of the most thermodynamically and kinetically stable class of boranes, the polyhedral anions, carbaboranes, and metalloboranes.

Collectively, boranes are compounds of very complex structure with $B-H^{\mu}-B$ and (open or closed structured) $B-B^{\mu}-B$ tricentered bonds, with H^{μ} equal to linking hydrogen atoms. Explanation of structures and unusual bonding of these compounds has furthered the theory of chemical bonding [67]. Nomenclature of these compounds has been confusing, however, and recommendations to clarify their structures have been made by IUPAC and ACS [68].

5.1.1. Physical and Chemical Properties

The lower boranes are very reactive to oxygen and moisture. Diborane(6), B_2H_6, and air mixtures containing 75 – 98 % diborane detonate in air when sparked, and concentrations over 0.5 vol % can lead to explosions in air or oxygen even without an ignition source. For investigations of oxidation see [69–71]. Pentaborane(9), B_5H_9, a volatile liquid, reacts readily with air and moisture. Solid decaborane(14), $B_{10}H_{14}$, should be handled in an inert atmosphere; exposure should be avoided because of the compound's toxicity (see Table 8). Solutions of $B_{10}H_{14}$ and carbon tetrachloride are explosive and more shock sensitive than nitroglycerine [72]. Some physical properties of the most commercially available boranes are given in Table 5 (see below).

As a Lewis acid, the borane(3), BH_3, group complexes with Lewis bases, analogous to coordinate compounds of the Werner type; the ligands NH_3, NH_2R, and NR_3 are common to the two cases [73]. The Lewis acid BH_3 combines with another borane (Lewis acid) to give the relatively stable borane(6); hence borane groups may be considered as operational "electron-deficient" bases as well as Lewis acids. How the nature of the base or nucleophile affects the cleavage of diborane is controversial and has been discussed [74]. With neutral bases the two modes of cleavage, symmetric or unsymmetric, may occur to give molecular or ionic fragments as shown in Figure 1 (see next page). Generally with large bases or weaker donor atoms symmetric cleavage occurs. Base displacement reactions readily occur and the relative stability of the complexes is:

Group V: P > N
Group VI: S > O
Group V > Group VI

Table 5. Physical properties of boranes

Formula	CAS registry no.	M_r	*mp*, °C	*bp*, °C (101.3 kPa)	$\Delta H°$ fusion, kJ/mol
B_2H_6	[*19287-45-7*]	27.69	−164.9	−92.6	35.5
B_4H_{10}	[*18283-93-7*]	53.32	−120	18	66.1
B_5H_9	[*19624-22-7*]	63.17	− 46.6	48	73.2
$B_{10}H_{14}$	[*17702-41-9*]	122.22	99.7	213	31.5

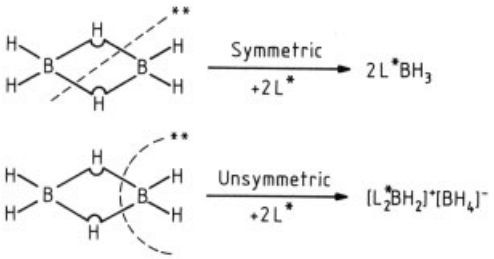

Figure 1. Cleavage of diborane
∗ L is any ligand. ∗∗ Modes of cleavage.

Lewis bases react with decaborane, without deprotonation, evolving hydrogen with the ligands coordinating at the 6,9 position. This species, $B_{10}H_{12}L_2$, is a key intermediate in the synthesis of the very stable and useful *closo*-boranes, $[B_{10}H_{10}]^{2-}$, and carbaboranes (where carbon is inserted in the 1,2 position of the polyhedron, i.e., 1,2-$C_2B_{10}H_{12}$).

Bridge hydrogens of the boranes are acidic as demonstrated by titration of $B_{10}H_{14}$ and by deuterium exchange [75], [76]. Abstraction of a proton forms hydropolyborate anions, such as $B_5H_8^-$. They are useful intermediates in the formation of higher boranes, heteroboranes, and metalloboranes. These anions, containing a B−B edge bond, can be reacted with diborane to expand the polyhedron [77].

5.1.2. Production

Diborane(6) is available from Callery Chemical Co., Callery, Pennsylvania (USA). The method preferred for the commercial preparation is the addition of boron trifluoride to a stirred solution of sodium borohydride in diethylene glycol dimethyl ether (diglyme):

$$3\ NaBH_4 + 4\ BF_3 \xrightarrow{\text{Diglyme}} 2\ B_2H_6 + 3\ NaBF_4$$

Diborane is shipped either as a compressed gas (ambient) or as a liquid (cylinder over packaged with dry ice). It is packaged as a hazardous material regulated by the DOT [78].

Tetraborane(10), B_4H_{10}, can be obtained by the action of acid on borides [79–81]. It has been prepared from the decomposition of diborane at 120 – 180 °C [82] and under pressure [83]. Tetraborane can also be prepared from the reaction of acid on commercially available $B_3H_8^-$ salts, e.g., NaB_3H_8 [*12429-74-2*].

Pentaborane is obtained by passing diborane through a hot tube. Careful control of temperature, pressure, and flow are required to obtain good yields and avoid further pyrolysis to higher hydrides.

Decaborane can be produced by several procedures, none of which give high yields; all require extensive preparatory time or equipment. It is best produced by the pyrolysis of diborane between 120 – 240 °C. The solids thus produced are removed from the reactor by dissolution and further purified by sublimation.

5.1.3. Uses

Most of the boranes are highly reactive. This property makes them useful and important chemicals, but also necessitates special care in their application.

Diborane serves as a strong but selective reducing agent in organic chemistry. It is an electrophilic reagent and reacts markedly different from the nucleophilic alkali-metal boron and aluminum hydrides. The rates of reduction with diborane were found to decrease in the order: carboxylic acids > olefins > ketones > nitriles > epoxides > esters > acid chlorides [84]. Reductions of several other organic substrates have been studied, for example, oximes [85], perfluorinated olefins [86], diene polymers [87], and organic acid amides [88].

Hydroboration, the addition of diborane or borane to olefins, has found great significance in preparative chemistry [65]. Hydroboration is applicable to multiple bonds between C−C, C−O, C−N, N−N, N−O. Diborane catalyzed by the presence of an ether adds, in an anti-Markovnikov mode, virtually quantitatively to the carbon – carbon double bond of olefins, forming an alkyl borane.

Other important derivatives of diborane are Lewis base adducts, carbonyl adducts, and oxygen-containing products. Further areas of application for diborane are the doping of semiconductor silicon and germanium. Diborane is readily decomposed at high temperature (600 – 700 °C) [89] to give fine-powder, high-purity elemental boron. Copyrolysis of diborane with hydrocarbons yields boron carbide,

and with NH_3 yields boron nitride. These materials, deposited on metal substrates as thin, hard layers, are used in ceramics, tools, and wear-resistant parts. Boron fibers are formed from the decomposition of diborane on substrates and used for light-weight structural components.

The large cross-sectional area of the ^{10}B isotope for neutron capture makes these compounds especially useful in nuclear applications. Borane is also supplied as stable liquid adducts $(CH_3)_2S \cdot BH_3$ [*13292-87-0*] (Callery Chemical Co., Aldrich Chemical Co.), and as a 1 M solution in tetrahydrofuran (THF) [*14044-65-6*] (Aldrich Chemical Co., Ventron Div. of Morton-Thiokol Inc.).

5.2. Amine Boranes

Most amine boranes are white solids, stable to air and moisture. Primary and secondary amine boranes are, however, thermally unstable above 70 °C.

Aliphatic amine boranes are soluble in a variety of protic and nonprotic solvents, in contrast to $NaBH_4$. Hydrolysis is slow under neutral conditions and increases in rate as the pH is lowered. Arylamine boranes are hydrolyzed much more readily.

Production. Preparation may be by reaction of diborane or dimethyl sulfide borane with an amine, or by reacting an ammonium salt of the desired amine with $NaBH_4$ in an ether [90]. The dimethyl sulfide borane route is the most convenient method of preparing small quantities of amine boranes [91].

Both $(CH_3)_2NH \cdot BH_3$ [*74-94-2*] and *tert*-$C_4H_9NH_2 \cdot BH_3$ [*7337-45-3*] are available in up to tonnage quantities, whereas smaller quantities of $(CH_3)_3N \cdot BH_3$ [*75-22-9*], $(C_2H_5)_3N \cdot BH_3$ [*1722-26-5*], $(C_2H_5)_2NH \cdot BH_3$ [*2670-68-0*], pyridine borane [*110-51-0*], and morpholine borane [*4856-95-5*] are commercially available (Callery Chemical Co., Bayer AG).

Uses. The largest quantities of amine boranes are consumed in the electroless plating of various metals, primarily nickel. Dimethylamine borane is the preferred reductant in North America, whereas diethylamine borane is used in Europe. The nickel thus plated by chemical reduction is amorphous and contains varying amounts of boron depending on the bath formulation [92]. This plating mainly provides corrosion protection and high solderability to electrical parts. Amine boranes have also been used to plate copper, gold, palladium, cobalt, and other metals [93].

Amine boranes are finding expanding use as mild reducing agents, comparable to cyanoborohydride in reducing power, but considerably less toxic. Amines of greater base strength give amine boranes of diminished reducing power; for example, $(CH_3)_3N \cdot BH_3$ is a relatively weak reducer, whereas $NH_3 \cdot BH_3$ is a moderately strong reducer.

It has been shown that $NH_3 \cdot BH_3$ and primary and secondary amine boranes are efficient stereoselective reducing agents for aldehydes and ketones in a variety of solvents [94]. Additionally, these reagents selectively reduce aldehydes in the presence of ketones or esters [95], [96].

Chiral amine boranes have been useful in the asymmetric reduction of prochiral ketones. In the presence of BF_3 etherate, borane complexes of sodium salts of amino acids gave enantiomeric excesses up to 62 % [97].

Pyridine borane in acidic media is a versatile and powerful reducing agent. It reduces aldehydes, ketones, oximes, Schiff bases, tosylhydrazines, and aryl alcohols [98].

Polymer-bound amine boranes (Amboranes, Alpha Ventron) have been proposed for the recovery of precious metals. They are effective, if slow, reducing agents for imines, enamines, aldehydes, and ketones [99].

Amine borane analogue of amino acids showed anti-inflammatory activity in rodents [100].

5.3. Borohydrides

Of the borohydrides or hydroboron ions, the most common and most completely studied is the tetrahydromonoborate(1−). The oxidation state of the central atom in the higher hydroboron ions is fractional in ions containing several boron atoms and the charge is listed as a suffix in parentheses [101]. Generally, the borohydrides are highly polarizable [102] and the bonding may

vary from essentially ionic to essentially covalent.

All of the alkali-metal tetrahydroborates have been prepared and are white crystalline solids with markedly different solubilities. Sodium tetrahydroborate(1−) [*16940-66-2*], $NaBH_4$, M_r 37.83, is stable in dry air, reacting slowly with moisture. Hydrolysis is catalyzed by acid, but can be terminated with sufficient sodium hydroxide [103]. Salts of the ions $B_3H_8^-$, $B_{11}H_{14}^-$, $B_{10}H_{10}^{2-}$, and $B_{12}H_{12}^{2-}$ are all white crystalline solids. The latter two salts are extremely stable thermally, especially their halo derivatives $B_{10}X_{10}^{2-}$, $B_{12}X_{12}^{2-}$.

Production. Sodium tetrahydroborate is made by the addition of methyl borate to hydrogenated sodium in mineral oil [104], [105]:

$$4\,Na + 2\,H_2 + B(OCH_3)_3 \xrightarrow[275\,^\circ C]{oil} NaBH_4 + 3\,NaOCH_3$$

Separation of the sodium methoxide from $NaBH_4$ was accomplished by such methods as extraction with amines [106], and extraction with water followed by a counterextraction with a solvent, such as 2-aminopropane [107]. Sodium borohydride is produced commercially by Ventron Div., Morton-Thiokol (USA) [108].

A simple preparation for $B_3H_8^-$, adaptable to large scale, involves the addition of B_2H_6 to $NaBH_4$ in diglyme at 100 °C [109]. Further addition of B_2H_6 to the $B_3H_8^-$ at higher temperature yields $B_{12}H_{12}^{2-}$ [110].

Uses. A concentrated aqueous solution of $NaBH_4$ stabilized by NaOH is used to reduce hydrogen sulfite to dithionite in the bleaching of ground wood pulp [111]. Also, $NaBH_4$ has extensive and important use as a reducing agent in organic synthesis. Other hydroborates can be obtained by reaction with $NaBH_4$ to give different reductive capabilities (e.g., $NaBH_3CN$ [*25895-60-7*]) [112].

Additional applications are extensive, e.g., electroless plating [113], blowing agent for cellular plastics [114], removal of heavy metals from waste streams [115], polymerization catalyst [116], and hydrogen source for fuel cells [117]. The $B_3H_8^-$, $B_{10}H_{10}^{2-}$, and $B_{12}H_{12}^{2-}$ may also find use in the latter application. The quaternary ammonium salt of $B_3H_8^-$ [*33055-82-2*] is used as a fogging agent in photographic film [118]. Chloro derivatives of $B_{10}H_{10}^{2-}$ and $B_{12}H_{12}^{2-}$ lithium salts have been used as battery electrolytes [119]; $B_{12}H_{11}SH^{2-}$ [*12448-23-6*] was useful in a study of neutron-capture therapy in brain tumors [120].

5.4. Carbaboranes

Carbaboranes are boron hydrides with one to four of the polyhedral sites occupied by carbon atoms. The carbon atom, like the cage boron atoms, can bond singly to hydrogen, carbon, etc. The number of known substituted carbaboranes is very large [121], [122] and has grown rapidly.

The largest body of literature is based on 1,2-dicarba-*closo*-dodecaborane(12) [*16872-09-6*], *o*-carborane, 1,2-$C_2B_{10}H_{12}$, M_r 144.23, *mp* 285–287 °C, and its derivatives (for cage numbering system see [121]).

Production. 1,2-Dicarba-*closo*-dodeca-borane(12) is prepared from $B_{10}H_{14}$, a Lewis base, such as dialkyl sulfide, and acetylene [123]. *C*-Substituted derivatives can be readily prepared from substituted acetylenes or from lithio reagents, such as 1,2-Li_2-1,2-$C_2B_{10}H_{10}$ [*22220-85-5*] [121].

Heating 1,2-$C_2B_{10}H_{12}$ under inert conditions at 450–500 °C gives 1,7-$C_2B_{10}H_{12}$; at 600 °C it gives a mixture of 1,7-$C_2B_{10}H_{12}$ and 1,12-$C_2B_{10}H_{12}$ [*20644-12-6*], sometimes called the meta and para isomers, respectively.

There is a large body of literature on the rest of the $C_2B_nH_{n+2}$ family ($n = 3-10$) [124], but the members having $n < 10$ are of little commercial significance.

One boron atom may be removed from 1,2-$C_2B_{10}H_{12}$ and 1,7-$C_2B_{10}H_{12}$ to yield the open-cage species *nido*-7,8-$C_2B_9H_{12}^-$ and *nido*-7,9-$C_2B_9H_{12}^-$, respectively. Further degradation yields *arachno*-boranes (e.g., $C_2B_nH_{n+6}$) as well as nido and closo species [125].

The nido cages have been compared with cyclopentadienide ion as ligands for transition metals. Indeed, a rather large number of transition-metal carbaborane complexes are known [126]. Complexes with two carbaborane cages sandwiching a transition metal are common, e.g., $Co(B_9C_2H_{11})_2^-$, whereas cages containing two

transition metals, usually also bonded to cyclopentadienide, are also known. Some smaller carbaboranes form multidecker complexes with transition-metal cyclopentadienyls [127].

Carbaboranes, like polyhedral boranes, undergo Friedel-Crafts halogenation, although halogenation is complete only for the chlorine derivative [121].

Uses. The early development of carbaborane substitution chemistry was propelled by the search for polymers that would be heat stable at very high temperatures. Polyesters, siloxanes, chains linked with Si, Ge, S, and others were investigated [121].

A short review of structure – property correlations for linear polycarborane – siloxanes has been published [128]. Much of the original work on poly-1,7-carboranylsiloxanes was done at Olin Corp., which formed a company (Dexsil Chemical Corp.) to market the materials under the trade name Dexsil. Early hopes were that markets would grow for high-temperature seals, gaskets, *O*-rings, and coatings. However, costs and therefore prices have remained too high to justify most such uses.

One application that may remain economically viable is the use of poly-1,7-carboranylsiloxanes as stationary phases for high-temperature gas chromatography. For example, Dexsil 300 GC is a low-melting solid that can be used up to 450 °C and even at 500 °C for short times. A bibliography of their use in gas chromatography is available [129].

Carbaboranes have been attached to halo-substituted cyclotriphosphazines. A heat-resistant linear polymer is formed when the monomer is heated [130].

If the residual Cl atoms are replaced with hydrocarbyloxy groups in the polymer, boron atoms can be removed by piperidine to form nido cages, which can complex with transition metals, such as Rh. These polymeric metal complexes might be useful as immobilized catalysts with high temperature stability [131].

A series of rhodacarborane complexes that exhibit high activity as hydrogenation or alkene isomerization catalysts has been developed [132]. One of the most active homogeneous hydrogenation catalysts ever reported is *closo*-1,3-[$\mu(\eta^2$-3-CH_2=$CHCH_2CH_2$)-3*H*-3-$P(C_6H_5)_3$-3,1,2-$RhC_2B_9H_{10}$ [*76166-14-87*], containing a chelating 3-butenyl side chain attached to the carborane cage. Hydrogenation produces a noncoordinating butyl group, freeing up an uncoordinated site on the rhodium [133]. The alkene – metalloborane catalyst may be bound to a suitable polymer for use as a heterogeneous catalyst [134].

Thermal reaction of 1,3-dicyanobenzene in the presence of 1,2-dicarba-*closo*-dodecaborane(12) produced *s*-triazines [135].

1-Alkyl-1,2-dicarba-*closo*-dodecaboranes(12)(especially the *n*-hexyl derivative) have found use in solid aluminum composite propellants as burning rate enhancers [136] based on OH- or CO_2H-terminated polybutadienes. Exceptionally high burning rates have been achieved when part of the NH_4ClO_4 oxidizer was replaced by carboranyl diferrocenylmethyl perchlorate or methyltricarboranylmethyl perchlorate [*61919-09-3*] [137]. A series of nido – closo dicage monoanions exhibited high antifungal activity. The combination of lipophilic *closo*-carboranyl groups with hydrophilic *nido*-dicarbollide anions may lead to the antimicrobial activity [138]. A review covers the physiology of various carbaborane derivatives [139].

Several carbaborane derivatives may find use in boron neutron-capture cancer therapy (BNCT), based on the high cross section of ^{10}B for thermal neutron capture. The products are an α particle and a ^7Li ion with an average kinetic energy of 3.8×10^{-13} J. Use of boranes in BNCT has been reviewed [139, pp. 610 – 631].

p-[1,2-Dicarba-*closo*-[1-^3H]-dodecaborane(12)-2-yl]benzenediazonium chloride [*82187-51-7*] was complexed with goat anti-CEA IgC (carcinoembryonic antigen immunoglobin C), and the antibody lost little immunoreactivity up to five carbaborane cages per molecule [140]. Thus carbaborane-substituted tumor-specific antibodies may be a good method of concentrating ^{10}B in tumors, a goal for many years.

An approach to treatment of hormone-sensitive tumors is to attach carbaborane cages to estradiol or testosterone [141]. Growth of human breast cancer cells was equally stimulated by 17-α-carboranylestradiol [*85197-15-5*], Carbestrol [142].

Carbaborane-substituted polyphosphonates have shown a high affinity for calcified tissues.

They may be useful for BNCT of calcific tumors [143].

6. Organoboranes

6.1. Alkyl- and Arylboranes

Physical Properties. Mono- and dialkylboranes usually associate as dimers through hydrogen bridges (see Fig. 1). However, if the alkyl group is bulky, i.e., sterically hindered, the mono- or dialkylborane can exist in monomeric form, e.g., thexylborane [*3688-24-2*]. The bulkiness of the alkyl group considerably affects the properties of the borane. Compounds containing large hindered groups often are solid products, and as the size of the group decreases, the normal state of the compound goes from solid to liquid to gas. Properties of selected organoboranes are given in Table 6.

Chemical Properties. Generally the organoboranes are very reactive with oxygen and total exclusion is recommended on handling. The low homologues (methyl to butyl) are spontaneously combustible (pyrophoric) in air. Although dilution of the alkylboranes to about 15 wt % in an inert solvent reduces the pyrophoricity, oxidation by air still occurs. The trialkylboranes do not form strong donor-acceptor complexes with ethers, but other stronger Lewis bases form complexes, and stabilization toward oxidation is dependent upon steric factors. The trialkylboranes are insoluble in and unreactive with water. They are not readily protonated by water or alcohols, amines, or mineral acids; however, they are readily protonated by carboxylic acids. They are soluble in and unreactive with aliphatic and aromatic hydrocarbons.

Triarylboranes are generally solid, less sensitive to oxygen, and also unreactive with water.

Triorganoboranes can be stored indefinitely at ambient temperature under an inert atmosphere. Decomposition occurs at elevated temperature: tetraethyldiborane [*12081-54-8*] is one of the pyrolysis products from triethylborane, and butene is observed from tri-*n*-butylborane at 100 °C after an extended period of time. Isomerization and cyclization occur at > 200 °C, giving complex mixtures. Alkylhaloboranes, like the trihalides, react with water, alcohols, and alkoxides, and cleave ethers to yield the corresponding derivatives of boron acids [146].

Production. Two general methods of producing trialkylborane compounds are alkylation and hydroboration. Alkylation of readily accessible boron compounds, borate esters, or preferably boron halides is accomplished with organometallic compounds of Mg (Grignard reagents), Zn, Al, Sn, and alkali metals:

$$3\,RMgX + BF_3 \xrightarrow{(C_2H_5)_2O} R_3B + 3\,MgXF$$
$$3\,R_2Zn + 2\,B(OR)_3 \longrightarrow 2\,R_3B + 3\,Zn(OR)_2$$
$$R_3Al + B(OR)_3 \longrightarrow R_3B + Al(OR)_3$$

where R is alkyl. Aluminum alkyls are least expensive and therefore the preferred alkylating agents in commercial production of trialkylboranes. The exothermic reaction can be controlled at < 100 °C and the product is easily distilled [147]. Grignard reactions are especially useful in producing trimethylborane as well as arylboranes. Triphenylborane has been prepared by reacting finely divided sodium metal with a haloaromatic and a secondary alkyl borate ester [148]; it is made commercially by Du Pont (USA). Hydroboration [66] can be conducted in various ways, i.e., by direct addition of diborane to the olefin at 30–50 °C, or by use of a Lewis base as a reaction medium, with THF

Table 6. Properties of selected organoboranes [144], [145]

Formula	CAS registry no.	*mp*, °C	*bp*, °C (kPa)	Density, g/cm^3 (°C)
$(CH_3)_3B$	[*593-90-8*]	−160	−21.8 (101.3)	0.63 (−100)
$(C_2H_5)_3B$	[*97-94-9*]	− 93	95 (101.3)	0.685 (20)
$(C_4H_9)_3B$	[*122-56-5*]	− 34	170 (29.6)	0.756 (20)
$[(CH_3)_2CHCH_2]_3B$	[*1116-39-8*]	− 88	88 (3.1)	0.731 (30)
$(C_6H_5)_3B$	[*960-71-4*]	142	203 (2.0)	–
$(C_2H_5)_2BCl$	[*5314-83-0*]	− 84.6	78.5 (101.3)	–
$(C_4H_9)_2BCl$	[*1730-69-4*]	–	173 (101.3)	0.879 (20)

and $(CH_3)_2S$ being preferred. Various alkylboranes are commercially available by this method (Callery Chemical Co., Aldrich Chemical Co., Ventron).

As the olefin becomes bulkier, fewer additions can be made. A tetrasubstituted olefin usually forms only the monoalkylborane. Lewis base adducts of the mono- and dichloroboranes yield the mono- and dichloroalkylboranes when reacted with olefins. These can be converted to their respective di- and monoalkylboranes by hydride addition [149]. Mixed trialkylboranes were obtained using this method [150].

Uses. Organoboranes are finding an ever increasing role in organic synthesis. Mono- and dialkylboranes have been used as stereoselective reducing agents, and asymmetric synthesis, hydroborations, and reductions have been accomplished using organoboranes [151]. Such reactions are most significant for their biological application in the pharmaceutical industry, and may mean a small but cost effective use of alkylboranes.

Significant volumes of trialkylboranes are used in catalytic polymerization of ethylene, vinyl derivatives (acetates and halides, acrylic acid, acrylonitrile, and methyl methacrylate) [152–154]. Free radicals are formed during autoxidation of the trialkylborane or during the decomposition of the resulting peroxides [155], [156]. Unique products have also been obtained with copolymers and graft polymers. Water-resistant dental and surgical adhesives having high binding power and being nonirritating to tissue have been developed using alkylborane as a catalyst [157], [158]. The amine complexes of triarylboranes can be used as catalysts for homo- and copolymerization of acrylic monomers.

Several reactions of organoboranes have possible applications in the petroleum industry [159]. Economics remain unfavorable because of high costs of the organoborane and equipment to implement recycle. Amine complexes of the trialkylboranes have been used as additives to increase the cetane rating of diesel fuels. The complexes also inhibit sludge formation in fuel oils.

Triethylborane, because of its ease of oxidation, provides rapid ignition of nonhypergolic fuel in aircraft, and serves as an additive to prevent flameout and increase flame speed.

Lithium and potassium hydride complexes of trialkylboranes are highly reactive and selective reducing agents, and have found small but growing commercial applications.

The tetraarylborates(1−) are used in chemical analysis because of their low solubilities with specific cations (K, Cs, Rb, Ag, Tl). This property can also be used to remove these materials from various chemical waste streams.

The production of organoboron compounds is relatively small compared to other organometallic chemicals. Expensive raw materials and low volume of production tend to keep manufacturing costs high.

6.2. Borinic and Boronic Acids

Borinic and boronic acids are the mixed hydroxyorganoboranes R_2BOH and $RB(OH)_2$, respectively. Two commercially available acids are: phenylboronic acid [*98-80-6*], phenylboric acid, $C_6H_5B(OH)_2$, M_r 121.93, *mp* 217–220 °C; and diphenylborinic acid–ethanolamine complex [*524-95-8*], 2-aminoethyl diphenylborinate, $(C_6H_5)_2BOCH_2CH_2NH_2$, M_r 225.10, *mp* 192–194 °C. Replacement of the organo group by hydroxyl changes the properties of the boranes dramatically. Most of them are stable, water soluble, less susceptible than triorganoboranes to autoxidation, and do not initiate olefin polymerization. Their acidity is dependent on the organo moiety, with the aryl derivatives being the more acidic. The anhydrides of borinic acids are the boroxines, trimeric cyclic $(RBO)_3$ compounds, whereas the anhydrides of the boronic acids are $R_2B{-}O{-}BR_2$.

Production. These acids have been prepared by Grignard synthesis. Borate esters give methyl-substituted acids [160]. Alkyl acids can readily be prepared from hydroboration of olefins using complexes of haloboranes followed by hydrolysis. Rearrangement of triorganoboranes with triorganoborates yields esters, which are susceptible to hydrolysis to the acids.

Uses. The acids are of great interest because they are utilized in the production and reactivity of biologically active compounds. Boronic acids have been prepared in high optical purity for

application in organic synthesis [161]. Specific boronic acids have been found to inhibit the catalyzed hydrolysis of enzymes [162]. Stabilizers are also formed from phenols and arylboronic acids. Benzeneboronic acid derivatives, such as cyclic condensation products with *o*-aminophenol and pyrocatechol, have a sterilizing effect on insects (flies) [163]. Also, compounds of benzeneboronic acid are stimulants for plant root growth. The acids are used in chromatographic columns to enhance separation. They also serve as inhibitors for the autoxidation of fuels and mineral oils.

6.3. Borazines

Borazine [*6569-51-3*], *s*-triazaborane, $B_3N_3H_6$, M_r 80.53, sometimes called inorganic benzene, is a low-boiling liquid (*bp* 55 °C, *mp* −58 °C), consisting of a six-membered ring containing alternating boron and nitrogen atoms, each bonded to a hydrogen atom.

Production. Borazine can be prepared by copyrolysis of $NaBH_4$ and NH_4Cl in a high-boiling glyme [164], by reduction of trichloroborazine [*933-18-6*] [165], or by pyrolysis of ammonia borane in diglyme [166].

A large body of literature exists on the preparation and chemistry of symmetrically and unsymmetrically substituted borazines [167]. *B*-Substituted trihaloborazines, readily prepared from BX_3 and NH_4Cl, react with Grignard reagents to give symmetrical trialkyl- or triarylborazines. Symmetrically substituted *N*-trialkyl- or *N*-triarylborazines are readily prepared from $LiBH_4$ and RNH_3Cl, or by pyrolysis of RNH_2BH_3.

Borazine is thermally stable at 0 °C, but decomposes slowly at ambient temperature. Borazine hydrolyzes slowly, but trihaloborazines hydrolyze quite rapidly. Hydrogen halides form 1 : 3 adducts with borazine, which decompose to trihaloborazines on warming.

Uses. Films that were deposited thermally from borazine vapor were found to be too reactive to be suitable for fabricating masks for X-ray lithography [168]. However, films containing predominately cubic boron nitride, CBN [*10043-11-5*], were deposited using ions extracted from a borazine plasma via an ion beam. Very hard, adherent films were deposited on ceramic, glass, stainless steel, etc. [169]. An obvious practical application is the hardening of tool surfaces to markedly extend their life.

A thermally deposited film of hexagonal BN [*10043-11-5*], from $B_3N_3Cl_6$ [*19087-72-0*] at 900 °C, protected SiO_2 tubes from halide corrosion [170].

A copolymer between 2,4,6-tributylborazine [*7325-06-6*] and an unsaturated *N,N′*-hexamethylenebiscarboxamide was useful for hardening epoxy resins [171]. Trichloroborazine was used as part of a catalyst mix for the polymerization of propene [172].

Flexible BN coatings on copper or copper-base alloys were formed from gaseous 2,4,6-trichloroborazine. The films were stable to 700 °C in air [173].

6.4. Boric Acid Esters

The first boric acid esters were prepared in 1846 by reactions of aliphatic alcohols with boron trichloride [174] and represent the earliest known examples of organoboron compounds. Simple boric acid esters have the formula $B(OR)_3$, where R can be alkyl or aryl; related compounds, such as boroxines, $(ROBO)_3$, as well as boric acid esters of di- and polyfunctional hydroxy compounds can be prepared readily and are included in the general category of boric acid esters. The IUPAC committee on boron chemistry suggested naming simple boric acid esters as trialkoxy- and triaryloxyboranes, and this nomenclature has been adopted by the American Chemical Society [175] and by Chemical Abstracts. However, this system is not normally used in industry or in the scientific literature, where these compounds continue to be referred to as boric acid esters, alkyl (or aryl) borates or orthoborates, trialkyl (or aryl) borates or orthoborates, and (rarely) boron trialkoxides or aryloxides. For example, the most common derivatives are usually referred to as methyl borate, ethyl borate, phenyl borate, etc. Extensive discussions of the chemistry of boric acid esters can be found in [176], [177].

Table 7. Properties of some boric acid esters

Compound	Name	CAS registry no.	*mp*, °C	*bp*, °C (at 101.3 kPa)	d_4^t	n_D^t
$B(OCH_3)_3$	methyl borate	[*121-43-7*]	−29	68.0–68.5	0.920^{20}	1.3548^{25}
$B(OCH_3)_3-CH_3OH$	methyl borate azeotrope	–	–	54.3	0.8804^{25}	1.3472^{25}
$B(OCH_2CH_3)_3$	ethyl borate	[*150-46-9*]	−84.8	117–119	0.859^{26}	1.3723^{25}
$B[OCH(CH_3)_2]_3$	isopropyl borate	[*5419-55-6*]	–	139–140	0.815^{23}	1.3750^{25}
$B(OCH_2CH_2CH_2CH_3)_3$	*n*-butyl borate	[*688-74-4*]	–	227	0.856^{25}	1.4077^{25}
$B(OC_6H_5)_3$	phenyl borate	[*1095-03-0*]	71–81 *	360–370 224–230 (at 2.3)	–	–
$B_2C_{12}H_{24}O_5$	2,2′-oxybis(4,4,6-trimethyl-1,3,2-dioxaborinane)	[*14697-50-8*]	–	114–115 (at 0.27)	1.013^{24}	1.4308^{25}
$B_2C_{18}H_{36}O_6$	2,2′-(1,1,3-trimethyltrimethylenedioxy)bis-(4,4,6-trimethyl-1,3,2-dioxaborinane)	[*100-89-0*]	–	274–276 170–172 (at 2.7)	0.932^{21}	1.4381^{25}
$B_2C_{12}H_{24}O_6$	2,2′-(1-methyltrimethylenedioxy)bis-(4-methyl-1,3,2-dioxaborinane)	[*2665-13-6*]	–	207–213 (at 2.3)	1.071^{25}	1.4464^{17}
$(CH_3OBO)_3$	methyl boroxine	[*102-24-9*]	–	decomp.	1.2286^{25}	–
$[(CH_3)_2CHOBO]_3$	isopropyl boroxine	[*10298-87-0*]	52–54	235–239 decomp.	–	–
$B(OCH_2CH_2)_3N$	triethanolamine borate	[*283-56-7*]	235–239	–	–	–

* Phenyl borate has been reported to melt over a wide range of temperatures from 38 to 146 °C

Physical Properties. Boric acid esters range from colorless, low-boiling liquids to high-melting solids. Physical properties of some of the more common materials are given in Table 7. For a more extensive compilation see [176], [178]. Boric acid esters are monomeric compounds generally soluble in organic solvents. They dissolve in water with hydrolysis to boric acid and the corresponding alcohol or phenol.

Chemical Properties. Alkyl boric acid esters from straight-chain alcohols and aryl boric acid esters are stable to relatively high temperatures. Branching in the alcohol chain leads to reduced stability.

Most boric acid esters react readily with water to yield boric acid and the respective alcohol or phenol, and this tendency to react with atmospheric moisture has seriously restricted their general utility. Reactivity toward water can be modified by either steric or electronic effects as shown by the improved hydrolytic stability of compounds, such as tri-*tert*-amyl borate [*22238-22-8*] and triethanolamine borate. Comparative hydrolysis rates have been reported for a number of boric acid esters [179].

Other reactions of boric acid esters include the formation of amine complexes, transesterification, usually by displacement of lower boiling alcohols by higher boiling alcohols or phenols, reactions with organometallic reagents to form alkylboranes, and reactions with metal hydrides to produce boron hydrogen compounds.

Preparation. Boric acid esters have been prepared in the laboratory by a variety of methods, but the standard approach involves the reaction of boric acid with alcohols or phenols:

$$B(OH)_3 + 3\,ROH \rightleftharpoons B(OR)_3 + 3\,H_2O$$

This equilibrium reaction, which favors the hydrolysis products boric acid and ROH, is displaced to the right by using excess alcohol and removing water by azeotropic distillation with hydrocarbon solvents. This procedure is complicated in the case of methyl and ethyl borates by the formation of azeotropes between the alcohols and the boric acid esters. A closely related method in which boric oxide is substituted for boric acid has also been used.

The relatively easy transesterification of boric acid esters results in ready displacement of lower boiling alcohols and the preparation of higher molecular mass alkyl and aryl derivatives.

A number of less practical laboratory procedures for boric acid esters have been reported, starting with boron halides, trisaminoboranes, and other boron-containing compounds.

Production, Economic Aspects. No published data are available on manufacturing processes for boric acid esters or on their annual production rates. Commercial usage is very limited except for the internal production of methyl borate by the Ventron Division of Morton-Thiokol. This ester is a key intermediate in the Ventron process for sodium borohydride. Some additional quantities of methyl borate and its methanol azeotrope are used in welding and brazing applications. Other commercially available boric acid esters include ethyl borate, *n*-propyl borate, isopropyl borate, *n*-butyl borate, and cresyl borate. A number of boric acid esters prepared from butylene and hexylene glycols are also available commercially, as well as methyl and isopropyl boroxines. Probably, all of these compounds are prepared by the standard reactions of boric acid with the appropriate alcohols or phenols.

Prices for the commercially available boric acid esters are relatively high because of their limited usage. At higher production volumes these prices would be reduced significantly. Methyl borate was available in 1985 at ca. $ 3/kg, and the other esters are generally in the range of $ 3 to $ 10/kg.

Uses. Methyl borate is the boric acid ester of greatest commercial interest because it is used internally for the production of sodium borohydride (and diborane). Methyl borate azeotrope is used as a gaseous flux for brazing and welding, particularly in Europe.

Although numerous patents have been issued on the use of boric acid esters as polymer additives, this application apparently has remained small unless there are proprietary uses that have not been disclosed. Application of small quantities of several derivatives as epoxy resin curing agents has continued with little measurable growth. Despite wide patent coverage, boric acid esters have not been used in any quantity for polymer stabilization. In the 1950s and 1960s boric acid esters derived from glycols were used extensively as gasoline additives. A mixture of these glycol derivatives is being manufactured by U.S. Borax & Chemical Corp. under the trade name Biobor JF for the control of microorganisms in fuel systems containing water. This product can be used successfully in jet fuels, diesel fuels, and fuel oil systems that have water contamination.

Borate esters have also been patented for use as flame retardants for cotton batting, modifiers for the oxidation of hydrocarbons, starting materials for the preparation of trialkylboranes, such as triethylborane and triphenylborane, wood preservatives, hydraulic fluids (particularly alkylated glycol borates), and lubricants.

Transportation, Handling. Shipping regulations for boric acid esters vary depending on the nature of the alcohol or phenol that is bound to boron. Lower molecular mass derivatives have very low flash points, and those derived from alcohols containing one to four carbon atoms must be labeled as flammable, and shipped and stored accordingly. Higher molecular mass compounds present no particular difficulty except in specific cases, such as the easily hydrolyzed borates derived from phenols. Boric acid esters in general hydrolyze rapidly and their exposure to the atmosphere should be minimized. Particularly reactive compounds should be transferred and stored in a moisture-free atmosphere of nitrogen or other inert gas.

Shipping in metal cans or drums is normally acceptable except in cases where hydrolysis can lead to a corrosive product, such as a halogenated alcohol.

Toxicology. The published toxicities of the most common boric acid esters vary from acute oral LD_{50} s of ca. 400 mg/kg in rats for aromatic borates to 1200 – 2100 mg/kg for simple aliphatic esters. These toxicities are directly related to the toxicity of the alcohol or phenol produced on hydrolysis.

7. Toxicology and Occupational Health

The toxicology of boron compounds has been reviewed in 1980 [180]. Experimental animal toxicity data, cases of human exposures, industrial

Table 8. Maximum concentrations in the work environment [184], [185]

	CAS	TWA		STEL		MAK	
	registry no.	ppm	mg/m^3	ppm	mg/m^3	ppm	mg/m^3
Diborane(6)	[*19287-45-7*]	0.1	0.1	–	–	0.1	0.1
Pentaborane(9)	[*19624-22-7*]	0.005	0.01	0.015	0.03	0.005	0.01
Decaborane(14)	[*17702-41-9*]	0.05	0.3	0.15	0.9	0.05	0.3

hygiene, and medical treatments are discussed. There are also other readily accessible sources of information [181], [182]. Diborane(6) is primarily a pulmonary irritant [183]; TLVs and MAKs are given in Table 8 [184]. Overexposure causes respiratory tract injury, and secondary infections may follow [186]. Prolonged exposure to low concentrations causes headaches and other central nervous system (CNS) effects [187]. Diborane is a flammable gas and must be handled in a closed system. Pentaborane(9) and decaborane(14) affect the CNS and can have effects that may be delayed for up to 48 h [187]. TLVs have been set (see Table 8) [184]. Pentaborane(9) is a liquid which should be treated as pyrophoric and must be handled in a closed system. Decaborane(14) is a solid but has sufficient vapor pressure at room temperature to make it a hazard by inhalation, as well as by the dermal and oral routes [188].

Animal toxicity testing of amine boranes has shown that they also affect the CNS [189]; however, no TLVs or MAKs have been set for them. Toxicity is affected by the number and kind of substituents and varies from compounds that are highly toxic to those with low toxicity [189], [190].

The few toxicity studies on borohydrides show that they are corrosive to the skin and eyes [191]. When ingested, sodium borohydride produced death from gas embolism in animals.

The lower trialkylboranes have some toxicity to small rodents at nonpyrophoric concentrations in air [180], [192].

Toxicity of carbaboranes depends mainly on the nature of the substituents [180], [193].

8. References

General References

1. *Gmelin,* **vol. 34,** sect. 1 & 9, 332 pp.
2. *Gmelin,* **vol. 53,** sect. 2 & 19, 341 pp.
3. *Gmelin,* **vol. 2(1),** pp. 255 – 316.
4. *Gmelin,* **vol. 2(2),** pp. 1 – 154.
5. N. N. Greenwood, B. S. Thomas in J. C. Bailar (eds.): *Comprehensive Inorganic Chemistry,* **vol. 1,** Pergamon Press, Oxford 1973, p. 956.
6. G. Urry in E. Muetterties (ed.): *The Chemistry of Boron and Its Compounds,* J. Wiley & Sons, Inc., New York 1967, pp. 325 – 375.
7. J. D. Cox, *J. Chem. Thermodyn.* **10** (1978) 903 – 906.
8. A. Finch, P. J. Gardiner in R. J. Brotherton, H. Steinberg (eds.): *Progress in Boron Chemistry,* **vol. 3,** Pergamon Press, New York 1970, p. 200.
9. E. W. Rothe, B. P. Mathur, G. P. Reck, *Inorg. Chem.* **19** (1980) 829 – 831.

Specific References

10. D. Mootz, M. Steffen, *Angew. Chem.* **92** (1980) 481; *Angew. Chem. Int. Ed. Engl.* **19** (1980) 483 – 484.
11. D. Mootz, M. Steffen, *Acta Crystallogr. Sect. B,* **B 37** (1981) 1110.
12. J. L. Gay-Lussac, L. J. Thenard, *Ann. Chim. (Paris)* **69** (1809) no. 1, 204 – 220.
13. B. Zawadzki, A. Bulinska, E. Wilk, *Mater. Ogolnopol. Symp. Zwiazki Fluorowe* 1980 (Publ. 1981) 113 – 117. *Chem. Abstr.* **95** (1981) 189416.
14. Kerr-McGee Chemical Corp., US 4125590, 1978 (L. F. Schmoyer).
15. Alcoa, US 4024221, 1977 (A. J. Becker, D. R. Careatti).
16. PPG Industries, Inc., US 4210631, 1980 (N. R. Delue, J. C. Crano).
17. PPG Industries, Inc., US 4213948, 1980 (J. C. Crano).
18. H. Kral, DE 2826747, 1980. *Chem. Abstr.* **92** (1980) 113123.
19. V. S. Khain, V. P. Val'kova, E. S. Kotelevets, *Zh. Neorg. Khim.* **22** (1977) 338.
20. J. Cueilleron, S. C. Viala, *J. Less Common Met.* **58** (1978) 123. *Chem. Abstr.* **89** (1978) 15952.
21. Showa Koji K. K., JP-Kokai 7768754, 1977 (H. Yasutomi, N. Takeshima, T. Matsuno, H. Momo). *Chem. Abstr.* **87** (1977) 206078.

22. Morita Kagaku Kogyo Co. Ltd., JP-Kokai 7896260, 1978 (T. Tatsuno, K. Momota). *Chem. Abstr.* **90** (1979) 11870.
23. Asahi Glass Co. Ltd., JP-Kokai 8047119, 1980 (M. Noshiro, T. Yarita, S. Kobayashi). *Chem. Abstr.* **94** (1981) 7204.
24. E. H. Vernot, J. D. MacEwen, C. C. Haun, E. R. Kinkead, *Toxicol. Appl. Pharmacol.* **42** (1977) 417 – 423. *Chem. Abstr.* **88** (1978) 84173.

General References
25. H. S. Booth, D. R. Martin: *Boron Trifluoride and Its Derivatives,* J. Wiley & Sons, New York, p. 87.
26. D. W. A. Sharp in M. Stacey, J. C. Tatlow, A. G. Sharpe (eds.): *Advances in Fluorine Chemistry,* **vol. 1,** Butterworths Scientific Publications, London 1960, p. 68.
27. *Gmelin,* **vol. 2(1),** p. 276.
28. *Gmelin,* **vol. 2(2),** p. 31.

Specific References
29. R. M. Adams, *Pure Appl. Chem.* **30** (1972) 683.
30. D. Mootz, M. Steffen, *Z. Anorg. Allg. Chem.* **482** (1981) 193.
31. N. V. Krivtsov, K. V. Titova, V. Ya. Rosolovskii, *Zh. Neorg. Khim.* **22** (1977) 679.
32. R. Wilmotte, J. Benezech, *Galvano-Organo* **45** (1976) 313.
33. C. J. Hill, R. P. Lash, *Anal. Chem.* **52** (1980) 24.
34. J. P. Wilshire, W. A. Brown, *Anal. Chem.* **54** (1980) 1647.
35. United States Steel Corp., US 3933605, 1976 (T. J. Butler, R. M. Hudson, C. J. Warning).
36. RCA Corp., US 3979238, 1976 (H. W. Justice).
37. Rockwell International Corp., US 4004957, 1977 (L. J. Quintana).
38. Y. Nakao, H. Nakeshima, M. Ohta, K. Furukawa, *J. Nucl. Sci. Technol.* **15** (1978) 76. *Chem. Abstr.* **88** (1978) 96218.
39. United States of America, US 4207124, 1980 (K. O. Christe).
40. Morita Chemical Industry Co., Ltd., JP-Kokai 76116060, 1976 (Y. Mochida, T. Tatsuno).*Chem. Abstr.* **87** (1977) 90347.
41. Dart Industries, Inc., US 4008162, 1977 (T. F. Korenowski, J. L. Penland, C. J. Ritzert).
42. Showa Koji K. K., JP-Kokai 7692561, 1976 (H. Yasutomi, N. Takeshima, T. Matsuno, H. Momo). *Chem. Abstr.* **85** (1976) 166268.
43. Nitto Electric Industrial Co., Ltd., JP-Kokai 82162692, 1982.*Chem. Abstr.* **98** (1983) 59412.
44. P. Langer, H. Kokesova, K. Gschewendtova, *Acta Endocrind. Copenhagen* **81** (1976) 516. *Chem. Abstr.* **85** (1976) 40748.
45. M. W. Curtis, C. H. Ward, *J. Hydrol. (Amsterdam)* **51** (1981) 359. *Chem. Abstr.* **95** (1981) 55868.
46. E. L. Muetterties: *The Chemistry of Boron and Its Compounds,* J. Wiley & Sons, New York 1967, pp. 647 – 667.
47. H.-Y. Chen, B. R. Conard, P. W. Gilles, *Inorg. Chem.* **9** (1970) 1776 – 1777.
48. H. Diercks, B. Krebs, *Angew. Chem.* **89** (1977) no. 5, 327; *Angew. Chem., Int. Ed. Engl.* **16** (1977) no. 5, 313.
49. E. G. Zhukhov, S. A. Dembovski, *Izv. Akad. Nauk SSSR Neorg. Mater.* **16** (1980) 37 – 41. *Chem. Abstr.* **92** (1980) 117432.
50. M. P. Morozova, G. A. Rybakova, *Vestn. Leningr. Univ., Fiz., Khim.* **23** (1968) no. 22, 161 – 163. *Chem. Abstr.* **71** (1969) 7162.
51. N. N. Greenwood, B. S. Thomas in J. C. Bailar (ed.): *Comprehensive Inorganic Chemistry,* **vol. 1,** Pergamon Press, New York 1973, pp. 909 – 916.
52. B. Krebs, H. U. Huerter, *Angew. Chem.* **92** (1980) 479; *Angew. Chem., Int. Ed. Engl.* **19** (1980) 481 – 482.
53. H.-Y. Chen, P. W. Gilles, *J. Am. Chem. Soc.* **92** (1970) 2309 – 2312.
54. J. H. Holloway, D. C. Puddick, G. M. Staunton, D. Brown, *Inorg. Chim. Acta* **64** (1982) L 209 – L 210.
55. R. R. Schumaker, E. M. Engler, *J. Am. Chem. Soc.* **99** (1977) 5521 – 5522.
56. K. P. Callahan, P. J. Durand, *Inorg. Chem.* **19** (1980) 3211 – 3217.
57. R. D. Baechler, S. K. Daley, *Tetrahedron Lett.* 1978, no. 2, 101 – 104.
58. E. R. Squibb & Sons, Inc., US 4018761, 1977 (C. M. Cimarusti, P. Wojtkowski, J. E. Dolfini).
59. A. Levasseur, R. Olazcuaga, M. Kbala, M. Zahir et al., *C.R. Séances Acad. Sci. Ser. 2* **293** (1981) 563 – 565. *Chem. Abstr.* **96** (1982) 186016.
60. H. Wada, M. Menetrier, A. Levasseur, P. Hagenmuller, *Mater. Res. Bull.* **18** (1983) no. 2, 189 – 193.
61. S. Susman, L. Boehm, K. J. Volin, C. J. Delbecq, US Appl. 375525, 1983. *Chem. Abstr.* **100** (1984) 124156.
62. S. Susman, L. Boehm, K. J. Volin, C. J. Delbecq, *Solid State Ionics* **5** (1981) 667 – 669.

63. A. Stock: *Hydrides of Boron and Silicon,* Cornell University Press, Ithaca, New York 1933.
64. H. I. Schlesinger, A. B. Burg, *J. Am. Chem. Soc.* **53** (1931) 4321.
65. H. C. Brown: *Hydroboration,* W. A. Benjamin Inc., New York 1962.
66. W. N. Lipscomb, *Science* **196** (1977) 1047.
67. W. N. Lipscomb, *Adv. Inorg. Chem. Radiochem.* **1** (1959) 156.
68. *Pure Appl. Chem.* **30** (1972) 683. R. M. Adams, *Inorg. Chem.* **7** (1968) 1945. J. B. Casey, W. J. Evans, W. H. Powell, *Inorg. Chem.* **22** (1983) 2245.
69. T. P. Fehlner, R. L. Strong, *J. Phys. Chem.* **64** (1960) 1526.
70. R. F. Porter, F. A. Grimm, *Adv. Chem. Ser.* **72** (1969) 100. *Chem. Abstr.* **68** (1968) 110276.
71. M. D. Carabine, R. G. W. Norrish, *Proc. R. Soc. (London), Ser. A* **296** (1967) 1.
72. A. E. Newkirk, H. R. Broadly, A. L. Marshall, General Electric Co. Report No. 55248, March 20, 1950.
73. R. W. Parry, L. J. Edwards, *J. Am. Chem. Soc.* **81** (1959) 3560.
74. E. Muetterties (ed.): *Boron Hydride Chemistry,* Academic Press Inc., New York 1975, Chapter 3.
75. G. A. Guter, G. W. Schaeffer, *J. Am. Chem. Soc.* **78** (1956) 3546.
76. M. F. Hawthorne et al., *J. Am. Chem. Soc.* **82** (1960) 1825.
77. H. D. Johnson, II, S. G. Shore, *J. Am. Chem. Soc.* **93** (1971) 3798.
78. Callery Chemical Co., Diborane Handling Bulletin,1982.
79. A. Stock, C. Massenz, *Ber. Dtsch. Chem. Ges.* **45** (1912) 3539.
80. E. Wiberg, K. Schuster, *Ber. Dtsch. Chem. Ges.* **67 B** (1934) 1807.
81. V. I. Mikheeva, V. Y. Markina, *Zh. Neorg. Khim.* **1** (1956) 619.
82. M. J. Klein, B. C. Harrison, I. Solomon, *Abstr. 131st ACS Meeting,* Miami, Fla., April, 1957.
83. C. R. Dillard, Ph. D. Thesis, University of Chicago, 1949.
84. H. C. Brown, W. Korytnik, *J. Am. Chem. Soc.* **82** (1960) 3869.
85. S. Ikeganic, S. Yamada, *Chem. Pharm. Bull. (Tokyo)* **14** (1966) no. 12, 1399. *Chem. Abstr.* **66** (1967) 65377.
86. R. L. Johnson, D. J. Burton, *Tetrahedron Lett. 1965,* no. 46, 4039 – 4084.
87. H. Ikeda, A. Kogure, K. Shuna, Y. Minoura, *Kogyo Kagaku Zasshi* **68** (1965) 1107. *Chem. Abstr.* **66** (1965) 56346.
88. E. M. Fedneva, V. N. Konoplev, V. D. Krosnaperova, *Zh. Neorg. Khim.* **11** (1966) 2051. *Chem. Abstr.* **66** (1967) 28367.
89. R. M. Adams: *Boron Metallo-Boron Compounds and Boranes,* Interscience, New York 1964, p. 249.
90. G. W. Shaeffer, E. R. Anderson, *J. Am. Chem. Soc.* **71** (1949) 2150.
91. Work at Callery Chemical Co., Callery, Pa., 1975 – 1980.
92. K. Stallmann, H. Speckhardt, *Metall (Berlin)* **35** (1981) 979. *Chem. Abstr.* **96** (1982) 56402.
93. References cited in Amine Borane Product Brochure,Callery Chemical Co., Callery, Pa.
94. G. Andrews, T. Crawford, *Tetrahedron Lett.* **21** (1980) no. 8, 693.
95. G. Andrews, *Tetrahedron Lett.* **21** (1980) no. 8, 697.
96. S. Karody, J. Amoto, R. Reaner, L. Weinstock, *J. Am. Chem. Soc.* **103** (1981) 6765.
97. N. Umino, T. Iwakuma, W. Itoh, *Chem. Pharm. Bull. (Tokyo)* **21** (1979) 1479.
98. Y. Kikugawa, *Ventron Alembic* No. 29, Ventron Div., Morton-Thiokol Inc., March, 1983, pp. 1 – 4.
99. B. Nazar, Ph. D. Thesis, Drexel University, 1981. *Chem. Abstr.* **95** (1981) 5880.
100. U.S. Dept. of the Army, US 4368194, 1983 (B. F. Spielvogel, A. T. McPhail, I. H. Hall).
101. R. Ewens, H. Bassett, *Chem. Ind. (London)* **68** (Feb., 1949) 131.
102. A. P. Altschuler, *J. Am. Chem. Soc.* **77** (1955) 5455.
103. Metal Hydrides, US 2970114, 1961 (R. W. Bragdon).
104. Mine Safety Appliances, US 2744810, 1956 (C. B. Jackson).
105. Callery Chemical Co., GB 774728, 1957 (W. H. Schecter).
106. A. E. Finhold, A. C. Bond, Jr., H. I. Schlesinger, *J. Am. Chem. Soc.* **69** (1947) 1199.
107. Ventron Div., Morton-Thiokol Inc., DE 1218413, 1959;DE 2429521, 1974 (E. R. Winiarczyk).
108. M. M. Cook, E. A. Sullivan, R. C. Wade: *13th Northeast Regional ACS Meeting,* Hartford, Conn., June, 1983.
109. D. F. Gaines, R. Schaeffer, F. Tebbe, *Inorg. Chem.* **2** (1963) 526.
110. H. C. Miller, N. E. Miller, E. L. Muetterties, *J. Am. Chem. Soc.* **85** (1963) 3885. *Inorg. Chem.* **3** (1964) 1456.

111. G. S. Pauson, L. E. Weill, *J. Inorg. Nucl. Chem.* **15** (1960) 184.
112. R. O. Hutchins, N. R. Natale, *Org. Prep. Proced. Int.* **11** (1979) 246.
113. Ventron Div., Morton-Thiokol Inc., US 2942990, 1960 (E. A. Sullivan).
114. Ventron Div., Morton-Thiokol Inc., US 2930771, 1960 (R. C. Wade).
115. E. A. Sullivan: *Third International Meeting on Boron Chemistry,* Munich and Ettal (FRG), July 5 – 9, 1976, Pergamon Press, Oxford 1977, p. 132.
116. GAF Corp., US 3306886, 1967 (F. Grosser, E. V. Hart, A. Schwartz).
117. Union Carbide, US 3374121, 1968 (J. N. Hgsett).
118. Eastman Kodak, FR 1548122, 1968 (C. C. Bard, H. W. Vogt).
119. W. Boden, *J. Electrochem. Soc.* **129** (1982) 1252.
120. H. Hatanka, *J. Neurol.* **209** (1975) 81.
121. R. N. Grimes: *Carboranes,* Academic Press, New York 1970, p. 181.
122. H. Beall in E. L. Muetterties (ed.): *Boron Hydride Chemistry,* Academic Press, New York 1975, pp. 302 – 347.
123. M. F. Hawthorne, T. D. Andrews, P. M. Garrett, F. P. Olsen et al., *Inorg. Synth.* **10** (1967) 95.
124. T. Onak in E. L. Muetterties (ed.): *Boron Hydride Chemistry,* Academic Press, New York 1975, pp. 348 – 382.
125. G. B. Dunks, M. F. Hawthorne, *Acc. Chem. Res.* **6** (1973) 124.
126. R. N. Grimes (ed.): *Metal Interactions with Boron Clusters,* Plenum Press, New York 1982, Chapters 1, 4, 5.
127. W. Siebert in R. W. Parry, G. Kodama (ed.): *Boron Chemistry-4,* Pergamon Press, Oxford 1980, pp. 81 – 94.
128. M. B. Roller, J. K. Gillham, *Pol. Eng. Sci.* **14** (1974) 567.
129. T. R. Lynn, J. W. Churchill, T. B. Lynn: *Biblio-graphy for Dexsil Applications,* Dexsil Chemical Corp., Hamden, Conn., 1982.
130. H. R. Allcock, A. G. Scopelianos, J. P. O'Brien, M. Y. Bernheim, *J. Am. Chem. Soc.* **103** (1981) 350.
131. H. R. Allcock, A. G. Scopelianos, R. R. Whittle, N. M. Tollefson, *J. Am. Chem. Soc.* **105** (1983) 1316.
132. M. F. Hawthorne, *Transition Met. Chem. Proc. Workshop,* 1981, pp. 299–306. *Chem. Abstr.* **98** (1983) 89412.
133. M. S. Delaney, C. B. Knobler, M. F. Hawthorne et al., *Inorg. Chem.* **20** (1981) 1341.
134. University of California, Berkley, US 4363747, 1982 (M. F. Hawthorne, M. S. Delaney).
135. T. J. Klingen, B. M. Johnson, *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 619.
136. U.S. Dept. of the Army, US 3764417, 1973 (W. E. Hill, L. R. Beason).
137. U.S. Dept. of the Army, US 3980710; US 4025913, 1977 (D. C. Sayles).
138. T. Totani, K. Aono, K. Yamamoto, K. Tawara, *J. Med. Chem.* **24** (1981) 1492.
139. W. Kliegel: *Bor in Biologie, Medizin und Pharmazie,* Springer Verlag, Berlin 1980, pp. 578 – 609.
140. E. Mizusawa, H. L. Dahlman, S. J. Bennett, M. F. Hawthorne et al., *Proc. Natl. Acad. Sci. USA* **79** (1982) 3011. *Chem. Abstr.* **97** (1982) 21838.
141. F. Sweet, *Steroids* **37** (1981) 223.
142. F. Sweet: *Abstr. First Int. Symposium on NCT,* Cambridge, Mass., Oct., 1983, paper VI-3.
143. Procter & Gamble Co., EP-A 82200784.5, 1982 (J. J. Benedict).
144. N. Hagehara, M. Kumada, R. Okawara (eds.): *Handbook of Organometallic Compounds,* W. A. Benjamin, Inc., New York 1968.
145. Brochure on Trialkyl Boranes, Callery Chemical Co., Callery, Pa., 1978.
146. H. C. Brown, S. W. Kulkarni, *J. Organomet. Chem.* **239** (1982) 41.
147. Texas Alkyls, Inc., US 3719695, 1973 (S. H. Eidt).
148. Du Pont, US 4045495, 1977 (N. Nazarenko, W. C. Seidel).
149. H. C. Brown, S. H. Kulkarni, *J. Organomet. Chem.* **218** (1981) 307.
150. P. J. Maddocks, A. Pelter, K. Rove, K. Smith et al., *J. Chem. Soc., Perkin Trans. I* 1981, 656.
151. H. C. Brown, P. K. Jadhav, A. K. Mandal, *Tetrahedron* **37** (1981) 3587.
152. A. Akimoto, *J. Appl. Polym. Sci.* **17** (1973) 2262.
153. Union Oil Co. of Calif., US 3663490, 1972 (A. M. Sarem).
154. K. Kojima, S. Iwabushi, E. Asami, Y. Iwata, *Chiba Daigaku Kogakubu Kenkyu Hokoku* **22** (1971) 55. *Chem. Abstr.* **77** (1972) 35059.
155. G. Sasnowsky, J. M. Brown, *Chem. Rev.* **66** (1966) 540.
156. P. B. Brindley, R. G. Pearson, *J. Polym. Sci.* (Part B), Polym. Lett. Ed. 6 (1968) no. 12, 831.
157. Mochida Pharmaceutical Co. Ltd., DE 2321215, 1973 (E. Masuhara, N. Tarumi, N. Nakabayashi, M. Baba et al.).
158. G-C Chemical Manuf. Co., Ltd., US 3527737, 1970.

159. H. C. Brown: *164th ACS Meeting,* Div. of Petro-leum Chem., New York, 1972.
160. R. T. Hawkins, W. J. Lennarz, H. R. Snyder, *J. Am. Chem. Soc.* **82** (1960) 3059.
161. H. C. Brown, P. K. Jadhav, M. C. Disai, *J. Am. Chem. Soc.* **104** (1982) 4304.
162. D. S. Matteson, K. M. Sadler, *J. Am. Chem. Soc.* **103** (1981) 5242.
163. J. A. Settepani et al., *J. Med. Chem.* **13** (1970) 128.
164. D. T. Haworth, L. F. Hohnstedt, *Chem. Ind. (London)* **20** (May 14, 1960) 559.
165. L. F. Hohnstedt, D. T. Haworth, *J. Am. Chem. Soc.* **82** (1960) 89.
166. Mine Safety Appliances, Inc., US 4150097, 1973 (W. V. Hough, C. R. Guibert, G. T. Hefferan).
167. K. Niedenzu, J. W. Dawson: *Boron Nitrogen Compounds,* Academic Press, New York 1965, pp. 85 – 126.
168. A. C. Adams, *J. Electrochem. Soc.* **128** (1981) 1378.
169. S. Shanfield, R. Wolfson, *J. Vac. Sci. Technol. A* **1** (1983) 324.
170. G. Constant, R. Feurer, *J. Less Common Met.* **82** (1981) 113.
171. State Scientific Research and Design Institute for Polymeric Adhesives, SU 519430, 1974 (K. N. Karadzhyan, G. A. Kararyan). *Chem. Abstr.* **85** (1976) 109524.
172. Sun Oil Co., DE 1236789, 1967 (J. L. Jezl, H. M. Khelghatian, L. Dentler).
173. Texas Instruments, Inc., US 3564565, 1971 (R. Haberecht, R. Patterson).
174. J. J. Ebelman, M. Bouquet, *Ann. Chim. Phys.* **17** (1846) 54 – 73. *Ann.* **60** (1846) 251 – 253.
175. "The Nomenclature of Boron Compounds," *Inorg. Chem.* **7** (1968) 1945 – 1964.
176. H. Steinberg: "Boron-Oxygen and Boron-Sulfur Compounds," *Organoboron Chemistry,* vol. 1, Interscience, New York 1964.
177. D. B. Green: "Boron, Part A," in J. W. Mellor: *A Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol. 5, Supplement 1, Longman, London 1980, pp. 703 – 720.
178. P. M. Christopher, W. H. Washington, *J. Chem. Eng. Data* **14** (1969) 437 – 439.
179. H. Steinberg, D. L. Hunter, *Ind. Eng. Chem.* **49** (Feb., 1957) 174 – 181.
180. W. Kliegel: *Bor in Biologie, Medizin und Pharmazie,* Springer Verlag, Berlin 1980, pp. 636 – 803.
181. G. D. Clayton, F. E. Clayton (ed.): *Patty's Industrial Hygiene and Toxicology,* J. Wiley & Sons, New York 1981, pp. 2978 – 3005.
182. R. L. Tatken, R. J. Lewis, Sr.: *Registry of Toxic Effects of Chemical Substances,* **vols. 1, 2, 3,** U.S. Dept. of Health and Human Services, NIOSH Publ. No. 83 – 107, 1981 – 1982.
183. C. C. Comstock, L. Feinsilver, L. H. Lawson, F. W. Oberst, Army Chemical Corps Medical Laboratories, Research Report No. 258, 1954.
184. ACGIH (ed.): *Threshold Limit Values (TLV),* ACGIH, Cincinnati, Ohio, 1985.
185. DFG (ed): *Maximale Arbeitsplatzkonzentrationen,* Verlag Chemie, Weinheim 1984.
186. E. M. Cordasco, R. W. Cooper, J. V. Murphy, C. Anderson, *Chest* **41** (1962) 68.
187. H. J. Lowe, G. Freeman, *AMA Arch. Ind. Health* **16** (1957) 523.
188. E. H. Krackow, *AMA Arch. of Ind. Hyg. Occup. Med.* **8** (1953) 335.
189. R. M. Adams: *Boron, Metallo-Boron Compounds and Boranes,* Interscience, p. 731. New York 1964.
190. I. H. Hall, C. O. Starnes et al., *J. Pharm. Sci.* **69** (1980) 1025.
191. C. T. Blaisdell, Army Chemical Corps Medical Laboratories, Research Report no. 351, 1955.
192. W. E. Reinhart, *Am. Ind. Hyg. Assoc. J.* **21** (1960) 389.
193. R. A. Spryshkova, L. I. Karaseva, V. A. Bratsev, N. G. Serebryakov, *Med. Radiol. (Mosk.)* **26** (1981) 62. *Chem. Abstr.* **95** (1981) 108411.

Boron Nitride → Boron Carbide, Boron Nitride, and Metal Borides
Boron Trifluoride → Boron Compounds
Brake Fluids → Hydraulic Fluids
Brandy → Spirits
Brass → Copper Alloys
Brazing Alloys → Soldering and Brazing

Bread and Other Baked Products

DIETMAR SIEVERT, Nestlé Product Technology Centre Orbe, Orbe, Switzerland

R. CARL HOSENEY, R& R Research Services, Inc., Manhattan, KS, United States

JAN A. DELCOUR, Laboratory of Food Chemistry, Katholieke Universiteit Leuven, Heverlee, Belgium

1. Introduction . . . 2
2. Wheat Types and their Uses . . . 2
3. Wheat Breeding and Biotechnology . . . 4
3.1. Traditional Breeding . . . 4
3.2. Biotechnology . . . 4
4. Milling of Wheat . . . 4
5. Wheat Flour . . . 5
5.1. Flour Constituents . . . 5
5.1.1. Starch . . . 5
5.1.2. Protein . . . 7
5.1.3. Lipids . . . 8
5.1.4. Nonstarch Polysaccharides (NSP) 8
5.1.5. Ash . . . 8
5.2. Chemical Flour Treatments . . . 8
6. Evaluation of Flours . . . 9
6.1. Physical Dough Tests . . . 9
6.1.1. Mixing Curves . . . 10
6.1.2. Flour Strength . . . 11
6.2. α-Amylase Activity . . . 11
7. Dough Formation . . . 11
7.1. Underlying Mechanisms of Dough Formation . . . 11
7.2. Role of Proteins in Dough Formation . . . 12
7.3. Thermodynamic View of Dough Mixing . . . 14
7.4. Aeration of Dough during Mixing . . . 14
8. Role of Bread Ingredients . . . 14
9. Bread and Dough Making Processes . . . 17
9.1. Straight Dough . . . 17
9.2. Sponge and Dough . . . 17
9.3. Liquid Preferments . . . 17
9.4. No-Time Processes . . . 18
9.5. Sour Dough . . . 18
9.6. Frozen Dough . . . 19
10. Gas Production and Retention . . . 20
10.1. Leavening Mechanisms . . . 20
10.2. Fermentation and Gas Production by Yeast . . . 20
10.3. Gas Production by Chemical Leavening . . . 21
10.4. Gas Retention . . . 23
11. Molding and Proofing . . . 23
12. Baking . . . 23
12.1. Transformation of Dough into Bread . . . 23
12.2. Crumb Grain Formation . . . 27
12.3. Crust Formation . . . 27
12.4. Changes of Flour Constituents . . . 28
13. Flavor of Baked Products . . . 28
14. Glass Transition and its Role in Baking . . . 29
15. Bread Varieties and Speciality Breads . . . 30
16. Soft Wheat Products . . . 31
16.1. Technology of Cookie and Cracker Production . . . 32
16.2. Cookies . . . 33
16.2.1. Cookie Types . . . 33
16.2.2. Underlying Mechanisms of Cookie Baking . . . 34
16.2.2.1. Texture Development of Cookies . . . 36
16.2.2.2. Development of the Cookie Surface Pattern . . . 37
16.3. Crackers . . . 37
16.4. Wafers . . . 38
16.5. American Biscuits . . . 38
16.6. Cakes . . . 39
16.6.1. Cake Making and Underlying Mechanisms of Cake Baking . . . 39
16.6.2. Role of Cake Ingredients . . . 40
16.7. Pastries . . . 42
16.7.1. Short Pastry . . . 42
16.7.2. Puff Pastry . . . 42
17. Retention of Baked Product Quality . . . 44
17.1. Staling . . . 44
17.2. Microbial Spoilage . . . 46
17.2.1. Retention of Quality . . . 46
17.2.2. Chemical Preservatives . . . 47
18. Trends in Baking . . . 47

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_331.pub2

18.1. U.S. Bakery Market 47
18.2. Western European Bakery Market 48
18.3. Consumer Demands and Product Trends 49
19. References 49

1. Introduction

Cereal grains have been the mainstays of human nutrition throughout history [1 – 4]. They are nutritious, palatable, and easy to store. It is easy to picture pre-historic man eating unmodified grains. Because of the hardness of the dried grains and the subsequent difficulty in chewing the grain, it would appear to be a small step to decide to soak the grains before eating them. Crushing the grain after it had been soaked for a time lead directly to production of gruel. Ubiquitous wild yeast and bacteria will inoculate a gruel setting over night. Placing a stiff gruel in hot ashes or on a hot rock results in a solid material with a much different taste than that of the starting gruel.

How the above events unfolded can only be the subject of conjecture as it occurred before recorded history. Evidence suggests that by the end of the Stone Age man was consuming flat baked products. All grains can be used to produce gruels and all gruels can be baked to produce flat unleavened bread. The step to lighter leavened bread required the use of wheat or, to a much lesser extent, rye. This is still true today.

The next step necessary for the production of baked products was the control of the sour dough fermentation and development of baking ovens. The oldest reported ovens were in ancient Babylonia. The ancient Egyptians had modern-looking ovens as early as 2700 B.C. and by 1750 B.C. and were producing bread leavened chemically and with yeast. Evidence suggests that sour dough bread was known in Europe by the late Bronze Age (800 B.C.) and was known throughout the known World by 100 B.C. During the next few centuries there were gradual improvements in oven design. The next large step came with the development of dough-mixing machines in the second half of the 18th century (the industrial revolution). The production of bakers' yeast in the 19th century resulted in a much shorter production schedule and one that could be automated into large plants.

2. Wheat Types and their Uses

Wheat is the most important cereal grain produced and traded in the world for the production of bread and other baked products. Wheat is a member of the grass family of plants. It is cultivated on more area of the earth's surface than any other crop and is man's major source of food. The world wheat production for 2005 is estimated at 626×10^6 t. This is slightly higher than the previous five-year average of 588×10^6 t. It is estimated that nearly two-thirds of the total wheat production is used for food; the remaining one-third is used for seed, feed, and non-food applications. Wheat is grown from as far north as the Arctic Circle in Finland and near the equator in parts of Africa. However, the major production is in the temperate zones. Major areas of production include: the Great Plains of the United States and Canada, Western and Eastern Europe, Russian Federation, Argentina, Australia, China, and the Punjab of India. There are three species of wheat grown commercially *Triticum aestivum* (common wheat), *T. durum* (durum wheat) and *T. compactum* (club wheat).

Wheat can be of either spring, sometimes called summer, or winter habit. Spring wheat is planted in the early spring as soon as the soil starts to warm and is harvested in the late summer. Winter wheat is planted in the fall. Harvest in the United States is in the early summer. In Europe, where the summers are much cooler and wetter, harvest does not occur until late in summer. In general, there is a negative relationship between yield and protein content of the grain. Thus, spring wheat with its lower yield will generally have higher protein content.

Wheats can vary in their hardness, i.e., the force required for crushing the kernels varies. In North America and in certain other parts of the world, the common (*T. aestivum*) wheats are classified as either hard or soft, and the *T. durum* wheats are referred to as durum wheats. In Europe and many other parts of the world, the common wheats are called soft and the durum

wheats are called hard. This has led to considerable confusion. When a European tells an American that they use soft wheat for bread making the American has the wrong idea. However, he or she is no more confused than the European when the American says we make our bread from hard wheat (durum in the mind of the European). The hardness of wheat is under genetic control [6]. Thus, crossing two hard wheats will give progeny that are also hard. Likewise crossing two soft wheats will lead to progeny that are soft. In addition to the genetic effect, the environment also has a major effect upon the hardness of the kernels. The measurement of hardness of wheat is much more difficult than it might appear at first. Wheat kernels have various sizes and shapes. They also have a pronounced crease. Those factors make it difficult to define the geometry of the test piece that is being crushed. Wheat kernel hardness can either be measured on individual kernels or on the bulk sample. An instrument which measures the individual hardness of 300 kernels and gives a distribution of the force required to crush the individual kernels is available. The technique is useful if mixtures of varying hardness are being measured. The most common way of measuring the hardness of a bulk sample is to grind the sample with a burr type mill and sift the resulting meal. Hard wheats will give a much coarser meal than will soft wheats. The reasons for differences in wheat endosperm hardness are not clear. However, major classes of *T. aestivum* and *T. durum* wheat hardness (texture) seem to be attributable to the presence, absence or sequence polymorphism of the proteins puroindolin a and puroindolin b [5].

Wheats also vary in their color. Most wheats are either red or white. The red color comes from a pigment in one layer of the pericarp cells. In white wheats this pigment is missing. Thus, white wheats are not really white but only nonpigmented. Purple and black wheats are known but they appear not to be grown commercially except in parts of China. The absence or presence of the pigmented layer is under genetic control and, thus, plant breeders can readily change the color of the wheat. From a quality standpoint, color is of relatively little importance. It can be important in products that have a very light color, for example, in some cookies, pasta, or wafer products where red bran specs are readily seen. Breakfast food manufacturer's also like white wheats for producing bran products. The bran from white wheat does not brown as readily as that from red wheat and gives more appealing products.

Common (*T. aestivum*) wheats are hexaploid and resulted from the prehistoric combination of three grass species each having 14 chromosomes. Common wheat thus has 42 chromosomes. Wheat then is made up of three genomes called A, B, and D. Durum wheat is tetraploid and, thus, has 28 chromosomes. It has the A and B genomes but not the D one. Since the gene controlling the hardness of common wheats is on the D genome, it is obviously missing in durum wheats. Perhaps as a result the durum wheats are very hard, much harder than the hard common wheats.

The durum (*T. durum*) wheats that are commercially grown are mostly spring wheats. They are white wheats, containing no pigments in the pericarp. Because of the hard translucent texture of the endosperm they appear amber in color and not chalky white such as found with soft wheats. Red durums are known and are used for feed. Winter durums are also known but not grown to any extent. The last species of wheat is the club wheat (*T. compactum*). In general, club wheats are very soft and are used mainly for cookie and cake flours in North America.

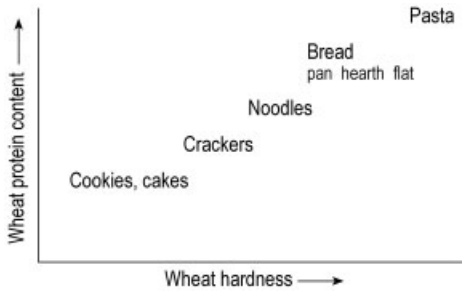

Figure 1. Illustration of wheat hardness and wheat protein content in relation to baked product type.

Wheat is used in countless food products. They include, for example, fermented (leavened) breads, bread specialties, cookies, crackers, cakes, pastries, breakfast cereals, pasta, noodles, and starch/gluten [7 – 10]. The overall use of wheat in relation to wheat hardness and wheat protein content is illustrated in Figure 1. Apart from product and process requirements (wheat

quality or suitability), the use of wheat is determined by factors such as availability, price, contracts and relationships with a supplier (wheat producer, trader, miller), geographic location of the wheat user and the supplier, and regulations (e.g., tariffs) that control the use of specific wheats. For the production of bioethanol from wheat starch see → Cereals, Cereal Products.

3. Wheat Breeding and Biotechnology

The primary efforts of cereal breeding are directed towards (1) increase in grain yield, (2) assembling resistance or tolerance to fungi, bacteria, insects, viruses etc, and (3) improvement of grain quality attributes (e.g., bread making potential) [11 – 13].

3.1. Traditional Breeding

Common wheat is a complicated cereal. It has 42 pairs of chromosomes compared to 14 for maize or rye. This also means it has three sets of genes for each function in the plant. However, all three of the genes may not be active. This complexity makes it difficult to obtain any desired function by conventional breeding. A good example of this is the production of waxy wheat (a wheat in which the starch is composed of amylopectin only with essentially no amylose) [14]. There are three genes controlling the production of amylose. To obtain a waxy variety all three of these genes must be null (not active). The probability of this occurring by chance is very small. It was only when the gene products (proteins) were identified that crosses between lines containing the various mutant genes resulted in waxy wheat.

3.2. Biotechnology

Rapid developments in plant biotechnology have led to the ability to transform or insert new DNA into plants [15 – 20]. Early products of genetic engineering technologies include plants with improved disease protection such as enhanced insect resistance and improved herbicide tolerance. As the world population continues to grow at a rapid rate the problem of producing sufficient food becomes more severe. To meet this challenge, a sustainable advancement in food production must be developed. Plant biotechnology, appears to be an answer to obtaining the sustainable increase in food production.

Indeed it can increase harvestable yield and improve quality of end-products. Specific quality targets of genetic engineered wheat have included the modification of gluten protein fractions, especially the high molecular weight glutenin subunits of wheat glutenin for improved bread making quality [21]. The starch content and structure could be modified, for example, to optimize the noodle-making quality of wheat flour or to extend the shelf life of baked products [22, 23]. Moreover, the nutritional value of wheat may be improved by altering the amino acid profile, e.g., by increasing the essential amino acid lysine. However, it seems likely it may take several years before new wheat improved through biotechnology is commercially available to wheat growers, millers, and food processors.

Consumer acceptance of foods obtained through biotechnology remains a very important issue [24]. An area where wheat improvement through biotechnology, such as wheat of increased starch content, may become more rapidly acceptable, is in the use as a raw material for bioethanol production.

4. Milling of Wheat

The wheat kernel is composed of three main morphological parts: the endosperm comprising starch and storage proteins (ca. 82 %), the bran (ca. 15.5 %) rich in dietary fiber, B vitamins, minerals and metabolizing proteins, and the germ (ca. 2.5 %) which is relatively high in oil. The first objective of milling is to separate these anatomical parts of the wheat kernel as cleanly as possible. This separation is achieved, for the most part, in the break system of the mill. The next objective is to transform the endosperm into materials of smaller particle size, i.e., flours. This is accomplished by the reduction system. The flour used for the production of baked products is essentially the endosperm of the wheat that is reduced to flour composed of a

range of particles, from very small (< 1 μm) up to about 200 μm. Soft wheat (North American terminology) produces a flour with much smaller average particle size than hard wheat (≈40 μm vs. 75 μm, same terminology). In hard wheat, the protein and starch tend to remain together in the flour and cannot be separated without additional grinding to reduce the particle size. As a result of milling cereals can be further processed into more palatable and desirable food. Milling, however, may also decrease the nutritional value by the removal of vitamins, minerals, and dietary fiber present in bran and germ. In recent years there has been a trend towards using whole wheat flours and meals comprising all parts of the wheat kernel for the production of highly nutritional baked products.

Flour Types. Many different flours are produced by the milling operation, one from each grinding step. Flour extraction rates are used to compare yields of flour and are expressed as a percentage of the production of milled products. Flour yields typically vary from 72 % for white flour to 100 % for whole-meal flour. Curves of ash and/or protein extraction (Fig. 2) are useful in determining the maximum yield that can be obtained and still meet the ash and protein specifications of the flour. A composite of all the flours produced by the mill is *straight grade flour*. Depending upon the details of the process it represents about 72 % of the total products. The flour produced at the head (first grinding steps) of the reduction system is *patent flour*.The term short patent is used for flour with the lowest ash content and represents about 45 % of the total products. This flour comes mainly from the center of the wheat kernel. A long patent flour may be up to 65 % of the total products. The term *cut-off flour* is used for the 20 % between the short and long patents. The 7 % between the long patent and the straight-grade flour is *low-grade flour* (also denoted as clear flour). This flour, which comes from the tail end (last grinding steps) of the break and reduction systems, has high ash and dark color. Another product from the mill is millfeed, i.e., bran and shorts; bran represents about 11 % of total products and shorts about 16 %. Germ can be an additional product, usually recovered at about 1.0 %. Developments in milling engineering such as improved process performance, complete process automation and optimization of product control have made flour milling a highly efficient and cost-effective operation [25].

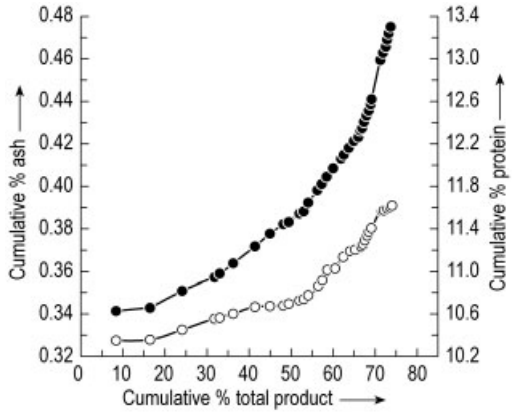

Figure 2. Curves of ash extraction (●–●) and protein extraction (o–o) for a commercial wheat flour mill.

5. Wheat Flour

5.1. Flour Constituents

Wheat flour constituents, specifically starch, proteins, lipids and nonstarch polysaccharides have been extensively studied [7, 8, 26 – 30].

5.1.1. Starch (→ Starch)

Starch is the most abundant component in wheat flour. The level varies but will average nearly 70 % on a dry weight basis. In all cereal grains the plants excess energy is stored as starch. The molecular composition and structure of starch has been examined in relation to starch properties [31 – 35]. Starch occurs only in the endosperm of the seed as small microscopic particles or granules. In wheat, the granules are predominately of two size ranges, i.e., large lenticular (lens shaped) and smaller spherical granules. The large granules average about 25 μm in the large dimension and about 10 μm in the other dimension. The smaller spherical granules average about 10 μm in diameter. The starch granules are tightly packed (density about 1.6 g/L) and partially crystalline. In excess water they absorb about 30 % of their weight and swell about 5 %.

The starch granules are composed of two classes of starch polymers: amylose and amylopectin. Amylose makes up about 25 % of the starch, the remaining 75 % is amylopectin. Amylose is a mostly linear polymer of anhydroglucose units linked α-1 $\rightarrow$ 4. In wheat the amylose polymer is composed of about 1000 to 2000 anhydroglucose units (degree of polymerization, DP 1000 to 2000) and has a molecular mass of about 160 000 to about 320 000. There is a small level ($<$ 1 %) of branching through α-1 $\rightarrow$ 6 bonds. Although amylose is slightly branched, it appears to act like a truly linear polymer.

Amylopectin molecules are very large (DP typically over 600 000) and highly branched polymers. Amylopectin is also predominately composed of α-1 $\rightarrow$ 4 bonds with typically about 5 % α-1 $\rightarrow$ 6 bonds as the branching points. The amylopectin molecules are responsible for the crystallinity in the granule. The crystallites are very small being about 50 $\times$ 200 nm.

Within the granule both the amylose and amylopectin appear to be ordered at right angles to the surface of the granule. This ordering gives the granule the property of birefringence.

When starch is heated in excess water at about 55 °C the crystallites start to melt, the granule swell, and the birefringence is lost over about a 7 °C temperature range (gelatinization). However, the loss of birefringence in individual granules occurs over a much smaller temperature range, showing that gelatinization is a granule per granule event. Gelatinization is essential in most baked products [31, 36 – 43]. The amylose content and the amylopectin branch chain-length distributions predominantly affect the gelatinization and pasting/gelation properties of the starch [44 – 46].

In processing of most baked products the water is limiting. In limited water systems, the temperature at which the starch granules start to melt remains the same, as long as the starch contains more than 30 % water. However, the temperature range over which the crystallites melt becomes much wider [47]. Thus, it is possible that the crystallites are not completely melted and birefringence is not completely lost even at 100 °C. Thus, starch occurs as two phases (soluble polymers and remnants of granules) in most baked products. The ratio of soluble polymers to remnants of granules depends on the water content of the system. Also important is the level of sugars in food systems. High levels of sugars will increase the melting point of the starch crystallites, thereby controlling the gelatinization temperature of starch [48, 49].

Upon cooling and storage, reassociation of the starch chains occurs [40, 50 – 58]. In this context, the term retrogradation is frequently used. It refers to the starch "going back" to its initial crystallinity. In the ungelatinized starch amylopectin is partially crystalline, and upon cooling and storing gelatinized starch both amylose and amylopectin show crystallinity, thus one should only refer to amylopectin crystallization as retrogradation and refer to amylose crystallization as such. The minimum chain length for chain association involving double helix formation and crystallization is estimated to be DP 10. By reducing the chain length through partial β-amylolysis, the retrogradation of amylopectin can be inhibited [59]. The retrogradation of starch is also affected by the presence of sugars, salt and lipids [49, 60 – 65].

Damaged Starch. Starch damage occurs as a result of roller milling when the starch granules are subjected to some shear. Such shear can disrupt the crystallites within the starch granule. As a result of crystallite disruption the starch granules will swell in water at room temperature [31, 66] and will be readily hydrolyzed by α-amylase.

The amount of shear that the granule is subjected to depends on a number of factors. The most important is probably the hardness of the wheat kernel. The harder the wheat the more shear is encountered during milling and, therefore, the more damaged starch is produced. There are both advantages and disadvantages of producing damaged starch during milling. In the case of hard wheat that is being milled into bread flour the damaged starch is an advantage. It increases the water absorption of the flour and, thus, gives a higher yield of dough. More dough is directly related to more bread. Thus, from the same amount of flour the baker obtains a higher amount of baked product.

In bread making the damaged starch swells and starts to solubilize during dough mixing. Starch swelling and solubilization potentially together with starch–gluten interactions increase the strength of the dough. Its greater strength makes it more difficult for the dough to be ex-

panded by the gases produced during fermentation. Therefore, the damaged starch that was beneficial at mixing, to increase water absorption, is detrimental at the end of fermentation. Addition of α-amylase, either as malt or amylase originating from microorganisms, is common at the mill. It has been suggested that amylase is added to produce sugars for the yeast. However, this does not appear logical as a large excess of sugar is added in many bread making formulas. The real reason these enzymes are added is to degrade the damaged starch, which is a readily accessible substrate for amylases. The destruction of the damaged starch is accompanied with a decrease in dough strength and allows the dough to expand much easier during proofing and baking. The result is the increased baked product volume found with amylase supplementation.

In many products traditionally made from soft wheat flours, such as most cookies, damaged starch is a strong detrimental factor. For cookies to spread in the oven during baking, the flour must be very low in damaged starch.

5.1.2. Protein

The protein content of wheat varies over a rather wide range (6 to 18 %) depending on both environmental and genetic factors. In general, as the protein content of wheat increases the starch content will decline. Wheat flour generally contains 1 % less protein than the wheat it is milled form. Thus, to produce a flour of 12 % protein wheat grains of about 13 % protein must be used as starting material. According to the classical Osborne fractionation procedure, cereal grain proteins can be classified into four groups on the basis of their extraction and solubility in a series of solvents: water (albumins), dilute aqueous salt solutions (globulins), alcohol–water mixtures (prolamins) and dilute acids or alkalis (glutelins) [67, 68]. In wheat flour about 15 % of the protein is soluble in water or aqueous salt solutions (albumins or globulins) and the remaining 85 % is storage protein (prolamins and glutelins). The storage proteins in wheat are forming the gluten [69]. Hence, wheat gluten is made up of two classes of proteins, prolamins called gliadin and glutelins called glutenin. It is the gluten protein that is largely responsible for wheat flour producing the viscoelastic dough that retains gas during fermentation resulting in the leavened loaf of bread [70 – 79]. Although there is some variation, in general, the gliadin makes up about 52 % of the gluten and the glutenin the remaining 48 %. Both of these two classes are complicated mixtures of individual proteins.

The *gliadin proteins* are relatively small molecules with an average molecular mass of about 44 000. More than 50 individual proteins have been identified in gliadin, although only 30 or so have been found in any one wheat line. The gliadin proteins consist of single polypeptide chains and, thus, all disulfide bonds would be internal (i.e., intermolecular disulfide bonding). Whereas a mixture of hydrated gliadin is viscous, sticky, and has little elastic properties, hydrated glutenin has strong elastic properties. The *glutenin proteins* are much larger in size than gliadins with an average molecular mass reported to be about 3×10^6. Some of the larger glutenin molecules have been reported to be as large as 20×10^6. The proteins are multichained and consist of many subunits. There are both low molecular mass and high molecular mass subunits. The molecular mass of the former varies from 30 000 to 60 000 and that of the latter from about 65 000 to 90 000.The subunits are linked together by disulfide bonds. Since the repeating unit in the interior of the subunits contains essentially no sulfur-containing amino acids, the disulfide bonds both intra- and intermolecular occur at the ends of the subunits. Apparently only one intramolecular bond exists between any two subunits in dough. Thus, glutenin is polymerized but not crosslinked. It appears that glutenin becomes crosslinked during baking and determines together with the embedded starch the rheological properties of bread crumb [80].

The amino acid composition of gluten at least in part explains its properties. The most prominent amino acid in gluten is glutamine, which makes up about 35 % of the total gluten. Glutamine in the polypeptide chain does not take on a charge, is quite hydrophilic, and easily forms hydrogen bonds. The second most prominent amino acid is proline (14 %). Proline contains a pyrrolidine ring structure and, thus produces a kink in the protein chain. This keeps the protein from forming a helix or similar structure. The level of hydrophobic amino acids in gluten is about 34 %. Only about 28 % hydrophobic

groups can be buried in the center of the protein, thus some of the hydrophobic groups are on the outside of the molecule and allow hydrophobic interactions. The low levels of positively charged amino acids (resulting from the high levels of glutamine and proline) and the absence of negatively charged amino acids result in a protein with a very low charge density. This means the polymers are only slightly mutually repelled and thus can readily interact with each other.

5.1.3. Lipids

Wheat, like all cereal grains, is low in fat or lipid material (ca. 2.0 %). The lipids are not uniformly distributed throughout the kernel. The germ contains the highest level of lipids (ca. 28 %) and the endosperm the lowest (ca. 1.5 %). The lipids in the endosperm occur in small lipid bodies. Flour lipids are a complicated mixture of compounds [81 – 83]. Of the 1.5 % lipids found in flour about one-half are nonpolar lipids (triacyl glycerides, fatty acids, sterols and their esters) and the other half are strongly polar (phospholipids and glycolipids). The polar lipids have strong surfactant properties and play a functional role in the production of baked products [84 – 86].

5.1.4. Nonstarch Polysaccharides (NSP)

The NSP in wheat flour make up the endosperm cell walls. In wheat flour apparently three classes of compounds occur: arabinoxylans, β-glucans with different linkage point, and a glycoprotein with the carbohydrate moiety being an arabinogalactan. All these compounds are dietary fiber constituents. A wide variation in the sizes and chemical structures is found within each of the classes [87 – 94]. While in older literature, the main NSP determining wheat functionality are referred to as pentosans, they are now more commonly and more correctly referred to as arabinoxylans.

Both water-extractable and water unextractable arabinoxylans are found in cereal grains. Water-extractable arabinoxylans are present in about equal levels (typically 0.5 %) in endosperm and bran. The content of unextractable arabinoxylans is much higher in the bran than in the endosperm. Therefore, whereas the level of water-extractable arabinoxylans is relatively independent of the flour extraction rate (ca. 0.5 %), the level of unextractable arabinoxylans in the milled product increases with the rate of flour extraction (ca. 2.0–8 % in milled products) [95]. The water-extractable arabinoxylans impart viscosity to dough, and their water-unextractable counterpart have a high water holding capacity.

Both the water-unextractable [96] and the water-extractable arabinoxylans [97] have been shown to interact with gluten [98]. The water-extractable arabinoxylans stabilize protein foams [92], affect gas retention and undergo oxidative gelation [99 – 102]. The gelation may be an important strengthening mechanism for dough during fermentation [103] The role of arabinoxylans and endoxylanases in bread making is reviewed in [104].

5.1.5. Ash

Typical wheat grains have an ash (mineral) content that ranges from 1.5 to 2.0 %. However, the ash is not distributed uniformly in the grain. The inner endosperm has a relatively low ash content (0.3 %), whereas the bran is higher in ash and may contain as much as 6 %. Thus, the ash content is a convenient and sensitive assay for the presence of bran in flour, as its determination is relatively simple and reproducible. However, wheats vary in the level of ash natively found in the endosperm and, thus, small variations in ash content do not necessarily imply the presence of different concentrations of bran in flours. In general, the ash itself does not affect flour properties, and, thus, it can be argued that the ash content alone has no meaning. However, high ash content indicates the presence of bran in the flour and bran contains compounds that generally are detrimental to bread making performance.

5.2. Chemical Flour Treatments

Flour may receive a number of treatments at the mill. Bleaching is very common in some countries. For bread and all-purpose flours, the most common bleaching agent is dibenzoyl peroxide, which is added as a dry powder and whitens flour

over a two-day period. It only bleaches flour pigments and has no effect on the baking properties. In countries where it is allowed, like the USA, cake flour is treated with chlorine gas. The gas destroys (bleaches) the pigments and gives very white flour. It also strongly affects the flour's functional properties.

Application of chlorine is detrimental to bread flours but beneficial to cake flours [105]. High-ratio cakes (cakes that contain more sugar than flour) cannot be made without treatment with chlorine gas (see Section 16.6.2). In countries where chlorine bleaching is not allowed, no high ratio cakes are produced. Chlorine reacts with almost all flour constituents thereby affecting their functional properties. The rate of reaction is quite rapid with flour and much slower with starch. However, fractionation and reconstitution studies have shown that the reaction with the starch is responsible for the improvement in baking properties. The amount of chlorine gas used is critical. When chlorine reacts with organic material, hydrochloric acid is usually produced. Thus, the pH of the flour is, generally, a good measure of the extent of reaction and therefore of the amount of chlorine added. Generally, an end pH of 4.7–4.9 is desired for most high-ratio cakes. The actual effect of chlorine appears to be its reaction with the starch to give a modified (oxidized) starch, which can swell to a greater extent than untreated starch. This gives a more viscous batter at the same temperature than with unoxidized flour. The increased viscosity of the batter keeps the cake from collapsing in the oven and, to some extent, after it is removed from the oven. Thus, with cake flours chlorine is both bleach and improving agent. Mills in the UK have produced special cake and biscuit flours from special low protein wheat varieties that are chlorinated to give a pH 4.5–5.5. However, chlorination is no longer allowed under EC rules and a substitute for chlorination has yet to be identified [106].

Bread flour also undergoes aging or maturing and concomitant improvement in baking performance with storage [107, 108]. The presumption has been made that aging is an oxidative process. If flour is stored long enough after milling, it slowly bleaches, supposedly because of air oxidation. Interestingly, high ratio cakes can be produced from nontreated cake flour if it is stored for about eight months. Reports suggest that also other mechanisms are involved in the aging of wheat and wheat flour. For example, the break down of the lipid bodies in the endosperm has been related to the aging phenomenon [109, 110]. In an attempt to speed up the aging process a number of oxidizing agents are added at the mill (azodicarbonamide, acetone peroxide, chlorine dioxide, and potassium bromate). These oxidants improve the bread-making potential of the flour, but do not replace aging.

6. Evaluation of Flours

Flour is the most important ingredient in essentially all baked goods. It is the flour that determines the rheological properties of both doughs and batters. Different kinds of baked goods require flours with different properties [111]. For example, flour that is good for bread production may not be satisfactory for producing cookies. The commercial value of wheat flour depends on a number of factors. The quantity and quality of protein, the ash content, color, moisture content, and level of enzyme activity are the most important factors. Flours for the production of yeast leavened products (bread, rolls, etc.) are characterized by relatively high protein contents (> 10.5 %), high water absorption, and moderate levels of enzyme activity. The flours should produce doughs with a good balance of elastic and viscous properties. For the production of chemically leavened products (cookies, cakes, crackers, etc.) wheat flours with low ash, relatively low protein content (8–10 %), and low contents of damaged starch are preferred [112].

There are many methods to evaluate flour quality [113 – 116]. In addition, a number of rapid methods mostly related to near-infrared reflectance (NIR) are becoming more popular [117, 118]. In the next sections evaluation techniques are discussed that are commonly used in the production of baked goods.

6.1. Physical Dough Tests

A number of dough tests are used in an attempt to predict the performance of wheat flours [119 – 121]. Unfortunately none of the tests is successful in making those predictions because

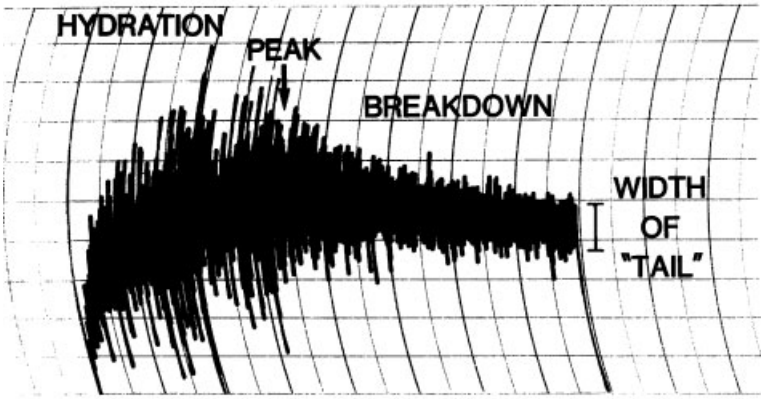

Figure 3. A mixograph curve showing the stages of flour hydration, dough development, and dough breakdown.

difference in flour quality manifests itself during baking. For example, if bread doughs made from flours of different baking quality are all proofed to the same height, they still give large differences in loaf volume after baking. Cookie doughs all cut with the same size cutter will spread to different diameters, and cake batter of the same weight will give different size cakes. Although physical dough tests may fail in predicting final baking performance, they can give useful information related to the processing behavior of flours.

6.1.1. Mixing Curves

Two machines that measure the mixing properties of flour-water doughs are the farinograph [122] and the mixograph [123]. Both machines are torque measuring devices. As flour and water are mixed the dough develops and offers more resistance to being mixed up to a point, beyond which the resistance decreases.

The *farinograph* has two sigma shaped mixing blades that run at a differential (one at 90 rpm and the other at 60 rpm). The mixing action is relatively gentle and is characterized as a kneading action. The mixing bowl is temperature controlled. The procedure for the farinograph calls for centering the farinograph curve (plot of torque vs. time) on an arbitrary 500 Brabender Units (BU) line. To do this, the amount of water in the dough is altered. From the farinograph curve important data are obtained. The curve provides the mixing time (time to peak resistance), the amount of water needed to center the curve on the 500 BU line (water absorption), and the tolerance to overmixing (stability). One problem with the farinograph is that some flours will give a false hydration peak (see Section 7.1) that is often mistaken for the mixing peak. The hydration peak is often more dominant than the mixing peak and with computer analyzed curves it is often taken as the mixing peak. In general, with North American flours farinograph mixing peaks of 3 min or less should be questioned.

The *mixograph* has a much more severe mixing action and, as a result, the peaks are easier to identify. The mixing action is a pulling stretching action instead of kneading. The mixograph allows the determination of the optimum water addition to a flour from the shape of the mixograph curve (Fig. 3). The curve height is related to the flour protein content. The mixing time (time to peak) and mixing tolerance (rate of curve decline after the peak) can be obtained from the curve. A wider tail of the mixograph curve indicates a better tolerance to abuse such as overmixing or overproofing the dough.

In general, strong flours in either instrument will have relatively long mixing times, with good stability, and wider curve widths. The mixing instruments will easily identify flour with good or poor strength and good or poor mixing tolerance. Perhaps the best use of the mixing curves is to check for flour consistency. If the same flour is to be delivered each day and the mixing curve changes it is reason to be concerned.

6.1.2. Flour Strength

Two instrument used to measure flour strength are the extensograph [124] and the alveograph [125]. For the *extensograph* dough is mixed in the farinograph, rounded, and formed into a log shape. The ends of the log are clamped into a holder and after a suitable relaxation time, the holder and dough are placed in the machine and a hook stretches the dough until it fails. The resulting curve gives a measure of the dough strength (curve height) and extensibility (curve length). A good dough will have a balance of strength and extensibility. More fundamental rheological measurements are also being used to measure flour strength [126 – 128].

The *alveograph* has its own mixer that mixes the dough and extrudes a dough sheet. The alveograph forces air under the sheet of dough to blow a dough bubble. The bubble is expanded until it fails and collapses. From the curve (a plot of pressure vs. time) the work used to blow the bubble can be calculated. More work is related to stronger flours.

6.2. α-Amylase Activity

The endogenous enzyme α-amylase is the most important enzyme in regard to bread production since it markedly influences baked product quality [129]. Three instruments are used to essentially measure the α-amylase activity of flours, i.e., the amylograph, the Rapid Visco Analyser, and the falling number apparatus. All three heat a flour–water slurry and measure the consistency/viscosity of the resulting slurry. As the level of α-amylase activity increases, starch is increasingly degraded during heating and a lower consistency/viscosity of the flour–water slurry is obtained.

7. Dough Formation

Production of bread products containing wheat flour begins by mixing flour, water, and various other ingredients to form a dough [130]. When wheat flour and water are mixed in various proportions, they form everything from a slurry when water is in large excess ($\geq 60\,\%$ water on a flour basis), to a dry but slightly cohesive powder when flour is in large excess ($< 35\,\%$ water). At intermediate water levels, the more liquid systems are described as batters, while at about 40–60 % water, the flour–water mixture is more like a sticky mess. However, when such a system is continuously stirred or mixed, remarkable changes occur. The system appears to become less wet and sticky, and a dough with cohesive and viscoelastic properties is formed. In a bread making dough the different flour constituents vary in their water uptake capacity. In a dough system starch (granular and damaged) takes up about 46 % of the water, protein about 32 %, and the arabinoxylans about 22 % [131].

7.1. Underlying Mechanisms of Dough Formation

Flour particles are highly hygroscopic and quickly absorb water. When water is added to the flour, the surfaces of the particles rapidly hydrate because water is in large excess compared to the surface area of the particles. Microscopy studies have shown that upon wetting of flour particles hydrated protein strands are formed at the air–flour–water interface due to surface tension forces [132 – 134, 269]. The strands quickly interact with each other to form physical crosslinks between flour particles.

In the absence of mixing, the only driving force for the water to move from the surface to the center of the flour particle is diffusion, which is slow. Mixing provides an additional mechanism by removing the hydrated surface of the particles and exposing new surfaces on flour particles for interaction with water. As more and more of the free water is used to hydrate the protein and starch, the resistance of the system to extension increases progressively and the height of the mixing curve gradually develops to a peak (Fig. 3). At this stage no intact flour particles are present and continuous protein films with adhering starch granules form a homogenous mixture (dough; Fig. 4). When the dough shows its maximum resistance to extension, it is said to be developed. All the flour particles are now hydrated. Particles from hard wheat flour are dense and water penetrates the particles only slowly whereas particles from soft wheat flour are less dense, smaller, and contain voids

that allow a more rapid moisture uptake. Since dough development is essentially a result of complete hydration of the flour particles and formation of a viscoelastic protein network, doughs from hard wheat flours need longer mixing times than doughs from soft wheat flours. The kinetics of dough development are also affected by the flour protein content. In the mixograph, low protein flours ($< 12\,\%$) require longer mixing times simply because they contain less protein. A flour protein content above 12 % does not affect mixing time. The overall positive relationship between wheat hardness and protein content has already been described (see Chap. 2). A dough that has been mixed to a peak can be referred, for example, as mixed dough, a dough with minimum mobility, or an optimally mixed dough. All of these imply that an end point has been reached, the point to which a dough should be mixed for producing a loaf of bread.

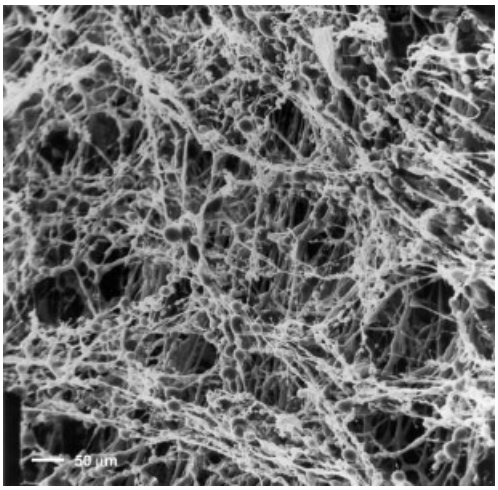

Figure 4. Scanning electron micrograph of a lyophilized mixed dough.

7.2. Role of Proteins in Dough Formation

A dough made from wheat flour is viscoelastic, it exhibits both viscous flow and rubber-like elastic recovery. The various rheological techniques used to follow the bread-making process are described in [136]. Flour protein content and quality are the major factors accounting for the viscoelastic properties of dough and thus dough mixing properties and bread making performance [70 – 73, 75 – 77, 79, 130, 131, 137, 138]. In general, flours with higher levels of protein ($> 12\,\%$) provide stronger doughs with higher elasticity, higher resistance to overmixing, and higher bread making potential. The weight-average molecular mass of the polymeric fraction (glutenin) appears to be particularly important [139].

The composition, structure, molecular size distribution, and aggregative behavior of gliadin [140, 141] and glutenin that together form the gluten mass determine the quality of the endosperm storage proteins. With respect to protein composition, scores have been assigned to the high molecular mass glutenin subunits to rank them in order of their effects on bread making quality [142]. The scores rank the subunits in order of their effects on gluten strength. The mechanism by which certain high molecular mass subunits confer to dough elasticity and bread making properties has been suggested to be related to their structural conformation. The repetitive domains of high molecular mass subunits have a β-helical conformation that may be intrinsically elastic, capable of expansion and contraction in response to stress, and therefore be of significance in the formation of gluten elasticity. It has also been suggested that the size of the glutenin protein may be responsible for its elastic properties [143]. Quantitative effects such as the level and ratio of high and low molecular mass glutenin subunits will also be superimposed on the intrinsic potentials.

In specific subunits the possibility of formation of intrachain disulfide bonds exists. The position and orientations of cysteine residues within the high molecular mass subunits capable of creating intramolecular disulfide bonds affect the generation of a high molecular mass disulfide-bonded polymeric glutenin structure [144 – 148]. This structure is a key factor determining dough strength and elasticity. A hierarchical arrangement of high and low molecular mass glutenin subunits has been proposed in relation to their proposed intrinsic contributions to dough strength [149]. This arrangement is based on the potential of a subunit to form branched polymers and the length of the repeating sequence domain. Branched polymers

and longer repeating sequence domains lead to stronger doughs.

A polymer science model has been proposed for the development of a continuous protein structure in a dough system. The model involves shear-thinning (i.e., slippage of polypeptide chains against one another), and crosslinks and/or entanglements between large glutenin polymers that are constantly breaking and reforming [79, 150]. This allows the polymers to relax after being stressed and thus, to lose their strain. A gluten–water dough has a smaller elastic modulus than a flour–water dough. It appears, therefore, that starch acts as a filler in the gluten polymer. Filled polymers generally have a larger elastic modulus than their unfilled counterparts. This situation is similar to using carbon black as a filler in the rubber for tires.

In general, dough development that is reflected by an increase in dough consistency and a peak at optimum development is followed by a dough breakdown, i.e., a decrease in dough consistency (Fig. 3). Various types of crosslinks such as sulfhydryl–disulfide interchange reactions, noncovalent hydrogen bonds between glutamine side chains, hydrophobic interactions, and chain entanglements affect the rheological dough properties. The decrease in dough consistency upon overmixing, i.e., dough weakening, is accompanied by a breaking down of the large macropolymer glutenin structure, possibly due to the rupture of disulfide bonds and loss of high molecular mass glutenin subunits [151, 152]. Chemical reducing agents such as cysteine, dithiothreitol, sodium bisulfite, and related compounds are quite effective in weakening the dough and shortening mixing time by breaking disulfide bonds. Flour contains the reducing agent glutathione [153]. In contrast, oxidation increases the strength of the dough [154] and the addition of oxidizing agents to a partially reduced dough restores the original dough strength. This treatment is accompanied by a return to the original glutenin molecular mass distribution and thus, provides evidence of the importance of sulfhydryl–disulfide interchange reactions in dough formation. The reaction of oxidizing agents such as iodate and of *N*-ethylmaleimide (NEMI) in dough appears to proceed via blocking of accessible and reactive thiol (SH) groups. This leads to an inhibition of sulfhydryl–disulfide interchange reactions and subsequently to a stabilization of aggregated wheat proteins [131, 155]. Bakers have taken practical advantage of these disulfide interchange reactions during dough processing. By adding minute quantities of reducing substances such as cysteine, dough can be developed much faster at lower mixing speeds, i.e., lower energy input. Such a developed dough is extremely fragile and, where allowed, is stabilized by addition of a slow-acting oxidizing agent such as potassium bromate.

Wheat flour doughs mixed in a nitrogen atmosphere or in the absence of the water-soluble flour fraction do not decrease in dough consistency as a result of overmixing. This observation implies that oxidative reactions mediated by water soluble compounds are involved in the dough breakdown mechanism. Addition of fumaric acid and related compounds to the water-soluble flour fraction reduce mixing time and also greatly increases the rate of dough breakdown [156]. Certain α, β-unsaturated carbonyl compounds such as fumaric, maleic, or ferulic acids have effects on overmixing similar to those of the sulfhydryl blocking reagent *N*-ethylmaleimide. The effect of the activated double-bond compounds on dough breakdown during overmixing can be reversed by lipoxygenase present in enzyme-active soy flour or by free radical scavengers [157]. The pH of the dough also effects mixing time, with lower pH giving a shorter mixing time and higher pH (up to pH 10) giving longer times [158]. This might be explained as an effect upon the charge of the proteins. The effect of low pH can be overcome by addition of salt (presumably by charge shielding). If a dough is mixed to optimum, allowed to relax and then mixed again it will redevelop and the resistance to extension will be restored to its original value. This implies that reversible and/or noncovalent molecular bondings are involved. Direct evidence for the presence of hydrogen bonds in gluten can be inferred from mixing experiments with heavy water (D_2O). Hydrophobic interactions can be related to the high level (30 %) of nonpolar amino acid residues in gluten. The very low charge density of gluten proteins also facilitates chain interactions.

Research is focusing on the interactions between flour constituents, specifically between the different gluten proteins themselves and the non-

protein flour components starch, lipids and arabinoxylans [29, 159].

7.3. Thermodynamic View of Dough Mixing

The phase behavior of biopolymer mixtures and excluded volume effects of macromolecules are critical factors influencing the functionality of dough [160]. During the mixing of flour and water, albumins, globulins, water-soluble starch, and arabinoxylans form a liquid aqueous phase. This is immiscible with the gluten phase. At first, bread dough contains two continuous phases: a gluten thixotropic gel phase and a liquid phase. Mechanical treatment greatly affects the structural–property relationship of doughs by establishing equilibrium between the coexisting phases: transformation of the continuous liquid protein–polysaccharide phase into the dispersed phase and orientation of the polypeptide chains of the gluten phase.

7.4. Aeration of Dough during Mixing

Incorporation of air during dough mixing is necessary to obtain a baked loaf of good volume, structure, and texture. As dough becomes cohesive, it starts to incorporate air and, thus, decreases in density. The small air bubbles created in the dough phase during mixing provide the nucleation sites to which the CO_2 produced by the yeast diffuses during fermentation. Without these nuclei the grain of the final bread would be very coarse with only a few large cells, because yeast is unable to create new gas cells. The viscoelastic texture of the dough is able to retain the CO_2 produced during proofing without rupturing. The total number of bubbles per cubic centimeter dough generated during mixing depends on the mixing conditions. With laboratory scale mixers operating under atmospheric pressure about 5,000–10,000 bubbles with a diameter of about 50–100 μm are incorporated into a dough [161].

Some doughs, as practiced in the Chorleywood bread making process (see Section 9.4), are mixed under vacuum. At first glance, this would appear to be detrimental to its quality. However, the vacuum is not complete. High vacuum cannot exist because dough contains water. The partial vacuum is beneficial for the baked product texture. Mixing under reduced pressure increases the size of existing bubbles and makes them easier to subdivide. The result is a larger number of bubbles which leads to a finer crumb grain.

8. Role of Bread Ingredients [162]

The minimum formula for bread is flour, yeast, salt, and water. If any one of these ingredients is missing, the product is not bread. Other ingredients that are often found in the formula are fat, sugar, milk or so-called milk replacers, oxidants, various enzyme preparations (including malted grain), surfactants, and additives to protect against molds. Each of the components in the formula performs a specific function in producing the finished loaf [7, 8, 163, 164]. The flour, of course, is the major component and is responsible for the structure of the bread.

Yeast. (baker's yeast, *Saccharomyces cerevisiae*) (→ Yeasts) is one of the fundamental ingredients; its major role is to convert fermentable carbohydrates into carbon dioxide and ethanol. The gases that result from that conversion provide the lift that produces a light, or leavened, loaf of bread. In addition to its gas production, the yeast has a marked effect on the rheological properties of dough. Furthermore, yeast fermentation affects texture and flavor characteristics of finished baked products.

Fat/Shortening. Fat has several functions in bread making [165]. In dough making, fat acts as a plasticizer. Thus, if the level of fat in the dough is increased, the amount of water must be reduced and vice versa to maintain an equal dough consistency. Fat increases the volume of the bread. The increase in volume is significant, usually about 10%. In bread that is to be stored for any significant period of time after baking, fat is a desirable ingredient due to its antistaling properties. Bread containing fat in the formula stays soft and more palatable for a longer period of time than does bread prepared without fat.

Sugar is added to the formula as a source of fermentable carbohydrate for the yeast, and if used in sufficient levels provides a slightly sweet taste to the bread. With the proper enzymes present in the dough, sufficient sugar is produced from the flour to maintain fermentation and added sugar is not necessary for gas production. However, under most production conditions, added sugar is used for fermentation. Most Americans will state, when asked, that they do not prefer sweet bread. However, if they are given a series of breads to taste and instructions to rank the breads according to flavor, their rankings will generally follow the raise in sugar level. This probably explains why the level of sugar in U.S. bread continues to increase. In other countries sugar is mainly added to the formula to induce yeast fermentation and to enhance the color of the crust.

Milk or milk solids were a common ingredient in U.S. bread in the recent past. The nutritional advantages of milk are well known. It also had widespread consumer appeal and, because the United States was producing a large excess of milk, it was a relatively cheap ingredient. The latter situation has now changed drastically, and milk is an expensive ingredient. Therefore, it is hardly used anymore. More common are milk replacers such as soy–whey mixtures. Bread for toasting, in general, contains higher levels of milk powder preparations to enhance browning reactions and thus color development during toasting.

Oxidants. The chemical oxidants that are added to the bread formula affect dough rheology and baked product quality [166]. Oxidants such as ascorbic acid [167 – 169], potassium bromate [170, 171], azodicarbonamide, and calcium peroxide, at ppm levels, improve dough strength and result in bread with better loaf volume and texture [154]. Several oxidants (potassium iodate and azodicarbonamide, for example), are fast-acting and are effective during mixing. Potassium bromate, on the other hand, has essentially no effect during mixing but affects dough rheology during fermentation and in the early stages of baking. At least part of the time-dependent effect of potassium bromate may be attributed to the change in dough pH during fermentation; potassium bromate reacts faster at lower pH. Ascorbic acid has both a rapid and a time-dependent reaction on dough rheology. The type and level of oxidants that can be used in bread varies from country to country. In the EU questions have been raised on the value on chemical flour treatments and use of additives. For example, most European countries do not allow the use of bromate for bread flour. However, the functionality of bromate is very complex and the baking industry is facing a challenge in replacing bromate [170].

Surfactants. Bread that must remain soft for a number of days usually contains a surfactant in addition to enzymes (amylases) [172]. The most common compounds for this usage are α-monoacyl glycerols, more commonly (but less correctly) referred to as monoglycerides; a common level of usage is 0.5 % based on the flour weight. Other surfactants are used as dough strengtheners. These are helpful in allowing the dough to withstand the mechanical stress of the bakery's processing lines. Examples of this type of surfactant are, e.g., sodium stearoyl lactylate (SSL), ethoxylated monoglycerides (EMG), and diacetyl tartaric acid esters of mono- and diglycerides (DATEM). They also are used at about 0.5 % of the flour weight.

Enzymes. There is a worldwide trend away from chemical additives in food processing. As a part of this trend, bakers are turning to enzymes as an alternative to chemicals [173 – 175]. Of course, enzymes have been used in baking throughout history. However, the use of enzymes as formula ingredients is relatively new. Often, enzymes have synergistic effects when used in combinations. These effects can be exploited to improve both, processing performance and end-product quality. One of the major advantages of enzymes is that they are considered as processing aids. They act during dough/batter preparation and in the early stages of baking before they are denatured and inactivated. As such, the enzymes need not to appear on the product label. However, some heat-stable enzymes may survive the baking process. If they fulfill a function in the final product, they need to be declared.

α-Amylase. The most popular enzyme added to baking formulas is α-amylase. Too much α-amylase activity result in a poor loaf with a sticky crumb. Too little α-amylase activity results in a

smaller loaf with a course crumb texture. Sound wheats are low in α-amylase. Flours produced from such wheats produce bread with a harsh crumb. To correct this problem, malted barley or malted wheat flour is added, usually at about 0.25 % of the flour weight. This treatment improves loaf volume and reduces the rough texture of the crumb. The proper treatment level can be determined by the amylograph, Rapid Visco Analyser, or falling number tests (see Section 6.2). In very lean formulas with no added sugar the α-amylase together with the endogenous β-amylase produces maltose that can be used for fermentation. If sugar is added to the formula, yeast does not utilize maltose. An important role of added α-amylase is to destroy damaged starch (see Section 5.1.1). Damaged starch has the advantage of increasing water absorption during mixing. However, it is detrimental later in the baking system. Its removal facilitates obtaining a loaf of good volume and texture. Another role of α-amylase that might be linked to the α-amylase related volume and texture improvements, is as an anti-staling aid, keeping bread from firming during storage [50, 176]. Apart from the action on damaged starch, the action of the α-amylase on regular starch is in a time window after the starch has gelatinized (ca. 65 °C) but before the enzyme is denatured. α-Amylase produces dextrins that remain in the final product as the natively present β-amylase which is more sensitive to heat, is no longer active. These formed dextrins may well interfere with the firming of bread.

Xylanases. Other enzymes that are finding more use in baking are xylanases that degrade wheat arabinoxylans [177]. Much progress has been made in understanding xylanase functionality. An important breakthrough was the finding that some xylanases preferentially solubilize water-unextractable arabinoxylans, while others preferentially hydrolyze the water-extractable components [104, 178, 179]. In addition, in the 1990s, it was found that wheat flour contains proteins which can inhibit xylanase enzymes [180 – 183]. In the particular case of bread making, the ideal xylanase is one which preferentially solubilizes water-unextractable arabinoxylan, leaves the water-extractable component intact, and is not inhibited [184]. In a dough system, xylanases appear to improve the rheological properties. This may be explained by the degradation of arabinoxylans that can interfere with the development of a continuous gluten network [175]. Appropriate xylanases are particularly useful in high fiber content breads. In a batter system, depending on their specific action, xylanases can decrease or increase the viscosity of the batter.

Lipoxygenase has several beneficial effects during baking [185, 186]. Lipoxygenase, usually from enzyme active soy or pea flour, has a strong bleaching action. In addition, the lipoxygenase action results in improved mixing tolerance, a strengthening of the dough, and increased loaf volume. Oxygen or hydrogen peroxide produced by lipoxygenase and other oxidative enzymes may play a role in redox reactions of proteins and arabinoxylans during mixing [187].

Glucose oxidase has been used to produce stronger dough. Glucose oxidase produces hydrogen peroxide that appears to be the active agent. This enzyme has been partially successful in replacing certain chemical oxidants in the bread making formula.

Transglutaminase, a protein crosslinking enzyme, has also been suggested to increase or maintain bread volume when other ingredients are substituted or reduced during dough preparation [188, 189].

Proteolytic Enzymes. Various proteolytic enzyme preparations are used by bakers, generally to weaken bucky (too elastic) doughs. Proteolytic enzymes can be divided into exo- and endo-acting types. The endo-enzymes affect dough rheology while the exo-enzymes do not. One problem with the action of endoproteolytic enzymes is that the activity continues as long as the dough phase lasts. This leads to disaster if there is down time. Under such conditions the dough will turn into a sticky mess that cannot be recovered. Proteolytic enzymes are also used in the preparation of batters (e.g., for wafer baking) to avoid gluten protein aggregation (lump formation) and to reduce batter viscosity. These modifications facilitate batter mixing, pumping and batter depositing on the baking plates. The effect of proteolytic enzymes can be attributed to the hydrolysis and subsequent molecular mass reduction of the wheat glutenin polypeptides [190].

Lipase has only been recently introduced into bread making. Beneficial effects include an increase in bread volume, a more uniform crumb

structure and improved softness during storage [175, 191]. It has been suggested that one of the mechanisms of action of lipase in bread making is related to the increase in gluten strength.

Salt/Water. Salt is generally used at levels of about 1–2 % based on flour weight. Salt appears to have two major functions. First is taste; bread made with no salt is quite tasteless. The second is to affect the dough's rheological properties. Salt makes dough stronger, presumably by shielding charges on the dough proteins. The last fundamental ingredient is water, which acts as plasticizer, solvent, and reaction medium.

Other Ingredients. A large number of other ingredients are found in bread formulations. Most, if not all, of these are added to affect the sensory attributes of the resultant bread and not for any technological functional effects. These items include such nongrain ingredients as eggs, honey, syrups, fruits (often dried), nuts, and various spices. Grain other than wheat can be used individually or in combination to add diversity, flavor, and eventually nutritional benefits by blending with wheat flour [192]. Nongrain plant materials (often fibers) are also used occasionally. In many parts of the world wheat flour is enriched for nutritional reasons. In the USA, since 1940, all enriched "family" flour and baker's bread flour have been required to be enriched with the vitamins thiamin, riboflavin, and niacin and with iron. The purpose is to replace part of the vitamins and minerals lost as a result of milling. In 1998, the addition of folic acid was made mandatory for enriched flours in the USA. All UK flours used for bread production must contain at least 0.24 mg of vitamin B, 1.6 mg of nicotinic acid, and 1.65 mg of iron per 100 g of flour. Lower extraction flours are also enriched with calcium carbonate.

9. Bread and Dough Making Processes

There are a large number of processes used to make dough and bread [7, 8, 193 – 196]. In this chapter only the most important ones will be covered.

9.1. Straight Dough

In a straight dough process all the formula ingredients are added together at the start and the mixture is mixed into optimally developed dough. The dough is then allowed to ferment. The fermentation time may vary but generally would be 2–3 h. After fermentation the dough is divided into loaf sized pieces of dough, rounded into a ball, given an intermediate proof time of 10–20 min (which allows the dough to relax), and then molded and panned. The dough cylinders are then allowed to proof for about 1 h before going to the oven. Bread produced by a straight dough procedure is strong, chewy, and tends to have an open grain. The straight dough procedure has a strong time dependence. If a large dough quantity is mixed, the bread produced from the start and end of the dough (perhaps 30 min of elapsed time) will be noticeably different in properties.

9.2. Sponge and Dough

In the sponge and dough process a sponge is prepared from part of the flour (ca. 65 %), part of the water, and the yeast. The sponge is only mixed enough to have a uniform mixture and then allowed to ferment for 3–4 h. After fermentation, the sponge is brought back to the mixer and mixed with the rest of the formula ingredients. At this stage the dough is mixed to optimum development. After mixing the dough is given a floor time (15–20 min) to allow it to relax. It is then processed as described for the straight dough (Section 9.1). The sponge and dough procedure requires more time, more space, and more personnel than the straight dough procedure. In general, it produces bread with a soft and uniform grain. The process is very tolerant, particularly to time. Thus, large batches can be processed with relatively little variation in bread quality. It is also generally believed to give a better bread flavor than the straight dough process.

9.3. Liquid Preferments

Preferments have become increasing popular in the last few years. The use of preferments saves

labor costs, space, and processing time. Preferments consist of water, yeast, sugar, inorganic salts, and flour. The ratio of water to flour is maintained so that the preferment is a pumpable, liquid mixture. Preferments precede dough fermentation and thus, they can be viewed as simply a more liquid sponge that is then combined with other formula ingredients to give a dough. In this sense, preferments are used in place of the sponge. Control of pH is important in preferments. The pH should be maintained between 4.5 and 5.2. The preferment is buffered by the flour, salts, and often milk powder. With no buffer, the pH of a nonflour brew will drop to about 2.5. An advantage of the preferment is that it can be chilled and held for 24 h without loss of activity. Of course, the quantities of flour and water must be adjusted to maintain the original formula. A single preferment can be used not only for bread but also for rolls, snack crackers, and sweet goods such as doughnuts.

9.4. No-Time Processes

These systems really do not require no time. What the term no time refers to is the amount of fermentation time. Several systems have been developed. The continuous system in the USA, the Chorleywood process in the U.K., and the Brimec process in Australia, to name a few. The latter two systems are still widely used in their respective countries. The continuous system in the United States is no longer widely used, although it once accounted for as much as 40 % of U.S. bread production. The *U.S. continuous system* generally used a preferment that was mixed with the other ingredients and developed with a high speed developer under pressure. The dough was then extruded directly into bread pans, proofed and baked. The bread produced had an extremely uniform grain and a very soft texture. Consumers in the U.S., however, appeared to view the bread produced as inferior and sales dropped even though the bread was priced below other bread on the market.

The *Chorleywood process* employs no preferment. It uses a high speed Tweedy mixer and develops the dough under a partial vacuum. The *Brimec process* also uses no preferment and a special mixer developed for this process. The mixer operates at a pressure of 150 kPa and the dough is extruded directly into the baking pan. All short time systems require high levels of oxidants. Commonly used oxidants in the systems are potassium iodate, potassium bromate, and ascorbic acid.

9.5. Sour Dough

Utilization of sour dough is an old-time leavening method in bread making. During sour dough fermentation a typical microflora develops that includes lactic acid bacteria (lactobacilli) and yeasts [197]. The yeasts are generally responsible for the leavening action via carbon dioxide production and for flavor precursor formation. Lactic acid bacteria lower the pH by producing lactic and acetic acid. A typical ratio is about 20 % acetic and the remainder lactic acid. The acidity protects against spoilage by inhibiting mold growth. The growth of rope bacteria is increasingly inhibited as the pH drops below 6.0 and the acids formed preserve the dough. The desired pH of the dough is in the range of 4.0 to 4.6. However, the pH must not become too low, because below about pH 3.7, putrefaction bacteria take over, developing bad odors. Generally this is not a problem, as flour is a relatively good buffer. In addition, lactic acid bacteria produce enzymes (α-amylase, phytase and proteolytic enzymes) that form low molecular mass carbohydrates, peptides, and amino acids that can act as flavor precursors.

For the production of rye bread an acidification is required. The acidity affects the swelling properties of rye flour constituents (proteins, arabinoxylans), controls the enzyme activity in the dough and improves the bread making performance (i.e., improved crumb grain, elasticity, slicing properties, flavor and taste perception, crumb/crust color, and shelf life). Compared to wheat bread rye bread prepared with sour dough has a denser loaf with lower volume, sour-aromatic taste and prolonged shelf life.

The sour dough process is a very complex biological system. Factors affecting this system include process variables such as fermentation temperature and time, and ingredient parameters such as flour type, flour ash content, media sources (availability of carbon, nitrogen-sources, vitamins), pH, water concentration and the presence of antimicrobial compounds. These

parameters affect the multiplication of lactobacilli and yeast, the amounts of acids formed, gas (CO_2, O_2) and ethanol concentrations, the proteolytic activity of lactobacilli, and thereby, the organoleptic properties of the baked product such as taste, flavor, texture, and crumb color.

Sour dough production traditionally includes three stages of fermentation: a fresh sour, a basic sour and a full sour (final dough for breadbaking). To initiate sour dough fermentation and to prepare the fresh sour, a spontaneous sour (spontaneously fermented flour-water mixture), a portion of an already developed, old sour dough or a commercial starter culture derived from natural sour dough fermentations can be used. As dough becomes old, the yeast is inactivted, as the pH is low. At the lower pH, bacteria are still quite active. With time, the food available for the organisms becomes limited; however, as new flour is added and the "starter" is fed or rebuilt, the bacteria become more predominant. In general, the gas-producing ability of sour-dough organisms is lower than that of commercial yeast. Consequently proof time of sour dough bread is often long, of the order of several hours. Various sour dough processes (multi-, two, single stage varying from 2 h up to 24 h) were designed to increase the growth of yeast and lactic acid bacteria to give the final sour dough (full sour) proper acidity (especially the lactic acid/acetic acid ratio) and dough consistency. Often, baker's yeast (*Saccharomyces cerevisiae*) is added to accelerate the leavening process. The use of organic acids (dough acidifiers) including lactic, acetic, tartaric, and citric acids alone or in combination with other additives such as sour dough concentrates can replace sour dough in single-stage processes or partly replace it in multistage processes. Apart from the usual batch process, technologies for continuous preparation of sour doughs have been developed [198]. A comprehensive documentation on sour dough technology, fermentation/microbiology and biochemistry is given in [199].

9.6. Frozen Dough

Frozen dough has gained popularity since the 1970s [200]. Frozen dough usage has been most successful for small goods, puff pastry, Danish, and croissants. Also successful have been bread systems where good control of both the temperature and time of frozen dough storage is attained. The temperature in the freezer should be kept constant and the freezing time as short as possible. Frozen doughs sold in the supermarket have been much less successful primarily because neither time nor temperature are fully controlled.

The deterioration of the baking quality of the dough that has been subjected to prolonged frozen storage has been mainly attributed to two factors: loss of yeast viability and dough strength. The reduction in yeast viability during frozen storage results in a loss of gassing power, longer proof times and, ultimately, reduced loaf volumes. In addition, reducing compounds such as glutathion released from the dead yeast cells cause depolymerization of gluten proteins, which may also be detrimental to dough rheology. The loss of dough strength during freezing has also been attributed to the physical disruption of the gas cell wall structure in the dough by ice crystal formation [201 – 204], which affects gas retention properties of the dough and final bread quality. Freezing and thawing or just variation in the storage temperature, even if it remains below freezing, is also detrimental to the quality (rheology) of frozen dough.

The baking performance of frozen dough can be optimized to a great extent through the use of appropriate formulations and processing conditions [200 – 202, 205, 206]. Standard practice for frozen dough production is to keep the dough cold, and the fermentation time as short as possible, increase the yeast level, and keep the time between freezing and usage as short as possible. Flour quality has also been implicated as a determinant of frozen dough quality [207]. Adhering to those guidelines will produce consistent bread from frozen dough.

There appears to be a trend away from frozen dough and toward par-baked, (i.e., partially baked) products as a process to give the consumer "fresh baked" products without going to the bakery.

10. Gas Production and Retention

10.1. Leavening Mechanisms

Leavening with gases is designed to control specific process requirements (e.g. batter and dough rheology, product density and volume) and to manufacture products with desired eating characteristics such as a light product texture and tender bite. As leavening gases are used: carbon dioxide, water and/or ethanol vapor, ammonia, and air. Leavening starts with the incorporation of air into the dough or batter during mixing to form nuclei of gas cells that will expand during baking. Air is therefore present and used in all baked products. The expansion of the initial gas cells can be achieved by the use of chemical aerating agents and/or yeast. Water is also present in all baked products and in all cases the evaporation of water in the dough or batter plays a role in the expansion process. However, mostly its leavening effect is quite limited because water has a relatively high boiling point. Water vapor is an effective leavener only when the product is heated at a very fast rate due to the rapid increase of water vapor pressure, as done in saltine cracker or wafer production. Yeast fermentation as biological leavening process via carbon dioxide and ethanol production is the major leavener for bread but is only rarely used in soft wheat products. Small amounts of yeast maybe used in some pretzel formulas. Most, if not all, soft wheat products are chemically leavened, that means, carbon dioxide is produced by chemical reactions of bicarbonate or carbonate with acids. The rheological effect of yeast is generally not desirable in soft wheat products. In soft wheat products, sodium and ammonium bicarbonates are the most common sources of carbon dioxide.

10.2. Fermentation and Gas Production by Yeast (→ Yeasts)

Yeast is a living organism that is inactive during storage. The inactivity is caused either by drying, in the case of active dry yeast, or by low temperature, in the case of compressed, crumbled, or cream yeast. Yeast, as it is produced commercially, is always contaminated with bacteria, mainly lactobacilli. These are quite important in saltine crackers and perhaps in some sour dough breads but do not appear to be important in regular bread making processes.

When yeast is incorporated into a dough, conditions are suitable for it to become active. Yeast is a versatile organism; it can ferment under either aerobic or anaerobic conditions [208]. The growth of yeast and the early stages of brewing are a aerobic processes, whereas bread fermentation is an anaerobic process. Thus, little growth of yeast occurs during dough fermentation. The yeast and bacteria rapidly consume the oxygen in dough as fermentation starts. Thereafter, the fermentation is anaerobic unless oxygen is added to the system (e.g., by remixing). The process of anaerobic fermentation of carbohydrates by yeast to the end products, ethanol and carbon dioxide, is summarized below:

$$C_6H_{12}O_6 + \text{yeast} \rightarrow 2\,CH_3CH_2OH + 2\,CO_2$$

In a system designed to maximize yeast activity, the substrate (sucrose or glucose) is utilized at a rate of 0.77–3.0 g of sugar per hour per gram of yeast solids. In a 3-h fermentation and 55-min proof, with compressed yeast (29 % solids) added at 2 % flour weight, that calculates to be 1.75–6.82 g of sugar per 100 g flour weight. Yeast utilizes sucrose following conversion to glucose and fructose by its efficient invertase enzyme system.

Carbon dioxide is retained in the bread dough in two phases: as a gas within the gas cells and dissolved in the aqueous phase. As carbon dioxide is produced it reacts to a small extent with water to produce carbonic acid and the pH decreases. Continued production of carbon dioxide soon leads to saturation of the aqueous phase. The apparent initial lag that is found in gas "production" curves for bread dough is because the aqueous phase of the dough must become saturated with carbon dioxide before its evolution or loss can be measured. Only after the aqueous phase has become saturated is the carbon dioxide available to leaven the system. Carbon dioxide solubility in water is inversely related to the temperature and affected by pH. At the pH of bread doughs, most of the carbon dioxide is present as CO_2 and only to a very small extent as H_2CO_3, HCO_3^-, or CO_3^{2-}. The carbon dioxide present in a fully proofed dough is only about 45 % of the total produced by fermentation. Part of it is lost during fermentation, punching, molding and proofing. The expansion of CO_2 in the

air cells and the CO_2 that comes out of the aqueous phase is insufficient to completely account for the increased volume of dough in the oven.

As fermentation proceeds, it is customary to punch or remix the dough, depending upon which baking system is used. The gas cells in the dough have become larger and larger as more gas is produced. Punching or remixing subdivides the gas cells to produce many smaller cells. During remixing, a large amount of CO_2 is lost to the atmosphere, but the important aspect of the process is the creation of the new gas cells. Another important benefit of punching or remixing is the mixing of the dough ingredients. Yeast cells are not mobile in dough. Therefore, they depend upon the diffusion of sugar to them. As fermentation proceeds, the diffusion distances become large, the concentration of this nutrient diminishes and along with it the rate of fermentation. Punching or remixing brings the yeast cells and fermentables together again. In zero- or short-time baking systems punching is not practiced, as the dough is not given sufficient time to expand. The net result is usually a coarser grain in the bread. A partial solution to this problem is to mix under partial vacuum. This increases the size of the bubbles and allows them to be subdivided into many small bubbles that contribute to a finer crumb texture.

In addition to producing gas, yeast also affects dough rheology. The effects of yeast on dough rheology can best be shown by a simple spread test [209]. Dough has both viscous-flow properties and elastic properties. Dough that has more viscous-flow properties has a large spread ratio (width of dough divided by height), whereas dough that is more elastic has a smaller one. A flour–water dough without yeast gives a larger spread ratio after 3 h than a yeast fermented flour–water dough (Fig. 5). This indicates that the viscous-flow properties are larger in flour–water dough. Thus, the addition of yeast to the formula causes a change in dough properties: the viscous-flow properties are reduced and the elastic properties are increased. It has been shown that the build-up of hydrogen peroxide produced by yeast during fermentation affects rheological dough properties [103].

Optimally fermented and oxidized dough has no viscous-flow properties under the force of gravity at the proof stage. Thus, the dough in the pan expands rather than flows to fill the pan. After fermentation, presumably because of the mechanical punching, the gluten fibrils appear to be aligned. The change in pH associated with fermentation time is also important to the dough's rheological properties. Dough just out of the mixer usually has a pH of about 6.0. During fermentation, the pH drops to 5.0. A rapid drop is caused at first by carbon dioxide dissolving in water to produce carbonic acid. A second factor is the slow production of organic acids by the bacteria in the dough. The flour itself and milk or soy proteins (if present in the formula) are good buffers and, therefore, help to control pH. A lower pH decreases the mixing time of dough. This is, at least in part, the reason for the shorter mixing time in a sponge-and-dough or preferment system than in a straight-dough system. However, the change in pH has little effect on the doughs spread ratio.

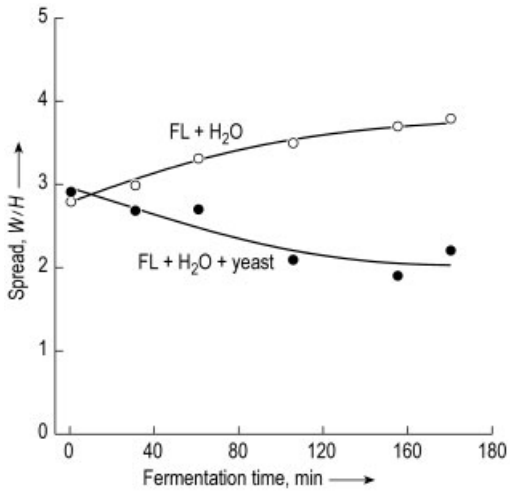

Figure 5. Effect of yeast on the spread ratio (W = width, H = height) of a fermenting dough
FL = flour

10.3. Gas Production by Chemical Leavening

Chemical leavening systems have been know since the first baking powder was formulated in 1838 in England and sodium bicarbonate is believed to have been used as a source of carbon dioxide since the 14th century. There are two basic types of chemical leavening: decomposition and chemical neutralization [210, 211]. Decom-

position, i.e. the breakdown of the leavening agent, takes place in the presence of moisture or heat. The most common example of decomposition is the use of ammonium bicarbonate that undergoes thermal decomposition at about 60 °C. When the product in the oven reaches this temperature, ammonium bicarbonate decomposes rapidly to ammonia, carbon dioxide, and water:

$$NH_4HCO_3 \rightarrow NH_3 + CO_2 + H_2O$$

Leavening by decomposition of ammonium bicarbonate can be used only in products that are thin and baked to low moisture levels (less than 5 %). If the product retains more than a few percent of water, it also retains the dissolved ammonia and even a small amount of ammonia in the product is associated with an off-flavor and makes the product inedible. Ammonium bicarbonate is used fairly widely in dry cookies and in some snack cracker type of products. It does have the advantage that it leaves no residual salt after decomposition. Residual salts can affect flavor, dough rheology or both.

Chemical neutralization is a process where a bicarbonate source is neutralized by an acid (HX) with release of carbon dioxide as leavening gas:

$$NaHCO_3 + HX \rightarrow NaX + H_2O + CO_2$$

Once the bicarbonate comes in contact with water in the dough or batter system, it quickly solubilizes and reacts with the acid under production of carbon dioxide. Since no ammonia is produced, this reaction is suited for leavening of high moisture baked goods. Carbon dioxide can exist as free CO_2, as HCO_3^- or CO_3^{2-}. The relative proportion of each is determined by the pH and temperature of the solution. The rate at which carbon dioxide is released from the bicarbonate depends on the rate at which the proton from the acid becomes available to react with the bicarbonate. No CO_2 is available if the pH stays above 8.0. Many soft wheat products end up with a pH near 7.0, where only a part of the carbon dioxide is in the gaseous state. To produce significant amounts of gas, the dough or batter must contain an acid. Many ingredients used in baking are sources of acid. Acidic fruits or buttermilk are obvious examples. If no acid is available natively, it is added to the formula. To determine the correct level of leavener and acid, the neutralizing value has been defined. It indicates the parts of weight of sodium bicarbonate that 100 parts by weight of leavening acid will neutralize. Thus, sufficient acidity to neutralize the bicarbonate and consequently, the release all of carbon dioxide is assured so that after baking there remains little, if any, unreacted bicarbonate or acid. Several acids are used in the baking industry. They all have distinct timing of release and reaction kinetics at various temperatures. Commonly used acids include cream of tartar (potassium hydrogen tartrate $KHC_4H_4O_6$), tartaric acid ($C_4H_6O_6$), δ-gluconolactone ($C_6O_6H_{10}$) and various sodium–aluminum–calcium phosphate compounds.

The most popular bicarbonate leavening agent is sodium bicarbonate (baking soda or soda). The popularity of sodium bicarbonate is due to its numerous advantages. It is relatively cheap, it is nontoxic, easy to handle, gives a relatively tasteless end product, and the commercially produced product is of high purity. Sodium bicarbonate decomposes when heated to about 90 °C in an aqueous system releasing half of the carbon dioxide released and the remainder becoming sodium carbonate.

Leavening systems used in industrial baking formulations are often tailor-made, comprising balanced blends of aerating chemicals to give a compound baking powder. These blends are adapted to the specific process requirements or constraints and to the desired organoleptic and technical product characteristics. They may combine various bicarbonates and acids and are either single- or double-acting. A double-acting baking powder is one that contains two acids, one that reacts (becomes soluble) at room temperature and another that reacts when the product is heated (heat-activated leavening agents). Encapsulated leavening agents perform by having the correct release and delivery kinetics. For example, encapsulated acids and encapsulated sodium bicarbonate are available as heat-activated leavening systems. Besides their leavening effect and, in some cases, their effect upon the product's taste by the salts produced during the leavening reaction, chemical leavening agents also can affect the rheology of the product. In general, di- or trivalent ions tend to increase the elasticity of the final product (especially cakes), and sulfate ions tend to decrease its elasticity. These ions presumably act by affecting ionic crosslinks with the proteins in the

batters. Dried starch is usually used as an inert diluent in baking powders. Its primary function is to physically separate the soda and acid particles and prevent their premature reaction.

10.4. Gas Retention

Bubble mechanics shows that the pressure P in a bubble is related to the radius of the bubble r and the interfacial tension γ by the following relationship:

$$P = \frac{2\gamma}{r}$$

Thus, in a system where the interfacial tension does not change, if r approaches zero, then the pressure required to start a new bubble is infinite. This implies that yeast can not create new bubbles in a dough system [212] and that air must be incorporated during mixing in the dough to provide preexisting bubbles (see Section 7.4).

In mixed bread dough, an insoluble but highly hydrated gluten protein system constitutes the continuous phase, with starch and air bubbles as the discontinuous phases. Also dispersed throughout the aqueous system are yeast cells, which ferment sugar and produce, among others, CO_2. The CO_2 is produced in the aqueous phase and saturates the water. Once the water is saturated, newly produced CO_2 must find a place to go. As discussed above, it cannot form new bubbles; thus, the CO_2 enters the preexisting air bubbles and increases the pressure. Dough has viscous flow properties and therefore, allows the bubble to expand to equalize the pressure. The CO_2 in the bubble cannot diffuse out because the aqueous phase surrounding the bubble is saturated with CO_2. Thus, the total volume of the dough mass is increased or, in other words, the dough is leavened. The use of ultrasound to investigate the cellular structure of bread and presumably dough has been described [213].

11. Molding and Proofing

After bulk fermentation, dough is divided into individual loaf-sized pieces, and is given a floor time. The purpose of this time is to allow the dough to relax. The dough is then ready for molding. The molding operation is essentially sheeting followed by curling, rolling, and application of pressure. As the dough is sheeted (passed between rolls to be flattened) during the various processes, it must be sheeted in different directions. Continued machining in one direction would align the protein fibrils and result in dough that is strong in one direction but weak in the direction at a 90° angle to the sheeting. After being molded, the loaf is panned. The dough is then ready for proofing. This is usually accomplished at 30–35 °C and at 85 % relative humidity. Because the dough now has only limited viscous-flow properties, it fills the pan by expansion. Proofing usually takes about 55–65 min; the dough increases greatly in volume. After proofing, the dough is ready for baking.

12. Baking

It is easy to tell the difference between dough and bread. The simple experiment of attempting to blow through dough shows that in dough gluten is continuous and the gas is discontinuous (in gas cells). Dough can be used to blow a bubble (e.g., in the alveograph). On the other hand, a slice of bread offers no resistance to the passage of air. Bread is not only gluten-continuous but also gas-continuous and is no longer capable of retaining gas [214].

However, the transformation of dough into bread, the conversion of the foam structure of the dough into the open sponge structure of the bread is a complex process involving irreversible changes (Fig. 6) [7, 133, 215 – 217]. Baking times and temperatures as well as baking temperature profiles vary widely, depending on the type of bread. Heat is transferred to the loaf mostly through the pan or baking hearth. In the crust and crumb, starch gelatinization, protein polymerization, and Maillard reactions (i.e., color and flavor generation) take place. Reactions inside the loaf (crumb) and at the surface (crust) are different because of the differences in moisture content and temperature.

12.1. Transformation of Dough into Bread

During the first few minutes in the oven, the dough expands rapidly in size (oven-spring or

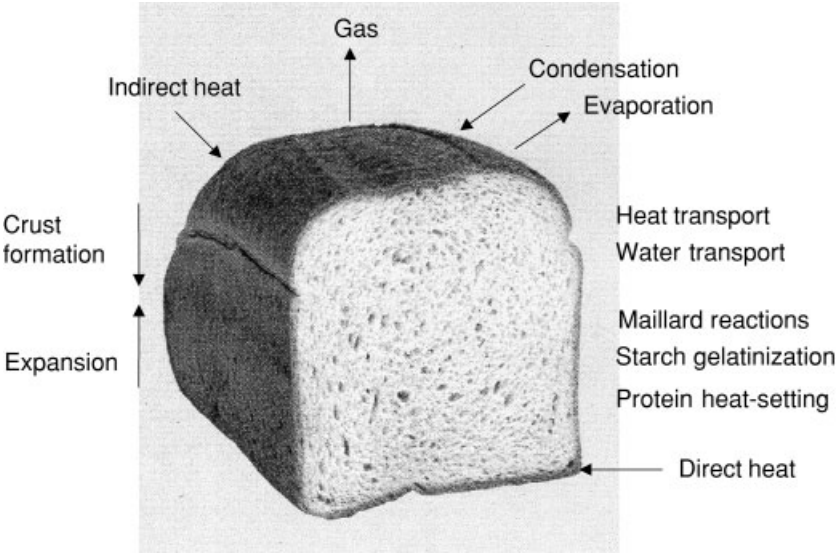

Figure 6. Changes during the transformation of dough into bread.

oven-rise). Several factors are responsible for the oven-spring. As temperature increases gases increase in volume; carbon dioxide becomes less soluble; yeast becomes quite active in the initial phase of baking producing more CO_2 and ethanol; evaporation of CO_2 and the water–ethanol mixture during heating in the oven contributes greatly to the overall expansion of the dough. Generally the oven-spring stage lasts less than 8 min; the remaining baking time is to ensure that the center of the loaf approaches 100 °C.

If pan bread is produced, most of the heat that is absorbed by the dough comes through the baking pan. Therefore, the rate at which the dough temperature increases depends on the heat transfer from the air and the baking surface to the baking pan. Dough does not conduct heat as well as the metal pan; therefore, a well-defined temperature gradient forms from the outside to the center of the loaf during baking. The internal structure of dough that has been in the oven for various periods of time shows a sharp demarcation between dough and bread regions of each loaf (Fig. 7). This line of demarcation is consistent with the temperature gradient known to develop during baking.

When bread dough is baked in an electrical resistance oven the temperature gradients are, essentially, removed and the dough heats uniformly. When dough is baked by this method and, at the same time, carbon dioxide loss from the loaf is measured along with the dough temperature, the temperature at which the dough starts to lose carbon dioxide can be determined (Fig. 8). The question then arises as to how flour protein content and quality affect loaf volume. The wheat flour dough retains gas as the gas can only diffuse slowly through the complex gluten network. At higher protein content there is more gluten to cover the surface of the starch and air cells, thus, the gluten phase between the air cells is thicker than with low protein doughs. As the loaf expands in the oven the gluten is thinned to up the point where it is no longer extensible (strain hardened, [218]). In practice dough retains most of the carbon dioxide until the temperature reaches about 72 °C. Only at that temperature large amounts of carbon dioxide are released from the wheat dough/bread. This indicates that the temperature induces the loss of dough's ability to retain carbon dioxide. The stability of gas cell walls during baking is related to strain hardening [219].

The protein content of the flour controls the degree of expansion of the loaf. A higher flour protein content and/or a better flour protein quality allow the production of more expanded and high volume baked goods. Therefore, the baking performance of flours can be improved by incorporating additional vital wheat gluten into the formula. Doughs from nonwheat flours that do not form a continuous viscoelastic gluten ma-

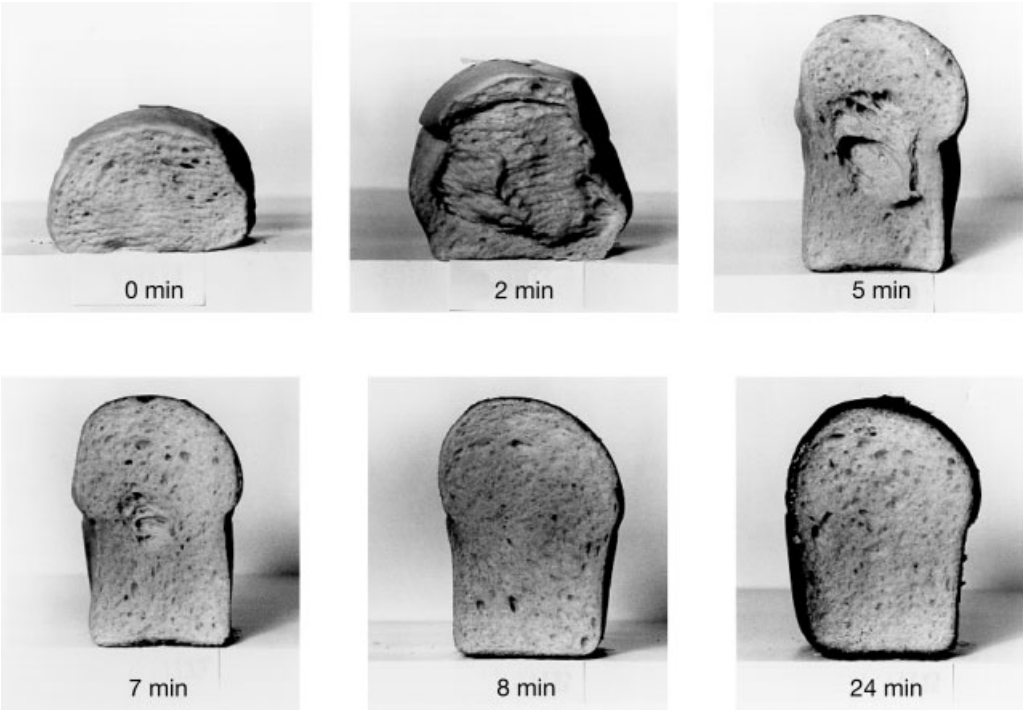

Figure 7. Cross sections of dough and bread before baking (0 min) and after 2 min, 5 min, 7 min, 8 min, and 24 min of baking.

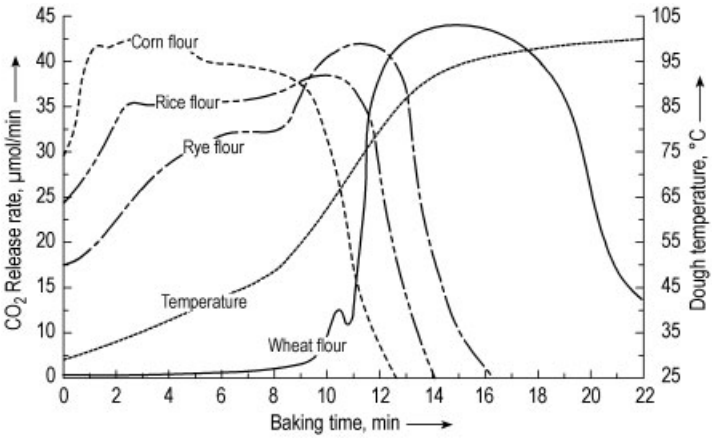

Figure 8. Effect of temperature on the CO_2 release during heating of cereal flour doughs
Dotted line represents temperature increase.

trix are unable to retain carbon dioxide during heating (Fig. 8) and are therefore not suitable for bread making. However, good quality bread can be made from rye flour if an appropriate acidification is used (see Section 9.5). For wheat flour, the relationship between protein content and loaf volume is linear over the range of 8–18 % protein [220]. This effect of flour protein content on loaf volume is almost universally accepted. The proof height (volume) of dough containing different levels of protein is similar but the difference in height (volume) becomes much greater during the early stages of baking [221]. Thus, during the early stages of baking dough containing more protein expands at a much faster rate than dough with less protein. The protein

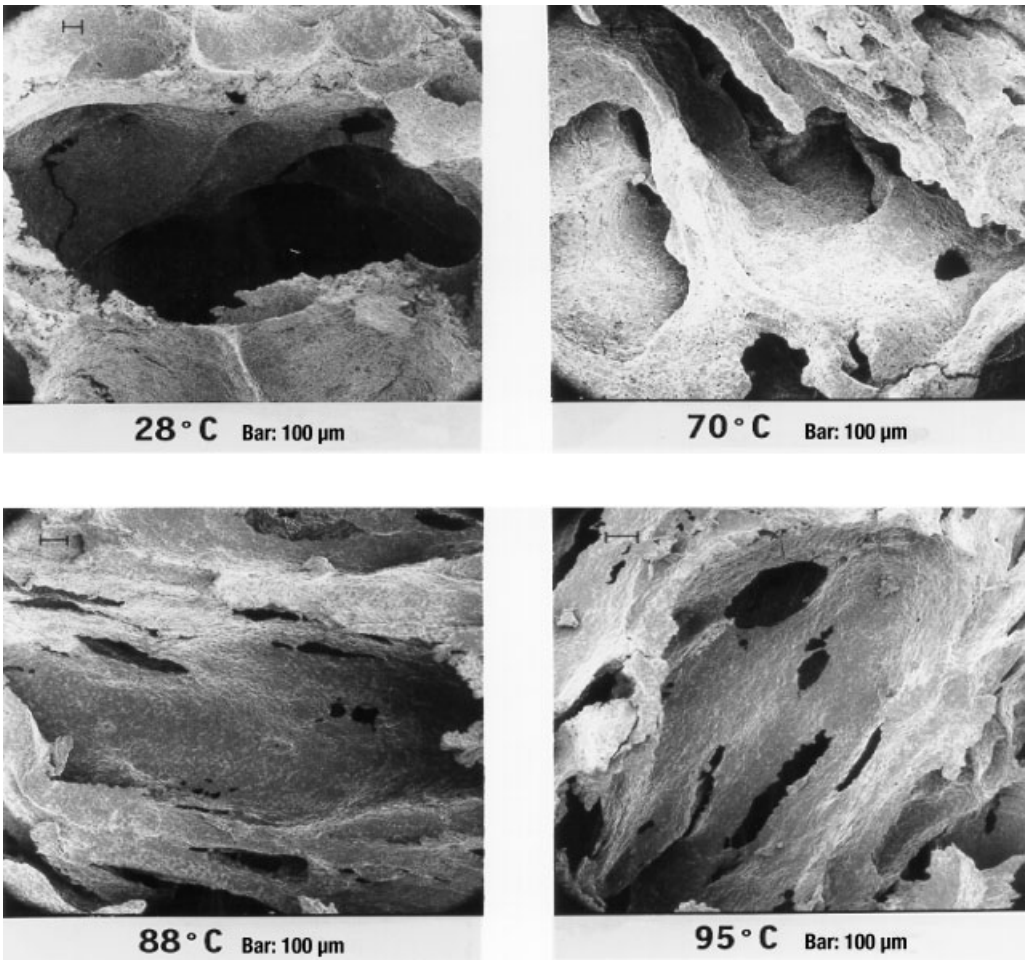

Figure 9. Scanning electron micrographs of the cross sections of dough (28 °C) and bread baked to 70 °C, 88 °C, and 95 °C.

content of the dough does not affect the loss of carbon dioxide during baking if the flour protein content is less than about 8.5 %.

By placing a thermocouple in dough as it baked in the electrical resistance oven it can be seen that the visual transformation from dough to bread occurs at 65 °C. However, the physical properties of dough continue to change it is heated to higher temperatures. Specifically the bread becomes stronger and more elastic. At 95 °C the bread has gone from a liquid to a solid state. The pressure becomes large enough to rupture the gas cell walls and produces the gas continuous sponge structure of baked bread. Scanning electron micrographs of dough/bread heated to below 70 °C show no tears in the gas cell walls while dough/bread heated to above 88 °C shows such tears (Fig. 9). The increase in strength of the bread structure, by itself, is not enough to keep the bread from collapsing or keyholing when it is removed from the oven. The pressure differential developed when the gas cools would cause the loaf to collapse. However, because the bread is gas continuous the pressure differential cannot develop and the bread does not collapse.

Occasionally, reference is found to changing the rate of baking. This is extremely difficult to do. The rate of baking is actually determined by the rate of heat penetration into the dough, and this cannot be changed. Increasing the oven temperature provides a larger gradient of temperature (ΔT) but that only speeds the rate slightly. The higher oven temperature results in larger

temperature and moisture gradients in the loaf but will not greatly change the time required to raise the center of the loaf to the desired temperature. Primarily the so-called Watt principle conveys heat to the interior of the loaf. The Watt principle states that moisture vapor will move to a cool surface where it condenses. In condensing, it gives up its latent heat and warms the cool surface. In a loaf of bread the moisture just under the crust is vaporized and moves toward the interior of the loaf (the cool surface) and condenses. The process is repeated until the center of the loaf reaches the desired temperature (near 100 °C). As a result of this process, immediately after coming from the oven, the center of the baked loaf will contain more moisture than what was present in the original dough. Of course, moisture is also lost from the dough surface during baking. The result is a large gradient of moisture content as the loaf comes from the oven. This gradient slowly disappears with time.

12.2. Crumb Grain Formation

The grain of a loaf of bread is not a random collection of cells but instead an organized system. In an aqueous environment with no added stress, gas cells are always spherical. They then have a minimum free energy and a minimum surface area. Spherical gas cells are clearly seen in cakes or in bread made with the continuous bread-making systems; therefore, there is no external force on the cells in those systems. In conventionally made bread, many of the cells are elongated, indicating that considerable force was put on the cells. The degree of elongation is a measure of the strength of the dough and a direct measure of the chewiness of the bread. With increasing flour strength, a trend toward increased loaf volume, more uniform crumb grain and stronger crumb was found [222].

During molding, the dough is sheeted and then rolled into a cylindrical form that causes elongation of the gas cells. If the dough is strong, the elongation persists through the proof stage. During baking, the crust on the exposed surface of the dough breaks at one or both sides where the crust meets the pan. As the break occurs, the loaf expands in that direction, giving a break and shred (i.e., an area where new dough is exposed to oven temperature as the loaf expands) on the exterior of the loaf and a nice slip plane (i.e., an area where the cells slide past each other as the dough expands) of elongated cells in the interior. If the loaf breaks on the other side, the grain is much less organized and more difficult, if not impossible, to judge. A fully oxidized and developed (matured) loaf has a large number of elongated cells, with a distinct slip plane. If the dough is underoxidized (termed green), many small round cells are evident and the slip plane is not distinct. Overoxidized dough gives loaves with larger round cells with thick cell walls and a very prominent slip plane, often containing large open cells. Thus, the grain of the loaf gives a reading on the oxidation requirement of the flour. In general, a larger volume and a more uniform crumb structure contribute to a softer baked product and improve the crumb softness during storage.

12.3. Crust Formation

When dough is placed in an oven, the surface that is exposed to the oven atmosphere skinsover and forms a crust almost immediately, since skin or crust formation is a result of surface drying. The expansion of the dough during baking is inhibited by the formation of the crust. Then, the moisture on the surface of the dough vaporizes rapidly, as it is in contact with dry, high-temperature air in the oven. The surface of the dough cools as the water vaporizes. If a thicker, heavier crust is desired, steam is generally added to the oven. The steam in the oven slows the rate of water vaporization, and the surface cooks to a greater extent, producing a thicker crust. The formation of a crust or skin has nothing to do with retention of gas, which is controlled by the properties of the interior of the dough and not by the surface.

Bread is a moist product; thus, although the baking temperatures exceed 200 °C, the temperature cannot exceed 100 °C unless pressure builds up in the product or the product becomes dry. The only part of the loaf that becomes dry is the crust, and this occurs late in the baking cycle. This is when browning takes place since the Maillard type browning reactions occur much faster in low moisture systems. Maillard reactions require the simultaneous presence of protein and reducing sugars. Sucrose is a nonre-

ducing sugar, but as stated above, yeast contains invertase, which rapidly produces glucose and fructose, both of which are reducing and brown quite readily. Other reactions in the crust might involve the thermal decomposition of biopolymers (starch, proteins).

12.4. Changes of Flour Constituents

Among the changes that are caused by heating a dough system are starch gelatinization and heat-setting of gluten polypeptides. In a bread dough, the starch starts to gelatinizes at about 65 °C. Gelatinization then proceeds over a wider temperature range dependent on water level/availability, formula ingredients, and baking temperature and time. Small amounts of amylose and degraded amylopectin are leached out of the starch granules and are released into the intergranular phase. Because of the limited amount of water in bread dough, essentially none of the starch is in solution [223]. Amylose can also form inclusion helices with polar lipids such as free fatty acids and monoacyl glycerols. Amylose in form of a single helix–lipid complex can not participate in double helix formation and thus, does not contribute to crumb firmness. This explains, in part, the softening effect of emulsifiers in bread production. Moreover, emulsifiers appear to retard crumb firming via reducing the swelling of starch. In addition, when efficient xylanases are used in the bread making recipe, significant levels of water-unextractable arabinoxylan are solubilized. During baking, the heat also induces the formation of disulfide crosslinks, thereby setting the gluten matrix. This process of gluten protein heat-setting contributes to the formation of the final bread structure and texture [80].

13. Flavor of Baked Products

The flavor of baked products is, together with texture and taste, the key sensory attribute responsible for the attractiveness of baked goods. Baked product flavor is generated from rather flavorless raw materials by two main processing steps, dough fermentation and the baking process. An important color and flavor generating reaction mechanism is the nonenzymatic Maillard reaction involving the reaction of reducing sugars with amino groups during baking [224]. Besides the processing conditions such as fermentation time or the type of microorganisms present in the dough, the recipe (type of flour, flour treatment, added fat, etc.) and storage conditions affect the overall aroma of the final product [225]. Volatile chemical compounds referred to as odorants elicit the overall bread aroma such as roasty–cracker-like, malty, caramel-like, sour, sweet, yeasty, or cake-like. Studies on volatiles that evoke the odor notes of bread in particular have shown that only a small number of volatiles are of significance in determining the flavor [226].

The overall odors of either wheat bread crust or crumb as well as those of rye bread crust or crumb are significantly different. The roasty smelling 2-acetyl-1-pyrroline (ACPY) is the key odorant evoking the pleasant note in the aroma profile of *wheat bread crust*. Ornithine, an amino acid occurring in baker's yeast was identified as the most effective flavor precursor of ACPY. In baguette production, the formation of the key odorants (ACPY and aldehydes) in the crust and thus the flavor profiles were shown to depend on the amount of yeast and the fermentation conditions [227].

In *wheat bread crumb*, three aldehydes and the two alcohols 3-methylbutanol (fermentation-like) and 2-phenylethanol (flowery) are the most potent odor-active compounds. The use of sour dough in wheat bread production leads to an increase in the concentration of these two alcohols and might contribute to the more intense flavor of sour dough- enriched wheat bread. The very low concentration of ACPY in wheat bread crumb might be explained by the nonavailability of the precursor ornithine. It is indeed possible that the high temperatures in the crust induce the disruption of yeast cells and consequently the release of ornithine, and that at lower temperatures in the crumb ornithin remains in the yeast cells.

ACPY disappears very rapidly after baking. This loss together with the formation of higher-stability odorants by lipid peroxidation, e.g., (*E*)-2-nonenal is involved in the formation of the stale off-flavor perceivable during storage of wheat bread.

In contrast to wheat bread crust, *rye bread crust* is very low in ACPY and much higher in

3-methylbutanal, methional, furanone and acetic acid. In *rye bread crumb*, phenylacetaldehyde, which is not present among the key odorants of wheat bread or rye bread crust, is an important odor active compound.

14. Glass Transition and its Role in Baking

The texture of baked products is strongly affected by their glass transition temperature (T_g). Regions in solid food products such as baked goods can exist in either a thermodynamically stable crystalline or in an amorphous, nonequilibrium state. In trying to understand the behavior of materials such as polymers (e.g., proteins, starch) and low molecular mass components such as sugar in their amorphous state, one of the most important properties to know is T_g [7, 228 – 238]. The T_g markedly determines many of the physical properties of a material. (For a definition of T_g, see → Plastics, General Survey, Section 6.2.5)

The overall concept of T_g and phase transitions can be summarized in the generalized 'state diagram' as shown in Figure 10. The state diagram defines the moisture content and temperature region at which a food domain shows glassy, rubbery or flow behavior. A food region composed of polymeric material has a higher T_g than a food region composed of lower molecular mass components. The level of plasticizer (e.g., water), along with the temperature, determines what state the region is in. The higher the concentration of plasticizer (water) in a food product, the lower is the glass transition temperature. Below the T_g curve, reactions occur very slowly because the food is in a glassy state and has a very high viscosity [239 – 241]. Molecular movement is decreased at this high viscosity except for small molecules such as oxygen or water vapor that can diffuse through larger pores. Above the T_g curve, viscosity decreases significantly. This allows increased movement of molecules, resulting in a soft/rubbery texture or, for example, in the rapid crystallization of sugar. From this plot it can be seen that if a product starts at point A in the glassy/crispy region and the temperature of the product is raised (Fig. 10 A → B), it will first become leathery, then rubbery, and eventually will flow. For example, for a bread with 37.4 % moisture a T_g of − 12 °C has been reported [242]. Softening starts at − 12 °C and then progresses continuously up to 100 °C. Likewise if the moisture content increases, the product can also go from glassy to leathery and rubbery (Fig. 10 A → C) even if the temperature remains constant.

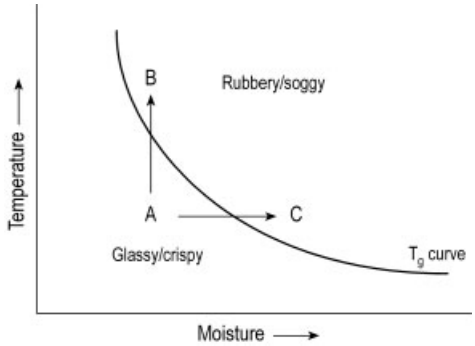

Figure 10. Idealized state diagram of a cereal material as a function of temperature and moisture
A is the assumed starting point of a crispy baked product showing a glassy/crispy to rubbery/soggy phase transition as a result of a temperature (A → B) or moisture (A → C) increase

These alterations in moisture and temperature can have positive and negative effects on consumer acceptance of texture and there are many examples of glass transitions in baked products: crispy crust formation in bread due to dehydration during baking; toasting bread drives moisture out and restores the glassy state to the crust and, indeed, to the rest of the bread if sufficient moisture is removed; pizza crust becomes soggy in the presence of sauce; a baked cone filled with ice cream becomes less crisp over time; saltine crackers, that are crispy at low moisture contents, become rubbery and soggy when absorbing moisture from other adjacent food components such as cheese, jam, or a fruit paste of high moisture activity; cookies become rubbery when taking up moisture during storage or chocolate chip cookies change from a soft desired texture into a hard undesirable texture because of sugar crystallization over time; staling of bread is associated with the nonequilibrium phenomenon of starch recrystallization during storage (see Section 17.1).

15. Bread Varieties and Speciality Breads

Today, a great variety of speciality breads is available in the Western world. The great preponderance of the bread produced in the United States is white pan bread. However, in the 1990s, nonwhite or speciality bread has increased rapidly in popularity.

In the United States white breads that are not baked in a pan are also speciality breads. Included in this type are hearth breads of the French, Italian, and Vienna types and a number of hard rolls. Other than the obvious difference of being baked free-standing on a hearth rather than supported in a pan, the difference between pan bread and these loaves as made in the U.S. is not large. In general, steam is used in the baking oven to produce a thicker crust. In some shops, the formula is somewhat leaner (lower in fat, sugar, and milk) but not nearly as lean as the flour, water, salt, and yeast loaves of their European namesakes.

Wheat breads, breads made from 100 % whole wheats, as well as combinations of whole wheat and white flour are also types of breads that are rapidly growing in significance. To be labeled as a whole-wheat loaf in the United States or Germany, the bread must be made from whole-wheat flour of specified granulation derived from the whole wheat kernel. Loaves of this type are smaller and more compact, with a high density and lower volume. Most wheat bread produced in North America contains partly whole-wheat meal and partly strong white flour. This results in a much better loaf volume and a lighter-tasting product. Many bakers also add vital wheat gluten to such formulas to aid in improving the volume. The formula, other than the type of flour, does not vary greatly from that used for white pan bread.

In Europe, bread production techniques and specialties are very diverse [243]. The selection of bread available is mainly based on the use of wheat and rye flours. At the same time, barley and oats have always had a considerable role in the production of baked goods, originally due to the scarcity of wheat and rye, nowadays to provide more variety in breads. The cultivation of rye plays an essential role in European bread making and has contributed significantly to the development of various types of bread. In the North America bread made from 100 % rye flour is rarely seen. Most rye bread sold in the USA is made with white wheat flour with enough added rye flour to attain a desired flavor and appearance.

Every country has its own food specialties and in the following, France, Germany and the United Kingdom are briefly presented as examples of the European bread culture [244, 245].

France. In France, about 80 % of the bread is still produced in small bakeries. The most common traditionally eaten bread is the baguette. It represents more than 50 % of the French bread production. The baguette is a white wheat bread characterized by a long, thin, crusty loaf. It should be fresh and crunchy and the shelf life varies between 4 and 6 h. The proportion of the crust comprises 40–50 % of the total bread and crumb has an open grain and excellent flavor. Baguette is produced with four basic materials: flour, water, salt and yeast. Bread improvers might be added to correct possible flour deficiencies: lipoxygenase-active bean flour to improve whiteness, ascorbic acid to strengthen the dough, malt to provide α-amylase activity, and lecithin to improve gas retention. Other French speciality breads include the fine wheat bread (pain de gruau), the Viennese bread, and biscottes.

Germany. Baked products are divided into two types: the production of bread allows a maximum of 10 parts sugar and/or fat to 90 parts of flour; products with more than 10 parts sugar and/or fat to 90 parts of flour are called fine bakery wares (Feine Backwaren) [243, 244]. Bread is divided into five main groups: wheat breads (at least 90 % wheat), mixed wheat–rye breads (min. 50 % wheat), mixed rye–wheat breads (min. 50 % rye), rye breads (at least 90 % rye), and bread specialties. Each of the first four groups was established based on the proportion of wheat and rye in the formula. These groups are further classified into subgroups according to the type of milled raw materials used: low extraction flour bread, meal bread, and wholemeal bread. Speciality breads can be prepared, for example, by adding nonbread grains (e.g., oat, barley, rice, maize), raw materials of plant origin (e.g., oilseeds, germs, raisins), materials of animal origin (e.g., milk, butter, yogurt, whey) or by using special baking techniques (e.g., wood-heated

oven, steam oven, stone oven, crisp bread). From the total bread consumption, about 15 % is eaten as small rolls and its consumption is further increasing. The so-called nonbread grain breads have become very popular. Also the production of dark breads such as wholemeal flour breads is on the rise.

United Kingdom. The typical UK sandwich-bread has a high volume, soft texture, very fine porous crumb structure and is also characterized by long shelf live properties. Most of the bread in the UK is produced with the Chorleywood bread process that includes a high-speed mixing of the dough in the presence of an oxidizing agent. Speciality breads comprises germ breads with the addition of 10–25 % wheat germs, high protein bread with wheat gluten, milk, or sometimes soy protein. Malt bread produced in the United Kingdom is a sticky, sweet, dark loaf that might also contain dried fruit pieces. UK rye bread is usually made from a mixture (50 : 50) of white wheat flour and rye flour.

Sour dough breads, both rye and wheat, are popular in Europe and certain areas of the United States. The U.S. San Francisco sour dough bread is made from white wheat flour using sour dough as leavening agent. Typical sour dough breads such as rye bread or mixed rye–wheat bread are often produced as round loaves and baked in ovens with steam to give a heavy crust. The sour dough technology is also applied to typical Italian baked products such as Panettone, Pandoro, Veneziana cakes, Pugliese bread, and brioches as well as to certain gingerbread cookies and crackers. Depending on the level of sour dough used, baked products made with sour dough are characterized by a sour-aromatic taste and better shelf life properties.

Flat Breads. Another major use of wheat flour is in the production of flat breads [246 – 248]. In Central America, North Africa, the Middle East, the Southern States of the former Soviet Union, India, and in Scandinavia these breads represent a major source of calories and protein. Flat breads are prepared from one flour or a flour mixture made from wheat, corn, rye, barley, sorghum, millet, or rice grain. Based on the cross-section of these products, they can be divided into two groups, single-layered and double-layered. Single layered flat breads include tortillas, chapati, foccacia, naan, pizza crust, and Scandinavian crisp bread. Wheat tortillas have exhibited increased consumption and market growth in the United States and large-scale commercial production has become common in Mexico and parts of the United States [249, 250]. Typical examples of double-layered flat breads are Arabic (pita) and baladi bread.

Steamed Bread. In some parts of Asia, especially in northern China, steamed bread is a staple item in the diet [251]. A large part of the country's domestic and imported wheat is used to produce steamed bread. Steamed bread is prepared by cooking or steaming a fermented dough. Therefore, steamed bread does not have, in contrast to Western-style breads, a crust but rather a soft, smooth, thin skin. The color of the external skin and internal crumb should be very white. Steamed bread has a symmetrical, well-rounded shape and a moist, elastic, springy crumb texture and is eaten as plain bread or filled with meat, vegetables, or sweet ingredients.

16. Soft Wheat Products

The diversity of products made from soft wheat flour (North American terminology) is amazing. Unlike hard wheat flour (same terminology), which is used for one major product, bread, and numerous minor ones, soft wheat has more than one major use. The baked products made from soft wheat flour can be grouped as follows, however, with a wide variation within each of these groups: (1) biscuits/cookies, crackers, and wafers, (2) cakes, (3) pastry products from pie crusts to puff pastries, and (4) waffles, pancakes, and donuts. According to this product variation, large differences exist in formulas (Table 1) and processing technologies. In the following, the discussion will concentrate on the biscuits/cookies, crackers, wafers, and pastry products to explain the technology and underlying principles of their production. Other nonbaked product applications for soft wheat flour include oriental noodles, thickeners for soups and soup mixes, manufacture of crumb for coating fish and meat products, and as the basis for

some breakfast cereals. All of the above products have better appearance and eating quality when made from soft rather than hard wheat flour. Soft wheat is used in these applications because (1) the stronger gluten and higher protein levels of hard wheat would reduce the quality of the above mentioned products, (2) soft wheat flours have lower levels of damaged starch and therefore lower water absorption and viscosity, and (3) soft wheat flours have a finer texture and smaller particle size distribution [10, 112, 252, 253].

Cookies/biscuits, crackers, and pastries contain generally high percentages of sugar and/or fat relative to flour and a relatively small percentage of water. This overall composition separates this group of baked products from breads, that contain relatively low levels of sugar and fat and intermediate levels of water, and from cake-type products that have high levels of fat and sugar, but also water. The low moisture content of cookies/biscuits and crackers confers a long shelf life to these products and also gives them a relatively high caloric density. Because of these properties, biscuits and crackers have a firm place in the diet of some parts of the world such as in South America. They are staple food and eaten often as bread replacer. Within the various types of cookies and crackers, large differences in composition can occur.

16.1. Technology of Cookie and Cracker Production [7, 254 – 258]

Ingredients for biscuit and cracker making include dry (e.g., wheat flour, sugar, starches), liquid (e.g., fats and oils, syrups) and minor ingredients (e.g., leavening agents, emulsifiers). Most large biscuit factories receive and store their major ingredients in bulk, i.e., not packed but in containers of a given size. The ingredients are pre-processed (sieving, filtering, blending) and then metered automatically into the dough mixer.

The majority of dough mixers used for the preparation of cookie and cracker doughs fall into one of three categories; i.e., vertical spindle mixers, high-speed mixers (horizontal drum mixers) and continuous mixers. *Vertical spindle mixers* operate at slow speed and therefore have a gentle mixing action. This type of mixer is widely used in the USA for mixing cracker doughs. *High-speed mixers* are widely used in modern cookie bakeries and are particularly useful for mixing stiff, tight doughs that require considerable amounts of work input. *Continuous mixers* are mainly used in large, dedicated production lines. After mixing, doughs are left to relax or ferment from 30–45 min for cookie doughs, to 2–3 h for chemically leavened crackers doughs and up to 24 h for yeast fermented cracker doughs. The dough then goes to a forming stage prior to baking.

Large scale biscuit and cracker baking is carried out in band ovens (traveling oven) varying in length from about 30 to about 150 m. The products travel through an oven tunnel on a continuous band of solid steel or a woven wire mesh which may be up to 1.2 m wide. The tunnel is heated internally either directly (e.g., by gas or electric heaters within the baking chamber–direct gas fire) or indirectly by hot air circulated through the oven from an external source (convection–indirect ovens) or by a combination of the two systems. Baking times vary from about 1 min to about 15 min and baking temperatures from about 200 °C to 240 °C according to the product. The completion of the baking process is generally judged by two product properties: its color and its moisture content. To better control process conditions and final product properties of cookies and crackers, the baking process has been studied by mathematical modeling [259, 260]. This approach takes into account parameters such as oven and cookie temperature, baking time, throughput, turbulence, humidity of the oven atmosphere, properties of dough pieces and cookies (e.g., cookie dimensional data, moisture, color) to build models for oven design and optimized and standardized finished product quality.

After baking, cookies/biscuits and crackers are transferred to a cooling band prior to being packed. In some cases, cookies/biscuits are subjected to further secondary processes after cooling such as coating or sandwiching (e.g., with chocolate, fat-based creams, jam, jelly) to achieve a greater variety of flavors, textures, and appearance.

Table 1. Average dough recipes of soft wheat products: parts of ingredient based on 100 parts of flour (by weight)

	Cookie	Cake	Pound cake	Short pasty	Puff pasty	Croissants
Flour	100	100	100	100	100	100
Dough fat	30	50	100	45	20	15
Lamininating fat	–	–	–	–	80	45
Water	25	140	–	23	52	58
Sugar	50	11	100	< 70	–	3
Others (e.g., egg, milk)	100	17	100	–	–	10

16.2. Cookies

In general, cookies are characterized, by a formula high in sugar and fat/shortening and relatively low in water. The products called "cookies" in the United States correspond to similar products called "biscuits" in Europe and the United Kingdom. The biscuits made in the United States are more accurately defined as chemically leavened bread. The diversity of cookie products is quite wide; cookies as well as crackers not only vary in formula but also in type of manufacture (Fig. 11) [258].

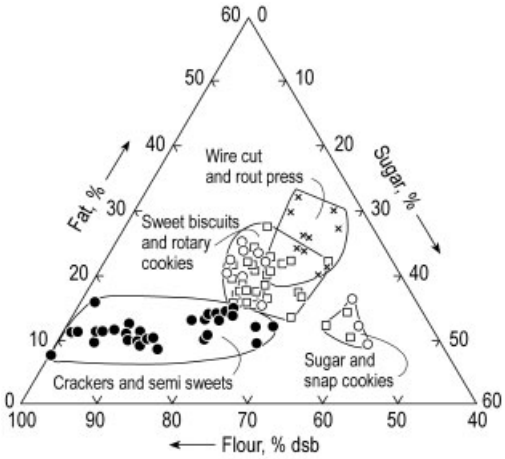

Figure 11. Composition of various types of cookies/biscuits and crackers
Proportions of flour solids, fat, and total dry sugars are expressed as percentages of the dry product. Hard doughs: •; short doughs, sheeted and cut: O; rotary molded: □; soft doughs, wire-cut: x [258]

In *soft cookie* doughs, in general, gluten should not be developed during mixing. Development of the gluten leads to tougher products and generally to cookies that will not spread. The retarding of gluten development is not much of a problem. The low water content (about 10–20 %), the high sugar and high fat content in most cookie formulas result in a dough in which the gluten is not developed since the hydration of protein is insufficient to allow gluten network formation. Therefore, the dough is of soft consistency and has little or no cohesive properties. In the United Kingdom, such dough is termed a *short dough* and or *short sweet dough* and is related to cake formulations except that it contains much less water. The majority of these short doughs are processed into the desired shapes by compression into dies engraved on a designed roller (rotary-mold cookies) or by extrusion and subsequent cutting to the required size by an oscillating wire (wire-cut cookies) or by a guillotine (bar-press cookies) (see Section 16.2.1).

The other basic type of cookie dough is referred to as *hard dough*. The hard doughs are related to bread doughs but are of much stiffer consistency and give products that are somewhat tougher with a firmer bite than soft dough products. Hard doughs contain more water (about 20–30 %) and less sugar than soft doughs, thereby allowing gluten development during mixing and subsequent processing. Therefore, these doughs exhibit, like bread doughs, viscoelastic properties and require rolling with heavy steel rolls to form dough sheets of the required thickness before the dough is cut out from it (cutting machine cookies: e.g., rotary-cut cookies). The gluten development stops spreading during baking of this type of cookie.

16.2.1. Cookie Types

Rotary-Mold Cookies. For this type of cookie, the dough is forced into molds on a rotating roll. As the roll completes a half turn, the dough is extracted from the cavity and placed on the band for baking. The consistency of the dough must be such that it will feed into the cavity with no voids but still be extracted from the cavity without being distorted. Formulations for rotary-mold cookies are characterized by fairly

high sugar and shortening levels and very low levels of water (< 20 % based on the flour and including the moisture in the flour). The typical dough is crumbly, lumpy, and stiff, with virtually no elasticity as the gluten in the dough does not develop during mixing. Therefore, the cookie does not rise during baking and the design embossed on the cookie will not be distorted. Much of the cohesiveness of this type of dough comes from the plastic shortening used. Rotary-mold cookies are economical to produce. A small amount of water is added to the dough and, therefore, less energy is required to remove it during baking.

Wire-Cut Cookies. A relatively soft dough is extruded through an orifice and cut to size, usually by a reciprocating wire. The dough must be cohesive enough to hold together, yet short enough to separate cleanly when cut by the wire. A typical formulation, based on flour weight, may contain 50–75 % sugar, 50–60 % shortening, and up to 15 % eggs. Wire-cut machines can handle a range of cookie doughs, and a wide variety of products can be made. Wire-cut cookies rise and spread or flow during the baking process and they are of much softer consistency than rotary mold cookies. The final size of the cookie is determined by the formula and the flour used. In addition to such common examples of this cookie type like sugar-snap, chocolate-chip, or oatmeal cookies, also fig or date bars can be coextruded and subsequently wire-cut.

Cutting-Machine Cookies. The process for this type of cookie includes both rotary-cut and stamped doughs. The dough is made into a continuous sheet and the product cut out from it. Typical examples are animal shaped cookies, gingerbread men, etc. The formula contains much more water than do formulas used with a rotary mold. The sugar content is relatively low compared to most other types of cookies. Because a sheet of dough is made, the gluten is developed in this system. The gluten development stops spread during baking of this type of cookie.

16.2.2. Underlying Mechanisms of Cookie Baking

In general, cookie quality includes (1) the size of the cookie, both the width and the height, (2) the eating quality and mouthfeel (i.e., cookie bite, good quality cookies must have a tender bite), and (3) the appearance including color and surface cracking (top grain).

To evaluate the effects of flour properties, formula ingredients and process parameters on cookie quality a cookie must be baked. Two American Association of Cereal Chemists (AACC) cookie formulations — the wire-cut formula and the sugar-snap formula — were developed to study the effects of flour quality (e.g., wheat cultivar differences) on cookie properties [261]. The most widely used baking procedure to evaluate soft wheat flours is the AACC sugar-snap cookie baking test. It is reproducible and very sensitive to changes in both flours and formulations. A prediction equation for sugar-snap cookie diameter that uses sucrose solvent retention capacity, wheat milling softness and flour protein content has been reported to be useful [262].

When a cookie dough is heated in an oven, a number of events occur that are related to structural and dimensional changes, loss of moisture, and the development of color and flavor [7, 224, 258]. These changes occur in three successive, albeit overlapping stages. In the first stage the dough piece increases in thickness due to aeration and moisture loss commences. In the second stage the product "sets", i.e., cookie spread stops and moisture loss reaches its maximum. In the third stage, coloring of the cookies occurs. Thermal events that occur during heating sugar-snap cookie dough have been attributed to the melting of fat, and the dissolution of sugar [7]. First, the fat (shortening) in the dough melts. Thus, the dough containing melted shortening is more free to flow under the force of gravity. The second thermal event is dissolution of the sugar. The formula for the simple sugar snap cookie is high in both sugar (60 % of flour weight) and shortening and relatively low in water (about 24 % of flour weight). As a result of those ratios and the creaming step (mixing the sugar and shortening) only about half the sugar is dissolved during mixing; the remainder stays in a crystalline form until the dough is heated. Then that sugar

dissolves. When sugar dissolves, it increases the volume of solution in the system. Each gram of sugar, when dissolved in 1.0 g of water, produces about 1.6 cm^3 of solution. The increase in total solution also has a pronounced effect on the cookie dough. The increased solvent content makes it sticky. Likewise, cookie dough prepared with sugar syrup rather than crystalline sucrose is sticky and does not machine well.

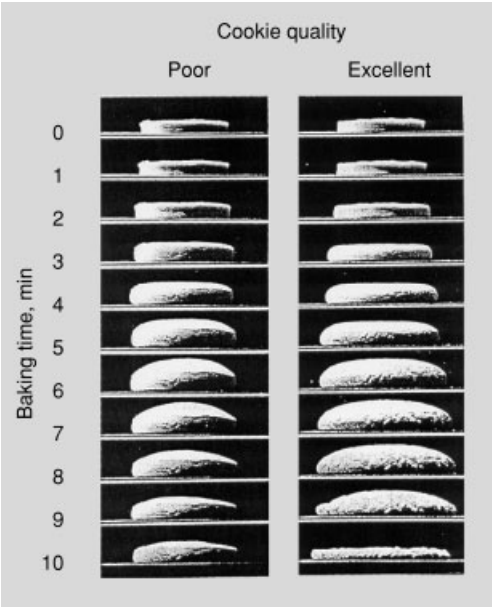

Figure 12. Time-lapse photographs of cookies made from excellent and poor-quality cookie flour (reprinted from [265]).

Time-lapse photography has been used to monitor the kinetics of changes in visual appearance (e.g., surface cracking) and geometry such as the development of cookie spread during the cookie baking process [263 – 265]. As demonstrated in Figure 12, the lateral and vertical dimensional changes of sugar-snap cookies during baking are markedly affected by the flour quality. A poor cookie flour (i.e., hard wheat flour with high protein content $> 10\,\%$) and an excellent cookie flour (i.e., soft wheat flour with low protein content, ca: 8–9 %) was used for cookie dough preparation. Early baking behavior (1–2 min) is similar for both poor- and excellent cookie flour doughs. Both types of cookies expand in width and height due to the production of leavening gases from the sodium–ammonium bicarbonate leavening system. By the third minute of the baking cycle, the cookie baked from excellent cookie flour spreads more rapidly and continues to spread for a longer period during baking than the cookie from the poor cookie flour. The cookie made from poor cookie flour is capable of withstanding collapse during baking [264]. In contrast, cookies made from excellent cookie flours cannot support their own weight against uncontrolled, structural collapse under gravitational force. The expansion (increased cookie diameter) and decreased cookie height are accompanied by progressive structural collapse throughout the baking cycle with the most dramatic phase of collapse occurring in the last minute of the 10 min baking cycle.

Cookie expansion and lateral flow have been related to the viscosity of the system. At baking temperatures the cookie dough made from excellent cookie flour has a lower viscosity than the cookie dough made from poor cookie flour and thus shows more expansion due to of leavening and faster flow under gravity. Since both, gravity and the amount of leavening agents are constant, the question arises of what are the underlying mechanisms of the viscosity of cookie doughs are. While starch and soluble starch in the wheat flour may affect cookie spread [263, 266], starch gelatinization can be ruled out as contribution to the different viscosities of the two cookie systems. Due to the low water and high sugar content of the cookie starch does not gelatinize or only to a limited extent. It appears that the viscosity changes during baking are related to the wheat proteins. The wheat proteins in a sugar-snap cookie formula are not developed into a continuous gluten network during mixing. When the dough is heated during baking, the wheat gluten proteins go through their glass transition (see Chap. 14), thereby gaining mobility that allows them to interact and to form a continuous gluten phase. The viscosity of this continuous gluten phase is sufficient to stop the flow of the cookie dough [267]. In cookie doughs from poor, i.e., hard wheat flours, this gluten network appears to be sufficiently strong to account for the resistance to expansion and flow (i.e., spread) during baking and the prevention of structural collapse [145, 147].

The overall importance of bread has favored the breeding and growing of harder wheat flours with higher protein contents. The proteins of

these hard wheat flours, specifically the high molecular mass glutenins, that form a strong gluten network are detrimental to the baking of cookies and crackers. Consequently, sulfite compounds (e.g., sodium metabisulfite $Na_2S_2O_5$) have become widespread in the cookie manufacture to rapidly break down wheat proteins by rupture of disulfide bonds. From the sulfite compounds added, traces may remain in the final product. This has led to a ban of these chemicals in most countries and to investigations into alternative ways of controlling dough rheology such as the use of proteolytic enzymes.

Apart from gluten structural effects on viscosity, sucrose as a major ingredient in the cookie system is likely to play a role in the differences of cookie diameter when hard and soft flours are used for cookie baking [238, 263, 264]. Sucrose can act as plasticizer but it is much less effective than water due to its larger molecular mass. Antiplasticizing agents increase the glass transition temperature (see Chap. 14). Therefore, in comparison to water, sucrose may have an effect by acting as an antiplasticizer. The T_g of a cookie dough made from excellent (soft, low protein) flour is raised to higher temperatures than the T_g of a cookie dough made from poor (hard, high protein) flour. Since setting, i.e., stop of cookie spread, is possible when the cookie system goes through T_g and thus into the flexible/rubbery region (where gluten proteins can form a network that increases the viscosity to a degree sufficient to stop flow of the dough), doughs from excellent wheats would have more time to spread, resulting in a larger cookie diameter [267].

16.2.2.1. Texture Development of Cookies

A cookie can be described as a brittle matrix in which gas bubbles of various size and shape are embedded. The mechanical properties and fracture behaviors of a cookie are affected by the volume fraction and distribution of each phase and by interactions between the various phases [268 – 270]. Recipe (Fig. 11) and process conditions strongly affect the textural properties of the final cookie product.

The nice tender bite that is associated with good quality cookies appears to be related to several factors:

1) The air volume fraction and the degree of inhomogeneities of the cookie (e.g., size and shape of gas cells, material thickness between gas cells) determine its textural properties. Fracture starts from large size defects (i.e., gas cells) [268]. This crack initiation is then followed by crack propagation.
2) The use of fat or shortening produces a short product (see Section 16.7.1, the word "shortening" came from this usage). Reduction of the fat content is generally accompanied by an increase in modulus of elasticity and fracture stress. In the brittle cookie matrix, fat acts as a filler at low concentrations (about 16 wt % fat) (fat dispersed system) or forms a partly continuous phase at high concentrations (about 28 wt %) (fat continuous system) [269]. Fat also affects the flavor release and delivery of a cookie.
3) The flour strongly impacts the textural properties of a cookie. Generally, a soft wheat flour gives products with a tender bite.
4) Type and physical state (glassy/crystalline) of sucrose: Essentially all cookies, except those that are dried to a very low moisture content, are soft and quite flexible when they come out of the oven. With time, they become firm and often brittle. This appears to be related to the crystallization of sucrose. When the sugar dissolves, the syrup gives the cookie flexibility. In order to maintain a soft cookie, only a sugar or other substance needs to be added that interferes with sucrose recrystallization.
5) The moisture content markedly affects cookie texture. The final moisture content of a freshly baked cookie is usually within the range of 1–5 % that is associated with a water activity (α_w) of about 0.1. This residual moisture however, is not uniformly distributed throughout the product. A moisture gradient exists with the highest moisture content occurring in a thin lamella near the product center and the lowest moisture content occurs at the periphery and product surface. This moisture gradient disappears by migration and redistribution of moisture from the center to the surface of the product during cooling and storage. Firm cookies, i.e., snap-cookies such as sugar- and ginger-snaps, are baked to lower moisture content than are soft cookies. For a cookie to remain soft it must have a relatively high moisture content after baking.

Using two doughs that are kept separate can make a dual-texture cookie. One contains sucrose and the other a crystallization inhibitor (corn syrup). To maintain softness, hygroscopic sugars such as glucose and fructose are very effective. To maintain the crisp or soft texture, cookies must be stored under low or high relative humidity, respectively, which requires appropriate packaging material.

16.2.2.2. Development of the Cookie Surface Pattern

For certain cookies such as sugar-snap cookies and gingersnaps, the development of a cracking pattern on the surface called 'top grain' is desired. This cracking pattern develops during baking and the phenomenon has been explained as follows. Moisture is lost from the surface of the cookie at a rapid rate during baking. The hot air in the oven has the capacity to hold a large level of water. As the water from the surface is lost, it is replaced by water diffusing from the interior of the cookie, and the sugar, not being volatile, is concentrated. Then, sugar will crystallize at the surface of the cookie and will no longer hold the water that gives a moist and moldable surface. Thus, the surface dries and as the leavening system expands the cookie the surface breaks to form the typical cracked pattern. Small levels of high-fructose corn syrup or certain other sugars interfere with sucrose crystallizing and thus destroy the cracking ability. It has also been proposed that the cookie collapse mechanism is associated with the increase of surface cracking, and thus, the development of top grain [264].

16.3. Crackers

As with cookies, the definition of crackers is quite broad. In general, crackers contain little or no sugar but moderate to high (10–20 %) levels of fat, based on flour weight (Fig. 11). The doughs generally contain low levels of water (20–30 %). In the USA the term cracker is applied to products made from flour and fat with contents of sugar varying from zero to about 40 % (flour weight basis) and includes at one extreme saltines and, at the other, products such as honey Grahams. In the UK the term cracker is traditionally reserved for products made from flour and fat with little, if any, added sugar. These products include cream crackers (made from fermented dough) and water biscuits. Another group of products known as semisweet biscuits, contain about 20–30 % sugar on flour weight basis. Although the formula of these products in terms of flour, fat, and sugar is not dissimilar to that of Graham type crackers the appearance and eating qualities of these UK products — which include rich tea, morning coffee, thin arrowroot, etc., crackers — bear no resemblance to American crackers.

Crackers are often prepared from yeast-fermented doughs. The best-known crackers made from fermented doughs in the USA are soda crackers, or saltines. The corresponding product in the UK is the cream cracker. Compared to cream crackers, soda crackers contain less fat and are slightly alkaline (pH > 7.0) due to the higher amount of sodium bicarbonate in the soda cracker formulation. Although very different in appearance and eating quality both products are usually made by a sponge and dough process with long fermentation times (up to 24 h total). These long fermentation times are associated with the specific texture and flavor development of these products. Usually, the 'sponge' made from flour, yeast and water is fermented for about 18 h. This fermentation step involves the growth of yeast added in the formulation and several species of lactic acid bacteria. Thereby, acidity is formed and the pH drops from about 6 to 4. At pH 4, a proteolytic enzyme present in flour has its optimum and appears to modify dough proteins so as to become less elastic.

After its fermentation, the sponge is mixed with fat, salt, sodium bicarbonate and the rest of the flour to form a dough which is allowed to ferment for further 5 h. The level of sodium bicarbonate in soda crackers doughs is sufficient to bring the dough pH to above 7. After 24 h the dough is ready to be processed. There is a trend to shorten the fermentation time but keeping the quality advantages of the sponge dough procedure. The application of liquid fermentation (lactic acid bacteria, yeast culture, flour and/or flour fractions, and water) has allowed to shorten the fermentation time to < 4 h and to improve

process control, to reduce floor space and capital investment (see Section 9.3).

After laminating, layering and sheeting the cracker dough is cut and docked. The purpose of docking is to pin the dough together (attach the top and bottom surfaces together) so it will not separate into layers. After docking, the dough is salted and baked for about 2.5 min at 260 °C to a final moisture content of about 2 %. The rapid heating during baking is accompanied by a rapid increase in water vapor pressure that causes puffing and separation of the layers of the cracker between the points tied together by the docking holes. Special packaging is required to maintain the low moisture level of the product, as even small increases in moisture decrease the desirability of the cracker.

Other cracker types include snack crackers that vary much more widely in formula than do soda crackers. In general, snack crackers do not contain yeast and are chemically leavened. They contain more shortening and much higher levels of flavoring materials. Often, the texture of snack crackers is denser than that of soda crackers.

16.4. Wafers

Wafers and wafer products are generally regarded by the consumer as biscuits and biscuit companies frequently manufacture them. Wafer making, however, has little in common with cookie or cracker formulations and process technology. The principle ingredients in wafer-making are wheat flour and water. Minor ingredients comprise small amounts of oil and/or lecithin, salt, sugar and aerating agents (sodium/ ammonium bicarbonate for nonfermented batters or yeast for fermented batters) that affect texture, flavor and color of the final wafer. The major function of bicarbonate is to adjust the pH, not to act as a leavening agent. The ingredients are mixed under high-speed to a fluid, homogenous batter containing up to 65 % water. The batter is distributed to the wafer-baking plate via a depositor. If gluten aggregates, strands and lumps are formed during batter mixing and pumping, they may cause blockages in the pipework, metering pumps, and batter depositors. Gluten aggregation can be controlled by the selection of flour quality [271] and adjusting process conditions (batter temperature, mixing speed). Often, commercial enzymes comprising a mixture of protease and arabinoxylanases are used in wafer manufacture to control batter viscosity and to avoid gluten aggregation (see Section page 15).

After batter depositing, the wafer is baked between a pair of heated, engraved metal plates for 1.5 to 3 min at about 180–200 °C. During wafer baking starch completely gelatinizes in the high-moisture environment. The water in the batter evaporates from the closed-plate system as steam through vent channels making water vapor the main aerating agent. At the end of the baking process, a dried (1–3 % moisture), highly aerated starch gel is formed with wheat protein interspersed in the starch matrix.

After baking, a filling is placed between the various wafer sheets followed by cutting the layers into "books". The filling generally consists of fat, sugar, flavors, and ground scrap product (rework) originating from wafer cutting. Variations in moisture contents among the wafers result in moisture transfers and a concomitant contraction and expansion of the wafer. This change in geometry can weaken the adhesion between the wafer and the filling or lead to cracking of a chocolate coating. Therefore, wafers are conditioned or matured, i.e., equilibrated under controlled atmosphere to ensure equilibration of the moisture in the wafer sheets before further processing.

Sugar wafers are made from batters containing a high level of sucrose (up to 60 % on flour weight). Such products, while still hot from the baking process, are quite flexible and may be rolled round a mandrel to form ice cream cones or folded to form a decorative fan or other shaped product that become firm and very brittle upon cooling.

16.5. American Biscuits

The American biscuit, actually a chemically leavened bread or bun, is typical to the United States and has only recently been introduced into the European market. In the USA it has become quite popular, particularly in fast-food establishments. The use of chemical leavening gives dough that results in a rather thick cell wall and a coarse grain in the final product. The flavor is strongly influenced by the taste of the soda

and the leavening acid. The production process is quite simple: the dough is mixed, sheeted to a desired thickness, cut, and baked.

A variation of the chemically leavened biscuit is the type sold refrigerated and packed in cans. The cans are actually foil-lined cardboard containers. The dough is mixed, sheeted, cut to size, and placed in the can. The volume of dough added to the can is much less than the volume of the can. The ratio of the two volumes must be carefully controlled. After the cans are sealed, they are placed in a proof box at a temperature high enough to trigger the leavening system but low enough not to hurt the dough properties. The dough then "proofs" i.e., expands to fill and caulk the can. The excess air present in the can is expressed through the can, which is permeable to air but not to dough. The dough builds a pressure of about 100 kPa (15 psi) and in this condition is quite stable for reasonable periods of time (60–90 d). Because the dough is not sterile, organisms will eventually destroy the quality of the biscuit. The bacteria can cause the pH drop, the pressure in the can to increases to dangerous levels, and the dough to become very short. Those factors limit the shelf life of the product.

The leavening acid for refrigerated biscuits must meet strict requirements. It must not react during mixing or sheeting of the dough but must react in the proofroom. This requirement is met only by the slow acting sodium acid pyrophosphate (SAPP). A number of different SAPPs are on the market. They vary in their reaction rates, i.e., in their release of CO_2. SAPPs are used widely in canned biscuits and in cake doughnuts, both of which have unique leavening requirements that are handled only by SAPPs. The major problem with SAPPs is the "aftertaste" that they leave in the mouth and on the teeth. Their so-called "pyro" taste is quite noticeable in these products. It apparently comes from the exchange of calcium from the teeth for the sodium in the disodium phosphate that results from the leavening reaction and is the result of the enzyme action that splits the pyrophosphate. Attempts to limit the effect of the disodium phosphate by adding various forms of calcium to the formula have been only partially successful.

16.6. Cakes

Cakes include a wide range of products of various types and sizes such as sponges, high ratio cakes, heavily fruited cakes, from cup cakes to wedding cakes and prepared cake mixes. Cakes, like cookies, are characterized by a high level of sugar in the formula. The major difference between the two is that cakes also contain relatively high levels of water. Because of the higher water level, starch can gelatinize and the cake sets during baking to a light, tender product that is characterized by a multitude of evenly distributed minute cells without any large holes [7, 210, 272].

16.6.1. Cake Making and Underlying Mechanisms of Cake Baking

Cake making starts from mixing a batter that is different from a cookie, cracker, or bread dough. Most important for all cake-making methods is the incorporation of small air bubbles into the batter since they act as nuclei for the expansion of the cake. The cake-making methods can be divided into two basic types, depending upon how the air is incorporated into the batter: multistage mixing and single-stage mixing.

Multistage Mixing. This is the classic procedure of high quality cake making, say, Madeira cake. It starts with a creaming step. Whipping air into the fat in finely distributed bubbles makes a fat–sugar cream. The air bubbles are dispersed in the continuous oil phase and stabilized by small, relatively stable β' fat crystals (needles of about 1 µm length; the polymorphic forms of fat crystals are named α, β' and β in order of increasing stability) distributed around the bubbles. Creaming steps are also used for certain types of cookies. Creaming offers a number of advantages; for example, a large number of air cells can be formed, leading to a fine cake texture. Also, the batter is stable and can sit for extended periods of time because the air is incorporated into the fat phase, where it is immobile and stable. Two and even three subsequent mixing steps then incorporate the liquids (egg, milk) and flour to form the final batter. This procedure allows the fat/air particles to be very finely distributed throughout the batter. The finer the distribution of fat and air, the better are the final cake volume and

crumb structure and texture. During baking, fat melts and separates from the air bubbles. The air bubbles move into the aqueous phase to generate a foam that is stabilized by egg protein. As heating proceeds, the air bubbles act as nuclei into which the carbon dioxide and water vapor diffuse to increase the volume of the batter. Then starch gelatinizes in the high sugar system, egg proteins coagulate, and the structure becomes set at about 95 °C.

Single-Stage Mixing. This type of cake making is typically used for sponge cake production. The liquids (water/milk and eggs) are mixed with the flour and the air is then incorporated into the aqueous phase. This is normally accomplished by mechanical means, i.e., high-speed mixing. When the air is incorporated into the aqueous phase, the fat (shortening) crystals of the cake formulation become coated with an interfacial layer of absorbed protein in the aqueous batter phase and then absorb to the air–water interface [273]. This mechanism helps to stabilize the large number of small air bubbles, to reduce coalescence of bubbles and the formation of larger bubbles during heating. Variations in bubble size and the presence of large bubbles result in an unstable mix, producing cakes of coarse and uneven texture. Gas can diffuse in a batter: gas diffuses out of the small bubbles into the aqueous phase and then into the larger bubbles because the pressure in a small bubble is larger than the pressure in a large bubble. Thus, there is a tendency for the small bubbles to disappear and the large bubbles to become larger. Because of the greater buoyancy the larger bubbles will overcome the high viscosity in the batter, rise to the surface and be lost. Hence, cake batters prepared by direct incorporation of air into the aqueous phase are not very stable and should not be allowed to sit for long periods of time before being baked.

A cake batter made by single-stage mixing can be regarded as an emulsion of the air-in-water type with air bubbles entrapped into the aqueous phase. During baking, the fat crystals melt to cover the internal surface of the bubble thus permitting bubble expansion without rupture. The volume of a cake batter is increased by about 3.5 times during baking as a result of the expansion of gases (air and carbon dioxide if baking powder is used) together with the vapor pressure (conversion of water into steam) within the bubble. Starch gelatinization and coagulation of egg proteins finally stop the process of bubble expansion and coalescence. Starch gelatinization is responsible for the dramatic increase in batter viscosity and is the first step in cake setting, i.e., the transformation from a liquid/foam to a solid sponge system. Because the molar sugar concentration is much lower in a cake than in a cookie formula, the starch can gelatinize during cake baking and prevents structural collapse of the finished product. The structure, volume and appearance of sponge cake depend on the levels of baking powder and emulsifier (monoacyl glycerol) and specific batter gravities as well as on ingredient-process interactions [274].

Box Cakes. All dry ingredients are sold to the customer in a box. Generally, the homemaker adds water, eggs, and oil to the mix and the batter is produced. The surfactants in the mix lower the interfacial tension and allow the air to be incorporated directly. Propylene glycol monostearate (PGMS) is a common surfactant used for this purpose. This type of cake mix must be run through a cake finisher that is essentially a grinder. Apparently, the purpose of the finishing is to bind the fat to the flour and to decrease the amount of free fat that might destabilize the aqueous foam produced during mixing. The box cakes produced are very fine and tender. They are not suitable for commercial production because they are very fragile and would not withstand handling during transport.

16.6.2. Role of Cake Ingredients

The ingredients in batter-type cakes perform specific functions with respect to batter properties and structure, texture and color of the cake.

Bakery Fats/Shortenings, and Surfactants. The fat is used in cake production for batter preparation and structure formation and it is essential for eating properties, flavor delivery, and shelf life of the cake [165]. During batter preparation, the fat causes the incorporation of air and controls, together with the emulsifiers and sugar, batter viscosity. While the solid

(plastic) shortening increases the viscosity during batter preparation, it acts as lubricant decreasing viscosity once it creates a liquid phase during heating in the oven. The viscosity of a cake batter is very important during the baking step. Too low a viscosity will allow large bubbles to rise to the surface and escape. It would also allow starch granules to accumulate at the bottom on the cake pan and produce a rubbery layer during baking.

The introduction of an emulsifying agent (mono- and diacyl glycerols) into a shortening enhances the emulsifying properties of the shortening and hence, complements its use. The batter increases in viscosity, has greater stability, and the air becomes more finely distributed, leading to a larger, moister cake. The batter containing the surfactants maintains a higher viscosity during the early stages of heating. In the later stages of heating, shortenings together with surfactants appear to lower the rate of increase in viscosity by reducing starch swelling. The use of these high-emulsifying fats enables to step up considerably the liquid and sugar level to produce lighter, sweeter, and moister cakes with improved shelf-life.

Sugars. Besides sugar's obvious function as sweetening agent, the amount of sugar in the cake formula affects batter viscosity, starch gelatinization temperature, as well as tenderness and color of the cake. The amount of sugar in the formula controls the onset of starch gelatinization and thereby delays starch swelling. As the level of sugar in the formula increases, the temperature at which starch gelatinizes increases. This mechanism allows more time for the product to rise, resulting in a cake with fine, uniform, porous structure and tender crumb. When the proportion of sugar in the formula is higher than the proportion of flour, the cake produced is called a *high ratio cake*. Due to the high level of sugar the starch gelatinization temperature increases to a point where the cake does not set and thus will collapse during baking. To avoid this phenomenon the flour must be treated with chlorine gas (see Section 5.2). The use of chlorine treated flour allows the use of 120–140 % sugar, based on the flour.

Reducing sugars act as color enhancers and color precursors by taking part in the Maillard browning reaction. Reducing sugars are often added to the formula in the form of milk (lactose) or fresh egg whites (glucose). Sucrose is nonreducing and, because the formula contains no yeast (and therefore no invertase) to hydrolyze sucrose to reducing sugars, it does not give browning. Also because the pH of the system is basic, no chemical hydrolysis of sucrose to glucose and fructose occurs. Therefore, when sucrose is the only sugar in the formula, no browning takes place and the cake is quite white. Invert syrups (a mixture of glucose and fructose obtained by hydrolysis of sucrose) are hygroscopic and used as humectants that increase the retention of moisture in the crumb, resulting in improved eating properties and an extended shelf-life. In addition to the use of sucrose in cake itself, it is an essential ingredient in various icings, toppings and fillings.

Egg Whites. The proteins in egg white increase viscosity at room temperature by facilitating the incorporating of air into the aqueous phase of the batter. During heating, egg white proteins denature, leading to a great increase in batter viscosity and help to set the cake structure. In excess, egg whites tend to produce a tough, rubbery cake texture.

Egg white is the most important component in the angel food cake formula. Generally, the eggs and sugar are whipped to a protein foam, and the flour is folded in carefully so as not to disrupt the foam. The function of the flour appears to be to provide starch that will gelatinize and thereby remove excess free water. A low level of flour, often diluted with wheat starch characterizes the formula. Cream of tartar (monopotassium tartrate) is added to reduce the pH and, thereby, to improve the whipping of the egg whites. The formula contains no leavening agent. Simply the air trapped in the egg white foam produces the leavening. Care must be taken not to introduce any fat into the formula, as fat destabilizes the foam. In fact, one cannot use plastic containers because they retain enough fat to destroy the foam.

Milk, because of the reducing sugar lactose, aids in browning or crust color. It is also generally considered to improve the cakes flavor.

Flour. Cake flours are of low extraction rate, low protein content, and milled to a fine parti-

cle size. Generally, the finer the particle size, the better the cake flour. Therefore, most cake flours are pin-milled to reduce the particle size. Pin-milling also increases the amount of damaged starch in the sample, but this does not appear to affect the cake flour quality. The improvement of flour functionality for cake making by chlorination has been discussed (see Section 5.2).

16.7. Pastries

Pastries can be divided into two product groups: short pastry and puff pastry [210].

16.7.1. Short Pastry

The term short pastry encompasses (1) sweet pastry such as jam tarts, fruit pies, choux paste, (2) unsweetened, savory pastry, and (3) richer pastries such as shortbread and confectioner biscuits. All these products contain significant proportions of fat in the recipe, usually between 40–50 g/100 g flour. The major function of fat is to shorten the texture and to make the product soft and tender so it melts in the mouth. Like in cake production, shortness is achieved by using fats/shortenings that avoid or reduce gluten development. Thus, the elastic gluten strands are shortened and a short paste or pastry dough results. As a tenderizer butter or margarine can be used themselves or in combination with lard, or a general-purpose shortening as the texture of the fat allows its easy distribution throughout the dough.

The high level of sugar in sweetened pastes reduces the availability of water for hydration of the protein and also reduces the tendency for gluten development during paste mixing. Baking powder might be added to short pastes to give a more open texture and a lighter eating product, especially when lower fat levels are used in the pastes since the tendency for gluten development is increased.

16.7.2. Puff Pastry

Puff pastry is a laminated bakery product with a light, flaky texture. It is believed to have originated in France but the art of preparing layered dough was perfected in Vienna, Austria. The puffing, i.e., the volume increase during baking, involves neither chemical or biological leaving but depends solely on the transformation of water to steam that becomes trapped between dough layer separated by thin films of fat resulting from the lamination process [210, 275].

The ability of the roll-in fat to keep dough layers separated primarily determines the quality of puff pastry. Laminating of fat into the dough is achieved by a series of reduction (sheeting) and folding steps that forms alternate and discrete layers of dough and fat. A dough sheet is covered with fat, folded, reduced in thickness by rolling and subjected to further folding and rolling to build up the required number of layers of thin dough sheets with a well developed gluten network separated by layers of fat. As the number of dough and fat layers increases and a multilayer sandwich structure develops, the specific height (mm/g paste) of the baked product first rises to a maximum and then declines. The optimum number of fat layers depends on the type of puff pastry product and the process conditions and should provide a product with a regular structure and flaky texture. Lower levels of layers might give greater specific heights of the final product but a lower quality pastry due to individual layers yielding a fragile, irregular structure with large voids. More laminations eventually result in thinner dough and fat layers causing coalescence of the dough phase and formation of capillaries and micropores through which the steam generated during baking can escape. A distribution of fat through the dough phase with a subsequent disruption of the required gluten network might occur. Both mechanisms, interconnections of the dough phase and interruption of the gluten network would decrease lift and final product volume.

Since the melting of the fat layer destroys the layer, it is imperative that the fat remains solid. Thus, it is a common practice to retard (allow to rest at low temperature) the dough between sheeting operations.

There are essentially three basic production methods for puff pastry (English, French, and Scottish method) that differ in the use of laminating fat. In all cases, the laminated dough is finally cut and formed into the required final shape. If appropriate, the pastry specific height can be controlled by docking of the paste before

baking. Pastry dough can be used for a wide variety of pastries such as baked pastry layers for torte cakes or quiche or as filled products such as Danish pastries, fruit pockets, creme horns, and strudel.

During the initial step of baking the laminating fat melts at 35–40 °C and forms a liquid fat layer. Then, water of the puff pastry dough will evaporate as steam that passes into the molten, liquid fat layer where it inflates large bubbles. The steam moves into the liquid fat because of the smaller pressure differential needed to inflate a bubble of steam in the liquid fat than in the dough phase. The expanding bubbles are trapped between the hardly penetrable multi-layer structure that act as barrier to the water vapor. The pressure generated by the bubbles causes a puffing effect that expands and stretches the elastic gluten network and subsequently leads to the lift. The volume increase of the product during baking may reach up to eight times the original dough volume. The final product structure is fixed by interconnections of dough strands via gelatinization and heat-setting of gluten proteins. All these events are responsible for the well-known open, flaky, and crispy texture of puff pastry.

Only about half of the water vapor generated exerts pressure to lift the pastry. The remaining water escapes as steam after the structure has set and cannot lift the pastry any further. To assure a good lift it, is essential that the integrity of the dough layers is as complete as possible. That means, for lamination until the puffing step the formation of uninterrupted fat separation layers between the dough sheets is required. If dough strands are interconnected the water vapor escapes to the outside through micropores and capillaries. The molten laminating fat partially soaks into the micropores and capillaries of the adjacent dough matrix. This fat acts as filling medium providing lubrication and tender, eating properties.

Ingredients for Puff Pastry Making.

Fat. The fat must give a tough and plastic texture at dough temperature (20 °C). This texture can be reduced to a thin continuous film when subjected to rolling and stretching [276]. This "roll-in-fat" must resist the mechanical stress of the sheeting process and should not be absorbed and penetrate into the dough since this would interfere with the development of the necessary viscoelastic gluten network. The continuous fat film must be a stable barrier separating the adjacent layers of dough, also in the initial stages of baking, to maintain a multi-laminate structure. To avoid an unpleasant waxy or greasy mouthfeel when the product is eaten, there is a limit of using fat with a melting point of ≤ 40 °C. Butter is an excellent roll-in-fat. It contributes to a pleasant taste, aroma, and mouthfeel when processed within a temperature range of 15–20 °C. Special margarines have been developed as roll-in fats. These puff pastry fats used for lamination are made from blends of animal and vegetable fats with high melting points (usually about 40 °C) and high solid to liquid ratios that show little change over a broader temperature range.

Fat also plays a role in puff pastry flavor formation and the type of fat (e.g., butter or shortening) generates different flavor notes. Differences in the concentration of two potent odorants formed during the autoxidation of linoleic acid are most likely the cause for the differences in the overall odors of butter and margarine pastry.

Flour. During lamination, the flour must provide the gluten structure to form a strong dough film that is able to entrap the water vapors as they are formed during baking. This facilitates the formation of large vacuoles to form, leading to a very open texture and large product volume. Therefore, the amount and quality of gluten proteins is essential for the integrity of the paste layers and the lift of the product during baking. Strong flours with protein contents of about 11.5–13.5 %, low extensibility, and high resistance to stretching will yield a better product lift and higher volume puff pastry with a very flaky texture than weaker flours with protein contents < 11 %, high extensibility and low resistance to stretching. Higher protein levels tolerate slightly more lamination, but may not produce a better pastry. The water absorption of the dough is usually 56–62 % of the flour weight. A well-formulated pastry dough has consistency comparable to that of a stiff bread dough and similar to that of roll-in fat during sheeting. Liquid or dried whole eggs not only add liquids to the dough but contribute also to dough strength and to an attractive golden-brown crust color via proteins participating in Maillard browning reactions. Soft doughs are difficult to process and

tend to slide over the roll-in fat during sheeting and the fat will break eventually into lumps that will not form the necessary continuous fat film. The flour specification required also depends on the plant design. After the lamination process the dough is rested for $\geq$ 15 min to allow relaxation of the developed gluten and to avoid shrinkage of the dough during baking. Thus, for short relaxation times weaker flours must be used, for long relaxation times stronger flours.

17. Retention of Baked Product Quality

17.1. Staling

Bread loses its desirability progressively with time it is out of the oven. Those undesirable changes that occur with time are collectively called staling: the crust takes up moisture from the air or the crumb and loses crispness and becomes tough; the crumb becomes inelastic, dry, opaque, firm, and crumbly and loses its coherence; flavor and taste become increasingly stale and off-flavor develops.

The changes that take place in the crust are clearly different from those that occur in the crumb. The toughening of the crust appears to be mainly associated with migration of water from the crumb to the crust [277]. When bread is freshly baked, the crust is dry and only contains 2–5 % moisture. Under these conditions the crust is crisp, friable and desirable. With the time, the moisture migrates to the crust, as an equilibrium is established. At the same time moisture is lost to the atmosphere, so the crust's property is largely preserved. However, if bread is placed in a plastic bag the loss of moisture from the crust is slowed. As a result of this moisture diffusion from crumb to crust, the moisture content of the crust increases. Consequently, the crust loses its friability and becomes tough, the biopolymers (carbohydrates, proteins) forming the crust go through their glass transition temperature and thus, change from a glassy, crispy state to one that is leathery or rubbery (see Chap. 14).

The molecular changes that occur in the bread during storage and that are responsible for the firming and staling phenomenon appear to be quite complex. It was shown about 150 years ago, that the firming of bread crumb is not a drying phenomenon [278]. Firming happens even though no moisture is lost.

In fresh bread crumb, a thin layer of gluten proteins separates swollen, gelatinized starch granules (Fig. 13). Staling of bread involves, among other things, the recrystallization of starch, specifically of the amylopectin fraction (retrogradation, see Section 5.1.1) [50]. Essentially none of the amylopectin becomes soluble during baking. Inside the remnants of the starch granules, amylopectin chains begin to reassociate into ordered structures by forming "junction zones" that consist of double helices. These junction zones serve as nucleation sites for the aggregation of double helices into crystallites (retrogradation). Thereby, a network is formed comprising starch chains passing through amorphous and crystalline regions inside the remnants of the granules. This network appears to contribute to the development of crumb rigidity. The increase in opaqueness of the crumb is presumably caused by the growth of the crystallites, which changes the refractive index.

The concepts of glass transition (see Chap. 14) and the nonequilibrium theory of crystallization (nucleation and propagation: crystal formation, and crystal growth) have been applied to explain the strong influence of the storage time and temperature on starch retrogradation [53, 57]. According to these concepts, the rate of amylopectin recrystallization depends upon the total water content through depression of the T_g and the storage temperature. A bread loaf in the refrigerator has a storage temperature closer to its T_g and the rate of crystal nucleation will be higher. Consequently, if the bread is removed and kept at room temperature, it will stale, i.e., firm faster. The kinetics of the structural changes of the starch in the crumb can be monitored, for example, by changes in the mobility and ordering of the starch chains with NMR, by the development of a β-type pattern ($\rightarrow$ Starch) in the X-ray diffraction pattern, or by the evolution of a melting endotherm at about 55 °C (staling peak) in the DSC (differential scanning calorimetry) profile.

Starch retrogradation has been viewed as the basic cause of bread staling. However, the rate of firming and the rate of starch retrogradation are not the same. In addition, heating can re-

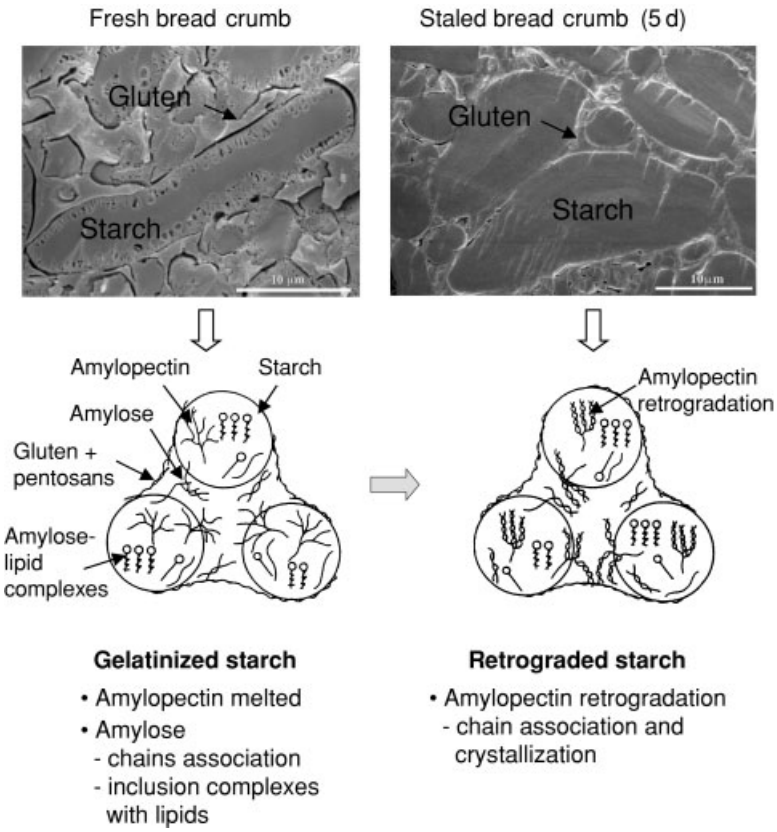

Figure 13. Changes of starch and moisture in bread crumb during storage.

verse the firming of bread crumb. This is one of the advantages of toasting bread; it is said to be refreshened. The retrograded amylopectin crystals melt at about 50 °C, but bread crumb continues to lose firmness as it is heated above 60 °C to about 100 °C. This argues that firming during storage and refreshening during heating is not completely explained by retrogradation of amylopectin.

Apparently, other factors such as water redistribution and starch–gluten interactions are also involved in the staling phenomenon [279, 280]. Cryo-scanning microscopy (Fig. 13), for example, can be used to draw conclusions on the state of water in the bread crumb. Freezable water can be detected by ice crystal formation induced by a slow freezing technique: large crystals indicate high freezable water content, small crystals a low water content. In fresh bread ice crystal formation can be observed in the periphery of the starch granule indicating higher water content at the granule surface than at the granule center. It could therefore be hypothesized that the granule surface is softer, more flexible and easier mechanically deformable than the center enabling a sliding of the starch granule surface over the gluten film. This would result in a softer crumb texture. In freezing stale bread ice crystals increasingly disappear from the inside to the outside of the granule suggesting that water, in part at least, becomes bound within the starch granule as a hydrate, or as nonfreezable water in amylopectin crystals that form upon aging. As a result, granules become more firm. In addition, the moisture gradient between the starch granule surface and the surrounding gluten phase would induce water redistribution from gluten to the starch. This loss of plasticizer (nonfreezable water) could cause adherence of the starch and gluten phase making sliding impossible. This again would contribute to a firming in the crumb texture.

A model for firming has been proposed that includes an effective crosslinking of the contin-

uous gluten phase with the gelatinized starch in bread [280, 281]. This model is based, in part, on the fact that gluten is the continuous phase in bread and that the firmness of bread is a long range phenomenon. Thus, by itself, it would not matter if the discontinuous starch phase is firm or flexible. It is too small to be detected. However, once an interaction between the gluten and the starch has been formed, firming of the remnant of the starch granules would help to firm the total network.

The degree and kinetics of staling of baked goods are affected by several factors such as origin and type of cereal raw material (wheat, rye, extraction rate, wholemeal flour), ingredients (water, fats, sugars, emulsifiers, ethanol), process conditions (baking time, temperature, humidity), product type (bread, buns, rolls, cake, etc), product size, and storage conditions (time, temperature, humidity) [282]. A number of factors alter the rate of staling and produce softer bread that retains its softness over time. A baked product with a high volume and fine crumb structure is softer and retains softness better than a product with lower volume and a coarse crumb grain. Rye breads, in general, stale less fast than wheat breads and baked products made from high extraction flours or wholemeal flours remain longer fresh than products from low extraction flours. Flours produced from partially waxy wheats have shown slower firming rates [22]. The functional properties of waxy wheats and their variation because of environmental and genotypic affect have been reported [283]. The use of fat, sugars, surfactants, and enzymes in the recipe or ethanol sprayed on the surface of baked goods (e.g., cake-type products) can retard staling and retains bread soft. The use of surfactants (emulsifiers) as antistaling additives [172] has been related, among others, to two factors: (1) surfactants form inclusion complexes with amylose and thus, prevent amylose chain association and crystallization [40, 62] and (2) in the presence of surfactants, starch does not swell as much as starch alone. Inclusion of shortening in the bread formula alters the staling rate. In addition, the use of a relatively heat-stable α-amylase in the bread formula retards staling [284] (see Chap. 8). The antifirming effect of these enzymes might be explained by improving loaf volume and crumb structure [175], "softening" the starch network [50], reducing the formation of amylopectin crystallites during storage [185] or by affecting starch–gluten interactions [281]. The conditions at which bread is stored are also important. Higher temperatures retard staling and cooler temperatures (above freezing) increase the rate of firming.

17.2. Microbial Spoilage

Flour contains many microorganisms and most of these are inactivated during baking. Bread and moist fine bakery wares, because of their high water activity and abundant available nutrients are good growth media for microorganisms. If these products are contaminated after baking they can spoil rapidly. One organism that appears to survive baking is the one causing rope. Rope occurs primarily in yeast-leavened products. Rope in bread starts to be noticed as early as 12 h after baking. The bread crumb develops a slightly bitter taste and a fruity odor. With time, the crumb becomes more discolored and the taste becomes repulsive as the bread crumb decomposes. The dark crumb loses its elasticity and can be pulled into fine, silky threads (i.e., rope). At this point the bread is inedible and can cause diarrhea and vomiting.

A baked product is free of fungi as it comes from the oven. However, it can be contaminated during cooling, slicing, or packaging. Mold can develop in about 3 d under ideal conditions. Breads with higher moisture content and thin crusts are more susceptible than those with low moisture and those with a hard, thick crust. Moldy bread has objectionable taste and smell and must be rejected due to safety reasons since the mold-inducing microorganisms may also produce mycotoxins.

17.2.1. Retention of Quality

The trend to larger bakeries and, therefore, transport of baked goods over longer distances has made retention of quality more important. Storage conditions aim at reducing moisture loss, contamination by microorganisms, and enzyme reactions. Recommended storage conditions are 15–20 °C and 65–75 % relative humidity for bread. Both heating and freezing can be used to preserve baked goods. Freezing allows baked

goods to be stored for long periods of time without staling or microbial damage. To reduce staling the product must be frozen and thawed rapidly as storage at just above freezing is subject to rapid firming (see Chap. 14). Freezing can also cause dehydration (freeze drying) of products. Thus, good packaging is important. In general, freezing is expensive. Preservation by heating (45–65 °C, 75–80 % relative humidity) is also quite effective at retarding staling and microbial growth but does have undesirable side effects. Bread with light colored crust might darken and the flavor is affected. Par-baking is also used to preserve baked products such as baguette-type goods, buns, and rolls. Par-baked products are partially baked, packed under modified atmosphere or treated with preservatives and are shelf-stable for some weeks. Fully baked in the consumer's oven they provide fresh baked products. Baked goods are often not packaged if they are expected to be sold and consumed rapidly. For hygiene reasons the goods may be wrapped in paper. Waxed paper does not retain moisture and the product can easily be contaminated with molds. Perforated paper/plastic bags allow moisture transfer from the product to the surrounding air outside of the bag. This mechanism facilitates the moisture evaporation from the product to the air outside and thus, avoids the retransfer of moisture to the crust and helps to keep the crust crispy for a longer period of time. It does not protect against mold contamination. For bread that is sliced and shipped over some distance, polyethylene bags are often used. They can withstand heating to 120 °C retain moisture well, extensible, and flexible. For long term storage (e.g., military use) bread can be baked in cans, which can then be sealed and sterilized.

17.2.2. Chemical Preservatives

The use of chemical preservatives is controlled in various countries by differing rules and regulations. In most countries, various propionates are used to inhibit mold and rope organisms. The most commonly used additive to inhibit mold growth in bread is calcium propionate (0.15–0.30 g/kg). Calcium propionate also retards yeast activity to some extent. Sorbic acid is very effective in stopping mold. However, it also is effective in inhibiting yeast activity. Thus, it cannot be used as a formula ingredient in bread. It is often used in chemically leavened systems or as a spray on the surface of baked products. Ethanol has also been used to preserve baked goods, especially cake-type products.

18. Trends in Baking

18.1. U.S. Bakery Market

Wheat flour per capita consumption in the United States has held steady in 2004–2005 at 60.9 kg. This is at an all-time low. After reaching a modern high of 66.4 kg per person in 2000, the consumption has declined (64.1 kg in 2001 and 62.3 kg in both 2002 and 2003). It appears that the damage done to the milling and baking industries by the "bad carbohydrate" diets persists even though those diets are no longer popular. As one might expect, the above trend has severe effects on the economy of the industry. The large bakers are still closing bakeries and producing more product with fewer plants. This means that the distance between the bakery and the consumer is greater and, therefore, also the time between baking and buying the bread by the consumer is longer.

A trend to high fiber breads continues but the rate of increase has slowed. The high fiber breads include added fiber from such sources as soy hulls, oat hulls, pea fiber, and many others. Whole meal breads also fall in this category. The increased sales of these types appear to be driven by two factors. The first is the perception that whole meal breads are good for the health. The healthy image appears to cover all dark breads. The second trend is based on the fact that the high fiber breads have reduced calorific value. The reduction in calorific value is because the fiber is not completely digestible and because the breads contain high levels of water. Multigrain breads also have a healthy image and sales are growing rapidly compared to other breads.

Bread shops, both free standing in shopping malls and in-store bakeries in supermarkets, continue to grow. Their rapid growth does not appear to be at the expense of the wholesale bakeries. There are, apparently, increased sales of bread which may be, at least partially, responsible for the increase in wheat flour consumption in the USA. Most bread shops sell primarily crusty

breads and most, if not all, of their product is sold within a few hours after being baked. Thus, the costumer has access to fresh bread. Most of the free-standing bread shops also sell sandwiches, salads, and soups in addition to the baked products. Much of the bread produced by in-store bakeries is produced from frozen dough. However, the growth of production of frozen dough appears to have slowed. Frozen dough may be giving up market share to par-baked breads.

A rapid growing product in the United States is the wheat flour tortilla. Sales of tortillas in 2004 were \$ 6.1×10^9 up from \$ 1.1×10^9 in 1997. The trend shows no sign of slowing. The tortilla is the 2nd most popular baked item in the United States after white pan bread. Cookie and cracker markets show little overall growth. The industry appears to be moving away from products that are marketed as "good for you" and concentrating on products with full flavor and things that are fun to eat.

18.2. Western European Bakery Market

European bakeries produce about 25×10^6 t of bread, about 8×10^6 t of which in an industrial environment, and 5×10^6 t of biscuits (cookies in the U.S.), cakes and pastries [285]. Germany is by far the largest producer of bread with a European market share of about 25 %, followed by Italy, France, the UK, and Spain. An average of about 65 kg of bread per capita per year are consumed in Europe with Germany, Denmark, Finland, Italy, and Belgium being the highest in consumption. The Germans and Austrians consume about 80 kg per year while in the United Kingdom and Ireland the annual per capita consumption is below 50 kg.

Bread plays a different role in the diet across Europe. In countries such as Germany and France bread is a major meal component, while in the UK it is seen as an accompaniment to other foods. In addition, different types of bread are favored in different countries. Germany, Norway, Belgium, and the Netherlands tend to prefer brown, wholemeal and mixed wheat/rye bread whereas white bread, mainly baguette, dominates the market in France.

While the bread market is relatively static as a whole there are some sectors where sales are increasing. This includes speciality and ethnic breads such as certain types of flat breads like tortillas, ciabatta, pita and naan bread. A fast growing sector within the bakery market in Europe is the morning goods market (e.g., croissants, rolls, buns, muffins, pancakes, bagels).

The UK, Germany, Italy, and France together contribute to more than 70 % of the total production of the 5×10^6 t of biscuits, cakes, and pastries produced in Europe. Within the biscuit, cake, and pastry sector, biscuits account for about 75 % of the total European production. The UK, Germany, France, and Italy are the greatest consumers of biscuits, cakes, and pastries with about 22 %, 22 %, 16 % and 15 %, respectively, of the total European consumption. In general, the market is stable (average European consumption about 10 kg/a) and is not expected to change dramatically in the future. In contrast to the purchase of bread, the purchase of biscuits, cakes and pastries is more impulse driven. In response, producers across Europe are improving the product appearance (e.g., packaging) and are searching for new merchandising and distribution channels to improve sales.

The production sites of bread and other baked products can be divided into three groups although definitions vary in different countries: (1) craft bakeries that use small-scale machinery and sell their products directly to the consumer, (2) plant bakeries that use large-scale equipment and produce a range of products that is distributed to wholesalers, supermarkets, restaurants etc., and (3) in-store bakeries that are either owned by supermarkets or franchised to a local baker.

Traditional craft bakeries remain dominant in Italy and France while industrial or plant bakeries are more dominant in northern Europe. Overall, plant bakeries are accounting for an increasing proportion (almost a third) of total bread production in Europe. Consolidation and expansion of market positions through mergers and acquisitions has been a noticeable trend. Generic standards for quality and safety management and quality assurance such as ISO registrations and the Hazard Analysis Critical Control Point (HACCP) concept have now been introduced into the bakeries, minimizing food safety issues. Experimental designs, mathematical modeling and knowledge-based expert systems offer new tools in production scheduling,

process control, and baked product development [286, 287].

18.3. Consumer Demands and Product Trends

Convenience, taste and freshness are expected to remain the driving purchases of bakery products [288 – 293]. Dough, batters, and bakery products are widely distributed in frozen, par-baked, and refrigerated form so that they are available at more outlets and can be baked or re-heated on the premises with "just baked" smell and taste. Convenience is essential and baked products whether packaged or just baked are now available at many nontraditional outlets like gas stations. The appeal of freshness is evidenced in the sales of household bread-making machines and the rising popularity of in-store bakeries.

Flatbreads are well suited to trendy sandwiches and flatbread sales are raising. The distinction between snacks and crackers is blurring and some products such as filled bagels make their way into the snack category since they provide a convenient and quick form of energy.

In addition, the concern and interest in a healthier diet will continue to increase and there is a greater demand from consumers to use more nutritional functional ingredients in products [294 – 297]. After the initial phase of curiosity and health-induced purchases, consumers continue to turn away from products like fat-free or low-fat sweet baked good varieties because they are first unhappy with the taste and second nonconvinced about the advertised, claimed, or initially personally perceived health effects of these products. To focus on nutrition, however, will become increasingly important to bakery food producers. In the EU, large research efforts have been launched to provide healthy bread products (see www.solfibred.com and www.healthgrain.org).

Within the health domain, bakery products play an increasingly important role as functional foods, i.e., foods that provide additional physiological effects beyond their traditional effects such as preventing and correcting classical nutritional deficiencies. Bakery products are ideal carriers for bioactive nutrients, since they are produced from cereal grains that contain intrinsically beneficial components and thus, these products can be considered as nutritional functional in nature. The grain components or functional nutrients with favorable properties that can be incorporated in baked products include, among others, dietary fibers/prebiotics like β-glucans, hemicelluloses and fructo-oligosaccharides, resistant starch, phytochemicals, probiotics, antioxidants, vitamins, and minerals [298 – 301].

The increase in health, social and environmental concern of the consumer represents an opportunity for bakers to diversify into good tasting products such as organic multi-grain, fiber/bran enriched breads.

19. References

1. R. Drapron, J. Potus, F. Laplume, P. Potus: *Notre pain quotidien*, Antenne Générale de Publication (AGP), Paris 1999.
2. H. Eiselen (ed.): *Brotkultur*, Vater und Sohn Eiselen-Stiftung, Ulm, DuMont, Cologne 1995.
3. A. Spicer: *Bread*, Applied Science Publ., London 1975.
4. W. Ziehr, E. Buhrer, M. Wahren: *Bread through the ages*, Lannoo, Tielt 1987.
5. C. F. Morris, A. N. Massa: "Puroindoline Genotype of the U. S, National Institute of Standards & Technology Reference Material 8441, Wheat Hardness," *Cereal Chem.* **80** (2003) 674–678.
6. Y. Pomeranz, P. C. Williams: "Wheat hardness: its genetic, structural, and biochemical background, measurement and significance," in *Advances in Cereal Science and Technology*, vol. **X**, American Association of Cereal Chemists, St. Paul, MN 1990, pp. 471–548.
7. R. C. Hoseney: *Principles of Cereal Science and Technology*, 2nd ed., American Association of Cereal Chemists, St. Paul, MN 1994.
8. Y. Pomeranz: *Wheat chemistry and technology*, 3rd ed., vol**s. I + II**, American Association of Cereal Chemists (AACC), St. Paul, MN 1988.
9. Y. Pomeranz (ed.): *Wheat is unique*, American Association of Cereal Chemists (AACC), St. Paul, MN 1989.
10. C. F. Morris, S. P. Rose: "Wheat," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal grain quality*, Chapman & Hall, London 1996, pp. 3–54.

11. L.-C. Autran, R. J. Hamer, J. J. Plijter, N. E. Pogna: "Exploring and improving the industrial use of wheats," *Cereal Foods World* **42** (1997) 216–227.
12. N. E. Pogna, R. Redaelli, T. Dachkevitch, A. Curioni, A. Dal Belin Peruffo: "Genetics of wheat quality and its improvement by conventional and biotechnological breeding," in W. Bushuk, V. F. Rasper (eds.): *Wheat production, properties and quality*, Blackie Academic & Professional, Chapman & Hall, Glasgow 1994, pp. 205–224.
13. C. W. Wrigley, C. F. Morris: "Breeding cereals for quality improvement," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal grain quality*, Chapman & Hall, London 1996, pp. 321–369.
14. R. A. Graybosch: "Waxy wheats origin, properties, and prospects," *Food Sci. Technol.* **9** (1998) 135–142.
15. O. Anderson: "Molecular approaches to cereal quality improvement," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal grain quality*, Chapman & Hall, London 1996, pp. 371–404.
16. M. N. Cline, M. A. Esfield: "New horizons for the amber waves — technologies boost the capabilities of wheat," *Cereal Foods World* **43** (1998) 5–10.
17. F. Katz: "Biotechnology-New Tools in Food Technology's Toolbox," *Food Technol.* **50** (1996) 63–65.
18. Y.-Y. Linko, P. Javanainen, S. Linko: "Biotechnology of bread baking," *Trends Food Sci. Technol.* **8** (1997) 339–344.
19. K. Liu: "Biotech crops: products, properties, and prospects," *Food Technol.* **53** (1999) 42–49.
20. J. Q. Wilkinson: "Biotech plants: from lab bench to supermarket shelf," *Food Technol.* **51** (1997) 37–42.
21. L. Rooke et al.: "Overexpression of a gluten protein in transgenic wheat results in greatly increased dough strength," *J. Cereal Sci.* **30** (1999) 115–120.
22. B.-K. Baik, C. S. Park, B. Paszcynska, C. F. Konzak: "Characteristics of noodles and bread prepared from double-null partical waxy wheat," *Cereal Chem.* **80** (2003) 627–633.
23. M. Båga et al., "Wheat starch modification through biotechnology," *Starch/Stärke* **51** (1999) 111–116.
24. T. J. Hoban: "Consumer acceptance of biotechnology in the United States and Japan," *Food Technol.* **53** (1999) 50–53.
25. E. S. Posner, A. N. Hibbs: *Wheat flour milling*, American Association of Cereal Chemists, St. Paul, MN 1997.
26. W. Bushuk, V. F. Rasper: *Wheat-Production, properties and quality*, Blackie Academic & Professional, Glasgow, imprint of Chapman & Hall, London 1994.
27. H. J. Cornell, A. W. Hoveling: *Wheat-Chemistry and Utilization*, Technomic Publishing Company, Inc., Basel 1998.
28. A.-C. Eliasson, K. Larsson: *Cereals in Breadmaking*, Marcel Dekker, New York 1993.
29. R. J. Hamer, R. C. Hoseney (eds.): *Interactions: The key to cereal quality*, American Association of Cereal Chemists, St. Paul, MN 1998.
30. R. Lasztity: *Cereal Chemistry*, Akadémiai Kiado, Budapest 1999.
31. T. Galliard (ed.): *Starch: properties and potential. Critical Reports on Applied Chemistry*, vol. **13**, Society of Chemical Industry, John Wiley & Sons, New York 1987.
32. M. J. Gidley, D. Cooke: "Aspects of molecular organization and ultrastructure in starch granules," *Biochemical Society Transactions* **19** (1991) 551–555.
33. A. Imberty, A. Buléon, V. Tran, S. Pérez: "Recent advances in knowledge of starch," *Starch/Stärke* **43** (1991) 375–384.
34. P. J. Jenkins, R. E. Cameron, A. M. Donald: "A universal feature in the structure of starch granules from different botanical sources," *Starch/Stärke* **45** (1993) 417–420.
35. H. F. Zobel: "Molecules to granules: a comprehensive starch review," *Starch/Stärke* **40** (1988) 44–50.
36. C. Biliaderis: "The structure and interactions of starch with food constituents," *Can. J. Physiol. Pharmacol.* **69** (1991) 60–78.
37. C. Biliaderis: "Non-equilibrium phase transitions of aqueous starch systems," in H. Levine, L. Slade (eds.): *Water Relationships in Food*, Plenum press, New York–London 1991, pp. 251–273.
38. C. Biliaderis: "Structures and phase transitions of starch in food systems," *Food Technol.* **46** (1992) 98–109, 145.
39. D. Cooke, M. J. Gidley: "Loss of crystalline and molecular order during starch gelatinsation: origin of the enthalpy transition," *Carbohydrate Res.* **227** (1992) 103–112.
40. A.-C. Eliasson: "Lipid-carbohydrate interactions," in R. J. Hamer, R. C. Hoseney (eds.): *Interactions: the key to cereal quality*,

Am. Assoc. of Cereal Chem., St. Paul, MN 1998, pp. 47–79.
41. R. C. Hoseney: "Chemical changes in carbohydrates produced by themal processing," *J. Chem. Edu.* **61** (1984) 308–312.
42. W. R. Morrison: "Starch lipids and how they relate to starch granule structure and functionality," *Cereal Foods World* **40** (1995) 437–446.
43. K. Münzing: "DSC studies of starch in cereal and cereal products," *Thermochimica Acta* **193** (1991) 441–448.
44. J. Jane, Y. Y. Chen, L. F. Lee, A. E. McPherson, K. S. Wong, M. Radosavljevic, T. Kasemsuwan: "Effects of amylopectin branch chain length and amylose content on the gelatinization and pasting properties of starch," *Cereal Chem.* **76** (1999) 629–637.
45. V. M. Leloup, P. Colonna, A. Buleon: "Influence of amylose-amylopectin ratio on gel properties," *J. Cereal Sci.* **13** (1991) 1–13.
46. P. L. Russel: "Gelatinisation of starches of different amylose/amylopectin content. A study by differential scanning calorimetry," *J. Cereal Sci.* **6** (1987) 133–145.
47. K. Ghiasi, R. C. Hoseney, E. Varriano-Marston: "Gelatinization of wheat starch. III. Comparison by differential scanning calorimetry and light microscopy," *Cereal Chem.* **59** (1982) 258–262.
48. A. Beleia, R. A. Miller, R. C. Hoseney: "Starch gelatinization in sugar solutions," *Starch/Stärke* **48** (1996) 259–262.
49. K. Kohyama, K. Nishinari: "Effect of soluble sugars on gelatinization and retrogradation of sweet potato starch," *J. Agric. Food Chem.* **39** (1991) 1406–1410.
50. R. E. Hebeda, H. F. Zobel: *Baked goods freshness*, Marcel Dekker, Inc., New York 1996.
51. R. Hoover: "Starch retrogradation," *Food Rev. Int.* **11** (1995) 331–346.
52. M. Langton, A.-M. Hermansson: "Microstructural changes in wheat starch dispersions during heating and cooling," *Food Microstructure* **8** (1989) 29–39.
53. V. J. Morris: "Starch gelation and retrogradation," *Trends Food Sci. Technol.* **1** (1990) 2–6.
54. Ph. Roulet, W. M. MacInnes, P. Würsch, R. M. Sancheu, A. Raemy: "A comparative study of the retrogradation kinetics of gelatinized wheat starch in gel and powder form using X-rays, differential scanning calorimetry and dynamic mechanical analysis," *Food Hydrocolloids* **2** (1988) 381–396.
55. P. L. Russel: "The ageing of gels from starches of different amylose/amylopectin content studied by differential scanning calorimetry," *J. Cereal Sci.* **6** (1987) 147–158.
56. D. Sievert, P. Würsch: "Amylose chain association studied by differential scanning calorimetry," *J. Food Sci.* **58** (1993) 1332–1334, 1345.
57. L. Slade, H. Levine: "Recent advances in starch retrogradation," in S. S. Stivala, V. Crescenzi, I. C. M. Dea (eds.): *Recent Developments in Industrial Polysaccharides*, Gordon and Breach Science, New York 1987, pp. 387–430.
58. K. J. Zeleznak, R. C. Hoseney: "The role of water in the retrogradation of wheat starch gels and bread crumb," *Cereal Chem.* **63** (1986) 407–411.
59. P. Würsch, D. Gumy: "Inhibition of amylopectin retrogradation by partial beta-amylolysis," *Carbohydrate Res.* **256** (1994) 129–137.
60. P. Cairns, M. J. Miles, V. J. Morris: "Studies of the effect of the sugars ribose, xylose and fructose on the retrogradation of wheat starch gels by X-ray diffraction," *Carbohydr. Polym.* **16** (1991) 355–365.
61. S.-M. Chang, L. C. Liu: "Retrogradation of rice starches studied by differential scanning calorimetry and influence of sugars, NaCl and lipids," *J. Food Sci.* **56** (1991) 564–566, 570.
62. Z. Czuchajowska, D. Sievert, Y. Pomeranz: "Enzyme-resistant starch. IV," Effects of complexing lipids, *Cereal Chem.* **68** (1991) 537–542.
63. Y. Hibi, S. Kitamura, T. Kuge: "Effect of lipids on the retrogradation of cooked rice," *Cereal Chem.* **67** (1990) 7–10.
64. K. J. I'Anson, M. J. Miles, V. J. Morris, L. S. Besford, D. A. Jarvis, R. A. Marsh: "The effects of added sugars on the retrogradation of wheat starch gels," *J. Cereal Sci.* **11** (1990) 243–248.
65. P. L. Russel, G. Olivier: "The effect of pH and NaCl content on starch gel ageing. A study by differential scanning calorimetry and rheology," *J. Cereal Sci.* **10** (1989) 123–138.
66. R. F. Tester, W. R. Morrison: "Properties of damaged starch granules. V. Composition and swelling of fractions of wheat starch in water at various temperatures," *J. Cereal Sci.* **20** (1994) 175–181.

67. P. R. Shewry: "Cereal grain proteins," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal grain quality*, Chapman & Hall, London 1996, pp. 228–250.
68. A. S. Tatham, P. R. Shewry, P. S. Belton: "Structural studies of cereal prolamins, including wheat gluten," in Y. Pomeranz (ed.): *Advances in Cereal Science & Technology*, vol. **X**, American Association of Cereal Chemists, St. Paul, MN 1990, pp. 1–78.
69. P. R. Shewry: "The synthesis, processing, and deposition of gluten proteins in the developing wheat grain," *Cereal Foods World* **44** (1999) 587–589.
70. W. Bushuk: "Flour proteins: structure and functionality in breadmaking," *Cereal Foods World* **30** (1985) 447–451.
71. W. Bushuk, R. Tkachuk: *Gluten proteins*, American Association of Cereal Chemists, St. Paul, MN 1990.
72. F. MacRitchie: "Identifying the baking quality related components of wheat flour," *Cereal Foods World* **34** (1989) 548–552.
73. F. MacRitchie: "Physicochemical properties of wheat proteins in relation to functionality," *Adv. Food Nutr. Res.* **36** (1992) 1–87.
74. F. MacRitchie, D. L. du Cros, C. W. Wrigleys: "Flour polypeptides related to wheat quality," in Y. Pomeranz (ed.): *Advances in Cereal Science & Technology*, vol. **X**, American Association of Cereal Chemists, St. Paul, MN 1990, pp. 79–145.
75. J. D. Schofield: "Wheat proteins: structure and functionality in milling and breadmaking," in W. Bushuk, V. F. Rasper (eds.): *Wheat production, properties and quality*, Blackie Academic & Professional, Chapman & Hall, Glasgow 1994, pp. 73–106.
76. P. L. Weegels, R. J. Hamer, J. D. Schofield: "Functional properties of wheat glutenin," *J. Cereal Sci.* **23** (1996) 1–18.
77. J.-C. Autran: "Recent perspectives on genetics, biochemistry and functionality of wheat proteins," *Trends Food Sci. Technol.* **4** (1993) 358–364.
78. N. E. Pogna, P. Tusa, G. Boggini: "Genetic and biochemical aspects of dough quality in wheat," *Adv. Food. Sci.* **18** (1996) 145–151.
79. C. W. Wrigley in R. J. Hamer, R. C. Hoseney (eds.): *Interactions: the key to cereal quality*, American Association of Cereal Chemists, St. Paul, MN 1998, pp. 17–46.
80. G. Attenburrow, D. J. Barnes, A. P. Davies, S. J. Ingman: "Rheological properties of wheat gluten," *J. Cereal Sci.* **12** (1990) 1–14.
81. D. Marion, L. Dubreil, P. J. Wilde, D. C. Clark: "Lipids, lipid-protein interactions and the quality of baked cereal products," in R. J. Hamer, R. C. Hoseney (eds.): *Interactions: the key to cereal quality*, American Association of Cereal Chemists, St. Paul, MN 1998, pp. 131–167.
82. W. R. Morrison: "Lipids," in Y. Pomeranz (ed.): *Wheat chemistry and technology*, 3rd ed., American Association of Cereal Chemists (AACC), St. Paul, MN 1988, pp. 373–439.
83. W. R. Morrison: "Wheat Lipids," in W. Bushuk, V. F. Rasper (eds.): *Wheat production, properties and quality*, Blackie Academic & Professional, Chapman & Hall, Glasgow 1994, pp. 128–142.
84. G. McCormack, J. Panozzo, F. Békés, F. MacRitchie: "Contributions to breadmaking of inherent variations in lipid content and composition of wheat cultivars. I. Results of survey," *J. Cereal Sci.* **13** (1991) 255–261.
85. G. McCormack, J. Panozzo, F. MacRitchie: "Contributions to breadmaking of inherent variations in lipid content and composition of wheat cultivars. II. Fractionation and reconstitution studies," *J. Cereal Sci.* **13** (1991) 263–261.
86. J. B. Ohm, O. K. Chung: "Relationships of free lipids with quality factors in hard winter wheat flours," *Cereal Chem.* **79** (2002) 274–278.
87. A. Basic, B. A. Stone: "Cell walls and their components in cereal grain technology," in Y. Pomeranz (ed.): *Advances in Cereal Science and Technology*, vol. **III**, American Association of Cereal Chemists, St. Paul, MN 1986, pp. 207–295.
88. M. S. Izydorczyk, C. G. Biliaderis: "Structural heterogeneity of wheat endosperm arabinoxylans," *Cereal Chem.* **70** (1993) 641–646.
89. A.-M. A. Loosveld, P. J. Grobet, J. A. Delcour: "Contents and structural features of water-extractable arabinogalactan in wheat flour fractions," *J. Agric. Food Chem.* **45** (1997) 1998–2000.
90. F. Meuser, P. Suckow: "Non-starch polysaccharides," in J. M. V. Blanshard, P. J. Frazier, T. Galliard (eds.): *Chemistry and Physics of Baking*, Royal Society of Chemistry, London 1986, pp. 42–61.
91. H. Neukom: "Chemistry and properties of the non-starchy polysaccharides (NSP) of wheat flour," *Lebensm. Wiss. Technol.* **9** (1976) 143–148.

92. D. K. Sarker, P. J. Wilde, D. C. Clarke: "Enhancement of protein foam stability by formation of wheat arabinoxylan-protein crosslinks," *Cereal Chem.* **75** (1998) 493–499.
93. L. Saulnier, N. Peneau, J.-F. Thibault: "Variability in grain extract viscosity and water-soluble arabinoxylan content in wheat," *J. Cereal Sci.* **22** (1995) 259–264.
94. D. Sievert, W. Seibel, E. Rabe, K. Pfeilsticker: "Einfluss der Verfahrenstechnik auf Ballaststofffraktionen von Getreideerzeugnissen," *Ernährung/Nutrition* **13** (1989) 75–81.
95. F. Meuser, P. Suckow, W. Kulikowski, U. Meißner: "Beitrag des Brotes und anderer Lebensmittel auf Getreidebasis an der Ballaststoffversorgung der Bevölkerung der Bundesrepublik Deutschland," *Getreide Mehl Brot* **38** (1984) 211–214.
96. M. Wang, T. van Vliet, R. J. Hamer: "Interaction of water unextractable solids ands xylanase with gluten proteins: Effect of wheat cultivar," *J. Cereal Sci.* **41** (2004) 251–258.
97. M. Wang, R. J. Hamer, T. van Vliet, G. Oudgenoeg: "Interaction of water extractable pentosans with gluten proteins: effect on dough properties and gluten quality," *J. Cereal Sci.* **36** (2002) 25–37.
98. S. P. Roels, C. M. Courtin, J. A. Delcour: "Endoxylanases and arabinoxylans in gluten isolated in a batter system," *J. Cereal Sci.* **33** (2001) 53–57.
99. C. F. Ciacco, B. L. D'Appolonia: "Characterization of pentosans from different wheat flour classes and of their gelling capacity," *Cereal Chem.* **59** (1982) 96–100.
100. M. S. Izydorczyk, C. G. Biliaderis, W. Bushuk: "Oxidative gelation studies of water-soluble pentosans from wheat," *J. Cereal Sci.* **11** (1990) 153–169.
101. M. S. Izydorczyk, C. G. Biliaderis: "Influence of structure on the physicochemical properties of wheat arabinoxylan," *Carbohydrate Polym.* **17** (1992) 237–247.
102. H. Neukom, H. U. Markwalder: "Oxidative gelation of wheat flour pentosans: a new way to cross-linking polymers," *Cereal Foods World* **23** (1978) 374–376.
103. Yue Liao, R. A. Miller, R. C. Hoseney: "Role of hydrogen peroxide produced by Baker's yeast on dough rheology," *Cereal Chem.* **75** (1998) 612–616.
104. C. M. Courtin, J. A. Delcour: "Arabinoxylans and endoxylanases in wheat flour bread-making," *J. Cereal Sci.* **35** (2002) 225–244.
105. A. C. Johnson, R. C. Hoseney: "Chlorine treatment of cake flours. IV. Effects of storing and heating nondefatted and defatted flours," *Cereal Chem.* **57** (1980) 92–93.
106. C. A. Thomasson, R. A. Miller, R. C. Hoseney: "Replacement of chlorine treatment for cake flour," *Cereal Chem.* **72** (1995) 616–620.
107. Y. Pomeranz: " Biochemical, functional, and nutritive changes during storage," in D. B. Sauer (ed.): *Storage of cereal grains and their products*, American Association of Cereal Chemists (AACC), St. Paul, MN 1992, pp. 55–141.
108. L. Wang, R. A. Flores: "The effects of storage on flour quality and baking," *Food Rev. Int.* **15** (1999) 215–234.
109. C. Campos-Grijalva: 1997. Age related changes in freshly harvested hard wheat. MS thesis. Kansas State University, Manhatten, KS, U.S.A.
110. US 5 194 276, 1993 (R. C. Hoseney, J. M. Faubion, K. Shelke).
111. W. C. Mailhot, J. C. Patton: "Criteria of flour quality," Y. Pomeranz (ed.): *Wheat Chemistry and Technology*, vol. **II**, American Association of Cereal Chemists (AACC), St. Paul, MN 1988, pp. 69–90.
112. W. T. Yamazaki, C. T. Greenwood (eds.): *Soft wheat: production, breeding, milling, and uses*, American Association of Cereal Chemists, St. Paul, MN 1981.
113. R. Andersson, M. Hämäläinen, P. Aman: "Predictive modelling of the bread-making performance and dough properties of wheat," *J. Cereal Sci.* **20** (1994) 129–138.
114. *AACC Approved Methods*, 10th ed., American Association of Cereal Chemists (AACC), St. Paul, MN.
115. *ICC-Standards: Standard methods of the International Association for Cereal Science and Technology (ICC)*, Verlag Moritz Schäfer, Detmold.
116. S. Salmon: *Manual of methods for wheat and flour testing*, Guideline No. 3, Campden & Chorleywood Food Research Association, Gloucestershire–Hertfordshire 1999.
117. S. R. Delwiche, R. A. Graybosch, C. J. Peterson: "Predicting protein composition, biochemical properties, and dough-handling properties of hard red winter wheat flour by near-infrared reflectance," *Cereal Chem.* **75** (1998) 412–416.

118. C. S. Gaines, W. R. Windham: "Effect of wheat moisture content on meal apparent particle size and hardness scores determined by near-infrared reflectance spectroscopy," *Cereal Chem.* **75** (1998) 386–391.
119. H. Faridi (ed.): *Rheology of wheat products*, American Association of Cereal Chemists, St. Paul, MN 1985.
120. J. R. Olivier, H. M. Allen: "The prediction of bread baking performance using the farinograph and extensograph," *J. Cereal Sci.* **15** (1992) 79–89.
121. C. E. Walker, J. L. Hazelton: "Dough rheological test," *Cereal Foods World* **41** (1996) 23–28.
122. B. L. D'Appolonia, W. H. Kunerth (eds.): *The Farinograph Handbook*, 3rd ed., American Association of Cereal Chemists, St. Paul, MN 1984.
123. K. F. Finney, M. D. Shogren: "A ten-gram mixograph for determining and predicting functional properties of wheat flours," *Baker's Dig.* **46** (1972) 32–35, 38–42, 77.
124. V. F. Rasper, K. R. Preston (eds.): *The Extensigraph Handbook*, American Association of Cereal Chemists, St. Paul, MN 1991.
125. H. Faridi, V. F. Rasper: *The Alveograph Handbook*, American Association of Cereal Chemists, St. Paul, MN 1987.
126. K. Autio, L. Flander, A. Kinnunen, R. Heinonen: "Bread quality relationship with rheological measurements of wheat flour dough," *Cereal Chem.* **78** (2001) 654–657.
127. M. Menkovska, D. Knezevic, M. Ivanoski: "Protein allelic composition, dough rheology, and baking characteristics of flour mill streams and wheat cultivars with known and varied baking qualities," *Cereal Chem.* **79** (2002) 720–725.
128. K. H. Sutton, L. D. Simmons: "Molecular level protein composition of flour mill streams from a pilot-scale flour mill and its relationship to product quality," *Cereal Chem.* **83** (2006) 52–56.
129. J. E. Krueger: "Enzymes of sprouted wheat and their possible technological significance," in W. Bushuk, V. F. Rasper (eds.): *Wheat production, properties and quality*, Blackie Academic & Professional, Chapman & Hall, Glasgow 1994, pp. 143–153
130. R.. C. Hoseney, D. E. Rogers: "The formation and properties of wheat flour doughs," *Critical Reviews in Food Science and Nutrition* **29** (1990) 73–93.
131. W. Bushuk: "Interactions in wheat doughs," in R. J. Hamer, R. C. Hoseney (eds.): *Interactions: the key to cereal quality*, American Association of Cereal Chemists, St. Paul, MN 1998, pp. 1–16.
132. T. Amend, H.-D. Belitz, "Microstructural studies of gluten and a hypothesis on dough formation," *Food Structure* **10** (1991) 277–288.
133. H.-D. Belitz, W. Grosch: *Lehrbuch der Lebensmittelchemie*, Springer Verlag, Berlin 1992, pp. 608–668.
134. J. E. Bernardin, D. D. Kasarda: "Hydrated protein fibrils from wheat endosperm," *Cereal Chem.* **50** (1973) 529–536.
135. T. Amend, H.-D. Belitz: "Microscopical studies of water/flour systems," *Z. Lebensm. Unters. Forsch.* **189** (1989) 103–109.
136. B. J. Dobraszczyk, M. P. Morgenstern: "Rheology and the breadmaking process," *J. Cereal Sci.* **38** (2003) 229–246.
137. P. S. Belton: "On the elasticity of wheat gluten," *J. Cereal Sci.* **29** (1999) 103–107.
138. B. S. Khatkar, A. E. Bell, J. D. Schofield: "The dynamic rheological properties of glutens and gluten sub-fractions from wheats of good and poor bread making quality," *J. Cereal Sci.* **22** (1995) 29–44.
139. E. Lemelin, G. Branlard, L. Salvo, V. Lein, T. Aussenac, J. Dayde: "Breadmaking stability of wheat flours: Relation between mixing properties and mokecular weight distribution of polymeric glutenins," *J. Cereal Sci.* **42** (2005) 317–326.
140. B. S. Khatkar, R. J. Fido, A. S. Tatham, J. D. Schofield: "Functional properties of wheat gliadins. I. Effects of mixing characteristics and bread making quality," *J. Cereal Sci.* **35** (2002) 299–306.
141. B. S. Khatkar, R. J. Fido, A. S. Tatham, J. D. Schofield: "Functional properties of wheat gliadins. II. Effects on dynamic rheological properties of wheat gluten," *J. Cereal Sci.* **35** (2002) 307–314.
142. P. I. Payne, M. A. Nightingale, A. F. Krattiger, L. M. Holt: "The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties," *J. Sci. Food. Agric.* **40** (1987) 51–65.
143. F. MacRitchie: "Conversion of a weak flour to a strong one by increasing the proporation of its high molecular weight gluten protein," *J. Sci. Food. Agric.* **24** (1973) 1325–1329.

144. D. Lafiandra, S. Masci, C. Blumenthal, C. M. Wrigley: "The formation of glutenin polymer in practice," *Cereal Foods World* **44** (1999) 572–578.
145. F. MacRitchie: "Wheat proteins: Characterization and role in flour functionality," *Cereal Foods World* **44** (1999) 188–193.
146. P. R. Shewry, A. S. Tatham: "Disulfide bonds in wheat gluten proteins," *J. Cereal Sci.* **25** (1997) 207–227.
147. P. R. Shewry, N. G. Halford, A. S. Tatham: "High molecular weight subunits of wheat glutenin," *J. Cereal Sci.* **15** (1992) 105–120.
148. C. W. Wrigley, F. Békés: "Glutenin-protein formation during the continuum from anthesis to processing," *Cereal Foods World* **44** (1999) 562–565.
149. D. Kasarda: "Glutenin polymers: the in vitro to in vivo transition," *Cereal Foods World* **44** (1999) 566–571.
150. P. L. Weegels, A. M. Pijpekamp, A. Graveland, R. J. Hamer, J. D. Schofield: "Depolymerisation and Re-polymerization of wheat glutenin during dough processing. I. Relationships between glutenin macropolymer content and quality parameters," *J. Cereal Sci.* **23** (1996) 103–111.
151. J. H. Skerrit, L. Hac, F. Békés: "Depolymerization of the glutenin macropolymer during dough mixing. I. Changes in levels, molecular weight distribution, and overall composition," *Cereal Chem.* **76** (1999) 395–401.
152. J. H. Skerrit; L. Hac, M. P. Lindsay, F. Békés: "Depolymerization of the glutenin macropolymer during dough mixing. II. Differences in retention of specific glutenin subunits," *Cereal Chem.* **76** (1999) 402–409.
153. D. Every, S. C. Morrison, L. D. Simmons, M. P. Ross: "Distribution of glutathione in millstreams and relationship to chemical and baking properties of flour," *Cereal Chem.* **83** (2006) 57–61.
154. R. A. Miller, R. C. Hoseney: "Effect of oxidation on the dynamic rheological properties of wheat-flour-water-doughs," *Cereal Chem.* **76** (1999) 100–104.
155. D. Sievert, H. D. Sapirstein, W. Bushuk: "Changes in electrophoretic patterns of acetic acid-insoluble wheat flour proteins during dough mixing," *J. Cereal Sci.* **14** (1991) 243–256.
156. L. F. Schroeder, R. C. Hoseney: "Mixograph studies. II. Effect of activated double-bond compounds on dough mixing properties," *Cereal Chem.* **55** (1978) 348–360.
157. J. S. Sidhu, P. Nordin, R. C. Hoseney: "Mixograph Studies. III. Reaction of fumaric acid with gluten proteins during dough mixing," *Cereal Chem.* **59** (1980) 159–163.
158. R. C. Hoseney, R. A. Brown: "Mixograph studies. V.Effect of pH," *Cereal Chem.* **60** (1983) 124–126.
159. F. Békés, P. W. Gras: "In vitro studies on gluten protein functionality," *Cereal Foods World* **44** (1999) 580–586.
160. V. Tolstoguzov: "Thermodynamic aspects of dough formation and functionality," *Food Hydrocolloids* **11** (1997) 181–193.
161. G. M. Campbell, C. D. Rielly, P. J. Fryer, P. A. Sadd: "Aeration of bread dough during mixing: effect of mixing dough at reduced pressure," *Cereal Foods World* **43** (1998) 163–167.
162. H. Goesaert, K. Brijs, W. S. Veraverbeke, C. M. Courtin, K. Gebruers, J. A. Delcour: "Wheat flour constituents: how they impact bread quality, and how to impact their functionality," *Trends. Food Sci. Technol.* **16** (2005) 12–30.
163. K. Autio, T. Laurikainen: "Relationships between flour/dough microstructure and dough handling and baking properties," *Trends Food Sci. Technol.* **8** (1997) 181–185.
164. S. A. Matz: *Ingredients for bakers*, Pan-Tech International, Inc., McAllen, TX 1996.
165. C. E. Stauffer: "Fats and oils in bakery products," *Cereal Foods World* **43** (1998) 120–126.
166. A. H. Bloksma, W. Bushuk: "Rheology and chemistry of dough," in Y. Pomeranz (ed.): *Wheat Chemistry and Technology*, vol. **II**, American Association of Cereal Chemists, St. Paul, MN 1998, pp. 131–217.
167. D. Every, L. Simmons, K. H. Sutton, M. Ross: "Studies on the mechanism of ascorbic-acid improver effect on bread using flour fractionation and reconstitution methods," *J. Cereal Sci.* **30** (1999) 147–158.
168. W. Grosch: "Mechanismus der Ascorbinsäurewirkung," *Getreide Mehl Brot* **52** (1998) 267–269.
169. W. Grosch, H. Wieser: "Redox reactions in wheat dough as affected by ascorbic acid," *J. Cereal Sci.* **29** (1999) 1–16.
170. W. G. Allen: "Alternative oxidants as dough conditioners," *Cereal Foods World* **44** (1999) 642–649.

171. B. Dupuis: "The chemistry and toxicology of potassium bromate," *Cereal Foods World* **42** (1997) 171–181.
172. E. Armero, C. Collar: "Crumb firming kinetics of wheat bread with antistaling additives," *J. Cereal Sci.* **28** (1998) 165–174.
173. H. Moonen: "Die Bedeutung von Enzymen für die Herstellung von Backwaren," *Getreide Mehl Brot* **52** (1998) 270–272.
174. B. Poldermans, P. Schoppink: "Contolling the baking process and product quality by enzymes," *Cereal Foods World* **44** (1999) 132–135.
175. J. Qi Si: "Synergistic effects of enzymes for breadmaking," *Cereal Foods World* **42** (1997) 802–807.
176. K. R. Morgan et al.: "Staling in starch breads: the effect of antistaling α-amylase," *Starch/Stärke* **49** (1997) 54–59.
177. G. Cleemput, C. Booij, M. Hessing, H. Gruppen, J. A. Delcour: "Solubilization and changes in molecular weight distribution of arabinoxylans and protein in wheat flours during bread-making, and the effects of endogenous arabinoxylan hydolysing enzymes," *J. Cereal Sci.* **26** (1997) 55–66.
178. C. M. Courtin, A. Roelants, J. A. Delcour: "Fractionation-reconstitution experiments provide insight in the role of endoxylanases in breadmaking," *J. Agric. Food Chem.* **47** (1999) 1870–1877.
179. C. M. Courtin, G. Gelders, J. A. Delcour: "The use of two endoxylanases with different substrate selectivity provides insight into arabinoxylan functionality in wheat flour breadmaking," *Cereal Chem.* **78** (2001) 564–571.
180. W. Debyser, G. Derdelinckx, J. A. Delcour: "Arabinoxylan solubilisation and inhibition of the barley malt xylanolytic system by wheat during brewing with wheat wholemeal adjunct: evidence for a new class of enzyme inhibitors," *J. Am. Soc. Brew. Chem.* **55** (1997) 153–156.
181. W. Debyser, W. J. Peumans, E. J. M. Vandamme, J. A. Delcour: "Triticum aestivum xylanase inhibitor (TAXI), a new class of enzyme inhibitor affecting breadmaking performance," *J. Cereal Sci.* **30** (1999) 39–43.
182. K. Gebruers, W. Debyser, H. Goesaert, P. Proost, J. Van Damme, J. A. Delcour: "Triticum aestivum L. endoxylanase inhibitor (TAXI) consists of two inhibitors, TAXI I and TAXI II, with different specificities," *Biochem. J.* **353** (2001) 239–244.
183. S. Sansen et al.: "Structural basis for inhibition of Aspergillus niger xylanase by Triticum aestivum xylanase inhibitor-I," *J. Biol. Chem.* **279** (2004) 36022–36028.
184. I. Trogh, J. F. Soerensen, C. M. Courtin, J. A. Delcour: "Impact of inhibition sensitivity on endoxylanase functionality in wheat flour breadmaking," *J. Agric. Food Chem.* **52** (2004) 4296–4302.
185. R. Kieffer, W. Grosch: "Verbesserung der Backeigenschaften von Weizenmehlen durch die Typ II-Lipoxygenase aus Sojabohnen," *Z. Lebensm. Unters. Forsch.* **170** (1980) 258–261.
186. D. Every, L. D. Simmons, M. P. Ross: "Distribution of redox enzymes in millstreams and relationship to chemiscal and baking properties of flour," *Cereal. Chem.* **83** (2006) 62–68.
187. L. Rakotozafy et al.: "Effect of adding exogenous oxidative enzymes on the activity of three endogenous oxidoreductases during mixing of wheat flour dough," *Cereal Chem.* **76** (1999) 213–218.
188. M. Motoki, K. Seguro: "Transglutaminase and its use for food processing," *Trends Food Sci. Technol.* **9** (1998) 204–210.
189. N. Bauer, P. Koehler, H. Wieser, P. Schieberle: "Studies on effects of microbial transglutaminase on gluten proteins of wheat II. Rheological properties," *Cereal Chem.* **80** (2003) 787–790.
190. N. Bombara, M. C. Anon, A. M. R. Pilosof: "Functional properties of protease modified wheat flours," *Lebensm. Wiss. Technol.* **30** (1997) 441–447.
191. B. Jensen, P. Drube: "Effekt und Wirksamkeit von Lipasen in Teig und Brot," *Getreide Mehl Brot* **52** (1998) 153–157.
192. J. K. Chavan, S. S. Kadam: "Nutritional enrichment of bakery products by supplementation with nonwheat flours," *Crit. Rev. Food Sci. Nutr.* **33** (1993) 189–226.
193. R. C. Hoseney: "Bread baking," *Cereal Foods World* **39** (1994) 180–183.
194. E. J. Pyler: *Baking science and technology*, 3rd ed., vol. **I + II**, Sosland Publishing Company, Merriam, KS 1988.
195. W. Seibel: "Recent research progress in bread baking technology," in W. Bushuk, V. F. Rasper (eds.): *Wheat-Production, properties and quality*, Blackie Academic & Professional, Glasgow, imprint of Chapman & Hall, London 1994, pp. 59–72.
196. C. A. Stear: *Handbook of breadmaking technology*, Elsevier Science Pub., Essex 1990.

197. M. Gobetti: "The sourdough microflora: interactions of lactic acid bacteria and yeasts," *Trends Food Sci. Technol.* **9** (1998) 267–274.
198. F. Meuser: "Development of fermentation technologies in modern bread factories," *Cereal Foods World* **40** (1995) 114–122.
199. F. Vogel: "Sourdough-a traditional product with a great future," *Adv. Food Sci. (CMTL)* **18** (1996) 130–244.
200. K. Kulp, K. Lorenz, J.-M. Brümmer: *Frozen and refrigerated doughs and batters*, American Association of Cereal Chemists, St. Paul, MN 1995.
201. P. T. Berglund, D. R. Shelton, T. P. Freeman: "Frozen bread dough ultrastructure as affected by duration of frozen storage and freeze-thaw cycles," *Cereal Chem.* **68** (1991) 105–107.
202. J. Räsänen, J. M. V. Blanshard, J. R. Mitchell, W. Derbyshire, K. Autio: "Properties of frozen wheat doughs at subzero temperatures," *J. Cereal. Sci.* **28** (1998) 1–14.
203. A. Baier-Schenk, S. Handschin, M. von Schonau, A. G. Bittermann, T. Bachi, B. Conde-Petit: "In situ observations of the freezing process in wheat dough by confocal laser scanning microscopy (CLSM): formation of ice and changes in gluten mnetwork," *J. Cereal Sci.* **42** (2005) 255–260.
204. S. Zounis, K. J. Quail, M. Wootton, M. R. Dickson: "Studying frozen dough structure using low-temperature scanning electron microscopy," *J. Cereal Sci.* **35** (2002) 135–148.
205. P. Gélinas, I. Deaudelin, M. Grenier: "Frozen dough: effects of dough shape, water content, and sheeting-molding conditions," *Cereal Foods World* **40** (1995) 124–126.
206. S. Kenny, K. Wehrle, M. Auty, E. K. Arendt: "Influence of sodium caseinate and whey protein on baking properties and rheology of frozen dough," *Cereal Chem.* **78** (2001) 458–463.
207. C. E. Perron, O. M. Lukow, W. Bushuk, F. Townley-Smith: "The blending potential of diverse wheat cultivars in a frozen dough system," *Cereal Foods World* **44** (1999) 667–672.
208. G. Reed, H. J. Peppler: *Yeast Technology*, Avi Publishing Co., Westport, CT 1973.
209. R. C. Hoseney, H. K. Hsu, R. C. Junge: "A simple spread test to measure the rheological properties of fermenting dough," *Cereal Chem.* **56** (1979) 141–143.
210. E. B. Bennion, G. S. T. Bamford in A. J. Bent (ed.): *The Technology of Cake Making*, Blackie Academic and Professional, Chapman & Hall, London 1997.
211. B. B. Heidolph: "Designing chemical leavening systems," *Cereal Foods World* **41** (1996) 118–126.
212. J. C. Baker, M. D. Mize: "The origin of the gas cell in bread dough," *Cereal Chem.* **18** (1941) 19–34.
213. H. M. Elmehdi, J. H. Page, M. G. Scanlon: "Using ultrasound to investigate the cellular structure of bread crumb," *J. Cereal Sci.* **38** (2003) 33–42.
214. J. C. Baker: "The permeability of bread by air," *Cereal Chem.* **16** (1939) 730–733.
215. Z. Gan, P. R. Ellis, J. D. Schofield: "Gas cell stabilization and gas retention in wheat bread dough," *J. Cereal Sci.* **21** (1995) 215–220.
216. P. L. Weegels, R. J. Hamer: "Temperature-induced changes of wheat products," in R. J. Hamer, R. C. Hoseney (eds.): *Interactions: the key to cereal quality*, American Association of Cereal Chemists, St. Paul, MN 1998, pp. 95–130.
217. Y. Yin, C. E. Walker: "A quality comparison of breads baked by conventional versus nonconventional ovens: a review," *J. Sci. Food Agric.* **67** (1995) 283–291.
218. T. Van Vliet, A. M. Janssen, A. H. Bloksma, P. Walstra: "Strain Hardening of dough as a requirement for gas retention," *J. Text. Studies* **23** (1992) 439–460.
219. B. J. Dobraszczyk, J. Smewing, M. Albertini, G. Maesmans, J. D. Schofield: "Extensional rheology and stability of gas cell walls in bread doughs at elevated temperatures in relation to breadmaking performance," *Cereal Chem.* **80** (2003) 218–224.
220. K. F. Finney, M. A. Barmore: "Loaf Volume and protein content of hard winter and spring wheats," *Cereals Chem.* **25** (1948) 291–312.
221. H. He, R. C. Hoseney: "Effect of the quantity of wheat flour protein on bread loaf volume," *Cereal Chem.* **69** (1992) 17–19.
222. M. C. Zghal, M. G. Scanlon, H. D. Sapirstein: "Effects of flour strength, baking absorption, and processing conditions on the structure and mechanical properties of bread crumb," *Cereal Chem.* **78** (2001) 1–7.
223. K. Ghiasi, R. C. Hoseney, D. R. Lineback: "Characterization of soluble starch from bread crumb," *Cereal Chem.* **56** (1979) 485–490.
224. M. S. Feather: "The mechanism of nonenzymatic browning reactions occuring during cookie and cracker making," in H.

Faridi (ed.): *The Science of Cookie and Cracker Production*, Chapman & Hall, New York–London 1994, pp. 439–453.
225. P. Schieberle: "Intense aroma compounds-useful tools to monitor the influence of processing and storage on bread aroma," *Adv. Food Sci. (CMTL)* **18** (1996) 237–244.
226. W. Grosch, P. Schieberle: "Flavor of cereal products-A review," *Cereal Chem.* **74** (1997) 91–97.
227. G. Zehentbauer, W. Grosch: "Crust aroma of baguettes.II. Dependence of the concentrations of key odorants on yeast level and dough processing," *J. Cereal Sci.* **28** (1998) 93–96.
228. J. M. V. Blanshard, P. J. Lillford (eds.): *The glassy state of foods*, Nottingham University press, Nottingham 1993.
229. A. Eisenberg: "The glassy state and the glass transition," in J. E. Mark, A. Eisenberg, W. W. Gressley, L. Mandelkern, J. L. Koenig (eds.): *Physical Properties of Polymers*, American Chemical Society, Washington, D.C. 1984, pp. 55–96.
230. R. W. Hartel: "Controlling sugar crystallization in food products," *Food Technol.* **47** (1993) 99–107.
231. T. P. Labuza, C. R. Hyman: "Moisture migration and control in multi-domain foods," *Trends Food Sci. Technol.* **9** (1998) 47–5.
232. T. R. Noel, S. G. Ring, M. A. Whittham: "Glass transitions in low-moisture foods," *Trends Food Sci. Technol.* **1** (1990) 62–67.
233. Y. Roos: "Characterization of food polymers using state diagrams," *J. Food Eng.* **24** (1995) 339–360.
234. Y. Roos, M. Karel: "Applying state diagrams to food processing development," *Food Technol.* **45** (1991) 66–71, 107.
235. H. Levine, L. Slade: "Water as plasticizer: Physico-chemical aspects of low-moisture polymeric systems," in F. Frankd (ed.): *Water Science Reviews*, vol. **3**, Cambridge University Press, Cambridge 1998, pp. 79–185.
236. H. Levine, L. Slade: "Glass transition in foods," in H. G. Schwartzberg, R. W. Hartel (eds.): *Physical Chemistry of Foods*, Marcel Dekker, New York 1992, pp. 83–222.
237. L. Slade, H. Levine: "Beyond water activity: recent advances based on an alternative approach to the assessment of food quality and safety," *Crit. Rev. Food. Sci. Nutr.* **30** (1991) 115–360.
238. L. Slade, H. Levine, J. Ievolella, M. Wang: "The glassy state phenomenon in applications for the food industry: application of the food science polymer approach to structure-function relationships of sucrose in cookie and cracker systems," *J. Sci. Food Agric.* **63** (1993) 133–176.
239. R. J. Nicholls, I. A. M. Appelqvist, A. P. Davies, S. J. Ingman, P. J. Lillford: "Glass transition and the fracture behaviour of gluten and starches within the glassy state," *J. Cereal Sci.* **21** (1995) 25–36.
240. G. Roudaut, G. Maglione, M. Le Mestre: "Relaxations below glass transition temperature in bread and its components," *Cereal Chem.* **76** (1999) 78–81.
241. G. Roudaut, G. Maglione, D. van Dusschoten, M. Le Mestre: "Molecular mobility in glassy bread: a multispectroscopy approach," *Cereal Chem.* **76** (1999) 70–77.
242. M. LeMeste, V. T. Huang, J. Panama, G. Anderson, R. Lentz: "Glass transition of bread," *Cereal Foods World* **37** (1992) 264–267.
243. F. Meuser, J.-M. Brümmer, W. Seibel: "Bread varieties in central Europe," *Cereal Foods World* **39** (1994) 222–230.
244. W. Seibel: "Wheat usage in Western Europe," in H. Faridi, J. Faubion (eds.): *Wheat end uses around the world*, American Association of Cereal Chemists (AACC), St. Paul, MN 1995, pp. 93–125.
245. W. Seibel: "Weizenmehlqualität und Produktvielfalt in Westeuropa," *Getreide Mehl Brot* **50** (1996) 316–319.
246. H. Faridi: "Flat breads," in *Wheat chemistry and technology*, 3rd ed., vol. **II**, American Association of Cereal Chemists (AACC), St. Paul, MN 1998, pp. 457–506.
247. J. Qarooni: *Flat bread technology*, Chapman & Hall LTD, New York 1996.
248. K. J. Quail: *Arabic bread production*, American Association of Cereal Chemists (AACC), St. Paul, MN 1996.
249. V. Dally, L. Navarro: "Flour tortillas: a growing sector of the U.S. food industry," *Cereal Foods World* **44** (1999) 457–459.
250. R. D. Waniska: "Perspectives on flour tortillas," *Cereal Foods World* **44** (1999) 471–473.
251. P. Yue, P. Rayas-Duarte: "An overview of steamed bread," *Cereal Foods World* **42** (1997) 210–215.
252. H. Faridi, C. Gaines, P. Finney: "Soft wheat quality in production of cookies and crackers," in W. Bushuk, V. F. Rasper (eds.): *Wheat*

production, properties and quality, Blackie Academic & Professional, Chapman & Hall, Glasgow 1994, pp. 154–168.
253. C. S. Gaines, A. Kassuba, P. L. Finney: "Influence of eight flours on the hardness of commercial cookies and crackers," *Cereal Foods World* **39** (1994) 160–167.
254. N. Almond: *Biscuits, Cookies and Crackers*, vol. **2**, Elsevier Applied Science, New York–London 1989.
255. H. Faridi (ed.): *The Science of Cookie and Cracker Production*, Chapman & Hall, New York–London 1994.
256. K. Kulp: *Cookie Chemistry and Technology*, The American Institute of Baking, Manhatten, KS 1994.
257. D. Manley: *Technology of Biscuits, Crackers and Cookies*, 2nd ed., Ellis Horwood, New York 1991.
258. P. Wade: *Biscuits, Cookies and Crackers*, vol. **1**, Elsevier Applied Science, New York–London 1988.
259. R. Lawson: "Mathematical modeling of cookie and cracker ovens," in H. Faridi (ed.): *The Science of Cookie and Cracker Production*, Chapman & Hall, New York–London 1994, pp. 387–438.
260. G. Trystam, M. Allache, F. Courtois: "Automatic control of the biscuit baking oven process," in T. Yano, R. Matsuno, K. Nakamura (eds.): *Developments in Food Engineering, Proceedings of the 6th International Congress on Engineering and Food*, Blackie Academic & Professional, an imprint of Chapman & Hall, London 1994, pp. 975–977.
261. C. S. Gaines, A. Kassuba, P. L. Finney: "Using wire-cut and sugar-snap formula cookie test baking methods to evaluate distinctive soft wheat flour sets: implications for quality testing," *Cereal Foods World* **41** (1996) 155–160.
262. C. S. Gaines: "Prediction of sugar-snap cookie diameter using sucrose solvent retention capacity, milling softness, and flour protein," *Cereal Chem.* **81** (2004) 549–552.
263. R. A. Miller, R. C. Hoseney: "Factors in hard wheat flour responsible for reduced cookie spread," *Cereal Chem.* **74** (1997) 330–336.
264. L. Slade, H. Levine: "Structure-function relationships of cookie and cracker ingredients," in H. Faridi (ed.): *The Science of Cookie and Cracker Production*, Chapman & Hall, New York–London 1994, pp. 23–141.
265. W. T. Yamazaki, D. D. Lord: "Soft wheat products," in Y. Pomeranz (ed.): *Wheat Chemistry and Technology*, 2nd ed., American Association of Cereal Chemists, St. Paul, MN 1971, pp. 743–776.
266. H. Yamamoto, S. T. Worthington, G. Hou, P. K. W. Ng: "Rheological properties and baking qualities of selected soft wheats grown in the United States," *Cereal Chem.* **73** (1996) 215–221.
267. R. A. Miller, R. Mathew, R. C. Hoseney: "Use of a thermomechanical analyzer. Study of an apparent glass transition in cookie dough," *J. Therm. Anal.* **47** (1996) 1329–1338.
268. A. Baltsavias, A. Jurgens, T. van Vliet: "Fracture properties of short-dough biscuits: effect of composition," *J. Cereal Sci.* **29** (1999) 235–244.
269. A. Baltsavias, A. Jurgens, T. van Vliet: "Properties of short-dough biscuits in relation to structure," *J. Cereal Sci.* **29** (1999) 245–255.
270. L. Pizza, P. Masi: "Development of crispness in cookies during baking in an industrial oven," *Cereal Chem.* **74** (1997) 135–140.
271. K. Tiefenbacher, M. Dobrovics: "Mehle für die Waffelherstellung," *Getreide Mehl Brot* **53** (1999) 349–353.
272. R. C. Hoseney, P. Wade, J. W. Finley: "Soft Wheat Products," in Y. Pomeranz (ed.): *Wheat chemistry and technology*, 3rd ed., vol. **II**, American Association of Cereal Chemists (AACC), St. Paul, MN 1998, pp. 407–506.
273. B. E. Brooker: "The stabilisation of air in cake batters — the role of fat," *Food Structure* **12** (1993) 285–296.
274. S. Cauvain, J. Cyster: *Sponge cake technology*, Campden & Chorleywood Food Research Association, Gloucestershire–Hertfordshire 1996.
275. G. W. Telloke: "Die Blätterteiglockerung," *Getreide Mehl Brot* **43** (1989) 116–121.
276. G. W. Telloke: "Optimierung physikalischer Ziehfetteigenschaften," *Getreide Mehl Brot* **48** (1994) 74–80.
277. Z. Czuchajowska, Y. Pomeranz: "Differential scanning calorimetry, water activity, and moisture contents in crumb and near-crust zones of bread during storage," *Cereal Chem.* **66** (1989) 305–309.
278. J. B. Boussingault: 1852. De la Transformation du Pain tender en Pain Rassis. Ann. Chem. Phys. III, Series XXXVI, p. 490.
279. D. Every, J. A. Gerrard, M. J. Gilpin, M. Ross, M. P. Newberry: "Staling in starch bread: the effect of gluten additions on specific loaf

volume and firming rate," *Starch/Stärke* **50** (1998) 443–446.
280. M. L. Martin, K. J. Zeleznak, R. C. Hoseney: "A mechanism of bread firming. I. Role of starch swelling," *Cereal Chem.* **68** (1991) 498–502.
281. M. L. Martin, R. C. Hoseney: "A mechanism of bread firming. II. Role of starch hydrolyzing enzymes," *Cereal Chem.* **68** (1991) 503–507.
282. B. K. Patel, R. D. Waniska, K. Seetharaman: "Impact of different baking process on bread firmness and starch properties in bread-crumb," *J. Cereal Sci.* **42** (2005) 173–184.
283. R. A. Graybosch, E. Souza, W. Berzonsky, P. S. Baenziger, O. Chung: "Functional properties of waxy wheat flours: genotypic and environmental," *J. Cereal Sci.* **38** (2003) 69–76.
284. T. Maeda, T. Hashimoto, M. Minoda, S.-i. Tamagawa, N. Morita: "Effects of mutant thermostable alpha-amylase on rheological properties osf wheat dough and bread," *Cereal Chem.* **80** (2003) 722–727.
285. P. Hy: "Baking new strategies," *Food Ingredients and Analysis International* (1998) nos. 3 + 4, 27–28, 30.
286. B. DeCindio, S. Correra: "Mathematical modelling of leavened cereal goods," *J. Food Eng.* **24** (1995) 379–403.
287. L. S. Young: "Getting the product right," *Food Manufacture* (1995) no. 4, 43–46.
288. W. L. Graham: "Real science and really great taste for bakery product success," *Cereal Foods World* **43** (1998) SR-10–SR-15.
289. J. Grider: "A snapshot of the U.S. baking industry," *Cereal Foods World* **41** (1996) SR-5–SR-6.
290. F. Katz: "Indulging in foods that make you feel good," *Food Technol.* **53** (1999) 46–67.
291. S. Kohn: "An update on the U.S. baking industry," *Cereal Foods World* **43** (1998) SR-4–SR-7.
292. S. Kohn: "Update on the U.S. baking industry," *Cereal Foods World* **44** (1999) SR-5–SR-7.
293. A. E. Sloan: "Top ten trends to watch and work on for the millenium," *Food Technol.* **53** (1999) 40–60.
294. S. L. DeFelice: "The nutraceutical revolution: its impact on food industry and R& D," *Trends Food Sci. Technol.* **6** (1995) 59–61.
295. S. L. DeFelice: "The time has come for nutraceutical cereals," *Cereal Foods World* **40** (1995) 51–52.
296. P. O'Carroll: "The shape of functional foods," *The World of Ingredients* (1999) Sept., 54–57.
297. A. E. Sloan: "Positive eating and problem treating: neutraceuticals and cereal-based foods in the 21st century," *Cereal Foods World* **44** (1999) 746–750.
298. W. Andlauer, P. Fürst, "Does cereal reduce the risk of cancer?" *Cereal Foods World* **44** (1999) 76–78.
299. R. Chaudari: "Foods for the future: the impact of functional foods in the cereal industry," *Cereal Foods World* **44** (1999) 94–95.
300. E. Rabe, D. Sievert: "Effects of baking, pasta production, and extrusion cooking on formation of resistant starch," *Europ. J. Clinical Nutrition* **46** (1992) Suppl. 2, Oct., S105–S107.
301. W. Seibel: "Funktionelle Lebensmittel auf Getreidebasis," *Getreide Mehl Brot* **52** (1998) 185–187.

Bromine

JACK F. MILLS, The Dow Chemical Co., Midland, Michigan 48640, United States

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 4
3.1. Bromine Addition to Alkenes and Alkynes 5
3.2. Bromination of Aromatic Compounds 5
3.3. Free-Radical Bromination 6
3.4. Production of Hydrogen Bromide . . . 6
3.5. Bromination Using Bromine Carriers 6
4. Production 6
4.1. Steaming-Out Process 7
4.2. Seawater Process 8
4.3. Electrolysis 9
4.4. Bromine from Recycle and Waste Brines 9
4.5. Purification 10
5. Quality Specifications and Testing . . 10
6. Storage and Transportation 10
7. Uses 11
8. Economic Aspects 11
9. Toxicology and Occupational Health . 12
10. References 13

1. Introduction

Bromine, Br [*7726-95-6*], a nonmetallic element of the halogen family, is a dark red-brown liquid at ordinary temperatures and pressures. It appears in group 7 A, period 4 of the periodic table between chlorine and iodine. The name is derived from the Greek $\beta\varrho\tilde{\omega}\mu o\sigma$, meaning stench. It vaporizes readily and has a sharp, penetrating odor. The electronic configuration is: $1s^2$, $2s^2$, $2p^6$, $3s^2$, $3p^6$, $4s^2$, $3d^{10}$, $4p^5$; the most common oxidation states are -1 and +5. Valences of +3, +1, and +7 are also observed.

Bromine was first discovered by A. J. BALARD in 1824. He had isolated the element while studying the flora of a salt marsh near Montpellier, France. He noticed a deposit of sodium sulfate which had crystallized in a pan containing mother liquor from common salt. Investigating uses for the waste liquors, BALARD observed that upon saturation with chlorine, the new red liquid could be obtained by distillation [1]. CARL LÖWIG [2] independently isolated bromine at Kreuznach, Germany, in 1825 and was in the process of studying its properties when BALARD's paper appeared. The first mineral to contain bromine was discovered in 1841 by BERTHIER; the mineral was apparently bromyrite (silver bromide).

The first commercial production of bromine was from native salt brines in 1846 at Freeport, Pennsylvania, by DAVID ALTER. The Dow process for extracting bromine from seawater began in 1934 near Wilmington, North Carolina.

Bromine does not occur in nature as the free element but is found solely as the bromide. There are only a few bromine-containing minerals, all silver ores, among which are bromyrite, AgBr; embolite, Ag(Cl, Br); and iodobromite, Ag(Br, Cl, I). Bromine usually follows chlorine through the geochemical cycle. Crystal rocks contain 2 – 3 ppm bromine.

Bromine also appears in variable amounts in the biosphere. The bromide concentration in terrestrial plants is about 7 ppm and in freshwater plants, 22 ppm. The amounts of bromide found in some edible foods are wheat flour, 19.7 ppm; potatoes, 10.3 ppm; oranges, 3 ppm; and apples, 4 ppm.

In the animal world, the highest bromide contents are found in sea life, such as fish, sponges, and crustaceans [3]. Many natural products containing organic bromine have been isolated, such as dibromoindigo (Tyrian purple) from sea snails, bromoform from red algae, and brominated phenols and diphenyl oxides from sponges and marine bacteria [4]. Animal tissues contain between 1 and 9 ppm bromide and blood bromide ranges between 5 and 15 ppm.

The most easily recoverable bromine occurs in the hydrosphere as soluble bromide salts. The concentration of bromine in seawater is about 65 mg/L. Other sources of extractable bromine occur in salt lakes, salines, and inland seas (Ta-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_391

ble 1). The Dead Sea is one of the richest sources, containing nearly 4 g/L of bromide at the surface and up to 6 g/L at deeper levels.

Table 1. Bromine occurrence

	Original content, g/L	After concentration, g/L
Sea water	0.065	
Inland seas		
Kharaghoda, Indian Ocean	0.25	6
Ssassyksee, Black Sea	0.28	1.5 – 4.5
Shebhka el Melah, Tunisia	2.5	6.8
Salt lakes		
Dead Sea, Israel	4 – 6	12 – 13
Ellon Sea, USSR	0.63 – 2.25	
Searles Lake, Calif. (USA)	0.85	
Brine wells		
Michigan, (USA)	2 – 3	
Arkansas, (USA)	4 – 5	
Yakutsk, Siberia, USSR	6 – 7 ($CaCl_2$ solution)	
Mineral salt deposits	wt %	
Rock salt	0.005 – 0.040	
Sylvite	0.117 – 0.300	
Carnallite	0.155 – 0.334	
Bischofit	0.467	
Tachydrite	0.438	
Hard salt	0.05 – 0.20	

The most important source of bromine is brine wells, which is the principal source in the United States. The richest brines are found in Arkansas and Michigan, with bromide content ranging from 2 to 5 g/L. The total bromine content in the crust of the earth has been estimated to be $10^{15} - 10^{16}$ t or 0.00016 % bromine.

2. Physical Properties

Bromine [*7726-95-6*], ^{20}Br, M_r 159.808 (for Br_2), stable isotopes 79 (50.54 %), 81 (49.46 %), is a dense, dark red, fuming, highly corrosive, and lacrimatory liquid. Its strong pungent odor is detectable at 1 ppm by volume in air. The color of bromine varies with temperature (yellow orange at 20 K, orange red, red brown, to almost black at its melting point). Liquid bromine is always dark red, whereas, bromine vapors are usually orange to red brown. The absorption band for liquid bromine lies between 417–420 nm [5].

The physical properties of elemental bromine are summarized in Table 2. The vapor pressure curve is shown in Figure 1.

Table 2. Physical properties of bromine [6–9]

fp, °C	−7.25
bp, °C	58.8
Density, g/cm^3, 15 °C	3.1396
20 °C	3.1226
25 °C	3.1055
30 °C	3.0879
Vapor density, g/L, 0 °C, 101.3 kPa	7.139
Refractive index, 20 °C	1.6083
25 °C	1.6475
Viscosity, mm^2/s (= cSt), 20 °C	0.314
30 °C	0.288
40 °C	0.264
50 °C	0.245
Surface tension, mN/m (= dyn/cm), 25 °C	40.9
Solubility parameter, 25 °C	11.5
Critical temperature, °C	311
Critical pressure, MPa (atm)	10.3 (102)
Thermal conductivity, W m^{-1} K^{-1}	0.123
Specific conductivity, $(\Omega \cdot cm)^{-1}$	9.10×10^{-12}
Dielectric constant, 25 °C, 10^5 Hz	3.33
Electrical resistivity, 25 °C, $\Omega \cdot cm$	6.5×10^{10}
Expansion coefficient at 20 – 30 °C, K^{-1}	0.0011
Compressibility, 20 °C (0 – 10 MPa)	62.5×10^{-6}
Heat of vaporization, J/g, 50 °C	187
Heat of fusion, J/g, −7.25 °C	66.11
Heat capacity, J/mol, 15 K [10]	7.217
30 K	22.443
60 K	36.33
240 K	57.94
solid, 265.9 K	61.64
liquid, 265.9 K	77.735
288.15 K	78.66
Electronegativity	3.0
Electron affinity, kJ	330.5

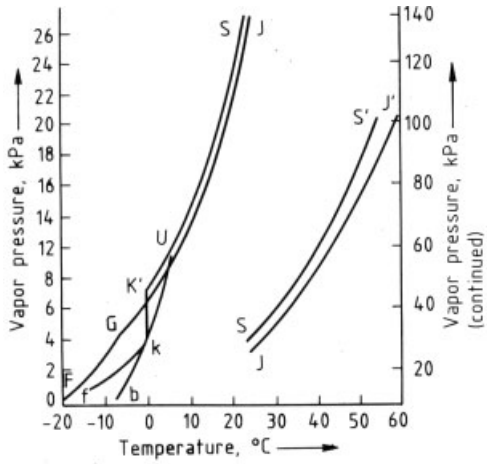

Figure 1. Vapor pressure curves for Br_2 and water saturated with Br_2
FGJ – JJ′: vapor-pressure curve of Br_2; f (b)k(K′)US – SS′: vapor-pressure curve of Br_2 – containing water; f – K: Br_2 hydrate + ice + vapor; b – K (metastable), k – U (stable): Br_2 hydrate + solution + vapor; k – K′ (metastable): ice + solution + vapor; K′ – U (metastable), U – S, S – S′ (stable): Br_2 + solution + vapor; for S – S′, J – J′ use right-hand axis

The density of liquid bromine (in g/L) in the range 0 – 50 °C can be calculated using the following equation (T is temperature in K):

$$\varrho = 4.1295 - 0.003447T$$

The dynamic viscosity, in Pa · s, can be determined (T is temperature in K):

$$\eta = \left(7.0274 \times 10^{-3}T^2 + 6.037T - 1368\right)^{-1}$$

An important reaction of elemental bromine is its dissociation:

$$Br_2\,(g) \longrightarrow 2\,Br\,(g) \qquad (1)$$

Some data for this reaction are shown in Tables 3 and 4. The free energy (149 640 J at 298.1 K) is calculated using:

$$\Delta F^\circ = 174170 - 10.886T\ln T + 0.0020934T^2 - 20934T^{-1}$$

Table 3. Equilibrium constants for Equation (1)

T, °C	650	800	900	1000
Log K	−5.983	−3.747	−2.854	−2.114

Table 4. Thermodynamic data for molecular and atomic bromine

	ΔH°, J	ΔG°, J	S°, J/K
Br	111788	82480	175.0
$Br_{2\,(g)}$	30564	2931	245.3
$Br_{2\,(l)}$	0	0	152.4

Solubility. Bromine and chlorine are miscible in all proportions, yielding an equilibrium mixture containing the interhalogen bromine chloride:

$$Br_2 + Cl_2 \rightleftharpoons 2\,BrCl \quad K_{eq} = 2.94 \text{ at } 25\,°C$$

This equilibrium mixture can be separated by distillation [10]. The density of bromine – chlorine mixtures (in g/L) can be calculated using the formula:

$$\varrho_T = 4.0916 - 0.0374w_{Cl} - 0.0033T$$

of chlorine in % and T the temperature in K. The solubility of bromine in water is low, 3.5 g/100 g at 20 °C. The solubility curve for bromine in water is represented in Figure 2.

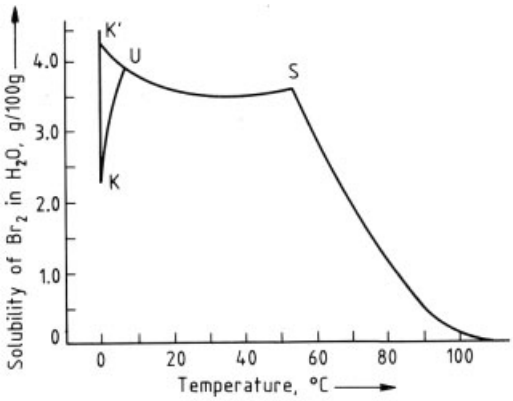

Figure 2. Solubility diagram for $Br_2 - H_2O$ at 101.3 kPa [11]
K′: 4.3 g (−0.58 °C), $Br_{2\,(l)}$ + ice + H_2O; K: 2.27 g (−0.3 °C), Br_2 (hydrate) + ice + H_2O; U: 3.86 g (5.84 °C), Br_2 (hydrate) + $Br_{2\,(l)}$ + H_2O; S: 3.58 g (54.3 °C), *bp*

Bromine forms a red crystalline hydrate, $Br_2 \cdot 8\,H_2O$, which decomposes at 5.8 °C [12]. Bromine hydrate and ice form a eutectic at −6.3 °C. Because of the slow formation of bromine hydrate, bromine in water solution is metastable and forms ice and bromine below −58 °C.

The density of water containing bromine (in g/mL) can be determined using the formula:

$$\varrho_T = 1.0995 + 0.0073w_{Br} - 0.000344T$$

concentration in % and T the temperature in K. Bromine can be recovered quantitatively from a water solution by sparging with air [13]. The solubility of bromine in water can be increased by adding bromide and chloride salts because of polyhalide complex formation (Table 5). Equilibrium constants for the formation of tribromide

Table 5. Solubility of bromine in various aqueous solutions at 25 °C [14], [15]

Solute	c_{solute}, g/L	c_{Br_2}, g/L	Solute	c_{solute}, g/L	c_{Br_2}, g/L
None	–	34.0	NaCl	58.5	55.9
NaBr	92.6	99.2	NaCl ∗	118.0	86.4
NaBr	206.0	248.0	KNO_3	101.2	29.0
NaBr	320.0	546.0	$NaNO_3$	85.1	28.0
KBr	11.9	49.3	K_2SO_4	91.2	24.8
KBr	59.5	119.0	Na_2SO_4	63.6	23.9
KBr	119.0	216.0	H_2SO_4	49.0	29.3
KBr	360.8	632.0	HBr ∗	33.7	108.8
KCl	74.6	57.4	HCl ∗	39.0	79.2

∗ At 20.6 °C

and pentabromide ions at 25 °C have been determined [16]:

$$Br_2 + Br^- \rightleftharpoons Br_3^- \quad K = 16.85$$
$$Br_3^- + Br_2 \rightleftharpoons Br_5^- \quad K = 1.45$$

The relationship between bromine solubility and the concentration of NaBr and KBr salts is shown in Figure 3.

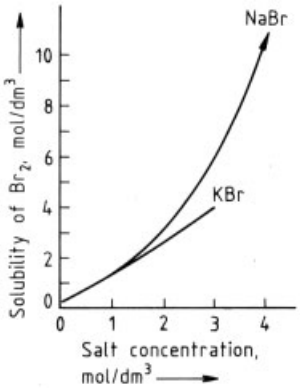

Figure 3. Bromine solubility in salt solutions at 25 °C

The solubility of water in liquid bromine is relatively low (380 ppm at 25 °C, Fig. 4). Commercial bromine usually contains less than 30 ppm water, but bromine has an affinity for moisture and may pick up additional water if improperly handled. For instance, at room temperature, it could reach a water content of 70 ppm if contacted with air at 20 % R.H. This could rise to 300 ppm if the bromine is exposed to air having 80 % R.H., a level that increases considerably bromine's corrosive properties [9].

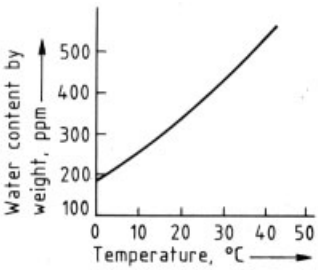

Figure 4. Solubility of water in bromine

Bromine is miscible with many organic solvents, such as carbon tetrachloride, ether, methanol, alkyl bromides, methylene chloride, chloroform, and acetic acid. Reactions can take place with some of these solvents under certain conditions.

Bromine itself is a good solvent for cesium bromide, as shown in Table 6 along with the solubility of other salts in liquid bromine.

Table 6. Solubility of various salts in liquid bromine, g/100 g solution [8]

LiBr	0.005	$SrBr_2$	0.002
NaBr	0.010	$BaBr_2$	0.005
KBr	0.0185	$NiBr_2$	0.0003
RbBr	0.058	$ZnBr_2$	0.0146
CsBr	19.3	$FeBr_3$	0.324
$MgBr_2$	0.002	$AlBr_3$	0.002
$CaBr_2$	0.003		

3. Chemical Properties

Many chemical reactions of bromine are a result of its strong oxidizing properties. The tendency of bromine to gain a single electron lies as expected between chlorine and iodine, as do most other chemical properties. Bromine is only a moderately strong oxidizing agent (Table 7).

Table 7. Standard potentials of bromine and various other oxidizing agents

I_2 (s) / $2\,I^-$	+ 0.53 V
Br_2 (l) / $2\,Br^-$	+ 1.07 V
Cl_2 (g) / $2\,Cl^-$	+ 1.36 V
F_2 (g) / $2\,F^-$	+ 2.85 V
Fe^{3+}/Fe^{2+}	+ 0.77 V
Ce^{4+}/Ce^{3+}	+ 1.44 V
MnO_4/Mn^{2+}	+ 1.52 V
Pb^{4+}/Pb^{2+}	+ 1.8 V

Corresponding to its position in the electromotive series, chlorine oxidizes bromide ion to elemental bromine. This reaction is the chemical basis for the production of bromine (see Chap. 4).

Of the bromine fluoride compounds (BrF, BrF_3, BrF_5), BrF_3 is of greatest practical importance because of its use in fluorinating organic substances.

Bromine undergoes *reduction* to bromide in water. The reducing agent can be either the water itself or the remaining bromine. In the first case, oxygen is formed:

$$2\,Br_2 + 2\,H_2O \longrightarrow 4\,H^+ + 4\,Br^- + O_2 \quad E^\circ_{298} = +0.25\,V$$

The potential for corresponding reactions is + 2.055 V for fluorine, + 0.544 V for chlorine,

and -0.28 V for iodine. Thus, bromine reacts less vigorously with water than fluorine and chlorine.

When bromine itself becomes the reducing agent, bromide and hypobromite are formed:

$$Br_2 + H_2O \longrightarrow H^+ + Br^- + HOBr \quad E^\circ_{298} = +0.525\ V$$

The equilibrium constant for this reaction is 5.8×10^{-9} at 25 °C and varies little with temperature [17]. This reaction can be suppressed by addition of acid. With a small amount of base, the reaction goes to completion. In strong base, the reaction occurs only in the cold. Warming under these conditions converts the hypobromite to bromate and bromide:

$$3\,NaOBr \longrightarrow NaBrO_3 + 2\,NaBr$$

With the addition of acid to the bromate–bromide mixture, bromine is again formed.

Ammonia is oxidized by bromine to nitrogen:

$$3\,Br_2 + 8\,NH_3 \longrightarrow 6\,NH_4Br + N_2 \quad \Delta H = -879.2\,kJ\ [18]$$

Bromine reacts with *hydrogen cyanide* to give cyanogen bromide:

$$Br_2 + HCN \longrightarrow BrCN + HBr$$

Carbonates react similarly to alkali, forming the bromide and bromate salts:

$$3\,Na_2CO_3 + 3\,Br_2 \longrightarrow 5\,NaBr + NaBrO_3 + 3\,CO_2$$

Sulfur dioxide is oxidized to sulfuric acid by bromine:

$$SO_2 + Br_2 + H_2O \longrightarrow 2\,HBr + H_2SO_4$$

The latter two reactions are used commercially in recovering or neutralizing bromine residues.

Bromine reacts with many *metals* to form bromides. Sodium is stable in dry bromine, but sodium vapor reacts vigorously. Potassium and cesium react violently with bromine. Bromine is also highly reactive with aluminum and titanium. Aluminum reacts with emission of light. Magnesium, silver, nickel, and lead become coated with their bromide, which prevents further reaction. This protective coating makes lead a useful material for containers of bromine. Other nonferrous metals, that are useful in handling bromine are nickel, tantalum, Hastelloy C, and monel and certain other copper alloys.

Moisture is important in the corrosion of metals by bromine. This is probably due to the hydrolysis products, hydrobromic and hypobromous acids. At moisture contents below 40 ppm, nickel containers can be used for transporting bromine [19]. Mercury is attacked by bromine and should not be used in instruments and gages exposed to bromine vapor. Dry bromine reacts slowly with iron to form a protective layer of ferric bromide; but when wet, a mixture of hydrated iron bromides is formed which does not adhere to the iron surface [20]. Other heavy metals, such as copper, manganese, chromium, antimony, cobalt, cadmium, and bismuth, react with bromine, although some metals require higher temperature.

Bromine also reacts with plutonium and uranium and offers the possible separation of UBr_4 and $PuBr_3$ by distillation. Titanium and zirconium are said to form H_2TiBr_6 and H_2ZrBr_6 in solution [21].

3.1. Bromine Addition to Alkenes and Alkynes

Bromine adds readily to unsaturated compounds. Such reactions are usually run at low temperature to avoid substitution side reactions. Although a catalyst is usually not required, ultraviolet radiation or high temperature may be used to accelerate the reaction.

Among the more important commercial products from bromine addition to unsaturated compounds are ethylene bromide, acetylene tetrabromide, 2,3-dibromopropanol, and hexabromocyclodecane. Ethylene bromide is the largest volume bromine-containing organic chemical produced (→ Bromine Compounds, Chap. 1.4.1.).

3.2. Bromination of Aromatic Compounds

Aromatic compounds can undergo three types of reactions with bromine: (1) addition, (2) substitution on side chains, and (3) substitution on the aromatic ring.

The light-catalyzed addition of bromine across aromatic double bonds is generally a slow

reaction, but use of chlorine increases the rate. Bromination of aromatic side chains is normally a free-radical process.

Electrophilic substitution on the aromatic ring is by far the most important type of aromatic bromination. In the presence of a catalyst, bromine reacts with aromatic compounds to give aryl bromides and hydrogen bromide:

$$ArH + Br_2 \longrightarrow ArBr + HBr$$

Suitable catalysts for aromatic brominations are Lewis acids, such as halides of aluminum, iron, zinc, or antimony. Activated aromatic compounds, such as phenols, anilines, and aromatic ethers, can be brominated without a catalyst. Bromine is more efficiently used in aromatic substitution reactions in situ generation of it from hydrogen bromide using chlorine:

$$2\,ArH + Br_2 + Cl_2 \longrightarrow 2\,ArBr + 2\,HCl$$

Preformed bromine chloride can also be used for the same purpose. Substitution reactions involving BrCl are much faster than ones using bromine alone [22].

3.3. Free-Radical Bromination

Bromination of saturated hydrocarbons and of alkyl side chains of aromatic compounds occurs by a free-radical chain reaction:

$$Br_2 \longrightarrow 2\,Br\cdot$$
$$RH + Br\cdot \longrightarrow R\cdot + HBr$$
$$R\cdot + Br_2 \longrightarrow RBr + Br\cdot$$

Dissociation of bromine molecules is needed to initiate this reaction. Bromine can be dissociated thermally, photolytically, by gamma radiation, or by use of peroxide initiators. Compared with chlorine, the bromine atom is less reactive and the hydrogen abstraction step is endothermic. As a result, reaction rates are relatively slow and the bromine atom exhibits high position selectivity. For example, in bromination of *n*-butane, secondary hydrogen atoms are replaced 82 times faster than primary ones [23]. Bromination of aromatic side chains is a much faster reaction than bromination of alkanes.

3.4. Production of Hydrogen Bromide

Hydrogen bromide [*10035-10-6*] is an important intermediate in producing various organic and inorganic bromides (→ Bromine Compounds). It is prepared commercially by the direct gas-phase reaction of hydrogen and bromine:

$$H_2 + Br_2 \longrightarrow 2\,HBr + \text{heat}$$

The reaction can be conducted without catalyst by maintaining a self-sustaining flame within an enclosed burner [24].

3.5. Bromination Using Bromine Carriers

Bromine complexes and bromoimides are useful for brominations where selectivity is of prime importance. The bromine carrier that has found the most use in laboratory brominations is *N*-bromosuccinimide (NBS) [*128-08-5*]. A unique application of NBS is in free-radical allylic brominations:

$$CH_2{=}CHCH_3 \xrightarrow{NBS} CH_2{=}CHCH_2Br$$

Other useful bromine carriers are dibromodimethylhydantoin, dioxane dibromide, pyridine hydrobromide dibromide, and various quaternary ammonium polybromides [25].

4. Production

Bromine production utilizes bromine-containing liquors (brines), which occur naturally, or result from the manufacture of potassium salts or from seawater.

In most brines, as well as in ocean water, the molar concentration of chloride ions is likely to be 200 to 700 times as large as that of bromide ions. Therefore, isolating bromine from these sources requires a separation method of high selectivity. Fortunately, such a method exists. It is based on the greater ease with which bromide can be oxidized compared to chloride, and the fact that the oxidation product, bromine, is quite volatile. Chlorine [*7782-50-5*] serves as the most economical and convenient oxidant. The normal

oxidation – reduction potentials ($E°$) applicable to the operation are:

$$1/2\,Br_2 + e^- \longrightarrow Br^- \quad E° = +1.09\,V$$
$$1/2\,Cl_2 + e^- \longrightarrow Cl^- \quad E° = +1.36\,V$$

Two general processes using chlorination for bromine recovery are based on this chemistry:

1) Steaming-out process for brines and waste liquors containing bromine
2) Blowing-out process for seawater.

4.1. Steaming-Out Process

Production of bromine from brines containing Br^- in concentrations of 1 – 5 g/L basically follows the process described by KUBIERSCHKY in 1906 [26].

Raw brine is warmed in a heat exchanger (a) and caused to pass successively through two packed towers (b), (c) by gravity flow (Fig. 5). In the upper (smaller) tower, it meets a recycle stream of gases from which chlorine and bromine are absorbed. Near the bottom of the lower tower (c), chlorine and steam are introduced; as they pass upward, the chlorine reacts with bromide in the brine, and a mixture of bromine vapor and chlorine (about 85:15 by weight) with steam is taken off the top. The water and most of the halogens are condensed (d), the liquid phases enters a gravity separator (e) and the gas goes to the upper tower. From the separator, water containing dissolved halogens is sent to the lower (steaming-out) tower and the heavier bromine layer, containing some chlorine, flows to a fractionating column (g). The chlorine vapors join the stream to the upper tower while the liquid bromine, about 99 % pure, is drawn off either for direct use in the manufacture of bromine compounds or to be further purified for sale or for more exacting uses.

The hot debrominated brine is treated to neutralize acidity (h) and to reduce free halogens, if necessary, and then is passed through the heat exchanger, where it warms the incoming brine.

More recent modifications of the original Kubierschky process are related generally to the larger scale of modern operations, to the materials of construction, which must be corrosion resistant, and to the procedures and instrumentation needed for control. Whereas originally most of the construction was stoneware, brick, or lead, now much of the material exposed to wet halogens is glass, tile, or other special ceramics or al-

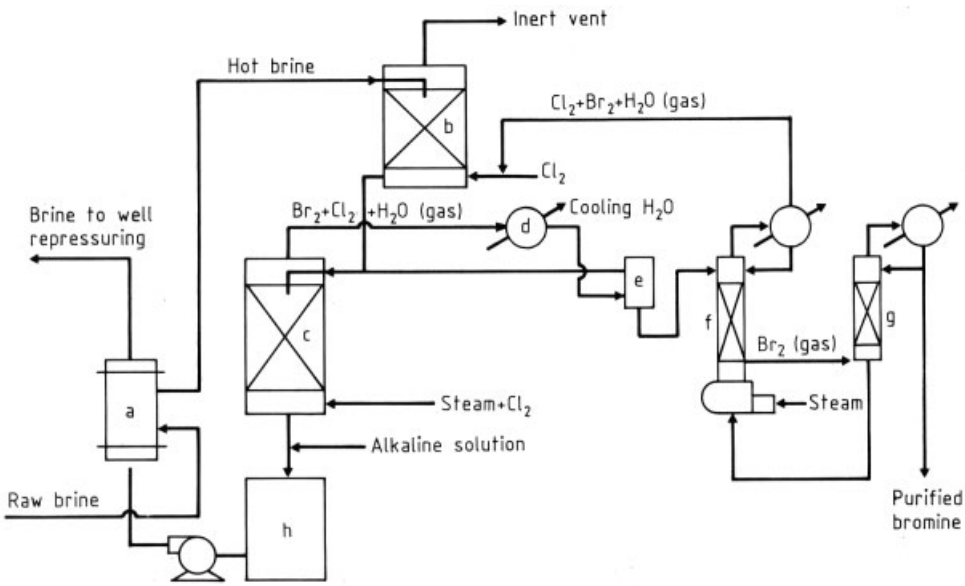

Figure 5. Bromine steaming-out process
a) Bromine cross heat exchanger; b) Chlorine absorber; c) Steaming-out tower; d) Condenser; e) Separator; f) Stripping column; g) Fractionating column; h) Neutralizer

loys; tantalum may be used for condensers [27]. The processes are controlled by measuring pH, oxidation – reduction potentials, flow rates, temperatures, and pressures at critical points.

Some acidity is generated when chlorine reacts with reducing substances, such as hydrogen sulfide, in the brine. This acidity adds to the efficiency of bromine liberation by preventing hydrolysis to hypobromous acid. Use of sulfuric acid is not advisable if the brine contains calcium or strontium because they may precipitate as sulfates and foul the tower packing, heat-exchange surfaces, or brine disposal wells. However, sulfates will form if the brine contains sulfides.

4.2. Seawater Process

The first economically successful recovery of bromine directly from the sea was carried out near Wilmington, North Carolina (USA), using an *air-blowing process* originally developed for brine operations by H. H. Dow [28]. The use of air rather than steam is necessary because the cost of steam to heat ocean water, with its bromine content of only about 65 mg/L, would be prohibitive. The Wilmington plant has since been abandoned, but the process is 1986 still employed, e.g., at Amlwich (Wales) and Hayle (Cornwall).

Water from the ocean is pumped to the top of blowing-out towers, with sulfuric acid and chlorine being added just above the pumps so that mixing takes place during the ascent. About 1.3 kg of 10 % sulfuric acid per ton of water is required to neutralize the natural hydrogen carbonates and bring the pH to 3.5; 15 % excess chlorine over the theoretical requirement is used [29].

Air is drawn up through the towers, sweeping out a mixture of bromine and chlorine (or bromine chloride) from the descending ocean water. The air is drawn next through absorber towers in which it is scrubbed counter-currently by sodium carbonate solution. The several reactions which take place may be summarized approximately by the equation:

$$3\,Na_2CO_3 + 2\,Br_2 + BrCl \longrightarrow NaBrO_3 + 4\,NaBr + NaCl + 3\,CO_2$$

To remove spray from the air, small packed chambers are interposed between the absorber towers and the fans. When the alkalinity of the scrubber solution is nearly depleted, the solution is brought first to a storage tank and then to a reactor in which it is treated with sulfuric acid and steamed to release bromine:

$$NaBrO_3 + 5\,NaBr + 3\,H_2SO_4 \longrightarrow 3\,Br_2 + 3\,Na_2SO_4 + 3\,H_2O$$

The bromine is condensed and used in the manufacture of ethylene dibromide.

In a modification of the process [30], [31], which was used for a number of years at Freeport, Texas, halogens blown from the ocean water are reduced with sulfur dioxide and absorbed in a water spray:

$$Br_2 + SO_2 + 2\,H_2O \longrightarrow 2\,HBr + H_2SO_4$$
$$BrCl + SO_2 + 2\,H_2O \longrightarrow HBr + HCl + H_2SO_4$$

The resulting aqueous solution of mixed acids is treated with chlorine in a conventional steam-out tower and the bromine is recovered. The mixture of hydrochloric and sulfuric acid remaining is used for acidifying the raw ocean water.

The *SO_2 process* has been described [32]; further details of its operation at Freeport and the automatic controls employed there are available [33]. Normal flow of incoming gulf water is 568 m^3/min; the rubber-lined riser pipe is about 12 m tall and 2.1 m in diameter. A glass electrode at the top of the riser measures the pH to within 0.02 unit and controls "butterfly" valves in a return line on the acid supply pumps. Oxidation potential is controlled at about 970 mV with a platinum electrode and a saturated calomel reference electrode, regulating the addition of chlorine. Oxidation potential measurements are also used to control the mixing of sulfur dioxide with the free halogen vapor stream and the chlorination of the reduced acid mixture.

Analytical methods used for bromide in brines and ocean water have been described [34]. Usually the van der Meulen method [35] or one of its various modifications is employed. *Economic factors* important in the location of seawater bromine plants, especially as they were related to the siting of the Amlwich operation, have been discussed [36]; temperature of the water has a significant effect on the efficiency of the blowing-out process [32].

Because of the large volume of ocean water processed, the bromine-depleted stream must

be discharged at some distance from the intake and in a favorable location from the standpoint of prevailing currents, to avoid dilution. Aside from this, *waste disposal* is not a problem since both the acidity and the free halogen content of the outgoing water are very low. In contrast, the waste stream from a strong brine process requires neutralization and, in some instances, reduction. In most cases, spent well brine is returned to the ground via disposal wells to maintain hydrostatic pressure within the formation. Acidity or free halogens should be absent to minimize corrosion of the system. Lime or caustic soda are customarily used for neutralizing acidity. The use of semiburnt dolomite with sodium thiosulfate has been recommended [37].

4.3. Electrolysis

Electrolysis was used in early production in Germany instead of chlorine to oxidize bromides. Two processes were developed in Germany. The *Wunsche electrochemical process* was based on the use of diaphragms to provide good current efficiency. The carbon cathodes were separated from the carbon anodes by porous clay cylinders, which allowed ion diffusion. The anode reaction product was bromine:

$$2\,Br^- \longrightarrow Br_2 + 2\,e^-$$

The cathode reaction produced hydrogen and magnesium hydroxide, which created a major problem of clogging the clay diaphragm pores:

$$Mg^{2+} + 2\,e^- + 2\,H_2O \rightarrow Mg(OH)_2 + H_2$$

The *Kossuth electrochemical process* used bipolar carbon plate electrodes with one side serving as anode and the other side as cathode. The power supply was connected to the end electrodes, and to dislodge the magnesium hydroxide deposits that built up on the cathodes, the current was periodically reversed for a short time. Despite this advantage over the Wunsche process, the Kossuth process had very low-current efficiency. The amount of power required related to bromine was 3000 kW h/t as compared to about 2000 kW h required to produce 500 – 600 kg of liquefied chlorine used to make 1 t of bromine by the Kubierschky process (Section 4.1.), which was being developed at the same time.

Another process for the recovery of bromine from seawater was developed on a commercial scale by Ethyl Corp. in the United States. The process involved the treatment of seawater with chlorine and aniline, and collection of the precipitated tribromoaniline. The company later used a 4200-t ship as a floating factory, which was unsuccessful.

Current bromine production methods are based on modified Kubierschky or steaming-out processes and Dow Chemical's blowing-out process.

4.4. Bromine from Recycle and Waste Brines

In the United States, a modified Kubierschky process has been used for recovering bromine as a byproduct from potassium chloride production liquors at Searles Lake [38]. The Searles Lake brine contains ca. 0.085 % bromine, but it is too alkaline and too heavily buffered for direct use in the conventional chlorine process of liberating bromine. However, since potassium bromide and potassium chloride are isomorphous, the bromide crystallizes with the potassium chloride. This mixture, containing up to 1.8 % bromide, is neutral and therefore useable. When chlorinating this mixture, the bromine is expelled and recovered while an equivalent amount of new potassium chloride replaces the bromide and thus effects the desired purification of the chloride product.

For economic reasons, increasing attention is being given to the development of methods for recovering bromine from byproduct bromides, such as those obtained in halogen displacement reactions. Many substitution reactions produce a kilogram of bromine in the form of byproduct hydrogen bromide for each kilogram of bromine substituted in the desired product. The hydrogen bromide, which is usually released as a gas, is absorbed in water or in an alkaline solution using a scrubber. The aqueous scrubber effluent is acidified and chlorinated in order to recover the bromine for recycle. In some cases, the hydrogen bromide is recovered and used as is or as the bromide salt in other applications.

4.5. Purification

Bromine produced by either the steaming or blowing processes ordinarily contains as impurities chlorine, water, halogenated hydrocarbons, and acids (HCl or HBr). Fractional distillation brings the purity to 99.5 – 99.7 %, long associated with commercial bromine. Improvement to the 99.9+ quality available at present may be done by proprietary methods, of which the following may be considered representative.

Maintaining bromine at 50 °C for 4 h with provision for removal of the vapor lowers the chlorine content from 0.12 to 0.06 % and moisture from 0.05 to 0.005 % [39]. Passing predried, vaporized bromine and oxygen through a tube packed with quartz or Vycor glass fragments and heating to 1000 °C decomposes organic halogen compounds, such as chloroform [40]. Heating bromine vapor to 370 – 540 °C decomposes organic impurities to HCl and to higher boiling compounds which can be removed by distillation [41]. A mixture of steam and bromine can be heated to similar temperatures. Upon cooling, the impurities remain largely in the aqueous phase [42]. Absorption of bromine by a bromide solution with subsequent steam stripping provides highly purified bromine with essentially no chlorine [43].

To avoid corrosion of lead, nickel, or monel containers, less than 30 ppm water should be present in bromine. Drying with concentrated sulfuric acid removes most of the moisture. Organic impurities can be removed by distilling bromine at 58 °C with 96 % sulfuric acid, ca. 2 L/1000 kg, washing the condensed bromine with water, and drying countercurrently with 98 % acid [44]. In addition drying can be done with molecular sieves, type 3 A – 5 A (0.3 – 0.5 nm) [45]. The laboratory preparation of extremely pure bromine for physical properties determination has been described [46].

5. Quality Specifications and Testing

Typical industrial product specifications for purified bromine are [47]:

Bromine, min.	99.98 %
Chlorine, max.	100 ppm
Water, max.	30 ppm
Nonvolatile residue, max.	30 ppm
Organics, total identifiable, max.	100 ppm

Bromine and bromine compounds can be analyzed by the following methods [48]. Liquid bromine and bromine solutions can be assayed by standard iodometric titration [49]. Bromine vapor can be analyzed by absorption in aqueous potassium iodide solution and the liberated iodine titrated iodometrically. A modification of this procedure uses a spectrophotometer to measure the iodine concentration by determining the absorption of the triiodide ion formed by addition of excess iodide [50].

A method for the separation of iodide, bromide, and chloride ions in any mixture involves selective oxidation and distillation, followed by potentiometric or spectrophotometric determination. Other methods for analyzing mixtures of halide ions include chromatography and precipitation.

Various procedures are used for the analysis of impurities in bromine. For example, chlorine content (η) in bromine can be calculated (in wt %) by accurately measuring its density (ϱ) and then using the formula:

$$\eta = 109.4 - 0.088234T - 26.738\varrho$$

A more accurate and convenient method for chlorine determination is ion chromatography [51]. Pure bromine is transparent in the infrared, so water and organic impurities can be measured by their IR absorption spectrum. The residue after evaporation of bromine is determined by heating a sample in a tared container on a steam bath, drying at 105 °C, and weighing.

The bromine content in organic compounds can be determined by neutron activation and X-ray fluorescence.

6. Storage and Transportation

Bromine is usually shipped in monel alloy drums constructed of 14-gage metal and designed to contain 38 L or 102 kg of liquid bromine. Larger bulk shipments of bromine use lead- or nickel-lined tank cars or tank trucks. The latter must be filled to a minimum 98 % of capacity before shipping to avoid inertia effects in transport.

The relatively high freezing point of bromine (−7.25 °C) may cause some problems in storage and shipping. If there is evidence that the bromine is frozen, although the cars are

usually insulated, warm water (below 54 °C) should be circulated through the heating coils. Steam should never be used, as bromine boils at 58.6 °C.

Storage tanks should be protected from moist air. Bromine that contains 30 ppm of water will pick up moisture from air having a dew point of −70 °C. Bromine in equilibrium with air having a normal summer dew point of 5 – 10 °C contains about 70 ppm of water. Corrosion of many bromine-handling metals is severe at this concentration of water. To keep out moisture, a very slight positive pressure (25 Pa) should be kept on bromine storage tanks by purging dry nitrogen or dry air into the tank vent line just above its connection to the tank. When pressurization of the tank is necessary for unloading, dry air should be used, and no more pressure applied than is required to lift the bromine to the highest point in the unloading system (7 kPa gage pressure lifts bromine 0.23 m).

For transferring bromine, plastic-lined steel is a versatile material that combines chemical resistance with reinforced strength. Fluorinated plastics, such as Kynar, Teflon, Kel-F, and Viton, find widespread use in equipment, valves, piping, and gaskets. Nickel, monel alloys, and Hastelloy C alloys may be used with dry bromine (less than 30 ppm water) but are severely attacked by wet bromine. Tantalum may be used with either wet or dry bromine up to 150 °C. Steel and stainless steel usually are unsatisfactory. However, the use of steel is possible when nitric oxide are present to prevent corrosion. Chemical lead- and glass-lined steel are probably the most versatile and practical materials for general use in storing and handling bromine. For additional information, see industrial product brochures on bromine handling and shipping [9].

7. Uses

The primary use for elemental bromine is in the manufacture of bromine compounds that have chemical and biological activity, high density, or fire retarding and extinguishing properties. Bromine products are well represented in such use areas as gasoline additives, flame retardants, agricultural chemicals, drilling fluids, photographic chemicals, sanitizers, dyes, pharmaceuticals, and others.

The largest use of bromine is as ethylene dibromide (EDB), a lead scavenger additive used in gasoline with tetraalkyllead compounds to reduce engine knocking. In the past 10 years, the EDB share of the United States bromine market dropped from 70 % to under 30 %, which can be attributed partly to government regulations issued in 1973, gradually reducing the lead content in gasoline. The use of EDB as a fumigant in agriculture is also on the decline for environmental reasons.

The high-density characteristics of bromine compounds are advantageously applied in hydraulic fluids, gage fluids, ore flotation, and drilling fluids. Calcium bromide, 53 % in aqueous solution, was introduced in 1972 as a dense fluid product (1.70 kg/L) for use as oil well packs and completion fluids [52]. The use of calcium bromide, zinc bromide, and sodium bromide as completion fluids in oil and gas well drilling grew rapidly in the 1970s, capturing 22 % of the elemental bromine market [53].

Another of the fastest growing uses for bromine is in the manufacture of flame retardants (→ Flame Retardants). Bromine fire retardants are benefiting from increased use of engineering plastics and hold an estimated 16 % of the elemental bromine market. Some of the major bromine products used as flame retardants are tetrabromobisphenol A, decabromodiphenyl oxide, and vinyl bromide.

Bromine is also used, either directly or indirectly, in sanitary preparations, water disinfection, in the desizing of cotton, in bleaching of pulp and paper, air conditioning absorption fluids, and hair waving compositions.

8. Economic Aspects

World bromine production for 1993 was reported to be about 3.73×10^5 t. The distribution for each major producing country is shown in Table 8.

Table 8. World bromine production

Country	$\times 10^3$ t			
	1982	1983 *	1984	1993
United States	181.9	172.4	153.0	163.0
Israel	69.9	69.9	99.0	135.0
Former States of USSR	68.0	68.0	68.0	19.0
United Kingdom	27.0	22.7	27.0	–
France	15.9	13.6	9.0	15.0
Japan	12.0	11.8	11.7	–
Federal Republic of Germany	4.0	3.6	3.2	–
Italy	0.6	0.5	0.4	–
Spain	0.4	0.4	0.4	–
India	0.4	0.4	0.4	–
Other countries	0.9	0.9	0.9	–
Total	381	364.2	373	366

* Production was down due to economic recession

9. Toxicology and Occupational Health

In the vapor state, bromine is similar to chlorine in physiological reactions. As a result of its oxidizing action, comparatively low concentrations of bromine vapor are highly irritating to the entire respiratory tract, the mucous membranes, and the eyes, producing responses, such as coughing and smarting of the eyes. The approximate exposure limits for bromine are given in Table 9.

Table 9. Exposure limits for Br_2 vapors [54–56]

	Vapor concentration, ppm
Maximum allowable concentration that is tolerated without disturbance	
prolonged exposure (several h)	0.1 – 1
short exposure (1/2 – 1 h)	4
Least detectable odor	≈3.5
Least amount causing irritation to the throat	15
Dangerous to life (short exposure)	40 – 60
Rapidly fatal on very short exposure (< 30 min)	1000
TLV: TWA	0.1 (0.7 mg/m^3)
STEL	0.3 (2 mg/m^3)
MAK	0.1 (0.7 mg/m^3)

Exposure to high concentrations of Br_2 vapor leads to inflammation and congestion of the respiratory system and edema of the lungs, which in severe cases can be fatal. The sharp odor of bromine gives warning against acutely dangerous exposure but not against concentrations which may be harmful as a result of prolonged exposure.

Liquid bromine presents a serious hazard in handling. It rapidly attacks the skin and other tissues to produce irritation and necrosis; the burns are not only painful but slow to heal. Liquid bromine can also cause severe eye damage.

For the average person, the ratio of bromine to chlorine is roughly about 1:100 in both blood and urine. Table salt has been implicated as the major source of dietary bromine, as it may contain as much as 0.1 % (1000 ppm) bromide [57]. The World Health Organization in 1967 stated that the maximum acceptable daily intake of bromide ion is to be 1.0 mg/kg.

Acute bromide intoxication is rare because amounts sufficient to elevate the serum bromide to a toxic level can hardly be ingested and retained without vomiting. Chronic bromide intoxication from the continued use of large doses (> 3 – 5 g/d) is known as bromism. The symptoms are skin rash, glandular excretions, gastrointestinal disturbances, and neurological disturbances. Bromide compounds have been used in the treatment of neurological disturbances as a sedative and in the control of epilepsy at doses of 3 – 5 g/d [58].

Dilute aqueous solutions of bromine exert considerable germicidal action; comparable studies suggest that it is commonly between chlorine and iodine in its activity against bacterial spores, although with some species, e.g., yeast and certain algae, it appears to be more toxic than chlorine. During disinfection, when ammonia compounds are present, the bromine (bromamine) species are much more toxic to

microorganisms than are chlorine (chloramine) species [59]. Also, the lower stability of bromamine residuals lowers their potential hazard to marine life by discharged effluents. In specific cases, the toxic action of bromine may well depend upon oxidation of enzyme functional groups, such as −SH and −SS−, or of the protein coat of the organism.

Safety Precautions. Safe handling of bromine demands the following measures: (1) rigorous avoidance of all contact of bromine with skin, eyes, and clothing, and (2) adequate ventilation with all possible precautions against inhaling the vapors when significant quantities of bromine are to be handled. Good practice requires that goggles, face shield, gloves, and plastic or rubber suit be worn and that suitable gas masks be available. Preventative health measures and first-aid procedures are given elsewhere [9].

In the case of a bromine spill, all unprotected personnel should be evacuated from the area contaminated either by liquid or by vapor releases. Containment of the spill should be given first priority. Liquid bromine can be contained using earth, sand, or inert material. Vapors can be contained by covering the spill with water. Bromine can then be neutralized or decontaminated with potassium carbonate solution (10 – 50 %), sodium carbonate, hydrogencarbonate, or sodium thiosulfate (hypo). Anhydrous ammonia (in cylinders) is useful for neutralizing any bromine fumes. It is also useful in locating bromine leaks.

10. References

1. A. J. Balard, *Ann. Chim. Phys.* **2** (1826) no. 32, 337.
2. C. Löwig: *Das Brom und seine chemischen Verhältnisse,* Heidelberg 1829.
3. P. J. Scheuer: *Chemistry of Marine Nature Products,* Academic Press, New York 1973.
4. D. J. Faulkner, *Tetrahedron* **33** (1977) 1421 – 1443.
5. P. Bovis, *J. Phys. Radium* **6** (1929) no. 10, 267.
6. A. J. Downs, C. J. Adams in A. F. Trotman-Dickenson (ed.): *Comprehensive Inorganic Chemistry,* Pergamon Press, New York 1973, p. 1107.
7. F. Yaron in Z. E. Jolles (ed).: *Bromine and its Compounds,* Ernest Benn Ltd., London 1966, pp. 43 – 49.
8. V. A. Stenger, *Angew. Chem. Int. Ed. Engl.* **5** (1966) no. 3, 280.
9. Dow Chemical, Bromine – Unloading, Storing, Handling, Form No. 101-2-80, 1980.
10. B. J. Karsten, *Z. Anorg. Allg. Chem.* **53** (1907) 379.
11. J. d'Ans, P. Hofer, *Angew. Chem.* **47** (1934) 71.
12. M. V. Stackelberg, *Naturwissenschaften* **36** (1949) 327.
13. M. Berthelot, *Ann. Chim. Phys.* **6** (1886) no. 7, 413.
14. A. Seidell in W. F. Linke (ed.): *Solubilities of Inorganic and Organic Compounds,* 3rd ed., **vol. 1,** Van Nostrand, Princeton, N.J., 1940, pp. 207 – 209.
15. A. Seidell in W. F. Linke (ed.): *Solubilities of Inorganic and Organic Compounds,* 4th ed., **vol. 1,** Van Nostrand, Princeton, N.J., 1958, pp. 440 – 444.
16. A. I. Popov in V. Gutman (ed.): *Halogen Chemistry,* **vol. 1,** Academic Press, New York 1967, p. 225.
17. H. A. Liebhafsky, *J. Amer. Chem. Soc.* **56** (1934) 1500; **61** (1939) 3513.
18. W. Moldenhauer, M. Burger, *Ber. Dtsch. Chem. Ges. B* **62** (1929) 1615.
19. D. E. Lake, A. A. Gunkler, *Chem. Eng. (New York)* **67** (1960) 136.
20. N. W. Gregory, B. A. Thackery, *J. Am. Chem. Soc.* **72** (1950) 3176.
21. J. R. M. Fernandez, A. B. Duran, *An. R. Soc. Esp. Fis. Quim. B* **55** (1959) 823; *Chem. Abstr.* **54** (1960) 11800.
22. J. F. Mills, J. A. Schneider, *Ind. Eng. Chem. Prod. Res. Dev.* **12** (1973) no. 3, 160 – 165.
23. P. C. Anson, P. S. Fredericks, J. M. Tedder, *J. Chem. Soc.,* Part 1, 1959 918.
24. Dow Chemical, US 2070363, 1937 (J. C. Kur).
25. J. J. McKettas (ed.): *Encyclopedia of Chemical Processing and Design,* vol. 5, Marcel Dekker Inc., New York 1977.
26. K. Kubierschky, DE 194657, 1906; *Z. Chem. Apparatenkd.* **3** (1908) 212; *Chem. Zentr.* **1** (1908) 2115.
27. F. Yaron in Z. E. Jolles (ed).: *Bromine and its Compounds,* Ernest Benn Ltd., London 1966, pp. 3 – 42.
28. Dow Chemical, US 460370 (1891).
29. L. C. Stewart, *Ind. Eng. Chem.* **26** (1934) 361.
30. Dow Chemical, US 2143223, 1939.
31. Dow Chemical, US 2143224, 1939.

32. C. M. Shigley, *J. Metals* **3** (January, 1951) 25.
33. P. Hart, *Bull. Agric. Mech. Coll. Tex.* **5 (3)** (1947) no. 2, 59; *Instruments* **20** (1947) 956.
34. J. Haslam, G. Moses et al., *Analyst. (London)* **75** (1950) 343, 352, 357, 371, 383.
35. J. H. van der Meulen, *Chem. Weekbl.* **28** (1931) 238.
36. H. Fossett, *Chem. Ind. (London)* 1971 no. 41, 1161.
37. H. Kloth, B. Peter, *Bergakademie* **22** (1970) no. 10, 628.
38. G. R. Robertson, *Ind. Eng. Chem.* **34** (1942) 133.
39. VEB Kombinat Kali, US 3615265, 1971 (R. Gartner).
40. U.S. Dept. of Army, US 2929686, 1960 (M. Codell, G. Norwitz).
41. Michigan Chemical, US 3314762, 1967 (L. H. Hahn).
42. Michigan Chemical, US 3642447, 1972 (L. H. Hahn, C. W. Dunbar).
43. Great Lakes Chemical, US 3145084, 1964 (L. R. Belohlav).
44. Kali u. Salz GmbH, DE 2034437, 1972 (R. Sell, O. Braun).
45. Kali u. Salz GmbH, DE 2034438, 1972 (R. Sell, G. Budan).
46. M. S. Chao, V. A. Stenger, *Talanta* **11** (1964) 271.
47. Dow Chemical, Bromine Sales Specifications, 1983.
48. Reference 7, P. G. W. Scott, M. L. Parker: "The Analytical Chemistry of Bromine and Its Compounds," pp. 739 – 801.
49. D. A. Skogg, D. M. West: *Fundamentals of Analytical Chemistry,* 4th ed., Saunders College Publishers, New York 1982, pp. 386 – 389.
50. J. J. Custer, S. Natelson, *Anal. Chem.* **21** (August, 1949) no. 8, 1005 – 1009.
51. P. F. Riegler, N. J. Smith, V. T. Torkelson, *Anal. Chem.* **54** (January, 1982) no. 1, 84 – 87.
52. Dow Chemical, Well Pack Fluid-14, Form No.101-221-83, 1983.
53. *Chemical Marketing Research* 1983 (June 6) 5 – 35.
54. M. B. Jacobs: *The Analytical Toxicology of Industrial Inorganic Poisons,* Interscience, New York 1967, pp. 655 – 657.
55. ACGIH (ed.): *Threshold Limit Values (TLV) 1985,* ACGIH, Cincinnati, Ohio, 1985.
56. DFG (ed.): Maximale Arbeitsplatzkonzentrationen *(MAK) 1982,* Verlag Chemie, Weinheim 1984.
57. C. L. Comar, F. Bronner: *Mineral Metabolism – An Advanced Treatise,* **vol. 2,** Academic Press, New York 1964, Part A, pp. 12 – 14.
58. F. A. Patty (ed.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 B,** J. Wiley & Sons, New York 1981, pp. 2965 – 2972.
59. J. F. Mills in J. D. Johnson (ed.): *Disinfection of Water and Wastewater,* Ann Arbor Science Publishers Inc., Ann Arbor, Mich., 1975, Chapter 6, pp. 113 – 144.

Bromine Compounds

Michael J. Dagani, Ethyl Corp., Baton Rouge, Louisiana 70821, United States (Sections 1 – 1.4.10, 1.5)

Henry J. Barda, Ethyl Corp., Baton Rouge, Louisiana 70821, United States (Section 1.4.11 – 1.4.11.4)

Theodore J. Benya, Ethyl Corp., Baton Rouge, Louisiana 70821, United States (Section 1.6)

David C. Sanders, Great Lakes Chemical Corp., West Lafayette, Indiana 47906, United States (Chap. 2)

1. Organic Bromine Compounds . 1
1.1. Physical Properties 2
1.2. Chemical Properties 2
1.2.1. Nucleophilic Displacement of Bromine 2
1.2.2. Displacement of Bromine by Metals 2
1.3. Production 4
1.3.1. Addition Reactions 4
1.3.2. Substitution Reactions 4
1.4. Commercial Products 5
1.4.1. 1,2-Dibromoethane 5
1.4.2. 1-Bromododecane 6
1.4.3. 1-Bromo-3-chloropropane 7
1.4.4. 1-Bromo-2-phenylethane 7
1.4.5. 1-(Bromomethyl)-3-phenoxybenzene 8
1.4.6. Bromoacetic Acid 8
1.4.7. Bromomethane 9
1.4.8. 3-Bromo-1-propene 10
1.4.9. Bromochloromethane and Dibromomethane 10
1.4.10. Trifluorobromomethane and Difluorobromochloromethane . . . 11
1.4.11. Flame Retardants 11
1.4.11.1. Tetrabromobisphenol-A 12
1.4.11.2. Decabromobiphenyl Oxide 14
1.4.11.3. Tetrabromophthalic Anhydride . . 15
1.4.11.4. Hexabromocyclododecane 15
1.5. Manufacturers 16
1.6. Toxicology and Occupational Health 16
2. Inorganic Bromine Compounds 19
2.1. Hydrogen Bromide and Bromides 20
2.1.1. Hydrogen Bromide 20
2.1.2. Bromides 20
2.2. Hypobromous Acid, Hypobromites 23
2.3. Bromous Acid, Bromites 23
2.4. Bromic Acid, Bromates 24
3. References 24

1. Organic Bromine Compounds

In almost all of its properties, both physical and chemical, bromine is intermediate between the other active halogens, chlorine and iodine (→ Bromine). These same intermediate characteristics are also found in organic compounds in which halogen is covalently bonded to carbon (→ Chlorine, → Chlorinated Hydrocarbons, → Iodine and Iodine Compounds). In general the organic compounds of bromine tend to be less stable and more reactive than the corresponding chlorine compounds, but more stable and less reactive than their iodine counterparts. Because of their unique physical, chemical, and biological properties, organic bromides have found wide use in numerous commercial products and industrial applications.

Annual production of organic bromine compounds is also intermediate between that of the organic compounds of chlorine and iodine, mainly because of the cost of the elemental halogens. On a mass basis the relative costs of chlorine, bromine, and iodine are about 1 : 4 : 78, but on a molar basis these relative costs rise to about 1 : 9 : 280, respectively. In spite of apparent economic disadvantages, the unique properties and greater reactivity of organic bromine compounds compared to their chlorine analogs are sufficient in many cases to justify the higher cost of bromine. Furthermore, recycle of byproduct inorganic bromides to bromine can almost eliminate the premium cost of bromine over chlorine.

In terms of bromine consumption, organic compounds accounted for about 75 – 80 % of elemental bromine production, and inorganic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_405

compounds accounted for the remainder in the early 1980s. Ethylene dibromide, largely because of its use as an additive for leaded gasoline, is still the largest volume organic bromine compound produced. Organic bromides are used in many other applications, such as flame retardants, chemicals for agriculture and water treatment, pharmaceuticals, disinfectants, flame extinguishing agents, and dyes. Bromine and its organic compounds are also widely employed as chemical intermediates for the production of other commercial organic compounds.

The scope of this article is limited to those organic bromides having industrial significance. General reference works should be consulted for the numerous other small-volume applications in industry and in basic and applied chemistry [1–6].

1.1. Physical Properties

Except for the gaseous lower aliphatic bromides, organic bromides are either liquids or solids at ambient temperature and pressure. In general the boiling points fall between those of the corresponding chlorine and iodine compounds, as do the melting points, relative densities, and refractive indexes. Simple organic compounds containing bromine are generally soluble in most common aliphatic and aromatic solvents, such that some bromine compounds are classified as solvents. On the other hand their water solubility is usually low and is roughly inversely proportional to their relative molecular mass [7].

Because many organic bromine compounds have been known for years, their physical properties have been documented [8]. Their spectroscopic and electronic properties, in conjunction with their chemical analysis, are discussed in general texts [9], [10]. Representative lists of aliphatic and aromatic bromine compounds and their physical properties are given in Tables 1 and 2, respectively.

1.2. Chemical Properties

The chemistry of organic bromine compounds is so extensive and varied that only a few of the reactions used in the chemical and related industries are briefly discussed. In general, organic bromides undergo these displacement and other chemical reactions with greater facility and usually in higher yield than the less expensive organic chlorides. This comparative advantage, coupled with their more efficient preparation, accounts for much of their industrial importance.

1.2.1. Nucleophilic Displacement of Bromine

Bromine bonded to tetravalent carbon is a good leaving group and is readily displaced by neutral or anionic nucleophiles to form a wide variety of other useful organic derivatives. Aliphatic bromides undergo these substitution reactions by either unimolecular (S_N1) or bimolecular (S_N2) reaction mechanisms that may be accompanied by competing elimination reactions (E 1 or E 2). The rates of substitution or elimination are influenced by a number of factors, such as the substrate structure, the presence of neighboring groups, the nature of the solvent, and the concentration of reactants, to cite a few. Appropriate reaction conditions can usually be chosen so that either substitution or elimination reactions will predominate:

$$RCH_2CH_2Br + Nuc^- \begin{cases} \rightarrow RCH_2CH_2Nuc + Br^- \ (S_N1 \text{ or } S_N2) \\ \rightarrow RCH{=}CH_2 + NucH + Br^- \ (E1 \text{ or } E2) \end{cases}$$

where Nuc^- is any nucleophile. In general, aryl and vinyl bromides undergo these reactions only with difficulty, although there are exceptions. Some examples of the many nucleophiles used to displace bromine in commercial processes are water, secondary and tertiary amines, cyanide and phenoxide ions, the carbanions of active methylene compounds, and trisubstituted phosphines and phosphites.

1.2.2. Displacement of Bromine by Metals

Both aliphatic and aromatic bromides can react with magnesium to form alkyl- and arylmagnesium bromides. These *Grignard reagents* are very useful reaction intermediates because they are capable of reacting with a wide variety of organic functional groups under mild reaction conditions. Organic bromides can also react with

Table 1. Physical properties of aliphatic bromine compounds

Compound	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (101.3 kPa)
Bromomethanes					
Bromomethane	[*74-83-9*]	CH_3Br	95.0	−93.6	3.6
Dibromomethane	[*74-95-3*]	CH_2Br_2	173.8	−52.6	97.0
Bromochloromethane	[*74-97-5*]	CH_2BrCl	129.4	−86.5	68.1
Tribromomethane	[*75-25-2*]	$CHBr_3$	252.8	8.3	149.5
Tetrabromomethane	[*558-13-4*]	CBr_4	331.6	90.5	189.5
Trichlorobromomethane	[*75-62-7*]	CCl_3Br	198.3	−5.6	104.7
Trifluorobromomethane	[*75-63-8*]	CF_3Br	148.9	−168.0	−57.8
Difluorobromochloromethane	[*353-59-3*]	CF_2BrCl	165.4	−159.5	−3.3
Bromoalkenes					
Bromoethylene	[*593-60-2*]	$CH_2{=}CHBr$	107.0	−137.8	15.8
1,1-Dibromoethene	[*593-92-0*]	$CH_2{=}CBr_2$	185.9	–	91.7
cis-1,2-Dibromoethene	[*540-49-8*]	$BrCH{=}CHBr$	185.9	−53.0	112.5
3-Bromo-1-propene	[*106-95-6*]	$CH_2{=}CHCH_2Br$	121.0	−119.4	71.3
2,3-Dibromo-1-propene	[*513-31-5*]	$CH_2{=}CBrCH_2Br$	199.9	–	141.5
trans-1,4-Dibromo-2-butene	[*821-06-7*]	$BrCH_2CH{=}CHCH_2Br$	213.9	53.5	205.0
Tribromoethene	[*598-16-3*]	$BrCH{=}CBr_2$	264.8	−52.9	164.0
Tetrabromoethene	[*79-28-7*]	$Br_2C{=}CBr_2$	343.7	56.5	226.0
Bromoalkanes					
Bromoethane	[*74-96-4*]	CH_3CH_2Br	109.0	−118.6	38.4
1,1-Dibromoethane	[*557-91-5*]	CH_3CHBr_2	187.9	−63.0	108.0
1,2-Dibromoethane	[*106-93-4*]	$BrCH_2CH_2Br$	187.9	9.8	131.4
1-Bromopropane	[*106-94-5*]	$CH_3CH_2CH_2Br$	123.0	−110.0	71.0
2-Bromopropane	[*75-26-3*]	$CH_3CHBrCH_3$	123.0	−89.0	59.4
1,2-Dibromopropane	[*78-75-1*]	$CH_3CHBrCH_2Br$	201.9	−55.2	140.0
1,3-Dibromopropane	[*109-64-8*]	$BrCH_2CH_2CH_2Br$	201.9	−34.2	167.3
1-Bromobutane	[*109-65-9*]	$CH_3(CH_2)_3Br$	137.0	−112.4	101.6
2-Bromobutane	[*78-76-2*]	$CH_3CH_2CHBrCH_3$	137.0	−111.9	91.2
1-Bromopentane	[*110-53-2*]	$CH_3(CH_2)_4Br$	151.0	−87.9	129.6
1-Bromohexane	[*111-25-1*]	$CH_3(CH_2)_5Br$	165.1	−84.7	155.3
1-Bromooctane	[*111-83-1*]	$CH_3(CH_2)_7Br$	193.1	−55.0	200.8
1-Bromodecane	[*112-29-8*]	$CH_3(CH_2)_9Br$	221.2	−29.2	240.6
1-Bromododecane	[*143-15-7*]	$CH_3(CH_2)_{11}Br$	249.2	−9.5	275.9
1-Bromotetradecane	[*112-71-0*]	$CH_3(CH_2)_{13}Br$	277.3	5.6	307
1-Bromohexadecane	[*112-82-3*]	$CH_3(CH_2)_{15}Br$	305.4	17–19	336
1-Bromooctadecane	[*112-89-0*]	$CH_3(CH_2)_{17}Br$	333.4	28.2	362

Table 2. Physical properties of aromatic bromine compounds

Compound	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (101.3 kPa)
Bromobenzene	[*108-86-1*]	C_6H_5Br	157.0	−30.8	156.0
1,2-Dibromobenzene	[*583-53-9*]	$C_6H_4Br_2$	235.9	7.1	225
1,3-Dibromobenzene	[*108-36-1*]	$C_6H_4Br_2$	235.9	−7	218
1,4-Dibromobenzene	[*106-37-6*]	$C_6H_4Br_2$	235.9	87.3	218.5
1-Bromo-4-chlorobenzene	[*106-39-8*]	ClC_6H_4Br	191.5	68.0	196.0
1-Bromo-4-fluorobenzene	[*460-00-4*]	FC_6H_4Br	175.0	−8	152
1-Bromo-4-iodobenzene	[*589-87-7*]	IC_6H_4Br	282.9	92.0	252
4-Bromophenol	[*106-41-2*]	BrC_6H_4OH	173.0	66.4	238
2-Bromotoluene	[*95-46-5*]	$BrC_6H_4CH_3$	171.0	−27.7	181.7
4-Bromotoluene	[*106-38-7*]	$BrC_6H_4CH_3$	171.0	28.5	184.4
4-Bromoanisole	[*104-92-7*]	$BrC_6H_4OCH_3$	187.0	9–10	223
3,5-Dibromosalicylic acid	[*3147-55-5*]	$Br_2C_6H_2(OH)CO_2H$	295.9	223	–
1,3,5-Tribromobenzene	[*626-39-1*]	$C_6H_3Br_3$	314.8	121–124	271
2,4,6-Tribromoaniline	[*147-82-0*]	$Br_3C_6H_2NH_2$	329.8	122	300
2,4,6-Tribromophenol	[*118-79-6*]	$Br_3C_6H_2OH$	330.8	95–96	282–290
4,4′-Dibromo-1,1′-biphenyl	[*92-86-4*]	$BrC_6H_4-C_6H_4Br$	312.0	164	355–360
1-Bromonaphthalene	[*90-11-9*]	$C_{10}H_7Br$	207.1	−6.2	281.0

lithium to produce the more reactive but more expensive organolithium compounds, which are more convenient to use than Grignard reagents in some cases. Because of economic considerations these organometallic compounds are predominantly used in the synthesis of high-value products, such as pharmaceuticals or perfumes.

1.3. Production

Organic bromine compounds can be produced by a number of different chemical reactions [1–6]; however, addition and substitution reactions are the methods most commonly employed in industrial processes.

1.3.1. Addition Reactions

The uncatalyzed addition of bromine to olefins is rapid and quantitative, and can be used to analytically determine the amount of unsaturation in a molecule (bromine number). In the case of unsymmetrical reagents, such as hydrogen bromide (HBr), the addition to unsaturated compounds can proceed by ionic or free-radical reaction mechanisms. Under ionic reaction conditions and in the absence of peroxides, the product is the one that is formed from the most stable carbocation intermediate generated by the addition of a proton to the double bond (normal or Markovnikov addition). Because carbocation stability follows the order tertiary > secondary > primary, a 2-bromoalkane is the expected product when HBr is added to a 1-alkene:

$$RCH{=}CH_2 \xrightarrow{H^+} R\overset{+}{C}HCH_3 \xrightarrow{Br^-} RCHBrCH_3$$

Under free-radical reaction conditions the reverse order of addition to olefins occurs (abnormal or anti-Markovnikov addition). Thus the reaction of HBr with a 1-alkene in the presence of a peroxide or other free-radical initiator produces a 1-bromoalkane. The reaction proceeds by radical chain-initiating and chain-propagating steps, and the direction is based on the most stable free radical (tertiary > secondary > primary) generated by the addition of a bromine atom to the olefin:

$$RCH{=}CH_2 \xrightarrow{Br\cdot} R\dot{C}HCH_2Br \xrightarrow{HBr} RCH_2CH_2Br + Br\cdot$$

The free-radical addition of HBr is unique because the other hydrogen halides undergo addition to olefins only by the ionic pathway.

1.3.2. Substitution Reactions

Bromine and other reagents containing bromine can react with various organic compounds by substitution reactions that involve the replacement of hydrogen by bromine. In the case of saturated hydrocarbon substrates, such as alkanes and alkyl aromatic compounds, the bromination reaction occurs by a free-radical chain reaction and requires thermal, photolytic, or other forms of initiation:

$$Br_2 \longrightarrow 2\,Br\cdot \text{ (initiation)}$$
$$RH + Br\cdot \longrightarrow R\cdot + HBr \text{ (propagation)}$$
$$R\cdot + Br_2 \longrightarrow RBr + Br\cdot \text{ (propagation)}$$
$$R\cdot + Br\cdot \longrightarrow RBr \text{ (termination)}$$

where R is alkyl or alkylaryl.

The reaction with most linear aliphatic hydrocarbons is not generally useful because hydrogen abstraction by bromine atoms occurs at random in the hydrocarbon, and the radical chains are short and not self-sustaining. Reaction selectivity is low because the carbon–hydrogen bond dissociation energy differences between the many secondary hydrogens in a long-chain alkane are negligible. Reaction selectivities are much greater with branched alkanes because of the greater energy differences between the different carbon–hydrogen bonds, such that tertiary alkyl bromides can be the predominant substitution product.

In the case of alkyl aromatic compounds, such as ethylbenzene, the substitution reaction is very selective because of the much greater energy differences between the radicals generated in the hydrogen abstraction step. Because the benzylic radical ($C_6H_5\dot{C}CH_3$) is much more stable than the primary radical ($C_6H_5CH_2CH_2\cdot$), 1-bromo-1-phenylethane is essentially the only product formed.

Aromatic compounds (ArH) undergo nuclear substitution reactions with bromine by a ratelimiting electrophilic addition to form an arenium ion intermediate followed by the rapid loss of a proton:

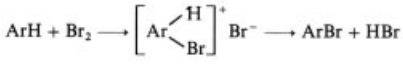

Replacement of hydrogen by bromine usually requires a Lewis acid catalyst when the aromatic compound contains deactivating (nitro) or weakly activating (alkyl) substituents. No catalyst is required for the bromination of more activated aromatics, such as phenol or aniline. The amount and orientation of bromine substitution is dependent on a number of factors, such as stoichiometry, catalyst activity, reaction conditions, and substituent effects.

An equimolar mixture of bromine and chlorine (bromine chloride) can also be used for these substitution reactions. In free-radical substitutions the overall reaction is faster than with bromine because the more energetic chlorine atoms lower the energy of activation in the hydrogen abstraction step. Although minor amounts of chlorinated products are usually produced, bromine utilization is improved because HCl rather than HBr is produced as the byproduct. This reagent is often used in place of bromine for both aliphatic and aromatic substitution reactions when there are not adequate facilities to recycle or use HBr.

Other substitution reactions that are often used in commercial chemical processes include: the replacement of allylic and benzylic hydrogen atoms by bromine using *N*-bromo compounds, such as *N*-bromosuccinimide and 1,3-dibromo-5,5-dimethylhydantoin; the substitution of hydroxyl groups of alcohols by bromine using phosphorus compounds, such as phosphorus tribromide or triphenylphosphonium dibromide; and the replacement of chlorine atoms of many chlorinated hydrocarbons by bromine with various Lewis acids and phase-transfer catalysts. Various reagents can also be used to substitute the hydrogen atoms adjacent to carbonyl groups to produce the α-bromo derivatives of aldehydes, ketones, carboxylic acids, and active methylene compounds.

1.4. Commercial Products

Of the numerous organic bromine compounds that are commercially produced, the individual compounds discussed in this chapter were selected because of their industrial importance and because they illustrate the most common patented production processes. The order of treatment is by chemical reaction classification and begins with addition reactions, followed by substitution reactions of bromine for hydrogen, hydroxyl, and halogen. Because brominated flame retardants consume a significant amount of bromine, the production of this growing and important class of compounds is treated separately.

Most of these commercial products have different product quality specifications because they are generally produced by several manufacturers. Additional physical and chemical properties, first aid and personal protection information, toxicity and safety data, and proper precautions for handling, storage, and transportation are available from the manufacturers.

1.4.1. 1,2-Dibromoethane

Production. 1,2-Dibromoethane [*106-93-4*], $BrCH_2CH_2Br$ (ethylene dibromide), is manufactured by the uncatalyzed addition of bromine to ethylene.

The commercial process is carried out in a glass-column reactor consisting of a lower packed section and an unpacked upper section containing a number of superimposed, high-capacity, coil heat exchangers. Liquid bromine is continuously added above the packed section while a slight excess of ethylene is continuously fed countercurrently from the bottom of the packed section. The exothermic reaction between ethylene and bromine occurs in the liquid phase on the surfaces of the cooling coils, and heat is removed at a rate sufficient to maintain a maximum temperature of 100 °C in this section of the column. Some reaction also occurs in the gas phase above the bromine feed where product is condensed and separated from the vent gas (ethylene, HBr, and inert material). As the crude liquid product passes downward through the packed section, it provides a contacting surface for rising ethylene to convert any residual dissolved bromine. Ethylene dibromide is continuously withdrawn from the reactor into a hold-up tank where it is irradiated with ultraviolet light to eliminate minor amounts of unconverted starting materials. The technical-grade product is obtained in 99.5 % purity as

a clear, colorless liquid in close to quantitative yield [11].

A continuous process that claims lower corrosivity and high output of ethylene dibromide involves the countercurrent addition of bromine and ethylene plus the continuous addition of a cold concentrated aqueous solution of inorganic bromide, which functions as the heat exchange medium [12]. A process that uses ethylene dibromide as the reaction medium is also reported to allow good control of the exothermic reaction between bromine and ethylene [13].

Uses. The largest market for ethylene dibromide is in gasoline. It acts as a scavenger for lead alkyls and prevents buildup of lead oxides in automobile engines. This use, which amounted to about 35 – 40 % of United States bromine consumption in 1981, has been declining since 1974 as a result of the lead phasedown mandated by the U.S. Environmental Protection Agency (EPA). Government agencies in many other countries are following this course of action to decrease the amount of lead in the atmosphere. The continued production of nonlead tolerant automobiles and conservation policies will more than likely result in further declines.

The other major application for ethylene dibromide is for soil and space fumigation, and the postharvest fumigation of fruit. It was used as a substitute for DBCP (1,2-dibromo-3-chloropropane), $BrCH_2CHBrCH_2Cl$ [*96-12-8*], which was banned by the EPA in the late 1970s because workplace exposure led to azoospermia; it was applied to control pests, such as soil nematodes, rodents, and most species of insects. In late 1983 the EPA took emergency action to cancel the pesticide registrations for ethylene dibromide formulations used in most soil fumigation applications in the United States because of groundwater contamination. Some minor fumigation uses were allowed to continue, but the predominant volume of ethylene dibromide used in soil fumigation will cease.

Smaller amounts of ethylene dibromide are used as a reaction intermediate in the manufacture of other chemicals, such as the reactive flame retardant monomer, vinyl bromide [*593-60-2*]. Minor amounts are used as a high-density, nonflammable solvent in a variety of applications.

1.4.2. 1-Bromododecane

Production. 1-Bromododecane [*143-15-7*], n-$C_{12}H_{25}Br$ (lauryl or n-dodecyl bromide), is produced by the free-radical addition of gaseous HBr to 1-dodecene. Almost all industrial processes to manufacture the higher 1-bromoalkanes containing 6 – 18 carbon atoms employ this addition reaction. The alternative manufacturing process, the reaction of the corresponding alcohols with HBr, is not as economical because of the higher cost and lower reactivity of the alcohols compared to the olefins.

The continuous process is conducted in the liquid phase by contacting 1-dodecene containing a free-radical initiator with anhydrous HBr in a glass-lined reactor. Teflon (polytetrafluoroethylene) or Kynar (poly(vinylidene fluoride)) reactor linings are suitable alternatives. The reaction is usually carried out under moderate pressure to provide good mass transfer. Reaction temperatures of less than 20 °C are usually necessary to obtain high regioselectivity to 1-bromododecane.

Various continuous reactor designs that have been used to produce high ratios of primary to secondary alkyl bromides at higher temperatures (20 – 30 °C) include a two-stage reactor [14], and a thin turbulent rising film reactor [15]. The latter reactor operates by forming a thin film of olefin in a vertical tubular reactor, which is contacted with an upward stream of gaseous HBr to maintain turbulence and initimate mixing. Because residence times are short and the exothermic reaction is easily controlled by cooling the large surface area of the reactor tube, a 22 : 1 molar ratio of 1-bromo- to 2-bromododecane is obtained at 27 °C. When HBr is added to the olefin under batch reaction conditions, the rate of the competing ionic addition reaction increases and the molar ratio of primary to secondary bromides falls to 8 : 1 at comparable temperatures (32 °C).

The crude reaction product can be either flashed or contacted with an aqueous solution of sodium hydroxide to remove excess HBr. The conversion of 1-dodecene and the yield of 1- and 2-bromododecane are essentially quantitative.

Free-radical initiators that have been employed for this reaction are preformed olefin ozonides [16], olefin peroxides generated by preblowing air through the olefin [17], oxygen

[18] or phosgene [19] cofed with HBr, organic peroxides alone [20] or in the presence of phase-transfer catalysts [21], molecular sieves [22], and ultraviolet light [23].

Uses. *n*-Alkyl bromides have assumed considerable importance as starting materials for the production of many types of organic derivatives. The major use for *n*-alkyl bromides with 12 – 16 carbon atoms is in the production of alkyldimethylamines by the nucleophilic displacement of bromide ion with dimethylamine. These amines are used as intermediates in the manufacture of fatty amine oxides and alkyl dimethyl benzyl quaternary ammonium salts, which are used in a wide range of products because of their biological (germicidal) or physical (cationic) properties. Minor amounts of *n*-alkyl bromides are used to produce insect pheromones, phasetransfer catalysts, and other useful specialty chemicals.

1.4.3. 1-Bromo-3-chloropropane

Production. The versatile synthetic intermediate 1-bromo-3-chloropropane [*109-70-6*], $Br(CH_2)_3Cl$ (trimethylene chlorobromide), is virtually always produced by the free-radical addition of anhydrous HBr to allyl chloride.

Continuous or batch-type processes in glass-lined or impervious polymer-coated equipment can be employed. Gaseous HBr is added under slight positive pressure to a solution of allyl chloride dissolved in an inert, low-dielectric solvent using a free-radical initiator, such as oxygen [24], molecular sieves [25], organic peroxides [26], or ultraviolet irradiation [27].

Because of the inductive effect of the neighboring chloride group, the addition reaction is not as rapid or exothermic as the corresponding hydrobromination of 1-propene. The competing ionic addition to form 2-bromo-1-chloropropane [*3017-95-6*] is retarded to a much greater extent than the free-radical reaction so that low reaction temperatures are not necessary to achieve a high ratio of primary to secondary bromide. Metal ions that accumulate during allyl chloride storage in metal tanks or drums are usually removed because they can promote ionic addition.

The yield of crude product, which typically consists of 95 % primary and 5 % secondary bromochloropropanes, is about 95 % based on converted allyl chloride. The minor isomer can easily be separated by distillation if product specifications call for higher purity.

Uses. Trimethylene chlorobromide is a useful reaction intermediate to prepare γ-chloropropyl derivatives because of the greater reactivity of the bromine atom compared to the chlorine atom in nucleophilic displacement reactions. For example, the cyanide displacement of trimethylene chlorobromide forms γ-chlorobutyronitrile, an intermediate used for the production of herbicides, pesticides, pharmaceuticals, and other commercial products.

1.4.4. 1-Bromo-2-phenylethane

Production. The useful intermediate, 1-bromo-2-phenylethane [*103-63-9*], $C_6H_5CH_2CH_2Br$ (β-phenethyl bromide), is usually manufactured by the free-radical addition of HBr to styrene.

The typical batch process is run in a glass-lined reactor by sparging gaseous HBr through a solution of styrene in an inert solvent containing a free-radical initiator. The reaction, which is rapid and moderately exothermic, can be controlled by introducing HBr at about the rate it is consumed. When styrene is completely converted, any excess HBr that is present can be recovered by various techniques prior to recovery of solvent and product by distillation. Styrene conversions are usually quantitative and β-phenethyl bromide yields are typically 95 %. Although most patents describe batch reactions, the process can also be carried out continuously.

The essential features necessary to obtain high selectivities to the desired β-isomer and negligible amounts of the undesired α-isomer, 1-bromo-1-phenylethane [*585-71-7*], are the use of dilute solutions of styrene (< 30 wt %) in an inert, nonpolar solvent, such as heptane; reaction temperatures in the range of 40 – 95 °C; and the absence of water and free-radical inhibitors. Organic peroxides [28], oxygen or air [29], azo compounds [30], and irradiation by ultraviolet light or by ^{60}Co [31] are examples of free-radical initiators that have been employed for this reaction.

Although β-phenethyl bromide can be prepared in comparable yields by reaction of the more expensive β-phenethyl alcohol with phosphorus tribromide [32] or with other reagents, this alternative method is seldom used because of economic considerations.

Uses. β-Phenethyl bromide is an important starting material for the production of various β-phenethyl derivatives, pharmaceuticals, fragrances, and other fine chemicals.

1.4.5. 1-(Bromomethyl)-3-phenoxybenzene

Production. The synthetic pyrethroid starting material, 1-(bromomethyl)-3-phenoxybenzene [*51632-16-7*], 3-$C_6H_5OC_6H_4CH_2Br$ (*m*-phenoxybenzyl bromide), is produced by the thermal side chain bromination of *m*-phenoxytoluene. The usual byproduct from this benzylic bromination is 1-(dibromomethyl)-3-phenoxybenzene [*53874-67-2*], 3-$C_6H_5OC_6H_4CHBr_2$ (*m*-phenoxybenzal bromide).

The batch reaction is carried out in a glass reactor equipped with an appropriate HBr recovery system by feeding 20 % molar excess bromine vapor to neat *m*-phenoxytoluene at 265 °C. The rate of addition should be slow enough to avoid high local concentrations of bromine so that only negligible amounts of undesired ring-brominated byproducts are formed. Metal contamination must be avoided for the same reason and also to prevent Friedel-Crafts alkylation reactions. This uncatalyzed thermal bromination process produces the benzyl and benzal bromides in a 25 : 1 molar ratio in about 65 % yield based on 90 % converted *m*-phenoxytoluene. The low amount of *m*-phenoxybenzal bromide produced is partly due to its thermal instability, which results in the formation of some oligomeric byproducts at these high reaction temperatures [33].

Other thermal processes for brominating *m*-phenoxytoluene employ a catalyst, such as phosphorus trichloride [34], or activation with ultraviolet radiation [35]; however, appreciable amounts of ring-brominated byproducts are formed in these batch reactions. Continuous bromination of *m*-phenoxytoluene with ultraviolet light at 225 °C is claimed to eliminate formation of ring-brominated byproducts, but the product distribution of the benzyl and benzal bromides was not reported [35].

Additional *m*-phenoxytoluene brominating agents that have been used are $BrCCl_2CCl_2Br$ [36] and *N*-bromosuccinimide [37]. The former reagent is reported to produce mostly the benzyl bromide; the latter claims the predominant formation of the benzal bromide.

The competitive and analogous free-radical chlorination of *m*-phenoxytoluene to produce the corresponding benzyl chloride has been reported, but the reaction is not as selective as bromination and more ring-halogenated byproducts are formed.

Uses. Both *m*-phenoxybenzyl and *m*-phenoxybenzal bromides are important starting materials for the production of synthetic pyrethroid reaction intermediates. They can be converted by various reaction sequences to the alcohol, *m*-phenoxybenzyl alcohol, or to the aldehyde, *m*-phenoxybenzaldehyde, the key intermediates common to these synthetic pyrethroids. Both the alcohol and the aldehyde must contain less than 0.5 % ring-brominated byproduct for use in this application, hence the importance of minimizing the aromatic substitution side reaction.

1.4.6. Bromoacetic Acid

Production. Bromoacetic acid [*79-08-3*], $BrCH_2CO_2H$, and other α-bromocarboxylic acids are usually prepared by the well-known Hell-Volhard-Zelinski reaction in which the α-hydrogens of the carboxylic acid are replaced by bromine. A catalyst, such as a phosphorus trihalide, is necessary to convert the acid to the acid halide, the actual reaction intermediate. If the acid chloride or bromide is available as a starting material, no catalyst is needed and only one equivalent of bromine is required versus the two equivalents necessary when the carboxylic acid is the starting material. The resulting bromoacetyl halide can either be hydrolyzed or esterified to produce bromoacetic acid or its esters.

High yields of acid bromides can be obtained by adding one equivalent of bromine to the carboxylic acid containing phosphorus tribromide at 25 °C [38]. In practice a portion of the acid is converted to the acid bromide at 25 – 30 °C and then the remainder of the two equivalents of

bromine is fed over time at higher temperatures to complete the reaction. The intermediate, bromoacetyl bromide [*598-21-0*], is hydrolyzed by the controlled addition of water and the resulting bromoacetic acid is separated from the major impurity, dibromoacetic acid [*631-64-1*], by distillation. The hydrogen bromide byproduct that is evolved during the reaction and the hydrolysis is usually recovered as a concentrated aqueous solution for subsequent use in other processes or for recycle to bromine. Bromine utilizations can be improved by using bromine chloride as the brominating agent to produce bromoacetyl chloride [*22118-09-8*], which is subsequently hydrolyzed to bromoacetic acid and HCl [39].

If the corresponding α-chlorocarboxylic acid is available, it can be converted to the α-bromo-carboxylic acid by reaction with gaseous HBr catalyzed by aluminum chloride [40] or with concentrated hydrobromic acid and an azeotropic solvent to remove water and hydrochloric acid [41]. The nucleophilic displacement reaction of α-chlorocarboxylic acids with inorganic bromides, such as KBr or NaBr, is an alternative method [42]. The reaction of chloroacetic esters with hydrobromic acid to produce bromoacetic acid and an alkyl bromide from the ester portion of the molecule is another variation of the reaction [43].

Uses. Bromoacetic acid as well as other α-bromocarboxylic acids and their esters are used in many chemical processes as starting materials or as reaction intermediates.

1.4.7. Bromomethane

Production. Bromomethane [*74-83-9*], CH_3Br (methyl bromide), is produced from methanol and hydrogen bromide, which is generated in situ from bromine and a reducing agent like sulfur or hydrogen sulfide.

The commercial process is conducted by continuously feeding streams of liquid bromine, elemental sulfur, and methanol to a reactor maintained at 75 – 85 °C under a slight positive pressure. Methyl bromide vapor (99 % purity) is continuously withdrawn from the reactor overhead system and scrubbed free of dimethyl ether by contacting the gas with a sulfuric acid scrubber. Excess methanol and sulfur are used because some methanol is entrained in coproduced sulfuric acid and because some sulfur is consumed in formation of a thio acid byproduct. Sulfuric acid containing thio acid, sulfur, and methanol is continuously purged from the reactor and either processed for resale or treated for waste disposal. By operating at a higher temperature and pressure than described in other similar processes, this process allows for good control of the exothermic reaction, retards the formation of dimethyl ether, and produces methyl bromide in high yield and purity [44]. Some methyl bromide is produced by the decomposition of a brominated ester of the thio acid, thus the stoichiometry deviates somewhat from that shown in Equation 1; therefore, this process produces a 6 : 1 molar ratio of methyl bromide to sulfur, another process using hydrogen sulfide as the reducing agent [45] has the advantage of producing an 8 : 1 ratio (Eq. 2):

$$6\,CH_3OH + 3\,Br_2 + S \longrightarrow 6\,CH_3Br + H_2SO_4 + 2\,H_2O \quad (1)$$

$$8\,CH_3OH + 4\,Br_2 + H_2S \longrightarrow 8\,CH_3Br + H_2SO_4 + 4\,H_2O \quad (2)$$

Methyl bromide can also be coproduced with other useful organic bromine compounds by the reaction of methanol solvent with byproduct hydrogen bromide. Thus, when a solution of phenol in methanol is reacted with bromine, aromatic substitution produces a 96 % yield of the flame retardant intermediate, 2,4,6-tribromophenol [*118-79-6*], and hydroxyl substitution of the solvent produces methyl bromide in 34 % yield [46]. The flame retardant tetrabromo-bisphenol-A [*79-94-7*] and methyl bromide are obtained in 95 % and 61 % yields, respectively, by the similar bromination of bisphenol-A in methanol [47]. Even though methyl bromide yields are much less than theory because of the equilibrium, $CH_3OH + HBr \rightleftharpoons CH_3Br + H_2O$, favorable economics are achieved because less methanol and HBr need to be recovered, and a saleable coproduct is produced.

Uses. Methyl bromide is a colorless, odorless, nonflammable, poisonous gas at ambient temperature and pressure. It is primarily used as a space and soil fumigant to control a wide range of animal and plant pests and disease

organ-isms. In some of these fumigation applications methyl bromide containing the odorant, chloropicrin (Cl_3CNO_2), competes with ethylene dibromide.

A smaller but important use of methyl bromide is in the synthesis of many pharmaceuticals and fine chemicals, either as a methylating agent or as the Grignard reagent, methylmagnesium bromide [*75-16-1*].

1.4.8. 3-Bromo-1-propene

Production. The useful intermediate, 3-bromo-1-propene [*106-95-6*], $CH_2{=}CHCH_2Br$ (allyl bromide), is usually produced commercially from allyl alcohol and various brominating agents. The laboratory method of reacting allyl alcohol with hydrobromic acid and sulfuric acid is probably the most used process even though other reactants, such as red phosphorus and bromine [48], can be employed.

The batch reaction of allyl alcohol with a 35 % molar excess of 60 % hydrobromic acid is carried out in a stirred, glass-lined reactor at reflux (68 °C) for 3 h. The aqueous phase containing excess hydrobromic acid is separated and the organic phase is washed successively with water and aqueous sodium carbonate. Flash distillation produces allyl bromide (98 – 99 % purity) in about 75 % yield. An additional 6 % yield of allyl bromide can be recovered from the aqueous phases by extraction. The major impurity, 1,2-dibromopropane [*78-75-1*] formed by addition of HBr to the product, is readily separated by the distillation. Although the yield is lower than that from the similar process employing sulfuric acid (92 % yield), the process using 60 % hydrobromic acid has the advantage of higher reactor loading and does not require recycle of sulfuric acid [49]. Allyl bromide can also be prepared in 50 % yield by the halogen exchange reaction of allyl chloride with excess hydrobromic acid in the presence of a copper bromide catalyst [50]. The allylic bromination of propene by *N*-bromo derivatives or by bromine in the gas phase are alternative methods; however, they are not widely used commercially.

Uses. Allyl bromide is a very reactive intermediate capable of undergoing facile displacement reactions with many nucleophiles to produce synthetic perfumes, pharmaceuticals, and other compounds containing the allyl functionality.

1.4.9. Bromochloromethane and Dibromomethane

Production. Bromochloromethane [*74-97-5*], CH_2BrCl (methylene chlorobromide), can be produced from dichloromethane by a halogen exchange reaction using either liquid bromine or gaseous HBr:

$$6\,CH_2Cl_2 + 3\,Br_2 + 2\,Al \longrightarrow 6\,CH_2BrCl + 2\,AlCl_3$$

$$CH_2Cl_2 + HBr \xrightarrow{AlCl_3} CH_2BrCl + HCl$$

The unisolated bromochloromethane can be further brominated to produce dibromomethane [*74-95-3*], CH_2Br_2 (methylene bromide), by adjusting the reactant stoichiometry and some process variables:

$$6\,CH_2BrCl + 3\,Br_2 + 2\,Al \longrightarrow 6\,CH_2Br_2 + 2\,AlCl_3$$

$$CH_2BrCl + HBr \xrightarrow{AlCl_3} CH_2Br_2 + HCl$$

In the process using bromine to produce bromochloromethane, a stirred reactor is charged with aluminum filings and one-sixth of the total dichloromethane charge, followed by a small amount of bromine to initiate the reaction. When the reaction temperature approaches 35 °C the remaining dichloromethane, which is used in excess, is added, followed by the slower addition of bromine at a rate sufficient to maintain a mild reflux. After completion of the reaction, the crude reaction mixture is pumped to a hydrolyzer vessel from which the halogenated methanes are steam distilled. The crude product, which is separated from water, consists of 90 % bromochloromethane and 10 % dibromomethane in addition to unconverted starting material. Fractional distillation is employed to separate all three components. The overall yield is about 85 % based on the bromine that was used [51].

In the process employing HBr to produce dibromomethane, three reactors are connected for passage of a gas in series and each is charged with dichloromethane and aluminum chloride catalyst (5 mol %). Gaseous HBr is fed into the liquid phase of the first vessel and the evolved

gas from each vessel is fed into the next vessel in sequence. The gas exiting from the last vessel is principally hydrogen chloride. When the contents of the first vessel consist of 99 % CH_2Br_2 and 1 % CH_2BrCl, the composition in the second vessel is 46 % CH_2Br_2, 47 % CH_2BrCl, and 7 % CH_2Cl_2, and the composition in the third vessel is 18 % CH_2Br_2, 55 % CH_2BrCl, and 27 % CH_2Cl_2. This overall product composition corresponds to a 91 % utilization of the HBr fed. At this point the first reactor is discharged for recovery of product, recharged with starting materials, and reconnected to function as the last vessel in a new reaction sequence. The HBr feed line is switched from the first reactor to the second reactor, which becomes the first vessel in the next cycle. Thus, the third reactor in the original sequence operates as the second vessel in the new sequence [40].

Other patented processes to produce these materials include the gas-phase bromination of methyl chloride with a mixture of HBr and chlorine [52] or a mixture of bromine and chlorine [53], and the liquid-phase nucleophilic displacement reaction of dichloromethane with inorganic bromides in aprotic [54] and protic [55] solvents.

Uses. The major use for bromochloromethane is as a fire-extinguishing fluid in portable systems. Dibromomethane is primarily used as a solvent and as a gauge fluid. Both compounds have been employed as reaction intermediates for the production of several herbicides and pesticides.

1.4.10. Trifluorobromomethane and Difluorobromochloromethane

Production. Both trifluorobromomethane [*75-63-8*], CF_3Br, and difluorobromochloromethane [*353-59-3*], CF_2BrCl, are usually produced by substitution reactions involving the replacement of either hydrogen or chlorine atoms of fluorocarbons by bromine.

These reactions are carried out in the gas phase at elevated temperatures (450 – 650 °C) by continuously feeding the reactants through a tube reactor. To obtain high reaction selectivities to the desired halomethane, the most important process variables are contact times and the proper ratios of reactants, which in turn are dependent on the temperature and whether or not a catalyst is employed. In most cases starting material conversions range from about 40 to 90 % with product selectivities of 90 % or higher.

Trifluorobromomethane can be produced along with byproduct HBr by the gas-phase substitution reaction of trifluoromethane with bromine in a tube reactor packed with catalysts, such as alumina [56], or aluminum oxide – fluoride containing nickel [57]. Alternatively the reaction can be carried out by cofeeding trifluoromethane and bromine containing minor amounts of either sulfur trioxide [58] or chlorine [59] at elevated temperatures.

Other patented processes to produce CF_3Br include the disproportionation reaction of difluorodibromomethane (CF_2Br_2) catalyzed by aluminum chloride [60], and the zinc bromide – carbon catalyzed chloride exchange reaction of trifluorochloromethane (CF_3Cl) with hydrogen bromide [61].

Difluorobromochloromethane is produced by similar halogen exchange and disproportionation reactions. The gas-phase reaction of CF_2ClH with bromine in a tube lined with alumina (560 °C) produces a 96 % yield of a mixture of CF_2BrCl and CF_2Br_2 (15 : 1 molar ratio) based on a 91 % conversion of starting material [62]. The byproduct, CF_2Br_2, can be separated and converted to CF_2BrCl by the gas-phase reaction with chlorine [63].

Uses. Both halogenated hydrocarbons are employed as fire-extinguishing agents, especially in high-value applications, such as electronics and computer installations. They are just as effective as bromochloromethane but are much less toxic. Trifluorobromomethane is mainly used in automatic fire-extinguishing systems, and CF_2BrCl is primarily used in portable fire extinguishers.

1.4.11. Flame Retardants (→Flame Retardants)

Fires in industrialized countries inflict a heavy toll in terms of economic loss, human suffering, and death. In response to this problem, the chemical industry has developed and commercialized flame retardants to reduce the flammability and

to improve the fire safety of polymeric materials. Flame retardants are incorporated into polymers either by physical blending techniques, *additive type,* or by chemical methods in which part of a conventional monomer is replaced with a flame retardant monomer in the polymerization process, *reactive type.* Although reactive flame retardants have the potential for greater performance in the polymer (migration and leaching are traditional problems with polymer additives in general), economic considerations have limited their acceptance mostly to thermoset polymers, while the additive types are widely used in both thermoplastic and thermoset polymers.

In the late 1950s pentabromobiphenyl oxide was the first aromatic bromine flame retardant to gain commercial significance. Since then the development of new flame retardants has focused on products with increased molecular mass, such as decabromobiphenyl oxide and poly(tribromostyrene). The benefits of higher mass are lower volatilization during processing and reduced surface migration during polymer use. Other significant parameters that have been improved over the years with the introduction of new flame retardants are enhanced thermal, light, and chemical stability, reduced toxicity, and improved physical and engineering properties of the resulting polymer formulations. The major markets for flame retardants are the electrical, electronic, appliance, automotive, textile, and furniture industries.

A number of industrially important brominated flame retardants used in polymers are listed in Table 3. Although manufacturing procedures for these materials are not generally published, those described in this section are from the patent literature and illustrate the different types of processes that can be employed.

1.4.11.1. Tetrabromobisphenol-A

Production. Tetrabromobisphenol-A (TBBPA), 4,4′-(1-methylethylidene)bis-(2,6-di-bromophenol) [*79-94-7*], the largest selling brominated flame retardant, is produced by the bromination of bisphenol-A. The reaction is an example of substitution on an activated aromatic system and does not require catalysis.

The bromination reaction is generally conducted in solvents, such as a halocarbon [64], a halocarbon and water [65–67], a halocarbon and 50 % hydrobromic acid [68], or aqueous alkyl monoethers [69]. Aqueous acetic acid is also a satisfactory medium [70], and acetic acid with added sodium acetate is reported to improve product color [71].

The HBr byproduct can either be recovered and recycled to recover bromine values or utilized in situ by means of one of two techniques. The first method involves the addition of an oxidizing agent to the reaction medium to convert evolved HBr to bromine, which is subsequently consumed in the reaction. Suitable oxidizing agents are hydrogen peroxide [72], [73], sodium bromate [74], sodium chlorate [75], or chlorine [76], [77]. In the case of chlorine, the amount and rate of addition must be carefully controlled to limit chlorination.

The other approach is the use of an alcohol as the solvent to coproduce an economically useful alkyl bromide by the reaction of byproduct HBr with the alcohol [78]. Methanol is the solvent of choice because methyl bromide is widely marketed as a fumigant.

In the typical batch reaction, bromine in slight excess is added to a stirred bisphenol-A solution in methanol in a suitable reactor at 25 – 30 °C. The exothermic reaction is controlled by the rate of bromine addition, which is dependent on the refrigeration capacity of the reactor cooling system. After bromine addition, the contents of the reactor are heated to reflux (65 °C) for a short period to allow the slower reaction between methanol and dissolved HBr to occur. During this postreaction period, methyl bromide and some HBr are vented through the overhead condenser system into a series of scrubbers that neutralize and dry methyl bromide, which is produced in about 60 % yield based on byproduct HBr. Water is then added to precip-itate TBBPA and the resulting slurry is cooled, filtered, washed, and dried. The yield of TBBPA is generally 94 – 97 % based on bisphenol-A [47].

Uses. Commercial TBBPA is available in two grades, an epoxy resin grade and a higher quality polycarbonate grade. TBBPA is a reactive flame retardant; both hydroxyl groups can be reacted with epichlorohydrin under basic conditions to form the diglycidyl ether, which is widely used in epoxy resin formulations. TBBPA is also used in polycarbonate and ether

Table 3. Physical properties of brominated flame retardants

Compound	CAS registry no.	Structure	M_r	Bromine content, %	*mp*, °C
Dibromoneopentyl glycol	[*3296-90-0*]	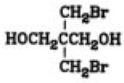	262.0	61.0	109–110
Tetrabromo-*p*-xylene	[*23488-38-2*]	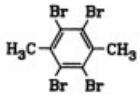	421.8	75.8	254–255
1-Dibromoethyl-3,4-dibromo-cyclohexane	[*3322-93-8*]	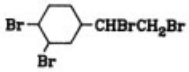	427.8	74.7	70 (65–68 *)
Tetrabromophthalic anhydride	[*632-79-1*]	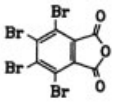	463.8	68.9	275–280 (270–276 *)
Tetrabromobisphenol-A	[*79-94-7*]	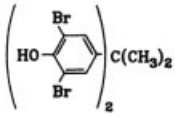	543.9	58.8	179 (179–181 *)
Pentabromoethylbenzene	[*85-22-3*]	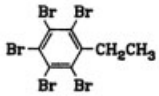	500.6	79.8	138 (136–138 *)
Pentabromochlorocyclo-hexane	[*87-84-3*]	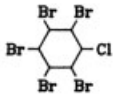	513.1	77.9	203
Pentabromobiphenyl oxide	[*32534-81-9*]	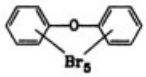	564.8	70.7	liquid
Hexabromocyclododecane	[*3194-55-6*]	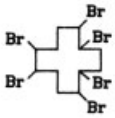	641.8	74.7	170–180 (185–195 *)
Hexabromobenzene	[*87-82-1*]	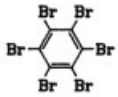	551.6	86.9	327
1,2-Bis(2,4,6-tribromo-phenoxy)ethane	[*37853-59-1*]	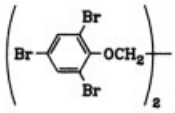	687.7	69.7	223–225

Table 3. Continued.

Compound	CAS registry no.	Structure	M_r	Bromine content, %	*mp*, °C
Octabromobiphenyl oxide	[*32536-52-0*]	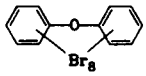	801.5	79.8	70–150
1,2-Bis(tetrabromo-phthalimido)ethane	[*32588-76-4*]	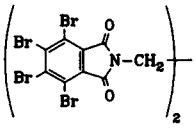	951.6	67.2	446–450
Decabromobiphenyl oxide	[*1163-19-5*]	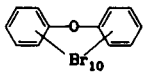	959.3	83.3	304 (300–315*)
1,2-Bis(pentabromo-phenoxy)ethane	[*61262-53-1*]	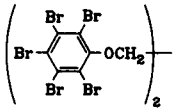	1003.3	79.6	322
Poly(tribromostyrene)	[*57137-10-7*]	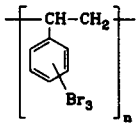	$(340.9)_n$	70.3	–

* *mp* reported by a different manufacturer

polyester resins requiring low color and good clarity.

1.4.11.2. Decabromobiphenyl Oxide

Production. Decabromobiphenyl oxide (DBBPO), decabromodiphenyl ether, 1,1′-oxybis(2,3,4,5,6-pentabromobenzene) [*1163-19-5*], is manufactured by the exhaustive bromination of phenyl ether. As the bromination reaction progresses, partial substitution of hydrogen by bromine deactivates the aromatic ring toward further substitution, thus necessitating more vigorous reaction conditions than those used in the preparation of TBBPA. Lewis acid catalysis and an excess of bromine are used to obtain high conversions of the less reactive, under-brominated intermediates to DBBPO.

Decabromobiphenyl oxide can be prepared at atmospheric pressure by reacting bromine with phenyl ether in ethylene dibromide solvent and in the presence of aluminum bromide catalyst [79]. The use of bromine in an organic solvent requires long reaction times, results in low productivity per reactor volume, and necessitates recycling the solvent. These limitations have led to the use of bromine as both the reactant and the solvent [80–82].

A typical batch process to produce DBBPO [82] involves the addition of phenyl ether over one hour to an agitated reactor containing 150 % excess of the stoichiometric amount of bromine and catalytic amounts of aluminum bromide. The reaction temperature is maintained at 44–55 °C throughout the addition. After a 2-h postaddition reaction at reflux (59 °C), the product is recovered by first steam distilling the excess bromine from the reactor and then filtering the aqueous slurry. The resulting solid cake is washed with water and oven dried at 110 °C. The crude DBBPO, which consists of about 96 %

deca-, 3 % nona-, and 1 % octabromobiphenyl oxides, is obtained in high yield and contains small amounts of occluded bromine and hydrogen bromide. These occluded impurities can be substantially removed by first grinding and subsequently heating the crude DBBPO [83].

Uses. Decabromobiphenyl oxide is an additive-type flame retardant that has excellent thermal stability under conditions necessary to process high-impact polystyrene, the polymer system in which it is most widely used. It is also suitable for applications in ABS, rubbers, and some polyolefins and epoxy resins.

1.4.11.3. Tetrabromophthalic Anhydride

Production. Tetrabromophthalic anhydride (TBPA), 4,5,6,7-tetrabromo-1,3-isobenzofurandione [*632-79-1*], is prepared by phthalic anhydride bromination. Phthalic anhydride is a deactivated aromatic compound and is unreactive toward bromine under the usual bromination reaction conditions. TBPA can be prepared by reacting phthalic anhydride and bromine in a mixture of concentrated sulfuric acid and 70 % hydrogen peroxide in the presence of catalytic amounts of iodine [84], or in chlorosulfonic acid containing a small amount of sulfur [85].

The most common reaction medium is fuming sulfuric acid containing 45 – 65 % sulfur trioxide (oleum) and catalytic amounts of iron powder and iodine. If a mixture of bromine and chlorine is employed as the brominating agent, an excess of bromine must be present to limit the incorporation of chlorine in the product [86]. Use of bromine as the halogenating agent requires an excess of sulfur trioxide because it oxidizes byproduct HBr to bromine during the course of the reaction.

The batch reaction is typically carried out by adding 90 % of the bromine (2.1 mol/mol of phthalic anhydride) to the anhydride in excess oleum containing the catalyst at 90 °C. The reaction is completed by heating the reaction mixture to 105 °C followed by the addition of the remaining bromine. Excess sulfur trioxide and bromine are then distilled from the reactor and crude product is recovered from the cooled distillation residue by filtration. The product is thoroughly washed with water to minimize residual sulfuric acid, dried, and obtained in 94 % yield based on phthalic anhydride [87].

Uses. Tetrabromophthalic anhydride is primarily employed as a reactive flame retardant in unsaturated polyester and epoxy resin formulations. It is also used as an intermediate in the manufacture of other flame retardants, such as diallyl tetrabromophthalate and tetrabromophthalate polyol esters.

1.4.11.4. Hexabromocyclododecane

Production. The aliphatic flame retardant, hexabromocyclododecane (HBCD), is produced by the addition reaction of bromine to *trans,trans,cis*-1,5,9-cyclododecatriene, the cyclic trimer of butadiene. The HBCD that is obtained from this reaction consists of a mixture of stereoisomers of 1,2,5,6,9,10-hexabromocyclododecane [*3194-55-6*] ranging from oils to an isomer with a 205 – 208 °C melting point. Commercially available HBCD contains a mixture of solid isomers and has a melting point range of 170 – 180 °C (185 – 195 °C).

Addition of bromine to the cyclic triene in nonpolar solvents results in substantial amounts of free-radical substitution reactions which lead to decreased yields and lower quality product. Use of polar solvents, such as ethanol, reduces the substitution reaction but can give rise to vicinal bromoethoxy byproducts. Mixed solvent systems, such as benzene – *tert*-butyl alcohol [88] or perchloroethylene – ethanol [89], can be employed to improve product purities and yields.

Another technique to reduce the formation of undesired byproducts is to add the bromine and cyclododecatriene simultaneously and separately to ethanol [90]. This process is carried out by separately feeding bromine and cyclododecatriene in a molar ratio of 3 : 1 to 96 % ethanol in a stirred reactor. The reaction temperature is kept at 20 – 30 °C and a 1 – 5 % excess of bromine is maintained during the 2.5 – 3.0 h addition. The precipitated product is recovered by filtration or centrifugation, and the mother liquor is recovered and reused in subsequent reactions after adding makeup ethanol. The total yield of dried HBCD from nine separate batches using recycled mother liquor is 92 % [90].

Uses. The main use for HBCD is as an additive-type flame retardant for extruded and expanded polystyrene foam. Other applications include crystal and high-impact polystyrene, SAN resins, adhesives, and coatings.

1.5. Manufacturers

Organic bromine compounds that are produced in annual volumes of 100 t or more are generally manufactured by companies that also produce elemental bromine. Manufacturers with organic bromine production facilities near their bromine-producing plants have lower production costs than nonbromine producers because bromine shipping and handling costs are minimized and byproduct inorganic bromides can be readily processed to recover bromine values.

The major integrated producers are Dow Chemical, Ethyl Corp., Great Lakes Chemical Corp., and their various subsidiary companies in the United States; the Dead Sea Bromine Group, a member of Israel Chemicals Ltd. (Israel); Chemische Fabrik Kalk GmbH, a subsidiary of Kali and Salz AG (Federal Republic of Germany); ATOCHEM, a subsidiary of Elf Aquitaine (France); and Asahi Glass Company Ltd. and Toyo Soda Manufacturing Company Ltd. (Japan). Associated Octel Company Ltd. (United Kingdom), which manufactures large volumes of ethylene dibromide, can also be included in this group of producers.

1.6. Toxicology and Occupational Health

Organic bromine compounds having adopted TLVs and MAKs are given in Table 4.

1,2-Dibromoethane. Ethylene dibromide (EDB) is severely irritating to skin, eyes, and lungs; acutely toxic; and a profound CNS depressant [93–99]; sensitizer [93], [94], [97]; mutagen [100–105]; and carcinogen [95], [106–111].

Reproductive studies in mammals have demonstrated a no-effect or reversible effect

Table 4. Maximum concentrations in the work environment [91], [92]

	TWA[a)]		STEL[a)]		MAK[a)]		BAT[b)]
	ppm	mg/m^3	ppm	mg/m^3	ppm	mg/m^3	mg/l^3
Bromomethane [*74-83-9*], skin	5	20	15	60	*	*	–
Tribromomethane [*75-25-2*], skin	0.5	5	–	–	*	*	–
Tetrabromomethane [*558-13-4*]	0.1	1.4	0.3	4	–	–	–
Bromochloromethane [*74-97-5*]	200	1 050	250	1 300	200	1 100	–
Trifluorobromomethane [*75-63-8*]	1 000	6 100	1 200	7 300	1 000	6 200	–
Difluorodibromomethane [*75-61-6*]	100	860	150	1 290	100	870	–
Bromoethane [*74-96-4*]	200	890	250	1 110	**	**	–
1,2-Dibromoethane [*106-93-4*], skin	*	*	–	–	**	**	–
1,1,2,2-Tetrabromoethane [*79-27-6*]	1	15	1.5	20	1	14	–
2-Bromo-2-chloro-1,1,1-trifluoroethane [*151-67-7*]	–	–	–	–	5	41	2,5
1,2-Dibromo-3-chloropropane [*96-12-8*]	–	–	–	–	**	**	–
Bromoethylene [*593-60-2*]	5 *	20 *	–	–	–	–	–

* Suspect of carcinogenic potential for humans.
** Carcinogenic in animal experiments
a) Data of 1983
b) Data of 2001

at nontoxic dose levels. Reduced egg weights of hens have been reported [112], [113]. Teratospermia in roosters could not be induced [113]; reproduction in rats could not be affected [114–116]. Ethylene dibromide has been reported to be nonteratogenic in rats [116]. Reversible spermicidal effects have been demonstrated in rats, rams, and bulls [117–122]. A dominant lethal effect in mice or rats, respectively, could not be demonstrated [123], [124].

Two metabolic pathways have been postulated. One involves a glutathione transferase mediated conjugation of EDB; the other is based on oxidative dehalogenation by microsomal enzymes yielding a free-radical intermediate [125].

Disulfiram (tetraethylthiuram disulfide) has been reported to potentiate EDB toxicity in mice and rats and carcinogenicity in rats [126–128]. Disulfiram, an inhibitor of aldehyde dehydrogenase, blocks detoxification of the EDB intermediate, 2-bromoacetaldehyde, which results in mutagenicity and DNA binding [129].

EDB has now been prohibited for use as a soil fumigant by EPA [130], and OSHA has proposed a reduction in permissible industrial exposure limits from 20 ppm to 0.1 ppm [131].

1-Bromododecane. No toxicology records exist in computer literature data bases (N.L.M., S.D.C., Dialog) from 1965 to 1984.

1-Bromo-3-chloropropane (TMCB). TMCB vapors cause severe eye and skin irritation, typical haloalkane cardiovascular effects [132], and CNS depression [133]. LC_{50} (fish): 75 mg/L/24 h [134]; BOD: 0.02 g/g, [135]; it is nonmutagenic (Ames); nondominant lethal; spermatogenesis was not impaired in mice treated at 300, 600, or 1200 mg/kg/d for 5 d [133].

1-Bromo-2-phenylethane. Phenylethylbromide has been reported to be metabolized by rats and rabbits to metabolites that are common to styrene and styrene oxide, and that are ultimately excreted in the urine [136], [137].

1-(Bromomethyl)-3-phenoxybenzene. 3-Phenoxybenzyl bromide is a skin sensitizer [133], [138]; LD_{50} (rats, oral): 5 g/kg male, and 2.3 g/kg female [133].

Bromoacetic Acid. Bromoacetic acid has an LD_{50} (oral, mice) of 100 mg/kg [139]; is irritating and corrosive to skin and mucous membranes [140], [141]. Biological in vivo or in vitro studies are based upon bromoacetate.

Bromoacetate administered as sodium bromoacetate to rabbits causes retinal degeneration, [142–144] and inhibits human erythrocyte carbonic anhydrase B [145].

Neuromuscular blockage by bromoacetate has been demonstrated [146]. Inhibition of neuroblastoma in A/J mice following intratumor injection has been reported. A mechanism involving irreversible binding of bromoacetate to cholinergic receptors on neuroblastoma cell membrane and/or to cholinesterase was proposed [147]. The antitumor action may be based on alkylation of DNA, not inhibition of cholinergic receptors [148].

The carcinogenic potential of chloroacetic acid has been studied in a 2-year (1982 – 1984) chronic gavage bioassay [149].

Bromomethane. Acute inhalation exposure causes pulmonary edema, convulsions, hyperthermia, and coma; chronic exposure can lead to neuropathies and behavioral disturbances. Significant changes occurred in rat brain monoamines following inhalation exposure to less than 15 ppm [150].

In addition to skin irritation and blistering, dermal exposure has been associated with systemic effects typical of inhalation exposure [151].

Methyl bromide has been reported to be mutagenic in a number of systems, but nonteratogenic [151]. It has been evaluated for carcinogenic potential by inhalation [149].

For further background, an excellent review article is recommended [151]. TLVs and MAKs are shown in Table 4. Ethyl Corp. has set an internal standard of 1 ppm, 8-h TWA [133].

3-Bromo-1-propene. Allyl bromide is extremely irritating and painful to the eyes, lungs, mucosal membranes, and skin [152].

Direct mutagenicity in the absence of activating enzymes has been demonstrated; the direct mutagenicity (allyl iodide > allyl bromide > allyl chloride) either decreased or was totally lost following addition of microsomal enzyme S-9 supernatant [153], [154]. Also in

vitro alkylation of DNA by allyl bromide was shown [155], [156].

Dibromomethane. The toxicity of dibromomethane has been summarized as follows [157]: more toxic than either methylene chloride or methylene chlorobromide, but less toxic than tribromomethane; can produce significant liver and kidney damage in animals; may cause cardiac arrhythmias; LD_{50} (rats, oral): >1 000 mg/kg; slightly irritating to the eyes and skin of rabbits and does not appear to be absorbed significantly even when applied repeatedly; vapors are anesthetic; disorder occurs in the protein-prothrombin and glycogenesis functions of the liver and in filtration capacity of the kidneys in animals; metabolized to carbon monoxide. Mutagenicity (Ames) has been demonstrated [158]; it has been reported as the most mutagenic among dihalomethanes [159].

Rats were observed to have elevated blood levels of carboxyhemoglobin following [^{13}C]-dichloromethane administration. It was later demonstrated in vitro that dihalomethanes were metabolized to carbon monoxide by a microsomal mixed function oxidase system [160], [161]. The reported rate and amount of carbon monoxide produced in vivo was as follows: dibromide > dichloride > chloride – bromide > diiodide [162]. The rate of metabolism by isolated rat hepatocytes was as follows: diiodide > dibromide > chloride – bromine = dichloride [163]. Further studies have shown that trihalomethanes are also metabolized to carbon monoxide [164]. The P450 pathway yields a reactive formyl halide intermediate which may either acylate tissue nucleophiles or decompose to carbon monoxide, hydrogen ion, and inorganic halide [164–167].

Dihalomethanes are also metabolized to formaldehyde, formic acid, and inorganic halide. Dihalomethanes were metabolized by rat liver homogenates to formaldehyde and organic halide [168]. An enzymatic pathway yielding a highly reactive intermediate capable of either alkylating tissue nucleophiles or undergoing metabolism to inorganic halide, formaldehyde, and formic acid was proposed [164], [166], [169]. Elevated urinary formic acid levels have been reported in workers exposed to high levels of dichloromethane [166]. The mutagenic activity of dibromo- and diiodomethane, involving an electrophilic intermediate, is mediated by microsomal enzymes and glutathione-S transferase [170].

Bromochloromethane. The most significant toxicological effects of bromochloromethane, as well as haloalkanes in general, are central nervous system depression [171] and impaired cardiac performance [172], [173], arrhythmias due to sensitization to epinephrine [174], [175], and decreased arterial blood pressure [176]. Two distinct activation in vivo pathways have been observed: a high-affinity, saturable oxidation reaction involving microsomal enzymes, and a much lower affinity pathway involving glutathione-S transferase [177]. The rate and amount of carbon monoxide metabolized is less than that following exposure to dibromomethane, but greater than dichloromethane [160]. Studies have also demonstrated adverse effects upon the liver, kidney, and adrenal cortex [178–180]; it is irritating to the eyes and skin [181]; LD_{50} (rats, oral), 5000 mg/kg [182]. Inhalation toxicity was reported to be greater than either dichloromethane or carbon tetrachloride [183]. The TLVs and MAKs are given in Table 4.

Trifluorobromomethane. Trifluorobromomethane causes central nervous system and respiratory depression [184], [185]; cardiac arrhythmias due to sensitization to epinephrine [174], [175]; reversible hypotension due to ganglionic blockage [176]; and impaired cardiac performance [172], [173].

The compound has been observed to inhibit acetylcholinesterase activity [186], [187]. For TLVs and MAKs, see Table 4.

Difluorobromomethane causes central nervous system depression [171], [188], [189], cardiac sensitization to epinephrine induced arrythymias, hypotension, and reduced myocardial performance [172], [173], [188], [189], and liver damage [190], [191].

Tetrabromobisphenol-A (TBBPA). Tetrabromobisphenol-A has a low order of acute toxicity [LD_{50} (rat, oral) > 5 g/kg, LD_{50} (rabbit, dermal) > 2 g/kg]; it is nonirritating to the eyes and skin (Draize scores of 0/24 h – 7 d and 0/24 h – 72 h, respectively); nonsensitizer; non-

mutagenic (Ames test); and nonbromacnegenic [133].

The uncoupling of oxidative phosphorylation by TBBPA in isolated rat liver mitochondria has been demonstrated; oxygen uptake decrease was also observed with the same concentration of biphenyl or biphenyl ether [192]. Adverse effects upon mitochondrial energy transfer reactions by polychlorinated biphenyls have been observed [193]. Also, TBBPA induced liver microsomal enzyme systems [194]. Ethyl Corp. has adopted a TLV of 2 mg/m^3, 8-h TWA [133].

Octabromobiphenyl Oxide (OBBPO). Preliminary studies have indicated that OBBPO causes developmental toxicity when administered as a corn oil suspension via gavage to pregnant rats on days 5 through 16 of gestation.

Administration of 26 mg/kg per day of the mixture, ranging from 6 bromine atoms to 9 bromine atoms per molecule, resulted in significant effects on the conceptuses including reduced average fetal body weights, increased embryo and fetal death (resorption), fetal malformation, and delayed skeletal ossification. The embryo and fetal no-effect level was defined at 2.5 mg/kg per day.

Decabromobiphenyl Oxide (DBBPO). Decabromobiphenyl oxide has also been shown to have a low order of acute toxicity [LD_{50} (rat, oral) and (rabbit, dermal) > 5 g/kg] [195]; nonirritating to the eyes [195], [196], but may cause transient irritation to the conjunctival membrane [196]; nonsensitizer to humans [197]; nonbromacnegenic [196]; nonteratogenic to rats [195], [196], [198]; no adverse reproductive effects in rats [198]; neither mutagenic [195] nor cytotoxic [198]; readily photodegraded by ultraviolet light [196]; and nonbioconcentrating in fish [196].

The principal target organs which may be adversely affected by DBBPO include the thyroid, liver, and kidney. Thyroid hyperplasia has been observed in rats following a 30-d feeding study at 80 and 800 mg/kg per day [196], [199]; no evidence of thyroid alterations was observed following a 2-year rat feeding study at 0.01 – 1.0 mg/kg per day [200].

Bromine concentration in the liver and adipose tissue of rats increased during a one-year feeding study at 0.1 mg/kg per day [198]; liver and kidney lesions were observed following a 30-d feeding study at dose levels > 80 mg/kg per day [196], [199]; liver alterations were observed following a rat inhalation study at 0.05 mg/L of air [195].

Liver enlargement in animals has been reported to occur with all brominated diphenyl ethers that are capable of inducing xenobiotic metabolism [201]. Neither DBBPO nor DBB (decabromobiphenyl) showed any porphyrinogenic potential [197]. There is evidence that porphyrinogenic effects of halogenated aromatic hydrocarbons are mediated through reactive intermediates, and apparently DBBPO is very resistant to metabolic degradation [202].

DBBPO was neither toxic nor carcinogenic in rats fed 0.01 – 1.0 mg/kg per day for 2 years [200]. DBBPO carcinogenicity potential is presently under evaluation [149].

Tetrabromophthalic Anhydride (TBPA) has a low order of acute toxicity [LD_{50} (rat, oral) and LD_{50} (rabbit, dermal) > 10 g/kg]; may cause moderate eye irritation (Draize score of 12/4/24 h); it is nonirritating to the skin, but is a potential sensitizer [133].

Neither TBPA [133] nor its hydrolysis product, tetrabromophthalic acid [203], are mutagenic (Ames).

Hexabromocyclododecane (HBCD) has a low order of acute toxicity: LD_{50} (rat, oral) >10 g/kg, LD_{50} (rabbit, dermal) >8 g/kg; LC_{50} (rat, 1 h): > 200 mg/L; nonirritating to eyes or skin (Draize scores of 0); nonmutagenic (Ames) [131].

2. Inorganic Bromine Compounds

The most important commercial inorganic bromine compounds are hydrogen bromide, hydrobromic acid, and the alkali bromides and bromates. The newest technology in the area is the utilization of aqueous inorganic bromide solutions (clear brines) by the petroleum industry as completion and workover fluids. Thousands of tons of zinc, calcium, and sodium bromide are used by the industry to help increase production rates and to extend well production life cycles [204].

2.1. Hydrogen Bromide and Bromides

2.1.1. Hydrogen Bromide

Hydrogen bromide [*10035-10-6*], HBr, M_r 80.9, is a colorless, corrosive, pungent gas that fumes on exposure to moist air and is extremely irritating to the eyes, nose, and throat. Physical properties of the anhydrous gas include:

mp	−89.6 °C
bp at 101.3 kPa	−66.8 °C
$t_{crit.}$	89.8 °C
$p_{crit.}$	8.509 MPa (85.1 bar)
heat of fusion at *mp*	31.1 kJ/kg
heat of vaporization at *bp*	215 kJ/kg
specific gravity (air = 1.00)	2.71

Hydrogen bromide is soluble in water, ethyl alcohol, other polar organic solvents, and acetic acid.

The anhydrous gas is produced industrially by burning bromine in the presence of excess hydrogen at 500 °C [205] or by reacting it with hydrogen at 375 °C over platinized silica gel or platinized asbestos catalyst. Generation of hydrogen bromide as a byproduct during bromination of aromatic compounds is another major source of the industrial gas. In either case, entrained bromine can be removed by scrubbing with phosphorous acid or by activated carbon [206]. The gas is liquefied by compression for storage and shipment in cylinders. Methods for laboratory preparation include reaction of bromine with tetrahydronaphthalene or with sulfur and water.

Anhydrous hydrogen bromide is commonly available in 99.8 % purity in cylinders containing 4.5, 13.6, or 59.1 kg at 2.2 MPa (22 bar) at 21.1 °C. The "Non-Flammable Gas" label is required for shipment.

Hydrogen bromide is utilized in the production of inorganic bromides by reaction with the appropriate aqueous solution or slurry of metal hydroxides, oxides, or carbonates; in the production of organic bromides by reaction with alkyl alcohols or alkenes; and as a catalyst for oxidations, alkylations, and condensations in organic chemistry.

In contact with metals in the presence of moisture, hydrogen bromide can produce hydrogen, which may form explosive mixtures with air. The acute inhalation LC_{50} for rats is 9527 mg/m^3 (2858 ppm). The TLV is 10 mg/m^3 (3 ppm); the MAK value is 20 mg/m^3; 16.7 mg/m^3 (5 ppm) is irritating to humans.

Hydrobromic acid is the aqueous solution of hydrogen bromide. The aqueous acid is commercially available as colorless 48 and 62 % solutions; they turn yellow on storage as the result of bromine formation caused by air oxidation or exposure to light. The "Corrosive" label is required for transportation by truck or rail.

Constant-boiling hydrobromic acid distills at 124.3 °C at atmospheric pressure and contains 47.6 wt % hydrogen bromide. The boiling point and hydrogen bromide concentration can be partially controlled by varying the pressure during distillation. These solutions are produced industrially by dissolution of hydrogen bromide in water. The densities of aqueous hydrobromic acid are shown in Table 5 (see below). A convenient laboratory preparation consists of distillation from a mixture of sodium or potassium bromide in dilute sulfuric acid. Hydrobromic acid is used industrially in the production of inorganic bromides, alkyl bromides, salts of organic nitrogen compounds, and as an alkylation catalyst.

The highly ionized aqueous solution can cause severe burns to the eyes, skin, and mucous membranes. Immediate removal by inundation with large amounts of water and subsequent medical attention are the recommended responses to accidental exposure.

2.1.2. Bromides

Metal bromides are routinely produced by reaction of hydrogen bromide or hydrobromic acid with metals, their hydroxides, carbonates, or oxides. Nonmetal bromides are produced by direct

Table 5. Densities of aqueous hydrobromic acid

c_{HBr}, wt %	10	20	30	40	50	60	65
d_4^{20}	1.0723	1.1579	1.2580	1.3772	1.5173	1.6787	1.7675

reaction with hydrogen bromide or bromine at elevated temperature.

The alkali-metal bromides, particularly those of sodium, potassium, and lithium, are important industrial chemicals. Manufacture of each of these compounds is by reaction of the hydroxide or carbonate with hydrogen bromide or with bromine and a reducing agent, such as ammonia, formic acid, or carbon [207–209]. Concentration, filtration, and drying produces crystalline materials, whereas spray-drying results in a powder. Calcium and zinc bromides are the major non-alkali-metal bromides used industrially, although several others are used selectively.

Sodium bromide [*7647-15-6*], NaBr, M_r 102.9, d_4^{26} 3.20, *mp* 747 °C, *bp* 1390 °C, is a bitter, salty-tasting hygroscopic white solid. For its solubility in water, see Table 6 (see below). Sodium bromide can be isolated as the anhydrous compound from aqueous solution above 50.7 °C. The solid product is available in 25-kg plastic-lined bags, and a 45 % solution is available from 55-gallon drums (208 L) up to tank car quantities.

Sodium bromide is utilized for the preparation of light-sensitive silver bromide for photographic applications. Another use is as a bromine – hypobromous acid precursor for treatment of swimming pools, hot tubs, and industrial water. The bioactive agent is released by reaction of the bromide with an appropriate oxidant, such as chlorine. Sodium bromide solutions are also utilized by the petroleum industry as workover and completion fluids. They are particularly useful in geological formations containing high levels of carbonate – bicarbonate, which would form insoluble solids upon exposure to calcium halide solutions. Pharmaceutical-grade sodium bromide is prescribed for its sedative properties.

Sodium bromide solutions and the solid sodium bromide are considered moderate eye irritants, but nonirritating to the skin. The acute oral LD_{50} in rats is greater than 5 000 mg/kg and the acute dermal LD_{50} in rabbits is greater than 2000 mg/kg.

Potassium bromide [*7758-02-3*], KBr, M_r 119.0, d_4^{25} 2.75, *mp* 730 °C, *bp* 1435 °C, is a white, slightly hygroscopic crystal or powder with a bitter salty taste. For solubility in water, see Table 6. Potassium bromide is utilized for the production of photographic-grade silver bromide. It also finds use in process engraving, lithography, and in medicine as a sedative. It is considered moderately toxic and no special handling techniques are required.

Lithium bromide [*7550-35-8*], LiBr, M_r 86.9, d_4^{25} 3.46, *mp* 550 °C, *bp* 1265 °C, is used extensively as a 55 % solution for water absorption in air conditioning applications, in the lithium – bromine battery, and in medicine as a sedative. For its water solubility, see Table 6.

Exposure to lithium bromide may result in mild eye and skin irritation, whereas ingestion of large doses may cause central nervous system depression.

Ammonium bromide [*12124-97-9*], NH_4Br, M_r 98.0, ϱ 2.43 g/cm^3, sublimes at 452 °C (*mp* 542 °C in a sealed capillary); for its solubility in water, see Table 6. Ammonium bromide is a hygroscopic white to yellow-white crystal or powder that is freely soluble in alcohol and acetone. It is made by the reaction of hydrobromic acid and ammonia or preferably by the direct reaction of ammonia with bromine in water:

$$8\,NH_3 + 3\,Br_2 \longrightarrow 6\,NH_4Br + N_2$$

For safety, this reaction is run with ammonia in slight excess. Concentration of solutions produces the crystalline product, whereas a powder results from spray-drying. A 45 % solution is also available commercially.

Table 6. Solubilities of NaBr, KBr, LiBr, and NH_4Br in water (in g/100 g H_2O) at different temperatures (in °C)

Compound	−20	−10	0	20	40	60	80	100	120
NaBr	71.8		79.5	90.5	106	118	118.3	121.2	125.3
KBr		48.8	54.0	65.6	76.1	85.9	95.3	105	115
LiBr			131	151.3			210.5	227.9	
NH_4Br			59.5	74.2	89.8	105	117	135	

Ammonium bromide is used by the photographic industry, in process engraving, in textile finishing, as a flame retardant for wood, and in medicine as a sedative. It is a mild eye and skin irritant with moderate human toxicity.

Calcium bromide [*7789-41-5*], $CaBr_2$, M_r 199.9, d_4^{25} 3.35, *mp* 730 °C (some decomposition), *bp* 810 °C, solubility 142 g/100 g water, is a white, hygroscopic, odorless, salty-tasting crystal or powder. The hexahydrate, $CaBr_2 \cdot 6\,H_2O$, M_r 295.9, d_4^{25} 2.295, *mp* 38 °C, *bp* 149 °C (decomp.), is also available commercially.

Common manufacturing techniques include reaction of calcium hydroxide, oxide, or carbonate with hydrobromic acid or bromine and a reducing agent (formic acid [208], formaldehyde [210]). Vacuum concentration, filtration, and drying results in a crystalline product, whereas spray-drying produces a powder.

The major use for calcium bromide is as a component of high-density clear drilling, completion, workover, and packing fluids [211], [212]. By reducing formation damage, the use of clear brines results in higher production rates and longer production life. Calcium bromide is used in oil fields as a 95 % minimum purity solid in 25-kg plastic-lined bags or as a 53 % aqueous solution of specific gravity 1.70 – 1.71 at 15.5 °C. Calcium bromide finds other applications in photography, in medicine, in sizing compounds, as a wood preservative, and in flame retardant blends.

The oral LD_{50} for male and female rats is 2447 mg/kg; the solid and solution are recognized as primary eye irritants.

Zinc bromide [*7699-45-8*], $ZnBr_2$, M_r 225.21, d_4^{25} 4.20, *mp* 394 °C, *bp* 650 °C, solubility 447 g/100 g water at 20 °C, is an extremely hygroscopic, granular powder with a metallic taste. Aqueous solutions are highly acidic (pH 1 – 4). Manufacture is by reaction of zinc oxide with hydrobromic acid or with bromine and a reducing agent [209], [210]. Reaction of bromine water with zinc metal is an acceptable alternative but less desirable procedure.

A major new use for zinc bromide is as a completion and workover fluid for oil field application [213]. The normal article of commerce to the petroleum industry is a 2 301-kg/m^3 (19.2-lb/gal) solution consisting of 55 % zinc bromide and 20 % calcium bromide. A potential major new use for 77 % zinc bromide solution is in the zinc – bromine storage battery, which is being considered for use in utility load leveling [214] and as a power source for electric vehicles [215]. Other uses include photography and rayon manufacture; an 80 % solution of highly pure zinc bromide is used in nuclear reactor viewing windows.

Zinc bromide is not a primary skin irritant but is a severe eye irritant. The oral LD_{50}, combined male and female rats, is 1047 mg/kg. No hazard label is required for truck and rail shipments of solid zinc bromide; however, the acidic zinc bromide solutions require the "Corrosive" label.

Aluminum bromide [*7727-15-3*], $AlBr_3$, M_r 266.71, ϱ 3.01 g/cm^3, *mp* 97.5 °C, *bp* 263 °C (99.6 kPa), reacts vigorously with water. Aluminum bromide is a white to yellow hygroscopic solid which must be stored under moisture-free conditions. Manufacture is by direct reaction of aluminum and bromine or aluminum hydroxide and hydrobromic acid. Aluminum bromide is a very reactive catalyst for Friedel-Crafts reactions, olefin polymerizations, hydrocarbon isomerization, and olefin alkylation of aromatic compounds. It is commonly used as a catalyst in combination with aluminum chloride.

Aluminum bromide is extremely corrosive to human tissues and should be handled accordingly.

Copper(I) bromide [*7787-70-4*], CuBr (Cu_2Br_2), M_r 143.45, ϱ 4.98 g/cm^3, *mp* 492 °C, *bp* 1345 °C, is slightly soluble in water. Copper(I) bromide is prepared by addition of sodium bromide and sulfur dioxide to a solution of copper(I) sulfate. It is used for the preparation of aromatic bromine compounds via the Sandmeyer reaction and has limited use as a catalyst in polymer chemistry.

Phosphorus tribromide [*7789-60-8*], M_r 270.70, d_4^{15} 2.85, *mp* −40 °C, *bp* 172.9 °C, is produced by the reaction of bromine with yellow phosphorus in PBr_3 at 130 °C [216]. Reaction of organic alcohols with phosphorus tribromide results in the production of alkyl bromides. The

tribromide also serves as a catalyst in certain organic reactions.

Phosphorus tribromide is corrosive to the skin via hydrolysis to hydrogen bromide. A "Corrosive" label is required for rail and air shipment.

Bromine trifluoride [*7787-71-5*], BrF_3, M_r 136.9, d_4^{25} 2.80, *mp* 8.8 °C, *bp* 135 °C, is a colorless liquid produced by mixing fluorine with bromine. It reacts vigorously with water, organic compounds, and metals. The high specific conductivity of $8 \times 10^{-3}\,\Omega^{-1}\,cm^{-1}$ is due to dissociation:

$$2\,BrF_3 \longrightarrow BrF_2^+ + BrF_4^-$$

Bromine trifluoride is a strong fluorinating agent, which can be used in organic synthesis and in formation of inorganic fluorides.

Bromine trifluoride is highly toxic, corrosive to all tissues, and reacts violently with water. Both "Oxidizer" and "Poison" labels are required for rail shipment.

Bromine pentafluoride [*7789-30-2*], BrF_5, M_r 174.9, d_4^{25} 2.47, *mp* −61 °C, *bp* 40.8 °C, is a fuming, colorless liquid produced by reacting bromine and fluorine in an iron or copper vessel at 200 °C. This material is used as a fluorinating agent and as the oxidizer component of some rocket propellants. Highly toxic and corrosive to the skin, bromine pentafluoride requires the "Oxidizer" label for rail shipment and the "Corrosive" label for air shipment. The TLV is 0.1 ppm (0.7 mg/m^3), with an STEL of 0.3 ppm (2 mg/m^3).

Bromine chloride [*13863-41-7*], BrCl, M_r 115.4, *fp* −66 °C, *bp* −5 °C, is an orange-yellow liquid prepared by mixing bromine and chlorine in equimolar quantities. The reaction is exothermic and provision should be made for heat removal during production. At room temperature about 60 % of the mixed halogen is present as bromine chloride. Bromine chloride is used in organic synthesis both for addition across olefinic bonds, producing the bromochloro compound, and for aromatic bromination, whereby the aromatic bromide and hydrogen chloride are produced. It also finds use in water treatment.

This toxic and irritating material is a strong oxidizing agent and requires the "Corrosive" label for shipment by rail.

2.2. Hypobromous Acid, Hypobromites

Hypobromous acid [*13517-11-8*], HBrO, is somewhat unstable, existing as a dilute solution at ambient temperatures. Aqueous solutions can be prepared by reacting yellow mercuric oxide with bromine water [217]:

$$HgO + 2\,Br_2 + H_2O \longrightarrow 2\,HOBr + HgBr_2$$

Hypobromous acid, generated in situ, is a strong bactericide and water disinfectant. Alkali-metal hypobromites are normally prepared by reacting bromine with alkali hydroxide in aqueous solution below 0 °C:

$$2\,NaOH + Br_2 \longrightarrow NaBrO + NaBr + H_2O$$

At higher temperature, hypobromite disproportionates to bromate and bromide:

$$3\,NaBrO \longrightarrow 2\,NaBr + NaBrO_3$$

Hypobromites are used by the textile industry as bleaching and desizing agents.

2.3. Bromous Acid, Bromites

Bromous acid [*7486-26-2*], HOBrO, can be produced in low concentration by the treatment of a silver nitrate solution with excess bromine or by bromine oxidation in a diaphragm cell. Sodium bromite [*7486-26-2*], $NaBrO_2$, has been isolated as 99.6 % pure $NaBrO_2 \cdot 3\,H_2O$. The compound is stable and can be stored for extended periods of time under normal conditions. It is made from a concentrated hypobromite solution at pH 11 – 12 and < 0 °C:

$$2\,NaBrO \longrightarrow NaBrO_2 + NaBr$$

The bromine reacts in solution with remaining hypobromite forming sodium bromide and sodium bromate until an equilibrium is established. Ammonia is added to decompose the remaining hypobromite,

$$3\,NaBrO + 2\,NH_3 \longrightarrow 3\,NaBr + 3\,H_2O + N_2$$

leaving a solution from which yellow sodium bromite trihydrate can be isolated by concentration and filtration. Barium, calcium, and lithium bromites can be prepared similarly [218]. Bromites are used industrially as desizing agents for textiles and in water treatment.

Table 7. Solubilities of $NaBrO_3$, $KBrO_3$, and $Ba(BrO_3)_2$ in water (in g/100 g H_2O) at different temperatures (in °C)

Compound	0	20	40	60	80	100
$NaBrO_3$	27.5	36.4	48.8	62.6	75.8	90.8
$KBrO_3$	3.1	6.9	13.1	22.0	33.9	49.7
$Ba(BrO_3)_2$	0.287	0.657	1.33	2.32	3.65	5.71

2.4. Bromic Acid, Bromates

Bromic acid [*7789-31-3*], $HBrO_3$, has been prepared by reaction of alkali bromates with dilute sulfuric acid and by electrolysis of bromine water.

The industrially important salts of bromic acid, sodium and potassium bromate, are commonly produced by electrolytic oxidation of the appropriate aqueous bromide in the presence of dichromate at 65 – 70 °C [219]. The cathode may be stainless steel or copper and the anode is commonly lead dioxide or iron.

Another production method consists of reacting the aqueous alkali hydroxides or carbonates with bromine at elevated temperature:

$$6\,NaOH + 3\,Br_2 \longrightarrow NaBrO_3 + 5\,NaBr + 3\,H_2O$$

The relatively high solubility of the bromides allows the bromates to be recovered from the reaction solutions by cooling and filtration.

Sodium bromate [*7789-38-0*], $NaBrO_3$, M_r 150.9, d_4^{17} 3.34, *mp* 381 °C (decomp., oxygen evolution), is a colorless crystal or powder; for its solubility in water, see Table 7. Commercially, the most important material is 99.5 % minimum assay available in 50- or 113.6-kg polyethylene-lined drums. This strong oxidizing agent ("Oxidizer" label required) can cause fires upon contacting organic materials and will release bromine upon acidification or in a fire situation. Considered a skin and eye irritant, this material exhibits an oral LD_{50} in rabbits of 250 mg/kg. Its primary use is in hair treatment formulations. It has also been used in mixtures with sodium bromide in mining applications.

Potassium bromate [*7758-01-2*], $KBrO_3$, white crystals or powder, M_r 167.0, $d_4^{17.5}$ 3.27, decomposes at ~ 370 °C; for its solubility in water, see Table 7. Available as pure (≧ 99.5 %) and reagent (≧ 99.8 %) grade, its major uses are in flour treatment, in hair wave solutions, and as an analytic standard.

Barium bromate [*10326-26-8*], $Ba(BrO_3)_2 \cdot H_2O$, M_r411.2, exists as white crystals, ϱ 4.0 g/cm^3, which lose their water of crystallization upon heating to 180 – 200 °C, with decomposition occurring at ≈ 270 °C; solubility in water is given in Table 7. In addition to the reaction of bromine and barium hydroxide, barium bromate may be produced by reaction of a soluble barium salt with an alkali bromide. The insoluble barium bromate may be removed by filtration. It has limited use as a corrosion inhibitor.

3. References

1. Z. E. Jolles (ed.): *Bromine and its Compounds,* Academic Press, New York 1966.
2. S. Patai (ed.): *The Chemistry of the Carbon – Halogen Bond,* J. Wiley & Sons, New York 1973, Part 1 & 2.
3. *Houben-Weyl* **5/4**.
4. W. K. R. Musgrave, vol. 1 A, 2nd ed., p. 478; W. J. Feast, W. K. R. Musgrave, vol. 3 A, p. 241, both in S. Coffey (ed.): *Rodds Chemistry of Carbon Compounds,* Elsevier, Amsterdam 1964, 1971, respectively.
5. W. J. Feast, vol. 1 AB, 2nd ed. supplement, p. 31; vol. 3A, 2nd ed. supplement, p. 141, in M. F. Ansell (ed.): *Rodds Chemistry of Carbon Compounds,* Elsevier, Amsterdam 1975, 1983, respectively.
6. R. D. Chambers, S. R. James in D. Barton, W. D. Ollis (eds.): *Comprehensive Organic Chemistry,* **vol. 1,** Pergamon, Oxford 1979, p. 493.
7. A. L. Horvath: *Halogenated Hydrocarbons: Solubility-Miscibility with Water,* Dekker, New York 1982.
8. R. R. Dreisbach: *Physical Properties of Chemical Compounds,* **vol. 1 – 3,** Advances in Chemistry Series, No. 15, 22, and 29, Amer. Chem. Soc., Washington, D.C., 1955, 1959, 1961.

9. E. F. Mooney, M. Goldstein, [1], in p. 802
10. J. Zabicky, S. Ehrlich-Rogozinski, A. G. Loudon, [2], p. 63, p. 223.
11. Associated Ethyl Co., GB 804995, 1958 (W. J. Read, D. B. Clapp, D. R. Stephens, J. E. Russell).
12. Dead Sea Bromine Co. Ltd., US 2921967, 1960 (F. Yaron).
13. Dow Chemical, US 2746999, 1956 (A. A. Gunkler, D. E. Lake, B. C. Potts).
14. Procter & Gamble Co., US 3336403, 1967 (A. Kessler).
15. Procter & Gamble Co., US 3396204, 1968 (C. B. McCarty, K. W. Thiele).
16. Procter & Gamble Co., US 3321538, 1967 (K. W. Thiele, G. R. Wyness, C. B. McCarty).
17. Gulf Research & Development Co., US 3699179, 1972 (J. P. Boyle, C. R. Murphy, W. W. Walsh).
18. Distillers Co., Ltd., US 3108141, 1963 (E. J. Gasson, D. J. Badley).
19. Great Lakes Chemical Co., US 3655788, 1972 (R. M. Thomas, F. R. Gerns).
20. Distillers Co., Ltd., GB 843234, 1960 (E. J. Gasson).
21. Continental Oil Co., US 3812212, 1974 (R. D. Gordon).
22. Shell Oil Co., CA 749982, 1967 (L. C. Fetterly, K. F. Koetitz).
23. Chevron Research Co., US 3519695, 1970 (S. Suzuki).
24. Thomas A. Edison, Inc., US 2412882, 1946 (C. B. Gardenier).
25. Shell International Research, NL-A 6410368, 1965.
26. Esso Research & Engineering Co., US 3847985, 1974 (J. Linder, L. L. Maravetz, G. M. Schmit).
27. Société Nationale des Pétroles d'Aquitaine, DE-OS 2530675, 1976 (M. Gellato, J. L. Seris, J. Suberlucq).
28. Dow Chemical, US 2935535, 1960 (A. A. Asadorian).
29. Du Pont, US 3321539, 1967 (F. J. Plesmid).
30. M.C.P. Chemical Processes Ltd., US 4228106, 1980 (M. Martan).
31. Dow Chemical, DE 1240059, 1967 (C. E. Grabiel, J. L. Russell, H. Volk, B. W. Wilkinson et al.).
32. Chemische Fabrik Kalk GmbH, GB 1103650, 1968.
33. Ethyl Corp., US 4326089, 1982 (K. A. Keblys).
34. Sumitomo, US 4104940, 1979 (Y. Ume, T. Matsuo, N. Itaya).
35. Shell Oil Co., US 4010087, 1977 (D. A. Wood, R. F. Mason).
36. Toyo Soda Manufacturing Co., Ltd., EP-A 19804, 1980 (K. Katsuragawa, H. Sakka, K. Kihara).
37. Roussel UCLAF, DE-OS 2810305, 1978 (J. Warnant, J. Jolly).
38. Chemische Fabrik Kalk GmbH, GB 1210475, 1970 (H. Jenkner, H. Nestler).
39. Zaklady Chemiczne "Organica-Azot", PL 99987, 1975 (J. Swietoslawski, A. Ratajczak). *Chem. Abstr.* **91** (1979) 91182 d.
40. Dow Chemical, US 2553518, 1951 (D. E. Lake, A. A. Asadorian).
41. Katayama Kagaku Kogyo Kenkyusho Co., Ltd., JP-Kokai 78105419, 1978 (O. Umekawa, K. Takemori, S. Katayama).
42. Katayama Kagaku Kogyo Kenkyusho Co., Ltd.,DE-OS 2640943, 1977 (O. Umekawa, K. Takemori, S. Katayama).
43. Chemische Fabrik Kalk GmbH, GB 1381919, 1975 (H. Jenkner, R. Karsten).
44. Degussa, GB 768893, 1957.
45. Degussa, US 2717911, 1955 (L. Huter, H. Veith).
46. Nippon Kayaku Co., JP-Kokai 74108003, 1974 (K. Matsuda, M. Sugino, S. Kaji).
47. Dow Chemical, US 3182088, 1965 (H. E. Hennis).
48. Whiffen & Sons Ltd., GB 836653, 1960 (P. F. Pasco, P. R. E. Lewkowitsch).
49. Ethyl Corp., unpublished.
50. Chemische Fabrik Kalk GmbH, DE-OS 2204194, 1973 (H. Jenkner, W. Buettgens).
51. I.G. Farbenind. AG, DE 727690, 1942 (O. Scherer, F. Dostal).
52. Solvay et Cie., DE 1283214, 1968 (U. Giacopelli, M. Manca).
53. Società Chimica Dell'Aniene SpA, GB 874062, 1959.
54. Shell International Research, FR 1441233, 1966.
55. Ethyl Corp., US 3923914, 1975 (P. Kobetz, K. L. Lindsay).
56. Du Pont, US 2875254, 1959 (F. J. Gradishar).
57. Hoechst, DE-OS 3046330, 1982 (H. Block, J. Mintzer, J. Wittman, J. Russow).
58. Central Glass Co., Ltd., JP Kokai 7976504, 1979 (N. Kouketsu, F. Inoue, T. Komatsu, K. Matsuoka).
59. Onoda Cement Co., Ltd., JP Kokai 7762208, 1977 (Y. Ikubo, K. Kunihiro).
60. Kali-Chemie AG, DE-OS 1913405, 1970 (H. Paucksch).

61. Kali-Chemie AG, BE 856233, 1977 (H. Boehn, K. H. Hellburg).
62. Hoechst, BE 631933, 1963.
63. Kali-Chemie AG, DE-OS 1946509, 1974 (H. Paucksch, J. Massonne).
64. Bayer, DE 1151811, 1964 (J. Nentwig).
65. Chemische Fabrik Kalk GmbH, FR 1412959, 1964.
66. Dow Chemical, US 3546302, 1970 (A. A. Asadorian, R. G. Tigner).
67. Asahi Glass Co., Ltd., JP-Kokai 7654538, 1976 (M. Ichimura, T. Nishiyama, K. Suzuki).
68. Chemische Fabrik Kalk GmbH, DE 1768444, 1972 (H. Jenkner, O. Rabe).
69. Dow Chemical, US 3363007, 1968 (T. E. Majewski, L. R. Collins).
70. Ethyl Corp., US 4013728, 1977 (D. R. Brackenridge).
71. Chemische Fabrik Kalk GmbH, DE-OS 2511981, 1976 (H. Jenkner).
72. Degussa, DE-OS 2227439, 1974 (W. Weigert, H. Hein, H. Mechler, E. Meyer-Simon et al.).
73. Produits Chimiques Ugine Kuhlmann, FR 2274586, 1976 (A. Isard, D. Pellet).
74. Chemische Fabrik Kalk GmbH, DE-OS 2613969, 1977 (H. Jenkner, R. Strang).
75. Zaklady Chemiczne "Organica-Azot", US 4112242, 1978 (J. Swietoslawski, A. Silowiecki, A. Ratajczak, B. Nocon et al.).
76. Chemische Fabrik Kalk GmbH, DE 1129957, 1962 (H. Hahn).
77. Società Italiana Resine SpA, DE-OS 2162859, 1972 (F. Montanari, B. Calcagno, L. Conti).
78. Dow Chemical, US 3029291, 1962 (A. J. Dietzler).
79. I.S.C. Chemical Ltd., GB 1472383, 1977 (L. J. Belf).
80. Produits Chimiques Ugine Kuhlmann, DE OS 2400455, 1975 (D. Balda, D. Pitiot).
81. Michigan Chemical Corp., US 3965197, 1976 (H. Stepniczka).
82. Great Lakes Chemical Corp., US 4287373, 1981 (J. A. Garman, R. I. Mamuzic, R. B. McDonald, J. L. Sands et al.).
83. Great Lakes Chemical Corp., US 4327227, 1982 (J. T. Ayres, D. L. McAllister, J. L. Sands).
84. Degussa, DE-OS 2250550, 1974 (H. Hein, K. Janzon, W. Weigert, C. Liedtke et al.).
85. Unassigned, DD 14750, 1958 (H. Ohle).
86. Chemische Fabrik Kalk GmbH, US 3194817, 1965 (H. Hahn).
87. Chemische Fabrik Kalk GmbH, US 3382254, 1968 (H. Jenkner, O. Rabe, R. Strang).
88. Cities Service Co., US 3833675, 1974 (J. Newcombe, A. O. Dotson).
89. Toa Gosei Chemical Industry Co., Ltd., JP 7505187, 1975 (T. Kowaguchi, N. Hisanega, H. Naito).
90. Chemische Fabrik Kalk GmbH, US 3558727, 1971 (H. Jenkner, O. Konigstein).
91. ACGIH (ed.): *Threshold Limit Values (TLV),* ACGIH, Cincinnati, Ohio, 1983.
92. DFG (ed.): Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte, Wiley-VCH, Weinheim 2001.
93. *Clin. Toxicol.* **14** (1979) 473 – 478.
94. V. K. Rowe, H. C. Spencer, D. D. McCollister, R. L. Hollingsworth et al., *AMA. Arch. Ind. Hyg. Occup. Med.* **6** (1952) 158 – 173.
95. S. F. Stinson, G. Reznik, J. M. Ward, *Cancer Lett.* **12** (1981) 121 – 129.
96. G. Reznik, S. F. Stinson, J. M. Ward, *Arch. Toxicol.* **46** (1980) 233 – 240.
97. *Vet. Hum. Toxicol.* **22** (1980) 101 – 103.
98. I. Jakobson, J. E. Wahlberg, B. Holmberg, E. Johansson, *Toxicol. Appl. Pharmacol.* **63** (1982) 181 – 187.
99. V. K. Rowe, H. C. Spencer, D. D. McCollister, R. L. Hollingsworth et al., *AMA. Arch. Ind. Hyg. Occup. Med.* **6** (1952) 158 – 173.
100. D. Clive, K. O. Johnson, J. F. S. Spector, A. G. Batson, *Mutat. Res.* **59** (1979) 61 – 108.
101. H. Bren, A. B. Stein, H. S. Rosenkranz, *Cancer Res.* **34** (1974) 2576 – 2579.
102. E. Vogel, J. L. R. Chandler, *Experientia* **30** (1974) 621 – 623.
103. U. Ranning, *Mutat. Res.* **76** (1980) 269 – 295.
104. E.-L. Tan, A. W. Hsie, *Mutat. Res.* **90** (1981) 183 – 191.
105. P. A. Brimer, E.-L. Tan, A. W. Hsie, *Mutat. Res.* **95** (1982) 377 – 388.
106. National Cancer Institute DHHS Publication No. (NIOSH) 81-1766, 1982, 1 – 163.
107. B. L. VanDuuren, B. M. Goldschmidt, G. Loewengart, A. C. Smith et al., *J. Natl. Cancer Inst.* **63** (1979) 1433 – 1439.
108. W. A. Olson, R. T. Habermann, E. K. Weisburger, J. M. Ward et al., *J. Natl. Cancer Inst.* **51** (1973) 1993 – 1995.
109. M. B. Powers, R. W. Voelker, N. P. Page, K. Weisburger, *Toxicol. Appl. Pharmacol.* **33** (1975) 171.
110. E. Nachtomi, D. S. R. Sarma, *Biochem. Pharmacol.* **26** (1977) 1941 – 1945.
111. K. D. Nitschke, R. J. Kociba, D. G. Keyes, J. J. McKenna, *Fund. Appl. Toxicol.* **1** (1981) 437 – 442.

112. A. Bondi, E. Olomucki, M. Claderon, *J. Sci. Food Agric. Abstr.* **6** (1955) 600 – 602.
113. E. Alumot, E. Nachtomi, O. Kempenish-Pinto, E. Mandel et al., *Poult. Sci.* **41** (1968) 1079 – 1985.
114. D. Amir et al., *Anim. Repro. Sci.* **6** (1983) 35 – 50.
115. D. Amir, E. Ben-David, *Ann. Biol. Anim. Biochim. Biophys.* **13** (1973) 165 – 170.
116. R. D. Short, J. L. Minor, J. M. Winston, J. Seifter et al., *Toxicol. Appl. Pharmacol.* **46** (1978) 173 – 182.
117. D. Amir, R. Volcani, *Nature (London)* **206** (1965) 99 – 100.
118. D. Amir, C. Esnault, J. C. Nicolle, M. Courot, *J. Reprod. Fertil.* **51** (1977) 453 – 456.
119. K. Edwards, H. Jackson, A. R. Jones, *Biochem. Pharmacol.* **19** (1970) 1783 – 1789.
120. D. Amir, *J. Reprod. Fertil.* **35** (1973) 519 – 525.
121. A. H. El-Jack, F. Hrudka, *J. Ultrastruct. Res.* **67** (1979) 124 – 134.
122. D. Amir, B. L. Gledhill, D. L. Garner, J. C. Nicholle et al., *Animal Repro. Sci.* **6** (1983) 35 – 50.
123. S. S. Epstein, E. Arnold, J. Andrea, W. Bass et al., *Toxicol. Appl. Pharmacol.* **23** (1972) 228 – 235.
124. R. D. Short, J. M. Winston, C. Hong, J. L. Minor et al., *Toxicol. Appl. Pharmacol.* **49** (1979) 97 – 105.
125. A. Tomasi, E. Albano, M. U. Dianzani, T. F. Slater, *FEBS Lett.* **160** (1983) 191 – 194.
126. L. C. K. Wong, J. M. Hong, C. B. Lee, C. C. Bhandari et al., *Pharmacologist* **20** (1978) 174.
127. L. C. K. Wong, J. M. Winston, C. B. Hong, H. Plotnick, *Toxicol. Appl. Pharmacol.* **63** (1982) 155 – 165.
128. A. M. El-Hawari, *Pharmacologist* **20** (1978) 213.
129. T. W. Shih, D. L. Hill, *Res. Commun. Chem. Pathol. Pharmacol.* **33** (1981) 449 – 461.
130. Federal Register 48, 1983, 49918 – 46248.
131. Federal Register 48, 1983, 45956 – 46003.
132. E. W. VanStee (ed.): *Target Organ Toxicology Series: Cardiovascular Toxicology,* Raven Press, New York 1982, pp. 281 – 326.
133. Ethyl Corp., *Toxicology Reports,* 1979, 1984.
134. A. L. Bridie, C. J. M. Wolff, M. Winter, *Water Res.* **13** (1979) 623 – 626.
135. [134] pp. 627 – 630.
136. S. P. James, D. A. White, *Biochem. J.* **104** (1967) 914 – 921.
137. B. F. Seuher, L. P. C. Delbressine, F. L. M. Smeets, I. A. P. Wagenaars-Zegers, *Xenobiotica* **9** (1979) 311 – 316.
138. K. S. Rao, J. E. Betso, K. J. Olson, *Drug Chem. Toxicol.* **4** (1981) 331 – 351.
139. *J. Pharmacol. Exp. Ther.* **86** (1946) 336.
140. J. Apffel, *Arch. Mal. Prof.* **32** (1971) 603 – 606.
141. *Merck* 9th ed., Merck & Co., Inc., Rahway, N.J., 1976.
142. D. R. Lucas, J. P. Newhouse, J. B. Davery, *Br. J. Ophthalmol.* **41** (1957) 313 – 316.
143. D. R. Lucas, J. P. Newhouse, *Br. J. Ophthalmol.* **43** (1959) 147 – 158.
144. A. Sorsby, J. P. Newhouse, D. R. Lucas, *Br. J. Ophthalmol.* **41** (1957) 309 – 312.
145. P. L. Whitney, *Eur. J. Biochem.* **16** (1970) 126 – 135.
146. B. V. Rama Sastry, C. Y. Chiou, *Toxicol. Appl. Pharmacol.* **17** (1970) 303.
147. C. Y. Chiou, *J. Pharm. Sci.* **67** (1978) 331 – 333.
148. C. E. Stratton, W. E. Ross, S. Chapman, *Biochem. Pharmacol.* **30** (1981) 1497 – 1500.
149. National Toxicology Program Management Status Report, Dec. 6, 1983.
150. T. Honma, A. Sudo, M. Miyagawa, M. Sato et al., *Neurobehav. Toxicol. Teratol.* **4** (1982) 521 – 524.
151. G. V. Alexeeff, W. W. Kilgore, *Residue Rev.* **88** (1983) 101 – 153.
152. R. E. Goddelin, H. C. Hodge, R. P. Smith, M. N. Gleason (eds.): *Clinical Tox. of Commercial Products,* 4th ed., Williams & Wilkins Co., Baltimore, Md., 1976, p. 113.
153. W. Liginsky, A. W. Andrews, *Teratogenesis Carcinog. Mutagen.* **1** (1980) 259 – 267.
154. E. Eder, T. Neudecker, D. Lutz, D. Henschler, *Biochem. Pharmacol.* **29** (1980) 993 – 998.
155. E. Eder, D. Lutz, D. Henschler, *Naunyn-Schmiedebergs Arch. Pharmacol.* **51** (1981) 316 – supplement.
156. E. Eder, T. Neudecker, D. Lutz, D. Henschler, *Chem. Biol. Interact.* **38** (1982) 303 – 315.
157. T. R. Torkelson, V. K. Rowe in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* **vol. 2 B,** 3rd ed., J. Wiley & Sons, New York 1981, pp. 3433 – 3601.
158. M. Bignami, G. Cardamine, P. Comba et al., *Mutat. Res.* **46** (1977) 243 – 244.
159. V. F. Simmer: "Structural Correlates of Carcino-genesis and Mutagenesis," *Second FDA Science Summer Symposium* , Washington, D. C., 1977, pp. 163 – 171.

160. V. L. Kubic, M. W. Anders, R. R. Engel, C. H. Barlow et al., *Drug Metab. Dispos.* **2** (1974) 53 – 57.
161. V. L. Kubic, M. W. Anders, *Drug Metab. Dispos.* **3** (1975) 104 – 112.
162. F. L. Rodkey, H. A. Collison, *Toxicol. Appl. Pharmacol.* **40** (1977) 39 – 47.
163. J. L. Stevens, J. H. Ratnayake, M. W. Anders, *Toxicol. Appl. Pharmacol.* **55** (1980) 484 – 489.
164. A. E. Ahmed, V. L. Kubic, J. L. Stevens, M. W. Anders, *Fed. Proc.* **39** (1980) 3150 – 3155.
165. V. L. Kubic, M. W. Anders, *Biochem. Pharmacol.* **27** (1978) 2349 – 2355.
166. A. E. Ahmed, M. W. Anders, *Biochem. Pharmacol.* **27** (1978) 2021 – 2025.
167. M. W. Anders, V. L. Kubic, A. E. Ahmed, *Xenobiotics* **440** (1978) 22 – 24.
168. L. A. Heppel, V. T. Porterfield, *J. Biol. Chem.* **176** (1948) 763 – 769.
169. A. E. Ahmed, M. W. Anders, *Drug Metab. Dispos.* **4** (1976) 357 – 361.
170. P. J. VanBladeren, D. D. Breimer, G. M. T. Rotteveel-Smijs, G. R. Mohn, *Mutat. Res.* **74** (1980) 341 – 346.
171. K. C. Back, E. W. VanStee, *Annu. Rev. Pharmacol. Toxicol.* **17** (1977) 83 – 95.
172. E. W. VanStee, A. M. Harris, M. L. Horton, K. C. Back, *Toxicol. Appl. Pharmacol.* **34** (1975) 62 – 71.
173. P. S. Beck, D. G. Clark, D. J. Tinston, *Toxicol. Appl. Pharmacol.* **24** (1973) 20 – 29.
174. J. H. Wills, P. Bradley, H. Kao, H. Grace et al., *Toxicol. Appl. Pharmacol.* **22** (1972) 305 – 306.
175. D. G. Smith, D. J. Harris, *Aerosp. Med.* **44** (1973) 198 – 201.
176. E. W. VanStee, K. C. Back, *Toxicol. Appl. Pharmacol.* **23** (1972) 428 – 442.
177. M. L. Gargas, M. E. Andersen, *Toxicol. Appl. Pharmacol.* **66** (1982) 55 – 68.
178. C. C. Comstock, R. W. Fogleman, F. W. Oberst, *Arch. Ind. Hyg. Occup. Med.* **7** (1953) 526.
179. T. R. Torkelson, F. Oyen, F. K. Rowe, *Am. Ind. Hyg. Assoc. J.* **21** (1960) 275.
180. B. Highman, J. L. Svirbely, W. F. von Oettingen, W. C. Alford, L. J. Pecora, *Arch. Pathol.* **45** (1948) 299.
181. *Toxic and Hazardous Ind. Chem. Safety Manual,* International Technical Information Institute, Tokyo, Japan, 1979, p. 80.
182. W. B. Deichmann: *Toxicology of Drugs & Chemicals,* Academic Press, New York 1969, p. 390.
183. J. L. Svirbely, B. Highman, W. C. Alford, W. F. von Oettinger, *J. Ind. Hyg.* **29** (1947) 382 – 389.
184. D. C. Call, *Aerosp. Med.* **44** (1973) 202 – 204.
185. R. A. Rhoder, K. L. Gabriel, *Toxicol. Appl. Pharmacol.* **21** (1972) 166 – 175.
186. W. Young, J. A. Parker, *Combust. Toxicol.* **2** (1975) 286 – 297.
187. V. A. Voronin, A. A. Denisenko, L. A. Linyucheva, N. M. Petushkov et al., *Gig. Tr. Prof. Zabol.* **2** (1982) 52 – 53.
188. E. W. VanStee, K. C. Back, *Toxicol. Appl. Pharmacol.* **25** (1973) 469.
189. P. S. Beck, D. G. Clark, D. J. Tinston, *Toxicol. Appl. Pharmacol.* **24** (1973) 20 – 29.
190. R. S. Weinstein, B. U. Pauli, J. Alroy, R. Fresco, *U.S. National Technical Information Service Report No. ADA 046051,* 1977, p. 19.
191. J. P. Murphy, E. W. VanStee, K. C. Back, *Toxicol. Appl. Pharmacol.* **41** (1977) 175.
192. B. Inouye, Y. Katayama, T. Ishida, M. Ogata et al., *Toxicol. Appl. Pharmacol.* **48** (1979) 467 – 477.
193. I. J. Stotz, Y. A. Greichus, *Bull. Environ. Contam. Toxicol.* **19** (1978) 319 – 325.
194. P. Lundberg: "Drug Oxidation," *Int. Symp. Microsomes,* 4th ed., 1980, p. 853.
195. R. Millischer, F. Girault, R. Heywood, G. Clark et al., *Toxicol. Eur. Res.* **2** (1979) 155 – 161.
196. J. M. Norris, J. W. Ehrmantraut, C. L. Gibbons, R. J. Kociba et al., *Appl. Polym. Symp.* **22** (1973) 195 – 219.
197. P. Koster, F. M. H. Debets, J. J. T. W. A. Strik, *Bull. Environ. Contam. Toxicol.* **25** (1980) 313 – 315.
198. J. M. Norris, R. J. Kociba, B. A. Schwetz, J. Q. Rose et al., *Environ. Health Perspect.* **11** (1975) 153 – 161.
199. J. M. Norris, J. W. Ehrmantraut, C. L. Gibbons, R. J. Kociba, *J. Fire Flammability Combust. Toxicol.* **1** (1974) 52 – 77.
200. R. J. Kociba, L. O. Frauson, C. G. Humiston, J. M. Norris et al., *J. Combust. Toxicol.* **2** (1975) 267 – 285.
201. G. P. Carlson, *Toxicol. Lett. (Netherlands)* **5** (1980) 19 – 25.
202. M. Goto, M. Hattori, K. Sugiura, *Chemosphere* **3** (1975) 177.
203. J. T. MacGregor, M. Friedman, *Mutat. Res.* **56** (1977) 81 – 84.
204. Z. E. Jolles (ed.): *Bromine and Its Compounds,* Academic Press, New York – London 1966.
205. W. F. Giaque, R. Wiebe, *J. Am. Chem. Soc.* **50** (1928) 2193.
206. Dow, US 2070263, 1937 (G. F. Dressel, O. C. Ross).

207. American Potash & Chemical Corp., US 2269733, 1942 (E. P. Pearson).
208. Great Lakes Chemical Corp., US 4083942, 1978 (D. C. Sanders).
209. A. S. Behrman, US 1916457, 1933 (A. S. Behrman).
210. Ethyl Corp., US 4248850, 1981 (K. A. Keblys).
211. C. M. Hudgins, R. L. McGlasson, E. D. Gould: "Developments in the Use of Dense Brines as Packer Fluids," *Drilling and Production Practice* 1961, 160 – 168.
212. J. H. Plonka: "New Bromide Packer Fluid Cuts Corrosion Problems," *World Oil* 1972 (April), 88 – 89.
213. Great Lakes Chemical Corp., US 4292183, 1981 (D. C. Sanders).
214. R. A. Putt: "Assessment of Technical and Economic Feasibility of Zinc/Bromine Batteries for Utility Load Leveling," Electric Power Research Institute (EPRI), EM-1059, Project 635-1, Gould Inc., Rolling Meadows, Ill., 1979.
215. A. L. Alexander, *Electric Vehicle Progress* **5** (1983) no. 6, 381.
216. Chemische Fabrik Kalk, DE 2103143, 1971.
217. A. J. Balard, *Ann. Chem. Phys.* **57** (1834) 225.
218. A. Massagli, A. Indelli, F. Pergola, *Inor. Chim. Acta* **4** (1970) no. 4, 593 – 596.
219. T. Osuga, K. Sugino, *J. Electrochem. Soc.* **104** (1957) no. 7, 448 – 451.

Bronchodilators → Antiasthmatic Agents
Broncholytic Agents → Cough Remedies
Bronze → Copper Alloys
Brown Coal → Coal
BTX Processing → Benzene

Butadiene

JOACHIM GRUB, Erdölchemie GmbH, Köln, Federal Republic of Germany (Chaps. 1 – 7)

ECKHARD LÖSER, Bayer AG, Wuppertal, Federal Republic of Germany (Chap. 8)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 5
4.1. Dehydrogenation of Butane and Butenes 6
4.2. Isolation of Butadiene from C_4 Steam Cracker Fractions 7
5. Specifications of Butadiene 10
6. Stabilization, Storage, and Transportation 11
7. Uses and Economic Importance 11
8. Toxicology 13
8.1. Experimental Toxicity Data 13
8.2. Metabolism 14
8.3. Human Data 14
9. References 15

1. Introduction

The name butadiene generally refers to 1,3-butadiene [*106-99-0*], C_4H_6, M_r 54.09. The thermodynamically less stable 1,2-butadiene [*590-19-2*], whichhas two cumulated double bonds, has only minor industrial importance. The1,3 isomer is economically the most important unsaturated C_4 hydrocarbon.

$CH_2=CH-CH=CH_2$ 1,3-Butadiene
$CH_2=C=CH-CH_3$ 1,2-Butadiene

CAVENTOU was the first to isolate butadiene, by the pyrolysis of amyl alcohol. BERTHELOT produced it by passing a mixture of acetylene and ethylene through a red-hot iron tube. In 1895 the structure of butadiene was elucidated by CIAMICIAN and MAGNAGHI. In 1886, ARMSTRONG and MILLER discovered butadiene in the products obtained on cracking petroleum.

Practical interest in butadiene and its derivatives began with the discovery by LEBEDEW in 1910 that butadiene forms rubberlike polymers. Butadiene was produced on a large scale in Germany prior to World War II and in the United States during the war. Simultaneously, methods were developed for the manufacture of useful butadiene polymers [1]. After the war, the work of ZIEGLER and NATTA, among others, on polymerization with organometallic catalysts led to higher quality rubber.

2. Physical Properties

Butadiene is a colorless gas under normal conditions. Some physical properties are summarized in the following [3], [4]:

bp at 101.3 kPa	− 4.4 °C
mp at 1013. kPa	− 108.9 °C
Critical temperature	425 K
Critical pressure	4.32 MPa
Critical molar volume	221 cm^3/mol
Density, liquid	
at 0 °C	0.646 g/cm^3
at 25 °C	0.616 g/cm^3
at 50 °C	0.582 g/cm^3
Gas density (air = 1)	1.87
Vapor pressure	
at − 4.4 °C	101.3 kPa
at 0 °C	120.0 kPa
at 25 °C	273.6 kPa
at 50 °C	537.9 kPa
at 75 °C	986.7 kPa
at 100 °C	1733 kPa
Viscosity, liquid	
at 0 °C	0.25 mPa · s
at 50 °C	0.20 mPa · s
Enthalpy of vaporization	
at 25 °C	20.86 kJ/mol
at − 4.4 °C	22.47 kJ/mol
Enthalpy of formation, gaseous, at 298 K, 101.3 kPa	110.0 kJ/mol
Free enthalpy of formation, gaseous, at 298 K, 101.3 kPa	151.5 kJ/mol
Enthalpy of combustion, gaseous, at 298 K, 101.3 kPa	2541.5 kJ/mol
Enthalpy of hydrogenation to butane, gaseous, at 298 K, 101.3 kPa	235.7 kJ/mol
Entropy of formation, liquid, at 298 K, 101.3 kPa	199.0 J mol^{-1} K^{-1}
Enthalpy of melting at 164.2 K, 101.3 kPa	7.988 kJ/mol

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_431

The technical data [2] important for reasons of safety are, above all, the flash point, − 85 °C, the ignition temperature, 415 °C, and the explosion limits when mixed with air and oxygen (Table 1). Unstabilized or insufficiently stabilized butadiene forms explosive peroxides with atmospheric oxygen.

Table 1. Explosion limits of butadiene in air (concentrations of butadiene stated) [5]

	At 101.3 kPa, 20 °C		At 490.4 kPa, 30 °C	
	vol %	g/m^3	vol %	g/m^3
Lower limit	1.4	31	1.4	150
Upper limit	16.3	365	ca. 22	ca. 2400

Table 2 lists azeotropic mixtures relevant to the distillation of butadiene-containing hydrocarbons.

Table 2. Binary azeotropic mixtures of 1,3-butadiene and of other C_4-hydrocarbons [6]

Mixture	*bp*, °C ∗	Composition
Butane/butadiene	min.	
trans-2-Butene/1-butyne		25.5 wt % 1-butyne
cis-2-Butene/vinylacetylene	min.	
Butadiene/2-butene	− 5.53	24.5 wt % 2-butene
Methylamine/vinylacetylene	− 6.8	2.5 wt % vinylacetylene
Ammonia/butadiene	− 37	45 wt % butadiene
Ammonia/1-butene	− 37.5	55 wt % 1-butene
Ammonia/isobutene	− 38.5	55 wt % isobutene
Ammonia/*n*-butane	− 37.1	55 wt % *n*-butane
Ammonia/isobutane	− 38.4	65 wt % isobutane
Methylamine/butadiene	− 9.5	58.6 wt % butadiene
Acetaldehyde/butadiene	5.0	94.8 wt % butadiene

∗ At 101.3 kPa.

Butadiene is sparingly soluble in water, see Table 3, soluble in methanol and ethanol, and very soluble in higher-boiling polar solvents, e.g., methylpyrrolidone.

Table 3. Solubility α of butadiene in water at 101.3 kPa (Bunsen's solubility coefficient) and solubility L of water in liquid butadiene

t, °C	α, m^3/m^3	L, g H_2O/kg butadiene
10	0.29	0.53
20	0.23	0.66
30	0.19	0.82
40	0.16	

3. Chemical Properties

Butadiene has two conjugated double bonds and therefore can take part in numerous reactions, which include 1,2- and 1,4-additions with itself (polymerization) and other reagents, linear dimerization and trimerization, and ring formation.

Polymerization. Polymerization by 1,2- and 1,4-addition is the most important reaction of butadiene [7].

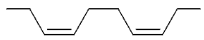

cis-1,4-addition

trans-1,4-addition

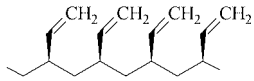

isotactic 1,2-addition

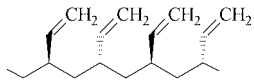

syndiotactic 1,2-addition

The nature of these polymers depends greatly on the way in which they are prepared and on the catalyst system employed (→ Rubber, 3. Synthetic, Chap. 3.1.2.; → Rubber, 3. Synthetic, Chap. 3.2.1.).

Other Addition Reactions. Butadiene undergoes 1,2- or 1,4-addition with many reagents. The relative proportions of the 1,4- and 1,2-addition products depend on the reaction conditions such as temperature, solvent, reaction time. The addition products are important intermediates in the manufacture of, for example, chloroprene and hexamethylenediamine. A typical example of an electrophilic addition to butadiene is the reaction with hydrogen chloride.

Chloroprene [*126-99-8*] (**1**) is produced by the chlorination of butadiene followed by isomerizationof the 1,4-isomers and alkaline dehydrochlorination (→ Chlorinated Hydrocarbons, Chap. 6.4.3.1.):

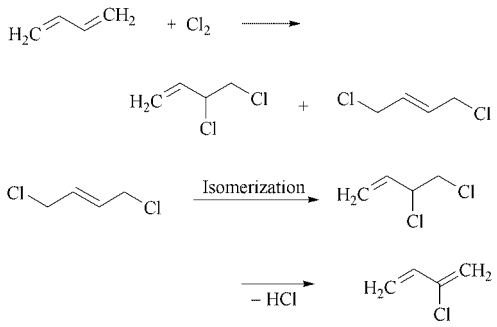

1

In the production of *adiponitrile* [*111-69-3*] (**2**) and *hexamethylenediamine* [*124-09-4*] (**3**), [8], [9] hydrogen cyanide reactswith butadiene in two steps, and the adiponitrile thus obtained is hydrogenated to give the diamine. Typical catalysts are Ni^0 phosphine and phosphite complexes with Al/Zn promotors. Typical reaction conditions are 90 – 150 °C and ambient pressure in THF.

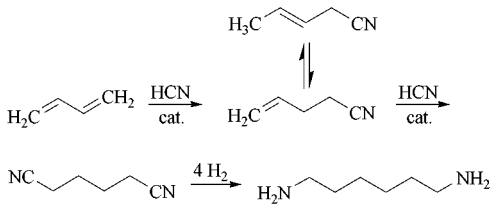

2 3

Several processes have been developed for the production of *1,4-butanediol* [*110-63-4*] from butadiene. In the three-stepMitsubishi process [10], [11], butadiene catalytically reacts with acetic acid to give 1,4-diacetoxy-2-butene (**4**), which in turn is hydrogenated to 1,4-diacetoxybutane and hydrolyzed to 1,4-butanediol. A similar process was reported by BASF [12].

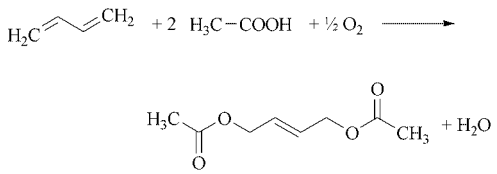

4

The Toyo Soda process [13] for the preparation of 1,4-butanediol involves the reaction of the butadiene chlorine addition products 1,4-dichloro-2-butene and 1,2-dichloro-3-butene with sodium acetate or formate to give 1,4-diacetoxy-2-butene or 1,4-diformyl-2-butene, which are then hydrolyzed and hydrogenated to 1,4-butanediol.

The direct oxidation of butadiene with air or oxygen, a method introduced by Eastman Chemical [14], is an economic process for the production of several C_4 derivatives. In the oxidation 3,4-epoxy-1-butene (**5**) is formed with high selectivity. Thermal rearrangement of **5** followed by hydrogenation affords tetrahydrofuran, which can be hydrolytically ring opened to give 1,4-butanediol. Hydrogenation of **5** affords ethyloxirane (**6**).

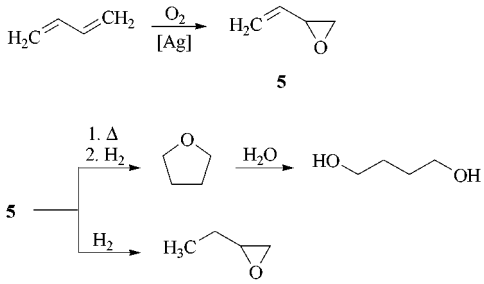

5 6

The *Diels–Alder reaction* is one of the best known reactions of butadiene [15]. The dienophile is usually an olefin with an activated double bond, which reacts with butadiene to give a cyclohexene ring. This addition reaction, which is exclusively a 1,4-addition, can also take place with a second molecule of butadieneto give 4-vinylcyclohexene [*100-40-3*] (**7**) [16]. The reaction can take place withor without a catalyst [17–19]. Vinylcyclohexene can be dehydrogenated or oxidized to give styrene [*100-42-5*] [20].

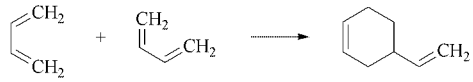

7

Two molecules of butadiene can cyclodimerize to form 1,5-cyclooctadiene [*111-78-4*] and three molecules of butadiene

can cyclotrimerize to 1,5,9-cyclododecatriene [*4904-61-4*] [21]. Typical catalysts are Ni^0 complexes with phosphine or phosphite ligands. 1,5-Cyclooctadiene and 1,5,9-cyclododecatriene are important intermediates in the production of higher polyamides. 1,5-Cyclooctadiene can also be used to produce polyoctenamers via hydrogenation to cyclooctene and metathesis reaction.

Butadiene readily undergoes a reversible 1,4-addition with sulfur dioxide to form the cyclic sulfone 2,5-dihydrothiophene 1,1-dioxide [*1708-32-3*] (**8**) [22], [23], which is converted into sulfolan [*126-33-0*] (**9**), a heat-stable and highly polar solvent, by catalytichydrogenation.

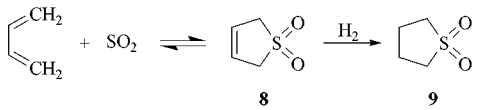

Linear Dimerization and Trimerization; Telomerization. Butadiene forms linear dimers or trimers in the presence of Ni, Co, Pd, or Fe catalysts [24]. Since the linear oligomers of butadiene differ from each other by a chain length of four carbon atoms, their separation is easier than that of ethylene oligomers. Dimerization results in the formation of 1,3,7-octatriene [*1002-35-3*] and 1,3,6-octatriene [25], [26], which are, however, industrially uninteresting. The aromatic hydrocarbons xylene and ethylbenzene [27] can be made by the dehydrocyclization of 1,3,7-octatriene. If the dimerization is carried out under hydrogenation conditions, 1,7-octadiene and 1,6-octadiene are formed [28], [29]. 1,7-Octadiene is converted into a C_{10} diol by hydroformylation and into a C_{10} diamine by hydrocyanation and subsequent hydrogenation.

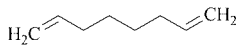

1,7-Octadiene

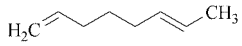

1,6-Octadiene

Dimerization of butadiene and simultaneous reaction with carbon monoxide and alcohol, followed by hydrogenation and saponification leads to pelargonic acid (*n*-nonanoic acid) [30], a starting material in the production of heat-resistant lubricants.

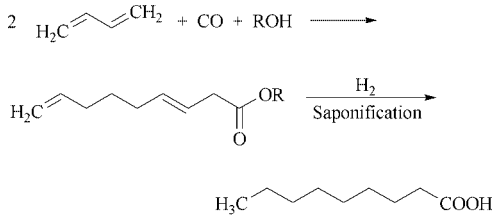

The telomerization of butadiene with a carboxylic acid [31] eventually results in 1-octene [*111-66-0*].

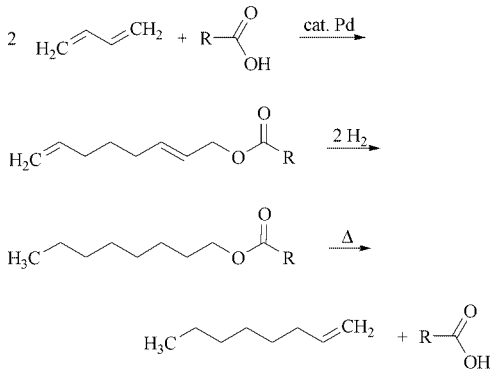

The telomerization of butadiene with water, a Kuraray process [32], and subsequent hydrogenation eventually results in *n*-octanol [*111-87-5*] with a selectivityof greater than 90 %.

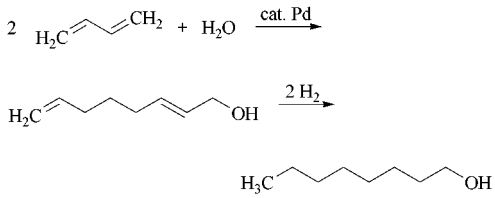

The telomerization of butadiene with ammonia [*7664-41-7*] in biphasic liquid–liquid systems with palladium acetateand trisodium 3,3′,3″-phosphanetriylbenzenesulfonate as catalyst yields in high selectivity the primary octadienylamines 2,7-octadienylamine (**10**) and 3-

amino-1,7-octadiene (**11**) [33]. In biphasic systems the consecutive reactions to give the secondary and tertiary amines are dramatically suppressed in comparison to monophasic systems. The industrially important *n*-octylamine (starting material for softeners, corrosion inhibitors, flotation agents) is obtained by hydrogenation of 2,7-octadienylamine.

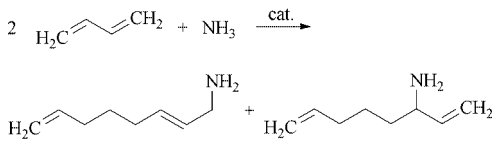

10 **11**

The telomerization of butadiene with carbon dioxide yields different products. Products, yields and selectivities depend on the catalyst type [34]. If $[Pd(acac)_2]/iPr_3P$ (acac = acetylactonoate) is used as the catalyst system, the δ-lactone **12** is formed in high yield and with high selectivity. This lactone could be of interest as a starting material for the synthesis of unsaturated and functionalized long-chain carboxylic acids.

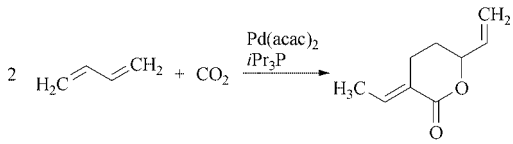

12

Selective hydrogenation of butadiene to butene and complete hydrogenation to butane are possible. Industrially, these reactions are carried out with mixed C_4 fractions in cases of butadiene oversupply. The hydrogenated C_4 fractions are recycled to the steam cracker.

Formation of Complexes. Butadiene reacts with numerous metal compounds to form complexes, e.g., with Cu(I) salts, which were formerly used in the extraction of butadiene from C_4 hydrocarbon mixtures. This method has been replaced by modern extractive distillation techniques (Section 4.2). Complexes with other transition metalls like Fe, Ni, Co, Pd, and Pt are also well known.

4. Production

A series of articles was published in 1942 describing all the methods used to produce butadiene [35]. Historically butadiene was industrially produced from acetylene by two processes (Fig. 1)

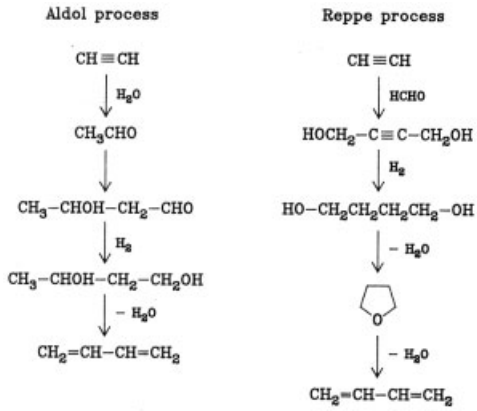

Figure 1. Production of butadiene from acetylene

These methods are no longer used as the production of acetylene requires much energy and is expensive. Only the first steps of the Reppe process leading to 1,4-butanediol and tetrahydrofuran are commercially employed today. The production from ethanol [36], [37], was first developed by IPATJEW and OSTROMISLENSKY and was modified by LEBEDEW into a commercial process.

$$2\,CH_3CH_2OH \longrightarrow CH_2{=}CHCH{=}CH_2 + 2\,H_2O + H_2$$

This process, which operates at 370 – 390 °C on $MgO-SiO_2$ or $SiO_2-Al_2O_3$ catalysts with a selectivity of up to 70 % can be of interest for countries with cheap ethanol or a large surplus of agricultural alcohol. Nowadays butadiene is preferentially isolated from C_4 fractions obtained by cracking naphtha and gas oil to ethylene (→ Ethylene). The dehydrogenation of *n*-butane or *n*-butenes plays a minor role and is only operated on a campaign basis when there is a large enough differential between feedstock and butadiene prices.

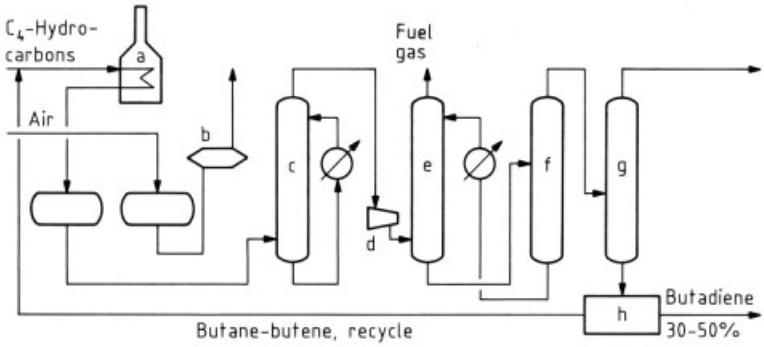

Figure 2. Houdry Catadiene process
a) Reactor; b) Waste-heat boiler; c) Quench column; d) Compressor; e) Absorber; f) Stripper; g) C_3 Column; h) C_4 Separation

4.1. Dehydrogenation of Butane and Butenes

The dehydrogenation reactions are endothermic, the following values being valid at 430 °C:

$$\text{Butane} \rightarrow \text{1-Butene} + H_2$$
$$\Delta H = 131\,\text{kJ/mol}$$
$$\text{Butane} \rightarrow \textit{cis}\text{-2-Butene} + H_2$$
$$\Delta H = 118\,\text{kJ/mol}$$
$$\textit{cis}\text{-2-Butene} \rightarrow \text{Butadiene} + H_2$$
$$\Delta H = 126\,\text{kJ/mol}$$

According to the Le Chatelier's principle, the yield is increased by decreasing the partial pressure of the reaction products. Practically, the dehydrogenation can be improved by conducting the reaction under vacuum or by addition of steam. The latter has the following advantages:

1) It reduces coking of the dehydrogenation catalyst
2) It provides the heat required for the endothermic dehydrogenation process
3) Steam can be easily separated from the reaction products by condensation

A further increase in yield is achieved by raising the reaction temperature.

Undesirable side reactions are cracking, isomerization, and polymerization. Compounds which tend to undergo these reactions are removed before the dehydrogenation process. As the dehydrogenation process does not proceed to completion, the reaction products must be separated, and the starting materials must be recovered and recycled into the dehydrogenation process.

Dehydrogenation of *n*-Butane. The best known one-step dehydrogenation is the Houdry Catadiene process (Fig. 2), which has been in operation on a commercial scale since 1943 [38–41]. In this adiabatic process, several packed-bed reactors, arranged parallel to each other, are operated alternatingly. Aluminum oxide mixed with approximately 20 % chromium oxide is the catalyst. *n*-Butane is subjected to dehydrogenation as such or in a mixture of *n*-butenes at 600 – 700 °C and 10 – 25 kPa. The use of high temperatures results in byproducts like $C_1 - C_3$ hydrocarbons, hydrogen, and carbon deposits on the catalyst. After 5 – 15 min of running time, the reactor is switched to regeneration. The heat generated by burning the coke residue during the regeneration phase is stored in the catalyst and in the added inert material and is then reused in the next reaction phase.

The concentration of butadiene at the outlet of the reactor is 15 – 18 %. During the recovery process, which includes absorption of the C_3 and C_4 hydrocarbons, compression, stripping, and separation from unconverted *n*-butane and *n*-butenes, the concentration of butadiene is increased to 30 – 50 %. Approximately 550 t of butadiene is obtained from 1000 t of *n*-butane.

Phillips Petroleum has developed a two-step dehydrogenation process [42], [43]:

1) *n*-Butane is catalytically dehydrogenated to butene at 600 °C and 1 bar on $Cr_2O_3 - Na_2O - Al_2O_3$.
2) The *n*-butenes are separated by extractive distillation with auxiliaries such as acetone, acetonitrile, and furfural.

3) The *n*-butenes are dehydrogenated to butadiene in an isothermic tubular reactor, heated with flue gas to 600 °C, at a pressure of 1 bar, with addition of superheated steam on a $Fe_2O_3 - K_2O - Al_2O_3$ catalyst.
4) Butadiene is recovered by extractive distillation with the above-mentioned auxiliaries and purified.

The advantages of this method are the longer running times without catalyst regeneration and higher butadiene yields (65 % based on butane).

Dehydrogenation of *n*-Butenes. *n*-Butenes are formed in the production of standard gasoline by catalytic cracking, in the production of olefins by thermal cracking, and in the dehydrogenation of LPG. Normally, butenes are a part of a C_4 hydrocarbon mixture. The C_4 paraffins can be separated from the C_4 olefins by extractive distillation, as routinely employed in the recovery of butadiene [42]. The boiling points of isobutene and *n*-butene are so close together that they can only be separated by special methods. The gas-phase BASF process employs 40 – 45 % H_2SO_4 to selectively convert isobutene into *tert*-butanol [48]. Today, separation is carried out by conversion to methyl *tert*-butyl ether or to isobutene oligomers [49–52]. Table 4 lists catalysts for the dehydrogenation of *n*-butenes to butadiene.

Oxidative Dehydrogenation of *n*-Butenes [53], [54]. The conversion and the selectivity of the dehydrogenation of *n*-butenes to butadiene can be significantly improved by removing the hydrogen from the equilibrium. The addition of oxygen causes the oxidation of hydrogen to water:

$$2\,C_4H_8 + O_2 \longrightarrow 2\,C_4H_6 + 2\,H_2O$$

The exothermic oxidation of hydrogen partially covers the heat requirements of the endothermic dehydrogenation reaction and, in addition, the oxygen, together with steam added during the reaction, leads to decreased catalyst coking. Especially productive methods are the Oxo-D process of Petro-Tex [55] and the O-X-D process of Phillips [56], [57].

The Oxo-D process was first applied on a large scale in 1965. The advantages of this method are the low consumption of steam and heating energy, high conversion and selectivity per reactor cycle, longer life span of the catalyst, and no necessity for catalyst regeneration. Petro-Tex achieved a 65 % conversion and a butadiene selectivity of 93 % by using a molar steam/butene ratio of 12/1.

Phillips has operated a plant in Borger (Texas) since 1970 based on the O-X-D process, producing 125 000 t/a of butadiene. Phillips also obtains a high conversion and a 88 – 92 % selectivity of butadiene.

The addition of oxygen to the dehydrogenation of butanes is not meaningful because at the high temperatures required, oxygen reacts with the reaction products to give undesired byproducts.

Oxidative Dehydrogenation of Butane with Iodine [58], [59]. Shell has developed a one-step dehydrogenation of butane to butadiene with iodine as the hydrogen acceptor. The addition of iodine enables a high conversion and yield of butadiene but causes serious corrosion problems in the plant.

4.2. Isolation of Butadiene from C_4 Steam Cracker Fractions

Today, the isolation of butadiene is based worldwide predominantly on butadiene-containing C_4 fractions from the steam cracking of ethane,

Table 4. Catalytic dehydrogenation of butenes [44–47]

Catalyst	Phillips 1490	Shell 205	Dow Type B
Composition	Fe_2O_3 – Bauxite	Fe_2O_3 – Cr_2O_3	Ca/Na phosphate – Cr_2O_3
Mole ratio H_2O/butene	9 – 12/1	8/1	18 – 20/1
Conversion per pass, %	26 – 28	27 – 33	$\leq$45
Selectivity, %	76	70	90
Regeneration after		24 h	30 min
Duration of regeneration		1 h	15 min

Table 5. Feedstocks and steam cracking yields (in wt %)

	Light hydrocarbons			Naphthas and heavier		
	Ethane	Propane	*n*-Butane	Medium range naphthas	Atm. gas oil	Light vacuum gas oil
Feedstocks						
Methane	0.6					
Ethane	95.2					
Propane	4.2	100				
n-Butane			100			
Paraffins				79.9		
Olefins				–		
Naphthenes				17.4		
Aromatics				2.7		
Total	100	100	100	100		
TBP cut range, °C					204 – 343	343 – 454
ASTM boiling range, °C				47 – 148		
Rel. density				0.692	0.844	0.901
Cracking yields						
Hydrogen and methane	15.1	29.7	23.8	17.7	12.1	9.5
Ethylene	77.7	42.0	40.0	34.0	25.9	20.5
Propylene	2.8	16.8	17.2	15.7	16.2	14.1
Butenes and butanes	0.8	1.3	6.7	4.3	4.8	6.3
Butadiene	1.9	3.0	3.5	4.7	4.6	5.4
Pyrolysis gasoline	1.7	6.6	7.1	18.8	18.4	19.3
Pyrolysis fuel oil	–	0.6	1.7	4.8	18.0	25.0
Total	100.0	100.0	100.0	100.0	100.0	100.0

Table 6. Typical analyses of C_4 -fractions (in vol %) [64], [65]

	Cracking intensity				Dehydrogenates
	moderate	medium	high	very high	
C_3 -Hydrocarbons	0.3	0.3	0.3	0.16	*
Butane	4.2	5.2	2.8	0.54	*
Isobutane	2.1	1.3	0.6	0.53	*
1-Butene	20.0	16.0	13.6	9.18	*
cis-2-Butene	7.3	5.3	4.8	1.61	*
trans-2-Butene	6.6	6.5	5.8	3.63	*
Isobutene	32.4	27.2	22.1	10.13	*
1,3-Butadiene	26.1	37.0	47.4	70.10	15 – 45
1,2-Butadiene	0.12	0.15	0.2	0.40	**
Methylacetylene	0.06	0.07	0.08	0.10	**
Vinylacetylene	0.15	0.3	1.6	2.99	**
Ethylacetylene	0.04	0.1	0.2	0.53	**
C_5 -Hydrocarbons	0.5	0.5	0.5	0.1	**

* sum of fractions = 85 – 55 vol%
** sum of fractions = 0.03 – 0.2 vol%

LPG, naphtha, gas oil, and other higher boiling hydrocarbon fractions to form ethylene and homologous compounds. In Western Europe, naphtha is primarily used as raw material. In the United States, however, liquefied petroleum gas (LPG) and natural gas liquids account for more than 70 % of the starting materials (→ Ethylene, Chap. 4.) [60].

Table 5 shows typical cracking yields and their relationship to the starting materials [61–63]. The influence of the cracking severity (temperature and duration) on the butadiene yield is minor compared to the influence exerted by the starting materials. However, the composition of the C_4 fraction itself is greatly dependent on cracking severity (Table 6).

Butadiene cannot be separated from this C_4 hydrocarbon mixture by simple distillation, because 1,3-butadiene and butane form an azeotrope (see Table 2). Hence, butadiene is isolated by liquid – liquid extraction or extractive distillation.

Table 7. Relative volatility of C_3- and C_4-hydrocarbons in comparison to butadiene (50 °C, infinite dilution) [70]

	Without solvent	NMP *	DMF *	Acetonitrile	DMAC *	Furfural
Butane	1.17	3.66	3.43	3.13	3.13	2.89
1-Butene	1.08	2.38	2.17	1.92	2.07	1.78
trans-2-Butene	1.23	1.90	1.76	1.59	1.71	1.20
cis-2-Butene	1.37	1.63	1.56	1.45	1.52	1.26
1,3-Butadiene	1.00	1.00	1.00	1.00	1.00	1.00
1,2-Butadiene	1.79	0.74	0.72	0.73	0.71	0.65
Methylacetylene	2.16	0.81	0.72	1.00	0.73	1.04
Ethylacetylene	1.62	0.42	0.42	0.48	0.44	0.52
Vinylacetylene	1.44	0.21	0.23	0.39	0.23	0.41

* NMP *N*-Methylpyrrolidone; DMF Dimethylformamide; DMAC Dimethylacetamide

CAA Liquid – Liquid Extraction [42]. The cuprous ammonium acetate (CAA) process, introduced by ESSO, is based on the ability of butadiene to form a complex with Cu^I compounds. Butadiene is separated from the C_4 hydrocarbon mixture by countercurrent extraction with Cu^I acetate in aqueous ammonia solution. This method is advantageous in the processing of C_4 fractions with a low butadiene and C_4 acetylene content, which is the case in dehydrogenation fractions [66]. A higher content of α-acetylenes (e.g., ca. 100 ppm of vinylacetylenes) strongly interferes with this process as a result of foam formation. Acetylenes, especially the unstable vinylacetylene, can be removed by selective hydrogenation before the isolation of butadiene [67–69].

Extractive Distillation. The method of choice for the isolation of butadiene from C_4 fractions is extractive distillation with selective organic solvents. The affinity of hydrocarbons to polar solvents depends directly upon their degree of unsaturation. A highly unsaturated hydrocarbon is not only more soluble in a polar solvent, but the solvent is also more effective in decreasing the volatility of the hydrocarbon (see Table 7).

The first plants using furfural as solvent were built in the United States during World War II (the process was developed by Phillips Petroleum) [42], [71]. The method introduced by Shell in the early 1950s uses acetonitrile as solvent. It was developed from a one-step procedure involving separate removal of acetylene into a two-step process [42]. Union Carbide has described a process based on dimethylacetamide [72]. Modern plants employ the BASF procedure [66], [73–76] with *N*-methylpyrrolidone or the DMF method [77], [78] of Nippon Zeon.

A flowsheet of the BASF process is shown in Figure 3 (see next page). *N*-Methylpyrrolidone (*bp* 203 °C at 101.3 kPa), containing 5 – 10 % water to increase the selectivity, is used as solvent. In the main washer (a), the more unsaturated hydrocarbons, such as acetylenes and diolefins, are dissolved in NMP, and the butenes/butanes are removed from the top of the washer. In a second extractive distillation step, in tower (c) and in the lower part of tower (b), the C_4 acetylenes and part of the 1,2-butadiene, which are more soluble in NMP than butadiene, are removed from the bottom of (c). The butadiene, which corresponds to specifications as far as alkanes, alkenes, and C_4 acetylenes are concerned, is removed from the top of tower (c) and is separated from the more volatile propyne in tower (h) and from the less volatile 1,2-butadiene and C_5 hydrocarbons in tower (i). The loaded solvent is led from the sump of tower (b) into tower (e) for complete degassing. Here the C_4 acetylenes are removed as a sidestream and, after selective hydrogenation, returned to tower (a). Recoveries of 99 – 100 %, based on the butadiene content of the C_4 feed, are achieved with selective hydrogenation of vinylacetylenes to butadiene [79]. An additional advantage of hydrogenation is the safe removal of the unstable C_4 acetylenes. Improper handling of C_4 acetylenes led to a severe explosion and complete destruction of a U.S. plant working with dimethylacetamide [80], [81]. Other efficient solvents for the extractive distillation of butadiene are given in [82]; e.g., methoxypropionitrile is used in some plants [85], [86].

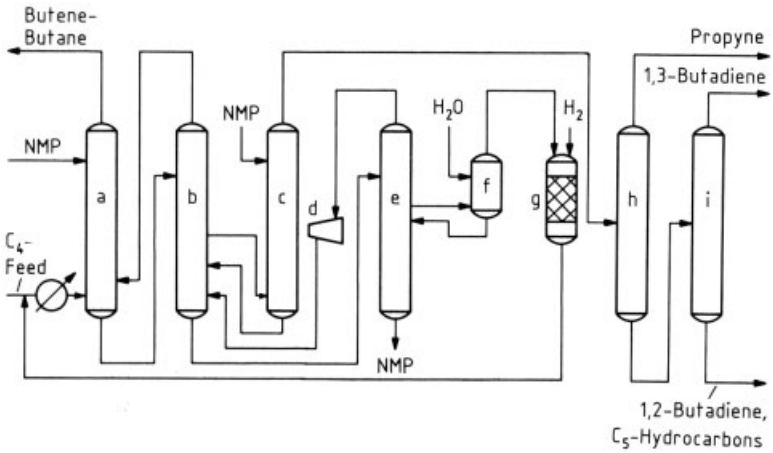

Figure 3. BASF Extractive distillation process with NMP
a) Main washer; b) Rectifier; c) After washer; d) Centrifugal compressor; e) Degassing tower; f) Water washer; g) Hydrogenation reactor; h) First distillation; i) Second distillation

In newer developments, C_4 acetylenes are removed by selective hydrogenation of the crude C_4 feedstock. UOP offers the KLP process, in which the crude C_4 stream is mixed with a stoichiometric amount of hydrogen and fed to a fixed-bed reactor containing a selective hydrogenation catalyst [83]. Temperatures are not allowed to exceed 85 °C. The reactor pressure is set to maintain the reaction mixture in the liquid phase. Because of catalyst fouling caused by polymeric materials that are produced during operation, catalyst regeneration by carbon combustion with air followed by catalyst reduction with hydrogen, is necessary every two months. The butadiene yield in comparison with a conventional two-stage extraction is increased from 95.5 to 99.5 %.

CDTECH offers the CDHYDRO process for acetylenes conversion in the crude C_4 feedstock [84]. The CDHYDRO process involves placing catalyst in the steam cracker debutaniser, adding hydrogen to the feed, and venting off any gases from the reflux receiver. Two modes of operation are available. In the first, the acetylenes concentration can be reduced to levels low enough to allow a single extraction system for butadiene purification, but with a loss of 3 – 4 % butadiene by hydrogenation. The second mode is only applicable with a double extraction unit. The acetylenes conversion is lower so that there is nearly no loss of butadiene. Unconverted acetylenes are extracted in the second stage and recycled to the debutanizer feed. Conversion of butadiene to butenes is less than 1 %.

5. Specifications of Butadiene

A typical butadiene specification to meet market requirements is listed in Table 8. The purity requirements have been intensified because the Ziegler-type catalysts employed in the stereospecific synthesis of rubbers are sensitive to,

Table 8. Typical specifications of butadiene

Compound	Specification	Test method
1,3-Butadiene	$\geqq$ 99.6 %	ASTM-D 2593-86
Propyne	max. 5 ppm	ASTM-D 2593-86
Propadiene	max. 2 ppm	ASTM-D 2593-86
Vinylacetylene	max. 5 ppm	ASTM-D 2593-86
Ethylacetylene	max. 2 ppm	ASTM-D 2593-86
Σ Acetylenes (calculated as vinylacetylene)	max. 20 ppm	ASTM-D 1020-76
Dimerics	max. 250 ppm *	ISO 7381
1,2-Butadiene	max. 30 ppm	ASTM-D 2593-86
Carbonyls (calculated as CH_3CHO)	max. 25 ppm	ASTM-D 4423-91
Sulfur	max. 1 ppm	ASTM-D 3246-76
Peroxides (calculated as H_2O_2)	max. 1 ppm	ASTM-D 1022-76
Methanol	max. 10 ppm	ISO 8174
H_2O	max. 5 ppm	ISO 6191
Residue	max. 10 ppm	ASTM-D 2593-86
Inhibitor TBC **	75 – 125 ppm	ASTM-D 2593-86
Oxygen in the gas phase	max. 0.1 vol %	ASTM-D 2593-86

* Maximum set by battery producers;
** TBC = *tert*-butylpyrocatechol

for example, allenes and acetylenes. The purity requirements are less stringent for butadiene used in emulsion polymerization, such as in the synthesis of styrene – butadiene rubber and latex [85].

6. Stabilization, Storage, and Transportation

Stabilization [78] involves protection against the action of oxygen and against polymerization. Butadiene forms dangerous polymeric peroxides [86], [87] with oxygen, which are viscous liquids sparingly soluble in liquid butadiene. They are deposited at the bottom of butadiene containers. Hence, the handling of butadiene requires complete exclusion of oxygen. In addition, inhibitors such as *p-tert*-butylcatechol (TBC) and 2,6-di-*tert*-butyl-*p*-cresol (TBK), are added because they are especially effective at scavenging radicals. They also prevent spontaneous polymerization of butadiene and can easily be removed by washing with aqueous sodium hydroxide. Aqueous sodium nitrite solution is also used as an antioxidant in the production of butadiene.

Butadiene can polymerize during production, storage, and transportion in three ways. Dimerization to vinylcyclohexene [88] is dependent on time and temperature and can be stopped by adding potassium cyanide solution to butadiene [89]. Second, butadiene can polymerize under the influence of oxygen and high temperatures, especially during manufacture, to give rubber-like polymers (fouling). Antioxidants such as TBC, exclusion of oxygen, low temperatures, and short exposures to higher temperatures reduce this polymerization [90]. Third, butadiene can polymerize, initiated by oxygen, high temperatures, and rust, to give a glassy, granular, opalescent, and very hard polymer (so-called popcorn) [90], [91]. It forms preferentially in off-stream tubes. Its growth is favored when popcorn seeds are present in the plant. The formation of popcorn in closed areas can lead to pressures of > 1000 bar, causing the bursting of off-stream tubes or containers. It can be prevented by the addition of TBC and by the careful elimination of all popcorn seeds. Popcorn is spontaneously inflammable in air and hence must be kept moist during its removal.

Liquified butadiene is stored at normal temperatures in liquid-gas containers or, preferably, for safety reasons at a temperature of 4 °C in almost pressureless containers. It is transported at normal temperatures and raised pressures in tankers, railroad tank wagons, and in ships. Especially in ships, temperatures of 4 °C and pressureless containers are required [92–94].

The following regulations must be followed when labeling and transporting butadiene:

- By road: Germany: GGVS, Class 2, no. 2F; Europe: ADR; USA: DOT regulations.
- By rail: Germany: GGVE, Class 2, no. 2F; Europe: RID; USA: DOT regulations.
- By inland shipping: GGVS/GGVE, Class 2, no. 2F; Europe: ADR.
- By sea freight: Germany: GGVSee: Class 2, no. 2F; International: IMDG Code.
- By air freight: International: IATA, Class 2, no. 2.1.

Butadiene is safely transported today in amonts greater than 100 000 t/a from, for example, Western Europe to Japan and to the United States.

7. Uses and Economic Importance

The majority of the butadiene produced worldwide is used as a monomer or co-monomer in the manufacture of synthetic rubber (→ Rubber, 3. Synthetic), above all for styrene – butadiene rubber and latex (SBR), polybutadiene rubber (BR), acrylonitrile – butadiene rubber and latex (NBR), and for chloroprene rubber (CR). Important plastics containing butadiene as a monomer component are impact-resistant polystyrene, a two-phase system consisting of polystyrene and polybutadiene; ABS polymers, consisting of acrylonitrile, butadiene, and styrene; and a copolymer of methyl methacrylate, butadiene, and styrene (MBS), which is used as a modifier for poly(vinyl chloride) (→ Polystyrene and Styrene Copolymers, Chap. 5.12.2.). In addition, butadiene is an intermediate in the synthesis of several important chemicals (Chap. 3).

The worldwide production capacity, production, and consumption of butadiene are summarized in Table 9 (see next page). Butadiene-containing C_4 steam cracker fractions are the

Table 9. Butadiene production, capacities, and consumption 1998 (in 10^3 t/a) [95]

	Production	Capacities	Consumption
USA/Canada	2020	2105	2218
Latin America	312	333	398
Western Europe	2000	2345	1752
Eastern Europe *	430	1080	452
Asia and Pacific	142	203	269
Middle East	160	267	33
Japan	1010	955	928
East Asia	1223	1430	1294
Africa			44
Total	7297**	8718***	7388

* Including the former Soviet Union. ** Production by extraction 7097×10^3 t/a, production by dehydrogenation 200×10^3 t/a.
*** Extraction capacity 8213×10^3 t/a, dehydrogenation capacity 505×10^3 t/a

starting material for more than 95 % of production. In Western Europe, naphtha is the principal raw material for steam cracking. High yields of butadiene are obtainied by this method, which explains the absence of dehydrogenation capacities. In the United States more than 90 % of the butadiene production is based on steam cracking. The available dehydrogenation plants are only used at times of butadiene shortage.

In the United States, Japan, and Western Europe, most of the butadiene rubber is required by the automobile industry. Table 10 shows the worldwide butadiene usage according to products [95]. In the 1950s and 1960s, new steam cracking plants were built worldwide, and the accumulation of butadiene as a byproduct of ethylene production resulted in a steady fall in the price of butadiene. The prices then dramatically rose after the oil crisis of 1973/74 and 1979/80. Figure 4 (see below) shows the effect on the butadiene prices in Western Europe. In 1986 a strong price drop, especially in Deutsche Mark, caused by a weak dollar, and a slightly weakened demand were noted.

Table 10. Butadiene usage in 1998 [95]

Application	Consumption	
	10^3 t/a	%
Styrene – butadiene rubber/latex	3440	46.5
Polybutadiene rubber	1986	27
Nitrile – butadiene rubber/latex	289	4
Polychloroprene rubber	189	2.5
Acrylonitrile – butadiene – styrene polymer	737	10
Adiponitrile	427	6
Others	320	4
Total	7388	100

In the early 1990s an oversupply of butadiene caused by increased cracking capacities led

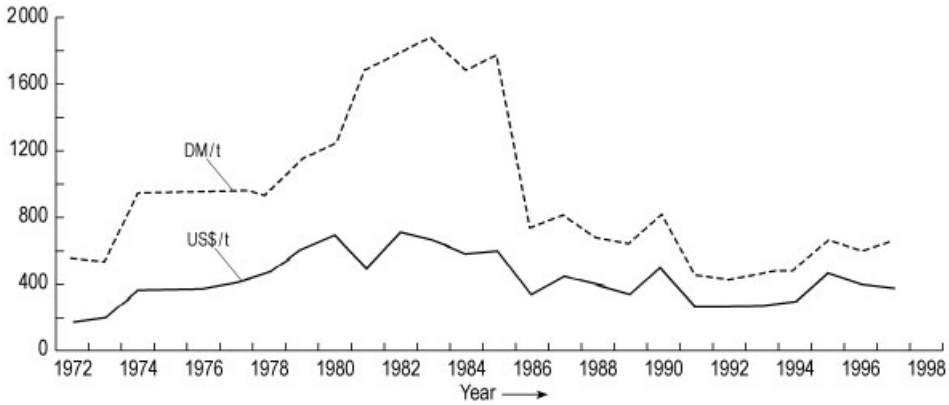

Figure 4. Development of the price of butadiene in Western Europe, 1972 till first quarter of 1998

to a further price drop. This resulted, for example, in the hydrogenation of butadiene surplusses and use as cracker feedstock. Co-cracking of butadiene-containing C_4 streams was also used. Only with the increasing butadiene demand in the mid-1990s did a slight price relaxation take place. Up to 2010, an annual growth rate of 3 % worldwide is estimated for butadiene.

Only increasing butadiene prices could stimulate the manufacture of butadiene by the dehydrogenation and oxidative dehydrogenation methods. Dehydrogenation plants require ready access to energy and raw materials and high butadiene prices if they are to compete with the butadiene from steam crackers.

8. Toxicology

Workers usually are exposed to butadiene by inhalation. Contact of liquid butadiene with skin may exceptionally occur.

8.1. Experimental Toxicity Data

Acute Toxicity. Butadiene is of low acute toxicity. Lethal concentrations are: LC_{50} 12.5 vol % (rat, inhal., 2 h) and 11.5 – 12 vol % (mouse, inhal., 2 h) [96], [97]. Like some other hydrocarbons butadiene causes narcotic effects after inhalation of high concentrations, sometimes preceded by excitation and hyperventilation [98].

Repeated Short-Term Exposure. Rats, guinea pigs, rabbits, and one dog were exposed to butadiene at concentrations of 0, 600, 2300, and 6700 ppm for 7.5 h per day, 6 days a week during 8 months with no significant progressive injury occurring; only the highest concentration slightly retarded the growth and caused light cloudy swelling of liver in some cases. The data on fertility, hematology, body weights, blood biochemistry, urinalysis, and histopathology did not reveal any adverse effect to the animals [98].

In a similar study on Sprague-Dawley rats no clinical sign of toxicity was observed with the exception of moderately increased salivation, particularly at higher concentrations (4000, 8000 ppm). No adverse effects on growth rate, food consumption, neuromuscular function, hematological and biochemical parameters were recorded in the second, sixth, or thirteenth week of the test. Histopathologic investigations of the rats also did not demonstrate any change [99].

Mice were exposed to butadiene at levels of 0, 625 – 8000 ppm during 15 days or 14 weeks with no effect on survival or pathology during 15 days. Only the body weight decreased slightly at the 8000 ppm level. Exposure for 14 weeks resulted in dose dependent lower body weights and markedly increased mortality. No other effect has been reported [100].

Long-term Exposure. Groups of Sprague-Dawley CD rats were exposed to butadiene in air (0, 1000, 8000 ppm) on 6 h/day, 5 days/week for up to 105 (females) or 111 (males) weeks. The effects were: transient nasal secretion, slight ataxia, early increased mortality (at 8000 ppm concentration), suggestive higher frequency of spontaneous nephropathy in males (8000 ppm), higher liver weight, and a slightly increased incidence of metaplasia in the lung (males at 8000 ppm). Various tumors were identified with higher incidence in the exposed groups, others appeared earlier than in controls [101].

The authors concluded that butadiene was a "weak oncogen" under the conditions of this experiment. It is, however, questionable whether the small increase in mostly benign tumors and tumors with high background incidence like mammary tumors are biologically significant at these extremely high exposure levels.

Groups of $B6C3F_1$ mice were exposed to butadiene in air (0, 625, 1250 ppm) on 6 h/day, 5 days/week. Following week 60 for males and week 61 for females the study was terminated because of excessive mortality. Toxic and proliferative lesions of the nasal cavity occurred at increased incidence in the 1250-ppm group, but no neoplastic lesions were observed in the nasal cavity. Butadiene was also associated with liver necrosis as well as ovarial and testicular atrophy.

There was an early onset and a high incidence of several tumors including malign ones in the exposed mice. The authors concluded that there was clear evidence of carcinogenicity for butadiene. Mice obviously are much more sensitive to butadiene than rats are with respect to toxicity and carcinogenicity [100].

Mutagenicity. Gaseous butadiene was mutagenic in the Ames test only in the presence of liver microsomal enzyme preparations. This suggests the formation of a mutagenic intermediate (see Section 8.2) [102], [103].

Embryotoxicity and Teratogenicity. Groups of rats were exposed from day 6 through day 15 of gestation to various concentrations of butadiene [104]. There was an effect on body weight of the dams. The fetuses showed reduction of fetal weight at 8000 ppm (highest concentration) and an increased incidence of normally occurring variations, especially in wavy ribs, indicating retarded development, most prominent in the high-level animals. Major malformations occurred in two pups of one litter of the 8000-ppm group only. The 1000-ppm group did not exhibit higher incidences of malformed fetuses [104].

8.2. Metabolism

In vitro. 1,2-Epoxy-3-butene is formed when butadiene is incubated with liver microsomes and NADPH. Further metabolism involving epoxide hydratase yields 3-butene-1,2-diol [105–109]. The diepoxybutane and epoxybutanediol as well as conjugation products are postulated as further intermediates:

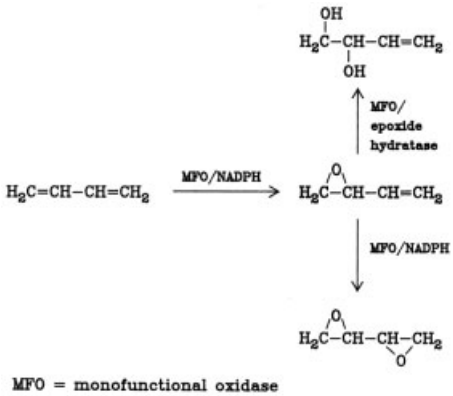

The formation of 1,2-epoxy-3-butene is significantly more pronounced in mouse liver tissue than in liver cells from rats, rhesus monkeys, and humans [109]. The experiments using lung tissues of the same species demonstrated the very rapid formation of 1,2-epoxy-3-butene in mice. However, rat lung preparations were capable of producing only one tenth of the amount seen in mouse lung preparations, whereas the epoxide could not be detected using monkey or human lung tissue. The formation of diepoxybutane could not be detected in any experiment [109].

In vivo. When rats were exposed to butadiene in a closed system, 1,2-epoxy-3-butene also was detected in the exhaled air [108], [109]. Mice in a similar experiment exhaled more of the monoepoxide than rats.

8.3. Human Data

Adverse effects after occupational exposure to butadiene are not reported in the literature. Volunteers exposed to a very high level of 8000 ppm of butadiene in air complained of eye irritation, blurring of vision, coughing, nasal congestion, and drowsiness. Subsequent repeated exposures did not suggest any cumulative effect [98], [110].

A hematology survey of workers at a styrene – butadiene rubber plant showed no pronounced evidence of hematological abnormalities in the peripheral blood. Exposure to butadiene was up to an average of 20 ppm in the tank farm area; in other departments less than 2 ppm occurred [111].

A retrospective study was conducted at two styrene – butadiene rubber plants involving 2756 male workers with an average duration of employment of approx. 10 years. No significant excess of total or cause-specific mortality has been found [111], [112]. Another mortality study covered a period of 36 years and reviewed a population of approx. 14 000 workers from styrene – butadiene rubber production facilities [114]. The study indicated that the overall mortality was lower as compared to general population (Healthy Worker Effect). There were no statistically significant differences in tumor mortality in total or for any specific cause of death.

Hygiene standards [115–117]. As a result mainly from the long-term mice study, butadiene is listed as an A 2 carcinogen with a TLV of 10 ppm by the ACGIH (1984/1985). The German MAK committee has withdrawn the MAK value, and butadiene is now listed in category

A 2 of the 1984 MAK list; a TRK value (lowest technically feasible level) will be established as a consequence.

The in vitro studies on the formation of critical intermediates in tissues from different species (mice, rat, monkey, man), however, suggest that the data from the long-term rat and especially from the mice study may not considered appropriate for direct use in risk evaluation for man.

9. References

1. G. S. Whitby: *Synthetic Rubber,* John Wiley & Sons, New York, 1954, Chapter 2.
2. Kühn, Birett: *Merkblätter Gefährliche Arbeitsstoffe*, 107. Erg.-Lfg. 5/98 – B 059-1.
3. D. R. Lide: *CRC Handbook of Chemistry and Physics,* 79th edition, CRC Press, New York, 1998.
4. C. L. Yaws et al., *Chem. Eng.* **83** (1976) no. 17, 79.
5. EC Material Safety Data Sheet Butadiene.
6. Am. Chem. Soc.: *Azeotropic Data III* (Advances in Chem. Ser. 116), Washington, D.C., 1973.
7. J. P. Kennedy et al.: *Polymer Chemistry of Synthetic Elastomers,* Interscience, New York 1969.
8. Du Pont, US 3 853 948, 1974; US 4 328 172, 1982; US 4 714 773, 1987;
9. Du Pont, US 3 461 167, 1969.
10. Mitsubishi Chem., DE-OS 2 345 160, 1974; DE-OS 2 424 539, 1974; DE-OS 2 505 749, 1975; DE-OS 2 510 088, 1975; DE-OS 2 510 089, 1975.
11. Y. Tanabe, *Hydrocarbon Process.* **60** (1981) no. 9, 187.
12. BASF, DE-OS 2 444 004, DE-OS 2 454 768 , 1976.
13. Toyo Soda, US 3 720 704, 1973.
14. Eastman, US 5 117 012, 1992.
15. J. Hamer: *1,4-Cycloaddition Reactions, The Diels-Alder-Reaction in Heterocyclic Synthesis,* Academic Press, New York 1967.
16. J. Tkatchenko, *J. Organomet. Chem.* **124** (1977) no. 3, C 39.
17. Esso Res. & Eng., US 3 454 665, 1969.
18. Borg-Warner, DE-OS 2 000 974, 1970.
19. Union Chimie Elf-Aquitaine, DE-OS 2 350 689, 2 350 690, 1974.
20. Maruzen Oil, US 3 502 736, 1970.
21. C. Wilke et al., *Angew. Chem.* **69** (1957) 397; **71** (1959) 574.
22. H. Staudinger, B. Ritzenthaler, *Ber. Dtsch. Chem. Ges.* **68** (1935) 455.
23. L. R. Drake et al., *J. Am. Chem. Soc.* **68** (1946) 2521.
24. R. Baker, *Chem. Rev.* **73** (1973) 487.
25. E. J. Smutny, *J. Am. Chem. Soc.* **89** (1967) 6793.
26. A. D. Josey, *J. Org. Chem.* **39** (1974) no. 2, 139.
27. T. Anstock et al., *Ind. Eng. Chem. Prod. Res. Dev.* **21** (1982) 415.
28. P. Roffio et al., *J. Organomet. Chem.* **55** (1973) 405.
29. C. U. Pittman, Jr., R. M. Hanes, *Ann. N.Y. Acad. Sci* . 239 (1974) 76.
30. Celanese, US 4 269 781, 1981.
31. Shell Oil, US 5 030 792, 1991.
32. Kuraray, EP 0 296 550, 1988.
33. T. Prinz, B. Drießen-Hölscher, *Angew. Chem. Int. Ed. Engl.* **35** (1996) 1708 – 1710.
34. A. Behr, R. He, K.-D. Juszak, C. Krüger, Y.-H. Tsay, *Chem. Ber.* **119** (1986) 991 – 1015.
35. G. Egloff, G. Hulla, *Oil Gas. J.* **41** (1942) no. 26, 40.
36. G. S. Whitby: *Synthetic Rubber,* J. Wiley & Sons, New York 1954, p. 86.
37. A. Talalay, M. Magat: *Synthetic Rubber from Alcohol,* Interscience, New York 1945.
38. R. G. Craig, J. M. Duffalo, *Chem. Eng. Prog.* **75** (1979) no. 2, 62.
39. R. G. Craig, E. A. White, *Hydrocarbon Process.* **59** (1980) no. 12, 111.
40. *Hydrocarbon Process.* **55** (1976) no. 9, 229.
41. H. A. Foster, *Hydrocarbon Process.* **52** (1973) no. 9, 119.
42. J. C. Reidel, *Oil Gas J.* **54** (1957) Nov. 11, 166, Dec. 2, 88, Dec. 9, 114, Dec. 16, 110, Dec. 23, 74.
43. K. K. Kearby: "Catalytic Dehydrogenation," in: *Catalysis*, vol. 3, Reinhold, New York 1955.
44. J. C. Reidel, *Oil Gas J.* **54** (1957) Dec. 9, 114.
45. P. M. Reilly, *Chem. Can.* **5** (1953) no. 3, 25.
46. T. H. Arnold, *Chem. Eng.* **68** (1961) Oct. 30, 90.
47. R. J. Harbour, 5th World Petr. Congr., Sect. IV, Paper 10, 121, New York 1959.
48. H. Kröper et al., *Erdöl Kohle Erdgas Petrochem.* **22** (1969) no. 10, 605.
49. V. Fattore et al., *Hydrocarbon Process.* **60** (1981) no. 8, 101.
50. J. A. Convers et al., *Hydrocarbon Process.* **60** (1981) no. 3, 95.
51. W. Krönig, G. Scharfe, *Erdöl Kohle Erdgas Petrochem.* **19** (1966) no. 7, 497.

52. D. Stadermann, *Chem. Techn.* **35** (1983) no. G, 290.
53. D. S. Alexander, 7th World Petr. Congr. Proc., vol. 5, Elsevier, London 1967.
54. P. A. Batist et al., *J. Catal.* **7** (1967) 33, 12 (1968) 45.
55. L. M. Welch et al., *Hydrocarbon Process.* **57** (1978) no. 11, 131.
56. T. Hudson, Jr. et al., *Hydrocarbon Process.* **55** (1974) no. 6, 133.
57. P. C. Husen et al., *Oil Gas J.* **69** (1971) Aug. 2, 60.
58. C. Wolf, *Chem. Week* 1966, May 28, 113.
59. R. B. Stobaugh, *Hydrocarbon Process.* **46** (1967) no. 6, 141.
60. H. Kuper, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 3, 119.
61. J. G. Freiling, A. A. Simone, *Oil Gas J.* **71** (1973) Jan. 1, 25.
62. H. Ritzer, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 3, 124.
63. L. F. Hatch, S. Matar, *Hydrocarbon Process.* **57** (1978) no. 3, 129.
64. B. Schleppinghoff, *Erdöl Kohle Erdgas Petrochem.* **27** (1974) 240.
65. F. Asinger: *Die petrochemische Industrie,* Akademie-Verlag, Berlin 1971, pp. 418 – 422.
66. H. K. Kroper et al., *Hydrocarbon Process.* **41** (1962) no. 11, 191.
67. H. Lauer, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) no. 6, 249.
68. M. Derrien et al., *Hydrocarbon Process.* **58** (1979) no. 5, 175.
69. G. Ritzert, W. Berthold, *Chem. Ind. Techn.* **45** (1973) no. 3, 131.
70. W. A. Gorschkow et al., *Khim. Prom.* **11** (1971) 807.
71. C. Buell, *Ind. Eng. Chem.* **39** (1947) no. 6, 695.
72. W. W. Coogler, Jr., *Hydrocarbon Process.* **46** (1967) no. 5, 166.
73. B. Hausdoerfer et al., *Chem. Ind. Techn.* **40** (1968) no. 23, 1147.
74. H. Klein et al., *Hydrocarbon Process.* **47** (1968) no. 11, 135.
75. U. Wagner, H. M. Weitz, *Ind. Eng. Chem.* **62** (1970) no. 4, 43.
76. K. Volkamer et al., *Erdöl Kohle Erdgas Petrochem.* **34** (1981) no. 8, 343.
77. Japanese Geon Co., US 3 436 436, 1969; US 3 436 438, 1969; GB 1 177 040, 1970.
78. S. Takao, *Hydrocarbon Process.* **45** (1966) no. 11, 151.
79. A. Lindner, *Chem. Ing. Techn.* **55** (1983) no. 1, 68.
80. S. Griffith, R. G. Keister, *Hydrocarbon Process.* **49** (1970) no. 9, 323.
81. J. H. Buehler et al., *Chem. Eng.* **77** (1970) Sept. 7, 77.
82. H. Asatani, W. Hayduk, *Can. J. Chem. Eng.* **61** (1983) Apr., 227.
83. H. Abrevaya, B. V. Vora and E. H. Page, M. J. Banach, "Selective Hydrogenation of C_4 Acetylenes by the KLP Process", in DMGK Tagungsbericht 9705: *C_4 Chemistry – Manufacture and Use of C_4Hydrocarbons*, pp. 99 – 106.
84. J. R. Adams, T. P. Hickey, "Catalytic Distillation for the Processing of Light Olefins, Particularly Etherification and Hydrogenation", in DMGK Tagungsbericht 9705: *C_4 Chemistry – Manufacture and Use of C_4 Hydrocarbons*, pp. 51 – 56.
85. A. Hahn et al., *Hydrocarbon Process.* **54** (1975) no. 2, 89.
86. Mitteilungsblatt Berufsgenossenschaft Chem. Ind., Jedermann-Verlag, Heidelberg 1980, Jan.
87. D. G. Hendry et al., *Ind. Eng. Chem. Prod. Res. Dev.* **7** (1950) 136.
88. R. F. Robey et al., *Ind. Eng. Chem.* **36** (1944) no. 1, 3.
89. Erdölchemie, DE-OS 2 051 548, 1972.
90. M. S. Kharasch et al., *Ind. Eng. Chem.* **39** (1947) 830.
91. G. Whitby: *Synthetic Rubber,* J. Wiley & Sons, New York 1954, p. 80.
92. Pittsburgh – Des Moiness Steel Co., *Chem. Metallurg. Eng.* 1942, Nov., 117.
93. J. H. Boyd, *Ind. Eng. Chem.* **40** (1948) no. 9, 1703.
94. Petroquimica Argentina, Informative Bulletin and Specification, no. 1, "Butadiene".
95. Parpinelli Tecnon, Milano, 1999.
96. I. B. Batkina, *Gig. Sanit.* **31** (1966) 18 – 22.
97. B. B. Shugaer, *Arch. Environ. Health* **18** (1969) 878–882.
98. C. P. Carpenter, C. B. Shaffer, C. S. Weil, H. F. Smyth, Jr., *J. Ind. Hyg. Toxicol.* **26** (1944) 69 – 78.
99. C. N. Crouch, D. H. Pullinger, I. F. Gaunt, *Am. Ind. Hyg. Assoc. J.* **40** (1979) 796 – 802.
100. Toxicology and Carcinogenesis Studies of 1,3-Butadiene in B6C3F1 Mice (Inhalation Studies), National Toxicology Program, Technical Report Series no. 288 (1984).
101. Hazleton Laboratories Europe, The Toxicity and Carcinogenicity of Butadiene Gas Administered to Rats by Inhalation for Approximately 24 Months, Final Report, vol. 1 – 4, Report no. 2653-522/2, Hazleton Labs., Harrogate, England, 1981.

102. C. De Meester, F. Poncelet, M. Roberfroid, M. Mercier, *Toxicol. Lett.* **6** (1980) 125 – 130.
103. F. Poncelet, C. de Meester, M. Duverger-van Bogaert, M. Lambotte-Vandepaer, M. Roberfroid, M. Mercier, *Arch. Toxicol. Suppl.* **4** (1980) 63 – 66.
104. Hazleton Laboratories Europe: "1,3-Butadiene: Inhalation Teratogenicity Study in the Rat," Final Report, Report no. 2788-522/3, Hazleton Labs., Harrogate, England, 1981.
105. E. Malvoisin, G. Lhoest, F. Poncelet, M. Roberfroid, M. Mercier, *J. Chromatogr.* **178** (1979) 419 – 425.
106. E. Malvoisin, M. Roberfroid, *Xenobiotica* **12** (1982) 137 – 144.
107. E. Malvoisin, M. Mercier, M. Roberfroid, *Adv. Exp. Med. Biol.* **136 A** (1982) 437 – 444.
108. H. M. Bolt, G. Schmiedel, J. G. Filser, H. P. Rolzhäuser, K. Lieser, D. Wistuba, V. Schurig, *J. Cancer Res. Clin. Oncol.* **106** (1983) 112 – 116.
109. U. Schmidt, E. Löser, *Archives of Toxicology,* in press.
110. R. H. Wilson, G. V. Hough, W. E. McCormick, *Ind. Med.* **17** (1948) 199 – 207.
111. H. Checkoway, T. M. Williams, *Am. Ind. Hyg. Assoc. J.* **43** (1982) 164 – 169.
112. T. J. Meinhardt, R. J. Young, R. W. Hartle, *Scand. J. Work Environ. Health* **4** (1978) 240 – 246.
113. T. J. Meinhardt, R. A. Lemen, M. S. Crandall, R. J. Young, *Scand. J. Work Environ. Health* **8** (1982) 250 – 259.
114. G. M. Matanoski, L. Schwartz, J. Sperrazza, J. Tonascia: *Mortality of Workers in the Styrene – Butadiene Rubber Polymer Manufacturing Industry,* Johns Hopkins University, School of Hygiene & Public Health, Baltimore MD, June 1982.
115. TLVs, Threshold Limit Values for Chemical Substances in the Work Environment Adopted by ACGIH for 1984 – 85.
116. DFG: *Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte,* Verlag Chemie, Weinheim 1984.
117. *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* Verlag Chemie, Weinheim 1984.

Butadiene Polymers → Rubber, 3. Synthetic
Butadiene Rubber → Rubber, 3. Synthetic
Butadiene Rubber → Styrene

Butanals

Boy Cornils, formerly Hoechst AG, Frankfurt/M., Federal Republic of Germany

Richard W. Fischer, Celanese Chemicals Europe GmbH, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany

Christian Kohlpaintner, Celanese Chemicals Europe GmbH, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Production 2
4. Stabilization, Testing, and Quality Specifications 4
5. Handling, Storage, and Shipment 5
6. Chemical Reactions and Applications . 5
7. Toxicology 8
8. References 8

1. Introduction

Butanals are saturated aliphatic C_4 aldehydes, C_4H_8O, M_r 72.11. There are two isomers: the straight-chain butyraldehyde [*123-72-8*] (butanal) and thebranched isobutyraldehyde [*78-84-2*] (2-methylpropanal, dimethylacetaldehyde).

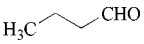

Butyraldehyde

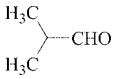

Isobutyraldehyde

Both aldehydes are found in minor concentrations in various essential oils, in plants, as well as in other natural substances [1], [2]. Both butyraldehyde and isobutyraldehyde are highly reactive intermediates that serve as versatile building blocks for many chemical syntheses, predominantly for the preparation of C_4-C_8 compounds.

The annual production volume of butyraldehydes is several million tones worldwide. The total production capacity for *n*-butanal is 6.8×10^6 t with a demand of 5.9×10^6 t. The annual growth rate is estimated to be 2–4 %. Production on an industrial scale is performed exclusively by hydroformylation (→ Oxo Synthesis) [3–5]. World-scale plants such as those operated by BASF, Oxeno (Degussa-Hüls), Celanese, and UCC have an annual output of $(200-600) \times 10^3$t.

2. Physical Properties

n-Butyraldehyde and isobutyraldehyde are colorless, flammable liquids with a pungent penetrating odor. Their vapors have a narcotic effect and irritate the mucous membranes. Both isomers are miscible with organic solvents, such as alcohols, ethers, and benzene. The solubilities of the butyraldehydes in water are collected in Table 1. Some physical data of the butyraldehydes are compiled in Table 2. Table 3 gives the composition of some binary and ternary azeotropes. Further physical properties are provided in [1], [2], and [6].

Table 2. Physical properties of butyraldehydes

	Butyraldehyde	Isobutyraldehyde
Melting point, °C	− 97	− 66
Boiling point, °C	74.8	64
Density at 20 °C, g/cm^3	0.803	0.788
Refractive index n_D^{20}	1.3805	1.3730
Viscosity at 20 °C, mPa s	0.45	0.45
Specific heat capacity at 25–30 °C, J g^{-1} K^{-1}	2.123	2.544
Heat of vaporization at the boiling point, J/g	436	409
Heat of combustion, kJ/mol	− 2478.7	− 2510
Crit. pressure, mPa	4.06	ca. 4.3
Crit. density, g/cm^3	0.259	
Crit. temperature, °C	248	ca. 267
Surface tension at 20 °C, mN/m	24.6	24
Dielectric constant at 20 °C	14.9	13.5
Vapor pressure at 20 °C, kPa	12	17
Dipole moment (vapor), Debye	2.72	
Coefficient of expansion at 20–30 °C, K^{-1}	0.0013	0.0014

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_447

Table 1. Solubility of butyraldehydes in water and of water in butyraldehydes (mass fractions in %)

Temperature, °C	Aldehyde in water		Water in aldehyde	
	Butyraldehyde	Isobutyraldehyde	Butyraldehyde	Isobutyraldehyde
0	9.8	9.3	3.2	2.1
10	8.6	7.8	2.8	2.0
20	7.6	6.7	2.6	1.9
30	6.8	5.8	2.4	1.9
40	6.1	5.0	2.4	2.0
50			2.4	2.1

Table 3. Azeotropic mixtures with butyraldehydes

Compound A	Compound B	Compound C	Boiling point of azeotrope, °C	Concentration of B (C) in the azeotrope, wt %
Butyraldehyde	ethanol		70.7	60.6
Butyraldehyde	water		68	8.8
Butyraldehyde	hexane		60	74
Butyraldehyde	ethanol	water	67.2	11 (9)
Isobutyraldehyde	water		60.5	6.0

3. Production

The only process used for manufacturing butyraldehydes is the hydroformylation (→ Oxo Synthesis) of propylene. The initially formed aldehyde can be hydrogenated to the corresponding alcohol, oxidized to the acid, or converted to other derivatives. Other routes, such as the hydrogenation of crotonaldehyde, prepared by the dimerization of acetaldehyde, and the dehydrogenation of butanols, are obsolete and not used on an industrial scale.

Hydroformylation of Propylene. In the hydroformylation process discovered by Otto Roelen at Ruhrchemie in 1938, a C_n olefin is treated with synthesis gas in the presence of a catalyst to form a C_{n+1} aldehyde. Transition-metal hydridocarbonyl complexes of the general formula $[HM(CO)_x]$ (M = Co, Rh, Ru) serve as a homogeneous catalysts. On an industrial scale, hydroformylation is carried out at 90 – 180 °C and 2.0 – 30 MPa [3–5].

Basically, all olefins, diolefins, or even functionally substituted olefins may be used as raw materials. On an industrial scale alkenes from C_2 to C_{18} are predominantly used. Propylene has gained particular significance, being converted to a mixture of *n*- and isobutyraldehyde. Although the reaction medium is homogeneous in most processes, a two-phase process (water/organic solvent) has also been developed, with advantageous catalyst separation and optimized energy consumption.

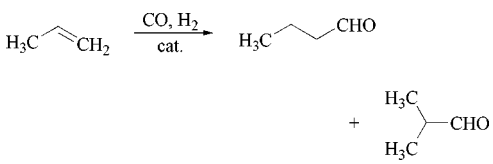

Depending on the oxo catalyst, the reaction conditions, and the process design, mixtures of butyraldehydes with a *n*/iso ratio from 1/1 to 20/1 are obtained. In addition to the aldehydes, which constitute 90 – 98 % of the reaction mixture, minor amounts of byproducts and secondary products of the reactive butyraldehydes are formed, such as butyl formate, butanols, and aldehyde dimers, trimers, and condensation products [3–5].

Oxo Technology [7]**.** The oxo reaction (hydroformylation) is the reaction of synthesis gas (CO/H_2, usually with a 1/1 molar ratio) with an alkene (see also → Oxo Synthesis). Terminal olefins yield a mixture of *n*-aldehydes and 2-methylaldehyde isomers, the so-called isoaldehydes. The reaction is catalyzed by rhodium or cobalt metal complexes. With unmodified hydrido rhodium or cobalt species the *n*/iso ratio is usually poor. By addition of ligands like triphenylphosphine (TPP) in homogeneous systems or trisulfonated triphenyl phosphine

(TPPTS) in water/organic two-phase systems the *n*/iso ratio is significantly enhanced. In particular at low temperatures, at high carbon monoxide partial pressure, and at a high ligand concentrations the formation of the straight-chain isomers is favored [3–5].

Early versions of the oxo process used cobalt catalysts at relatively high pressures (20 – 35 MPa), whereby $[HCo(CO)_4]$ is the active catalyst species. The typical *n*/iso ratio is about 4/1 with propylene as feedstock. By the mid-1900s various companies (e.g., BASF, ICI, Eastman) had developed their own versions of the cobalt process. For short-chain olefins like propylene, the cobalt technology has almost fully been substituted by rhodium-based processes. Only for higher and in particular branched olefins is cobalt still the most widely applied metal. The second-generation, phosphine-modified cobalt hydroformylation catalysts were developed in the 1960s by Shell.

In the 1970s UCC and Celanese developed a low-pressure oxo process (LPO) with Rh as metal and triphenylphosphine (TPP) as ligand. This process operates at a lower pressure and yields a n/iso ratio of about 10/1. Versions of this process with a gas recycle step are commonly known as LP Oxo Mk I (early version) or Mk II. A further proposal is the LPO Mk III process with improved raw material usage and a liquid catalyst recycle. The product/catalyst separation is performed in a separate vessel in which the aldehyde vapors are flashed off and the liquid solution containing the rhodium catalyst is recycled to the reactor. In this operating mode, the aldehyde condensation products serve as a solvent for the catalyst. However, the application of all versions is mostly limited to propylene and butene hydroformylation. Many plants have been converted from high-pressure cobalt to LPO technology for new or additional butyraldehyde (and thus butanol and 2-ethylhexanol) capacity.

By the mid-1980s Ruhrchemie (now Celanese) and Rhône-Poulenc developed a two-phase process with a water-soluble rhodium/TPPTS complex. The major advantages are the simple and mild catalyst recycle, the high *n*/iso ratio of ca. 20/1 and the beneficial low energy consumption of the whole process [3–5], [8].

The fourth generation of oxo catalysts used for the production of butyraldehyde and especially of higher aldehydes (and thus less important for synthesis of butyraldehyd) is known as the LPO Mk IV version. This process applies highly active bis-phosphite-modified rhodium catalysts which are suitable for both propylene and raffinate II feedstocks for butyraldehyde and valeraldehyde, respectively. The high catalyst activity allows this process to be run in single-pass mode with no olefin-recycle requirements. The potential of this process is its flexibility to convert propylene or butenes in the same reactor system.

Process Technology. For homogeneous reaction systems, the standard processing of the crude hydroformylation product, containing *n*- and isobutyraldehyde, rhodium metal, ligand, and byproducts, is generally the distillation of the reaction mixture to separate the aldehydes from the catalyst solution. To minimize catalyst decomposition, phosphine-modified cobalt or rhodium catalysts are separated from the reaction product by fast flash distillation. The less volatile catalysts remain at the bottom and are recycled to the hydroformylation reactor without further treatment. However, optimized catalyst lifetime and thus improved activity and selectivity requires the discharge of a small portion of the catalyst for workup and addition of fresh catalyst.

In contrast to the homogeneous hydroformylation processes, the Ruhrchemie/Rhône-Poulenc process (RCH/RP) is designed on a completely different reaction and catalyst recovery system. In principle the two-phase RCH/RP process applies a stirred tank reactor, followed by a phase separator and a stripping column. The reactor containing the aqueous catalyst solution (Rh/TPPTS) is fed with propylene and synthesis gas. The crude oxo product is passed on to the phase separator, where gaseous components are removed and the liquids separated into two phases, the aqueous catalyst solution and the organic aldehyde phase. The catalyst solution is recycled to the reactor via a heat exchanger system, which generates steam. This results in one of the advantages of the two-phase technology: the high energy efficiency. The P/Rh ratio is kept above 20/1 to ensure an *n*/iso ratio of at least 95/5. Almost no rhodium is leached out of the

water phase. The organic phase containing the crude aldehydes is stripped with synthesis gas to remove dissolved propylene, which is recycled to the reactor. The raw aldehyde mixture is split into iso- and *n*-butyraldehydes in a distillation unit.

Hydrogenation of Crotonaldehyde. Before the discovery of the hydroformylation reaction, the aldol condensation of acetaldehyde to give crotonaldehyde followed by hydrogenation to *n*-butyraldehyde was the most important method of producing *n*-butyraldehyde.

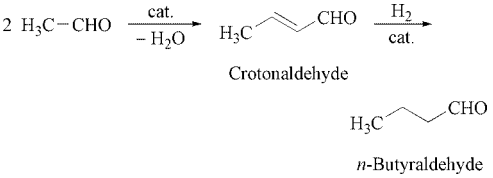

n-Butanol is formed as a byproduct by hydrogenation of the aldehyde moiety during the selective hydrogenation of the double bond of crotonaldehyde. Selective hydrogenation is performed in gas- or liquid-phase processes. Typical catalysts applied are supported catalysts on the basis of Cu, Ni, Pd, Ru, Fe, Cr, or Mn [1], [9–14].

Up to the 1950s this route was still the preferred method for producing *n*-butyraldehyde [15] on an industrial scale, but the convenient one-stage oxo processes replaced crotonaldehyde hydrogenation almost completely. Only countries with massive ethanol production based on natural resources, such as Brazil, might rediscover the crotonaldehyde route for producing *n*-butyraldehyde [15].

Dehydrogenation of Butanols. Butanols formed in other processes, such as Fischer – Tropsch synthesis [16], alcoholic fermentation of sugar with *Clostridium saccharobutylacetonium liquefaciens* or *Bacillus amylobacter* [17], and Reppe carbonylation [18], can be dehydrogenated to butyraldehydes.

$$CH_3CH_2CH_2CH_2OH \rightleftharpoons CH_3CH_2CH_2CHO + H_2$$

The reaction is carried out in the gas phase with a heterogeneous catalyst [19]. The addition of hydrogen acceptors such as O_2, either pure or as air, increases the aldehyde yield but also the formation of byproducts. Zinc- and/or chromium-doped copper catalysts are particularly effective [20], [21].

Fischer – Tropsch synthesis generates small amounts of *n*-butyraldhyde, which is of importance in some local markets [16], [22].

4. Stabilization, Testing, and Quality Specifications

Stabilization. The reactivity of butanals is determined by the polarity of the carbonyl group. The inductive effect of the carbonyl group promotes α-proton abstraction, which allows a broad variety of consecutive intermolecular reactions, with addition type reactions as the most important examples. Thus it is recommended to handle aldehydes during distillative work-up, storage and shipment only under the following precautions:

- Inert gas atmosphere for all reaction and work-up steps
- Storage with exclusion of UV light
- Storage at constant and moderate temperature
- Conversion into thermally stable compounds like acetals, which are easily cleaved back to the aldehydes
- Addition of stabilizers and antioxidants

Testing. The most important method of testing and continuous control of product quality (Table 4) is gas chromatography with a low injection temperature to avoid decomposition of impurities, e.g., aldols.

In the analysis of butyraldehydes it is important to detect trimeric aldehydes, which are formed readily, especially in the absence of stabilizers. Since these trimerization reactions are catalyzed by acids, contact with acidic compounds must be strictly avoided during the course of aldehyde processing. During the analysis of water in butyraldehyde by Karl Fischer titration, special precautions must be taken to avoid condensation to acetals, which is accompanied by formation of water. This is achieved by the use of a solvent mixture composed of four parts pyridine and one part ethylene glycol. The density and refractive index of the water-containing product are strongly affected by the

Table 4. Delivery specifications for butyraldehydes [23]

	Method	Butyraldehyde		Isobutyraldehyde	
		moist	dry	moist	dry
Butyraldehyde, wt %	GLC	$\geqq$ 96.0	$\geqq$ 99.0	$\leqq$ 0.5	$\leqq$ 0.2
Isobutyraldehyde, wt%	GLC	$\leqq$ 0.3		$\geqq$ 95.0	$\geqq$ 99
Density at 20 °C, kg/L	DIN 51757, ASTM D 1298	0.807 – 0.812	0.801 – 0.803	0.790 – 0.794	0.786 – 0.790
Refractive index n_D^{20}	DIN 51423, ASTM D 1747	1.380 – 1.382		1.373 – 1.376	1.372 – 1.374
Color, Pt-Co scale	DIN 53409, ASTM D 1209	$\leqq$ 15	$\leqq$ 10	$\leqq$ 15	$\leqq$ 15
Acid number, mg KOH/g	DIN 53402, ASTM D 1613	$\leqq$ 3.0	$\leqq$ 2	$\leqq$ 3.0	$\leqq$ 2.0
Water, wt%	Karl Fischer	$\leqq$ 3.0	$\leqq$ 0.2	$\leqq$ 2.0	$\leqq$ 0.1
Distillation range (95 vol%) at 101.3 kPa, °C	ASTM D 1078	67 – 80	ca. 74 – 76 ($\triangle t \leqq 2$ °C)	57 – 70	63 – 67

equilibrium hydrate concentration, which is established slowly. During determination of the acid number care must be taken to exclude atmospheric oxygen. The boiling point of the butyraldehydes is strongly affected by traces of water.

Quality Specifications. Table 4 outlines the standard quality specifications for butyraldehydes. The limits are largely determined by the intended use and by specific customer requests.

5. Handling, Storage, and Shipment

The butyraldehydes are stored and shipped in containers of corrosion-resistant steel, aluminum, glass, or ceramic, as well as specially coated tanks of normal steel and drums with removable polyethylene inserts. Even with water contents below 0.1 % and storage under inert gas, *n*-butyraldehyde dissolves iron from steel, causing discoloration of the product.

On contact with air butyraldehyde is oxidized readily to the butyric acids. Therefore, storage under inert gas is mandatory. Even traces of acids catalyze the trimerization of isobutyraldehyde. In general this may be prevented by the addition of a suitable amine (e.g., triethanolamine) or other bases [24].

Owing to the low flash points and ignition temperatures (Table 5) as well as the high volatility of these aldehydes the safety regulations are to be strictly observed.

The butyraldehydes are shipped in tank cars or trucks, tankers, or drums. The classification and designations for the various modes of transport must be observed (Table 5).

6. Chemical Reactions and Applications

Owing to their reactivity butyraldehydes are important starting materials for a broad variety of products [30]. Due to the high reactivity of the carbonyl group, aldehydes undergo many organic chemical reactions. Typical aldehyde reactions are:

- Addition to the C=O bond without substitution of the oxygen atom
- Addition to the C=O bond under substitution of the generated hydroxyl group
- Substitution of the oxygen functionality
- Formation of enol derivatives
- Oxidation
- Reduction
- C–C bond cleavage
- C–C bond formation.

Figure 1 outlines the reactions applied on industrial scale.

Butyraldehydes are important intermediate products for a vast number of C_4 products, including *n*-butanol, *n*-butyric acid, *n*-butylamines, and butyl acrylates, as well as for the plasticizer alcohol 2-ethylhexanol (C_8) and for C_6 products such as 2-ethylbutyric acid. Isobutyraldehyde is converted mainly to isobutanol and neopentyl glycol. Isobutyric acid is a potential intermediate for methacrylic acid (→ Methacrylic Acid and Derivatives, Chap. 3.3.).

Hydrogenation. The butyraldehydes are converted to the corresponding butanols by catalytic hydrogenation, whereby both liquid-phase and gas-phase processes are used. The

Table 5. Safety information for butyraldehydes∗

		Butyraldehyde	Isobutyraldehyde
Flash point, closed cup (DIN 51755)	°C	− 13	< − 18
Lower and upper explosive limit in air at 101.3 kPa	vol%	1.94 – 12.5	1.6 – 11
Autoignition temperature (DIN 51794, ASTM D 2155)	°C	190	165
Odor threshold		low	low
LD_{50} (oral, rat)	g/kg	2.5 – 5.9	2.8 – 3.7
LD_{50} (dermal, rabbit)	g/kg	3.56	3.7 – 7.1
LC_{Lo} (inhal., rat)	mg/15 min	23.9	23.2
LD_{Lo} (subcut., rat)	g/kg	10	
Marking according to Regulations of the European Community	Symbol	F	F∗∗
	R-Phrases	11	11∗∗
	S-Phrases	(2)-9-29-33	9-16-29-33-36∗∗
Hazard classification for transport:			
RID/ADR (Class, Item Number)		3 b	3 b
IMDG-Code (Class, UN No., Packing Gp.)		3.2, 1129, II	3.1, 2045, II
IATA-DGR (ICAO Techn. Instr.) (Class, UN No., Packing Gp.)		3, 1129, II	3, 2045, II
Mail (FRG)		max. 125 cm^3 per receptacle and max. 500 cm^3 per consignment	
DOT (USA)		flammable liquid	

∗ Vapor pressure at 20 °C, solubility of butyraldehydes in water and vice versa: see Tables 1 and 2
∗∗ No EEC regulation; marking corresponds to *n*-butyraldehyde.

mild, highly selective gas-phase hydrogenation is carried out on nickel or copper supported catalysts.

Oxidation. Pure oxygen or air converts butyraldehydes to carboxylic acids in high yields, the reaction taking place in the presence or even absence of a catalyst. Transition metals of Groups 5 – 10, Cu, Ag, Ce, the alkali metals, and alkaline earth metals are suitable catalysts. If oxidation is carried out in the liquid phase, the metal naphthenates or carboxylates are used. In gas-phase processes, generally metal oxides, deposited on carriers such as silica gel, kieselguhr, silicates, quartz, or alumina, are used [1]. For further information, see [1], [31–33], [34].

The use of specific metal oxide catalysts allows isobutyraldehyde to be oxidized at high temperature directly to methacrolein, or methacrylic acid, or a mixture of the two [1], [35–38]. By variation of the reaction conditions, acetone or isopropyl alcohol is accessible [1], [39]. *n*-Butyric anhydride is also produced by oxidation of *n*-butyraldehyde [1].

Addition and Condensation Reactions. In a large number of addition and condensation reactions butyraldehydes react with themselves or other carbonyl compounds, amines, alcohols, nitriles, and other functional groups. On an industrial scale, the most important synthesis of this kind is the homocondensation of *n*-butyraldehyde in the presence of a basic catalyst. The initially formed 2-ethylhexenal is converted by hydrogenation to 2-ethylhexanol, the alcohol component of the most important PVC plasticizer, dioctyl phthalate (DOP) [1], [3–5], [40], [41]. Mild reaction conditions during the aldolization step allow butyraldol (2-ethyl-3-hydroxyhexanal) to be isolated, which can be hydrogenated to 2-ethyl-1,3-hexanediol.

The condensation product of *n*-butyraldehyde, 2-ethylhexenal, also is selectively hydrogenated in the presence of a precious metal catalyst to yield 2-ethylhexanal, which is converted to 2-ethylhexanoic acid or 2-ethylhexylamine [1], [3–5], [40].

Condensation of *n*-butyraldehyde with two moles of formaldehyde and subsequent hydrogenation yields trimethylolpropane [1], [42], [43]. Phenol and *n*-butyraldehyde condense to

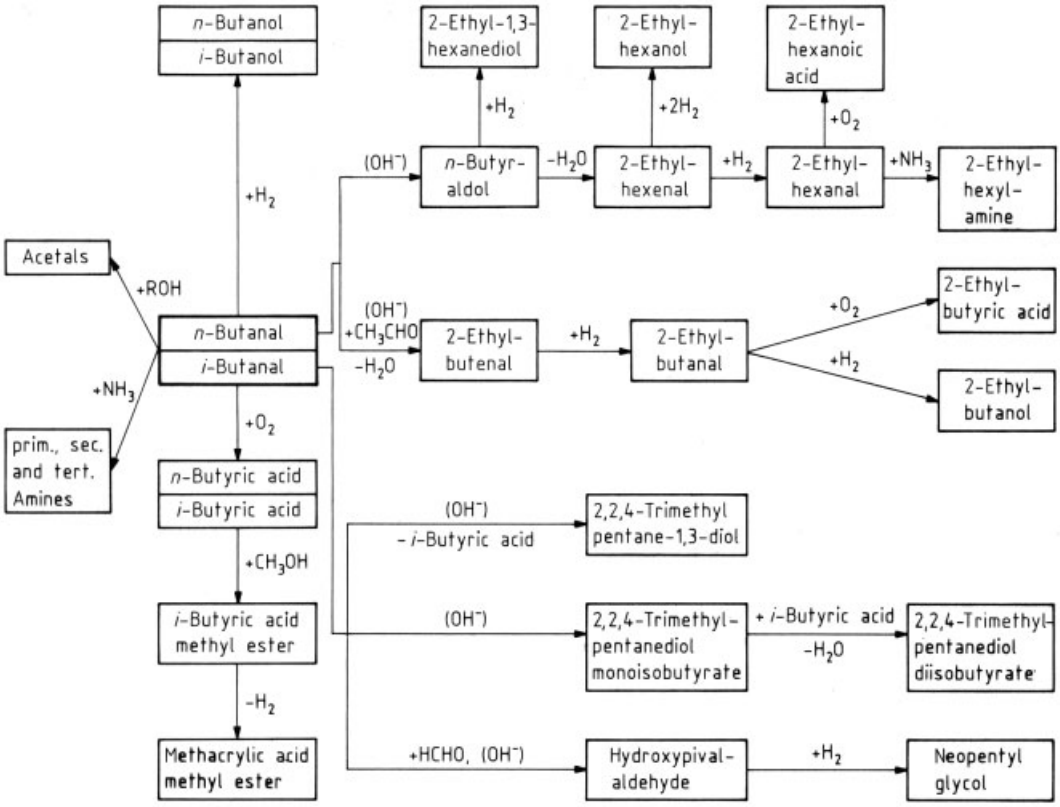

Figure 1. Reactions of butyraldehydes

yield oil-soluble paints, and condensation products with urea are used in alcohol-soluble paints [1].

Condensation of *n*-butyraldehyde with acetaldehyde gives 2-ethylbutyl products [44]; their properties and applications are described in [45].

With poly(vinyl alcohol) *n*-butyraldehyde condenses to poly(vinyl butyral), PVB, which is used in the manufacture of laminated safety glass or as a bonding agent in paints [1].

Condensation products of *n*-butyraldehyde with aniline, butylamine, thiourea, diphenylguanidine, or methylthiocarbamate are used on a large scale as vulcanizing accelerators for rubber [1].

Reductive amination, that is, the reaction with hydrogen and ammonia or amines, provides access to primary, secondary, and tertiary butylamines [1], [46–48].

The homocondensation of isobutyraldehyde leads via the intermediate isobutyraldoxane (4-hydroxy-5,5-dimethyl-2,6-diisopropyl-1,3-dioxane) to isobutyraldol or secondary products such as 2,2,4-trimethylpentane-1,3-diol, 2,2,4-trimethylpentane-1,3-diol mono- and diisobutyrate, and 2,2,4-trimethylpentane-1-ol [1], [49–52].

The reaction of isobutyraldehyde with formaldehyde forms primary hydroxypivalaldehyde, a starting material for products with a neopentyl backbone. Especially neopentyl glycol, obtained by hydrogenation of hydroxypivalaldehyde, has become increasingly important [53], [54]. The wide variety of its uses is described in [55].

D,L-Hydroxy-β,β-dimethylbutyrolactone (pantolactone), is obtained by reaction of isobutyraldehyde, formaldehyde, and sodium cyanide [1], [56], [57].

The reaction of isobutyraldehyde with urea produces slow-release fertilizers (→ Fertilizers, Chap. 4.4.2.2.) [1]. Isobutyraldehyde serves as a

raw material for the syntheses of the amino acids D,L-valine and D,L-leucine [1].

The Tishchenko reaction is used to obtain isobutyl isobutyrate, a specialty solvent. Acetals of the butyraldehydes with various alcohols serve as solvents for cellulose, resins, and rubber [1], [54].

The crossed Tishchenko reaction of *n*-butyraldehyde with acetaldehyde provides commercial ester mixtures. Butyraldehydes and hydroxylamine yield the corresponding aldoximes or their salts [1], [54].

Other Reactions. Isobutyraldehyde can be isomerized to *n*-butyraldehyde and methyl ethyl ketone or converted to diisopropyl ketone [1], [54]. The chlorination of butyraldehydes yields fragrances and pharmaceutical intermediates [1], [54], [58–60]. *n*-Butyraldehyde is used as a catalyst component for special polymerization reactions [61–64]. Isobutyraldehyde can be converted to isobutene on special molybdenum catalysts [65].

Additional reactions and applications can be found in [1], [54] and references therein.

7. Toxicology

Butyraldehydes have low acute toxicity (Table 5). Exposure limits have not yet been established. Butyraldehyde liquids and vapors strongly irritate the skin, eyes, and respiratory organs, possible due to rapid oxidation to the carboxylic acids on contact with air. The symptoms are coughing, watering eyes, and a burning sensation on the skin and in the nose and throat. On exposure to higher doses, the CNS is impaired, resulting in depression, dizziness, headaches, and a slight narcotic effect. The repugnant smell of butyraldehydes occasionally causes nausea and vomiting. Subcutaneous injection of *n*-butyraldehyde causes hemolysis.

8. References

1. *Ullmann,* 4th ed., **9,** p. 42; *Ullmann,* 5th ed., **A4,** p. 447.
2. *Kirk-Othmer,* 3rd. ed., **vol. 4,** p. 376.
3. B. Cornils in J. Falbe (ed.): *New Syntheses with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980.
4. B. Cornils, E. Wiebus, *CHEMTECH* **25** (1995) no. 1, 33.
5. B. Cornils, *Org. Proc. Res. &. Dev.* **2** (1998) no. 2, 121.
6. J. McGarry, *Ind. Eng. Chem.* **22** (1983) 313.
7. C. D. Frohning, C. W. Kohlpaintner in B. Cornils, W. A. Herrmann (eds.): *Homogeneous Catalysis with Organometallic Compounds* , VCH, Weinheim, 1996, pp. 29 – 104.
8. B. Cornils, W. A. Herrmann (eds.): *Aqueous-Phase Organometallic Catalysis,* Wiley-VCH, Weinheim 1998.
9. V. N. Kulakov et al., SU 172744, 1964.
10. Usines de Melle, DE 814444, 1949.
11. BASF AG, DE-OS 2839474, 1980; DE-OS 3130805, 1983 (M. Horner et al.).
12. V. Macho, CS 131903, 1966.
13. L. S. Pakhomova et al., SU 335227, 1969.
14. Chem. Werke Hüls, DE-OS 3151086, 1982 (L. Fischer et al.).
15. O. Horn in *Ullmann,* 3rd ed., **vol. 4,** p. 796.
16. H. Schulz, J. H. Cronjéin *Ullmann,* 4th ed., p. 329.
17. B. Cornils, E. Zilly in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** p. 88.
18. see [15], p. 20.
19. L. Bexten, J. Weber in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** p. 281.
20. C. D. Frohning, G. Horn in J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive,* G. Thieme Verlag, Stuttgart 1978, p. 63.
21. Hoechst AG: *Katalysatoren Hoechst,* Frankfurt 1979, p. 44.
22. C. D. Frohning et al. in J. Falbe (ed.); *Chemierohstoffe aus Kohle,* G. Thieme Verlag, Stuttgart 1977, p. 289.
23. Hoechst AG: *Delivery specifications,* Frankfurt 1984. Chemische Werke Hüls AG: *Data sheet for n-butyraldehyde, dry,* 1983.
24. Ruhrchemie AG, DE 2905267, 1981 (J. Weber, V. Falk, C. Kniep).
25. Allied Chem. Corp., US 2808429, 1954.
26. Universal Oil Prod. Co., US 2683174, 1949.
27. V. M. Polyakov et al., *Khim. Promst. (Moscow)* 1972, no. 10, 48, 743; *Chem. Abstr.* **78** (1973) 42812.
28. W. Rupilius in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** p. 325.
29. J. Simonik, L. Beranek, *J. Catal.* **24** (1972) no. 2, 348.
30. C. E. Loeffler, L. Stautzenberger, I. D. Unruh, *Encycl. Chem. Process. Des.* 1977, no. 5, 358 – 405.

31. Mitsubishi Chemical Industries Co. Ltd., DE-OS 2928002, 1980 (T. Masuko, S. Fukaya, N. Murai, J. Noma); JP 78105412, 1977 (T. Maki); JP 78105413, 1977 (T. Maki); JP 7909203, 1977 (T. Maki); JP 78108915, 1977 (T. Maki).
32. Chisso Corp., JP 7378116, 1973 (M. Fukin, T. Hirai, T. Ohashi, S. Furukawa, I. Koga).
33. A. Benning, L. Mußler, H. Tummes in F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** p. 558.
34. B. Lappe, E. Schulz, in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **vol. 1,** VCH Verlagsgesellschaft, Weinheim 1996, p. 424.
35. Mitsubishi Rayon Co., Ltd., JP 7882715, 1978 (H. Matsuzawa, M. Kobayashi, H. Ishii, M. Kato); JP 7214085, 1972 (T. Kita, C. Ishii); DE-OS 2633593, 1978 (T. Onoda, M. Otake).
36. Union Carbide Corp., DE-OS 2836309, 1979 (D. W. McNeil, B. Phillips).
37. Chisso Corp., JP 7946705, 1979 (M. Fukui, N. Otake, N. Nagata).
38. Johnson, Matthey PLC, GB 2094782, 1982 (E. Shutt).
39. Rhône-Progil, US 3987103, 1976 (G. Gobrou, C. Falize, H. Dufour).
40. B. Cornils, A. Mullen, *Hydrocarbon Process.* 1980 (11) 93.
41. Ruhrchemie AG, DE-OS 2713434, 1978 (H. Tummes, H. Noeske, B. Cornils, W. Kascha); DE-OS 2437957, 1976 (G. Kessen).
42. Bayer AG, DE-OS 2702582, 1978 (O. Immel, H. H. Schwarz, O. Weissel, H. Krimm).
43. L. Cairati et al., *Chim. Ind. (Milan)* **63** (1981) no. 11, 723 – 725.
44. Mitsubishi Chemical Industries Co., Ltd., JP 7424891, 1974 (A. Matsukuma, I. Takakishi, K. Yoshida); JP 7406887, 1974 (J. Takakishi, K. Yoshida).
45. B. Cornils, H. Feichtinger, *Chem. Ztg.* **101** (1977) 107 – 117.
46. Texaco Inc., US 4299985, 1981 (J. F. Knifton, P. H. Moss).
47. BASF AG, DE-OS 2725669, 1978 (K. Merkel et al.).
48. B. Cornils, H. Feichtinger in J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive,* G. Thieme-Verlag, Stuttgart 1978, p. 45.
49. BASF AG, DE-OS 820518, 1979 (F. Merger, H. J. Förster).
50. Eastman Kodak Co., US 4225726, 1980 (D. L. Morris, A. W. McCollun).
51. Chisso Corp., JP 7715582, 1977 (J. Tsuchiya et al.); JP 7323411, 1973 (R. Takakashi, J. Tsuchiya); JP 7334574, 1973 (S. Matsumoto, K. Gunji, T. Hamahata).
52. Hoechst AG, DE-OS 3102826, 1982 (H. Baltes, E. J. Leupold).
53. Eastman Kodak Co., DE-OS 2652224, 1977 (B. W. Palmer, H. N. Wright).
54. H. J. Hagemeyer, G. DeCroes: *The Chemistry of Isobutyraldehyde and its Derivatives,* Tennessee Eastman Comp., Kingsport, Tenn., 1953.
55. B. Cornils, H. Feichtinger, *Chem. Ztg.* **100** (1976) 504 – 514.
56. BASF AG, DE-OS 2758883, 1979 (H. Distler, W. Goetze).
57. VEB-Jenapharm, GB 1345459, 1974 (J. Schmidt et al.); DE-OS 2228641, 1973 (J. Schmidt, H. Grunert, C. Weigelt).
58. Allied Chemical Corp., US 4096187, 1978 (J. H. Bonfield, A. Murthy, D. Pickens).
59. Dow Chemical Co., US 3801645, 1974 (D. A. Dalman).
60. Shell International Research Maatschappij, DE-OS 2641356, 1977 (A. Reinink, J. Grendelman).
61. Daicel Ltd., JP 7411480, 1974 (A. Tanaka et al.).
62. Thiokol Chemical Corp., DE 1545025, 1974 (R. M. Gobran, S. W. Osborn).
63. Japan Synthetic Rubber Co. Ltd., DE-OS 2444681, 1975 (M. Ikeda et al.).
64. Ube Industries Ltd., DE-OS 2445776, 1975 (H. Ueno et al.).
65. Toa Nenryo Kogyo K. K., JP 7748603, 1977 (K. Kaneko, H. Furukawa).

Butanediols, Butenediol, and Butynediol

Heinz Gräfje, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1.1)

Wolfgang Körnig, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1.2)

Hans-Martin Weitz, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1.3)

Wolfgang Reiss, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1.3)

Guido Steffan, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 2.1)

Herbert Diehl, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 2.1)

Horst Bosche, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2.3)

Kurt Schneider, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2.3)

Heinz Kieczka, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 3)

1. 1,4-Diols 1
1.1. 2-Butyne-1,4-diol 1
1.2. 2-Butene-1,4-diol 3
1.3. 1,4-Butanediol 4
2. Other Butanediols 6
2.1. 2,3-Butanediol 6
2.2. 1,3-Butanediol 7
2.3. 1,2-Butanediol 7
3. Toxicology 7
4. References 8

1. 1,4-Diols

1.1. 2-Butyne-1,4-diol

Physical Properties. 2-Butyne-1,4-diol [*110-65-6*], $HOCH_2-C{\equiv}C-CH_2OH$, M_r 86.09, *mp* 58 °C, *bp* 150 °C (at 1.8 kPa), is a colorless, hygroscopic, orthorhombic crystalline compound, which is readily soluble in water (374 g in 100 g at 25 °C) and in strongly polar solvents, e.g., alcohol, acetone, but only slightly soluble in ether and almost insoluble in hydrocarbons.

Heat of vaporization	50.28 kJ/mol, 584 kJ/kg (at 101.3 kPa)
Heat of combustion	2204 kJ/mol, 25.5 kJ/g
Flash point (DIN 51758)	152 °C
Ignition point	335 °C

Chemical Properties. 2-Butyne-1,4-diol decomposes relatively slowly between 160 and 200 °C, but at temperatures significantly above 200 °C violent decomposition can occur. Explosive decomposition can be caused or accelerated by alkali hydroxide, strong anhydrous acids, and certain heavy-metal salts, particularly those of mercury. 2-Butyne-1,4-diol is sensitive to oxidation.

Substitution of the hydroxyl groups is analogous to that of other diols. In the case of substitution by halogens very reactive compounds are formed. The reaction with bifunctional compounds leads to linear polymers only. The carbon – carbon triple bond undergoes the usual addition reactions of acetylene. Two moles of chlorine, but only one mole of bromine or iodine, are added per mole of diol. Chlorination in the presence of hydrochloric acid gives 2,3-dichloro-3-formylacrylic acid (mucochloric acid) [*87-56-9*]:

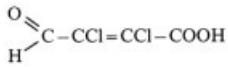

The addition of water, catalyzed by mercury salts, leads to the rather unstable 2-oxobutanediol.

Production. The synthesis of 2-butyne-1,4-diol, based on the work of Reppe and his coworkers [1–3], is carried out by reaction under pressure of acetylene and an aqueous solution of formaldehyde, catalyzed by copper acetylide:

$$2\,CH_2O + HC{\equiv}CH \longrightarrow HOCH_2-C{\equiv}C-CH_2OH \quad \Delta H = -100.5\ \text{kJ/mol}$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_455

This reversible reaction proceeds via propargyl alcohol, 2-propyn-1-ol [*107-19-7*], which mostly remains on the catalyst and thus only-partially goes into solution. The reaction is first order in respect to formaldehyde and, in industrial processes, effectively zero order in respect to acetylene.

The prime objective in the synthesis is to minimize the formaldehyde concentration in the resulting butynediol solutions. In principle, it is possible to react rather concentrated formaldehyde solutions partially, distill off the unreacted formaldehyde under pressure, and reintroduce it into the reactor. However, this process is complex and detrimental to the catalyst. Usually, the production of the butynediol is carried out continuously in cascades of three to five reactors. In these reactors the catalyst is present either as a fixed bed of strands 3 – 5 mm in diameter and 10 mm in length (Fig. 1) or as a fluidized bed of particles of diameter 0.1 – 2 mm. In the first case the formaldehyde solution, together with gaseous acetylene, is fed from the top [4]. In the fluidized-bed process the required amount of acetylene is dissolved in the solution and the solution is then fed from the bottom, suspending the catalyst (Fig. 2) [5].

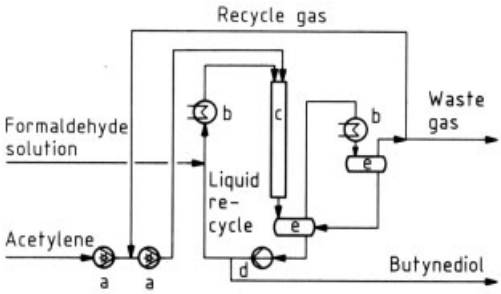

Figure 1. Synthesis of butynediol, fixed-bed reactor
a) Liquid seal pump; b) Cooler; c) Reactor; d) Recycle pump; e) Gas – liquid separator

In either case the reactors are equipped with liquid-circulating systems. In the fixed-bed process the volume of the liquid is determined by external cooling. In the fluidized-bed process the volume of the circulating liquid is determined by the amount of acetylene which is introduced by absorbing columns or injectors.

The catalyst is 3 – 6 % bismuth(III) oxide and 10 – 20 % copper(II) oxide mostly supported on silica. Copper(II) oxide is converted by acetylene and formaldehyde at 60 – 90 °C to copper(I) acetylide, which complexes further acetylene [6] and thus forms the catalyst proper. Bismuth oxide inhibits the formation of water-insoluble polymers, the "cuprenes", from oligomeric acetylene complexes. Conditions of synthesis: 80 – 100 °C; 30 – 50 % aqueous formaldehyde solution; partial pressure of acetylene 2 – 6 bar, pH 5 – 8. Dilute alkali is added to the cycle solution to adjust the pH.

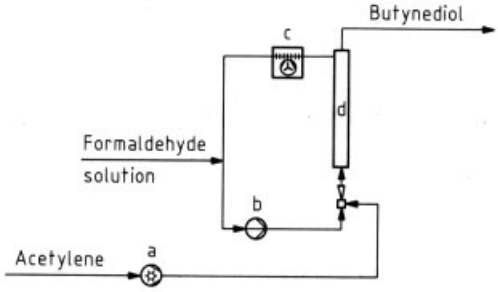

Figure 2. Synthesis of butynediol, fluidized-bed reactor
a) Liquid seal pump; b) Recycle pump; c) Air cooler; d) Reactor

The product solution contains 33 – 55 % butynediol, 1 – 2 % propargyl alcohol, 0.4 – 1 % unreacted formaldehyde, and 1 – 2 % byproducts. These byproducts are mainly high-boiling materials, along with some sodium formate. The butynediol can be recovered from this solution by vacuum distillation; the propargyl alcohol is recovered in an azeotrope.

Handling of High-Pressure Acetylene. Acetylene can explode or even detonate if the pressure rises above 1.4 bar (→ Acetylene). In butynediol plants this deflagration is often ignited by copper acetylide that was accidentally deposited in the head space; if it dries up, it can explode even upon slight vibration. Because other sources of ignition can never be excluded special precautions are required when handling acetylene at pressures above 1.4 bar. When filling all equipment with metal packings and all pipes with bundles of thin tubes (max. diameter 12.5 mm) the transition from deflagration to detonation (50 – 100 fold increase in pressure) can be excluded. Explosion (10 – 12 fold increase in pressure), however, cannot completely be avoided; this is accounted for by planning a 12-fold pressure for the whole apparatus [1]. When planning such plants the technical rules

for handling acetylene [7] and the publication of SARGENT [8] are recommended.

Environmental Protection. The off-gas, which mainly consists of acetylene and other hydrocarbons, is burnt. Waste-water problems arise when regenerating the catalyst. The water used for rinsing is treated biologically after traces of catalyst have been deposited. Spent catalyst must always be wet. Copper and bismuth can be recovered from the catalyst by acid treatment.

Analysis. Because of the high boiling points and the instability of the compounds involved, GC analysis is of limited value only. The determination of byproducts by titration or by fractional distillation (and subsequent GC analysis of the fractions) is time consuming but more reliable.

Storage and Transportation. The storage temperature should be kept below 40 °C and storage times longer than a few months should be avoided because the butynediol flakes tend to set up.

Uses. Approximately 95 % of the butynediol is hydrogenated in aqueous solution to 1,4-butanediol and occasionally to 2-butene-1,4-diol. In addition, butynediol is an intermediate for the production of 3-chloro-2-butynyl *N*- (3-chlorophenyl)carbamate, a herbicide. Solvay reacts butynediol with alkylene oxides and obtains a flame retardant after subsequent addition of bromine to the triple bond. Considerable amounts of butynediol are added to electroplating baths (nickel and copper) as brightening agent. In steel inhibitor picklings, the diol acts as a corrosion inhibitor.

Commercial Products. Butynediol is traded as 34 % aqueous solution by GAF and BASF (Golpanol BOZ liquid or Korantin BH liquid with hexamethylenetetramine being added to the latter). The anhydrous product is also traded as flakes. Products containing residues (GAF) are 97 – 98 % pure. Distilled butynediol (BASF) is 99 % pure: Butynediol pure cryst., Korantin BH solid, Golpanol pure solid.

1.2. 2-Butene-1,4-diol

Physical Properties. 2-Butene-1,4-diol [*110-64-5*], $HOCH_2-CH{=}CH-CH_2OH$, $C_4H_8O_2$, M_r 88.1, *mp* 10 °C, *bp* 240 °C (at 1013 mbar), *bp* 140 °C (at 24 mbar), is a pale yellow liquid, which is soluble in water. The product should be distilled at reduced pressure because it deteriorates above 180 °C.

Specific heat capacity	2.71 J g^{-1} K^{-1}
Flash point	127 °C
Ignition temperature	335 °C

Vapor pressure

t, °C	110.4	123.7	140	171.3	240
p, mbar	4.4	9.6	24	101.7	1013

Chemical Properties. Butenediol undergoes the typical reactions of alcohols, such as ester formation with anhydrides or acid chlorides, substitution of the hydroxyl groups by halogen, and the formation of polyesters with dicarboxylic acids. Typical reactions of the double bond are addition of halogen, epoxidation with hydrogen peroxide, copolymerization with vinyl esters, and the classic Diels-Alder reaction with dienes, e.g., with hexachlorocyclopentadiene.

Production. 2-Butene-1,4-diol is produced industrially by partial hydrogenation of 2-butyne-1,4-diol in aqueous solution and in the presence of a palladium catalyst, which generally is poisoned with zinc or other metals [9]. The main product of this synthesis is *cis*-2-butene-1,4-diol. Raney nickel can also be used as a catalyst for the hydrogenation [10]. Another method for the production of 2-butene-1,4-diol is the acetoxylation of butadiene followed by hydrolysis [11].

Analysis. The 2-butene-1,4-diol concentration in the product is determined by gas chromatography, the water content by the Karl Fischer method.

Storage and Quality. 2-Butene-1,4-diol must be stored in rust-free containers. Typical commercial specifications are: purity 98.5 %, *mp* 10 °C, color light to slightly yellow, water content < 0.5 %.

Uses. 2-Butene-1,4-diol is used for the production of endosulfan, chlorinated bicyclo-[2.2.1]heptene-(2)-bis(oxyalkylene-5,6) sulfite, an insecticide [12]; for the production of pyridoxine (vitamin B6) [13]; and in mixtures with other compounds as bactericide.

Economic Aspects. Worldwide 4000 – 5000 t/a of 2-butene-1,4-diol are produced, most of which is used for the production of crop protection products and vitamins.

1.3. 1,4-Butanediol

Physical Properties. 1,4-Butanediol [*110-63-4*], $HOCH_2CH_2CH_2CH_2OH$, $C_4H_{10}O_2$, M_r 90.12, is a colorless, almost odorless, hygroscopic liquid (see Table 1), readily soluble in water, alcohols, ketones, glycol ethers, and glycol ether acetates, less soluble in diethyl ether and esters, and not miscible with aliphatic and aromatic hydrocarbons and chlorinated hydrocarbons.

Chemical Properties. 1,4-Butanediol is readily cyclized in acid medium to give tetrahydrofuran. Dehydrogenation in the presence of copper – zinc – aluminum catalysts gives butyrolactone.

1,4-Butanediol reacts with monocarboxylic acids to give diesters [14]. Esterification with dicarboxylic acids and their derivatives leads to partially crystalline, linear, thermoplastic polymeric esters [15], [16].

At ca. 200 °C, 1,4-butanediol reacts with ammonia or an amine over nickel or cobalt catalysts and in the presence of hydrogen to give pyrrolidine or pyrrolidine derivatives [17], [18]. Phosgene reacts with 1,4-butanediol at −5 °C to give the bis(chloroformate) of butanediol [19], [20]. Acrylonitrile adds to 1,4-butanediol at 20 – 100 °C and in the presence of catalytic amounts of alkali to give 1,4-bis(2-cyanoethoxy)butane [21]. Like other alcohols, 1,4-butanediol can be vinylated, giving the divinyl ether [22].

Production. 1,4-Butanediol is made on a large industrial scale by continuous hydrogenation of the 2-butyne-1,4-diol [23] over modified nickel catalysts [24], [25]. The one-stage flow process is carried out at 80 – 160 °C and 300 bar according to Figure 3.

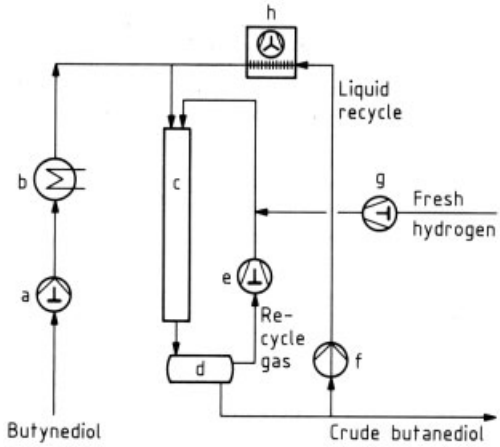

Figure 3. Hydrogenation of 2-butyne-1,4-diol
a) Piston pump; b) Preheater; c) Hydrogenation reactor; d) Gas – liquid separator; e) Recycle compressor; f) Canned motor pump; g) Hydrogen compressor; h) Air cooler

An aqueous solution of 2-butyne-1,4-diol (30 – 50 %), together with carbon monoxide-free hydrogen and recycled reaction mixture, which acts as medium for dissipation of heat, is lead over a reduced nickel – copper – manganese catalysts on silica gel strands. The initial temperature in the reactor is 80 °C; the temperature must not exceed 170 °C. In order to obtain a better distribution of the liquid, hydrogen is also circulated.

The raw product contains methanol, propanol, and butanol as byproducts as well as traces of 2-methyl-1,4-butanediol, hydroxybutyraldehyde, acetals, and triols. The reactor effluent is worked up to pure 1,4-butanediol by fractional distillation [26].

The hydrogenation of butynediol can also be carried out in two process steps. In the first stage, for example, at 40 bar, mainly 2-butene-1,4-diol is obtained, which is then completely hydrogenated at 300 bar in the second stage.

A further variant is the low-pressure hydrogenation at ca. 20 bar over a suspended Raney nickel catalyst, followed by hydrogenation at 120 – 140 °C and 140 – 210 bar on a fixed-bed contact [24].

Other Processes. 1,4-Butanediol can also be made from raw materials other than acetylene. Allyl alcohol and synthesis gas ($CO + H_2$) can

Table 1. Physical properties of 1,4-butanediol

mp	20.2 °C			
bp	230.5 °C (at 101.3 kPa)			
Density ϱ	1.017 g/cm^3 (at 20 °C); 1.0154 g/cm^3 (at 25 °C)			
Critical temperature t_c	446 °C			
Critical pressure p_c	41.2 bar			
Vapor pressure: *t*, °C	60	100	140	180	200
p, kPa	ca. 0.031	0.47	4.08	21.08	41.5
Heat of fusion ΔH_f	16.3 kJ/mol ± 5 %				
Heat of vaporization: *t*, °C	131.4	193.2	215.6	230.5	
ΔH_v, kJ/mol	68.2	59.4	57.8	56.5	
Specific heat capacity: *t*, °C	20	50	100	150	
c, J g^{-1} K^{-1}	2.2 ± 2 %	2.46 ± 2 %	2.9 ± 3 %	3.33 ± 4 %	
Specific heat capacity of a 50 % aq. solution: *t*, °C	20	50	75	100	
c, J g^{-1} K^{-1}	3.4 ± 2 %	3.56 ± 2 %	3.69 ± 2 %	3.82 ± 3 %	
Heat of combustion ΔH_c	2585 kJ/mol				
Thermal conductivity: *t*, °C	30	50	70	100	
λ, W m^{-1} K^{-1}	0.2100	0.2091	0.2083	0.2069	
Thermal conductivity of a 50 % aq. solution: *t*, °C	20	50	100	150	
λ, W m^{-1} K^{-1}	0.3601	0.3694	0.3886	0.3984	
Viscosity η	91.56 mPa · s (at 20 °C); 71.5 mPa · s (at 25 °C)				
Refractive index n_D	1.4460 (at 20 °C); 1.4446 (at 25 °C)				
Dielectric constant ε	31.4				
Flash point	134 °C				

be converted to 1,4-butanediol according to a process invented by Daicel [27], [28].

Mitsubishi uses a three-step process [29], [30]: (1) the catalytic reaction of butadiene and acetic acid yields 1,4-diacetoxy-2-butene [*18621-75-5*]; (2) subsequenthydrogenation gives 1,4-diacetoxybutane; and (3) hydrolysis leads to 1,4-butanediol:

$$CH_2{=}CH{-}CH{=}CH_2 + 2\,CH_3COOH + 1/2\,O_2 \longrightarrow CH_3OOC{-}CH_2{-}CH{=}CH{-}CH_2{-}COOCH_3$$

$$\xrightarrow{H_2} \xrightarrow{H_2O} 2\,CH_3COOH + HO{-}CH_2CH_2CH_2CH_2{-}OH$$

BASF also has patented a similar process, in which acetic acid first adds to butadiene; the product then isomerizes to 1,4-diacetoxy-2-butene [31], [32].

According to a process practiced by Toyo Soda [33] chlorine first adds to butadiene to form a mixture of 1,4-dichloro-2-butene [*764-41-0*] and 3,4-dichloro-1-butene [*760-23-6*]. This mixturereacts with sodium acetate to form 1,4-diacetoxy-2-butene, which is subsequentlyhydrogenated directly to 1,4-butanediol.

In a patent described by Shell butadiene reacts with a peroxide to diperoxybutene, which is then converted to 1,4-butanediol by hydrogenation [34].

1,4-Butanediol also is a starting material for the production of *tetrahydrofuran* [*109-99-9*], which can be madedirectly from butadiene according to a process developed by Chevron [35]: 1,2-epoxy-3-butene is made from butadiene and a peroxide and subsequent isomerization yields 2,5-dihydrofuran [*1708-29-8*], whichis hydrogenated to tetrahydrofuran:

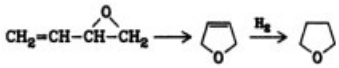

Other products derived from 1,4-butanediol are butyrolactone and tetrahydrofuran. In addition, both 1,4-butanediol and its derivatives are used in the textile, leather, photographic, food, and pharmaceutical industries.

The worldwide capacity for 1,4-butanediol is ca. 512 000 t/a in 1995. European producers are BASF and GAF/Chemische Werke Hüls; producers in the United States are Du Pont, GAF, and BASF Wyandotte. Whereas all these producers use acetylene as starting material, a plant at Mitsubishi Chemical Industries has been making 1,4-butanediol from butadiene since 1982.

Storage and Quality. 1,4-Butanediol can be stored indefinitely. The product is noncorrosive and therefore can be transported in cast iron containers. When it is stored for longer periods, storage tanks of steel or aluminum are necessary in order to avoid traces of iron in the product. In

this case, a cover of dry nitrogen also is recommended. Typical commercial specifications are: purity 99.5 – 99.8 %, *mp* 19.9 – 20.0 °C.

Uses and Economic Aspects. 1,4-Butanediol is a versatile intermediate for the chemical industry. The most important area of application is the production of polyurethanes and poly(butylene terephthalate), → Polyurethanes. Among the polyurethanes produced from 1,4-butanediol, cellular and compact elastomers are of prime importance. Poly(butylene terephthalate) is processed particularly to plastic materials and hot-melt adhesives, but is used also for the production of plastic films and fibers.

2. Other Butanediols

2.1. 2,3-Butanediol

2,3-Butanediol [*513-85-9*], 2,3-butylene glycol, exists in three stereoisomericforms:

Previously, 2,3-butanediol was obtained by bacterial fermentation of hexoses and pentoses. By pyrolysis of the diacetate very pure butadiene can be obtained, which has been used in the

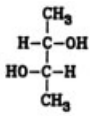

2*S*,3*S*–2,3–Butanediol
dextrorotatory

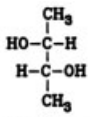

2*R*,3*R*–2,3–Butanediol
[*24347–58–8*]
levorotatory

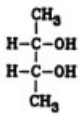

(*R*,*S*)–2,3–Butanediol
[*5341–95–7*]
optically inactive

United States for the production of synthetic rubber. Today butenes from crack gases are the raw material.

Physical Properties. The 2,3-butanediols, $C_4H_{10}O_2$, M_r 90.12, are colorless and odorless, strongly hygroscopic, oily liquids or crystals with a sweet taste. They are miscible with water and easily soluble in low molecular mass alcohols and ketones. For further physical data, see Table 2.

Chemical Properties [36]. The dehydrogenation of 2,3-butanediol yields acetoin [*513-86-0*] and diacetyl [*431-03-8*]. Dehydration leads chiefly to 2-butanone. The oxidation, e.g., with periodate, gives acetaldehyde; the reaction can be used for analytical determination. 2,3-Butanediol forms cyclic esters, acetals, and ketals. With diisocyanates, polyurethanes are obtained.

Production. After removal of butadiene and isobutene from crack gases, a C_4 hydrocarbon fraction, called C_4 raffinate II, is obtained, which contains approximately 77 % butenes and 23 % of a mixture of butane and isobutane. By chlorohydrination of this fraction with a solution of chlorine in water and subsequent cyclization of the chlorohydrins with sodium hydroxide, a butene oxide mixture of the following composition is obtained:

55 % *trans*-2,3-butene oxide, 30 % *cis*-2,3-butene oxide, 15 % 1,2-butene oxide.

Hydrolysis of this mixture (50 bar, 160 – 220 °C, reaction enthalpy $\Delta H = -42$ kJ/mol) yields a mixture of butanediols which are separated by vacuum fractionation. In order to avoid the formation of polyethers during the hydrolysis, an excess of water must be used. The fractionation of the butanediols is easier than that of the butene oxides. By this reaction sequence, *meso*-2,3-butanediol is obtained from *trans*-2-butene via *trans*-2,3-butene oxide; the racemic mixture of *R,R*- and *S,S*-2,3-butanediol is formed analogously from *cis*-2-butene via *cis*-2,3-butene oxide. For a discussion of stereochemistry of this reaction, see also [37].

Analysis, Storage. The butanediol mixture is analyzed best by GC (polyethylene glycol as stationary phase).

Storage and transportation do not provide any problems. However, the hygroscopicity of the product requires some precautions.

Uses. 2,3-Butanediol (at least 80 % meso isomer, the rest racemic mixture) is used as a cross-linking agent for naphthalene-1,5-diisocyanate

Table 2. Physical properties of 2,3-butanediols

	(*R,S*)-2,3-Butanediol	2*R*, 3*R*-2,3-Butanediol	2*S*, 3*S*-2,3-Butanediol	Racemate (*R,R*- and *S,S*-2,3-Butanediol)
mp, °C	35.5 – 36.5	19		7.6
bp (976 mbar), °C	181.7	179 – 180 ∗	179 – 182 ∗	176.7
bp (13.16 mbar), °C	83.5	77.3 – 77.4		75.3 – 75.6
d_4^{25}	0.9939	0.9873	0.9872	
n_D^{25}	1.43719	1.43095	1.4306	1.43109
$[\alpha]_D^{25}$		−13.16° ∗	+11.8° ∗	
viscosity (35 °C), mPa · s	65.6	21.8		
mp of bis(4-nitrobenzoates), °C	193 – 193.5	143 – 143.5		128 – 128.5

∗ Theoretically *bp* and absolute amount of rotation of enantiomers should be the same.

in the production of specific hard-rubber products (Vulkollan). Derivatives of 2,3-butanediol are important as insecticides (Sapecron: acetal with *N*-methylcarbamate of salicylic aldehyde) and as intermediates in the pharmaceutical industry. 2,3-Butanediols have some interest as humectants and in the synthesis of polymers and plasticizers.

2.2. 1,3-Butanediol

1,3-Butanediol [*107-88-0*], M_r 90.12, *bp* 207.5 °C (at 1013 mbar), *bp* 103 – 104 °C (at 10 mbar), d_4^{20} 1.0053, n_D^{20} 1.4410, is miscible with water and ethanol. The molecule has one center of chirality, and the data given above refer to the racemate. (*R*)-(−)-1,3-Butanediol [*6290-03-5*] has a specific rotation (ethanol) of $[\alpha]_D^{25} = -18.8°$.

1,3-Butanediol is produced as an intermediate in the manufacture of butadiene from acetaldol (→ Butadiene). This process has, however, been abandoned by most companies. The compound is mainly used as a component of special polyester resins.

2.3. 1,2-Butanediol [38]

Physical Properties. 1,2-Butanediol [*584-03-2*], $C_4H_{10}O_2$, M_r 90.12, *mp* −50 °C, *bp* 195 – 196.9 °C (at 101.3 kPa), is a colorless liquid, d_4^{20} 1.0023, n_D^{20} 1.4382. The diol is miscible with water in all proportions, readily soluble in alcohols, slightly soluble in ethers and esters, and insoluble in hydrocarbons. The dynamic viscosity at 20 °C is 73 mPa · s and the flash point 107 °C.

Chemical Properties. 1,2-Butanediol is a typical glycol and readily forms acetals and ketals; with dicarboxylic acids or anhydrides polyesters are formed and with diisocyanates polyurethanes.

Production. 1,2-Butanediol is synthesized by addition of water to 1,2-epoxybutane [*106-88-7*]:

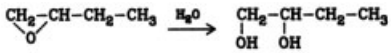

The hydration reaction is exothermic (ΔH= −93 kJ/mol). A 10 to 20-fold molar excess of water is used to suppress polyether formation. The reaction is carried out either without a catalyst at 160 – 220 °C and 10 – 30 bar or in the presence of catalysts below 160 °C and only slightly above atmospheric pressure. Sulfuric acid or strongly acid ion exchange resins are used as catalysts [38]. Depending on the excess of water, selectivity is 70 – 92 %. Higher ethers of 1,2-butanediol are formed as byproducts.

Uses. The field of application for 1,2-butanediol is not very broad. It is used mainly as a solvent and intermediate.

3. Toxicology

1,2-Butanediol and 1,3-Butanediol. The median lethal dose (LD_{50}) of 1,2-butanediol is 16 g/kg (rat, oral) [39]; the corresponding value of 1,3-butanediol is 29.6 g/kg [40], [41]. Repeated administration of 1,3-butanediol to rats

(2 a, 10 wt % in the feed) and to dogs (2 a, 3 wt % in the feed) did not reveal any toxic effects [42]. In high doses, 1,3-butanediol has a narcotic effect with a specific depressive effect on the central nervous system [43], [44]. 1,3-Butanediol is approved as a food additive by the U.S. Food and Drug Administration [45].

2,3-Butanediol has an LD_{50} value of 8.9 g/kg (mouse, oral) [40].

1,4-Butanediol, however, with an LD_{50} value of 1.78 g/kg [39] or 1.5 g/kg [46] (both rat, oral) should be classified as harmful according to the EC Council Directive. The observed depressive effect on the central nervous system is attributed to the metabolite γ-hydroxybutyric acid.

None of the butanediol isomers either irritate or sensitize the skin; however, they cause strong (1,2-butanediol) to slight (1,3- and 1,4-butanediol) eye irritation [39].

2-Butene-1,4-diol with an LD_{50} value of about 860 mg/kg (rat, oral) is harmful. Eye and skin irritation are not observed, even after prolonged contact. Rats could inhale air saturated with 2-butene-1,4-diol at 20 °C for 8 h without showing any symptoms [46]. Repeated intake by rats (5 % in the feed) led to death after 7 – 11 days; 30 % in the feed led to the animals' death after four days [47].

1-Butene-3,4-diol with an LD_{50} value of approximately 1.5 g/kg (rat, oral) is also harmful. An 8-h inhalation of air saturated with the compound at 20 °C by rats does not lead to any effects. The animals' skin was not irritated after contact for several hours; however, irritation to the eyes may occur.

2-Butyne-1,4-diol is to be classified as "toxic" having an LD_{50} value of about 100 mg/kg (rat, oral). Strong irritation to the skin – especially after contact for several hours – and to the eyes were observed. An 8-h inhalation of air saturated with the compound at 20 °C by rats did not lead to any adverse effects [46].

Neither TLV nor MAK values have been established for the compounds cited above.

4. References

1. W. Reppe: *Chemie und Technik der Acetylen-Druckreaktionen,* 2nd ed., Verlag Chemie, Weinheim 1952.
2. C. J. S. Appleyard, J. F. C. Gortshore: Manufacture of Butynediol at IG Ludwigshafen, Bios-Report 367, OTS-Report PB 28556, US Department of Commerce.
3. D. L. Fuller, A. O. Zors, H. M. Weir: The Manufacture of Butynediol from Acetylene and Formaldehyde, Fiat-Report no. 296, 1946, OTS-Report PB 80334 US Department of Commerce.
4. BASF, DE-AS 2040501, 1970 (G. Boettger, H. Hoffmann, W. Reiß, L. Schuster, H. Toussaint).
5. BASF, DE-AS 2421407, 1974 (J. Dehler, H. Hoffmann, W. Reiß, R. Schnur, S. Winderl, P. Zehner).
6. W. Reppe, *Justus Liebigs Ann. Chem.* **596** (1955) 8.
7. *Technische Regeln Acetylen,* Carl Heymanns Verlag, Köln.
8. H. B. Sargent, *Chem. Ing.* 1957 (Feb.) 250.
9. BASF, DE 2431929, 1981.
10. GAF, DE-AS 1139832, 1961.
11. K. Takehira, H. Mimoun, I. Sérée de Roche, *J. Catal.* **58** (1979) 155 – 169.
12. Hoechst, DE-AS 1015797, 1954.
13. R. A. Firestone, E. E. Harris, W. Reuter, *Tetrahedron* **23** (1967) no. 2, 943 – 955.
14. I.G. Farbenind., FR 889079, 1943.
15. Wingfoot Corp., GB 630992, 1949.
16. J. Nelles, O. Bayer, W. Tischbein, F. Baehren, US 2417513, 1947.
17. W. Reppe, C. Schuster, E. Weiss, US 2421650, 1947.
18. J. H. Paden, H. Adkins, *J. Am. Chem. Soc.* **58** (1936) 2487.
19. I.G. Farbenind., FR 905141, 1945.
20. US Department of Commerce: *The Continuous Production of 1,4-Butanediol Dichlorocarboxylic Ester,* PB 58619.
21. L. H. Smith: *Synthetic Fiber Development in Germany,* Textile Research Institute, New York 1964.
22. I.G. Farbenind./BASF, DE 679607, 1939.
23. I.G. Farbenind./BASF, DE 858094, 1937.
24. GAF, US 3449445, 1967 (F. E. Wetherill).
25. BASF, DE 2917018, 1979 (K. Baer, W. Reiss, W. Schroeder, D. Voges).
26. S. H. H. Chow, J. D. Verbsky, US 3852164, 1974.
27. *Chem. Ind. (Düsseldorf)* **32** (1980) 275.
28. M. Tamura, S. Kumano, *Chem. Econ. Eng. Ref.* **12** (1980) no. 9, 32.
29. Mitsubishi Chem., DE-OS 2345160, 1974; DE-OS 2424539, 1974; DE-OS 2504637, 1975; DE-OS 2510088, 1975; DE-OS 2510089, 1975; DE-OS 2505749, 1975.

30. Y. Tanabe, *Hydrocarbon Process.* **60** (1981) no. 9, 187.
31. BASF, DE-OS 2444004, 1976.
32. BASF, DE-OS 2454768, 1976.
33. Toyo Soda, US 3720704, 1973.
34. Shell Oil, US 4384146, 1983.
35. Chevron Research, US 3932468, 1976.
36. *Beilstein,* **1 (3)** 2178, 2180, 2181, 2183; **1 (4)** 2524, 2525.
37. C. E. Wilson, H. J. Lucas, *J. Am. Chem. Soc.* **58** (1936) 2396.
38. K. Szafraniak, J. Myszkowski, A. Zielinski, W. Pyc, PL 79662, 1978; *Chem. Abstr.* **92** (1980) 22031 n; *Przem. Chem.* **52** (1973) no. 11, 744 – 746; *Chem. Abstr.* **80** (1974) 47385 w.
39. V. K. Rowe et al.: *Patty's Industrial Hygiene and Toxicology,* **vol. 2 C,** Wiley Interscience Publ., New York 1982, p. 3874.
40. L. Fischer et al., *Z. Gesamte Exp. Med.* **115** (1949) 22.
41. A. Loeser, *Pharmazie* **4** (1949) 263.
42. R. A. Scala et al., *Toxicol. Appl. Pharmacol.* **10** (1967) 160.
43. G. D. Frye et al., *J. Pharmacol. Exp. Ther.* **216** (1981) 306.
44. G. S. Stoewsand et al., *Proc. Int. Congr. Nutr. 7th 1966* **4** (1967) 1082–7.
45. W. R. Hewitt et al., *Toxicol. Appl. Pharmacol.* **64** (1982) 529.
46. BASF, unpublished results 1959 – 1981.
47. H. Schlüssel, *Naunyn Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **221** (1954) 67.

Butanes → Hydrocarbons

Butanols

Heinz-Dieter Hahn, Ruhrchemie Aktiengesellschaft, Oberhausen, Federal Republic of Germany (Chaps. 1 – 8)

Georg Dämbkes, Ruhrchemie Aktiengesellschaft, Oberhausen, Federal Republic of Germany (Chaps. 1 – 8)

Norbert Rupprich, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chap. 9)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Production 4
5. Quality Requirements and Control . 7
6. Storage and Transportation 7
7. Uses 7
8. Economic Aspects 9
9. Toxicology and Occupational Health 9
9.1. 1-Butanol 9
9.2. 2-Butanol 10
9.3. 2-Methyl-1-propanol 10
9.4. 2-Methyl-2-propanol 11
10. References 11

1. Introduction

Butanols (butyl alcohols) are aliphatic saturated C_4 alcohols (C_4H_9OH, M_r 74.12). There are four structural isomers of the alcohols: two primary, one secondary, and one tertiary; as there is an asymmetric C atom in the secondary alcohol, there are two stereoisomers of 2-butanol.

Formula	Nomenclature	Common name
$CH_3CH_2CH_2CH_2OH$	1-butanol [*71-36-3*]	*n*-butanol
$CH_3CH(CH_3)CH_2OH$	2-methyl-1-propanol [*78-83-1*]	isobutanol
$CH_3CH_2\overset{*}{C}H(OH)CH_3$	2-butanol [*78-92-2*]	*sec*-butanol
$CH_3-C(CH_3)_2-OH$	2-methyl-2-propanol [*75-65-0*]	*tert*-butanol

In addition to the systematic IUPAC nomenclature, other, nonsystematic names for the various butanols are in common use.

1-Butanol (*n*-butanol) occurs in nature in compound form. It also occurs, sometimes in high concentrations, in fusel oils obtained by fermentation. The first industrial production of 1-butanol, around 1912, was based on the discovery of the bacterium *Clostridium acetobutylicum Weizmann,* which causes carbohydrates to ferment to give mainly acetone and 1-butanol.

The continually increasing demand for 1-butanol could only be met by developing new manufacturing processes, such as the hydrogenation of crotonaldehyde formed by the aldolization of acetaldehyde; the Reppe synthesis (propylene carbonylation); and, in particular, the process by which most 1-butanol is produced, the hydrogenation of *n*-butyraldehyde, easily obtainable by the hydroformylation of propylene.

2-Methyl-1-propanol (isobutanol) occurs in natural products as well as in fusel oils (74 % of the total alcohol in the product of molasses fermentation), from which it can be isolated. The isobutyl oil synthesis, a reaction related to methanol synthesis from CO and H_2, first made it possible to produce larger quantities of 2-methyl-1-propanol. In the 1980s, nearly all 2-methyl-1-propanol is produced by the oxo synthesis (propylene hydroformylation). This straight forward production process and the attractive prices have led to a considerable expansion of the market for isobutanol.

The secondary and tertiary butanols are obtained by hydration of the corresponding unsaturated hydrocarbons, the butenes, whereas dehydration of the butanol yields the corresponding butene.

Thus, 2-butanol (*sec*-butanol), the simplest alcohol with an asymmetric carbon atom, can be obtained by the acid-catalyzed hydration

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_463

of 1- and 2-butene whereas hydration of 2-methylpropene (isobutene) yields 2-methyl-2-propanol (*tert*-butanol). The ready availability of 1- and 2-butene and 2-methylpropene from olefins derived from petroleum cracking has made the commercial application of these reactions possible. Most 2-butanol produced is further processed to 2-butanone (methyl ethyl ketone).

The rapid development of 2-methyl-2-propyl methyl ether as an additive for low-lead gasolines has increased the availability of 2-methyl-2-propanol, from which it can be derived as by-product.

2. Physical Properties

With the exception of 2-methyl-2-propanol (*mp* 25.6 °C) butanols are colorless liquids. Butanols exhibit a characteristic odor; their vapors have an irritant effect on mucous membranes and a narcotic effect in higher concentrations.

All butanols are completely miscible with common organic solvents. Only 2-methyl-2-propanol is completely miscible with water. The major characteristic physical properties of butanols are compiled in Table 1.

See Chapter 6 for safety data, such as flash points, ignition temperatures, and lower explosive limits.

For further physical properties of butanols see also [1]. The thermodynamic properties of all butanols have been examined in detail; cf. [2], [3] for 1-butanol; [4] for 2-methyl-1-propanol; [5–7] for 2-butanol; [8], [9] for 2-methyl-2-propanol.

References [10–12] give an outline of liquid – vapor equilibria for the distillation of butanols. H. A. Rizk and N. Youssef [13] as well as G. E. Rajala and J. Crossley [14] have investigated the dielectric properties of butanol in detail. D. Davis [15] has set up nomograms for the densities of aqueous solutions of 1-butanol, 2-methyl-1-propanol, and 2-methyl-2-propanol.

Further physical properties of the systems comprising water with the various butanols have been described by several authors; cf. [16–18] for 1-butanol/water, [19], [20] for 2-butanol, and [21] for 2-methyl-2-propanol.

For the phase equilibria of the ternary systems isobutyraldehyde/2-methyl-1-propanol (isobutanol)/water and 2-butanol/2-methyl-2-propanol (*tert*-butanol)/water cf. [22] and [23].

Table 2 shows the composition of some binary azeotropic mixtures. For other azeotropes see [24–26].

3. Chemical Properties

The chemical properties of butanols are mainly those of primary, secondary, and tertiary alcohols. In particular the reactive primary alcohols serve as starting materials for a wide range of reactions. The following deserve special mention:

Dehydration. The catalytic dehydration of alcohols for the preparation of olefins is a well-established reaction [27]. Of the butanols 2-methyl-2-propanol (*tert*-butanol) undergoes this reaction most readily. Even with dilute sulfuric acid water is split off on warming, leading to the formation of very pure 2-methylpropene (isobutene) [*115-11-7*] [28]. As 2-methylpropene is readily obtainable from the C_4-fractions of olefins from cracked petroleum this reaction is no longer of industrial importance, however. See Chap. 4 for the manufacture of pure 2-methylpropene. The dehydration of 2-methyl-1-propanol (isobutanol) similarly leads to the formation of 2-methylpropene.

The dehydration of 2-methyl-1-propanol to 2-methylpropene is carried out in the presence of γ-Al_2O_3 catalyst at about 300 – 350 °C with practically quantitative conversion and a 2-methylpropene selectivity of over 90 %. In an analogous manner 1-butanol can be dehydrated to a mixture of 1-butene [*106-98-9*] and *cis*- and *trans*-2-butene [*624-64-6*], [*590-18-1*] [29], [30].

Selective dehydration of 1-butanol in mixtures with 2-methyl-1-propanol can be carried out with calcium zeolite molecular sieves [31], due to the difference in the spatial structure of the two butanols.

Also, at lower temperature dibutyl ether [*142-96-1*] can be formed from 1-butanol in the presence of dehydration catalysts [30]. The branched butanols show little tendency to undergo this reaction.

Table 1. Physical properties of butanols

	1-Butanol	2-Methyl-1-propanol	2-Butanol	2-Methyl-2-propanol
mp, °C	− 89.3	− 107.9	− 114.7	+ 25.6
bp, °C	117.7	107.9	99.5	82.55
Density d_4^{20}	0.8098	0.8027	0.8065	0.7867
Refractive index n_D^{20}	1.3991	1.3959	1.39719	1.3841
Viscosity at 20 °C, mPa · s	3.0	4.0	4.2 (15 °C)	3.3 (30 °C)
Specific heat, J g^{-1} K^{-1}	2.437 at 30 – 80 °C	2.5263 at 30 – 80 °C	2.81 at 20 °C	3.035
Heat of vaporization, J/g	591.64	578.83	562.75	535.78
Heat of fusion, J/g	125.2			91.61
Heat of combustion, kJ/g	36.111	35.981		35.588
Critical pressure, hPa	48.4	48		
Critical temperature, °C	287	265	265	235
Surface tension at room temp., mN/m	22.3	23.0	23.5	
Dielectric constant at room temperature	17.8	18.8	15.5	11.4
Evaporation number (ether = 1)	33	24		
Solubility in water				
at 20 °C, wt %	7.7	8.5	12.5	miscible
at 30 °C, wt %	7.08	7.5	18	miscible
Solubility of water in butanol				
at 20 °C, wt %	20	15		miscible
at 30 °C, wt %	20.62	17.3	36.5	miscible
Rotation $[\alpha]_D^{20}$			(+) (*S*)-2-butanol + 13.73° (−) (*R*)-2-butanol − 13.79°	

Table 2. Azeotropic mixtures [24], [25], [26]

Compound A	Compound B	*bp*, °C comp. B	*bp*, °C azeotrope	wt % butanol in azeotrope
1-Butanol	water	100	92.7 [26] 93.0 [24]	57.5 [26], 55.5 [24] 63.0 [25]
1-Butanol	acetic acid 1-butyl ester	126.2	117.6	67.2
1-Butanol	2-methyl-1-propanol	99.5		no azeotrope
1-Butanol	cyclohexane	81.4	79.8	10.0
1-Butanol	toluene	110.6	105.6	27.0
2-Methyl-1-propanol	water	100	89.8	67 [26], 70 [24]
2-Methyl-1-propanol	acetic acid 2-methylpropyl ester	116.5	107.6	92/95 [26], 55 [24]
2-Methyl-1-propanol	cyclohexane	81.4	78.1	14.0
2-Methyl-1-propanol	toluene	110.6	101.2	44.5
2-Butanol	water	100	87.0 [26], 88.5 [24]	73.2 [26], 68 [24]
2-Butanol	acetic acid 2-methylpropyl ester	112.2		no azeotrope
2-Butanol	2-butanone	79.6		no azeotrope
2-Butanol	cyclohexane	81.4	76.0	18.0
2-Butanol	toluene	110.6	95.3	55.0
2-Methyl-2-propanol	water	100	79.9	88.24
2-Methyl-2-propanol	cyclohexane	81.4	71.3	37.0

Oxidation. As primary and secondary alcohols 1-butanol, 2-methyl-1-propanol, and 2-butanol can be dehydrogenated to the corresponding carbonyl compounds. Dehydrogenation can be carried out even at low reaction temperatures with oxidizing agents such as manganese(IV) oxide in sulfuric acid, nitric acid, chromic acid, or selenium dioxide [32].

Even without oxidizing agents the use of suitable catalysts (e.g., Cu) at higher temperatures enables the alcohols to be dehydrogenated to aldehyde or ketone until the thermodynamic equilibrium has been established [33–36]. 2-Butanol dehydrogenates to 2-butanone (methyl ethyl ketone) [*78-93-3*] at about 350 °C [37].

As a tertiary alcohol 2-methyl-2-propanol cannot be selectively dehydrogenated; splitting

of the molecule occurs under extreme conditions.

Primary alcohols can be oxidized to carboxylic acids under other reaction conditions. In addition to the standard oxidizing agents, this oxidation can be accomplished by the reaction of the alcohol with alkali hydroxide at 275 °C. In this reaction, which is based on the observations of GUERBERT, the alkali salt of butyric acid [*107-92-6*] is formed together with hydrogen and 2-ethylhexanol [*104-76-7*] (cf. also [38]).

The Koch-Haaf reaction can be used to prepare trimethylacetic acid [*75-98-9*] from 2-methyl-1-propanol and 2-methyl-2-propanol by the addition of carbon monoxide in the presence of sulfuric acid [39]. With readily available 2-methyl-1-propanol (isobutanol) from the oxo synthesis, trimethylacetic acid can be obtained in 89 % yield [40].

Alkylation. Butanols can be employed in a wide variety of alkylation reactions. *N*-alkyl-, *N,N*-dialkyl-, or *N,N,N*-trialkylamines are obtainable with ammonia and amines. The use of isomeric mixtures of alcohols leads to the corresponding mixed amines. Ring alkylation of aromatic hydrocarbons proceeds with butanol in the presence of Friedel-Crafts catalysts [41].

Esterification. Butanols can be converted in the usual manner into butyl esters with inorganic and organic acids. The reaction is generally carried out in the presence of acid catalysts. In certain cases excess of acid which is to be esterified can serve as the catalyst. The rate of ester formation greatly depends on the structure of the carboxylic acid and of the alcohol. The primary butanols react more rapidly than 2-butanol, which in turn reacts more quickly than 2-methyl-2-propanol. For the formation of esters with acid chlorides and acid anhydrides, as well as other esterification processes, see [42].

4. Production

Of the many available processes for the preparation of butanols, the following have achieved industrial importance:

1-Butanol:

propylene hydroformylation (oxo synthesis)
Reppe synthesis
crotonaldehyde hydrogenation

2-Methyl-1-propanol:

propylene hydroformylation
catalytic hydrogenation of carbon monoxide
homologization reaction

2-Butanol:

n-butene hydration

2-Methyl-2-propanol:

2-methylpropene hydration
byproduct of propylene oxide and methyl *tert*-butyl ether production

Oxo Synthesis. The most important process for the manufacture of 1-butanol and 2-methyl-1-propanol is propylene hydroformylation with subsequent hydrogenation of the aldehydes formed.

In the oxo reaction (hydroformylation) carbon monoxide [*630-08-0*] and hydrogen [*1333-74-0*] are added to a carbon – carbon double bond in the liquid phase in the presence of catalysts (hydrocarbonyls or substituted hydrocarbonyls of Co, Rh, or Ru) (→ Oxo Synthesis) [43]. In the first reaction step aldehydes are formed with one more C-atom than the original olefins. For olefins with more than two C-atoms, isomeric aldehyde mixtures are normally obtained. In the case of propylene these consist of 1-butanal and 2-methylpropanal.

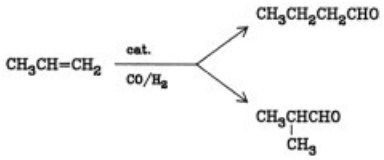

There are several variations of the hydroformylation process, the differences being in the reaction conditions (pressure, temperature) as well as the catalyst system used. The classic high-pressure process exclusively used until the beginning of the 1970s operates at pressures of 20 – 30 MPa (200 – 300 bar) CO/H_2 and temperatures of 100 – 180 °C. The catalyst is Co. It

leads to about 75 % 1-butanol and about 25 % 2-methyl-1-propanol.

The new process developments of the past few years have led to a clear shift in the range of products. The processes operating at relatively low pressures (1 – 5 MPa , 10 – 50 bar) use modified Rh-catalysts. The isomeric ratios achieved are about 92 : 8 [44] or 95 : 5 [45] 1-butanol to 2-methyl-1-propanol. However, by the use of unmodified Rh the percentage of 2-methyl-1-propanol can be increased to about 50 %. For details on the oxo synthesis and propylene hydroformylation see [25], [38], [46] and → Oxo Synthesis.

Catalytic hydrogenation of the aldehydes leads to the formation of the corresponding alcohols.

As only primary alcohols can be obtained via the oxo synthesis, it is not possible to produce 2-butanol and 2-methyl-2-propanol by this process.

Reppe Process. 1-Butanol and 2-methyl-1-propanol can also be produced on a commercial scale by the carbonylation of propylene [*115-07-1*] developed by Reppe. In this process, developed in 1942, [47], [48], olefins, carbon monoxide, and water are made to react under pressure in the presence of a catalyst (tertiary ammonium salt of polynuclear iron carbonyl hydrides). The difference between this process and the classic Co-catalyzed hydroformylation is that at low temperature (about 100 °C) and low pressure (0.5 – 2 MPa, 5 – 20 bar) alcohols are formed directly from the olefin.

As with the oxo synthesis the carbon monoxide can be added to both C-atoms of the double bond which means that, when propylene is used, 1-butanol and 2-methyl-1-propanol are obtained in a ratio of 86 : 14.

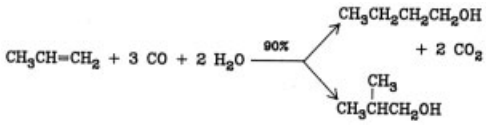

The catalyst is sensitive both to air and to higher temperatures at which, in the presence of water and CO_2, it decomposes into iron carbonate. To achieve adequate reaction rates the catalyst, carbonyl triferrate, must be present in concentrations of $\gg$ 10 % in the reaction solution; this is attained by the presence of dissolving agents (*N*-alkylpyrrolidine); see [49], [50], [51] for details of the process.

The Reppe process has not been as successful as propylene hydroformylation with Co catalysts despite favorable n/iso ratios in the reaction products and milder reaction conditions [52]. This can be attributed to the more expensive process technology.

Hydrogenation of Crotonaldehyde [*4170-30-3*]. Until the mid 1950s the manufacture of 1-butanol based on acetaldehyde was the preferred process. With the development of the oxo synthesis, however, it has lost its importance even in Japan and the USA and is no longer in use.

The individual steps of the process are as follows:

aldol condensation

$$2\,CH_3CHO \longrightarrow CH_3CH(OH)CH_2CHO$$

splitting off of water

$$CH_3CH(OH)CH_2CHO \longrightarrow CH_3CH=CHCHO + H_2O$$

hydrogenation

$$CH_3CH=CHCHO + 2\,H_2 \longrightarrow CH_3CH_2CH_2CH_2OH$$

Acetaldehyde is aldolized to acetaldol at normal temperature and pressure in the presence of alkaline catalysts. With conversions of about 60 % the acetaldol yield is about 95 %. Unreacted acetaldehyde can be distilled off and recycled. The removal of water and formation of aldehyde is brought about by acidification of the acetaldol with acetic acid or phosphoric acid and subsequent distillation, whereby the crotonaldehyde is obtained almost quantitatively as the top product.

Various gas- and liquid-phase processes have proved their value for the hydrogenation of crotonaldehyde to 1-butanol. Copper catalysts are particularly useful. About 1000 kg of 1-butanol can be obtained from 1350 kg of acetaldehyde [53].

The future economic importance of the crotonaldehyde hydrogenation process will depend to a large extent on developments in the raw materials market. As “crude oil” raw material becomes scarcer and more expensive as a basis

for the oxo synthesis, the alternative route based on ethanol from fermentation will become increasingly competitive. The biomass substrate for ethanol production is almost inexhaustible. The ethanol is first dehydrogenated to form acetaldehyde, from which the synthesis can proceed.

Such a process route is an increasingly interesting alternative for tropical countries with large supplies of cheap biomass as well as for the more developed countries of the third world which do not have their own oil resources [54].

Catalytic Hydrogenation of Carbon Monoxide. In addition to the CO/H_2-based methanol synthesis developed commercially in the 1920s, other analogous processes were developed for the manufacture of higher alcohols. Compared with the methanol synthesis, however, these processes found only limited application. Nowadays they have no commercial importance.

The most significant process for the production of 2-methyl-1-propanol was the BASF isobutanol oil synthesis. Reaction products with about 50 % methanol [*67-56-1*] and 11 – 14 % 2-methyl-1-propanol as well as various other components were obtained from CO and H_2 at temperatures of about 430 °C and pressures of about 30 MPa (300 bar) in the presence of a KOH-modified methanol synthesis catalyst [55], [56]. BASF ceased its isobutyl oil synthesis in 1952 when 2-methyl-1-propanol and its secondary product 2-methylpropene could be obtained more economically using the oxo synthesis or via a petrochemical route. Further comprehensive experiments to obtain the desired butanol from the cheap raw materials methanol and ethanol with the homologization reaction have so far proved fruitless. Butanols have so far only been obtained in small amounts as byproducts. A review is given in [57].

Olefin Hydration. 2-Butanol (*sec*-butanol) and 2-methyl-2-propanol (*tert*-butanol) can be obtained by the acid-catalyzed addition of water to 1- or 2-butene [*106-98-9*], [*107-01-7*] and 2-methylpropene [*115-11-7*] respectively. The water adds at the double bond in accordance with the Markovnikov rule. The difference in olefin reactivity can be used for the selective manufacture of 2-butanol or 2-methyl-2-propanol from technical C_4-fractions where the two alcohols are prepared in turn and isolated. 2-Methylpropene can be completely converted to 2-methyl-1-propanol with 65 % sulfuric acid. Hydration of *n*-butene, on the other hand, requires 75 – 80 % H_2SO_4 [58]. Because of the tendency of the olefins to polymerize readily, the requisite reaction temperatures must not be exceeded.

In addition to this indirect hydration another process variant is known to industry in which the olefins are hydrated directly at elevated temperatures (about 200 – 300 °C) and higher pressures (10 – 35 MPa, 100 – 350 bar) [59], [60].

$$CH_3CH{=}CHCH_3 + H_2O \overset{\text{cat.}}{\rightleftharpoons} CH_3CH_2\underset{\underset{OH}{|}}{C}HCH_3$$

Various catalysts have been described for direct hydration, such as tungsten oxide [61] and aqueous aluminum hydroxide gel suspensions [62]. For the hydration of 1-butene and 2-butene the best results are obtained at temperatures between 220 and 300 °C or in the case of 2-methylpropene, at 190 – 245 °C [63]. Direct hydration of the olefins has not yet established itself as a serious competitor to indirect hydration. For process improvement see [64].

Other Processes. 2-Methyl-2-propanol (*tert*-butanol) is formed during the manufacture of propylene oxide from 2-methylpropane [*75-28-5*] in an amount of about 1.2 t per ton of propylene oxide [27, pp. 834, 1005]. In an intermediate stage *tert*-butyl hydroperoxide [*75-91-2*] is formed from isobutane. Another process, in which 2-methyl-2-propanol is obtained as a by-product, is the splitting of methyl *tert*-butyl ether to highly pure 2-methylpropene or 2-methyl-2-propanol. Similarly the methyl *tert*-butyl ether synthesis with slight traces of water in the feedstock (methanol and C_4-olefins) produces 2-methyl-2-propanol as a byproduct [65].

In the literature a number of other methods for the preparation of butanol are cited which are generally of no significance other than for laboratory preparation. A comprehensive review is to be found in B. Cornils and E. Zilly [38].

5. Quality Requirements and Control

On an industrial scale the four different butanols are generally marketed as pure substances and not as isomer mixtures. Because of their many applications both as commodities and as materials for the manufacture of special fine chemicals a high standard of purity is demanded for the butanols.

The quality is mainly controlled by determining the physical and chemical properties of the alcohols, such as density, refractive index, color, water content, hydroxyl value, acid value, carbonyl value, and distillation range. The characteristic values are normally determined according to the ASTM [66], DIN [67], or the BS methods [68].

Table 3 lists the standard specifications which are stipulated by the American Society for Testing and Materials for 1-butanol, 2-butanol, and 2-methyl-1-propanol.

The B. S. I. (British Standards Institution) has issued specifications for 1-butanol (BS 508: 1966) and 2-butanol (BS 1993: 1968).

The best way to verify isomer purity and detect traces of impurities is by gas-chromatographic analysis. Such analyses can be carried out rapidly and with good reproducibility, on a routine basis, enabling the use of a high sampling frequency.

Separation can be effected by the use of both polar and non-polar packed columns or appropriate capillary columns.

For general purity analyses a thermal conductivity detector system is used; trace (impurity) analysis is carried out using a flame ionization detector system. Experience has shown that nitrogen, hydrogen, and helium are suitable carrier gases.

The degree of purity of commercial 1-butanol, 2-methyl-1-propanol, and 2-butanol is at least 99 wt %. 2-Methyl-2-propanol is supplied both as a product of at least 99.8 wt % purity and as an azeotropic mixture with 11 to 12 wt % water.

In addition to characteristic value determination and gas-chromatographic analysis a number of special analyses and tests may be carried out depending on the intended application, such as esterification tests with phthalic anhydride and color tests with concentrated sulfuric acid or $KMnO_4$.

6. Storage and Transportation

Butanols can be stored or dispatched in untreated mild steel or enamelled steel drums provided that ingress of moisture is prevented. Stainless steel containers are also suitable. The butanols are transported in rail and road tank cars, and in drums as well as in tanker vessels and containers. It is important to observe the national and international safety regulations relating to aviation, ocean-going and inland waterway shipping, rail and road transport as well as the particular safety requirements for butanols.

Some important data relating to the safe handling, storage, and transport of butanols are compiled in Table 4.

7. Uses

1-Butanol. 1-Butanol is used principally in the field of surface coating. In the USA, for instance, approx. 85 % of 1-butanol production finds an outlet in this sector of the market. It is used either directly as a solvent for varnishes or it is converted into derivatives which then serve as solvents or monomer components.

1-Butanol cannot be used directly as a solvent in the production of nitrocellulose lacquers; in admixture with toluene, ethanol, or various esters, however, it is an excellent thinner.

The addition of 5 – 10 % of 1-butanol is also valuable in overcoming "blushing" (unwanted white opacity) which may arise when large quantities of thinners, particularly volatile solvents, are used.

1-Butanol is also useful for regulating the viscosity and improving the flow properties of varnishes and for the prevention of "streaking" in paints and lacquers based on spirit-soluble gums and resins.

Although not itself a solvent for substances such as polystyrene and chlorinated rubber, 1-butanol can be successfully used in quantities up to 20 % as a diluent for the commonly used solvents for these substances, which are mainly the esters of the saturated carboxylic acids; in particular, the acetates.

The acrylic ester of 1-butanol has become increasingly important during the last decade.

Table 3. ASTM Specifications for butanols [66]

ASTM-Standard:	1-Butanol D 304-80	2-Methyl-1-propanol D 1719-81	2-Butanol D 1007-80	ASTM Standard
Sampling				D 268-80
Specific gravity	0.810 – 0.813	0.802 – 0.804	0.807 – 0.809	D 268-80
Color (platinum – cobalt scale)	< 10	< 10	< 10	D 1209
Distillation range	117.7 °C within a range of 1.5 °C	107.9 °C within a range of 2 °C	98.0 – 101.0 °C	D 1078
Nonvolatile matter	< 0.005 g/100 mL	< 0.005 g/100 mL	< 0.005 g/100 mL	D 1353
Odor	characteristic, nonresidual	characteristic, nonresidual	characteristic, nonresidual	D 1296
Water (Karl Fischer)	< 0.1 wt %	< 0.2 wt %	< 0.5 wt %	D 1364
Miscibility with vol% heptane (99 %) at 20 °C	19	19	19	D 1476
Acidity (as free acetic acid)	< 0.005 wt % = 0.047 mg KOH/g sample	< 0.003 wt % = 0.028 mg KOH/g sample	< 0.002 wt % = 0.019 mg KOH/g sample	D 1613

Table 4. Safety data for butanols [25], [69–71]

	UN-No.	Flash point, °C	Ignition limits in air, vol%	Ignition temperature, °C	Danger class, IMDG-code	MAK [a], ppm	TLV [b], ppm	TWA [c], ppm
1-Butanol	1120	+34	1.4 – 11.3	380	3.3	100	50	100
2-Methyl-1-propanol (isobutanol)	1212	+28	1.7 – 10.9	430	3.3	100	75	100
2-Butanol (*sec*-butanol)	1121	+24		390	3.3	100	150	150
2-Methyl-2-propanol (*tert*-butanol)	1122	+11	2.3 – 8	470	3.2	100	150	100

[a] For the Federal Republic of Germany.
[b] Threshold Limit Value.
[c] Time-weighted average concentration.

It has become an essential component of latex paints since such coatings are not only robust but can also be produced relatively cheaply. The 1-butyl esters of phthalic, adipic, sebacic, oleic, azelaic, stearic, and phosphoric acids are used as plasticizers and additives for surface coatings. The most important of these products is di-1-butyl phthalate (DBP). Consumption is, however, stagnating and in some countries it decreased slightly within the last years.

In the USA in particular, but also in the Federal Republic of Germany and Belgium, 1-butanol is used for the manufacture of butylamines.

1-Butanol has numerous applications in the plastics and textile sector. It is used as a coagulation bath for spinning acrylic fibers and in the dyeing of poly(vinyl alcohol) fibers.

2-Methyl-1-propanol. The fields of application of 2-methyl-1-propanol (isobutanol) closely resemble those of 1-butanol. Indeed, because of its lower price it can often replace 1-butanol and is accordingly used as a solvent, diluent, and additive for nitrocellulose and synthetic resins; wetting agent; cleaner additive; and component of printing inks and related products. Its capacity for dissolving resins (ketone, phthalate, urea, and melamine – formaldehyde resins) is of commercial importance. As a solvent for phenol – formaldehyde resins, however, 2-methyl-1-propanol is not as effective as 1-butanol.

Several esters exhibit excellent solvent properties; in particular, isobutyl acetate is a very good solvent for fats, chlorinated rubber, polystyrene, and coumarone resins. The esters of phthalic acid, adipic acid, some dicarboxylic acids as well as phosphoric acid are used as plasticizers, particularly for PVC and its copolymers and for cellulose derivatives.

The isobutyl esters of 2,4-dichloro- and 2,4,5-trichlorophenoxyacetic acid exhibit herbicidal activity. For the anti-freezing properties of 2-methyl-1-propanol in gasoline see [72].

Also of interest is the use of 2-methyl-1-propanol as an extracting agent in the recovery of ammonium phosphate [73].

2-Butanol. Practically all of the 2-butanol produced is dehydrogenated to 2-butanone (MEK or methyl ethyl ketone) which has found considerable application because of its good solvent properties and its favorable boiling point (79.57 °C). 2-Butanol itself is also used as a solvent: in particular, when mixed with aromatic hydrocarbons it is especially suitable as a solvent for alkyd resins and ethylcellulose lacquers. The ability of 2-butanol to dissolve both water and oils is applied in the manufacture of brake fluids and cleaning agents.

2-Butanol is converted into amines by treatment with ammonia. The acetate is recommended as a good solvent for nitrocellulose; the xanthate is used in ore flotation. 2-Butanol also finds application in the manufacture of perfumes, dyestuffs, fruit essences, and wetting agents.

2-Methyl-2-propanol. Consumption of 2-methyl-2-propanol is small in comparison with the other butanols, although its use appears to be growing. The solvent sector is again the major field of application. 2-Methyl-2-propanol is also used as an agent for introducing the *tert*-butyl group into organic compounds (e.g., *tert*-butylphenol for the preparation of oil-soluble resins and antioxidants, trinitro-*tert*-butyltoluene as artificial musk) and as a starting material for the preparation of peroxides (polymerization catalysts).

2-Methyl-2-propanol competes with 2-propanol as a gasoline additive to prevent carburetor freezing. The methylether can be used as an antiknock agent in gasoline. Its use as a fuel additive increased from 175 000 t in 1982 to 400 000 t in 1983.

8. Economic Aspects

Both in the industrialized countries and worldwide, there is considerable excess capacity for the manufacture of 1-butanol and 2-methyl-1-propanol, the two major products in this group of substances (Table 5).

Table 5. Butanol capacities in the leading producer countries in 1984

	1-Butanol/∗ 2-methyl-1-propanol, 1000 t	2-Butanol, 1000 t
United States	700	
F.R. of Germany	300	60
Western Europe		422 ∗∗
Japan	206	146

∗ Any 1-butanol production plant also can be used for the manufacture of 2-methyl-1-propanol;
∗∗ 1982

Plant utilization was approx. 60 – 70 % in 1983.

The uses can vary considerably from one region to another. As well as direct use as a sol vent the main areas of application are in plasticizers, butyl acetate, acrylic esters, and glycol esters (see Table 6).

Table 6. Butanol consumption by end use, %

	Japan (1982)	USA (1980)
Plasticizers	13	8
Solvents	18	17
Butyl acetate	20	9
Butyl acrylate	24	29
Glycol ethers and esters	13	23
Miscellaneous	12	14

9. Toxicology and Occupational Health [74–78]

The main effects of exposure to excessive concentrations of all butanols are irritation of the mucous membranes and depression of the central nervous system. Animal studies have shown low acute oral, dermal, and inhalation toxicity. There is insufficient data to make a reliable assessment of the long-term effects in animals. The limited amount of documented experience in humans suggests that excessive exposure can cause irritation, headache, nausea, fatigue, and dizziness.

9.1. 1-Butanol

The lowest median lethal dose (LD_{50}) published in the literature is 790 mg/kg (rat, oral) and 4200 mg/kg (rabbit, dermal) [74]. Inhalation of

high concentrations of 1-butanol by experimental animals results in pulmonary and eye irritation, incoordination, and narcosis. The lowest LC_{50} published in the literature is 8000 ppm (rats, inhal., 4 h) [74]. Application of drops of 1-butanol into the rabbit eye produces severe corneal irritation. 1-Butanol has a slight to moderate irritating effect on rabbit skin [75]. 1-Butanol was found to be non-mutagenic in the Ames test [75].

Conjunctivitis and keratitis, sometimes accompanied by small intracorneal vacuoles, were reported among workers exposed to 1-butanol vapors. Intracorneal vacuoles proved to be reversible after exposure had ceased [76]. In a ten-year-study, workers exposed to 1-butanol at concentrations of 200 ppm and above developed corneal inflammation associated with a burning sensation, blurring of the vision, lacrimation, and photophobia. The mean erythrocyte count was slightly lowered. There were only rare complaints of irritation when the concentration was 100 ppm [75].

Further data indicate that eye irritation may be expected when the concentration of 1-butanol is above 50 ppm [76]. According to another study humans exposed to 1-butanol showed mild irritation already at 25 ppm, followed by headaches at 50 ppm [77]. Workers exposed to 80 ppm experienced a greater hearing loss (hypoacusia) in comparison with an unexposed group [77]. Dermatitis of the fingers and hands upon exposure to 1-butanol is also reported [76].

Exposure limits:

TLV (ceiling limit) 50 ppm (skin) [77],
OSHA air standard (TWA) 100 ppm [74],
MAK value 100 ppm.

9.2. 2-Butanol

The lowest LD_{50} published in the literature is 4400 mg/kg (rat, oral) [77]. 2-Butanol, like the other butanols, is a depressant to the central nervous system. The lowest lethal concentration published in the literature is 10670 ppm for mice exposed for 225 min; repeated exposure of mice to 5330 ppm produced narcosis without death; and no acute signs of intoxication were observed in mice exposed to 1650 ppm for 420 min [75].

2-Butanol caused severe corneal injury when applied to the rabbit eye but was not found to be irritating to the rabbit skin [75].

According to one report industrial exposure to about 100 ppm has not resulted in any difficulties [77]. Another study concludes that irritation of the eyes, nose, and throat, headache, nausea, fatigue, and dizziness have been experienced from excessive exposure to 2-butanol [75].

Exposure limits:

TLV 100 ppm (TWA); 150 ppm (STEL) [77],
OSHA air standard 150 ppm (TWA),
MAK value 100 ppm.

9.3. 2-Methyl-1-propanol

The lowest published LD_{50} is 2460 mg/kg (rat, oral) [75]. A dermal LD_{50} of 4250 mg/kg in the rabbit indicates a low toxicity due to local application [75]. The lowest LC_{50} for 2-methyl-1-propanol is 8000 ppm (rats, inhal., 4 h) [77]. The compound produces narcosis in mice that had been intermittently exposed to 6400 ppm for a total of 136 h. The narcotic dosage is followed by slight organic changes in the liver and kidneys. Repeated exposure of mice to 2125 ppm for 9.25 h did not cause any adverse effect [76].

Application of 2-methyl-1-propanol to the rabbit eye causes moderate to severe irritation. A moderate irritation was observed after 24-h exposure of the rabbit skin [74].

The results of long-term studies in which a small number of rats received the material either orally or subcutaneously suggest that 2-methyl-1-propanol may have a carcinogenic potential [75]. Exposure of *E. coli* bacteria resulted in mutagenesis [74].

In humans, there was no evidence of eye irritation during repeated 8-h exposure to levels in the order of 100 ppm [75]. Industrial exposure to mixed vapors of butyl acetate and 2-methyl-1-propanol has caused vacuolar keratitis in several workers but it is not known which of the substances was responsible [76]. Application of 2-methyl-1-propanol to human skin resulted in slight erythema and hyperemia [76].

Exposure limits:

TLV 50 ppm (TWA), 75 ppm (STEL) [77], OSHA air standard 100 ppm (TWA) [74], MAK value 100 ppm.

9.4. 2-Methyl-2-propanol

The lowest LD_{50} is reported to be 3500 mg/kg (rat, oral) [74]. The acute effect of the bad-smelling 2-methyl-2-propanol is that of a narcotic. Prolonged contact with rabbit skin failed to cause irritation [75]. The testing of the compound by *NTP* (National Toxicology Programm, U.S. Department of Health and Human Services) for carcinogenesis bioassay began in June 1983 [74].

Following application to human skin there was no reaction other than slight erythema and hyperemia. Irritating effects, headache, nausea, fatigue, and dizziness are reported as symptoms of excessive exposure [75].

Exposure limits:

TLV 100 ppm (TWA), 150 ppm (STEL) [77], OSHA air standard 100 ppm (TWA) [74], MAK value 100 ppm.

10. References

1. R. W. Gallant, *Hydrocarbon Process.* **46** (1967) no. 1, 183 – 189. R. W. Gallant: *Physical Properties of Hydrocarbons,* **vol. 1,** Gulf Publ. Comp., Houston, Texas, 1968, pp. 84 – 91.
2. H. A. Gundry, A. J. Head, G. B. Lewis, *Trans. Faraday Soc.* **58** (1962) 1309 – 1312.
3. C. Mosselman, H. Dekker, *Recl. Trav. Chim. Pays-Bas* **88** (1969) no. 3, 257 – 265.
4. J. F. Counsell, E. B. Lees, J. F. Martin, *J. Chem. Soc. A* 1968, no. 8, 1819 – 1823.
5. N. S. Berman, Doctoral Thesis (1962), Univ. of Texas, University Microfilms, Ann Arbor, Michigan; *Chem. Eng. Prop.* **59** (1963) no. 1, 85.
6. R. J. L. Andou, J. E. Connett, J. F. Counsell, E. B. Lees, J. F. Martin, *J. Chem. Soc. A* 1971, no. 4, 661 – 664.
7. D. Ambrose, C. H. S. Sprake, *J. Chem. Soc.* 1971, no. 9, 1261 – 1262.
8. E. T. Beynon, Jr., J. J. McKetta, *J. Phys. Chem.* **67** (1963) 2761 – 2765.
9. E. T. Beynon, University Microfilms, Ann Arbor, Michigan, Order no. 65-8025, 75 pp.; *Diss. Abstr.* **26** (1965) no. 2, 930; *Chem. Abstr.* **63** (1965) no. 12, 16196.
10. W. B. Kay, W. E. Donham, *Chem. Eng. Sci.* **4** (1955) 1 – 16.
11. M. Hirata, *Kagaku To Seibutsu* **2** (1964) 311 – 316; *Chem. Abstr.* **64** (1966) 12203.
12. E. P. Ageev, G. M. Pančenkov, V. A. Spivak, *Zh. Fiz. Khim.* **41** (1967) no. 5, 1155 – 1157.
13. H. A. Rizk, N. Youssef, *Z. Phys. Chem. Frankfurt/M* 58 (1/4) (1968) 100 – 113; Z. Phys. Chem. Leipzig NF 244 (1970) no. 3/4, 165 – 172.
14. G. E. Rajala, J. Crossley, *Can. J. Chem.* **49** (1971) no. 22, 3617 – 3622.
15. D. Davis, *Ind. Chem.* **39** (1963) no. 2, 75.
16. J. J. Kipling, *J. Colloid Sci.* **18** (1963) 502 – 511.
17. D. Hessel, G. Geiseler, *Z. Phys. Chem. Leipzig* **229** (1965) no. 314, 199 – 209.
18. V. M. Komarov, B. K. Kricevcov, *Zh. Prikl. Khim. Leningrad* **39** (1966) 2604 – 2606.
19. A. I. Alzybejewa, W. P. Beloussow, N. W. Owtracht, A. G. Moratschewski, *J. Phys. Chem. (Moscow)* **38** (1964) 1242 – 1247.
20. G. Schneider, C. Russo, *Ber. Bunsenges. Phys. Chem.* **70** (1966) no. 9/10, 1008 – 1014.
21. H. Arm, *Helv. Chim. Acta* **45** (1962) 1803 – 1806.
22. A. G. Moračevskij, N. A. Smirnova, R. V. Lyzlova, *Zh. Prikl. Khim. Leningrad* **38** (1965) 1212 – 1267.
23. A. G. Moračevskij, Z. P. Popovic, *Zh. Prikl. Khim. Leningrad* **38** (1965) 2129 – 2131.
24. *Handbook of Chemistry and Physics,* 55th ed., CRC Press, Inc., Cleveland, Ohio, 1974 /75.
25. Hoechst AG/Ruhrchemie AG: *Oxo Synthesis Products (Aldehydes, Alcohols, Acids),* 2nd ed., 1971.
26. L. H. Horsley, W. S. Tamplin: "Azeotropic Data III," *Adv. Chem. Ser.* **116** (1973) .
27. F. Asinger: *Die Petrochemische Industrie,* Akademie-Verlag, Berlin 1971, p. 302.
28. B. C. Gates, J. S. Wisnouskas, H. W. Heath, *J. Catal.* **24** (1972) no. 2, 320 – 327.
29. H. Pines, W. O. Haag, *J. Am. Chem. Soc.* **82** (1960) 2471.
30. H. Knözinger, R. Köhne, *J. Catal.* **5** (1966) 264 – 270.
31. P. B. Weisz, *Erdöl Kohle Erdgas Petrochem.* **18** (1965) no. 7, 525 – 527.

32. L. Bexten, J. Weber in: *Methodicum Chimicum,* **vol. 5** (1975) p. 277.
33. E. Buckley, J. D. Cox, *Trans. Faraday Soc.* **63** (1967) 895.
34. A. Amariglio, X. Duval, *Bull. Soc. Chim. Fr.* 1966, 2100 – 2106.
35. K. J. Miller, J. L. Wu, *J. Catal.* **27** (1972) no. 1, 60 – 63.
36. U. R. Rao, R. Kumar, N. R. Kuloor, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 9.
37. *Hydrocarbon Process. Pet. Refin.* **44** (1965) no. 11, 234.
38. B. Cornils, E. Zilly in: *Methodicum Chimicum,* **vol. 5** (1975) p. 84.
39. H. Koch, W. Haaf, *Justus Liebigs Ann. Chem.* **618** (1958) 251 – 266.
40. Ruhrchemie AG, DE-OS 1793369, 1968 (J. Falbe, B. Cornils, J. Meis, H. Tummes, J. Weber).
41. L. M. Kozlova, I. A. Romaan, *Izv. Akad. Nauk Latv. SSR Ser. Khim.* 1968, no. 6, 727 – 730.
42. *Methodicum Chimicum,* **vol. 5 (1975) p.** 637.
43. J. Falbe: *Carbon Monoxide in Organic Synthesis,* Springer Verlag, Berlin-Heidelberg-New York 1970.
44. UCC, DE-AS 2062703, 1970. Johnson Matthey Ltd., DE-AS 1939322, 1969.
45. RP-Industries, DE 2627354, 1976. Stamicarbon, EP 40891, 1980.
46. *Ullmann,* 3rd ed., supplement vol., p. 87.
47. W. Reppe, H. Vetter, *Justus Liebigs Ann. Chem.* **582** (1953) 133.
48. N. V. Kutepow, H. Kindler, *Angew. Chem.* **72** (1960) 802.
49. H. Kindler, K. Eisfeld, *Ind. Chim. Belge* **32** (1967) 650.
50. *Ullmann,* 3rd ed., **vol. 5,** p. 134.
51. BASF, DE-AS 1114796, 1960; DE-AS 1114797, 1960.
52. *Hydrocarbon Process.* **50** (1971) no. 11, 138.
53. Kyowa Hakko Kogyo, *Hydrocarbon Process.* **48** (1969) no. 11, 160.
54. W. Swodenk, *Chem. Ing. Tech.* **55** (1983) 683.
55. *Winnacker-Küchler,* 3rd ed., **vol. 3,** pp. 349, 386.
56. H. Friz, *Chem. Ing. Tech.* **40** (1968) 999.
57. H. Bahrmann, W. Lipps, B. Cornils, *Chem. Ztg.* **106** (1982) 249.
58. H. Kropf, *Chem. Ing. Tech.* **38** (1966) 837.
59. ICI, US 2815391, 1953.
60. M. Sittig: *Chemicals from C_4-Hydrocarbons,* Noyes Development Corp., Park Ridge 1966, pp. 73 – 79.
61. ICI, US 3164641, 1959.
62. Montecatini, US 3164641, 1959.
63. Gulf Research & Development Co., US 3006970, 1958.
64. Jap. Oil, DE-OS 1944392, 1968 (T. Horie).
65. CWH AG, DE 2629769, 1976 (F. Obenaus et al.); Hydrocarbon Process. 61 (1982) no. 3, 121 – 123. *Hydrocarbon Process.* **60** (1981) no. 8, 101 – 106.
66. *Annual Book of ASTM Standards,* Section 6, vol. 06.03: "Paint-Fatty Oils and Acids, Solvents, Miscellaneous; Aromatic Hydrocarbons; Naval Stores," American Society for Testing and Materials, Philadelphia 1984.
67. *DIN-Taschenbuch 30,* Beuth-Vertrieb GmbH, Berlin-Köln-Frankfurt 1971.
68. *British Standard Yearbook,* British Standards Institution, Inc., Royal Charter, London 1971.
69. Hoechst AG: *"Lösemittel Hoechst,"* VOC L, Frankfurt 1984. Hoechst AG: *"Hoechst Solvents, Manual for Laboratory and Industry,"* 5th ed., Frankfurt 1975.
70. G. Sorbe, *GIT-Fachz. Lab.* **16** (1972) no. 5, 598.
71. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* 2nd ed., Deutscher Eichverlag, Berlin 1963.
72. T. Hammerich, H. Schildwächter, *Erdöl Kohle Erdgas Petrochem.* **18** (1965) no. 12, 972 – 976.
73. *Chem. Eng. Int. Ed.* **79** (1972) 32 – 36.
74. R. J. Lewis, R. L. Tatken (eds.): *Registry of toxic effects of chemical substances, National Institute for Occupational Safety and Health (RTECS ONLINE data base)* DIMDI, 10. 01. 84.
75. V. K. Rowe, S. B. McCollister in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 C,** Wiley Interscience, New York 1982, pp. 4571-4588.
76. E. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, Amsterdam-London-New York 1965, pp. 342 – 355.
77. American Conference of Governmental Industrial Hygienists: *Documentation of Threshold Limit Values for substances in workroom air,* Cincinnati, Ohio 1980, vol. 4, pp. 54, 55, 235.
78. A. I. Cederbaum, G. Cohen, *Biochem. Biophys. Res. Commun.* **97** (1980) no. 2, 730 – 736.

2-Butanone

WILHELM NEIER, Deutsche Texaco AG, Moers, Federal Republic of Germany

GUENTER STREHLKE, Deutsche Texaco AG, Moers, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Production 3
4.1. Catalytic Dehydrogenation of *sec*-Butyl Alcohol (SBA) in the Gaseous Phase 3
4.2. Liquid-Phase Oxidation of *n*-Butane . 4
4.3. Direct Oxidation of *n*-Butenes (Hoechst-Wacker Process) 4
5. Quality, Storage, Transportation . . . 5
6. Uses 5
7. Economic Aspects 6
8. Toxicology 6
9. References 7

1. Introduction

2-Butanone [*78-93-3*], methyl ethyl ketone, MEK, is the second link in the homologous series of aliphatic ketones and, next to acetone, the most important commercially produced ketone. 2-Butanone is produced primarily by dehydrogenation of 2-butanol, analogous to the production of acetone by dehydrogenation of gaseous isopropyl alcohol on copper, zinc, or bronze catalysts at 400 – 550 °C. At 80 – 95 % *sec*-butyl alcohol conversion, MEK selectivity is > 95 %. In some cases MEK can be produced in the same facilities as acetone. Butenes (dehydration) and higher ketones (autocondensation) are byproducts. In 1995, 730 000 t of MEK were produced worldwide.

2-Butanone is produced by ARCO (US), AKZO (NL), Biochimica (ES), BP (GB), Carbochlor (Argent.), Celanese (US, CA, Mexico), Esso (GB, US), Maruzen (JP), PCUK (FR), Shell (FR, NL, US, CA), Texaco (FRG), Tonen (JP), UCC (US), Sasol (ZA). Interest in MEK as a solvent for paints and adhesives has been growing since the 1980s. In general, MEK is considered to be a competitor for ethyl acetate, especially as a low-boiling solvent. It has broad application as a solvent for nitrocellulose, cellulose acetate – butyrate, ethylcellulose, acrylic resins, vinylacetates, and vinylchloride – vinylacetate copolymer (based on synthetic surface-coating preparation). It is favored as a lacquer solvent because of its low viscosity, high solids concentration, and great diluent tolerance.

Moreover, MEK can be used as an activator for oxidative reactions, as a selective extractant, as a special solvent for dewaxing mineral oil fractions, and as a chemical intermediate. (see Table 1).

Table 1. Physical data of 2-butanone [9], [10]

M_r	72.11
mp, °C	− 86.9
bp, °C	79.6
Relative density,	
d_4^{20}	0.8045
d_{20}^{20}	0.80615
Refractive index, n_D^{20}	1.3788
Evaporation number (ether=1)	2.6
Critical temperature, °C	262.45
Critical pressure, MPa	4.15
Critical density, ϱ, g/L	0.270
Dynamic viscosity at 20 °C, mPa · s	0.323
Surface tension at 20 °C, mN/m	24.6
Molar heat c_p at 23.8 °C, J mol^{-1} K^{-1}	160.8
Heat of fusion, J/mol	7456
Heat of vaporization at 79.6 °C, J/mol	31.2
Combustion enthalpy at constant pressure, 25 °C, kJ/mol	2444.3
Thermal conductivity, W m^{-1} K^{-1}	
at 0 °C	0.150
20 °C	0.145
50 °C	0.137
Solubility at 20 °C	
2-butanone in water, mass fraction, %	27.5
water in 2-butanone, mass fraction, %	12.5
Flash point (DIN 51755), °C	−1
Explosion limits in air at 20 °C, 101.3 KPa,	
lower, volume fraction, %	1.8
upper, volume fraction, %	11.5
Ignition temperature, °C	505
Ignition class (VED)	G 1
Explosion class (VED)	1
Electric conductivity at 20 °C, Ohm^{-1} cm^{-1}	5×10^{-8}
Dipole moment, Debye	3.18
Dielectric constant of the liquid at 20 °C	15.45

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_475

2. Physical Properties

2-Butanone, methyl ethyl ketone, MEK, $CH_3COCH_2CH_3$, is a relatively mobile, colorless liquid. Its typical odor resembles that of acetone. The compound is only partially water-miscible, whereas it is completely miscible with most organic solvents. 2-Butanone forms binary and ternary azeotropic mixtures in combination with water and several other organic solvents (see Table 2).

Table 2. Binary azeotropic mixtures containing MEK [11]

Second component	Mass fraction of MEK	*bp* at 1013 hPa, °C
Water	88.7	73.4
Benzene	37.5	78.4
n-Hexane	29.5	64.3
n-Heptane	73.0	77.0
Cyclohexane	40	72
1,3-Cyclohexadiene	40	73
Methanol	30.0	63.5
Ethanol	60.9	74.0
Isopropyl alcohol	68.0	77.5
tert-Butyl alcohol	73	77.5
Ethyl acetate	18.0	77.0
Methyl propionate	52	79.25
Propyl formate	55	79.45
Chloroform	96.0	79.65
Carbon tetrachloride	71.0	73.8
Carbon disulfide	15.3	45.85
Propyl mercaptan	75	55.5
Thiophene	55.0	76.0
Ethyl sulfide	20	77.5

2-Butanone does not form a binary azeotropic mixture with toluene, *m*-xylene, *n*-butanol, isobutanol, *sec*-butyl alcohol, 4-methyl-2-pentanol, allyl alcohol, acetic acid methyl ester, acetic acid isopropyl ester, acetic acid *n*-butyl ester, acetic acid isobutyl ester, 1,1-dichloroethane, isobutyl chloride, propyl bromide, ethyl iodide, trichloroethylene, dichlorobromomethane, isobutyl bromide, formic acid, and acetic acid.

2-Butanone forms ternary azeotropes with water/benzene and water/carbon tetrachloride [11].

3. Chemical Properties

Under normal conditions and in the absence of atmospheric oxygen MEK is stable. Care must be taken after prolonged storage because peroxides may form in the presence of oxygen [12].

2-Butanone is unsaponifiable and, unlike esters, does not form corrosive products upon hydrolysis. It is heat and light stable. It decomposes only after prolonged UV exposure (yielding ethane, methane, carbon monoxide, ethylene, and diacetyl) [13].

Diacetyl [*431-03-8*] is formed by oxidation with airin the presence of special catalysts [14].

Methyl ethyl ketone peroxide [*19393-67-0*], a polymerization catalyst, is formed by oxidation with a 30 % solutionof hydrogen peroxide [15]. Nitric acid and other strong oxidants oxidize MEK to a mixture of formic and propionic acids [15].

sec-Butyl alcohol [*78-92-2*] is obtained by catalytic reduction with hydrogen [15]. It can also be formed by electrolytic reduction in sodium acetate solution or by reduction with ammonium amalgam or lithium aluminum hydride. 3,4-Dimethyl-3,4-hexanediol is obtained by electrolytic reduction in an acidic medium or by reduction with magnesium amalgam [15].

Methyl ethyl ketone forms addition products with hydrogen cyanide as well as with sodium and potassium hydrogen sulfites. In an alkaline medium MEK condenses with aldehydes to form higher unsaturated ketones. Condensation with formaldehyde to form methyl isopropenyl ketone [*563-80-4*], an intermediatefor further syntheses, is of particular interest. During base-catalyzed autocondensationin the liquid phase and during gase-phase condensation on alkalinized copper catalysts, the carbonyl group reacts with the methyl group, whereas during acid-catalyzed condensation the methylene group in α-position to the carbonyl group is attacked [16].

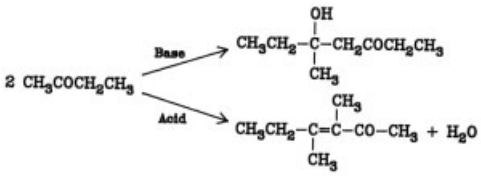

Methyl ethyl ketone and citral [*5392-40-5*] condense toform methylpseudoionone that can be cyclized to methylionone, a compound used for producing synthetic violet perfume.

During condensation with low-molecular aldehydes (during base-catalyzed and acid-catalyzed aldolization) the α-position of the carbonyl group is first occupied [17], [18].

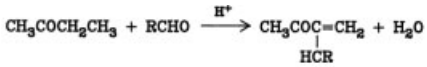

Thus, the base-catalyzed aldolization with less than amounts of formaldehyde yields 2-methyl butane-1-ol-3-one [17] and exhaustive hydroxymethylation, with reduction of the carbonyl group (crossed Cannizzaro reaction), produces desoxyanhydroeneaheptite [19].

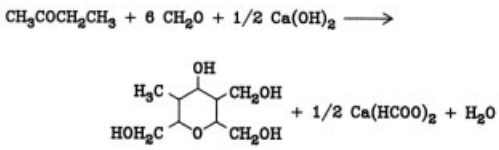

When MEK is reacted with primary and secondary alcohols, higher ketones are obtained. Reaction with *sec*-butyl alcohol gives ethyl amyl ketone [*106-68-3*] [20].

Methyl ethyl ketone reacts with polyoxy compounds or epoxides to form cyclic products.

Amyl nitrite [*110-46-3*] attacks the CH_2 groupin α-position to the carbonyl group and yields themonooxime of diacetyl.

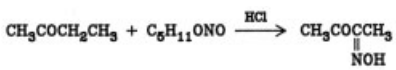

The keto group reacts with amino groups with elimination of water. In combination with hydroxylamine [*7803-49-8*], methyl ethylketoxime, an antiskinning agent, is formed.

Condensation of MEK with aliphatic esters and anhydrides gives β-diketones.

Phenols react with MEK to form oxyphenylene compounds. In combination with phenol, 2,2-hydroxyphenyl butane is obtained, a homolog of hydroxyphenyl propane (Bisphenol A [*80-05-7*]), an important materialfor the production of synthetic resins.

Methyl ethyl ketone can be halogenated in the α-position. Methyl ethyl ketone reacts with Grignard compounds to form tertiary alcohols. With acetylene in the presence of sodium amide 3-methyl-1-pentyn-3-ol [*77-75-8*] is formed. *N*-Methyl-formyl-aminobutaneis obtained from MEK plus *N*-methylformamide. The Reformatzky reaction produces the β-oxyester from monobromine-substituted esters [20].

4. Production

Most MEK (88 %) is produced today by dehydrogenation of *sec*-butyl alcohol (SBA). SBA can easily be produced by hydration of *n*-butenes (from petrochemically produced C_4 raffinates) in a two-step process (catalyst: sulfuric acid), or in a single-step process by direct addition of water, acidic ion-exchange resins being used as a catalyst [21]. The remaining 12 % MEK is produced by processes in which liquid butane is catalytically oxidized, giving both acetic acid and MEK [22].

The direct oxidation of *n*-butenes (Hoechst-Wacker process, Maruzen process, [23], [24]) has not been generally accepted, because of undesired byproducts.

The *sec*-butylbenzenehydroperoxide route giving phenol and MEK by acid-catalyzed splitting [25] is uneconomical. The autoxidation of liquid *sec*-butyl alcohol, giving MEK and hydrogen peroxide [26], and the catalytic oxidative hydration of gaseous *n*-butenes [27] are also uneconomical. The oxidation of *n*-butenes with ethylbenzenehydroperoxide to form butylene oxides, and subsequent hydration and formation of ketones appears to be moderately attractive [28]. Styrene, *n*-butanol, and MEK are obtained in coupled production.

4.1. Catalytic Dehydrogenation of *sec*-Butyl Alcohol (SBA) in the Gaseous Phase

The catalytic dehydrogenation of SBA is an endothermic reaction (51 kJ/mol). The equilibrium constant for SBA can be calculated as follows [29]:

$$\log K_{\mathrm{p}} = -2.790 \times T^{-1} + 1.51 \log T + 1.865$$
$$(T = \text{reaction temperature, in K})$$

The MEK concentration in the reaction mixture increases with the temperature and reaches its maximum at approx. 350 °C [30].

Copper [31], zinc [32], or bronze [33] are used as catalysts in gas-phase dehydrogenation. The latter two require high dehydrogenation temperatures (400 °C). Dehydration of SBA to butenes takes place as a side reaction on

zinc oxide. Platinum on alumina [34], copper or chromium [35], [36] as well as copper, and zinc on alumina [34] are recommended as dehydrogenation catalysts for aqueous SBA.

Commercially used catalysts are reactivated by oxidation, after 3 to 6 months use. They have a life expectancy of several years. Catalyst life and alcohol conversion are impaired by contamination with water, butene oligomers, and di-*sec*-butyl ether [37].

Deutsche Texaco developed a process in which practically anhydrous *sec*-butyl alcohol has been produced since 1983 by direct hydration of *n*-butene, catalyzed by acidic ion-exchange resin; it is then dehydrogenated on a copper-based precipitating catalyst at 240 – 260 °C under normal pressure [21], [31]. The LHSV (4 L/L of catalyst · h), the conversion (90 – 95 %), and the catalyst life until reactivation becomes necessary (3 – 4 months) render the process economically attractive. (see Figure 1).

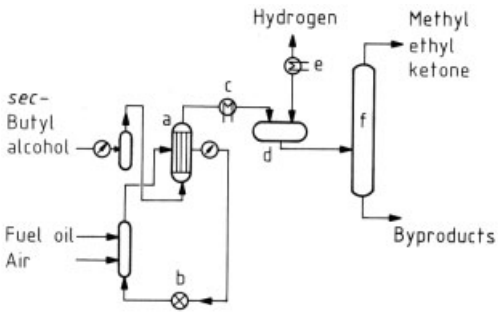

Figure 1. Gaseous-phase dehydrogenation of *sec*-butyl alcohol (Deutsche Texaco AG process)
a) Reactor; b) Oil circulation heating; c) Condenser; d) Separator; e) Refrigerator; f) Distillation

sec-Butyl alcohol is dehydrogenated in a multi-tube reactor. The reaction heat (51 kJ/mol) is supplied by heat transfer oil. The reaction products leave the reactor as a gas and are split into liquid crude MEK and hydrogen on cooling. The hydrogen is purified by further cooling. The reaction is highly selective. Autocondensation to higher ketones (e.g., 5-methyl-3-heptanone) is much lower in comparison to acetone. In addition, the dehydration to butenes on copper catalysts is for the most part prevented.

During reaction and subsequent treatment, practically no waste disposal problems arise.

Table 3 lists further processes for the production of MEK by gas-phase hydrogenation of *sec*-butyl alcohol.

4.2. Liquid-Phase Oxidation of *n*-Butane

2-Butanone is a byproduct in the liquid-phase oxidation of *n*-butane to acetic acid. Autoxidation of *n*-butane takes place in liquid phase according to a radical mechanism yielding MEK as an intermediate and acetic acid as the end product. The continuous plug flow process developed by Union Carbide allows the partial collection of MEK intermediate [42]. MEK and acetic acid (mass ratio 0.15 – 0.23 : 1.0) are obtained by non-catalyzed liquid-phase oxidation at 180 °C and 5.3 MPa (53 bar) with remixing. Continuous oxidation under plug flow conditions at 150 °C, 6.5 MPa (65 bar), and a residence time of 2.7 min forms MEK and acetic acid at mass ratios of up to 3 : 1 [42]. Celanese uses acetic acid as a solvent and cobalt acetate and sodium acetate as homogeneously dissolved catalysts [43]. It is a batch process performed at 160 – 165 °C and 5.7 MPa (57 bar). MEK and acetic acid are obtained at a mass ratio of 0.4 : 1.0.

4.3. Direct Oxidation of *n*-Butenes (Hoechst-Wacker Process)

In the direct oxidation of *n*-butenes according to the Hoechst-Wacker process, oxygen is transferred in a homogeneous phase onto *n*-butenes using a redox salt pair, $PdCl_2$/2 CuCl [44], [45]. The salt pair is subsequently reoxidized.

$$n\text{-}C_4H_8 + PdCl_2 + H_2O \longrightarrow CH_3COC_2H_5 + Pd + 2\ HCl$$

$$Pd + 2\ CuCl_2 \rightleftharpoons PdCl_2 + 2\ CuCl$$

n-Butenes can be converted into the following reaction products (conversions of up to 95 % are attained):

Reaction products:	Selectivities, mol %
MEK	86
n-Butyraldehyde	4
Chlorinated products	6
Carbon dioxide	1

The main disadvantages are: formation of chlorinated butanones and *n*-butyraldehyde and corrosion caused by free acids.

Table 3. Further processes for producing methyl ethyl ketone by gase-phase dehydrogenation of *sec*-butyl alcohol

Company	Catalyst	H_2O content, vol %	Temperature, °C	Pressure, MPa	Conversion, %	Selectivity, mol %	Yield, mol %	Reference
Standard Oil	ZnO/Bi_2O_3	0	400	0.1 – 0.3			80	[38]
Esso Research & Eng.	$ZnO/Na_2CO_3/Al_2O_3$	0	413		96 – 97	97	93 – 94	[39]
Maruzen Oil	bronze	0	390	0.3	80	99	79.2	[33]
Knapsack Griesheim	CuO/CrO	0	270 – 320		88 – 93	93 – 96		[40]
Toyo Rayon	$CuO/NaF/SiO_2$		300		96	100		[41]
Ruhrchemie	60 % Cu, Cr_2O_3, MgO, 12 % SiO_2, 10 % H_2O		260		90	100	90	[35]
Veba-Chemie	22 % Cu, 8 % $BaCrO_4$, 2 % Cr_2O_3, 0.5 % Na_2O, 61 % SiO_2	90.4	180		57.3 [a]	63 [b]	97.8 [c]	[36]
Shell-Chemie	0.05 % Pt/Al_2O_3	0	358	0.6	92.5 – 93.5	96.3	90	[37]
Shell-Chemie	5 % Cu, 5 % Cr/Al_2O_3	0	286	0.6	81 – 85	86 – 92	73 – 77	[37]

[a] Relative to *sec*-butyl alcohol.
[b] Relative to di-*sec* -butyl ether.
[c] Total yield.

The Maruzen process is similar [46], [47]. Oxygen is transferred by an aqueous solution of palladium sulfate and ferric sulfate.

Other processes employing the same oxygen transfer principle were developed by Consortium für Elektrochemie [48] and Eastman Kodak [49]. For further processes, see [50–53].

5. Quality, Storage, Transportation

It is possible to produce high-purity MEK (DIN 53247, ASTM D 740, BS 1940).

The present sales specifications of Deutsche Texaco AG are listed in Table 4.

Table 4. Sales specifications for MEK

Purity, wt %	GC	min. 99.7
d_4^{20}	(DIN 51757)	0.804 – 0.806
d_{20}^{20}		0.805 – 0.807
Boiling range, °C	DIN 51751	within 0.5
	ASTM D 1078	incl. 79.6
Acidity (as acetic acid), mg KOH/g		0.0001
Water content, wt %,	DIN 51777	
	ASTM D 1364	max. 0.1
Color (Pt – Co, APHA),	ASTM D 1209	max. 10
Nonvolatile matter, wt %		max. 0.002

A stable dilute potassium permanganate solution indicates high purity (permanganate time according to ASTM D 1363).

Storage life of MEK is limited. Carbon steel containers (ST 3529) are suitable for short-term storage and transportation. Stainless steel (316 SS) or containers with a tin lining are recommended for long-term storage. Once autocatalysis has started, it continues even if storage is continued in inert containers. During long-term storage the formation of peroxide must be prevented. Since MEK is somewhat hygroscopic, water is absorbed from the air.

The following regulations for transportation of MEK must be observed [54].

IMDG-Code: D 3308, Kl. 3.2 UN no. 1193, RID, ADR, ADNR: class 3, Rn 301, 2301, and 6301 respectively, no. 1 a, category Kl n; European Council, Yellow Book 78/79: no. 606-002-00-3; European Communities: Guideline/D VgAst, no. 606-002-00-3; UK: Blue Book, Fla.L. IMDG-Code E 3080; USA: CFR 49, 172.101, Fla.L.; IATA RAR: art. no. 726 Fla.L.

6. Uses

MEK is an important solvent with properties similar to those of acetone. MEK has the following advantages in comparison to other solvents with comparable rates of evaporation: very high power of dissolution, high ratio of dissolved matter to viscosity, miscibility with a large number of hydrocarbons without impairing the solids content or viscosity, favorable volume/mass ratio due to its low density.

The following natural substances, plastics, and resins can be dissolved in MEK: rosin, ester resins, pentaerythritol ester resins, Congo ester, dammar (dewaxed), nitrocellulose, low-molecular cellulose acetate, cellulose acetobutyrate, cellulose acetostearate, methylcellulose, epoxy resins, nearly all alkyd and phenolic resins, polyvinylacetate, vinylchloride/acetate mixed polymerizates, vinylchloride/vinylidene chloride mixed polymerizates, coumarone – indene resins, sulfonamide resins, cyclohexanone resins, acrylic resins, polystyrene, chlorinated rubber, polyurethane.

Cellulose triacetate, high-molecular cellulose acetate, poly(vinylchloride), poly(vinylbutyral), polysulfide rubber cannot be dissolved in MEK. Shellac is only partially soluble.

Other areas of application are production of synthetic leather, transparent paper, printing inks, aluminum foil lacquers; degreasing of metal surfaces; extraction of fats, oils, waxes, natural resins; dewaxing of mineral oils [55].

In contrast to its uses as a solvent, use as a chemical feedstock is of minor importance despite the great number of possible reaction; however, condensation with formaldehyde to obtain methyl isopropenyl ketone, autocondensation to form ethyl amyl ketone, and mixed condensation with acetone to obtain methyl amyl ketone are of interest. Methyl ethyl ketoxime, used as an antiskinning agent in lacquers, is of minor importance. Methyl ethyl ketone peroxide is used as a polymerization initiator for unsaturated polyesters. Diacetyl serves as a butter flavorer. The perfume industry reacts MEK with citral to obtain perfume components such as methylpseudoionone. Since 1962 MEK is permitted as an alcohol denaturant in the Federal Republic of Germany (by decree of *Bundesmonopolverwaltung* in Offenbach).

7. Economic Aspects

Table 5 shows the 1979 sales (in %) for the various markets in the Federal Republic of Germany.

The worldwide consumption of MEK in 1979 amounted to 642000 t (Western Europe 166000 t, Eastern Europe 18000 t, North America 315000 t, Central and South America 34000 t, Asia, Australia, Oceania 100000 t, Africa 9000 t).

Table 5. Sales survey (FRG, 1979)

	%
Paints, lacquers, printing inks, aluminum foil lacquers	40
Coating and printing of plastics	20
Chemical industry (incl. sound carrier), pharmaceutical industry	13
Adhesives	11
Miscellaneous	16

MEK is also available under the following names:

German: Ethylmethylketon, Acetonersatz, β-Ketobutan, 2-Butanon

English: Butanone, 2-butanone, MEK, methylacetone, meetco

French: Méthyléthylcétone, butane-2-one, ethylméthylcétone, MEC

8. Toxicology

The odor threshold of MEK is 10 ppm; both, MAK and TLV are established at 200 ppm. The inhalation of MEK vapor has narcotic effects. The vapor irritates the eyes and the nasal and pharyngeal mucous membranes [54]. Frequent and prolonged contact with liquid MEK causes skin moisture loss and slight irritation [56]. Sensitive persons may develop dermatoses [57]. Liquid MEK temporarily irritates the eye and corneas [56].

MEK is usually absorbed through the respiratory tract. It may also be absorbed by the skin, but the cutaneous LD_{50} in rabbits is above 8 mL/kg [57]. The MEK metabolism has been studied in guinea pigs. MEK is both reduced to 2-butanol and oxidized to 3-hydroxy-2-butanone. However, unlike 2-hexanone and *n*-hexane, which are further oxidized to form neurotoxic 2,5-hexanedione after oxidation of the ω-1 C atom, the hydroxybutanone is not further oxidized, but converted to 2,3-butanediol [58].

Animal tests have shown that the neurotoxic effect of 2-hexanone may be potentiated by simultaneous administration of MEK [59], [60]. The chronic inhalation of 200 ppm (MAK, TLV) does not seem to be harmful [61].

Even workroom concentrations of 500–700 ppm over an extended period of time do not cause permanent damage.

The LD_{50} (oral, rat) is 2500–3400 mg/kg [62], [63].

Toxic concentrations for water organisms [54]: average lethal concentration for fish: 5600 mg/L. Maximum permissible concentrations for *Pseudomonas putida:* 1150 mg/L, for *Scenedesmus quadricanda:* 4300 mg/L, for *Microcystis aeruginosa:* 120 mg/L. For small crabs (*Daphnia magma*) the LC_0 is 2500 mg/L, the LC_{50} is 8890 mg/L, and the $LC_{100} > 10000$ mg/L.

9. References

General References

1. *Beilstein* **1,** 668, **1(1),** 347, **1(2),** 726, **1(3),** 2770, **1(4),** 3243.
2. F. Asinger: *Die Petrolchemische Industrie,* Akademie Verlag, Berlin 1971, p. 1086.
3. J. Mellan: *Ketones,* Chemical Publ., New York 1968.
4. Winnacker-Küchler: *Organische Technologie II,* **vol. 6,** Hanser Verlag, München 1982, pp. 77, 81.
5. *Römpps Chemie Lexikon,* 8th ed., **vol. 1,** p. 541.
6. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin 1980.
7. *Kirk-Othmer,* 3rd ed., **vol. 13,** p. 894.
8. *Ullmann,* 4th ed., **vol. 14,** p. 193 ff.

Specific References

9. VDI-Wärmeatlas 1983, VDI-Verlag Düsseldorf.
10. A. Weissberger: *Organic Solvents,* **vol. II, 3rd ed.,** Wiley-Interscience, New York 1970.
11. L. H. Horsley: *Azeotropic Data,* Advances in Chemistry Series 116, Am. Chem. Soc., Washington D.C. 1973.
12. O. L. Mageli, J. R. Kolczynski, Lucidol Division, Wallace & Tiesman Inc., *Encycl. Polym. Sci. Technol.* 1964–1972, vol. 9, 831; N. A. Milas, *J. Am. Chem. Soc.* **81** (1959) 5824.
13. W. Davis Jr., *Chem. Rev.* **40** (1947) 240–244.
14. *Beilstein* **(Syst. No. 95) 1,** 769, **1(1),** 397, **1(2),** 824, **1(3),** 3098.
15. *Beilstein* **1,** 668, **1(1),** 347, **1(2),** 726, **1(3),** 2770, **1(4),** 3243.
16. A. T. Nielsen, *Organic Reactions,* **vol. 16,** J. Wiley & Sons, New York 1968, p. 20.
17. Rheinpreußen AG, CA 605368, 1955 (W. Grimme, J. Wöllner).
18. Rheinpreußen AG, DE 1198814, 1963 (J. Wöllner, F. Engelhardt).
19. J. R. Roach, H. Wittcoff, S. E. Miller *J. Am. Chem. Soc.* **69** (1947) 2651.
20. Technical Brochure LB 11, Deutsche Shell, Frankfurt (Main) 1976.
21. Erdöl Informations-Dienst A. M. Stahmer, vol. 37, no. 28 (1984).
22. Union Carbide, US 3196182, 1965 (R. N. Cox).
23. Hoechst, US 3215734, 1965 (E. Katzschmann).
24. Maruzen Oil, JP 46-2010, 1971.
25. Union Carbide, DE-OS 2300903, 1973 (F. P. Wolf).
26. N. V. de Bataafsche Petroleum Maatschappij, DE 935503, 1955.
27. Stamicarbon, NL 69, 016990, 1971 (J. W. Geus, J. H. Kruit, P. P. Nobel).
28. SRI-Report, PEP-Review no. 80-3-1 (1981), Menlo Park, California.
29. H. J. Koll et al., *J. Am. Chem. Soc.* **67** (1945) 1084.
30. C. Padovani et al., *Riv. Combust.* **5** (1951) 81.
31. Rheinpreußen AG, DE 1147933, 1958 (R. Langheim, H. Arendsen).
32. Esso, US 2885442, 1959 (W. J. G. McCullock, I. Uirshenbaum).
33. Maruzen Oil, Technical Brochure, Aug. 1969.
34. Shell, DE-OS 2028350, 1970 (B. Stouthamer, A. Kwantes).
35. Ruhrchemie AG, DE-OS 2347097, 1973 (W. Rottig, C. O. Frohning, H. Liebern).
36. Veba-Chemie, DE-OS 1913311, 1969 (W. Ester, W. Heitmann).
37. Shell, DE 2028350, 1970 (B. Stouthamer, A. Kwantes).
38. Standard Oil, US 2436970, 1948 (V. F. Mistretta).
39. Esso, US 2835706, 1958 (C. E. Cordes).
40. Knapsack-Griesheim, DE 1026739, 1958 (W. Opitz, W. Urbanski).
41. Toyo Rayon, JP 43-3163, 1968.
42. Union Carbide, US 3196182, 1965.
43. Celanese, US 2704294, 1955 (C. S. Morgan, N. C. Robertson).
44. Consortium f. Elektrochemie, GB 878777, 1961.
45. J. Smidt et al., *Erdöl Kohle Erdgas Petrochem.* **16** (1963) 560.
46. Maruzen Oil, JP 46-2010, 1971.

47. Maruzen Oil, DE 1951759, 1970 (N. Irinchijma, H. Taniguchi).
48. Consortium f. Elektrochemie, US 3080425, 1963 (J. Smidt, W. Hafner, R. Jira).
49. Eastman Kodak, GB 1099348, 1968 (H. J. Hagemeyer, F. C. Canter).
50. Union Carbide, DE-OS 2300903, 1973 (F. P. Wolf).
51. Union Carbide, DE-OS 2421168, 1974 (B. J. Argento, E. A. Rick).
52. F. Conssenant, US 2829165, 1958.
53. C. S. Cronan, *Chem. Eng. Int. Ed.* **67** (1960) Feb. 8, 63.
54. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin-Heidelberg-New York 1980.
55. W. P. Gee et al., *Refine. Nat. Gasoline Manuf.* **15** (1936) no. 6, 205.
56. H. F. Smyth Jr., *Am. Ind. Hyg. Ass. Q.* **17** (1956) 129.
57. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 2nd rev. ed., **vol. 2,** Wiley Interscience, New York 1962.
58. Di Vincenzo et al., *Toxicol. Appl. Pharmacol.* **36** (1976) 511.
59. K. Saida et al., *J. Neuropath. Exp. Neurol.* **35** (1976) 207.
60. S. N. Ducket, *Experientia* **30** (1974) 1283.
61. M. T. Okawa, *U.S. Nat. Techn. Inform. Serv. P. B. Rep.* 1973, no. 229 166/46 A, Avail NTIS, Gov. Rep. Announce (US), 74 (10) 51 (1974).
62. E. T. Kimura, M. E. Donn, W. P. Dodge, *Toxicol. Appl. Pharmacol.* **19** (1971) 699.
63. N. J. Sax: *Dangerous Properties of Industrial Materials,* 3rd ed., Van Nostrand Reinhold, New York 1969, p. 495.

1,4-Butenediol → Butanediols, Butenediol, and Butynediol
Butene Polymers → Polyolefins

Butenes

FRITZ OBENAUS, Hüls AG, Marl, Federal Republic of Germany

WILHELM DROSTE, Hüls AG, Marl, Federal Republic of Germany

JOACHIM NEUMEISTER, Hüls AG, Marl, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Resources and Raw Materials 4
5. Upgrading of Butenes 6
6. Quality Specifications 8
7. Analysis 8
8. Storage and Transportation 9
9. Uses and Economic Data 9
10. Toxicology 12
11. References 12

1. Introduction

Butenes are unsaturated olefinic hydrocarbons, C_4H_8, M_r 56.1080. There are four isomers:

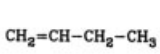

1 1–Butene

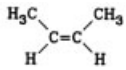

2 *cis*–2–Butene

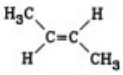

3 *trans*–2–Butene

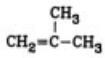

4 2–Methylpropene, Isobutene

"Butylenes", the older name for "butenes", is still used today; **4** is frequently referred to as "isobutylene". The designation "*n*-butenes" refers to mixtures of **1, 2,** and **3.**

All the butenes, which do not exist as natural products, have been known for more than 100 years, but remained in very limited use and importance. Scarce availability has been for long the reason. However with the growth of cracking processes in crude oil refining and for ethylene production, butenes have been obtained as coproducts in huge quantities. Since then the complicated nature of the raw C_4 streams, containing besides the four butene isomers also the two butane isomers and multiple unsaturated C_4 hydrocarbons, has been the main barrier for specific chemical use of the butenes.

Until the mid 1980s availability of butene mixtures exceeds by far worldwide the production of the single butene isomers and their derivatives. But during the 1980s the sharp price increase of hydrocarbons accelerated the development of economic separation processes, which opened for the butenes access to appropriate upgraded use.

2. Physical Properties

Butenes are colorless, flammable gases at room temperature and atmospheric pressure. They are completely miscible with alcohols, ethers, and hydrocarbons [1]. Butenes are only slightly water soluble and water is only slightly butene soluble [2]. Important physical properties are summarized in Table 1. Other thermodynamic properties and transport regulations are reported in [3–5].

3. Chemical Properties

Butenes behave as typical olefins. The main reactions are acid-catalyzed addition reactions, isomerization, and polymerization. The four isomers, including two isomers with nonterminal double bonds and one branched-chain olefin, show differences in their chemical behavior: while the 2-butenes as the lowest olefins with nonterminal double bonds generally show a minor chemical activity, isobutene, the lowest branched-chain olefin, exhibits higher reactivity, especially in addition and polymerization reactions. This difference in reactivity permits, e.g.,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_483

Table 1. Physical properties of butenes

			1-Butene [*106-98-9*]	*cis*-2-Butene [*590-18-1*]	*trans*-2-Butene [*624-64-6*]	Isobutene [*115-11-7*]
Melting point (101.3 kPa)		°C	− 185.35	− 138.92	− 105.53	− 140.34
Boiling point (101.3 kPa)		°C	− 6.25	+ 3.72	+ 0.88	− 6.90
Critical temperature		°C	146.45	162.43	155.48	144.75
Critical pressure		MPa	4.02	4.20	4.10	4.00
Critical density		g/cm^3	0.240	0.234	0.238	0.239
Density of liquid at 25 °C		g/cm^3	0.5888	0.6154	0.5984	0.5879
Density of gas at 0 °C, 101.3 kPa		kg/m^3	2.582	2.591	2.591	2.582
Vapor pressure at						
	0 °C	kPa	127.3	87.9	98.4	130.3
	20 °C	kPa	252.9	181.2	199.7	257.0
	40 °C	kPa	457.4	337.5	367.8	462.8
	60 °C	kPa	766.7	579.6	626.0	774.3
	80 °C	kPa	1207.8	931.3	999.2	1219.0
	100 °C	kPa	1807.1	1416.4	1512.1	1824.7
Vapor pressure (Antoine equation constants) ∗						
	temp. range	°C	− 82 to + 13	− 73.4 to + 23	− 76 to + 20	− 82 to +12
	A		6.84290	6.86926	6.86952	6.84134
	B		926.10	960.10	960.80	923.20
	C		240.00	237.00	240.00	240.00
Heat of vaporization at saturation pressure						
	at 25 °C	J/g	358.7	394.5	380.3	366.9
	at *bp*	J/g	390.6	416.2	405.6	394.2
Isobaric specific heat at 25 °C						
	gas in ideal state	J kg^{-1} K^{-1}	1528	1408	1566	1589
	liquid at 101.3 kPa	J kg^{-1} K^{-1}	2299	2250	2276	2336
Enthalpy of formation ΔH_f^0 at 25 °C, 101.3 kPa		kJ/mol	− 0.04	− 6.91	− 11.1	− 16.9
Free enthalpy of formation ΔG_f^0 at 25 °C, 101.3 kPa		kJ/mol	71.38	65.98	63.10	58.11
Heat of combustion [to H_2O (liquid) and CO_2 (gas)] at constant pressure and 25 °C		kJ/mol	− 2719.1	− 2712.3	− 2708.1	− 2702.3
Ignition temperature (DIN 51794)		°C	384	325	325	465
Flammability limits in air at 20 °C, 101.3 kPa						
	lower	vol %	1.6	1.7	1.7	1.8
	higher	vol %	9.3	9.7	9.7	8.8

∗ $\log_{10} p = A - B/(t+C)$, where p is in mm Hg and t in °C; to convert mm Hg to kPa, divide by 7.528.

the separation of the other isomers through selective reaction of isobutene. The differences in activity (isobutene ≫ 1-butene > 2-butenes) are due to different electron densities, polarities, and steric effects.

Hydration. Acid-catalyzed hydration of butenes is one of the commercially most important processes. Both gas- and liquid-phase processes are used. For kinetic and mechanistic studies see references [6], [7] (→ Butanols). Isobutene yields *tert*-butyl alcohol (TBA). Sulfuric acid (45 wt %) is commonly used as a protonating agent [8]. This process has also been carried out in the presence of sulfonated styrene – divinylbenzene ion-exchangers [9]. The latter liquid-phase process has fewer side reactions, such as formation of di- and triisobutene, less equipment corrosion, and fewer environmental problems. Hydration is commercially used for separating isobutene from mixed butenes.

Hydration of isobutene-free *n*-butenes to form *sec*-butyl alcohol (SBA) is catalyzed under more severe conditions usually by sulfuric acid [10] or more recently by acid ion-exchange resins [11].

Etherification. The acid-catalyzed addition of alcohols to butenes yields alkyl butyl ethers (→ Ethers, Aliphatic). Reaction of isobutene with methanol, yielding methyl *tert*-butyl ether

(MTB or MTBE), is of technical importance. Liquid phase and ion exchange resins as catalysts are commonly used [12]. The etherification of *n*-butenes requires more severe conditions and is of no commercial importance.

Halogenation. 1-Butene reacts with halogens at room temperature to give 1,2-dihalogenbutane and reaction with 2-butenes gives 2,3-dihalogenbutane. Allylic substitution occurs at higher temperatures (> 200 °C). This reaction is of no practical importance. Isobutene readily reacts with chlorine at low temperatures forming methallyl chloride $ClCH_2C(CH_3)=CH_2$ [13].

Hydroformylation. Hydroformylation of butenes in the presence of cobalt or rhodium catalysts gives valeric aldehydes and the corresponding amyl alcohols. *n*-Pentanol and 2-methylbutanol are formed from *n*-butenes, while isobutene gives only 3-methylbutanol [14].

Hydrocarboxylation. The catalyzed reaction of butenes with carbon monoxide and water yields carboxylic acids. Particularly isobutene readily forms pivalic acid $(CH_3)_3CCOOH$ in the presence of strong acid (Koch reaction [15]).

Isomerization. With increasing temperatures the following reactions take place:

cis-2-butene $\rightleftharpoons$ *trans*-2-butene (1)

1-butene $\rightleftharpoons$ 2-butenes (2)

n-butenes $\rightleftharpoons$ isobutene (3)

While *cis/trans*-isomerization (Eq. 1) occurs at ambient temperatures in the presence of catalysts, carbon rearrangement (Eq. 3) requires temperatures of about 450 °C. A large number of catalysts, e.g., Lewis acids, Brønsted acids, metal oxides, and zeolites, are effective. A review of catalysts and equilibrium constants as a function of temperature is given in [16].

Polymerization, Oligomerization. The two types of reactions depend on the catalyst system used:

a) Coordination catalysts (Ziegler-Natta)
b) Liquid or solid Brønsted and Lewis acids

For a review see [17].

Only 1-butene yields high-molecular material such as isotactic or atactic poly-1-butene using Ziegler-Natta catalysts. Isotactic polymers [18] and copolymers with ethylene or propene are produced industrially.

Acidic substances induce poly- or oligomerization by forming carbenium ions [19]. Molecular mass is dependent on many reaction variables, especially temperature: increasing temperature generally results in lower molecular masses.

Polymerization of highly pure isobutene to polyisobutene occurs in an inert solvent at temperatures between − 10 and − 100 °C [20]. Copolymerization with 1 – 3 % isoprene under similar conditions yields butyl rubber. Polymerization of isobutene in butene mixtures in the presence of $AlCl_3$ between − 10 and + 80 °C yields polybutenes with a few percent *n*-butene and isobutane as chain terminators. The molecular masses lie between 300 and 2500. About 80 –95 % of the isobutene is usually converted to polymer [21].

Oligomerization of isobutene to dimers and trimers is done by extracting isobutene from mixed butenes using 65 – 70 wt % sulfuric acid and subsequent heating to 100 °C (cold acid process). An alternative process uses acid ion-exchange resins instead of sulfuric acid (Bayer). The resulting isobutene oligomers, mainly 2,2,4-trimethylpentenes, contain isomeric *n*-butene cooligomers, which increase with higher isobutene conversions. A good part of the accompanying 1-butene is isomerized to 2-butenes [22].

n-Butenes are commercially oligomerized by various homogeneous or heterogeneous catalysts. The Dimersol X Process (IFP), which is commercially applied in Japan (Nissan Oil), uses homogeneous Ziegler-Natta catalysts and forms, in addition to higher oligomers mainly 3-methylheptenes [23]. The Hüls/UOP Octol Process is an example of a heterogeneous catalyzed process which oligomerizes *n*-butenes by different mechanisms depending on the catalyst. The dimers are either mainly 3,4-dimethylhexenes or 3-methylheptenes containing considerable amounts of *n*-octenes and 3,4-dimethylhexenes. A commercial plant is on stream in the Federal Republic of Germany (Hüls).

Highly branched oligomers for fuel are produced by cooligomerizing complete butene

streams alone or in the presence of propene to polymer gasoline (Cat Poly; UOP Catalytic Condensation Process) [24]. The process uses phosphoric acid on kieselguhr.

Dehydrogenation. Dehydrogenation of *n*-butenes is important in the USA for the production of 1,3-butadiene. The reaction is carried out discontinuously at 600 – 700 °C over a chromium/alumina catalyst (Catadiene process, Houdry [25]). Catalytic oxidative dehydrogenation processes in the presence of air and steam running continuously at lower temperatures were later developed (Oxo-D process, Petrotex [26]).

Oxidation. Air oxidation of isobutene [27] or *tert*-butyl alcohol [28] over complex mixtures of transition metal oxides gives methacrolein, which is oxidized in a second reactor to methacrylic acid. The selectivities are between 70 – 80 % at 80 % conversion. Two plants in Japan are in operation (Asahi Glass, Mitsubishi Rayon).

Oxidation of isobutene with ammonia and oxygen to produce methacrylonitrile (ammoxidation) is an extension of SOHIO's propene-based acrylonitrile technology.

n-Butenes and mixed butenes can be oxidized over V_2O_5 /P_2O_5-catalysts to maleic anhydride with selectivities of about 50 – 60 mol %. The temperatures are between 350 and 450 °C. Any isobutene present in the feed is rapidly oxidized to CO, CO_2, and water.

Three plants, one in the Federal Republic of Germany (Bayer [29]) and the others in Japan (Mitsubishi Chemical [30], Nichiyu Chemical) are presently on stream.

Catalytic gas-phase oxidation of *n*-butenes in the presence of steam yields acetic acid [31]. Higher yields can be achieved by liquid-phase oxidation of *sec*-butyl acetate, which is formed in a separate reaction of *n*-butenes with acetic acid [32].

Disproportionation (Metathesis). The Phillips Triolefin Process, which was developed for the production of ethylene and *n*-butenes from propylene, can be reversed to react 2-butenes with ethylene to yield propene. The most common catalyst systems are tungsten, molybdenum, or rhenium complexes; for a review see [33].

Prins Reaction. The acid-catalyzed reaction of isobutene with formaldehyde is the most important synthetic route for isoprene production. Several plants using two-step processes (IFP, Bayer, Kuraray) via intermediate 4,4-dimethylmetadioxane formation are operating in the Former States of USSR, Japan, and Romania [34].

Alkylation. Friedel-Crafts alkylation of aromatics, primarily phenol, *p*-cresol, and catechol, with isobutene in the presence of acid catalysts yields *tert*-butylaromatics. The formation of the thermodynamically favored para-substituted products can be minimized by varying the reaction time, amount of catalyst, and temperature [35].

Alkylation for motor fuel involves reacting mixed butenes with excess isobutane in the liquid phase over sulfuric or hydrofluoric acid catalyst [36]. The main product, aside from numerous other isomers, is 2,2,4-trimethylpentane (isooctane).

4. Resources and Raw Materials

Butenes are rarely made purposely. Almost all butenes are mainly byproducts of gasoline and ethylene manufacture. The operation of these processes is governed by the demand for the major products, while the quantity and composition of the butene streams are left to vary.

Petroleum refining is the largest butene source in the world. Most butenes are generated in catalytic crackers, which convert vacuum gas oils into gasoline and medium distillates. Thermal refining processes (visbreaking, thermal cracking, coking) also byproduce butenes, but their volume is inconsequential compared to the quantity from cat crackers.

The butane – butene ("B – B") streams from catalytic cracking may account for 10 to 13 wt % of fresh feed depending on cracking severity and the catalyst. The butene portion may vary somewhat, but generally comprises about 50 % of the total weight. This is illustrated together with a breakdown of the butenes in Table 2.

Table 3. Estimated world primary production of butenes in 1984 (10^3 t)

Area	Source	Sum of butenes	Iso-butene	*n*-Butenes
North America	St.Cr.	1130	590	540
	Ref.	13600	4420	9180
	Dehydro	30	–	30
	TBA	250	250	–
	C_2H_4	40	–	40
	Total	15050	5260	9790
South America and Caribbean	St.Cr.	220	120	100
	Ref.	2300	710	1590
	Dehydro	60	–	60
	Total	2580	830	1750
Western Europe	St.Cr.	1750	920	830
	Ref.	2350	730	1620
	TBA	50	50	–
	Total	4150	1700	2450
Eastern Europe	St.Cr.	620	330	290
	Ref.	930	320	610
	Dehydro	650	50	600
	Total	2200	700	1500
Asia and Australia	St.Cr.	1060	550	510
	Ref.	1680	530	1150
	Dehydro	40	–	40
	C_2H_4	10	–	10
	Total	2790	1080	1710
Africa and Middle East	St.Cr.	60	30	30
	Ref.	1260	390	870
	Total	1320	420	900
World	St.Cr.	4840	2540	2300
	Ref.	22120	7100	15020
	Dehydro	780	50	730
	TBA	300	300	–
	C_2H_4	50	–	50
	Total	28090	9990	18100

Table 2. Typical composition (in wt%) of C_4 fractions from fluid catalytic cracking (FCC) with zeolite catalyst and from naphtha steam cracking (St.Cr.) under medium severity

Component	FCC	St.Cr.
Isobutane	37	2
Isobutene	15	26
1-Butene	12	14
1,3-Butadiene	<0.5	43
n-Butane	13	6
trans-2-Butene	12	5
cis-2-Butene	11	4

Even though steam crackers account worldwide for only a small percentage of the total butene production, the products from these facilities are more important for chemical use than butenes produced in refineries. The yield of pyrolysis C_4 in olefin plants is highly dependent on the type of feedstock. Liquid feedstocks, i.e., naphtha and gas oil, are converted up to 12 wt % to pyrolysis C_4, while light hydrocar-

bon feedstocks, i.e., ethane and propane yield only roughly 2 or 4 wt %, respectively. More severe cracking conditions favor ethylene formation and lower the higher olefin and butene yields [37]. Typical composition of pyrolysis C_4 obtained by naphtha steam cracking under medium severity is shown in Table 2.

Catalytic dehydrogenation of *n*-butane is only of some importance in Eastern Europe as a source in *n*-butenes, which usually serve as intermediates in the butadiene production. Formerly, *n*-butenes obtained by dehydrogenation were also a major source of butadienes in the USA, but were replaced by the increased amounts of butadiene from steam cracking.

Increasing availability of butanes as condensate from natural gas or gas from crude-oil production (field butanes) in countries far from the consumers has generated plans for new dehydrogenation plants primarily for the production and conversion of isobutene to methyl *tert*-butyl ether (MTB) [38]. Three developed catalytic processes – Houdry's Catofin [39], UOP's Oleflex [40], and Phillip's STAR [41] – are offered for commercial use.

tert-Butyl alcohol (TBA), made by Arco (formerly Oxirane) as a coproduct in its propene oxide plants in the USA and the Netherlands [42], is another source of isobutene.

Ethylene oligomerization processes for production of alpha-olefins with 6 – 20 carbon atoms yield about 10 wt % 1-butene as a byproduct which can be isolated in high purity. Three plants in the USA (Ethyl Corp., Gulf, Shell) and one in Japan (Mitsubishi Chemical) produce 1-butene via this route. The Alphabutol process of IFP for selective dimerization of ethylene to 1-butene is offered for commercial use [43].

Worldwide butene availability, especially as crude mixtures, is enormous. Estimated world primary production of butenes (1984) is given in Table 3. A breakdown into particular areas and sources shows remarkable differences. More than half of the total is generated in North America, but the production of the more valuable butenes from steam cracking is highest in Western Europe.

5. Upgrading of Butenes

Upgrading production of pure butene isomers, their derivatives, or a concentrate of the *n*-butenes, with the above mentioned exceptions, always requires separation from the C_4 mixtures. The pyrolysis C_4 mixture from steam cracking after removal of butadiene, which is recovered in almost all cases by extractive distillation, forms a butene-rich stream known as raffinate-I (Table 4). This is the preferred petrochemical feedstock. Where demand exceeds supply, the more diluted "B – B streams" from cat crackers are also processed.

Table 4. Typical composition (in wt %) of the C_4 fraction from naphtha steam cracking after (1) butadiene extraction (raffinate-I), (2) conversion of isobutene to MTB ∗ (raffinate-II), (3) isobutane removal and selective hydrogenation ∗ (*n*-butene concentrate)

Component	Raff.-I	Raff.-II	*n*-Butene conc.
Isobutane	3	6	< 0.05
Isobutene	45	< 0.1	< 0.1
1-Butene	25	45	48
1,3-Butadiene	< 0.5	< 0.8	< 0.001
n-Butane	11	19	20
trans-2-Butene	9	17	18
cis-2-Butene	7	13	14

∗ Related to the Hüls process (Fig. 1)

Butenes cannot be separated by mere distillation. Their boiling points are too close together, especially those of isobutene and 1-butene, being the isomers which have established substantial outlets for high purity qualities. Extractive distillation is sometimes used to separate butenes from butanes, but cannot separate butene isomers. Molecular sieve adsorption processes offer potential for isolation of pure butenes [44], but are not yet in commercial use.

Isobutene Separation. All industrially used butene separation processes are based on the higher chemical reactivity of isobutene, the key substance in the butene mixtures. The isobutene derivatives are easily separated and can be split back during a subsequent step to produce pure isobutene or be recovered for particular use. In early stages several similar processes for isobutene separation by reactive extraction with sulfuric acid (45 to 70 wt %) were applied [45]. Isobutene can be isolated in more than 99 wt % purity as *tert*-butanol (TBA) or dimer/trimer

mixture depending on the acid concentration and type of acid regeneration. Only the latter is formed by the so-called "cold acid process", in which higher acid concentrations are added. This process is still widely used to reduce isobutene to < 0.2 wt % in the spent C_4 mixture.

The difficulty and high cost of the traditional processes to separate isobutene has been the main factor in retarding growth of butenes for specific chemical use. Within the 1980s new processes have been introduced that are superior form the view points of economics, environment and energy conservation.

The etherification of isobutene with methanol in the C_4 mixtures to form methyl *tert*-butyl ether (MTB) is the key step in these processes. This very selective reaction takes place in the liquid phase between 40 and 100 °C using a macroporous cation exchange resin as a catalyst. Hüls was the first to modify its raffinate-I based MTB production in 1978 to achieve more than 99.9 % isobutene conversion [46]. All other C_4 components in the co-produced raffinate-II remain unchanged (Table 4). The broad acceptance of MTB as a valuable octane booster opened an unlimited outlet for all surplus isobutene. This, together with the ease of complete isobutene conversion to MTB, makes the corresponding *n*-butenes available, especially as raffinate-II at low cost for specific use. Even though several processes for splitting MTB by reversing synthesis to high-purity isobutene are available [47], none are in commercial use.

Table 5. Isobutene and 1-butene composition (in wt%) in high-purity commercial products (Hüls)

Component	Isobutene	1-Butene
Isobutene	99.98	0.15
1-Butene	0.005	99.70
2-Butenes	0.01	< 0.01
Butanes	0.005	0.15
1,3-Butadiene	< 0.001	< 0.001
tert-Butanol	< 0.0005	< 0.0001
Water	< 0.003	< 0.001
Sulfur	< 0.0001	< 0.0001

Another new process introduced by Hüls in 1981 selectively hydrates isobutene to *tert*-butanol (TBA) under non-corrosive liquid-phase conditions. A cation exchanger acts as catalyst under conditions similar to the MTB process. Equilibrium and incomplete phase miscibility limits the conversion to TBA to about 90 %. Due to the different equilibriums, decomposition of TBA is more favored than that of MTB. The Hüls process decomposes in a second step well below 150 °C heterogeneously catalyzed in liquid phase TBA to high-purity isobutene (Table 5) [9], [48].

***n*-Butene Concentration.** Once isobutene content has been reduced to the raffinate-II level shown in Table 4, recovery of high-purity 1-butene, which is worldwide in growing demand, is possible by fractionation. However, a preceding hydrotreating step is necessary to convert selectively butadiene and traces of acetylenes to achieve the desired product quality (Table 5) [49].

2-Butene is the preferred feedstock for butadiene production by dehydrogenation and is therefore sometimes concentrated. It is of no commercial importance as a high-purity product. Because 1-butene reacts in most cases to the same products as obtained from 2-butene (hydration, esterification, oxidation, oligomerization) the use of 2-butene in mixed *n*-butene concentrates has proved to be a convenient feed for chemical production.

An example of *n*-butene concentrate composition is illustrated in Table 4. Higher concentration can be obtained by removing butanes through extractive distillation or the more recently commercialized molecular sieve adsorption method [50]. Lower concentrations, especially *n*-butane mixtures in the bottom of 1-butene fractionation, are also useful for producing *n*-butene derivatives.

Complementing Process Integration. Whether a butene upgrading process is economical or not depends largely on suitable process combinations. A rather simple combination is e.g. the production of MTB in front of an alkylation of FCC C_4's, whereby the contained isobutene is higher upgraded and alkylation of the remaining butenes becomes independant of the availability of additional isobutane. A steady outlet of the products to the gasoline pool is in such cases generally no problem. More sophisticated systems are required to properly upgrade the more highly valued raffinate-I butenes. Continuous processing of the entire C_4 stream requires that varying demand for different isomers or derivatives be compensated by

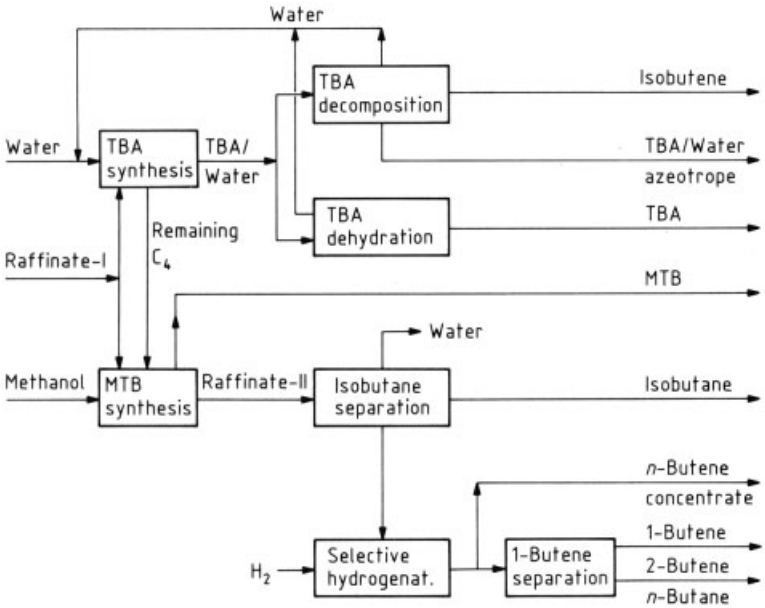

Figure 1. Hüls integrated raffinate-I processing

flexible production schedules that allow production of alternate products. This leads to highly integrated process systems. Figure 1 shows the Hüls scheme for raffinate-I processing.

Raffinate-I is fed in a variable ratio to the synthesis of TBA and MTB. Primary product of the TBA synthesis is a TBA/water mixture that is either decomposed to give high-purity isobutene or purified to give TBA/water azeotrope and dehydrated TBA. The raffinate-I not required for TBA/isobutene is conjointly used with remaining C_4 compounds from TBA synthesis in MTB production. Whereas TBA is more easily converted to pure isobutene, MTB is the more economic gasoline improver and according to its thermodynamically favored synthesis isobutene conversion is accomplished for the total stream, resulting in virtually isobutene-free raffinate-II (Table 4).

Isobutane is separated together with water from the raffinate-II by fractionation. Selective hydrogenation of the bottom product leads to *n*-butene concentrate with a composition shown in Table 4. Fractionation of the *n*-butene concentrate yields polymerization grade 1-butene. The 2-butene/*n*-butane mixture remaining after 1-butene separation is like any surplus of *n*-butene concentrate suitable for the production of di-*n*-butenes through oligomerization.

6. Quality Specifications

Due to different production processes and applications, no standard quality specifications for the butenes are available. The two isomers which have become commodities as high purity products show in general a purity greater than 99 wt %. Table 5 illustrates the typical isobutene and 1-butene composition as obtained through the Hüls processes.

7. Analysis

Analysis of pure butenes and butene mixtures containing *n*-butane, isobutane, 1,3-butadiene, small amounts of C_3- and C_4 acetylenes, and diolefins are generally performed by gas – liquid chromatography. A routine GLC-analysis using a packed column is described in the ASTM Method D-1717.

An Al_2O_3 capillary column (Chrompack) permits the simultaneous determination of all C_1- to C_5-components combined with improved detection (usually $\leqq 5$ ppm).

The water content of butenes often plays an important role in further processing. It can be determined either by Karl Fischer titration [51] or more precisely by a commercial on-line hygrometer (e.g., Panametrics).

All other impurities are determined by general standard methods.

8. Storage and Transportation

Butenes are usually stored and transported as pressurized liquids.

Most countries have specific legislation pertaining to the erection and operation of tank farms for LPG as butenes.

The following regulations pertaining to the transport of dangerous goods must be observed while transporting butenes:

GGVE/GGVS:

class 2, number 3 b

RID/ADR:

class 2, number 3 b

ADNR:

class 2, number 6, category F

IMDG-Code/GGVSee:

class 2, UN-No. 1012 (*n*-butenes)
class 2, UN-No. 1055 (isobutene)

IATA-DGR/ICAO-Code:

class 2 (3) UN-No. 1012 (*n*-butenes)
class 2 (3) UN-No. 1055 (isobutene)

The transport containers for butenes (steel cylinders, steel drums, road tankers, pressure gas tank wagons and tank containers) must conform to these regulations in respect to construction materials, design, fittings, and marking. They must be approved for use at not less than 1 MPa (10 bar) test pressure.

9. Uses and Economic Data

The estimated world butene consumptions in 1984 are shown according to regions in Table 6. More than half of the butenes produced worldwide are utilized as alkylate and polymer gasoline. One third is used without any conversion as fuel, mainly as fuel gas or blendstock for gasoline. Only about 15 % of the butenes are converted to specific chemicals. Even the greater part of the isobutene transformed to pure MTB is ultimately burned as a gasoline component.

Table 6 shows some remarkable regional differences. In North America, Western Europe, Eastern Europe, and Japan, most butenes are converted into higher value gasoline components and chemical products, whereas in industrially less developed regions, butenes are usually directly used without further conversion for their calorific value.

Alkylate Gasoline. In North America nearly 80 % of butenes are converted into alkylate gasoline, whereas alkylate gasoline plays a lesser role in other regions. However, since more than half of all butenes are produced in North America, almost 50 % of the butenes used worldwide are converted into alkylate gasoline, the largest outlet for butenes.

Polymer Gasoline. The oligomerization of butenes, in most cases mixed with propene, gives highly branched C_6-C_{12} -olefins [24]. It is the oldest method for reconverting the low boilers obtained from catalytic cracking into higher molecular mass gasoline components. This was an important process, particularly in the 1940s, for the production of aviation fuel. Today, it has been largely replaced by alkylation, which gives higher octane rating products.

Straight Fuel. Butenes are also used directly for such diverse applications as a blendstock in gasoline, LPG, and as plant fuel. The least demand exists in North America, where paraffinic light hydrocarbons are readily available from natural gas liquids and butenes are more highly valued as alkylation feedstock. Consumption is higher in areas as Europe, where LPG production is more or less restricted to refineries.

Methyl tert-Butyl Ether (MTB). MTB is almost exclusively admixed to carburetor fuels as an octane booster. It showed the highest growth rate of all butene derivatives during the last decade [52]. Since the isobutene in raffinate-I is already largely utilized, isobutene from catalytic cracking or other sources – upstream of alkylation – is increasingly used as feedstock for the production of MTB [53].

Table 6. Estimated world butene consumptions in 1984 (10^3 t)

Area	Product	Sum of butenes	Iso-butene	*n*-Bu-tenes
North America	Alkylate gasoline	11860	3690	8170
	Polymer gasoline	390	120	270
	Straight fuel uses	1220	410	810
	MTB	500	500	–
	Di-/triisobutene	30	30	–
	Butyl rubber	250	250	–
	Polybutene	250	250	–
	1,3-Butadiene	60	–	60
	1-Butene in polyolef.	180	–	180
	2-Butanol (SBA)	250	–	250
	Special chemicals	210	100	110
	Total	15200	5350	9850
South America and Caribbean	Alkylate gasoline	260	90	170
	Polymer gasoline	30	10	20
	Straight fuel uses	2100	665	1435
	MTB	20	20	–
	Polybutene	5	5	–
	1,3-Butadiene	60	–	60
	1-Butene in polyolef.	10	–	10
	2-Butanol (SBA)	15	–	15
	Special chemicals	30	10	20
	Total	2530	800	1730
Western Europe	Alkylate gasoline	900	280	620
	Polymer gasoline	180	60	120
	Straight fuel uses	1710	380	1330
	MTB	370	370	–
	TBA	10	10	–
	Di-/triisobutene	120	120	–
	Butyl rubber	190	190	–
	Polybutene	130	130	–
	1-Butene in polyolef.	10	–	50
	2-Butanol (SBA)	200	–	200
	Special chemicals	180	100	80
	Total	4040	1640	2400
Eastern Europe	Alkylate gasoline	280	100	180
	Straight fuel uses	910	190	720
	MTB	90	90	–
	Butyl rubber	60	60	–
	Isoprene	250	250	–
	1,3-Butadiene	600	–	600
	Special chemicals	10	10	–
	Total	2200	700	1500
Asia and Australia	Alkylate gasoline	360	140	220
	Polymer gasoline	90	30	60
	Straight fuel uses	1960	730	1230
	MTB	10	10	–
	TBA	40	40	–
	Di-/triisobutene	20	20	–
	Butyl rubber	50	50	–
	Polybutene	30	30	–
	Isoprene	20	20	–
	1,3-Butadiene	40	–	40
	1-Butene in polyolef.	30	–	30

Table 6. (Continued)

Area	Product	Sum of butenes	Iso-butene	*n*-Bu-tenes
	2-Butanol (SBA)	100	–	100
	Special chemicals	40	10	30
	Total	2790	1080	1710
Africa and Middle East	Alkylate gasoline	130	35	95
	Straight fuel uses	1190	385	805
	1-Butene in polyolef.	10	–	10
	Total	1330	420	910
World	Alkylate gasoline	13790	4335	9455
	Polymer gasoline	690	220	470
	Straight fuel uses	9090	2760	6330
	MTB	990	990	–
	TBA	50	50	–
	Di-/triisobutene	170	170	–
	Butyl rubber	550	550	–
	Polybutene	415	415	–
	Isoprene	270	270	–
	1,3-Butadiene	760	–	760
	1-Butene in polyolef.	280	–	280
	2-Butanol (SBA)	565	–	565
	Special chemicals	470	230	240
	Total	28090	9990	18100

tert-Butanol (TBA). Large quantities of TBA from isobutane are obtained as a byproduct of propene oxide production by ARCO [42]. It is predominantly used in its impure form as blending component for gasoline [54] or to produce pure isobutene (see Table 3).

Smaller quantities of pure TBA are obtained by direct hydration of isobutene-containing C_4 cuts. This process, only is in use in Japan and the Federal Republic of Germany, serves also for the isolation of pure isobutene via TBA. That portion is not included in the figures in Table 6. This pure TBA is in Japan predominantly used for the production of methyl methacrylate [28]. Other countries use TBA as a solvent for the production of *tert*-butyl peroxides or as a stabilizer for chlorinated hydrocarbons among others.

Di- and Triisobutenes. The decreased use of sulfuric acid extraction for isobutene separation has resulted in di- and triisobutene production for chemical use only.

Diisobutene is mainly used for the production of isononanol and octylphenol, which are required as intermediates for plasticizers and detergents, respectively.

Butyl Rubber. More than half of the 1 million tons of high-purity isobutene produced in 1984 was converted to butyl rubber. Butyl rubber, which is made by copolymerizing high-purity isobutene with small amounts of isoprene, is used mainly as a tire inner liner because of its excellent air impermeability.

Also included are the smaller amounts of so-called polyisobutene, which does not contain any comonomers and is used in films, among others [20].

Polybutene. Polybutene is a polymerization product of isobutene obtained directly from raffinate-I or the catalytic cracker C_4 cut and contains therefore about 2 % *n*-butene. The largest portion is in the 1000 molecular mass range and is used as a precursor for lube oil additives [55]. Polymers with higher molecular masses are used as caulks and sealants.

Isoprene. Isoprene can be directly extracted from the steam cracker C_5 cut or synthesized by reacting formaldehyde with pure isobutene. Even though this process is only used in Eastern Europe, especially in the Former States of USSR, and in one plant in Japan, isoprene production via isobutene is the second largest outlet of pure isobutene.

Butadiene. Although butadiene production via *n*-butene dehydrogenation is the single largest *n*-butene consumer for chemical use, less than 15 % of worldwide butadiene production in 1984 was based on dehydrogenation of *n*-butenes. Even in the United States, where dehy-

drogenation was once dominating, it is presently only used to augment the supply when butadiene from extractive distillation of steam cracker C_4 cut is in short supply. Butadiene production by two-step dehydrogenation via *n*-butane is used in countries with an ample butane supply, such as the Former States of USSR and Mexico.

1-Butene in Polyolefins. High-purity 1-butene demands have multiplied during the 1980s and are expected to continue to grow at this rate [56]. Most is used as a comonomer for linear low-density polyethylene (LLDPE), which contains up to 10 %. However, 1-butene is also used as a comonomer for modifying high-density polyethylene (HDPE), which contains up to 4 %.

Shell in the United States uses about 20000 t of high-purity 1-butene for the production of poly-1-butene per year.

sec-Butanol (SBA). SBA made by hydration is after butadiene quantitatively the second most important product from *n*-butenes. Raffinate-II is a typical feed material. Almost all the SBA is further converted by catalytic dehydrogenation into methyl ethyl ketone (MEK), which is mainly used as a solvent and diluent in the surface-coating and plastic industries.

Special Chemicals. Butenes are used to produce various special chemicals.

Isobutene derivatives include, among others, phenols, such as 4-*tert*-butylphenol, 2,6-di-*tert*-butyl-4-methylphenol (BHT), 2- or 3-*tert*-butyl- 4-hydroxyanisole (BHA), most of which are used as stabilizers. Other isobutene derivatives worth mentioning are *tert*-butylamine (intermediate for lubricating oil additives), *tert*-butylmercaptan (gas odorant), pivalic acid, 3-methylbutyric acid (isovaleric acid), 3-methyl-1-butanol, methallyl chloride, and triisobutylaluminum.

With a few exceptions, such as *n*-butene 1,2-oxide (stabilizer for chlorinated hydrocarbons) *n*-butene derivatives are generally derived from mixtures (raffinate II or *n*-butene concentrate). These include oligomers (intermediates for plasticizers, detergents, and lube oil additives), maleic anhydride, the oxo derivatives *n*-valeric acid, and 2-methylbutyric acid, and the corresponding alcohols 1-pentanol and 2-methyl-1-butanol, which are primarily used as solvents.

10. Toxicology

A specific toxic action of the butenes is not known. They have, however, a narcotic action at higher concentrations causing vomiting, giddiness, and intoxication. Butene gas in high concentrations, as built up by rapid evaporation, displaces atmospheric oxygen so that there is a danger of suffocation. The symptoms, however, disappear rapidly in fresh air.

Skin contact with the liquid causes frost bite due to the high negative heat of vaporization. Toxicological data are available only for isobutene.

Acute toxicity of isobutene:

LC_{50}, inhalation, rats: 620 g/m^3 4 h
LC_{50}, inhalation, mice: 415 g/m^3 2 h [57]

Chronic toxicity of isobutene:
Isobutene shows little chronic toxicity because rats inhaling high doses (up to 8000 ppm) for 90 days show no evidence of any toxicological effects. Evidence of mutagenic or cell transforming activity (Ames Test, Mouse Lymphoma Assay, Cell Transformation Study) is also absent [58].

11. References

1. K. Smeykal, H. Lütgert, *Chem. Tech. Leipzig* **14** (1962) 202 – 207; *Chem. Abstr.* **57** (1962) 10103 h.
2. W. F. Hoot, A. Azarnoosh, J. J. McKetta, *Pet. Refin.* **36** (1957) no. 5, 255 – 256.
3. ASTM Committee D-2: Physical Constants of Hydrocarbons C_1 to C_{10}, ASTM Data Series Publication DS 4 A, American Society for Testing and Materials, Philadelphia 1971.
4. Technical Data Handbook-Petroleum Refining, American Petroleum Institute, New York 1970.
5. R. W. Gallant: *Physical Properties of Hydrocarbons,* **vol. 1,** Gulf Publishing Comp., Houston 1968, pp. 11, 157.
6. F. Axel, H. J. Warnecke, H. Langemann, *Monatsh. Chem.* **113** (1982) 541 – 546; *Chem. Abstr.* **97** (1982) 91333 c.
7. M. Katsuno, *Kogyo Kagaku Zasshi* **44** (1941) 275 – 282; *Chem. Abstr.* **44** (1950) 7752 b.
8. H. Kröper, K. Schlömer, H. M. Weitz, *Hydrocarbon Process.* **48** (1969) no. 9, 195 – 198.

9. *Chem. Eng. N.Y.* **90** (1983) no. 25, 60.
10. *Hydrocarbon Process.* **54** (1975) no. 11, 118.
11. Deutsche Texaco AG, EP 0051164, 1982 (W. Neier, W. Webers, R. Ruckhaber, G. Osterburg, W. Ostwald).
12. F. Obenaus, W. Droste, *Erdöl Kohle Erdgas Petrochem.* **33** (1980) 271 – 275.
13. W. Reeve, D. H. Chambers, C. S. Prickett, *J. Am. Chem. Soc.* **74** (1952) 5369.
14. B. Cornils: "Hydroformylation. Oxo Synthesis, Roelen Reaction" in J. Falbe (ed.): *New Synthesis with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980, p. 149.
15. H. Bahrmann: "Koch Reactions" in J. Falbe (ed.): *New Synthesis with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980, p. 372.
16. V. R. Choudhary, *Chem. Ind. Dev.* **8** (1974) no. 7, 32 – 41; *Chem. Abstr.* **82** (1975) 5131 w.
17. D. Stadermann, H. Hartung, G. Heublein, (Part I) *Chem. Tech. Leipzig* **35** (1983) no. 4, 169 – 174; *Chem. Abstr.* **98** (1983) 180218 g; (Part II) *Chem. Tech. Leipzig* **35** (1983) no. 6, 290 – 296; *Chem. Abstr.* **99** (1983) 38819 f.
18. P. W. de Leeuw, C. R. Lindegreen, R. F. Schimber, *Chem. Eng. Prog.* **76** (1980) no. 1, 57 – 63.
19. A. Onopchenko, B. L. Cupples, A. N. Kresge, *Ind. Eng. Chem. Prod. Res. Dev.* **22** (1983) no. 2, 182 – 191.
20. H. Güterbock: *Polyisobutylen und Isobutylen-Mischpolymerisate,* Springer Verlag, Berlin 1959, p. 62.
21. D. Mark, A. R. Orr, *Pet. Refin.* **35** (1956) no. 12, 185 – 186.
22. G. Scharfe, *Hydrocarbon Process.* **52** (1973) no. 4, 171 – 173.
23. J. F. Boucher, G. Follain, D. Fulop, J. Gaillard, *Oil Gas J.* **80** (1982) no. 13, 84 – 86.
24. P. C. Welnert, P. Egloff, *Pet. Process.* **3** (1948) 585 – 586, 589 – 590, 592 – 593.
25. R. G. Craig, J. M. Duffalo, *Chem. Eng. Prog.* **75** (1979) no. 2, 62 – 65.
26. L. M. Welch, R. J. Croce, H. F. Christmann, *Hydrocarbon Process.* **57** (1978) no. 11, 131 – 136.
27. D. J. Hucknell: *Selective Oxidation of Hydrocarbons,* Academic Press, London-New York 1974, pp. 104 – 111.
28. T. Hasuike, H. Matsuzawa, *Hydrocarbon Process.* **58** (1979) no. 2, 105 – 107.
29. G. Lenz, *Chem. Anlagen + Verfahren* **76** (1976) no. 7, 27 – 29; *Chem. Abstr.* **86** (1977) 5860 r.
30. *Eur. Chem. News* **25** (1974) no. 630, 30.
31. Chemische Werke Hüls, DE 1279011, 1968 (R. Brockhaus) (= US 3431297).
32. Chemische Werke Hüls, DE 3019932, 1981 (R. Brockhaus, H. Rademacher, B. Scholz)
33. R. Streck, *Chem. Ztg.* **10** (1975) 397 – 413; *Chem. Abstr.* **84** (1976) 73132 p.
34. M. Hellin, M. Davidson, *Inf. Chim.* **206** (1980) 163 – 184; *Chem. Abstr.* **94** (1981) 139101 t.
35. S. H. Patinkin, B. S. Friedmann: "Alkylation of Aromatics with Alkenes and Alkanes" in G. A. Olah (ed.): *Friedel-Crafts and Related Reactions,* vol. II, Interscience Publishers, Inc., New York 1964, Part 1, p. 1.
36. J. Weitkamp, S. Maixner, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) 523 – 529.
37. B. Schleppinghoff, *Erdöl Kohle Erdgas Petrochem.* **27** (1974) 240 – 245.
38. *Eur. Chem. News* **42** (1984) no. 1128, 18.
39. R. G. Craig, E. A. White, *Hydrocarbon Process.* **59** (1980) no. 12, 111 – 113.
40. V. V. Vera, F. Imai, *Hydrocarbon Process.* **61** (1982) no. 4, 171 – 174.
41. F. M. Brinkmeyer, D. F. Rohr, M. E. Olbrich, L. E. Drehmann, *Oil Gas J.* **81** (1983) no. 13, 75 – 78.
42. *Eur. Chem. News* **37** (1981) no. 1004, 23.
43. D. Commereuc, Y. Chauvin, J. Gaillard, J. Léonard et al., *Hydrocarbon Process.* **63** (1984) no. 11, 118 – 120.
44. A. J. de Rosset, J. W. Priegnitz, D. J. Korous, D. B. Broughton, Presentation to the ACS 175th National Meeting at Anaheim, March 12 – 17, 1978. M. S. Adler, D. R. Johnson, Presentation to the AIChE 85th National Meeting at Philadelphia, June 4 – 8, 1978.
45. Chemische Werke Hüls, DE 2362115, 1975 (W. Zerrweck) (= US 3979474).
46. *Oil Gas J.* **77** (1979) no. 1, 76 – 77.
47. V. Fattore, M. M. Mauri, G. Oriani, G. Paret, *Hydrocarbon Process.* **60** (1981) no. 8, 101 – 106.
48. Chemische Werke Hüls, EP-A 0082937, 1983 (F. Obenaus, B. Greving, H. Balke, B. Scholz) (= US 4423271)
49. Chemische Werke Hüls, DE-OS 3143647, 1983 (F. Obenaus, F. Nierlich, O. Reitemeyer, B. Scholz) (= US 4517395).
50. *Oil Gas J.* **82** (1984) no. 31, 109.
51. E. Scholz, *Fresenius Z. Anal. Chem.* **303** (1980) 203 – 207.
52. I. Steel, Unsaturated C4's – A West European Producer's View, Presentation to the ECMRA Meeting at Oslo, Oct. 11 – 13, 1982.

53. R. N. Davis, R. S. Andre, *Proc. Refin. Dep. Am. Pet. Inst.* **59** (1980) 59 – 64; *Chem. Abstr.* **93** (1980) 242.2345.
54. E. G. Guetens, Jr., J. M. DeJovine, G. J. Yogis, *Hydrocarbon Process.* **61** (1982) no. 5, 113 – 117.
55. Chem. Systems, Synthetic Lubricants, Report No. 77 – 6 (1978) 173.
56. Buten-1 LLDPE, *Chem. Week* **129** (1981) Dec. 16, 30 – 34.
57. B. B. Shugaer, *Arch. Environ. Health* **18** (1969) no. 6, 878 – 882; *Chem. Abstr.* **71** (1970) 47775 n.
58. Hazleton Laboratories Europe Ltd., Report no.2916 – 13/11 (November 1982); unpublished.

Butter → Milk and Dairy Products
Butyl Rubber → Rubber, 3. Synthetic
1,4-Butynediol → Butanediols, Butenediol, and Butynediol
Butyraldehydes → Butanals
Butyric Acid → Carboxylic Acids, Aliphatic

Butyrolactone

Wolfgang Schwarz, BASF AG, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 7)

Jürgen Schossig, BASF AG, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 7)

Roland Rossbacher, BASF AG, Ludwigshafen, Federal Republic of Germany (Chap. 8)

Hartmut Höke, Weinheim, Federal Republic of Germany (Chap. 8)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 2
5. Quality Specifications and Analysis . . . 4
6. Storage and Transportation 4
7. Uses 4
8. Toxicology and Occupational Health . . 4
9. References 5

1. Introduction

γ-Butyrolactone or γ-hydroxybutyric acid lactone [*96-48-0*], $C_4H_6O_2$, M_r 86.09, became industrially available in the 1940s as a result of the work of W. Reppe and his colleagues at BASF.

To date, the Reppe process (see Chap. 4) has remained the most important process for the production of butyrolactone. Butyrolactone is important as an intermediate in the manufacture of pyrrolidone derivatives and as a solvent for polymers and agrochemicals.

Methods of synthesis of historical and preparative interest are summarized in *Beilstein* [1] and in *Houben-Weyl* [2]. In the mid-1960s butyrolactone was discovered in nature, as a normal metabolic product in animals [3].

2. Physical Properties

Butyrolactone, *mp* −43.5 °C, *bp* 204 – 206 °C (at 101.3 kPa), is a colorless, slightly hygroscopic liquid having a faint odor. Butyrolactone is miscible in all proportions with water, alcohols, esters, ethers, ketones, and aromatic hydrocarbons. It has limited miscibility with linear and cyclic aliphatic hydrocarbons. It is an excellent solvent for numerous polymers and a selective solvent for lower hydrocarbons. Some physical properties of γ-butyrolactone are listed below.

Density	
at 0 °C	1.15 g/cm^3
at 20 °C	1.13 g/cm^3
at 40 °C	1.11 g/cm^3
Critical temperature $T_{\mathrm{crit.}}$	436.5 °C
Critical pressure $p_{\mathrm{crit.}}$	3.35 MPa
Vapor pressure	
at 20 °C	0.04 kPa
at 50 °C	0.3 kPa
at 100 °C	3.6 kPa
at 150 °C	22.5 kPa
at 205 °C	101.3 kPa
Heat of vaporization at 206 °C	535 kJ/kg
Heat of solution in water	2500 J/mol
Specific heat capacity c_p (l)	
at 25 °C	1680 J kg^{-1} K^{-1}
at 100 °C	1850 J kg^{-1} K^{-1}
at 200 °C	2200 J kg^{-1} K^{-1}
Specific heat capacity c_p (g)	
at 100 °C	1275 J kg^{-1} K^{-1}
at 200 °C	1575 J kg^{-1} K^{-1}
at 300 °C	1820 J kg^{-1} K^{-1}
Evaporation number (DIN 53170)	216
Flash point (DIN 51758)	100 – 101 °C
Ignition temperature (DIN 51794)	455 °C
Heat of combustion at constant volume or constant pressure	234 kJ/g
Explosion limits	
lower	2.7 vol %
upper	15.6 vol %
Refractive index n_D^{20}	1.4352
Viscosity	
at 20 °C	1.90 mPa · s
at 50 °C	1.25 mPa · s
at 80 °C	0.86 mPa · s
at 110 °C	0.63 mPa · s
at 140 °C	0.49 mPa · s
at 170 °C	0.40 mPa · s

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_495

Surface tension (20 °C)	44.6×10^{-5} N/cm
Dielectric constant (20 °C)	39.1
Thermal conductivity (25 – 65 °C)	0.276 J m^{-1} s^{-1} K^{-1}
Conductivity	0.7 µS/cm

3. Chemical Properties

In aqueous solution there is an equilibrium between the lactone and free hydroxybutyric acid, which lies 100 % on the lactone side at 0 °C, and about 80 % at 100 °C. In the presence of one molar equivalent of alkali, the equilibrium is displaced 100 % to the hydroxybutyric acid side; therefore, butyrolactone can be determined by titration with a base. At pH 7 the compound is extremely stable. Butyrolactone reacts with bases, hydrogen halides, and alcohols (in the presence of acids), undergoing ring cleavage and giving derivatives of γ-hydroxybutyric acid. It reacts with ammonia, amines, carbonyl compounds, and halogens in the α-position without ring cleavage. For example, butyrolactone condenses with ethyl acetate [*141-78-6*] to give α-acetobutyrolactone, an intermediate in the production of vitamin B_1:

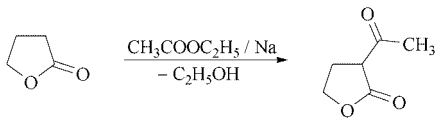

Cleavage of the lactone ring gives derivatives of γ-substituted butyric acid. For example, butyrolactone reacts with sodium sulfide to give the rubber additive thiodibutyric acid [*5152-99-8*]:

Reaction with phenols gives phenoxybutyric acids [*6303-58-8*]:

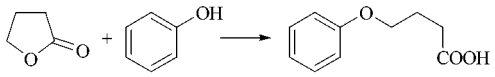

Replacement of the ring oxygen atom by nitrogen yields pyrrolidone derivatives, a reaction often used in industry. For example, butyrolactone reacts with methylamine [*74-89-5*] to give *N*-methylpyrrolidone [*872-50-4*]:

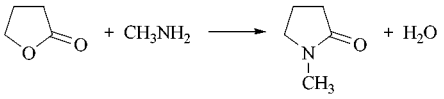

The chemistry of butyrolactone, and its use in a variety of syntheses, was investigated, above all, by W. Reppe [4] and others [5]. Reactions, generally typical of lactones, are described in *Houben-Weyl* [2].

Homopolymerization of butyrolactone can be effected only under very high pressure (2000 MPa) [6].

4. Production

Dehydrogenation of 1,4-Butanediol [*110-63-4*] (→ Butanediols, Butenediol, and Butynediol). The Reppe process for manufacturing butyrolactone involves theendothermic dehydrogenation of 1,4-butanediol in the gas phase. This process is used by BASF, ISP, and Lyondell.

$\Delta H = +61.6$ kJ/mol

Preheated 1,4-butanediol vapor is introduced into a hot stream of circulating hydrogen and passed at atmospheric pressure through a bed of copper catalyst at temperatures between 180 and 240 °C (Figure 1). The yield of butyrolactone is approximately 95 %. The reaction takes place via γ-hydroxybutyraldehyde [*25714-71-0*] [7].

The byproduct hydrogen off-gas requires only simple purification before reuse (e.g., catalytic methanization of carbon monoxide impurities). The crude butyrolactone separated from the recycle gas stream contains small amounts of byproducts, including 1,4-butanediol, butyric acid, and high boilers, from which butyrolactone is separated by distillation.

Butyrolactone itself is noncorrosive and can be handled in carbon steel apparatus. However, where parts of the synthesis or distillation vessels and pipes come into contact with hot crude product containing butyric acid, they must be made of stainless steel.

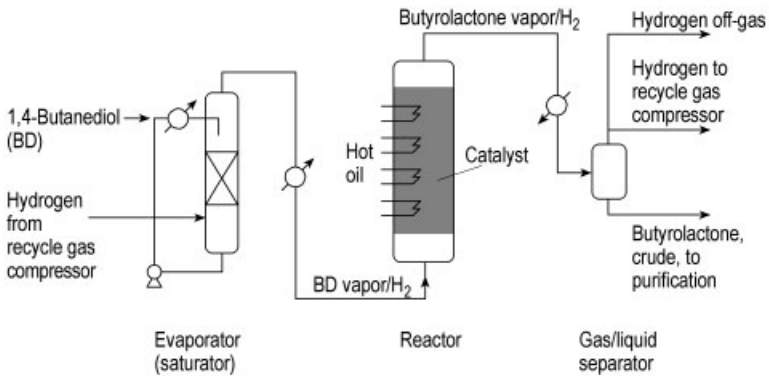

Figure 1. Production of γ-butyrolactone by dehydrogenation of 1,4-butanediol

Hydrogenation of Maleic Anhydride [*108-31-6*]. In the preparation of butyrolactone by hydrogenatingmaleic anhydride, molten maleic anhydride is fed into a preheated circulatingstream of hydrogen and passed under a pressure of 6 – 12 MPa at 160 – 280 °C over a nickel-containing catalyst [8].

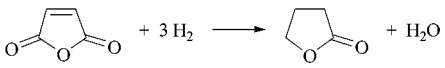

$\Delta H = -211$ kJ/mol

The reaction takes place via succinic anhydride [*108-30-5*] and can, by choice of the conditions, be continued to produce tetrahydrofuran [*109-99-9*]. The excess hydrogen is washed with water and recycled tothe synthesis. Byproducts contained in the butyrolactone are separated out of the circulating gas: propanol [*71-23-8*], butanol [*71-36-3*], propionic acid [*79-09-4*], and butyric acid [*107-92-6*]. The butyrolactone is separated from these by distillation.

Because of the acids present, both the synthesis apparatus and the distillation apparatus must be made of stainless steel. The Japanese manufacturer Mitsubishi Chemical Corporation [9] uses this process.

Hydrogenation of Maleic Esters. New processes for the production of 1,4-butanediol and tetrahydrofuran starting from maleic anhydride via dimethyl maleate have been developed in the past few years (→ Tetrahydofuran). They offer the possibility of extracting butyrolactone, which is an intermediate in these processes.

In a process developed by Kvaerner Process Technology (KPT, London) [10] dimethyl maleate [*624-48-6*] is producedin a first step from maleic anhydride and methanol with a strongly acidicion exchanger as catalyst. The resulting dimethyl maleate is hydrogenated in the gas phase on a Cu-containing catalyst at a pressure of 2 – 8 MPa at 150 – 250 °C and gives a mixture of 1,4-butanediol, tetrahydrofuran, butyrolactone, and a small amount of the intermediate dimethyl succinate [*106-65-0*].

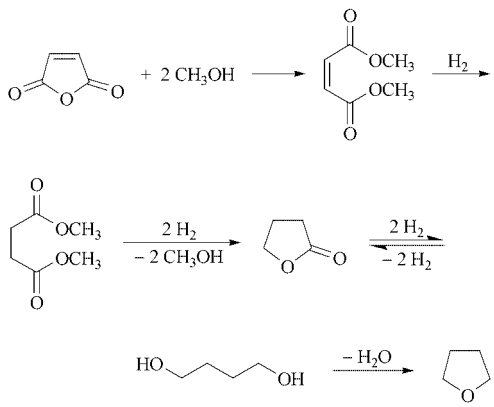

Butyrolactone and dimethyl succinate can be recovered as an azeotrope and recycled to the hydrogenation stage to obtain complete conversion to 1,4-butanediol and tetrahydrofuran. Alternatively the azeotrope can be refined by distillation to recover pure butyrolactone. The amount of butyrolactone depends on the pressure and temperature in the hydrogenation step, which influence the equilibrium between 1,4-

butanediol and butyrolactone. Under the conditions described above it may vary from 5 to 50 %.

The new process has been licensed by KPT several times. The first commercial plants using this process are expected to come on stream in 2000.

A proprietary process practised by Eurodiol, a Belgian company of the SISAS group also starts from dimethyl maleate, which is hydrogenated in the gas phase at 1 – 2 MPa to give a mixture of butyrolactone and tetrahydrofuran in variable proportions. Butyrolactone and tetrahydrofuran are recovered as pure products by distillation, while the byproduct azeotrope butyrolactone/dimethyl succinate can be recycled for full conversion to butyrolactone and tetrahydrofuran or hydrogenated in a subsequent hydrogenation step in the liquid phase to give 1,4-butanediol and additional tetrahydrofuran.

Other Processes. Processes via tetrahydrofuran [11], dihydrofuran [12], acetylene [13], [14], butynediol [15], olefins [16–18], butadiene [8], or by carbonylation [19–21] are not industrially important.

Producers. Butyrolactone is manufactured by BASF (Ludwigshafen, Germany and Geismar, USA), ISP (Calvert City and Texas City, USA), Lyondell (Channelview, USA), MCC (Mizushima, Japan) and Eurodiol (Feluy, Belgium).

5. Quality Specifications and Analysis

Standard-grade butyrolactone is about 99.7 % pure. The main impurities in the Reppe-based process are 1,4-butanediol, butyric acid, and water; in the maleic ester based processes, dimethyl succinate, 4-hydroxybutyl butyl ether, and water. Electronic-grade butyrolactone is about 99.9 % pure and is subject to stringent restrictions regarding the content of specific ions such as Fe ions, Na^+, Cl^-, NO_3^-.

Butyrolactone is usually analyzed by gas chromatography. The butyric acid content can be determined by titration with alkali. The water content is determined by coulometric titration. Cations and anions are analyzed by atomic spectroscopy and ion chromatography, respectively.

6. Storage and Transportation

Butyrolactone can be stored indefinitely. The pure product is noncorrosive and therefore can be stored and transported in carbon steel containers. Butyrolactone becomes slightly yellow on prolonged storage in carbon steel vessels without nitrogen padding. In this case stainless steel containers are recommended. The hygroscopicity of the product requires some precautions. Butyrolactone attacks concrete vigorously.

For detailed transport information (classification, regulatory data etc.), see the safety data sheet [22].

7. Uses

The main use of butyrolactone is as an intermediate in the synthesis of *N*-methylpyrrolidone (NMP) [23], pyrrolidone [24], herbicides (e.g., MCPB = γ-2-methyl-4-chlorophenoxybutyric acid) [25], growth regulators (e.g., α-(4-methylbenzylidene)-γ-butyrolactone [*5418-24-6*] [26]), α-acetobutyrolactone (a vitamin B_1 intermediate) [27], and the rubber additive thiodibutyric acid [28]. Butyrolactone is used as a solvent for polymers [29], [30] and as a polymerization catalyst [31]; in hairwave compositions [32] and sun lotions [33]; and in pharmaceuticals [34]. It is also used in printing inks, e.g., for ink-jet printing [35], as an extractant in the petroleum industry [36], as a stabilizer for chlorohydrocarbons [37] and phosphorus-based pesticides [38], and as a nematocide [39]. More recent applications are in the electronics field as a cosolvent for capacitor electrolytes [40] and as a cosolvent for photoresists [41].

8. Toxicology and Occupational Health

γ-Butyrolactone is readily absorbed from the intestine and through the skin [42], [43]; the half-life was about 20 min in the blood of rats after oral administration [43]. Metabolism and excretion is rapid; γ-hydroxybutyric acid is the primary metabolite. Following i.v. injection in rats, ca. 60 % of radiolabeled lactone was detected in the expired air as CO_2

within 2.5 h with the peak level about 20 min after application [44]. In human volunteers who received 1 g of γ-butyrolactone in water, (*S*)-3,4-dihydroxybutyrate, glycolic acid, γ-hydroxybutyrate, and an isomer of 4-hydroxy-3-oxobutyrate were analyzed in the urine 4 h after application [45].

γ-Butyrolactone has a low acute toxicity: LD_{50} data in the literature vary between 800 and 1800 mg/kg (rat, mouse) and 500 and 1700 mg/kg (guinea pig). The dermal LD_{50} in guinea pigs is 5600 mg/kg [46], [47], [48]. The inhalative LC_{50} (rat, 4-h) lies above the highest tested concentration which did not cause mortality (>5.1 mg/L) [49]. Inhalation (8-h) of air saturated with γ-butyrolactone at room temperature produced no symptoms in rats [50].

γ-Butyrolactone causes depression of the CNS by means of its metabolite γ-hydroxybutyric acid and, in addition, anesthetic effects on the peripheral nervous system [51], [52].

In adult human volunteers, a dose of 2500 mg per person (estimated dose: 35 – 45 mg/kg) induced a 1-h sleep [53]. When used as an anesthetic in children, 66 mg/kg of γ-butyrolactone resulted in deep sleep with rare cases of respiratory depression [54]. Higher specific oral doses, accidentally ingested by adults or infants, are reported to have caused unconsciousness within minutes due to CNS depression, occasionally coma, as well as brachycardia and respiratory depression. The patients had to be provided supportive medical treatment, but apparently recovered completely within ca. 24 h [55], [56].

In rabbits, γ-butyrolactone does not cause primary skin irritation but is irritating to the eye [50]. It does not cause skin sensitization in guinea pigs [57]. In humans, no skin reactions were seen after patch-testing of γ-butyrolactone for skin-contact allergy in 200 volunteers [47]. Lotions containing γ-butyrolactone are well tolerated when applied to the skin [51]. The substance served as successful topical analgesic for the relief of pain arising from local injuries of the musculoskeletal system, including rheumatic complaints, or from headache (e.g., migraine) [51].

Prolonged administration of γ-butyrolactone in drinking water for four weeks at a daily dosage of 3 g/kg led to depression of the CNS in rats. Apart from a slight reduction in body weight gain, the animals showed no further health effects [58]. The low toxicity can be explained by the fast metabolism and elimination. In animal studies with repeated oral administration by gavage to rats and mice, somewhat higher toxicity was observed than with application via drinking water (bolus effect) [59].

The majority of the large number of mutagenicity and genotoxicity studies provides evidence that γ-butyrolactone is devoid of genotoxic potential [47]. Several repeated dose [58], [60] and lifelong carcinogenicity studies [47], [59] failed to reveal any signs of chronic organotoxic effects by any route of application. Two valid two-year (lifetime) carcinogenicity studies in rats and mice receiving up to 450 or 525 mg kg^{-1} d^{-1}) of γ-butyrolactone by gavage (5 d per week in corn oil) showed no evidence of tumorigenic potential. The slight increase of pheochromocytomas in low-dosed male mice was interpreted by the authors as an equivocal response [59]. These findings principally confirmed the consistently negative results found in early dermal and oral studies [61], [62], [47] and therefore allow the conclusion that no significant carcinogenic potential is attributable to γ-butyrolactone.

Prenatal exposure to γ-butyrolactone gave no evidence of substance-related developmental toxic effects either by the oral route (rats, 10 – 500 mg kg^{-1} d^{-1}) [63] or by inhalation (rabbits, 0.5 – 5 mg/L air) [64].

In Europe and the United States, no specific occupational exposure limits for γ-butyrolactone have been established.

Experimental acute toxicity data and its ready degradability suggest that γ-butyrolactone is unlikely to pose any risk to the environment [65], [66].

9. References

1. *Beilstein,* **7 (1),** 234, **7 (2),** 286, **17/5 (3/4),** 4159, **17/9 (5),** 7.
2. *Houben-Weyl,* **6/2,** 561 – 852.
3. G. Quadbeck, *Dtsch. Med. Wochenschr.* **90** (1965) 403.
4. W. Reppe, *Justus Liebigs Ann. Chem.* **596** (1955) 163 – 224; *Chem. Ing. Tech.* **22** (1950) 361 – 373; *Neue Entwicklungen auf dem Gebiet der Chemie des Acetylens und*

Kohlenoxids, Springer Verlag, Berlin – Göttingen – Heidelberg 1949, pp. 45 – 47.
5. S. Kano, S. Shibuya, T. Ebata, *Heterocycles* **14** (1980) 661 – 711.
6. F. Korte, W. Glet, *J. Polym. Sci. Part B* **4** (1966) 685 – 689.
7. S. Oka, *Bull. Chem. Soc. Jpn.* **35** (1962) 986 – 989.
8. Mitsubishi Petrochemical, DE-OS 1593073, 1966; DE-OS 1901870, 1969 (T. Asano, J. Kanetaka).
9. J. Kanetaka, T. Asano, S. Masumune, *Ind. Eng. Chem.* **62** (1970) 24 – 32. T. Yoshimura, *Chem. Eng. N.Y.* **76** (11. Aug. 1969) 70 – 72; *Chem. Week* **104** (1969) 63 – 72. S. Minoda, M. Miyajima, *Hydrocarbon Process.* **49** (1970) no. 11, 176 – 178.
10. M. W. M. Tuck, M. A. Wood, C. Rathmell, P. H. E. Eastland: "Butane to Butanediol: The emergence of a new Process Route," AIChE 1994 Spring International Meeting.
11. Quaker Oats, US 3074964, 1961 (A. P. Dunlop, E. Sherman). H. Hara, JP-Kokai 7887347, 1978.
12. BASF, DE 4 339 269, 1993 (R. Pinkos, R. Fischer).
13. BASF, WO 9 707 111, 1995 (M. Heider et al.).
14. BASF, DE 19 530 549, 1995 (M. Heider, T. Ruehl, J. Henkelmann, S. Stutz).
15. Y. Shvo, Y. Blum, *J. Organomet. Chem.* **238** (1982) C 79 – C 81.
16. Toa Nenryo Kogyo K.K., JP-Kokai 75 154 237, 1975 (Y. Okumura, Y. Nagashima).
17. Nat. Dist. and Chem. Corp., US 4 247 467, 1978 (J. H. Murib).
18. Standard Oil, US 4 238 357, 1980 (Th. Haase, F. A. Pesa).
19. Texaco, US 3 061 614, 1958 (W. M. Sweeney, J. A. Patterson).
20. The British Petroleum Company, EP 176 370, 1984 (H. Alper, D. J. H. Smith).
21. ARCO Chem. Technology LP, US 5 401 857, 1994 (D. Armstead, R. A. Grey).
22. BASF Sicherheitsdatenblatt "Butyrolacton dest.", version 11.02, 1999.
23. I.G. Farbenind., DE-AS 694043, 1938 (C. Schuster, A. Seib).
24. BASF, GB 1312463, 1973 (W. Himmele, E. Hofmann, H. Hoffmann).
25. *Chem. Age London* **71** (1954) 1239.
26. S. Huneck, K. Schreiber, DD 112884, 1975.
27. W. Reppe, *Justus Liebigs Ann. Chem.* **596** (1955) 163 – 164.
28. BASF, DE 917247, 1952 (W. Reppe, H. Friedrich).
29. Phillips Petroleum, US 2 958 677, 1960 (R. F. Kleinschmidt).
30. DuPont, US 3 000 844, 1961 (I. M. Bro).
31. Mitsubishi Petrochemical, JP 8129881, 1981.
32. Shiseido, DE-OS 2421248, 1974 (K. Ono, K. Torii, T. Ozawa).
33. BASF, EP-A 44970, 1982 (F. Thoemel, W. Hoffmann, D. Degner).
34. W. Klunk, A. C. McKeon, *Science* **217** (1982) 1040 – 1042. W. Klunk, D. F. Covey, *Mol. Pharmacol.* **22** (1982) 431 – 437. Grissmann Chemicals, GB 2028653, 1980 (E. Diethalm).
35. Whittaker Corp., DE-OS 2936241, 1980 (I. R. Mansukhani).
36. Esso Research, US 3013962, 1958 (Ch. N. Kimberlin Jr., W. J. Mattox). P. Alessi, I. Kikic, *Int. Congr. Scand. Chem. Eng. Proc.* 1980, 366 – 373.
37. DuPont, US 2958712, 1958 (F. W. Starks).
38. Dow Chemical, JP-Kokai 7986618, 1979.
39. GAF, US 3086907, 1958 (F. A. Hessel).
40. Nippon Chemicon Corp., EP 543 311, 1991 (Y. Yokoyama, T. Tsuji, M. Ue, T. Sato).
41. Samsung Electronics Corp., DE 19 720 413, 1997 (S. Chon et al.).
42. H. L. Fung et al., *J. Pharm. Sci.* **68** (1979) 1198 – 1200.
43. J. Lettieri et al., *Res. Comm. Chem. Pathol. Pharmacol.* **22** (1978) 107 – 118.
44. R. H. Roth, N. J. Giarman, *Biochem. Pharmacol.* **15** (1966) 1333 – 1348.
45. C. R. Lee, *Biochem. Med.* **17** (1977) 284 – 291.
46. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man:* γ-Butyrolactone, vol. 11, Lyon 1976, 231 – 239.
47. BG Chemie, *Toxicol. Bewertungen* **7** (1989).
48. J. Larsen, E. Soderlund: "γ-Butyrolactone," Nat. Inst. of Public Health, Norway 1992.
49. Monsanto Agric. Comp., Study Rep. No. 85061, NTIS/OTS 0534527, 1986.
50. BASF, unpublished results (1960).
51. E. Szirmai et al., *Pharmazie* **44** (1989) 570 – 571.
52. E. H. Roth et al., *Int. J. Neuropharmacol.* **5** (1966) 421 – 428.
53. E. H. Jenney et al., *Pharmacologist* **4** (1962) 166.
54. B. Root, *Anesthesiol.* **26** (1965) 259 – 260.
55. M. B. Andersen, B. Netterstrom, *Ugeskr. Laeger.* **154** (1992) 3064.

56. Arco Chem. Comp., Communication to U.S. EPA on Data from Nordic Poison Centers, NTIS/OTS 0538281, 1992.
57. G. D. Clayton, F. E. Clayton (eds.): "γ-Butyrolactone," *Patty's Industrial Hygiene and Toxicology,* 2nd ed., **vol. 2,** Interscience Publ., New York 1967, pp. 1824 – 1825.
58. M. C. Nowycky, R. H. Roth: "Chronic γ-Butyro lactone (GBL) Treatment: A Potential Model of Dopamine Hypoactivity," *Naunyn Schmiedebergs Arch. Pharmacol.* **309** (1979) 247 – 254.
59. National Toxicology Program (NTP): Toxicology and carcinogenicity studies of gamma-butyrolactone [CAS No. 96-48-0] in F344/N rats and B6C3F1 mice (gavage studies), NTP Techn. Rep. 406, PB92-189323, U.S. DHHS, 1992.
60. GAF Chem. Corp., Subacute feeding studies with butyrolactone in rats and dogs, Report FDRL (Food and Drug Res. Lab., Inc.), Wayne, NJ 1970.
61. B. L. Van Duuren et al.: "Carcinogenicity of Epoxides, Lactones, and Peroxy Compounds," *J. Nat. Canc. Inst.* **31** (1963) 41 – 55.
62. G. Rudali et al.: "A propos de l'action cancérigène de la γ-butyrolactone chez les souris," *C.R. Acad. Sc. Paris, Série D,* **282** (1976) 799 – 802.
63. T. Kronevi et al., *Pharmacol. Toxicol.* **62** (1988) 57 – 58.
64. BASF AG, unpublished results, 1993.
65. V. C. Applegate et al.: Special Sci. Rep. – Fisheries N. 207, U.S. Dept. Interior, Washington 1957.
66. BASF AG: unpubl. results (0009/88) and (89/176), 1988/1990.

Cadmium and Cadmium Compounds

KARL-HEINZ SCHULTE-SCHREPPING, Degussa AG, Bonn, Federal Republic of Germany (Chaps. 1 – 6)

MAGNUS PISCATOR, Karolinska Institute, Stockholm, Sweden (Chap. 7)

1. Introduction 1
2. Cadmium Metal 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.3. Occurrence, Raw Materials 2
2.4. Refining 2
2.4.1. The Starting Materials 2
2.4.2. Leaching of the Starting Materials and Cementation 3
2.4.3. Electrolytic Recovery 3
2.4.4. Distillation 4
2.4.5. Other Processes 4
2.4.6. Recycling 5
2.5. Quality Specifications and Analysis 5
2.6. Uses 5
3. Alloys 6
4. Cadmium Compounds and Complexes 7
4.1. Properties 7
4.2. Individual Compounds 7
4.3. Storage and Transportation 9
5. Environmental Protection 9
5.1. Emission 9
5.2. Waste Water 9
5.3. Air 10
5.4. Soil 10
6. Economic Aspects 11
7. Toxicology and Occupational Health 11
8. References 15

1. Introduction

Cadmium [*7440-43-9*], Cd, atomic number 48, A_r 112.40, belongs to the zinc subgroup of the periodictable along with mercury. In its compounds it has an oxidation state of +2. There are eight natural isotopes, with mass numbers ranging from 106 – 116. The most abundant are ^{114}Cd (29 %) and ^{112}Cd (24 %). There are also a number of metastable isotopes.

STROHMEYER discovered cadmium in 1817 in the course of investigating zinc carbonate. He recognized that the yellow color of a sample of zinc oxide produced by roasting was due to the presence of an unknown metal oxide. Because this new element also occurred in the zinc ore calamine, a name derived from the Latin word "cadmia," he named the new element cadmium.

Unlike some other heavy metals, such as lead or mercury, which have been used since ancient times, cadmium has been refined and utilized only relatively recently. After its discovery more than a century elapsed before the metal or its compounds were employed to any significant extent. Only since the 1940's production and consumption have risen distinctly. The primary uses are electroplated cadmium coatings, nickel – cadmium storage batteries, pigments, and stabilizers for plastics. Publicity about the toxicity of cadmium has affected the consumption.

2. Cadmium Metal

2.1. Physical Properties

Cadmium is a soft, ductile, silver-white metal. Like tin, it makes a grating sound when bent. The most important physical properties are given below.

mp		320.9 °C
bp		767 ± 2 °C
Vapor pressure at	218 °C	0.133 Pa
	302 °C	6.666 Pa
	392 °C	133.32 Pa
	485 °C	1333.2 Pa
	611 °C	13.332 kPa
	727 °C	66.660 kPa
	765 °C	101.33 kPa

Density	8.64 g/cm^3
Specific heat at 20 °C	0.230 J g^{-1} K^{-1}
Heat of fusion	55 J/g
Heat of vaporization at 767 °C	890 J/g
Thermal conductivity at 18 °C	92 J m^{-1} s^{-1} K^{-1}
Electrical resistivity at 18 °C	7.5 × 10^{-6} Ω cm
Brinell hardness	22 – 24

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_499

Coefficient of linear expansion at 20 °C	$31 \times 10^{-6}\ K^{-1}$
Ionization potential	8.96 V
Electrochemical potential	− 0.40 V
Cross section for thermal neutrons	$2450 \times 10^{-28}\ m^2$

The crystal structure is distorted closest packed hexagonal.

2.2. Chemical Properties

Cadmium is stable in air; only a slight loss in luster occurs after an extended period of time in air. When it is heated, initially yellow to brown colors develop as a thin oxide layer forms. If the metal is heated to volatilization, it burns with a red-yellow flame to form brown cadmium oxide, which is poisonous. Cadmium dissolves readily in nitric acid but only slowly in hydrochloric or sulfuric acid and not at all in bases. Zinc displaces it from solution.

2.3. Occurrence, Raw Materials

Cadmium is widely distributed. It occurs in the earth's crust with a content estimated to be between 0.08 and 0.5 ppm. In top soil cadmium content usually lies between 0.1 and 1 ppm.

The best known cadmium mineral is greenockite, cadmium sulfide (77.6 % Cd). One mineral form of cadmium carbonate (61.5 % Cd) was named otavite after its discovery site, the Tsumeb Mine, in the Otavi deposit of South-West Africa. In Sardinia, pure cadmium oxide (87.5 % Cd) was found in a zinc deposit. However, none of these cadmium minerals is of industrial importance because the deposits are too small.

Only zinc minerals in which cadmium is found as an isomorphic component, with concentrations ranging from 0.05 to 0.8 %, averaging about 0.2 %, have economic significance for cadmium recovery [2]. In addition, lead and copper ores contain small amounts of cadmium, which can be separated during the roasting and smelting processes.

2.4. Refining

Figure 1 summarizes the individual process steps and their combination. Detailed description, with emphasis on the distillation process, can be found in [3].

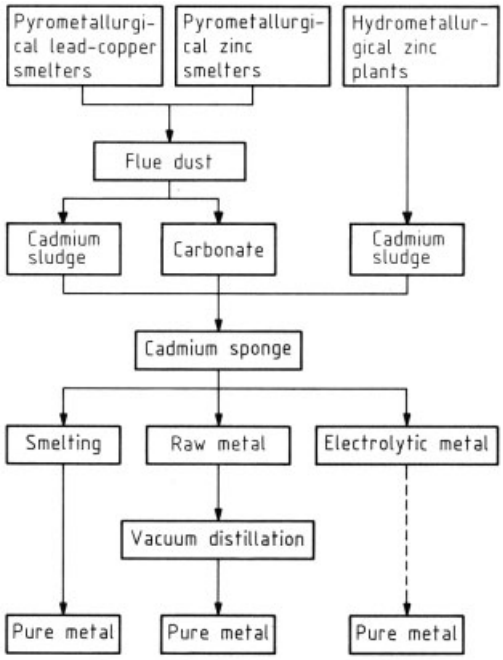

Figure 1. Processes for the production of cadmium metal

2.4.1. The Starting Materials

The flue dust on which the volatile cadmium collects when zinc, copper, and lead ores are heated in air are the primary starting material for cadmium recovery and refining. In many cases it is necessary to recirculate this dust to obtain higher cadmium concentrations [4].

In the Outokumpu Zinc Works in Finland, roasting material contains 0.19 – 0.24 % Cd. The cadmium content is extracted in the third step of the zinc process and results in a cake consisting of 90 % cadmium and 10 % zinc. This is the starting material for further processing [5].

If the primary flue dust is reduced in a rotary oven, lead and zinc remain in the oven, while the cadmium is volatilized and enriched in the secondary flue dust.

In the Berzelius Metal Works, Duisburg, Federal Republic of Germany, complex lead–zinc ores are refined by the Imperial smelting process. When the concentrate is roasted in the sintering furnaces, cadmium-containing flue dust

is produced. This is leached in an acid solution, and the cadmium is subsequently precipitated as cadmium carbonate. The cadmium in secondary raw materials, such as steel mill dust, leaching sediments, and sludge from water treatment, can be processed at the same time. However, before they are mixed with the concentrates they are enriched in a rotating tube furnace (Waelz process).

In hydrometallurgical zinc refining, the cadmium-containing zinc concentrate is leached with sulfuric acid. The cadmium is removed from the solution together with copper by reduction with zinc dust to give a metallic sludge. Similar cadmium sludges form in the purification of zinc solutions used in the production of zinc sulfide pigments or zinc carbonate. These cadmium sludges are the most important starting materials for cadmium refining today. The secondary starting materials mentioned above can be added to the primary raw materials without causing large alterations in the refining process.

Also of economic significance is the recycling of used nickel – cadmium storage batteries; cadmium-containing alloys, such as Sn – Cd metallizing alloys from the manufacture of rectifiers; defective industrial batches; residues from the production of cadmium pigments; and cadmium-containing fluorescent materials. Special processes are usually necessary to convert the cadmium content into a form that can be processed normally. The most important processes of this type are discussed in Section 2.4.6.

2.4.2. Leaching of the Starting Materials and Cementation

Cadmium is extracted from the enriched starting materials by oxidizing solutions. The extraction must be carried out in such a way that all the cadmium as well as the zinc dissolves. The procedure of Ruhr Zinc, Datteln, Federal Republic of Germany, is described here as an example. The liquid from electrolytic baths, which has a sulfuric acid concentration of 25 – 30 g/L, is conveyed into 50 m^3 tanks. In order to oxidize the cadmium, either air is blown in or manganese dioxide is added. Fresh cadmium sludge or cadmium carbonate is added to decrease the acidity of the solution, and a copper-containing precipitate is removed with a filter press. Zinc, cadmium, and thallium are almost completely dissolved in the solution along with almost no copper, cobalt, nickel, and arsenic [6].

Copper chloride mixed with other metal chlorides or iron(III) chloride can be used as an oxidant instead of air [7], [8]. Dust from the sintering process can also be roasted to give sulfates and then be extracted in a similar fashion with water [9].

Independent of the further processing of the cadmium, it is precipitated from the solution as a metal sponge by reduction with zinc. If necessary the solution is purified first.

The cadmium raw solution resulting from the leaching step at Ruhr Zinc contains about 30 g/L of zinc and small amounts of cobalt, nickel, thallium, and copper in addition to the 30 – 40 g/L of cadmium. The cadmium is precipitated in a 50 m^3 tank at a temperature of 60 °C. In order to avoid excess zinc in the cadmium sponge, only 80 % of the stoichiometric amount of the zinc is added initially. The remaining zinc dust is added in small portions until the cadmium concentration of the solution is below 0.1 g/L. After decanting, the cadmium sponge is left in the reaction container. The precipitation is carried out three times, principally to reduce the amount of zinc. After the final precipitation, the sponge is washed, and the wash water is recycled to the zinc plant. The precipitated sponge contains 90 % cadmium and only 2.5 % zinc [6].

The procedure of the Outokumpu Zinc Works starts with a purified cadmium solution because of the nature of the succeeding metal recovery (see Section 2.4.5). Copper is removed with zinc powder, and lead is removed with barium carbonate. The separation of the cadmium sponge from the solution, however, scarcely deviates from the Ruhr Zinc processes. The greatest difference is that the cementation is interrupted at cadmium concentrations of 3 – 5 g/L, and the solution recirculated. Together with a very careful wash, this reduces the residual content of zinc in the compacted cadmium-sponge briquettes to 0.5 – 2.0 % [5].

2.4.3. Electrolytic Recovery

The largest amounts of cadmium are recovered electrolytically. It is customary to extract the cadmium sponge in the same tank with

the cadmium-containing electrolytic acid. Air is supplied at a temperature of 80 °C for 6 – 10 h. Concentrated sulfuric acid is added to keep the concentration of free sulfuric acid at 100 g/L. Manganese dioxide can also be used as oxidant [6].

The time required to dissolve the cadmium sponge at room temperature can be reduced by blowing air through special reactors. The relationship between solid and liquid, the amount of air, and the pH are carefully controlled to reduce the time and temperature required for the dissolution [10].

The electrolytic deposition of cadmium requires that the impurities that cause problems during the electrolysis or that are deposited as impurity with cadmium be removed prior to electrolysis. Traces of copper can be removed by addition of small quantities of sodium sulfide in the form of copper sulfide or by cementation with zinc dust. Thallium is removed as thallium dichromate by addition of potassium dichromate. Thallium can also be removed as thallium(III) hydroxide by addition of potassium permanganate. Any arsenic is eliminated as iron arsenate by oxidation with the potassium permanganate. The lead remaining in solution is removed by adding strontium carbonate to occlude lead carbonate [6].

The anode consists of lead, which does not dissolve, and the cathode is sheet aluminum. Today rotating cathodes, which allowed high current density, have disappeared because current densities of 100 A/m^2, permitting economic electrolysis, can also be attained with fixed cathodes. However, the cathode deposit must be stripped after 12-h intervals to avoid low yields caused by short circuits between cathode and anode. The current yield is $\approx$ 93 %; the energy consumption is 1250 kWh per tonne of cadmium cathode sheet deposited. Careful operation and the use of hide glue and sodium silicate allow cadmium of high purity to be produced. Typical impurity levels in electrolytic cadmium sheet are 5 – 10 ppm Cu, 15 – 20 ppm Pb, 5 – 10 ppm Tl, and 5 ppm Zn.

Some sheet is used as is. Otherwise, it is melted under a cover of sodium hydroxide to prevent oxidation and to reduce the amount of zinc. The molten metal is poured into molds to produce marketable shapes.

2.4.4. Distillation

After mixing with a reducing agent such as pulverized coke, the cadmium sponge can be briquetted. The briquettes can be distilled under atmospheric pressure to a raw metal still containing thallium and zinc. These impurities can be separated from the fused metal by addition of ammonium chloride and caustic soda.

Distillation under reduced pressure avoids a number of disadvantages of distillation at atmospheric pressure. The differences between the vapor pressure of cadmium and those of the most important accompanying elements allows Preussag (Federal Republic of Germany) to produce a 99.99 % pure metal in a single step. The raw cadmium can contain up to 3 % lead and thallium as well as 1 % copper [11]. Every hour, 100 kg of pure cadmium is produced at a distillation pressure of 0.7 – 2.7 kPa and a temperature of 420 – 485 °C. The presence of thallium, the most volatile impurity, at a concentration exceeding 10 ppm in the distillate indicates that the distillation is not functioning properly.

The cadmium distillation process of Hoboken Overpelt, Belgium, allows the processing of cadmium – zinc alloys containing 15 – 20 % Zn. In a two-step distillation procedure, both zinc and cadmium are obtained highly pure (99.995 %) [12].

2.4.5. Other Processes

The cadmium sponge of the Outokumpu Zinc Works contains 0.5 – 2 % zinc, 13 ppm lead, and 15 ppm copper. The briquettes pressed out of this material are smelted with sodium hydroxide. In seven hours 2.2 t of cadmium is obtained using a 110-kW furnace. The refined metal is drawn in a continual casting machine to a cadmium wire, which is then cut to rod. These cadmium rods have a purity of at least 99.95 %; the impurities (max.) consisting of 20 ppm Cu, 10 ppm Zn, 5 ppm Tl, 5 ppm Fe, and 15 ppm Pb [5].

The removal of cadmium from solution by ion-exchange has been described [13]. After stripping, the cadmium is processed further by cementation and smelting. Complexation of the cations to give anionic cadmium species allows selective separation of the cadmium even from

copper and iron [14]. Aliphatic carboxylic acids [15] and *o*-aminomethylphenols [16] selectively extract cadmium from aqueous solution.

2.4.6. Recycling

The purpose of the recycling procedures described here is the production of intermediate products – by special procedures – that can then be fed into the normal cadmium production processes.

The recovery of *cadmium – tin alloys*, about 65 % Sn and 35 % Cd, usually begins by dissolving the scrap in nitric acid. The tin stays in the residue as stannic acid. An impure carbonate is precipitated from the cadmium nitrate solution with soda.

The residue and the rejected batches from the production of *cadmium pigments* containing cadmium sulfide and selenide are dissolved in strongly oxidizing acid. The cadmium is recovered by cementation or as the carbonate. A variation of this process is described in [17].

In order to recycle the cadmium from *nickel – cadmium batteries* and similar production scrap, the parts are dissolved in agitated 2 wt % sulfuric acid. A reasonably good separation from the dissolved nickel can be attained by taking advantage of the lower stability of the ammine complexes of cadmium: cadmium can be precipitated from ammonium hydroxide solution as carbonate, whereas nickel remains in solution [18]. Alternatively, nickel can be extracted from the ammonia solution with chelating reagents before cadmium carbonate is precipitated [19].

Sludge from the treatment of cadmium-containing waste water can often be injected directly into the refinery process (Section 2.4.1). Preliminary, conventional wet chemical enrichment is possible but is usually omitted for economic reasons [20]. Several methods for conversion of electroplating waste water to sludge not requiring special handling can be found in [21].

A plant was constructed by Sab-Nife in Sweden for recycling used batteries and scrap at the same location where nickel – cadmium batteries are produced. Cadmium-containing waste and filter dust are leached with acid, and the cadmium is deposited by electrolysis. Used batteries are dismounted in a semiautomatic process. From the iron- and cadmium-containing plates, the cadmium is distilled under a reducing atmosphere at 850 °C and returned to the production department. Even sealed cells can be processed by distillation after the plastic is destroyed by pyrolytic pretreatment in a slightly oxidizing atmosphere [22].

2.5. Quality Specifications and Analysis

The quality of the cadmium depends on the production process. However, for 99.99 % cadmium, the following maximum impurity levels are generally accepted [11]: 1 ppm Cu, 5 ppm Pb, 5 ppm Tl, and 2 ppm Zn. For special purposes > 99.9998 % cadmium is available.

The maximum amounts of impurities and the analysis procedures are given in ASTM C-752. Although the classical analytical methods for impurities are still used, atomic absorption spectroscopy is the most common method [23].

2.6. Uses

Cadmium metal is used for cadmium coatings, nickel – cadmium batteries, and to a limited extent for reactor construction (see page 6).

Electrodeposited Cadmium Coatings. Cadmium coatings are superior to other metallic coatings because of their optimal combination of properties: corrosion protection, ductility, frictional behavior, and soldering potential. Cadmium coatings have considerable significance in *automotive construction,* where disk brake calipers, bearing races, screws, nuts, bolts, springs, tubing connectors, contacts, and steering parts are all cadmium plated. In *machine construction and factory construction*, the surface of fastening elements, especially screws, is coated. Although the corrosion protection is most important, the frictional properties are also important. For example, even screws that are extremely tight must last for years and then be removable without difficulty. Precision screws and other connecting elements are subject to similar requirements. In the machines of chemical plants they are subjected to corrosion in addition. Aircraft parts not made from stainless steel are usually cadmium coated.

For the often necessary *attachment of different metals*, such as copper alloys, stainless steel, and titanium alloys to aluminum or magnesium alloys, cadmium coating is an absolute prerequisite to avoid corrosion. Cadmium plating is often required in military procurement contracts to ensure that the weapons and equipment function reliably over many years.

Nickel – Cadmium Accumulators (→ Batteries, Chap. 3.2.). The second most widely used storage battery is based on cadmium and nickel. In the charged state the negative electrode consists of cadmium powder, and the positive electrode consists of nickel(III) hydroxide. As the battery discharges, cadmium hydroxide and nickel(II) hydroxide form. Charging reestablishes the original composition. The starting material for the production of the negative electrode is usually cadmium metal, which is dissolved in nitric acid and then precipitated as cadmium hydroxide. In another technique, cadmium hydroxide is repeatedly precipitated into the pores of sintered bodies, which become the negative electrode. A further variation uses electrodeposited cadmium powder, which is mixed with an electrical conductor such as graphite. A properly designed accumulating system does not release gas, even when overcharged, and can be sealed gastight. This development opened the way for new applications, such as rechargeable accumulators not larger than a single primary cell for use in portable electronic products.

In addition to the general advantages of accumulators, nickel – cadmium cells have a long life and are insensitive to low temperature and storage in the uncharged condition. Nonsealed nickel – cadmium accumulators are used in airplanes; telephone, telegraph, and radar stations; computer installations; and emergency power stations. Gastight cells are used in transportable pocket computers, razors, flashlights, etc.

Other Uses. Cadmium compounds, stabilizers for plastics, and pigments (→ Pigments, Inorganic) are generally produced from cadmium metal. The metal is dissolved in a mineral acid, usually sulfuric acid. Dissolution requires considerable time even at high temperatures. Increasingly, cadmium sheets from electrodeposition are used because they, unlike metal blocks, have a larger surface area and dissolve more quickly. Because they do not have to be melted, production costs are lower. In many cases, the high purity of the sheet allows direct use of the cadmium solution without additional purification, especially as the noble impurities remain as residue.

A small amount of cadmium is converted into cadmium sulfide – copper sulfide *solar cells* (→ Photoelectricity), which directly convert light into electrical energy. The yield at present is ≈ 10 %. A yield of 15 % is thought to be attainable. It is expected that these cells can be produced inexpensively. Other cadmium compounds – cadmium telluride, cadmium selenide, and the combination cadmium sulfide – indium phosphide – have also been tested successfully.

The amounts of cadmium consumed for the most important uses are given in Table 2.

3. Alloys

Numerous binary and tertiary cadmium alloys, often with complicated phase diagrams, are known. For practical purposes, cadmium does not dissolve in molten iron or aluminum. A tabular summary of cadmium alloys is available [24], and there is special literature about cadmium alloys [25]. A number of cadmium alloys, as a rule produced simply by melting, are used industrially. They can be divided by their principal use: solder, electrical conductors, and other.

Solder. Cadmium forms alloys with silver, copper, and zinc that have low melting points and are widely applicable for the hard soldering of metals. The cadmium lowers the melting point. More than 80 % of the cadmium used in alloys is used for this purpose.

Aluminum or aluminum alloys can be soldered to copper, brass, or stainless steel by lead alloys that contain antimony, tin, and 2.6 – 17 wt % cadmium [26], [27]. Similar alloys that contain between 0.6 and 60 % cadmium join glass and ceramics at temperatures under 180 °C [28]. For soldered connections on gold-plated electrical conductors, cadmium-containing zinc – lead – indium alloys are recommended [29], [30]. Occupational-medicine aspects of cadmium-containing soldering are treated in [31].

Electrical Contacts and Conducting Wire. The addition of cadmium or cadmium oxide to silver, often the contact material in electrical switches, reduces the metal loss due to sparking. In addition, the welding tendency is decreased. Cadmium (0.8 – 1 wt %) improves the mechanical properties of copper wire without reducing its electrical conductivity. This is especially useful in the case of the overhead wires supplying power to electrically powered locomotives [32], [33].

Other Alloys. Cadmium is a component of a few bearing alloys based on tin, copper, and aluminum, usually in amounts under 1 %. Wood's alloy and Lipowitz's alloy, which contain between 6 and 14 % Cd, are used as fusible alloys in automatic fire-protection sprinklers and other devices. Cadmium in the form of a 5 % alloy with silver and indium is put to use in the control rods of nuclear reactors because ^{113}Cd has a particularly large neutron capture cross section. A cadmium – selenium – indium alloy is an important component of a white-light laser for optical data processing systems [25].

4. Cadmium Compounds and Complexes

4.1. Properties

Generally water-soluble cadmium compounds are colorless. Aluminum and zinc can be used to precipitate cadmium metal from solutions of cadmium salts. From weakly acid or neutral solutions, cadmium sulfide is precipitated by hydrogen sulfide, sodium sulfide, and sodium hydrogensulfides. The temperature, concentration, and presence or absence of Cl^- determine the color of the precipitate: lemon yellow, orange, or even red. Sodium selenide precipitates black cadmium selenide. Mixed precipitates of cadmium sulfide and cadmium selenide are brown. Sodium hydroxide gives white cadmium hydroxide, which is insoluble in excess hydroxide. The precipitate formed by ammonia redissolves in excess reagent to form the complex $[Cd(NH_3)_6]^{2+}$. This complex also forms when ammonia is added to precipitates such as the hydroxide or oxalate. Sodium carbonate produces insoluble cadmium carbonate. Sodium cyanide reacts with aqueous cadmium solutions to give cadmium cyanide, which redissolves in excess reagent to form $[Cd(CN)_4]^{2-}$. This complex is weak enough that the cadmium can be completely precipitated with hydrogen sulfide. Cadmium solutions react with hexacyanoferrate(II) to give a white precipitate, whereas they react with hexacyanoferrate(III) to give a yellow precipitate. Cadmium sulfite is only slightly soluble and can be precipitated from cadmium solution with sodium sulfite. All insoluble cadmium compounds can be dissolved in mineral acids. A few cadmium compounds fluoresce or phosphoresce, the silicate yellow to rose, the borate rose. Silver- or copper-activated cadmium – zinc sulfide gives colors spread over the entire spectrum.

4.2. Individual Compounds

The properties and uses of numerous cadmium compounds are described in [24]. The following text is limited to the most important industrial cadmium compounds.

Cadmium oxide [*1306-19-0*], CdO, M_r 128.41. The amorphous formis yellow red, brown red, or brown black, the color depending on the particle size and the stoichiometry. It has a density of 6.95 g/cm^3. The amorphous oxide is insoluble in water and bases but is readily soluble in dilute acids, ammonia, ammonium salt solutions, and sodium cyanide solutions.

The lustrous black cubic crystals of cadmium oxide have a density of 8.15 g/cm^3. They sublime at $\approx 700\,°C$. The low specific resistance at $0\,°C$, $5.5 \times 10^{-3}\,\Omega\,cm$, is caused by an excess of Cd^{2+} ions, which makes CdO a semiconductor.

Cadmium oxide is produced by evaporation of cadmium metal and oxidation of the vapor. It can also be obtained by thermal decomposition of cadmium nitrate or carbonate or by oxidation of molten cadmium by an oxidizing agent. Commercial cadmium oxide should be completely soluble in sodium cyanide solution and contain no heavy-metal or sulfur impurities.

Cadmium oxide is used as a catalyst in oxidation – reduction reactions, dehydrogenation, cleavage, polymerization, the production of multiply unsaturated alcohols, hydrogenation of

unsaturated fatty acids, and as a mixed catalyst component to produce methanol from carbon monoxide and water. Further uses include resistant enamels, metal coatings for plastics, heat-resistant plastics, selenium ruby glass, starting material for other cadmium compounds, and stabilizers for poly(vinyl chloride). Cadmium oxide combined with an alkali-metal cyanide is the salt mixture used in the baths for cadmium electroplating. Cadmium oxide is a component of batteries. It is temperature resistant and together with silver useful in heavy-duty electrical contacts. Cadmium oxide improves the behavior of some high-temperature plastics.

Cadmium hydroxide [*21041-95-2*] , $Cd(OH)_2$, M_r 146.43, isa colorless powder with a hexagonal, layered lattice. The precipitation of fine-grained cadmium hydroxide is only possible from nitric acid solution, basic salts often resulting from precipitation from other solutions. In the presence of halogenide ions, X^-, the complex $[CdX_4]^{2-}$ forms. In fact, precipitation of $Cd(OH)_2$ can be prevented by a high concentration of alkali-metal chloride. Cadmium hydroxide is a component of cadmium – nickel accumulators and silver – cadmium batteries. Cadmium hydroxide often replaces the oxide as the starting material for other cadmium compounds.

Cadmium carbonate [*513-78-0*] , $CdCO_3$, M_r 172.42, has a density of 5.3 g/cm^3. (The density of $CdCO_3 \cdot H_2O$ is 4.3 g/cm^3.) The white crystalline powder with an orthorhombic lattice of the calcite type decomposes above 360 °C. Often precipitation with sodium carbonate produces hydroxy products, which contain water difficult to remove. Cadmium carbonate is a starting material for the production of cadmium pigments and other cadmium salts.

Cadmium chloride [*10108-64-2*] , $CdCl_2$, M_r 183.32, has a density of 4.05 g/cm^3. The colorless, lustrous orthorhombic crystals, which melt at 568 °C to a liquid that boils at 967 °C, are slightly soluble in water. The hydrates are $CdCl_2 \cdot H_2O$, $CdCl_2 \cdot 21/2\,H_2O$, and $CdCl_2 \cdot 4\,H_2O$. Cadmium chloride is produced by reaction of molten cadmium and chlorine gas at 600 °C or by dissolving cadmium metal or the oxide in hydrochloric acid, subsequently vaporizing the solution. The salt is used in electroplating. An aqueous solution absorbs hydrogen sulfide. Molten cadmium chloride dissolves cadmium metal. Upon cooling, the metal precipitates.

The significance of cadmium chloride as a commercial product is declining. It occurs, however, as an intermediate in the production of cadmium-containing stabilizers and pigments, which are often obtained from a cadmium chloride solution itself obtained from cadmium metal, oxide, hydroxide, or carbonate.

Cadmium nitrate tetrahydrate [*10325-94-7*] , $Cd(NO_3)_2 \cdot 4\,H_2O$, M_r 308.49, has a density of 2.46 g/cm^3. The small colorless deliquescing crystals readily dissolve in water. They have a *mp* of 59.9 °C, and the liquid has a *bp* of 132 °C. There are three hydrates, including the tetrahydrate, and the anhydrous salt (stability ranges):

$Cd(NO_3)_2 \cdot 9\,H_2O$ (−16 to 3.5 °C)
$Cd(NO_3)_2 \cdot 4\,H_2O$ (3.5 to 48.7 °C)
$Cd(NO_3)_2 \cdot 2\,H_2O$ (48.7 to 56.8 °C)
$Cd(NO_3)_2$ (> 56.8 °C)

Cadmium nitrate is produced by dissolving cadmium metal in nitric acid, purifying if necessary, concentrating, and crystallizing. It is used for the production of red lusters in glass and porcelain and in cadmium – nickel sinter plates of storage batteries.

Cadmium sulfate [*10124-36-4*] , $CdSO_4$, M_r 208.46, 4.7 g/cm^3. The melting point of the anhydrous salt is 1000 °C. Anhydrous cadmium sulfate is produced by melting cadmium with ammonium or sodium peroxodisulfate. A saturated aqueous solution contains 76 g $CdSO_4$ per 100 g of water at 0 °C, 77.2 g at 20 °C, 69.4 g at 74.5 °C, and 58.0 g at 100 °C.

$CdSO_4 \cdot 8/3\,H_2O$ [*7790-84-3*], M_r 256.52, ϱ 3.09 g/cm^3. The colorless monoclinic crystals, which effloresce in air, are soluble in water (see above). This hydrate melts at 41.5 °C in its own water of crystallization, converting to the monohydrate.

Cadmium sulfate monohydrate [*13477-20-8*], $CdSO_4 \cdot H_2O$, M_r 226.48, ϱ 3.79 g/cm^3, which is the form usually marketed, is produced by evaporating a cadmium sulfate solution above 41.5 °C.

Cadmium sulfate is used in electroplating and as a starting material for pigments, stabilizers, and other cadmium compounds that can be precipitated from aqueous solution. It is used to produce fluorescent materials and in analysis. A cadmium sulfate solution is a component of the Weston cell, which has an almost constant voltage of 1.0186 V.

Cadmium cyanide [*542-83-6*] , $Cd(CN)_2$, M_r 164.448, ϱ 2.226 g/cm^3, is cubic and isostructural with $Zn(CN)_2$. The solubility in water is 17 g per liter at 15 °C. The solubility in sodium cyanide solution is greater because tricyanocadmium ions, $Cd(CN)_3^-$, form. Cadmium cyanide is produced from dilute cyanic acid and cadmium hydroxide by evaporation or by precipitation from a concentrated solution of cadmium salt and alkali-metal cyanide. Cadmium cyanide and its mixtures with an alkali-metal cyanide are used in electroplating. Cadmium cyanide baths for the electroplating of cadmium metal coatings have the advantage that they are easy to work with. Metal removed from the solution as electroplate is replaced by dissolution of the cadmium anode. These cadmium anodes should be at least 99.95 % pure.

4.3. Storage and Transportation

The labeling for transport of cadmium compounds in the Federal Republic of Germany and the European Economic Community are similar [34], [35]. The rules specify the use of certain symbols and directions for safety precautions. The international provisions for sea and air transport have been made compatible in 1984 [36], [37], although special regulations for individual lands are often still very different. Frequently there are even different provisions for the various types of transportation or routes.

5. Environmental Protection

Cadmium has been ubiquitously distributed in the natural environment for millions of years, and information about this natural distribution in soil and water can be found in general literature, e.g., [24]. Industrial production of cadmium has affected the total distribution in soil and water only insignificantly, although in some restricted areas environmental problems have developed.

More than 90 % of the total nonsmoker intake of cadmium is reported to be through food (see Chap. 7). All sources of cadmium that contaminate cultivated soil, whether by air, fertilizer, or water, should be reduced as much as possible.

Cases of cadmium poisoning caused by the contamination of the environment are described in page 15.

5.1. Emission

Cadmium emission in the Federal Republic of Germany is given in Table 1. Summaries for the United Kingdom [38], the European Economic Community [39], and the United States [40] are difficult to compare because different parameters have been selected. However, this literature cites its sources, so comparison based on tabulations in these sources is possible. Data about industrial emission in the European Economic Community and extensive discussion of the technologies to reduce emissions have been collected [41].

5.2. Waste Water

Cadmium ions are normally separated from sewage by precipitation with an alkali-metal hydroxide or carbonate, with subsequent separation of the sediment. The required limit of 1 mg/L, however, is only attainable when the pH is adjusted to 9.5 or higher, and the precipitate is thoroughly separated from the solution. New regulations go even further and the "Cadmium Guidelines" for the European Economic Community require concentration limits of 0.3 – 0.5 mg/L after 1986 and $\leqq$ 0.2 mg/L after 1989 [42]. These limits can only be met by manufacturers and consumers with special methods [41]. Especially useful for producers of cadmium and cadmium compounds are the coprecipitation of cadmium in so-called collectors and subsequent filtration or special treatment during production and before combination with cadmium-free waste water. In special cases, ion exchange, precipitation, for example, with a sulfide compound, or electrolytic separation is

Table 1. Cadmium emission in the Federal Republic of Germany

Source (Year)	Emission, t/a			
	Air	Soil	Water	Total ∗
Production of Cd metal and recycling (1980)	5 – 7	–	62	69
Pigments (1980)	0.17	–	1.2	1.4
Stabilizers (1980)	0.05	–	0.18	0.23
Electroplating (1981)	–	–	0.7	0.7
Batteries (1982)	0.31	–	0.9	1.2
Alloys (1982)	0.5	–	–	0.5
Glass (1982)	3	–	–	3
Fossile fuels (1982)<	5	–	–	5
Iron and steel (1982)	5	–	–	5
Cement (1982)	0.3	–	–	0.3
Fertilizer (1982)	–	35	0.5	35
Waste-water sludge (1982)	0.1	3.6 – 14.4 ∗∗	–	14
Man-made deposits (1982)	–	–	–	–
Waste burning (1982)	2.9 – 3.8	–	–	3.8
Automobile exhaust (1982)	< 0.5	–	< 0.5	1.0
Total, max.	*26*	*49*	*66*	*140*

∗ Maximum.
∗∗ Forty percent of the 1.8×10^{-6} t of waste water sludge produced each year, or 0.72×10^{-6} t, is used agriculturally. If the sludge contains, on the average, 5 ppm Cd, this corresponds to the minimum estimate, 3.6 t Cd. Twenty ppm corresponds to the maximum estimate, 14.4 t Cd.

used. Clearly the most important factor is the reduction of the amount of cadmium carried into the waste water. Consequently the guidelines of the European Community concentrate on reduction of the amount of cadmium in waste water.

The constant tightening of the standards for waste water purification is intended to prevent excessive concentrations of cadmium in the sludge produced in biological waste-water treatment. When the concentration of heavy metals is too high, the sludge can no longer be used in agriculture. In the Federal Republic of Germany, maximum cadmium content in sludge is restricted to 20 mg per kilogram of dry waste, along with a limitation on the total amount of waste [43]. Similar requirements are in preparation for the European Economic Community.

5.3. Air

Practically all cadmium emission sources are in the form of the oxide because this compound is formed in pyrometallurgical production, high-temperature processing, and combustion, either directly or via the metal. Initially the oxide is finely divided, but these fine particles can be made to agglomerate into coarse grains in suitable apparatus. However, this specialized apparatus is often not practical, and other methods, which also work well enough with fine particles, must be used. The laws vary from land to land. In *TA-Luft* (Technical Regulations for Air) the Federal Republic of Germany requires that not more than 0.2 mg Cd per m^3 of air be released.

Larger gas effluents are usually subject to an electrostatic purification, for example, the removal of particles from gas exhausts of power plants and garbage disposal plants. For low cadmium contents this separation suffices. For high cadmium concentrations filtration is used. The filters can be made from metal screen, ceramic, or a variety of textiles. Regulations and gas release must be carefully adapted to conditions. Washing, especially with Demisters, has proved useful for removal of water-soluble aerosols [44].

5.4. Soil

Cadmium reaches agricultural soil via airborne particles; wastes, particularly sludge; and phosphate fertilizers. The importance of these three has been summarized and described for the Federal Republic of Germany [45].

The normal content of an agricultural field lies between 0.1 and 1 mg Cd per kilogram of soil (0.1 – 1 ppm). Maximum 3 ppm in a 30-cm layer

of top soil is thought allowable, corresponding to 13.5 kg/ha. The contribution to this reservoir per year is 7 g/ha from the air and 6 g/ha from phosphate fertilizer. The removal by harvest is about 4 g/ha. The annual increase is then 9 g/ha. This means that a soil with a content of 4.5 kg/ha, corresponding to 1 ppm, would reach the upper limit after 1000 years.

A number of studies deal with the disposal of sludge [46]. Even then there are still 265 years until the upper limit is reached; therefore, there is time to solve the problem of cadmium emissions. The national regulations for the production of waste-water sediment are tabulated in [47].

However, there are already problems associated with excessive cadmium contents in soil on the grounds of old mining and smelting plants. In the immediate vicinity of the sources of emission, the amount of cadmium in the soil exceeds that thought allowable. Here the use of the adjacent agricultural land must be limited, but there is no large-scale danger. Apparently man-made waste deposits do not create a cadmium source for plants since transfer of cadmium to the soil does not take place.

In addition to concern about the amounts of cadmium [48], the complicated mechanism of transfer from the soil to the plant must be considered. A transport model, from soil into the plant, has been proposed [39].

6. Economic Aspects

Since 1970 the production of the non-communist countries has remained rather constant, 12 – 15 kt/a. The increase in European production was balanced by the decrease in North America. Consumption has remained in equilibrium with production. The amounts of cadmium for various uses have varied since 1970, affected by technical, economic, and environmental factors [49]. The cadmium price from 1970 to the beginning of the 1980s varied strongly, from \$ 0.50 to \$ 4 per pound. The consumption in the Western world between 1970 and 1982 is tabulated in Table 2. The figures are totals of the published statistics for the Federal Republic of Germany, Japan, the United Kingdom, and the United States, which together account for about two thirds of non-communist consumption.

7. Toxicology and Occupational Health

For more than a century there have been many reports on acute poisoning caused by cadmium compounds. Cadmium uptake occurred either via inhalation during occupational exposures or by oral poisoning due to ingestion of contaminated food or beverages [50]. In 1974 it was established [51] that long-term inhalation of cadmium oxide dust could cause a syndrome characterized by damage to the pulmonary and renal systems. Since then many studies have been made in various countries, establishing that occupational exposure to cadmium compounds can cause adverse effects, especially in the kidneys.

During the last decades it was also recognized that cadmium can cause renal effects in the general population. In Japan, a large population is exposed to cadmium via food, especially rice; in one area this resulted in an epidemic of chronic cadmium poisoning with severe bone damage, called *Itai-Itai disease.*

Whereas the exposure levels were reduced considerably in many industries, there is a growing concern about risks for the general population around the world. This is not only caused by emissions from industry. Waste disposal, especially by incineration, the increasing use of phosphate fertilizers, and pH changes in soil and water caused by acid rain are other factors of concern. For more detailed information on different aspects of occurrence, metabolism, and effects of cadmium, see [50], [52–56].

Exposure Levels. The average normal daily intake via food is 10 – 20 µg in most countries; considerably higher values, from 30 to above 200 µg/d, have been reported from certain areas in Japan [50], [52], [54], [55]. The highest cadmium concentrations are found in some basic foodstuffs such as wheat and rice and in liver, kidney, and certain seafoods. Drinking water generally is a minor source. Atmospheric levels of cadmium in rural areas are less than 1 ng/m^3, in urban areas 1 – 10 ng/m^3, and in certain industrialized areas a yearly average of up to 50 ng/m^3 has been found [50]. Atmospheric exposure to cadmium therefore is of minor importance.

Table 2. Consumption of cadmium, t

Use	1970	1973	1976	1979	1982
Plating	3096	3273	3578	3326	2205
Batteries	680	1580	2106	2202	2180
Pigments	2021	3231	2466	2578	1819
Stabilizers	1947	1892	1195	1184	947
Alloys	642	877	519	400	488
Total	*8386*	*10853*	*9864*	*9690*	*7639*

Tobacco may contain 1 – 2 µg Cd per gram, and smoking twenty cigarettes a day has been estimated to result in an inhalation of about 3 µg. Because a large part of that dose will be absorbed, exposure via cigarette smoking may contribute to the internal dose as much as that absorbed from food [50], [52], [55].

Metabolism. In order to understand the toxicity of cadmium some basic facts about the metabolism of this metal must be known.

The fate of inhaled cadmium compounds depends on particle size and solubility. Finely divided cadmium oxides, especially fumes, deposit in the lower respiratory tract; because of their relatively high solubility about 30 % of the inhaled amount is absorbed. Cadmium sulfide and sulfoselenide are relatively slightly soluble, and they are not absorbed to the same extent. (Mucociliary transport to the gut occurs, but the absorption from the gut probably is only a few percent.)

Ingested cadmium generally is absorbed to a few percent [50], [57]. Nutritional deficiencies, e.g., in iron and calcium, will cause higher absorption. Women with severe iron depletion absorbed up to 20 % of the cadmium ingested orally [57].

Cadmium absorbed from the lungs or the gut initially is stored mainly in the liver. Exposure to cadmium induces the synthesis of metallothionein, a low molecular mass, cysteine-rich protein, which strongly binds cadmium but also zinc and copper [58]. The liver has a high capacity to synthesize this protein, and even at very high exposures most of the cadmium in the liver will be bound to metallothionein. The protein breaks down relatively fast, but continuous synthesis ensures that cadmium does not escape to bind to other structures.

In the blood, cadmium is found in the cells and plasma, the concentrations being approximately equal. However, with increasing cadmium levels in the body more cadmium will be found in the cells. In the red blood cells cadmium is bound to several proteins, a major part to metallothionein. In plasma only a minor part is bound to metallothionein. Plasma metallothionein easily passes the glomeruli, and as a protein it will be reabsorbed in the proximal tubules. In the kidneys the protein then breaks down, and cadmium is released. The kidney can also synthesize metallothionein, which ensures that cadmium is trapped in an inert form. However, when the synthesizing capacity is exceeded, cadmium ions will be released and toxic effects may occur.

The cadmium excretion from the kidney is very small. Long retention times in liver and kidney, as well as in other tissue, lead to an accumulation of cadmium in the kidneys from birth to middle age. The newborn is virtually free from cadmium because the placenta is an efficient barrier. The gross biological half-life of cadmium in the body has been estimated to be about twenty years [50].

The average cadmium concentration in the liver of adults is about 1 mg/kg wet weight; the concentration in the renal cortex is about 20 mg/kg in most European countries and in North America [55], [56], [59]. In Belgium higher concentrations, on the average 40 mg/kg in renal cortex, have been reported [60]. The highest concentrations in members of the general population have been found in Japan [55], [56], [59], [60].

Smokers generally have about twice as high cadmium concentrations in the renal cortex (about 25 mg/kg wet weight) as compared with nonsmokers [59], [60].

In the blood the cadmium concentrations in nonsmokers are generally less than 1 µg/L, whereas smokers have up to 5 µg/L [60]. The difference between smokers and nonsmokers is also obvious in the urine. Nonsmokers in Eu-

rope and North America excrete about 0.4 µg/g creatinine, whereas smokers excrete about twice as much [59], [61]. This urinary excretion is related to the level of cadmium in the body and is useful for monitoring general populations as well as most persons that are occupationally exposed to cadmium. The total body burden in adult nonsmokers generally is less than 10 mg [59], whereas that of the newborn is only a few micrograms.

Acute Toxicity. Acute cadmium poisoning by inhalation has mainly been caused by accidental exposure to cadmium fumes. The highest risk is in welding, cutting, or soldering operations when cadmium-containing materials are treated, especially if the worker is unaware cadmium is present. There may be a latent period of up to 24 h after exposure before symptoms occur. Chemical pneumonitis occurs, which may lead to lung edema and general symptoms, in some cases with lethal outcome [50], [55], [62], [63].

The concentration of cadmium in the air has not been measured at the time of accidents. Based on cadmium analysis of lung tissue from people dying after acute exposure it has been estimated that the lethal dose corresponds to 1 mg/m^3 as cadmium fume for eight hours. Concentrations of 0.5 – 1 mg Cd/m^3 for a couple of hours may cause pneumonitis [53]. A short-term exposure limit for cadmium oxide fumes and respirable dust of 0.25 mg Cd/m^3 has been recommended to prevent acute lung reactions [53].

Single *oral* doses of about 10 – 15 mg may cause gastrointestinal disturbances, but the lethal dose is probably several hundreds of milligrams [53], [55].

Chronic Toxicity. The chronic effects after long-term inhalation are mainly seen in lungs and kidneys [50], [51], [55]. In earlier times anemia, anosmia, and yellow lines on the teeth were common among cadmium workers but are uncommon today. Emphysema was a major finding in some earlier studies involving exposure to cadmium oxide fume and dust [50]. Because of the decrease in exposure, lung dysfunction is nowadays rare, and if it occurs the symptoms are mild. Figure 2 shows how air concentrations of cadmium have decreased during the years in a Swedish factory, where emphysema was seen in the 1940s [51], [64]. It has been concluded that to prevent any pulmonary effect of cadmium the time-weighted average concentration of respirable dust in air should not exceed 20 µg/m^3 [53].

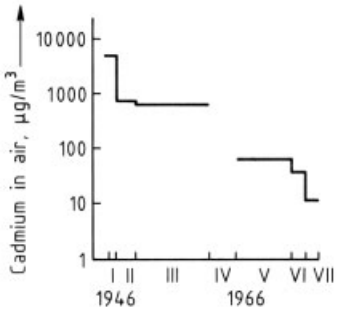

Figure 2. Concentration of cadmium in workroom air from 1946 to 1977
Arithmetic means of personal and stationary samples were obtained within seven subperiods. Data covering period IV, 1961 – 1964, are not available.

The critical effect of cadmium is *renal tubular dysfunction.* In 1950 it was discovered that workers exposed to cadmium oxide dust in an alkaline battery factory had an unusual type of proteinuria [51]. The majority of the proteins were of low molecular mass. Studies of renal function showed that the glomerular filtration rate and the concentrating capacity were decreased. The proteinuria in chronic cadmium poisoning is a tubular proteinuria, caused by a decrease in the reabsorption of filtered plasma proteins in the proximal tubules [50].

The urinary excretion of protein was found to be related to exposure time [65]. The critical cadmium concentration in renal cortex was estimated to be 200 mg/kg wet weight [50]. This value was confirmed by further studies [66–68] and is to be compared with the normal concentrations reported in the previous section. A critical urine concentration of about 10 µg of cadmium per gram of creatinine has also been established, which generally is a good indicator of the body burden [50], [67–69].

The best dose – response relationship seems to be between cadmium concentration in the liver, determined by in vivo neutron activation, and the degree of tubular dysfunction [67]. At liver concentrations above 30 mg/kg wet weight proteinuria appeared; this corresponds to a cad-

mium concentration in renal cortex of about 200 mg/kg.

The increased excretion of low molecular mass proteins is regarded as the first sign of renal tubular effects caused by cadmium. This is also consistent with the fact that cadmium primarily is deposited in the proximal tubules, because the cadmium-containing metallothionein is reabsorbed.

At this early stage there are no measurable changes in the excretion of other substances. However, if exposure goes on and the renal levels increase, other functions may be disturbed. In advanced cases of chronic cadmium poisoning the excretion of glucose, amino acids, and phosphate is increased (Fanconi syndrome). Disturbances in mineral metabolism may cause the formation of renal stones. For detailed discussion, see [70], [71].

The *diagnosis* of chronic cadmium poisoning must be based on occupational history, exposure levels in air, levels of cadmium in blood and urine, and protein analysis in urine.

The tubular damage is irreversible, and *prevention* is thus more important than diagnosis. The determination of cadmium in whole blood gives information about recent exposure, but is not always a good indicator of the body burden or risk for renal dysfunction, especially in smokers. Cadmium in urine should be used to monitor this risk. As long as the urinary excretion of cadmium is below 5 µg per gram of creatinine there should be no risk for the kidneys. At concentrations of 5 – 10 µg per gram of creatinine, exposure should be minimized to prevent further accumulation in the kidneys.

Monitoring is also necessary after exposure has ceased. Intensive exposure during relatively short periods may lead to a high cadmium concentration in the liver. Cadmium will then be released from the liver and transported to the kidneys for a long time. The renal concentrations may reach a critical level many years after the last exposure.

The concentration of cadmium in air is seldom of help in evaluating risks. The long biological half-life of cadmium will lead to a continuous increase in renal levels over many years, and past exposure often is more important than present exposure. Changes in workplace exposure during several decades are shown in Figure 2. Figure 3 demonstrates that cadmium blood levels vary greatly. This is most obvious with smokers. The natural cadmium content of cigarettes does not explain this finding, which is caused instead by contamination of cigarettes or pipe tobacco during work and poor personal hygiene, as illustrated in Figure 4 [64].

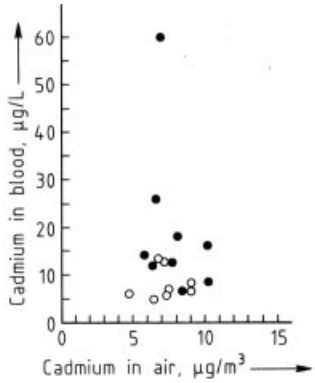

Figure 3. Cadmium in blood vs. cadmium in air
Arithmetic means of 4 – 5 samples of blood cadmium and 9 – 11 personal samples of airborne cadmium dust for 17 alkaline battery workers. Smokers are represented by the solid circles; nonsmokers, by open circles.

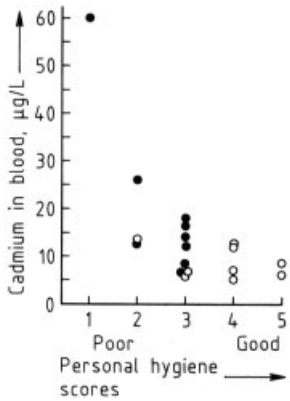

Figure 4. Cadmium in blood in relation to the personal hygiene score
Arithmetic means of five samples of cadmium in blood from nine smokers (solid circles) and of 4 – 5 samples from nine nonsmokers (open circles).

Therefore, it is imperative to prevent smoking in the workshops and the carrying of tobacco in work clothes; such measures will have much more impact than small reductions of cadmium concentrations in air. Dust accumulation must be prevented in the workshop to minimize exposure caused by contamination of work clothes

and hand-to-mouth transfer. The latter may lead to considerable gastrointestinal exposure [72].

The threshold limits for cadmium in air differ among countries, varying from 10 µg Cd/m^3 in Finland to 100 µg/m^3 (total dust) in East European countries. The low values, 10–20 µg/m^3, should prevent renal damage, but as indicated above, smoking and poor personal hygiene may increase the internal exposure.

Carcinogenicity. Cadmium has been implicated as a carcinogenic agent, especially with regard to cancer of the prostate. The evidence is weak and based on deaths from cancer of the prostate. However, this cancer is extremely common among elderly men, and it often does not give any symptoms [73], [74]. There is therefore no evidence that cadmium induces this cancer; diet and hormonal factors are more important determinants. However, it cannot be excluded that in heavily exposed workers cadmium might have affected zinc metabolism in the prostate or caused hormonal changes. It is extremely unlikely that low exposures will have any effect on the prostatic function.

According to animal experiments long-term exposure to cadmium chloride aerosols causes lung cancer in rats [75]. The concentrations of cadmium in air were relatively low, 12.5–50 µg/m^3, but exposures were for 23 h a day. Humans are generally not exposed to cadmium chloride. In the Federal Republic of Germany cadmium chloride has been listed as a carcinogen. There have been a number of studies on the occurrence of lung cancer among workers exposed to cadmium oxide dust, but there are no conclusive results indicating that cadmium oxide causes lung cancer in humans [73].

Poisoning from Contamination of the General Environment. Contamination of the general environment has caused cadmium poisoning in certain areas of Japan [64]. Cadmium levels in rice caused daily intakes of several hundreds of micrograms, which should be compared to the "normal" intake of 10–20 µg in most countries in Europe and North America. The exposure in Japan has caused renal tubular dysfunction of the same type as described in the section on chronic toxicity (page 13). In one area in the Toyama prefecture the cadmium exposure resulted in the development of severe bone damage, Itai-Itai disease. Mainly multiparous women above 40 years of age were affected. The males in the area had tubular dysfunction but generally did not develop the bone disease. The women had poor nutritional status with deficient intakes of calcium, vitamin D, and protein, probably also of iron. Cadmium alone can hardly cause such severe bone changes, but combined with severe calcium deficiency the result became disastrous. It is not likely that Itai-Itai disease will occur again, but as mentioned tubular dysfunction is common in several other areas of Japan.

There are some contaminated areas in Europe, where signs of renal disease have been sought [76], [77]. No definite conclusions can yet be drawn, but the investigated populations had higher cadmium levels than generally found in Europe [77], [78].

Most domestic animals are not at risk for cadmium poisoning. An exception is the horse. Adult horses generally have about five times higher cadmium concentrations in the renal cortex than adult humans in the same area [79].

8. References

General References

1. Bundesverband der Deutschen Industrie e.V. (ed.): *Cadmium – Eine Dokumentation,* **No. 154,** Köln 1982.

Specific References

2. A. v. Röpenack, *Erzmetall* **35** (1982) 534.
3. *Raffinationsverfahren in der Metallurgie '83,* Section D 4, pp. 139–156, Verlag Chemie, Weinheim 1983.
4. D. Blana, *Erzmetall* **32** (1979) 262–266.
5. V. Sipilä, *Erzmetall* **32** (1979) 527–529.
6. *Metal Bulletin Monthly*, April 1983, 115–117.
7. Soc. Minière et Métallurg. de Penarroya, US 4230487, 1979.
8. Cyprus Metallurg. Proc. Corp., GB 1511323, 1974.
9. R. E. Sheppard, A. O. Martel, *Min. Eng. (Littleton, Colo.)* **24** (1972) 80.
10. I. Pajak et al., *Pr. Inst. Met. Niezelaz.* **8** (1979) no. 4, 185–191.
11. G. Schenker, *Erzmetall* **32** (1979) 524–526.
12. C. A. Lemaître, *Erzmetall* **32** (1979) 530–531.
13. A. Cornea et al., *Rev. Chim. (Bucharest)* **29** (1978) no. 11, 1031–1034.

14. Nederlandse Centrale Organisatie Voor Toegepast Natuurwetenschappelijk Onderzoek, GB 2014122 A, 1978.
15. M. Verhaege, *Bull. Soc. Chim. Belg.* **87** (1978) no. 9, 651 – 657.
16. Berol Kemi AB, GB 2001618 A, 1977.
17. K. Hanusch, *Metall (Berlin)* **35** (1981) 911 – 912.
18. Degussa, DE 2001985, 1970.
19. Nife Jungner AB, GB 1475863, 1975.
20. T. B. Lloyd, K. J. Wise, *Ed. Proc. Int. Cadmium Conf. 3rd,* Miami 3 – 5 Feb. 1981, Cadmium Association, London.
21. R. Kammel, H.-W. Lieber, *Galvanotechnik* **68** (1977) 710 – 715.
22. N.-E. Barring, *Ed. Proc. Int. Cadmium Conf. 4th,* Munich 1 – 4 March 1983, Cadmium Association, London.
23. E. Jackwerth, S. Salewski, *Fresenius Z. Anal. Chem.* **310** (1982) 108 – 110.
24. M. Farnsworth: *Cadmium Chemicals,* Int. Lead Zinc Research Organisation, New York 1980.
25. J. F. Cole, D. S. Carr, *Ed. Proc. Int. Cadmium Conf. 4th,* Munich 1 – 4 March 1983, Cadmium Association, London.
26. Soc. Anonyme des Usines Chausson, US 3969110, 1974.
27. Soc. Anonyme des Usines Chausson, GB 2019440 A, 1978.
28. Asahi Glass Co., US 4106930, 1972.
29. V. Prakash et al., *Proceedings, Solder Seminar,* Bangalore, India, 5 – 6 Nov. 1979, Lead Development Association, London.
30. H. Heinzel, K. E. Saeger, *DVS Ber.* 1976, no. 40, 55 – 58.
31. M. Mahler, K. F. Zimmermann, *Schweißen + Schneiden* **34** (1982) 277 – 281.
32. A. K. Woollaston, M. S. Stamford, *Metall. Met. Form.* **44** (1977) 100 – 101, 103 – 104.
33. S. T. Udeshi, A. A. Sahay, *Proceedings, Cadmium Seminar,* Udaipur, India, 26 Feb. 1980, Cadmium Association, London.
34. Arbeitsstoffverordnung, Bundesgesetzblatt 1982, part I, p. 144.
35. Guideline 79/831/EEC, 18 Sept. 1979, Official Journal of the European Community no. I 259/10, 15 Oct. 1979.
36. International Maritime Organisation London: International Maritime Dangerous Goods – Code (IMDG-Code) Amendment 20-82.
37. International Civil Aviation Organisation (ICAO): Technical Instructions for the Safe Transport of Dangerous Goods by Air. Addition 1984.
38. Department of the Environment: *Cadmium in the Environment and its Significance to Man,* Pollution Paper no. 17, Her Majesty's Stationery Office, London 1980.
39. M. Hutton: *Cadmium in the European Community,* contract no. 333, ENV U.K., London 1982.
40. R. Coleman et al., *Sources of Atmospheric Cadmium,* contract no. 68-02-2836, prepared for U.S. Environmental Protection Agency, 1979.
41. A. Rauhut, *Industrial Emissions of Cadmium in the European Community,* European Community Study contract no. ENV/223/74 E, Oct. 1978.
42. Guideline 83/514/EEC, 26 Sept. 1983, Official Journal of the European Community no. L 291/1, 24 Oct. 1983.
43. Klärschlammverordnung – AbfKlärV, 25 June 1982, Bundesgesetzblatt 1982, part I, 26 June 1982, p. 734.
44. K. H. Schulte-Schrepping, *Cadmium, Ed. Proc. Int. Cadmium Conf. 2nd,* Cannes 6 – 8 Feb. 1979, Metal Bulletin, London.
45. K. H. Schulte-Schrepping, *Ed. Proc. Int. Cadmium Conf. 4th,* Munich 1 – 4 March 1983, Cadmium Association, London.
46. A. Kloke: *Landwirtschaftliche Forschung,* Sonderheft 39, Kongreßband 1982, I. D. Sauerländer's Verlag, Frankfurt, p. 302.
47. D. Purves: *International Conference Heavy Metals in the Environment,* Amsterdam Sept. 1981, CEP Consultants Edinburgh U.K.
48. A. Cottenie: *International Conference Heavy Metals in the Environment,* Amsterdam Sept. 1981, CEP Consultants Edinburgh U.K.
49. D. A. Temple, D. N. Wilson: *Ed. Proc. Int. Cadmium Conf. 4th,* Munich 1 – 4 March 1983, Cadmium Association, London.
50. L. Friberg, M. Piscator, G. F. Nordberg, T. Kjellström: *Cadmium in the Environment,* 2nd ed., Chemical Rubber Co., Cleveland 1974.
51. L. Friberg, *Acta Med. Scand. Suppl.* **138** (1950) 240.
52. Commission of the European Communities: *Evaluation of the Impact of Cadmium on the Health of Man,* Luxembourg 1977.
53. World Health Organization: "Recommended Health-based Limits in Occupational Exposure to Heavy Metals," *Tech. Rep. Ser.* **647** (1980) .
54. K. Tsuchiya (ed.): *Cadmium Studies in Japan – A Review,* Elsevier – North Holland Biomedical Press, Amsterdam 1978.

55. L. Friberg, T. Kjellström, G. Nordberg, M. Piscator in L. Friberg, G. Nordberg, V. Vouk (eds.): *Handbook on the Toxicology of Metals,* Elsevier – North Holland Biomedical Press, Amsterdam 1979, pp. 355 – 381.
56. M. Piscator in A. S. Prasad (ed.): *Clinical, Biomedical, and Nutritional Aspects of Trace Elements,* A. R. Liss, Inc., New York 1982, pp. 521 – 536.
57. P. R. Flanagan, J. S. McLellan, J. Haist, G. Cherian, et al., *Gastroenterology* **71** (1978) 841 – 846.
58. J. H. R. Kägi, M. Nordberg (eds.): *Metallothionein,* Birkhäuser, Basel 1979.
59. T. Kjellström, *EHP Environ. Health Perspect.* **28** (1979) 169 – 197.
60. L. Friberg, M. Vahter, *Environ. Res.* **30** (1983) 95 – 128.
61. C.-G. Elinder, T. Kjellström, L. Linnman, G. Pershagen, *Environ. Res.* **15** (1978) 473 – 484.
62. P. A. Lucas, A. G. Jariwalla, J. H. Jones, J. Gough, et al., *Lancet 1980,* **vol. II,** 205.
63. J. R. Patwardhan, E. S. Finch, *Med. J. Aust.* **1** (1976) 962 – 966.
64. E. Hassler: "Exposure to Cadmium and Nickel," Dissertation, Karolinska Institute, Stockholm 1983.
65. M. Piscator, *Arch. Environ. Health* **4** (1962) 607 – 621.
66. K. J. Ellis, W. D. Morgan, I. Zanzi, S. Yasamura, et al., *J. Toxicol. Environ. Health* **7** (1981) 691 – 703.
67. H. Roels, R. R. Lauwerys, J. P. Buchet, A. Bernhard, et al., *Environ. Res.* **26** (1981) 217 – 240.
68. H. Roels, R. Lauwerys, A. N. Dardenne, *Toxicol. Lett.* **15** (1983) 357 – 360.
69. R. Lauwerys, H. Roels, M. Regniers, J. P. Buchet, et al., *Environ. Res.* **20** (1979) 375 – 391.
70. A. Bernhard, J. P. Buchet, H. Roels, P. Masson, et al., *Eur. J. Clin. Invest.* **9** (1979) 11 – 22.
71. M. Piscator in H. Zumkley (ed.): *Spurenelemente,* Thieme Verlag, Stuttgart 1983, pp. 81 – 97.
72. E. Adamsson, M. Piscator, K. Nogawa, *EHP Environ. Health Perspect.* **28** (1979) 219 – 222.
73. M. Piscator, *EHP Environ. Health Perspect.* **40** (1981) 107 – 120.
74. M. Piscator in D. Wilson, R. A. Volpe (eds.): *Ed. Proc. Int. Cadmium Conf. 3rd,* Miami 3 – 5 Feb. 1981 , Cadmium Association, London 1982, pp. 135 –137.
75. S. Takenaka, H. Oldiges, H. König, D. Hochrainer, et al., *J. Natl. Cancer Inst.* **70** (1983) 367 – 371.
76. M. Carruthers, B. Smith, *Lancet 1979,* **vol. I,** 845 – 847.
77. H. R. Roels, R. R. Lauwerys, J.-P. Buchet, A. Bernhard, *Environ. Res.* **24** (1981) 117 – 130.
78. T. C. Harvey, D. R. Chettle, J. H. Fremlin, I. K. Al Haddad, et al., *Lancet 1979,* **vol. I,** 551.
79. C.-G. Elinder, L. Jönsson, M. Piscator, B. Rahnster, *Environ. Res.* **26** (1981) 1 – 21.

Caffeine → Purine Derivatives
Calciferols → Vitamins
Calcined Phosphates → Phosphate Fertilizers

Calcium and Calcium Alloys

Stephen E. Hluchan, BCI, Cambridge, MA, 02138, United States
Kenneth Pomerantz, Cornell University Medical College, New York, NY, United States (Chap. 6)

1. Introduction 1
2. Physical and Chemical Properties . . 1
3. Production 2
4. Uses 3
5. Occupational Health and Safety 5
6. Biological Relevance 7
6.1. Introduction 7
6.2. Chemical Properties of Ca Important in Biological Functions 7
6.3. Dietary Considerations; Distribution in the Body 8
6.4. Regulation of Plasma Calcium by Hormones 8
6.5. Control of Plasma Calcium by Target Organs 9
6.6. Physiology of Intracellular Calcium Trafficking 10
6.7. Physiology of Bone 10
6.8. Disturbances in Plasma Calcium Homeostasis 11
6.9. Diseases of Bone 11
7. References 12

1. Introduction

Calcium metal was discovered in 1808 independently by Sir Humphry Davy and by J. J. Berzelius and Poutin. Its name derives from the Latin "calx," for lime. Calcium belongs to the alkaline earth elements. Though a metallic element, it is highly reactive and never found in its elemental form in nature. Pure Calcium metal is silvery white. It is oxidized easily and in air forms a white coating; ignited, the metal burns with a brilliant white light and a yellow-red flame.

Calcium is the fifth most abundant element in the Earth's crust and constitutes more than 3 % of the whole. Calcium is the fourth most abundant element in the lunar highlands. Some important, naturally occurring compounds are the carbonate limestone ($CaCO_3$), gypsum ($CaSO_4 \cdot 2\ H_2O$), fluorite (CaF_2), and complex silicates.

2. Physical and Chemical Properties

Calcium [*7440-70-2*], $_{20}Ca$, A_r 40.08; stable isotopes 40, 42, 43, 44, 46, 48. Its major properties are:

Density (20 °C)	1.55
Melting point	838 °C
Boiling point	1440 °C
Specific heat (0 – 100 °C)	0.624 J g^{-1} K^{-1}
Heat of fusion	217.7 J/g
Heat of vaporization	4187 J/g
Thermal expansion (0 – 400 °C)	22.3×10^{-6} K^{-1}
Electrical resistivity (0 °C)	3.91×10^{-6} Ω cm
Thermal conductivity (20 °C)	1.26 W cm^{-1} K^{-1}
Lattice constant (Fcc)	0.5582 nm

Calcium is relatively unstable in moist air, rapidly forming a hydration coating. It can be stored in dry air (less than 30 % R.H.) at room temperature. Calcium reacts spontaneously with water to form $Ca(OH)_2$ and hydrogen gas; when finely divided, it will ignite in air, burn at a high temperature and produce a nitride.

One of the alkaline earth group of metals, group 2, calcium exists in the face-centered cubic (fcc) form at room temperature, transforms into a body-centered cubic (bcc) structure at 448 °C, and melts at 838 °C.

The predominate stable isotope of calcium is ^{40}Ca. Calcium exhibits only one valence state, +2, in all of its reactions. It is slightly less reactive than barium and strontium in the same series. Calcium is a very ductile metal and can be formed by casting, extrusion, rolling, etc. Table 1 presents mechanical properties of calcium.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_515.pub2

Table 1. Mechanical properties of calcium metal

Mechanical properties	Annealed	Cold worked
Tensile strength, N/mm^2	4800	11500
Yield strength, N/mm^2	1370	8500
Elongation, %	51	7
Modulus of elasticity, MN/mm^2	2.2 – 2.6	
Hardness, Rockwell B	16 – 18	

Calcium is noted for its high reactivity, especially the high heat of formation of some of its compounds. This property enables calcium to be used as a reducing agent in the preparation of metals such as chromium, thorium, zirconium, and uranium. Examples are given in Table 2.

Table 2. Heats of formation ΔH_f of calcium compounds

Compound	CAS registry no.	ΔH_f, kJ/mol
$CaBr_2$	[*7789-41-5*]	– 675.3
$CaCl_2$	[*10043-52-4*]	– 795.5
CaF_2	[*7789-75-5*]	–1215.4
CaH_2	[*7789-78-8*]	– 188.8
CaI_2	[*10102-68-8*]	– 535.1
Ca_3N_2	[*12013-82-0*]	– 432.1
CaO	[*1305-78-8*]	– 636.0
CaO_2	[*1305-79-9*]	– 659.4
Ca_3P_2	[*1305-99-3*]	– 504.5
CaS	[*20548-54-3*]	– 482.7

The low density and relatively low electrical resistivity make calcium one of the most efficient electrical conductors on a mass basis. At 20 °C, calcium conducts 16.7 % more electricity than aluminum, and at 100 °C, it conducts 21.6 % more electricity through one centimeter length and one gram mass of the respective metal. Compared to copper, calcium conducts 150 % higher current density at 20 °C, and 197 % higher at 100 °C. Calcium metal is not used as a conductor on Earth simply because it reacts spontaneously when it comes in contact with oxygen, but in the vacuum environment of space, calcium becomes attractive.

3. Production

The electrowinning of metals, the electrochemical production of metals from their compounds, is the oldest industrial electrolytic process. Alkaline earth metals were first extracted by the English chemist Humphrey Davy, using electrolysis of molten salts of sodium hydroxide. In the electrolytic cell, calcium metal collects at the cathode. The metal obtained by this method is typically 98 % pure. By further subliming, it can be purified to levels beyond 99 %.

The major producers of calcium metal are China, approximately 10 000 – 12 000 t/a, primarily at the Shaanxi Zhonghe Special Metal Plant, Hanzhong Shaanxi; Russia, 6000 – 8000 t/a, at the Chepetsky Mechanical Plant, Glazov, Udmurt; the USA, 2000 – 4000 by Minteq International Inc. at Canaan, Connecticut; Canada, at Timminco Ltd; and France, at Pechiney. World capacity was approximately 24 000 t/a in 2005. The USA accounts for about half of the world's consumption of calcium. Total world capacity utilization is close to 80 %.

The process used in Russia is the electrolysis of molten salts of sodium hydroxide; the same process is believed to be used in China. While electrowinning is the more expensive process, it produces a higher purity metal when properly implemented.

In the USA and Canada calcium is produced by thermal reduction of lime with aluminum. The reactants, high-calcium lime of small particle size and aluminum powder, are compacted into briquettes to ensure intimate contact of the reactants. The briquettes are charged into a high-temperature alloy retort. The open ends of the retort protrude from the furnace and are cooled by water jackets to condense the calcium vapor. The retorts are then sealed and the reaction vessel is evacuated to 0.1 Pa or less; it is then heated to 1200 °C. The reaction proceeds for approximately 8 h, the vacuum is broken, and the condensed, 23 – 34 kg pieces of calcium metal, known as crowns, and calcium aluminate residue are removed The aluminum reduces the lime producing calcium metal vapor. The calcium then is removed from the reaction by condensation, thus allowing the reaction to continue in the desired direction. Limestone $CaCO_3$ is quarried and calcined to form calcium oxide. The calcium oxide is then ground and dry-blended with the desired amount of finely divided aluminum. This mixture is compacted into briquettes for reactants then placed in horizontal tubes, i.e., retorts, made of heat-resistant steel and heated to 1100 – 1200 °C at a pressure less than 13 Pa. The commercial-grade metal resulting from this process is suitable for most industrial applications such as steelmaking.

A further vacuum distillation step is required to obtain high-purity calcium metal, used in more specialized refining applications such as the production of neodymium for magnetic alloys. This second operation reduces the level of contaminants. Commercial-grade calcium is placed at the bottom of a large vertical retort made of heat-resistant steel; the vessel has a water-cooled condenser at the top. The retort is sealed and evacuated to a pressure of less than 6.6 Pa; the bottom is heated to 900 – 925 °C. Calcium distills to the condensing section and leaves behind less volatile impurities. Table 3 shows the purity after the first and second purification steps.

Table 3. Chemical analysis of typical commercial and redistilled calcium

Element	Commercial grade, wt %	Redistilled grade, wt %
Mg	0.50	0.50
N	0.08	0.02
Al	0.30	0.001
Fe	0.008	0.001
Mn	0.01	0.002
Co	n.a.	0.0002
Li	n.a.	0.0001
Be	n.a.	0.0001
Cr	n.a.	0.0002
B	n.a.	0.0001
Ca and Mg	99.5	99.9

4. Uses

Steelmaking. The major use of calcium is to improve the quality of steel (→ Steel). For decades, calcium-containing ferroalloys have been used as tap-stream additions to the molten metal, or calcium compounds were injected through a refractory lance by using argon as carrier gas. Although the benefits of using pure calcium metal were known, they were difficult to obtain; calcium is highly volatile, boiling well below steelmaking temperatures.

Calcium is important in steel chemistry because it is a strong oxide and sulfide former; furthermore, it has the uncommon ability to alter the oxides and sulfides to give liquid lime aluminates and underformed sulfides. Treatment with calcium modifies the melting point of inclusions, which rapidly float out of the steel; in addition, treatment alters the morphology of any remaining inclusions, rendering them spherical in shape, very small, and finely dispersed. The result is a fundamental quality improvement, especially in castability and the mechanical properties: formability, drawing, impact, tensile, machinability, resistance to cracking and tearing, and improved surface and internal cleanliness. Calcium also improves resistance to hydrogen-induced cracking in line pipe associated with high-sulfur oil and gas pipelines.

In the early 1970s, calcium wire feeding technology was introduced. A steel-clad calcium wire is fed through a delivery system which propels the wire well below the surface of the molten steel. The steel cladding protects the calcium until it reaches a depth where the ferrostatic pressure suppresses vaporization.

The use of metallic calcium to refine steel became common in Japan and the USA in the 1980s. Cored-wire injection technology was applied in ladle metallurgy, tundish, and mold injection for continuous casting. This led to the adoption of calcium injection for improved castability and superior product quality in thin slab casting. This casting technology, which can produce continuous 50 – 70 mm steel slabs, was developed by SMS and MDH in Germany during the mid-1980. Nucor, USA, established the first commercial-scale thin-slab caster in 1989.

Thin-slab casting employs a specially designed nozzle that is susceptible to alumina buildup, develops progressive blockage, serious clogging, and leads to termination of the casting process. Calcium has proved the most effective method to prevent these problems. With appropriate sequential process steps, treatment with calcium allows inclusion-shape control and gives much cleaner steel.

For large-tonnage use and ladle treatment, a wire-lance system combines the advantages of wire feeding, gas control of fluid dynamics, and treatment with pure calcium. This has resulted in the efficient production of improved-quality ultraclean steels.

Maintenance-Free Automotive Battery. In the maintenance-free automotive battery (→ Batteries, Chap. 4.2.2), calcium improves electrical performance and battery life. The antimony – lead alloy used in the conventional lead – acid battery is replaced with a 0.1 % Ca – Pb grid alloy. The calcium in "calcium batteries" refers to the lead alloy used in the production of the grids for plate making. The alloy is typ-

ically still 99 % lead with alloying metals. Calcium alloy grids afford low water loss and low self-discharge.

Calcium improves the conductivity and current capability of the cell; it significantly reduces gassing, permitting the cell to be closed, thus preventing water loss and extending life. The hydrogen overvoltage increases with calcium – lead alloy grids, so at a given rate of overcharge at the same temperature, an antimonial battery will consume more water than a calcium battery. The resulting battery will not require water addition, i.e., maintenance, during its operating life under normal operating conditions. Their lower internal resistance provides a higher charge-rate acceptance and a small increase in performance. The life and performance of calcium – lead batteries is enhanced by tight control of the manufacturing process and the specifications of the grid alloy for the positive plate.

Calcium – lead alloy grids were prone to expansion and cracking that lead to premature battery failure. The use of calcium alloys and roll-hardening the positive grid reduced the problem. A calcium – aluminum alloy forms a eutectic with lead, and rapid solidification of the alloy makes continuous grid making possible.

Lead – calcium alloys have replaced lead – antimony alloys in a number of casting applications. These alloys contain 0.03 – 0.15 % Ca. Aluminum is added to calcium – lead and calcium – tin – lead alloys as a stabilizer for calcium. Adding tin to lead or lead alloys increases hardness and strength, but lead – tin alloys are more commonly used in type metals and solders for their good melting, casting, and wetting properties. Tin gives the alloy the ability to wet and bond with metals such as steel and copper; unalloyed lead has poor wetting characteristics.

Magnet Materials. High-energy-density magnetic materials are produced using calcium (→ Magnetic Materials). Samarium – cobalt magnets with energy products of between 8 and 16×10^4 T·A/m have found applications in miniature transducers and other devices requiring high energy or volume restrictions. The reaction is:

$$3\,Sm_2O_3 + 10\,Co_3O_4 + 49\,Ca\ (g) \rightarrow 6\,SmCo_5 + 49\,CaO$$

The production of neodymium – iron – boron magnets involves the use of calcium metal as a reductant. This is an important development because these magnets have energy products approaching 40×10^4 T·A/m, over twice that of samarium – cobalt. In addition, the availability of the raw materials is significantly greater than for samarium – cobalt; both the quantity and stability of raw material supply limited the growth of samarium – cobalt magnets. Therefore, neodymium – iron – boron magnet materials are less expensive to produce.

The energy density is sufficient to permit the replacement of armature windings in motors, transducers, and generators with permanent magnets. The weight of an automotive starting motor was reduced from 3.6 to 1.8 kg; the size and cost of the motor were commensurately reduced as well. This technology is finding extensive applications in automotive, computer peripheral, medical, home appliance, and military markets.

There are two methods of producing the neodymium – iron – boron raw material, a neodymium – iron alloy. One is the calcium thermal reduction method:

$$NdF_3 \text{ or } NdCl_3 + Ca + Fe \rightarrow NdFe + \text{byproducts}$$

or

$$Nd_2O_3 + Ca + Fe \rightarrow NdFe + \text{byproducts}$$

The second method is by electrolysis:

$$NdCl_2 + \text{electricity} \rightarrow Nd_2 + Cl_{2(v)}$$

or

$$NdF_3 + Nd_2O_3 + \text{electricity} \rightarrow Nd + \text{byproducts}$$

The calcium thermal reduction process is preferred. It has the greatest flexibility in producing the basic neodymium – iron and alloy variations. The byproducts are not toxic as they are in the electrolytic process. The calcium process is less capital intensive and readily scaled in production volume to meet market demands.

Iron Making. Injection of calcium wire has been adopted by many foundries for iron treatment. Again, the calcium is used for desulfurization, inoculation spherodizing, and alloying.

Calcium ferroalloys are used in the production of nodular iron castings. In magnesium ferrosilicon, calcium reduces the reactivity, enhances nucleation, and improves morphology. The ratio of calcium to magnesium varies from

0.15 to 0.50. Pieces of the ferroalloy are placed in a protected pocket cut in the refractory lining of the ladle prior to tap. The molten iron is then poured into the ladle where it reacts with the alloy. The treated, nodularized iron is then cast into molds.

In the in-mold process, granularized ferroalloy is placed in a special reaction chamber formed in the channels of the mold. This permits the reaction to occur when the iron is cast; it enhances the effectivity of the ferroalloy and results in improved castings. The process can be automated for high productivity.

Other Applications of Calcium and Calcium Compounds. Alloys of calcium also are used to deoxidize magnesium, to strengthen lead electrodes, and to produce special aluminum alloys. Calcium is also used to improve the mechanical properties of lead (→ Lead Alloys). A proprietary process for improving the integrity and formability of the lead used in yacht keels was developed in Australia.

In the *debismuthizing of lead* by the Kroll-Betterton process, calcium metal is combined with bismuth which then floats out in a dross:

$$3\,Pb_3Ca + 2\,Bi \rightarrow Bi_2Ca_3 + 9\,Pb$$

Lead ores are thus refined to commercial soft lead with 0.02 % or less Bi (→ Lead).

Calcium metal is reacted with zirconium fluoride to refine *zirconium*; the high heat of reaction melts the zirconium. The zirconium ingot produced by this method is remelted under vacuum for purification. The resulting metal is leached with acetic acid to remove the lime produced as a byproduct of the reaction.

Uranium fluoride or oxide can be reduced with calcium. To produce *thorium* and *uranium*, the oxides are mixed with a stoichiometric excess of calcium and reduced under an atmosphere of argon. Auergesellschaft during World War II made the uranium metal for the atomic bomb project by using metallic calcium to reduce uranium oxide to uranium metal. In 1945 the Russians dismantled and took the calcium plant to Russia for refining uranium. Calcium produced in China was used in obtaining nuclear materials as well.

Calcium metal is readily hydrided for use as a portable source of *hydrogen gas*. It also is used in the production of the B-complex vitamin calcium pantothenate [*137-08-6*].

Precipitated calcium carbonate is used to replace wood pulp in paper and as an antacid.

Many calcium compounds are used in the foodstuffs, pharmaceuticals and medical industries. Calcium lactate $Ca(C_3H_5O_3)_2 \cdot 5\ H_2O$, calcium diphosphate (phosphate of lime), $CaHPO_4 \cdot H_2O$, and tricalcium phosphate $Ca_3(PO_4)_2$ are used to supply calcium and phosphorus to foods. The last-named also finds application as a polishing agent in dentifrices and as an antacid. Calcium lactobionate, a white powder, provides a suspending agent in pharmaceuticals.

Calcium monophosphate, $CaH_4P_2O_8 \cdot H_2O$ is used as a leavening agent in baking. Anhydrous monocalcium phosphate $CaH_4(PO_4)_2$ is used in pre-mixed flour mixes.

Calcium sulfite $CaSO_3 \cdot 2\ H_2O$ provides a bleach for paper pulp and textiles. As a bleach, it is a disinfectant.

Calcium silicate $CaO \cdot SiO_2$ is used as a reinforcing agent in rubber. This compound also helps manipulate the viscosity of liquids, commercially as a filler in paints and coatings. In coatings it modifies the reflectivity resulting in reduce sheen, a matte surface. Calcium metasilicate $CaO \cdot SiO_3$ (wollastonite) is used as a brightener for paints and paper coatings, a filler for plastics, a coating for welding rods, and for electrical insulators, tile, and ceramics.

Calcium acetate $Ca(C_2H_3O_2) \cdot 2\ H_2O$, (known as lime acetate and vinegar salts) is employed in liming rosin and the manufacture of metallic soaps and synthetic resins.

Limestone is an essential element in Portland cements. Oyster shells ($CaCO_3$) are a commercially important source of calcium for animal feeds.

5. Occupational Health and Safety

Calcium metal reacts with water and acids to form hydrogen gas, calcium oxide and calcium hydroxide; the reaction is exothermic. If calcium metal contacts moisture in the eyes, on the skin, or in the respiratory tract, severe corrosive irritation may result. Inhalation of dust or fume may cause severe respiratory irritation, cough, difficulty in breathing, and chemical pneumonitis. Contact with skin causes irritation and possible corrosion damage.

The substance is severely irritating to the eyes and may injure eye tissue if not promptly removed.

Ingestion may cause acute irritation or burns to the mouth, throat, and stomach; calcium may cause vomiting. Pre-existing chronic respiratory, skin, or eye diseases may be aggravated.

The symptoms of inhalation include severe irritation of respiratory tract. Skin and eye contact symptoms are severe irritation.

There are no known adverse health effects resulting from long-term exposure to calcium metal

First Aid. *Inhalation:* Remove to fresh air. If breathing is difficult, give oxygen. Seek immediate medical attention.

Skin contact: Remove calcium metal immediately with a dry cloth. Wash the area with large amounts of water until all the chemical is removed.

Eye contact: Immediately flush eyes with running water for 15 min. Lift upper and lower eyelids occasionally. Get immediately medical attention.

Ingestion: If victim is conscious, give large amounts of water to dilute the alkali. Do not induce vomiting. Never give anything by mouth to an unconscious person.

Note to physician: All treatment should be based on observed signs and symptoms of distress in the patient. Consideration should be given to the possibility that overexposure to materials other than calcium metal may have occurred.

Firefighting Measures. *Extinguishing method:* Do not use water, foam, or halogenated hydrocarbons such as Halon or carbon tetrachloride to extinguish fire. Use only graphite powder, soda ash, powdered sodium chloride, or an appropriate metal-fire-extinguishing dry powder, such as Met-L-X. For large fires, withdraw from the area and let the fire burn.

Firefighting procedures: Firefighters should wear self-contained breathing apparatus with full face piece operated in the pressure-demand or positive-pressure mode. Firefighters should move containers from the fire area if this can be done without risk. Do not use water or foam. Use dry powder only.

Fire and explosion hazards: Water reacts dangerously with calcium metal and is not recommended as an extinguishing agent for fires that involve it. If water must be used, prevent it from coming into direct contact with calcium metal. If contact is unavoidable, apply the water in flooding amounts to safely absorb the heat that will be generated.

Hazard: Calcium metal is extremely dangerous when wet. Calcium metal forms calcium hydroxide and hydrogen gas resulting in an explosion hazard when wet. Calcium metal forms CaO, quicklime, when it burns. It reacts with wet extinguishing agents such as water, halogens, and possibly carbon dioxide.

Accidental Release Measures. Do not touch spilled calcium metal. Wear protective apparel. Do not smoke or place flame or ignition sources near spill area. Do not allow water to touch spilled calcium metal or to get inside containers. Use a cover, a plastic sheet for example, to prevent water or rain from dissolving spilled calcium metal or to prevent its spreading. Isolate hazard area and keep nonessential personnel away from spill or leak site. Shovel small dry spills into a dry container and cover it tightly. Move containers away from spill to a safe area. Take up small spills with sand or an absorbent and contain as described above. Dike the flow of large calcium metal and water spills with soil, sandbags, or concrete. Keep the waste from entering drains and open sewers. Wear full protective gear.

Handling and Storage. *Storage:* Store calcium metal in a sealed container away from water, acids, or organic compounds. Protect containers against physical damage. Calcium metal will generate heat upon contacting water.

Handling: Avoid damaging container.

Exposure Controls and Personal Protection. *Ventilation:* Provide adequate exhaust ventilation to meet exposure limit requirements. An exhaust filter system may be required to avoid environmental contamination.

Respiratory protection: Calcium metal does not have established exposure limits. Wear a positive pressure air-supplied respirator in situations where there may be a potential for airborne exposure.

Hand protection: Impervious gloves.

Eye protection: Chemical goggles and/or face shield.

Other protection: Wear rubber apron or other impervious clothing to prevent contact with skin.

6. Biological Relevance

6.1. Introduction

Calcium is indispensable for life; it is required for maintenance of structure and metabolism. Examples of the structural role of calcium include the elaborate structure of coral [8], the secreted shells of mollusks, and the internal skeletons of vertebrates. Calcium is also necessary for many biological functions, the most prominent being muscle contraction, nerve conduction, and regulation of blood clotting. To control these functions, all organisms have mechanisms for maintaining critical intracellular calcium levels. Organisms with a circulatory system also have mechanisms to control extra-cellular calcium levels as well as calcium uptake and excretion. The role of calcium in biological systems has been the subject of a number of books [9, 10]. This chapter focuses on the molecular, metabolic and structural aspects of calcium in vertebrate biology.

6.2. Chemical Properties of Ca Important in Biological Functions

Calcium has two properties that render it indispensable for life. First, Ca ions complex with organic compounds, particularly proteins, thereby affecting their function. Second, Ca salts display a wide range of solubility, making it possible for vertebrates to possess a skeleton.

Calcium – Protein/Lipid Interactions. Calcium mediates many biological functions through its ability to bind specifically to proteins and lipids. Calcium binds to proteins by three principal mechanisms. First, it can directly bind the carboxyl groups of glutamic and aspartic acid residues to alter enzyme function. Second, it can interact with specific phosphorylated serine, tyrosine, or threonine residues. Third, specific amino acids that have undergone vitamin K-dependent γ-carboxylation effectively chelate Ca. Proteins may use one or more of these mechanisms to bind Ca. For example, the digestive protein trypsin has a locus of glutamic and aspartic residues that bind Ca, and this changes the conformation of the enzyme to enhance its activity [11, 12]. Osteocalcin (a bone matrix protein) possesses a sequence with three γ-carboxyglutamic acid residues that enable it to bind Ca minerals [13]. The bone matrix proteins osteopontin and bone sialoprotein are rich in aspartic acid, glutamic acid, and phosphorylated serine and/or threonine residues, which permit them to bind calcium phosphate crystals [14 – 18]. Calcium-dependent enzymes can be activated by two principal mechanisms. Many enzymes are activated by binding Ca directly. Other calcium-dependent enzymes do not bind Ca, but are activated by noncovalent association with calcium-binding proteins. Calcium-binding proteins which fulfill this co-enzyme function include calmodulin, annexin, and calcineurin.

To optimize the ability of proteins to bind calcium, the relevant amino acid residues are localized in specific domains which may be conserved between species and throughout evolution. An example domain is known as "EF-hand", which is a single helix – loop – helix structure that binds Ca within the 12-amino acid loop [19]. The proteins possessing EF hand motifs have undergone significant and turbulent molecular evolution. Starting as an archetypal protein containing one EF hand, the genes coding these proteins have undergone duplications, deletions, and retroversions to yield a family of calcium-binding proteins with a wide array of functions and a variable number of EF hands [19 – 25]. A representative calcium-binding protein of this type is calmodulin, which has four EF hands that recapitulate the canonical EF hand sequence observed in slime molds, paramecium, plants, and mammals, including humans [22, 26]. Calcium also complexes with cell membrane phospholipids, which can anchor cell surface-associated proteins on the basis of electrostatic interaction between Ca ion and the negatively charged head groups of cell membrane phospholipids.

Solubility Range of Calcium Salts. Primary calcium phosphate $Ca(H_2PO_4)_2$ is extremely soluble. Secondary calcium phosphate $CaHPO_4$ has a solubility of about 2.0 mM and makes up

about 85 % of the Ca in extracellular fluid. Tertiary calcium phosphate $Ca_3(PO_4)_2$ has a solubility of about 100 μM, and most of it is found in bone in the form of hydroxyapatite [27]. Ultimately, the control of Ca crystallization is determined by these different solubilities and by biological factors, such as calcium-binding and phospholipid vesicles, which can regulate calcium crystal formation.

6.3. Dietary Considerations; Distribution in the Body

The age-adjusted daily calcium requirement is as follows: 0 – 0.5 years, 400 mg; 0.5 – 1.0 years, 600 mg; 1 – 10 years, 800 mg; 11 – 24 years, 1200 mg; 25 and above, 800 mg. The bulk of dietary calcium is derived from dairy and grain products (76 %), the remainder coming from vegetables, protein foods, fruit, sugar, fats, and oils [28]. Lactose enhances the calcium absorption, but fiber and fat, and certain plant constituents reduce its bioavailability. The use of calcium supplements to compensate for poor dietary intake is controversial, since uptake by the gut is highly regulated. Furthermore, the absorption of Ca supplements is dependent on their solubility: the citrate, maleate and lactate salts have high bioavailability, whereas the oxalate is poorly absorbed [28].

Phosphorus is an essential dietary ingredient and a major bone mineral component. Phosphorus is present in most foods, and most adults consume greater than 100 % of required daily allowances [29].

Vitamin D is essential to normal Ca homeostasis and is derived from dietary sources and from *de novo* synthesis in peripheral tissues. The required daily allowance for vitamin D is 10 mg/d; it is found in fortified milk, cod liver oil, and certain fish [30].

The adult human contains 1.0 – 1.3 kg of Ca, about 99 % of which is located in the skeleton, 0.5 % in the teeth, and the remainder in soft tissues and extracellular fluids. Within the plasma compartment, 48 % is found in the free (ionized) form, 48 % is bound to plasma proteins, and 3 % is complexed as citrate, phosphate, or bicarbonate. A simplified schematic of whole-body Ca flux is shown in Figure 1.

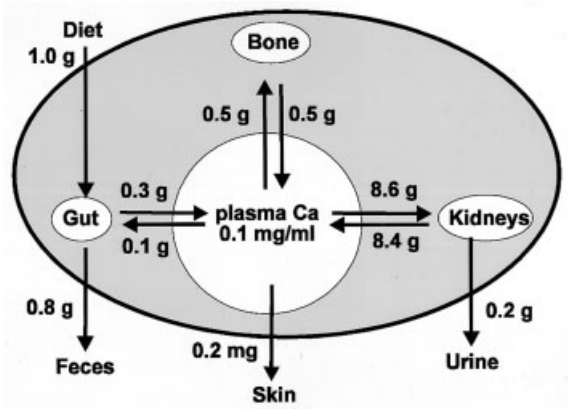

Figure 1. Whole body distribution of calcium [31]
Plasma calcium pool is in equilibrium with Ca input (from the intestine, bone and kidney) and Ca output (urine, sweat and feces). There is a steady-state balance in Ca flux to and from bone. While a significant amount of Ca is filtered out by the kidney, most is reabsorbed, and the net loss is usually compensated for by intestinal absorption

6.4. Regulation of Plasma Calcium by Hormones

Cells and tissues have developed mechanisms for compartmentalizing Ca within extracellular and intracellular pools, and can regulate the flux of Ca between these pools very precisely. Hormonal regulation of gastrointestinal absorption, renal absorption, and bone mineral Ca flux provides a system through which plasma calcium levels remain within narrow limits. Three major hormones regulate Ca flux from the gut, kidney and bone.

Parathyroid hormone (PTH) is a major short-term regulator of plasma calcium level and is synthesized by the parathyroid gland. Cell surface receptors for Ca in the parathyroid gland are activated when plasma calcium level falls, resulting in PTH secretion [32]. PTH is also secreted in response to catacholamines, vitamin D, and calcitonin. The net effect of PTH is to increase plasma calcium level by interacting with target cells in the kidney, gut, and bone to stimulate Ca entry into the plasma pool [33].

Vitamin D is responsible for long-term maintenance of plasma calcium level. It is derived from the diet or generated *de novo* form cholesterol following exposure of the skin to ultraviolet

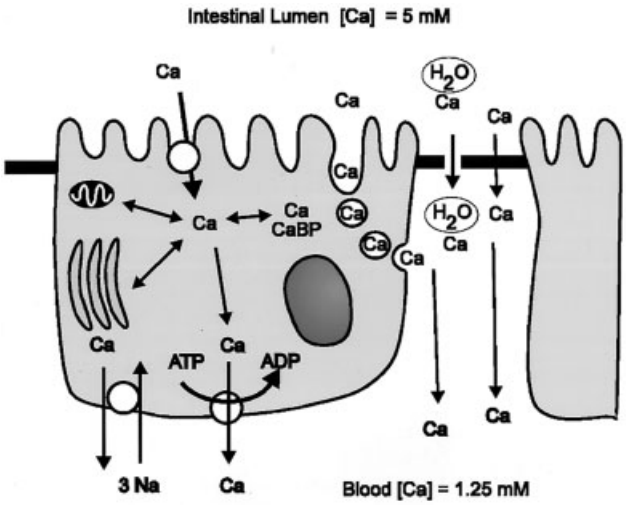

Figure 2. Intestinal cell calcium transport
Two mechanisms allow for transcellular Ca transport. First, Ca is transported through the apical surface by a Ca transport protein (calbindin D_{28K}). It is then transported through the cell bound to Ca-binding proteins, and is extruded into extracellular milieu along basolateral cell surface by vitamin D-inducible ATP-dependend Ca pump in association with Na/Ca exchanger and Na/K pump [37]. Second, calibindin-D_{28K} immobilizes Ca on the cell membrane, which invaginates to form vesicles that carry Ca to basolateral surface. Here, the vesicles fuse with the plasma membrane, discharging Ca into the extracellular milieu [38]. Two processes also allow for paracellular Ca transport. Ca flows passively down a concentration gradient (from 0.5 to 1.25 mM) through structures between intestinal cells called tight junctions that are semipermeable to Ca. In addition, Ca can also be "dragged" by solvent water. By the latter two passive processes, Ca is absorbed, albeit with low efficiency [39].

light [34]. Vitamin D undergoes hydroxylation in the liver to form 25-(OH)-D and then by the kidney to 1, 25-$(OH)_2$-D, which is the major active metabolite regulating Ca homeostasis. Circulating vitamin D_2 has a half-life of about 1.5 d. The major target organs of vitamin D_2 include the gut, kidney, and bone [35].

Calcitonin is a short 32 amino acid peptide secreted by specialized cells in the thyroid gland. Calcitonin secretion is stimulated (or inhibited) by acute increases (or decreases) in plasma calcium concentration. Its plasma half-life is about 10 min. Calcitonin inhibits bone resorption, thus antagonizing the effect of PTH. The role of calcitonin in regulating plasma calcium levels is minor relative to PTH and vitamin D [36].

6.5. Control of Plasma Calcium by Target Organs

In the *intestine*, Ca is absorbed as the free ion with an efficiency of approximately 30 – 32 %. Calcium traverses the intestinal lining either by transcellular flux, which is mediated by calcium-transport proteins or vesicular transport, or a second, passive flux between intestinal cells, known as paracellular flux (Fig. 2).

The *kidneys* are the principal regulators of plasma calcium level, and filter about 10 g of Ca per day. In the renal glomeruli, Ca is filtered out of the blood into the proximal tubule; however, 70 % is reabsorbed in this portion of the kidney by paracellular transport. In addition, about 20 % of the filtered Ca is reabsorbed in the thick ascending limb of Henle by paracellular transport and calcitonin-induced active transport. The remaining Ca is absorbed by the distal convoluted tubes in a PTH- and vitamin D-dependent manner due to enhanced expression of calbindin-D_{28K}. The urinary filtrate thus contains approximately 1 % of the filtered Ca load [39].

Bone contributes to the steady-state plasma calcium level, which is regulated by PTH and vitamin D. PTH promotes bone resorption by activating cells within the bone matrix (osteoclasts) committed to matrix destruction, allowing Ca access to plasma [29]. However, PTH also promotes bone formation, probably due to its action on cells lining bone (osteocytes). Vitamin D has multiple effects on Ca trafficking: it enhances

deposition of Ca into bone by stimulating the maturation of primitive bone cells (osteoblasts) to osteocytes, but also acts in concert with PTH to stimulate osteoclast activity. Thus, PTH and vitamin D act together to increase the rate of bone and mineral turnover without altering net bone mineral content or mass [40].

6.6. Physiology of Intracellular Calcium Trafficking

The optimum intracellular calcium concentration is 0.1 μM, which is 10 000 times less than the extracellular calcium level. Accordingly, cells developed mechanisms to control Ca flux between the extra- and intracellular pools. The major restriction of extracellular Ca to the intracellular milieu is the plasma membrane which has a low passive permeability to Ca. However, the plasma membrane also has Ca channels which allow Ca entry in response to changes in voltage, stretching, or hormonal stimulation [41]. At least one type of channel is tonically active and permits Ca to the extracellular milieu, as described above [42]. In addition, an intracellular tubular structure known as the smooth endoplasmic reticulum (SER) serves as a source of intracellular Ca. A membrane-bound Ca/ATPase transports cytoplasmic Ca into the SER, which lowers cytoplasmic calcium level. In contrast, release of calcium from the SER is mediated by two receptors, one of which is responsive to hormonal stimulation, while the other is activated by the Ca itself [43]. The cytoplasmic calcium level is thus a reflection of the state of activation of the cell. Collectively, calcium-dependent enzymes control nearly every aspect of cell function, including nucleotide metabolism, glycogen metabolism, electrolyte balance, neuromuscular irritability, neural function, muscle contraction, and proliferation.

6.7. Physiology of Bone

Bone is a connective tissue that has the ability to become calcified. Structurally, bone consists of hydroxyapatite deposited within an organic matrix. Hydroxyapatite is hard, rigid, and brittle, and is effective mainly in resistance to compression. However, the organic matrix, over 90 % of which is collagen type I, is flexible and tension resistant. Thus, bone has the tensile strength of cast iron, but has only one-third the weight and is ten times as flexible. Bone is organized into two general types of structures. Compact bone which forms the shafts of the long bones and the outer surface of nearly all other bones, is highly calcified (80 – 90 %). Trabecular bone forms the ends of the shafts of the long bones and most of the bones in the spine, and is 15 – 20 % calcified [44].

Cellular Basis of Bone Formation and Turnover. Bone undergoes continuous remodeling due to the activity of bone-forming cells (osteoblasts, osteocytes), and bone-resorbing cells (osteoclasts). Osteoblasts secrete connective tissue proteins to form unmineralized bone matrix. Mature osteoblasts, known as osteocytes, lie within small cavities that are completely surrounded by bone, and are responsible for bone deposition. These cells are activated by vitamin D, calcitonin, platelet-derived growth factor, transforming growth factor- β, insulin-like growth factor, androgens, and low physiological amounts of glucocorticoids [45]. In contrast, osteoclasts degrade bone, and are activated by PTH, vitamin D, interleukin-1, tumor necrosis factor, and prostaglandin E_1. Conversely, osteoclast activity is inhibited by estrogens [46].

There are two general mechanisms by which bone is laid down. In the first, bone formation occurs in the absence of a pre-existing structure. Primordial cells differentiate directly into osteoblasts, which secrete soluble collagen into the extracellular space. The soluble collagen then self-assembles to form fibrils, which are stabilized by extensive intermolecular cross-linking. Initially, the bone matrix is randomly oriented (or woven), and hydroxyapatite deposition occurs in irregular patches by the formation of matrix vesicles as a result of exocytosis from the plasma membrane of osteoblasts. Mineralization occurs on these vesicles. and as the crystals grow, the vesicle is subsequently destroyed. Bones which are formed by this method include flat bones, such as the skull, jaw, and ribs. The second method of bone formation occurs on a previously existing cartilaginous structure as an initial supporting structure which undergoes initial calcification. Remodeling of this calcified matrix by osteoclasts and osteoblasts eventually

replace the calcified cartilage in the formation first of woven bone, and then compact bone. Mineralization in compact bone occurs within the tightly spaced organic matrix within vesicle formation. This process occurs in the development of the long bones of the legs and arms [47].

Remodeling of bone occurs throughout life. During growth, the midshaft of long bones undergoes growth on the outer surface, and resorption at the inner surface to maintain net bone thickness. At puberty, the growth plates present on both ends of long bone close and prevent further bone lengthening. As bone ages, the organic matrix becomes brittle and weak, and undergoes microfractures. Bone remodeling continuously replaces old matrix with new matrix in a steady-state process of osteoclastic destruction of damaged bone and generation of new bone by osteoblasts. This aspect of bone physiology is the subject of numerous books [44, 48].

Mineralization of Bone. The mineral in bone is hydroxyapatite, and makes approximately 60 – 65 wt % of bone. For the hexagonal form of hydroxyapatite, the content of the unit cell is $Ca_{10}(PO_4)_6(OH)_2$. *In vitro*, hydroxyapatite can be generated by mixing $CaCl_2$ with KPO_4. An unstable amorphous phase (ACP1) is initially formed, followed by the appearance of a second amorphous phase (ACP2), which is less soluble than ACP1. ACP2 then transforms to $Ca_4H(PO_4)_3$, which polymerizes into deficient hydroxyapatite [49, 50]. Other ions can influence hydroxyapatite formation. Fluoride generates two crystal forms: hydroxyapatite covered with fluorapatite, and fluorapatite covered with hydroxyapatite [51]. In contrast, magnesium, carbonate, and cadmium inhibit hydroxyapatite formation by substitution into the apatite crystals, preventing further crystallization. The inhibition of crystal growth by magnesium can be attenuated by fluoride ion [52, 53]. In vivo, hydroxyapatite crystal formation and attachment to collagen is regulated by calcium-binding proteins. Osteopontin and bone sialoprotein initiate and regulate crystal growth and mediate the turnover of bone in areas of growth [54, 55]. In compact bone, hydroxyapatite is maintained by osteocalcin [14, 56, 57] and osteonectin [58 – 60].

6.8. Disturbances in Plasma Calcium Homeostasis

Hypercalcemia. Excessive calcium intake may cause increased calcium levels. However, this is rare, since intestinal paracellular Ca transport is rather inefficient, and at least 2 – 3 grams of additional Ca must be consumed to cause toxicity. Hypercalcemia is usually due to excessive intake of vitamin D, increased PTH secretion (primary hyperparathyroidism), or a consequence of bone destruction secondary to tumor metastasis to bone. Hypercalcemia is manifested by deposition of Ca salts into the heart, blood vessels, kidneys, and skeletal muscles, anorexia, nausea, vomiting, memory loss, confusion, muscle weakness, increased urine excression, and dehydration, and metabolic bone disease (see below).

Hypocalcemia. may be due to inadequate dietary intake of Ca, vitamin D, or insufficient sunlight. The most common endocrine cause is inadequate or nonexistent secretion of PTH (primary hypoparathyroidism). Hypocalcemia is also caused by defects of the PTH receptor in the target cells and is characterized by a resistance to peripheral tissue to PTH rather than by a low amount of active hormone (pseudohyporparathyroidism). The features of hypocalcemia include neuromuscular excitability, which may result in tetany and cardiac conduction defects.

6.9. Diseases of Bone

The strength of bone depends in part on the molecular structure and organization of its constituent mineral crystals in their organic matrix [61]. Most common diseases of bone cause increased propensity for fracture, and are due to a failure in the ability to generate adequate matrix or hydroxyapatite.

Osteoporosis is characterized by a reduction in mineral content per unit volume of bone and is caused by reduced 17- β-estradiol secretion (Type I) or a variety of mechanisms including deficiencies in calcium, vitamins C, D or K, hypogonadism, hyperthyroidism, glucocorticoid excess, rheumatoid arthritis, and immobilization (Type II) [62]. Treatment of osteoporosis includes Ca supplements, vitamin D and bis-

phosphates.[63]. Sodium fluoride is no longer recommended for patients with osteoporosis.

As well as its beneficial effects on the skeleton, calcium supplements may favorably affect serum lipids in postmenopausal women and older men. In addition, there is evidence that calcium consumption is inversely associated with cardiovascular disease in postmenopausal women. In the past, estrogen replacement was considered a primary therapy for the prevention of postmenopausal osteoporosis. Estrogen had the additional advantages of controlling menopausal symptoms and presumptive prevention or delay of cardiovascular disease. However, data now indicates that estrogen-progestin therapy does not reduce the risk of coronary heart disease, and increases the risk of breast cancer, stroke, and venous thromboembolic events.

Osteomalacia is characterized by loss of bone mineral Ca and connective tissue matrix as a result of calcium or vitamin D deficiency. The bony matrix is replaced by poorly ossified fibrous tissue.

Osteogenesis imperfecta is a genetic disease due to defects in the synthesis of Type I collagen, resulting in impaired fibril formation and ossification.

Pagets' disease of bone is caused by osteoclastic hyperactivity, resulting in widespread bone resorption.

Osteopetrosis is characterized by excessive bone ossification due to the relative absence of osteoclasts. Narrowing and obliteration of bone marrow reduces bone distensibility.

Metastatic Calcification. Calcification also occurs in soft tissues, and may have serious consequences. A major cause of soft tissue calcification is a long-standing hypercalcemia. Aberrant calcification can also occur as a consequence of cell death and necrosis. Atherosclerosis is also accompanied by dystrophic calcification, and may be due to the ability of plaque-associated lipids (cholesterol and phospholipids) [64, 65] or bone-related proteins [66, 67] to nucleate hydroxyapatite. Calcification of the vascular wall causes loss of elasticity and disruption of laminar blood flow, which can lead to plaque rupture and thrombosis.

7. References

General References

1. *Metals Handbook*, 9th ed., vol. 1. American Society for Metals, Metals Park, Ohio, 1978.
2. J. W. Robison: "Ladle and Mold Treatments with Steel-Clad Metallic Calcium Wire," *Scaninject III*, MEFOS, Lulea, Sweden, 1983, pp. 35.1 – 35.23.
3. D. C. Brown: "High Quality Strand Cast Steel," *33 Metal Producing* 1982 (Sept.) 62 – 66.
4. R. D. Prengaman: *Advanced Lead Alloys for Maintenance-Free Batteries*, Lead Development Assoc., London, England, 1980.
5. T. Ototani, N. Grant, *Calcium Clean Steel*, Springer-Verlag, Berlin, Heidelberg, 1986.
6. S. E. Hluchan, *The Future in Calcium Treatment — Wire Lance System*, AISE, PA 1980.
7. S. E. Hluchan: *Calcium Alloys for Battery Grid Production*, Lead Development Assoc., London England 1983.
8. D. M. Roy, S. K. Linnehan, *Nature* **247** (1974) 220.
9. J. B. Anderson, S. C. Garner (eds.): *Calcium and Phosphorus in Health and Disease*, CRC Press, New York 1995.
10. *Calcium*, Springer Verlag, New York 1988.
11. H. D. Bartunik, L. J. Summers, H. H. Bartsch, *J. Mol. Biol.* **210** (1989) 813.
12. E. Chiancone, T. Drakenberg, O. Teleman, *J. Mol. Biol.* **185** (1985) 201.
13. N. Nakao, Y. Nishiucjhi, T. Kimure, S. Sakaibara, *Pept. Res.* **7** (1994) 171.
14. A. L. Boskey et al., *Bone Miner.* **22** (1993) 147.
15. G. K. Hunter, C. L. Kyle, *Biochem. J.* **300** (1994) 723.
16. J. Sodek et al., *Ann. N. Y. Acad. Sci.* **760** (1995) 223.
17. G. K. Hunter, H. A. Goldberg, *Biochem. J.* **302** (1994) 175.
18. G. K. Hunter, H. A. Goldberg, *Proc. Natl. Acad. Sci. U.S.A.* **90** (1993) 8562.
19. S. Nakayama, R. H. Kretsinger, *Annu. Rev. Biophys. Biomol. Struct.* **23** (1994) 473.
20. S. Nakayama, R. H. Kretsinger, *Mol. Biol. Evol.* **10** (1993) 221.
21. S. Nakayama, N. D. Moncrief, R. H. Kretsinger, *Biochim. Biophys. Acta* **1160** (1992) 8.
22. S. Nakayama, R. H. Kretsinger, *J. Mol. Evol.* **36** (1993) 477.
23. S. Nakayama, N. D. Moncrief, R. H. Kretsinger, *J. Mol. Evol.* **34** (1992) 416.

24. J. E. Murphy-Ullrich, T. F. Lane, M. A. Pallero, E. H. Sage, *J. Cell Biochem.* **57** (1995) 341.
25. E. H. Sage et al., *J. Cell Biochem.* **57** (1995) 127.
26. M. C. Kilhoffer, T. J. Lukas, D. M. Watterson, J. Haiech, *Biochim. Biophys. Acta* **1160** (1992) 8.
27. R. A. Young, *Clin. Orthop.* (1975) 249.
28. *Handbook of Vitamins, Minerals, and Hormones*, 2nd ed., Van Nostrand Reinhold, New York 1981.
29. C. W. Bales, M. K. Drezner in W. N. Kelley (ed.): *Textbook of Internal Medicine*, J. B. Lippincott, Philadelphia 1992.
30. Recommended Dietary Allowances, National Academy Press, Washington D.C. 1989, p. 10.
31. U. Karbach, *J. Nutrition* **122 (suppl. 3S)** (1992) 672.
32. E. M. Brown et al., *Nature* **366** (1993) 575.
33. S. Garner in J. B. Anderson, S. Garner (eds.): *Calcium and Phosphorus in Health and Disease*, CRC Press, New York 1995, p. 157 – 176.
34. M. R. Huassler et al., *Rec. Prog. Hormone Res.* **44** (1988) 263.
35. H. F. DeLuca, *Adv. Exptl. Med. Biol.* **196** (1986) 361.
36. L. A. Austin, H. Heath, Jr., *New Eng. J. Med.* **304** (1981) 269.
37. N. C. Blumenthal, A. S. Posner, *Calcif. Tissue Res.* **13** (1973) 235.
38. A. S. Posner, *Arch. Biochem. Biophys.* **124** (1968) 604.
39. M. E. Kay, R. A. Young, *Nature* **204** (1964) 150.
40. W. N. Suki, D. Rouse in B. M. Brenner, F. C. Rector (eds.): *The Kidney*, 4th ed., Saunders, Philadelphia 1991.
41. R. Vieth, *Bone Miner.* **11** (1990) 267.
42. R. Y. Tsien, *Ann. Rev. Cell Biol.* **6** (1990) 715.
43. E. Carafoli, *J. Biol. Chem.* **267** (1992) 2115.
44. T. Meyer, L. Stryer, *Ann. Rev. Biophys. Chem.* **20** (1991) 153.
45. J. A. Albright, R. A. Brand (eds.): *The scientific Basis of Orthopaedics*, Appleton & Lange, Norwalk CI 1987.
46. J. D. Termine, *Ciba Found. Symp.* **136** (1988) 178.
47. H. C. Blair, P. H. Schlesinger, F. P. Ross, S. L. Teitelbaum, *Clin. Orthop. Rel. Res.* **294** (1993) 7.
48. E. Bonucci, *Clin. Orthop. Rel. Res.* **78** (1971) 108.
49. *The Mechanical Adaptations of Bones*, Princeton Univ. Press, Princeton N.J. 1984.
50. M. R. Christoffersen, J. Christoffersen, W. Kibalczyc, *J. Crystal Growth* **106** (1990) 349.
51. M. R. Christoffersen, J. Christoffersen, *J. Crystal Growth* **121** (1992) 617.
52. H. Tohda, M. Okazaki, T. Yanagisawa, J. Takahashi, *Biomaterials* **16** (1995) 945.
53. M. Okazaki, *Biomaterials* **12** (1991) 831.
54. N. C. Blumenthal et al., *Calcif. Tissue Int.* **56** (1995) 316.
55. E. M. Raif, M. F. Harmand, *Biomaterials* **14** (1993) 978.
56. H. I. Roach, *Cell Biol. Int.* **18** (1994) 617.
57. R. A. Atkinson et al., *Eur. J. Biochem.* **232** (1995) 515.
58. C. Chenu et al., *J. Cell Biol.* **127** (1994) 1149.
59. P. Mauerer et al., *J. Mol. Biol.* **26** (1995) 347.
60. J. D. Termine et al., *Cell* **26** (1981) 99.
61. P. Mauerer et al., *Eur. J. Biochem.* **205** (1992) 233.
62. W. J. Landis, *Bone* **16** (1995) 533.
63. N. C. Binkley, J. W. Suttie, *J. Nutrition* **125** (1995) 1812.
64. M. Tsuchimoto et al., *Jpn. J. Pharmacol.* **66** (1994) 25.
65. D. Hirsch, R. Azoury, S. Sarig, H. S. Kurth, *Calcif. Tissue Int.* **52** (1993) 94.
66. D. Hirsch, W. J. Landis, R. Azoury, S. Sarig, *Calcif. Tissue Int.* **50** (1992) 261.
67. T. Ikeda et al., *J. Clin. Invest.* **92** (1993) 2814.
68. S. Hirota et al., *Am. J. Pathol.* **143** (1993) 1003.

Calcium Antagonists

Raimund Mannhold, Abteilung für Klinische Physiologie, Physiologisches Institut, Universität Düsseldorf, Federal Republic of Germany (Chaps. 1 – 4)

Werner Seitz, BASF Aktiengesellschaft, Main Laboratory, Ludwigshafen, Federal Republic of Germany (Chap. 5)

1. Introduction 1
2. Molecular Pharmacology 1
2.1. Sites of Action 1
2.2. Structure – Activity Relationships . . . 3
3. Pharmacodynamics of Calcium Antagonists in Heart and Smooth Muscle 5
4. Clinical Application 6
5. General Aspects of Chemistry 8
5.1. Aralkylamines 9
5.2. 1,4-Dihydropyridines 10
5.3. Diphenylalkylamines 11
6. References 12

1. Introduction

During the 1980s the important bioregulatory functions of calcium ions (Ca^{2+}; in short: calcium) have emerged. In a normal resting cell the concentration of free calcium is rather low, ranging from 10^{-8} to 10^{-6} mol/L. Calcium concentrations within the cell are effectively controlled by calcium channels and calcium pumps located in the plasmalemma and internal membranes. External stimuli are able to trigger the opening of calcium channels, whereby the internal calcium concentration rises to about 10^{-5} mol/L. Thus, binding of calcium to intracellular calcium-binding proteins is enabled. Activated binding proteins can now interact with target proteins, stimulate their enzymatic activities, and trigger various cellular events. After a kinetic delay, the channels close and calcium pumps restore the initial calcium concentration (Fig. 1). The ubiquitous role of calcium as second messenger in numerous cellular events ranges from muscle contraction in smooth and striated muscle through glycogen metabolism to synthesis and release of neurotransmitters.

Amongst the numerous compounds interfering with calcium pathways by different mechanisms and at different sites, two groups are of major importance: the calmodulin antagonists first described by Weiss [1] and the calcium antagonists first described by Fleckenstein [2]. This review is mainly focused on the molecular, pharmacological, and physicochemical properties of the calcium antagonists.

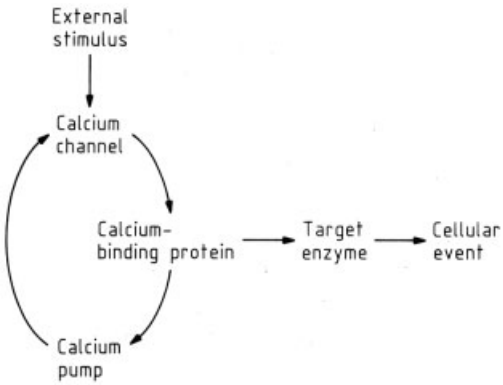

Figure 1. Scheme of calcium regulation in a eukaryotic cell

2. Molecular Pharmacology

2.1. Sites of Action

Calcium antagonists exert their effects by specific, i.e., receptor-mediated modes of action. The existence of different receptor sites according to the pronounced chemical heterogeneity of calcium antagonistic compounds is experimentally evidenced. Stereoselectivity is one strong indicator for a specific drug action and has been documented for the racemic calcium antagonists verapamil, gallopamil (D 600), prenylamine, and chiral dihydropyridine derivatives.

Calcium channels are multisubunit metalloproteins of a molecular mass of about 180000

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_519

Table 1. Classification of calcium antagonists [19–21]

Class I		Class II	Class III
Class I A	Class I B		
Nifedipine	(−)-Prenylamine	(−)-Gallopamil	Diltiazem
(±)-Nisoldipine	(+)-Prenylamine	(+)-Gallopamil	Bencyclane
(±)-Niludipine	Tiapamil	(−)-Verapamil	KB 944
(−)-Nitrendipine	Flunarizine	(+)-Verapamil	
(+)-Nitrendipine	Fendiline		
(±)-Felodipine			

[3], [4]. For calcium antagonists of the 1,4-dihydropyridine type receptor sites within the calcium channel have been identified by various groups for smooth muscle [5], [6], heart muscle [7–12], skeletal muscle [13], [14], and brain [9–12], [15–18] with the aid of radioligand binding studies. According to these investigations dihydropyridines exhibit rapid, reversible, saturable, and stereoselective binding to their receptor sites, with high affinities in the nanomolar range. Chemical characterization has revealed heat sensitivity and protein nature of the binding sites probably associated with phospholipids. Furthermore, the binding sites require divalent cations in order to bind dihydropyridines, as demonstrated by the effects of different chelators. Reconstitution of binding occurs by the addition of Mn^{2+}, Ca^{2+}, and Mg^{2+} (in decreasing order). Calcium antagonists were classified according to their binding characteristics to the dihydropyridine binding site [19–21] (Table 1). Class I contains the dihydropyridine derivatives and in addition flunarizine, prenylamine, fendiline, and tiapamil. The latter differ from the dihydropyridines by becoming weaker competitors in the presence of diltiazem whereas the dihydropyridines become more potent competitors. Class II compounds are verapamil, gallopamil, and their enantiomers; they exhibiting biphasic competition profiles representing negative heterotropic allosteric regulators. Class III drugs are positive heterotropic allosteric regulators of the dihydropyridine binding represented by the compound diltiazem. However, it remains to be clarified whether the allosteric sites regulating dihydropyridine binding are identical with the pharmacologically relevant binding sites of verapamil and diltiazem. In the case of the dihydropyridines significant correlations between binding data of a series of nifedipine analogues and their inhibition of mechanical responses was obtained [6]. Ligand binding, therefore, is occurring at the biologically relevant receptor sites.

Further substantiation of the receptor concept for calcium antagonists of the dihydropyridine type was obtained by the development of calcium agonists interacting with the same binding sites. Positive inotropic and vasoconstrictor properties were reported for the compound Bay k 8644, which could be competitively antagonized by nifedipine [22]. By means of patch clamp analysis it could be demonstrated that Bay k 8644 increases the mean opening time of the single calcium channel. The agonistic compound has been obtained by simply exchanging the ester group in position 3 of the dihydropyridine ring by a nitro substituent (see Fig. 3).

Various groups favor *calmodulin* as a receptor candidate for at least some of the actions exerted by calcium antagonists [23–27]. Calmodulin is a ubiquitous calcium-dependent regulator protein, which is involved in numerous Ca^{2+}-mediated cellular events. The pharmacological importance of calmodulin inhibition has first been investigated by the group of B. Weiss, indicating the calmodulin-binding properties of phenothiazines as a biochemical basis of their antipsychotic effects [1]. The key role of calmodulin in the regulation of smooth muscle contraction, the potent vasodilator properties, and the probably intracellular location of at least some sites of calcium antagonistic drug action [28] led to the assumption that relationships between vasodilating and calmodulin-binding properties of some calcium antagonists exist [23], [24], [27]. But only in fendiline and prenylamine, binding constants in the micromolar range were found. The comparably weak binding of verapamil and diltiazem probably excludes a contribution of calmodulin inhibition to the vasodilating properties of these compounds.

The effects of calcium and calmodulin antagonists on the calmodulin-dependent contraction of chemically skinned vascular smooth muscle from rabbit renal arteries were investigated [29]. According to these experiments the vasodilator properties of calcium antagonists cannot be explained by interference with the calmodulin system. Thus, the pharmacological importance of calmodulin inhibition for smooth muscle relaxation by calcium antagonists remains to be clarified by further investigations; however, a subclassification of calcium antagonists based on their differing potency on calmodulin-regulated enzyme activities was developed (Table 2).

Table 2. Subclassification of calcium antagonists, CaM=Calmodulin

Anti CaM activity with putative functional importance	Weak anti CaM activity	Inhibition of CaM and target enzymes
Prenylamine Fendiline Bepridil	Verapamil Diltiazem	Dihydropyridine derivatives

2.2. Structure – Activity Relationships

Beyond stereoselectivity one of the most important criteria in elucidating specific drug actions is the existence of clear-cut structure – activity relationships. The number of chemically rather diverse compounds with putative calcium antagonistic properties steadily increases. Some groups therefore try to differentiate between *specific* and *non-specific* calcium antagonists. Experimental evidence for such a classification is derived from electrophysiological [30] and physicochemical investigations [31]. The following correlations between negative inotropic potency and lipophilicity were obtained [31]: the potency of non-specific compounds increases with increasing lipophilicity whereas specific calcium antagonists exhibit an inverse dependence on lipophilicity (Fig. 2).

Consistently, at least in heart muscle, specific calcium antagonists are represented by the compounds verapamil, gallopamil, nifedipine, diltiazem, prenylamine, and fendiline.

The negative inotropic potency of calcium antagonists strongly depends on stimulation rate; the determination of amplitude – frequency relationships is therefore an appropriate criterion to derive qualitative structure activity relationships for these compounds. Figure 3 summarizes the results obtained from investigations of the two most important calcium antagonists verapamil and nifedipine in isolated cat papillary muscle. In nifedipine [32] the following *essential* moieties could be detected: (1) the dihydropyridine ring, (2) the secondary nitrogen in the heterocycle, (3) a space-filling substituent in the *para*-position of the dihydropyridine ring, and (4) an ester group in 3′-position of the heterocycle. The *ortho* nitro substituent and the 5′-ester structure have been shown to be *non-essential.* The position of the substituent on the phenyl ring is of major importance for the drug potency, which decreases from *ortho* through meta to *para* substitution. These results are in good accordance with in-vivo investigations [33] and they have been substantiated by different other groups in intestinal smooth muscle [6] and in vivo for antihypertensive potency [34].

In the verapamil molecule [28], [35–38] the nitrile group [39], both benzene rings, and the tertiary amino nitrogen are *essential* drug moieties, whereas the isopropyl group and the aromatic substituents are *non-essential. Meta* substitution in the benzene ring near the asymmetric carbon is optimal for potency; *ortho* and *para* substitution diminish potency probably because of steric effects. According to extended Hückel molecular orbital calculations the benzene ring and the nitrile group are coplanar in the most stable conformation [35–41]. These results agree well with other in-vivo observations indicating that halogen-substituted derivatives are mostly effective in lowering blood pressure [42].

Rather limited data are available on structure-activity analyses of diltiazem. The chemical requirements for the vasodilating properties of this compound in the anesthetized dog were investigated [43]. This revealed the importance of the amino nitrogen, parallel to that in the verapamil molecule. Also in the case of diltiazem the *N*-dimethyl derivative (secondary amine) as well as quaternization products were either far less active or not active compared to the parent compound. In contrast, ring substitution affected the potency in an opposing manner compared to verapamil. Halogen substitution (strong negative inductive effects) decreases the potency. Carbonyl substituents in the 3-position of the thiaz-

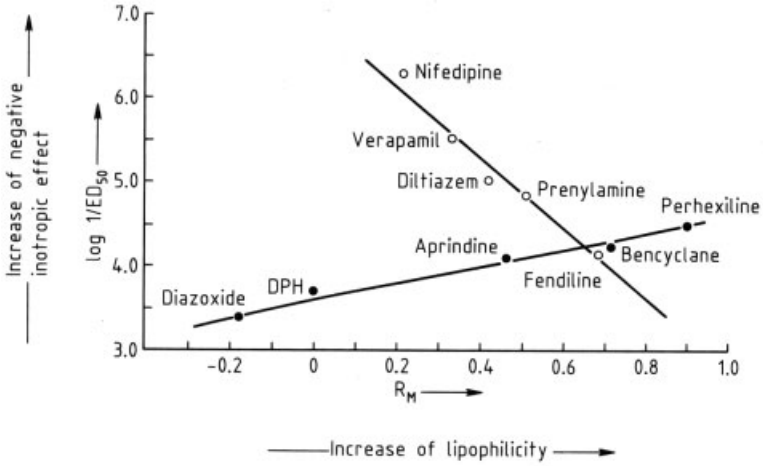

Figure 2. Correlations between lipophilicity (R_M) and negative inotropic potency (log $1/ED_{50}$) of specific calcium antagonists (open circles) and non-specific negative inotropic compounds (closed circles)

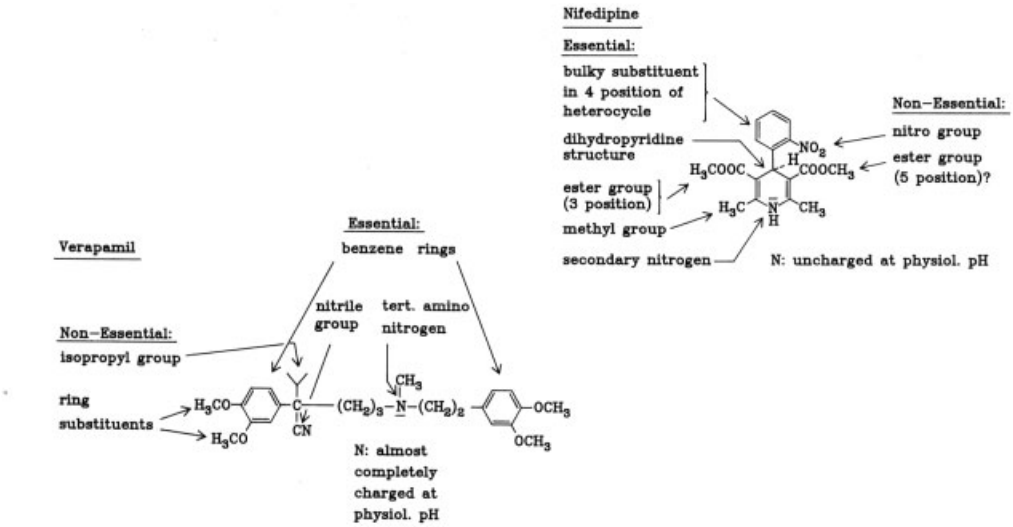

Figure 3. Discrimination between essential and non-essential drug moieties of the two calcium antagonists verapamil and nifedipine

epine ring strongly increase the vasodilating potency of diltiazem.

The electronic, lipophilic, and steric requirements for drug – receptor interactions of calcium antagonists have been investigated by means of quantitative structure – activity relationships (Hansch approach) [37]. Correlations of the negative inotropic potency of a series of verapamil and nifedipine derivatives with the respective physicochemical substituent parameters revealed the following results:

1) Steric and lipophilic properties determine the calcium-antagonistic potency of nifedipine.
2) Potency of verapamil derivatives mainly depends on electronic properties, whereas increasing steric bulk of ring substituents could lead to steric hindrance of the drug – receptor interaction.

These results prove the existence of different receptor sites for calcium antagonists.

3. Pharmacodynamics of Calcium Antagonists in Heart and Smooth Muscle

Coronary heart disease, arrhythmias, and hypertension are the main indications for the therapeutic application of calcium antagonists. Antianginal efficacy is based on direct dilatation of the coronary vessels (vasospastic angina) and on a diminution of preload and afterload due to a decreased peripheric resistance (see also, → Antihypertensives). Antihypertensive potency of calcium antagonists is likewise founded on the vasodilator properties on resistance vessels. Antiarrhythmic properties are explained, on the one hand, by impairment of pathologic calcium-mediated slow action potentials and, on the other hand, by their inhibitory action at the atrioventricular node (i.e., supraventricular tachyarrhythmia).

Most calcium antagonists exhibit negative inotropic actions in vitro, i.e., they diminish the contractile force of the isolated myocardium. However, the cardiac output of a heart in situ is not necessarily reduced by these compounds. This fact has been attributed to reflex sympathetic discharge compensating the direct negative inotropic action of calcium antagonists. The mode and site of interference of calcium antagonists with excitation – contraction coupling (e.c. coupling) is well understood. Because e.c. coupling involves a large set of potential-dependent and time-dependent calcium movements, differentiation of calcium antagonists is easily accomplished by investigating the time and voltage dependence of their action.

The time dependence of inotropic drug effects is simply characterized by measuring contraction amplitudes at varied stimulation frequencies as compared to control conditions. This was done for numerous negative inotropic compounds [44–48]. According to their influence on the amplitude – frequency relationship (AFR) one can differentiate the following three groups:

1) drugs preferentially acting at high stimulation frequencies, such as verapamil, diltiazem, tiapamil, and fendiline (inversion of AFR)
2) drugs with a negative inotropic action fairly independent on stimulation frequency as, for instance, nifedipine and prenylamine (parallel shift of AFR)
3) drugs with a preferred action at low stimulation frequencies, such as bepridil and bencyclane (accentuation of AFR)

Correlations between these pharmacological properties and lipophilic parameters (R_M-value) [31] have elucidated an inverse dependence on lipophilicity for group 1 and 2 compounds, whereas compounds of group 3 exhibit a positive correlation with lipophilicity probably consistent with non-specific membrane perturbing actions.

At the cellular level, the antiarrhythmic action of calcium antagonists is based on the inhibitory effects of these compounds on potential-dependent transmembrane ion currents in working myocardium and in pacemaker tissues. In this context sodium- and calcium-mediated fast action potentials in working myocardium have to be differentiated from mainly calcium-carried "slow" action potentials, which occur physiologically in impulse generating and propagating tissue and pathologically in hypoxic myocardial areas.

Electrophysiological studies of calcium antagonists on sinoatrial (SA) and atrioventricular (AV) node are of main clinical relevance for the antiarrhythmic therapy with these compounds (→ Antiarrhythmic Drugs). In *isolated SA nodes,* both (+)- and (−)-isomers of verapamil and gallopamil (0.2 – 1.0 μg/mL) reduce the discharge rate to the point of complete suppression of automaticity; different mechanisms are responsible for the effects. The (−)-isomers (0.3 – 0.6 μg/mL) slightly reduce the slope of the slow diastolic depolarization, causing a more effective depression of the maximum velocity of depolarization (MVD) and of the nodal conduction velocity until partial or complete nodal conduction blocks occur. The (+)-isomers (1 – 2 μg/mL) do not affect MVD or nodal conduction, but obviously shift the threshold voltage for the fast depolarization to less negative voltages.

In isolated SA nodes nifedipine (3×10^{-7} to 1×10^{-6} mol/L) decreases the discharge rate by reducing both the rate of the slow diastolic depolarization and the maximal diastolic potential until intranodal conduction blocks occur.

In humans , calcium antagonists normally have no significant effect on sinus node function. However, in patients with sinus node dysfunction, the same drugs may induce dangerous bradycardia or sinoatrial block. These different effects seem to be at least in part caused by the different responsiveness of the normal and diseased sinus node to changes in automatic tone via the baroreceptor reflex. In the AV node, conduction and refractoriness are prolonged by each calcium antagonist to a different degree. In this regard, gallopamil and verapamil have a very strong effect, whereas nifedipine in normal doses does not influence AV-nodal conductivity, and tiapamil and diltiazem have an intermediate position between these groups.

For excellent reviews on this topic see [49], [50].

Calcium antagonists exhibit highly selective actions in smooth muscle preparations. This is revealed by their preferential inhibition of depolarization-induced responses described for nitrendipine, nifedipine, and verapamil [51–53].

Agonist-induced responses exhibit varied sensitivity to calcium-antagonistic compounds. The sensitivity of calcium antagonists presumably corresponds to the relative extent to which calcium is mobilized through potential-dependent or receptor-operated channels and intracellular stores. Receptor-operated channels may exhibit different sensitivities to calcium antagonists depending on receptor and tissue under investigation. Thus, some agonist-induced contractions have been shown to be equally or even more sensitive to calcium antagonists as compared to potassium-mediated contractions. Equieffective inhibition of norepinephrine and potassium responses by verapamil were shown in the rat mesenteric artery [54]. Corresponding results were obtained for diltiazem in rabbit mesenteric artery [55].

Originally, calcium channel blockers have been defined as specific and competitive antagonists of calcium flow into heart and smooth muscles. Accordingly, mechanical responses and activated calcium uptake are inhibited by calcium antagonists in nearly the same concentration ranges [56–59].

Selective inhibitory action of calcium antagonists in vascular smooth musculature is presumably based on (1) the presence of different mechanisms for activation, (2) the presence of different groups of calcium channels activated by different stimuli, (3) the distinction between calcium channels in different tissues and (4) the distinction between calcium channels in different vascular smooth muscles [60].

Tissue Selectivity. Norepinephrine-induced contractions in rat aorta are more strongly inhibited by nifedipine than by flunarizine, whereas nifedipine is less active in mesenteric artery [61]. Serotonin-induced contractions are highly sensitive to nimodipine in rabbit basilar artery, but insensitive in saphenous artery [63]. One of the most striking examples of tissue selectivity is concerned with calcium antagonists of the piperazine type (cinnarizine, flunarizine). These compounds relax vascular smooth muscle with high potency, but their calcium-antagonistic potency in heart muscle is rather negligible.

Stimulus Selectivity. The strikingly different sensitivities of agonist-induced and potassium-induced contractions to calcium antagonists have already been mentioned. In this context the use-dependent interaction of calcium entry blockers with voltage-dependent channels represents marked differences to the interaction with receptor-operated channels. The phenomenon of use-dependent drug actions has first been detected for local anesthetics [64] in sodium channels and later on for calcium antagonists in heart muscle [65] and in smooth muscle preparations [59].

In addition to the inhibition of calcium entry interaction by calmodulin, the following alternative mechanisms of the vasodilatation by calcium antagonists are proposed: inhibition of cAMP phosphodiesterase [66], stimulation of Na^+- und K^+-activated ATPase [67], or a stimulation of the calcium pump [68]. Calcium antagonists must be differentiated further into subgroups, e.g., according to their effects [69] or to the occurrence of subtypes of channels [70]. The latter could respond to the striking selectivity of cinnarizine [71] and flunarizine [72] for smooth musculature.

The main pharmacodynamic properties of calcium antagonists are summarized in Table 3.

4. Clinical Application

The spectrum of clinical indications for calcium-antagonistic compounds is permanently widen-

Table 3. Differentiation of basic pharmacodynamic properties of calcium antagonists. Data in the last column on tissue selectivity for vasodilator potency (coronary artery, K^+-contracture) indicate the factor by which smooth muscle is more sensitive as compared to heart muscle.

Drug	Negative inotropic action				Use-dependent action		Electrophysiological action		Vasodilating action	
	EC_{50}*, mol/L	Frequency dependence	Stereoselectivity	Active species	Development of negative inotropic effect	Dissociation from specific sites	Negative chronotropic potency	Negative dromotropic potency	EC_{50}*, mol/L	Tissue selectivity
Nifedipine	5.1×10^{-7}	–	(–)	base	–	(–)	+	(+)	8.13×10^{-9}	63
Verapamil	3.5×10^{-6}	+	(–) isomer	cation	+	+	+	(+)	2.0×10^{-7}	17.5
Diltiazem	9.8×10^{-6}	+	d-cis isomer	cation	+	presumably +	+?	(+)?	2.5×10^{-7}	39
Prenylamine	2.0×10^{-5}	–	(–) isomer	cation	+	presumably –	–/+	(+)?	1.29×10^{-6}	15.5
Fendiline	5.0×10^{-5}	+	(–) isomer	cation	+	presumably +	?	?	3.6×10^{-6}	14

* EC = effective concentration.

ing. But it is not possible to list therapeutic applications that are common to all calcium antagonists. The different therapy spectra of the calcium antagonists reflect the differing molecular modes of action of these compounds. Verapamil, for example, is extraordinarily suitable in the treatment of several forms of dysrhythmias as well as hypertrophic obstructive cardiomyopathy, whereas nifedipine is not indicated in these diseases. Thus, the following brief summary of the main clinical applications should be considered in terms of variant validity for the individual members of the calcium antagonist family. Calcium antagonists have been found to be useful in treating a wide variety of cardiovascular disorders [73]:

Angiospastic angina (Prinzmetal)
Angina pectoris, any type
Dysrhythmias
Arterial hypertension
Left ventricular failure
Acute myocardial infarction
Cardiac preservation
Hypertrophic obstructive cardiomyopathy
Cerebral vasospasm

One of the most prominent indications of calcium antagonists is referred to as vasospastic or Prinzmetal's angina. In stable angina calcium antagonists apparently are not superior to β-blockers. Combination of calcium antagonists with β-blockers has revealed an impressively increased efficacy as compared to monotherapy.

Calcium antagonists are established in the treatment of arterial hypertension [73], especially when complications by coronary heart disease and/or impaired left ventricular performance are involved. In a comparative study on nifedipine and verapamil the acute and chronic effects of these compounds on blood pressure, heart rate, plasma catecholamines, renin – aldosterone, and plasma volume in patients with mild to moderate hypertension were investigated [74]. Both drugs did not induce renin stimulation or fluid retention. Nifedipine, in contrast to verapamil, increased adrenergic activity which favors the combination of this drug with β-blockers. According to the vasodilator properties of calcium antagonists a clinically important observation indicates that vascular smooth muscle shows markedly increased sensitivity with aging [61]. This was also found in hypertensive patients [62].

The first clinical empirical reports on calcium antagonists were concerned with the antiarrhythmic properties of these compounds. Verapamil is now well established as the drug of choice for treating supraventricular tachycardia. Further indications are related to atrial fibrillation and flutter and sinus tachycardia. Diltiazem exhibits antiarrhythmic properties in part resembling the properties of verapamil; doses required to produce antiarrhythmic effects with nifedipine are not achieved in clinical use.

Additional indications for calcium antagonists comprise cardiomyopathy, cerebral vasospasm, and other vasospastic syndromes.

5. General Aspects of Chemistry

From a chemical point of view the calcium antagonists are a most heterogeneous group of compounds. On the basis of structural features and lipophilicities (Table 4) three main subgroups [75], [76] can be differentiated:

1) aralkylamines, e.g., verapamil, gallopamil, bepridil, tiapamil, diltiazem, caroverine
2) 1,4-dihydropyridines, e.g., nifedipine, nicardipine, nitrendipine, nimodipine, nisoldipine, and further derivatives
3) diphenylalkylamines, e.g., cinnarizine, flunarizine, prenylamine, fendiline, perhexiline, terodiline

Table 4. Physicochemical properties of calcium antagonists. The last column refers to the degree of ionization of the respective antagonist at physiological pH (7.4).

Compound	R_M	pK	% ionized
Nifedipine	0.21	–	–
Verapamil	0.33	8.73	95.5
Prenylamine	0.51	9.47	99.2
Fendiline	0.68	9.33	98.8
Diltiazem	0.42	8.06	82.1
Perhexiline	0.89	11.09	99.9
Bepridil	–	9.39	98.9
Tiapamil	0.11	8.48	92.3

The 1,4-dihydropyridines form a rather extensive group of calcium antagonists for which quantitative structure – activity relationships have been proposed (see Section 2.2). They are easily distinguished from the other two groups. Whereas the aralkylamines and

diphenylalkylamines are protonated at physiological pH, most of the dihydropyridines are uncharged at pH 7.4 (Table 4). The lipophilic character is quantitatively expressed by the R_M values determined by reversed-phase TLC. The diphenylalkylamines can be considered as another chemical subgroup with a bulky lipophilic diphenyl moiety separated by two or three carbon atoms from a basic nitrogen atom.

5.1. Aralkylamines

Verapamil and Gallopamil. Verapamil [*52-53-9*] [77], (±)-5-[*N*-(3,4-dimethoxyphen-ethyl)-*N*-methylamino]-2-(3,4-dimethoxyphenyl)-2-isopropyl-valeronitrile, $C_{27}H_{38}N_2O_4$, M_r 454.6; the hydrochloride [*152-11-4*]melts at 139.5 –140.5 °C.

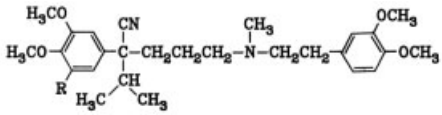

Trade names: Isoptin (Knoll), Cardibeltin (Pharma Schwarz), Cordilox (Abbott), Calan (Searle), Vasolan (Eisai); contain verapamil hydrochloride.

Gallopamil [*16662-47-8*] [77], 5-[*N*-(3,4-dimethoxyphenethyl)-*N*-methylamino]-2-(3,4,5-trimethoxyphenyl)-2-isopropyl-valeronitrile, $C_{28}H_{40}N_2O_5$, M_r 484.6; the hydrochloride [*16662-46-7*]melts at 145 – 148 °C.

Trade name: Procorum (Chemische Werke Minden); contains gallopamil hydrochloride.

Bepridil [*64706-54-3*] [78], (±)-*N*-(3-isobutoxy-2-pyrrolidin-1-yl-propyl)-*N*-phenyl-benzylamine, $C_{24}H_{34}N_2O$, M_r 366.3; the monohydrochloride monohydrate [*74764-40-2*] melts at 88 – 91 °C.

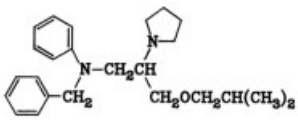

Bepridil is a long-acting (half-life 42 h) antianginal agent with antiarrhythmic properties. The pharmacological and clinical efficacy of bepridil is reviewed [79].

Trade name: Cordium (Organon); contains bepridil monohydrochloride monohydrate.

Diltiazem [*42399-41-7*] [80], *cis*-(+)-3-acetoxy-5-(2-dimethylaminoethyl)-2,3-dihydro-2-(4-methoxyphenyl)-1,5-benzothiazepin-4 (5*H*)-one, $C_{22}H_{26}N_2O_4S$, M_r 414.5; the hydrochloride [*33286-22-5*]melts at 212 °C (decomp.).

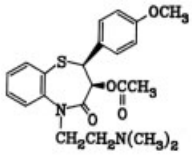

Trade names: Herbesser (Tanabe Seiyaku), Dilzem (Goedecke), Cardizem (Marion); contain diltiazem hydrochloride.

Caroverine [*23465-76-1*] [81], 1-(2-diethylaminoethyl)-3-(4-methoxybenzyl)-1,2-dihydro-2- quinoxalinone, $C_{22}H_{27}N_3O_2$, M_r 365.5, *mp* 69 °C.

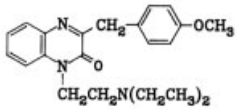

Caroverine is being used clinically as a spasmolytic. Caroverine fumarate is now under development (pre-registration) as calcium antagonist. The pharmacological properties are reviewed [82].

Trade names: Spadon (Donau-Pharmazie), Spasmium (Medichemie, CH).

Tiapamil [*57010-31-8*] [83], *N*-(3,4-dimethoxyphenethyl)-2-(3,4-dimethoxyphenyl)-*N*-methyl-*m*-dithiane-2-propylamin-1,1,3,3-tetra-oxide, $C_{26}H_{37}NO_8S_2$, M_r 555.7; the hydrochloride [*57010-32-9*]melts at 167 – 169 °C.

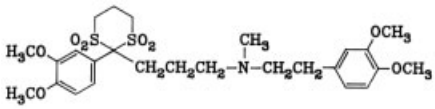

Tiapamil is structurally related to verapamil. It is a calcium antagonist with antianginal, antiarrhythmic and antihypertensive activities. First launch is expected for 1986.

Proprietary name: Larocord (Hoffmann-La Roche).

5.2. 1,4-Dihydropyridines

Nifedipine [*21829-25-4*] [84], dimethyl 1,4-dihydro-2,6-dimethyl-4-(2-nitrophenyl)pyridine- 3,5-dicarboxylate, $C_{17}H_{18}N_2O_6$, M_r 346.3, *mp* 172 – 174 °C. The chemical structures of nifedipine and other dihydropyridines are shown in Table 5.

Trade names: Adalat (Bayer), Procardia (Pfizer), Duranifin (Durachemie), Pidilat (Giulini Pharma).

Nicardipine [*55985-32-5*] [85], (±)2-(*N*-methylbenzylamino)ethyl methyl 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)-pyridine-3,5-dicarboxylate, $C_{26}H_{29}N_3O_6$, M_r 479.5; the hydrochloride [*54527-84-3*] is isolated in two crystalline forms: α-form, *mp* 168 – 170 °C and β-form, *mp* 179 – 181 °C.

Nicardipine hydrochloride is a calcium antagonistic vasodilator with short half-life, being developed for use in hypertension and angina. It was launched 1981 for cerebrovascular insufficiency and hypertension [86].

Trade names: Perdipine (Yamanouchi), Nicodel (Mitsui).

Nitrendipine [*39562-70-4*] [87], (±)-ethyl methyl 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)pyridine-3,5-dicarboxylate, $C_{18}H_{20}N_2O_6$, M_r 360.2, *mp* 158 – 159 °C, acts as a coronary vasodilator with pronounced antihypertensive efficacy. The duration of action is significantly longer than that of nifedipine [88]. Nitrendipine was launched 1985 as a once-daily antihypertensive without tachycardial side effects.

Trade names: Bayotensin (Bayropharm), Baypress (Bayer).

Nimodipine [*66085-59-4*] [89], (±)-isopropyl 2-methoxyethyl 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)pyridine-3,5-dicarboxylate, $C_{21}H_{26}N_2O_7$, M_r 418.2, *mp* 125 °C, differs from nifedipine and nitrendipine by its predominant cerebral vasodilating effects. It prevents the spasm of brain vessels without considerable decrease of blood pressure. Nimodipine is under clinical development and appears effective in treatment after subarachnoid hemorrhage, cerebrovascular insufficiency, and migraine [90].

Trade name: Nimotop (Bayer).

Nisoldipine [*63675-72-9*] [91], (±)-isobutyl methyl 1,4-dihydro-2,6-dimethyl-4-(2-nitrophenyl)pyridine-3,5-dicarboxylate, $C_{20}H_{24}N_2O_6$, M_r 388.4, *mp* 140 – 142 °C. The pharmacological properties of nisoldipine differ substantially from those of nifedipine and nitrendipine. It acts preferentially on vascular smooth muscle, leading to lowering of the total peripheral resistance. Nisoldipine is the most potent vasodilator known to act by inhibiting transmembrane calcium influx. As a powerful and fast-acting vasodilator without significant side effects nisoldipine should be useful in treating cardiac failure and hypertension [92].

Manufacturer: Bayer, first launch expected 1987.

Niludipine [*22609-73-0*] [93], bis-(2-propoxyethyl) 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)pyridine-3,5-dicarboxylate, $C_{25}H_{34}N_2O_8$, M_r 490.6, *mp* 83 – 86 °C, is a nifedipine analogue with greater vasodilatory activity and with less cardiac depression. The results of a clinical trial suggest that niludipine is a safe antianginal calcium antagonist with broad effectiveness for various types of angina pectoris [94].

Manufacturer: Bayer, first launch expected 1986.

Felodipine [*72509-76-3*] [95], (±)-ethyl methyl 1,4-dihydro-2,6-dimethyl-4-(2,3-dichlorophenyl)pyridine-3,5-dicarboxylate, $C_{18}H_{19}Cl_2NO_4$, M_r 384.3, *mp* 145 °C, is a peripheral vasodilator with antihypertensive activity. Compared to nifedipine it appears more potent, shows less cardiodepressive potency, and the effect seems longer-lasting. Felodipine appears to be an active antianginal drug and is effective in the treatment of hypertension and congestive heart failure [96].

Manufacturer: Hässle (Sweden), phase III clinical trials, first launch 1985.

Table 5. Structural formulas of some 1,4-dihydropyridine derivatives

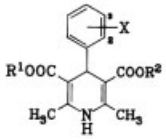

Drug	X	R^1	R^2
Nifedipine	$2-NO_2$	CH_3	CH_3
Nicardipine	$3-NO_2$	CH_3	$(CH_2)_2N(CH_3)(CH_2C_6H_5)$
Nitrendipine	$3-NO_2$	CH_3	C_2H_5
Nimodipine	$3-NO_2$	$CH(CH_3)_2$	$(CH_2)_2OCH_3$
Nisoldipine	$2-NO_2$	CH_3	$CH_2CH(CH_3)_2$
Niludipine	$3-NO_2$	$(CH_2)_2O(CH_2)_2CH_3$	$(CH_2)_2O(CH_2)_2CH_3$
Felodipine	2,3 – Cl	CH_3	C_2H_5

Nilvadipine [*75530-86-6*] [97], (±)-5-isopropyl 3-methyl 2-cyano-6-methyl-4-(3-nitrophenyl)

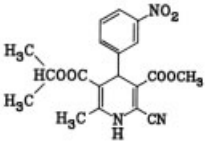

-1,4-dihydropyridine-3,5-dicarboxylate, $C_{19}H_{19}N_3O_6$, M_r 385.4, *mp* 130 – 131 °C, is a new dihydropyridine derivative and possesses the advantage over nifedipine of having light stability in solution. Compared with other dihydropyridines, the duration of the effects of nilvadipine is 2 to 3 times longer. Nilvadipine will be beneficial in various cardiocirculatory disorders and particularly promising in the treatment of cerebrovascular spasm and angina pectoris [98].

Manufacturer: Fujisawa Pharmaceutical; phase III clinical trials.

Darodipine [*72803-02-2*] [99], diethyl 4-(2,1,3-benzoxadiazol-4-yl)-1,4-dihydro-2,6-dimethylpyridine-3,5-dicarboxylate, $C_{19}H_{21}N_3O_5$, M_r 371.2, *mp* 153 – 155 °C.

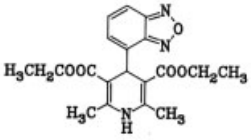

Darodipine lowers blood pressure and heart rate without depressing an ejection phase parameter of myocardial contractility. The wide separation between negative chronotropic and negative inotropic effects could present a therapeutic advantage of darodipine [100].

Manufacturer: Sandoz, phase II clinical trials.

5.3. Diphenylalkylamines

Prenylamine and Fendiline. Prenylamine [*390-64-7*] [101], (±)-*N*-(2-benzhydrylethyl)-α-methyl-phenethylamine, $C_{24}H_{27}N$, M_r 329.5, *mp* 38 – 40 °C; the lactate [*69-43-2*] melts at 137 – 140 °C.

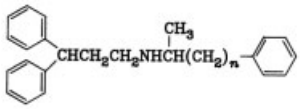

Prenylamine: $n = 1$
Fendiline: $n = 0$

Proprietary names: Segontin (Hoechst), Synadrin (Hoechst, UK).

Fendiline [*13042-18-7*] [102], (±)-*N*-(2-benzhydrylethyl)-α-methyl-benzylamine, $C_{23}H_{25}N$, M_r 315.5; the hydrochloride [*13636-18-5*] melts at 204 – 205 °C.

Proprietary name: Sensit (Thiemann).

Perhexiline [*6621-47-2*] [103], (±)-2-(2,2-dicyclohexylethyl)piperidine, $C_{19}H_{35}N$, M_r 277.5; the maleate [*6724-53-4*] melts at 189 °C.

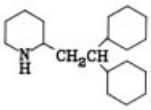

Proprietary name: Pexid (Merrell).

Cinnarizine and Flunarizine. Cinnarizine [*298-57-7*] [104], (*E*)-1-cinnamyl-4-(diphenylmethyl)piperazine, $C_{26}H_{28}N_2$, M_r 368.5; the hydrochloride melts at 192 °C (decomp.).

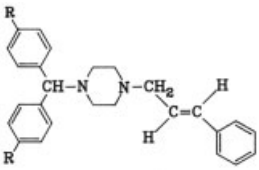

Cinnarizine: R = H
Flunarizine: R = F

Trade name: Stutgeron (Janssen).

Flunarizine [*52468-60-7*] [105], (*E*)-1-cinnamyl-4-[bis-(4-fluorophenyl)methyl]piperazine, $C_{26}H_{26}F_2N_2$, M_r 404.5; the dihydrochloride [*30484-77-6*]melts at 251.5 °C.

Trade name: Sibelium (Janssen).

Terodiline [*15793-40-5*] [106], (±)-*N*-*tert*-butyl-1-methyl-3,3-diphenylpropylamine, $C_{20}H_{27}N$, M_r 281.4; the hydrochloride[*7082-21-5*] melts at 178 – 180 °C.

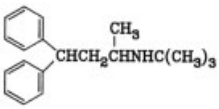

Trade name: Bicor (Recip, Sweden).

6. References

1. R. M. Levin, B. Weiss, *Mol. Pharmacol.* **13** (1977) 690.
2. A. Fleckenstein, H. Kammermeier, J. J. Döring, H. J. Freund, *Z. Kreislaufforsch.* **56** (1967) 716. A. Fleckenstein: *Calciumantagonism in Heart and Smooth Muscle,* Wiley Interscience, New York 1983.
3. D. R. Ferry, A. Goll, H. Glossmann, *Naunyn Schmiedeberg's Arch. Pharmacol.* **323** (1983) 292.
4. J. C. Venter, C. M. Fraser, C. Y. Jung, G. Bolger, D. J. Triggle, *J. Biol. Chem.* **258** (1983) 9344.
5. C. R. Triggle, D. K. Agrawal, G. T. Bolger, E. E. Daniel, C. Y. Kwan, E. M. Luchowski, D. J. Triggle, *Can. J. Physiol. Pharmacol.* **60** (1982) 1738.
6. G. T. Bolger, P. Gengo, R. Klockowski, E. Luchowski, H. Siegel, R. A. Janis, A. M. Triggle, D. J. Triggle, *J. Pharmacol. Ther.* **225** (1983) 291.
7. P. Bellemann, D. R. Ferry, F. Lübbecke, H. Glossmann, *Arzneim. Forsch./Drug Res.* **31** (1981) 2064.
8. J. G. Sarmiento, R. A. Janis, R. A. Colvin, D. J. Triggle, A. M. Katz, *J. Mol. Cell. Cardiol.* **15** (1983) 135.
9. K. M. M. Murphy, S. H. Snyder, *Eur. J. Pharmacol.* **77** (1982) 201.
10. F. J. Ehlert, E. Itoga, W. R. Roeske, H. J. Yamamura, *Biochem. Biophys. Res. Commun.* **104** (1982) 937.
11. P. Bellemann, D. Ferry, F. Lübbecke, H. Glossmann, *Arzneim. Forsch./Drug Res.* **32** (1982) 361.
12. H. Glossmann, D. R. Ferry, F. Lübbecke, R. Mewes, F. Hofmann, *J. Receptor Res.* **3** (1983) 45.
13. D. R. Ferry, H. Glossmann, *FEBS Lett.* **148** (1982) 331.
14. H. Glossmann, D. R. Ferry, *Naunyn Schmiedeberg's Arch. Pharmacol.* **323** (1983) 279.
15. P. J. Marangos, J. Patel, Ch. Miller, A. M. Martino, *Life Sci.* **31** (1982) 1575.
16. K. M. M. Murphy, R. J. Gould, B. L. Largent, S. H. Snyder, *Proc. Natl. Acad. Sci.* **80** (1983) 860.
17. R. J. Gould, K. M. M. Murphy, S. H. Snyder, *Proc. Natl. Acad. Sci.* **79** (1982) 3656.
18. D. R. Ferry, H. Glossmann, *Naunyn Schmiedeberg's Arch. Pharmacol.* **321** (1982) 80.
19. H. Glossmann, D. R. Ferry, in: A. Fleckenstein, K. Hashimoto, M. Herrmann, A. Schwartz, L. Seipel (eds.) *Drug Development and Evaluation, vol. 9, New Calcium Antagonists, Recent Developments and Prospects*, Gustav Fischer Verlag, Stuttgart-New York 1983, pp. 63 – 68.

20. H. Glossmann, D. R. Ferry, F. Lübbecke, R. Mewes, F. Hofmann, *Trends Pharmacol. Sci.* **3** (1982) 431.
21. D. R. Ferry, H. Glossmann, *Br. J. Pharmacol.* **78** (1983) 81 p.
22. M. Schramm, G. Thomas, R. Towart, G. Franckowiak, *Nature* **303** (1983) 535.
23. H. Hidaka, M. Asano, T. Tanaka, *Mol. Pharmacol.* **20** (1981) 571.
24. H. Hidaka, T. Yamaki, M. Naka, T. Tanaka, H. Hyashi, R. Kobayashi, *Mol. Pharmacol.* **17** (1979) 66.
25. M. Kanamori, M. Naka, M. Asano, H. Hidaka, *J. Pharmacol. Exp. Ther.* **217** (1981) 494.
26. P. M. Epstein, K. Fiss, R. Hachisu, D. M. Andrenyak, *Biochem. Biophys. Res. Commun.* **105** (1982) 1142.
27. J. D. Johnson, P. L. Vaghy, T. H. Crouch, J. D. Potter, A. Schwartz, in H. Yoshida, Y. Hagihara, S. Ebashi (eds.): *Advances in Pharmacology and Therapeutics II. Vol. 3, Cardio-Renal and Cell Pharmacology,* Pergamon Press, Oxford-New York 1982, p. 121.
28. R. Mannhold, R. Steiner, W. Haas, R. Kaufmann, *Naunyn Schmiedeberg's Arch. Pharmacol.* **302** (1978) 217.
29. V. A. W. Kreye, J. C. Rüegg, F. Hofmann, *Naunyn Schmiedeberg's Arch. Pharmacol.* **323** (1983) 85.
30. M. Kohlhardt, B. Bauer, H. Krause, A. Fleckenstein, *Pflügers Arch. Gesamte Physiol. Menschen Tiere* **335** (1972) 309.
31. R. Rodenkirchen, R. Bayer, R. Mannhold, *Progr. Pharmacol.* **5** (1982) 9.
32. R. Rodenkirchen, R. Bayer, R. Steiner, F. Bossert, H. Meyer, E. Möller, *Naunyn Schmiedeberg's Arch. Pharmacol.* **310** (1979) 69.
33. B. Loev, M. M. Goodman, K. M. Snader, R. Tedeschi, E. Macko, *J. Med. Chem.* **17** (1974) 956.
34. F. Bossert, H. Horstmann, H. Meyer, W. Vater, *Arzneim. Forsch./Drug Res.* **29** (1979) 226.
35. R. Mannhold, Dissertation, Universität Düsseldorf 1977.
36. R. Mannhold, R. Bayer, *Naunyn Schmiedeberg's Arch. Pharmacol.* **293** (1970) R 21.
37. R. Mannhold, R. Steiner, R. Rodenkirchen, R. Kaufmann, IUPAC-IUPHAR Symposium on Biological Activity and Chemical Structure, Noordwijkerhout (NL) 1977.
38. R. Mannhold, P. Zierden, R. Bayer, R. Rodenkirchen, R. Steiner, *Arzneim. Forsch./Drug Res.* **5** (1981) 773.
39. R. Mannhold, unpublished results.
40. H.-D. Höltje, R. Mannhold, R. Rodenkirchen, R. Bayer, *Naunyn Schmiedeberg's Arch. Pharmacol.* **316** (1981) R 33.
41. H.-D. Höltje, *Arch. Pharm. (Weinheim)* **315** (1982) 317.
42. H. Haas, E. Busch, *Arzneim. Forsch./Drug Res.* **17** (1967) 257.
43. T. Nagao, M. Sato, H. Nakajima, A. Kiyomoto, *Chem. Pharm. Bull.* **21** (1973) 92.
44. R. Bayer, R. Hennekes, R. Kaufmann, R. Mannhold, *Naunyn Schmiedeberg's Arch. Pharmacol.* **290** (1975) 49.
45. R. Bayer, R. Rodenkirchen, T. Ehara, R. Mannhold, *Naunyn Schmiedeberg's Arch. Pharmacol.* **293** (1976) R 21.
46. R. Rodenkirchen, *Naunyn Schmiedeberg's Arch. Pharmacol.* **307** (1979) R 40.
47. R. Mannhold: *Kalziumantagonisten vom Typ der aliphatischen Amine und strukturverwandte herzaktive Pharmaka. Untersuchungen zur pharmakologischen und physikalisch-chemischen Charakterisierung.* Verlag Dr. Peter Mannhold, Düsseldorf 1983, p. 1.
48. R. Bayer, R. Rodenkirchen, R. Kaufmann, J. H. Lee, R. Hennekes, *Naunyn Schmiedeberg's Arch. Pharmacol.* **301** (1977) 29.
49. L. Seipel, G. Breithardt, *Cardiol.* **69** (1982) Suppl. 1, 105.
50. N. Taira, in A. Fleckenstein, K. Hashimoto, M. Herrmann, A. Schwartz, L. Seipel (eds.): *Drug Development and Evaluation, vol. 9, New Calcium Antagonists, Recent Developments and Prospects,* G. Fischer Verlag, Stuttgart-New York 1983, p. 37.
51. R. Towart, *J. Cardiovasc. Pharmacol.* **4** (1982) 895.
52. A. Grün, A. Fleckenstein, *Arzneim. Forsch./Drug Res.* **22** (1972) 334 – 344.
53. S. Mras, N. Sperelakis, *Blood Vessels* **18** (1981) 196.
54. K. Kondo, H. Suzuki, T. Okuno, M. Suda, T. Saruta, *Arch. Int. Pharmacodyn.* **245** (1980) 211.
55. C. Cauvin, K. Saida, C. van Breemen, *J. Cardiovasc. Pharmacol.* **4** (1982) 287.
56. C. Cauvin, R. Loutzenhiser, C. van Breemen, *Ann. Rev. Pharmacol. Toxicol.* **23** (1983) 373.
57. K. D. Meisheri, O. Hwang, C. van Breemen, *J. Membr. Biol.* **59** (1981) 19.
58. S. Thorens, G. Häusler, *Eur. J. Pharmacol.* **54** (1979) 79.

59. T. Godfraind, D. Dieu, *J. Pharmacol. Exp. Ther.* **217** (1981) 520.
60. C. van Breemen, A. Mangel, M. Fahim, K. D. Meisheri, *Am. J. Cardiol.* **49** (1982) 507.
61. T. Godfraind, R. C. Miller, *Circ. Res.* **52** (1983) Suppl. I, 81.
62. F. R. Bühler, L. Hulthen, W. Kiowski, P. Bolli, in *Clin. Science* (Suppl. I), 9th Scientific Meeting of the ISH, Mexico City, Feb. 21 – 24, 1982.
63. R. Towart, *Circ. Res.* **48** (1981) 650.
64. P. Heistracher, *Naunyn Schmiedeberg's Arch. Pharmacol.* **269** (1971) 199.
65. T. Ehara, R. Kaufmann, *J. Pharmacol. Exp. Ther.* **207** (1978) 49. L. Rosenberger, D. J. Triggle, in G. B. Weiss (ed.): *Calcium and Drug Action,* Plenum Press, New York 1978, p. 3.
66. N. Sakamoto, M. Terai, H. Maeno, *Biochem. Pharmacol.* **27,** (1978) 1269.
67. M. Pan, R. A. Janis, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **41** (1982) 1483.
68. K. Hermsmeyer, R. Mason, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **41** (1982) 1483.
69. M. Spedding, *J. Cardiovasc. Pharmacol.* **4** (1982) 973.
70. P. V. Kaplita, D. J. Triggle, *Biochem. Pharmacol.* **32** (1983) 65.
71. A. Broekaert, T. Godfraind, *Eur. J. Pharmacol.* **53** (1979) 281. M. B. Emanuel, J. A. Will, *Proc. Roy. Soc. Med.* **70** (1977) Suppl. 8, 7.
72. J. M. van Nueten, J. van Beek, P. A. J. Janssen, *Arzneim. Forsch./Drug Res.* **28** (1978) 2082. M. Borgers, E. Ghoos, F. Thoné, J. M. van Nueten, *Blood Vessels* **17** (1980) 123.
73. P. D. Henry, *Am. J. Cardiol.* **46** (1980) 1047.
74. E. Agabiti-Rosei, C. Alicandri, M. Beschi, M. Castellano, R. Fariello, G. Romanelli, M. L. Muiesan, L. Corea, G. Muiesan, in T. Godfraind, A. Albertini, R. Paoletti (eds.): *Calcium Modulators,* Elsevier Biomedical Press, Amsterdam 1982, p. 257.
75. M. Spedding, *Trends Pharmacol. Sci.* **6** (1985) 109 – 114.
76. P. A. van Zwieten, *Arzneim. Forsch./Drug Res.* **35 (I),** (1985) no. 1 a, 298 – 301.
77. Knoll AG, DE 1154810, 1961 (F. Dengel).
78. CERM, DE-OS 2310918, 1973 (R. Y. Mauvernay et al.), US Re 30577, 1981 (Busch et al.).
79. J. S. Alpert, P. Coumel, K. Greeff, D. M. Krikler, W. J. Remme, E. Schönbaum, C. W. Verduyn, *Pharmatherapeutica* **4** (1985) no. 4, 195 – 222.
80. Tanabe Seiyaku, DE 1805714, 1968.
81. Donau-Pharmazie, US 3028384, 1962 (Zellner et al.).
82. Y. Ishida, H. Ozaki, S. Shibata, *Br. J. Pharmacol.* **71** (1980) 343 – 348.
83. Hoffmann-La Roche, CH 18030/73, 1973.
84. Bayer, DE 1670827, 1967 (F. Bossert, W. Vater).
85. Yamanouchi, JP-Kokai 74109384, 1967 (M. Iwanami et al.); DE-OS 2407115, 1974 (M. Murakami et al.).
86. T. Takenaka, J. Handa, *Int. J. Clin. Pharmacol. Ther. Toxicol.* **17** (1979) 1 – 11.
87. Bayer, DE 2117573, 1971 (H. Meyer, F. Bossert).
88. L. Andren, L. Hansson, L. Oroe, T. Ryman, *J. Cardiovasc. Pharmacol.* **4** (1982) , Suppl. 3, 387 – 391.
89. Bayer, DE 2117571, 1971 (H. Meyer et al.); DE-OS 2815578 (1979).
90. K. Satoh, M. Kawada, Y. Wada, N. Taira, *Arzneim. Forsch./Drug Res.* **34 (I)** (1984) no. 5, 563 – 568.
91. Bayer, DE-OS 2549568, 1975 (E. Wehinger et al.).
92. S. Kazda et al., *Arzneim. Forsch./Drug Res.* **30 (II)** (1980) 2144 – 2162.
93. Bayer, DE-OS 2670824, 1967 (F. Bossert, W. Vater).
94. H. Yasuda et al., *Arzneim. Forsch./Drug Res.* **34 (I)** (1984) no. 5, 614 – 619.
95. A. B. Hässle, EP 7293, 1979 (P. B. Berntsson et al.).
96. G. Leonetti et al., *J. Cardiovasc. Pharmacol.* **6** (1984) 392 – 398.
97. Fujisawa Pharmaceutical, DE-OS 2940833, 1979 (Y. Sato).
98. G. J. Gross, D. C. Warltier, H. F. Hardman, *J. Cardiovasc. Pharmacol.* **6** (1984) 61 – 67.
99. Sandoz, EP-A 150, 1978 (P. Neumann).
100. R. P. Hof, G. Scholtysik, *J. Cardiovasc. Pharmacol.* **5** (1984) 176 – 183.
101. Hoechst, DE 1111642, 1958 (G. Erhart).
102. Chinoin, DE 1171930, 1962.
103. Richardson-Merrell, US 3038905, 1960.
104. Janssen, DE 1086235, 1957.
105. Janssen, US 3773939, 1968.
106. Recip, GB 923942, 1963.

Calcium Carbide

BERNHARD LANGHAMMER, SKW Trostberg AG, Trostberg, Federal Republic of Germany

1.	**Introduction**	1
2.	**Physical Properties**	2
3.	**Chemical Properties**	2
4.	**Production**	2
4.1.	**Raw Materials and Their Preparation**	3
4.2.	**Electrothermic Reduction**	4
4.3.	**Furnaces**	4
4.3.1.	Single-Phase Furnaces	4
4.3.2.	Three-Phase Furnaces	4
4.3.3.	Half-Covered Furnaces	5
4.3.4.	Closed Furnaces	5
4.4.	**Operation**	7
4.5.	**Furnace Gas**	8
4.6.	**Mass and Energy Balance**	9
5.	**Further Treatment**	9
5.1.	**Cooling**	9
5.2.	**Size Reduction**	10
5.3.	**Quality Specifications**	10
5.4.	**Transport and Storage**	11
6.	**Environmental Protection**	11
7.	**Uses**	11
8.	**Economic Aspects**	12
9.	**Toxicology**	12
10.	**References**	12

Based on the article in Ullmann's Fifth Edition by FRIEDRICH-WILHELM KAMPMANN, WILHELM PORTZ, FRIEDRICH-KARL FRORATH, HEINZ OETTEL and BERNHARD LANGHAMMER

1. Introduction

Of all the carbides (→ Carbides), calcium carbide [*75-20-7*] is the most important industrially. It is still the basis of the acetylene welding-gas industry where raw materials and energy are available and petroleum is in short supply. Calcium carbide is produced in a number of countries as the starting material for the products of acetylene chemistry. The fertilizer cyanamide (see→ Cyanamides) and derived products are still produced to a limited extent. Besides these classical applications, calcium carbide is used for desulfurization of pig iron and steel, and in foundry technology, where it is mixed with other metal treatment additives.

History. The chemical composition of calcium carbide was first recognized in 1862 by WÖHLER. He carried out a synthesis starting from a calcium – zinc alloy and coal. Before this, various others had prepared calcium carbide in the laboratory, usually making use of the action of an electric arc on coal or coke products in the presence of calcium compounds. Towards the end of the last century the industrial production of calcium carbide in electrothermic furnaces was developed. Independently of each other, MOISSAN in France and MOREHEAD and WILLSON in the United States laid the foundations, in 1885 – 1895. The year 1895 saw the first production carbide furnace, operated by the Willson Aluminum Company, Spray, United States.

In 1897 BERTOLUS proposed the use of multiphase electric current for industrial furnaces. In 1904 HELFENSTEIN constructed a furnace that used a three-phase current with three electrodes. These first furnaces had a power rating of 3000 – 7500 kW, and evenhad provisions for waste-gas recovery (Helfenstein barrier); within fifty years, 40 MW furnaces were in operation.

Modern large electric furnaces, with power ratings up to 60 MW, have Söderberg electrodes and are completely closed so that the carbon monoxide byproduct can be recovered. Development in recent years has been in the direction of reliable, medium-sized closed furnaces more suited to the market situation and available raw material.

An alternative route to calcium carbide is the carbothermic process, which was operated during 1940 – 1960 by BASF in Germany, but has not been used since.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_533

2. Physical Properties

Calcium carbide [*000075-20-7*], CaC_2, M_r 64.10, is colorless and odorless when pure. Owingto the presence of impurities, various colors, from gray to brownish black, occur in technical-grade carbide, at times even violet or red-brown tones. The typical odor of technical-grade calcium carbide is that of phosphine, which is formed in moist air from traces of calcium phosphide. The melting point of pure calcium carbide is 2160 °C [1]; above 2500 °C complete decomposition takes place [2], [3]. Four modifications of calcium carbide are known:

- Phase I: tetragonal, stable between 25 and 440 °C
- Phases II and III: lower symmetry, stable below 440 °C
- Phase IV, face-centered cubic, stable above 440 °C

Commercial calcium carbide is mainly phase I. Phase II is formed in some pure preparations [4–8]. The heat of transformation of calcium carbide at 440 °C is 8.68 J/g [9].

The CaC_2 – CaO phase diagram (Fig. 1) [1] is a simple eutectic system without compound formation or solid solutions; the pure components are not miscible with each other to a significant extent in the solid state. Above the liquidus line there is complete miscibility. Technical-grade calcium carbide has a lower liquidus temperature owing to the presence of impurities.

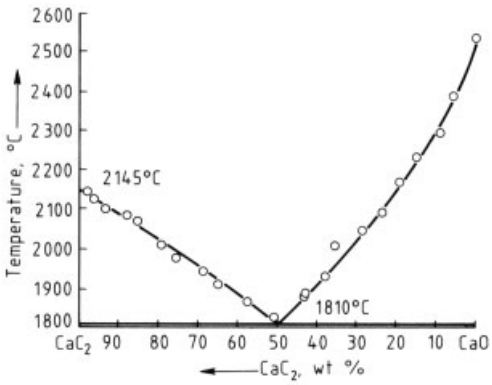

Figure 1. Phase diagram CaC_2 – CaO

The average specific heat capacity between 20 and 440 °C is 1047 J kg^{-1}K^{-1}; the heat of fusion of pure calcium carbide is 1470 J/g [10].

The standard enthalpy of formation of CaC_2 from the elements at 298 K and the standard entropy of CaC_2 at 298 K are $\Delta H = -59.4$ kJ/mol and $S = 70.3$ J mol^{-1} K^{-1} [11], [12], [13].

The electrical conductivity of calcium carbide depends strongly on the degree of purity and increases with increasing temperature. Values between 100 and 1000 Ω^{-1} have been reported [14].

3. Chemical Properties

Commercially important reactions of calcium carbide are conversion to calcium cyanamide

$$CaC_2 + N_2 \longrightarrow CaCN_2 + C \quad \Delta H^{298} = -289 \text{ kJ/mol}$$

hydrolysis to acetylene

$$CaC_2 + 2\,H_2O \longrightarrow C_2H_2 + Ca(OH)_2 \quad \Delta H^{298} = -128.4 \text{ kJ/mol}$$

and the reaction with sulfur in molten iron

$$CaC_2 + [S]_{Fe} \longrightarrow CaS + 2\,[C]_{Fe}$$

Acetylene is produced smoothly at temperatures below 300 °C. Pure calcium carbide yields 346.89 L of dry acetylene per kg CaC_2 at 0 °C and 101.3 kPa (372.66 L at 15 °C, 101.3 kPa (1013 mbar), saturated with water vapor). At red heat, calcium oxide, hydrogen, and carbon dioxide are formed [15], [16]. The hydrogenation of calcium carbide under pressure produces calcium hydride and methane as well as other hydrocarbons [17].

Calcium carbide is soluble in molten salts, such as lithium chloride, sodium potassium chloride, and calcium chloride fluoride. Various reductions can be carried out in these solutions [18]. Electrolysis of calcium carbide dissolved in a molten salt yields calcium at the cathode and carbon at the anode.

4. Production

The preparation of pure CaC_2 is only of academic interest. The industrial production of technical-grade calcium carbide is based on the reaction of quicklime with carbonaceous materials such as coke and anthracite:

$$CaO + 3\,C \longrightarrow CaC_2 + CO \quad \Delta H^{298} = +465\,\text{kJ/mol}$$

The reaction rate becomes appreciable only above 1600 °C. Above 2000 °C undesirable side reactions become noticeable, especially

$$CaC_2 \longrightarrow Ca + 2\,C$$

$$CaC_2 + 2\,CaO \longrightarrow 3\,Ca + 2\,CO$$

For these reasons, all industrial processes operate between 1800 and 2100 °C and give a technical-grade calcium carbide with a CaC_2 content of ca. 80 %. The remainder is mainly calcium oxide, but there are also other impurities from the raw materials, principally SiO_2 and Al_2O_3.

The energy for the production of the calcium carbide must be introduced at very high temperatures. There are two possible routes:

1) Melting of the carbonaceous material with CaO in a short electrothermic shaft furnace
2) Combustion of carbonaceous material with pure or enriched oxygen in the presence of CaO in a shaft furnace

The development of the second method has now been discontinued after about 50 years because the synthesis gas byproduct can be made more cheaply from petroleum. In Japan, research has been undertaken to combine methods 1 and 2 [19], but all the calcium carbide produced today is manufactured by method 1.

4.1. Raw Materials and Their Preparation

The raw material or burden that is introduced into the electrothermic furnace consists of a carbonaceous component and a lime-bearing component. The physical properties (particle size, particle strength, conductivity, etc.) and the chemical properties (content of volatiles, ash content, etc.) of the raw material depend on the type of carbide furnace and on the desired calcium carbide quality. Smaller open furnaces do not impose such strict demands because during operation the conditions inside the furnace may be observed and corrective action can be taken, either manually or mechanically. With high-capacity closed furnaces, the properties of the raw materials must allow the furnace to be operated without outside intervention for at least a few days.

The grain-size distribution of the raw material must enable the carbon monoxide to escape. In the reaction zone of the furnace the lime dissolves in the carbide melt, while the carbon component reacts as a solid grain to produce calcium carbide in the liquid. Therefore lime with a large grain size can be used without affecting its reactivity, but for the carbon component a finer grain size is usually necessary.

Modern furnaces are fitted with hollow electrodes through which fine-grained raw material can be blown directly into the molten zone. These manufacturing units only impose an upper limit to the grain size; the necessary fineness is obtained by sieving.

Impurities, particularly metal oxides, cause side reactions that waste energy. The principal impurities are iron oxides, SiO_2, Al_2O_3, MgO, and compounds of nitrogen, sulfur, and phosphorus. Oxides of iron and magnesium are completely reduced, the others only partially. The iron combines with silicon to give FeSi, which is liquid at reaction temperatures. It is partly emulsified in the carbide and can be tapped off. The vaporization of magnesium and its oxidation by carbon monoxide in the cooler parts of the furnace lead to buildup of dust and solid incrustations [20]. The oxides SiO_2 and Al_2O_3 are generally dissolved in the carbide – CaO melt along with the nitrides, sulfides, and phosphides formed from the nitrogen, sulfur, and phosphorus compounds.

The following carbonaceous materials are available: coke from bituminous coal or lignite, anthracite, petroleum coke, and charcoal. Coke from bituminous coal is used worldwide for calcium carbide production.

Uncalcined anthracite and petroleum coke can be used only as a part of the charge in a closed furnace. If the volatile content is reduced below 2 % by calcination, they can be used as the sole charge. This is done in South Africa, a country rich in anthracite. Charcoal can be used only in small open furnaces.

Today for closed furnaces coke with a grain size of 3 – 25 mm is used. The fraction below 3 mm is blown in through the hollow electrodes. Small open furnaces can also use coarser grain sizes. Before introduction, the coke must be

dried below 2 % moisture content. Ash content below 15 % is necessary to meet the quality standards for the calcium carbide product in many countries.

The lime-containing component of the raw material mixture is predominantly quicklime. Manufacturing plants with large closed furnaces specify a grain size of 6 – 50 mm, and the material below 6 mm is blown in through the hollow electrodes. For closed furnaces, the magnesium content should be below 2 %. The lime must be burnt so as to be as soft as possible, but resistant enough for transport and storage. This is achievable in modern lime kilns. At the same time the CO_2 content is reduced to less than 2 %.

The calcium hydroxide left behind after acetylene generation from calcium carbide, after granulation or briquetting with calcination, is also a cheap raw material. However, this lime recycling is limited by the build-up of impurities.

Typical examples of compositions of raw materials for calcium carbide production are listed in Tables 1 and 2.

Table 2. Composition of lime for calcium carbide production (wt %)

	New lime	Recycled lime
Total CaO	93.8	89.0
Free CaO	93.6	–
$CaSO_4$	0.4	–
SiO_2	1.3	2.9
Fe_2O_3	0.4	0.4
Al_2O_3	0.5	2.4
MgO	0.8	0.8
CO_2	2.7	3.0
P_2O_5	0.015	0.01
Loss on ignition (CO_2)	0.5	0.5

4.2. Electrothermic Reduction

The raw material mixture is heated to a high temperature in a short shaft furnace with an electric current. Investigations of the reaction have led to the conclusion that it takes place in two steps:

$$CaO + C \longrightarrow Ca + CO$$

$$Ca + 2\,C \longrightarrow CaC_2$$

The principal reaction in the furnace takes place in a $CaC_2 - CaO$ melt. Electric current, energy supply, temperature, and chemical reactions interact in a manner that is not yet completely understood. According to [21] the rate-determining step is the transport of the CaO dissolved in the melt to the surface of the coke grain, where the CaC_2 is formed. This CaC_2 then immediately dissolves in the melt. The liquid calcium carbide is tapped from the bottom of the furnace.

4.3. Furnaces

Carbide furnaces can be divided into various classes according to their electrical supply (single-phase, three-phase), the arrangement of the electrodes (serial, symmetrical), and the means of excluding air (open, half-covered, closed).

4.3.1. Single-Phase Furnaces

High-performance single-phase furnaces were developed in 1928 by MIGUET. This first furnace had a power rating of 15 MW. These furnaces are operated with low voltage and high current; optimal arrangement of the electric circuit produces a power factor ($\cos\varphi$) of up to 0.95.

4.3.2. Three-Phase Furnaces

The majority of today's carbide furnaces are operated with three-phase current. They can be classified according to the arrangement of the electrodes.

Serial. In a serial furnace the center lines of the three electrodes lie in a straight line, or they form an isosceles triangle. Rectangular and oval furnaces were formerly equipped exclusively with package electrodes made from precoked coal; later, rounded rectangular *Söderberg electrodes* were preferred. The serial furnace has advantages over the round furnaces as regards feeding and distribution of the burden and can be operated half covered without any problems. Water-cooled barriers between the electrodes not only provide for even distribution of the burden but also promote cooling of the upper part of the solid material, hence reducing its conductivity.

Table 1. Composition of carbonaceous raw materials for calcium carbide production (wt %)

	Metallurgical coke	Anthracite	Petroleum coke
Fixed carbon	89.0	80.0	81.6
Volatile carbon	0.8	6.0	9.0
Moisture	0.8	8.0	8.0
Sulfur	0.8	0.8	1.9
Ash	9.4	6.0	0.3

The unsymmetrical arrangement of the secondary current conductors leads to an inductive effect and consequently to dead and live phases with respect to the middle phase. Output and electrode consumption differ at the three electrodes owing to the differences in power consumption. (For the theoretical relationships, see [22–25].) Difficulty can occur in discharging the melt at the dead phase; this can be alleviated, for example, by addition of lime. A further disadvantage compared with circular furnaces is the unfavorable ratio of furnace capacity to surface area, which leads to greater heat loss.

Furnaces of this type still exist and are operated at up to 40 MW, sometimes with hollow electrodes (see page 5). However, all modern carbide furnaces are now symmetrical three-phase furnaces.

Symmetrical. In symmetrical three-phase furnaces the electrodes and as far as possible the transformers together with the high current conductors are arranged symmetrically about the center. In a circular furnace vessel the electrodes are situated at the corners of an equilateral triangle [26]. Well-designed circular furnaces have the advantage that high power factors can be achieved. At a sufficiently high power, molten zones are formed at all three electrodes and overlap at the center of the furnace. The tapping points should be situated as near as possible to the electrodes. On account of the high temperature, the furnace wall at the tapping point must be water-cooled or constructed of highly heat resistant material.

4.3.3. Half-Covered Furnaces

According to the equation

$$CaO + 3\,C \longrightarrow CaC_2 + CO\ (g)$$

one mole of CO is produced for each mole of calcium carbide. In the first years of carbide production this gas was simply burnt at the open furnace. Because of the air pollution they cause, large furnaces of this type have largely been retired from service.

In Germany, HELFENSTEIN attempted in 1910 [27] to construct above the linear furnace rooflike covers made of welded, water-cooled steel tubes between the electrodes. These gas collectors can also consist of ceramic material traversed by cooling coils. The liberated gas can be collected in this way. For the symmetrical three-phase furnace the construction of a gas collection hood is more difficult.

4.3.4. Closed Furnaces

Closed three-phase furnaces permit 100 % gas recovery and in normal operation do not cause any pollution of the environment by escaping dust-laden carbon monoxide. A schematic of a typical closed furnace is shown in Figure 2.

Hearth. The hearth consists of a welded steel outer casing stiffened with steel girders and ribs. The horizontal steel sheets are fixed in a ring shape round the hearth for reinforcement. A natural air current should flow around all sides.

The base of the hearth is lined with carbon blocks and rammed carbon to a height of about 1.5 m. The side walls are lined with refractory bricks or rammed refractory material. The region of the tapping point, because of the especially high temperatures, is more strongly constructed or provided with water cooling.

Electrodes. For supplying electrical energy to the molten region of the carbide furnace at ca. 2000 °C, carbon is the only suitable material. The electrodes must have high electrical and

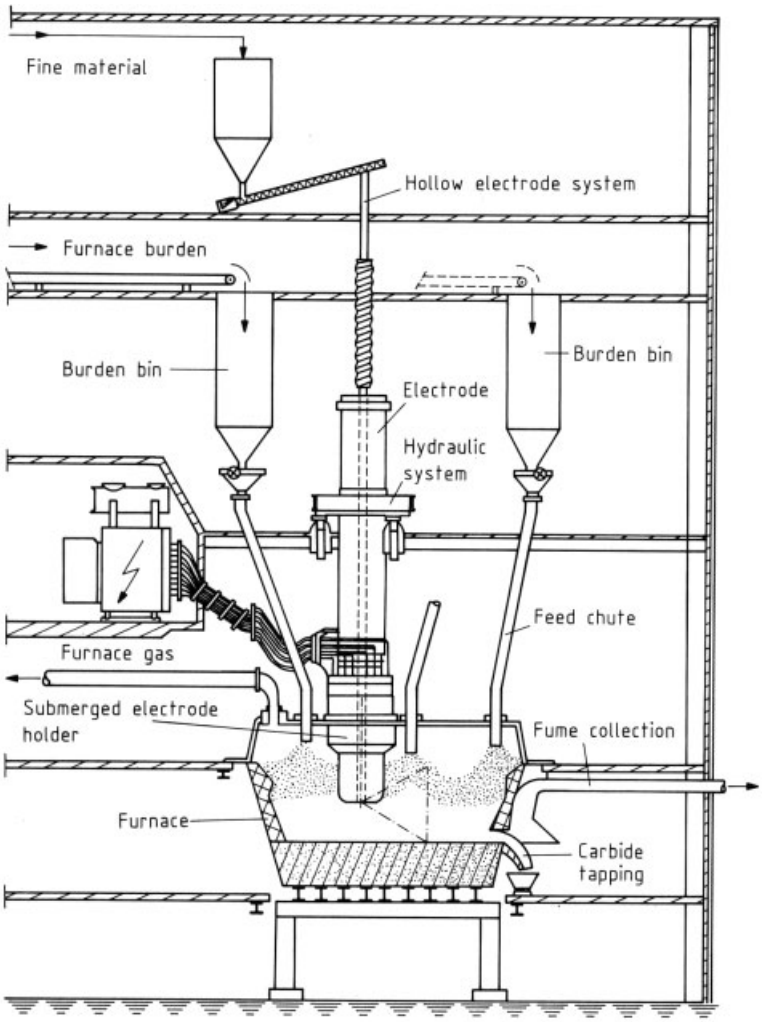

Figure 2. Closed carbide furnace

thermal conductivity and adequate mechanical strength.

The *prebaked carbon electrode*, which was at one time the only type used, has with few exceptions been replaced by the Söderberg electrode, first tried in 1921 [28]. In open rectangular furnaces prebaked packaged electrodes were used. Round electrodes are manufactured up to 1.5 m in diameter and 2.5 m in length. The current density achievable with prebaked carbon exceeds 10 A/cm^2.

The *self-baking Söderberg electrode* consists of a sheet steel cylinder with internal vertical ribs. This casing is filled with electrode paste. Söderberg electrodes are the most economical type of electrode for furnace ratings from about 10 MW upwards. The electrode paste is manufactured from coke, calcined anthracite, or petroleum coke and a tar/pitch binder. The most important property of the paste is freedom from breakage, even under high or variable load and at low consumption. The consumption of the Söderberg electrode is compensated by welding on a new steel casing and filling with new electrode material during furnace operation. The consumption of Söderberg paste can range from

15 to more than 20 kg per tonne of calcium carbide, especially if the electrode is not hollow.

The *hollow Söderberg electrode* has been used increasingly since 1961 [9], [30]. Lime and coke fines are introduced directly into the molten region through a central pipe by a carrier gas such as carbon monoxide or nitrogen. The carbide-forming reaction causes a temperature reduction and cools the tip of the electrode; a reduction of the rate of electrode consumption by up to 50 % is achieved, and cheap lime and coke fines can be used. Prebaked electrodes can also be provided with a central hole and operated as hollow electrodes. Large furnaces with power loads higher than 30 MW should always be equipped with hollow electrodes.

The electrode connection for continuous electrodes consists of water-cooled metal contact clamps, adjusted exactly to the shape of the electrode to ensure optimum transfer of the electric current. Up to ten pressure plates are pressed against the round electrode casing by a water-cooled pressure ring. A remote-controlled slipping device adapts the slipping rate to the rate of consumption of the electrode, so that the baking zone of the Söderberg electrode does not drop below the clamps. Otherwise, burn-through of the casing and breakage of the electrode can occur. The baking zone can be monitored by thermocouples, mechanical sounding, or by measuring the voltage between the casing and a steel rod in the electrode. The electrode length is an important factor for furnace operation and can be measured by mechanical sounding or weighing.

A new development is the modular electrode holder, in which the electrode ribs extrude through the casing into the contact assembly [31]. The contact clamps and pressure ring are replaced by a slipping unit at each rib. There are no radial forces acting on the electrode, and heat removal from the electrode is low.

The *electrical supply* to industry is today exclusively three-phase alternating current.

In the oil-cooled transformers the oil is circulated and cooled indirectly by water. The oil pressure is usually higher than the water pressure to prevent water from entering the oil circuit in the event of any leakage. On account of the occasional violent load alterations of the furnaces, such as may be caused by variable current demand and variable resistance of the burden, the power of the transformer must be controllable with the largest possible load variation. For this purpose motor-driven sequence switches are necessary. The operating voltage of today's carbide furnaces is in the region of 100 – 300 V. The secondary current (up to 150 kA) is carried by water-cooled copper conductors or flexible cables.

A power diagram is used to represent the specific operating conditions of a three-phase furnace. Apart from the negligible ohmic resistance of the high-current cables and the electrodes, this includes the constant, calculable reactance of the transformer, cables, and electrodes, as well as the variable resistance (impedance) of the furnace itself [23–25]. The power must be variable between wide limits for practical operation, e.g., starting up and shutting down the furnace, adjusting the electrical load to suit the conditions in the furnace (temperature, pressure), or today's usual wide-ranging operating schedule for electricity consumption (day/night usage, etc.).

For the single-phase furnace a current regulator is sufficient; for the multiphase furnace, impedance control, i.e., the control of apparent furnace resistance, by setting the immersion depth of the electrodes is necessary.

Certain types of closed hood construction allow for raising and lowering of the feed chutes so that, in addition to supplying the feed material, they can be used to alter the configuration of the surface of the burden and hence the furnace resistance.

4.4. Operation

On start up, the carbide furnace is charged with a layer of coarse coke. Then the electrodes, at a low voltage, are brought towards the coke until contact takes place, which can be seen by the appearance of electric arcs between individual pieces of coke. When the black material is at red heat, the furnace bottom is gradually fed with burden and the power is slowly increased.

Smooth operation of the furnace requires smooth feeding of burden. Any variations in the optimal composition of the burden are usually compensated by addition of lime. This also necessitates a desirable deeper immersion of the electrodes in the melt owing to the low conductivity of the lime. When the electrodes are raised too high, carbide is produced in the upper layers

of the burden, which leads to increased danger of spurting from the melt. The mass of solid material must remain permeable to gas at all times.

Closed three-phase furnaces have numerous feed chutes, which are arranged around the electrodes and automatically supply the burden. Special attention must be paid to the composition and grain size of the black – white mix to prevent any segregation. The feed must above all contain no dust, in order to reduce the danger of crusting. However, if cheap, very impure raw materials are used, crust and slag formation cannot be eliminated, and they appear on the lower electrical connectors and the hood. In such circumstances and also in the case of segregation of the feed materials, short circuits can be produced and can destroy the parts affected. A compromise is necessary between cheap raw materials and avoiding damage to the furnace.

The gas pressure inside the furnace hood should be about equal to atmospheric pressure. A lower pressure causes atmospheric oxygen to leak into the furnace, severe excess pressure leads to carbon monoxide contaminating the environment and can endanger operating personnel.

Where large units or groups of furnaces are in production, the entire furnace operation is usually computer-controlled from a separate control room.

To tap off the calcium carbide, the furnace is opened mechanically or by burning a tapping hole with an electric arc. The tapping electrode is a baked electrode, 10 – 12 cm thick, which is supplied with electricity from its own transformer, a furnace transformer, or both. The molten calcium carbide flows at 1700 – 2100 °C, depending on its purity, through cooled tapping spouts into cast-iron ladles that can be handled on a carousel arrangement or continuous track. Owing to its high melting point and the rapid loss of heat by radiation, technical-grade calcium carbide (80 % CaC_2) solidifies very quickly so that the tapping holes must be kept open either mechanically or manually with steel rods. When tapping is completed, the hole can be closed with a clay plug or plugging carbide is propelled or blown into the tapping hole until it solidifies. The fumes arising from the tapping hole are removed by fume extraction equipment. As these fumes have the character of an aerosol (particle size, 20 % $< 10\,\mu m$), purification is expensive.

During tapping the process operators must be protected from the hot gases, dust, and spurting carbide with the aid of chain curtains, wire netting, and steel shields installed in front of the tapping point. The tapping men wear aluminum-covered fireproof coats, dark glasses, and protective visors, in addition to fire-retardant working clothes.

Modern tapping machines equipped with a hydraulic manipulator can carry out all tapping operations. They are fitted with interchangeable tools such as tapping electrodes, reaming rods, and plugging devices. In this way, the tapping personnel can be reduced to one or two men per shift.

The reduction (especially of Fe_2O_3 and SiO_2) by carbon yields iron and silicon, which react to give ferrosilicon. This FeSi, on account of its appreciably higher density, collects under the carbide melt and normally flows away on tapping the furnace. As the melting point of this alloy is however several hundred kelvin below that of the carbide melt, there is a danger that FeSi could flow through cracks in the extremely hot furnace lining and cause damage in the area surrounding the furnace before it solidifies. In every carbide plant constant attention is paid to this problem.

4.5. Furnace Gas

Composition. The composition of the furnace gas depends on the raw material. If moist coke is used, the hydrogen content increases on account of the water-gas reaction, $C + H_2O \rightleftharpoons CO + H_2$. The result is the same when any appreciable amounts of volatiles in the black material are thermally cracked. Incompletely calcined limestone leads to the production of carbon monoxide as a result of the reaction $CO_2 + C \rightleftharpoons 2\,CO$, which consumes both coke and energy.

A sudden increase in hydrogen content suggests a water leak in the furnace. If the carbon dioxide and nitrogen contents increase, then the hood can no longer be relied upon to be gastight. If atmospheric oxygen enters the furnace hood, combustion occurs and increases the temperature of the cooling water and gas. Explosions are to be expected only on ignition of an unburnt mixture of oxygen and carbon monoxide when the furnace is cool.

A typical analysis of the purified gas (vol %) follows:

CO	80 – 90	CH_4	1
H_2	6 – 15	N_2	2 – 7
CO_2	0.5 – 3.0		

The heat of combustion is ca. 11 720 kJ/m^3 (STP).

Gas Purification. The carbon monoxide obtained from half-covered or closed furnaces must be dedusted. Both dry and wet dedusting methods are used.

In the dry dedusting process, the gas is, e.g., filtered by self-supporting candle filters made of ceramic fibers [33]. Their surfaces are cleaned off-line by jet pulses of previously cleaned gas or nitrogen. A final dust content of < 1 mg/m^3 can be attained. Finally the hot furnace gas is cooled in a heat exchanger. The dust from the filter has the composition 35 – 40 % CaO, 15 – 25 % MgO, 8 % $Fe_2O_3 + Al_2O_3$, 7 % SiO_2, 10 % CO_2, 9 % C and it is difficult to use. Owing to the presence from time to time of soluble cyanide the disposal of dust also is a problem. However, there are methods of roasting that remove cyanide.

In the wet process the gas is passed through several washing towers and sprayed with circulating water. In the last stage, Theisen washers reduce the dust content to ca. 5 mg/m^3 at STP. Cyanide in the slurry is precipitated as an iron complex, and the slurry is separated in thickeners, usually hydrocyclones. The clear liquor that overflows is reused.

4.6. Mass and Energy Balance

A modern large furnace without hollow electrodes requires per tonne of 300-L calcium carbide 950 kg lime (94 % CaO), 550 kg coke (dry, 10 % ash), 20 kg electrode material, and 3100 kWh electricity [26].

The typical analysis (wt %) of a commercial calcium carbide that yields 300 L of acetylene per kilogram of calcium carbide follows:

CaC_2	80.5	Fe	0.2
CaO (free)	12.9	S	0.5
Si	1.3	C	0.3
Al	1.1		

The enthalpy of reaction (ΔH) for such calcium carbide is 465 kJ/mol, corresponding to ca. 2000 kWh per tonne. In practice, for 300-L technical-grade calcium carbide, depending on purity of raw materials, furnace type, and loading, about 3000 – 3300 kWh/t are required. The difference is accounted for by the heat contents of the molten carbide and furnace gas and the heat lost to the cooling systems.

5. Further Treatment

5.1. Cooling

Most carbide plants use the block casting method, in which the liquid calcium carbide is tapped off into a ladle or crucible and solidifies to a block. The block is allowed to cool and is then broken up, ground, and sieved to the required size.

Generally, a cast-iron crucible receives the liquid carbide at ca. 1900 °C from the tapping point. After a short time, a solid crust forms on the walls of the crucible and isolates the iron from the hot liquid core of the block. The heat passing through this crust is then radiated from the iron. To prevent cracking of the crucible, this process is regulated so that the temperature of the iron does not exceed 600 °C, mainly by ensuring that the mass ratio of iron to carbide exceeds 2 : 1. Depending on the power of the furnace, crucibles with a capacity between 0.4 and 2.0 t are used.

For further cooling of the block there are several possibilities. If small molds are used, the block is lifted or tipped from the mold after ca. 2 h and then cools further outside the mold, either on a block transporter, or on steel conveyor belts that transport it to the crusher. The cooling times are ca. 20 h.

Higher power carbide furnaces are generally equipped with crucibles of capacity over 1 t. In one operating procedure the filled crucible is placed in a cooling room, and after 25 – 30 h the contents of the crucible are tipped into the crusher. This procedure has the disadvantage that for a 55 MW carbide furnace, 400 1.5 t crucibles must be ready for use. As the weight of the crucible must be at least twice the weight of the contents, this means an investment in about 1200 t of cast iron for crucibles.

A simpler procedure is to use a conical crucible open at the bottom and placed on a steel plate. After ca. 2 h the crucible can be lifted from its bottom plate. The block is then pushed onto an inclined plane and onto a band conveyor which transports it to the crusher. Conveyor tapping systems, which allow continuous transport of the calcium carbide from the furnace to the crusher, have been developed [34].

All attempts to use even a part of the heat content of the calcium carbide block (ca. 33 % of the energy used in production) have failed.

5.2. Size Reduction

Because calcium carbide reacts with atmospheric water vapor, further processing of the block is started when its average temperature reaches ca. 400 °C, to minimize the losses of acetylene during processing. The method of processing depends on the intended use of the calcium carbide. The first step is always to crush the block, usually to < 100 mm. Then the desired size grading is produced by sieving, further crushing, and, if necessary, milling in impact mills. Magnetic impurities (iron and ferrosilicon) must be removed from the calcium carbide. For this purpose, magnetic separators are installed in all stages of the size reduction process.

5.3. Quality Specifications

In many countries the quality requirements are laid down in standard specifications:

United States	Federal Specification 0-6-101 a (July, 1949)
Australia	Standards Association of Australia, Publication No. K 49 (1931)
Belgium	Norme Belge, NBN 211-01 (01.72)
France	Norme Française, NF T 24-001 (10.54 + 09.55)
Germany	DIN 53922 (07.79)
United Kingdom	British Standard 642 (1965, 11.66)
India	Indian Standard IS 1040 (1978)
Israel	Israel Standard 267 T [250 × 4], SJ 446 (07.62)
Japan	Japanese Industrial Standard JIS K 1901 (1974)
Mexico	Norme Official K 23 (1945)

Normally these standards specify sieve analysis, dust content, gas generation, and phosphorus content.

The principal criterion of the quality of calcium carbide is the amount of acetylene produced per unit mass of calcium carbide by reaction with water – the so-called *liter number*, because it is a direct measure of the CaC_2 content of the carbide.

The method of determining this gas yield is laid down in the standard specifications. The British standard calculates the dry gas volume at atmospheric pressure and 15 °C, whereas the German standard assumes that the gas is saturated with water vapor at 15 °C. (Such differences explain the differences in the liter numbers.)

When calcium carbide is crushed, the impurities concentrate in the finer fractions, and thus the minimum liter number of commercial calcium carbide depends on the grain size. For the coarsest size fraction the British standard requires 295 L/kg, and the German standard 300 L/kg. For smaller size fractions, e.g., 4 – 7 mm, the British standard is 275 L/kg, and the German standard 260 L/kg.

Because the gas determination methods in all specifications are time-consuming, industry has developed methods of testing for liter number by using small amounts of calcium carbide in more compact apparatus. In this way a large number of tests of sufficient accuracy can be carried out in a short time, and calcium carbide producers can continuously monitor their production.

The most important *size ranges* specified in individual standards are (in mm): 2 – 4, 4 – 7, 7 – 15, 15 – 25, and 25 – 50. The British standard also gives the ranges 1 – 2, 4 – 80, 1 – 80, and 0 – 80. It also mentions the well-known commercial form 14 ND (0.4 – 1.7 mm), without making it a standard.

The phosphine content of the acetylene produced is specified in all the standards. The maximum is in the region of 0.1 vol %. Detailed methods for the determination of sulfur and of materials insoluble in dilute sulfuric acid are also given, but no upper limits are set.

For the determination of the quality of the granular calcium carbide, samples must be taken correctly. All standard specifications therefore give precise instructions regarding sampling and sample preparation.

5.4. Transport and Storage

Owing to the fact that calcium carbide produces a flammable gas in the presence of water, it is classed as a hazardous industrial material. Therefore, in most countries special regulations concerning packing and storage have been issued.

In Britain "The Storage of Carbide of Calcium", published by the British Acetylene Association, states the regulations.

According to the "Verordnung über Acetylenanlagen und Calciumcarbidlager", anyone in Germany who stores calcium carbide must register [35], [36]. The requirements for these storage sites are stated in "Technische Regeln für Acetylenanlagen und Calciumcarbidlager" (TRAC 301, 1982). In addition, the provisions of the "Arbeitsstoffverordnung" (industrial materials regulations) are applicable.

The following international regulations govern the transport of calcium carbide [37–39]:

Land transport (ADR/RID)
  Class 4.3
  Item 17b
  UN no. 1402
Sea transport (IMDG)
  Class 4.3
  UN no. 1402
  PG II
  MFAG 705
  EMS 4.3-03
Air transport (ICAO-TI, IATA-DGR)
  Class 4.3
  UN no. 1402
  PG II

Labelling (EU directives):

Calcium Carbide
EEC no. 200-848-3
Symbol: F – Highly Flammable
R phrases
  R15: Contact with water liberates extremely flammable gases
S phrases
  S8: keep container dry
  S43: In case of fire use dry sand. Never use water.

In general, steel containers are permitted for transport of calcium carbide, e.g., drums or kegs. Under certain conditions transport in cardboard cartons is allowed, provided that the carbide is also sealed in airtight bags.

For the transport of larger amounts of calcium carbide, containers with a capacity of 1 – 20 t have been introduced. Fine granular calcium carbide is also conveyed in tankers that can be discharged pneumatically. For storage and transport in the open, only suitably equipped containers are allowed, not drums or kegs.

6. Environmental Protection

The raw materials used in manufacturing calcium carbide are nonpoisonous and present no hazard to the environment. During production, transport, and preparation of these raw materials, dust can cause a nuisance in the vicinity of the factory and at the workplace and is therefore removed by air extraction and retained by means of filters.

Precautions, especially breathing protection, must be taken when handling hot Söderberg material. The carbide furnaces of the open construction emit lime and coke dust from the surface of the burden. As a rule, open furnaces are fitted with large extraction hoods so that these emissions, together with lime dust arising during tapping, can be extracted and filtered. Closed furnaces emit dust only on tapping. This dust, because of the extremely small particle size, makes high demands on the filtration equipment.

Carbon monoxide gas from open furnaces burns directly at the surface of the feed material to harmless carbon dioxide. From closed furnaces it is usually used as fuel gas. The traces of cyanide in the furnace gas are converted into harmless compounds by special treatment in the gas purification equipment. The residues from wet purification are usually dumped.

7. Uses

Calcium carbide has three main uses: the manufacture of acetylene (→ Acetylene, Chap. 4.3.4.), reaction with nitrogen to give cyanamide, see → Cyanamides, and the desulfurization of pig iron, cast iron, and steel. Since 1960 the use of carbide-based acetylene has declined worldwide as a raw material for heavy chemicals due to the widespread use of petrochemicals. Only a few chemical products are still produced from acetylene. These include butynediol, a component of plastics and fibers, and acetylene black, used in dry batteries.

Acetylene from calcium carbide still holds a firm position in the field of gas welding. The

acetylene cylinder, in spite of the rapid development of techniques of joining materials, is indispensable, especially in pipeline construction, on account of its adaptability and mobility.

Calcium carbide has an important application in metallurgy because it reacts with sulfur. The increasingly strict standards for steel have led to requirements for lower sulfur contents. The lowest sulfur contents can be achieved by desulfurizing the hot metal before the converter. Injection of calcium carbide into the melt was introduced in 1970 and is carried out in a transport ladle. Generally the sulfur content can be reduced from 0.2 % to 0.001 %. Since the reaction of calcium carbide with sulfur takes place on the surface of the solid carbide, the calcium carbide should be ground to $< 100\,\mu m$. Combinations of calcium carbide with magnesium are increasing in importance. A gas-evolving component is added to the desulfurization mixture to improve the reactivity of the calcium carbide in the hot metal [40], [41].

Consumption of calcium carbide for hot metal desulfurization in 1997 was: Europe 100 000 t/a; USA 15 000 t/a; India, Japan, and South Africa 30 000 t/a.

In foundry technology, coarse-grained calcium carbide is used in combination with additives such as fluorspar. The calcium carbide is stirred into the ladle after tapping and before casting.

8. Economic Aspects

Between 1962 and 1965 the world output of calcium carbide reached its maximum, nearly 10×10^6 t/a. Since then the decline has been steady. This decline has been accompanied by a shift of the main production centers.

Table 3 lists production for the years 1962, 1972, 1982, and 1990. In the 1960s and 1970s, there was a severe decline in North America, Western Europe, and Japan on account of conversion to petroleum-based products and reduced output of calcium cyanamide for fertilizer, whereas Eastern Europe showed an increase in calcium carbide production. In the 1980s, however, replacement of calcium carbide by petrochemical raw materials began in these countries, too. The rapid decline in calcium carbide production in the last few years has been due to the increasing availability of petroleum in many countries, including South Africa. Production in the former German Democratic Republic has ceased. In Central and South America small installations still take advantage of favorable local conditions. In Asia, especially China, calcium carbide production has increased in the last few decades.

Table 3. World calcium carbide production, 1000 t/a

Region	1962	1972	1982	1990
Western Europe	2540	1410	660	500
Eastern Europe	2200	3300	2600	1710
America	14 00	710	470	570
Asia and Africa	1860	2080	2670	2080
World production	*8000*	*7500*	*6400*	*4860*

The general trend of calcium carbide production will remain stagnant or in a slightly downward direction worldwide.

9. Toxicology

The toxicological dangers of technical-grade calcium carbide arise only as a result of impurities, which on reaction with water give rise to the formation of hydrides of phosphorus, etc. (→ Acetylene, Chap. 8.).

10. References

1. R. Juza, H. U. Schuster, *Z. Anorg. Allg. Chem.* **311** (1961) 62 – 75.
2. G. W. Healy, *J. Met.* **18** (1966) 643.
3. H. H. Franck, *Z. Anorg. Allg. Chem.* **232** (1937) 75 – 111.
4. M. Atoji, *J. Chem. Phys.* **54** (1971) 3514.
5. M. A. Bredig, *J. Phys. Chem.* **46** (1942) 801.
6. M. von Stackelberg, *Z. Phys. Chem. Abt.* **B 9** (1930) 437.
7. N. G. Vannerberg, *Acta Chem. Scand.* **15** (1961) 769.
8. N. G. Vannerberg, *Acta Chem. Scand.* **16** (1962) 1212.
9. O. Kubaschewski, E. L. L. Evans, *Metallurgical Thermochemistry,* Pergamon Press, London 1967.
10. *Winnacker-Küchler,* 3rd ed., **vol. 1,** p. 426; **4th ed., vol. 2,** p. 609.
11. G. Geiseler, W. Büchner, *Z. Anorg. Allg. Chem.* **343** (1966) 286.

12. K. K. Kelley, *Bull. US Bur. Mines* **477** (1950) 101.
13. L. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin-Heidelberg-New York, Verlag Stahleisen, Düsseldorf 1972.
14. W. Rettkowski, Chr. Geilhufe, K. H. Rüdiger, *Chem. Tech. (Leipzig)* **28** (1976) 588.
15. E. Biesalski, H. van Eck, *Angew. Chem.* **41** (1928) 278.
16. O. Ruff, E. Förster, *Z. Anorg. Allg. Chem.* **131** (1923) 321.
17. Bayer, DE 944487, 1954.
18. W. A. Barber, C. L. Sloan, *J. Phys. Chem.* **65** (1961) 20 – 26.
19. S. Kiritani, T. Nishimaki, *CEER Chem. Econ. Eng. Rev.* **16** (1984) no. 5.
20. K. Feldmann, *Chem. Ing. Tech.* **41** (1969) 199.
21. K. Budde, A. Strauss, B. Schmidt, *Chem. Tech. (Leipzig)* **28** (1976) 585.
22. R. Toussig: *Die Industrie des Calciumcarbides,* Knapp, Halle 1930.
23. J. Wotschke: *Grundlagen des elektrischen Schmelzofens,* Knapp, Halle 1933.
24. E. Kluss: *Einführung in die Probleme des elektrischen Lichtbogen- und Widerstandsofens,* Springer Verlag, Berlin 1951.
25. E. Kluss, *Z. Elektrotech.* **72** (1951) 115.
26. AG für Stickstoffdünger, DE 545696, 1925; DE 571711, 1928.
27. A. Helfenstein, DE 226956, 1910.
28. Det Norske Aktieselskab for Elektrokemisk Industrie, DE 324741, 1919.
29. Union Carbide Corp., US 2996360, 1958.
30. D. E. Hamby, *J. Met.* **19** (1967) 45.
31. Elkem, DE 3 141 108, 1983.
32. W. Weiss, L. Bökelmann, L. Iltzsche, H. Kleiner, *Chem. Tech. (Leipzig)* **36** (1984) 89.
33. SKW Trostberg AG, DE-OS 19 611 119, 1997.
34. Denki Kagaku Kogyo, DE 4 210 387, 1994.
35. Verordnung über Acetylenanlagen und Calciumcarbidlager 27. 02. 1980, BGBl. I, p. 173.
36. Technische Regeln für Acetylenanlagen und Calciumcarbidlager (TRAC), July 1971.
37. European agreement concerning the international carriage of dangerous goods by road (ADR), 07. 1968.
38. Agreement concerning the international carriage of dangerous goods by rail as annex I of the international agreement on railroad freight traffic (RID, CIM), 09. 1977, revised 01. 09. 1983.
39. International regulation for inland navigation and for ocean navigation and for aviation (IATA).
40. SKW Trostberg, DE 2 741 588, 1977.
41. SKW Trostberg, Thyssen Stahl, EP 0 164 592, 1985.

Calcium Carbonate → Lime and Limestone

Calcium Chloride

Robert Kemp, Allied Corp., Morristown, New Jersey, 07960, United States

Suzanne E. Keegan, Allied Corp., Morristown, New Jersey, 07960, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 3
4. **Production** 3
5. **Environmental Protection** 4
6. **Quality Specifications** 4
7. **Chemical Analysis** 4
8. **Storage and Transportation** 5
9. **Uses** 5
10. **Economic Aspects** 6
11. **Toxicology and Occupational Health** . 7
12. **References** 7

1. Introduction

Calcium chloride is found in small quantities, along with other salts, in seawater and in many mineral springs. It is found in higher concentrations in natural brines in California and Michigan (USA), in the mineral antarcticite [*29854-80-6*], $CaCl_2 \cdot 6\,H_2O$; in the yellowcolored mineral tachydrite [*12194-70-6*] (from the Greek word meaning easily soluble), $2\,MgCl_2 \cdot CaCl_2 \cdot 12\,H_2O$; and in the mineral chlorocalcite [*12418-72-3*], $KCaCl_3$. Calcium chloride is reportedly also contained in small amounts in the mineralcarnallite, [*63919-72-2*], $KMgCl_3$ [1–4].

Calcium chloride appears to have been discovered as early as the 15th century but apparently received little attention or study until the latter part of the 18th century. All of the work was done with laboratory prepared samples because it was not commercially produced until after the development of the Solvay ammonia–soda process in the mid-1800s. This process was originally designed to produce soda ash (crude sodium carbonate). Calcium chloride was considered a waste product until uses for it were developed some years later. Its importance has grown continuously to the extent that it is now invaluable for various uses, including ice and dust control, oil well drilling, refrigeration, coal thawing, and food processing. Natural brines account for as much as 70–75 % of United States calcium chloride production in 1983 [5], [6].

2. Physical Properties

Calcium chloride, $CaCl_2$, M_r 110.99, is a white, odorless, extremely water soluble salt that forms hydrates with properties as indicated in Table 1.

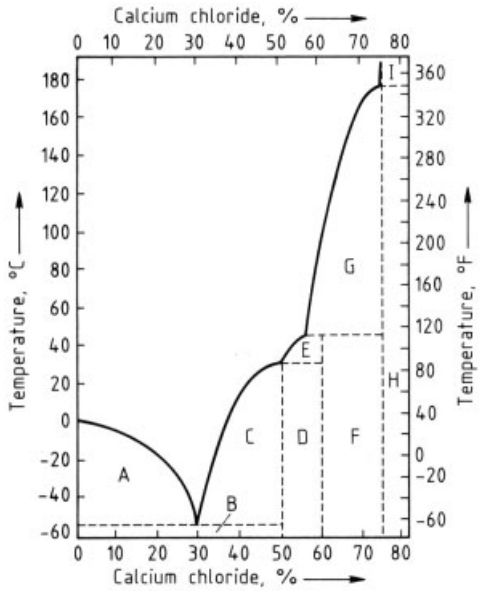

Figure 1. Solubility of pure calcium chloride in water–freezing point diagram (Courtesy Allied Corp.)
Data for areas A–J given in Table 2.

Calcium chloride is highly soluble in water over a wide temperature range, with maximum solubility of approximately 75 % at temperatures above 175 °C. Although calcium chloride dissolves readily in water to form highly concentrated solutions, due account must be taken of the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_547

Table 1. Properties of pure calcium chloride, anhydrous and hydrates (Courtesy Allied Corp.)

Property	$CaCl_2$ [*10043-52-4*]	$CaCl_2 \cdot H_2O$ [*22691-02-7*]	$CaCl_2 \cdot 2H_2O$ [*10035-04-8*]	$CaCl_2 \cdot 4H_2O$ [*25094-02-4*]	$CaCl_2 \cdot 6H_2O$ [*7774-34-7*]
$CaCl_2$, %	100.00	86.03	75.49	60.63	50.66
M_r	110.99	129.01	147.02	183.05	219.08
mp, °C	772	260	176	45	30
bp, °C	1670	–	–	–	–
ϱ (25 °C), g/cm^3	2.22	2.24	1.85	1.83	1.72
Heat of formation (18 °C), kJ/kg	−7190	−8610	−9490	−10980	−11911
Heat of fusion, kJ/kg	256 (772 °C)	–	–	–	170 (30 °C)
Heat of hydration (anhydrous salt), kJ/kg	–	−240	−295	−600	−840
Heat of solution ∗ (hydrate, 291 °C), kJ/kg	−680	−380	−290	−50	+80
Specific heat (hydrate), J K^{-1} kg^{-1}	670	840	1170	1340	1420

∗ One mol solute in 400 mol of water, except monohydrate, which is in 300 mol

temperature at which solid hydrates precipitate from these solutions. A detailed diagram of the solid and liquid phases, showing temperatures and concentrations of calcium chloride and water, is given in Figure 1. The solubility data in this diagram were obtained with pure calcium chloride. The saturation curve and the freezing points of solutions of commercial calcium chloride up to a concentration of 29.8 % $CaCl_2$ are practically the same as for pure solutions; above 29.8 % $CaCl_2$, the curve follows the same general pattern (and direction) but falls slightly to the left of the curve for pure calcium chloride. The difference in the curves is small between 29.8 % and 50 % $CaCl_2$ but increases somewhat at higher concentrations.

The effect of a small percentage of sodium chloride (NaCl) and other minor impurities is to raise the saturation temperature slightly for any given concentration of calcium chloride. Thus, a 40 % $CaCl_2$ solution becomes saturated when cooled to ca. 13 °C, whereas a pure 40 % $CaCl_2$ solution is saturated at ca. 12 °C.

Calcium chloride forms solutions with high densities. This property is commonly used to determine the strength of calcium chloride brines. Density data for commercial calcium chloride solutions are presented in Table 3.

Calcium chloride is extremely hygroscopic and deliquescent when the vapor pressure of the air equals or exceeds that of the saturated solution at the prevailing temperature. Data on calcium chloride solutions in equilibrium with air at various temperatures and relative humidities are presented in Figure 2.

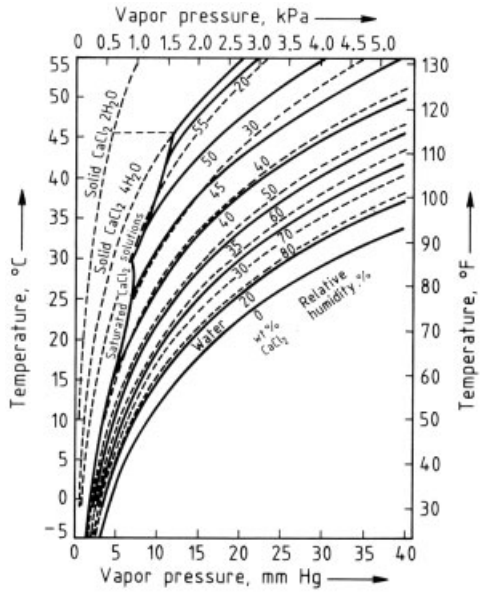

Figure 2. Vapor pressure of calcium chloride solutions (Courtesy Allied Corp.)

Freshly prepared solutions of commercial calcium chloride are somewhat alkaline due to the presence of a small amount of lime (CaO). Although these solutions may exhibit a relatively high pH (pH $\leqq$ 10.3), they have little acid-neutralizing value because of the small percentage of lime present.

Calcium chloride solutions exposed to air gradually absorb carbon dioxide to form calcium

Table 2. Data for Figure (Courtesy Allied Corp.)

Area	Saturation curve for	wt % $CaCl_2$		Temperature °C			Liquid phase contains wt % $CaCl_2$		Solid phases	Solid phase transition at	
		from	to	at and below	from	to	from	to		wt % $CaCl_2$	°C
A	ice	0	29.8		0	−55	0	29.8	ice		
B		0	50.7	−55			none		ice $CaCl_2 \cdot 6H_2O$	29.8	−55
C	$CaCl_2 \cdot 6H_2O$	29.8	50.7		−55	+29.8	29.8	50.3	$CaCl_2 \cdot 6H_2O$		
D		50.7	60.6	+29.8			none		$CaCl_2 \cdot 6H_2O$ $CaCl_2 \cdot 4H_2O$	50.3	+29.8
E	$CaCl_2 \cdot 4H_2O$	50.3	60.6		+29.8	45.3	50.3	56.6	$CaCl_2 \cdot 4H_2O$		
F		60.6	75.5	45.3			none		$CaCl_2 \cdot 4H_2O$ $CaCl_2 \cdot 2H_2O$	56.6	45.3
G	$CaCl_2 \cdot 2H_2O$	56.6	75.5		45.3	175.5	56.6	74.9	$CaCl_2 \cdot 2H_2O$		
H		75.5	86.0	175.5			none		$CaCl_2 \cdot 2H_2O$ $CaCl_2 \cdot H_2O$	74.9	175.5
I	$CaCl_2 \cdot H_2O$	74.9	86.0		175.5	260.0	74.9	86.0	$CaCl_2 \cdot H_2O$		
J	Area above saturation curve unsaturated solutions								none		

carbonate from the lime, thus reducing the pH of freshly prepared solutions [5], [7].

Table 3. Densities and crystallizing temperatures of commercial calcium chloride solutions (Courtesy Allied Corp.)

c_{CaCl_2}, wt %	ϱ (15.5 °C), g/cm^3	Crystallizing temperature, °C
0	0.999	0
2	1.017	− 0.9
4	1.033	− 1.9
6	1.051	− 2.9
8	1.068	− 4.1
10	1.086	− 5.4
12	1.104	− 7.1
14	1.124	− 9.2
16	1.144	−11.6
18	1.164	−14.5
20	1.185	−18.0
25	1.238	−29.4
30	1.294	−46.0
35	1.349	− 9.8
40	1.409	+13.3

∗ Courtesy Allied Chemical Corp.

3. Chemical Properties

Calcium chloride is a source of soluble inorganic calcium and reacts with carbonates, fluorides, and sulfates to form insoluble or moderately soluble salts. Calcium chloride forms water soluble compounds with ammonia, e.g., $CaCl_2 \cdot 8NH_3$, and with alcohol, e.g., $CaCl_2 \cdot C_2H_5OH$. Calcium chloride reacts with sodium tungstate [*13472-45-2*], Na_2WO_4, to form calcium tungstate [*7790-75-2*], $CaWO_4$, also known as synthetic scheelite [7–11].

4. Production

Calcium chloride is produced in commercial quantities by several processes: (1) refining of natural brines, (2) the reaction of calcium hydroxide with ammonium chloride in Solvay (synthetic) soda ash production, and (3) the reaction of hydrochloric acid with calcium carbonate. The refining of natural brines and calcium chloride recovered from synthetic soda ash production together account for over 90 % of calcium chloride production [6].

Natural brines in California and Michigan contain a mixture of chloride salts of calcium, magnesium, and sodium. Magnesium is removed by precipitating magnesium hydroxide, $Mg(OH)_2$, with lime. The solution remaining after filtration is concentrated. Sodium chloride precipitates because it is only slightly soluble in calcium chloride solutions.

In the *ammonia – soda* (Solvay) *process,* ammonia [*7664-41-7*] serves as a catalystfor the

reaction of sodium chloride [*7647-14-5*] with calciumcarbonate [*471-34-1*] (→ Sodium Carbonates). Although this process is actually quite complicated, it can besummarized as follows:

$$2\,NaCl + CaCO_3 \xrightarrow{NH_3} Na_2CO_3 + CaCl_2$$

The calcium chloride solutions produced by the purification of natural brines and by the Solvay process are relatively dilute. Part of the solution is concentrated by evaporation to 30 – 45 % calcium chloride and marketed as such. An additional part is concentrated to ca. 75 % solids, corresponding to calcium chloride dihydrate, $CaCl_2 \cdot 2\,H_2O$. This material is flaked and dried to the commercial calcium chloride dihydrate (77 – 82 % $CaCl_2$) and anhydrous (94 – 97 % $CaCl_2$) products [12–14]. Anhydrous pellets can be produced by substituting an agglomeration procedure in place of the flaker [15], [16].

5. Environmental Protection

Calcium chloride is not generally considered harmful to the environment. However, as with any chemical, calcium chloride spills should be promptly contained and disposed of in accordance with local ordinances.

The taste threshold of calcium chloride in potable water is reported as 150 – 350 ppm. However, depending on the calcium hardness of the water, even concentrations as low as 50 ppm can be objectionable [17].

In high concentrations (>1 000 ppm), calcium chloride retards plant growth and can damage foliage. Because calcium is a nutrient for plants, these effects are most likely caused by accumulation of the chloride ion in the plant tissue. High chloride concentrations are seldom found in U.S. water supplies, even in areas of high salt usage for dust and ice control [17–19].

Calcium chloride concentrations of 10 000 to 20 000 ppm in water have been shown to be hazardous to animals and fish because of differences in osmotic pressure. The effects vary from reduced growth rate and impaired reproduction to death. Such high concentrations are readily detected by the presence of an unpleasant saline taste. In general, the environmental effects of calcium chloride are quite similar to those of rock salt (sodium chloride) [17], [20].

6. Quality Specifications

The major uses of calcium chloride (ice and dust control) do not require high chemical purity. However, the calcium chloride must be free of materials that are harmful to the environment. Commercial calcium chloride from the three main processes contains low levels of heavy metals and organic compounds and is relatively innocuous.

Standard grades of calcium chloride are listed in Table 4 [21–23].

Total alkali chlorides are generally the impurities of concern to calcium chloride producers. These impurities, essentially sodium chloride with smaller amounts of potassium chloride, are known to be harmful when included in cement. Because calcium chloride is utilized in cement, its total alkali chlorides content must be kept at a reasonable level.

In general, commercial calcium chloride meets the requirements of the Food Chemicals Codex. Because of the high quality of commercial calcium chloride, the FDA has granted variances covering some commercial grades of calcium chloride for use in vegetable brining and apple dipping. However, direct addition of commercial grades of calcium chloride to foods, such as cheese, continues to require food-grade material.

7. Chemical Analysis

Routine analytical tests performed on calcium chloride include: assay, total alkali chlorides, percent sodium chloride, percent potassium chloride, and various granulation tests [24–27].

The *concentration* of calcium chloride is most easily determined by an ethylenediaminetetraacetic acid (EDTA) titration. This procedure determines total calcium, which is expressed as calcium chloride. In the determination of total alkali chlorides, sodium chloride, and potassium chloride, atomic absorption spectroscopy is most widely used. Sodium and potassium are determined individually, and these results are combined and expressed as total alkali chlorides.

Table 4. Calcium chloride standards

	ASTM [a] D 98-80			AASHTO [b] M 144-78		Food Chemicals Codex [c]	
	77 %	90 %	94 %	77 %	94 %	77 %	94 %
Calcium chloride, % min.	77	90	94	77	94	74.7 to 80.8	93
Total alkali chlorides, % max. as NaCl	6.8	8	8	2	5	–	–
Magnesium, % max. as $MgCl_2$	0.43	0.5	0.5	0.5	0.5	–	–
Magnesium and alkali salts, % max.	–	–	–	–	–	4	5
Other impurities excluding water, % max.	0.85	1	1	1	1	–	–
Arsenic, ppm max. as As	–	–	–	–	–	3	3
Fluoride, % max. as F	–	–	–	–	–	0.004	0.004
Heavy metals, % max. as Pb	–	–	–	–	–	0.002	0.002
Lead, ppm max. as Pb	–	–	–	–	–	10	10

[a] American Society of Testing Materials specification is based on 90.5 % calcium chloride; for comparison, the values shown have been adjusted and are on an "as received" basis.
[b] American Association of State Highway & Transportation Officials.
[c] Food Chemicals Codex actually specifies a range of 99 – 107 % for calcium chloride dihydrate assay; for comparison, this range has been converted to actual calcium chloride content

8. Storage and Transportation

Calcium chloride is not classified as a hazardous material by the DOT and is, therefore, not subject to specific handling regulations [28]. The various modes of transportation include: bulk rail cars (30 – 90 t), bulk trucks (up to 20 t), box rail cars, and van and flatbed trailers. Depending on whether the shipments are in liquid or dry form, there are several variations to the above general classifications.

Dry bulk calcium chloride can be stored in bins fabricated from most construction-grade steel materials. Care should be taken to minimize moisture. Venting should be limited to times of filling and discharging calcium chloride from the storage bin. Liquid calcium chloride can be stored in either horizontal or vertical cylindrical tanks constructed of steel. Fiberglass and plastic may also be used within limits of strength and temperature.

Because calcium chloride is hygroscopic, the common safety precautions involved in the handling of chemicals should be observed: wearing gloves, boots, long-sleeve shirts, and safety glasses, and if dry products are being handled, dust protection must be insured.

9. Uses

Calcium chloride's versatility is related to its unique combination of physical properties: moisture attraction and retention, high solubility, high heat of solution, and freezing point depression of solutions. Calcium chloride is most widely known for its deicing and dust-controlling abilities.

The largest market for calcium chloride (30 % of total production) is for *deicing* of roads, sidewalks, and parking lots (see Table 5). Calcium chloride melts ice at temperatures as low as −51 °C (−60 °F). Because it liberates heat when exposed to moisture, ice is melted quickly, usually within 15 – 30 min of application. Calcium chloride is also used in conjunction with rock salt (sodium chloride) to enhance and sustain the effectiveness of the rock salt.

Table 5. Calcium chloride end uses in the United States

	%
Deicing	30
Dust control, road stabilization	25
Industrial (refrigerant, coal thawing, etc.)	15
Oil and gas drilling fluids	10
Concrete	5
Tire ballast	4
Miscellaneous	11

Dust control accounts for ca. 25 % of calcium chloride production. Its hygroscopic and deliquescent properties make it ideal for this use. Calcium chloride absorbs moisture from the air and forms a solution that acts to coat dust particles and bind them together, greatly reducing dusting. Calcium chloride solutions are slow to evaporate because of their low vapor pressure

and are, therefore, useful in the compaction of road surfaces.

Calcium chloride is used in the *cement* and *concrete* industries (→ Cement and Concrete). Addition of 1 – 2 % calcium chloride accelerates the setting time of concrete, resulting in earlier strength development. Calcium chloride should not be considered as an antifreeze in concreting; however, the addition of calcium chloride to concrete mixes poured at temperatures below 21 °C (70 °F) largely offsets the retarding effects of the lower temperatures [29].

Calcium chloride solutions, because of their low freezing points, are used extensively as heat-transfer media in *food processing* (→ Foods, 2. Food Technology). Calcium chloride brine greatly increases the heat-transfer rate compared to chilled air or to a sodium chloride brine. Contact time between the brine and the various food molds is decreased, resulting in higher production rates [30].

Calcium chloride is also used in the food industry to increase the firmness of fruits and vegetables and to prevent spoilage during processing.

In the *petroleum* and *petrochemical industries* calcium chloride is used as a desiccant for hydrocarbons. It is also used in drilling fluids, packer fluids, completion fluids, and workover fluids in oil well drilling.

Other uses for calcium chloride include:

Adhesives	– humectant – lowers gel temperature
Animal feed supplement	– source of calcium
Tractor tire weight	– lowers freezing temperature of water – improves traction by increasing tire weight
Highway shoulder and base stabilization	– retains moisture which im proves compaction of soils
Freezeproofing of coal	– lowers freezing point of residual water
Paper manufacture	– increases web strength of corrugated media
Rubber manufacture	– coagulates latex emulsions
Steel and pig iron manufacture	– eliminates alkalies that attack furnace refractory
Wastewater treatment	– precipitates fluorides – breaks oil emulsions – densifies floc

10. Economic Aspects

The majority of the world's calcium chloride is a byproduct from synthetic (Solvay) soda ash production. About 1 t of calcium chloride (anhydrous basis) is produced with each ton of synthetic soda ash. Synthetic soda ash production worldwide was ca. 20.8×10^6 t in 1979, and $31{,}5 \cdot 10^6$ t in 1993. A similar quantity of calcium chloride was also produced. However, most of this material is produced as a 5 – 10 % waste solution and is simply discarded. World production of calcium chloride (waste and refined) from the Solvay process is shown by continent in Table 6 [31].

Table 6. World calcium chloride production from the Solvay process (1979)

Continent	Production (100 % basis), 10^3 t
Europe *	15 480
Asia	3 390
North America	1 460
Australia	290
South America	210
Africa	10
Total	20 840

* Includes the USSR

In the United States, synthetic soda ash has been largely replaced by natural material from Sweetwater County, Wyoming. As a result, natural brines now account for ca. 70 % of the U.S. calcium chloride production. The major calcium chloride producers and their capacities are shown in Table 7.

Table 7. United States calcium chloride producers (1983)

Producer	Capacity (100 % basis), $\times 10^3$ t
Allied	263
Dow	900
Hill Brothers	4
Leslie Salt	27
National Chloride	13
Occidental	6
Reichhold	6
Texas United	32
Wilkinson Salt	15
Total	1266

Because of the widespread use of calcium chloride in deicing and dust control, demand for this chemical is erratic (demand, 100 % basis, 1000 t):

Year	1975	1979	1980	1981	1982
Demand	550	690	550	670	590

Demand is expected to grow at a rate of 3 % per year through 1986 [6], [32].

11. Toxicology and Occupational Health

Calcium chloride is not generally considered toxic; however, prolonged exposure may be detrimental to health. Contact with skin may cause irritation. Calcium chloride can irritate or burn eyes. Inhalation of product dust may irritate nose, throat, or lungs. Ingestion could irritate the mouth, esophagus, or stomach. In rats calcium chloride has an acute oral LD_{50} of 1 g/kg. No TLV or MAK has been established for calcium chloride in air [33].

12. References

1. J. C. Bailar, H. J. Emeleus, R. Nyholm, A. F. Trotman Dickenson (ed.): *Comprehensive Inorganic Chemistry,* 1st ed., **vol. 1,** Pergamon Press, Oxford 1973, pp. 630 – 633.
2. C. Palache, H. Berman, C. Frondel: *The System of Mineralogy,* 7th ed., **vol. 2,** J. Wiley & Sons, New York 1951, pp. 91 – 96.
3. G. E. Dunning, J. F. Cooper, Jr., *Am. Mineral.* **54** (1969) 1018 – 1025.
4. J. F. Thorpe, M. A. Whiteley: *Thorpe's Dictionary of Applied Chemistry,* 4th ed., **vol. 2,** Longmans, Green & Co., New York 1938, pp. 212 – 214.
5. "Calcium Chloride," *Technical and Engineering Service Bulletin No. 16* , Allied Corp., Morristown, N.J., 1978.
6. "Calcium Chloride," *Chemical Products Synopsis,* Mansville Chemical Products, Mansville, N.Y., June, 1983.
7. National Research Council: *International Critical Tables,* McGraw-Hill, New York 1928.
8. G. F. Hüttig, *Z. Anorg. Allg. Chem.* **123** (1922) 31 – 42; *Chem. Abstr.* **16** (1922) 4153.
9. W. H. Nebergall, F. C. Schmidt, H. F. Holtzclaw, Jr.: *College Chemistry with Qualitative Analysis,* 2nd ed., D. C. Heath & Co., Boston 1963, pp. 654 – 655.
10. H. Stephen, T. Stephen (ed.): *Solubilities of Inorganic and Organic Compounds,* **vol. 1,** Macmillan Publ. Co., New York 1963, Part 1, pp. 243 – 261.
11. D. J. Treskon: "Tungsten," *Chemical Economics Handbook,* Stanford Research Institute, Menlo Park, Calif., 1977, Section 789.1000 – 789.1001 M.
12. T.-P. Hou: *Manufacture of Soda With Special Reference to the Ammonia Process,* 2nd ed., Reinhold Publ. Co., New York 1942, pp. 252 – 253.
13. Solvay Process Co., CA 314878, 1927 (C. Sundstrom).
14. Dow Chemical Co., US 1660053, 1928 (A. K. Smith).
15. Dow Chemical Co., US 2646343, 1953 (W. R. Bennett, L. N. Carmouche).
16. Dow Chemical Co., US 4076776, 1978 (W. G. Moore).
17. J. E. McKee, H. W. Wolf (ed.): *Water Quality Criteria,* 2nd ed., The Resources Agency of California State Water Quality Control Board, Publication No. 3-A, pp. 152 – 153.
18. F. W. Holmes, J. H. Baker: "Salt Injury to Trees," II. Sodium and Chloride in Roadside Sugar Maples in Massachusetts," *Phytopathology* **56** (June, 1966) no. 6, 633 – 636 (reprint).
19. F. E. Hutchinson: *The Influence of Salts Applied to Highways on the Levels of Sodium and Chloride Ions Present in Water and Soil Samples,* U.S. Department of the Interior, Project No. A-007-ME, U.S. Gov't. Printing Office, Washington, D.C., June, 1969.
20. "Primary Skin Irritation Test-Rabbits," Hazelton Laboratories, Inc., March 1963, unpublished paper.
21. R. P. Lukens, J. L. Cornillot, R. A. Prieman, D. J. Felty et al.: *Annual Book of ASTM Standards,* ASTM, Philadelphia 1981, Part 14, pp. 183 – 186.
22. American Association of State Highway and Transportation Officials, *Standard Specification for Calcium Chloride: M 144-78,* AASHTO, Washington, D.C., 1978.
23. National Research Council: *Food Chemicals Codex,* 3rd ed., National Academy Press, Washington, D.C. 1981, pp. 47 – 49.
24. Allied Corp., Assay of Calcium Chloride, CA 4-12 A, revised June, 1983.
25. Allied Corp., Determination of Sodium and Potassium in Calcium Chloride, CA 4-56, June, 1983.

26. Reference 21, Part 30, pp. 1041 – 1046.
27. Reference 21, Part 35, pp. 644 – 647.
28. General Services Administration: "Transportation," *Title 49 Code of Federal Regulations,* U.S. Gov't. Printing Office, Washington, D.C., 1982, Part 172.
29. National Ready Mixed Concrete Association: *Cold Weather Ready Mixed Concrete,* Silver Spring, Md., 1968, pp. 11 – 12.
30. *ASHRAE Handbook 1981 Fundamentals,* American Society of Heating, Refrigerating, and Air Conditioning Engineers, Inc., Atlanta, Ga., 1981, pp. 18.2 – 18.5.
31. *The Economics of Soda Ash,* 1st ed., Roskill Information Services Ltd., London 1981, pp. 20 – 26.
32. "Calcium Chloride," *Chemical Profiles,* Schnell Publishing Co., New York April, 1982.
33. N. I. Sax: *Dangerous Properties of Industrial Materials,* 5th ed., Van Nostrand Reinhold Co., New York 1979, pp. 460 – 461.

Calcium Cyanamide → Cyanamides
Calcium Fluoride → Fluorine Compounds, Inorganic
Calcium Hydrogencarbonate → Lime and Limestone
Calcium Hydroxide → Lime and Limestone
Calcium Hypochlorite → Chlorine Oxides and Chlorine Oxygen Acids
Calcium Nitrate → Nitrates and Nitrites
Calcium Oxide → Lime and Limestone
Calcium Phosphate → Phosphoric Acid and Phosphates

Calcium Sulfate

FRANZ WIRSCHING, Gebrüder Knauf Westdeutsche Gipswerke, Iphofen, Federal Republic of Germany

1. Introduction 1
2. The $CaSO_4 - H_2O$ System 2
2.1. Phases 2
2.2. Laboratory Synthesis 4
2.3. Industrial Dehydration of Gypsum 4
2.4. Energy Aspects 5
2.5. Structure, Mixed Compounds, Solubility 5
3. Occurrence, Raw Materials 6
3.1. Gypsum and Anhydrite Rock 6
3.2. Flue Gas Desulfurization (FGD) Gypsum 7
3.3. Phosphogypsum 9
3.4. Fluoroanhydrite 10
3.5. Other Byproduct Gypsums 10
4. Production 10
4.1. Natural Gypsum to Calcined Products 10
4.2. FGD Gypsum to Calcined Products 16
4.2.1. β-Hemihydrate and Multiphase Plasters from FGD Gypsum 16
4.2.2. α-Hemihydrate Plaster from FGD Gypsum 17
4.3. Phosphogypsum to Calcined Products 18
4.4. Anhydrite Plaster 19
5. Use and Properties of Gypsum Plasters and Products and Anhydrite Plasters 20
5.1. Hydration, Setting, Hardening . . . 20
5.2. Prefabricated Gypsum Building Components 22
5.3. Gypsum Plaster 23
5.4. Other Uses 24
5.5. Properties of Gypsum Building Products Installed in Situ 26
6. Material Testing and Chemical Analysis 27
6.1. Standards 27
6.2. Testing 27
6.3. Chemical Analysis 28
6.4. Phase analysis 28
7. Economic Aspects 28
8. Recycling and Disposal 29
9. References 29

1. Introduction

Both gypsum [*13397-24-5*], $CaSO_4 \cdot 2\,H_2O$ [*10101-41-4*], and anhydrite [*14798-04-0*], $CaSO_4$ [*7778-18-9*], are widely distributed in the earth's crust. Only in volcanic regions are gypsum and anhydrite rock completely absent. In addition, large quantities of gypsum and anhydrite are obtained from industrial processes when flue gases are desulfurized or calcium salts are reacted with sulfuric acid. In these cases calcium sulfate is usually obtained as a moist, fine powder.

Gypsum is useful as an industrial material because (1) it readily loses its water of hydration when heated, producing partially or totally dehydrated calcined gypsum, and (2) when water is added to this calcined gypsum, it reverts to the original dihydrate – the set and hardened gypsum product. These two phenomena, dehydration and rehydration, are the basis of gypsum technology:

Dehydration

$$CaSO_4 \cdot 2\,H_2O + \text{heat} \longrightarrow CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O$$

$$CaSO_4 \cdot 2\,H_2O + \text{heat} \longrightarrow CaSO_4 + 2\,H_2O$$

Rehydration

$$CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O \longrightarrow CaSO_4 \cdot 2\,H_2O + \text{heat}$$

$$CaSO_4 + 2\,H_2O \longrightarrow CaSO_4 \cdot 2\,H_2O + \text{heat}$$

The apparent density of the rehydrated calcium sulfate is much lower, 500 – 1500 kg/m^3, than the density of the calcium sulfate rock, 2300 kg/m^3.

Gypsum and anhydrite are nontoxic.

Both gypsum plaster and lime were used as mortar in antiquity. Gypsum was called *gatch* in Persian, *gypsos* in Greek, and *gypsum* in Latin. The Iranians, Egyptians, Babylonians, Greeks, and Romans were familiar with the art of working with gypsum plaster, examples being the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a04_555

walls of Jericho, the pyramid of Cheops, the palace of Knossos, and the decorated interior walls of Pompeii. In Germany gypsum plaster was used as mortar in walls and buildings during the early Middle ages, e.g., the monastery at Walkenried in the Harz Mountains. It gained popularity tremendously, reaching its peak during the Baroque and Rococo periods. Examples are the Wessobrunn school for stucco workers [17] and the stucco decorations in Charlottenburg palace, Berlin. The expansion of the cement industry in the second half of the nineteenth century also considerably increased the use of gypsum.

Over the centuries the gypsum industry has developed empirically out of the old craft of gypsum plastering. The distinctions between gypsum plaster and lime, however, remained obscure up to the eighteenth century. Research into the principles of gypsum technology was begun in 1765 by LAVOISIER, and has continued to this day. However, a craft so firmly steeped in tradition was slow to accept scientific conclusions, and only in the last few decades has gypsum manufacture developed into a modern industry [1].

2. The $CaSO_4 - H_2O$ System

2.1. Phases

The $CaSO_4 - H_2O$ system is characterized by five solid phases. Four exist at room temperature: calcium sulfate dihydrate, calcium sulfate hemihydrate, anhydrite III, and anhydrite II. The fifth phase, anhydrite I, only exists above 1180 °C [19], and it has not proved possible to produce a stable form of anhydrite I below that temperature. Table 1 characterizes the phases in the $CaSO_4 - H_2O$ system. The first four phases are of interest to industry.

Calcium sulfate dihydrate, $CaSO_4 \cdot 2\,H_2O$, is both the starting material before dehydration and the final product after rehydration. Calcium sulfate hemihydrate [*10034-76-1*], $CaSO_4 \cdot 1/2\,H_2O$, occurs in two different forms, α and β, representing two limiting states [20]. They differ from each other in their application characteristics, their heats of hydration, and their methods of preparation (see Table 8) [21]. The α-hemihydrate (Fig. 1) consists of compact, well-formed, transparent, large primary particles. The β-hemihydrate (Fig. 2) forms flaky, rugged secondary particles made up of extremely small crystals.

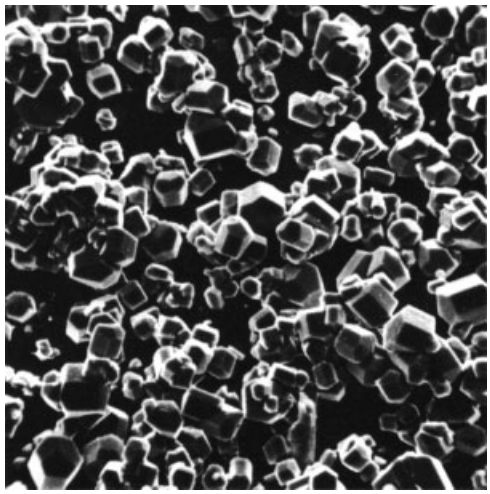

Figure 1. α-Hemihydrate (scanning electron micrograph) 200×

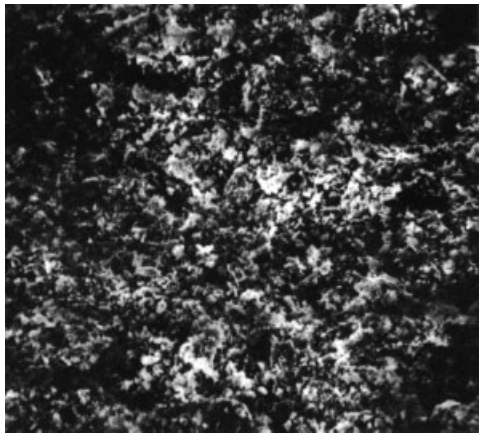

Figure 2. β-Hemihydrate (scanning electron micrograph) 200×

LEHMANN et al. [22] postulated three limiting stages for anhydrite III, also known as soluble anhydrite: β-anhydrite III, β-anhydrite III$'$, and α-anhydrite III. The three stages of anhydrite III were characterized by X-ray analysis, differential thermal analysis, scanning electron microscopy, mercury porosimetry, and measurement of the specific surface area. Anhydrite II is the naturally occurring form and also that

Table 1. Phases in the $CaSO_4$–H_2O system

Characteristic	Calcium sulfate dihydrate	Calcium sulfate hemihydrate	Anhydrite III	Anhydrite II	Anhydrite I
Formula	$CaSO_4 \cdot 2H_2O$	$CaSO_4 \cdot 1/2\,H_2O$	$CaSO_4$	$CaSO_4$	$CaSO_4$
Molecular mass M_r	172.17	145.15	136.14	136.14	136.14
Thermodynamic stability, °C	< 40	metastable∗	metastable∗	40 – 1180	> 1180
Forms or stages		two forms: α β	three stages: β-anhydrite III β-anhydrite III′ α-anhydrite III	three stages: AII-s, slowly soluble anhydrite AII-u, insoluble anhydrite AII-E, *Estrichgips*	
Other names, often based on the application	gypsum raw gypsum synthetic gypsum chemical gypsum byproduct gypsum set gypsum hardened gypsum FGD gypsum	*α-form:* α-hemihydrate autoclave plaster α-plaster *β-form:* β-hemihydrate stucco plaster β-plaster plaster of Paris [*26499-65-0*]	soluble anhydrite	raw anhydrite natural anhydrite anhydrite synthetic anhydrite chemical anhydrite byproduct anhydrite calcined anhydrite	high-temperature anhydrite
Synthesis conditions: temperature, °C, and atmosphere	< 40	α-form: > 45, from aqueous solution β-form: 45 – 200 in dry air	α- and β-AIII: 50 and vacuum or 100 in air β-AIII′: 100 in dry air	200 – 1180	> 1180
Production temperature, °C	< 40	α-form: 80 – 180 β-form: 120 – 180	β-AIII and β-AIII′: 290 α-AIII: 110	300 – 900, specifically AII-s: < 500 AII-u: 500 – 700 AII-E: > 700	not produced commercially

∗Metastable in dry air.

Table 2. Physical properties of the $CaSO_4 \cdot x\,H_2O$

Property	Calcium sulfate	Calcium sulfate hemihydrate		Anhydrite III	Anhydrite II	Anhydrite I
	dihydrate	α-form	β-form			
Water of crystallization, wt %	20.92	6.21	6.21	0.00	0.00	0.00
Density ϱ, g/cm^3	2.31	2.757	2.619 – 2.637	2.580	2.93 – 2.97	undetermined
Hardness, Mohs	1.5	–	–	–	3 – 4	–
Solubility in water at 20 °C, g $CaSO_4$ per 100 g of solution	0.21	0.67	0.88	hydrates to the hemihydrate	(0.27)	–
Refractive indices						
n_α	1.521	1.559∗		1.501	1.570	undetermined
n_β	1.523	1.5595∗		1.501	1.576	
n_γ	1.530	1.584		1.546	1.614	
Optical character	+			+	+	undetermined
Optical orientation	$n_\beta \parallel b$ $n_{\gamma/c} = 52°$	$c \parallel n_\gamma$		$c \parallel n_\gamma$	$n_\alpha \parallel c$ $n_\beta \parallel c$	undetermined
Axial angle $2V$	58 – 60 °	14 °		≈ 0 °	42 – 44 °	undetermined
Lattice symmetry	monoclinic	monoclinic		orthorhombic	orthorhombic	cubic
Space group	*I* 2/*a*	*I* 121		*C* 222	*Amma*	undetermined
Lattice spacing, nm						
a	0.5679	1.20317		1.20777	0.7006	undetermined
b	1.5202	0.69269		0.69723	0.6998	
c	0.6522	1.26712		0.63040	0.6245	undetermined
β	118.43°	90.27°				

∗ Average.

produced from calcining the dihydrate, hemihydrate, and anhydrite III at elevated temperature.

The most important physical properties of the calcium sulfate phases are shown in Table 2.

2.2. Laboratory Synthesis

The thermodynamic stability ranges for the calcium sulfate phases are shown in Table 1. Below 40 °C, i.e., under normal atmospheric conditions, only calcium sulfate dihydrate is stable. The other phases are obtained at higher temperatures by progressive dehydration of the calcium dihydrate in the following order:

dihydrate → hemihydrate → anhydrite III → anhydrite II

Under normal atmospheric conditions hemihydrate and anhydrite III are metastable, and below 40 °C in the presence of water or water vapor they undergo conversion to the dihydrate, as anhydrite II does. However, between 40 °C and 1180 °C anhydrite II is stable.

To synthesize pure phases in the laboratory, β-hemihydrate is made from the dihydrate by heating at a low water-vapor partial pressure, i.e., in dry air or vacuum, between 45 °C and 200 °C. Further careful heating at 50 °C in a vacuum or up to ca. 200 °C at atmospheric pressure produces β-anhydrite III.

At very low water-vapor partial pressure, if water vapor is released rapidly and particle size is small, β-anhydrite III′ forms directly from the dihydrate, without formation of an intermediate hemihydrate. The specific surface area of such β-anhydrite III′ can be up to ten times that of β-anhydrite III.

α-Hemihydrate is obtained from the dihydrate at high water-vapor partial pressure, e.g., above 45 °C in acid or salt solutions, or above 97.2 °C in water under pressure (e.g., 134 °C, 3 bar, 4 h). Further careful release of water at 50 °C in a vacuum or at 100 °C under atmospheric pressure yields α-anhydrite III.

Anhydrite III is difficult to prepare pure because anhydrite II begins to form above 100 °C, and anhydrite III reacts readily with water vapor to form hemihydrate.

The β-hemihydrates from β-anhydrite III and β-anhydrite III′ differ in their physical properties [23]. Therefore, hemihydrates from β-anhydrite III′ should be designated as β-hemihydrate′. α-Anhydrite III absorbs water vapor to form α-hemihydrate. Likewise, the hemihydrates, in humid air, reversibly adsorb up to 2 % of their weight in water without converting to dihydrate. This nonstoichiometric water in the hemihydrate can be completely removed by drying at 40 °C.

Gypsum dehydration kinetics have been investigated by several authors. Neutron and X-ray powder diffraction studies have shown that the dehydration (and hydration) mechanism is strongly topotactic in the temperature range of 20 – 350 °C [24], [25]. In neutron thermodiffractometry experiments it was found that gypsum decomposes to $CaSO_4(H_2O)_{0.5}$, then to anhydrite III, and finally to anhydrite II. With high local steam pressure, a subhydrate with 0.74 H_2O was found [25]. According to another paper three phases of the α-hemihydrate type can be prepared as pure samples: $CaSO_4(H_2O)_{0.6}$, $CaSO_4(H_2O)_{0.5}$, and anhydrite III. The crystal structures of these phases were determined; data for $CaSO_4(H_2O)_{0.5}$ and for anhydrite III are listed in Table 2. The structure of the subhydrate $CaSO_4(H_2O)_{0.6}$ was found to be monoclinic, space group $I121$, with a = 1.19845, b = 0.69292, c = 1.27505 nm and β = 90 °. The density is 2.74 g/cm^3 [26]. In general the hemihydrate $CaSO_4(H_2O)_{0.5}$ is considered to be the kinetically most stable subhydrate. Table 2 also lists the well-established crystallographic data of gypsum [27] and of anhydrite II [28], [29].

Anhydrite II is formed at temperatures between 200 °C and 1180 °C. Above 1180 °C, anhydrite I forms; below 1180 °C it reverts to anhydrite II.

Another mechanism for conversion of gypsum directly to anhydrite II has been found in the catalytic action of small quantities of sulfuric acid on moist, finely divided gypsum at 100 – 200 °C. In this case anhydrite II with orthorhombic crystal structure is produced by neoformation [30].

2.3. Industrial Dehydration of Gypsum

Industrially it is most important that dehydration is achieved in the shortest time with the lowest energy consumption, i.e., that the costs be held to a minimum. Because of kinetic inhibitions calcination is carried out at much higher

temperatures than those used in the laboratory (Table 1). Rarely are pure phases produced during manufacture; rather, mixtures of phases of the $CaSO_4 - H_2O$ system are produced. Three types of calcined anhydrite II (anhydrous gypsum plaster or overburnt plaster) are manufactured, depending on burn temperature and time:

1) Anhydrite II-s (slowly soluble anhydrite), produced between 300 and 500 °C
2) Anhydrite II-u (insoluble anhydrite), produced between 500 and 700 °C
3) Anhydrite II-E (partially dissociated anhydrite; floor plaster, *Estrichgips*), produced above ca. 700 °C

In use the difference among these products lies in the rates of rehydration with water, which for anhydrite II-s is fast, for anhydrite II-u slow, and for anhydrite II-E in between, a little faster than anhydrite II-u (see Fig. 8). Transitions between these different stages of reaction are possible.

Anhydrite II-E consists of a solid mixture of anhydrite II and calcium oxide formed by the partial dissociation of anhydrite into sulfur trioxide and calcium oxide when raw gypsum is heated above 700 °C. The presence of impurities lowers the normal dissociation temperature of anhydrite II, ca. 1450 °C.

2.4. Energy Aspects

KELLY et al. made a thorough study of the thermodynamic properties of the $CaSO_4 - H_2O$ system [31], [32]. Tables 3 and 4 list the heats of hydration and dehydration of the various phase changes that are of industrial significance.

2.5. Structure, Mixed Compounds, Solubility

Crystal Structure. All structures in the $CaSO_4 - H_2O$ system consist of chains of alternate Ca^{2+} and SO_4^{2-} ions. These $CaSO_4$ chains remain intact during phase changes. Calcium sulfate dihydrate has a layered structure, and the water of crystallization is embedded between the layers. When calcium sulfate dihydrate is dehydrated to hemihydrate, a tunnel structure is formed with wide channels parallel to the $CaSO_4$ chains. In the subhydrates the tunnels are filled with water molecules; in anhydrite III they are empty. The relative ease of escape of these water molecules explains the facile conversion from subhydrate to anhydrite III. Anhydrite II exhibits closest packing of ions, which makes it the densest and strongest of the calcium sulfates. However, lacking empty channels, it reacts only very slowly with water. The α- and β-forms of calcium sulfate hemihydrate exhibit no crystallographic differences. The β-forms of calcium sulfate hemihydrate are active states of the α-form [33].

Isomorphism. Isomorphic incorporation of chemical compounds into the lattice of $CaSO_4 - H_2O$ phases is of interest in connection with FGD gypsums and phosphogypsums. Isomorphic incorporation of calcium hydrogenphosphate dihydrate occurs because $CaHPO_4 \cdot 2\,H_2O$ has lattice spacings almost identical with those of $CaSO_4 \cdot 2\,H_2O$ [34]. Monosodium phosphate, NaH_2PO_4, can also be incorporated into the gypsum lattice [35]. KITCHEN et al. [36] consider isomorphism between the ions AlF_5^{2-} and SO_4^{2-} a possibility. EIPELTAUER [20], [37] reports on further isomorphism, incorporation of the anion FPO_3^{2-} as well as the incorporation of Na^+, up to 0.2 % in the hemihydrate lattice, but only 0.02 % in the dihydrate lattice. Chlorides are not incorporated.

Double and Triple Salts. Gypsum forms double salts with alkali-metal and ammonium sulfates, some of which occur in nature, such as syngenite, $CaSO_4 \cdot K_2SO_4 \cdot H_2O$.

There are triple sulfates of calcium with the divalent ions of the iron and zinc subgroups and of manganese, copper, and magnesium along with the univalent alkali metals, also including ammonium. An example is the well-known polyhalite, $2\,CaSO_4 \cdot MgSO_4 \cdot K_2SO_4 \cdot 2\,H_2O$, which also occurs in natural salt deposits.

Ettringite, $3\,CaO \cdot Al_2O_3 \cdot 3\,CaSO_4 \cdot 32\,H_2O$, is important in cement chemistry [38], as are syngenite [39] and thaumasite [40].

Adducts. Gypsum forms sulfuric acid adducts in concentrated sulfuric acid: $CaSO_4 \cdot 3\,H_2SO_4$ and $CaSO_4 \cdot H_2SO_4$ [41].

Table 3. Heat of hydration

Phase change		Heat of hydration per mole (gram) of dihydrate at 25 °C, J
β-$CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O$ (l)	$\longrightarrow CaSO_4 \cdot 2\,H_2O$	$19\,300 \pm 85$ (111.9 ± 0.50)
α-$CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O$ (l)	$\longrightarrow CaSO_4 \cdot 2\,H_2O$	$17\,200 \pm 85$ (100.00 ± 0.50)
β-$CaSO_4$III $+ 2\,H_2O$ (l)	$\longrightarrow CaSO_4 \cdot 2\,H_2O$	$30\,200 \pm 85$ (175.3 ± 0.50)
α-$CaSO_4$III $+ 2\,H_2O$ (l)	$\longrightarrow CaSO_4 \cdot 2\,H_2O$	$25\,700 \pm 85$ (149.6 ± 0.50)
$CaSO_4$II $+ 2\,H_2O$ (l)	$\longrightarrow CaSO_4 \cdot 2\,H_2O$	$16\,900 \pm 85$ (98.0 ± 0.50)

Table 4. Heat of dehydration

Phase change	Heat of dehydration per mole or tonne of dehydration products at 25 °C	
	J/mol	kJ/t
$CaSO_4 \cdot 2\,H_2O \longrightarrow \beta$-$CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O$ (g)	86 700	597 200
$CaSO_4 \cdot 2\,H_2O \longrightarrow \alpha$-$CaSO_4 \cdot 1/2\,H_2O + 3/2\,H_2O$ (g)	84 600	582 700
$CaSO_4 \cdot 2\,H_2O \longrightarrow \beta$-$CaSO_4$III $+ 2\,H_2O$ (g)	121 800	895 700
$CaSO_4 \cdot 2\,H_2O \longrightarrow \alpha$-$CaSO_4$III $+ 2\,H_2O$ (g)	117 400	863 100
$CaSO_4 \cdot 2\,H_2O \longrightarrow CaSO_4$II $+ 2\,H_2O$ (g)	108 600	798 000

Calcium sulfate dihydrate can combine with four molecules of urea to form an addition compound.

Solubility and Supersaturation. Gypsum is slightly soluble in water (see Table 2). The solubilities of the various forms of calcium sulfate are strongly affected by the presence of other solutes. The references [4], [7] can be consulted for particulars on the solubility of gypsum in acids, especially sulfuric acid, phosphoric acid, and nitric acid. Gypsum is readily soluble in glycerol and sugar solutions and in aqueous solutions of chelating agents, such as EDTA. All calcium sulfates have a tendency to supersaturation [4].

3. Occurrence, Raw Materials

3.1. Gypsum and Anhydrite Rock

Gypsum and anhydrite deposits are found in many countries. They originated from supersaturated aqueous solutions in shallow seas, which evaporated and deposited first carbonates, then sulfates, and finally chlorides, i.e., in order of increasing solubility [5], [6].

Gypsum and anhydrite are almost always found in the Permian, Triassic, and Tertiary formations. The Permian includes the Upper Permian (Zechstein deposits, Germany). The Triassic consists of Lower (*Buntsandstein*), Middle (*Muschelkalk*), and Upper Triassic (*Keuper*). The Tertiary deposits include the Eocene (Spain, Persia, Near East), the Oligocene (Paris), and the Miocene deposits (Mediterranean area, North africa). The Jurassic and Cretaceous formations, which lie between the Triassic and Tertiary, are almost devoid of useful gypsum deposits, as are the older pre-Permian formations – at least in the Old World – as, for example, the Carboniferous formation.

The accepted view is that gypsum, $CaSO_4 \cdot 2\,H_2O$, was formed upon initial crystallization. The best known deposits of primary gypsum are those of the Paris Basin and the areas around the Mediterranean. Under increased pressure and thus temperature, anhydrite could have developed locally from the original gypsum [42], [43]. However, uptake of surface water can convert the anhydrite back into gypsum. This gypsum is therefore a secondary rock, especially in formations older than the Middle Tertiary [44].

The various gypsum and anhydrite deposits differ in purity, structure, and color. The major impurities are calcium carbonate (limestone), dolomite, marl, clay; less frequently, silica, bitumen, glauberite, syngenite, and polyhalite.

Table 5. Chemical composition of some calcium sulfates

	Natural gypsums		Natural anhydrites		Phosphogypsums			Fluoroanhydrite
	1	2	3	4	5	6	7	8
Mineralogical analysis								
$CaSO_4 \cdot 2H_2O$	91.6	96.4	6.0	3.8	93.0	97.6	96.5	0.0
$CaSO_4$ II	6.4	0.9	88.8	83.7	0.0	0.0	0.0	94.7
$MgCO_3$	0.1	0.6	2.0	3.6	0.0	0.0	0.0	0.0
$CaCO_3$	1.9	1.3	3.1	5.3	0.0	0.0	0.0	0.0
Chemical analysis								
Combined water	19.1	20.1	1.3	0.8	19.5	19.2	19.2	0.0
SO_3	46.4	45.4	55.1	51.0	43.2	45.4	46.9	56.4
CaO	33.6	32.5	40.3	38.7	32.2	32.5	32.1	40.8
MgO	0.05	0.28	0.95	1.71	0.01	0.01	0.01	0.13
SrO	0.07	0.16	0.07	0.14	0.06	2.05	0.05	0.0
$Fe_2O_3 + Al_2O_3$	0.01	0.08	0.01	0.39	0.27	0.70	0.93	0.27
HCl-insoluble residue	0.10	0.47	0.10	2.49	1.51	1.00	0.68	0.69
Na_2O	0.02	0.01	0.04	0.11	0.47	0.14	0.02	0.03
Total P_2O_5	0.0	0.0	0.0	0.0	1.01	0.11	0.30	0.0
F	0.0	0.0	0.0	0.0	1.76	0.14	0.29	0.92
Others								0.11 ZnO 0.69 K_2O
Organics	0.0	0.0	0.0	0.0	0.08	0.03	0.04	0.0
Ignition loss	0.54	0.71	2.08	4.45	1.38	0.34	0.43	0.14
pH	6–7	6–7	6–7	6–7	3	9	4	10–12

1) Zechstein gypsum; 2) Keuper gypsum; 3) Zechstein anhydrite; 4) Keuper anhydrite; 5) From Moroccan raw phosphate, dihydrate process; 6) From Kola raw phosphate, dihydrate/hemihydrate process, after rehydration to dihydrate; 7) From Moroccan raw phosphate, hemihydrate/dihydrate process; 8) From hydrofluoric acid production.

If these are present in the gypsum and anhydrite from its formation, they are called *primary impurities*. A distinction is drawn between carbonate – sulfate deposits, mainly found in the Upper Permian (Zechstein), and the clay – sulfate deposits, which predominate in the Triassic [45]. *Secondary impurities* are formed during exposure to materials flushed into cracks and leached cavities, but may also be introduced into the rock as waste material during mining. Gypsum is readily soluble in water, even if it is only slightly soluble, and therefore gypsum rock is leached by surface water. The chemical composition of some calcium sulfates is shown in Table 5.

Workable gypsum and anhydrite rock are structurally quite distinct from one another. The most important types of native gypsum are sparry gypsum, also called gypsum spar or flaky gypsum (selenite), fibrous gypsum, alabaster (grainy gypsum), gypsum rock (common gypsum, also massive gypsum), porphyritic gypsum, earthy gypsum (gypsite in the United States), and gypsum sand (contaminated with Glauber's salt $Na_2SO_4 \cdot 10H_2O$). Anhydrite, always crystalline, can either be sparry (anhydrite spar), coarse to close grained, or even rod shaped.

Gypsum can be pure white. If it contains iron oxide, it is reddish to yellowish. If clay and/or bitumen is present, it is gray to black. Very pure anhydrite is bluish white, but usually it is gray with a bluish tinge. The white veins sometimes found on the boundary between gypsum and anhydrite consist largely of glauberite ($CaSO_4 \cdot Na_2SO_4$) or Glauber's salt ($Na_2SO_4 \cdot 10H_2O$).

3.2. Flue Gas Desulfurization (FGD) Gypsum

FGD gypsum is produced from the desulfurization of combustion gases of fossil fuels (coal, lignite, oil) in power stations as a product with a specification and quality standards. The internationally accepted definition of FGD gypsum is: "Gypsum from flue gas desulfurization plants (FGD gypsum, desulfogypsum) is moist, finely

divided, crystalline, high-purity calcium sulfate dihydrate, $CaSO_4 \cdot 2\,H_2O$. It is specifically produced in a flue gas desulfurization process incorporating after lime(stone) scrubbing, a refining process involving oxidation followed by gypsum separation, washing, and dewatering" [46]. FGD gypsum is produced in four stages:

1) Desulfurization: In a first circuit the dedusted flue gas is sprayed with a limestone suspension in counterflow. The reaction, which takes place at pH 7 – 8, produces insoluble calcium sulfite:

$$SO_2\ (g) + CaCO_3\ (s) + 1/2\,H_2O \longrightarrow CaSO_3 \cdot 1/2\,H_2O\ (s) + CO_2\ (g)$$

The following stages (2 – 4) are recovery operations in which the FGD gypsum is formed and isolated.

2) Forced oxidation: In a second circuit the calcium sulfite reacts spontaneously with atmospheric oxygen at pH 5, initially forming soluble calcium bisulfite, which is then oxidized to calcium sulfate dihydrate:

$$CaSO_3 \cdot 1/2\,H_2O\ (s) + O_2\ (g) + 3\,H_2O \longrightarrow 2CaSO_4 \cdot 2\,H_2O\ (s)$$

During this procedure the gypsum crystals grow by continuous circulation to the required average size of 30 – 70 µm.

3) Gypsum separation: In a hydrocyclone, the calcium sulfate dihydrate crystals produced in the quencher slurry are separated and solid impurities are removed.

4) Gypsum washing and dewatering: The calcium sulfate crystals are separated from the process water in filters or centrifuges and washed with clean wash water to remove unwanted water-soluble chlorides and sodium and magnesium salts.

The resulting FGD gypsum contains < 10 wt % free moisture. About 5.4 t of gypsum is produced per tonne of sulfur in the fuel. Figure 3 shows a typical flow diagram of a flue gas desulfurization plant for producing FGD gypsum.

Sometimes FGD gypsum from lignite-fired power stations is darker than that from coal-fired power stations. Therefore a special "overflow cleaning" operation was developed to improve the color of lignite FGD gypsum. Such lignite FGD gypsum meets international quality specifications and can be used for all applications. This could be important, for example in Germany, where lignite FGD gypsum will predominate in the future [47]. A large-scale investigation has been carried out to compare FGD gypsum to natural gypsum in the production, processing, utilization, and disposal of building materials. The conclusion was that the differences in chemical composition and in trace element content between FGD gypsum and natural gypsum are insignificant. FGD gypsum can be used for the production of all building materials [48].

Table 6 lists quality standards for FGD gypsum.

Table 6. Quality standards for FGD gypsum

Quality parameters	Quality standards
Free moisture, wt %	< 10.0
Calcium sulfate dihydrate, wt %	> 95.0*
Magnesium oxide, water soluble, wt %	< 0.10
Sodium oxide, water soluble, wt %	< 0.06
Chloride, wt %	< 0.01
Sulfur dioxide, wt %	< 0.25
pH value	5 – 9
Color	white**
Odor	neutral
Toxicity	nontoxic

*The reduction of the calcium sulfate dihydrate content by inert elements is not detrimental to the different areas of application.
**Different color values may apply depending on the use of the FGD gypsum and the final products.

FGD gypsum is classified as a product and not as a waste, and it is therefore not included in the European Waste Catalogue (EWC). Furthermore, FGD gypsum is not included in the OECD lists for the Control of Transfrontier Movements of Wastes Destined for Recovery Operations. The EU customs code of FGD gypsum is 2520 20 90. The PRODCOM code of FGD gypsum in the international classification of goods for the production statistics of the EU Member States is 2653 10 003. FGD gypsum is used directly in the gypsum industry (and small quantities in the cement industry) without any further treatment, and is equivalent to natural gypsum.

The predicted quantities of FGD gypsum from coal-fired and lignite-fired power stations in Europe, the United States and Canada, and Japan and Taiwan in 1996 are listed in Table 7 [47].

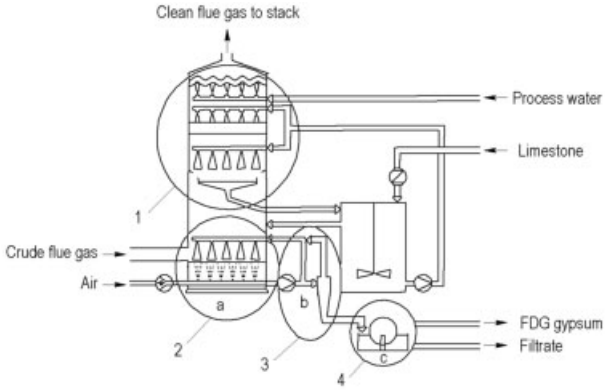

Figure 3. Flow sheet of a flue gas desulfurization plant for producing FGD gypsum with the four operation stages 1 to 4 (see text for explanation) a) Quencher; b) Hydrocyclone; c) Filter

Table 7. Predicted quantities of FGD gypsum in 1996

Country	Quantity, 10^3 t
Germany	4 920
Denmark	340
Finland	190
United Kingdom	1 200
Italy	40
Netherlands	360
Austria	100
Poland	770
Russia	60
Czech Replublic	40
Ukraine	85
Total for Europe	8 225
USA and Canada	3 000
Japan and Taiwan	2 300

In Germany the quantities of FGD gypsum will rise from 4.9×10^6 t in 1996 to 6.3×10^6 t in the year 2000 due to lignite FGD gypsum from new power stations in the former Democratic Republic. In Spain, France, Italy and Turkey, fairly large quantities are to be expected after 1996. The quantities of FGD gypsum produced in Japan and Taiwan of ca. 2.3×10^6 t are expected to remain stable after 1996 [85]. The quantities of FGD gypsum predicted for the United States and Canada will probably be attained later than predicted above, i.e., only after 1996 [49].

3.3. Phosphogypsum

By far the largest amount of byproduct gypsum is obtained from the production of wet phosphoric acid from phosphate rock (usually fluorapatite) and sulfuric acid (see → Phosphoric Acid and Phosphates, Chap. 1.2.2.)

$$\underset{(504)}{Ca_5(PO_4)_3F} + \underset{(5\times98)}{5\,H_2SO_4} + \underset{(10\times18)}{10\,H_2O} \longrightarrow \underset{(5\times172)}{5\,CaSO_4\cdot 2\,H_2O} + \underset{(3\times98)}{3\,H_3PO_4} + \underset{(20)}{HF}$$

Worldwide, about 100×10^6 t/a of phosphogypsum is currently obtained. Phosphogypsum is a waste and therefore listed in the EWC as follows [51]:

06 00 00:	wastes from inorganic chemical processes
06 09 01:	phosphogypsum

In the OECD lists for the Control of Transfrontier Movements of Wastes, phosphogypsum is listed as follows [52]:

Amber List AB 140:	gypsum arising from chemical industry processes

This phosphogypsum is a moist, fine powder with a free water content of ca. 20 – 30 % and considerable amounts of impurities, the exact impurities and their amounts depending on the rock and the specific process. The radioactive substances present in small amounts in sedi-

mentary phosphate rock are partly transferred to the phosphogypsum as ^{226}Ra, leading to slightly increased radioactivity of such gypsums [14]. About 1.7 t of gypsum is obtained per tonne of raw phosphate, corresponding to 5 t of gypsum per tonne of phosphorus pentoxide produced.

Currently < 2 % of the phosphogypsum obtained is utilized. The principal difference in using phosphogypsum compared to FGD gypsum is the need to refine phosphogypsum by recovery operations (washing, flotation, and recrystallization) to remove harmful impurities and to modify its unsuitable particle shape. However, it is simply cheaper to use natural gypsum or FGD gypsum than to purify and process phosphogypsum. The quantity of phosphogypsum obtained annually is nearly equal to world demand for natural gypsum and anhydrite of ca. 100×10^6 t/a [7], [50].

3.4. Fluoroanhydrite

Fluoroanhydrite is a byproduct in the production of hydrofluoric acid from fluorspar and sulfuric acid, to the extent of $(2-3) \times 10^6$ t/a ($n \times M_r$):

$$\underset{(78)}{CaF_2} + \underset{(98)}{H_2SO_4} \longrightarrow \underset{(136)}{CaSO_4} + \underset{(40)}{2\,HF}$$

One tonne of fluorspar produces 1.75 t of anhydrite. Fluoroanhydrite is a product; it is used as a raw material for the gypsum industry only in Germany and some of its neighbors.

3.5. Other Byproduct Gypsums

Calcium sulfate dihydrate is formed in small quantities in the production or treatment of organic acids (e.g., tartaric acid, citric acid, and oxalic acid) or inorganic acids (e.g., boric acid). All these acids are produced by reaction of their calcium salt with sulfuric acid.

The amounts of industrial byproduct gypsums are expected to increase from treating sulfate-containing waste waters and neutralizing dilute sulfuric acids that are produced in the processing of minerals (e.g., zinc ore) and the manufacture of pigments (e.g., titanium dioxide). Especially titanogypsum is obtained in large quantities and can be used in the gypsum industry if it is produced in the titanium dioxide industry with specification and quality standards similar to FGD gypsum. Several 100 000 t are already used in the gypsum industries of Japan, the United States, and Europe [85].

Gypsum scrap from the ceramic industry and metal foundries has likewise not been used so far as a raw material for the gypsum industry, although efforts continue to be made in that direction [53]. The quantities are small and recovery operations are very costly because of the impurities originating from slip casting.

4. Production

Although a dozen industrial processes for the preparation of cementitious calcium sulfates are well known, only a few are used extensively in the industrial countries. The processes characterized by low capital and operating costs, simple design and operation, robust and long-lasting equipment, and uniform high-grade calcined gypsum product subsequently used to manufacture gypsum building components and gypsum plaster are discussed in the following chapter.

Since the two major raw materials – natural gypsum and FGD gypsum – are processed differently, they are discussed separately, as is the refining and processing of phosphogypsum.

4.1. Natural Gypsum to Calcined Products

Natural gypsum is mined by the open-pit method [54] and by underground mining [55]. In open-pit mining the gypsum is recovered by drilling and blasting, mostly on one or more levels. In deep mining, chamber blasting is used [56]. The blasted rock consists of large lumps, containing 0 – 3 % free moisture (quarry water). The amounts of explosives needed are about 250 g/t for open-cast mining and 400 g/t for underground mining.

The coarse rock is conveyed to the crushing plants, normally at the gypsum works, by means of trackless loading and haulage gear. Impact crushers, jaw crushers, and single-roll crushers with screen and oversize return are all suitable for coarse size reduction. Impact pulverizers or

roller mills are used for intermediate size reduction; and hammer mills, ball mills, or ring-roll mills are used for fine grinding. The degree of size reduction is determined by the calcining unit or the intended use of the gypsum:

rotary kiln	0 – 25 mm
kettle	0 – 2 mm
kettle with combined drying and grinding unit	0 – 0.2 mm
conveyor kiln	4 – 60 mm
gypsum and anhydrite for cement	5 – 50 mm

If transport by ship or truck to distant gypsum or cement plants is necessary, the material is reduced to required size for further treatment prior to transport.

Cleaning or beneficiating the blasted rock is no longer common practice, except for gypsum with much Glauber's salt, which is readily removed by leaching with water. This is still done in Britain [54].

Often the mined and crushed rock is homogenized before being calcined. This is done in homogenizing plants with capacities of about one-week's production.

The final stage in gypsum open-pit mining is the restoration of the quarry site and its reinclusion in the environment of the area.

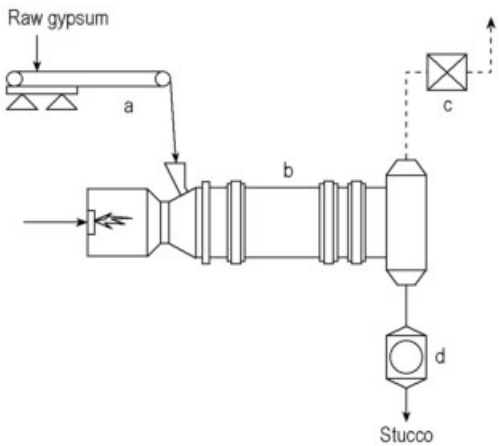

Figure 4. Production of stucco (plaster of Paris, β-hemihydrate plaster) by the rotary kiln process
a) Weigh-belt feeder; b) Rotary kiln with combustion chamber; c) Cyclone (filter); d) Pulverizer

β-Hemihydrate Plasters. Stucco and plaster of Paris are prepared by dry calcination between 120 and 180 °C, either in directly fired rotary kilns or else in indirectly heated kettles, to produce the β-hemihydrate. The kettles may be upright or horizontal.

The *rotary kiln* (capacities up to 600 t/d) is particularly suitable for calcining granular (0 – 25 mm) gypsum rock. This, without predrying, is fed continuously to the kiln in cocurrent flow with the hot gases with a weigh-belt feeder and cellular wheel sluice (Fig. 4). The hot gases are produced in a brick-lined combustion chamber ahead of the gypsum feed. The shell of the rotary kiln is not brick-lined but consists of thick steel plates with steel inserts to ensure an even distribution of gypsum particles across the furnace cross section.

High thermal efficiency is achieved by the direct transfer of heat from the hot gases to the gypsum. The residence time of the gypsum rock in the kiln is self-regulating. It increases with the size of the grains, but complete calcination can be achieved regardless of particle size, to produce β-hemihydrate of uniform quality and invariant properties after fine grinding and air classifying. Any gypsum not dehydrated to hemihydrate at start-up or shut-down is collected in a bin and added in small quantities to the calcined gypsum. Rotary kiln plants are fully automatic and usually can be controlled by one worker.

The externally heated *kettle*, of capacity up to 760 t/d, is the first of a series of similar calcining units heated indirectly. The gypsum is predried and ground to < 2 mm. In modern plants the fine grinding, which used to follow the calcining process, is combined with the predrying and carried out in drying – grinding units to produce particles < 0.2 mm (Raymond mill, Claudius Peters mill, attritor). The properties of the product hemihydrate are shown in Table 8. The phase composition of plaster of Paris is shown in Table 9.

In the 1960s [58], [59] the kettle process was improved and made continuous, especially in England (British Plaster Board Ltd) and North America. Predried and finely ground gypsum is fed continuously from above. As it dehydrates, the gypsum sinks to the bottom, the water vapor set free keeping the gypsum bed fluidized. The calcined gypsum settles at the bottom of the kettle, where it is continuously discharged through a pipe connected to the side of the kettle.

Figure 5 shows a flow diagram for such a plant. The thermal balance is not quite as good

Table 8. Examples of characteristic properties of calcined gypsum [57]

Mode of production	Type of plaster	g plaster per 100 g water	Water to plaster ratio	Setting time, min		Strength of hardened plaster,∗ N/mm^2			Density of hardened specimen, kg/m^3
				initial	final	flexural	compressive	hardness	
Rotary kiln	β-hemihydrate	137	0.73	13	28	4.8	11.2	19.1	1069
Conveyor kiln	multiphase plaster	167	0.60	6	35	5.1	15.5	25.9	1225
Kettle	β-hemihydrate	156	0.64	9	22	5.2	14.0	26.8	1133
Autoclave	α-hemihydrate	263	0.38	10	22	12.8	40.4	92.0	1602

∗40 × 40 × 160 mm specimen.

Table 9. Phase compositions, in %, of overburnt plaster,plaster of Paris, and multiphase plaster∗

	Plaster of Paris (rotary kiln)	Overburnt plaster (conveyor kiln)	Multiphase plaster, a mixture of plaster of Paris and overburnt plaster
Dihydrate, $CaSO_4 \cdot 2\,H_2O$	0	0	0
β-Hemihydrate, β-$CaSO_4 \cdot 1/2\,H_2O$	75	6	27
β-Anhydrite III, β-$CaSO_4$ III	20	19	15
Anhydrite II, $CaSO_4$ II∗∗	5	75	58

∗ Impurities not considered. ∗∗ Consisting of reaction stages AII-s (slowly soluble A), AII-u (insoluble A), and AII-E (*Estrichgips*).

as for directly fired rotary kilns, but the waste gases are less dust laden. Up to 760 t/d of hemihydrate plaster can be produced in this type of plant.

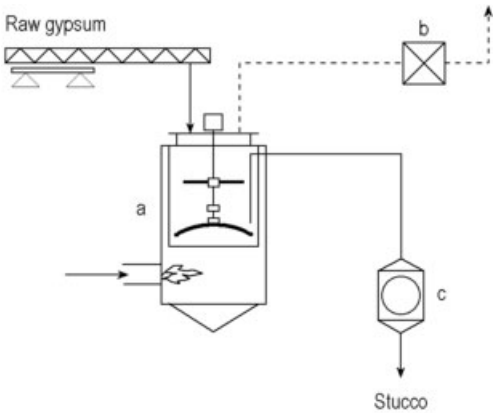

Figure 5. Production of stucco (plaster of Paris, β-hemihydrate plaster) by the continuous kettle process
a) Kettle with combustion chamber and ascending pipe; b) Cyclone (filter); c) Pulverizer

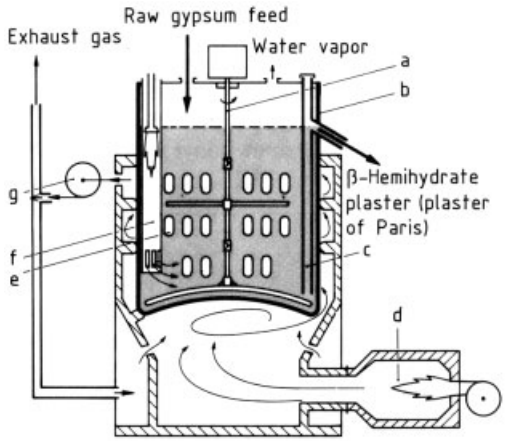

Figure 6. Kettle with submerged combustion burner for continuous production of β-hemihydrate (plaster of Paris)
a) Agitator; b) Outer casing of kettle; c) Ascending pipe for stucco plaster discharge; d) Combustion chamber; e) Cross pipes for hot gases; f) Submerged pipe with submerged combustion burner; g) Ventilating fan for exhaust and circulating gas

The *submerged combustion* method – more specifically indirect heat transmission of the externally heated boiler casing combined with direct heat transmission – has been introduced. The advantages are high thermal efficiency and

low energy consumption. Kettles heated only by submerged combustion are being tested at present. Existing kettles can also be fitted with submerged combustion burners, which reduces energy consumption and increases the flow rates. Figure 6 shows a kettle with such a burner that is designed for continuous operation [59].

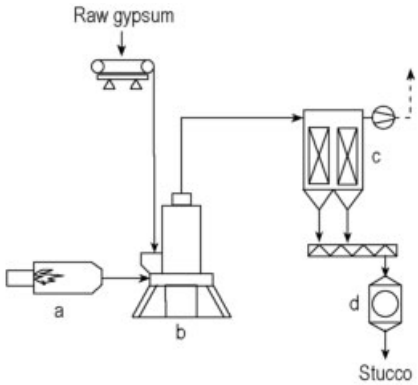

Figure 7. Production of stucco (plaster of Paris, β-hemihydrate plaster) by the grinding – calcining process
a) Combustion chamber; b) Grinding – calcining unit; c) Separating cyclone; d) Pulverizer

Batch horizontal kettles, with capacities of 5 – 10 t/h, were popular in France (Beau kettles).

Almost the total world production of β-hemihydrate is carried out in plants of the types just described. Today in North america indirectly heated kettles and directly heated rotary kilns are in use in a proportion of about 5 : 1, whereas earlier only kettles were used. In Britain the indirectly heated kettle is widely used; in France, the horizontal kettle; and in Germany, rotary kilns and kettles.

Most of the β-hemihydrate or stucco produced is used in gypsum building components; to a lesser extent β-hemihydrate is used in special building plaster.

Other calcining processes for the manufacture of β-hemihydrate have only occasionally been used industrially. In a grinding – calcining process [60], the raw gypsum is continuously crushed and calcined to β-hemihydrate in a single unit (Claudius Peters; Fig. 7).

A few plants produce β-hemihydrate and to some extent also multiphase plaster in a fluidized bed [61–63] The dried and finely ground gypsum rock is calcined in the stream of hot gases (Thyssen – Rheinstahl). Another development is the indirectly heated, continuously operated horizontal kettle. Externally it resembles a rotary kiln. In this unit the hot gases first pass through a central tube in cocurrent flow and

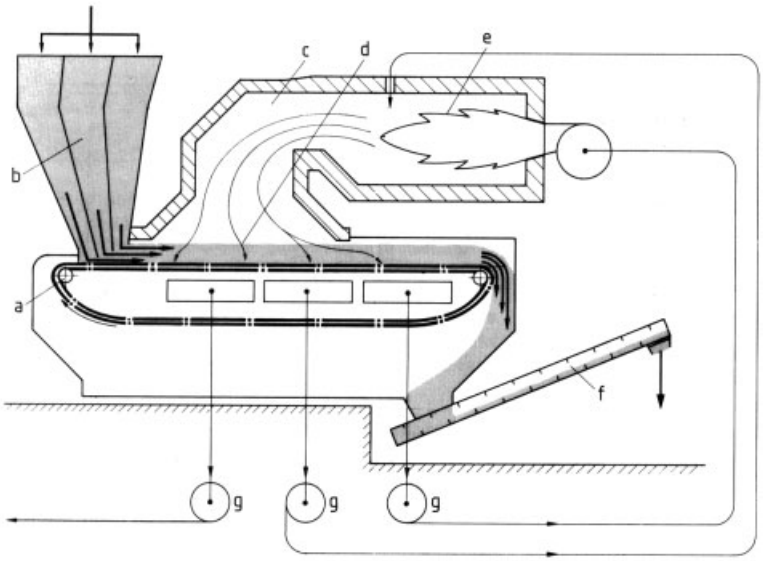

Figure 8. Conveyor kiln for the production of anhydrous plaster
a) Conveyor grate; b) Feed hoppers; c) Calcining hood; d) Layers of gypsum; e) Combustion chamber; f) Discharger; g) Circulating and cooling air fans for exhaust gas

then through further heating tubes in countercurrent flow. Similar is the continuously operated Holoflite calcining unit, using hot oil or superheated steam as the heat-exchange medium. In this case the gypsum is moved along on a screw conveyor that has hot oil passing through it [64].

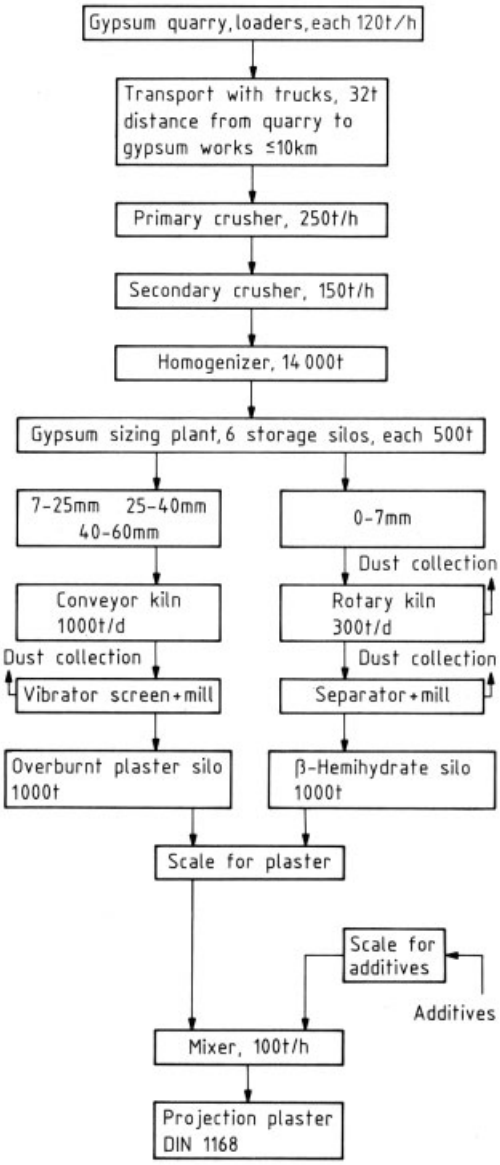

Figure 9. Flow diagram for production of projection plaster with a conveyor kiln

Calcined Anhydrous Plasters and Multiphase Plasters. Overburnt plasters are produced in dry calcining processes at temperatures between 300 and 900 °C. For the most part they are calcined anhydrite. There are three reaction stages:

1) AII-s, slowly soluble anhydrite, < 500 °C
2) AII-u, insoluble anhydrite, 500 – 700 °C
3) AII-E, *Estrichgips*, > 700 °C

The properties and final phase compositions of overburnt plaster are shown in Tables 8 and 9.

In overburnt plaster, and also in multiphase plaster, these three stages of reaction have to have a definite ratio, which is determined jointly by the raw material and the calcining process. In some of these processes [1] the gypsum is calcined in such a way that the anhydrite phase is obtained along with the hemihydrate. In other processes the overburnt plaster is produced separately and mixed with β-hemihydrate afterwards. These plasters are designated as multiphase plasters, *Putzgips* in Germany (see Tables 8 and 9).

A modern calcining unit for overburnt plaster and multiphase plaster is the *conveyor kiln* developed by Knauf [65] (Fig. 8). Today, capacities up to 1200 t/d are usual, making it the most efficient kiln used for gypsum. Before being fed to the conveyor kiln, the gypsum rock is crushed to size 4 – 60 mm and split into three or four sieve fractions (e.g., 7 – 25 mm, 25 – 40 mm, and 40 – 60 mm or 4 – 11 mm, 11 – 25 mm, 25 – 40 mm, and 40 – 60 mm). The fractions are piled on the continuous conveyor grate, the smallest on the bottom. The grate, which moves continuously at a speed of 20 – 35 m/h, passes through a calcining hood. There the hot gases are drawn through the gypsum bed by exhaust fans. The top layer can reach a temperature up to 700 °C; the bottom layer, up to 300 °C. The temperature of the heat-resistant plates of the conveyor grate does not exceed 270 °C. The gypsum is not mixed during calcination, and therefore little dust is formed and no dust collection system is needed. About half the hot gases are discharged into the atmosphere as waste gas, issuing from the stack at about 100 °C; the remainder (at ca. 270 °C) is recirculated to the combustion chamber together with air (ca. 230 °C) drawn through the calcined gypsum bed to cool it. The thermal efficiency of such a conveyor kiln is high, greater than 70 %. In Germany and elsewhere conveyor kilns with a total capacity of more than 2×10^6 t/a are currently in operation.

Figure 9 shows a flow diagram of a process producing construction plaster. In addition to a conveyor kiln it also uses a *rotary kiln* in which raw gypsum of small particle size (0 – 4 mm or 0 – 7 mm) is calcined to β-hemihydrate and subsequently mixed with anhydrous plaster that has been ground to the required fineness. In this way, multiphase plasters are produced. They differ from each other in their setting properties and also from β-hemihydrate plaster in that they are coarser (see Table 9 for composition, Table 8 for properties, and Figure 10 for representation as three-component diagram) [1].

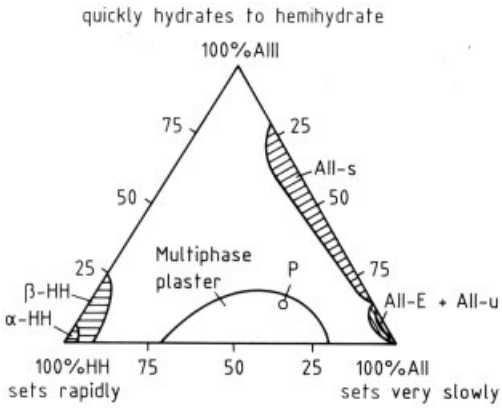

Figure 10. Three-phase composition diagram of calcined gypsums [1]
Point P indicates the composition of a multiphase plaster consisting of 27 % β-hemihydrate, 15 % anhydrite III, and 58 % anhydrite II, the composition given in Table 9 for multiphase plaster

Multiphase plasters are used to finish interior walls and ceilings. Since 1965 machine-applied plaster has been produced in large quantities (see Section 5.3) by incorporating chemical additives to one such multiphase plaster. Even pure construction plasters (Saarland construction plaster, southern German *Doppelbrandgips*), which have been used in Germany for years, but which recently have declined in importance, can be prepared in these types of plants. This is also true for the construction plasters used in Latin countries, North Africa, and the Near East [66].

Other calcining plants producing anhydrous plasters operate on the principle of a *countercurrent rotary kiln* (Vernon). For example, in France an overburnt plaster, *surcuit*, is produced and mixed with β-hemihydrate to produce *Paris construction plaster* [67]. In such a countercurrent rotary kiln, the gypsum enters at the end opposite to the combustion chamber and travels into hot gases. The dehydrating gypsum reaches its maximum temperature, (ca. 500 °C), just before exiting the kiln. These kilns, with capacities of 15 – 30 t/h, are sometimes linked together and used to produce anhydrous plaster and β-hemihydrate simultaneously [68]. This improves the efficiency and offers the opportunity of cooling the anhydrous plaster obtained at very high temperatures advantageously. The problem of cooling has also been solved by making use of the planetary cooling system of the cement industry.

α-Hemihydrate. α-Hemihydrate can be produced by wet calcining processes, either under elevated pressure in autoclaves or at atmospheric pressure in acids or aqueous salt solutions between ca. 80 and 150 °C. However, only the autoclave processes have so far achieved industrial importance for the small amounts produced.

α-Hemihydrate from natural gypsum is nearly always batch-processed. For instance, gypsum rock (particle size 150 – 300 mm, >95 % $CaSO_4 \cdot 2\,H_2O$, rocklike) is put in wire baskets, and either stacked in upright autoclaves or wheeled into horizontal autoclaves with capacities of 0.5 – 10 m^3. The autoclaves are heated directly or indirectly with steam at 130 – 135 °C. The heating is carried out so that after about four hours a pressure of 4 – 5 bar has built up. The autoclave is then emptied; the α-hemihydrate formed is immediately transferred in the baskets to a chamber to be dried at 105 °C under atmospheric pressure. It is subsequently ground to the desired particle size and size distribution. Variation in pressure and temperature during dehydration and drying can be used to affect the properties of the products (see Table 8).

α-Hemihydrate is often mixed with β-hemihydrate. For that reason processes have been developed that produce a mixture of α- and β-hemihydrates in a single operation [69]. In one such process the α-hemihydrate is dried in a rotary kiln, which at the same time serves to both calcine the β-hemihydrate and blend the two types of material.

Process Control. Dry gypsum calcining processes are usually equipped with dry *dust collection* systems. The amount of dust formed

depends on the type of calcining unit. For the most part electrostatic filters are used for dust collection, but recently mechanical filters have also been used. Sieves and grinders, as well as conveying equipment, for calcined gypsum are fitted with flat screen-type filters for internal dust collection. The dust collected is added to the calcined gypsum.

The water liberated in the process is discharged as vapor through outlet stacks into the atmosphere. There are no wastes or byproducts. No environmental problems are encountered, provided the processing of the gypsum plaster proceeds normally.

The *energy consumption* of a gypsum plant is the sum of the fuel used in the calcination of the gypsum and the electric power needed to operate the machinery. Table 10 shows the energy consumption of various types of processes based on experience. Mainly oil and gas have been used as the fuel. However, coal is being used increasingly, chiefly in grate firing, pulverized coal firing, and fluidized firing systems [70]. To improve the energy consumption of these units, the design is optimized and the waste heat is utilized. At the same time large modern gypsum plants are increasingly using automatic process control.

4.2. FGD Gypsum to Calcined Products

FGD gypsum is used directly as a major raw material by the gypsum and the cement industries.

4.2.1. β-Hemihydrate and Multiphase Plasters from FGD Gypsum

FGD gypsum is a finely divided powder with less than 10 % surface moisture. It therefore differs considerably in its state of aggregation from natural gypsum and requires treatment not needed for natural gypsum.

To produce β-hemihydrate for the manufacture of gypsum plasterboards and other gypsum building components, FGD gypsum must be dried before calcination. This can be carried out in cocurrent drying units in which the hot gases come into direct contact with the moist gypsum (e.g., flash dryers), or in tube dryers heated indirectly with steam. After drying, the gypsum is calcined to β-hemihydrate in kettels [71], [72].

FGD gypsum used for the production of multiphase plaster and gypsum building plaster must be treated further. Its particle structure, which ranges from cubic (bulk density 1200 kg/m^3) to lath or rod shaped (bulk density 500 kg/m^3), and its narrow particle size distribution (40 – 60 µm) are unsuitable for the production of multiphase plaster [73]. It is this particle structure that is the cause of its thixotropy if it is used as multiphase plaster. As plaster it also lacks smoothness, and its volume yield is too variable.

Agglomeration processes have been developed to modify the particle structure of FGD gypsum and convert it into a lump product. Of all the agglomeration processes – briquetting with a compacting press, granulating with extrusion presses, pelletizing – *briquetting* has proved the most effective and is the most widely accepted. In this process the flue-gas gypsum is compacted in the dry state without addition of bonding agents or additives to produce briquettes 2 cm thick and 6 cm long, comparable to natural gypsum in their mechanical properties [74].

Their point strength exceeds 500 N, and their apparent density is ca. 2.15 g/cm^3. These briquettes are exceptionally abrasion resistant. They can be stored in the open and are unaffected even by frost or rain. In this form all types of FGD gypsum have a bulk density of ca. 1.1 t/m^3 and can be used alone or mixed with natural gypsum.

Figure 11 shows a flow diagram for a plant processing flue-gas gypsum by drying and briquetting. Several of these plants operate in Germany.

The FGD gypsum briquettes can be processed to multiphase gypsum plaster in the calcining units used for natural gypsum.

The extra energy input per tonne of dry $CaSO_4 \cdot 2\,H_2O$ consists of 550 000 kJ of thermal energy and 12 kWh of electrical energy for drying and 10 kWh of electrical energy for agglomeration into briquettes.

Another method for modifying the unsuitable particle structure of calcined FGD gypsum is fine grinding, which also gives a product suitable for production of multiphase plaster and gypsum building plaster [75].

The value of FGD gypsum can be established from a cost comparison between processing FGD gypsum and natural gypsum. A second

Table 10. Examples of characteristic energy data in manufacturing calcined gypsum

Parameter	Rotary kiln (β-hemihydrate)	Kettle (β-hemihydrate)	Conveyor kiln (*Putzgips*)	Autoclave (α-hemihydrate)
Calcined gypsum capacity, t/d	600	760	1200	150
Theoretical energy requirement, kJ/t	598 600	584 000	770 400	560 000
Practical energy requirement, kJ/t	892 000	800 000	1 100 000	1 590 000
Thermal efficiency, %	67	73	70	35
Moisture content of gypsum rock, %	1.5	1.5	1.6	0.0
Crude gypsum purity, %	90.0	90.0	94.0	95.0
Water of crystallization, raw gypsum, %	18.8	18.8	19.7	19.9
Water of crystallization, calcined gypsum, %	3.0	5.5	1.0	5.8

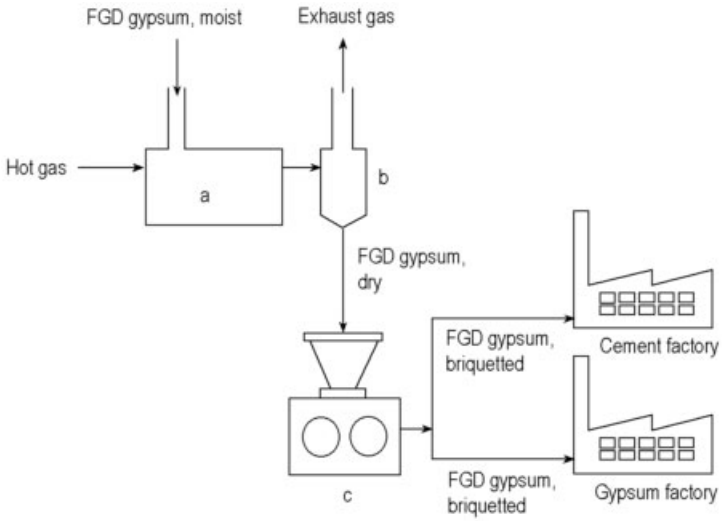

Figure 11. Flow sheet of a FGD gypsum drying and briquetting plant
a) Dryer; b) Cyclone; c) Press

major factor is the transport of the FGD gypsum from power station to gypsum plant. In comparison with the transport of natural gypsum from the quarry to the gypsum plant, the transport of FGD gypsum is associated with additional costs, e.g., for intermediate storage due to seasonal differences in production of the power station and consumption of the gypsum plant.

For a gypsum plant the long-term security of supply of FGD gypsum is vital. In times of FGD gypsum shortages it must therefore be possible to fall back on natural gypsum. The surplus quantities of FGD gypsum that will be available in the future in parts of Europe should be placed in intermediate storage locally in artificial deposits as a future raw material reserve. An environmental compatibility study is at present being prepared. With such a gypsum reserve it would be possible to fulfill the requirement of the gypsum industry for an assured long-term supply of raw material over the life-time of the power station [47], [76].

4.2.2. α-Hemihydrate Plaster from FGD Gypsum

FGD gypsum is an ideal raw material for manufacturing α-hemihydrate plaster by the continuous autoclave process (Nitto Gypsum – Knauf). The gypsum is used in its original moist, finely divided state without predrying. In the continuous autoclave process, shown schematically in Figure 12, FGD gypsum having a free water content of 10 % is slurried with water (one part gypsum, two parts water) and pumped continuously into the autoclave where it is dehydrated to α-hemihydrate under controlled conditions (135 °C, 2 h). Additives in the suspen-

sion change the crystal habit of the α-plaster and yield a product of defined, consistent properties. The α-hemihydrate produced is withdrawn as an aqueous suspension and dewatered in a vacuum filter. The product, which has a free water content of ca. 10 – 15 % is immediately dried at ca. 150 °C in an indirectly heated dryer and ground. This dry α-hemihydrate can then be used for all types of α-hemihydrate plasters and products.

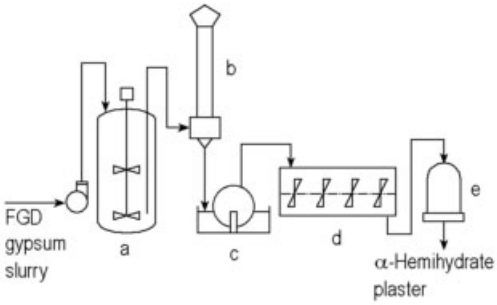

Figure 12. Production of α-hemihydrate from FGD gypsum by the continuous autoclave process (Nitto Gypsum – Knauf)
a) Autoclave; b) Expander; c) Vacuum filter; d) Dryer; e) Pulverizer

Today the continuous autoclave process is also used industrially for the processing of finely divided natural gypsum and synthetic gypsum.

Another α-hemihydrate plaster process has recently been developed especially for utilizing lignite FGD gypsum (ProMineral – SICOWA). The moist, finely divided FGD gypsum is mixed with auxiliaries, pressed into stable stackable blocks and converted into α-hemihydrate plaster in large hot pressure autoclaves. The resulting blocks of α-gypsum are dried and ground. Important applications for this type of α-gypsum are the manufacture of self-levelling floor plaster, floor slabs, and mining mortar. Other continuous autoclave processes have ceased operations within a few years (Giulini), never got beyond the pilot stage (ICI), or remained dormant in patents. Processes in which α-hemihydrate is produced without pressure in sulfuric acid, nitric acid, or salt solutions are found in the literature, but they never came into operation. A patent describes a method for producing α-hemihydrate directly from the calcium sulfite of a scrubbing liquor [77]. All efforts to produce an α-hemihydrate plaster similar in its properties to β-hemihydrate plaster have so far been unsuccessful.

4.3. Phosphogypsum to Calcined Products

Moist, finely divided phosphogypsum is a waste from the production of wet phosphoric acid [13]. It is available as filter cake with a free water content of 20 – 30 wt %. The content of impurities is high and the impurities must be removed or rendered harmless by recovery operations such as washing, flotation, and recrystallization before phosphogypsum can be used by the gypsum industry. All phosphogypsums contain inorganic impurities such as phosphates, silicofluorides, and sodium salts. They affect the properties of the gypsum. Organic impurities and small amounts of radiation (^{226}Ra) are present in sedimentary phosphate rocks (Morocco), part of which appears in the gypsum. Magmatic phosphate rocks (Kola) contain no organic impurities and no radioactive matter. The unsuitable crystal shape of all phosphogypsums must be modified by fine grinding or by recrystallization.

The one- and two-stage processes that have been developed for the manufacture of wet phosphoric acid produce different qualities of phosphogypsum. In the *one-stage* processes, phosphogypsum is obtained as dihydrate (dihydrate process, e.g., that of Prayon, altogether representing 84 % of all phosphogypsum produced) or hemihydrate (hemihydrate process, e.g., that of VEBA [78] or Fisons [79], but representing less than 1 %).

In the *two-stage* phosphoric acid processes, either gypsum or hemihydrate is produced in the first stage. In the second stage it is converted to another state of hydration before being removed by filtration. These processes include the hemihydrate – dihydrate process (e.g., Nissan [80]; 15 % of all phosphogypsum produced) and the dihydrate – hemihydrate process (e.g., Central Prayon [81], less than 1 %). Generally, gypsum obtained from the two-stage processes is of better quality in regard to inorganic impurities such as phosphate, fluorine, and sodium.

The gypsum industry is faced with the need for recovery operations if it attempts to use phosphogypsum [82], [83]. Removal of organic impurities, which discolor the gypsum, and of water-soluble inorganic contaminants, which cause efflorescence, involves first remashing the phosphogypsum with water, then sub-

jecting it to flotation, classification (hydrocyclone), thorough washing, and filtering. The radioactivity must be decreased to < 370 bq/kg ($< 10 \times 10^{-9}$ Ci/kg). Up to 5 t of water per tonne of phosphogypsum is consumed. Gypsum purified in this way is obtained as filter cake with a free water content of 20 – 30 %.

For production of β-hemihydrate intended for gypsum building components the filter cake is dried (e.g., rapid dryer from Hazemag, or the contact dryer from Serapic) and then calcined to β-hemihydrate in the same way as finely ground natural gypsum. Today this method is still the principal one used by the gypsum industry in Japan, where this method was developed in 1940 by Yoshino.

For use in gypsum building plaster and multiphase plaster, phosphogypsum is made unsuitable by its particle shape, fineness, and isomorphous acid phosphate impurities. Its particle shape and fineness seriously impair the workability of the multiphase plaster, i.e., the plaster is thixotropic. The cocrystallized acid phosphates cause calcined products to develop *lime sensitivity*, which interferes with setting and development of the strength [84].

In order to overcome these deficiencies the phosphogypsum is dried, calcined, and after addition of aqueous calcium hydroxide suspension simultaneously agglomerated and recrystallized in a pelletizer. Such alkaline-recrystallized pellets are used as starting material for gypsum building plaster and multiphase plaster.

In one process *phosphohemihydrate gypsum* is the starting material. If the process is carried out with adequate care, the phosphohemihydrate is sufficiently pure that the gypsum recovery can be dispensed with. The fine phosphohemihydrate, having a residual water content of ca. 20 – 25 %, is mixed immediately with a calcium hydroxide suspension or calcium hydroxide powder so that the calcium sulfate crystallizes into coarse-grained dihydrate. Part of the moisture and all of the acid phosphate are bound chemically, and the particle size and structure are satisfactory.

The phosphogypsum pellets or lumps thus produced can then be calcined and further processed in rotary kilns or, after grinding, in kettles to produce a β-hemihydrate plaster similar in composition to plaster of Paris. They can be converted on a conveyor grate into overburnt plaster, which consists of coarse particles and is comparable to multiphase plaster. These plasters can be processed into all kinds of machine-applied and premixed plasters.

Such processing methods were developed and put into operation by Knauf in 1962 and 1970. Most other processes designed to use phosphogypsum (e.g., those of Rhône-Progil, Charbonnages de France, Chimie-Air Industrie, Imperial Chemical Industry, Buell) have not resulted in viable commercial operations.

The recovery operations required by phosphogypsum involve additional capital expenditure and operating costs, jeopardizing its competitiveness with natural gypsum and generally rendering phosphogypsum uneconomical for commercial use. The repercussions felt after the first energy crisis, in 1973, because the extra treatment is energy intensive, were an additional set-back.

Unlike flue gas gypsum producers, phosphoric acid producers do not pay for the extra treatment required by the gypsum manufacturer. In summary, the conclusion must be drawn that the chance of phosphogypsum ever being utilized on a large scale worldwide is extremely remote, even though disposal of phosphogypsum is becoming more of a problem [50].

Only Japan has so far managed to continuously use phosphogypsum, an accomplishment favored by its total lack of natural gypsum resources [85]. South Korea also uses phosphogypsum as a source of gypsum.

4.4. Anhydrite Plaster

Anhydrite plaster is produced by grinding anhydrite rock in tube mills or impact pulverizers to a particle size below 0.2 mm. Activators to promote setting are added together with the gauging water. However, the very fine grinding is expensive.

The activators are mixtures of alkali-metal or heavy-metal salts and calcium hydroxide, up to ca. 2 wt % of the anhydrite. Acid activators, e.g., potassium hydrogensulfate or iron(II) sulfate, can also be used [86].

Fluoroanhydrite, a dry fine powder, is neutralized with calcium hydroxide and ground very finely for use as an anhydrite plaster. Sulfates, e.g., potassium sulfate and zinc sulfate, and cal-

cium hydroxide or Portland cement are activators [87], which are usually added and mixed with the anhydrite powder in the factory.

Natural anhydrite and fluoroanhydrite differ from each other in crystal structures. Fluoroanhydrite consists of very small primary crystals that have been agglomerated to secondary particles with a high specific surface area and high reactivity, whereas natural anhydrite consists of large primary particles, which are rendered reactive by fine grinding.

A recently devoloped process for the manufacture of thermal anhydrite from FGD gypsum is the suspension gas calcining process (POLCAL process). The plant consists of one or more preheating cyclones, a reactor with integrated separating cyclone, and a cooler. The conversion temperature is $> 700\,^{\circ}C$, and the residence time of the small gypsum particles in the reactor is a few seconds. Thermal anhydrite from FGD gypsum is used as a binder component in self-levelling floor plasters [47].

5. Use and Properties of Gypsum Plasters and Products and Anhydrite Plasters

5.1. Hydration, Setting, Hardening

Calcium sulfate hemihydrate, anhydrite III, and anhydrite II undergo hydration under ambient conditions, converting into calcium sulfate dihydrate. If hydration is carried out with just enough water to produce a homogeneous, fluid, stable, nonsedimenting slurry, then this mixture sets and hardens because the calcium sulfate dihydrate forms needles that intergrow and interlock.

Much research has gone into the mechanism of hydration. Around 1900, Le Chatelier [88], [89] established the theory of crystallization, which gained universal acceptance [90], [91]. According to this theory the calcium sulfate hemihydrate in water first forms a saturated solution, about 8 g/L at 20 °C. However, this solution is actually supersaturated, because at 20 °C calcium sulfate dihydrate has a solubility of only 2 g/L, and $CaSO_4 \cdot 2\,H_2O$ precipitates.

Cavazzi [93], and later Baykoff [94], put forward a colloid theory, which states that the hydration proceeds via a colloidal intermediate involving formation of a type of gel or an adsorption complex between the calcium sulfate hemihydrate and water. However, this gel has not yet been demonstrated experimentally. The Le Chatelier theory of crystallization has been supplemented by the detection of topochemical reactions taking place during hydration [95]. The two mechanisms are not mutually exclusive, because transformation into dihydrate in both cases takes place via the solution phase. The topochemical hydration is an internal hydration of the hemihydrate particles taking place over very short distances.

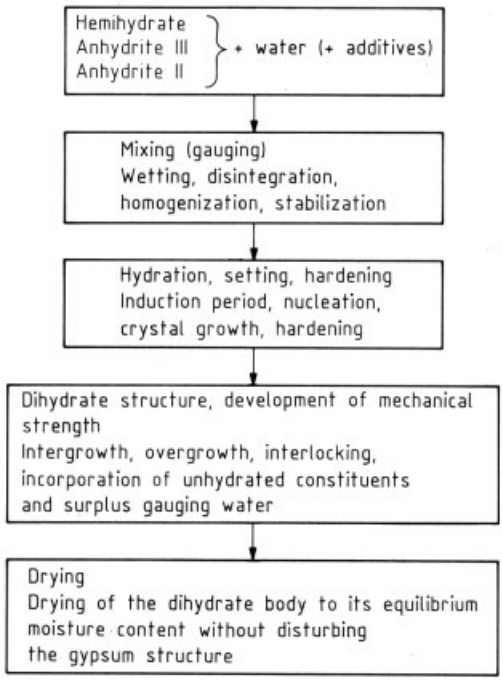

Figure 13. Stages of hydration, setting, and hardening of calcium sulfates

Formation of dihydrate crystals conforms to the laws of nuclei formation and crystal growth (Fig. 13). Mixing and wetting of the hemihydrate powder, which causes disintegration of the hemihydrate particles, is followed by a short induction period, after which nuclei begin to form from the supersaturated solution. The accumulation of very small dihydrate crystals with much excess water has been described by Krönert et al. as clustering [96]. Subsequently, after this nucleation, crystal growth begins, which at least initially is accompanied by continuous recrys-

tallization [92]. The rate of nuclei formation is proportional to the relative supersaturation (von Weimarn's theory), and the rate of crystal growth is proportional to the absolute supersaturation (Nernst – Noyes equation). Hemihydrate is converted directly into dihydrate; there are no intermediate stages. Anhydrite III is converted via the hemihydrate, and anhydrite II is converted directly into dihydrate, without anhydrite III or hemihydrate intermediates. If the proportion of water is correct for setting and hardening, the slurry hardens by forming a dihydrate structure, a final crystallization, which according to LUDWIG et al. [97] consists of intergrown, overgrown, and interlocking dihydrate crystals and inclusions of unhydrated components. Excess water can be removed by drying.

The rate of hydration of β-hemihydrate as shown in Figure 14 is indicated by combined water content, intensity of the X-ray diffraction, and rise in temperature due to the heat of hydration [96].

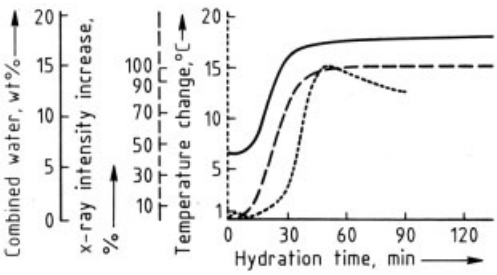

Figure 14. Rate of hydration of β-hemihydrate as indicated by combined water content (—-), intensity of X-ray diffraction (– – –), and temperature increase from heat of hydration (. . .) [98]

There are many ways in which these processes of hydration, setting, and hardening can be applied in practice. Parameters for characterizing these processes are the water-to-plaster ratio on mixing, the consistency of the mixture, the initial and final set, the rate of strength development, and the strength and density of the final dry gypsum product.

The method of manufacture of the plaster influences the gypsum technology to a very large extent. For instance, β-hemihydrate from a rotary kiln requires more water to produce a fluid slurry of uniform consistency than does plaster from a kettle. The latter, in turn, requires more water than multiphase plaster, which, in turn, requires more than autoclave plaster. This water-to-plaster ratio (water capacity of the gypsum plaster), an inverse of the quantity of gypsum plaster in grams per 100 g of water, is related to the strength and density of the set and hardened gypsum product.

α-Plasters, which are workable with little water, can be simply turned into gypsum products of high strength and high density. However, these products are avoided in the building industry on account of their brittleness.

β-Plaster and multiphase plaster require more water than α-plaster to obtain a fluid consistency. They produce products of average strength, higher elasticity, and lower densities, and are used universally in the construction industry.

The particle shape, particle-size distribution, and specific surface area also determine the quantity of water required to give a specific consistency. Generally, very fine plaster requires more water than coarse-grained plaster. Rod-shaped particles also increase the amount of water needed. β-Hemihydrate particles may disintegrate on first contact with water, breaking up into a multitude of very fine loose particles, thus changing their particle size distribution [99]. The particle size distribution of multiphase plasters determines their workability to a great extent.

When stored, calcined gypsums are subject to changes in their properties, called *aging.* This aging is caused to some extent by the uptake of water vapor from the air. The degree of aging affects the water needed for given consistency; more water is needed for fresh calcined gypsum than for aged. If a considerable amount of water vapor is adsorbed, dihydrate nuclei may form, accelerating the hydration process. However, the reactions taking place on aging have not yet been entirely elucidated [98], [100]. Natural aging of calcined gypsum produces gradual changes in the properties of the plaster over a period of months. To avoid this, methods have been developed to bring about aging artificially, so that the plaster undergoes no significant changes during storage. One process is called aridization: calcium chloride or similar salts are added in quantities up to about 0.2 wt % to the raw gypsum before calcination [101]. Aging of calcined gypsum is also achieved by injecting small quantities of water containing a wetting agent to prevent the formation of a dihydrate [102], [103].

The mixing and gauging of calcined gypsum with water to form a slurry of specific consistency can be affected by various wetting agents. Most of these, called *plasticizers* or water-reducing agents, lower the water demand. They include alkylarylsulfonates, lignosulfonates, or melamine resins [104]. It is also possible to increase the water requirement by adding flocculating agents, e.g., polyethylene oxide [105]. Chemicals that thicken, e.g., cellulose and starch ethers, can be added to stabilize the water – plaster slurry or prevent sedimentation and segregation; however, these have little effect on water demand [106].

Setting and hardening can be accelerated or retarded by numerous additives [107], [108]. Many inorganic acids and their salts are useful as *accelerators*, especially sulfuric acid and its salts. Calcium sulfate dihydrate is regarded as a special additive in this respect. Finely divided, it acts as a strong accelerator and therefore must be completely removed when raw gypsum is calcined. The accelerating effect of these substances is due to an increase in solubility and the rate of dissolution of the calcined gypsum and to an increase in the rate of nuclei formation.

Retarders are usually organic acids [109] and their salts and organic colloids that are the decomposition and hydrolysis product of biopolymers such as proteins as well as salts of phosphoric acid or boric acid. The mechanism of retardation varies: high molecular mass colloids prolong the induction period because they are nuclei poisons. Other retarders slow down the rate of dissolution of the hemihydrate or the growth of the dihydrate crystals. The hydration of anhydrite II usually does not have to be retarded since it is slow enough and almost always requires acceleration.

In every case, temperature affects the rate of hydration of plaster, the rate increasing up to ca. 30 °C, after which it decreases [110].

The *strength* of set dried gypsum is directly proportional to its density, and therefore depends only on its porosity or, less directly, on the water-to-plaster ratio and the size and structure of the pores [111], [112]. The strength is affected by moisture or additives without a change in density. The strength of gypsum products with a moisture content exceeding 5 % is only about one half that of air-dried gypsum products. When a gypsum product dries, the strength begins to increase below 5 % moisture content, becomes evident around 1 % moisture content, and reaches its final value in the region of its equilibrium moisture content [11].

Persistently moist conditions reduce strength, because crystalline and textural changes, especially recrystallization, take place as a result of the solubility of gypsum in water [113]. The deformation or creep of moist gypsum products under mechanical stress is likewise the result of structural change. Additives induce a change in structure by changing the crystal habit of the dihydrate so that without a change in density strength is changed even in the dry state.

An extreme example of reducing strength is the effect of citric acid, commonly used as a retarder. Used sparingly, less than 0.1 %, it does have a retarding effect and lessens the strength only slightly. More, say, above 0.2 %, changes the crystal habit of the dihydrate to such an extent that hardening of the gypsum is no longer possible because the crystals no longer interlock and intergrow [114].

Murat [115] has studied the morphology of natural and synthetic calcium sulfate dihydrates with up-to-date techniques and the effects of additives upon the crystal habit of the dihydrate.

5.2. Prefabricated Gypsum Building Components

Prefabricated gypsum building components are manufactured in large quantities: plasterboards, partition panels, ceiling tiles, and fiber-reinforced boards. These are light, porous, dry, and nonbrittle products possessing excellent workability. β-Hemihydrate is the starting material for all these because it sets quickly and easily meets the building industry's demand for certain properties for the finished product.

The plants for the manufacture of gypsum building components are usually built adjacent to gypsum works.

Gypsum *plasterboards* are large thin gypsum panels covered with cardboard; they have a density of 750 – 950 kg/m^3. They are manufactured by feeding β-hemihydrate plaster into a continuous mixer from controlled feeding devices, mixing it continuously with water and additives, e.g., adhesives, to form a homogeneous and rapidly setting slurry. This slurry is spread on a con-

tinuous sheet of cardboard about 0.5 mm thick (200 – 300 g/m^2). The slurry is then covered with a second sheet of cardboard and passed over a molding platform to be shaped into a completely encased strip, 1.20 – 1.25 m wide and 9.25, 12.5, 15.0, 18.0, or 25.0 mm thick. This strip of gypsum plasterboard, initially soft, hardens within minutes and is cut into separate panels. These panels, approximately one-third of whose weight is free water, are dried immediately in a continuous tunnel dryer heated indirectly with steam or directly with gas or oil. The finished gypsum plasterboard, consisting of a gypsum core tightly encased and bonded to cardboard, is considered to be a laminated building material.

Many different types of gypsum plasterboard are manufactured, depending on their intended use. Distinctive features are size, edge configuration, weight, water resistance, structural behavior, and strength. Gypsum plasterboards with specific fire-resistant properties incorporate fiberglass. The noncombustible gypsum lightweight board makes use of a laminate of glass fiber mat with a lined-on glass silk scrim to replace the cardboard [116]. Modern plants for the manufacture of plasterboards have a capacity of about (20 – 40) $\times$ 10^6 m^2/a and an annual consumption of 150 000 – 300 000 t of hemihydrate plaster.

Plasterboard is used in interior finishing, e.g., for wall and ceiling paneling. It is screwed to wooden or metal frames or pasted on masonry or concrete with special building plaster (adhesive plaster, the German *Ansetzgips*). The joints are covered and finished paperless or with special paper and joint filler to form a smooth surface. Plasterboard can also be used for the construction of dismountable partitions and lightweight dividing walls having various characteristics (weight, sound proofing, fire resistance) in concrete or steel-framed buildings as well as in prefabricated dwellings. Plasterboard is factory processed into special-sized panels, coffers, or lightweight laminated panels with intermediate layers of polystyrene or polyurethane (insulating board). Multilayered plasterboards are used for dry floorings and for the construction of lift shafts. Factory-made tiles with decorative surface finishes such as plastic sheet, special coats of paint, or aluminum foil to prevent water vapor transmission are available.

Gypsum *partition panels* consist of set gypsum plaster. To produce them β-hemihydrate plaster is mixed with water (water-to-plaster ratio 0.9 – 1.0) and the slurry, which sets quickly, is poured into molds. After 5 – 8 min the panels are removed from the molds and dried [117].

The standard size of gypsum partition panels is 500$\times$666 mm, with a thickness of 60, 70, 80, or 100 mm and a density of 700 – 900 kg/m^3. They are used in interiors as lightweight dividing walls, the tongue and groove joints being bonded with joint plaster. The partition walls can be single or multilayered. Characteristic features are low weight, average to good sound insulation, and excellent fire resistance.

Gypsum *ceiling tiles* are produced by mixing β-hemihydrate, water, and small amounts of glass fiber and pouring the slurry into rubber molds. The molds allow for individual designs.

Ceiling tiles are used as decorative tiles, ventilation tiles, heating tiles, or sound-proofing tiles with mineral wool bonded on the back. They are screwed to a base frame and fitted into the framework. They are normally 625 $\times$ 625 mm in size and weigh between 10 and 20 kg/m^2.

Gypsum *fiber board* and fiber-reinforced gypsum elements are another group of gypsum building components manufactured from hemihydrate plaster and paper, glass, or other fibers [118–122]. Glass fiber can be incorporated as mat or web [123]. These components contain up to 15 wt % fiber, evenly distributed in the plaster. They have an apparent density between 800 and 1200 kg/m^3.

Factories for the manufacture of gypsum partition panels and ceiling tiles have capacities up to 10^6 m^2/a, consuming about 50 000 t/a of hemihydrate plaster, and those for gypsum fiberboard up to 10 $\times$ 10^6 m^2/a, consuming the same amount of hemihydrate. The most widely used gypsum fiberboard is Fermacell, which is produced in central Europe in quantities of 20 $\times$ 10^6 m^2/a.

5.3. Gypsum Plaster

Calcined gypsum is used for plastering. Multiphase plaster (see Table 9) is most suitable because its phase composition results in quick initial setting and gradual final setting; smoothness, plasticity, and high coverage; single-coat appli-

cation; rapid drying of the finished plasterwork; and suitable density and high strength.

β-Hemihydrate plaster without additives is not suitable for plastering because the initial setting occurs too late and the final setting too early. Also its particle size distribution is not suitable for plaster. However, additives permit β-hemihydrate to be made into certain types of plastering material, known as retarded hemihydrate gypsum plaster and premixed lightweight gypsum plaster. Because of inadequate setting properties, lack of smoothness, and insufficient coverage, α-plasters are not used in construction [11], [124].

Over the past twenty years considerable progress has been made in Germany towards development of a factory-made machine-applied gypsum plaster. Ordinary multiphase plaster is used as the starting material. Mixed with additives and aggregates it is processed into a material that can be used for continuous machine application in single coats. The rate of setting and the hardening process have been adjusted to suit the longer working cycle for material applied over a large surface area. Too rapid water uptake is controlled by adding water-retaining hydrophilic colloids such as methyl cellulose. In this way the soft plaster slurry remains plastic until the work has been completed, resulting in better bonding to the base. The uniformity of the properties and the homogeneity of these plasters are particularly good when they are manufactured in large plants.

Such plants are equipped with high-capacity mixing and handling devices (see Fig. 9) and can produce up to 400 000 t/a. At the construction site the dry plaster is fed continuously into the plastering machine where it is mixed with water at a metered rate to form a slurry applied in a single coat to the base by means of a screw pump and hose. A smooth and even surface is obtained with an electrically powered felting tool (e.g., Power Float, Gipsoglätt).

Plastering machines operated inside buildings are supplied with plaster directly from outdoor bins or containers by pneumatic conveyor pumps through flexible hoses.

Premixed gypsum plasters (bonding plaster, lightweight plaster), factory-processed from β-hemihydrate plaster (or multiphase plaster) containing additives and aggregates, e.g., expanded perlite or vermiculite, are worked batchwise on site. The plasters are mixed with water and applied in one coat on all types of concrete and masonry. Premixed plasters already contain the necessary additives to ensure good workability. Only water need be added by the plasterer. The application properties of these plasters are listed in Table 11.

In some countries, especially in Great Britain, the two- or three-coat method of plastering is still employed. The undercoat is browning plaster, a factory-processed hemihydrate plaster containing a retarder, which is mixed with sand or expanded perlite either in the factory or on the building site. It takes several hours to set, i.e., far longer than premixed plaster. The next day a smooth finishing coat of plaster of Paris and hydrated white lime is applied.

In France, Spain, some countries in North Africa and the Near East, and to some extent in southwest Germany, especially in the Saarland, pure multiphase plaster containing no additives (Paris construction plaster, southern German *Doppelbrandgips*, Saarland multiphase plaster) is used for single-coat and multicoat plasterwork. In countries with a dry climate, e.g., Iran, multiphase plaster is also used as mortar for brickwork and for external claddings.

5.4. Other Uses

Anhydrite capable of setting and α-hemihydrate plaster are extensively used as binders for mining mortar and as flooring plaster, uses where their brittleness is not a handicap [125].

Substantial amounts of natural anhydrite are consumed by the mining industry for the construction of roof supports for galleries and ventilating air structures in coal mines. Dry *mining anhydrite* having a specified particle size distribution, e.g., 0 – 6 mm, and containing an accelerator is used in underground coal mines worked by the long-wall method. With the aid of large blowers the anhydrite is conveyed pneumatically through hoses and wetted at the site of use before being sprayed in layers to form a barrier. The anhydrite barriers rapidly develop a high initial compressive strength of about 15 N/mm^2, their strength exceeding 40 N/mm^2 after 28 d. They serve to support the gallery and roofs in the roadways [126]. α-Hemihydrate plaster is also used as a binder for mining mortars [127].

Table 11. Application properties of projection plaster, bonding plaster, and premixed plaster∗ [10]

Property	Projection plaster	Bonding plaster	Lightweight plaster
Water : plaster ratio	0.45 – 0.65	0.60 – 0.80	0.55 – 0.75
Initial setting, min	60 – 120	40 – 90	40 – 90
Final setting, min	120 – 240	60 – 120	60 – 120
Strength of set and hardened gypsum, N/mm^2			
Flexural strength	1 – 2	1 – 2	1 – 2
Compressive strength	2.5 – 5	2.5 – 4	2.5 – 4
Apparent density of dry gypsum, kg/m^3	1000 – 1200	850 – 1000	900 – 1100
Coverage, m^2 per 100 kg of plaster	> 60	> 120	> 110

∗ Tests conform to DIN standard 1168 [57].

Finely ground natural anhydrite, processed fluoroanhydrite and α-hemihydrate plaster, all capable of setting, are used as binders for self-levelling floor plasters. When used for laying floor screeds, the cementitious material is first mixed with sand in the proportions of 2 : 5 and then batch-mixed with water in a pan mixer until it attains a stiff consistency. Piston pumps and hoses pump the mix to the different floors and rooms, where it is spread on the floors. More recently α-hemihydrate plaster, with or without additives, mixed with water to form a fluid slurry, has been used as machine-applied floor plaster (self-levelling floor plaster). Since calcium sulfate floor screeds do not shrink, they can be laid free of joints and cracks over a large surface area.

Substantial quantities of natural gypsum, FGD gypsum, and anhydrite rock are used as a retarder for Portland and blast-furnace *cements.* In compliance with German (DIN) standards, the maximum permissible calcium sulfate content of cement is 3 wt % SO_3 for Portland cement and 4 wt % SO_3 for blast-furnace cement [128]. Gypsum rock, FGD gypsum briquettes, or anhydrite rock, or a mixture with a particle size range 5 – 50 mm, is directly added to the cement clinker before grinding. It is then present in the cement in a finely divided reactive form [129]. More than 50 % of the entire gypsum and anhydrite rock production is consumed by the cement industry (see Chap. 7).

In the Russia a *gypsum cement* mortar has been manufactured by calcination of raw gypsum in combination with certain types of cement and active silica. Because of its water-resistant properties, this mortar was used for the production of prefabricated washrooms, occasionally also for exterior wall elements of prefabricated cottages. However, pure gypsum building components have not found widespread use externally because of the solubility of gypsum in water [12], [130], [131].

α-Hemihydrate is used as a high-strength *molding material* for roofing tile, cast metal, and dental materials [15]. Mixed with β-hemihydrate, α-hemihydrate is the favored molding plaster – especially for ceramics – because its expansion on setting and the strength and absorbency of the final mold can be varied by varying the $\alpha : \beta$ hemihydrate ratio [132]. Addition of high molecular mass compounds perhaps improves the stability of gypsum molds without adversely affecting the plaster. Molding plaster is also used to make surgical casts and orthopedic bandages.

Calcium sulfate dihydrate is used as a soil stabilizer and to displace sodium in soils too high in sodium salts, e.g., those soils flooded by seawater. It also serves as a fertilizer in soils deficient in sulfur, especially in North America [133].

The classical Müller-Kühne process (gypsum – sulfuric acid – cement process) has once again become of interest for treating secondary materials such as unrefined calcium sulfite and calcium sulfate from FGD or SAP spray absorption residues (OECD Amber List AB 150) [52], gypsum wallboard or plasterboard from demolished buildings (OECD Green List GG 20) [52], or fly ash. Today in Wolfen, Germany, 100 000 t/a of cement is produced by the Müller – Kühne process [134].

Unrefined calcium sulfite and calcium sulfate from FGD or SAP spray absorption residues, together with fly ash and activating additives, can be converted to a stabilized back-filling material for landfills and mining caverns [135].

The Merseburg process for the manufacture of ammonium sulfate from natural gypsum or phosphogypsum has never been important because the demand for ammonium sulfate has always been met from the large quantities arising as byproduct in the chemical industry [136].

Anhydrite III is an effective drying agent (Drierite) in laboratories and in industry.

Fillers made from FGD gypsum have been investigated thoroughly and their application in adhesives, paints and plastics gave good results. In principle, it is possible to produce calcium sulfate fillers from FGD gypsum, but the economic situation compared to competitive fillers is unsatisfactory [137]. Also coating pigments produced from FGD gypsum for the paper industry have been developed, showing new fields of application of FGD gypsum outside the construction industry [138]. Natural gypsum and anhydrite are insufficiently pure and therefore not suitable for producing fillers and coating pigments. In the beverage industry, especially breweries, pure natural gypsum is used to obtain and standardize the desired water hardness (→ Beer, Chap. 2.5.). Processes developed in the United States and Japan for the manufacture of calcium sulfate whiskers have not yet become established [139].

Pure finely ground anhydrite has found a ready market in the glass industry as a substitute for sodium sulfate because of its low price.

5.5. Properties of Gypsum Building Products Installed in Situ

Gypsum building materials are suitable for the construction of non-load-bearing interior finishes in dry locations. Gypsum building components and prefabricated units are dried at the factory. Gypsum plasters dry after application, with ventilation within a few days. These plasters adhere to the base well [140], [141]. Once gypsum building materials are sufficiently dry, i.e., moisture content < 2 %, they can be decorated immediately with interior coatings, wallpaper, ceramic tiles, and other facings. The surfaces of gypsum building materials are inert [142].

Table 12 summarizes some of the most important properties of gypsum building products installed in situ [10], [143]. Set and hardened gypsum has low density and high pore volume and thus has a low thermal conductivity. The large pore size allows rapid absorption and desorption of water vapor, resulting in high breathing capacity. Gypsum walls and ceilings in living and working rooms feel warm to the touch; they do not sweat. Figure 15 shows a scanning electron micrograph of the structure of set and hardened gypsum.

Table 12. Gypsum building materials and examples of their properties

	Plasterboard, 9.5 mm thick	Finished gypsum projection plaster
Apparent density and porosity		
Density, kg/m^3	900	800 – 1200
Pore volume, %	60∗	50 – 65
Pore radius, µm	ca. 99 % > 0.05	ca. 99 % > 0.05
Climatic properties		
Heat conductivity λ, WK^{-1}m^{-1} (DIN 4108)	0.21	0.35
Heat penetration coefficient b (equivalent)		
1/2-h test, J s$^{-1/2}$ m^{-2} K^{-1}	412	1200
2-h test	489	1400
Resistance to transmission of water vapor μ (DIN 4108)	8	10
Water vapor absorption coefficient a, m/h		
untreated surface	2.29	
rough fiber coated	2.98	2.5
wallpapered	2.66	
Heat expansion and equilibrium moisture content		
Linear coefficient of thermal expansion α_t, K^{-1}	(13 – 20) $\times 10^{-6}$	20 $\times 10^{-6}$
Equilibrium moisture content at 20 °C and 65 % relative humidity, wt %	0.6 – 1.0	0.3
Modulus of elasticity, N/mm^2	2000 – 2800	2800

∗ Core.

Gypsum building materials have volume stability and undergo only slight changes in size with variations in temperature or moisture content. Their equilibrium moisture content is below 1 %; however, permanent wetting and continuous exposure to temperatures above 60 °C change the crystal structure, and should be avoided.

Gypsum building materials promote fire protection because of the combined water of the dihydrate, which, in case of fire, evaporates, not allowing the other surface of the gypsum to reach much above 100 °C. Plasterboard and fiber-reinforced gypsum boards retain their structural

integrity during exposure to fire because of the fibers in the material [9], [144].

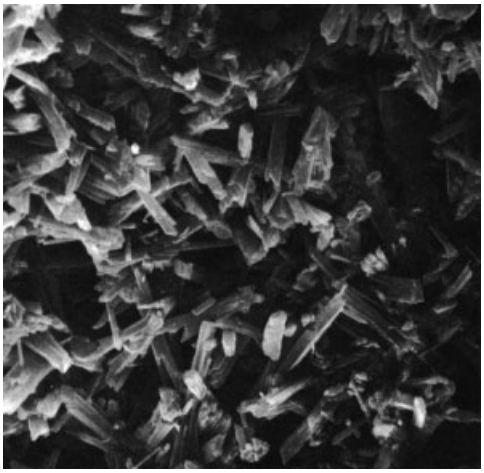

Figure 15. Structure of set and hardened gypsum (scanning electron micrograph). The structure shown in the photograph of set hemihydrate plaster is typical of rehydrated gypsum: the characteristic features are determined by needle-like particles that interlock and intergrow and by the high degree of porosity

6. Material Testing and Chemical Analysis

6.1. Standards

Today there are international standards for gypsum products (ISO – International Standards Organisation), European standards (CEN – Comité Européen de Coordination des Normes), and national standards (e.g., DIN, Germany; AFNOR, France; B.S., United Kingdom; ASTM, United States; J.I.S., Japan).

Eurogypsum, the Association of European Gypsum Industries, through its scientific and technical committee, constantly exchanges information with the ISO [145] and RILEM, the International Union of Testing and Research Laboratories for Materials and Structures. The German standards for gypsum comprise material standards [57], [146–149] and application standards [150–163].

EN standards for gypsum building materials are in preparation. In future, quality management and assurance sytems that comply with ISO 9000 – 9004 will be introduced into the gypsum industry.

6.2. Testing

The DIN standard test methods for gypsum building plasters comprise tests on particle size by sieving, initial set with a water : plaster ratio defined by the water capacity of the plaster or the spread in a flow table test, flexural strength, compressive strength, and hardness. A test for bond strength, e.g., for bonding plaster to concrete, is also recommended. Gypsum partition panels and gypsum ceiling tiles are characterized by apparent density, dimensions, weight, and breaking load. Plasterboards are characterized by dimensions, weight, breaking load, and permanent deflection after specified loading and unloading.

A method for determining the Blaine specific surface area is used for comparative testing [164]. A simple and inexpensive device for the determination of impact strength can be used for comparative and routine testing, especially for building components [165], [166].

The ceramic industry determines additional properties, such as setting and final expansion and water absorption coefficient. A standardized vacuum mixer is used to gauge the gypsum slurry and to remove air bubbles from the mixture [167].

Classification of gypsum building materials with regard to their behavior during fires is set out in DIN 4102. Gypsum building materials come within Class A, rated as noncombustible building materials. Materials with no combustible components (such as plaster, partition panels, and ceiling tiles) are classified as Class A 1. Gypsum plasterboard is classed as Class A 2. For materials classified under Class A 2 there are limits for smoke density and toxicity.

Gypsum building materials and components that comply with DIN 4102 are ideal for structures ranging from fire-retardant to highly fire-resistant walls, partitions, ceilings, steel columns, and supporting beams. Test certificates issued for built-in components are based on compliance with Classes F 30, F 60, F 90, F 120, and F 180 – fire retardant to highly fire resistant. The figures 30 – 180 denote the fire resistance rating (period of time in minutes) of the components for a given temperature – time curve.

6.3. Chemical Analysis

Conventional chemical analysis [3] is used in arbitrational analysis. Free water content in raw gypsum, calcined gypsum, and set and hardened gypsum is determined by drying at 45 °C. Combined water content is determined by drying at 360 °C.

Calcium sulfate in solution is qualitatively determined by precipitation of syngenite, $CaSO_4 \cdot K_2SO_4 \cdot H_2O$.

Today the following methods of analysis are used for routine quantitative testing in the laboratories of the gypsum industry: complexometric titration of calcium and magnesium; flame photometry for sodium and potassium; and atomic absorption for sodium, potassium, magnesium, calcium, strontium, iron, aluminum, and sulfate, the last indirectly, via barium [168]. The phosphates contained in phosphogypsums are determined spectrophotometrically or gravimetrically with ammonium molybdate. Water-soluble, citrate-soluble, and total phosphates are distinguished. Fluorine can be determined titrimetrically. Organic substances can be determined by oxidation with potassium permanganate. acidity is determined by titration with alkali using methyl orange or bromophenol blue. Calcium sulfite is determined via SO_2; chlorine, by the Volhard method or with a chloride electrode.

6.4. Phase analysis

Phase analysis differentiates between dihydrate, hemihydrate, anhydrite III, and anhydrite II, including its three reaction stages AII-s, AII-u, and AII-E. This differentiation of calcined gypsum according to phases is not possible by chemical analysis.

Phase analysis can be carried out by X-ray diffraction, infrared spectroscopy, microscopy, and calorimetry. However, it takes considerable experience to reach ± 5 % accuracy [169]. The conventional method in any industrial laboratory is the gravimetric determination utilizing the differing rates of hydration [1], [3], [170], [171].

The determination of dihydrate is not especially accurate, particularly when small quantities (< 5 %) are involved. In such cases, however, an exact determination is possible by differential thermal analysis [172], [173].

The determination of anhydrite II by exposure to moisture is a matter of definition. Thus the anhydrite hydrating within 3 d is designated as AII-s, anhydrite hydrated in 7 d as AII-u, and anhydrite that remains nonhydrated after 7 d is assigned to the nonhydrated constituents and can be determined by an additional measurement. (By the addition of a 1 % aqueous potassium sulfate solution in place of distilled water the seven-day period can be reduced to a few hours [1].) The stages AII-s and AII-E can be distinguished from each other by measurements of the pH value, which for AII-E is > 9, but is otherwise 6.

Distinction between α- and β- hemihydrates is not possible by hydration methods. Today α-hemihydrate and β-hemihydrate are usually determined by microscopy, the quantitative proportions are simply estimated.

7. Economic Aspects

The estimated world consumption of natural gypsum and natural anhydrite in the gypsum and cement industries was ca. 95×10^6 t/a in 1995.

The cement industry, as the largest consumer of raw gypsum and raw anhydrite, used ca. 55×10^6 t/a for a world cement production of 1400×10^6 t in 1995.

The gypsum industry uses ca. 40×10^6 t/a natural gypsum and anhydrite and ca. 5×10^6 t/a FGD gypsum. Less than 2×10^6 t/a of phosphogypsum is used, mainly in Japan, Korea, and Europe. Titanogypsum and fluoroanhydrite consumption is less than 10^6 t/a worldwide.

The most important gypsum building materials are gypsum plasterboards with a per capita consumption of 8 m^2/a in the United States; 6 m^2/a in Japan; and 2 m^2/a in the EU, with the highest consumption in France and Great Britain (ca. 3 m^2/a), and the lowest in Spain and Portugal (0.35 m^2/a). Machine-applied plaster is important in Central Europe, and multiphase plaster in Spain, Italy, North Africa, and the Near East.

Since 1985 the demand for calcium sulfate has increased at an average annual rate of 5 % in Asia, 3 % in Western Europe, and 1 % in North America [174].

The gypsum and cement industries will remain the prime consumers. In only a few countries, perhaps China and the countries in Eastern Europe, the states around the Persian Gulf, and some rapidly developing areas in the Far East, is gypsum consumption expected to rise in response to increased building activities.

8. Recycling and Disposal

As with other materials, the life cycle of the products of the gypsum building materials industry consists of winning of raw materials and their treatment, production, use, and recycling or disposal:

- Gypsum-containing *production residues* are recycled in the production plant by returning them to the production stream.
- Building site *residues from construction plasters* that are free of foreign components and sorted according to type can be returned to the producer or sent to regional collection facilities.
- Regional collection facilities are planned for building site *residues from gypsum building components* that are free of foreign components. These residues will be treated and then returned to the production process.
- *Residues from demolition and renovation* are to be collected at treatment plants and so processed that they become suitable for recycling. Plants of this type are already in operation in Germany.
- Used *gypsum molds and models* from the ceramics industry, provided they are free of foreign components, can be taken back by agreement with the gypsum supplier and used as secondary raw materials.
- Used *models from dental laboratories* and from modelling can be treated in the same way
- Non-recyclable gypsum-containing residues can be disposed of in landfills for building wastes. In Germany this is governed by the TA-Siedlungsabfall and the Kreislaufwirtschafts- und Abfallgesetz [10], [175], [176].

In general, for the disposal of gypsum-containing wastes in landfills, the influence of the gypsum residues on ground water and surface water is comparable to the influence of natural gypsum in deposits.

9. References

General References

1. H. E. Schwiete, A. N. Knauf: *Gips. – Alte und neue Erkenntnisse in der Herstellung und Anwendung der Gipse,* Merziger Druckerei u. Verlags-GmbH, Merzig 1969.
2. *Winnacker – Küchler,* 4th ed., **vol. 3,** pp. 262 – 277.
3. A. Voellmy, W. Albrecht: "Die Prüfung der Gipse und Gipsmörtel," in O. Graf (ed.): *Handbuch der Werkstoffprüfung,* 2nd ed., **vol. 3,** Springer Verlag, Berlin 1957, pp. 520 – 576.
4. *Gmelin,* system no. 28, Calcium, Main B 3 (1961), pp. 675 – 785.
5. A. W. Groves: *Gypsum and anhydrite, Overseas Geological Surveys,* Her Majesty's Stationery Office, London 1958.
6. F. Lotze: *Steinsalz und Kalisalze, Geologie. Die wichtigsten Lagerstätten der "Nichterze",* 2nd ed., 1st part, Verlag Gebr. Bornträger, Berlin 1957.
7. A. V. Slack: *Phosphoric acid,* Marcel Dekker, New York 1968, part I, part II.
8. M. Sekiya: *Gypsum,* K. K. Gihodo, Tokio 1964.
9. K. Volkart: *Bauen mit Gips,* Bundesverband der Gips- u. Gipsbauplatten-Industrie, Darmstadt 1981.
10. *Gips-Datenbuch,* Bundesverband der Gips- und Gipsbauplattenindustrie, 1995.
11. W. Albrecht: *Stuckgips und Putzgips, Fortschritte und Forschungen im Bauwesen,* no 15, Franckh'sche Verlagshandlung, Stuttgart 1953.
12. A. V. Volzenskij, A. V. Ferronskaja: *Gips, Bindemittel und Erzeugnisse (Technologie, Eigenschaften, Anwendung),* Stroijizdat, Moskau 1974.
13. S. D. Ewjentschik, A. A. Nowikow: *Phosphogips und seine Anwendung,* Moskau ″Chimija″ 1990.
14. *Die Strahlenexposition von außen in der Bundesrepublik Deutschland durch natürliche radioaktive Stoffe im Freien und in Wohnungen unter Berücksichtigung des Einflusses von Baustoffen,* Der Bundesminister des Innern, Bonn 1978.

15. G. Franz: *Dentalgipse,* Hanser Verlag, München – Wien 1981.
16. H. Hanusch: *Gipskartonplatten Trockenbau-Montagebau-Ausbau,* Verlagsgesellschaft Rudolf Müller, Köln-Braunsfeld 1978.
17. G. Beard: *Stuck – Die Entwicklung plastischer Dekoration,* Schuler Verlagsgesellschaft, Herrsching 1983.
18. M. Murat, M. Foucault: "Sulfates de Calcium et Matériaux Dérivés," *Colloques Internationaux de la RILEM,* Paris 1977.

Specific References

19. R. M. Gruver, *J. Am. Ceram. Soc.* **34** (1951) 353 – 357.
20. E. Eipeltauer, *Zem. Kalk Gips* **11** (1958) 264 – 272, 304 – 316.
21. W. C. Riddell, *Rock Prod.* **53** (1950) 68 – 71, 102.
22. H. Lehmann, K. Rieke, *Tonind. Ztg. Keram. Rundsch.* **97** (1973) 157 – 159.
23. S. M. Mehta: *Dissertation,* Techn. Universität, Clausthal 1974.
24. K. Reisdorf, W. Abriel, *Neues Jahrbuch Miner. Abh.* **157** (1987)35 – 46.
25. W. Abriel, K. Reisdorf, J. Parnnetier, *J. Solid State Chem.* **85** (1990) 23 – 30.
26. C. Bezou, A. Nonat, J-C. Mutin, A. Nørlund Christensen, M. S. Lehmann, *J. Solid State Chem.* **117** (1995) 165 – 176.
27. B. F. Pedersen, D. Semmingsen, *Acta Crystallogr. Sect. B* **38** (1982) 1074 – 1077.
28. A. Kirfel, G. Will, *Acta Crystallogr. Sect. B* **36** (1980) 2881 – 2890.
29. G. C. H. Cheng, J. Zussman, *Acta Crystallogr.* **16** (1963) 767 – 769.
30. F. Wirsching, R. Hüller, B. Limmer, *Zem. Kalk Gips* **47** (1994) 278 – 286.
31. K. K. Kelley, J. C. Southard, C. T. Anderson: *U. S. Bur. Mines Tech. Paper* **625** (1941).
32. A. Kruis, H. Späth, *Tonind. Ztg. Keram. Rundsch.* **75** (1951) 341 – 351, 395 – 399.
33. H.- J. Kuzel, M. Hauner, *Zem. Kalk Gips* **40** (1987) 628 – 632.
34. Nippon Kokan Kabushiki, GB 1016007, 1962 (K. Araki).
35. H. J. Förster, *Chem. Ing. Tech.* **44** (1972) 969 – 972.
36. D. Kitchen, W. J. Skinner, *J. Appl. Chem. Biotechnol.* **21** (1971) 53 – 55, 56 – 60, 65 – 67.
37. E. Eipeltauer, *Tonind. Ztg. Keram. Rundsch.* **97** (1973) 4 – 8.
38. U. Ludwig, *Zem. Kalk Gips* **21** (1968) 81 – 90, 109 – 119, 175 – 180.
39. S. Sprung, *Zem. Kalk Gips* **27** (1974) 259 – 267.
40. G. Leifeld, W. Münchberg, W. Stegmaier, *Zem. Kalk Gips* **23** (1970) 174 – 177.
41. D. Hass, E. Kemnitz, B. Grunze, R. Sekowski, H. Worzala, *Z. Anorg. Allg. Chem.* **525** (1985) 173 – 178.
42. H. Borchert, E. Baier, *Neues Jahrb. Mineral. Abh.* **86** (1953) 103 – 154.
43. E. Posnjak, *Am. J. Sci.* **238** (1940) 559 – 568.
44. A. Herrmann, *Silik. J.* **3** (1964) 443 – 466.
45. D. Ostwald, *Tonind. Ztg. Keram. Rundsch.* **78** (1954) 137 – 142, 173 – 177.
46. F. Wirsching, R. Hüller, R. Olejnik, *Zem. Kalk Gips* **47** (1994)65 – 69; *Zem. Kalk Gips* **47**(1994) 683 – 688.
47. H. Hamm, *Zem. Kalk Gips* **47** (1994) 443 – 451.
48. J. Beckert, H. J. Einbrodt, M. Fischer, *Bericht und gutachterliche Stellungnahme zu FGD Gypsum and Natural Gypsum,* VGB-Forschungsstiftung Essen; Bundesverband der Gips- und Gipsbauplattenindustrie Darmstadt (1990).
49. *Proceedings from the 4th International Conference on FGD and Synthetic Gypsum,* Ortech Corporation, Mississauga, Ontario, Canada.
50. W. Kumpf, K. Maas, H. Straub: *Handbuch Müll- und Abfallbeseitigung: Rückstände aus der Phosphorsäureproduktion,* **vol. 5,** E. Schmidt Verlag, Berlin 1964, no. 8581 (1982).
51. Official Journal of the European Communities L5/22, Vol. 37, 7 January 1994.
52. Official Journal of the European Communities L288/36-46, Vol. 37, 9 November 1994.
53. E. Eipeltauer, S. Stojadinovic, *Ber. Dtsch. Keram. Ges.* **37** (1960) 442 – 447.
54. J. R. Gunn, *Gypsum J.* **49** (1968) 14 – 18.
55. W. Kreuter, *Zem. Kalk Gips* **27** (1974) 222 – 225.
56. P. Joest, *Ind. Steine Erden* **71** (1961) 23 – 29; *Glückauf* **92** (1956) 504 – 506.
57. DIN 1168,Gypsum building plasters, latest version.
58. BPB Ind., DE 1258321, 1962 (R. C. Blair).
59. A. J. T. Ward, *Zem. Kalk Gips Ed. B* **33** (1980) 594 – 600.
60. W. Lahl, H. E. Schwiete, *Zem. Kalk Gips* **12** (1959) 345 – 351, 582.
61. J. Steinkuhl, O. Wiechmann, K. Moldan, *Zem. Kalk Gips* **25** (1972) 383 – 386.
62. E. Eipeltauer, Ch. Stein, *Zem. Kalk Gips* **52** (1963) 45 – 53.

63. P. Sörgel, J. Bergmann, G. Fietsch, *Silikattechnik* **22** (1971) 225 – 230.
64. National Gypsum Co., US 2788960, 1954 (S. D. Skinner).
65. Gebrüder Knauf, DE 1143430, 1961 (A. N. Knauf).
66. H. Hamm, F. Wirsching, *Zem. Kalk Gips* **27** (1974) 226 – 229.
67. P. Landrieu, J. Gibaru, C. Collomb, *Zem. Kalk Gips* **17** (1964) 455 – 460.
68. D. Roddewig: "Erfahrungen beim Brennen von Gips im Kompaktdrehofen," *15th Eurogypsum Congress,* Venetia, Sept. 1982.
69. Roddewig & Co., CH 445359, 1964 (H. Roddewig, sen.).
70. D. Böcker, H. Kreusing, *Zem. Kalk Gips Ed. B* **34** (1981) 221 – 226.
71. H. Rennen, M. Grunwald, G. Hilscher, *Zem. Kalk Gips* **42** (1989) 217 – 222.
72. H. Jurkowitsch, R. Hüller, *Zem. Kalk Gips* **43** (1990) 583 – 588.
73. H. Scholze, M. Hurbanic, R. Conradt, *Zem. Kalk Gips* **42** (1989) 248 – 251.
74. F. Wirsching, *Umwelt* **13** (1983) 435 – 438.
75. A. N. Knauf, *Zem. Kalk Gips* **36** (1983) 271 – 274.
76. H. Hamm, *Zem. Kalk Gips* **44** (1991) 593 – 604.
77. Dravo Lime Company, US 5312609, 1992 (U. W. Collegs).
78. A. Kurandt, *ISMA Tech. Conf.* 1974.
79. Fisons, GB 1135951, 1966 (N. Robinson).
80. Nissan Kakasu KKK, US 3653826, 1968 (T. Ishihara).
81. Société de Prayon, DE 1567821, 1966 (E. Pavonet).
82. "Getting Rid of Phosphogypsum I – IV," *Phosphorus Potassium* **87** (1977) 37; **89** (1977) 36; **94** (1978) 24; **96** (1978) 30.
83. *International Symposium on Phosphogypsum,* 5 – 7 Nov. 1980, Florida Institute of Phosphate Research (47 Papers).
84. J. Beretka, T. Brown, *J. Chem. Technol. Biotechnol.* **32** (1982) 607 – 613; **33 A** (1983) 299 – 308.
85. "Supply-Demand Trend of Gypsum, Lime and Cement in 1994", *Inorganic Materials* **2** (1995) no. 258, 426 – 438.
86. D. Israel, *ZKG International* **49** (1996) 228 – 234.
87. K. Schaupp, K. Metz, *Zentralbl. Industriebau* **4** (1965) 180 – 184.
88. M. H. Le Chatelier, *C. R. Hebd, Séances Acad. Sci.* **96** (1883) 1668 – 1671.
89. J. H. Van't Hoff, E. F. Armstrong, W. Hinrichsen, F. Weigert, G. Just, *Z. Phys. Chem. Stoechiom. Verwandtschaftsl.* **45** (1903) 257 – 306.
90. J. R. Clifton, *Report NBS-TN* **755** (1973) 1 – 28.
91. M. J. Ridge, J. Beretka, *Rev. Pure Appl. Chem.* **19** (1969) 17 – 44.
92. W. Krönert, P. Haubert, *Zem. Kalk Gips* **25** (1972) 553 – 558.
93. A. Cavazzi, *Z. Chem. Ind. Kolloide* **12** (1913) 196 – 201.
94. M. Baykoff, *C. R. Hebd, Séances Acad. Sci.* **182** (1926) 128 – 129.
95. K. W. Fischer, *Wiss. Z. Hochsch. Archit. Bauwes. Weimar* **10** (1963) 351 – 371.
96. W. Krönert, P. Haubert, unpublished results, RWTH Aachen.
97. U. Ludwig, J. Kuhlmann, *Tonind. Ztg. Keram. Rundsch.* **98** (1974) 1 – 4.
98. H. Lehmann, H. Mathiak, P. Kurpiers, *Ber. Dtsch. Keram. Ges.* **50** (1973) 201 – 204.
99. M. K. Lane, *Rock Prod.* **71** (1968) no. 3, 60 – 63, 108; **71** (1968) no. 4, 73 – 75, 116, 117.
100. R. A. Kuntze, *Mater. Res. Stand.* **7** (1967) 350 – 353.
101. Certain-Teed Products Corp., US 2067762, 1934 (G. A. Hoggatt).
102. B. Lelong, *Zem. Kalk Gips Ed. B* **37** (1984) 205 – 218.
103. Gebrüder Knauf, DE-AS 2023853, 1970 (A. N. Knauf).
104. A. Aignesberger, H. Krieger, *Zem. Kalk Gips* **21** (1968) 415 – 419.
105. Imperial Chem. Ind., GB 1049184, 1963 (K. G. Cunningham).
106. Imperial Chem. Ind., DE 1126792, 1959 (K. G. Cunningham).
107. G. Benz, *Stuckgewerbe* 1969, no. 12, 533 – 544.
108. A. N. Knauf, W. Krönert, P. Haubert, *Zem. Kalk Gips* **25** (1972) 546 – 552.
109. T. Koslowski, U. Ludwig: *Zitronensäure,* Institut für Gesteinshüttenkunde der RWTH, Aachen (1983).
110. K. Aichinger, B. Wandser, *Zem. Kalk Gips* **1** (1948) 33 – 37, 50 – 51.
111. H. Engelke, *Zem. Kalk Gips* **32** (1979) 560 – 568.
112. M. Rößler, *Dissertation,* Techn. Universität, Clausthal 1983.
113. B. Wandser, *Zem. Kalk Gips* **15** (1962) 437 – 438.
114. M. J. Ridge, H. Surkevicius, *J. Appl. Chem.* **11** (1961) 420 – 434.

115. M. Murat, *Tonind. Ztg. Keram. Rundsch.* **97** (1973) 160 – 164; **98** (1974) 33 – 37, 73 – 78.
116. G. Neuhauser, *Bundesbaublatt* **31** (1982) 566 – 569.
117. E. Aeppli, *Eurogypsum,* Stockholm 1972.
118. N. W. Knauf, DE 1104419, 1957 (A. N. Knauf).
119. FERMA Gesellschaft für Rationelle Fertigbaumethoden und Maschinenanlagen mbH & Co., CH 505674, 1969 (K. Schäfer).
120. M. A. Ali, F. J. Grimer, *J. Mater. Sci.* **4** (1969) 389 – 395.
121. J. Kazimir, *Tonind. Ztg. Keram. Rundsch.* **91** (1967) 22 – 25.
122. G. Kossatz, K. Lempfer, *Holz Roh. Werkst.* **40** (1982) 333 – 337.
123. G. Kossatz, *Baustoffindustrie* **9** (1966) 1 – 5.
124. A. N. Knauf, *Stuckgewerbe* 1961, no. 3, 1 – 4.
125. Forschungsgemeinschaft Bauen und Wohnen: *Estriche im Hochbau,* **no. 80,** Verlagsgesellschaft R. Müller, Köln – Braunsfeld 1966.
126. F. Henrich, *Glückauf* **107** (1971) 1 – 13; *Neue Bergbautech.* **9** (1979) 409.
127. P. Thien, *Glückauf* **128** (1992) 750 – 757.
128. DIN 1164,Portland-, Eisenportland-, Hochofen- und Traßzement, latest version.
129. H. E. Schwiete, U. Ludwig, P. Jäger, *Zem. Kalk Gips* **17** (1964) 229 – 236.
130. T. Matyszewski, G. Ambrozewicz, *Baustoffindustrie, Ausg. A* **16** (1973) no. 5, 18 – 20.
131. Stroitelnych i Njerudnych Materialow, DE 1241330, 1963 (P. W. Lapschin).
132. W. Reingen, *Zem. Kalk Gips* **27** (1974) 252 – 258.
133. J. C. Rinehart, G. R. Blake, J. C. F. Tedrow, F. E. Bear, *Bull. N.J. Agric. Exp. Stn.* **772** (1953).
134. "The WSZ Schwefelsäure und Zement GmbH - Unique or a Model for the Future?" *ZKG International* **49** (1996) A33 – A34.
135. F. Wirsching, E. Weißflog, *VGB Kraftwerkstechnik* **68** (1988)1131 – 1141; 1269 – 1278.
136. E. Sacher, *ISMA Tech. Conf.* 1968.
137. F. Wirsching, R. Hüller, H. Hamm, H. Hoffmann, A. Pürzer, *ZKG International* **48** (1995) 241 – 256.
138. H. Hamm, K. Kraft, J. Trummer, F. Wirsching, *Wochenblatt für Papierfabrikation* **123** (1995) 88 – 94.
139. Franklin Key, Inc., US 3822340, 1974 (J. J. Eberl).
140. H. Scholze, M. Hurbanic, H. Engelke, *Zem. Kalk Gips Ed. B* **34** (1981) 318 – 338.
141. H. Scholze, *Tonind. Fachber.* **108** (1984) 170 – 172.
142. H. Haagen, *Farbe Lack* **87** (1981) 543 – 550.
143. F. Wirsching, *ZKG International* **44** (1991)248 – 252.
144. D. Schumann, *Bauwirtschaft* **26** (1972) 88 – 97.
145. Ch. Collomb, *Zem. Kalk Gips* **17** (1964) 451 – 454.
146. DIN 4208, Anhydrite Binder, latest version.
147. DIN 18163, Gypsum Partition Panels, latest version.
148. DIN 18169, Gypsum Ceiling Panels, latest version.
149. DIN 18180, Gypsum Plasterboards, latest version.
150. DIN 4102, Behavior of Building Materials and Components in Fire (Parts 1–4), latest version.
151. DIN 4103, Non-Loadbearing Partitions (Parts 1, 2+4), latest version.
152. DIN 4108, Thermal Insulation in Buildings (Parts 1–5), latest version.
153. DIN 4109, Noise-Control in Buildings, latest version.
154. DIN 18168, Light Ceiling Linings and Underceilings (Parts 1+2), latest version.
155. DIN 18181, Gypsum Plasterboards for Building Construction, latest version.
156. DIN 18182, Accessories for the use of Gypsum Plasterboards (Parts 1–4), latest version.
157. DIN 18183, Lightweight Partitions of Gypsum Plasterboards, latest version.
158. DIN 18184, Gypsum Plaster Sandwich Boards, latest version.
159. DIN 18350, Plaster and Stucco Works, latest version.
160. DIN 18550, Plaster and Rendering (Parts 1–4), latest version.
161. DIN 18555, Testing of Mortars with Mineral Binders (Parts 1 – 3+7), latest version.
162. DIN 18560, Floor Screeds in Building Construction (Parts 1 – 4), latest version.
163. DIN 13911, DIN EN 26873Dental Materials; Gypsum; Requirements, Testing, latest version.
164. L. Chassevent, *Zem. Kalk Gips* **15** (1962) 509 – 512.
165. F. Wirsching, W. Poch, *Zem. Kalk Gips* **27** (1974) 240 – 244.
166. Wissenschaftlich-Technische Kommission von Eurogypsum, *Tonind. Ztg. Keram. Rundsch.* **97** (1973) 145 – 157.
167. K. Litzow: *Handbuch der Keramik, Gruppe Va, 7,* Verlag Schmid, Freiburg i. Brg. 1969.

168. G. Neuhauser, *Zem. Kalk Gips* **27** (1974) 240 – 244.
169. H. Lehmann, *Tonind. Ztg. Keram. Rundsch.* **91** (1967) 8 – 14.
170. *Ullmann,* 4th ed., **vol. 12,** p. 312.
171. *Phasenanalyse von Gips,* Forschungsvereinigung der Gipsindustrie e.V., Darmstadt (1988).
172. R. A. Kuntze, *Mater. Res. Stand.* **2** (1962) 640 – 642.
173. V. Schlichenmaier, *Tonind. Ztg. Keram. Rundsch.* **98** (1974) 223.
174. A. Mields, *Intern. Cem. Rev.,* Nov. 1995, 20 – 29; 31 – 35.
175. TA-Siedlungsabfall, Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit, Bonn, June 1993.
176. Kreislaufwirtschafts- und Abfallgesetz – KrW-/AbfG, Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit, Bonn, September 1994.

Californium → Transuranium Elements

Cancer Chemotherapy

Bernhard Kutscher, ASTA Medica AG, Frankfurt am Main, Federal Republic of Germany

Gregory A. Curt, National Cancer Institute, Bethesda, Maryland 20205, United States

Carmen J. Allegra, National Cancer Institute, Bethesda, Maryland 20205, United States

Robert L. Fine, National Cancer Institute, Bethesda, Maryland 20205, United States

Hamza Mujagic, National Cancer Institute, Bethesda, Maryland 20205, United States

Grace Chao Yeh, National Cancer Institute, Bethesda, Maryland 20205, United States

Bruce A. Chabner, National Cancer Institute, Bethesda, Maryland 20205, United States

1. Introduction . . . 2
2. Antimetabolites . . . 3
2.1. Methotrexate . . . 3
2.1.1. Mechanism of Action and Mechanisms of Resistance . . . 3
2.1.2. Analogs . . . 5
2.2. Fluoropyrimidines . . . 6
2.2.1. Mechanism of Action . . . 6
2.2.2. Mechanisms of Resistance . . . 7
2.2.3. Other Fluoropyrimidines . . . 7
2.3. 5-Azacytidine . . . 8
2.3.1. Mechanism of Action . . . 8
2.3.2. Mechanism of Resistance . . . 8
2.3.3. New Analogs . . . 8
2.4. Cytosine Arabinoside (Ara-C) . . . 9
2.4.1. Mechanisms of Resistance . . . 9
2.4.2. New Analogs . . . 10
2.5. Deoxycytidine and Analogs . . . 10
2.6. 2-Halopurines and Analogs . . . 11
2.7. 6-Mercaptopurine and 6-Thioguanine . . . 12
2.7.1. Mechanism of Action . . . 12
2.7.2. Mechanism of Resistance . . . 13
2.7.3. New Analogs . . . 14
3. Alkylating Agents . . . 14
3.1. Nitrogen Mustard . . . 14
3.1.1. Mechanism of Action . . . 14
3.1.2. Mechanisms of Drug Resistance . . . 15
3.2. Melphalan . . . 15
3.2.1. Mechanism of Action . . . 15
3.2.2. Mechanism of Resistance . . . 15
3.3. Cyclophosphamide . . . 16
3.3.1. Mechanism of Action . . . 16
3.3.2. Mechanism of Resistance . . . 16
3.4. Chlorambucil . . . 16
3.5. Thio-TEPA . . . 16
3.6. Ifosfamide . . . 17
3.7. Estramustine . . . 17
3.8. Nitrosoureas . . . 17
3.8.1. Mechanism of Action . . . 18
3.8.2. Mechanisms of Resistance . . . 18
3.8.3. Analogs . . . 18
3.9. Procarbazine . . . 19
3.9.1. Mechanism of Action . . . 19
3.9.2. Mechanisms of Resistance . . . 20
3.10. Dacarbazine . . . 20
3.11. Hexamethylmelamine . . . 20
3.12. Mitomycin-C . . . 20
4. Anthracyclines . . . 21
4.1. Mechanism of Action . . . 22
4.2. Mechanism of Resistance . . . 23
4.3. Analogs . . . 23
5. Intercalating Anthracenes and Analogs . . . 24
5.1. Mitoxantrone . . . 24
5.2. Analogs . . . 24
6. Antitumor Antibiotics Other than Anthracyclines . . . 25
6.1. Actinomycin D . . . 25
6.2. Bleomycin . . . 25
6.2.1. Analogs . . . 26
6.2.2. Mechanism of Action . . . 26
6.3. DNA Interactive Natural Products 26
7. Antitubulin Agents . . . 28
7.1. Vinca Alkaloids . . . 28
7.1.1. Vincristine and Vinblastine . . . 28
7.1.2. Vindesine . . . 29
7.1.3. Vinorelbine . . . 29
7.2. Podophyllotoxin and Its Derivatives . . . 30
7.3. Camptothecin and Analogs . . . 30
7.4. Taxoids . . . 31
7.5. Epothilone A and B . . . 33
8. Heavy-Metal Complexes . . . 33
8.1. *cis*-Platinum . . . 33
8.1.1. Mechanism of Action . . . 34

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_001

8.1.2. Mechanisms of Resistance 34
8.2. Carboplatin 35
8.3. Analogs 35
9. Hormonally Active Anticancer Drugs/Antihormones 36
9.1. Antiestrogens 36
9.1.1. Antagonists 36
9.1.2. Tamoxifen, Toremifene 36
9.1.3. Analogs 37
9.2. Aromatase Inhibitors 38
9.3. Antiandrogens 39
9.3.1. Flutamide 39
9.3.2. Nilutamide 40
9.3.3. Bicalutamide 40
9.4. LHRH Analogs 41
9.4.1. LHRH Agonists 41
9.4.1.1. Leuprorelin Acetate 42
9.4.1.2. Goserelin 43
9.4.2. LHRH Antagonists 43
9.4.2.1. Receptor Assays 44
9.4.2.2. Peptidomimetics 44
10. Signal Transduction Inhibitors . . 45
10.1. Enzyme Inhibitors 45
10.2. Phospholipid – Based Antineoplastics 45
11. Economic Aspects 46
12. References 46

1. Introduction

Malignant tumors represent one of the most common human diseases worldwide. Based on an estimation made in the United States, cancer will become the leading cause of death in the year 2000 [1].

Unfortunately, the subset of human cancer types that are amenable to curative treatment still is rather small. Although there is a tremendous progress in understanding the molecular events that lead to malignancy and many agents are known that effectively kill cancer cells, progress in development of clinically innovative drugs that can cure humans is slow [2], [3].

The heterogeneity of malignant tumors with respect to their genetics, biology, and biochemistry as well as primary or treatment-induced resistance to therapy hamper curative treatment [4], [5].

Searching for antineoplastic agents with improved selectivity to malignant cells remains the central task for drug discovery and development [5], [6].

According to a survey published in 1997 more then 315 drugs are under development in the United States for the treatment of cancer. This figure includes 42 drugs for treatment of lung cancer, 58 for breast cancer [8], [9], 60 for treatment of skin tumors (60), 36 for prostate cancer, and 35 for colon cancer.

In 1997 more than 1500 Americans are expected to die of cancer each day and more than a million new cases will be diagnosed with overall medical costs of $\$35 \times 10^9$. The total disease costs are estimated to sum up to more than $\$100 \times 10^9$ per year in treatment expenses and lost wages.

Increase in the incidence of cancer, mostly as a result of an aging population, is the driving growth in the marketplace. However, new medicines like hormone-analogs have lead to decrease in severity of the side effects of cancer therapy and have spurred wider use. According to a review [11] of ca. 90 approved anticancer drugs, more than 60 % are of natural origin or modeled on natural products parents. Cancer chemotherapeutics can be grouped according to their pharmacological and mechanistic profiles into

Antimetabolites
Alkylating agents
DNA-intercalating (agents) antibiotics
Mitose inhibitors
Signaltransduction-inhibitors

Two primary events in cell proliferation are DNA replication and cell division [12], [13]. The cell cycle has been divided into four sequential phases (G_0, G_1, S, M) and the cytostatics vary in the way they interfere with the cell cycle. Phasespecific agents/drugs, e.g., those interacting in G_1, S- or M-phase, are the mitose-inhibitors vincristine or vinblastine or the antimetabolites methotrexate or cytarabine, attacking in the S-phase. Alkylating drugs such as cyclophosphamide, cisplatin, or carboplatin inhibit and damage the cell in all phases and are thus phase unspecific. In general cells in the resting phase (G_0) are insensitive.

Approximately 70 % of patients diagnosed as having cancer have metastatic disease, i.e., dis-

ease that has spread beyond the primary site at the time of diagnosis [15]. However, the steady progress made in the treatment of cancer with drugs has contributed to the curing of an increasing proportion of patients with metastatic disease. The greatest change and improvements in cancer treatment have occurred because of the discovery and clinical development of drugs and the demonstration that metastatic cancer can be cured by these agents. Drugs, such as cytoxan, adriamycin, vincristine, *cis*-platinum, and bleomycin, all developed since 1960, are now regularly used by physicians to treat patients who would have been considered incurable a short time ago.

This article details the pharmacology and clinical use of the major classes of anticancer agents, including (1) antimetabolites, (2) alkylating agents, (3) anthracyclines and analogs, (4) other antitumor antibiotics, (5) antitubulin agents, (6) platinum complexes, (7) antihormones, and (8) signaltransduction-inhibitors. In each case, special emphasis is given to the progress made in developing clinically useful drugs and analogs that retain antitumor activity while decreasing host toxicity.

2. Antimetabolites

2.1. Methotrexate

Folic acid analogs comprise a class of antineoplastic agents of which methotrexate has gained the most widespread clinical use. These agents were the first to produce impressive remissions in acute leukemia [16] and cures in choriocarcinoma in women [17].

Reduced folates (tetrahydrofolates) are the biologically active form of folates required as cosubstrates in one-carbon transfer reactions. Included in these reactions are several important enzymatic steps in the de novo synthesis of purines and pyrimidines.

Methotrexate (NSC-740) [*59-05-2*], *N*-(4-{(2,4-diamino-6-pteridinyl)methyl]methylamino}benzoyl)-L-glutamic acid, $C_{20}H_{22}N_8O_5$, M_r 454.46, is a 2,4-diamino, N^{10}-methyl analog of folic acid that is capable of inhibiting certain folate-requiring reactions.

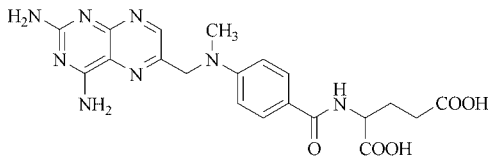

Methotrexate

Most importantly, methotrexate can inhibit dihydrofolate reductase (DHFR, $K_i = 10^{-11}$ M), a key enzyme for the maintenance of biologically active intracellular reduced folate pools. The folate-requiring reactions utilize reduced folates, and all reactions except that catalyzed by thymidylate synthase maintain folates in a reduced state during carbon transfer. Thymidylate synthase, which catalyzes the methylation of deoxyuridylate to thymidylate (required for DNA synthesis), requires the transfer of a carbon group from the folate cofactor N^{5-10}-methylene tetrahydrofolate with resultant oxidation of the folate to dihydrofolic acid. Oxidized folates must be reduced to the tetrahydro form by DHFR to be useful for intracellular metabolism. Inhibition of DHFR following methotrexate exposure ultimately leads to depletion of intracellular reduced folates. Cessation of first thymidylate and then purine nucleotide synthesis occurs as an indirect effect of methotrexate on reduced-folate levels. Methotrexate metabolites (polyglutamates) may also have direct inhibitory effects on folate-requiring enzymes, e.g., thymidylate synthase [18], and these effects may be important in inducing cytotoxicity.

2.1.1. Mechanism of Action and Mechanisms of Resistance

Transport. At concentrations less than 10 μM, methotrexate and reduced folates enter cells via an energy-dependent, temperature-sensitive carrier mechanism [19]. The affinity for this carrier has been variously reported to fall between 1 and 10 μM for tumor cell lines [19–21] and to be 87 μM for normal intestinal epithelial cells [22]. These differences in efficiency of transport may account for some of the selectivity of methotrexate for neoplastic cells. In addition to this carrier-mediated transport system, there exists a second, low-affinity trans-

port mechanism that is poorly understood but appears to play a role in the transport of drug when concentrations exceed 20 μM [21], [23]. Methotrexate and reduced folates do not compete for uptake by this process, and it may represent a means for drug entry in cells resistant to low doses of drug by virtue of defective transport by the high-affinity mechanism. In some models, the sensitivity of a cell to methotrexate can be directly correlated with efficiency of drug transport, i.e., sensitive cells have a greater capacity for drug transport and longer intracellular retention of drug when compared to methotrexate-resistant cells [24]. Because decreased membrane transport may play a role in clinical drug resistance, a number of analogs with high lipid solubility have been developed that can circumvent a transport deficit. Methotrexate esters, diaminopyrimidines, and triazenates have been synthesized and used with success against experimental cell lines with defective methotrexate transport.

Intracellular Metabolism. Once inside the cell, naturally occurring folates may be metabolized to *polyglutamates;* that is, additional glutamyl moieties are added to the terminal glutamate present on the parent compound. This process allows for selective intracellular retention of the polyglutamated forms and an increased affinity for certain folate-requiring enzymes, such as thymidylate synthase, and for enzymes required for the de novo production of purine nucleotides [25]. Like the naturally occurring folates, intracellular methotrexate is also polyglutamated with the addition of from one to four additional glutamyl residues [26]. This process has been demonstrated in a variety of tissues, including human breast cancer cell lines [26], normal human liver [27], and murine leukemia cells [28]. The polyglutamates of methotrexate are selectively retained by the cells and appear to have an enhanced inhibitory potential for certain enzymes [18]. The inhibitory capacity of methotrexate polyglutamates for dihydrofolate reductase appears to be somewhat greater than that of the parent drug. The selective retention of methotrexate polyglutamates may be critical for the delayed cytotoxicity exhibited by cells capable of polyglutamate synthesis. In vitro experiments using MCF-7 human breast cancer cells have demonstrated that the intracellular retention and duration of binding to dihydrofolate reductase are directly related to the length of the polyglutamate tail [29].

Interaction with Dihydrofolate Reductase. The binding of methotrexate to dihydrofolate reductase has been extensively investigated by X-ray crystallographic and amino acid sequencing studies [30–34]. Methotrexate binds in a stoichiometric fashion to a hydrophobic pocket in the target enzyme DHFR [35]. The binding affinity depends on multiple factors, including pH, salt concentration, and NADPH concentration, and has been reported to be ca. 10 pM [36]. In the cell, methotrexate is a reversible inhibitor capable of being displaced by high concentrations of substrate. Thus, free intracellular drug in excess of the cellular dihydrofolate reductase binding capacity is required to maintain complete inhibition of the enzyme and thereby produce and maintain a state of reduced-folate depletion. If an excess of intracellular drug is not maintained, the intracellular reduced-folate pool recovers through enzymatic reduction of oxidized folates and cellular metabolism resumes.

Cellular resistance to methotrexate has been most commonly associated with an increase in dihydrofolate reductase activity. In general, the amplified enzyme is identical to the native protein in its affinity for methotrexate; however, altered methotrexate affinity has been reported to correlate with sensitivity to methotrexate in a series of murine leukemias [36]. Increased reductase activity and resistance to methotrexate has also been demonstrated in a number of cell lines made resistant in vitro by stepwise increases in drug concentration [37], [38], and in human tumor samples from clinically resistant tumors [39]. The increased enzyme levels can be correlated to gene amplification that may take the form of small new pieces of chromosomal material, called double minutes, or of large chromosomes, referred to as homogeneously staining regions (HSRs). The former variety of amplification imparts relatively unstable resistance, which requires the ongoing selective pressure of drug presence to be maintained [40], whereas the HSRs represent a more durable form of amplification and thus resistance. Several investigators [41–43] have successfully transvected amplified reductase genes into normal hematopoietic cells, allowing greater marrow resistance to metho-

trexate, an important dose-limiting toxicity of the drug.

Determinants of Cytotoxicity. Methotrexate is an S-phase-specific agent whose cytotoxic effects are determined by drug concentration and duration of cell exposure. These effects may be altered by the cellular milieu. The toxic effects of methotrexate can be completely reversed by exogenous administration of the end products (purines and thymidine) whose de novo synthesis is inhibited by methotrexate treatment. Also, the synthesis of these products may resume if an exogenous source of reduced folates is provided. These data provide the rationale for the treatment of patients with high-dose methotrexate and subsequent administration of a reduced folate in the form of leucovorin calcium (N^5-formyl tetrahydrofolic acid) as "rescue." The reversal of methotrexate cytotoxicity by reduced folates is a competitive process. The reasons for the competitive nature of this relationship are unclear but may be the result of a shared membrane transport system. In addition, methotrexate may have direct inhibitory effects on enzymes other than dihydrofolate reductase that require competitive levels of the folate cosubstrate to overcome the inhibition.

2.1.2. Analogs

Many new analogs to methotrexate have been developed in an effort to circumvent the cellular resistance that occurs with prolonged methotrexate exposure. As mentioned, drugs with increased lipid solubility have been successful in treating transport-resistant cells in vitro [44]. The lipophilic derivative metoprine and variations of the 10-deazaaminopterin series are the most interesting additions. Of the 10-deaza series, an ethyl sub-stitution at the 10-position imparts a marked increase in cytotoxicity when compared to methotrexate [45]. The analog possesses an improved membrane transport ability while retaining a high affinity for dihydrofolate reductase. A new antineoplastic agent is piritrexim isothionate [*79483-69-5*], 6-[(2,5-dimethoxyphenyl)methyl]-5-methylpyrido[2,3-d]pyrimidine-2,4-diamine mono-2-hydroxyethanesulfonate, $C_{17}H_{19}N_5O_2$ [46], that inhibits dihydrofolate reductase (DHFR). Further lipophilic DHFR-inhibitors are trimetrexate [*82952-645*] [47], and edatrexate [48] which are clinically studied and have shown activity, e.g., in non-small cell lung cancer. Thymidylate synthase (TS) is the rate-limiting enzyme in the anabolism of thymidine resulting in the incorporation into DNA. Raltitrexed (company codes: ZN-1694, D-1694, ICI-D1694) [*112887-68-0*] [49], is currently under clinical investigation with response rates in colon and breast cancers of up to 30 % [50], [51]. Myelosuppression seems to be the predominant dose limiting toxicity.

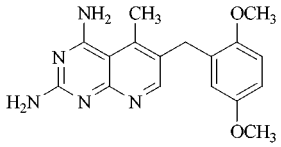

Piritrexim

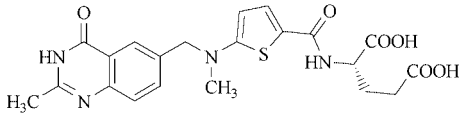

Raltitrexed

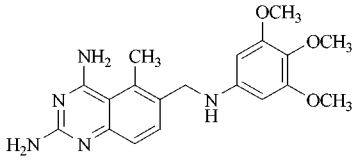

Trimetrexate

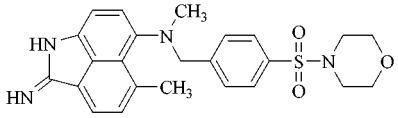

AG 331

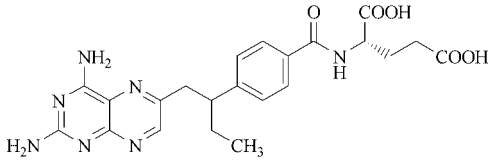

Edatrexate

Crystallographic data and computer-assisted drug design led to the development of thymidylate synthase (TS) inhibitors of the type of AG 331 [52].

Finally, CB 3717, a potent inhibitor of thymidylate synthase, is toxic for cell lines with altered dihydrofolate reductase; and homofolate, a de novo purine inhibitor requiring di-

hydrofolate reductase for activation, is effective in reductase-amplified lines [44].

2.2. Fluoropyrimidines

5-Fluorouracil (NSC-19893) [*51-21-8*], 5-fluoro-2,4-(1*H*, 3*H*)-pyrimidinedione, 5-FU, $C_4H_3FN_2O_2$, M_r 130.08, is a fluorinated pyrimidine whose structural formula resembles thymine; the hydrogen in the 5-position of the naturally occurring pyrimidine being replaced by fluorine. The *synthesis* of 5-FU in 1957 [15] represents the first successful effort in the rational design of anticancer drugs [53], and was predicated on the earlier observation that malignant cells selectively utilized uracil (and possibly toxic uracil analogs) in vivo [16], [54].

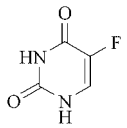

5-Fluorouracil

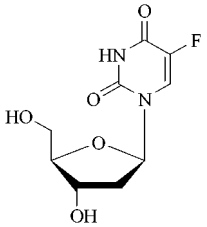

5-FUDR

Since the original synthesis of 5-FU and its nucleoside 5-FUDR (NSC-27640) [*50-91-9*], 2′-deoxy-5-fluorouridine, floxuridine, $C_9H_{11}FN_2O_5$, M_r 246.21, much has been learned about the mechanism of action of the fluoropyrimidines. These drugs are useful in the treatment of a wide range of human malignancies.

2.2.1. Mechanism of Action

Both 5-FU and 5-FUDR are prodrugs that require intracellular metabolism to their respective nucleotides for cytotoxicity. The pathways for fluoropyrimidine activation are shown in Figure 1. Each drug is enzymatically activated by different routes to FdUMP, FUMP, or FUTP, and each of these fluorinated nucleosides has different mechanisms of cytotoxicity.

Thymidine phosphorylase converts 5-FU to the deoxyribonucleotide 5-FUDR, which is then phosphorylated by thymidine kinase to yield 5-FdUMP. In the presence of methylene tetrahydrofolate, 5-FdUMP forms a stable ternary complex with thymidylate synthetase (TS), inhibiting this critical enzyme to cause "thymineless death." As expected, cytotoxicity is prevented in the presence of exogenous thymidine in these cells with intact salvage pathways. Inhibition of TS has long been considered the principal mechanism of 5-FU cytotoxicity, but it has also been demonstrated that 5-FU can be converted to 5-FUMP, either by orotic acid phosphoribosyl-transferase (OPRTase) in the presence of phosphoribosyl pyrophosphate (PRPP), or by stepwise conversion to the ribonucleotide 5-FUR (by uridine phosphorylase) followed by formation of 5-FUMP by uridine kinase. This intermediate can be converted to 5-FdUMP by ribonucleotide reductase to inhibit TS. Alternatively, 5-FUMP can be phosphorylated to 5-FUTP, which may be fraudulently incorporated into RNA to induce cytotoxicity. In a number of tumor models, loss of clonogenic capacity is directly correlated with the extent of incorporation of 5-FUTP into RNA [55], [56]. This RNA-specific toxicity is not reversed by thymidine. However, the precise mechanism of RNA-induced cell kill is speculative. The most consistent structural effect of 5-FU exposure is impaired processing of ribosomal RNA [57]. This concept is not supported by current evidence, however, as neither the synthesis nor the translation of messenger RNA (mRNA) appears affected by documented 5-FUTP incorporation into mRNA. Using human colon carcinoma cells propagated in vitro, neither quantitative nor qualitative differences in the translational products (polypeptides) of 5-FUTP-containing mRNA could be demonstrated [58]. However, small nuclear RNA species responsible for exon recognition during RNA splicing do contain significant quantities of uridylic acid [59]. Specific substitution of 5-FUTP into this RNA fraction may be critical for 5-FU toxicity.

As shown in Figure 1, 5-FdUMP can be further phosphorylated to 5-FdUTP, which can be incorporated into tumor cell DNA [60], [61]. This mechanism of drug action has been particularly difficult to appreciate, since the fraudulent base is quickly excised from DNA by the enzymes uracil-DNA glycosylase and dUTP nucleotidohydrolase. Thus, when tumor cells are incubated at low (0.1 μM) concentrations

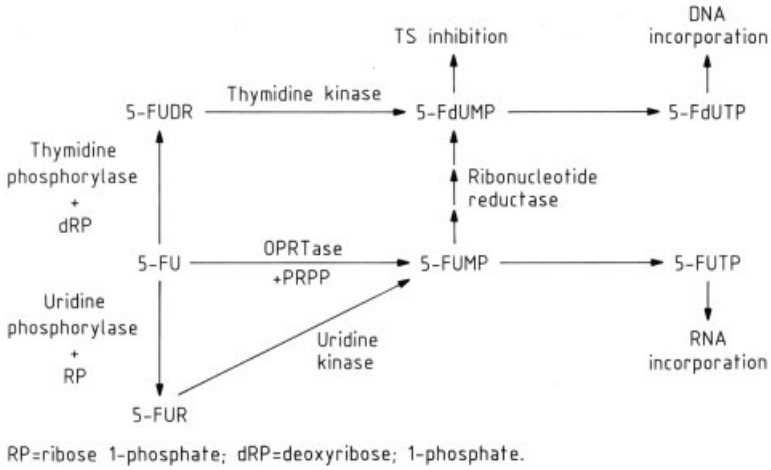

Figure 1. Pathways for fluoropyrimidine activation

of 5-FUDR, single-strand DNA shifts to lower molecular mass species, suggesting excision of the fluorinated base [62]. However, the actual presence of 5-FdUTP in cellular DNA cannot be detected until tumor cells are exposed to higher drug concentrations. Apparently, the importance of 5-FU incorporation into DNA varies from tumor to tumor, with human promyelocytic leukemia cells incorporating nearly 100-fold more 5-FdUTP into DNA than mouse leukemia cells [63]. The importance of this pathway both to tumor cell cytotoxicity and therapeutic index remains to be elucidated.

2.2.2. Mechanisms of Resistance

Tumor cells selected for in vitro resistance may demonstrate a deletion of critical drug-activating enzymes, including uridine kinase [64], orotic acid phosphoribosyltransferase [65], [66], and uridine phosphorylase [67]. Methylene tetrahydrofolate is required for 5-FdUMP inhibition of TS, and decreased availability of intracellular folates has been involved as a mechanism of 5-FU resistance [68].

In addition, alterations in the target enzyme TS can result in drug resistance. Resistant cells have been described with altered thymidylate synthetase having decreased affinity for FdUMP [69–71]. In addition, increased specific activity of TS has been reported in drug-resistant fibroblasts [72] and tumor cells [73], [74]. Whether elevated levels of target protein represent the end result of specific gene amplification (as has been documented for methotrexate resistance) remains to be determined.

2.2.3. Other Fluoropyrimidines

Attempts to develop fluoropyrimidines with improved therapeutic indexes have resulted in the synthesis and clinical trial of so-called *masked fluoropyrimidines.* The masked fluoropyrimidine 5′-deoxy-5-fluorouridine is a nontoxic prodrug that is converted to 5-FU by pyrimidine nucleoside phosphorylase. Because this enzyme may be present to a greater degree in some tumor cells than in normal human bone marrow, an improved therapeutic index can be demonstrated in vitro using breast, sarcoma, leukemia, and colon carcinoma cells [75]. Since 5′-deoxy-5-fluorouridine requires conversion to 5-FU and is in itself nontoxic, tumor cells that lack phosphorylase activity are resistant to the masked compound, while remaining cross-sensitive to 5-FU [76].

Ftorafur is a second masked fluoropyrimidine that is less myelosuppressive than 5-FU. However, the drug has a higher incidence of gastrointestinal and nervous system toxicity, which is probably due to organ-specific localization of activating enzymes [77]. In early clinical studies, ftorafur has shown antitumor activity in patients with 5-FU refractory colorectal and breast cancer [78].

2.3. 5-Azacytidine

5-Azacytidine (NSC-102816) [*320-67-2*], 4-amino-1-β-D-ribofuranosyl-1,3,5-triazine-2(1*H*)-one, 5-azacitidine, $C_8H_{12}N_4O_5$, M_r 244.21, is a pyrimidine analog that was first isolated as a fermentation product from *Streptoverticillium* cultures [79] and was chemically synthesized in Czechoslovakia in 1964 [80]. Structurally, 5-azacytidine differs from cytidine by the substitution of nitrogen in the 5 position of the pyrimidine nucleus.

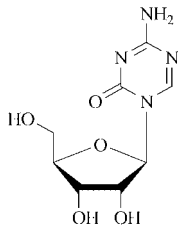

5-Azacytidine

2.3.1. Mechanism of Action

5-Azacytidine shares the facilitated transport system for cytidine for entry into cells [81] and must be phosphorylated to exert cytotoxic effects. Conversion to the monophosphate is catalyzed by the enzyme uridine – cytidine kinase [82], and this is likely the rate-limiting step for drug activation [83]. 5-Azacytidine monophosphate inhibits the enzyme orotidylate decarboxylase and interferes with de novo pyrimidine biosynthesis [84]. Subsequent metabolism of the monophosphate to azacytidine di- and triphosphate, catalyzed by cytidine monophosphate kinase and nucleoside diphosphate kinase, occurs rapidly and does not appear to be a rate-limiting step in drug activation [81].

The diphosphate of azacytidine is a substrate for ribonucleotide reductase and dAzaCTP (deoxyribonucleotide triphosphate of azacytidine) for DNA polymerase, allowing direct incorporation of drug into DNA [85]. This pathway may be critical for cytotoxicity because active DNA synthesis correlates with drug sensitivity in vitro [86]. In addition, incorporation of 5-azacytidine into DNA may affect gene expression. Mammalian DNA contains ca. 5 % of incorporated cytosine methylated in the 5 position [87]; methylation appears to inhibit gene expression. Thus the globin gene is hypomethylated in bone marrow as compared to other tissues [88–91]. The DNA containing even low levels of 5-azacytidine is a potent inhibitor of the enzyme responsible for cytosine methylation, i.e., DNA cytosine methyltransferase [92], [93]. The enzyme inhibition is disproportionately great compared with the small amount of incorporated fraudulent base, and appears to result from formation of a stable complex between 5-azacytidine residues and the methyltransferase, similar to the complex formed with thymidylate synthase and FdUMP [94]. Thus, in vitro treatment with 5-azacytidine can induce DNA hypomethylation and differentiation of murine cells [95], [96]. More recently these observations have been extended to clinical medicine. Azacytidine treatment in a patient with severe β-thalassemia could stimulate gamma-globin synthesis by inducing hypomethylation and expression of the gamma-globin gene [97].

After conversion to a triphosphate, azacytidine also competes with CTP for incorporation into RNA [98] and inhibits maturation of ribosomal and transfer RNA [99]. This causes disassembly of polyribosomes [100] and interferes with protein synthesis [101].

2.3.2. Mechanism of Resistance

Cellular resistance to 5-azacytidine may be the result of either decreased drug activation or possibly increased drug degradation. Drug metabolism by cytidine kinase appears to be rate-limiting in drug activation; deletion of this enzyme has been reported in 5-azacytidineresistant cells in vitro [81], [102]. Cytidine deaminase degrades 5-azacytidine to 5-azauridine; however, the role of this enzyme in drug resistance remains uncertain. For example, drug toxicity may be dependent on deamination; bacteria incapable of forming 5-azauridine from 5-azacytidine are resistant to the drug [103].

2.3.3. New Analogs

In clinical trials 5-azacytidine has demonstrated consistent antileukemic activity, inducing complete remissions in a significant number of

heavily pretreated patients with acute myelogenous leukemia [104]. Acute dose-limiting toxicities associated with bolus drug administration (severe gastrointestinal symptoms, fever, life-threatening hypotension) may be ameliorated by administering the drug via constant intravenous infusion [105]. However, 5-azacytidine is chemically unstable, undergoing ring opening at the 5,6-imino double bond to form *N*-formylamidinoribofuranosylguanylurea, which further decomposes to ribofuranosylurea [106]. The halflife of this decomposition is 4 h, complicating the task of strict dosage control of prolonged infusions.

To circumvent the problem of aqueous instability, the hydrolytically susceptible 5,6-imino bond of 5-azacytidine was reduced to produce *dihydro-5-azacytidine* (NSC-264880) [*62402-31-7*], 4-amino-5,6-dihydro-1-β-D-ribofuranosyl-1,3,5-triazine-2(1*H*)one · monohydrochloride, $C_8H_{14}N_4O_5 \cdot HCl$, M_r 282.7 [107]. This compound has completed phase I trials in the 1980s with preliminary evidence of activity in lymphoma; in addition it exhibits the unusual dose-limiting toxicity of chest pain at the maximally tolerated dose [108].

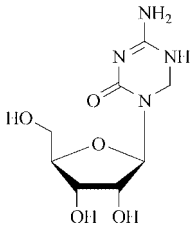

Dihydro-5-azacytidine

2.4. Cytosine Arabinoside (Ara-C)

Ara-C (NSC-63878) [*147-94-4*], 4-amino-1-β-D-arabinofuranosyl-2(1*H*)-pyrimidone, cytarabine, cytosine arabinoside, $C_9H_{13}N_3O_5$, M_r 243.22, an antimetabolite that is a structural analog of cytidine, differs from the physiologic nucleoside by the epimeric configuration of the β-*trans*-hydroxyl group at the 2′ position of the sugar. The drug is transported into cells by a carrier-mediated process with shared affinity for deoxycytidine [109]. Once ara-C has entered the cell, cytotoxicity is dependent on formation of the triphosphate ara-CTP, which is responsible for inhibition of DNA synthesis.

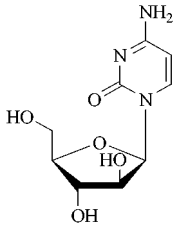

Cytosine arabinoside

The precise *mechanism* by which ara-CTP inhibits DNA synthesis remains uncertain. Ara-CTP does inhibit both DNA polymerase α [110] and β [111]. The former is essential for DNA synthesis, the latter for DNA repair. Ara-CTP can be directly incorporated into DNA as well, and this pathway correlates strongly with cytotoxicity. The extent of drug incorporation into DNA is proportional to cell kill in both acute myelocytic and promyelocytic leukemia cells [112]; incorporation and cytotoxicity can be modulated by thymidine [113]. Also ara-C-substituted DNA is unstable under conditions of alkaline elution, suggesting drug-induced strand fragility and breakage [114]. Ara-CTP incorporation also directly blocks strand elongation [115] and causes premature strand termination [116]. This results in accumulation of DNA peaks of small sizes, suggesting that preexisting DNA may be nicked following exposure to ara-C [117]. A further effect of ara-C exposure is inhibition of DNA repair, as determined by alkaline elution studies in L 1210 cells [118]. As might be expected, cells pretreated with ara-C are more sensitive to ionizing radiation, suggesting potentiation of radiation damage by inhibition of DNA repair [119], [120–123].

2.4.1. Mechanisms of Resistance

A number of mechanisms of resistance to ara-C have been described. Ara-C itself is a prodrug that is metabolized first to ara-CMP by deoxycytidine kinase, then to ara-CDP by pyrimidine nucleoside monophosphate kinase, and finally to ara-CTP by nucleoside diphosphate kinase. Resistant cell lines lack the initial, rate-limiting activating enzyme [115], [124].

Since the 1980s, HPLC has been used to separate and semipurify deoxycytidine kinase, pyrimidine nucleoside monophosphate kinase, and pyrimidine nucleoside diphosphate kinase from cell extracts [125]. In murine leukemia cells selected for ara-C resistance, deoxycytidine kinase was found in lower specific activity, and similar results have been reported for other systems [126].

In addition, expansion of the intracellular pool of the physiologic substrate deoxycytosine triphosphate can inhibit ara-C activation by feedback inhibition of the initial activating enzyme [127]. Moreover, increased dCTP pools may also compete directly with ara-CTP for DNA polymerase. Experiments have shown a relationship between duration of leukemia remission and the ability of tumor cells obtained from patients to form and retain ara-CTP in vitro [128], but the mechanisms underlying this relationship remain to be established.

Although some workers have reported that increased drug catabolism by cytidine deaminase may underlie both de novo and acquired drug resistance in patients with leukemia [129], a definite clinical correlation between response and levels of this enzyme has not been established in several studies [130–132].

2.4.2. New Analogs

Several ara-C analogs have been rationally designed with the goal of overcoming specific mechanisms of resistance. Compounds with lipophilic side chains are relatively resistant to inactivation by cytidine deaminase. Another analog with lipophilic modifications in the side chain is enocitabine (NSC-239336, BHAC) [*055726-47-1*] [133], [134]. N^4-Behenoyl-ara-C (BHAC) has undergone clinical trial in patients with acute leukemia [135]. Despite its lipophilicity, the drug does not enter the cerebrospinal fluid and concentrates in bone marrow and red blood cells. N^4-Palmitoyl-ara-C can be administered orally and appears to be more active than the parent compound in preclinical models [136]. Although developed as an antiviral agent, 2′-fluoro-5-iodo-1-β-D-arabinofuranosylcytosine (FIAC) has significant antitumor activity [137]. Interestingly, this compound may be relatively cytotoxic for cells with high levels of cytidine deaminase because the catabolism product, FIA-uracil, is more toxic than the parent compound [138].

2.5. Deoxycytidine and Analogs

The pyrimidine antimetabolite gemcitabine (LY-188011, dFdC) [*095058-81-4*], 2'-deoxy-2',2'-difluorocytidine, $C_9H_{11}F_2N_3O_4$ is an analog of deoxycytidine and a result of a program initiated at Lilly Research to synthesize fluorinated D-ribose and fluorinated nucleosides [139–142]. The difluorinated analog of cytarabine, gemcitabine was identified as novel antimetabolite with a broad spectrum of antitumor activity.

Trade names: Gemicitabine monohydrochloride [*122111-03-9*] is on the market as Gemzar in the indication of palliative treatment of locally advanced or metastatic non-small cell lung cancer.

Mechanism of Action. Gemcitabine shows good activity against human leukemic cell lines, a number of murine solid tumors, and human tumor xenografts [143–146]. Gemcitabine was significantly more cytotoxic than cytarabine in Chinese hamster ovary cells. The major cellular metabolite is the 5'-triphosphate of gemcitabine. The cytotoxicity was competitively reversed by deoxycytidine, suggesting that the biological activity required phosphorylation by deoxycytidine kinase [145].

Tumor-bearing mice were treated with either gemcitabine or cytarabine (20 mg/kg). DNA synthesis reached 1 % of control levels upon administration of gemcitabine. The greater accumulation of gemcitabine-5'-triphosphate compared with cytarabine-5'-triphosphate may cause greater cytotoxicity and therapeutic activity [146].

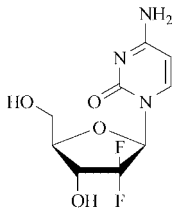

Gemcitabine

Further gemcitabine may enhance its own cytotoxic effects by self-potentiation mechanisms

that act on, e. g., deoxycytidine monophosphate deaminase, deoxycytidine kinase or on DNA-synthesis [147], [148].

2.6. 2-Halopurines and Analogs

Cladribine (2-CdA, RWJ-26251) *004291-63-8*, is a purine deoxyribonucleoside with remarkable antileukemic activity. It represents a significant advance over existing therapy because it is given as a single 7-day continuous treatment, thus minimizing the side effects observed with multiple treatments. Remission rates of up to 89 % lasting for up to 25 months were observed in clinical trials in patients with hairy cell leukemia [149–152]. The antimetabolite used for first line treatment of hairy cell leukemia [153].

Trade name: Leustatin.

Fludarabine phosphate (NSC-312887, 2-F-ara-AMP) [*075607-67-9*] [154], 2-fluoro-9-(5-*O*-phosphono-β-D-arabino-furanosyl)-9*H*-purin-6-amine, $C_{10}H_{13}FN_5O_7P$, M_r 365.21, another cytotoxic purine antimetabolite, acts via inhibition of DNA synthesis. The product is used for treatment of patients with chronic lymphocytic leukemia.

Trade names: Fludara (Berlex), Benefluor (Schering AG).

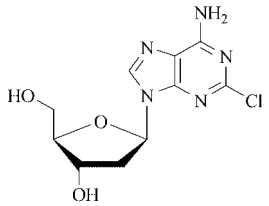

Cladribine

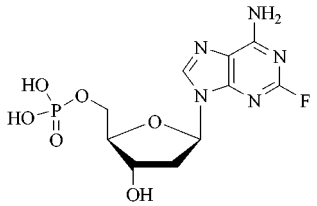

Fludarabine

Mechanism of Action. Fludarabine and its soluble derivatives interfere with phosphorylation, e.g., in L 1210 cells. Fludarabine behaves more like an analog of deoxycytidine than adenine or deoxyadenine as indicated by reports [155] demonstrating that the presence of fluorine in the 2-position of the adenine ring alters its function as a substrate for deaminase and nucleoside kinases. This results in differences in biological activity and metabolism. Halogenation does not simply block deamination, but also influences the enzyme that carries out the phosphorylation, as a result cytotoxicity is increased [156]. Fludarabine phosphate may selectively inhibit the incorporation of thymidine and uridine into the DNA molecule by inhibiting both ribonucleotide reductase [157] and DNA polymerase [158]. The maximum tolerated dose (MTD) in heavily pretreated patients with advanced malignancy/solid tumors on the daily regimen was about 15 mg/m^2. Granulocytopenia and thrombocytopenia were dose-limiting [159–161].

Pentostatin. The cytotoxic and immunosuppressant pentostatin (NSC-218321, CI-825, PD-81565, YK-176, 2-deoxycoformycin, 2'-dCF) [*063677-95-2*], (R)-3-(2-deoxy-β-D-erythro-pentofuranosyl)-3,6,7,8-tetrahydroimidazo[4,5-d][1,3]diazepin-8-ol, $C_{11}H_{16}H_4O_4$, M_r 268.13, *mp* 220 – 225 °C, can be isolated from the fermentation broth of *Streptomyces antibioticus* NRRL 3238 [162], [163] or *Aspergillus nidulanus* Y 176-2 or *Emericella* [164]. The adenosine nucleoside analog pentostatin, is the most potent inhibitor of adenosine deaminase, which is an important and ubiquitous cellular enzyme. The inhibition of this enzyme leads to accumulation of dATP which inhibits ribonucleotide reductase and thus DNA synthesis. Pentostatin was launched for treatment of hairy-cell leukemia refractory to α-interferon.

Trade names: Nipent (Parke-Davis, Lederle), Coforin (Katetsuken).

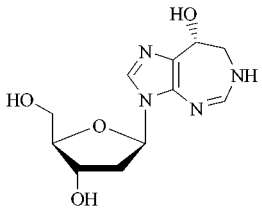

Pentostatin

Mechanism of Action. The highest activity of adenosine deaminase is found in lymphoid tissue as well as in the malignant cells of acute lymphoblastic leukemia [165]. Since pentostatin is the most potent inhibitor of adenosine deaminase it was expected to possess antitumor properties against certain malignancies, especially acute leukemias, chronic myelogenous leukemia, and lymphomas. Surprisingly, when tested against murine tumor models and also against various tumor cell cultures no antitumor effect was found [166]. However, in a phase I clinical trial the compound was found to produce a drop in lymphoblast count and showed antitumor activity in acute leukemia and lymphoma. The toxicity observed consisted mostly of CNS side effects including nausea, hepatic and renal dysfunction. In combination with vidarabin (Ara-A) (→ Chemotherapeutics, Chap. 4.6.1.) superiority to monotherapy was demonstrated in various studies [167], [168]. Durable complete remission was observed after pentostatin treatment in patients with hairycell leukemia resistant to α-interferon [169–174].

2.7. 6-Mercaptopurine and 6-Thioguanine

The purine antimetabolites, 6-mercaptopurine (NSC-755) [*50-44-2*], 1,7-dihydro-6*H*-purine-6-thione, 6-MP, $C_5H_4N_4S$, M_r 152.19; and 6-thioguanine (NSC-63878) [*154-42-7*], 2-amino-1,7-dihydro-6*H*-purine-6-thione, 6-TG, $C_5H_5N_5S$, M_r 167.21, have been used for several decades in the treatment of leukemia and certain other neoplastic diseases. The 6-thiopurine analog, 6-MP, was first synthesized in 1952 [175]; subsequently the same workers synthesized 6-TG in 1955 [176]. The clinical evaluation of 6-MP in the treatment of acute leukemia and chronic myelocytic leukemia was carefully studied in 1953 [177]. Shortly thereafter, the 2-amino analog of 6-MP, 6-TG, was submitted for clinical evaluation as an antitumor agent [178].

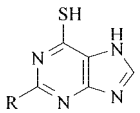

R = H, 6-Mercaptopurine
R = NH_2, Thioguanine

6-Mercaptopurine is used for maintenance therapy of acute lymphocytic and acute myelogenous leukemia, and 6-TG is used primarily for remission induction in acute myelogenous leukemia. Although 6-MP and 6-TG have an important role in chemotherapy of leukemia patients, the purine analogs produce low-response rates in patients with solid tumors, lymphomas, and chronic lymphocytic leukemia.

2.7.1. Mechanism of Action

6-Mercaptopurine has been used as an antineoplastic and immunosuppressive agent for decades, but the precise mechanism by which it exerts its cytotoxic effects has not yet been established. Both 6-MP and 6-TG must be converted to their thiol nucleotide form, which is the active cytotoxic moiety. The conversion is catalyzed by hypoxanthine guanine phosphoribosyltransferase (HGPRT) [179–181] and the reaction is dependent on the phosphoribosylpyrophosphate (PRPP) level in the cells [182–186]. 6-Thiouric acid (6-TU) is the major catabolic product of 6-MP. The rapid conversion of 6-MP to 6-TU by xanthine oxidase [187], [188] in leukemic cells may be a possible mechanism of 6-MP resistance [182], [189], [190]. The antineoplastic effect of 6-TG is similar to that of 6-MP. In its nucleotide form, 6-TG inhibits de novo purine biosynthesis and purine interconversions [191–195].

The cytotoxicity of 6-MP and 6-TG has been linked to (1) the interference with de novo purine biosynthesis and purine interconversions, (2) the inhibition of in vitro RNA synthesis, and (3) the incorporation into DNA during S phase, resulting in a deformation of the DNA.

The interference with de novo purine biosynthesis by 6-MP is regulated by 6-MP nucleotides. These nucleotides inhibit the enzyme 5-phosphoribosylpyrophosphate amidotransferase that catalyzes the initial reaction in the purine biosynthetic pathway [196]. The 6-MP nucleotides also inhibit conversion of inosine monophosphate (IMP) to adenine monophosphate (AMP) and to xanthine monophosphate (XMP), and limit the availability of XMP to form guanine monophosphate (GMP), thereby interfering with the sup-

ply of purine precursors for nucleic acid synthesis [181]. In the 1980s two new findings pertaining to the cytotoxicity mechanism in tumor cells have been reported. Studies of human lymphoma revealed that 6-MP was a potent inhibitor of cellular RNA synthesis and that 6-thioITP inhibited both the RNA polymerase I and RNA polymerase II activities of these cells [197]. These data suggested that direct inhibition of the enzymes mediating transcription by 6-thio-IMP may be one of the mechanisms for the cytotoxic action of 6-MP in human tumor cells. Using 6-TG as a cytotoxic agent resulted in severe chromosome damage in wild-type CHO cells [198]. Gross unilateral chromatid damage resulted, and the unilateral nature of this damage was probably due to malfunction of 6-TG-containing DNA as a replication template.

2.7.2. Mechanism of Resistance

Several mechanisms of resistance to these agents have been described in experimental tumors and relate to the pathways of antimetabolite activation and degradation. A decreased HGPRT activity in tumor cells diminishes antimetabolite activation, and this resistance pathway has been reported by several workers [199–202]. However, the HGPRT-regulated mechanism as a basis for drug resistance in human leukemic cells is relatively uncommon [203]. The resistance of 6-MP is related to an increase in alkaline phosphatase in sarcoma cells [204]. Increased alkaline phosphatase, a membrane-bound enzyme that converts the active mononucleotide to 6-thioinosine and inorganic phosphate, has also been reported in human leukemia patients resistant to drug treatment [205], [206]. Another enzyme in the degradation pathway of 6-MP that needs to be considered for drug efficacy and bioavailability is xanthine oxidase, which is responsible for converting 6-MP to 8-OH-6-MP and subsequently to thiouric acid, which is excreted through the urine.

To understand further the bioavailability and pharmacokinetics of thiopurines, the effect of allopurinol on 6-MP catabolism was studied. Allopurinol, an analog of hypoxanthine, enhances the therapeutic efficacy of 6-MP by inhibiting xanthine oxidase. The urinary excretion of 6-MP metabolites is markedly reduced in patients treated with allopurinol. Furthermore, allopurinol increased the plasma level of 6-MP in rabbits [207]. The data suggest that inhibition of 6-MP catabolism by allopurinol may contribute to a greater availability of 6-MP to tissues. Studies of the effect of allopurinol on the kinetics of oral and intravenous 6-MP in Rhesus monkeys and in humans demonstrated that allopurinol pretreatment resulted in a nearly 400 % increase in peak plasma concentration of 6-MP in monkeys and a 500 % increase in humans, but only when 6-MP was administered orally [208]. Allopurinol pretreatment had no effect on the kinetics of intravenously administered 6-MP. This difference is due to the action of allopurinol on liver or intestinal xanthine oxidase and inhibition of first-pass metabolism of oral 6-MP. This finding may explain the low and variable plasma levels of mercaptopurine in patients with acute lymphoblastic leukemia treated with oral 6-MP [209]. Although these studies emphasize the catabolic pathway for purines, all the purine and pyrimidine metabolic enzymes may be important to the bioavailability and activation of antimetabolites.

To clarify the mechanisms of resistance to thiopurines and potential drug interactions in tumor cells, studies have focused on the thiopurine-resistant cell lines deficient in HGPRT (L 1210) and the regulation of PRPP formation in thiopurine-resistant cell lines.

The thiopurines are inactive in the base form and must be converted to their respective nucleotides. This activation step requires PRPP as the cofactor and HGPRT as the enzyme to catalyze the conversion to nucleotide. A major biochemical effect of methotrexate is the suppression of purine biosynthesis and expansion of the PRPP pool. Studies of the cytotoxic and biochemical interaction of methotrexate and 6-TG in L 1210 mouse leukemia cells demonstrated that methotrexate can markedly enhance 6-TG activity [210]. Preexposure of cells to methotrexate resulted in a large increase in cytotoxic potency of 6-TG, whereas simultaneous exposure caused an antagonism of 6-TG cytotoxic activity. Although PRPP pools were not measured quantitatively, the effect of methotrexate preexposure was to increase PRPP pools, enhance the activation of 6-TG to 6-TG monophosphate, and thereby increase its incorporation into RNA.

2.7.3. New Analogs

Alkyl disulfide derivatives have been used as masked compounds of 6-mercaptopurines. The antitumor effect of these derivatives has been measured by their ability to decrease the degradation of 6-MP monophosphate. Of the seven 6-alkyl disulfide derivatives of 6-MP and 6-TG in L 1210 leukemia cells tested, decyl derivatives of both 6-MP and 6-TG were the most effective with therapeutic ratios as high as 50 and 48, while those of parent compounds were 6.2 and 5.0, respectively [211]. Partial circumvention of thiopurine resistance may have resulted from cellular uptake of intact acylated bis(6-MP-9-β-D-ribofuranoside-5′)pyrophosphate derivatives in 6-MP-resistant human cell lines deficient in HGPRT (L 1210) [212].

The 6-MP-resistant sublines of P 388 and L 1210 leukemia are also sensitive to two new *purine antagonists:* 5-carbamoyl-1*H*-imidazol-4-yl piperonylate and 4-carbamoylimidazolium-5-olate [213]. These two new purine analogs kill 6-MP-resistant cells by suppressive de novo purine synthesis. The activation of those new purine analogs is mediated by adenine phosphoribosyltransferase.

3. Alkylating Agents

Historically, alkylating agents were important in the early development of cancer chemotherapies. Victims of sulfur mustard gas exposure in World War I were found to have severe lymphoid aplasia as well as pulmonary irritation [214]. This led to clinical trials of the related, but less toxic, nitrogen mustard derivative, which produced tumor regressions in lymphoma patients [215].

Clinical use of nitrogen mustard today is mostly limited to the treatment of lymphomas, especially Hodgkin's disease, where it is used in a multidrug regimen called MOPP (nitrogen mustard, vincristine, procarbazine, and prednisone). Three widely used derivatives of nitrogen mustard used in patients with malignancies today include melphalan, cyclophosphamide, and chlorambucil (Sections 3.2, 3.3, 3.4).

The mechanism by which alkylating agents act can be classified as either S_N1 or S_N2. In the *S_N1 reaction,* a highly reactive intermediate forms initially and quickly reacts with a nucleophile to produce an alkylated product. This reaction follows first-order kinetics because the rate-limiting step is formation of the intermediate. The *S_N2 reaction* is a second-order reaction and thus is dependent on the concentration of both the alkylating agent and its target nucleophile [216]. In general, alkylating agents that react by an S_N1 mechanism, such as nitrogen mustard, are less selective in their reactions than S_N2 agents, but this rule is not always true. Selectivity also depends upon membrane permeability, charge, and reactivity of the drug.

3.1. Nitrogen Mustard

3.1.1. Mechanism of Action

Nitrogen mustard (NSC-762) [*51-75-2*], mechloroethamine, HN2, $CH_3N(CH_2CH_2Cl)_2$, $C_5H_{11}Cl_2N$, M_r 156.07, is activated through loss of one of its chlorines. The α carbon then reacts with the nucleophilic nitrogen to form the positively charged, highly reactive, cyclic aziridinium compound, which is attacked by nucleophiles to give the initial alkylated product. The second chlorine can also leave, initiating a second alkylation, which produces a cross-linked alkylation between two nucleophiles.

Because HN2 bonds covalently to many biologic molecules, such as DNA, RNA, and proteins, the alkylated sites responsible for its cytotoxicity are difficult to determine. However, studies have shown that cytotoxicity is likely to result from inhibition of DNA synthesis by damaging the DNA template [217–219]. The DNA molecule is rich in potential sites for alkylation, including the phosphate groups in the sugar phosphate backbone structure, and the oxygen and nitrogen sites in the purine and pyrimidine bases. However, the tendency for alkylation to occur in the N-7 position of guanine is enhanced. This may be mediated by the increased nucleophilic characteristics of the N-7 deoxyguanosine due to base stacking and charge transfer [220]. Other preferred sites of alkylation in decreasing order are the N-1 of adenosine, the N-3 of cytidine, and the N-3 of thymidine [221].

Bifunctional alkylating agents, such as HN2, produce intra- and interstrand cross-linking bet-

ween DNA in the double-helix structure preferentially at the N-7 guanosine site. Thus, these bifunctional compounds are more effective antitumor agents than their monofunctional analogs; however, increasing the number of alkylating sites on the agent beyond two does not appear to increase antitumor activity [222]. This evidence suggests that DNA cross-linking is critical for alkylator activity. Further evidence for the importance of cross – linking to cytotoxicity comes from alkaline elution studies that can detect low levels of cross-linking in cells exposed to minimal doses of bifunctional alkylating agents [223].

In contrast, monofunctional agents, such as procarbazine and dacarbazine, do not produce DNA interstrand cross-links and appear to exert toxicity by producing single-strand DNA breaks. The increased carcinogenicity seen with some monofunctional alkylating agents may be due to incorrect base pair substitution by DNA repair enzymes that could result in malignant transformation.

3.1.2. Mechanisms of Drug Resistance

Several mechanisms have been elucidated for nitrogen mustard and other bifunctional alkylating agents. First, resistant cells with defective drug transport have been described. Nitrogen mustard enters cells by an active transport system that is physiologically utilized for choline transport. Lymphoma cells resistant to nitrogen mustard demonstrate decreased drug uptake by this specific active transport site, which also decreases its uptake of choline [224]. Second, cytosolic increase in nonprotein sulfhydryl levels [225] and higher nonprotein-bound thiol compounds that could inactivate the drugs before they reach the nucleus have been found in nitrogen mustard-resistant cells [226]. Third, enhanced repair of DNA cross-linking by repair enzymes has been demonstrated in vitro [227] and in vivo [228].

3.2. Melphalan

Melphalan (NSC-8806) [*148-82-3*], 4-[bis(2-chloroethyl)amino]-L-phenylalanine, L-phenylalanine mustard, $C_{13}H_{18}Cl_2N_2O_2$, M_r 305.20, was rationally designed and synthesized as a phenylalanine derivative of nitrogen mustard with the aim of obtaining increased specificity against melanoma tumor cells that utilize phenylalanine or tyrosine to produce melanin.

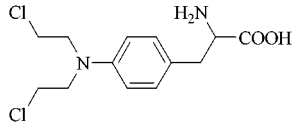

Melphalan

Although this compound does not exhibit specific antimelanoma activity, it has a broad-spectrum cytotoxicity in multiple myeloma, breast cancer, and lymphomas.

3.2.1. Mechanism of Action

Melphalan is a bifunctional alkylating agent but differs from HN2 mainly by the presence of an aromatic ring. This ring reduces the nucleophilicity of the nitrogen atom by withdrawing electrons, making the drug less reactive. Thus, it can be taken orally. It retains its alkylating activity but is more selective than nitrogen mustard because it is less likely to form the unstable and highly reactive aziridine intermediate indiscriminately. Like nitrogen mustard, it forms DNA cross-links that are critical for cytotoxic effect. Its cellular uptake is also mediated by an active, energy-dependent transport mechanism shared with leucine and glutamine uptake [218]. Thus, high concentrations of leucine or glutamine can reduce the cytotoxicity of melphalan in a marrow colony-forming unit assay [229]. Another transport mechanism, although less active, can be used by melphalan, and this mechanism is shared by the neutral amino acids alanine, cysteine, and serine [218].

3.2.2. Mechanism of Resistance

Decreased transport of melphalan into drug-resistant leukemia cells has been demonstrated and correlated to the melphalan-resistant phenotype. Specifically, a mutation in the higher velocity transport system has been suggested that results in a decreased affinity of the carrier protein for melphalan and leucine [230]. Evidence

for the other two mechanisms of resistance seen with nitrogen mustard, i.e., increased intracellular thiol compounds and increased DNA cross-linking repair also, exists.

3.3. Cyclophosphamide

Cyclophosphamide (NSC-26271) [*50-18-0*], *N,N*-bis(2-chloroethyl)tetrahydro-2*H*-1,3,2-oxazaphosphorin-2-amine 2-oxide, $C_7H_{15}Cl_2N_2O_2P$, M_r 261.10.

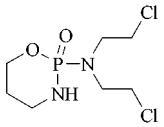

Cyclophosphamide

This widely used bifunctional, cyclic alkylating agent has important clinical use in lymphomas, leukemias, sarcomas, carcinomas of breast and ovary, as well as childhood malignancies. The compound was rationally designed based on data that tumor cells possess high concentrations of enzymes capable of cleaving the P−N bond. This reaction would activate drug by release of the potent antitumor agent phosphoramide mustard. Cyclophosphamide requires hepatic activation by oxidase enzymes. First, it is metabolized by liver microsomes to hydroxycyclophosphamide, which is spontaneously tautomerized to aldophosphamide. Aldophosphamide reaches peripheral tissues and tumors where it is hydrolyzed to yield the active antitumor agent phosphoramide mustard and acrolein. Acrolein has very weak antitumor activity and, when concentrated within the bladder by excretion, can cause hemorrhagic cystitis [216].

3.3.1. Mechanism of Action

Phosphoramide mustard can undergo similar bifunctional alkylation as nitrogen mustard. Also, because of its need for metabolic conversion for biologic activity, cyclophosphamide can be given orally or intravenously.

3.3.2. Mechanism of Resistance

Similar mechanisms of resistance occur for cyclophosphamide as for nitrogen mustard. Also, defective metabolic conversion by the hepatic microsomal system could serve to decrease the bioavailability of aldophosphamide to tumor tissue, but it is not known whether this is significant.

3.4. Chlorambucil

Chlorambucil (NSC-3088) [*305-03-3*], 4-[bis(2-chloroethyl)amino]benzenebutanoic acid, $C_{14}H_{19}Cl_2NO_2$, M_r 304.23. This drug is a close congener of melphalan and exhibits similar stability because of the electron-withdrawing properties of the aromatic ring. It is given orally and has proven efficacy in treating chronic lymphocytic leukemias, multiple myeloma, and lymphomas.

Chlorambucil is thought to have activation properties and mechanisms of resistance similar to melphalan.

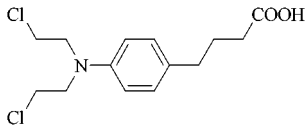

Chlorambucil

3.5. Thio-TEPA

Triethylenethiophosphoramide (NSC-6396) [*52-24-4*], 1,1′,1″-phosphinothioylidynetrisaziridine, thio-TEPA, $C_6H_{12}N_3PS$, M_r 189.23. This agent is representative of alkylating agents that have two or more aziridine rings. It has clinical activity against the same tumors as nitrogen mustard and has been used clinically in carcinomas of the breast and ovary. Thio-TEPA is also indicated for intrathecal therapy of meningeal carcinomas. The reactivity of the aziridine groups is increased by protonation; thus thio-TEPA is most active at low pH and has been used to cause sclerosis of malignant pleural effusions that often have an acidic pH.

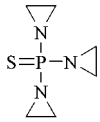

Thio-TEPA

The mechanisms of action and resistance are similar to those of nitrogen mustard.

3.6. Ifosfamide

Ifosfamide, or isofosphamide (NSC-109724) [*3778-73-2*], *N*,3-bis(2-chloroethyl)-tetrahydro-2*H*-1,3,2-oxazaphosphorin-2-amine 2-oxide, $C_7H_{15}Cl_2N_2O_2P$, M_r 261.07, is an analog of cyclophosphamide [231]. It has approximately one-third the alkylating activity of cyclophosphamide and requires hepatic microsomal conversion to its active form. Also, the rates of conversion by metabolism are similar in both drugs, but less biologically active alkylating moieties are formed in ifosfamide [232]. It has shown promising results in refractory pediatric bone and soft tissue sarcomas, refractory testicular tumors, and Wilms' tumor in children [233], [235].

Trade names: Ifex, Holoxan (ASTA Medica)

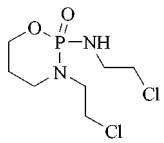

Ifosfamide

3.7. Estramustine

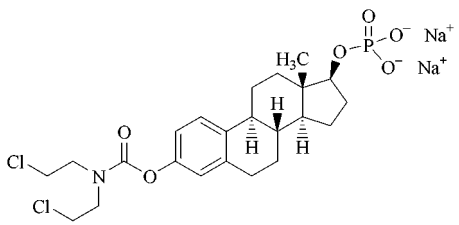

Estramustine

From the numerous analogs few found a way into the clinic or to the market. However, estramustine [*489-15-0*] was successfully launched for treatment of prostate cancer [234].

Trade name: Estracyt (Pharmacia/Upjohn)

3.8. Nitrosoureas

The clinically useful nitrosoureas include carmustine (NSC-409962) [*154-93-8*], *N,N′*-bis(2-chloroethyl)-*N*-nitrosourea, BCNU,$C_5H_9Cl_2N_3O_2$, M_r 214.04; lomustine (NSC-79037) [*13010-47-4*], *N*-(2-chloroethyl)-*N′*-cyclohexyl-*N*-nitrosourea, CCNU, $C_9H_{16}ClN_3O_2$, M_r 233.69; methylcyclohexylchloroethylnitrosourea (NSC-94941) [*52662-76-7*], 1-(2-chloroethyl)-3-(4-methylcyclohexyl)-1-nitrosourea, methyl-CCNU, $C_{10}H_{18}ClN_3O_2$, M_r 248; streptozotocin (NSC-85998) [*18883-66-4*], 2-deoxy-2{[(methylnitrosoamino)carbonyl]amino}-D-glucopyranose, $C_8H_{15}N_3O_7$, M_r 265.22; and chlorozotocin (NSC-178248) [*54749-90-5*], 2-({[(2-chloroethyl)nitrosoamino]carbonyl}amino)-2-deoxy-D-glucose, DCNU, $C_9H_{16}ClN_3O_7$, M_r 313.69. Streptozotocin is a naturally occurring nitrosourea derived from *Streptomyces acromogenes.*

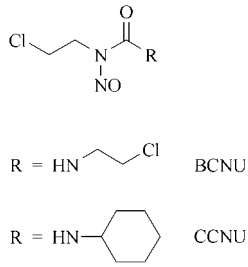

This group of agents was developed by careful structure – function studies based on the antitumor activity of methyl-CCNU [235–237]. The chloroethyl derivatives were found to possess increased activity and a capacity to cross the blood – brain barrier because of their lipophilic nature. Each of these agents is capable of undergoing alkylating reactions with biologic molecules in a manner similar to the classic mustards through the formation of highly reactive chloroethyl carbocations. With BCNU, each molecule of drug may undergo two such reactions to produce nucleic acid strand breaks, and DNA – DNA and DNA – protein cross-links

[238]. Monofunctional CCNU and methyl-CCNU cross-link by initial carbocation formation and alkylation followed by loss of the chloride substituent and formation of a second reactive carbocation. These alkylation reactions appear to be the major mode of cytotoxicity for these agents [239], [240]. With the exception of chlorozotocin, the nitrosoureas may also form an isocyanate compound that may play a role in the toxic side effects of these agents [241], [242], but has little importance in the antitumor effect. In support of this view is the fact that chlorozotocin retains its potent cytotoxic capacity while producing little or no isocyanate compound and reportedly has less marrow toxicity. Experimental evidence raises additional questions concerning the role of isocyanates, which may enhance the antitumor activity of these compounds. Methylnitrosourea, which cannot alkylate DNA to produce cross-linking, produces alterations of nuclear protein in a manner similar to BCNU, and isocyanates are considered a possible explanation for these effects from both agents [243]. Also, sensitivity of a Walker tumor line made resistant to bifunctional alkylating agents can be restored by simultaneous treatment with an isocyanate-producing agent [244].

3.8.1. Mechanism of Action

Because of their lipophilicity, nitrosoureas enter cells by passive diffusion as opposed to an active transport mechanism common to the classic alkylating agents [245]. Once inside the cell, alkylating agent exposure results in pancellular covalent binding of drug to proteins, nucleic acids, and to a variety of smaller intracellular molecules. Which of these reactions is critical for cytotoxicity remains uncertain, but the majority of evidence points to interaction directly with DNA as the focal point of cytotoxicity [246–248]. The 7 position of guanine is particularly susceptible to alkylation, and accounts for the majority of the total alkylation of DNA [249], [250]. Since the chloroethyl nitrosoureas are each capable of two independent alkylations [238], DNA can be cross-linked by either interstrand or intrastrand processes [251]. Multiple studies using a variety of alkylating agents have confirmed that DNA cross-linking, which leads to inactivation of the DNA template, may well be the key mechanism of cytotoxicity [252–254]. Monofunctional DNA alkylations also occur following nitrosourea exposure and these must also be considered cytotoxic as monofunctional alkylating agents, incapable of cross-linking, retain cytotoxic activity. Monofunctional alkylations may produce single-strand DNA breaks by endonuclease cleavage at apurinic sites produced by the alkylation and repair process [255].

3.8.2. Mechanisms of Resistance

The mechanisms of cellular resistance to nitrosoureas remain unclear, although defective drug transport, as has been demonstrated for the classic alkylating agents, would be unlikely given the lack of need for an active membrane transport system. In human glioblastoma cells, increased activity of a specific excision enzyme, guanine-O^6-alkyltransferase, is correlated with in vitro resistance to nitrosoureas [256], and repair mechanisms would seem a likely mechanism since mammalian cells are capable generally of such repair [257].

3.8.3. Analogs

New agents have been synthesized that contain a sugar moiety similar to streptozotocin. Analogs that contain mannose, glucose, ribose, maltose, and galactose have all been produced, with the maltosyl derivative being exceptionally active in a variety of tumors tested [258], [259]. Sugar alcohols, such as mannitol, linked to nitrosoureas retain their cytotoxicity but appear to protect against the marrow toxicity induced by the parent nitrosourea compounds. A number of derivatives with di- and tripeptides containing alkyl nitrosoureas have also been introduced but have met with only moderate preclinical success [260].

There is continuous effort to obtain new nitrosoureas with higher efficiency and/or lower toxicity and several new compounds are undergoing clinical trials. In 1987 ranimustine (NSC-270561, MCNU) [*058994-96-0*], was marketed as Cymerin (Tokyo Tanabe) [261].

The phosphonoalanine derivative fotemustine (S-10036) [*092118-27-9*],

(±)-diethyl[1-[3-(2-chloroethyl)-3-nitroso-ureido]ethyl]phosphonate, $C_9H_{19}ClN_3O_5P$, M_r 315.69, *mp* 85 °C was approved as Muphoran (Servier) for treatment of disseminated malignant melanoma [262], [263]. Side effects of the related compound BCNU (carmustine) have been linked to the inhibitory effect on a major enzyme of the glutathione pathway, the cytosolic glutathione reductase. Fotemustine has no inhibitory effect on cytosolic glutathione reductase, indicating that fotemustine has a lower toxicity than carmustine [264–266]. In addition it has been shown to have lower mutagenicity in the Ames and micronucleus tests compared to BCNU [267]. In the clinical studies the major toxic effects of fotemustine were thrombocytopenia and leukopenia, which were delayed and reversible, nausea and vomiting being mild [268].

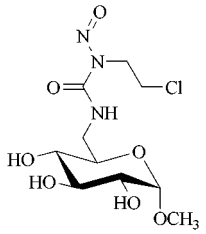

Ranimustine

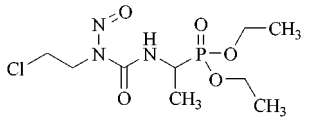

Fotemustine

3.9. Procarbazine

Procarbazine (NSC-77213) [*671-16-9*], *N*-(1-methylethyl)-4-[(2-methylhydrazino)methyl]-benzamide, $C_{12}H_{19}N_3O$, M_r 221.30, is one of a number of substituted hydrazine compounds originally synthesized as monoamine oxidase inhibitors in the early 1960s [269], and found to possess antineoplastic activity, particularly in the treatment of Hodgkin's disease [270]. Procarbazine is nontoxic as the parent compound but undergoes rapid chemical and metabolic degradation to intermediate products that, through a variety of mechanisms, are capable of cytotoxicity [271].

A crucial step in the activation of procarbazine appears to be the production of the azo analog *N*-isopropyl-α-(2-methyldiazeno)-*p*-toluamide, which is catalyzed by hepatic microsomal cytochrome P450 [272], [273]. Further metabolism by hepatic microsomes leads to the production of methyl- and benzyl-azoxy metabolites [272], [273]. Cytotoxic alkylating compounds may be formed from these metabolites via hydroxylation reactions, and these may play the major role in cytotoxicity. This sequence of metabolism and formation of the alkylating intermediates is consistent with the time course of appearance of the active species isolated in the serum and then excreted [274], [275]. In addition to the formation of these alkylating intermediates, methyl- and benzyl-azoxy metabolites may result in free radicals through the formation of diazenes [276], which in the presence of oxygen form free radicals and nitrogen. However, free-radical formation is not likely a major cause of cytotoxicity, because drug activity is preserved when cells are exposed to procarbazine under conditions that do not support the formation of free radicals [277]. The degree of toxicity induced by alkylation and by free-radical formation is unknown.

3.9.1. Mechanism of Action

The exact mechanism by which procarbazine produces cytotoxicity is unknown; however, the effects of its action have been well studied at the cellular level. Chromosome breaks and translocations have been demonstrated in vivo in Ehrlich ascites and L 1210 leukemia cells [278]. Inhibition of nucleic acid and protein synthesis and of a variety of enzymes has also been documented following procarbazine treatment. Inhibition of transfer and nuclear RNA synthesis occurs 2 h after procarbazine exposure and lasts for up to 24 h [279]. Thymidine incorporation into DNA is inhibited concomitantly with protein synthesis, reaching maximum inhibition in 12 – 16 h [271], [280]. Although many potentially cytotoxic events have been associated with procarbazine administration, it is not clear which of these effects causes cell death.

3.9.2. Mechanisms of Resistance

Resistance mechanisms for procarbazine are poorly understood, and no detailed studies have been reported that illustrate the typical cellular resistance encountered with alkylating agents. Since the drug enters cells by simple diffusion, resistance is unlikely to involve altered transport mechanisms [281]. New pieces of chromosomal material were found in Ehrlich ascites cells made resistant to procarbazine [271], and these may represent gene amplification, perhaps encoding for a target enzyme or detoxifying enzyme.

3.10. Dacarbazine

Dacarbazine (NSC-45388) [*4342-03-4*], 5-(3,3-dimethyl-1-triazenyl)-1*H*-imidazole-4-carboxamide, DTIC, $C_6H_{10}N_6O$, M_r 182.18, was synthesized in the late 1950s as an analog of 5-aminoimidazole-4-carboxamide, an intermediate in de novo purine synthesis. The drug is a product of rational synthesis, designed as a false intermediate capable of inhibiting de novo purine synthesis. Despite this theoretical basis for antitumor action, DTIC does not function as a purine analog; instead, it is extensively metabolized to a methylating agent [282]. Similar to procarbazine, DTIC must undergo activation by a microsomal oxidase to form a compound that can spontaneously produce a methyl diazonium ion intermediate that is probably the active metabolite. However, more recent evidence suggests a further metabolism to *N*-hydroxymethyl diazonium ion may be responsible for the selective antitumor effect of the drug [283]. Evidence for methylation of nucleic acids has been demonstrated in tissue culture [284]. An additional metabolic pathway involves the light sensitivity of the drug. Exposure to ultraviolet energy converts the parent compound to metabolites with moderate cytotoxicity in vitro [285], but this probably does not represent an important pathway for cytotoxicity in vivo.

Mechanism of Action. The mechanism of action of DTIC has not been systematically investigated, but it appears that the drug may act during any phase of the cell cycle [286] and can produce inhibition of RNA, DNA, and protein synthesis.

3.11. Hexamethylmelamine

Hexamethylmelamine (NSC-13875) [*645-05-6*], *N,N,N′,N′,N″,N″*-hexamethyl-1,3,5-triazine-2,4,6-triamine, altretamine, $C_9H_{18}N_6$, M_r 210.27, was synthesized in 1951 [287]. It has consistent antitumor activity in a variety of solid tumors, including ovarian, lung, and breast cancer.

The compound is almost insoluble in water, and thus, must be administered orally. Neither the mechanism of action nor the products of metabolic breakdown have been firmly established. Following administration of hexamethylmelamine, a spectrum of *N*-demethylation species has been isolated in the urine. The triazine ring appears unaffected by metabolism, as is evident by almost complete recovery of the intact ring in the urine using ring-labeled compound [288]. Existing evidence suggests possible formation of an alkylating species through *N*-demethylation [289] or the formation of *N*-methylol intermediates by hydroxylation of the parent compound [290] to account for the cytotoxic effects of the drug. *N*-Methylol derivatives have cytotoxic effects in vitro, but whether or not these derivatives are formed in vivo is unclear.

Analogs. Pentamethylmelamine (NSC-118742) [*35832-09-8*], *N,N,N′,N′,N″*-pentamethyl-1,3,5-triazine-2,4,6-triamine, $C_8H_{17}N_6$, M_r 196, is the most commonly used analog of hexamethylmelamine and differs by the absence of a single methyl group. Its major advantage is aqueous solubility, allowing an intravenous formulation. Its metabolism, toxicity, and antitumor activity parallel those of hexamethylmelamine [291], [292]. Other metabolites of hexamethylmelamine with varying numbers of methyl groups also possess antitumor activity that, in general, is directly proportional to the number of methyl groups on the triazene ring [292], [293].

3.12. Mitomycin-C

The mitomycins are a family of antibiotics isolated from *Streptomyces caespitosus*.

Mitomycin-C (NSC-26980, MIT-C) [*50-07-7*] [1aR-(1aα,8β,8aα,8bα)]-6-amino-8-[(aminocarbonyl)oxymethyl]-1,1a,2,8,8a,8b-hexahydro-8a-methoxy-5-methylazirino-[2',3':3,4]pyrrolo[1,2-a]indole-4,7-dione, $C_{15}H_{18}N_4O_5$, has DNA-alkylating properties [294]. Mitomycin is applied for treatment of stomach, breast, and gynecological cancers [295], [296]. Dose-limiting are leucopenia and thrombopenia.

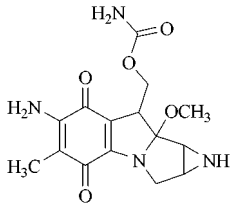

Mitomycin

4. Anthracyclines (→ Antibiotics, Chap. 3.2.3.)

Doxorubicin (NSC-123127) [*23214-92-8*], 10-[(3-amino-2,3,6-trideoxy-α-L-lyxo-hexopyranosyl)oxy]-7,8,9,10-tetrahydro-6,8,11-trihydroxy-8-(hydroxyacetyl)-1-methoxy-5,12-naphthacenedione, adriamycin, $C_{27}H_{29}NO_{11}$, M_r 543.54; and daunorubicin (NSC-82151 for HCl salt) [*20830-81-3*], 8-acetyl-10-[(3-amino-2,3,6-trideoxy-α-L-lyxo-hexopyranosyl)oxy]-7,8,9,10-tetrahydro-6,8,11-trihydroxy-1-methoxy-5,12-naphthacenedione, daunomycin, $C_{27}H_{29}NO_{10}$, M_r 527.51. Doxorubicin is used in the treatment of breast cancer, sarcoma, lymphoma, and small-cell lung cancer. Daunorubicin is used more commonly in acute myelocytic and lymphocytic leukemias.

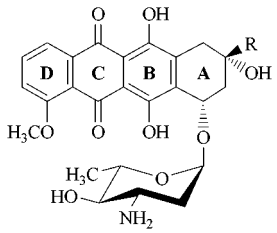

R = $COCH_2OH$, Doxorubicin
R = $COCH_3$, Daunorubicin

Meanwhile epirubicin (4'-epiadriamycin, pidorubicin, IMI-28) [*56420-45-2*], (8S-*cis*)-10-[(3-amino-2,3,6-trideoxy-β-L-arabino-hexopyranosyl)oxy]-7,8,9,10-tetrahydro-6,8,11-trihydroxy-8-(hydroxyacetyl)-1-methoxy-5,12-naphthacenedione, is the most commonly used antineoplastic antibiotic for breast cancer treatment. The total turnover exceeded \$ 200 × 10^6 in 1996. The compound can be synthesized by classical chemical synthesis [297–299].

Trade names: Farmorubicin and Pharmorubicin (Farmitalia).

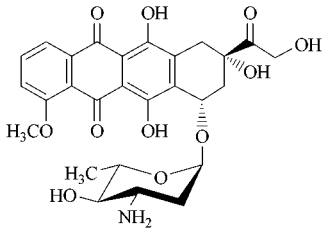

Epirubicin

In addition idarubicin (NSC-256439, IMI-30, DMDR) [*058957-92-99*] and pirarubicin (THP-ADM) [*072496-41-4*], (8S-*cis*)-10-[[3-amino-2,3,6-trideoxy-4-*O*-(tetrahydro-2H-pyran-2-yl]-α-L-lyxo-hexopyranosyl]oxy]-7,8,9,10-tetrahydro-6,8,11-trihydroxy-8-(hydroxyacetyl)-1-methoxy-5,12-naphthacenedione, $C_{32}H_{37}NO_{12}$, M_r 627,64 are available as idamycin (Adria) and pinorubicin (Nippon Kayaku), respectively.

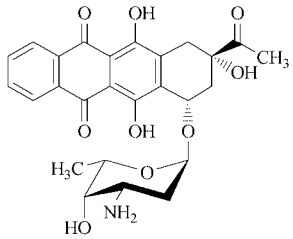

Idarubicin

Pirarubicin is the 4'-*O*-tetrahydropyranyl analog of adriamycin and can be synthesized from adriamycin [300], [301]. The acute cardiac toxicity was significantly less than that of adri-

amycin and general toxicity lower than that of other analogs [302, 303, 303]. The main indications for pirarubicin are cancer of the bladder, head and neck, and cervix.

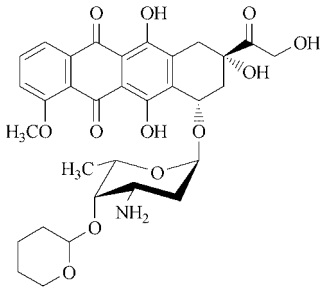

Pirarubicin

The anthracyclines are derived from *Streptomyces* species and are structurally tetracyclic chromophore antibiotics [305]. They are classified by their chromophore, otherwise known as aglycone, structure. The sugar most commonly attached to the aglycones is daunosamine, but other sugars may be involved, and these are mentioned in Section 4.3. The basic tetracyclic aglycone structure of the anthracyclines shares many characteristics with the hydroxyanthraquinones, which are ubiquitous in nature.

4.1. Mechanism of Action

DNA–RNA Binding. The exact mechanism of cytotoxicity by anthracyclines is unknown, but they do have multiple and distinct toxic effects that may kill a tumor cell in one or more ways [306].

Initially these drugs were found to bind DNA by intercalation between base pairs perpendicular to the long axis of the double helix, with the major binding occurring between the B and C rings of the drugs with the bases above and below [307]. The daunosamine sugar is thought to bind ionically with the sugar-phosphate backbone of DNA. The binding association constant is between 10^5 and 10^6 M^{-1}. Intercalation of the DNA causes a partial unwinding of the helix and thus disrupts DNA polymerases and transcription. However, these experiments were done with DNA in vitro; DNA in vivo is organized into chromatin, which is DNA wrapped around a series of histone core particles. Also, it has been established that the anthracyclines can affect every DNA function, including initiation, chain elongation, DNA synthesis, DNA repair, and RNA synthesis [308], [309]. Experimentally these compounds can cause sister chromatid exchanges, single- and double-strand breaks, and alkylation of DNA [310], [311]. Interestingly, some results strongly suggest that inhibition of DNA synthesis is not essential for cell kill. New anthracycline analogs, such as aclacinomycin A, which selectively inhibits preribosomal RNA synthesis and not DNA synthesis, retain cytotoxicity. This suggests other non-DNA-mediated mechanisms of cytotoxicity, such as preribosomal RNA synthesis [309].

Free-Radical Generation. Free-radical formation (highly reactive compounds with an unpaired electron) occurs during the metabolism of anthracyclines. When the microsomal enzyme P 450 reductase or xanthine oxidase interacts with and reduces the ketone oxygen in ring B to O^-, a semiquinone radical intermediate is formed. This interacts with oxygen to produce the superoxide radical with regeneration of the original anthracycline structure [312]. The superoxide radical can serve as a substrate for superoxide dismutase to form hydrogen peroxide, which can interact with the superoxide molecules to form hydroxyl radicals [313]. The superoxide and hydroxyl radicals can interact with and damage cells, especially the hydroxyl radical, which is one of the most reactive substances known. Hydroxyl radicals can also react with purine or pyrimidine bases, amines, and thiols. This free-radical formation is responsible for cardiac damage seen with chronic use of the anthracyclines in the treatment of human malignancies. Evidence exists in many animal models, as well as in humans, that superoxide and hydroxyl radical formation occurs in cardiac tissue, leading to lipid peroxidation of mitochondria and sarcosomes [314], [315]. Since mitochondria account for more than 40 % of cardiac muscle mass, as well as being the major source for ATP needed for contraction, and are coupled to calcium release during the action potential, one can easily visualize how these agents mediate cardiac toxicity. However, it now seems apparent that cardiac tissue lacks catalase and that doxorubicin destroys glutathione peroxidase [315], [316]. Agents that can scavenge free-radicals are

under active investigation. Interestingly, there is no evidence to date to link anthracycline free-radical formation to its antitumor activity so that analogs incapable of free-radical formation may demonstrate improved therapeutic index.

Membrane Interactions. Anthracycline binding to cell membranes appears to be an important mechanism for cytotoxicity. Changes in membrane glycoproteins, transmembrane flux of ions, and membrane morphology have been demonstrated in a variety of cells. Doxorubicin binds most tightly to cardiolipin, a phospholipid found in high concentration in mitochondrial and tumor cell membranes but in low concentration in normal cell membranes [317]. Also, membrane redox potential changes occur with drug binding, and may promote free-radical generation. Perhaps the most interesting and important finding to date on this subject is that doxorubicin, covalently linked to beads to prevent cell entry of the drug, retains cytotoxic effects [318]. This suggests that doxorubicin does not need to enter cells or interact with DNA to mediate cell kill.

Another potential mechanism of action may be the dissociation of the cytochrome oxidase electron transport chain for ATP generation [319]. Cytochrome c oxidase requires cardiolipin for activity, and doxorubicin can remove cardiolipin from the enzyme complex, thus inactivating the enzyme.

Metal Ion Chelation. The anthracyclines are capable of chelating ions, including copper, calcium, magnesium, zinc, and iron. Of the metal ions, the tightest complex seems to be with iron(III). Doxorubicin – iron chelates can act as a redox catalyst for electron transfer from glutathione to oxygen, which leads to formation of cytotoxic oxygen radicals. This reaction can also utilize hydrogen peroxide and superoxide, leading to hydroxyl radical formation. Thus, evidence exists that the anthracycline – iron complex can mediate free-radical formation capable of lipid peroxidation and cell damage [320]. Whether this is a mechanism of tumor cell kill is unknown, but the phenomenon may be important to cardiac toxicity.

4.2. Mechanism of Resistance

Probably the most common and important mechanism by which tumor cells become resistant to the anthracyclines is decreased net intracellular accumulation. Many studies have shown that doxorubicin-resistant tumor cells are capable of effluxing the drug more efficiently than their parent sensitive cells [355]. In fact, this mechanism of drug resistance may be responsible for resistance to a variety of structurally unrelated compounds with different modes of antitumor action, such as the vinca alkaloids and actinomycin D. This has been termed pleiotropic drug resistance, and several laboratories have shown the reversal of this resistant phenotype by co-incubation with various calcium channel blockers with doxorubicin [356]. These compounds increase the net intracellular drug accumulation in resistant cells, but the precise mechanism of this action is uncertain.

4.3. Analogs

As mentioned in Section 4.1, anthracycline tumor toxicity is probably related to intercalation with DNA, preribosomal RNA inhibition, and/or membrane binding effects, while cardiac toxicity may be due to free-radical formation by drug and drug – metal complex. Thus, analogs have been developed with less potential for free-radical formation. One such drug in early clinical development is aclacinomycin A (NSC-208734), which has an aklavinone aglycone structure and is derived as a fermentation product of *Streptomyces* [321]. This drug lacks the 11-hydroxyl, and therefore has about 10 % the potential to generate free-radicals in the P 450 reductase system. Although it does not bind DNA as well as doxorubicin, it retains significant antitumor activity.

The National Cancer Institute (USA) has screened hundreds of anthracycline analogs for antitumor activity. From these studies it can be concluded that (1) the amino sugar is not required for activity, (2) disaccharide analogs are generally more active than the parent saccharide in inhibiting RNA synthesis, and (3) in the aglycone, substituents in the 7 and 9 positions are important for activity. Studies have also shown

that if the 4′-hydroxyl group is removed, cardiac toxicity is lessened significantly [322–324]. Thus, it seems that the amino sugar and/or the 4′-hydroxyl group are major determinants of cardiac toxicity in the anthracyclines.

5. Intercalating Anthracenes and Analogs

5.1. Mitoxantrone

Mitoxantrone hydrochloride (NSC-301739, DAD, CL-232315) [*070476-82-3*], 1,4-dihydroxy-5,8-bis[(2-(2-hydroxyethyl)amino)ethyl]amino-9,10-anthracenedione, is a new type of antineoplastic agent. Mitoxantrone is active against breast cancer, acute leukemia, lymphoma, cervix carcinoma, and liver cell cancer. Unlike the anthracyclines that have a red color, the anthracenediones are deep blue. Mitoxantrone is structurally similar to adriamycin but without the aminosugar at C9 and can be synthesized starting from 1,8-dihydroxyanthrachinone [325–329].

Trade names: Novanthrone (Lederle / American Cyanamide); Onkotrone (ASTA Medica AWD).

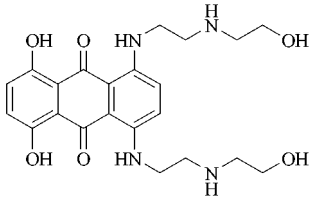

Mitoxantrone

Mechanism of Action. The quinone structure of mitoxantrone was recognized as being similar to that of adriamycin, having a lower cardiotoxic potential. However, its discovery was both a result of serendipity and of rational drug development [325]. The exact mechanism of action by which mitoxantrone exerts its cytotoxic effects has not been fully defined. The cytotoxicity is most likely associated with the action on chromosomal elements, resulting in DNA damage and leading to inhibition of nucleic acid synthesis and the eventual death of the cell. When compared to doxorubicin on an equimolar basis mitoxantrone proved to be six to seven times more potent in inhibiting the incorporation of uridine and thymidine into DNA [330]. Mitoxantrone is a cell phase nonspecific agent and has wide spectrum of activity against several experimental animal tumors. Cross-resistance to adriamycin was not always seen [331–333]. Adriamycin-like cardiac toxicity was not found in comparative studies using rats, dogs, rabbits, and monkeys maybe partially due to inhibition of free radical formation and due to the lack of the amino sugar moiety [334], [335]. Although mitoxantrone is not entirely free of cardiac toxicity in humans, minimal nausea and vomiting was observed [336–341].

5.2. Analogs

Bisantrene hydrochloride (NSC-337766, ADD, CL-216942) [*071439-68-4*], 9,10-anthracenedicarboxaldehyde bis[(4,5-dihydro-1H-imidazol-2-yl] dihydrochloride, $C_{22}H_{22}N_8$’2 HCl, M_r 471.39, is an intercalating anthracenebishydrazine cytostatic [342–344]. The product was launched for treatment of acute non-lymphocytic leukemia [345].

Trade names: Zantrene and Cyabin (Lederle).

As solubility is a problem, bisantrene prodrugs with enhanced water solubility (e.g., 199344, see below) were developed [346–348].

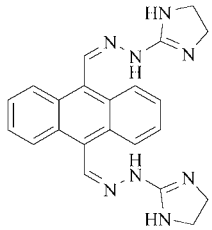

Bisantrene

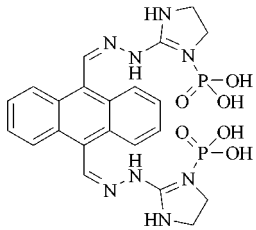

199344

Further intercalating agents such as amsacrine (NSC-156303/NSC-249992, m-AMSA, SN-11841) [*051264-14-3*] [349], [350], or nitracin [*4533-39-5*] [351] have the acridine structure in common.

Trade names: Amsacrine is marketed as Amsidine, Amecrin, and Ansidyl (Parke-Davis) [352–354].

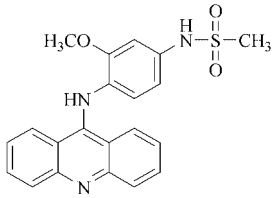

Amsacrine

6. Antitumor Antibiotics Other than Anthracyclines

The antitumor antibiotics are a group of microbial products capable of inhibiting tumor growth. This class of antitumor agents has been extensively studied as to mechanism of action and has a rather broad spectrum of activity. In contrast to antibacterial antibiotics, the therapeutic index of these drugs tends to be narrow and toxicity to normal host tissues is considerable.

6.1. Actinomycin D (→ Antibiotics, Chap. 3.7.6.)

The actinomycins are a family of antibiotics derived from *Actinomyces* broths during the 1940s [357]. Structurally, compounds in this class share a common phenoxone ring and two cyclic pentapeptides. The natural products differ among themselves in the amino acid composition of the peptide chains, but only actinomycin D is used in the clinical treatment of cancer, where it demonstrates reproducible activity against gestational choriocarcinoma and Wilms' tumor.

Actinomycin D (NSC-3053) [*50-76-0*], dactinomycin, $C_{62}H_{86}N_{12}O_{16}$, M_r 1255.47.

Mechanism of Action. Actinomycin interacts with DNA through "pseudo-intercalation" at deoxyguanylyl-3′,5′-deoxycytidine sequences [358], [359], resulting in inhibition of DNA-directed RNA synthesis and inhibition of protein synthesis [360]. Resistance to actinomycin D is linked to cross-resistance against other structurally unrelated amphiphilic drugs with dissimilar mechanisms of action as part of the phenomenon of pleiotropic drug resistance (see Section 4.2). Certainly it had been previously demonstrated that the accumulation of drug is greater in sensitive cells [361], [362], suggesting that alterations in transport at the membrane level might be important to resistance.

6.2. Bleomycin (→ Antibiotics, Chap. 3.7.1.)

Bleomycins (NSC-125066) are a group of antitumor antibiotics initially isolated from broths of *Streptomyces verticillus* [363].

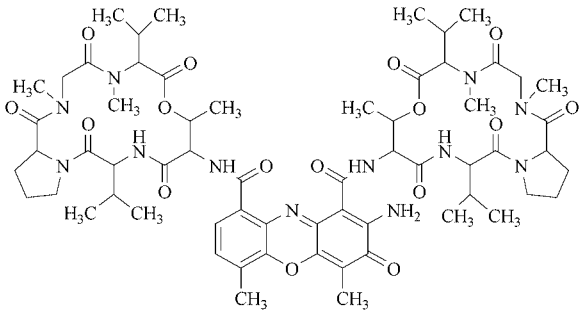

Actinomycin D

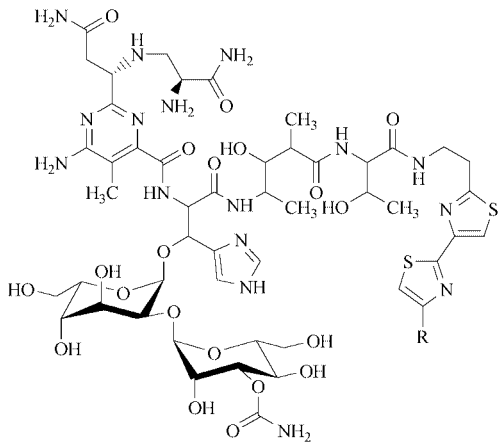

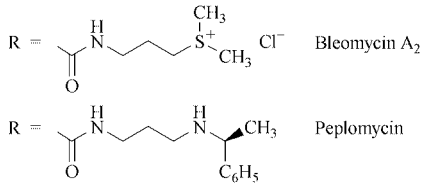

The fermentation product consists of approximately 12 different components clinically marketed as bleomycin. Each is a peptide with a low relative molecular mass (ca. 1500), all containing bleomycinic acid but differing in terminal alkylamine groups. The clinical product is approximately 60–70 % bleomycin A_2 (NSC-146842), N^1-[3-(dimethylsulfonio)propyl]bleomycinamide, $C_{55}H_{84}N_{17}O_{21}S_3$, M_r 1416; and 20–30 % bleomycin B_2. Other analogs comprise about 5 % of the total. The drug is highly active against germ cell neoplasm of the testis.

6.2.1. Analogs

Peplomycin [*68247-85-8*], N^1-{3-[(1-phenylethyl)amino]propyl}bleomycinamide, $C_{61}H_{88}N_{18}O_{21}S_2$, M_r 1473.62, a biosynthetic bleomycin analog, demonstrates significantly reduced pulmonary toxicity in rodents [364]. The toxicology studies are supported by clinical trials in Japan (initiated in 1981) and France (1983), and suggest that peplomycin has a greater therapeutic index than the parent compound and may replace bleomycin in the clinic [365]. In addition, total synthesis of bleomycin described in the 1980s will likely lead to other analogs and a further understanding of the drug's mechanism of action [367]. A new drug delivery system which comprises peplomycin absorbed on to small activated carbon particles was introduced as slow release formulation suggesting to decrease the systemic toxicity [366].

6.2.2. Mechanism of Action

Bleomycin enters cells slowly. Labeling studies demonstrate that the drug is first detected on the cell membrane and reaches the nucleus only after several hours [368]. Bleomycin kills cells by producing single- and double-strand DNA breaks. Bleomycin first binds guanine bases in DNA through the amino terminal peptide of the drug [369]. Free-radical formation occurs through oxidation of a bleomycin–Fe(II) complex to Fe(III), which catalyzes the reduction of molecular oxygen to superoxides and free hydroxyl radicals [370]. In vitro resistance to the drug appears to be mediated either by defective accumulation [371] or by increased intracellular drug degradation by a specific bleomycin hydrolase [372–374].

The clinical toxicity of bleomycin is unusual. The drug has little hematopoietic toxicity; its major dose-limiting toxicities are to the lungs and skin. Acute and chronic pneumonitis followed by progressive pulmonary fibrosis appears to be dose dependent, with risk increasing significantly in patients receiving a cumulative dose of more than 450 mg. This toxic endpoint may be due to the terminal amines of the parent compound. This hypothesis is supported by preclinical studies [375] as well as by early clinical studies of peplomycin. The maximum-tolerated cumulative dose of this terminal amino acid-substituted bleomycin analog has yet to be determined, although during phase I analysis, pulmonary toxicity was not observed until weekly doses exceeded 10 mg/m^2 [376].

6.3. DNA Interactive Natural Products

Analogs related to the natural product CC-1065 like adozelesin, bizelesin, and carzelesin bind in the minor grove of the DNA and form a covalent adduct with adenine [377]. Adozelesin

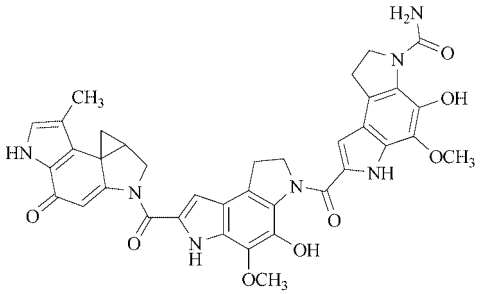

CC-1065

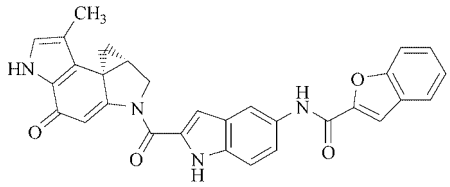

Adozelesin

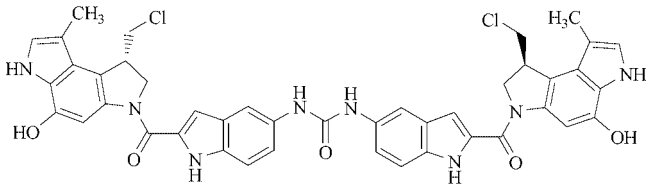

Bizelesin

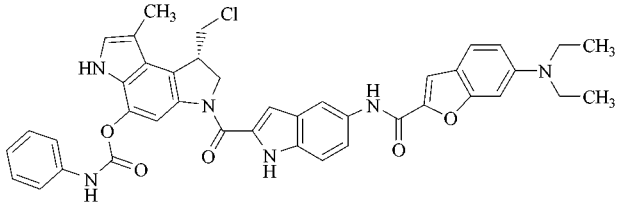

Carzelesin

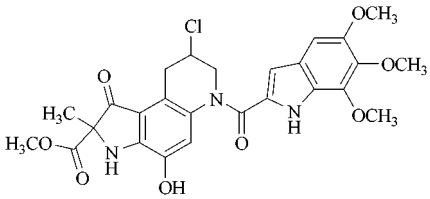

Duocarmycin C1

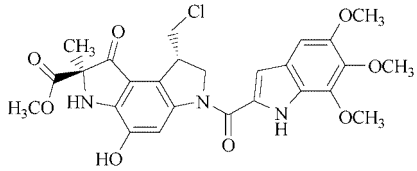

Pyrindamycin A

[*110314-48-2*] [378] is the clinically farthest advanced agent, while the prodrug carzelesin [*119813-10-4*] [379] and the dimer bizelesin [*129655-21-6*] [380] demonstrated impressive preclinical activity.

Further modified cyclopropylpyrroloindoles (CPI) forming covalent adducts with DNA are duocarmycin A [*118292-35-6*] [381], [382], pyrindamycin [*118292-36-7*] and FCE 24517 [383], [384].

A series of new DNA cleaving molecules based on the reactive enediyne moiety are the anticancer antibiotics calicheamicin [*113440-58-7*] [385], esperamicin [*114797-28-3*] [386], and dynemicin.

Simplified enediyne-type compounds damage DNA upon activation by chemical or biological means and are extremely potent cytotoxic agents in vitro [387], [388].

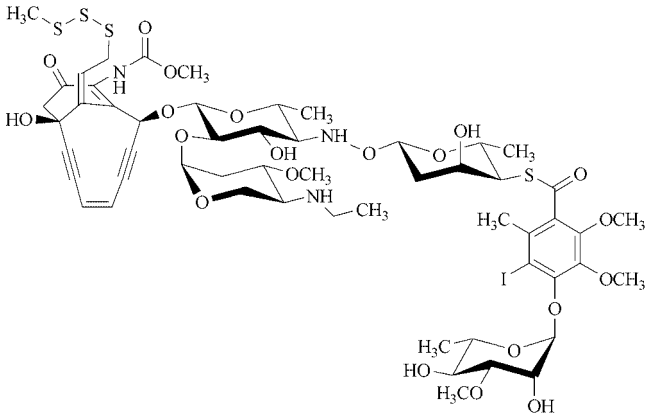

Esparamicin A1

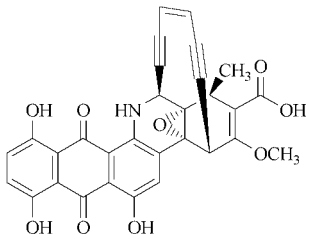

Dynemicin A

7. Antitubulin Agents

Tubulin-containing structures are important for diverse cellular functions, including chromosome segregation during cell division, intracellular transport, development and maintenance of cell shape, cell motility, and possibly distribution of molecules on cell membranes. The drugs that interact with tubulin are heterogeneous in structure. A common characteristic of these agents is binding to tubulin, causing its precipitation and sequestration to interrupt many important biologic functions that depend on the microtubular class of subcellular organelles. The tubulin – drug aggregates can be visualized by the indirect immunofluorescence technique as brightly stained cytoplasmic paracrystals. Of the tubulin binders, those that are important in cancer medicine include vinca alkaloids, podophyllotoxins, and taxoids.

7.1. Vinca Alkaloids (→ Alkaloids, Chap. 11.3.4.)

Vinca alkaloids are dimeric indole derivatives isolated from the periwinkle plant, *Catharantus roseus.* Of the whole family of more than 70 naturally occurring alkaloids, only a few have cytotoxic activity.

7.1.1. Vincristine and Vinblastine

The molecular structures of the dimeric *Catharantus roseus* alkaloids vincristine (NSC-67574) [*57-22-7*], 22-oxovincaleukoblastine, $C_{46}H_{56}N_4O_{10}$, M_r 824.94, Oncovin (sulfate); and vinblastine (NSC-49842) [*865-21-4*], vincaleukoblastine, $C_{46}H_{58}N_4O_9$, M_r 811.00, Velban (sulfate hydrate), are very similar. Both are formed of multiringed units, vindoline and catharantine, linked by a carbon – carbon bridge. They differ only in the nature of the substituent on the vindoline nitrogen atom.

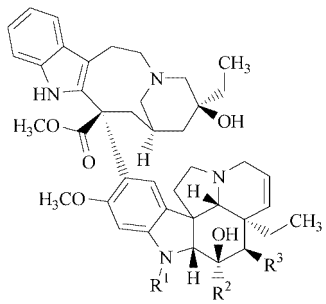

R^1 = CHO, R^2 = CO(OCH$_3$), R^3 = CO(CH$_3$), Vinristine
R^1 = CH$_3$, R^2 = CO(OCH$_3$), R^3 = CO(CH$_3$), Vinblastine
R^1 = CH$_3$, R^2 = CO(NH$_2$), R^3 = OH, Vindesine

Cellular Pharmacology. As yet it is not clear how vinca alkaloids cross cell membranes. Some data suggest an energy-dependent transport system [389], while other data suggest simple diffusion across membranes [390]. However, passive diffusion is important only at drug concentrations exceeding 100 μM. Transport of vincristine is completely inhibited by vinblastine, suggesting a common carrier.

Vincristine and vinblastine exert their biologic effect through binding to tubulin. This occurs in interphase (late S and G_2), producing a transient G_2 block [391]. They do not affect DNA synthesis directly. The metabolic consequences of tubulin binding include polyploidy, nuclear fragmentation, and inhibition of cytokinesis, which occurs after prolonged drug exposure. Present studies suggest that sensitivity to these drugs increases progressively as cells approach mitosis and that cells at the end of the cycle at the time of exposure are likely to exhibit the greatest degree of mitotic disorganization [391].

Vincristine and vinblastine share a common binding site on each tubulin monomer, with the binding affinity (K_d) of about 1.6×10^{-6} M. The drug concentrations necessary to produce 50 % cell kill in vitro are ca. 4×10^{-8} M [392]. Malignant lymphocytes appear more susceptible than normal lymphocytes, presumably because of high tubulin content expressed on the surfaces of leukemic cells [393]. Other metabolic effects include inhibition of DNA and RNA and protein synthesis. However, these effects are exerted only at very high drug concentrations (1000 times greater than those achieved in vivo) and probably are secondary phenomena.

Drug Interactions. Both vincristine and vinblastine potentiate the effect of methotrexate through their blockade of methotrexate exit from cells [394], [395]. Some amino acids (glutamine, aspartic acid, ornithine, citrulline, and arginine) completely reverse the cytotoxic effect of vinblastine in tissue culture [396].

Mechanisms of Resistance. Vinca resistance may arise through mutations in tubulin, leading to decreased drug binding. Vinca-resistant cells may also share cross-resistance with antitumor antibiotics through a separate mechanism of resistance. The precise mechanism of this pleiotropic drug resistance remains to be clarified. However, tubulin is not responsible for that phenomenon because there is very little difference in affinity binding of colchicine to tubulin isolated from drug-sensitive and drug-resistant cells [397]. The appearance of a novel glycoprotein on the membrane of resistant cells and the accelerated drug efflux – leading to impaired drug accumulation – seem, at present, to be the possible mechanisms responsible for resistance [398].

7.1.2. Vindesine

Vindesine (NSC-245467) [*53643-48-4*], 3-(aminocarbonyl)-O^4-deacetyl-3-de(methoxycarbonyl)vincaleukoblastine, deacetylvinblastineamide, $C_{43}H_{55}N_5O_7$, M_r 753.95 (structure, Section 7.1.1), is a vinblastine metabolite. It possesses antitumor activity that is similar to vincristine's rather than that of its parent compound, vinblastine. Vindesine demonstrates better activity than vinblastine in some tumor models (Gardner lymphosarcoma, Ridgway osteogenic sarcoma, and mammary carcinoma). In the past several years vindesine has attracted considerable attention among clinical investigators and has been intensively investigated. Response rates vary, with the highest responses observed in the highest lymphatic malignancies [399]. The drug is less neurotoxic than vincristine and appears to be active in vincristine-resistant tumors [400], [401].

7.1.3. Vinorelbine

Vinorelbine (KW-2307, NVB) [*071486-22-1*] 3',4'-didehydro-4'-deoxy-C'-norvincaleukoblastine, $C_{45}H_{54}N_4O_8$, M_r 778.94, is a semisynthetic vinca alkaloid. Starting from anhydrovinblastine the compound is obtained as 5'-noranhydrovinblastine in three steps [402], [403]. The product was launched for the treatment of non-small cell lung cancer.

Trade names: Navelbine (Pierre-Fabre/Glaxo), Ennades (Farmitalia).

Mechanism of Action. Vinorelbine was selected for drug development due to its high affinity for tubulin and its ability to prevent tubulin

polymerization [403–405]. The compound induced total depolymerization of microtubules in P 388 murine leukemia cells, possibly via stimulation of microtubular protein synthesis [403]. Vinorelbine and vincristine were equally active against L 1210 leukemia in mice, while vinblastine had no significant effect. Vinorelbine exerted significant antitumor activity in the vincristine-resistant cell line P 377/VCR and low cross-resistance to other vinca alkaloids was observed [406], [407].

In clinical studies leukopenia was dose-limiting, no thrombocytopenia was observed [408]. Efficacy in non-small cell lung cancer was demonstrated [409–411].

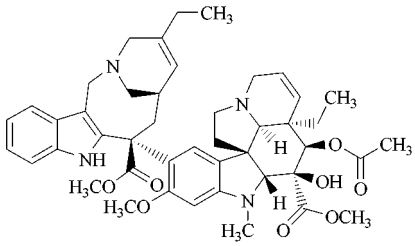

Vinorelbine

7.2. Podophyllotoxin and Its Derivatives

Podophyllin is a complex mixture of crystalline compounds derived from the mayapple plant. The active agent derived from this plant product, podophyllotoxin [*518-28-5*], shares a common binding to tubulin with colchicine, and the morphological effects of podophyllotoxin and colchicine exposure are indistinguishable [412]. The semisynthetic glycoside derivatives of podophyllotoxin, VP-16-213 (NSC-141540) [*33419-42-0*], etoposide, and VM-26 (NSC-122819) [*29767-20-2*], teniposide, have reproducible clinical activity against testicular cancer, small-cell lung cancer, and lymphomas [413].

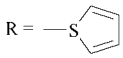

R = CH_3

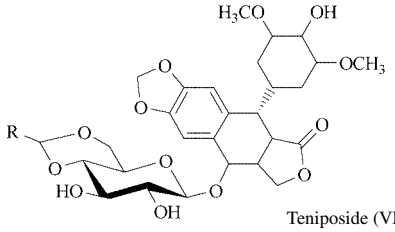

Teniposide (VM-26)

Despite structural similarities between the semisynthetic derivatives and parent natural product, neither VP-16 nor VM-26 binds tubulin. Evidence from a number of laboratories suggests that VP-16 exerts cytotoxic effects by causing DNA breakage [414], [415]. Selective double-strand DNA breaks, which are particularly lethal, appear to be caused by these agents [416].

The precise mechanism for VP-16 and VM-26 DNA damage is unknown, but the fact that the presence of a 4′-hydroxyl group is critical for cytotoxicity suggests formation of free-radical intermediates. Indeed, free-radical scavengers appear to be able to protect from podophyllin derivative cytotoxic effects in vitro [417]. Formation of DNA breaks may also be due to specific inhibition of the DNA repair enzyme topoisomerase II by both VP-16 and VM-26 [418]. It is likely that further characterization of topoisomerase II and the effects of its inhibition will result in the development of new leads in antitumor drug development.

7.3. Camptothecin and Analogs

Camptothecin (→ Alkaloids, Chap. 11.3.5.) (CPT) [*7689-03-4*], an alkaloid isolated from the Chinese plant *Camptotheca accuminata* in 1966 [419] was proved to have antineoplastic activity in various tumor systems [420]. The clinical use of CPT was, however, limited because of its high toxicity and low solubility in water. New derivatives have been synthesized carrying substitutents in 7,9,10- or 11-position of the ring A.

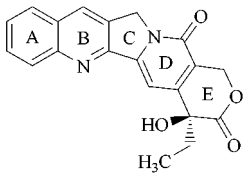

Campothecin

Irinotecan (NSC-616348, CPT-11, DQ-2805) [*097682-44-5*] [421], (+)-7-ethyl-10-[4-(1-piperidyl)-1-piperidyl]carbonyloxycamptothecin, $C_{33}H_{38}N_4O_6$, M_r 586.69 and topotecan (NSC-609699, SKF-S-104864-A, E-89/001) [*119413-54-4*] [422], (S)-10-dimethylaminomethyl-9-hydroxycamptothecin, $C_{23}H_{23}N_3O_5$, M_r 457.91 are camptothecin derivatives with topoisomerase I-inhibitory activity. As antineoplastic alkaloids, they can be synthesized semisynthetically starting from CPT [423]. In addition both agents have intercalating properties and are water soluble. Irinotecan hydrochloride was launched for the treatment of small cell and non-small cell lung cancer and cancers of the uterine, cervix, and ovaries.

Trade names: Campto (Rhône-Poulenc Rorer/Yakult Honsha) Topotecin (Daiichi Seiyaku).

Unlike irinotecan, topotecan is not a prodrug and does not require metabolic activation.

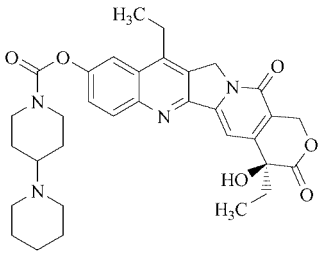

Irinotecan

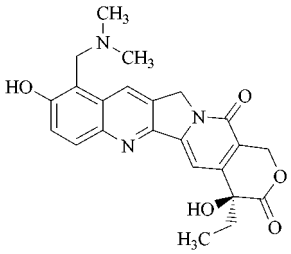

Topotecan

Mechanism of Action. Topoisomerase (topo I) is an ubiquitous nuclear enzyme involved in the regulation of essential cellular functions by relieving torsional DNA strain. Relaxation of supercoiled DNA is achieved through a series of topo I-mediated reactions. The main physiological role of topo-I is in sensing and subsequently releasing the positive supercoiling generated ahead of the moving transcription apparatus. A number of malignancies including acute leukemia, blasts, colon, esophageal, and ovarian cancers contain increased topo-I levels as compared to normal tissues [424], [425]. Topotecan binding to topo-I-DNA adducts results in markedly decreased rates of nick resealing and in delayed enzyme release, leading to increased numbers of strand breaks [426].

In animal tumor models including xenografts of human cancer lines topotecan and irinotecan exhibit a wide spectrum of antineoplastic activity [427–429]. Cell lines overexpressing P-glycoprotein display low levels of resistance to topotecan through decreased drug retention, more pronounced resistance is found in cell lines containing low levels of topo I [430], [431].

Dose-limiting toxicity in patients were stomatitis and esophagitis as well as neutropenia [432].

7.4. Taxoids

Paclitaxel (NSC-125973, BMS-181339) [*33069-62-4*], $C_{47}H_{51}NO_{14}$, M_r 853.9, is a natural antineoplastic taxane derivative originally isolated from the plant *Taxus brevifolia* [433] and subsequently synthesized due to limited supply from the bark of endangered yew trees. Starting from 10-deacetyl baccatin III (isolated from the renewable twigs and needles of *Taxus baccata*) the compound can be prepared in three steps utilizing protected *N*-benzoyl-(2R,3S)-3-phenylisoserine as key building block [434–436]. Various synthetic approaches are described and patented, including three total synthesis of paclitaxel [437–440].

Paclitaxel has a novel mechanism of action. Unlike vincristine or vinblastine, which bind tubulin to inhibit tubulin polymerization, paclitaxel stabilizes microtubular structures [441]. Indicated for the treatment of primary ovarian cancer in combination with cisplatin and for

metastatic ovarian cancer where standard therapy has failed, paclitaxel has been marketed since 1993 as Taxol (Bristol-Myers Squibb) and was the best-selling anticancer agent in 1996. Taxol does not share a common binding site with other antitubulin agents and may instead bind to and stabilize polymerized tubulin [442].

A semisynthetic analog of paclitaxel, docetaxel (NSC-628503, RP-56976) [*114977-28-5*], (2R,3S)-*N*-carboxy-3-phenylisoserine, *N-tert*-butyl ester, 13-ester with 5β, 20-epoxy-1,2α-4,7β,10β,13α-hexahydroxytax-11-en-9-one 4-acetate 2-benzoate, $C_{43}H_{53}NO_{14}$, M_r 807.9, is a derivative with an *N-tert*-butyloxycarbonyl-(2R,3S)phenylisoserine C-13 side chain. Docetaxel is more potent than paclitaxel in vitro [443], [444]. Impressive clinical results have been reported for the treatment of ovarian, breast, and bronchial cancers with docetaxel [445]. The compound was launched in 1996 for the treatment of locally advanced breast cancer or relapse during anthracycline therapy of NSCLC and breast cancer.

Trade name: Taxotere (Rhône-Poulenc Rorer).

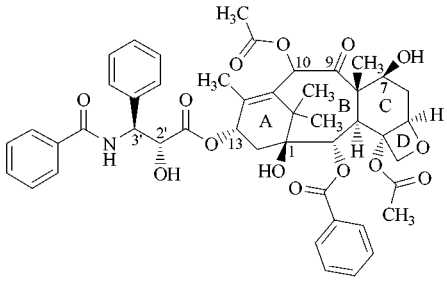

Paclitaxel

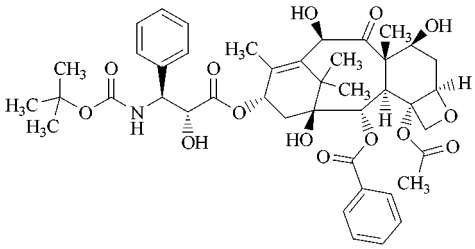

Docetaxel

Mechanism of Action. More than two decades after its isolation and the elucidation of its complex structure and cytotoxic activity [433] interest in paclitaxel raised again when S. Horwitz et al. [446] reported on the novel mechanism of action. As spindle poison paclitaxel promotes the polymerization of tubulin to microtubules and stabilizes them against depolymerization, whereas vinca alkaloids induce microtubule disassembly (see Fig. 2). Thus, with paclitaxel the dynamic equilibrium of assembly and disassembly of microtubules is shifted in favor of the polymer, preventing cell division [447]. Paclitaxel binds preferentially to the β-tubulin subunit. This binding is reversible and the site is different from the binding sites of vinca alkaloids, colchicine and podophyllotoxin [448–450]. If paclitaxel is present, tubulin polymerizes without exogenous GTP and these stabilized, rigid microtubules cannot be disassembled. As a result, the dynamic organization of the cell is interrupted which leads to irreversible damage in rapidly dividing cells. Further evidence has been reported that the antiproliferative activity of paclitaxel is caused by additional effects [451–453].

A major impediment in the development of taxol as a drug was its poor water solubility [454]. With the solubility enhancer Cremophor EL hypersensitivity reactions, including hypotension, urticaria, and dyspnea, occurred in patients during rapid infusion. To cope these allergic reactions, 24-hour infusions and pretreatment with dexamethasone, diphenhydramine, or cimetidine was recommended [455].

The dose-limiting *toxicity* of paclitaxel is neutropenia. Several other toxic effects such as diarrhea, nausea, and emesis are less common. Docetaxel shares many toxic effects with paclitaxel such as dose-limiting neutropenia, alopenia, myalgias, and mucositis. In addition fluid retention and cutanous toxicities are observed [445], [460], [461]. Because of the low water solubility of paclitaxel and docetaxel, the synthesis of more soluble taxoid prodrugs and smaller analogs has become an interesting area of research (see Section 7.5).

Structure – Activity Relationship. A large number of taxoid analogs has been synthesized with emphasis to enhance biological activity and to improve the water solubility [436]. C-13

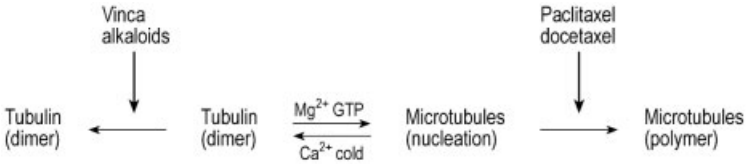

Figure 2. Mechanism of action of paclitaxel and vinca alkaloids

side chain depleted analogs such as baccatin III and its derivatives [456] as well as *N*-benzoyl-(2R,3S)-phenylisoserine are inactive. Simplified side chains at C-13 (like acetic, crotonic, or phenylacetic acid) possess reduced activity.

3'-Cyclohexyl-3'-dephenylpaclitaxel has a similar cytotoxicity to paclitaxel. Further modifications in the aromatic 3'-phenyl group gave compounds that were equipotent with paclitaxel. The compound with a 3'-(*p*-methoxyphenyl) has a slightly increased activity. Compounds with different substituents in the *N*-benzoyl part were similar to paclitaxel if these substituents were aromatic, aliphatic substituents reduced cytotoxicity.

As the 2'-hydroxyl group is essential for maximal biological activity [457] esterification leads to a total loss of activity in the microtubule assay, whereas cytotoxicity remains unchanged. Thus various amino acid esters were produced as prodrugs.

Notable loss of cytotoxicity was observed with A-ring modified analogs and oxidation at C-10 or C-7. 7-Acetylpaclitaxel and the C-7 epimer were similar in their ability to inhibit cell proliferation. All derivatives without the intact oxetane moiety are inactive [458]. Further, all C-4 modified analogs were devoid of activity underscoring the vital importance of these functional groups [459].

7.5. Epothilone A and B

The high cytotoxicity and good stabilization of microtubule raised interest in the natural products epothilone A [*152044-53-6*] and B [*152044-54-7*] originally isolated from myxobacteria *Sorangium cellulosum* [462]. Since Bollag et al. [463] reported on the mechanism of action, which resembles that of paclitaxel numerous reports were published on total synthesis [464–468] and biology of epothilones. Their unique capability to inhibit taxol resistant tumor cell lines [470] and their good solubility in water are the biggest advantages as compared to paclitaxel. Their in vivo activity is similar to that of paclitaxel [469].

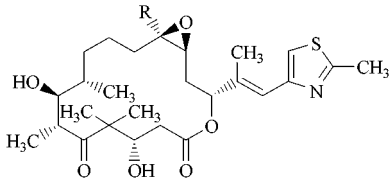

R = H, Epothilone A
R = CH_3, Epothilone B

8. Heavy-Metal Complexes

8.1. *cis*-Platinum

cis-Platinum (NSC-119875) [*15663-27-1*], *cis*-diamminedichloroplatinum, cisplatin, $PtCl_2(NH_3)_2$, M_r 300.05, was the first heavy-metal compound to be introduced into clinical cancer chemotherapy. The discovery of this unique agent was predicated on the serendipitous observation that bacterial growth was inhibited when culture medium was subjected to an alternating current using platinum electrodes [471]. Moreover, the spent medium itself developed bacteriocidal characteristics, even in the absence of electrical current. Detailed analysis confirmed that, of the several platinum species produced by electrolysis, it was the cis isomer of $PtCl_2(NH_3)_2$ that had antibacterial activity. In 1969, Rosenberg and co-workers reported that *cis*-platinum also had potent antitumor activity in murine tumors [472], and phase I clinical studies with the drug were initiated two years later [473], [474]. *Cis*-platinum, in combination with other agents, has led to highly active and often curative regimens in patients with testic-

ular, ovarian, and head and neck cancer [475], [476].

8.1.1. Mechanism of Action

As with the alkylating agents, one of the major factors mediating the cytotoxicity of *cis*-platinum is probably the formation of cross-links between opposing strands of DNA (interstrand cross-links), linkage within a single strand of DNA (intrastrand cross-links), or the formation of linkages dependent on the hydrolysis of *cis*-platinum in solution. While the covalent stability of $Pt-NH_3$ bonds is high, both chlorides are good leaving groups and can be displaced by water or hydroxyl ions to form positively charged, aquated platinum species that avidly react with nucleophilic sites on macromolecules, especially the N^7 position of guanine and the N^3 position of cytosine [477]. Formation of the active intermediate is inhibited in the presence of Cl^-; in plasma the Cl^- concentration is sufficient to inhibit aquation of the drug, which has an in vitro plasma half-life of several hours [478]. However, this process occurs rapidly in the intracellular milieu, where Cl^- concentrations are low.

The cis configuration is central to cytotoxicity; the trans isomer is devoid of antitumor effects. This fact suggests that of the reactions caused by the aquated platinum species, formation of intrastrand cross-links is most important (intrastrand cross-links cannot be formed by the inactive trans compound).

As a result of nucleophilic attack on macromolecules, *cis*-platinum causes changes in the structural conformation of DNA [479] as well as intra- and interstrand cross-links. These changes inhibit RNA transcription from the DNA template [480] and, probably more important, directly inhibit DNA synthesis itself [481]. In addition, *cis*-platinum can react with tumor cell membrane to cause presentation of new antigenic determinants [482]. Whether this mechanism is related to tumor cell recognition and immune response remains to be determined.

As a heavy-metal-based compound, *cis*-platinum has unique clinical toxicities, including renal tubular damage, severe nausea and vomiting, high-tone hearing loss, and peripheral neuropathy. Less common are myelosuppression, hemolytic anemia, hypomagnesemic tetany, allergic reactions, and hepatotoxicity. Nephrotoxicity was dose limiting in early clinical trials and appeared to be due to tubular reabsorption of active platinum species causing proximal and distal tubule necrosis [483–485]. Tubular damage has been reported to cause defective reabsorption of magnesium, resulting in hypomagnesemia (which may result in tetany) [486], [487], as well as inappropriate renal loss of calcium, potassium, and phosphorus [488]. The acute renal toxicities of *cis*-platinum appear to be secondary to activation of the renin – angiotensin system, resulting in reduced renal blood flow and glomerular filtration [484], [489]. Thus, this toxicity can be mitigated by high-volume diuresis [490]. In 1984 it has been demonstrated that even high doses of *cis*-platinum are well tolerated when administered with high-volume chloresis, which not only dilutes urinary platinum levels but also prevents leaving of the chloride groups to form the toxic aquated molecule [491]. When administered by this regimen, the limiting side effects of *cis*-platinum become neurotoxicity and myelosuppression, while renal function is remarkably spared.

8.1.2. Mechanisms of Resistance

Although bacterial resistance to *cis*-platinum appears to be due to increased efficiency of DNA repair [492], the mechanisms of tumor cell resistance are less clear. Alkaline elution studies have demonstrated a direct relation between DNA cross-linking and tumor cell resistance [493–495], and a *cis*-platinum-resistant murine leukemia line has been described in which cross-links are formed at a reduced rate [496]. However, it remains uncertain whether this resistance is secondary to accumulation, impaired activation, or altered DNA repair processes. As with the alkylating agents in general, high levels of metallothionein have been reported in *cis*-platinum-resistant cells [497]. This sulfhydryl-rich protein is known to protect from metal toxicity by specifically binding to platinum, cadmium, and other heavy metals [498].

Because the cellular mechanisms of resistance to *cis*-platinum are so poorly characterized, new platinum analogs in clinical trial have

been selected in an attempt to reduce host toxicity with the aim of improving therapeutic index [499].

8.2. Carboplatin

Carboplatin (NSC-241240, CBDCA) [*839805-03-3*], *cis*-diammine[1,1-cyclobutanedicarboxylato-(2)-*O,O'*]platinum(II), $C_6H_{12}N_2O_4Pt$, M_r 371.3, is a second generation cisplatin analog without significant nephrotoxicity or neurotoxicity and with less emetic potential than the parent compound [500–503]. Clinical trials show activity against several tumor types. Carboplatin is especially effective in treatment of ovarian and small-cell lung cancer [504], [505]. The minimal emetogenic dose of *cis*-platinum in dogs is 9 mg/m^2, whereas for carboplatin the dose is 624 mg/m^2. In addition, this compound retains antitumor activity in *cis*-platinum-restistant murine leukemia [506], demonstrating consistent antitumor activity in patients with ovarian cancer in the absence of either ototoxicity, nephrotoxicity, or neuropathy [507], [508]. Carboplatin entered the market in 1989 and is the leading platinum complex cancer drug with sales of $ 373 × 10^6 worldwide in 1996. Carboplatin is indicated for the treatment of ovarian cancer and sales have benefited from the drug's use in combination with Taxol [509–512].

Trade names: Paraplatin (Bristol-Myers Squibb), Carboplat (BMS).

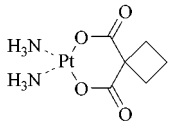

Carboplatin

8.3. Analogs

The synthesis and development of "second-generation" platinum compounds such as carboplatin has modified the problem of nephrotoxicity. Iproplatin (NSC-256927, CHIP) [*34348-60-2*], dichlorodihydroxy-bis-(2-propanamine)platinum $C_6H_{20}Cl_2N_2Pt$, M_r 418 [514], nedaplatin (NSC-375101D, 254-S), [*095734-82-0*], *cis*-diamine(glycolato-O^1,O^2)platinum, $C_2H_8N_2O_3Pt$, M_r 303.19 [515], and oxaliplatin (OHP) [*61825-94-3*], [Sp-4-2-(1R-*trans*)](1,2-cyclohexanediamine-*N,N'*)[ethanedioxato(2$^-$)-*O,O'*]-platinum, $C_8H_{14}N_2O_4Pt$ [516] were clinically investigated in depth. Nedaplatin entered the market in 1997 for the treatment of head and neck, small-cell lung, non-small cell lung, esophageal, bladder, testicular, ovarian, and uterine cervical cancers [517–520].

Trade name: Agupla (Shionogi).

Further studies of additional "third-generation" compounds are centering on the elimination of toxicity, enhanced therapeutic activity, non-cross-resistance and selective drug delivery [521]. Ormaplatin (NSC-363812, U-77233) [*62816-98-2*] [±(*trans*)]-tetrachloro(1,2-cyclohexanediamine-*N,N'*]platinum, $C_6H_{14}Cl_4N_2Pt$ and Lobaplatin (D-19466) [*135558-11-1*] are representatives of a series of platinum complexes in clinical development [522–524].

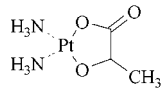

Nedaplatin

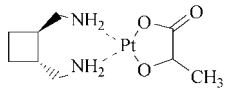

Lobaplatin

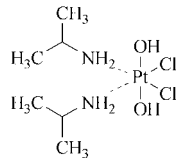

Iproplatin

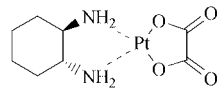

Oxaliplatin

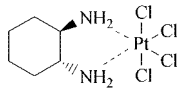

Ormaplatin

Orally active cisplatin-analogs with higher therapeutic ratio and specifically high anticancer activity are the challenge in the search for new "fourth-generation" derivatives. Platinum(IV)-complexes like JM 216, drug targeting approaches, e.g., the use of ligands with hormone receptor-binding affinity [525] or intercalating structures [526], and special formulation techniques are the major focus.

9. Hormonally Active Anticancer Drugs/Antihormones

Hormones and in particular, the sex hormones were the first growth factors discovered to be involuntary helpers of cancer. Female breast cancer and male prostate cancer are the best known examples of tumors acknowledged to be hormone-dependent. Shutting down the main production site of the sex hormones estrogen and testosterone either by removing the ovaries or by castration is a well-known and often effective therapy; however, these procedures can be problematic due to the concomitant psychological stress. Modern hormone therapy for advanced breast cancer and prostate cancer attempts to spare the patient such irreversible operative procedures for as long as possible by using hormone antagonists. Examples are the antiestrogens or LHRH antagonists, which hinder deployment of the hormone itself and thus its growth-promoting activity.

9.1. Antiestrogens (→ Hormones, Chap. 3.5.4.)

9.1.1. Antagonists

Estrogens can induce hormone-dependent human breast carcinoma and stimulate tumor growth. Reduced estrogen production is correlated with a lower risk of breast cancer and in particular with tumor regression. Hormonally active drugs are considered to be the treatment of first choice for advanced breast cancer, unless metastatic complications require immediate aggressive chemotherapy.

9.1.2. Tamoxifen, Toremifene

The antiestrogen tamoxifen (ICI 46474) [*10540-29-1*], (Z)-2-[*p*-(1,2-diphenyl-1-butenyl)phenoxy]-*N,N*-dimethylethylamine, $C_{26}H_{29}NO$, has become the standard first-line agent in postmenopausal patients [527–529]. The product is indicated for the treatment of breast cancer with worldwide sales of $\$561 \times 10^6$ in 1996.

Trade name: Nolvadex (Zeneca).

However, primary and secondary resistance to tamoxifen treatment requires the introduction of new analogs with improved therapeutic activity for hormon-dependent neoplasia. Toremifene (NK-622, Fc-1157a) [*089778-26-7*] [530], [531], 2-[*p*-[(Z)-4-chloro-1,2-diphenyl-1-butenyl]phenoxy]-*N,N*-dimethylethylamine, $C_{26}H_{28}ClNO$, M_r 405.97, raloxifene [532], and droloxifene [533] were investigated clinically.

Toremifene is a novel triphenylethylene derivative structurally related to tamoxifen [527–529]. Almost all compounds related to tamoxifen contain an alkylaminoethoxy side chain, which seems to be essential for their binding to the estrogen-receptor (ER) and antiestrogenic activity [536].

Trade names: Toremifene has been launched as Fareston (Farmos, ASTA Medica) and Estrimex (Adria) for treatment of postmenopausal breast cancer.

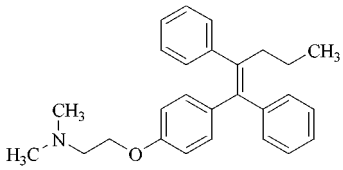

Tamoxifen

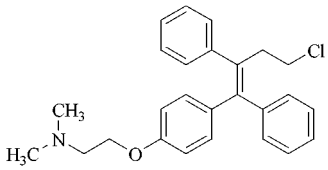

Toremifene

Mechanism of Action. Tamoxifen and toremifene bind to estrogen receptors in the cytosol, are translocated to the nucleus, and block estrogen-induced cell proliferation. However, specific differences in the drug profiles exist, indicating improved therapeutic properties of toremifene compared with tamoxifen. These differences include lower intrinsic estrogenic activity, longer nuclear retention, no retinal damage or neoplastic liver changes. Toremifene is more active against dimethylbenz[a]anthracene (DMBA)-induced rat mammary cancer

and unlike tamoxifen, it inhibited the growth of an ER-negative transplantable mouse uterus sarcoma, although the antitumor effect of this compound was preferentially directed against estrogen-dependent tumors of the mammary gland and endometrium [534], [535]. In clinical trials no concrete side effects or pathological clinical chemistry values were observed in most patients [537], [538]. Some patients complained of light hot waves, sweating, nausea, and transient vertigo.

Cytotoxicity in vitro, dose related activity in ER-positive (i.e., cancer cells containing estrogen receptors) and ER-negative tumor models suggest that additional mechanism like growth-factor production may be triggered by toremifene [539].

9.1.3. Analogs

Raloxifene (LY-139481) [*82640-04-8*] is a benzo(b)thien-3yl-antiestrogen formerly under development for treatment of breast cancer. It mimics the effects of estrogen on the skeleton and is therefore effective in the prevention of postmenopausal osteoporosis [540]. Faslodex (ICI-182780, ZD-182780) [*129453-61-8*] is a pure, steroidal estrogen antagonist with oral anticancer activity [541], [542]. The in vivo antitumor activity of faslodex in xenografts of MCF-7 and Br 10 human breast cancers in mice was equivalent to that of tamoxifen. Miproxifene phosphate (Tat-59) [*115767-74-3*] [543], panomifene (Gyki-13504) [*77599-17-8*], idoxifene [544] (SB-223030, CB-7432) [*116057-75-1*] and droloxifene [*83647-29-4*] are more potent than tamoxifen against estrogen-dependent tumors in mice.

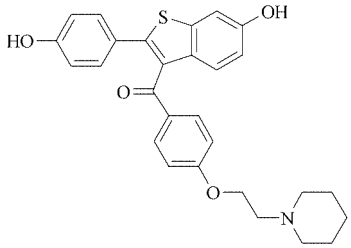

Raloxifene

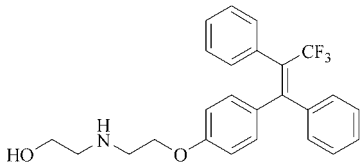

Panomifene

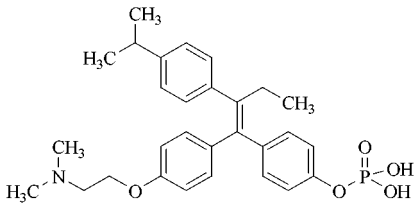

Miproxifene phosphate

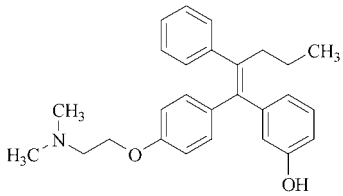

Droloxifene

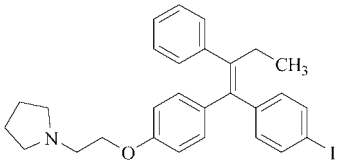

Idoxifene

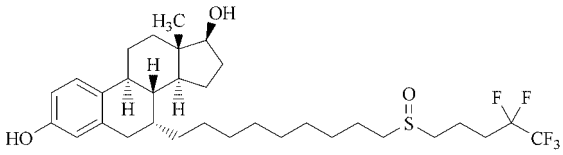

Faslodex

9.2. Aromatase Inhibitors (→ Hormones, Chap. 3.5.4.)

All endocrine therapies inhibit endogenous estrogen production or the interaction between estrogens and cellular estrogen receptors. Both ablative and additive hormone therapies are equal in efficiency.

Aromatase inhibitors block cellular estrogen synthesis and thus induce, particularly during postmenopause, a marked decrease in estrogen production. The inhibition of estrogen synthesis is caused by a suppression of the enzyme aromatase which converts androstenedione to estrone [545].

Aminoglutethimide [*125-84-8*], (±)-2-(4-aminophenyl)-2-ethylglutarimide, an unspecific aromatase inhibitor which also suppresses adrenal desmolase and 11-β-hydrolase, was the first aromatase inhibitor in the clinic and on the market.

Trade name: Orimeten (Ciba-Geigy).

Aminoglutethimide induces a decrease in cortisol, followed by an increase in ACTH [546]. Therefore new powerful specific aromatase inhibitors without influence on adrenal steroid synthesis were developed.

Formestane (4-OHA; CGP-32349) [*000566-48-3*], 4-hydroxyandrost-4-ene-3,17-dione, $C_{19}H_{26}O_3$, M_r 302.41, is an androstane derivative with highly specific aromatase inhibition [547]. The substance is used for the treatment of advanced breast cancer in postmenopausal women [548–553].

Trade name: Lentaron (Ciba-Geigy).

Atamestane, exemestane, and NKS-01 are orally active steroids.

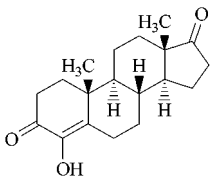

Formestane

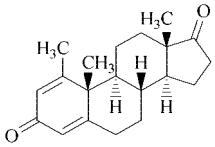

Atamestane

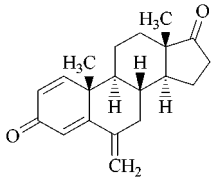

Exemestane

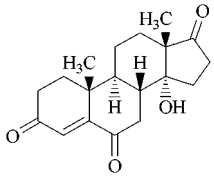

NSK-01

In addition to steroidal also nonsteroidal, in particular, imidazole/triazole derivatives were investigated as aromatase inhibitors. Anastrazole (ICI-D1033, ZD-1033) [*120511-73-1*] [554], 2,2'-[5-(1H-1,2,4-triazol-1-ylmethyl)-1,3-phenylene]-bis(2-methylpropionitrile), $C_{17}H_{19}N_5$, M_r 293.37, fadrozole (CGS-16949A) [*102676-96-0*] [555], (±)-4-(5,6,7,8-tetrahydroimidazo[1,5-a]pyridin-5-yl)benzonitrile, $C_{14}H_{13}N_3$, M_r 259.74, letrozole (CGS-20267) [*112809-51-5*] [556], 4,4'-(1H-1,2,4-triazol-1-ylmethylene)bis[benzonitrile], $C_{17}H_{11}N_5$, and vorozole (R83842), [*129731-10-8*] [557], (+)-6-[(4-chlorophenyl)-1H-1,2,4-triazol-1-ylmethyl]-1-methyl-1H-benzotriazole, $C_{16}H_{13}ClN_6$, are highly selective nonsteroidal aromatase inhibitors without intrinsic androgenic or estrogenic properties.

Trade names: Anastrazole and Fadrazole are marketed as Arimedex (Zeneca) and Arensin (Ciba-Geigy), respectively, for treatment of postmenopausal breast cancer.

Another aminoglutethimide analog in development is rogletimide [558] (pyridoglutethimide) [*121840-95-7*], (±)-3-ethyl-3(4-pyridinyl)-2,6-piperidinedione, $C_{12}H_{14}N_2O_2$.

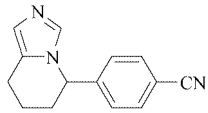

Fadrazole

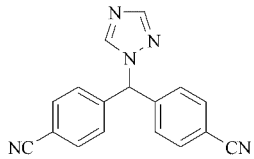

Letrozole

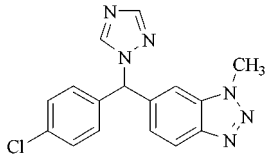
Vorozole

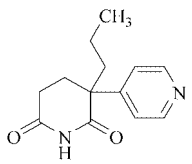
Rogletimide

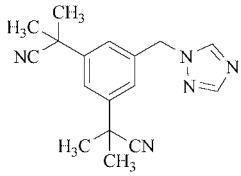
Anastrazole

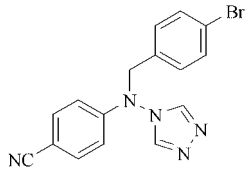
YM-511

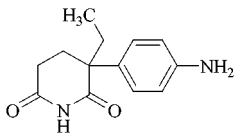
Aminoglutethimide

Mechanism of Action. Fadrazole [559] and anastrazole [560–567] for example, were found to be potent and specific aromatase inhibitors with neither androgenic nor estrogenic activity. Fadrazole was 180 times more potent than aminoglutethimide as an aromatase inhibitor. Anastrazole inhibited human placental aromatase with an $IC_{50} = 15$ nM (IC_{50} = inhibitory concentration). In animal studies fadrazole was found to lower serum estrogen, raise luteinizing hormone (LH) level, and reduce uterine weight as a result of aromatase inhibition. In addition, administration of 2 mg per kilogram body weight caused regression of DMBA-induced mammary tumors in female rats. In humans half-life of, e.g., anastrazole was more than 30 h. No serious side effects were reported and no significant effects on cortisol or aldosterone secretion was observed.

9.3. Antiandrogens

Prostate cancer is primarily a disease of the elderly and the second most common cancer in men in the United States. Ever since HUGGINS and HODGES demonstrated the partial androgen dependence of most prostatic tumors more than 50 years ago, androgen deprivation has become the commonly used initial treatment for prostate cancer [568], [569].

The major circulating androgen in man is testosterone, 90 % of which is produced in the testis. In addition, a small amount of androgen is produced by the adrenal gland under the control of ACTH. (see Fig. 3).

Four ways for androgen deprivation exist:

1) Removal of organs by surgery
2) Interference with control mechanism
3) Inhibition of biosynthesis
4) Competitive inhibition of androgens at the receptor site

The group 2 approach with, e.g., LHRH agonists and antagonists will be discussed in Section 9.4. Examples of group 3 include compounds (e.g., ketoconazole), which inhibit the synthesis of adrenal androgens [570] as well as inhibitors of the enzymes 5α-reductase [571] and aromatase [572]. "True" antihormones are compounds of group 4, based on the definition that an antiandrogen is a substance which binds to the target tissue androgen receptor and prevents the stimulatory effects of androgens. Although the exact mechanism of action of antiandrogens is not totally understood, an important feature is the competitive inhibition of the binding to the cytosol receptor. The first antiandrogen to be used clinically was the synthetic steroidal antiandrogen cyproterone acetate [*427-51-0*] [573]. However, cyproterone acetate also exhibits other steroidal activities, it is, e.g., a potent progestin, exhibits weak antigonadotrophic activity and has glucocorticoid like properties. Beyond the success in treating prostate cancer the steroidal properties are largely responsible for the fluid retention and thrombosis seen in patients [574]. Therefore, nonsteroidal antiandrogens are expected to have advantages by avoiding steroid-related side effects [575].

9.3.1. Flutamide

Flutamide (Sch-13521, NK-601, FTA) [*013311-84-7*], 2-methyl-*N*-[4-nitro-3(trifluormethyl)phenyl]-propanamide, $C_{11}H_{11}F_3N_2O_3$ was discovered in the early 1970s. Unlike the steroids it can easily be synthesized [576], [577]

and is devoid hormonal activities. Total sales of the product were $ 271 × 10^6 in 1996.

Trade names: Eulexin and Drogenil (Schering-Plough).

Mechanisms of Action. In a comparative study in castrated rats flutamide was shown to be equipotent to cyproterone acetate as an antiandrogen [578]. Several groups had observed that flutamide was a more potent antiandrogen in vivo than in vitro and suggested the involvement of an active metabolite [579], [580]. The major metabolite was identified as 2-hydroxyflutamide analog (Sch-16423) [*52806-53-8*] and high levels of 2-hydroxyflutamide in the plasma led to the conclusion that this was the active form of flutamide [581]. Although the plasma levels of testosterone increased on flutamide treatment, the levels of testosterone and dihydrotestosterone in androgen target tissue were reduced [582]. Flutamide exerts its antiandrogenic action by blocking the binding of androgens to the cytosolic androgen receptor and / or inhibiting the nuclear binding of androgens in the target tissue [581].

In patients improvements were seen in pain relief, prostatic enlargement and induration, reduction of metastases, and increase in body weight and phosphatase. Most common side effects of flutamide therapy are gynecomastia and breast tenderness [584]. More recent trials use a combination of flutamide with LHRH agonists (see Section 9.4). This is the concept of maximal androgen withdrawal in which the LHRH agonist wipes out androgens of testicular origin and the antiandrogen blocks the action of androgens of adrenal origin at the androgen receptor [585–587]. The therapeutic benefits seem to be greatest in patients with minimal disease at the start of treatment.

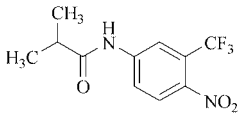

Flutamide

9.3.2. Nilutamide

Nilutamide has been discovered bei Roussel-Uclaf.

Trade name: Anandron (Roussel-Uclaf).

Mechanism of Action. Nilutamide competitively inhibits binding of androgens to the cytosolic androgen receptor. Administration over 7 days to immature, castrated male rats, nilutamide inhibited the increase of prostate weights induced by testosterone in a dose-dependent manner [588]. In rat pituitary cells nilutamide reverses the inhibition of LHRH-induced LH release elicited by dihydrotestosterone. It is probably because of these effects that nilutamide is recommended for use only in surgically or medically castrated males [589]. Nilutamide has demonstrated an antiandrogenic action in several animal tumor models [590].

Single dose kinetics of nilutamide in volunteers (100 mg) indicate a half-life of 43 ± 3 h, compared to 5.2 h for flutamide (200 mg) [591].

Clinical trials with nilutamide have concentrated on combination therapy with surgical or medical castration. The side effects observed in patients include hot flushes, nausea, vomiting, and visual problems [592].

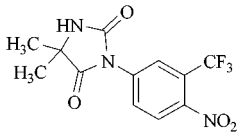

Nilutamide

9.3.3. Bicalutamide

Bicalutamide (ICI-176334) [*090357-06-5*], (±)-4-[3-(4-fluorophenylsulfonyl)-2-hydroxy-2-methylpropionamido]-2-(trifluoromethyl)benzonitrile, $C_{18}H_{14}F_4N_2O_4S$, M_r 430.37 was discovered at ICI / Zeneca and selected from more than 1000 compounds as having the desired properties of a pure, nonsteroidal, peripherally selective antiandrogen [593], [594]. The product was launched for the treatment of advanced prostate cancer in combination with LHRH analogs or surgical castration.

Trade name: Casodex (Zeneca).

Mechanism of Action. Bicalutamide inhibits the binding of the synthetic androgen [^{3}H]-R-1881 to both rat prostate and pituitary cytosol androgen receptors. The substance

binds some 50 times less effectively than dihydrotestosterone and about 100 times less effectively than R-1881 to the prostate androgen receptor, Its affinity for the prostate receptor is about four-fold higher, that for the pituitary receptor ten times higher than that of hydroxyflutamide [595]. In vivo studies revealed that bicalutamide is about five times as potent as flutamide after oral application. Bicalutamide did not cause a significant elevation in LH or testosterone at any of the doses tested, whereas flutamide elicited increases. Half-life of bicalutamide in prostate cancer patients who received 10, 30 or 50 mg/d bicalutamide was around 6 d. The compound was well tolerated in all doses [596–598].

Nilutamide and bicalutamide offer advantages over flutamide because of their long half-lifes and sustained serum levels on once-daily dosing. The latter is essential to prevent androgen stimulation.

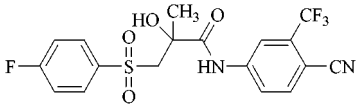

Bicalutamide

9.4. LHRH Analogs (→ Peptide and Protein Hormones, Chap. 2.1.)

The releasing hormone gonadorelin (GnRH, synonymous with LHRH, luteinizing hormone-releasing hormone or gonadoliberin), together with its specific receptor, plays a central role in neuroendocrinology [599]. The decapeptide LHRH is formed in the cell bodies of hypothalamic neurons and is secreted in pulses into the blood stream [600–603]. Ultimately it stimulates secretion of the sex-specific hormones in the testes and ovaries. Specific receptors for LHRH and synthetic analogs are also present in the pituitary gland and other tissues (for example, tumor cells) [604] and organs.

Three concepts for therapeutic application have emerged. The first is the restoration of normal physiology by administration of LHRH by infusion pump to promote fertility in men and women who are infertile due to defective endogenous LHRH secretion. Second, long-lasting LHRH agonists (so-called superagonists) are used in a depot form, which bring about desensitization of the pituitary receptors and thus interrupt the signal cascade. This results in a biochemical "castration", which opens up new therapeutic possibilities for hormone-dependent diseases such as prostate cancer, breast cancer, and endometriosis. Although superagonists are generally well tolerated, they have the disadvantage that hormone secretion (estrogen, testosterone) is initially stimulated before the depletion of receptors or "down" regulation can take place, and thus the illness temporarily worsens [605]. This has led to the development of the third concept: use of LHRH antagonists [606] (see Fig. 3). In the late 1980s about 5000 LHRH analogs had been synthesized worldwide and tested in vitro or in vivo [607–609]. Whilst LHRH agonists have been on the market for about ten years, the LHRH antagonists that have been developed farthest are still in clinical testing [610–614].

9.4.1. LHRH Agonists

The first years after the discovery of the gonadorelins were marked by the search for more active agonists, since the therapeutic potential of gonadorelins as, for example, antitumor agents or in gynecology, was apparent [606], [615]. Such superagonists bring about a very effective reversible inhibition of the release of steroidal sex hormones. The exchange of glycine at postion 6 ($glycine_6$) of the native LHRH for other, always D-configured, amino acids, is common to all modern superagonists; some have a *C*-terminal ethylamide (buserelin, leuprorelin) or azaglycinamide (goserelin) residue instead of glycinamide. Eight to ten amino acids of the LHRH sequence are thus conserved in all clinically relevant superagonists; by exchange at a maximum of two positions, the biological activity or hormone suppression in tumor patients, can be increased by a factor of up to 100 on subcutaneous application [606], [615]. Table 1 summarizes the most important derivatives.

Buserelin (Profact, Suprecur) [*57982-77-1*], leuprorelin (Lupron, Carcinil, Enatone) [*53714-56-0*], triptorelin (Decapeptyl) [*57773-63-4*], and goserelin (Zoladex) [*65807-02-5*] (trade names in Germany in parantheses) are the products on the market with the

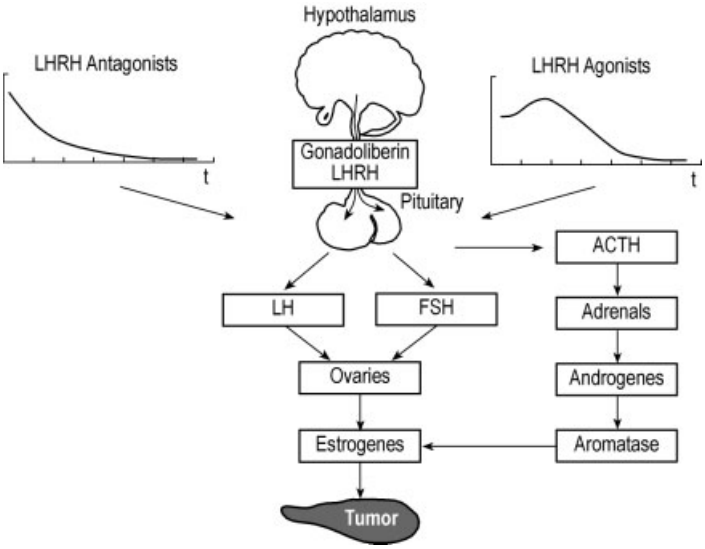

Figure 3. Antitumor activity of LHRH agonists and antagonists in femaleFSH = follicle stimulating hormone; ACTH = adrenocorticotropic hormone

Table 1. Structure of LHRH agonists on the market (given are only those amino acid residues that are different in LHRH).

Name (Co.)	Structure									
	1	2	3	4	5	6	7	8	9	10
LHRH	Glp[a]	His	Trp	Ser	Tyr	Gly	Leu	Arg	Pro	Gly-NH_2
Buserelin (Hoechst)						D-Ser(*t*Bu)				Gly-NHEt
Nafarelin						D-(2)Nal				
Leuprorelin (Abbott, Takeda)						D-Leu				Gly-NHEt
Goserelin (Zeneca)						D-Ser(*t*Bu)				Azagly-NH_2
Histrelin (Ortho)						D-His(Bzl)				Gly-NHEt
Triptorelin (Ferring)						D-Trp				

[a] Glp = pyroglutamic acid.

highest turnover. The Lupron line of products is the leading anticancer hormone drug with worldwide sales in 1996 of $ 810 × 10^6 followed by Zoladex with sales of $ 563 × 10^6.

Annual production is less than 100 kg for buserelin and significantly over 100 kg for the market leader, leuprorelin. At this order of magnitude, substances are only produced by classical organic preparative synthesis (fragment condensation in solution). Hoechst, for example, synthesizes the nonapeptide buserelin from the units pyroglutamic acid-histidine (Glp-His), tryptophan-serine-tyrosine (Trp-Ser-Tyr), and D-Serine-tert-butyl ether-leucine-arginine-proline-NHEt (D-Ser(*t*Bu)-Leu-Arg-Pro-NHEt); the tri- and tetrapeptide units are coupled to form the corresponding *C*-terminal heptapeptide, and then the *N*-terminal dipeptide is condensed with this to form the complete sequence. Control of the physicochemical process parameters, such as concentrations, precipitation, separations, reaction temperature profiles, and purification techniques, is important for successful scale-up of peptide synthesis to a technical scale [617].

9.4.1.1. Leuprorelin Acetate

Because leuprorelin inhibits the synthesis of androgen and estrogen, the drug blocks the growth of hormone-dependent tumors by shutting down testosterone production. Leuprorelin is indicated for the treatment of advanced prostate cancer, as an alternative to castration

for the treatment of endometriosis, and for the presurgical management of patients with anemia caused by benign fibroid tumors.

Trade name: Lupron (Abbott, Takeda).

9.4.1.2. Goserelin

As certain prostate tumors grow in response to testosterone, goserelin prevents the production of testosterone in testes and is therefore indicated for the treatment of prostate cancer, advanced breast cancer, and endometriosis.

Trade name: Zoladex (Zeneca).

9.4.2. LHRH Antagonists

Common to the intrinsic activity of all superagonists is the initial temporary stimulation of gonadotropin release. Soon after the use of highly active agonists became an established therapy, a search began for corresponding antagonists, which do not bring about an initial hormone release, to avoid this therapeutically counterproductive effect. A final big hurdle for clinical use of highly active antagonists was the inherent anaphylactic potential of these peptides. Starting from the sequence of native LHRH, the individual positions of the peptide chain were examined in rapid succession for their contribution to biological activity. Particular attention was paid to side effects. The most effective early improvements in antagonistic activity were achieved by using D-phenylalanine (D-Phe) instead of histidine at position 2, by D-amino acids at position 6 instead of Gly, and the exchange of *C*-terminal glycine for D-alanine (D-Ala^{10}). Further stepwise optimization led to the sequence scheme now usual for all modern antagonists of D-Nal^{1}-D-Cpa^{2}-D-Pal^{3} (Nal = 2-naphthylalanine, Cpa = Phe(4-Cl), Pal = 3-pyridylalanine) as hydrophobic cluster, a D-configured aromatic or aliphatic, yet hydrophilic, aminocarboxylic acid at position 6, and the *C*-terminal hydrophilic sequence Xxx^{8}-Pro^{9}-D-Ala^{10} where Xxx is either arginine or isopropyllysine. Excellent documentation of the stepwise optimization can be found in [618], [619].

Antagonists of the second generation caused formation of temporary edemas of the face and extremities in animal experiments, due in part to massive histamine release by mast cell degranulation. Cyanosis and respiratory impairment were also observed [620]. The cause of these intolerable side effects is thought to be the combination of D-arginine at position 6 with the three aromatic amino acids at the N-terminus of the sequence [621]. For the desired biological antagonist potency, a D-configured basic amino acid is necessary at position 6. A. V. Schally et al. achieved the breakthrough to highly active antagonists free of side effects with the derivatives SB-75 (cetrorelix) [*120287-85-6*] and SB-88 [*120287-93-6*], by incorporating hydrophilic, nonbasic amino acids with side-chain carbamoyl functions at position 6 (D-Citrulline^{6}, D-homocitrueline) [621], [622].

Today, e.g., cetrorelix is *manufactured* exclusively by classical fragment condensation on a kilogram scale. Two synthesis strategies proceed via either the *N*-terminal tripeptide (D-Nal^{1}-D-Cpa^{2}-D-Pal^{3}) or the *C*-terminal tripeptide (Arg^{8}-Pro^{9}-D-Ala^{10}) and the complementary heptapeptide (Ser^{4}-D-Ala^{10}) or (Nal^{1}-Leu^{7}) to the protected decapeptide with *tert*-butyl side chain protection. The deprotection with hydrochloric acid is followed by final purification by preparative HPLC. The *C*- and *N*-terminal functionalization of the acetylated decapeptide amide, necessary for biological activity and to avoid rapid enzymatic degradation, is introduced at the level of the terminal tripeptide. Functionalization is achieved by acetylation of the free α-amino group of naphthylalanine^{1} with acetylhydroxysuccinimide, or amminolysis of the resulting alanine^{10} methyl ester in alcoholic ammonia [623].

Folkers et al. were able to improve active antagonists successively by consistent and systematic modifications both at the relevant sequence positions and the side chain substituents; the most successful were complex substitutions at the positions 5, 6 and, in part, 8 [624–628]. An important contribution was made by Rivier et al., who synthesized the decapeptide "azaline" with novel modifications at positions 5 and 6, where aminotriazole-substituted *p*-aminophenylalanine or lysine are positioned [629–631]. Azaline B [*134457-28-6*] is probably one of the most active antagonists presently available worldwide.

DEGHENGHI published a highly active decapeptide sequence with minimal histamine release and good water solubility. This structure, known as antarelix [*151277-78-5*], differs from SB-75 (cetrorelix) in that it has homocitrulline^6 instead of citrulline^6 and isopropylysine^8 instead of arginine^8 [633]. By using Lys(*i*Pr)^8, residual potential for histamine release can be further reduced. Organon is developing the antagonist ganirelix (RS26306) [*124904-93-4*] under license from Syntex; this is a decapeptide with novel alkyl-modified D- and L-homoarginine units at positions 6 and 8 [634]. With ramorelix (HOE 013) [*127932-90-5*], Hoechst (HMR) has a peptide antagonist with a sugar – amino acid unit (*O*-α-L-rhamnosyl-D-serine^6), that has improved water solubility [632]. Schering has also synthesized peptide antagonists, using nonproteinogenic amino acids, such as ε-dialkylated lysine or benzodiazepine aminocarboxylic acids [635], [636]. Abbott's A-76154 [*136989-30-5*] is an octapeptide antagonist with LHRH receptor affinity of a similar order of magnitude to that of active decapeptides such as A-75998 [*135215-95-1*] or "Nal-Glu" [*103733-02-4*] [644].

9.4.2.1. Receptor Assays

At present, selected peptides are being tested in vivo by measuring the testosterone concentration in male rat serum. Here, a single subcutaneous application of a potent LHRH antagonist leads to persistent testosterone suppression. The activity in animals generally correlates well with the binding affinity determined in vitro on human receptors. In cases where no in vivo activity was observed despite high binding affinity, the peptide probably had pharmacokinetic characteristics that played a decisive role. In the future, transgenic animals will also be very important for in vivo testing [645–647].

9.4.2.2. Peptidomimetics

For some years, increased efforts have also been made to find substances with affinity to the LHRH receptor that do not have the characteristic substance-specific properties and also disadvantages of peptides (short half-life, lack of bioavailability), yet have a high binding affinity. Ideally, such substances should be able to be administered orally, be sufficiently stable in the organism, and possess favorable pharmacological parameters comparable to peptide antagonists. Clinical investigations of the influence of the antimycotic ketoconazole on prostate cancer and testosterone suppression indicate that it may have a LHRH-mimetic effect [648], [649]. Further investigations by Abbott showed weak antagonistic activity both in vitro and in vivo for ketoconazole and its modified analogs; however, the biological activity of such derivatives is most probably not mediated mainly by LHRH antagonism [650], [651]. Complex, highly substituted nitrogen heterocycles seem to have LHRH antagonistic potential [652]. In patent literature benzodiazepines, benzodiazepinones, heterocyclic benzo-substituted alkylamines, and thienopyridine carboxylic acid derivatives are described as LHRH receptor antagonists with receptor-binding inhibition at submicromolar concentrations, suitable amongst other uses as antitumor agents for hormone-dependent tumors [647], [653]. A lead structure from the latter substance class is in extensive pharmacological trials [654–656].

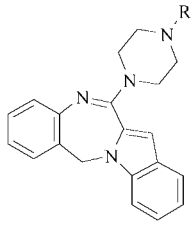

McNeil, US 4 678 784

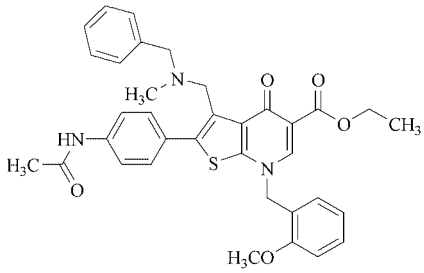

Takeda, WO 95/28405

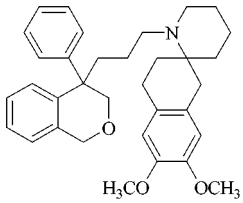

Takeda, EP 679 642 A1

10. Signal Transduction Inhibitors

Anticancer drug discovery has been directed away from agents that affect cells by producing DNA damage towards modulators of signal transduction pathways that have become unregulated or aberrant in malignant transformation [657].

10.1. Enzyme Inhibitors

Binding of growth factors [658], such as epidermal growth factor (EGF) [659] to a membrane bound receptor tyrosine kinase results in dimerization and autophosphorylation of tyrosine residues on the protein surface. As result the GTP-bound form of *ras* (→ Toxicology, Chap. 3.7.3.1.) [660] undergoes a conformational change on its surface, enabling it to bind to several effector molecules leading to the activation of transcription factors involved in DNA synthesis [661]. An overexpression of the EGF-receptor is observed in various types of human cancers.

Mechanism-based screens have identified the natural products bryostatin [662] [*83314-01-6*] as partial agonist of protein kinase C and fumagillin [663] [*23110-15-8*] as interfering with tumor-induced neovascularization thus inhibiting angiogenesis. FR-111142 as analog of fumagillin is reported to be less toxic and more active than the parent compound [664]. A potent inhibitor of *ras* protein farnesylation [665] is pepticinnamin E [666], and peptides related to the CAAX-tetrapeptide (where C is cysteine, A is valine, isoleucine, or leucine, and X is methionine or serine). In order to prepare completely nonpeptide and potentially more stable inhibitors a series of potent peptidomimetic inhibitors were designed [667].

Specific natural and synthetic kinase inhibitors are in clinical development as anticancer agents [668–670].

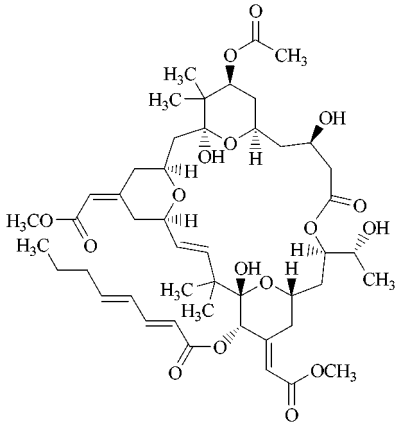

Bryostatin 1

10.2. Phospholipid – Based Antineoplastics

Ether phospholipids and lysophospholipids are naturally occurring derivatives of phospholipids of the cell membrane with many interesting properties. For example, the ether phospholipid 1-*O*-alkyl-2-acetyl-sn-glycero-3-phosphocholine (platelet activating factor, PAF) [671] causes platelet aggregation and dilation of blood vessels and lysophosphatidic acid (LPA), the simplest natural phospholipid, is a potent mitogen. The observation that alkyllysophospholipids exerted experimental antitumor activity in vitro and in vivo [672], [673] led to systematic modification of the structures and subsequent clinical investigation of, e.g., ilmofosine [*83519-04-4*] [675]. In the course of modification it was found that an analog without glycerol backbone was equally effective.

Miltefosine (D-18506, hexadecylphosphocholine), [*058066-85-6*] [676], 2-[[(hexadecyloxy)-hydroxyphosphinyl]oxy]-*N,N,N*-trimethylethanaminium hydroxide (inner salt); $C_{21}H_{46}NO_4P$, M_r 407.57 is a new phospholipid-based antineoplastic agent which

exerts substantial antitumor activity in appropriate models. Because of its special pharmacology – high activity against mammary carcinoma and low toxicity to normal tissue – it is an ideal candidate for topical treatment of cutanous breast cancer metastases. The substance was launched in 1993 as a topical formulation for palliative treatment of refractory skin metastases of breast cancer [676].

Trade name: Miltex (ASTA Medica).

A variety of derivatives were synthesized and characterized, some of which appeared to be considerably less toxic to the gastrointestinal tract [677]. One of them, perifosine (D-21266) [*157716-52-4*] is currently in clinical trials.

Mechanism of Action. Miltefosine shows highly selective antineoplastic activity in DMBA- and *N*-methyl-*N*-nitrosourea (MNU)-mammary carcinoma of the rat. The activity is not due to an antiestrogenic effect in these estrogen-dependent tumor models [678]. The treatment was tolerated and no overt toxic symptoms were observed. The mechanism of action of miltefosine is most probably different from that of other chemotherapeutic agents [679]. Incorporation of miltefosine and metabolites into biological membranes was demonstrated, affecting the interactions of receptor proteins or membrane associated enzymes involved in growth control and cellular signalling [680]. The reduction in tumor mass was accompanied by morphological changes compatible with the induction of differentiation [675], [681]. Furthermore, the decrease of tumors was mainly due to cell loss by apoptosis.

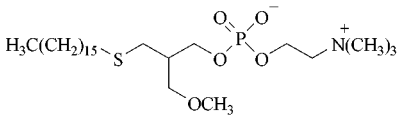

Ilmofosine

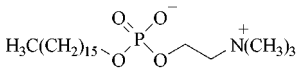

Miltefosine

In clinical trials mild dryness and flaking of the skin were noted, but no systemic toxicity was reported after topical administration. Several phase I and II studies in patients that were treated with oral capsules were terminated early due to gastrointestinal intolerance [682].

11. Economic Aspects

The world cancer drug market is expected to reach $ 14 × 10^9 by 2000 [10], [14]. Total sales of the top 100 anticancer drugs generated a turnover of $ 4.66 × 10^9 in 1996. Of the worlds eight top-selling anticancer drugs, four – the prostate cancer-drugs flutamide (Eulexin), bicalutamide (Casodex), leuprorelin (Lupron) and goserelin (Zoladex) – are merely palliative, yet have combined annual sales of $ 1.7 × 10^9, whereas sales of the breast cancer drugs tamoxifen (Nolvadex) and paclitaxel (Taxol) are approaching $ 500 × 10^6 and $ 850 × 10^6, respectively.

Taxol, BMS's drug for treating ovarian and breast cancers, was the biggest selling anticancer agent in 1996, with sales of $ 813 × 10^6.

12. References

1. I. B. Weinstein, *Cancer Res. (Suppl.)* **51** (1991) 5080.
2. R. Doll, *Eur. J. Cancer* **26** (1990) 5007.
3. C. LaVecchina, F. Levi, F. Lucchini, S. Garattini, *Anti-Cancer-Drugs* **2** (1991) 215.
4. D. Kessel, *In Vivo* **8** (1994) 829.
5. T. Tsurno, A. Toninda, *Anti-Cancer-Drugs* **6** (1995) 213.
6. G. Eisenbrand, S. Lauch-Birkel, W. C. Tang, *Synthesis* (1996) 1246.
7. G. B. Elion, *Cancer Res.* **45** (1985) 2943.
8. *Scrip* 1997, July 15, p. 20.
9. *Chem. Marketing Report,* July 14, 1997, p. 19.
10. *Med. Ad-News,* May 1996, p. 56.
11. G. M. Cragg, D. J. Newman, K. M. Snader, *J. Nat. Prod.* **60** (1997) 52 – 60.
12. K. Nasmyth, *Science* **274** (1996) 1643 – 1645.
13. G. Dratta, H. Pagano, *Annu. Rep. Med. Chem.* **31** (1996) 241 – 248.
14. *Pharma Business,* July/August 1997, p. 62.
15. S. A. Rosenberg: "Principles of Surgical Oncology," in V. T. DeVita, S. Hellman, S. A. Rosenberg (eds.): *Cancer: Principles and Practice of Oncology,* J. B. Lippincott Co., Philadelphia 1982, pp. 93 – 102.
16. S. Farber, L. K. Diamond, R. D. Mercer, R. F. Sylvester, V. A. Wolff, *N. Engl. J. Med.* **238** (1948) 787 – 793.

17. R. Hertz, G. T. Ross, M. B. Lipsett, *Am. J. Obstet. Gynecol.* **86** (1963) 808 – 814.
18. R. L. Kisliuk, Y. Gaumont, C. M. Baugh, J. M. Galivan, G. F. Maley, F. Maley: "Inhibition of Thymidylate Synthetase by Poly-gamma-glutamyl Derivatives of Folate and Methotrexate," in R. L. Kisliuk, G. M. Brown (eds.): *Chemistry and Biology of the Pteridines,* Elsevier/North Holland, New York 1979, pp. 431 – 435.
19. I. D. Goldman, N. S. Lichtenstein, V. T. Oliverio, *J. Biol. Chem.* **243** (1968) 5007 – 5017.
20. F. M. Sirotnak, R. C. Donsbach, *Cancer Res.* **36** (1976) 1151 – 1158.
21. R. D. Warren, A. P. Nichols, R. A. Bender, *Cancer Res.* **38** (1978) 668 – 671.
22. P. L. Chello, F. M. Sirotnak, D. M. Dorck et al., *Cancer Res.* **37** (1977) 4297 – 4303.
23. B. T. Hill, B. D. Bailey, J. C. White et al., *Cancer Res.* **39** (1979) 2440 – 2446.
24. F. M. Sirotnak, R. C. Donsbach, *Cancer Res.* **35** (1975) 1737 – 1744.
25. J. J. McGuire, J. R. Bertino, *Mol. Cell Biochem.* **38** (1981) 19 – 48.
26. R. L. Schilsky, B. D. Bailey, B. A. Chabner, *Proc. Natl. Acad. Sci. USA* **77** (1980) 2919.
27. S. A. Jacobs, C. J. Derr, D. G. Johns, *Biochem. Pharmacol.* **26** (1977) 2310 – 2313.
28. V. M. Whitehead, *Cancer Res.* **37** (1977) 408 – 412.
29. J. Jolivet, B. A. Chabner, *J. Clin. Invest.* **72** (1983) 773 – 778.
30. S. V. Gupta, N. J. Greenfield, M. Poe et al., *Biochemistry* **16** (1977) 3073 – 3079.
31. H. Nakamura, J. Littlefield, *J. Biol. Chem.* **247** (1972) 179 – 187.
32. D. A. Matthews, R. A. Alden, J. T. Bolin et al.:"X-ray Structural Studies of Dihydrofolate Reductase," in R. L. Kisliuk, G. M. Brown (eds.): *Chemistry and Biology of the Pteridines,* Elsevier/North Holland, New York 1979, pp. 465 – 470.
33. P. A. Charlton, D. W. Young, B. Birdsall et al., *Chem. Commun.* **20** (1979) 922 – 924.
34. D. A. Matthews, R. A. Alden, J. T. Bolin et al., *Science* **197** (1977) 452 – 455.
35. W. L. Werkheiser, *Cancer Res.* **23** (1963) 1277 – 1285.
36. B. A. Kamen, W. Whyte-Bauer, J. R. Bertino, *Biochem. Pharmacol.* **32** (1983) 1837 – 1841.
37. R. J. Kaufman, J. R. Bertino, R. T. Schimke, *J. Biol. Chem.* **253** (1978) 5852 – 5860.
38. E. W. Alt, R. E. Kellems, J. R. Bertino et al., *J. Biol. Chem.* **253** (1978) 1357.
39. G. A. Curt, J. Jolivet, B. D. Bailey, D. N. Carney, B. A. Chabner, *Biochem. Pharmacol.* **33** (1984) 1682 – 1685.
40. G. A. Curt, D. N. Carney, K. H. Cowan, J. Jolivet et al., *N. Engl. J. Med.* **308** (1982) 199.
41. M. J. Cline, H. Stang, K. Mercola et al., *Nature (London)* **284** (1980) 922 – 925.
42. M. Bar-Eli, H. D. Stang, K. E. Mercola, M. J. Cline, *Somatic Cell Genet.* **9** (1983) 55 – 67.
43. F. Carr, W. D. Medina, S. Dube, J. R. Bertino, *Blood* **62** (1983) 180 – 185.
44. H. Diddens, D. Niethammer, R. C. Jackson, *Cancer Res.* **43** (1983) 5286 – 5292.
45. F. M. Sirotnak, J. I. Degraw, F. A. Schmid, L. J. Goutas et al., *Cancer Chemother. Pharmacol.* **12** (1984) 26 – 30.
46. R. P. Hertzberg, R. K. Johnson, *Annu. Rep. Med. Chem.* **28** (1993) 167 – 176.
47. G. F. Fleming, R. L. Schilsky, *Semin. Oncol.* **19** (1992) 707.
48. K. Y. Shum et al., *J. Clin. Oncol.* **6** (1988) 446.
49. A. L. Lackman et al., *Cancer Res.* **51** (1991) 5579.
50. D. Cunningham, *Cancer Res.* **53** (1991) 810.
51. A. L. Jackman, D. C. Farrugia, W. Gibson, *Eur. J. Cancer* **13A** (1995) 1277 – 1282.
52. M. D. Varney et al., *J. Med. Chem.* **35** (1992) 663.
53. C. Heidelberger, N. K. Chandhavi, P. Dannenberg et al., *Nature (London)* **179** (1957) 663 – 666.
54. R. J. Rutman, A. Cantarow, K. E. Paschkis, *Cancer Res.* **14** (1954) 119 – 126.
55. R. L. Glazer, L. S. Lloyd, *Molec. Pharmacol.* **21** (1982) 468 – 473.
56. D. W. Kufe, P. P. Major, *J. Biol. Chem.* **256** (1981) 9802 – 9806.
57. K. V. Hadjiolova, Z. G. Naydenova, A. A. Hadjiolova, *Biochem. Pharmacol.* **30** (1981) 1861 – 1866.
58. R. I. Glazer, K. D. Hartman, *Molec. Pharmacol.* **23** (1983) 540 – 546.
59. Y. Oshima, M. Itoh, N. Okada, T. Miyata, *Proc. Natl. Acad. Sci. USA* **78** (1981) 4471 – 4474.
60. D. W. Kufe, P. P. Major, E. M. Egan, E. Leh, *J. Biol. Chem.* **256** (1981) 8885 – 8889.
61. P. V. Danenberg, C. Heidelberger, M. A. Mulkins, A. R. Peterson, *Biochem. Biophys. Res. Commun.* **102** (1981) 654 – 659.
62. Y. C. Cheng, K. Nakayama, *Molec. Pharmacol.* **23** (1983) 171 – 174.
63. M. Tanaka, K. Kimura, S. Yoshida, *Cancer Res.* **43** (1983) 5145 – 5150.
64. P. Reichard, O. Skold, G. Klein et al., *Cancer Res.* **22** (1962) 235 – 243.

65. P. Reyes, T. C. Hall, *Biochem. Pharmacol.* **18** (1969) 1587 – 1590.
66. D. K. Kasbeker, D. M. Greenberg, *Cancer Res.* **23** (1963) 818 – 825.
67. P. Reichard, O. Skold, G. Klein, *Nature (London)* **183** (1959) 939 – 941.
68. J. A. Houghton, S. J. Masada, J. O. Phillips et al., *Cancer Res.* **42** (1982) 144 – 149.
69. C. Heidelberger, G. Kaldos, K. L. Mukherjee et al., *Cancer Res.* **20** (1960) 903 – 909.
70. M. M. Lastieboff, B. Kedzioska, W. Rode, *Biochem. Pharmacol.* **32** (1983) 2259 – 2267.
71. A. R. Bapat, C. Zarow, P. V. Dannenberg, *J. Biol. Chem.* **258** (1983) 4130 – 4136.
72. C. C. Rossana, L. G. Rao, L. F. Johnson, *Mol. Cell. Biol.* **2** (1982) 1118 – 1125.
73. F. Baskin, S. C. Carlin, P. Kraus et al., *Molec. Pharmacol.* **11** (1975) 105 – 117.
74. D. G. Priest, B. E. Ledford, M. T. Doig, *Biochem. Pharmacol.* **29** (1980) 1549 – 1553.
75. R. D. Armstrong, E. Cadman, *Cancer Res.* **43** (1983) 2525 – 2528.
76. J. L. Au, Y. M. Rustum, J. Minonda, B. I. Sriva-stava, *Biochem. Pharmacol.* **32** (1983) 541 – 545.
77. Y. M. El Sayed, W. Sadee, *Cancer Res.* **43** (1983) 4039 – 4044.
78. E. J. Ansfield, G. J. Kallas, J. P. Simpson, *J. Clin. Oncol.* **1** (1983) 107 – 110.
79. F. Sorm, A. Piskala, A. Cihak et al., *Experientia* **20** (1964) 202 – 203.
80. A. Piskala, F. Sorm, *Collect. Czech. Chem. Commun.* **29** (1964) 2060 – 2075.
81. P. G. W. Plagemann, M. Behrens, D. Abraham, *Cancer Res.* **38** (1978) 2458 – 2466.
82. J. C. Drake, R. G. Stoller, B. A. Chabner, *Biochem. Pharmacol.* **26** (1977) 64 – 66.
83. T. Lee, M. Karon, R. L. Momparler, *Cancer Res.* **34** (1974) 2481 – 2488.
84. A. Cihak, *Collect. Czech. Chem. Commun.* **39** (1974) 3782 – 3792.
85. J. Doskel, V. Paces, F. Sorm, *Biochim. Biophys. Acta* **145** (1967) 771 – 779.
86. L. H. Li, E. J. Olin, T. J. Fraser et al., *Cancer Res.* **30** (1970) 2770 – 2775.
87. B. F. Vanyosbin, S. G. Tkacheug, A. W. Belozinsky, *Nature (London)* **225** (1976) 948 – 949.
88. J. L. Mandell, P. Chambon, *Nucleic Acids Res.* **7** (1979) 2081 – 2103.
89. J. D. McGhee, G. D. Gineler, *Nature (London)* **280** (1979) 419 – 420.
90. C. Scheir, T. Maniatis, *Proc. Natl. Acad. Sci. USA* **77** (1980) 6634 – 6638.
91. L. H. Van der Ploeg, R. A. Flavill, *Cell* **19** (1980) 947 – 958.
92. O. Niwa, T. Sugarhar, *Proc. Natl. Acad. Sci. USA* **78** (1981) 6290 – 6294.
93. S. Friedman, *Molec. Pharmacol.* **19** (1980) 314 – 320.
94. D. V. Santi, C. E. Garrett, P. J. Barr, *Cell* **25** (1983) 9 – 10.
95. P. G. Constantinides, P. A. Jones, W. Geness, *Nature (London)* **267** (1977) 364 – 366.
96. P. A. Jones, S. M. Taylor, *Cell* **20** (1980) 85 – 93.
97. T. J. Ley, J. DeSimone, N. P. Anagnov et al., *N. Engl. J. Med.* **307** (1982) 1469 – 1475.
98. J. Vesely, A. Chiak, *Pharmacol. Therap.* **2** (1978) 813 – 840.
99. T. Lee, M. R. Kovan, *Biochem. Pharmacol.* **25** (1976) 1737 – 1742.
100. A. Cihak, H. Vessela, F. Sorm, *Biochem. Biophys. Acta* **166** (1968) 277 – 279.
101. A. Cihak, J. Vesely, *Biochem. Pharmacol.* **21** (1972) 3257 – 3265.
102. J. Vesely, A. Cihak, F. Sorm, *Int. J. Cancer* **2** (1967) 639 – 646.
103. J. Doskacil, F. Sorm, *FEBS Lett.* **2** (1974) 30 – 32.
104. D. D. Von Hoff, M. Slavik, F. M. Muggia, *Ann. Intern. Med.* **85** (1976) 237 – 245.
105. Z. H. Israili, W. R. Vogler, E. S. Mingidi et al., *Cancer Res.* **36** (1971) 1453 – 2461.
106. J. A. Beisler, *J. Mol. Chem.* **21** (1978) 204 – 208.
107. J. A. Beisler, M. M. Abbasi, J. S. Driscoll, *Cancer Treat. Rep.* **60** (1976) 1671 – 1674.
108. G. A. Curt, J. A. Kelley, R. L. Fine et al., *Proc. Am. Assoc. Clin. Oncol.* **3** (1984) 37.
109. P. G. W. Plagemann, R. Marz, R. M. Wolhvete, *Cancer Res.* **38** (1978) 978 – 989.
110. A. Iwagaki, T. Nakamura, G. Wakisaka, *Cancer Res.* **29** (1969) 2169 – 2176.
111. M. Y. Chu, G. A. Fisher, *Biochem. Pharmacol.* **11** (1962) 423 – 430.
112. P. P. Major, E. M. Egan, G. P. Beardsley, M. D. Minden et al., *Proc. Natl. Acad. Sci. USA* **78** (1983) 3235 – 3239.
113. P. P. Major, L. Sargent, E. M. Egan, D. W. Kufe, *Biochem. Pharmacol.* **30** (1981) 2221 – 2224.
114. P. P. Major, E. M. Egan, D. Herrick, D. W. Kufe, *Biochem. Pharmacol.* **31** (1982) 861 – 866.
115. M. Y. Chu, G. A. Fisher, *Biochem. Pharmacol.* **14** (1965) 333 – 341.
116. M. R. Atkinson, M. Deutscher, A. Kornberg, A. Russel et al., *Biochemistry* **8** (1969) 4897 – 4904.

117. P. A. Diskwel, F. Wanka, *Biochim. Biophys. Acta* **520** (1978) 461 – 471.
118. R. J. Frann, C. M. Egan, D. W. Kufe, *Leuk. Res.* **7** (1983) 243 – 249.
119. J. F. Ward, E. I. Jones, W. F. Blakely, *Cancer Res.* **44** (1984) 59 – 63.
120. S. R. Bähring-Kuhlmey, *Drugs of Today* **13** (1977) 475.
121. A. A. Claesen et al., *Tetrahedron Lett.* (1966) 3499.
122. P. Major et al., *Proc. Natl. Acad. Sci. USA* **78** (1981) 3235.
123. R. Preston, *Treat. Carcinogen Mutagen* **1** (1980) 147.
124. P. Drahovsky, W. Kreis, *Biochem. Pharmacol.* **19** (1970) 940 – 944.
125. W. Kreis, J. Graham, L. A. Damin, *Biochem. Pharmacol.* **31** (1982) 3831 – 3837.
126. J. Balzarini, E. De Cleireg, *Mol. Pharmacol.* **23** (1983) 175 – 181.
127. A. W. Harris, E. C. Reynolds, L. R. Finch, *Cancer Res.* **39** (1979) 538 – 541.
128. Y. Rustum, H. Priesler, *Cancer Res.* **39** (1979) 42 – 49.
129. C. D. Stuart, P. J. Burke, *Nat. New Biol.* **233** (1971) 109 – 110.
130. P. Chang, P. H. Wiernik, S. D. Reich et al.: "Prediction of Response to Cytosine Arabinoside and Daunorubicin in Acute Nonlymphocytic Leukemia," in F. Mandelli (ed.): *Therapy of Acute Leukemias,* Lombardo-Editore, Rome 1977, pp. 148 – 159.
131. J. F. Smyth, A. B. Robins, C. L. Leese, *Eur. J. Cancer* **12** (1976) 567 – 573.
132. M. N. H. Tattersall, U. K. Ganeshaguru, A. V. Hoffbrand, *Br. J. Haematol.* **27** (1974) 39 – 46.
133. M. Aoshima et al., *Cancer Res.* **37** (1977) 1481.
134. F. Cabaninas, *Drugs of the Future* **5** (1980) 12, 603.
135. T. Ueda, T. Nakamura, S. Ando et al., *Cancer Res.* **43** (1983) 3412 – 3416.
136. K. Hoie, T. Tsuro, K. Naganuma, S. Tsukagoshi et al., *Cancer Res.* **44** (1984) 172 – 177.
137. C. W. Young, R. Schneider, B. Leyland-Jones et al., *Cancer Res.* **43** (1983) 5006 – 5009.
138. Y. C. Cheng, R. S. Tan, J. L. Ruth, G. Dutschman, *Biochem. Pharmacol.* **32** (1983) 726 – 729.
139. L. W. Hertel, J. S. Kroin, J. W. Misner, J. M. Tustin, *J. Org. Chem.* **53** (1988) 11, 2406 – 2409.
140. Eli Lilly, EP 122707, 1984 (L. W. Hertel).
141. Eli Lilly, EP 184365, 1986 (G. B. Grindey, L. W. Hertel).
142. Eli Lilly, EP 306190, 1989T. S. Chou, P. Heath). L. E. Patterson).
143. S. Chubb et al., *Proc. Am. Assoc. Cancer Res.* **28** (1987) Abst. 1282.
144. D. Y. Bouffard, L. F. Momparler, R. L. Momparler, *Eur. J. Pharmacol.* **183** (1990) 2, Abst. 032.
145. V. Heinemann, L. W. Hertel, G. B. Grindley, W. Phukett, *Cancer Res.* **48** (1988) 14, 4024 – 4031.
146. M. N. Serradel, J. Castañer, *Drugs of the Future* **15** (1990) 8, 794 – 797.
147. V. Heinemann et al., *Proc. Am. Assoc. Cancer Res.* **30** (1989) Abst 2204.
148. H. H. Hansen, *Ann. Oncol.* **120** (1994) 519 – 528.
149. E. Beutler, *Lancet* **340** (1992) 8825, 952.
150. H. M. Bryson, E. M. Sorkin, *Drugs* **46** (1993) 5, 872.
151. E. H. Estey et al., *Blood* **79** (1992) 4, 882.
152. S. P. Smith, *Hosp. Formul.* **28** (1993) 7, 621.
153. C. P. Robinson, *Drugs Today* **29** (1993) 6, 379.
154. US Dept. Of Health US 4357324, 1982 (J. A. Montgomery, A. T. Shortnacy).
155. R. W. Brockman, Y. C. Cheng, F. M. Schabel, J. M. Montgomery, *Cancer Res.* **40** (1980) 3610 – 3615.
156. J. A. Montgomery, *Cancer Res.* **42** (1982) 3911 – 3917.
157. W. C. Tseng et al., *Mol. Pharmacol.* **21** (1982) 474 – 477.
158. S. J. Hopkins, *Drugs of the Future* **10** (1985) 1, 20.
159. *Annu. Drug Data Rep.* (1992) 633 – 634.
160. J. S. Whelan et al., *Brit. J. Cancer* **64** (1991) 1, 120.
161. H. S. Hochster et al., *J. Clin. Oncol.* **10** (1992) 1, 28.
162. H. W. Dion et al., *Ann. NY Acad. Sci.* **284** (1977) 21 – 29.
163. Parke Davis, US 3923785, 1975 (A. Ryder et al.).
164. P. W. Woo et al., *J. Heterocycl. Chem.* **11** (1974) 4, 641 – 643.
165. J. F. Smith et al., *Cancer Chemother. Pharmacol.* **1** (1978) 49 – 51.
166. J. F. Smith et al., *Proc. Am. Assoc. Cancer Res.* **20** (1979).
167. G. A. Le Page et al., *Cancer Res.* **36** (1976) 1481 – 1485.
168. M. N. Seradel, J. Castañer, *Drugs of the Future* **6** (1981) 419 – 420.
169. M. Blick et al., *Am. J. Hematol.* **33** (1990) 3, 205 – 209.

170. S. Bernhard et al., *Med. Pediatr. Oncol.* **19** (1991) 4, 276.
171. A. D. Ho et al., *J. Natl. Canver Inst.* **82** (1990) 17, 1416 – 1420.
172. E. H. Kraut et al., *J. Clin. Oncol.* **7** (1989) 2, 168 – 172.
173. R. S. Witte et al., *Invest. New Drug* **10** (1992) 1,49.
174. B. J. Kane, J. G. Kuhn, M. K. Roush, *Ann. Pharmacother.* **26** (1992) 939 – 947.
175. G. B. Elion, E. Burgi, G. H. Hitchings, *J. Am. Chem. Soc.* **74** (1952) 411.
176. G. B. Elion, G. H. Hitchings, *J. Am. Chem. Soc.* **77** (1955) 1676.
177. J. H. Burchenal, M. L. Murphy, R. R. Ellison et al., *Blood* **8** (1953) 965.
178. G. H. Hitchings, G. B. Elion, *Ann. N.Y. Acad. Sci.* **60** (1954) 195.
179. R. W. Brockman, C. S. Debavadi, P. Stutts, D. J. Hutchison, *J. Biol. Chem.* **236** (1961) 1471 – 1479.
180. J. D. Davidson, *Cancer Res.* **20** (1960) 225 – 232.
181. G. B. Elion, *Fed. Proc.* **26** (1967) 898 – 904.
182. J. F. Henderson, M. K. Y. Khoo, *J. Biol. Chem.* **240** (1965) 2349 – 2357.
183. F. F. Snyder, J. F. Henderson, *Can. J. Biochem.* **51** (1973) 943 – 948.
184. P. C. L. Wong, J. F. Henderson, *Biochem. J.* **129** (1972) 1085 – 1094.
185. T. Fields, L. Brox, *Can. J. Biochem.* **52** (1974) 441 – 446.
186. C. D. Green, D. W. Martin Jr., *Proc. Natl. Acad. Sci. USA* **70** (1973) 3698 – 3702.
187. G. B. Elion, S. Callahan, W. Rundles, G. H. Hitchings, *Cancer Res.* **23** (1963) 1207 – 1217.
188. G. H. Hitchings, *Cancer Res.* **23** (1963) 1218 – 1225.
189. T. Nakamura, *Blood & Vessel (Jpn.)* **2** (1971) 237 – 248.
190. T. Higuchi, T. Nakamura, G. Wakisaka, *Blood &Vessel (Jpn.)* **3** (1972) 313 – 318.
191. P. W. Allan, L. L. Bennett Jr., *Biochem. Pharmacol.* **20** (1971) 847 – 852.
192. A. Hampton, *J. Biol. Chem.* **238** (1963) 3068 – 3074.
193. R. J. McCollister, W. R. Gilbert, D. M. Ashton, J. B. Wyngaarden, *J. Biol. Chem.* **239** (1964) 1560 – 1563.
194. R. P. Miech, R. E. Parks Jr., J. H. Anderson Jr., A. C. Sartorelli, *Biochem. Pharmacol.* **16** (1967) 2222 – 2227.
195. A. R. P. Paterson, D. M. Tidd: "6-Thiopurines," in A. C. Sartorelli, D. G. Johns (eds.): *Handbook of Experimental Pharmacology,* vol. 38, Springer-Verlag, Berlin 1975, Part 2, pp. 384 – 403.
196. D. L. Hill, L. L. Bennett Jr., *Biochemistry* **8** (1969) 122.
197. R. T. Kavahata, L. F. Chuang, C. A. Holmberg, B. I. Osburn et al., *Cancer Res.* **43** (1983) 3655 – 3659.
198. J. Maybaum, H. G. Mandel, *Cancer Res.* **43** (1983) 3852 – 3856.
199. J. S. Lazo, K. M. Huang, A. C. Sartorelli, *Cancer Res.* **37** (1977) 4250.
200. C. K. Carrico, A. C. Sartorelli, *Cancer Res.* **37** (1977) 1868.
201. K. Kim, W. J. Blechman, V. G. H. Riddle, A. B. Pardee, *Cancer Res.* **41** (1981) 4529.
202. J. D. Strobel-Stevens, S. M. El Dareer, M. W. Trader, D. L. Hill, *Biochem. Pharmacol.* **31** (1982) 3133.
203. M. Rosman, H. E. Williams, *Cancer Res.* **33** (1973) 1202.
204. M. K. Wolpert, S. P. Damle, J. E. Brown et al., *Cancer Res.* **31** (1971) 1620.
205. M. Rosman, M. L. Lee, W. A. Creasey et al., *Cancer Res.* **34** (1974) 1952.
206. E. M. Scholar, P. Calabresi, *Biochem. Pharmacol.* **28** (1979) 445.
207. L. Tterlikkis, J. L. Day, D. A. Brown, E. C. Schroeder, *Cancer Res.* **43** (1983) 1675 – 1679.
208. S. Zimm, J. M. Collins, D. O'Neill, B. A. Chabner et al., *Clin. Pharmacol. Therap.* **34** (1983) 810 – 817.
209. S. Zimm, J. M. Collins, R. Riccardi, D. O'Neill et al., *N. Engl. J. Med.* **308** (1983) 1005 – 1009.
210. R. D. Armstrong, R. Vera, P. Snyder, E. Cadman, *Biochem. Biophys. Res. Commun.* **109** (1982) 595.
211. M. Inomata, F. Fukuoka, A. Hoshi, K. Kuretani et al., *J. Pharmacobiodyn.* **4** (1981) 928.
212. D. M. Tidd, I. Gibson, P. D. G. Dean, *Cancer Res.* **42** (1982) 3769.
213. M. Inaba, M. Fukui, N. Yoshida, S. Tsukagoshi et al., *Cancer Res.* **42** (1982) 1103.
214. C. P. J. Adair, H. J. Bragg, *Ann. Surg.* **93** (1931) 190.
215. L. P. Jacobson, C. L. Spurr, E. S. O. Barron et al., *J. Am. Med. Assoc.* **132** (1946) 126.
216. M. Colvin: "The Alkylating Agents," in B. A. Chabner (ed.): *Pharmacologic Principles of Cancer Treatment,* W. B. Saunders, Philadelphia 1982, pp. 276 – 308.

217. N. O. Goldstein, R. J. Rutman, *Cancer Res.* **24** (1964) 1363.
218. R. W. Ruddon, J. M. Johnson, *Mol. Pharmacol.* **4** (1968) 258.
219. G. P. Wheeler, J. A. Alexander, *Cancer Res.* **29** (1969) 98.
220. P. Brookes, P. D. Lawley, *Biochem. J.* **80** (1961) 486.
221. C. C. Price, G. M. Gaucher, P. Koneru et al., *Biochim. Biophys. Acta* **166** (1968) 327.
222. A. Loveless, W. C. J. Ross, *Nature (London)* **166** (1950) 1113.
223. K. W. Kohn, L. C. Erickson, R. A. G. Ewig et al., *Biochemistry* **15** (1976) 4629.
224. R. J. Rutman, E. H. C. Chun, F. A. Lewis, *Biochem. Biophys. Res. Commun.* **32** (1968) 650.
225. I. Hirono, *Nature (London)* **186** (1960) 1059.
226. G. Calcutt, T. A. Connors, *Biochem. Pharmacol.* **12** (1963) 839.
227. R. A. G. Ewig, K. W. Kohn, *Cancer Res.* **37** (1977) 2114.
228. L. C. Ericson, G. Laurent, N. A. Sharkey et al., *Nature (London)* **288** (1980) 727 – 729.
229. D. T. Vistica, J. N. Toal, M. Rabinowitz, *Biochem. Pharmacol.* **27** (1978) 2865.
230. W. R. Redwood, M. Colvin, *Cancer Res.* **40** (1980) 1144.
231. W. P. Brade, J. Engel, *Drugs Today* **20** (1984) 491 – 496.
232. P. J. Creaven, L. M. Allen, D. A. Alford et al., *Clin. Pharmacol. Therap.* **16** (1974) 77.
233. T. A. Conners: "Alkylating Agents, Nitrosourea, and Alkyltriazines," in H. M. Pinedo, B. A. Chabner (eds.): *Cancer Chemotherapy Annual 5,* Elsevier, Amsterdam 1983, pp. 30 – 65.
234. A. Mittelman et al., *J. Urol. (Baltimore)* **115** (1976) 409.
235. W. A. Skinner, H. F. Gram, M. O. Green et al., *J. Med. Pharmaceut. Chem.* **2** (1960) 299.
236. K. A. Hyde, E. Acton, W. A. Skinner et al., *J. Med. Pharmaceut. Chem.* **5** (1962) 1.
237. F. M. Schabel, T. P. Johnston, G. S. McGaleb et al., *Cancer Res.* **23** (1963) 226.
238. K. W. Kohn, *Cancer Res.* **37** (1977) 1450.
239. R. J. Weinkam, D. F. Deen, *Cancer Res.* **42** (1982) 1008.
240. J. Mendel, R. Thust, H. Schwartz, *Arch. Geschwulstforsch.* **52** (1982) 371.
241. B. J. Bowdon, J. Grimsley, H. H. Lloyd, *Cancer Res.* **34** (1974) 194.
242. L. C. Panasci, D. Green, R. Nagourney et al., *Cancer Res.* **37** (1977) 2615.
243. J. M. Dornish, I. Smith-Kielland, *FEBS Lett.* **139** (1981) 41.
244. K. D. Tew, A. L. Wang, *Molec. Pharmacol.* **21** (1982) 729.
245. A. Begleiter, H. Y. P. Lam, G. J. Goldenberg, *Cancer Res.* **37** (1977) 1022.
246. N. G. Goldstein, R. J. Rutman, *Cancer Res.* **24** (1964) 1363.
247. R. W. Ruddon, J. M. Johnson, *Molec. Pharmacol.* **4** (1968) 258.
248. J. J. Roberts, T. P. Brent, A. R. Crathorn, *Eur. J. Cancer* **7** (1971) 515.
249. W. P. Tong, D. B. Ludlam, *Biochim. Biophys. Acta* **608** (1980) 174.
250. C. C. Price, G. M. Gaucher, P. Koneru et al., *Biochim. Biophys. Acta* **166** (1968) 327.
251. P. Brookes, P. D. Lawley, *Biochem. J.* **80** (1961) 486.
252. C. B. Thomas, R. Osieka, K. W. Kohn, *Cancer Res.* **38** (1978) 2448.
253. C. C. Erickson, M. O. Bradley, S. M. Ducore et al., *Proc. Natl. Acad. Sci. USA* **77** (1980) 467.
254. R. A. G. Ewig, K. W. Kohn, *Cancer Res.* **38** (1978) 3197.
255. W. G. Verly, Y. Paquette, *Cancer J. Biochem.* **50** (1972) 217.
256. L. C. Ericson, G. Laurent, N. A. Sharkey et al., *Nature (London)* **288** (1980) 727 – 729.
257. A. R. Crathorne, J. J. Roberts, *Nature (London)* **211** (1966) 150.
258. K. Tsujihara, M. Ozeki, T. Morikawa, M. Kawamori et al., *J. Med. Chem.* **25** (1982) 441.
259. Y. Akaike, Y. Arai, H. Takuchi, H. Satoh, *Gann* **73** (1982) 480.
260. W. J. Zeller, M. Berger, G. Eisenbrand, W. Tang et al., *Arzneim.-Forsch.* **32** (1982) 484.
261. J. R. Prous, *Drug News Perspect* **1** (1988) 1, 35.
262. ADIR, FR 2536075, 1984 (C. Cudennec, G. Lavielle).
263. M. N. Serradel, J. Castañer, *Drugs of the Future* **14** (1989) 11, 1042 – 1046.
264. S. Filippeschi et al., *Anticancer Res.* **14** (1988) 11, 1351 – 1354.
265. D. Khayat et al., *Cancer Res.* **47** (1987) 6782 – 6785.
266. J. A. Boutin et al., *Eur. Cancer Oncol.* (1989).
267. P. Deloffre, C. A. Cudennec, G. Lavielle, J. P. Bizzari, *15th Int. Cong. Chemother.,* Istanbul 1987, Abst. 755.
268. J. Jacquillat et al., *Proc. Am. Assoc. Cancer Res.* **30** (1989) Abst. 1088.
269. P. Zeller, H. Gutmann, B. Hegedus et al., *Experientia* **19** (1963) 129.
270. V. T. DeVita, A. Serpick, P. Carbone, *Ann. Intern. Med.* **73** (1970) 542.

271. J. Gutterman, A. Huang, P. Hochstein, *Proc. Soc. Exp. Biol. Med.* **130** (1979) 797.
272. R. J. Weinkam, D. A. Shiba, *Life Sci.* **22** (1978) 937.
273. D. L. Dunn, R. A. Lubet, R. A. Prough, *Cancer Res.* **39** (1979) 4555.
274. D. E. Schwartz, W. Bollag, P. Obrecht, *Arzneim.- Forsch.* **17** (1967) 1389.
275. D. Reed, F. Dost, *Proc. Am. Assoc. Cancer Res.* **7** (1965) 57.
276. M. Baggiolini, B. Dewald, H. Aebi, *Biochem. Pharmacol.* **18** (1969) 2187.
277. C. Pueyo, *Mutat. Res.* **67** (1979) 189.
278. E. Therman, *Cancer Res.* **32** (1972) 1111.
279. W. Kreis, *Proc. Am. Assoc. Cancer Res.* **7** (1966) 39.
280. A. Sartorelli, S. Tsunamura, *Proc. Am. Assoc. Cancer Res.* **6** (1965) 55.
281. H. Lam, A. Begleiter, W. Stein et al., *Biochem. Pharmacol.* **27** (1978) 1883.
282. J. L. Skibba, D. D. Beal, G. Ramirez et al., *Cancer Res.* **30** (1970) 147.
283. P. Farina, A. Gescher, J. A. Hickman, J. K. Horton et al., *Biochem. Pharmacol.* **31** (1982) 1887.
284. T. L. Loo, G. E. Housholder, A. H. Gerulath et al., *Cancer Treat. Rep.* **60** (1976) 149.
285. D. D. Beal, J. L. Skibba, K. K. Whitnable et al., *Cancer Res.* **36** (1976) 2827.
286. A. H. Gerulath, T. L. Loo, *Biochem. Pharmacol.* **21** (1972) 2335.
287. D. W. Kaiser, I. T. Thurston, J. R. Dudley et al., *J. Am. Chem. Soc.* **73** (1951) 2984.
288. J. F. Worzalla, B. D. Kaima, B. M. Johnson et al., *Cancer Res.* **34** (1974) 2669.
289. J. F. Worzalla, B. D. Kaima, B. M. Johnson et al., *Cancer Res.* **33** (1973) 2810.
290. B. C. V. Mitchley, S. A. Clarke, T. A. Connors et al., *Cancer Treat. Rep.* **61** (1977) 3.
291. M. Ames, G. Powis, J. S. Kovach et al., *Cancer Res.* **39** (1979) 5016.
292. L. M. Lake, E. E. Grunden, B. M. Johnson, *Cancer Res.* **35** (1975) 2858.
293. A. J. Cumber, W. C. J. Ross, *Chem. Biol. Interact.* **17** (1977) 349.
294. V. Iver, W. Szybalski, *Science (Washington)* **145** (1964) 55 – 58.
295. L. Littlefield et al., *Mut. Res.* **81** (1981) 377.
296. J. Mac Donald et al., *Ann. Intern. Med.* **93** (1980) 533.
297. *Drugs Today* **20** (1984) 489; **21** (1985) 420.
298. *Drugs of the Future* **8** (1983) 402; **9** (1984) 371; **10** (1985) 420.
299. F. Arcamone et al., *J. Med. Chem.* **18** (1975) 703.
300. Microbiochem. Res. Found., US 4303785, (H. Naganawa, T. Takenchi, H. Umezawa).
301. H. Umezawa et al., *J. Antibiot.* **32** (1979) 1082 – 1085.
302. M. N. Serradel, J. Castañer, *Drugs of the Future* **8** (1983) 7, 610 – 611.
303. T. Nishimura et al., *J. Antibiot.* **33** (1980) 737 – 743.
304. D. Dantchev et al., *J. Antibiot.* **32** (1979) 1085 – 1086.
305. H. Brockmann, *Fortschr. Chem. Org. Naturst.* **50** (1973) 121.
306. C. E. Myers: "Anthracyclines," in B. A. Chabner (ed.): *Pharmacologic Principles of Cancer Treatment,* W. B. Saunders, Philadelphia 1982, pp. 416 –434.
307. R. Phillips, A. DiMarco, F. Zunino, *Eur. J. Biochem.* **85** (1978) 487.
308. R. B. Painter, *Cancer Res.* **38** (1978) 4445.
309. G. Daskal, C. Woodard, S. T. Crooke et al., *Cancer Res.* **38** (1978) 467.
310. W. E. Ross, L. A. Zwelling, K. W. Kohn, *Int. J. Radiat. Oncol., Biol. Phys.* **5** (1979) 1221.
311. B. Sinha, R. H. Sik, *Biochem. Pharmacol.* **29** (1980) 1867.
312. K. Handa, S. Sato, *Gann* **66** (1975) 43.
313. N. R. Bachur, S. L. Gordon, M. V. Gee, *Cancer Res.* **38** (1977) 1745.
314. R. Ogura, H. Toyama, T. Shimada et al., *J. Appl. Biochem.* **1** (1979) 325.
315. J. H. Doroshow, G. Y. Locker, C. E. Myers, *J. Clin. Invest.* **65** (1980) 128.
316. N. W. Revis, N. Marusic, *J. Mol. Cell. Cardiol.* **10** (1978) 945.
317. T. R. Tritton, S. A. Murphree, A. C. Sartorelli, *Biochem. Biophys. Res. Commun.* **84** (1978) 802.
318. T. R. Tritton, G. Yeh, *Science* **217** (1982) 248.
319. E. Goormaghtigh, R. Brasseur, J. M. Ruysschaert, *Biochem. Biophys. Res. Commun.* **104** (1982) 314.
320. C. E. Myers, C. Simone, L. Gianni et al., *Proc. Am. Assoc. Cancer Res.* **22** (1981) 112.
321. T. Oki, *Jpn. J. Antibiot.* **30** (1977) 570.
322. R. K. Y. Zee-Cheng, C. C. Cheng, *J. Med. Chem.* **21** (1978) 291.
323. L. H. Patterson, B. M. Gandecka, J. R. Brown, *Biochem. Biophys. Res. Commun.* **110** (1983) 399.
324. E. D. Kharasch, R. F. Novak, *Biochem. Biophys. Res. Commun.* **108** (1982) 1346.
325. S. S. Legha, *Drugs Today* **20** (1984) 12, 629 – 638.

326. *Drugs of the Future* **5** (1980) 234; **6** (1981) 316; **7** (1982) 357; **8** (1983) 64; **9** (1984) 380; **10** (1985) 429.
327. American Cyanamid, USP 4197249, 1980 (F. E. Durr, K. C. Murdock).
328. R. K.-Y. Zee-Cheng, C. C. Cheng, *J. Med.-Chem.* **22** (1979) 1024.
329. K. C. Murdock et al., *J. Med. Chem.* **22** (1979) 9, 1024 – 1030.
330. F. I. Durr, R. E. Wallace, R. V. Citarella, *Cancer Treat Rev.* **10** (1983) 3 – 11.
331. R. E. Wallace, K. C. Murdock, R. B. Angier, F. E. Durr, *Cancer Res.* **39** (1979) 1570 – 1574.
332. R. K. Johnson et al., *Cancer Treat. Rep.* **63** (1979) 425 – 439.
333. D. D. Von Hoff, C. A. Coltuan, B. Forseth, *Cancer Res.* **41** (1981) 1853 – 1855.
334. B. M. Henderson et al., *Cancer Treat. Rep.* **66** (1982) 1139 – 1143.
335. B. M. Sparano, G. Gordon, C. Hall, M. J. Iatropoulos, J. F. Noble, *Cancer Treat. Rep.* **66** (1982) 1145 – 1158.
336. K. C. Anderson et al., *Cancer Treat. Rep.* **67** (1983) 435 – 438.
337. F. C. Schell et al., *Cancer Treat. Rep.* **66** (1982) 1641 – 1643.
338. C. B. Pratt et al., *Cancer Treat. Rep.* **67** (1983) 85 – 88.
339. D. V. Unverferth et al., *Cancer Treat. Rep.* **67** (1983) 343 – 350.
340. M. S. Aapro, D. S. Alberts, *Invest. New Drugs* **2** (1984) 329 – 330.
341. S. A. Taylor, B. T. Tranum, D. D. Von Hoff, J. J. Constanzi, *Invest. New Drugs* **3** (1985) 67 – 69.
342. J. A. Elliott et al., *Anti-Cancer Drug Res.* **3** (1989) 4, 271 – 282.
343. R. F. Marschke et al., *Med. Pediat. Oncol.* **16** (1988) 4, 269 – 270.
344. K. Hillier, *Drugs of the Future* **6** (1981) 12, 762; **7** (1982) 896.
345. K. C. Murdock et al., *J. Med. Chem.* **25** (1982) 505.
346. American Cyanamid, EP 338372, 1989 (V. J. Lee, K. C. Murdock)
347. American Cyanamid, US 4900838, 1990 (K. C. Murdock)
348. K. C. Murdock et al., *J. Med. Chem.* **36** (1993) 15, 2098.
349. US Dept. Of Health, USP 25157, 1981 (H. Dubicki, J. L. Parsons, F. W. Starks).
350. B. F. Cain et al., *J. Med. Chem.* **18** (1975) 1110.
351. *Annu. Drug Data Report* (1981) 177.
352. *Drugs of the Future* **5** (1980) 277; **8** (1983) 535.
353. L. Steinherz et al., *Cancer Treat. Rep.* **66** (1982) 483.
354. J. Kolwas, *Drugs Today* **15** (1979) 200.
355. G. A. Curt, N. J. Clendeninn, B. A. Chabner, *Cancer Treat. Rep.* **68** (1984) 87 – 99.
356. T. Tsuruo, H. Lida, M. Nojiri, *Cancer Res.* **43** (1983) 2905.
357. S. Waksman, H. B. Woodruff, *Proc. Soc. Exp. Biol. Med.* **45** (1940) 609 – 611.
358. H. M. Sobell, S. C. Jam, T. D. Sakore et al., *Nat. New Biol.* **231** (1971) 200.
359. F. Takusagawa et al.: *12th Congress of the IUCR Associated Meeting on Molecular Structure and Biologic Activity,* Buffalo, New York, August 26 – 28, 1981, p. 55.
360. E. Reich, R. M. Franklin, A. J. Shatkin et al., *Proc. Natl. Acad. Sci. USA* **48** (1982) 1238.
361. H. S. Schwartz, E. Godoyien, R. Y. Ambaye, *Cancer Res.* **287** (1968) 192.
362. M. N. Goldstein, K. Hamm, E. Amrod, *Science* **151** (1966) 1555 – 1556.
363. H. Umezawa, K. Meada, T. Takeuchi et al., *J. Antibiot. Ser. A.* **19** (1966) 200 – 206.
364. B. T. Sikic, Z. H. Siddik, T. E. Gram, *Cancer Treat. Rep.* **64** (1980) 659 – 667.
365. T. Takita, Y. Muraoka, H. Umezawa: "Bleomycin and Peplomycin," in H. M. Pinedo, B. A. Chabner (eds.): *Cancer Chemotherapy Annual 6,* Elsevier, Amsterdam 1984, pp. 85 – 90.
366. A. Hagiwara et al., *Anti-Cancer Drug Res.* **2** (1988) 319 – 324.
367. S. Saito, Y. Umezawa, T. Yoshioka et al., *J. Antibiot. (Tokyo)* **36** (1983) 92 – 95.
368. J. Fugimito, H. Higashi, G. Kosaki, *Cancer Res.* **36** (1976) 2248 – 2251.
369. M. Chien, A. P. Grollman, S. B. Horwitz, *Biochemistry* **16** (1977) 3641 – 3645.
370. W. J. Caspay, C. Niziak, D. A. Lanzo et al., *Mol. Pharmacol.* **16** (1979) 256 – 260.
371. S. Biables, J. R. Warr, *Genet. Res.* **34** (1979) 269 – 279.
372. M. Mayaki, T. Ono, S. Hori et al., *Cancer Res.* **35** (1975) 2015 – 2018.
373. S. Akiyama, M. Kuwano, *J. Cell Physiol.* **107** (1981) 147 – 153.
374. S. Akiyama, K. Ikezaki, H. Kunamochi et al., *Biochem. Biophys. Res. Commun.* **101** (1981) 55 – 60.
375. W. E. G. Mueller, M. Geisort, R. K. Zahn et al., *Eur. J. Cancer Clin. Oncol.* **19** (1983) 665 – 669.
376. P. G. Sorensen, M. Dorth, H. H. Hansen, *Eur. J. Cancer Clin. Oncol.* **19** (1983) 319 – 323.

377. R. K. Johnson, R. P. Hertzberg, *Annu. Rep. Med. Chem.* **25** (1990) 129.
378. P. A. Aristoff et al., *Invest. New Drugs* **7** (1989) 364.
379. J. P. Mc Govren et al., *Invest. New Drugs* **7** (1989) 448.
380. M. A. Mitchell, P. D. Johnson, M. G. Williams, P. A. Aristoff, *J. Am. Chem. Soc.* **11** (1989) 6428.
381. K. Goni et al., *Jpn. J. Cancer Res.* **83** (1992) 113.
382. D. L. Boger, M. S. S. Palanki, *J. Am. Chem. Soc.* **114** (1992) 9318.
383. M. Fontana et al., *Anti-Cancer Drug Res.* **7** (1992) 131.
384. D. Volpe et al., *Invest. New Drugs* **10** (1992) 255.
385. N. Zein, M. Poncin, R. Nilakantan, G. A. Ellestad, *Science* **244** (1989) 697.
386. B. H. Long et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 2.
387. K. C. Nicolaou, *Science* **256** (1992) 1172.
388. K. C. Nicolaou, W. M. Dai, *J. Am. Chem. Soc.* **114** (1992) 8908.
389. W. A. Bleyer, S. A. Frisby, V. T. Oliverio, *Biochem. Pharmacol.* **24** (1979) 633 – 639.
390. R. A. Bender, W. D. Kornreich, *Proc. Am. Assoc. Cancer Res.* **22** (1981) 227.
391. H. Mujagic, S. S. Chen, R. Geist et al., *Cancer Res.* **43** (1983) 3591 – 3597.
392. D. V. Jackson, R. A. Bender, *Cancer Res.* **39** (1979) 4341 – 4349.
393. R. Schrek, *Am. J. Clin. Path.* **62** (1974) 1 – 7.
394. R. A. Bender, W. A. Bleyer, S. A. Frisby et al., *Cancer Res.* **35** (1975) 1305 – 1308.
395. R. F. Zager, S. A. Frisby, V. T. Oliverio, *Cancer Res.* **33** (1973) 1670 – 1676.
396. I. S. Johnson, H. F. Wright, G. H. Svoboda et al., *Cancer Res.* **20** (1980) 1016 – 1022.
397. G. A. Curt, B. D. Bailey, H. Mujagic et al., *Proc. Am. Assoc. Cancer Res.* **25** (1984) 337.
398. W. T. Beck, M. C. Cirtain, J. L. Lefko, *Cancer Treat. Rep.* **67** (1983) 875 – 879.
399. R. B. Sklaroff, D. Straus, C. Young, *Cancer Treat. Rep.* **63** (1979) 793 – 794.
400. G. Mathe, J. L. Misset, F. deVassal et al., *Cancer Treat. Rep.* **62** (1978) 805.
401. M. Bayssas, J. Gouveia, F. deVassal et al., *Cancer Res.* **74** (1980) 91 – 99.
402. R. Z. Andriamialisoa, N. Langlois, Y. Langlois, P. Potier, *Tetrahedron* **36** (1980) 3053 – 3060.
403. M. N. Serradel, J. Castañer, *Drugs of the Future* **11** (1986) 7, 575 – 577.
404. M. R. Paintrand, I. Pignot, *J. Electron. Microsc.* **32** (1983) 2, 115 – 124.
405. P. Mangeney et al., *J. Org. Chem.* **44** (1979) 3765 – 3768.
406. P. Maral, C. Bourut, E. Chenn, G. Mathe, *Cancer Chemother. Pharmacol.* **5** (1981) 3, 197 – 199.
407. R. Maral, C. Bourut, E. Chenn, G. Mathe, *Cancer Lett.* **22** (1984) 49 – 54.
408. G. Mathe, P. Reizenstein, *Cancer Lett.* **27** (1985) 285 – 293.
409. M. Besenval et al., *Semin. Oncol.* **16** (1989) 2, 37.
410. A. Depierre et al., *Semin. Oncol.* **16** (1989) 2, 26.
411. J. B. Sorensen, *Drugs* **44** (1992) 60.
412. P. B. Schiff, A. S. Kende, S. B. Horwitz, *Biochem. Biophys. Res. Commun.* **85** (1978) 737 – 740.
413. R. A. Bender, B. A. Chabner: "Tubulin Binding Agents: Epipodophyllotoxin," in B. A. Chabner (ed.): *Pharmacologic Principles of Cancer Treatment,* W. B. Saunders, Philadelphia 1982, pp. 263 – 266.
414. J. D. Loike, S. B. Horwitz, *Biochemistry* **15** (1976) 5435 – 5438.
415. D. K. Kalwinsky, A. T. Look, J. Ducore, A. Fridland, *Cancer Res.* **43** (1983) 1592 – 1596.
416. A. J. Wozniak, W. E. Ross, *Cancer Res.* **43** (1983) 130 – 135.
417. A. J. Wozniak, B. S. Glisson, K. R. Hande, W. E. Ross, *Cancer Res.* **44** (1984) 626 – 629.
418. B. H. Long, A. Minocha, *Proc. Am. Assoc. Cancer Res.* **24** (1983) 321.
419. M. E. Wall, M. C. Wani, C. E. Cook, K. H. Palmer, *J. Am. Chem. Soc.* **88** (1966) 3888.
420. J.-C. Cai, C. R. Hutchinson, *Chem. Heterocyl. Compd.* **25** (1983) 753.
421. Yakult Honsha, EP 137145, 1985 (T. Miyasaka et al.).
422. Smith Kline Beecham Corp., EP 321122, (J. C. Boehm, S. M. Hecht, K. G. Holden, R. K. Johnson, W. D. Kingsbury).
423. W. D. Kingsbury et al., *J. Med. Chem.* **34** (1991) 98 – 107.
424. N. Osheroff, *Pharmacol. Ther.* **41** (1989) 223 – 41.
425. L. Liu, J. Wang, *Proc. Natl. Acad. Sci. USA* **84** (1987) 7024 – 7027.
426. A. J. Ryan, S. Squires, H. L. Strutt, R. T. Johnson, *Nucl. Acid Res.* **19** (1991) 12, 3295 – 3300.
427. M. N. Serradel, J. Castañer, R. M. Castañer, *Drugs of the Future* **12** (1987) 3, 207.
428. C. H. Spiridonis, *Drugs of the Future* **20** (1995) 5, 483 – 489.

429. W. J. Slichenmyer, E. K. Rowinsky, R. C. Donehower, S. H. Kaufmann, *J. Natl. Cancer Inst.* **85** (1993) 4, 271 – 291.
430. C. B. Hendricks et al., *Cancer Res.* **52** (1992) 8, 2268 – 2278.
431. W. K. Eng et al., *Mol. Pharmacol.* **38** (1990) 4, 471 – 480.
432. D. Abigerges et al., *J. Natl. Cancer Inst.* **86** (1994) 446.
433. M. C. Wani et al., *J. Am. Chem. Soc.* **93** (1971) 2325 – 2327.
434. J.-N. Denis et al., *J. Am. Chem. Soc.* **110** (1988) 5917 – 5919.
435. J.-N. Denis, A. Correa, A. E. Greene, *J. Org. Chem.* **55** (1990) 1957 – 1959.
436. M. Hepperle, G. I. Georg, *Drugs of the Future* **19** (1994) 573 – 584.
437. K. C. Nicolaou, W. M. Dai, R. K. Guy, *Angew. Chem.* **106** (1994) 38 – 69.
438. K. C. Nicolaou et al., *Nature* **367** (1994) 630 – 634.
439. R. A. Holton et al., *J. Am. Chem. Soc.* **116** (1994) 1599 – 1600.
440. J. J. Masters et al., *Angew. Chem.* **107** (1995) 1883.
441. P. B. Schiff, S. B. Horwitz, *Proc. Natl. Acad. Sci. USA* **77** (1980) 1561 – 1564.
442. P. B. Schiff, S. B. Horwitz, *J. Cell Biol.* **91** (1981) 479 – 483.
443. M. A. Bissery, D. Guenard, F. Gueritte-Voegelein, F. Lavelle, *Cancer Res.* **51** (1991) 4845 – 4852.
444. I. Ringel, S. B. Horwitz, *J. Natl. Cancer Inst.* **83** (1991) 288 – 291.
445. R. Pazdur et al., *Cancer Treat. Rev.* **19** (1993) 351 – 386.
446. P. B. Schiff, J. Faut, S. B. Horwitz, *Pharmacol. Ther.* **25** (1984) 83 – 124.
447. S. Rao, J. J. Manfredi, S. B. Horwitz, I. Ringel, *J. Natl. Cancer Inst.* **84** (1992) 785 – 788.
448. S. Rao, S. B. Horwitz, I. Ringel, *J. Natl. Cancer Inst.* **84** (1992) 785 – 788.
449. J. Parness, S. B. Horwitz, *J. Cell. Biol.* **91** (1981) 479 – 487.
450. P. B. Schiff, S. B. Horwitz, *Biochemistry* **20** (1981) 3247 – 3252.
451. A. H. Ding, F. Porten, E. Sanchez, C. F. Nathan, *Science* **248** (1990) 370 – 372.
452. C. Bogdan, A. Ding, *J. Leukocyte Biol.* **52** (1992) 119 – 121.
453. M. E. Stearns, M. Wang, *Cancer Res.* **52** (1992) 3776 – 3781.
454. D. M. Vyas et al., *Biorg. Med. Chem. Lett.* **3** (1993) 1357 – 1360.
455. E. K. Rowinsky et al., *Semin. Oncol.* **20** (1993) 3, 1 – 15.
456. H. Lataste et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 4090 – 4094.
457. J. Kant et al., *Biorg. Med. Chem. Lett.* **3** (1993) 2471 – 2474.
458. G. Samaranayake, N. F. Magri, C. Jitransgri, D. G. Kingston *J. Org. Chem.* **56** (1991) 5114 – 5119.
459. A. Datta, L. Jayasinghe, G. I. Georg, *J. Med. Chem.* (1994) 4258 – 4260.
460. J. L. Fabre, D. Lolli-Tonelli, L. H. Spiridonidis, *Drugs of the Future* **20** (1995) 5, 464.
461. *Annual Data Report* (1995) 765.
462. G. Höfle et al., *Angew. Chem.* **108** (1996) 1671; *Int. Ed. Engl.* **35** (1996) 1567.
463. D. M. Bollag et al., *Cancer Res.* **55** (1995) 2325 – 2333.
464. A. Bolag et al., *Angew. Chem.* **108** (1996) 2976; *Int. Ed. Engl.* **35** (1996) 2801.
465. D. Meng et al., *J. Amer. Chem. Soc.* **119** (1997) 2733.
466. K. C. Nicolaou et al., *Angew. Chem.* **108** (1996) 2534.
467. K. C. Nicolaou et al., *Nature* **387** (1997) 268 – 272.
468. D. Schinzer et al., *Angew. Chem.* **109** (1997) 543 – 544.
469. D.-S. Su et al., *Angew. Chem.* **109** (1997) 2178.
470. R. J. Kowalski et al., *J. Biol. Chem.* **272** (1997) 2534.
471. B. Rosenberg, L. Van Camp, T. Krigas, *Nature (London)* **205** (1965) 698 – 700.
472. B. Rosenberg, L. Van Camp, J. E. Troska et al., *Nature (London)* **222** (1969) 385 – 386.
473. A. H. Rossof, R. E. Slayton, C. P. Perlia, *Cancer* **30** (1972) 1451 – 1455.
474. D. J. Higby, H. J. Wallace Jr., J. F. Holland, *Cancer Chemother. Rep.* **57** (1973) 459 – 463.
475. L. H. Einhorn, S. D. Williams, *N. Engl. J. Med.* **300** (1979) 289 – 292.
476. A. W. Prestayko, J. C. D'Aousst, B. F. Issel et al., *Cancer Treat. Rev.* **6** (1979) 17 – 24.
477. W. M. Scovell, T. O'Connor, *J. Am. Chem. Soc.* **99** (1977) 120 – 126.
478. A. F. LeRoy, R. J. Lutz, R. L. Dedrick et al., *Cancer Treat. Rep.* **63** (1979) 59 – 64.
479. G. L. Cohen, W. R. Bauer, J. K. Berton et al., *Science* **203** (1979) 1014 – 1016.
480. R. C. Srivastava, J. Froelich, G. L. Eichhorn, *Biochimie* **60** (1979) 879 – 881.
481. J. J. Roberts, A. J. Thomson, *Prog. Nucleic Acid Res. Mol. Biol.* **22** (1979) 71 – 133.

482. B. Rosenberg, *Naturwissenschaften* **60** (1973) 399 – 408.
483. T. F. Slater, M. Ahmed, S. A. Ibrahim, *J. Clin. Hematol. Oncol.* **7** (1977) 534 – 539.
484. N. E. Madias, J. T. Harrington, *Am. J. Med.* **65** (1978) 307 – 311.
485. D. C. Dobyan, J. Levi, C. Jacobs et al., *J. Pharmacol. Exp. Therap.* **213** (1980) 551 – 556.
486. R. L. Schilsky, T. Anderson, *Ann. Intern. Med.* **90** (1979) 929 – 931.
487. F. A. Hayes, A. A. Green, N. Jenzen et al., *Cancer Treat. Rep.* **63** (1979) 547 – 549.
488. S. Davis, W. Kessler, B. M. Haddad et al., *J. Med.* **11** (1980) 133 – 137.
489. M. Dentino, F. L. Luft, M. N. Yum et al., *Cancer* **41** (1978) 1274 – 1279.
490. K. K. Chang, D. J. Higby, E. S. Henderson et al., *Cancer Treat. Rep.* **61** (1977) 367 – 371.
491. R. F. Ozols, B. Corden, J. Jacobs et al., *Ann. Intern. Med.* **100** (1984) 19 – 24.
492. J. Drobnik, M. Urbankova, A. Krekulova, *Mut. Res.* **17** (1973) 13 – 20.
493. L. A. Zwelling, S. Michaels, H. Schwartz, *Cancer Res.* **41** (1981) 640 – 649.
494. J. Ducore, L. Zwelling, K. Kohn, *Proc. Am. Assoc. Cancer Res.* **21** (1980) 267.
495. L. C. Erikson, L. A. Zwelling, J. M. Ducore, *Cancer Res.* **4** (1981) 2791 – 2794.
496. K. Micetich, S. Michaels, G. Jude et al., *Proc. Am. Assoc. Cancer Res.* **22** (1981) 252.
497. A. Bakka, L. Endresen, A. B. S. Johnson et al., *Toxicol. Appl. Pharmacol.* **61** (1981) 215 – 226.
498. L. R. Beach, R. D. Palmiter, *Proc. Natl. Acad. Sci. USA* **78** (1981) 2110 – 2114.
499. C. R. Wilkenson, P. J. Cox, M. Jones et al., *Biochimie* **60** (1978) 851 – 857.
500. *Drugs of the Future* **8** (1983) 489; **9** (1984) 463; **10** (1985) 497; **11** (1986) 499.
501. *Drugs Today* **22** (1986) 255.
502. Research Corp., DE 2 329 485, 1973 (M. J. Cleare, J. D. Hoeschele, B. Rosenberg).
503. R. C. Harrison et al., *Inorg. Chim. Acta* **46** (1980) 215.
504. A. H. Calvert et al., *Cancer Chemother. Pharmacol.* **9** (1982) 140.
505. B. D. Evans et al., *Proc. Ass. Cancer Res.* **24** (1983) 154.
506. CBDCA, Clinical Brochure, Investigational Drug Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, Md., 1980.
507. B. D. Evans, K. S. Raju, A. H. Calvert, S. U. Harland et al., *Cancer Treat. Rep.* **67** (1983) 997 – 1001.
508. P. J. Creaven, S. Madajewicz, L. Pendyala et al., *Cancer Treat. Rep.* **67** (1983) 795 – 798.
509. C. Sternberg et al., *Cancer Treat. Rep.* **69** (1985) 1305 – 1307.
510. R. Canetta et al., *Cancer Treat. Rep.* **63** (1985)2107 – 2109.
511. I. N. Olver et al., *Cancer Treat. Rep.* **70** (1986) 421 – 422.
512. A. P. Kyriazis et al., *Cancer res.* **45** (1985) 2012 – 2015.
513. K. R. Harrap, M. Jones, C. R. Wilkenson et al.: "Antitumor Toxic and Biochemical Properties of *cis*-Platinum and Eight Other Platinum Analogs," in A. W. Prestayko, S. T. Crooke, S. K. Carter (eds.): *cis-Platinum: Current Status and New Developments (Pap Symposium),* Academic Press, New York 1980, pp. 193 – 212.
514. P. Chang, *Drugs of the Future* **8** (1983) 364; **10** (1985) 7, 561.
515. *Drugs of the Future* **12** (1987) 11, 1029 – 1031.
516. *Annu. Drug Data Rep.* **8** (1986) 6, 590.
517. Shionogi & Co., EP 216362 1987 (H. Kagawa, K. Shima, T. Tsukada).
518. K. Hirabayashi, E. Okada, *Cancer* **71** (1993) 9, 2769.
519. M. Koenuma et al., *Clin. Rep.* **29** (1995) 12, 259.
520. N. Uchida et al., *Clin. Rep.* **29** (1995) 12, 269.
521. I. H. Krakoff, 5th *Intl. Symp. Platinum Cancer Chemother*., Abano Therme, 1987, Abst. L7.
522. B. K. Bhuyan et al., *Cancer Commun.* **3** (1991) 2, 53.
523. ASTA Medica AG, EP 324154 1989 (J. Engel et al.).
524. R. Voegeli, E. Günther, P. Aulenbacher, J. Engel, P. Hilgard, *Drugs of the Future* **17** (1992) 883 – 886.
525. A. M. Otto et al., *Pharm. Pharmacol. Lett.* **1** (1992) 103 – 106.
526. W. A. Denny et al., *J. Med. Chem.* **35** (1992) 2983 – 2987.
527. *Drugs Today* **10** (1974) 74.
528. ICI, GB 1013907 1962.
529. G. R. Bedford, D. N. Richardson, *Nature* **212** (1966) 733.
530. M. N. Serradel, J. Castañer, *Drugs of the Future* **11** (1986) 5, 398 – 400.
531. Farmos Group, EP 95875 (R. J. Tiovola et al.).
532. J. T. Pento et al., *Drugs of the Future* **9** (1984) 7, 5.
533. R. Löser, P.-St. Jamak, K. Seibel, *Drugs of the Future* **9** (1984) 3, 186.

534. S. Kallio et al., *Cancer Chemother. Pharmacol.* **17** (1986) 103 – 108.
535. L. Kangas et al., *Cancer Chemother. Pharmacol.* **17** (1986) 109 – 113.
536. V. C. Jordan, B. Gosden, *Mol. Cell. Endocrinol.* **27** (1982) 27, 921.
537. S. P. Robinson et al., *Eur. J. Cancer Clin. Oncol.* **24** (1988) 12, 1817.
538. R. Valavaara et al., *Eur. J. Cancer Clin. Oncol.* **24** (1988) 4, 785.
539. S. R. Ebbs et al., *Lancet II* (1987) 621.
540. R. F. Kauffmann et al., *J. Pharmacol. Exp. Ther.* **280** (1997) 146 – 153.
541. M. A. Ferreira, M. M. Caramona, L. M. Celeste, *Proc. Br. Pharmacol. Soc.* (1995) Abs. 267 P
542. M. Dukes, *Pharm. J.* **257** (1996) 176.
543. A. Hoshi, *Drugs of the Future* **16** (1991) 3, 217.
544. *Annu. Drug Data Report* **15** (1993) 4, 378.
545. A. Brodie, *J. Steroid Biochem. Mol. Biol.* **40** (1991) 1 – 3, 255.
546. S. A. Wells et al., *Cancer Res.* **Suppl. 42** (1982) 3454.
547. R. D. Burnett, D. N. Kirk, *J. Chem. Soc. Perkin Trans.* **I** (1973) 1830.
548. R. C. Coombes, *Eur. J. Cancer* **28A** (1992) 12, 1941.
549. D. Cunningham et al., *Brit. J. Cancer* **55** (1987) 331.
550. J. H. Davis et al., *Brit. J. Cancer* **66** (1992) 1, 139.
551. M. Dowsett et al., *Cancer Res.* **49** (1989) 1306.
552. M. Dowsett et al., *Eur. J. Cancer* **28** (1992) 2 – 3, 415.
553. R. C. Stein et al., *Cancer Chemother. Pharmacol.* **26** (1990) 75.
554. ICI, US 4935437, 1990 (P. N. Edwards, M. S. Lange).
555. Ciba Geigy AG, US 4617307, 1986 (L. J. Browne).
556. J. Prous, *Drugs of the Future* **19** (1994) 4, 335 – 337.
557. J. Castañer, *Annu. Drug Data Rep.* **13** (1991) 8, 716.
558. A. B. Foster et al., *J. Med. Chem.* **28** (1985) 200.
559. J. T. Pento, *Drugs of the Future* **14** (1989) 9, 843 – 845.
560. J. Prous et al., *Drugs of the Future* **20** (1995) 1, 30 – 32.
561. R. E. Steele et al., *Steroids* **50** (1987) 147.
562. M. Dukes et al., *Proc. Am. Assoc. Cancer Res.* **33** (1992) Abst. 1677.
563. P. V. Plourde et al., *Breast Cancer Res. Treat.* **30** (1994) 1, 103 – 111.
564. K. Schieweck, A. S. Bhatnagar, A. Matter, *Cancer Res.* **48** (1988) 834.
565. K. Schieweck et al., *Proc. Am. Assoc. Cancer Res.* **29** (1988) Abst. 968.
566. M. Dowsett et al., *Breast Cancer Res. Treat.* **27** (1993) 1 – 2, Abst. 87.
567. P. V. Plourde et al., *Proc. Am. Soc. Clin. Oncol.* **12** (1993) Abst. 165.
568. C. Huggins, C. V. Hodges, *Cancer Res.* **1** (1941) 293.
569. J. A. Smith, *J. Urol.* **137** (1987) 1 – 10.
570. J. Trachtenberg, *J. Urol.* **132** (1984) 61.
571. G. H. Rasmusson et al., *J. Med. Chem.* **27** (1984) 1690 – 1701.
572. M. R. Robinson, B. S. Thomas, *Brit. Med. J.* **4** (1971) 391 – 394.
573. F. Neumann, *J. Steroid Biochem.* **19** (1983) 391 – 402.
574. B. J. Furr, *Clin. Oncol.* **2** (1988) 581 – 590.
575. H. Tucker, *Drugs of the Future* **15** (1990) 3, 255 – 265.
576. *Drugs Today* **20** (1984) 296.
577. *Drugs of the Future* **1** (1976) 108; **8** (1983) 270.
578. R. O. Neri, E. A. Peets, *J. Steroid Biochem.* **6** (1975) 815 – 819.
579. W. I. P. Mainwaring, F. R. Mangan, P. A. Feherty, M. Freifeld, *Med. Cell. Endocrinol.* **1** (1974) 113 – 128.
580. S. Liao, D. K. Howell, T. M. Chang, *Endocrinology* **94** (1974) 1205 – 1209.
581. B. Katchen, S. Buxbaum, *J. Clin. Endocrinol. Metab.* **41** (1975) 373 – 379.
582. J. Geller et al., *Prostate* **2** (1981) 309 – 314.
583. E. A. Peets, M. F. Henson, R. O. Neri, *Endocrinology* **94** (1974) 532 – 540.
584. R. Neri, N. Kassem, *Prog. Cancer Res. Ther.* **31** (1984) 507 – 518.
585. F. Labrie et al., *Oncol.* **2** (1988) 597 – 619.
586. F. Labrie, A. Dupont, M. Ciguere, *Brit. J. Urol.* **61** (1988) 341 – 346.
587. A. Belanger, A. Dupont, F. Labrie, *J. Clin. Endocrinol. Metab.* **59** (1984) 422 – 426.
588. J.-P. Raynaud, C. Bonne et al., *Prostate* **5** (1984) 299 – 311.
589. T. Ojasoo, *Drugs of the Future* **12** (1987) 763 – 770.
590. L. Proulx, F. Labrie, *Prostate* **5** (1984) Abst. 429.
591. F. Labrie, A. Dupont, A. Belanger, *Prostate* **4** (1983) 579.
592. C. Harnois, A. Dupont, F. Labrie, *Brit. J. Opthalmol.* **70** (1986) 471 – 473.

593. H. Tucker, G. J. Chesterson, *J. Med. Chem.* **31** (1988) 885 – 887.
594. H. Tucker, J. W. Crook, G. J. Chesterson, *J. Med. Chem.* **31** (1988) 954 – 959.
595. B. J. A. Furr et al., *J. Endocrinol.* **113** (1987) R7 – 9.
596. S. N. Freeman, W. I. P. Mainwaring, B. J. A. Furr, *J. Endocrinol.* (1986) Suppl. III, 155.
597. C. J. Tyrell, *Prostate* (Suppl. 4) **20** (1992) 97.
598. G. Wilding et al., *Proc. Am. Assoc. Clin. Oncol.* **10** (1991) Abst. 593.
599. A. V. Schally, S. M. McCann, *Fertil. Steril.* **64** (1995) 452 – 453.
600. A. V. Schally, A. Arimura, A. J. Kastin, *Science* **173** (1971) 1036 – 1038.
601. H. Matsuo, Y. Baba, M. Nair, A. Arimura, A. V. Schally, *Biochem. Biophys. Res. Commun.* **43** (1971) 1334 – 1339.
602. K. Amoss et al., *Biochem. Biophys. Res. Commun.* **44** (1971) 205 – 210.
603. Y. Baba, H. Matsuo, A. V. Schally, *Biochem. Biophys. Res. Commun.* **44** (1971) 459 – 463.
604. G. Emons, A. V. Schally, *Hum. Reprod.* **9** (1994) 1364 – 1370.
605. P. M. Conn, W. Crowley, *New Engl. J. Med.* **324** (1991) 93 – 103.
606. G. F. Weinbauer, E. Nieschlag in K. Höffgen (ed.): *Peptides in Oncology* Springer, Heidelberg, 1992, pp. 113 – 136.
607. A. S. Dutta, *Drugs of the Future* **13** (1988) 43 – 57.
608. M. Filicori, C. Flamingi, *Drugs* **35** (1988) 63 – 82.
609. J. J. Nestor, B. H. Vickery, *Annu. Rep. Med. Chem.* **23** (1988) 211 – 220.
610. M. T. Goulet, *Annu. Rep. Med. Chem.* **30** (1995) 169 – 178.
611. T. Reissmann et al., *Hum. Reprod.* **20** (1995) 1974 – 1981.
612. M. J. Karten in W. F. Crowley, P. M. Conn (eds.): *Modes of Action of GnRH and GnRH Analogs,* Springer, Heidelberg 1992, pp. 277 – 297.
613. G. Flouret et al., *Pept. Sci.* **1** (1995) 89 – 105.
614. P. M. Conn, W. F. Crowley, *Annu. Rev. Med.* **45** (1994) 391.
615. A. V. Schally in J. F. Holand et al. (eds): *Cancer Medicine* 3rd ed., Lee & Febiger, Philadelphia, PA 1993, pp. 827 – 840.
616. A. V. Schally in P. Belfort, J. Pinotti, T. K. Eskes (eds.): *Advances in Gynecology and Obstetrics Vol. 6,* Parthenon, Cornforth 1989, pp. 3 – 22; b) Scrip 22, 1995, 2066.
617. A. Friedrich, G. Jaeger, K. Radscheit, R. Uhmann, *Pept. Proc. Eur. Pept. Symp. 22*nd 1992/1993, 47 – 49.
618. B. Kutscher et al., *Angew. Chem. Int. Ed. Engl.* **36** (1997) 2148 – 2161.
619. R. L. Barbieri, *Trends Endocrinol. Metab.* **3** (1992) 30 – 34.
620. R. Schmidt, K. Sundaram, R. B. Thau, C. W. Badrin, *Contraception* **29** (1984) 283 – 289.
621. S. Bajusz et al., *Int. J. Pept. Protein Res.* **32** (1988) 425 – 435.
622. J. Pinski et al., *Int. J. Pept. Protein Res.* **45** (1995) 410 – 417.
623. A. Kleemann et al., *Proc. Akabori Conf. Ger. Jpn. Symp. Pept. Chem. 4*th 1991, 96 – 101; b) F. R. Kunz, T. Müller, K. Drauz, *Proc. Akabori. Conf. Ger. Jpn. Symp. Pept. Chem. 5*th 1994, 15 – 16.
624. A. Ljugquist et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 8236 – 8240.
625. J. Leal et al., *Drugs of the Future* **16** (1991) 529 – 537.
626. A. Janecka, T. Janecki, C. Bowers, K. Folkers, *J. Med. Chem.* **37** (1994) 2238 – 2241.
627. A. Janecka et al., *Med. Chem. Res.* **1** (1991) 306 – 311.
628. A. Janecka, T. Janecki, C. Bowers, K. Folkers, *Int. J. Pept. Protein Res.* **44** (1994) 19 – 23.
629. P. Theobald et al., *J. Med. Chem.* **34** (1991) 2395 – 2402.
630. J. Rivier et al., *J. Med. Chem.* **35** (1992) 4270 – 4278.
631. J. E. Rivier et al., *J. Med. Chem.* **38** (1995) 2649 – 2662.
632. K. Stoeckemann, J. Sandow, *J. Cancer Res. Clin. Oncol.* **119** (1993) 457 – 462.
633. R. Deghenghi, F. Boutignon, P. Wüthrich V. Lenaerts, *Biomed. Pharmacother.* **47** (1993) 107 – 110.
634. J. Nester, et al., *J. Med. Chem.* **35** (1992) 3942 – 3948.
635. J. Mulzer in E. Ottow, U. Schöllkopf, B. G. Schulz (eds.): *Stereoselective Synthesis* , Springer, Heidelberg, 1994, pp. 37 – 61.
636. J. Mulzer et al., *Angew. Chem.* **106** (1994) 1813 – 1815; *Angew. Chem. Int. Ed. Engl.* **33** (1994) 1737 – 1739.
637. F. Haviv et al., *J. Med. Chem.* **32** (1989) 2340 – 2344.
638. F. Haviv et al., *J. Med. Chem.* **36** (1993) 928 – 933.
639. Abbott Laboratories, PCT/US 95/02410, 1995 (F. Haviv).
640. Abbott Laboratories, WO 95/04540, 1995 (F. Haviv).
641. Abbott Laboratories, WO 94/14841, 1994 (F. Haviv).

642. Abbott Laboratories, WO 94/13313, 1994 (J. Greer).
643. Tap Pharmaceuticals, US-A 5300492, 1994 (F. Haviv).
644. F. Haviv et al., *J. Med. Chem.* **37** (1994) 701 – 707.
645. T. Beckers, K. Marheineke, H. Reiländer, P. Hilgard, *Eur. J. Biochem.* **231** (1995) 535 – 543.
646. R. P. Millar, C. A. Flanagan, R. C. Milton, J. A. King, *J. Biol. Chem.* **264** (1989) 21007 – 21013.
647. G. A. McPherson, *J. Pharmacol. Methods* **14** (1985) 213 – 228.
648. J. Trachtenberg, A. Pont, *Lancet* **2** (1984) 433 – 435.
649. S. Bhasin et al., *Endocrinology* **118** (1986) 1229 – 1232.
650. B. De et al., *J. Med. Chem.* **32** (1989) 2036 – 2038.
651. Abbott Laboratories, US-A 4992421, 1991 (B. De).
652. McNeillab Inc., US-A 4678784, 1987 (C. Y. Ho).
653. Takeda Chemical Industries, WO 95/28405, 1995 (S. Furuya).
654. Takeda Chemical Industries, WO 96/34012A1, 1996 (C. Kitada).
655. Merck, WO 97/21435, 1997 (M. Goulet).
656. Takeda Chemical Industries, WO 97/14697, 1997 (S. Furuya).
657. C. Unger, *Drugs of the Future* **22** (1997) 12, 1337 – 1345.
658. G. Powis, *Pharmacol. Ther.* **62** (1994) 57 – 95.
659. W. J. Fantl, D. E. Johnson, L. T. Williams, *Annu. rev. Biochem.* **62** (1993) 453 – 481.
660. F. McCormick, *Nature* **363** (1993) 15 – 16.
661. C. A. Lang-Carter et al., *Science* **260** (1993) 315 – 319.
662. D. Rea et al., *Proc. 7th NCI-EORTC Symp. On New Drugs in Cancer Ther.* Amsterdam (1992), p. 62.
663. D. Ingber et al., *Nature* **348** (1990) 555.
664. T. Ozsuka et al., *J. Antibiotics* **45** (1992) 348.
665. Y. Reiss et al., *Cell* **62** (1990) 81 – 88.
666. S. Omura, D. van der Pyl, *Cell* **46** (1993) 222.
667. S. M. Sebti, A. D. Hamilton, *Drug Discovery Technol.* **3** (1998) 26 – 33.
668. A. Levitski, A. Gazit, *Science* **267** (1995) 1782 – 1788.
669. A. Levitski, *Eur. J. Biochem.* **226** (1994) 1 – 13.
670. R. T. Abraham, M. Aquarone, A. Anderson, *Biol. Cell.* **83** (1995) 105.
671. D. J. Hanahan, *Ann. Rev. Biochem.* **55** (1986) 483 – 509.
672. W. E. Berdel, *Onkologie* **13** (1990) 245 – 250.
673. W. E. Berdel et al., *Anticancer Res.* **1** (1981) 345 – 352.
674. G. Rodriguez et al., *Proc. Am. Assoc. Cancer Res.* **33** (1992) 262.
675. C. Unger, H. Eibl, *Lipids* **26** (1991) 1412.
676. P. Hilgard, J. Engel, *Drugs Today* **Suppl. B.,** (1994) 30.
677. P. Hilgard et al., *Cancer Chemother. Pharmacol.* **32** (1993) 90 – 95.
678. P. Hilgard, J. Stekar, C. Unger, *Proc. Annu. Meet. Am. Assoc. Cancer Res.* **31** (1990) A2457.
679. J. Engel et al., *Drugs of the Future* **13** (1988) 10, 948 – 951.
680. C. Geilen et al., *Eur. J. Cancer* **27** (1991) 12, 1650 – 1653.
681. R. Hass et al., *Cancer Res.* **52** (1992) 1445 – 1450.
682. R. Becher et al., *Onkologie* **16** (1993) 1, 11.

Candles

See also → Waxes, Chap. 4.7.

FRANZ WILLHÖFT, Rosengarten-Nenndorf, Federal Republic of Germany

RUDOLF HORN, Hamburg, Federal Republic of Germany page 2

Records show that candles were first used around the eighth century B.C.

Until about 1850, candles were made out of raw [*8006-40-4*] or bleached [*8012-89-3*] beeswax and animal fats (beef tallow). Spermaceti [*8002-23-1*], paraffin [*64742-43-4*], and stearin [*11099-07-3*] have since been added to the list of raw materials.

Raw Materials. The raw materials used in the production of candles can be classified as basic materials and additives.

Basic materials: paraffin, beeswax, stearin
Additives: microcrystalline waxes, hard waxes, opacifiers, coloring agents, and other such additives as polyethylene, resins, and perfumes

Refined paraffin, obtained from mineral oil and having a low oil content, is the principal raw material used today in the fabrication of candles (→ Waxes). It is a mixture of normal paraffins, isoparaffins, and cycloparaffins (naphthenes). The percentage of each component in the mixture depends on the type of crude oil and on the extent of refining prior to wax isolation. The *n*-paraffins shrink severely on solidification, and at temperatures below the solidification point they form relatively large, uniform crystals. The isoparaffins contract less on solidification than the *n*-paraffins and form smaller, less uniform crystals. The cycloparaffins, which can be composed of one or more rings and of branched or unbranched chains, usually solidify in an amorphous mass. Only a few form small, needle-shaped crystals.

Beeswax (→ Waxes) is a mixture which contains approximately 12 % *n*-paraffins in addition to hydrocarbon waxes, free fatty acids, cerotic acid [*506-46-7*], cerotic esters, and cholesteryl esters. It is plastic andhighly malleable, but it does not smear. It is hard and brittle under cold-conditions, an indication that the softening point and the melting point lie far apart on the temperature scale. Beeswax is used to produce wax sheets which are then processed into candle decorations and altar candles.

Stearin is used in pure form or in mixtures for candle making. It contains primarily palmitic acid [*57-10-3*] and stearic acid [*57-11-4*] (C_{16} and C_{18} fatty acids) and gives the candle an opaque, white, and very full appearance.

Microcrystalline waxes are hydrocarbon waxes with a very fine crystalline structure.

Hard waxes (→ Waxes) are either natural, particularly carnauba wax [*8015-86-9*], or synthetic products.Polymethylene [*68649-50-3*] (synthesized using the Fischer-Tropschtechnique) or polyethylene [*9002-88-4*] (obtained using the Zieglersynthesis) mainly are used. When added to the raw materials, they give candlesan opaque appearance, and in trace amounts they give a good hardness.

Opacifiers include stearin, hard waxes, polyethylenes, and, above all, organic and inorganic pigments. Air, blown into the wax material, also acts as an opacifier.

Candle material can be *colored* by adding either 0.1 – 0.8 % of an oil- or paraffin-based dye or about 0.5 – 2.0 % of a pigment to the basic materials.

Wick. The wick is made of braided cotton and can be either flat or round. Usually raw yarn is employed, hence the wick subsequently must be washed, bleached, prepared, spun, and dried. Ammonium salts (e.g., ammonium nitrate, ammonium phosphate, and ammonium sulfate), potassium chloride, potassium nitrate, borax, and boric acid prevent the wick from glowing after the candle has been put out. They regulate the state of the wick, ensure complete burning, and act as oxygen carriers. The wick controls the melting, evaporation, and burning of the candle

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_029

material. It conveys the liquid wax from the cup to the burning zone. The capillarity of the wick should be sufficient to enable the liquid wax to rise fast enough and yet not stream down the sides of the candle.

Manufacturing Technique. Although substantially improved models of the well-known candle molding or candle drawing machines are still available, a powder press (one- to eight-stanchion press) is normally used. The paraffin powder is so intensely compressed in a cylinder that at a compression of approximately 1.7 : 1 a solid, rough candle is formed. Methods used in the extrusion of plastics have also been applied to the processing of wax; in no other method used in the manufacture of candles are the demands on the raw materials so critical. The granules are pressed through a cylinder by means of a screw conveyer and subsequently forced through a perforated disc into a slightly constrictive compression chamber. The end of the machine is joined to a small, heated, cylindrical tube from which a ribbon of the still malleable candle continuously issues.

In the latest method employing a horizontal piston press, the granules (powder, chunks, beads) are fed into the press and compressed with each cycle of a hydraulic piston. The cycle time and throughput are variable. The candle appears discontinuously through the bore that determines its diameter.

The wick can also be dipped into the wax melt using special machines and the process repeated until the candle acquires the required diameter. The appearance of rough candles produced in this manner can subsequently be improved by dipping them into a coloring bath.

Economic Aspects. The production of candles in the Federal Republic of Germany increased from 8 kt in 1955 to 46.5 kt in 1983. The per capita consumption in 1983 was almost 1 kg.

The candle production in Western Europe in 1979 amounted to 165 kt; per capita consumption was about 430 g. The production in the United States in 1979 was 100 kt.

References

General References

1. M. Faraday: *The Chemical History of a Candle,* Chatto & Windus, London 1908.

Candy → Confectionery

Caprolactam

JOSEF RITZ, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 – 8)
HUGO FUCHS, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 – 8)
HEINZ KIECZKA, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 9)
WILLIAM C. MORAN, BASF Corporation, Freeport, Texas, United States (Chaps. 2 – 8)

1. History 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 2
4.1. Processes via Cyclohexanone 3
4.1.1. Cyclohexanone Oxime 3
4.1.2. Beckmann Rearrangement to Caprolactam 4
4.1.3. Improvement of the Oximation Process 6
4.1.4. Gas-Phase or Catalytic Rearrangement of Cyclohexanone Oxime 8
4.1.5. Modified Oximation and Beckmann Rearrangement 9
4.2. Photooximation 9
4.3. Production from Toluene 11
4.4. Recovery from Nylon-6 Waste 13
4.5. Former Commercial Routes 13
4.6. New Routes in Development 14
4.7. Environmental Protection 16
5. Quality Specifications 16
6. Analysis 17
7. Storage and Transportation 18
8. Economic Aspects 18
9. Toxicology 19
10. References 19

1. History

ε-Caprolactam (2-oxohexamethylenimine, hexahydro-1*H*-azepin-2-one) has been known since the 19th century. S. GABRIEL and T. A. MAAS synthesized caprolactam in 1899 by cyclization of ε-aminocaproic acid. O. WALLACH synthesized caprolactam by Beckmann rearrangement of cyclohexanone oxime. Commercial interest increased in 1938, when P. SCHLACK of IG Farbenindustrie produced the first spinnable polymer by polycondensation of caprolactam. Since then, caprolactam has gained importance. Large-scale industrial production has increased rapidly. In 1989, world production reached ca. 2×10^6 t, in 1994 ca. 2.2×10^6 t [8].

Table 1. Solubility of caprolactam (wt %)

Solvent	Solubilities at			
	20	30	40	50 °C
Water	82	86.5	90	93.5
Toluene	26	36.5	51	66.5
Ethyl acetate	24.2	33.3	48.5	66.2
Methyl ethyl ketone	34.6	45.7	59.2	72.9
Cyclohexanone	34.6	42.2	54.5	68.2
Cyclohexane	2	2.5	7	18.5

2. Physical Properties

Caprolactam [*105-60-2*], $C_6H_{11}ON$, is a white, hygroscopic, crystalline solid with a characteristic odor.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_031

Its basic properties are summarized below:

M_r	113.16
mp, °C	69.2
bp, °C	
at 101.3 kPa	268.5
at 6.7 kPa	174
at 1.3 kPa	134
at 0.4 kPa	111
Density, kg/L	
at 120 °C	0.9829
at 100 °C	0.9983
at 80 °C	1.0135
Viscosity, mPa · s	
at 120 °C	2.93
at 100 °C	4.87
at 80 °C	8.82
Specific heat, kJ kg^{-1} K^{-1}	
at 150 °C	2.345
at 80 °C	2.135
Heat of fusion, kJ/kg	123.5
Heat of polycondensation, kJ/kg	138
Heat of vaporization, kJ/kg	
at 268 °C	481
at 168 °C	574
at 105 °C	628
Vapor pressure, kPa	
at 268 °C	101.3
at 168 °C	5.3
at 105 °C	0.25
Flash point, °C	139.5
Ignition temperature, °C	375
Lower explosion limit	1.4 vol% at 135 °C
Upper explosion limit	8.0 vol% at 180.5 °C
Thermal conductivity coefficient, kJ m^{-1} h^{-1} K^{-1}, at 76 – 183 °C	0.5
Coefficient of volume expansion, K^{-1}, at 80 – 90 °C	0.00104

Caprolactam is soluble in polar and aromatic solvents and slightly soluble in high molecular mass aliphatic hydrocarbons. The solubility of caprolactam in various solvents is summarized in Table 1.

3. Chemical Properties

Polymerization is caprolactam's most important chemical property. The ring is hydrolyzed at 260 – 270 °C. Linear polymer chains are formed by polycondensation [9]. Caprolactam also reacts directly by polyaddition with the polymer chains. These reactions lead to an equilibrium between the polymer and caprolactam favoring a 90 % conversion to polymer. Caprolactam can also be polymerized by anionic polymerization at low moisture contents, preferably less than 100 ppm. A catalyst and cocatalyst system is necessary. The reaction temperature is lower than in hydrolytic polymerization.

Caprolactam can undergo all cyclic amide reactions, such as oxidation, hydrolysis, *N*-alkylation, phosphogenation, and nitration.

Molten caprolactam absorbs atmospheric oxygen. Oxygen reacts at 75 °C with caprolactam to produce very small amounts of peroxide. Above 100 °C adipic acid imide is formed [10].

This reaction is catalyzed by traces of heavy metals.

Caprolactam is heat stable in the absence of oxygen [11].

Caprolactam is quantitatively converted to ε-aminocaproic acid by hydrolysis with aqueous acids or alkalis [12]. Treatment with gaseous methanol in the presence of a dehydrating catalyst gives *N*-methyl-ε-caprolactam, a versatile solvent [13]. Caprolactam reacts with phosgene to give chloroformic acid ε-caprolactim ester. Nitration gives nitrocaprolactam, and subsequent reduction amino-caprolactam. L-Lysine is formed from amino-caprolactam by hydrolysis and resolution of the racemate [14]. Reactions with other chemical reagents are of no industrial importance.

4. Production

Many methods for caprolactam production have been developed. All commercial processes are based on benzene or toluene from BTX extract streams. Figure 1 summarizes the various processes.

Large-scale industrial processes are, without exception, multistage processes in which ammonium sulfate and sometimes organic compounds are formed as byproducts.

Cost-efficiency depends a great deal on the byproducts. Large quantities of ammonium sulfate are undesirable in the manufacture of caprolactam. The profitability of a process is dependent on this.

Large-scale industrial processes employ cyclohexanone, cyclohexane, or toluene as starting materials.

About 90 % of the caprolactam is produced by using the conventional cyclohexanone process. Cyclohexanone is obtained by catalytic oxidation of cyclohexane with air, or by hydro-

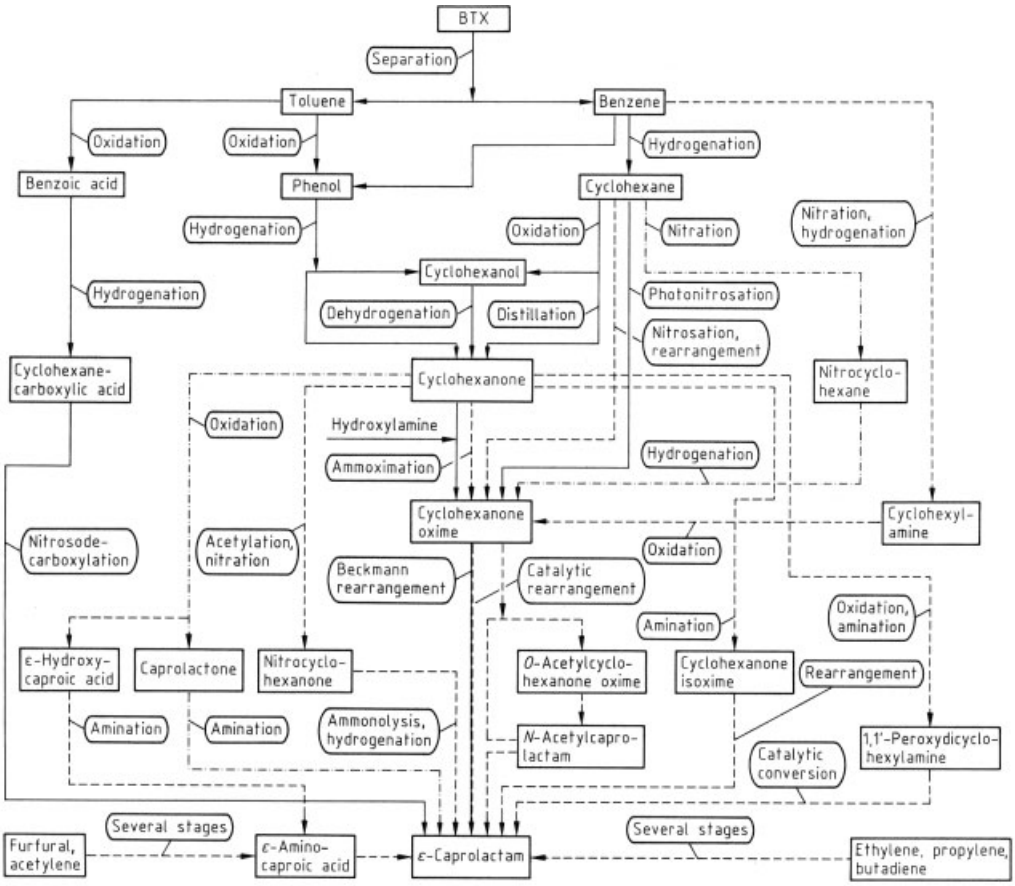

Figure 1. Routes to caprolactam
—- routes in commercial use; —· —· routes once in commercial use; - - - - - - routes in laboratory development

genation of phenol and dehydrogenation of the cyclohexanol byproduct. The conversion of cyclohexanone to cyclohexanone oxime followed by Beckmann rearrangement gives caprolactam. The yield from cyclohexanone approaches 98 %.

About 10 % of caprolactam is produced by photonitrosation of cyclohexane or by nitrosation of cyclohexanecarboxylic acid in the presence of sulfuric acid; in the latter process cyclohexanecarboxylic acid is produced from toluene.

4.1. Processes via Cyclohexanone

4.1.1. Cyclohexanone Oxime

In caprolactam processes using cyclohexanone, water containing cyclohexanone oxime is formed and isolated as an intermediate.

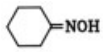

Physical Properties. Cyclohexanone oxime is a white crystal with a slightly pungent odor. Typical properties are listed below:

M_r	113.16
mp, °C	
water content 0 wt %	89.4
water content 3 wt %	74.6
water content 5 wt %	68.4
water content 7 wt %	65.7
bp, °C	
at 101.3 kPa	209.8
at 6.7 kPa	130
at 1.3 kPa	96
at 0.4 kPa	74
Density, kg/L	
at 100 °C	0.969
at 90 °C	0.976

Viscosity, mPa · s	
at 100 °C	4.34
at 90 °C	6.11
Specific heat, kJ kg^{-1} K^{-1}	
at 90 – 120 °C	1.766
Heat of fusion, kJ/kg	98.8
Heat of rearrangement, kJ/kg	1815
Heat of vaporization, kJ/kg	
at 204 °C	490.3
at 168 °C	499.5
at 105 °C	540
Vapor pressure, kPa	
at 204 °C	101.3
at 168 °C	28.0
at 105 °C	2.0
Flash point, °C	100
Ignition temperature, °C	285

Cyclohexanone oxime can be distilled without decomposition only under reduced pressure. For unknown reasons, violent decomposition may occur [15]. Traces of acid can initiate spontaneous rearrangement.

Chemical Properties. The Beckmann rearrangement to caprolactam is the most important industrially used reaction. The catalytic reduction of cyclohexanone oxime with hydrogen gives cyclohexylhydroxylamine and cyclohexylamine [15]. Hydrolysis with mineral acids gives cyclohexanone and the corresponding hydroxylamine salt.

Production. The reaction of cyclohexanone with hydroxylamine is the best known method for the production of cyclohexanone oxime.

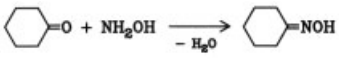

Hydroxylamine is used in its sulfate or phosphate form.

Some manufacturers produce hydroxylamine by the modified Raschig process using ammonium salts as starting materials. Oximation is then carried out with aqueous hydroxylammonium sulfate solution containing ammonium sulfate. The sulfuric acid liberated is neutralized with ammonia to form ammonium sulfate. The amount of ammonium sulfate formed during oximation is about 2.7 tons per ton of cyclohexanone oxime [16].

The BASF process (Fig. 2) for the production of hydroxylammonium sulfate solution generates about 0.7 tons of ammonium sulfate per ton of cyclohexanone oxime during the oximation step. A similar process has been developed by Inventa [17]. BASF and Inventa obtain the hydroxylammonium sulfate solution by hydrogenation of nitric oxide over a platinum catalyst in the presence of dilute sulfuric acid. The hydroxylammonium sulfate solution is reacted with cyclohexanone and ammonia in an oxime reactor.

This reaction is conducted with thorough mixing at 85 – 90 °C in a weak acidic solution.

Cyclohexanone oxime is obtained as a moist melt. It is separated from the aqueous ammonium sulfate solution in a separator drum. The purified aqueous ammonium sulfate solution is free of substances that would be undesirable in fertilizer- or technical-grade salt uses.

4.1.2. Beckmann Rearrangement to Caprolactam

All caprolactam manufacturers use sulfuric acid or oleum as a rearrangement medium. The rearrangement is more complete in concentrated sulfuric acid. Excess sulfur trioxide further increases the speed of the rearrangement [18].

The rearrangement, which includes opening of the cyclohexyl ring, is a very rapid, highly exothermic reaction.

Since the Beckmann rearrangement is highly exothermic, molten cyclohexanone oxime and concentrated oleum (27 %) are introduced simultaneously into a relatively large amount of already rearranged product.

Molar ratios of cyclohexanone oxime and oleum may range from 1 to 1.05 [19]. Sulfur trioxide binds the water in the moist cyclohexanone oxime to form sulfuric acid and catalyzes the rearrangement. Cooling reduces the heat of reaction. The reaction gives the sulfate of caprolactam in excess sulfuric acid. It is hydrolyzed by neutralization with ammonia or ammonia water. The reactions take place according to the following equations:

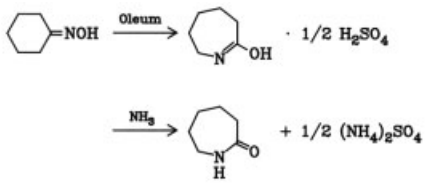

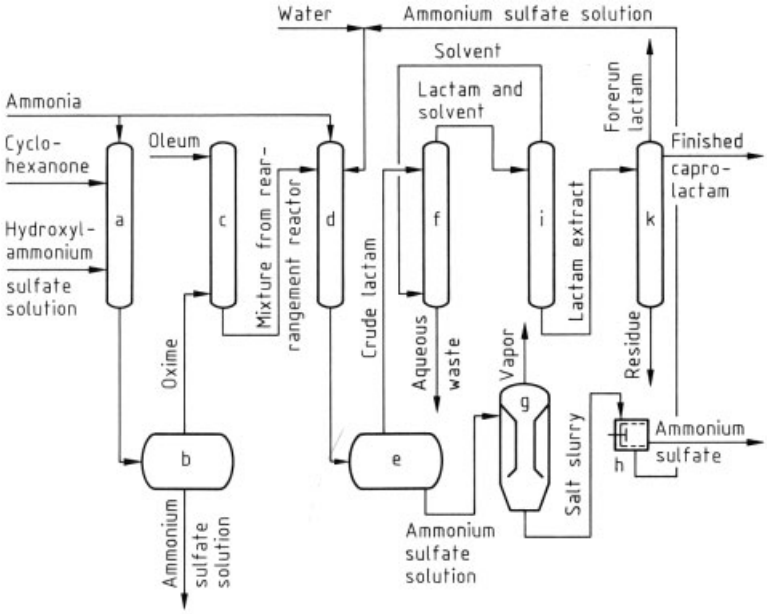

Figure 2. BASF caprolactam production
a) Oximation; b) Oxime separation; c) Rearrangement; d) Neutralization; e) Crude-lactam separation; f) Extraction; g) Crystallization; h) Centrifuge; i) Solvent distillation; k) Lactam distillation

The rearrangement mixture is removed from the reactor at the rate it is formed. It is neutralized in ammonia and water to prevent precipitation of solid ammonium sulfate. The reaction product is then separated from the lighter phase above the concentrated ammonium sulfate solution as 70 % aqueous crude caprolactam. The crude caprolactam is removed in a separator drum and fed to an extractor.

Caprolactam is extracted with solvents such as benzene, toluene [20], or chlorinated hydrocarbons. The virtually caprolactam-free aqueous solution is discharged into a waste treatment plant or incinerated.

Concentrated ammonium sulfate solution is further evaporated and crystallized.

The caprolactam-containing solvent is fed into a distillation column. Solvent-free caprolactam is further distilled under reduced pressure. Very pure caprolactam is obtained in molten form. It is either used as such or solidified in a flaker. A total of 4.5 tons of ammonium sulfate per ton of caprolactam is obtained from both oximation and neutralization with the Raschig process. The BASF or Inventa process produces a total of 2.5 tons of ammonium sulfate per ton of caprolactam in oximation and neutralization.

Caprolactam yields can be improved and undesirable byproduct formation reduced by multistage rearrangement and the adjustment of reaction conditions in each stage. Acid concentration, free sulfur trioxide concentration, amount of oxime added, reaction temperature, and degree of mixing are critical variables at each stage [21].

Some processes differ from the BASF process with regard to purification of crude caprolactam.

In the Allied Chemical process, solvent extraction is followed by crystallization from aqueous caprolactam solution.

In the DSM process, the crude caprolactam is first extracted with a solvent and then reextracted with water. Before being distilled, the caprolactam extract is subjected to a number of physicochemical purifications, such as hydrogenation and ion exchange treatment [17].

Polimex developed a different approach by combining crystallization from aqueous caprolactam solution and oxidizing treatment. This

physical and chemical purification process is followed by distillation [22].

Inventa licenses caprolactam plants that utilize crystallization as a final purification process.

4.1.3. Improvement of the Oximation Process

Oximation processes which form no or little ammonium sulfate have been developed by DSM [23], [24] and BASF [25]. These processes comprise a combination of hydroxylamine and oxime manufacture.

Hydroxylamine–Phosphate Oxime Process (HPO Process of DSM; Fig. 3) The process proceeds according to the following equation:

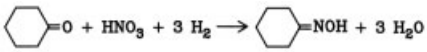

Oximation is conducted in a hydroxylamine phosphoric acid buffer solution. After separation of the oxime, the remaining aqueous ammonium phosphate buffer solution is recycled to hydroxylamine synthesis and concentrated. The reactions involved are described below:

1) Reduction of the phosphoric acid/ammonium nitrate buffer solution with hydrogen, and formation of hydroxylammonium phosphate:

$$NH_4NO_3 + 2\ H_3PO_4 + 3\ H_2 \longrightarrow$$
$$[NH_3OH]^+\ [H_2PO_4]^- + NH_4H_2PO_4 + 2\ H_2O$$

Palladium on graphite or alumina carrier is used as a catalyst.

2) Oxime formation:

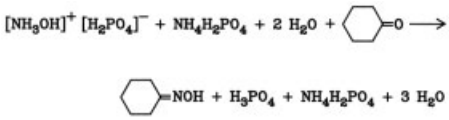

3) After the oxime has been separated, the nitrate ions consumed are replaced by addition of 60 % nitric acid.

$$H_3PO_4 + NH_4H_2PO_4 + 3\ H_2O + HNO_3 \longrightarrow$$
$$2\ H_3PO_4 + NH_4NO_3 + 3\ H_2O$$

The hydroxylamine formed by catalytic hydrogenation of nitrate ions reacts with free phosphoric acid in the buffer solution at pH 1.8 to form a hydroxylammonium phosphate solution. The reaction takes place in special columns; the unreacted hydrogen is separated from the catalyst suspension in a separator, and is recycled to the reaction via a compressor.

After the catalyst has been filtered and recycled, the hydroxylamine buffer solution reacts with cyclohexanone in the oximation reactor to produce cyclohexanone oxime. Toluene is used as the solvent and phosphoric acid is liberated.

Oximation takes place in a cascade of mixers and separators with a countercurrent process at pH 2. Conversion is 98 %.

The remaining 2 % of unreacted cyclohexanone is oximated at pH 4.5 with about 3 % of the mainstream hydroxylamine. Ammonia is added. The toluene, containing about 30 % oxime, is separated from the aqueous buffer solution. The cyclohexanone oxime is freed from toluene and fed into the Beckmann rearrangement process. The distilled solvent is recycled to the oximation process [26]. To avoid poisoning of the catalyst, residual cyclohexanone and oxime are extracted from the separated buffer solution. Toluene is used as the extracting agent and the process is carried out in a packed pulsed column. The exhausted buffer solution still contains dissolved toluene, which is stripped with steam. In addition, the excess ammonium ions formed as a reduction byproduct in the buffer solution must be removed. This is achieved with nitrous gases from the combustion of ammonia. This reaction proceeds according to the following equation:

$$2\ NH_4H_2PO_4 + NO + NO_2 \longrightarrow 2\ N_2 + 3\ H_2O + 2\ H_3PO_4$$

The reaction takes place in a decomposition column, and the nitrogen generated passes into the vent gas. The excess nitrogen oxides are absorbed in a downstream column and recycled for hydroxylamine synthesis.

Oximation without Neutralization or Acidic Oximation (BASF). The reactions involved are summarized by the following basic formula [25]:

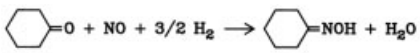

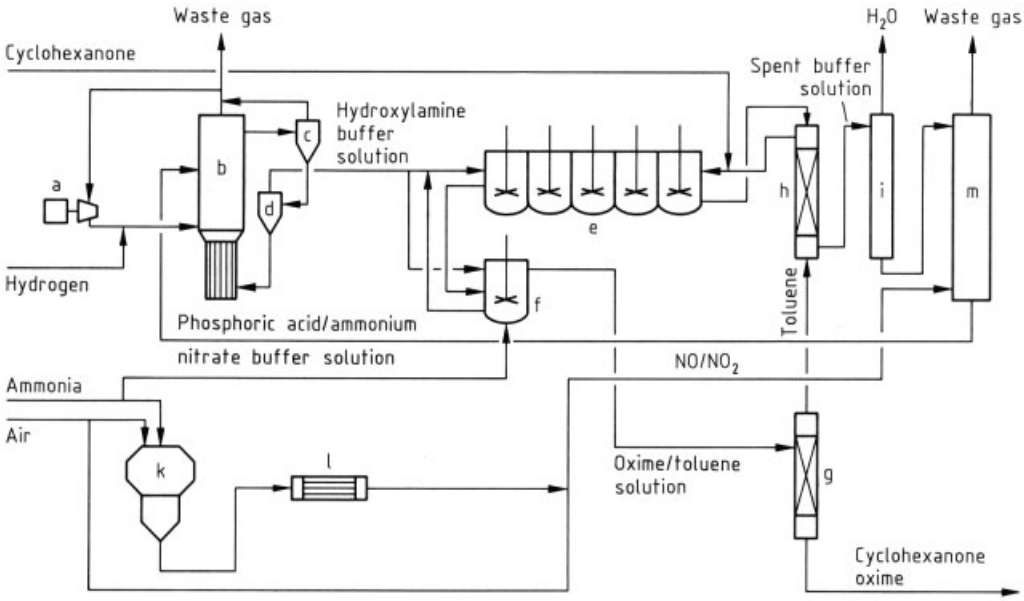

Figure 3. DSM HPO hydroxylamine and cyclohexanone oxime production
a) Compressor; b) Hydroxylamine generator; c) Separation; d) Filtration; e) Oximation; f) Neutralization; g) Solvent distillation; h) Extraction; i) Toluene stripping; k) Ammonia combustion; l) Condenser; m) Decomposition and absorption column

1) Nitric oxide is catalytically hydrogenated in an ammonium hydrogen sulfate solution:

$$NO + 3/2\ H_2 + (NH_4)HSO_4 \longrightarrow (NH_3OH)(NH_4)SO_4$$

Platinum on graphite is used as catalyst.

2) Cyclohexanone oxime is formed by reacting cyclohexanone with ammonium hydroxylammonium sulfate:

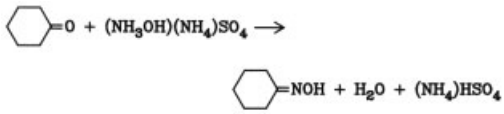

Whereas the sulfuric acid set free in BASF's classic process is neutralized with ammonia for separation of cyclohexanone oxime, further neutralization of ammonium hydrogen sulfate is not required for cyclohexanone oxime recovery. Hence, the ammonium hydrogen sulfate is recycled directly into hydroxylamine production (Fig. 4).

After filtration and separation of the catalyst, the ammonium hydroxylammonium sulfate solution is fed into the oximation process. Oximation takes place in a special column by means of countercurrent flow of cyclohexanone and hydroxylamine solution with 97 – 98 % conversion.

Temperature is maintained above the melting point of cyclohexanone oxime.

Complete conversion of cyclohexanone is achieved with classic after-oximation.

Some ammonium sulfate is produced was by-product in both hydroxylamine synthesis and after-oximation. The ammonium level can be kept constant with a purge to after-oximation. The corresponding sulfate deficit is made up by addition of sulfuric acid.

The aqueous phase of oximation contains ammonium hydrogen sulfate as well as traces of cyclohexanone and cyclohexanone oxime. The recycled ammonium hydrogen sulfate solution must be free of residual carbon. This is achieved in a stripping tower, where residual cyclohexanone oxime is decomposed and cyclohexanone stripped. The resulting ammonium hydrogen sulfate solution is virtually free of organic material and recycled to hydroxylamine production.

Acidic oximation gives 0.1 ton ammonium sulfate per ton of cyclohexanone oxime.

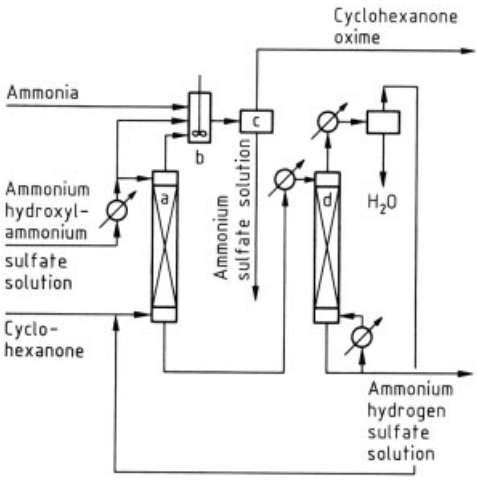

Figure 4. BASF acidic oximation
a) Oximation; b) After-oximation; c) Ammonium purge; d) Stripping column

4.1.4. Gas-Phase or Catalytic Rearrangement of Cyclohexanone Oxime

Interest in heterogeneous catalytic rearrangement of cyclohexanone oxime was prompted by economic considerations in order to avoid the ammonium sulfate generated in Beckmann rearrangement.

In 1938, Du Pont proposed a process in which the cyclohexanone oxime vapors are passed over water-eliminating catalysts such as oxides of silicon, aluminum, titanium, or magnesium, preferably in the presence of ammonia, under atmospheric or reduced pressure at 250 – 350 °C.

In the early 1950s, Leuna-Werke demonstrated surface catalysis [27]. This work led to worldwide activity on catalytic rearrangement by such companies as BASF, Bayer, Solvay, BP Chemicals, Mitsubishi, Asahi, and Honshu.

An additional catalytic rearrangement process was developed in the former States of the USSR.

Figure 5 shows the process developed and proven by BASF in a large-scale pilot plant. The catalyst system contains boric acid. It is advantageous to have the catalyst in a fluidized bed during reaction and regeneration. A fluidized flow system is also possible [28].

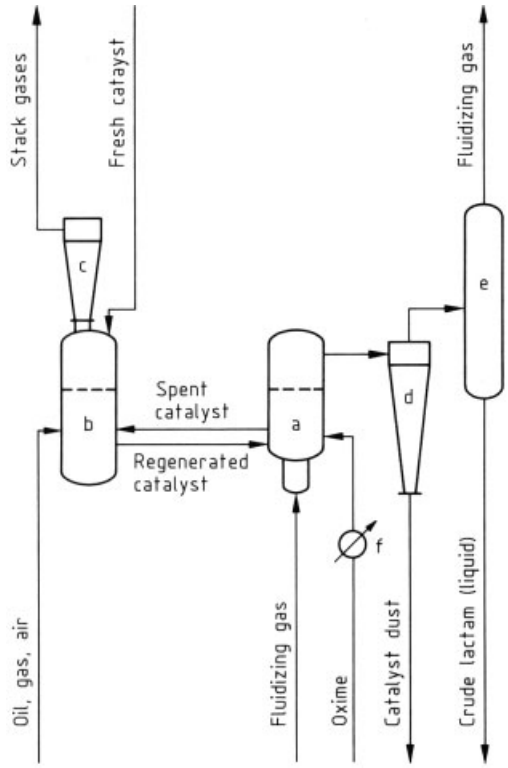

Figure 5. Catalytic oxime rearrangement
a) Fluidized-bed reactor; b) Catalyst regenerator; c) Cyclone; d) Cyclone; e) Condensation; f) Pre-evaporation

The catalyst can be transported by means of a pneumatic conveyor from the fluidized-bed reactor to the catalyst regenerator and back. In the catalyst regenerator, the catalyst is regenerated with air at 700 – 900 °C [28–30]. Dust is removed from the stack gases in a cyclone. The catalyst can be cooled to the reaction temperature before recycling.

The fluidized-bed reactor can be operated under atmospheric, reduced, or elevated pressure by introducing a certain amount of inert gas as a fluidizing agent to the bottom of the reactor.

The cyclohexanone oxime can be injected into the 275 – 375 °C fluidized bed as a gas or liquid via nozzles. It can be introduced as a solid with the aid of an inert-gas stream [31–33]. The pre-evaporation of cyclohexanone oxime is particularly advantageous.

The recovery of rearrangement heat supplies the heat of fusion and evaporation.

Catalyst dust deposited in the cyclone can be reused for the production of fresh catalyst. The

hot vapors from the cyclone can be condensed on a surface condenser or in a column [34]. The inert gas required for fluidization can be circulated. Crude caprolactam obtained by catalytic rearrangement (yield up to 96 %, based on cyclohexanone oxime) is less pure and the impurities are different from those in caprolactam produced by Beckmann rearrangement. It is therefore purified by combined extraction, crystallization, and distillation methods, as described in references [35–45].

However, caprolactam has not yet been produced catalytically on an industrial scale.

4.1.5. Modified Oximation and Beckmann Rearrangement

It has been proposed that cyclohexanone be reacted with hydroxylammonium nitrate instead of hydroxylammonium sulfate. In this case, ammonium nitrate is formed as byproduct in oximation [46].

Another process bypasses ammonium salts as byproducts in oximation. The acid obtained during the reaction of cyclohexanone with hydroxylammonium salts is removed by electrodialysis [47].

The Beckmann rearrangement of cyclohexanone oxime can be carried out in polyphosphoric acid [48] instead of oleum. The rearrangement mixture is neutralized in an organic solvent in which caprolactam and polyphosphoric acid are soluble. The monoammonium dihydrogen phosphate and diammonium monohydrogen phosphate formed during neutralization with ammonia are separated.

It has also been proposed that the rearrangement of cyclohexanone oxime be carried out in acetic acid over acidic ion exchangers [49]. Mixtures of oleum and phosphoric acid, sulfur trioxide in liquid sulfur dioxide, thionyl chloride, phosgene, hydrofluoric acid in acetic anhydride, and chlorosulfonic acid in chlorinated hydrocarbons can also be used [50].

Attempts have been made to neutralize the mixture obtained from rearrangement in oleum with metal oxides, such as calcium oxide, instead of ammonia [51]. It is proposed that the resulting calcium sulfate is then decomposed thermally into recyclable calcium oxide and sulfur dioxide. Diluting the rearrangement reaction mixture with water and extracting the caprolactam with a solvent such as phenol is another method for avoiding the production of ammonium sulfate [52].

Another approach is to dilute the rearrangement mixture with aqueous ammonium sulfate or ammonium dihydrogen phosphate resulting in splitting caprolactam sulfate into caprolactam and sulfuric acid and then extracting caprolactam [53]. The resulting sulfuric acid solution containing ammonium salts is used in phosphoric acid production. It has been proposed for economic reasons to decompose ammonium sulfate from rearrangement back into ammonia and sulfur dioxide containing gas streams [54].

4.2. Photooximation

In the 1950s, Toray industries developed a photochemical process for the production of caprolactam (Fig. 6).

Photonitrosation (PNC) converts cyclohexane to cyclohexanone oxime dihydrochloride followed by Beckmann rearrangement.

The cyclohexane is reacted with nitrosyl chloride to give cyclohexanone oxime hydrochloride:

1) Preparation of nitrosylsulfuric acid from nitrous gases obtained from combustion of ammonia, and sulfuric acid:

$$2\ H_2SO_4 + NO + NO_2 \longrightarrow 2\ NOHSO_4 + H_2O$$

2) Preparation of nitrosyl chloride by reaction with hydrogen chloride:

$$NOHSO_4 + HCl \longrightarrow NOCl + H_2SO_4$$

3) Photochemical reaction:

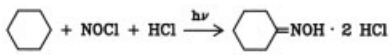

The industrial photonitrosation process is based on the development of efficient photoreactors. Toray designed an immersion lamp with a high radiation efficiency and capacity as well as long life. In order to remove the short-wave radiation below 365 nm, which contributes to tar formation on the lamps, either an absorbent is

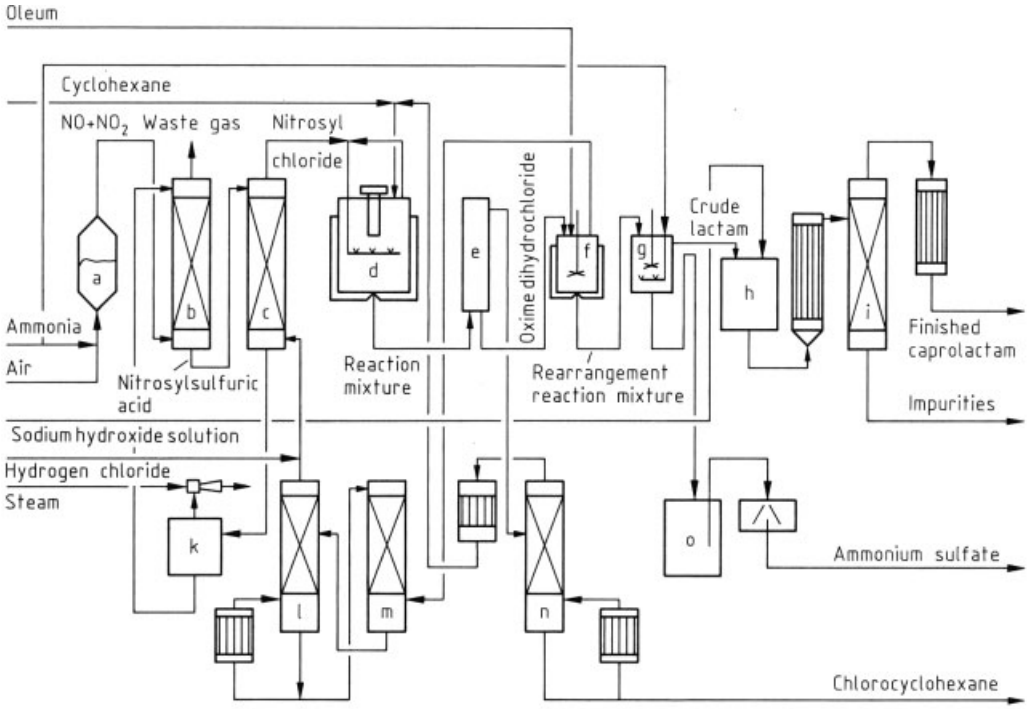

Figure 6. Toray PNC caprolactam production
a) Ammonia combustion; b) Nitrosylsulfuric acid generator; c) Nitrosyl chloride generator; d) Photonitrosation; e) Cyclohexane/cyclohexanone oxime separation; f) Rearrangement; g) Neutralization; h) Chemical treatment; i) Drying and lactam distillation; k) Dewatering of sulfuric acid; l) Hydrogen chloride regenerator; m) Hydrogen chloride recovery; n) Cyclohexane recovery; o) Ammonium sulfate recovery

added to the cooling water, or the light source is surrounded by a glass filter [55].

The cyclohexanone oxime produced by photonitrosation of cyclohexane is separated as the dihydrochloride in the presence of excess hydrogen chloride. This compound exists in the form of oil droplets and forms a heavier lower phase in cyclohexane; this lower phase is subjected to Beckmann rearrangement with excess sulfuric acid or oleum to give caprolactam. Hydrogen chloride is set free and recycled. The rearrangement reaction mixture is neutralized with ammonia water to give crude lactam and ammonium sulfate.

PNC process description: Nitrous gases are produced in an ammonia burner and cooled. They then react with sulfuric acid to form nitrosylsulfuric acid. This is brought into contact with excess hydrogen chloride to give nitrosyl chloride. The remaining sulfuric acid is dehydrated and then either recycled into the nitrosyl chloride production stage, or circulated. The gaseous nitrosyl chloride/hydrogen chloride mixture is passed through liquid cyclohexane. The conversion to oxime dihydrochloride is carried out in a photoreactor with actinic light from cooled mercury immersion lamps. This product separates at the bottom as a heavy oil. In order to prevent the deposit of oxime salt and resinous coating, the walls of the lamp cooler are periodically washed with concentrated sulfuric acid [56]. The thermal energy emitted from the light source must be removed by cooling water. Since nitrosyl chloride, hydrogen chloride, and oxime dihydrochloride are extremely corrosive, the photoreactor is made of titanium [57] or lined with enamel or PVC.

Unreacted nitrosyl chloride is removed from the photoreactor with the excess hydrogen chloride and recycled into the photochemical reaction. The oily oxime dihydrochloride is then rearranged with oleum. Unreacted cyclohexane is purified during cyclohexane regeneration and reused. Hydrogen chloride is liberated during

Beckmann rearrangement if carried out under the same conditions as with pure oxime which is then absorbed during the hydrogen chloride regeneration stage in dilute hydrochloric acid. The solution is concentrated in the hydrogen chloride recovery stage and recycled into nitrosyl chloride production. The mixture from the rearrangement reaction is neutralized with ammonia water and the aqueous crude caprolactam is purified. This includes a chemical process, drying, and distillation. In cyclohexane regeneration, a small chlorocyclohexane residue is obtained. The ammonium sulfate solution is crystallized by evaporation. In the PNC Toray process, 1.55 tons of ammonium sulfate are produced per ton of caprolactam [58].

4.3. Production from Toluene

Snia Viscosa developed a toluene-based process in 1960 [59](Fig. 7).

Caprolactam is obtained in three steps:

1) Catalytic oxidation of toluene with air to benzoic acid:

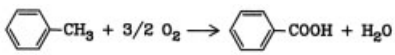

2) Hydrogenation of benzoic acid to cyclohexanecarboxylic acid:

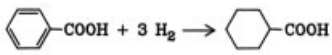

3) Nitrosodecarboxylation of cyclohexanecarboxylic acid to caprolactam in the presence of oleum:

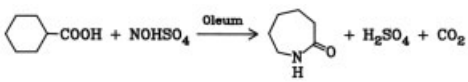

During the last stage, nitrosocyclohexane and cyclohexanone oxime are formed in situ. Cyclohexanone oxime immediately undergoes rearrangement to caprolactam.

Snia process description: The oxidation of toluene with air is carried out in liquid phase using a cobalt catalyst at 160 – 170 °C and 0.8 – 1 MPa (8 – 10 bar) pressure. The yield is well above 90 % of theory. The gases from the oxidation reactor contain mainly nitrogen with small amounts of oxygen, carbon dioxide and carbon monoxide. They are cooled to 7 – 8 °C in order to recover unreacted toluene. The water accumulated during the reaction is removed in a separator drum and toluene is recycled to the reactor. In addition to toluene and cobalt catalyst, the liquid reaction product contains about 30 % benzoic acid and also various intermediates and byproducts.

Most of the toluene is removed by distillation, and the remaining concentrated solution is given into a rectification column. The lower-boiling intermediates and the remaining toluene are removed at the top and reused. Benzoic acid in vapor form is removed from the rectification column as a side stream. The high-boiling byproducts leave the column as a residue. They include the cobalt salt, which is reprocessed. The benzoic acid obtained is suitable for hydrogenation without any further purification.

Hydrogenation of benzoic acid to cyclohexanecarboxylic acid is carried out in liquid phase in the presence of a palladium on graphite catalyst. A series of stirred reactors are used with a temperature of about 170 °C and pressure of 1 – 1.7 MPa (10 – 17 bar). Conversion is 99.9 % with a commensurate yield. The catalyst is centrifugally separated from the liquid reaction product, mixed with benzoic acid, and reused in hydrogenation. Cyclohexanecarboxylic acid is distilled under reduced pressure. Any catalyst still present remains in the residue. A 73 % nitrosylsulfuric acid solution in sulfuric acid is used for the nitrosation of cyclohexanecarboxylic acid. It is obtained by absorption of nitrous gases (from ammonia combustion) in concentrated sulfuric acid or oleum. Cyclohexanecarboxylic acid is premixed with oleum at ambient temperature and fed into a multistage nitrosation reactor. Exactly defined amounts of cyclohexanecarboxylic acid and nitrosylsulfuric acid are reacted at each stage, ensuring the complete reaction of the nitrosating agent. The conversion rate of the cyclohexanecarboxylic acid is maintained at about 50 %. One mole of carbon dioxide per mole of caprolactam is set free along with small amounts of other gases such as nitric oxide, carbon monoxide, sulfur dioxide, and nitrogen. In order to maintain the reaction and remove the reaction heat, this process

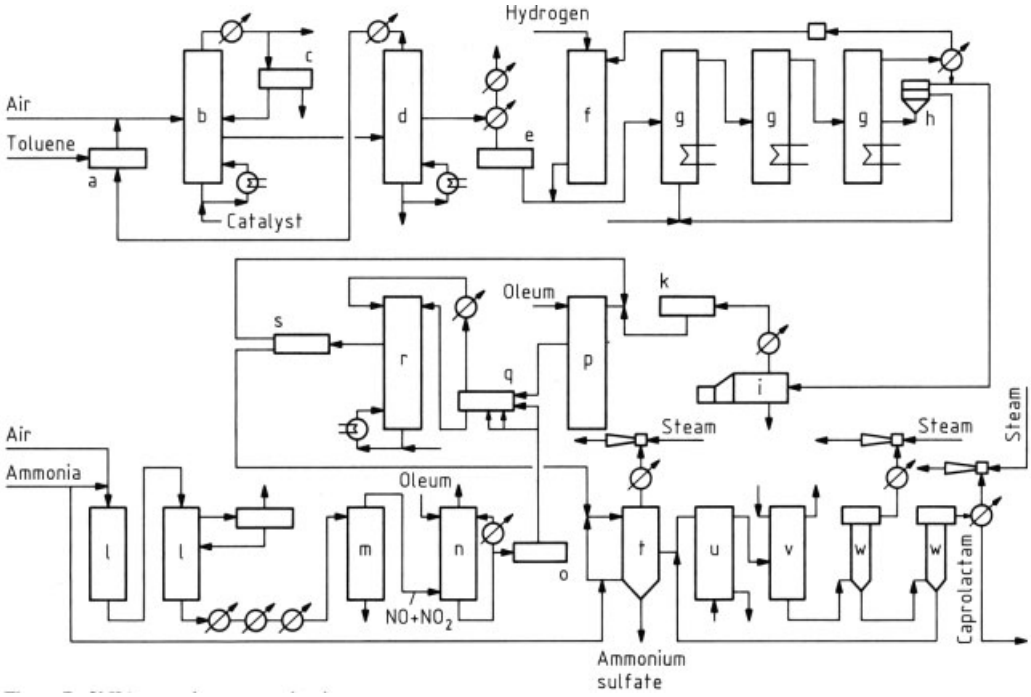

Figure 7. SNIA caprolactam production
a) Toluene tank; b) Oxidation; c) Separation; d) Rectification; e) Benzoic acid tank; f) Benzoic acid/hydrogen mixture; g) Benzoic acid hydrogenation; h) Removal of catalyst; i) Cyclohexanecarboxylic acid distillation; k) Cyclohexanecarboxylic acid tank; l) Ammonia combustion; m) Separation; n) Nitrosylsulfuric acid generator; o) Nitrosylsulfuric acid tank; p) Cyclohexanecarboxylic acid/oleum mixture; q) Rearrangement; r) Hydrolysis; s) Solution of cyclohexanecarboxylic acid in cyclohexane; t) Neutralization and ammonium sulfate crystallization; u) Solvent extraction; v) Water extraction; w) Lactam distillation

is carried out in boiling cyclohexane under atmospheric pressure. Depending on the organic acid, less than 10 % of the yield is lost.

Sulfonic acids are byproducts. During the reaction, 3.3 moles of sulfuric acid per mole of caprolactam are consumed. Nitrosation with subsequent rearrangement is the most important part of the process [60].

Products from the nitrosation/rearrangement step are hydrolyzed with water at low temperatures. Cyclohexane extracts unreacted cyclohexanecarboxylic acid. The final residues are extracted from the sulfuric acid solution, which has been diluted with water, by countercurrent extraction using cyclohexane as the solvent. The recovered cyclohexanecarboxylic acid is freed from the solvent and recycled to the process. The acidic caprolactam solution containing excess sulfuric acid is then fed into the neutralization stage, where it is neutralized with ammonia. Neutralization is carried out directly in a crystallizer under reduced pressure. Two liquid layers are formed, a saturated ammonium sulfate solution and a concentrated aqueous caprolactam solution. The caprolactam solution is purified in several stages. It is first separated from the water-soluble byproducts by toluene extraction. An aqueous caprolactam solution is then obtained by countercurrent extraction of the caprolactam–toluene solution with water. The toluene-soluble byproducts remain in the organic layer. Water is removed from the aqueous caprolactam solution and the product distilled to give pure caprolactam.

Crystallization of the separated saturated ammonium sulfate solution gives 4.1 tons of ammonium sulfate per ton of caprolactam.

Research centered on reduction of ammonium sulfate production:

Cyclohexanecarboxylic acid is treated with oleum in the first version of this process to produce pentamethylene ketene.

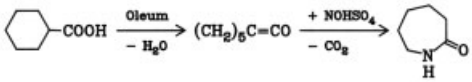

The pentamethylene ketene is then nitrosodecarboxylated to caprolactam in the presence of oleum. Ammonium sulfate is decreased to 2 tons per ton of caprolactam by this route [61].

The second version produces no ammonium sulfate as a result of a modification in the separation procedure: Caprolactam dissolved in sulfuric acid is extracted by diluting the solution with small amounts of water, which is then extracted with an alkylphenol. The remaining sulfuric acid is subjected to thermal cracking, which destroys the impurities, and the SO_2 formed is recycled into the process. Aside from not generating ammonium sulfate, this version is also advantageous in that impurities do not present waste disposal problems [62].

4.4. Recovery from Nylon-6 Waste

Solid nylon-6 waste is depolymerized in a kettle operation with the aid of a cracking catalyst and superheated steam [63].

The mixture of steam and caprolactam leaving the kettle is fed to an evaporation column, where the mixture is condensed and concentrated under atmospheric pressure. The concentrate is purified by adding an oxidizing agent that converts impurities to readily removable compounds. The caprolactam is purified through distillation.

The washwater obtained from extraction of nylon-6 chips contains oligomers of caprolactam as well as caprolactam. The oligomers are separated after preconcentration in a thin-film evaporator and then depolymerized in the same manner as the solid waste. The resulting aqueous caprolactam solution is purified chemically. Subsequent distillation gives water, a fraction containing low-boiling compounds, and caprolactam as the principal product.

Oligomers can also be depolymerized by other methods [64], [65]. The washwater of nylon-6 chip production is concentrated at a sufficiently high temperature to maintain the oligomers in solution.

The caprolactam-oligomer solution is fed into a fixed- or fluidized-bed reactor. The oligomers are cracked at 275 – 350 °C.

A special aluminum oxide catalyst is used and a yield of 95 %, relative to the oligomers in the feed, is obtained. After cracking, the caprolactam is purified by conventional methods [65].

4.5. Former Commercial Routes

Union Carbide has developed a process for the manufacture of caprolactam based on the following reaction sequence [66]:

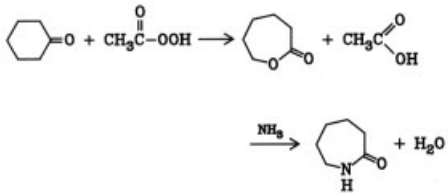

This reaction can be carried out by two methods:

Cyclohexanone is reacted with peracetic acid in an anhydrous medium, such as acetone. With an excess of cyclohexanone, ε-caprolactone is obtained in a yield of about 90 %.

Oxidation can also be carried out in situ. Cyclohexanone is reacted with air at 25 – 50 °C with the simultaneous introduction of acetaldehyde. An excess of cyclohexanone is also used in this case. Manganese, cobalt, platinum, palladium, vanadium, and zirconium salts and their oxides on carriers are examples of suitable oxidation catalysts. The acetic acid formed as a byproduct during the oxidation (about 1.1 kg per kg of lactam) is separated from the ε-caprolactone by distillation. The conversion of caprolactone to caprolactam with ammonia can either be carried out in anhydrous or aqueous medium, preferably at 350 – 425 °C at increased pressure. An excess of ammonia is used. The caprolactam formed during the ammonolysis of caprolactone in aqueous medium is isolated by extraction. The total yield is 65 – 70 %.

Union Carbide operated a plant with a capacity of about 20 000 t/a using this method.

Several companies have developed processes for the production of caprolactam using nitrocyclohexane as a starting product. Nitrocyclohexane is converted to cyclohexanone oxime by

partial reduction with hydrogen in the presence of silver oxide, zinc oxide, or chromium oxide catalysts in an aqueous phase [67], [68]. The yield can be increased by using lead-modified platinum and palladium catalysts [69]. The reduction is carried out under increased pressure at 100 – 150 °C. The byproducts include cyclohexylamine.

Du Pont operated a caprolactam plant based on catalytic reduction of nitrocyclohexane to cyclohexanone oxime with a capacity of about 25 000 t/a. Rearrangement was carried out in the conventional manner. The plant was shut down.

The catalytic conversion of nitrocyclohexane to caprolactam is a modified method never used industrially. The reaction is carried out over borophosphoric acid catalysts in the gas phase at 150 – 450 °C [70].

4.6. New Routes in Development

Several attempts are being made to develop new routes using cheaper raw materials or more economical reaction steps or minimizing byproducts and waste.

Hydrogen Peroxide Processes. Inventa describes a process for the production of cyclohexanone oxime by oxidation of cyclohexylamine with hydrogen peroxide in the presence of catalysts and stabilizers [71]. Salts of acids of group VI B metals are used as catalysts. The yield based on cyclohexylamine is 98 %. An adduct of 1 mol of cyclohexylamine and 1 mol of cyclohexanone oxime is formed as an intermediate, which is hydrolyzed to cyclohexylamine and cyclohexanone oxime.

Researchers of the former states of the USSR succeeded in catalytically converting cyclohexanone with hydrogen peroxide in the presence of ammonia in aqueous solution to cyclohexanone oxime [72]. Toa Gosei developed a process based on this approach [73]. The reaction is carried out with an excess of ammonia at 10 – 30 °C under atmospheric pressure. Tungsten and tin compounds are used as catalysts. The cyclohexanone oxime yield is said to be well above 90 %. The cyclohexanone oxime is obtained from the reaction mixture by extraction with organic solvents. About 95 % of the catalyst can be recovered from the aqueous solution. Catalyst still in solution is adsorbed on carbon.

Ammonium sulfate byproduct is formed by both the Inventa and Toa Gosei process during rearrangement of the cyclohexanone oxime.

Another hydrogen peroxide process known as the 1,1′-Peroxydicyclohexylamine Process has been demonstrated by BP Chemicals [74] and Degussa [75]:

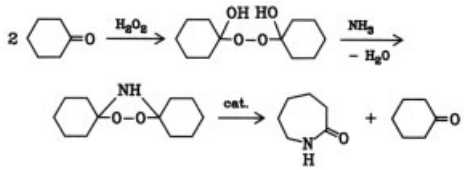

Cyclohexanone is reacted with hydrogen peroxide in a molar ratio of 2 : 1 in the presence of an organic solvent, such as methanol, ethanol, or isopropanol, and a stabilizer, such as sodium ethylenediaminetetraacetate, to give 1,1′-dihydroxydicyclohexyl peroxide.

The peroxide reacts with ammonia to form 1,1′-peroxydicyclohexylamine. Salts of organic acids are used as accelerators. The reaction temperature for the first stage is 25 °C, while the second stage reacts at about 25 – 45 °C; the reaction can also be carried out in a single stage. 1,1′-Peroxydicyclohexylamine is converted to caprolactam and cyclohexanone in the presence of catalysts such as lithium halides; molten caprolactam is used as a solvent. The reaction mixture is then distilled at 100 – 115 °C and 1.3 – 2 kPa (13 – 20 mbar). Cyclohexanone is recycled and caprolactam is purified in a conventional manner.

This process proceeds without any byproduct formation.

Oxidation of Cyclohexylamine with Elemental Oxygen. Allied proposed the oxidation of cyclohexylamine to cyclohexanone oxime with elemental oxygen [76].

Cyclohexylamine vapor is brought into contact with oxygen. Helium is used as a diluent to keep cyclohexylamine and oxygen concentration outside the explosion limits. This gas is passed over silica gel catalyst at temperatures between 120 – 250 °C and adequate pressure. A yield of about 11 % per pass of the reaction mixture over the catalyst is reported.

The cyclohexanone oxime obtained can be processed to caprolactam in the usual way.

Direct Oximation of Cyclohexanone. Allied introduced a process for ammoximation of cyclohexanone [77].

Ammonia, air, and cyclohexanone react in the presence of silica, alumina, or gallia catalysts at temperatures between 50 and 500 °C. A solvent can be employed to remove cyclohexanone oxime from the catalyst.

A 40 % yield of cyclohexanone oxime is obtained.

Bis(nitrosocyclohexane) Process. Monsanto described a process for the production of cyclohexanone oxime using electrical discharge [78]. In this process, cyclohexane is reacted with either nitric oxide, alkyl nitrites, or nitroalkanes to form bis(nitrosocyclohexane). If the reaction is carried out at about 100 °C, the bis(nitrosocyclohexane) undergoes in situ rearrangement to give cyclohexanone oxime.

Pressure, temperature, ratio of reactants, residence time, and current density are important prerequisites for obtaining a good yield. Other companies have developed similar processes.

ε-Hydroxycaproic Acid – ε-Caprolactone Process. Caprolactam production via ε-hydroxycaproic acid and ε-caprolactone has long been known [79].

Teijin described a process for ε-hydroxycaproic acid production by oxidation of cyclohexanone with air in the presence of such catalysts as azoisobutyronitrile or cyclohexyl peroxide. Substantial amounts of adipic acid are formed as byproduct. The yield of ε-hydroxycaproic acid and adipic acid is about 80 % [80].

Several companies have attempted to produce caprolactam from ε-hydroxycaproic acid. ε-Hydroxycaproic acid and its esters, including ε-caprolactone, can be converted, either in the liquid or in the gas phase, to caprolactam using ammonia.

Suitable solvents for carrying out the reaction in the liquid phase are water, alcohols, chloroform, or dioxane. Nickel, cobalt [81], copper, palladium, platinum [82], etc. are suitable catalysts. The reaction is carried out between 140 and 475 °C and usually under pressure. The yield of caprolactam after three cycles is 96 – 98 % [83].

In the gas-phase process, ε-caprolactone or the esters of ε-hydroxycaproic acid together with hydrogen, ammonia and water are passed over catalysts at temperatures of about 120 – 320 °C. The catalysts are titanium dioxide, alumina, or silica and copper or nickel [84], or copper oxide/chromium oxide [85]. The caprolactam yield is 90 % or higher.

The production of ε-caprolactone from hexane-1,6-diol, and its conversion to caprolactam, has also been described [86].

2-Nitrocyclohexanone Process. Another process forming no byproducts has been proposed by Techni-Chem [17], [87].

Cyclohexanone is reacted with excess acetic anhydride in the presence of ketene at 140 °C to give cyclohexenyl acetate. The catalysts are sulfuric or hydroiodic acid. The amount of ketene is such that the acetic acid formed during acetylation of cyclohexanone with acetic anhydride is reconverted to acetic anhydride with ketene. The reaction mixture is removed from the residue by flash distillation under reduced pressure.

Unreacted cyclohexanone and acetic anhydride are separated and recycled. Cyclohexenyl acetate is converted to 2-nitrocyclohexanone in an exothermic reaction at 30 – 50 °C in excess acetic anhydride and concentrated nitric acid (acetyl nitrate). Excess acetyl nitrate is removed under reduced pressure, absorbed by acetic anhydride, and recycled to the nitration stage. 2-Nitrocyclohexanone, acetic anhydride, and the resulting acetic acid are separated. Acetic acid is recycled into the ketene generator and acetic anhydride is returned to the nitration stage.

2-Nitrocyclohexanone is then quantitatively cracked with excess aqueous ammonia at 40 – 60 °C to give the ammonium salt of ε-nitrocaproic acid. The aqueous solution of the ammonium nitrocaproate is hydrogenated in the presence of a Raney nickel catalyst at 100 °C and 2 MPa (20 bar). The conversion to ε-aminocaproic acid is virtually quantitative. After the removal of catalyst and ammonia, the 5 – 25 % aqueous ε-aminocaproic acid solution is converted to caprolactam at 300 °C and 10 MPa (100 bar). Caprolactam is then extracted from unreacted ε-aminocaproic acid. ε-Aminocaproic acid is returned to the cyclization

process. The process is illustrated by the following equation:

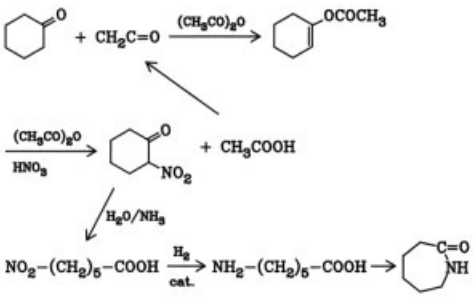

Teijin described a process for the manufacture of caprolactam from 2-nitrocyclohexanone. In this process, ring cleavage, hydrogenation, and cyclization are carried out in one step: 2-Nitrocyclohexanone is reacted with hydrogen in the presence of water and ammonia at 150 – 300 °C over hydrogenation catalysts [88].

3,3-Pentamethyleneoxaziridine Process. Leuna-Werke developed a process for the production of caprolactam by heating 3,3-pentamethyleneoxaziridine (cyclohexanone isoxime) [89].

This cyclohexanone isoxime is obtained by reacting excess cyclohexanone in aqueous ammonia solution in the presence of a water-immiscible organic solvent, with hypochlorite solution at − 10 to + 50 °C.

The cyclohexanone isoxime solution is then heated to 250 – 350 °C in an organic solvent. The distillate contains caprolactam. The conversion of 3,3-pentamethyleneoxaziridine to caprolactam can also be carried out in sulfuric acid.

Catalytic Deacetylation of *N*-Acetylcaprolactam. Kanebo discussed the preparation of caprolactam by treating *O*-acetylcyclohexanone oxime in steam over silica – alumina, silica – magnesia, or silica – zirconia catalysts. The reaction is carried out at 150 – 450 °C and reduced pressure [90].

N-Acetylcaprolactam is converted to caprolactam by reaction with cyclohexanone oxime in the presence of a catalyst. Caprolactam and *O*-acetylcyclohexanone oxime are separated by distillation.

O-Acetylcyclohexanone oxime is recycled to rearrangement [91].

Acetylene, Ethylene, Propylene, Butadiene, or Furfural based Processes. Figure 8 summarizes potential routes for the production of caprolactam.

However, none of these processes is used commercially.

4.7. Environmental Protection

Atmospheric and liquid wastes from caprolactam production require disposal methods that take local and legal restrictions into consideration:

1) Crude caprolactam extraction produces a vent gas stream containing traces of organic solvent which must be reduced to acceptable levels by physical methods such as carbon adsorption.
2) In all processes using Beckmann rearrrangement heavy bottoms in crude caprolactam extraction containing ammonium sulfate and other sulfur compounds are formed. They can be incinerated to produce sulfuric acid.
3) The residue of finished caprolactam distillation can be incinerated.

5. Quality Specifications

Industrial caprolactam is 99.9 – 99.94 % pure. Main contaminant is usually water amounting to 0.04 – 0.1 %.

High purity is requested by the users who operate processes sensitive to quality fluctuations.

Standard specifications for polymerization and subsequent processing to fibers and plastics are listed in Table 2

Table 2. Standard specifications of caprolactam (BASF)

Solidification point	min. 69 °C
Moisture content	max. 0.05 %
Permanganate absorption number	max. 5
Volatile bases	max. 0.5 meq/kg
Content of free bases	max. 0.1 meq/kg
Content of free acids	max. 0.05 mg/kg
Oxime content	max. 5 ppm
Color number of 50 % aqueous solution	max. 5 APHA
Absorbance 290 nm (50 % aqueous solution; cuvette length: 1 cm)	max. 0.05

Aqueous solutions of caprolactam should be clear in all concentration ranges

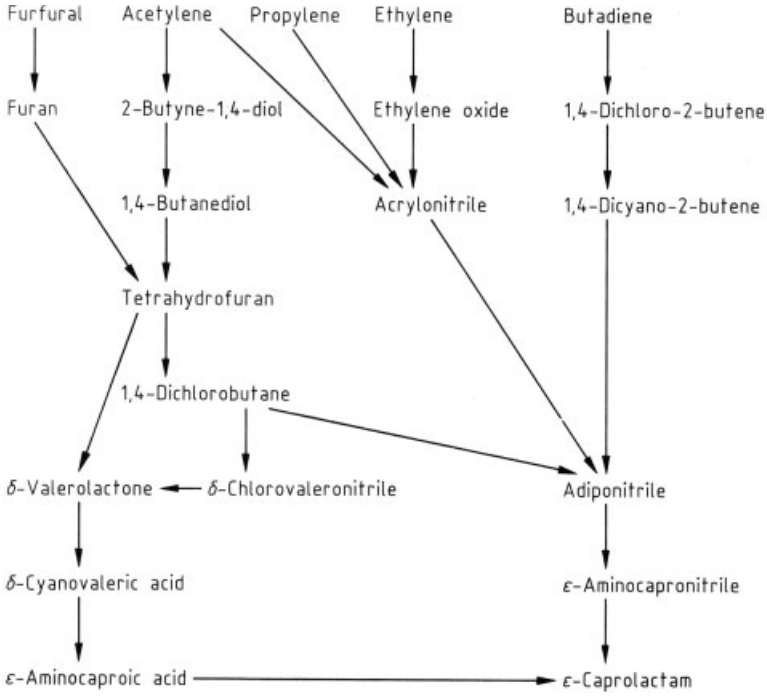

Figure 8. Routes for caprolactam production from furfural, acetylene, propylene, ethylene, and butadiene

The most important quality criteria are the ppm amounts of impurities which can be oxidized with potassium permanganate, as well as the basic impurities which are present in free or volatile form.

Quality specifications and the methods used for determining these values vary from manufacturer to manufacturer.

6. Analysis

Appearance and Solubility. Caprolactam melt is clear. A clear solution is obtained by dilution with water in all concentrations.

Permanganate Numbers. Two different approaches are used.

1) Visual comparison with standard solution: 1 mL of 0.01 N potassium permanganate solution is added to 100 mL of aqueous caprolactam solution at 20 °C. 0.01 N is equivalent to $c(1/5\,KMnO_4) \triangleq 0.01$ mol/L. The time taken for the color to change to that of the standard solution is measured. Typical values are shown in Table 3.
2) Photometric method: The absorbance of a 1 wt % caprolactam solution (50 mL or 100 mL aqueous solution) is measured 250 or 600 s after the addition of 0.01 N potassium permanganate (1 mL or 2 mL) at a specific wavelength and compared to water treated with permanganate.

Table 3. Permanganate numbers of caprolactam

Method	Concentration of the caprolactam/water solution	Standard solution composition in 1 L of water	Permanganate number in s
1	1 wt %	2.5 g $Co(NO_3)_2 \cdot 6H_2O$ + 0.01 g $K_2Cr_2O_7$	10000 to 40000
2	1 wt %	3 g $CoCl_2 \cdot 6H_2O$ + 2 g $CuSO_4 \cdot 5H_2O$	3600 to 4000
3	3 wt %	3 g $Co(NO_3)_2 \cdot 6H_2O$ + 0.012 g $K_2Cr_2O_7$	5000 to 10000

The extinction times 100 is referred to as the Klett-Summerson value or permanganate absorption number (Table 4).

Table 4. Permanganate absorption number of caprolactam

Method	Wave-length	Time, s	Extinction × 100
1	410 nm	250	Klett-Summerson value: max. 7
2	420 nm	600	permanganate absorption number: max. 5

Permanganate numbers can be converted into permanganate absorption numbers. A permanganate number of 10000 s (method 3/ Table 3) corresponds to a permanganate absorption number of 4 (method 2/Table 4).

Volatile Bases. Volatile bases are separated by steam distillation in the presence of aqueous sodium hydroxide, collected in a receiver, and determined by titration.

Methods differ in the caprolactam concentrations and the amounts of sodium hydroxide solution used. The amounts of volatile bases are expressed either in meq/kg or in mg of NH_3/kg.

Color Number. The color number is expressed in APHA units, and corresponds to the platinum – cobalt number, also referred to as Hazen color number or degree. It is used to determine the absorbance or extinction of a 50 wt % aqueous caprolactam solution. It is measured in comparison to standard solution or by photometric determination of the extinction compared to distilled water. The latter measurement is carried out in a 5-cm cell at wavelengths of 390 and 465 nm. The standard value amounts to 5 – 10 APHA units.

Solidification Point. The solidification point is determined by melt analysis. This analysis is related to the moisture content of caprolactam. The standard value amounts to 68.8 – 69.0 °C (corrected), corresponding to a moisture content of 0.04 – 0.1 % water.

The moisture content itself is determined by titration with Karl-Fischer solution.

Free Bases or Acids. The free base or acid content is measured by acidimetric titration to pH 7.0 of an aqueous caprolactam solution.

Residual Cyclohexanone Oxime. Cyclohexanone oxime is measured colorimetrically after hydrolysis to hydroxylamine, oxidation, diazotization, and formation of a colored diazonium compound or other methods.

Absorbance at 290 nm. The absorbance of a 50 % aqueous solution of caprolactam is measured at a wavelength of 290 nm and corrected to a cuvette length of 1 cm.

The International Organisation for Standardization (ISO) is attempting to standardize analytical methods for caprolactam. Methods for "Determination of crystallizing point – ISO 7060-1982", "Determination of colour of 50 % aqueous caprolactam solution – ISO DIS 8112-1983", and "Determination of absorbance at a wavelength of 290 nm – ISO 7059-1982" have been finalized and recommended.

7. Storage and Transportation

Molten caprolactam is stored and shipped in stainless steel or aluminum containers at 75 – 80 °C in the absence of moisture and under a blanket of nitrogen containing less than 10 ppm of oxygen.

Flaked caprolactam is supplied in polyethylene bags for general purposes or in aluminum coated paper bags for anionic polymerization. Flake caprolactam should be stored in dry warehouses and protected against direct sunlight. It is advisable to keep the ambient temperature below 45 °C (wherever possible) and the relative humidity below 65 %. In areas with high temperatures and humidity, an additional outer reinforced paper bag can be used for protection.

8. Economic Aspects

The United States, Japan, the Soviet Union, Belgium, the Netherlands, Italy, and Germany are the main producers of caprolactam. With the exception of France and the United Kingdom virtually all industrialized nations possess caprolactam facilities of their own.

World capacity is summarized in Table 5.

Table 5. Caprolactam capacities (kt/a) [8]

	1984	1989	1991	1994
Western Europe	845	775	841	843
FRG	–	120	160	180
USA	505	595	582	685
Latin America	115	–	–	–
Asia	615	–	–	–
Japan	(455)	469	531	519
Eastern Europe	800	–	–	–
World caprolactam capacity	2880	1959	2114	2227

Some 90 % of all caprolactam produced is processed to filament and fiber. Nearly 10 % is used for the production of plastics. Only small quantities are used for chemical syntheses.

9. Toxicology

Caprolactam is of relatively low acute toxicity: LD_{50} 1155 mg/kg (rat, oral) [92] resp. 1660 mg/kg (rat, oral); the symptoms of acute intoxication are tonoclonic convulsions. Rabbits and cats are more sensitive to caprolactam than rats [93].

Repeated epicutaneous application (50 % ether solution; 10 times) and intracutaneous injection (0.1 % in physiological NaCl solution) did not cause local irritation or sensitization to the skin of guinea pigs [93]. Other animal experiments, however, indicated weak sensitization [94]. Repeated instillation of 3 – 4 drops of a 5 % or 10 % solution into the conjunctival sac of the rabbit eye caused no irritation [95].

At the workplace caprolactam is usually inhaled as a vapor or dust. Repeated inhalations (7 times, 7 h) of dust containing 118 – 261 mg/m^3 was tolerated by guinea pigs without any symptoms except for occasional coughing [96]. In a subchronic inhalation study rats were exposed to relatively high concentrations of caprolactam (vapor: 500 mg/m^3; aerosol: 120 – 150 mg/m^3) 45 times for 4 h per day: the genital cycle of female rats was prolonged, the formation of follicles was retarded, and the reproduction of corpus luteum was stimulated [97].

Caprolactam was not teratogenic after oral administration of 1/10 of the LD_{50} (116 mg/kg) to rats from day 6 to 15 of gestation [98].

Effects on kidneys after chronic exposure are being discussed as the most important toxic effect [99].

Short-term screening for mutagenic/carcinogenic potential on bacterial and mammalian cells (Ames test, CHO test, cell transformation test) was negative for caprolactam [100].

When caprolactam was administered in the feed to mice (7500 and 15000 ppm) and rats (3750 ppm) over a period of two years no carcinogenic effect was observed; only the body weight gain was retarded with increasing concentrations [101].

Long-term experience in the handling of caprolactam has demonstrated that skin and eye irritation may occur. Skin sensitization is apparently very rare [102]. Long-term exposure of female workers to caprolactam (concentration: "a multiple of 10 mg/m^3") led to irregularities of menstruation and disorders during the course of pregnancy [97]. Another publication, however, reports no adverse health effects at workplace concentrations of 40 – 60 mg/m^3 [103].

Exposure levels are established at 1 mg/m^3 (TLV for dust) and 20 mg/m^3 (5 ppm; TLV for vapor) in the United States and at 25 mg/m^3 (MAK) in the Federal Republic of Germany.

10. References

General References

1. *Houben-Weyl,* **10/1,** 897 – 1016; **10/4,** 1 – 308, 449 – 472; **11/2,** 529 – 564.
2. M. Sittig: Chemical Process Monograph no. 21, *Caprolactam and Higher Lactams,* Noyes Development Corp., Park Ridge, N.J. 1966.
3. "Caprolactam Production: A Survey of Current Technology," *Eur. Chem. News* (1976) April 30; "Process Survey: Caprolactam," Eur. Chem. News (1969) May 2.
4. J. J. McKetta et al. in *Encyclopedia of Chemical Processing and Design,* **vol. 6,** Marcel Dekker, New York – Basel 1978, 72 – 95.
5. *Kirk-Othmer,* 3rd ed., **vol. 18,** 425 – 436.
6. W. Rösler et al.: "Entwicklungstendenzen und technologische Fortschritte bei der Produktion von ε-Caprolactam," *Chem. Tech. Leipzig* **30** (1978) no. 2, 67 – 73.
7. C. van de Moesdijk: "De Industriele Bereiding van Caprolactam, Het Monomer van Nylon 6," *PT-Procestech.* **36** (1981) no. 3, 147 – 153.

Specific References

8. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 5th ed., Wiley-VCH, Weinheim, 1998.
9. H. Hopff et al.: *Die Polyamide,* Springer Verlag, Berlin 1954, pp. 7, 23.
10. A. Rieche et al., *Z. Chem.* **3** (1963) 443 – 452.
11. A. Rieche et al., *Kunststoffe* **57** (1967) 49 – 52.
12. *Houben-Weyl,* **11/2,** 565.
13. K. Wehner et al., *Chem. Tech. Leipzig* **33** (1981) 193.

14. *Eur. Chem. News* (1966) June 3, 36.
15. *Houben-Weyl,* **10/4,** 236.
16. A. H. Jubb, *Educ. Chem.* **8** (1971) 23 – 25.
17. *Eur. Chem. News* (1969) May 2, Caprolactam Supplement.
18. O. Wichterle et al., *Collect. Czech. Chem. Commun.* **16** (1951) 591 – 602.
19. H. Hopff et al.: *Die Polyamide,* Springer Verlag, Berlin 1954, pp. 61 – 62.
20. BASF, DE 1194863, 1961.
21. Allied Chemical Corporation, US 3914217, 1975.
22. A. Krzysztoforski et al., *Chem. Prod.* (1977) Dec. 8.
23. J. Damme et al., *Chem. Eng. N.Y.* **79** (1972) July 10, 54 – 55. A. H. de Roij et al., *Chemtech* May 1977, 309 – 315.
24. Stamicarbon, DE-OS 2106385, 1971.
25. BASF, DE 2508247, 1975.
26. Stamicarbon, DE 2029114, 1970.
27. W. Dawydoff, *Chem. Tech. Leipzig* **7** (1955) 647 – 655.
28. BASF, DE 1195318, 1961.
29. BASF, DE-AS 1242620, 1961.
30. Solvay et Cie, BE 695573, 1967.
31. BASF, DE 1227028, 1962.
32. BASF, EP 17945, 1980.
33. Bayer, DE 2641414, 1976.
34. BASF, DE-OS 1445549, 1964.
35. Tore K. K., JP 24783, 1967.
36. Inventa AG, CH 317460, 1953; CH 484142, 1967.
37. Kanegafuchi Boseki K. K., JP 26300, 1966.
38. V. Kaleb, *Sb. Vys. Sk. Chem. Technol. Praze K. Chem. Inz.* Chemische Industrie (Czech) 19/44,4,162 (1969).
39. S. Stempel et al., Chemische Industrie (Czech) 19/44, 10467 (1969).
40. Zimmer Aktiengesellschaft, DE-OS 1620755, 1965.
41. Chemicke Zavody Na Slovensko, GB 666717, 1949.
42. Bayer, DE-OS 1930218, 1969.
43. Allied Chemical Corporation, US 3406167, 1965.
44. VEB Chemieanlagenbau Erfurt, DE-OS 2035859, 1970.
45. Stamicarbon, DE 953168, 1954.
46. BASF, DE-OS 2100034, 1971.
47. BASF, DE-OS 2062436, 1970.
48. BASF, DE-OS 1545617, 1965.
49. Inventa AG, CH 394212, 1962.
50. *Houben-Weyl,* **8,** 669 – 670; **11/2,** 529 – 564.
51. Toyo Rayon, DE-OS 1916149, 1969.
52. Bayer, GB 1021709, 1964.
53. Stamicarbon, DE-OS 1620468, 1965; 2129657, 1971.
54. Stamicarbon, DE-OS 2704561, 1977.
55. G. A. Turner, *Chem. Process. (London)* **15** (1969) 4 – 6.
56. Toyo Rayon, DE-AS 1150975, 1960.
57. Y. Ito et al., *Ann. N.Y. Acad. Sci* . 155 (1969) 618 – 624.
58. *Hydrocarbon Process.* **62** (1983) Nov., 85.
59. Snia Viscosa, IT 603606, 1960; 604795, 1960; 608873, 1960.
60. M. Taverna et al., *Hydrocarbon Process.* **49** (1970) 137 – 145.
61. G. Siali, *Hydrocarbon Process.* **53** (1974) July, 124; *Hydrocarbon Process.* **54** (1975) Jan., 83.
62. A. Heath, *Chem. Eng. N.Y.* **81** (1974) no. 15, 70.
63. R. Conrad, *Chem. Ing. Tech.* **45** (1973) 1510.
64. SU 176680, 1964.
65. BASF, DE 3030735, 1980.
66. Union Carbide, US 3000880, 1959; 3025306, 1960; 3064008, 1960.
67. C. Grundmann et al., *Angew. Chem.* **62** (1950) 556 – 560.
68. Stamicarbon, DE-AS 1012910, 1953.
69. Du Pont, DE-AS 1073490, 1957.
70. Du Pont, US 2634269, 1951.
71. Inventa AG, DE 939808, 1952.
72. D. L. Lebedew et al., *J. Gen. Chem. USSR (Engl. Transl.)* **30** (1960) no. 92, 1631.
73. Toa Gosei Chem. Ind., DE-AS 1274124, 1967.
74. BP Chemicals Ltd., DE-OS 1695503, 1967; 1770477, 1968; 1803872, 1968.
75. Degussa, DE-OS 2003269, 1970; 2004440, 1970.
76. Allied Chemical Corporation, EP-A 43445, 1981.
77. Allied Chemical Corporation, US 4163756, 1979; 4281194, 1981; 4225511, 1980.
78. Monsanto, DE-OS 2027810, 1970.
79. Anorgana GmbH, DE 935544, 1953.
80. Teijin Ltd., JP 7210368, 1967; 7211736, 1969; 7211413, 1967.
81. Teijin Ltd., JP 6614983, 1964.
82. Société d'Electrochimie d'Ugine, FR 1411872, 1964.
83. Stamicarbon, NL 6412868, 1964.
84. Teijin Ltd., DE-OS 2111216, 1971.
85. Kanegafuchi-Boseki K. K., FR 1506874, 1966.
86. W. Reppe: *Neuere Entwicklungen auf dem Gebiet der Chemie des Acetylens und Kohlenoxyds,* Springer Verlag, Berlin – Göttingen – Heidelberg 1949, 38.

87. The Techni-Chem Comp., DE-OS 1940809, 1969.
88. Teijin Ltd., DE-OS 1931121, 1969.
89. VEB Leuna Werke "Walter Ulbricht", DE-OS 1961474, 1969; 1961473, 1969; 2055165, 1970.
90. *Chem. Eng. News* **51** (1973) no. 15, 114.
91. Kanebo Co. Ltd., DE-OS 2307302, 1973.
92. G. Bornmann, A. Löser, *Arzneim. Forsch.* **9** (1959) 9.
93. BASF, unpublished results (1965/66).
94. T. P. Ivanova et al., *Farmakol. Toksikol. (Kiev)* **8** (1973) 178.
95. F. Hohensee, *Faserforschung* **1** (1951) 299.
96. M. W. Goldblatt et al., *Br. J. Ind. Med.* **11** (1954) 1.
97. A. P. Martynova et al., *Gig. Tr. Prof. Zabol.* **11** (1972) 9.
98. BASF, unpublished results (1978).
99. M. A. Friedman et al., *Food Cosmet. Toxicol.* **18** (1980) 39.
100. E. J. Greene et al., *Environ. Mutagen.* **1** (1979) 399.
101. National Toxicology Program (USA), Report 1982, NIH Publ. 81 – 1770, NTP-80-26.
102. H. Janson, *Zentralbl. Haut Geschlechtskr.* **26** (1959) 37.
103. Z. Zwierzchowski et al., *Med. Pr.* **18** (1967) no. 4, 357.

Capryl Alcohol → Alcohols, Aliphatic

Carbamates and Carbamoyl Chlorides

Dithiocarbamic Acid and Derivatives is a separate keyword.

PETER JÄGER, BASF Aktiengesellschaft, Ludwigshafen/Rh., Federal Republic of Germany (Chaps. 1 – 4)

COSTIN N. RENTZEA, BASF Aktiengesellschaft, Ludwigshafen/Rh., Federal Republic of Germany (Chaps. 2 – 4)

HEINZ KIECZKA, BASF Aktiengesellschaft, Ludwigshafen /Rh., Federal Republic of Germany (Chap. 5)

1. Introduction 1
2. Salts of Carbamic Acid 1
3. Esters of Carbamic Acids 1
4. Carbamoyl Chlorides 3
5. Toxicology and Occupational Health . . 6
6. References 7

1. Introduction

Carbamic acid, H_2NCOOH, the half amide of carbonic acid, does not exist as the free acid but forms numerous stable metal salts, esters, halides, amides (i.e., urea and urea derivatives), and simple or mixed anhydrides. A large number of derivatives result from replacement of the hydrogen atoms on the nitrogen atom by organic radicals.

2. Salts of Carbamic Acid

The salts of carbamic acid (carbamates) are relatively unstable. In aqueous solution they hydrolyze slowly at room temperature, and more rapidly on heating, to the corresponding metal carbonates and ammonia. Heating of the anhydrous solid produces cyanates from alkali-metal carbamates, and cyanamide from alkaline earth-metal carbamates.

Of the unsubstituted salts only *ammonium carbamate* (→ Ammonium Compounds, Chap. 4.2.) is made industrially. It is present in technical-grade ammonium carbonate (hartshorn, sal volatile) and is an intermediate, which is not isolated, in the industrial synthesis of urea.

3. Esters of Carbamic Acids

Carbamic acid esters are referred to as carbamates, in the same way as the metal salts, or by the trivial name urethane.

Physical Properties. The unsubstituted and the *N*-phenyl carbamates are generally crystalline, whereas the simple mono and di *N*-alkyl substituted compounds are liquids at ambient temperature. The physical data of the simple esters are shown in Table 1.

Preparation. There is no outstanding synthetic route to the esters. Depending on the carbamate to be prepared and the availability of the starting materials, one of the following methods has an economic advantage.

During the preparation of *N*-monosubstituted esters temperatures of 60 – 80 °C should not be exceeded, in order to prevent decomposition into isocyanate and alcohol.

Alcoholysis of Carbamoyl Chlorides:

$$R^1R^2N-COCl + R^3OH \xrightarrow{-HCl} R^1R^2N-COOR^3$$

In order to obtain good yields the reaction should be carried out in the presence of a base,such as a tertiary amine or pyridine, or in the presence of an acylation catalyst, such as 4-dimethylaminopyridine [5].

Aminolysis of Chloroformates:

$$2\,R^1R^2NH + ClCOOR^3 \longrightarrow R^1R^2N-COOR^3 + R^1R^2NH{\cdot}HCl$$

Since the second mole of amine neutralizers the hydrogen chloride no other base is required; however, acylation catalysts may be used if necessary. The reaction can also be carried out with tertiary amines with the formation of an alkyl chloride [6].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_051

Table 1. Physical properties of simple carbamates

Substituent	Formula	M_r	CAS reg. no.	*mp*, °C	*bp*, °C at 101.3 kPa	Density, g/cm^3	State at 20 °C
Methyl	$H_2N-COOCH_3$	75.07	[598-55-0]	54	177	1.136 (56 °C)	solid
Ethyl	$H_2N-COOC_2H_5$	89.09	[51-79-6]	48.1	185	1.0599 (48.2 °C)	solid
n-Propyl	$H_2N-COOC_3H_7$	103.12	[627-12-3]	60	196		solid
Isopropyl	$H_2N-COOCH(CH_3)_2$	103.12	[1746-77-6]	92.4	181 (94.6 kPa)	0.995 (66 °C)	solid
n-Butyl	$H_2N-COOC_4H_9$	117.15	[592-35-8]	54	204		solid
Isobutyl	$H_2N-COOCH_2CH(CH_3)_2$	117.15	[543-28-2]	67	207	0.947 (77.7 °C)	solid
n-Pentyl	$H_2N-COOC_5H_{11}$	131.17	[638-42-6]	55.5–62.5	116.7 (1.73 kPa)		solid
n-Hexyl	$H_2N-COOC_6H_{13}$	145.20	[2114-20-7]	65			solid
Phenyl	$H_2N-COOC_6H_5$	137.14	[622-46-8]	143		1.079 (60 °C)	solid
Methyl *N*-methyl	$CH_3NH-COOCH_3$	89.09	[6642-30-4]		158 (101.9 kPa)	1.065	liquid
Methyl *N*-ethyl	$C_2H_5NH-COOCH_3$	103.12	[6135-31-5]		165 (102.4 kPa)		liquid
Methyl *N,N*-dimethyl	$(CH_3)_2N-COOCH_3$	103.12	[7541-16-4]		131	1.012 (15 °C)	liquid
Ethyl *N,N*-dimethyl	$(CH_3)_2N-COOC_2H_5$	117.15	[687-48-9]		147	0.972 (15 °C)	liquid
1-Naphthyl *N*-methyl	$CH_3NHCOOC_{10}H_7$	201.22	[63-25-2]	142	(decomp.)		solid

Reaction of Urea or Urea Nitrate with Alcohols in the Presence of Heavy Metals [7]:

$$H_2N-CO-NH_2 + ROH \xrightarrow{ZnCl_2} H_2N-COOR + NH_3$$

All the lower unsubstituted carbamates (e.g., methyl or ethyl carbamate) are prepared by this method.

Reaction of Alcohols with Isocyanates:

$$R^1NCO + R^2OH \longrightarrow R^1NH-COOR^2$$

Alkyl or aryl isocyanates yield monosubstituted carbamates, whereas isocyanic acid, HNCO, yields unsubstituted esters [8]. For the substituted isocyanates the order of reactivity is *prim.* alkyl > *sec*-alkyl > *tert*-alkyl > phenyl. Lewis bases such as triethylamine and Lewis acids such as dibutyltin dilaurate [9] are useful catalysts.

Transamidation of Simple Carbamates with Higher Amines:

$$R^1NH_2 + H_2N-COOR \xrightarrow{Cat.} R^1NH-COOR + NH_3$$

Zinc acetate is commonly used as the catalyst [10].

Transesterification of Lower Alkyl Carbamates with Higher Alcohols:

$$R^1R^2N-COOR^3 + R^4OH \xrightarrow{Cat.} R^1R^2N-COOR^4 + R^3OH$$

Aluminum isopropylate is commonly used as the catalyst [11].

Aminolysis of Dialkyl Carbonates:

$$R^1O-COOR^1 + HNR^2R^3 \rightarrow R^2R^3N-COOR^1 + R^1OH$$

Rearrangements of ω-Alkoxyalkyl- and ω-Phenoxyalkylcarbamoyl Chlorides [12]:

$$R^1O(CH_2)_nN\begin{matrix}COCl\\R^2\end{matrix} \longrightarrow Cl(CH_2)_nN\begin{matrix}COOR^1\\R^2\end{matrix}$$

Oxidative Amination of Carbon Monoxide:

$$R^1NH_2 + CO + R^2OH + 1/2\,O_2 \xrightarrow{Cat.} R^1NH-COOR^2 + H_2O$$

Preferred catalysts are iron compounds [13] in the presence of small amounts of a noble metal.

Reductive Carbonylation of Organic Nitro Compounds:

$$C_6H_5NO_2 + CO + R^1OH \xrightarrow{Cat.} C_6H_5NH-COOR^1$$

The usual catalyst is palladium together with an iron compound [14].

Uses. Carbamates are used mainly as crop protection agents and pharmaceuticals, with the greater proportion being used as insecticides, herbicides, and fungicides. For examples, see page 4.

Of the simple carbamates, the methyl ester and the ethyl ester (urethane) still have a certain importance, especially as starting materials for the preparation of higher carbamates. The condensation product of ethyl carbamate and formaldehyde is used to give a crease-resistant finish to textiles [15].

4. Carbamoyl Chlorides

Carbamoyl chlorides, also known as carbamic acid chlorides, are industrially important as intermediates because of their ready availability and of their reactivity.

Physical Properties. The simple carbamoyl chlorides are colorless liquids or solids, usually with pungent odor. The physical properties of the most important carbamoyl chlorides are given in Table 2.

Chemical Properties. Carbamoyl chloride, $H_2N-COCl$, is unstable; it may be stabilized by Lewis acids (e.g., aluminum chloride) as adducts and in this form can be used in the Friedel-Crafts acylation of arenes [16]. Since many of the monosubstituted carbamoyl chlorides also deteriorate on storage, the corresponding isocyanates are used instead. Disubstituted carbamoyl chlorides are stable and can be stored for prolonged periods.

In their chemical behavior the carbamoyl chlorides resemble the carboxylic acid chlorides. Their most important reaction is the nucleophilic displacement of the chlorine atom.

Since carbamoyl chlorides are less reactive than carboxylic acid chlorides, acid acceptors or acylation catalysts such as 4-dimethylaminopyridine are generally used in reactions involving the liberation of hydrogen chloride. Table 3 gives an overview of the reactions of carbamoyl chlorides.

Hydrolysis of carbamoyl chlorides yields the corresponding substituted ammonium chloride and carbon dioxide instead of the unstable carbamic acid. This reaction is used to remove traces of carbamoyl chloride vapor from exhaust gases or to render residues harmless.

Production. In industry phosgene, $COCl_2$, is usually the starting material.

$$COCl_2 + R^1R^2NH \longrightarrow R^1R^2N-COCl + HCl$$

R^1, R^2 = alkyl, alkoxy, aryl

At low reaction temperatures (cold phosgenation) the hydrogen chloride liberated removes half the amine as the ammonium salt $R^1R^2NH \cdot HCl$. Complete conversion of valuable amines into the corresponding carbamoyl chlorides is achieved by adding a base, e.g., pyridine.

The same is achieved by the use of higher temperatures (about 90 – 100 °C, hot phosgenation) in which case the hydrogen chloride is removed from the reaction mixture as a gas. Crystallization of the substituted ammonium chloride during the course of the reaction can lead to difficulties with stirring or to blockage of the reactor. Sometimes, however, the amine salt is used as the starting material. Phosgenation is carried out at room temperature because the salt, unlike the free amine, does not react with the carbamoyl chloride already formed to produce a substituted urea:

$$R^1R^2NH + R^1R^2N-COCl \longrightarrow R^1R^2N-CO-NR^1R^2 + HCl$$

To repress urea formation, excess phosgene is used together with an appropriate amount of solvent as a diluent. Good mixing must also be employed to prevent any local excess of amine.

The solvent used should be a good solvent for phosgene, e.g., toluene, chlorobenzene, or ethyl acetate [19], and should be easily separable from the carbamoyl chloride, i.e., the boiling point difference must be great enough. This separation is not required when the carbamoyl chloride serves as the reaction medium [25].

Table 2. Physical properties of some carbamoyl chlorides

N-Substituents	Formula	M_r	CAS reg. no.	*mp*, °C	*bp*, °C	Density,
	$H_2N-COCl$	79.49	[*463-72-9*]	ca. 50	61–2	
Methyl	$CH_3NH-COCl$	93.51	[*6452-47-7*]	90	93–4	
Dimethyl	$(CH_3)_2N-COCl$	107.54	[*79-44-7*]	−33	167	1.17
Ethyl	$C_2H_5NH-COCl$	107.54	[*41891-13-8*]		92.3	
Diethyl	$(C_2H_5)_2N-COCl$	135.59	[*88-10-8*]	−32	186	1.073
n-Propyl	$n\text{-}C_3H_7NH-COCl$	121.57	[*41891-16-1*]		109 (2.67 kPa)	
Di-*n*-propyl	$(n\text{-}C_3H_7)_2N-COCl$	163.65	[*27086-19-7*]	−21	100–4 (1.6 kPa)	1.023
Diisopropyl	$(iso\text{-}C_3H_7)_2N-COCl$	163.65	[*19009-39-3*]	56–7		
Di-*n*-butyl	$(n\text{-}C_4H_9)_2N-COCl$	191.70	[*13358-73-1*]		ca. 245	0.991
Di-*sec*-butyl	$(sec\text{-}C_4H_9)_2N-COCl$	191.70	[*36756-72-6*]	−42–6	151 (0.6 kPa)	1.009
Diisobutyl	$(iso\text{-}C_4H_9)_2N-COCl$	191.70	[*38952-42-0*]	−18	105–9 (0.13 kPa)	0.985
Di-*n*-hexyl	$(n\text{-}C_6H_{13})_2N-COCl$	247.81	[*27086-21-1*]		162–4 (0.13 kPa)	
Di-*n*-octyl	$(n\text{-}C_8H_{17})_2N-COCl$	303.92	[*27086-22-2*]		200–2 (0.13 kPa)	
Methoxy-methyl	$(CH_3O)(CH_3)N-COCl$	123.54	[*30289-28-2*]		86 (5.0 kPa)	
Hexamethylene	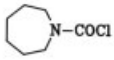	161.63	[*27817-35-2*]	−66	135 (2.67 kPa)	1.159
Oxydiethylene	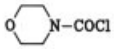	149.58	[*15159-40-7*]	ca. −50	116 (2.0 kPa)	1.32
Phenyl	$C_6H_5NH-COCl$	155.58	[*2040-76-8*]	59		
Diphenyl	$(C_6H_5)_2N-COCl$	231.68	[*83-01-2*]	85	138 (0.13 kPa)	

The amine also reacts in the gas phase with phosgene, if necessary in a diluent inert gas (e.g., nitrogen), at temperatures above 250 °C [26]. Carbamic acid chloride can be prepared in this way from ammonia and phosgene at 400–450 °C [16].

Phosgene also reacts under mild conditions (50–60 °C) with tertiary amines containing at least one lower alkyl group (preferably methyl or ethyl) to give alkyl halides as byproducts [27]:

$$COCl_2 + R^1R^2NR^3 \longrightarrow R^1R^2N-COCl + R^3Cl$$

R^3 = methyl or ethyl

Carbamoyl chlorides can also be obtained by the addition reaction of phosgene to the C=N double bond of azomethines [28–30]. α-Chloro- or α,β-unsaturated carbamoyl chlorides are formed, depending on the reaction conditions:

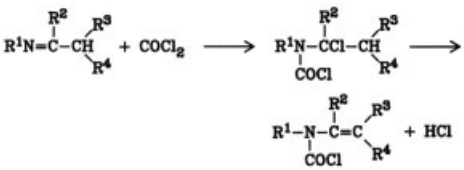

In the laboratory, carbamoyl chlorides may be prepared by the addition of hydrogen chloride to isocyanates:

$$RNCO + HCl \longrightarrow RNH-COCl$$

or by the chlorination of disubstituted formamides with sulfur dichloride, SCl_2, sulfuryl chloride, SO_2Cl_2 [31], phosphorus trichloride, PCl_3, or thionyl chloride, $SOCl_2$ [32]:

$$R^1R^2N-CHO \longrightarrow R^1R^2N-COCl$$

or by the action of phosphorus oxychloride, $POCl_3$, or phosphorus pentachloride, PCl_5, on carbamic acid esters [33]:

$$R^1R^2N-COOR^3 \longrightarrow R^1R^2N-COCl$$

as well as by carbonylation of chloramines [34]:

$$R^1R^2N-Cl + CO \longrightarrow R^1R^2N-COCl$$

Table 3. Chemical reactions of carbamoyl chlorides [16–24]

$R^1R^2N-COCl + H_2O$ Carbamoyl chloride	$\longrightarrow$	$R^1R^2NH \cdot HCl + CO_2$ Substituted ammonium chloride
$+ R^3OH$ Alcohol	$\xrightarrow[-HCl]{}$	$R^1R^2N-COOR^3$ Carbamid acid ester
$+ R^3R^4C=N-OH$ Oxime	$\xrightarrow[-HCl]{}$	$R^1R^2N-COO-N=CR^3R^4$ Oxime carbamate
$+ R^3-CO-CH_2R^4$ Aldehyde or Ketone	$\xrightarrow[-HCl]{}$	$R^1R^2N-COO-\overset{\overset{R^3}{\vert}}{C}=CHR^4$ Enol carbamate
$+ R^3-\overset{O}{\diagup\!\!\diagdown}CH-CH_2$ Epoxide	$\longrightarrow$	$R^1R^2N-COO-\overset{\overset{R^3}{\vert}}{C}H-CH_2Cl$ 2-Chloroalkyl carbamate
$+ R^3SH$ Mercaptan (Thiol)	$\xrightarrow[-HCl]{}$	$R^1R^2N-CO-SR^3$ Thiol carbamate
$+ 2\,R^3R^4NH$ Amine	$\xrightarrow[-R^3R^4NH \cdot HCl]{}$	$R^1R^2N-CO-NR^3R^4$ Urea
$+ R^3NH-NR^4R^5$ Hydrazine	$\longrightarrow$	$R^1R^2N-CO-\overset{\overset{R^3}{\vert}}{N}-NR^4R^5$ Semicarbazide
$+$ Azole	$\xrightarrow[-HCl]{}$	$R^1R^2N-CO-N$ Carbamoyl azole
$+ ArH \xrightarrow[-HCl]{AlCl_3} ArCONR^1R^2 \xrightarrow[-R^1R^2NH]{+H_2O} ArCOOH$ Aromatic — Aromatic acid amide — Aromatic acid		
$+ R^3C\equiv CNa$ Acetylene derivative	$\xrightarrow[-NaCl]{}$	$R^3C\equiv C-CONR^1R^2$ Acetylenecarboxylic acid amide
$+ R^3-COOH(Na)$ Carboxylic acid or its alkali-metal salt	$\xrightarrow[-HCl\ (-NaCl)]{} R^1R^2N-CO-O-CO-R^3 \xrightarrow[-CO_2]{}$	$R^1R^2N-COR^3$ Carboxylic acid amide
HCl- or $R^1R^2R^3CCl$-elimination		
$R^1NH-COCl$ Monosubstituted carbamoyl chloride	$\xrightarrow{Heat}$	$R^1NCO + HCl$ Isocyanate
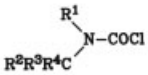 Disubstituted carbamoyl chloride with tertiary alkyl group	$\xrightarrow{Heat}$	$R^1NCO + R^2R^3R^4CCl$ Isocyanate

Uses. As is apparent from the reactions shown in Table 3, carbamoyl chlorides are intermediates with many uses. Of primary industrial importance are the reactions with alcohols, phenols, and oximes to give carbamic acid esters; with thiols (mercaptans) to give thiocarbamates; with amines and hydroxylamines to give substituted ureas; and with imidazoles and triazoles to give carbamoyl azoles. All these are important classes of compounds in crop protection (→ Fungicides, Agricultural; → Weed Control; → Insect Control). The following crop protection materials are examples of products derived from carbamoyl chlorides:

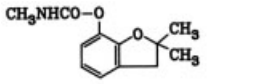

Carbofuran (insecticide)

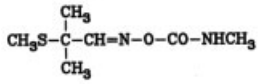

Temik (insecticide)

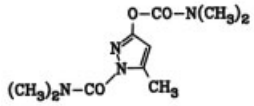

Dimetilan (insecticide)

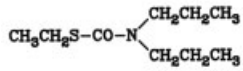

Eptam (herbicide)

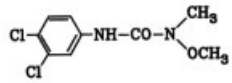

Linuron (herbicide)

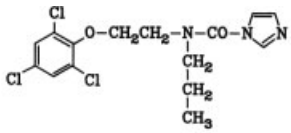

Sportak (fungicide)

Handling and Transportation. The industrially important lower carbamoyl chlorides in particular are chemically and biologically (see Chap. 5) active substances. The selection of suitable apparatus and corrosion resistant materials (e.g., glass, polytetrafluoroethylene, glass-lined steel, or, if necessary, tantalum or Hastelloy) requires much care and due attention must also be given to the various regulations existing in different countries. In the Federal Republic of Germany, for example, dimethylcarbamoyl chloride and *N*-chloroformylmorpholine (morpholinocarbamoyl chloride) are specifically mentioned in Appendix II No. 1 of the regulations dealing with dangerous materials [35] as well as in Appendix II of the regulations dealing with noxious emissions [36].

Most of the details about technical safety data are to be found in [37–39].

5. Toxicology and Occupational Health

Ethyl Carbamate. There is little data available on the acute effects of this compound: the median lethal dose (LD_{50}) upon oral administration in mice is 2500 mg/kg [40].

The most important toxicological effect is its *carcinogenicity.* Ethyl carbamate has proved to be carcinogenic in mice, rats, and hamsters upon oral, inhalative, subcutaneous, or intraperitoneal administration [41]. Lung tumors, lymphomas, hepatomas, melanomas, tumors of the blood vessels, and skin tumors (upon treatment of the skin) have occurred in the tests. The carcinogenic effects were observed upon administration of the following doses: lifelong treatment of rats with 1 ppm ethyl carbamate in drinking water; lifelong treatment of hamsters with 0.2 % ethyl carbamate in drinking water; two treatments (once a week) of the skin of mice with 120 mg ethyl carbamate, followed by eighteen applications (once a week) of 0.3 mL of a 0.5 % preparation in croton oil; inhalative treatment of mice (20 – 60 min per d) with aerosols containing varying amounts of ethyl carbamate, for a maximum of 14.5 weeks. Even short-term treatments, for example, a ten-day treatment of young mice with 0.4 % ethyl carbamate in drinking water, led to the development of tumors. Further investigations have confirmed the types of effects described here.

Ethyl carbamate has been classified by the MAK commission as being a carcinogen in animal experiments and is listed in group III A/2.

Diethylcarbamoyl Chloride. The median lethal dose (LD_{50}) upon oral administration to rats is about 200 mg/kg [42]. The compound is irritating to skin and eyes. The inhalation of an atmosphere saturated with its vapors at 20 °C is lethal to rats after 3 h [43]. Diethylcarbamoyl chloride has proved to be mutagenic on *E. coli* strains in mutagenicity tests [44]. It is weakly mutagenic in high doses in the Ames test [45]. In view of these results and on account of the structural similarity to dimethylcarbamoyl chloride, a carcinogenic potential cannot be ruled out. The MAK commission has listed this substance in group III B.

Dimethylcarbamoyl Chloride. The LD_{50} is 1170 mg/kg upon oral administration to rats and 350 mg/kg upon intraperitoneal administration to mice. The compound is irritating to the skin and eyes. All the test animals survived an 8-min inhalation test in an atmosphere saturated with

vapors of the substance at 20 °C. Longer exposure times led to death. No sensitization was observed in animal experiments [46].

Dimethylcarbamoyl chloride proved to be a mutagen in numerous short-term tests for mutagenicity. They included the Ames test [47], [48], cell transformation test [49], sister chromatid exchange [50], [52], and micronucleus test [51]. The compound caused damage to the DNA of bacteria, fungi, and plants, but not in mammalian cells in vitro. It was mutagenic in bacteria, fungi, and plants, as well as in mammalian cells in vitro; however, it was not mutagenic in insects. Chromosomal aberrations were observed in fungi and plants, and in mammalian cells in vitro and in vivo, but not in humans. A review is given in [53].

Dimethylcarbamoyl chloride is carcinogenic in animal experiments. Mice developed tumors at the site of application upon percutaneous treatment with 2 mg substance in 0.1 mL acetone, three times weekly, for 492 d. Its tumorigenic effect was confirmed in another dermal application study. The subcutaneous application of dimethylcarbamoyl chloride to mice (weekly injection of 5 mg substance in 0.05 mL tricapryline over a period of 26 weeks) also caused local tumors and – as with percutaneous treatment – reduced the survival rate. Intraperitoneal treatment of mice (weekly dose of 1 mg substance in 0.05 mL tricapryline over a period of 450 d) led to increased incidences of lung tumors [54]. Lifelong inhalation of 1 ppm of dimethylcarbamoyl chloride over 6 h every day led to squamous cell carcinomas of the nasal mucosa in 51 % of the hamsters employed. In a comparable study using more sensitive rats, 96 % of the animals developed tumors of the respiratory tract [55]. The MAK commission has listed this substance in group III A/2; in the LV list dimethylcarbamoyl chloride is in group A 2.

Diphenylcarbamoyl Chloride. The LD_{50} in rats after a single oral administration is above 2000 mg/kg. It is not irritating to the skin and eyes but caused sensitization in animal experiments [56].

Morpholinocarbamoyl Chloride. Its acute toxicity in rats is relatively small. The LD_{50} after oral administration of the substance in peanut oil is about 3500 mg/kg.

The compound is only weakly mutagenic in the Ames test. However, in view of the instability of the compound in aqueous solutions, it is doubtful whether it is only a weak mutagen.

Weekly subcutaneous injection of rats with 300 mg/kg in peanut oil over more than 140 d led to tumors around the site of injection after 140 d. This treatment caused the death of all the animals within 340 d. As many as 58 of the 60 test animals in the test developed fibrosarcomas at the site of injection, and two animals developed adenocarcinomas of the mamma. Weekly subcutaneous treatment with 100 mg/kg in peanut oil over 140 d led to the death of all the rats within 335 d. As many as 59 of the 60 animals showed a sarcoma at the site of injection; in contrast, there were only three incidences of tumors among the sixty control animals treated with peanut oil alone. Morpholinocarbamoyl chloride has been listed in group III A/2 by the MAK commission [57].

6. References

General References

1. *Beilstein* **3,** 20 – 31; **3 (1),** 9 – 15; **3 (2),** 18 – 27; **3 (3),** 39 – 65; **3 (4),** 37 – 80.
2. *Houben-Weyl,* **8,** 111 – 118, 137 – 149; **E (4),** 36 – 64, 142 – 192.
3. P. Adams, F. Baron: "Esters of Carbamic Acid," *Chem. Rev.* **65** (1965) 567 – 602.
4. G. Scheuerer: "Carbamate als Agrarchemikalien," *Fortschr. Chem. Forsch.* **9** (1967/68) 254 – 294.

Specific References

5. G. Höfle, W. Steglich, H. Vorbrüggen, *Angew. Chem.* **90** (1978) 602 – 615; *Angew. Chem. Int. Ed. Engl.* **17** (1978) 569 – 583.
6. M. Matzner, R. P. Kurkjy, R. J. Cotter, *Chem. Rev.* **64** (1964) 645.
7. A. Paquin, *Z. Naturforsch.* **1** (1946) 518 – 523.
8. Ota-Seiyaku Co., JP 4716433, 1972.
9. T. Francis, M. P. Thorne, *Can. J. Chem.* **54** (1976) 24.
10. BASF, EP 18583, 1981.
11. W. M. Kraft, *J. Am. Chem. Soc.* **70** (1948) 3569.
12. R. Banks, R. F. Brookes, D. H. Godson, *J. Chem. Soc. Perkin Trans. 1* 1975, 1836 – 1840.
13. BASF, DE-OS 2910132, 1979.
14. Bayer, DE-OS 2819826, 1978.

15. R. Arceneaux, J. Fricke, Jr., J. Reid, G. Gautreaux, *Am. Dyest. Rep.* **50** (1961) 849 – 853.
16. H. Hopf, H. Ohliger, *Angew. Chem.* **61** (1949) 183 –185.
17. B. F. Filipasic, R. Patarcity, *Chem. Ind. (London)* 1969, 166 – 167.
18. F. Boberg, G. Schultze, *Chem. Ber.* **88** (1955) 275 –280.
19. The Boots Co., GB 1469772, 1973.
20. G. A. Olah, J. Olah: *Friedel-Crafts and Related Reactions,* **vol. III,** Interscience, New York 1964, pp. 1262 – 1267.
21. R. Epsztein, *C. R. Hebd. Séances Acad. Sci.* **240** (1955) 989 – 990.
22. SU 229489, 1967 (A. Zalikin, Y. Sterpikheev).
23. J. Lawson, Jr., J. Croom, *J. Org. Chem.* **28** (1963) 232 – 235.
24. J. Tilley, A. Sayigh, *J. Org. Chem.* **28** (1963) 2076 –2079.
25. Ciba-Geigy, DE-OS 2206365, 1972.
26. R. Slocombe, E. Hardy, J. Saunders, R. Jenkins, *J. Am. Chem. Soc.* **72** (1950) 1888 – 1891.
27. H. Babad, A. G. Zeiler, *Chem. Rev.* **73** (1973) 75 – 91.
28. BASF, DE-OS 1901542, 1969.
29. Bayer, DE-OS 2146069, 1971.
30. H. Kiefer, *Synthesis* 1972, 39 – 42.
31. U. Hasserodt, *Chem. Ber.* **101** (1968) 113 – 120.
32. N. Schindler, W. Plöger, *Chem. Ber.* **104** (1971) 969 – 971.
33. O. Schmidt, *Ber. Dtsch. Chem. Ges.* **36** (1903) 2479.
34. T. Saegusa, T. Tsuda, Y. Isegawa, *J. Org. Chem.* **36** (1971) 857 – 860.
35. Verordnung über gefährliche Arbeitsstoffe (Arb-StoffV), publ. 11th February 1982.
36. Zwölfte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Störfall-Verordnung) – 12. BImSchV.
37. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin – Heidelberg – New York 1983.
38. *Encyclopedia of Occupational Health and Safety,* International Labour Organisation, Geneva 1983.
39. *Handbook of Toxic and Hazardous Chemicals,* Noyes Publications, Park Ridge, New Jersey, 1981.
40. *Arzneim. Forsch.* **9** (1959) 595.
41. *IARC Monogr. Eval. Carcinog. Risk Chem. Man* **7** (1974) 111 – 140.
42. BASF, unpublished results, 1977.
43. BASF, unpublished results, 1970.
44. J. F. Finklea, National Institute for Occupational Safety and Health; Current Intelligence Bulletin no. 12, July 7, 1976.
45. BASF, unpublished results, 1978.
46. W. Hey et al., *Zentralbl. Arbeitsmed. Arbeitsschutz* **24** (1974) 71 – 77.
47. J. F. Finklea, National Institute for Occupational Safety and Health; Current Intelligence Bulletin no. 11, July 7, 1976.
48. J. McCann et al., *Proc. Nat. Acad. Sci. USA* **72** (1975) 979 – 983.
49. C. Heidelberger, *Mutat. Res.* **114** (1983) 283 – 385.
50. E. L. Evans et al. in P. J. de Serres, J. Ashby (eds.): *Progress in Mutation Research. Evaluation of Short-Term Tests for Carcinogenesis,* **vol. 1,** Elsevier, Amsterdam 1981, Chap. 49.
51. B. Kirkhart in F. J. de Serres, J. Ashby (eds.): *Progress in Mutation Research. Evaluation of Short-Term Tests for Carcinogenesis,* **vol. 1,** Elsevier, Amsterdam 1981, Chap. 67.
52. A. T. Natarajan et al. in F. J. de Serres, J. Ashby (eds.): *Progress in Mutation Research. Evaluation of Short-Term Tests for Carcinogenesis,* **vol. 1,** Elsevier, Amsterdam 1981, Chap. 50.
53. *IARC Monogr. Suppl.* **4** (1982) 118.
54. *IARC Monogr. Eval. Carcinog. Risk Chem. Man* **12** (1976) 77 – 84.
55. A. R. Sellakumar et al., *J. Environ. Pathol. Toxicol.* **4** (1980) 107 – 115.
56. BASF, unpublished results, 1967.
57. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentration (MAK) und Biologische Arbeitsstofftoleranzwerte (BA),* VCH Verlagsges., Weinheim 1985.

Carbazole

Gerd Collin, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany
Hartmut Höke, Weinheim, Federal Republic of Germany
Jörg Talbiersky, Dorsten, Federal Republic of Germany

1. **Physical Properties** 1
2. **Chemical Properties** 1
3. **Production** 1
4. **Analysis** 2
5. **Uses and Economic Aspects** 2
6. **Derivatives** 2
7. **Toxicology and Environmental Aspects** 3
8. **References** 3

Carbazole (dibenzopyrrole) [*86-74-8*], $C_{12}H_9N$, was first found in coal tar by C. Graebe and C. Glaser in 1872.

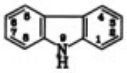

1. Physical Properties

M_r 167.21; *mp* 246 °C, *bp* at 101.3 kPa 354.8 °C; $d_4^{18} = 1.1035$ g/cm^3; plates or tables; sublimable; readily soluble in acetone and pyridine, slightly soluble in diethyl ether and ethyl alcohol, barely soluble in chloroform, acetic acid, carbon tetrachloride, and carbon disulfide, soluble in concentrated sulfuric acid, almost insoluble in water (ca. 1 mg/L at 25 °C); heat of fusion 176.3 kJ/kg; heat of combustion 3.719×10^4 kJ/kg at 25 °C.

2. Chemical Properties

The N-hydrogen atom of carbazole which is a secondary amine can be substituted with alkali metals. During halogenation, nitration, and sulfonation, substitution initially takes place at the 3- and 6-positions; the 1,3,6- and 1,3,6,8-derivatives are formed under more rigorous conditions.

Hydrogenation yields 1,2,3,4-tetrahydro-, hexahydro-, or dodecahydrocarbazole. Oxidation with chromates gives 3,3′-dicarbazyl, with permanganate 9,9′-dicarbazyl. Carboxylation with alkali and carbon dioxide gives, depending on the temperature, carbazole-3- or carbazole-1-carboxylic acid. Alkylation yields *N*-alkylcarbazoles.

3. Production

High-temperature coal tar contains an average of 0.9 % carbazole. Carbazole is obtained as a co-product in the production of anthracene. Due to its higher boiling point and better solubility, it can be separated from anthracene by extraction or complexation [6] with pyridine, ketones, benzene/methanol, *N*-methylpyrrolidone, dimethylacetamide, dialkyl sulfoxides, or dialkylformamides, by azeotropic distillation with ethylene glycol, or by selective hydrogenation of the anthracene [7] (→ Anthracene). The chemical separation of carbazole by means of potassium hydroxide or concentrated sulfuric acid fusion has become uneconomical. Pyridine mother liquors obtained from the preparation of pure anthracene yield pure carbazole by means of concentration and recrystallization from chlorobenzene.

Carbazole can be synthesized by converting cyclohexanone azine to octahydrocarbazole and subsequent dehydrogenation [8], by reductive cyclization of 2-nitrobiphenyl [9, 10], or by dehydrogenation and cyclization of diphenylamine [11 – 14], *o*-aminobiphenyl [15 – 17], or

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_059.pub2

N -cyclohexylideneaniline [18]. These synthetic processes are not yet of industrial significance, because sufficient quantities of carbazole can be prepared from coal tar. A modern method to produce carbazole from coal tar is solvent-free, high-temperature melt crystallization (→ Anthracene).

4. Analysis

Gas – liquid chromatography is suitable for the quantitative determination of carbazole in complex mixtures, such as in coal tar or coal-tar fractions [19]. Thin-layer chromatography can also be used to identify carbazole along with other compounds [20]. Purity is determined by UV spectroscopy. As little as 0.005 % carbazole in anthracene can be detected by phosphorescence spectroscopy. Small quantities of phenanthrene do not interfere with the results [21].

5. Uses and Economic Aspects

Carbazole is obtained from coal tar in amounts of more than 3500 t/a, primarily for dye synthesis.

The blue sulfur dye Hydron Blue [22] is commercially synthesized by condensing carbazole with *p*-nitrosophenol and subsequent sulfurization (→ Sulfur Dyes). Its 2-hydroxy-3-carboxylic acid amides (Naphtol AS-LB, -SG, -SR; → Azo Dyes), anthraquinone vat dyes from phthaloyl carbazoles (→ Anthraquinone Dyes and Intermediates), styryl dyes from *N*-ethylcarbazole (→ Methine Dyes and Pigments), dioxazine dyes and pigments from *N* -ethyl-3-aminocarbazole (→ Azine Dyes ; → Pigments, Organic), and reactive dyes [23] can also be synthesized from carbazole.

The insecticide Nirosan, 1,3,6,8-tetranitrocarbazole, has been commercially available since 1939. *N* -Vinylcarbazole is polymerized to give poly(vinylcarbazole) (Luvican, Polectron) [24], which is thermally and chemically stable and has a high softening temperature and low dielectric loss and photoconductivity. These properties make it useful in electrical industry and for electrostatic copying methods [24 – 28]. Methacrylate carbazole polymers can be used as photorefractive materials. These materials combine photoconductivity and electro-optic properties. Applications are high-density optical data storage, optical image processing, dynamic holography, and optical computing [29, 30]. Carbazole compounds can also be used as host materials for triplet emitters in light-emitting diodes [31, 32]. Carbazole reacts with phenols and formaldehyde in the presence of acidic catalysts to give novolaks. These can be cured with hexamethylenetetramine to form highly heat – resistant polymers [33]. Concrete plasticizers are obtained by co-condensation of carbazole with phenols and formaldehyde and subsequent sulfonation [34]. Thermal condensation of 3,6-diaminocarbazole with dicarboxylic acid gives elastic and thermally stable polyamides [35].

6. Derivatives

N-Ethylcarbazole [*86-28-2*], $C_{14}H_{13}N$; M_r 195.27; *mp* 68 °C; by ethylation of carbazole – potassium with diethyl sulfate [36].

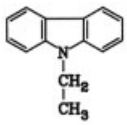

1,3,6,8-Tetranitrocarbazole [*4543-33-3*], $C_{12}H_5O_8N_5$; M_r 347.20; *mp* 312 °C; by nitration of carbazole with a mixture of nitric and sulfuric acid.

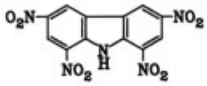

N-Vinylcarbazole [*1484-13-5*], $C_{14}H_{11}N$; M_r 193.25; *mp* 65 °C; from carbazole – potassium with ethylene oxide or vinyl chloride, or from carbazole and acetylene [24].

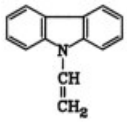

7. Toxicology and Environmental Aspects

Toxicology. The LD_{50} of carbazole ranges from 200 mg/kg (intraperitoneal, mice) to more than 5000 mg/kg (oral, rats) [37, 38]. Hence, it can be considered as nontoxic. Folliculitis and comedos were reported by workers with carbazole contact [39]. In animal studies, no skin- and eye-irritating potential was observed [40, 41], even under photoactivating conditions [42]. In rats and rabbits, carbazole is glucuronized and excreted in the urine [43]. There is some evidence for carbazole inducing carcinomas in the liver and forestomach of $B6C3F_1$ mice fed high doses of ca. 270 – 1050 mg (kg bw)$^{-1}$ d^{-1} for 96 weeks [44, 45], whereas no tumorigenic activity was noted in the liver and lung of newborn-mice following threefold intraperitoneal injection of carbazole [46]. Carbazole is not classifiable as to its carcinogenicity to humans (Group 3) [47]. Carbazole was not mutagenic in numerous bacterial assays [48 – 52]. No abnormalities in the intrauterine development of rats was observed following dermal administration of up to 250 mg (kg bw)$^{-1}$ from day 0 to 20 of gestation [52].

9-Ethylcarbazole produced a weak increase in the mutation rate in the bacterial tester strain *Salmonella typh.* TA100 after metabolic activation [49], but was inactive in a different complete Ames test [50].

Environmental Fate and Toxicity. Carbazole is biodegradable by adapted microorganisms in water [53, 54] and soil [55], but elimination of the dissolved organic carbon (DOC) to more than 90 % (according to OECD guideline 301E) is compatible with ready biodegradability [56]. In a model ecosystem including degradation and distribution processes, >99.5 % of initial carbazole disappeared from the water phase within 33 d, resulting in a spectrum of unidentified polar and nonpolar conversion products. N-methylation, N-acetylation, and presumably ring oxidation were relevant metabolic pathways in the test organisms (alga, snail, fish, and mosquito) which accumulated part of carbazole [57].

Photodegradation of carbazole in water is rapid; the elimination half-life was measured to be 2.9 h under simulated sunlight conditions [58].

The moderate water-partition coefficient (log P_{ow} = 2.58, 3.34 – 3.84 [56, 57, 59, 60]) indicates a significant bioconcentration potential.

However, bioconcentration factors (BCF) were found to be relatively low at ca. 125 and 500 in fish [57, 61] and about 110 and 115 in daphnia [62, 63].

Based on the molecular structure and photochemical properties, carbazole is predicted to be nonphototoxic to fish [64]. Exposure of daphnia to 0.11 mg/L for 48 h combined with 24-h continuous UV irradiation failed to produce mortality [62].

Acute aquatic toxicity was observed in fish and daphnia above water solubility: An approximate 48-h LC_{50} of about 1 mg/L is reported for saltwater fish (striped mullet) [65], and a 96-h LC_{50} for freshwater fish (zebra fish) of 4.31 mg/L [56]. In daphnia, the EC_{50} was >100 mg/L [56].

For the protozoa *Tetrahymena pyriformis*, the EC_{50} (72-h) was found to be nominally 20 mg/L [66].

High aquatic toxicity in the range of water solubility was found in algae (*Scenedesmus subspicatus*) with a mean inhibitory concentration IC_{50} (96-h) of 0.214 mg/L [56]. However, in another test using the alga *Scenedesmus acuminatus*, no toxic effect was noted up to 10 mg/L [67].

8. References

General References

1. *Beilstein*, **20**, 433;
   20(1), 162;
   20(2), 279;
   20(3), 3824.
2. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung*, Verlag Glückauf, Essen 1951, pp. 181–187.
3. H.-G. Franck, G. Collin: *Steinkohlenteer*, Springer Verlag, Berlin - Heidelberg - New York 1968, pp. 56–58 180–182.
4. *Erzeugnisse aus Steinkohlenteer*, Rütgerswerke, Frankfurt 1958.
5. J. A. Joule in: A. R. Katritzky (ed.), *Advances in Heterocyclic Chemistry*, Vol. 35, Academic Press, Orlando 1984, pp. 83–198.

Specific References

6. J. Polaczek, M. Jamorez, A.M. Szafranski, *Erdöl Kohle-Erdgas-Petrochem.* **41** (1988) 113.
7. Celanease GmbH and Rütgers VFT GmbH, DE 19757530, 1997 (P. Heymanns, K. Schalapski, E. Fuhrmann, J. Talbiersky).
8. BASF, DE 1158518, 1958 (M. Seefelder, H. Maisack).
9. Allied Chem. & Dye Corp., US 2508791, 1948 (M. S. Larrison).
10. M. Tashiro, JP-Kokai 80122760, 1980.
11. BASF, DE 937590, 1952 (H. Friedrich, O. Stichnoth, H. Waibel).
12. American Marietta, US 2921942, 1956 (H. M. Grotta).
13. Schering, DE 2418503, 1974 (B. Akermark, L. Eberson, E. Jonsson, E. Pettersson).
14. Bayer, DE 19633609, 1996 (H.-J. Buysch, A. Klausener, R. Langer, U. Notheis).
15. Monsanto, US 2479211, 1949 (C. Conover).
16. Ciba, DE 889592, 1951 (D. Porret).
17. H. Suhr, U. Schöch, G. Rosskamp, *Chem. Ber.* **104** (1971) 674–676.
18. Snam Progetti, NL 7115166, 1971 (M. M. Mauri, P. A. Moggi, U. Romano).
19. H. D. Sauerland, *Brennst. Chem.* **45** (1964) 55–56.
20. E. Sawicki, H. Johnson, K. Kosinski, *Microchem. J.* **10** (1966) 72–102.
21. M. Zander, *Angew. Chem.* **76** (1964) 922; *Angew. Chem. Int. Ed. Engl.* **3** (1965) 755.
22. Leopold Cassella & Co GmbH, DRP 218371, Kl.22d, 29, 1908.
23. ICP, Warschau, DE 4303359 A1, (J. Pielichowski, R. Chrzaszcz, J. Polaczek, S. Galka).
24. J. M. Pearson, M. Stolka: *Poly(N-vinylcarbazole)*, **Polymer Monographs no. 6**, Gordon and Breach Science Publishers, New York - London - Paris 1981.
25. S. Sandler, W. Karo: "Polymerisation Reactions of N-Vinyl carbazole and Related Monomers", *Polymer Synthesis 3*, Vol. 3, 2nd ed., Academic Press, 1996.
26. W. Klöpffer, *Kunststoffe* **61** (1971) 533–539.
27. U. König, V. Stepanek, *Kunststoffe* **69** (1979) 223–227.
28. Xerox Corporation, US 5055366, 1991 (R. C. U. Yu, R. P. Premo)
29. K. Diduch, M. Wübbenhorst, S. Kucharski, *Synth. Met.* **139** (2003), 515–520.
30. J. Sanetra, D. Bogdal, M. Warzala, *A. Boron, Chem. Mat.* **14** (2002) 89–95.
31. B. K. Runner et al., *J. Am. Chem. Soc.* **126** (2004) 6035
32. A. van Dijken et al., *J. Am. Chem. Soc.* **126** (2004) 771.
33. Rütgerswerke, DE 2033015, 1970 (J. Omran, H.-G. Franck, M. Zander).
34. Rütgerswerke, DE 3210458, 1982 (J. Omran, M. Zander).
35. Rütgerswerke, DE 2125128, 1971 (J. Omran, H.-G. Franck, M. Zander).
36. Farbwerke Hoechst, DE 2132961, 1971 (T. Papenfuhs).
37. N. J. Sax: *Dangerous Properties of Industrial Materials*, Van Nostrand Reinhold Comp., New York - London 1979, p. 468.
38. E. Eagle, A. J. Carlson, *Pharm. Exp. Ther.* **99** (1950) 450–57.
39. L. Jirásek, Z. Stáva, *Prakt. Lek.* **35** (1955) 34–37.
40. HRC: unpublished report no. 332b, Ruetgerswerke AG, 1979.
41. HRC: unpublished report no. 332c, Ruetgerswerke AG, 1979.
42. I. E. Kochevar, R. B. Armstrong, J. Einbinder et al., *Photochem. Photobiol.* **36** (1982) 65–69.
43. S. R. Johns, S. E. Wright, *J. Med. Chem.* **7** (1964) 158–161.
44. H. Tsuda, A. Hagiwara, M. Shibata et al., *J. Nat. Cancer Inst.* **69** (1982) 1383–1387.
45. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans*, **32** (1983) 239–245.
46. E. H. Weyand, J. Defauw, C. A. McQueen et al., *Food Chem. Toxicol.* **31** (1993) 707–715.
47. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans*, **71** (1999) 1319–1323.
48. D. Anderson and J. A. Styles, *Br. J. Cancer* **97** (1978) 924–930.
49. E. J. LaVoie, G. Briggs, V. Bedenko, D. Hoffmann, *Mut. Res.* **101** (1982) 141–150.
50. I. Florin, L. Rutberg, M. Curvall, C. R. Enzell, *Toxicology* **18** (1980) 219–232.
51. D. A. Kaden, R. A. Hites, W. G. Thilly, *Cancer Res.* **39** (1979) 4152–4159.
52. S. M. Dutson, G. M. Booth, G. B. Schaalje et al., *Environ. Toxicol. Chem.* **16** (1997) 2113–1117.
53. J. M. Foght, D. L. Gutnick, D. W. S. Westlake, *Appl. Environ. Microbiol.* **55** (1989) 36–42.
54. J. G. Mueller, D. P. Middaugh, S. E. Lantz, P. J. Chapman, *Appl. Environ. Microbiol.* **57** (1991) 1277–1285.
55. R. J. Grosser, D. Warshawsky, J. R. Vestal, *Environ. Toxicol. Chem.* **14** (1995) 375–382.

56. T. Knacker, H.-J. Schallnaß, U. Klaschka, J. Ahlers, *J. ESPR Environ. Sci. Pollut. Res.* **2** (1995) 179–187.
57. P. Lu, R. L. Metcalf, E. M. Carlson, *Environ. Health Perspec.* **24** (1978) 201–208.
58. T. Mill, M. R. Mabey, B. Y. Lan, A. Baraze, *Chemosphere* **10** (1981) 11281–11290.
59. C. Hansch, A. Leo, *Substituents Constants for Correlation in Chemistry and Biology*, Wiley, New York, 1979, App. II., p. 274.
60. P. de Voogt, G. A. van Zijl, H. Govers, U. A. T. Brinkman, *J. Planar Chromatogr.* **3** (1990) 24–33.
61. P. de Voogt, B. van Hattum, P. Leonards et al., *Aquatic Tox.* **29** (1991) 169–194.
62. J. L. Newsted and J. P. Giesy, *Environ. Toxicol. Chem.* **6** (1987) 445–461.
63. G. R. Southworth, J. J. Beauchamps, P. K. Schmieder, *Bull. Environ. Contam. Toxicol.* **23** (1979) 73–78.
64. J. T. Oris and J. P. Giesy, *Chemosphere* **16** (1987) 1395–1404.
65. P. Thomas, H. W. Wofford, J. M. Neff, *Aquatic Toxicol.* **1** (1981) 329–342.
66. T. W. Schultz, M. Cajina-Quezada, J. N. Dumont, *Arch. Environ. Contam. Toxicol.* **9** (1980) 591–598.
67. P. L. A. van Vlaardingen, W. J. Steinhoff, P. de Voogt, W. A. Admiraal, *Environ. Toxicol. Chem.* **15** (1996) 2035–2042.

Carbides

Boron Carbide, Boron Nitride, and Metal Borides; Calcium Carbide; and Silicon Carbides are separate keywords.

HELMUT TULHOFF, Hermann C. Starck Berlin, Werk Goslar, Goslar, Federal Republic of Germany

1. Survey 1
1.1. Saltlike Carbides 1
1.2. Metal-like Carbides 3
1.3. Diamond-like Carbides 3
1.4. Carbides of Nonmetallic Elements . . 3
1.5. Crystal Structure 3
1.6. General Production Processes 4
1.7. Uses 5
2. Metal-like Carbides of Industrial Importance 5
2.1. Tungsten Carbide 5
2.2. Titanium Carbide 9
2.3. Tantalum Carbide 10
2.4. Niobium Carbide 11
2.5. Zirconium Carbide 12
2.6. Hafnium Carbide 12
2.7. Vanadium Carbide 13
2.8. Chromium Carbide 13
2.9. Molybdenum Carbide 14
3. Mixed Carbides 15
3.1. Tungsten – Titanium Carbide 16
3.2. Other Mixed Carbides 16
3.3. Carbonitrides 17
3.4. Mixed Carbonitrides 17
4. Carbides of the Iron Group and Manganese 17
5. Complex Carbides 18
6. References 18

1. Survey

Most of the elements form binary compounds with carbon, all of which can be called carbides. The properties of these carbides are very different; therefore, like binary hydrides and nitrides, the carbides should be classified into groups. To avoid too many subdivisions, the following four types of carbides may be defined:

1) saltlike carbides of metallic elements, e.g., CaC_2
2) metal-like carbides of metallic elements, e.g., WC
3) diamond-like carbides, e.g., B_4C
4) carbides of nonmetallic elements, e.g., CO

This classification suggests another group: the elements that do not react with carbon, e.g., Sn.

Generally, the four groups of carbides can not be strictly separated from each other. Numerous carbides are in intermediate positions between these groups. One example is BeC_2 [*57788-94-0*]. It is a typical saltlike carbide and is decomposed by water. On the other hand, it may be viewed as a diamond-like carbide because of its hardness and other properties resembling those of SiC.

Figure 1 surveys the four types of carbides in the form of a periodic table. Elements that do not form binary compounds with carbon, or are not known to form carbides, are not shown. The carbides of the iron group and manganese are a subgroup of the metal-like carbides.

1.1. Saltlike Carbides

Saltlike carbides of metallic elements are the carbides of the elements of groups 1 – 3 and 11 – 13 (I – III, both A's and B's) of the periodic table, the lanthanides and actinides included. Exceptions are Ga, In, and Tl, which do not form carbides, and B_4C, which is a typical diamond-like carbide.

The saltlike carbides – also called ionic carbides – are attacked by water to form hydrocarbons. Most of these carbides form acetylene, e.g.:

$$CaC_2 + 2\,H_2O \longrightarrow Ca(OH)_2 + C_2H_2$$

These carbides can be viewed as salts of acetylene and may be called acetylides. The crystals contain C_2^{2-} anions.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_061

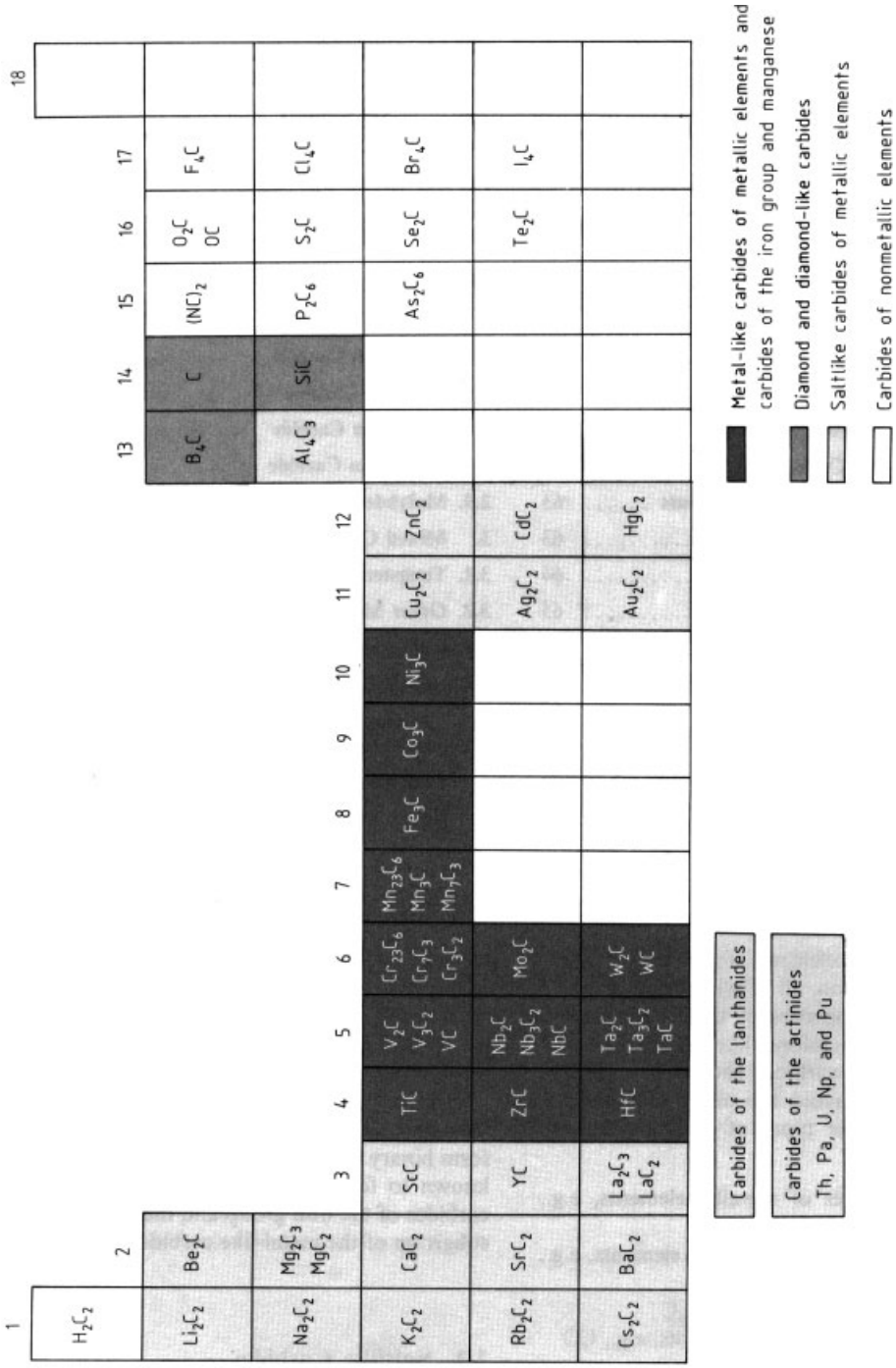

Figure 1. Survey of binary compounds of carbon with the elements

The carbides Be_2C and Al_4C_3 form pure methane when hydrolyzed:

$$Be_2C + 4\,H_2O \longrightarrow 2\,Be(OH)_2 + CH_4$$

$$Al_4C_3 + 12\,H_2O \longrightarrow 4\,Al(OH)_3 + 3\,CH_4$$

In the crystal lattice of these carbides, the carbon atoms are isolated from each other, in contrast to the C_2 groups of the acetylides. The Be_2C lattice is antiisotypical to that of CaF_2.

The carbide MgC_2 can be decomposed by heating to form Mg_2C_3 and graphite. Hydrolysis of Mg_2C_3 yields propyne:

$$Mg_2C_3 + 4\,H_2O \longrightarrow 2\,Mg(OH)_2 + CH_3{-}C{\equiv}CH$$

In their carbides the lanthanides and actinides are mainly divalent. During hydrolysis they become trivalent, and hydrogen is formed in this reaction:

$$M^{2+} + H^+ \longrightarrow M^{3+} + H$$

This hydrogen reacts with the acetylene also formed to produce a mixture of acetylene, methane, ethylene, and hydrogen.

Whereas the saltlike carbides of groups 1 and 2 are transparent and are not electrical conductors, the lanthanide and actinide carbides show some metallic behavior, an indication of a state intermediate between saltlike and metal-like carbides. The electrical conductivity and metallic luster may be due to the fact that the metals are divalent in their carbides and the third valence electron is available for metallic bonding.

One other subgroup of saltlike carbides should be mentioned: the alkali-metal – graphite compounds. They are formed by absorption of molten Na, K, Rb, and Cs by graphite. Compositions such as MC_8, MC_{16}, and MC_{60} are known. These compounds are quite likely not chemical compounds, but merely adsorptional compounds, and perhaps better not called carbides.

1.2. Metal-like Carbides

Metal-like carbides of metallic elements are the carbides of the transition elements of groups 4, 5, and 6 of the periodic table. These carbides, also called metallic carbides, are not attacked by water. The metallic character of these compounds is shown in their high thermal and electrical conductivity as well as in their metallic luster.

All the metallic carbides are stable at room temperature and resist attack by dilute acids as well as by alkaline and organic liquids. Their hardness and wear resistance are utilized in the cemented carbides (→ Hard Materials), which are sintered products of the carbides with cobalt or other metals. Because of their industrial significance, these carbides are described in more detail in Chapter 2.

The carbides of Mn, Fe, Co, and Ni are generally included in the metal-like carbides, although they are really better classified as a group on their own. These carbides are in an intermediate position between the metal-like carbides and the saltlike carbides. Their crystal structures are quite different from the structures of the metal-like carbides and the saltlike carbides. The pure compounds are attacked by water or dilute acids.

1.3. Diamond-like Carbides

Diamond-like carbides include, strictly speaking, only B_4C and SiC. They are called diamond-like because of their extreme hardness, which is exceeded only by diamond itself. Sometimes the very hard Be_2C is included in the diamond-like carbides. However, its hardness cannot be used industrially, because of its decomposition by water.

1.4. Carbides of Nonmetallic Elements

Such carbides as CO, CS_2, and CCl_4, the carbides of nonmetallic elements, have covalent, molecular character and are not discussed in this article.

1.5. Crystal Structure

The lattice structure of most carbides can be deduced from the structure of their most important group, the metal-like transition-metal carbides. Basically these carbides are cubic or hexagonal closest packings of metal atoms with the smaller carbon atoms in the interstitial sites. Therefore,

the transition-metal carbides can also be called interstitial carbides.

In 1931 HÄGG [7] reported that the structure of the transition-metal carbides is determined by the radius ratio r of the metal and carbon atoms $r = r_C/r_{metal}$. When r is less than 0.59, the metals form the simple structures just described, with the carbon atoms located at the octahedral interstices. If all interstices are occupied in a body-centered cubic (bcc) metal lattice, the result is the face-centered cubic (fcc) sodium chloride structure. All the carbides of transition-metal groups 4 and 5 crystallize in this B1 lattice. Tungsten carbide has a simple hexagonal structure with all of the trigonal prismatic interstitial sites occupied by carbon.

The B1 carbides, principally TiC, ZrC, HfC, and VC, tend to form defect structures in which the interstitial sites are not completely filled. Broad homogeneity ranges are the result, but some substructures with overlapping homogeneity ranges are indicated [8].

When only one-half of the octahedral interstitial sites are occupied in an hexagonal-closest-packed (hcp) metal structure, the subcarbides – V_2C, Nb_2C, Ta_2C, Mo_2C, and W_2C – are obtained. This is a simplified interpretation, and in fact the subcarbides are more complex structures, as was shown by NOWOTNY [9]. Indeed, these structures are sometimes called Nowotny phases, to contrast them with the simpler Hägg phases.

When Hägg's ratio exceeds 0.59, the simple phases can no longer be formed as before. Close to 0.59 and in the case of low carbon content, there are the compounds $Cr_{23}C_6$ and $Mn_{23}C_6$, which can still be viewed as interstitial structures. For higher values of r and higher carbon content, more complex structures, no longer interstitial compounds, are formed: M_3C, M_7C_3, M_3C_2. These stoichiometries are primarily found in the iron group. These more complex structures are less metallic than the Hägg phases. Hardness, melting point, and chemical resistance are markedly lower.

The structures of the saltlike carbides can also be deduced from the Hägg phases. When there are more carbon atoms than octahedral interstitial sites in the metal lattice, pairs of carbon atoms are formed. The CaC_2 type is a tetragonally deformed B1 structure. The dicarbides of the lanthanides and actinides crystallize in this system. They lack metallic characteristics. The bcc carbides, M_2C_3, also contain C_2 groups, e.g., U_2C_3.

The structures of the diamond-like carbides SiC and B_4C differ from all structures described thus far. The carbide SiC has an expanded diamond lattice, whereas B_4C crystallizes in a rhombic lattice containing B_{12} icosahedrons and C_3 chains.

1.6. General Production Processes

There are a number of general methods of producing carbides:

1) Nearly all carbides can be prepared at high temperature by direct reaction from the metal powder mixed with lamp black or graphite, e.g.:

   $$W + C \longrightarrow WC$$

   Generally the temperature is in the range 1000 – 1500 °C, and special furnaces are used. A protective atmosphere or vacuum is needed.
2) Instead of the pure metal, the oxide or hydride can be carburized with solid carbon:

   $$Ta_2O_5 + 7\,C \longrightarrow 2\,TaC + 5\,CO$$

   Large amounts of gas result from this reaction. Both processes 1 and 2 are solid-state reactions.
3) Carbides with high melting points can be prepared by a modified aluminothermic process:

   $$3\,Cr_2O_3 + 6\,Al + 4\,C \longrightarrow 2\,Cr_3C_2 + 3\,Al_2O_3$$
4) Instead of solid carbon, gaseous carbon compounds, such as CO or CH_4, can be used. This process is important in the steel industry, where mainly iron, chromium, and manganese carbides are formed during fusion:

   $$3\,Fe + 2\,CO \longrightarrow Fe_3C + CO_2$$
5) Reaction of metal chlorides with hydrocarbons in a hydrogen atmosphere produces carbides:

   $$ZrCl_4 + CH_4 \longrightarrow ZrC + 4\,HCl$$

   This method is used to produce layers of carbides on other materials by chemical vapor deposition (CVD → Thin Films).

Most of the saltlike carbides are prepared by processes 1 and 2, by heating the metals or oxides with carbon, e.g.:

$$CaO + 3\,C \longrightarrow CaC_2 + CO$$

However, another, quite different process can also be used:

Reaction of acetylene with metals or salts dissolved in water or liquid ammonia or suspended in inert organic fluids forms simple or complex acetylides. Most are metastable but are stabilized by H_2O, NH_3, or acetylene itself. Cautious decomposition produces individual carbides:

$$Li\,(s) + NH_3\,(l) + C_2H_2 \longrightarrow LiHC_2 \cdot NH_3 + 0.5\,H_2$$

$$LiHC_2 \cdot NH_3 \longrightarrow LiHC_2 \longrightarrow Li_2C_2$$

Today this last method is primarily of laboratory interest.

1.7. Uses

The uses of carbides are as diverse as the types of carbides. Most important from an economic point of view are the carbides of the transition metals of groups 4, 5, and 6. The most important of these is tungsten carbide. The hardness and chemical resistance of these carbides are the basis for their use by the tool and machine industry in multitudinous applications as *cemented carbides.* The terms *hardmetals* and cemented carbides are synonymous. The term *cermet* is also used for some or all of these composites. Furthermore the term *tungsten carbide* is used because WC is the main constituent in most of these materials. These cemented carbides are sintered products of one or more carbides with a metallic binder, preferably the metal cobalt. There are many different combinations of carbides and binder metals. Factories producing cemented carbides are found in every industrial country; the world's annual production is estimated at ca. 20 000 t in 1985. Some tungsten carbide combined with Cu or Ag is used in electrical contacts and in fuel cells.

The carbides of manganese and iron are never used alone like the harder transition-metal carbides, but rather are formed in alloys during fusion. These carbides, especially cementite, Fe_3C, are of fundamental importance because the individual carbides and the binary mixed carbides with V, Cr, Mo, and W are responsible for the hardness of steel, Stellites, and related alloys.

The most important saltlike carbide is CaC_2 (→ Calcium Carbide). One-half of the world's annual production, several million tons, is converted to cyanamide (→ Cyanamide), which is used as fertilizer. Some 20 % is used for acetylene production (→ Acetylene, Chap. 4.3.4.), and the remainder is used in steelmaking as a carburizing additive (→ Steel).

The monocarbides and dicarbides of uranium and thorium are used as nuclear fuels in high-temperature reactors. These carbides are not used as hard materials, although they do have some metallic character. Other saltlike carbides do not have industrial importance.

The carbides B_4C and SiC are used in large quantities as abrasives (→ Abrasives; → Boron Carbide, Boron Nitride, and Metal Borides, → Silicon Carbide). Heating elements and many heat-resistant parts are made from SiC.

2. Metal-like Carbides of Industrial Importance

The important individual carbides of the transition metals and the mixed carbides of these metals are described in detail in the following. Tungsten carbide, WC, because of its importance in cemented carbides, or hardmetals, is described first. Thereafter, TiC, TaC, and NbC, which are also basic hard carbides, are described. The physical properties of these four carbides are given in Table 1. Finally, the carbides of Zr, Hf, V, Cr, and Mo are described. These last carbides are used only as additives in cemented carbides and have less importance. Their physical properties are given in Table 2.

2.1. Tungsten Carbide

There are two hexagonal carbides in the tungsten – carbon system (Fig. 2): the monocarbide, WC, and the subcarbide [*12070-13-2*], W_2C. The hexagonal WC, also called α-WC, decomposes at its incongruent melting point of 2776 °C. Its range of homogeneity is extremely narrow: from 49.5 to 50.5 mol % C.

Table 1. Physical properties ∗ of WC, TiC, TaC, and NbC

Property	WC	TiC	TaC	NbC
CAS Registry Number	[*12070-12-1*]	[*12070-08-5*]	[*12070-06-3*]	[*12069-94-2*]
Molecular mass M_r	195.87	59.91	192.96	104.92
Carbon content (theory), wt %	6.13	20.05	6.23	11.45
Crystal structure	hex., Bh	fcc, B1	fcc, B1	fcc, B1
Lattice constants, pm	*a* 291 *c* 284	432.8	445.5	447.0
Density, g/cm^3	15.7	4.93	14.48	7.78
Melting point *mp*, °C	2776	3067	3985	3610
Microhardness, kg/mm^2	1200 – 2000	≈ 3000	1800	2000
Transverse rupture strength, MPa	550	240 – 390	350 – 400	300 – 400
Modulus of elasticity, GPa	696	451	285	338
Specific heat, J mol^{-1} K^{-1}	39.8	47.7	36.4	36.8
Heat of formation ΔH_{298}, kJ/mol	− 40.5	− 183.7	− 148.9	− 141.0
Coefficient of thermal conductivity, W m^{-1} K^{-1}	121	21	22	14
Coefficient of thermal expansion, 10^{-6} K^{-1}	*a* 5.2 *c* 7.3	7.74	6.29	6.65
Electrical resistivity, μΩ · cm	22	68	25	35
Superconductive transition temperature, K	10.0	1.15	10.3	11.1
Hall constant, 10^{-4} cm^3 A^{-1} s^{-1}	− 21.8	− 15.0	− 1.1	− 1.3
Magnetic susceptibility, 10^{-6} emu/mol	+ 10.0	+ 6.7	+ 9.3	+ 15.3

∗ Properties given without a temperature are for room temperature.

Table 2. Physical properties ∗ of ZrC, HfC, VC, Cr_3C_2, and Mo_2C

Properties	ZrC	HfC	VC	Cr_3C_2	Mo_2C
CAS Registry Number	[*12020-14-3*]	[*12069-85-1*]	[*12070-10-9*]	[*12012-35-0*]	[*12069-89-5*]
Molecular mass M_r	103.23	190.51	62.96	180.05	203.91
Carbon content (theory), wt %	11.64	6.30	19.08	13.33	5.89
Crystal structure	fcc, B1	fcc, B1	fcc, B1	orthorh., $D5_{10}$	hex. L′3
Lattice constants, pm	469.8	464.8	416.5	*a* 1147 *b* 554 *c* 283	*a* 300 *c* 473
Density, g/cm^3	6.46	12.3	5.36	6.68	9.18
Melting point *mp*, °C	3420	3930	2650	1810	2520
Microhardness, kg/mm^2	2700	2600	2900	1400	1500
Modulus of elasticity, GPa	348	352	422	373	533
Specific heat, J mol^{-1} K^{-1}	37.8	37.4	32.3	32.7	30.3
Heat of formation ΔH_{298}, kJ/mol	−196.8	−209.6	−100.8	−94.2	−46.0
Coefficient of thermal conductivity, W m^{-1} K^{-1}	20.5	20.0	38.9	19.1	21.5
Coefficient of thermal expansion, 10^{-6} K^{-1}	6.73	6.59	7.2	10.3	7.8
Electrical resistivity, μΩ · cm	42	37	60	75	71
Superconductive transition temperature, K	>1.2	>1.2	>1.2	>1.2	2.78
Hall constant, 10^{-4} cm^3 A^{-1} s^{-1}	−9.42	−12.4	−0.48	−0.47	−0.85
Magnetic susceptibility, 10^{-6} emu/mol	−23	−25.2	+26.2	–	–

∗ Properties given without a temperature are for room temperature.

The subcarbide, W_2C, probably has three modifications. The highest temperature modification melts without decomposition at ≈ 2800 °C. The "eutectic" of W_2C and WC is known as cast or fused tungsten carbide. The solid eutectic mixture is sometimes incorrectly called W_2C. In addition to these two industrial carbides, there is a substoichiometric face-centered cubic WC_{1-x} phase, also called β-WC, which is unstable at room temperature, forming only above 2530 °C. The phase diagram is made more complicated by this additional compound and the W_2C modifications, mainly in the high-temperature range. Although many have investigated the tungsten – carbon system, unanswered questions remain.

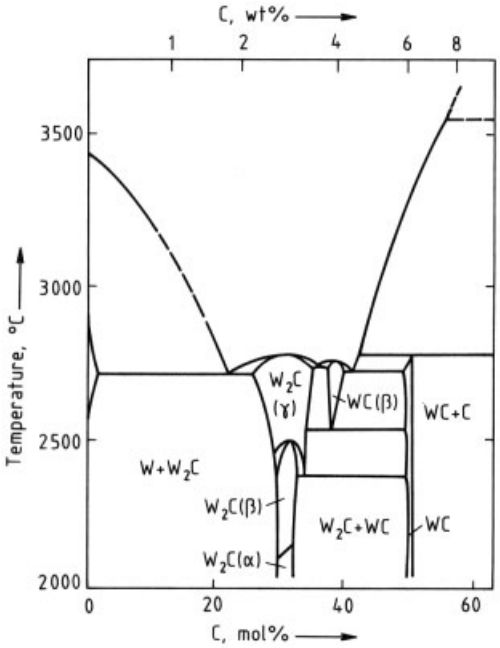

Figure 2. Tungsten – carbon phase diagram

Properties. Commercial monocarbide, WC, the raw material for the powder-metallurgy industry, is a gray metallic powder. Its average grain size is between 0.5 and 20 µm. In addition, very small quantities are prepared with smaller or coarser size for special applications. The carbide is insoluble in water and dilute acids, but is dissolved by hot mixtures of HNO_3 and HF. It is oxidized in air above 600 °C. Although it is stable in dry hydrogen up to its melting temperature, wet hydrogen decarburizes it. Chlorine attacks WC above 400 °C, while fluorine attacks WC at room temperature.

Hardness, combined with high modulus of elasticity, is the most important mechanical property of WC. The microhardness is anisotropic [10], [11] and, because of this, values ranging between 1000 and 2500 kg/mm^2 can be found in the literature.

The chemical resistance of the subcarbide W_2C is less than that of WC. The subcarbide is dissolved by HNO_3 – HF mixtures even at room temperature. It may be distinguished from WC by its reaction with alkaline potassium hexacyanoferrate (Murakami's reagent): W_2C turns yellow to brown, whereas WC remains gray. The microhardness of the subcarbide is higher than that of WC, but W_2C is not used alone industrially because it is too brittle.

Because eutectic W_2C – WC is prepared by a fusion process, it is not produced as a powder in the micron range. The grains are much coarser, up to several millimeters. The carbon content can vary from 3.5 to 4.5 wt %, corresponding to 50 – 90 % of W_2C or 10 – 50 % of WC in this "eutectic".

Preparation. Most of the world's annual production of 15 000 – 20 000 t of WC is made by direct carburization of tungsten metal with carbon. Mixtures of metal and lamp black, or even graphite, are heated to temperatures between 1400 and 2000 °C in a hydrogen atmosphere or vacuum. Electrical walking-beam or pusher-type furnaces or gas-fired muffle furnaces are used. Carbon tube furnaces are needed for the high-temperature range, and batch-type induction furnaces are needed for vacuum processing.

After purity, the most important property of the carbide is grain size, because the grain size significantly affects the mechanical properties of WC products. Fine-grained powders cannot be made from coarser powders only by milling. Intensive milling changes carbon and oxygen contents, the shape of the grains, and the grain size distribution. Therefore, the grain size is better determined by the processing parameters during reduction and carburization: temperature, reaction time, humidity, flow rate of the hydrogen, and several other factors. Most importantly, the grain size of the starting material must be selected to produce the desired end product. Generally, powders become coarser when converted from oxide to metal to carbide.

The chemical nature of the starting materials and the intermediate steps also affect the physical properties of the final carbide. Possible starting materials are tungstic acid [*7783-03-1*], H_2WO_4, and ammonium paratungstate or APT [*11120-25-5*], $(NH_4)_{10}W_{12}O_{41} \cdot 5H_2O$. The intermediates are yellow oxide (WO_3), blue oxide (W_4O_{11}, simplified), and brown oxide (WO_2). Various ways of processing are illustrated in Figure 3 by the flow sheet, which contains 12 different production lines.

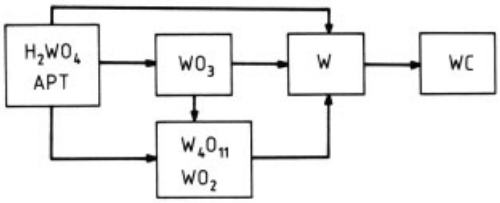

Figure 3. Various production lines leading to WC
APT stands for ammonium paratungstate

The following are typical production lines:

1) Fine tungstic acid powder is reduced directly to metal by dry hydrogen at 750 °C. Metal particles with an average size of 0.7 – 0.8 µm are obtained. These are carburized at 1400 °C to produce 1-µm carbide.
2) Ammonium paratungstate is calcined at 700 °C in a stream of nitrogen to give blue oxide, which is reduced at 800 °C. The metal is carburized at 1400 °C to produce carbide of 2 – 5 µm.
3) Ammonium paratungstate is roasted at 800 °C in air to produce WO_3. This is reduced by wet hydrogen at 950 °C. Carburizing at 1600 °C produces carbide of 10 – 20 µm.

Tungsten oxides can also be carburized directly with carbon, e.g.,

$$WO_3 + 4\,C \longrightarrow WC + 3\,CO$$

In this case, the intermediate metal step is omitted. The disadvantage is the difficulty in obtaining the correct carbon content in the carbide, since the CO itself reacts with oxide to form metal to a degree that cannot be calculated. Therefore, this process is used only to produce technical grades. Tungstic acid, ammonium paratungstate, or even scheelite (natural or artificial) can be treated in the same way.

$$CaWO_4 + 4\,C \xrightarrow{1800\,^\circ C} CaO + 3\,CO + WC$$

However, when scheelite is heated with carbon, the resulting cake must be leached with acid to isolate the carbide, and this carbide is of low quality because of its high level of impurities.

Tungsten metal can be carburized by carbon-containing gases, usually carbon monoxide or methane. Gas-phase carburization is done preferably in the temperature range 800 – 900 °C. Therefore, the grains do not become much coarser. For example, WC of 0.3 – 0.4 µm is obtained from 0.3-µm metal. Such fine carbides are often called submicron carbides. Tungsten oxide can also be reduced and carburized in one step by CO or CH_4, but the product is always slightly deficient in combined carbon. When CO is used, water is not present in the furnace atmosphere as a byproduct of reduction. Because grain growth of tungsten during reduction is induced mainly by water vapor and high temperature, extremely fine WC can be made. These powders can be used as catalysts in fuel cells.

Another method of preparing fine WC is the reaction of tungsten metal or oxide with CH_4 and H_2 in a plasma reactor [12]. Carbide having a grain size of 0.1 µm or less, sometimes called ultrafine carbide, is obtained. The plasma technology and the use of such ultrafine WC are still being developed.

Numerous other ways to prepare WC have been developed. Some are modifications of the processes just described; others are entirely different. Most still need to be improved and are not yet in use on an industrial scale:

1) In the Axel – Johnson process, tungsten ore, ferrotungsten, or tungsten scrap is treated with chlorine to form WCl_6, which is reduced by H_2 in a gas-phase reaction. The fine metal formed by this reaction is carburized by a conventional process [13].
2) A mixture of WO_3 and carbon is heated in a two-stage rotary furnace. In the first stage the oxide is reduced to metal in a stream of nitrogen, and in the second stage WC is formed at higher temperature in a stream of hydrogen [14].
3) A mixture of WO_3, Co_3O_4, and carbon is reduced in H_2. After carburization, the mixture of WC and cobalt metal can be sintered directly to cemented carbides. The WC grains have a uniform, fine size, which is a result of the coreduction of the oxides [15].

All methods of preparation described thus far are solid-state or gas-phase reactions. The preparation of cast tungsten carbide is the only method involving fusion. A mixture of tungsten metal and carbon or tungsten carbide is heated in a carbon tube or high-frequency furnace to ca. 2800 °C. The molten eutectic is quenched in

water or otherwise cooled rapidly to produce a fine crystalline structure.

Uses. Tungsten carbide is by far the carbide most used in cemented carbides: About 90 % of the world's production of carbide tools are tungsten carbide-based sinter alloys. Of these, 50 % are so-called straight grades, tungsten carbide – cobalt products consisting of 70 – 95 wt % of tungsten carbide. There are many, many uses for cemented carbides of various compositions. The greatest demand is for cutting and drilling tools, mining machinery, and wear-resistant parts of all kinds. Some examples illustrate the broad field of applications: milling cutters, cutting tips and drills, sawing teeth and blades, drawing and heading dies, rolls, nozzles, sealing rings, balls for ball mills, balls for ballpoint pens, tire studs, and even scratch-proof watchcases. Protective surface coatings are made from cast tungsten carbide.

The mechanical properties of the cemented carbides depend primarily on the grain size of the tungsten carbide. Generally speaking, smaller grain sizes produce greater hardness but lower crack resistance. The cobalt content also affects the mechanical properties, and the properties of the cemented carbide can be adjusted to the mechanical requirements over a wide range.

Toxicology. Tungsten carbide and the other carbides of the transition metals are not known to be toxic in themselves. However, nearly all of these carbides are used in combination with cobalt metal, and cobalt dust is carcinogenic. Therefore, mixtures for the powder-metallurgical preparation of cemented carbides are classified as dangerous materials in some countries.

2.2. Titanium Carbide

The face-centered cubic monocarbide TiC is the only carbide in the titanium – carbon system. It melts without decomposition at ca. 3000 °C. Its range of homogeneity is very broad, ranging from 35 mol % to just below 50 mol % carbon. The composition with the highest melting point and the largest lattice constant contains less carbon than stoichiometric TiC (see Fig. 4). Because of this, undesired low-carbon phases cannot be formed during sintering of cemented carbides when TiC is present, unlike the case for the straight WC – Co grades. Titanium carbide forms solid solutions with all other cubic transition-metal carbides of groups 4 and 5. In addition, it is the host lattice for hexagonal WC in the most important industrial solid-solution carbide, (W,Ti)C (see Section 3.1).

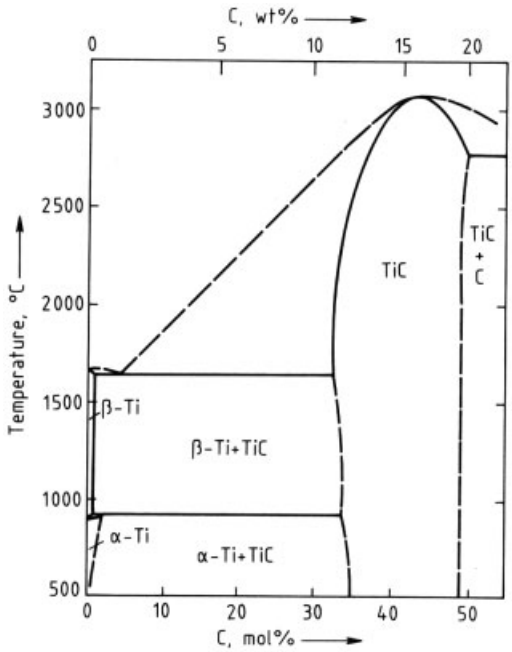

Figure 4. Titanium – carbon phase diagram [3]

At elevated temperatures, TiC and Ti metal react with oxygen and nitrogen to form TiO and TiN, the structures of which are isotypical to the lattice of TiC. Therefore, many TiC powders contain small amounts of N and O, to an extent of 1 % or more, and may be viewed as Ti(C, N, O) mixed crystals.

Commercial TiC is a gray powder usually having an average grain size of 2 – 10 µm. It is very resistant to acids, oxidation, and heat. However, it is dissolved easily by mixtures of HNO_3 and HF. In hydrogen it can be heated to its melting point without decomposition. Titanium carbide is the hardest of all the transition-metal carbides.

Preparation. Most commercial TiC is made by the reaction of TiO_2 with carbon. Intimate mixtures of pure TiO_2 and carbon are heated

to 2000 °C or above in a hydrogen atmosphere. Large quantities of CO are produced. After the resulting cake is milled, the material contains up to 1 wt % each of free carbon, nitrogen, and oxygen. The amounts of these elements must be reduced in a second step, usually a vacuum heating process. Nitrogen and oxygen contents are decreased to less than 0.1 wt % each. The free carbon content is usually in the range 0.2 – 0.4 wt %, and the combined carbon is 19.5 wt % max., somewhat less than the stoichiometric content of 20.05 wt %.

Titanium metal can be carburized with carbon. Titanium sponge, or even finely divided scrap, is used. The process is exothermic, and therefore, exact temperature control is not possible. As a result, the cake is sometimes extremely dense and can merely be broken down or milled only with difficulty. Carbon, nitrogen, and oxygen contents must be adjusted in a subsequent process.

Very coarse, comparatively pure TiC is prepared by the auxiliary metal bath technique [16]. Titanium metal, ferrotitanium, or even titanium alloy scrap is dissolved along with graphite in molten metal, preferably iron or nickel. After this mixture is cooled, TiC is isolated by dissolving the auxiliary metal with a nonoxidizing acid (menstruum process).

Extremely fine TiC is made in a plasma reaction of $TiCl_4$, H_2, and CH_4.

Layers of TiC on other materials may be produced by controlled vapor deposition from a mixture of $TiCl_4$, H_2, and CH_4.

Uses. Titanium carbide is the hardest carbide of the commercial transition-metal carbides, but it is too brittle to be used alone. However, it is the most important additional carbide in tungsten-based cemented carbides for cutting steel. Although toughness is decreased a little by the addition of TiC, the hardness and especially the heat resistance are increased significantly. Normal steel-cutting grades contain 5 – 30 wt % of TiC. Furthermore, TiC is the basic carbide for the formation of solid solutions with all other transition-metal carbides used in cemented carbides. In tungsten-free hardmetals, TiC is used in combination with molybdenum carbide and nickel binder metal.

Titanium carbide in combination with steel alloy forms a special type of hard alloy called Ferro-TiC. Tungsten-based cemented carbides can be replaced by this material in some cases. TiC was the first carbide material used for coatings on cutting tips made from normal cemented carbides. Even though the thickness of such a layer is in the range of only a few microns, the life of the cutting tools is increased markedly. Some TiC is used in combination with oxides in ceramic cutting tools (Al_2O_3 – TiC).

2.3. Tantalum Carbide

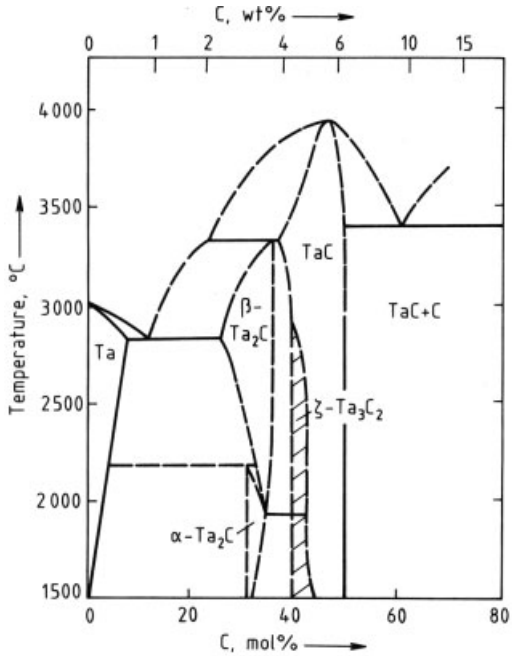

Figure 5. Tantalum – carbon phase diagram [3]

In general, the phase relationships in the systems of group 5 metals and carbon are complex. The system Ta – C (Fig. 5) is typical for the group. This system is characterized by several subcarbides, with lower carbon contents, in addition to the monocarbide, TaC. The face-centered cubic monocarbide melts without decomposition near 4000 °C, one of the highest melting points known. The broad range of homogeneity extends from 43 to 50 mol % C. The subcarbide Ta_2C decomposes at its incongruent melting point. There are two modifications, a high-temperature phase, with disordered L′3 structure, and a low-temperature phase,

of C6 type. The transformation temperature is near 2000 °C. In addition, there is a metastable Ta_3C_2, which is sometimes called the Brauer or ζ-phase. Similar phases are also found in the V – C and Nb – C systems, although the structure of the carbides is still not completely resolved [17].

The monocarbide, TaC, is the only phase of commercial interest. It is a brown powder, usually of 1 – 5 µm average grain size. Sintered and polished pieces have a yellow-golden sheen. The chemical resistance is high. The monocarbide is stable in nonoxidizing acids, although it is attacked easily by a mixture of HNO_3 and HF and by oxidizing salt melts. It can be heated up to 3000 °C in hydrogen or nitrogen, but it is oxidized rapidly in air at 800 °C.

Preparation. The method most used for preparation is based on the reaction of tantalum oxide with carbon:

$$Ta_2O_5 + 7\,C \longrightarrow 2\,TaC + 5\,CO$$

Intimate mixtures of oxide and carbon are pressed into graphite boats and heated at 1700 °C in hydrogen. Usually the product is deficient in bound carbon, and this must be adjusted in a second step. Tantalum metal can also be used for direct carburization with carbon. Very pure monocarbide for scientific use is obtained by the reaction of tantalum hydride with carbon.

A commercial grade of lower purity can be made by melting ferrotantalum or tantalum-containing scrap and slag in a metal bath with an excess of carbon. After the mixture is cooled, the auxiliary metal is dissolved with acid to free tantalum carbide. An additional step to adjust the carbon content is also necessary in this process.

Uses. Because of its extremely high melting point, some TaC is used in high-temperature techniques, but the main application is in hardmetals. Although TaC is the most expensive of all the carbides normally used in cemented carbides, consumption is still increasing because of the marked improvement in the properties of cemented carbides containing TaC. The world's annual demand can be estimated to be ca. 500 t.

There are two quite different reasons for this use of TaC. First, small amounts of TaC, in the range of 0.2 – 2.0 wt %, are added to straight WC – Co grades in which fine-grained WC, 1.5 µm or less, is used. In these grades, often there is an undesired grain growth of the carbide phase during sintering because of the sintering time and temperature and probably some still unknown factors. This grain growth is inhibited to a great extent by TaC. Although there are other, cheaper compounds that retard grain growth, TaC is preferred because it is the only compound known to have no negative effects. The second reason for using TaC in cemented carbides is based on the great improvement in cutting tools, mainly in long-chipping steel cutting grades. In this second case the TaC content ranges from 2 to 15 wt %, and even more in some special cases. Thermal shock resistance, hot hardness, and resistance against cratering and oxidation are all increased markedly.

2.4. Niobium Carbide

The phase relationships in the system niobium – carbon are quite similar to those in the system tantalum – carbon. However, because of the close similarity, there is some doubt about the correctness of the phase diagram. The face-centered cubic monocarbide, NbC, melts without decomposition at ca. 3600 °C. It has a range of homogeneity from ca. 40 to almost 50 mol %; the stoichiometric value of 50 % is never reached. The subcarbide Nb_2C decomposes at its melting point of ca. 3000 °C. It exists in several modifications, the number and structure of which are a point of uncertainty. The other open question is the existence of an additional ζ-phase, Nb_3C_2.

The monocarbide, NbC, is the only phase used industrially. It is a gray-brown powder of normally less than 10 µm average grain size. Sintered pieces show a lavender tint. The chemical reactivity is similar to that of TaC, but NbC is less resistant to nitrogen. Heating the carbide in ammonia produces the nitride.

Preparation. Niobium monocarbide, NbC, is made in the same way as TaC: by carburization of the oxide, hydride, or metal at 1500 °C. The auxiliary metal bath technique can also be used. Most NbC is not produced as a single pure compound because demand for pure NbC

is small. Nearly all of the NbC is used in combination with TaC; therefore, mixtures of Nb_2O_5 and Ta_2O_5, in various ratios, are carburized together. The resulting products are true mixed crystals of NbC and TaC. Any ratio can be prepared, but the usual commercial compositions contain 10, 20, or 50 wt % NbC.

Uses. Only small quantities of pure NbC are needed. Some is used in special grades of cemented carbides in combination with Al_2O_3 (cermets). Another use is the reduction of Nb_2O_5 by NbC to niobium metal, a process carried out at 1600 °C under vacuum or hydrogen.

Most of the NbC is used in combination with TaC in hardmetals. When used with TaC, the NbC improves the properties of the sintered material just like pure TaC does. At the same time, NbC is much less expensive than TaC. However, NbC is said to decrease the strength when it is added in large amounts. The limit and the degree of toughness loss are not known exactly. In any case, NbC is never used alone, and the content of NbC in the TaC ranges from less than 1 to 50 wt %.

2.5. Zirconium Carbide

The monocarbide ZrC is the only compound in the zirconium – carbon system. It has a face-centered cubic crystal structure, and the range of homogeneity reaches from 38 to 50 mol % C. Zirconium carbide melts without decomposition at ca. 3400 °C. The phase boundaries in the Zr – C system are extremely sensitive to oxygen, nitrogen, and probably even more to other impurities. On the other hand, ZrC is difficult to prepare free from oxygen and nitrogen because the lattices of ZrC, ZrO, and ZrN are isotypical. As a result there have been many misinterpretations of the phase diagram in the past, and some disagreements still must be clarified.

The carbide is a gray powder. Its chemical resistance is somewhat lower than that of TiC. It is dissolved by cold HNO_3 or a cold mixture of H_3PO_4 and dilute H_2SO_4. It can be heated in hydrogen up to its melting point, but it is attacked by oxygen at 500 °C. When the carbide is heated in nitrogen above 1500 °C, the nitride is formed.

Preparation. Zirconium carbide is made mostly by heating mixtures of ZrO_2 and carbon at 1800 – 2400 °C in hydrogen or under vacuum. The carbon content must be adjusted in a second step. The metal, in the form of a sponge, or the hydride can be carburized with carbon at a temperature as low as 1400 – 1600 °C. Up to 1800 °C, the carbide getters oxygen, and oxygen-free material is difficult to obtain. Very pure ZrC, for scientific use, can be made by the gas-phase reaction of $ZrCl_4$, H_2, and a hydrocarbon.

Use. Up to now only small quantities of ZrC have been used in hardmetals. This may be due to the comparatively high price of ZrC as well as to its insufficient heat resistance. ZrC forms solid solutions with all other transition-metal carbides. Therefore, the demand of ZrC may increase. Other than use in cemented carbides, there is no use of importance.

2.6. Hafnium Carbide

The hafnium – carbon system and the properties of the carbide are similar to those of zirconium. The only phase is the face-centered cubic monocarbide, HfC. The broad range of homogeneity extends from 37.5 to 50 mol % C. The carbide melts at ca. 3900 °C without decomposition. Its chemical reactivity seems to be similar to that of ZrC, but little information is to be found in the literature.

Preparation. Hafnium dioxide can be carburized like ZrO_2 in hydrogen or under vacuum at 1800 – 2200 °C. If hafnium metal or hydride is used for carburization, a temperature of 1600 – 1700 °C is sufficient. When the oxide is the starting material, a second step for adjustment of carbon content and reduction of the oxygen content is necessary. This second step is often not neces-sary if the metal or hydride is the starting material.

Uses. Hafnium oxide or metal is a byproduct in the production of zirconium for nuclear reactors. Therefore, hafnium is available in sufficient quantities, and HfC has become attractive for cemented carbides. Tantalum carbide, TaC,

in steel-cutting tools or as grain-growth inhibitor may be replaced by HfC. An HfC – NbC solid solution seems to be the most effective. Coatings of HfC on normal hardmetal tools increase the oxidation resistance. Such coatings are made by chemical vapor deposition (CVD) with $HfCl_4$ and a carbonizing gas.

2.7. Vanadium Carbide

Vanadium forms the same phases with carbon as tantalum and niobium. The face-centered cubic monocarbide, VC, exists over a broad range of homogeneity, from 43 to 49 mol % C. It melts without decomposition at ca. 2800 °C. There are two modifications of the subcarbide, V_2C, the orthorhombic low-temperature α-phase and the hexagonal high-temperature β-phase. The latter decomposes on melting. Furthermore, there is a metastable ζ-phase, V_3C_2. Often a V_4C_3 phase has been described in the literature, but this phase is only found in vanadium-containing steel alloys. Probably it is not a distinct phase, but rather a solid solution of carbide, oxide, and nitride, the structures of which are isotypical.

The only phase of commercial interest is the monocarbide, VC. It is a gray powder with a grain size usually of several micrometers. It is resistant to cold acids, except HNO_3, but it is easily dissolved by hot oxidizing acids. The monocarbide can be heated to its melting point in hydrogen, but in air it is oxidized rapidly at 800 °C.

Preparation. Vanadium monocarbide is made mostly by heating V_2O_3 or V_2O_5 with carbon at 1500 – 1700 °C in hydrogen. Ammonium vanadate also can be used as the starting material. A second treatment under vacuum is necessary in every case to adjust the carbon content and to reduce the oxygen level. Because of the great stability of the V(C,O,N) mixed crystal, oxygen-free material is difficult to prepare. Very pure VC is best made by the reaction of vanadium metal with carbon under vacuum.

Uses. Vanadium monocarbide is too brittle to be used alone in cemented carbides. Some special grades were made with VC – TiC mixed crystals with nickel or iron binder, but this was done temporarily only when there was a shortage of tungsten. Small quantities of VC are used to inhibit grain growth in tungsten carbide – cobalt hardmetals. The effectiveness is higher than that of TaC, but the toughness of the sintered material is lower when > 0.5 % VC is added. Large quantities of VC are contained in steel alloys where it forms during melting.

2.8. Chromium Carbide

The phase relationships in the chromium – carbon system are quite different from those of the other metals of group 6 as well as those of the metals of groups 4 and 5. Hägg's ratio of the atomic radii is 0.609 in the case of chromium and carbon; thus, the critical figure of 0.59 is exceeded and a simple closest packed structure can no longer form. There are three chromium carbides in the system (Fig. 6). The cubic carbide $Cr_{23}C_6$ [*12105-81-6*] is a complex $D8_4$ type with 116 atoms in the unit cell. It decomposes on melt-ing at ca. 1500 °C. This $Cr_{23}C_6$ is sometimes formulated incorrectly as Cr_4C. The hexagonal Cr_7C_3 [*12075-40-0*] melts without decomposition at ca. 1800 °C, whereas the orthorhombic Cr_3C_2 [*12012-35-0*] decomposes at its melting temperature of ca. 1900 °C. All three phases have very narrow ranges of homogeneity. There are hints of the existence of one or more additional phases in the high-temperature range. These phases and some uncertainties in the phase boundaries between the known carbides are still a matter of discussion. Very little is known about the properties of $Cr_{23}C_6$ and Cr_7C_3 because these carbides are not prepared as pure compounds and are never used alone. They form during melting of steel and ferrous alloys, and they exist probably in the form of mixed crystals with iron and other metallic carbides. The carbide Cr_3C_2 is the only phase produced as such. It is a gray powder of a grain size normally less than 10 µm. It is insoluble in cold HCl, but dissolves in hot oxidizing acids and in H_2O_2. It has the greatest resistance to oxidation of all metal-like carbides. It is stable in air up to 1000 °C because of a very dense and firm oxide layer that forms on its surface.

Preparation. The carbide Cr_3C_2 is made by heating mixtures of Cr_2O_3 and carbon up to 1600 °C in hydrogen. Below 1300 °C, primarily Cr_7C_3 is formed. The following equations

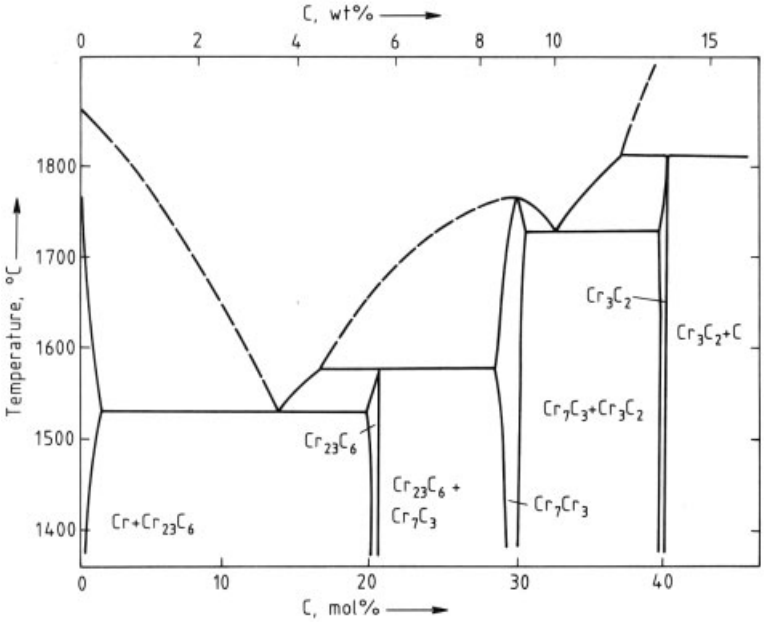

Figure 6. Chromium – carbon phase diagram [6]

demonstrate how complicated the process of carburization is:

$$3\,Cr_2O_3 + 13\,C \longrightarrow 2\,Cr_3C_2 + 9\,CO$$

$$5\,Cr_2O_3 + 27\,Cr_3C_2 \longrightarrow 13\,Cr_7C_3 + 15\,CO$$

$$Cr_2O_3 + 3\,Cr_7C_3 \longrightarrow Cr_{23}C_6 + 3\,CO$$

$$2\,Cr_2O_3 + 3\,Cr_3C_2 \longrightarrow 13\,Cr + 6\,CO$$

Oxygen-free Cr_3C_2 with the stoichiometric carbon content is difficult to obtain if the starting material is the oxide. Very pure Cr_3C_2 can be made by the carburization of chromium metal powder.

Uses. Some Cr_3C_2 is used in hardmetals in special tools with great resistance to acids and salts. In these grades the carbide is bound with nickel. Small quantities of Cr_3C_2 are used as a grain-growth inhibitor in WC – Co cemented carbides. Considerable amounts of the eutectic $Cr_7C_3 - Cr_3C_2$ are used in welding electrodes for hard facing.

The greatest demand for chromium carbides is in steel, Stellites, and related alloys. In such cases, pure chromium carbides are not used; instead, chromium metal is added to the melt together with carbon-containing additives.

2.9. Molybdenum Carbide

Although much work has been done on the molybdenum – carbon system, there are still uncertainties and disagreements. The existence of at least four phases seems to be assured (Fig. 7). The hexagonal Mo_2C is the only phase stable at room temperature. Its range of homogeneity is very narrow and lies between 33 and 34 mol % C. The orthorhombic Mo_2C phase is stable only above 1475 °C. It melts without decomposition at ca. 2400 °C. Two carbon-rich phases, ca. 39 mol % C, exist only at high temperature, a hexagonal one above 1655 °C and a cubic one above 1960 °C. Both phases are designated as MoC_{1-x} by some authors and as MoC or Mo_3C_2 by other authors. Below their decomposition temperatures these phases break down into Mo_2C and C. Hexagonal MoC_{1-x} is isotypical with WC and can be stabilized by the inclusion of tungsten. More phases have been observed in the system, but probably all of them were oxygen-containing mixed phases.

The only phase of commercial interest is the hexagonal Mo_2C. It is a gray powder in the micron range. It is resistant to nonoxidizing acids but is dissolved by HNO_3 or by hot H_2SO_4. It

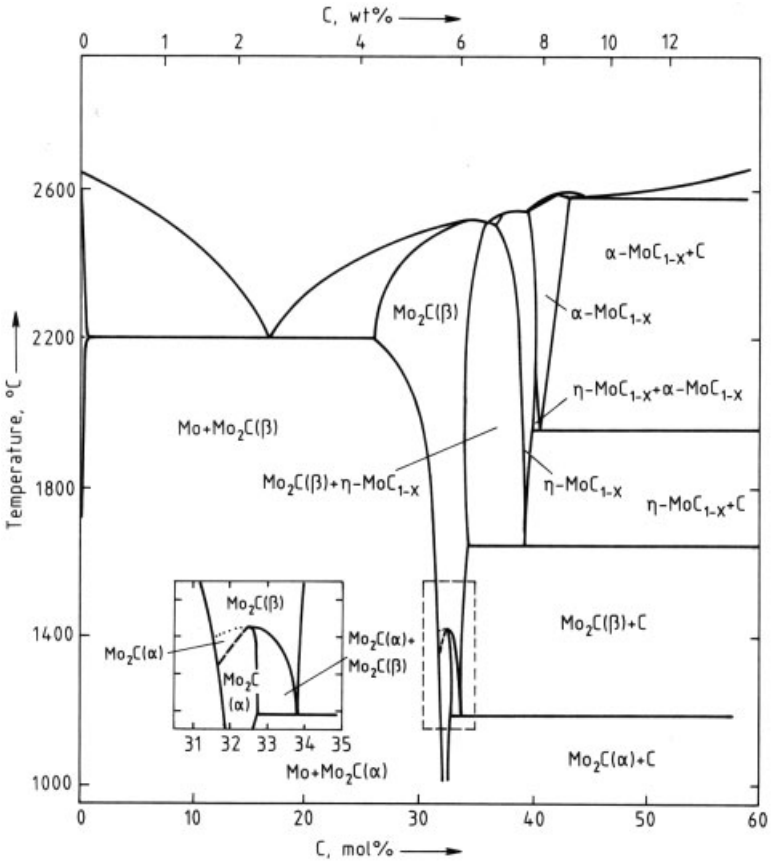

Figure 7. Molybdenum – carbon phase diagram [6]

is stable in hydrogen, but it is oxidized in air at 500 °C.

Preparation. Although MoO_3 or MoO_2 can be carburized with carbon at 1500 °C, a carbide with the correct carbon content and a low oxygen content is difficult to obtain. Pure Mo_2C is best made by heating molybdenum metal powder with carbon in hydrogen at ca. 1500 °C.

Uses. Mo_2C is used in special cemented carbide grades containing TiC and nickel metal. Such grades were the first tungsten-free hardmetals. Attempts have been made to replace tungsten partially with Mo – W mixed crystals. Most Mo_2C is used in steel alloys, where it forms during melting.

3. Mixed Carbides

The commercial carbides of groups 4, 5, and 6 form numerous mixed carbides with each other. The formation of these solid solutions depends on the lattice constants of the carbides and corresponds to the Hume – Rothery rule on the atomic volumes. Only a few of the metallic carbides do not conform to these conditions and, thus, do not form a continuous series of solid solutions. The cubic monocarbides of the metals of groups 4 and 5 are completely miscible, except the pairs ZrC – VC and HfC – VC, which are soluble in each other only to a limited extent. Limited miscibility is found between TiC and the carbides of Cr, Mo, and W, but there are still some uncertainties about the TiC – WC system. Solid solutions of three or more carbides also exist. The host lattice of the mixed carbides is usually TiC. The

hexagonal WC has only a negligible capability to receive cubic carbides into solid solution.

The use of mixed crystals in cemented carbides offers several advantages. Mixed crystals are harder and tougher than single, unalloyed carbides. The contents of oxygen, nitrogen, and free graphite are distinctly lowered by autopurification during the diffusion process; the wettability by cobalt and other binder metals is increased.

The temperatures for the preparation of mixed crystals are ca. 500 °C higher than the normal sintering temperatures for hardmetals. Therefore, an exact formation of mixed crystals cannot be attained by mixing the individual carbides with the binder metal before sintering: the mixed carbides must be prepared in a separate process.

The mixed carbides can be prepared by the processes used for the preparation of the single carbides. Mixtures of the oxides and carbon are heated up to 1800 – 2000 °C in hydrogen, mostly in high-frequency induction furnaces. A second step under vacuum is always necessary to adjust the carbon content. The mixed metal powders can be treated with carbon in the same way. In addition, the reaction of metal oxide with another metal carbide and additional carbon is used, for example:

$$TiO_2 + WC + 3\,C \longrightarrow 2\,(Ti,W)C + 2\,CO$$

Very pure mixed carbides are best made by heating mixtures of the single carbides at 2000 °C under vacuum. The process of diffusion can be accelerated by the addition of cobalt, nickel, iron, or chromium metal in the range 0.5 – 1.0 wt %:

$$TiC + TaC \xrightarrow{Ni} 2\,(Ti,Ta)C$$

Pure mixed crystals are also made by the auxiliary metal bath technique (menstruum process, see page 9).

3.1. Tungsten – Titanium Carbide

The most important mixed carbide in cemented carbides is tungsten – titanium carbide. The ternary system is not yet known in all its details. The uncertainties in the Ti – C system caused by the isotypical TiO and TiN phases are observed in the Ti – W – C system as well. The solid solubility of WC in the cubic lattice of TiC is limited and depends on the temperature:

Temperature, °C	1500	2000	2400
WC, wt %	60	80	90
WC, mol %	31	55	73

Above 2600 °C, there is probably complete miscibility.

Saturated mixed crystals prepared at high temperatures become supersaturated when used in hardmetal production because of the comparatively low sintering temperatures, only 1400 – 1500 °C. Therefore, very fine WC crystals precipitate in the metal binder phase, thus strongly influencing the properties of the hardmetal. The mechanisms are still a matter of discussion. The W – Ti mixed crystals used in industry generally contain 50 or 70 wt % W.

3.2. Other Mixed Carbides

Another important mixed carbide for cemented carbides is tungsten – tantalum carbide. The solubility of cubic TaC in hexagonal WC is negligible, but WC has a limited solubility in TaC that depends strongly on the temperature:

Temperature, °C	1500	1800	2000	2500
WC, wt %	10	20	27	70

Because the solubility of WC in TaC decreases rapidly as the temperature is lowered, WC crystals always precipitate from the solid solutions during cooling. Therefore, preparation of single-phase (W,Ta)C mixed crystals is almost impossible, and unlike other mixed carbides, (W, Ta) C is usually marketed with the additional designation *double phase.*

The (W,Ti,Ta)C mixed crystal may be viewed as a combination of (W,Ti)C and (W,Ta)C mixed crystals. It is used in considerable quantities in cutting tools for steel and related long-chipping materials. Sometimes this mixed crystal is called "triple carbide", which is incorrect because in nearly every case, TaC – NbC mixtures, not pure TaC, are used. The system W – Ti – Ta – Nb – C is not yet known in all its details. There are some

isothermal sections in the quasiternary system WC – TiC – TaC [18], but the effects of the addition of NbC to this system are not known exactly.

Another important mixed crystal is titanium – molybdenum carbide (Ti,Mo)C. It is used in tungsten-free hardmetals with nickel binder for special steel cutting tools.

From the nine commercial carbides of the transition metals, 36 combinations of double carbides can be formed. Some of these are attracting growing interest, for example (W,Mo)C as a partial substitute for WC, or (Zr,Hf)C and (Zr,Nb)C as substitutes for TaC. All the double carbides are being intensively investigated [19], [20]. Mixed crystals of three transition-metal carbides, such as (Ti,Nb,Ta)C, number a total of 84, and 126 combinations are possible from four carbides, for example, (Ti,W,Hf,Zr)C. The knowledge about most of these systems is still quite limited.

3.3. Carbonitrides

The face-centered cubic monocarbides of groups 4 and 5 are isotypical with the face-centered cubic nitrides of the same groups, and because of this, there is complete miscibility between a metal carbide and metal nitride. The situation for Ta (C,N) is more complex because the usual form of TaN is hexagonal, the cubic TaN being stable only at high temperature and pressure.

With some limitations all carbonitride compounds can be used in cemented carbides [21]. A special application is scratchproof watchcases, because some carbonitrides are colored. The tint depends on the C : N ratio. For example, Nb (C,N) is violet at high carbon contents but yellow at high nitrogen contents. Another special application is coating of normal cemented carbides with layers of carbonitrides, usually Ti (C,N), by chemical vapor deposition (CVD). Carbonitrides can best be made by heating mixtures of the single carbides and nitrides in argon or under vacuum at 1600 – 1800 °C.

3.4. Mixed Carbonitrides

By methods similar to those used in the production of solid solutions of carbides or carbonitrides, a large number of mixed crystals can be prepared with various metals and nonmetals in one lattice. Only a few of them have been investigated up to now, and still fewer are used commercially. However, interest is growing. For example, (W,Mo)(C,N) with a nickel binder has properties comparable to those of WC – Co cemented carbides and can be used as a partial substitute for tungsten. The mixed carbonitride (Ti,Mo)(C,N) with a nickel alloy binder can be made to have an extremely fine carbide structure because of a spinodal decomposition of the carbonitrides. Both (Ti,W)(C,N) and (Ti,Ta)(C,N) have excellent heat-resistant properties.

4. Carbides of the Iron Group and Manganese

The carbides of Fe, Co, Ni, and Mn are usually classified as metallic carbides or metal-like carbides like the carbides of groups 4, 5, and 6. However, in fact, they are different from the transition-metal carbides, and their metallic characteristics are less pronounced. Hardness, melting points, and electrical conductivity are all distinctly lower. The crystal structures are not the simple interstitial Hägg phases, but rather much more complex structures, similar to those of the chromium carbides.

The carbides of Fe, Co, Ni, and Mn are neither prepared nor used alone. They are formed in iron and steel alloys during the melting process, and they can be isolated from these products by anodic oxidation of the metals. The carbides are important because they are the hardening phases in steel alloys, Stellites, cast iron, and related materials.

In the iron – carbon system there is probably only one phase, the orthorhombic Fe_3C, which is called cementite [*12011-67-5*]. Preparation of the pure carbide from the elements has not been achieved up to now. When Fe_3C is isolated electrolytically from alloys and sintered with cobalt metal, it decomposes. In alloy steels, the "iron carbide" is mostly included in mixed crystals with chromium carbides: $(Fe,Cr)_{23}C_6$, $(Fe,Cr)_7C_3$, and $(Fe,Cr)_3C_2$.

Cobalt and nickel form only the carbides Co_3C and Ni_3C, which are isotypical to Fe_3C. In the manganese – carbon system, three carbides

are formed: Mn_3C is isotypical to Fe_3C, and Mn_7C_3 and $Mn_{23}C_6$ are isotypical to the corresponding chromium carbides.

5. Complex Carbides

A great number of ternary and quaternary phases can be formed between carbon and two or three metals, one a transition metal. In addition, the elements S, P, and As can be included.

These so-called complex carbides are a group of their own. They are not solid solutions of one carbide in the lattice of another carbide. Each has its own typical structure, which in all cases is much more complicated than the simple Hägg phases of the transition-metal carbides.

Numerous complex carbides have been investigated [9], and all were found to contain octahedral or, less often, trigonal prismatic M_6C groups. M is always a transition metal, and six such atoms surround a central carbon atom. The octahedrons are linked by common corners, edges, or faces. The resulting interstitial sites can be occupied by other metals. Many distinct crystal structures can be formed under these conditions. The most important are perowskite carbides such as Ti_3AlC, β-Mn carbides such as Ta_3Al_2C, $\varkappa$-carbides such as $W_{16}Ni_3C_6$, η-carbides such as W_3Co_3C, H-phases or Cr_2AlC-type carbides such as Zr_2SC, V_3AsC-type carbides such as Cr_3PC, and Mn_5Si_3-type carbides such as $Nb_5Ga_3C_{0.2}$. Complex carbides can be best prepared by heating mixtures of the single carbides and metals for an extended period of time. Mechanical pressure or gas pressure is helpful.

Of commercial interest are mainly the η-carbides, which are formed in alloy steels and in Stellites. In hardmetals, η-carbides, such as W_3Co_3C and W_4Co_2C, form because of carbon deficiency, these phases causing a decrease in toughness.

6. References

General References

1. R. Kieffer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963.
2. R. Kieffer, F. Benesovsky: *Hartmetalle,* Springer Verlag, Wien 1965.
3. E. K. Storms: *The Refractory Carbides,* Academic Press, New York 1967.
4. L. Toth: *Transition Metal Carbides and Nitrides,* Academic Press, New York 1971.
5. W. B. Pearson: *Handbook of Lattice Spacings and Structures of Metals and Alloys,* **vol. 1 and 2,** Pergamon Press, Oxford 1958 (vol. 1) and 1967 (vol. 2).
6. E. Rudy: *Compendium of Phase Diagram Data,* **AFML-TR-65-2,** part 5, 1969.

Specific References

7. G. Hägg, *Z. Phys. Chem. Abt. B* **12** (1931) 33 – 56.
8. H. Nowotny, F. Benesovsky, *Planseeber. Pulvermetall.* **16** (1968) 204 – 214.
9. H. Nowotny, *Angew. Chem.* **84** (1972) 973 – 982; *Angew. Chem. Int. Ed. Engl.* **11** (1972) 906 – 915.
10. D. N. French, D. A. Thomas, *Trans. Metall. Soc. AIME* **233** (1965) 950 – 952.
11. O. Rüdiger, G. Ostermann, H. Kolaska, *Tech. Mitt. Krupp Forschungsber.* **28** (1970) no. 2, 33 – 54.
12. E. Neuenschwander, *J. Less-Common Met.* **11** (1966) 365 – 375.
13. L. Ramquist in H. H. Hausner (ed.): *Modern Developments in Powder Metallurgy Processes,* **vol. 4,** Plenum Press, New York 1970, pp. 75 – 84.
14. M. Miyake, *Prepr. Eur. Symp. Powder Metall. 5th 1978* 1978 – 1979, 93 – 98.
15. K. Ushiyima, *Powder Metall. Int.* **11** (1979) 158 – 160.
16. G. Jangg, R. Kieffer, L. Usner, *J. Less-Common Met.* **14** (1968) 269 – 277.
17. G. Brauer, R. Lesser, *Z. Metallkd.* **50** (1959) 8.
18. Ch. Chatfield, *Powder Metall. Int.* **15** (1983) 18 – 19.
19. H. Holleck, *Metall (Berlin)* **35** (1981) 999 – 1004.
20. H. Holleck, *Metall (Berlin)* **35** (1981) 1246 – 1253.
21. R. Kieffer, P. Ettmayer, M. Freudhofmeier, *Metall (Berlin)* **25** (1971) 1335 – 1342.

Carbohydrates

FRIEDER W. LICHTENTHALER, Institut für Organische Chemie, Technische Universität Darmstadt, Darmstadt, Federal Republic of Germany

1. Introduction 1
2. Monosaccharides 2
2.1. Structure and Configuration 3
2.2. Ring Forms of Sugars: Cyclic Hemiacetals 4
2.3. Conformation of Pyranoses and Furanoses 5
2.4. Structural Variations of Monosaccharides 6
3. Oligosaccharides 7
3.1. Common Disaccharides 7
3.2. Cyclodextrins 10
4. Polysaccharides 10
5. Nomenclature 15
6. General Reactions 16
6.1. Hydrolysis 16
6.2. Isomerization 16
6.3. Decomposition 17
7. Reactions at the Carbonyl Group 17
7.1. Glycosides 17
7.2. Thioacetals and Thioglycosides . . 18
7.3. Glycosylamines, Hydrazones, and Osazones 18
7.4. Chain Extension 19
7.5. Chain Degradation 20
7.6. Reductions to Alditols 21
7.7. Oxidation 22
8. Reactions at the Hydroxyl Groups 22
8.1. Ethers 22
8.2. Esters of Inorganic Acids 23
8.3. Esters of Organic Acids 24
8.4. Acylated Glycosyl Halides 24
8.5. Acetals 25
9. Carbohydrates as Organic Raw Materials 26
9.1. Microbial Synthesis 27
9.2. Chemical Conversions 27
9.2.1. Furan Derivatives 27
9.2.2. Pyrones and Dihydropyranones . . . 30
9.2.3. Sugar-Derived Unsaturated *N*-Heterocycles 31
9.2.4. Sugar-Based Surfactants 34
9.2.5. Hydrophilic Monomers of Polyamides 36
10. References 37

1. Introduction

Terrestrial biomass constitutes a multifaceted conglomeration of low and high molecular mass products, exemplified by sugars, hydroxy and amino acids, lipids, and biopolymers such as cellulose, hemicelluloses, chitin, starch, lignin and proteins. By far the most abundant group of these organic products and materials, in fact about two thirds of the annually renewable biomass, are carbohydrates, i.e., a single class of natural products. As the term 'carbohydrate' (German *'Kohlenhydrate'*; French *'hydrates de carbone'*) implies, they were originally considered to consist solely of carbon and water in a 1:1 ratio, in recognition of the fact that the empirical composition of monosaccharides can be expressed as $C_n(H_2O)_n$. Today, however, the term is used generically in a much wider sense, not only comprising polysaccharides, oligosaccharides, and monosaccharides, but substances derived thereof by reduction of the carbonyl group (alditols), by oxidation of one or more terminal groups to carboxylic acids, or by replacement of one or more hydroxyl group(s) by a hydrogen atom, an amino group, a thiol group, or similar heteroatomic groups. A similarly broad meaning applies to the word 'sugar', which is often used as a synonym for 'monosaccharide', but may also be applied to simple compounds containing more than one monosaccharide unit. Indeed, in everyday usage 'sugar' signifies table sugar, which is sucrose (German *'Saccharose'*; French *'sucrose'* or *'saccharose'*), a disaccharide composed of the two monosaccharides D-glucose and D-fructose.

Carbohydrates appear at an early stage in the conversion of carbon dioxide into organic compounds by plants, which build up carbohydrates from carbon dioxide and water by photosynthesis. Animals have no way of synthesizing carbo-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_079

hydrates from carbon dioxide and rely on plants for their supply. The carbohydrates are then converted into other organic materials by a variety of biosynthetic pathways.

Carbohydrates serve as sources (sugars) and stores of energy (starch and glycogen); they also form a major portion of the supporting tissue of plants (cellulose) and of some animals (chitin in crustacea and insects); they play a basic role as part of the nucleic acids DNA and RNA. Other carbohydrates are found as components of a variety of natural products, such as antibiotics, bacterial cell walls, blood group substances, glycolipids, and glycoproteins, the latter, due to their multifaceted carbohydrate-based recognition phenomena, forming the basis of glycobiology.

In the last decade, a large collection of books on carbohydrate chemistry and biochemistry have appeared, ranging from comparatively brief introductions [1–3] to more elaborate monographs [4–7] and multivolume comprehensive treatises [8], [9]. They are recommended as more profound sources of information.

2. Monosaccharides

The generic term 'monosaccharide' denotes a single sugar unit without glycosidic connection to other such units. Chemically, monosaccharides are either polyhydroxyaldehydes or *aldoses* (e.g., glucose) or polyhydroxyketones or *ketoses* (e.g., fructose), the ending 'ose' being the suffix to denote a sugar. Monosaccharides are classified according to the number of carbon atoms they contain, i.e., hexoses and ketohexoses (or hexuloses) of the general formula $C_6H_{12}O_6$ or pentoses and pentuloses ($C_5H_{10}O_5$). Subdivisions are made according to functional groups which may also be present, for example, aminohexoses ($C_6H_{13}O_5N$), deoxyhexoses ($C_6H_{12}O_5$), and hexuronic acids ($C_6H_{10}O_7$). Monosaccharides with fewer (trioses, tetroses) or more carbon atoms (heptoses, octoses, etc.) are rare.

D-*Glucose* (→ Glucose and Glucose-Containing Syrups, Chap. 2.) [*50-99-7*], also known as dextrose, blood sugar, or grape sugar ('*Traubenzucker*' in German), is a pentahydroxyhexanal, hence belonging to the class of aldohexoses (see Section 2.1). Glucose can be considered the parent compound of the monosaccharide family, because it is not only the most

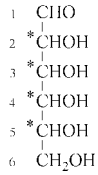

Figure 1. Structural formula of aldohexoses, of which due to the four chiral centers (marked by ∗) 16 stereoisomers are possible

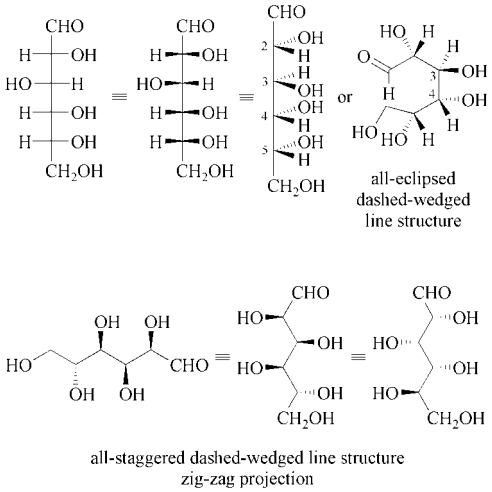

Figure 2. Configurational representations of the linear (acyclic) form of D-glucose: Traditional Fischer projection formula (top left) and its transformation into the more realistic dashed-wedged line depictions with the six-carbon chain in zigzag arrangement.

abundant monosaccharide in nature but also the one most extensively studied. It occurs as such in many fruits and plants, in concentrations of 0.08 – 0.1 % in human blood, and constitutes the basic building unit of starch, cellulose, and glycogen. Other ubiquitous aldohexoses are D-*mannose* [*3458-28-4*], occurring naturally mainly in polysaccharides ('mannans', e.g., from ivory nut) and D-*galactose* [*59-23-4*], a frequent constituent of oligosaccharides, notably lactose and raffinose, and of primary cell wall polysaccharides (pectins, galactans, arabinogalactans). A corresponding isomeric 2-ketohexose is D-*fructose* [*57-48-7*] (→ Fructose), the sweetest natural sugar, which occurs in many fruits and in honey, and, glycosidically linked, in sucrose and the polysac-

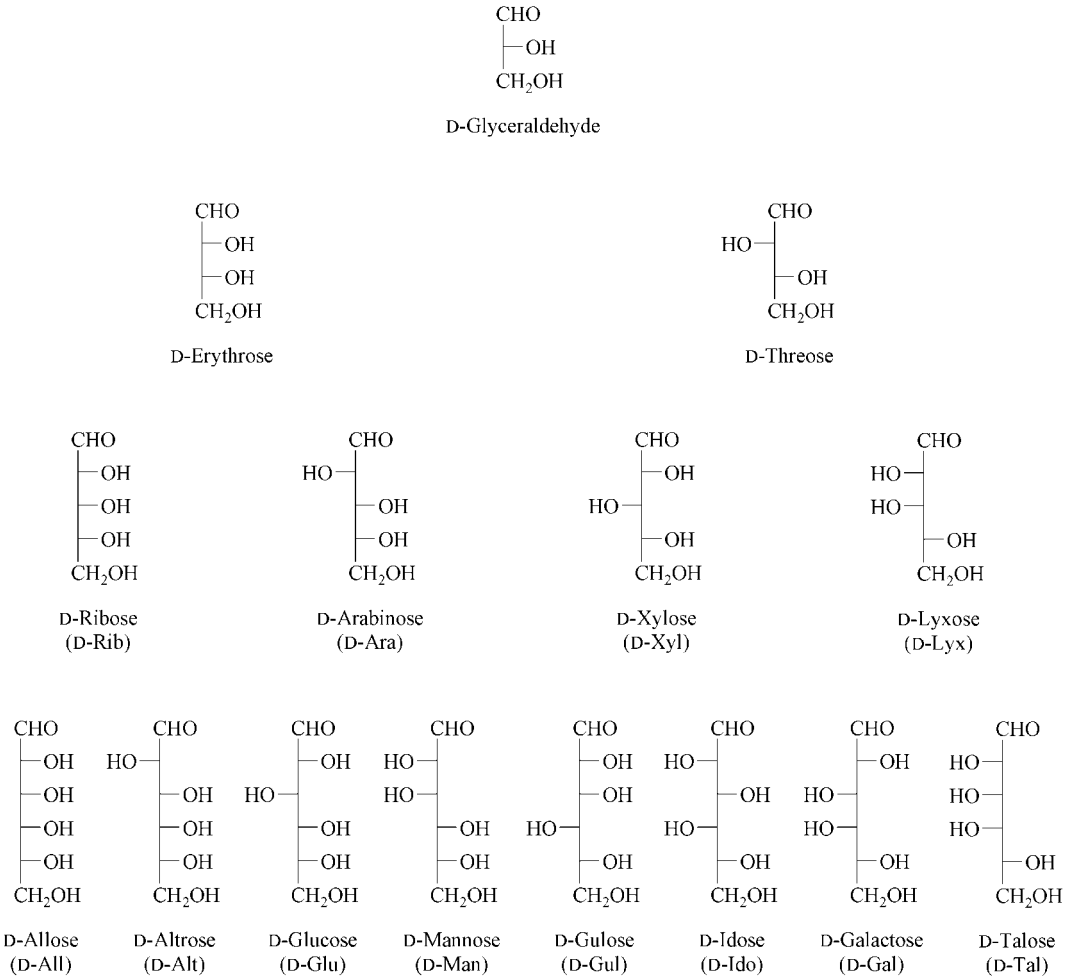

Figure 3. The D-aldose family tree (up to aldohexoses) in their acyclic forms: Common names and Fischer projection formulas, with secondary hydrogen atoms omitted for clarity.

charide inulin, a reserve carbohydrate for many plants (chicory, Jerusalem artichoke). Other important natural sugars are the aldopentose D-*ribose* [*50-69-1*], which constitutes a building block of the ribonucleic acids, L-*arabinose*, widely distributed in bacterial polysaccharides, gums and pectic materials, and D-*xylose* [*58-86-6*], of widespread occurrence in pentosans ('xylans') that accumulate as agricultural wastes (cottonseed hulls, corn cobs).

2.1. Structure and Configuration

D-Glucose, the most abundant monosaccharide, has the molecular formula $C_6H_{12}O_6$ as shown by elemental analysis and molecular mass determination. As evidenced from ensuing reactions (see below) this is consistent with a six-carbon, straight-chain pentahydroxyaldehyde of the following structural formula, an aldohexose in carbohydrate notation (Fig. 1).

This structure contains four asymmetric centers, thus $2^4 = 16$ stereoisomers exist, which can be grouped into eight pairs of enantiomers, and classified as D- and L-sugars. In the D-sugars, the highest numbered asymmetric hydroxyl group (C-5 in glucose) has the same configuration as the asymmetric center in D-glyceraldehyde and, likewise, all L-sugars are configurationally derived from L-glyceraldehyde. A convenient way

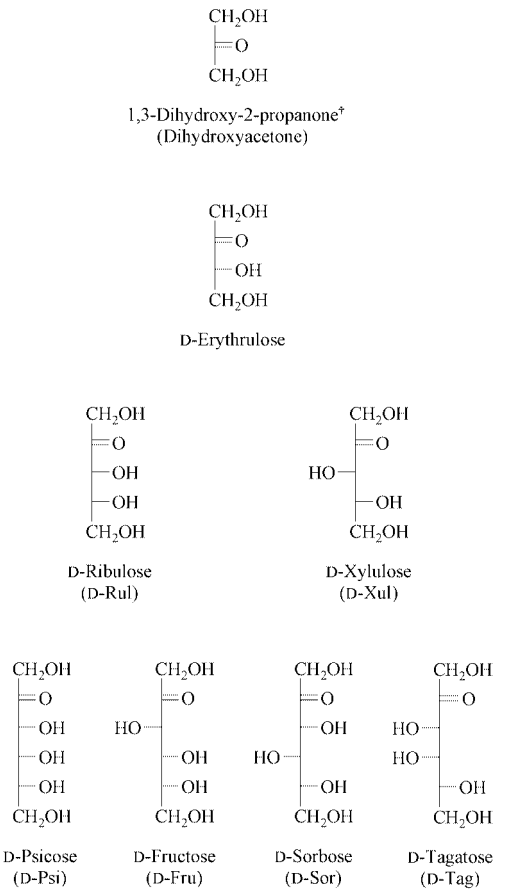

Figure 4. The D-ketohexose (or D-hexulose) family tree: Trivial names, systematic designation (in brackets) and Fischer projection formulas.
† Not regarded as being a sugar, due to absence of an asymmetric carbon atom.

to show configurational relationships was introduced by EMIL FISCHER in 1891 [10], [11], now termed Fischer projection formula (Fig. 2), as it – literally – projects tetrahedral space relationships into a plane. The resulting formulas are simple to write and easy to visualize, yet they require the setting up of conventions: The carbon chain of a sugar is oriented vertically and to the rear with the aldehyde group at the top; hydrogen atoms and hydroxyl groups at the asymmetric carbon atoms stand out in front. The resulting three-dimensional model is then imagined to be flattened and the groups are laid on the plane of the paper. If the lower-most asymmetric center (C-5 in glucose) has the OH group to the right, it is considered to have the D-configuration. FISCHER's decision to place the OH group of natural glucose to the right, hence D-glucose, was purely arbitrary, yet proved to be a fortunate one, since much later, in 1951, it was proven by special X-ray structural analysis [12] that he had made the right choice.

The D-aldose family tree is shown in Figure 3, comprising five of the most important monosaccharides, the aldopentoses D-ribose and D-xylose, and the hexoses D-glucose, D-mannose, and D-galactose, each having the hydroxyl group at the highest-numbered stereocenter (at the bottom) pointing to the right. Likewise, all L-aldoses are configurationally derived from L-glyceraldehyde, entailing a family tree with the lowest OH group to the left; the respective projection formulas being, in essence, mirror images to those in Figure 3.

A similar system is used to build up the series of ketohexoses or hexuloses, i.e., monosaccharides with a keto group at C-2, which therefore contain one asymmetric carbon atom less (Fig. 4).

2.2. Ring Forms of Sugars: Cyclic Hemiacetals

In the solid state and in solution monosaccharides exist in a cyclic hemiacetal form, ring closure corresponding to reaction between the aldehyde group and either the C-4-OH or C-5-OH. Cyclization involving *O*-4 results in a five-membered ring structurally related to furan and therefore designated as a *furanose*, whilst hemiacetal formation with *O*-5 gives rise to an essentially strain-free, hence sterically more favored, six-membered ring, a derivative of pyran, hence termed a *pyranose*. Either ring formation generates a new asymmetric carbon atom at C-1, the anomeric center, thereby giving rise to diastereomeric hemiacetals which are called and labeled α and β. For visualization of the cyclic hemiacetal forms of sugars, HAWORTH, in 1928 [13], introduced his projection formula, in which the rings are derived from the open-chain form and drawn as lying perpendicular to the paper with the ring oxygen away from the viewer. To facilitate this mode of viewing, the front part is usually accentuated by wedges as shown in Figure 5 for the β-anomers of pyranose and furanose forms of D-glucose. The projections devised by MILLS

in 1954 [14], corresponding to those customary for terpenes and steroids, are also very useful for revealing the stereochemistry of sugars in their cyclic hemiacetal forms: the ring is placed in the plane of the paper with solid or broken wedge-shaped lines to show the orientation of substituents, i.e., OH and CH_2OH groups.

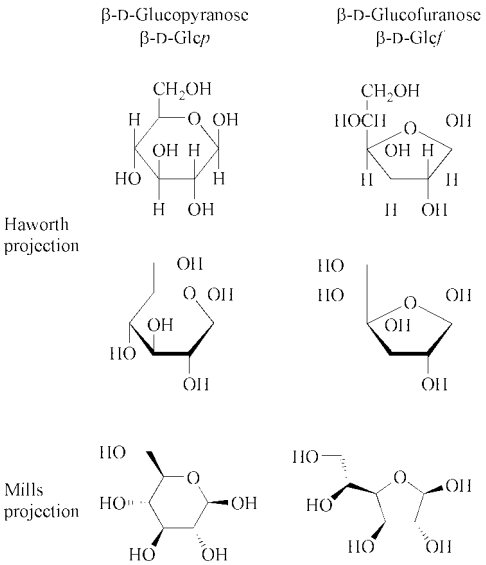

Figure 5. Haworth and Mills projection formulas for the β-anomers of D-glucopyranose and D-glucofuranose (in the formula at the center and at the bottom, the carbon and C-hydrogen atoms are omitted for clarity).

2.3. Conformation of Pyranoses and Furanoses

The concepts of conformation are fundamental to a proper understanding of the structure-property relationships of carbohydrates, most notably of the regio- and stereoselectivities of their reactions. The conformational analysis of monosaccharides is based on the assumption that the geometry of the pyranose ring is essentially the same as that of cyclohexane and, analogously, that of furanoses the same as that of cyclopentane – a realistic view, since a ring oxygen causes only a slight change in molecular geometry. Hence, the rhombus-shaped Haworth formulas which imply a planar ring, and the equally flat dashed-wedged line configurational depictions by Mills (Fig. 5) are inadequate to represent the actual three-dimensional shape of the rings and the steric orientation of the ring substituents (OH and CH_2OH groups). For the six-membered pyranose ring a number of recognized conformers exist [15]: two chairs ($^1C_4$, $^4C_1$), six boats (e.g., $^{1,4}B$ and $B_{1,4}$ in Fig. 6), six skews and twelve half-chairs (e.g., $^oS_2$ and $^5H_4$ forms).

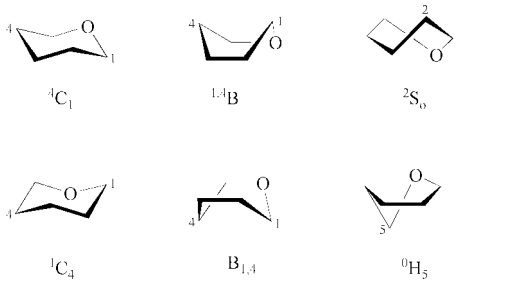

Figure 6. Conformational forms of pyranose rings: chair (C), boat (B), skew (S) and half-chair (H).
To designate each form, the ring atom numeral lying above the plane of reference appears as a superscript preceding the letter, those below the plane are written as subscripts and follow the letter.

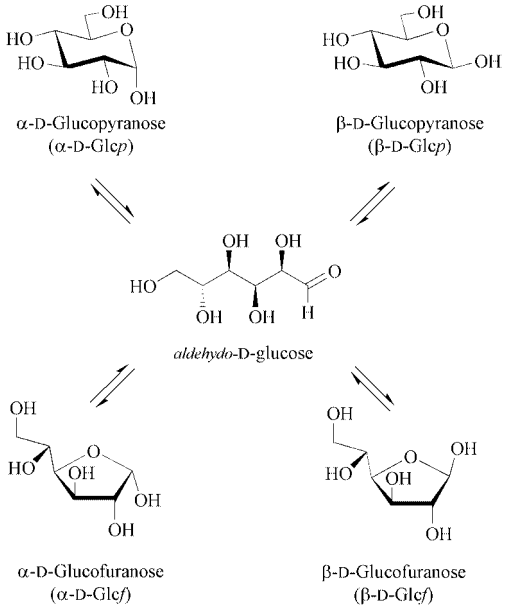

Figure 7. Cyclic hemiacetal forms of D-glucose in configurational representation. In solution, these forms rapidly interconvert through the energetically unfavorable acyclic form; in water at 25 °C the two pyranoid forms are nearly exclusively adopted, the equilibrium mixture amounting to 62 % of the β-p and 38 % of the α-p anomers. From water, D-glucose crystallizes in the α-pyranose form.
The six-membered (pyranose) ring is denoted by the symbol p after the three-letter symbol for the monosaccharide (for example, Glc$p$), the five-membered (furanose) ring correspondingly is signated by an f (e.g., Glc$f$).

Although there are exceptions, most aldohexoses adopt the chair conformation that places the bulky hydroxymethyl group at the C-5 terminus in the equatorial position. Hence, β-D-hexopyranosides are predominantly in the $^4C_1$ chair conformation, since each of the alternative forms outlined in Figure 6, most notably the $^1C_4$ chair, are energetically less favored. For glucose, this preference means that, in the α-form, four of the five substituents are equatorial, and one is forced to lie axial; in the β-form, all substituents are equatorial (Fig. 7). This situation is unique for glucose; the other seven D-aldohexoses contain one or more axial substituents.

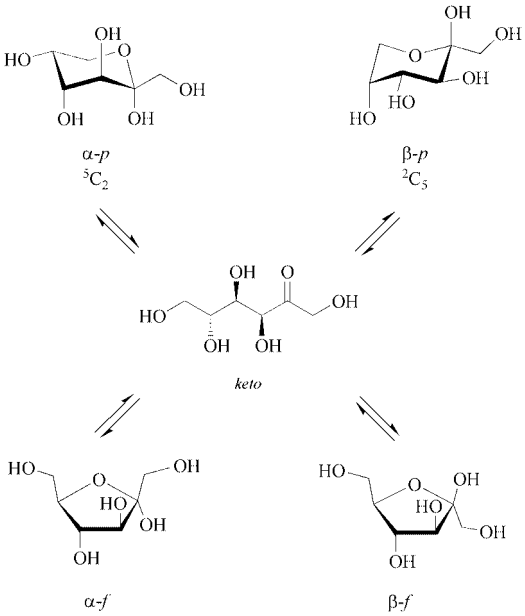

Figure 8. Forms of D-fructose in solution. In water, the major conformers are the β-pyranose (β-*p*, 73 % at 25 °C) and β-furanose (β-*f*, 20 %) forms [17]. On crystallization from water, D-fructose adopts the $^2C_5$ chair conformation in the crystal lattice as evidenced by X-ray analysis [16].

The hexulose counterpart to the conformational forms of D-glucose is the D-fructose isomerization scheme depicted in Figure 8. Whilst the crystalline product is the β-D-fructopyranose in the $^2C_5$ chair conformation as evidenced by X-ray analysis [16], on dissolution in water, equilibration is essentially instantaneous to yield a mixture mainly containing the β-*p*-form (73 % at 25 °C, the only sweet one in fact), together with the β-*f*- (20 %), α-*f*- (5 %) and α-*p*-forms (2 %) [17]. The acyclic form through which equilibration occurs is present to a minute extent only.

The principal conformations of the furanose ring are the *envelope* (*E*) – one atom lying above or below a plane formed by the other four ring atoms – or the *twist* (*T*) arrangement, in which three ring atoms are in a plane and the other two above and below, respectively. As energy differences between the various *E* and *T* conformations are small, the form actually adopted depends on the type of ring substitution (hexoses, hexuloses, pentoses), their configuration, their solvation and the type of intra- or intermolecular hydrogen bonding present. Accordingly, the exact conformation of an individual furanose is usually not known – except for the crystalline state when an X-ray structural analysis is available. Thus, the planar Haworth and Mills projection formulas are the preferred way of drawing furanose forms (Fig. 9).

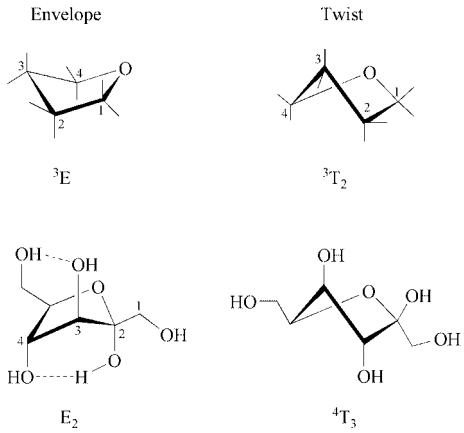

Figure 9. The envelope conformation (top left) is the ^3E form as defined by the C-3 atom lying above the plane formed by the other ring atoms. The defined plane for the twist form (top right) is the triangle given by C-1, C-4, and O-4, entailing the conformational description $^3T_2$. In aprotic solvents (dimethylsulfoxide) D-fructose populates the E_2 envelope conformation to a substantial extent [17], whilst in crystalline sucrose, the β-D-fructofuranose portion adopts the $^4T_3$ twist form [18], [19] (bottom entries).

2.4. Structural Variations of Monosaccharides

Sugars may possess functionalities other than hydroxyl groups. Amino sugars are aldoses, which have a hydroxyl group replaced by an

amino functionality, e.g., D-*glucosamine* (2-amino-2-deoxy-D-glucose), which is one of the most abundant. In its *N*-acetylated form (*N*-acetyl-D-glucosamine), it is a constituent of the polysaccharide chitin (→ Chitin and Chitosan), that forms the hard shells of crustaceans and other anthropods, but also appears in mammalian glycoproteins and links the sugar chain to the protein. Monosaccharides lacking a hydroxyl group at the terminal C-6, i.e., 6-deoxy-sugars, are likewise of wide occurrence, for example, L-*rhamnose* (6-deoxy-D-mannose) is found in plant and bacterial polysaccharides whereas L-*fucose* (6-deoxy-D-galactose) is present in combined form in animals, plants, and microorganisms. 2-Deoxy-D-*erythro*-pentose (2-deoxy-D-ribose) is the exceedingly important sugar component of DNA, various mono-, di- and trideoxy sugars are constituents of many antibiotics, bacterial polysaccharides, and cardiac glycosides.

The uronic acids are aldoses that contain a carboxylic acid chain terminating function, and occur in nature as important constituents of many polysaccharides. The D-gluco compound, D-*glucuronic acid*, was first isolated from urine (hence the name), in which it occurs as glycosides and glycosyl esters of toxic substances that the body detoxifies in this way.

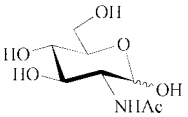

N-Acetyl-D-glucosamine
2-acetamido-2-deoxy-D-glucopyranose
(D-GlcNA*cp*)

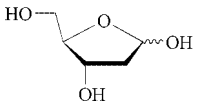

2-Deoxy-D-ribose
2-deoxy-D-*erythro*-pentafuranose
(D-dRib*f*)

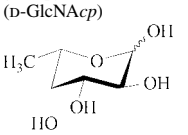

L-Fucose
6-deoxy-L-galactopyranose
(L-Fuc*p*)

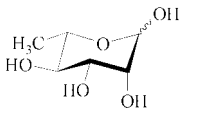

L-Rhamnose
6-deoxy-L-mannopyranose
(L-Rha*p*)

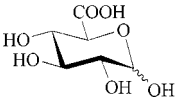

D-Glucuronic acid
(D-GlcA*p*)

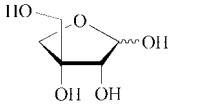

D-Apiose
3-C-hydroxymethyl-D-*glycero*-tetrose (D-Api*f*)

Branched-chain sugars, i.e., saccharides with a non-linear carbon chain, are comparatively uncommon, the more widely occurring being D-*apiose* (3-C-hydroxymethyl-D-*glycero*-tetrose), richly present in polysaccharides of parsley and duckweed [20], and D-*hamamelose* (2-C-hydroxymethyl-D-ribose), a component of the bark of witchhazel [21].

3. Oligosaccharides

Oligosaccharides are compounds in which monosaccharide units are joined by glycosidic linkages, i.e., simple polymers containing between two and ten monosaccharide residues. Accordingly, there are disaccharides – a disaccharide composed of two hexopyranoses can have 5120 distinguishable isomeric forms – trisaccharides, tetrasaccharides, etc. They may be further subdivided into homo- (consisting of only one type of sugar) and hetero-oligosaccharides, and into those that are reducing (presence of a free hemiacetal group) or non-reducing. A comprehensive listing of the di-, tri-, and higher oligosaccharides known by 1990 is available [22].

3.1. Common Disaccharides

Sucrose, affectionately called "the royal carbohydrate" [23], is a non-reducing disaccharide because its component sugars, D-glucose and D-fructose, are glycosidically linked through their anomeric carbon atoms: Sucrose is a β-D-fructofuranosyl α-D-glucopyranoside (see Fig. 10). It is widely distributed throughout the plant kingdom, is the main carbohydrate reserve and energy source and an indispensable dietary material for humans (→ Sugar). For centuries, sucrose has been the world's most plentiful produced organic compound of low molecular mass, the present annual production from sugar-cane and sugar beet being an impressive 130×10^6 t [24].

α,α-Trehalose, a non-reducing D-glucosyl D-glucoside occurs extensively in the lower species of the plant kingdom (fungi, young mushrooms, yeasts, lichens, and algae). In baker's yeast it accounts for as much as 15 % of the dry mass, in the metabolic cycle of insects it circulates like glucose does in the mammalian

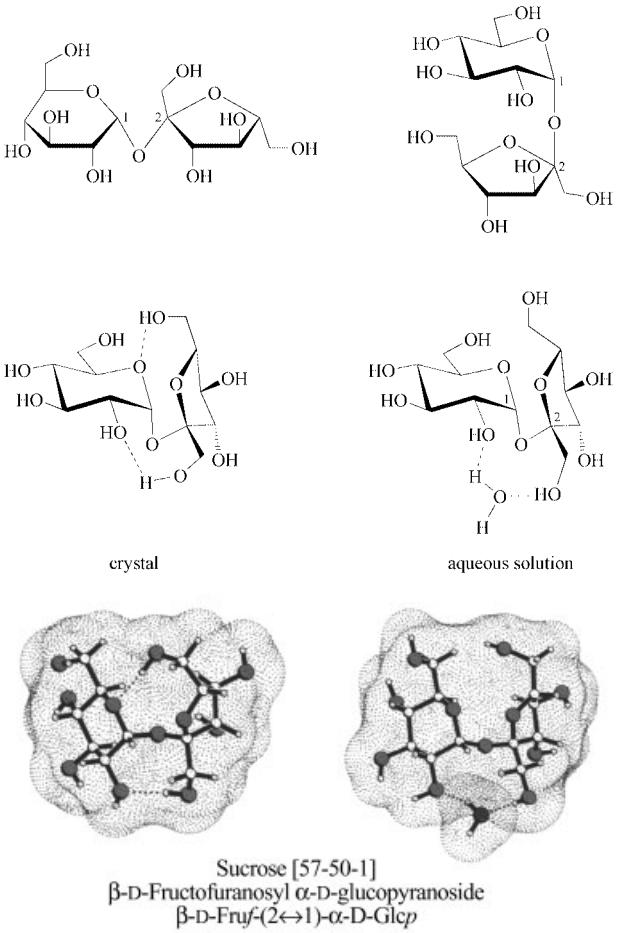

Figure 10. Common structural representations of sucrose (top entries), the molecular geometry realized in the crystal featuring two intramolecular hydrogen bonds between the glucose and fructose portion [18], [19] (bottom left), and the sterically similar disposition of the two sugar units towards each other in aqueous solution form, caused by hydrogen bonding through a 'water bridge' [25]. The bottom entries show the solvent-accessible surfaces (dotted areas) of the crystal form (left) and the form adopted in water [25] (right), clearly demonstrating that sucrose has an unusually compact overall shape, more so than any other disaccharide.

cycle. Similarly nonreducing, due to being a galactosylated sucrose, is the trisaccharide *raffinose*, distributed almost as widely in the plant kingdom as sucrose, yet in lower concentration (e.g., less than 0.05 % in sugar beets).

There are only very few naturally occurring oligosaccharides with a free anomeric hydroxyl group, which therefore possess reducing properties. The most important example is *lactose* (milk sugar, → Lactose and Derivatives), an ingredient of the milk of mammals (up to 5 % in cows). As it is produced on an industrial scale from whey, it represents the only large-scale

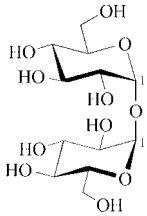

α,α-Trehalose [*99-20-7*]
α-D-Glucopyranosyl
α-D-glucopyranoside
(α-D-Glc*p*[1↔1]α-D-Glc*p*)

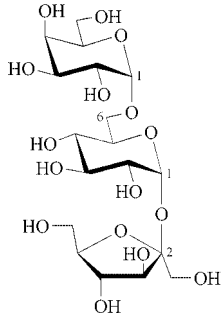

Raffinose [*512-69-6*]
α-D-Galactopyranosyl-
(1→6)-sucrose
(α-D-Gal*p*-(1→6)-α-D-
Glc*p*(1↔2)-β-D-Fru*f*)

available sugar derived from animal rather than plant sources. Uses include human food, pharmaceuticals, and animal feeds. The reducing gluco-disaccharides *cellobiose* and *maltose* (malt sugar) are chemical or enzymatic hydrolysis products of the polysaccharides cellulose and starch, respectively, and, hence are not regarded as native oligosaccharides.

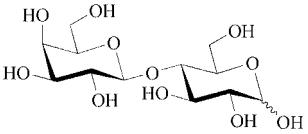

Lactose [*63-42-3*]
β-D-Galactopyranosyl-(1→4)-D-glucopyranose
(β-D-Gal*p*-[1→4]-D-Glc*p*)

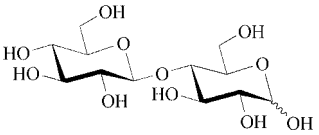

Cellobiose [*528-50-7*]
β-D-Glucopyranosyl-(1→4)-D-glucopyranose
(β-D-Glc*p*-(1→4)-D-Glc*p*)

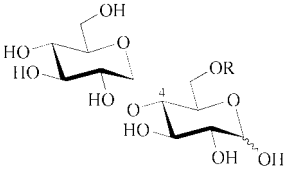

Maltose [*69-79-4*]
α-D-Glucopyranosyl-(1→4)-D-glucopyranose
(α-D-Glc*p*-(1→4)-D-Glc*p*)

Isomaltulose (palatinose, → Sugar Alcohols, Chap. 5.1.) and *lactulose*, both produced in fairly large amounts from sucrose and lactose, respectively, are 6-*O*-glucosyl- and 4-*O*-galactosyl-fructoses. The sucrose → isomaltulose transformation, industrially realized at a 40 000 t/a-scale [26], is effected by a *Protaminobacter rubrum*-induced glucosyl shift from the anomeric fructosyl oxygen to its *O*-6, taking place in mostly intramolecular fashion via a closed-shell intermediate [27], whilst the generation of lactulose from lactose, presently running at a 12 000 t/a level, comprises a base-promoted C-1 → C-2 carbonyl shift. Most of the isomaltulose produced is subsequently hydrogenated to isomalt (→ Sugar Alcohols, Chap. 5.2.), a low-calorie sweetener with the same taste profile as sucrose, lactulose (→ Lactose and Derivatives, Chap. 2.1.) has medical and pharmaceutical applications, mainly for treating intestinal disorders.

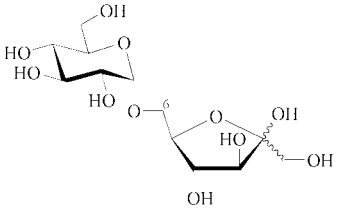

Isomaltulose [*13718-94-0*]
α-D-Glucopyranosyl-(1→6)-D-fructofuranose
(α-D-Glcp-(1→6)-D-Fru*f*)

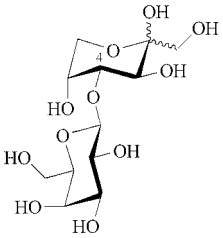

Lactulose [*4618-18-2*]
β-D-Galactopyranosyl-(1→4)-D-fructopyranose
(D-Galp-β(1→4)-D-Fru*f*)

Other Heterooligosaccharides. Heterooligosaccharides of considerably higher complexity occur in large variety in plants, animals and microorganisms where they are covalently bound to proteins ('glycoproteins') and lipids ('glycolipids') or other hydrophobic entities, and, as such, are implemented in a range of key biological processes: cell-cell recognition, fertilization, embryogenesis, neuronal development, hormone activities, the proliferation of cells and their organization into specific tissues, viral and bacterial infection and tumor cell metastasis [28–30]. Red blood cells, for example, carry carbohydrate antigens which determine blood group types in humans: type A people have the tetrasaccharide in Figure 11 with R = NHAc (i.e., a GalNAc residue) as a key antigen linked by lipid components to the surfaces of red blood cells; in type B blood, the tetrasaccharide determinant is exceedingly similar – formal replacement of NHAc by OH, i.e., GalNAc by

Gal – yet on mixing with type A blood leads to clumping and precipitation [31].

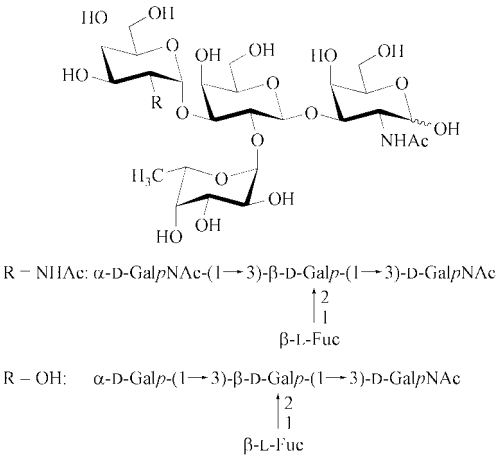

Figure 11. Human blood groups determinants: Differentiation between type A (R = NHAc) and B (R = OH) is effected by relatively simple changes within a branched tetrasaccharide linked to lipid components on the surface of red blood cells.

All *N*-glycoproteins (*N*-glycans) share the peptide-linked pentasaccharide fragment in Figure 12, consisting of three mannose units in a branched arrangement and two GlcNAc residues, of which the terminal one is *N*-glycosidically linked to an asparagine moiety of the protein. Branching out from this uniform core region are monosaccharides and oligosaccharide chains of high structural diversity leading to multiple types of branched and unbranched glycoproteins [32].

3.2. Cyclodextrins (→ Cyclodextrins)

Although discovered more than 100 years ago, the cyclic glucooligosaccharides termed cyclodextrins (based on dextrose, which is an old name for glucose) remained laboratory curiosities until the 1970s when they started to be used commercially [33]. Their large-scale production is based upon the degradation of starch by enzymes elaborated by *Bacillus macerans* ('CGTases'), involving excision and reconnection of single turns from the helical α-(1→4)-glucan (amylose) chain (cf. Fig. 13) to provide cyclic α-(1→4)-linked glucooligosaccharides with six, seven and eight glucose units. They are named α-, β- and γ-cyclodextrin, respectively.

Cyclodextrins are truncated cones with well-defined cavities. All secondary hydroxyl groups are located at the wider rim of the cone leaving the primary CH_2OH groups to protrude from the narrower opening. The respective cavities, as exemplified by that of α-cyclodextrin with its six glucose units (Fig. 14, [34], [35]), are distinctly hydrophobic in character, and show an amazing propensity to form stable complexes with a large variety of equally hydrophobic, sterically fitting guest molecules by incorporating them into their cavities [33], [34], changing the physical and chemical properties of the included guest. The features and properties of the resulting cyclodextrin inclusion compounds has led to the exploitation of cyclodextrins for a wide variety of purposes: as drug carriers [36], [37], as stationary phases for the separation of enantiomers [38], [39], as building blocks for supramolecular structures [40], and as enzyme models [41].

4. Polysaccharides

The bulk of the annually renewable carbohydrate-biomass are polysaccharides (glycans), such as cellulose, hemicelluloses, chitin, starch, and inulin. Invariably composed of monosaccharide units, they have high molecular masses and, hence, differ significantly in their physical properties. The majority of naturally occurring polysaccharides contain 80 – 100 units, with a few though made up of considerably more.

Cellulose (→ Cellulose) [*9004-34-6*] is an unbranched glucan composed of β-(1→4)-linked D-glucopyranosyl units (see Fig. 15) with an average molecular mass equivalent to about 5000 units. It is the most abundant organic material found in the plant kingdom, forming the principal constituent of the cell walls of higher plants and providing them with their structural strength. Cotton wool is almost pure cellulose, but in wood, the other chief source of the polymer, cellulose is found in close association with other polysaccharides (mainly hemicelluloses) and lignin. X-ray analysis and electron microscopy indicate that these long chains lie side by side in bundles, held together by a multiplicity of hydrogen bonds between the numerous neighboring OH groups. These bundles are twisted to-

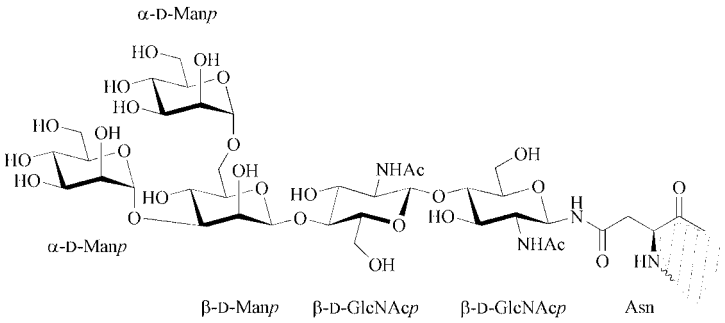

Figure 12. Central core region common to all *N*-glycoproteins is a pentasaccharide, *N*-glycosidically linked to the carbamido nitrogen of an asparagine moiety (Asn) within the peptide chain

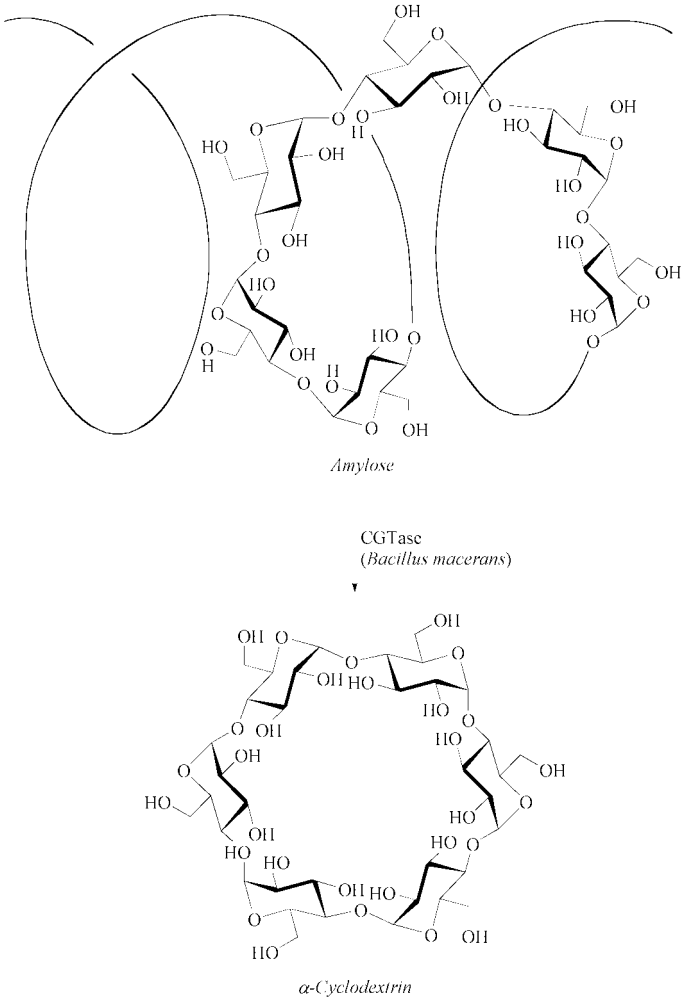

Figure 13. Sketch representation of a left-handed, single-stranded helix of V_H-amylose (top), and of α-cyclodextrin (bottom), which de facto represents a single turn of the amylose helix excised and re-connected by *Bacillus macerans*-derived enzymes (CGTases). The close analogy allows to consider V_H-amylose as a tubular analog of α-cyclodextrin.

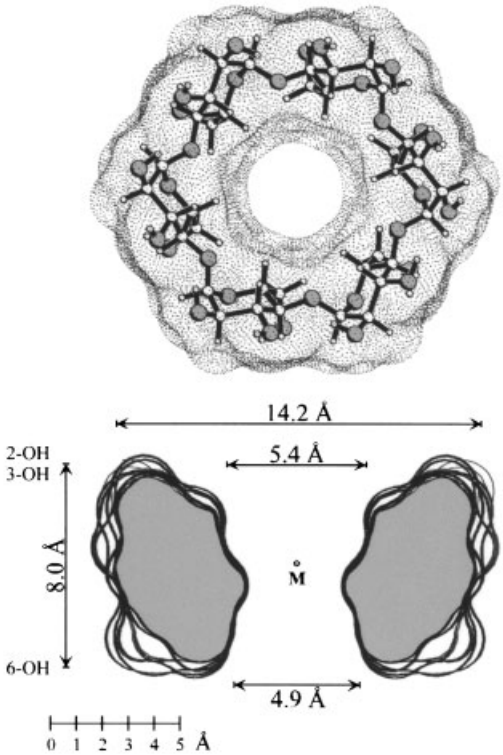

Figure 14. Top: Ball-and-stick model representations of the X-ray-derived solid-state structure of α-cyclodextrin, together with its solvent-accessible surface, shown as a dotted pattern.
Bottom: Cross section contour of a plane perpendicular to the macrocycle's mean plane with approximate molecular dimensions [34], [35].

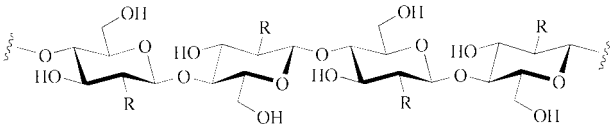

Figure 15. Structural representations of segments of cellulose (R = OH), chitin (R = NHAc), and chitosan (R = NH_2).

gether to form rope-like structures, which themselves are grouped to form the fibers that can be seen. In wood (→ Wood, Chap. 1.) these cellulose "ropes" are embedded in lignin to give a structure that has been likened to reinforced concrete.

Chitin (→ Chitin and Chitosan) is a polysaccharide composed of β-(1→4)-linked 2-acetamido-2-deoxy-D-glucopyranosyl residues (cf. Fig. 15), of which about one out of every six is not acetylated. Chitin is the major organic component of the exoskeleton (shells) of insects, crabs, lobsters, etc. and, hence, an abundant byproduct of the fishing industries. *Chitosan*, a related water-soluble polysaccharide in which the vast majority of residues is not acetylated (i.e., a β-(1→4)-linked chain of 2-amino-2-deoxy-D-glucose residues), can be obtained from chitin by deacetylation in concentrated sodium hydroxide solution.

Starches (→ Starch). The principal food-reserve polysaccharides in the plant kingdom are starches. They form the major source of carbohydrates in the human diet and are therefore of great economic importance, being isolated on an industrial scale from many sources. The two components, amylose and amylopectin, vary in relative amount among the different sources from less than 2 % of amylose in waxy maize to about 80 % of amylose in amylomaize (both

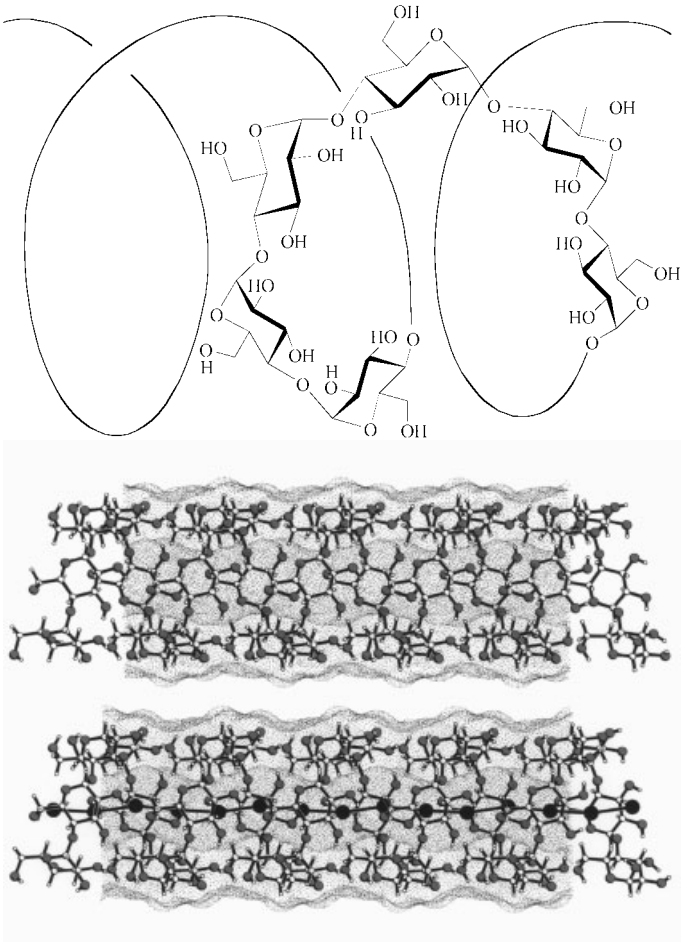

Figure 16. Sketch illustration of a left-handed, single stranded helix of V_H-amylose (top) and more detailed representation of its molecular geometry based on X-ray diffraction data [42] and calculation of the solvent-accessible contact surfaces [43], indicated by dots with ball-and-stick models superimposed.
Center: The channel generated by the helical arrangement of the α-D-glucose residues, of dimensions corresponding to those of the cavity of α-cyclodextrin, is clearly apparent. The outside surface area of V_H-type amylose is uniformly hydrophilic (in conformity with its solubility in water) whereas the center channel is as distinctly hydrophobic – predestined to incorporate equally hydrophobic guests such as iodine or fatty acids [43].
Bottom: A linear polyiodide chain embedded into the channel, corresponding to the intense blue starch-iodine complex [43].

corn starches), but the majority of starches contain between 15 and 35 % amylose.

Amylose [*9005-82-7*] is made up of long chains, each containing 100 or more α-(1→4)-linked glucopyranosyl units which due to the kink in every α-glycosidic linkage tend to coil to helical segments with six glucose units forming one turn (see sketch in Fig. 13). Amylose is the fraction of starch that gives the intense blue color with iodine, which has its cause in the trapping of iodine molecules within the hydrophobic channel of the helical segments of the polysaccharide (Fig. 16).

Amylopectin [*9037-22-3*] is also an α-(1→4)-glucan, yet there is a branch point via *O*-6 about every 25 units. The molecular size of amylopectin is of the order of 10^6 D-glucose residues, making it one of the largest naturally occuring molecules. The secondary structure is characterized by a several hundred linear chains of about 20 – 25 glucose units each, which are connected in a variety of arrangements to give

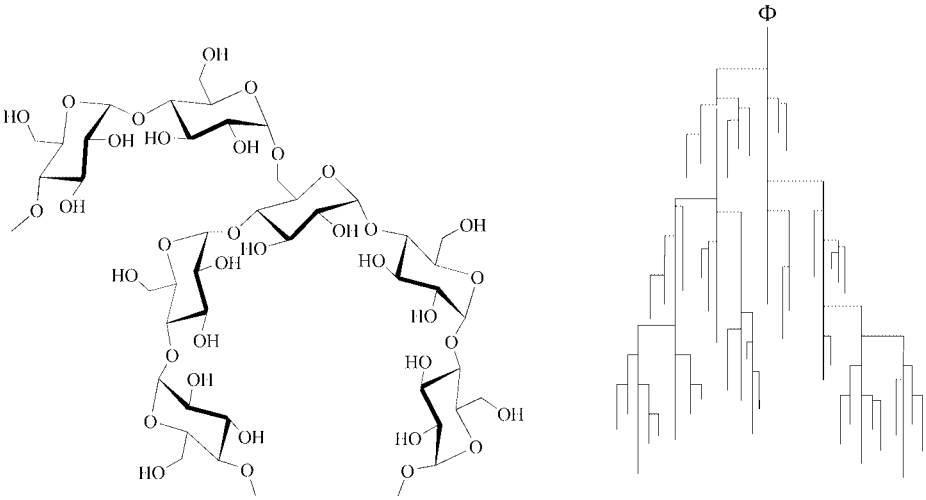

Figure 17. Schematic representation of a section of amylopectin with an $\alpha(1\rightarrow6)$-branch of a helical chain of $\alpha(1\rightarrow4)$-glucopyranosyl residues (left), with the tassel-on-string model of its higher level structure (ϕ = reducing end)

clusters for which the tassel-on-a-string model (see Fig. 17) has been proposed. With iodine, amylopectin produces only a dull-red color, indicating that the short linear chain portions cannot coil effectively to the helices required for formation of inclusion complexes.

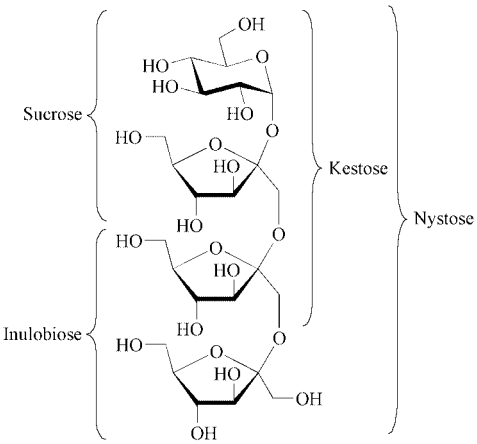

Figure 18. Nystose fragment of inulin, showing subfragments corresponding to sucrose, inulobiose and 1-kestose. Commercial inulins, e.g., those isolated from chicory, have a degree of polymerization far below that found in other polysaccharides, their molecular sizes ranging from around 5 to 30 units [43].

Dextrans (→ Dextran) are linear water-soluble α-(1→6)-glucans with only occasional branches via *O*-2, *O*-3 or *O*-4. They are generated from sucrose by a large number of organisms, of which *Leuconostoc mesenteroides* is used to produce the slightly branched commercial dextran, used clinically as a plasma volume expander.

Inulin is a polysaccharide composed of $\beta(1\rightarrow2)$-linked D-fructofuranose units with varying chain length of about 15 – 30 units. It is present to the extent of 30 % or more in various plants such as dahlia or Jerusalem artichoke where it replaces starch either partially or completely as the food storage carbohydrate [44], [45]. The structure of inulin is unique in leaving no 'reducing end', as this is glycosidically blocked by an α-D-glucopyranose residue – a sucrose unit in fact (Fig. 18).

Other Polysaccharides (→ Polysaccharides). A plethora of other homo- and heteropolysaccharides are abound in nature, most notably D-*xylans* (hemicelluloses with linear chains of β-(1→4)-D-xylopyranosyl units), *pectins* (principal constituent D-galacturonic acid), plant gums (building blocks D-galactose, L-arabinose, L-rhamnose) and various algal and microbial polysaccharides with, in part, unusual sugar units: L-guluronic and D-mannuronic acids in alginates, glucuronic acid and pyruvate acetals in *agar*, sulfated galactosyl residues in

carrageenans, or ribitol phosphates in *teichoic acids*. Excellent accounts on this subject have been given [46], [47].

5. Nomenclature

According to common practice, trivial names are used for monosaccharides and for many naturally occurring oligosaccharides. With the development of carbohydrate chemistry, however, and ever-increasing numbers of newly defined compounds, it has become necessary to introduce a semisystematic nomenclature which has been approved by the joint commission of IUPAC (International Union of Pure and Applied Chemistry) and IUB (International Union of Biochemistry) [48]. This nomenclature is based on the classical names for monosaccharides which appear, written in italics, as a "configurational prefix". For example, D-*xylo*, L-*arabino*, D-*gluco* refer to the distribution of asymmetric carbon atoms along a carbon chain of any length, designating the configuration of the corresponding monosaccharide.

Monosaccharides with an aldehydic carbonyl or potential aldehydic carbonyl group are called aldoses; those with a ketonic or potential ketonic carbonyl group, ketoses, with the chain length given by the root, such as pentose, hexose, or heptose and pentulose, hexulose, etc. In ketoses the position of the keto group is indicated by the position number. D-Fructose is systematically named D-*arabino*-2-hexulose.

Replacement of a hydroxyl group by hydrogen is indicated by the prefix deoxy, e.g., L-rhamnose is a 6-deoxy-L-aldohexose of *manno*-configuration. Replacement of a hydroxyl group by any other substituent is formally regarded as going via the deoxysugar. Thus a sugar with an amino group instead of OH is called an amino-deoxysugar. Formation of ether groups, most commonly methyl, is indicated by adding '*O*-methyl-' to the front of the name preceded by the number of the carbon atom whose hydroxyl group has been etherified. Esters, for example, acetates are shown by adding either '*O*-acetyl-' before the name or 'acetate' after it, in each case again preceding it with the appropriate carbon number.

The ring size is indicated by a suffix: pyranose for six-membered rings, furanose for five-membered rings, and pyranulose for six-membered ketose rings. The six-membered cyclic hemiacetal of D-fructose is named D-*arabino*-2-hexopyranulose. The symbol α or β for the anomeric configuration is always written together with the configurational symbol D or L (α-D, β-D, α-L, β-L).

Names of oligosaccharides are formed by combining the monosaccharide names, usually the trivial names. The nonreducing disaccharide sucrose is β-D-fructofuranosyl α-D-glucopyranoside. The endings "yl" and "ide" describe the fructose part as the aglycone and the glucose part as the glycone in this "glycoside". It is thus clearly indicated that both sugars are glycosidically linked by their anomeric hydroxyl groups. In reducing oligosaccharides the reducing monosaccharide is the root, and all attached monosaccharide units are named as substituents. The disaccharide lactose is therefore named β-D-galactopyranosyl-(1$\rightarrow$4)-D-glucopyranose. Position numbers and arrows indicate a β-configurated glycosidic bond between the anomeric hydroxyl group (carbon atom 1) of D-galactose (glyconic part) and the hydroxyl group at carbon atom 4 of D-glucose (aglyconic part).

For description of more complex oligosaccharides an abbreviation system has come into use - as for oligopeptides and oligonucleotides - that is unambiguous and practical: each monosaccharide is abbreviated by a three-letter symbol, comprising the first three letters of its trivial name, i.e.:

Glucose	Glc	Xylose	Xyl
Fructose	Fru	Arabinose	Ara
Galactose	Gal	Ribose	Rib
Mannose	Man	Deoxyribose	dRib
Fucose	Fuc	Glucosamine	GlcN
Rhamnose	Rha	*N*-Acetylglucosamine	GlcNAc

The anomeric configuration and D- or L-affiliation is written before the three-letter acronym, the ring size (*p* for pyranose, *f* for furanose) is added to the end, followed by the intersaccharidic linkage position in the case of oligosaccharides. Sucrose (Fig. 10), accordingly, is β-D-Fru*f*-(2$\rightarrow$1)-α-D-Glc*p*, lactose β-D-Gal*p*-(1$\rightarrow$4)-D-Glc*p*.

6. General Reactions

6.1. Hydrolysis

The hydrolysis of disaccharides such as sucrose and lactose, or polysaccharides like starch and cellulosic materials to their free component sugars is of great importance in the food and fermentation industries. It can be effected by enzymes called glycosidases or by acid treatment.

Enzymatic hydrolysis proceeds with high specificity towards both the sugar and the configuration at the anomeric center. Maltase, an α-D-glucosidase obtainable from barley malt, catalyses the hydrolysis of α-linked di-, oligo- and polysaccharides (sucrose, maltose, starch, dextrans), whereas the almond emulsin-derived enzyme is a β-glucosidase cleaving only β-linked glycosides.

Acid-induced hydrolysis of glycosides requires comparatively harsh conditions, standards being 1 M sulfuric acid at 100 °C for 4 h for hexose-containing polysaccharides and 0.25 M H_2SO_4 at 70 °C for pentosans. Partial degradation of the resulting monosaccharides can usually not be avoided whatever conditions are used [49], [50]. Thus, aside from enzymatic hydrolysis, acid hydrolysis of *starch* is industrially performed on a 10^6 t/a basis, the resulting D-glucose being used in liquid form (corn syrup) as a sweetener (→ Glucose- and Glucose-Containing Syrups, Chap. 4.). Another significant industrial product, in fact the only large-volume organic chemical prepared from carbohydrate sources (~ 200 000 t/a), is furfural (2-furaldehyde) [51], [52]. The technical process involves exposure of agricultural or forestry wastes to aqueous acid and fairly high temperatures, the *pentosans* first being hydrolyzed to pentoses, then undergoing cyclodehydration (→ Furan and Derivatives, Chap. 3.).

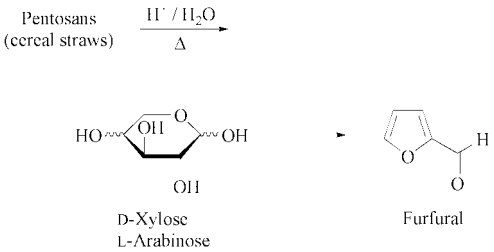

Similarly accessible by acid-induced elimination of three moles of water from *fructose* or *inulin hydrolysates* is 5-hydroxymethylfurfural (HMF) [26], [53].

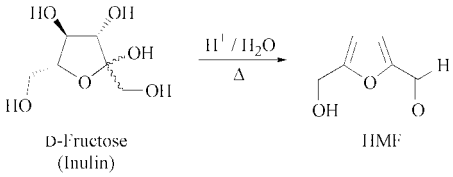

When using nonaqueous conditions, e.g., DMSO as the solvent and a strongly acidic resin, the fructose part of disaccharides such as isomaltulose can similarly be converted into the respective, glucosylated HMF-derivative (GMF) without cleaving the acid-sensitive glycosidic linkage [54].

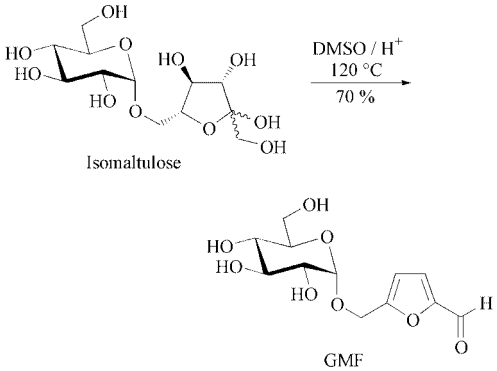

The trisaccharide *raffinose*, a storage carbohydrate in many plants can be cleaved enzymatically with α-galactosidase into sucrose and galactose. This reaction is used in the beet sugar industry to increase the yield of sucrose, as well as to improve the digestibility of food from leguminous plants. Raffinose can also be fermented by baker's yeast to form melibiose (α-D-Gal*p*-(1→6)-D-Glc*p*).

6.2. Isomerization

Under basic conditions aldoses isomerize to their C-2 epimers and the corresponding ketoses. Specific conditions may be applied for the preparation of particular products. In 0.035 % aqueous sodium hydroxide at 35 °C for 100 h, for example, either of the three following sugars is

converted into an equilibrium mixture containing D-glucose (57 %), D-fructose (28 %), and D-mannose (3 %). This interconversion is known as the *Lobry de Bruyn – van Ekenstein rearrangement* [55], which occurs by enolization of either sugar to the 1,2-enediolate – the mechanism being best visualized in the Fischer projection formulae (Fig. 19). In favorable cases alkali-promoted stereoisomerizations can be of preparative use, especially when the starting sugar is relatively abundant and when structural features minimize competing reactions. Thus lactulose can be satisfactorily made by epimerization of lactose (→ Lactose and Derivatives, Chap. 2.1.2.), or maltulose from maltose [56], [57], using either sodium hydroxide alone or with borate or aluminate as coreagents.

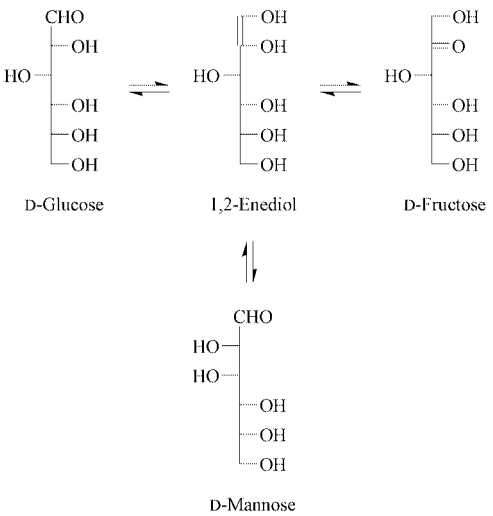

Figure 19. Lobry de Bruyn – van Ekenstein rearrangement

Alternatively, C-2-epimerization without ketose involvement can be induced by use of molybdate under mildly acidic conditions. This remarkable transformation (*Bilik reaction*) [58] involves a C-1 / C-2 interchange within the carbon skeleton.

6.3. Decomposition

Exposure of carbohydrates to high temperatures leads to decomposition (dehydration) with darkening (caramelization). This can be used to produce the caramel color, e.g., the color of cola beverages. Thermal decomposition in the presence of amino acids (*Maillard reaction* [59]) is responsible for many color- and flavor-forming reactions, such as in baking of bread and roasting of meat or coffee. The highly complex Maillard reaction, elicited during cooking or the preservation of food, involves condensations, Amadori-type rearrangements of glycosylamine intermediates, and degradations. The dark-colored products formed are responsible for the nonenzymic browning observed with various foodstuffs.

7. Reactions at the Carbonyl Group

In solution, reducing sugars establish an equilibrium between their pyranoid and furanoid hemiacetal forms via the open-chain carbonyl species. Although the latter is present only to a very minor extent, equilibration between the different forms is fast, so that reducing sugars undergo the typical carbonyl reactions with *O*-, *N*-, *S*-, and *C*-nucleophiles.

7.1. Glycosides

With alcohols in the presence of acid catalysts reducing sugars give the respective full acetals, called glycosides (Fischer glycosidation) [60]. Depending on the distribution of furanoid and pyranoid tautomeric forms in the reaction mixture, not only glycosides with different ring sizes, i.e., glycopyranosides and glycofuranosides, can result, but also the α- and β-anomers of each. Thus, when D-glucose is heated with methanol in the presence of anhydrous hydrogen chloride, pure crystalline methyl α-D-glucopyranoside can be isolated in 90 % yield, whilst the same reaction with D-galactose yields a mixture of the two furanoid and pyranoid methyl galactosides, from which the methyl α-D-galactopyranoside can be obtained in crystalline form in 41 % yield only.

Although the Fischer glycosidation presents one of the easiest means for preparing glycosides, synthesis of more complex members of this series, particularly the construction of the biologically important heterooligosaccharides widely distributed in nature, requires the use of more sophisticated methodologies. These preparative techniques generally involve the

coupling of suitably OH-group protected glycosyl donors (i.e., glycosides with an anomeric leaving group) with an alcohol component – usually a mono-, di-, or oligosaccharide in which the hydroxyls carry protecting groups [61] except for the one to be glycosylated ('the glycosyl acceptor'). Effective glycosyl donors, derived from D-glucose, are listed in Table 1. They represent the presently most suitable donors for achieving glycosidic bond-forming reactions with high stereocontrol: glycosyl chlorides and bromides [62], 2-oxoglycosyl bromides [63], [64], anomeric phosphates [65] and trichloracetimidates [66], thioglycosides [67], glycosyl sulfoxides [68], and 1,2-anhydrides [69], some of these methodologies being amenable to combinatorial and solid phase synthesis [70].

Table 1. Established glycosyl donors for the stereoselective synthesis of oligosaccharides

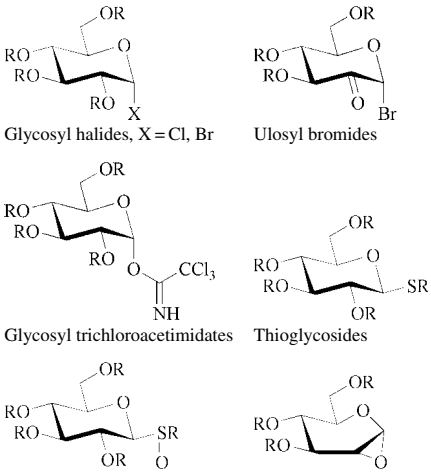

For further details on this subject, presently under intense further exploration, some recent general treatments [29], [71–73] are recommended.

7.2. Thioacetals and Thioglycosides

Sugars react rapidly with alkanethiols in the presence of acid catalysts at room temperature to give acyclic dialkyl dithioacetals as the main products [74], and therefore the reaction is markedly different from the Fischer glycosidation. These open-chain compounds can be used to prepare monosaccharide derivatives with a free carbonyl group, such as 2,3,4,5,6-penta-*O*-acetyl-D-glucose:

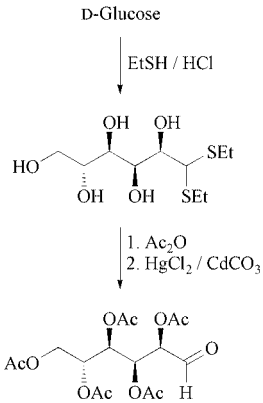

1-Thioglycosides, established glycosyl donors in oligosaccharide syntheses (upon activation with methyl trifluoromethanesulfonate or other promoters), have to be prepared indirectly, e.g., from peracylated pyranoses (or their 1-halides) by exposure to thiols in the presence of BF_3 etherate or zinc chloride:

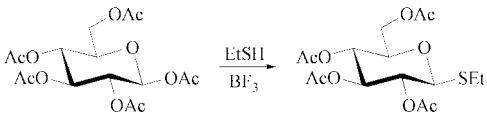

7.3. Glycosylamines, Hydrazones, and Osazones

Aldoses condense with ammonia and with primary and secondary amines upon loss of water – reactions that are analogous to the Fischer glycosidation. The initial condensation products appear to be the open-chain aldimines which then cyclize to the *glycosylamines* – also called *N*-glycosides, Thereby, the pyranose forms are preferentially adopted as these are thermodynamically more stable. Accordingly, D-glucose reacts with aniline in methanol to the α- and β-*N*-glucopyranosides:

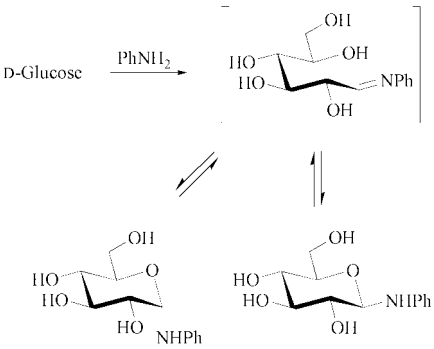

Acids also catalyze a transformation called the *Amadori rearrangement* [75] which often accompanies attempts to prepare glycosylamines from aldoses and amines. This reaction is related to the *Lobry de Bruyn-van Ekenstein reaction* of aldoses involving the rearrangement of *N*-alkylamino-D-glucopyranosides into 1-alkylamino-1-deoxy-D-fructoses (Fig. 20).

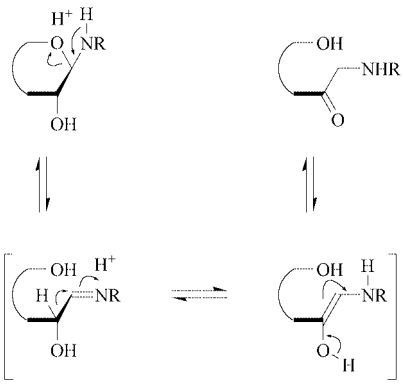

Figure 20. Amadori rearrangement of glycosyl amines induced by acid catalysis: D-Glucopyranosylamine is converted into 1-alkylamino-D-fructose [74]

Glycosylamine derivatives are probably involved in the complex Maillard reaction [59], whereby sugars, amines, and amino acids (proteins) condense, rearrange, and degrade during cooking or the preservation of food. Hydrazones and osazones result when aldoses or ketoses are reacted with hydrazine or arylhydrazines, the product depending on the conditions used [76]. With hydrazine acetate in highly acidic medium in the cold, hydrazones are formed, which initially adopt the acyclic structure, but tautomerize in aqueous solution to the cyclic glycosylhydrazine forms. However, when free sugars are treated with an excess of phenylhydrazine, the reaction proceeds further to give – in a formal oxidation of the vicinal 2-OH – the highly crystalline, water-insoluble phenylosazones which contain two phenylhydrazine residues per molecule, with a third phenylhydrazine molecule being converted into aniline and ammonia. As C-2 of a sugar is involved in this process, D-glucose, D-mannose, and D-fructose yield the same product:

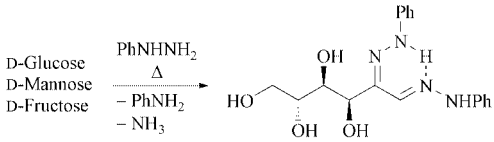

D-Glucosazone

This result played a fundamental role in Emil Fischer's elucidation of the configurational interrelationships of the sugars, eventually leading [10], [11] to the sugar family trees depicted in Figures 3 and 4.

7.4. Chain Extension

The carbonyl group offers excellent opportunities for extension of the sugar chains and the formation of 'higher' sugars. However, only a few carbon nucleophiles can be applied directly to the free sugars, i.e., without protection of the OH groups. The classical methods comprise the addition of cyanide ion (*Kiliani – Fischer extension*) [77] and of nitromethane under suitable alkaline conditions [78]. In either case, the cyano and nitromethylene group newly introduced can be converted into an aldehyde functionality by hydrolysis of the diastereomeric cyanohydrins to aldonic acids, lactonization and subsequent reduction, or by applying the *Nef reaction* to the nitroalditols, thus providing methods for ascent of the sugar series. For example, the rare sugar D-allose can be readily prepared from D-ribose [79] (Fig. 21), whereas the nitromethane addition approach allows the acquisition of the equally scarce hexoses L-glucose and L-mannose from L-arabinose [80] (Fig. 22).

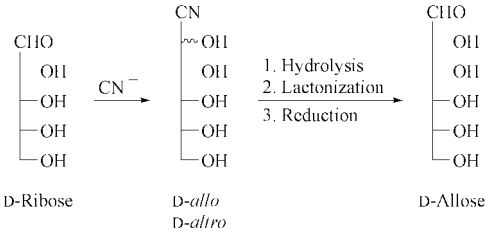

Figure 21. The Kiliani – Fischer cyanohydrin synthesis with D-ribose: the approximate 1:1 mixture of the 2-epimeric D-*allo*- and D-*altro*-cyanohydrins can be separated at the aldonic acid stage, subsequent reduction of the D-allonic acid in the form of its 1,4-lactone then providing D-allose (34 % overall yield)

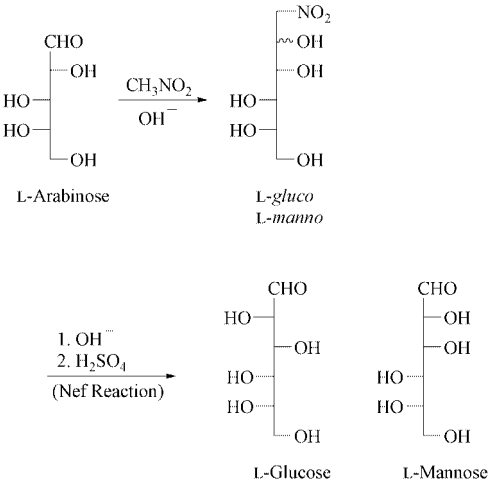

Figure 22. Nitromethane addition to L-arabinose in alkaline medium generates a mixture of the 2-epimeric L-nitroalditols (of L-*gluco* and L-*manno* configuration) which upon separation are subjected to the Nef reaction [78]

A special case of chain extension by nitromethane is the cyclization of sugar-derived dialdehydes – readily and quantitatively obtained from anomerically blocked glycopyranosides or furanosides by periodate oxidation – to give 3-nitrosugars [81], [82]. As exemplified for methyl β-D-glucopyranoside (Fig. 23), a mixture of 3-nitrohexosides is primarily obtained from which the major product, the D-*gluco* isomer crystallizes. Subsequent catalytic hydrogenation then provides the 3-amino-3-deoxy-D-glucoside [83].

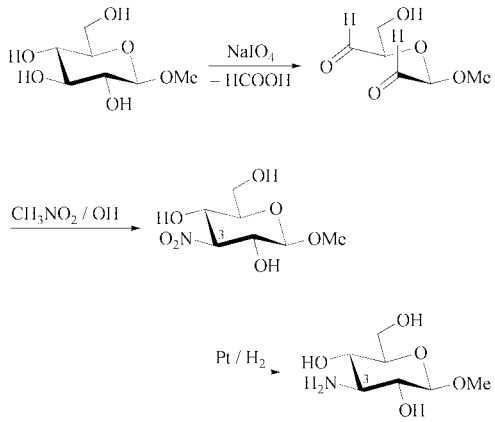

Figure 23. Conversion of methyl β-D-glucoside into its 3-amino-3-deoxy-derivative via the dialdehyde-nitromethane cyclization approach [81]

This nitromethane cyclization sequence can be extended to nitroalkanes, e.g., nitroethane or even nitroacetate [84], thus providing a ready access – upon hydrogenation of the nitro group – to 3-methyl- or 3–carboxy-branched 3-aminosugars.

7.5. Chain Degradation

The removal of a terminal carbon atom from a sugar or sugar derivative to leave an aldehyde group is realizable in a variety of ways but the yields are often poor. The most practical approach involves conversion of an aldose to the corresponding dialkyl dithioacetal (mercaptal) by reaction with an alkanethiol, then oxidation to the bis-sulfone with a peracid. Treatment of the bis-sulfone with dilute ammonia causes expulsion of the stabilized bis(ethylsulfonyl)methyl carbanion and gives the aldose with one carbon atom less. This three-step protocol smoothly converts D-glucose, for example, into D-arabinose [85]:

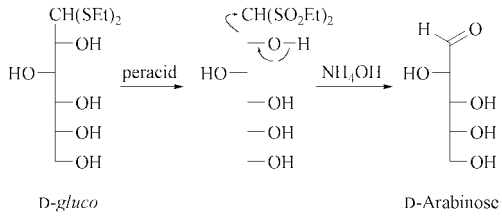

7.6. Reductions to Alditols

Aldoses and ketoses can readily be reduced to alditols (Table 2) with the generation of a new alcoholic group from the carbonyl functions, their names being derived from the respective aldose by replacing the 'ose' suffix with 'itol'. Thus, reduction of D-glucose gives D-glucitol. Originally, sodium amalgam was the reducing agent most commonly used for these reductions, but now it has been superseded by others, particularly sodium borohydride in aqueous solution, or for alkali-sensitive sugars, by $NaBH_3CN$ in acetic acid. High pressure hydrogenation of aldoses and ketoses over rare metal catalysts, especially nickel is used for the commercial preparation of alditols (→ Sugar Alcohols).

Table 2. Low-caloric, non-cariogenic sugar alcohol sweeteners, obtained by catalytic hydrogenation of the parent aldoses. Recommended nomenclature [48] for sorbitol is D-glucitol (hence D-Glc-ol). Being a *meso* compound, xylitol requires no D- or L-prefix .

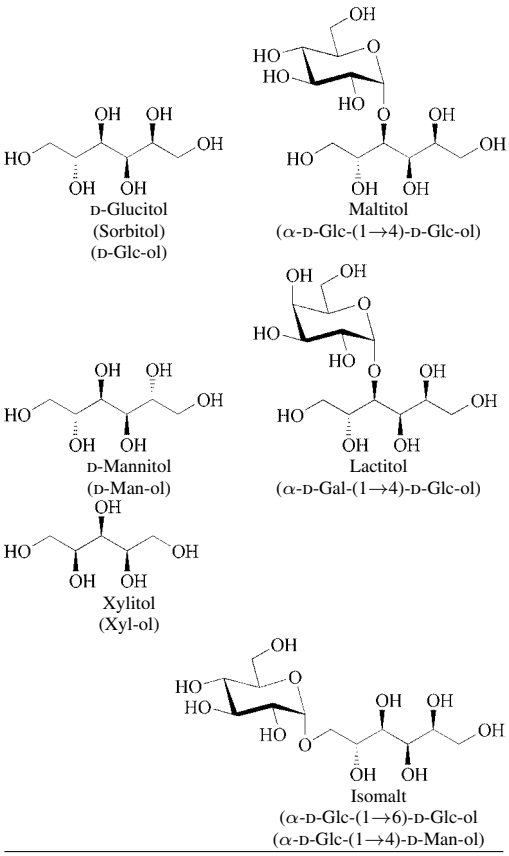

d-Glucitol [*50-70-4*] (→ Sugar Alcohols, Chap. 3.), common name sorbitol, produced at a level of 900 000 t/a worldwide, has a sweet taste and is used in foods for diabetics [86], [87]. It is also the synthetic precursor for ascorbic acid (vitamin C), with about 20 % of the annual production sorbitol going to this use. Sorbitol is used as a humectant in cosmetic and pharmaceutical formulations and in foods. It is also applied as an alcoholic component in the preparation of rigid polyurethane foams. Fatty acid esters of monoanhydrosorbitol (1,4-sorbitan) are widely used as emulsifiers and non-ionic surfactants. The mono- and dinitrate esters of 1,4 : 3,6-dianhydrosorbitol (isosorbide [*652-67-5*]) are coronary vasodilators.

d-Mannitol [*87-78-5*] (→ Sugar Alcohols, Chap. 4.), is prepared by hydrogenation of the fructose portion of invert sugar [86], [88], which yields a mixture of mannitol and sorbitol. In contrast to sorbitol, mannitol is not hygroscopic; world production in 2000 was approximately 30 000 t. Mannitol is used in the manufacture of dry electrolytic condensers and synthetic resins; in the pharmaceutical industry as a diluent for solids and liquids and in the preparation of the vasodilator mannitol hexanitrate; in the food industry as an anticaking and free-flow agent, and as a lubricant, stabilizer, and nutritive sweetener.

Other sugar alcohols mostly used as sweeteners, are *xylitol*, obtained by catalytic hydrogenation of D-xylose, which in turn is acquired from wood xylans or maize cobs by acid hydrolysis, and a series of disaccharide alcohols, each manufactured analogously from the respective parent disaccharide: *maltitol* (from partially hydrolyzed starch syrup; Lycasin [88] contains a high proportion thereof), *lactitol* and, most notably, *isomalt*, which due to its mild, pleasant sweetness, ready crystallizability and excellent thermal stability appears presently the most prevailing. Isomalt, also called "Palatinit", consists of an approximate 1:1 mixture of α-D-glucosyl-(1→6)-D-sorbitol and α-D-glucosyl-(1→1)-D-mannitol (see Table 2). The latter forms a dihydrate, the two water molecules being attached to the mannitol portion in a hydrogen-bonded water bridge [89]. Isomalt is produced from sucrose through *Protaminobacter rubrum*-induced

isomerization to isomaltulose ('palatinose') and subsequent catalytic high-pressure hydrogenation (→ Sugar Alcohols, Chap. 5.2.).

7.7. Oxidation

Controlled stoichiometric oxidations of carbohydrates to yield glyconic acids or their derivatives are limited to aldoses. Such oxidations can be carried out almost quantitatively either enzymatically by dehydrogenases or oxidases or chemically with bromine or iodine in buffered solution. Under these conditions, D-glucose – through its pyranose form prevailing in solution – is directly converted into the 1,5-lactone of D-gluconic acid (i.e., the internal ester rather than the free acid), which on addition of base is converted to the salt (in open-chain form). However, by crystallization from aqueous solution it is possible to obtain the free acid or the 1,4-lactone. For different aldonic acids the amounts of each form present at equilibrium vary with structure and with pH of the solution, in contrast to the free sugars, the five-membered ring lactones are relatively favored.

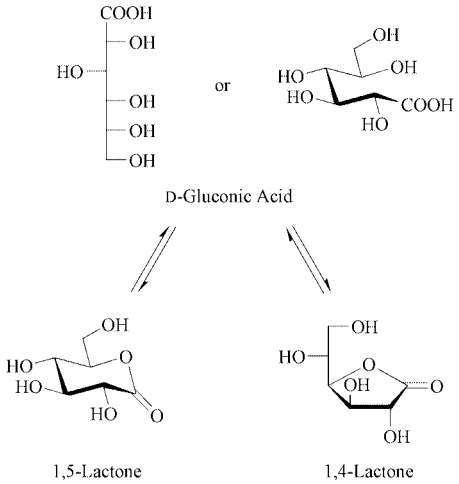

Strong nitric acid appears to be one of the few oxidants which is able to also oxidize the terminal primary hydroxyl group of aldoses but leave the secondary hydroxyl groups unchanged. D-Glucose treated with this reagent gives D-*glucaric acid* [90], its name being derived by replacing the ending 'ose' in the sugar by 'aric acid'. Aldaric acids can form mono- or dilactones, in the case of D-glucaric acid, the well crystallizing form is the furanoid 1,4-lactone:

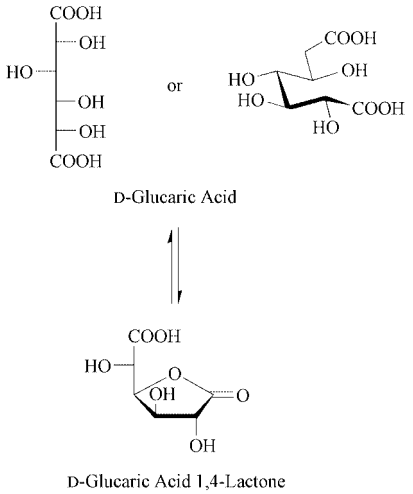

Under the influence of very strong oxidizing agents such as potassium dichromate or permanganate, sugars suffer oxidative degradation. Hence, these agents are of no preparative use.

8. Reactions at the Hydroxyl Groups

8.1. Ethers

The most simple compounds of this type are *methyl ethers* which occur in a range of natural carbohydrates. Methyl esthers belong to the most stable *O*-substituted sugar derivatives, such that per-*O*-methylated hexoses can even be distilled. Traditionally, the labeling of free OH-groups in polysaccharides is effected by methylation, structural analysis being then based on the *O*-methyl sugars obtained on hydrolysis. Methyl ethers are conveniently prepared from methyl bromide, iodide, or sulfate in polar aprotic solvents such as dimethylformamide or dimethylsulfoxide. Agents for deprotonation of the hydroxyl group and for binding the mineral acids liberated include alkali hydroxides or hydrides and barium or silver oxide. With high molecular mass carbohydrates, quantitative deprotonation is best carried out with sodium or potassium methylsulfinyl methanide (the conjugate base of dimethyl sulfoxide) [91].

Benzyl ethers are amongst the most commonly used protecting groups in carbohydrate chemistry [92], as the *O*-benzyl moiety is easily removed by hydrogenolysis (Pd/C, H_2) to yield the respective alcohol and toluene [61]. For the preparation of benzyl ethers, traditional methods involve such reagents as benzyl halides in combination with sodium hydroxide, sodium hydride, or silver oxide.

Triphenylmethyl (trityl) ethers are used mainly for the temporary substitution of primary hydroxyl groups and are usually prepared using trityl chloride in pyridine. Under mild acidic conditions, e.g., acetic acid or boron trifluoride in methanol, the trityl ethers are readily cleaved.

Trimethylsilyl ethers, although extremely sensitive both to base- and acid-catalyzed hydrolysis, are often used in analytical and preparative carbohydrate chemistry. The pertrimethylsilyl ethers of monosaccharides and small oligosaccharides are relatively volatile, highly lipophilic, and thermostable, and therefore, ideal derivatives for gas chromatographic analysis. The trimethylsilyl ethers are rapidly formed in pyridine by using a mixture of hexamethyldisilazane and trimethylchlorosilane [93].

Cellulose ethers (→ Cellulose Ethers) are generally manufactured by the Williamson synthesis: reaction of sodium cellulose (prepared by treating cellulose with 20 to >50 % sodium hydroxide) with an organic halide such as chloromethane or sodium monochloroacetate. The latter reagent produces sodium carboxymethyl cellulose (NaCMC), which is widely used, for example, as a thickening agent in foods. Worldwide production of NaCMC is in the range of several hundred thousand tons per year.

8.2. Esters of Inorganic Acids

Phosphoric acid esters of sugars play vital roles in such fundamental processes as the biosynthesis and metabolism of sugars and, hence, are present in every organism, the most important esters being D-glucose 1-phosphate [*59-56-3*], D-glucose 6-phosphate [*56-73-5*], and D-fructose 1,6-diphosphate [*488-69-7*]. In addition, phosphates of D-ribose and its 2-deoxy derivative form fundamental components of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) and of various coenzymes (→ Nucleic Acids).

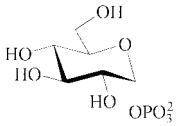
α-D-Glucose-1-phosphate

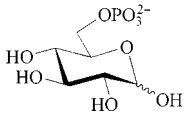
D-Glucose-6-phosphate

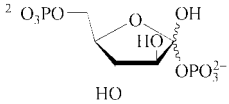
D-Fructose-1,6-diphosphate

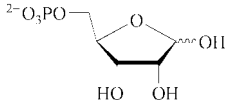
D-Ribose-5-phosphate

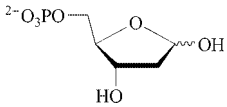
2-Deoxy-D-ribose-5-phosphate

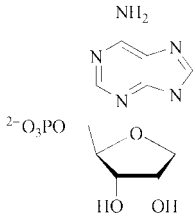
Adenosine-5-phosphate (AMP)

Both chemical and enzymic methods are available for the synthesis of specific phosphates. Chemically, anomeric phosphate esters are usually prepared either from glycosyl halides or other glycosyl donors by reaction with silver dibenzyl phosphate, whilst phosphorylation of nonanomeric hydroxyl groups is effected with specifically blocked sugar derivatives and diphenyl or dibenzyl phosphorochloridate [94]. Biochemically, phosphates are produced by the action of phosphatases on provided substrates [95].

Sulfate Esters. Sulfate groups are present in many biologically important polysaccharides, such as heparin and chondroitin sulfate. Sulfated monosaccharides can be prepared from suitable monosaccharide derivatives by reaction with chlorosulfuric acid in pyridine [96].

Nitrate esters of carbohydrates [97] are not found in nature, yet a large variety ranging from monoesters to peresters have been prepared, favorable conditions being cold nitric acid/acetic anhydride for nonanomeric OH groups, whilst anomeric nitrate esters are accessible via reaction of acyl glycosyl halides with silver nitrate. Sugar mono- and dinitrates are stable crystalline

compounds, such as, e.g., adenosine mononitrate (AMN), the nitrate analog to AMP [98], and the dinitrate of 1,4:3,6-dianhydro-D-glucitol ('isosorbide dinitrate'), which is in broad pharmaceutical use as a coronary vasodilator [99]:

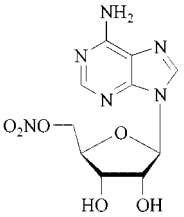

Adenosine 5-nitrate (AMN)

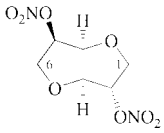

Isosorbide dinitrate

More highly substituted derivatives are heat and shock sensitive, such as, e.g., mannitol hexanitrate or nitrate esters of cellulose (nitrocellulose). These contain as many as three ONO_2 groups per glucose unit. The product with about 13 % nitrogen is the well-known guncotton (→ Cellulose Esters, Chap. 1.), whereas celluloid is nitrocellulose containing about 10 % nitrogen plasticized with camphor; it is one of the oldest known plastics. This plastic was once the principal photographic and movie film, but has been replaced by other films because of its high flammability.

8.3. Esters of Organic Acids

For the esterification of the hydroxyl group of free or partially otherwise blocked sugars, acyl halides or acid anhydrides are usually used, e.g., acetic anhydride/sodium acetate or zinc chloride or acetic anhydride/pyridine readily yield the respective *peracetates*.

Perbenzoylation can be effected with benzoyl chloride in pyridine or benzoyl cyanide in acetonitrile with triethylamine as the catalyst. Tertiary OH groups, present in ketoses or branched-chain sugars, usually require the addition of 4-(dimethylamino)pyridine.

Whereas peracetates and perbenzoates of simple sugars are important intermediates for the preparation of the respective glycosyl halides and, hence, acylated glycals and hydroxyglycals (see Section 8.4), those of some polysaccharides are of industrial relevance. Acetate esters of cellulose are manufactured on a large scale, whereby the degree of acetylation determines the solubility and use: the triacetate (3.0-acetate) is soluble in chloroform, the 2.5-acetate in acetone, and the 0.7-acetate in water. These esters, as well as mixed cellulose acetate/propionate and acetate/butyrate are widely used in the production of lacquers, films, and plastics (→ Cellulose Esters, Chap. 2.1.).

Polysaccharide esters in which the carbohydrate portion is the acid component occur in the plant kingdom in fruits, roots, and leaves. For example, *pectins* are high molecular mass polygalacturonic acids joined by α-(1→4)-glycosidic links, in which some of the carboxylic acid groups are esterified with methanol (→ Polysaccharides, Chap. 3.). In the production of fruit juices the formation of methanol, which can be liberated through the action of pectinesterases, should be avoided. Pectins in which 55 – 80 % of the carboxyl groups are esterified are called high-methoxyl pectins (HM-pectins), and have the important property of gelling at very low concentrations ($\approx$0.5 %) in water in the presence of sugars and acid. Low-methoxyl (LM, $<$50 % of the carboxyl groups esterified) pectins form gels with divalent cations such as a Ca^{2+}; 0.5 % of a low-methoxyl pectin can bind 99.5 % of the water in the gel matrix. These pectins can be used as gelling agents in the production of jellies from fruit juices.

8.4. Acylated Glycosyl Halides

Per-*O*-acylated monosaccharides can be converted smoothly into glycosyl halides by dissolving them in cold solutions of the hydrogen halide in glacial acetic acid (acetates of acid-sensitive oligosaccharides may undergo cleavage of glycosidic bonds). Because of the dominance of the anomeric effect in the pyranosyl cases, the anomer with axial halide is substantially preferred. Accordingly, on acylation and subsequent HBr-treatment, usually performed as a one pot operation, D-glucose yields the 2,3,4,6-tetra-*O*-acetyl-α-D-glucopyranosyl bromide ('acetobromoglucose', R = Ac in Fig. 24) or its benzoylated, pivaloylated (R = *tert*-BuCO) or benzylated (R = $C_6H_5CH_2$) analogues [62]. These halides are commonly used directly for glycosylation reactions, which is the basis of the traditional Koenigs-Knorr procedure [100], or

converted into more elaborated glycosyl donors (see Table 1).

Glycosyl halides are also of significance in terms of the use of monosaccharides as inexpensive enantiopure starting materials for the construction of complex, non carbohydrate natural products [101–103], which usually require the reduction of the number of chiral centers paired with the introduction of olefinic or carbonyl unsaturation. Treatment of glycosyl halides with zinc/acetic acid [104] – or, preparatively more efficient zinc/1-methylimidazole in ethyl acetate under reflux [105] – results in reductive elimination to give the glycal, in Figure 24 illustrated with the formation of tri-*O*-acetyl-D-glucal. Simple 1,2-elimination of hydrogen bromide using diethylamine in acetonitrile in the presence of tetrabutylammonium bromide or by 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) in DMF [106], [107] yields the respective 2-hydroxyglycal esters. The D-glucose-derived benzoylated example in Figure 24 is an ester of the enol form of 1,5-anhydro-D-fructose.

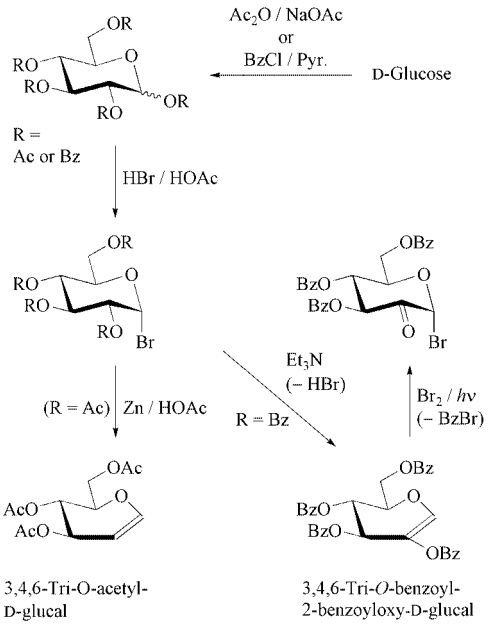

Figure 24. Formation of glycosyl and 2-oxoglycosyl ('ulosyl') bromides from peracylated monosaccharides, as exemplified for the D-glucose case, and their conversions into glucal and 2-hydroxyglucal esters

Endowed with high crystallinity and shelf stability, the hydroxyglycal esters are of considerable preparative interest not only for the generation of a plethora of other unsaturated compounds, e.g., pyranoid enones [108] and enolones [102], [109], but also as precursors for the highly versatile ulosyl bromides, produced in high yields simply by exposure to NBS or bromine in the presence of ethanol [110], [111]. The utility of these ulosyl bromides as glycosyl donors in the straightforward synthesis of β-D-mannosides has been amply demonstrated [62], [63], [112].

8.5. Acetals

Acetals are generally derived from the reaction of an aldehyde or ketone – benzaldehyde and acetone being the most common – with a geometrically suitable diol grouping, of which there is a large variety in free sugars, glycosides, and alditols [113–115]. The reactions are normally carried out in the reagent aldehyde or ketone as solvent with an electrophilic catalyst (H_2SO_4 or $ZnCl_2$). Acetal formation under these conditions is thermodynamically controlled and usually very specific. Ketones such as acetone or cyclohexanone predominantly bridge vicinal diols to form five-membered cyclic products (1,3-dioxolanes) as exemplified by the di-*O*-isopropylidene derivatives of D-glucose ('diacetone-glucose'), D-galactose and D-mannitol:

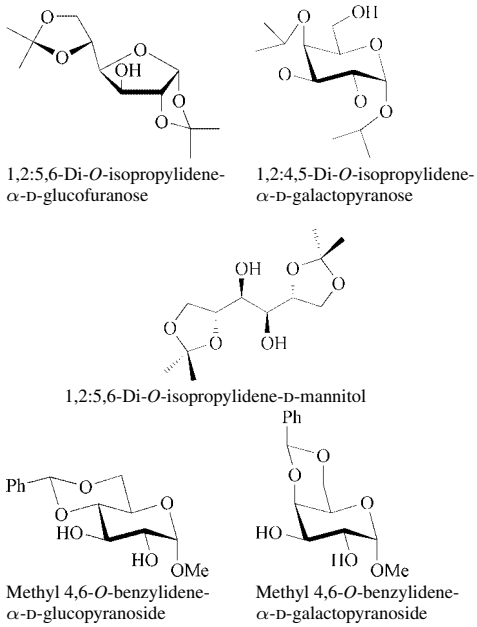

Aldehydes, however, show a distinct preference for 1,3-diols, as illustrated by the six-membered 4,6-*O*-benzylidene acetals of methyl D-glucoside and D-galactoside.

Introduction of cyclic acetal groups into sugars is simple and satisfactory in terms of yields. As cyclic acetals are stable towards alkali, the entire armory of organic reactions requiring basic conditions can be applied, and due to their ready removal with mild acid (e.g., 90 % aqueous trifluoroacetic acid at room temperature), they provide indispensable intermediates in preparative carbohydrate chemistry.

9. Carbohydrates as Organic Raw Materials

As our fossil raw materials are irrevocably decreasing – the end of cheap oil is realistically prognosticated for 2040 at the latest [116–118] – and as the pressure on our environment is building up, the progressive changeover of chemical industry to renewable feedstocks emerges as an inevitable necessity [118], [119].

The terrestrial biomass is considerably more complex than fossil raw materials, constituting a multifaceted accumulation of low and high molecular mass products. Carbohydrates, the most abundant of these materials, aside their traditional uses for food, lumber, paper and heat, are the major biofeedstocks to develop industrially and economically viable organic chemicals that are to replace those derived from petrochemical sources.

The bulk of the annually renewable carbohydrate-biomass consists of polysaccharides, yet their nonfood utilization is confined to textile, paper, and coating industries, either as such or in the form of simple esters and ethers. Organic commodity chemicals, however, are low molecular mass products; hence, they are more expediently acquired from low molecular mass carbohydrates than from polysaccharides. This in turn means that polysaccharides usually must be hydrolyzed before being further processed to organic commodity chemicals.

Table 3 lists the availability and bulk-quantity prices of the eight least expensive sugars – all well below €10/kg – as compared to some sugar-derived compounds and basic chemicals from petrochemical sources. The result is stunning, since the five cheapest sugars, some sugar-alcohols, and sugar-derived acids are not only cheaper than any other enantiopure product, such as hydroxy- or amino acids, but they compare favorably with basic organic bulk chemicals such as acetaldehyde or aniline. Actually, the first three of these sugars, sucrose, glucose, and lactose, are in the price range of some of the standard organic solvents.

Table 3. Annual production volume and prices of simple sugars, sugar-derived alcohols, and acids as compared to some petrochemically derived basic chemicals and solvents

	World production∗, t/a	Price ∗∗, €/kg
Sugars		
Sucrose	130 000 000	0.30
D-Glucose	5 000 000	0.60
Lactose	295 000	0.60
D-Fructose	60 000	1.00
Isomaltulose	50 000	2.00
Maltose	3000	3.00
D-Xylose	25 000	4.50
L-Sorbose	60 000	7.50
Sugar alcohols		
D-Sorbitol	900 000	1.80
D-Xylitol	30 000	5.00
D-Mannitol	50 000	8.00
Sugar-derived acids		
D-Gluconic acid	60 000	1.40
L-Lactic acid	> 100 000	1.75
Citric acid	500 000	2.50
L-Tartaric acid	35 000	6.00
Amino acids		
L-Lysine	40 000	5.50
L-Glutamic acid	500 000	7.00
Basic chemicals		
Aniline	1 300 000	0.95
Acetaldehyde	900 000	1.15
Adipic acid	1 500 000	1.70
Solvents		
Methanol	25 000 000	0.15
Toluene	6 500 000	0.25
Acetone	3 200 000	0.55

∗ Reliable data are only available for the world production of sucrose, the figure given referring to the crop cycle 2000/2001 [23]. All other data are average values based on estimates from producers and/or suppliers, as the production volume of many products is not publicly available.
∗∗ Prices given are those attainable in early 2002 for bulk delivery of crystalline material (where applicable) based on pricing information from sugar industry (sugars) and *The Chemical Market Reporter* **2002**, no. 2, 16 – 19 (acids, basic chemicals, and solvents). The listings are intended as a benchmark rather than as a basis for negotiations between producers and customers. Quotations for less pure products are, in part, sizably lower, e.g., for the commercial sweetener "high fructose syrup", which contains up to 95 % fructose, and, thus, may readily be used for large-scale preparative purposes.

Despite their large-scale accessibility, chemical industry, at present, utilizes these mono- and disaccharides only to a minor extent as feed-

stock for organic chemicals. This is amply documented by the fact that of the 100 major organic chemicals manufactured in the USA in 1995 [120], only seven were derived from biofeedstocks, and five of these – ethanol, sorbitol, citric acid, lysine, and glutamic acid – used carbohydrates as the raw material source. Intense efforts within the last decade to boost the production of organic chemicals from the sugars in Table 3 [121–127] have not basically changed this picture.

There are various reasons for that: at present, the use of fossil raw materials is more economic and the process technology for conversion of petrochemical raw materials into organic chemicals is exceedingly well developed and basically different from that required for transforming carbohydrates into products with industrial application profiles. This situation originates from the inherently different chemical structures of the two types of raw materials, of which the essence is manifested in their structure-based names.

Fossile Resources:	Renewable Resources:
HYDRO-CARBONS	CARBO-HYDRATES
C_nH_{2n+2}	$C_n(H_2O)_n$
oxygen-free,	overfunctionalized with
lacking functional groups	hydroxyl groups

Our fossil resources are *hydrocarbons*, distinctly hydrophobic, oxygen-free, and lacking functional groups, annually renewables are *carbohydrates*, overfunctionalized with hydroxyl groups and pronouncedly hydrophilic in nature. Needless to say, that the methods required for converting carbohydrates into viable industrial chemicals are diametrically opposed to those prevalent in petrochemical industry: microbial transformations or chemical processing with reduction of oxygen content and introduction of C=O and/or C=C unsaturation, or preferably both.

9.1. Microbial Synthesis

Microbial processing by direct fermentation of carbohydrates can be used to synthesize a large number of organic chemicals, yet only a few have clearly reached commodity status. They are *ethanol* (fermentation ethanol as distinguished from synthetic ethanol produced by hydration of ethylene), *citric acid* (from molasses by *Aspergillus niger*), *glutamic acid* and *lysine* (from glucose by *Brevibacterium lactofermentum* and others), and *gluconic acid* (from glucose by *Gluconobacter suboxydans*). Together, these chemicals have a sizable annual production (cf. Table 3), however, questions of how and when other microbially synthesized commodity chemicals will play a larger role in the world's chemical markets remain to be answered.

By contrast, there is a broad range of specialty chemicals that are manufactured today by microbial conversion of carbohydrate feedstocks, usually products with molecular structures too complex for conventional chemical synthesis. Examples are antibiotics (penicillins, kanamycins, tetracyclins) and vitamin C, B_2 (riboflavin), and B_{12} (cyanocobalamin). The possibility, however, for production of further high-value-added specialty chemicals or pharmaceuticals from carbohydrates is almost unlimited as long as the appropriate organism and substrate are allowed to interact under suitable conditions [128].

9.2. Chemical Conversions

Presently, the dominant methods of converting carbohydrates to organic chemicals – be it bulk, intermediate or fine chemicals, pharmaceuticals, agrochemicals, high-value-added specialty chemicals, or simply enantiopure building blocks for organic synthesis – are chemical.

9.2.1. Furan Derivatives

Furfural appears to be the only large-volume chemical (200 000 t/a) produced from carbohydrate sources, usually agricultural and forestry wastes [51], [52] (→ Furan and Derivatives, Chap. 3.). Due to its ready accessibility, the ensuing chemistry of furfural is well developed, providing a host of highly versatile industrial chemicals by simple straightforward operations (Fig. 25): furfuryl alcohol and its tetrahydro derivative by hydrogenation of furfural, furfurylamine through reductive amination, furoic acid by oxidation, and furanacrylic acid (Perkin reaction) or furylidene ketones via aldol condensations. Furfural is also the key chemical for the

commercial production of furan (through catalytic decarbonylation) and of tetrahydrofuran by hydrogenation, thereby providing a biomass-based alternative to its petrochemical production via dehydration of 1,4-butanediol. Further importance of these furanic chemicals results from their ring-cleavage chemistry, which has led to a variety of other established chemicals, e.g., levulinic acid (4-oxovaleric acid) and maleic anhydride [52], [129].

The bulk of the furfural produced is used as foundry sand linker in the refining of lubricating oil, and, together with furfurylalcohol and its tetrahydro derivative, enters into condensations with formaldehyde, phenol, acetone, or urea to yield a variety of resins of complex, ill-defined structures, yet excellent thermosetting properties, most notably high corrosion resistance, low fire hazard and extreme physical strength [52], [130].

5-(Hydroxymethyl)furfural (HMF). Like many petroleum-derived basic organic chemicals, e.g., adipic acid and hexamethylenediamine, 5-hydroxymethylfurfural (HMF) is a six-carbon commodity with high industrial potential, and, thus, has been termed "a key substance between carbohydrate chemistry and mineral oil-based industrial organic chemistry" [26]. It is readily accessible from fructose, or inulin hydrolysates by acid-induced elimination of three moles of water [53] and even a pilot-plant process has been developed [26].

Of high industrial potential as intermediate chemicals are the various HMF-derived products (Fig. 26), for which well worked-out, large scale-adaptable production protocols are available. Of these products, 2,5-bis(hydroxymethyl)furan, 5-hydroxymethyl-2-furoic acid, and the 2,5-dicarboxylic acid have extensively been exploited for the preparation of furanoic polyesters [136]. The diol has been reacted with various aliphatic and aromatic diacids; the ethyl ester of 5-hydroxymethylfuroic acid, upon polycondensation, gave a mixture of linear and cyclic products, whereas the furan diacid has been poly-esterified with a series of aliphatic diols or bisphenols. Even an all-furanic polyester has been successfully prepared from its respective monomeric components [136].

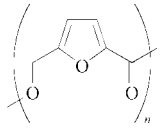

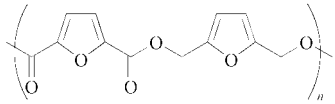

Another obvious outgrowth from the furan-2,5-dicarboxylic acid and the respective diamine was the generation of furanic polyamides, as they could potentially replace adipic or terephthalic acid, and, correspondingly, hexamethylenediamine or *p*-diaminobenzene in polyamides (see Section 9.2.5).

Despite this impressive array of useful HMF-derived intermediate chemicals, it is, as of now, not produced on an industrial scale. Obviously, the economic preconditions are not yet favorable enough. A recent assessment of the economics of HMF against competitive petrochemical raw materials [137] gives ample evidence thereof: prices of naphtha and ethylene are in the €150 – 400/t range, those of crude inulin or fructose (≈ €1000/t) give rise to an HMF-marketing price of at least €2500 per ton – too expensive at present for a bulk-scale industrial product. Accordingly, as long as the economic situation favors fossil raw materials, applications of HMF lie in high value-added products, such as pharmaceuticals or special niche materials. Prototype for this outlet is Ranitidine [138], an efficient antiulcer drug due to its potent oral inhibition of histamine-induced gastric acid secretion:

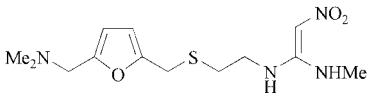

Ranitidine

Furans with a Tetrahydroxybutyl Side Chain. Another simple, one-step entry from hexoses to more highly substituted furans involves their $ZnCl_2$-mediated reaction with 1,3-dicarbonyl compounds such as ethyl acetoacetate or 2,4-pentanedione. As only the first

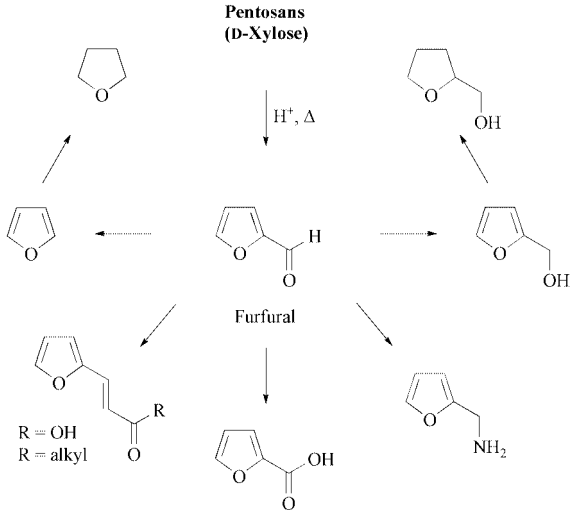

Figure 25. Versatile furanic commodity chemicals derived from pentosans in agricultural wastes (corn cobs, oat hulls, wood chips, bagasse)

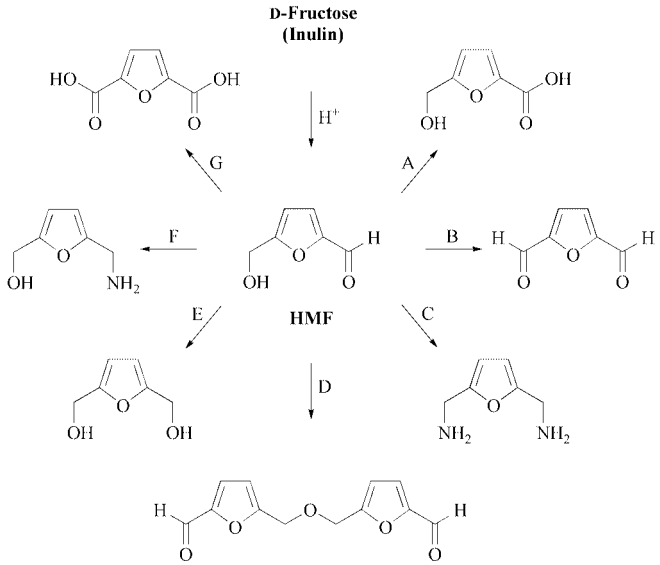

Figure 26. Versatile intermediate chemicals derived from hydroxymethylfurfural (HMF)

Key	A	Ag_2O, 100 °C, 75 % [130]	E	Pt, C / H_2, quant. [132]
	B	$BaMnO_4$, 93 % [130]	F	Ni/H_2, NH_3, 72 % [133]
	C	NH_2OH, then Ni/H_2, 33 % [130]	G	Pt, C / O_2, pH 7. 91 % [134]
	D	TsOH (H_2O↑), 89 % [131]		

two sugar carbons contribute to the formation of the furan, a distinctly hydrophilic tetrahydroxylbutyl side chain is introduced into the heterocycle. D-Glucose, e.g., smoothly and efficiently provides furans with D-*arabino*-configuration in the polyol fragment [139], [140], which can be shortened oxidatively to the dicarboxylic acid or a variety of other furanic building blocks:

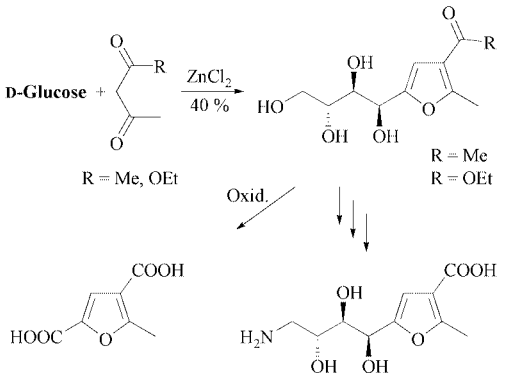

9.2.2. Pyrones and Dihydropyranones

The bulk scale-accessible mono- and disaccharides listed in Table 3 are preferentially adopting the pyranose cyclohemiacetal forms, from which efficient reaction pathways lead to an unusually large variety of unsaturated pyranoid building blocks, such as pyrones, dihydropyrans, and dihydropyranones.

The γ-pyrone kojic acid is readily obtained from D-glucose either enzymatically by *Aspergillus oryzae* (growing on steamed rice) [141] or chemically via pyranoid 3,2-enolones [102], [109]. An isomeric α-pyrone is produced from D-glucose by oxidation to D-gluconic acid and acetylation (1 h, 80 °C, 90 %) [142]. Both, at present, are of little significance as starting materials for preparative purposes, despite a surprisingly effective route to cyclopentanoid products [143] which are surmised to have industrial potential.

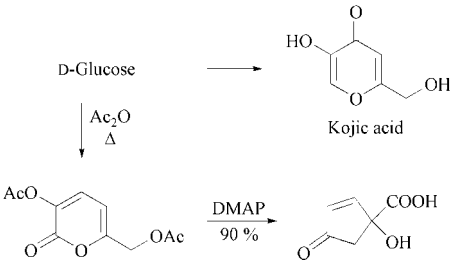

Table 4. Enantiopure six-carbon building blocks readily accessible in 3 – 5 straightforward steps from D-glucose via D-glucal ester as key intermediate (R = Ac, Bz) [107], [144] (R' = Me, Et)

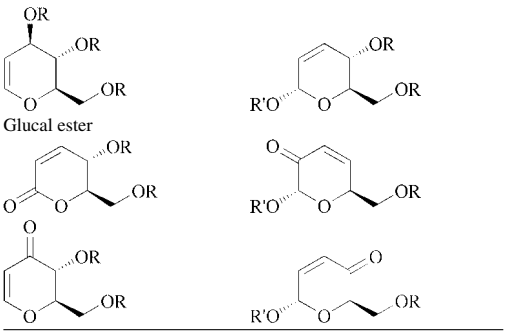

Of higher interest, at least with respect to their extensive use for the total synthesis of non carbohydrate natural products [101], [102], are the bevy of enantiopure six-carbon building blocks of the dihydropyran and dihydropyranone type in Tables 4 and 5, all compounds depicted being readily accessible from D-glucose in no more than three to five straightforward steps [102], [144].

Table 5. Pyranoid building blocks from D-glucose via hydroxyglucal esters (R = Ac, Bz; R′ = Me, Et) [106] as key intermediates.

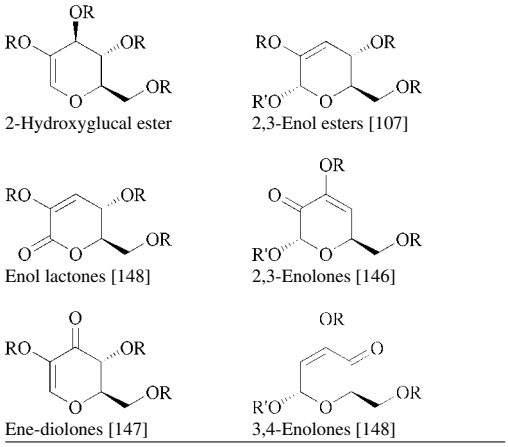

Some of these pyranoid building blocks are accessible even more directly, e.g., levoglucosenone, which has been used for the synthesis of a diverse variety of natural products [149]. Although the yield attainable from pyrolysis of waste paper [150] is low (3 – 4 %), relative large quantities can be amassed quickly. Similarly convenient are the preparations of the three glycosylated dihydropyranones requiring two, three, and four steps from maltose, sucrose, and lactose, respectively:

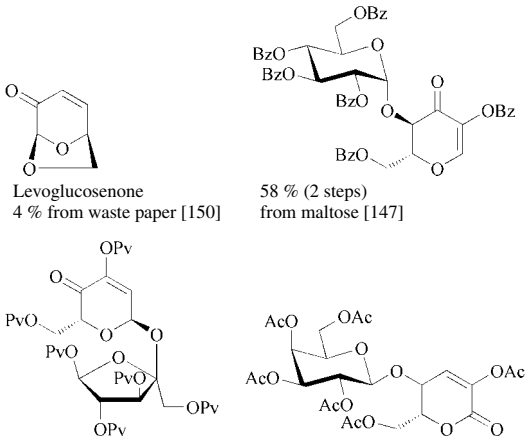

Ac = acetyl; Bz = benzoyl; Pv = pivalolyl

All of these pyranoid building blocks are enantiopure, and have a unique, highly diverse array of functional groups to which the huge arsenal of preparative organic methods can be applied directly, reflecting the extensive use of these building blocks in the total synthesis of non carbohydrate natural products in enantiopure form [101], [102], [149]. They have found little use, though, as high-value-added specialty chemicals. However, if suitable targets and appropriate preparative outlets can be found, particularly along pharmaceutical objectives towards biologically hopeful compound libraries via combinatorial techniques, these pyranoid building blocks are apt to become a plethora of attractive, industrially relevant specialty chemicals.

9.2.3. Sugar-Derived Unsaturated *N*-Heterocycles

Although transformation of sugars into trace amounts of *N*-heterocycles occurs extensively on exposure of foods to heat (Maillard reaction [59]), and despite the fact that various nitrogen heterocycles have been generated from saccharide derivatives [151], synthetic procedures meeting preparative standards are exceedingly scarce. Recent improvements of existing procedures and the development of new methodologies have led to the more ready access to various *N*-heterocycles from carbohydrates. Examples are imidazoles, pyrroles, pyrazoles, pyridines, and quinoxalines which due to their derivation from sugars have hydrophilic side chains.

Pyrroles. The formation of pyrroles by heating a glycerol solution of the lactose-derived ammonium salt of galactaric acid [152] over a free flame [153] appears to be the production process from a carbohydrate source giving the highest yield (40 %). However, this process does not seem to be utilized industrially, neither in this nor in modified form.

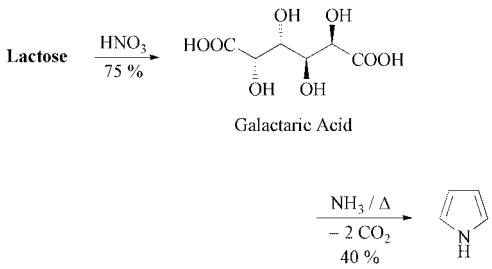

2,5-Disubstituted pyrroles are accessible from carbohydrate sources via HMF in a preparatively straightforward reaction sequence, involving photooxidative furan ring opening and cyclization of the saturated 2,5-diketones with ammonia or amines [154]:

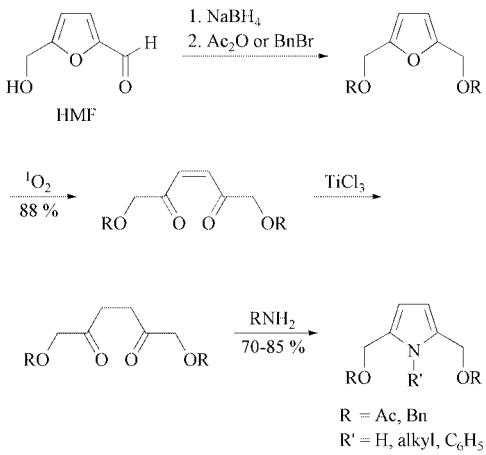

These reaction sequences can directly be transferred to GMF, leading to pyrroles carrying an additional glucosyl residue [154]. Pyrroles with an equally hydrophilic tetrahydroxybutyl substituent are available from D-glucosamine by exposure to acetylactone or ethyl acetoacetate under mildly basic conditions [155] or in a one-pot reaction from D-fructose by heating with acetylacetone and ammonium carbonate in DMSO [156].

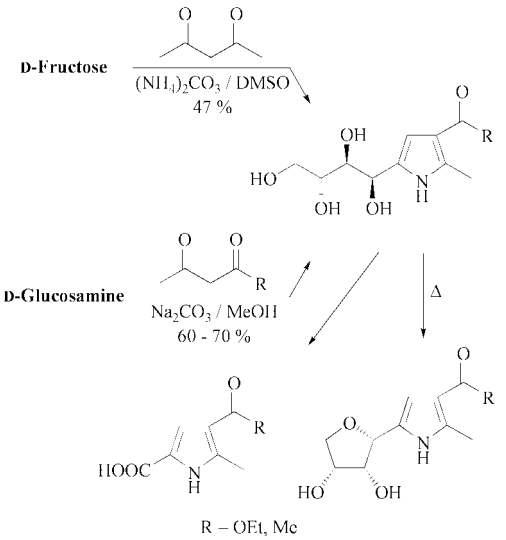

The hydroxylated side chain can, of course, be oxidatively shortened to give a variety of simple pyrrole building blocks, cyclized to a furanoid ring. These compounds may be considered as *C*-nucleosides [155].

Pyrazoles. An expeditious four-step approach to 1-phenylpyrazol-3-carboxaldehydes with a 5-hydroxymethyl, 5-dihydroxyethyl, or a 5-glucosyloxymethyl substituent has been elaborated starting from D-xylose [157], D-glucose, and isomaltulose [158], respectively.

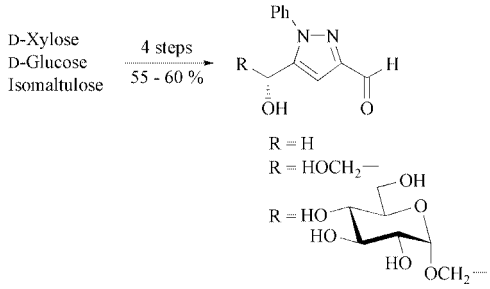

The osazone of D-xylose, nearly quantitatively formed on reaction with phenylhydrazine, straightforwardly cyclizes to the pyrazole upon addition to refluxing acetic anhydride. Subsequent removal of the *N*-acetylphenylhydrazone residue with formaldehyde/acetic acid and de-*O*-acetylation provides a pyrazole-aldehyde (57 % overall yield from D-xylose), a versatile heterocyclic building block, useful in the synthesis of pharmaceuticals or monomers for the production of polyamides and polyesters, e.g., in the form of its the diamino and diol derivatives [157]:

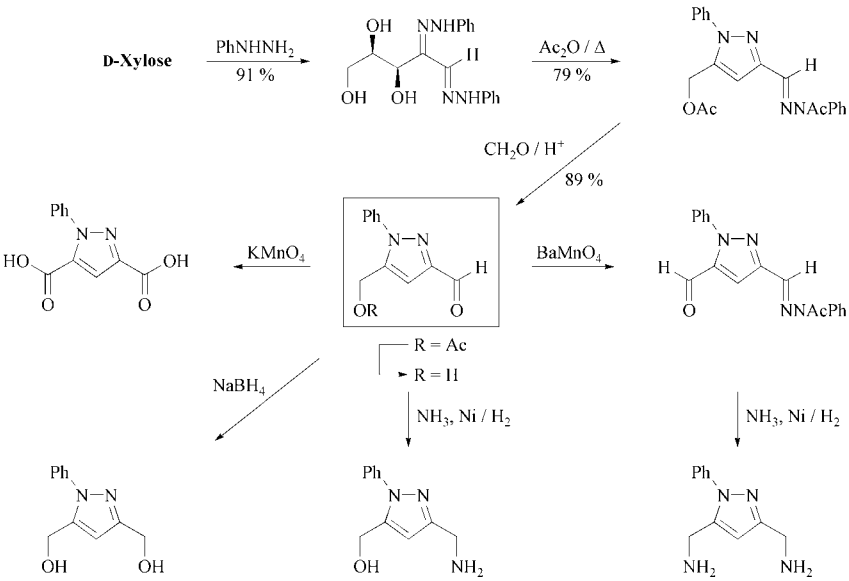

Imidazoles. Various imidazoles carrying hydrophilic substituents in the 4-position are readily accessible in one-pot procedures from the standard monosaccharides. Of those, the formation of 4-hydroxymethylimidazole by Cu(II)-promoted reaction of monosaccharides with formaldehyde and conc. ammonia [159] is rather unique, because obviously retroaldolization to glyceraldehyde and dihydroxyacetone is involved. The retroaldol fission can be partially suppressed when heating a monosaccharide, D-fructose for instance, with formamidinium acetate in liquid ammonia in a pressure vessel [160], or with formamidinium acetate in the presence of boric acid and hydrazine [161]. The latter reaction obviously proceeds via a boric acid complex of the bishydrazone of D-glucosone (2-ketoglucose).

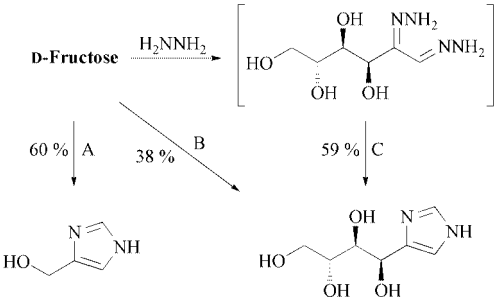

Conditions
A CH_2O, aq. NH_3, $CuCO3/Cu(OH)_2$, 2 h, 100 °C, [158]
B $HC(NH)NH_2 \cdot HOAc$/liq. NH_3, 15 h, 75 °C, [159]
C $N_2H_4/HC(NH)NH_2 \cdot HOAc$/aq. HOAc, H_3BO_3, 3 h reflux, [160]

These conditions can be readily applied to pentoses or disaccharides, as exemplified with D-xylose [160] and isomaltulose [161] in for one-pot procedures acceptable yields:

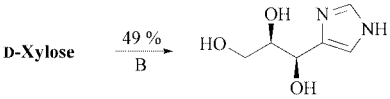

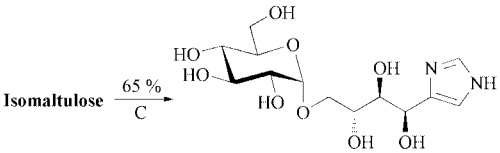

3-Pyridinols. The conversion of pentosans or pentoses into 3-pyridinol can be effected in a practical three-step sequence, involving acid-induced dehydration to furfural, reductive amination to furfurylamine, and subsequent oxidation wit hydrogen peroxide [134], [162], the last step conceivably proceeding through the stage of a 2,5-dihydroxy-2,5-dihydrofurfurylamine, which gives the pyridine nucleus via dehydration to a 5-aminopentenal intermediate and intramolecular aldimine formation. The pyridinol is an important intermediate in the preparation of herbicides and insecticides [163] as well as cholinergic drugs of the pyridostigmine type.

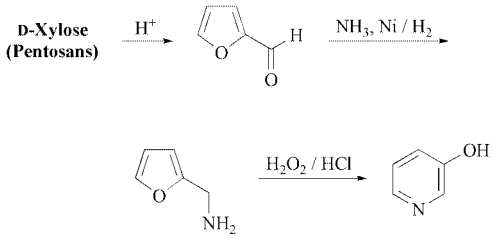

For the conversion of furfurylamines with oxidizable hydroxyl groups, e.g., those derived from fructose via HMF, the multistep process to the hydroxymethyl-pyridinol can be effected in a one-pot procedure simply by treatment with bromine in water-methanol at 0 °C [164] (see bottom of the next page).

Quinoxalines. Useful one-pot procedures are also available for the conversion of various monosaccharides into tetrahydroxybutyl substituted quinoxalines, the preparatively most favorable conditions seem to be reaction of fructose with hydrazine, *o*-phenylenediamine and boric acid in dilute acetic acid with bubbling oxygen through the solution [165], the decisive intermediate being the bishydrazone of D-glucosone:

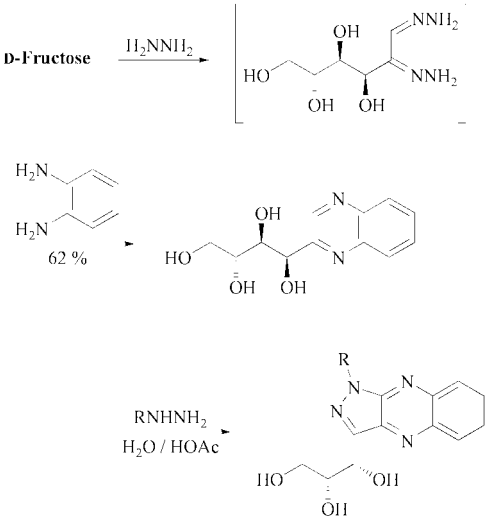

R = H (45 %) [166], R = Ph (90 %) [167]

On briefly refluxing the quinoxaline in aqueous acid with excess hydrazine or phenylhydrazine, a surprising oxidative cyclization takes place to form the trihydroxypropyl-substituted flavazols [166], [167].

9.2.4. Sugar-Based Surfactants

Utilization of cheap, bulk-scale accessible sugars as the hydrophilic component and fatty acids or fatty alcohol as the lipophilic part provides nonionic surfactants which are nontoxic, nonskin-irritating and fully biodegradable. The industrially relevant surfactants along this route are fatty acid esters of sorbitol and sucrose, fatty acid amides of 1-methylamino-1-deoxy-D-glucitol, and, most pronounced in terms of volume produced, fatty alcohol glucosides, the so-called alkyl polyglucosides (→ Surfactants, Chap. 7.4.; → Laundry Detergents, Chap. 3.1.2.6.).

Fatty Acid Esters of Sorbitol (Sorbitan Esters). Sorbitol, well accessible by catalytic hydrogenation of D-glucose (see Table 3), readily undergoes dehydration to "sorbitan" which constitutes a mixture of sorbitol, its 1,4-anhydro and 1,4:3,6-dianhydro derivatives, the exact composition depending on the conditions employed, usually acid catalysis. Esterification of this mixture is generally carried out with fatty acids or their methyl esters at 200 – 250 °C with 0.1 N NaOH. Depending on the amount of fatty acids used, sorbitan monoester ('SMS', R = C_{16}/C_{18} acyl), di- or triester ("SMT") are formed:

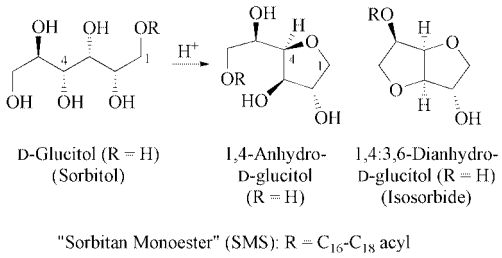

"Sorbitan Monoester" (SMS): R = C_{16}-C_{18} acyl

SMS is dispersible in water and soluble in fats and oils and has low hydrophilic lipophilic balance (HLB) values, besides as a surfactant SMS finds use as an emulsifier in desserts [168].

Sucrose Fatty Acid Esters. Surfactants based on sucrose have gained limited acceptance industrially due to the instability of its

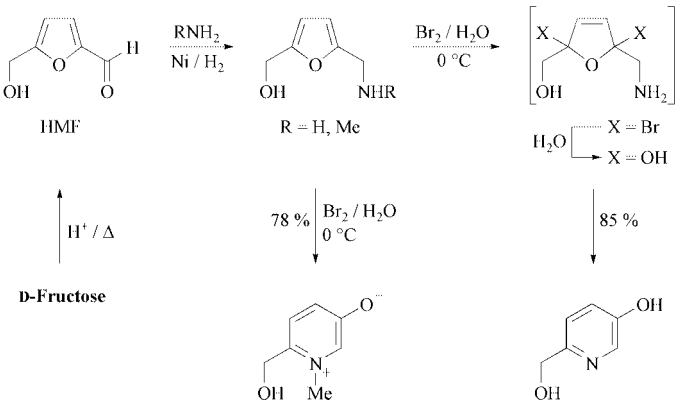

intersaccharidic bond, and its insolubility in conventional organic solvents. This requires the use of the rather toxic and expensive dimethylformamide or dimethylsulfoxide for effecting transesterification with fatty acid methyl esters and complete removal of the solvents. Alternately, transesterification of fats with sucrose without a solvent is also possible. In either case, the resulting sucrose monoester (if 1:1 molar ratios have been used in the esterification process) is not a defined product acylated exclusively at the primary glucose-6-OH, as indicated in the formula, but at the other primary and some of the secondary OH groups as well:

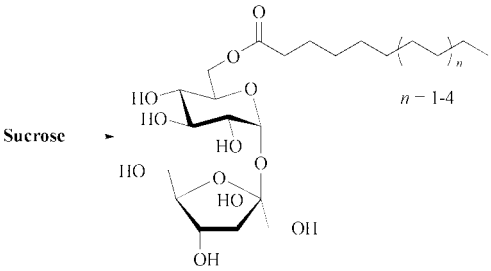

Sucrose fatty acid monoesters, presently produced at a 4000 t/a only, are mostly used in the food industry as emulsifiers and in cosmetic formulations because of their attractive physiological properties.

Whilst sucrose fatty acid mono- or diesters are biodegradable, fully or nearly completely esterified analogs (octa- or heptaesters) are not. The sucrose molecule is fully encased in a mantle of long-chain alkyl groups, so that enzymes have no way of binding to the sugar, which is a prerequisite for degradation. Thus, Olestra [169], a sucrose fatty acid hepta-/octaester, which has the consistency of an oil or a solid depending on the nature of fatty acid used, can be used as a noncaloric fat ('fat-free fat' [170]) in food applications including frying, due to its thermal stability.

***N*-Methyl-*N*-acyl-glucamides (NMGA).** Reductive amination of D-glucose with methylamine smoothly generates the respective aminoalditol, 1-methylamino-1-deoxy-D-glucitol, which on amidation with fatty acids gives the corresponding fatty acid amides, carrying a methyl group and a pentahydroxylated six-carbon chain at the amido nitrogen.

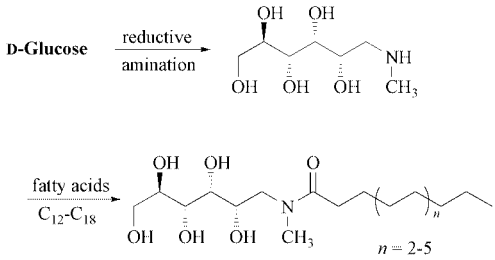

The NMGA's possess highly advantageous ecological and toxicological properties which brought about their use as surfactants, cleansing agents and cosmetic applications [171], [172] (→ Laundry Detergents, Chap. 3.1.2.5.).

Alkylpolyglucosides (APG) [173]. Presently produced in two plants with a capacity of up to 40 000 t/a each, APG's are by far the most important nonionic surfactants. Alkylpolyglucosides are fatty alcohol glucosides with an alcohol chain length normally between C_8 and C_{14}. Their industrial synthesis either comprises a direct acid-catalyzed Fischer glycosidation of glucose (in the form of a syrupy starch hydrolysate) or starch itself. The alternate process consists of two stages, the first being Fischer glycosidation with *n*-butanol to butyl glycosides which are subsequently subjected to acid-promoted transacetalization:

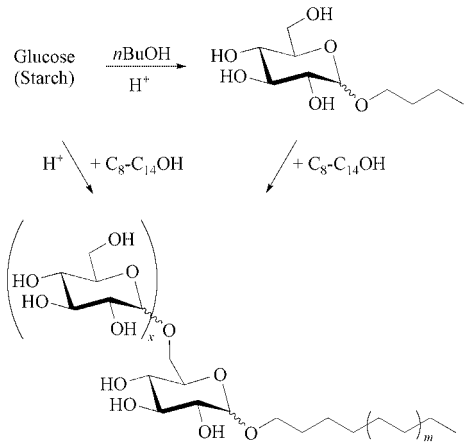

Alkyl Polyglucosides
(x = 0.3-0.7; m= 4-7)

APG's are not skin-irritating, have good foaming properties, and are readily biodegradable, hence are widely used in manual dishwash-

ing detergents and in formulations of shampoos, hair conditioners, and other personal care products.

9.2.5. Hydrophilic Monomers of Polyamides

Polyamide production worldwide amounted to about 5.8×10^6 t in 1998 [174]. More than 90 % of these polyamides are based on six-carbon monomers, i.e., caprolactam (nylon 6), and adipic acid/hexamethylenediamine (nylon 66).

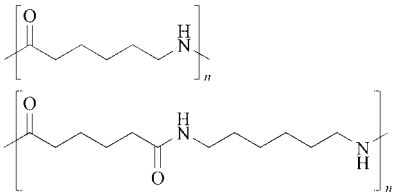

As six-carbon compounds in the form of hexoses are abundantly available in nature, substantial efforts have been made to derive monomers suitable for polyamidation from the bulk scale-accessible hexoses. This approach becomes particularly evident, when considering the large variety of amincarboxylic acids, dicarboxylic acids, and diamines (Table 6), that are accessible from the common six-carbon sugars [175], [176].

Of the myriad of possible combinations of these sugar-derived monomers either with themselves or with the common, petrochemically derived diamines and dicarboxylic acids, an immense number have been realized. In the following only a few of the respective polyamides are exemplarily covered.

Solution or interfacial polycondensation of galactaric acid dichloride in its acetylated form with various aliphatic and aromatic diamines resulted in a series of polyamides [175], [177], the one resulting from 1,6-diaminohexane resembling a nylon-6,6 in which half of the hydrogens of the methylene chain have been substituted by acetoxy groups (R = Ac). These can be deacylated with aqueous ammonia to give the tetra-hydroxylated nylon 66:

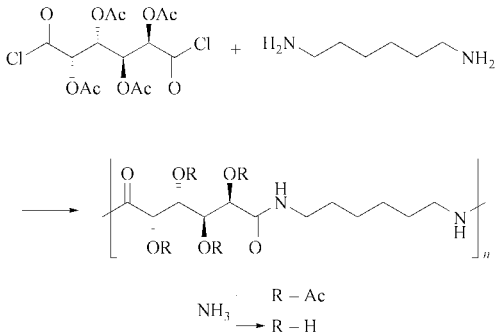

Table 6. Sugar-derived six-carbon building blocks suitable for polyamidation [175], [176]

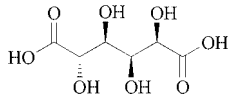	D-Glucaric acid	(Glucose)
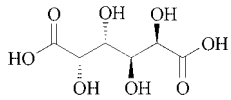	D-Galactaric acid	(Lactose)
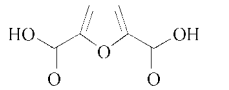	Furan-2,5-dicarboxylic acid	(Fructose)
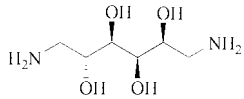	1,6-Diamino-1,6-dideoxy-D-glucitol	(Glucose)
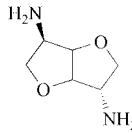	2,5-Diamino-1,4:3,6-dianhydro-sorbitol	(Glucose)
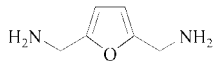	2,5-Bis(aminomethyl)furan	(Fructose)
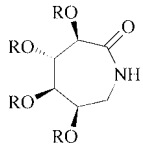	6-Amino-D-gluconic acid lactam (R = H, CH_3)	(Glucose)

In the case of D-glucaric acid, the use of its 3,6-lactone monomethyl ester proved advantageous to generate stereoregular polyglucaramides, effected with an impressive array of aliphatic and aromatic diamines [178]:

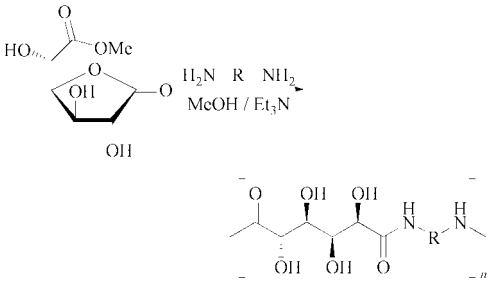

R – $-(CH_2)_x-$ with x – 2-12; $-CH_2C_6H_4CH_2-$;

$-(CH_2)_3NH(CH_2)_3-$; $-(CH_2)_3N(CH_2)_3-$; (CH_3)

$-(CH_2)_2O(CH_2)_2O(CH_2)_2-$; $-CH_2CH(CH_2)_2-$ (CH_3)

Sugar-based "quasi-aromatic" monomers for polyamides, i.e., the furan-2,5-dicarboxylic acid, appear particularly relevant as they embody the potential to replace petrochemically derived terephthalic or isophthalic acid in the present industrial products. Similar potential pertains to the furanic 1,6-diamine as a substitute for *p*-phenylenediamine. Indeed, a series of such furanic polyamides has been prepared [179] using the furan dicarboxylic acid and aliphatic as well as aromatic diamines. Of these, the polyamide resulting from condensation of furan-2,5-dicarboxylic acid with *p*-phenylenediamine, an analog of Nomex and Kevlar, has particularly promising decomposition and glass temperature parameters [180], distinctly better than those found for the all-furanic polyamide:

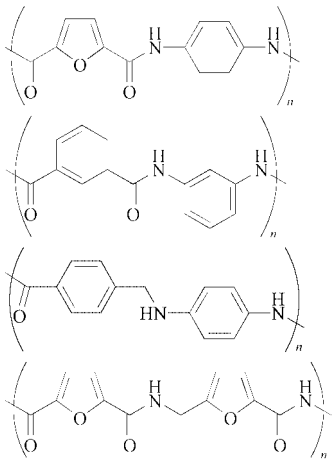

Nomex

Kevlar

In spite of the impressive array of highly useful products – with respect to their application profiles they compare on the market favorably with the well-known polyamides – none of these sugar-derived polyamides is, at present, produced on an industrial scale. The reasons are purely economic because the products derived from fossil raw materials are still cheaper by a factor of 5 on the average. Eventually though, with the end of cheap oil being prognosticated for 2040 at the latest [116], and the increasing pressure on the environment, this at present untoward situation for products from carbohydrate feedstocks will be changing.

10. References

1. J. Lehmann: *Kohlenhydrate, Chemie und Biologie,* Thieme, Stuttgart/New York, 1996, 372 pp.; *Carbohydrates, Chemistry and Biology,* Thieme, Stuttgart/New York, 1998, 274 pp.
2. T. K. Lindhorst: *Essentials of Carbohydrate Chemistry and Biochemistry,* Wiley-VCH, Weinheim, 2000, 217 pp.
3. R. V. Stick: *Carbohydrates: The Sweet Molecules of Life,* Acad. Press, San Diego, 2001, 256 pp.
4. P. M. Collins, R. J. Ferrier: *Monosaccharides, their Chemistry and their Roles in Natural Products,* Wiley, Chichester/New York, 1995, 574 pp.
5. J. F. Kennedy (ed.): *Carbohydrate Chemistry,* Clarendon Press, Oxford, 1988, 678 pp.
6. G.-J. Boons (ed.): *Carbohydrate Chemistry,* Blackie Academic, London/Weinheim, 1998, 508 pp.
7. J. F. Robyt: *Essentials of Carbohydrate Chemistry,* Springer, Heidelberg–New York, 1998, 399 pp.
8. B. Ernst, G. W. Hart, P. Sinay (eds.): *Carbohydrates in Chemistry and Biology,* **Vol. 1-4.** Wiley-VCH, Weinheim, 2000, 2340 pp.
9. B. Fraser-Reid, K. Tatsuta, J. Thiem (eds.): *Glycoscience, Chemistry and Chemical Biology,* **Vol. I-III,** Springer, Heidelberg, 2001, 2850 pp.
10. E. Fischer, "Ueber die Configuration des Traubenzuckers und seiner Isomeren, *Ber. Dtsch. Chem. Ges.* **24** (1891) 2683 – 2687.
11. F. W. Lichtenthaler, "Emil Fischer's Proof of the Configuration of Sugars: A Centennial Tribute," *Angew. Chem.* **104** (1992)

1577-1593; *Angew. Chem. Int. Ed. Engl.* **31** (1992) 1541 – 1556.
12. J. M. Bijvoet, A. F. Peerdemann, A. J. van Bommel, "Determination of the Absolute Configuration of Optically Active Compounds by Means of X-Rays. *Nature* **168** (1951) 271 – 272.
13. W. N. Haworth: *The Constitution of Sugars,* Arnold, London, 1929, 104 pp.
14. J. A. Mills, "The Stereochemistry of Cyclic Derivatives of Carbohydrates," *Adv. Carbohydr. Chem.* **10** (1955) 1 – 53.
15. IUPAC-IUB Joint Commission (JCBN), "Conformational Nomenclature for five- and six-membered Ring Forms of Monosaccharides and their Derivatives," *Pure Appl. Chem.* **53** (1981) 1901 – 1905; *Carbohydr. Res.* **297** (1997) 20 – 23.
16. J. A. Kanters, G. Roelofsen, B. P. Alblas, I. Meinders, "The Crystal Structure of β-D-Fructose with Emphasis on the Anomeric Effect," *Acta Cryst.* **B33** (1977) 665 – 672.
17. B. Schneider, F. W. Lichtenthaler, G. Steinle, H. Schiweck, "Distribution of Furanoid and Pyranoid Tautomers of D-Fructose in Dimethylsulfoxide, Water and Pyridine via Anomeric Hydroxyl Proton NMR Intensitites," *Liebigs Ann. Chem.* (1985) 2454 – 2464.
18. G. M. Brown, H. A. Levy, "Sucrose: Precise Determination of Crystal and Molecular Structure by Neutron Diffraction," *Science* **141** (1963) 921 – 923; *Acta Cryst.* **B29** (1973) 790 – 797.
19. J. C. Hanson, L. C. Sieker, L. H. Jensen, "Sucrose: X-Ray Refinement and Comparison with Neutron Refinement," *Acta Crystallogr.* **B29** (1973) 797 – 808.
20. R. R. Watson, N. S. Orenstein, "Chemistry and Biochemistry of Apiose," *Adv. Carbohydr. Chem. Biochem.* **31** (1975) 135 – 184.
21. J. Yoshimura, "Synthesis of Branched-chain Sugars," *Adv. Carbohydr. Chem. Biochem.* **42** (1984) 69 – 134.
22. A. Lipták, P. Fügedi, Z. Szurmai, J. Harangi: Handbook of Oligosaccharides, Vol. I (Disaccharides, 474 pp.), Vol. II (Trisaccharides, 267 pp.), Vol. III (Higher Oligosaccharides, 179 pp.), CRC Press, Boca Raton/Boston, 1990.
23. A. Hugill: *Introductory Dedicational Metaphor to Sugar and All That. A History of Tate & Lyle,* Gentry Books, London, 1978.
24. International Sugar Organization, "World Sugar Production 2000/2001," *Zuckerind. (Berlin)* **126** (2001) 300, 577.
25. S. Immel, F. W. Lichtenthaler, "The Conformation of Sucrose in Water: A Molecular Dynamics Approach," *Liebigs Ann. Chem.* (1995) 1925 – 1937.
26. H. Schiweck, M. Munir, K. M. Rapp, B. Schneider, M. Vogel, "Use of Sucrose as an Industrial Bulk Chemical and Production of Isomaltulose and Isomalt," *Zuckerind. (Berlin)* **115** (1990) 555 – 565.
27. F. W. Lichtenthaler, P. Pokinskyj, S. Immel, "Sucrose as a Renewable Raw Material. New Reactions via Computer Simulation of Hydroxyl Group Reactivities," *Zuckerind. (Berlin)* **121** (1996) 174 – 190.
28. A. Varki, "Biological Roles of Oligosaccharides," *Glycobiology* **3** (1993) 97 – 130.
29. H. M. I. Osborn, T. H. Khan: *Oligosaccharides. Their Synthesis and Biological Roles,* Oxford University Press, Oxford, 2000, 112 pp.
30. H.-J. Gabius, S. Gabius (eds.): *Glycosciences–Status and Perspectives,* Chapman & Hall, London/Weinheim, 1997, 631 pp.
31. J. Kopitz: "Glycolipids: Structure and Function," in [30], pp 163 – 189.
32. N. Sharon, H. Lis: "Glycoproteins: Structure and Function," in [30], pp 133 – 162.
33. J. Szejtli, "Introduction and General Overview of Cyclodextrin Chemistry," *Chem. Rev.* **98** (1998) 1743 – 1753.
34. F. W. Lichtenthaler, S. Immel, "Towards Understanding Formation and Stability of Cyclodextrin Inclusion Complexes. Visualization of their Lipophilicity Patterns," *Starch/Stärke* **48** (1996) 145 – 154.
35. All 3D-structures can be viewed on the www at: http://caramel.oc.chemie.tu-darmstadt.de/immel/3Dstructures.html; MOLCAD graphics are available at: http://caramel.oc.chemie.tu-darmstadt.de/immel/molcad/gallery.html.
36. J. Szejtli, "Medical Applications of Cyclodextrins," *Med. Res. Rev.* **14** (1994) 353 – 386.
37. J. K. Uekama, F. Hitayama, T. Ivie, "Cyclodextrin Drug Carrier Systems," *Chem. Rev.* **98** (1998) 2045 – 2076.
38. V. Schurig, H. P. Novotny, "Gaschromatographic Separation of Enantiomers on Cyclodextrin Derivatives," *Angew. Chem.* **102** (1990) 969 – 986; *Angew. Chem. Int. Ed. Engl.* **29** (1990) 939 – 958.

39. S. Li, W. C. Purdy, "Cyclodextrins and their Applications in Analytical Chemistry," *Chem. Rev.* **92** (1992) 1457 – 1470.
40. G. Wenz, "Cyclodextrins as Building Blocks for Supramolecular Structures and Functional Units," *Angew. Chem.* **106** (1994) 851 – 870; *Angew. Chem. Int. Ed. Engl.* **33** (1994) 803 – 822.
41. I. Tabushi, "Design Synthesis of Artificial Enzymes," *Tetrahedron* **40** (1984) 269 – 292.
42. T. L. Bluhm, P. Zugenmaier, "Detailed Structure of the V_H-amylose-Iodine Complex: A linear polyiodide chain," *Carbohydr. Res.* **89** (1981) 1 – 10.
43. S. Immel, F. W. Lichtenthaler, "The Hydrophobic Topographies of Amylose and its Blue Iodine Complex," *Starch/Stärke* **52** (2000) 1 – 8.
44. A. Fuchs (ed.): *Inulin and Inulin-containing Crops,* Elsevier, Amsterdam/New York, 1993, 417 pp.
45. A. Fuchs, "Potentials for Non-food Utilization of Fructose and Inulin," *Starch/Stärke* **39** (1987) 335 – 343.
46. J. F. Kennedy, C. A. White: "The Plant, Algal and Microbial Polysaccharides," in J. F. Kennedy (ed.): *Carbohydrate Chemistry* Oxford Science Publ., Oxford 1988, pp 220 – 262.
47. R. J. Sturgeon: "The Glycoproteins and Glycogen," in J. F. Kennedy (ed.): *Carbohydrate Chemistry* Oxford Science Publ., Oxford, 1988, pp 263 – 302.
48. IUPAC-IUB Joint Commission, "Nomenclature of Carbohydrates (1996 Recommendations)," *Pure Appl. Chem.* **68** (1996) 1919 – 2008, or *Carbohydr. Res.* **297** (1997) 1 – 92, or *Adv. Carbohydr. Chem. Biochem.* **52** (1997) 47 – 177.
49. J. Szejtli: *Säurehydrolyse glycosidischer Bindungen. Einfluß von Struktur und Reaktionsbedingungen auf die Säurespaltung von Glykosiden, Disacchariden, Oligo- und Polysacchariden,* VEB Fachbuchverlag, Leipzig 1976, 399 pp.
50. C. J. Biermann, "Hydrolysis and Other Cleavages of Glycosidic Linkages in Polysaccharides," *Adv. Carbohydr. Chem. Biochem.* **46** (1988) 251 – 271.
51. O. Theander, D. A. Nelson, "Aqueous, High-Temperature Transformation of Carbohydrates towards Biomass Utilization," *Adv. Carbohydr. Chem. Biochem.* **46** (1988) 273 – 332.
52. W. J. McKillip, "Furan and Derivatives," *Kirk-Othmer Encyclopedia Chemical Technology,* **11** (1981) 501 – 527.
53. B. F. M. Kuster, "Manufacture of 5-Hydroxymethylfurfural," *Starch/Stärke* **42** (1990) 314 – 312.
54. D. Martin, T. Weber, H. Schiweck, "5–(α-D-Glucosyloxymethyl)-furfural: Preparation from Isomaltulose and Exploration of its Ensuing Chemistry", *Liebigs Ann. Chem.* (1993) 967 – 974.
55. S. J. Angyal, "The Lobry de Bruyn-Alberda van Ekenstein Transformation and Related Reactions," *Topics Current Chem.* **215** (2001) 1 – 14.
56. JP 73-49938, 1971 (I. Machida, J. Kanaeda, H. Miki, S. Kubomura, H. Toda, T. Shirioshi). *Chem. Abstr.* **79** (1973) 126744v.
57. F. W. Lichtenthaler, S. Rönninger, "α-D-Glucopyranosyl-D-fructoses. Distribution of Furanoid and Pyranoid Tautomers in Water, DMSO and Pyridine," *J. Chem. Soc. Perkin Trans. 2* (1990) 1489 – 1497.
58. L. Petruš, M. Petrušová, Z. Hricovíniová, "The Bílik Reaction," *Topics Current Chem.* **215** (2001) 16 – 41.
59. F. Ledl, E. Schleicher, "Die Maillard-Reaktion in Lebensmitteln und im menschlichen Körper," *Angew. Chem.* **102** (1990) 597 – 626; *Angew. Chem. Int. Ed. Engl.* **29** (1990) 565 – 594.
60. A. F. Bochkov, G. E. Zaikov: *The Chemistry of the O-Glycosidic Bond,* Pergamon Press, Oxford, 1979, 210 pp.
61. T. Ziegler: "Protecting Group Strategies for Carbohydrates," in G.-J. Boons (ed.): *Carbohydrate Chemistry,* Blackie Academic, London/Weinheim, 1998, pp 21 – 45.
62. M. Nitz, D. R. Bundle: "Glycosyl Halides in Oligosaccharide Synthesis," in B. Fraser-Reid, K. Tatsuta, J. Thiem (eds.): *Glycoscience* **Vol. II,** 2002, pp 1497 – 1544.
63. F. W. Lichtenthaler, E. Kaji, T. Schneider-Adams, "2-Oxo- and 2-Oximino-glycosyl Halides. Versatile Glycosyl Donors for the Construction of β-D-Mannose- and β-D-Mannosamine-containing Oligosaccharides," *Trends Glycosci. Glycotechnol.* **5** (1993) 121 – 142; *J. Org. Chem.* **59** (1994) 6728-6734 and 6735 – 6738; *Carbohydr. Res.* **305** (1998) 293 – 303.
64. M. Nitz, D. R. Bundle, "Synthesis of Di- to Hexasaccharide 1,2-Linked β-Mannopyranan

Oligomers and a Terminal S-Linked Tetrasaccharide Congener," *J. Org. Chem.* **66** (2001) 8411 – 8423.
65. E. R. Palmacci, O. J. Plante, P. H. Seeberger, "Oligosaccharide Synthesis in Solution and on Solid Support with Glycosyl Phosphates," *Eur. J. Org. Chem.* (2002) 595 – 606.
66. R. R. Schmidt, W. Kinzy, "Anomeric-oxygen Activation for Glycoside Synthesis: The Trichloroacetimidate Method," *Adv. Carbohydr. Chem. Biochem.* **50** (1994) 21 – 124.
67. P. J. Garegg, "Thioglycosides as Glycosyl Donors in Oligosaccharide Synthesis," *Adv. Carbohydr. Chem. Biochem.* **52** (1997) 179 – 205.
68. L. Yan, D. Kahne, "Generalizing Glycosylation. Synthesis of Blood Group Antigens Le^a, Le^b, and Le^x," *J. Am. Chem. Soc.* **118** (1996) 9239 – 9248.
69. S. J. Danishefsky, M. T. Bilodean, "Glycals in Organic Synthesis. Comprehensive Strategies for the Assembly of Oligosaccharides," *Angew. Chem.* **108** (1996) 1482 – 1522; *Angew. Chem. Int. Ed. Engl.* **35** (1996) 1380 – 1419.
70. K. Fucase: "Combinatorial and Solid Phase Methods in Oligosaccharide Syntheses," in B. Fraser-Reid, K. Tatsuta, J. Thiem (eds.): *Glycoscience,* **Vol. II** Springer, Heidelberg – New York, 2001, pp 1661 – 1694.
71. G.-J. Boons: "Strategies and Tactics in Oligosaccharide Synthesis," in G.-J. Boons (ed.): *Carbohydrate Chemistry* Blackie Academic, London/Weinheim, 1998, pp. 175 – 222.
72. Ref. [8], Vol. 1, pp 1 – 583.
73. G. Magnusson, U. J. Nilson: "Regio- and Stereoselective Methods of Glycosylation," in B. Fraser-Reid, K. Tatsuta, J. Thiem (eds.): *Glycoscience,* **Vol. II** Springer, Heidelberg – New York, 2001, pp 1543 – 1588.
74. J. D. Wander, D. Horton, "Dithioacetals of Sugars," *Adv. Carbohydr. Chem. Biochem.* **32** (1976) 15 – 123.
75. T. J. Wrodnigg, B. Eder, "The Amadori and Heyns Rearrangements," *Topics Curr. Chem.* **215** (2001) 115 – 152.
76. H. S. El Khadem, A. F. Fatiadi, "Hydrazine Derivatives of Carbohydrates and Related Compounds," *Adv. Carbohydr. Chem. Biochem.* **55** (1999) 175 – 263.
77. C. S. Hudson, "The Kiliani-Fischer Cyanohydrin Synthesis," *Adv. Carbohydr. Chem.* **1** (1945) 1 – 36.
78. J. C. Sowden, "The Nitromethane and 2-Nitroethanol Syntheses," *Adv. Carbohydr. Chem.* **6** (1951) 291 – 318.
79. F. L. Humoller, "β-D-Allose," *Methods Carbohydr. Chem.* **1** (1962) 102 – 104.
80. J. C. Sowden, "L-Glucose and L-Mannose," *Methods Carbohydr. Chem.* **1** (1962) 132 – 135.
81. F. W. Lichtenthaler, "Cyclizations of Dialdehydes with Nitromethane," *Angew. Chem.* **76** (1964) 84-97; *Angew. Chem. Int. Ed. Engl.* **3** (1964) 135 – 148.
82. H. H. Baer, "The Nitro Sugars," *Adv. Carbohydr. Chem. Biochem.* **24** (1969) 67 – 138.
83. F. W. Lichtenthaler, "Amino Sugars and Amino Cyclitols via Cyclization of Dialdehydes with Nitromethane," *Methods Carbohydr. Chem.* **6** (1972) 250 – 260.
84. F. W. Lichtenthaler, "Branched-chain Amino Sugars via Dialdehyde-Nitroalkane-Cyclization," *Fortschr. Chem. Forschg.* **14** (1970) 556 – 577.
85. D. L. MacDonald, H. O. L. Fischer, "The Degradation of Sugars by Means of their Disulfones," *J. Am. Chem. Soc.* **74** (1952) 2087 – 2090.
86. P. J. Sicard: "Hydrogenated Glucose Syrups, Sorbitol, Mannitol and Xylitol," in G. G. Birch, K. J. Parker (eds.): *Nutritive Sweeteners* Applied Science Publ., London/New Jersey, 1982, pp 145 – 170.
87. C. K. Lee, "The Chemistry and Biochemistry of the Sweetness of Sugars," *Adv. Carbohydr. Chem. Biochem.* **45** (1990) 199 – 351.
88. P. J. Sicard, P. Leroy: "Mannitol, Sorbitol and Lycasin; Properties and Food Applications," in T. H. Grenby, K. J. Parker, M. G. Lindley, (eds.): *Developments in Sweeteners-2* Appl. Science Publ., London – New York, 1983, pp 1 – 26.
89. H. J. Lindner, F. W. Lichtenthaler, "Extended Zig-zag Conformation of 1-O-α-D-Glucopyranosyl-D-mannitol," *Carbohydr. Res.* **93** (1981) 135 – 140.
90. C. L. Mehltretter, "D-Glucaric Acid," *Methods Carbohydr. Chem.* **2** (1963) 46 – 48.
91. H. E. Conrad, "Methylation of Carbohydrates with Methylsulfinyl Anion and Methyl Iodide in DMSO," *Methods Carbohydr. Chem.* **6** (1972) 361 – 364.
92. C. M. McCloskey, "Benzyl Ethers of Sugars," *Adv. Carbohydr. Chem.* **12** (1957) 137 – 156.
93. C. C. Sweely, R. Bentley, M. Makita, W. W. Wells, "Gas-Liquid-Chromatography of

Trimethylsilyl Derivatives of Sugars," *J. Am. Chem. Soc.* **85** (1963) 2497–2507.
94. C. E. Ballou, D. L. MacDonald, "Phosporylation with Diphenyl Phosphorochloridate," *Methods Carbohydr. Chem.* **2** (1963) 270 – 272; 273 – 276; 277 – 281; 282 – 288.
95. C.-H. Wong, G. M. Whitesides: *Enzymes in Synthetic Organic Chemistry,* Elsevier Science, Oxford, 1994.
96. R. L. Whistler, W. W. Spencer, J. N. BeMiller, "Sulfates," *Methods Carbohydr. Chem.* **2** (1963) 298 – 303.
97. J. Honeyman, J. W. W. Morgan, "Sugar Nitrates," *Adv. Carbohydr. Chem.* **12** (1957) 117 – 135.
98. F. W. Lichtenthaler, H. J. Müller, "Preparation of Nucleoside 5'-Nitrates," *Synthesis* (1974) 199 – 201.
99. L. A. Silvieri, N. J. DeAngelis: In K. Florey (ed.): *Profiles of Drug Substances,* **Vol. 4,** Acad. Press, New York, pp 225 – 244.
100. K. Igarashi, "The Koenigs-Knorr Reaction," *Adv. Carbohydr. Chem. Biochem.* **34** (1977) 243 – 283.
101. S. Hanessian: *Total Synthesis of Natural Products: The Chiron Approach,* Pergamon, Oxford/New York, 1983.
102. F. W. Lichtenthaler: "Building Blocks from Sugars and their Use in Natural Product Synthesis. A Review with Procedures." in R. Scheffold (ed.): *Modern Synthetic Methods,* **Vol. 6,** VCH Verlagsgesellschaft, Weinheim, 1992, pp 273 – 376.
103. G. J. Boons, K. J. Hale: *Organic Synthesis with Carbohydrates,* Sheffield Acad. Press, Sheffield, 2000, 336 pp.
104. E. Fischer, "Ueber neue Reduktionsprodukte des Traubenzuckers: Glucal und Hydroglucal," *Ber. Dtsch. Chem. Ges.* **47** (1914) 196 – 210; *Methods Carbohydr. Chem.* **2** (1963) 405 – 408.
105. L. Somsák, I. Nemeth, "A Simple Method for the Synthesis of Acylated Pyranoid Glycals under Aprotic Conditions," *J. Carbohydr. Chem.* **12** (1993) 679 – 684.
106. M. G. Blair, "The 2-Hydroxyglycals," *Adv. Carbohydr. Chem.* **9** (1954) 97 – 130; *Methods Carbohydr. Chem.* **2** (1963) 411 – 414.
107. R. J. Ferrier, "Modified Synthesis of (2-Hydroxylglycal) Esters," *Methods Carbohydr. Chem.* **6** (1972) 307 – 311.
108. N. L. Holder, "The Chemistry of Hexuloses," *Chem. Rev.* **82** (1982) 287 – 33.
109. F. W. Lichtenthaler, "Sugar Enolones–Synthesis, Reactions of Preparative Interest, and γ-Pyrone Formation," *Pure Appl. Chem.* **50** (1978) 1343 – 1362.
110. F. W. Lichtenthaler, U. Kläres, M. Lergenmüller, S. Schwidetzky, "Glycosyl Donors with a Keto or Oximino Function next to the Anomeric Centre: Practical Preparation and Evaluation of their Selectivites in Glycosidation," *Synthesis* (1992) 179 – 184.
111. F. W. Lichtenthaler, T. Schneider-Adams, "3,4,6-Tri-O-benzyl-α-D-arabino-hexopyranos-2-ulosyl Bromide–a Glycosyl Donor for the Efficient Generation of β-D-Mannosidic Linkages," *J. Org. Chem.* **59** (1994) 6728 – 6734.
112. F. W. Lichtenthaler, T. Schneider-Adams, S. Immel, "A Practical Synthesis of β-D-Xyl-(1$\rightarrow$2)-β-D-Man-(1$\rightarrow$4)-D-GlcOMe, a Trisaccharide Component of the Glycosphingolipid of Hyriopsis schlegelii," *J. Org. Chem.* **59** (1994) 6735 – 6738.
113. R. F. Brady, "Cyclic Acetals of Ketoses," *Adv. Carbohydr. Chem. Biochem.* **26** (1971) 197 – 278.
114. A. N. deBelder, "Cyclic Acetals of the Aldoses and Aldosides," *Adv. Carbohydr. Chem. Biochem.* **34** (1977) 179 – 241.
115. D. M. Clode, "Carbohydrate Cyclic Acetal Formation and Migration," *Chem. Rev.* **79** (1979) 491 – 513.
116. C. J. Campbell, J. H. Laherrère, "The End of Cheap Oil," *Sci. Am.*1998, March, 60 – 65.
117. D. H. Klass: "Fossil Fuel Reserves and Depletion," *Biomass for Renewable Energy, Fuels and Chemicals,* Acad. Press, San Diego, 1998, pp 10 – 19.
118. C. Okkerse, H. van Bekkum, "Fossil to Green," *Green Chem.* (1999) 107 – 114.
119. National Research Council, USA. Biobased Industrial Products: Priorities for Research and Commercialization, Natl. Acad. Press, Washington, 2000.
120. D. L. Klass: "Organic Commodity Chemicals from Biomass," in *Biomass for Renewable Energy, Fuels and Chemicals,* Acad. Press, San Diego, 1998, pp 495 – 546.
121. F. W. Lichtenthaler: *Carbohydrates as Organic Raw Materials,* VCH Verlagsgesellschaft, Weinheim, 1991, 350 pp.
122. G. Descotes: *Carbohydrates as Organic Raw Materials II,* VCH Verlagsgesellschaft, Weinheim, 1993, 270 pp.
123. M. Eggersdorfer, S. Warwel, G. Wulff (eds.): *Nachwachsende Rohstoffe: Perspektiven für*

die Chemie, VCH Verlagsgesellschaft, Weinheim, 1993, 388 pp.
124. H. van Bekkum, H. Röper, A. Voragen: *Carbohydrates as Organic Raw Materials III,* VCH Verlagsgesellschaft, Weinheim, 1996, 306 pp.
125. H. Eierdanz (ed.): *Perspektiven nachwachsender Rohstoffe in der Chemie,* VCH Verlagsgesellschaft, Weinheim, 1996.
126. F. W. Lichtenthaler, S. Mondel, "Perspectives in the Use of Low-molecular-weight Carbohydrates as Organic Raw Materials," *Pure Appl. Chem.* **69** (1997) 1853 – 1866.
127. F. W. Lichtenthaler, "Towards Improving the Utility of Ketoses as Organic Raw Materials," *Carbohydr. Res.* **313** (1998) 69 – 89.
128. S. David, C. Augé, C. Gautheron, "Enzymic Methods in Preparative Carbohydrate Chemistry," *Adv. Carbohydr. Chem. Biochem.* **49** (1991) 176 – 238.
129. A. R. Katritzky, C. W. Rees, (eds.): *Comprehensive Heterocyclic Chemistry,* **Vol. 4,** Pergamon, Oxford, 1984, pp 599 – 712.
130. A. Gandini, M. N. Belgacem, "Furans in Polymer Chemistry," *Prog. Polym. Sci.* **22** (1997) 1203 – 1279, and references therein.
131. T. El-Hajj, A. Masroua, J.-C. Martin, G. Descotes, "5-Hydroxymethylfurfural and Derivatives," *Bull. Soc. Chim. Fr.* (1987) 855 – 860.
132. C. Larousse, L. Rigal, A. Gaset, "Thermal Degradation of 5-Hydroxymethylfurfural," *J. Chem. Technol. Biotechnol.* **53** (1992) 111 – 116.
133. V. Schiavo, G. Descotes, J. Mentech, "Hydrogenation of 5-Hydroxymethylfurfural," *Bull. Soc. Chim. Fr.* **128** (1991) 704 – 711.
134. N. Elming, N. Clauson-Kaas, "6-Methyl-3-pyridinol from 2-Hydroxymethyl-5-aminomethyl-furan," *Acta Chem. Scand.* **10** (1956) 1603 – 1605.
135. DE 3.826.073, 988 (E. Leupold, M. Wiesner, M. Schlingmann, K. Rapp); *Chem. Abstr.* **113** (1990) 23678t.
136. A. Gandini, M. N. Belgacem, "Furanic Polyesters," *Progr. Polym. Sci.* **22** (1997) 1232 – 1238.
137. M. Kunz: "5-Hydroxymethylfurfural, a Possible Basic Chemical for Industrial Intermediates," in A. Fuchs (ed.): *Inulin and Inulin-containing Crops* Elsevier, Amsterdam, 1993, pp 149 – 160.
138. US 4 497 961 1980 (J. W. Clitherow); *Chem. Abstr.* **102** (1985) 166604z.
139. F. García-Gonzáles, "Reactions of Monosaccharides with β-Ketoesters," *Adv. Carbohydr. Chem.* **11** (1956) 97 – 143.
140. A. J. Moreno-Vargas, J. Fuentes, J. Fernández-Bolanos, I. Robina, "Synthesis of Hetarylene-carbopeptoids," *Tetrahedron Lett.* **42** (2001) 1283 – 1285.
141. A. Beélik, "Kojic Acid," *Adv. Carbohydr. Chem.* **11** (1956) 145 – 183.
142. C. Nelson, J. Gratzl, "Conversion of D-Glucono-1,5-lactone into an α-Pyrone," *Carbohydr. Res.* **60** (1978) 267 – 273.
143. K. Tajima, "Cyclopentenones from 3-Acetoxy-6-acetoxymethyl-2-pyrone," *Chem. Lett.* (1987) 1319 – 1322.
144. M. Bols: *Carbohydrate Building Blocks,* Wiley, New York, 1996, 182 pp.
145. F. W. Lichtenthaler, S. Rönninger, P. Jarglis, "Expedient Approach to Pyranoid Ene and Enol Lactones," *Liebigs Ann.* (1989) 1153 – 1161.
146. F. W. Lichtenthaler, U. Kraska, "Preparation of Sugar Enolones," *Carbohydr. Res.* **58** (1977) 363 – 377.
147. F. W. Lichtenthaler, S. Nishiyama, T. Weimer, "2,3-Dihydropyranones with Contiguous Chiral Centers," *Liebigs Ann. Chem.* (1989) 1163 – 1170.
148. F. W. Lichtenthaler, S. Ogawa, P. Heidel, "Synthesis of Unsaturated Hexopyranosid-4-uloses," *Chem. Ber.* **110** (1977) 3324 – 3332.
149. Z. J. Witczak, "Synthesis of Natural Products from Levoglucosenone," *Pure Appl. Chem.* **66** (1994) 2189 – 2192.
150. F. Shafizadeh, R. H. Furneaux, T. T. Stevenson, "Reactions of Levoglucosenone," *Carbohydr. Res.* **71** (1979) 169 – 191.
151. H. El Khadem, "N-Heterocycles from Saccharide Derivatives," *Adv. Carbohydr. Chem.* **25** (1970) 351 – 405.
152. B. A. Lewis, F. Smith, A. M. Stephen, "Galactaric Acid," *Methods Carbohydr. Chem.* **2** (1963) 38 – 46.
153. S. M. McElvain, K. M. Bolliger, "Pyrrole," *Org. Synth.,* **Coll. Vol. 1,** 1941, 473 – 474.
154. F. W. Lichtenthaler, A. Brust, E. Cuny, "Hydrophilic Pyrroles, Pyridazines and Diazepinones from D-Fructose and Isomaltulose," *Green Chemistry* **3** (2001) 201 – 209.
155. F. García-Gonzáles, A. Gomes Sanchez, "Reactions of Amino Sugars with β-Dicarbonyl Compounds," *Adv. Carbohydr. Chem.* **20** (1965) 303 – 355.

156. A. Rozanski, K. Bielawski, J. Boltryk, D. Bartulewicz, "Simple Synthesis of 3-Acetyl-5-(tetrahydroxybutyl)-2-methylpyrrole," *Akad. Med. Juliana Marchlewskiego Bialymstoku* (1991) 35 – 36, 57 – 63; *Chem. Abstr.* **118** (1992) 22471m.
157. V. Diehl, E. Cuny, F. W. Lichtenthaler, "Conversion of D-Xylose into Hydrophilically Functionalized Pyrazoles," *Heterocycles* **48** (1998) 1193 – 1201.
158. M. Oikawa, C. Müller, M. Kunz, F. W. Lichtenthaler, "Hydrophilic Pyrazoles from Sugars," *Carbohydr. Res.* **309** (1998) 269 – 279.
159. R. Weidenhagen, R. Hermann, "4-Hydroxymethylimidazole," *Ber. Dtsch. Chem. Ges.* **70** (1937) 570 – 583; *Org. Synth.,* **Coll. Vol. III,** (1955) 460 – 462.
160. J. Streith, A. Boiron, A. Frankowski, D. Le Nouen, H. Rudyk, T. Tschamber, "One-Pot Synthesis of Imidazolosugars," *Synthesis* (1995) 944 – 946.
161. S. Rapp, F. W. Lichtenthaler: unpublished.
162. GB 862 581, 1961 (N. Elming, S. V. Carlsten, B. Lennart, I. Ohlsson); *Chem. Abstr.* **56** (1962) 11574g.
163. EP. 227045, EP 227046, 1987(V. Koch, L. Willms, A. Fuß, K. Bauer, K. H. Bieringer, H. Buerstell); *Chem. Abstr.* **107** (1987) 175892, 134217.
164. C. Müller, V. Diehl, F. W. Lichtenthaler, "3-Pyridinols from Fructose and Isomaltulose," *Tetrahedron* **54** (1998) 10703 – 10712.
165. H. Ohle, M. Hielscher, "Darstellung von Tetraoxybutyl-chinoxalin," *Ber. Dtsch. Chem. Ges.* **74** (1941) 13 – 19.
166. H. Ohle, G. A. Melkonian, "Flavazol, ein neuer Heterocyclus aus Zuckern," *Ber. Dtsch. Chem. Ges.* **74** (1941) 279 – 285.
167. H. Ohle, A. Iltgen, "Flavazole," *Ber. Dtsch. Chem. Ges.* **76** (1943) 1 – 14.
168. N. J. Krog in K. Larsson, S. E. Friberg, (eds.): *Food Emulsions* Marcel Dekker, New York/Basel, 1990.
169. Procter & Gamble Co., US 3 600 168, 1968 (F. H. Mattson, R. A. Volpenhein); *Chem. Abstr.* **75** (1971) 139614v.
170. "Fat-free Fat," (Cover Story) *Time Magazine* (1996) Jan. 8.
171. H. Kelkenberg, "Sugar-based Detergents. New Components for Cleaning Compositions and Cosmetics," *Tensides, Surfactants, Detergents* **25** (1988) 8 – 13.
172. P. Jürges, A. Turowski: "Vergleichende Untersuchung von Zuckerestern, N-Methylglucamiden und Glycosiden am Beispiel von Reinigungsprodukten," in H. Eierdanz (ed.): *Perspektiven nachwachsender Rohstoffe in der Chemie* VCH Publ., Weinheim, 1996, pp 61 – 70.
173. W. von Rybinski, K. Hill, "Alkyl Polyglycosides–Properties and Applications of a New Class of Surfactants," *Angew. Chem.* **110** (1998) 1394 – 1412; *Angew. Chem. Int. Ed.* **37** (1998) 1328 – 1345.
174. H.-P. Weiß, W. Sauerer, "Polyamides," *Kunststoffe* **89** (1999) 68 – 74.
175. J. Thiem, F. Bachmann, "Carbohydrate-derived Polyamides," *Trends Polym. Sci.* **2** (1994) 425 – 432.
176. O. Varela, H. A. Orgueira, "Synthesis of Chiral Polyamides from Carbohydrate-derived Monomers," *Adv. Carbohydr. Chem. Biochem.* **55** (1999) 137 – 174.
177. E. M. E. Mansur, S. H. Kandil, H. H. A. M. Hassan, M. A. E. Shaban, "Synthesis of Carbohydrate-containing Polyamides and Study of their Properties," *Eur. Polym. J.* **26** (1990) 267 – 276.
178. L. Chen, D. E. Kiely, "Synthesis of Stereoregular Head/Tail Hydroxylated Nylons Derived from D-Glucose," *J. Org. Chem.* **61** (1996) 5847 – 5851; US 5 329 044, 1994; *Chem. Abstr.* **122** (1994) 56785.
179. A. Gandini, M. N. Belgacem, "Furanic Polyamides Chemistry," *Progr. Polym. Sci.* **22** (1997) 1238 – 1246.
180. A. Mitiakoudis, A. Gandini, "Poly(p-phenylene)-2,5-furandicarbonamide and Conversion into Filaments and Films," *Macromolecules* **24** (1991) 830 – 835.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Carbon

Standard Article

Otto Vohler[1], Gabriele Nutsch[2], Gerd Collin[3], Ferdinand von Sturm[4], Erhard Wege[5], Wilhelm Frohs[6], Klaus-Dirk Henning[7], Hartmut von Kienle[8], Manfred Voll[9], Peter Kleinschmit[10], Otto Vostrowsky[11], Andreas Hirsch[12]

[1]Sigri GmbH, Meitingen, Federal Republic of Germany
[2]Technische Universität Ilmenau, Ilmenau, Federal Republic of Germany
[3]DECHEMA e.V., Frankfurt am Main, Federal Republic of Germany
[4]Sigri GmbH, Meitingen, Federal Republic of Germany
[5]Sigri GmbH, Meitingen, Federal Republic of Germany
[6]SGL Carbon GmbH, Meitingen, Federal Republic of Germany
[7]CarboTech Aktivkohlen GmbH, Essen, Federal Republic of Germany
[8](formerly Degussa AG, Hanau), Heusenstamm, Federal Republic of Germany
[9]Degussa AG, Werk Kalscheuren, Hürth, Federal Republic of Germany
[10]Degussa AG, Zweigniederlassung Wolfgang, Hanau, Federal Republic of Germany
[11]Institut für Organische Chemie der Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Federal Republic of Germany
[12]Institut für Organische Chemie der Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a05_095
Article Online Posting Date: January 15, 2002

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Diamond
2.1. General Sources
2.2. Properties and Structure of Natural Diamond
2.3. Diamond Synthesis
2.3.1. HPHT Synthesis
2.3.2. The Diamond Chemical Vapor Deposition (CVD) Process
2.3.3. Diamond CVD Techniques
2.3.3.1. Nonthermal Plasma CVD
2.3.3.2. Thermal Plasma Diamond CVD
2.3.3.3. Other CVD Techniques
3. Graphite
3.1. Single Crystal
3.2. Natural Graphite
3.3. Artificial Graphite
4. Carbon and Graphite Materials
4.1. Carbon and Graphite Artifacts
4.1.1. Manufacture
4.1.2. Properties
4.1.3. Applications
4.1.4. Economic Aspects
4.2. Other Classes of Carbon
4.2.1. Glassy Carbon

4.2.2. Pyrocarbon and Pyrographite
4.3. Graphite Compounds
4.3.1. Surface Complexes
4.3.2. Graphite Intercalation Compounds
4.4. Expanded Graphite and Graphite Foil Derived from Intercalation Compounds
5. Activated Carbon
5.1. General Aspects
5.2. Carbonaceous Adsorbents
5.2.1. Types of Carbonaceous Adsorbents
5.2.2. Chemical Properties
5.2.3. Mechanical Properties
5.2.4. Adsorption Properties
5.2.5. Quality Control
5.3. Production
5.3.1. General Aspects
5.3.2. Raw Materials
5.3.3. Activating Furnaces
5.3.4. Methods of Activation
5.3.4.1. Chemical Activation
5.3.4.2. Gas Activation
5.3.5. Granular and Pelletized Carbons
5.3.6. Carbon Molecular Sieves
5.3.7. Further Treatment
5.3.8. Impregnation
5.4. Applications
5.4.1. Gas-Phase Applications
5.4.1.1. Solvent Recovery
5.4.1.2. Process-Gas and Air Purification
5.4.1.3. Gas Separation
5.4.1.4. Gasoline Vapor Adsorption
5.4.1.5. Flue Gas Cleaning
5.4.2. Liquid-Phase Applications
5.4.2.1. Water Treatment
5.4.2.2. Micellaneous Liquid-Phase Applications
5.4.3. Impregnated Activated Carbon
5.4.4. Catalysts and Catalyst Supports
5.5. Regeneration and Reactivation
5.6. Economic Aspects
6. Carbon Black
6.1. Physical Properties
6.2. Chemical Properties
6.3. Raw Materials
6.4. Production Processes
6.4.1. Furnace Black Process
6.4.2. Gas Black and Channel Black Processes
6.4.3. Lamp Black Process
6.4.4. Thermal Black Process
6.4.5. Acetylene Black Process
6.4.6. Other Manufacturing Processes
6.4.7. Oxidative Aftertreatment of Carbon Black
6.4.8. Environmental Problems
6.5. Testing and Analysis
6.6. Storage and Transportation
6.7. Uses
6.8. Economic Aspects
6.9. Toxicology and Occupational Health
7. Fullerenes
7.1. History
7.2. Production

7.2.1. Vaporization of Graphite
7.2.2. Fullerene Synthesis by Combustion
7.2.3. Pyrolysis of Naphthalene
7.3. Structure and Nomenclature
7.4. Physical Properties
7.5. Chemical Properties
7.5.1. Addition Reactions
7.5.2. Hydrogenation
7.5.3. Halogenation
7.5.4. Reduction
7.6. Endohedral Fullerenes
7.7. Carbon Nanotubes and Nanostructures
7.8. Application Properties of Fullerenes

[Top of Page]

1. Introduction

Otto Vohler, Gabriele Nutsch, Gerd Collin, Ferdinand von Sturm

Carbon is the sixth element of the periodic table with an atomic mass of 12.01115 ± 0.00005. It consists of 98.89 % of the isotope ^{12}C, 1.11 % of the isotope ^{13}C, and several unstable isotopes with mass numbers between 9 and 16. The isotope ^{14}C, which is formed in the atmosphere by the reaction $^{14}N(n,p)^{14}C$, is used in archaeology for age determination and tracing of chemical compounds [29]. The half-life of this β emitter is about 5570 years.

Carbon is the thirteenth most abundant element. The earth's crust, including the lithosphere, hydrosphere, and atmosphere, contains about 3×10^{16} t of carbon, which corresponds to 0.09 wt % or 0.15 atom %, most of which is present as compounds, and only a small portion in elementary form. The carbon contents of sea water, the atmosphere, and the biosphere are about 4×10^{13} t, 4×10^{11} t, and 3×10^{11} t, respectively.

Carbon exists in several allotropic modifications (Fig. 1). The modifications differ from one another in the arrangement of the carbon atoms in their crystals (Figs. 2 and 3).

Figure 1. Allotropic modifications of carbon

Figure 2. The three most important allotropic forms of solid elemental carbon and their main derivates [30]

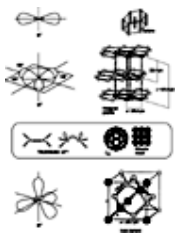

Figure 3. Bonding hybridization and corresponding crystal structure of the carbon allotropes [30]

Diamond is the metastable phase of carbon under standard conditions and can be transformed into graphite by heating above 2000 K [31] (see Fig. 4). The transformation of graphite into diamond (diamond synthesis) is only possible at high temperatures and high pressures [32]. The atoms in the diamond lattice are singly bonded. The four symmetrical sp^3 σ-bonds point to the corners of a regular tetrahedron (Fig. 3). Therefore, diamonds have a rigid and isotropic structure and are usually cubic. Hexagonal diamond (londsdaleite) was first found as a product of diamond synthesis, but meanwhile it has also been discovered in meteorites [33]. According to Fig. 4 diamond can be produced at low pressure by chemical vapor deposition (CVD; see Section The Diamond Chemical Vapor Deposition (CVD) Process).

Graphite is characterized by its layered lattice structure due to association by π-bonds perpendicular to the σ-bonds layers. Hence, graphite is highly anisotropic. However, only a few natural graphites and hot-worked pyrographite samples are nearly perfect crystals. Most carbon materials deviate more or less from the perfect crystal structure. The differentiation between graphite and carbon, a term used for the less perfect structures such as metallurgical coke or charcoal, is only for the sake of convenience. Graphite is the thermodynamically most stable allotrope of carbon at room temperature and pressure (see Fig. 4). It is characterized by double bonds between sp^2 carbon atoms (see Fig. 3). At atmospheric pressure, graphite vaporizes without melting at ca. 3895 – 4020 K. The heat of sublimation depends on the vapor composition. It amounts to 711, 825, and 788 kJ/mol for mono-, di-, and triatomic molecules, respectively. At 4000 K, about 80 % of carbon vapor consists of C_3 molecules. By using a laser-heating technique, a graphite – liquid – vapor equilibrium pressure of 10.7 ± 0.2 MPa at a temperature of about 4000 K was found [34].

Chaoite, a member of the carbyne group, was found in small quantities together with graphite in the rocks of the Nördlinger Ries in Germany, which were transformed by meteoritic impact [35]. Small quantities of other members of the carbyne group were found together with graphite in carbon layers formed by vaporized carbon, in laser-heated carbon, and also in meteorites. The transformation of *sp*-hybridized carbon molecules (carbyne) into diamond by shock wave experiments has been investigated [36]. A common feature of the numerous members of this group are triple bonds [37-39].

Fullerenes are the most recently discovered allotropic form of carbon [40]. In October 1996, R. E. SMALLEY, R. F. CURL, and H. W. KROTO received the Nobel prize for their discovery of the C_{60} cage molecule in 1985. However, the method for their bulk production wass discovered only in 1990 [41]. These all-carbon cage molecules constitute the third allotropic form of carbon. They are composed of more or less deformed sp^2-bonded carbon atoms. Fullerenes are formed in condensing carbon vapor in which linear carbon chains link together and form graphitic sheets. These free-floating sheets cannot tie up the dangling bonds on their edges. Elimination of the dangling bonds by curling up serves the physical tendency to reach the lowest energy level available. Thus, the fullerenes are a pure, finite form of carbon. The two other forms, diamond and graphite, are infinite network solids.

Giant fullerenes in the form of *nanotubes* and *buckyonions* have further expanded the realm of this new form of carbon. The fullerene-related nanotubes are formed by a single-layer graphitic network of hexagons (Fig. 2). The presence of pentagonal or heptagonal carbon rings renders their physical properties different from those of graphite. There are no sp^3 bonds, grain boundaries, or edges, and they contain no hydrogen [30], [42].

Fullerene derivatives can be created by adding endo- or exohedral substituents. This opens the possibility of obtaining a wide range of building blocks for developing new materials with unique properties.

[Top of Page]

2. Diamond [43-45]

Otto Vohler, Gabriele Nutsch, Ferdinand von Sturm, Erhard Wege

2.1. General Sources

For many centuries, natural diamond has been appreciated as a valuable gem and investment. The first center of mining was India, which was replaced by Brazil in the early 1700s, South Africa in the mid-1800s, and most recently Kongo (formerly Zaire), Australia, Russia, Botswana, and Namibia. Output amounted to ca. 55 × 10^6 ct in 1998, but this does not satisfy the increasing demand for industrially used diamonds. Table 1 lists the diamond mining output of some of major producing countries and the total world mining production in 1998.

The sources of natural diamonds are probably ultrabasic rocks, especially eclogite, which consists essentially of omphacite and pyrope. The famous blue ground of Kimberlite slots is a decomposed eclogite rock containing about 0.5 ct of diamond per tonne. Such primary deposits are found in South Africa, Russia, and Tanzania. Most of occurences of diamond are secondary deposits. From so-called placers, of which the Lüderitzbay deposit in Namibia is an example, diamond is mined explosively.

Less than the half of the mined diamonds can be used for gems; the majority are used for industrial purposes. The prospering interest in industrial diamonds is based on their combination of outstanding properties besides their high hardness, which makes them highly suitable for many applications. Synthetic diamond has been produced commercially for over 30 years [46-48] by high-pressure, high-temperature (HPHT) techniques (see Fig. 4) Synthetic industrial diamond is available in far larger quantities than its natural diamond counterpart. The estimated capacity for synthetic diamond of 100 × 10^6 ct in 1982 increased to 350 × 10^6 ct in 1990 and to more than 500 × 10^6 ct in 1999 (see Table 2). Consequently, synthetic diamond accounts for more than 90 % of the industrial diamond used. Various reports estimate the total industrial diamond output worldwide to be more than 600 × 10^6 ct, valued at (600 – 800) × 10^6 dollars. More than 30 countries produced industrial diamond in 1998. In addition to the countries listed in Table 2, Korea produces synthetic diamond, but specific data on their output could not be confirmed. China and Russia also presumably produce far more than the output given in the table. Thus, two-thirds of industrial diamond output is concentrated in four countries: the United States, Ireland (possibly the world's largest producer), Russia, and South Africa.

Table 1. World diamond mining output in 1998, 10^6 ct

Country	Gem	Industrial	Total
Australia	18.897	23.096	41.993

Russia	9.250	9.250	18.500
Congo	3.000	15.000	18.000
Botswana	11.000	5.000	16.000
Rep. of South Africa	5.360	6.000	11.300
Angola	3.600	0.400	4.000
Namibia	1.300		1.300
Brazil	0.700	0.600	1.300
China	0.230	0.900	1.130
Ghana	0.125	0.505	0.630
Other	1.955	0.852	2.807
World	55.400	61.600	117.000

Table 2. Synthetic diamond, estimated world production, 10^6 ct

Country	1995	1997	1999
Belarus	25.0	25.0	25.0
China	15.5	16.0	16.5
Czech Republic	5.0	5.0	3.0
France	3.0	3.0	3.0
Greece	1.0	0.75	0.75
Ireland	60.0	60.0	60.0
Japan	32.0	32.0	32.0
Poland	0.256	0.26	0.2
Romania	5.0	5.0	3.0
Russia	80.0	80.0	80.0
Slovakia	5.0	5.0	3.0
South Africa	60.0	60.0	60.0
Sweden	25.0	25.0	25.0
Ukraine	8.0	8.0	8.0
United States	115.0	125.0	208.0
Total	440.0	451.0	467.0

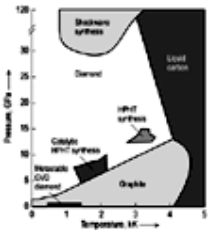

Figure 4. Phase diagram of different carbon modifications

Interest in industrial diamond has been further increased by the more recent discovery [49], [50] that it is possible to produce polycrystalline diamond films or coatings by a wide variety of chemical vapor deposition (CVD) processes from a hydrocarbon gas in an excess of hydrogen. This CVD diamond can show mechanical, tribological, and even electronic properties comparable to those of natural diamond.

2.2. Properties and Structure of Natural Diamond

Diamond is a colorless mineral, generally characterized by its hardness (Mohs hardness 10) and brilliance after cutting. The density is 3.515 g/cm^3 (cf. graphite 2.266 g/cm^3). The lattice of the cubic modification is face-centered with the lattice constant a = 0.3567 nm. The atomic arrangements in diamond is shown in Figure 3. Diamond is tetrahedrally bonded, and all atoms have equal bond lengths of 0.1544 nm and are linked by overlapping sp^3 hybrid orbitals (sp^3 – sp^3 σ-bond). The bonding angle of the covalent hybrid bonds is 109.28°. This contrasts with the hexagonal structure of graphite with trigonal bonding (Fig. 3). The sp^2 – sp^2 σ-bonding is supplemented by p – p π-bonds oriented perpendicularly to the σ-bonds. All σ-bonds lie in the same plane, and the bond angle is 120°. The stacking sequence for the carbon atoms in the diamond structure is ABCABC (Fig. 5 b). It shows a chair configuration (hatched) of puckered hexagonal rings in both the vertical and horizontal stacking directions. The first carbon atom that starts nucleating a new {111} diamond layer is shown in the Figure 5 a (top). By addition of a further carbon atom the basic unit of the structure is obtained. This is also the nucleation kernel for a new plane.

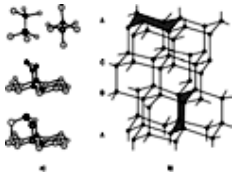

Figure 5. Stacking of carbon atoms in diamond crystals

The color of diamonds can vary from white to pink, yellow, light blue, and even black (carbonado), depending on the incorporated impurities. On the basis of the different characteristics of the IR absorption spectra of natural diamond, a distinction is made between Type I and Type II diamond, with several subgroups. A majority of gem-quality diamonds are of Type I a and contain substitutional nitrogen impurities up to 0.5 atom %. These diamonds have a characteristic, slightly greenish yellow color. Only about 1 % of natural diamonds contains impurities in the form of substitutionally isolated nitrogen atoms. These diamonds are classified as Type I b. All synthetic diamonds belong to this group. Their nitrogen atoms did not havesufficient time to form aggregates, in contrast to the natural diamonds of Type I a, most of which remain deep in the earth's mantle after their formation and are exposed to extremely high temperatures and pressures. During this phase, the majority of the nitrogen atoms can form aggregates.

Diamonds without measurable substitutional nitrogen atoms or aggregates are classified as Type II a. They are very rare in nature; less than 1 % of all diamonds belongs to this group. Their absorption bands are caused only by the vibrational absorption of the lattice that is common to all diamonds. They are often colorless or light pink. Type II a diamonds are electrical insulators like Type I, but have excellent optical and thermal properties. Type II b diamonds containing ca. 1 ppm of boron atoms and even smaller amount of nitrogen impurities are the purest form of naturally occuring diamonds. They are usually blue to bluish gray in color. The presence of boron atoms makes these diamonds to electrical semiconductors.

Diamond crystals cleave well parallel to (111). The cutting hardness is different for the various crystal faces, and this must be taken into consideration when cutting. The coefficient of friction depends on the purity of the surfaces. With increasing degasification, the coefficient rises from about 0.1 to 0.6 [51], [52].

Diamond is known for its combination of its outstanding properties: it has extreme mechanical hardness (ca. 90 GPa, or Mohs hardness 10), it is the strongest known material with the highest bulk modulus (1.2×10^{12} N m^{-2}) and lowest compressibility (8.3×10^{-13} m^2 N^{-1}). Its thermal expansion coefficient at room temperature of 0.8×10^{-6} K^{-1} is comparable with that of low-expansion iron – nickel alloys (Invar). The specific heat is 6.2 J g^{-1} K^{-1} at 300 K, 13.55 J g^{-1} K^{-1} at 500 K and 21.6 J g^{-1} K^{-1} at 1000 K. Diamond Types I and II a are good electrical insulators, with a room-temperature resistivity of about 10^{16} Ω cm. However, diamond can be doped, for example with boron (Type II b), to vary its resistivity over the range 10 – 10^6 Ω cm, so that it becomes a semiconductor with a wide band gap of 5.4 eV.

Diamonds are excellent conductors of heat. Type II a diamond has the highest known thermal conductivity at room temperature of 2×10^3 W m^{-1} K^{-1}, which is five times higher than that of copper. Diamond has a broad optical transparency from the deep - UV to the far - IR“ndash” OK? (Korektur nicht eindeutig auch im Folgenden) region of the electromagnetic spectrum. Moreover, it is biologically compatible and very resistant to chemical erosion and abrasive wear at room temperature. At higher temperatures, it is attacked by oxygen, oxidizing agents, and carbide-forming elements. At temperatures of about 2000 K and above, diamond is converted to graphite, both internally and on the crystal surfaces. The rate of phase transformation increases rapidly with increasing temperature [53].

2.3. Diamond Synthesis

According to the phase diagram (Fig. 4) there are only a few ways to synthesize diamond: by shockwaves (mostly used in Japan), by high pressure and high temperature (HPHT synthesis), and by chemical vapor deposition (CVD) at relatively low pressures and moderate temperatures.

2.3.1. *HPHT Synthesis*

Figure 4 shows how the state of carbon phases depends on temperature and pressure [54]. Diamond is a metastable phase at standard temperature and pressure, and graphite is the only thermodynamically stable carbon phase. The transformation of graphite into diamond is possible at high temperatures and pressures [55-57]. Above 12.5 GPa and temperatures of about 3000 K, graphite spontaneously collapses into polycrystalline diamonds.

The difficulty in the diamond synthesis process has always been the simultaneous achievement of sufficiently high pressures and temperatures. These conditions require special apparatus which was first developed by P. W. BRIDGMAN [58] in the form of a self-sealing high-pressure packing which operated without leaking and permitted hydraulic operation, namely, the unsupported area seal. For the invention of the apparatus to produce extremely high pressures, BRIDGMAN was awarded the Nobel Prize in 1946.

The General Electric process [46-48] uses a modified Bridgman apparatus [59], [60]. In the “catalytic” process, graphite is transformed into diamond by first dissolving it in molten metals and then cystallizing it above equilibrium pressure. Nickel is the most widely used metal for dissolving carbon, for which it has a high solubility (ca. 4 wt % at 1500 °C and 5.7 GPa). Under these conditions, graphite is metastable, and diamond, which is less soluble (ca. 3.6 wt %) crystallizes out. In the past it was assumed that the metals used (e.g., nickel, manganese, iron, cobalt, palladium, platinum, and alloys thereof) act as catalysts. The term catalyst is still frequently used, but it is now clear that the metals serve purely as solvents [61], and that the growth process is analogous to the traveling-zone solvent technique [62].

Any form of carbon can be used as the starting material [63]. Naphthalene initially forms graphite; paraffin wax and sugar are

converted directly to diamond. Even peanuts are transformed to the extent of up to 60 % into diamond at 2000 K and 10 GPa in 5 min.

The belt and multipiston diamond – synthesis apparatus is shown in Figure 6. It consists of two truncated conical pistons and a prestressed cylinder assembly, which forms the reaction chamber. The circular gaps between the sections are filled with conical gaskets made of pyrophyllite, a hydrous aluminium silicate, coated with a thin dusting of hematite (Fe_2O_3) to further increase the friction. The gaskets act as seals, pressure transmitters, and thermal and electric insulators. By using strongly tapered pistons (a and c in Fig. 6) a large piston motion becomes possible without squeezing the pyrophyllite gasket too thin. The compressed gasket in turn provides additional strength by supporting the pistons. This arrangement is suitable for a large volume of starting material, which consists of stacked graphite and metal plates or a mixture of graphite powder and nickel or other group 8 – 10 metal powder, and is heated directly by resistance heating. The current is supplied via pistons, current rings and nickel plates. Since the graphite has the highest resistance in the current path, it is heated to the desired temperature. The nickel melts, dissolves the graphite, and begins to crystallize out diamond. The molten metal film is less than 0.1 mm thick, and the movement of the carbon through this metal film is quite rapid. In a few minutes the nickel film moves through the graphite from both ends to the hotter center, and all graphite is converted to diamond. The pressure is maintained until after current has been cut off and the sample cooled to prevent reversion to graphite. After cooling, the diamonds are separated by dissolving the metal with acids. The average size of the synthetic diamonds is about 0.05 – 0.5 mm. The lifetime of this apparatus with a diameter of up to 40 mm is ca. 100 to 1000 cycles when properly operated. Smaller diamonds are produced only by shockwaves (see Fig. 4).

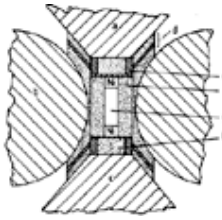

Figure 6. Belt-type pressure apparatus [139]

a) Top piston; b) Cylinder; c) Bottom piston; d) Gasket assembly; e) Current ring; f) Sample holder; g) Graphite; h) Insulating pill

Diamond crystals can be controlled in size and mechanical properties by varying the process parameters such as the form and nature of graphite, the metal solvent used, temperature, pressure, and time. Grit up to 0.5 mm in size is produced without crushing. Also the shape of the crystals can be controlled by the temperature. Because of this size- and shape-control ability, synthetic diamond grit is generally superior to natural diamond. Cutting blades made from GE diamonds have been manufactured for many years [64], [65]. For cutting rocks or glass, blocky grit sintered in a metal matrix is used. For shaping hard metals and tungsten carbide, a friable grit in a resinoid matrix is preferred. For polishing pastes, the grit is often mixed with olive oil, lithium stearate, and some amorphous silica. There are also diamond compounds for polishing based on water with polyethylene glycol, potassium hydroxide in methanol, and amorphous silica. There is no difference in hardness between synthetic and natural diamond.

About 15 years after the first announcement of industrial diamonds, single crystals in gem quality were also grown by this method [63]. The equipment is similar to that described above, but the large, high-quality crystals must be grown very slowly (about one week is necessary to grow a 5 mm, 1 carat, high-quality diamond crystal) and under much more carefully controlled and stable conditions. This makes gem diamond growth so difficult and expensive.

2.3.2. *The Diamond Chemical Vapor Deposition (CVD) Process*

Chemical vapor deposition (CVD) of diamond films has developed dramatically [66-74] since its initiation in the 1950s [49], [50]. The diamond film deposition process utilizes temperature and pressure conditions under which graphite is the stable form of carbon (see Fig. 4). However, the kinetic factors of the CVD process allow crystalline diamond to be produced by the typical net reaction:

$$CH_4(g) + \text{activation energy} \rightarrow C(\text{diamond}) + 2\,H_2(g)$$

In addition to methane, a wide variety of carbon-containing substances have been used as precursor materials. The diamond CVD process involves chemical reactions in the gas phase as well as at the solid surfaces which cause depositions onto the surfaces (Fig. 7). Usually, the CH_3 radical is regarded as the precursor. This depends on the boundary conditions at the substrate. With high-energy plasma jets, atomic carbon is the most important precursor species.

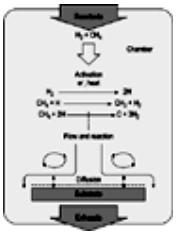

Figure 7. Diamond CVD process (after [71])

All CVD techniques used for producing diamond films require:

- a particular gas atmosphere and hence, a closed reaction chamber
- a means for activating the gas-phase carbon-containing precursor molecules
- a substrate temperature of about 400 to 1000 °C

– dissolution of the precursor gas in an excess of hydrogen

While HPHT synthesis requires special industrial equipment used by only few companies, the CVD of diamond films can be performed by using more or less the laboratory equipment. Deposition methods are diverse.

The Role of Hydrogen. Hydrogen is a very important species for diamond growth [75]. Hydrogen plays two main roles. First, molecular hydrogen suppresses the formation of aromatics in the gas phase; second, atomic hydrogen activates the surface.

With pure CH_4, growth of a layer also occurs, but it consists of both diamond and unwanted carbon phases (Fig. 8). With hydrogen alone, only etching of both graphite and diamond occurs. The graphitic phases are always more readily etched by reaction with hydrogen atoms than the diamond phase. Therefore, hydrogen atoms should be present in higher than equilibrium concentration to obtain successful diamond growth and etching of non-diamond carbon phases.

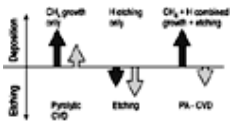

Figure 8. Schematic depiction of the growth of diamond (dark arrows) and removal of unwanted carbon phases (light arrows)

Due to the excess of hydrogen the diamond surface is almost fully saturated with hydrogen during the growth process (step 1 in Fig. 9). Desorption of bound hydrogen by incoming atomic hydrogen creates vacant sites or surface radicals (step 2 in Fig. 9). These vacant sites can be refilled by atomic hydrogen (step 3 a in Fig. 9), or H atoms can be added to unsaturated carbon – carbon bonds. Thereby, conversion into sp^3 bonds at low temperatures or gasification of sp^2 bonds at high temperatures takes place, and diamond growth occurs (step 3 b in Fig. 9).

Figure 9. C, H bounded C and H atoms, Ⓗ•, ©• excited atoms

Without activation of the gas by plasma or other energetic sources, the concentration of atomic hydrogen, and therefore the number density of active sites, is low. Hence, the film – growth rate is very low.

The Role of Substrate Temperature. The deposition of polycrystalline diamond films occurs in a surface temperature range of about 850 – 1500 K. The lower limit approaches about 700 K and is nearly independent of the deposition technique [72], [73], [76], but it depends on the precursor used, the growth rate is very low, and non-carbon phases are implanted into the layer [77], [78]. However, the growth rate can be enhanced by using halogenated precursors [79]. The surface chemistry depends strongly on the surface temperature. At low temperatures, the rate of creation of active sites is low. Increasing the temperature also increases this creation rate. Above about 1500 K the number density of active sites decreases due to thermal decomposition and subsequent graphitization. The sp^2 bonds are rapidly etched by the hydrogen atoms.

At too low surface temperatures the condensation of aromatics can lead to amorphous carbon phases. Therefore, an efficient diamond growth rate can be controlled by the surface temperature, the absolute number density of the carbon-containing species and atomic hydrogen, and the ratio of the two.

By varying the substrate temperature, the properties of the diamond film can be tailored. It varies from well-defined facets to irregular forms or round shapes. Variations in the diamond morphology with the substrate temperature are shown in Figure 10 for the DCTP jet. With increasing substrate temperature, it changes from predominantly more {111} to {100} facets, which was also found for an RF-ICTP Jet [80] and for DCTP CVD [81]. The crystallite size also increases with increasing substrate temperature [81-86], as is shown in Figure 10. However, the mechanism of diamond growth is very complex, and contrary results were also published [87]. Therefore, the substrate temperature must be kept constant during the deposition time and must be as uniform as possible across the entire surface area to minimize or prevent radial variations in the morphology of the crystallites. In the thermal plasma jet diamond CVD the substrate must be intensively cooled. Special designs of substrate holder exist [88], [89] by which either the different thickness or/and the varying materials influence the heat transfer to the cooling fluid (water is mostly used, but also ethylene glycol). To adjust for varying heat fluxes from the plasma jet to the substrate, special cooling systems have been developed [90], [91].

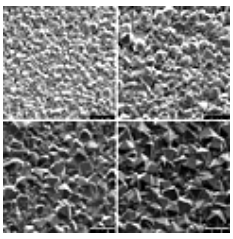

Figure 10. Morphology of diamond films manufactured with a DCTP jet and 5 vol % CH_4 in H_2 at various substrate temperatures

a) 780 °C; b) 850 °C; c) 950 °C; d) 1050 °C

The Role of the CH_4/H_2 Ratio. The morphology and therefore the quality of the diamond films depend on the ratio of the precursor dilution in hydrogen. The morphology can be changed in a similar way as for the substrate temperature with an

increasing methane/hydrogen ratio (Fig. 11). The ratios vary considerably with the deposition method used. The ratios of more than 5 vol % given in [76], [91], [92] are relatively high compared to those reported in [93-95] with less than 5 vol %. For epitaxial growth [96] or with nonthermal plasma techniques, concentrations of less than 1 vol % CH_4 based on the H_2 content are used.

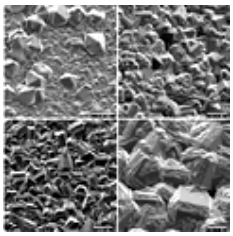

Figure 11. Morphology of DCTP diamond films at varying volume content ratios of CH_4 to H_2 at a substrate temperature of 1050 °C

a) 4 vol % CH_4; b) 5 vol % CH_4; c) 7.5 vol % CH_4; c) 10 vol % CH_4

Precursors and Substrates. A wide variety of precursor and substrate materials has been tested. Methane has been most frequently used as precursor material. Other carbon-containing substances [97], [98] such as aliphatic and aromatic hydrocarbons, alcohols, amines, ethers, are also used:

1. Gaseous
    a. Methane
    b. Ethane
    c. Ethylene
    d. Acetylene
2. Liquid
    a. Methanol
    b. Ethanol
    c. 1-Propanol
    d. 1-Butanol
    e. Acetone
    f. Benzene
    g. Cyclohexane
3. Halogenated
    1. Dichloromethane
    2. Trichloromethane
    3. Carbon tetrachloride
    4. Trifluoromethane
    5. Tetrafluoromethane

More recently, halogenated hydrocarbons have been recommended, mainly because of indications of enhanced diamond nucleation density [99] and lower substrate temperature [100]. Substrate temperatures below 300 °C are possible. An overview on halogenated precursors is given in [101], [102]. However, the use of fluorine restricts the choice of substrate materials; for example, silicon is always damaged, even if only a small amount is added. Chlorine-containing compounds prohibit the use of stainless steel vessels.

The variety of substrate materials used are listed in in the following:

1. Carbons
    a. Amorphous carbon
    b. Graphite
    c. Diamond
2. Metals
    a. Aluminum
    b. Titanium
    c. Molybdenum
    d. Nickel
    e. Copper
    f. Tungsten
    g. Tantalum
    h. Platinum
    i. Germanium
    j. Iridium
    k. Silicon
    l. Tool steel (with interfacial stainless steel layer)
3. Ceramics
    a. Boron carbide
    b. Cubic boron nitride
    c. Silicon carbide
    d. Silicon nitride
    e. Titanium carbide
    f. Zirconium carbide
    g. Hafnium carbide
    h. Chromium carbide

i. Molybdenum carbide
j. Cemented carbide, such as WC – Co compound
k. Alumina
l. Silica

4. Polymers
   a. Polyimide (with interfacial layer)

Materials which easily form carbides are particularly suitable for diamond deposition. Due to the high solubility of carbon in tool steel, interfacial layers have to be deposited, such as tungsten or silicon carbide and silicon nitride.

Silicon single crystals, commonly taken from the microelectronics industry, are the most frequently used substrate materials in laboratory reactors. However, well-adhering, high-quality diamond films are necessary to realize coatings on industrially important substrate materials, such as ceramics, steels, nickel alloys, cemented carbides, and most alloys containing transition metals. Basically, all samples are diamond-ground and washed with alcohol in an ultrasonic bath. Different pretreatments have to be performed on these as-ground substrates to enhance nucleation and adhesion. Chemical and mechanic pretreatments increase the surface roughness, which improves both the nucleation density and the mechanical anchoring of the film. A review on different chemical and mechanical pretreatments is given in [103], especially for cemented carbides.

Another means to improve the nucleation density is a secondary discharge to the substrate. It is used to enhance the chemical nonequilibrium in the boundary layer in front of the substrate. In this way, the substrate can be cleaned when negative biased. The diamond growth rate is improved by positive biasing [104] due to the increased flux of atomic hydrogen to the substrate surface.

2.3.3. *Diamond CVD Techniques*

Today a wide range of methods exists to prepare CVD diamond layers. The techniques can be categorized according to the specific method of producing chemically active species, such as thermal decomposition, hot-filament technique, combustion process, plasma-assisted processes, and beam technologies (Fig. 12). The plasma-assisted processes include nonthermal plasmas, such as the direct current (DC) glow discharges, capacitively and inductively coupled 13.45 MHz radio frequency (CC- and IC-RF) plasmas, microwave (µW) plasmas excited by 915 MHz and 2.45 GHz radiation, as well as thermal plasmas (TP), such as the thermal inductively coupled radio frequency plasmas (IC-RFTPs) or thermal direct current plasmas (DCTPs) which produce plasma jets with high velocities and temperatures.

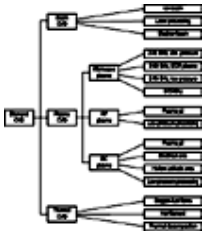

Figure 12. Diamond CVD technologies (from [97])

Beam technologies, such as electron or ion beams, and lasers are not often used for CVD diamond film production. Initially, thermal decomposition was used for diamond film production. The gas and substrate temperatures must be high enough to decompose the precursor gas molecules, and the resulting growth rates are too low to be industrially interesting. Therefore, plasma CVD methods were developed. Energy transfer from activated species of the plasma to precursor molecules leads to their decomposition and to chemical reactions already in the gas phase. In this way the deposition rate can be enhanced without increasing the substrate temperature.

2.3.3.1. *Nonthermal Plasma CVD*

Plasmas generated under high vacuum conditions are characterized by different temperatures of electrons and heavy particles such as ions, atoms, molecules, and radicals [105]. The low pressure is tantamount to low number densities. The long mean free path enables the electrons to gain high energies, but energy transfer to the heavy particles is reduced due to the low collision probability. Therefore, plasmas at pressures lower than 0.01 MPa are no longer in thermodynamic equilibrium. Such plasmas have the advantage of a large and homogeneous plasma volume; the disadvantages are the low gas velocities, thick boundary layers in front of the substrate [106], and medium growth rates (see Fig. 13).

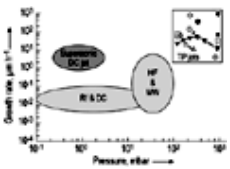

Figure 13. Linear growth rate of various plasma jet techniques in comparison to other techniques, such as HF (Hot Filament), MW (microwave), RF and DC (radio-frequency and d.c. nonthermal plasma processes)

The *2.45 GHz microwave discharge* was the first and is still most frequently used diamond CVD technique with plasma discharge. An overview on the different microwave reactor types used for diamond CVD is given in [107]. The NIRIM reactor type was the first reactor reported [108] and operates with a power of up to 1500 W and pressures typically below 1 kPa. The carbon-containing gas mixture consists of 1 % methane in hydrogen, and the growth rate is about 0.5 µm/h for substrate temperatures around 1000 °C. The main drawback of this reactor type is the smallness of the substrate, which is only 2 – 3 cm in diameter. In the dome-shaped bell-jar reactor [110], three-inch silicon wafers are coated with polycrystalline diamond

at substrate temperatures of 900 – 1000 °C, pressures of about 1 kPa, and 1 % methane in hydrogen gas at a growth rate of about 0.2 – 0.3 µm/h. ASTex improved and commercialized the bell-jar reactor in the 1990s for using four-inch-diameter substrates. The highest growth rate reached with a high-pressure microwave source (HPMS) reactor with 5 kW and about 10 kPa amounts to 14 µm/h. The drawbacks are connected with the 2.45 GHz frequency: the power density, plasma size, and therefore the diamond deposition rate are limited. Nonuniform temperature distributions and film thickness, as well as variations in phase purity are possible. Shifting to 915 MHz can help to overcome the power limitation and to increase the geometric uniformity.

On the basis of mathematical simulations, the more recently developed ellipsoid reactor [109] for large-scale diamond film production was designed. A variety of different reactor sizes and designs have been explored, especially for 915 MHz. The plasma has the shape of a hemisphere sitting on the substrate. The design was commercialized by the AIXTRON Company. By using the AIX P6 (6 kW; Fig. 14) at 2.45 GHz and, in particular, the 915 MHz AIX P60 (60 kW) [111], free-standing polished diamond windows up to six inches in diameter (Fig. 15) can be produced with a maximum growth rate of 12 µm/h at a maximum pressure of 20 kPa.

Figure 14. Ellipsoidal microwave reactor AIX P6 [111]

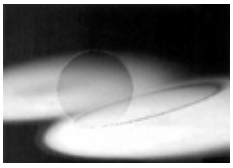

Figure 15. Optical-window-quality free-standing diamond films [110]

As is shown in Figure 13, the growth rate can be increased also in the lower pressure region by means of *DC supersonic jet synthesis* [112]. Here a distinction is made between the DC plasma jet, which is more or less a thermal plasma jet, and the DC supersonic jet, which is also arc-heated but has a large deviation from thermodynamic equilibrium. Typically, the arc is generated inside a torch, similar to a DC plasma torch. The jet leaves the converging/diverging nozzle (Fig. 16) with a minimum diameter of about 1 mm in the reactor chamber, maintained at a pressure of 2 – 3 kPa [113], which is much lower than that in the torch (ca. 0.6 MPa). The high pressure difference is responsible for the gas dynamic conversion of thermal arc energy into high kinetic energy of the jet, which expands into the reactor with supersonic velocities (1 – 10 km/s) [114]. The temperatures in the jet are not higher than 2 kK, and reach the highest values of up to 5 kK in the bow shock near the substrate [115]. This kind of jet shows a large deviation from the local kinetic and chemical equilibria. The high flow velocities lead to small transport timescales relative to the chemical kinetic timescales [116]. Therefore, the recombination of atomic hydrogen is reduced and the carbon-utilization rate is high. The growth rate ranges between 6 [117] and 10 µm/h [118], which is much higher compared to the other nonthermal plasma techniques operating in the same pressure region (see Fig. 13). This results from the thinning of the boundary layer thickness due to the high flow velocities.

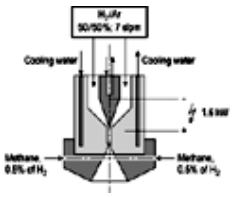

Figure 16. Schematic of the supersonic DC jet torch [115]

2.3.3.2. *Thermal Plasma Diamond CVD*

A promising alternative to the conventional low-pressure gas discharges is the use of high-pressure thermal plasma jets. Plasma jets are generated by electric discharges at moderate to high pressure (10 – 100 kPa and higher), so that the plasma leaves the nozzle with both high temperature (> 6 kK) and high velocity (often > 500 m/s). Due to the high plasma temperatures, the reactant decomposition process is much more intensive. All of the molecular hydrogen is dissociated into atomic hydrogen, the most important precursor in the diamond deposition process. The delivery of the atomic hydrogen to the substrate is enhanced because of the reduced thickness of the boundary layer due to the relatively high gas velocities. Thus, the highest growth rates are obtained with thermal plasma jets, as is shown in Figure 13.

The plasma jets can be categorized in accordance with the method by which the electric discharge is sustained. The radio frequency inductively coupled discharge and the direct current discharge, shown schematically in Figure 17, are the most frequently used plasma jets.

Figure 17. RF-ICTP jet and DC-TP jet

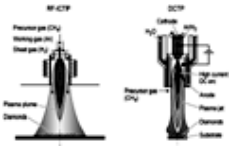

The radio-frequency inductively coupled thermal plasma (RF-ICTP) is an electrodeless plasma and is therefore almost free of contaminants. Two separate gas flows are necessary for maintaining the RF discharge: the sheath gas for discharge stabilization and for preventing direct contact between the plasma and the plasma-confining wall, and the central gas flow for the working gas. The features of the RF-ICTP are the large volume, and consequently the low flow velocities of the plasma exiting the torch, with or without a nozzle. The diameter of the plasma jet leaving the torch ranges from about 30 [119], [120] through 76 [121], up to 100 mm [122], depending on the torch geometry [124] and the reactor pressure [125]. The velocities at the nozzle exit decrease with increasing diameter and range from 30 m/s [126] up to 250 m/s [127]. Supersonic plasma jets generated by RF-ICTP with converging/diverging nozzles are also possible [128]. An other feature of the RF-ICTP is the possibility for axial precursor injection by means of a water-cooled feeding probe. Moreover, there are no restrictions on the kind of gases used.

The direct current thermal plasma (DCTP) is generated in a nontransferred plasma torch by the high-current arc generated between the cathode, usually made from tungsten, and an anode shaped as a nozzle. The nozzle diameter ranges from 3 to 8 mm, depending on the power level used. The argon – hydrogen gas mixture is heated by the high-current arc and leaves the anodic nozzle as a currentless plasma jet with a much higher velocity in comparison to the RF-ICTP jet [119], [120]. Impinging onto the surface, the DCTP jet also provides a higher momentum. However, the small diameter of the DCTP leads to small deposition areas. The deposition on extensive or three-dimensional substrates requires movement of the substrate. In contrast to the RF-ICTP torch, precursor injection must occur outside of the torch to prevent contamination of the tungsten electrode by carbon.

The main differences between the DCTP jet and the RF-ICTP jet are the temperature distributions in the axial and radial directions. The DCTP jet has steep gradients in both radial and axial directions. In the case of the RF-ICTP jet, the temperatures in both the axial and radial directions decrease much less within the plasma core than in the thermal DC plasma jet. The most important difference between the DCTP and ICTP jets is the thickness of the boundary layer in front of the substrate. In general, the boundary layer thickness of the ICTP jet is on the same order of magnitude as in the low-pressure plasma techniques. The higher the velocity of the jet impinging the substrate, the thinner is the boundary layer. In the case of the DC plasma jet, the boundary layer extends only to few tenths of millimeters. Therefore, chemical reactions cannot occur inside the boundary layer (“frozen” chemistry). Thus, the high temperatures at the end of the boundary layer of about 4000 K generate a lot of atomic hydrogen and carbon. Therefore, atomic carbon is the most important species for diamond deposition with the DCTP Jet [106].

The experimental set-up for Plasma Jet Diamond CVD (Fig. 18) consists of the plasma torch with its power supply, the gas supply with mass flow controllers for exact control of the gas flow and composition, the reactor chamber, the control unit for cooling water, and the automatic registration unit for plasma parameters (chamber pressure and substrate temperature, measured by a thermocouple in the substrate holder and by a pyrometer on the substrate surface) [129].

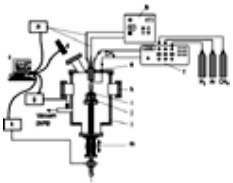

Figure 18. DCTP jet diamond CVD equipment

a) Cooling water temperature measurement; b) DC power supply; c) Automatic registration unit; d) Pyrometer; e) Plasma torch; f) Mass flow controller; g) Pressure measurement; h) Window; i) Substrate; j) Water atomizer; k) Thermocouple (substrate temperature); l) Reactor wall, water-cooled; m) Substrate holder, movable

2.3.3.3. *Other CVD Techniques*

The *hot filament* deposition method is a simple technique and is therefore widely used. The principal task of the hot filaments is to dissociate molecular hydrogen into atomic hydrogen. Bare metal wires generate atomic hydrogen more efficiently than silicon carbide filaments. Atomic hydrogen is generated by thermal dissociation of molecular hydrogen at the hot solid filament surface. The upper operating temperature (ca. 2200 °C) of the filament restricts the quantity of atomic hydrogen produced. To overcome this, low pressures (e.g., < 10 kPa) are used to enhance its production and transport to the substrate. The substrate distance is relatively small (ca.1 – 20 mm). However, the low gas-phase concentration permits only relatively low growth rates compared to the plasma jet techniques (Fig. 13). Because of its low cost and its very simple construction, the hot filament process has been the most frequently used method since it was first reported [130]. The methane content based on the flow rate of molecular hydrogen is always less than 5 % [131]. A higher methane content lowers substantially the rate of recombination of atomic hydrogen at the substrate surface. Besides the relatively low growth rate, another drawback is filament recrystallization due to exposure to the carbon-containing atmosphere. The wire materials, such as tungsten and tantalum, change their structure as a result of carburization and hydrogen embrittlement. Cracks are formed, and the lifetime is limited to about 20 h, depending on the operating temperature [132]. Rhenium filaments allow carbon to diffuse rapidly through them without forming carbides. Carburized rhenium is ductile [133] and therefore has a longer lifetime. Single filaments are often used in laboratory reactors; multifilament reactors are more common for industrial applications. An overview on the hot filament process and scale-up for industrial purposes (see Fig. 19) is given in [134].

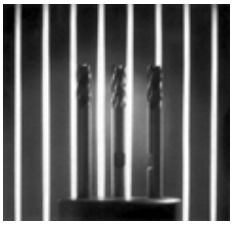

Figure 19. Drill bits in a hot filament reactor [135]

The *oxy-acetylene combustion flame* method was first used in 1988 [136]. The mechanism of diamond growth in combustion CVD is similar to the other diamond deposition systems, even though the composition of acetylene flames differs much from the common CH_4/H_2 mixtures. The combustion flame is divided into three regions: the inner flame, the acetylene feather, and the outer diffusion flame. The substrate is placed in the acetylene-rich feather region. The ratio of acetylene to oxygen is the most important parameter for diamond growth. At values of unity, the acetylene burns completely. Heat and atomic hydrogen are provided by the chemical energy of the oxidation of acetylene. A certain supersaturation of acetylene is essential. The highest diamond quality is obtained in a slightly rich acetylene flame with a supersaturation of up to 15 % (Fig. 20). Combustion CVD is possible at atmospheric pressure and reduced pressure. The lower the pressure, the lower is the growth rate. The growth rate at atmospheric pressure is about 40 µm/h, and at 6 kPa about 4 µm/h. Different types of burners are used. More recently, flat-flame burners with a uniform axial gas velocity have been used to enhance the deposition area while simultaneously improving the film uniformity [137]. Material deposited by combustion CVD has not been tested extensively. However, the quality is similar to that produced by hot filaments or plasma systems. There are not so many applications. Deposition onto steel has been tested with nitrided chromium as buffer layer [138]. Combustion CVD necessitates lower investment costs compared to the plasma CVD processes, but these saving are overridden by the high running costs due to the high quantities of acetylene and its relatively high costs.

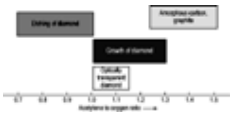

Figure 20. Regions for diamond growth in combustion CVD

[Top of Page]

3. Graphite

Otto Vohler, Gerd Collin, Ferdinand von Sturm, Erhard Wege

Graphite has been used for the production of refractory crucibles since at least the 15th century, and for pencils since about 1550. It was frequently mixed with optically similar minerals such as galena, molybdenite, and even lead (lead pencil). ABRAHAM GOTTLOB WERNER (1750 – 1817) named it after the Greek word " Γ ρ α φ ε ι ν" (to write). Since then, the history of graphite has been closely associated with the development of electricity. Sir HUMPHRY DAVY, in 1800, was probably the first to use carbon electrodes for his studies on the electric arc. The increasing demand for electric arc light carbons since about 1870 gave rise to several processes for the manufacture of artificial carbon (CARRÉ, BRUSH, SIEMENS). At the beginning of the 20th century, important electrolytic and electrothermic processes using carbon or graphite electrodes became available, e.g., aluminum electrolysis (HALL, HÉROULT), alkali–chlorine electrolysis (CASTNER et al.), calcium carbide (WILLSON), electric steel (HÉROULT), and silicon carbide (ACHESON). Progress has been made in customizing the carbon and graphite grades and increasing their usefulness for new and more rigorous applications.

3.1. Single Crystal

The ideal crystal lattice of hexagonal graphite is illustrated in Figure 21. Each carbon atom has three nearest neighbors within the layer plane at a distance of 0.1421 nm and a bonding angle of 120° [140]. Within the layer planes, the bond energy due to the trigonal sp^2 hybridization of the bonding orbitals is far greater than that of the nonhybridized orbitals which form weak π-bonds vertically to the layer planes. The layer or basic planes of the hexagonal modification are stacked in the sequence ABAB..., and those of the rhombohedral modification are in the sequence ABCABC... . Natural graphite normally contains both modifications, in some cases up to 30 % of the rhombohedral modification. The proportion of this modification may be increased by milling and shearing and decreased by thermal annealing. The space groups of the graphite modifications are D^4_{6h} – $P6_3/mmc$ and D^5_{3d} – $R\bar{3}m$, respectively.

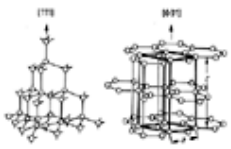

Figure 21. Lattice of the cubic diamond and the hexagonal graphite crystal

The interlayer spacing depends on the degree of disorder and amounts to 0.3354 nm in an ideal crystal, 0.344 nm for crystals with a completely random stacking sequence (turbostratic structure), and even more for less ordered structures. One indicator of the disorder is the width of the X-ray diffraction lines. Most theories explaining the broadening are based on the assumption that aligned layer stacks only correlate in parallel. The broadening is then due to translation and rotation of the layers [141-143].

Other models only accept translations parallel to the layer planes, caused by clusters between the layers and larger vacancies within the layers [144]. These defects yield to deformations of the layers and deviations from the ideal AB sequence, respectively. Due to their layered structure, all vectorial properties are to a large extent controlled by direction. Some properties of single graphite crystals, their dependence on the crystal direction, and for reasons of comparison, some common properties of diamond are shown in Table 3. The anisotropy of thermal expansion is particularly remarkable since the crystal expands in the *c* direction and contracts in the *a* direction at temperatures below 400 °C. Parallel to the basic planes, graphite is a metallic conductor whose resistivity increases with rising temperature. In the *c* direction, graphite behaves like a semiconductor. The thermal conductivity strongly decreases with lower temperature. At room temperature, the conductivity in the *a* direction is better than that of copper. The properties shown in Table 3 strongly depend on the type of samples; therefore the original publications or manuals should be consulted for a deeper understanding [145-149].

Table 3. Properties of graphite and diamond (single crystals)

	Graphite		Diamond*
	***c* direction**	***a* direction**	
Density, g/cm^3	2.266		3.515
Coefficient of linear thermal expansion (15 – 150 °C), 10^{-6} K^{-1}	28.3	– 1.5	0.8
Thermal conductivity [145], W cm^{-1} K^{-1}	0.04 – 0.06	10 – 15	20
Resistivity [146], Ω · cm	1	50×10^{-6}	10^{20}
Magnetic susceptibility [147], 10^{-6} cm^3 g^{-1}	– 21	– 0.3	
Elastic modulus [145], GPa	36	1000	1000
Mohs hardness [148]	9	0.5	10

* The properties of diamond crystals vary considerably

3.2. Natural Graphite

Occurrence and Properties. Numerous hypotheses on the genesis of natural graphite existed, but today most of them are forgotten. It is currently supposed that organic matter was the origin of most natural graphites with the probable exception of the Sri Lanka deposits. Graphite inclusions in plutonic rocks can be traced back to the accumulation of insoluble carbon within the residual melt. The decomposition of inorganic compounds, e.g., carbide, carbonyl, and cyano compounds, is less important for the genesis of graphite deposits.

Crystallinity, refractoriness, and other properties of natural graphite from different deposits may vary enormously due to differences in the precursor materials and the conditions of metamorphosis. In general, natural graphite is a ductile soft mineral which easily cleaves parallel to the basic layer. The flaky grades (plumbago or vein graphite) show a typical metallic luster; the microcrystalline “earthy” grades have a duller black appearance. Table 4 shows a classification [150], [151]. The most common accompanying minerals are quartz, mica, and feldspars; accessory minerals may include pyrite, magnetite, pyrrhotite, and calcite.

Table 4. Classification of natural graphite

	Deposit	Carbon content, %	Average crystallite diameter, mm	Origin
Macrocrystalline flakes	Brazil Federal Republic of Germany (Kropfmühl) China Finland Malagasy Republic Norway Zimbabwe Russia South Africa Czechoslovakia	< 60	< 0.1	syngenetic cata- and mesozonal metamorphism of sapropelites

Macrocrystalline lumps	Sri Lanka	< 100	< 0.01	epigenetic, probably pneumatolytic
Mesocrystalline	Austria Czechoslovakia	30 – 90	< 0.001	syngenetic metamorphism of sapropelites
Microcrystalline	China Italy Korea Mexico Austria	30 – 90	< 0.001	syngenetic, epizonal metamorphism of coals

Mining and Concentration. Natural graphite is mined in open-pit and underground mines. The crude ore contains several impurities in various quantities. Only some of the Sri Lanka deposits contain types with carbon contents up to 100 % [152].

The crude ore is separated in some places by hand. Crude ores with a carbon content sufficiently high for industrial use are only crushed, dried, graded, and if necessary, milled. Flake graphites must be enriched, preferably by using flotation processes. Since graphite flakes float easily, this process is successfully used even for low-grade ores. A modern flotation plant is illustrated in Figure 22.

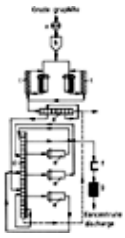

Figure 22. Flow sheet of a flotation plant (Kropfmühl, Germany) [151]

a) Crusher; b) Silo; c) Rod mill and bowl classifier; d) Flotation cells; e) Grinder; f) Centrifuge; g) Drier

The separation of microcrystalline crude ore is more complicated. Since graphite and mineral impurities are intensively intermingled the separation is accomplished only after the crude has been finely ground. During flotation, the graphite-containing foam is possibly polluted by hurled mineral particles. Only if grain size, flotation agents, and other process parameters completely correspond to the properties of the crude ore is enrichment possible. It is reported that by using tar oil, pine oil, or float oil as a flotation agent and a water glass as a trigger for the nongraphitic components in a process including four flotation steps, microcrystalline graphite with more than 92 % carbon has been extracted [153]. Other proven flotation agents are petroleum, paraffinic oils, cresole, creosote oil, and aromatic amines.

Carbon enrichment by flotation yields up to about 96 % C for flake graphites. The mineral components of graphite may be further reduced if the graphite is purified by treatment with hydrofluoric acid, smelted sodium carbonate, or chlorine and fluorine at temperatures up to 3000 °C. The ash content can be reduced by this processes to less than 50 mg/kg [154], [155].

Uses. Natural graphites vary widely in flakiness, refractoriness, toughness, durability, bulk density, etc. and meet the requirements of various processing industries: 24 % are used in foundries, 24 % in the steel industry, 15 % for the production of crucibles, 12 % in the electrical field, 8 % in the refractory industry, 5 % for lubricants, 4 % for pencils, and 3.5 % for batteries [156].

Graphite crucibles are used in foundries for melting steel, nonferrous metals, and precious metals. The most suitable are clay-bound crucibles containing flake graphite grades with a carbon content above 85 %. Bricks for open-hearth furnaces [157], ladle plugs, and tape spouts are other examples for the use of natural graphite in metallurgy. When compared with other refractory materials, graphite-containing grades have generally better thermal conductivity, thermal shock resistance, and resistance against corrosion. Clay-bound graphite crucibles have been widely substituted by pitch-bound silicon carbide–graphite crucibles.

In the production of dry cells and storage cells, graphite powder is mixed with the active mass to increase the electrical conductivity. The conductivity of natural graphite powder is also used in plastics (against static charge buildup), explosives (against ignition by friction), and conductive lacquers.

Natural graphite is used to a large extent as a lubricant. Fine-grained graphite powders often having a particle size less than 1 µm are suspended in liquid carrier mediums, such as oils, greases, alcohols, or water. Most self-lubricating bearings and packings contain natural graphite as a lubricant; pencils function due to the gliding and lubrication of their graphite component. Other applications that use these properties are the smoothing of the surface of molds to facilitate the removal of castings after cooling, blackwashes, and additives decreasing the friction during the compaction of powders.

Natural graphite is also used for carburization, deoxidation, and the covering of metal melts. Plastics, ceramics, porcelain, and concrete are colored with graphite, particularly with microcrystalline grades, and it also serves as a pigment in temperature-resistant paints. Well-crystallized flake graphites are a precursor for graphite intercalation compounds,

expanded graphite, and graphite foils (see Sections Graphite Compounds and Expanded Graphite and Graphite Foil Derived from Intercalation Compounds).

Purified flake graphite has been proposed as a moderator and reflector material for nuclear fission reactors [158]. To some extent, natural graphite is used for the production of fuel elements for high-temperature nuclear reactors [159].

The world production of natural graphite amounts to about 500×10^3 t [156]. The most important countries mining graphite are the following:

flaky graphite grades: Former Soviet Union, China, Malegasy, Sri Lanka, Germany, and Norway
earthy graphite grades: Korea, Mexico, and Austria.

3.3. Artificial Graphite

In general, artificial graphite is produced by heating carbon compounds or carbon powder to 2500 – 3000 °C [160], [161]. This treatment increases the structural order, depending on the structure of the precursor materials. One distinguishes between soft (or graphitizing) carbons and hard (or nongraphitizing) carbons by the fact that the preordering of the first type is distinctly better than that of the last. Examples of both types are petroleum coke and charcoal, respectively. Usually, any carbon material heated above ca. 2500 °C is designated as an artificial graphite, irrespective of its real crystal structure. The thermal graphitizing process can be accelerated by catalysts. Although catalytic graphitization had already been studied by ACHESON, it is of little practical significance. Other methods for the production of artificial graphite are the vapor deposition of volatile carbon compounds or the condensation of carbon vapor (pyrographite) [162], crystallization of carbon melts at pressures above 15 MPa [163], precipitation from oversaturated metal melts (kish graphite) [164], the decomposition of carbide crystals and nitrogenation of calcium carbide [165], as well as the electrolytic decomposition of carbonate melts [166]. All of these processes are economically less significant than thermal graphitization.

All carbon-containing materials are, in principle, suited as a raw material for the production of carbon and graphite. In the first step, the raw material is decomposed by pyrolysis, and in the second step, the intermediate carbon product can be graphitized by further heat treatment. During pyrolysis, the intermediate material may be solid, liquid, or gaseous; it is obvious that the mobility of the molecules determines the degree of microstructural ordering.

It is, therefore, customary to distinguish between solid-, liquid-, and gas-phase pyrolysis, although transitions between the pure reaction types exist. Examples of the *solid-phase pyrolysis* is chars and vitreous or glassy carbon which is produced by pyrolysis of cross-linked polymers. Due to the structural disorder and the small average crystal size of these polymers, the ability of these materials to be graphitized is low. The precursor material of carbon fibers, e.g., thermally stabilized polyacrylonitrile or pitch, also does not melt during pyrolysis, but an ordered microstructure is achieved by stretching the fibrous material below the pyrolysis temperature and preserving this degree of order by the above-mentioned stabilization treatment [167].

An example of *liquid-phase pyrolysis* is petroleum coke, which is the most important raw material for the manufacture of carbon and graphite products. BROOKS and TAYLOR first discovered the formation of optically anisotropic spherulites (mesophase) during the pyrolysis of petroleum pitch, coal tar pitch, and softening coal [168]. During heat treatment, pitch gradually forms an isotropic melt and develops small spherulites with increasing temperatures, probably by a process of homogeneous nucleation, the diameter of which increases with temperature and to a lesser extent with residence time. The spherulites are nematic liquid crystals composed of large lamellar aromatic molecules. The isotropic phase is finally exhausted, and the spherulites touch each other and coalesce. Formation and structure of mesophase are essentially governed by three limiting conditions:

1. the shape of the primary molecules must favor the formation of liquid crystals
2. the fluidity of the system
3. the reaction or condensation rate must be smaller than the ordering rate

Other factors influencing the structure of the mesophase which eminently affects structure and properties of coke materials are the amount of nonmelting constituents and/or impurities [169].

The chemistry of the pyrolytic conversion of hydrocarbon compounds to solid carbon involves numerous chemical reactions. Generalized schemes have been developed by studying simpler models [170], [171]. The major reactions involved are:

bond cleavage and formation of free radicals
molecular rearrangement
thermal polymerization
aromatic condensation
elimination of side-chains and hydrogen

A general scheme is shown in Figure 23 with structural models of the reaction products. Due to the formation of volatile products during pyrolysis and structural transformation, pores and shrinkage cracks develop on solid carbon. Carbon must be heated to ca. 3000 °C to be transformed into graphite. The degree of graphitization, which is an indication for a perfect graphite crystal, the spatial distribution of the crystallites, and the amount and distribution of porosity determine the properties of the final product. Lattice disorder as well as porosity can be reduced by high-pressure treatment at temperatures above 2500 °C [172].

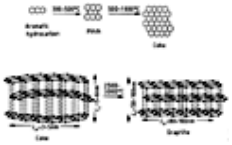

Figure 23. Reaction scheme for carbonization and graphitization

Gas-phase pyrolysis is characterized by thermal cracking of gaseous hydrocarbon compounds and deposition of carbon on a substrate. Examples are pyrocarbon (see Section Pyrocarbon and Pyrographite), carbon films, and carbon black. Stress annealing of pyrocarbon at high temperature yields nearly perfect graphite crystals.

[Top of Page]

4. Carbon and Graphite Materials

Gerd Collin, Ferdinand von Sturm, Erhard Wege, Wilhelm Frohs

Most artificial carbon and graphite materials were developed in response to the growing demand for high-temperature electrodes and corrosion resistant materials.

DAVY employed carbon electrodes cut from charcoal for his electric arc studies (1801). FOUCAULT prepared electrodes from retort carbon which is deposited on the upper parts of the retorts during the production of coke and water gas (1844). Since these products were far from uniform the next step was to improve uniformity by applying powdered carbon as a raw material, e.g., caking coals, powdered coke, retort carbon, and lampblack (SCHOENBEIN, BUNSEN 1840 – 1842, STAITE, EDWARDS 1846, LEMOLT 1849, CARRÉ 1867). The powders needed to be bound with a carbonizable binder, e.g., concentrated sugar solutions or tar. In 1876, BRUSH introduced petroleum coke, which is now the most important raw material due to its uniformity and graphitizability.

DESPRETZ (1849), BERTHELOT (1870), and MOISSAN (1890) discovered the transformation of carbon into graphite when carbon is heated in an electric arc. GIRARD and STREET (1893) attempted to graphitize carbon rods by using this treatment.

The first useful processes for graphitization based on direct resistance heating of the artifacts were proposed by CASTNER (1895) and ACHESON (1895).

Carbon and graphite materials are characterized by their thermal and electrical conductivity, which resembles that of metals, and their thermal and chemical stability, which is similar to that of refractory materials. Their preferred applications are, therefore, high-temperature electrodes and conductors used in the presence of corrosive media and refractories with considerable thermal conductivity. When raw materials and process variables are changed, carbon and graphite products can be adjusted within a broad range to meet all usage requirements.

4.1. Carbon and Graphite Artifacts

4.1.1. *Manufacture*

Most methods for the manufacture of carbon artifacts consist of a two-stage process. In the first step, raw materials, especially coke, are produced by pyrolysis of carbon-containing substances, e.g., pitches and coals. In the second step, the unmeltable coke, together with a binder, is processed to carbon in a manner similar to the manufacture of ceramic materials. The carbon can be transformed into graphite by subsequent heat treatment.

Petroleum coke, pitch coke, metallurgical coke, anthracite, and carbon black are commonly used as raw materials. Coal tar pitches are the most commonly used binder material; less common are petroleum pitches and resins. Mixtures of these materials are shaped by extrusion, vibration molding, die molding, or isostatic molding and are baked at temperatures of ca. 1000 °C in an inert or reducing atmosphere and may be graphitized by further heating to 3000 °C. The manufacturing process is shown in Figure 24.

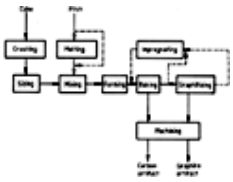

Figure 24. Carbon and graphite manufacturing, flow diagram

Raw Materials. The range of possible raw materials extends from well-ordered *natural graphites* and *needle cokes* with highly preordered two-dimensional layer stacks to less ordered materials such as *chars* and certain types of *carbon blacks.* The properties of carbon and graphite products and their performance depend to a large extent on the properties of the raw material. Therefore, the selection of raw materials is a very important step in the manufacture of carbon and graphite.

Petroleum coke, the most frequently used raw material, is produced from residues of atmospheric- and vacuum-distilled crude oil, from catalytically or thermally cracked fractions, and residues from the production of ethylene [173]. The principal method is the delayed coking process (⟶ Petroleum Coke). Some of this delayed coke is high in sulfur and metallic impurities and, therefore, is unsuited for the production of high-quality carbon grades [174]. Due to the increasing requirements for high-power graphite electrodes, the demand for well-ordered needle coke with an acicular or needle-like microstructure has increased. Needle coke precursors are the hydrodesulfurized more aromatic refinery streams [175] from

which the asphaltene components have been removed by cracking and distillation, e.g., decant oils derived from catalytic cracking processes [176], [177]. The shortage of suitable crude oils for the production of needle coke gave rise to more flexible processes with regard to feedstocks [178]. Coal tar pitches and coal extracts have also proven to be precursor materials for needle cokes [179], [180]. Graphite artifacts containing needle coke are characterized by outstanding thermal shock resistance and low resistivity.

Normal feedstock gives less anisotropic petroleum coke, which is used in large quantities for the production of carbon anodes [174]. Isotropic coke structures are obtained from feedstocks containing fast-reacting components [181]. Examples of isotropic coke types are gilsonite and fluid coke, a residue of a fluidized-bed coking process [182]. Isotropic cokes are of interest for nuclear reactor graphite grades.

Pitch coke, which is produced by coking of heat-treated or blown coal tar pitch, is less ordered than average petroleum cokes and has greater hardness and strength. Coking takes place in a chamber furnace or, more recently, in a delayed coker [183] for the production of needle coke from coal tar pitch.

Carbon black and *lampblack* are used predominantly as additives to increase bulk density, resistivity, and polishability of the carbon or graphite material. Brush grades may contain blacks exclusively. In such cases, the black is processed with a binder to an intermediate product called carbon black coke [184]. Additives for reducing electrical resistivity are artificial and natural graphite.

Metallurgical coke or *anthracite* are useful for the production of refractory linings (blast furnaces) and cathodes for aluminum production. Metallurgical coke is also used as a packing material in baking and in graphitization furnaces.

Calcination. Most raw materials, especially delayed-coking petroleum coke, must be calcined to decrease the volatile matter which forms during the baking process of the carbon material and causes severe shrinkage. During calcination, water and volatiles are removed by dehydrogenation and dealkylation; the coke attains its final porosity and its density is increased. The coke quality is determined by the maximum calcination temperature, calcination time, and the gaseous atmosphere during processing [185]. Even the thermal expansion of the coke is influenced by the calcination conditions, as a unique system of microcracks is developed which absorb, to some extent, the expansion of the crystallites; that results in a smaller expansion of the coke grains [186]. Maximum temperature of calcination is ca. 1200 – 1400 °C for petroleum coke and even higher for anthracite (electrocalcination).

Calcination is generally performed in rotary kiln furnaces consisting of a slightly inclined tube 2 – 3 m in diameter and 50 – 70 m in length. The residence time of the coke is ca. 2 h; the evolved volatiles are burned to deliver the energy required to raise the temperature of the coke and kiln to the calcination temperature. Within the last few years, the thermal efficiency has been improved so that additional energy is no longer needed. The kiln contains a cooling zone in which the calcined coke is cooled before discharging.

Rotary hearth furnaces are another furnace type used for coke calcination [187]. According to this process, the coke is heated on a disk that rotates slowly around a vertical axle. Vertical retort systems are of less significance.

If higher calcination temperatures are required, gas-fired furnaces may be used, but are less effective than electric calciners. These furnaces generate heat by resistance heating, which heats the carbon material to temperatures above 1600 °C. Most notably, anthracite is calcined by this process. A similar process has been proposed for the desulfurization of petroleum coke [188]. Properties of cokes are shown in Table 5.

Table 5. Properties of filler materials

	Needle coke	Petroleum coke (regular)	Pitch coke	Metallurgical coke
Microstructure	coarse fibrous	mosaic	coarse mosaic	heterogeneous, fine isotropic
Density (calcined at 1300 °C), g/cm^3	2.10 – 2.15	2.11 – 2.13	2.07 – 2.11	1.9 – 2.0
Ash content, %	0.05 – 0.3	0.1 – 0.3	0.2 – 0.4	7 – 11
Sulfur content, %	0.2 – 0.9	0.6 – 1.8	<0.6	<1.0
Coefficient of linear thermal expansion (measured with graphite rods, 20 – 200 °C), 10^{-6} K^{-1}	< 0.8	1.0 – 2.0	2.6 – 3.0	
Coefficient of volume thermal expansion, 10^{-6} K^{-1}	3 – 5	6 – 10	8 – 14	9 – 11
Anisotropy of thermal expansion	>2	1.5	1.4	1.2

Binder Materials. The purpose of binder materials is to wet the surface of carbon grains, give mixtures of high plasticity when mixed with grains, and give high-yield carbon residue after pyrolysis. Coal tar pitches, the residues from the distillation of crude tar produced while coking bituminous coal, are most suited. The distillation of tar is generally performed continuously in atmospheric columns and under vacuum. In special cases the pitch residue can be heat treated at about 400 °C to increase the content of quinoline-insoluble and toluene-insoluble matter, which influences the coking value and binding behavior [189]. The treatment is limited by the rising softening point. Characteristic properties of some pitch grades are illustrated in Table 6 (see ⟶ Tar and Pitch). Pitch grades with higher softening points are preferably used as binders for the production of molded graphite grades. The softening point represents a certain viscosity, but gives no information on the slope of the viscosity – temperature curve, an important factor for mixing and forming [190]. Since an increasing viscosity hampers flowability, the softening point of pitches used as impregnating agents should be lower than that of binder pitches. Another characteristic of pitches is their distillation curve, which should change only gradually. Petroleum pitches processed by dealkylation and dehydrogenation, with subsequent condensation and polymerization of low-boiling decant oils and aromatic extracts, have also been proposed as a binder and impregnating material [191]. Due to the lack of tar in the United States, petroleum pitches are widely used for impregnation.

Table 6. Properties of pitch binders and impregnants

	Binders for carbon anodes		Binder	Impregnating agent
	Prebaked	Söderberg	for graphite electrodes	
Softening point (Mettler), °C	110 – 130	110 – 130	100 – 115	65 – 90
Density, g/cm^3	1.32	1.32	1.32	1.25
Quinoline insolubles, %	6 – 10	6 – 10	6 – 10	< 5
Toluene insolubles, %	20 – 30	20 – 30	20 – 30	10 – 20
Coking value (Alcan), %	> 60	> 60	55 – 60	40 – 50

The coking value of pitch binders can be increased by adding condensation agents, e.g., sulfur, Lewis acids, etc. [192], [193]. To improve the flowability of the mixtures, lubricants are added that are insoluble in pitch such as oil, stearic acids, and metal soaps [194].

Thermosetting resins are also reliable binders and impregnating agents, although their use is limited to the production of smaller carbon and graphite shapes because of higher costs [195]. Phenolic and furan resins are frequently used.

Grinding and Sizing. The bulk density of carbon and graphite shapes depends on the porosity of the coke grains and the packing density of these grains [196]. These variables are controlled by crushing and grinding the coke and sizing the milled product. Large grains are generally used for the production of large shapes and smaller ones for small bodies. The maximum grain size is ca. 15 – 20 mm for large electrodes and less than 20 µm for spectroscopic carbon grades. The microstructure of the artifacts improves with decreasing grain size.

Unfortunately, the binder requirement increases with decreasing grain size, and higher binder contents imply problems during baking. This problem can be solved with a multistep process using pellets or agglomerates which are baked, partially ground, and again mixed with a binder whose proportion can then be relatively small. This process is also advantageous for controlling the anisotropy of the artifact, which is primarily determined by the shape factors of the coke grains [197],[198].

The calcined coke discharged from a silo is ground and sized in a highly automated process using roll crushers, roller pulverizers, as well as ball mills and jet mills. The product is sized by vibration screens, jarring screens, or separators. Pitch binders are usually melted, but the pitch may also be ground in swing-hammer mills.

Mixing. Mixing is generally performed at temperatures at least 50 K above the softening temperature of the pitch binder. The proportion of the filler material and binder determines the fluidity of the mixture, which also depends on the grain size of the filler material and the forming procedure. The average binder content is about 15 – 30 %. A green mixture consists primarily of filler, binder, and air bubbles [199]. During the mixing process, the bubbles must be removed and the binder should at least partly penetrate the pores of the coke particles. High mixing temperatures accelerate the spreading of the binder films on the solid surface [200].

The different filler materials stored in separate bins are discharged, weighed, and fed in the proper proportion to the mixer [201]. Before the filler is blended with binder, it is customary to homogenize the filler components and heat the filler to a temperature above the softening point of the pitch binder.

Kneaders with rotating sigma blades, ribbon blenders, and jet mixers, which are heated electrically, with natural gas, steam, or thermal oil, are customary. Mixing time is 20 – 60 min and the capacity of these mixers is 800 – 2000 L. Traditional sigma-blade mixers have been partially replaced by fast-running intensive mixer systems that allow mixing times of less than 15 min [202]. Continuous kneading extruders are used especially for the production of carbon anodes [203]. The homogeneity of the mix can be improved by using a vacuum or increased pressure [204]. Trapped air is removed and the impregnation of the porous coke particles increased. In another process, the filler is dispersed in a solution of binder material which is precipitated through addition of a further liquid in which the binder is insoluble and which is miscible in the solvent [205]. This process coats the particles with a uniform binder film.

If necessary, the lubricants are added to the mixture shortly before completing the mixing process. The mixture must be

cooled to the forming temperature in either hopper or barrel mixers, cooling drums, or conveyers. Mixing a liquid coolant, e.g., water, into the green mix and vaporizing the coolant results in a more uniform temperature distribution [206], [207]. Since this process can be carried out in a closed system, tar fumes are not emitted.

Forming. Due to the high production rate, elongated shapes are usually formed by *extrusion.* The cross-section may be circular, rectangular, or more complicated, e.g., for finned tubes. The cross-section ranges from a few square millimeters (battery carbon, spectroscopic carbon) to nearly 1 m^2 (electric furnace electrode). The hot carbon mixture is charged into the cylindrical chamber (mud cylinder) of a conventional hydraulic press [208], evacuated, and compacted. The mixture is forced through a forming die by an advancing ramhead or piston. The mixture may also be fed into the die by one or two augers. The mud chamber is up to ca. 1.5 m in diameter and up to 3 m in length, and the pressing force up to 60 MN. Auger extrusion allows a continuous extrusion process, but the extruded form often has a spiral internal structure. Although this structure may be destroyed by special gratings in the forming die to reorient the mix flow, this process is used only to form shapes of a small diameter [209]. Proposals for continuous piston extrusion have not yet been realized on a large scale [210].

The flow profile of the mixture also affects the structure of the forms including pore structure, anisotropy, and crack formation [211]. Even more important is the shape of the coke grains, which are aligned parallel to the flow lines. The structure and anisotropy can be controlled to a certain degree by changing the converging section of the die between the mud cylinder and the forming die [212]. The wall friction may be reduced by forcing lubricants through small openings in the wall [213]. For the production of tubes and other hollow bodies, a mandrel is positioned in the forming die.

The extruded form must be cut into sections of appropriate length. A cutting device which moves parallel and more or less synchronously with the formed body is used [214]. To avoid deformation, the bodies are cooled from the extrusion temperature (20 – 30 K above the softening point of the binder) to room temperature by spraying or dipping into water [215].

Molding. Smaller bodies, as well as some of the anode blocks used in the electrolysis of aluminum, are frequently produced by molding. Due to the smaller amount of binder, the shrinkage of these bodies can be better controlled and forming to the final size is possible (press-to-size). Pressures as high as 200 MPa (2 kbar) are necessary for compacting. The relationship of height to cross-section of the body is limited by the wall friction, which can at least be partly compensated by pressing at a temperature above the softening point of the binder or by adding solid or liquid lubricants. Crank, eccentric, and hydraulic presses are used, with the force of the press plunger reaching 140 MN.

As a consequence of the lower impact of flowability, the composition of the mix can be varied to a larger extent than in the extrusion process. Molding is, therefore, particularly well suited for the production of carbon and graphite bodies for such sensitive electrical and mechanical applications as brushes, electrodes for electric discharge machines, bearings, etc. For the production of larger bodies, this process must compete with isostatic pressing and especially vibration molding.

Isostatic Molding. In contrast to other forming methods, the compacting forces are transmitted by a fluid acting equally from all directions. Accordingly, the density gradient is smaller and the uniformity of the pressed form is better. The dies (bags) consist of an elastomeric material, e.g., silicon. They are usually evacuated during or after being filled with the pressing powder and are attached to the pressure tank (dry bag) or are freely removable (wet bag). The pressure tank is up to 1 m in diameter and more than 2.5 m in height. The pressure is up to 200 MPa (2 kbar) [216]. When the powder is being compacted, all anisometric particles are aligned perpendicular to the direction of force, yielding a quasiisotropic body. This is most advantageous for nuclear graphites, particularly spherical graphite forms containing nuclear fuel materials [217]. Compared to other forming methods, the overall anisotropy of the shape is small even when the pressing powder contains anisometric particles such as natural graphite or needle coke [218], [219].

Vibration Molding. During the last decades, the use of this method has steadily increased, particularly for the production of carbon electrodes used for aluminum electrolysis and the production of other large equipment containing coarsely grained carbon. A compacting machine consists of a vibrator and a mold which are connected. The damping capacity of viscous carbon mixtures is higher than that of dry granular materials. To achieve sufficient compaction, the pressing mixture must be loaded with top weights [220] and then evacuated [221]. The vibration frequency is ca. 1000 – 2000 cycles per minute, which may be changed eventually to compensate changing damping capacity during compaction. One advantage of this forming method is its great variability, which makes it possible to compact mixtures containing small as well as large amounts of binder. Even large cylindrical bodies can be formed [222]. A continuous compaction method for these bodies has been proposed. Molding vibrators are completely automated by using molds mounted on turning or pushing tables [223]. Jar molding, a vibration molding method with a low frequency but large amplitude, is used to form large-diameter carbon electrodes.

Other Forming Methods. Other methods are less common. Carbon shapes with a high bulk density can be produced by explosion impact [224] or by implosion into a mold [225]. Spin molding is used especially for mixtures containing clay for the production of graphite crucibles. A mold containing the mixture is rotated at a speed of several hundred cycles per minute while a plunger is moving axially into the mixture [226]. Tamping, a vibration molding with a low frequency, but large amplitude, is almost exclusively used for the compaction of carbon-containing cements used for joining carbon cathodes and carbon blocks.

Baking. In the baking process, the binder material of the shaped artifact is pyrolyzed, and a binder coke is formed which joins together any coke particles. Volatile products are generated which must be removed from the carbon form. The process consists of the following steps: (1) the heat transfer from the outside into the form, (2) the pyrolysis reactions, and (3) the removal of volatile products from the form [227]. Heat transfer and the diffusion of the volatile products developed during baking are influenced by layers of packing material — mostly granular metallurgical coke — needed to prevent deformation of the bodies at temperatures between the softening point of the binder and binder coke formation. The packing material also protects the bodies against oxidation. In addition to the mass loss, the bodies undergo complicated volume changes during baking which depend on the type and amount of the binder, the heating rate, and the volume of the body [228] (Fig. 25).

These changes, as well as pore formation, require certain heating programs to avoid crack formation which is detrimental to the quality of the product. The increase in temperature within the body should ensure that the transfer capacity of volatile pyrolysis products is larger than or equal to the quantity of products formed at each temperature interval. Since the formation rate of volatiles is largest in the interval between 300 and 500 °C, the heating rate in this range is as low as 0.5 – 4 K/h. At higher temperatures, considerably higher heating rates are tolerated. The final temperature lies between 750 and 1300 °C, depending on the further treatment of the form.

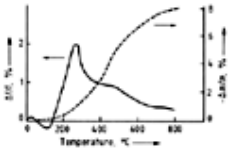

Figure 25. Variation in length and mass of pitch-bound carbon shapes during baking

At normal baking conditions, about 30 – 50 wt % of the original binder is lost as volatile products. This loss can be decreased by adding Lewis acids, sulfur, and other condensating substances to the binder [229], or by performing the heat treatment under pressure [230], [231]. Pressure baking not only increases the coking residue, but also cuts production time from weeks to minutes, thus enabling the automation of carbon production to a degree not attainable with other baking processes.

In general, all furnaces are suited for baking carbon bodies, provided the temperature is adjustable. Each furnace must have devices for collecting and purifying the volatile pyrolysis products, e.g., electrostatic precipitators or incinerators [232]. Ring furnaces, in which several chamber furnaces are coupled, are most frequently used for baking carbon bodies. For the baking of sensitive grades, e.g., large electrodes, single-chamber furnaces are customary. Tunnel kilns are only used for rebaking impregnated articles.

Ring Furnace. Ring furnaces consist of ca. 10 – 30 identical brick chambers or pits that are connected to each other to form an endless chain of individual units. A distinction must be made between open-top and closed chambers. In both types, several process steps occur simultaneously: one section of the furnace is loaded with green forms, in a second section the green forms are preheated by combustion or waste gases streaming in through flues located beside the chambers, in a third section the chambers are fired by gas or oil burners, in a fourth section the chambers are cooled by air streaming to the burners through flues, and in a fifth section the cooled baked forms are unloaded (Fig. 26). After a defined period of time, the burners are shifted to the neighboring chamber in the direction of the gas flow, the loaded chamber is coupled to the combustion gas stream, and the outermost chamber is unloaded. Ring-type furnaces, in which the heating section rotates stepwise through the furnace, are counterflow heat exchangers, which allow better fuel economy than other furnaces. To improve heat exchange, the flues contain baffles. In the closed-type ring furnace, the chambers are divided into subsections by hollow brick walls acting as flow channels. The temperature distribution within the chambers depends on a uniform gas velocity [233], [234]. The temperature adjustment is complicated since the burner and the suction cannot be individually adjusted. A more rapid cooling can be achieved without affecting the entire furnace by spraying water or vapor into one or more of the chambers [235], [236].

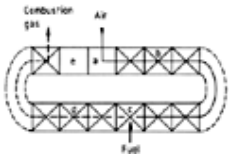

Figure 26. Closed ring furnace (schematically)

a) Unloading chamber; b) Cooling section; c) Heating chamber; d) Preheating section; e) Loading chamber

The baking cycle is relatively short if identical shapes are baked and thin layers of packing materials are used. These conditions are fulfilled by carbon anodes, for which a firing curve is shown in Figure 27. The capacity of large ring furnaces amounts to about 50×10^3 t/a.

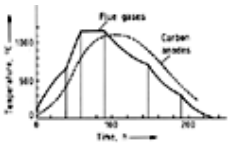

Figure 27. Ring furnace, heating cycle [236]

Single-Chamber Furnace. Several different types, usually containing a square chamber, are in use. To improve gas circulation and heat transfer, some types are equipped with fire or flue pipes or with channeled walls [237]. Some furnaces are open, but most types are covered by arched covers. The green bodies are often transferred to steel saggars, which are distributed in the chamber according to a given pattern. The thickness of the packing mass layer between the saggar wall and the green form can be kept small, thus improving heat transfer and the adjustability of the system. This enables the baking of sensitive carbon grades.

The efficiency of single-chamber furnaces can be improved by using a chamber bottom which can be moved out of the furnace for loading and unloading (car bottom furnace, bulk furnace) [238]. For greater energy efficiency, two or more furnaces act together for recovering the heat of the combustion gas.

Tunnel Kiln. Tunnel kilns are used only for baking small carbon bodies and for rebaking pitch-impregnated carbon artifacts, although handling is simpler and the fuel efficiency obviously better. On the other hand, special precautions must be taken to

avoid condensation of the volatile pyrolysis products (especially tars) and to get a stable flow system from the flue gases. Therefore, the tunnel must be divided into several subsections [239].

Other Furnaces. Electric resistance heating (resistor baking) at high pressure remains of limited use [230]. The short baking cycle of a few minutes duration at pressures of ca. 500 MPa (5 kbar) is increased to several hours at pressures of ca. 0.5 MPa (5 bar) [240]. To overcome the considerable resistance of the green form, it is also possible to bake with convection heat and continue with resistance heating after some pyrolysis [241].

After the baking process, the adhering packing mass is removed from the baked shapes by cleaning machines.

Graphitization. Graphitization is the transformation of somewhat disordered carbon material into the well-ordered graphite structure; the heat treatment of carbon shapes at temperatures above ca. 2600 °C is referred to by this term. During this heat treatment, the average interlayer distance decreases and the height and width (L_a and L_c) of the ordered areas increase [242-244], whereby the attainable order or degree of graphitization depends on the structure of the precursor material. Structural ordering is a relatively rapid process (Fig. 28), while the rate of volume shrinkage decreases with increased heating [246]. All volume properties, such as bulk density, resistivity, strength, etc., depend not only on the maximum temperature, but also on the heating rate. The heating rate of larger specimens may be limited by insufficient thermal shock resistance [247]. The actual rate is shown in Figure 29. During heat treatment, the impurities within the artifacts react with the carbon to form carbides, which mostly decompose and vaporize at higher graphitizing temperatures.

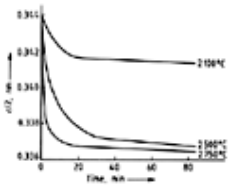

Figure 28. Changing of interlayer spacing during graphitization [245]

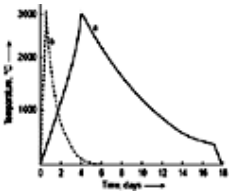

Figure 29. Temperature cycles of the Acheson furnace (a) and Castner furnace (b)

At temperatures of ca. 1600 – 1700 °C, the sulfur contained in the carbon bodies is promptly emitted, thereby increasing its volume (puffing) [228]. To avoid the formation of cracks, puffing must be reduced by adding inhibitors such as powdered iron, iron oxide, or compounds of alkali and alkaline-earth metals which delay sulfur emission. Graphitization may also be performed at lower temperatures by adding carbide-forming elements to the green mixture, e.g., boron, silicon, iron, nickel, or vanadium; this is called *catalytic graphitization.* The different solubility of ordered and nonordered carbon in the carbide phase is the driving force [248], [249]. In practice, this process is discredited by the remaining carbide impurities and the formation of larger pores due to carbide decomposition.

The grade and rate of graphitization may be increased by viscous deformation of the carbon artifacts at temperatures above 2500 °C, known as “hot-worked graphite” [250]. Although the properties of these artifacts are promising, the process has not been introduced on a larger scale due to the great expense.

Graphitization is performed almost exclusively with Acheson and Castner furnaces. For the production of graphite electrodes, Acheson furnaces have been increasingly replaced by Castner furaces. Inductive and radiation heating is used only for the production of specialities.

Acheson Furnace. The furnace invented by E. G. ACHESON [161] consists of a horizontal bed of refractory materials, furnace heads carrying the current load to the furnace, and long sidewalls made of concrete blocks, steel sheets, or steel grids. The bottom is cooled by air and loaded with a layer of insulating materials, e.g., granular silicon carbide, metallurgical coke, sand, or sawdust. The same materials are used for insulating the sidewalls and the surface of the furnace. Carbon black is used as an insulating material, especially if a lower ash content of the graphite product is desired (thermal purification). Carbon bodies that are to be graphitized are stacked between the furnace heads, usually horizontally and transversely to the flow of the electric current. The bodies are separated from each other and surrounded by a granular resistor medium, e.g., coke – graphite mixtures. Since the resistivity of the furnace depends essentially on the distance between the bodies and the packing density of the resistor material, the furnace must be carefully loaded. The capacity of Acheson furnaces is max. about 200 t.

The furnace heads consist of solid masonry with water-cooled graphite electrodes, which are connected by bus bars or shunts to transformers or rectifier transformers. The electric load is 0.1 – 1 MW/t. Because the resistance of the furnace decreases with the progressing graphitization, the voltage, usually between 50 and 250 V, must be reducible. The maximum current is ca. 100 kA. The power input is usually controlled by an empirical program since direct temperature measurement, especially at higher temperatures, is less precise [251]. Figure 29 illustrates the temperature changes during the graphitization process. The energy requirement of large furnaces is 3 – 4 kWh per kg of graphite material.

Acheson furnaces are firm constructions well-suited for the graphitization of all possible shapes and grades. Less satisfying

are the high energy demand, the relatively small production rate per square meter, the inhomogeneity of the product, and the difficulty of emission control.

Castner Furnace. This type of furnace, which was proposed by H. Y. CASTNER [160], does not contain granular resistor materials. One or more carbon artifacts are clamped between head electrodes, these forming the path for the heating current (lengthwise or longitudinal graphitization). One or both ends of the column are subjected to a mechanical or hydraulic pressure device to maintain a tight connection during the heating cycle. Insulation is similar to that of the Acheson furnace. Much higher heating rates are possible, and heat losses are accordingly smaller. The heating time can be less than 10 h, and the energy requirement is about 2 – 3 kWh per kg of graphite material [252]. In this process, the contact resistivity between the head electrodes and the abutting artifacts must be kept low by compensating for length changes of the column due to thermal expansion, puffing, and graphitization shrinkage. It is customary to insert sheets of a plastic temperature-resistant material between the abutting faces and push the faces against each other. Due to their small size, Castner furnaces are well-suited as mobile units on tracks and for moving to separate loading and unloading stations [253].

In addition to the lower energy requirement, the homogeneity of the product is also an impressive feature of this process. Less favorable is its limitation to certain sizes and grades.

Induction Furnace. Small induction furnaces have been used for several years for the graphitization of smaller carbon bodies [254]. The graphitization of large artifacts up to 2 m in diameter has also been studied. Induction furnaces make it possible to perform continuous graphitization. Although the electrical load of the induction coil is limited to about 2 MW, an adequate production rate and energy efficiency can be achieved. The frequency that is used depends on the specific electrical resistivity of the artifact and its diameter. The frequency for larger diameters is several hundred and for small diameters several thousand cycles per minute. The diameter of the induction coil must correspond to the diameter of the artifact. This is a considerable drawback of this method. By using separate susceptors or combining several shapes in a certain pattern to one unit, this problem may be eased [255].

Radiation Heating. Furnaces using thermal radiation from graphite heating elements are less sensitive to the size and grade of the carbon artifacts and are also suited for continuous graphitizing processes [256], [257]. The heating rate may be as high as 3000 K/h, and the energy requirement is about 2.5 kWh/kg without heat recovery. Since the temperature of the heating elements is obviously above 3000 °C, there is considerable vaporization of carbon (average loss about 10^{-6} g cm^{-2} s^{-1}). Therefore, the process is limited to smaller sizes.

Purification. Coke and binder materials used for the production of graphite artifacts contain various quantities of impurities. The impurities and contaminants of the artifacts resulting from production must be removed for some uses, e.g., nuclear graphite, spectroscopic carbons, etc. Thermal, thermochemical, and chemical purification processes must be differentiated. The graphitization process is itself a simple thermal process. Its efficiency may be increased by adding getter substances such as carbon black to the insulating material [258]. Gaseous halogens and halogen compounds are blown into the graphitizing furnace for further purification at temperatures above 1000 °C. They form volatile halogen compounds with nonvolatile impurities; this is known as the thermochemical process [259]. In another process, solid halides (e.g., NaF) are distributed in the graphitizing furnace and vaporized during heating [260]. The flue gases are cleaned in scrubbers with caustic soda solutions. Chemical purification processes were originally applied to natural graphite to remove gangue. The treatment of graphite ore with soda or fluorine is an example. Fine-grained natural graphite used in the manufacture of nuclear graphite is treated with alkali metal hydrogen fluorides [261]. The ash content of nuclear graphite grades is as high as several hundred mg/kg and that of spectroscopic carbons as low as 1 mg/kg.

Machining. Graphite is readily machinable. Lubricants or cutting fluids are not necessary. It can be sawed, turned, bored, drilled, planed, shaped, and milled. Surface roughness decreases in this sequence from ca. 25 – 125 to 1 – 6 µm. A fine-grained grade is always smoother due to the smaller pores revealed through cutting. The same machines are generally used as in metalworking. Tool speed and depth of cut are greater, while feed and cutting power are lower (5 – 30 N/mm^2 compared to ca. 2000 N/mm^2 for steel) [262]. The hardness of carbon grades is considerably greater than that of graphite grades, and the tool wear is about 100 times greater. Graphite electrodes for EDM and other complicated shapes are satisfactorily drilled with NC machines [263].

Graphite electrodes are frequently machined on production lines consisting of stations for turning the cylinder surface, as well as for cutting and grinding the threaded sockets and abutting surfaces. The connecting nipples are cut to length, bored, filled with pitch pins, and threaded [264]. Graphite apparatus and equipment are machined to precise specifications and often consist of numerous single sections which must be joined together. This is achieved with curable resin mixtures containing powderous carbon or graphite. Soldered joints are used to join carbon and graphite shapes to electrical conductors.

Impregnation and Surface Coating. About 25 % of the volume of carbon or graphite artifacts consist of pores which are sometimes detrimental to the bulk properties. The pores may be filled at least partially or closed by impregnation with a solid material whereby strength, conductivity, and resistance against oxidation are improved. Pitch is the most frequently used impregnating agent. Porous artifacts are charged in a vacuum autoclave after being preheated well above the softening point of the impregnating pitch. The autoclave is evacuated to remove the air from the pores of the artifact and is filled with molten coal tar pitch or petroleum pitch. The pressure is increased to ca. 1.0 MPa (10 bar). To pyrolyze the incorporated pitch, the impregnated artifacts must be heated to ca. 800 °C, a process known as rebaking. Impregnated artifacts have a higher bulk density and, therefore, improved strength and conductivity. Emission control of vaporous pitch components during impregnation requires special provisions [265]. Artifacts are generally impregnated and rebaked before graphitization. The coke that has been incorporated into the pores becomes porous, and the artifact again becomes permeable a result of pyrolysis. For impermeable artifacts, e.g., for nuclear fuel elements, impregnation must be repeated several times [266].

Graphite grades used at lower temperatures may be impregnated with curable resins, particularly phenol – aldehyde and furan resins. The tolerated application temperature of these grades is at the most ca. 200 °C. Since these resins are resistant to corrosion, the impermeable material is widely used for the production of heat exchangers and other chemical apparatus

[267]. When polytetrafluoroethylene wax is used as an impregnating medium, the application temperature can be raised above 200 °C [268].

The pores of carbon and graphite bodies may also be impregnated with carbon formed by chemical vapor deposition (CVD) processes [269]. The rate of deposit is limited, since otherwise only pyrocarbon would be deposited on the outer surface of the artifact [270]. Carbon and graphite bodies are impregnated for special effects with lubricants (brushes), metals (bearings), and oxidation inhibitors, e.g., phosphates and borates [271].

Coating the surface of carbon and graphite artifacts with refractory materials, e.g., carbides, silicides, borides, or nitrides, also protects against oxidation. Coatings are brushed or sprayed onto the surface, and the deposited layer is sintered, or the coating is applied by flame or plasma spraying, chemical vapor deposition (CVD), or physical vapor deposition (PVD) [272].

Emission Control. The most important emissions during production of carbon and graphite products are coke dust (crushing, machining), coal tar pitch volatiles (mixing, baking, impregnating), and sulfur dioxide (graphitizing). Of primary hygienic concern are the high-molecular-mass polycyclic compounds which are present in the air as aerosols [273]. Particulate emissions are limited to 75 mg/m^3 including 50 mg/m^3 tarry substances at the most [274].

Current Problems. As with other materials, the production of graphite faces three major problems: energy requirements, increased product quality, and environmental protection. The production of 1 t of graphite requires ca. 2.6 t of fuel oil; the heat content of graphite together with auxiliary materials equals that of ca. 1.9 t of fuel oil, amounting to a total of 4.5 t [275]. The production of graphite is, therefore, very sensitive to the oil supply. Efforts must be made to minimize this dependency. Only a fraction of this amount of energy is on disposal (Table 7). The energy requirement could be decreased considerably by continuous baking and graphitizing processes, which are more energy efficient. These processes would also facilitate energy recovery as well as the handling of the material and product [276].

Table 7. Energy requirements (GJ/t)

Baking	
Single-chamber furnace	6 – 15
Ring furnace,	
open	4 – 10
closed	2 – 8
Graphitization	
Acheson furnace	11 – 15
Castner furnace	7 – 11

During the past decades, the demand for carbon and graphite products has increased considerably. Product quality needed to be, and must be still further, improved by applying advanced raw materials and decreasing the variation in processing, e.g., by developing continuous production processes.

4.1.2. *Properties*

Artificial graphite bodies differ more or less from the graphite single crystal. This difference is due to structural disorder, imperfect alignment, and porosity. Other parameters influencing the properties of graphite artifacts are the size and size distribution of the coke grains, the shape of the grains, the pore size distribution, the interaction between the components, and the content of impurities. Each parameter can be changed through the selection of raw materials and production methods. For instance, needle coke consists of highly ordered and aligned crystallites, regular petroleum coke is less ordered, and carbon black products are not ordered at all. The anisotropy of graphite artifacts containing these raw materials decreases in that sequence. The alignment of anisometric grains is also influenced by the forming methods. Anisometric particles are more perfectly aligned by the extrusion process than by vibration compacting. The microporosity of the artifacts that is developed during cooling from graphitizing temperatures affects the strength and thermal expansion [277].

Physical Properties. The variability of carbon and graphite products is illustrated in Table 8. The mechanical properties are similar to those of ceramic materials, whereas the thermal conductivity and electrical resistivity correspond to those of metals.

Table 8. Properties of carbon and graphite products (with grain)

Carbon block for furnace linings	Carbon anode	Graphite electrode		Graphite die for continuous casting	Resin-impregnated graphite for chemical application	Glassy carbon	Carbon fiber (HM)
		normal	UHP				

Bulk density, g/cm^3	1.55	1.52 – 1.60	1.55 – 1.65	1.65 – 1.75	1.70 – 1.75	1.85	1.55	1.90
Open porosity, %	17	15 – 18	20 – 25	18 – 20	15 – 18	0	0	0
Young's modulus, kN/mm^2	11	12 – 15	6 – 8	7 – 9	10 – 12	15	25	400
Flexural strength, N/mm^2	10	10 – 20	7 – 10	10 – 15	30 – 40	35 – 40	70	3000
Resistivity, Ω µm	50	25 – 35	8 – 10	5 – 8	10 – 15	15	50	1
Thermal conductivity, $W\ m^{-1}\ K^{-1}$	3	3 – 5	150 – 160	160 – 180	120 – 140	120 – 130	5	> 100
Coefficient of thermal expansion, $10^{-6}\ K^{-1}$	3.5	2.5 – 3.5	0.7 – 1.4	0.5 – 1.1	1.2 – 1.6	3.5	3.5	– 1.0

The unique combination of refractory and metallic properties is useful for electrothermal and electrolytic processes performed at higher temperatures. In progressively loaded graphite artifacts, a rising number of microcracks is generated, propagating preferentially normal to the applied stress and causing relief of internal stress. The propagation of the cracks is restricted by pores or structural elements which are orientated transversely to the direction of propagation. When through linking of microcracks a critical crack length is reached macrofailure will occur. A model of fracture and stress distribution depending on microstructural features is described in [278]. Macroscopically, the Weibull approach can be used to give a reasonable prediction of failure. The fracture toughness K^c of common graphite grades is about 1 – 15 $MN/m^{3/2}$ [279], [280].

Graphite is a semimetal, in which the highest filled valence band overlaps the lowest empty conduction band [281]. There is no simple relationship between the conductivity of single graphite crystals and that of polygranular artifacts. The resistivity (with grain) of the latter is more than 20 times greater than that of a single crystal (in the *a* direction), and it decreases with temperature up to ca. 350 °C for graphite grades and up to ca. 1000 °C for carbon grades. The sign of the temperature coefficient of resistivity changes above these temperatures due to the fact that increased scattering of the charge carriers compensates their rising number in the conduction band. The thermal conductivity of carbon and graphite artifacts is determined by photon scattering at the crystal boundaries [282]. Conductivity is at its maximum at about room temperature. More informative for users of carbon and graphite artifacts may be the change of these properties with temperature. Strength and Young's modulus of graphite artifacts reach a maximum at temperatures of ca. 2500 °C, where they are ca. 50 – 100 % larger than at room temperature. Above 1400 °C, the specific strength (strength/density) of graphite is better than that of metals and most other refractory materials. The coefficient of emissivity amounts to ca. 0.8 – 0.9. The temperature dependence of strength and several other properties is illustrated in Figure 30. Types used at high temperatures are often exposed to sudden temperature changes or large temperature gradients, both inducing mechanical stresses. Table 9 shows the favorable thermal-stress-resistance factor *R′* of graphite compared to that of several refractory materials.

$$R' = \frac{k\sigma}{(1-\mu)E\alpha}$$

where k = thermal conductivity, μ = Poisson's number, E = Young's modulus, α = coefficient of thermal expansion, and σ = tensile strength.

Table 9. Average thermal-stress-resistance factor *R′* (W/m)

Graphite	50 000
Titanium carbide	1 000
Magnesia	50
Alumina	1

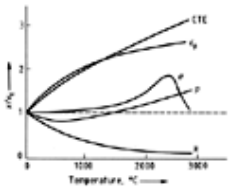

Figure 30. Relative properties of polygranular graphite grades vs. temperature (schematic)

CTE = coefficient of thermal expansion; c_p = specific heat; σ = tensile strength; k = thermal conductivity; ϱ = electrical resistivity

Chemical Properties. Carbon and graphite products are resistant to most corrosive substances, except strong oxidants, substances forming intercalation compounds, and carbide-forming substances [283]. At higher temperatures, graphite is, in fact, less reactive to oxygen than most metals. Since the reaction products CO_2 and CO are volatile, no protective oxide film can be formed. The reactivity depends primarily on the degree of crystalline order. The binder phase, which is normally less ordered, is preferably attacked. At lower temperatures, the reaction is not limited to the outer surface of the artifact but includes practically the entire inner surface, due to pore diffusion. With increasing temperature, the depth of the reaction zone decreases. In air, the average loss of small graphite artifacts in the temperature range between 500 and 600 °C is ca. 1 % per day. Impurities such as alkali metals, vanadium, and lead accelerate the reaction rate considerably. However, most carbon and graphite products are used at high temperatures in oxygen-containing atmospheres. The burn-off at these conditions may be decreased by impermeable protective coatings, e.g., carbides or silicides, or by the addition of substances which inhibit oxidation [272].

Other properties that are important to special applications of carbon and graphite products, e.g., lubrication properties, are described in Section Applications.

4.1.3. *Applications*

Carbon and graphite artifacts are mainly used in metallurgy, electrochemistry, chemistry, and mechanical and electrical engineering.

Electrodes for Arc Furnaces. About 90 % of the graphite products are electrodes used for the production of electric arc furnace steel. The production of electric steel has been increased by the rise in productivity, the development of small steel works (mini mills), and direct reduction methods (DR) supplying the electric furnaces with processing materials other than scrap. The addition of thin-slab casting to mini mills has opened up new appllications in the automotive industry. In 1999, electric furnaces accounted for some 33 % of the worldwide raw steel output, as opposed to to 25 % in 1989. Graphite electrodes were required to meet the rising demands of increased electric load up to ca. 30 A/cm^2.

Graphite electrodes up to 750 mm in diameter consist of cylindrical sections up to 2700 mm in length, which are joined with tapered, threaded nipples. High-accuracy machining is a prerequisite for a well-functioning joint. The screw joint is secured by pitch or resin cement bridges. This material is stored in the recesses of the nipple and flows when the electrodes are heated in at least part of the thread clearance. At higher temperatures, the cement carbonizes and prevents the electrode sections from loosening [284], [285]. The surface temperature of the electrode increases along the entire length of the electrode to 2300 – 2800 °C [286]. The burn-off through oxidation is proportional to the surface area and is ca. 5 kg m^{-2} h^{-1} at normal operating conditions [287]. Carbon is also vaporized from the arc spot at the tip of the electrode. The loss is proportional to I^2 (I = electric current). Accidental sources of loss are breakages of the electrode sections. Depending on the operating conditions, the specific graphite consumption is approximately 1.5 – 3.5 kg per tonne of steel. With D.C. furnaces consumption figures of < 1 kg of graphite per tonne of steel have been achieved. Surface and tip losses range between 30 and 70 % each. The relative breakage loss is ca. 5 – 20 %. Figure 31 shows graphite electrodes removed during recharging of an arc furnace.

Figure 31. Withdrawn graphite electrodes during recharging of the arc furnace

To reduce oxidation losses, the electrode surface can be coated with refractories and oxidation-resistant metals [288], [289]. For the same reason, water-cooled electrode systems have been introduced. In such a system, the upper section of the electrode consists of a water-cooled metal cylinder instead of graphite [290], [291]. The tip losses can be reduced by using longer arcs and weaker current. The breakage losses caused by stresses due to temperature changes and by bending moments due to scrap motions and vibration are decreased by using high-quality electrodes (low electrical resistivity and high thermal shock resistance and flexural strength) and relieving the stress. The electrodes can be upgraded by using selected needle coke grades and impregnating the electrodes with pitch. In general, breakage losses of impregnated electrodes are only about one-third of that for nonimpregnated types. Severe thermal stresses are induced in the joint components — nipple and electrode socket — due to the difference in properties of these components [292], [293]. To avoid any undue increase in thermal stresses, the stress must be relieved by special shaping of the joint and its clearance [294], [295]. It was also proposed to relieve stresses in the electrode section, e.g., by slotting the electrode and even the nipple [296]. Cost-saving measures will focus on a further reduction in power-on times, which can be achieved by scrap preheating, post-combustion of CO, and increased injection of oxygen and coal. These measures will require improved stability of the graphite electrode material.

Electrodes for Electrothermal Reduction Furnaces. Electrothermal reduction furnaces are used for the production of calcium carbide, ferroalloys, phosphorus, silicon, and refractory oxides. Due to the relatively low current of ca. 6 – 10 A/cm^2, adequate results can be achieved with carbon electrodes, particularly Söderberg electrodes (self-baking electrodes). Söderberg electrodes consist of a sheet metal casing that is filled with a pasty mixture of electrode raw materials. The lower part of the electrode is baked by the excess furnace heat. In accordance with the consumption rate of the electrode — ca. 20 – 30 kg per ton of product — the electrode is lowered stepwise, a new casing sheet is joined to the upper part, the casing is filled with the mixture, and the middle section is baked. Hollow electrodes enable central charging, which in turn improves furnace operation [298]. Baked carbon electrodes consist of several sections that are connected by threaded joints. One end

of each section is nipple-shaped, while the other end is socket-shaped with an appropriate thread. Due to their higher electric conductivity, semigraphite electrodes are also used for reduction furnaces, especially for the production of silicon [299]. The combination of a graphite core surrounded by Söderberg paste has also gained market shares [297]. Electrodes for the fusion of tap holes generally consist of graphite.

Electrodes for Molten Salt Electrolysis. Carbon anodes and cathodes are used for the electrolytic production of aluminum. The anodes are partially submerged into a cryolite melt containing several percent of alumina. During electrolysis, the anodically produced oxygen reacts with the anode material and CO and CO_2 are formed. The ratio CO_2 : CO determines the current efficiency. Due to the operation temperature of 950 – 970 °C, the parts of the anode extending out of the melt are attacked by oxygen in the air. Consumption greatly depends on the properties of the carbon anodes, especially their permeability and reactivity, which are a function of the structural disorder and the impurities catalyzing oxidation, e.g., alkali metals [300]. As in reduction furnaces, Söderberg anodes are used in addition to prebaked anodes, which are now favored, since the waste gas of furnaces operated with prebaked anodes does not contain tars and other pyrolysis products, which complicate the gas cleaning. The average consumption of prebaked anodes is ca. 0.47 kg per kg of Al (theoretically 0.33 kg per kg of Al), and that of Söderberg anodes is ca. 0.55 kg per kg of Al. The difference is partially due to the baking loss when the Söderberg mixture is carbonized. Current supply rods are connected to the anode by carbon-containing cements or metal solders. The edge length of the anodes is up to 1500 mm (Fig. 32).

Figure 32. Carbon anode

The bottom and sidewalls of the Héroult–Hall cells, which protect the pot and the ceramic insulation against the cryolite melt, also consist of carbon material. Carbon blocks that form the bottom have longitudinal grooves in which cathodic bus bars are inserted. Bars and carbon blocks are joined in a manner similar to the anodic current supply rods. The joint between neighboring carbon blocks needs special attention. Different joining methods have been developed using special cements [301]. Cell life depends almost exclusively on the resistance of the carbon bottom. During operation, sodium and aluminum penetrate the carbon blocks, thereby forming intercalation compounds and aluminum carbide, which results in swelling and crack formation. These reactions are probably responsible for the long-term upheaval of the carbon bottom and the penetration of molten aluminum into the bus bars [302]. Both effects determine the cell's lifetime. Some reliable tests are available for predicting the serviceability of carbon grades [303]. If semigraphite blocks are used and the width of the joints is reduced, cell lifetime is increased to more than 3000 d [304]. Therefore, anthracite cathodes are increasingly being replaced by semigraphite or semigraphitized cathodes.

Due to its better resistance to chlorine, graphite anodes instead of carbon anodes are used in processes requiring molten chloride-containing electrolytes. The winning of aluminum from aluminum chloride melts [305], of alkali and alkaline-earth metals, of rare-earth metals, and of fluorine are examples. The anodes have often complicated shapes, e.g., slotted, hollow cylinders. Even though graphite reacts with fluorine in alkali fluoride melts, the lifetime of graphite anodes is adequate due to the formation of CF_x layers on its surface.

Electrodes for Aqueous Electrolysis. During the past decades, titanium anodes have been almost completely substituted for graphite anodes. Titanium anodes that are activated with platinum or a platinum compound often achieve a lower overvoltage than graphite anodes and do not “burn off” in the presence of oxygen-containing reaction products. Accordingly the use of graphite anodes for chlorine, chlorate, and hydrochloric acid electrolysis, for the production of manganese dioxide and hydrogen peroxide, for electrolytic pickling and polishing processes, and for protection against cathodic corrosion is decreasing. Electrodes for electroorganic synthesis are less affected [306].

Linings, Casting Molds. Due to their temperature resistance and their resistance to most metal melts and slags, carbon and graphite blocks are widely used as a lining for metallurgical furnaces. For use in blast furnaces, resistance to alkali attack and wear is required. For each refractory material, a critical temperature, which must not be exceeded during furnace operation, has been defined. To obtain sufficient lifetime from the lining carbon, grades with high critical temperature should be used (refractory approach), or the cooling of the lining must be increased by using semigraphite or graphite grades (thermal approach) [307]. The actual thermal and mechanical load of each part of the furnace must be calculated before the appropriate lining materials are selected. In Europe, the linings usually consist of large bricks (e.g., 600×800×2600 mm), in the United States smaller bricks are preferred. The number of joints can be reduced in the first case, whereas smaller bricks can be more easily adapted to the shape of the furnace. Carbon, semigraphite, and graphite linings are limited to the lower stack, bosh, hearth, and the bottom, whose life expectancy has been increased to more than 10 years. Lining and steel jacket are thermally coupled by a carbonaceous ramming mass with the necessary thermal conductivity. Other shaft furnaces with carbon linings are used for the production of ferroalloys, carbides, corundum, phosphorus, as well as melting furnaces for slag and zinc, and cupola furnaces.

Due to its thermal conductivity, thermal shock resistance, and low wettability by molten metals and slags — the wetting angle for most metals is 120 – 140° — graphite is an outstanding material for foundry molds, ingot molds, and crucibles for the casting of aluminum, magnesium, iron, steel, copper, brass, bronze, and other metals [308]. Continuous casting devices for steel generally consist of materials containing natural graphite and corundum. For nonferrous metals, high-density and high-strength graphite grades are used. Railway wheels are die cast in graphite dies. The station time is about 1 min and the die requires slight refinishing only after about one hundred casts [309]. Graphite molds are also used for precision casting [310]. All molds must have a smooth surface and must be uniformly permeable.

Graphite is also used for the production of cemented carbide and diamond tools, such as die, boat, crucible, and auxiliary devices for the production of semiconductor materials [311], for the refining of metal melts by feeding gases into the melt, and as a resistance heating element. Granular carbon and graphite are common resistor materials in Acheson furnaces for the production of graphite and silicon carbide. The resistivity depends on the packing density, which may vary to a great extent. Graphite rods, tubes, or slotted hollow cylinders have a more uniform resistance and are, therefore, preferred as heating elements for most other furnaces [312]. Graphite heating elements must be used in an inert atmosphere or under vacuum, keeping in mind that the vapor pressure of carbon is ca. 0.5 kPa at 3000 °C.

Process Equipment. One outstanding property of carbon and graphite is their resistance to corrosion. Graphite also has a relatively high thermal conductivity and is an interesting material for heat exchangers and chemical equipment provided that its permeability can be considerably reduced. The sealing of graphite artifacts is achieved by impregnation, especially with thermosetting resins such as phenol aldehyde or furfuryl alcohol resins. The permeability of impermeable graphite grades is less than 10^{-8} cm^2/s. The temperature at which it can be used is, of course, restricted by the decomposition temperature of the resins, but the limit of 200 °C is adequate for most chemical processes. To reach a higher application temperature, the impregnating media must be transformed into carbon and the loss of pyrolysis products must be compensated by a second impregnation and carbonization, etc. Usually two repetitions are needed for an impermeable allcarbon shape. The same result can be achieved by depositing pyrocarbon into the pores.

Examples of impervious carbon and graphite grades are absorbers, columns, heat exchangers, pumps, valves, pipes, fittings, towers, tower packing, and rupture disks (Table 10) [313]. Contrasted with metals, stresses are only dissipated by microfracture. Stress peaks must be carefully avoided through proper design [314]. This is especially true for all joints. The permissible operation pressure can be increased by fiber-reinforced types [315] and erosion resistance by certain coatings [316]. Standard quality is guaranteed through strict and extensive inspection.

Table 10. Carbon and graphite processing equipment

Equipment*	Max. size		Max. pressure (MPa)	Favored application
1. Heat exchanger				
Shell and tube	length	10 m	0.6 (1.2)	large heat exchange areas;
heat exchanger	diameter	2 m		fouling and encrusting substances
	exchange area	1000 m^2		
Block heat exchanger	length	9 m	0.6 (1.5)	smaller heat exchange areas;
	diameter	1.5 m		narrow spaces; high heat transfer rates
	exchange area	500 m^2		
Plate heat exchanger	length	2.3 m	0.8	high transfer rates; small pressure drops;
	width	0.7 m		small temperature differences
	length	2.5 m		
	exchange area	50 m^2		
2. Columns and vessels	diameter	1.5 m	0.2	mass transfer
3. Equipment for	115 t of		0.1	synthesis of anhydrous HCl gas;
the production of	HCl gas			production of pure hydrochloric acid
hydrochloric acid	per day			
4. Rupture disk	diameter	0.5 m	8.0	safety devices for chemical equipment
5. Centrifugal pumps	400 m^3/h		1.1	booster pump for corrosive liquids

* Maximum temperature: pos. 1, 200 °C; pos. 2 and 5, 180 °C; pos. 3 and 4, 350 °C.

Nuclear Reactors. Carbon has a low absorption cross-section for thermal neutrons and a favorable moderation coefficient. Since Fermi's Chicago pile reactor, graphite has been used successfully as a moderator and reflector material for gas-cooled reactors. The operation temperature is limited by the nature of the cooling media and is far lower for CO_2-cooled Magnox reactors than for helium-cooled high-temperature reactors (HTR). The neutron absorption cross-section is particularly

affected by such impurities as boron and rare-earth metals, which must be reduced by purification. The material must be processed under “clean conditions” [163]. During neutron irradiation, graphite undergoes considerable dimensional changes, which depend on structural disorder and the coefficient of anisotropy. The degree of damage can be considerably reduced through selection of raw materials and processing conditions [317]. The dimensional changes of the structural elements of graphite during the lifetime of an HTR are now small enough to be compensated by structural design.

Mechanical Engineering Applications. The self-lubricating characteristic of carbon is a fundamental property. The coefficient of static friction is ca. 0.1 – 0.25 at normal humidity. Carbon and graphite are, therefore, widely used as mechanical seals and bearings if the use of lubricants is not possible due to contamination (food), contact by aggressive substances (chemical industry), or operation at high temperatures. Carbon and graphite radial and axial bearings and bushings are used, for example, in immersed pumps and circulation pumps for feed water and cooling fluids. These bearings do not require maintenance for years [318]. Carbon and graphite sealing rings are used to prevent fluid leakage along rotating shafts and vanes [319]. Graphite and carbon vanes are used in rotary compressors and pumps. Depending on the different heat conductivity requirements, wear resistance, and mechanical load, the samples are usually either impregnated with resins, glasses, or metals (babbit, antimony, silver), or coated with resistant silicon carbide by reacting silicon or silicon monoxide with the carbon [320].

Miscellaneous Electrical Applications. Carbon and graphite are classical materials for electric contacts and electrodes whose range of application has been considerably expanded. Arc light carbons had a comparatively short history as a street illuminant. They are now almost exclusively used in film projectors. Color, temperature, and brightness have been improved by the introduction of carbons consisting of a carbon shell and a core containing alkali salts and rare-earth metal compounds [321]. Further developments of arc light carbons are arc cutting and welding rods. The arc cutting and gouging process requires the striking of an arc between the carbon electrode and a metal workpiece. A high-velocity air stream is directed to the base of the arc blowing the molten metal from the workpiece. The operation velocity depends on the electric power and the geometry of the air streams [322]. Welding carbons are preferred for spot-welding sheet metal.

For spectroscopic emission analysis, the sample is vaporized and the atoms are excited in an arc between two carbon electrodes. These electrodes tolerate only little contamination, which is also true for crucibles and boats used in atomic absorption spectroscopy. The electrically heated, mostly tubular crucibles are often coated with an impermeable pyrographite layer which is roughened to increase its reactivity [323]. Graphite tool electrodes are juxtaposed with a workpiece in an electric-discharge machine (EDM), both immersed in a dielectric liquid. The workpiece is eroded by a series of electric pulses. The graphite electrode loss decreases with the rising bulk density of the graphite grade. The smoothness of the eroded surface depends on grain and pore size of the electrode [324]. The size of EDM electrodes, which can be easily machined, ranges between a few millimeters and more than 1 m.

Graphite electrodes are also used as anodes for mercury-vapor rectifiers and power tubes. Sparking can be avoided if the electrode consists of high-density graphite grades which were heat treated under vacuum at temperatures above 2000 °C [325]. Graphite grids for high-power electron tubes improve their power handling capacity [326].

Carbon is applied in various types of primary batteries (battery carbon) and fuel cells. Particularly fuel cells have gained renewed interest for automotive and stationary applications. Carbon does not enter into the cell reactions, but acts as a corrosion-resistant conductor, as an electrocatalyst in air-consuming electrodes, and as a catalyst carrier [327]. In Leclanché cells, the carbon rod is surrounded by a depolarizing mixture containing MnO_2 and graphite powder or carbon black. The rod is often impregnated with waxes or fats to prevent evaporation of the electrolyte and permit the diffusion of gases. Leak proof cells are provided with additional seals.

Stacked carbon plates or disks are used as pressure-sensitive pile resistors. Examples are voltage regulators for sewing machines and illumination equipment. The high contact resistance of carbon granules, their invariability against atmospheric conditions, and the reproducibility of the resistance – pressure curve make carbon granules a satisfactory material for microphones. The granules are often stabilized by pretreatment through vibration or surface oxidation. Carbon diaphragms improve the frequency response with these granules (microphone carbon).

Carbon and graphite are important materials for use as electrical contacts. Carbon contacts have an excellent arc resistivity, they neither weld nor form isolating oxide films. Examples of stationary contacts are switches and circuit breakers for sliding contacts, as well as brushes and current collectors for trolley shoes and wheels [328]. Pantograph shoes are sliding contacts that transfer electric energy from a wire or rail to a moving device (vehicle). Carbon shoes spare trolley cables, wear slowly, have adequate resistance to impact, and cause little radio interference. The shoes are adjusted to special pantographs which maintain nearly constant contact pressure. Examples are electric locomotives (engines), tramways, trolley busses, and mobile cranes.

In most electric motors and generators, the revolving part of the machine and the stationary part are electrically connected by brushes. Brushes must be able to “absorb” the roughness of the commutator or slip ring surface, control film formation on the metallic counterpart, and absorb current pulses which result during current reversal (commutation). The diversified requirements are satisfied by numerous brush grades which can be adjusted to the various operating conditions. Typical examples are the following:

1. Carbon brushes with good polishing properties (for small electrical loads and circumferential speeds)
2. Soft natural graphite brushes with a great damping decrement (for high circumferential speeds)
3. Carbon – graphite brushes with good commutation properties (for small machines)
4. Resin-bonded graphite brushes (for machines with high commutation loads and average circumferential speeds, fractional horsepower brushes)
5. Impregnated electrographite brushes (power and industrial brushes)
6. Copper – graphite brushes (for machines with high electrical loads, automotive DC starter brushes)
7. Brushes containing lubricants for high-altitude use (aircraft and aerospace brushes)

4.1.4. *Economic Aspects*

By far the largest quantity of carbon and graphite products is used as carbon anodes in the aluminum industry and graphite electrodes in the steel industry, respectively. The average anode requirement is about 500 kg per ton of aluminum and that of graphite electrodes is 1.5 – 3.5 kg per ton of steel, both with decreasing tendency. The calculated consumption rate is shown in Table 11. The carbon and graphite requirement for other uses is much smaller, probably less than 15 % of the quantities shown in Table 11.

Table 11. Consumption of carbon anodes and graphite electrodes (in 10^3 t)

Year	Carbon anodes for aluminum electrolysis				Graphite electrodes for electric steel production			
	World	USA	Japan	Germany	World	USA	Japan	Germany
1950	750	220	10	14	116	38	5	2
1960	2260	900	65	84	294	58	30	15
1970	5150	1800	360	152	645	124	102	30
1980	8200	2300	550	365	1000	186	125	37
1985	7700	2900	110	380	1000	133	110	33
1999	12 500	1900	10	225	930	140	70	30

4.2. Other Classes of Carbon

Carbon and graphite products described in Section Carbon and Graphite Artifacts consist of an agglomeration of single carbon particles and can be regarded as a polygranular structure. Industrially important products have been manufactured from essentially monogranular carbon materials [329]. Carbon fibers are an important member of this group (⟶ Fibers, 5. Synthetic Inorganic).

4.2.1. *Glassy Carbon*

Glassy carbon is formed by pyrolysis of thermosetting resins, e.g., phenol, formaldehyde, furan, polyimide, and polyphenylene resins. The use of cellulosic or hemicellulosic materials as a precursor is less common [330]. For the production of glassy carbon artifacts, liquid resins are poured into molds, dry resins are molded or extruded, and the shapes are heat treated first for curing and then for pyrolysis. Porosity and pore diameter of the thermoset resins are very small, which complicates the diffusion of gaseous pyrolysis products. Although the heating rate during this process is small, the permissible thickness of the artifacts is at most 5 mm. Material shrinkage due to a mass loss of ca. 40 % during pyrolysis is also a restricting factor [331].

Glassy carbon is impermeable to liquids and gases and its fracture exhibits glossy shell-like planes. The highly disordered structure consists of a network of branched and twisted ribbons and microfibrils, which are interconnected at several points with an average height and length of ca. 3 and 10 nm, respectively. The ribbons surround micropores of 5 – 10 nm diameter, and the total porosity is ca. 40 % [332]. It is possible to produce glassy carbon grades with a specific BET surface and pore diameter by selected raw materials and heat treatment. These products are useful as molecular sieves or microfilters [333].

Due to its cross-linked microstructure, glassy carbon is isotropic and stronger and harder than polygranular carbon grades. The oxidation rate is considerably lower, since most pores are not accessible to reagents. These properties are advantageous for laboratory equipment, e.g., crucibles, boats, bowls, trays, and tubes for evaporation and smelting, as well as for the decomposition of materials for chemical and spectroscopic analysis (Fig. 33). Resistance-heated tubes are used in atomic absorption spectroscopy, molds for glass forming, electrodes for organic electrochemical synthesis, and other shapes for numerous other uses [335-337]. Due to its compatibility with human tissue, glassy carbon has also been successfully used for implants, e.g., bile ducts, and as stimulating electrodes for cardiac pacemakers [338], [339].

Figure 33. Laboratory equipment made of glassy carbon [334]

The carbonization of foamed resins, which are either thermoset or thermally stabilized by cyclization, yields foamed carbon. In particular, phenolic and polyurethane foams, hollow microspheres consisting of the same material, or thermally stabilized pitch are the initial raw materials [340], [341]. The latter must be mixed with a binder before carbonization. The thermal treatment is the same as for the production of normal glassy carbon. The carbon foam consists of a network of thin carbon cell walls (several micrometers) which surround comparatively large pores. The bulk density of carbon foam is less than 0.1 g/cm^3, and the thermal resistance is, therefore, almost as low as that of foamed resins. Carbon foam differs from these materials in that it can be used at temperatures in excess of 3000 °C, provided that oxidative atmospheres are excluded.

Foamed glassy carbon is an excellent insulating material for high-temperature furnaces, e.g., vacuum sinter furnaces or carburation furnaces. Above ca. 2000 °C, it is advisable to cool the insulation from the outside to reduce energy loss through radiation. Other uses are casting cores, substrates for catalysts, porous electrodes, and filters.

4.2.2. *Pyrocarbon and Pyrographite*

Pyrocarbon and pyrographite are products of the decomposition of gaseous carbon-containing compounds, a process called chemical vapor deposition (CVD). Parameters influencing this process are the type and concentration of the carbon donor, the temperature of pyrolysis, contact time, and gas flow [162]. Hydrocarbon compounds are favored as reaction gases, especially methane, propane, and propylene. At low temperature and low concentration, the deposited pyrocarbon layers have a laminar structure. At high concentration and a low temperature gradient, the gas is supersaturated and optically isotropic layers are formed. High temperature and low partial pressure promotes the formation of granular and columnar microstructures. The pyrocarbon is deposited on substrates, which have the contour of the desired final body. The deposition rate is ca. 10 – 2000 µm/h. The bulk density depends primarily on the deposition temperature; it is lowest at temperatures of ca. 1400 °C (1.3 g/cm^3) and increases to ca. 2.23 g/cm^3 at 2700 °C. Due to structural variation, it is possible to adjust the coefficient of thermal expansion of the layer to that of the substrate, thereby minimizing stresses caused by temperature gradients.

Highly oriented pyrographite (HOPG) is obtained by thermal or stress annealing of pyrolytic carbons [342]. This process decreases the proportion of disordered layers considerably. Interlayer spacing, which is the distance between the single graphite layer planes, can be as low as 0.3354 nm. Accordingly, the properties of this type of pyrographite are similar to a single graphite crystal. HOPG plates are used as flat and bent X-ray monochromators, neutron monochromators, and for the polarization of high-energy photons [172]. Nuclear fuel particles for high-temperature nuclear reactors are coated with less-ordered pyrolytic carbons impermeable to fission products at high operating temperatures (coated particles) [162], [343]. The carbon is deposited in a fluidized-bed, and the deposition conditions are altered so that the coating consists of several layers differing in microstructure and properties.

Since pyrolytic carbon is compatible with human tissue and has a high degree of reproducibility, it is an interesting material for bioengineering [344]. Alloyed pyrolytic carbon grades which contain a percentage of silicon are particularly useful. They are used for prosthetic heart valves, hip-joint prostheses, and dental prostheses. As mentioned above, pyrolytic carbon can be deposited in porous solids to reduce their permeability and improve the strength of the porous material [269], [270]. Fibrous materials such as carbon fabrics and carbon felts are an important infiltration substrate. The deposited pyrolytic carbon forms a closed matrix which is reinforced by the fibrous component. This composite (carbon fiber-reinforced carbon) is less sensitive to shear stresses than normal pyrolytic graphite. It is widely used in rocket nozzles, nose cones, heat shields, prostheses, and friction materials, e.g., brakes and clutches.

4.3. Graphite Compounds

4.3.1. *Surface Complexes*

Carbon and graphite form surface complexes or groups through surface reaction or chemisorption when brought in contact with certain gases and liquids [345], [346]. The proportion of these surface groups increases with rising structural disorder and surface area. This proportion is high for carbon blacks and chars and low for single graphite crystals. The total amount of gas adsorbed by graphite grades is ca. 100 – 600 mL/g. The adsorption of oxygen begins at room temperature, forming a chemisorbed surface layer. At higher temperature or in the presence of liquid oxidizing agents, functional groups such as carboxyl, phenolic hydroxyl, and carbonyl are formed. Lactones, quinoid structures, and hydroperoxides are less frequent. The highest desorption rate is achieved at temperatures between 600 and 800 °C.

Surface oxides and physically adsorbed compounds undoubtedly determine the friction properties of carbon and graphite shapes; e.g., graphite wear increases rapidly if the amount of surface “impurities” falls short of a critical threshold. On the other hand, it is still not always clear whether these surface groups have an appreciable effect on the binding strength of carbon grains or fibers distributed in a continuous binder matrix or if an increased binding strength is the consequence of interlocking or keying-in of roughened adjacent surfaces.

4.3.2. *Graphite Intercalation Compounds*

At relatively low temperatures, graphite forms intercalation compounds with numerous substances. Graphite oxide with the ideal formula $C_8O_2(OH)_2$ was the first compound investigated [347]. The following rules are valid for these intercalation compounds: 1) only highly ordered graphite or graphitizable materials yield homogeneous intercalation compounds [348]; 2) the atomic arrangement within the carbon layers is preserved; 3) the intercalated substance is regularly distributed between the carbon layers; and 4) most compounds can be reconverted into graphite.

The reagents capable of intercalation with graphite can be divided into two classes: electron donors (alkali metals, alkaline-earth metals, lanthanides, metal compounds containing hydrogen or polar molecules, and aromatic compounds) and electron acceptors (halogens, halides, oxyhalides, and acids) [349]. The number of electrons in the conduction band changes accordingly, eventually reducing the electrical resistivity considerably. Due to reagent penetration, the graphite crystal swells in the *c* direction and the identity period *I* (in nanometers) changes according to $I_{(n)} = d_i + (n - 1)0.3354$, where 0.3354 nm is the interlayer spacing of graphite, d_i is the period of a first-stage intercalation compound, and *n* is the stage number. “First-stage” compound implies, that all intervals between the carbon layers are occupied. In a second-stage compound, each intercalated layer is separated by two carbon layers, etc. (Fig. 34). A defined sublattice is attached to each stage conserving the tridimensional structure [350].

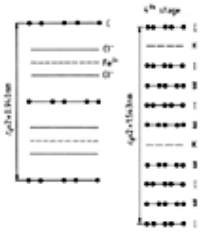

Figure 34. Graphite intercalation compounds, layer sequence (schematically) for C_7FeCl_3 (first stage) and $C_{48}K$ (fourth stage)

For the production of graphite intercalation compounds, graphite powder is heat treated in the presence of a gaseous or liquid reagent. The pressure at which intercalation begins depends on the polarity and the structural disorder of the graphite [351]. An inverse reaction of intercalation is the desorption of the intercalant. Graphite intercalation compounds have only limited stability. Compounds containing electron donors are easily oxidized when exposed to air — some are even pyrophoric — and compounds containing electron acceptors decompose easily or are strongly hygroscopic. The use of graphite intercalation compounds is, therefore, very limited. This is not true for graphite fluoride, CF_x, since this compound is adequately stabilized due to the formation of sp^3 bonds. The carbon layers do not remain flat but become rippled.

The use of graphite intercalation compounds as catalysts had already been proposed at an early time [349]. Catalysts of this type have been studied for use in the synthesis of ammonia, Fischer–Tropsch synthesis, reduction and oxidation of organic compounds, halogenation, initiation of polymerization, and hydrodesulfurization [352]. The proposed application of graphite intercalation compounds as low-resistance conductors has especially been studied with SbF_5 and $SbCl_5$ intercalants [353], [354]. At room temperature, the electrical resistivity is as low as 3×10^{-8} Ω m. The powdery material must be filled in copper tubes, the grains must be aligned parallel to the axis of the tube and the tube must be kept free from moisture. The increase in conductivity does not seem to compensate these costly operations. On a small scale, graphite fluoride is used as a lubricant [355].

4.4. Expanded Graphite and Graphite Foil Derived from Intercalation Compounds

Graphite intercalation compounds are starting materials for the production of expanded graphite particles. Intercalation compounds expand considerably during decomposition due to rapid heating. Under certain conditions, the volume increases more than 100 times [356]. Luzi observed that the expanded vermiform particles, which are segmented like the leaves of a partially opened book, are plastic and can be easily pressed into flexible sheets and other shapes [357]. The addition of a thermosetting binder to the expanded graphite particles to form bearings, seals, packings, brushes, etc. from the mixture was also proposed [358]. This process waited until about 1960, when the strong dependence of the strength of the molded product on the degree of the particle's expansion had been discovered. The expansion parallel to the *c* axis of the graphite crystal must be at least 80 times the original thickness to achieve the proper interlocking of the graphite particles necessary to form flexible graphite sheet material with sufficient strength [359].

Highly ordered graphite flakes with stack heights of at least 75 nm are used as a raw material [360]. The precursor flakes are usually treated with an oxidizing agent, e.g., mixtures of sulfuric acid and nitric acid at ca. 100 °C. After the graphite intercalation compound is formed – in this example graphite hydrogen sulfate – the particles are thoroughly rinsed with water and then rapidly heated to ca. 1000 °C. The gas pressure of the vaporized and trapped intercalants partially contributes to the exfoliation of the stacked layers. The exfoliated particles (expanded graphite) can be molded to sheets, laminates, and other useful shapes in the absence of a binder. During compression, the particles interlock and form nonrigid bonds.

Properties. The large anisotropic coefficient is due to the nearly complete alignment of the layer planes perpendicular to the compression force. Sheets are good thermal and electrical conductors parallel to the surface and semiinsulators in perpendicular planes. Graphite foils have considerable flexibility; they can, for example, be wrapped around cylinders, and they are impervious to fluids even at high temperatures. Like other graphite grades, this product is resistant to temperature, aging, and most corrosive media. Graphite foils have a smooth surface, they are under most conditions resistant to wear, and the coefficient of friction is about 0.08 – 0.1. They are not wetted by melts of glass, ceramic, or metal.

Applications. Flexible graphite is an extraordinarily good packing and sealing material, which will probably replace asbestos due to the lower health risks. Examples are cylinder-head gaskets reinforced with metallic inserts and stuffing box seals consisting of flexible graphite tapes wrapped around the shaft of a pump or valve [361], [362]. Preshaped lamellar rings are used as gaskets, e.g., for exhaust system piping. Other applications are tube linings, vessels and molds for protection against temperature and corrosive media, blackwash as protection for hot-press molds, and heat shields often in combination with carbon felt sheets. Rupture disks lined with graphite foils are workable at higher temperatures than resin-impregnated graphite disks and are more resistant to aging [363]. Heating elements can be easily cut from graphite foils and attached to high-temperature furnaces [364].

[Top of Page]

5. Activated Carbon

Klaus-Dirk Henning and Hartmut von Kienle

5.1. General Aspects

Definition. Activated carbon [7440-44-0] is the collective name for carbonaceous adsorbents which are defined as follows [365]: Activated carbons are nonhazardous, processed carbonaceous materials having a porous structure and a large internal surface area. They can adsorb a wide variety of substances, that is, they are able to attract molecules to their internal surface and therefore act as adsorbents. The pore volume of activated carbons is generally greater than 0.2 mL/g. The

internal surface area is generally greater than 400 m^2/g. The width of the pores ranges from 0.3 to several thousand nanometers.

All activated carbons [366-370] are characterized by a ramified pore system (Fig. 35) in which pores of various sizes, such as mesopores (d = 2 – 50 nm), micropores (d = 0.8 – 2.0 nm) and submicropores ($d \leq$ 0.8 nm) branch off from macropores ($d \geq$ 50 nm).

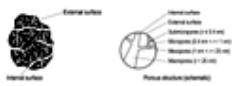

Figure 35. Schematic model of activated carbon

X-ray investigations show that the carbon is mainly in the form of very small crystallites with a graphite structure. However, the typical graphite arrangement of the carbon platelets one above the other is absent. The crystalline regions are only 0.7 – 1.1 nm thick and 2.0 – 2.5 nm in diameter, a considerably smaller size than is observed in graphite (see Table 12).

Table 12. Typical dimensions of crystal regions

	L_a, nm	L_b, nm	$c/2$, nm
Natural graphite	210	94	0.351
Activated carbon	2.0 – 2.5	0.7 – 1.1	0.35 – 0.37

This means that in each crystallite there are usually only three or four layers of carbon atoms with about 20 – 30 carbon hexagons in each layer. The spaces between the crystallites are filled with amorphous carbon which is bonded in three dimensions with other atoms, especially oxygen. The irregular arrangement of the carbon atoms is broken up by numerous cracks and fissures (pores), which are often idealized as cylindrical in shape. The large number of very fine pores (micropores and submicropores) gives activated carbon a large inner surface, which is the basis of its remarkable adsorption properties.

A common feature of these adsorbents is an amorphous structure with a high carbon content and a hydrophobic surface properties. Activated carbon is made on commercial scale from carbon-containing raw materials (wood, peat, coconut shells, lignite, hard coal) by chemical activation or gas activation. Activated carbon is commercially available in shaped (cylindrical pellets), granular, or powdered form.

History. The medical use of charcoal was described as early as 1550 B.C. in an ancient Egyptian papyrus and later by HIPPOCRATES and PLINY THE ELDER. The products described at that time and later in the 1700s were of varying effectiveness and included wood, blood, and animal charcoals.

The decolorization of sugar solutions by bone black was first carried out commercially in England in 1811. Bone black, however, consists mainly of calcium phosphate and contains only a small amount of carbon. Strictly speaking, it is not a carbon. Methods of obtaining decolorizing charcoals from plant materials were first set out in British patents in 1856 – 1863. The first industrially manufactured activated carbons in the proper sense were Eponit decolorizing carbons, which have been produced since 1909 according to a patent of R. VON OSTREJKO [371] by heating wood charcoal with steam and carbon dioxide in a furnace specifically designed for the purpose.

In 1911 in the Netherlands, Norit NV began commercial activation of peat by using steam. The chemical activation of sawdust with zinc chloride to produce Carboraffin was first described in a patent of the Austrian Association for Chemical and Metallurgical Production [372]. This process was first operated in Aussig, Czech Republic, and in 1915 by Bayer. In the United States during World War I, the activation of coconut charcoal for gas masks was developed. The many and varied applications of activated carbon are amply illustrated by the existence of over 1500 manufacturing patents worldwide [367].

During the first decades of this century, activated carbon was used mainly for the purification of products of the chemical, pharmaceutical, and food industries; purification of drinking water was also an important application from the outset. It is increasingly used for the prevention of environmental pollution and for meeting the constantly increasing demands for purity of natural and synthetic products.

5.2. Carbonaceous Adsorbents

5.2.1. *Types of Carbonaceous Adsorbents*

Due to the wide variation in the properties of activated carbon there is no comprehensive nomenclature or standardization. The product group can be characterized by appearance, pore radius distribution, or by typical applications.

Classification by *appearance*:

- Powdered activated carbon (PAC)
- Granular activated carbon (GAC)
- Cylindrical pellets
- Spherical pellets

- Activated carbon fibers (ACF)
- Activated coke

Classification by *pore radius distribution*:
- Activated coke
- Activated carbon
  - fine-pore
  - medium-pore
  - wide-pore
- Carbon molecular sieves (CMS)

Classification by *field of application*:
- Carbon molecular sieve (CMS)
- Decolorizing carbon
- Water-treatment carbon
- Catalyst carbon
- Drinking-water carbon
- Solvent-recovery carbon

Carbonaceous adsorbents are usually characterized by appearance as powder, granules, or formed shapes such as cylindrical or spherical pellets. Grouping only according to end use is not meaningful because the requirements are so varied. Also no useful classification is possible on the basis of raw materials or production methods. In practice these products are often called, for example, powdered decolorizing carbon, granular carbon for water treatment, or pelletized catalyst carbon.

In industrial practice, carbonaceous adsorbents are often classified by their pore radius distributions into activated carbon, activated coke, and carbon molecular sieves [373]. Some information on characteristic data and applications of commercial activated cokes and carbon molecular sieves are given in Tables 13, 14, and 15.

Table 13. Gas-phase applications and typical data of carbon-based adsorbents

Adsorbent	Activated carbon, fine-pore	Activated carbon, medium-pore	Activated carbon, wide-pore	Activated coke	Carbon molecular sieves
Typical applications	intake air and exhaust air cleanup, odor control, adsorption of low-boiling hydrocarbons	solvent recovery, adsorption of medium-boiling hydrocarbons	adsorption and recovery of high-boiling hydrocarbons	dioxin and furan adsorption, SO_2 and NO_x removal	N_2 and O_2 recovery from air, CH_4 from biogas
Compacted density,* kg/m^3	400 – 500	350 – 450	300 – 400	500 – 600	620
Apparent density,* kg/m^3	800	700	600	900	ca. 1000
True density, kg/m^3	2100 – 2200	2100 – 2200	2100 – 2200	1900	2100
Pore volume for pore size					
d < 20 nm mL/g	0.5 – 0.7	0.4 – 0.6	0.3 – 0.5	0.05 – 0.1	0.2
d > 20 nm mL/g	0.3 – 0.5	0.5 – 0.7	0.5 – 1.1	0.2 – 0.3	> 0.3
Specific surface area, m^2/g	1000 – 1200	1200 – 1400	1000 – 1500	< 400	< 100
Specific heat capacity, J/kgK	850	850	850	850	850

*For powdered activated carbon, only applicable with limitations.

Table 14. Liquid-phase applications and typical data of carbon-based adsorbents

Adsorbent	Activated carbon, fine-pore	Activated carbon, medium-pore	Activated carbon, wide-pore	Activated coke
Typical applications	dechlorination, removal of micropollutants, gold recovery, decaffeination	potable and wastewater purification	decolorization, wastewater purification	deozonisation, removal of manganese and iron
Compacted density, * kg/m^3	400 – 500	350 – 450	300 – 400	500 – 600
Apparent density, * kg/m^3	800	700	600	900
True density, kg/m^3	2100 – 2200	2100 – 2200	2100 – 2200	1900
Pore volume for pore size				
d < 20 nm mL/g	0.5 – 0.7	0.4 – 0.6	0.3 – 0.5	0.05 – 0.1
d > 20 nm mL/g	0.3 – 0.5	0.5 – 0.7	0.5 – 1.1	0.2 – 0.3
Specific surface area, m^2/g	800 – 1200	800 – 1300	800 – 1400	< 400
Specific heat capacity, J/kgK	850	850	850	850

*For powdered activated carbon, only applicable with limitations.

Table 15. Applications and characteristic data of activated carbon in catalysis

Adsorbent	Activated carbon	
	fine-pore	medium- and wide-pore
Applications	catalyst for chemical reactions	catalyst support for impregnation
Compacted density, * kg/m^3	400 – 500	300 – 450
Apparent density, * kg/m^3	800	600 – 700
True density, kg/m^3	2100 – 2200	2100 – 2200
Pore volume for pore size *d* < 20 nm mL/g	0.5 – 0.7	0.3 – 0.6
Pore volume for pore size *d* > 20 nm mL/g	0.3 – 0.5	0.5 – 1.1
Specific surface area, m^2/g	1000 – 1200	1000 – 1500
Specific heat capacity, J/kgK	850	850

*For powdered activated carbon, only applicable with limitations.

Figure 36 shows the schematic pore diameter distribution of activated carbon, activated coke, and carbon molecular sieves. Carbon molecular sieve clearly exhibits narrower pores than activated carbon and activated coke. The pore diameters are matched to the molecular sizes of the gases to be adsorbed.

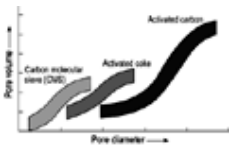

Figure 36. Schematic pore diameter distributions of carbonaceous adsorbents

Activated carbon typically exhibits pore volumes of well above 25 cm^3/100 g and pore diameters of less than 2 nm. The specific inner surface area of commercial activated carbons ranges between 500 m^2/g and 1500 m^2/g. Due to its hydrophobic character, activated carbon is particularly suited to the adsorption of nonpolar organic substances, which has led to a broad range of applications in air pollution control and water treatment. In addition, catalytic reactions occur on activated carbon surfaces. For this reason activated carbon is used commercially as catalyst or catalyst support.

Activated coke is manufactured from lignite or hard coal and has typical pore volumes of up to 25 cm^3/100 g and specific surface areas of up to 400 m^2/g. Depending on the source material and the manufacturing process, adsorptive or catalytic characteristics may predominate. Activated coke is used for removing SO_2 and dioxins from waste and flue gases.

Carbon molecular sieves have micropore diameters that range from 0.50 to 1.00 nm with pore volumes of up to 20 cm^3 /100 g. The separation effect of carbon molecular sieves (CMS) is based on differing rates of diffusion into the pore system. For this purpose the pore diameters of the CMS are matched carefully to the diameters of the molecules to be separated. Thus, a kinetic separation effect (sieve effect) is obtained. Larger molecules are adsorbed at a substantially lower rate in spite of higher equilibrium loads. For example, oxygen is adsorbed ten times faster than nitrogen although its molecular diameter is only slightly smaller.

5.2.2. *Chemical Properties*

Activated carbon contains not only carbon, but also small amounts of oxygen, nitrogen, sulfur and hydrogen, which are chemically bonded in the form of various functional groups, such as carbonyl, carboxyl, phenol, lactone, quinone, and ether groups [370], [374], [375]. These surface oxides are sometimes derived from the raw material or they can be formed during or after the activation process by the action of air or water vapor. They usually have acidic character (seldom basic), and they give to the activated carbon the character of a solid acid or base. Under suitable conditions, surface sulfides and carbon – chlorine compounds can be formed. These surface chemical properties play a significant role in adsorption and catalysis.

All the raw materials used for the production of activated carbons contain mineral components which become concentrated during the activation process. Moreover, the inorganic chemicals used in chemical activation are often only incompletely removed. The ash content of many products is reduced by water or acid washing. Hence, commercial products contain from a few tenths of a percent up to 20 % ash. The main constituents of the ash are salts of the alkali and alkaline earth metals, mostly carbonates and phosphates, together with silica, iron, and aluminum oxides.

5.2.3. *Mechanical Properties*

Performance characteristics for activated carbon are generally expressed in relation to its mass. If figures on a volume basis are needed, bulk density or tapped density figures are also needed to carry out the conversion calculation.

The bulk density is very much dependent on the filling technique, the geometry of the vessel used, and the grain size of the material. Therefore, a defined tapping or shaking process is usually included in the case of the higher tapped densities [365].

Another important feature of activated carbon is its grain size distribution. This determines the resistance of a layer of activated carbon to the flow of a liquid or gas (Fig. 37).

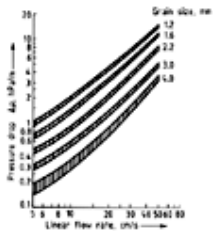

Figure 37. Pressure drop of molded activated carbon with different particle diameter (20 °C, 1-m layer, dense packing)

The fineness of grinding of a powdered carbon affects the filtration properties.

For filling a high adsorption tower or for any other static loading, the granules or pellets must be resistant to crushing. For transportation procedures such as pneumatic delivery, the abrasion resistance is very important. Care must also be taken that the thermal and chemical resistance properties are adequate to withstand any severe temperature variations or aggressive environments such as oxidizing atmospheres.

5.2.4. *Adsorption Properties*

The adsorption properties of activated carbon [366-370], [376-380] depend principally on its inner surface area, which in commercial products is 500 – 1500 m^2/g. To make use of the inner surface which is provided by the walls of the pores, the accessibility of this surface is important, that is, the pore size and the pore size distribution. This is determined by various methods, mostly from nitrogen adsorption isotherms, and is represented as integral or differential distribution curves (Fig. 38).

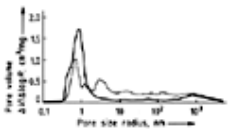

Figure 38. Pore size distribution of different activated carbons

—— Gas-phase activated carbon (narrow pores); · · · · · Liquid-phase activated carbon (wide pores)

The term “adsorption“ refers to the accumulation of gaseous or dissolved components on the surface layer of a solid (the adsorbent) [373], [375-379]. On their surface activated carbons have “active sites“ where the binding forces between the neighboring carbon atoms are not fully saturated and adsorption of foreign molecules takes place. Depending on the intensity of interactions between adsorbent and the component to be adsorbed, an adsorption enthalpy of 5 – 65 kJ/mol is released. With increasing temperature the quantity of the matter adsorbed in the equilibrium state decreases; in gas phase the influence of temperature on adsorption capacity is greater than in liquid phase. The reverse of the adsorption process is called desorption.

The adsorption mechanisms are classified as chemisorption and physisorption. Physisorption is reversible and involves only physical interaction forces (van der Waals forces). Chemisorption is characterized by higher interaction energies which result in a chemical modification of the adsorbed component.

The adsorption capacity (loading) of an adsorbent for a given component is normally represented as a function of the component's concentration c in the gas (or liquid) for the equilibrium condition at constant temperature, known as the adsorption isotherm $x = f(c)T$.

There are a variety of approaches derived from different model assumptions for the quantitative description of adsorption isotherms (see also ⟶ Adsorption – Thermodynamics, Equilibria, and Heat of Adsorption, ⟶ Air). The Langmuir isotherm is based on the assumption of ideal monolayer adsorption. The BET isotherm additionally takes into account multilayer adsorption. However, their underlying assumption of a homogeneous surface structure is not applicable to adsorbents whose structure is characterized by active sites with widely differing bonding energies. In such cases, the empirical Freundlich isotherm is often useful. According to the Freundlich isotherm, the logarithmic adsorbent loading increases linearly with the partial pressure of the component to be adsorbed in the carrier gas.

However, commercial adsorbents do not have a smooth surface but are highly porous solids with a very irregular and rugged inner surface. This fact is taken into account by the potential theory which forms the basis of the Dubinin isotherm. It describes the logarithm of the adsorpt volume V, which is determined via the density of the liquid phase of the component to be adsorbed as a function of the relative saturation $(\log p/p_s)^n$. According to [376], [377], this equation is applicable to some wide-pore activated carbons with an exponent of $n = 1$, while an exponent of $n = 2$ results for a variety of activated carbons with organic vapors and certain carbon molecular sieves.

At adsorption temperatures below the critical temperature of the component to be adsorbed, the adsorbent pores may fill up with liquid adsorpt. This phenomenon is known as capillary condensation and enhances the adsorption capacity of the adsorbent. Assuming cylindrical pores, capillary condensation can be quantitatively described with the aid of the Kelvin equation, the degree of pore filling being inversely proportional to the pore radius.

In the liquid phase, the empirical Freundlich isotherm is a very helpful tool, as the position and gradient of this isotherm allows conclusions to be drawn regarding carbon properties in practical applications. Often, specific substances like phenol, iodine, alkylbenzenesulfonate, or methylene blue are used for characterizing the adsorptive properties of activated carbons.

For commercial adsorption processes not only the equilibrium value, but also the rate at which it is achieved (adsorption kinetics) is of decisive importance. The adsorption kinetics are determined by the following series of individual steps:

- Transfer of molecules to the external surface of the adsorbent
- Boundary layer film diffusion
- Diffusion into the particle
- Actual adsorption step

Adsorption kinetics in the liquid phase are much slower than in gas phase; therefore adsorber design is different for liquid and gas phase applications (see Table 16).

Table 16. Typical design data of adsorbers for adsorptive gas and water purification

Parameter	Gas purification	Water purification
Carbon particle size, mm	3 – 5	0.5 – 2.5
Depth of adsorbent bed, m	0.5 – 1.5	2 – 15
Mass transfer zone, m	0.05 – 0.3	0.5 – 5
Superficial velocity, cm/s	10 – 50	0.03 – 0.4 *
Residence time, s	1 – 15	1800 – 7200 **

*1 – 15 m/h.
**0.5 – 2 h.

5.2.5. *Quality Control*

Methods of quality control are of great importance for both technical and commercial practice. Some tests are used in production control to check the basic properties of final or intermediate products. Special requirements are placed on acceptance tests, since they may be used in disputes between buyer and supplier. The buyer wishes to be certain of the material's suitability for his purpose and uses tests in which the conditions simulate as closely as possible his operational conditions. For example, active carbon for sugar refineries is tested for its decolorizing efficiency on a molasses solution to

determine the adsorption properties but also side effects due to the ash content and pH value. A gas-mask carbon is tested dynamically for its service time with the substances that are to be removed in practice [365], [368].

Physical and Mechanical Tests [365]. *Bulk density* (DIN-ISO 787 11, ASTM D2854) is defined as the mass of a unit volume of the sample in air, including both the pore system and the voids between the particles. It is expressed in kg/m^3 on a dry basis. The bulk density of activated carbon depends on the shape, size, and density of the individual particles. Bulk density data are useful for the estimation of tank or packing volume.

The *particle density*, otherwise known as Hg density, is defined as the mass of a unit volume of the carbon particle, including its pore system, normally expressed in g/cm^3. The particle density is an important characteristic of granular carbon that is used for the determination of bed porosity or void fraction. This in turn, is necessary for the determination of numerous other properties. Under the conditions of this method, pores $r \geq 7$ µm or more in radius are filled and therefore do not contribute to the density of the particle.

The *absolute* or *helium density* is defined as the mass of a unit volume of the solid carbon skeleton that is inaccessible to He, normally expressed in g/cm^3.

The *pressure drop* gives information about the resistance to flow of a gas through a pelleted or granular carbon layer. The pressure drop over a packed bed is adequately defined by a modified form of the semi-empirical Ergun equation as a function of the shape and size of the particles and the temperature, pressure, and superficial velocity of the gas. It is expressed in pascals per meter of carbon bed length. Air is passed through an activated carbon layer of given length, and the pressure drop is measured as a function of gas velocity. The resistance to the flow of liquids or gases has a practical significance, since it is related to the filtration resistance of a carbon layer. The relationship between flow rate and pressure drop is shown in Figure 37.

Particle size is an important property that influences the flow characteristics, adsorption kinetics and catalytic behavior of granular activated carbon layers. The grain size distribution of granular activated carbons (ASTM D2862) is determined by using standard sieves and motor-driven sieving apparatus. All common methods provide for mechanical separation with standard sieves, the aperture of which is expressed in millimeters or mesh. Determination of particle size by sieving is not applicable to extruded activated carbons. The fineness of powdered carbons can be determined by elutriation or by laser-beam scattering. Equipment such as the Coulter counter or sedimentation tests lead to incorrect results because the individual particles of activated carbon vary in electrical conductivity and density.

Mechanical strength is an important factor in most technical applications of granular activated carbon. Mechanical strength tests are modified to conform with the technical requirements: it differs for active carbon for gas masks, in which the granules suffer attrition, from that for solvent recovery, for which coherence of the granules is critical. The determination of mechanical strength simulates the resistance to abrasion or attrition under practical conditions. Depending on various practical requirements, there are many different test methods using vibration, impact, rotary motion, or motion as in a fluidized bed. A variety of tests are available for the evaluation of the mechanical strength of granular activated carbon. In these tests the change in particle size distribution or the amount of fines produced is determined. An overview of the wide range of hardness tests, which test different aspects of the mechanical strength and hence cannot be related to one another mathematically is given in the following:

Ball-mill hardness:	The activated carbon is abraded for a given time in a horizontal cylinder with steel or ceramic balls under prescribed conditions
Abrasion strength:	The activated carbon is abraded by an iron rod in a horizontal rotating cylindrical sieve of given dimensions for a prescribed time
Impact hardness:	The activated carbon particles are broken by dropping a weight onto a sample under controlled conditions
Ball-pan hardness (ASTM D 3802):	The activated carbon is shaken for a given time in a pan together with a given number of steel balls of known diameter
Crushing strength:	The pressure required to crush a granule of activated carbon (not applicable for broken granules)
Impact hardness (fluidized bed):	The activated carbon is pneumatically agitated for a standard time in a vertical cylinder, the top of which is equipped with an impact plate.
Attrition and abrasion resistance (ASTM D4058):	The activated carbon is rotated for a period of time in a cylindrical drum having a single baffle
Stirring abrasion (AWWA B604):	The activated carbon is abraded by a T-shaped stirrer in a special abrasion unit
Ro-Tap abrasion (AWWA B604):	The activated carbon is shaken with steel balls in the testing pan of a Ro-Tap sieve machine

Chemical and Physicochemical Tests. *Moisture content* (ASTM D2867) is determined by heating the sample in air in an oven at constant temperature to constant weight (3 h at 150 °C). The oven-drying method is used when water is the only volatile material present in the activated carbon. A xylene distillation method is used when the carbon is known or suspected to be heat-sensitive or to contain water-insoluble volatile compounds as well as water.

Ash content (ASTM D2866) is determined by ignition of the sample to constant weight in a muffle furnace (air circulation) at 650 ± 25 °C. Frequently only the water-soluble or acid-soluble part of the ash is determined. In cases where certain cations or anions have an adverse effect on the products, they can be determined by first extracting with hydrochloric acid, nitric

acid, or water in a prescribed way, and then analyzing the extract. Arsenic, cyanide, and sulfide must be determined on the activated carbon itself.

Volatile matter is determined by heating the sample at 900 °C for 7 min. The percentage of volatile matter is calculated from the weight loss of the sample, corrected for moisture content.

Ignition temperature (ASTM D3466) is determined by exposing a sample of carbon to a heated air stream, the temperature of which is slowly increased until the carbon ignites. The temperature of the carbon bed and of the air entering the bed are recorded, and ignition is defined as the point at which the carbon temperature suddenly rises above the temperature of the air entering the bed. The test provides a basis for comparing the ignition characteristics of different carbons, or the change in ignition characteristics of the same carbon after a period of service.

The *self-ignition test* (IMDG Code, Class 4.2) determines whether a sample of chemically activated carbon ignites at 140 °C under specified conditions. Steam-activated carbon is considered to be non-self-ignitable.

Activated carbon bearing inorganic and chemically active groups on its surface may alter the pH of liquids to which it is added. A predictive standard test giving, a good approximation of actual conditions has been devised. The acid or alkali content is determined by boiling an aqueous suspension and measuring the pH of the filtered or decanted aqueous extract. The pH of the extract is defined as the *pH value* of the activated carbon.

Adsorption Measurements. The most common method of measuring adsorption properties of activated carbon is by the determination of the BET surface, e.g., according to DIN 66 131. However, the numerical value thus obtained has only a limited practical significance, because in a practical situation, the molecules adsorbed are usually to large to reach the inner surface of the very small pores due to the large size of their molecules, whereas determination of the nitrogen isotherm [381] or assessment of porosity by using mercury vapor [382] gives deep penetration into the pore structure. It is easier to estimate the porosity by measuring the amount of benzene or cyclohexane vapor adsorbed by activated carbon at 20 °C and at a range of partial pressures. The difference between two measured values of the isotherm represents a certain pore volume, and limiting values of pore sizes can be related to this. An example of the single-point method is the determination of the carbon tetrachloride retentivity according to ASTM [383].

Carbons for gas masks are characterized by the breakthrough times (holding times or service times), which are determined by using certain test substances such as chloropicrin.

For many practical applications involving gas-phase adsorption, the total adsorption capacity is of minor interest in comparison with the adsorption capacity after regeneration. In such cases, the “working capacity” after several cycles of adsorption and desorption is quoted.

5.3. Production

5.3.1. *General Aspects*

Nearly all carbon-containing materials can be used for the manufacture of activated carbon; e.g., wood, nut shells, fruit stones, peat, charcoal, brown coal, lignite, bituminous coal, mineral oil products, and some waste materials. Cellulose and organic polymers are used for the manufacture of fiber and spherical activated carbon [366-370]. These starting materials vary considerably with regard to the extent to which they can be activated; e.g., calcined petroleum coke or high-temperature coke from coal are difficult materials, while wood charcoal is easily activated. In addition, the purity of the activated carbon produced as well as its pore size distribution is very much dependent on the starting material.

There are two principal methods of activation, i.e., that which uses chemicals and that which uses gases. *Chemical activation* is based on the dehydrating action of certain substances, e.g., phosphoric acid [7664-38-2] or zinc chloride [7646-85-7], mostly on uncarbonized starting materials such as sawdust or peat. A temperature of 400 – 1000 °C is usually used. After removal of the chemicals, e.g., by extraction, the porous and active carbon structure of the raw material remains. Cellulose fibers or woven materials produce products having the same physical form, but made of activated carbon. Polyacrylonitrile fibers are also used.

Gas activation entails the use of gases containing combined oxygen, such as steam or carbon dioxide. At temperatures of 800 – 1000 °C, some of the carbonaceous starting material is decomposed, producing numerous extremely fine pores or cracks. The inner surface area of the carbon determines its degree of activation. The yield is, therefore, dependent on the degree of activation. A high degree of activation is associated with a low yield, and this can be between 20 and 60 %.

Production of activated carbon can be carried out in rotary kilns, multiple hearth furnaces, or furnaces of the vertical-shaft or fluidized-bed type, each type being suitable for a particular particle size of starting material. At the present time, internally heated rotary kilns are most commonly used, since these are suitable for production of activated carbon of a large range of particle sizes from powder and granular material up to cylindrical pellets. The activation of coarse materials such as softwood and beech wood charcoal is carried out in a shaft furnace. The product is usually ground to a powder for use as a decolorizing agent. Fluidized-bed furnaces are suitable for granules and cylindrical pellets.

Important factors that determine the adsorption properties of an activated carbon are the pore volume, the pore size distribution, and the type of functional groups on the surface (surface oxides). The large inner surface area is mainly due to the micropores. Consequently, particular attention is paid to these during manufacture. Pore volume and pore size are affected by the type of starting material and the heating process. Thus, gas activation of coconut shell charcoal always gives a high proportion of fine pores, while the same process with softwood charcoal yields a product with open pores. Chemical activation produces carbons with extremely high proportions of meso- and micropores.

5.3.2. *Raw Materials*

Wood, sawdust, peat, straw, and other cellulose-containing materials are usually only treated by chemical activation.

The direct gas activation of uncarbonized products is possible, although usually with these raw materials, a carbonization process (possibly at low temperature) is added. This has the advantage that the partly carbonized intermediate product can be screened, giving a standard particle size. Norit NV (Netherlands) carries out gas activation of peat without producing any partially carbonized intermediate material. In a further development of the classical procedure of OSTREJKO [371], activated carbon is produced directly from uncarbonized carbonaceous material by heating to 840 – 900 °C in an atmosphere of combustion products that contains hydrocarbons but no oxygen [372].

Raw materials very suitable for gas activation are wood charcoal, nut shell charcoal, and coke from brown coal or peat. *Wood charcoal* [7440-44-0] is obtained by carbonization of pieces of wood from beech, spruce, or pine in large-capacity batch retorts or continuous vertical retorts (⟶ Charcoal). Carbonized coconut shells are used by many producers in Europe, Japan, and the Far East.

Bituminous coals are suitable for gas activation to varying extents. A useful criterion is their carbon, oxygen, and hydrogen contents. The diagram in Fig. 39 gives an approximate guide to these values [384]. The high-carbon materials graphite and anthracite are very difficult to activate. Bituminous coals with high oxygen and hydrogen contents can have troublesome sticking and swelling properties. In these cases, a preoxidation is usually carried out before gas activation [385], [386]. This is done with air or oxygen-containing gases at temperatures between 150 and 350 °C. This causes 5 – 30 % oxygen to be taken up. The optimum temperature is 220 – 250 °C. This procedure is carried out as long as necessary to cause complete disappearance of the swelling and sticking properties; as much as 5 h may be required.

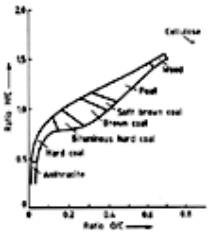

Figure 39. Characterization of different raw materials

Bituminous coal may be more easily activated after addition of mineral acids such as phosphoric acid. For this purpose, the raw material is finely ground and mixed with a few percent of acid before being formed into shapes. This process, which was developed by Carborundum [387], is operated in the United States by Ceca. It uses a combination of chemical and gas activation. In some types of bituminous coal, the quality of the activated product is adversely affected by the high ash content; however, a small ash content that includes potassium compounds, for example, can have a catalytic effect and, when the material is preoxidized, can lead to intensive gasification of the coal. In these cases the ash content of the raw material can be reduced by grinding and flotation. The preoxidation can be carried out on the powder or after granulation. Some companies which activate bituminous coal: the Calgon Carbon Corp. (United States), Norit (United States), and the CarboTech Aktivkohlen GmbH (Germany).

Brown coals and the related lignites offer an inexpensive and readily activated starting material for the production of activated carbons. Their relatively high ash and sulfur contents are detrimental. Therefore, methods have been developed for removing ash from brown coal before coking, e.g., by treatment with an oil – water mixture. The ash goes into the aqueous phase, while the coal remains in the oil phase. By this means the ash content can be reduced by 80 – 90 %. Lignite is activated in high tonnages by Norit in the United States according to a process developed by ICI United States [388].

Many publications and patents describe the production of activated carbon from *mineral oil products*. Petroleum sludges, fly ash, old tires, domestic refuse, and sewage sludge have also been proposed as raw materials for activated carbon, but none of these products have attained commercial importance [389], [390]

5.3.3. *Activating Furnaces*

Shaft furnaces originally consisted of simple vertical chambers with smooth walls made of refractory bricks. Heating is external. As the process was developed, the mixing of the feed materials was improved and, consequently, so was the reaction with the activating gas. This was done by installing ceramic attachments in the form of gratings or replaceable louvers which can control the direction and velocity of the gas stream within the furnace. The reaction gases (hydrogen and carbon monoxide) can be removed at various levels. The temperature of the furnace, usually 5 – 8 m high, can be controlled by means of a number of burners or afterburners (Fig. 40) [391]. Shaft furnaces can be used for the reactivation of exhausted activated carbon [392].

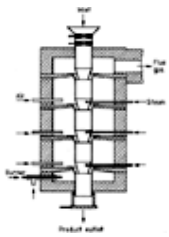

Figure 40. Shaft furnace

Rotary kilns are the most commonly used activating furnaces. Due to the length of the kiln and the high temperatures necessary for gas activation, direct heating is the only feasible method when the materials of construction of the kiln are

considered. To be able to control gas composition and temperature throughout the entire length of the kiln, several burners and gas supply lines are distributed along and around the kiln casing, e.g., in the American Norit furnace. Figure 41 shows an arrangement including lifters to give improved mixing of the feed material. By means of a variable steam injection rate, the water vapor content and, therefore, the activation rate can be further controlled.

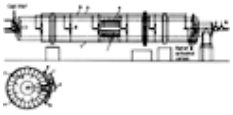

Figure 41. Rotary kiln for steam-activation process

a) Steam; b) Gas; c) Air; d) Burner; e) Brick lining; f) Lifters

Multiple-hearth furnaces (Fig. 42) with rotating arms and stationary floors on each stage are used by several firms such as Calgon in the United States and Belgium. This type of furnace is also operated for the purpose of reactivation, e.g., by Windhoek Municipality (Republic of South-West Africa).

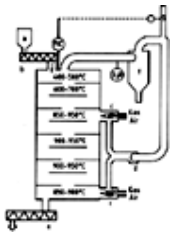

Figure 42. Multiple hearth furnace

a) Raw material silo; b) Inlet; c) Burner; d) Off-gas suction; e) Outlet for activated carbon; f) Dust collector; g) Off-gas stack

Fluidized-bed furnaces offer the advantage of extremely intensive heat and mass transfer. This means not only that the activating gases are quickly brought into contact with the raw material, but also that the waste gases are just as quickly removed. Furnaces have been developed which are operated continuously (Fig. 43) and in which several fluidized beds are run in series. The activating gases may be introduced into the spaces between the fluidized layers and also into the circulatory system. Thus, it is possible to achieve a preoxidation in the first stage with oxygen-containing gas and then to carry out the actual activation with oxygen-free gases in a second stage. The activation in the fluidized bed is so intensive that usually only fragile products with poor resistance to abrasion are obtained, which are processed to give decolorizing carbon in powder form. By maintaining certain conditions, in particular by keeping the water vapor content of the gas below 0.6 kg/m^3 and by maintaining a neutral or slightly reducing atmosphere, it is possible to produce abrasion-resistant granules.

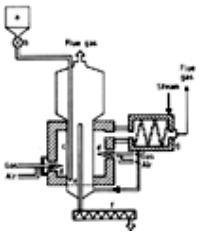

Figure 43. Fluidized-bed furnace

a) Raw material silo; b) Inlet; c) Combustion chamber (indirect heating); d) Burner; e) Gas distribution plate; f) Outlet for activated carbon; g) Heat exchanger

5.3.4. *Methods of Activation*

5.3.4.1. *Chemical Activation*

Zinc Chloride Process. In the classical zinc chloride process for the chemical activation of carbonaceous materials, 0.4 – 5.0 parts of zinc chloride as a concentrated solution are mixed with 1 part peat or sawdust. The mixture is then dried and heated to 600 – 700 °C in a rotary kiln. The product is washed with acid and water, and the zinc salts are recovered. In some cases, chemical activation is followed by steam activation to obtain additional fine pores. In spite of the efficiency and simplicity of the process, it is in decline because of the problems of environmental contamination with zinc compounds.

Phosphoric acid [7664-38-2] can be used to treat either uncarbonized or carbonized raw materials, and the process is operated by Ceca (France), Hooker (Mexico), and Norit (United Kingdom and United States). Finely ground raw material such as sawdust is mixed with a phosphoric acid solution, forming a pulp. This is dried and heated to 400 – 600 °C in a furnace such as a rotary kiln. The phosphoric acid is then extracted, sometimes after neutralizing it to give phosphate salts, and the material is dried, giving an activated carbon which usually has finer pores than the zinc chloride product. Activation with a combination of phosphoric acid and steam is also possible.

As with zinc chloride activation, a highly active decolorizing carbon is obtained by a rapid process in high yield and at a relatively low reaction temperature. However, the cost of recovering the activating chemicals is high. Activation by phosphoric acid has become more popular and there is no doubt that improved methods of phosphoric acid recovery have contributed to this. These innovations have hardly been reported in the literature; the know-how is not divulged by the producers.

Other Chemicals. In the literature, many chemicals have been proposed for the activation of carbonaceous raw materials, but none of them attained industrial importance.

5.3.4.2. *Gas Activation*

In gas activation, carbonaceous material is treated at elevated temperatures with suitable gases, the most common being steam, carbon dioxide, and mixtures thereof. Experiments using graphite have established the reaction velocities: steam has

been shown to be 8 times as reactive as carbon dioxide. Both gases behave as mild oxidizing agents at 800 – 1000 °C, there being several simultaneous reactions:

$$H_2O + C \rightarrow CO + H_2 \quad \Delta H = +117\ \mathrm{kJ}$$

$$2H_2O + C \rightarrow CO_2 + 2H_2 \quad \Delta H = +75\ \mathrm{kJ}$$

$$CO_2 + C \rightarrow 2\,CO \quad \Delta H = +159\ \mathrm{kJ}$$

Due to the endothermic character of these reactions, the carbon particles must be brought into intimate contact with the activating gas. This must be hotter than the required reaction temperature; otherwise, the necessary heat energy will not be provided, or only with difficulty. Below 800 °C, the reaction velocity is so seriously reduced that the activation process ceases for all practical purposes. A useful improvement to the heat supply can be obtained by combustion of gases produced during activation:

$$CO + 0.5\,O_2 \rightarrow CO_2 \quad \Delta H = -285\ \mathrm{kJ}$$

$$H_2 + 0.5\,O_2 \rightarrow H_2O \quad \Delta H = -238\ \mathrm{kJ}$$

Modern furnace construction takes advantage of this fact by introducing oxygen and air at suitable points, which at the same time has the effect of regenerating the activating gases. Thus, the best way of dealing with the carbon monoxide and hydrogen that are produced is by burning them off in the reactor itself. This is necessary for the additional reason that these gases reduce the velocity of activation, carbon monoxide to a noticeable extent and hydrogen very markedly. The reaction mechanism for the gas activation of carbon with steam or carbon dioxide is characterized by an initial adsorption of these gases with subsequent oxidation of the carbon surface as the rate-determining step:

$$C + CO_2 \rightarrow C(O) + CO$$

$$C(O) \rightarrow CO$$

where C(O) signifies surface oxide.

The retarding action of the carbon monoxide and hydrogen can be attributed to the formation of C(CO) and C(H) surface complexes, the latter in particular being much more stable than the C(O) surface complex. Thus, the active sites which could adsorb oxygen are blocked by hydrogen.

Oxygen or air are unsuitable as activating gases. In a mixture with steam or inert gas, small amounts of oxygen lead to activated material with very large pores. Oxygen reacts with carbon about 100 times as fast as carbon dioxide. This reaction velocity is even further increased by potassium salts, so that potassium-containing raw materials react so vigorously when oxygen is present in the gas that an uncontrolled combustion takes place without producing activation. The chemical condition of the carbon surface, especially the presence of larger or smaller amounts of carbonyl and carboxyl groups, can determine the adsorption properties of the activated carbon and, very importantly, its properties as a catalyst.

Acidic surface oxides are formed by heating activated carbon in air or oxygen for a short period below the ignition temperature. If carbon is first heated to 1000 °C and then allowed to react with air at room temperature, basic surface oxide groups are formed, although the quantity of these is at most much less than the quantity of acidic groups.

It has long been known that the gas activation of carbonaceous materials is accelerated by small amounts of various compounds, e.g., salts of alkali and alkaline earth metals, almost all chlorides, sulfates, acetates, and carbonates, as well as most acids and hydroxides. The most important catalysts used industrially are caustic potash and potassium carbonate. Amounts between 0.1 % and 5 % are used. The activation accelerators can be used in solid form mixed with the finely powdered carbonaceous substances or added as solutions, sometimes followed by molding into shapes and low-temperature carbonization.

If bituminous coal is activated with addition of alkali metal salts, the gas mixtures containing carbon dioxide must be used for the activation process in preference to pure steam.

In addition to these accelerators, the patent literature also refers to compounds of iron, manganese, and aluminum.

The theory of catalytically accelerated activation is described in [394] and [395].

5.3.5. *Granular and Pelletized Carbons*

For a number of applications in liquid and gas purification, the activated carbon must be provided in the form of grains or granules; therefore, the raw material used to produce it must be similarly shaped, e.g., wood charcoal or coconut shell charcoal. Alternatively, the raw material, such as coal, can be pulverized, briquetted by using a binder, and finally carbonized. Before activation, the material is broken down to the required particle size. For applications requiring a carbon with a high mechanical strength, it is often an advantage to use carbon which has been specially preshaped, i.e., pelletized. In some cases, the hardness is dependent not only on the raw material and binder, but also on the degree of activation and porosity. Highly activated carbons have a high pore volume and, therefore, low density and strength. Similarly, products with open pores are not as hard as those with fine pores.

Carbons in pellet form are produced from finely powdered raw material as follows. The powder is first mixed with a binder in a heated paste mixer to give a flowable mass. This material is then extruded to form strands ≥ 1 mm in cross section, or may be formed discontinuously in a cylinder press (Fig. 44). These strands, sometimes after drying, are then broken into short

lengths, the length of each piece being approximately equal to its diameter. These are then either chemically activated or carbonized at 400 – 500 °C and finally gas-activated. Suitable binders are coal tar, wood tar, lignosulfonic acids, or mixtures of phenols and aldehydes or their condensation products. Bases such as caustic soda or chalk neutralize the acidic groups of the tar and improve gas activation.

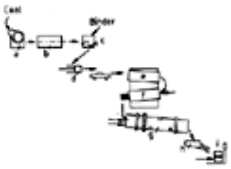

Figure 44. Flow sheet for production of pelletized activated carbon

a) Crusher; b) Mill; c) Kneader; d) Extrusion; e) Drying; f) Carbonization; g) Activation; h) Screening; i) Packing

A simplified flow sheet of the CarboTech process is shown in Figure 45. The specially de-ashed hard coal feed is finely ground, and partially oxidized with air. Then the coal dust is mixed with a binder. The binder is added to produce a plastic coal/binder mixture that can be used to form shaped extrudates of the desired diameter. The extrudates are then carbonized to activated coke in a rotary kiln at 900 – 1000 °C. In a final activation step, the carbon skeleton of the activated coke is partially gasified by steam activation in a multistage fluidized-bed furnace.

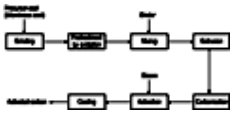

Figure 45. Production steps of formed activated carbon

The production of pelletized activated carbon from material which is already in an activated state, such as powdered decolorizing carbon, is at the present time of little commercial importance. In the beverage and food industries, products of this type have found increased application due to their dust-free nature. To make the fullest use of the available adsorption capacity of agglomerates of this type in an aqueous medium, it is necessary to disintegrate them.

There are a number of patents describing the production of microporous carbon spheres from pitch. The process involves several stages: melting, dispersing, oxidizing with air to render the material insoluble, and finally, activation by steam.

5.3.6. *Carbon Molecular Sieves*

Carbon molecular sieves for air separation are produced from coal, coconut shells, or resins. For example, bituminous coal is ground to a fine dust and oxidized in air at a temperature below the ignition temperature. The oxidized coal is mixed with a binder and shaped into pellets with diameter of 2.5 mm. The pellets are carbonized in a special rotary kiln.

The final step is treatment with hydrocarbon under cracking conditions to deposit carbon exactly on pore openings [396]. The pore diameter is then smaller than that of the initial material, and the diameter of the “bottle necks“ are on the same range as those of nitrogen and oxygen molecules. Oxygen molecules can penetrate much quicker than the nitrogen molecule into the pores. Therefore, the most of the nitrogen is recovered, while almost all of the oxygen is adsorbed.

5.3.7. *Further Treatment*

Many firms produce low-ash varieties of activated carbon by removal of various impurities by washing with water or acids such as hydrochloric or nitric acid. If the activated carbon is to be used for the production of fine chemicals or pharmaceutical preparations, a particularly thorough washing procedure is required, and activated carbons for use as catalysts or catalyst support require similar treatment. Basic constituents and accelerators such as caustic potash may be neutralized by acid or washed out with water. Active carbons with fine pores, made by steam activation at 800 °C, can be further activated by air at 500 – 600 °C in the presence of alkali, giving improved decolorizing ability. However, this two-stage activation process has no known practical importance. The most important result of a postoxidation of activated carbon by nitrous gases, in particular nitrogen dioxide, is the formation of additional surface oxides rather than any further activation. The opposite effect, i.e., the removal of chemically bound oxygen, is possible by treating activated carbon with hydrogen at 200 – 500 °C. Lowering of sulfur content can be achieved by the action of steam and hydrogen. In the literature a method is described of producing activated carbon with a low iron content. This involves converting the iron into volatile compounds by treating the carbon while it is still hot with suitable gases or vapors such as halogens, halogen compounds, or carbon monoxide.

5.3.8. *Impregnation*

For cost-effective removal of certain impurities contained in gases (e.g., hydrogen sulfide, mercury, and ammonia), the adsorption capacity and removal rate must be substantially increased by impregnation of the activated carbon with suitable chemicals. When these chemicals are deposited on the internal surface of the activated carbon, the removal mechanism also changes. The impurities are no longer removed by adsorption but by chemisorption [397].

For the manufacture of impregnated activated carbon, an activated carbon of suitable quality for the particular application is impregnated with solutions of salts or other chemicals which, after drying or other aftertreatment steps, remain on the internal surface of the activated carbon. As well as soaking impregnation, spray impregnation can be used. In that case the activated carbon is sprayed in a rotary kiln or in a fluidized bed under defined conditions. The impregnated wet activated carbon must be dried in, for example, a rotary kiln or fluidized-bed drier. After the drying step, most impregnated activated carbons can be used industrially. In some applications the impregnating agents are present in the form of hydroxides, carbonates, chromates, or nitrates and must be subjected to thermal aftertreatment at higher temperatures (150 – 400 °C) to decompose the anions. Depending on the application, various activated carbons (pellets, granules, powders) are impregnated with suitable organic or inorganic chemicals. Homogeneous distribution of the impregnating agents on the internal surface of an

activated carbon is important. Furthermore, blocking of the micropores and macropores must be avoided so that the impregnating agent remains accessible for the reactants.

5.4. Applications

Since the range of applications for carbonaceous adsorbents in the gas and liquid phases is very broad only a small selection can be dealt with here. The major applications are in water treatment, gas purification, food processing, gold recovery, and solvent recovery (Fig. 46).

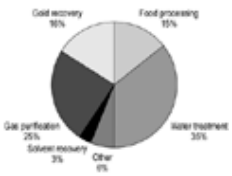

Figure 46. Gas- and liquid-phase applications of carbonaceous adsorbents

5.4.1. *Gas-Phase Applications*

The majority of gas- and vapor-phase applications of activated carbon are in process gas purification, air purification, catalysis, flue gas purification, solvent recovery, and automotive emission control, and personal protection (Fig. 47).

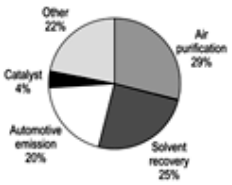

Figure 47. Gas-phase applications of carbonaceous adsorbents

5.4.1.1. *Solvent Recovery*

Activated carbon can be used advantageously for the removal of organic vapors from gases, its performance being good even at very low partial pressures. This may be deduced from the linear adsorption isotherm plotted in Figure 48. For example, a pelletized activated carbon is able to take up ca. 18 % of its mass of toluene from a current of air that contains only 0.11 g of toluene per m^3. This represents an enrichment factor of over 400 000 : 1. This ability to concentrate a substance is particularly useful for producing extremely pure gases, as well as for protecting the environment [366], [398-402].

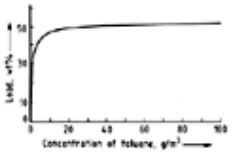

Figure 48. Linear adsorption isotherm for toluene

When activated carbon is used for the recovery of solvents, which usually occurs at concentrations between 1 and 20 g/m^3, efficiencies of > 90 % are sought. The cost of regeneration, whether by steam, hot gas, or electrical heating, is usually small in comparison to the value of the recovered solvent. The charge of activated carbon retains its effectiveness for a long time if the regeneration is carried out at a sufficiently high temperature. However, fine material which is formed by attrition and thermal or chemical stress must be removed and replaced from time to time. A survey of some areas of application of solvent recovery is given in Table 17.

Table 17. Solvent recovery

Industry	Typical solvents
Plastic film and foil	ether, acetone, methyl ethyl ketone, alcohols, methylene chloride, tetrahydrofuran, cyclohexanone
Printing	toluene, petroleum spirit, trichloroethene, *n*-hexane
Metal degreasing	trichloroethane, trichloroethene, tetrachloroethene
Rubber	petroleum spirit, benzene, toluene
Viscose and rayon	carbon disulfide
Dry cleaning	tetrachloroethene, fluorochloro hydrocarbons
Synthetic leather and fibers	alcohol, acetone, hexane, toluene, esters, dimethylformamide
Adhesives	petroleum spirit, hexane, toluene

In industry, the solvent recovery is carried out in vertical or horizontal adsorbers, and by arranging these in parallel a continuous operation is possible by changing over from one to another (Fig. 49) [373]. The gas flow is usually in an upward direction, whereas the desorption with hot steam is in the opposite direction. In Figure 50 a typical temperature – time graph for a cycle is given. Usually, fixed-bed adsorbers have a bed thickness of 0.8 – 2 m. A more recent development is the use of fluidized-bed reactors with finely granulated activated carbon [402]. Another development is the use of activated carbon fibers. Both methods are, however, rarely used.

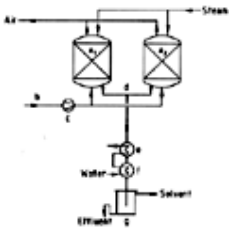

Figure 49. Flow sheet of a solvent recovery unit

a_1) Adsorber 1; a_2) Adsorber 2; b) Exhaust air; c) Blast; d) Desorption; e) Condenser; f) Cooler; g) Separator

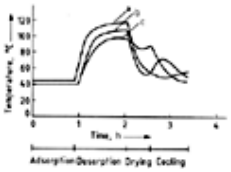

Figure 50. Temperature diagram for the carbon bed of a solvent recovery unit

a) Upper part; b) Middle; c) Lower part

Practical solvent recovery systems use gas flow rates of 0.2 – 0.5 m/s. Lower flow rates would lead to better utilization of the adsorption capacity of the carbon, but there is a danger that the heat of adsorption is not carried away, and therefore overheating and even ignition of the carbon charge can occur. Typical operating data for solvent recovery plants and design ranges are given in the following:

Air velocity	0.2 – 0.4 m/s
Air temperature	20 – 40 °C
Bed height	0.8 – 1.5 m
Steam velocity	0.1 – 0.2 m/s
Time cycle per adsorber	
Adsorption	2 – 6 h
Drying (hot air)	0.2 – 0.5 h
Cooling (cold air)	0.2 – 0.5 h
Solvent concentration	1 – 10 g/cm^3
Solvent adsorbed per cycle	10 – 25 wt %
Steam/solvent ratio	(2 – 5):1
Energy	50 – 600 kWh/t solvent
Cooling water	30 – 100 m^3/t solvent
Activated carbon	0.5 – 1 kg/t solvent

Recovery units have at least two, but more usually three or four adsorbers which pass successively through the stages of the operation cycle. While adsorption takes place in one or more of them, desorption, drying, and cooling are carried out in the others.

5.4.1.2. *Process-Gas and Air Purification*

Many *gas purification* processes use activated carbon, e.g., for production of pure gases in the chemical industry, in protection against poison gas, in air conditioning, for removal of oil from compressed air, and in purification of waste air. Small traces of unwanted gases or vapors are adsorbed onto activated carbon which is often improved by being specially impregnated, in which case regeneration with recovery of the extracted materials is impossible. Activated carbon is also used to remove resin-forming and other hydrocarbons from gases before passing them over sensitive molecular sieves or catalysts [366], [373], [396-401].

Small amounts of hydrogen sulfide can be converted to elemental sulfur in the presence of oxygen by means of activated carbon which has been impregnated with potassium iodide [401]. This reaction is used in the viscose industry for purification of waste air. In this case, it is combined with carbon disulfide recovery and is known as the Sulfosorbon process [401], [403]. The Sulfren process uses sulfur dioxide as an oxidizing agent.

Sulfur compounds such as carbon disulfide, carbonyl sulfide, and organic thiols are removed from moist gases containing excess oxygen at high temperature on alkaline carbon. This is the Desorex process. The adsorbate is hydrolyzed and oxidized to sulfate, which can be washed out.

In *air-conditioning* installations, activated carbon is used for the purification of air drawn in from outside, e.g., in airports, near chemical plants, or for environmentally controlled rooms of hospitals or museums. By the use of activated carbon, the amount of necessary cold outside air can be reduced. Because of the large quantities of air to be handled for such large spaces, a small resistance to flow is required, and therefore, thin layers of carbon are used. Often exchangeable cartridge filters are used, and complete filter elements made of specially shaped activated carbon bodies are available. Composite materials have been produced, in which powdered activated carbon is bonded onto polyurethane foam or some other suitable carrier. For the removal of some substances, impregnated activated carbons like those in industrial respirators are used.

In *nuclear power installations*, activated carbon impregnated with iodine compounds is used to remove radioactive iodine compounds from the air exhausted to the atmosphere. This takes place by isotope exchange. In some countries impregnation with the base tetraethylenediamine (TEDA) is used for the same purpose; in this case the iodine compounds are removed by salt formation. Activated carbon is also used in off-gas delay beds. Deep beds of activated carbon with very fine pores adsorb radioactive gases such as krypton and xenon long enough for the isotopes to decay to safe levels of radioactivity, after which they can be released into the atmosphere.

Since World War I, activated carbon filters have been used by the armed forces for *respirators*, and this use has since been extended to protection against hazardous gases in industry. For both applications, virtually complete removal of impurities is necessary, and therefore, only fine-grained activated carbon can be used so as to make a filter as compact as possible. Table 18 shows how the service time depends on the grain size. For respirators, impregnation of activated carbon with chromium and copper salts is a well-proven technology; due to the toxicity of chromium compounds the latest impregnations are based on copper and molybdenum salts. These substances have a strong oxidizing action and high reactivity toward chlorine, hydrogen cyanide, and their derivatives. Other substances used for impregnation are caustic potash (acidic gases), zinc salts (ammonia and hydrogen sulfide), and iodine compounds (mercury vapor).

Table 18. Effect of grain size on the service time of a respirator fitted with an activated carbon filter (test gas: chloropicrin in moist air)

Diameter of carbon granules	mm	0.8	1.2	1.6	2.5	4.0
Service time	min	75	60	35	18	5

Activated carbon and impregnated activated carbon is also used for many odor emission control applications: wood chip drying (pinene, terpene), plastic processing (styrene, benzene, etc.), home application (kitchen hood, refrigerator). Apart from physisorption, chemisorption on impregnated activated carbon (Section Impregnated Activated Carbon) can be applied, especially for H_2S and mercaptan removal.

In cigarette filters and in attachments for tobacco pipes, activated carbon is used to reduce the nicotine and tar content of the smoke.

5.4.1.3. *Gas Separation*

In 1960 the first pressure-swing adsorption (PSA) plants for gas drying, gas purification and gas separation were built. All PSA processes have in common that adsorption is operated at a higher and desorption at a lower total pressure. In the majority of processes the adsorption pressure is markedly higher than atmospheric pressure. For desorption the pressure is either reduced to atmospheric pressure or vacuum is applied to lower the pressure below atmospheric pressure. In some processes, desorption is boosted by a flushing cycle. The pressure-swing adsorption cycle comprises the following steps:

- Adsorption at higher pressure
- Desorption by pressure reduction
- Flushing with product gas
- Pressure build-up with raw gas or product gas to adsorption pressure

Intervals between these individual steps are quite short (30 s to a few minutes) and depend on the separation process and plant design.

Several types of product gases can be obtained during the absorption or desorption step, see Table 19 [396], [405].

Table 19. Selected application fileds for pressure-swing processes for gas separation

Separation problem	Adsorbent **	Production phase *
Gas drying	Al_2O_3, SiO_2, ZMS	A
Hydrogen from coke-oven or reformer gas	CMS, ZMS, Al_2O_3	A
Helium from diving gases	CMS, ZMS, Al_2O_3	A
Nitrogen from air	CMS	A
	ZMS	D
Oxygen from air	CMS	D
	ZMS	A

Methane from biogas	CMS	A
Carbon dioxide from exhaust gas	CMS, ZMS	D

*A = adsorption phase, D = desorption phase.
**Al_2O_3 = aluminium oxide; CMS = carbon molecular sieves; SiO_2 = silica; ZMS = zeolitic molecular sieves.

For gas separation, carbon molecular sieves or activated carbons with extremely fine pores and molecular sieve properties are used. From a mixture of hydrogen, carbon monoxide, and methane, it is possible to produce highly pure hydrogen by adsorption of CO and CH_4 under pressure (1.5 – 4.0 MPa). If pure hydrogen is then passed through in the opposite direction at lower pressure (0.2 – 0.5 MPa), the adsorbed substances are again desorbed. Since the holding time of an adsorber which operates in this PSA mode is in the range of 30 s to a few minutes, a large number of adsorbers (8 – 16) are used in rotation, so that pressure changes can be minimized between the adsorbers and compression energy can be saved. The method can be used to separate oxygen and nitrogen, as well as for production of synthetic natural gas from biogas (biologically produced methane).

5.4.1.4. *Gasoline Vapor Adsorption*

Gasoline vapors evaporate during manufacture, distribution, refueling, and running of cars and enter the environment. Vapor recovery units are installed at tank farms and distribution terminals of refineries. These waste air streams, saturated with organic vapors, are often cleaned by combined processes [406]:

- Absorption and pressure-swing adsorption
- Membrane permeation and pressure-swing adsorption
- Condensation and adsorption

To avoid pollution of the environment by gasoline vapor from motor vehicles, installation of an activated carbon filter in the ventilation port of the gasoline tank is required in the United States, the EU, and Japan. When a car is left standing in the hot sun, for example, the gasoline that evaporates from the tank is adsorbed by the activated carbon and then desorbed again when the vehicle runs and fresh air for the engine is drawn through the carbon cartridge.

5.4.1.5. *Flue Gas Cleaning*

For SO_2 and NO_x removal from flue gases from power plants and waste incineration plants many activated carbon/activated coke processes have been developed. The adsorption capacity of activated carbon and activated coke for sulfur dioxide from flue gas is only a few percent by weight. Therefore, processes are based on the ability of activated coke to oxidize sulfur dioxide in the presence of oxygen, forming sulfuric acid.

$$2\,SO_2 + O_2 + 2\,H_2O \rightarrow 2\,H_2SO_4$$

The spent activated coke can be regenerated thermally at 400 – 500 °C in a desorber; the carbon of the activated carbon skeleton is used as a reactant.

$$2\,H_2SO_4 + C \rightarrow 2\,SO_2 + CO_2 + 2\,H_2O$$

This carbon consumption results in a larger inner surface area and higher catalytic activity, converting the activated coke to activated carbon. The SO_2-rich gas can be processed to elemental sulfur or sulfuric acid.

The catalytic properties of activated carbon are used for catalytic NO reduction by addition of gaseous ammonia.

$$4\,NO + 4\,NH_3 + O_2 \rightarrow 4\,N_2 + 6\,H_2O$$

Activated coke processes for simultaneous SO_2 and NO_x removal are the Mitsui Mining Process, the Sumitomo Heavy Process, and the Uhde/Bergbau-Forschung/Mitsui Process [407]. In the Sulfacid process, which is designed for SO_2 removal, regeneration by water extraction continuously yields dilute sulfuric acid [401].

In waste incineration plants, powdered activated carbon is injected into the flue gas stream through an atomizer. The loaded adsorbent is collected in a fabric filter; further pollutant removal takes place in the filter cake deposited on the fabric [408].

5.4.2. ***Liquid-Phase Applications***

Liquid-phase applications are estimated to account for over two-thirds of world activated carbon consumption; both granular and powdered activated carbons are in use [366], [368], [369], [399], [410].

There are many processes available for treatment of liquids and solutions with activated carbon which have found wide industrial application. In the *batch contact unit operation*, powdered activated carbon is added to the liquid being treated, either directly or (more usually) as a previously prepared suspension. The temperature is raised to reduce the viscosity and hence also the diffusion time. Equilibrium is normally reached after 15 – 30 min, after which the mixture is filtered, usually after adding a filtration aid such as diatomite.

In *continuous-layer filtration*, the liquid is pumped through a prepared bed of powdered carbon, which is usually made more permeable by the addition of diatomite. Due to the short contact time, the purification process is usually incomplete, and because of the limited amount of carbon present in the bed, the operating life is short. The method is used for liquids with only low levels of impurities or for purification of working liquids which can be pumped so as to bypass the filter, e.g., in

electroplating, in swimming baths, or in dry cleaning machines.

Percolation through granular carbon is used particularly in the field of water purification and decolorization. This continuous process has the advantage that large charges of carbon can be employed, particularly when several filters are arranged in series. Depending on the concentration and viscosity of the solution, the contact time in the percolation process can range from 10 min to more than 1 h. In some applications catalytic side reactions such as oxidation may occur, or changes of pH can be brought about by the ash content of the carbon.

Activated carbon and activated coke are often used in combination with other filter materials in multilayer filters for wastewater and drinking water treatment [411].

5.4.2.1. *Water Treatment*

Water treatment (35 % of world consumption) can be divided into drinking water, industrial and municipal wastewater, and groundwater.

In *drinking water*, activated carbon is used to remove unpleasant odors and tastes and reduce the concentration of compounds constituting a health hazard (pesticides, chlorinated hydrocarbons, etc.).

Powdered activated carbon is often used to solve temporary pollution problems and is added to the water as a slurry at the same time or just before adding of flocculant. After a suitable contact time, the powdered activated carbon is removed with the flocculant by sedimentation.

When granular activated carbon is used to purify drinking water, after a rapid gravity sand filtration (removal of suspended solids) the percolation process is used almost exclusively, and in addition to the adsorptive purification, the catalytic decomposition of the oxidizing agents chlorine, chlorine dioxide, and ozone plays an important role. The filtration velocities are between 5 and 20 m/h, which for a layer depth of 2 – 4 m corresponds to a contact time of 6 – 48 min. The backwashing velocities are 40 – 50 m/h. After a service life of 6 – 24 months, the spent carbon is reactivated in a on-site reactivation plant or in the supplier's reactivation plant.

Figure 51 compares the different treatment steps of drinking water with powdered (PAC) and granular activated carbon (GAC).

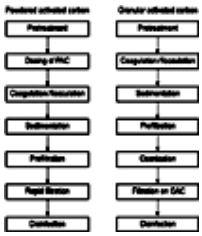

Figure 51. Drinking-water treatment with powdered and granular activated carbon

The complex composition of *industrial wastewaters* and dumpsite leachates mostly requires a combination of various purification methods to obtain the required purity economically. A combination of activated carbon and biological purification is suitable in many cases. Powdered activated carbon (PAC) is added to aerobic or anaerobic biological treatment plants to adsorb toxic contaminants and stabilize the biological activity. There is then no impairment of the biological activity, and the impurity-loaded carbon is removed along with the microorganisms and usually incinerated. PAC is also used independently in treatment plants [412]. Granular activated carbon filter beds are often used as a tertiary treatment after conventional secondary biological treatment. Industrial wastewaters may exhibit several times higher concentrations of organic substances than drinking water. Even though these differences in concentration are accommodated in plant design (contact time 0.5 – 4 h, filtration velocities of 2 – 6 m/h), the adsorption lifetime of the activated carbon filters is reduced from several months to few days. Accordingly, cost-effective use of activated carbon adsorbers only became possible with the development of processes for activated carbon regeneration in the 1980s. *Dump-site leachates* are often purified by biological treatment followed by a filtration or membrane permeation process and an activated carbon step. The main purpose of adsorption on activated carbon is removal of organic halogen compounds.

5.4.2.2. *Micellaneous Liquid-Phase Applications*

Both powdered and granular activated carbon made from a variety of raw materials are used in *food and beverage processing*. Wood-based chemically activated carbons are preferable for the removal of large color bodies and other high molecular mass impurities. Peat- and coal-based steam-activated carbons are used for decolorization and removal of unpleasant tastes, odors, and other low and medium molecular mass impurities. Microporous coconut-shell activated carbons are less efficient in decolorization. Powdered grades are used in batch processes which already require a filtration stage and where the dosage needs to be varied according to different process conditions. The powdered activated carbon is mixed with the liquid to be purified. When the impurities have been adsorbed, the carbon is removed from the solution by filtration or sedimentation, and is then discharged. Granular activated carbons are used where high volumes of liquids of a consistent quality or grade are continuously processed in large quantities. Fixed-bed and moving bed adsorbers are used, and the spent activated carbon can be reactivated on site or by the carbon supplier.

Some products which are purified with activated carbon follow:

Alcoholic beverages	Beer

	Wine
	Vodka
	White rum
	Whisky
Soft drinks	Dechlorination and removal of taste and organic matter from water
	Decolorisation of sugar and fruit juices
Sugar and sweeteners	Cane sugar
	Beet sugar
	Polysaccharides
	Glucose, lactose, maltose, fructose corn syrups
	Xylitol
	Aspartame
Decaffeinated coffee	Caffeine removal by recirculating CO_2 or hot water and adsorption on activated carbon
Edible oils	Coconut oil
	Palm oil
	Fish oil
	Soybean oil
	Olive oil
Flavorings	Yeast extract
	Monosodium glutamate
	Hydrolysed vegetable
	Proteins
Chemicals and pharmaceutical products	Paraffins
	Waxes
	Phosphoric acid
	Sodium hydroxide
	Citric acid
	Gelatin
	Pectin
	Quinine
	Insulin
	Antibiotics
	Sulfonamides

Activated carbon processes have been developed *gold recovery* from low-grade ores. In the carbon-in-pulp process (CIP) a suspension of the ore is treated with cyanide to produce gold cyanide, which is then adsorbed onto granular activated carbon (⟶ Gold, Gold Alloys, and Gold Compounds – Recovery of Gold with Carbon . By using multistage countercurrent adsorption, the gold cyanide complex is concentrated on the activated carbon. The carbon is then separated by mechanical sieving and subjected to a elution process under slight pressure. Gold is recovered from the activated carbon by elution followed by electrowinning. After thermal reactivation the carbon can be re-used in the process. Activated coconut carbon or extruded granular carbon, with high attrition and abrasion resistance, are used.

One of the oldest uses is in *medicine* for the adsorption of harmful bacteria and their metabolic products in the gastrointestinal tract. For blood dialysis treatment in cases of kidney and liver diseases or poisoning, granular activated carbon with a semipermeable coating is used.

5.4.3. *Impregnated Activated Carbon*

Impregnated activated carbon is predominantly used in the following applications:

- Gas purification
- Civil and military gas protection
- Catalysis

For these applications the manufacturers offer various qualities of impregnated activated carbon. Table 20 lists frequently used impregnated activated carbons. A given impregnating agent is frequently used for various purification tasks [397].

Table 20. Commercial grades of impregnated activated carbon [397]

Impregnation

Chemicals	Quantity, wt %	Activated carbon*	Examples for applications
Sulfuric acid	2 – 25	F 1 – 4 mm Ø	ammonia, amine, mercury
Phosphoric acid	10 – 30	F 1 – 4 mm Ø	ammonia, amine
Potassium carbonate	10 – 20	F 1 – 4 mm Ø	acid gases (HCl, HF, SO_2, H_2S, NO_2), carbon disulfide
Iron oxide	10	F 1 – 4 mm Ø	H_2S, thiols, COS
Potassium iodide	1 – 5	F 1 – 4 mm Ø	H_2S, PH_3, Hg, AsH_3, radioactive gases/radioactive methyl iodide
Triethylenediamine	2 – 5	F 1 – 2 mm Ø	radioactive gases/radioactive methyl iodide
		G 6 – 16 mesh	
Sulfur	10 – 20	F 1 – 4 mm Ø, G	mercury
Potassium permanganate	5	F 3 + 4 mm Ø	H_2S from oxygen-lacking gases
Manganese IV oxide		G 6 – 16 mesh	aldehyde
Silver	0.1 – 3	F 3 + 4 mm Ø	F: phosphine, arsine
	0.05 – 0.4	G 8 – 30 mesh	G: domestic drinking water filters (oligodynamic effect)
Zinc oxide	10	F 1 – 4 mm Ø	hydrogen cyanide
Chromium – copper – silver molybdenum salts	10 – 20	F 0.8 – 3 mm Ø	civil and military gas protection
		G 12 – 30 mesh	phosgene, chlorine, arsine
		G 6 – 16 mesh	chloropicrin, sarin, and other nerve gases
Mercury (II) chloride	10 – 15	F 3 + 4 mm Ø	vinyl chloride synthesis, vinyl fluoride synthesis
Zinc acetate	15 – 25	f 3 + 4 mm Ø	vinyl acetate synthesis
Noble metals	0.5 – 1.0	F, G, P	organic synthesis, hydrogenation
(palladium, platinum)	0.5	G 2 – 5 mm	purification of terephthalic acid

*F = pelletized activated carbon, G = granulated activated carbon, P = powdered activated carbon, Ø = pellet diameter.

Potassium iodide promotes the action of activated carbon as an oxidation catalyst and thus allows catalytic oxidation of hydrogen sulfide to sulfur or of phosphine to phosphoric acid. The same impregnation technique is used to extract radioactive methyl iodide and other gaseous compounds arising in nuclear installations. Hydrogen sulfide and formaldehyde can be oxidized to nonhazardous substances with the aid of activated carbon impregnated with manganese dioxide; at high temperatures the formaldehyde is oxidized not just to formic acid, but directly through to carbon dioxide. Impregnation with iron salts and conversion to iron(III) oxide enables the removal of divalent sulfur compounds from gas mixtures low in oxygen. About 1.5 times the stoichiometric amount of oxygen is required to oxidize these compounds [401]. Activated carbon treated with elemental sulfur is used to eliminate mercury vapor from natural gas, hydrogen, and other gases [409]. Silver-impregnated activated carbon is used for purification of drinking water due to its oligodynamic effect.

5.4.4. *Catalysts and Catalyst Supports*

In addition to its adsorption properties, activated carbon has catalytic properties which allow commercial use as catalyst and catalyst support. The catalytic action of activated carbon is due to the crystalline structure of the skeleton, which consists of a mixture of amorphous and graphitic carbon. On the rims of the layers, there are many unsaturated edges and ridges which act as lattice vacancies. On the internal activated carbon surface, there are surface oxides, which have been identified as carbonyl groups, lactone groups, phenolic hydroxyl groups, and carboxyl groups [370], [375]. These surface oxides can participate in redox reactions, and they are a reason for the effectiveness of activated carbon catalysts in oxidation reactions. The type and quantity of surface oxides play a part in such oxidation reactions as the conversion of sulfur dioxide to sulfur trioxide or when activated carbon is incorporated into dry cells in conjunction with manganese dioxide for the depolarization process. Activated carbon serves as a catalyst in the synthesis of phosgene from carbon monoxide and chlorine and in the synthesis of sulfuryl chloride from sulfur dioxide and chlorine. The catalytic activity of the carbon is associated with the π electrons in a similar manner as graphite crystals catalyze the trimerization of cyanogen chloride to cyanuric chloride. In each case, weakening of the double bonds in the reactants occurs. In addition, the industrial process in which hydrogen chloride is removed from chloroethane in the presence of activated carbon presumably has a similar mechanism.

By incorporation of heteroatoms the surface of activated carbon can be modified to achieve very high activity for redox reactions such as the decomposition of chloramines and the oxidation of H_2S or SO_2. When these catalytic/adsorptive carbons are used for removal of sulfur compounds, the end product is sulfuric acid, which can be washed out with water.

Some types of activated carbon are used as catalyst support for industrial syntheses of vinyl acetate and vinyl chloride. Activated carbon is particularly suitable as catalyst support because, in contrast to oxidic catalyst supports, it has no Lewis acidity and therefore does not lead to unwanted polymerization of the vinyl monomers. Vinyl acetate is produced by vinylation of acetic acid with acetylene in a heterogeneous catalytic gas-phase reaction in presence of a zinc acetate/activated carbon catalyst at 160 – 240 °C. Vinyl chloride can be produced by direct catalytic addition of hydrogen chloride to ethyne at 120 – 150 °C in the presence of an activated carbon impregnated with about 10 % of mercury(II) chloride is used as catalyst.

Catalysts for various purposes are obtained by depositing noble metals on activated carbon. A well-known example is palladium-coated activated carbon powder [7440-05-3], a typical hydrogenation catalyst for use in suspension.

In the Merox process for oxidation of mercaptans in mineral oil, activated carbons impregnated with cobalt phthalocyanine [3317-67-7] are used.

5.5. Regeneration and Reactivation

The use of activated carbon for gas and water cleaning in industries and environmental protection is particularly economical if suitable processes for the regeneration of spent activated carbon are available. Generally for the regeneration of the spent carbon pressure swing, thermal desorption, extraction, or thermal reactivation processes are used (Table 21) [366-369], [410].

Table 21. Regeneration of spent activated carbon

Process	Basic concept	Main applications
Pressure-swing adsorption	adsorption at higher pressure, desorption at low pressure	hydrogen purification, air separation (oxygen/nitrogen), methane enrichment from biogas, gasoline vapor recovery
Thermal desorption	steam desorption at 100 – 150 °C, inert gas desorption at 150 – 250 °C	solvent recovery, (toluene, alcohols, hydrocarbons, etc.)
Extraction	elution of organics by solvents or alkali	phenols by alkaline extraction, sulfur extraction with CS_2
Thermal reactivation	selective gasification of carbonized residues and carbon deposits by steam at 800 – 900°C	spent activated carbon from water treatment and other applications

If the adsorbate is bound reversibly onto the surface of the adsorbent the activation energy E_{des} is required for desorption only. For low values of E_{des} (10 – 30 kJ/mol), as in gas separation processes, desorption can be achieved by lowering the pressure in a PSA process.

For adsorptive removal and recovery of solvent, a higher energy of desorption (30 – 60 kJ/mol) is needed for desorption. Usually, superheated steam or hot inert gas is passed through the carbon bed, and this also removes the desorbed substances from the intergranular spaces.

Apart from thermal desorption, extraction with solvents can be used to remove adsorbed substances from used carbons. For example, the elemental sulfur produced in the Sulfosorbon process can be extracted by carbon disulfide. In the literature solvents such as dimethylformamide, acids, and alkali solutions are described [366], [367].

The success of activated carbon in drinking-water and wastewater treatment has been due to the development of regenerable activated carbon types and the development of effective processes for the regeneration of the spent carbon. Spent activated carbon used in water treatment usually contains many different kinds of organic substances. A complete thermal desorption of all adsorbed substances is not possible. The decomposition temperatures of high molecular mass compounds are below the desorption temperature. Thus a residual load consisting of pyrolysis products and carbon deposits remains on the internal surface of the activated carbon. These spent activated carbons are thermally regenerated at 800 – 850 °C in a process that is similar to the production of activated carbon and hence more properly called reactivation. Independent of the reactor type used, the spent activated carbon passes through four main steps during the temperature rise to 850 °C:

≤ 100 °C	water removal
100 – 350 °C	desorption of low-boiling organic compounds
350 – 800 °C	thermal cracking of high molecular mass compounds
800 – 850 °C	gasification of the pyrolysis product and carbon deposits by H_2O and CO_2

Investigations have shown that carbonaceous deposits formed by thermal desorption treatment of the residual load show a higher reactivity due to their disturbed structure than the carbon surface of the activated carbon. Under such conditions, the loaded activated carbon can be selectively treated with the water gas shift reaction at ca. 800 °C, which restores the original surface structure and adsorption performance of the activated carbon. Drying, desorption, thermal cracking, and gasification can be performed in a single industrial reactor. All types of furnaces used for activation of activated carbon can be used for

reactivation. Fluidized-bed reactors, rotary kilns, and multiple-hearth reactors are widely used.

Some methods for regenerating powdered carbons have been proposed in the past, but these have not yet attained widespread practical application. To date, powdered activated carbon has only been reactivated in a few cases.

5.6. Economic Aspects

World consumption of activated carbon was estimated at 450 000 t in 1998, and capacity utilization of the activation plants at 70 – 80 %. Percentage consumption (1998) by region was estimated as United States 37 %, Europe 32 %, Japan 18 %, rest of the world 13 %. World demand was forecast to rise by 2 – 3 % per annum. Areas of potential growth are flue gas purification in Europe and Japan, water treatment in the United States, Europe, and Japan, and automotive emission control.

The regeneration of spent granular activated carbon (Europe: 50 000 – 60 000 t/a) is a growing market in all areas, but increased regeneration will lead in a fall in the rate of growth in demand for virgin material.

Worldwide there are more than 100 activated carbon producers, but the ten largest producers account for about 70 % of world capacity. About 45 % of production is powdered activated carbon and about 55 % is granular and extruded activated carbon (pellets). The largest producers, divided by country, are given in the following:

America	
USA	Anticarb
	Barnebey and Sutcliffe Corp.
	Calgon Carbon Corp.
	Norit America Inc.
	Westvaco Corp.
Brasil	Industrias Quimicas Carbomafra
Mexico	Clarimex SA de CV
	Nobrac Mexicana SA de CV
Europe	
Belgium	Chemviron Carbon
France	Ceca
	Pica
Germany	A.U.G.
	CarboTech Aktivkohlen GmbH
Netherlands	Norit
UK	Norit UK
Asia	
China	Datong Yuanghua Activated Carbon Plant
	Ningxia Huahui Activated Carbon Co.
	Ningxia Longde Activated Carbon plant
	Shanxi Detong Minerals Bureau
	Huai Yu Shan Activated Carbon Group
	Tianjin Anfull Chemical Co. Ltd.
	Zhejiang Shuichang Activated Carbon Co.
India	Indian Dyestuff Industries
	Indo German Carbon Ltd.
Indonesia	PT. Ikaindo
Japan	Futamura Chemical Industries
	Kuraray Chemicals
	Mitsubishi Chemicals
	Sankyo Sangyo
	Takeda Chemical Industries
Malaysia	Century Chemical Works
	Pasific Activated Carbon
Philippines	Cenapro Chemical
	Davao Central Chemical
	Philippine Activated Carbon
	Phileppines Japan Activated Carbon
Sri Lanka	Bieco-Link Carbon
	Haycarb
	Tajit
Taiwan	China Activated Carbon Industries

Taiwan Active Carbon Industries
Carbokarn Thailand

[Top of Page]

6. Carbon Black

Manfred Voll and Peter Kleinschmit

Carbon black [1333-86-4] and soot are formed either by pyrolysis or by partial combustion of vapors containing carbon. Soot as an unwanted byproduct of combustion (e.g., in chimneys or diesel engines) is a poorly defined material. Besides carbon black particles, it often contains significant amounts of ash and large amounts of polycyclic aromatic hydrocarbons (PAH) [413]. Residual hydrocarbons, which can be determined by extraction with solvents (e.g., toluene), can account for 30 wt %.

On the other hand, the term “carbon black” is used for a group of well-defined, industrially manufactured products, which are produced under carefully controlled conditions. The physical and chemical properties of each grade of carbon black are kept within narrow specifications. Carbon black is one form of highly dispersed elemental carbon with extremely small particles. Depending on the production process and the raw materials, carbon black also contains chemically bound hydrogen, oxygen, nitrogen, and sulfur.

Due to its excellent pigmentation properties, especially its light stability and universal insolubility, carbon black has been used as a black pigment since early times. It was produced for this purpose by burning oils, fats, or resinous materials. The flame was either quenched on a cool surface (impingement black) or cooled in special chimneys (lamp black), where the carbon black was deposited.

Both processes are still used for the production of carbon black. The channel black process, a process for making impingement blacks, has been used in the United States since the end of the 19th century. This process, which has now been abandoned because of economic and environmental considerations, used natural gas as raw material. A similar process for the production of impingement blacks, the “Degussa gas black process,” is still used today.

The increasing demand for carbon black led to new production processes. The most important process today is the furnace black process. It was developed in the United States in the 1930s and substantially improved after World War II. It is a continuous process, which allows the production of a variety of carbon black grades under carefully controlled conditions. Nearly all rubber grades and a significant part of pigment-grade carbon blacks are now manufactured by the furnace black process. Nevertheless, other processes, such as gas black, lamp black, thermal black, and acetylene black processes, are still used for the production of specialties.

While carbon black was exclusively used as a pigment until the beginning of this century, its use as an active filler in rubber was the starting point for a new rapidly expanding application. In the production of automobile tires, it was found that treads filled with carbon black had a markedly higher abrasion resistance than those filled with zinc oxide. This discovery, together with increasing use of motor vehicles, was the basis for the present importance of carbon black as a filler in rubber.

Today at least 35 different grades of carbon black are used as fillers in rubber, and about 80 are used in pigments or special applications. The total world production in 1994 was 6×10^6 t, of which ca. 90 % was used in the rubber industry.

6.1. Physical Properties

Morphology. Electron micrographs show that the primary particles of carbon black are almost spherical. In general, a larger number of such primary particles builds aggregates in the form of chains or clusters. In practice, the degree of aggregation is called the “structure” of carbon black. These aggregates tend to agglomerate.

The mean primary particle diameter, the width of the particle size distribution, and the degree of aggregation can be varied within relatively wide ranges (Figs. 52-54) by varying the production process and several process parameters.

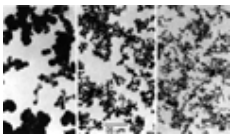

Figure 52. Carbon blacks of different primary particle sizes and specific surface areas

From left to right: lamp black (mean primary particle diameter 95 nm, BET surface area 21 m^2/g), furnace black (27 nm, 90 m^2/g), finely divided gas black (13 nm, 320 m^2/g)

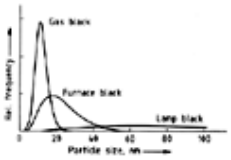

Figure 53. Particle distribution curves for the carbon blacks of Figure 52

Figure 54. Furnace blacks of different aggregation degrees

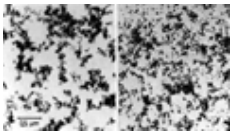

The diameter of the primary particles ranges from 5 to 500 nm. Diffraction patterns produced by the so-called phase-contrast method in high-resolution electron microscopy show that the spherical primary particles are not amorphous (Fig. 55). They consist of relatively disordered nuclei surrounded by concentrically deposited carbon layers [414]. The degree of order increases from the center to the periphery of each particle, a phenomenon important to the understanding of the chemical reactivity of carbon black.

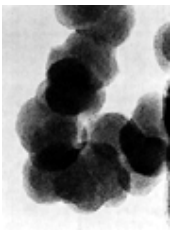

Figure 55. Phase-contrast electron micrograph of a carbon black aggregate

The carbon atoms within each layer are arranged in almost the same manner as in graphite. The layers are almost parallel to each other; however, the relative position of these layers is random, so that there is no order as in the *c* direction of graphite (“turbostratic structure”) [415]. X-ray diffraction permits the determination of “crystalline” regions within the carbon black primary particle. These regions are parts of more extended layers, not isolated crystallites. X-ray diffraction reflexes are observed wherever parts of at least three layers are parallel and equidistant. For most carbon blacks, these “crystalline” regions are 1.5 – 2.0 nm in length and 1.2 – 1.5 nm in height, corresponding to 4 – 5 carbon layers [416]. The fraction of “crystalline” or well-ordered carbon in carbon blacks varies according to oxidation kinetic studies between 60 and 90 %.

The morphology of carbon black primary particles indicates that during formation of carbon black, the first nuclei of pyrolyzed hydrocarbons condensate from the gas phase. Thereafter, further carbon layers or their precursors are adsorbed onto the surface of the growing particle. Due to this adsorption, the new layers are always orientated parallel to the existing surface. In the case of high-structure carbon blacks, several particles are joined by collision while they grow. Aggregates are formed by further carbon deposits on these initially loose agglomerates. Polyacetylene seems to play a role in the formation of precursors when aliphatic hydrocarbons are used as starting materials. With aromatic raw materials, however, it is more likely that aromatic degradation products are the intermediates.

The carbon layers of carbon black rearrange to a graphitic order, beginning at the particle surface at temperatures above 1200 °C. At 3000 °C, graphite crystallites are formed and the carbon black particles assume polyhedral shape.

Specific Surface Area. The specific surface area of industrial carbon blacks varies widely. While coarse thermal blacks have specific surface areas as small as 8 m^2/g, the finest pigment grades can have specific surface areas as large as 1000 m^2/g. The specific surface areas of carbon blacks used as reinforcing fillers in tire treads lie between 80 and 150 m^2/g. In general, carbon blacks with specific surface areas > 150 m^2/g are porous with pore diameters of less than 1.0 nm. The area within the pores of high-surface-area carbon blacks can exceed the outer (geometrical) surface area of the particles.

Adsorption Properties. Due to their large specific surface areas, carbon blacks have a remarkable adsorption capacity for water, solvents, binders, and polymers, depending on their surface chemistry. Adsorption capacity increases with increasing specific surface area and porosity. Chemical and physical adsorption not only determine wettability and dispersibility to a great extent, but are also most important factors in the use of carbon blacks as fillers in rubber as well as in their use as pigments. Carbon blacks with high specific surface areas can adsorb up to 20 wt % of water when exposed to humid air. In some cases, the adsorption of stabilizers or accelerators can pose a problem in polymer systems.

Density. Density measurements using the helium displacement method yield values between 1.8 and 2.1 g/cm^3 for different grades of carbon black. A mean density value of 1.86 g/cm^3 is commonly used for the calculation of electron microscopic surface areas. Graphitization raises the density to 2.18 g/cm^3. The lower density with respect to graphite (2.266 g/cm^3) is due to slightly greater layer distances.

Electrical Conductivity. The electrical conductivity of carbon blacks is inferior to that of graphite, and is dependent on the type of production process, as well as on the specific surface area and structure. Since the limiting factor in electrical conductivity is generally the transition resistance between neighboring particles, compression or concentration of pure or dispersed carbon black, respectively, plays an important role. Special grades of carbon black are used to give polymers antistatic or electrically conductive properties. Carbon blacks with high conductivity and high adsorption capacity for electrolyte solutions are used in dry-cell batteries.

Light Absorption. The wide use of carbon blacks as black pigments is due to their absorption of visible light. The absorption rate can reach 99.8 %. The black color can take on a bluish or brownish tone, depending on light scattering, wavelength, the type of carbon black, and the system into which the carbon black is incorporated. Infrared and ultraviolet light is also absorbed. Therefore, some carbon blacks are used as UV stabilizers in plastics.

6.2. Chemical Properties

Chemical Composition and Surface Chemistry. The global chemical composition according to elemental analysis is within the following limits:

carbon. 80.0 – 99.5 wt %
hydrogen. 0.3 – 1.3 wt %
oxygen. 0.5 – 15.0 wt %
nitrogen. 0.1 – 0.7 wt %
sulfur. 0.1 – 0.7 wt %

depending on the manufacturing process, raw material, and possible chemical aftertreatment. The ash content of most furnace blacks is < 1 wt %. The ash components can result from the raw material, the salts which are injected to control the structure, and salts in the process water. The ash content of gas blacks is less than 0.02 %.

The surface of carbon blacks contains certain amounts of polynuclear aromatic substances. These are strongly adsorbed and can only be isolated by continuous extraction with solvents, e.g., boiling toluene. For most industrial pigment-grade carbon blacks, the amount of extractable material is below the limit defined by the food laws.

The hydrogen in carbon black is bound as CH groups at the edge of the carbon layers. Nitrogen seems to be primarily integrated into the aromatic layer system as heteroatoms.

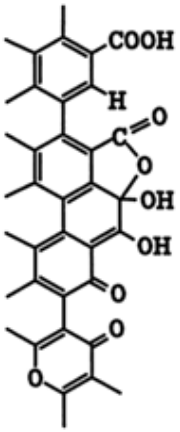

The oxygen content of carbon blacks is of great importance for their application. Oxygen is bound to the surface in the form of acidic or basic functional groups. The amount of surface oxides and their composition depend on the production process and an eventual aftertreatment. Furnace blacks and thermal blacks, which have been produced in a reducing atmosphere, contain about 0.2 – 2.0 wt % oxygen in the form of almost pure, basic surface oxides. Gas and channel blacks, which are manufactured in the presence of air, contain up to 8 wt % oxygen. In this case, the greater part of the oxygen is contained in acidic surface oxides and only a small portion in basic oxides. The amount of acidic surface oxides can be increased by oxidative aftertreatment, in which oxygen contents of up to 15 wt % can be obtained.

The surface oxides are destroyed at high temperatures. Due to this fact the weight loss at 950 °C (“volatiles”) is a rough indication of the oxygen content of a carbon black.

The pH measured in an aqueous slurry is another indication of the degree of oxidation. In general, the pH is > 7 for furnace blacks (low oxygen content, basic surface oxides), 4 – 6 for gas blacks, and 2 – 4 for oxidized carbon blacks (high oxygen content, acidic surface oxides with a high amount of polar functional groups). Further organic reactions, e.g., alkylation, halogenation, esterification, can be carried out with the surface oxides to modify the surface properties.

Oxidation Behavior. Industrial carbon blacks do not spontaneously ignite when stored in air at 140 °C according to IMCO Code [417]. When ignited in air, carbon black glows slowly. In contrast to coal, dust explosions are not observed under normal test conditions [418]. However, ignition sources of extremely high energy, e.g., a gas explosion, may induce a secondary dust explosion in air.

For modification of the application properties with regard to surface oxidation, see Section Oxidative Aftertreatment of Carbon Black .

6.3. Raw Materials

Mixtures of gaseous or liquid hydrocarbons which can be vaporized represent the raw materials preferable for the industrial production of carbon black. Since aliphatic hydrocarbons give lower yields than aromatic hydrocarbons, the latter are primarily used. The best yields are given by unsubstituted polynuclear compounds with 3 – 4 rings. Certain fractions of coal tar oils and petrochemical oils from petroleum refinement or the production of ethylene from naphtha (aromatic concentrates and pyrolysis oils) are materials rich in these compounds. These aromatic oils, which are mixtures of a variety of substances, are the most important feedstocks today. Oil on a petrochemical basis is predominant. A typical petrochemical oil consists of 10 – 15 % monocyclic, 50 – 60 % bicyclic, 25 – 35 % tricyclic, and 5 – 10 % tetracyclic aromates.

Important characteristics determining the quality of a feedstock are the C/H ratio as determined by elemental analysis and the BMCI [419] (Bureau of Mines Correlation Index), which is calculated from the density and the mid-boiling point or from the

density and the viscosity. Both C/H ratio and BMCI values give some information on the aromaticity and therefore the expected yield. Further characteristics are viscosity, pour point, temperature of solidification, alkaline content (due to its influence on the carbon black structure), and sulfur content, which should be low because of environmental and corrosion considerations.

Natural gas, which was previously the predominant feedstock for the production of channel blacks and furnace blacks in the United States, has lost its importance for economic reasons. Only thermal blacks are produced with natural gas. However, natural gas is still the most important fuel in the furnace black process, although other gases and oils are used in some cases. In several patents, recycled tail gas, in combination with oxygen or oxygen-enriched air, has also been proposed as a fuel, but has not gained any commercial importance.

Acetylene, due to its high price, is used only for the production of highly specialized conductivity blacks and battery blacks.

6.4. Production Processes

A summary of the most important production processes is given in Table 22. In general, the processes are divided into two groups: those employing partial combustion and those based on pure pyrolysis. This nomenclature is somewhat misleading insofar as the carbon black resulting from the partial combustion process is also formed by pyrolysis. The two types of processes differ in that air is used in the one to burn part of the feedstock, thus producing the energy required to carry out the pyrolysis, whereas in the other heat is generated externally and introduced into the process.

Table 22. Summary of the manufacturing processes and feedstocks for the production of carbon black

Chemical process	Manufacturing process	Feedstock
Incomplete combustion	furnace black process	petrochemical oils and coal tar oils
	Degussa gas black process	coal tar oils
	channel black process	natural gas
	lamp black process	petrochemical and coal tar oils
Thermal cracking	thermal black process	natural gas, oil
	acetylene black process	acetylene

The furnace black process is currently the most important production process. It accounts for more than 95 % of the total worldwide production. The advantages of the furnace black process are its great flexibility, which allows the manufacture of various grades of carbon black, and its better economy compared to other processes. The following comparison makes this apparent: for similar grades of carbon black, the production rate of one flame is ca. 0.002 kg/h for channel black, ca. 0.2 kg/h for gas black, and ca. 2000 kg/h for a modern furnace black reactor. However, in spite of the more advantageous furnace black process, the production processes listed in Table 22 (except for the channel black process) are still in use for the production of special carbon blacks which cannot be obtained via the furnace black process.

6.4.1. *Furnace Black Process*

In the past decades, the rapidly expanding automobile industry required increasing numbers of tires with various characteristics. This led not only to the development of new rubber grades, but also to the development of new carbon blacks required by the increasingly refined application processes and to the development of a new and better manufacturing process, the furnace black process. Unlike the old channel black process, this process allows the production of virtually all grades of carbon black required by the rubber industry. It also meets the high economic and ecological requirements of our times.

The furnace black process was developed in the United States in the 1920s, and since then, it has been greatly refined. It is a continuous process, carried out in closed reactors, so that all inputs can be carefully controlled [420]. Today most semireinforcing rubber blacks (carcass or soft blacks), with specific surface areas of 20 – 60 m^2/g, and the active reinforcing blacks (tread or hard blacks; see Table 25), with specific surface areas of 65 – 150 m^2/g, are manufactured by this process, as well as an increasing number of pigment-grade carbon blacks with much higher specific surface areas and smaller particle sizes. In addition to the specific surface area, other quality specifications such as structure, as measured by DBP absorption, and application properties such as abrasion resistance, modulus, and tear strength for usage in rubber, or jetness and tinting strength for color blacks, can also be systematically varied in the furnace black process by adjusting the operating parameters. This flexibility is necessary to meet the very narrow specifications required by customers.

The heart of a furnace black production plant is the furnace in which the carbon black is formed. The feedstock is injected, usually as an atomized spray, into a high-temperature zone of high energy density which is achieved by burning a fuel (natural gas or oil) with air. The oxygen, which is in excess with respect to the fuel, is not sufficient for complete combustion of the feedstock, the majority of which is therefore pyrolyzed to form carbon black at 1200 – 1900 °C. The reaction mixture is then quenched with water and further cooled in heat exchangers, and the carbon black is collected from the tail gas by a filter system.

Figure 56 shows a schematic drawing of a furnace black production line. The feedstock, preferably petrochemical or carbochemical heavy aromatic oils, some of which begin to crystallize near ambient temperature, is stored in heated tanks equipped with circulation pumps to maintain a homogeneous mixture. Oil is conducted to the reactor by means of rotary pumps via heated pipes and a heat exchanger, where it is heated to 150 – 250 °C to obtain a viscosity appropriate for atomization. Various types of spraying devices are used to introduce the feedstock into the reaction zone. An axial oil injector with a spraying nozzle which produces a hollow-cone spray pattern, is often used. One- and two-component atomizing nozzles [421] are in use, air and steam being the preferred atomizing agents in the latter case. In other reactors, however, the feedstock is injected as a plurality of coherent or atomized streams into the accelerated combustion gases perpendicular to the direction of stream [422].

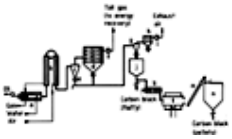

Figure 56. Furnace black process

a) Furnace black reactor; b) Heat exchanger; c) Collecting or agglomerating cyclone; d) Bag filter; e) Carbon black outlet to pneumatic conveying system; f) Tail gas blower; g) Collector; h) Exhaust air filter; i) Blower for the pneumatic conveying air; j) Fluffy black storage tank; k) Pelletizer; l) Dryer drum; m) Conveying belt; n) Storage tank for carbon black pellets

As the carbon black structure may be reduced by the presence of alkali metal ions in the reaction zone [423], alkali metal salts, preferably aqueous solutions of potassium hydroxide or potassium chloride, are often added to the make oil in the oil injector. Alternatively, the additives may be sprayed separately into the combustion chamber. In special cases, other additives, e.g., alkaline-earth metal compounds which increase the specific surface area are introduced in a similar manner.

The high temperature necessary for pyrolysis is obtained by burning fuel in excess air in a combustion chamber. Natural gas is still the fuel of choice, but other gases, e.g., coke oven gases or vaporized liquid gas, are occasionally used. Various oils, including the feedstock, are occasionally used as fuel for economic reasons. Depending on the type of fuel, special burners are used to obtain fast and complete combustion.

The air required for combustion is compressed by rotating piston compressors or turbo blowers. The air is preheated in heat exchangers by the hot carbon-black-containing gases leaving the reactor. This saves energy and thus improves the carbon black yield. Preheated air temperatures of 500 – 700 °C are common.

Important progress has been made on the reactor throughput: A production plant with a capacity of 20 000 t/a (2.5 t/h) was originally run with as many as 12 furnaces. In the last decades they have been replaced by a single high-performance reactor for the same capacity. Modern plants are a one-stream configuration with one unit for each process step (reactor, collecting system, pelletizer, dryer). From a technical point of view, even larger units could be built. However, due to the great variety of carbon black grades required, the capacity of one unit is economically limited by the frequency of switching over to other grades and the amount of off-grade carbon black which may be produced during this procedure.

The reactors of modern furnace plants vary considerably in internal geometry, flow characteristics, and the manner in which fuel and feedstock are introduced. Nevertheless, they all have the same basic process steps in common: producing hot combustion gases in a combustion chamber, injecting the feedstock and rapidly mixing it with the combustion gases, vaporizing the oil, pyrolyzing it in the reaction zone, and rapidly cooling the reaction mixture in the quenching zone to temperatures of 500 – 800 °C.

Schematic drawings of some typical modern furnace black reactors are shown in Figure 57. They all have a gas-tight metal jacket. The reaction zone is coated with a ceramic inner liner, generally on an alumina base, which is stable to temperatures of ca. 1800 °C. Several quenching positions allow the effective reaction volume of the reactor to be adjusted. This allows variation of the mean residence time of the carbon black at the high reaction temperature. Typical residence times for reinforcing blacks are 10 – 100 ms.

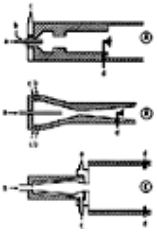

Figure 57. Furnace black reactors

A) Restrictor ring reactor; B) Venturi reactor [424]; C) Reactor with high-speed combustion chamber [425]a) Feedstock; b) Fuel; c) Combustion air; d) Quench

Most furnace black reactors are arranged horizontally. They can be up to 18 m long with an outer diameter of up to 2 m. Some vertical reactors are used especially for the manufacture of certain semireinforcing blacks [426] (Fig. 58). For further reactors, see [427].

Figure 58. Vertical reactor for manufacturing semireinforcing blacks

a) Fuel inlet; b) Oil injector; c) Air conduit; d) Water spray; e) Outlet to the collecting system

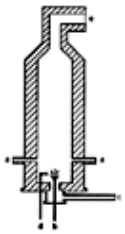

The properties of carbon blacks depend on the ratios of fuel, feedstock, and air, which therefore must be controlled carefully [428]. The particle size of the carbon black generally decreases with increasing amounts of excess air relative to the amount needed for the complete combustion of the fuel. Since the excess air reacts with the feedstock, a greater amount of air leads to higher oil combustion rates, resulting in rising temperatures in the reaction zone. As a consequence, the nucleation velocity and the number of particles formed increase, but the mass of each particle and the total yield decrease. This allows semireinforcing carbon blacks to be manufactured with better yields than active reinforcing carbon blacks.

The yields, which depend on the carbon black type and the type of feedstock, range between 50 and 65 % for semireinforcing blacks and 40 and 60 % for reinforcing blacks. High-surface-area pigment blacks with markedly smaller particle size than rubber blacks give lower yields.

Other parameters influencing carbon black quality are the manner in which the oil is injected, atomized, and mixed with the combustion gases, the type and amount of additives, the preheating temperature of the air, and the quench position. As long as the carbon black is in contact with the surrounding gases at the high reaction temperature, several reactions on the carbon surface occur (e.g., Boudouard reaction, water gas reaction), so that the chemical nature of the carbon black surface is modified with increasing residence time. When quenched to temperatures below 900 °C, these reactions are stopped and a certain state of surface activity is frozen. Carbon black surface properties can also be further modified by varying the pelletizing and drying conditions (see below).

Typical processing data for reactors with a carbon black output of 10 000 t/a (1250 kg/h) of tread black and of 14 000 t/a (1750 kg/h) of carcass black are listed in Table 23. These data show that the total mass put through the reactor amounts to 10 – 16 t/h. Although this is done at high streaming velocities (up to 800 m/s) and high temperatures (up to 1800 °C), modern high-performance reactors can have lifetimes of two years and more.

Table 23. Processing data for high-performance furnace black reactors

		Semireinforcing carbon black	Reinforcing carbon black
Natural gas	m^3/h	300 – 550	280 – 440
Air	m^3/h	7 000 – 10 000	6 000 – 7 500
Oil	kg/h	2 500 – 3 300	2 000 – 3 000
Carbon black	kg/h	1 500 – 2 000	1 000 – 1 500
	t/a	12 000 – 16 000	8 000 – 12 000
Yield	kg/100 kg of oil	50 – 65	40 – 60

The mixture of gas and carbon black leaving the reactor is cooled to 250 – 350 °C in heat exchangers by counterflowing combustion air and then conducted into the collecting system. Formerly, a combination of electroflocculators and cyclones or cyclones and filters were used [420], [429]. Currently, simpler units are preferred. Generally, the collecting system consists of only one high-performance bag filter with several chambers, which are periodically purged by counterflowing filtered gas or by pulse jets. Occasionally, an agglomeration cyclone is installed between the heat exchanger and the filter. Depending on the capacity of the production unit, the filter may contain several hundred bags with a total filter area of several thousand square meters. Usual filter loads are on the order of 0.2 – 0.4 $m^3\ m^{-2}\ min^{-1}$. Since the filtered gas contains 25 – 40 vol % water vapor, most filters operate at temperatures above 200 °C to avoid condensation. The residual carbon black content in the off-gas is less than 10 mg/m^3.

Because of the reducing atmosphere and the high temperatures in the reactor, the tail gas, which consists of 25 – 40 vol% water vapor, 40 – 50 vol % nitrogen, and 3 – 5 vol % carbon dioxide, also contains a certain amount of combustible gases, the amount of which depends on the feedstock and the processing conditions. These gases include 5 – 10 vol% carbon monoxide, 5 – 10 vol % hydrogen, and small amounts of methane and other hydrocarbons. The lower heating value lies between 1700 and 2100 kJ/m^3. The energy remaining in the tail gas can be calculated by using the typical overall energy balance of the furnace black process shown in Figure 59. The gas must be burned for environmental reasons, and its energy is used, e.g., for heating dryer drums or for the production of steam and electricity.

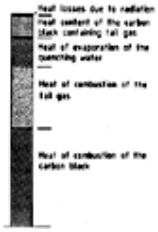

Figure 59. Typical energy balance for the manufacture of a reinforcing black

The fluffy carbon black coming out of the filter is pneumatically conveyed into a first storage tank. Small amounts of solid impurities (“grit,” e.g., iron, rust, or coke particles) are either removed by magnets and classifiers or milled to an appropriate consistency.

Freshly collected carbon black has an extremely low bulk density of 20 – 60 g/L. To facilitate handling and further processing by the customer, it must be compacted. Densification by “outgassing” — a process by which the carbon black is passed over porous, evacuated drums — is the weakest form of compacting and allows the carbon black to retain its powdery state [430]. This form of compacting is used for certain pigment blacks which must retain easy dispersibility.

Other pigment blacks and the rubber blacks are compacted by granulation. Two processes are used: dry and wet pelletization.

Dry pelletizing is a simple and energy-saving method, but it does not work with all types of carbon black. It is mainly used for color blacks. Dry pelletization is carried out in rotating drums, where the powdery carbon black rolls to form small spheres.

The wet pelletization process is used for the majority of rubber blacks. Carbon black, water, and small amounts of additives (e.g. molasses, ligninosulfonates) are mixed in special pelletizers [431]. They usually consist of a horizontal cylinder ca. 3 m long and 0.7 – 1 m in diameter, in the axis of which a pin shaft rotates at 300 – 750 rpm (Fig. 60). The water containing the pelleting agents dissolved in it is injected via spray nozzles. The density of the pelletized material is ca. 10 times that of the original carbon black. DBP absorption is also reduced during this process. The pellet crush strength and some application properties in rubber can be influenced by the type and amount of the pelleting agent. The size of the pellets is ca. 1 – 2 mm.

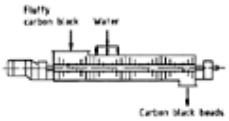

Figure 60. Pelleting machine

The carbon black leaving the pelleting machine contains ca. 50 wt % water. It is dried in dryer drums indirectly heated by burning tail gas. Dryer drums with a capacity of 2000 kg of carbon black per hour are 15 – 20 m long and 2 – 3 m in diameter. They are rotated at 5 – 15 rpm. Drying temperatures, generally between 150 and 250 °C, allow further modification of the carbon black properties.

The dried carbon black is transported via conveyor belts and elevators to the storage tank or packing station. Bulk densities of wet-pelletized carbon blacks are between 250 and 500 g/L.

A flow diagram summarizing the complete furnace black process is shown in Figure 61.

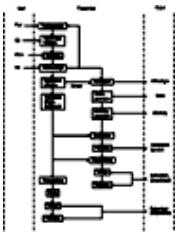

Figure 61. Furnace black process

6.4.2. *Gas Black and Channel Black Processes*

The channel black process, used in the United States since the late 1800s, is the oldest process for producing small-particle-size carbon blacks on an industrial scale. Originally, the first reinforcing blacks were also produced by this process. In 1961, world production of channel black was about 120 000 t. Due to low profitability and environmental difficulties, the last production plant in the United States was closed in 1976. Natural gas was used as the feedstock. The carbon black yield was only 3 – 6 %.

In Germany, where natural gas was not available in sufficient amounts, the gas black process was developed in the 1930s. It is similar to the channel black process, but uses coal tar oils instead of natural gas. Yields and production rates are much higher with oil-based feedstock; this process is still used to manufacture high-quality pigment blacks with properties comparable to those of channel blacks. The gas black process has been used by Degussa on an industrial scale since 1935.

Originally, gas black was primarily used for the reinforcement of rubber. Today, almost all grades are used as color blacks in printing inks, plastics, lacquers, and coatings. High-quality oxidized gas blacks are of special interest, e.g., in deep black lacquers and coatings.

In the gas black process (Fig. 62), the feedstock is partially vaporized. The residual oil is continuously withdrawn. The oil vapor is transported to the production apparatus by a combustible carrier gas (e.g., hydrogen, coke oven gas, or methane). Air may be added to the oil – gas mixture for the manufacture of very small particle size carbon black. Although this process is not as flexible as the furnace black process, various grades of gas black can be made by varying the relative amounts of carrier gas, oil, and air. The carbon black properties are also influenced by the type of burners used.

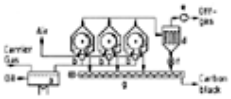

Figure 62. Degussa gas black process

a) Oil evaporator; b) Burner; c) Cooling drum; d (Bag filter; e) Blower; f) Rotary valve; g) Conveying screw

A gas black apparatus consists of a burner pipe approximately 5 m long, which carries 30 – 50 diffusion burners. The flames burn in contact with a water-cooled drum, where about half of the carbon black formed is deposited. This black is continuously scraped off and transported by a screw to a pneumatic conveying system. The gas black apparatus is surrounded by a steel housing open at the bottom. At the top, fans extract the off-gas into filters, which collect the carbon black suspended in the gas. The amount of air entering the apparatus can be regulated by valves in the exhaust pipes.

Several gas black apparatus are combined to form one production unit. The whole “group” is fed by one oil vaporizer. The production rate and the yield of an apparatus depend on the grade of carbon black produced. For a typical RCC black, the production rate is 7 – 9 kg/h and the yield is 60 %. The yield for high-quality color blacks is considerably lower (10 – 30 %).

To remove possible impurities, the gas black is classified and then densified, pelletized, or submitted to an oxidative aftertreatment (see Section Oxidative Aftertreatment of Carbon Black), depending on its intended use. Since gas blacks are formed in the presence of excess air, their surface is oxidized. Acidic surface oxides are predominant.

6.4.3. *Lamp Black Process*

The lamp black process is the oldest industrial-scale production process [420], [429].

Currently, only a few plants still produce rather coarse blacks (mean particle diameter ca. 100 nm) with special properties. They are used as nonreinforcing or semireinforcing blacks in rubber goods and as tinting black with a low pigment separation tendency.

The lamp black process is only partially continuous. The feedstock — oil with a high aromatic hydrocarbon content — is burned in flat steel vessels up to 1.5 m in diameter (Fig. 63). The oil is continuously introduced into the vessel to keep a constant feedstock level. The off-gas containing carbon black is sucked into a conical exhaust pipe, which is coated with a ceramic inner liner and leads to the cooling and collecting system. The properties of the carbon black can be influenced to some extent by variation of the distance between the vessel and the exhaust and the amount of air sucked into the apparatus. One lamp black apparatus can produce 100 kg/h. The production process must be interrupted at certain time intervals to remove coke-containing residues from the vessels.

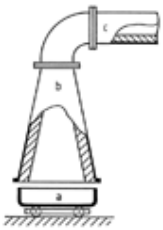

Figure 63. Lamp black process

a) Vessel filled with feedstock; b) Conical exhaust pipe; c) Pipe leading to the collecting system

6.4.4. *Thermal Black Process*

Some special processes for producing carbon black are based on the thermal decomposition of lower gaseous hydrocarbons in the absence of air.

The thermal black process, which was developed in the 1930s, is still used for the production of coarse carbon blacks (nonreinforcing carbon blacks) for special applications in the rubber industry. Contrary to the above described processes, energy generation and the pyrolysis reaction are not carried out simultaneously.

A thermal black plant consists of two furnaces, which are used in alternate heating and production periods of ca. 5-min duration (Fig. 64). Each of the cylindrical furnaces (4 m in diameter and 6 m high) contains a network of heat-resistant bricks. They are heated with natural gas and air. At a temperature of ca. 1400 °C, the air is switched off and only feedstock is introduced for pyrolysis. Since this reaction is endothermic, the temperature falls. At about 900 °C, a new heating period is necessary.

Figure 64. Thermal black process

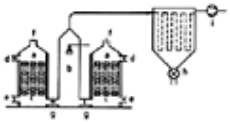

a) Thermal black reactor; b) Cooler; c) Filler bricks; d) Inlet for the feedstock; e) Inlet for the fuel; f) Outlet for the burned fuel; g) Outlet for the pyrolysis products; h) Carbon black outlet; i) Blower

The products leaving the furnace, carbon black and nearly pure hydrogen, are cooled by injecting water into an ascending channel. The carbon black is separated in the collecting system.

Carbon blacks of lower particle size can be produced by diluting the natural gas with recycled hydrogen. Fine thermal blacks (FT blacks) with mean primary particle sizes of 120 – 200 nm were manufactured in this way in the past. Medium thermal blacks (MT blacks) with mean particle sizes of 300 – 500 nm are still produced and are obtained by using undiluted feedstock. The yield of MT blacks is about 40 % with respect to the total amount of feedstock and fuel used.

Thermal blacks are used for mechanical rubber goods with high filler contents. Cheaper products (clays, milled coals, and cokes), however, have become increasingly important as substitutes for economic reasons. The total production of thermal blacks is, therefore, decreasing.

6.4.5. *Acetylene Black Process*

Acetylene and mixtures of acetylene with light hydrocarbons are the raw materials for a process that has been used since the early 1900s. Unlike other hydrocarbons, the decomposition of acetylene is highly exothermic (ΔH = – 230 kJ/mol).

The discontinuous explosion process is the oldest technical process. It was mainly used for the production of color blacks. Continuous processes were later developed with production rates up to 500 kg/h [432]. Acetylene or acetylene-containing gases are fed into a preheated, cylindrical reactor with a ceramic inner liner. Once ignited, the reaction is maintained by the decomposition heat that is evolved. The carbon black is collected in settling chambers and cyclones. Approximately 95 – 99 % of the theoretical yield is obtained.

The primary particles of acetylene black have different shapes than those of other carbon blacks (Fig. 65). As the increased order in the *c* direction of the crystalline regions indicates, folded sheets of carbon layers are the main structural component. Because of their relatively high price, the application of acetylene blacks is limited to special uses, e.g., in dry cells,. Total worldwide production is ca. 40 000 t/a.

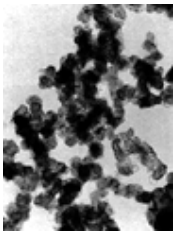

Figure 65. Electron micrograph of acetylene black

6.4.6. *Other Manufacturing Processes*

In a plasma, hydrocarbon vapors may be almost quantitatively decomposed into carbon and hydrogen [433]. Many producers of carbon black have done research in this field. According to numerous patent specifications, this method can be used to make small-particle carbon blacks with new properties. However, an economical plasma-based commercial process is not yet known.

The Hüls electric arc process was the only large-scale process using plasma reactions in which large quantities of carbon black were produced as a byproduct of the production of acetylene. The particles of the Hüls arc carbon black resemble those of acetylene black. The mean primary particle size is ca. 35 nm. Today, this kind of carbon black is no longer used as a pigment.

Since the price of both feedstocks and fuels, and thus, the profitability of the carbon black production processes, is highly dependent on the petrochemical industry, several attempts have been made to find new raw materials. Processes for obtaining carbon black directly from coal [434] or for isolating carbon black from used tires, for example, have been studied. None of them, however, has been of any commercial importance up to now. On the other hand, clay, milled coal, and coke have found limited use as substitutes for very coarse carbon blacks, primarily thermal blacks and some SRF blacks. The increasing use of precipitated silicas in tires and mechanical rubber goods, mostly in combination with organosilane coupling agents, which originally was indicative of an increasing search for new non-oil-based fillers, has led to new rubber properties.

6.4.7. *Oxidative Aftertreatment of Carbon Black (Fig. 66)*

Oxygen-containing functional groups on the surface of carbon blacks strongly influence their application properties. High contents of volatiles, i.e., high concentrations of surface oxides, decrease the vulcanization rate and improve the flow characteristics of inks. The gloss of lacquers and coatings is increased, the color tone is shifted from brownish to bluish, and jetness often increases.

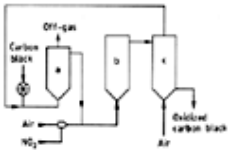

Figure 66. Equipment for the oxidative aftertreatment of carbon black in a fluidized bed

a) Fluidizing vessel; b) Reaction vessel; c) Desorption vessel

Due to the production conditions, only gas blacks (and channel blacks) are covered to a certain extent with acidic surface oxides. Furnace blacks contain only small amounts of oxygen in the form of basic surface oxides.

To amend their color properties, some color blacks are aftertreated by oxidation on a commercial scale. Depending on the oxidizing agent and the reaction conditions selected, different types of surface oxides are formed in varying quantities.

The simplest method of oxidizing the carbon black surface is by aftertreating it with air at 350 – 700 °C. However, the degree of oxidation is limited. Higher contents of surface oxides and better process control are achieved with nitric acid [435], mixtures of NO_2 and air [436], ozone, or sodium hypochlorite solutions [437] as oxidizing agents. As a rule, all strongly oxidizing agents may be used, either as a gas or in solution. Most surface oxidations of carbon black are carried out at elevated temperatures.

Oxidized carbon blacks may contain up to 15 wt % oxygen. They are strongly hydrophilic. Some of them form colloidal solutions spontaneously in water. In polar printing ink systems, lacquers, and coatings, a better wettability and dispersibility is achieved through surface oxidation [438], thus reducing binder consumption.

Surface oxidation of carbon black with nitric oxide and air can be carried out industrially in a fluidized-bed reactor [436]. A suitable aftertreatment unit consists of a preheating vessel, in which the carbon black is fluidized and heated, a reaction vessel to carry out the surface oxidation, and a desorption vessel, in which adsorbed nitric oxide is removed. Typical reaction temperatures lie between 200 and 300 °C. Depending on the degree of oxidation, the residence time can amount to several hours. The nitric oxide acts primarily as a catalyst, the oxygen in the air being the genuine oxidizing agent.

Oxidation of powdery black with ozone is also carried out on a commercial scale. Another common method of surface oxidation was carried out during pelletization. Instead of water, nitric acid was used as pelletizing agent. The surface was oxidized while the wet beads were dried at elevated temperature [439].

6.4.8. *Environmental Problems*

The furnace black process, the most economical and commercially advanced production method for carbon black, does not form toxic solid or liquid byproducts under normal operating conditions. The tail gas containing hydrogen and carbon monoxide was formerly emitted into the atmosphere, but is now burned, the energy being used partially within the process itself, e.g., for heating dryer drums in the wet-pelletizing process and for generating steam or electricity, so that emissions of carbon black and inflammable gases are avoided.

Although their economic importance is decreasing, progress with respect to environmental problems has also been made for the older processes that are still in use.

6.5. Testing and Analysis

The chemical composition of carbon blacks (see Section Chemical Properties), as determined by common elemental analysis methods, is of little significance for predicting their properties. Special characteristic properties are, therefore, determined for the characterization and quality control of carbon blacks. Traces of heavy metals are determined spectroscopically in the ash. Copper and manganese ions, etc., are of special interest to the rubber industry because of their interference with the aging process in rubber goods.

Electron Microscopy. Electron microscopy is one of the most important physical methods for the characterization of finely divided solids. It allows direct viewing of the shape and morphology of particles in this order of magnitude, primary particle size, particle size distribution, and aggregation.

To determine the mean primary particle size and particle size distribution, the diameters of 3000 – 5000 particles are measured on electron micrographs of known magnification. Spherical shape is anticipated for calculations. However, since the primary particles generally build up larger aggregates, the results may be somewhat uncertain. The specific “electron microscopic surface area” can be calculated from the primary particle size distribution. This value refers only to the outer (geometrical) surface of the particles. For porous carbon blacks the electron microscopic surface area is lower than the specific surface area according to BET (see below).

An attempt was made to find characteristic values for the type and degree of aggregation using electron micrographs [440]. However, neither visual comparisons with standard aggregates nor automatic picture analyses have led to a practical method for a quantitative characterization of the carbon black structure on a routine basis.

High-resolution phase-contrast pictures and X-ray diffraction are used to elucidate the internal structure of single primary particles.

Sorption Analysis. Specific surface areas and porosity can be calculated from the adsorption isotherm of nitrogen at – 196 °C. The BET method [441] is generally accepted for the evaluation of specific surface area (m^2/g). The two-parameter equation is applicable to carbon black. The BET surface area comprises the outer surface area as well as the surface area of

the pores. The outer surface area can be calculated from the adsorption isotherm by using the de Boer t-plot method.

Porosity of carbon blacks can be detected by the de Boer *t*-plot method [442]. The total surface area and the geometrical surface area outside the pores can be determined separately. Special attention must be paid to the selection of a suitable master *t* curve. Due to the small diameters of most carbon black primary particles, methods for the determination of mesopores are of no importance.

Special Analytical Test Methods. Test methods which on the one hand resemble certain physicochemical methods, but on the other hand already give some indication of application properties are summarized as special analytical test methods in Table 24. They can be carried out within a short time and are used for characterization as well as production control. Since the results are influenced by the test conditions, these test methods are standardized.

Table 24. Special analytical test methods for carbon black

Test method	Unit	Standard	Remark
Iodine adsorption	mg/g	ASTM D 1510; ISO 1304	amount of iodine adsorbed from aqueous solution as a measure for the specific surface area; not applicable for oxidized carbon blacks
CTAB surface area	m^2/g	ASTM D 3765; ISO 6810	amount of cetyl triammonium bromide adsorbed from aqueous solution as a measure for the specific outer surface area
BET surface area	m^2/g	ASTM D 3037; DIN 66 131, 66 132	total specific surface area calculated from the nitrogen adsorption isotherm by using the BET equation
External surface area	m^2/g	ASTM D 5816	STSA (statistical thickness surface area) calculated from the nitrogen adsorption isotherm
Aggregate dimension		ASTM D 3849	determination of aggregate dimensions (unit length, width, etc.) by electron microscope image analysis
DBP absorption	mL/100 g	ASTM D 2414; ISO 4656; DIN 53 601	determination of the wetting point with dibutyl phthalate in a special kneader as a measure for the carbon black structure
24M4-DBP absorption	mL/100 g	ASTM D 3493; ISO 6894	determination of DBP absorption after repeated compressing at high pressure as a measure for the permanent structure
Oil absorption	%	DIN ISO 787/5	percentage of linseed oil needed to make a barely flowable paste
Jetness, blackness value		DIN 55 979	light absorption of a carbon black paste in linseed oil; determination by visual comparison against standard blacks or by measuring the absolute light remission (DIN)
Tinting strength	%	ASTM D 3265; DIN ISO 787/16, 24; DIN 53 204; DIN 53 234	ability of a carbon black to darken a white pigment in a linseed oil paste; the tinting strength is the weight percentage of the standard carbon black with respect to the tested black to obtain the same gray tone; different standard white pigments and carbon black concentrations are used according to ASTM and DIN ISO
Volatiles	%	ISO 1126; DIN 53 552	weight loss when calcined at 950 °C for 7 min
Heating loss (moisture)	%	ASTM D 1509; DIN ISO 787/2	weight loss on drying at 125 °C for 1 h (ASTM) or 2 h at 105 °C (DIN ISO)
pH		ASTM D 1512; DIN 53 200	pH of an aqueous slurry of carbon black; pH is mainly influenced by surface oxides

Extractables	%	DIN 53 553	amount of extractable material (usually by boiling toluene) in at least 8 h
Toluene discoloration		ASTM D 1618; ISO 3858	light absorption (transmission) of a toluene solution of the extracted material
Ash content	%	ASTM D 1506; DIN 53 586	amount of noncombustible material after burning the carbon black at 675 °C (DIN) or 550 °C (ASTM, DIN)
Sulfur content	%	ASTM D 1619; DIN 53 584	
Sieve residue	%	ASTM D 1514; DIN ISO 787/18	amount of coarse impurities that cannot be purged through a testing sieve by water
Pour density	g/L	ASTM D 1513; DIN 53 912	measure for the densification of carbon black
Tamped density	g/L	DIN ISO 787/11	similar to bulk density; however, void volume is reduced by tamping
Pellet size distribution		ASTM D 1511	determination by means of sieve shaker
Fines content	%	ASTM D 1508; DIN 53 583	only for pelletized blacks; percentage passing through a sieve of 125 µm mesh width
Pellet crush strength	g	ASTM D 5230; ASTM D 3313	individual pellet hardness

Application Tests. Physicochemical and special analytical test methods allow the classification of carbon blacks and a rough estimation of their application properties. Exact data on the application properties of a carbon black in a special system, e.g., plastic material or a rubber mixture, can only be given by application tests under nearly practical conditions.

6.6. Storage and Transportation

The majority of the carbon black produced (up to 80 %) is transported as bulk material; the rest is handled in bags. Generally, the large stocks of pelletized furnace blacks for the rubber industry are stored, by the producers as well as the customer, in coated steel storage tanks consisting of one or more cells with a capacity of 100 – 1000 t (300 – 3000 m^3). Carbon black is shipped in containers or silo trucks with a capacity of up to 20 t. Recently, big bags and steel bins with a capacity of up to about 1 t have gained in importance. Smaller amounts of rubber blacks and especially pigment blacks are stored and transported in paper or plastic bags stacked on palettes. In some cases, bags with special coatings are used.

6.7. Uses

The most important grades and groups of carbon black, some typical characteristics, and the principal applications are summarized in Tables 25 and 26. A general overview of the most important application fields is given in Figure 67.

Table 25. Rubber blacks

ASTM No.	Iodine adsorption, mg/g	DBP absorption, mL/100 g	Use in natural and synthetic rubber
N 115	160	113	tire treads, rubber goods with high abrasion resistance (also as blend with N 220 and N 330 and N 326)
N 121	121	132	tire treads for high-performance passenger cars
N 125	117	104	tire treads for trucks, off-road tires
N 220	121	114	tire treads for trucks, tank pads, conveyor-belt covers
N 234	120	125	tire treads for high-performance cars, tank pads, conveyor belt covers
N 326	82	72	tire carcasses, steel cord adhesion compounds, mechanical rubber goods

N 330	82	102	tire treads, tire carcasses (blended with less-active carbon blacks), tire sidewalls, mechanical rubber goods
N 335	92	110	tire treads for passenger cars
N 339	90	120	tire treads for passenger cars, abrasion-resistant mechanical rubber goods
N 347	90	124	tire treads for passenger cars, abrasion-resistant mechanical rubber goods
N 351	86	120	tire treads, abrasion-resistant mechanical rubber goods
N 356	92	154	tire treads, cushion gum
N 375	90	114	tire treads for passenger cars and trucks, tank pads, conveyor-belt covers
N 539	43	111	tire carcasses, mechanical rubber goods with good dynamic properties, extrusion compounds
N 550	43	121	tire carcasses, mechanical rubber goods with good dynamic properties, extrusion compounds
N 650	36	122	tire carcasses, extrusion compounds
N 660	36	90	tire carcasses, molded goods
N 683	35	133	tire carcasses, tire sidewalls, mechanical rubber goods with good dynamic properties, extrusion compounds
N 762	27	65	mechanical rubber goods with excellent dynamic properties
N 765	31	115	tire carcasses, molded goods, extrusion compounds
N 772	30	65	mechanical rubber goods with excellent dynamic properties
N 774	29	72	tire carcasses, mechanical rubber goods with excellent dynamic properties, molded goods

Table 26. Pigment blacks

Type*	Blackness in linseed oil	Tinting strength**, %	Volatiles, %	BET surface area, m^2/g	Mean primary particle size, nm	pH	Application
Gas black							
HCC	>270	120 – 122	4 – 6	250 – 350	13 – 15	4 – 5	extremely deep
HCC ox.	>270	112 – 116	15 – 25	300 – 500	13 – 17	2 – 3	black paints and coatings of all types, plastics, fibers
MCC	250 – 270	120 – 122	4 – 6	150 – 200	15 – 20	4 – 5	coatings, plastics, fibers
MCC ox.	250 – 270	112 – 118	13 – 18	200 – 300	15 – 20	2 – 4	coatings, plastics, fibers
RCC	240 – 250	105 – 115	4 – 6	90 – 110	25 – 30	4 – 5	paints, coatings, printing inks,
RCC ox.	240 – 250	100 – 110	10 – 15	100 – 200	25 – 30	3 – 4	carbon paper, carbon ribbons, plastics, fibers
Lamp black							
	208 – 210	25 – 30	1 – 2	18 – 25	90 – 110	7 – 9	tinting black for coatings and plastics
Furnace black							
HFC	>255	115 – 125	0.5 – 1.5	250 – 300	14 – 15	9 -	high-jet coloring of

						10	plastics
MFC	250 – 260	120 – 126	0.5 – 1.5	150 – 200	16 – 18	9 – 10	high-jet coloring of plastics
RCF	240 – 250	95 – 120	0.5 – 1.5	80 – 120	20 – 30	9 – 10	plastics, printing inks,
RCF ox.	240 – 250	110 – 120	2 – 3	100 – 120	20 – 30	2.5 – 3.5	carbon paper, carbon ribbons, fibers, UV stabilization of polyolefins
LCF	215 – 235	60 – 100	0.5 – 1	30 – 65	30 – 60	9 – 10	plastics, tinting black,
LCF ox.	215 – 235	65 – 105	1.5 – 2.5	30 – 65	30 – 60	2.5 – 4	printing inks

*HCC = high color channel, MCC = medium color channel, RCC = regular color channel, HCF = high color furnace, MCF = medium color furnace, RCF = regular color furnace, LCF = low color furnace.
**Reference: IRB 3 = 100 %.

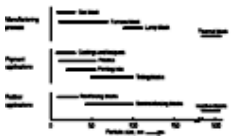

Figure 67. Mean particle sizes and typical applications of various carbon blacks

About 90 % of the carbon black produced is used by the rubber industry as a reinforcing filler in tires, tubes, conveyor belts, cables, rubber profiles, and other mechanical rubber goods. Furnace blacks are predominantly used in rubber processing. Fine-particle-size carbon blacks (reinforcing blacks) are used for the production of rubber mixtures with high abrasion resistance (e.g., tire treads). Coarser carbon blacks (semireinforcing blacks) are used in rubber mixtures requiring low heat buildup during dynamic stress (e.g., carcass mixtures). Very coarse carbon blacks (nonreinforcing blacks) are incorporated into mixtures with high elasticity and good extrusion properties.

Table 25 lists furnace rubber blacks and their main application fields.

According to the commonly used ASTM classification (ASTM D 1765 and D 2516), rubber blacks are characterized by one letter and a three-digit number. The letter indicates the influence of the carbon black on the vulcanization process (N = normal curing, and S = slow curing). The first digit of the number should characterize the primary particle size or the specific surface area (Table 27). The other two digits are free and are used to identify individual grades within the group. Within each carbon black group, considerable variation in the combination of properties is possible, e.g., by varying the carbon black structure (DBP absorption) and surface activity.

Table 27. Classification of rubber blacks according to ASTM D 1765 and ASTM D 2516

Group no. (first digit of the three-digit ASTM number)	Average particle size, nm	Specific surface area*, m^2/g
0	1 – 10	>155
1	11 – 19	125 – 155
2	20 – 25	110 – 140
3	26 – 30	70 – 90
4	31 – 39	43 – 69
5	40 – 48	36 – 52
6	49 – 60	26 – 42
7	61 – 100	17 – 33
8	101 – 200	–
9	201 – 500	–

* The ranges of the specific surface area may vary, depending on the structure of the carbon blacks.

Quantitatively, the pigment blacks are substantially less important than the rubber blacks. They are used for the manufacture of printing inks, coloring plastics, fibers, lacquers, coatings, and paper (see Table 26). Oxidized carbon blacks are frequently used in the printing ink and coating industry. While high-color gas blacks are still predominant in lacquers and coatings, furnace blacks are becoming more and more important in plastics, coatings, and printing inks.

Besides their two main uses as reinforcing fillers and pigments, small amounts of carbon blacks are used by the electrical industry to manufacture dry cells, electrodes, and carbon brushes. Special blacks are used to give plastics antistatic or

electrical conduction properties. Another application is the UV stabilization of polyolefins [444].

6.8. Economic Aspects

In 1995, approximately (6 – 6.5)×10^6 t of carbon black was produced, and production capacity was estimated to be nearly 8×10^6 t/a (Table 28). The annual growth rate, on the average 7.9 % per annum between 1965 and 1975 in the United States, has since decreased substantially, primarily for rubber blacks due to longer tire life and the fact that the car market is reaching saturation. Therefore, the future growth rate of the carbon black market is expected to be rather limited and will not exceed 1 – 2 % per annum.

Table 28. Carbon black production capacity (1996)

Country or region	Capacity, 10^3 t
North America	1815
Western Europe	1310
Eastern Europe	1545
Asia	2630
South America	480
Africa, Australia	185
Total	**7965**

Nearly 40 % of the total world production capacity for carbon black is concentrated in the United States and Western Europe. A detailed survey of the capacities in Western Europe is given in Table 29.

Table 29. Carbon black production capacity in Europe (1995/96)

Country	Capacity, 10^3 t
United Kingdom	180
Germany	277
France	280
Italy	200
Netherlands	160
Spain	90
Sweden	40
Others	83
Total	**1310**

More than 90 % of the total amount of carbon blacks produced are used as reinforcing fillers in rubber, of which 65 – 70 % go into the tire industry and an additional 25 – 30 % are needed for the production of mechanical rubber goods. Less than 10 % of all carbon blacks produced are used for nonrubber purposes.

6.9. Toxicology and Occupational Health

As far as toxicology is concerned, one must distinguish between soot, which is formed by the uncontrolled combustion of coal and oil, and carbon black, which is industrially produced under precisely defined conditions. Commercial carbon blacks are characterized by an atomic ratio H : C of < 0.1, low ash content, and high adsorption capacity. The soluble organic fraction (extractable materials) is less than 0.5 wt %.

The “chimney-sweep cancer” described by P. POTT in England as early as 1775, which is basically related to soot but not to carbon black, was the starting point of intensive research on analytical test methods for polycyclic aromatic hydrocarbons (PAH) in carbon blacks. Increasingly refined test methods have been developed (e.g., column, thin-layer, and paper chromatography, gas chromatography, HPLC, UV and fluorescence spectrophotometry, and mass spectroscopy) to detect traces of such substances and to investigate the ability of carbon black surfaces to adsorb and to desorb PAHs.

Many years of experience in the carbon black producing and processing industry have clearly shown that there is no health hazard attributable to this product. Occasional allegations that carbon blacks impair human health have been refuted. In most cases it was possible to attribute these claims to confusion between carbon blacks and soot or coal dust. Carbon black differs markedly from soot, the unwanted, uncontrolled byproduct of combustion found in chimneys and the ambient air [445], [446].

Commercial carbon blacks (not aftertreated) consist of more than 95 % of carbon, and small amounts of hydrogen, oxygen and sulfur. They have an atomic ratio H:C of < 0.1 and a low ash content. In most carbon blacks the total toluene-extractable matter is less than 0.15 %, only part of which consists of polycyclic aromatic hydrocarbons (PAHs). Due to this low content and the highly adsorbent properties of carbon black, the extractable matter is bound very tightly on the surface and can only be removed with powerful solvents [447]. In contrast, soots vary widely in composition and properties. Besides carbon they contain variable amounts of inorganic and organic byproducts. Because of their uncontrolled genesis the content of extractable matter is often very high (> 25 %) and consists, among others, of a wide range of PAHs. The high content and loose binding of the PAHs is to a large extent responsible for the mutagenicity and carcinogenicity of soots.

Detailed studies of workers in the carbon black industry and among carbon black consumers in the United States and Europe have shown that carbon black causes neither structural damage to the lungs nor lung tumors. No association between cumulative carbon black exposure and the incidence of respiratory disease could be detected [448-450].

In small-animal tests no irritation of the mucous membranes or eyes was observed. Also no changes could be detected when the skin was exposed to carbon black. Conversely, animals exposed to the benzene extract of the same carbon black developed a significant number of malignant tumors [451]. If adsorbed on the carbon black surface, the PAHs are biologically inactive. Tests in five systems for genetic activity of a furnace black containing 294 ppm toluene-extractable PNAs indicated that no mutagenic activity could be attributed to whole carbon black [452]. The results of several studies provided no indications of pathological effects in the gastrointestinal tracts when carbon blacks were ingested in rodents [453].

Harmful materials other than PAHs, such as polychlorinated bi- and triphenols, polychlorinated dioxins, and polychlorinated hydrofurans, were not found in carbon blacks. Nitrosamines could not be detected in carbon black, but they may be formed in rubber compounds if rubber chemicals containing secondary amines are used. The total amine content of carbon black is less than 0.01 %, and the aromatic amine content is therefore even lower. Heavy metal content does not exceed 0.002 %. Carbon blacks therefore conform to all known regulations that limit these impurities.

Lifelong inhalation studies on various species, involving titanium dioxide, iron oxide, talc, diesel soot, as well as carbon black, have shown that these fine dusts, hitherto regarded as inert are toxic to the lung [454], [455]. Of the various species used in the studies, including rats, mice and hamsters, only the rat developed identifiable lung tumors. In all cases the rats were exposed to high dust levels for prolonged periods of time, to the point that they experienced a phenomenon termed “lung overload“. The effect is not restricted to carbon black. It appears from the results of long-term inhalation studies that any insoluble, low-toxicity particle will cause lung tumors in rats if deposited chronically at high enough doses [456]. Extrapolating the incidence of lung tumors in rats inhaling inert insoluble particles, such as carbon black, to human responses must be seriously questioned. The tests were carried out under such extreme overload conditions, never experienced in the workplace today, that the lungs of the rats were unable to clear themselves during nonexposure periods.

Such high carbon black levels could not occur in modern carbon black production plants. Most of the carbon black producing countries have adopted the U.S. TLV of 3.5 mg/m^3 [457]. Extensive sampling in a number of carbon black plants in Europe [458] resulted in an occupational exposure standard (OES) for carbon black of 3.5 mg/m^3 (8-h TWA) and 7 mg/m^3 (STEL). In an exhaustive review in 1984, carbon black was classified by IARC [459] in Group 3: *Not classifiable as to its carcinogenicity to humans*. This classification was confirmed in 1987. However, based on the results of the inhalation studies with rats, IARC [460] reclassified carbon black in 1995 in Group 2B: *The agent is possibly carcinogenic to humans*. The criteria applied are: Inadequate evidence in humans and sufficient evidence in experimental animals.

Only if strong evidence can be provided that the mechanism of carcinogenicity in rats does not operate in humans might a new reclassification of carbon black by IARC in Group 3 be possible.

Food Contact Regulations. Carbon blacks, which according to their application properties may be used for food products, cosmetics, drinking water pipes, food packaging materials, and toys, must comply with local regulations. The testing methods for the approval of carbon blacks for such applications are different in different countries. The general aim, however, is to give limitations for the content of PAHs and of heavy metals. In general, the amount of heavy metals is far below the limitations. With respect to PAHs, either the total amount of extractable materials or the content of special species is limited. Since industrial carbon blacks contain only trace amounts of strongly adsorbed PAHs, the majority of commercial grades comply with these regulations.

[Top of Page]

7. Fullerenes

Otto Vostrowsky and Andreas Hirsch

7.1. History

It has been known since the early work of HAHN et al. on nuclear fission products [461] that in the gas phase at 3000 – 4000 °C under the conditions of a high-frequency arc carbon tends to form C_n clusters. Cluster ions up to C_{15} were detected [462]. In 1970 OSAWA recognized that the bowl-shaped corannulene **1** is part of a soccer-ball framework, and postulated the existence of the stable C_{60} cluster **2** [463] and discussed its possible aromatic properties [464].

1 **2**

In 1984 it was observed that upon laser vaporization of graphite, large all-carbon clusters C_n with n = 30 – 190 can be produced [465]. Only ions with even numbers of carbon atoms were observed in the mass spectra of these carbon clusters, among which were C_{60} and C_{70}. Their identity, however, was not recognized at this time. The breakthrough came in 1985 [466] when KROTO visited the Rice University in Houston, where SMALLEY and coworkers had developed a technique [467] for generating clusters by focusing a pulsed laser on a solid (in this case graphite) and studying them by mass spectrometry. The original goal of KROTO and SMALLEY was to simulate the conditions under which carbon nucleates in red-giant stars. Under specific conditions of cluster formation, a peak attributed to C_{60} (m/z = 720) and a less intense peak attributed to C_{70} (m/z = 840) exhibit pronounced intensity in the spectra. Conditions were found under which the mass spectra were completely dominated by the C_{60} signal. KROTO and SMALLEY concluded that the stability of C_{60} is due to its spherical structure, that is, a truncated icosahedron with I_h symmetry (Fig. 68). The substance was named buckminsterfullerene due to its geometrical similarity with the geodesic domes constructed by the American architect BUCKMINSTER FULLER and its discovery was reported to *Nature* on September 12, 1985 [466]. Later the enhanced intensity of the peak for C_{70}, another stable fullerene, was understood. Although buckminsterfullerene C_{60} had been discovered, a method for its synthesis in macroscopic amounts remained to be found.

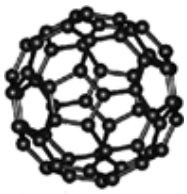

Figure 68. Ball-and-stick model of buckminsterfullerene

The second breakthrough in the fullerene research was achieved by KRÄTSCHMER and HUFFMAN [468]. Their intention was to produce laboratory analogues of interstellar dust by vaporization of graphite electrodes in a helium atmosphere [469]. At a certain helium pressure, the IR spectra of the soot generated by the vaporization of graphite showed four sharp absorptions besides the continuum of regular soot [470]. These absorptions were close to the positions predicted for buckminsterfullerene [471]. The fullerenes were isolated from the soot by sublimation or extraction. This allowed the determination of spectroscopic and crystallographic data, as well as experiments with ^{13}C-enriched material, and moved C_{60} from the world of elusive objects known only from mass spectroscopy to that of material science. Higher fullerenes can also be obtained in appreciable amounts by this technique.

The discovery of C_{60} led to an explosive growth in research into the properties of this new class of carbon allotropes. More than a thousand publications on fullerenes appear each year, dealing with their chemical and physical properties and material science aspects.

The discovery of the fullerenes was honored by the award of the Nobel prize for Chemistry in 1996 to KROTO, CURL, and SMALLEY.

7.2. Production

7.2.1. *Vaporization of Graphite*

Resistive Heating of Graphite. The resistive heating of graphite was the first method for producing macroscopic quantities of fullerenes [468]. The apparatus first used by KRÄTSCHMER et al. for the production of C_{60} was a bell jar connected to a pump and gas inlet. In the jar two graphite rods — one sharpened to a conical point and the other with a flat end — are kept in contact by a spring. The apparatus is evacuated, filled with 140 mbar of helium, and an electric current passed through the rods. This heats the point of contact to 2500 – 3000 °C and forms a black smoke, which condenses on the bell jar and a smoke catcher. The fullerenes are extracted from the soot with toluene in a yield of 10 – 15 %. The HUFFMAN – KRÄTSCHMER process has been patented, and the first computer-controlled, fully automated fullerene production systems are now commercially available.

More simply constructed modifications of this fullerene generator have been developed [472-474], in which vertical graphite electrodes are fed by gravity. This, together with low costs, make them more attractive to synthetic chemists. A simple benchtop reactor gives a yield of fullerenes based on evaporated graphite in the range of 5 – 10 % [474].

Arc Heating of Graphite. Arc vaporization [475-477] of graphite was first developed by SMALLEY [476]. The tips of two sharpened graphite rods are kept at contact-arcing distance [476], so that the electric power is dissipated in an electric arc rather than in ohmic heating. This method allows efficient evaporation of carbon; the yield of fullerenes is about 15 %. The

disadvantage of this method is the intense UV radiation originating from the arc plasma. The absorption of UV light by fullerenes produces a triplet state T_1 with a lifetime of a few microseconds [478]. In the T_1 state, the fullerene is an open-shell system that can readily react with other carbon species C_n to give a nonvaporizable insoluble product [476].

Solar Heating of Graphite. The problem of intense UV radiation is avoided by the use of solar furnaces as fullerene generators [476], [478], which lead to far lower exposure of the fullerenes to radiation than arc vaporization or resistive heating. In the solar generator developed by SMALLEY [476], sunlight is focused by parabolic mirrors onto the tip of a graphite rod under an argon atmosphere in a Pyrex tube. The condensing carbon vapor quickly moves away from the intensive sunlight, cools in the upper regions of the Pyrex tube, and subsequently deposits on the upper walls. Although fullerenes can be obtained in this way, the efficiency of the generator is not very high.

Inductive Heating of Graphite. Fullerenes can also be produced by direct inductive heating of a carbon sample in a boron nitride support [479]. Evaporations at 2700 °C in a helium atmosphere lead to the formation of a fullerene-containing soot, which can be collected on the cooler parts of the Pyrex reactor. This method allows continuous operation by keeping the graphite sample continuously in the heating zone. The evaporation of 1 g of graphite gives 80 – 120 mg of fullerene extract within 10 min.

7.2.2. *Fullerene Synthesis by Combustion*

The existence of fullerenes in sooty flames was first revealed by mass spectrometry [480], and the production of fullerenes in optimized sooty flames was shown to be feasible [481]. Premixed laminar benzene – oxygen – argon flames operated at different pressures, temperatures, and carbon to oxygen ratios have been used. Besides fullerenes and soot, polyaromatic hydrocarbons are formed. The yield of fullerenes and the C_{70}/C_{60} ratio strongly depend on the operating mode. The yield of C_{60} and C_{70} is in the range of 0.003 – 9 % of the soot mass. Expressed as percentage of fuel carbon, the yields vary from 2×10^{-4} to 0.3 %. The C_{70}/C_{60} ratio varies over the range 0.26 to 5.7, which is much larger than that observed for graphite-vaporization methods (0.02 – 0.18). This ratio tends to increase with increasing pressure.

7.2.3. *Pyrolysis of Naphthalene*

Fullerenes can also be obtained by pyrolysis of naphthalene at 1000 °C in an argon stream [482]. The naphthalene skeleton is a monomer of the C_{60} structure. The fullerenes are formed by dehydrogenative coupling reactions. Primary reaction products are polynaphthyls with up to seven naphthalene moieties. Full dehydrogenation leads to both C_{60} and C_{70} in yields less than 0.5 %. Hydrofullerenes such as $C_{60}H_{36}$ have been observed as side products by mass spectrometry.

7.3. Structure and Nomenclature

Fullerenes and fullerene-like nanostructures are defined as polyhedral cages made up entirely of five- and six-membered rings. The structural principle obeys Euler's theorem, which states that for the closure of a spherical network of N hexagons, twelve pentagons are required, with the exception of $N = 1$. Therefore, fullerenes contain $2(10 + N)$ carbon atoms. Hence, the smallest conceivable fullerene is C_{20}, and all fullerenes must contain an even number of carbon atoms. Since it is energetically unfavorable in unsaturated molecules for two pentagons to be adjacent to each other, the pentagons should be isolated by hexagons (*isolated pentagon rule* IPR [483]). For soccer-ball-shaped C_{60} only one IPR isomer is possible; it is the smallest stable fullerene. The rugby-ball-like C_{70} also has only one IPR isomer and is the next stable homologue, followed by the higher fullerenes C_{76}, C_{78}, C_{80}, C_{82}, C_{84}, C_{90}, and C_{96}. With increasing number of hexagons, the number of possible fullerene isomers rises dramatically. For C_{78}, five IPR isomers and more than 20 000 general isomers [484], [485] can be imagined. The number of strain-free fullerene isomers that are possible becomes considerably lower if only IPR structures are allowed. Taking into account that open-shell structures are avoided [485] and that the number of double bonds in pentagons are minimized [486], magic numbers n for stable fullerenes C_n can be predicted as n = 60, 70, 72, 76, 78, 84, . . . [485] (Fig. 69).

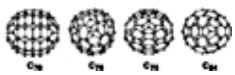

Figure 69. Ball-and-stick models of C_{70}, C_{76}, C_{78}, and C_{84}

C_{60} belongs to the highly symmetric point group I_h (120 symmetry operations). Many of the higher fullerenes are chiral, and the number of possible chiral isomers also increases with increasing cage size. Because of its D_2 symmetry, C_{76} exists in two enantiomeric forms. Two of the five possible IPR structures of C_{78} are chiral, as are ten of the 24 possible IPR structures of C_{84}. These carbon allotropes are rare examples for chiral modifications of pure elements.

Since the IUPAC names of the fullerenes are too lengthy and complicated for general use [487], simpler designations have been developed. For example C_{60} can be named [60-I_h]fullerene, which sufficiently describes its structure. The number 60 denotes the number of carbon atoms, and I_h the symmetry group. Since for C_{60} and C_{70} only one stable IPR isomer exists, their names simplify to [60]fullerene and [70]fullerene. The indication of the symmetry group, however, is important for higher fullerenes, for which different isomers coexist, e.g., [78-D_3]fullerene and [78-C_{2v}]fullerene.

7.4. Physical Properties

The *heat of formation* of [60]fullerene is 42.54 kJ/mol [488] and that of [70]fullerene 40.40 kJ/mol per C atom [489]. It is expected that upon increasing the size of the fullerenes their energy content will asymptotically reach that of graphite [488],

[490].

The *solubility* of [60]fullerene has been investigated systematically [491], [492]. It is essentially insoluble in polar solvents, and sparingly soluble in alkanes. In aromatic solvents and in carbon disulfide appreciable solubilities are observed. A significant increase of the solubility is observed in going from benzenes to naphthalenes. Curiously, its solubility passes through a maximum with increasing temperature [491]. The solubility (in mg/mL) of [60]fullerene in various solvents [491] is as follows:

n-Pentane	0.005
n-Hexane	0.043
Cyclohexane	0.036
n-Decane	0.071
Decalins	4.6
Dichloromethane	0.26
Carbon disulfide	7.9
Dichloromethane	0.26
Chloroform	0.16
Tetrachloromethane	0.32
Tetrahydrofuran	0.000
Benzene	1.7
Toluene	2.8
Tetralin	16
Benzonitrile	0.41
Anisole	5.6
Chlorobenzene	7.0
1,2-Dichlorobenzene	27
1-Methylnaphthalene	33
1-Chloronaphthalene	51
Acetone	0.001
Methanol	0.000

Fullerenes form solvates with many organic solvents. Slow crystallization of C_{60} from benzene gives the solvate $C_{60}(C_6H_6)_4$ [493]. Cocrystallates with ferrocene [494], bis(ethylenedithio)tetrathiafulvalene [495], $[Fe_4CO_4(\eta^5\text{-}C_5H_5)_4]$ [496] iodine [497], 1,4-hydroquinone [498], sulfur (S_8) [499], calixarene [500], and γ-cyclodextrin [501] have been characterized.

The *UV/VIS spectrum* [472] of [60]fullerene exhibits strong absorption bands at 213, 257 and 329 nm, as well as some forbidden transitions in the visible region of the spectrum, which are due to forbidden singlet – singlet transitions and are responsible for the purple color of solutions of [60]fullerene in toluene.

Solutions of [60]fullerene in hydrocarbon solvents are magenta, and those of [70]fullerene are port-wine red. In the same solvents, [76]fullerene gives yellow-green solutions, as does [84]fullerene; [82]fullerene has a less greenish tinge. Solutions of [78]fullerene are golden chestnut brown.

In the solid state, C_{60} molecules crystallize in a face-centered cubic (fcc) *crystal structure* [502] with a lattice constant of 1.417 nm. The density in the solid state is 1.72 g/cm^3 [468]. At 260 K, a first-order phase transition from the fcc to a simple cubic (sc) structure is observed [503], accompanied by a lattice contraction (lattice constant 1.410 nm). At 90 K the molecules rearrange to attain the best global minimum, that is a higher order phase transition. At this temperature, the rotational degrees of freedom are frozen, and the solid structure is designated as a pseudo-fcc structure (Fig. 70). [70] Fullerene is also rotationally disordered at room temperature and undergoes a transition to a low-temperature phase [504].

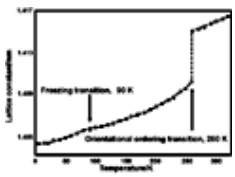

Figure 70. Phase transitions and variation with temperature of the cubic lattice constant of solid [60]fullerene

[60]Fullerene is an insulator but exhibits *electrical conductivity* when doped with electropositive metals. Crystalline phases have been obtained by exohedral doping with alkali and alkaline earth metals, in which the dopants enter the host crystal structure in interstitial lattice positions [505], [506]. The metal atoms are diffused into solid C_{60}, and donate electrons to C_{60} to form C_{60}^{n-} anions. A large variety of alkali metal fullerides M_nC_{60} have been synthesized and studied with respect to their physical and solid-state properties [507-512]. Alkali metal fullerides exhibit superconductivity [513]. The fulleride superconductors have the general formula M_3C_{60} and crystallize in a fcc lattice (except for Cs_3C_{60}), in which the three metal

ions are located in the two octahedral and the tetrahedral site between the spheres. The highest T_C was found for Cs_3C_{60} (40.0 K) [514], followed by $RbCs_2C_{60}$ and Rb_2CsC_{60} (33.0 K and 31.3 K, respectively) [515]. Na_3C_{60} is not superconducting, which is attributed to the fact that below 260 K a reversible structural transition leading to a two-phase mixture of Na_2C_{60} and Na_6C_{60} takes place [512].

Reaction of C_{60} with the organic donor TDAE [tetrakis(dimethylamino)ethylene] leads to the charge-transfer complex $[TDAE^+]C_{60}^-$, which has ferromagnetic properties [516], [517]. This compound exhibits the highest magnetic ordering temperature T_C = 16.1 K of organic molecular ferromagnets.

7.5. Chemical Properties

Reviews on the synthetic organic chemistry of fullerenes can be found in [518-522].

7.5.1. *Addition Reactions*

The additions of nucleophiles [518], [519], [523] and radicals [524] to the electrophilic [6,6] double bonds ([6,6] denotes the junction of two hexagons) are among the earliest observed reactions in fullerene chemistry. On addition of nucleophiles (Nu) the initially formed anionic intermediates $Nu_nC_{60}^{n-}$ can be stabilized by addition of electrophiles E^+ to give substituted fullerenes $C_{60}ENu$ or by an S_Ni reaction to give methanofullerenes. Suitable sources of carbon nucleophiles are Grignard or organolithium reagents RMgX and RLi, which form the anions RC_{60}^- to $R_nC_{60}^{n-}$ as primary intermediates [523], [525], [526]. Their protonation yields the hydrofullerene derivatives $C_{60}H_nR_n$ (Fig. 71). Highly alkylated anions $R_nC_{60}^{n-}$ have been quenched with MeI to form $C_{60}Me_nR_n$ [525]. The generation of carbon nucleophiles by deprotonation of α-halo esters or α-halo ketones in the presence of fullerenes leads to cyclopropanation [527]. Cyclopropanated fullerenes are also formed on treating of C_{60} with phosphonium ylides [528]. [60]Fullerene reacts readily with primary and secondary aliphatic amines [523], [529], but often complex mixtures of amino adducts are obtained. Isomerically pure mono adducts are obtained by the reaction of secondary diamines such as *N,N*-dimethylethylenediamine, piperazine, or homopiperazine with C_{60} [530] (Fig. 71).

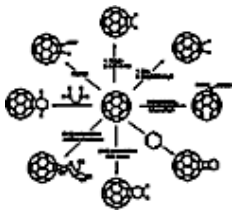

Figure 71. Typical addition reactions of [60]fullerene

Like olefins, [60]fullerene undergoes many thermal, photochemical, and electron-transfer-activated cycloadditions. In [4 + 2] cycloadditions, the [6,6]-bonds of C_{60} act as dienophiles [518], [523], [531]. Methano-bridged fullerenes are accessible bycycloaddition of diazomethanes, diazoactates, and diazoamides to C_{60} [518], [523]. The reaction of azides with C_{60} gives aza-bridged fullerenes [532] (Fig. 71).

The addition of benzyne and electron-rich alkynes to [60]fullerene or the photopolymerization of C_{60} leads to [2 + 2]-cycloadducts [533]. [2 + 1]-Cycloadducts are formed by addition of carbenes or of silylenes to [60]fullerene [518], [522].

The facile nucleophilic addition of primary and secondary amines to C_{60} has been used to synthesize fullerene-based polymers and polymer-bound C_{60} [534]. Self-assembled monolayers of covalently bound C_{60} have been synthesized by treatment of C_{60} with $(MeO)_3Si(CH_2)_3$-modified iridium tin oxide surfaces [535] and cysteamine-modified gold surfaces [536].

The strongly electron withdrawing fullerenes behave like “radical sponges”. Radicals can readily add to [60]- and [70]fullerene to form fullerenyl radicals. Radicals in which up to 34 methyl groups are attached to the cage have been produced [529]. A number of alkyl-, benzyl-, phenyl-, and thiomethyl-substituted fullerenyl radicals have been obtained [529], [537]. The addition of radicals has been used to synthesize fullerene polymers and perfluoroalkylated fullerenes.

7.5.2. *Hydrogenation*

Dihydrofullerene $C_{60}H_2$ is synthesized by hydroboration [538], hydrozirconation [539], and zinc/acid reduction [540]. Of the 23 possible regioisomers, only the 1,2-adduct is formed. The soluble $[(\eta^5\text{-}C_5H_5)_2ZrClC_{60}H]$ complex obtained by hydrozirconation undergoes further hydrozirconation [539]. Hydrolysis of the higher adducts leads to $C_{60}H_4$ and $C_{60}H_6$ as mixtures of regioisomers. Hydroboration of [70]fullerene followed by hydrolysis gives two isomers of $C_{70}H_2$. Exhaustive hydrogeneration of C_{60} leads to mixtures of unstable polyhydrofullerenes $C_{60}H_{2n}$ [473], [541]. Attempts to synthesize the fully hydrogenated $C_{60}H_{60}$ have so far failed.

7.5.3. *Halogenation*

[60]Fullerene is fluorinated by XeF_2 [542] or fluorine gas [543] to give mixtures of polyfluorofullerenes $C_{60}F_{2n}$ (e.g., n = 15 – 22). A defined regioisomer of $C_{60}F_{48}$ was synthesized by reaction of solid C_{60} with fluorine gas in the presence of NaF

[544]. Higher degrees of fluorination are obtained by reacting perfluorinated fullerenes with fluorine under UV radiation [544]. The most extensively fluorinated species is $C_{60}F_{102}$, which requires rupture of the fullerene framework. Polychlorination of C_{60} can be achieved with chlorine gas at 250 – 400 °C [545] or liquid chlorine at – 35 °C [546]. The chlorofullerenes are less stable than their fluorine analogues. Treatment of [60]fullerene with liquid bromine gives the yellow-orange, crystalline $C_{60}Br_{24}$ [547]. Reaction of C_{60} with iodine occurs on irradiating a $C_{60} – I_2$ mixture at 130 °C to give $C_{60}I_2$ [548].

7.5.4. *Reduction*

Reductions were the first chemical transformations carried out with fullerenes. Electron transfer reactions with electropositive metals, and organic donor reagents, as well as electrochemical reductions have been used to prepare fulleride salts. The fulleride ions readily undergo subsequent reactions with electrophiles.

Fullerenes are readily reduced by alkali and alkaline earth metals. Thus, titration of a suspension of C_{60} in liquid ammonia with a solution of Rb in ammonia or ultrasound-aided treatment with Li metal in THF generates fullerides C_{60}^{n-} (n = 1 – 5 for Rb^+, n = 1 – 24 for Li^+). Defined compounds M_nC_{60} with M = Na, K, Rb, Cs and n = 2, 3, 4, and 6 [507-510] as well as mixed alkali compounds $M_n^1M_n^2C_{60}$ can be obtained [510]. Sodium and potassium fullerides can also be synthesized by solid-state reaction of C_{60} with NaH, KH, $NaBH_4$, or KBH_4 [549]. Other techniques apply vapor-transport procedures [550] or binary alloys [509] as the alkali metal source. A bulk reduction of C_{60} with the less electropositive mercury was also demonstrated [551].

Fulleride salts are also formed with organic or organometallic donor molecules. The reduction with donors like (tetraphenylporphinato)chromium(II), cobaltocene, bis(*N*-methylimidazole)(tetraphenylporphinato)tin(II), tetrakis (dimethylamino)ethylene (TDAE) or $[Fe^I(C_5H_5)(C_6Me_6)]$ leads to fullerene-based charge transfer complexes, some of which show remarkable electronic and magnetic properties [531], [552].

7.6. Endohedral Fullerenes

The hollow fullerene molecules are potential hosts for trapping smaller guest molecules in the spherical carbon cage. Such structures were first obtained with lanthanide guests and are usually referred to as endohedral fullerenes. Their discovery in 1985 stemmed from the observation of La-related peaks in the mass spectra of material that was obtained by laser vaporization of graphite soaked in $LaCl_3$ solution [553]. They were formulated as $La@C_{60}$ and $La_2@C_{60}$. (To assign the position of the metal with an appropriate nomenclature an “@” is added to the name, and endohedral heteroatoms are placed to the left of the @, whereas heteroatoms of the fullerene framework are positioned to the right of the @.)

Endohedral metal fullerene complexes are prepared by laser and arc vaporization of graphite impregnated with metal salts, metal oxides or carbides [554]. Among others, Li, K, Rb, Cs, Ca, Sc, Y, La, Ce, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Lu, U, Ti, Zr, Hf, Fe, and Co complexes can be prepared in this way, with various fullerene cage sizes. For the larger fullerenes, the endohedral addition of up to four metal ions has been demonstrated by mass spectrometry [555]. The interior of C_{82}, which is approximately 8 Å in diameter, can readily accommodate three rare earth trivalent ions [556].

Atoms of the noble gases He or Ne are incorporated inside C_{60} by arc vaporization of carbon electrodes in the presence of these gases [557]. The first helium compounds $He@C_n$ (n = 60, 70, 76, 78, 84) were prepared by heating fullerenes to 600 °C in a helium atmosphere at 3000 atm [558]. Also Ne, Ar, Kr, and Xe were successfully introduced into fullerenes [558] by heating under high-pressure or by the impact of high-energy ion or atom beams [559]. The endohedral fullerene $N@C_{60}$ [560] is produced by bombardment of C_{60} with N_2^+ and N^+ ions or in a glow discharge reactor [560].

7.7. Carbon Nanotubes and Nanostructures

In 1991 IIJIMA observed by high-resolution transmission electron microscopy the formation of multilayered *carbon nanotubes* on the surface of a carbon arc cathode [562]. These tubules, which consist of several coaxial tubes and a hollow core, have typical lengths of up to 1000 nm; the outer diameter ranges between 2 and 20 nm, and the inner diameter between 1 and 3 nm. The longest reported *carbon nanotube* length was 700 nm for a 0.9 nm diameter tubule [563]. The *multi-walled nanotubes* MWNT (“buckytubes”) can consist of 12 to 50 coaxial tubes, each of which is formally a graphite sheet rolled to a cylinder and closed at each end by a hemifullerene end cap (Fig. 72).

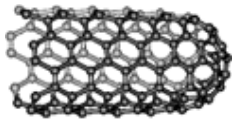

Figure 72. Singe-wall nanotube sealed at one end by a hemi-[60]fullerene cap

Carbon nanotubes have been generated by carbon arc discharge [562], [564], pyrolysis of hydrocarbons [565] or triazine over catalysts [566], carbon vapor deposition [567], condensed phase electrolysis [568], and carbon ion bombardment [569]. For the preparation of single-wall nanotubes, the carbon anode is modified by addition of a transition metal, which serves as a catalyst [570]. In 1996, an improved method for making single-wall nanotubes became available [571]; the tubes are obtained in bundles in about 50 % yield (cf. 1 % yield of soluble fullerene material). Instead of single tubes, ropes consisting of 10 – 100 parallel nanotubes are found. The smallest bundle is a microbundle consisting of 10 to 100 aligned nanotubes of almost equal length.

Carbon nanotubes are extremely strong; the Young's modulus of individual MWNTs is 0.4 – 3.7 TPa (experimental) [572], and a value of 5.5 TPa has been calculated [573] for MWNTs and SWNTs. Single-walled tubes are remarkably flexible and

bend into curved arcs with radii as small as 20 nm. Nanotubes bent to an angle of 30 ° develop kinks, but the hexagonal pattern remains intact up to 110 ° [574].

Carbon nanocones are found on the caps of nanotubes and also as free-standing structures generated in a carbon arc. Concentric hollow carbon spheres with onion-like structure [575] are formed upon intense electron beam irradiation of carbon particles, by shock-wave treatment of carbon soot, from carbon exposed to a plasma torch, by laser melting of carbon, and by annealing nanodiamonds at temperatures above 1000 °C. These *carbon onions* ("buckyonions") contain no dangling bonds and are remarkably stable. An innermost sphere with an inner diameter of 7.1 Å, corresponding to the diameter of the C_{60} molecule, was reported in one case.

7.8. Application Properties of Fullerenes

Solid fullerenes are compressible because of the weak van der Waals forces between the fullerene molecules in the lattice. Various high-pressure phases of C_{60} and C_{70} have been characterized [576]. When compressed to 70 percent of its original size, C_{60} becomes more than twice as hard as diamond. Thus, it could be applied as a shock absorber for high-energy collision or for reinforcing metals or plastics. Cross-linked polymers and high-performance polymer films can be obtained by cross-linking C_{60} with resins. These materials are very hard and thermally extremely stable.

In C_{60} – polymer composites, fast photoinduced electron transfer occurs from the polymer to the C_{60} molecule and photoconductivity is observed [577]. The efficiency of organic photoconductors could be improved by addition of C_{60} [577], [578]. This discovery may lead to applications in xerography, laser printers, fax machines.

Fullerene-based transistors have been constructed [579], and a variety of $M/C_{60}/M$ (M = metal) diodes and field effect transistors (FETs) proposed. Proposed application in the fabrication of microelectronics, include films for patterning, surfaces of uniform electric potential for electronic applications [580], passivation of surfaces [581], and adhesives for bonding silicon wafers [582].

The possibility of using C_{60} for hydrogen storage [583] might be useful not only in energy storage cells but also in high-energy-density primary batteries.

Investigations of the lubricating properties of fullerene films [584] showed that fullerenes may be of interest as high-temperature lubricants and anticorrosion coatings. The most important application of fullerenes in lubrication engineering is as additives in oils and greases [585]. By total fluorination of C_{60} a fully fluorinated "teflonball" $C_{60}F_{60}$ should be accessible which is assumed to represent the slickest molecular lubricant known to man. However, highly fluorinated C_{60} is unstable in the presence of traces of water [586].

[Top of Page]

References

Acknowledgment. The authors of Chapter Carbon black would like to thank GERHARD KÜHNER (formerly Degussa AG, Frankfurt) for his contribution to Section Toxicology and Occupational Health

General References

1. *Gmelin*, System no. 14 Carbon, Sect. B, parts 1, 2, and 3.
2. A. Pacault (ed.): *Les Carbones*, **vols. 1 and 2**, Masson et Cie., Paris 1965.
3. A. R. Ubbelohde, F. A. Lewis: *Graphite and its Crystal Compounds*, Clarendon Press, Oxford 1960.
4. G. M. Jenkins, K. Kawamura: *Polymeric Carbons – Carbon Fibre*, Glass and Char, Cambridge University Press, Cambridge-London-New York-Melbourne 1976.
5. K. Arndt: *Die künstlichen Kohlen für elektrische Öfen*, Elektrolyse und Elektrotechnik, 2nd ed., Springer Verlag, Berlin 1932.
6. A. S. Fialkov: *Graphitic Materials (in Russian)*, Energia, Moscow 1979.
7. J. F. Tschalych: *Herstellung von Kohle- und Graphitelektroden*, VEB-Verlag für Grundstoffind., Leipzig 1961.
8. C. L. Mantell: *Carbon and Graphite Handbook*, Interscience Publ., New York-London-Sidney-Toronto 1968.
9. T. Ishikawa, T. Nagaoki: *Recent Carbon Technology*, JEC Press Inc., Cleveland 1984.
10. L. C. F. Blackman: *Modern Aspects of Graphite Technology*, Academic Press, London 1970.
11. R. E. Nightingale (ed.): *Nuclear Graphite*, Academic Press, New York-London 1962.
12. E. I. Shobert II: *Carbon Brushes*, Chemical Publ., New York 1965.
13. *Winnacker-Küchler*, vol. **3**, pp. 278 – 308.
14. J. B. Donnet, R. C. Bansal, M.-J. Wang (eds.): *Carbon Black*, 2nd ed., Marcel Dekker, New York 1993.
15. H. Ferch: "Pigmentruße" in U. Zorll (ed.): *Die Technologie des Beschichtens*, Vincentz Verlag, Hannover 1985.
16. H. W. Kroto, J. E. Fischer, D. E. Cox: *The Fullerenes*, Pergamon Press, Oxford 1993.
17. W. E. Billups, M. A. Ciufolini (eds.): *Buckminsterfullerenes*, VCH, New York 1993.
18. R. S. Ruoff, K. W. Kadish (eds.): *Fullerenes: Recent Advances in the Chemistry and Physics of Fullerenes and Related Materials*, The Electrochemical Society Proceedings Series, Pennington, NJ, from 1994.
19. J. Dettmann: *Fullerene – Die Buckyballs erobern die Chemie*, Birkhäuser, Basel 1994.
20. A. Hirsch: *The Chemistry of the Fullerenes*, Thieme, Stuttgart – New York 1994.

21. J. Baggott: *Perfect Symmetry*, The Accidental Discovery of Buckminsterfullerene, Oxford University Press, Oxford 1994.
22. P. Ball: *Designing the Molecular World – Chemistry at the frontier*, Princeton University Press, 1994.
23. H. Aldersey-Williams, *The Most Beautiful Molecule*, The Discovery of the Buckyball, Wiley, New York 1995.
24. P. W. Fowler, D. E. Manolopoulos, *An Atlas of Fullerenes*, Clarendon Press, Oxford 1995.
25. R. Taylor (ed.), *The Chemistry of Fullerenes*, **Advanced Series in Fullerenes, vol. 4**, World Scientific, Singapore – New Jersey – London – Hong Kong 1995.
26. W. Krätschmer, H. Schuster, *Von Fuller bis zu Fullerenen*, Vieweg, Braunschweig 1996.
27. M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund, *Science of Fullerenes and Carbon Nanotubes*, Academic Press, San Diego 1996.
28. T. W. Ebbesen (ed.), *Carbon Nanotubes*, CRC Press, Boca Raton 1997.

Specific References

29. E. Broda, T. Schönfeld, *Österr. Chem. Ztg.* **54** (1953) 209.
30. H. Marsh, F. Rodriguez-Reinoso (eds.): *Science of Carbon Materials*, University of Alicante, Alicante 2000.
31. V. R. Howes, *Proc. Phys. Soc. London* **80** (1962) 648.
32. F. P. J. Bundy, *Chem. Phys.* **38** (1963) 631.
33. H. M. Hannemann, H. M. Strong, F. P. Bundy, *Science* **155** (1967) 995.
34. D. M. Haaland, *Carbon* **14** (1976) 357. Links
35. A. el Goresey, G. Donnay, *Science* **161** (1968) 363.
36. R. B. Heinmann, S. E. Evsyukov, L. Kavan (eds.): *Carbyne and Carbynoid Structures*, Kluwer Academic Publ., Deventer 1999
37. A. G. Whittacker, *Carbon* **17** (1979) 21.
38. A. G. Whittaker, M. N. Neudorfer, E. J. Watts, *Carbon* **21** (1983) 597. Links
39. V. M. Melnitchenko, B. N. Smirnov, V. P. Varlakov, Y. N. Nikulin, *Carbon* **21** (1983) 131. Links
40. H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, R. E. Smalley, *Nature* **318** (1985) 162 – 163. Links
41. W. Krätschmer, L. D. Lamb, K. Fostiropoulos, D. R. Huffman, *Nature* **347** (1990) 354.
42. H. E. Spear, J. P. Dismukes (eds.): *Synthetic Diamond*, Wiley, New York 1994.
43. H. Marsh, E. A. Heintz, F. Rodriguez-Reinoso (eds.): *Introduction to Carbon Technologies*, Publ. University of Alicante, Alicante 1997.
44. M. A. Prelas, G. Popovici, L. K. Bigelow (eds.): *Handbook of Industrial Diamond*, Marcel Dekker, Inc., New York 1997.
45. B. Dischler, C. Wild (eds.): *Low-Pressure Synthetic Diamond*, Springer-Verlag, Berlin 1998.
46. General Electric Comp., US 2 992 900, 1961 (H. P. Bovenkerk).
47. General Electric Comp., FR 1 341 561, 1963 (M. R. H. Wentorf).
48. ASEA, FR 1 346 568, 1963.
49. B. V. Spitsyn, B. V. Derjaguin, patent application 10.7.56.
50. US 3 0301 87 and 3 030 188, 1962 (W. Eversole)
51. J. E. Field (ed.): *The Properties of Diamond*, Academic Press, London 1979.
52. R. Berman, *Physical Properties of Diamond*, Clarendon Press, Oxford 1965.
53. V. R. Howe, *Proc. Phys. Soc. London* **80** (1962) 648.
54. F. P. Bundy, *J. Geophys. Res.* **85 (B12)** (1980) 6930.
55. J. Abrahamson, *Carbon* **12** (1974) 111. Links
56. F. P. Bundy, *J. Chem. Phys.* **38** (1963) 631. Links
57. F. P. Bundy, H. M. Strong, R. H. Wentdorf in P. L. Walker, P. A, Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **10**, Marcel Dekker, New York 1973.
58. P. W. Bridgman, *Sci. Am.* **193** Nov. 1955, p. 42.
59. F. P. Bundy, H. T. Hall, H. M. Strong, R. H. Wentdorf, Jr., *Nature* **176** (1955) 51. Links
60. H. P. Bovenkerk, F. P. Bundy, H. T. Hall, H. M. Strong, R. H. Wentdorf, Jr., *Nature* **184** (1959) 1094. Links
61. Y. A. Litvin, *Inorganic Materials* **4** (1968) 144.
62. K. Nassau, “Synthesis of Bulk Diamond ”, in: R. F. Davis (ed.): *Diamond Films and Coatings*, Noyes Publications, Park Ridge 1993.
63. R. H. Wentdorf, Jr., *Phys. Chem.* **69** (1965) 3063.
64. K. Subramanian, V. R. Shanbhag: “Abrassive Applications of Diamond” in M. A. Prelas, L. K. Bigelow (eds.): *Handbook of Industrial Diamonds and Diamond Films*, Marcel Dekker, New York 1998, "chapt. 28".
65. W. Morrow, *Lapidary J.* **17** (1963) 922.
66. R. C. DeVries, *Annu. Rev. Mater. Sci.* **17** (1987) 161.
67. J. C. Angus, C . C. Hayman, *Science* **241** (1988) 913.
68. K. E. Spear, *J. Am. Ceram. Sci.* **72** (1989) 171.
69. F. J. Celii, J. E. Butler, *Annu. Rev. Phys. Chem.* **42** (1991) 643. Links
70. R. E. Clausing, L. L. Horton, J. C. Angus, P. Koidl (eds.): *NATO ASI Series, Series B: Physics*, Vol. **266**, 1990.
71. J. E. Butler, R. L. Woodin, *Philos. Trans. R. Soc. Lond.* **A342** (1993) 209.
72. K. E. Spear, M. Frenklach, *Pure Appl. Chem.* **66** (1994) 1773. Links

73. K. E. Spear, M. Frenklach, *Synthetic Diamond:Emerging CVD Science and Technology*, Wiley, New York 1994.
74. H. Liu, D. S. Dandy, *Diamond Chemical Vapor Deposition: Nucleation and Early Growth Stages*, Noyes Publications, Park Ridge, NJ, 1995.
75. B. V. Deryaguin, D. V. Fedossejev, *Growth of Diamond and Graphite from the Gas Phase*, Nauka, Moscow 1977.
76. Z. P. Lu, J. Heberlein, E. Pfender, *Plasma Chem. Plasma Proc.* **12** (1992) 35.
77. I. Watanabe, T. Matsushita, K. Sasahara, *Jpn. J. Appl. Phys.* **31** (1992) 1428.
78. Y. Liou, A. Inspektor, R. Weimer, R. Messier, *Appl. Phys. Lett.* **55** (1989) 631. Links
79. I. Schmidt, C. Benndorf, *Diamond and Related Materials* **7** (1998) 266. Links
80. J. W. Lindsay, J. M. Larson, S. L. Girshick, *Diamond and Related Materials* **6** (1997) 481.
81. Z. P. Lu, J. Heberlein, E. Pfender, *Plasma Chem. Plasma Proc.* **12** (1992) 55 – 69.
82. Status and Applications of Diamond and Diamond-Like Materials, National Materials Advisory Board (NMAB)-445, Academy Press, 1995.
83. A. B. Harker, J. F. Natale, Diamond Depositions, Science and Technology, Aug. 1991, 12 – 13.
84. K. E. Spear, M. Frenklach, *Pure Appl. Chem.* **66** (1984) 1773.
85. A. Boudina et al., *Proc. Diamond Films*, 1991, Nizza.
86. W. Zhu, R. Messier, A. R. Badzian, *Proc. 1st Int. Symp. Diamond & Diamond-like Films*, Electrochem. Soc. Pennington, NJ, 1989, 66.
87. W. A. Yarbrough, R. Messier, *Science* **274** (1990) 688.
88. Q. D. Zhuang, H. Guo, J. Heberlein, E. Pfender, *Diamond and Related Materials* **3** (1994) 319. Links
89. S. K. Baldwin, T. Owano, M. Zhao, C. H. Kruger, *Diamond and Related Materials* **6** (1997) 202 – 206.
90. M. T. Bieberich, S. L. Girshick, *Plasma Chem. Plasma Process.* **16** (1996) Suppl., pp. 157/s – 168/s.
91. M. Breiter, Ch. Doppleb, K. H. Weiss, G. Nutsch, *Diamond and Related Materials* **9** (2000) 333 – 336. Links
92. J. Heberlein, E. Pfender, *2nd Plasma-Technik-Symposium*, 1992, Vol. **1**, p. 187.
93. J. M. Larson, S. L. Girshick, Proc. 14th Int. Symp. on Plasma Chemistry, 1999, Prague, 1663.
94. A. Higa, A. Hatta, T. Ito, M. Toguchi, A. Hiraki, *Diamond Films and Technology* **6** (1996) 277.
95. A. Hirata, M. Yoshikawa, *Diamond and Related Materials* **4** (1995) 1163.
96. A. Higa, A. Hatta, T. Ito, T. Maehama, *Jpn. J. Appl. Phys. 2* **35** (1996) L577. Links
97. P. K. Bachmann *et al.*, *Diamond and Related Materials* **1** (1991) 1021 – 1034.
98. P. K. Bachmann, H. Lade, D. Leers, D. U. Wiechert, *Diamond and Related Materials* **3** (1994) 79.
99. R. A. Rudder, *et al.*, *Appl. Phys. Lett.* **59** (1991) 791. Links
100. M. S. Wong, C. H. Wu, *Diamond and Related Materials* **1** (1992) 369. Links
101. Assmann, M. Heberlein, J. E. Pfender, *Diamond and Related Materials* **8** (1999) 1 – 16.
102. Assman, M. Nelson, K. Heberlein, J. E. Pfender, *Proc. ISPC 14*, 1999, Praha, p. 1657.
103. F. Deuerler, H. van den Berg, R. Tabersky, A. Freundlieb, M. Pies, V. Buck, *Diamond and Related Materials* **5** (1996) 1478 – 1489.
104. S. K. Baldwin, T. G. Owano, C. H. Kruger, *Appl. Phys, Lett.* **67** (1995) 194.
105. M. I. Boulos, P. Fauchais, E. Pfender, *Thermal Plasmas*, Plenum Press, New York 1995.
106. B. W. Yu, L. Si, S. Girshick, *J. Appl. Phys.* **75** (1994) 3914 – 3923. Links
107. P. K. Bachmann, “Microwave Plasma Chemical Vapor Deposition of Diamond”, in M. A. Prelas *et al.* (eds.): *Handbook of Industrial Diamonds and Diamond Film* Marcel Dekker, New York 1997.
108. M. Kamo, Y. Sato, S. Matsumoto, N. Seteka, *J. Cryst. Growth* **62** (1983) 642. Links
109. M. Füner, C. Wild, P. Koidl, *Surface & Coatings Techn.* **74 – 75** (1995) 221 – 226.
110. P. K. Bachmann *et al.* in A. Badzian *et al.* (eds.): Diamond and Diamond-like Materials, *MRS Symp. Proc.* **EA-15** (1988) p. 99.
111. AIXTRON AG, Aachen, Germany, Prospectus.
112. M. A. Capelli, “Arcjet Synthesis of Diamond”, in M. A. Prelas *et al.* (eds.): *Handbook of Industrial Diamonds and Diamond Film*, Marcel Dekker, New York 1997.
113. D. S. Dandy, M. E. Coltrin, *Appl. Phys. Lett.* **66** (1995) 391 – 393. Links
114. H. M. Loh, M. A. Capelli, *Surface and Coatings Techn.* **54/55** (1992) 408 – 413.
115. Juchmann, W. Luque, J. Wolfrum, J. J. B. Jeffries, *Proc. Gas Discharge Applications*, 1997, Greifswald, Germany, pp. 346 – 349.
116. H. M. Loh, M. A. Capelli, *Proc. 3rd Int. Symp. Diamond Materials*, Vol. **93 – 17**, pp. 17 – 23, Electrochem. Soc., Honolulu, HI, 1993.
117. J. Laimer, H. Pauser, H. Störi, R. Haubner, B. Lux, *Diamond and Related Materials* **6** (1997) 406 – 410.
118. M. H. Loh, J. G. Liebeskind, M. A. Capelli, AIAA 94–3233, *30th Joint Propulson Conference*, Indianapolis, IN, 1994.
119. G. Nutsch, B. Dzur, *Elektrowärme Int.* **B** (1995) B201.
120. G. Nutsch: “Anwendung des Thermischen Plasmas in der Werkstofftechnik,” Habilitationschrift, TU Ilmenau, 1995.
121. R. Hernberg, T. Lepistö, T. Mäntylä, T. Sternberg, J. Vattulainen, *Diamond and Related Materials* **1** (1992) 255. Links
122. M. Kohzaki, K. Uchida, K. Higuchi, S. Noda, *Jpn. J. Appl. Phys. Part 2* **32** (1993) L 438.
123. Löschen??? M. Rahmane, G. Soucy, M. I. Boulos, *J. High Temp. Chem. Processes* **3** (1996).
124. S. L. Girshick, C. Li, B. W. Yu, H. Han, *Diamond and Related Materials* **2** (1993) 1090. Links

125. M. Hollenstein, M. Rahmane, M. J. Boulos, *Proc. 14th Int. Symp. Plasma Chem.*, Prague, 1999, p. 257.
126. B. Dzur, G. Nutsch, *Proc. UTSC 1999*, Essen.
127. T. G. Owano, D. G. Goodwin, C. H. Kruger, M. A. Capelli, 10th Int. Symp. Plasma Chemistry (ISPC 10) 1991, Bochum, Germany, 3. 1–8 pp 1 – 6.
128. E. Theophile, F. Gitzhofer, M. I. Boulos, Proc. ISPC, Prague 1999, pp. 2145 – 2148.
129. M. Breiter, G. Nutsch: "Diamond and Releated Materials", Proc 10th Europ. Conf. Diamond & Diamond-like Materials, Prague 1999, pp. 333 – 336.
130. S. P. Chauhan, J. C. Angus, N. C. Gardner, *J. Appl. Phys.* **47** (1976) 1031 – 1033.
131. L. Schaefer, C. P. Klages, U. Meier, K. K. Honighaus, *J. Appl. Phys. Lett.* **58** (1991) 571 – 573.
132. C. T. Lynch (ed.): *Practical Handbook of Materials Science*, CRC Press, Boca Raton 1989.
133. G. B. Gaines, C. T. Sims, R. I. Jaffee, *J. Electrochem. Soc.* **106** (1959) 881 – 885. Links
134. C. P. Klages: "Untersuchungen zu Herstellung und Eigenschaften von Diamantschichten (Study of Production and Properties of Diamond Coatings), Habilitationsschrift, TU Braunschweig, 1998.
135. Fraunhofer Institut Schicht- und Oberflächentechnik, Diamanttechnologie, Prospectus.
136. Y. Hirose, S. Amanuma, K. Komaki, *J. Appl. Plugs.* **68** (1990) p. 6401.
137. C. A. Wolden, Z. Sitar, R. F. Davis, "Combustion Flame Deposition of Diamond" in B. Dischler, C. Wild (eds.): *Low-Pressure Synthetic Diamond*, Springer Series in Materials Processing, Springer-Verlag, Berlin 1998, pp. 41 – 58.
138. J. G. Buinsters, F. M. Bouwelen, J. J. Schermer, W. J. P. van Enckevort, J. J. ter Meulen, *Diamond and Related Materials* **9** (2000) 341 – 345.
139. R. Berman: *Physical Properties of Diamond*, Clarendon Press, Oxford 1965.
140. J. D. Bernal, *Proc. R. Soc. London Ser.* **106** (1924) 749.
141. R. Franklin, *Acta Crystallogr.* **4** (1951) 253.
142. J. Mering, J. Maire, *J. Chim. Phys. Phys. Chim. Biol.* **57** (1960) 803.
143. C. R. Houska, B. E. Warren, *J. Appl. Phys.* **25** (1954) 1503. Links
144. W. Ruland in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **4**, Marcel Dekker, New York 1968, p. 1.
145. B. T. Kelly: *Physics of Graphite*, Applied Science Publishers, London 1981.
146. A. K. Dutta, *Phys. Rev.* **90** (1953) 187.
147. E. Poquet, N. Lumbroso, J. Hoaran, A. Marchand, A. Pacault, D. E. Soule, *J. Chim. Phys. Phys. Chim. Biol.* **57** (1960) 866.
148. R. O. Grisdale, A. C. Pfister, W. van Roosbrock, *Bell Syst. Tech. J.* **30** (1951) 271.
149. W. N. Reynolds: *Physical Properties of Graphite,* Elsevier, New York 1968.
150. H. Spatzek, F. Mayr-Melnhof, *Carbon '72*, Int. Carb. Conf., Baden-Baden 25 – 30 Jun. 1972, p. 351.
151. K. Erhard, personal communication.
152. L. Mackles, R. A. Heindl, L. E. Mong, *J. Am. Ceram. Soc.* **36** (1953) no. 8, 266. Links
153. H. Spatzek, *Montan. Rundsch. Sonderh. Steine und Erden* 1955, 313; 1967, 34.
154. B. R. Bjørnstad, NO 56 633, 1935.
155. Graphitwerk Kropfmühl, DE U1413, 1951.
156. L. Pettifer, *Ind. Miner. (London)* **156** (1980) 19.
157. K. Andrews, *Fachber. Hüttenprax. Metallweiterverarb.* **20** (1982) 905.
158. K. Lipsdorf, R. Schäfer, G. Spencer, E. Wege in: *Fuel Element Fabrication 1*, Academic Press, London-New York 1961, p. 179.
159. Nukem GmbH, GB 1 137 939, 1965.
160. J. H. Castner, GB 19 089, 1893.
161. E. G. Acheson, US 568 323, 1895.
162. J. C. Bokros in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **5**, Marcel Dekker, New York 1969, p. 1.
163. L. M. Currie, V. C. Hamister, H. G. MacPherson, *Proc. Int. Conf. Peaceful Uses At. Energy* **8** (1956) 451.
164. B. Lux, *Giessereiforschung* **22** (1970) 65, 161.
165. H. P. Boehm, *Carbon* **16** (1978) 385. Links
166. United Aircraft Corp., FR 1 433 002, 1965 (W. E. Arnoldi, P. J. Birbara, S. Russell).
167. NRDC, GB 1 110 791, 1964 (W. Johnson, L. N. Philips, W. Watt).
168. J. D. Brooks, G. H. Taylor in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **4**, Marcel Dekker, New York 1968, p. 243.
169. H. Marsh, P. L. Walker in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **15**, Marcel Dekker, New York 1979, p. 229.
170. E. Fitzer, K. Müller, W. Schaefer in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **7**, Marcel Dekker, New York 1971, p. 237.
171. I. C. Lewis, *Carbon* **20** (1982) 519. Links
172. A. R. Ubbelohde, D. A. Young, A. W. Moore, *Nature (London)* **198** (1963) 1192.
173. R. Marrett, J. W. Stadelhofer, H. Marsh, *Chem. Ing. Tech.* **55** (1983) no. 1, 1.
174. S. Wilkening, *Erzmetall* **29** (1976) no. 6, 255.
175. G. C. Hughes, B. J. Doerksen; A. M. Engleholm, S. Romero, 3rd International Conference on Refining Processing, AIChE Spring National Meeting, March 5 – 9, 2000, Atlanta, GA, pp. 67 – 112.

176. Great Lakes Carbon Corp., US 2 775 549, 1954 (F. L. Shea, Jr.,
177. J. L. White in R. F. Gould (ed.): *Petroleum Derived Carbons*, American Chem. Soc., Washington 1976, p. 282.
178. K. Hayashi, M. Nakaniwa, M. C. Sze, A. A. Simone, *Erdöl Kohle Erdgas Petrochem.* **30** (1977) no. 2, 65.
179. I. Romey, H. Glaser, R. Marrett, H. Tillmanns, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) 562.
180. G. M. Kimber, A. Brown, J. N. Kirk: "Carbon '80," *3rd Int. Carb. Conf.,* Baden-Baden Jun. 30 – Jul. 4, 1980, p. 443.
181. Bergwerksverband, Sigri Elektrographit, US 4 289 604, 1977 (G. Kölling, G. Pietzka, I. Romey, H. Tillmanns).
182. Kaiser Aluminium & Chemical Corp., CA 666 055, 1960 (L. F. Hunt).
183. Lummus Comp., US 3 617 515, 1969 (W. J. Bloomer).
184. Bergwerksverband, DE 958 278, 1955 (H. Echterhoff, G. Heinze).
185. C. B. Scott, H. O. Folkins, *J. Met.* **24** (1972) 7, 25.
186. M. Kakuta, H. Tanaka, J. Sato, K. Noguchi, *Carbon* **19** (1981) 347. Links
187. Salem Brosius Comp., US 3 227 627, 1962.
188. Superior Graphite Comp., US 3 684 446, 1970.
189. W. Gemmeke, G. Collin, M. Zander in J. J. Miller (ed.): *Light Metals 1978*, vol. **1**, AIME, New York 1978, p. 355.
190. R. W. Wallouch, H. N. Murty, E. A. Heintz, *Carbon* **10** (1972) 729. Links
191. J. W. Newman in C. J. McMinn (ed.): *Light Metals 1980*, AIME, New York 1979, p. 503.
192. Siemens-Plania AG, DE 969 619, 1958 (F. Jeitner, E. Nedopil).
193. Hoechst, DE 900 569, 1953 (O. Peter).
194. Great Lakes Carbon Corp., US 2 500 209, 1950 (F. L. Shea, L. H. Juel).
195. C. H. Riesz, S. Susman, *Proc. Conf. Carbon 4th 1959* 1960, 609.
196. S. Mrozowski, *Proc. Conf. Carbon 1st, 2nd 1953 and 1955* 1956, 195.
197. United Kingdom Atomic Energy, US 3 492 382, 1965.
198. P. G. Ruopp, L. D. Loch, L. A. Joo, *13th Biennial Conf. on Carbon*, Extended Abstracts and Program, Irvine, CA, 18 – 22 Jul. 1977, p. 426.
199. J. Okada, Y. Takeuchi, *Proc. Conf. Carbon 4th 1959* 1960, 657.
200. J. Schultz, Z. Kessaissia, *Carbon* **20** (1982) 395. Links
201. D. Sell, *Aufbereit. Tech.* **19** (1978) 17.
202. B. Hohl, *Eurocarbon 1998*, Strasbourg, France, p. 173.
203. *Aluminium (Düsseldorf)* **59** (1983) 258.
204. H. B. Ries, *Aufbereit. Tech.* **23** (1982) 8.
205. Kernforschungsanlage Jülich, US 4 009 143, 1972 (H. Luhleich, H. Nickel, P. Pflaum, F. Dias).
206. Baker Perkins Inc., US 3 570 588, 1968 (D. A. Wheeler).
207. Swiss Aluminium Ltd., GB 2 046 143, 1980 (W. K. Fischer, F. Keller, M. Hanni).
208. W. Giesen, *Int. Z. Elektrowärme* **20** (1962) no. 6, 282.
209. Union Carbide Corp., US 3 196 486, 1962 (E. G. Shesler, A. Bowen, R. F. Burden).
210. Siemens-Planiawerke, DE 632 410, 1934 (R. Rhee).
211. B. F. Jones, M. G. Wright, *Carbon* **8** (1970) 689. Links
212. Great Lakes Carbon Corp., US 3 676 535, 1969 (L. H. Juel).
213. Sigri Elektrographit, DE 2 555 983, 1975 (J. Semmler).
214. Sigri Elektrographit, US 4 369 683, 1981 (H. Bieg, H. Jonas, H. Rosswurm, W. Demmer, L. Döbel).
215. Klöckner-Humboldt-Deutz, DE 1 928 424, 1969 (K. Wilkens).
216. D. Ermel, *Sprechsaal* **115** (1982) no. 10, 955.
217. Nukem Nuklear-Chemie und Metallurgie, GB 1 137 013, 1967.
218. R. Bigot, J. Fourre: "Carbon '72," *Int. Carb. Conf.,* Baden-Baden 25. – 30. Jun. 1972, p. 384.
219. F. Kurka, H. Persicke: "Carbon '72," *Int. Carb. Conf.,* Baden-Baden 25. – 30. Jun. 1972, p. 388.
220. Gebr. Siemens & Co., DE 390 607, 1921.
221. Siemens-Plania, DE 498 443, 1925.
222. S. Wilkening, *13th Biennial Conf. on Carbon. Extended Abstracts and Program*, Irvine, CA, 18. – 22. Jul. 1977, p. 330.
223. W. Hirt, S. Wilkening, *Aluminium (Düsseldorf)* **57** (1981) 663.
224. Du Pont, CA 670 708, 1960 (J. D. Chrisp, G. E. D. Jenkins).
225. VEB Elektrokohle Lichtenberg, DE 2 820 722, 1978 (G. Ostermay *et al.*).
226. Union Carbide Corp., US 2 892 218, 1955 (K. B. McGhee, N. J. Johnson).
227. K. J. Hüttinger, *Chem. Ing. Tech.* **44** (1972) 849.
228. I. Letizia, F. Calderone, *High Temp.-High Pressures* **13** (1981) no. 3, 297.
229. Hoechst, DE 1 102 030, 1959 (O. Peter, H. Reckhard).
230. Union Carbide Corp., US 2 965 931, 1957 (L. B. Alden, F. P. Holloway).
231. Union Carbide Corp., US 3 092 437, 1958 (M. B. Carter, R. C. Stroup, J. J. Vadla, K. McGhee).
232. G. Schiele, *Aluminium (Düsseldorf)* **43** (1967) 171.
233. Sigri Elektrographit, DE 2 643 764, 1976 (M. Ulrich, K. W. F. Etzel, W. Krohe, B. Rösch).

234. M. E. de Fernández, J. Marletto, H. Martirena in E. M. Adkins (ed.): *Light Metals 1983*, AIME, New York 1982, p. 805.
235. Sigri Elektrographit, DE 2 008 206, 1970 (E. Nedopil, W. Krohe).
236. Alcan Res. & Development, EP 0 019 425, 1982 (D. N. Holdner). P. Hyvernat, P. Fayet, J. L. Lemarchand in E. M. Adkins (ed.): *Light Metals 1983*, AIME, New York 1982, p. 821.
237. Great Lakes Carbon Corp., US 3 142 482, 1961 (W. M. Kenan).
238. Sigri Elektrographit, GB 1 312 856, 1971 (W. Krohe, B. Rösch).
239. H. J. Feist, FR 2 465 977, 1980.
240. Klöckner-Humboldt-Deutz, DE 2 731 760, 1977 (F. Isenhardt, E. Schultze-Rhonhof, R. Hesse).
241. Aluminium-Industrie-AG, CH 288 579, 1951.
242. J. Maire, J. Mering in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **6**, Marcel Dekker, New York 1970, p. 125.
243. D. B. Fischbach in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **7**, Marcel Dekker, New York 1971, p. 7.
244. A. Pacault in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **7**, Marcel Dekker, New York 1971, p. 107.
245. E. Wege, *Ber. Dtsch. Keram. Ges.* **43** (1966) 224.
246. R. W. Marek, W. E. Parker, J. H. Cox, *Carbon* **6** (1968) 349. Links
247. S. Sato, K. Kawamata, H. Awaji, M. Osawa, M. Manaka, *Carbon* **19** (1981) 111. Links
248. E. Fitzer, W. Weisweiler, *Chem. Ing. Tech.* **44** (1972) 972.
249. H. Marsh, A. P. Warburton, *J. Appl. Chem.* **20** (1970) 133.
250. H. Böder, E. Fitzer, *Carbon* **8** (1970) 453.
251. Japan Atomic Energy Res. Institute, DE 1 646 611, 1967 (S. Sato, T. Honda, H. Ito).
252. M. Utzat, *Forschungsbericht T83–162-Nichtnukleare Energietechnik*, Bundesministerium für Forschung und Technologie, August 1983.
253. Great Lakes Carbon Corp., WO 83/00549, 1983 (B. Y. Karagoz).
254. Carbone Lorraine, FR 1 271 206, 1961.
255. A. Burmeister: “Carbon '80,” *Int. Carb. Conf.,* Baden-Baden 30. Jun. – 4. Jul. 1980, p. 481.
256. H. G. Hopp, H. Schlesinger, *High Temp.-High Pressures* **5** (1973) no. 4, 339.
257. R. F. Markel, *11th Biennial Conf. on Carbon. Extended Abstracts and Program*, Gatlinburg, 4. – 8. Jun. 1974, p. 136.
258. Siemens-Plania, DE 1 085 863, 1958 (A. Ragoß, E. Nedopil).
259. C. A. Odening, J. C. Bowmann, *Proc. 1st Conf. Ind. Carbon and Graphite*, London 24. – 26. Sept. 1957, Soc. of Chem. Ind. 1958, p. 537.
260. Pechiney, Commissariat à L'Energie Atomique, FR 1 143 902, 1956.
261. Hobeg, DE 2 337 428, 1973 (K.-G. Hackstein, P. Venet, G. Spener).
262. S. M. Lelcuk, *Stanki Instrum.* **39** (1968) 22.
263. F. Smeling, *Stahl Eisen* **102** (1982) 771.
264. Sigri Elektrographit, DE 3 147 793, 1981 (A. Sigmund).
265. J. H. Feist, DE 3 147 040, 1981.
266. L. W. Graham, W. Watt, W. Johnson, P. A. P. Arragon, M. S. T. Price, *Proc. Conf. Carbon 5th 1961*, 1963, 387.
267. F. Brandmair, *Werkst. Korros.* **17** (1966) 10.
268. Sigri Elektrographit, DE 1 927 828, 1969 (A. Pelz, G. Pietzka, W. Ulsamer).
269. W. V. Kotlensky in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **9**, Marcel Dekker, New York 1973, p. 173.
270. O. Vohler, P. L. Reiser, E. Sperk, *Carbon* **6** (1968) 397. Links
271. A. Baranski *et al.*, *Carbon* **20** (1982) 401. Links
272. O. Rubisch, O. Vohler, *Tech. Rundsch.* **41** (1964) 33.
273. *The Measurement of Employee Exposures in Aluminum Reduction Plants*, International Primary Aluminum Institute, London 1982.
274. *VDI-Richtlinie 3467*, “Herstellung von Werkstoffen aus Kohlenstoff und Elektrographit”, Düsseldorf 1998.
275. F. Schieber, W. Lauterbach, T. Taube, *Sprechsaal* **114** (1981) 271.
276. P. Lefrank, G. Schuster, A. Treugut, *Fachber. Hüttenpraxis Metallweiterverarb.* **20** (1982) 164.
277. S. Mrozowski, *Proc. Conf. Carbon 1st, 2nd 1953 and 1955* 1956, 31.
278. R. A. Meyer, J. E. Zimmer, *Aerosp*, Rep. No. ATR-74 (7425)-3 (1974) 1.
279. J. E. Brocklehurst in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **13**, Marcel Dekker, New York 1977, p. 145.
280. G. M. Jenkins in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **11**, Marcel Dekker, New York 1974, p. 189.
281. I. L. Spain in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **8**, Marcel Dekker, New York 1973, p. 1.
282. B. R. T. Kelly in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **5**, Marcel Dekker, New York 1969, p. 119.
283. P. Oliver, *Chem. Eng. (N.Y.)* **59** (1952) no. 9, 276.
284. Union Carbide, US 2 735 705, 1954 (H. V. Johnson, N. J. Johnson).
285. Sigri Elektrographit, DE 1 565 076, 1965 (F. Schipke).

286. T.-L. Weng, E. J. Seldin, *Carbon* **15** (1977) 391. Links
287. B. Bowman, *JFE J. Four Electr. Ind. Electrochim.* **9** (1982) 23.
288. Metalurgitchen Zavod “Lenin,” DE 1 271 007, 1962 (A. J. Waltschev, T. N. Waltschewa).
289. D. Ameling, H. Jaunich, J. Schiffarth, F. J. Strunck, *Stahl Eisen* **103** (1983) 1271.
290. J. Isenberg-O'Loughlin, *33 Metal Producing* **20** (1982) no. 11, 45.
291. K. D. Unger, H. Bauer, J. Mühlenbeck, D. Zöllner, *Stahl Eisen* **104** (1984) 11.
292. K. G. Middleton, P. M. Bralden, *Proc. Br. Ceram. Soc.* **32** (1982) 355.
293. W. Frohs, *Metallurg. Plant Technol.* 1999, April 2.
294. Sigri Elektrographit, DE 2 555 688, 1975 (J. Semmler).
295. Union Carbide Corp., US 4 290 709, 1979 (E. J. Seldin, T.-L. Weng).
296. Union Carbide Corp., US 2 603 669, 1948 (C. H. Chappell).
297. US 4 575 856, 1986 (J. A. Persson).
298. K. Eisenächer, *Chem. Ind. (Düsseldorf)* **35** (1983) 435.
299. *Metal Bull. Monthly* **155** (1983) 29.
300. F. Keller, W. K. Fischer in J. E. Andersen (ed.): *Light Metals 1982*, AIME, New York 1981, p. 729.
301. Schweizerische Aluminium AG, Sigri Elektrographit, EP 0 027 534, 1980 (W. Eckel, E. Schulz).
302. M. Sørle, H. A. Øye, *Metall (Berlin)* **36** (1982) 635.
303. J. J. Vadla, H. C. Quandt, D. J. Rich, *Metall (Berlin)* **34** (1980) 1102.
304. K. W. F. Etzel, F. Brandmair, P. Aeschbach, H. Friedli in E. M. Adkins (ed.): *Light Metals 1983*, AIME, New York 1982, p. 885.
305. A. S. Russell, *J. Met.* **33** (1981) no. 6, 132.
306. F. Beck: *Elektroorganische Chemie*, Verlag Chemie, Weinheim, Germany 1974.
307. R. H. Herron, C. R. Beechan, *Am. Ceram. Eng. Sci. Proc.* **3** (1981) 1126.
308. B. Goetze, U. Rosenblatt, *13th Biennial Conf. on Carbon. Extended Abstracts*, Irvine, CA, 18. – 22. Jul. 1977, p. 334.
309. H. H. Hursen, *Met. Prog.* **83** (1963) 67.
310. L. Bunch, *Mater. Eng.* **97** (1983) no. 1, 53.
311. G. E. Ross, D. W. Brown, *Proc. 1st Conf. Ind. Carbon and Graphite*, London 24. – 26. Sept. 1957, Soc. of Chem. Ind. 1958, p. 618.
312. W. Krüger, *Elektrowärme Int. Ed. B* **33** (1975) 184.
313. J. R. Schley, *Chem. Eng. (N.Y.)* **81** (1974) no. 4, 144; **81** (1974) no. 6, 102.
314. H. Würmseher, A. Swozil, *Z. Werkstofftech.* **9** (1978) 19.
315. Sigri Elektrographit, DE 3 116 309, 1981 (A. Swozil).
316. Sigri Elektrographit, DE 3 117 187, 1982 (A. Swozil, W. Krätschmer, G. Ullmann).
317. W. Delle, K. Koizlik, H. Nickel: *Graphitische Werkstoffe für den Einsatz in Kernreaktoren,* part 2, Verlag K. Thiemig, München 1983.
318. H. Wiemer, H. Stein, *Ringsdorff-Werke Tech. Mitt.* **10** (1980) 1.
319. P. E. Lyddon, R. K. Hurden, *Proc. 1st Conf. Ind. Carbon and Graphite*, London 24. – 26. Sept. 1957, Soc. of Chem. Ind. 1958, p. 579.
320. R. R. Paxton, *Mater. Eng.* **97** (1983) no. 3, 35.
321. W. Finkelnburg: *Hochstrom-Kohlebogen*, Springer Verlag, Berlin 1948.
322. Union Carbide, US 4 315 126, 1980 (P. R. Schaefer, D. W. Carkhuff).
323. N. V. Philips' Gloeilampenfabrieken, GB 2 072 640, 1981 (B. Lersmacher, W. F. Knippenberg).
324. L. Weiss, *Werkstatt Betr.* **111** (1978) no. 1, 7.
325. Brown, Boveri & Cie, CH 371 524, 1959 (T. Wasserrab).
326. Varian Associates, US 4 277 718, 1979 (M. B. Schrader).
327. F. v. Sturm: *Elektrochemische Stromerzeugung*, Verlag Chemie, Weinheim, Germany 1969.
328. R. Holm: *Electric Contacts Handbook*, Springer Verlag, Berlin 1967.
329. E. Fitzer, *Angew. Chem.* **92** (1980) 375.
330. The General Electric Comp., GB 889 351, 1958 (H. W. Davidson).
331. B. Lersmacher, H. Lydtin, W. F. Knippenberg, *Chem. Ing. Tech.* **42** (1970) 659.
332. E. Fitzer, W. Schaefer, *Carbon* **8** (1970) 353. Links
333. P. L. Walker, *Carbon* **10** (1972) 369. Links
334. *Chem. Labor Betr.* **28** (1977) 422.
335. Carbone-Lorraine, DE 2 034 960, 1970 (A. Mouchot).
336. Eastman Kodak Comp., US 3 844 755, 1972 (M. A. Angle, G. L. Bender, G. E. Blair, C. C. Maier).
337. M. Mitzlaff, J. Cramer, R. Pistorius, *Chem. Ing. Tech.* **54** (1982) 601.
338. Sigri Elektrographit, EP 0 004 299, 1979 (G. Gistinger, D. Wolter).
339. Sigri Elektrographit, Siemens, DE 2 613 072, 1976 (G. Richter, E. Weidlich, K. Mund, H. Böder).
340. C. R. Schmidt, J. M. Schreyer, A. L. Coffey, *Carbon* **7** (1969) 443.
341. Kureha K.K., DE 2 451 019, 1974 (H. Nagai, K. Katori).

342. A. W. Moore in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **11**, Marcel Dekker, New York 1973, p. 69.
343. W. Delle, K. Koizlik, H. Nickel: *Graphitische Werkstoffe für den Einsatz in Kernreaktoren,* part I, Verlag K. Thiemig, München 1979.
344. J. C. Bokros, L. D. La Grange, F. J. Schoen in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **9**, Marcel Dekker, New York 1973, p. 103.
345. H. P. Boehm, M. Voll, *Carbon* **8** (1970) 227, 741; **9** (1971) 473, 481.
346. J. B. Donnet, *Carbon* **20** (1982) 267. Links
347. B. C. Brodie, *Philos. Trans. R. Soc. London* **149** (1859) 249.
348. D. Berger, B. Carton, A. Métrot, A. Hérold in P. L. Walker, P. A. Thrower (eds.): *Chemistry and Physics of Carbon*, vol. **12**, Marcel Dekker, New York 1975, p. 1.
349. W. Rüdorff, E. Stumpp, W. Spriessler, F. W. Siecke, *Angew. Chem.* **75** (1963) 130.
350. A. Hérold in F. Lévy (ed.): *Intercalated Layered Materials*, D. Reidel Publ. Comp., Dordrecht 1979, p. 323.
351. J. G. Hooley in P. L. Walker (ed.): *Chemistry and Physics of Carbon*, vol. **5**, Marcel Dekker, New York 1969, p. 321.
352. M. S. Whittingham, L. B. Ebert in F. Lévy (ed.): *Intercalated Layered Materials*, D. Reidel Publ. Comp., Dordrecht 1979, p. 533.
353. T. C. Wu, F. L. Vogel, *Carbon* **20** (1982) 225. Links
354. L. Streifinger, H. P. Boehm, R. Schlögl, P. Pentenrieder, *Carbon* **17** (1979) 195. Links
355. Millipore Corp., US 4 342 679, 1980 (L. Abrahams, T. P. J. Izod).
356. W. Luzi, *Z. Naturwissenschaften* **64** (1892) 224.
357. W. Luzi, GB 9922, 1891.
358. Condensite Comp. of America, US 1 137 373, 1913 (J. W. Aylsworth).
359. Union Carbide, US 3 404 061, 1963 (J. H. Shane, R. J. Russel, R. A. Bochman).
360. M. B. Dowell, *12th Biennial Conf. on Carbon. Extended Abstracts and Programs*, Pittsburgh 28. Jul. –1. Aug. 1975, pp. 31, 35.
361. General Motors, DE 2 404 544, 1974 (W. G. Meyers).
362. Union Carbide, US 4 068 853, 1976 (D. L. Schnitzler).
363. Sigri Electrographit, GB 1 338 747, 1971 (H. Linder, E. Schwarz).
364. Degussa, US 4 180 697, 1977 (E. Schumann, O. Bohländer).
365. European Council of Chemical Manufacturers' Federations (CEFIC): *Test methods for activated carbon*, CEFIC Brussels 1986, p. 7.
366. H. Kienle von, E. Bäder: *Aktivkohle und ihre industrielle Anwendung*, Enke-Verlag, Stuttgart 1980.
367. A. Yehaskel: *Activated Carbon*, Manufacture and Regeneration, Noyes DATA Corp., Park Ridge, NJ 1978.
368. M. Smisek, S. Cerny: *Activated Carbon*, Elsevier Publ., Amsterdam – London – New York 1970.
369. P. N. Cheremisinoff, F. Ellerbusch (eds.): *Carbon Adsorption Handbook*, Ann Arbor Science. Publishers, MI 1978.
370. R. Ch. Bansal, J.-B. Donnet, F. Stoeckli: *Activated Carbon*, Marcel Dekker Inc., New York 1988.
371. DE 136 792, 1901 (R. v. Ostrejko).
372. Österr. Verband für chem. und metall. Produkte, AT 68 169, 1913.
373. VDI 3674, Waste gas cleaning by adsorption. Process and waste gas cleaning, 1998.
374. V. A. Garten, D. E. Weiss, J. B. Willis, *Aust. J. Chem.* **10** (1957) 295. Links
375. H. P. Boehm, E. Diehl, W. Heck, R. Sappock, *Angew. Chem.* **76** (1964) 742.
376. H. Jüntgen: “Grundlagen der Adsorption,” *Staub Reinhalt. Luft* **36** (1976) no. 7, 281, 324.
377. H. Seewald: “Technisch verfügbare Adsorbentien,” *Vortragsveröffentlichung Haus der Technik*, vol. **404**, Vulkan-Verlag, Essen 1977, pp. 24, 34.
378. W. Kast: *Adsorption aus der Gasphase*, VCH Verlagsgesellschaft, Weinheim, Germany 1988.
379. S. Brunauer, P. H. Emmet, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309. Links
380. International Union of Pure and Applied Chemistry (IUPAC): *Manual of Definition*, Terminology and Symbols in Colloid and Surface Chemistry, IUPAC 1972.
381. R. W. Cranston, F. A. Inkley, *Adv. Catal.* **9** (1957) 14.
382. H. Jüntgen, M. J. Schwuger, *Chem. Ing. Tech.* **38** (1966) 1271.
383. ASTM D 3467–99, Standard Test Method for Carbon Tetrachloride Activity of Activated Carbon, 1999.
384. J. Meunier in: *Vergasung fester Brennstoffe und oxydative Umwandlung von Kohlenstoffen*, Verlag Chemie, Weinheim, Germany 1962, p. 41.
385. O. Großkinsky, *Glückauf* **86** (1950) 988.
386. Calgon Corp., BE 839918, 1976.
387. The Carborundum Co., DE 2 624 530, 1976; DE 2 625 625, 1976.
388. H. J. Fornwalt, W. A. Helbig, G. H. Scheffler, *Br. Chem. Eng.* **8** (1963) 546 – 550.
389. Kureha Kakagu, Kogyo, DE 2 146 100, 1971.
390. West Virginia Pulp & Paper Comp., DE 1 467 195, 1964.
391. US 2 966 447, 1957 (H. J. Walter).
392. Takeda Chemical Ind., DE 2 400 991, 1982.

393. Standard Oil Comp., US 3 833 514, 1971.
394. M. M. Dubinin, W. A. Gordejwa, L. T. Jefimova, B. A. Onusaitis, Z. A. Rjabikowa, *Zh. Pril. Kmm. (Leningrad)* **43** (1970) no. 10, 2219 – 2225.
395. C. Hinskelwood, *BCURA Gaz.* **27** (1956) 3 – 10.
396. H. Jüntgen, K. Knoblauch, K. Harder, *Fuel* **60** (1981) Sept., 817 – 822.
397. K.-D. Henning, S. Schäfer, *Gas Sep. Purification* **7** (1993) no. 4, 235 – 240.
398. H. Menig: *Luftreinhaltung durch Adsorption*, Absorption und Oxidation, Deutscher Fachschriften Verlag, Wiesbaden 1977.
399. J. W. Hassler: *Purification with Activated Carbon*, Chemical Publishing Co. Inc., New York NY 1974.
400. H. Krill, H. Menig: "Lösungskonzepte für Lösungsmittel-Vermeidung und Verwertung," *Entsorgungspraxis Spezial* (1989) no. 6, 20, 26.
401. K. Storp: "Abluft- und Abgasentschwefelung durch Adsorption und Katalyse an Aktivkohlen," *Dechema Monogr.* **64** (1970) 91, 94.
402. H. W. Bräuer: "Abluftreinigung beim Flugzeugbau, Lösemittelrückgewinnung in der Wirbelschicht," *UTA Umwelt Technologie Aktuell* (1993) no. 2, 115, 120.
403. K. Storp, *Chemiefasern* **18** (1968) 173.
404. J. Klein, K.-D. Henning: *Fuel Magazine* **63** (1984) 1064 – 1067.
405. E. Richter, E. Pilarczyk, K.-B. Harder, K. Knoblauch, *Erdöl Erdgas Kohle* **102** (1986) no. 11, 512 – 516.
406. W. N. Tuttle, *Port Technology International* **1** (1995) no. 1, 143, 146.
407. K. Knoblauch, *Erzmetall*, **33** (1980) no. 2, 109, 114.
408. J. Gottschalk: "Flugstromverfahren – ein Verfahren der Kohlenstoffadsorptionstechnik zur Minimierung der Restemissionen," *Abfallwirtschaftsjournal* **4** (1992) no. 12, 997, 1001.
409. K.-D. Henning, K. Keldenich, K. Knoblauch, J. Degel, *Gas Sep. Purif.* (1988) March, 20, 22.
410. H. Jüntgen, J. Klein, *Energy Sources* **2** (1976) no. 4, 311 – 330.
411. K.-D. Henning, J. Degel, J. Klein, K. Knoblauch, *gwf-wasser/abwasser* **127** (1986) no. 6, 272 – 282.
412. R. G. Rice, C. M. Robson: *Biological Activated Carbon*, Ann Arbor Science Publishers, Ann Arbor, MI 1982.
413. A. I. Medalia, D. Rivin, *Carbon* **20** (1982) no. 6, 481 – 492. Links
414. H. P. Boehm, *Farbe + Lack* **79** (1973) no. 5, 419 – 432.
415. R. D. Heidenreich, W. M. Hess, L. L. Ban, *J. Appl. Crystallogr.* **1** (1968) 1; K. A. Burgess, C. E. Scott, W. M. Hess, *Rubber Chem. Technol.* **44** (1971) no. 1, 230 – 248. Links
416. W. C. Wake (ed.): *Fillers for Plastics*, Iliffe Books, London 1971.
417. Inter-Governmental Maritime Consultive Organization, London, *International Maritime Dangerous Goods Code*, 11 – 75, 4081 – 4082.
418. *Rep. Invest. U.S. Bur. Mines* **6598** (1965) .
419. *U.S. Bur. Mines Technical Paper* **610** (1940).
420. H. W. Davidson *et al.*: *Manufactured Carbon*, Pergamon Press, Oxford-New York 1968.
421. Degussa, DE-OS 2 410 565, 1974.
422. Cabot Corp., DE-OS 2 507 021, 1975.
423. Cabot Corp., US 3 010 794, 3 010 795, 1958.
424. Cities Service Co., US 3 490 869, 1966.
425. Cabot Corp., DE-OS 2 507 021, 1975.
426. Cities Service Co., DE-OS 1 592 853, 1967.
427. Ashland Oil & Refining Co., US 3 649 207, 1969; Phillips Petroleum Co., US 3 986 836, 1974.
428. G. Kuehner, G. Dittrich, *Chem. Ing. Tech.* **44** (1972) 11.
429. *Ullmann*, 3rd ed., vol. **14**, pp. 793 – 810.
430. Degussa, DE 895 286, 1951; DE 1 129 459, 1960.
431. Cities Service Co., US 3 593 371, 1968; Phillips Petroleum Co., US 3 956 445, 1974; Cabot Corp., DE-OS 2 004 493, 1970; Cabot Corp., GB 1 370 704, 1972.
432. Shawinigan Chem. Ltd., US 2 453 440, 1944; US 2 492 481, 1948; J. Wotschke, K. Paasch, *Schweiz. Arch.* **5** (1949) 173.
433. Cabot Corp., US 3 342 554, 1963.
434. M. S. Jyengar, *Chem. Technol.* **2** (1972) no. 9, 565 – 569.
435. Degussa, DE 742 664, 1940; Cabot Corp., US 2 420 810, 1941.
436. Degussa, GB 895 990, 1958.
437. Cabot Corp., US 2 439 442, 1943.
438. T. C. Patton (ed.): *Pigment Handbook,* vol. 3, J. Wiley & Sons, New York 1973.
439. Cabot Corp., US 3 383 232, 1968; Phillips Petroleum Co., US 3 870 785, 1975.
440. W. M. Hess, L. L. Ban, G. C. McDonald, *Rubber Chem. Technol.* **42** (1969) 1200; J. C. Motte, *Rev. Gen. Caoutch. Plast.* **49** (1971) no. 5, 417 – 421.
441. S. Brunauer, P. H. Emmett, E. J. Teller, *J. Am. Chem. Soc.* **60** (1938) 309. Links
442. J. H. de Boer *et al.*, *J. Catal.* **4** (1965) 319 – 323, 643 – 648. Links
443. W. Schumacher, *FATIPEC Congr. 13th* 1976, 582 – 587.

444. R. Bode, *Kautsch. Gummi Kunstst.* **36** (1983) no. 8, 660 – 676.
445. D. Rivin, R. G. Smith, *Rubber Chem. Technol.* **55** (1982) no. 3, 707 – 761. Links
446. J. M. Robertson, R. G. Smith, *Patty's Industrial Hygiene and Toxicology*, 4th ed., vol. **2D**, "chap. 28" Wiley Interscience, New York 1994, p.2395.
447. ECBECB (European Committee for Biological Effects of Carbon Black), Bulletin No. 7 (1992), *Polyaromatic Hydrocarbons in Carbon Black.*
448. J. M. Robertson, T. H. Ingalls, *Am. Ind. Hyg. Assoc. J.* **50** (1989) 510. Links
449. W. A. Crosbie, *Arch. Environ. Health* **35** (1986) 346.
450. J. M. Robertson, T. H. Ingalls, *Arch. Environ. Health* **35** (1980) 181. Links
451. C. A. Nau, J. Neal, V. A. Stembridge, *Am. Med. Assoc. Arch. Ind. Health* **18** (1958) 511.
452. C. J. Kirwin *et al.*, *J. Toxicol. Environ. Health* **7** (1981) 973. Links
453. F. Buddingh, M. J. Bailey, B. Wells, J. Maesemmeyer, *Am. Ind. Hyg. Assoc. J.* **42** Links
454. J. L. Mauderly et al., in: *Health Effect Research Report No. 68,* Health Effect Institute, Cambridge, Massachusetts (1994).
455. U. Heinrich *et al.*, *Inhal. Toxicol.* **7** (1995) 533. Links
456. G. Oberdorster, in: *Toxic and Carcinogenic Effects of Solid Particles in the Respiratory Tract*, ILSI Press, Washington, DC 1994, p. 335.
457. American Conference of Governmental Industrial Hygienists, *Documentation of the Threshold Limit Values for Substances in Workroom Air — Carbon Black*, rev. ed., ACGIH, Cincinnati, OH 1991, p.220.
458. K. Gardiner, W. N. Trethowan, J. M. Harrington, *et. al.*, *Ann. Occup. Hyg.* **5** (1992) 477.
459. IARC (Lyon): “Evaluation of Carcinogenic Risk of Chemicals to Humans”, Vol. 33, Carbon Black (1984); confirmed in March,1987.
460. IARC (Lyon): “Evaluation of Carcinogenic Risk of Chemicals to Humans”, Vol. 65, Carbon Black, 1996.
461. O. Hahn, F. Strassmann, J. Mattauch, H. Ewald, *Naturwissenschaften* **30** (1942) 541. Links
462. J. Mattauch, H. Ewald, O. Hahn, F. Strassmann, *Z. Phys.* **120** (1943) 598.
463. E. Osawa, *Kagaku* **25** (1970) 854.
464. Z. Yoshida, E. Osawa (eds.): *Aromaticity*, Kyoto 1971 p. 174.
465. E. A. Rohlfing, C. M. Cox, A. Kaldor, *J. Chem. Phys.* **81** (1984) 3322. Links
466. H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, R. E. Smalley, *Nature* **318** (1985) 162. Links
467. T. G. Dietz, M. A. Duncan, D. E. Powers, R. E. Smalley, *J. Chem. Phys.* **74** (1981) 6511. Links
468. W. Krätschmer, L. D. Lamb, K. Fostiropoulos, D. R. Huffman, *Nature* **347** (1990) 354.
469. D. R. Huffman, *Adv. Phys.* **26** (1977) 129. Links
470. W. Krätschmer, K. Fostiropoulos, D. R. Huffman, *Chem. Phys. Lett.* **170** (1990) 167.
471. S. J. Cyvin, E. Brendsdal, B. N. Cyvin, J. Brunvoll, *Chem. Phys. Lett.* **143** (1988) 377. Links
472. H. Ajie *et al.*, *Phys. Chem.* **94** (1990) 8630.
473. R. L. Whetten *et al.*, *Mater. Res. Soc., Symp. Proc.* **206** (1990) 639.
474. A. S. Koch, K. C. Khemani, F. Wudl, *J. Org. Chem.* **56** (1991) 4543. Links
475. L. P. F. Chibante, A. Thess, J. M. Alford, M. D. Diener, R. E. Smalley, *J. Phys. Chem.* **97** (1993) 8696.
476. R. E. Haufler *et al.*, *J. Phys. Chem.* **94** (1990) 8634.
477. W. A. Scrivens, J. M. Tour, *J. Org. Chem.* **57** (1992) 6932; Links K. Kikuchi *et al.*, *Chem. Lett.* **1991**, 1607
478. R. E. Haufler *et al.*, *Chem. Phys. Lett.* **179** (1991) 449. Links
479. G. Peters, M. Jansen, *Angew. Chem.* **104** (1992) 240; *Angew. Chem. Int. Ed. Engl.* **(31)** 1992 223.
480. P. Gerhardt, S. Löffler, K. H. Homann, *J. Phys. Chem.* **137** (1987) 316; P. Gerhardt, K. H. Homann, *J. Phys. Chem.* **94** (1990) 5381.
481. J. B. Howard, J. T. McKinnon, Y. Makarovsky, A. Lafleur, M. E. Johnson, *Nature* **352** (1991) 139; Links J. B. Howard, J. T. McKinnon, M. E. Johnson, Y. Makarovsky, A. Lafleur, *J. Phys. Chem.* **96** (1992) 6657; J. B. Howard *et al.*, *Carbon* **30** (1992) 1183; Links C. J. Pope, J. A. Marr, J. B. Howard, *J. Phys. Chem.* **97** (1993) 11001.
482. R. Taylor, G. L. Langley, H. W. Kroto, D. R. M. Walton, *Nature* **366** (1993) 728. Links
483. T. G. Schmalz, W. A. Seitz, D. J. Klein, G. E. Hite, *Chem. Phys. Lett.* **130** (1986) 203; Links H. W. Kroto, *Nature*, **329** (1987) 529. Links
484. D. E. Manolopoulos, J. C. May, *Chem. Phys. Lett.* **181** (1991) 105. Links
485. P. W. Fowler, D. E. Manolopoulos, R. C. Batten, *J. Chem. Soc., Faraday Trans.* **87** (1991) 32.
486. R. Taylor, *Tetrahedron Lett.* (1991) 3734.
487. H. W. Kroto, A. W. Allaf, S. P. Balm, *Chem. Rev.* **91** (1991) 1213. Links
488. H. D. Beckhaus *et al.*, *Angew. Chem.* **103** (1992) 69; *Angew. Chem. Int. Ed. Engl.* **31** (1992) 63.
489. H. D. Beckhaus *et al.*, *Angew. Chem.* **106** (1994) 1033; *Angew. Chem. Int. Ed. Engl.* **33** (1994) 996.
490. R. F. Curl, *Philos. Trans. R. Soc.* **A, 343** (1993) 19.
491. R. S. Ruoff, D. S. Tse, R. Malhorta, D. C. Lorents, *J. Phys. Chem.* **97** (1993) 3379.
492. N. Sivaraman *et al.*, *J. Org. Chem.* **57** (1992) 6077; Links W. A. Scrivens, J. M. Tour, *J. Chem. Soc. Chem. Comun.* (1993) 1207.
493. M. F. Meidine, P. B. Hitchcock, H. W. Kroto, R. Taylor, D. R. M. Walton, *J. Chem. Soc., Chem. Commun.* (1992) 1534; A. L. Balch, J. W. Noll, M. M. Olmstead, *J. Chem. Soc., Chem. Commun.* (1993) 56.

494. J. D. Crane *et al.*, *J. Chem. Soc., Chem. Commun.* (1992) 1764.
495. A. Izuoka *et al.*, *J. Chem. Soc., Chem. Commun.* (1992) 1472.
496. J. D. Crane, P. B. Hitchcock, *J. Chem. Soc., Dalton Trans.* (1993) 2537. Links
497. P. R. Birkett, C. Christides, *J. Chem. Soc., Perkin Trans.* **2** (1993) 1407.
498. O. Ermer, *Helv. Chim. Acta* **74** (1991) 1339. Links
499. G. Roth, P. Adelmann, *J. Phys. I France* **2** (1992) 1541; H. B. Bürgi, P. Venugopalan, D. Schwarzenbach, F. Diederich, C. Thilgen, *Helv. Chim. Acta* **76** (1993) 2155.
500. J. L. Atwood, G. A. Koutsantonis, C. L. Raston, *Nature* **368** (1994) 229. Links
501. T. Anderson, K. Nilson, M. Sundahl, G. Westman, O. Wennerström, *J. Chem. Soc., Chem. Commun.* (1992) 604.
502. R. M. Flemming et al., in G. S. Hammond, V. J. Kuck (eds.): *Fullerenes: Synthesis, Properties, and Chemistry of Large Carbon Clusters* American Chemical Society Symposium Series 481, Washington D.C. 1992 p. 25.
503. P. A. Heiney *et al.*, *Phys. Rev. Lett.* **66** (1991) 2911; Links A. Dworkin *et al.*, *C.R. Acad. Sci. Paris* **312 II** (1991) 979; J. S. Tse, D. D. Klug, D. A. Wilson, Y. P. Handa, *Chem. Phys. Lett.* **183** (1991) 387. Links
504. P. A. Heiney, *J. Phys. Chem. Solids* **53** (1992) 1333. Links
505. R. C. Haddon, *Acc. Chem. Res.* **25** (1992) 127 and references therein.
506. A. R. Kortan *et al.*, *Nature* **355** (1992) 529; Links A. R. Kortan *et al.*, *Nature*, **360** (1992) 566. Links
507. K. Holczer *et al.*, *Science* **252** (1991) 1154; Links P. W. Stephens *et al.*, *Nature* **351** (1991) 351. Links
508. O. Zhou *et al.*, *Nature* **351** (1991) 462. Links
509. S. P. Kelty, C.-C. Chen, C. M. Lieber *Nature* **351** (1991) 223.
510. T. Tanigaki, T. W. Ebbesen, *Nature* **352** (1991) 222; Links T. Tanigaki *et al.*, *Nature* **356** (1992) 419. Links
511. Y. Murakami *et al.*, *Fullerene Sci. and Tech.* **1** (1993) 351; Q. Zhu *et al.*, *Phys. Rev. B* **47** (1995) 13948; D. M. Poirier, D. W. Owens, J. H. Weaver, *Phys. Rev. B* **51** (1995) 1830. Links
512. M. J. Rosseinsky *et al.*, *Nature* **356** (1992) 416. Links
513. A. F. Hebard *et al.*, *Nature* **350** (1991) 660. Links
514. T. T. M. Palstra, *Solid State Commun.* **93** (1995) 327. Links
515. K. Tanigaki, S. Kuroshima, J. Fujita, T. W. Ebbesen, *Appl. Phys. Lett.* **63** (1993) 2351; Links R. M. Fleming *et al.*, *Nature* **352** (1991) 787. Links
516. P.-M. Allemand *et al.*, *Science* **253** (1991) 301. Links
517. K. Tanaka *et al.*, *Phys. Lett. A* **164** (1992) 221; Links K. Tanaka *et al.*, *Phys. Rev. B* **47** (1993) 7554; Links R. Seshadri, A. Rastogi, S. V. Bhat, S. Ramasesha, C. N. R. Rao, *Solid. State Commun.* **85** (1993) 971; Links P. W. Stephens, *Nature* **356** (1992) 383; H. Klos *et al.*, *Chem. Phys. Lett.* **224** (1994) 333. Links
518. A. Hirsch, *Synthesis* (1995) 895. Links
519. A. Hirsch, *J. Phys. Chem. Solids* **58** (1997) 1729. Links
520. F. Diederich, C. Thilgen, *Science* (1996) 271. Links
521. F. Wudl, *Acc. Chem. Res.* **25** (1992) 157. Links
522. F. Diederich, L. Isaacs, D. Philip, *Chem. Soc. Rev.* (1994) 243. Links
523. F. Wudl *et al.* in G. S. Hammond, V. S. Kuck (eds.): *Fullerenes: Synthesis*, Properties, and Chemistry of Large Carbon Clusters; American Chemical Society Symposium Series 481, Washington, DC 1992 p. 161.
524. P. J. Krusic *et al.*, *Science* **254** (1991) 1183. Links
525. A. Hirsch, A. Soi, H. R. Karfunkel, *Angew. Chem.* **104** (1992) 808; *Int. Ed. Engl.* **31** (1992) 766.
526. A. Hirsch, T. Grösser, A. Skiebe, A. Soi, *Chem. Ber.* **126** (1993) 1061; P. J. Fagan, P. J. Krusic, D. H. Evans, S. A. Lerke, E. Johnston, *J. Am. Chem. Soc.* **114** (1992) 9697. Links
527. C. Bingel, *Chem. Ber.* **126** (1993) 1957.
528. H. J. Bestmann, D. Hadawi, T. Röder, C. Moll, *Tetrahedron Lett.* (1994) 9017. Links
529. A. Hirsch, Q. Li, F. Wudl, *Angew. Chem.* **103** (1991) 1339; *Int. Ed. Engl.* **30** (1991) 1309; R. Shedari, A. Govindaraj, R. Nagarajan, T. Pradeep, C. N. R. Rao, *Tetrahedron Lett.* **23** (1992) 2069; A. Skiebe, A. Hirsch, H. Klos, B. Gotschy, *Chem. Phys. Lett.* **220** (1994) 138. Links
530. K.-D. Kampe, N. Egger, M. Vogel, *Angew. Chem.* **105** (1993) 1203; *Angew. Int. Ed. Engl.* **32** (1993) 1174; *Liebigs Ann.* (1995) 115.
531. M. Tsuda, T. Ishida, T. Nogami, S. Kurono, M. Johashi, *J. Chem. Soc., Chem. Commun.* (1993) 1296; M. Prato *et al.*, *J. Am. Chem. Soc.* **115** (1993) 1594; Links J. Averdung, J. Mattay, *Tedrahedron Lett.* , in press; B. Kräutler, M. Puchberger, *Helv. Chim. Acta* **76** (1993) 1626; Y. Rubin, S. Kahn, D. I. Freedberg, C. Yeretzian, *J. Am. Chem. Soc.* **115** (1993) 344; Links Y.-Z. an, J. L. Anderson, Y. Rubin, *J. Org. Chem.* **58** (1993) 4799; Links S. Kahn, A. M. Oliver, M. N. Paddon-Row, Y. Rubin, *J. Am. Chem. Soc.* **115** (1993) 4919; P. Belik, A. Gügel, J. Spikermann, K. Müllen, *Angew. Chem.* **105** (1993) 95, *Angew. Chem. Int. Ed. Engl.* **32** (1993) 78; F. Diederich *et al.*, *Helv. Chim. Acta* **76** (1993) 2445; Links A. Gügel, A. Kraus, P. Belik, K. Müllen, *Angew. Chem.* **106** (1994) 601, *Angew. Chem. Int. Ed. Engl.* **33** (1994) 559; M. Iyoda, F. Sultana, S. Sasaki, M. Yoshida, *J. Chem. Soc., Chrm. Commun.* (1994) 1929.
532. M. Prato, Q. Li, F. Wudl, V. Lucchini, *J. Am. Chem. Soc.* **115** (1993) 1148. Links
533. S. H. Hoke, II, *et al.*, *J. Org. Chem.* **57** (1992) 5069; Links X. Zhang, A. Romero, C. S. Foote, *J. Am. Chem. Soc.* **115** (1993) 11024. Links
534. K. E. Geckeler, A. Hirsch, *J. Am. Chem. Soc.* **115** (1993) 3850; Links A. O. Patil, G. W. Schriver, B. Carstensen, R. D. Lundberg, *Polym. Bull.* (1993) 645.
535. K. Chen, W. B. Caldwell, C. A. Mirkin, *J. Amer. Chem. Soc.* **115** (1993) 1193.
536. W. B. Caldwell, K. Chen, C. A. Mirkin, S. J. Babinec, *Langmuir* (1993) 1960.

537. P. J. Krusic *et al.*, *J. Am. Chem. Soc.* **113** (1991) 6274. Links
538. C. C. Henderson, P. A. Cahill, *Science* **259** (1993) 1885. Links
539. S. Ballenweg, R. Gleiter, W. Krätschmer, *Tetrahedron Lett.* **34** (1993) 3737. Links
540. T. F. Guarr, M. S. Meier, V. K. Vance, M. Clayton, *J. Amer. Chem. Soc.* **115** (1993) 9862.
541. M. R. Banks *et al.*, *J. Chem. Soc., Chem. Commun.* (1993) 1149; C. Rüchardt *et al.*, *Angew. Chem.* **105** (1993) 609; *Angew. Chem. Int. Ed. Engl.* **32** (1993) 584; M. I. Attalla, A. M. Vassallo, B. N. Tattam, J. V. Hanna, *J. Phys. Chem.* **97** (1993) 6329; K. Shigematsu, K. Abe, M. Mitani, K. Tanaka, *Chem. Express* **8** (1993) 37.
542. J. H. Holloway *et al.*, *J. Chem. Soc.; Chem. Commun.* (1991) 966.
543. H. Selig *et al.*, *J. Am. Chem. Soc.* **113** (1991) 5475; Links K. kniaz *et al.*, *J. Am. Chem. Soc.* **115** (1993) 6060. Links
544. A. A. Gakh *et al.*, *J. Am. Chem. Soc.* **116** (1994) 819; Links A. A. Tuinman, A. A. Gakh, J. L. Adcock, R. N. Compton, *J. Am. Soc.* **115** (1993) 5885.
545. G. A. Olah *et al.*, *J. Am. Chem. Soc.* **113** (1991) 9385. Links
546. F. N. Tebbe *et al.*, *J. Am. Chem. Soc.* **113** (1991) 9900. Links
547. F. N. Tebbe *et al.*, *Science* **256** (1992) 822; Links P. R. Birkett, P. B. Hitchcock, H. W. Kroto, R. Taylor, D. R. M. Walton, *Nature* **357** (1992) 479. Links
548. T. Zenner, H. Zabel, *J. Phys. Chem.* **97** (1993) 8690; Q. Zhu *et al.*, *Nature* **355** (1992) 712. Links
549. D. W. Murphy *et al.*, *J. Phys. Chem. Solid.* **53** (1992) 1321.
550. J. P. McCauley, Jr., *J. Am. Chem. Soc.* **113** (1991) 8537. Links
551. P. Boulas *et al.*, *J. Electrochem. Soc.* **140** (1993) L130. Links
552. A. Penicaud *et al.*, *J. Am. Chem. Soc.* **113** (1991) 6699; P W. Stephens *et al.*, *Nature* **355** (1992) 331; Links G. Sparn *et al.*, *Solid State Commun.* **82** (1992) 779; Links F. Wudl, J. D. Thompson, *J. Phys. Chem. Solids* **53** (1992) 1449; Links K. Yoshizawa, T. Sato, K. Tanaka, T. Yamaba, K. Okahara, *Chem. Phys. Lett.* **213** (1993) 498; Links J. Stinchcombre, A. Penicaud, A. Penicaud, P. Bhyrappa, P. D. W. Boyd, C. A. Reed, *J. Am. Chem. Soc.* **115** (1993) 5212; C. Bossard *et al.*, *J. Chem. Soc., Chem. Commun.* (1993) 333.
553. J. R. Heath *et al.*, *J. Am. Chem. Soc.* **107** (1985) 7779. Links
554. Y. Chai *et al.*, *J. Phys. Chem.* **95** (1991) 7564; R. D. Johnson, D. S. Bethune, C. S. Yannoni, *Accounts of Chem. Res.* **25** (1992) 98.
555. R. E. Smalley in G. S. Hammond, V. J. Kuck, (eds.): *Large carbon Clusters* American Chemical Society, Washington, DC 1991 p. 91. R. E. Smalley, *Mater. Sci. Eng. B:* **19** (1993) 1.
556. E. G. Gillan, *J. Phys. Chem.* **96** (1992) 6869.
557. M. M. Ross, J. M. Callahan, *J. Phys. Chem.* **95** (1991) 5720; K. A. Caldwell, D. E. Giblin, M. L. Gross, *J. Am. Chem. Soc.* **114** (1992) 3743. Links H. Schwarz, T. Weiske, D. K. Böhme, J. Hrusák, in W. E. Billups, M. A. Ciufolini (eds.): *Buckminsterfullerenes* VCH, New York 1993 p. 257.
558. M. Saunders, H. A. Jiménez-Vásquez, R. J. Gross, R. J. Poreda, *Science* **259** (1993) 1428; Links M. Saunders *et al.*, *J. Am. Chem. Soc.* **116** (1994) 2193. Links
559. R. Shimshi, R. J. Gross, M. Saunders, *J. Am. Chem. Soc.* **119** (1997) 1163.
560. T. A. Murphy *et al.*, *Phys. Rev. Lett.* **77** (1996) 1075. Links C. Knapp, K. P. Dinse, B. Pietzak, M. Waiblinger, A. Weidinger, *Chem. Phys. Lett.* **272** (1997) 433.
561. B. Pietzak *et al.*, *Chem. Phys. Lett.* **279** (1997) 259.
562. S. Iijima, *Nature*, **354** (1991) 56. Links
563. M. Endo, *Chemtech* **18** (1988) 568.
564. T. W. Ebbesen, P. M. Ajayan, *Nature* **358** (1992) 220; Links T. W. Ebbesen *et al.*, *Chem. Phys. Lett.* **209** (1993) 83. Links
565. S. Amelinckx *et al.*, *Science* **265** (1994) 635; Links M. Endo, *Carbon* **33** (1995) 873. Links P. R. Birkett, C. Christides, *J. Chem. Soc., Perkin Trans.* **2** (1993) 1407.
566. M. Terrones, *et al.*, in: H. Kuzmany, J. Fink, M. Mehrings, S. Roth, (eds.): *Molecular Nanostructures*, Proceedings of the International Winter School on Electronic Properties of Novel Materials, World Scientific Publishing Co., Ltd., Singapore 1998 p. 375.
567. M. Endo *et al.*, *J. Phys. Chem. Solids* **54** (1993) 1841. Links
568. W. K. Hsu *et al.*, *Nature* **377** (1995) 687; Links W. K. Hsu *et al.*, *Chem. Phys. Lett.* **261** (1996) 161.
569. Z. Y. Kosakovskaya, A. Chernozatonskii, E. A. Fedorov, *JETP Lett. (Pis' ma Zh. Eksp. Teor.)* **56** (1992) 26.
570. D. S. Bethune *et al.*, *Nature* **363** (1993) 605; Links S. Iijima, T. Ichihashi, *Nature* **363** (1993) 605; C. H. Kiang, W. A. Godhard, *III* R. Beyers, D. S. Bethune, *Carbon* **33** (1995) 903. Links
571. A. Thees, *Science* **273** (1996) 483. Links C. Journet, in H. Kuzmany, J. Fink, M. Mehring, S. Roth, (eds.): *Molecular Nanostructures*, Proceedings of the International Winter School on Electronic Properties of Novel Materials, World Scientific Publishing Co., Ltd., Singapore 1998 p. 386.
572. M. M. J. Treacy, T. W. Ebbesen, J. M. Gibson, *Nature* **381** (1996) 678. Links
573. B. I. Yacobson, C. J. Brabec, J. Breeenholc, *Phys. Rev. Lett.* **76** (1996) 2511. Links
574. S. Iijima, C. Brabec, A. Maiti, J. Bernholc, *J. Chem. Phys.* **104** (1996) 2089.
575. D. Urgate, *Nature* **359** (1992) 707; Links D. Urgate, *Chem. Phys. Lett.* **207** (1993) 473; D. Ugrate, *Carbon* **33** (1995) 989; H. Kroto, *Nature* **359** (1992) 670.
576. A. Lundin, B. Sundqvist, *Europhys. Lett.* **27** (1994) 463; Links H. Kuwamura *et al.*, *Phys. Chem. Solids* **54** (1993) 1675.
577. N. S. Sariciftci, *Appl. Phys. Lett.* **62** (1993) 585; Links N. S. Sariciftci, A. J. Heeger, *Int. J. Mod. Phys. B* **8** (1994) 237. Links

578. A. Watanabe, O. Ito, *J. Chem. Soc., Chem. Commun.* **11** (1994) 1285.
579. J. Paloheimo, H. Isotalo, J. Kastner, H. Kuzmany, *Synth. Metals* **56** (1993) 3185; K. Hoshimono, S. Fujimoro, S. Fujita, *Jpn. Appl. Phys. Part 2, Lett.* **32** (1993) L1070; R. C. Haddon *et al.*, *Appl. Phys. Lett.* **67** (1995) 121. Links
580. J. B. Camp, R. B. Schwarz, *Appl. Phys. Lett.* **63** (1993) 445.
581. Y. B. Zhao, D. M. Poirier, R. J. Pechmann, J. H. Weaver, *Appl. Phys. Lett.* **64** (1994) 557.
582. Q.-Y. Tong, C. B. Eom, U. Gösele, A. F. Hebard, *J. Electrochem. Soc.* **141** (1994) L137. Links
583. D. Koruga *et al.*, *Fullerene C_{60} History*, Physics, Nanobiology, Nanotechnology Horth-Holland, Amsterdam 1993 "Chapter 10".
584. Y. J. Blau, C. E. Haberlin, *Thin Solid Films* **219** (1992) 129. Links B. Bhushan, B. K. Gupta, G. W. Vancleef, C. Capp, J. V. Coe, *Appl. Phys. Lett.* **62** (1993) 3253. Links B. Bhushan, B. K. Gupta, *J. Appl. Phys.* **75** (1994) 6156. Links
585. B. K. Gupta, B. Bhushan, *Lubrication Engineering* **50** (1994) 524. Links
586. R. Taylor *et al.*, *Nature* **355** (1992) 27.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Carbon Dioxide

SUSAN TOPHAM, Imperial Chemical Industries, Agricultural Division, Billingham, United Kingdom

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 5
4. Occurrence in Nature 7
4.1. Lithosphere 7
4.2. Atmosphere 7
4.3. Hydrosphere 7
4.4. Biosphere 8
4.4.1. Life Cycle 8
4.4.2. Radiocarbon Dating 8
5. Environmental Effects 8
6. Production 8
6.1. Carbon Dioxide from Ammonia and Hydrogen Plants (Process Gas CO_2) 8
6.1.1. Potassium Carbonate Process 10
6.1.2. Ethanolamine Process 11
6.2. Carbon Dioxide from Flue Gases . 11
6.3. Carbon Dioxide from Fermentation 11
6.4. Carbon Dioxide from Carbonates 12
6.5. Carbon Dioxide from Natural Sources 12
7. Purification of Carbon Dioxide before Liquefaction 12
8. Liquefaction of Carbon Dioxide . 12
9. Production of Dry Ice 13
10. Quality Specifications and Analysis 14
11. Commercial Forms, Storage, and Distribution 15
11.1. Solid Carbon Dioxide 15
11.2. Liquid Carbon Dioxide 15
12. Uses 16
12.1. Uses of Gaseous Carbon Dioxide . 16
12.2. Uses of Solid Carbon Dioxide . . . 16
12.3. Uses of Liquid Carbon Dioxide . . 16
12.3.1. Use as an Inert Environment 16
12.3.2. Promoting the Growth of Plants . . . 17
12.3.3. Heat Transfer Medium in Nuclear Power Stations 17
12.3.4. Refrigerant 17
12.3.5. Uses Based on the Solubility of Carbon Dioxide 17
12.3.6. Chemical Uses 17
12.3.7. Other Uses 18
13. Economic Aspects 18
14. Safety Precautions 18
15. References 19

1. Introduction

The existence of carbon dioxide has been known since primitive times [1]; PLINY wrote of lethal vapors from caverns in the first century A.D. [2]. It was first recognized as a distinct gas, however, by VAN HELMONT (1577 – 1644), who prepared the gas by various routes and demonstrated that it was the same gas as that which issued from caverns and mines [3]. He obtained it by fermentation, by the action of acids on carbonates, and by the burning of charcoal, and he studied many of its properties. Soon after, HOFFMANN, while investigating gas escaping from effervescent mineral waters, observed that the gas was acidic by its effect on blue vegetable coloring materials [4]. BLACK, in 1757, showed that during respiration part of the atmospheric air was chemically changed, and discovered the lethal effect such "fixed air" had on animal life [5]. CAVENDISH made various observations on the chemical properties of "fixed air" [6], and PRIESTLEY, who lived near a brewery, made further observations on the physical properties of the gas, finding that pressure favored dissolution of the gas in water [7]. The French chemist LAVOISIER was the first to prove the composition of "fixed air" by showing that it was produced when carbon is heated in oxygen [8]. The resultant gas, which he named carbonic acid, was found to contain 23.5 – 28.9 parts by mass carbon and 71.1 – 76.5 parts by mass oxygen [9].

The commercial exploitation of carbon dioxide began with attempts to produce artificial mineral waters (solutions of CO_2 in water), which were thought to have medicinal properties. The industrial development of the gas really began, however, with the experiments of FARADAY on the liquefaction of gases. He succeeded in liquefying CO_2 in a bent glass tube [10]. THILORIER repeated these experiments on a large scale [11]. Two iron retorts connected by a flexible

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_165

coupling were used; the first contained a reaction mixture of sodium hydrogen carbonate and sulfuric acid. This was agitated, and the carbon dioxide thus produced was liquefied under pressure in the second retort. Despite a spectacular explosion, which killed his assistant, THILORIER succeeded in carrying out extensive experiments on the expansion, vapor pressure, density, and enthalpy changes of the liquid CO_2 during evaporation. He was also the first experimenter to produce solid carbon dioxide, which, on the evaporation of liquid carbon dioxide, appeared as a white, flocculant, easily compressible mass. He observed that it disappeared by slow evaporation without first melting to form a liquid (i.e., it sublimed).

In 1845, FARADAY made larger quantities of liquid and solid carbon dioxide by means of a hydraulic pump [12]. He also used solid carbon dioxide, in a mixture with ether, as a refrigerant for use in further gas liquefaction studies. Soon after, NATTERER developed a mechanical compressor (a forerunner of the modern multistage type), which was used to make liquid carbon dioxide [13]. As uses for the new liquid were proposed, compressors were developed to produce greater quantities. The first factory for the production of liquid carbon dioxide was established by RAYDT in Germany in 1884, and the development of the industry continued as more uses were found [14]. In particular, its uses in ice-making and refrigeration and as a fire extinguisher became important. Much carbon dioxide issued from the earth naturally in places such as Germany. This could be used to make liquid CO_2, but to avoid the high costs of transporting cylinders of liquefied gas, new production methods were required. The first facility was established in Berlin in 1889 to make carbon dioxide from flue gases.

Since the turn of the century, a great many uses for carbon dioxide have been identified, and several other methods of manufacture have assumed commercial importance.

2. Physical Properties

Carbon dioxide [*124-38-9*], CO_2, M_r 44.010, is a colorless, odorless, nonflammable gas with a slightly sour taste. The regions of existence of the three phases are shown in Figure 1, the entropy – temperature diagram, and Figure 2, the enthalpy – pressure diagram [15]. Table 1 presents vapor pressure, density, and thermal data [16]. Detailed tables of physical, thermodynamic, and transport properties of solid, liquid, and gaseous carbon dioxide, together with a complete record and critical analysis of the original measurements, are available [17]. The physicochemical properties of carbon dioxide are summarized in references [18] and [19].

Gas density (0 °C, 0.1 MPa)	1.977 kg/m^3 (STP)
Specific density (compared with air)	1.5291
Molar heat c_p at 25 °C	37.13 J mol^{-1} K^{-1}
Entropy S_0	213 J mol^{-1} K^{-1}
Heat of formation ΔH	−393.51 kJ/mol
Free energy of formation ΔG_0	−394.2 kJ/mol
Specific heat c_p (gas, 0.1 MPa)	
at 0 °C	0.8277 J g^{-1} K^{-1}
at 20 °C	0.8459 J g^{-1} K^{-1}
at 60 °C	0.8833 J g^{-1} K^{-1}
Specific heat c_p (gas, 20 °C)	
at 1 MPa	0.9225 J g^{-1} K^{-1}
at 4 MPa	1.473 J g^{-1} K^{-1}
Molar heat (under saturated vapor)	
solid, at −123 °C	40.1 J mol^{-1} K^{-1}
at −63 °C	53.6 J mol^{-1} K^{-1}
liquid, at −43 °C	82.0 J mol^{-1} K^{-1}
at −13 °C	90.0 J mol^{-1} K^{-1}
at +17 °C	130.2 J mol^{-1} K^{-1}
Critical data	
Temperature	31.04 °C
Pressure	7.383 MPa
Density	468 kg/m^3
Triple point	
Temperature	−56.57 °C
Pressure	518 kPa
Heat of vaporization	347.86 J/g
Heat of fusion	195.82 J/g
Sublimation point	
Temperature	−78.92 °C
Pressure	98.07 kPa
Heat of sublimation	573.02 J/g
Gas constant R	8.48 J K^{-1} mol^{-1}
Correction factor χ_0 for the ideal gas equation	-9.2×10^{-6}
Van der Waals' constants for molar volume	
a	3.648 $\times10^{-7}$ MPa cm^6 mol^{-2}
b	42.672 cm^3/mol

$p-V$ values between 0 – 200 °C and 0 – 3.0 MPa, [20, p. 855]

Second virial coefficient B in atm^{-1} between −50 °C and 1000 °C [20, p. 861]

Compressibility factor Z in $pV = ZRT$ [21]

Isothermal throttling effect (at 30 °C and 243 kPa) $\varepsilon = \left(\frac{\delta H}{\delta p}\right) T$	−9.04 cm^3/g
Adiabatic throttling effect	

Table 1. Vapor pressure, density, and thermal data for saturated carbon dioxide [16]

Temperature, °C	Pressure, kPa	Density, kg/m^3 Condensed phase	Density, kg/m^3 Gaseous phase	Enthalpy, J/g Condensed phase	Enthalpy, J/g Gaseous phase	Entropy, J g^{-1} K^{-1} Condensed phase	Entropy, J g^{-1} K^{-1} Gaseous phase
−100	13.97	1595	0.4298	45.636	630.951	2.5104	5.8908
− 90	37.27	1582	1.0902	56.940	637.231	2.5749	5.7435
− 80	89.62	1525	2.515	68.664	642.674	2.6373	5.6095
− 78.92	98.07			70.045	643.176	2.6435	5.5948
− 70	198.1	1545	5.386	82.061	646.861	2.7043	5.4860
− 60	409.7	1522	10.962	99.227	649.236	2.7863	5.3671
− 56.57	517.9	1513	13.797	105.507	649.373	2.8156	5.3273
− 56.57	517.9	1177.8	13.797	301.450	649.373	3.7200	5.3273
− 50	683.1	1152.6	18.03	314.052	651.340	3.7765	5.2883
− 40	1005.1	1113.6	26.21	333.227	653.853	3.8594	5.2348
− 30	1428.1	1072.7	37.04	352.487	655.485	3.9389	5.1854
− 20	1969.1	1029.3	51.33	372.322	656.407	4.0168	5.1380
− 10	2646.6	981.9	70.33	393.936	655.653	4.0976	5.0924
0	3481.7	928.5	96.26	418.680	653.685	4.1868	5.0472
10	4498.8	864.2	134.59	445.894	647.237	4.2781	4.9894
20	5728.9	777.7	193.90	477.295	632.625	4.3827	4.9128
31.04	7383.4	468.0	468.0	558.938	558.938	4.6465	4.6465

(Continued)

(Joule-Thomson effect) at 101.3 kPa

$\mu = \left(\frac{\delta T}{\delta p}\right) H = \frac{-\varepsilon}{c_p}$

at −50 °C	24.130 °C MPa^{-1}
at −25 °C	16.500 °C MPa^{-1}
at 0 °C	12.900 °C MPa^{-1}
at +50 °C	8.950 °C MPa^{-1}
Adiabatic throttling effect at 20 °C	
at 0.1013 MPa	11.050 °C MPa^{-1}
at 2.026 MPa	11.355 °C MPa^{-1}
at 6.078 MPa	1.435 °C MPa^{-1} (condensation)
Thermal expansion coefficient γ	
gas, at 0 °C, 98.1 kPa	375.1×10^{-5} K^{-1}
20 °C, 98.1 kPa	374.2×10^{-5} K^{-1}
liquid, −50 to 0 °C	495×10^{-5} K^{-1}
solid at triple point	185.5×10^{-5} K^{-1}
Thermal conductivity at 25 °C, 0.1 MPa	1.64×10^{-4} W cm^{-1} K^{-1}
Viscosity (constants in the Sutherland formula) at equilibrium liquid – gas at 20 °C	
gas	20.3×10^{-6} Pa s
liquid	70.1×10^{-6} Pa s
Dielectric constant ε	
liquid −5 °C	1.60
liquid 0 °C	1.58
ε −1 (gas, 0 °C, 0.1 MPa)	989×10^{-6}
CO_2 diffusion	
in air at 0 °C	0.1420 cm^2 s^{-1}
40 °C	0.1772 cm^2 s^{-1}
in H_2 at 27 °C	0.806 cm^2 s^{-1}
127 °C	1.272 cm^2 s^{-1}
277 °C	2.09 cm^2 s^{-1}

Thermal conductivity of binary gas mixtures (of CO_2 with CO, N_2, O_2, H_2) see ref. [20, p. 1470]

The effect of overpressures of foreign gases (H_2, N_2, air, etc.) on the vapor pressures of liquid and solid CO_2 see ref. [22]

Vapor pressures and boiling point curves of binary mixtures (CO_2 with CH_4, C_2H_4, C_2H_6, C_3H_8, SO_2) see ref. [20, p. 975]

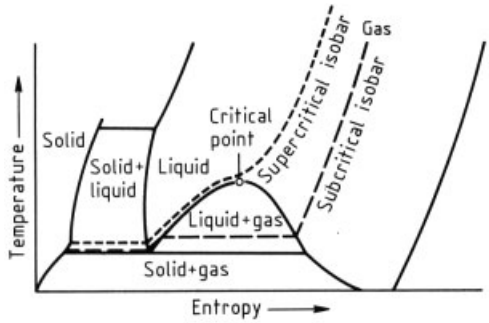

Figure 1. Phase diagram of carbon dioxide, 0 – 1200 MPa and −100 to +200 °C

Of all liquids, water is almost the least effective medium for dissolving CO_2. The solubility in many other liquids is much higher. Table 2 shows a few examples.

Solubility of CO_2 in water: Henry's law is approximately valid up to CO_2 partial pressures of about 500 kPa. At low temperature and high pressure there is a severe deviation (see Table 3 [23], [24]).

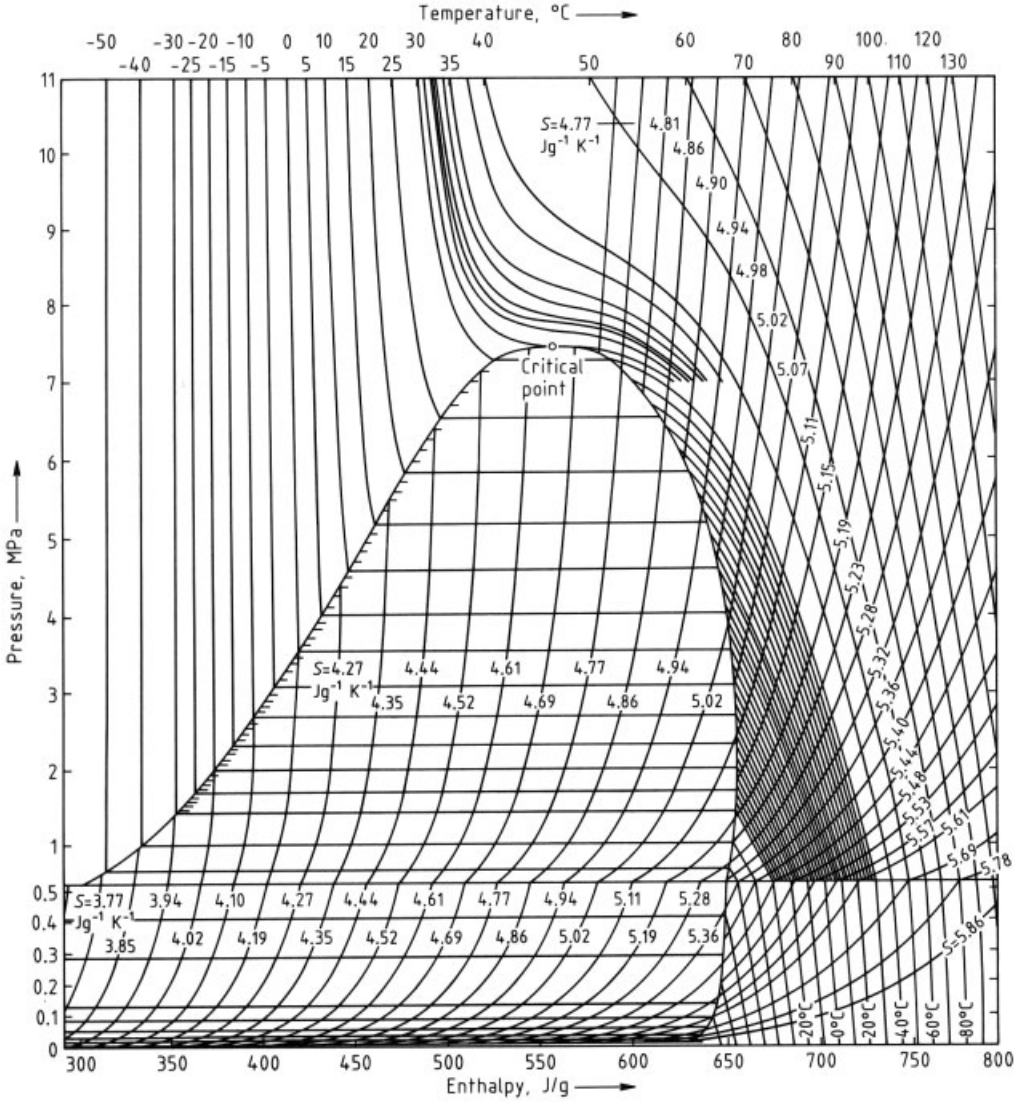

Figure 2. Enthalpy – pressure – temperature – entropy diagram for carbon dioxide (entropy, S, has the units $\mathrm{J\,g^{-1}\,K^{-1}}$)

Table 2. Solubility of carbon dioxide in various solvents at different temperatures

Solvent	Quantity dissolved, mL/g (STP) Temperature, °C											
	−80	−60	−40	−20	0	10	15	20	25	30	35	40
Acetone	460	127	50	24	13	10.5		8.2		6.6		5.4
Ethanol	100	40.4	28		5.3	4.3		3.6		3.2		
Benzene						2.9	2.81	2.71	2.69	2.59	2.43	
Methanol	220	66	24.5	11.4	6.3	5.0		4.1		3.6		3.2
Toluene	21	8.7	4.4	3.0	3.5	3.4	2.7	3.0	2.48	2.8		
Xylene		7.8	4.9	2.6	1.9		2.48	2.31	2.15			
Heptane								2.8	2.63		2.34	
Methyl acetate	350	101	41	20.5	11.5	9.2		7.4		6.0		
Diethyl ether	300	90	36	17.5	9.6	7.8		6.3				

Table 3. Solubility of carbon dioxide in water

Pressure, kPa	Amount dissolved, mL/kg (STP) Temperature, °C		
	10	20	30
101	1137	835	634
507	1103	804	620
1013	1062	778	602
1521	1024	753	587
2027	987	732	570

Heat of solution of gaseous CO_2 in water:

at 18 °C	19.92 kJ/mol
at 25 °C	20.30 kJ/mol

The solubility in ice at −5 °C to −20 °C is much less than in water at 0 °C [25].

The relationship between the dew point and the water content of CO_2 is shown in Figure 3.

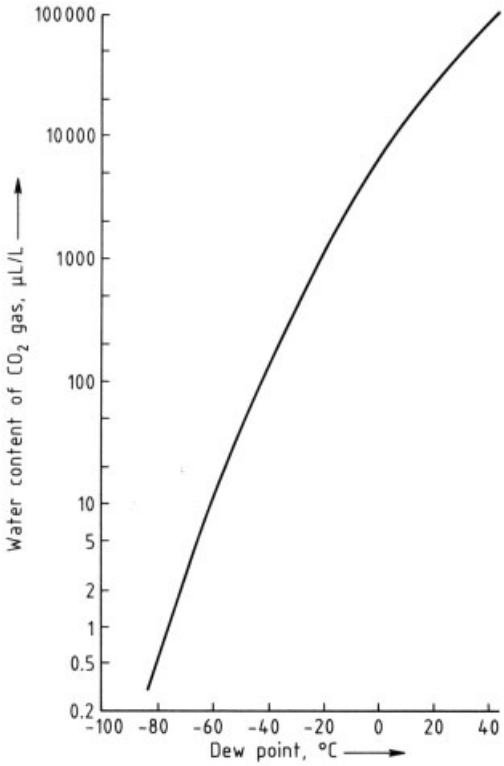

Figure 3. Dew point curve for carbon dioxide

The solubility of CO_2 in salt solutions and acids is given in ref. [20, p. 1210].

3. Chemical Properties

At normal temperature, gaseous carbon dioxide is not very reactive. The carbon dioxide molecule is relatively stable and does not readily break down into simpler compounds. Some decomposition may be effected, however, by the use of high temperature, ultraviolet light, or electrical discharge.

$$CO_2 \rightleftharpoons CO + 0.5\,O_2$$

Reactions between carbon dioxide and other substances are generally effected only at high temperature or by use of catalysts. Reduction to carbon monoxide [*630-08-0*] occurs by the reaction with hydrogen, the reverse of the water-gas shift reaction.

$$CO_2 + H_2 \rightleftharpoons CO + H_2O$$

Carbon monoxide is also formed by reaction with carbon [*7440-44-0*] at elevated temperatures.

$$CO_2 + C \rightleftharpoons 2\,CO$$

The latter reaction is of considerable importance in the smelting of iron ore; coke and limestone are also added to the furnace. The limestone first decomposes to form carbon dioxide, which is then reduced to carbon monoxide by the hot coke. Reduction of the iron ore to the metal is then effected by the carbon monoxide.

Carbon dioxide may be reduced to carbon by the action of metals (e.g., aluminum [*7429-90-5*] or magnesium [*7439-95-4*]) at high temperature; the reduction can be controlled so that it goes only part way (e.g., by using tin as the metal). Data for gas reactions in which carbon dioxide is either formed or decomposed are shown in Table 4 [26].

Several other reactions involving carbon dioxide are of commercial importance. Its reaction with ammonia [*7664-41-7*] forms ammonium carbamate [*1111-78-0*], which when dehydrated yields urea [*57-13-6*], a compound of considerable importance as a concentrated fertilizer and as a reagent in the plastics industry.

$$CO_2 + 2\,NH_3 \longrightarrow NH_2COONH_4$$

$$NH_2COONH_4 \longrightarrow NH_2CONH_2 + H_2O$$

Table 4. Some thermodynamic data for carbon dioxide

Temperature	Per mole CO_2 in the ideal gas state				Equilibrium constant K_p for formation of CO_2 from graphite and CO			
	Specific heat c_p	Enthalpy function $\frac{I-I_0}{T}$	Entropy S	Free enthalpy function $\frac{G-I_0}{T}$	$-R\ln K_p = \frac{\Delta I_0}{T} + \Sigma v_i \frac{G-I_0}{T}$			
					$C + O_2 \rightarrow CO_2$		$CO + \frac{1}{2}O_2 \rightarrow CO_2$	
	$p_{CO_2} = 0$	$p_{CO_2} = 0$	$p_{CO_2} = 101.3$ kPa	$p_{CO_2} = 0$	$\Delta I_0 = -393.43$ kJ/mol		$\Delta I_0 = -279.54$ kJ/mol	
K	$J\,K^{-1}\,mol^{-1}$	$J\,K^{-1}\,mol^{-1}$	$J\,K^{-1}\,mol^{-1}$	$J\,K^{-1}\,mol^{-1}$	log K_p	K_p	log K_p	K_p
0	0	0	0	0				
298.15	37.154	31.428	−182.356	213.782	69.0915	1.234×10^{69}	45.0437	1.106×10^{45}
300	37.237	31.465	−182.549	214.012	68.6680	4.656×10^{68}	44.7397	5.492×10^{44}
400	41.328	33.440	−191.873	225.313	51.5365	3.439×10^{51}	32.4096	2.568×10^{32}
500	44.640	35.360	−199.572	234.934	41.2582	1.812×10^{41}	25.0054	1.013×10^{25}
600	47.357	37.140	−206.150	243.291	34.4011	2.518×10^{34}	20.0649	1.161×10^{20}
800	51.498	40.242	−217.274	257.518	25.8266	6.709×10^{25}	13.8948	7.848×10^{13}
1000	54.407	42.797	−226.544	269.341	20.6768	4.751×10^{20}	10.1992	1.582×10^{10}
1500	58.573	47.461	−244.765	292.310	13.7986	6.290×10^{13}	5.2944	1.970×10^{5}
2000	60.709	50.535	−258.954	309.488	10.3258	2.117×10^{10}	2.8633	7.299×10^{3}
3000	62.802	54.345	−280.976	335.321	6.8437	6.977×10^{6}	0.5065	3.210

The liberated water can combine with the carbamate to produce ammonium carbonate [*506-87-6*] or ammonium hydrogencarbonate [*1066-33-7*].

$$NH_2COONH_4 + H_2O \longrightarrow (NH_4)_2CO_3$$

$$(NH_4)_2CO_3 \longrightarrow NH_4HCO_3 + NH_3$$

A gaseous mixture of carbon dioxide, carbon monoxide, and hydrogen [*1333-74-0*] can be converted to methanol [*67-56-1*] at moderate temperature and pressure by means of a copper – zinc catalyst.

$$CO_2 + 3\,H_2 \underset{\text{cat.}}{\rightleftharpoons} CH_3OH + H_2O$$

Addition of carbon dioxide to sodium phenolate [*139-02-6*] yields the sodium salt of salicylic acid [*54-21-7*], which is important in the manufacture of aspirin.

The solid octahydrate of carbon dioxide, $CO_2 \cdot 8\,H_2O$ [*27592-78-5*], can be formed if the gas is kept in contact with water at low temperature and high pressure.

Carbon dioxide in water solution is somewhat more reactive. Carbon dioxide dissolves in water to form a weak acid, carbonic acid [*463-79-6*], which undergoes reactions typical of such an acid. The following equilibria are set up:

$$CO_2\,(aq) \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons 2\,H^+ + CO_3^{2-}$$

In a solution of carbon dioxide in water, only ca. 0.1 % exists in the form of carbonic acid. The apparent primary dissociation constant (4.31×10^{-7} mol/L at 25 °C) suggests that the carbon dioxide solution is only weakly acidic in nature. This figure is based on the assumption that all the carbon dioxide dissolved and undissociated is present as H_2CO_3, but this is wrong. If the dissociation constant were based on the actual amount of H_2CO_3 present, then the acid would be regarded as moderately strong. (The actual primary dissociation constant is about 4×10^{-4} mol/L at 25 °C.) The apparent secondary ionization constant is 5.61×10^{-11} mol/L at 25 °C. At atmospheric pressure the pH of a carbon dioxide solution is 3.7; at a pressure of 560 kPa it is 3.3 and remains at this level at still higher pressures.

The rate at which carbon dioxide comes into equilibrium with H_2CO_3 and its dissociation products is slow, but the neutralization of H_2CO_3 by hydroxyl ions occurs instantaneously:

at pH < 8	$CO_2 + H_2O$	$\rightleftharpoons H_2CO_3$	slow
	$H_2CO_3 + OH^-$	$\rightleftharpoons HCO_3^- + H_2O$	instantaneous
at pH > 10	$CO_2 + OH^-$	$\rightleftharpoons HCO_3^-$	slow
	$HCO_3^- + OH^-$	$\rightleftharpoons CO_3^{2-} + H_2O$	instantaneous

at pH 8 – 10 both mechanisms operate.

As a dibasic acid, carbonic acid can react to form hydrogencarbonates ($M^{I}HCO_3$) and carbonates ($M^{II}CO_3$). All hydrogencarbonates are

readily soluble in water and yield carbon dioxide when heated; these properties are important in the industrial recovery of carbon dioxide. All carbonates are relatively insoluble in water except those of the alkali metals.

4. Occurrence in Nature

Carbon dioxide occurs naturally in the earth's lithosphere, hydrosphere, and atmosphere. A balance exists between the processes by which it is produced and used up.

4.1. Lithosphere

The lithosphere is that part of the earth not covered by the oceans. An estimated 5.5×10^{16} t of carbon dioxide is bound in the earth's crust in the form of carbonates of calcium (chalk), magnesium (dolomite), and other metals and as a constituent of many organic compounds. All igneous rocks, when heated strongly under vacuum, emit a mixture of gases, the principal constituents of which are usually carbon dioxide, water, and hydrogen. Rocks containing carbonate easily yield carbon dioxide when heated as do bituminous shale materials. Gases constantly escape from the surface of the earth at volcanoes and eruptive vents as a result of the action of heat on the rocks of the earth's crust. Carbon dioxide is a constituent of nearly all of these gaseous mixtures, in concentrations ranging from a few percent to nearly pure CO_2. Many of these sources have long been known to have adverse effects on living organisms.

In addition to that given off as a gas, much carbon dioxide escapes from the earth in natural spring water. Two main classes of such water are known: those which are saturated with carbon dioxide under a partial pressure (as in normal air) and those containing carbon dioxide dissolved under pressure such that, as they come to the surface, the pressure is released and they effervesce. Carbon dioxide has been dissolved in these waters by a process occurring inside the earth. Because of the presence of the carbon dioxide, these waters are acidic and thus more likely to dissolve rock, hence the name "mineral waters".

4.2. Atmosphere

The atmosphere includes all of the gases surrounding the earth. Carbon dioxide is present in the earth's atmosphere at a level of about 0.03 % by volume. The exact level is dependent on the exact position on the earth's surface at which the sample is taken, the quantity of vegetation present, and the time of day (the level being higher at night). The total amount of carbon dioxide in the atmosphere is about 2.3×10^{12} t.

Processes which expel carbon dioxide into the atmosphere are numerous. In addition to the evolution of the gas from the earth via springs, wells, and volcanoes, various other mechanisms produce carbon dioxide. Combustion of carbonaceous material, respiration of plants and animals, decay of organic material, and industrial processes such as lime-burning and the production of hydrogen and ammonia all generate gaseous carbon dioxide.

There are also various processes that remove carbon dioxide from the air, e.g., photosynthesis, weathering of rocks, and chemosynthesis by certain bacteria. The interrelationship of these generating and consuming processes is described in Section 4.4.

4.3. Hydrosphere

The hydrosphere comprises the oceans, seas, lakes, streams, and other bodies of water upon the earth. The concentration of carbon dioxide in the seas and oceans varies widely, but an average figure may be taken as five volumes of gas per ten thousand volumes of water. The total CO_2 content of the seas and oceans is about 1.4×10^{14} t.

In sea water, carbon dioxide exists as carbonates, hydrogencarbonates, carbonic acid, and as the dissolved gas itself. The amount of carbonate is usually fixed, but the hydrogencarbonate is labile and will give up gaseous carbon dioxide if the temperature of the water increases or the partial pressure of carbon dioxide above it diminishes. A dynamic equilibrium thus exists whereby the colder parts of the ocean absorb CO_2 and the warmer parts expel it to the atmosphere.

4.4. Biosphere

4.4.1. Life Cycle

The biosphere comprises all living things. Carbon dioxide is an essential substance in the life cycle of plants and animals and, as such, is basic to all life on earth. It is absorbed from the atmosphere by the leaves of plants. By using energy from sunlight and the accelerating influence of chlorophyll [*1406-65-1*], it reacts with water to form glucose [*50-99-7*].

$$6\,CO_2 + 6\,H_2O \xrightarrow[\text{chlorophyll}]{h\nu} C_6H_{12}O_6 + 6\,O_2 \qquad \Delta H = 2803\ \text{kJ/mol}$$

Glucose is subsequently converted by the plant into sugar and starch. All animals depend ultimately on vegetable matter for their food, and convert these carbon compounds built up by plants into carbon dioxide and water. Energy is simultaneously released. This reaction is the reverse of photosynthesis.

$$C_6H_{12}O_6 + 6\,O_2 \longrightarrow 6\,CO_2 + 6\,H_2O \quad \Delta H = -2803\ \text{kJ/mol}$$

Thus, carbon is being continually exchanged between the bodies of plants and animals. Atmospheric carbon dioxide is the medium of this exchange.

4.4.2. Radiocarbon Dating

An interesting and useful application of atmospheric carbon dioxide is the determination of the age of organic matter. Naturally occurring carbon contains 98.89 % ^{12}C, 1.11 % ^{13}C, and traces of ^{14}C. Cosmic rays interact with nitrogen in the upper atmosphere to form ^{14}C. This then decays with a half-life of 5570 years, sufficient time for it to become thoroughly mixed into all living organisms, which are in equilibrium with atmospheric carbon dioxide. When such organisms die, the ^{14}C contained within them decays and is not replenished. By comparing this quantity with the proportion of radioactive carbon existing in present-day organisms, the age of the specimen can be determined with an accuracy of ± 5 %.

5. Environmental Effects

Other processes serve to increase significantly the level of carbon dioxide in the atmosphere, in particular the combustion of fossil fuels. Since the advent of the industrial revolution, the burning of oil, natural gas, and coal has increased dramatically, and measurements of atmospheric carbon dioxide levels over the last 25 years have shown that there is an annual increase on the order of 1 mg/kg [27]. Although it is generally acknowledged that atmospheric carbon dioxide levels are increasing, scientists disagree as to the likely effects of such an increase [28]. One theory is based on the ability of carbon dioxide to absorb and transmit radiation. Carbon dioxide is transparent to the sun's incoming shortwave radiation, but it absorbs the longwave radiation emanating from the earth's surface. An increase in the atmospheric carbon dioxide level could cause an increase in the earth's temperature by this mechanism, known as the "greenhouse effect" (the glass in a greenhouse performs the same function as the carbon dioxide). The ultimate effects of this temperature rise might be radical changes in climate and rainfall patterns and a rapid rise in the sea level caused by the melting of polar ice. Other scientists argue that there may be other contributing factors in the regulation of the earth's temperature and that the carbon dioxide question should not be addressed in isolation.

6. Production

Much of the carbon dioxide generated in the world is a byproduct of ammonia and hydrogen production which make much more carbon dioxide than is ever recovered. Other sources are still exploited, but these are generally less efficient, and financially less attractive.

6.1. Carbon Dioxide from Ammonia and Hydrogen Plants (Process Gas CO_2)

The process for the production of ammonia is outlined in Figure 4. Desulfurization of the hydrocarbon feedstock (e.g., natural gas, naphtha) is carried out before catalytic steam reforming of the hydrocarbon to give a gaseous mixture of

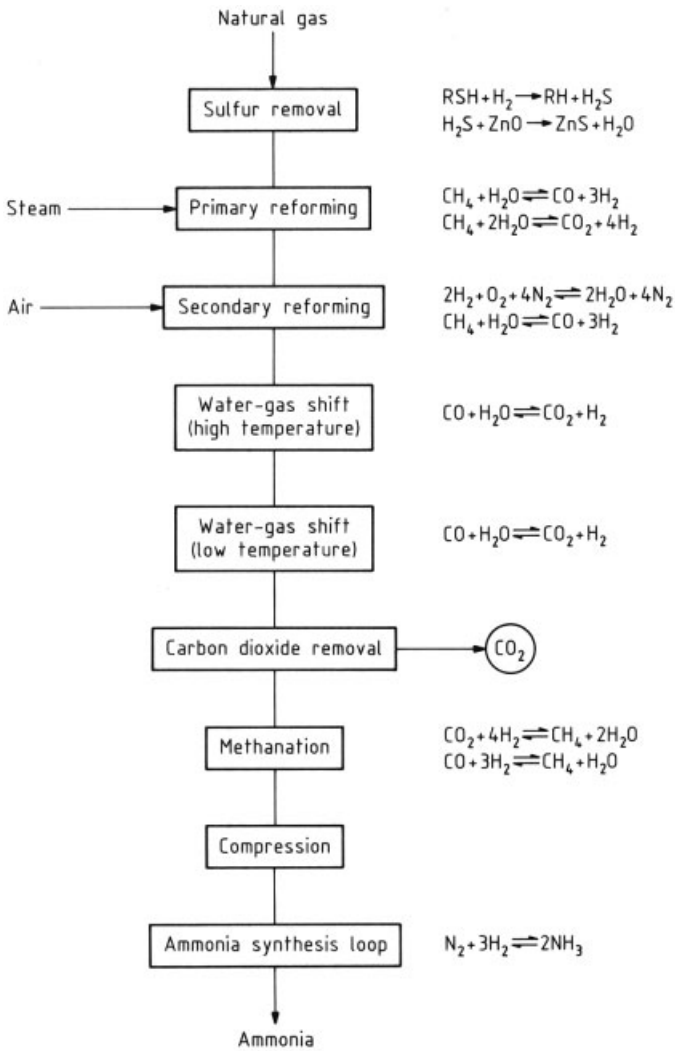

Figure 4. Schematic representation of an ammonia plant

hydrogen, carbon dioxide, and carbon monoxide. Air is added and further steam reforming is effected in a gaseous mixture that then also contains nitrogen [*7727–37–9*]. Because only hydrogen and nitrogen are required to make ammonia, the carbon oxides are removed from the gas stream. Most of the carbon monoxide is catalytically converted to carbon dioxide, and the latter is removed by dissolution under pressure.

In the manufacture of hydrogen, the process is very similar, with the obvious omission of the addition of the air. Thus, in both cases, carbon dioxide must be removed before the process-gas stream can be used. Several methods are available; all are based on the absorption of carbon dioxide into a solution, under pressure, which flows countercurrent to the process-gas stream. The absorbed gas is liberated in a separate vessel by raising the temperature and lowering the pressure above the solution. Under the normal process conditions, the solubility of carbon dioxide in water is too low for this to be used as the absorbent. The solubility is enhanced by adding a solute to the water so that chemical combination of carbon dioxide with the solution is effected.

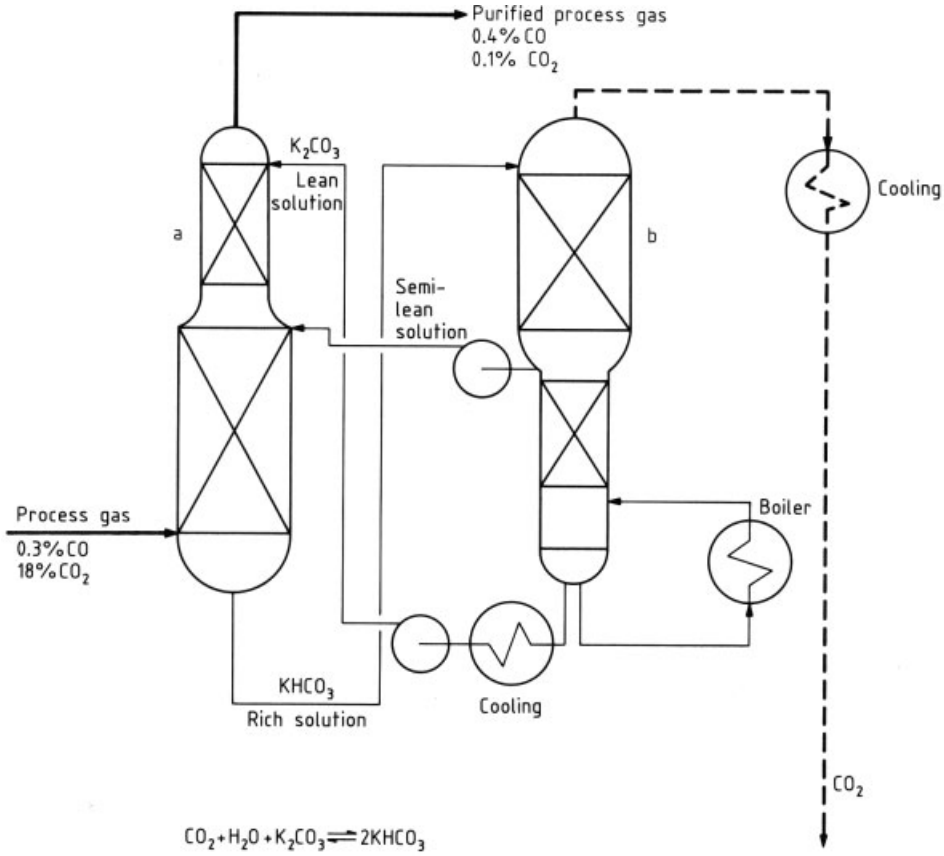

Figure 5. Carbon dioxide removal systema) Absorber; b) Regenerator

Alkali metal carbonates and ethanolamines have been employed in this capacity [29], [30].

6.1.1. Potassium Carbonate Process

The potassium carbonate [*584–08-7*] process is based on a reversible reaction.

$$K_2CO_3 + H_2O + CO_2 \rightleftharpoons 2\,KHCO_3$$

The rate of carbon dioxide absorption increases with decreasing temperature, increasing partial pressure of the gas, and increasing concentration of potassium carbonate solution. Figure 5 shows the usual arrangement of absorber and regenerator. The absorption should take place at a lower temperature than that of regeneration to maximize the amount of gas removed and recovered. This is achieved by a two-stage absorption and regeneration system. The solution containing carbon dioxide is pumped from the base of the absorber to the top of the regeneration column, supplied with a reboiler at its base. The solution is returned to the absorber from two points on the regenerator. Part of the solution is withdrawn from an intermediate point and is only partially regenerated; this is pumped to an intermediate point on the absorber. The remainder is withdrawn from the base of the regenerator. It is more completely regenerated, and is pumped, after it is cooled, to the top of the absorber. The carbon dioxide and water vapor mixture released in the regenerator is cooled. The water condenses out and the carbon dioxide is passed to the liquefaction plant.

Most commercial potassium carbonate systems also employ a catalyst to accelerate the rates of absorption and desorption of carbon dioxide. Corrosion inhibitors are also added to

the solution. Three processes are in most common use: the Benfield process (using diethanolamine [*111-42-2*] as catalyst and vanadium pentoxide [*1314-62-1*] as corrosion inhibitor), the Vetrocoke process (using arsenic trioxide [*1327-53-3*] as catalyst), and the Catacarb process (using unspecified additives).

6.1.2. Ethanolamine Process

The ethanolamine process, developed by the Girdler Corporation, uses an aqueous solution of ethanolamine instead of potassium carbonate. Monoethanolamine [*141-43-5*] reacts with carbon dioxide and water to form monoethanolamine carbonate [*21829-52-7*].

$$2\,HOC_2H_4NH_2 + H_2O + CO_2 \rightleftharpoons (HOC_2H_4NH_3)_2CO_3$$

6.2. Carbon Dioxide from Flue Gases

Carbon dioxide is a component of all flue gases produced by the complete combustion of carbonaceous fuels. Typical concentrations of carbon dioxide in such gases are 10 – 18 vol %. A schematic diagram of the process for recovering the carbon dioxide fraction of the gases is shown in Figure 6.

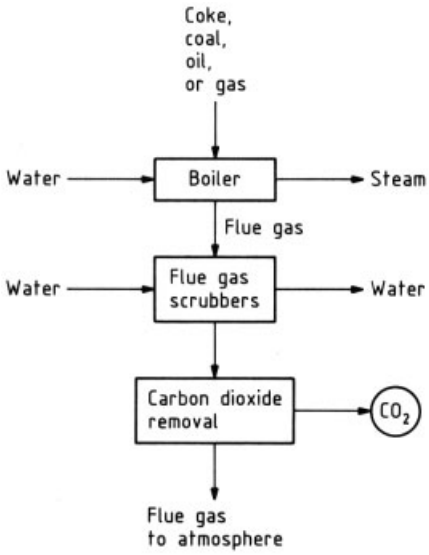

Figure 6. Carbon dioxide production from flue gases

The flue gases, after being cooled and cleaned by passing through a water scrubber, are passed through an alkaline carbonate solution or an amine solution which absorbs carbon dioxide. Unlike carbon dioxide from an ammonia plant, the product obtained from flue gases is generally contaminated with small amounts of sulfur compounds.

Although manufacture of carbon dioxide by this method was once of a considerable commercial importance, it is now seldom economically viable. It is, however, being revived in the United States for enhanced oil recovery projects, which require vast quantities of CO_2 [30], [32]. A refined process for recovering CO_2 from flue gases has been developed jointly by Dow and Procon [33]. It involves first mixing the flue gas with natural gas in an enrichment boiler. This mixture burns and thus removes the oxygen contained in the flue gas. In this way additional CO_2 is produced. The gas from the flue of the boiler is then fed to an absorption – regeneration system. The CO_2 is removed by using a concentrated aqueous solution of an amine, which also contains corrosion inhibitors, and is liberated by using steam supplied from the enrichment boiler [34].

A further development in the removal of carbon dioxide from flue gases, announced by Dow, involves a two-stage process. Caustic soda [*1310-73-2*] or soda ash [*497-19-8*] is first used to remove sulfur dioxide [*7446-09-5*], before the gas stream is passed through an amine solution which removes the carbon dioxide. An unspecified solvent is used, which, it is claimed, enables the attainment of two to three times the carbon dioxide loading achievable with conventional systems [35].

6.3. Carbon Dioxide from Fermentation

Large quantities of carbon dioxide are generated by fermentation processes, and up to 80 % of this gas may be recoverable. Before being suitable for further use, however, the carbon dioxide must be freed of the impurities inherent in this method of manufacture, i.e., hydrogen sulfide [*7783-06-4*], sulfur dioxide, and various organic compounds such as aldehydes, acids, and higher alcohols and diols. Two general methods

are available to purify fermentation carbon dioxide. Both use water scrubbers to remove the bulk of the entrained material. The impurities are then taken out by passing through either an activated charcoal bed or solutions of potassium permanganate [*7722-64-7*] and potassium dichromate [*7778-50-9*]. The first method relies on adsorption, whereas the second involves chemical reactions. In the second case the gas must be treated further downstream to remove oxidation products formed in the earlier stages as well as any traces of the reagents used in the purification.

6.4. Carbon Dioxide from Carbonates

The production of carbon dioxide by the calcination of calcium carbonate [*471-34-1*] and magnesium carbonate [*546-93-0*] is no longer economically important.

6.5. Carbon Dioxide from Natural Sources

Natural sources of gases containing very high concentrations of carbon dioxide are present in various parts of the world. Often the gas is so pure that it only requires drying before liquefaction. The main impurities present in this source of carbon dioxide are methane [*74-82-8*] and hydrogen sulfide. These sources are becoming more important in areas where enhanced oil recovery is being practiced [32], [34], [36].

7. Purification of Carbon Dioxide before Liquefaction

Unlike carbon dioxide derived from fermentation processes, which requires considerable purification as described, most of the gas from other sources is very pure. It contains only trace amounts of foreign materials, usually sulfur compounds. Because the ultimate use of the gas often involves foodstuffs, these trace impurities must be removed since they often impart distinctive tastes and odors. This is achieved by the use of activated carbon. Drying of the gas is also very important, because insufficiently dried material tends, on liquefaction, to choke the lines in the plant. Activated alumina, silica gel, or molecular sieve beds are used for drying. Most plants are multistreamed and are equipped with beds that can be reactivated by means of hot carbon dioxide, nitrogen, air, or steam.

The level of foreign gases present in the carbon dioxide stream must be minimized to reduce the energy required to liquefy and concentrate the circulating gas. The solubility of gases in liquid carbon dioxide decreases with decreasing temperature. If any are present, they separate out easily during the liquefaction stage.

8. Liquefaction of Carbon Dioxide

The critical parameters of carbon dioxide, $T_{crit} = 31\,°C$, $p_{crit} = 7.4\,MPa$, show that the gas may be liquefied at any temperature between 31 °C and its triple point (−56.6 °C) by compression to the required liquefaction pressure at that particular temperature and removal of the heat of condensation. It is, therefore, possible to liquefy CO_2 near the critical temperature by using a pressure of 7.4 MPa and water as the coolant, but because it is often difficult to guarantee a constant water supply at temperatures below 20 °C, this method is less favored than one in which a refrigerant is used. In this second method, the gas is compressed under 2 – 4 MPa before being cooled to subzero temperatures by means of a refrigerant. The choice of refrigerant obviously is a function of the cost and availability of the likely materials at their point of use. The proximity of many carbon dioxide liquefaction plants to ammonia production facilities means that ammonia is often the preferred refrigeration medium. Fluorocarbons and brine are used in some plants. A typical carbon dioxide liquefaction plant is shown in Figure 7.

Raw carbon dioxide piped from the ammonia plant is cooled by water and then passed to the suction side of three parallel, three-stage, electrically driven reciprocating compressors. After compression under oil-free conditions, the gas is scrubbed with water to remove traces of organic material formed in catalytic stages of the purification train of the ammonia process. Two further purification stages are then carried out, and traces of oil and sulfur compounds are removed in an activated charcoal bed. The gas then passes through one of a pair of parallel driers containing

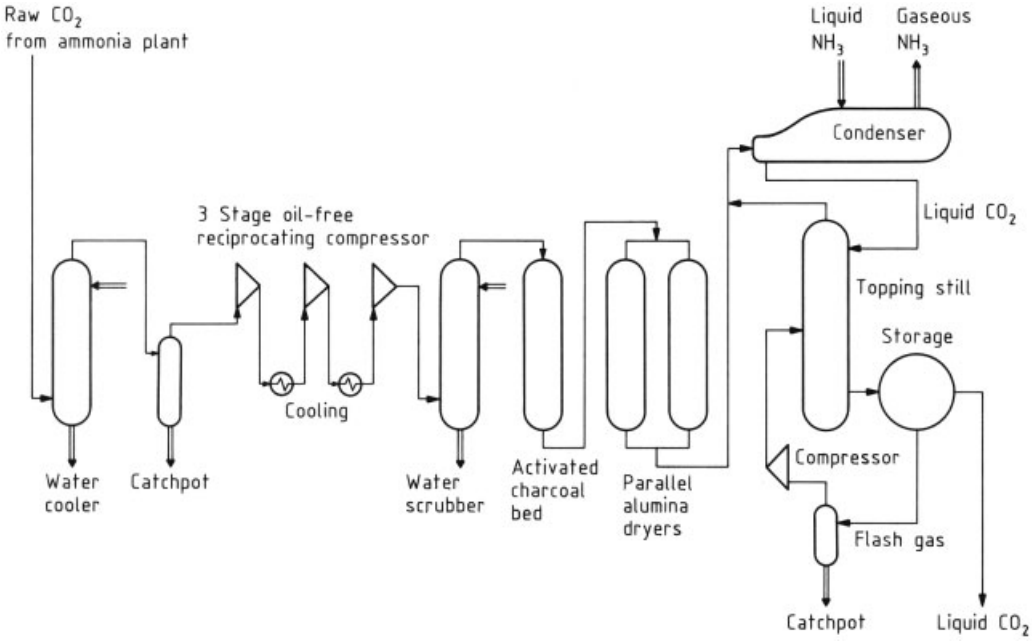

Figure 7. Schematic representation of a CO_2 liquefaction plant

activated alumina. The water content is reduced to mg/kg levels. (The alumina absorbent is periodically regenerated by passing a countercurrent of warm, dry carbon dioxide through the bed.) The dried, purified carbon dioxide is then liquefied in a double-pass condenser, using liquid ammonia at $-12\,^{\circ}C$ as refrigerant. The gaseous ammonia resulting from the transfer of heat is passed back to the plant ammonia-gas main. The liquefied carbon dioxide is passed down a stripping column in which entrained permanent gases (hydrogen, nitrogen, and methane) are removed by a countercurrent flow of purer carbon dioxide gas from recycle compressors. This stripping gas is the flash gas from the storage vessels of liquid carbon dioxide. The carbon dioxide from the top of the column, now rich in permanent gases, joins the main gas stream from the compressors and goes on to the condenser. Purging of the permanent gases is effected by means of a pressure-controlled blow-off to the atmosphere.

The pressure is lowered and the liquid carbon dioxide is sent to storage tanks, from which both road and rail tankers may be filled. The liquid may also be processed further to form dry ice.

9. Production of Dry Ice

Dry ice is thus named because it sublimes without residue. It is formed as a snow when liquid carbon dioxide is allowed to expand to atmospheric pressure. The density of this snow is 0.4 – 0.6 kg/L. Commercial manufacture of dry ice involves the densification of this loose snow by means of a hydraulic press to form solid blocks of material of density 1.5 – 1.6 kg/L. A typical unit for the production of dry ice is shown in Figure 8.

Liquid carbon dioxide is piped from the liquid storage facility, and the pressure is reduced in a suitable vessel. The resultant flash gas is passed back to the gas system. A measured dose of the liquid is then fed into one of several presses, where it is allowed to expand. Solid carbon dioxide forms along with some carbon dioxide gas. This flash gas is also returned to the gas system. When off-gassing is complete, the solid material is compressed and ejected as a large block, which is conveyed on a belt to an automated system of band saws. The block is cut into smaller pieces, the final size being $25 \times 19 \times 19$ cm. The dry ice is then weighed, wrapped, and packed into large, insulated containers. If smaller slices

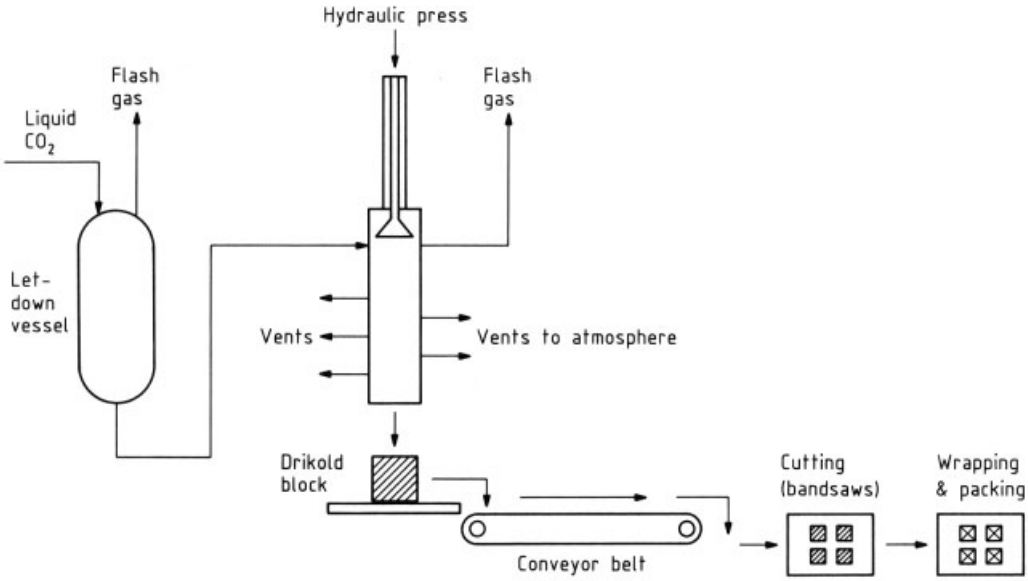

Figure 8. Schematic representation of the ICI Drikold plant for the production of dry ice

are required (e.g., for use on aircraft), further automated slicing and wrapping is carried out.

It is important that the carbon dioxide used in this process is free from moisture, which can block the lines and cause the formation of crumbly dry ice.

Continuous loss of carbon dioxide by sublimation and consequent diminution in size of the blocks is inevitable during storage of dry ice. It is necessary, therefore, to schedule production carefully so that the product is made as near as possible to the time it is used.

Since considerable handling of solid carbon dioxide is involved in the manufacture of dry ice, automation of the plants is common, to minimize costs, maximize efficiency, and improve hygiene.

Solid carbon dioxide is also made in pellet form. Typical pellet dimensions are 1 cm in diameter and 1 – 3 cm in length. Pellets are advantageous to customers who would normally break up the larger blocks before use, but pellets have the disadvantage of being much more susceptible to sublimation because of their greater surface-to-volume ratio. They also have a lower bulk density (0.8 – 1.0 kg/L) and, thus, occupy more storage space.

10. Quality Specifications and Analysis

Most of the carbon dioxide produced is of a final purity greater than 99.7 % by volume, and purities of 99.9 % by volume should be achievable if sufficient care is taken with the purification stages. Carbon dioxide from ammonia and hydrogen plants is usually of a high degree of purity before being subjected to purification and liquefaction; therefore, use of this source of starting material facilitates the production of very pure final product. Refined carbon dioxide from natural sources often still contains a small quantity of methane.

The necessary degree of purity is dependent on the final use to which the carbon dioxide will be put. Because a large proportion is used in the food and drink industries, the major criteria for the quality of the carbon dioxide are that it should be free of odor and taste. Thus, all contaminants which could contribute to these two properties (e.g., sulfur compounds, oils, and hydrocarbons) should be removed, preferably to less than mg/kg levels. The main impurities present to any significant degree in the final purified product are usually nitrogen, oxygen [*7782-44-7*], and argon [*7440-59-7*] (from

air), and hydrogen and carbon monoxide (if the carbon dioxide came from a process-gas source).

To meet the quality specifications described above, the purified carbon dioxide is analyzed regularly, using both on-line techniques and laboratory analyses of samples taken from the plant. Permanent gases and hydrocarbons are usually monitored by chromatographic techniques, and sulfur compounds, oil, and other trace impurities are determined by using standard analytical procedures. The moisture content of the carbon dioxide is determined by hygrometry.

The use of carbon dioxide in the food and drink industries is subject to the foodstuffs regulations in effect in the particular country of use. In some cases, it is necessary to carry out microbiological tests.

11. Commercial Forms, Storage, and Distribution

Carbon dioxide is marketed as a liquid or solid, and since the early 1960s the sale of liquid CO_2 has exceeded that of the solid form.

11.1. Solid Carbon Dioxide

Solid CO_2, commonly called dry ice, is formed below the triple point of CO_2, and there are a variety of patented devices for its production. Generally, a conversion of about 60 % liquid to solid is achieved, the flash gas being recycled. The product snow is either compressed into blocks of 10 – 25 kg or extruded into pellets of various sizes. Because dry ice continually sublimes even in well-insulated containers, it is rarely stored at the point of manufacture, but is made and distributed on demand. Dry ice is distributed in insulated containers of varying capacity that should be designed to keep losses to less than 3 % per day. Individual blocks or packs are sold from supply depots.

11.2. Liquid Carbon Dioxide

Liquid CO_2 is generally stored in bulk at −18 °C and 2.1 MPa in insulated vessels fitted with refrigeration units, level gauges, and relief valves. The vessels are normally fabricated of stainless or low-carbon mild steel. If the latter is used, national design codes may impose the need for stress relief to guard against low-temperature embrittlement. The lowest in-service temperature possible is that of solid CO_2, −78 °C. The imposition of furnace stress relief could limit the capacity of a mild-steel vessel to about 250 t.

Liquid CO_2 is distributed in cylinders or in bulk. In cylinders, CO_2 is at ambient temperature and high pressure. Cylinders are filled by means of a compressor to a filling ratio dependent on the expected storage temperature, i.e., temperate climate ratio = 0.75, tropical climate ratio = 0.67.

$$\text{Filling ratio} = \frac{\text{weight of liquid } CO_2 \text{ permitted}}{\text{water capacity by weight of cylinder}}$$

Stringent national design codes cover the manufacture of cylinders, and there are also codes of practice for their storage and handling. All CO_2 cylinders should be fitted with bursting discs. Bulk liquid CO_2 is distributed by road or rail in insulated containers at the same temperature and pressure as in storage. Generally, these containers are not refrigerated, and they are usually designed to allow for a pressure rise of 50 – 100 kPa per day. They are fitted with relief valves. Road tankers are fitted with transfer pumps and are of a capacity limited only by the gross vehicle weight. Rail tankers are not normally fitted with transfer pumps and can have a payload of up to 200 t. The capacity of storage tanks of the customer can be 1 – 250 t. These vessels are usually insulated, refrigerated, and fitted with pressure gauges and relief valves. Transfer from road or rail tanker to customer is normally achieved by equalizing the pressure between the two vessels in the gas phase and then pumping. As a rule, the consumer requires CO_2 in the gas phase and can achieve this either by direct removal of the gas phase from his storage vessel or, should the quantity required exceed the conversion rate of liquid to gaseous CO_2 by the natural heat leak of the vessel, convert liquid to gas by means of a steam, hot water, or electrically heated vaporizer.

12. Uses

12.1. Uses of Gaseous Carbon Dioxide

A large proportion (ca. 50 %) of all the carbon dioxide recovered is used at the point of production to make further chemicals of commercial importance, mainly *urea* and *methanol.* In these applications the carbon dioxide is used in its gaseous form, and its chemical properties are of prime importance. A further important use of carbon dioxide close to the source of the gas is in *enhanced oil recovery* in the United States [32]. In this application, CO_2 is added to aging oil wells to increase production of oil. To bring crude oil to the surface, a driving force must be present. In new wells this is provided by gases dissolved in the oil. As the well ages, other sources of energy are needed. Water is often used, and finally, a miscible gas, such as carbon dioxide, is added to reduce the interfacial forces of the entrained oil droplets and, thus, mobilize the remaining reserves. The source of carbon dioxide for this application if often natural wells adjacent to the oil well. Extensive CO_2 gathering and distribution facilities are required for this application [37]. Subsequent separation and recycling of the used CO_2 from mixtures also containing hydrocarbons is often carried out [38]. In this application, membranes have been used as the first stage of a three-stage removal process, while the second and third stages use amine solutions. The membrane system consists of many hollow fibers, over which the gas mixture passes; CO_2 is preferentially adsorbed on the surface of the fiber and diffuses through the membrane material into the hollow center. Up to 60 % of the CO_2 present in the gas mixture can be removed by this first stage.

The remainder of the carbon dioxide generated worldwide is converted to either liquid or solid form for use elsewhere, or is vented to the atmosphere, since transportation of gaseous carbon dioxide is not economically viable.

12.2. Uses of Solid Carbon Dioxide

Dry ice was originally the more important of the two nongaseous forms of carbon dioxide. Its use first became popular in the United States in the mid-1920s as a refrigerant for the preservation of food, and in the 1930s it became an important factor in the growth of the ice cream industry. Solid carbon dioxide was favored over the liquid form, which could only be moved about in cylinders. Road tankers had not been developed at that time.

The refrigeration range of dry ice extends to −78 °C at atmospheric pressure (lower temperatures can be achieved at reduced pressures). The reasons for the emergence of solid carbon dioxide as a refrigerant were threefold: it leaves no residue on evaporation, is not heavy, and may be allowed to come into direct contact with food because it is nontoxic. After World War II, changes in compressor design and the availability of special low-temperature steels made it possible to liquefy carbon dioxide on a very large scale. Liquid carbon dioxide consequently began to replace dry ice in many applications. The uses for which dry ice is still favored are essentially small-scale applications, where weight is an important factor, e.g., in the transportation of refrigerated food by road and air.

12.3. Uses of Liquid Carbon Dioxide

The uses of liquid carbon dioxide are many, and range from those in which the chemical composition is important to those in which it is not. In some applications, the liquid carbon dioxide may be converted to the gaseous form immediately prior to use.

12.3.1. Use as an Inert Environment

In these applications, carbon dioxide is used instead of an air atmosphere because the presence of air would cause an undesirable effect.

In the handling and transportation of foodstuffs, oxidation leading to a loss of flavor, or the growth of bacteria can be prevented by using carbon dioxide. Beer and lager are canned and bottled under carbon dioxide, and these liquids may also be transported by using carbon dioxide under pressure. Coffee is packed under carbon dioxide, and fruit, vegetables, and cereals are transported in an atmosphere of the gas. In such cases, the nontoxic nature of carbon dioxide is obviously very important. It competes with nitrogen as an inert medium in this application,

and the choice of gas is dependent on the relative costs of the two gases at the particular location.

Carbon dioxide is particularly suitable for use in fire extinguishers because it is heavier than air and extremely stable. It may be used on fires for which water cannot be used effectively, e.g., electrical or oil fires. Because carbon dioxide evaporates without residue and is nontoxic, its use does not result in further damage or contamination to the materials being treated.

Carbon dioxide is used extensively as a shielding gas in arc welding [39], in which its function is to protect the welding zone from the deleterious effects of oxygen, nitrogen, and hydrogen. Only a few gases possess the required shielding properties, and the alternatives to carbon dioxide (rare gases, particularly argon) are more expensive. Carbon dioxide may be used either in place of the rare gas or in conjunction with it, depending on the material to be welded.

A further use of carbon dioxide as an inert medium is in the movement of liquids which are either flammable or susceptible to deterioration by oxidation.

12.3.2. Promoting the Growth of Plants

The importance of carbon dioxide in the growth of plants has already been outlined. This is exploited by fruit and vegetable growers (particularly in the United Kingdom and the Netherlands), who introduce the gas into their greenhouses to give levels greater than that normally present in air (e.g., 1000 mg/kg compared with 300 mg/kg). The plants respond by raising their assimilation rate of carbon dioxide, and an increase in yields of about 15 % is achievable [40].

12.3.3. Heat Transfer Medium in Nuclear Power Stations

Carbon dioxide is used in certain nuclear power reactors as the intermediate heat transfer medium. It transfers heat from the fission processes to steam or boiling water in heat exchangers. It thus serves as a reactor coolant while indirectly generating steam for electrical power. Carbon dioxide is favored in this application because it does not become significantly radioactive and is available in large quantities with a high degree of purity. It is also substantially cheaper than alternative gases such as helium [*7440-59-7*]. Some radiolysis to CO and O may occur, but catalytic recombination units are fitted into the gas stream to regenerate the CO_2.

12.3.4. Refrigerant

The advantages of carbon dioxide as a refrigerant in the food industry have been described in Section 12.2. Liquid carbon dioxide is used extensively to freeze food (particularly in the United States) and also for its subsequent storage and transport.

Molded rubber and plastic components may be trimmed (deflashed) by rapid freezing with liquid carbon dioxide and tumbling in a barrel. The flash, being frozen and brittle, may easily be removed by this process, leaving smoothly trimmed items. This method is cheaper and quicker than hand or machine trimming.

12.3.5. Uses Based on the Solubility of Carbon Dioxide

Carbon dioxide has a moderate solubility in water, and use is made of this property in the production of both nonalcoholic and alcoholic effervescent drinks. This was the first major application of carbon dioxide.

The use of carbon dioxide in the aerosol industry is increasing steadily. This application depends on its solubility in the product to be dispensed [41]. The nontoxic nature of carbon dioxide makes it particularly suitable for aerosols of foodstuffs (e.g., whipped cream).

12.3.6. Chemical Uses

In the carbon dioxide – silicate process for the production of foundry molds and cores, use is made of the chemical reaction between carbon dioxide and silica, which serves to bind together the sand grains [42]. The mold is ready for use immediately, and intricate molds can be made in a very short time.

Sodium salicylate, one of the intermediates in the manufacture of aspirin, is made by the reaction of carbon dioxide with sodium phenolate.

Carbonation of softened waters is carried out by using carbon dioxide to eliminate the precipitation of insoluble lime compounds.

Carbon dioxide also finds use in the production of basic lead carbonate [*598-63-0*] and sodium, potassium, and ammonium carbonates and hydrogen carbonates. It is used as a neutralizing agent in mercerizing operations in the textile industry because it is more convenient to use than sulfuric acid [*7664-93-9*].

12.3.7. Other Uses

Liquid carbon dioxide is used in a coal mining process, in which it is fed down the mine shaft and on subsequent rapid expansion shatters the coal. High-pressure extraction of natural products, using liquid carbon dioxide as a solvent, may be used to isolate particular flavorings and fragrances.

Carbon dioxide has a multitude of other uses that account for only a relatively small amount of the product's use. Anesthetizing of animals prior to slaughter, cryobranding of animals, generating fog for theatrical productions, freezing of benign tumors and warts, lasers, production of lube oil additives, tobacco processing, and preburial sanitization are examples of such uses.

13. Economic Aspects

Only about half of all the carbon dioxide generated is recovered. Of this, approximately half is used at the source for the production of urea and methanol, or in enhanced oil recovery. The remainder is sold in the liquid or solid form.

Estimates of the total amounts of liquid and solid carbon dioxide produced throughout the world are of the order of several millions of tons per year in the United States, approximately one million tons per year in Europe (Great Britain, Germany, and France each produce a few hundred thousand tons per year), and approximately half a million tons per year in Japan. The proportion of this total that is made up of solid carbon dioxide has declined steadily since the 1950s, because liquid carbon dioxide has taken over many of its applications. Dry ice is estimated in 1985 to account for 10 – 20 % of the total.

The cost of transporting carbon dioxide is very high compared with the cost of production; thus, the vast majority of the carbon dioxide recovered is used in the country where it is made.

The major supply of carbon dioxide is dependent on the operation of plants for the production of other primary products (e.g., ammonia), the demand for which obviously depends on other factors (e.g., the variable demand for fertilizer). The value of the carbon dioxide produced is thus intimately linked to these other factors. For example, in 1985, the United States produced and consumed about four million tons of carbon dioxide worth more than $ 260 000 000 [43].

14. Safety Precautions

Carbon dioxide is a colorless, odorless, noncombustible, and nontoxic gas. Because it is heavier than air, it spreads along the ground. Although it forms a portion of exhaled air, high concentrations of carbon dioxide in respirable air can lead, without significant observable signs, to asphyxia. The TLV-TWA is 0.5 % by volume and the TLV-STEL is 1.5 % by volume [44]. This figure may be raised to 3.0 % by volume if the notice of intended changes (for 1984 – 85) proposed by the ACGIH is adopted. At 5 % by volume, there is an approximately threefold increase in breathing rate, and concentrations above this level rapidly lead to unconsciousnes and death. It is necessary, therefore, to ensure that adequate ventilation is provided in establishments where carbon dioxide is handled to prevent the accumulation of the gas in the respirable air.

Carbon dioxide monitoring equipment is usually provided in areas where there is a likelihood of building up significant concentrations of the gas. Alarms sound if the concentration of carbon dioxide exceeds 0.5 %.

When solid carbon dioxide is handled, there is a risk of unpleasant skin burns, which are a severe form of frostbite. Thick gloves should be worn to prevent contact of the refrigerant with the skin. Medical attention should be sought for such burns.

15. References

1. E. L. Quinn, C. L. Jones: "Carbon Dioxide," *A.C.S Monograph Series,* Reinhold Publ. Co., New York 1936.
2. Pliny the Elder, *Historia Naturalis* **2** (A.D. 77) 95.
3. J. B. van Helmont: *Ortus Medicinae,* ca. 1650.
4. Fr. Hoffmann: *De Methode Examinandi Aquas Salubres Lugduni Batavorum,* 1708.
5. J. Black: *Experiments upon Magnesia Alba, Quicklime, and Other Alcaline Substances,* Alembic Club, Edinburgh 1777.
6. H. Cavendish, *Philos. Trans. R. Soc. London* **56** (1766) 141.
7. J. Priestley, *Philos. Trans. R. Soc. London* **62** (1772) 127.
8. A. L. Lavoisier: *"Opuscules Physiques et Chimiques,"* Paris 1774. English translation: "Essays Physical and Chemical," translated with notes and appendix by Thomas Henry, 2nd ed., Frank Cass, London 1970.
9. A. L. Lavoisier, P. S. Laplace, *Mem. Acad. Sci. Paris histoire et memoires* (1780) 359, (1784) 387.
10. M. Faraday, *Philos. Trans. R. Soc. London* **113** (1823) 193.
11. M. Thilorier, *Ann. Chim. Phys. (2)* **60** (1835) 427.
12. M. Faraday, *Philos. Trans. R. Soc. London* **135** (1845) 155.
13. J. Natterer, *J. Prakt. Chem.* **31** (1844) 375.
14. W. Raydt, *Polytech. Notizbl.* **37** (1882) 196; DE 33168, 1885.
15. F. Cramer, *Chem. Ing. Techn.* **27** (1955) 487.
16. K. Ražnjević: *Handbook of Thermodynamic Tables and Charts,* Hemisphere Publishing Corp., Washington-London 1976 (Originally published as Termodinamičke Tablice i dijagrami by K. Ražnjević. Translated by Dr. Marijan Bošković and Professor R. Podhorsky).
17. M. P. Vukalovich, V. V. Altunin: *Thermophysical Properties of Carbon Dioxide,* Collets, London 1968 (Translated under the Direction of U. K. Atomic Energy Authority from 'Teplofizicheskie Svoistva Dvuokisi Ugleroda', Atomizdat, Moscow 1965).
18. A. L. Horvath: *Physical Properties of Inorganic Compounds (SI Units),* Arnold, London 1975.
19. S. Angus, B. Armstrong, K. Renck: *Carbon Dioxide – International Tables of the Fluid State (IUPAC),* **vol. 3,** Pergamon Press, Oxford-New York, 1976.
20. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* 3rd ed., vol. 1, Springer Verlag, Berlin-Heidelberg-New York 1967.
21. F. Schuster, *GWF Gas Wasserfach* (1974) no. 54.
22. *Ullmann,* 3rd ed., **vol. 9,** p. 751.
23. A. M. P. Tans, *Ind. Chem.* **38** (1962) 411.
24. *Landolt-Börnstein,* **II2 b,** 1 – 156, 157 (22261 + 22263).
25. R. E. Essery, R. Gane, *J. Inst. Brew.* **58** (1952) 129.
26. E. Schmidt: *Thermodynamik,* 6th ed., Springer Verlag, Berlin-Göttingen-Heidelberg 1956.
27. S. Terra, *EPRI J.* **3** (1978) no. 6, 22.
28. J. Williams: *Carbon Dioxide, Climate and Society,* Pergamon Press, Oxford 1978. R. A. Kerr, *Science* **232** (1986) 573.
29. A. V. Slack, G. R. James: *Ammonia,* Part II, Marcel Dekker, New York 1974, p.
30. *Nitrogen* **141** (1983) Jan. – Feb., 34.
31. C. Giavarini, M. Moresi: *Chem. Ind. (London)* 1982, Aug. 21, 601.
32. *Chem. Week* **117** (1975) Aug. 13, 71.
33. *Chem. Eng. News* **60** (1982) Oct. 11, 33.
34. *Oil Gas J.* **83** (1985) Jul. 15, 147.
35. *Chem. Eng.* **92** (1985) Sep. 30, 17.
36. *Oil Gas J.* **82** (1984) Oct. 22, 72.
37. *Oil Gas J.* **83** (1985) Jul. 8, 17.
38. *Oil Gas J.* **83** (1985) Sep. 9, 80.
39. A. A. Smith: *CO_2 Shielded Consumable Electrode Arc Welding,* British Welding Research Association, Keyworth and Fry, Lincoln, England 1962. L. M. Gourd: *Principles of Welding Technology,* Edward Arnold, London, 1980, p. 24. J. E. Jones, D. L. Olson, *Chemical Engineering* **93** (1986) 131.
40. D. W. Hand, *Sci. Hortic. (Canterbury, Engl.)* **33** (1982) 14.
41. Chemotron Corp. of Chicago, *Can. Chem. Process.* **59** (1975) Jul., 4.
42. K. E. L. Nicholas: *CO_2 Silicate Process in Foundries,* British Cast Iron Research Association, Alvechurch, Birmingham, Westerham Press, Kent, England, 1972.
43. *Chem. Eng. News* **63** (1985) Jul. 29, 10.
44. American Conference of Government Industrial Hygienists: *Threshold Limit Values for Chemical Substances and Physical Agents in the Work Environment and Biological Exposure Indices with intended Changes for 1984 – 85,* Cincinnati, USA, 1984.

Carbon Disulfide

MANCHIU D. S. LAY, Stauffer Chemical Co., Dobbs Ferry, New York 10522, United States

MITCHELL W. SAUERHOFF, Stauffer Chemical Co., Dobbs Ferry, New York 10522, United States

DONALD R. SAUNDERS, Stauffer Chemical Co., Dobbs Ferry, New York 10522, United States

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Thermodynamics and Rates of Reaction 4
4.2. Production from Charcoal and Sulfur 4
4.3. Production from Methane and Sulfur 5
4.4. Other Processes 5
5. Environmental Aspects 6
6. Quality Specifications and Analytical Methods 6
7. Storage and Transportation 7
8. Uses and Economic Aspects 7
9. Toxicology and Occupational Health . 8
10. Carbonyl Sulfide 10
11. References 10

1. Introduction

Carbon disulfide [*75-15-0*], CS_2, is an important industrial chemical with an extensive and well-developed chemistry. LAMPADIUS discovered carbon disulfide in 1796 while heating a mixture of iron pyrite and carbon. Industrial interest in this compound dates back to 1839, when it was prepared by SCHRÖTTER using charcoal and sulfur in a heated retort [1]. Its solvent power was recognized early, and for some time carbon disulfide was widely used for the extraction of fats and oils. It was, however, not until introduction in the early 1900s of the viscose rayon process, which uses carbon disulfide to solubilize cellulose, that it began to assume the role of a large-scale industrial chemical. By the late 1960s, competition from petroleum-based synthetic fibers began to affect the earlier rapid growth of rayon; the 1970s witnessed a decline of the rayon market and correspondingly that of carbon disulfide. Usage of carbon disulfide has since stabilized, and a small market growth is projected for the future.

Of the 1.1×10^6 t of carbon disulfide produced worldwide in 1984, over 75 % was used by the regenerated cellulose industry. The rest was consumed in production of carbon tetrachloride and miscellaneous organosulfur compounds used as rubber chemicals, flotation agents, and pesticides.

2. Physical Properties

Carbon disulfide is a volatile, dense liquid with a wide flammability range in air (1 – 50 vol %) [2]. It has an unusually low autoignition point (100 °C) and a low flash point (– 30 °C). Its evaporation rate into air is 1.6 times that of diethyl ether [3]. It is an excellent solvent for many organic compounds and readily dissolves sulfur, phosphorus, iodine, waxes, rubber, and resins. The commercial product has a pungent odor characteristic of sulfur compounds.

Some of the most useful physical, thermochemical, and transport properties of carbon disulfide are as follows (also see Table 1) [1], [4–13]:

M_r	76.13
fp	– 111.6 °C
bp at 101.3 kPa	46.23 °C
Density,	
liquid at 25 °C	1260 kg/m^3
saturated gas at 20 °C	8.07 kg/m^3
Critical temperature	273 °C
Critical pressure	7600 kPa
Surface tension at 20 °C	32.3 mN/m
Refractive index, liquid at 20 °C	1.62546
Electrical conductivity at 18 °C	$78 \times 10^{-19}\ \Omega^{-1}\ cm^{-1}$
Dielectric constant at 20 °C	2.63
Interfacial tension, $CS_2 - H_2O$	48.2 mN/m
Heat of combustion at 25 °C	1032 kJ/mol
Heat of formation,	
gas at 25 °C, 101.3 kPa	117.1 kJ/mol
Heat of fusion	4.390 kJ/mol
Heat of vaporization at 25 °C	27.41 kJ/mol

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_185

Specific heat, liquid at 25 °C	76.45 J mol^{-1} K^{-1}
Specific heat,	
gas at 25 °C, 101.3 kPa	46.2 J mol^{-1} K^{-1}
Entropy, gas at 25 °C, 101.3 kPa	237.83 J mol^{-1} K^{-1}
Free energy of formation,	
gas at 25 °C, 101.3 kPa	66.9 kJ/mol
Thermal conductivity,	
liquid at 25 °C, 101.3 kPa	0.162 W m^{-1} K^{-1}
gas at 25 °C, 101.3 kPa	0.0078 W m^{-1} K^{-1}
Viscosity,	
liquid at 25 °C, 101.3 kPa	0.36 mPa · s
gas at 25 °C, 101.3 kPa	0.0098 mPa · s
Diffusivity,	
gas at 0 °C, 101.3 kPa	
self-diffusion	0.0472 cm^2/s
diffusion in air	0.0892 cm^2/s
diffusion in CO_2	0.063 cm^2/s
liquid at 25 °C, 101.3 kPa	
self-diffusion	4.1×10^{-5} cm^2/s

Vapor pressure

t, °C	−10	0	10	20	30	46.22 (*bp*)	60	80
p, kPa	10.48	16.95	26.41	39.76	58.03	101.3	155.7	272.2
t, °C	100	125	150	175	200	225	250	273 (t_{crit})
p, kPa	450.0	785.8	1247	1930	2854	4108	5693	7615 (p_{crit})

Table 1. Mutual miscibilities of carbon disulfide – water [6]

Temperature, °C	CS_2 in H_2O, g/100 g	H_2O in CS_2, g/100 g
0	0.242	–
10	0.230	0.0086
20	0.210	0.012
30	0.185	0.017
40	0.111	0.022

3. Chemical Properties

The chemistry of carbon disulfide is extensive and well-investigated [1], [5], [14–16]. Comprehensive surveys of the inorganic chemistry of carbon disulfide have appeared [1], [5]. A detailed review of the organic reactions of carbon disulfide is given in reference [15].

Carbon disulfide reacts with sulfides, alcohols, ammonia and amines, and chlorine to give industrially important derivatives and intermediates. Aqueous alkalies react slowly with carbon disulfide to give the trithiocarbonate and carbonate:

$$3\,CS_2 + 6\,NaOH \longrightarrow 2\,Na_2CS_3 + Na_2CO_3 + 3\,H_2O$$

Trithiocarbonates are also formed with metal sulfides:

$$K_2S\ (\text{aqueous}) + CS_2 \longrightarrow K_2CS_3$$

Carbon disulfide reacts with alcoholic alkalies to produce xanthates (salts of dithiocarbonates) [17]:

$$ROH + NaOH \longrightarrow RONa + H_2O$$

$$RONa + CS_2 \longrightarrow ROCSSNa$$

This reaction is the basis for the viscose process in which cellulose is converted to the xanthate and regenerated as fibers and films:

$$(\text{cellulose})ONa + CS_2 \longrightarrow (\text{cellulose})OCSSNa$$

$$2\,(\text{cellulose})OCSSNa + H_2SO_4 \longrightarrow 2\,(\text{cellulose})OH + 2\,CS_2 + Na_2SO_4$$

Carbon disulfide reacts with ammonia and, depending on ammonia concentration and on temperature, forms ammonium dithiocarbamate (**1**) [*513-74-6*], ammonium trithiocarbonate (**2**) [*13453-08-2*], and ammonium thiocyanate (**3**) [*1762-95-4*] [14]:

$$CS_2 + 2\,NH_3 \longrightarrow \underset{\mathbf{1}}{NH_2CSSNH_4}$$

$$CS_2 + 6\,NH_3 + 3\,H_2O \longrightarrow \underset{\mathbf{2}}{2\,(NH_4)_2CS_3 + (NH_4)_2CO_3}$$

$$2\,CS_2 + 4\,NH_3 \longrightarrow \underset{\mathbf{3}}{(NH_4)_2CS_3 + NH_4(SCN)}$$

In the vapor phase and at higher temperature (150 – 160 °C), thiourea [*62-56-6*] is obtained via decomposition of the thiocyanate [14]:

$$NH_4(SCN) \longrightarrow SC(NH_2)_2$$

Reactions with primary and secondary amines produce principally the dithiocarbamate salts [18]:

$$2\,RNH_2 + CS_2 \longrightarrow RNHCSSNH_3R$$

$$2\,R_2NH + CS_2 \longrightarrow R_2NCSSNH_2R_2$$

In the presence of alkalies, many commercially important salts of dithiocarbamic acid are obtained:

$$R_2NH + CS_2 + NaOH \longrightarrow R_2NCSSNa + H_2O$$

Sodium methyldithiocarbamate [*137-42-8*], $CH_3NHCSSNa$, an important soil fumigant, is produced by the reaction of carbon disulfide with methylamine and an aqueous solution of caustic soda. In a similar reaction with ethylenediamine, the fungicide sodium ethylene bis(dithiocarbamate) [*142-59-6*] is the reaction product:

$$NH_2CH_2CH_2NH_2 + 2\,CS_2 + 2\,NaOH \longrightarrow \begin{array}{l} CH_2NHCSSNa \\ | \\ CH_2NHCSSNa \end{array} + 2\,H_2O$$

Two important vulcanization accelerators are products in the reaction of aniline with carbon disulfide. In the absence of sulfur, thiocarbanilide [*102-08-9*] is obtained:

$$C_6H_5NH_2 + CS_2 \longrightarrow C_6H_5NH\overset{\overset{\displaystyle S}{\|}}{C}NHC_6H_5 + H_2S$$

With sulfur, 2-mercaptobenzothiazole [*149-30-4*] (MBT) is produced:

$$C_6H_5NH_2 + CS_2 + S \longrightarrow$$ 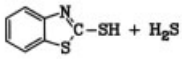 $$+ H_2S$$

Thiurium disulfides, a family of compounds used as fungicides and vulcanization accelerators, are formed by oxidizing dithiocarbamates [18]:

$$2\,R_2NCSSNa + H_2O_2 + H_2SO_4 \longrightarrow (R_2NCSS\text{-})_2 + Na_2SO_4 + 2\,H_2O$$

Chlorination of carbon disulfide gives rise to a host of chlorinated products [19]. At temperatures between 5 and 30 °C and in the absence of iron, iodine-catalyzed chlorination of carbon disulfide gives good yields of trichloromethanesulfenyl chloride [*594-42-3*]:

$$2\,CS_2 + 5\,Cl_2 \longrightarrow 2\,Cl_3CSCl + S_2Cl_2$$

Depending on the degree of chlorination, some carbon tetrachloride [*56-23-5*], sulfur dichloride [*10545-99-0*], and thiophosgene [*463-71-8*] are also formed.

In the presence of iron and metal chlorides and at higher temperature (70 – 100 °C), carbon tetrachloride and sulfur chlorides are exclusively obtained [14]:

$$CS_2 + 3\,Cl_2 \longrightarrow CCl_4 + S_2Cl_2$$

Further reaction of the sulfur monochloride [*10025-67-9*] with carbon disulfide produces more carbon tetrachloride and sulfur:

$$CS_2 + 2\,S_2Cl_2 \rightleftharpoons CCl_4 + 4\,S$$

These two reactions are the basis for the commercial production of carbon tetrachloride from the chlorination of carbon disulfide [20]. The reaction with chlorine is very fast and proceeds to completion. Carbon disulfide and sulfur (I) chloride react more slowly. The stoichiometric reaction is equilibrium limited at about 70 % at 90 °C.

Carbon disulfide reacts readily with water in the presence of alumina and oxide catalysts above 150 °C to form carbon dioxide and hydrogen sulfide:

$$CS_2 + 2\,H_2O \longrightarrow CO_2 + 2\,H_2S$$

Carbonyl sulfide is an intermediate in this hydrolysis reaction:

$$CS_2 + H_2O \longrightarrow COS + H_2S$$

$$COS + H_2O \longrightarrow CO_2 + H_2S$$

At 200 – 300 °C, equilibrium conversion of carbon disulfide is nearly complete [21].

At temperatures above 150 – 200 °C and in the presence of metal sulfide catalysts, notably molybdenum and nickel sulfides, carbon disulfide is reduced by hydrogen to give various amounts of methanethiol, dimethyl sulfide, methane, and hydrogen sulfide [22].

4. Production

Carbon disulfide is commercially manufactured by the reaction of sulfur with charcoal or methane. Ethane, propane, and propene have been used to a limited extent. Since the methane process was first introduced in the early 1950s, it has steadily supplanted the older charcoal process, which is no longer a factor in carbon disulfide manufacture in the United States, Europe, and Japan. In areas where natural gas or methane is not readily available or when plant size is relatively small, the charcoal process still supplies local viscose rayon requirements.

4.1. Thermodynamics and Rates of Reaction

The reactions of sulfur with either carbon or methane have been evaluated for the following equations [23]:

$$C_{(s)} + 2\,S_{(g)} \longrightarrow CS_{2\,(g)}$$

$$CH_{4\,(g)} + 4\,S_{(g)} \longrightarrow CS_{2\,(g)} + 2\,H_2S_{(g)}$$

The maximum yield from carbon was found to be 91 % at 720 °C. For the methane reaction, complete conversion was attained at 300 °C. In practice, higher temperature is required to provide economic reaction rates. The reactions are endothermic over most of the temperature range practiced commercially; the energy supplied is taken up mainly by vaporization and superheating of sulfur [23–25].

The thermodynamics of sulfur dissociation [7], [23], [24] determine the concentration of the reactive diatomic sulfur molecule, S_2; high temperature and low pressure favor high S_2 concentrations [7]. Equilibration rates of the vapor species, S_n ($n = 1-8$), are relatively fast so that equilibrium values of S_2 may be used in rate equation formulation.

Homogeneous and heterogeneous rate data for the reaction of methane and sulfur have been obtained for temperatures between 500 and 700 °C [26], [27–30]. The rate equation that best fits these data is one that is first order with respect to methane and S_2:

$$\text{rate} = k\,c_{CH_4}\,c_{S_2}$$

where k is the reaction rate constant. The following are the Arrhenius equations for the rate constants, based on reference [26]. For the homogeneous reaction:

$$k = 1.9 \times 10^6 \exp(-52000/RT)\ \text{mol}\,\text{h}^{-1}\,\text{cm}^{-3}\,\text{kPa}^{-2}$$

For the silica gel catalyzed reaction:

$$k = 7.7 \times 10^2 \exp(-28000/RT)\ \text{mol}\,\text{h}^{-1}\,\text{g}^{-1}\,\text{catal.}\,\text{kPa}^{-2}$$

Other values for these rate constants as well as space velocities for the catalyzed reaction have been reported [27–31]. At 600 °C, the homogeneous rate is about 1 % that for the silica gel catalyzed reaction.

4.2. Production from Charcoal and Sulfur

The basic process for the reaction of charcoal with sulfur has changed little since its inception in the mid 1840s. Externally heated retorts are still being used, although equipment design has undergone much improvement as better materials of construction have become available. Details on retort design are available in references [32], [33].

In retort plants [34], [35], molten sulfur is vaporized and superheated either before or after addition to the retort. Reaction between the superheated sulfur vapor and a stationary bed of charcoal at 850 – 900 °C takes place inside the retort at slightly above atmospheric pressure. The product gas, consisting of carbon disulfide, some sulfur, hydrogen sulfide, carbonyl sulfide, and inerts, passes through a train of condenser, scrubber, and oil absorber where the crude carbon disulfide is recovered; the pure product is obtained by further treatment in distillation columns. The tail gas, made up primarily of hydrogen sulfide, carbonyl sulfide, and inerts, is either incinerated and scrubbed with a caustic solution or routed to a sulfur recovery plant.

Typically, each retort can produce 1 – 3 t/d of carbon disulfide. High-temperature corrosion of the cast iron retorts limits their useful life to less than 1 year. The raw materials used in the retorts should be low in ash and residue to minimize the frequency of reactor cleanouts. Fouling of the retorts also restricts heat transfer and shortens retort life. Charcoal quality is extremely important because less reactive material requires higher temperature for satisfactory operation; charcoal derived from hardwood is preferred for this reason. The charcoal is normally precalcined at 500 °C to remove volatiles that could result in formation of objectionable byproducts. Thermal efficiency of retorts is ca. 25 %; carbon and sulfur yields are generally below 90 % [34–36].

Electrical resistance heating in brick-lined electrothermal reactors provides an alternative heat source for the reaction of charcoal with sulfur [37]. Where cheap power is available, the electrical method may be more economic. Power consumption as a function of carbon disulfide produced is ca. 1200 kW h/t. Other advantages

offered by electrothermal reactors are a larger capacity per reactor (up to 10 t/d) and a longer reactor life; furthermore, a wider range of solid carbon feed can be used because electrothermal reactors can generate higher temperatures.

4.3. Production from Methane and Sulfur

As demand for carbon disulfide rapidly increased in the 1940s, limitations of the charcoal processes provided the incentive for development of a more efficient large-scale method for its manufacture. Readily available cheap natural gas and favorable thermodynamics [21], [23], [31] led to the successful development of the process based on the catalyzed reaction of methane and sulfur [21], [27–31], [38], [39]. During the period 1943 – 1965, many patents were issued that covered various aspects of the process [38–54].

The commercial methane process [35], [55], [56] is based primarily on work described by the patents issued to the Pure Oil Co. [41–45]. The rights to these patents were acquired by the Food Machinery Corp. and Stauffer Chemical Co., and in 1953, the first methane process plant was built by FMC in South Charleston, West Virginia. This was followed by a Stauffer plant at LeMoyne, Alabama, three years later. By the mid 1960s all carbon disulfide in the United States was manufactured by this method; many similar plants soon followed in Europe and elsewhere [35]. Over 85 % of the world's production capacity is based on methane.

Figure 1 shows a typical methane process flow sheet. Purified natural gas containing over 95 % methane and low concentrations of propane and heavier hydrocarbons is preheated to about 250 °C in the convection section of the reaction furnace. The furnace is a typical box-type pyrolysis furnace commonly used in the petrochemical industry, with radiant and convection sections [57]. The hot methane gas is mixed with liquid sulfur and introduced into the horizontal heating coil located in the radiant section. There the liquid sulfur is vaporized, and the temperature of the mixture is raised to 550 – 650 °C. Operating pressure is typically 400 – 700 kPa. Alternatively, the methane and sulfur may be introduced separately and at different locations of the coil. A slight excess of sulfur (5 %) is used to ensure good methane conversion and to minimize carbon and tar formation.

To withstand the high-temperature corrosive environment, cast high-alloy stainless steel is used as tube material. Tube life is normally 2 – 3 years. Packed, brick-lined reactors are connected to the heating coil to provide additional residence time and housing for catalyst beds. Reaction takes place both in the tubes and in the reactors; almost all the hydrocarbons are converted by the time the gas leaves the last reactor.

Gas exiting the reaction furnace system passes through the sulfur condenser where most of the excess sulfur is separated. Next the cooled gas goes through a sulfur scrubber and then to the CS_2 scrubber in which the remaining sulfur is removed. The relatively sulfur-free gas then enters water-cooled condensers for bulk separation of carbon disulfide. The remaining carbon disulfide is finally removed from the hydrogen sulfide gas in the oil absorber. The hydrogen sulfide gas exiting the oil absorber contains 1 – 2 % methane and less than 1000 ppm carbon disulfide. This gas is routed to a Claus sulfur recovery plant where it is converted back to sulfur.

Crude carbon disulfide from the condensers, including that recovered in the stripper, is sent to the stabilizer column for removal of low-boiling impurities. The bottoms from the stabilizer column are next processed in the purification column where high-boiling contaminants are separated from the product. A caustic and water wash complete the product purification operations, and the finished product is pumped to check tanks and storage.

4.4. Other Processes

The use of different feedstocks for the manufacture of carbon disulfide is a subject of much economic interest. Raw materials evaluated include the following: methane and hydrogen sulfide [58], [59]; coke and coal with sulfur dioxide and hydrogen sulfide [60–62]; liquid hydrocarbons with sulfur [63–65]; hydrocarbons with sulfur and sulfur dioxide [66], [67]; chlorinated hydrocarbons and sulfur [68]; carbon monoxide and sulfur [36]; and organosulfur compounds with sulfur [69]. Most of these alternatives are unattractive because of poor yield, undesirable

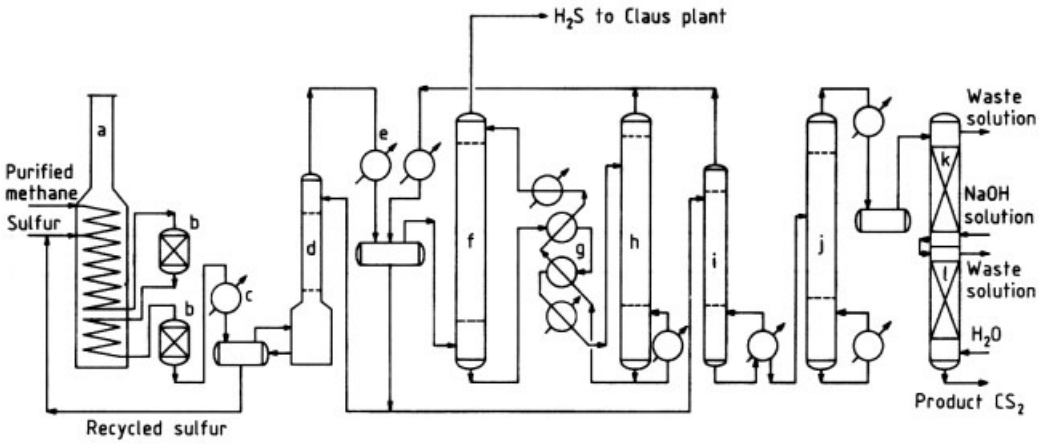

Figure 1. Methane process for CS_2 production
a) Reaction furnace; b) Catalytic reactors; c) Sulfur condenser; d) CS_2 scrubber; e) CS_2 condenser; f) CS_2 absorber; g) Oil interchangers; h) CS_2 stripper; i) Stabilizer column; j) Purification column; k) Caustic wash column; l) Water wash column

byproducts, or difficulties in processing due to tar and polymer formation [70].

Important refinements to the basic methane process include the use of high-pressure (2000 kPa) condensation and fractionation for product recovery and purification [54], and modifications to the reaction furnace system to allow the use of olefinic feedstocks [71], [72].

5. Environmental Aspects

Carbon disulfide in the atmosphere is of biogenic as well as industrial origin [73]. Tropospheric concentrations of carbon disulfide have been reported as 3×10^{-5} ppm in remote areas and $7\times10^{-5}-30\times10^{-5}$ ppm in urban and industrial locations [74]. Natural release includes diffuse emission from soil [75] and volcanic eruption [73]. Industrial sources consist mainly of release from gasification and pyrolysis of carbonaceous fuel in the presence of sulfur compounds, from sulfur recovery (Claus) plant tail gas, and from industrial users of carbon disulfide [73], [74].

Investigations on the fate of carbon disulfide in the atmosphere show that the principal oxidation products are sulfur dioxide, carbonyl sulfide, and carbon monoxide [73], [76], [77]. Eventually these compounds degrade to carbon dioxide and sulfates.

Carbon disulfide may be removed from gas streams by oil absorption or adsorption with activated carbon [78], [79]. A common method used in the treatment of carbon disulfide and carbonyl sulfide in sulfur recovery plant tail gas involves hydrogenation and hydrolysis of these compounds to carbon dioxide and hydrogen sulfide. The hydrogen sulfide is then recovered by absorption in alkanolamine solutions [78], [80].

6. Quality Specifications and Analytical Methods

Carbon disulfide produced commercially in modern plants is generally of high purity and meets most of the specifications listed in Table 2.

Analytical methods for detecting carbon disulfide are detailed in references [83], [84]. A common procedure, sensitive to 1 ppm, is based on the reaction of carbon disulfide with diethylamine in the presence of a cupric salt to give the characteristic yellow cupric diethyldithiocarbamate. Infrared spectrophotometry is sensitive to 5 ppm, and gas chromatography can provide detection to 0.5 ppm.

Table 2. Carbon disulfide specifications [81–83]

Property	Grade		
	Commercial	ACS reagent	Viscose
Residue, %	0.002	0.002	0.01
Relative density *d*, 20 °C	1.261 – 1.265	–	1.260 – 1.270
Color, APHA	20	10	colorless
Dissolved sulfides			
lead acetate test	negative	–	–
mercury drop test	bright	bright	–
Sulfite and sulfate	–	0.002 % SO_2	–
Boiling range	46 – 47 °C	1 – 95 mL < 0.5 °C; 95 mL – dryness < 0.5 °C	98.5 % should distill between 46 and 47 °C
Water, %	–	0.05	–

7. Storage and Transportation [81], [85], [86]

The toxicity and extreme flammability and volatility of carbon disulfide make it one of the most hazardous chemicals to handle. Facilities using carbon disulfide must be designed to operate, at all times, as a closed system and above atmospheric pressure to avoid the influx of air. Whenever possible, these plants should be located outdoors or in open structures away from flames and other sources of ignition. Equipment should be properly grounded to prevent accumulation of static electricity. Water deluge systems around processing and storage areas should be provided; enclosed units should be designed with proper ventilation of the downdraft or lateral types.

Areas where appreciable concentration of carbon disulfide may be present should be designated as Class 1 hazardous locations as defined by the National Electrical Code [87]; electrical specifications should be in accordance with the requirements of the code.

Carbon disulfide is stored in properly diked above-ground steel tanks or in tanks confined in open-top water-flooded concrete pits. Underwater tanks are water padded, whereas water or inert gas padding is used in above-ground tanks. Transfer of carbon disulfide from storage tanks may be by padded pressure or pumps [81].

Carbon disulfide is classified by the United States Department of Transportation (DOT) as a Class 1 flammable liquid (identification number 1131). Use of the red label for flammable liquid is required. Transport of this chemical should follow DOT regulations [88].

8. Uses and Economic Aspects

In 1984, an estimated 1.1×10^6 t of carbon disulfide was produced worldwide in the following areas (production figures in 10^3 t/a):

North America	235
Western Europe	260
Eastern Europe	250
Latin America	50
Asia	160
Africa	10
Former States of USSR	120
Total	1085

In 1991 1.3×10^6 t of carbon disulfide was produced worldwide in the following areas (production figures in 10^3 t/a):

North and Central America	190
South America	54
Europe and Africa	822
Asia	221
Total	1287

[89]

The regenerated cellulose industry is by far the largest user of carbon disulfide. Over 65 % of world production is used in the manufacture of rayon (ca. 0.32 kg of carbon disulfide is consumed for each kilogram of rayon produced). Some 10 % is used in making cellophane, and another 10 % is chlorinated to form carbon tetrachloride (about 25 % of world carbon tetrachloride is produced by this method). The remaining 15 % is distributed among rubber chemicals, flotation agents, pesticides, and miscellaneous chemical reagents. Table 3 lists uses of carbon disulfide and some of its derivatives.

Table 3. Uses of carbon disulfide and its derivatives

Compound	Application	Reference
Carbon disulfide	reaction and extraction solvent, catalyst presulfidation agent, oil well solvent, cold vulcanization additive for rubber, synthetic chemical	[33], [81]
Sodium cellulose xanthate	precursor for rayon, cellophane, and regenerated cellulose products; chelating agent	[17]
Alkyl xanthates	flotation agent, lubricant additive, analytical reagent	[17], [83]
Carbon tetrachloride	fluorocarbon refrigerant, degreasing and cleaning solvent	[20]
Trichloromethanesulfenyl chloride	intermediate for fungicide and pesticide manufacture	[90], [20], [125]
Sodium, potassium, and ammonium thiocyanate	fungicide, electrolysis and electroplating solution additive, corrosion inhibitor, photographic emulsion stabilizer, metal complexing agent, fabric dye enhancer, herbicide, insecticide, soil sterilant, intermediate for pharmaceuticals	[91]
Dithiocarbamates	fungicide, insecticide, soil fumigant, herbicide and defoliant, rubber vulcanization accelerator	[33], [92], [93]
Thiocarbanilide	rubber vulcanization accelerator	[94], [95]
Benzothiazoles	rubber vulcanization accelerator	[94], [95]
Thiazolsulfenamides	delayed rubber vulcanization accelerator	[95]

Historical production and price data in the United States are shown in Table 4. Competitive pressure from synthetic fibers and films will continue to limit market growth in the regenerated cellulose industry. Improvement in carbon disulfide usage will be slow and must come from increased use in agricultural and specialty chemicals manufacture.

Table 4. United States carbon disulfide production and price

Year	$/t	Production, 10^3 t/a
1965	94	345
1970	97	330
1975	145	220
1980	240	170
1984	420	210

World production capacity for carbon disulfide is shown in Table 5. Excess capacity will continue in the United States and Europe.

Table 5. World production capacity for carbon disulfide, 1000 t/a, 1984∗

Area	Charcoal process	Methane process
North America United States and Canada	–	335
Western Europe United Kingdom, France, Federal Republic of Germany, Belgium, Spain	–	360
Eastern Europe Poland, Rumania, Bulgaria, German Democratic Republic	40	260
USSR	40	120
Japan	–	100
India	50	–
China	20	–
Latin America Argentina, Brazil, Mexico, Columbia, Chile, Peru	60	15
Others Egypt, South Africa, Australia	15	–
Total	225	1190

∗ References [40], [96], and various trade sources

9. Toxicology and Occupational Health

The toxicology data base on carbon disulfide is extensive. Citations include hundreds of experimental investigations on animals and many clinical and epidemiological reports on humans.

Mutagenicity and Genotoxicity. The mutagenic and genotoxic potential of carbon disulfide has been evaluated in vitro and in vivo [97–100]. Carbon disulfide does not exhibit mutagenic activity in bacteria (*S. typhimurium* and *E. coli*) with or without the presence of an activation system. Additional in vitro tests, including host-mediated assay, unscheduled DNA synthesis in human fibroblasts, and primary cultures of human leukocytes, are inconclusive [101], [102].

However, the significance of these tests cannot be properly evaluated because of methodological problems including the lack of proper positive controls. Therefore, insufficient data are available to evaluate the mutagenic and genotoxic potential of carbon disulfide.

Human Health Effects. Carbon disulfide toxicity in humans is well-documented in the literature. It was first recognized as a workplace hazard by the French in the 1850s. Subsequent studies demonstrated effects on the central and peripheral nervous systems and cardiovascular system. Carbon disulfide can induce polyneuropathy and encephalopathy. Poisoning may be acute or chronic.

Symptoms of polyneuropathy include paresthesia, cramps, muscle weakness, pain, distal sensory loss, and neurophysiological impairment [103–110]. Exposure to high levels of carbon disulfide for prolonged periods is necessary to produce the polyneuropathy. However, exposure at levels of 65 – 170 ppm for less than 1 year are reported to result in many of the symptoms described [103]. *Symptoms of encephalopathy* may include headache, sleep disturbances, general fatigue, loss of libido, impotence, organic dementia, and Parkinsonism [103–110]. Onset of symptoms range from several months to several years. The early stage of encephalopathy is characterized by depression with sudden attacks of hallucinations. Visual disturbances that may result from high levels of carbon disulfide exposure include abnormal color vision, loss of visual acuity, loss of accommodation, and retinopathy [111].

Cardiovascular alterations in humans may also result from carbon disulfide exposure. A number of studies [112–114] reveal vascular atherosclerotic changes following long-term exposure to carbon disulfide (30 – 90 mg/m^3). Below 10 ppm, cardiovascular effects are not apparent.

The toxic effects of carbon disulfide require that precautions should be taken to limit exposure in the workplace. In the United States, the TLV-TWA as adopted by ACGIH for exposure to carbon disulfide is 30 mg/m^3, 10 ppm [115]; the MAK value is the same. The federal standard (OSHA, 1985) is 60 mg/m^3.

Disposition. The primary route of human carbon disulfide exposure is inhalation. Uptake in the blood is rapid and carbon disulfide is transported to tissues. Carbon disulfide possesses significant lipophilicity while also binding to proteins and amino acids. In humans, most ($>$ 70 %) of a systemically absorbed dose is rapidly converted to metabolites, followed by urinary excretion. Less than 30 % of carbon disulfide is exhaled and very small quantities are excreted unchanged [116].

Subchronic Toxicity. A 90-d inhalation toxicity study using Fischer 344 rats, Sprague-Dawley rats, and $B_6C_3F_1$ mice (exposed simultaneously in the same chambers) was conducted with carbon disulfide vapor [117]. Evaluation of the hematology examinations (Fischer 344 and Sprague-Dawley rats) indicated effects on red cells and platelets. In addition, high-dose animals of both sexes exhibited a mild increase in neutrophils and a decrease in lymphocytes. Serum chemistry examinations at the high dose did not reveal dramatic changes. Organ weight data revealed a depression of brain as well as liver weight.

Histopathology studies revealed axonal swelling of spinal cord nerve fibers in high-dose males, high-dose females, and possibly mid-dose males. In addition, males and females from the high-dose group had segmental degeneration of fibers in the sciatic nerve. Other lesions that oc-curred were incidental and not related to treatment.

Biochemical Mechanisms. Understanding the mechanism of action of carbon disulfide toxicity may lead to early clinical intervention in cases of suspected poisoning. Although a number of hypotheses have been proposed, the biochemical mechanism(s) remain unknown. A metabolite of carbon disulfide may be responsible for its hepatotoxic effects [117], [118]. Sulfur is liberated during conversion of carbon disulfide to carbon dioxide. The liberated sulfur may bind to cellular components or critical enzymes, giving rise to toxicity.

Research has also focused on related areas, including the role of carbon disulfide in chelating dithiocarbamates, binding and depleting essential trace elements, disturbing vitamin

metabolism [73], turnover of catecholamines [119], and metabolizing lipids [120–122].

10. Carbonyl Sulfide

Carbonyl sulfide [*463-58-1*], carbon oxysulfide, COS, is a colorless, flammable gas that is sparingly soluble in water but very soluble in alcoholic alkalies [123]. Selected *physical properties* are:

M_r	60.07
mp	− 138.8 °C
bp at 101.3 kPa	− 50.2 °C
Heat of vaporization at 101.3 kPa	18.50 kJ/mol
Vapor density at 101.3 kPa, 25 °C	2.485 kg/m^3
Critical temperature	105 °C
Critical pressure	6179.4 kPa
Heat of formation at 25 °C	− 142.1 kJ/mol
Free energy of formation at 25 °C	− 164.3 kJ/mol

The chemistry of carbonyl sulfide has been reviewed [14], [123], [124]. *Production* of carbonyl sulfide may be by the reaction of carbon monoxide with sulfur [123], reduction of sulfur dioxide with carbon [62], or hydrolysis of carbon disulfide [21]. Small amounts of carbonyl sulfide invariably are formed when carbonaceous fuel is pyrolyzed in the presence of oxygen, steam, and sulfur compounds [78]; the removal of carbonyl sulfide from these gas streams is the subject of many studies [78], [80].

Carbonyl sulfide is widely dispersed in the earth's atmosphere; concentrations between 4.7×10^{-4} and 5.3×10^{-4}ppm have been reported [74]. Carbonyl sulfide has been suggested as a major causative agent in atmospheric sulfur corrosion [74].

The *commercial importance* of carbonyl sulfide is limited; it is not produced in large quantity, and is used principally for small-scale syntheses and experiments. The use of carbonyl sulfide as a raw material for chemical manufacture has been proposed [126].

Carbonyl sulfide appears to play an important role in mediating the *toxicity* of carbon disulfide. Carbon disulfide is metabolized in vitro [127] and in vivo [128] to carbonyl sulfide. Carbonyl sulfide is further metabolized by hepatic carbonic anhydrase to hydrogen sulfide [128].

Death from acute inhalation of carbonyl sulfide can occur within 45 min [129] at concentrations of 20 000 ppm. A single intraperitoneal injection of 50 mg/kg resulted in 100 % mortality. Presently, complete subchronic toxicity studies on carbonyl sulfide are unavailable. There is one report, however, describing effects following subchronic administration. Rabbits [112] were continuously exposed for 7 weeks to ca. 50 ppm COS. Within five days, three rabbits died and two were sacrificed in a moribund condition. There were no further mortalities. Cholesterol (serum) elevations were noted in exposed rabbits. Histopathology did not reveal changes in the lungs or in the intima of the arteries in the organs evaluated.

11. References

1. G. Gattow, W. Behrendts: *Carbon Sulfides and Their Inorganic and Complex Chemistry,* **vol. 2,** Georg Thieme Publishers, Stuttgart 1977, pp. 8 – 58.
2. M. G. Zabetakis, *Bur. Mines Bull.* **627** (1965) 74.
3. *Handbook of Organic Industrial Solvents,* 5th ed., Alliance of American Insurers, Chicago, Ill., 1980, p. 66.
4. J. S. Gordon, *Chem. Eng. Data* **6** (1961) 390.
5. *Gmelin,* system no. 14, vol. D 4 (1977) pp. 31 – 197.
6. *International Critical Tables,* **vol. 2 (1927)** p. 174; **vol. 3 (1928)** pp. 231, 248, 287.
7. K. C. Mills: *Thermodynamic Data For Inorganic Sulfides, Selenides and Tellurides,* Butterworths, London 1974, pp. 181 – 184.
8. O. L. I. Brown, G. G. Manov, *J. Phys. Chem.* **59** (1937) 500.
9. D. A. Stull, H. Prophet, *JANAF Thermochemical Tables,* 2nd ed., NSRDS-NBS 37, U.S. Govt. Printing Office, Washington, D.C., 1971.
10. R. W. Gallant: *Physical Properties of Compounds,* **vol. 2,** Gulf Publishing Co., Houston, Tex., 1970, pp. 178 – 184.
11. R. N. Perry, D. Green (eds.): *Perry's Chemical Engineers' Book,* 6th ed., McGraw-Hill, New York 1984, section 3, pp. 247 – 259.
12. D. R. Stull, *Ind. Eng. Chem.* **39** (1947) 517.
13. L. J. O'Brien, W. J. Alford, *Ind. Eng. Chem.* **43** (1951) 506.
14. E. E. Reid: *Organic Chemistry of Bivalent Sulfur,* **vol. 4,** Chemical Publishing Co., New York 1961, pp. 422 – 445.

15. M. Yokoyama, T. Imamoto, *Synthesis* 1984, 797 – 824.
16. D. N. Jones (ed.): *Comprehensive Organic Chemistry,* vol. 3, Pergamon Press, New York 1979.
17. S. R. Rao: *Xanthates and Related Compounds,* Marcel Dekker, New York 1971.
18. G. D. Thorn, R. A. Ludwig: *The Dithiocarbamates and Related Compounds,* Elsevier, Amsterdam, New York 1962, p. 61.
19. A. D. Raskina, B. A. Suvovov, V. I. Zetkin, Z. M. Kolesnikov, E. V. Zakharov, *Zh. Prikl. Khim.* **45** (1972) 672.
20. J. S. Sconce: *Chlorine: Its Manufacture, Properties and Uses,* ACS Monograph Series No. 154, Reinhold Publ. Co., New York 1962, pp. 358 – 363.
21. R. F. Bacon, E. S. Boe, *Ind. Eng. Chem.* **37** (1945) 469.
22. Phillips Petroleum Co., US 3880933, 1975 (D. H. Kubicek).
23. D. R. Stull, *Ind. Eng. Chem.* **41** (1949) 1968.
24. J. Chao, *Hydrocarbon Process.* **59** (1980, Nov.) 217.
25. J. R. West, *Ind. Eng. Chem.* **42** (1950) 713.
26. B. John, Ph. D. Thesis, University College of Swansea, 1966.
27. R. A. Fisher, J. M. Smith, *Ind. Eng. Chem.* **42** (1950) 704.
28. H. O. Folkins, E. Miller, A. Henning, *Ind. Eng. Chem.* **42** (1950) 2202.
29. R. C. Forney, J. M. Smith, *Ind. Eng. Chem.* **43** (1951) 1841.
30. G. W. Nabor, J. M. Smith, *Ind. Eng. Chem.* **45** (1953) 1272.
31. C. M. Thacker, E. Miller, *Ind. Eng. Chem.* **36** (1944) 182.
32. O. Kausch: *Der Schwefelkohlenstoff,* Julius Springer, Berlin 1929.
33. J. F. Thorpe, M. A. Whitely: *Thorpe's Dictionary of Applied Chemistry,* 4th ed., **vol. 11,** Longmans' Green & Co., London 1938, pp. 328 – 344.
34. W. J. Bushell, *Chem. Ind. (London)* 1961 (Sept. 16) 1465.
35. *Sulphur* 1967, no. 76, 28.
36. H. N. Madon, R. F. Strickland-Constable, *Ind. Eng. Chem.* **50** (1958) 1189.
37. *Sulphur* 1980, no. 150, 32.
38. Shell Dev. Co., US 2187393, 1940 (M. De Simo).
39. Barium Reduction Corp., US 2474067, 1949 (L. Priesman).
40. *Sulphur* 1977, no. 139, 23.
41. Pure Oil, US 2330934, 1943 (C. M. Thacker).
42. Pure Oil, US 2411236, 1946 (C. M. Thacker).
43. Pure Oil, US 2536680, 1951 (H. O. Folkins, E. Miller).
44. Pure Oil, US 2565215, 1951 (H. O. Folkins, E. Miller).
45. Pure Oil, US 2568121, 1951 (H. O. Folkins, C. A. Porter, E. Miller, H. Hennig).
46. FMC, US 2616793, 1952 (H. O. Folkins, E. Miller).
47. FMC, US 2668752, 1954 (H. O. Folkins, E. Miller).
48. FMC, US 2709639, 1955 (H. O. Folkins, E. Miller, H. Hennig).
49. FMC, US 2857250, 1958 (R. W. Timmerman, A. G. Draeger, J. W. Getz).
50. FMC, US 2882130, 1959 (D. J. Porter).
51. FMC, US 2882131, 1959 (J. W. Getz, R. W. Timmerman).
52. FMC, US 3079233, 1963 (C. J. Wenzke).
53. FMC, US 3087788, 1963 (D. J. Porter).
54. FMC, US 3250595, 1966 (D. R. Olsen).
55. H. W. Haines, Jr., *Ind. Eng. Chem.* **55** (1963) 44.
56. C. M. Thacker, *Hydrocarbon Process.* **49** (1970) 124.
57. L. F. Albright, B. L. Crynes, W. H. Corcoran (eds.): *Pyrolysis: Theory and Industrial Practice,* Academic Press, New York 1983, pp. 368 – 375.
58. Phillips Petroleum Co., US 2468904, 1949 (C. R. Wagner).
59. Pan American Petroleum Corp., US 2788262, 1957 (W. A. Adcock, W. C. Lake).
60. A. J. Owen, K. W. Sykes, D. J. D. Thomas, *Trans. Faraday Soc.* **47** (1951) 419.
61. A. J. Owen, K. W. Sykes, D. J. D. Thomas, *Trans. Faraday Soc.* **49** (1953) 1198.
62. C. W. Siller, *Ind. Eng. Chem.* **40** (1948) 1227.
63. W. J. Thomas, R. F. Strickland-Constable, *Trans. Faraday Soc.* **52** (1956) 972.
64. W. J. Thomas, S. C. Naik, *Trans. Inst. Chem. Eng.* **48** (1970) nos. 4 – 6, 129.
65. C. M. Thacker, *Hydrocarbon Process.* **49** (1970) 137.
66. W. J. Thomas, B. John, *Trans. Inst. Chem. Eng.* **45** (1967) T119.
67. Rhône-Progil, US 3932595, 1976 (J. Bertheus, G. Schwahofer).
68. Andretech Corp., US 4397829, 1984 (Chem Wu Chi).
69. Rhône-Poulenc Industries, US 3978198, 1976 (B. Buathier, A. Combes, J. Jourdan).
70. W. A. Bryce, C. Hinshelwood, *J. Chem. Soc.* 1949, 3379.

71. Société Progil, US 3436181, 1969 (J. Berthoux, J. P. Guillet, G. Schneider).
72. Société Progil, US 3699215, 1972 (P. Gerin, L. Louat, J. P. Guillet).
73. T. O. Peyton, R. V. Steele, W. R. Mabey: "Carbon Disulfide, Carbonyl Sulfide: Literature Review and Environmental Assessment," Stanford Research Institute, EPA Report PB-257947, National Technical Information Service, Washington, D.C., 1976.
74. T. E. Graedel, G. W. Kammlott, J. P. Franey, *Science* **212** (1981) 663.
75. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., Van Nostrand Reinhold Co., New York 1983, p. 340.
76. J. Heicklen, W. P. Wood, K. J. Olszyna, E. Cehelnik: "Chemical Reactivity: Urban Atmosphere," *Proc. Symp. 1969,* Elsevier, New York 1971.
77. K. J. Olszyna, J. Heicklen, *J. Phys. Chem.* **74** (1970) 4188.
78. A. L. Kohl, F. C. Riesenfeld: *Gas Purification,* McGraw-Hill, New York 1974, pp. 701, 708.
79. R. J. Grant, M. Manes, S. B. Smith, *AIChE J.* **8** (1962) 403.
80. G. Astarita, D. W. Savage, A. Bisio: *Gas Treating with Chemical Solvents,* J. Wiley & Sons, New York 1976, p. 252.
81. Stauffer Chemical Co., CS2 Brochure, Westport, Conn., 1975.
82. *Reagent Chemicals, American Chemical Society Specifications,* 5th ed., Amer. Chem. Soc., Washington, D.C., 1974.
83. J. H. Karchmer (ed.): *The Analytical Chemistry of Sulfur Compounds,* J. Wiley & Sons, New York 1970, Part 1, pp. 353 – 369.
84. M. B. Jacobs: *The Analytical Toxicology of Industrial Inorganic Poisons,* Interscience, New York 1967, pp. 552 – 560.
85. "Carbon Disulfide," Chemical Data Sheet SD-12, Manufacturing Chemists Assoc., Washington, D.C., 1967.
86. "Unloading Flammable Liquids From Tank Cars," Manual TC-4, Manufacturing Chemists Assoc., Washington, D.C., 1967.
87. *National Fire Codes,* vol. 7, National Fire Protection Assoc., Quincy, Mass., 1985.
88. "Hazardous Materials Regulations of the Department of Transportation," ICC no. BOE-600-B, Bureau of Explosives, Washington, D.C., 1982.
89. *Kirk-Othmer* 4th ed., **vol 5** Wiley, New York, 1993.
90. Standard Oil Co., US 2553771, 1951 (S. A. R. Kittleson, C. Yowell, H. L. Yowell).
91. A. A. Newman (ed.): *Chemistry and Biochemistry of Thiocyanic Acid and Its Derivatives,* Academic Press, New York 1975, p. 222.
92. Stauffer Chemical Co., US 2766554, 1965 (S. C. Dorman, A. B. Lindquist).
93. J. Ashworth, G. A. Rodgers, G. G. Briggs, *Chem. Ind.* (1979, Feb.) 90.
94. S. C. Stinson, *Chem. Eng. News* 1983 (Apr. 25) 23.
95. A. Y. Coran, *Chemtech.* **8** (1983, Feb.) 105.
96. *Sulphur* 1985, no. 179, 24.
97. M. Donner, K. Flack, K. Hemminki, M. Sorsa, *Mutat. Res.* **91** (1981) 163 – 166.
98. A. Hedenstedt, U. Rannug, C. Ramel, C. A. Wachtmerster, *Mutat. Res.* **68** (1980) 313 – 325.
99. National Toxicology Program, National Toxicology Plan 79-9, vol. 26, Washington, D.C., 1979, p. 58.
100. Personal Communication from Dr. Errol Zeiger to *CRC Crit. Rev. Toxicol.,* Sept. 15, 1981.
101. R. P. Beliles, D. J. Brusick, F. J. Mecler, NIOSH Contract no. 210-77-0047, 1980.
102. S. Zajaczek, L. Solecka, *Med. Proc.* **28** (1977) no. 6, 461 – 465.
103. E. C. Vifliani, *Br. J. Ind. Med.* **11** (1954) 235 – 244.
104. R. Lillis, *Behav. Toxicol.* **51** (1974) 59.
105. A. Seppalainen, M. Tolenen, P. Karli, H. Hanninen et al., *Work, Environ., Health* **9** (1972) 71 – 75.
106. A. Seppalainen, M. Tolenen, *Work, Environ., Health* **11** (1974) 145 – 153.
107. C. Vasilescu, *Eur. Neurol.* **14** (1976) 447 – 457.
108. C. Vasilescu, A. Florescu, *J. Neurol.* **224** (1980) 59 – 64.
109. R. Gilioli, C. Bulgheroni, P. Betrazzi, A. M. Cula et al., *Med. Lav.* **69** (1978) 130 – 143.
110. B. Knare, B. Kolmodin-Hedman, H. C. Persson, J. M. Goldberg, *Work, Environ., Health* **11** (1974) 49.
111. A. Hamilton, H. L. Hardy: *Industrial Toxicology,* 4th ed., John Wright PSG, Lilliton, Mass., 1983, p. 575.
112. S. Hernberg, T. Partanen, C. H. Nordman, P. Sumari, *Br. J. Ind. Med.* **27** (1970) 313 – 325.
113. J. R. Teller, R. S. F. Schilling, J. N. Morris, *Br. Med. J.* **4** (1970) 407 – 411.
114. M. Tolonen, S. Hernberg, M. Nurminen, *Br. J. Ind. Med.* **32** (1975) 1 – 10.

115. ACGIH (ed.): *Threshold Limit Values (TLV)* , ACGIH, Cincinnati, Ohio, 1985 – 86.
116. Environmental Health Criteria 10, Carbon Disulfide, WHO, Geneva, 1979.
117. Chemical Industry Institute of Toxicology, Report 62063, Research Triangle Park, N.C., 1983.
118. E. J. Bond, F. DeMatteis, *Biochem. Pharmacol.* **18** (1969) 2531 – 2549.
119. F. DeMatteis, *Pharmacol. Ther., Part B* **2** (1978) 693 – 725.
120. L. Majos, *Rev. Environ. Health* **2** (1975) 65 – 80.
121. W. Laurman, T. Wronska-Nofer, *Med. Pr.* **28** (1977) 77 – 82.
122. L. Paterni, G. Pusic, S. Teodori, *Folia Med.* **41** (1958) 705 – 722.
123. R. J. Ferm, *Chem. Rev.* **57** (1957) 621.
124. *Gmelin,* system no. 14, vol. D 5, (1977) pp. 2 – 140.
125. G. Sosnovsky, *Chem. Rev.* **58** (1957) 509.
126. E. Field, E. A. Swakon: "Carbonyl Sulfide as a Chemical Building Block," *Amer. Chem. Soc. Symp. on New Sulfur Chemistry,* Los Angeles, Mar. 31, 1974.
127. R. R. Dalvi, R. E. Poore, R. A. Neal, *Life Sci.* **14** (1974) 1785.
128. C. P. Chengelis, R. A. Neal, *Toxicol. Appl. Pharmacol.* **55** (1980) 198 – 202.
129. E. Hayashi, *Oro Yakuri* **5** (1971) 435.

Carbon Fibers → Fibers, 5. Synthetic Inorganic

Carbonic Esters

HANS-JOSEF BUYSCH, Bayer AG, Krefeld, Federal Republic of Germany

1. Introduction 1
2. Properties 1
3. Production 2
3.1. Phosgenation of Alcohols 2
3.2. Phosgenation of Phenols 4
3.3. Oxidative Carbonylation of Alcohols 5
3.4. Oxidative Carbonylation of Phenols 7
3.5. Oxidative Carbonylation of Alcohols with NO_x 8
3.6. Carbonic Esters from Oxalic Esters 9
3.7. Electrochemical Oxidative Carbonylation of Alcohols 9
3.8. Reaction of CO_2 with Oxiranes . . . 9
3.9. Reactions of Alcohols with Urea . . . 9
3.10. Reaction of Phenols with Urea 10
3.11. Miscellaneous Reactions of CO_2 to Form Carbonates 10
3.12. Transesterification 11
3.13. Cyclic Carbonic Esters 12
4. Environmental Protection and Toxicology 13
5. Quality Specifications 14
6. Analysis 14
7. Storage and Transportation 14
8. Uses 14
8.1. Direct Uses 14
8.2. Use for Chemical Synthesis 15
9. Economic Aspects 17
10. References 18

1. Introduction

The monoesters of carbonic acid can only be isolated as salts, simple anhydrides, or mixed anhydrides with carboxylic acids [1]. In the past, they have aroused little industrial interest. On the other hand, the symmetrical and unsymmetrical diesters derived from aliphatic and aromatic hydroxy compounds are widely used, especially as intermediates and solvents.

2. Properties

The industrially important dialkyl carbonic esters (**1**, R = alkyl) are all colorless liquids, and some of them have pleasant odors. Cyclic aliphatic and diaryl carbonic esters are colorless crystalline compounds with relatively low melting points, whereas cyclic aromatic carbonates have high melting points, up to almost 400 °C.

1

2

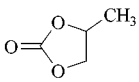

3

Most carbonic esters are insoluble or dimethyl- and diethylcarbonat sparingly soluble in water. Ethylene carbonate, 1,3-dioxolan-2-one **2** and propylene carbonate, 4-methyl-1,3-dioxolan-2-one **3** dissolve readily in water and are also soluble in organic solvents, particularly polar solvents. The other water-insoluble carbonates likewise dissolve in polar organic, solvents such as esters, ketones, alcohols, ethers, and substituted aromatic hydrocarbons.

The lower aliphatic carbonates form azeotropic mixtures with several organic solvents [2]. Dimethyl carbonate and ethylene carbonate form azeotropic mixtures with methanol and ethylene glycol, the alcohols from which they are respectively derived [3], [4].

Ethylene and propylene carbonates are extensively used as solvents for aromatic compounds, polymers, and saltlike compounds [5].

The physical properties of industrially important carbonic esters, including ethylene and propylene carbonate, are presented in Table 1. Properties of other cyclic carbonate, are listed in Table 2.

The ester structure and a strong tendency to form carbon dioxide play dominant roles in the chemistry of carbonic esters. Hydrolysis, which becomes more difficult with increasing molecular mass of the ester, leads to the formation of hydroxy compounds and carbon dioxide. Apart from massive steric effects, transesteri-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_197

Table 1. Properties of industrially important carbonic esters

Ester	*mp*, °C	*bp*, °C/*p*, kPa	d_4^{20}	n_D^{20}	Flash point, °C
Dimethyl carbonate [*616-38-6*]	4	90.2/101.3	1.073	1.3687	14 (closed cup)
Diethyl carbonate [*105-58-8*]	−43	125.8/101.3	0.9764	1.3843	33 (closed cup)
Diallyl carbonate [*15022-08-9*]		97/8.13		1.4280	
Diethylene glycol bis(allyl carbonate) [*142-22-3*]	− 4	160/0.27	1.143	1.4503	177 (open cup)
Diphenyl carbonate [*102-09-0*]	78.8	302/101.3	1.1215_4^{87}		168 (closed cup)
Ethylene carbonate [*96-49-1*]	39	248/101.3	1.3218_4^{39}	1.4158^{50}	
Propylene carbonate [*108-32-7*]	−48.8	242/101.3	1.2069_{20}^{20}	1.4189	

fication between carbonic esters and hydroxy compounds adheres to the following rule: the more nucleophilic hydroxy compound displaces the less nucleophilic compound, and if both hydroxy compounds have the same, or nearly the same, nucleophilicity, then the less volatile compound displaces the more volatile one. Hence, even with lower molecular mass alcohols, diaryl carbonates form dialkyl carbonates and the less nucleophilic phenols, but lower molecular mass dialkyl carbonates react with higher molecular mass alcohols to give the higher molecular mass dialkyl carbonates; the lower boiling alcohol is removed by distillation. In this case, transesterification occurs in steps with the intermediate formation of a mixed ester. The reaction of dialkyl carbonates with phenols to give diaryl carbonates, contrary to the above rule, can be achieved by using suitable conditions and catalysts [6]. It proceeds, however, at a relatively slow rate.

Carbonic esters react with primary and secondary amines to give urethanes and ureas [7]. At higher temperature and in the presence of catalysts, aliphatic carbonates lose carbon dioxide and act as alkylating agents, e.g., for amines [8] and phenols [9]. Rapid oxyethylation can be achieved in this way by using five-membered cyclic carbonates [10]. In the absence of compounds to be alkylated, these cyclic carbonates can be broken down to oxiranes and carbon dioxide, which is the reverse of the formation reaction [11]. Aliphatic and aromatic cyclic carbonate can easily be polymerized to a variety of homopolymers and copolymers with high molecular masses [12–15], [17–22]. Open-chain carbonic esters can be broken down catalytically at high temperatures to form alcohols, carbon dioxide, and olefins [23]. Dialkyl carbonates undergo Claisen condensation, e.g., with aliphatic carboxylic esters to form malonates [24].

3. Production

Carbonic esters have attracted increasing attention and been the focus of intensive research since ca. 1975 [25]. Figure 1 presents an overview of the various routes for producing carbonic esters and some important associated reactions. Carbon dioxide is the least expensive feedstock but also has the lowest reactivity, and it must therefore be transformed into more reactive compounds such as urea and cyclic carbonates. Although CO_2 can add to alcohols to form hemiesters, these are only present in low concentrations and, on account to their instability, their conversion to the diesters could not yet be achieved satisfactorily.

To avoid intermediate stages, the more expensive carbon monoxide is used. Either carbonic acid esters are formed by direct oxidative carbonylation of oxalic diesters which transform into carbonic esters by decarbonylation. Oxidation is performed with oxygen, if necessary via the nitric acid ester stage, or electrochemically. The latter technique, however, is still in its infancy.

For other methods of synthesizing carbonic esters, see [1], [26, p. 956, 964 – 966], [27].

3.1. Phosgenation of Alcohols

The phosgenation of hydroxy compounds is currently the most important method for producing

Table 2. Properties of cyclic carbonic esters

Ester		CAS no.	*mp*, °C	*bp*, °C/*p*, kPa
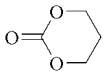	1,3-dioxan-2-one	[*31852-84-3*]	47 – 48	105/0.3
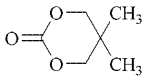	5,5-dimethyl-1,3-dioxan-2-one	[*29035-08-3*]	110	112/0.3
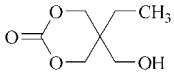	5-ethyl-5-(hydroximethyl)-1,3-dioxan-2-one	[*38802-97-0*]	41	ca. 180/0.1
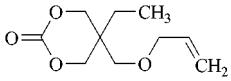	5-[(allyloxi)methyl]-5-ethyl-1,3-dioxan-2-one	[*3536-64-9*]		101 – 106/0.07
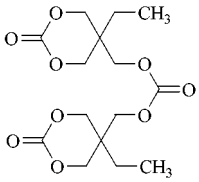	5,5′-[carbonyl bis-(oxymethylene)]-bis[5-ethyl-1,3-dioxan-2-one]	[*84056-44-0*]	248 – 254	
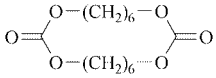	1,3,10,12-tetraoxo-cyclooctadecan-2,11-dione	[*82613-63-6*]	131	260/0.03
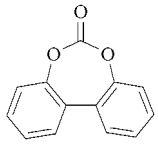	dibenzo [*d,f*]-[1,3] dioxepin-6-one	[*7623-38-3*]	99 – 100	
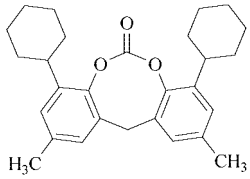	4,8-dicyclohexyl-2,10-dimethyl-12*H*-dibenzo [*d,g*] [1,3] dioxocin-6-one	[*133126-34-8*]	208	225 – 232/0.1
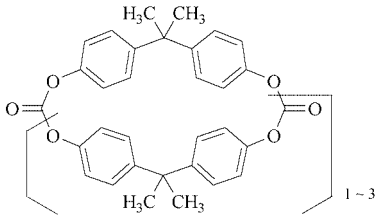	cyclic Bisphenol A oligocarbonate	$n = 1$ [*4437-79-0*] $n = 2$ [*811-43-8*] $n = 3$ [*21194-07-0*]	330 – 335 345 – 350 368 – 372	

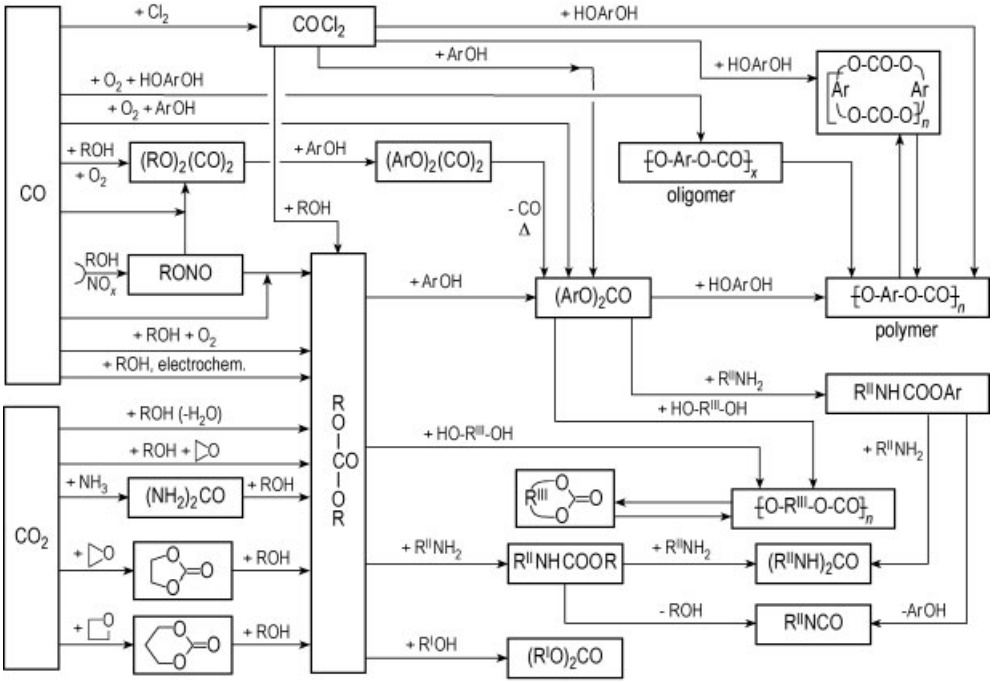

Figure 1. Synthetic routes to carbonic esters

carbonic esters [28]. Diethyl carbonate (DEC), one of the most frequently used aliphatic carbonates, is produced by passing gaseous phosgene in boiling ethanol, in a molar ratio of 1 part of phosgene to 2.5 – 4 parts of ethanol, into a glass or enamel apparatus consisting of a heatable reaction vessel with a reflux condenser and a distillation setup. The reaction gives essentially quantitative yields (≥ 99 %) and also forms hydrogen chloride. Excess ethanol is removed from the reaction mixture by distillation, and DEC is obtained as the residue.

To recover high purity (≥ 99.8 %) DEC that is free of chlorine (≤ 10 mg/kg), the crude product is distilled. If necessary, a base such as an alkali or alkaline earth metal carbonate, is used in the distillation. The yield based on phosgene is ≥ 95 %.

Dimethyl carbonate (DMC) is formed by the reaction of MeOH with phosgene or methyl chloroformate in the presence of a concentrated sodium hydroxide solution in a two-phase reaction in high yields and purity [29]. Other alcohols can also be phosgenated [30–32], [26, p. 953], [33–35]. As DMC is now more easily accessible via the direct oxidative carbonylation of MeOH, phosgenation is losing its attractiveness in this application.

3.2. Phosgenation of Phenols

Although diphenyl carbonate (DPC) produced by the oxidative carbonylation of phenol [36–38], the technology for this reaction has not yet been optimized. For this reason, the familiar [39–41] process of phosgenation of phenol has been subjected to renewed scrutiny with a view to improving the process and product quality (Cl-free). For example, phosgenation of aqueous sodium phenolate in the presence of DPC as solvent for newly formed DPC [25], [42] gave a high yield. Dissolving the phosgene in toluene before it enters the reaction vessel is said to increase safety, reduce costs, and minimize the Cl content of the DPC [43–45]. New purification methods have also been developed, e.g., rinsing with hot water and distillation [46].

From an early stage, there was interest in a homogeneously catalyzed phenol phosgenation process [47], [48] to avoid the formation of salts. The catalysts included salts of Al, Ti, Sn,

Zn, and Fe [48–52]; quaternary ammonium salts [53]; heterocyclic N bases [47]; and hydroxypyridines [54], [55]. The reaction rate (phosgene consumption) was highest at 130 – 150 °C [52], presumably because the increasing conversion of phosgene and its decreasing solubility with temperature are at an optimum in this range. Phosphorus compounds [56], [57] and higher alkylpyridines [58] make it possible to stop the phosgenation of phenol at the chloroformate stage. With other catalysts, such as certain salts of the alkali, rare earth, and alkaline earth metals, and metal oxides of Groups 2 and 12, phenyl chloroformate undergoes disproportionation to yiel DPC and phosgene. Further reaction with PhOH to yield DPC is also possible at higher temperatures, e.g., in the presence of phosphorus compounds [59], [60].

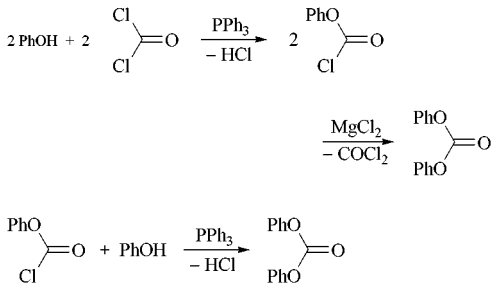

For this reason, a little more phenol is added at the end of the reaction to further lower the phenyl chloroformate content [45]. After rinsing with alkali and distillation, Cl-free DPC is obtained [61].

The need to separate homogeneous catalysts, usually by complex methods, led to a search for heterogeneous catalysts, including polymer-bonded pyridines [62] and phosphines [59], [60], activated carbon [63], metal salts on inert carriers [64], aluminum oxides [65], which had previously been considered inert, alumosilicates [66], silicates of various metals [67], and oxides of metals of Groups 2 – 5 [68], [69], ceramic precursors such as carbides, borides, nitrides [70], [71], and metallates of the metals of Groups 3 – 14 [70], [71].

Phenol phosgenation can be carried out in the liquid phase and in the gas phase [72] in a fully continuous reaction at 250 – 350 °C that gives almost complete phosgene conversion [73]. Complete phosgene conversion is also achieved by using an adiabatic procedure on a solid-bed catalyst, e.g., γ-Al_2O_3. The result is high-purity DPC in excellent yield and an electrolyzable hydrochloric acid [74–76], that is virtually no waste products (Fig. 2).

3.3. Oxidative Carbonylation of Alcohols

Homogeneous Catalysis. The reaction of alcohols with CO and oxygen to form dialkyl carbonates would seem both economically and ecologically desirable. This is reflected in the work on this reaction since the 1960s [77–82] using various CO-binding metal salts with an oxidative catalytic effect. The focus eventually concentrated on copper chlorides in the presence of various additives. Under certain conditions [83], [84–89] palladium compounds generate oxalic esters, sometimes as mixtures with carbonic esters, and are more difficult to reoxidize.

For this reason, work concentrated on the Cu system [90–93]. Non polar aprotic solvents [90], [94–98], N-bases, onium salts and alkali and alkaline earth metal salts [99–112] as well as additions of salts of Pd, of ferrous metals and the rare earths, complexing agents, and weak acids [113–115] are said to accelerate the reaction and improve the selectivity.

Some of the catalyst systems are highly complex, unstable, and generally difficult to separate from the reaction mixture. It therefore made sense to concentrate only on the effective Cu compounds. Using just a slurry of copper chloride [116], [117–120] in MeOH, it was possible to develop a workable DMC process with almost full selectivity for MeOH and 80 % selectivity for CO. ENI in Ravenna has used this process to produce DMC since 1983 [116–119] (Fig. 3). The plant was originally designed with an annual capacity of 5000 t and was expanded in 1993 to ca. 12 000 t [328].

The reaction follows this simplified redox cycle.

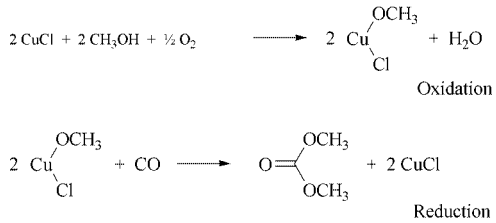

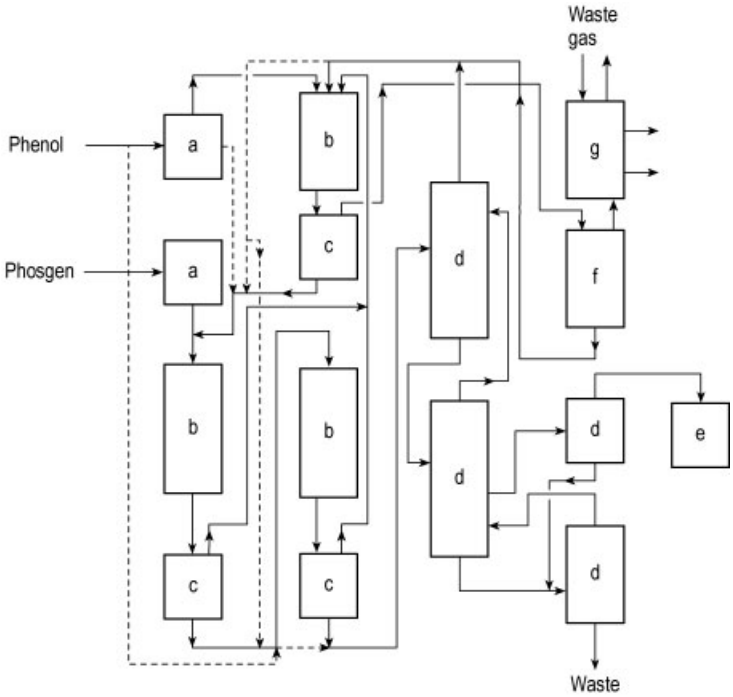

Figure 2. DPC by phosgenation of phenol [75]
a) Heat exchanger; b) Reactor; c) Degassing; d) Distillation column; e) DPC storage tank; f) Cold trap; g) HCl absorber.

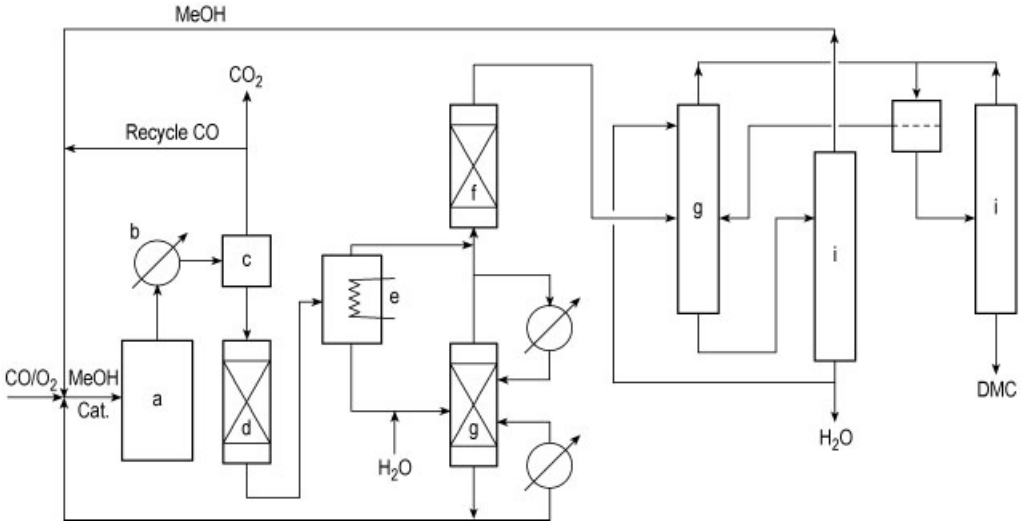

Figure 3. Oxidative Carbonylation of MeOH to DMC [328]
a) Reaction vessel; b) Heat exchanger; c) Gas separator; d) Ion exchanger; e) Evaporator; f) Aluminum oxide bed; g) Stripper; h) Extractive distillation; i) Distillation columns

The main by product is CO_2, along with small quantities of methyl chloride and dimethyl ether. Water in high concentrations disrupts the reaction as it displaces MeOH from the Cu complex. However, small quantities (0.5 – 1 mol per mole Cu salt) accelerate the formation of DMC [120]. The loss of chloride as HCl and methyl chloride is countered by the addition of small quantities of $CuCl_2$ [121], HCl [122], [123], halogens, or oxidants [124], [125]. The obtained azeotrope-forming mixtures of DMC, MeOH, and H_2O can be separated by extraction distillation, the addition of azeotrope-breaking liquids, pressure distillation, and membrane separation processes [126–128], [129–133].

The sparingly soluble catalyst sediments poorly and can only be isolated for reuse by filtration. However, in a CO atmosphere, it quickly sediments in an easily separable form [134]. The catalyst can also be left in the reaction vessel if the feedstock streams and pressures are adjusted so that DMC, MeOH, and H_2O are removed and the quantity and concentration of feedstock and products in the reaction vessel remain constant [135–138]. With > 30 % DMC in the liquid output, DMC recovery is easier if the DMC/MeOH azeotrope can flow back into the reaction vessel. The CO, used in surplus, is also returned to the reaction vessel and freed wholly or partially from byproducts such as HCl, CH_3Cl, and CO_2 [139], [140].

Heterogeneous Catalysis. Oxidative carbonylation on insoluble or supported catalysts, primarily in the gas phase, was first investigated with copper chloride complexes applied to activated carbon [141]. The inadequate selectivity and stability were improved by the addition of alkali and alkaline earth metal salts, and regeneration with HCl was proposed [142]. Dow attempted to develop a process on this basis [143], with supports such as aluminum oxides, silicates, and zeolites [144–146].

Other supports are based on carbon or metal oxides with complex ligands, and with the addition of salts of alkali and alkaline earth metals, bases, and borates [147–159]. This resulted in selectivities of up to 90 % for CO and 99 % for MeOH and more stable catalysts.

An interesting variant is the copper chloride-catalyzed oxidative carbonylation of alcohols in a Cu_2Cl_2/KCl melt (120 – 180 °C), performed continuously in a type of distillation column. The melt is continuously removed, dewatered, and then returned to the column [160–162]. Very good selectivity is achieved at low conversion rates. The apparatus must be made from corrosion-resistant materials.

In contrast to copper chlorides, whether in solution or in the molten state, active cobalt complexes are noncorrosive and less sensitive to water. Ligands include, e.g., Schiff bases with salicylaldehydes, oximes, pyridines, phenanthroline, acetyl acetone, and oxyquinoline, and solvents are, e.g., amides, nitriles, sulfones, sulfoxides, carbamates, and ureas [163–165]. The catalysts can be applied to supports and used in the gas phase [166]. However, ligands which are stable to oxidation are as yet unavailable [167].

3.4. Oxidative Carbonylation of Phenols

Homogeneous Catalysis. Analogous to the oxidative carbonylation of alcohols, diphenyl carbonate (DPC) can be synthesized from PhOH and Pd compounds [36–38]. The reaction can be steered to yield DPC or phenyl salicylate [171–173]. Water has a far more disruptive effect here than in MeOH carbonylation and must be quickly rendered harmless by using molecular sieves [36], [168–170]. Important to the smooth process of DPC formation are N bases, especially sterically hindered bases, complex ligands for Pd, and suitable redox-active metal complexes, especially of Mn and Co [36–38], [168–178].

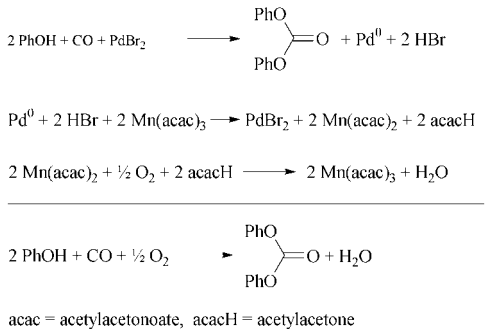

However, it is also claimed that it is possible to form DPC in the absence of bases [179], [180]. Instead of redox-active metals, iodides

[180] and benzophenones and quinones as cocatalysts [179], [181–183] are recommended. A catalyst system with a Co salicylaldimine complex, Pd acetate, terpyridine, hexaalkylguanidinium bromide, and an aprotic polar solvent such as N-methylpyrrolidone or DPC [184–186] accelerates the reaction while maintaining a selectivity of around 98 % with DPC contents of around 20 – 30 %. The selectivity decreases outside this range. Rare earth metal compounds, hydroxypyridines, and alkali or alkaline earth metal halides in aprotic polar solvents are also claimed to yield high selectivity [187], [188]. To ensure reproducible results, the Pd catalyst must be pretreated and activated with CO [189].

Instead of using molecular sieves, which require laborious regeneration, the reaction water can be flushed out efficiently with excess CO, e.g., in serially arranged bubble columns or by continuous withdrawal, dewatering at reduced pressure, and recycle of a partial stream of the reaction mixture [190–192]. That all can be achieved at a Pd content of only around 260 ppm in the reaction mixture. If gas throughput, total pressure and partial pressures of the reaction gases are kept constant, the space – time yield increases [193], [194].

The following byproducts of oxidative carbonylation have been identified: salicylate [171–173], *o*-phenylene carbonate and bromophenols [195], and phenoxyphenol [196]. The addition of Co compounds is claimed to completely suppress the bromination of PhOH [197].

Heterogeneous Catalysis. If the oxidative carbonylation of phenol is to be brought to technical maturity, a simple and continuous mode of operation and work-up, full catalyst recovery and regeneration, a heterogeneous catalyst, and efficient DPC purification must be developed. Various proposals have been made.

The reaction can be performed continuously in a multiplate rectification column [198] with side feeds to several reaction vessels that are also supplied with reaction gas. The reaction mixture is then returned to column at a point above which it was removed [199]. A bubble column cascade also allows continuous operation [190–192].

The DPC can be removed from the reaction mixture by vacuum distillation [200], but this will results in thermal damage to the catalyst system. However, the use of crystallization processes, perhaps in combination with distillation, minimizes this damage, and the catalyst can be recovered in a usable form [201–206].

Another mild method is to extract DPC from the reaction mixture with certain solvents [207]; the catalyst remains in the phenolic phase. Both this method and crystallization are benefited by a high DPC content in the reaction mixture [208]. The deactivated catalyst can be regenerated by oxidation [209].

It is difficult to convert complex systems like the ones used here into efficient heterogeneous catalysts. For this reason, work began with supported Pd [210], [211]. The addition of CsBr or nitriles to the reaction mixture had a favorable effect. Pd and Cu compounds with hydroxypyridine on activated carbon gave high selectivity for CO [212]. In other cases, activated carbon was simply added to the system, and considerable DPC yields were achieved [213], [214].

Insoluble oxides, nitrides, and carbides act as promoters [215]. It is possible to fix Pd and cocatalysts on supports as highly active products with long-term stability and thus develop a continuous process of DPC synthesis including DPC isolation and the recovery of the mother liquor [216–219].

3.5. Oxidative Carbonylation of Alcohols with NO_x

Palladium compounds, valued because of the mild carbonylation conditions, are easier to reoxidize with nitrites than with O_2 and alkyl nitrites are simple to manufacture. Hence, this route was investigated for DMC synthesis [220–223]; the reaction is ideally performed with heterogeneous Pd catalyst in the gas phase. Depending on the reaction conditions and the catalyst used, carbonates and/or oxalates are formed [224], [225].

To improve the selectivity for DMC, various catalyst supports such as aluminum oxides, metal phosphates and, in particular, zeolites [226–231] were investigated. These were doped with heavy metal, alkali and alkaline earth, and nonmetal compounds [232–234] and special ligands for Pd [235–237]. Chlorinated compounds are used to regenerate the catalysts [238–241]. The resulting systems have outstanding stability and selectivity, with particular interest attracted

by the Pd/Cu catalysts [242–245]. Concepts for a complete process were developed [246–251]. At the end of 1992, Ube started a DMC plant with a capacity of 3000 t/a using this process. Expansion to several tens of thousands of tons was planned [252].

$$CO + 2\,CH_3ONO \longrightarrow CH_3OC(=O)OCH_3 + 2\,NO$$

$$2\,CH_3OH + 2\,NO + 1/2\,O_2 \longrightarrow 2\,CH_3ONO + H_2O$$

3.6. Carbonic Esters from Oxalic Esters

Alcohols react with CO and O_2 or alkyl nitrites under certain conditions [84–89], [224], [225], [253–255] with highly selective formation of dialkyl oxalates. It would therefore seen obvious to exploit this CO fixation in carbonate synthesis, namely, by transesterifying dimethyl oxalate with PhOH to form diphenyl oxalate, followed by decarbonylation.

$$2\,CO + 2\,CH_3ONO \longrightarrow CH_3OCOCOOCH_3 + 2\,NO$$

$$CH_3OCOCOOCH_3\ 2\,PhOH \rightleftarrows PhOCOCOOPh + 2\,CH_3OH$$

$$PhOCOCOOPh \longrightarrow PhOCOOPh + CO$$

This transesterification requires conditions and catalysts similar to the manufacture of DPC from DMC and PhOH [256]. Diphenyl oxalate can be isolated and purified via a 1 : 2 adduct with PhOH [257]. The decarbonylation of diphenyl oxalate takes place at high temperatures and is catalyzed by heavy metal, alkali, alkaline earth, rare earth, and nonmetal compounds from Group 15, especially P compounds [258–265], with selectivities of up to 99 % and DPC yields of up to 96 %. Decarbonylation can also be carried out in the gas phase [266–268]. Concepts have been developed for transesterification, decarbonylation, and an overall process [269–271].

3.7. Electrochemical Oxidative Carbonylation of Alcohols

The oxidative carbonylation of alcohols is also possible at the anode of a suitable electrochemical apparatus [272–275]. High current efficiency can be attained but nothing is known about the cost – benefit ratio of this process.

3.8. Reaction of CO_2 with Oxiranes

The catalytic reaction of CO_2 with oxiranes directly yields five-membered cyclic carbonates (dioxolanones). Transesterification with alcohols allows the synthesis of other carbonates from inexpensive CO_2. The byproducts are 1,2-diglycols, especially ethylene glycol (EG), which is produced in enormous quantities worldwide by hydrolyzing ethylene oxide (EO) [16]. This process provides EO and EG producers with economical access to carbonic acid derivatives by using CO_2 from EO synthesis. The reaction of CO_2 with oxiranes, especially EO and propylene oxide, takes place under pressure with catalysts such as alkali metal iodides, ammonium and phosphonium salts, Lewis acids, heavy metal complexes, organometallic compounds, and ion exchangers with ammonium or phosphonium groups [276–281], [26], [282], [283]. Under milder conditions, the reaction takes place in the presence of bifunctional catalysts such as pairs of alkali metal and zinc halides [284], [285] and affords very high selectivity. A new process has been developed using this type of catalyst in which EO and CO_2 in EC as the reaction medium are converted under mild adiabatic conditions to EC in high yield. After purification, the quality of the product is excellent [284–289] (Fig. 4).

The combination of CO_2 and EO to form EC is reversible in the presence of the above-mentioned catalysts and others [290].

Like EO, other oxiranes can be converted into dioxolanones by reaction with CO_2 [291–293].

3.9. Reactions of Alcohols with Urea

At elevated temperatures, ammonia is gradually displaced from urea by alcohols, initially forming carbamates and then carbonates [294–296]. Depending on the temperature and the duration of the reaction, the two steps can be performed separately and both carbamates and carbonates obtained in high yields (98 %) [297–299]. Both

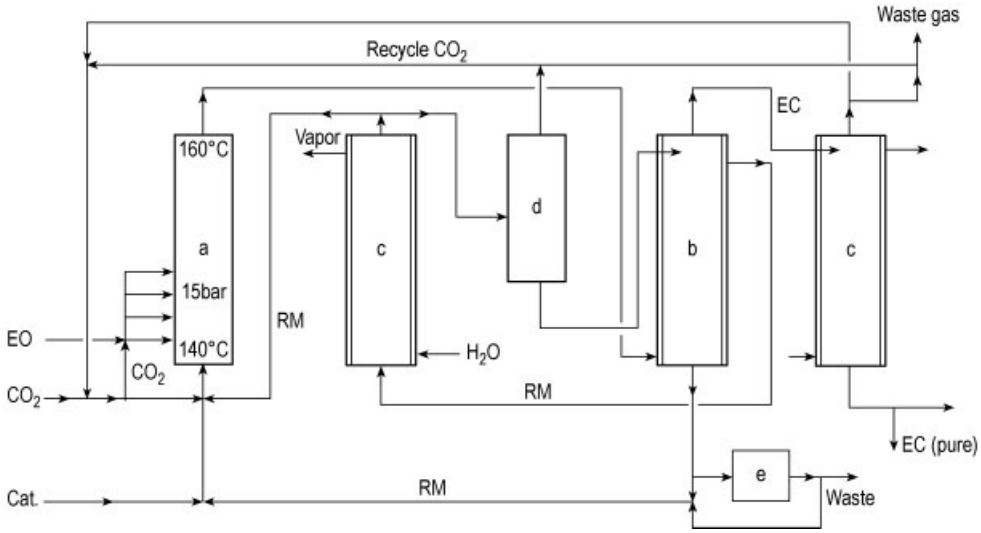

Figure 4. Production of EC from EO and CO_2 [286]
a) Reaction vessel, adiabatic; b) High-speed evaporator with heating jacket; c) Heat exchanger for steam generation; d) Depressurizing; e) Catalyst regeneration.
RM = Reaction mixture

reactions are catalytic. Nitrogen bases and salts, oxides, hydroxides, and alkoxides of alkali, alkaline earth, and many heavy metals, especially Sn, have been suggested as catalysts. High-boiling solvents are claimed to accelerate the reaction [298], [299]. Unfortunately, the reaction is virtually limited to higher alcohols, yet this route can be used to produce DMC by first reacting 1,2- or 1,3-diol with urea to form a cyclic carbonate (e.g., EC), followed by treatment with MeOH to form DMC [300], [301]. In both these steps, the selectivity is 95 – 99 %. As this synthesis always yields large quantities of ammonia as byproduct, it should preferably be located near a urea manufacturing facility.

3.10. Reaction of Phenols with Urea

The reaction of PhOH with urea generates NH_3 but little DPC [302]. Cyanic acid, formed above 130 °C, is quickly taken up by nucleophilic alcohols to form carbamates. However, this does not apply to the weakly nucleophilic PhOH and the cleavage of urea results in the usual derivatives of cyanic acid. To obtain DPC, the PhOH must be treated with urea intermediates such as DMC or urea derivatives of less nucleophilic and volatile amines than NH_3. Thus, diphenyl urea yields diphenyl carbamate and DPC [303].

3.11. Miscellaneous Reactions of CO_2 to Form Carbonates

Although technically interesting, less work has been done on the reaction of CO_2 with olefins to form dioxolanones [304–306] in the presence of oxidizing metal ions such as $Au^{(III)}$ and $Tl^{(III)}$, and on the addition of CO_2 to prop-1-yn-3-ols to form 5-methylene-1,3-dioxolan-2-ones [307], i.e., five-membered cyclic carbonates, which can be transesterified with MeOH to yield DMC, and on the reaction of CO_2 with alcohols and their acetals, ketals or orthoesters [308–310] in the presence of Sn compounds, whereby acetals, ketals, ortho esters, and molecular sieves act as water scavengers and convert them into aldehydes, ketones, and esters that can be reused. In this context, EO could also be termed a water scavenger when treated with CO_2 and, e.g., MeOH, to form EG and DMC [311], [312].

In the presence of certain Cd, Mg, and particularly Zn catalysts CO_2 undergoes alternating copolymerization with cyclic ethers, especially EO, to give polycarbonates [313]. Carbon diox-

ide is therefore becoming increasingly attractive in industrial chemistry [314], [315].

3.12. Transesterification

Since transesterification is an equilibrium reaction, it is possible in principle to convert one carbonate into another. However, in practice, this may prove difficult, require special conditions, or even fail.

Transesterification of Ethylene and Propylene Carbonates with Alcohols. Transesterification with EC runs particularly smoothly and quickly [316], [317]. Because of the significance of DMC, much work has been carried out on its synthesis, and the results have also been applied to carbonates of other alcohols. A wide variety of catalysts have been proposed [318]: alkali and alkaline earth metals as salts, oxides, and bound to ion exchangers [319–326], heavy metal compounds [327–334] rare earth compounds [335], [336]; N and P bases, their quaternary salts, and anion exchangers based thereon with various cations, especially those of weak acids [337–345]; derivatives of Ti, Zr, and Sn [326], [346], [347]; and catalysts consisting of two metal salts or of a metal salt and an onium salt which are highly effective in catalyzing the formation of EC from EO and CO_2 as well as transesterification [284], [285].

The course of the reaction is strongly dependent on the choice of catalyst. In the transesterification with Dowex MSA-1, an anion exchanger in chloride form with ammonium groups, the DMC yield is 99 % at 48 % conversion. Amberlyst A-21, a very similar exchanger, gives a yield of only 92 % and 18 % conversion under otherwise identical conditions [348]. Another anion exchanger in carbonate form gave 45 % conversion and $>$ 99 % yield at a water content of less than 160 ppm, but only 5 % conversion with the same yield at a water content of 200 ppm [349].

The technical environment for transesterification also influences its cost-effectiveness. If EC and MeOH are reacted together in cocurrent flow on a fixed bed catalyst [350], [351] conversion at best corresponds to the equilibrium. Therefore a difficult separation of feedstock and products by distillation is required, for both EG with EC and DMC with MeOH form azeotropes which must be broken [3], [4], [126–133], [352–354]. Only great care in catalyst manufacture prevents leaching of the heterogeneous catalysts. Otherwise, the resulting contamination must be removed before distillation [355], [356]. Transesterification and separation can be carried out simultaneously in distillation columns. A heterogeneous catalyst is placed on the plates, and the dwell time in the column is extended by removing a sidestream and passing it through a reaction vessel. In the case of incomplete conversion, water is fed into the bottom of the column to hydrolyze the residual EC [357–359]. If the catalyst activity and the process parameters are properly balanced, counterflow transesterification in a column yields a distillate that contains more DMC than the MeOH/DMC azeotrope at the top of the column and a virtually EC-free EG at the bottom [360], [361].

Dimethyl carbonate can also be produced by other transesterification reactions, e.g., by methanolysis of Bisphenol A polycarbonate (recycling) [362].

Transesterification of Phenols with Aliphatic Carbonates. For the entirely phosgene-free manufacture of Bisphenol A polycarbonate, the DPC must also be produced in a similar way. This can be achieved by the transesterification of DMC as, under transesterification conditions, EC oxyethylates PhOH with decarboxylation. However, the reaction of DMC with PhOH is slow and, because of the weak nucleophilicity of PhOH, the equilibrium is almost entirely on the DMC side. It is therefore necessary to ensure that the equilibrium is achieved as quickly as possible by the use of very active catalysts with high selectivity and that the equilibrium is quickly shifted by suitable modes of operation. The reaction takes place in three catalytic steps:

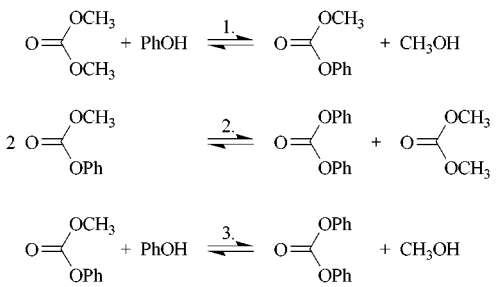

Steps 1 and 3 are slower than step 2. However, it would seem more economical to favor step 3 over step 2, as the DMC already bound to the PhOH is otherwise lost again. There is only limited potential to accelerate the reaction by the application of higher temperatures as, from around 200 °C, etherification (anisole) and carboxylation (salicylic acid derivatives) reduce the selectivity. For this reason, the reaction conditions and the catalysts used must be balanced accordingly [6], [363], [364]. The most important catalysts are Lewis acids and complexes of Ti, Zr, V, Zn, Fe, and Sn [365–367], Pb compounds [368], [369], alkali and alkaline earth salts of silicic, boric, phosphoric, and orthotungstic acids [370], alkyl tin compounds and titanates [363], [364], [371–377], some of which develop excellent activity and selectivity. Oxides, mixed oxides, silicates, and phosphates of numerous metals [363], [364], [378–380] have been less widely investigated.

Due to the unfavorable kinetic and thermodynamic conditions, the transesterification of DMC to DPC requires the use of complex process technology. For example, the use of a continuous distillation column has been recommended, whereby DMC and MeOH are removed at the top, and the higher-boiling products at the bottom for further work-up [381], [382] as have a system of several serially arranged reaction vessels with a column setup where DMC is fed in at the base of the reaction vessels [383–385], and a combination of two continuous distillation columns, in which the second column receives the bottom product of the first, and the first column the head product of the second [386]. Here, one or both columns can be linked to reaction vessels by sidestream removal and return feeds [387–390]. To decrease the DMC content of of the distillate below that the MeOH/DMC azeotrope, the distillate containing the azeotrope can be passed a number of times in counterflow through a column to yield high-percentage MeOH [391]. In the case of the two-column system, the distillate containing DMC from the first column is passed into the second, and a high yield of phenyl methyl carbonate is obtained at the bottom of the second [392], and is then converted to DPC by disproportionation (Fig. 5).

However, the transesterification can be steered such that almost pure DPC is discharged at the bottom of the second column [394]. The same applies for a combination of a counterflow and a reactive distillation column, and a good DMC conversion is achieved [395].

Transesterification in bubble column cascades yields an outstanding space – time yield [396], [397].

Further steps are required to separate the catalysts [398], [399] and the byproducts from DPC [201–206], [208], [400–402].

Where acetic anhydride or ketene is available, the transesterification process can follow the sequence:

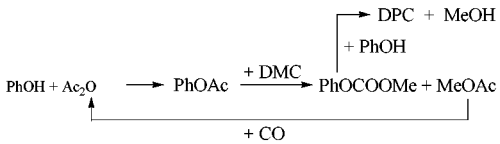

A number of variations are possible [403–409] which can be applied to bis-phenols to yield aromatic polycarbonates [410–413].

The direct transesterification of DMC with bis-phenols to form bis-phenol monomethylcarbonate and bis-phenol dimethylcarbonate or polycarbonate is more difficult [414–417].

DPC is also produced in a good yield by the phenolysis of aromatic polycarbonates (recycling) [418–420].

3.13. Cyclic Carbonic Esters

Cyclic carbonates can be produced in similar ways to the open-chain carbonates, i.e., by phosgenation [421–423], oxidative carbonylation [424] or transesterification [425–432] of dihydroxy compounds with, e.g., DMC, and also from urea [433–435]. However, these have the peculiarity that the main reaction product is usually a polymer or a mixture of monomers, oligomers, and polymers from which the monomers must be recovered by additional steps, usually by a combination of thermal degradation and distillation [436–438].

All aliphatic dihydroxy compounds react rapidly with phosgene to form bis-chloroformates, which with another diol in the presence of bases or at elevated temperatures are converted to the corresponding carbonates with cleavage of HCl. As working with bases pro-

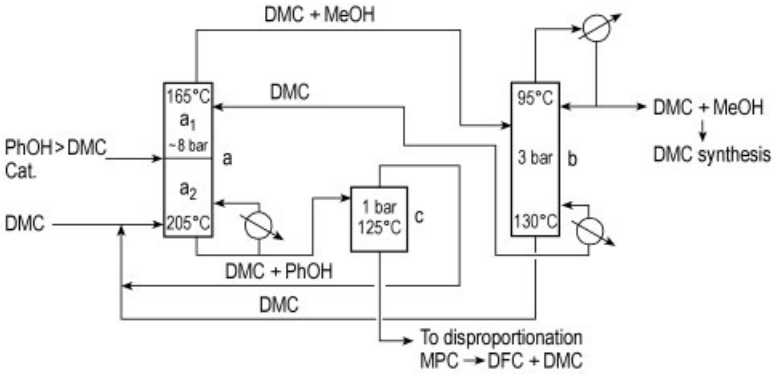

Figure 5. PhOH/DMC transesterification to methyl phenyl carbonate (MPC) [392]
a) Reaction distillation; a1) Reaction; a2) Rectification; b) Rectification; c) Flash distillation

duces large quantities of salts, the latter method is preferred [423], although small amounts of chlorinated byproducts are formed which are sometimes difficult to remove. For this reason, transesterification is the method of choice for producing high-purity carbonates [439]. Until now, oxidative carbonylation has been of little significance here. Transesterification can be carried out with aliphatic carbonates, especially DMC and DEC, but also with EC and DPC, the latter being recommended if working with diols having high melting points and at higher reaction temperatures as the melts are otherwise too viscous [440].

This process is used to manufacture cyclic carbonates from, e.g., 1,3-propanediol [425–435], 1,4-butanediol [425–432], 1,6-hexanediol [433–435], diethylene glycol [425–432], triethylene glycol [425–432], neopentyl glycol [423], [425–435, 439], pentaerythrite [425–432], trimethylol propane [425–432], [440], glycerin [425–432].

Cyclic aromatic carbonates can be produced above all by phosgenation and transesterification of, e.g., Bisphenol A [441–445], *o,o′*-dihydroxybiphenyl [446], 2,2′-dihydroxy-1,1′-bisnaphthalene [447], bis(2-hydroxy-3-cycloalkyl-5-alkylphenyl) alkanes [448], bis(2,4-dihydroxy-3-methylphenyl)methane [449], various tetraphenols [449–452].

4. Environmental Protection and Toxicology [453–456]

The safety precautions taken when working with phosgene must be taken during the normal production of carbonic esters. The exhaust gases are subjected to hydrolysis and then burned. Wastewater is freed of all undesired substances, e.g., phenol, by extraction and desorption.

Although carbonic esters are not generally considered dangerous substances, they must not enter the atmosphere or the water supply. Vapors must be removed by adequate exhaust systems, and spilled liquids must be absorbed onto an appropriate material, which is subsequently burned.

All important carbonic esters are flammable substances. In particular, the short-chain, aliphatic esters can be easily ignited because of their low flash points. Carbonic esters form explosive mixtures with air.

Handling precautions:

Eyes:	Safety goggles or face shield
Skin:	Protective clothing and gloves
Inhalation:	Adequate and effective ventilation
Ingestion:	No eating, drinking, smoking during handling

First aid:

Eyes:	Rinsing with plenty of water for least 10 min, consult physician
Skin:	Removal of contaminated clothing, washing with plenty of water, consult physician
Inhalation:	Move to fresh air, call physician
Ingestion:	Vomiting if conscious, immediately to hospital or physician

Toxicological Data.

Diethyl carbonate

LD_{50} (s.c.)	8500 mg/kg (rat)
LD_{50} (oral)	> 5000 mg/kg (rat)
TDL_0	500 mg/kg (mouse)
LC_{50} (inhalation)	> 21 mg/kg (mouse)

No irritation after contact with skin or mucous membranes (rabbit).

Dimethyl carbonate

LD_{50} (oral)	12.9 g/kg (rat)
LD_{50} (intraperitoneal)	1600 mg/kg (rat)
LD_{50} (cutaneous)	2500 mg/kg (rat)
LC_{50} (inhalation)	140 mg/L (4 h, rat)

Irritation possible after contact with skin and especially eyes.

Diphenyl carbonate

LD_{50} (cutaneous)	> 500 mg/kg (rat)
TDL_0 (oral)	28 g/kg (mouse)
TDL_0 (s.c.)	1000 mg/kg (mouse)

Prolonged and repeated skin contact can cause irritation and dermatitis. Prolonged inhalation of dust can irritate and affect the respiratory tract.

Ethylene carbonate

LD_{50} (oral)	10 g/kg (rat)

Inhalation of vapors (8 h, rats) does not cause death.

Propylene carbonate

LD_{50} (oral)	29 g/kg (rat)

Diglycol bis(diallyl carbonate)

LD_{50} (oral)	349 mg/kg (rat)
LD_{50} (cutaneous)	3038 g/kg (rabbit)
LC_{50} (inhalation)	> 0.73 mg/L (1 h, rat)

5. Quality Specifications

Carbonic esters are isolated in pure form by distillation. Dialkyl carbonates of ≥ 99.6 % purity are obtained as colorless liquids, free of water and containing less than 0.05 % Cl. DPC of 99.9 % purity, recovered in the molten state, has a color index of 20 – 30 (APHA) and a maximum phenol content of 0.05 %. If it is produced by transesterification of PhOH and DMC, there may be traces (≤ 0.1 %) of phenyl methyl carbonate.

Diethylene glycol bis(allyl carbonate) is more than 94 % pure, contains ≤ 1 % volatiles ($< 150\,°C$/5 mm Hg), has a color index of ≤ 30 (APHA), and a viscosity at 25 °C of $\leq 22 \times 10^{-6}\ m^2/s$.

6. Analysis

Gas chromatography is the simplest and most efficient method of analysis applicable to most impurities in carbonic esters. Titrations are used to detect water and traces of hydrogen chloride. Chloroformate can be detected by titration or gas chromatography [457].

7. Storage and Transportation

Because carbonic esters are not corrosive they can be stored, in the absence of moisture, in stainless steel tanks for at least one year. DPC can be stored in the molten state at 150 °C for approximately one month. However, allyl carbonates can only be stored for a limited time because of their tendency to polymerize. The permissible period of storage depends primarily on the temperature. Dialkyl carbonates are transported in steel barrels or tank cars. DPC is delivered as flakes in polyethylene sacks or in the molten state in tank cars.

8. Uses

8.1. Direct Uses

All of the main simple carbonates – DMC, DEC, EC, PC and even DPC in the molten state – are excellent solvents for a variety of substances such as cellulose ethers and esters [26], [454–456], pharmaceutical and cosmetic preparations [458–465], natural and synthetic resins and polymers [5], and salts. They serve as extractants [466–469], plasticizers, spinning dopes for

synthetic fibers, electrolytes [466–475], and additives for cleansing agents [476]. The addition of DMC improves the phenol yield when hydrocracking tar [477]. It also has a positive effect on fuel combustion [478–481] and as a foaming agent [482], [483]. Higher carbonates and cyclic aliphatic carbonates are used in the formulation of flavors and fragrances [484–486], and certain cyclic aromatic carbonates for thermal recording materials [487]. Cyclic carbonates are also used as reactive thinners in RIM polyurethanes [488], as activators in polyurethane coatings [489], epoxy [490], [491] and phenol – formaldehyde resins [492], and in silicate-based binders for molding sand [493]. Further applications are described in [466–469, 494, 495].

8.2. Use for Chemical Synthesis [1], [26], [496–501]

Transesterification. The transesterification of EC or DMC with aliphatic diols can be turned to aliphatic polycarbonates with certain molecular masses and OH numbers for the manufacture of polyurethanes [502–504]. Oligocarbonates based on $C_{12} - C_{44}$ alcohols and triethylene glycol yield high-quality lubricants and additives [505–508]. The carbonates of oximes or cholines with DPC are used as intermediates for bleaching and cleansing agents, and those of specially substituted phenols and alkynols for crop protection products [509–514], [515], [516].

Unsymmetrical carbonates can be produced from alcohols and symmetrical carbonates or from two different symmetrical carbonates [517–519], and monomethylcarbonates of dihydroxy compounds [520], [521] from DMC and the corresponding diols. Because of the outstanding properties, particularly of optical plastics made from allyl and vinyl carbonates, these are produced in various forms with particularly high purity by the transesterification of diols, allyl alcohol, and vinyl esters with DMC [393], [522–529].

Of major significance is the transesterification of bis-phenols, especially Bisphenol A, with DPC to form high-quality thermoplastic aromatic polycarbonates [530], [531]. Attempts have also been made to produce these polycarbonates with DMC or DEC as the CO donor, either by direct transesterification [532–534], stepwise by first synthesizing a Bisphenol A bis(methyl carbonate) and then reacting this with itself [535–537] to split off DMC, or by reaction of Bisphenol A diesters such as BPA diacetate with DMC, splitting off methyl acetate [538], [539]. Polycarbonate synthesis is achieved in all these cases, but none have gained industrial significance, presumably, secondary reactions result in unsatisfactory reproducibility.

The hydrolysis of EC produces a very pure EG in almost quantitative yield [277]. By comparison, EO yields 82 – 85 % EG and also polyglycols.

Alkylation. Nucleophilic substances are alkylated by DMC with CO_2 elimination. These substances include phenols, thiols [9], [540–548], amines [8], [549], [550], amides and urethanes [551], [552], alcohols [553–555], even if these contain trialkylamino groups (i.e., usually, under similar conditions, the trialkylamino groups are quaternized and/or split off, but not in this case) [556], [557], and carboxylic acids and their anhydrides [558–562]. In all cases, the corresponding carbonates are produced, often as isolable intermediates, which split off CO_2 under more extreme conditions. In this way, DMC can be used to form methyl ethers, amines, amides and esters, DEC the ethyl compounds, EC the hydroxyethyl compounds [10], and diallyl carbonate the corresponding allyl compounds. DMC and DEC also react with SiO_2 and TiO_2 in the presence of alkali hydroxide in a solid-gas reaction at 250 – 350 °C to form $Si(OCH_3)_4$, $Si(OEt)_4$, $Ti(OCH_3)_4$ and $Ti(OET)_4$ [563–566] in high yields.

Ethylene carbonate is often preferred to ethylene oxide for oxyethylation reactions because it is much easier to handle.

DPC and substituted DPC reacts analogously with, e.g., terephthalic acid to form its diphenyl ester and 4,4′-dinitrodiphenyl carbonate, which is converted to the 4,4′-dinitrodiphenyl ether and CO_2 at higher temperatures (risk of explosion).

Cyclic carbonates eliminate CO_2 catalytically to form cyclic ethers such as oxiranes, oxetanes, dioxanes, crown ethers, etc. by intramolecular alkylation [11], [567–570].

Active CH bonds are alkylated by carbonates, with substitution of the H atom by the alkyl group. Examples are phenylacetonitriles, phenyl

acetates [571], butyrolactone [572], malonic diesters [24], and acetoacetic esters.

Catalysts for the alkylation are bases, e.g., alkoxides, K_2CO_3, 4-dimethylaminopyridine, quaternary ammonium salts, platinum group salts, aluminum oxides, basic Mg compounds, and SnO. The reactions usually run smoothly and produce excellent yields, with CO_2 and the feedstock alcohol as the byproducts. These can be easily recovered for reuse. The same also applies to the other reactions discussed here, which is the reason for the strong growth in interest from industry in the chemistry of carbonic esters. Many of the substances produced by these alkylation processes are commercial products that are often still produced with other alkylation agents, some of which are toxic and produce salts.

Carboxylation. Under mild conditions in the presence of equivalent quantities of strong bases, active CH groups react with carbonates to yield carboxylic esters. This applies to substances such as ketones, nitriles, aliphatic nitro compounds and esters, which form, e.g., malonic esters [314], [315], [496–501].

Further considerable potential is provided by the transformation of amines to ureas and, in particular, urethanes. Both groups yield substances for crop protection [573], [574], pharmaceutical preparations [575], [576], and, by reaction of amines with EC [577], [578] lubricant additives [579]. The methyl isocyanate required for the synthesis of crop protection products is obtained by treating DPC with methylamine to form methyl phenyl urethane for cleavage [580], [581]. Bayer and ENI manufacture methyl isocyanate in this way.

The reaction of diamines with carbonates to form bis-urethanes and diisocyanates for polymers has been the focus of a great deal of work, since high selectivity and very high purity are required. Catalysts for urethane formation are Co complexes, organotin compounds, zinc salts and in α-position branched carboxylic acids [582–584], lead, titanium and zirconium derivatives [585–587], which also catalyze the reaction of ureas with carbonates to form urethanes. The secondary reactions of alkylation are virtually excluded, especially when working with Ti alkoxides and zinc acetate.

Diarylphosphinic acid is claimed to be an excellent catalyst for the reaction of DPC with aromatic amines [588]. Aliphatic mono- and diamines react virtually quantitatively with DPC, even under mild conditions, to yield *O*-phenylurethanes, which can be cleaved easily to yield isocyanates [589], [590]. The corresponding reaction with DMC to form *O*-methylurethanes requires catalysis with alkali metal alkoxides [591]. Daicel and Degussa-Hüls plan to use this process in the joint production of isophorone diisocyanate for polyurethane coatings.

Cyclization. Analogous to the reactions of diols to yield cyclic carbonates, many other compounds with two reactive groups such as 2-aminoalcohols, α- and β-hydroxycarbonamides, malonic diamides, 1,2- and 1,3-diamines yield known heterocyclic substances which are used in pharmaceutical and crop protection applications [592]. It is even possible to cyclize polyamidic acids to form polyimides [593].

Pyrolysis. Heating longer chain carbonates to $>300\,^{\circ}C$ yields olefins. ω-Hexenol is distilled off from hexanediol-1,6-polycarbonate, and methanol, CO_2 and ω-decenol (an insect attractant [23] in ca. 75 % yield) from 1,10-decanediol monomethyl carbonate.

At much lower temperatures, methanol and CO_2 are split off from the methyl carbonates of the β-hydroxyethyl compounds with the structure $HO–(CH_2)_2–X$ (X = N, S, O, P), which are easily produced by oxyethylation, evidently benefited by the neighboring group effect. This provides access to easily polymerizable vinyl phosphonates, *N*-vinyl heterocyclene [594–596] and *N*-vinylphthalimide [13], [597] for flame-resistant polymers and special ion exchangers.

Polymerization of Cyclic Carbonates. The aromatic cyclic carbonates produced from the bis-phenols, especially those produced from Bisphenol A, polymerize in the presence of many catalysts to form thermoplastics with medium to very high molecular masses and very high quality [12–16]. Compounds with two carbonate rings can be used for branching and achieving flame resistance [598]. Depending on their structure, other aromatic cyclic carbonates undergo statistical or block copolymerization with cyclic Bisphenol A carbonates to yield thermoplastics with high glass transition tempera-

tures [599], [600]. An unsymmetrical cyclic carbonate yields an optically active polycarbonate [601].

A wider variety of options are offered by the polymerization of aliphatic cyclic carbonates [602]. Catalyzed with cationic, anionic, or organometallic compounds, in solution or in the molten state, they form copolymers with widely differing properties and, in some instances, very high molecular masses [17–22]. Lactams yield copolymers which, depending on the monomer content, contain urethane, amide, ester, and carbonate groups and can be used to improve the flow properties in the melt and increase the notch impact resistance of polyamides [603–605]. Other polymers obtained from cyclic carbonates increase the impact and fracture resistance of poly(alkylene terephthalates) [606]. The addition of bicyclic carbonates results in polymerization with cross-linking to yield hard, elastic, and impactresistant thermosets with very good glass adhesion [437], [607], [608]. Polymers with OH end groups made from neopentyl glycol carbonate are suitable for the manufacture of PU rubber [609]. Cyclic carbonates produced from derivatives of trimethylpropane offer further options for the manufacture of new homopolymers and copolymers [610], [611]. The copolymers with allylic double bonds can be radically cross-linked [612]. Furthermore, copolymers have been produced with aromatic cyclic carbonates, acrylates, styrene, polyethers, and siloxanes, the potential applications of which are by no means exhausted [613–615].

An interesting new field of application for polymers and copolymers made from six-membered cyclic carbonates, lactones, lactide and glycolide is in medicine as these polymers are biocompatible, biodegradable and resorbable. They are thus suitable for drug-release implants, sutures, and tissue and repair materials [616–619]. Polymers made from cyclic carbonates are recyclable and are easily cleaved into monomers by thermal degradation [620], [621].

Cyclic aliphatic carbonates, especially that based on trimethylolpropane, are now used in the manufacture of high-quality polyurethane coatings [622–625].

EC can also be polymerized [626]. However, the reaction of EO with CO_2, which yields polymers directly [627], [628], is more important. These have a good barrier effect to oxygen and moisture and are suitable for the manufacture of packaging films [629–631].

9. Economic Aspects [117–119], [120], [252], [632–653]

Large amounts of carbonates do not reach the market as they are used primarily as intermediates. However, this could change if DMC with an oxygen content of 53 % were used to increase the octane number of gasoline or to purify diesel exhaust. Today, DMC and DPC are the most important carbonates in terms of volume. DMC is produced by ENI (ca. 12 000 t/a; oxidative carbonylation) for a number of products, General Electric in Spain (ca. 60 000 t/a) and General Electric Plastics in Japan (ca. 18 000 t/a; both using the ENI process), SNPE in France (ca. 1000 t/a; via phosgene), in Tongling, China (1000 t/a; PC/MeOH transesterification), Ube (12 000 t/a; oxidative carbonylation with NO_x, with planned expansion to 50 000 t/a), Nanjing Chem. Ind. in China (600 t/a), Hoechst Celanese, PPG, Mitsubishi, and BASF.

General Electric has purchased a license from ENI for the DMC and DPC processes and is working with these in two Bisphenol A polycarbonate plants operating the melt process (130 000 t/a in Spain, 45 000 t/a in Japan with plans to expand to 260 000 t/a in Spain). This is equivalent to a total DPC volume of ca. 280 000 t/a. Other DPC manufacturers are Bayer, BASF, ENI, SNPE, PPG, Mitecs, and Mitsubishi.

In addition to Bisphenol A polycarbonate, DPC is also used in the synthesis of methyl isocyanate (crop protection products) and 1,6-hexanediol polycarbonate. DMC can also be used to produce aliphatic polycarbonates, lubricants, methylated phenols, crop protection products, pharmaceuticals and isophorone diisocyanate, as is intended by a joint venture between Daicel and Degussa-Hüls AG.

EC and PC are used as solvents and extractants but also in other reactions, e.g., in DMC synthesis in Nanjing and Tongling. Dow, Degussa-Hüls, BASF, Arco, and Texaco also produce cyclic carbonates, and Ube produces glycerin carbonate.

DEC, primarily used as a solvent, is manufactured by SNPE, BASF, Bayer, Mitsubishi, Wuxian (China), and Essex Chem. (USA).

The kilogram price of the carbonates available in large quantities is around $ 1 – 3. Special carbonic esters can command higher prices.

The economic significance of carbonic esters will continue to increase since they are intermediates with a wide range of uses, are easy to handle, and satisfy the growing demand for environmentally friendly processes and high-quality products.

10. References

1. *Houben-Weyl,* **E4,** 64 – 112.
2. *Beilstein,* **3 (III)**, 4, 6.
3. Azeotropic Data, Advances in Chemistry Series 116, American Chem. Soc.
4. J. Chem. Eng. Data, Int. Data Series, Sel. Data Mixtures, Ser. A, 1992/1993.
5. Jefferson Chem. Comp. Techn. Bulletin (1960).
6. Bayer, EP 880, 1977 (H. Krimm, H.-J. Buysch, H. Rudolph).
7. N. Yamazaki, T. Iguchi, F. Higashi. *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) no. 3, 835 – 841.
8. Anic S.p.A., DE-OS 3007196, 1979 (G. Jori, U. Romano).
9. Anic S.p.A., GB-A 2026484, 1978 (G. Jori, U. Romano).
10. Mellon Inst. of Ind. Res., US 2448767, 1942 (W. W. Carlson).
11. Union Carbide Corp., EP-A 47474, 1980 (C. H. McMullen).
12. A. Horbach, H. Vernaleken, K. Weirauch, *Makromol. Chem.* **181** (1980) 111.
13. General Electric, US 4 345 062, 1981 (D. J. Brunelle).
14. D. J. Brunelle et al., *Polym. Prepr., Amer. Chem. Soc., Div. Poly. Chem.* **30** (1989) no. 2, 569.
15. General Electric, EP 400 478 A3, 1990 (H. O. Krabbenhoft, E. B. Boden).
16. J. Haggin, *Chem. Eng. News* (1992) May 4, 25.
17. J. Matsuo, K. Aoki, F. Sanda, T. Endo, *Macromolecules* **31** (1998) no. 14, 4432.
18. A. C. Albertson, H. Sjoeling, *J. Macromol. Sci., Pure Appl. Chem.* **A29** (1992) no. 1, 43.
19. H. R. Kricheldorf, J. Jensen, *J. Macromol. Sci., Chem.* **A26** (1989) no. 4, 631.
20. T. Ariga, T. Takata, T. Endo, *Macromolecules* **30** (1997) no. 4, 737.
21. Bayer, EP 236 862 B1, 1987 (E. Leitz et al.).
22. Bayer, EP 209 722 B1, 1986 (P. Mues, D. Kerimis, H. P. Müller, H.-J. Buysch).
23. National Distillers & Chem. Corp., DE-OS 3331929, 1982 (C. Blewett).
24. H. Krauch, W. Kunz: *Reaktionen der organischen Chemie,* 5th ed., Hüthig Verlag, Heidelberg 1976, pp. 370 – 372. Mallinckrodt Chem. Works, US 2454360,1948 (V. H. Wallingford, A. H. Homeyer).
25. General Electric, EP 228 672 A2, 1985 (M. Janatpour, S. J. Shafer).
26. A.-A. G. Shaik, S. Sivaram, *Chem. Rev.* **96** (1996) 951 – 976.
27. Tokuyama Soda Co., DE-OS 2838701, 1977 (S. Kazuo, S. Toshiaki).
28. H. Babad, A. G. Zeiler, *Chem. Rev.* **73** (1973) no. 1, 81.
29. Bayer, EP 21211, 1979 (H.-J. Buysch, H. Krimm, H. Böhm).
30. PPG Industries, US 4273726, 1979 (S. Altuglu).
31. Pennwalt Corp., DE-OS 2926354, 1978 (J. R. Angle, U. D. Wagle, D. C. Reid).
32. M. Matzner, R. R. Kurkjy, R. J. Cotter, *Chem. Rev.* **64** (1964) 645.
33. Bayer, EP 57 360, 1981 (H. Krimm, H.-J. Buysch).
34. Bayer, EP 369 249 A3, 1988 (N. Schön, H.-J. Buysch).
35. L. Piech, *J. Amer. Chem. Soc.* **73** (1951) 5779.
36. General Electric Co., DE-OS 2738437, 1977 (J. E. Hallgren).
37. General Electric Co., DE-OS 2815501, 1978 (A. J. Chalk).
38. J. E. Hallgren, G. M. Lucas, R. O. Matthews, *J. Organomet. Chem.* **204** (1981) no. 1, 135 – 138.
39. R. S. Hanslick, W. F. Bruce, A. Mascitti, *Org. Synth.* **33** (1953) 74.
40. Bayer, DE 1101386, 1956 (L. Bottenbruch, H. Schnell).
41. R. E. Stenseth, R. M. Schisla, J. W. Baker, *J. Chem. Eng. Data* **9** (1964) 390 – 397.
42. Teijin, JP 6 293 708 A, 1993 (K. Masumoto, K. Kyoshige).
43. Mitsubishi, JP 9 012 507 A, 1995 (Mitsukami et al.).
44. Mitsubishi, JP 9 012 511 A, 1995 (Mitsukami et al.).
45. Mitsubishi, JP 9 100 256 A, 1995 (Y. Hara, M. Tojima, H. Tsuchisada, H. Koto).
46. General Electric, EP 633 241 B1, 1993 (R. B. Ehlinger).

47. Bayer, DE-OS 2447348, 1974 (H.-J. Buysch, H. Krimm).
48. General Electric, US 2 362 865, 1941 (S. Tyron, W. S. Benedict).
49. UCC, US 3 234 262, 1962 (R. R. Kurkjy, M. Matzner, R. J. Cotter).
50. UCC, US 3 234 263, 1962 (R. R. Kurkjy, M. Matzner, R. J. Cotter).
51. T. Arine, N. Mori, Y. Tsurumaki, *Chem. Express.* **9** (1993) no. 6, 377.
52. Dow, US 5 167 946, 1990 (M. Mullins, A. T. Chamberlin, R. Galvan).
53. Dow, US 2 837 555, 1956 (J. M. Lee).
54. Dow, US 5 183 919, 1991 (C. J. Rand).
55. Mitsubishi, JP 9 024 278 A2, 1995 (Y. Hara, H. Koto, M. Higshijima).
56. Bayer, EP 40 741, 1980 (G. Rauchschwalbe, H. U. Blank, K. Mannes, D. Mayer).
57. BASF, EP 542 117, 1991 (T. Wettling, J. Henkelmann, J. Troetsch-Schaller, H. Köhler).
58. BASF, EP 542 132 A1, 1991 (H. Köhler, T. Wettling, W. Franzischka, L. Hüpfer).
59. Dow, EP 567 677, 1989 (C. L. Rand).
60. Dow, WO 9 525 083, 1994 (R. Galvan, M. J. Mullins).
61. Mitsubishi Chem., JP 10 245 366 A2, 1997 (Y. Katsro, H. Tsuchisada, M. Kawai).
62. Mitsubishi, JP 9 110 803 A2, 1995 (Y. Hara, H. Koto, M. Tojima).
63. Bayer, EP 483 632 A2, 1990 (H.-J. Buysch, N. Schön, G. Jeromin).
64. Dow, EP 516 355 A3, 1991 (A. D. Harley, J. Puga).
65. Bayer, EP 635 477 A1, 1993 (P. Ooms, N. Schön, H.-J. Buysch).
66. Bayer, EP 635 476 A, 1993 (P. Ooms, N. Schön, H.-J. Buysch).
67. Dow, US 5 239 105 A1, 1992 (P. G. Pews, R. G. Bowmann).
68. Bayer, EP 645 364 A1, 1993 (P. Ooms, N. Schön, H.-J. Buysch).
69. Mitsubishi, JP 9 000 923, 1995 (Y. Hara, H. Koto, M. Higashijima).
70. Bayer, EP 722 930 A1, 1995 (P. Ooms, H.-J. Buysch, N. Schön).
71. Bayer, EP 691 326 A1, 1994 (P. Ooms, H.-J. Buysch).
72. Dow, EP 500 786 A1, 1989 (A. D. Harley, S. King, C. L. Rand).
73. Bayer, EP 808 821 A2, 1996 (H.-J. Buysch et al.).
74. Bayer, EP 757 029 A1, 1996 (P. Ooms, H.-J. Buysch, S. Kühling, G. Zaby).
75. Bayer, EP 791 574 A1, 1997 (P. Ooms, H.-J. Buysch, S. Kühling, G. Zaby).
76. Bayer, EP 784 048 A1, 1997 (H.-J. Buysch, K. Hallenberger, G. Jeromin, S. Kühling, P. Ooms).
77. National Distillers, US 3 114 762, 1960.(J. L. Mador, A. U. Blackham).
78. Union Oil, US 3 227 740, 1963 (D. M. Fenton).
79. Mitsui Toatsu, JP 45 024 966, 1967 (H. Ota).
80. SNAM Progetti, US 3 846 468, 1970 (E. Perotti, G. Cipriani).
81. Röhm, US 3 952 045, 1973 (W. Gänzel, K. Kabs, G. Schröder).
82. SNAM Progetti, US 3 980 690, 1973 (G. Ciprani, E. Perotti).
83. F. Rivetti, U. Romano, *J. Organomet. Chem.* **174** (1979) 221 – 226.
84. Atlantic Richfield, US 4 005 128, 1975 (L. R. Zehner, R. W. Sauer).
85. D. M. Fenton, P. J. Steinwand, *J. Org. Chem.* **39** (1974) no. 5, 701.
86. Röhm, US 4 620 029 A, 1983 (G. Schröder, E. Baumgartner).
87. BP, GB 2 148 881 A, 1983 (S. F. Davison).
88. U. Romano, F. Rivetti, *Chim. Ind. (Milan)* **62** (1980) 7.
89. Atlantic Richfield, US 4 451 666, 1982 (J. A. Sofranko, A. M. Gaffney).
90. Shell Int. Res., EP-A 71286, 1981 (E. Drent).
91. U. Romano, R. Tesei, M. M. Mauri, P. Rebora, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) no. 3, 396 – 403.
92. Bayer, DE-OS 3016187, 1981 (G. Stammann, R. Becker, J. Grolig, H. Waldmann).
93. General Electric Co., US 4360477, 1981 (J. E. Hallgren, G. M. Lucas).
94. Texaco, US 4 636 576, 1985 (A. Bhattacharya, J. T. Nolan).
95. Texaco, US 5 001 252, 1987 (A. Bhattacharya).
96. Daicel, EP 452 997 B1, 1986 (S. Yokota, H. Koyama, H. Kojima).
97. Texaco, US 4 638 076, 1985 (A. Bhattacharya).
98. BASF, EP 413 217 B1, 1989 (K. Jörg, R. Kummer, F.-J. Müller).
99. Texaco, EP 217 651, 1985 (A. Bhattacharya, J. T. Nolan).
100. Daicel, EP 259 788 A2, 1986 (S. Yokota, H. Koyama, H. Kojima).
101. Daicel, EP 354 970 A1, 1988 (S. Yokota, Y. Tanaka, H. Miyake).
102. Daicel, JP 2 040 347 A, 1988 (H. Suzuki, H. Kojima).
103. Daicel, JP 2 004 737 A, 1988 (S. Yokota, H. Suzuki).
104. Texaco, US 4 785 130, 1984 (A. Bhattacharya).
105. Texaco, US 4 879 266, 1988 (A. Bhattacharya).

106. Mitsubishi, EP 532 861, 1990 (E. Watanabe, K. Murayama, K. Ida, K. Wada, Y. Kasori).
107. Daicel, JP 5 097 773 A2, 1991 (Y. Tanaka, M. Sugawara).
108. Mitsubishi, JP 5 246 954 A2, 1992 (T. Ookawa, T. Aoki).
109. Denki, JP 625 625 A2, 1993 (T. Arai, H. Kaneko, K. Aoki).
110. Daicel, JP 2 218 647, 1989 (Y. Yoshida, C. To, B. Ko, H. Ko).
111. Daicel, JP 5 017 410 A2, 1991 (Y. Tanaka, M. Sugawara, S. Oda, H. Kojima).
112. Röhm, EP 90 977 B2, 1982 (E. Baumgärtner, G. Schroeder, S. Besecke).
113. Dow, US 4 604 242, 1984 (A. D. Harley, J. L. Curnutt, D. T. Doughty).
114. Mitsubishi, JP 8 092 169 A2, 1994 (J. Yasumaru, T. Masunaga, M. Shiraishi).
115. Daicel, EP 521 480 A2, 1991 (H. Tanigawa).
116. *Chem. Eng. (N.Y.)* **90** (1983) no. 9, 15.
117. M. M. Mauri, U. Romano, F. Rivetti, *Ing. Chim. Ital.* **21** (1985) nos. 1 – 3, 6.
118. Anic, US 4 218 391, 1980 (U. Romano, R. Tesei, G. Cipriani, L. Miucci).
119. L. Cassar, *Chim. Ind. Milan* **72** (1990) 18.
120. Enichem Synth., EP 365 083 A1, 1988 (U. Romano, F. Rivetti).
121. Enichem Synth., EP 366 177 B1, 1988 (U. Romano, F. Rivetti).
122. Enichem Synth., EP 534 545 A2, 1991 (U. Romano, F. Rivetti).
123. JGC Corp., JP 5 255 200 B4, 1992 (N. Yamada).
124. Enichem Synth., EP 445 891 A1, 1990 (D. Delledonne, F. Rivetti).
125. Denki, JP 5 310 644, 1992 (T. Arai, H. Kaneko, M. Yamada).
126. Bayer, EP 894, 1977 (H.-J. Buysch, H. Krimm, H. Rudolph).
127. BASF, DE-OS 2706684, 1977 (W. Himmele, K. Fischer, G. Kaibel, K. Schneider, R. Irnich).
128. Anic S.p.A., DE-OS 2607003, 1975 (U. Romano).
129. Daicel, JP 2 212 456 A1, 1989 (K. Koga et al.).
130. Enichem Synth., BE 1 005 927 A3, 1992.
131. Daicel, JP 6 092 906 A2, 1992 (S. Oda, N. Yoshida).
132. Bayer, DE 4 234 525 A1, 1992 (A. Nickel, W. Arlt, I. Janisch, P. Wagner, A. Klausener).
133. Hüls AG, DE 4 408 697 A1, 1994 (S. Sridhar).
134. Bayer, EP 555 697 A1, 1992 (J. Rechner, A. Klausener, H.-J. Buysch, P. Wagner).
135. JCJ, EP 134 668 A1, 1983 (G. D. Short, M. S. Spencer).
136. BASF, EP 413 215 B1, 1989 (K. Jörg, F. J. Müller, W. Harder, R. Kummer).
137. Enichem Synth., EP 460 732 A1, 1991 (N. Di Muzio, C. Fusi, F. Rivetti, G. Sasselli).
138. Enichem Synth., EP 460 735 A3, 1991 (G. Paret, G. Donati, M. Ghirardini).
139. Enichem Synth., EP 634 390, 1994 (F. Rivetti, U. Romano, G. Garone, M. Ghirardini).
140. IGC Corp., US 5 780 663 A, 1996 (K. Mori, T. Koyama).
141. Dow, US 4 625 044 A, 1984 (G. L. Curnutt).
142. Dow, WO 8 707 601 A, 1986 (G. L. Curnutt).
143. *Chem. Eng. News* 1987, Sept. 28, 26.
144. Dow, US 5 387 708 A, 1993 (J. Puga, M. E. Jones, D. C. Molzahn, G. E. Hartwell).
145. Dow, US 5 391 803 A, 1992 (M. M. Olken, M. E. Jones, S. T. King).
146. S. T. King, *J. Catal.* **161** (1996) 530.
147. JGC Corp., WO 9 015 791, 1990 (K. Mori et al.).
148. JGC Corp., EP 528 498 A2, 1991 (T. Koyama, M. Tonosaki, N. Yamada, K. Mori).
149. JGC Corp., JP 7 194 983 A2, 1994 (T. Koyama, K. Mori).
150. JGC Corp., JP 9 010 591 A2, 1995 (K. Takiguchi, K. Mori).
151. JGC Corp., JP 2 767 146 B2, 1989 (K. Yamamoto, T. Toida, K. Mori, N. Yamada).
152. Quantum Chem. Corp., US 5 171 874, 1989 (D. W. Smith, S. D. Landau).
153. Denki, JP 5 255 201 A2, 1992 (T. Arai, H. Kaneko, K. Aoki).
154. Denki, JP 6 092 908 A2, 1992 (T. Arai, M. Mazushima, K. Aoki).
155. Mitsubishi, JP 8 169 863 A2, 1994 (N. Yamagata et al.).
156. Daicel, EP 603 766 A1, 1992 (H. Matsuda, S. Oda, M. Ohno).
157. Daicel, EP 584 785 A2, 1992 (A. Horiguchi, S. Oda).
158. Daicel, EP 607 943 B1, 1993 (S. Oda, M. Ohno).
159. Daicel, EP 727 409 A1, 1993 (S. Oda, M. Ohno).
160. Bayer, EP 544 162 A1, 1991 (Z. Kricsfalussy, H. Waldmann, H.-J. Traencker, M. Zlokarnik, R. Schomaecker).
161. Bayer, EP 636 601 A1, 1993 (Z. Kricsfalussy, H. Waldmann, H.-J. Traencker).
162. Bayer, EP 659 731 A1, 1993 (Z. Kricsfalussy, H. Steude, H. Waldmann, K. Hallenberger, W. Wagner).
163. D. Masi, T. Funaioli, *J. Organomet. Chem.* **1991,** 417.
164. Enichem Synth., EP 463 678 B1, 1991 (D. Delledonne, F. Rivetti, U. Romano).

165. Enichem Synth., EP 574 082, 1992 (F. Rivetti, U. Romano).
166. Enichem Synth., EP 558 128 A2, 1992 (F. Rivetti, D. Delledonne, D. Dreoni).
167. F. Rivetti, U. Romano, D. Delledonne, "Designing Chem. Eur. Green Chemistry, Chapt. 6," *Symp. Amer. Chem. Soc., Symp. Ser. 626* (1996) 78.
168. General Electric, DE OS 2 738 487, 1976 (A. J. Chalk).
169. General Electric, DE OS 2 949 936, 1978 (J. E. Hallgren).
170. General Electric, DE OS 2 815 512, 1978 (J. E. Hallgren).
171. General Electric, DE OS 2 738 488, 1976 (J. E. Hallgren).
172. General Electric, DE OS 2 738 520, 1976 A. J. Chalk).
173. General Electric, DE OS 2 738 519, 1976 (J. E. Hallgren).
174. General Electric, US 4 349 485, 1979 (J. E. Hallgren).
175. General Electric, EP 583 937 A1, 1992 (E. J. Pressman, J. A. King Jr.).
176. General Electric, EP 583 935 B1, 1992 (R. P. Joyce, J. A. King Jr., E. J. Pressman).
177. General Electric, EP 867 428 A1, 1997 (E. J. Pressman, S. J. Shafer).
178. J. E. Hallgren, R. O. Matthews, *J. Organomet. Chem.* **175** (1979) 135.
179. General Electric, EP 350 700, 1988 (TC.-T. Chang).
180. Asahi, JP 95 068 178 B2, 1987 (N. Fukuoka, H. Kogawa, T. Watanabe).
181. General Electric, US 5 142 086, 1991 (J. A. King Jr., T. E. Krafft, G. R. Faler).
182. Idemitsu, US 5 336 803, 1991 (H. Kezuka, F. Okuda).
183. A. Vavasori, L. Toniolo, *J. Mol. Catal. A. Chem.* **139** (1999) nos. 2 – 3, 109. Mitsubishi, JP 6 271 509 A2, 1993 (H. Iwane, K. Yoneyama, H. Miyashiro).
184. General Electric, EP 867 428 A1, 1997 (E. J. Pressman, S. J. Shafer).
185. General Electric, US 5 760 272 A1, 1997 (E. J. Pressman, S. J. Shafer).
186. General Electric, EP 858 991 A1, 1997 (E. J. Pressman, S. J. Shafer).
187. Mitsubishi, EP 736 512 A1, 1993 (H. Iwane, H. Miyagi, S. Imada, S. Seo, T. Yoneyama).
188. Mitsubishi, JP 9 110 804 A2, 1995 (Y. Hara, M. Tojima).
189. Bayer, EP 654 461 B1, 1993 (H.-J. Buysch, C. Hesse, J. Rechner, J. Dohm).
190. Bayer, US 5 498 742, 1994 (H.-J. Buysch et al.).
191. Bayer, US 5 625 091, 1995 (H.-J. Buysch, C. Hesse, J. Rechner).
192. A. Gelbein, R. Piccolini, *Chemtech.* (1996) May, 30.
193. General Electric, WO 9 303 000, 1991 (J. A. King Jr., P. D. Mackenzie, E. J. Pressman).
194. Teijin, JP 10 045 677 A2, 1996 (H. Yoshisato).
195. Idemitsu, JP 6 211 750 A2, 1993 (F. Okuda).
196. Mitsubishi, EP 614 867 B1, 1992 (H. Iwane, H. Miyagi, S. Imada, S. Seo, T. Yoneyama).
197. Mitsubishi, US 5 726 340 A, 1996 (M. Tagaki, N. Miyagi, Y. Ohgomori, H. Iwane).
198. Asahi, JP 4 257 546 A, 1991 (S. Fukuoka).
199. Asahi, JP 4 261 142 A, 1991 (S. Fukuoka).
200. General Electric, EP 507 546 B1, 1991 (J. A. King Jr., G. R. Faler, T. E. Krafft).
201. General Electric, EP 583 936 A1, 1992 (S. J. Shafer).
202. General Electric, US 5 756 801 A, 1997 (S. J. Shafer, E. J. Pressman).
203. General Electric, EP 757 030 A1, 1995 (G. L. Kissinger).
204. Bayer, EP 687 666 A1, 1994 (H.-J. Buysch, J. Rechner, C. Mendoza-Frohn, N. Schön).
205. Bayer, EP 801 052 A1, 1996 (H.-J. Buysch, J. Rechner, H.-P. Wirges).
206. Bayer, EP 801 053 A1, 1996 (H.-J. Buysch, C. Hesse, J. Rechner, H.-P. Wirges).
207. Bayer, EP 790 231 A1, 1996 (H.-J. Buysch, C. Hesse, J. Rechner).
208. Bayer, EP 807 619 A1, 1996 (H.-J. Buysch, C. Hesse, J. Rechner).
209. Bayer, EP 806 243 A1, 1996 (H.-J. Buysch, C. Hesse, J. Rechner).
210. Mitsui, JP 5 025 095 A2, 1991 (T. Fujita, Y. Kiso, T. Nagata, H. Iwasaki).
211. Mitsubishi, EP 572 980 A2, 1992 (M. Mazukami, K. Hayashi, K. Jura, T. Kawaki).
212. Mitsubishi, JP 8 169 863 A2, 1994 (N. Yamagata et al.).
213. Mitsubishi, JP 8 092 168 A2, 1994 (K. Kujira, M. Takagi, H. Myagi, H. Hayashi, J. Ookago).
214. Mitsubishi, JP 8 089 810 A2, 1994 (K. Kujira, M. Takagi, H. Myagi, H. Hayashi, J. Ookago).
215. Bayer, EP 736 511 A1, 1995 (H.-J. Buysch, C. Hesse, J. Rechner).
216. Bayer, EP 736 324 A2, 1995 (H.-J. Buysch, C. Hesse, J.-D. Jentsch, J. Rechner, E. Zirngiebl).
217. Bayer, EP 736 325 A2, 1995 (H.-J. Buysch, C. Hesse, J.-D. Jentsch, J. Rechner, E. Zirngiebl).
218. Bayer, US 5 821 377, 1996 (H.-J. Buysch, C. Hesse, J. Rechner).
219. A. Gelbein, R. Piccolini, *Chemtech.* (1999) Feb., 31.

220. Ube, JP 56 164 145 A2, 1980.
221. Mitsubishi, JP 6 075 447, 1983 (Y. Miyamori et al.).
222. Toa Nenryo, JP 6 094 943, 1983 (T. Shakahibara et al.).
223. Toa Nenryo, JP 6 011 443, 1983 (Y. Okumura et al.).
224. Toa Nenryo, JP 6 018 051, 1984 (Y. Okumura et al.).
225. Ube, EP 425 197 A2, 1989 (K. Nishihira, K. Mizutare, S. Tanaka).
226. Ube, EP 501 507 A2, 1991 (K. Nishihira, S. Tanaka, K. Kodama, T. Kaneko).
227. Ube, JP 8 020 561 A2, 1994 (K. Oodan, K. Ishichi, T. Kurato, Y. Yamazaki).
228. Ube, JP 7 118 210 A2, 1994 (T. Matsuzaki, K. Ishichi).
229. Ube, EP 559 212 A1, 1993 (K. Nishihira et al.). Bayer, EP 559 001 A2, 1992 (H. Landscheidt, A. Klausener, E. Wolters, L. Puppe).
230. Bayer, EP 558 996 A1, 1992 (H. Landscheidt, A. Klausener, E. Wolters, L. Puppe).
231. Bayer, EP 632 583 A1, 1993 (J.-D. Jentsch, A. Klausener, H. Landscheidt, E. Wolters, E. Zirngiebl).
232. Ube, EP 503 618 A1, 1991 (T. Matsuzaki, T. Simamura, S. Fujitsu, Y. Toriyahara).
233. Bayer, EP 538 676 A1, 1991 (H. Landscheidt, E. Wolters, A. Klausener, H.-U. Blank, U. Birkenstock).
234. Bayer, EP 464 460 A2, 1990 (A. Klausener, H. Landscheidt, H.-U. Blank, U. Kipshagen).
235. Mitsubishi, JP 5 255 197 A2, 1992 (H. Watanabe, K. Murayama).
236. Bayer, EP 634 387 A1, 1993 (J.-D. Jentsch et al.).
237. Bayer, DE 4 323 684 A1, 1993 (H. Landscheidt, A. Klausener).
238. Bayer, EP 523 508 B1, 1992 (H. Landscheidt, A. Klausener, E. Wolters, H.-U. Blank, U. Birkenstock).
239. Ube, WO 9 411 335 A1, 1992 (T. Kurafuji, N. Manada, M. Murakami, Y. Yamamoto).
240. Ube, EP 565 076 A1, 1992 (T. Matsuzaki, T. Shimamura, Y. Toriyahara, Y. Yamasaki).
241. Bayer, EP 634 388 A1, 1993 (A. Hable et al.).
242. Ube, WO 9 206 066 A1, 1991 (T. Matsuzaki, T. Shimamura, S. Fujitsu, Y. Toriyhara).
243. Ube, EP 581 240 A1, 1992 (N. Manada, M. Murakami, K. Abe, Y. Yamamoto, T. Yamafuji).
244. Ube, JP 7 196 581 A2, 1994 (H. Yoshimoto).
245. Y. Yamamoto, T. Matsuzaki, K. Ohdan, Y. Okamoto, *J. Catal.* **161** (1996) 577.
246. Ube, EP 501 374 A1, 1991 (K. Nishihira, Y. Yoshida, S. Tanaka).
247. Ube, EP 523 728 A2, 1991 (K. Nishihira, Y. Yoshida, S. Tanaka).
248. Ube, EP 742 198 A1, 1995 (K. Nishihira, Y. Yoshida, S. Tanaka, Y. Asada).
249. Bayer, EP 634 386 A1, 1993 (H. Landscheidt, P. Wagner, A. Klausener, E. Wolters).
250. Bayer, EP 634 389 A2, 1993 (A. Klausener, H. Landscheidt, R. Langer, P. Wagner).
251. Bayer, EP 614 878 A1, 1993 (A. Klausener, Z. Kricsfalussy, H. Landscheidt, P. Wagner).
252. *Jap. Chem. Week,* 1992, Jan. 9.
253. Ube, EP 802 175 A1, 1996 (T. Matsuzaki, M. Hidaka).
254. Ube, EP 832 872 A1, 1996 (K. Harada, K. Nishihira, R. Sugise, S. Tanaka).
255. X.-Z. Jiang, *Platinum Metals Rev.* **34** (1990) no. 4, 178.
256. Ube, JP 9 143 123 A2, 1995 (K. Nishihira, S. Tanaka, Y. Nishida, S. Fujitsu).
257. Ube, EP 776 882 A1, 1995 (K. Nishihira, S. Tanaka, Y. Nishida, S. Fujitsu).
258. Ube, JP 10 101 619 A2, 1996 (K. Harada, Y. Suginose, T. Karafuji, S. Araida, T. Sumita).
259. Ube, JP 10 114 723 A2, 1996 (K. Harada, Y. Suginose, K. Kashiwagi, T. Sumita).
260. Ube, JP 10 101 621 A2, 1996 (K. Harada, Y. Suginose, T. Sumita, S. Araida, T. Karafuji).
261. Ube, JP 10 114 718 A2, 1996 (K. Harada, Y. Suginose, K. Kashiwagi, T. Sumita).
262. Ube, JP 10 114 717 A2, 1996 (K. Harada, Y. Suginose, Y. Inbe, T. Sumita, S. Yamamoto).
263. Ube, JP 10 053 563 A2, 1996 (K. Harada, Y. Suginose, K. Kashiwagi, T. Shimakawa).
264. Ube, EP 737 665 A1, 1995 (K. Harada et al.).
265. Ube, EP 834 495 A1, 1996 (K. Harada et al.).
266. Ube, JP 9 255 628 A2, 1996 (Y. Yamamoto, T. Sugimoto, T. Kurato).
267. Ube, JP 9 295 959 A2, 1996 (Y. Yamamoto, T. Sugimoto, T. Kurafuji).
268. Ube, JP 8 325 206 A2, 1995 (K. Harada, R. Sugise, Y. Inbe, T. Sumita).
269. Ube, EP 795 539 A1, 1996 (K. Nishihira et al.).
270. Ube, EP 832 872 A1, 1996 (K. Nishihira, S. Tanaka, K. Harada, R. Sugise).
271. Ube, EP 834 496 A1, 1996 (T. Doi et al.).
272. Allied Chem. Corp., US 4 131 521 A, 1977 (D. Cipris, J. L. Mador).
273. General Electric, US 4 310 393 A, 1980 (R. A. Bell).
274. Daicel, JP 6 073 582 A2, 1992 (K. Ootsuka).
275. A. Galia, G. Filardo, O. Scialdone, M. Musacco, G. Silvestri in S. Torii (ed.): *Novel Trends in Electroorg. Synth. 3rd ed.,* 1997 [*Pap. Int. Symp.*] Springer, Tokyo 1998, pp. 3 – 6.

276. W. J. Peppel, *Ind. Eng. Chem.* **50** (1958) 767 – 770.
277. Union Carbide Corp., US 4314945, 1977 (C. H. McMullen, J. R. Nelson, C. B. Ream, J. A. Sims Jr.).
278. N. Ryoki, N. Akira, M. Harno, *J. Org. Chem.* **45** (1980) no. 19, 3735 – 3738.
279. Halcon SD Group Inc., DE-OS 3244456, 1981 (R. J. Harvey, H. M. Sachs).
280. Standard Oil Co., EP-A 69494, 1981 (J. E. Rinz, C. Paparizos, D. R. Harrington).
281. BASF, DE 1169459, 1960 (W. Münster, E. Dreher).
282. G. Rokicki, W. Kuran, *Monatsh. Chem.* **115** (1984) 205.
283. M. Ratzenhofer, H. Kisch, *Angew. Chem. Int. Ed. Engl.* **19** (1980) 317.
284. Bayer, EP 499 924 A1, 1991 (H.-J. Buysch, A. Klausener).
285. Bayer, EP 543 249 A1, 1991 (F.-J. Mais, H.-J. Buysch, C. Mendoza-Frohn, A. Klausener).
286. Bayer, EP 546 428 A1, 1991 (P. Wagner, H.-J. Buysch, A. Klausener, F.-J. Mais, C. Mendoza-Frohn).
287. Bayer, EP 636 620 A1, 1993 (C. Mendoza-Frohn, H.-P. Wirges).
288. Bayer, EP 678 498 B1, 1994 (C. Mendoza-Frohn, P. Wagner, H.-J. Buysch).
289. Bayer, EP 628 553 A1, 1993 (P. Wagner, C. Mendoza-Frohn, H.-J. Buysch).
290. Halcon SD, EP 180 387 B1, 1985 (R. M. Weinstein).
291. G. Rokicki, M. Lewandowski, *Angew. Makromol. Chem.* **148** (1987) 53.
292. Höchst, DE 3 600 602 A1, 1986 (G. Brindöpke).
293. A. Baba, H. Kashiwagi, H. Matsuda, *J. Heterocycl. Chem.* **26** (1989) no. 6, 1659.
294. Bayer, EP 13957, 1979 (W. Heitz, P. Ball).
295. Bayer, EP 13958, 1979 (W. Heitz, P. Ball).
296. BASF, EP 41622, 1981 (W. Harder, F. Merger, F. Towae).
297. Exxon, WO 9 517 369, 1993 (R. Y. Salch, R. C. Michaelson, E. Suciu, B. Kuhlmann).
298. Mitsubishi, DE 19 749 860 A1, 1996 (M. Mizukami, Y. Arai, H. Harada, T. Oshida, H. Ohgi).
299. Mitsubishi, EP 866 051 A1, 1997 (M. Mizukami, Y. Arai, H. Harada, T. Oshida, H. Ohgi).
300. Mitsubishi, EP 638 541 A1, 1993 (M. Doya, K. Kimizuka, Y. Kanbara).
301. Mitsubishi, US 5 489 702, 1993 (D. Masaharu, K. Kenicki, K. Yutoka).
302. Idemitsu, JP 9 104 658 A, 1995 (M. Ito).
303. Chang Chun Plastics, EP 709 363 A1, 1994 (K.-Y. Hwang, Y.-Z. Chen, C.-C. Chu, H.-T. Liao).
304. Halcon SD, DE 3 219 828 C2, 1981 (J. E. Jacobson).
305. Halcon SD, EP 118 248 A1, 1983 (J. E. Jacobson).
306. M. Aresta, A. Ciccarese, E. Quaranta, *J. Mol. Catal.* **41** (1987) 355.
307. BASF, EP 175 241 A1, 1984 (K. Schneider, W. Best).
308. Bayer, EP-A 85347, 1982 (J. Genz, W. Heitz).
309. Showa Denko, JP 7 033 715 A2, 1993 (K. Ko, F. Ogata).
310. RWE, Ges. f. Forsch. u. Entw., WO 9 422 805 A1, 1993 (A. Wagner, W. Löffler, B. Haas).
311. Bayer, DE 2748718, 1977 (H.-J. Buysch, H. Krimm, H. Rudolph).
312. BP, WO 8 403 701 A1, 1984 (M. J. Green).
313. E. J. Beckmann, *Science* **283** (1999) no. 2, 946.
314. M. Aresta, E. Quaranta, *Chemtech. March* 1997, 32.
315. A. Behr in U. Romano (ed.): *Aspects of Homogeneous Catalysis,* **vol. 6,** D. Reidel Publishing Co., 1988, p. 59.
316. Bayer, EP 1082, 1977 (H.-J. Buysch, H. Krimm, H. Rudolph).
317. Bayer, EP 1083, 1977 (H.-J. Buysch, H. Krimm, H. Rudolph).
318. J. F. Knifton, R. G. Duranleau, *J. Mol. Catal.* **67** (1991) 389.
319. Daicel, JP 3 044 354 A2, 1989 (K. Honda).
320. Mitsubishi, JP 6 048 993 A2, 1992 (M. Dotani, T. Ookawa).
321. Nippon Catal. Chem., JP 6 107 601 A2, 1992, JP 6 238 165 A2, 1993 (M. Kirishiki, Y. Onda, H. Tsuncki).
322. Mobil Oil, US 5 498 743 A, 1994 (S. S. Shi, M. Wu, T. Yan).
323. Chiyoda, JP 7 048 319 A, 1992 (T. Kondoh, Y. Okada, F. Tanaka, S. Assoka, S. Yamamoto).
324. Mitsui Toatsu, JP 6 345 696 A2, 1993 (K. Inoue, H. Ookubo).
325. Mitsui Toatsu, JP 6 343 871 A2, 1993 (K. Inoue, H. Ookubo).
326. UCC, EP 478 073 A2, 1990 (K. S. Wayne, R. B. Claude).
327. Mitsubishi, JP 8 059 560 A2, 1 994 N Shimizu, J. Ookago, M. Imanari.
328. Enichem Synth., EP 685 453 A1, 1995(F. Rivetti, D. Delledonne).
329. Mitsubishi, JP 8 073 411 A2, 1994 (N. Shimizu, J. Ookago, M. Imanari).

330. Mitsubishi, JP 8 073 410 A2, 1994 (N. Shimizu, J. Ookago, M. Imanari).
331. Mitsubishi, JP 8 059 559 A2, 1994 (N. Shimizu, J. Ookago, M. Imanari).
332. Mitsubishi, JP 8 059 558 A2, 1994 (N. Shimizu, J. Ookago, M. Imanari).
333. Asahi, JP 4 009 356 A2, 1990 (M. Tojo, S. Fukuoka).
334. Nippon Catal. Chem., JP 6 239 806 A2, 1993 (Y. Urano, M. Kirishiki, Y. Onda, H. Tsuneki).
335. Nippon Shokulai, EP 599 287 A1, 1992 (Y. Urano, M. Kirishiki, Y. Onda, H. Tsuneki).
336. Mitsubishi, JP 8 176 071 A2, 1994 (N. Shimizu, M. Takiguchi, J. Ookago, M. Imanori). Tajin, JP 9 110 744 A2, 1996 (A. Yoshisato, T. Yoshitomi).
337. Texaco, US 5 214 182 A, 1991 (J. F. Knifton).
338. Bayer, DE 4 208 875 A1, 1992 (F.-J. Mais, H.-J. Buysch).
339. Bayer, EP 543 234 A2, 1991 (F.-J. Mais, H.-J. Buysch).
340. Daicel, JP 3 044 353 A, 1989 (K. Honda et al.).
341. Mitsui Toatsu, JP 6 336 461 A2, 1993 (K. Inoue, H. Ookubo).
342. Mitsui Toatsu, JP 6 336 462 A2, 1993 (K. Inoue, H. Ookubo).
343. Mitsui Toatsu, JP 6 343 870 A2, 1993 (K. Inoue, H. Ookubo).
344. Asahi, JP 63 238 043 A2, 1993 (M. Tojo, S. Fukuoka).
345. Asahi, JP 4 103 561 A2, 1990 (M. Tojo, S. Fukuoka).
346. Daicel, JP 63 140 066 A2, 1988 (S. Oda).
347. Texaco, EP 255 252 B1, 1986 J. F. Knifton).
348. Asahi, JP 63 238 043 A2, 1987 (M. Tojo, S. Fukuoka).
349. Asahi, JP 3 109 358 A2, 1989 (H. Minora, H. Nakajima).
350. Texaco, EP 298 167 B1, 1987 (R. G. Duranleau, E. C. Y. Nieh, J. F. Knifton).
351. *Chem. Eng. News,* 1992, May 4, 25.
352. Texaco, US 5 152 898 A, 1989 (C. R. Bartels).
353. Bayer, EP 581 115 A2, 1992 (J.-F. Mais, P. Wagner, H.-J. Buysch).
354. Amoco, US 5 489 703, 1995 (M. A. Pacheco, F. D. Darrington, J. C. Reier, B. D. Alexander).
355. Asahi, JP 4 103 561 A2, 1990 (H. Sasaki, K. Komiya, S. Fukuoka).
356. Asahi, JP 454 156 A, 1990 (H. Sasaki, K. Komiya, S. Fukuoka).
357. Asahi, JP 4 198 141 A, 1990 (S. Fukuoka).
358. Asahi, JP 9 176 061 A2, 1 995 M Tojo, S. Fukuoka, M. Kawamura).
359. Asahi, JP 889 025 A1, 1 995 M Tojo, S. Fukuoka, M. Kawamura).
360. Bayer, DE 4 129 316 A1, 1991 (H.-J. Buysch, A. Klausener, R. Langer, F.-J. Mais).
361. Bayer, EP 569 812 A1, 1992 (P. Wagner, F.-J. Mais, H.-J. Buysch, R. Langer, A. Klausener).
362. Bayer, US 5 266 716, 1993, US 5 440 066 A, (H.-J. Buysch, N. Schön, S. Kühling, H. Hähnsen).
363. Mitsui Petroch., EP 338 760 A, 1988 (Y. Kiso, Y. Matsunaga).
364. General Electric, EP 780 361 A, 1995 (S. Inoki et al.).
365. SNAM Progetti, DE-OS 2 528 412, 1974 (G. Illuminati, U. Romano, R. Tesei).
366. General Electric, DE-OS 3 445 553, 1983 (V. Mark).
367. Asahi, JP 57 183 745 A2, 1981.
368. Asahi, JP 57 176 932 A2, 1981.
369. Asahi, JP 1 093 560 A2, 1987 (S. Fukuoka, R. Deguchi, M. Tojo).
370. Asahi, JP 3 236 354 A2, 1990 (S. Fukuoka, R. Deguchi).
371. Bayer, DE 2 736 062, 1977 (H. Krimm, H.-J. Buysch, H. Rudolph).
372. Bayer, DE 2 736 063, 1977 (H. Krimm, H.-J. Buysch, H. Rudolph).
373. Mitsubishi, JP 7 963 023, 1977 (N. Yomazaki, S. Nakahama, K. Endo).
374. General Electric, DE 3 445 555 A1, 1983 (V. Mark).
375. Daicel, JP 62 277 315 A2, 1986 (Y. Harano).
376. Bayer, EP 444 293 A1, 1990 (N. Schön, H.-J. Buysch, W. Ebert).
377. General Electric, DE 3 308 921 A1, 1983 (J. E. Hallgren).
378. Nippon Shokubai, WO 9 517 371 A1, 1993 (H. Tsuneki, M. Kirishiki, K. Watanabe, Y. Onda).
379. Mitsubishi, JP 10 139 736 A2, 1996 (K. Hasegawa, T. Suzuki, M. Inaba).
380. Teijin, JP 9 241 217 A2, 1996 (A. Yoshisato, T. Yoshitomo).
381. Daicel, EP 591 923 B1, 1992 (K. Murata, K. Kawahashi, M. Watibiki).
382. Asahi, EP 461 274 B1, 1990 (S. Fukuoka, M. Tojo, M. Kawamura).
383. Mitsubishi, JP 6 263 694 A2, 1993, JP 6 345 697 A2, 1993 (H. Iwane, S. Seo, M. Takagi, M. Inaba).
384. Enichem Synth., BE 1 006 181 A4, 1993 (F. Rivetti).
385. Bayer, EP 608 710 B1, 1993 (N. Schön et al.).
386. Asahi, JP 4 211 038 A2, 1990 (S. Fukuoka).
387. Asahi, JP 4 230 242 A2, 1990 (S. Fukuoka, M. Tojo, M. Kawamura).
388. Asahi, JP 4 230 243 A2, 1990 (S. Fukuoka, M. Tojo, M. Kawamura).

389. Asahi, JP 4 224 547 A2, 1990 (S. Fukuoka, M. Tojo, M. Kawamura).
390. Asahi, JP 4 235 951 A2, 1991 (S. Fukuoka, M. Tojo, M. Kawamura).
391. Bayer, EP 572 870 A1, 1992 (P. Wagner, N. Schön, H.-J. Buysch).
392. Enichem Synth., EP 785 184 A1, 1996 (F. Rivetti, U. Romano, R. Paludetto).
393. A. Guyot et al., *Eur. Polymer. J.* **20** (1984) no. 8, 811.
394. Bayer, EP 582 930 A2, 1992 (N. Schön, H.-J. Buysch, P. Wagner, R. Langer).
395. Bayer, EP 582 931 A2, 1992 (N. Schön, R. Langer, H.-J. Buysch, P. Wagner).
396. Bayer, EP 614 877 A1, 1993 (J. Rechner, N. Schön, P. Wagner, H.-J. Buysch, S. Kabelac).
397. Mitsubishi, JP 6 345 697 A2, 1993 (H. Iwane, S. Seo, M. Tagaki, M. Inaba).
398. Bayer, EP 560 159 A1, 1992 (N. Schön, H.-J. Buysch, J. Rechner).
399. Idemitsu, JP 9 169 704 A2, 1995 (T. Fujii, T. Ishibashi, F. Yamakawa, N. Fujikawa).
400. General Electric, EPA 781 760 A1, 1995 (S. Inoki, M. Tanaka, Y. Motoyama, K. Uno).
401. Mitsubishi, JP 8 059 815 A2, 1995 (T. Okamoto, A. Hirashima, M. Mizukami, K. Hayashi).
402. Mitsubishi, JP 9 110 805 A2, 1995 (M. Inaba, K. Sawa, T. Tanaka).
403. General Electric, US 4 533 504, 1982 (D. A. Bolon, T. B. Gorczya, J. E. Hallgren).
404. Dow, WO 9 209 555 A1, 1990 (H. E. Tuinstra, C. L. Rand).
405. Mitsubishi, EP 760 359 A1, 1995 (T. Oshida, H. Oghi, H. Harada).
406. Mitsubishi, JP 9 241 219 A2, 1996 (M. Mizukami, H. Harada).
407. Mitsubishi, JP 10 036 321 A2, 1996 (M. Mizukami, H. Harada).
408. Nippon Shokubai, EP 684 221 A1, 1994 (H. Tsuneki, Y. Onda, A. Moriya, H. Yoshida).
409. Nippon Shokubai, EP 854 132 A1, 1996 (H. Yoshida et al.).
410. General Electric, US 4 452 968 A, 1983 (D. A. Bolon, J. E. Hallgren).
411. BP, EP 229 773 A, 1987 (B. Hudson).
412. Dow, US 4 973 728 A, 1987 (H. E. Tuinstra, C. L. Rand, E. P. Woo).
413. Dow, US 5 276 134 A, 1990 (H. E. Tuinstra, C. L. Rand).
414. Bayer, EP 39 452, 1980 (H. Krimm, H.-J. Buysch, H. Rudolph).
415. Asahi, JP 1 047 740 A2, 1987 (Y. Haruyuki).
416. Mitsubishi, JP 5 255 198 A2, 1992 (H. Iwane, S. Seo, T. Suzuki).
417. Asahi, EP 855 384 A1, 1996 (K. Komiya, M. Tojo, S. Fukuoka).
418. Bayer, DE 4 220 412 A1, 1992 (H.-J. Buysch, N. Schön, S. Kühling).
419. Bayer, DE 4 326 906 B1, 1993 (H.-J. Buysch, N. Schön, J. Rechner, H. Hähnsen, S. Kühling).
420. General Electric, US 5 336 814 A, 1993 (S. J. Shafer).
421. Bayer, US 4 568 755 A, 1983 (P. Mues, B. Brassat, H.-J. Buysch).
422. Bayer, EP 57 360 A2, 1982 (H. Krimm, H.-J. Buysch).
423. Bayer, EP 369 249 B1, 1988 (N. Schön, H.-J. Buysch).
424. SNAM Progetti, DE 2 222 488 B2, 1972 (G. Cipriani, E. Perotti).
425. Bayer, EP 86 358 A1, 1982 (H. Krimm, H.-J. Buysch).
426. Bayer, DE 3 418 091 A1, 1984 (W. Richter, P. Mues, H.-J. Buysch).
427. Bayer, DE 3 418 092 A1, 1984 (W. Richter, P. Mues, H.-J. Buysch).
428. Mitsui Petroch., US 4 880 942, 1987 (Y. Kiso, M. Kuga, M. Imagawa).
429. Arco, US 5 091 543 A, 1990 (A. R. Grey).
430. Henkel, WO 9 309 111 A3, 1991 (J. M. Renga, F. D. Coms, E. R Humphreys).
431. Daicel, JP 8 134 060 A2, 1 994 T Tabuchi, T. Fujii).
432. Daicel, JP 5 086 052 A2, 1991 (K. Fujiwara).
433. Bayer, EP 57 825 B1, 1981 (H. Krimm, H.-J. Buysch).
434. Mitsubishi, EP 638 541 B1, 1994 (M. Doya, K. Kimizuka, Y. Kanbara).
435. Mitsubishi, JP 7 330 756 A2, 1994 (M. Dotani, Y. Kanbara, T. Ookawa).
436. Bayer, DE-OS 3103140, 1981 (H. Krimm, H.-J. Buysch).
437. Bayer, US 4440937, 1981 (H. Krimm, H.-J. Buysch).
438. Bayer, DE-OS 3103137, 1981 (H. Krimm, H.-J. Buysch).
439. Bayer, DE 4 109 236 A1, 1991 (N. Schön, H.-J. Buysch, E. Leitz, K.-H. Ott).
440. Bayer, EP 816 354 A1, 1996 (H.-J. Buysch, G. Fengler, K.-H. Neumann, P. Wagner).
441. H. Schnell, L. Bottenbruch, *Makromol. Chem.* **57** (1962) 1.
442. D. J. Brunelle, E. Boden, T. G. Shannon, *J. Amer. Chem. Soc.* **112** (1990) no. 6, 2399.
443. D. J. Brunelle, M. F. Garbauskas, *Macromolecules* **26** (1993) no. 11, 2764.

444. S. A. Dobrowski, H. G. H. Edwards, A. F. Johnson, U. Meekum, L. Shields, *Polymer* **35** (1994) no. 22, 4707.
445. C. A. Klug, W. Zhu, K. Tasaki, J. Schäfer, *Macromolecules* **30** (1997) no. 6, 1734.
446. H. R. Kricheldorf, J. Jensen, *Eur. Polym. J.* **25** (1984) no. 13, 1273.
447. T. Takata, H. Matsuoka, T. Endo, *Chem. Lett.* **1991,** no. 12, 2091.
448. Bayer, DE 4 015 163 A1, 1990 (H. Keul et al.).
449. General Electric, US 4 767 840 A, 1987 (T. G. Shannon, D. J. Brunelle).
450. General Electric, EP 186 893 B1, 1985 (N. R. Rosenquist).
451. N. R. Rosenquist, P. L. Fontana, *Polym. Prepr. (Amer. Chem. Soc. Div. Polym. Chem.)* **30** (1989) no. 2, 577.
452. General Electric, EP 205 042 B1, 1996 (N. R. Rosenquist, T. L. Evans).
453. Registry of Toxic Effects of Chemical Substances, NIOSH, Washington, USA, 1980.
454. *Hawley's Condensed Chemical Dictionary,* van Nostrand Reinhold, 1987.
455. *Kirk-Othmer,* 4th ed., **vol. 5,** p. 77. ff.
456. R. E. Lenga (ed.): *The Sigma-Aldrich Library of Chemical Safety Data,* 2nd ed., Sigma-Aldrich Corp., 1988.
457. B. J. Gudzinowicz, *Anal. Chem.* **37** (1965) no. 8, 1051 – 1053.
458. *Manufacturing Chemist* **6B** (1992) no. 11, 36.
459. *SPC, Soap Perfumery and Cosmetics* **71** (1998) no. 3, 60. *SPC, Soap Perfumery and Cosmetics* **65** (1992) no. 12, 65.
460. US 5 342 911 A, 1993 (A. Bresciani).
461. D. Lang, H. Erdmann, R. D. Schmid, *GBF Monogr.* **17** (1992) 239.
462. D. Lang, H. Erdmann, R. D. Schmid, *Enzyme Microb. Technol.* **14** (1992) no. 6, 479.
463. F. R. Terrell et al., *J. Food. Sci.* **58** (1993) no. 5, 1132.
464. M. Bialer et al., *Eur. J. Pharm. Sci.* **2** (1994) no. 3, 239.
465. E. Samara, D. Avnir, D. Iadkani, M. Bialer, *Biopharm. Drug. Dispos.* **16** (1995) no. 3, 201.
466. Exxon, US 5 492 618 A, 1994 (G. M. Kramer, E. R. Ernst, A. M. Fraga).
467. "Jeffsol. Ethylene and Propylenecarbonate," Huntsman Technical Data Sheet, 1993.
468. "Ethylenecarbonate, Propylenecarbonate and Glycerinecarbonate," Alkylenecarbonate Seminar T. M. Huntsman Corp., Huntsman Speciality Chemicals Corp.
469. Eniricerche SpA, EP 612 829 B1, 1994 (C. Savastano, R. Cimino).
470. Technol. Finance Corp., DE 4 328 755 A1, 1993 (T. M. Makepeace, G. R. June).
471. Sanjo Electric Co, EP 582 410 A1, 1993 (M. Fujimoto, M. Takahashi, K. Nishio).
472. FMC Corp., WO 9 517 346 A1, 1993 (D. J. Salomon, D. W. Barnette, R. A. Barnett).
473. Bell. Commun. Res., WO 9 420 996 A1, 1993 (A. S. Gozdz, C. M. Schmutz, J. M. Tarascon).
474. H. Berhil, N. Lebrun, A. Tranchant, R. Messina, *J. Power Sources* **55** (1995) no. 2, 205.
475. Motorola, WO 9 837 589 A1, 1998 (G. Venugopal, A. Anani, D. Vitsas).
476. Boeing Co., US 5 098 594, 1992 (P. A. Duscher, K. Wash). BoeingCo., US 5 204 026, 1993 (P. A. Duscher).
477. Y. Sato, Y. Kodera, T. Kamo, S. Kushijama, *Prepr. Symp. – Am. Chem. Soc., Div. Fuel Chem.* **43** (1998) no. 2, 340.
478. Quantum Chem. Corp., US 5 171 874 A, 1989 (D. W. Smith, S. D. Landau).
479. Arco, US 5 206 408, 1993 (F. J. Liotta Jr.).
480. Ube, EP 728 835 A1, 1995 (T. Murayama, M. Yamashita, H. Yoshitake).
481. Exxon, EP 448 238 A, 1990 (G. Shaw, P. K. Ladwig).
482. Bayer, DE 4 213 973 A1, 1992 (P. Ooms, H.-J. Buysch, W. Beer, L. Puppe).
483. Bayer, EP 900 811 A1, 1997 (G. Burkhart, M. Klincke, A. Weier).
484. Internat. Flavors and Fragrances Inc., US 4402985, 1981 (R. M. Boden, M. Licciardello).
485. *Food Chemical News* **38** (1996) no. 35, 16.
486. Int. Flavors and Fragances, US 4 405 646, 1982 (R. M. Boden, M. Licciardello).
487. Mitsui Toatsu, JP 5 077 553 A2, 1991 (N. Nakatsuka, Y. Totani).
488. Miles, US 5 149 458, 1992 (N. H. Nodelmann).
489. ICI, US 5 112 877, 1992 (M. C. Baker et al.).
490. G. Rokicki, E. Wojciechowski, *J. Appl. Polym. Sci.* **41** (1990) 647.
491. J. Daly, A. Britten, A. Garton, P. D. McLean, *J. Appl. Polym. Sci.* **40** (1990) 845.
492. H. Pizzi, A. Stephanu, *J. Appl. Polym. Sci.* **49** (1993) 2157.
493. Foseco Intern. Ltd., US 4 416 694, 1983 (J. Stevenson, J. Machin, D. L. Dyke).
494. Symposium on Advances in Petrochemical Technology 182nd Nat. Meet. of American Chem. Soc. 25. Aug. 1981.
495. Ethylene carbonate: A Comprehensive Review of Properties, Reactions and Uses; Dow Chemicals Technical Data.
496. M. M. Mauri, U. Romano, F. Rivetti, *Ing. Chim. Ital.* **21** (1985) nos. 1 – 3, 6.

497. Y. Ono, *Pure Appl. Chem.* **68** (1996) no. 2, 367.
498. Hüls AG, Ethylencarbonat, Propylencarbonat, company brochure.
499. M. Lissel, A. R. Rohani-Dezfuli, *Kontakte (Darmstadt)* **1990,** no. 1, 20.
500. U. Romano, F. Rivetti, *Chim. Oggi.* **1984,** 37.
501. M. Selva, P. Tundo, *Green Chemistry* **1998,** 87.
502. Bayer, DE 2 523 352 A1, 1975 (H.-J. Buysch, H. Krimm, H. Rudolph).
503. Bayer, DE 2 546 534 A1, 1975 (H.-J. Buysch, H. Krimm, H. Rudolph).
504. Mitsui Petrochem., JP 2 284 918 A2, 1989 (Y. Kiso, K. Shimamoto).
505. Henkel, EP 854 853 A1, 1996 (E. Reuter, W. Knörr, B. Gutsche).
506. Henkel, WO 9 210 462 A1, 1992 (A. Westfechtel, F. Bongardt, A. Ansmann).
507. Exxon, EP 277 007 A, 1987 (R. D. Tack, R. Andrews, F. Rodger, S. J. Ayres).
508. Agip Petroli SPA, Anic SPA, EP 89 709, 1982 (R. Koch, U. Romano).
509. Chemagro Corp., US 3348939, 1964 (D. W. Gier).
510. M. Pianka, P. Sweet, *J. Sci. Food. Agric.* **19** (1968) 676 – 681.
511. M. Pianka, *J. Sci. Food Agric.* **17**(1966) 47 – 56. Fabriek van Chem. Prod., US 3198824, 1961.
512. Boehringer, DE-OS 2054225, 1970 (H. M. Becher, R. Sehring).
513. Boehringer, DE-OS 2060825, 1970 (H. M. Becher, R. Sehring).
514. PPG Industries, US 4022609, 1970 (D. E. Hardies, J. K. Rinehart).
515. Allied Signal, WO 9 304 037 A1, 1991 (H. Ramanathan, C. T. Mathews).
516. Unilever, EP 475 512 A1, 1990 (J. L. Kerschner, S. M. Jureller).
517. BP, EP 487 208 A1, 1990 (R. W. Ambler, J. N. Stewart).
518. Mitsubishi, US 5 760 273 A, 1995 (M. Inaba, K. Hasegawa, N. Kiyomi, Y. Okago).
519. Wilson Greatbatch Ltd., EP 885 874 A1, 1998 (H. Gan, M. Palazzo, E. S. Takeuchi).
520. Mitsui Petrochem., JP 2 284 919 A2, 1989 (Y. Kiso, K. Shimamoto).
521. Asahi, JP 1 047 740 A2, 1987 (H. Yoneda, J. Fukawa).
522. Company Publication of PPG Industries, Chemicals Group, One Gateway Centre, Pittsburgh, PA 15222, USA.
523. PPG Industries, US 4398008, 1981 (M. S. Misura).
524. PPG Industries, EP 80339, 1981 (C. W. Eads, J. C. Crano).
525. PPG Industries, DE 3149499, 1980 (J. C. Crano, R. L. Haynes).
526. Air Products & Chem. Inc., BE 856911, 1977 (D. D. Dixon, F. L. Herman).
527. JSP Investments, WO 9 113 052 A1, 1990 (K. C. Liu et al.).
528. Societé Francaise d'Organo Synthese, EP 274 953 A1, 1988 (Y. Proux, M. Pellegrina).
529. Anic SPA, EP 35 304 A3, 1980 (U. Romano, G. Jori).
530. L. Bottenbruch (ed.): *Kunstoffhandbuch,* **vol. 3,** Hauser 1994, p. 1.
531. S. G. Allen: *Comprehensive Polymer Science,* **vol. 5,** Pergamon Press, 1989, p. 345.
532. Bayer, EP 39 452, 1980 (H. Krimm, H.-J. Buysch, H. Rudolph).
533. Daicel, JP 2 284 917, 1989 (S. Kanno et al.).
534. General Electric, US 5 807 964 A, 1997 (S. Sivaram, A. G. Shaikh).
535. Asahi, JP 63 205 318 A, 1987 (T. Watanabe).
536. Asahi, JP 63 223 035 A, 1987 (T. Watanabe).
537. Asahi, JP Appl. 62 159 099 A, 1987 (S. Fukuoka).
538. BP, WO 8 900 590 A1, 1988 (B. Hudson).
539. Mitsui Petrochem., JP 5 255 198 A2, 1992 (H. Iwane, S. Seo, T. Suzuki).
540. BASF, DE-OS 2807762, 1978 (F. Merger, F. Towae, L. Schroff).
541. PPG Industries, EP-A 104598, 1982 (R. B. Thompson).
542. PPG Industries, EP-A 104601, 1982 (K. J. Sienkowski).
543. Bayer, DE 4 133 824 A1, 1991 (D. Arlt, H.-J. Buysch, J. Dohm, L. Puppe).
544. P. Tundo, F. Trotta, G. Moraglio, F. Ligorati, *Ind. Eng. Chem. Res.* **27** (1988) 1565.
545. P. Tundo, F. Trotta, G. Moraglio, F. Ligorati, *Ind. Eng. Chem. Res.* **28** (1989) 881.
546. Bayer, DE 4 138 245 A1, 1991 (K. Zander et al.).
547. Bayer, DE 4 306 964 A1, 1993 (G. Fennhoff, W. Jakob).
548. A. G. Shaikh, S. Sivaram, *Macromol. Rep. A* **32** (1995) no. 7, 1053.
549. UCC, EP 476 783 A1, 1990 (S. W. King).
550. Bayer, DE 3 202 090 A1, 1982 (E. Eimers, D. Margotte, R. Dhein, K. Kraft, W. Klöker).
551. Bayer, EP 410 214 A1, 1990 (H. Reiff, D. Dietrich).
552. Bayer, EP 531 673 B1, 1991 (A. Botta, H.-J. Buysch, O. Immel, L. Puppe).

553. R. Lakhmiri, P. Lohste, D. Sino, *Tetrahedron. Lett.* **30** (1989) 4669.
554. UCC, EP 478 076 A3, 1990 (S. W. King).
555. UCC, EP 476 784 A1, 1990 (S. W. King, (K. D. Olson).
556. UCC, EP 478 075 B1, 1990 (S. W. King, K. D. Olson, B. C. Ream).
557. UCC, EP 476 785 B1, 1990 (S. W. King).
558. Bayer, GB 958798, 1960.
559. Sumitomo Chem. Co., JP-Kokai 80100399, 1979; *Chem. Abstr.* **94** (1981) 653386.
560. Bayer, DE-AS 1280240, 1966 (A. Böckmann, H. Vernaleken, L. Bottenbruch, H. Rudolph, H. Schnell).
561. General Electric, EP 44 509 B1, 1980 (D. W. Fox, B. A. Kaduk).
562. Bayer, DE 4 311 424 A1, 1993 (L. Heuer, W. Jöntgen, A. Klausener).
563. E. Suzuki, M. Akiyama, Y. Ono, *J. Chem. Soc. Chem. Commun.* 1992, 176.
564. Y. Ono, M. Akiyama, E. Suzuki, *Chem. Mater.* **5** (1993) 442.
565. M. Akiyama, E. Suzuki, Y. Ono, *Inorg. Chim. Acta* **207** (1993) 259.
566. *Comeline Chem. a. Mater.* (1991) Sept. 24, 2.
567. Bayer, DE 950850, 1954 (H. Schnell, K. Raichle, W. Biedermann).
568. Texaco Dev. Corp., US 4374259, 1979.
569. Texaco Dev. Corp., GB 2092127, 1981.
570. UCC, EP 476 786 A3, 1990 (S. W. King, K. D. Olson).
571. P. Tundo, M. Selva, C. A. Marques, "Designing Chem. Eur. Green Chemistry Chap. 7," *Symp Amer. Chem. Soc., Symp. Ser. 626* (1996) 80.
572. BASF, EP 602 515 B1, 1992 (R. Fischer, R. Pinkos).
573. Kureha, EP 562 576 A1, 1992 (Y. Kanda, H. Arabori).
574. ENI, I 1 237 207, 1993 (M. Arerta, E. Quaranta).
575. Sankyo, WO 9 309 085 A1, 1991 (S. Ishihara et al.).
576. A. Hubbuch, *Kontakte (Darmstadt)* (1979) no. 3, 14 – 19.
577. S. Morrell, *Eur. Rubber J.* **164** (1982) no. 2, 39.
578. BASF, EP-A 2526, 1977 (G. Hamprecht, K. Fischer, O. Woerz).
579. Mobil Oil, WO 9 847 987 A1, 1998 (J. T. Carey, H. Oumar-Mahamat).
580. BASF, EP 50739, 1980 (V. Schwendemann, D. Mangold).
581. ENI, EP 190 466 B1, 1985 (F. Rivetti, F. Mizia, G. Garone, U. Romano).
582. Dow Chem., EP-A 65026, 1980 (A. E. Gurgiolo).
583. ENI, EP 391 473 A1, 1989 (C. Calderoni, F. Mizia, F. Rivetti, U. Romano).
584. Ube, WO 9 835 936 A1, 1997 (K. Harada, R. Sugise, K. Kashiwagi, T. Matsuura).
585. Bayer, EP-A 48371, 1980 (H.-J. Buysch, H. Krimm, W. Richter).
586. Bayer, EP-A 48927, 1980 (H.-J. Buysch, H. Krimm, W. Richter).
587. Z.-H. Fu, Y. Ono, *J. Mol. Catal.* **91** (1994) 399.
588. M. Aresta, C. Berloco, E. Quaranta, *Tetrahedron* **51** (1995) 8073.
589. Wolff Walsrode AG, EP 511 572 A1, 1991 (H.-J. Buysch, K. König, A. Klausener, K. Szablikowsky, J. Breckwoldt).
590. Asahi, JP 1 230 550 A2, 1988 (M. Seko, S. Fukuoka, T. Dosono).
591. Daicel, JP 2 311 452 A2, 1989 (T. Yanagii, K. Murata, T. Maruyama).
592. Bayer, EP 425 948 B1, 1990 (K. Müller, P. Balczinski, H. J. Santel, R. R. Schmidt).
593. Univ. North Carolina, WO 9 213 024 A1, 1992 (T. D. Dubois, F. M. Tranjan, S. M. Bobbio).
594. Bayer, EP 124 819 A1, 1983 (H.-J. Buysch, P. Mues).
595. Bayer, EP 150 380 B1, 1983 (H.-J. Buysch, P. Mues).
596. Bayer, EP 131 801 A1, 1983 (H.-J. Buysch, R. Klipper, P. M. Lange, P. Mues).
597. Bayer, EP-A 87659, 1982 (H. Krimm, H.-J. Buysch, P. M. Lange, R. Klipper).
598. General Electric, EP 187 640 B1, 1985 (N. R. Rosenquist).
599. Bayer, DE 4 109 278 A1, 1991 (H. Keul et al.).
600. Bayer, DE 4 109 279 A1, 1991 (H. Keul et al.).
601. T. Takata et al., *J. Am. Chem. Soc.* **120** (1998) no. 18, 4530.
602. *Encyclopedia of Polymer Science and Engineering,* 2nd. ed, **vol. 11,** John Wiley & Sons, New York 1988, 649.
603. Bayer, EP 50 811 B1, 1981 (H. Krimm, H.-J. Buysch, A. Botten).
604. Bayer, EP 141 950 B1, 1983 (W. Richter, H.-J. Buysch, B. Brassat, J. Merten, H. Haupt).
605. B. Wurm, H. Keul, H. Höcker, *Macromolecules* **25** (1992) 2977.
606. Bayer, EP 208 107 A1, 1985 (N. Schön et al.).
607. Bayer, US 45 687 553, 1983 (P. Mues, B. Brassat, H.-J. Buysch).
608. Bayer, EP 188 204 A1, 1985 (P. Mues, H.-J. Buysch, B. Brassat, H. Heine).
609. Ethyl Corp., US 4 423 205 A, 1982 (S. J. Rajan).

610. S. Kühling, H. Keul, H. Höcker, N. Schön, H.-J. Buysch, *Makromol. Chem.* **192** (1991) 1193.
611. S. Kühling et al., *Macromolecules* **24** (1991) 4229.
612. H. Keul, S. Kühling, H. Höcker, *Makromol. Chem.* **191** (1990) 1611.
613. W. Hovestadt, H. Keul, H. Höcker, *Makromol. Chem.* **192** (1991) 1409.
614. A. J. Müller, H. Keul, H. Höcker, *Europ. Polym. J.* **27** (1991) 1323.
615. A. J. Müller, H. Keul, H. Höcker, *Polym. Int.* **33** (1994) no. 2, 197.
616. Boehringer Ingelheim, EP 422 523 A2, 1989 (K. R. Müller, B. Buchholz, J. Hess).
617. United States Surgical Corp., EP 499 205 A3, 1992 (M. P. Chesterfield, R. R. Muth, J. Kennedy).
618. Fraunhofer Gesellschaft, EP 778 304 A3, 1995 (G. Rafler, M. Jobmann, B. Müller).
619. J. Cai, K. J. Zhu, S. L. Yang, *Polymer* **39** (1998) no. 18, 4409.
620. J. C. McNeill, A. Rincon, *Polym. Degrad. Stab.* **24** (1989) no. 3, 171.
621. A. C. Albertson, A. Loefgren, M. Sjoeling, *Makromol. Chem., Macromol. Symp.,* (1993) 73.
622. Nippon Paint, EP 687 667 A1, 1994 (T. Morimoto, S. Nakano).
623. Bayer, EP 703 230 A1, 1994 (W. Hovestadt, H.-J. Buysch, L. Schmalstieg, H. Blum, N. Schön).
624. Daicel, EP 687 697 A3, 1995 (T. Fujiwa et al.).
625. Daicel, JP 10 330 452 A2, 1997 (T. Fujiwa).
626. G. Rokicki, P. Jezewski, *Polym. J.* **20** (1988) no. 6, 499.
627. S. Inoue, *Macromol. Chem. Rapid. Commun.* **1** (1980) 775.
628. K. Soga, *J. Polym. Sci., Polym. Chem. Ed.* **17** (1979) 415.
629. Air Products a. Chem., US 4 142 021, 1977 (D. D. Dixon, M. E. Ford).
630. Air Products a. Chem., US 4 137 280, 1978 (D. D. Dixon, M. E. Ford).
631. Air Products a. Chem., US 4 104 264, 1977 (D. D. Dixon, M. E. Ford).
632. *Chem. Eng. News* (1992) Jan. 15, 26.
633. Detroid Diesel, SAE Paper 912325 (1991).
634. Union Oil, US 5 004 480, 1988 (D. D. Kanne et al.).
635. *Europa Chemie* **34** (1987) Dec. 9, 595. *Europa Chemie* **35/36** (1989) Dec. 20, 585.
636. *Chem. Market. Reporter* **236** (1989) no. 23, 4.
637. ENI Chem. Annual Rep. 1990, (1991) June 25, 68.
638. *Eur. Plast. News* (1990) May 1, 52.
639. *Jpn. Chem. Week* **31** (1990) no. 1567, 2.
640. *Chem. Market. Peporter* **242** (1992) no. 15, 3, 4, 24.
641. *Chem. Week* (1991) June 26, 32; (1992) Oct. 14, 64.
642. *Chem. Eng. News* (1992) May 4, 25; (1992) Jan. 15, 26.
643. *Chem. Week* (1992) May 20, 40; (1990) Apr. 25, 7, 8.
644. *Chem. Week* (1989) Oct. 25, 48.
645. *Comline Chem. a. Mater.* (1989) Oct. 6, 5.
646. *Chem. Market. Reporter* **246** (1994) no. 8, 12.
647. *Chemistry in Britain* **30** (1994) no. 12, 970.
648. *HAPPI* **31** (1994) no. 8, 127.
649. *China. Chem. Reporter* **9** (1998) no. 37, 18.
650. *Eur. Chem. News* **69** (1998) no. 1840, 36.
651. *Eur. Chem. News* **69** (1998) no. 1825, 27.
652. *Chem. Rundschau* **51** (1998) no. 14, 2.
653. *China. Chem. Reporter* **10** (1999) no. 9, 18.
654. UCC, US 3 234 261, 1962 (R. R. Kurkjy, M. Matzner, R. J. Cotter).

Carbon Monoxide

Synthesis gas → Gas Production

JÜRGEN BIERHALS, Messer Griesheim, Duisburg, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
3.1. Molecular Structure of Carbon Monoxide 3
3.2. Coordination with Transition Metals 3
3.3. Chemical Reactivity 3
4. Production 4
4.1. Laboratory-Scale Production 4
4.2. Industrial Production 4
4.2.1. Gasification of Coal 5
4.2.2. Steam Reforming 5
4.2.3. Partial Oxidation of Hydrocarbons . . 5
4.3. Separation of Carbon Monoxide from Gas Mixtures 5
4.3.1. Separation by Reversible Complexation 6
4.3.2. Cryogenic Separation 7
4.3.3. Pressure Swing Adsorption and Membrane Separation 8
5. Environmental Protection 9
6. Quality Specifications 10
7. Analysis 10
8. Storage and Transportation 10
9. Uses 11
9.1. Production of Inorganic Chemicals 11
9.2. Production of Organic Chemicals . 11
9.3. Application as a Merchant Gas . . . 12
10. Toxicology and Occupational Health 12
11. References 13

1. Introduction

Carbon monoxide [*630-08-0*], CO, is a toxic gas generated by the incomplete combustion of carbon-containing compounds. It was first prepared and isolated in 1776 by J. M. F. DE LASSONE by heating a mixture of zinc oxide and charcoal. However, its chemical composition was only recognized 20 years later by W. CRUIKSHANK and confirmed by F. CLEMENT and J. B. DESORMES (1801).

Considerable amounts of carbon monoxide enter the atmosphere from natural processes. According to research done in the 1970s less than 10 % comes from man-made emissions and 80 % of the estimated 3.8×10^9 t originates from oxidation of methane emitted by decaying organic matter [1]. Recent research estimates the emission from the combustion of fossil fuels at 600×10^6 t and the total emission at 2.5×10^9 t [2]. Substantial concentrations of CO can be detected in the air of major urban areas, primarily due to emissions from vehicle exhausts and domestic heating. Environmental regulations and concerns are strong incentives to reduce the amount of CO released to the atmosphere.

Carbon monoxide can also be found in abundance in the atmosphere of other planets of our solar system and in interstellar space [3], [4].

Industrial production of CO is based mainly on coal gasification or steam reforming of natural gas or petrochemicals. Major applications are as a reducing agent for the production of metals, in the production of hydrogen by the water-gas shift reaction, and for the carbonylation of organic substrates such as alcohols, amines, and esters.

Mixtures of hydrogen and carbon monoxide (syngas) are used as feedstocks of growing importance for the large-scale production of several chemicals such as methanol [*67-56-1*] and aliphatic alcohols and aldehydes (oxo synthesis). Market studies show that demand for this application will increase dramatically up to 2010 [5].

2. Physical Properties

Carbon monoxide, CO, is a colorless, odorless, tasteless, and highly flammable gas at room temperature and atmospheric pressure. In the solid state, two different allotropic phases are

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_203

Table 1. Vapor pressure versus absolute temperature for carbon monoxide∗

T, K	90.0	95.0	100.0	110.0	115.0	120.0	125.0	130.0
p, kPa	238.5	368.9	544.6	1065.4	1428.6	1874.0	2413.6	3060.9

∗ Pressure values have been calculated by using [7].

known at low temperature. Cubic α-CO exists below 61.55 K, and hexagonal β-CO above that temperature. The melting point of solid carbon monoxide is 68.15 K (at 15.35 kPa) and increases with pressure according to the following equation [6]:

$$T = 68.146 + 2.229 \times 10^{-7} p - 2.55 \times 10^{-16} p^2$$
$$\text{overtherange } p = 0.1 - 25\text{MPa}.$$

Table 2. Physical properties of carbon monoxide

Molecular mass M_r	28.011
Critical temperature	132.91 K
Critical pressure	3.4987 MPa
Boiling point (at 101.3 kPa)	81.63 K
Triple point temperature (at 15.35 kPa)	68.15 K
Phase transition point (at 3.75 kPa)	61.55 K
Density of gas (STP)	1.250 kg/m^3
Density at critical point	301.0 kg/m^3
Density of liquid (at 81.63 K)	788.6 kg/m^3
Density of solid, hexagonal (at 65 K)	929.0 kg/m^3
Density of solid, cubic (at 20 K)	1028.8 kg/m^3
Heat capacity of gas (298 K, 101.33 kPa)	
c_p	29.142 J mol^{-1} K^{-1}
c_v	20.769 J mol^{-1} K^{-1}
Heat capacity of liquid (76 K)	60.351 J mol^{-1} K^{-1}
Heat of vaporization (at 81.63 K)	6.042 kJ/mol
Heat of fusion at triple point	837.3 J/mol
Heat of sublimation at triple point	7.366 kJ/mol
Heat of transition	632.11 J/mol
Free energy of formation of gas (at 298 K), ΔG_f^0	− 137.381 kJ/mol
Enthalpy of formation of gas, ΔH_f^0	− 110.63 kJ/mol
Entropy of gas (at 298 K, 101.33 kPa), S^0	197.89 J mol^{-1} K^{-1}
Viscosity of gas (at 273 K)	16.62 μN s/m^2
Refractive index of gas (at 273 K, 546.1 nm)	1.0003364
Surface tension of liquid (at 80 K)	9.8 mN/m
Thermal conductivity of gas (STP)	23.15 mW m^{-1} K^{-1}
Thermal conductivity of liquid (at 80 K)	0.1428 W m^{-1} K^{-1}
Dielectric constant of gas (at 298 K, 101.33 kPa)	1.000634
Electric conductivity of liquid (at 85 K)	9.43 × 10^{-19} S/m
Autoignition temperature	ca. 880 K
Flammability range in air	12.5 – 74.2 vol %
Flammability range in oxygen	15.5 – 93.9 vol %

The boiling point at atmospheric pressure is 81.63 K and the vapor pressure (in MPa) of liquid CO can be calculated by using the following equation [7]:

$$\log_{10} p = A/T - B \log_{10} T + CT + D$$

where $A = 612.904$, $B = 12.968013$, $C = 0.02709872$, and $D = 29.09142$ over the range $T = 90 - 130$ K.

Selected values are given in Table 1. Other physical and thermodynamic values are presented in Table 2 [8–12]. Carbon monoxide is sparingly soluble in water and in organic solvents (Table 3) [13–15].

Table 3. Solubility of carbon monoxide in various solvents

Solvent	T, K	Solubility, vol %∗
Water	273	0.03537
Water	293	0.02319
Water	323	0.01615
Methanol	293	0.1830
Ethanol	293	0.1901
Acetic acid	293	0.1689
Acetone	293	0.2538
Toluene	293	0.1742
Carbon tetrachloride	293	0.2142
Chloroform	293	0.1897
Diethyl ether	293	0.3907
Perfluoroheptane	298	0.416

∗ Volume of gaseous CO (at 101.33 kPa) absorbed by one volume of solvent.

3. Chemical Properties

Carbon monoxide is an unsaturated molecule. Its chemistry is dominated by its reducing properties and its ability to combine with a large variety of molecules.

3.1. Molecular Structure of Carbon Monoxide

Carbon monoxide is isoelectronic with N_2 and the predicted electronic configuration of the valence shell is:

$$\left(\sigma_{\mathrm{s}}^{\mathrm{b}}\right)^2 (\sigma_{\mathrm{s}}^{*})^2 \left(\pi_{\mathrm{x,y}}^{\mathrm{b}}\right)^4 \left(\sigma_{\mathrm{z}}^{\mathrm{b}}\right)^2$$

The energy levels of the molecular orbitals are shown in Figure 1. There are three bonds between the carbon and oxygen atoms: one sigma bond and two pi bonds. A very short C – O bond length 112.82 pm, a large bond dissociation energy 1.070 MJ/mol [16], and a high infrared stretching frequency $\tilde{\nu}_{\mathrm{CO}} = 2143.274\,\mathrm{cm}^{-1}$ (gas), $\tilde{\nu}_{\mathrm{CO}} = 2138.44\,\mathrm{cm}^{-1}$ (solid, 20 K) [17] are consistent with a carbon – oxygen triple bond. The first ionization potential is 1.351 MJ/mol (14 eV).

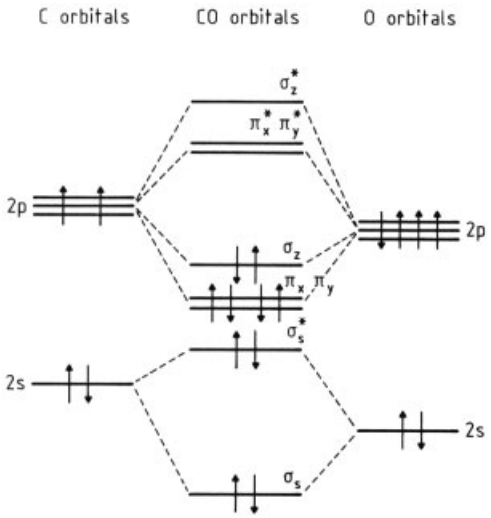

Figure 1. Simple molecular orbital diagram for CO [16]

3.2. Coordination with Transition Metals

Carbon monoxide is an extraordinarily widely used ligand (carbonyl) of utmost importance in coordination chemistry [18]. Synergic bonding occurs due to the slight donor character of the σ_z bonding orbital (σ donation from the carbonyl to an empty orbital of the metal having the correct symmetry) and the presence of two empty low-lying $\pi_{x,\,y}*$ antibonding orbitals which may act as strong π acceptors (π back donation from a filled π orbital of the metal), as depicted in Figure 2. The net effect is an electron density transfer from the σ_z bonding orbital to the $\pi_{x,\,y}*$ antibonding orbitals of the carbonyl ligand that results in an increase of the C–O bond length and a lowering of the infrared carbonyl stretching frequency compared with that of free carbon monoxide. Moreover, the stabilization of the $\pi_{x,\,y}*$ antibonding orbitals of the carbonyl ligand and prepare it for nucleophilic attack; this is an essential step in the activation of carbon monoxide by transition metals. Coordination can occur on a single metal center (monodentate carbonyl) with a linear or quasilinear M – C – O geometry (**1**) (Fig. 3). Numerous complexes having bridging (**2**) or capping carbonyl ligands(**3**) have been described and play a very important role in polymetallic cluster chemistry [19–21], as well as in the chemisorption of carbon monoxide on surfaces [22]. Carbonyl ligands have also been used extensively as internal probes for electronic effects in transition-metal complexes because their infrared stretching frequencies are highly sensitive to the charge density on the metal [23].

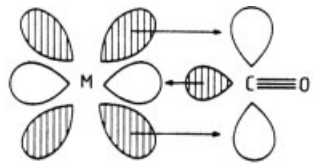

Figure 2. Synergic binding of CO to transition metals Arrows indicate the electron donation from filled (hatched) to empty orbitals

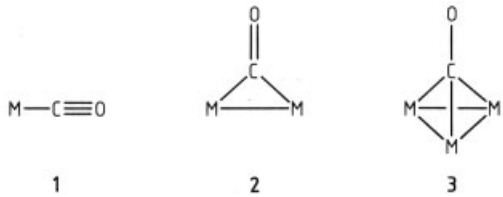

Figure 3. Some common geometries of CO coordination with transition metals

3.3. Chemical Reactivity

Carbon monoxide is a metastable molecule. It is almost inert under mild conditions, but it be-

comes very reactive when subjected to high pressure or elevated temperature, and it can be easily activated by catalysts. Carbon monoxide disproportionates to form carbon and carbon dioxide (Boudouard equilibrium).

$$2\,CO \rightleftharpoons CO_2 + C$$

$$\Delta H^0_{298} = -172.47\,\text{kJ/mol};\ \Delta G^0_{298} = -120.02\,\text{kJ/mol}$$

At atmospheric pressure, the extent of carbon monoxide disproportionation is very low below about 670 K, the temperature at which the thermodynamic equilibrium is almost completely displaced toward carbon and carbon dioxide, whereas less than 1 % of carbon dioxide is observed at equilibrium above 1250 K.

This disproportionation reaction, which is favored by increased pressure, is catalyzed by a large number of metals or metal oxides and is responsible for the undesirable formation of carbon deposits in places such as blast furnaces.

Carbon monoxide is extremely flammable. It has a large flammability range (see Table 2) and can form explosive mixtures in air. Carbon monoxide burns with a blue flame.

$$CO + 0.5\,O_2 \rightarrow CO_2$$

$$\Delta H^0_{298} = -282.98\,\text{kJ/mol};\ \Delta G^0_{298} = -257.19\,\text{kJ/mol}$$

It may also react violently with oxidants. Hopcalite, a mixture of the oxides of manganese, copper, cobalt, and silver, has been used in gas-mask elements to oxidize carbon monoxide. Also important are ceramic-supported transition metals of Groups 8 – 10, especially palladium, platinum, and rhodium, which are used as catalytic converters to reduce carbon monoxide emissions in vehicle exhaust [24].

The carbon monoxide reactions of greatest industrial importance are described in Chapter 9.

4. Production

Carbon monoxide is formed by the incomplete combustion of carbonaceous materials or by the reduction of carbon dioxide [25], [26]. It can also be formed by the decomposition of organic compounds such as aldehydes [27]. Carbon monoxide can also be recovered from the off-gas of several industrial processes such as blast furnace processes or calcium carbide synthesis.

4.1. Laboratory-Scale Production

Although readily commercially available in high purity, carbon monoxide can be prepared by the slow addition of concentrated formic acid (90 %, $d = 1.2$) to concentrated sulfuric acid at about 350 K [28], [29]. Traces of sulfur dioxide and carbon dioxide can be removed by passing the gas through a column filled with potassium hydroxide pellets.

4.2. Industrial Production

Incomplete combustion of carbonaceous materials is commonly employed for the manufacture of commercial quantities of carbon monoxide. The only process that can produce carbon monoxide with a very low content of hydrogen and nitrogen is the Otto process, which is an interesting possibility for smaller plants [30]. In most production processes the initial product is a gas mixture containing carbon monoxide. The most important processes are:

1) Gasification of coal
2) Steam reforming/CO_2 reforming (for light hydrocarbons up to naphtha)
3) Partial oxidation of hydrocarbons (for hydrocarbons heavier than naphtha)

To obtain the pure gases, the mixtures must be separated. The main gas that is produced together with carbon monoxide is hydrogen. Many applications need a synthesis gas that is a combination of carbon monoxide with hydrogen. Therefore the separation of carbon monoxide or the adjustment of the carbon monoxide hydrogen ratio are the following production steps. The CO can be separated by various procedures:

1) Reversible complexation (copper ammonium salt wash) at elevated pressure, followed by desorption by pressure release
2) Cryogenic separation
   Low-temperature partial condensation and fractionation
   Liquid methane scrubbing and separation
3) Pressure-swing adsorption
4) Permeable membranes

4.2.1. Gasification of Coal

Coal gasification is the term for several methods for the conversion of coal into gaseous reaction products in generators of different construction.

Coal is gasified at an elevated temperature when reacted with steam and air or preferably oxygen [31–39]. The main reaction is the endothermic water – gas reaction:

$$C + H_2O \rightleftharpoons CO + H_2 \quad \Delta H^0_{298} = +131.30\,\text{kJ/mol}$$

This is coupled with exothermic carbon combustion:

$$C + O_2 \rightleftharpoons CO_2 \quad \Delta H^0_{298} = -394.36\,\text{kJ/mol}$$

Several other reactions are involved in the process. Among these are the previously mentioned Boudouard equilibrium and the subsequent water-gas shift reaction.

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad \Delta H^0_{298} = -41.16\,\text{kJ/mol}$$

The composition of the raw gas is highly dependent on the C : H_2O : O_2 ratio, the temperature, and the nature of the coal feedstock.

Three types of coal gasifiers, using either fixed-bed, fluidized-bed, or entrained-bed technology, have been operated commercially for many years (Table 4). A Texaco unit (entrained-bed, operating pressure of up to 8 MPa) started operating in 1983 consuming coal at a rate of 900 t/d. Several new gasifiers have been developed by various companies, including Shell and Dow Chemical [40–48]. The feasibility of the gasification of low-grade fossil fuels such as shale oil and tar sands has also been investigated [49].

4.2.2. Steam Reforming

Carbon monoxide can also be produced by steam treatment of natural gas or light petroleum fractions (naphtha) [50–52].

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2 \quad \Delta H^0_{298} = +206.15\,\text{kJ/mol}$$

$$\text{or } C_n H_2{}^{n+2} + n\,H_2O \rightleftharpoons n\,CO + (2n+1)\,H_2$$

These highly endothermic reactions are usually carried out at 1000 – 1100 K with external heating, under a pressure of 1.5 – 4 MPa in the presence of a catalyst. The catalyst is nickel promoted with K_2O, deposited on a refractory support (e.g., ICI process). Sulfur-free feedstocks (less than 0.2 mg/kg) are required to avoid poisoning of the catalyst.

The reaction of a hydrocarbon with carbon dioxide is analogous to steam reforming with water vapor. Therefore, it is often called CO_2 reforming. The reformer fuel, for example LPG, together with carbon dioxide is exposed to a high temperature in the presence of a catalyst and is thereby transformed by an endothermic reaction into a mixture of carbon dioxide, water, hydrogen and carbon monoxide. The carbon dioxide removed from the raw synthesis gas at the outlet of the reformer can be recycled for the reforming process; thus, an H_2/CO ratio of 3/1 can be achieved. By adding imported CO_2 a further decrease of the H_2/CO ratio is possible. In practice the lower limit is 1.2 [53].

4.2.3. Partial Oxidation of Hydrocarbons

Hydrocarbons, e.g., methane, can be partially oxidized by using a limited amount of oxygen.

$$CH_4 + 1.5\,O_2 \rightleftharpoons CO + 2\,H_2O \quad \Delta H^0_{298} = -519.33\,\text{kJ/mol}$$

Such reactions are performed at 1500 – 1800 K and 3 – 8 MPa without a catalyst.

Several processes have been designed (BASF-Lurgi, Texaco, Hydrocarbon Research, Shell, etc.) to use petroleum feedstocks ranging from natural gas to asphalt [54].

4.3. Separation of Carbon Monoxide from Gas Mixtures

The previously mentioned preparations and the recovery of CO from various industrial processes (blast furnace processes, acetylene synthesis, etc.) yield gas mixtures containing 30 – 85 % of carbon monoxide. Usually a preliminary purification step is performed to remove acidic contaminants (CO_2, COS, H_2S, mercaptides) by using one of the numerous basic washing processes [55]. Pure carbon monoxide is obtained either by reversible complexation or cryogenic separation.

Table 4. Characteristics of commercially operated gasifiers

Parameter	Gasification process		
	Lurgi (fixed- bed)	Winckler (fluidized- bed)	Koppers- Totzek (entrained-bed)
Feed size, mm	6 – 40	1	< 0.1
H_2O/O_2 ratio	5 – 6	1.5 – 2	0.3
Gasification temperature, K	900 – 1200	1100 – 1400	1700 – 2200
Dry raw gas composition∗			
CO, %	18 – 24	30 – 50	55 – 66
H_2, %	37 – 40	35 – 46	21 – 32
CO_2, %	25 – 32	13 – 25	7 – 12
CH_4, %	9 – 12	1 – 18	0.1
N_2, %	0.5 – 1	0.5 – 1.5	1

∗ Dependent on the nature of the feed.

4.3.1. Separation by Reversible Complexation

Copper(I) salt solutions are able to absorb carbon monoxide to form carbonyl complexes [56]. This property is used in two industrial processes to selectively separate carbon monoxide from a mixture of gases. The copper – ammonium salt wash developed by BASF uses aqueous ammonia solutions of copper carbonate, formate, acetate, lactate, or a mixture thereof [57]. The absorption – desorption reaction is thought to be an equilibrium between free and coordinated carbon monoxide that is highly dependent on the temperature, pressure, and acidity of the medium.

$$[Cu(NH_3)_2]^+ + CO + NH_3(aq) \rightleftharpoons [Cu(NH_3)_3(CO)]^+$$

The solution composition is adjusted to ensure maximum solubility of copper(I) and thereby increase the productivity (volume of carbon monoxide absorbed per unit volume of solution).

A simplified flow diagram of the copper – ammonium salt wash process is shown in Figure 4.

The gas containing carbon monoxide is contacted with the copper – ammonium salt solution in countercurrent at 283 – 303 K and a pressure of up to 20 MPa in a tower packed with Raschig rings. Carbon monoxide, which can be extracted down to a few mg/kg, is recovered in the regenerator by heating the rich solution to 350 K, usually at atmospheric pressure. The vapor pressure of ammonia at this temperature is significant and a scrubber can be used to wash the off-gas.

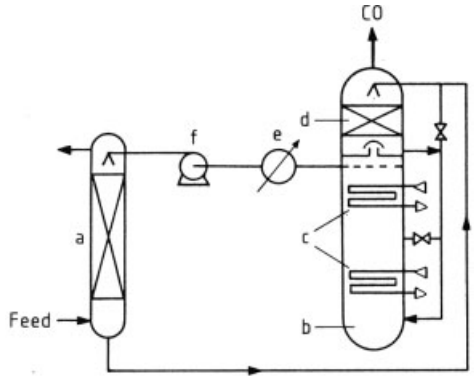

Figure 4. Simplified flow diagram of the copper – ammonium salt wash process
a) Absorption tower; b) Regenerator; c) Heaters; d) Ammonia scrubber; e) Cooler; f) Lean-solution recycle pump

The Cosorb process, developed in the early 1970s by Tenneco Chemicals, is based on a similar principle (Fig. 5) [58–60]. However, the use of copper(I) organometallic complexes in organic solvents (preferably $CuAlCl_4$ in toluene) as carbon monoxide carriers provides several advantages such as lower corrosion, higher complex stability, and higher purity of the gas produced. Nevertheless, several impurities such as water or sulfur compounds (SO_2, H_2S, COS, etc.) destroy the copper complex. A very careful purification of the feed gas is therefore required.

The mixture of gases is washed with a countercurrent flow of the copper complex solution in the absorber (on a packed column or in trays) at a moderate temperature of 303 – 318 K and a pressure of 0.7 – 3 MPa. The rich solution is then

decompressed to release physically dissolved gases, e.g., H_2, N_2, which can be either recycled in the absorber or eliminated. The carbon monoxide is stripped in the regenerator by lowering the pressure to ca. 0.15 MPa and heating the solution to 408 K. Solvent losses in the absorber and regenerator overhead gas flows are minimized by using a combination of compression, refrigeration, and activated charcoal recovery. Carbon monoxide can be obtained with a purity up to 99.7 % and yield of 95 – 99 %.

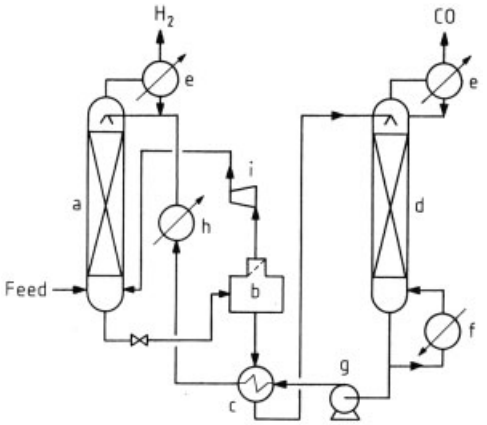

Figure 5. Simplified flow diagram for the Cosorb process
a) Absorption tower; b) Liquid – gas separator; c) Heat exchanger; d) Regenerator; e) Reflux condenser; f) Reboiler; g) Lean-solution recycle pump; h) Cooler; i) Compressor

Some new CO-complexing compounds and liquid absorption processes are under investigation [61].

4.3.2. Cryogenic Separation

Processes for the cryogenic separation of pure carbon monoxide from a mixture of gases have been developed by, among others, l'Air Liquide [62] and Linde [63]. The feed gas must be carefully purified to eliminate any traces of contaminants such as water and carbon dioxide that would freeze in the unit and thus cause clogging problems. This is usually achieved by pretreatment with molecular sieves. Carbon monoxide and nitrogen are difficult to separate by distillation, and feed gases containing nitrogen are therefore not suitable for production of high-purity carbon monoxide unless a nitrogen – carbon monoxide separation device is included in the process.

The *low-temperature partial-condensation process* (Fig. 6) utilizes the difference in boiling points between carbon monoxide and hydrogen. The purified mixture of gases is cooled to approximately 80 K at 3 – 4 MPa by means of heat exchange with the exhaust gases. Carbon monoxide and methane, which may possibly be present in the feed, condense and are recovered as liquid in the first separator. In the second separator, which is operated at a lower pressure, dissolved hydrogen is released, recovered overhead, and recycled to the feed gas after recompression. The liquid phase, rich in carbon monoxide, is evaporated in the heat exchanger to yield gaseous carbon monoxide of 98 – 99 % purity ($CO + N_2 > 99.5$ %). If the feed contains too much methane for the required purity, it is possible to introduce a carbon monoxide – methane separation column after the second separator to obtain a final concentration of methane as low as 0.1 % in the product gas.

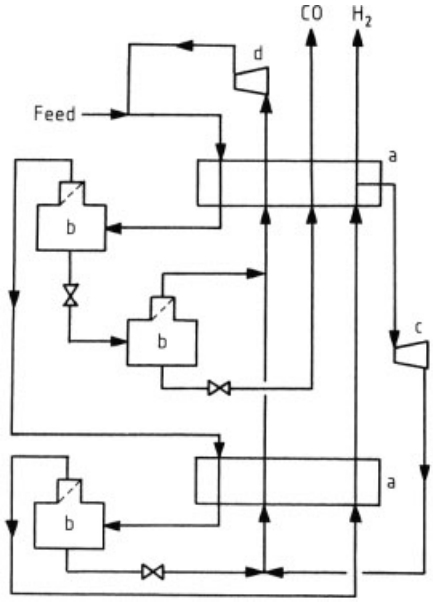

Figure 6. Simplified flow diagram for the partial condensation process for obtaining carbon monoxide and hydrogen
a) Heat exchanger; b) Liquid – gas separator; c) Hydrogen expansion turbine; d) Recycle compressor

The hydrogen-enriched phase collected overhead from the first separator is further cooled to ca. 70 K to condense more carbon monoxide. At this stage the gas phase contains 96 – 98 % pure hydrogen, which is passed through the heat ex-

changer to cool the incoming flow of impure hydrogen. The necessary refrigeration is provided by the decompression of a part of the hydrogen in an expansion turbine and the vaporization of the carbon monoxide rich liquid phase coming out of the third separator. The cryogenic part, usually referred to as the cold box, includes heat exchangers, phase separators, and, in some cases, fractionating columns. Perlite or rock wool is used to ensure efficient thermal insulation.

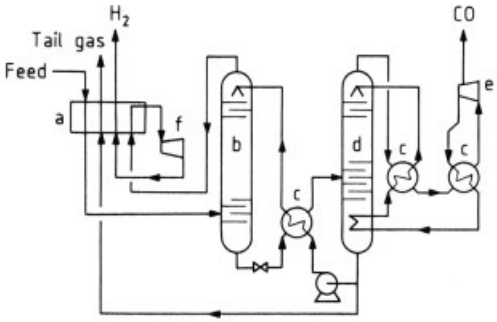

Figure 7. Simplified flow diagram for the methane wash process for obtaining carbon monoxide and hydrogen
a) Heat exchanger; b) Methane absorption tower; c) Heat exchanger; d) Regenerator; e) Carbon monoxide recycle compressor; f) Hydrogen expansion turbine

The *liquid methane wash process* (Fig. 7) takes advantage of the considerable solubility of carbon monoxide in cooled liquid methane. The feed gas at a pressure of about 4 MPa is cooled to 92 K in a heat exchanger to condense most of the carbon monoxide and methane. The gas phase, after countercurrent scrubbing with liquid methane to extract carbon monoxide (to a few mg/kg), is collected overhead and decompressed to ca. 1.5 MPa in an expansion turbine. This provides the refrigeration for the incoming gas flow, which contains 97 – 99 % pure hydrogen contaminated with small amounts of methane (1 – 1.5 %). The condensed phase from the scrubber is fed after partial decompression to a carbon monoxide – methane separation column operating at 0.1 – 0.2 MPa. Liquid methane is recovered at the bottom of the column and recompressed. Part of it is vented as a low-pressure fuel to ensure the cooling of the primary heat exchanger. The main stream is recycled through the top of the scrubber. Carbon monoxide is recovered overhead from the separation column with an excellent yield of 97 – 99 % and circulated to provide cryogenic power for the unit. Its purity is about 97 – 99 % ($CO + N_2 > 99.8$ %). Hydrogen is the main impurity along with a small amount of methane.

4.3.3. Pressure Swing Adsorption and Membrane Separation

Pressure swing adsorption (PSA) and membrane separation are used for lower production capacities if CO purities up to 99 % are required. The attainable CO content depends very much on the raw gas composition and is affected by the presence of nitrogen. Both processes are more suitable for recovery of pure hydrogen than for CO. Pressure swing adsorption is used today almost exclusively in new steam reforming hydrogen plants.

Molecular sieve adsorbents for the removal of specific components from a gas mixture are selected on the basis of their relative adsorptive power. The PSA process uses multiple adsorbent beds to provide continuous flow under constant-pressure conditions. Molecular sieves adsorb preferentially carbon monoxide, carbon dioxide, methane, and nitrogen from the raw synthesis gas. Other CO-selective adsorbents have been investigated [61]. In the PSA process the absorber swings between high pressure for absorption and low pressure for desorption.

An excellent solution for the simultaneous production of carbon monoxide and hydrogen is a combination of cryogenic partial condensation followed by a PSA unit [64].

Two commercial PSA plants have been in operation in Japan since 1989 for the recovery of CO from converter gases of steel plants [65].

Depending on the use of the syngas the desired composition ranges from 100 % hydrogen (for the production of hydrogen or ammonia) to 100 % CO for carbon monoxide or acetic acid production.

The first application of membrane processes in syngas separation was the adjustment of carbon monoxide to hydrogen ratio. Monsanto has developed a partial separation process that uses silicon-coated polysulfone hollow-fiber gas permeators (trade name: PRISM) [66–72]. Each gas permeator consists of an array of 10 000 – 100 000 hollow fibers embedded in a permeable and nonselective matrix. Hydrogen diffuses faster, permeates the fibers, and is re-

covered at low pressure, whereas the adjusted $CO-H_2$ mixture is obtained as the high-pressure residual gas (Fig. 8).

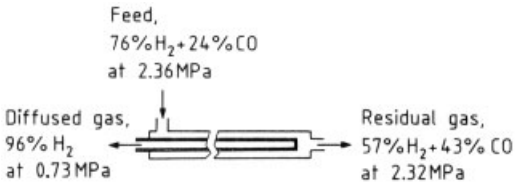

Figure 8. Principle of hollow-fiber gas separators (each contains 10 000 – 100 000 hollow fibers)
Data are taken from ref. [67]

For production of pure CO from natural gas reformed syngas a two-stage membrane system has been demonstrated to yield 98.9 % purity.

For both syngas ratio adjustment and CO production cellulose acetate was found to provide higher selectivity and permeability compared to a polysulfone membrane under the same operating conditions.

5. Environmental Protection

The concentration of CO in the atmosphere in the northern hemisphere is currently about 140 ppb ($\mu L/m^3$), compared to a pre-industrial concentration of 90 ppb [2]. Model calculations predict an approximate doubling of the CO concentration by 2100 [73].

According to research done in the 1970s, 3.0×10^9 t of the 3.77×10^9 t in the atmosphere originate from oxidation of methane emitted by decaying organic matter and only 270×10^6 t from man-made emissions [1]. Traces of CO are emitted during composting of grass and green waste. Quantitative measurements have been performed [74]. Recent estimates of the emission from the combustion of fossil fuels are around 600×10^6 t [2].

The lifetime of the CO molecule in the atmosphere is currently (2000) two months and is estimated to increase to 2.6 months by 2100 [73]. The lifetime-limiting reaction in the troposphere is [75]:

$$CO + OH \longrightarrow CO_2 + H$$

An increase in CO concentration is likely to cause a decrease in OH and O_3 in the troposphere, and this would result in extensive chemical feedback. The global budget and trends of CO are therefore of considerable environmental importance.

Air pollution by carbon monoxide is restricted almost entirely to urban and highly industrialized areas. CO plays an important role in photochemistry of regional and urban environments where, depending on the concentration of nitrogen oxides and hydrocarbons, CO can produce ozone, which is implicated in photochemical smog and haze.

During the past years all leading industrialized countries have significantly intensified their efforts to lower air pollution. (e.g., Clean Air Act amendments of 1970, 1977 and 1990 in the USA, TA Luft in Germany). One of the consequences is the substantially improved exhaust gas purification for vehicles by catalytic converters. By 1995 over 85 % of all new gasoline-powered cars sold worldwide had either a three-way or oxidation-only catalytic converter [76]. Today in urban main roads CO levels of about 5 ppm can be observed. Inside the motor vehicle the mean level of CO is about 20 ppm [77].

In addition, households and small-scale fuel consumers switched to liquid and gaseous fuels with considerably better emission values, and for industrial combustion units improved burner designs were developed.

The success of emission control in the industrial countries is exemplified for Germany and the United States. In Germany carbon monoxide emissions increased until the beginning of the 1970s. Since then a continual decrease was observed. From 1990 to 1994 the emission of CO decreased from 10.743×10^6 t/a to 6.738×10^6 t/a. The contribution of traffic decreased slightly from 60.4 % to 58.7 %. But despite the fall in absolute emissions, traffic remains the most significant cause of CO pollution.

In the USA nationwide emissions of CO from on-road vehicles fell by 33.4 % from 1970 to 1995. Since 1981 each car sold outside the state of California emitted only 2.1 g/km. Prior to 1968 each new passenger car emitted 48 g/km [76].

According to the EC Exhaust Emission Directive 94/12/EC (phase 2) the emission for gasoline engines is set at 2.2 g/km (new type approval from 1.1.96) From January 2005 (phase 4) the emission will be limited to 1.0 g/km

[78]. Based on the EC guidelines for the reduction of emissions of motor vehicles, the German Umweltbundesamt (Federal Department for the Environment) has calculated a reduction in CO emissions of 56 % between 1988 and 2005.

6. Quality Specifications

Carbon monoxide of high purity is available in cylinders as the compressed gas (10 – 20 MPa) in volumes of 0.1 – 9.5 m^3 (STP). No standard specifications have been issued by the Compressed Gas Association, and each supplier uses its own quality label. Carbon monoxide is sold in different purities [79–82]:

Commercial (technical, standard) grade 98 – 99 %
C. P. grade 99.3 – 99.5 %
Ultrahigh-purity grade 99.8 %
Research grade 99.99 – 99.997 %

Contaminants include primarily nitrogen, hydrogen, methane, oxygen, water, carbon dioxide, and argon. Carbonyls are normally not specified. If iron or nickel carbonyls are of special concern, aluminum cylinders and brass valves are better suited than carbon steel cylinders with stainless steel valves.

7. Analysis

Environmental pollution control has been an incentive to develop fast and reliable methods for the determination of traces of carbon monoxide. Nondispersive infrared spectrometry takes advantage of the strong absorption of CO at 4.7 µm. The lower limit of detection is about 0.5 mg/kg. Water vapor and, to a lesser extent, CO_2 interfere [83]. Semicontinuous measurement of carbon monoxide can be carried out by using gas chromatography with different detectors, depending on the required detection limit [84]. A reduction gas detector (RGD) is especially suited for the detection of traces of CO. It is specific to gases that can reduce a bed of mercury oxide. The released mercury can be detected by ultraviolet absorption at ppb levels (DIN 50450-7, 1995). Another common detection method uses the conversion of carbon monoxide into methane with a catalyst (DIN 50450-6, 1994). An absolute method, which is independent of calibration gases, is complete oxidation with I_2O_5 and titration of the elemental iodine (VDI 2459, Blatt 7, 1994). A turbidimetric method can determine 2 – 1800 mg/kg CO [85] by reducing a solution of a silver salt of *p*-sulfaminobenzoic acid to a colloidal silver derivative; absorbance is measured at 425 or 600 nm. No interferences are known for this method. Other analytical techniques include catalytically oxidizing carbon monoxide to carbon dioxide by using Hopcalite, platinum-group metals, or miscellaneous catalysts and monitoring the heat of reaction. The colorimetric tubes in some portable analyzers use the wet oxidation with I_2O_5, whereby a reactive adsorbent turns greenish brown as it adsorbs the iodine formed. Semiquantitative analysis can be performed by using a solution of palladium chloride, which is reduced by carbon monoxide to a black precipitate of metallic palladium.

8. Storage and Transportation

The standard quality of carbon monoxide has considerably improved during the last decades, especially in regard to moisture and sulfur-containing impurities. Dry, sulfur-free carbon monoxide is satisfactorily compatible with steel and other commonly used metals at pressures up to 15 MPa or even 20 MPa (these filling pressures are used for commercially available carbon monoxide in gas cylinders). Moisture and sulfur-containing impurities in carbon monoxide increase the corrosion of steel at all temperatures. Only highly alloyed chrome steels are sufficiently resistant to corrosion by CO containing these impurities. Iron, nickel and other metals can react even at room temperature with carbon monoxide at elevated pressures to form carbonyls in small quantities. Copper and copper – silver alloys are the only materials suitable for applications at high temperature and pressure. Carbon monoxide is also compatible with synthetic organic materials such as polyethylene, poly(vinyl chloride), Kel F, Teflon, and Viton. However, the use of rubber or Neoprene is forbidden.

Carbon monoxide is produced or recovered almost exclusively for on-site uses. Although available in the United States in bulk quantities

as a liquified gas, it is much more commonly transported as a compressed gas in tube trailers or cylinders (steel or, more frequently, aluminum or its alloys). In most countries, the cylinders must be recertified every five years.

There is no color international standardization for the marking of cylinders. In the European community there is a new standard (EN 1089-3) for the color marking of cylinders. Carbon monoxide container heads are painted with yellow, or yellow and red. The color of the cylindrical part is generally not specified. In Germany the gas industry agreed to use gray or the new or the old color of the shoulder. For CO this means that the cylindrical part is painted in gray or red.

Transportation of carbon monoxide is subject to a number of regulations [86–89]. According to transport regulations in Europe the containers must bear labels for flammable and toxic substances (with a skull and cross-bones). Since October 1999 in the US the DOT requires the labels "Flammable Gas" and "Inhalation Hazard" (with the poison symbol).

9. Uses

Carbon monoxide, as a pure compound or as a constituent of syngas ($CO + H_2$), is increasingly being used on a very large scale for the production of chemical intermediates.

9.1. Production of Inorganic Chemicals

As a component in the raw gas mixture after gasification carbon monoxide is used in the production of hydrogen in the water-gas shift reaction (→ Hydrogen, Chap. 4.1.).

Carbon monoxide is one of the reducing agents acting on oxides in blast furnaces. Direct reduction of iron ore to sponge iron by using syngas as an alternative technology has made rapid progress [90]. Some transition metals react under pressure with carbon monoxide to yield metal carbonyls such as $[Ni(CO)_4]$, $[Fe(CO)_5]$, $[M(CO)_6]$, (where M = Cr, Mo, or W), $[Co_2(CO)_8]$, etc. These are used as precursors for making homogeneous catalysts or for the preparation of very pure metals; e.g., the extractive metallurgy of nickel uses $[Ni(CO)_4]$ as an intermediate in the carbonyl refining processes (→ Nickel, Chap. 6.3.).

The reaction of carbon monoxide with chlorine, catalyzed by activated charcoal, yields phosgene, an intermediate in the synthesis of isocyanates.

$$CO + Cl_2 \rightarrow COCl_2$$

Phosgene or a mixture of carbon monoxide and chlorine is used to prepare aluminum chloride by the chlorination of bauxite [91].

9.2. Production of Organic Chemicals

Carbon monoxide is widely used in the production of organic chemicals [92–105].

From Pure Carbon Monoxide. The major use of pure carbon monoxide is in the preparation of acetic acid [*64-19-7*] by catalytic carbonylation of methanol [*67-56-1*]. This consumes more than 500 000 t/a.

$$CH_3OH + CO \rightarrow CH_3COOH$$

Cobalt promoted with iodides is used as the catalyst in the BASF process (470 – 520 K, up to 70 MPa, 90 % selectivity based on CH_3OH). The rhodium-based Monsanto process requires much less severe conditions (440 – 450 K, 1 – 4 MPa, 99 % selectivity) [106]. Processes using cheaper catalysts such as nickel (Rhône-Poulenc, Mitsubishi-Gas, Halcon) have been tested at the pilot-plant stage. Analogously, a plant that uses the carbonylation of methyl acetate [*79-20-9*] (Eastman Kodak-Halcon) is producing acetic anhydride [*108-24-7*] at the rate of 250 kt/a [107]. Other uses of pure carbon monoxide include the production of formic acid [*64-18-6*] and methyl formate [*107-31-3*] in the presence of a basic catalyst (350 – 420 K, 1 – 20 MPa).

$$CO + ROH \rightarrow HCOOR \quad R = H, CH_3$$

N,N-Dimethylformamide [*68-12-2*] is prepared from dimethylamine [*124-40-3*] under similar conditions.

$$CO + HN(CH_3)_2 \longrightarrow HCON(CH_3)_2$$

Acrylic acid [*79-10-7*] and propanoic acid [*79-09-4*] are also prepared by the hydroxy carbonylation (Reppe carbonylation) of acetylene [*74-86-2*] and ethylene [*74-85-1*], respectively, in the presence of a nickel catalyst (BASF).

$$HC \equiv CH + CO + H_2O \xrightarrow[4-5\ MPa]{450-470\ K} CH_2 = CH-COOH$$

and

$$CH_2 = CH_2 + CO + H_2O \xrightarrow[20-25\ MPa]{540-590\ K} CH_3-CH_2-COOH$$

From Syngas. A large variety of chemicals, ranging from saturated hydrocarbons to oxygenated compounds, are produced using syngas as a feedstock. The most important applications of syngas are in the production of

- Methanol (→ Methanol, Chap. 5.)
- Hydrocarbons (→ Coal Liquefaction, Chap. 2.2., → Waxes, Chap. 5.)
- Linear aliphatic aldehydes (→ Oxo Synthesis)

9.3. Application as a Merchant Gas

Most carbon monoxide is used on site directly after production. That which is sold as a merchant gas must be filled, transported, and for many applications also purified [108].

The demand for carbon monoxide as a merchant gas has grown strongly in the last few years. Applications include various test runs (e.g., turbines) with a high consumption for a limited time and smaller demands for industrial applications and research projects. Purities of up to 99.999 vol % are required for special purposes. In some cases, carbonyls must be removed by special purification techniques [109].

Uses for cylinder gases include:

- As a plasma etching gas in the electronic industry for selective oxide etching [109]
- As a component in calibration gas mixtures for environmental monitoring systems and for vehicle exhaust measuring systems
- In medicine as a component in gas mixtures for lung function tests
- As a component in CO_2 laser gas mixtures

10. Toxicology and Occupational Health

Carbon monoxide is a particularly dangerous chemical because it cannot be detected by the natural senses of the body [110], [111]. It is toxic because it competes successfully with oxygen for the binding sites of hemoglobin (Hb), the oxygen-carrying hemoprotein in the blood of mammals.

$$Hb + O_2 \rightleftharpoons HbO_2$$

$$Hb + CO \rightleftharpoons HbCO$$

The affinity of hemoglobin for carbon monoxide, calculated from the pressure of CO required for half-saturation, i.e., [Hb] = [HbCO], is 200 – 300 times higher than the affinity for oxygen [112]. The noxious effect of carbon monoxide is thus to deprive tissues of their required oxygen supply by hindering the oxygen-carrying activity of hemoglobin. Other hemoproteins can also be affected by carbon monoxide. Myoglobin, responsible for the storage of oxygen in muscles, is one example. Cytochrome P 450, the widespread monooxygenase involved in the metabolism of several essential biomolecules (e.g., steroids) and used for the detoxification of exogenous compounds by oxidative degradation, is another. At equilibrium the percentage of carboxyhemoglobin in blood can be estimated. The percentage of HbCO is 0.16 times the concentration of CO (expressed in mg/kg) [113]. However, equilibrium is reached slowly, requiring 5 – 10 h, depending on the physical activity of the subject. Low levels of carboxyhemoglobin are always present in the blood due to the endogenous production of carbon monoxide (ca. 0.45 mg/h) by the oxidative catabolism of heme proteins. Rarely are physiological effects observed if the proportion of carboxyhemoglobin is below 10 % (the HbCO level is frequently 4 – 5 % for smokers). The first toxic symptoms, a slight headache and shortness of breath, usually appear at about 20 % HbCO. Above 40 % HbCO, severe headache, weakness on exertion, dizziness, hearing and vision troubles, nausea, and eventual collapse are observed in conjunction with a cherry-red coloration of the skin. Unconsciousness and possibly death

can occur for a carboxyhemoglobin proportion of 60 – 70 %, and above 80 – 90 %, death occurs within minutes [114].

Carbon monoxide is not metabolized in the body, and the only mode of detoxification after poisoning is the displacement of carboxyhemoglobin with pure oxygen. The partial pressure of the oxygen is important for the effectiveness of detoxification; hyperbaric (2.0 to 2.5 bar) oxygen is advantageous. The Compressed Gas Association [86] still recommends the following procedure: If available, a mixture of oxygen with 7 – 10 vol % carbon dioxide will hasten the elimination of carbon monoxide from the bloodstream by acting as a powerful respiratory and cardiac stimulant, inducing deep breathing and rapid ventilation of the lungs. The victim will generally breathe more deeply with the administration of this mixture than with pure oxygen. Inhalation of this mixture should continue for at least 15 – 30 min. This constitutes the basis of emergency therapy. Severe carbon monoxide intoxication is generally believed to be harmless if limited to a brief period of time and if the appropriate treatment is quickly provided. However, chronic exposure to lower concentrations can result in irreversible brain damage and long-lasting heart trouble.

The TLV-TWA (ACGIH) is 25 mL/m^3 (29 mg/m^3) [115]. The exposure limit value for Germany is 30 mL/m^3, and the corresponding value for France is 50 mL/m^3.

The inhalation of concentrations of 400 mL/m^3 will result in headache and discomfort within 2 – 3 h. Inhalation of 4000 mL/m^3 concentration in air proves fatal in less than 1 h. Inhalation of high concentrations can cause sudden collapse with little or no warning [86] According to EC classification carbon monoxide may cause harm to the unborn child. CO is also on the California Proposition List 65 ("Carbon monoxide contains a chemical known to the State of California to cause birth defects or other reproductive harm.")

If carbon monoxide must be handled in a closed building, continuous monitoring of its concentration in the air with an automatic detection device is recommended [116]. Efficient ventilation is the simplest means of preventing carbon monoxide poisoning.

11. References

1. *Chem. Eng. News* **50** (1972) no. 27, 2.
2. M. A. K. Khalil, mJ. P. Pinto, M. J. Shearer, *Chemosphere: Global Science Change* **1** (1999) nos. 1 – 3,IX.
3. G. D. Roth (ed.): *Handbuch für Sternfreunde,* **Band 2**, Springer, Heidelberg 1989, p. 510.
4. K. Lodders, B. Fegley, *Icarus* **112** (1994) 368.
5. H. H. Gunardson, J. M. Abrardo, *Hydrocarbon Process.* **78** (1999) no. 4, 87.
6. K. Clusius, U. Piesbergen, E. Varde, *Helv. Chim. Acta* **47** (1960) 2059.
7. A. Michels, T. Wassenaar, T. N. Zwietering, *Physica (Amsterdam)* **18** (1952) 160.
8. J. Amiel in P. Pascal (ed.): *Nouveau Traité de Chimie Minérale,* **vol. 8,** Masson, Paris 1968, p. 542.
9. *Encyclopédie des Gaz,* L'Air Liquide, Division Scientifique, Elsevier Scientific Publ. Comp., Amsterdam 1976, p. 317.
10. L. N. Canjar, F. S. Manning: *Thermodynamic Properties and Reduced Correlations for Gases,* Gulf Publ. Comp., Houston/Texas 1967, p. 134.
11. A. L. Horvath: *Physical Properties of Inorganic Compounds,* E. Arnold, London 1975, p. 169.
12. NBS Tables of Chemical Thermodynamic Properties, *J. Phys. Chem. Ref. Data Suppl.* **11** (1982) 2.
13. J. Wisniak, M. Herskowitz: *Solubility of Gases and Solids,* Elsevier, Amsterdam 1984, p. 269.
14. E. Wilhelm, R. Battino, *Chem. Rev.* **73** (1973) 1.
15. W. F. Linke: *Solubilities of Inorganic and Metal-Organic Compounds,* 4th ed., **vol. 1,** American Chemical Society, Washington, D.C., 1958, p. 453.
16. R. L. Dekoch, H. B. Gray: *Chemical Structure and Bonding,* The Benjamin/Cummings Publishing Comp., Menlo Park/California 1980, p. 255.
17. A. G. Maki, *J. Chem. Phys.* **35** (1961) 931.
18. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 4th ed., J. Wiley & Sons, New York 1980, p. 82.
19. B. F. G. Johnson: *Transition Metal Clusters,* J. Wiley & Sons, Chichester 1980.
20. J. A. Connor: "Metal Carbonyls," in: *Organometallic Chemistry,* A Specialist Periodical Report, vol. 11, The Royal Society of Chemistry, London 1983, p. 164, and previous issues of this series.

21. R. Ugo, R. Psaro, *J. Mol. Catal.* **20** (1983) 53.
22. M. A. Vannice: "Catalytic Activation of Carbon Monoxide on Metal Surfaces," in J. R. Anderson, M. Boudart (eds.): *Catalysis,* **vol. 3,**Springer Verlag, Berlin 1982, p. 139.
23. C. A. Tolman, *Chem. Rev.* **77** (1977) 313.
24. J. Wei, *Adv. Catal.* **24** (1975) 57.
25. K. Weissermel, H. J. Arpe: *Industrial Organic Chemistry,* Verlag Chemie, Weinheim 1978, p. 15.
26. R. P. A. Sneeden in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, p. 1.
27. S. G. Davies: *Organotransition Metal Chemistry: Applications to Organic Synthesis,* Pergamon Press, Oxford 1982, p. 394.
28. A. I. Vogel: *Practical Organic Chemistry,* 3rd ed., Longman, London 1970, p. 185.
29. G. H. Coleman, D. Craig, *Org. Synth. Coll.* **2** (1943) 583.
30. H. Gössling, D. Werner, *Ullmann*, 4th ed., vol. 14, p. 583.
31. B. Cornils in H. Grünewald (ed.): *Chemistry for the Future,* Pergamon Press, Oxford 1984, p. 305.
32. M. Ferreti: *La Valorisation du Charbon,* Techniques & Documentation, Paris 1982.
33. J. A. Marker, J. R. Backhurst: *Fuel and Energy,* Academic Press, New York 1981, p. 91.
34. H. D. Schilling, B. Bonn, U. Krauss: *Coal Gasification,* Graham and Trotman, London 1981.
35. E. K. Goeke, R. E. Wetzel in L. E. St. Pierre, G. R. Brown (eds.): *Future Sources of Organic Raw Materials,* Chemrawn I, Pergamon Press, Oxford 1980, p. 155.
36. K. K. Koh, *Hydrocarbon Process.* **61** (1982) no. 6, 85.
37. L. J. Buividas, *Chem. Eng. Prog.* **77** (1981) no. 5, 44.
38. A. M. Brownstein: *Trends in Petrochemical Technology,* The Petroleum Publ. Comp., Tulsa 1976.
39. M. B. Sherwin, M. E. Frank: *Chemicals from Coal and Shale – an R & D Analysis for National Science Foundation,* Chem. System Inc., New York 1975.
40. M. K. Schad, C. F. Hafke, *Chem. Eng. Prog.* **79** (1983) no. 5, 45.
41. J. Haggin, *Chem. Eng. News* **61** (1983) no. 43, 26.
42. H. Staege, *Hydrocarbon Process.* **61** (1982) no. 3, 92.
43. W. Konkol, P. Ruprecht, B. Cornils, R. Durrfeld, J. Langhoff, *Hydrocarbon Process.* **61** (1982) no. 3, 97.
44. C. W. Schwartz, L. K. Rath, M. D. Freir, *Chem. Eng. Prog.* **78** (1982) no. 4, 55.
45. K. A. Theis, E. Nitsche, *Hydrocarbon Process.* **61** (1982) no. 9, 233.
46. J. A. Gray in: *Oxygen in the Metal and Gaseous Fuel Industries,* Special Publication no. 32, The Chemical Society, London 1978, p. 83; F. C. Schora, ibid., p. 103.
47. "Shell Coal Gasification"*Hydrocarbon Process.* **63**(1984) no. 4, 96.
48. J. P. Henley, D. G. Sundstrom: "Initial Experience of the Commercial Dow Coal Gasification Plant"; Dow Chemical company publication (1987).
49. J. D. Wall, *Hydrocarbon Process.* **61** (1982) no. 6, 93.
50. H. D. Marsch, H. J. Herbort, *Hydrocarbon Process.* **61** (1982) no. 6, 101.
51. F. Marschner, H. J. Renner, *Hydrocarbon Process.* **61** (1982) no. 4, 176.
52. J. P. Van Hook, *Catal. Rev. Sci. Eng.* **21** (1980) 1.
53. G. Kurz, *Chemie-Technik* **7** (1978) no. 9, 407.
54. H. J. Madsack, *Hydrocarbon Process.* **61** (1982) no. 7, 169.
55. Gas Processing Handbook. *Hydrocarbon Process.* **63** (1984) no. 4, 51; **61** (1982) no. 4, 85.
56. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 4th ed., J. Wiley & Sons, New York 1980, p. 810.
57. A. L. Kohl, F. C. Riesenfeld: *Gas Purification,* 2nd ed., Gulf Publishing Comp., Houston/Texas 1974.
58. D. J. Maase, P. M. Duke, J. W. Cates, *Hydrocarbon Process.* **61** (1982) no. 3, 103.
59. D. G. Walker, *CHEMTECH* 1975, 308.
60. D. J. Maase, *Chem. Eng. (N.Y.)* **82** (1975) no. 8, 52.
61. N. N. Dutta, G. S. Patil, *Gas Sep. Purif.* **9** (1995) 277.
62. L'Air Liquide – Département de Construction et de Vente Mécanique: *Plants for Separation of Various Gases, Hydrogen and Carbon Monoxide,* Champigny 1983.
63. W. Forg, *Linde Ber. Tech. Wiss.* **15** (1970) 20.
64. W. Bals, R. Fabian, *Linde Ber. Tech. Wiss.* **47** (1980) 22. M. Streich, V. Linge, *Linde Ber. Tech. Wiss.* **52** (1982) 18.
65. F. Kasuye, T. Tsuji, *Gas Sep. Purif.* **5** (1991) 242.
66. *Membr. Sep. Technol. News* **2** (1984) no. 9, 1.

67. D. L. MacLean, R. S. Narayan, *Proc. World Hydrogen Energy Conf. 4th* 1982, 837.
68. J. M. S. Henis, M. K. Tripodi, *Science (Washington, D.C.)* **220** (1983) 1.
69. G. Schulz, H. Michele, U. Werner, *Chem. Ing. Tech.* **54** (1982) 351.
70. H. K. Lonsdale, *J. Membr. Sci.* **10** (1982) 81.
71. J. L. Fox, *Chem. Eng. News* **60** (1982) no. 45, 7.
72. W. A. Bollinger, D. L. MacLean, R. S. Narayan, *Chem. Eng. Progr.* **78** (1982) no. 10, 27.
73. C. Wang, R. G. Prinn, *Chemosphere: Global Science Change* **1** (1999) 73.
74. H. J. Hellebrand, *J. Agricult. Engin. Res.* **69** (1998) 365.
75. *Science* **173** (1971) 141.
76. P. G. Flachsbart, R. G. Prinn, *Chemosphere: Global Science Change* **1** (1999) 301.
77. http://www.umweltbundesamt.de/uba-info-daten/unten.htm
78. ″Exhaust emissions″, Mercedes Benz Company Brochure (1997).
79. Internet Catalogue Praxair: http://www.praxair.com/.
80. Internet Catalogue Air Liquide: http://www.airliquide.com/us/.
81. Internet Catalogue Air Products: http://www.airproducts.com.
82. Internet Catalogue Messer: http://www.spezialgase.de/spezialgasekatalog/
83. *Annual Book of ASTM Standards,* **vol. 11.**03, ASTM, Philadelphia 1983, p. 163.
84. (185) J. D. Hogan (ed.): *Specialty Gas Analysis,* Wiley-VCH, Weinheim, 1997p. 43.
85. *Annual Book of ASTM Standards,* **vol. 11.**03, ASTM, Philadelphia 1983, p. 320.
86. Compressed Gas Association: *Handbook of Compressed Gases,* 4th ed., Van Nostrand Reinhold Comp., New York 1999.
87. Règlement International Concernant le Transport des Marchandises Dangereuses par Chemin de Fer (R.I.D.), Bern.
88. Accord Européen Relatif au Transport International des Matières Dangereuses par Route (A.D.R.).
89. Règlement de La Haye 1. 02. 1939.
90. *Direct Reduction of Iron Ore: A Bibliographical Survey,* The Metal Society, London 1979.
91. US 4425310, 1984 (P. D. Bowerman, M. F. Lucid). Public Service Co. of New Mexico, US 4285548, 1981 (J. W. Erickson) Mineral Process Licensing Corp. B. V., PCT Int. Appl. WO 8301612, 1983 (A. J. Wickens, J. M. W. Turner).
92. G. Henrici-Olive, S. Olive: *The Chemistry of the Catalyzed Hydrogenation of Carbon Monoxide,* Springer Verlag, Berlin 1984.
93. I. Tkatchenko, *Inf. Chim.* **249** (1984) 115.
94. A. Aquilo, J. S. Alder, D. N. Freeman, R. J. H. Voorhoeve, *Hydrocarbon Process.* **62** (1983) no. 3, 57.
95. W. Keim: *Catalysis in C-1 Chemistry,* D. Reidel Publ., Dordrecht 1983.
96. R. A. Sheldon: *Chemicals from Synthesis Gas,* D. Reidel Publ., Dordrecht 1983.
97. R. P. A. Sneeden in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, London 1982, p. 19.
98. I. Tkatchenko in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, London 1982, p. 101.
99. D. L. King, K. K. Ushiba, T. E. Whyte Jr., *Hydrocarbon Process.* **61** (1982) no. 11, 131.
100. D. T. Thompson in R. Pearce, W. R. Patterson (eds.): *Catalysis and Chemical Processes,* Leonard Hill, London 1981.
101. W. R. Moser (ed.): *Catalysis of Organic Reactions,* Marcel Dekker, New York 1981.
102. P. S. Ford (ed.), "Catalytic Activation of Carbon Monoxide," *ACS Symp. Ser.* **152** (1981).
103. W. Jones (ed.): *Catalysis in Organic Syntheses,* Academic Press, London 1980.
104. J. Falbe (ed.): *New Synthesis with Carbon Monoxide,* Springer Verlag, Berlin 1980.
105. R. Eisenberg, D. E. Hendriksen, *Adv. Catal.* **28** (1979) 79.
106. D. Forster, *Adv. Organomet. Chem.* **17** (1979) 255.
107. H. W. Cooper Jr., R. C. Hart, *Chem. Eng. Prog.* **78** (1982) no. 4, 72; **80** (1984) no. 5, 2.
108. J. Reininghaus, R. Schmidt, G. Schroeder, U. Zintel, *gas aktuell (Messer magazine)* **59** (2000) 4.
109. J. T. Snow, C. Ma *Eur. Semicond.* **20** (1988) no. 10, 39.
110. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., Van Nostrand Reinhold Comp., New York 1984, p. 643.
111. E. Meyer: *Chemistry of Hazardous Materials,* Prentice-Hall Inc., Englewood Cliffs/New Jersey 1977.
112. *A. Antonini, M. Brunori: Hemoglobin and Myoglobin in Their Reactions with Ligands,* North Holland Publ. Comp., Amsterdam 1971.

113. J. R. Goldsmith, S. A. Landaw, *Science (Washington, D.C.)* **162** (1968) 1352.
114. R. D. Stewart, *Annu. Rev. Pharmacol.* **15** (1975) 409.
115. *Threshold Limit Values and Biological Exposure Indices* ACGIH, Cincinatti, OH1998.
116. M. L. Woebkenberg, *Appl. Occup. Environ. Hyg.* **13** (1998) 567.

Carbonylation

Werner Bertleff, BASF Aktiengesellschaft, Ludwigshafen, Germany
Michael Roeper, BASF Aktiengesellschaft, Ludwigshafen, Germany
Xavier Sava, BASF Aktiengesellschaft, Ludwigshafen, Germany

1. Introduction 1
2. Catalysts 2
2.1. Nickel Catalysts 3
2.2. Cobalt Catalysts 4
2.3. Palladium Catalysts 5
2.4. Rhodium and Iridium Catalysts 7
3. Carbonylation of Alkynes 7
3.1. Acrylic Acid from Acetylene 7
3.2. Other Carbonylation Reactions of Alkynes 7
4. Carbonylation of Alkenes 8
4.1. Monoalkenes 10
4.2. Substituted Alkenes 11
4.3. Polyenes 12
5. Carbonylation of Alcohols and Ethers 15
5.1. BASF Acetic Acid Process 15
5.2. Monsanto and Related Low-Pressure Acetic Acid Processes 15
5.3. Carbonylation of Other Alcohols and Ethers 16
6. Carbonylation of Organic Halides . . 18
6.1. Aryl Halides 18
6.2. Alkyl Halides 19
6.3. Allyl Halides and Vinyl Halides 19
7. Other Carbonylation Reactions 20
7.1. Amines 20
7.2. Aromatic Nitro Compounds 21
7.3. Esters 21
7.4. Lactones and Aldehydes 22
7.5. Amidocarbonylation of Aldehydes . . 22
8. References 23

Carbonylation is a general method to produce carboxylic acid derivatives by homogeneous transition metal catalyzed reaction with CO. Today, more than 70 % of synthetic acetic acid production is based on rhodium- or iridium-catalyzed methanol carbonylation. Capacity of world-scale plants using this technology are in the range of 500×10^3 tons per year. But also alkynes, alkenes, esters, and organic halides can be successfully carbonylated. More recently, cationic palladium complexes with chelating bisphosphines have been developed that show great promise. At mild conditions, exceedingly high selectivities as well as high activities can be achieved and commercialization of this technology has been announced. Amidocarbonylation to produce amino acid derivatives is another technology that has found great interest and carbonylation of aromatic halides is about to establish itself as a standard procedure for pharmaceutical intermediates. The updated article on carbonylation covers the major developments in this field.

1. Introduction

Reactions with carbon monoxide [*630-08-0*] rank among the most selective methods of catalytically introducing oxygen-containing functional groups into organic substrates. Walter Reppe at BASF introduced the term *carbonylation* to describe a number of reactions he discovered from 1938 onward [1 – 3]. Since then, any reaction in which carbon monoxide, alone or together with other components (e.g., water, alcohols, or amines), is introduced into an organic molecule in the presence of a catalyst is called carbonylation. Organic substrates can be unsaturated or saturated; in the second case they must contain functional groups such as hydroxyl, alkoxycarbonyl, amino, or halogen.

Carbonylation covers a wide variety of reactions. They may be classified into (1) Reppe reactions, (2) hydroformylations, and (3) Koch carbonylations.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_217.pub2

Reppe Reactions. The term Reppe reactions describes reactions of organic substrates with carbon monoxide, sometimes in the presence of nucleophiles that contain acidic hydrogen; these reactions are catalyzed by metal carbonyls or their precursors. The reaction of unsaturated compounds with carbon monoxide and water is also known as *hydrocarboxylation*; similarly, that of carbon monoxide and alcohols is known as *hydroesterification*.

The Reppe reactions yield carboxylic acids, esters, anhydrides, amides, lactones, and isocyanates. Starting materials are alkynes, alkenes, esters, alcohols, organic halides, ethers, lactones, aldehydes, amines, or aromatic nitro compounds. Many of the carbonylation reactions of unsaturated compounds also require nucleophiles containing acidic hydrogen, such as water, alcohols, ammonia, amines, thiols, or carboxylic acids [4].

Hydroformylation. In 1938 Otto Roelen at Ruhrchemie discovered the synthesis of aldehydes and alcohols from olefins, carbon monoxide, and hydrogen. This reaction has for historical reasons always been treated separately as *hydroformylation* or *oxo synthesis*. Although it is covered by the generic term carbonylation, it is not treated in this article (→ Aldehydes, Aliphatic and Araliphatic; → Oxo Synthesis).

Koch Carbonylation. The carbonylation of olefins, alcohols, or aldehydes, by using strong acids such as H_2SO_4, $H_3PO_4 - BF_3$, or $HF - SbF_5$ as catalysts, is known as the Koch synthesis (also called Koch-Haaf reaction) [4, pp. 123 – 146], [5]. It yields predominantly branched isomers of carboxylic acids: *Neo Acids* (Exxon), *Versatic Acids* (Resolution, formerly Shell). Du Pont has also developed a process for the production of glycolic acid based on the Koch carbonylation of formaldehyde. The economic importance of these products is far less than that of the compounds obtained by Reppe reactions; the Koch synthesis is not dealt with in this article.

Also not treated is the base-catalyzed carbonylation of alcohols or amines to yield alkyl formates and formamides. The most prominent examples are methyl formate, an intermediate for the industrial synthesis of formic acid, and dimethylformamide (→ Formic Acid).

History. Many carbonylation processes were discovered in the period 1938 – 1957 by Reppe and his collaborators at BASF, and have subsequently been developed further to industrial-scale processes. Examples are the syntheses of acrylic acid, propionic acid, and acetic acid [3].

Before the term carbonylation was introduced, reactions of olefins with carbon monoxide and water to give carboxylic acids or their esters were described in numerous patents. However, these reactions used substances of large surface area as catalyst carriers and acids such as phosphoric acid, heteropoly acids, boron trifluoride, metal halides, or similar catalysts; they required high pressures (700 – 900 bar) and high temperatures and are no longer of practical importance. Further development of these carbonylation reactions led to implementation of the Koch synthesis.

Metal carbonyls were first used as carbonylation catalysts by Reppe. Their great advantage over the older catalysts is that structural isomerization hardly takes place during the reaction. Metals of the eighth subgroup have proved to be most effective as catalysts. It was possible to optimize the catalyst metal, the ligands, and the promoters for nearly every carbonylation reaction in such a way that the reactions took place under milder conditions than previously known. Nevertheless, these processes could only be realized industrially after the development of appropriate reactor materials because of the corrosive properties of the reaction media.

2. Catalysts

Metal carbonyls or metal carbonyl hydrogen compounds or their substituted derivatives are most suitable as catalysts for carbonylation reactions. The most effective catalyst metals are Fe, Co, Ni, Ru, Rh, Pd, Os, Ir, Pt, and Cu. However, it can also be advantageous to use more readily accessible metal compounds as "precatalysts," provided that they are capable of forming metal carbonyl complexes under the reaction conditions. Examples of these are simple or complex salts (mainly halides and acetates), oxides, or the metals themselves in finely divided form. When halides are used, the nature of the halogen is often a decisive factor, especially for nickel and cobalt: the catalytic activity increases from the

chloride through the bromide to the iodide; fluorides are inactive. Table 1 gives a survey of reactants in common industrial carbonylations.

Table 1. Summary of the catalyst metals generally used for carbonylation reactions

Raw materials	Products	Catalyst metals
Alkynes	unsaturated carboxylic acids and derivatives	Pd, Ni, Fe, Co
	unsaturated acid chlorides	Rh, Ru
	unsaturated dicarboxylic acids and derivatives	Pd, Co
Alkenes	saturated carboxylic acids and derivatives	Pd, Co, Ni, Pt, Fe, Ru, Cu
	unsaturated carboxylic acids and derivatives	Pd
	alcohols	Fe
Dienes	unsaturated monocarboxylic acids or saturated dicarboxylic acids or their esters	Pd, Co, Pt
	unsaturated dicarboxylic acid esters	Pd – Cu, Pd – V, Pd – Fe
Alcohols	carboxylic acids	Rh, Ir, Co, Ni
Diols	dicarboxylic acids	Rh, Cu, Ir
Alkyl halides and aryl halides	carboxylic acids and derivatives	Pd, Rh, Ni, Co, Fe
Allyl halides	carboxylic acid halides	Pd, Ni
Ethers	carboxylic acid esters	Rh, Ni, Co, Pd, Fe
Esters	anhydrides and carboxylic acids	Rh, Ni, Co,
Aldehydes	hydroxycarboxylic acids	Ni, Co
Nitro compounds	isocyanates and ureas	Pd, Rh, Ni, Co, Fe
Amines	formamides and ureas	Co, Ni, Pd

More recent work is focused on Rh or Ir catalysts for the carbonylation of alcohols, ethers, and esters and on Pd catalysts for the carbonylation of unsaturated substrates.

The following chemical reactions always take place at the central metal in the course of the catalysis (Fig. 1):

1) With saturated substrates, oxidative addition to the metal occurs, formally increasing its oxidation state by two; with unsaturated substrates the metal hydride adds to the π-electron system. Both reactions result in the formation of a metal – carbon σ-bond.
2) The next stage is the formation of an acyl – metal intermediate by migratory insertion of CO.
3) The final stage is the reductive elimination or solvolysis of the organometallic complex to yield the product.

When optimizing the reaction conditions, one must take into account the stability of the corresponding carbonyl complex; it varies considerably with the ligands, carbon monoxide pressure and temperature [4, p. 15].

All the metal carbonyls used as catalysts for Reppe-type reactions are extremely toxic. Because of this, great care must be taken when handling these substances, especially the volatile compounds $Ni(CO)_4$, $HCo(CO)_4$, and $Fe(CO)_5$.

2.1. Nickel Catalysts

For the carbonylation of alkynes, catalysts based on nickel and hydrogen halides are found to be far superior to all other catalysts. Nickel has lost its outstanding position with the decline in the economic importance of acetylene as a feedstock, because in some cases, nickel is less effective for the carbonylation of olefins, alcohols, or organic halides than other metals.

The catalytically active species is always a substituted nickel carbonyl complex. The reaction takes place at CO pressures of ≤ 100 bar and temperatures below 250 °C; at higher pressures the formation of tetracarbonyl nickel [*13463-39-3*], $Ni(CO)_4$, results in inactivity. Because nickel carbonyls are formed extremely readily under an atmosphere of CO, it is sufficient to use binary or complex nickel salts as starting materials. The use of $Ni(CO)_4$ enables carbonylation reactions to be carried out stoichiometrically in a convenient manner without using CO [6]. The reaction mechanism of the nickel-catalyzed carbonylation of alkynes is shown in the scheme on the next page (under Fig. 1):

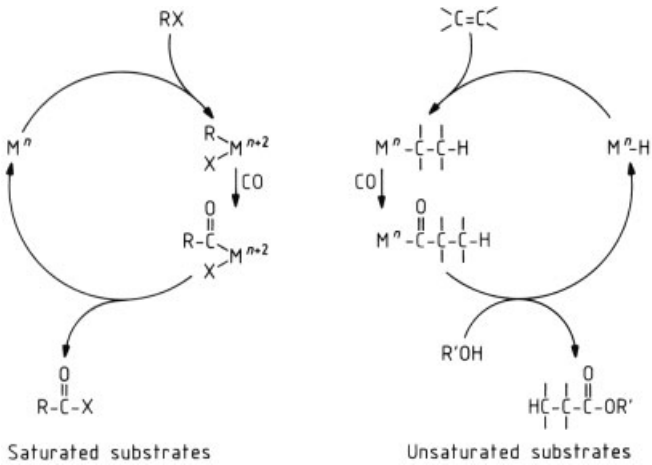

Figure 1. Catalytic cycle in carbonylation with saturated substrates or unsaturated substrates
R, R′ = alkyl, aryl; X = halide

Catalyst formation:

$NiX_2 + 5\ CO + H_2O \longrightarrow Ni(CO)_4 + 2\ HX + CO_2$

$Ni(CO)_4 + HX \longrightarrow HNi(CO)_2X + 2\ CO$

Carbonylation:

$HNi(CO)_2X + R{-}C{\equiv}CH \longrightarrow H_2C{=}C(R){-}Ni(CO)_2X$

$H_2C{=}C(R){-}Ni(CO)_2X + CO \longrightarrow H_2C{=}C(R){-}C(=O){-}Ni(CO)_2X$

$H_2C{=}C(R){-}C(=O){-}Ni(CO)_2X + R'OH \longrightarrow H_2C{=}C(R){-}C(=O){-}OR' + HNi(CO)_2X$

$R{-}C{\equiv}CH + CO + R'OH \longrightarrow H_2C{=}C(R){-}C(=O){-}OR'$

where X = halide; R, R′ = alkyl

When nickel halides are used as catalysts, traces of water are required to accelerate the formation of the active species. In absolutely anhydrous media, the carbonylation reaction only starts when water is formed in side reactions.

2.2. Cobalt Catalysts

The main field of use for cobalt catalysts in industrial processes is the hydroformylation of olefins under high-pressure conditions to yield aldehydes or alcohols. Cobalt catalysts may also be used for the carbonylation of monoolefins and diolefins and of alcohols to give carboxylic acids or their derivatives. The cobalt catalysts are used in the form of octacarbonyl dicobalt [*10210-68-1*], $Co_2(CO)_8$, or, as precursors of the latter, cobalt acetate or the cobalt salt of the carboxylic acid to be synthesized. The preferred reaction conditions are 150 – 250 °C and 150 – 200 bar for alkenes, and 500 bar or higher for alcohols. The active species in each case is hydrido tetracarbonyl cobalt [*16842-03-8*], $HCo(CO)_4$ [7]; the formation of this is accelerated by adding bases (pyridine, picoline, quinoline) or hydrogen:

Catalyst formation:

$2\ Co(OOCR)_2 + 8\ CO + 3\ H_2 \longrightarrow 2\ HCo(CO)_4 + 4\ RCOOH$

or

$Co_2(CO)_8 + 2\ C_5H_5N + CO + H_2O \longrightarrow 2\ C_5H_5N^+Co(CO)_4^- + CO_2$

$\Downarrow\Uparrow$

$C_5H_5N + HCo(CO)_4$

Catalytic cycle:

$HCo(CO)_4 + R'{-}CH{=}CH{-}CH_3 \rightleftharpoons$

$R'{-}CH{=}CH{-}CH_3 \rightleftharpoons R'{-}CH_2{-}CH(Co(CO)_n){-}CH_3 \rightleftharpoons$

$R'{-}CH_2{-}CH{=}CH_2 \rightleftharpoons R'{-}CH_2{-}CH_2{-}CH_2{-}Co(CO)_n \xrightarrow{CO,\ R''OH}$

$R'{-}CH_2{-}CH_2{-}CH_2{-}COOR'' + HCo(CO)_4$

where R, R′, and R″ = alkyl and n = 2 or 3

In the presence of hydrogen, hydroformylation takes place as a competitive reaction; therefore, the preforming of the catalyst with CO – H_2 often is carried out as a separate stage in industrial processes (precarbonylation) [8].

The addition of tertiary amines increases not only the reaction rate, but also the proportion of linear isomer in the product: coordination of the amine to the cobalt causes considerable steric hindrance for the further reaction of internal double bonds. As a result of this, these olefins isomerize to give α-olefins, which is also catalyzed by cobalt (for catalytic isomerization of double bonds, see Chap. 4). On the other hand, high concentrations of base suppress the carbonylation of monoolefins, because the base also competes with the alkene as a ligand.

The *carbonylation of conjugated dienes* requires a fairly high concentration of tertiary amines to break up the intermediate π-allyl complex [9]:

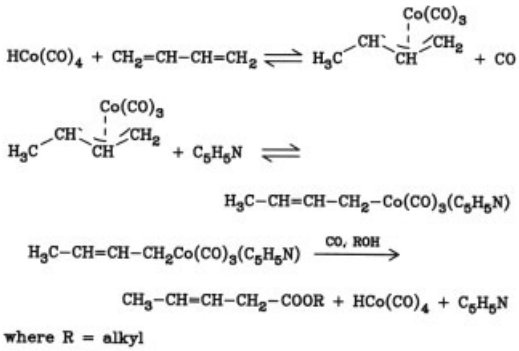

As the concentration of base increases, the reaction rate reaches a plateau and then remains unchanged if the concentration is increased further.

2.3. Palladium Catalysts

In general, carbonylation reactions with palladium catalysts take place under much milder conditions ($\leq$ 150 bar and 140 °C) than those with nickel or cobalt catalysts [10]. Considerable efforts have been devoted to develop more stable palladium catalysts, and the role of mono- and bidentate ligands as well as the role of coordinating and noncoordinating anions has been studied in detail [11]. Several large-scale industrial processes have been developed and palladium is the catalyst of choice for fine chemicals synthesis such as Ibuprofen. It has also helped to establish carbonylation as a standard synthetic tool in active ingredients production for agro and pharmaceuticals. Examples for Pd catalysts for the alternating polymerization of ethylene and CO to generate polyketone, for the carbonylation of alkenes, and of alkadienes are shown below.

Homogeneous catalysts based on palladium generally have a very narrow range of optimal activity and stability. Their suitability for a specific reaction depends, to a far greater extent than with the other metals, on finding a suitable combination of ligands, promoters, and solvents. The reason for this is that no simple carbonyl complexes of palladium exist, so that stabilization of the catalyst always has to be effected by adding further ligands, such as phosphines, amines, or nitriles. Furthermore, Pd^{2+} is a strong oxidizing agent, and even CO may cause its reduction to the metal, especially in an acid medium. This may be counteracted by adding such reoxidants as Cu^{2+} or Fe^{3+}.

In practice, homogeneous palladium catalysts are usually complicated multicomponent systems, and catalyst handling is problematic in continuous processes. In many cases the separation of the catalyst leads to deactivation and palladium losses. Therefore, the number of industrial processes that use palladium as the catalyst metal is still limited. To facilitate product catalyst separation the use of aqueous phase organometallic catalysis could be an option [12].

Besides the carbonylation of olefins, halides, or alcohols to carboxylic acid derivatives (see Section 4.1, Chaps. 5 and 6), palladium catalysts may also be employed for the oxidative carbonylation of monoolefins to unsaturated carboxylic acids and esters. The oxidative carbonylation of monoolefins requires a halide as promoter. In nucleophilic solvents, a stoichiometric reaction takes place in which Pd^{2+} is simultaneously reduced to Pd^0. It affords either monocarboxylic or dicarboxylic acid derivatives, the latter in the presence of strong nucleophiles (alkoxides or amines) [13]:

Table 2. Selectivity control by variation of palladium catalyst structure*

Catalyst	Ligand	Ligand structure	Substrate	Product	Ref.
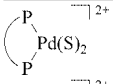	dppp	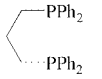	ethylene, propylene	polyketone	34
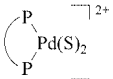	dtbpm	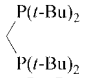	ethylene, propylene	polyketone	35
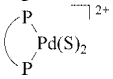	dtbpe	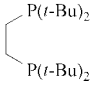	ethylene	diethylketone	33
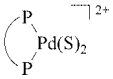	dtbpx	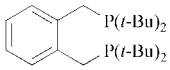	ethylene	methylpropionate	35
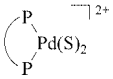	dpa-3	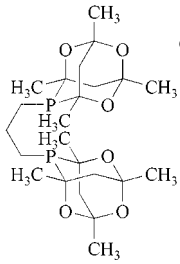	ethylene	methylpropionate	33 35
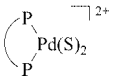	bcope	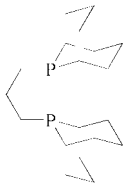	butadiene	methyl-3-pentenoate	74
$L_2Pd(S)_2^{2+}$	PPh_3	PPh_3	ethylene	methylpropionate	11
$L_2Pd(S)_2^{2+}$	2-$PyPPh_2$	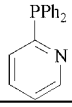	propyne	methylmethacrylate	21

* S = Solvent.

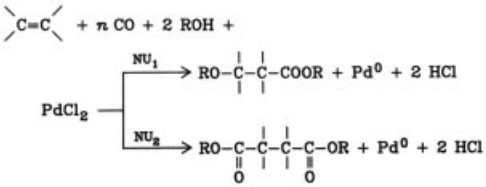

where n = 1,2; NU_1 = nucleophile, e.g., alcohol
NU_2 = strong nucleophile, e.g., alkoxide or amine; R = alkyl

In the presence of an oxidizing agent, such as Cu^{2+}, Fe^{3+}, or quinones, Pd^0 is reoxidized to Pd^{2+}, and the synthesis then follows a catalytic course, provided that the oxidizing agent is in turn reoxidized with oxygen.

In weakly nucleophilic or non-nucleophilic media the addition to the double bond does not take place and unsaturated carboxylic acid derivatives are obtained [14]. This reaction is then called *oxidative carbonylation*. It may be summarized as follows:

$$R{-}CH{=}CH_2 + PdCl_2 + CO + R'OH \longrightarrow R{-}CH{=}CH{-}COOR' + Pd^0 + 2\,HCl$$

$$2\,FeCl_3 + Pd^0 \longrightarrow 2\,FeCl_2 + PdCl_2$$

$$2\,FeCl_2 + 2\,HCl + 1/2\,O_2 \longrightarrow 2\,FeCl_3 + H_2O$$

where R, R' = alkyl

The primary catalytic cycle at the palladium center includes the following steps:

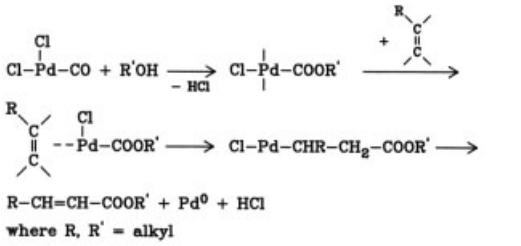

An industrial application of the oxidative carbonylation is not yet in sight.

Palladium catalysts are also used for the *carbonylation of organic halides*. Normally a Pd(0) – phosphine complex, such as $Pd(CO)[P(C_6H_5)_3]_3$ or $Pd[P(C_6H_5)_3]_4$, or a complex Pd(II) halide, such as $PdCl_2[P(C_6H_5)_3]_2$ or a bis(π-allyl)palladium chloride, is employed. The carbonylation of organic iodides takes place under conditions as mild as 60 – 100 °C and 1 – 15 bar [15]; the reaction mechanism is shown below:

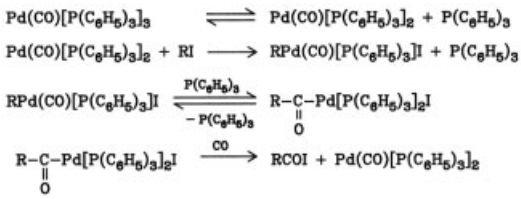

2.4. Rhodium and Iridium Catalysts

The preferred processes to produce acetic acid as well as acetic anhydride use noble-metal carbonylation technology. The first Rh-based process to acetic acid is known as the Monsanto process. Today, about 8×10^6 tons of acetic acid are produced by this technology and plant capacities exceeding 500 000 t/a are regarded as world scale. Anionic complexes such as $[Rh(CO)_2I_2]^-$ or $[Ir(CO)_2I_2]^-$ are used in the presence of high concentrations of methyl iodide and the oxidative addition of the latter to Rh is the rate-determining step [16]. The significant lower costs of iridium and the remarkable robustness of its carbonyl complexes have favoured its commercial application recently.

3. Carbonylation of Alkynes

3.1. Acrylic Acid from Acetylene

The first catalytic carbonylation reaction achieved by Reppe was the reaction of acetylene with CO and H_2O to give acrylic acid [*79-10-7*], using $Ni(CO)_4$ as the catalyst [17]. If water is replaced in this reaction by alcohols, thiols, secondary amines, or carboxylic acids, the corresponding esters, thioesters, amides, or mixed anhydrides of acrylic acid are obtained. In general, pressures below 100 bar and temperatures below 250 °C are required for this reaction, if such catalysts as $Ni(CO)_4$ or nickel salts (halides or complex salts) are present.

The synthesis of acrylic acid from acetylene was originally the industrially most important Reppe reaction. Since the middle of the 1960s, however, its importance has declined steadily, compared with that of the oxidation of propene [3, p. 296, 18]. An acetylene-based acrylic acid plant was operated by BASF in Ludwigshafen (110 000 t/a) until recently. It employed the catalyst system $NiBr_2$ /CuI. The synthesis took place at 100 bar and 220 – 230 °C, yielding acrylic acid with a selectivity of 90 %, based on acetylene.

The Reppe acrylic acid process is available in three versions: (1) in a stoichiometric version, in which $Ni(CO)_4$ acts as the source of CO and, together with HCl, as the catalyst; (2) in a stoichiometric – catalytic version, in which the $Ni(CO)_4$ is employed in a stoichiometric amount, but the $NiCl_2$ produced is rereduced to $Ni(CO)_4$ in an intermediate stage; and (3) in the catalytic version (BASF plant), in which the catalyst is not worked up [6]. Only the last two of these have achieved industrial importance.

3.2. Other Carbonylation Reactions of Alkynes

The carbonylation of alkynes in the presence of nickel catalysts always involves a formal *cis* addition of H and COOH to the triple bond. Markovnikov or anti-Markovnikov products result, depending on the nature of the substituent:

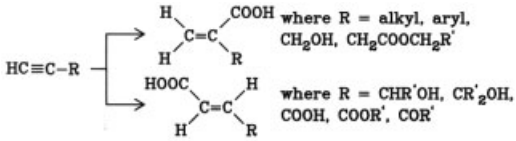

The homologous series of alkylacetylenes thus yields, virtually without exception, the corresponding 2-alkylacrylic acids. The yields decrease as the length of the alkyl radical increases [19]. Accordingly, the slowly reacting acetylenecarboxylates yield fumarates when a palladium compound and a hydrogen halide are used as catalysts [20].

A highly active catalyst system for the methoxycarbonylation of propyne to yield methyl methacrylate (MMA) has been developed by Drent and co-workers from Shell [21]. The cationic palladium catalyst is formed in situ from palladium acetate, an excess (10–40-fold on Pd) of a monophosphine, and a strong acid such as methylsulfonic acid. 2-Pyridyldiphenylphosphine (2-PyPPh$_2$) was the preferred ligand and at mild conditions such as 45 °C and 10 – 60 bar of CO about 40 000 mol (MMA) mol^{-1} h^{-1} (Pd) MMA were formed with a selectivity of almost 99%.

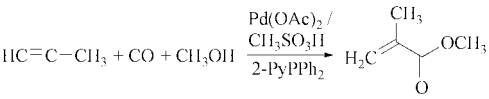

Depending on the reaction conditions and the catalysts a variety of products can be obtained from alkyne carbonylation, such as methyl 2,4-pentadienoate (**1**) [22], maleic acid esters (**2**) [23], or acetylenecarboxylic acid esters (**3**) [24]:

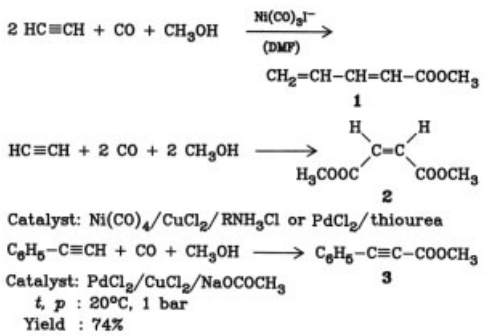

Polyunsaturated alkynes (enynes or diynes) react slower than simple alkynes and require drastic reaction conditions.

A few special carbonylation reactions of alkynes yield cyclic products, some of which are natural products or biologically active. Examples are the α-methylene lactone (**4**) [25] or the substituted furanone (**5**) [26, 27]:

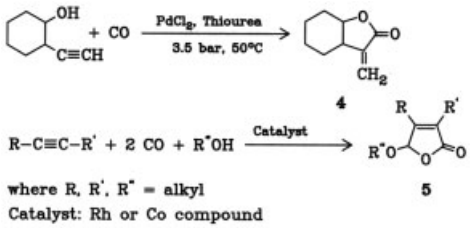

4. Carbonylation of Alkenes

In principle, the same catalyst metals may be used for the carbonylation of alkenes as for the carbonylation of alkynes, but higher temperatures and pressures are required. In contrast to alkyne carbonylation, the advantages of nickel as the catalyst metal are much less pronounced in alkene carbonylation. Cobalt, palladium, platinum, rhodium, and ruthenium give equally good or even better results [28]. The reaction conditions, yields, and selectivities can be influenced to a considerable extent by the choice of ligands and promoters.

Selectivities are in most cases markedly inferior to those of alkyne carbonylation. Usually mixtures of carboxylic acids or esters are obtained. The reason for this is isomerization of the olefinic substrate, because all carbonylation catalysts isomerize olefinic double bonds to some extent [29].

Nickel has the smallest isomerization activity. The product mixture obtained reflects almost exclusively the initial position of the double bond. If α-olefins are used, the introduction of the functional group into the 2-position is slightly preferred (60 %) over the 1-position (40 %) [30].

Cobalt is a much better isomerization catalyst than nickel. When the cheaper internal olefins or olefin mixtures are carbonylated, the products are mainly derived from the terminal olefin regardless of the position of the double bond in

the substrate. In contrast to nickel, the carbonylation introduces the functional group mainly into the 1-position [30], and the linear carboxylic acid, for example, predominates in the product mixture (*n*-fraction: 55 – 60 %).

The proportion of linear product can be increased to over 80 % by adding bases (bulky ligands, such as pyridine or picoline; cf. Section 2.2) [31].

In the presence of *palladium* and *platinum* catalysts, the reactions take place at moderate temperatures (below 140 °C); this facilitates the carbonylation of substrates that are sensitive to polymerization, such as butadiene or styrene. However, platinum catalysts are only effective for α-olefins [32]. As mentioned earlier, considerable progress has been achieved in the improvent of palladium catalysts (see also Table 2). For instance, complete product selectivity control has been achieved in the methoxycarbonylation of ethylene [11, 33, 34]. Cationic palladium complexes with monodentate ligand such as PPh_3 or chelating ligands with bulky substituents such as 1,3-bis(di-1,3,5,7-tetramethyl-2,4,8-trioxa-6-phospha-adamant-6-yl)propane (dpa-3), 1,3-bis(di-*tert*-butylphosphino)propane (dtbpp) or 1,2-bis(di-*tert*-butylphosphinomethyl)benzene (dtbpx) yield with very high activity methylpropionate with a selectivity exceeding 99%. *Diethyl ketone* is obtained, when 1,2-bis(di-*tert*-butylphosphino)ethane (dtbpe) is used in the presence of strong acids. However, if 1,3-bis(diphenylphosphino)propane (dppp) or 1,1-bis(di-*tert*-butylphosphino)methane (dtbpm) is used to generate the cationic complexes, *polyketone*, a perfect alternating copolymer of ethylene and CO, precipitates as a snow white solid. Molecular masses (M_r) of about 20 000 g/mol are achieved and more than 10^6 moles of ethylene per mole of Pd are converted. The influence of ligands to control carbonylation selectivity has been reviewed [35]. Palladium catalysts with chelating phosphines have also been developed for the isomerizing carbonylation of internal olefins, introducing the carboxylic acid into the 1-position with high selectivity (97 %) [36].

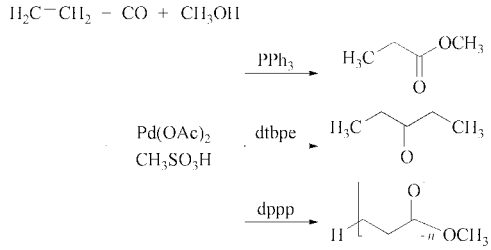

For the palladium catalyzed carbonylation of unsaturated substrates, formic acid, oxalic acid, as well as alkyl formates have been proposed as the carbonylating agent, replacing carbon monoxide [37].

Simple or complex *rhodium* and *iridium* salts, such as those containing tertiary phosphines as ligands, also catalyze the carbonylation of olefins in the temperature range 100 – 130 °C and at a pressure of 1 bar or more [29, 38].

In general, hydroesterification (CO + ROH) takes place more slowly than hydrocarboxylation (CO + H_2O). The opposite applies to palladium catalysts, however. The reactivity of the olefin decreases as the degree of branching and substitution increases. Therefore, tetraalkylethylenes cannot be carbonylated by the Reppe reaction. Under the conditions of the Koch synthesis, however, they are readily carbonylated [4, pp. 123 – 146].

The reaction of olefins with carbon monoxide and alcohols can be accelerated by adding hydrogen, water (formation of hydrogen by the shift reaction), or acids (carboxylic acids having up to three carbon atoms or a hydrogen halide). These additives promote the formation of the carbonyl hydride, which is the real effective catalyst.

In the presence of nucleophiles other than water or alcohols, acid chlorides, amides, thioesters, or anhydrides are obtained:

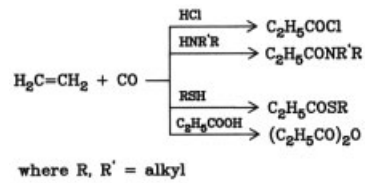

4.1. Monoalkenes

Ethylene. Catalytic carbonylation of ethylene [*74-85-1*] in the presence of water and nickel propionate gives propionic acid [*79-09-4*] in a yield of 95 %. The following process is operated by BASF: 200 – 240 bar, 270 – 320 °C, silver-lined reactor [5, 39] (→ Propionic Acid and Derivatives). When propionic acid is used as the solvent in this reaction, the product is propionic anhydride [2, p. 38], [40]. Much milder conditions such as pressures of < 10 bar and temperatures of 100 – 150 °C can be applied, if palladium catalysts with chelating ligands such as dpa-3 are used [41]:

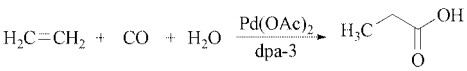

Lucite has developed the Alpha process to produce MMA by methoxycarbonylation of ethylene to methyl propionate followed by reaction with formaldehyde. A commercial plant with a capacity of 120 000 t/a MMA has been announced which is expected to be commissioned in 2007 in Singapore [42]:

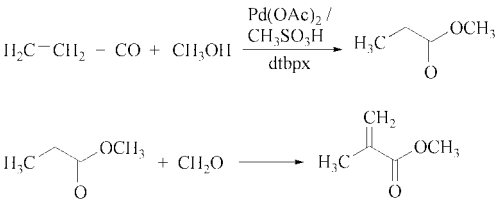

The oxidative carbonylation of ethylene using a $PdCl_2$ – $CuCl_2$ catalyst in acetic anhydride yields a mixture of acrylic acid [*79-10-7*] and β-acetoxypropionic acid [*4272-12-2*]; the second can be converted into acrylic acid by heating [43, 44]. If this reaction is carried out in the presence of secondary amines, acrylamides are obtained [45]. The synthesis of dialkyl succinate from ethylene using palladium catalysts takes place in the presence of a nucleophile, cf. Section 2.3:

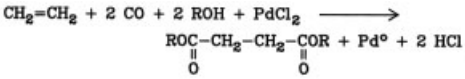

When $Co_2(CO)_8$ is used as catalyst, diethyl ketone [*96-22-0*] is obtained from ethylene and carbon monoxide in the presence of a hydrogen donor at 165 °C and 70 bar in a yield of 97 % [46]. A very efficient diethylketone synthesis is achieved at 40 bar CO/H_2 (1:1) with a $Pd/MeSO_3H$/dtbpe catalyst [33]. At a rate of 80 000 mol mol^{-1} h^{-1} (Pd) the diethylketone selectivity reaches 98%.

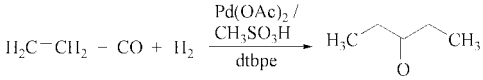

A further reaction of ethylene leads to five-membered cyclic lactones (100 – 200 °C; 600 – 800 bar). The catalyst is a palladium compound in combination with a hydrogen halide. The yields of the products **6, 7,** and **8** in this reaction depend on the conditions. At elevated temperature lactone **7** is formed preferentially [48]:

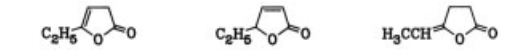

6 3–Hexen–1,4–olide **7** 2–Hexen–1,4–olide **8** 4–Hexen–1,4–olide

Palladium and chelating phosphines are equally very efficient catalysts for the strictly alternating copolymerisation of ethylene and carbon monoxide. The synthesis of polyketone by use of cationic palladium/dppp catalysts has been scaled up by Shell into a commercial operation of about 20 000 t/a. The polymer was intended to serve the engineering plastics market due to its excellent high temperature properties which could be adjusted by using mixtures of ethylene and propene as the olefin. Unfortunately, market success did not meet the expectations and the plant was dismantled. The same process was studied by BP at pilot plant scale.

The same kind of catalytic systems (a palladium(II) precursor, a chelating phosphine and an acid) in methanol as the solvent can thus lead very selectively to methyl propionate or to polyketone. The selectivity of this transformation is controlled by the substitution scheme of the chelating phosphine [35].

Propene. Depending on the catalyst, isobutyric acid [*79-31-2*] and butyric acid [*107-92-6*] are obtained in varying proportions when propene [*115-07-1*] is carbonylated [49]:

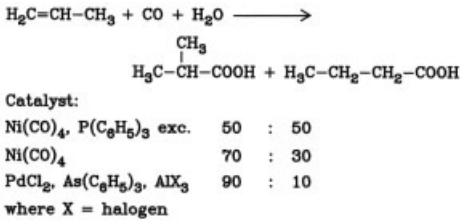

Carbonylation in the presence of methanol with ruthenium carbonyls as catalyst at 190 °C and 20 bar gives a mixture of methyl butyrates, the proportion of the linear isomer being less than 50 % [50].

The carbonylation of propene to give isobutyrates, with subsequent dehydrogenation by means of sulfur [51], I_2– O_2 [52], or O_2 [53], may offer a commercially interesting route to methyl methacrylate [*80-62-6*]:

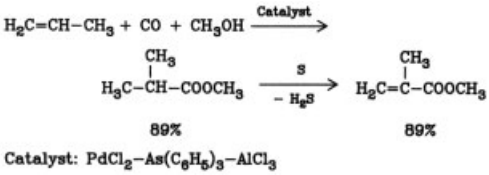

The *synthesis of butanol* from propene and CO is not a carbonylation, but a hydroformylation reaction, according to the definition given in Chapter 1. However, it is often mentioned in connection with Reppe reactions. This synthesis yields a mixture of 85 % of 1-butanol [*71-36-3*] and 15 % of 2-butanol [*78-92-2*] from propene; the catalyst is $H_2Fe(CO)_4$ in *N*-butylpyrrolidine; conditions: 100 °C and 15 bar. The butanol process is no longer operated, but some successful attempts have been made to generalize this reaction by adding a rhodium compound to the catalyst.

Styrene. Because of its strong tendency to polymerize, styrene [*100-42-5*] must be carbonylated at temperatures around 100 °C. Cobalt and palladium catalysts are sufficiently active within this temperature range. The products formed depend considerably on how the palladium catalyst is modified [54]:

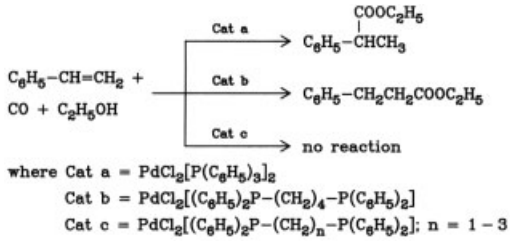

Heptene and Octene. When $SnCl_2$ – $PdCl_2[P(C_6H_5)_3]_2$ is used as catalyst at 70 °C and 140 bar, 1-heptene [*592-76-7*], carbon monoxide, and methanol yield a mixture of methyl octanoates with a 93 % fraction of the linear isomer [*111-11-5*] [55]. The analogous reaction of 1-octene [*111-66-0*] using $FeCl_2$ – $PdCl_2$ as catalyst gives 1-nonanoate [*1731-84-6*] and 2-nonanoate [*2177-86-8*] in a ratio of 4 : 5 [56]. Higher olefins can be alkoxycarbonylated with regioselectivities exceeding 99 % to the linear alkyl caboxylates by use of cationic Pd/dtbpx catalysts [36, 57]:

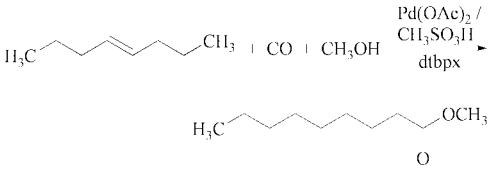

Surfactant Alcohols from Higher Olefins. The Reppe butanol synthesis cannot be applied to long-chain olefins containing nonterminal double bonds. However, a simple route to the esters of long-chain carboxylic acids is afforded by the carbonylation of these olefins in the presence of cobalt salts of fatty acids in combination with pyridine or β-picoline at about 170 °C and 130 bar [58]. This yields the corresponding methyl carboxylates, which can then be hydrogenated to surfactant alcohols; the proportion of linear isomers is 79 % and the overall conversion 70 – 80 % [59]. This route could develop into an economic alternative to the conventional oxo process.

4.2. Substituted Alkenes

Unsaturated Amides. Carbonylation of unsaturated amides in the presence of cobalt

catalysts affords five- or six-membered cyclic imides:

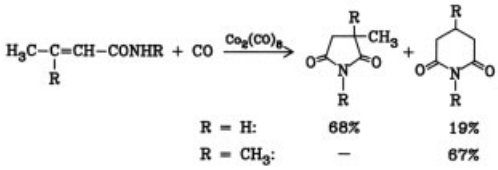

Unsaturated Amines. When unsaturated amines are carbonylated in the presence of $Co_2(CO)_8$, *N*-alkyl pyrrolidinones are obtained. Carbonylation of allyl halides in the presence of primary amines and a rhodium compound affords the same products [60]:

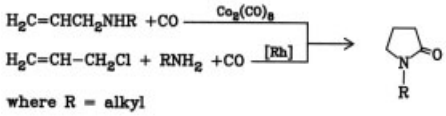

Unsaturated Alcohols. Five- or six-membered lactones are obtained by carbonylation of unsaturated primary alcohols [4]:

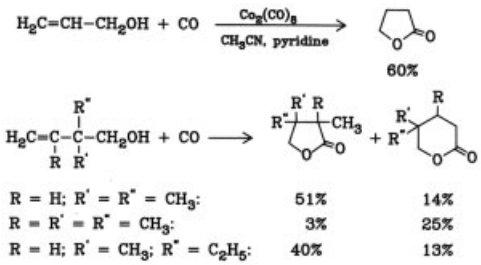

Unsaturated Esters, Acids, and Nitriles. 3-Pentenoic esters as well as pentenoic acid can be obtained by carbonylation of butadiene (see below). 3-Pentennitrile is the primary product of the first stage of butadiene hydocyanation to yield adipodinitrile, a process carried out at large scale at Invista, formerly Du Pont. These substrates can be carbonylated at surprisingly mild conditions (below 10 bar) to yield almost quantitatively (>98 %) the linear difunctional products dialkyladipate, adipic acid, and 5-cyanoalkylvalerate, respectively. The nitriloester can be hydrogenated to yield ε-caprolactam. Thus, these carbonylations are interesting routes to produce nylon intermediates [36, 61, 62] (see bottom of page).

Carbonylation of Oleic Acid. A mixture of isomers composed of 2-octyl-1,9-nonanedicarboxylic acid (**9**), 2-nonylsebacic acid (**10**), and 2-heptyl-1,10-decane-dicarboxylic acid (**11**) is obtained, in a yield of over 90 %, from oleic acid, carbon monoxide, and water in the presence of $Ni(CO)_4$ and iodine as catalyst. Isooleic acid and undecylenic acids react similarly [2, p. 38].

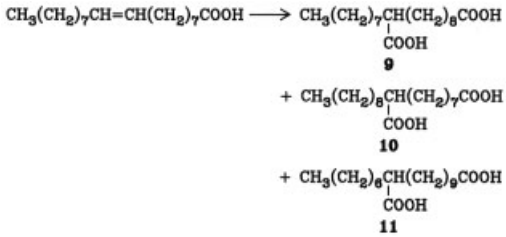

4.3. Polyenes

Butadiene. Temperatures above 140 °C are required for adequate activity of the *nickel catalysts* in the carbonylation of butadiene [*106-*

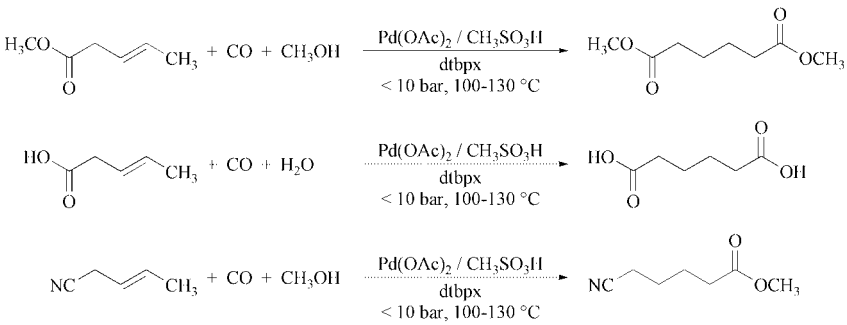

99-0]. However, at these temperatures butadiene dimerizes to a high degree to give 4-vinylcyclohexene or polymerizes in the course of the reaction. For this reason the carbonylation gives a mixture of branched and straight-chain monocarboxylic and dicarboxylic acids derived from vinylcyclohexene [2, p. 38] [40].

The reaction takes place at lower temperatures if *palladium catalysts* are used [13, 63]. 3-Pentenoyl chloride [*17066-20-5*] and Pd^0 are obtained from butadiene and CO in a stoichiometric reaction catalyzed by $PdCl_2$.

In the presence of alcohols, the catalytic reaction gives C_5 or C_9 esters, depending on the catalytic system:

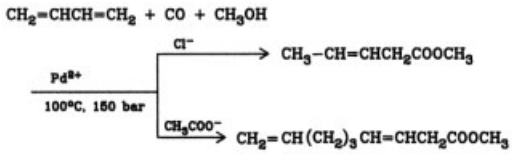

The nature of the anion determines the products formed in an extremely selective manner [64]. Halides prevent the palladium-catalyzed dimerization of butadiene. The addition of nitrogen bases (pyridine, alkylamines) [65], tetraalkylammonium, or tetraalkylphosphonium salts [66] increases the activity of the catalyst.

The use of palladium halides in combination with hydrogen halides increases both the life of the catalyst and the activity of the system. Palladium chloride is preferred; although it has a lower activity than the bromide or iodide, it displays the highest selectivity to the 3-pentenoic acid ester [67]. However, the initial high activity of the catalyst is only retained for a short period, and considerable loss of palladium takes place.

Carbonylation in the presence of water gives the free carboxylic acids. As a rule, higher temperatures are necessary for this reaction. 3-Pentenoic acid [*5204-64-8*] is obtained in a 70 % yield when $Co_2(CO)_8$ – pyridine is used at 160 °C and 250 bar [68], or in an 86 % yield when a palladium compound is used in the presence of tetraalkylammonium salts at 100 °C and 145 bar [69].

A mixture of 4-pentenoic and 3-pentenoic acid esters, in which the 4-isomer predominates, is obtained by means of *rhodium catalysts* [70]. The carbonylation of butadiene or substituted butadienes to give pentenoic acid esters takes place with very high regioselectivity, at 120 – 140 °C and above 200 bar, when *cobalt catalysts*, e.g., $Co_2(CO)_8$ and such nitrogen bases as pyridine, picoline, or isoquinoline are used as catalysts. The selectivity of the system increases as the pyridine concentration is increased (for the reaction mechanism, see Section 2.2). The reactivity decreases in the following series:

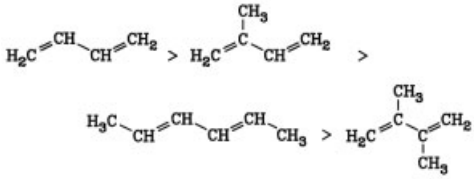

The double carbonylation of butadiene to give adipic acid [*124-04-9*] or its esters is described by many authors. Rhodium catalysts (yield 40 %) [71] or cobalt catalysts [72] are suitable. However, the drastic conditions (200 – 250 °C, 250 – 300 bar) required for this one-stage synthesis are very detrimental to selectivity.

On the other hand, a two-stage process using cobalt catalysts has been developed at pilot plant scale by BASF. It offers a route to adipic acid from the butadiene-containing C_4 cut [9, 73]:

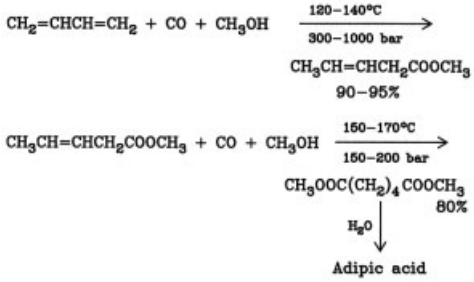

The advantage of this high-pressure two-stage process is that it is possible to optimize both the reaction conditions and the catalyst to suit the individual reactions. The *first stage* requires fairly high concentrations both of cobalt and of pyridine to ensure adequately rapid carbonylation of butadiene; at the same time the temperature may be kept low and both the reaction rate and the selectivity increase as the pressure increases. In the *second stage* the concen-

tration of pyridine must be low because it has an inhibitory effect on the carbonylation of the pentenoic acid ester. The 3-pentenoic acid ester rapidly isomerizes to form the 4-pentenoic acid ester under the conditions of the second stage. Dimethyl adipate [*627-93-0*] is formed with an overall selectivity of ca. 70 %. The catalyst precursor employed is a cobalt carboxylate, from which the actual catalyst, in both stages $HCo(CO)_4$, is formed.

More recently, improved cationic palladium catalysts have been developed both for the methoxycarbonylation [74] and hydroxycarbonylation of butadiene [75] running at pressures of about 50 bar. While Pd/bcope is suited for alkoxycarbonylation, Pd/dtbpx is the preferred catalyst for hydroxycarbonylation. Since the latter catalyst can also be used at reduced pressure for 3-pentenoic acid hydroxycarbonylation (see above), both stages of adipic acid synthesis can be carried out without change of the catalyst system. Adipic acid can be removed from the product mixture by filtration and the yield exceeds 90% based on butadiene. This process has been demonstrated by Shell at laboratory scale.

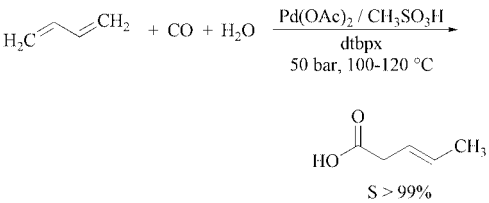

The *oxidative carbonylation* of butadiene using palladium catalysts takes place at moderate temperature.

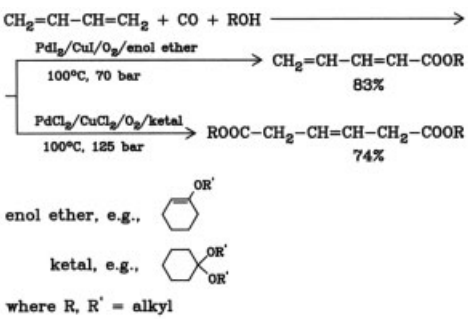

The palladium(0) formed in the reaction is reoxidized by Cu^{2+}, Fe^{3+}, or V^{5+}; enol ethers, acetals, or ketals are employed to bind the water of oxidation, which is harmful to the catalyst [76]. However, the life of the catalyst is short, and there is considerable formation of byproducts.

Acetyldienes are formed from butadiene, CO, and methyl iodide in the presence of $NaCo(CO)_4$ or $Co_2(CO)_8$ by phase transfer catalysis under ambient reaction conditions [27]:

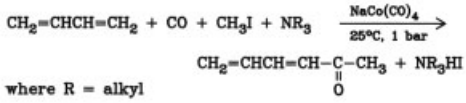

Allenes. In the presence of ruthenium carbonyls, allene [*463-49-0*] gives methyl methacrylate [*80-62-6*] or dimethyl 1,1-dimethyl-3-methyleneglutarate [77]:

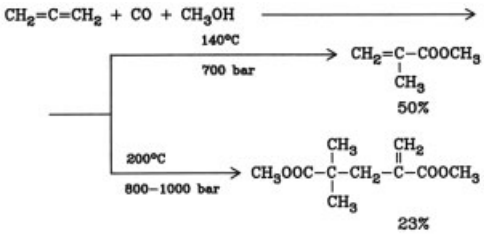

The carbonylation of tetraphenylallene gives a mixture of 1,1,3-triphenylindene (**12**), 2-diphenylmethyl-3-phenylindone (**13**), and 2,2,4-triphenylnaphthalenone (**14**) [4].

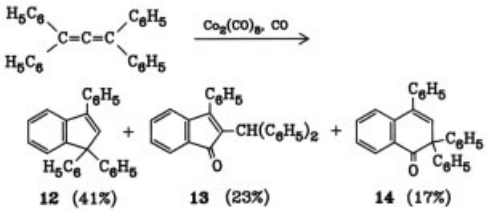

Nonconjugated Polyenes. The carbonylation of nonconjugated dienes using $HCo(CO)_4$ as catalyst gives mixtures of unsaturated and saturated ketones [4]; the overall product yields depend on the distance between the double bonds in the polyene. The reaction proceeds most advantageously if the double bonds are separated by one or two carbon atoms:

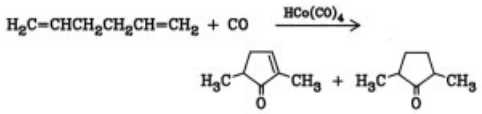

In palladium-catalyzed reactions with CO in the presence of water or alcohols, nonconjugated dienes behave like monoolefins: the double bonds are carbonylated independently and successively [78]:

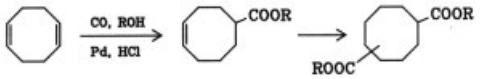

where R = alkyl

1,5,9-Cyclododecatriene [*2765-29-9*] cannot be carbonylated by conventional methods (high temperatures); instead, ring-contraction products are obtained. Only palladium catalysts can be used under conditions sufficiently mild to enable one, two, or all three double bonds to be carbonylated [10].

5. Carbonylation of Alcohols and Ethers

The carbonylation of alcohols to give carboxylic acids is by far the most important application of carbonylation in industry. The most important individual reaction of this type is the synthesis of acetic acid [*64-19-7*] from methanol. World production of this chemical is at present about 10×10^6 t/a. The bulk of this is obtained by the carbonylation of methanol, using rhodium or iridium catalysts based on the Monsanto process (→ Acetic Acid). In addition, the cobalt-catalyzed BASF process is still carried out at a relatively small plant in Ludwigshafen.

5.1. BASF Acetic Acid Process

The BASF process was optimized at the beginning of the 1950s. It requires high pressures (> 500 bar) and temperatures of about 250 °C [79]. The catalyst precursor is cobalt iodide [*15238-00-3*], CoI_2, and HI (or the less corrosive KI) in water. $Co_2(CO)_8$ is formed in the course of the reaction, and is then converted to $HCo(CO)_4$ in the presence of hydrogen (formed in side reactions). The active catalytic system is the combination of $HCo(CO)_4$ and HI [80]:

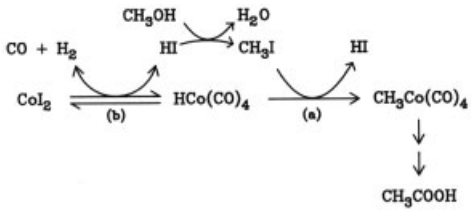

The high pressure is required to favor reaction (a), which yields the product, over reaction (b), which deactivates the catalyst.

In the past, some research work was carried out by BASF and Shell with the aim of reducing the temperature to 80 – 200 °C and the pressure to 70 – 300 bar [81].

5.2. Monsanto and Related Low-Pressure Acetic Acid Processes

The Monsanto low-pressure process was optimized in the mid-1960s. It uses a rhodium catalyst [82]. At present, about 75 % of the world's acetic acid is made by this process or by processes derived therefrom; the fraction is increasing, because further plants using this technology will start up in the next few years [83].

The process operates under very mild conditions. The reaction already takes place at a CO pressure of 1 bar, but commercial plants use a pressure of 30 bar and a temperature of approximately 175 °C. The selectivity is 98 %, based on methanol, which is unusually high for catalytic reactions.

The kinetics have been investigated in detail and were shown to depend first order on methyl iodide and rhodium/iridium concentration. Remarkably, neither methanol nor CO partial pressure are part of the rate law [84]. The oxidative addition of methyl iodide to Rh^+ constitutes the rate determining step. In addition, the iodide has the function of protecting Rh^+ against reduction by CO [85]. The following mechanism has been proposed for this reaction:

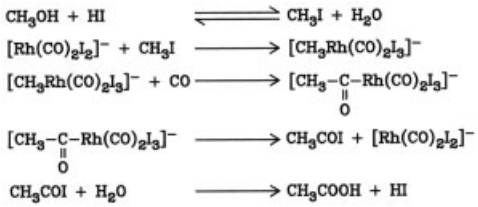

One problem of the Monsanto process is the high methyl iodide concentration required in the process. Therefore, numerous attempts have been made to absorb the rhodium onto carriers and to employ it in a halogen-free form, but it proves to be a poor heterogeneous catalyst [86]. On the other hand, nickel on zeolites or activated carbon gives a 95 % yield of acetic acid at 300 °C and 11 bar [87].

The most successful Monsanto process uses anionic rhodium iodide carbonyl complexes in the presence of large amounts of methyl iodide at water concentrations of at least 8 % to avoid catalyst decomposition and formation of insoluble rhodium salts. Unfortunately, the rhodium complexes applied also catalyze the water gas shift reaction, which results in loss of CO and side product formation such as acetaldehyde derivatives and propionic acid. Further developments have led consequently to the "low-water" processes, working at below 8 % water [16, 88]. These processes feature increased space–time yields and further improved selectivities. The Cativa process patented by BP used iridium catalysts, promoted by a second metal carbonyl, probably ruthenium. It is speculated that the role of the promoter is to abstract iodide to form a neutral iridium acetyl intermediate. In the Celanese low-water process, lithium iodide is used to stabilize rhodium at low water conditions and this is also the case in the acetic anhydride process by Eastman chemicals. A joint development by UOP and Chiyoda has led to a liquid phase process utilizing a heterogeneous polyvinylpyridine supported rhodium iodide catalyst [89].

5.3. Carbonylation of Other Alcohols and Ethers

In general, the carbonylation of primary and secondary alcohols takes place with good yields. Tertiary alcohols mainly yield olefins, as a result of dehydration. Diols may also be carbonylated to give the corresponding dicarboxylic acids or esters, but as a rule, in poorer yields and selectivities than monoalcohols. Branched isomers, monocarboxylic acids, hydroxycarboxylic acids, or lactones are formed as byproducts. Table 3 gives a survey of further carbonylation reactions of alcohols; see also [4, p. 116].

The oxidative carbonylation of methanol has found industrial application. The CuCl-catalyzed synthesis of dimethyl carbonate by the Enichem process is used at Ravenna, Italy, and at Chiba, Japan, with a combined capacity of 20 000 t/a. The product from the latter plant is used to generate diphenyl carbonate, an intermediate in the polycarbonate process by General Electric [44].

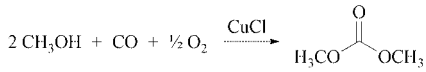

Ube Industries has perfected a process for the oxidative coupling of CO to give oxalic acid esters [90]. Although this is not a Reppe carbonylation in the classical sense, the reaction is related to the oxidative carbonylation of olefins (see Section 2.3). Alkyl nitrites are used as the oxidizing agent in the process; this avoids the handling of molecular oxygen. The synthesis is carried out in the presence of homogeneous [91] or heterogeneous [92] palladium catalysts at 70 °C and 60 – 70 bar; it yields dimethyl oxalate [*553-90-2*] or dibutyl oxalate [*2050-60-4*] in a selectivity of 95 %:

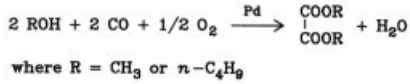

Hydrogenation of the oxalate affords a potential route to ethylene glycol [*107-21-1*], independent of ethylene.

Allyl Alcohol. The carbonylation of allyl alcohol [*107-18-6*] using palladium catalysts in the presence of alcohols gives 3-butenoic acid esters [93]. In the absence of alcohols, allyl vinylacetate [*1745-31-9*] is obtained in 88 % yield at 80 °C and 200 bar, when a catalyst containing $PdCl_2$, $P(C_6H_5)_3$, and $SnCl_2$ [60] is used:

Table 3. Carbonylation of alcohols

Alcohol	Catalyst	Reaction conditions	Product (yield)
CH_3OH	Ni/I_2	50 – 150 bar, 150 – 280 °C	CH_3COOH (95 %)
C_2H_5OH	Ni/I_2	120 bar, 250 °C	C_2H_5COOH (76 %)
$(CH_3)_2CHOH$	Ni/I_2	300 bar, 300 °C	isobutyric acid (68 %)
$C_6H_5CH_2OH$	Rh/I_2	71 bar, 175 °C	phenylacetic acid (83 %)
$HOCH_2CH_2OH$	$Ni(CO)_4/I_2$	200 bar, 250 °C	succinic acid (15 %)
$HO(CH_2)_4OH$	$Ni(CO)_4/I_2$	200 bar, 260 °C	adipic acid (69 %) *
$HO(CH_2)_4OH$	$RhCl_3/HI$	70 bar, 200 °C	adipic acid (38 %) *
$HO(CH_2)_5OH$	$Ni(CO)_4/I_2$	200 bar, 250 °C	pimelic acid (94 %) *
$HO(CH_2)_6OH$	$Ni(CO)_4/I_2$	200 bar, 260 °C	suberic acid (90 %) *

* Including the branched isomers.

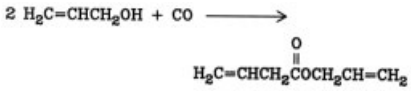

Benzyl Alcohols. An industrial benzyl alcohol carbonylation is the Ibuprofen synthesis which is often cited as a textbook example of green chemistry because of its atom efficiency. The process was developed in a joint venture by Hoechst-Celanese and Boots and is now operated by BASF. The feedstock is isobutylbenzene which is acylated in the para position with acetic anhydride by use of HF as the catalyst. The corresponding acetophenone is hydrogenated to the benzyl alcohol which is carbonylated by use of $PdCl_2/PPh_3$ in concentrated HCl. This process is carried out at a scale of >3000 t/a at Bishop/Texas and is the most important synthesis of the inflammatory drug Ibuprofen [94].

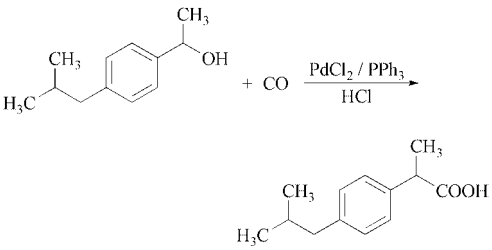

Aliphatic Ethers. When symmetrical aliphatic ethers react with carbon monoxide, the corresponding carboxylic acid esters are formed; small amounts of free carboxylic acids are always present in the reaction mixture, unless the reaction is carried out under strictly anhydrous conditions. Nickel (in the form of metal, carbonyl, or complex), cobalt, or iron (in the form of salts or on carriers) can be employed as catalysts; a halide is required as promoter [3]:

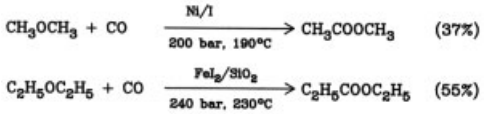

Cycloaliphatic Ethers. Carbonylation of cyclic ethers leads to various products, depending on the nature of the ether [97]. Epoxides first yield hydroxycarboxylic acids [98], but these eliminate water at fairly high reaction temperatures with the formation of α,β-unsaturated carboxylic acids or esters. β-Hydroxy esters can be used to produce 1,3-diols via hydrogenation.

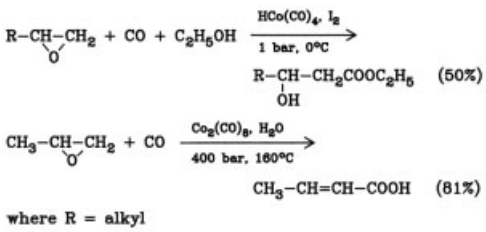

In absence of water or alcohols, epoxides may also be carbonylated to β-lactones. For example β-butyrolactone is obtained with high selectivity (98 %) by the carbonylation of propylene oxide catalyzed by a cobalt carbonyl complex and a Lewis acid as co-promoter ($BF_3 \cdot Et_2O$ or a Schiff base-aluminium complex) [95]. Using a cobalt carbonyl and 3-hydroxypyridine as the ligand, ethylene oxide and carbon monoxide may be copolymerized to a poly(3-hydroxypropionate) [96]. Molecular masses up to a few thousand grams per mole can be obtained.

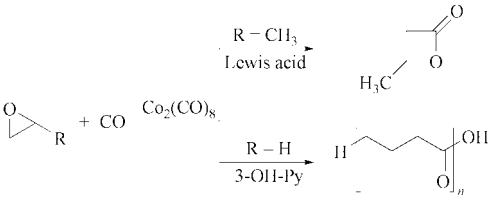

Butyrolactone [*96-48-0*] may be obtained by carbonylation of oxetane [*503-30-0*] [97]:

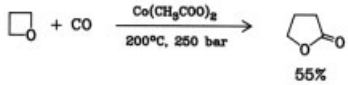

The carbonylation of tetrahydrofuran [*109-99-9*] leads to adipic acid [*124-04-9*] [99] or aldehyde derivatives [100], depending on the catalytic system:

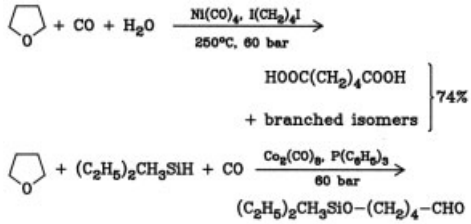

6. Carbonylation of Organic Halides

Saturated and unsaturated organic halides can be carbonylated to give carboxylic acid halides. The corresponding acids, esters, or amides are obtained in the presence of water, alcohols, or amines, respectively. In general, the reactions take place at moderate temperature (up to 150 °C) and at only slightly increased pressure; suitable catalysts are $Ni(CO)_4$, nickel(0) – phosphine complexes, palladium(0) – phosphine complexes, palladium – allyl complexes, $RhCl_3$, or $NaCo(CO)_4$. The use of $Co_2(CO)_8$ as catalyst normally requires high temperatures and pressures, but in that case, carbonylation also takes place under conditions as mild as 20 °C and atmospheric pressure if phase transfer catalysts are present [101].

The selectivity of the reactions with organic halides depends on the stability of the carbon – halogen bond to cleavage into radicals [102]. The reaction of these radicals with the solvent and radical recombinations lead to further byproducts in addition to those of the carbonylation reaction (ketones, diketones, and aldehydes) [103]. Due to the mild conditions that can be applied with palladium catalysts, this method has found increasingly application in fine-chemical synthesis [104, 105]. Gaseous carbon monoxide can be at least partially replaced by use of formic acid or formates as the carbonylating agent [106].

6.1. Aryl Halides

In the presence of such bases as $Ca(OH)_2$, aryl halides are converted into the calcium salts of the corresponding aromatic carboxylic acids. Reaction conditions are as mild as atmospheric pressure and 100 °C in polar, aprotic solvents, such as dimethylformamide (DMF) or dimethyl sulfoxide (DMSO), if $Ni(CO)_4$ is the catalyst. The reaction also takes place in less polar solvents, but it then requires higher temperatures and pressures. In tetrahydrofuran (THF), benzil [*134-81-6*] is obtained in 80 % yield. Phthalic anhydride [*85-44-9*] and benzene [*71-43-2*] are formed from bromobenzene in the presence of $Ni(CO)_4 - Na_2CO_3$ [107]:

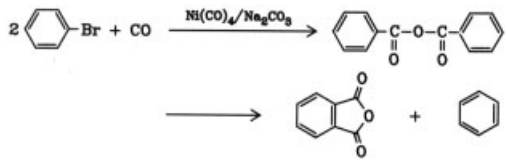

Aroyl halides are obtained in high yield at 80 bar and 160 °C when palladium halides are used as the catalyst [108]. In the presence of water, alcohols, or amines, aryl halides are also carbonylated to aryl carboxylic acids, esters, or amides by using palladium catalysts with chelate phosphines as the ligand. A stoichiometric amount of base (or preferably in excess) is necessary to capture the hydrogen halide formed by the reaction. The order of reactivity of aryl halides is directly correlated to the strength of the aryl–halide bond (bond dissociation energy): ArI > ArBr > ArCl. Aryl fluorides are unreactive. The reactions conditions are quite mild: less than 30 bar and < 150 °C. Aryl halides substituted by

electron attracting groups are by far more reactive than those substituted by electron donating groups, since the rate-determining step of the reaction is usually the oxidative addition of the aryl halide to the Pd(0)–phosphine complex.The progress in aryl halide carbonylation has been reviewed by Beller [109 – 111].

6.2. Alkyl Halides

Alkyl halides react with carbon monoxide only in strongly polar solvents such as dimethylformamide or ethanol [112]. In general, alkyl halides are appreciably more difficult to carbonylate than aryl halides.

The carbonylation of benzyl halides is easier than that of alkyl halides but not as easy as that of aryl halides. With $Ni(CO)_4$ as catalyst, it has been thoroughly investigated; the products depend very much on the solvent [113]:

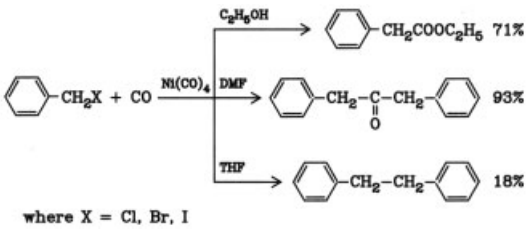

A catalytic process for the synthesis of phenylacetic acid [*103-82-2*] from benzyl chloride uses $Ni(CO)_4$ as catalyst. The product is formed with a selectivity of 95 % at 60 bar and 80 °C [114]. Phenylacetic acid or its esters can be made under even milder conditions when $[Fe(CO)_3NO]^-$ is used as the catalyst; yields between 80 and 90 % are achieved at 1 – 10 bar and temperatures below 80 °C [115]. In the laboratory phenylacetic acid esters can be made from substituted or unsubstituted benzyl halides with very high yields under catalysis by rhodium and aluminum alkoxides [116]. A process in which $Co_2(CO)_8$ and sodium methoxide are used as the catalyst is carried out on an industrial scale at 10 bar and 50 – 60 °C [112, 117]. α-Chlorocarboxylates are carbonylated to dicarboxylates by using $Co_2(CO)_8$ as the catalyst in the presence of a stoichiometric amount of base (potassium carbonate, sodium alkoxide). Dynamit Nobel has developed a synthesis of dimethyl malonate based on the carbonylation of methyl chloroacetate; this process runs under mild conditions (<100 °C, <20 bar CO) with a good selectivity (95 %). Dialkyl malonates and alkyl phenylacetates are produced by these carbonylation processes at a scale of several thousand tons per year:

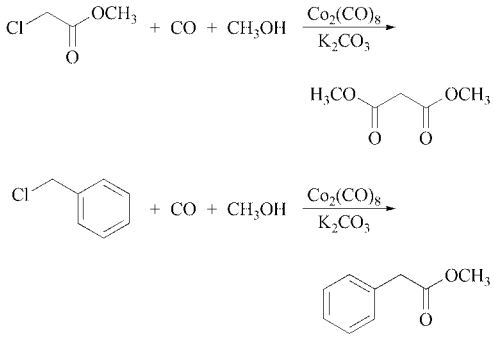

The formation of phenylpyruvic acid [*156-06-9*] from benzyl chloride in a yield of 82 % is a rare example of a double carbonylation [118]:

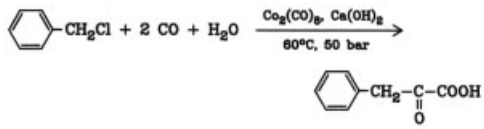

Primary alkyl bromides can be carbonylated to give aldehydes in over 90 % yield when $Na_2Fe(CO)_4 - P(C_6H_5)_3$ is used as catalyst; secondary halides and benzyl halides react less readily [119]. With the same catalyst, it is also possible to couple different alkyl halides to give unsymmetrical ketones; stoichiometric amounts of $Na_2Fe(CO)_4$ are required in this case [120].

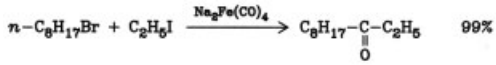

6.3. Allyl Halides and Vinyl Halides

When various catalysts are used, the carbonylation of allyl halides or allyl acetate in the presence of an alcohol gives high yields of 3-butenoic acid esters [121]. In the presence of ethylene, carbonylation is accompanied by chain-lengthening [122]:

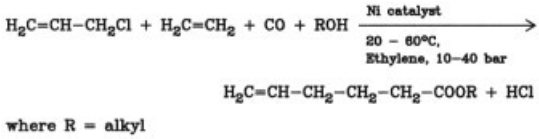

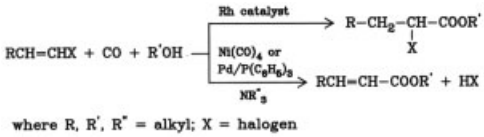

A similar insertion reaction employing butadiene is catalyzed by palladium (70 °C; 40 –50 bar); it gives 3,7-octadienoic acid chloride or ester [123]. If allyl derivatives are carbonylated in the presence of acetylene by means of $Ni(CO)_4$ and water or methanol at 20 °C, the acetylene adds onto the allyl halide and 2-*cis*-5-dienoic acids or esters [124] or sorbic acid esters [125] are obtained.

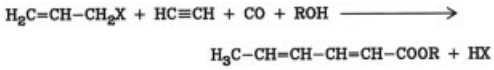

On the other hand, 2-cyclohexen-1-one [*930-68-7*] and phenol [*108-95-2*] are formed if the same reaction is carried out without the alcohol in aprotic solvents. If catalytic amounts of HCl are added, allyl alcohol, allyl ethers, or allyl esters can be used as the starting material instead of allyl halides in the reactions mentioned above.

A mixture of 1-chloro-2-butene and 3-chloro-1-butene gives 3-pentenoyl chloride [*17066-20-5*] in a virtually quantitative yield, when palladium – allyl complexes in benzonitrile are used as catalyst [126].

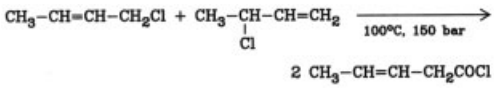

Vinyl Halides. Carbonylation of the double bond to give 2-chloropropionates takes place predominantly when unmodified rhodium or palladium catalysts are used. Carbonylation of the carbon – halogen group, however, which leads to acrylic acid esters, takes place when the catalyst is $Ni(CO)_4$ or a palladium compound together with $P(C_6H_5)_3$ in the presence of proton scavengers (amines or alkoxides) [60, 127]:

7. Other Carbonylation Reactions

7.1. Amines

Carbonylation of primary and secondary aliphatic or aromatic amines normally gives a mixture of *N*-formyl derivatives and ureas [128]. In the case of aliphatic amines, the product ratio can be controlled by varying the reaction conditions. With the exception of osmium and iridium, all metals of the eighth subgroup have been found suitable as catalysts for this reaction. The following equations show how the product distribution depends on the catalytic system [129 – 131]:

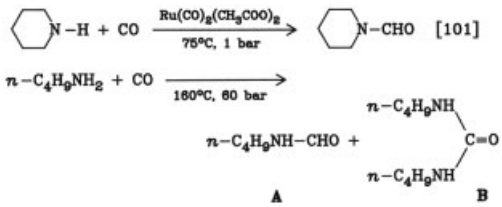

Catalyst	A	B	
$[Rh(CO)_2Cl]_2$	35	: 65	
$[Rh(CO)_2Cl]_2/P(CH_3)_3$ (1:6)	100	: 0	
Se	0	: 100	[102]

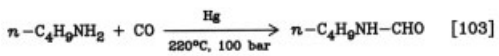

When *tertiary amines* are carbonylated, aliphatic and aromatic amines give different reaction products. Aliphatic amines yield *N*-formyl derivatives, with the elimination of olefin, whereas aromatic amines produce carboxamides [132]. Tetracarbonyl nickel or complex nickel or cobalt compounds, such as $(NR_4)_2CoBr_4$ or $(PR_4)_2CoBr_2I_2$ (R = alkyl), are suitable as catalysts. Temperatures range from 100 to 250 °C; pressures are 100 – 700 bar.

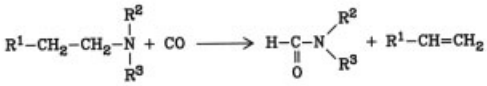

where R^1, R^2, R^3 = alkyl

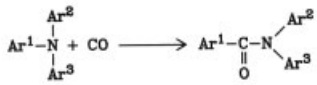

where Ar^1, Ar^2, Ar^3 = aryl

Carbonylation by means of enneacarbonyl diiron [*15321-51-4*], $Fe_2(CO)_9$, in a stoichiometric reaction offers a route to isocyanates starting from *N,N*-dichloroamines or *N,N*-dichloroamides [133] . The oxidative carbonylation of amines, using Pd^{2+} – Cu^{2+} catalysts, to give ureas [14] or formamides [134] is more favourable thermodynamically than the nonoxidative reaction. The catalytic cycle is maintained by reoxidizing Pd^0 by means of air in the presence of Cu^{2+}.

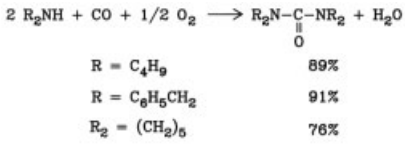

These reactions are of importance as potential routes to isocyanates; however, the yield of urea and formamide conversion to isocyanates is not yet satisfactory [135, 136]. The direct catalytic carbonylation of primary amines to give isocyanates, using $PdCl_2$ as catalyst, has also been described [134], but this reaction takes place very slowly and with poor yields. At present, all routes to isocyanates via oxidative carbonylation of amines appear less promising than those in which nitro compounds are used as the starting material (see following Section).

7.2. Aromatic Nitro Compounds

Dinitrotoluene (the mixture of 2,4- and 2,6-isomers) or nitrobenzene are used as educts for carbonylation reactions carried out with the aim of producing toluylene diisocyanate (TDI) [*26471-62-5*] or methylene di(phenylisocyanate) (MDI) [*101-68-8*], respectively. The overall reaction takes place as follows:

Both single-stage processes, in which isocyanates are synthesized directly from aromatic nitro compounds, and two-stage processes, in which urethanes are intermediates, have been described [137]. At present the two-stage processes is more promising. The single-stage synthesis requires pressures of over 200 bar and temperatures between 190 and 200 °C. A high concentration of catalyst is required. The most active catalysts are a combination of a palladium compound, a nitrogen donor ligand, and a Lewis acid. Thus, when a catalyst system containing $PdCl_2$, quinoline, and $GeCl_4$, is used, toluylene diisocyanate is obtained from dinitrotoluene with a selectivity of 60 % [138]. Rhodium, nickel, cobalt, or iron have also been described as catalysts [139].

The advantages of the two-stage process are the milder conditions required (170 °C; 100 – 150 bar) and, particularly, the fact that the carbonylation stage requires only about one-tenth the amount of catalyst used in the single-stage process. $PdCl_2$ together with various cocatalysts [140] and selenium together with pyridine and sodium acetate or lithium acetate have been described as catalysts [141, 142]. In the case of palladium, stability problems are encountered. Selenium, on the other hand, is difficult to remove from the reaction products. The route followed by the two-stage isocyanate synthesis is shown below:

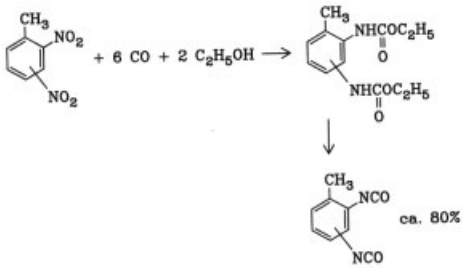

7.3. Esters

The carbonylation of carboxylic acid esters yields anhydrides, or the free carboxylic acids in the presence of water. The catalysts employed

are mainly based on nickel, cobalt, or rhodium [97]. The carbonylation of methyl acetate [*80-62-6*] to give acetic anhydride [*108-24-7*] is an important industrial reaction (→ Acetic Anhydride and Mixed Fatty Acid Anhydrides). If $[(C_6H_5)_3PC_4H_9]_2NiBr_4$ is used as catalyst, acetic anhydride is obtained at 180 – 190 °C and 650 – 700 bar in 43 % yield [143]. If nickel metal together with methyl iodide and basic cocatalysts, such as lithium acetate and lutidine [144], or a system composed of nickel, CH_3I, $Sn(CH_3COO)_2$, and CH_3COOLi [145] is employed, the reaction takes place at 180 – 200 °C and 80 bar with a yield of 60 – 70 %. Acetic anhydride can be obtained in comparable yields at 20 – 50 bar and 160 – 180 °C if $RhCl_3$ and methyl iodide as well as promoters, such as $Cr(CO)_6$, zirconium, hafnium, lithium iodide, or trialkyl arsenic [146], are used.

In 1983 Eastman Kodak started operation of an acetic anhydride plant with a capacity of 230 000 t/a, which is based on the carbonylation of methyl acetate (Halcon licence) and uses coal as the feedstock [147]. Additional capacity of 50 000 t/a will come on stream in 2008 at a acetyls plant owned by Sipchem, a joint venture in Jubail Industrial City, Saudi Arabia. This plant also uses Eastmans coproduction technology and will also produce 460 000 t/a of acetic acid. The mechanism of methyl acetate carbonylation is similar to the related methanol carbonylation. Salt promoters like LiI and a low partial pressure of H_2 are essential to maintain catalyst activity [12]. Plants utilizing the carbonylation technology for acetic anhydride are generally "swing" plants that can produce both acetic acid and acetic anhydride. BP Chemicals also operates such a plant in Hull, United Kingdom.

7.4. Lactones and Aldehydes

Lactones can be carbonylated, like cyclic ethers, to give dicarboxylic acids (see Section 5.3). They are probably formed as intermediates in the carbonylation of cyclic ethers. Glutaric acid [*110-94-1*] is obtained from γ-butyrolactone, and adipic acid [*124-04-9*] from δ-valerolactone [148].

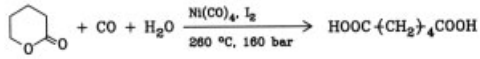

Aldehydes. The reactions of aldehydes with carbon monoxide are hitherto limited to formaldehyde [*50-00-0*]. This reaction requires the presence of proton-donating solvents and yields glycolic acid [*79-14-1*] or its esters. The catalysts used are nickel, cobalt, or iron [149] or alumosilicates [150]. The reaction conditions are 180 – 250 °C and more than 100 bar; the selectivities achieved are higher than 90 %. A better route from formaldehyde to glycolic acid is the Koch Carbonylation (see Chap. 1) [5, p. 41].

7.5. Amidocarbonylation of Aldehydes

The cobalt carbonyl catalyzed carbonylation of aldehydes in the presence of alkylamides is a versatile route to *N*-acyl α-amino acids [137, 152]. The reaction proceeds via aminal intermediates and can be coupled with aldehyde generating reactions like hydroformylation, epoxide rearrangement, or allyl alcohol rearrangement in one stage syntheses:

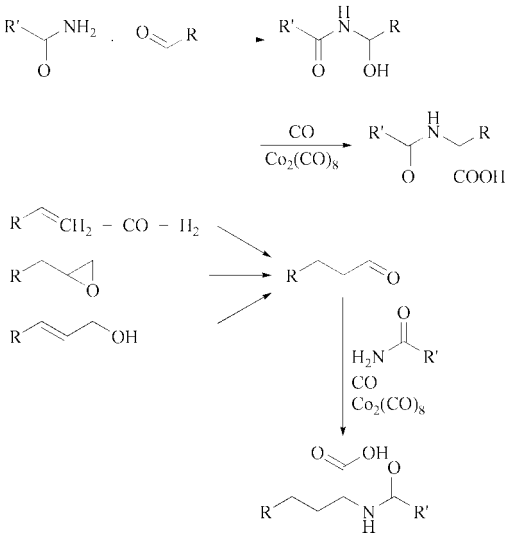

Amidocarbonylation of formaldehyde yields glycine derivatives. If this reaction is carried out in the presence of *N*-methyl laurylamide, *N*-lauroyl sarconisate is obtained, a product which is used as a detergent. If long-chain olefins

are amidocarbonylated in the presence of acetamide, long chain α-acetamido acids are obtained, which are also of interest as detergents:

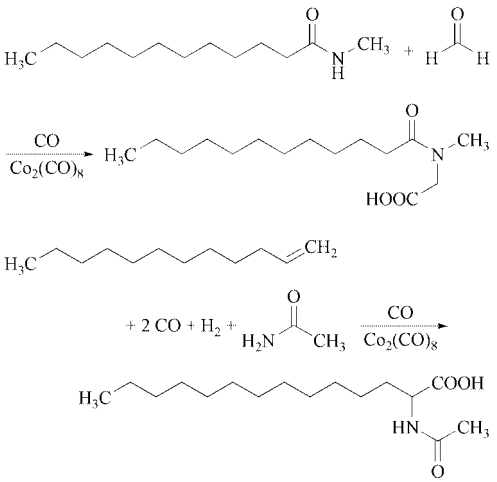

Selectivities of up to 95 % can be achieved and the stability of the cobalt carbonyl can be improved by ligands such as 1,3-bis(diphenylphosphino)propane. Further potential applications are the synthesis of intermediates for sweeteners such as aspartame, food additives such as glutamic acid, and of chelating agents such as the polyamido acids. An advantage of this synthesis is that it can substitute HCN or halogenated intermediates. Palladium catalysts have also been investigated for this reaction. They offer a higher catalyst activity by a factor of about 100 over cobalt and a turnover number of 60 000 has been achieved in the case of *N*-acetylleucine. Palladium catalysts also allow the synthesis of hydantoins from substituted ureas as the amide equivalent and of arylglycines from substituted benzaldehydes [153].

8. References

1. W. Reppe: *Neue Entwicklungen auf dem Gebiet der Chemie des Acetylens und Kohlenoxids*, Springer Verlag, Berlin 1949, pp. 94 – 126.
2. W. Reppe et al.: "Carbonylierung I – VI," *Liebigs Ann. Chem.* **582** (1953) 1 – 161.
3. A. Mullen in J. Falbe (ed.): *New Syntheses with Carbon Monoxide*, Springer Verlag, Berlin-Heidelberg-New York 1980, pp. 243 – 308.
4. J. Falbe: *Carbon Monoxide in Organic Synthesis*, Springer Verlag, Berlin-Heidelberg-New York 1970, pp. 15, 78 – 122.
5. K. Weissermel, H. J. Arpe: *Industrial Organic Chemistry*, 4th ed., Verlag Chemie, Weinheim, Germany 2003.
6. *Ullmann*, 4th ed., **9**, 155.
7. M. Orchin, *Acc. Chem. Res.* **14** (1981) 259.
8. BASF, DE 2103454, 1971; BE 757608, 1971; DE 2139630, 1971.
9. A. Matsuda, *Bull. Chem. Soc. Jpn.* **46** (1973) 524.
10. K. Bittler, N. v. Kutepow, D. Neubauer, H. Reis, *Angew. Chem.* **80** (1968) 352; *Angew. Chem. Int. Ed. Engl.* **7** (1968) 329. J. Tsuji: *Organic Syntheses with Palladium Compounds*, Springer Verlag, Berlin-Heidelberg-New York 1980.
11. G. Kiss, *Chem. Rev.*, **101** (2001) 3435 – 3456.
12. M. Beller, J. G. E. Krauter, in: B. Cornils, W. A. Herrmann (eds.): *Aqueous-Phase Organometallic Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2004, pp. 606 – 612.
13. J. K. Stille, D. E. James in S. Patai (ed.): *The Chemistry of Double-bonded Functional Groups*, Suppl. A., part 2, Wiley, New York 1977, p. 1099.
14. V. A. Golodov, Y. L. Sheludyakov, D. V. Sokolsky in M. Tsutsui (ed.): *Fundamental Research in Homogeneous Catalysis*, **vol. 3**, Plenum Press, New York 1979, pp. 239 – 256.
15. A. Schoenberg, I. Bartoletti, R. F. Heck, *J. Org. Chem.* **39** (1974) 3318. T. A. Weil, L. Cassar, N. Foá in I. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls*, **vol. 2**, Wiley, New York 1977, pp. 517 – 543.
16. P. Torrence, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 104 – 136.
17. W. Reppe: "Carbonylierung I," *Liebigs Ann. Chem.* **582** (1953) 1.
18. P. M. Kohn, *Chem. Eng. (N.Y.)* **86** (1979) 49.
19. A. Yakubovic, E. Volkova, *Dokl. Adad. Nauk SSSR* **84** (1952) 1183; *Chem. Abstr.* **47** (1953) 3278 a. J. Happel, S. Umemura, Y. Sakakibara, H. Blanck, S. Kunichika, *Ind. Eng. Chem. Process Des. Dev.* **14** (1975) 44.
20. J. Tsuji, T. Nogi, *J. Org. Chem.* **31** (1966) 2641.

21. E. Drent, W. W. Jager, J. J. Keijsper and F. G. M. Niele, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 316 – 328.
22. M. Foá, L. Cassar, *Gazz. Chim. Ital.* **102** (1972) 85.
23. J. Tsuji in E. Taylor, H. Wynberg (eds.): *Advances in Organic Chemistry*, **vol. 6**, Interscience, New York 1969, pp. 109, 177.
24. J. Tsuji, M. Takahashi, T. Takahashi, *Tetrahedron Lett.* 1980, 849.
25. J. R. Norton, K. E. Shenton, J. Schwartz, *Tetrahedron Lett.* 1975, 51.
26. T. Mise, P. Hong, H. Yamazaki, *Chem. Lett.* 1981, 993.
27. H. M. Colquhoun, *Chem. Ind. (London)* 1982, 747.
28. J. Falbe, *J. Organomet. Chem.* **94** (1975) 213.
29. D. Forster, A. Hershman, D. E. Morris, *Catal. Rev. Sci. Eng.* **23** (1981) 89 – 105.
30. B. Fell, *Fette Seifen Anstrichm.* **76** (1974) 193.
31. N. S. Imyanitov, D. M. Rudkovskii, *J. Appl. Chem. USSR (Engl. Transl.)* **41** (1968) 157.
32. J. F. Knifton, *J. Org. Chem.* **41** (1976) 793, 2885.
33. R. I. Pugh, E. Drent, in: A. Sen (ed.): "Catalytic Synthesis of Alkylene-Carbon Monoxide and Cooligomers," *Catalysis by Metal Complexes* **27** (2003) 9 – 35.
34. E. Drent, J. A. M. Broekhoven, and P. H. M. Budzelaar, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 344 – 361.
35. Z. Freixa, P. W. N. M. van Leeuwen, *Dalton Transactions* (2003) 1890 – 1901.
36. Shell, WO 2001068583, 2000.
37. B. El Ali, H. Alper, in: E. Negishi (ed.) *Handbook of Organopalladium Chemistry for Organic Synthesis*, Hoboken 2002, 2333 – 2349.
38. Monsanto, DE 1941501, 1969;
DE 2310808, 1972;
DE-OS 2324765, 1972.
39. BASF, DE 1259867, 1964;
BE 651532, 1964;
FR 1406194, 1964;
DE 1249867, 1964;
BE 674069, 1965;
GB 1123367, 1965.
40. R. E. Brooks, W. F. Gresham, J. V. Hardy, J. M. Lupton, *Ind. Eng. Chem.* **49** (1957) 2004.
41. BASF, WO 2002026382, 2000.
42. Lucite, WO 2005079981, 2004.
43. D. M. Fenton, K. L. Olivier, G. Biale, *Prepr. Div. Pet. Chem. Am. Chem. Soc.* **14** (1969) C 77–84.
44. A. Klausener, J.-D. Jentsch, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 164 – 182.
45. Union Oil, US 3523971, 1970 (G. Biale).
46. K. Murata, A. Matsuda, *Bull. Chem. Soc. Jpn.* **54** (1981) 2089.
47. E. Drent, P. H. Budzelaar, *Chem. Rev.* **96** (1996) 663.
48. BASF, DE-AS 1276029, 1964;
BE 673867, 1965.
49. S. Kunichika, S. Sakakibara, T. Okamoto, K. Takagi, *Bull. Chem. Soc. Jpn.* **44** (1971) 3405.
Du Pont, DE-OS 2739096, 1977 (E. Squire, F. Waller).
50. G. Sbrana, G. Braca, F. Piacenti, G. Mariano, M. Bianchi, *Chim. Ind. (Milan)* **54** (1972) 117.
51. Asahi Chem., DE 2129920, 1971;
DE 2208580, 1972.
Du Pont, DE-OS 2739096, 1978.
52. Du Pont, US 2719171, 1955.
BASF, DE-OS 2064576, 1972.
53. American Cyanamid, DE-OS 2118904, 1972.
54. Y. Sugi, K. Bando, S. Shin, *Chem. Ind. (London)* 1975, 397.
55. Texaco, DE-OS 2303118, 1972 (J. F. Knifton).
56. Union Oil, US 3661949, 1969 (D. M. Fenton).
57. C. J. Rodriguez, D. F. Foster, G. R. Eastham, D. J. Cole-Hamilton, *Chem. Comm.* (2004) 1720 – 1721.
58. B. Fell, W. Seide, V. Vogt, F. Asinger: "Die Hydrocarboxylierungsreaktion ungesättigter Verbindungen mit Nickelcarbonyl- und Kobaltcarbonyl-Komplex-katalysatoren," *Forschber. Landes Nordrhein-Westfalen* 1973, no. 2357, 91 – 123.
59. P. Hofmann, *Fette Seifen Anstrichm.* **85** (1983) 126.
60. J. F. Knifton, *J. Organomet. Chem.* **188** (1980) 223.
61. Shell, WO 2001068583, 2000;
BASF, WO 2002046143, 2000.
62. Lucite, WO 2004014834, 2002.
63. J. F. Knifton, *Ann. N.Y. Acad. Sci.* **333** (1980) 264.
64. J. Tsuji, *Acc. Chem. Res.* **6** (1973) 8.
65. Texaco, US 4124617, 1977 (J. F. Knifton).
66. Rhône-Poulenc, EP 75524, 1981.
67. S. Hosaka, J. Tsuji, *Tetrahedron* **27** (1971) 3821.

68. N. S. Imyanitov, B. E. Kuvaev, D. M. Rudkovskii, *Zh. Prikl. Khim. (Leningrad)* **38** (1965) 2558.
69. Rhône-Poulenc, F 2529885, 1982.
70. Esso, US 3161672, 1961; US 3253018, 1962. Phillips, US 3746747, 1973.
71. BASF, DE-OS 2037782, 1970.
72. N. S. Imyanitov, D. M. Rudkovskii, *Zh. Org. Khim.* **2** (1966) 231.
73. BASF, DE 2630086, 1976; DE 2713195, 1977; DE 2741511, 1977; DE 2837815, 1978; EP 3306, 1979.
74. Shell, WO 2000056695, 1999.
75. Shell, WO 2004103948, 2003.
76. Atlantic Richfield, US 4236023, 1980; US 4189599, 1980.
77. T. J. Kealy, R. E. Benson, *J. Org. Chem.* **26** (1961) 3126.
78. J. Tsuji, S. Hosaka, J. Kiji, T. Susuki, *Bull. Chem. Soc. Jpn.* **39** (1966) 141.
79. N. v. Kutepow, W. Himmele, H. Hohenschutz, *Chem. Ing. Tech.* **37** (1965) 383.
80. H. Hohenschutz, N. v. Kutepow, W. Himmele, *Hydrocarbon Process.* **45** (1966) no. 11, 141. T. Mizoroki, M. Nakayama, *Bull. Chem. Soc. Jpn.* **41** (1968) 1628.
81. BASF, DE-OS 2303271, 1973. Shell, DE 2400534, 1974.
82. J. F. Roth, J. H. Craddock, A. Hershman, F. E. Pauli, *CHEMTECH* 1971, 600.
83. *Chemical Week*, June 30/July 7 (2004) 40; *Chemical Week*, January 18 (2006) 17–20.
84. D. Forster, *J. Am. Chem. Soc.* **98** (1976) 846. K. M. Webber, B. C. Gates, W. Drenth, *J. Catal.* **47** (1977) 269. J. Hjortkjaer, V. W. Jensen, *Ind. Eng. Chem. Prod. Res. Dev.* **15** (1976) 46.
85. D. Forster, T. C. Singleton, *J. Mol. Catal.* **17** (1982) 299.
86. G. Henrici-Olivé, S. Olivé: *The Chemistry of the Catalyzed Hydrogenation of Carbon Monoxide*, Springer Verlag, Berlin-Heidelberg-New York 1984, p. 209.
87. K. Fujimoto, K. Omata, T. Shikada, H. Tominaga, *Ind. Eng. Chem. Prod. Res. Dev.* **22** (1983) 436.
88. J. H. Jones, *Platinum Metals Review* **44** (2000) 94 – 105.
89. N. Yoneda, S. Kusano, M. Yasui, P. Pujado, S. Wilcher, *Appl. Catal. A: General* **221** (2001) 253 – 265.
90. *Eur. Chem. News* **37** (1981) no. 1000, 4.
91. Ube Ind., US 3994960, 1976; US 4138587, 1979.
92. Ube Ind., US 4229591, 1980; JP 82–42 656, 1982.
93. BASF, DE 1221224, 1963. T. Rüll, *Bull. Soc. Chim. (France)* 1964, 2680.
94. M. Beller and A. Zapf, in: E. Negishi (ed.): *Handbook of Organopalladium Chemistry for Organic Synthesis*, Hoboken 2002, pp. 1209 – 1222.
95. J. T. Lee, P. J. Thomas, H. Alper, *J. Org. Chem.* **66** (2001) 54254; Y. D. Y. L. Getzler, V. Kundnani, E. B. Lobkovsky, G. W. Coates, *J. Am. Chem. Soc.* **126** (2004) 6842 – 6843; M. Almendinger, F. Molnar, M. Zintl, G. A. Luinstra, B. Rieger, *Chem. Eur. J.* **11** (2005) 1.
96. Shell, EP 577206, 2002.
97. F. Piacenti, M. Bianchi in I. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls*, **vol. 2**, Wiley, New York 1977, pp. 1 – 42.
98. Y. Takegami, C. Yokokawa, Y. Watanabe, H. Masada, *Bull. Chem. Soc. Jpn.* **37** (1964) 672.
99. M. Orchin, W. Rupilius, *Catal. Rev.* **6** (1972) 85.
100. S. Murai, N. Sonoda, *Angew. Chem.* **91** (1979) 896; *Angew. Chem. Int. Ed. Engl.* **18** (1979) 837.
101. H. Alper, *Pure Appl. Chem.* **52** (1980) 607.
102. A. V. Kramer, J. A. Osborn, *J. Am. Chem. Soc.* **96** (1974) 7832. T. T. Tsou, J. K. Kochi, *J. Am. Chem. Soc.* **101** (1979) 6319.
103. L. Cassar, M. Foá, *J. Organomet. Chem.* **51** (1973) 381.
104. M. Mori, in: E. Negishi (ed.): *Handbook of Organopalladium Chemistry for Organic Synthesis*, Hoboken 2002, p. 2663.
105. R. Skoda-Földes, L. Kollár, *Current Organic Chemistry* **6** (2002) 1097 – 1119.
106. T. Morimoto and K. Kakiuchi, *Angew. Chem. Int. Ed.* **43** (2004) 5580 – 5588.
107. W. W. Prichard, *J. Am. Chem. Soc.* **78** (1956) 6137.
108. National Distillers, US 3452090, 1969 (I. L. Mador, J. A. Scheben).
109. A. Zapf, M. Beller, *Topics in Catalysis* **19** (2002) 101 – 109.
110. M. Beller, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 145 – 156.
111. M. Beller, A. F. Indolese, *Chimia* **55** (2001) 684 – 687.

112. M. El-Chahawi, U. Prange, *Chem. Ztg.* **102** (1978) 1.
113. E. Yoshisato, S. Tsutsumi, *J. Org. Chem.* **33** (1968) 869.
114. Montecatini, IT 845993, 1969 (L. Cassar, M. Foá).
115. BASF, DE 2521610, 1975.
116. H. Alper, S. Antebi, J. B. Woell, *Angew. Chem.* **96** (1984) 710;
Angew. Chem. Int. Ed. Engl. **23** (1984) 729.
117. Dynamit Nobel, DE 2240398, 1972;
DE-AS 2553931, 1975.
DE 2359963, 1973.
DE 2524389, 1975.
118. Rhône-Poulenc, BE 837401, 1975;
DE 2600541, 1976;
EP 108698, 1982.
Ethyl Corp., WO 8402699, 1983;
US 4481368, 1983;
US 4481369, 1983.
119. M. P. Cooke, Jr., *J. Am. Chem. Soc.* **92** (1970) 6080.
120. J. P. Collman, N. W. Hoffman, *J. Am. Chem. Soc.* **95** (1973) 249.
121. G. P. Chiusoli, *Pure Appl. Chem.* **52** (1980) 635.
122. G. P. Chiusoli, L. Cassar in I. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls*, **vol. 2**, Wiley, New York 1977, p. 297.
123. D. Medema, R. van Helden, C. F. Kohll, *Inorg. Chim. Acta* **3** (1969) 225.
124. G. P. Chiusoli, L. Cassar, *Angew. Chem.* **79** (1967) 177.
National Distillers, US 3627827, 1971 (J. A. Scheben).
125. Montecatini, FR 1358900, 1964 (G. P. Chiusoli, S. Merzoni).
126. W. T. Dent, R. Long, G. Whitfield, *J. Chem. Soc.* 1964, 1588.
127. L. Cassar, M. Foá, *Chim. Ind. (Milan)* **51** (1969) 673.
128. C. W. Bird, *Chem. Rev.* **62** (1962) 283.
129. G. L. Rempel, W. K. Teo, B. R. James, D. V. Plackett in R. Lai, E. Ucciani (eds.): "Homogeneous Catalysis III," *Adv. Chem. Ser.* 1974, no. 132, 166.
J. J. Byerley, G. L. Rempel, N. Takebe, *Chem. Soc. D.* 1971, 1482.
130. D. Durand, C. Lassan, *Tetrahedron Lett.* 1969, no. 28, 2329.
131. Toyo Rayon Co., JP 69–10250, 1966 (J. Tsuji, N. Iwamoto).
Inst. Français du Pétrole, DE-OS 1902560, 1968 (C. Lassan, Y. Chauvin, G. Lefebvro).
132. Shell Oil, US 3407231, 1966 (K. Nozaki).
BASF, DE 948056, 1952.
133. H. Suzuki et al., *Chem. Lett.* 1975, 641.
134. E. W. Stern, M. L. Spector, *J. Org. Chem.* **31** (1966) 596.
135. Quimco, GB 1359032, 1974 (C. J. Hearsey);
US 3898259, 1975 (C. J. Hearsey).
136. Sun Ventures, US 3960914, 1976 (J. E. Lyons).
137. M. Dugal, D. Koch, G. Naberfeld, C. Six, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 1214 – 1125.
138. Olin Mathieson, US 3632827, 1972;
GB 1245480, 1971.
139. V. I. Manov-Yuvenskii, B. K. Nefedov, *Usp. Khim.* **50** (1981) 889;
Chem. Abstr. **95** (1981) 79321 y.
M. W. Ranney: "Isocyanate Manufacture," *Chem. Proc. Rev*, no. 63, Noyes Data Corp., Park Ridge 1972.
140. Mitsui Toatsu, DE 2555557, 1976;
JP-Kokai 53–15338, 1978.
141. Mitsui Toatsu, DE-OS 2614101, 1976.
142. Atlantic Richfield, GB 1402379, 1975;
US 3895054, 1975;
US 3956360, 1976.
143. BASF, US 2729651, 1956.
144. Mitsubishi Gas Chem., EP 55622, 1980.
145. Mitsubishi Gas Chem., EP 70180, 1981.
146. Halcon, BE 839322, 1975;
BE 886853, 1979;
GB 2067556, 1979;
US 4251458, 1979.
Eastman Kodak, BE 891211, 1980.
147. J. L. Ehrler, B. Juran, *Hydrocarbon Process.* **61** (1982) no. 2, 109.
148. W. Reppe, H. Kröper, N. v. Kutepow, H. J. Pistor, *Liebigs Ann. Chem.* **582** (1953) 72.
149. S. K. Bhattacharyya, D. Vir: *Advances in Catalysis IX*, Academic Press, New York 1957.
150. Mitsubishi Gas Chem., JP 58 046035, 1981.
151. H. Wakamatsu, J. Uda, N. Yamakami, *J. Chem. Soc., Chem. Commun.* (1971) 1540.
152. J. F. Knifton, in: B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis*, 2nd. ed., Wiley-VCH, Weinheim 2002, pp. 156 – 164.
153. A. J. van Wangelin, H. Neumann, D. Gördes, M. Beller, in: M. Beller, C. Bolm (eds.): *Transition Metals for Organic Synthesis*, vol. 1, 2nd. ed., Wiley-VCH, Weinheim 2004, pp. 133 – 148.

Carboxylic Acids, Aliphatic

Many industrially important aliphatic carboxylic acids and derivatives are treated in separate articles: Acetic Acid; Acrylic Acid and Derivatives; Adipic Acid; Chloroacetic Acids; Citric Acid; Crotonaldehyde and Crotonic Acid; Dicarboxylic Acids, Aliphatic; Ethylenediaminetetraacetic Acid and Related Chelating Agents; Fatty Acids; Formic Acid; Glyoxylic Acid; Hydroxycarboxylic Acids, Aliphatic and Oxocarboxylic Acids; Lactic Acid; Maleic and Fumaric Acid; Malonic Acid and Derivatives; Mercaptoacetic Acid and Derivatives; Methacrylic Acid and Derivatives; Nitrilotriacetic Acid; Oxalic Acid; Propionic Acid and Derivatives; Sorbic Acid; Tartaric Acid.

Wilhelm Riemenschneider, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Natural Sources 2
5. Production 5
5.1. Saturated Monocarboxylic Acids . . 5
5.1.1. Aldehyde Oxidation 5
5.1.2. Carboxylation of Olefins, Koch Process 6
5.1.3. Oxidation of Alkanes 6
5.1.4. Alkali Fusion of Alcohols 7
5.2. Unsaturated Monocarboxylic Acids 7
6. Environmental Protection 7
7. Quality Specifications and Analysis 7
8. Storage and Transportation 8
9. Uses 8
10. Specific Aliphatic Carboxylic Acids 9
10.1. Butyric Acids 9
10.2. Valeric Acids 9
10.3. Octanoic Acids 10
10.4. 2-Ethylhexanoic Acid 10
10.5. Nonanoic Acids 10
10.6. Isodecanoic Acid 11
10.7. Pivalic Acid 11
10.8. Versatic Acids and Neo Acids 11
10.9. Propiolic Acid 11
11. Trade Names, Economic Aspects . . 11
12. Toxicology and Occupational Health 12
13. Derivatives 13
13.1. Acyl Halides 13
13.2. Anhydrides 13
13.3. Lactams 13
13.4. Halogenated Carboxylic Acids . . . 14
14. References 14

1. Introduction

Aliphatic carboxylic acids have the general formula

where R is H or a straight-chain or branched-chain alkyl group. The first three members of this homologous series, formic, acetic, and propionic acid, are exceptionally important and are discussed in separate articles. Similarly, acids containing 12 or more carbon atoms are described separately under → Fatty Acids. The unsaturated acids acrylic, methacrylic, crotonic, and sorbic acid are also treated separately, as are chloroacetic and oxo- and hydroxycarboxylic acids.

The continuing scientific importance of the $C_4 - C_{11}$ carboxylic acids arises from their functions in the metabolism of plants and animals (see Chap. 4). The great commercial value of these acids, and particularly of their salts and esters, is based on their synthetic utility (see Chap. 9).

Numerous aliphatic carboxylic acids were first obtained from natural sources in the 19th century along with many other natural products. However, these natural sources, which usually yielded straight-chain acids containing an even number of carbon atoms, have since been replaced by large-scale synthetic operations (see Chaps. 5 and 10).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_235

2. Physical Properties

The properties of the important saturated and unsaturated aliphatic monocarboxylic acids are listed in Table 1 which also includes typical values of commercial mixtures.

Carboxylic acids are usually colorless. The *n*- and isoalkane carboxylic acids up to C_9 are liquids, as are the highly branched acids up to C_{13}, except for pivalic acid.

The volatility and characteristic odor of carboxylic acids decrease with increasing molecular mass. The odors of butyric and valeric acids are generally perceived as unpleasant; higher homologues and commercial mixtures have fainter smells. The lower saturated carboxylic acids are miscible with water; solubility decreases rapidly from C_5 up. Higher carboxylic acids are almost insoluble in water but miscible with most organic solvents. They are partially associated in polar solvents. In aqueous solution, carboxylic acids behave as weak acids; their strength changes only slightly with increasing molecular mass. The dissociation constants are between 1 and 2×10^{-5}.

3. Chemical Properties

The carboxyl group is the target of most reactions [1], [2] and is hardly affected by the carbon chain. However, the carboxyl group activates the α-position, which is also an important reaction site.

Alkali metal salts of carboxylic acids are readily obtained by neutralization. They undergo only limited hydrolysis in aqueous solutions, which are nearly neutral. Salts of the lower carboxylic acids are soluble in water. Alkali metal salts are poorly soluble in organic solvents; heavy metal salts, on the other hand, are soluble in organic solvents and can be prepared from the alkali metal salts. Zinc salts can be produced by treating the metal with the acids at high temperature. Esters are formed by an acid-catalyzed reaction between carboxylic acids and alcohols or olefins. Polyols such as glycerol or cellulose may react completely or partially. In the second case mixtures of esters are formed. Because of steric hindrance, esterification becomes more difficult with increasing chain length or branching. Addition of carboxylic acids to acetylene gives vinyl esters capable of polymerization.

In the presence of halogenating compounds, e.g., thionyl halides, phosphorus halides, or phosphorus oxyhalides, the highly reactive, unstable acyl halides, a valuable class of synthetic intermediates, are formed. With ammonia or amines the acyl halides give amides, with alcohols esters, and with carboxylic acids anhydrides. These derivatives can also be obtained directly from the carboxylic acids.

Ketenes are obtained from the lower carboxylic acids or anhydrides by pyrolysis; the higher homologues are obtained by dehydrochlorination of the acyl chlorides.

In the presence of such catalysts as aluminum oxide, thorium oxide, or boron phosphate, monocarboxylic acids give ketones with loss of carbon dioxide and water:

$$2\ \mathrm{RCOOH} \xrightarrow{\text{Cat.}} \mathrm{R\underset{\underset{O}{\|}}{C}R} + \mathrm{CO_2} + \mathrm{H_2O}$$

Reactions involving an unsubstituted carbon chain usually require more severe conditions than reactions of the carboxyl group. Halogenation often produces mixtures but sites of chain-branching are attacked preferentially. For example, isobutyric acid gives mostly the α-chloro derivative. Nitric acid oxidation gives dicarboxylic acids with some chain degradation. Permanganate oxidation produces oxalic and acetic acids under alkaline and acidic conditions, respectively.

4. Natural Sources

Low concentrations of aliphatic acids are widely distributed in nature. The free acids occur in plants, e.g., formic acid in stinging nettles and pine needles. Butyric acid is found in vegetable oils and in animal fluids, such as sweat, tissue fluids, and milk fat. Isobutyric acid is found in carob bean, isovaleric acid in valerian root. Other acids occur in fruit.

In animals, free carboxylic acids are important metabolites in the breakdown of carbohydrates, fats, and proteins, in the so-called tricarboxylic acid cycle (citric acid cycle). Nearly all naturally occurring acids have an even num-

Table 1. Physical properties of C_4-C_{13} monocarboxylic acids

Systematic name	CAS registry number	Trivial name	Molecular formula	M_r	*fp* or *mp*, °C	*bp*[a], °C	ϱ[b], g/cm³	n_D^{20}[c]	η[d]	S[e]	A[f]
Straight-chain acids											
n-Butanoic acid	[*107-92-6*]	*n*-butyric acid	$CH_3CH_2CH_2COOH$	88.11	− 5.5	164.0	0.9629^{15}	1.4003	1.6		
n-Pentanoic acid	[*109-52-4*]	*n*-valeric acid	$CH_3(CH_2)_3COOH$	102.1	−34.5	186.3	0.9437^{15}	1.4105	2.19	4	14
n-Hexanoic acid	[*142-62-1*]	caproic acid	$CH_3(CH_2)_4COOH$	116.2	− 3.9	205.3	0.9313^{15}	1.4188			
n-Heptanoic acid	[*111-14-8*]	enanthic acid	$CH_3(CH_2)_5COOH$	130.2	− 7.5	223.0	0.9222^{15}	1.4245			
n-Octanoic acid	[*124-07-2*]	caprylic acid	$CH_3(CH_2)_6COOH$	144.2	16.3	239.3	0.9088^{20}	1.4280			
n-Nonanoic acid	[*112-05-0*]	pelargonic acid	$CH_3(CH_2)_7COOH$	158.2	12.5	253–254	0.9068^{18}			0.026	
n-Decanoic acid	[*334-48-5*]	capric acid	$CH_3(CH_2)_8COOH$	172.3	31.3	268.4	0.8858^{40}	1.4285			
n-Undecanoic acid	[*112-37-8*]	undecylic acid	$CH_3(CH_2)_9COOH$	186.3	29.3	284.6	0.9905^{25}	1.4202^{70}			
n-Dodecanoic acid	[*143-07-7*]	lauric acid	$CH_3(CH_2)_{10}COOH$	200.4	44	225*	0.8690^{50}	1.4304^{50}			
Branched-chain acids											
2-Methylpropanoic acid	[*79-31-2*]	isobutyric acid	$(CH_3)_2CHCOOH$	88.11	−46.1	154.7	0.9529^{15}	1.3964	1.1	21	37
2-Methylbutanoic acid	[*116-53-0*]	α-methylbutyric acid	$CH_3CH_2CH(CH_3)COOH$	102.1	ca. −80	177	0.938^{20}		2.1	4.5	66
3-Methylbutanoic acid	[*503-74-2*]	isovaleric acid	$(CH_3)_2CHCH_2COOH$	102.1	−37	176.5	0.9308^{15}	1.4053	2.45	2.5	11
2,2-Dimethylpropanoic acid	[*75-98-9*]	pivalic acid	$(CH_3)_3COOH$	102.1	35	163.7	0.905^{50}			22	
2-Ethylbutanoic acid	[*88-09-5*]	α-ethylbutyric acid	$CH_3CH_2CH(C_2H_5)COOH$	116.2	−15.3	193	0.923^{20}		3.3	1.8	3.6
2-Methylpentanoic acid	[*97-61-0*]	isocaproic acid	$CH_3(CH_2)_2CH(CH_3)COOH$	116.2		196	0.921^{20}	1.414	2.9	1.5	2.8
2-Ethylhexanoic acid	[*149-57-5*]	α-ethylcaproic acid	$CH_3(CH_2)_3CH(C_2H_5)COOH$	144.2		220	0.908^{20}		7.8	0.2	1.2
2-*n*-Propylpentanoic acid	[*99-16-1*]	di-*n*-propylacetic acid	$(CH_3CH_2CH_2)_2CHCOOH$	144.2	ca. −60	229	0.906^{20}	1.425	8.1	0.2	1.2

Table 1. (Continued)

Systematic name	CAS registry number	Trivial name	Molecular formula	M_r	*fp* or *mp*, °C	*bp*[a], °C	ϱ[b], g/cm³	n_D^{20}[c]	η[d]	S[e]	A[f]
Commercial mixtures											
Isomeric C_8 acids	[*25103-52-0*]	isooctanoic acid	$C_7H_{15}COOH$	144.2	ca. −80	225–238	0.912–0.919	1.430	7.5	0.4	1.8
Isomeric C_9 acids	[*26896-18-4*]	isononanoic acid	$C_8H_{17}COOH$	158.2	ca. −70	232–245	0.895–0.902	1.430	11.2	0.3	1.3
Isomeric C_{13} acids	[*25448-24-2*]	isotridecanoic acid	$C_{12}H_{25}COOH$	214.4	ca. −46	280–315 (90%)	0.897–0.902	1.446	49.0	<0.1	0.3
Mixtures of C_9–C_{11} acids	[*71700-95-3*]	versatic acid 911				140–162**	0.92	1.447			
Unsaturated carboxylic acids											
2,2-Dimethylpropenoic acid	[*541-47-9*]	2,2-dimethylacrylic acid	$(CH_3)_2C{=}CHCOOH$	100.1	69	199	1.066^{24}			1.5	
Propynoic acid	[*471-25-0*]	propiolic acid	$CH{\equiv}CCOOH$	58.1	9	144 (decomp.)	1.139^{15}				

[a] at 1013 mbar (DIN 53 171)
[b] (DIN 51 757)
[c] (DIN 51 423, Bl. 2)
[d] Viscosity, mPa s (DIN 51 550)
[e] Solubility in water at 20 °C, wt%
[f] Absorption of water at 20 °C, wt%

* At 1.33 kPa.
** At 270 kPa.

ber of carbon atoms, which suggests a common synthetic pathway utilizing two-carbon units.

Esters occur more frequently than free acids and are important constituents of perfumes and essential oils. Esters of glycerol with various aliphatic acids are found in animals and plants. However, natural sources of carboxylic acids up to C_{11} have little or no commercial importance, because they frequently occur in low concentration only and their isolation is difficult.

5. Production [1], [2]

The four important commercial methods of production are (1) aldehyde oxidation, (2) carboxylation of olefins (Koch synthesis), (3) paraffin oxidation, and (4) alkali fusion of alcohols. The hydrolysis of natural oils such as castor and soybean oil for heptanoic acid and undecanoic acid is less important.

At present, biochemical processes are of limited importance. Butyric acid is produced from sugar or starch by *Bacillus butylicus,* from methanol by *Butyribacterium methylotrophicum* [3], or from CO or CO_2 by *Eubacterium limosum* [4]. However, commercial interest in biochemical methods may increase in the near future.

Similarly, catalytic air oxidation of alcohols has not yet attracted the interest of industry. Examples are the oxidation of 3-methyl-1-butanol to isovaleric acid (82 % yield), of 2-ethyl-1-hexanol to 2-ethylhexanoic acid (84 % yield), and of 1-decanol to *n*-decanoic acid (89 % yield) [5].

Although hydroformylation of methanol gives excellent yields of acetic acid, higher alcohols give much poorer yields. Thus, 1-heptanol gave only a 66 % yield of 1-octanoic acid with a ruthenium catalyst [6]. Homologation with palladium complexes and iodides gave similar results [7] (→ Carbonylation).

Carboxylic acids are also obtained by careful oxidation at low temperature and low concentration with one of the following agents: ozone, nitric acid, chlorine, chromic acid, permanganate, or periodate. However, these reagents are used especially when other functional groups are present in the molecule, and the reactions are largely limited to laboratory operations.

High-pressure hydrocarboxylation can be used for the conversion of propylene into isobutyric acid (main product) and *n*-butyric acid in the presence of a palladium catalyst [8].

5.1. Saturated Monocarboxylic Acids

5.1.1. Aldehyde Oxidation

The principal method for the commercial production of $C_4 - C_{13}$ carboxylic acids is catalytic oxidation of the corresponding aldehydes:

$$RCHO + 1/2\ O_2 \longrightarrow RCOOH$$

This highly exothermic reaction liberates between 250 and 300 kJ/mol.

The starting materials are produced catalytically from olefins with CO and H_2 by the oxo process, which was developed in 1938 by Ruhrchemie AG, Oberhausen, in Germany [9] (→ Oxo Synthesis). Virtually any olefin can be used in this process, giving access to a wide range of aldehydes oxidizable to the corresponding carboxylic acids. Other aldehydes can be produced via aldol condensation, increasing the range of accessible carboxylic acids. Because of the complex nature of the olefinic raw materials, the higher carboxylic acids obtained (C_8 and higher) are usually mixtures of branched-chain products.

The aldehydes are oxidized by air or oxygen with or without a catalyst. Liquid-phase oxidation without solvent is preferred, although vapor-phase oxidation is feasible. Side reactions are suppressed when the temperature is only slightly increased. Thorough dispersion of oxygen in the liquid medium is important, especially in the absence of catalysts [10].

Effective catalysts are the salts of metals that appear in more than one oxidation state, e.g., silver, cerium, cobalt, chromium, copper, iron, manganese, molybdenum, nickel, and vanadium. For liquid-phase oxidation, these metals are used as soaps or naphthenates, or in the form of complexes, such as $K_3\,[Fe(CN)_5 \cdot H_2O]$ [11]. In vapor-phase oxidation, oxides are predominantly used, supported on silicates, diatomaceous earth, silica gel, quartz, or alumina. Catalysts containing manganese or copper or both can be reclaimed by precipitation with oxalic acid [12].

Where side reactions or decomposition may occur during oxidation, as in the case of 2-ethylhexanal, selectivity is improved by using the alkali or alkaline earth salts of weak acids [13].

The following conditions are optimal for the oxidation of isobutyraldehyde to isobutyric acid: (1) temperature 50 °C, residence time ca. 3 h, and 0.02 g catalyst per mole of isobutyraldehyde; (2) temperature 60 °C, residence time 50 min, and 0.06 g catalyst per mole of isobutyraldehyde. The catalyst in each case is a mixture of cobalt, manganese, and copper acetates (1 : 1 : 1); oxygen is passed through in slight excess. Conversion is 100 %, selectivity 90 % [14]. A low pressure gives isobutyric acid with more desirable properties [15].

Peroxycarboxylic acids appear to be intermediates in both uncatalyzed and catalyzed liquid-phase oxidation:

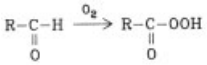

Aldehyde oxidation is usually carried out in stainless steel, but glass and enamelled vessels are also used (see Chap. 8).

5.1.2. Carboxylation of Olefins, Koch Process

Carboxylic acids are produced directly from olefins at high pressure as follows [16], [17] (→ Carbonylation):

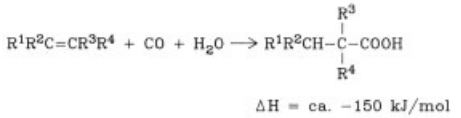

This eliminates the oxo-aldehyde step.

Reppe carbonylation, limited commercially to the conversion of ethylene into propionic acid, is carried out in the presence of metal carbonyls. The Koch carbonylation employs proton catalysis and is usually accompanied by double bond and structure isomerization to give branched, predominantly tertiary acids [18]. Mixtures are mainly produced; the proportion of tertiary acids depends on reaction conditions. For example, at 80 °C, a CO pressure of 10 MPa (100 bar), and prolonged residence time, all butene isomers are converted mainly into pivalic acid (trimethylacetic acid).

In addition to olefins (except for ethylene) and cycloalkenes, alcohols and alkyl halides may also serve as starting materials. Strong mineral acids, such as H_2SO_4, HF, or H_3PO_4, alone or in combination with BF_3, or SbF_5, serve as catalysts; anhydrous conditions are required to give high yields.

The commercial process employs two stages. In the first stage, using a stirred vessel cascade, CO is added to the olefin in the presence of the catalyst, at 20 – 80 °C and 5 – 10 MPa (50 – 100 bar); the preferred catalyst is $H_3PO_4 \cdot BF_3$. In the second stage water is added. The reaction mixture separates into a product phase containing the carboxylic acid and into a second phase that contains the catalyst which can then be recycled. The crude acids are washed with sulfuric acid, hydrogen carbonate, and citric acid, and are then distilled. Tertiary carboxylic acids are formed at high conversion rates with 85 – 95 % selectivity. Carboxylic acids derived from the dimerized olefins are produced as byproducts. The use of alcohol instead of water in the second stage produces esters.

The Koch process is employed commercially by Shell, Exxon (Enjay) [19], [20], Kuhlmann, and Du Pont. In addition to isobutene which gives pivalic acid, diisobutene and $C_6 - C_8$ and $C_8 - C_{10}$ cuts are also used as raw materials. On the basis of these olefins an economical and widespread production has been developed. The products are marketed as Versatic Acids (Shell), Neo Acids (Exxon), and CeKanoic Acids (Kuhlmann). The Koch synthesis is conducted in Hastelloy-C or lined pressure vessels (see Chap. 8).

5.1.3. Oxidation of Alkanes

The oxidation of paraffin waxes, polyethylenes, or polypropylenes with oxygen or air produces mixtures of straight- and branched-chain monocarboxylic acids ranging from formic acid to acids with the same number of carbon atoms as the starting material.

The reaction takes place above 100 °C in a melt or an aqueous suspension. The pressure is slightly above atmospheric; a catalyst may be

employed. The heat of reaction must be effectively removed.

This method is not selective; it produces large amounts of CO_2 and modest yields of products which are inseparable mixtures of acids, but can be used as such.

In this way, Rhône-Poulenc produces butyric acid in large amounts as a byproduct of naphtha oxidation.

A variant of alkane oxidation is the vapor-phase cracking of high-molecular alkanes in the presence of CO_2 at 500 – 700 °C with very short residence times [21], [22].

5.1.4. Alkali Fusion of Alcohols

This method is still used occasionally to produce carboxylic acids of intermediate chain length in good yields:

$$RCH_2OH + NaOH \longrightarrow RCOONa + 2\,H_2$$

Oxo alcohols from C_8 are possible starting products; their carbon skeleton is preserved. The process is carried out with NaOH or NaOH–KOH mixtures at a temperature of 250 – 300 °C (ca. 2 MPa) in stainless steel or copper pressure vessels [23]; [24].

5.2. Unsaturated Monocarboxylic Acids

Acrylic, methacrylic, crotonic, and sorbic acids are discussed in separate articles; other unsaturated carboxylic acids with less than 12 carbon atoms are not of commercial importance. However, the following two processes are employed on a small scale:

Dimethylacrylic acid is obtained in excellent yields from acetone and ketene:

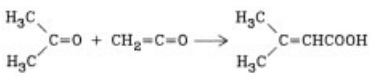

Homologous ketones and ketenes react similarly.

Propiolic acid, acetylenecarboxylic acid, H−C≡C−COOH, is produced by oxidizing propargyl alcohol at a lead anode in the presence of sulfuric acid [25]. The temperature is kept below 20 °C; the electrolytes are separated from each other by membranes. 2-Butyne-1,4-diol gives acetylenedicarboxylic acid, HOOC−C≡C−COOH.

6. Environmental Protection

The strong and unpleasant odors of many of these acids call for special protective measures. For example, the odor threshold of *n*-butyric acid is 8.8×10^{-13} g/L. These acids are readily removed from waste gas emanating from plants and storage tanks by an alkaline wash. Mixtures with aldehydes, e.g., from oxo syntheses, are removed by burning in a waste-gas flare or in a muffle furnace.

The degrees of degradation and BOD_5 values in a biological treatment plant have been determined for sodium carboxylates in wastewater using the Test for Biological Degradation from OECD-Guidelines 302 B (see Table 2).

Table 2. Biodegradation of sodium carboxylates [a]

Acid	Degradation, in 10 – 15 days, %	BOD_5 [b]
Isobutyric	>95	1160
Isovaleric	>95	1140
2,3-Dimethylbutyric	>96	<10
2,2-Dimethylpentanoic	>95	<10
Isoheptanoic	95	1350
Isooctanoic [c]	96	650
Isononanoic	95	125
Isodecanoic	79	690
Tridecanoic	48	40
Dimethylacrylic	96	590

[a] Straight-chain monocarboxylates are easily degraded.
[b] Biochemical oxygen demand: mg O_2 per liter of waste-water; sum over 5 days.
[c] Mixture.

7. Quality Specifications and Analysis

Manufacturers' specifications for the European market are given in Table 3. Specifications for the world market are similar.

Titration with aqueous alkali is satisfactory for many purposes; the preferred indicators are bromobenzene green, bromophenol blue, or phenolphthalein. The less soluble higher acids are titrated in aqueous methanol or ethanol or with methanolic KOH. Lauric acid is a suitable calibration standard.

Table 3. Manufacturers' specifications for commercial aliphatic carboxylic acids

Acid	Content, wt % min.	Boiling range, 2 – 97 % at 101.3 kPa, °C	Density at 20 °C, g/cm^3	Refractive index n_D at 20 °C	Hazen Color Index, max. APHA	Water content, wt %, max.	Freezing or melting temperature, °C
n-Butyric	99.5	161 – 164	0.956 – 0.959	1.398 – 1.399	15	0.1	− 5
n-Valeric	99	184 – 188	0.937 – 0.940	1.408 – 1.409	20	0.1	−32
Isobutyric	99.4	151 – 155	0.946 – 0.950	1.393 – 1.394	15	0.1	−46
2-Methylbutyric	99.3	176	0.935	1.406	5	0.05	<−60
Isovaleric	98.5	175 – 177	0.925 – 0.927	1.403 – 1.404	15	0.2	ca. −26
2-Methylpentanoic	99	196	0.921	1.414	ca. 10	0.1	<−60
2-Ethylbutyric	99	192 – 196	0.921 – 0.924	1.413 – 1.414	15	0.2	−13
2-Ethylhexanoic	99.2	226 – 230	0.905 – 0.907	1.425 – 1.426	15	0.1	−60
Di-*n*-propylacetic	99.5	225 – 229	0.903 – 0.906	1.424 – 1.426	20	0.1	<−60
Isooctanoic ∗	99	225 – 238	0.913 – 0.917	1.428 – 1.431	20	0.1	<−60
Isononanoic ∗∗	99	233 – 245	0.897 – 0.901	1.429 – 1.431	20	0.1	<−60
Isodecanoic ∗	99	252 – 269	0.903 – 0.906	1.436 – 1.438	35	0.1	ca. −60
Isotridecanoic ∗	99		0.897 – 0.902	1.446 – 1.447	50	0.2	−46

∗ Mixture of isomers.
∗∗ Ca. 90 % 3,5,5-trimethylhexanoic acid.

Titration against a glass–calomel electrode also gives excellent results; a graph is used to calculate the acid concentration [26].

Mixtures can be analyzed with great precision by gas chromatography; higher acids are first converted into methyl esters. The stationary phase in the GC column may contain Carbowax 20M impregnated with terephthalic acid, or polyglycol esters such as type Sp 1200/H_3PO_4 (10 : 4 %). Helium or nitrogen is used as carrier gas.

Paper or thin-layer chromatography [27] and spectrophotometric methods [28] are also used.

8. Storage and Transportation

Most aliphatic carboxylic acids are corrosive and flammable; flammability properties are given in Table 4. Carboxylic acids must be stored and transported in corrosion-resistant containers. Stainless steel and aluminum are used, but aluminum is subject to attack at certain concentrations. Polyethylene-lined metal drums are suitable for the lower acids. For acids with more than six carbon atoms, drums coated with acid-resistant paint are used. The higher-melting acids, e.g., lauric acid, are transported in paper bags or fiber drums lined with polyethylene.

Table 4. Flammability properties of aliphatic carboxylic acids [a]

Acid	Flash point, °C (Abel-Pensky)	Ignition temperature, °C
n-Butyric	69	425
Isobutyric	59	420
n-Valeric	86	360
2-Methylbutyric [b]	77	495
3-Methylbutyric	74	440
2-Methylpentanoic	95	390
2-Ethylbutyric	87	345
2-Ethylhexanoic	114	320
2-*n*-Propylpentanoic	113	350
Isooctanoic [c]	118	440
Isononanoic	120	405
Isodecanoic	128	355
Isotridecanoic	150	340
Dimethylacrylic [b]	93	465

[a] Ignition group G 2 according to VDE 0165, unless otherwise stated.
[b] Ignition group G 1.
[c] Mixture of isomers.

9. Uses

Carboxylic acids may undergo many reactions and have a wide variety of applications [1].

Esters, particularly those of branched-chain acids, are widely employed as solvents. The lower esters are good solvents for nitrocellulose and other cellulose derivatives. Esters have low polarity and low water solubility. Large quan-

tities are used in perfumes [29]. They are the principal constituents of many essential oils. Because of their flavor and aroma, the lower esters are also known as fruit ethers.

Carboxylic acids are incorporated into plastics. Much of the acids produced from oxo-process aldehydes is used for the manufacture of alkyd resins. Long-chain carboxylic acid groups increase solubility in organic solvents, compatibility with softeners, suppleness, and toughness.

Copper salts of branched-chain acids are important preservatives. Copper and zinc salts are used as fungicides and insecticides. Cobalt, lead, manganese, and iron salts are employed as drying agents for paints and printing inks. Because of their good miscibility with solvents and binding agents oxo-process acids give drying agents with a higher metal content than that afforded by naphthenic acids.

Carboxylic acids are also used as intermediates for textile chemicals, dyes, drugs, plastics, and agricultural chemicals. Amides and other derivatives serve as corrosion inhibitors, detergents, flotation aids, and oil additives.

10. Specific Aliphatic Carboxylic Acids

10.1. Butyric Acids

n-Butyric acid is produced commercially by the liquid-phase oxidation of *n*-butyraldehyde with oxygen. In the continuous process conversion is 97 – 98 % and selectivity over 95 %. The product, a colorless mobile liquid, is purified by vacuum distillation. Its penetrating and persistent odor is pungent when concentrated, and rancid or sweaty when diluted. The acid is miscible at room temperature with water and most organic solvents. It is steam-volatile and forms an azeotrope with water (18.5 wt % of butyric acid) boiling at 99.4 °C. Butyric acid esters are used to manufacture lacquers, plastics, and perfumes. Cellulose butyrates are more soluble in organic solvents than cellulose acetate and more miscible with resins and softeners.

Cellulose acetate butyrate is used for plastics with good physical stability and melting behavior. These plastics are suitable for molding. Being highly resistant to humidity, heat, and light, cellulose acetate butyrate gives excellent cable coatings and fishing tackle (→ Cellulose Esters).

Glycerol butyrates and *N*-arylbutyramides serve as softeners in cellulose-based plastics. The esters of monoalcohols have a pleasant, fruity aroma and are used in perfumes.

Isobutyric acid (2-methylpropanoic acid) is produced almost exclusively by the oxidation of isobutyraldehyde, a cheap byproduct of the hydroformylation of propene. This oxidation is less selective than that of *n*-butyraldehyde. Isobutyric acid is a colorless, clear liquid with a characteristically pungent and unpleasant odor. Above 26 °C, isobutyric acid and water are completely miscible. At 20 °C the aqueous phase contains 22 wt % of isobutyric acid and the organic phase contains 55 wt % of water. Miscibility with most organic solvents is also good. The water–isobutyric acid azeotrope boils at 99.3 °C and contains 21 wt % of the acid.

Isobutyric acid is valuable as a polar solvent [30], as starting material for diisopropyl ketone [31], and as a herbicide [32]. Its salts are used as textile chemicals, tanning compounds, stabilizers, catalysts, and preservatives. The esters are used in perfumes, as solvents, and as special oils for aircraft turbines.

10.2. Valeric Acids

n-Valeric acid, *n*-pentanoic acid, is a clear liquid with a characteristic, unpleasant odor. It is produced by the oxidation of *n*-valeraldehyde with oxygen. Its miscibility with water is limited, but it is fully miscible with common organic solvents.

Many esters of *n*-valeric acid with the lower alcohols are used as solvents and in aromas and perfumes [33]; esters with higher alcohols are used as softeners [34]. The ethyl ester is a lubricant. The sodium and calcium salts and some derivatives are used as rodenticides or fungicides.

2-Methylbutyric acid is produced by oxidation of 2-methylbutanal. Its properties and uses resemble those of *n*-valeric acid.

Valeric acid. A mixture of *n*-valeric acid (65 %) and 2-methylbutyric acid (35 %) is obtained directly by oxidation of the valeraldehyde mixture produced by oxo synthesis from

n-butene. In its properties and uses it resembles *n*-valeric acid.

Isovaleric acid, 3-methylbutyric acid, is a clear liquid with a highly unpleasant odor. It is produced by oxidation of isovaleraldehyde, which is obtained from isobutene. It is sparingly soluble in water (ca. 2.5 %), but readily soluble in organic solvents. It is used to make fungicides, rodenticides, and sedatives, narcotics, and other drugs. The esters are used as solvents and softeners and in perfumes [29]. The ammonium salts in particular are known for their rodenticidal properties.

10.3. Octanoic Acids

n-Octanoic acid, caprylic acid, is produced from oxo aldehydes. It is used in hydraulic fluids, machining oils, flotation agents [35], and as a wood preservative [36].

Isooctanoic acid is produced by the oxidation of the corresponding aldehyde. The commercial product is a mixture of methyl-branched isomers. It is a clear colorless liquid with a weak odor, practically insoluble in water, but miscible with most organic solvents. It is used in alkyd resins for high-quality lacquers.

Esters of isooctanoic acid, especially those of glycols and polyglycols, and amides are used as plasticizers for PVC polymers. Esters with glycols are widely employed in lubricating oils.

The lithium, magnesium, and aluminum salts are used in lubricants and the heavy metal salts in binding agents. Other derivatives are used as corrosion inhibitors, detergents, flotation aids, oil additives, hydraulic fluids, and preservatives. Uses are similar to those of 2-ethylhexanoic acid.

10.4. 2-Ethylhexanoic Acid

2-Ethylhexanoic acid is produced by oxidation of 2-ethyl-1-hexanol or 2-ethylhexanal; the latter is obtained in 95 % yield by hydrogenation of 2-ethyl-2-hexenal, which is itself formed by the aldol condensation of *n*-butyraldehyde (→ Aldehydes, Aliphatic and Araliphatic). In this way propene becomes the source of 2-ethylhexanoic acid. Its accessibility and utility have made it one of the acids with the highest production capacities worldwide. It is of particular importance in the production of alkyd resins used for baking enamels because of their exceptional stability even at temperatures over 200 °C and their good weathering and aging stability.

Esters of 2-ethylhexanoic acid, especially those of diglycols, triglycols, and polyethylene glycol, are used as plasticizers and stabilizers for PVC polymers. Peroxides of 2-ethylhexanoic acid are catalysts in the production of low-density polyethylene.

The lead, cobalt, manganese, and zinc salts are soluble in hydrocarbons. Such solutions serve as drying agents for lacquers and enamels. The iron, nickel, and cobalt salts are used as stabilizers for silicones. The copper salt is a fungicide for marine and other applications.

The acid and its derivatives are also used in the manufacture of lubricants, detergents, flotation aids, and corrosion inhibitors, as catalysts for polyurethane foaming, for solvent extraction [37], and for dye granulation [38]. Versatic Acid 10 (Shell) is used similarly.

10.5. Nonanoic Acids

Pelargonic acid , *n*-nonanoic acid, is manufactured via the oxo process and from oleic acid in 96 % yield by double-bond cleavage with ozone, nitric acid [39], or chlorine [40].

Like isononanoic acid, it is used in the manufacture of lubricating oils, peroxides, perfumes, and catalysts for alkyd resins.

Isononanoic acid, produced by the oxidation of isononyl aldehyde, is a mixture of isomers, containing more than 90 % 3,5,5-trimethylhexanoic acid. The starting aldehyde is obtained by the hydroformylation of diisobutene. The acid is also obtained directly from diisobutene by the Koch synthesis in 83 – 86 % sulfuric acid at 20 °C and a CO pressure of 7 MPa (70 bar); the main product in this case is 2,2,4,4-tetramethylpentanoic acid.

Isononanoic acid is a clear, colorless liquid with a weak odor. It is almost insoluble in water, but absorbs about 1.6 % water at 20 °C. It is miscible with most organic solvents. It is used for the same applications as isooctanoic acid; its higher molecular mass is sometimes advantageous.

Isononanoic acid is used in alkyd resins for high-quality lacquers and coatings. As such, it

provides excellent mechanical properties, surface hardness, color stability up to 200 – 230 °C, excellent gloss, and no yellowing.

Isononanoic acid esters are good stabilizers. Glycol and polyglycol esters are used in hydraulic fluids; they are also suited as lubricants under extreme conditions, e.g., in jet aircrafts, because of their good viscosity/temperature coefficient over a wide range. Metal salts find application in lubricating greases and as oxidative drying agents for lacquers. The copper salt is a fungicide. Derivatives are used as corrosion inhibitors, detergents [41], flotation aids, and catalysts.

Isononanoic acid can be replaced by Versatic Acid 10 (Shell) and Cardura E (Shell) or by CeKanoic 9 (Kuhlmann).

Table 5. Trade names and manufacturers of carboxylic acids

Chemical name	Trade name	Manufacturer
Valeric acid mixture	Neopentanoic Acid 5	Exxon
Valeric acid mixture	Versatic Acid 5	Shell
Octanoic-, nonanoic-, and decanoic acid mixtures with isomers	Neodecanoic Acid 10	Exxon
	Versatic Acid 10	Shell
	Cardura E	Shell
Isooctanoic acid mixture	CeKanoic 8	Kuhlmann
Isononanoic acid mixture	CeKanoic 9	Kuhlmann
Isodecanoic acid mixture	CeKanoic 10	Kuhlmann
Nonanoic and undecanoic acid mixture	Versatic Acid 911	Shell

10.6. Isodecanoic Acid

Isodecanoic acid is produced by oxidizing the mixture of aldehydes obtained from tripropylene in the oxo process. It is used in drying agents, PVC stabilizers, and alkyd resins; the copper salt is a wood preservative [42]. Isodecanoic acid may be replaced by versatic acid 10 or neodecanoic acid.

10.7. Pivalic Acid

Pivalic acid, trimethylacetic acid, is produced from isobutene in ca. 95 % yield by the Koch synthesis at 40 – 60 °C at a CO pressure of 5 MPa (50 bar) and with HF as catalyst [43–45]. Isobutanol or *tert*-butyl alcohol also serve as starting material [46].

Chain branching confers unusual properties on pivalic acid. The esters are very resistant to hydrolysis; some are heat-resistant lubricants. The poly(vinyl)esters form highly reflective lacquers with good weathering properties. The acid is used in cosmetics and for analgesic, anti-inflammatory, and antispasmodic preparations. It can replace *n*-valeric acid in various applications.

10.8. Versatic Acids and Neo Acids

These two acids, designated by trade mark (see Chap. 11), are prototypes of a group of highly branched carboxylic acids, mostly of 5 – 13 carbon atoms. They are produced on a large scale by the Koch synthesis; each product is a mixture of branched isomers and acids with varying numbers of carbon atoms. The name (e.g., Neononanoic Acid) or a numerical suffix (e.g., Versatic Acid 10) ordinarily refers to the number of carbon atoms in the constituent of highest molecular mass.

All these acids are liquids with a weak odor, miscible with organic solvents, but only slightly soluble in water. They can be vacuum distilled and are stable to about 250 °C. The exceptional thermal stability and the resistance to oxidation and hydrolysis are caused by the bulky tertiary alkyl groups of these acids or esters. They are therefore suitable as components of synthetic lubricants or hydraulic fluids. The vinyl esters are used as comonomers and plasticizers in vinyl chloride resins. The glycidyl esters are alkyd resin modifiers. The metal salts are drying agents.

10.9. Propiolic Acid

Propiolic acid, acetylenecarboxylic acid, is produced electrochemically from propargyl alcohol and used in agriculture and in the control of nematodes [47].

11. Trade Names, Economic Aspects

Most carboxylic acids are sold under their chemical names; exceptions are listed in Table 5. The 1980 consumption figures and 1983/84 production capacities for the most important products with a known capacity of more than 50 t/a are given in Table 6 [48].

Table 6. Economic aspects of aliphatic monocarboxylic acids

	Manufacturers [a]; production capacities (1983/1984), t/a			1980 Consumption, t		
	Western Europe	USA	Japan	Western Europe	USA	Japan
n-Butyric acid	Hoechst [b] Hüls BASF Rhône-Poulenc	Eastman; max. 50 000 Celanese; 5 000	Daicel; 250	10 000	11 800	90
Isobutyric acid	Hoechst BASF Hüls Rhône-Poulenc	Eastman; 500		<1 000		
n-Pentanoic acid (*n*-valeric acid and iso acids)	Hoechst BASF	Eastman UCC Exxon (as neo-pentanoic acid)	Daicel	1 000	6 800	50
n-Hexanoic acid (*n*-caproic acid)					550	
2-Ethylhexanoic acid	Hoechst BASF/Beroxo Berol-Kemi CdF Romania	Eastman; 25 000 UCC	Toyo Gosei Daisan Kasei Sanken Kako	10 000	14 100	1 900
Heptanoic acid (enanthic acid)	Kuhlmann Atochem			11 000	10 900	very little
Octanoic acid (caprylic acid)	Hoechst Kuhlmann	Procter & Gamble; 5000			1 000	
Nonanoic acid (pelargonic acid)	Unilever Akzo Henkel Hoechst	Givaudan Celanese Emery		15 000	9 800	very little
Decanoic acid (capric acid)	Shell	Procter & Gamble; 500 Ashland; 500			500	
Pivalic acid	BASF Shell	Exxon Shell			6 800	110
Versatic Acids	Shell	Shell		20 000		
Neo Acids		Exxon; 15 000		5 000	6 800	110
Lauric acid		Ashland; 5000 Capital City; 5000 Procter & Gamble; 500 Humko Sheffield; 500 Glymo Chem.; 500				
2-Chloropropanoic acid		Dow; max. 25 000			500	

[a] With total capacities over 50 t/a.
[b] Total carboxylic acid production capacity, 30 000 t/a [51]

12. Toxicology and Occupational Health

The saturated aliphatic carboxylic acids of 4 – 11 carbon atoms are not known as human poisons [49]; large doses are harmful when applied locally. The lower homologues are corrosive. With increasing chain length and decreasing water solubility, they become less irritating. They can be detected even in low concentrations by the pungent and sweaty odor of the lower homologues and the paraffinic odor of the higher ones. Inhalation of high concentrations can cause breathing difficulties and lung edema. Carboxylic acids are classified as dangerous substances; see also [50]. Chlorination increases acid strength and toxicity. Protective goggles and gloves must be worn when handling carboxylic acids to prevent contact with skin or clothing. If contact occurs, the affected

area should be liberally rinsed with water and the clothing changed. Chlorocarboxylic acids require special precautions.

The lethal doses LD_{50} for rats are given in Table 7. Neither MAK values nor TLVs have been established, indicating that harmful concentrations are not likely to arise in the workplace.

13. Derivatives

13.1. Acyl Halides

Only the acyl chlorides have commercial importance. They are obtained from carboxylic acids and their salts by halogenation with PCl_5, $POCl_3$, $SOCl_2$, SO_2Cl_2, or $COCl_2$. Alternatively, olefins are treated with hydrochloric acid and carbon monoxide at high pressure in the presence of palladium catalysts:

$$CH_2{=}CH_2 + CO + HCl \longrightarrow CH_3CH_2COCl$$

Vinylacetyl chloride [*1470-91-3*], $CH_2{=}CH{-}CH_2COCl$, can be produced from carbon monoxide and allyl chloride; acid bromides can be obtained by the same method.

Isobutyryl fluoride [*430-92-2*], $(CH_3)_2CHCOF$, can be prepared in 95 % yield from isopropyl fluoride [*420-26-8*], $(CH_3)_2CHF$, using CO and anhydrous HF at 0 – 80 °C and 1 – 5 MPa (10 – 50 bar) [52].

Acid chlorides have a pungent odor. They are thermally stable but easily hydrolyzed.

13.2. Anhydrides

Anhydrides of the lower carboxylic acids are mobile-to-oily liquids. Viscosity increases with molecular mass. The higher anhydrides are solids. From C_4 to C_7 the unpleasant odor resembles that of the carboxylic acids; the higher homologues possess little or no odor. In water, the lower members hydrolyze readily to the corresponding carboxylic acid; the higher members require prolonged heating.

Anhydrides can be obtained by the reaction of acyl chlorides with carboxylic acids or their salts or directly by oxidation of aldehydes with oxygen. An important synthesis of the higher homologues makes use of mixed anhydrides with acetic acid. These are formed from the acid by reaction with acetic anhydride or ketene. They react with a second mole of acid to give the symmetrical anhydride and one mole of acetic acid that is removed by a distillation; the removal may be accelerated by passing a stream of inert gas through the mixture. The process may be carried out in one step; yields are over 90 %:

$$RCOOH + CH_2{=}C{=}O \longrightarrow RCOOCOCH_3$$

$$RCOOCOCH_3 + RCOOH \longrightarrow (RCO)_2O + CH_3COOH$$

Another method is the addition of acetic anhydride to α-olefins using di-*tert*-butyl peroxide as the catalyst. For the reaction of olefins with CO and water to produce anhydrides, see → Carbonylation. The higher anhydrides are used mainly in the acylation of cellulose, e.g., propionate and butyrate (→ Cellulose Esters).

Table 7. Toxicity of aliphatic carboxylic acids

Acid	LD_{50} [a], mg/kg
n-Butyric	2940
n-Valeric	1120
Isobutyric	280
Isovaleric	2000
Pivalic	900
2-Methylpentanoic	2040
n-Hexanoic	3000
2-Ethylbutyric	2200
2-Ethylhexanoic	3000
Di-*n*-propylacetic [b]	670
n-Nonanoic	3200
Isooctanoic [c]	1.41
Isononanoic [d]	3135
n-Decanoic	129
Isodecanoic [c]	3.73
Dimethylacrylic	3650

[a] Rats, oral.
[b] 2-*n*-Propylpentanoic acid [51].
[c] Mixture of isomers.
[d] Ca. 90 % 3,5,5-trimethylhexanoic acid.

13.3. Lactams

Lactams are internal amides of aminocarboxylic acids; the lactams of 3-, 4-, and 5-aminocarboxylic acids are more stable than the acids themselves. They are produced by the reaction of ammonia with chlorocarboxylic acids, hydroxycarboxylic acids, or lactones. The principal industrial process is the Beckmann rearrangement of ketoximes, used to produce caprolactam and laurolactam (→ Caprolactam).

High yields of pyrrolidones and piperidones (five- and six-membered lactams, respectively) are obtained by the reaction of unsaturated amines with carbon monoxide in the presence of a cobalt carbonyl catalyst [53].

The most important reaction of lactams is polymerization to produce polyamide fibers (Nylon) and plastics.

The principal lactams of commercial importance are caprolactam (BASF, Du Pont, Toyo Rayon, Snia Viscosa, Union Carbide, Allied Chemical), pyrrolidone, and *N*-methylpyrrolidone (BASF); the last is used as a solvent and extractant, for example, for the removal of olefins and butadiene from crack gas mixtures (BASF process) or of aromatics from petroleum (Lurgi Arosolvan process). *N*-Vinylpyrrolidone polymers are used in cosmetics.

13.4. Halogenated Carboxylic Acids

Mono-, di-, and trichloroacetic acid are produced commercially on a large scale and are discussed in a separate article (→ Chloroacetic Acids). The higher chloro- and fluorocarboxylic acids are important as intermediates and agents for pest control.

Chlorocarboxylic acids are produced by liquid- and gas-phase chlorination at ca. 300 – 400 °C; mixtures usually result. Liquid-phase reaction is accelerated by UV light. Tertiary carbon atoms react preferentially, but chlorinating agents such as $POCl_3$, PCl_5, and SO_2Cl_2 favor α-substitution. In addition, chlorination of aldehydes in aqueous solution produces α-chloroisobutyric and other α-chloro acids [54]. Addition of HCl to α,β-unsaturated acids gives β-chloro acids. For example, a high yield of 3-chloropropionic acid [*107-94-8*] is produced from concentrated HCl and acrylic acid at 60 – 100 °C [55]. Halogen addition yields α,β-dihalopropionic acid.

The most important chlorinated aliphatic acid, 2,2-dichloropropionic acid [*75-99-0*] [56], can be made by chlorination of propionic acid, propionaldehyde, or propanol (→ Propionic Acid and Derivatives) [57]. It is a colorless liquid, *bp* 186 – 189 °C, d_4^{23} 1.389.

The sodium salt, CH_3CCl_2COONa, is known as the herbicide Dalapon [*127-20-8*] (Dow Chemical), a hygroscopic powder which decomposes at 166.5 °C and is soluble in water to the extent of 90 g/100 mL at 25 °C. The calcium and magnesium salts are also highly soluble in water. The oral toxicity (LD_{50}) of Dalapon in rats is 6600 – 9300 mg/kg. It is a soil herbicide and a selective, systemic leaf herbicide and is used against perennial grasses [58]. However, its importance in Europe is decreasing.

Less important products are 3-chlorobutanoic acid [*1951-12-8*] and 3-chloro-2,2-dimethylpropanoic acid [*13511-38-1*] (Kodak).

The Koch synthesis with unsaturated halohydrocarbons appears likely to become more important [59]. For example, chloropivalic acid [*13511-38-1*] can be obtained in ca. 70 % yield from methallyl chloride using a BF_3 catalyst.

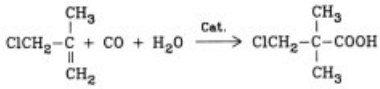

The chlorine of chloroacetic acid and other α-chloro acids may be replaced, e.g., by amino groups. β-Chloro acids can be converted into unsaturated acids by eliminating HCl; γ- and δ-chloro acids yield lactones.

Fluorocarboxylic Acids. Short-chain perfluorocarboxylic acids are used as intermediates, and the longer-chain ones, from about C_6, in surfactants. These compounds are discussed in a separate article (→ Fluorine Compounds, Organic).

14. References

General References

1. Chem. Technology Rev. no. 9, *Fatty Acids Synthesis and Applications,* Noyes Data Corp., Park Ridge, N.J., 1973.
2. L. F. Hatch, S. Matar: "From Hydrocarbons to Petrochemicals, Part 12, Chemicals from C_4's," *Hydrocarbon Process.* **57** (1978) no. 8, 153 – 165.

Specific References

3. Wisconsin Alumini Res., US 4 425 432, 1984 (J. G. Zeikus, L. H. Lynd).
4. University Patents Inc., US 4 377 638, 1983 (M. P. Bryant, B. Genthner).

5. S. Miyazaki, Y. Suhara, *J. Am. Oil. Chem. Soc.* **55** (1978) 536 – 538.
6. Texaco, DE 2 930 280, 1979 (J. F. Knifton); US 4 334 092, 1980 (J. F. Knifton).
7. Texaco, US 4 334 093, 1980 (J. F. Knifton); 4 334 094, 1980 (J. F. Knifton).
8. Standard Oil, EP-A 33 422, 1981 (F. A. Pesa, T. A. Haase); EP-A 52 419, 1982 (F. A. Pesa, T. A. Haase).
9. B. Cornils in J. Falbe (ed.): *New Syntheses with Carbon Monoxide,* Springer Verlag, Berlin-Heidelberg-New York 1980, pp. 1 – 181.
10. Ruhrchemie, DE 1 083 800; DE 1 154 454, 1959.
11. Ruhrchemie, DE-OS 2 931 154, 1981 (B. Cornils, W. Devin, J. Weber).
12. Celanese, US 4 246 185, 1981 (F. Wood, Jr.)
13. BASF, DE 950 007, 1951.
14. L. Nowakowski, M. Kaledkowka, *Zh. Prikl. Khim. (Leningrad)* **57** (1984) no. 11, 2569 – 2574; *Chem. Abstr.* **102** (1985) 61 767 b.
15. *Chem. Eng. (N.Y.)* **84** (1977) no. 12, 110 – 115.
16. H. Koch et al., *Brennst. Chem.* **36** (1956) 321; K. E. Möller, *Brennst. Chem.* **47** (1966) 15.
17. H. Bahrmann in J. Falbe (ed.): *New Syntheses with Carbon Monoxide.* Springer Verlag, Berlin-Heidelberg-New York 1980, pp. 372 – 408.
18. *Kirk-Othmer,* 3rd ed., **vol. 4,** pp. 863 – 871.
19. W. J. Ellis, C. Ronning, Jr., *Hydrocarbon Process.* **44** (1965) 139.
20. *Chem. Process. (Chicago)* **45** (1982) no. 11, 14.
21. Chevron, US 4 016 185, 1977 (J. Wilkes).
22. BP Chemicals, GB 1 563 202, 1980 (K. R. Dobson, M. Lynn).
23. Chemie Linz, GB 1 373 275, 1974.
24. Daisan Kasei, JP 75/96 517, 1975 (T. Nishida, H. Maruyama, S. Ushirode).
25. P. Jäger, H. Hannebaum, H. Nohe (BASF), *Chem. Ing. Tech.* **50** (1978) no. 10, 787 – 789.
26. L. Lykken, *Anal. Chem.* **22** (1950) 398 – 505.
27. *Methodicum Chimicum,* **vol. 1,** G. Thieme Verlag, Stuttgart 1973, p. 1010.
28. A. R. Baldwin, H. E. Longenecker, *Oil Soap (Chicago)* **22** (1945) 151 – 153.
29. D. L. J. Opdyke: "Monographs on Fragrance," *Food Cosmet. Toxicol.* **16** (1978) Suppl. 1, 839 – 841; **17** (1979) Suppl., 735 – 741, 841 – 843; **19** (1981) Suppl. 2, 237 – 245.
30. Hoechst, DE 1 352 351, 1964; DE 1 443 826, 1962.
31. H. J. Arpe, J. Falbe, *Brennst. Chem.* **48** (1967) 69.
32. P. Luberg, U. Härter, *Arzneim. Forsch.* **20** (1970) 1874.
33. Ch. Morel, *Soap Perfum. Cosmet.* **23** (1950) 732, 932.
34. Hoechst, DE 912 397, 1951.
35. US Depart. of the Interior, US 4 192 737, 1980 (P. Thompson, J. Huiatt, D. Seidel).
36. B. D. Hager, US 4 061 500, 1977 (B. D. Hager).
37. Rustenburg Refiners, GB 2 065 092, 1981 (R. Grant).
38. Ciba Geigy, DE 2 921 646, 1979 (J. A. Macphersan, J. A. Stirling, J. R. Robert, J. M. McCrae).
39. V. W. Advani, B. Y. Rao, *J. Oil. Technol. Assoc. India (Bombay)* **8** (1976) no. 2, 27 – 30; *Chem. Abstr.* **86** (1977) 74 857 g.
40. Nippon Oil, JP-Kokai 80/77675, 1980.
41. BASF, DE 2 754 359, 1979 (R. Baur, K. Oppenländer, D. Stickigt).
42. Australia Commonwealth Sc. and Res. Org. AU 77/9923, 78/35331 (D. F. McCarthy).
43. Shell, DE 1 924 151, 1969.
44. Ruhrchemie, DE 1 793 369, 1968.
45. S. Pawlenko, *Chem. Ing. Tech.* **40** (1968) 52.
46. Mitsubishi Gas, JP-Kokai 80/12870, 1980; 81/97245, 1981.
47. Kanesho Co., JP 83/52089, 1983.
48. T. Gibson, O. Kamatari, J. Wahlen: *Chemical Economics Handbook,* Marketing Res. Report "Oxo Chemicals", SRI-International, Dec. 1981.
49. D. W. Fasset in *Patty,* 2nd ed., **vol. II,** p. 1962; D. Guest, G. Katz, B. D. Astill, in *Patty,* 3rd ed., **vol. 2 C, 4901 – 4987,** 1982.
50. TSCA Toxic Substances Control test, Issue 1978.
51. *Chem. Mark. Rep.* 1977, March 7, 3.
52. Ashland Oil, US 4 451 670, 1984 (D. Grote, R. Grimm, R. Norton).
53. J. Falbe, H. J. Schulze-Steinen, F. Korte, *Chem. Ber.* **98** (1965) 1923.
54. Continental Oil, US 3 661 986, 1969; US 3 661 986, 1969.
55. Mitsui Toatsu Chem., JP 83/124 738, 1983.
56. R. Wegler, L. Eue: *Chemie des Pflanzenschutzes und der Schädlingsbekämpfungsmittel,* **vol. 2,** Springer Verlag, Berlin 1970, pp. 263 – 264
57. L. de Buyck et al., *Bull. Soc. Chim. Belg.* **89** (1980) no. 6, 441 – 458; *Chem. Abstr.* **94** (1981) 15 141 m.
58. Brit. Crop Prot. Council: *The Pesticide Manual,* 7th ed., London 1983, pp. 154 – 155.
59. K. E. Müller, *Brennst. Chem.* **47** (1966) 14.

Carboxylic Acids, Aromatic

Benzoic Acid, Cinnamic Acid, Hydroxycarboxylic Acids, Aromatic, Naphthalene Derivatives, Phthalic Acid and Derivatives, Salicylic Acid, Terephthalic Acid and Dimethyl Terephthalate, and Isophthalic Acid are separate keywords

FREIMUND RÖHRSCHEID, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1.	**Introduction**	1
2.	**Physical and Chemical Properties**	1
3.	**Production**	5
3.1.	**Carboxylation with Carbon Dioxide**	5
3.2.	**Dicarboxylic Acids Formed by Rearrangement**	5
3.3.	**Side-Chain Chlorination of Methylbenzenes**	5
3.4.	**Oxidation with Nitric Acid**	5
3.5.	**Oxidation with Molecular Oxygen in the Vapor Phase**	6
3.6.	**Oxidation with Molecular Oxygen in the Liquid Phase**	6
3.6.1.	Solvent-Free Oxidation	6
3.6.2.	Oxidation in Acetic Acid	7
4.	**Polycarboxylic Acids and Anhydrides**	7
4.1.	**Trimellitic Acid and Trimellitic Anhydride**	7
4.2.	**Pyromellitic Acid and Pyromellitic Dianhydride**	7
4.3.	**Other Benzenepolycarboxylic Acids and Anhydrides**	9
4.4.	**Naphthalic Acid and Naphthalic Anhydride**	9
4.5.	**3,4,9,10-Perylenetetracarboxylic Acid and Perylene Dianhydride**	9
4.6.	**2,6-Naphthalenedicarboxylic Acid**	10
4.7.	**1,4,5,8-Naphthalenetetracarboxylic Acid and 1,8 : 4,5-Naphthalenetetracarboxylic Dianhydride**	10
5.	**Economic Aspects**	10
6.	**Toxicology**	10
7.	**References**	11

1. Introduction

The production of virtually all industrially important aromatic carboxylic acids involves either the partial oxidation of aromatic alkyl groups or the partial oxidative degradation of the aromatic ring with atmospheric oxygen. The oxidation of the methyl groups of xylene exemplifies the former case.

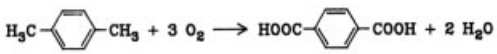

The partial oxidation of naphthalene is an example of the oxidation of the ring.

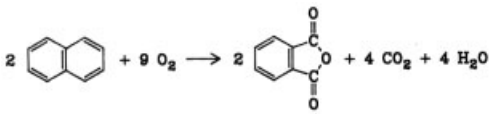

Such oxidizing agents as dichromate, permanganate, hypochlorite, chlorine, and nitric acid, which were used a few decades ago, have been replaced in almost all cases by atmospheric oxygen. The advantage of air oxidation is that only carbon dioxide and water are formed as the main byproducts. The progress made in modern oxidation techniques has gone hand in hand with the development of the synthesis of benzoic acid and of the two major products phthalic acid and terephthalic acid.

2. Physical and Chemical Properties

The physical properties of the most important aromatic carboxylic acids are given in Table 1. The monocarboxylic acids of benzene and naphthalene are, without exception, crystalline solids melting at temperatures above 100 °C. The melting points of the dicarboxylic and polycarboxylic acids are usually higher than those of the monocarboxylic acids. The acidity of benzoic acid (pK_a 4.17) and of the naphthoic acids (α: pK_a 3.7, β: pK_a 4.15) is a little greater than that of acetic acid (pK_a 4.74). The acidity is enhanced by electron-attracting substituents,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_249

especially those in the ortho position. Hence, in aromatic polycarboxylic acids, the acidity of the first carboxyl group is increased stepwise by the addition of further carboxyl groups (phthalic acid: pK_a 3.00, pyromellitic acid: pK_a 1.92).

In general, the chemistry of the carboxyl group of aromatic carboxylic acids does not differ from that of the carboxyl group of aliphatic carboxylic acids. Both form salts, esters, anhydrides, acid amides, and acid halides (→ Carboxylic Acids, Aliphatic). For this reason, only the most important reactions and some special features are discussed in detail here.

The esterification and amidation reactions of bifunctional carboxylic acids that yield polyesters and polyamides are of considerable industrial importance. The most important polyester, poly(ethylene terephthalate), is prepared by a two-stage reaction. Terephthalic acid or dimethyl terephthalate reacts with ethylene glycol at 200 °C in the molten state to give diglycol terephthalate:

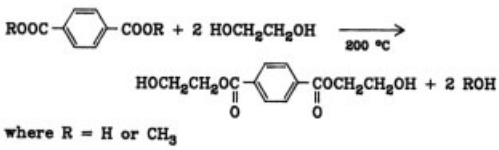

In the next stage, conducted under vacuum conditions at approximately 280 °C and in the presence of an antimony catalyst, polycondensation occurs which releases ethylene glycol.

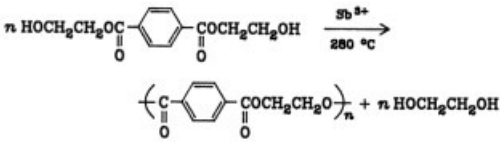

Among the polyamides, the condensation polymers of terephthalic acid and *p*-phenylenediamine (Kevlar) and of isophthalic acid and *m*-phenylenediamine (Nomex) are of practical importance because they can be converted into flameproof, temperature-resistant fibers.

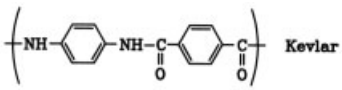

Aromatic polycarboxylic acids readily undergo cyclization to give anhydrides if a strainless five- or six-membered ring can be formed. The reaction is carried out by heating at low pressure or by boiling with acetic anhydride.

The five-membered cyclic anhydrides react readily with water, alcohol, or ammonia. They undergo only partial esterification with alcohols, but the product can be completely esterified at a higher temperature.

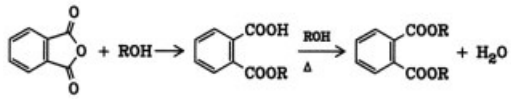

They react with ammonia or primary amines to form acid amides, which when heated form imides.

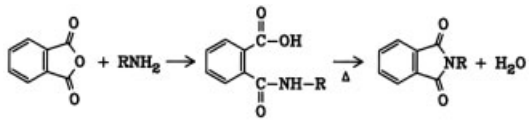

The acid amides or acid imides can be converted to the *o*-aminocarboxylic acids by a sodium hypochlorite – sodium hydroxide solution; e.g., phthalimide can be converted to anthranilic acid. This is an example of the Hofmann degradation.

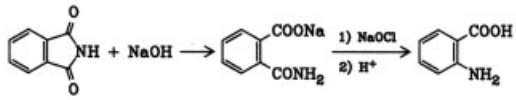

The six-membered cyclic anhydrides, which have the structure of naphthalic anhydride, form readily and are resistant to hydrolysis. They react easily with ammonia or primary amines to give imides.

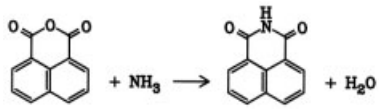

The reaction with *o*-phenylenediamine yields imidazoles. The alkaline salts of aromatic dicarboxylic acids can undergo carboxyl group rearrangements at high temperature to form thermodynamically more stable structures (Henkel reaction).

Table 1. Aromatic carboxylic acids

Acid	Common name	Empirical formula	M_r	Chemical structure	*mp*, °C	*bp*, °C	Solubility in 100 g water, g	
							at 25 °C	at 100 °C
Benzenecarboxylic acid [*65-85-0*]	benzoic acid	$C_7H_6O_2$	122.12		122.374	249	0.34	2.19 (75 °C)
2-Hydroxybenzenecarboxylic acid [*69-72-7*]	salicylic acid	$C_7H_6O_3$	138.12	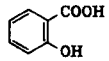	162	sublimes at 135 °C (650 Pa) sublimes at 150 °C (1930 Pa)	0.22	2.2 (80 °C)
trans-3-Phenyl-2-propenoic acid [*621-82-9*]	*trans*-cinnamic acid	$C_9H_8O_2$	148.16	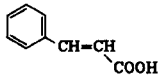	134	300 (101.3 kPa)	0.055	
1,2-Benzenedicarboxylic acid [*88-99-3*]	phthalic acid	$C_8H_6O_4$	166.13	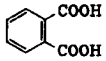	210 (forms anhydride)		0.7	19.0
1,3-Benzenedicarboxylic acid [*121-95-5*]	isophthalic acid	$C_8H_6O_4$	166.13	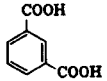	348 (sublimes)		0.012	0.25
1,4-Benzenedicarboxylic acid [*100-21-0*]	terephthalic acid	$C_8H_6O_4$	166.13	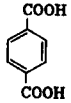	> 300 (sublimes)		0.0019	0.035
1,2,4-Benzenetricarboxylic acid [*528-44-9*]	trimellitic acid	$C_9H_6O_6$	210.14	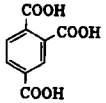	238 (forms anhydride)		2.1	60
1,3,5-Benzenetricarboxylic acid [*554-95-0*]	trimesic acid	$C_9H_6O_6$	210.14	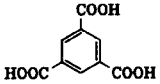	368		0.27	6.4
1,2,3-Benzenetricarboxylic acid [*569-51-7*]	hemimellitic acid	$C_9H_6O_6$	210.14	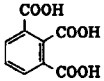	197 (forms anhydride)		very soluble	
1,2,4,5-Benzenetetracarboxylic acid [*89-05-4*]	pyromellitic acid	$C_{10}H_6O_8$	254.15	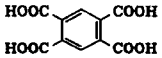	280 (forms dianhydride at 290 °C and 1.7 kPa)		1.5	> 30

Table 1. (Continued). Aromatic carboxylic acids

Acid	Common name	Empirical formula	M_r	Chemical structure	*mp*, °C	*bp*, °C	Solubility in 100 g water, g	
							at 25 °C	at 100 °C
1,2,3,4-Benzenetetracarboxylic acid [*476-73-3*]	mellophanic acid	$C_{10}H_6O_8$	254.15	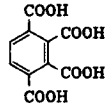	244–246 (decomp.)		soluble	very soluble
1,2,3,5-Benzenetetracarboxylic acid [*479-47-0*]	prehnitic acid	$C_{10}H_6O_8$	254.15	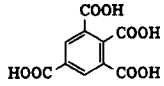	263–266		soluble	
Benzenepentacarboxylic acid [*1585-40-6*]		$C_{11}H_6O_{10}$	298.16	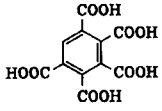	238–239		soluble	very soluble
Benzenehexacarboxylic acid [*517-60-2*]	mellitic acid	$C_{12}H_6O_{12}$	342.17	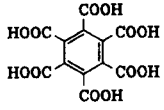	290 (in closed capillary) 285 decomp.		soluble	very soluble
1,8-Naphthalenedicarboxylic acid [*518-05-8*]	naphthalic acid	$C_{12}H_8O_4$	216.20	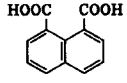	forms anhydride > 60 °C		sparingly soluble	
2,6-Naphthalenedicarboxylic acid [*1141-38-4*]		$C_{12}H_8O_4$	216.20	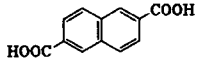	> 340		soluble in aqueous ethanol	
1,4,5,8-Naphthalenetetracarboxylic acid [*128-97-2*]		$C_{14}H_8O_8$	304.22	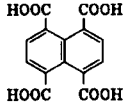	forms dianhydride > 80 °C		moderately soluble	
3,4,9,10-Perylenetetracarboxylic acid [*81-32-3*]		$C_{24}H_{12}O_8$	428.36	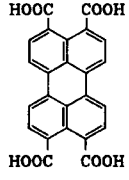	forms dianhydride		very soluble	

The copper-catalyzed oxidative decarboxylation of benzoic acid to form phenol is of considerable commercial importance. In the process developed by Dow Chemical, a mixture of air and steam at 220 – 250 °C and ca. 200 kPa is passed into molten benzoic acid. The resultant phenol leaves the reactor as a vapor. A selectivity of approximately 85 % is achieved using a copper benzoate catalyst.

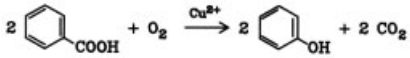

3. Production

3.1. Carboxylation with Carbon Dioxide

Specific carboxylic acids can be prepared by introducing a carboxyl group directly into the aromatic ring. This is usually carried out by reacting carbon dioxide with an aryl Grignard reagent or an aryl lithium compound. However, this reaction is of limited use.

In the Kolbe–Schmitt reaction, carbon dioxide at approximately 0.5 MPa (5 bar) reacts at 130 – 180 °C with the alkali salt of a phenol or naphthol to give a hydroxycarboxylic acid (→ Salicylic Acid, → Hydroxycarboxylic Acids, Aromatic).

3.2. Dicarboxylic Acids Formed by Rearrangement

Procedures involving carboxyl group rearrangements are known as the Henkel I and Henkel II procedures. They are named after the company that discovered them [1].

In the Henkel I procedure, the potassium salt of a dicarboxylic acid undergoes rearrangement to give its thermodynamically more stable form. For example, dipotassium phthalate is converted into dipotassium terephthalate by Zn–Cd-catalyzed rearrangement under carbon dioxide at 1.5 MPa (15 bar) and 430 – 440 °C [2].

The Henkel II procedure is based on the disproportionation of aromatic carboxylic acids. For example, potassium benzoate can be converted into dipotassium terephthalate and benzene with a 95 % selectivity by using the same reaction conditions as described above for the Henkel I procedure [3]. Although these methods have been applied by Teijin (Henkel I) and Mitsubishi Chemical (Henkel II) to produce terephthalic acid, they are used today only in the production of special carboxylic acids, e.g., 2,6-naphthalenedicarboxylic acid [*1141-38-4*].

3.3. Side-Chain Chlorination of Methylbenzenes

For the preparation of acid chlorides, the side-chain chlorination of methyl-substituted aromatic compounds is often highly advantageous [4]. Indeed, photochlorination of toluenes and xylenes at 80 – 140 °C in glass, lead, or nickel reactors forms trichloromethyl-substituted aromatic compounds. The reaction must be carried out with complete exclusion of any trace of iron. The product in turn reacts with one equivalent of water or carboxylic acid over an iron(III) chloride or zinc chloride catalyst to yield the acid chloride.

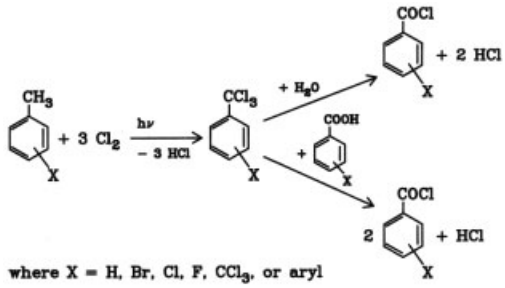

The acid chloride can be hydrolyzed with water to form the corresponding carboxylic acid.

3.4. Oxidation with Nitric Acid

The oxidation of methyl-substituted aromatic compounds to the corresponding carboxylic acids requires long heating (> 160 °C) with dilute nitric acid [5]. These conditions are chosen to reduce nitration. A supply of air at 3 – 4 MPa can lower the consumption of nitric acid. The reaction is catalyzed by ammonium vanadate. Even if the oxidation conditions are optimum, a marked nitration of the aromatic compound

occurs. This method was at one time very important, and is still used by AB Bofors Nobel Kemi to oxidize nitrotoluenes, chloronitrotoluenes, and chlorotoluenes.

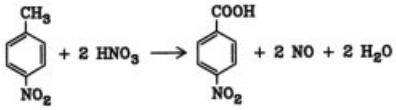

3.5. Oxidation with Molecular Oxygen in the Vapor Phase

The vapor-phase air oxidation of alkyl-substituted aromatic compounds or of multinuclear aromatic compounds to form aromatic carboxylic acids (e.g., *o*-xylene or naphthalene to phthalic anhydride) proceeds at 350 – 500 °C over a vanadium pentoxide catalyst. The catalyst is usually modified with titanium dioxide or sodium sulfate and is supported on a silica gel or silicon carbide carrier [6], [7]. The oxidation can take place in a fixed-bed reactor or in a fluidized-bed reactor. In the former case, the catalyst is firmly held in place in a multitubular reactor with 10000 or more tubes of about 1.5 cm inner diameter. Large amounts of heat are generated by the reaction, necessitating the use of a eutectic molten salt for cooling the tubes. The hydrocarbon reactant and air are passed into the reactor as a gaseous mixture. This procedure is used by Chemische Fabrik von Heyden, Veba Chemie, Koppers, BASF, Monsanto, and Rhône-Poulenc, and is suitable for the oxidation of *o*-xylene (at 380 – 400 °C, molar yield 78 %), naphthalene (at 400 °C, molar yield approximately 80 %), acenaphthene, and durene.

In 1944, the oxidation procedure using a continuous fluidized-bed reactor was introduced on a large industrial scale in the United States. It was used by Badger–Sherwin-Williams for the production of phthalic anhydride from naphthalene (see chapter 1). In this process liquid naphthalene is injected into the fluidized solid catalyst at 350 – 380 °C, using preheated air to maintain the flow. The main advantages of this technique are better control of the reaction heat and minimization of flammability problems. The yields obtained by it are lower than those in the fixed-bed process, especially when the starting material is *o*-xylene. The Badger Company, United Coke–Foster Wheeler, and ICI use the fluidized-bed reactor for the oxidation of naphthalene.

The gases leaving the reactor are passed through switch condensers where the product is almost quantitatively deposited on the cooled tubes. Hot oil is in turn passed through the tubes, and the product is melted off. The exhaust gas must be scrubbed before being discharged into the atmosphere. The vapor-phase oxidation gives molar yields of up to 80 % and is suitable for the production of maleic anhydride, benzoic acid, phthalic anhydride, naphthalic anhydride, and pyromellitic anhydride.

3.6. Oxidation with Molecular Oxygen in the Liquid Phase

Many carboxylic acids are destroyed by the high temperatures required in the gas-phase oxidation technique or they are just not sufficiently volatile. These acids can be effectively produced by catalytic oxidation in the liquid phase. The starting materials are alkyl-substituted aromatic compounds, and the reaction is carried out either in a column or in a pressure reactor with effective mixing. Although acetic acid is generally the most useful solvent, oxidation can also occur in the absence of solvent. Cobalt or manganese salts are commonly used as catalysts.

3.6.1. Solvent-Free Oxidation

A process developed by Snia Viscosa involves the oxidation of toluene to benzoic acid at 165 °C and 1 MPa (10 bar) and uses cobalt carboxylates as catalysts. A 20 – 30 % conversion and a selectivity of 90 % are achieved [8–10].

The Dynamit Nobel–Hercules procedure is used to make dimethyl terephthalate (DMT). This joint oxidation of *p*-xylene and the methyl ester of toluic acid at 150 – 170 °C and 400 – 800 kPa (4 – 8 bar) yields toluic acid and the monomethyl ester of terephthalic acid. This mixture is converted by pressure esterification with methanol at 200 – 250 °C to toluic acid methyl ester and DMT. The latter is isolated by distillation and recrystallized from methanol. The toluic acid methyl ester is then returned to the oxidation reactor along with fresh xylene [11].

3.6.2. Oxidation in Acetic Acid

Oxidation without addition of a solvent is in many ways disadvantageous because the resulting carboxylic acids have high melting points, and, therefore, only limited conversion rates can be attained. However, 100 % conversion is possible when working in acetic acid. The resultant carboxylic acids usually precipitate as crystals, e.g., terephthalic acid and *p*-nitrobenzoic acid.

The main catalysts are soluble cobalt and manganese carboxylates. However, only in a few cases do these metal ions alone give a satisfactory degree of conversion. To facilitate oxidation of chlorotoluene, *p*-nitrotoluene, or toluic acid to the corresponding carboxylic acids, the following additional requirements must be met.

In the Amoco procedure, the addition of bromide to cobalt acetate and manganese acetate in acetic acid produces a highly reactive catalyst. *p*-Xylene is oxidized at 180 – 210 °C and 2 – 3 MPa (20 – 30 bar) using 0.1 wt % cobalt to a very pure terephthalic acid with a yield of over 95 % [12–14]. This catalyst is also suitable for the preparation of tricarboxylic and tetracarboxylic acids of benzene [15]. The reactors must be coated with a resistant material (enamel, titanium, nickel-chromium-molybdenum alloys such as Hastelloy-C4) because bromide ion is highly corrosive.

Carboxylic acids can be efficiently obtained by the *cooxidation* of alkyl-substituted aromatic compounds with reagents (acetaldehyde, butane, 2-butanone) which form peroxy radicals during the course of the reaction. Two procedures for the synthesis of terephthalic acid are based on this principle. The liquid-phase oxidation procedure operated by Eastman Kodak uses acetaldehyde [16], [17]. In the Toray process, a modification of the Kodak method, paraldehyde is employed instead of acetaldehyde. Butane and 2-butanone have been used as cooxidizing agents by Mobil Oil and Olin Mathieson [18]. These methods can also be efficiently applied to the oxidation of other alkyl-substituted aromatic compounds.

4. Polycarboxylic Acids and Anhydrides

4.1. Trimellitic Acid and Trimellitic Anhydride

Trimellitic acid crystallizes from water as needles (Table 1). It dissolves to the extent of 1.0 g in 100 g of acetic acid at 25 °C [19]. Trimellitic anhydride reacts with acetic acid and water and is hygroscopic (Table 2).

Trimellitic anhydride alone is of industrial importance. Its preparation is best achieved by the liquid-phase catalytic oxidation of 1,2,4-trimethylbenzene (pseudocumene) with air (Amoco method) [15], [20], [21]. An alternative process involves oxidation with 7 % nitric acid at 170 – 190 °C and 2 MPa, but it is inefficient compared with the air oxidation method [22].

The largest amount of trimellitic anhydride is used to make plasticizers for poly(vinyl chloride). These plasticizers have a lower volatility than phthalate plasticizers. Polyimides and polyesters of trimellitic anhydride have a high thermal resistance and are used in the production of wire enamels, coatings, and baking varnishes. The anhydride is an important curing agent in the production of epoxy resins. Polyester resins of trimellitic anhydride are widely used as paint resins for water-based and solvent-based coatings.

4.2. Pyromellitic Acid and Pyromellitic Dianhydride

Pyromellitic acid crystallizes from water as a dihydrate, triclinic plates, *mp* 242 °C (Table 1). Pyromellitic dianhydride is very hygroscopic (Table 2).

Pyromellitic acid was originally made by Du Pont by oxidizing 1,2,4,5-tetramethylbenzene (durene) with nitric acid. Today, the only method employed industrially is the vapor-phase oxidation of durene, trimethylisopropylbenzene, or dimethyldiisopropylbenzene to the dianhydride [23]. The oxidation conditions are similar to those used for *o*-xylene. The yield is approximately 60 %. Most of the applications require very pure pyromellitic dianhydride. Thus, the

Table 2. Anhydrides of aromatic carboxylic acids

Compound	Common name	Empirical formula	M_r	Chemical structure	*mp*, °C	*bp*, °C
1,3-Isobenzofurandione [*85-44-9*]	phthalic anhydride	$C_8H_4O_3$	148.12		131.2	215 (440 Pa) 284.5 (101.3 kPa)
1,3-Dihydro-1,3-dioxo-5-isobenzofuran-carboxylic acid [*552-30-7*]	trimellitic anhydride	$C_9H_4O_5$	192.13	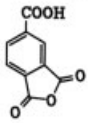	168	390
1,3-Dihydro-1,3-dioxo-4-isobenzofuran-carboxylic acid [*3786-39-8*]	hemimellitic anhydride	$C_9H_4O_5$	192.13	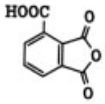	196	
Benzo[1,2-*c*:3,4-*c'*]-difuran-1,3,4,6-tetrone [*4435-60-3*]	mellophanic dianhydride	$C_{10}H_2O_6$	218.12	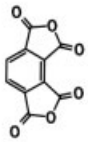	198 sublimes	
Benzo[1,2-*c*:4,5-*c'*]difuran-1,3,5,7-tetrone [*89-32-7*]	pyromellitic dianhydride	$C_{10}H_2O_6$	218.12	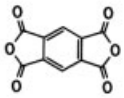	285	390
Benzo[1,2-*c*:3,4-*c'*:5,6-*c''*]-trifuran-1,3,4,6,7,9-hexone [*4253-24-1*]	mellitic trianhydride	$C_{12}O_9$	288.13	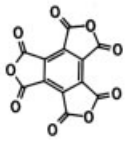	310 (de-comp.)	sublimes at 200 °C (400–500 Pa)
Naphtho[1,8-*c,d*]pyran-1,3-dione [*81-84-5*]	naphthalic anhydride	$C_{12}H_6O_3$	198.18		276	422; sublimes at 215 °C (440 Pa)
[2]Benzopyrano[6,5,4-*def*][2]benzopyran-1,3,6,8-tetrone [*81-30-1*]	1,8:4,5-naphthalenetetra-carboxylic dianhydride	$C_{14}H_4O_6$	268.18		> 300	sublimes at 320 °C (400 Pa)
Perylo[3,4-*c,d*:9,10-*c',d'*]dipyran-1,3,8,10-tetrone [*128-69-8*]	perylene dianhydride	$C_{24}H_8O_6$	392.33		> 300	

raw substance must be freed from all impurities. Because of the high boiling point, purification by distillation is expensive and involves a large loss of material. Therefore, methods have been developed for the purification of pyromellitic dianhydride. Inert gases are passed through the raw product at a temperature above the dew point of the impurities (175 – 275 °C) [24], [25]. Complexes of dioxane or xylene are formed, isolated, and decomposed [26], and the impurities are dissolved and removed [27–29]. The only commercial producer is Chemische Werke Hüls AG with a capacity of 500 t/a.

The major part of the pyromellitic dianhydride produced is used in the production of polyimides that have excellent electrical and physical properties and a high heat resistance. Polyimide resins can be used for films, fibers, molding compounds, varnishes, and wire coatings. They are resistant to a constant temperature of 300 °C and a transient temperature of 500 – 600 °C. The trade names of these Du Pont products are Kapton (condensation product of pyromellitic dianhydride and 4,4′-diaminophenyl ether), Pyre-M 11, Pyralin, and Vespel.

4.3. Other Benzenepolycarboxylic Acids and Anhydrides

1,3,5-Benzenetricarboxylic acid, trimesic acid, is formed by the oxidation of 1,3,5-trimethylbenzene (mesitylene). This is carried out on an experimental scale (4.5 t/a) by the American Bio-Synthetic Corp. The applications are the same as those of trimellitic anhydride.

1,2,3-Benzenetricarboxylic acid (hemimellitic acid), hemimellitic anhydride, 1,2,3,4-benzenetetracarboxylic acid (mellophanic acid), mellophanic dianhydride, 1,2,3,5-benzenetetracarboxylic acid (prehnitic acid), benzenepentacarboxylic acid, benzenehexacarboxylic acid (mellitic acid) and mellitic trianhydride can all be obtained either by the air oxidation of the corresponding polymethylbenzenes using the Amoco process or by oxidation with nitric acid. The preparation of mellitic acid can also be achieved by the nitric acid oxidation of certain types of brown coal. None of these compounds is of commercial importance.

4.4. Naphthalic Acid and Naphthalic Anhydride

Naphthalic acid crystallizes as needles (Table 1). Naphthalic anhydride crystallizes from ethanol as needles (Table 2).

Naphthalic acid is prepared by the gas-phase oxidation of acenaphthene with air at 450 – 550 °C over a vanadium pentoxide catalyst. Yields of approximately 85 % are obtained. The mass-to-mass ratio of air to acenaphthene should be 80 – 100 to 1 [30]. The oxidation of acenaphthene can also be achieved in the liquid phase in acetic acid at 100 – 150 °C in the presence of cobalt, manganese, and bromide ion [31].

Naphthalic anhydride is primarily converted into imides or imide derivatives. Naphthalimide [*81-83-4*] (*mp* 300 °C) is formed by heating the anhydride in concentrated aqueous ammonia. Naphthalic anhydride is the starting material in the production of imidazole dyes. Imide derivatives are employed as optical brighteners for synthetic resins and fibers. Most naphthalimide is converted into derivatives of 3,4,9,10-perylenetetracarboxylic acid.

4.5. 3,4,9,10-Perylenetetracarboxylic Acid and Perylene Dianhydride

When heated, deep red 3,4,9,10-perylenetetracarboxylic acid (Table 1) yields red perylene dianhydride (Table 2).

When naphthalimide is heated to 220 °C with caustic potash [*1310-58-3*] and sodium acetate [*127-09-3*], the potassium salt of the leuco form of perylenetetracarboxylic diimide is formed. This product is subsequently oxidized in aqueous solution to form perylene diimide [*81-33-4*], $C_{24}H_{10}N_2O_4$, M_r 390.36, Perylene Bordeaux. The diimide in turn forms perylene dianhydride on hydrolysis with concentrated sulfuric acid at 215 °C [32], [33].

The diimide and the dianhydride are starting materials for the synthesis of aliphatic- and aromatic-substituted diimides. The latter constitute an important class of red dyes and pigments used in the production of color-fast plastics and coatings. Approximately 100 t/a of the dyes and 1000 t/a of the pigments are manufac-

tured worldwide (→ Naphthalimide Dyes and Pigments).

4.6. 2,6-Naphthalenedicarboxylic Acid

2,6-Naphthalenedicarboxylic acid, is prepared by using the Henkel procedure starting with the potassium salt of 1,8-naphthalenedicarboxylic acid or 2-naphthoic acid [34], [35].

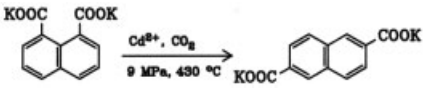

The salt hydrolyzes readily to form the corresponding acid. Polycondensation of 2,6-naphthalenedicarboxylic acid with diols yields highly ductile and heat-resistant polyesters, which are used in the manufacture of films and fibers [36]. The product is suitable for magnetic tapes and electrical insulation materials.

4.7. 1,4,5,8-Naphthalenetetracarboxylic Acid and 1,8 : 4,5-Naphthalenetetracarboxylic Dianhydride

1,4,5,8-Naphthalenetetracarboxylic acid readily loses water when heated above 80 °C to form 1,8 : 4,5-naphthalenetetracarboxylic dianhydride.

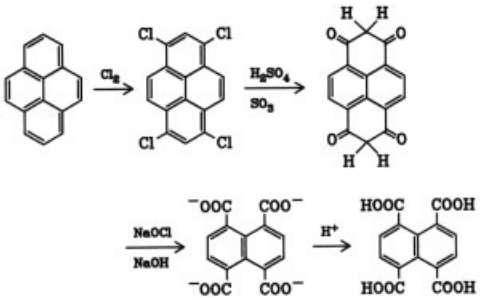

The tetracarboxylic acid is obtained by the oxidative degradation of pyrene from coal tar. The oxidation proceeds in a series of steps [37], [38].

A new, technically feasible, procedure is based on the reaction of acenaphthene with diketene in liquid HF to yield 1,2-dihydro-7-methyl-cyclopenta [*cd*] phenalen-5-one [*42937-13-3*], $C_{16}H_{12}O$ [39].

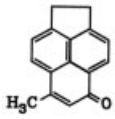

This compound in turn can be converted to the dianhydride by air oxidation in acetic acid in the presence of ammonium vanadate and nitric acid as catalysts [40].

Most of the naphthalenetetracarboxylic dianhydride is condensed with two equivalents of *o*-phenylenediamine to give a mixture of *cis*- and *trans*-1,4,5,8-naphthoylene-bis(benzimidazole) (Indanthrene Scarlet GG), which when separated into the isomers yields Perinone Orange (trans), Indanthrene Brilliant Orange GR [*4424-06-0*], and Perinone Red (cis), Indanthrene Bordeaux RR [*4216-02-8*]. Both are used as pigments and vat dyes.

5. Economic Aspects

The benzene- and naphthalenepolycarboxylic acids are speciality items and are produced in comparatively small amounts. Amoco is the sole manufacturer of trimellitic anhydride in the United States; it has a production capacity of 25 kt/a. In 1983, the United States used 14 kt trimellitic acid ester as a special softener [41].

Chemische Werke Hüls is the sole supplier of pyromellitic anhydride. The production capacity is 500 t/a. The 1200 – 2000 t naphthalic anhydride produced in the EEC is all made by Chemische Fabrik Weyl (Germany). 1,4,5,8-Naphthalenetetracarboxylic acid is produced by Hoechst and by Toms River Chemical. The production capacity is less than 1 kt/a.

6. Toxicology

As a result of their acidity and their ability to dehydrate, the carboxylic acids and their anhydrides may cause severe irritation to the pulmonary tract, skin, eyes, and mucous membranes. Some compounds have additional toxic properties. Exposure to trimellitic acid may result in noncardiac pulmonary edema and immunological sensitization [42]. Pyromellitic

acid has an intraperitoneal LD_{50} in mice of 300 mg/kg. Pyromellitic acid inhibits the enzyme aconitase [43]. Handling precautions include effective ventilation, protective clothing, and goggles if exposure to dust is expected.

7. References

1. B. Raecke, *Angew. Chem.* **70** (1958) 1 – 5.
2. Teikoku, GB 975113, 1964. Teijin, JP 4220525, 1967; JP 4420091, 1969.
3. Mitsubishi Chemical, JP 4719537, 1972; JP 4724549, 1972. Phillips Petroleum, DE 2230173, 1972 (Yu-Lin Wu, P. S. Hudson) (= US 3781341, 1973).
4. *Winnacker-Küchler,* 4th ed., **vol. 6,** pp. 165 – 169.
5. E. Bengtsson, B. Holm, *Chem. Ztg.* **104** (1980) 349 – 351. *Houben-Weyl,* **4/1 a,** 660 – 670.
6. H. Suter: *Phthalsäureanhydrid und seine Verwendung,* Steinkopff Verlag, Darmstadt 1972.
7. M. S. Wainwright, N. R. Foster: "Catalysts, Kinetics, and Reactor Design in Phthalic Anhydride Synthesis," *Catal. Rev. Sci. Eng.* **19** (1979) no. 2, 211 – 292.
8. Snia Viscosa, IT 657641, 1963 (L. Notarbartolo, G. Messina, V. Muench, R. Mattone); IT 803424, 1968 (G. Sioli, A. Salatini, S. Sanchioni).
9. *Chem. Age (London)* **117** (1978) Sept. 29, 15.
10. *Hydrocarbon Process.* **56** (1977) no. 9, 134; **58** (1979) no. 9, 145.
11. E. Katzschmannn, *Chem. Ing. Tech.* **38** (1966) 1 – 10. *Hydrocarbon Process.* **64** (1985) no. 11, 129
12. R. Landau, A. Saffer, *Chem. Eng. Prog.* **64** (1968) 20 – 26.
13. Y. Ichikawa, Y. Takeuchi, *Hydrocarbon Process.* **51** (1972) no. 9, 103 – 108.
14. Toray Industries, DE-OS 2749638, 1976 (H. Morimoto, Y. Tohda, H. Torigata, K. Nakaoka).
15. N. E. Ockerbloom, *Hydrocarbon Process.* **51** (1972) no. 4, 114–118.
16. Eastman Kodak, US 3240803, 1961 (B. Thompson, S. D. Neely).
17. Toray Industries, JP-Kokai 7834738, 1978 (H. Torikata, H. Morimoto, H. Eguchi, K. Nakaoka).
18. W. F. Brill, *Ind. Eng. Chem.* **52** (1960) 837.
19. P. Stecher: *Trimellitic Anhydride and Pyromellitic Dianhydride,* Noyes Data, Park Ridge, N.J., 1971.
20. *Trimellitic Anhydride. Data Sheet,* Amoco Chem. Corp., Chicago 1979. Standard Oil Indiana, US 3819659, 1974 (R. H. Baldwin, D. E. Burney, P. H. Towle, D. G. Micklewright); US 3931304, 1976 (G. L. Wampfler).
21. Mitsubishi Gas Chem. Co., DE-OS 3044752, 1981 (T. Suzuki, K. Katahara, S. Naito, T. Tsuji).
22. Bergwerksverband, *Hydrocarbon Process* . 50 (1971) no. 11, 214.
23. Gelsenberg-Benzin AG, DE 1246707, 1963; DE 1518917, 1965; DE 1568682, 1966; DE 1618461, 1967. Veba-Chemie, DE-OS 1593536, 1966.
24. Princeton Chem. Res. Inc., DE-AS 1468851, 1965 (J. F. McMahon).
25. Hüls AG, DE-OS 2751979, 1979 (H. Haferkorn).
26. Rhône-Poulenc, GB 1226695, 1971 (M. Balme, B. Rollet).
27. Veba Chemie, DE-OS 1930011, 1969 (G. Ibing, K. Neubold).
28. Gelsenberg-Benzin AG, DE-AS 1289039, 1965 (K. Peterlein).
29. Princeton Chem. Res., US 3592827, 1971 (R. I. Bergmann).
30. Rütgers Werke, BE 525660, 1956. BASF, US 3708504, 1973 (O. Kratzer, H. Suter, F. Wirth).
31. Nippon Kayaku K. K., DE-OS 2846432, 1980 (R. Hasegawa).
32. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* 4th ed., Akademie Verlag, Berlin 1966, p. 924. *Bios Final Report* no. 1484, p. 21.
33. K. Venkataraman: *The Chemistry of Synthetic Dyes,* vol. 2, Academic Press, New York-London 1952.
34. B. Raecke, H. Schirp, *Org. Synth. Coll.* **5** (1973) 813 – 816.
35. I. I. Kiiko, *Sov. Chem. Ind. (Engl. Transl.)* **7** (1975) 818 – 821.
36. *Chem. Week* **112** (1973) no. 17, 43.
37. H. E. Fierz-David, L. Blangey: *Grundlegende Operationen der Farbenchemie,* 8th ed., Springer Verlag, Wien 1952, pp. 219 – 221.
38. IG-Farbenind., DE 604445, 1934 (H. Vollmann, M. Corell).
39. Hoechst, DE 2209692, 1972; DE 2262858, 1972 (K. Eiglmeier).
40. Hoechst, DE 2343946, 1973 (F. Röhrscheid).
41. *Mod. Plast. Int.* **13** (1983) Nov., 35.

42. M. Sittig: *Hazardous and Toxic Effects of Industrial Chemicals,* Noyes Data, Park Ridge, N.J., 1979.

43. F. J. Letkiewiez, *US Environ. Prot. Agency NTIS* **PB-248** (1975) 838.

Carboxymethyl Cellulose → Cellulose Ethers
Carburization → Metals, Surface Treatment

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Carcinogenic Agents

Standard Article

Wolfgang Pfau[1]

[1]Department of Toxicology and Environmental Medicine of the Fraunhofer Society and Department of Toxicology, Hamburg University Medical School, Hamburg, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a05_261
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Carcinogenesis
3. Genotoxic Mechanisms
3.1. Metabolic Activation
3.2. DNA Binding
3.3. DNA Repair
3.4. Molecular Targets in the Genome
3.4.1. Oncogenes
3.4.2. Tumor-Suppressor Genes
4. Nongenotoxic Mechanisms
4.1. Tumor Promoters
4.2. Hormones
4.3. Peroxisome Proliferators
5. Identification of Carcinogenic Agents
5.1. Animals Tests (in vivo)
5.2. In vitro Assays
5.3. Modeling
5.4. Epidemiological Studies
5.5. Molecular Epidemiology
6. Predisposing Factors
7. Classes of Carcinogenic Agents
7.1. *N*-Nitroso Compounds
7.2. Benzene
7.3. Polycyclic Aromatic Hydrocarbons
7.4. Nitroaromatic Compounds
7.5. Aromatic Amines
7.6. Halogenalkyls
7.7. Aldehydes
7.8. Oxygen Damage
7.9. Alkylating Agents
7.10. Metals
7.11. Natural Compounds
7.12. Hormones
7.13. Mixtures
7.14. Mineral Fibers
7.15. Viruses and Other Biological Agents
7.16. Nonionizing Radiation

7.17. Ionizing Radiation
7.18. Electromagnetic Fields
8. Human Carcinogens
9. Forms of Exposure
9.1. Occupational Exposure
9.2. Environmental Exposure
9.3. Lifestyle
9.4. Diet
9.5. Drugs
9.6. Exposure Levels in Developing Countries
9.7. Safe Handling of Carcinogenic Agents

[Top of Page]

1. Introduction

Cancer incidences are increasing worldwide and in the industrialized countries malignant tumors are a leading cause of death, second only to cardiovascular diseases. Improvements in cancer treatment, surgery, pharmacological and radiation therapy, have not been able to reduce the threat that this illness is posing to the general public.

The term cancer covers a broad spectrum of malignant diseases that affect the blood-forming system in the bone marrow (leukemia, 4 % of human cancers), soft tissues and bones (sarcoma, 2 %), the lymphatic system (lymphoma, 5 – 9 %) and, most often, epithelial tissues including the skin, lung, gastrointestinal tract, and glands (carcinoma, 85 %).

Intensive investigations into the genesis of neoplasms have revealed a number of clues at the molecular level that may help in prevention and treatment of cancer.

[Top of Page]

2. Carcinogenesis

In most cases the induction of cancer is the consequence of low dose chronic exposure to carcinogenic agents and the process of carcinogenesis may take several years or even decades. A broad spectrum of mechanistic variations has been discovered. However, generally molecular biological investigations have confirmed the concept of carcinogenesis as being a multi-step process which is generally divided in the phases of initiation, promotion, and progression. On each of these phases carcinogenic agents may exert their specific activity and it has been possible in a few types of tumors to assign specific genetic changes in the process of transforming a normal, healthy cell into a malignant clone of tumorous cells (Fig. 1).

Figure 1. Carcinogenesis is a multi-step process. At any stage of this development carcinogenic agent may have an impact

Knowledge of the mechanisms of carcinogenesis may help to intervene in this process and reduce the incidence rates of cancer.

It is a biological fact that an increase in the number of tumors induced and/or a decrease of the latent period is observed with increasing dose of a carcinogen. For carcinogenic agents with a genotoxic mechanism, however, a no-effect dose cannot be defined. On the other hand, for carcinogens that act via nongenotoxic mechanisms thresholds are assumed. Thus, for the purpose of risk assessment it is pivotal to understand mechanisms of action in order to define possible threshold levels of exposure [1].

[Top of Page]

3. Genotoxic Mechanisms

The majority of carcinogenic agents acts as genotoxic agents, i.e., by damaging cellular DNA. These agents are termed mutagenic (see ⟶ Mutagenic Agents) or clastogenic.

3.1. Metabolic Activation

With few exceptions (see Section Alkylating Agents) carcinogenic organic chemicals (termed procarcinogens) are metabolically activated in the body to reactive metabolites (Fig. 2). These metabolic transformations are considered accidental in the general metabolism which improves hydrophilicity of foreign agents to facilitate their excretion from the body.

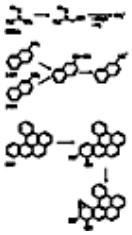

Figure 2. Metabolic activation of chemical carcinogens: Formation of electrophilic ultimate carcinogens. *N*-Nitrosodimethylamine (dimethylnitrosamine, DMNA) is activated via α-hydroxylation to a carbenium ion; polycyclic aromatic amines (2-aminonaphthalene, 2-AN) are oxidized and polycyclic aromatic nitro compounds (2-nitronaphthalene, 2-NN) are reduced to form a nitrenium ion; polycyclic aromatic hydrocarbons (dibenz[a,l]pyrene, DBP) are oxidized to form a dihydrodiol derivative and subsequently vicinal dihydrodiol epoxides in the fjord region of the molecule

Reactive metabolites formed may include radicals and most commonly electrophilic species (proximate or ultimate carcinogens). These may intermediately release positively charged ions that may react with nucleophilic components in the cell, such as H_2O or glutathione (a tripeptide containing cystein) resulting in a detoxification. These reactive species may also bind to cellular proteins leading to fatal cellular injury or the formation of immunogenic antigens. However, these electrophilic metabolic intermediates may also covalently modify the DNA in the cell nucleus.

3.2. DNA Binding

Several nucleophilic sites in the DNA have been identified as targets for covalent modification by electrophilic agents. DNA bases may be modified with preferential binding sites of alkylating agents (A), aromatic amines (AA) or polycyclic aromatic hydrocarbons (PAH).

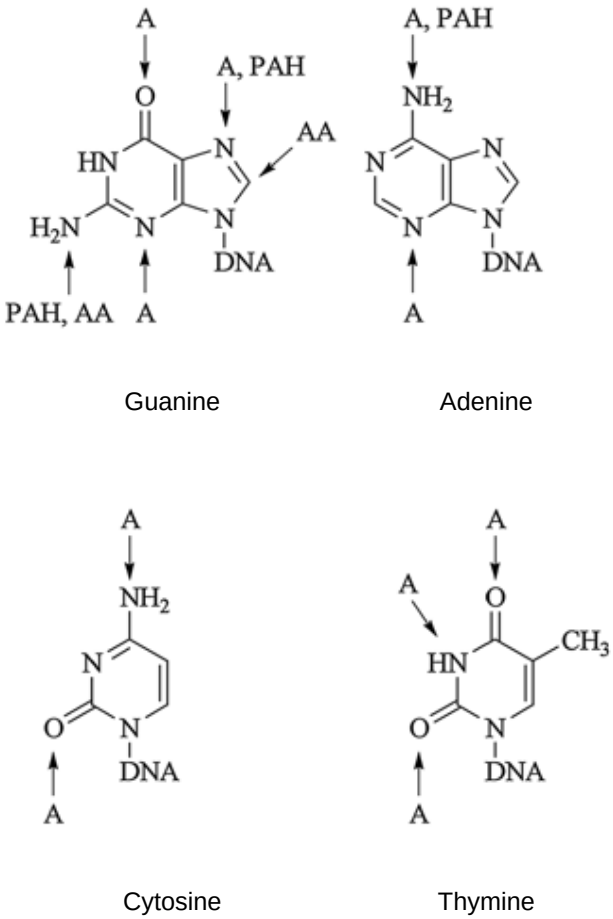

The most prominent target of DNA binding is the base guanine where binding has been detected at positions N^7, O^6, N^2, and C^8, depending on the carcinogenic agent involved. These covalent modifications have been termed DNA adducts and may yield mutations in the genome upon DNA replication. The ability of an agent to form covalent DNA adducts is, thus, an indicator of its genotoxic activity and this damage may lead to the development of neoplasms. Although the formation of these DNA adducts is not sufficient for tumor development, all genotoxic agents identified as human carcinogens have been shown to bind covalently to DNA.

A form of indirect DNA damage is produced by reactive oxygen species (e.g., hydroxyl radicals) which may be formed by oxidants (Section Oxygen Damage) or by ionizing radiation (see Section Ionizing Radiation).

3.3. DNA Repair

The integrity of the genome is pivotal for the survival of a cell. Therefore, a highly sophisticated system of removal and repair of DNA lesions is at work in a cell to identify and remove modifications. Indeed, mechanisms have been identified (involving

the tumor suppressor gene *p53*) that arrest DNA replication and cell division in order to allow sufficient time for this repair to occur. Deficient DNA repair is often associated with an increased risk for developing cancer.

3.4. Molecular Targets in the Genome

Genetic targets for carcinogenic agents have been proposed that are termed proto-oncogenes or tumor suppressor genes. A number of these have been found to be frequently altered in tumor cells.

3.4.1. *Oncogenes*

Proto-oncogenes (⟶ Toxicology) are genes that encode for proteins implicated in vital functions of a cell. These gene products, e.g., growth factors, protein kinases, and receptors, are involved in signal transduction mechanisms. Certain mutations induced in these genes result in an activation of these proto-oncogenes: point mutations (change of one letter in the DNA code), amplifications (enhanced expression, i.e., production, of the gene product) or translocations (also resulting an increase in expression).

3.4.2. *Tumor-Suppressor Genes*

When mutations in a tumor-suppressor gene of a cell lead to a deletion of the encoded protein or the production of a nonfunctional protein this may result in the cell's expansion and unrestrained growth [2] (⟶ Toxicology).

[Top of Page]

4. Nongenotoxic Mechanisms

There are a number of carcinogenic compounds that induce cancer without possessing genotoxic activity. These agents act by various mechanisms and are summarized as *epigenetic carcinogens*. In contrast to the DNA damaging (genotoxic) agents, for epigenetic carcinogens doses can be established below which no adverse effects are expected to occur. These agents are often carcinogenic via mechanisms that involve acute toxic effects, thus, threshold levels of exposure can be defined.

4.1. Tumor Promoters

Tumor promoters are not carcinogenic on their own, but when applied subsequently to a (genotoxic) tumor initiator increase the number of tumors per animal or shorten the latency period of tumor development in an experimental setting. Thus, tumor promotion is an operational definition rather than a mechanistic concept. The biological phenomenon of tumor promotion is very specific for the animal species tested and the organ site of tumorigenesis, where mouse skin and rat liver are the most commonly tested organ sites. Mechanisms under consideration are mitogenesis (increased cell proliferation) and inflammatory processes.

Different from genotoxic mechanisms, promotion is considered to be reversible and threshold values for promoters are assumed to exist.

Prominent promoting agents are phorbol (4 β,9 α,12 β,1 α,20-pentahydroxy-1,6-tigliadien-3-one) esters [such as tetradecanoyl phorbol acetate (TPA) isolated from *Euphorbiaceae*], phenobarbital, saccharin, halogenated hydrocarbons (including DDT and other organochloro pesticides, polyhalogenated biphenyls (PCBs) and polyhalogenated dibenzodioxins/dibenzofurans (TCDD and congeners).

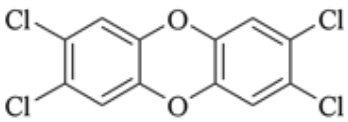

TCDD

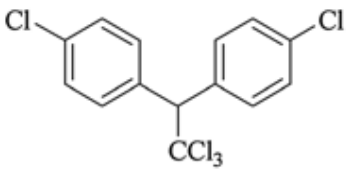

DDT

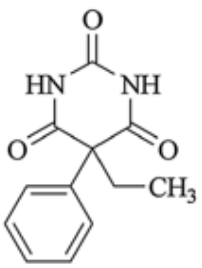

Phenobarbital

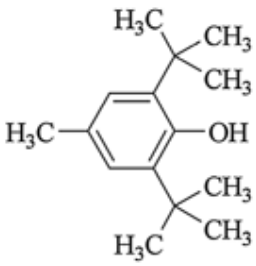

BHT
2,6-di-*tert*-butyl-4-methylphenol
(butylared hydroxytoluene)

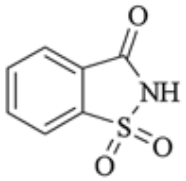

Saccharin

4.2. Hormones

Sexual hormones, mainly estrogens and progesterons, have been shown in animal experiments to increase the number of tumors at certain organ sites such as mammary gland, prostate, and endometrium. Tumor incidences in these organs are dependent on the sex; therefore, it is believed that a promoting effect of sex hormones is involved. The lifetime estrogen dose is considered a major factor for human breast cancer risk.

Steroid hormones are essential for the growth, differentiation, and function of many organs. Thus, the induction of cell proliferation is believed to be the mechanism by which steroid hormones act as carcinogenic agents.

In addition, genotoxic mechanism have been reported for estrogen, its metabolites, and related drugs (diethylstilbestrol, tamoxifen, cyproterone acetate).

A number of anthropogenic environmental contaminants have been discovered to have affinity to the estrogen or androgen receptor and to either mimic (agonists) or inhibit hormonal activity (antagonist). These xenoestrogens (hormonal disrupters) include polyhalogenated biphenyls, halogenated dibenzodioxins, and nonylphenol. Additionally, estrogen, its metabolites and anticontraceptive drugs (e.g., ethinyl estradiol) have been detected in sewage effluents.

Moreover, it has been shown that naturally occurring compounds may bind to the estrogen receptors. These compounds include flavonoids such as genistein and daidzein.

However, these exogenous hormones have generally a much lower affinity to the receptor than the endogenous hormones. Currently, it is unclear if these low levels of xenoestrogens and phytoestrogens have an impact on human endocrine homeostasis. A possible contribution to human cancer risk is a matter of speculation.

4.3. Peroxisome Proliferators

A considerable number of nongenotoxic agents that induce hepatic tumors in the rat have been shown to increase the number of peroxisomes in the rat liver. These cell organelles are responsible for the catabolism of long-chain fatty acids and cholesterols. During catabolism reactive oxygen species are produced in the hepatic cell and these may be the ultimate carcinogenic agents. Whether this mechanism may also play a role in human liver is a matter of dispute. Peroxisome proliferators include plasticizers (e.g., phthalates), halogenated organic solvents (1,1,2-trichloroethylene) and blood lipid lowering drugs (clofibrate, fenofibrate).

[Top of Page]

5. Identification of Carcinogenic Agents

In order to reduce cancer risk it is necessary (1) to identify the relevant agents and (2) to minimize exposure to these agents. The different approaches to identify the carcinogenic activity of an agent are outlined below. Ideally, both, experimental and human epidemiological data are available.

5.1. Animals Tests (in vivo)

In the National Toxicology Program in the United States and as a result of legislative requirements for new drugs and chemicals of a production volume of more than 1000 t/a in Europe, long term animal carcinogenicity studies are carried out. These two-year animal experiments are generally performed with rats, using several doses of the compound, including the maximum tolerated dose (MTD) which induces marginal toxic effects in the test animals. In the 1990s this concept has been challenged since this dosing regimen is generally several orders of magnitude higher than the anticipated human exposure and may produce many false positives because of toxic effects induced at the high doses [3].

Since this type of long term study is both labor- and cost-intensive, several approaches have been developed to introduce short-term in vitro or in vivo assays that may point towards a possible carcinogenic activity of a test substance.

5.2. In vitro Assays

Although the process of carcinogenesis may generally not be mimicked in a test tube or petri dish, several in vitro assays that are very sensitive in detecting genotoxic activity have been developed. These tests include microbial mutagenicity assays (e.g., the Ames test using various strains of *Salmonella typhimurium*, ⟶ Toxicology – The Ames Test for Bacterial Mutagenicity) and genotoxicity tests in mammalian cells. These in vitro assays are generally used as prescreening tests and may give hints on a possible carcinogenic activity of a test compound. They may also help in elucidating the mechanism of action of a rodent carcinogen.

More closely resembling the in vivo tumor induction are cell transformation tests. Here, the transformation of apparently normal cells into tumor cells is recorded as endpoint. These transformation tests are capable of detecting both, genotoxic and nongenotoxic carcinogens.

5.3. Modeling

Computer-assisted analysis of the results from several hundred rodent bioassays on the carcinogenic potency of organic chemicals has helped to identify certain structural elements as carcinogenic alerts [4]. These functional elements ("alerting structures") are listed in Table 1.

Table 1. Potential structural alerts for mutagenicity

Alkyl esters of phosphonic or sulfonic acids	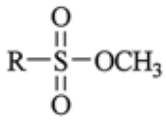
Aromatic nitro groups	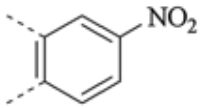
Aromatic amino groups	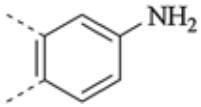
Aromatic hydroxylamines (+ esters)	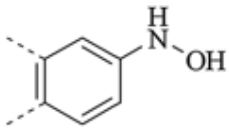
Aromatic ring-N-oxides	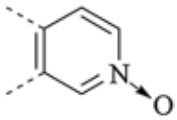
Aromatic azo groups	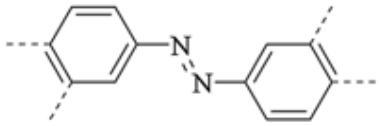
Aromatic azides	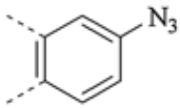
Aromatic mono- or dialkyamino groups	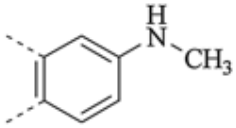

Alkyl hydrazines

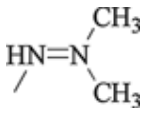

Alkyl aldehydes

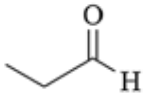

N-Hydroxymethyl groups

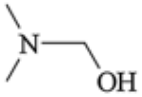

Monohaloalkenes

Halogenated methanes (X = H, F, Cl, Br, I)

CX_4

N- or *S*-mustard (β-haloethyl)

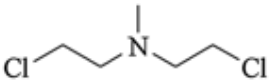

N-Chloramines

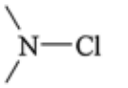

Propiolactone

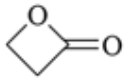

Aromatic aziridinyl derivatives

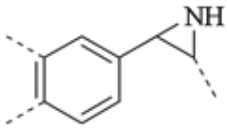

Primary alkyl halides

CH_2Cl

Carbamates

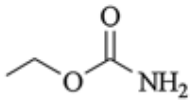

N-Nitroso alkylamines

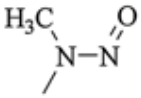

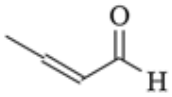

Michael reactive a, β-unsaturated aldehydes

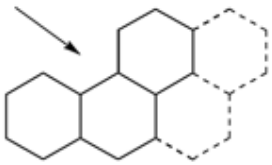

Polycyclic aromatic hydrocarbons (bay regions)

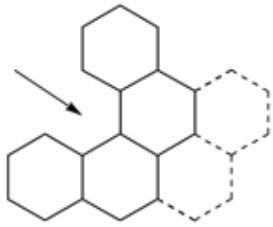

Polycyclic aromatic hydrocarbons (fjord region)

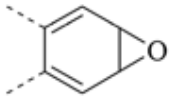

Aliphatic and aromatic epoxides

However, even highly sophisticated calculation methods are presently not capable to replace biological testing.

5.4. Epidemiological Studies

Since experimental studies with possibly carcinogenic compound can, for obvious ethical reasons, not be performed with humans, data on human carcinogenicity have to be deduced by extrapolation from animal experiments, in vitro experiments (using human tissue or cells) and epidemiological studies.

Exposures may be occupational, environmental or accidental (see Chap. Forms of Exposure). Together with experimental studies such results may also provide insights into the mode of carcinogenic activity and biological plausibility. This combined approach is the most secure way of identifying a potential carcinogenic hazard [5].

5.5. Molecular Epidemiology

Conventional epidemiological investigations have been refined by mathematical improvements. However, as was pointed out by TAUBES [6] in a controversial paper these studies are only suitable to detect low risks for the general population and cannot be employed for individual risk assessment. A novel approach that takes into account the individual exposure levels and predisposing (e.g., genetic) factors has been termed molecular epidemiology [7]. By analyzing the internal dose (such as serum levels of a given compound) and/or metabolic capacities of individuals, this approach may help to identify both, subpopulations at risk and agents that pose a significant risk for human cancer.

[Top of Page]

6. Predisposing Factors

Several types of cancer are inheritable diseases. Since most cancer incidences occur at a late age past the reproductive age, the selective mechanisms of evolution do not help to eradicate these predispositions.

These inheritable forms of cancer account for less than 10 % of all cancer cases. More importantly, the predisposition to develop a malignant tumor may be inherited but is clinically manifest only when other factors such as exposure to specific carcinogenic agents occur. The mechanisms leading to an elevated individual risk include certain enzyme activities in the pathway of metabolic activation such as an enhanced enzyme activity with regard to toxification, a reduced activity in detoxification steps, or defects in DNA repair capacity. Individual risk is also a function of the levels of exposure to carcinogenic agents, the duration of this exposure, and the time in life when this exposure occurs (higher risk at young age) [7].

[Top of Page]

7. Classes of Carcinogenic Agents

7.1. *N*-Nitroso Compounds

Included in this group are the directly acting *N*-nitrosourea derivatives and the *N*-nitrosodialkylamines. The latter group requires metabolic activation (α-hydroxylation) to release an alkylating species that modifies the cellular DNA (see Fig. 2). *N*-Nitrosamines are among the most potent carcinogenic agents ever tested and *N*-nitrosodimethylamine has been detected in a variety of foods, industrial products, and the environment.

N-Nitrosamines may also be formed endogenously in the acidic environment of the stomach by nitrosation of alkylamines with nitrite. *N*-Nitroso derivatives of nicotine and related alkaloids contained in tobacco are formed upon combustion of tobacco (termed tobacco specific nitrosamines) and are considered major contributors to the carcinogenic potential of tobacco smoke.

7.2. Benzene

Benzene is a constituent of crude petroleum and motor gasoline and is produced in large quantities in the chemical industry as a synthetic intermediate. Both experimental and epidemiological studies have identified benzene as a human carcinogen inducing different forms of leukemia.

7.3. Polycyclic Aromatic Hydrocarbons

When organic matter such as fossil fuels, wood, tobacco, or waste is combusted, polycyclic aromatic hydrocarbons (PAH) are formed, especially under oxygen deficient conditions.

Members of this substance group were among the first individual compounds to be identified as carcinogenic agents. The carcinogenic potency varies among different PAHs depending on the number of aromatic rings and structural features. The most potent PAH are among the five- to six-ring compounds with benzo[*a*]pyrene being the prototype of a carcinogenic PAH. It was shown that PAHs require a multi-step metabolic activation process which results in the formation of dihydrodiol epoxide derivatives (Fig. 2). Genotoxic potency of these dihydrodiol epoxides is enhanced when activation occurs in a *bay region* or in a *fjord region* of a PAH molecule (Table 1) (see also ⟶ Tar and Pitch – Toxicology). Thus, PAHs with these structural features such as benzo[*a*]pyrene, dibenz[*a*,*h*]anthracene, or dibenz[*a*,*l*]pyrene have been shown to be most potent carcinogenic PAHs. Furthermore, biological activity of these diol epoxides varies with on the conformation of the different hydroxyl substituents.

7.4. Nitroaromatic Compounds

Associated with the occurrence of PAHs are their nitro derivatives. Relatively high levels of these compounds are observed in emissions of diesel-fueled motor vehicles. These compounds also require metabolic activation, the enzymatic process involves reduction of the nitro group to a hydroxyl function that results in a nitrenium ion as ultimate carcinogenic species (Fig. 2).

7.5. Aromatic Amines

Similar to aromatic nitro compounds aromatic amines are also transformed metabolically to hydroxyl derivatives, albeit via an oxidative pathway (Fig. 2). This class of carcinogenic compounds includes the monocyclic aromatic amines (aniline and its derivatives), polycyclic aromatic amines (e.g., 4-aminobiphenyl, ABP) and heterocyclic aromatic amines (e.g., aminoimidazo quinoxalines). A number of these aromatic amines are found in tobacco smoke and polycyclic amines such as 2-aminonaphthaline (β-naphthylamine), benzidine, and 4-aminobiphenyl have been associated with the increased risk of urinary bladder cancer observed among cigarette smokers. The high incidence of bladder cancer among workers in the dye industry has also been attributed to polycyclic impurities in aniline, namely 2-aminonaphthalene and 4-aminobiphenyl, but recently ortho-toluidine has also been considered as etiologic agent in these occupational cancers.

7.6. Halogenalkyls

Halogenated aliphatic hydrocarbons are widely used as solvents and industrial intermediates. These include halomethanes like CCl_4 and $CHCl_3$ that have been shown to induce liver tumors in rodents via nongenotoxic mechanisms. Dichloromethane induces tumors in lung and liver of mice; the significance of these findings for human cancer risk has been questioned.

Halogenated unsaturated hydrocarbons induce tumors in the kidneys of exposed rodents. A genotoxic mechanism involving a β-lyase mediated metabolic activation has been discovered.

Vinyl chloride (chloroethylene) the monomer of poly(vinyl chloride) (PVC) is an established liver carcinogen inducing hemangiosarcoma of the liver of exposed individuals.

Polyhalogenated cyclic hydrocarbons including pesticides (e.g., DDT and hexachlorocyclohexane), polyhalogenated biphenyls, polyhalogenated dibenzodioxins and dibenzofurans are exceptionally persistent in the environment. They are accumulated in the fatty tissues of animals and are enriched in the food chain. These compounds may act by various mechanisms as (nongenotoxic) tumor promoting agents.

Haloethers, such as bis(chloromethyl) ether or bis(chloropropyl) ethers, have been detected in substantial amounts in the environment as persistent contaminants. These ethers have a considerable carcinogenic potential.

7.7. Aldehydes

Formaldehyde is an ubiquitous chemical intermediate used in several fields of health sciences, in the production of synthetic resins, plastics and chip- and fiberboard. Formaldehyde is mutagenic and induces tumors in the rat (nasal cavity) following high level exposure. It has been suggested that the carcinogenic effect of formaldehyde in rodents is due to high-dose cytotoxic effects and there is inadequate epidemiological evidence of a carcinogenic potential in humans. Acetaldehyde and glutaraldehyde must be viewed similarly.

Unsaturated aldehydes such as crotonaldehyde may react with DNA bases to form cyclic DNA adducts and have been shown to be strongly carcinogenic in rodents.

7.8. Oxygen Damage

Reactive oxygen species are formed as a result of physiological cellular function. Glutathione and several enzymatic systems detoxify these reactive species (e.g., $OH^{\bullet}$) and guard the cellular function up to a certain level. Additional oxidative stress is imposed by chemicals that undergo one-electron oxidation or reduction in the cell forming radical species. This may lead to oxidative damage of DNA (e.g., 8-oxoguanine) that is usually effectively repaired but can also result in mutations. Indirectly, the oxidative damage of cellular lipids may lead to premutagenic lesions via modification of DNA by lipid peroxidation products. Oxidative damage of DNA has been proposed as a mechanism of heavy metal carcinogenesis. The carcinogenic effect of ionizing radiation is believed to be effected by the formation of reactive oxygen species.

Antioxidative compounds (such as α-tocopherol, vitamin E) may have an anticarcinogenic effect because they detoxify the reactive oxygen derivatives.

7.9. Alkylating Agents

Mustard gas (sulfur mustard) was used in large quantities in chemical warfare. Both experimental data and studies from occupational exposure indicate this DNA alkylating agent to be carcinogenic to humans. Similarly, there is sufficient evidence that alkylating antineoplastic drugs (cyclophosphamide ⟶ Cancer Chemotherapy – Cyclophosphamide, melphalan ⟶ Cancer Chemotherapy – Melphalan) increase human cancer risk (urinary bladder, leukemia) because the modification (alkylation) of cellular DNA is not restricted to the tumor but causes genotoxic damage in various organs of the patient.

7.10. Metals

There is indisputable evidence that beryllium, chromium, nickel, and arsenic represent a carcinogenic hazard to humans (pulmonary carcinoma and cancer of other organ sites), when present in certain well defined physicochemical forms. The evidence for the carcinogenicity of cadmium is less clear (probable targets are lung and prostate) although in the laboratory chronic inhalative exposure resulted in lung tumors in rats. Either way, cadmium compounds are very toxic in other respects (kidney and bones) so that they should be used with particular care.

The mechanism of metal carcinogenesis has tentatively been described as a genotoxic pathway since metal salts have been shown to be mutagenic in several in vitro short term tests.

Earlier misconceptions were based on a lack of mutagenic activity of these metals in the Ames test, the most widely used mutagenicity test. Other metals that are carcinogenic in laboratory animals include cobalt, iron, lead, titanium, and zinc.

With the exception of arsenic, metal carcinogenesis seems not to be a problem of environmental contamination at trace levels but is relevant to occupational exposure.

7.11. Natural Compounds

The pollution of the environment with synthetic chemicals has supported the notion that these anthropogenic substances are the major cause of human cancer. Natural compounds are generally considered as safe by the general public. However, as has been brought forward rather pointedly by AMES and coworkers there is a considerable number of substances of natural origin that have a strong carcinogenic potency. These authors hypothesize that natural agents are indeed far more relevant to human carcinogenesis than environmental contamination with synthetic substances.

Aflatoxins are produced by molds (*Aspergillus flavis*). Among these, aflatoxin B_1 is one of the most potent liver carcinogens in the rat and acts via a genotoxic pathway: after metabolic activation the epoxide reacts with guanine residues in the DNA. Molecular epidemiological studies have shown that in certain regions of east Africa (The Gambia and Senegal) and provinces of China the endemic high incidences of liver cancer are associated with aflatoxin-contaminated food in a synergistic combination with hepatitis B viral infection.

Carcinogens have also been detected in plants, e.g., phorbol esters [such as tetradecanoyl phorbol acetate (TPA) isolated from *Euphorbiaceae* or Aquilid A from bracken fern (*Pteridium aquilinum*)] and herbal drugs such aristolochic acid derived from birthwort (*Aristolochia* species), pyrrolizidine alkaloids (*Senecio ssp*), or anthraquinones (e.g., senna leaves). Nicotine in tobacco is a constituent of carcinogenic tobacco smoke (see Section Aromatic Amines).

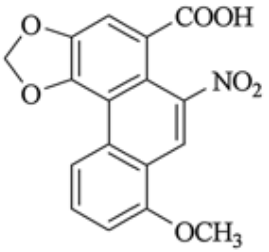

Aristolochic acid I

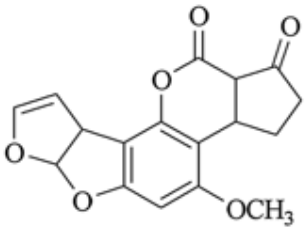

Aflatoxin B_1

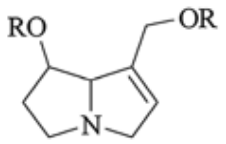

Pyrrolizidine

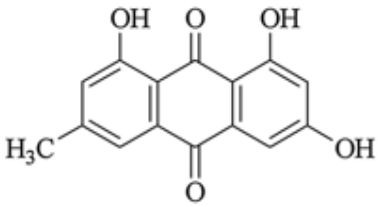

Emodin
(anthraquinone derivative)

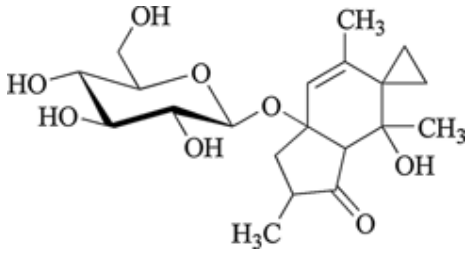

Aquilid A

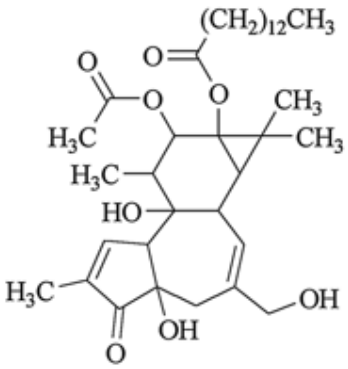

TPA

When proteinaceous food is grilled, fried, or broiled, heterocyclic aromatic amines are formed via the Maillard reaction. Among these, the aminoimidazo quinoxalines exhibit extremely strong mutagenic responses in the Ames test and have also been shown to induce tumors at several organ sites in rodents. Thus these heterocyclic amines are considered to be involved in human cancer at organ sites (colon, breast, prostate, pancreas) that are associated with a diet high in fried or grilled meat.

7.12. Hormones

See also Section Hormones.

Hormones include both endogenous sexual hormones, hormonal drugs, and environmental compounds with affinity to hormonal receptors. The latter can be of natural origin (phytohormones) or originate from anthropogenic environmental contaminants. The affinity to the hormonal receptors varies between different compounds by several orders of magnitude.

7.13. Mixtures

Humans are most often exposed to mixtures of carcinogenic agents, such as mixtures of polycyclic aromatic hydrocarbons in smoke or coal tar, mixtures of heterocyclic amines in fried meat or even complex mixtures of different types of chemical agents. These may act in an additive, synergistic (overadditive), or inhibitory manner. Since the composition of these complex mixtures is variable (for example, tobacco smoke or environmental pollution) it is difficult to predict quantitatively the biological potency of such complex mixtures. One of the most complex and most important chemical mixtures is the human diet.

7.14. Mineral Fibers (⟶)

Dust of the naturally occurring rock asbestos, particularly blue asbestos (crocidolite) can induce mesothelioma (malignant tumors of the pleura) following inhalative exposure. The physical properties of the needle-like fibers (length > 5 µm, diameter < 3 µm) and the very long biological half-life are crucial for the activity. Thus, other fibers (erionite, a zeolite mineral) with similar properties have also carcinogenic potential.

Other synthetic inorganic fibers (rock wool, glass fibers, ceramic fibers) are less persistent in the body; probably for this reason the carcinogenic potential of these is less certain.

7.15. Viruses and Other Biological Agents

Viral infection is a major risk factor for human cancers, chronic infection with specific viruses being implicated in the genesis of up to 20 % of tumors.

Retroviruses.
HTLV-I. has been implicated as a cause of adult T-cell leukemia/lymphoma; less than 4 % of individuals infected with this virus develop the tumor, and then only several decades after the initial infection.

HIV-1, HIV-2. The increased incidence of lymphoma and Karposi's sarcoma in patients infected with human acquired immune deficiency syndrome (AIDS) viruses is almost certainly the result of the chronic immunosuppression that occurs in the disease, rather than a direct oncogenic effect.

DNA and RNA Viruses. Chronic infection with human papilloma virus (HPV), acquired as a result of sexual transmission , has been implicated as a cause of cervical, anal, and vulval cancer. HPV is present in more than 60 % of cervical carcinomas, and additional risk factors such as herpes simplex virus and smoking, seem to be necessary.

Hepatitis B Virus (HBV) and Hepatitis C Virus (HCV). Epidemiological evidence links chronic infection with HBV or HCV to the development of hepatocellular carcinoma. It has been suggested that these viruses may induce liver cancer indirectly, acting as a promoter by causing chronic inflammatory liver disease.

Herpes Viruses. Epstein Barr virus (EBV) may be involved in the pathogenesis of a number of human tumors, including Burkitt's lymphoma, Hodgkin's disease and salivary gland, urogenital or nasopharyngeal carcinoma. In AIDS patients, lymphomas (immunoblastic and nervous system) are also associated with this virus.

Parasites. Schistosomiasis, a parasitic infection with different types of the flatworm *Schistosoma* is endemic in different tropical regions of the world and is associated with an increased risk of cancer of the urinary bladder. Chronic inflammation of the bladder and metabolic changes leading to increased endogenous formation of *N*-nitrosamines is being considered as mechanism of action.

Bacterial Infection. Infection with *Heliobacter pylori* is believed as causally involved in human stomach cancer.

7.16. Nonionizing Radiation

Skin cancer is the most common neoplasm in Caucasians in the United States with a lifetime risk nearly equal to that of all other cancers combined. Sun exposure is the major environmental agent implicated in induction of nonmelanoma skin cancer (basal cell carcinoma or squamous cell carcinoma).

The initiating event is the DNA damage induced by UV radiation: cross-linking of two thymidine bases to form a cyclobutane ring (2 + 2-cycloaddition). This hypothesis is corroborated by the observation that patients with the rare inherited disorder *Xeroderma pigmentosum* have a skin cancer risk 1000 times higher than that of the general population, probably due to a defect in DNA nucleotide excision repair.

With the reduction of the stratospheric ozone layer, UV exposure is increasing and skin cancer rates are rising. Preventing severe sunburn especially at young age and regular medical checkups are important to minimize the risk.

7.17. Ionizing Radiation

Long term follow-up of survivors of the atomic bomb releases in Hiroshima and Nagasaki, as well as of patients that received extensive radiation therapy provided epidemiological evidence that ionizing radiation can induce cancer in any organ in which cancer occurs naturally. In particular the rate of leukemia, cancer of the breast, lung, and thyroid are increased with latency periods up to 20 years. Most pronounced effects were observed when the exposure to a single high dose of radiation occurred at a young age.

Radiation may cause DNA damage indirectly: formation of reactive radicals by radiochemical homolysis of water molecules and induction of DNA strand breaks by these reactive intermediates.

The relationship between cancer and low-level radiation is a subject of much controversy. The considerable level of natural background radiation has been said to exceed the low levels of exposures such as in the vicinity of nuclear power or reprocessing plants. Recently, the observation of inheritable genomic instability induced by low amounts of radiation has been discussed.

7.18. Electromagnetic Fields

The increase of electrical appliances in the home, the widespread use of mobile phones, high tension lines and the constant exposure to computer monitors and television sets results in an exposure of the human body to electromagnetic fields. These fields are not without biological effects; however, the proposed carcinogenic (promoting) effects of extremely low frequency (0 – 300 Hz) electromagnetic fields have not been definitely confirmed experimentally.

[Top of Page]

8. Human Carcinogens

The evidence provided by epidemiological data is currently taken as the only acceptable proof for a given compound to be a human carcinogen. It is a matter of controversy whether or to what extent experimental data, e.g., from rodent assays may be regarded as adequate data to classify a compound a human carcinogen.

It has been pointed out that standard testing at the MTD results in a number of false positives, e.g., many agents are labeled carcinogenic only because of the cytotoxic action at unrealistic high doses [1], [3]. The International Agency for Research on Cancer (IARC) has evaluated a number of agents [5]. These are listed below and are classified according to the IARC categorization scheme:

Carcinogenic to humans (Group 1): a causal relationship has been established between the exposure and human cancer

Chemicals and groups of chemicals
Aflatoxins
4-Aminobiphenyl
Arsenic and Arsenic compounds
Asbestos

Benzene
Benzidine
Bis(chlormethyl)ether
Chromium compounds (hexavalent)
Erionite
Mustard gas (sulfur mustard)
2-Naphthylamine
Nickel compounds
Radon and its decay products
Talc containing asbestiform fibers
Vinyl chloride

Mixtures

Alcoholic beverages
Betel quid with tobacco
Coal-tar pitches
Coal tar
Mineral oils
Shale oils
Soots
Tobacco products, smokeless
Tobacco smoke

Drugs

Analgesic mixtures containing phenacetin
Azathioprine
N,*N*-Bis(2-chloroethyl)-2-naphthylamine (Chlornaphazine)
1,4-Butanediol dimethanesulfonate (myleran)
Chloroambucil
1-(2-Chloroethyl)-3-(4-methylcyclohexyl)-1-nitrososurea (Methyl-CCNU)
Cyclophosphamide
Cyclosporin
Diethylstilbestrol
Melphalan
8-Methoxypsoralen plus UV radiation
Estrogen replacement therapy
Estrogens, nonsteroidal
Estrogens, steroidal
Thiotepa
Treosulfan

Biological and physical factors

Hepatitis B virus
HTLV 1
Ionizing radiation
Ultra-violet radiation

Probably carcinogenic to humans (Group 2A): a positive association has been observed, a causal interpretations credible, but chance bias or confounding factors cannot be ruled out with confidence; sufficient evidence of carcinogenicity in animals);

Chemicals and groups of chemicals

Acrylonitrile
Benzo[*a*]anthracene
Benzidine-based dyes
Benzo[*a*]pyrene
Beryllium and beryllium compounds
Cadmium and cadmium compounds
p-Chlorotoluidine
Dibenz[*a*,*h*]anthracene
Diethyl sulfate
Dimethylcarbamoyl chloride
Dimethyl sulfate
Epichlorohydrin
Ethylenedibromide
Ethylene oxide
N-Ethyl-*N*-nitrosourea
Formaldehyde
4,4'-Methylene-bis(2-chloroaniline) (MOCA)
N-Methyl-*N*'-nitro-*N*-nitrosoguanidine (MNNG)
N-Methyl-*N*'-nitrosourea
N-Nitrosodiethylamine
N-Nitrosodimethylamine
Propylene oxide
Silica, crystalline
Styrene oxide
Tris(2,3-dibromopropyl)phosphate
Vinyl bromide

Mixtures

Creosotes

Diesel engine exhausts
Polychlorinated biphenyls
Drugs
Adriamycin
Anabolic steroids
Azacytidine
N,N'-bis(2-chloroethyl)-*N*-nitrosourea (BCNU)
Chloramphenicol
N-(2-chloroethyl)-*N'*-cyclohexyl-*N*-nitrosourea (CCNU)
Chlorototocin
Cisplatin 5-Methoxypsoralen
Nitrogen mustard
Phenacetin
Procarbazine hydrochloride
Tris(1-aziridinyl)phosphine sulfide (Thiotepa)
Biological factors
Clonorchis sinensis
Epstein Barr virus
Herpex simplex virus 2
Human immunodeficieny viruses
Human papilloma virus
Opisthorchis viverini
Schistosoma haematobium

Possibly carcinogenic to humans (Group 2B): sufficient evidence of carcinogenicity in animals, but no adequate data on cancer in exposed humans)
Not classifiable as to its carcinogenicity to humans (Group 3)
Probably not carcinogenic to humans (Group 4)

Other classification schemes by other organizations exist including those of the EPA, the EC, or the Chemical Manufacturers Association. These evaluations agree with respect to the classification of compounds that are known human carcinogens but place different emphasis on the results of animal or in vitro studies. These lists of agents with an established carcinogenic effect may help to avoid or at least reduce exposure to these substances. However, it cannot be assumed that substances not included in these lists are nonhazardous; it may simply be a lack of data on humans.

[Top of Page]

9. Forms of Exposure

9.1. Occupational Exposure

Historically, human occupational exposure to relatively high doses of chemical carcinogens has provided insight into the etiology of human cancer. The earliest observations were reported by POTT (1775) on an increased risk of scrotal cancer in chimney sweeps or by REHN (1897) who observed an elevated incidence of tumors in the urinary bladder among workers of the dye industry [8].

Whereas for high exposure levels at the workplace (asbestos, metals) often a clear epidemiological evidence exists for the carcinogenicity of certain substances, such causal relationships are difficult to verify for environmental (low level) exposures. In the developed countries, but less so in the developing or newly industrialized countries [9], safety precautions in the industry have been implemented by law and have resulted in a reduced levels of exposure and subsequently in lower incidences of occupational cancer.

9.2. Environmental Exposure

Many carcinogenic agents identified in an occupational setting can also be found in the general environment at lower levels. One example of these is asbestos. Whether the low levels of carcinogenic compounds in the general environment pose a significant cancer risk is a matter of intensive debate. It has been brought forward that the risk calculated from animal carcinogenicity data and environmental exposure levels is very low and may statistically effect only few people. On the other hand, individual differences in susceptibility and combination effects resulting from exposure to complex mixtures must be considered.

For genotoxic compounds the general assumption is that there is no threshold. For nongenotoxic carcinogenic agents biological considerations, that take into account the mechanisms of action, point towards the existence of a threshold dose. However, determination of threshold levels for chronic biological effects is a formidable task [1], [5].

9.3. Lifestyle

Perhaps the most common cause of human cancer is tobacco smoking and other forms of tobacco abuse (chewing, sniff). It has been estimated that over 90 % of lung cancer and significant proportions of cancers at other organ sites (urinary bladder, pancreas, gastrointestinal tract, upper respiratory passages) are attributable to excessive tobacco smoking. Overall, about one-third of all cancer deaths are believed to result from this habit [1], [5], [10]. Accordingly, it has been convincingly demonstrated that smoking cessation reduces the cancer risk significantly.

The carcinogenic effect of excessive intake of alcoholic beverages is not as readily apparent. However, the risk for cancer of the oral cavity or the larynx increases markedly when an individual smokes tobacco and abuses alcoholic beverages. Similarly, a synergistic effect has been observed for liver cancer when individuals infected with hepatitis B virus consume alcoholic beverages excessively [5].

Viral infections (e.g., HIV, HPV) are associated with promiscuity. Here, the use of protective measures upon sexual intercourse has been shown to reduce the risk of an infection (see Section Viruses and Other Biological Agents)

9.4. Diet

Dietary factors have been considered to be most important among the different contributors to human cancer [10], [11]. In the diet, carcinogenic agents can be found such as PAH and heterocyclic amines in grilled meat, carcinogenic plant products (safrol) or mycotoxins (see Section Natural Compounds). A high calorie diet and/or a diet high in fat (especially saturated animal fat) has been shown to increase tumor formation in experimental animals but epidemiological evidence is insufficient. There are also speculations on factors with anticarcinogenic activity such as antioxidants, vitamins, and fibers.

9.5. Drugs

Among the cytostatic drugs used in the chemotherapy of neoplasms several classes of compounds exist that have been shown to induce tumors in rodents or even lead to the induction of secondary tumors in treated patients. These include alkylating agents [mitomycin C, cyclophosphamide and nitrosourea derivatives (BCNU)], and anthracyclines (doxorubicin).

Genotoxic activity has been reported for hormone drugs (cyproterone acetate, tamoxifen) while a postulated increase in tumor incidences as a result of continued use of contraceptive drugs was not confirmed by epidemiological studies. Diethylstilbestrol, a synthetic estrogen used in the 1960s to prevent spontaneous abort, induced carcinoma in the children of treated patients.

Various drugs such as phenacetin that have been identified as carcinogenic agents have been withdrawn from the market. A number of herbal drugs contain carcinogenic components including aristolochic acid, anthraquinones, and pyrrolizidine alkaloids (see Section Natural Compounds).

9.6. Exposure Levels in Developing Countries

To some extent, industry is being forced to develop cleaner and safer processes and products in North America and Europe. However, in the rapidly industrializing countries of the Third World, corporate responsibility is not compelled by public awareness, regulation, and compensation laws. It has been pointed out not only by industry critics but also the International Labor Office and the United Nations Center on Transnational Corporations that global corporations are operating more polluting and dangerous plants in Third World countries [12]. Carcinogenic or otherwise toxic chemicals, e.g., pesticides such as DDT, banned for certain uses or voluntarily withdrawn from markets in the United States or Europe, are being produced and marketed in Africa, Asia, and Latin America. In small-scale industries that account for the bulk of industrial enterprises in developing and newly industrialized countries the standards of industrial hygiene, safety and, waste management are low [9].

With the industrialization the western lifestyle is adopted in these countries; thus, developing and newly industrialized countries are the fastest growing markets for the tobacco industry. It has been predicted that here the rates of lung cancer are going to increase in an exponential way in the years to come [13].

9.7. Safe Handling of Carcinogenic Agents

For laboratory handling of chemical carcinogens, the U.S. National Research Council has provided useful guidelines [14]. Because the toxicological properties of most laboratory chemicals are undetermined, virtually all operations should be performed in well-ventilated fume-cupboards. Care should be taken to avoid contact with the skin and to avoid ingestion even of traces. Additional precautions are to be taken for the handling of chemicals with high carcinogenic potency. National or federal legislation regulates the appropriate labeling of carcinogens, suspected carcinogens (see Table 1) and other hazardous chemicals; these should be handled only by trained persons.

[Top of Page]

References

1. L. Tomatis *et al.*: “Avoided and avoidable risks of cancer,” *Carcinogenesis* **18** (1997) 97 – 105. Links
2. T. Jacks, R. A. Weinberg: “Cell-cycle control and its watchman,” *Nature* **381** (1996) 643 – 644. Links
3. B. N. Ames, L. S. Gold: “Chemical Carcinogenesis: Too many rodent carcinogens,” *Proc. Natl. Acad. Sci. USA* **87** (1994) 7772 – 7776. Links
4. C. D. Klassen (ed.): *Casaret and Doull's Toxicology*, 5th ed., Pergamon Press, New York 1996.
5. L. Tomatis (ed.): *Cancer*, Causes, Occurrence and Control, IARC Scientific Publications 100, Lyon 1990.
6. G. Taubes: “Epidemiology faces its limits,” *Science* **269** (1995) 164 – 169. Links E. L. Wynder, *Am. J. Epidem.* **143** (1996) 747 – 749.
7. F. P. Perera: “Molecular epidemiology: Insights into cancer susceptibility, risk assessment, and prevention,” *J. Natl. Cancer Inst.* **88** (1996) 496 – 509. Links
8. H. Marquardt, S. G. Schäfer (eds.): *Lehrbuch der Toxikologie*, BI Wissenschaftsverlag, Mannheim 1994.
9. D. C. Christiani *et al.*: “Occupational health in developing countries: Review of the research needs,” *Am. J. Ind. Med.* **17** (1990) 393 – 401. Links
10. R. Doll: “Nature and nurture: possibilities for cancer control,” *Carcinogenesis* **17** (1996) 177 – 184. Links
11. World Cancer Research Fund: *Food, Nutrition and the Prevention of Cancer: a global perspective* American Institute for Cancer Research, Washington 1997.
12. B. I. Castleman: “The migration of industrial hazards,” *Int. J. Occup. Environ. Health* **1** (1995) 85 – 96.
13. S. J. Jay: “The emerging tobacco epidemic in China,” *J. Amer. Med. Ass.* **279** (1998) 1346 – 1347.
14. National Research Council: *Prudent Practices for Handling Hazardous Chemicals in Laboratories*. National Academy Press, Washington, DC, 1981, pp. 30 – 56.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Cardiac Glycosides and Synthetic Cardiotonic Drugs

Hans P. Albrecht, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Karl-Heinz Geiss, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Cardiac Glycosides 3
2.1. Chemical Structure 3
2.1.1. Cardenolides 5
2.1.2. Bufadienolides 7
2.2. Isolation, Chemical Synthesis 8
2.3. Pharmacology and Clinical Therapy 8
2.3.1. Pharmacological Investigation 8
2.3.2. Mode of Action 9
2.3.3. Pharmacokinetics 9
2.4. Clinically Important Cardiac Glycosides 10
2.4.1. Glycosides with Long Duration of Action 10
2.4.2. Glycosides with Medium Duration of Action 11
2.4.3. Glycosides with Short Duration of Action 11
3. Nonglycosidic Inotropic Agents . . . 12
3.1. Sympathomimetics 12
3.2. Phosphodiesterase Inhibitors 14
3.3. Agents with Direct Stimulation of Adenylate Cyclase 15
3.4. Agents Acting on the Calcium Channels 15
4. References 16

1. Introduction [1–5]

Blood circulation is effected by the rhythmic sequence of atrial and ventricular contraction (systole) and relaxation (diastole) of the heart. *Heart failure* is an abnormality in cardiac performance that prevents the heart from pumping blood at a rate commensurate with the metabolic and circulatory demands of the body.

Various compensatory mechanisms maintain a relatively normal cardiac output in the early stages of insufficiency; they are responsible for the symptoms of chronic congestive heart failure described in the following.

Sodium and water retention increase the blood volume and lead to elevated venous pressure, pulmonary vascular congestion, and peripheral edema. Increased adrenergic stimulation produces an increase in heart rate (tachycardia) and a constriction of blood vessels (vasoconstriction). In left-side heart failure, pulmonary capillary pressure increases. Edema and venous congestion increase the work of breathing, resulting in shortness of breath (dyspnea), fatigue, and weakness. If the capillary pressure exceeds a critical value, life-threatening acute pulmonary edema develops. Elevated systemic venous pressure in right-side heart failure is accompanied by peripheral congestion, resulting in ankle edema and enlargement of the liver.

Congestive heart failure develops secondary to an increase (1) in cardiac ventricular pressure and (2) in the size of cardiac ventricles. This occurs if the heart must cope for an extended period of time with an abnormally high demand upon cardiac contractility. This demand may be caused by cardiovascular disease (hypertension, myocardial ischemia), cardiomyopathy, or congenital heart disease. Knowledge of the cause of heart failure at the cellular and molecular level is based on growing information on the contractile proteins of the heart muscle. Hypotheses focus on disturbances in mitochondrial energy production and conversion of chemical energy into mechanical work in the contraction process.

Congestive heart failure affects about five million people in the United States. The New York Heart Association has developed a classifi-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_271

cation of patients on the basis of the relationship between symptoms and the amount of physical effort required to provoke them:

Class I. No limitations: ordinary physical activity does not cause undue fatigue, dyspnea, or palpitation.
Class II. Slight limitation of physical activity: patients are comfortable at rest. Ordinary physical activity results in fatigue, palpitation, dyspnea, or angina.
Class III. Marked limitation of physical activity: although patients are comfortable at rest, less than ordinary activity will lead to symptoms.
Class IV. Inability to carry on any physical activity without discomfort: symptoms of congestive heart failure are present even at rest. With any physical activity increased discomfort is experienced.

Therapy [6] should be aimed primarily at the determination and possibly the elimination of the underlying cause, e.g., by surgical correction of valvular lesions, alleviation of infection or arrhythmia, or chemotherapeutic treatment of hypertension. Treatment of heart failure usually starts with simple measures; these include decreasing the demands on the heart by reducing physical activities and restricting the intake of sodium. Drug treatment of congestive heart failure must increase cardiac output and relieve pulmonary congestion and peripheral edema. This may be achieved by the following:

1) treatment with diuretics to reduce volume overload (→ Antihypertensives)
2) stimulation of myocardial contractility with cardiotonic agents, particularly with cardiac glycosides
3) adjustment of peripheral circulation with peripheral vasodilators (→ Antihypertensives)

Drugs that inhibit the angiotensin converting enzyme (ACE inhibitors) are the vasodilators of choice for the treatment of congestive heart failure.

Positive Inotropic Agents. Drugs that stimulate myocardial contractility are called positive inotropic agents, or cardiotonic agents. The oldest and still most important drugs for the therapy of congestive heart failure are the cardiac glycosides.

Such classical sympathomimetics as epinephrine are not used for the treatment of chronic heart failure; their inotropic activity is high but only transient, and it is combined with an increase in heart frequency, vasoconstriction, and an excessive myocardial oxygen demand.

Since the 1970s new positive inotropic agents including orally active sympathomimetic agents and a group of drugs that act predominantly by increasing cyclic adenosine monophosphate (cAMP) – either as inhibitors of phosphodiesterase type III (amrinone, milrinone) or by stimulating cAMP formation (forskolin) – have been investigated thoroughly. Although some of these compounds are used for the treatment of acute heart failure, their utility in the long-term treatment of congestive heart failure could not be established.

Mode of Action. All hypotheses concerning the mode of action of positive inotropic agents are based on the assumption that these drugs affect contractility by raising the intracellular Ca^{2+} concentration. This effect is best understood by considering the structural elements of the heart that are responsible for contractility [7–9]. The fundamental contractile units of heart muscle cells – the longitudinally arranged myofibrils – are made up of the contractile proteins *actin* and *myosin* and the regulatory proteins *tropomyosin* and *troponin*. Contraction and relaxation are effected by sliding the thin filaments, which contain actin aggregates, with respect to the thick filaments of myosin chains (sliding filaments theory). In the relaxed state troponin and tropomyosin inhibit the interaction of actin and myosin. Initiation of the contracting process occurs after binding of Ca^{2+} to troponin and subsequent change of the position (conformation) of the troponin–tropomyosin complex to allow for the interaction of actin and myosin.

This hypothesis accounts for the well-known importance of Ca^{2+} in the contraction process in heart muscle cells. Rapid increase in intracellular Ca^{2+} concentration from 10^{-7} mol/L during relaxation to 10^{-5} mol/L during contraction is regulated by the action potential of cardiac cells. Increase in available intracellular Ca^{2+} is triggered by a slow influx of Ca^{2+} during the plateau phase of the myocardial action potential. Together with an increase in intracellular Na^{+}, this causes the release of additional Ca^{2+} from

internal stores. Because of the importance of free Ca^{2+} for excitation–contraction coupling, the regulation of the intracellular concentration of Ca^{2+} explains all phases of the mechanism of action of cardiac glycosides and other inotropic agents.

Despite their different modes of action, all inotropic agents act by increasing the force of contraction of the heart to produce a higher cardiac output. Figure 1 shows ventricular function curves in healthy humans and in those with congestive heart failure. In the second group a normal cardiac output *N* can only be achieved at the expense of an elevated end diastolic pressure (Frank–Starling mechanism). Treatment with cardiac glycosides or other inotropic agents shifts the curve of the failing heart upward toward normal so that the required output *N* can be delivered at a lower end diastolic pressure and without the symptoms of congestive heart failure.

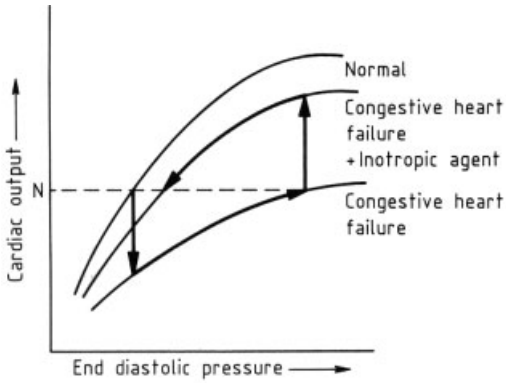

Figure 1. Ventricular function curves of the normal heart and the failing heart before and after treatment with cardiac glycosides according to ref. [10]

2. Cardiac Glycosides

Cardiac glycosides (synonyms: digitalis, cardiac steroids, cardiac poisons) are closely related natural products, found in a variety of plants [11], [12]. Some of them have been used as poisons and drugs since ancient times. The systematic use of cardiac glycoside preparations for the therapy of cardiac insufficiency and associated edema dates back to 1785, when W. WITHERING published his paper "*An Account of the Foxglove and Its Medical Uses.*" Of the more than 400 different cardiac glycosides that have been isolated from natural sources, only about 10 are used in therapy.

The pharmacology of cardiac glycosides involves a complex set of direct and indirect effects [13–16]. Their most important effect is to enhance the force and velocity of myocardial contraction (*positive inotropic effect*). In addition, cardiac glycosides decrease the rate of the failing heart (*negative chronotropic effect*); since ventricular tachycardia is common during heart failure, this effect is of great value in practical therapy. Digitalis depresses atrioventricular nodal conductivity (*negative dromotropic effect*), but it may also increase myocardial excitability and automaticity (*positive bathmotropic effect*), leading to the development of arrhythmias.

Because of their very narrow therapeutic range, therapy with cardiac glycosides is difficult. In some patients toxic symptoms are observed at doses that are required to compensate the failing heart at least partially. The toxicity of cardiac glycosides comprises both cardial and extracardial effects. Cardiac toxicity produces arrhythmias, leading in severe cases to ventricular fibrillations with subsequent systolic arrest and death. Extracardial symptoms include gastrointestinal disorders, mental changes, and visual disturbances.

Despite the introduction of new therapeutic strategies in the treatment of congestive heart failure and some uncertainty concerning their risk – benefit ratio and their long-term effects on survival [17], cardiac glycosides are still widely used, particularly in the elderly. In the United States digoxin (Lanoxin) belongs to the ten most frequently prescribed drugs [18].

2.1. Chemical Structure [11], [19–24]

The structures of cardiac glycosides are represented by one of the general formulas (**1**) to (**3**), in which a steroid (aglycone) is linked through an alcoholic hydroxyl group with the hemiacetal form of a mono-, di-, tri-, or tetrasaccharide:

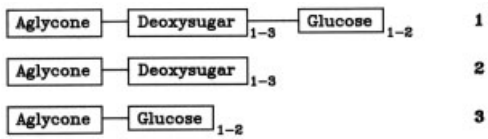

Aglycones. Cardioactivity resides in the steroidal aglycone, also called the genin. The saccharide moiety modifies potency and contributes mainly to the pharmacokinetic properties. Characteristic structural features of cardioactive steroids are the cis fusion of the rings A/B and C/D, a tertiary hydroxyl group at C-14, and an axially oriented hydroxyl group at C-3 which mediates the attachment of the sugar residue. An unsaturated lactone ring is joined to the steroid nucleus at C-17 (see structure of digitoxigenin).

Steroids carrying an α,β-unsaturated five-membered lactone ring (butenolide ring) and their corresponding glycosides are called *cardenolides.* The best known representative of this type is *digitoxigenin*:

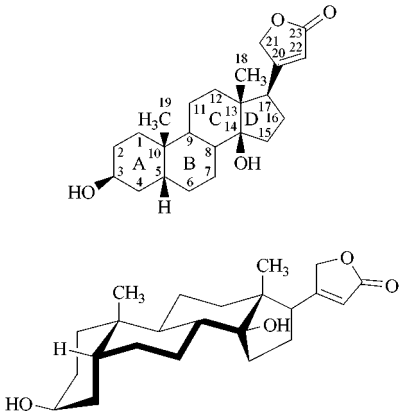

A six-membered, doubly unsaturated α-pyrone ring is characteristic of the *bufadienolides,* e.g., bufalin (3,14-dihydroxybufa-20,22-dienolide):

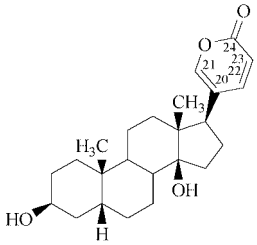

All naturally occurring cardioactive steroids are derived from these two basic structures. Additional hydroxyl groups may be at C-1, C-5, C-11, C-12, and C-16. Replacement of the C-10 methyl group by hydroxymethyl or an aldehyde function is found in certain glycosides of the strophanthus series. Another variation is the introduction of a double bond between C-4 and C-5. A survey of the most important aglycones from which essentially all naturally occurring glycosides are derived is given in Table 1.

Sugars of Cardiac Glycosides. [25]. Sugars include D-glucose, L-rhamnose, and a number of rare 6-deoxy- and 2,6-dideoxy sugars and their 3-*O*-methyl derivatives. The sugars of the D series are linked to the genin by a β-glycosidic bond and adopt the C 1 conformation; the L sugars occur as α-glycosides in the 1 C conformation (Table 2) [26], [27].

Table 2. Sugars of cardiac glycosides. d-Sugars are represented in the C 1 conformation of their β-d-hexopyranose form, l-sugars in the 1 C conformation of their α-l-hexopyranose form

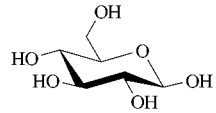

β-D-Glucose

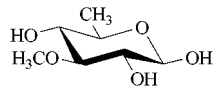

β-D-Thevetose (6-deoxy-3-*O*-methyl-β-D-glucose)

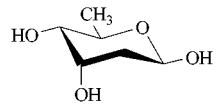

β-D-Digitoxose (2,6-dideoxy-β-D-ribo-hexose)

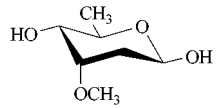

β-D-Cymarose (2,6-dideoxy-3-*O*-methyl-β-D-ribo-hexose)

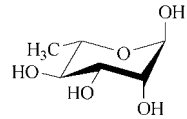

α-D-Rhamnose

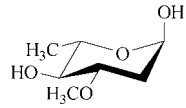

α-L-Oleandrose 2,6-dideoxy-3-*O*-methyl-α-L-arabino-hexose)

In cardiac glycosides derived from oligosaccharides, the sugars are joined either through a $1 \rightarrow 4$ or a $1 \rightarrow 6$ glycosidic linkage. Representative examples of the different structures of naturally occurring cardiac glycosides are digitoxin, *k*-strophanthoside, and scillaren (see Table 3).

The natural glycosides have one or two molecules of glucose attached to the end of the saccharide chain. They are called *genuine* or *primary glycosides.* These glucose residues are easily split off by specific glucosidases, which occur in plant sources of cardiac glycosides, to give the *secondary glycosides.* This enzymatic conversion may take place during prolonged storage

Table 1. Aglycones of cardiac glycosides

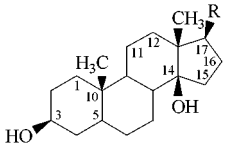

Cardenolides: R =

Bufadienolides: R = 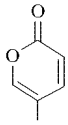

Aglycone	CAS reg. No.	Substituents						
		C-1	C-4	C-5	C-10	C-11	C-12	C-16
Cardenolides								
Digitoxigenin	[*143-62-4*]	H	H	β-H	CH_3	H	H	H
Acovenosigenin	[*639-15-6*]	β-OH	H	β-H	CH_3	H	H	H
Digoxigenin	[*1672-46-6*]	H	H	β-H	CH_3	H	β-OH	H
Gitoxigenin	[*545-26-6*]	H	H	β-H	CH_3	H	H	β-OH
Oleandrigenin	[*465-15-6*]	H	H	β-H	CH_3	H	H	β-OAc
Uzarigenin	[*466-09-1*]	H	H	α-H	CH_3	H	H	H
Corotoxigenin	[*468-20-2*]	H	H	α-H	CHO	H	H	H
Canarigenin	[*1778-88-7*]	H			CH_3	H	H	H
Cannogenin	[*510-63-4*]	H	H	β-H	CHO	H	H	H
Periplogenin	[*514-39-6*]	H	H	β-OH	CH_3	H	H	H
Strophanthidin	[*66-28-4*]	H	H	β-OH	CHO	H	H	H
Strophanthidol	[*560-54-3*]	H	H	β-OH	CH_2OH	H	H	H
Ouabagenin	[*508-52-1*]	β-OH	H	β-OH	CH_2OH	α-OH	H	H
Bufadienolides								
Bufalin	[*465-21-4*]	H	H	β-H	CH_3	H	H	H
Hellebrigenin	[*465-90-7*]	H	H	β-OH	CHO	H	H	H
Scillarenin	[*465-22-5*]	H			CH_3	H	H	H
Scilliglaucosidin	[*510-62-3*]				CHO	H	H	H

or in the course of the extraction procedures unless care is taken to inactivate these enzymes.

2.1.1. Cardenolides

Digitalis Glycosides. The leaves and seeds of the purple foxglove (*Digitalis purpurea*) and the woolly foxglove (*Digitalis lanata*) from the Scrophulariaceae family are the most important sources of cardiac steroids. The dried leaves contain up to 0.3 % (*D. purpurea*) or 1 % (*D. lanata*) of a complex mixture of cardioactive glycosides.

The main genuine glycosides of *D. lanata* are lanatosides A, B, and C, which are derived from the aglycones digitoxigenin, gitoxigenin, and digoxigenin, respectively. Their oligosaccharide part is made up of three molecules of digitoxose and one glucose at the end of the chain. The 3-hydroxyl group of the third digitoxose (3‴-position) is acetylated.

The main glycosides of *D. purpurea* are purpurea glycosides A and B, together with the 16-formyl derivative of purpurea glycoside B. Their structures differ from those of the lanatosides only by the absence of the acetyl group. Remarkably, a glycoside derived from digoxigenin, corresponding to deacetyl-lanatoside C, is not present in the purple foxglove. Some reactions and interconversions of digitalis glycosides are shown in Figure 2.

Table 3. Representative examples of naturally occurring glycosides

Digitoxin

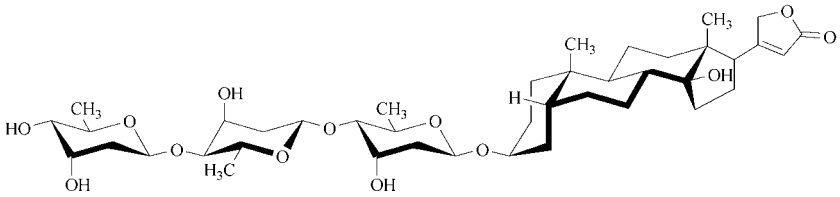

k-Strophanthoside

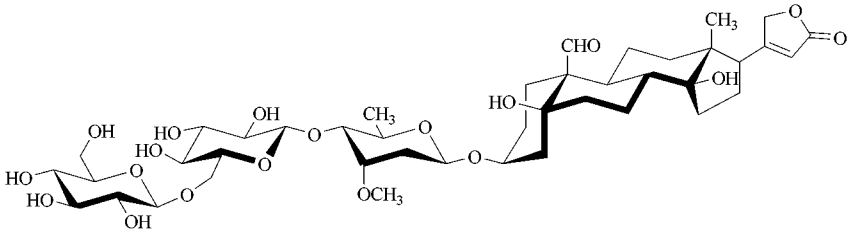

Scillaren

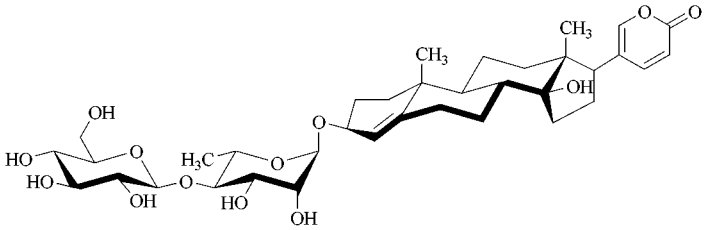

Structurally related to the digitalis glycosides is oleandrin [*465-16-7*], a gitoxigenin derivative from the white oleander (*Nerium oleander*):

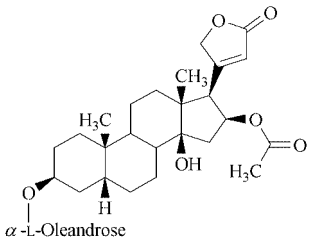

Oleandrin

The seeds of the South American tree *Thevetia neriifolia* (yellow oleander) are an especially rich source of glycosides.

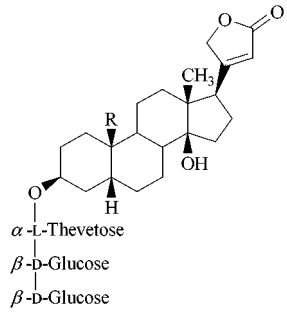

Thevetin A: R = CHO
Thevetin B: R = CH_3

In addition to the main glycoside thevetin B [*27127-79-3*], the seeds contain thevetin A [*37933-66-7*] which gives peruvoside upon enzymatic removal of glucose.

Strophanthus Glycosides. Extracts from the seeds and wood of trees of the genus *Strophanthus,* growing in tropical regions of Africa and Asia, contain powerful cardiac glycosides and have long been used as arrow poisons. Ouabain (*g*-strophanthin) [*11018-89-6*] is readily isolated from the seeds of *S. gratus* in 3 – 4 % yield.

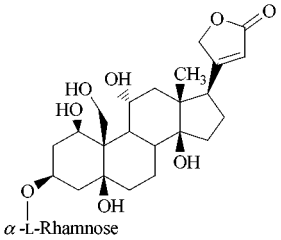

Ouabain

The genuine glycoside from *Strophanthus kombé* is *k*-strophanthoside [*33279-57-1*]. Stepwise enzymatic degradation with glucosidase and strophanthobiase leads to *k*-strophanthin-β [*560-53-2*] and then to cymarin [*508-77-0*]:

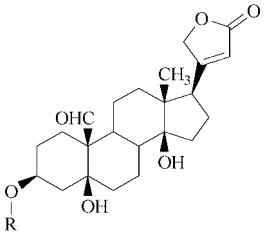

k-Strophantoside: R = β-D-cymarose-(β-D-glucose)$_2$
k-Strophantoside-β: R = β-D-cymarose-β-D-glucose
Cymarine: R = β-D-cymarose

Also derived from the same aglycone and structurally related to the strophanthus glycosides is convallatoxin [*508-75-8*], a glycoside found in the leaves and blossoms of lily of the valley (*Convallaria majalis*):

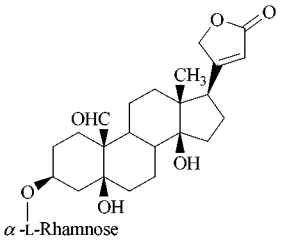

Convallatoxin

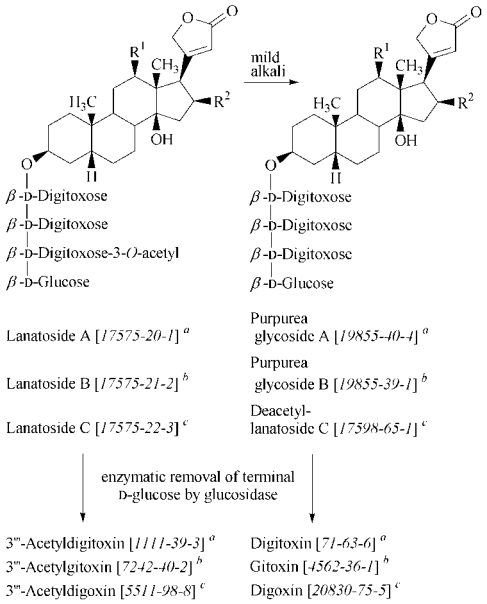

Figure 2. Main glycosides from *Digitalis lanata* and *Digitalis purpurea*
[a]R^1, R^2 = H; [b]R^1 = H, R^2 = OH; [c]R^1 = OH, R^2 = H

2.1.2. Bufadienolides

Bufadienolides are found as glycosides in some plants of the Liliaceae and Ranunculaceae families. Related to these glycosides are the toad poisons, which occur as the free genins or are combined with suberylarginine. The most important source of bufadienolide glycosides is the white squill (*Scilla maritima, Urginea maritima*) from the Mediterranean region and India. The bulb of the white squill contains up to 0.3 % of a mixture of cardioactive components, principally scillaren [*124-99-2*]. Proscillaridin [*466-06-8*] is obtained upon enzymatic hydrolysis of this mixture with scillarenase:

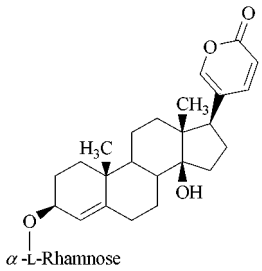

Proscillaridin

Scilliroside, the main component of the red squill, is not used therapeutically, but is worth mentioning because of its high toxicity to rats.

The roots of the Christmas rose (*Helleborus niger*) yield glycosides derived from the aglycone hellebrigenin [*465-90-7*].

2.2. Isolation, Chemical Synthesis

Efficient isolation procedures, structural elucidations, and total syntheses are due to pioneering work on the chemistry of cardioactive steroids in the groups of A. Windaus, W. A. Jacobs, R. C. Elderfield, H. Wieland, T. Reichstein, A. Stoll, R. Tschesche, K. Meyer, F. Sondheimer, H. Ruschig, D. Satoh, G. R. Pettit, K. Wiesner, and G. Stork.

Isolation is tedious because cardiac glycosides occur in plants only in small amounts and as complex mixtures. In addition, the crude plant material usually contains enzymes that split off glucose from the native glycosides, leading to partial degradation. Inhibition of these enzymes with ammonium salts is essential to isolate the genuine glycosides.

For the isolation of the therapeutically important secondary glycosides, the finely ground material is moisturized and exposed to the enzymes at 30 – 37 °C until glucose is completely removed. Extraction procedures, usually followed by precipitation of tannic acid and related phenolic products with lead salts, afford a crude mixture of cardioactive compounds, which is further purified by chromatography and/or crystallization. For representative examples of isolation procedures, see Section 2.4.

Formerly, mixtures of glycosides or crude plant extracts were used in therapy; these have been replaced by chemically pure drugs today.

Total syntheses of cardiac steroids and their corresponding glycosides have been accomplished but are not used commercially. Chemical modification of the sugar part of natural glycosides has led to derivatives with improved pharmacokinetic properties, in particular better oral absorption. Efforts to achieve better differentiation of cardiotonic and cardiotoxic properties, that is, to obtain compounds with a wide safety margin, by chemical modification have not been successful.

2.3. Pharmacology and Clinical Therapy

2.3.1. Pharmacological Investigation [13], [28–30]

Experimental models for evaluating cardiotonic effects include the papillary muscles of cats or guinea pigs and isolated organs, such as the guinea pig auricle [32], [33]. These preparations do not contract spontaneously, but are stimulated through electrodes at a certain frequency. Contractility is determined by measuring the isometric contraction after the addition of cumulatively increasing doses of cardiac glycosides to the medium.

Other models are the perfused heart preparation of Langendorff and the heart–lung preparation [33]. These models resemble the intact animal more closely and provide more detailed information about therapeutic and toxic parameters as well as oxygen consumption in the presence and absence of cardiac glycosides.

Cardiotonic effects of glycosides in cultures of heart cells were observed as early as 1955. Since improved techniques for the observation and registration of single cells have become available, this method is gaining attention.

Inotropic effects in the intact, anesthetized animal (cat or dog) are evaluated under continued infusion by measuring *intracardiac pressure.* The maximum rate of pressure rise (dp/dt_{max}) in the left ventricle serves as an index for inotropy (Fig. 3). By monitoring the electrocardiogram one can determine both inotropy and the dose at which arrhythmias occur. Continued infusion finally leads to cardiac arrest (lethal dose). By measuring these parameters, the *therapeutic range,* that is the ratio between the dose that produces the therapeutic effect and the dose that causes toxic side effects, can be determined [34].

The lethal dose (LD) is determined by intravenous infusion until cardiac arrest either in anesthetized cats (method of Hatcher) or in anesthetized guinea pigs (methods of Knaffl-Lenz); this parameter has formerly been used for the assessment of cardiac glycosides. Currently, toxic parameters (Table 4) are employed mainly for the determination of intestinal absorption and duration of action. A common procedure for measuring intestinal absorption is

to administer a sublethal dose of the glycoside orally or intraduodenally and then to determine the supplementary intravenous dose needed to produce cardiac arrest. Duration of action can be determined in a similar way.

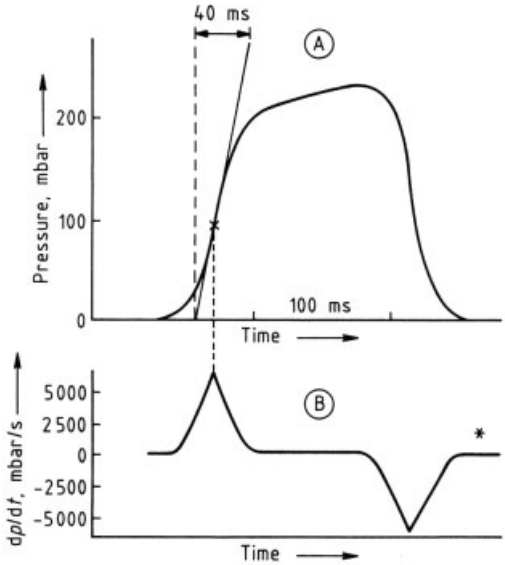

Figure 3. Schematic representation of the left ventricular pressure (A) and its derivative (B)
The maximum rate of pressure rise dp/dt_{max} in this case is 266 mbar/40 ms = 6650 mbar/s [31]

Table 4. Cardiotoxicity of some glycosides (intravenous administration, cat)

Glycoside	Lethal dose, mg/kg
Digitoxin [*71-63-6*]	0.32 – 0.42
Gitoxin [*4562-36-1*]	0.59 – 0.80
Oleandrin [*465-16-7*]	0.20
Digoxin [*20830-75-5*]	0.23 – 0.28
3‴-Acetyldigoxin [*5511-98-8*]	0.36 – 0.47
4‴-Acetyldigoxin [*5355-48-6*]	0.37 – 0.43
4‴-Methyldigoxin [*30685-43-9*]	0.33 – 0.38
k-Strophanthoside [*33279-57-1*]	0.13 – 0.19
Convallatoxin [*508-75-8*]	0.07 – 0.08
Cymarin [*508-77-0*]	0.11 – 0.13
Peruvoside [*1182-87-2*]	0.15
Ouabain [*11018-89-6*]	0.09 – 0.13
Proscillaridin [*466-06-8*]	0.16 – 0.18

2.3.2. Mode of Action [14], [35], [36]

The membrane-bound Na^+- and K^+-dependent adenosine triphosphatase (Na^+, K^+-ATPase) has been identified as the site of action of digitalis at the molecular level. Under physiological conditions this enzyme hydrolyzes adenosine triphosphate (ATP) to achieve the exchange of intracellular Na^+ for extracellular K^+ against concentration gradients. Cardiac glycosides bind to a specific receptor site of the enzyme at the external surface of the membrane-bound enzyme. Inhibition of the active transport system by glycosides leads to an increase in intracellular Na^+ and a decrease in intracellular K^+. The accumulation of Na^+ in the heart muscle cell is linked to an increased influx of Ca^{2+}, which is regulated by the Na^+/Ca^{2+} carrier system. As a consequence more Ca^{2+} is available for interaction with the myofibrils.

According to this hypothesis, supported by numerous investigators [37] inhibition of the enzyme must be basic to the inotropic action of cardiac glycosides. At higher digitalis concentrations the pump capacity is lowered and may eventually fall below the limit required to maintain ionic homeostasis. The resulting persistent increase in intracellular Na^+, combined with loss of K^+ and heavy uptake of Ca^{2+} from external sources, causes the symptoms of cardiac glycoside intoxication.

There is no contention that Na^+, K^+-ATPase is the binding site for glycosides or that inhibition of this enzyme is responsible for their toxicity; however, it is still controversial whether pump inhibition is causally related to the development of inotropy. Opponents of the Na^+, K^+-ATPase hypothesis stress the enormous difference between the concentration of $10^{-6} - 10^{-7}$ mol/L that is needed to produce an inotropic effect in isolated organs and the 100 – 1000 times lower therapeutically effective concentration of cardiac glycosides in humans. Assumption of significant contributions of extracardiac, neuronal effects that arise, e.g., from inhibition of neuronal Na^+, K^+-ATPases might possibly explain this discrepancy.

2.3.3. Pharmacokinetics [30], [38], [39]

Because of the narrow therapeutic range of cardiac glycosides, knowledge of their pharmacokinetic behavior is essential for safe therapy. Cardiac glycosides are very similar in their pharmacodynamic actions, but they vary considerably in their pharmacokinetic properties, according to their different lipophilicities. Such differences include absorption after oral application, onset of action, plasma protein binding, duration of action, rate of elimination, and mode of excretion.

Absorption. Treatment of congestive heart failure apart from emergencies requires continued medication; therefore, only orally active glycosides are used in therapy. Gastrointestinal absorption takes place predominantly in the small intestine and varies among different glycosides. Important parameters for absorption are (1) water solubility as a prerequisite for dissolution in the gastrointestinal tract and (2) high lipophilicity to facilitate rapid diffusion through the membrane of the intestinal mucosal wall. Polar molecules such as ouabain are very poorly absorbed and can only be administered intravenously, whereas the nonpolar digitoxin is taken up from the gastrointestinal tract nearly quantitatively.

The *absorption ratio* has been defined as the therapeutically effective fraction of the total amount of orally given drug. Variations in the rate and extent of dissolution of different commercially available digoxin tablets have been found to be responsible for widely differing serum levels of the drug (for bioavailability, see [40]). With the development of improved pharmaceutical preparations and of bioavailability specifications, the problem of variable absorption ratios has been solved.

The extent of *plasma protein binding* in the blood depends on the hydrophobicity of the glycoside. Nonpolar digitoxin is bound to the extent of more than 90 %, digoxin is bound to about 30 %, and ouabain shows less than 10 % binding.

Metabolism. Biotransformation also differs among the various glycosides according to their lipid solubilities. Hydrophobic digitoxin is metabolized extensively in the liver, but digoxin is only slightly degraded (about 10 %), and the polar ouabain is excreted unchanged. Metabolism of digitoxin in the liver occurs by successive cleavage of the three digitoxose molecules and is accompanied by hydroxylation at C-12 of the steroid. The very slow removal of the terminal digitoxose molecule accounts for the long duration of action.

Duration of Action. Digitalis glycosides and their metabolites are excreted mainly via the kidney. As a consequence elimination is slower in patients with renal insufficiency. This effect is particularly seen with digoxin and its semisynthetic derivatives. In contrast, excretion of digitoxin and proscillaridin is independent of renal function.

For practical therapy the duration of action is of great importance. The *decay ratio* defines the amount of the total body pool in percent that in the course of 24 h loses its effect on the heart because of inactivation and/or elimination. Alternatively, the *half-life* (the time required for a given blood level of the drug to decrease to 50 %) is commonly used to specify the duration of action of a particular glycoside.

The *oral maintenance dose* is the amount of glycoside that must be given daily to maintain a steady therapeutic level.

2.4. Clinically Important Cardiac Glycosides [41–43]

2.4.1. Glycosides with Long Duration of Action

Cardiac glycosides derived from digitoxigenin are characterized by a long duration of action, almost complete absorption from the gastrointestinal tract, and a very slow onset of action. Lanatoside A and acetyldigitoxin are both easily metabolized to digitoxin; they do not exhibit any advantages over digitoxin and are no longer important in therapy.

Digitoxin [*71-63-6*], 3β-[(*O*-2,6-dideoxy-β-D-ribo-hexopyranosyl-(1 $\rightarrow$ 4)-*O*-2,6-dideoxy-β-D-ribo-hexopyranosyl-(1 $\rightarrow$ 4)-2,6-dideoxy-β-D-ribo-hexopyranosyl)oxy]-14-hydroxy-4β,14β-card-20(22)-enolide (structure 3), $C_{41}H_{64}O_{13}$, M_r 764.92, *mp* 252 °C, $[\alpha]_D^{20} = +17°$ (chloroform : methanol 1 : 1), is isolated from *Digitalis purpurea* and purified [44–46]. Digitoxin is useful for maintenance therapy because its long half-life (5 – 9 days) provides a sustained therapeutic effect even if a dose is missed. For the same reason toxic reactions are not easy to manage. Elimination is independent of renal function.

Absorption ratio 90 – 100 %, decay ratio 7 %, effective dose level 2 mg, maintenance dose 0.15 mg.

Trade names: Crystodigin (Lilly), Digimerck (Merck).

Gitoxin, which can also be obtained from *Digitalis purpurea* and *D. lanata,* is not bioavailable because of its extreme insolubility in water and organic solvents. Prodrugs, such as pentaacetylgitoxin [47] and pentaformylgitoxin [48], are readily metabolized in humans to the cardioactive 16-acyl derivatives.

Pengitoxin [*7242-04-8*], pentaacetylgitoxin, $C_{51}H_{74}O_{19}$, M_r 991.13, *mp* 151 – 155 °C, $[\alpha]_D^{20} = +14°$ (pyridine); for synthesis see [49]. Absorption ratio 60 – 65 %, decay ratio 10 %, effective dose level 2.5 – 3 mg, maintenance dose 0.4 mg. Pengitoxin is comparable to digitoxin but is absorbed to a smaller extent.

Trade names: Pentagit (ASTA Medica AWD).

2.4.2. Glycosides with Medium Duration of Action

Digoxin and its derivatives are therapeutically the most useful and most widely used digitalis glycosides. Their pharmacokinetic properties are intermediate between the nonpolar, orally well-absorbed digitoxin with its long half-life and the polar, short-acting strophanthus glycosides.

Digoxin [*20830-75-5*] (structure Fig. 2), $C_{41}H_{64}O_{14}$, M_r 780.92, *mp* 265 °C, $[\alpha]_D^{20}$ +13.3° (pyridine), is isolated from *Digitalis lanata* [50]. Absorption ratio 70 %, decay ratio 20 %, effective dose level 2 mg, maintenance dose 0.5 mg.

Digoxin is regarded by most as the drug of choice for the treatment of congestive heart failure. The reasons are the very rapid onset of action (5 – 30 min after intravenous and 2 – 3 h after oral administration), the good absorption ratio, and, above all, the medium half-life which makes toxic reactions easier to manage. Digoxin can be administered either intravenously or orally, and the maintenance dose is thus easily established after emergency intravenous use. Digoxin is excreted mainly in the urine; in patients with impaired renal function, the half-life is prolonged.

Trade names: Digacin (Beiersdorf), Lanicor (Roche), Lanoxin (Glaxo Wellcome).

To improve the oral absorption of digoxin, a series of digoxin derivatives have been synthesized by acylation or methylation of the terminal digitoxose residue.

α-Acetyldigoxin [*5511-98-8*], 3‴-acetyldigoxin, $C_{43}H_{66}O_{15}$, M_r 823.0, *mp* 230 °C (decomp.), $[\alpha]_D^{20} = +21°$ (pyridine), is prepared from lanatoside C (Fig. 2) by enzymatic hydrolysis of the D-glucose unit [51] or from digoxin by reaction with triethylorthoacetate and subsequent partial hydrolysis of the orthoacetate [52]. Absorption ratio 80 %, decay ratio 20 %, maintenance dose 0.4 mg.

β-Acetyldigoxin [*5355-48-6*], 4‴-acetyldigoxin, $C_{43}H_{66}O_{15}$, M_r 823.0, *mp* 240 °C (decomp.), $[\alpha]_D^{20} = +31.5°$ (pyridine), is synthesized by selective acetylation using acetic anhydride in pyridine [53].

Trade names: Novodigal (Beiersdorf).

Medigoxin [*30685-43-9*], metildigoxin, β-methyldigoxin, 4‴-methyldigoxin, 16′-methyldigoxin, $C_{42}H_{66}O_{14}$, M_r 794.97, *mp* 226 – 229 °C, is made from digoxin by methylation with dimethyl sulfate in the presence of barium hydroxide and aluminum oxide [54]. Absorption ratio 95 %, decay ratio 22 %, maintenance dose 0.2 – 0.3 mg. Medigoxin is characterized by nearly complete absorption and a very rapid onset of action (5 – 20 min after oral administration).

Trade names: Lanitop (Roche).

2.4.3. Glycosides with Short Duration of Action

The strophanthus glycosides are characterized by a rapid onset and a very short duration of action. When toxic symptoms arise, they are relatively transient. Because of their high polarity, they are practically not absorbed when administered orally. Currently they are used only for intravenous therapy. The squill glycoside proscillaridin and its semisynthetic derivative meproscillarin combine the assets of the short duration of action of ouabain with satisfactory (proscillaridin) to good (meproscillarin) absorption. Cymarin is also orally active; it is derived from *k*-strophanthoside by enzymatic hydrolysis of two glucose units.

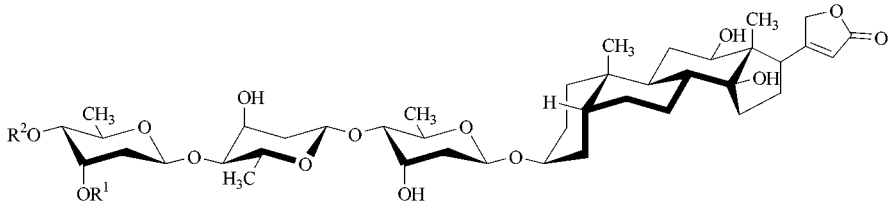

Digoxin: R^1, $R^2 = H$
3‴-Acetyldigoxin: $R^1 = COCH_3$, $R^2 = H$
4‴-Acetyldigoxin: $R^1 = H$, $R^2 = COCH_3$,
4‴-Methyldigoxin: $R^1 = H$, $R^2 = CH_3$,

Ouabain [*11018-89-6*], *g*-strophanthin, 3β-[(6-deoxy-α-L-mannopyranosyl)oxy]-1β,5,11α,14,19-pentahydroxy-5β,14β-card-20(22)-enolide octahydrate, $C_{29}H_{44}O_{12} \cdot 8\,H_2O$, M_r 728.78, *mp* 198 – 202 °C (decomp.), $[\alpha]_D^{20} = -30$ to $-33°$ (water); for its isolation, see [55]. Decay ratio 40 %, maintenance dose 0.25 mg (intravenous). Ouabain is used when rapid therapeutic benefit is required, especially in acute congestive heart failure.

Trade names: Ouabain-Nativelle (Nativelle, Wilcox), Strodival (Herbert).

Cymarin [*508-77-0*], *k*-strophanthin-α, $C_{30}H_{44}O_9$, M_r 548.65, *mp* 148 °C, $[\alpha]_D^{20} = +39°$ (methanol), is isolated from the seeds of *Strophanthus kombé* after enzymatic digestion [56]. Absorption ratio 33 %, decay ratio 40 %, maintenance dose 2.8 mg.

Proscillaridin [*466-06-8*], $C_{30}H_{42}O_8$, M_r 530.64, *mp* 227 – 230 °C, $[\alpha]_D^{20} = -93.5°$ (aqueous methanol), is isolated from *Scilla maritima* [57]. Absorption ratio 30 – 35 %, decay ratio 40 – 50 %, maintenance dose 1 – 1.5 mg. Because of its short duration of action, toxic symptoms after overdosage disappear within one day. Elimination is independent of renal function.

Trade names: Talusin (Knoll).

Meproscillarin [*33396-37-1*], proscillaridin-4′-methyl ether, $C_{31}H_{44}O_8$, M_r 544.66, *mp* 213 – 217 °C, $[\alpha]_D^{20} = -94°$ (methanol); for its synthesis, see ref. [58]. Absorption ratio 80 – 90 %, decay ratio 40 %, maintenance dose 0.5 – 0.75 mg. Meproscillarin and its metabolites are mainly eliminated by bilary secretion. The rapid elimination is not influenced by renal function. The bioavailability of meproscillarin is substantially improved compared to that of proscillaridin.

Trade names: Clift (Knoll).

3. Nonglycosidic Inotropic Agents

Because of the narrow therapeutic range and the side effects of digitalis glycosides, an intensive worldwide search for new positive inotropic agents that act by other mechanisms occurred over the past three decades. A hetereogeneous group of positive inotropic drugs were evaluated. Although these substances demonstrated improved hemodynamics and an improvement of symptoms in the short-term treatment of heart failure [59], [60], [61], [62], their use for long-term treatment has not been advocated due to a high incidence of adverse events and increased mortality that were observed in large clinical trials [63].

3.1. Sympathomimetics

Stimulation of β_1-adrenoreceptors, located at the extracellular side of the sarcolemma, results in increased production of cyclic adenosine monophosphate (cAMP) by stimulation of adenylate cyclase. Cyclic adenosine monophosphate then activates the slow inward Ca^{2+} channels, increasing the force of contraction, and accelerates the uptake of Ca^{2+} by the sarcoplasmic reticulum, decreasing the duration of contraction (relaxant effect) [64].

The clinical usefulness of the standard β-sympathomimetics such as epinephrine, norepinephrine, and isoproterenol is limited by their strong effects on heart rate and blood pressure. The more selective dopamine and dobutamine,

like the β-sympathomimetics mentioned above, are only effective by intravenous administration and have very short half-lives.

Dopamine. In patients with congestive heart failure, dopamine increases not only cardiac output, but also heart rate, pulmonary capillary wedge pressure, and systemic blood pressure. In addition, increased renal blood flow and consequent increased diuresis have been observed [65], [66].

Dopamine [*51-61-6*], 4-(2-aminoethyl)-1,2-benzenediol, $C_8H_{11}NO_2$, M_r 153.18, is made by demethylation of 3,4-dimethoxyphenethylamine [67]; the hydrochloride [*62-31-7*] melts at 241 °C (decomp.).

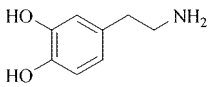

Trade names (hydrochloride): Catabon (Nikken), Dopamin (ASTA Medica AWD, Fresenius, Rhône-Poulenc Rorer), Dopamine (Abbott), Inovan (Kyowa Hakko), Intropin (American Critical Care, Du Pont).

Dobutamine, a derivative of dopamine, has comparable beneficial effects in the treatment of congestive heart failure. In contrast to dopamine, it reduces mean pulmonary wedge pressure and does not change blood pressure significantly or may even lower it [65].

Dobutamine is indicated in patients requiring short-term inotropic support. Because of its short half-life, which makes administration by infusion necessary, its use tends to be limited primarily to the treatment of cardiogenic shock.

Dobutamine [*34368-04-2*], (±)-4-[2-[3-(4-hydroxyphenyl)-1-methylpropyl]aminoethyl]-1,2-benzenediol, $C_{18}H_{23}NO_3$, M_r 301.38, is made by reductive condensation of 3,4-dimethoxyphenethylamine with 1-(4-methoxyphenyl)-3-butanone and demethylation [68]; the hydrochloride [*49745-95-1*] melts at 184 – 185 °C.

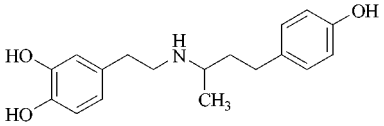

Trade names (hydrochloride): Dobutrex (Lilly).

Ibopamine [69], a prodrug form of *N*-methyldopamine, has been developed to meet the need for an orally effective positive inotropic agent for the treatment of chronic heart failure. Ibopamine is orally effective at daily doses of 100 – 350 mg with a duration of action of 4 – 5 h [70].

Ibopamine [*66195-31-1*], 4-[2-(methylamino)ethyl]benzene-1,2-diol diisobutyrate, $C_{17}H_{25}NO_4$, M_r 307.39, is made by reaction of epinine (deoxyepinephrine) with benzyl chloroformate, subsequent esterification with isobutanoyl chloride, and debenzylation [71].

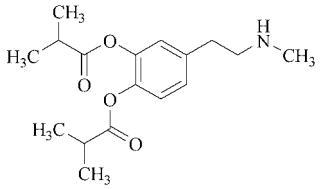

Trade names (hydrochloride): Inopamil (Astra, Zambon), Scandine (Zambon).

Prenalterol is a partial β_1-agonist for mild to moderate congestive heart failure [74]. Prenalterol increases cardiac output and stroke work index; it decreases left ventricular filling pressure and systemic vascular resistance with a slight increase in heart rate and, therefore, favorably affects the balance between myocardial oxygen supply and demand. Long-term studies have revealed a decrease in responsiveness: after 3 – 6 months the effects of prenalterol were no longer significant [72]. This down-regulation of β-adrenoreceptors arises from a decreased sensitivity or a reduction in the number of β-adrenoreceptors after prolonged stimulation and seems to be a general problem.

Prenalterol [*57526-81-5*], (−)-(*S*)-4-[2-hydroxy-3-[(1-methylethyl) amino] propoxy] phenol, $C_{12}H_{19}NO_3$, M_r 225.29, *mp* 127 – 128 °C, for its synthesis, see [73]; hydrochloride [*61260-05-7*].

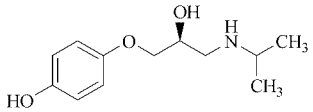

Trade names (hydrochloride): Hyprenan (Astra).

Xamoterol is another partial β_1-adrenoreceptor agonist used in the treatment of mild to moderate heart failure [74]. The drug is an agonist when sympathetic activity is low, and an antagonist when activity is high, reducing heart rate during exercise [75].

Xamoterol [*81801-12-9*], (±)-*N*-(2-[2-hydroxy-3-(4-hydroxyphenoxy)propyl]amino-ethyl)-4-morpholinecarboxamide, $C_{16}H_{25}N_3O_5$, M_r 339.39; the hydrogen fumarate [*69630-19-9*] melts at 168 – 169 °C. For synthesis, see [76].

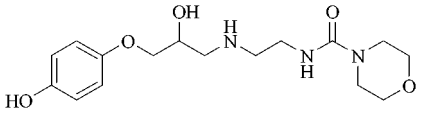

Trade names (hydrogen fumarate): Corwin (Zeneca).

Denopamine is a selective partial β_1-adrenoreceptor agonist, developed by Tanabe Seiyaku as an orally active cardiostimulant [77].

Denopamine [*71771-90-9*], (*R*)-α-([2-(3,4-dimethoxyphenyl)ethylamino]methyl)-4-hydroxybenzenemethanol, $C_{18}H_{23}NO_4$, M_r 317.18; the hydrochloride [*64299-19-0*] melts at 130 – 140 °C. For synthesis, see [78].

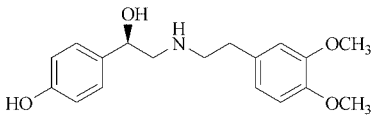

Trade names: Denopamine, Kalgut (both Tanabe).

Dopexamine is a dopamine analogue for the treatment of acute and congestive heart failure [79–81].

Dopexaminel [*86197-47-9*], 4-[2-[[6-[(2-phenetyl)amino]hexyl]amino]ethyl]-1,2-benzenediol, $C_{22}H_{32}N_2O_2$, M_r 356.51, dihydrochloride [*38029-10-6*]. For synthesis, see [82]

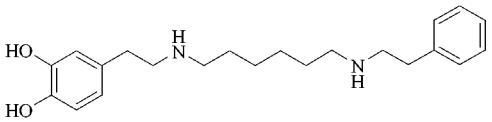

Trade names: Dopexamine (Fujisawa); dihydrochloride: Dopacard (Rhône-Poulenc Rorer).

Pirbuterol and salbutamol belong to a group of β_2-adrenoreceptor agonists. They are used as selective bronchodilators (→ Antiasthmatic Agents), but they also enhance cardiac contractility. However, vasodilation rather than direct positive inotropic effects probably is the dominant mechanism of action [59], [60], [83].

Pirbuterol [*38677-81-5*], 2-hydroxymethyl-6-(1-hydroxy-2-*tert*-butylaminoethyl)pyridine-3-ol, $C_{12}H_{20}N_2O_3$, M_r 240.30; the dihydrochloride [*38029-10-6*] melts at 182 °C (decomp.) For synthesis, see [84].

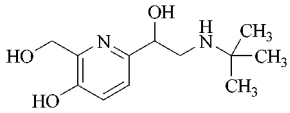

Trade names (dihydrochloride): Exirel (Pfizer) (as bronchodilator).

3.2. Phosphodiesterase Inhibitors

Inhibition of phosphodiesterases, which are responsible for the degradation of cyclic adenosine monophosphate (cAMP), increases the concentration of cAMP and the Ca^{2+} level in the myocardial cell. The following inotropic agents have the common mechanism of selective inhibition of a cAMP-specific phosphodiesterase (PDE III) with a high affinity for cAMP, which is not stimulated by calmodulin [85]. These PDE III inhibitors have positive inotropic and vasodilatory properties. Phosphodiesterase inhibitors were found to be very effective in the treatment of acute cardiac failure [86–88] and although they looked very promising initially, they proved to be very disappointing later on. These compounds were found to increase overall mortality during long-term treatment in several extensive studies and they should not be used for that purpose [89–91].

Amrinone improves cardiac performance by enhancing cardiac contractility and by decreasing systemic vascular resistance and left ventricular filling pressure with little rise in heart

rate and slight decrease in blood pressure. Because of the high incidence of side effects, such as gastrointestinal disturbances and thrombocytopenia, the development of the oral form has ceased. The intravenous form has been introduced with strong restrictions on use [59], [60], [61], [86].

Amrinone [*60719-84-8*], 5-amino-(3,4′-bipyridine)-6(1*H*)one, $C_{10}H_9N_3O$, M_r 187.20, *mp* 294 – 297 °C (decomp.), is made by reaction of γ-picoline with dimethylformamide–$POCl_3$, condensation with cyanoacetamide, hydrolysis to the amide, and reaction with bromine – sodium hydroxide [87].

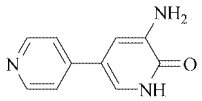

Trade names: Inocore, Wincoram (both Sanofi-Winthrop).

Milrinone has been shown to be effective in the treatment of severe heart failure[88]. Its mode of action is similar to amrinone, however, it is about 10 – 30 times more potent than amrinone. Milrinone shows a high level of side effects that limit its use to acute therapy.

Milrinone [*78415-72-2*], 1,6-dihydro-2-methyl-6-oxo(3,4′-bipyridine)-5-carbonitrile, $C_{12}H_9N_3O$, M_r 211.23, *mp* > 300 °C, is made by reaction of 1-(4-pyridine)propan-2-one with dimethylformamide dimethylacetal and condensation with cyanoacetamide [92].

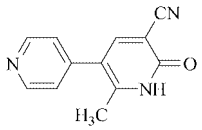

Trade names: Corotrope, Primacor (both Sanofi-Winthorp).

Enoximone – administered intravenously or orally – markedly improves left ventricular function in patients with severe congestive heart failure [93]. Cardiac side effects and gastrointestinal disturbances may limit the clinical utility of enoximone [98].

Enoximone [*77671-31-9*], 4-methyl-3-(4-methylthiobenzoyl)imidazolin-2-one, $C_{12}H_{12}N_2O_2S$, M_r 248.16, *mp* 255 – 258 °C (decomp.), is made by Friedel-Crafts acylation of 3-methylimidazolin-2-one [94].

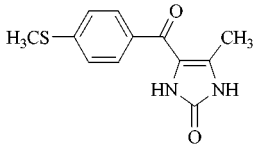

Trade names: Enoximone, Perfane (both Hoechst Marion Roussel).

3.3. Agents with Direct Stimulation of Adenylate Cyclase

Boforsin, a labdane diterpenoid, is able to stimulate the enzyme adenylate cyclase directly. This agent has positive inotropic, antihypertensive, and platelet aggregation inhibitory activity. Clinical trials with intravenous boforsin showed only a small increase in dp/dt_{max}, but strong increase in cardiac output, possibly as a result of its vasodilating properties [59], [60], [95], [96]. Development of boforsin as vasodilator and cardiotonic agent by Hoechst Marion Roussel has been discontinued.

Boforsin (forskolin) [*66575-29-9*], 5-acetoxy-3-ethenyldodecahydro-6, 10, 10b-trihydroxy-3, 4a, 7, 7, 10a-pentamethyl-[3*R*-(3α, 4aβ, 5β, 6β, 6aα, 10α, 10bα)]-1*H*-naphtho[2,1-*b*]pyran-1-one, $C_{22}H_{34}O_7$, M_r 410.49, melts at 230 – 232 °C; for its isolation, see [96].

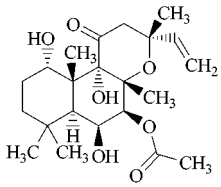

3.4. Agents Acting on the Calcium Channels

Increasing the opening time of slow Ca^{2+} channels in heart muscle cells stimulates cardiac contractility by increasing the intracellular level of Ca^{2+}. This effect is seen with agents such as *BayK 8644* [97], which is closely related to

the calcium antagonists of the dihydropyridine type (→ Calcium Antagonists). Contraction of smooth muscle cells, producing a blood pressure increase, prevented the development of BayK 8644 as an inotropic agent. Pharmacological evaluation of its congener BAY-Y-5959 indicates ongoing interest in this mode of action [98].

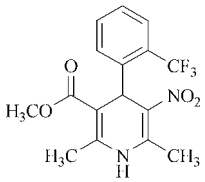

4. References

1. E. Braunwald (ed.): *Heart Disease,* W. B. Saunders Co., Philadelphia-London-Toronto 1980.
2. D. T. Mason (ed.): *Congestive Heart Failure, Yorke Medical Books,* Dun-Donnelley Publishing Corp., New York 1976.
3. A. S. Nies in K. L. Melmon, H. F. Morelli (eds.): *Clinical Pharmacology, Basic Principles in Therapeutics,* 2nd ed., Macmillan Publishing Co., New York 1978, p. 259.
4. P. Heimburg in H. Hornbostel, W. Kaufmann, W. Siegenthaler (eds.): *Innere Medizin in Praxis und Klinik,* **vol. 1,** Thieme Verlag, Stuttgart 1973, p. 3.
5. E. Braunwald in E. Braunwald (ed.): *Heart Disease,* W. B. Saunders Co., Philadelphia-London-Toronto 1980, p. 493.
6. T. W. Smith, E. Braunwald in E. Braunwald (ed.): *Heart Disease,* W. B. Saunders Co., Philadelphia-London-Toronto 1980, p. 509.
7. I. Mirsky, D. N. Ghista, H. Sandler (eds.): *Cardiac Mechanisms,* J. Wiley, New York-London-Sydney-Toronto 1974.
8. P. G. Waser, M. C. Schaub in K. Greeff (ed.): Cardiac Glycosides. Experimental Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/1** Springer Verlag, Berlin-Heidelberg-New York 1981, p. 437
9. J. Gergely, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **35** (1976) 1283.
10. D. T. Mason, *Am. J. Cardiol.* **32** (1973) 437.
11. G. Baumgarten: *Die herzwirksamen Glykoside,* VEB G. Thieme, Leipzig 1963.
12. R. Hegnauer, *Planta Med.* **19** (1970) 138.
13. K. Greeff (ed.): Cardiac Glycosides. Experimental Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/1** Springer Verlag, Berlin-Heidelberg-New York 1981.
14. R. Thomas, P. Gray, J. Andrews in: B. Testa (ed.): *Advances in Drug Research,* **vol 19**, Academic Press, London 1990, p. 311.
15. R. E. Thomas in: M. E. Wolf (ed.): *Burger's Medicinal Chemistry and Drug Duscovery,* **vol 2** John Wiley & Sons, New York 1996, p. 231.
16. B. F. Hoffmann, J. T. Bigger, in: A. Goodman Gilman, W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 814.
17. M. Gheorgiade, B. Pitt, *Am Heart J.* **134** (1996) 3.
18. G. Reuben in: C. G. Wermuth (ed.): *The Practise of Medicinal Chemsitry,* Academic Press, London 1996, p. 903.
19. Ch. Tamm, *Fortschr. Chem. Org. Naturst.* **13** (1956) 137; **14** (1957) 71.
20. P. G. Marshall in: *Rodd's Chemistry of Carbon Compounds,* vol. IID, Elsevier, Amsterdam 1970, p. 360.
21. *Elsevier's Encyclopaedia of Organic Chemistry,* **vol. 14 Suppl.**, Springer Verlag, Berlin-Heidelberg-New York 1969, p. 4377.
22. L. F. Fieser, M. Fieser: *Steroids,* Reinhold Publishing Corp., New York 1959, p. 727.
23. H. P. Albrecht in R. Ikan (ed.): *Naturally Occurring Glycosides,* John Wiley & Sons, Chichester 1999, p. 83.
24. D. Deepak, S. Srivastava, N. K. Khare, A. Khare, *Fortschr. Chem. Org. Naturst.* **69** (1996) 71.
25. T. Reichstein, E. Weiss, *Adv. Carbohydr. Chem.* **17** (1962) 65.
26. W. Klyne, *Biochem. J.* **47** (1950) 1 i.
27. H. Kubinyi, *Arch. Pharm. Ber. Dtsch. Pharm. Ges.* **304** (1971) 701.
28. B. Katzung in A. Burger (ed.): *Selected Pharmacological Testing Methods,* Medicinal Research Series vol. 3, Marcel Dekker, New York 1968, p. 193.
29. R. H. Thorp, L. B. Cobbin: *Cardiac Stimulant Substances,* Academic Press, New York-London 1967.
30. *Experimental Pharmacology,* Springer Verlag, Berlin-Heidelberg-New York 1981. K. Greeff (ed.): Cardiac Glycosides. Pharmacokinetics

and Clinical Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/2** Springer Verlag, Berlin-Heidelberg-New York 1981.
31. K. Greeff: *Probleme der klinischen Prüfung herzwirksamer Glykoside,* Steinkopff, Darmstadt 1968.
32. M. Reiter in K. Greeff (ed.): Cardiac Glycosides. Experimental Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/1** Springer Verlag, Berlin-Heidelberg-New York 1981, p. 153.
33. K. Greeff, D. Hafner in K. Greeff (ed.): Cardiac Glycosides. Experimental Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/1** Springer Verlag, Berlin-Heidelberg-New York 1981, p. 160.
34. H. Bahrmann, K. Greeff in K. Greeff (ed.): Cardiac Glycosides. Experimental Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/1** Springer Verlag, Berlin-Heidelberg-New York 1981, p. 117.
35. J. L. Suschitzky, E. Wells in P. G. Sammes (ed.): *Comprehensive Medicinal Chemistry,* **vol. 2** Pergamon Press, Oxford 1990, p. 205.
36. K. R. H. Repke, R. Megges, J. Weiland, R. Schoen, *Angew. Chem. Int. Ed.* **34** (1995) 282.
37. K. R. H. Repke, W. Schönfeld, *Trends Pharmacol. Sci.* **5** (1984) 393.
38. B. H. Marks, A. M. Weisler (eds.): *Basic and Clinical Pharmacology of Digitalis,* C. C. Thomas Publ., Springfield, Ill., 1972.
39. T. W. Smith, E. Haber, *N. Engl. J. Med.* **289** (1973) 1063.
40. T. R. D. Shaw in K. Greeff (ed.): Cardiac Glycosides. Pharmacokinetics and Clinical Pharmacology, in *Handbook of Experimental Pharmacology,* **vol. 56/2** Springer Verlag, Berlin-Heidelberg-New York 1981, p. 169.
41. E. F. Reynolds, A. B. Prasad (eds.): *Martindale, The Extra Pharmacopoeia,* 28th ed., The Pharmaceutical Press, London 1982, p. 531.
42. A. Kleemann, J. Engel: *Pharmazeutische Wirkstoffe,* 2nd ed., Thieme Verlag, Stuttgart-New York 1982.
43. S. Budavari (ed.): *The Merck Index,* 12th ed., Merck & Co, Whitehouse Station, NJ 1996.
44. A. Uskert, *Justus Liebigs Ann. Chem.* **638** (1960) 199.
45. A. Stoll, E. Angliker, F. Barfuss, W. Kussmaul, J. Renz, *Helv. Chim. Acta* **34** (1951) 1460.
46. Wyeth, US 2449673, 1946 (H. Rosen); US 2557916, 1948 (H. Rosen, D. Hill, C. T. Ichniowski); US 2615884, 1949 (H. Rosen, C. T. Ichniowski).
47. R. Megges, H. J. Portius, K. R. H. Repke, *Pharmazie* **32** (1977) 665.
48. *Drugs Exp. Clin. Res.* **6** (1980) 3.
49. Deutsche Akademie der Wissenschaften, Berlin, DE-AS 1252202,1963 (K. Repke, G. Baumgarten, R. Megges).
50. Richter Gedeon, DE-AS 1183627, 1961 (L. Wolf, A. Uskert, E. Bittner).
51. Heilmittelwerke Wien, DE 1668239, 1968 (M. Königstein, J. Schläger, K. Drexler).
52. C. H. Boehringer, DE-OS 2010422, 1970 (W. Loesel, H. Merz).
53. G. Haberland, *Arzneim. Forsch.* **15** (1965) 481.
54. Boehringer Mannheim, DE 1961034, 1969 (W. Voigtländer, F. Kaiser).
55. H. Thoms, C. Mannich, *Ber. Dtsch. Pharm. Ges.* **14** (1904) 104.
56. Goedecke, DE 1920177, 1969 (W. Herrmann, G. Satzinger, F. Zimmermann, F. Esser); DE 2050457, 1970 (W. Herrmann).
57. Knoll, US 3361630, 1964 (W. Steidle).
58. H. Kubinyi, *Arzneim. Forsch.* **28** (1978) 491.
59. J. A. Bristol, D. B. Evans, *Med. Res. Rev.* **3** (1983) 259.
60. A. E. Farah, A. A. Alousi, R. P. Schwarz, Jr., *Annu. Rev. Pharmacol. Toxicol.* **24** (1984) 275.
61. D. N. Sharpe, *Methods Find. Exp. Clin. Pharmacol.* **6** (1984) 215.
62. M. D. Taylor, I. Sircar, R. P. Steffen, *Ann. Rep. Med. Chem.* **22** (1987) 85.
63. C. V. Leier, *Am. J. Cardiol.* **69** (1992) 120 G.
64. A. M. Katz in P. Greengard, G. A. Robinson (eds.): *Advances in Cyclic Nucleotide Research,* **vol. 11,** Raven Press, New York 1979, p. 303.
65. J. Lipman, M. Plit, *S. Afr. Med. J.* **65** (1984) 506 – 509.
66. J. E. F. Reynolds (ed.): *Martindale, The Extra Pharmacopoeia,* 28th ed., The Pharmaceutical Press, London 1982, p. 9.
67. C. Schöpf, H. Bayerle, *Justus Liebigs Ann. Chem.* **513** (1934) 196. G. Hahn, K. Stiel, *Ber. Dtsch. Chem. Ges. A.* **69** (1936) 2640. Pierre Fabre S. A., FR-OS 2332748, 1977 (G. Mouzin, H. Causse).
68. Eli Lilly, DE 2317710, 1973; US 3987200, 1975 (R. R. Tuttle, J. Mills).
69. J. M. Henwood, P. A. Todd, *Drugs* **36** (1988) 11.
70. G. C. Reffo, M. Turin, A. Gabellini, C. Fovattini, *Eur. J. Clin. Pharmacol.* **26** (1984) 19.

71. SIMES S.p.A., DE 2734678, 1977 (C. Casagrande, G. Ferrari).
72. H. Lambertz, J. Meyer, R. Erbel, *Circulation* **69** (1984) 298. B. E. Strauer, I. Bohn, B. Hahn, A. Kment, U. Motz, *J. Cardiovasc. Pharmacol.* **6** (1984) 491.
73. Ciba-Geigy, DE-OS 2503968, 1975 (K. A. Jaeggi, H. Schröter, F. Ostermayer).
74. H. Sato et al., *Circulation* **75** (1987) 213.
75. N. Mizutani, T. Kobayashi, T. Watanabe, *Brit. J. Clin. Pharmacol.* **38** (1994) 285.
76. ICI, DE-OS 2822473, 1978 (B. G. Main, J. J. Barlow).
77. Y. Takatsu et al., *Jpn. Circulation J.* **51** (1987) 611.
78. *Drugs of the Future* **7** (1982) 407.
79. R. A. Brown et al., *Brit. J. Pharmacol.* **85** (1985) 599.
80. R. A. Brown et al., *Brit. J. Pharmacol.* **85** (1985) 609.
81. G. Svensson, A. Sjogren, L. Erhardt, *Eur. Heart J.* **7** (1986) 697.
82. Fisons, EP-OS 72061, 1983 (J. B. Farmer, F. Ince, R. A. Brown, J. Dixon).
83. *Drugs of the Future* **8** (1983) 65; **7** (1982) 61.
84. Pfizer Inc., EP-OS 58069, 1982 (B. W. Cue, S. S. Massett); DE 2204195, 1972 (W. E. Barth).
85. R. E. Weishaar, J. A. Bristol in P. G. Sammes (ed.): *Comprehensive Medicinal Chemistry,* **vol 2** Pergamon Press, Oxford 1990, p. 501.
86. A. Ward, R. N. Brogden, R. C. Heel, T. M. Speight, G. S. Avery, *Drugs* **26** (1983) 468.
87. Sterling Drug Inc., GB 2070600, 1981 (K. O. Gelotte, C. J. Opalka); GB 2070008, 1981 (K. O. Gelotte, E. D. Parady); DE-OS 2646469, 1976 (G. Y. Lesher, C. J. Opalka).
88. A. A. Alousi, J. M. Canter, M. J. Montenaro, D. J. Fort, R. A. Ferrari, *J. Cardiovasc. Pharmacol.* **5** (1983) 792; A. A. Alousi, G. P. Stankus, J. C. Stuart, L. H. Walton, *J. Cardiovasc. Pharmacol.* **5** (1983) 804.
89. P. Nony et al., *Eur. J. Clin. Pharmacol.* **46** (1994) 191.
90. R. DiBianco et al., *N. Engl. J. Med.* **320** (1989) 677.
91. J. K. Aronson in M. N. K. Dukes (ed.): *Myler's Side Effects of Drugs,* 13th ed., Elsevier, Amsterdam 1996, p. 447.
92. Sterling Drug Inc., DE-OS 3044568, 1980 (G. Y. Lesher, R. E. Phillion, D. F. Page, C. J. Opalka).
93. B. F. Uretsky, T. Generalovich, P. S. Reddy, R. Salerni et al., *Am. J. Cardiol.* **54** (1984) 357.
94. R. A. Schnettler, R. C. Dage, J. M. Grisar, *J. Med. Chem.* **25** (1982) 1477.
95. N. J. de Souza, A. N. Dohadwalla, J. Reden, *Med. Res. Rev.* **3** (1983) 201.
96. S. V. Bhat et al., *J. Med. Chem.* **26** (1983) 486.
97. M. Schramm, G. Thomas, R. Towart, G. Franckowiak, *Nature (London)* **303** (1983) 535.
98. N. Sato et al., *Am. J. Physiol.* **272** (1997) H753.

Cardiovascular Drugs

The following are separate keywords: Antiarrhythmic Drugs, Blood, Antihypotensives, Antihypotensives, Calcium Antagonists, Cardiac Glycosides and Synthetic Cardiotonic Drugs, Coronary Therapeutics, β-Receptor Blocking Agents

Matthias Hollmann, Mannheim, Federal Republic of Germany

Hans Dieter Lehmann, Knoll AG, Ludwigshafen, Federal Republic of Germany

Hans P. Albrecht, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Marco Thyes, Knoll AG, Ludwigshafen, Federal Republic of Germany

1. General Survey 1
2. Classification 3
3. Antiatherosclerotic Drugs (Survey) 3
4. Blood Lipid Affecting Drugs . . . 5
4.1. Introduction 5
4.2. Drugs Affecting Lipoprotein Catabolism 6
4.2.1. Ion-Exchange Resins and Sitosterol 6
4.2.2. Thyroxine Analogs 7
4.3. Drugs Affecting Lipoprotein Synthesis 7
4.3.1. Nicotinic Acid and Derivatives . . . 7
4.3.2. Aryloxyisobutyric Acids, Derivatives, and Analogs 8
4.3.3. Probucol 9
4.3.4. HMG-CoA Reductase Inhibitors . . 10
5. Platelet Affecting Drugs 11
5.1. Mechanisms of Platelet Aggregation 12
5.1.1. Physiological Situation 12
5.1.2. Pathophysiological Situation 13
5.2. Testing of Platelet Function 14
5.3. Specific Platelet Affecting Drugs . 15
5.3.1. Inhibitors of Prostaglandin Synthesis 15
5.3.1.1. Cyclooxygenase Inhibitors 15
5.3.1.2. Thromboxane Synthase Inhibitors . 16
5.3.2. Substances Influencing Cyclic 3′,5′-Adenosine Monophosphate 16
5.3.3. Inhibition of Platelet-Specific Agonist–Receptor Interaction 18
5.3.4. Calcium Entry Blockers 20
6. Blood Flow Affecting Drugs 21
6.1. Factors Influencing Blood Viscosity 21
6.2. Testing of Flow Characteristics of Blood 21
6.3. Blood Rheology Affecting Drugs . 22
6.3.1. Hemodilution 22
6.3.2. Defibrinogenation 22
6.3.3. Rheological Improvement of Red Blood Cells 23
7. References 24

1. General Survey

Cardiovascular diseases are among the most widespread of all illnesses and the most commonly occurring diseases in industrial countries. In the United States and Germany, 40 – 50 % of the annual deaths are caused by cardiovascular disorders [1], [2]. Their increase since 1945 is alarming. In some countries, such as the United States, Canada, Great Britain, and Scandinavia, this trend has reached its peak [3]. Results of the international Monica-Study (Monitoring of Trends and Determinants of Cardiovascular Disease) of the WHO show for 1985 – 1995 in most areas a distinct decrease in cardiovascular mortality, especially due to myocardial infarction by up to 50 % [4].

The treatment of these diseases plays a dominant role in medicine throughout the world, especially in the industrialized countries [5], [6], e.g., in the United States cardiovascular diseases accounted for $ 274 × 10^9 of medical expenses in 1998 [7].

In industrial nations the type of drug treatment practiced by physicians differs considerably. In Germany, for example, cardiac glycosides and antihypotensive drugs are used more frequently than in English speaking countries. Nevertheless, cardiovascular drug therapy plays a dominant role everywhere, both in terms of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_289

numbers of prescriptions and expenditures for both in- and outpatients; it represents 10 – 15 % of the total costs for drugs in Germany [4], [5].

The principal function of the cardiovascular system is the exchange of materials among organs by the blood circulation. The heart, arteries, arterioles, capillaries, and veins together with the blood contained therein represent an integrated pumping and tubing system. Cardiovascular disorders can be caused by a disturbed single malfunction or an impaired interaction of multiple functions of this system. For instance, a decrease in the pumping capacity of the heart, dilatation of the blood vessels, or decrease in the quantity of blood being transported can all have the same effect: a decrease in blood pressure, which causes insufficient blood circulation. Morphological disorders, for example, by inflammation or degeneration of the heart or the blood vessels, or an abnormal blood composition (clotting factors, blood lipids) can occur as primary damage or as a result of functional disorders.

Historical Development. HARVEY discovered the blood circulation in 1628; on this basis the most important pathoanatomical, pathophysiological, and diagnostic discoveries were made over a period of about 300 years. The first systematic pharmacotherapeutic data were collected with cardiac glycosides (WITHERING, 1785); these drugs continue to be of central importance. After the development of structural and synthetic organic chemistry and experimental pharmacology in the middle of the 19th century other essential cardiovascular drugs became available:

1) Drugs for the regulation of heart action (quinidine 1918, procainamide 1953, ajmaline 1959)
2) Drugs for the treatment of angina pectoris (nitro compounds before 1900)
3) Drugs for the lowering of blood pressure (reserpine 1952, hydralazine 1953, chlorothiazide 1959, α-methyldopa 1960)

However, these drugs affected only symptoms of the diseases; some were weak or toxic and therefore had a limited effect. They were incapable of controlling the increasing cardiovascular morbidity and mortality. This stimulated cardiologists after 1945 to devote themselves to research on atherosclerosis, which was recognized as being the most important pathological condition associated with cardiovascular diseases. The importance of changes in the clotting system (thrombosis, platelet aggregation) and such metabolic disorders as diabetes mellitus, increase in lipid and uric acid concentrations in the blood, and obesity was recognized following the fundamental studies by VIRCHOW, v. ROKITANSKY, and ANITSCHKOW (ca. 1850 – 1900). Also, the necessity for such medical intervention as dietary and drug prophylaxis at an early stage was perceived [1]. This knowledge finally led to the extensive, long-term administration of the following types of drugs:

1) Blood lipid lowering (clofibrate 1963)
2) Blood glucose lowering (tolbutamide 1956, glibenclamide 1969)
3) Anticoagulatory (coumarin derivatives, phenprocoumon 1953)
4) Platelet aggregation inhibiting (acetylsalicylic acid 1970)

Since the 1950s progress has been made in the development of intensive cardiac electrotherapy and operative therapy. In addition, major clinical–pharmacological improvements resulted from development of drugs such as lidocaine (1969), methyldigoxin (1971), and isosorbide dinitrate (1959). Other milestones in cardiovascular drug treatment are the introduction of β-receptor blockers (propranolol 1964, numerous others since), calcium channel blockers (verapamil 1963), angiotensin converting enzyme (ACE) inhibitors (captopril 1969) and HMG-CoA reductase inhibitors (compactin 1976/78 and mevinolin 1980 – see Section 4.3.4). The β-receptor blockers receive increasing criticism because of their undesired side effects, principally effects on levels of blood lipids and blood glucose. On the other hand, some β-receptor blockers have been shown to be useful in chronic heart failure which had been considered a contraindication for these drugs for many years. Calcium channel blockers and ACE inhibitors are becoming increasingly important in the therapy of coronary heart disease, hypertension, cardiac arrhythmias, and other cardiovascular disorders, particularly since the introduction of verapamil, nifedipine and captopril in the United States in 1981.

2. Classification

There are complex pathoanatomical, pathophysiological, and pathogenetic relationships in the cardiovascular system, and very diverse drug effects result. Therefore, a classified survey of the biological and medical effects of cardiovascular drugs is given in this chapter. Cross references are made to other articles in this encyclopedia that treat cardiovascular or related drugs to facilitate the introduction to and orientation within this field.

Cardioactive drugs are agents that affect cardiac structures, including the coronary vessels. *Positive inotropic drugs* enhance contractile force and are primarily represented by the *cardiac glycosides* and other positive inotropic drugs (→ Cardiac Glycosides and Synthetic Cardiotonic Drugs) as well as sympathomimetics (→ Antihypotensives). *Antiarrhythmic Drugs* have a regulating effect on an abnormally slow, fast, or irregular heart action (→ Antiarrhythmic Drugs). *Coronary therapeutics* correct any imbalance between the amount of oxygen supplied to and consumed by the heart muscle by means of diverse mechanisms, some of which also affect the peripheral vessels (→ Coronary Therapeutics). This group also comprises the *β-receptor blocking drugs* and the *calcium channel blockers,* which are also treated in separate articles because of their importance and the number of other conditions in which they are indicated (→ Calcium Antagonists; → β-Receptor Blocking Agents); both types of drugs have an antiarrhythmic and an antihypertensive action.

Vasoactive drugs primarily induce functional or structural vascular changes. *Blood pressure affecting drugs* primarily affect peripheral arteries and arterioles by changing the tension of the vascular wall. *Antihypertensives* (→ Antihypertensives) decrease blood pressure by acting either directly on the vessel wall or indirectly by central nervous neurohormonal mediation; the same effect is also exerted by β-receptor blocking drugs, calcium channel blockers, and to a certain extent by diuretics. The decrease in blood pressure by diuretics may be caused partly by a decrease in blood volume. On the other hand, blood pressure can be increased by increasing the volume, e.g., with plasma or plasma substitutes (→ Blood; → Dextran). The medical value of antihypotensive therapy (→ Antihypotensives) by increasing the tension of the venous or arterial wall (sympathomimetics), especially by long-term oral administration, is regarded very differently from country to country.

Blood flow affecting drugs (see Chap. 6) are intended to normalize blood circulation in specific regions – in the arteries, the veins, and specific organs, e.g., in migraine treatment. Their effects can involve muscular mechanisms of the vessels like those of some antihypertensives, but some also improve the rheology of the blood, e.g., dextran or fibrinogenases from snake venom.

Finally, blood flow can be improved by preventing or dissolving blood clots (anticoagulants, fibrinolytics; → Blood, Chap. 2.2.3., and see Section 5). This effect leads to the therapeutic approaches that affect morphologically and structurally fixed changes in the vessel wall accompanying atherosclerosis (see Chap. 3); this therapy is more difficult but of increasing importance.

The complexity of the cardiovascular system and its disorders – often with multifarious etiology and increasing age-dependent multimorbidity – is the reason for the use of combination drugs in cardiovascular therapy in some areas and countries [4], [5].

3. Antiatherosclerotic Drugs (Survey)

The predominant cardiovascular diseases are directly or indirectly related to arteriosclerotic (atherosclerotic) vascular changes. Atherosclerosis is a principal cause of death, at least in the industrialized nations; in these countries the fight against atherosclerosis is a major issue [1]. Today, antiatherosclerotic strategy is based on a whole group of dietetic, behavioral, and medicinal methods, reflecting the complex etiology and pathogenesis of atherosclerosis.

Lipid Hypothesis. A number of discoveries led to the formulation of the lipid hypothesis [1], [2], [8–14]; these include:

1) Pathoanatomical findings (high concentrations of cholesterol, especially in the initial stages, beginning with soft "atheromas")

2) Random and systematic epidemiological observations (changes in incidence of atherosclerosis coinciding with changes in consumption of cholesterol and saturated fatty acids during and following times of war, Framingham Study)
3) Clinical findings (early atherosclerosis in certain types of familial hyperlipidemia)

The lipid hypothesis was subsequently modified by the concept of *risk factors*. According to this concept, increased levels of lipids in the blood are a decisive pathogenetic factor in coronary and peripheral atherosclerosis; these levels are also an indirect factor in apoplexy to the extent that preexisting coronary disease is a risk factor. Blood cholesterol has been demonstrated to be the most definite risk factor; this is further refined by differentiating among the cholesterol contents of various lipoproteins [13], [14], [22] (see also Chap. 4). Although the lipid hypothesis has been confirmed qualitatively and pathogenetically, the quantitative correlation between blood lipid levels and atherosclerosis is poorer than had been expected initially. Lipids have proven to be only one factor in a large group of causes; in specific cases a main cause can be assigned, e.g., cigarette smoking in peripheral angiopathy and hypertension in apoplexy. About half of the cases of coronary heart disease can be explained by such major risk factors as hyperlipidemia, hypertension, and cigarette smoking. Further factors include diabetes mellitus, hyperuricemia, and obesity [1], [4], [14].

The lipid hypothesis was the subject of much controversy mainly from the prophylactic – therapeutic point of view, in particular with regard to drug treatment. Early and some very extensive studies, e.g., the WHO Clofibrate Study [15], yielded only partially positive results on decrease in morbidity, e.g., by cardiac infarction, by the use of blood lipid lowering drugs. With regard to mortality and the benefit : expense : risk ratio, the results were disappointing both for primary prevention (symptom-free persons) and for secondary prevention (relapse prevention after myocardial infarction).

Therefore, further efforts focussed on practically oriented studies with individual drug therapy for *multiple risk factors* such as normalization of blood lipid levels, blood pressure, weight, or cigarette smoking. Several studies of this type demonstrated a significant reduction in mortality, but they also led to controversy about biometrics and interpretation, e.g., the so-called Mr. Fit Study (MRFIT, Multiple Risk Factor Intervention Trial) [16–19]. Also, the decrease in cardiovascular mortality observed in the United States and other countries can only be explained quantitatively by public and individual multifactor intervention on the part of physicians (treatment for infarction, hypertension, hyperlipidemia, changes in smoking habits, etc.) [2], [4].

Antiatherosclerotic Drugs. Such pathogenetic and prophylactic – therapeutic complexity makes the definition of antiatherosclerotic drugs difficult. This group of drugs comprises substances that primarily influence the atherogenic risk factors and secondary complications; it would include such drugs as antihypertensives and anticoagulants. However, the definition of antiatherosclerotic drugs should be more precisely limited to those drugs that intervene in the atherosclerotic vascular processes themselves. Today no such substances are available that have proven successful in clinical trials, in spite of claims to the contrary (e.g., the so-called chelate therapy for binding and mobilizing arterial calcium deposits). Calcium channel blockers and ACE inhibitors may be considered as promising antiatherosclerotic drugs by improving so-called "endothelial dysfunction" [20], [21].

Today blood lipid affecting drugs (see Chap. 4) are the nearest yet to true antiatherosclerotic drugs. As has been shown repeatedly by angiography and other tests, these drugs slow down or decrease atherosclerotic changes at least in cases of pronounced hyperlipidemia [1], [22], [23]. Of these early studies the Coronary Primary Prevention Trial [23] demonstrated a significant decrease (30 %) in the number of infarction deaths in 1900 men with familial hypercholesterolemia during seven years of treatment with a cholesterol-lowering ion-exchange resin (see Section 4.2.1). This study exerted considerable impact by setting standards for testing blood lipid lowering/antiatherosclerotic therapeutic and preventive (primary and secondary) drug regimes, e.g. administration of fibrates (gemfibrozil, Helsinki Heart Study 1987) [24], nicotinic acid, colestipol, lovastatin (SCOR, i.e., Regression of Coronary Atherosclerosis Study

of the University of California 1990) [25] and other statins [26–30] (see Section 4.3.4). All these studies not only proved the efficacy of specific drugs/drug regimes but – taken together – built up step by step overwhelming evidence in favor of the lipid hypothesis.

4. Blood Lipid Affecting Drugs

4.1. Introduction [31–35]

Abnormal, elevated concentrations of such blood lipids as cholesterol, cholesterol esters, triglycerides, and phospholipids are called *hyperlipidemia.* Such increased levels reflect quantitative disorders in the metabolism of lipids. Lipids do not circulate freely in the blood stream because they are insoluble in water. Instead, they are transported as part of protein complexes, the *lipoproteins* [1], [11]. Hyperlipidemias always correspond to changes in lipoprotein composition and pattern. Therefore, they are also more precisely called *hyperlipoproteinemias.*

Knowledge of the different types of lipoproteins, which are continuously synthesized, interconverted, and degraded, is essential for the understanding and treatment of hyperlipidemias. Lipoproteins are divided into five main groups according to their density, composition, and electrophoretic mobility (Table 1):

1) Chylomicrons carry mainly triglycerides.
2) *Very low density lipoproteins* (VLDL) also carry mainly triglycerides.
3) *Low density lipoproteins* (LDL) contain most of the plasma cholesterol and are considered to have a very high atherogenic potential. LDL are derived from metabolic breakdown of VLDL via IDL (see below).
4) *Intermediate density lipoproteins* (IDL) normally occur only in low concentration.
5) *High density lipoproteins* (HDL) appear to have an important function in the elimination of cholesterol from tissues. High levels of HDL have been linked with a reduced incidence of atherosclerosis, indicating a protective effect of this class of lipoproteins.

In addition, a specific lipoprotein called Lipoprotein (a) was found to be strongly associated with atherosclerosis [36], [37]. The blood level of this lipoprotein is genetically determined and so far, therapeutically hardly changeable.

Following a proposal of Fredrickson [11], hyperlipoproteinemias are classified in terms of abnormalities of lipoprotein patterns, comprising *primary* (genetically determined, familial) and *secondary* (symptomatic, e.g., diabetic) disorders (Table 2). Total plasma cholesterol and triglyceride concentrations reflect the composition of the lipoprotein classes that are affected in different types of hyperlipoproteinemias. Increased levels of plasma triglycerides are caused by chylomicrons or an increase in VLDL, whereas increased cholesterol levels are usually associated with increased concentrations of LDL, e.g., in type II.

Table 2. Types of hyperlipoproteinemias according to Fredrickson [11] (For abbreviations, see text)

Type	Lipoprotein class having increased blood levels
I	chylomicrons
II a	LDL
II b	LDL and VLDL
III	abnormal IDL
IV	VLDL
V	chylomicrons and VLDL

The discovery and study of the *LDL receptor* [38], a transmembrane glycoprotein, has increased understanding of the cellular basis of cholesterol homeostasis and the development of atherosclerosis. Apoprotein B-100, which is embedded in the large hydrophilic coat of LDL, is recognized by the LDL receptor on the cell; the lipoprotein is then bound to the LDL receptor and internalized into the cell by receptor-mediated endocytosis. In the cell LDL is eventually metabolized to yield cholesterol that regulates its intracellular concentration by a complex feedback system: cholesterol suppresses the enzyme 3-hydroxy-3-methyl glutaryl coenzyme A reductase (HMG-CoA reductase), thereby inhibiting its own synthesis; cholesterol also activates cholesterol acyltransferase so that any excess of cholesterol can be reesterified. Most important, cholesterol inhibits the synthesis of the LDL receptor, thereby preventing further LDL from entering the cell. The decrease in the number of LDL receptors slows down the rate at which LDL is removed from the circulation; consequently blood levels of atherogenic LDL

Table 1. Properties of plasma lipoproteins

Lipoproteins	Density, g/mL	Electrophoretic mobility ∗	Diameter, nm	Proteins, wt %	Cholesterol, wt %	Triglycerides, wt %
Chylomicrons	<0.95	–	100 – 1000	2	5	85
VLDL	0.95 – 1.006	pre-β	25 – 75	5 – 10	25	50
IDL	1.006 – 1.019	pre-β or β	25	15 – 20	35	10
LDL	1.019 – 1.063	β	20 – 25	20 – 25	45	10
HDL	1.063 – 1.210	α	5 – 12	40 – 55	20	10

∗ Electrophoretic mobilities are expressed in comparison to α- and β-globulins.

rise. The central role that LDL receptor deficiency plays in the development of atherosclerosis is documented in patients suffering from familial hypercholesterolemia (FH) [39].

There is some controversy as to what plasma cholesterol and triglyceride levels should be considered "normal" at different ages. However, there exists general agreement that cholesterol and triglyceride levels exceeding 200 mg/100 dL should give rise to thorough and differentiated diagnostic measures to identify or exclude an increased atherogenic risk pattern and decide upon the necessity of therapeutic intervention [40]. Dietary measures and weight normalization are the first choices in the treatment of elevated levels of plasma lipids. Drugs should be used only if these measures alone are unsuccessful or cannot be fulfilled consistently. Drug therapy (and drug development) should not be aimed only at decreasing total serum cholesterol and triglycerides. The goal is to reduce VLDL and especially LDL while keeping HDL at constant or, preferably, increased levels.

Drug Development. Screening for potential drugs [41], [42] is usually performed in rats or rabbits fed either a normal, controlled diet or a diet rich in cholesterol, triglycerides, or carbohydrates in order to induce hyperlipoproteinemias. Changes in plasma levels of cholesterol and triglycerides after repeated oral application are determined by standard analytical procedures. Assays allow differentiating between atherogenic VLDL and LDL and nonatherogenic HDL cholesterol [43]. For further details regarding animal models, see [44].

Toxicity studies are of paramount importance in the evaluation of potential drugs because of the preventive nature of antiatherosclerotic therapy. Because compounds that lower blood lipid levels are found quite frequently by the above methods, lack of toxicity is the true criterion for selecting appropriate compounds for drug development.

Mode of Action. The mechanisms of action of antilipidemic agents are not always fully elucidated. The drugs currently in use can be classified according to their main mode of action:

1) The first group comprises agents that affect *lipoprotein catabolism.* This group includes substances that interfere with intestinal absorption and reabsorption of lipids during enterohepatic circulation as well as compounds that directly enhance the rate of metabolic degradation.
2) The second group includes agents that affect *lipoprotein production*; they inhibit different stages of lipid synthesis and lipoprotein formation.

4.2. Drugs Affecting Lipoprotein Catabolism

4.2.1. Ion-Exchange Resins and Sitosterol

Several strongly basic anion-exchange resins have been found useful for the treatment of type II hyperlipoproteinemia; they remain completely unchanged in the gastrointestinal tract. They act by binding bile acids and preventing their enterohepatic circulation; in this way, hepatic conversion of cholesterol into bile acids is increased. In addition, the rate of LDL catabolism is accelerated. Sitosterol, which is absorbed to the extent of less than 5 %, has a structure very similar to that of cholesterol. It inhibits cholesterol absorption competitively.

Colestyramine [*11041-12-6*] is the chloride of a pharmaceutically acceptable form of the Dowex 1-X2resin with quaternary ammonium groups incorporated into a copolymer of styrene and divinylbenzene [45]. The average molecular mass of the insoluble resin is higher than 10^6.

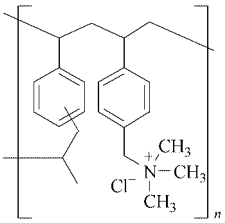

Trade names: Quantalan, Questran (both Bristol-Myers Squibb).

Colestipol [*26658-42-4*] and its hydrochloride [*37296-80-3*] are copolymers of diethylenetriamine and 1-chloro-2,3-epoxypropane. For preparation, see [46].

Trade names (hydrochloride): Cholestabil (Fournier), Colestid (Pharmacia-Upjohn).

β-Sitosterol [*83-46-5*], 22,23-dihydrostigmasterol, $C_{29}H_{50}O$, M_r 414.69, *mp*140 °C, is isolated from natural sources, e.g., tall oil, wheat and rice germ oil [47]. Partial synthesis is from stigmasterol by selective hydrogenation [48].

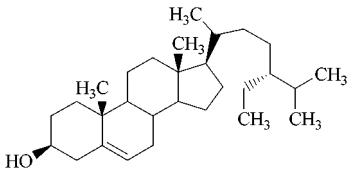

Trade names: Cytellin (Lilly), Sito-Lande (Synthelabo).

4.2.2. Thyroxine Analogs

Thyroid hormones increase the catabolic rate of cholesterol and the elimination of LDL from the plasma, but LDL synthesis remains unchanged. Development of analogs has been aimed at achieving hypolipidemic activity without increasing the basal metabolic rate. Despite considerable success, calorigenic effects and cardiotoxicity are still a problem. In general, thyroxine analogs are only rarely used as lipid lowering drugs.

Dextrothyroxine [*51-49-0*], D-thyroxine, $C_{15}H_{11}I_4NO_4$, M_r 776.93, *mp* 237 °C. For synthesis, see [49]. The sodium salt [*137-53-1*] is mainly used in therapy.

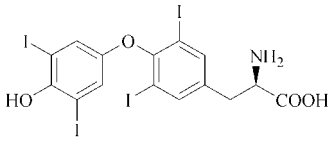

Trade name (sodium salt): Dynothel (Henning).

4.3. Drugs Affecting Lipoprotein Synthesis

4.3.1. Nicotinic Acid and Derivatives

Nicotinic acid decreases the rate of synthesis of VLDL and LDL, increases LDL catabolism, and inhibits lipolysis. It is effective in type II to type V hyperlipoproteinemias. To minimize such side effects as flush or pruritus and to prolong the duration of action prodrugs and timed-release formulations have been developed. A new analog with significant advantages is acipimox which is effective at lower doses than nicotinic acid and encumbered with fewer side effects [50].

Nicotinic acid [*59-67-6*], pyridine-3-carboxylic acid, $C_6H_5NO_2$, M_r 123.11, *mp* 236.6 °C, is synthesized by the reaction of paraldehyde with ammonia to form 5-ethyl-2-methlpyridine [51] and subsequent oxidation with nitric acid [52], see also → Pyridine and Pyridine Derivatives.

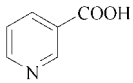

Trade names: Niacin (numerous suppliers), Nicobid, Nicolar (both Rhône-Poulenc Rorer).

Xantinol nicotinate [*437-74-1*], pyridine-3-carboxylic acid and 3,7-dihydro-7-[2-hydroxy-3-[(2-hydroxyethyl)-methylamino]propyl]-1,3-dimethyl-1-*H*-purine-2,6-dione (1:1), $C_6H_5N_1O_2$ and $C_{19}H_{26}N_6O_6$, M_r 434.35, *mp* 180 °C. Preparation according to [53].

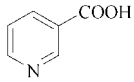

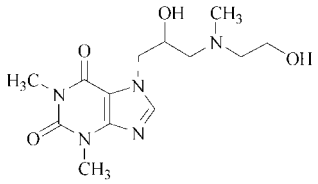

Trade name: Complamin (Smith Kline Beecham).

Inositol nicotinate [*6556-11-2*], myo-inositol hexa-3-pyridinecarboxylate, $C_{42}H_{30}N_6O_{12}$, M_r 810.73, *mp* 254 °C. For synthesis see [54].

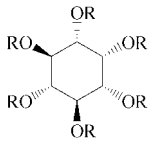

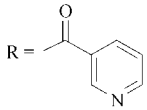

Trade names: Hexanicit (Astra), Nicolip (Henning).

Acipimox [*51037-30-0*], 2-carboxy-5-methylpyrazine 4-oxide, $C_6H_6N_2O_3$, M_r 154.13, *mp* 177 – 180 °C; for synthesis see [55].

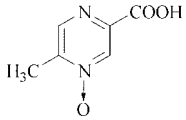

Trade names: Olbemox (Farmitalia Carlo Erba, Pharmacia-Upjohn), Olbetam (Farmitalia Carlo Erba).

4.3.2. Aryloxyisobutyric Acids, Derivatives, and Analogs

The class of compounds treated in this section comprises numerous substances with similar chemical, pharmacological, and clinical properties, which act primarily as antihypertriglyceridemic agents. The decrease in cholesterol levels is only moderate; therefore, these drugs are used mainly in type IIb, III, and type IV hyperlipoproteinemias. The oldest and still most widely used representative is clofibrate [56], which has been on the market since 1963. Mechanism of action of the fibrates is not yet completely understood. They increase the rate of VLDL catabolism by stimulating lipoprotein lipase activity and/or hepatic lipase activity [57], thereby lowering mainly plasma triglycerides. Other effects are inhibition of cholesterol synthesis and increase in excretion of neutral sterols [58].

Etofibrate combines the structural elements of nicotinic acid and clofibrate; therefore, it is used in all types of hyperlipidemias. The clofibrate analogs procetofene, bezafibrate, and gemfibrozil have only a small influence on the total cholesterol level, but they have been reported to lower selectively the LDL cholesterol level [25], [44], [59].

Clofibrate [*637-07-0*], 2-(4-chlorophenoxy)-2-methylpropanoic acidethyl ester, $C_{12}H_{15}ClO_3$, M_r 242.71, bp_{20} 148 – 150 °C, is synthesized by reaction of 4-chlorophenol with acetone and chloroform under alkaline conditions to form clofibric acid and subsequent esterification [60–62].

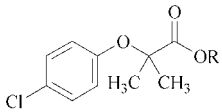

R = H: clofibric acidR = C_2H_5: clofibrate

Trade names: Atromid (American Home Products, Zeneca), Regelan (Zeneca).

Simfibrate [*14929-11-4*], 2-(4-chlorophenoxy)-2-methylpropanoic acid 1,3-propanediyl

ester, $C_{23}H_{26}Cl_2O_6$, M_r 469.36, *mp* 51 – 53 °C; for synthesis see [63].

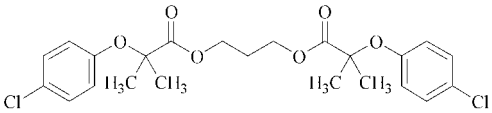

Trade name: Cholesolvin (Yoshitomi).

Procetofene [*49562-28-9*], fenofibrate, 2-[4-(4-chlorobenzoyl)phenoxy]-2-methylpropanoic acid 1-methylethyl ester, $C_{20}H_{21}ClO_4$, M_r 360.84, *mp* 80 – 81 °C; for synthesis, see [64].

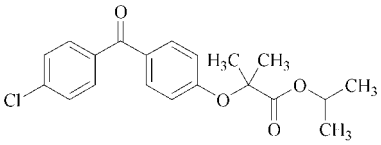

Trade name: Lipantyl (Fournier).

Bezafibrate [*41859-67-0*], 2-[4-[2-[(4-chloro-benzoyl)amino]ethyl]phenoxy]-2-methylpropanoic acid, $C_{19}H_{20}ClNO_4$, M_r 361.83, *mp* 155 – 156 °C; for synthesis, see [65].

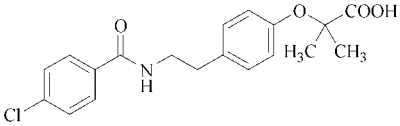

Trade name: Cedur (Roche).

Gemfibrozil [*25812-30-0*], 5-(2,5-dimethylphenoxy)-2,2-dimethylpentanoic acid, $C_{15}H_{22}O_3$, M_r 250.35, *mp* 61 – 63 °C; for synthesis, see [66], [67].

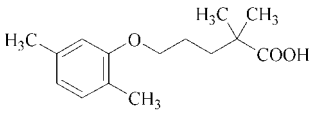

Trade names: Gevilon, Lopid (both Warner-Lambert).

Etofibrate [*31637-97-5*] [*31637-97-5*], 2-hydroxyethyl nicotinate 2-(4-chlorophenoxy)-2-methylpropionate (ester), $C_{18}H_{18}ClNO_5$, M_r 363.79, *mp* 100 °C; for synthesis, see [68].

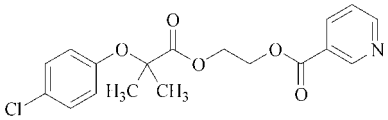

Trade name: Lipo-Merz (Merz).

Etofyllin clofibrate [*54504-70-0*], theofibrate, 2-(4-chlorophenoxy)-2-methylpropanoic acid 2-(1,2,3,6-tetrahydro-1,3-dimethyl-2,6-dioxo-7*H*-purin-7-yl) ethyl ester, $C_{19}H_{21}ClN_4O_5$, M_r 420.85, *mp* 133 – 135 °C; for synthesis, see [69].

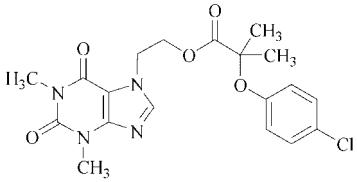

Trade name: Duolip (Mepha, Merckle).

Ciprofibrate [*52214-84-3*], 2-[4-(2,2-dichlorocyclopropyl)phenoxy]-2-methyl-propanoic acid, $C_{13}H_{14}Cl_2O_3$, M_r 289.16, *mp* 114 – 116 °C; for synthesis, see [70].

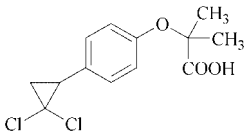

Trade name: Lipanor (Sanofi).

4.3.3. Probucol

Probucol [*23288-49-5*], 4,4′-[(1-methylethylidene)bis(thio)]bis[2,6-bis(1,1-dimethylethyl)phenol], $C_{31}H_{48}O_2S_2$, M_r 516.84, *mp* 125 – 127 °C, lowers cholesterol by an as yet only partially elucidated mechanism, which involves inhibition of cholesterol synthesis and increase in the elimination of bile acids. Probucol decreases LDL and it has been reported also to have an antioxidative effect on LDL which may retard atherosclerotic progression [71]. In addition to decreasing LDL levels, probucol also lowers HDL cholesterol levels. This fact may require a careful evaluation of its usefulness in

antiatherosclerotic therapy. The oral absorption of probucol is low (5 – 10 %). For synthesis, see [72].

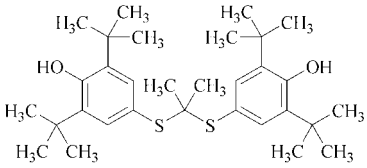

Trade names: Lorelco, Lurselle (both Hoechst Marion Roussel).

4.3.4. HMG-CoA Reductase Inhibitors

(CSE, Inhibitors, Statins). Compactin [73], [74] and mevinolin [75] are fungal metabolites isolated from the culture broths of *Penicillium citrinum* and *Aspergillus terreus,* respectively. Both compounds lower plasma cholesterol levels. They act by inhibiting β-hydroxy-β-methyl-glutaryl (HMG)-CoA-reductase, an enzyme that catalyzes an early rate-limiting step in cholesterol biosynthesis, i.e., reduction of HMG-CoA to mevalonic acid – the key substance for cholesterol biosynthesis. Following the demonstration of blood cholesterol reduction by this mechanism in humans [85], [86] intensive research and development led to introduction of lovastatin into therapy as the first drug of this class (1989). Since then, five additional compounds have been introduced, i.e., simva-, prava-, fluva-, atorva- and cerivastatin. Showing similar mechanism of action, these compounds differ in kinetics and metabolism (bioavailability <5 to 60 %). The most active compound so far is cerivastatin with effective daily dosages below 1 mg compared to 10 – 160 mg for the other compounds. The considerable medical and market success of the statins as the most widely used lipid lowering drugs is based not only on the intrinsic properties of these compounds. All statins have to a considerable extent contributed to the increase in knowledge regarding pathophysiology, clinic features and therapy of lipid disorders by producing more or less extended and well designed clinical trials for therapy and prevention in a wide area of indications: primary and secondary hyperlipoproteinemias, primary and secondary prevention, coronary heart disease, stroke, etc., with highly differentiated subgroups. Important studies are listed as follows: MARS = Monitored Atherosclerosis Regression Study (lovastatin) [28], 4S = Scandinavian Simvastatin Survival Study [27], WOS = West of Scotland Study (pravastatin) [26], CARE = Cholesterol and Recurrent Events Trial (pravastatin) [29], FLUENT = Fluvastatin Long-Term Extension Trial [30], AVERT = Atorvastin Versus Revascularization Treatments (1999, in progress).

Lovastatin [*75330-75-5*], (2*S*)-2-methylbutanoicacid (1*S*, 3*R*, 7*S*, 8*S*, 8*aR*)-1,2,3,7,8,8*a*-hexahydro-3,7-dimethyl-8-[2-[(2*R*,4*R*)-tetrahydro-4-hydroxy-6-oxo-2*H*-pyran-2-yl]-ethyl]-1-naphthalenyl ester, $C_{24}H_{36}O_5$, M_r 404.55, *mp* 174.5 °C. Lovastatin has been discovered and isolated as metabolite of *Monascus ruber*. [76] and from *Aspergillus terreus* [75].

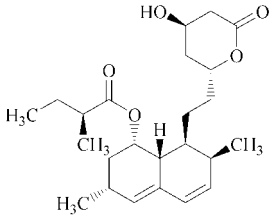

Trade names: Mevacor, Mevinacor (both Merck & Co.).

Simvastatin [*79902-63-9*], 2,2-dimethylbutanoic acid (1*S*, 3*R*, 7*S*, 8*S*, 8*aR*)-1,2,3,7,8,8*a*-hexahydro-3,7-dimethyl-8-[2-[(2*R*,4*R*)-tetrahydro-4-hydroxy-6-oxo-2*H*-pyran-2-yl]-ethyl]-1-naphthalenyl ester, $C_{25}H_{38}O_5$, M_r 418.57, *mp* 135 – 138 °C. Simvastatin is a once-daily analogue of lovastatin. For synthesis see [77], [78].

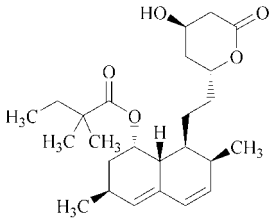

Trade names: Denan (Boehringer Ingelheim), Lodales (Sanofi), Zocor (Dieckmann, Merck & Co).

Pravastatin [*81093-37-0*], (*βR*,*δR*,1*S*,2*S*,6*S*,8*aR*)-1,2,6,7,8,8*a*-hexahydro-*β*,*δ*,6-trihydroxy-2-methyl-8-[(2*S*)-2-methyl-1-oxobutoxy]-1-naphthaleneheptanoic acid, $C_{23}H_{36}O_7$, M_r 424.53. Pravastatin is the bioactive metabolite of mevastatin. For its preparation by microbial hydroxylation of mevastatin [79] at the 6-position see [80].

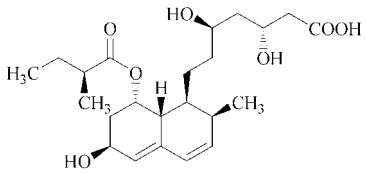

Sodium salt [*81131-70-6*], $C_{23}H_{35}O_7Na$, M_r 446.52.

Trade names (sodium salt): Pravachol, Pravasin (both Bristol-Myers Squibb), Mevalotin (Sankyo).

Fluvastatin [*93957-54-1*], (3*R*,5*S*,6*E*)-7-[3-(4-fluorophenyl)-1-(1-methylethyl)-1*H*-indol-2-yl]-3,5-dihydroxy-6-heptenoic acid, $C_{24}H_{26}FNO_4$, M_r 411.47, is a synthetic HMG-CoA reductase inhibitor, prepared according to [81].

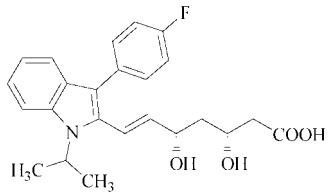

Sodium salt [*93957-55-2*], $C_{24}H_{26}FNaNO_4$, M_r 434.46, *mp*194 – 197 °C.

Trade names (sodium salt): Canef, Cranoc (both Astra), Lescol (Novartis), Lochol (Tanabe Seiyaku).

Atorvastatin [*134523-00-5*], (*βR*,*δR*)-2-(4-fluorphenyl)-*β*, *δ*-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-1*H*-pyrrole-1-heptanoic acid, $C_{33}H_{35}FN_2O_5$, M_r 558.65. For synthesis see [82].

Calcium salt [*134523-03-8*], (*βR*,*δR*)-2-(4-fluorphenyl)-*β*, *δ*-dihydroxy-5-(1-methylethyl)-3-phenyl-4-[(phenylamino)carbonyl]-1*H*-pyrrole-1-heptanoic acid calcium salt, $C_{66}H_{68}CaFN_4O_{10}$, M_r 1155.38.

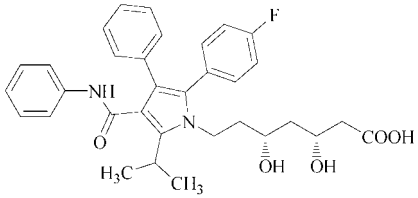

Atorvastatin calcium is the only HMG-CoA reductase inhibitor approved for type III and IV lipid disorders.

Trade names (calcium salt): Lipitor (Warner-Lambert, Pfizer), Sortis (Warner-Lambert).

Cerivastatin [*145599-86-6*], (3*R*,5*S*,6*E*)-7-[4-(4-fluorophenyl)-5-methoxymethyl)-2,6-bis(1-methylethyl)-3-pyridinyl]-3,5-dihydroxy-6-heptenoicacid, $C_{26}H_{34}FNO_5$, M_r 459.55. For synthesis see [83].

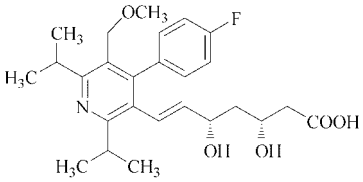

Sodium salt [*143201-11-0*], $C_{26}H_{33}FNaO_5$, M_r 481.54. Preparation of the sodiumsalt [84].

Trade names (sodium salt): Baycol (Bayer, Smith Kline Beecham), Lipobay (Bayer), Zenas (Fournier).

5. Platelet Affecting Drugs

Platelet affecting drugs inhibit platelet aggregation and therefore prevent the formation of occluding platelet aggregates and thrombi. They are increasingly used in various thromboembolic cardiovascular diseases of mainly atherosclerotic origin, including myocardial infarction, stroke, transitory ischemic attacks and venous thrombosis. The effectivity of such therapy is confirmed [87]. A further indication is the prevention of ischemic complications in patients undergoing vascular surgery (e.g. percutaneous coronary revascularization). The therapy is based on the discovery that the aggregation of platelets – normally a physiologically important initial step in hemostasis maintaining the integrity of the vascular system – acquires

pathological significance when it becomes excessive, because it then can induce thrombotic occlusion and thus impairment of blood flow. To understand the mode of action of these agents, a knowledge of the very complex physiological and pathophysiological mechanisms that lead to platelet aggregation is necessary (see also → Blood, Chap. 2.2.3.).

5.1. Mechanisms of Platelet Aggregation

5.1.1. Physiological Situation

Under physiological conditions, when the vessel wall has been injured, the platelets immediately adhere to the exposed subendothelium, forming a monolayer. More platelets subsequently adhere to this monolayer to form an aggregate, which mechanically plugs the damaged site. This process is based on an activation of the platelets, which occurs in several stages: 1) an initial stickiness of the surface, 2) a shape-change, 3) aggregation, and 4) secretion. This activation can be triggered by numerous substances acting on specific receptors of the plasma membrane. The collagen of the subendothelium, which is exposed by vascular injuries, is of particular importance. Other important activation factors are the endogenous mediators released from activated platelets: adenosine diphosphate (ADP), the arachidonic acid metabolite thromboxane A_2 (TXA_2), and the platelet activating factor (PAF) [88]. Epinephrine (adrenaline) and norepinephrine (noradrenaline) probably play a special role; they are only weak aggregation inducers, but they lead to a massive intensification of the aggregation triggered by other inducers [89], [90].

Increased stickiness of the platelet surface is a prerequisite for adhesion and aggregation. The processes underlying this event involve a series of interactions between activated platelet receptors and adhesive proteins which are widely unknown in details [91]. The subsequent shape change consists of a transition from a disc shape to a globular form bearing pseudopodia-like protrusions, by which intercellular contact is facilitated. This shape-change is due to a contractile process mediated by actomyosin.

After the appearance of stickiness and shape-change, additional factors are necessary for aggregation to occur. These are a sufficient rate of blood flow (to bring about platelet contact) and sufficient amounts of calcium ions and fibrinogen in the plasma. A prerequisite to the aggregation process is finally the activation of glycoprotein IIb/IIIa receptors on the platelet surface which bind circulating fibrinogen (and other adhesive proteins) leading to the formation of fibrinogen bridges between the platelets [92].

Various substances are released during the aggregation process, either from storage in specific granules or through immediate biosynthesis. Storage compounds are: ADP, ATP, 5-hydroxytryptamine (5-HT), and calcium ions from the dense granules; and fibrinogen, platelet factor 4, β-thromboglobulin, and the platelet-derived growth factor (PDGF) from the α-granules. Immediately synthesized substances are the various arachidonic acid metabolites (in particular TXA_2) and the platelet activating factor (PAF).

The released substances induce a further avalanche-like increase in platelet aggregation (by ADP and TXA_2), a vascular contraction (by TXA_2 and 5-HT), which creates conditions of flow dynamics needed for the preservation of the platelet aggregate, and finally a promotion of blood clotting (by fibrinogen and platelet factor 4), which stabilizes the aggregate by fibrin precipitation and changes it into a thrombus. An integral part of this stabilization process is the contraction of actomyosin, which causes not only the secretion, but also the subsequent retraction of the clot.

An essential link between receptor stimulation and platelet reaction is the increase in the free calcium ion concentration in the platelets by release from intracellular stores and by inflow from extracellular space. The hydrolysis of inositol phospholipids in the cell membrane, in particular the catalytic formation of inositol 1,4,5-triphosphate (via phospholipase C), which mobilizes Ca^{2+} from intracellular stores, plays a central role in the increase of cytoplasmic Ca^{2+} concentration [93–96]. In addition, the cytoplasmic calcium ion concentration is increased by the influx of extracellular calcium ions into the cytoplasma. These processes can be reversed by various mechanisms, which finally restore the basal calcium ion concentration. Thus, the calcium ion-triggered platelet homeostasis can be regulated by intracellular cyclic 3′,5′-adenosine monophosphate (cAMP) which arises from ATP

by the action of adenylate cyclase. Activation of this enzyme (through different prostaglandins, among others) leads to an increase of the cAMP level, and thus to a sequestration of calcium ions within the dense tubular system [97–99].

Platelet activation is accompanied by a stimulation of lipid metabolism in the platelet membrane; the products of this process promote aggregation, see Figure 1. In addition to the breakdown of inositol phospholipids mentioned in the preceding paragraph the following processes also occur: formation of ether lipids (platelet activating factor; PAF) from choline alkyl ether [100] and liberation of arachidonic acid from phosphatidylethanolamine and phosphatidylcholine.

Arachidonic acid undergoes further changes [101], [102]. It is transformed by cyclooxygenase into the cyclic prostaglandin endoperoxides G_2 and H_2 (PGG_2 and PGH_2), which are further transformed into the extraordinarily potent aggregation inducer thromboxane A_2 (TXA_2; half-life 30 s) through the action of thromboxane synthetase. The manner in which the arachidonic acid metabolites 12-L-hydroxyperoxy- and 12-L-hydroxy-5,8,10,14-eicosatetraenoic acid (HPETE/HETE) formed by the action of lipoxygenase act on the platelets is not yet fully understood.

The importance of the arachidonic acid metabolites for platelet physiology and pathophysiology has gained a new aspect with the discovery of prostaglandin I_2 (PGI_2), or prostacyclin, in the vessel wall, particularly in the endothelial layer [103]. PGI_2 is a short-lived substance ($t_{1/2} = 2$ min) with powerful antiaggregatory and vasodilatory actions. Its action opposes that of TXA_2, which possesses both proaggregatory and strong vasoconstrictor properties. According to VANE's hypothesis, under normal conditions the formation of PGI_2 protects the intact vascular wall from any activation and aggregation of platelets by antagonizing TXA_2. However, injury of the vessel wall with loss of endothelium leads to a deficiency in PGI_2, with consequent displacement of equilibrium between pro- and antiaggregatory forces; this then results in platelet activation and thrombus formation under the influence of TXA_2.

5.1.2. Pathophysiological Situation

The pathophysiologically relevant formation of occluding platelet aggregates and thrombi occurs mainly at the site of chronic vascular wall alterations associated with atherosclerosis. These focal alterations start from damage to the endothelial cell layer of arteries, the "early lesions of the intima". This damage is caused by unknown mechanisms through such risk factors as overweight, hyperlipidemia, hypertension, and smoking. As a reaction, platelet accumulation as a rather early event and increased growth of smooth muscle cells occur, which results in a thickening of the vessel wall and thus narrowing of the vessel lumen. The process is reversible up to this point. However, if there is an elevated blood lipid level (hyperlipidemia), the smooth muscle cells will increasingly incorporate fat and eventually die. A necrosis of the fat-rich vessel wall, the irreversible *atherosclerotic plaque*, is the final result. Lesion of the intima as well as plaque formation initiate platelet activation and thereby induce thrombus formation.

Platelet aggregation is to be regarded as an essential pathogenic factor in the development of the atherosclerotic process [104]. According to experimental evidence, the *platelet-derived growth factor* (PDGF) plays an essential role in atherosclerosis. PDGF is a cationic protein with a molecular mass of about 30 000, liberated in the course of the release reaction; it is characterized by its strong stimulating action on the proliferation of arterial smooth muscle cells and fibroblasts in tissue cultures [105]. Whether such an action also takes place in vivo is as yet hypothetical. However, the release of PDGF may be the mechanism by which the atherosclerotic process is perpetuated: in a vicious circle PDGF liberation leads to progression of vessel wall alterations, inducing further platelet aggregation and renewed PDGF liberation. Additionally, the discovered macrophage-derived growth factor, which appears to be different from PDGF, may also play a role in atherogenesis [106], [107].

According to experimental studies on animals and clinical investigations, platelet affecting drugs can markedly reduce the proliferation of smooth muscle cells and the formation of preatherosclerotic and atherosclerotic lesions in vivo [106–109].

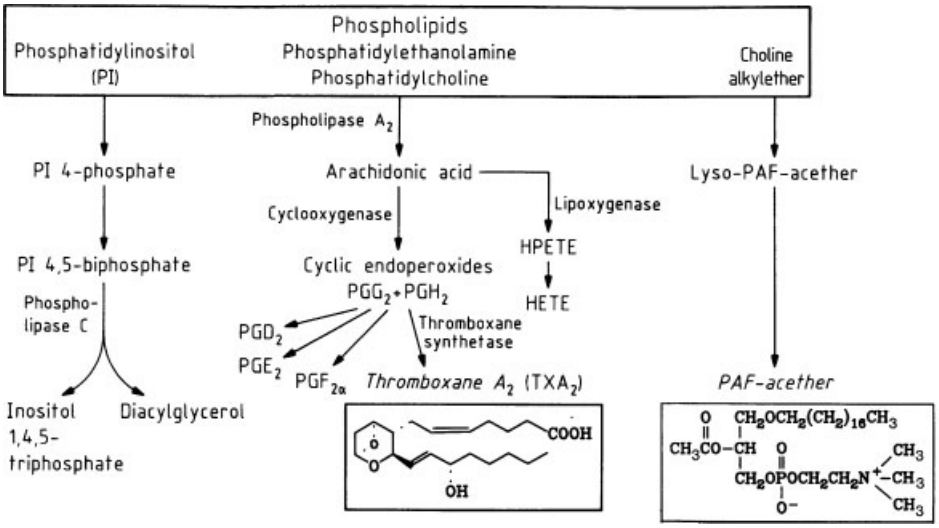

Figure 1. Pathways of lipid metabolism in activated platelets
(For abbreviations, see text)

In many cardiovascular diseases, there is an increased tendency for platelets to aggregate, and spontaneous formation of aggregates in the blood is observed. The causes of this effect are still rather obscure, but they are possibly linked to hyperlipidemia and an increased plasma level of catecholamines [110–114].

The significant *vasoconstrictor action* of the aggregators TXA_2 and 5-HT released from the platelets alone is probably of pathogenetic significance. This effect may participate in the development of circulatory shock, myocardial ischemia, and sudden death [115–117].

The pathophysiological significance of platelets is strongest in the arterial vascular bed because of the conditions of flow dynamics existing there and the positive correlation between blood flow rate and aggregation [118]. Consequently, in the venous vascular bed, platelet aggregation is less important in the development of thrombosis than clotting factors, but antiplatelet therapy reduced significantly the risk of deep venous thrombosis and pulmonary embolism [87].

5.2. Testing of Platelet Function

Platelet function is tested for clinical or experimental reasons preferably in vitro by means of platelet aggregation because good quantitative and reproducible methods exist [119]. For this purpose platelet-rich plasma (PRP) obtained from venous blood is used and aggregation is induced by the addition of such compounds as ADP, collagen, thrombin, PAF, epinephrine, or 5-HT. In addition, spontaneous aggregation can be studied [120]. Aggregation is manifested optically by clarification of the initially turbid plasma. This clarification can be followed photometrically in an aggregometer; as an index the aggregation rate, its maximum, or the area under the aggregation curve is used. In addition, whole blood aggregations can be measured by impedance technique [121]. Platelet affecting drugs can be studied either after addition in vitro or they are administered to humans or animals and aggregation is determined in plasma or blood samples obtained at different times after the administration. The extent of the aggregation inhibition is evaluated by comparison with a control group [122].

By suitably modified techniques the initial shape-change of the platelets and the secretion (e.g., of ATP) can also be studied [123]. The adhesiveness of platelets is determined, for example, by measuring the decrease in platelet count after running platelet-rich plasma through a tube filled with glass beads.

For special experimental investigations washed or gel-filtered platelets are used; in this case, the plasma is replaced by electrolyte solutions [124].

Several methods are available for the study of aggregation or its inhibition in vivo, e.g., the induction of thrombocytopenia by intravenous injection of an inducer, such as ADP, to produce intravascular aggregation; the subsequent decrease in platelet count with time is plotted to establish the kinetics of the process [125].

Finally, for the determination of an antithrombotic action, the vessel wall can be damaged by several means or an artificial arterio-arterial or arterio-venous shunt is inserted and thrombus formation can be followed gravimetrically, by radioactive markers, histologically or by measuring the flow reduction [126–131].

5.3. Specific Platelet Affecting Drugs

5.3.1. Inhibitors of Prostaglandin Synthesis [132]

Thromboxane A_2 (TXA_2) is one of the most active mediators of aggregation (see Section 5.1). The aggregation induced by low concentrations of collagen depends on TXA_2 formation, whereas the initial adhesion and the related release reaction occur without participation of TXA_2. Platelet aggregation caused by thrombin and PAF, as well as the sensitization of the platelets by epinephrine, are also independent of TXA_2. Therefore, substances that inhibit prostaglandin synthesis can prevent only one of the ways by which platelets are able to mediate thrombogenesis.

Prostaglandin synthesis inhibitors are classified into two categories:

1) *Cyclooxygenase inhibitors* (acetylsalicylic acid/aspirin)
2) *Thromboxane synthase inhibitors* (no drug on the market)

5.3.1.1. Cyclooxygenase Inhibitors

Aspirin irreversibly inhibits platelet cyclooxygenase through acetylation. Because platelets are unable to synthesize proteins, the inhibitory effect persists until the end of the life span of the platelets, i.e., for about 7 days [133]. The cyclooxygenase inhibition is not tissue-specific for platelets but specificity can be achieved by low oral doses and controlled release formulations of aspirin to restrict the drug – platelet interaction to the presystemic, portal circulation [134].

Cyclooxygenase inhibitors prevent not only the formation of TXA_2, but also that of prostaglandin endoperoxides, e.g., PGH_2 and PGD_2 (Fig. 1), which have partly pro-, partly antiaggregatory potential, as well as that of PGI_2 in the vessel wall. The latter action is undesirable because of the antithrombotic properties of PGI_2 (see page 17). In the aortic intima, however, cyclooxygenase is about 60 – 250 times less sensitive to aspirin than in platelets; in addition, the inhibition is reversible there [135–137]. Although a thrombogenic action has not been demonstrated for higher doses of aspirin, lower doses are recommended in order to limit the inhibition of PGI_2 synthesis (325 mg/day or as low as 0.45 mg kg^{-1} day^{-1}) [138–140].

Clinical trials showed that the use of aspirin in survivors of myocardial infarction reduces the risk of reinfarction and death significantly [87], [141]. It is also effective in certain forms of cerebral vascular disease, however, the optimal dosage is less clear [142], [87]. Furthermore aspirin is used at present as a standard treatment in patients undergoing coronary revascularization [87], [143]. Aspirin as an antiplatelet drug is most widely administered in a medium dose of 75 – 325 mg/d which reduces the risk of gastrointestinal intolerance or bleeding and results in an acceptable risk – benefit ratio for most patients. The gastrointestinal intolerance can be further diminished by using enteric-coated formulations [143].

From the mode of action it is obvious that aspirin is not an optimal antiplatelet drug, especially in the indication of coronary revascularization, since important pathways in platelet activation which are of relevance in such patients, e.g., activation by ADP and thrombin, expression of GP IIb/IIIa receptors, are not affected. The efficacy of antiplatelet therapy can therefore be improved if aspirin is combined with drugs which interfere with the activation processes mentioned.

Acetylsalicylic acid [*50-78-2*], aspirin, 2-(acetyloxy)benzoic acid, salicylicacid acetate, is described in more detail under → Salicylic Acid.

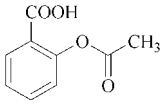

5.3.1.2. Thromboxane Synthase Inhibitors

The undesirable effect of cyclooxygenase inhibitors on prostacyclin (PGI_2) formation has stimulated the development of thromboxane synthase inhibiting drugs. These substances selectively inhibit the transformation of endoperoxides to TXA_2; in this manner they shift the metabolism of the endoperoxides toward the formation of, e.g., PGH_2 and PGD_2 (in platelets, Fig. 1), and PGI_2 (in endothelial cells). It was subsequently shown that the persisting formation of the proaggregatory PGH_2 is of relevance for aggregation [144], but can be inhibited by TXA_2 receptor antagonists. Therefore, some expectations were directed to ridogrel, a potent inhibitor of the TXA_2 synthase that also possesses TXA_2/prostaglandin endoperoxide receptor blocking properties [145]. However, ridogrel was not superior to aspirin in patients with myocardial infarction after thrombolysis therapy in preventing early phase rethrombosis, but might be more effective in preventing new ischemic events [146]. The missing clear superiority of ridogrel to aspirin might be explained by different mechanisms, e.g., an insufficient capacity of atherosclerotic arteries to produce prostacyclin. More controlled prospective clinical trials are required to assess the therapeutic value of such substances especially in combination with aspirin.

Ridogrel [*110140-89-1*], R-68070, (*E*)-5-[[[3-pyridinyl[3-(trifluoromethyl)phenyl] methylene]amino]oxy]pentanoic acid, $C_{18}H_{17}F_3N_2O_3$, M_r 366.34, *mp* 70.3 °C; for synthesis, see [147].

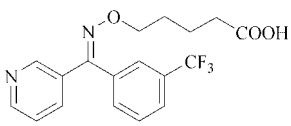

5.3.2. Substances Influencing Cyclic 3′,5′-Adenosine Monophosphate

An increase in cyclic adenosine monophosphate (cAMP) prevents the initial shape change of the platelets, their adhesion to surfaces, the aggregation, and the release reaction, regardless of the type of stimulator. As a consequence of this broad activity pattern, substances influencing cAMP are of potential therapeutic interest.

The concentration of cAMP in the platelets can be increased (1) by stimulation of adenylate cyclase, which is responsible for the transformation of ATP into cAMP, or (2) by inhibition of cAMP-phosphodiesterase, which catalyzes the breakdown of cAMP to AMP (Fig. 2). The adenylate cyclase is stimulated via different receptors by prostacyclin (PGI_2) and the less potent PGD_2. Papaverine and dipyridamole inhibit cAMP-phosphodiesterase. Such inhibitors reinforce the action of PGI_2 and other adenylate cyclase activators; this reinforcement is generally obtained at lower doses than those that are necessary for inhibition of aggregation. Prostacyclin (PGI_2) and dipyridamole are clinically established as platelet aggregation inhibitors.

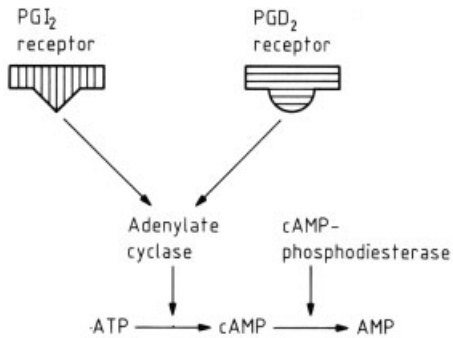

Figure 2. Mechanisms influencing the concentration of cAMP in platelets

Prostacyclin/Epoprostenol. At present prostacyclin (PGI_2) is the most active platelet aggregation inhibitor; the ADP-induced aggregation is inhibited by 50 % at a plasma concentration of 0.5 ng/L [148]. In addition, PGI_2 causes, both in humans and in animals, a systemic vasodilation with dose-dependent blood

pressure lowering, accompanied by an increase in heart rate; furthermore, a stabilization of lysosomes has been shown. The effects of PGI_2 are receptor -mediated and indications exist that these receptors may become down-regulated if there is a significant increase in PGI_2 production (e.g., acute myocardial infarction, unstable angina) [149], [150]. Decreased sensitivity of human platelets to PGI_2 also occurred during long-term intraarterial PGI_2 infusion in patients with peripheral vascular disease [151]. Prostacyclin has a very short biological half-life and must be administered as a continuous intravenous infusion. In humans, flush and, in most cases, headache develop, which disappear after the infusion is interrupted or slowed down [152]. No encouraging results were obtained with PGI_2 in a randomized, double-blind, placebo-controlled trial in patients undergoing percutaneous coronary revascularization. The drug was unable to inhibit platelet aggregation and to prevent restenosis [153].

Epoprostenol [*35121-78-9*], PGI_2, prostacyclin, 5-[(1*S*,3*Z*,5*R*,6*R*,7*R*)-7-hydroxy-6-[(1*E*,3*S*)-3-hydroxy-1-octenyl]-2-oxabicyclo[3.3.0]-oct-3-ylidene]pentanoic acid, (5*Z*,9α,11α,13*E*,15*S*)-6,9-epoxy-11,15-dihydroxyprosta-5,13-dien-1-oic acid, $C_{20}H_{32}O_5$, M_r 352.48, is unstable in aqueous solution [154]: its half-life at 37 °C is 2 min [155]; the monosodium salt [*61849-14-7*] is stable as a solid and insolution [156].

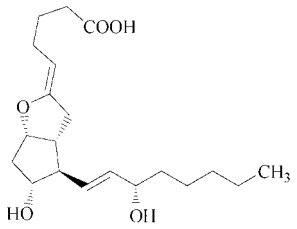

Prostacyclin is synthesized from prostaglandin $F_{2\alpha}$ systems [156–161].

Trade name monosodium salt: Flolan (Glaxo Wellcome).

Chemically stable PGI_2 derivatives such as carbacyclin [162] and iloprost [163] are easier to handle than prostacyclin. Carbacyclin [*69552-46-1*] has the same pharmacological profile as PGI_2 but only ca. 3 % of its activity. Iloprost is equipotent to PGI_2 with reference to platelet activity in vitro, but possibly superior to it on the basis of a reduced blood pressure lowering activity. In an experimental study in dogs iloprost was uneffective to prevent reocclusion of coronary arteries after thrombosis due to electrical injury and thrombolysis [164]. Until now there is no definite evidence that PGI_2 derivatives might be of clinical value as antithrombotics or as adjunct to aspirin. The development of carbacyclin has been discontinued and iloprost and other PGI_2 derivatives are at present used as blood flow increasing drugs.

Iloprost [*78919-13-8*], ciloprost, 5-[(1*S*,3*E*,5*S*,6*R*,7*R*)-7-hydroxy-6-[(1*E*,3*S*,4*RS*)-3-hydroxy-4-methyl-1-octen-6-ynyl]bicyclo[3.3.0]oct-3-ylidene]pentanoic acid, $C_{22}H_{32}O_4$, M_r 360.50; for synthesis, see [165–167].

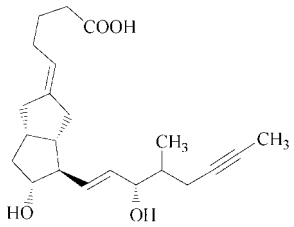

Trade names: Endoprost, Ilomedin, Ilomedine (all Schering).

Dipyridamole. Dipyridamole, originally introduced as a coronary vasodilator, has an antithrombotic effect in animals [168]. The inhibition of platelet aggregation in vitro using platelet-rich plasma is weak and requires high concentrations [169]; a strong antiaggregatory action is nevertheless found when the test is carried out on whole blood [170], [171]. In whole blood the drug probably acts by increasing the plasma concentration of adenosine because it inhibits the re-uptake of adenosine by red blood cells. Adenosine itself stimulates adenylate cyclase via an external adenosine receptor, thereby causing an increase of cAMP. This effect is enhanced by an inhibition of platelet cAMP phosphodiesterase. Dipyridamole reduces the size of experimentally induced myocardial infarction [172]. Platelet adhesion to the subendothelium is prevented by dipyridamole in animals under experimental conditions [173], [174]. Evidence

exists for both a protective and a promoting action in experimentally induced atherosclerosis [175], [176].

Inhibition of thrombus formation by dipyridamole has been repeatedly demonstrated in humans. The superiority of a combination of dipyridamole and aspirin compared with aspirin in humans has been demonstrated in multicenter studies on the protective action on myocardial reinfarction [177] and on cerebrovascular thrombosis [178].

Dipyridamole [*58-32-2*], 2,2′,2″,2‴-[(4,8-di-1-piperidinylpyrimido[5,4-*d*]pyrimidine-2,6-diyl)dinitrilo]tetraethanol, $C_{24}H_{40}N_8O_4$, M_r 504.64, *mp* 163 – 165 °C, can be obtained from 2,4,6,8-tetrahydroxypyrimido[5,4-*d*]pyrimidine [179]:

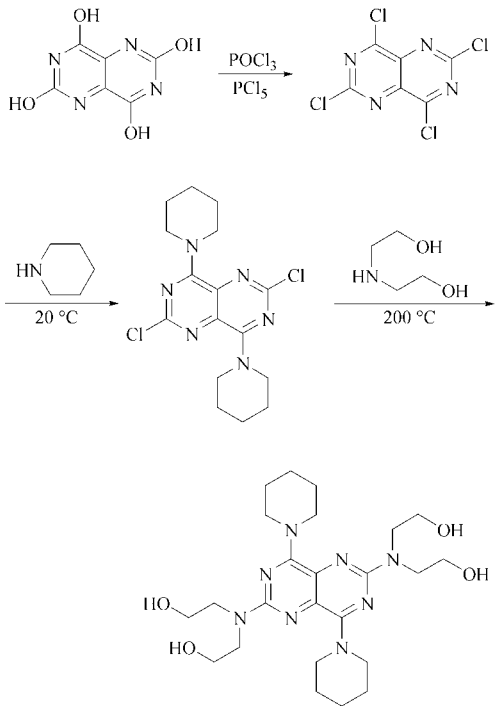

For another route to dipyridamole, see [180].

Trade names: Agilease (Isei), Anginal (Yamanouchi), Cleridium (Marcofina), Coronarine (Negma), Corosan (Farmacologico Milanese), Coroxin (Malesci), Curantyl (Berlin-Chemie), Novodil (OFF), Perkod (Lab. Biogalénique), Persantin (Boehringer Ingelheim), Persantine (Boehringer Ingelheim); in combination with acetylsalicylic acid as Asasantin (Boehringer Ingelheim).

5.3.3. Inhibition of Platelet-Specific Agonist–Receptor Interaction

The platelets are activated by substances that interact with specific receptors in the plasma membrane. The inhibition of this interaction limits activation. Such a mechanism is apparently responsible for the effect of ticlopidine and the thromboxane receptor antagonists.

Ticlopidine. Ticlopidine is a long-acting inhibitor of platelet aggregation; all types of aggregation are inhibited [181–183]. The action of ticlopidine appears with a delay of 24 – 48 h and persists according to investigations in rats up to the end of the life span of the platelets (the half-life is about 2 days). Ticlopidine appears to change the platelet membrane irreversibly so that the binding of ADP and fibrinogen is greatly reduced. This effect is independent of the prostaglandin synthesis and the cAMP level. Furthermore, the substance has no influence on the formation of PGI_2 in the vessel wall. It prevents the proliferation of the intima, which occurs after experimental removal of the endothelium in the rabbit aorta [184]. The activity of the substance on platelet aggregation and adhesion has been demonstrated in humans [185], [186]; the maximum activity appears only after several days of administration. The effectivity of ticlopidine for the secondary prevention of stroke was shown [187]. However, in a ticlopidine aspirin stroke study ticlopidine was only slightly more effective than aspirin in reducing the incidence of stroke and death (17 versus 19 %) [188].

In some patients a change in the white blood cell count (granulopenia) was found as a side effect; the substance is therefore preferentially used in patients with intolerance to aspirin.

Ticlopidine [*55142-85-3*], 5-[(2-chlorophenyl)methyl]-4,5,6,7-tetrahydrothieno[3,2-*c*]pyridine, $C_{14}H_{14}ClNS$, M_r 263.79; hydrochloride [*53885-35-1*], *mp* 190 °C (decomp.), can be synthesized from thieno[3,2-*c*]pyridine [189]:

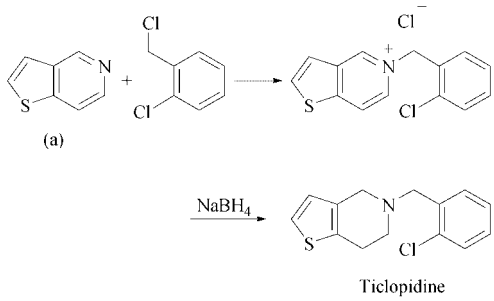

(see a: [190])

According to another method, ticlopidine is obtained by reacting 4,5,6,7-tetrahydrothieno-[3,2-*c*]pyridine with 2-chlorobenzyl chloride in the presence of potassium carbonate [189, b], [191]:

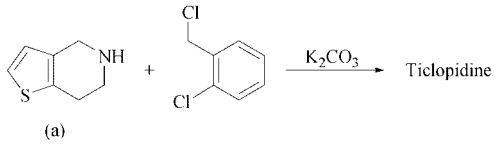

(see a: [190, b], [191], [192])

Another route leading to ticlopidine is shown in the following scheme [193]:

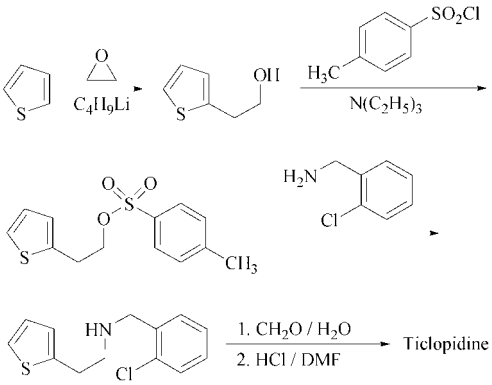

For further syntheses of ticlopidine, see [190, b], [194], [195].

Trade names (hydrochloride): Anagregal (Gentili), Antigreg (Piam), Aplaket (Rottapharm), Clox (Caber), Fluilast (Boniscontro e Gazzone), Klodin (Savio I.B.N.), Opteron (Therabel Pharma), Panaldine (Daiichi Seiyaku), Parsilid (Crinos), Ticlid (Sanofi Winthrop, Roche), Ticlopidina Dorom (Dorom), Tiklid (Sanofi Winthrop), Tiklyd (Sanofi Winthrop).

Clopidogrel, chemically related to ticlopidine, similarly inhibits the ADP receptor by selective binding, affects fibrinogen binding with about sixfold higher potency and additionally inhibits thromboxane mediated aggregation [196], [197]. After oral administration the substance is rapidly absorbed and undergoes metabolic activation in the liver. In experimental studies clopidogrel reduced the formation of both arterial and venous thrombi and was more effective than aspirin to prevent reocclusion after thrombolysis [198], [199]. However, in a large multicenter study (CAPRIE) in patients at risk of ischemic stroke, myocardial infarction, or vascular death, clopidogrel was only of slightly greater benefit than aspirin [200].

Clopidogrel [*113665-84-2*], methyl (+)-(*S*)-α-(2-chlorophenyl)-6,7-dihydrothieno[3,2-*c*]-pyridine-5(4*H*)-acetate, $C_{16}H_{16}ClNO_2S$, M_r 321.83; hydrogen sulfate [*135046-48-9*], *mp* 184 °C, $[\alpha]_D^{20}$ +55.10° (c = 1.891, methanol), can be prepared by condensing 4,5,6,7-tetrahydrothienol[3,2-*c*]pyridine with racemic methyl α-chloro-(2-chlorophenyl)acetate in the presence of potassium carbonate [192], [201] and optical resolution of the product obtained [202]:

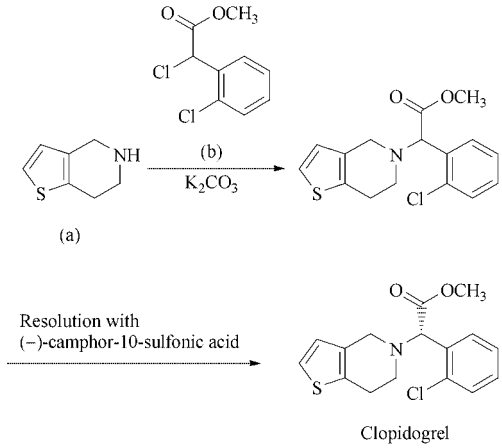

(see a: [190, b], [192], [191]; b: [203])

For another route to the product from the condensation step of the synthesis shown above, see [204].

Alternatively, clopidogrel can be obtained as shown below [205]:

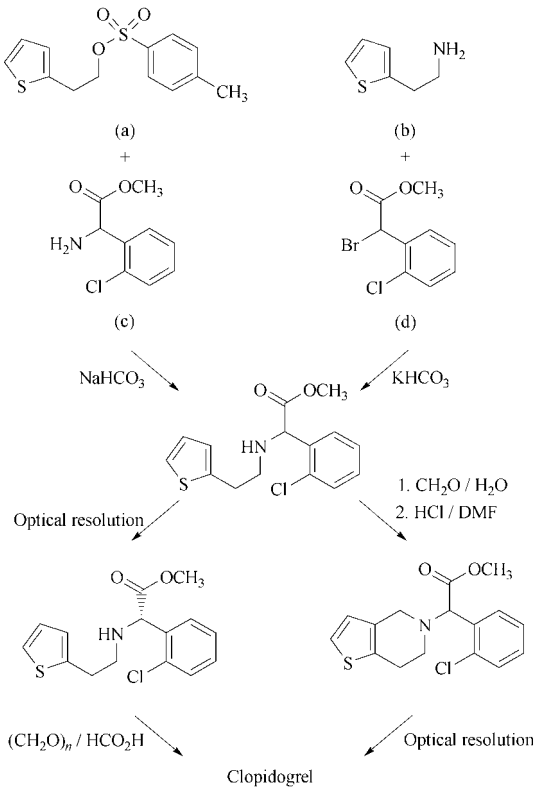

(see a: [193]; b:[192], [191], [207], [208]; c: [206]; d: [204])

For further possibilities to synthesize clopidogrel see [205].

Trade names (hydrogen sulfate): Iscover (Bristol-Myers Squibb), Plavix (Sanofi, Sanofi Winthrop, Sanofi Pharma Bristol-Myers Squibb SNC).

Abciximab is a mouse – human monoclonal antibody fragment which blocks reversibly the platelet GP IIb/IIIa receptor and thus inhibits platelet aggregation regardless of the trigger mechanism. A maximal antiplatelet effect of 80 % receptor blockade is achieved with intravenous bolus injection of 0.25 mg/kg followed by infusion of 0.125 $\mu g kg^{-1} min^{-1}$. After treatment platelet function recovers over several days, but the risk of bleeding is diminished within hours [209]. In large clinical studies abciximab has been investigated as an adjunct to aspirin and heparin in patients undergoing percutaneous coronary revascularization. It was shown that bolus plus infusion of the drug significantly reduced the incidence of ischemic complications (death, myocardial infarction, recurrent ischemia) and improved standard treatment with aspirin and heparin [210]. With respect to the bleeding risk abciximab was so far only tested in high-thrombotic-risk patients. At the moment several new orally active inhibitors of the GP IIb/IIIa receptors are under evaluation [132].

Abciximab [*143653-53-6*], 7E3, c7E3 Fab, immunoglobulin G (human – mousemonoclonal c7E3 clone p7E3V_Hh$C_{\gamma 4}$ Fab fragment anti-humanglycoprotein IIb/IIIa receptor), disulfide with human – mouse monoclonal c7E3 clone p7E3V_KhC_K light chain; for preparation, see [211].

Trade name: ReoPro (Lilly).

5.3.4. Calcium Entry Blockers

Calcium entry blockers (→ Calcium Antagonists) inhibit the transmembrane transport of calcium ions and thus the increase in cytoplasmic calcium ion concentration which is essential for cell activation. Such an action is established in heart muscle cells and smooth muscle cells [212]. In platelets, low concentrations of *verapamil* inhibit calcium uptake and subsequent intracellular epinephrine-induced calcium release and platelet aggregation. Higher concentrations of verapamil are necessary for the inhibition of ADP- and collagen-induced platelet aggregation. The anti-aggregatory effect of verapamil has been demonstrated in both animal experiments and humans [213], [214].

Nifedipine and *diltiazem* also have antiaggregatory activity [215], [216]. There are pronounced differences between these three blockers in their activity patterns against different types of aggregation. These differences are due to the existence of different types of calcium channels or of different ways by which calcium channels are opened [217], [218]. In humans, both verapamil and nifedipine inhibit the forma-

tion of TXA_2 in platelets parallel to the inhibition of aggregation [219]. The knowledge on the antithrombotic action of calcium channel blockers is still very fragmentary. For details on these drugs, see → Calcium Antagonists.

6. Blood Flow Affecting Drugs

Blood flow affecting drugs are compounds that are used for the treatment of occlusive arterial diseases that involve persistent underperfusion of the microcirculation. Such blood flow disorders are generally due to an increased flow resistance. This may be caused either by a decreased vascular cross-section – by passive mechanical obstacles (e.g., atherosclerotic and thrombotic processes) as well as by active vascular constriction (vascular spasm) – or by an increased viscosity of the blood. Compounds used for the remittence of atherosclerotic and thromboembolic processes, as well as for the elimination of vascular spasms are treated elsewhere; therefore, this section describes only compounds that influence the viscosity and hence the flow characteristics of the blood (blood rheology affecting drugs).

6.1. Factors Influencing Blood Viscosity

The viscosity of the blood is determined predominantly by the concentration of red blood cells (hematocrit), and their deformability [220]. Concentration and composition of the plasma proteins have less influence on the viscosity. If one standardizes the relative viscosity of water as 1, that of plasma is about 1.5, and that of blood with a hematocrit of 45 % is about 4.0.

Blood is a non-Newtonian fluid. Therefore, its viscosity is not constant, but is determined by the flow conditions in the vascular network. With decreasing flow rate, the viscosity increases disproportionally; in the case of stasis, viscosity values are so high that the blood loses its ability to flow. This increase in viscosity is the consequence of cross-linking of the red blood cells, which is brought about by fibrinogen, and the formation of rouleau-like aggregates.

At constant flow rate the blood viscosity depends on the vascular diameter. The viscosity falls off to a minimum with decreasing vascular diameter and attains a value that is only about 15 – 20 % higher than that of pure plasma under equivalent flow conditions. In capillaries the red blood cells thus contribute less to the viscosity than in the vessels of the macrocirculation. This optimization of flow dynamics is due to a decrease of the hematocrit in the capillaries to less than half of the systemic hematocrit (Fahraeus-Lindqvist effect) as well as to a deformation of the red blood cells. The high degree of flexibility of the cell membrane and the cell contents and the small cell volume in comparison to the cell surface area are prerequisites for the high degree of deformability of the red blood cells. This effect permits red blood cells to pass through vessels whose diameter is only 40 % that of the cell diameter.

Deformation and flow orientation are made possible through a permanent rotatory movement of the cell membrane around the fluid cell contents (tank track rotation). If the red blood cell loses its flexibility, it will be retained in the microcirculation and subsequently removed by phagocytes. This process normally occurs during the aging of red blood cells in the spleen.

Different factors lead to a loss of red blood cell flexibility, and thus can be of pathogenetic significance. These are acidosis, cell aggregation, change in shape, membrane stiffening, increase of intracellular hemoglobin concentration, and alteration of the surface-to-volume relationship. Hyperviscosity of the blood can moreover be caused by an increase in the hematocrit and an elevation of the protein concentration of the plasma [221]. Hemorheological factors are of pathogenetic significance in such hematological diseases as polycythemia and sickle cell anemia; these factors probably are also very significant in blood flow disorders such as cerebral ischemia, intermittent claudication, myocardial ischemia, and diabetogenic microangiopathy [222].

6.2. Testing of Flow Characteristics of Blood

The flow characteristics of the blood are almost exclusively tested in vitro; only a few experimental investigations under in vivo conditions have been carried out [223]. The clinically important parameters are total blood viscosity and

the extent of aggregation and deformability of the red blood cells. Because of the particular flow dynamics of blood in vivo, investigation of the blood in the rotational viscosimeter is more informative than investigation in the capillary viscosimeter [224], [225]. The following parameters considerably influence the test results: blood sampling technique, type and dosage of anticoagulants, hematocrit, temperature, and age of the blood sample.

One of the ways by which the dynamics of red blood cell aggregation can be assessed is in the rheoscope by determination of the re-aggregation after hydrodynamic dispersion [226].

The flexibility of the red blood cells is preferably determined by filtration of the red blood cell suspension at constant pressure through a pore membrane; the pores of the membrane have a smaller diameter than those of the red blood cells, and their thickness and dimensions must be clearly defined [227].

6.3. Blood Rheology Affecting Drugs

At present the rheological characteristics of the blood can be improved pharmacologically in three ways: by hemodilution, by a decrease of the plasma fibrinogen level, and by optimization of the rheological characteristics of the red blood cells.

6.3.1. Hemodilution

Hemodilution is accomplished through intravenous infusion of crystalloid solutions containing plasma expanders, mostly low molecular mass dextran or human albumin (→ Blood; → Dextran). Depending on the state of cardiac and renal sufficiency, the infusion is carried out in an isovolemic manner (with simultaneous blood withdrawal) or a hypervolemic manner (without blood withdrawal). Persisting hemodilution therapy is only possible by continuous infusion during several days. The blood viscosity is decreased by this measure because the red blood cell count is reduced (decrease in hematocrit of up to 30 %). The blood supply of organs is clearly improved by this method. Such a therapy is effective in acute cerebral ischemia and in peripheral arterial blood supply disorders [228].

6.3.2. Defibrinogenation

The plasma fibrinogen level can be specifically decreased by the parenteral administration of the snake venom enzymes *ancrod* and *batroxobin* [229], [230]. These proteolytic enzymes split off various fibrinopeptides (A, AP, and AY) from fibrinogen, which subsequently polymerize to an unstable form of fibrin that is removed from the circulation by the reticuloendothelial system. This results in a dose-dependent hypofibrinogenemia leading to a decrease in blood viscosity; a decrease in red blood cell aggregation contributes substantially to this effect. As a consequence of the fibrinogen deficiency, platelet aggregation is also diminished [231].

In humans the plasma fibrinogen level begins to fall shortly after the subcutaneous administration of the enzymes and reaches a minimum after some hours. Enteral application has no effect. A decrease of the fibrinogen level down to 70 mg/100 mL is therapeutically desirable; a reduction beyond this value increases the risk of bleeding. To maintain the action, the drug must be administered once or twice a day. After drug withdrawal the plasma fibrinogen level increases slowly and reaches the initial level after several days. With prolonged use the activity of the drug is lost due to the formation of antibodies, because both ancrod and batroxobin are antigenically active. The treatment of humans should therefore not exceed 7 weeks. If resistance occurs during treatment, the action can be maintained by switching from one enzyme to the other because the two enzymes are immunologically different (no cross-reaction). Ancrod and batroxobin are active in different obstructive arterial disorders [229], [232], [233]. Studies in patients with acute ischemic stroke showed that ancrod is safe and effective (more likely than placebo) to achieve functional recovery [234].

Ancrod [*9046-56-4*], *Agkistrodon* serine proteinase, *Agkistrodon rhodostoma* venom proteinase, is a defibrinogenatingenzyme derived from the venom of *Agkistrodon* (new name: *Calloselasma*) *rhodostoma* (Boie), the Malayan pit viper. For its separation from the crude venom, see [235].

The enzyme is a glycopolypeptide with a high carbohydrate content (36 wt %) and a molecular

mass of about 35 400. The polypeptide consists of 18 amino acids [235, a].

When stored at 4 °C either in the lyophilized state (light powdery texture [236]) or as a dilute solution in physiological saline, ancrod is stable for very long periods. At 37 °C its activity is decreased to 50 % within 4 – 6 h [229].

Trade name: Viprinex (Knoll Pharma Inc.).

Batroxobin [*9039-61-6*], *Bothrops* venom proteinase, is a defibrinogenating enzyme derived from the venom of *Bothropsatrox,* a pit viper found in various species in South and Central America [230]. The properties of the agent vary with the subspecies of *Bothrops atrox* from which it is isolated [230], [237, b]. Batroxobin used in therapy is prepared from the subspecies *moojeni* [230]. For the purification of batroxobin from the crude venom of *Bothrops atrox moojeni,* see [237–240].

The enzyme from the venom of *Bothrops atrox moojeni* is a glycopolypeptide with a high carbohydrate content (27 wt %) and a molecular mass of about 29 000. The polypeptide consists of 18 amino acids [237, b].

When stored at 4 °C as a dilute solution in physiological saline, batroxobin is stable for more than 2 years. There is no significant loss in activity if batroxobin is heated in glycerol for 1 h at 100 °C [230].

Trade names: Botropase (Ravizza), Reptilase (Difa-Cooper, Pharma-développement).

6.3.3. Rheological Improvement of Red Blood Cells

Pentoxifylline reduces the viscosity of blood predominantly through an attack on the red blood cells [241]. The flexibility of the red cells is improved, especially in patients with peripheral blood flow disorders. This effect is probably linked to the observed increase in energy-rich phosphate (ATP) in the cell. In addition, the level of fibrinogen in the plasma decreases, which lowers the tendency of the red cells to aggregate [242]. The following additional effects probably contribute to the therapeutic efficacy of the substance: fibrinolytic action [242], inhibition of platelet aggregation, and stimulation of prostacyclin formation [243]. Pentoxifylline has been found to be active in numerous peripheral blood flow diseases [244].

Pentoxifylline [*6493-05-6*], 3,7-dihydro-3,7-dimethyl-1-(5-oxohexyl)-1*H*-purine-2,6-dione, 1-(5-oxohexyl)theobromine, $C_{13}H_{18}N_4O_3$, M_r 278.31, *mp* 105 °C, can be obtained by the methods shown below [245], [246].

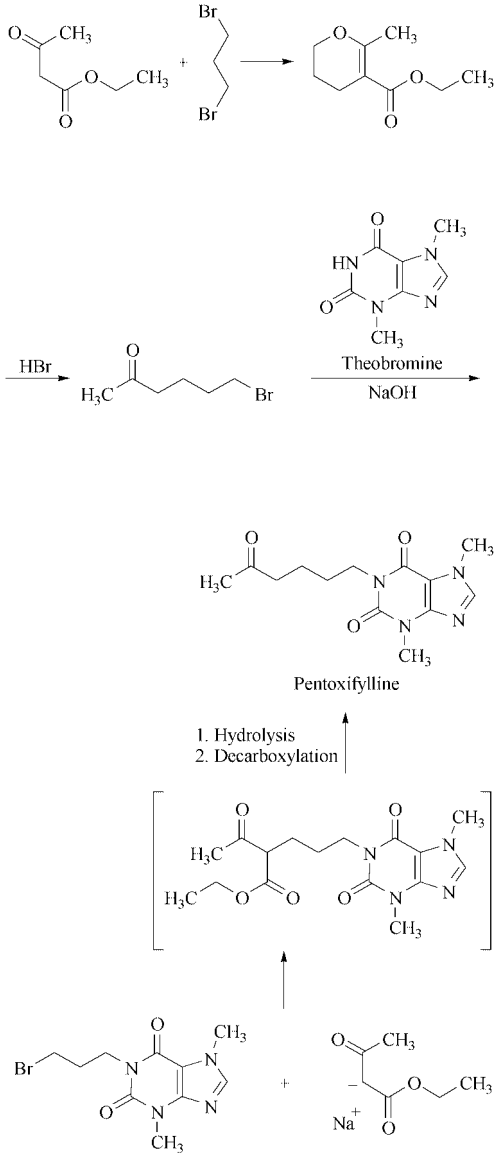

(see a: [247])

For other syntheses, see [248–250].

Trade names:Agapurin (medphano), Azupentat (Azupharma), Claudicat (Byk Gulden), durapental (Merck dura), Pento AbZ (AbZ Pharma), Pentoflux (Bouchara), Pentohexal (Hexal), PENTO-PUREN (Isis Puren), pentox von ct (ct-Arzneimittel), Ralofekt (Arzneimittelwerk Dresden, Temmler Pharma), Rentylin (FUISZ PHARMA), Torental (Hoechst), Trental (Hoechst Marion Roussel).

Buflomedil is able to restore the pathologically altered tissue circulation and thus normalize the oxygen supply under experimental conditions [251]. This effect probably depends on (1) a vascular action in the microcirculatory region of the affected tissue and (2) an improved fluidity of the blood caused by increasing the flexibility of the red cells and by inhibiting platelet aggregation [252], [253]. A decrease of oxygen consumption in the tissue caused by this drug has so far only been demonstrated experimentally [254]. In a multicenter study, the value of this substance in obstructive arterial diseases has been established [255].

Buflomedil [*55837-25-7*], 4-(1-pyrrolidinyl)-1-(2,4,6-trimethoxyphenyl)-1-butanone, $C_{17}H_{25}NO_4$, M_r 307.39; hydrochloride [*35543-24-9*], *mp* 192 – 193 °C. Buflomedil is made by treating 1,3,5-trimethoxybenzene with 4-(1-pyrrolidinyl)butyronitrile in the presence of HCl (Houben – Hoesch reaction) [256]:

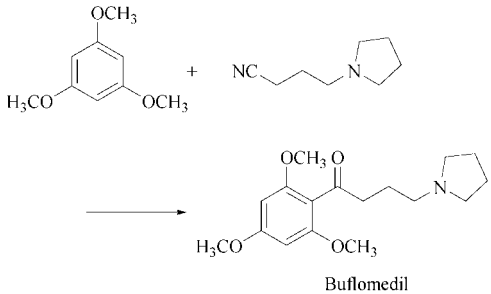

Buflomedil can also be synthesized by reacting 1,3,5-trimethoxybenzene with 4-chlorobutyryl chloride in the presence of zinc chloride (Friedel – Crafts acylation) and treating the resulting product with pyrrolidine [257]:

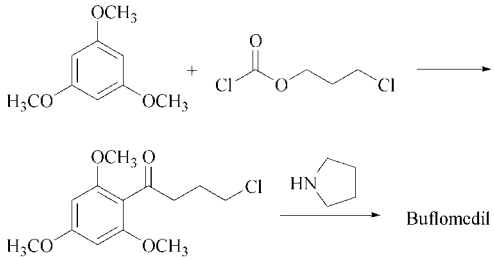

Trade names (hydrochloride): Bufedil (Abbott), Bufene (Ist. Chim. Inter.), Buflan (Pierrel), Buflo AbZ (AbZ-Pharma), BUFLO-PUREN (Isis Puren), Buflocit (CT), Buflofar (Farge), Buflohexal (Hexal), Defluina (Rhône-Poulenc Rorer), Emoflux (Metapharma), Flomed (Pulitzer), Fonzylane (Lafon), Irrodan (Biomedica Foscama), Loftyl (Abbott), Medil (Crosara), Perfudan (Piam).

7. References

1. *Atherosclerosis* IV, Springer Verlag, Berlin-Heidelberg-New York 1977, pp. XXI – XXXVIII.
2. W. B. Kannel, *JAMA J. Am. Med. Assoc.* **247** (1982) 877 – 880.
3. R. Cooper, *N. Engl. J. Med.* **304** (1981) 1259 – 1265.
4. St. N. Willich, H. Löwel, W. Mey, Chr. Trautner, *Deutsches Ärzteblatt* **96** (1999) B378 – B383.
5. M. Hollmann, *Med. Klin. (Munich)* **77** (1982) 597 – 602.
6. U. Schwabe, D. Paffrath, *Arzneiverordnungsreport 1998,* Springer-Verlag, Berlin – Heidelberg 1999.
7. J. W. Hay, W. M. Yu, T. Ashraf, *Pharmacoeconomics* **15** (1999) 47 – 74.
8. A. Keys, *Atherosclerosis (Shannon, Irel.)* **23** (1975) 149 – 192.
9. J. Stamler, *Circulation* **58** (1978) 3 – 19.
10. W. B. Kannel, W. P. Castelli, T. Gordon, P. M. McNamara, *Ann. Intern. Med.* **74** (1971) 1 – 12.
11. D. S. Fredrickson, R. I. Levy, R. S. Lees, *N. Eng. J. Med.* **276** (1967) 34 – 44, 94 – 103, 148 – 156, 215 – 225, 273 – 281.
12. A. Keys, *Acta Med. Scand.* **207** (1980) 153 – 160.
13. S. H. Epstein, *Ther. Umsch* . 37 (1980) 947 – 953.

14. S. H. Epstein, S. Gutzwiller, H. Howald, W. Junod, W. Schweizer, *Schweiz. Med. Wochenschr.* **109** (1979) 1171 – 1180.
15. M. S. Oliver, J. A. Heady, J. N. Morris, J. Copper, *Lancet* 1980, 379 – 385.
16. The Multiple Risk Factor Intervention Trial (MRFIT), *JAMA J. Am. Med. Assoc.* **235** (1976) 825 – 837.
17. The Multiple Risk Factor Intervention Trial Research Group, *JAMA J. Am. Med. Assoc.* **248** (1982) 1465 – 1477.
18. The Multiple Risk Factor Intervention Trial Research Group, *JAMA J. Am. Med. Assoc.* **263** (1990) 1795 – 1801.
19. G. Schettler, *Deutsches Ärzteblatt* **88** (1991) B1580 – B1583.
20. J.-L. Rouleau et al., *J. Am. Coll. Cardiol.* **1** (1983) no. 6, 1453 – 1460.
21. J. Östergen: "The role of ACE inhibitors in the prevention and treatment of atherosclerosis," *Int. J. Clin. Pract. Suppl.* **94** (1998) May.
22. R. R. Malinow, *Circulation* **64** (1981) 1 – 3. G. Noseda, C. Fragiacomo, R. Fumagalli, R. Paoletti (eds.): *Lipoproteins and Coronary Atherosclerosis,* Elsevier Biomedial Press, Amsterdam-New York-Oxford 1982.
23. B. M. Rifkind, Lipid Research Clinics Program, *JAMA J. Am. Med. Assoc.* **251** (1984) 351 – 364.
24. M. H. Frick et al., *N. Engl. J. Med.* **317** (1987) 1237 – 1245.
25. J. P. Kane et al., *JAMA J. Am. Med. Assoc.* **264** (1990) 3007 – 3012.
26. J. Shepherd et al., *N. Engl. J. Med.* **333** (1995) 1301 – 1307.
27. Scandinavian Simvastatin Survival Study Group, *Lancet* **344** (1994) 1383 – 1389.
28. D. H. Blankenhorn et al., *Ann. Int. Med.* **119** (1993) 969 – 976.
29. F. M. Sacks et al., *N. Engl. J. Med.* **335** (1996) 1001 – 1009.
30. M. H. Davidson, FLUENT Investigators Group, *Am. J. Med.* **96** (1994) Suppl. 6A, 41S – 44S.
31. D. Kritchevsky (ed.): "Hypolipidemic Agents," *Handbook of Experimental Phamacology,* **vol. 41,** Springer Verlag, Berlin-Heidelberg-New York 1975.
32. R. Paoletti, C. J. Glueck (eds.): *Lipid Pharmacology,* **vol. 2,** Academic Press, New York-San Francisco-London 1976.
33. M. S. Brown, J. L. Goldstein in A. Goodman Gilman, W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 874.
34. S. D. Larsen, C. H. Spilman, *Annu. Rep. Med. Chem.* **28** (1993) 217.
35. H. N. Ginsberg, Y. Arad, I. J. Goldberg in M. Antonaccio (ed.): *Cardiovascular Pharmacology,* 3rd ed., Raven Press, New York 1990, p. 485.
36. K. Berg, *Acta Path. Microbiol. Scand.* **59** (1963) 369 – 382.
37. A. Murai, T. Miyahaya, N. Fujimoto, M. Matsuda, M. Kameyama, *Atherosclerosis* **59** (1986) 199 – 204.
38. J. L. Goldstein, M. S. Brown, *Ann. Rev. Biochem.* **46** (1977) 897.
39. J. L. Goldstein, T. Kita, M. S. Brown, *New Engl. J. Med.* **309** (1983) 288. J. B. Stanbury, J. B. Wyngaarden, D. S. Fredrickson (eds.): *The Metabolic Basis of Inherited Disease,* McGraw Hill, New York 1982.
40. "Lipid-Liga-Empfehlungen", *Deutsches Ärzteblatt – Sonderbeilage* (1996) Heft 4.
41. D. Kritchevsky in D. Kritchevsky (ed.): "Hypolipidemic Agents," *Handbook of Experimental Phamacology,* **vol. 41,** Springer Verlag, Berlin–Heidelberg–New York 1975, p. 216.
42. E. Granzer in G. Erhart, H. Ruschig (eds.): *Arzneimittel,* **vol. 2,** Verlag Chemie, Weinheim 1972, p. 421.
43. C. E. Day, P. E. Schurr, D. E. Emmert, R. E. TenBrink et al., *J. Med. Chem.* **18** (1975) 1065.
44. D. M. Bailey, J. D. Prugh, C. S. Rooney, R. L. Smith in H. J. Hess (ed.): *Annual Reports in Medicinal Chemistry,* **vol. 18,** Academic Press, New York-London 1983, p. 161.
45. Merck Sharp & Dohme, US 3 383 281, 1968.
46. Upjohn Comp., DE-OS 1 927 336, 1969 (N. A. Nelson, G. E. Van den Berg); DE-OS 2 053 585, 1971 (D. Lednicer, C. Y. Peery).
47. *The Merck Index,* 12th ed., Merck & Co., 1996, 8697.
48. S. Bernstein, E. S. Wallis, *J. Org. Chem.* **2** (1938) 341.
49. H. Nahm, W. Siedel, *Chem. Ber.* **96** (1963) 1.
50. P. Tornvall, G. Walldius, *J. Int. Med.* **230** (1991) 415.
51. A. Nenz, M. Pieroni, *Hydrocarbon Process.* **47** (1968) 139.
52. Lonza, DE-OS 2 046 556, 1971 (A. Stocker et al.).
53. Wuelfing, DE 1 102 750, 1961 (W. Bestian); *Chem. Abstr.* **56** (1962) 11602.

54. C. O. Badgett, C. F. Woodward, *J. Am. Chem. Soc.* **69** (1947) 2907.
55. V. Ambrogi et al., *Eur. J. Med. Chem.* **15** (1980) 157.
56. J. M. Thorp, W. S. Waring, *Nature (London)* **194** (1962) 948.
57. J. P. Desager, Y. Horsmans, C. Vandenpla et al., *Atherosclerosis* **124** (1996) Suppl., S65 – S73.
58. D. Stahlberg, *Cardiovascular Drug Reviews* **10** (1992) 259 – 279.
59. D. M. Bailey, M. N. Cayen, M. A. Kallai-Sanfacon in H. J. Hess (ed.): *Annual Reports in Medicinal Chemistry,* **vol. 15,** Academic Press, New York-London 1980, p. 162.
60. P. Galimberti, A. Defrancesci, *Gazz. Chim. Ital.* **77** (1947) 431.
61. M. Julia, M. Ballarge, G. Tchernoff, *Bull. Soc. Chim. Fr.* 1956, 777.
62. Imperial Chemical Industries, US 3 262 850, 1966 (W. G. Mors, J. M. Thorp, W. S. Waring).
63. Yoshitomi Pharmaceutical Industries, NL-A 6 600 044, 1966.
64. R. Sornay, J. Gurrieri, C. Tourne, R. F. Renson et al., *Arzneim. Forsch.* **26** (1976) 885.
65. Boehringer Mannheim, DE-OS 2 149 070, 1973 (E. C. Witte, K. Stach, M. Thiel, F. Schmidt et al.).
66. Parke, Davis and Co., DE-OS 1 925 423, 1969 (P. L. Creger).
67. Warner Lambert Comp., US 4 126 637, 1978 (O. P. Goel, W. M. Pearlman).
68. Merz & Co., DE-OS 1 941 217, 1971 (A. Scherm, D. Peteri).
69. Merckle, DE 2 308 826, 1974 (G. Metz, M. Specker); *Chem. Abstr.* **82** (1975) 4310.
70. Sterling, DE 2 343 606, (D. K. Phillips); *Chem. Abstr.* **80** (1974) 133048.
71. T. E. Carew, G. C. Schwenke, D. Steinberg, *Proc. Nat. Acad. Sci. USA* **84** (1987) 7725 – 7729.
72. M. B. Neuworth, R. J. Laufer, J. W. Barnhart, J. A. Sefranka et al., *J. Med. Chem.* **13** (1970) 722.
73. *Drugs of the Future* **3** (1978) 662.
74. A. Endo, M. Kuroda, Y. Tsujita, *J. Antibiot.* **29** (1976) 1346.
75. A. W. Alberts, J. Chen, G. Kuron, V. Hunt et al., *Proc. Natl. Acad. Sci. USA* **77** (1980) 3957.
76. A. J. Endo, *J. Antibiot.* **32** (1979) 852; Sankyo, DE 3 006 215, 1980 (Y. Tsujita, K. Tanazawa, K. Furuya, K. Masao, S. Iwado); *Chem. Abstr.* **94** (1981) 119471.
77. Merck, EP 33 538, 1981 (A. K. Willard, R. L. Smith, W. F. Hoffman); *Chem. Abstr.* **95** (1981) 219 968.
78. W. F. Hoffman et al., *J. Med. Chem.* **29** (1986) 849.
79. A. Endo, *J. Med. Chem.* **28** (1985) 401.
80. Sankyo, DE 3 122 499, 1981 (A. Terahara, M. Tanaka); *Chem. Abstr.* **97** (1982) 4664.
81. Sandoz, WO 8 402 131, 1984 (F. G. Kathawala); *Chem. Abstr.* **102** (1984) 24475.
82. Warner-Lambert, EP 409 281, 1991 (B. D. Roth); *Chem. Abstr.* **115** (1991) 29107; K. L. Baumann et al., *Tetrahedron Lett.* **33** (1992) 2283.
83. Bayer, DE 4 040 026, 1992 (R. Angerbauer et al.); *Chem. Abstr.* **117** (1992) 131084.
84. Bayer, DE 617 019, 1994 (R. Angerbauer, R. Grosser, W. Hinsken, J. Rehse); *Chem. Abstr.* **121** (1994) 280555.
85. A. Yamamoto, H. Sudo, A. Endo, *Atherosclerosis* **35** (1980) 259.
86. J. A. Tobert, G. Hitzenberger, W. R. Kukovetz, I. B. Holmes et al. *Atherosclerosis* **41** (1982) 61.
87. Antiplatelet Trialists′ Collaboration, *Br. Med. J.* **308** (1994) 81, 159, 235.
88. a) J. L. Gordon in J. L. Gordon (ed.): *Platelets in Biology and Pathology,* Elsevier/North-Holland Biomedical Press, Amsterdam 1981, p. 1. b) D. Ogston: *The Physiology of Hemostasis,* Croom Helm, London 1983.
89. D. C. B. Mills, G. C. K. Roberts, *J. Physiol. (London)* **193** (1967) 443.
90. M. C. Scrutton, R. B. Wallis in J. L. Gordon (ed.): *Platelets in Biology and Pathology,* Elsevier/North-Holland Biomedical Press, Amsterdam 1981, p. 179.
91. G. H. van Zanten, P. G. de Groot, J. J. Sixma in F. von Bruchhausen, U. Walter (eds.): *Platelets and their factors,* Springer-Verlag Berlin 1997, p. 61.
92. E. F. Plow, M. H. Ginsberg, *Prog. Hemost. Thromb.* **9** (1989) 117.
93. M. M. Billah, E. G. Lapetina, P. Cuatrecasas, *J. Biol. Chem.* **256** (1981) 5399. E. G. Lapetina, *Life Sci.* **32** (1983) 2069.
94. F. A. O'Rourke, S. P. Halenda, G. B. Zavoico, M. B. Feinstein, *J. Biol. Chem.* **260** (1985) 956.
95. D. E. MacIntyre, M. Bushfield, A. M. Shaw, *Fed. Europ. Biochem. Soc.* **188** (1985) 383.
96. D. E. MacIntyre, A. M. Shaw, M. Bushfield, L. J. MacMillan, A. McNicol, W. K. Pollock, *Nouv. Rev. Fr. Hematol.* **27** (1985) 285.

97. P. Massini, R. Käser-Glanzmann, E. F. Lüscher, *Thromb. Haemostasis* **40** (1978) 212.
98. J. M. Gerrard, D. A. Peterson, J. G. White in J. L. Gordon (ed.): *Platelets in Biology and Pathology,* Elsevier/North-Holland Biomedical Press, Amsterdam 1981, p. 407.
99. M. B. Feinstein, J. J. Egan, R. I. Sha'afi, J. White, *Biochem. Biophys. Res. Commun.* **113** (1983) 598.
100. Y. Denizot in F. von Bruchhausen, U. Walter (eds.): *Platelets and their factors,* Springer-Verlag Berlin 1997, p. 483.
101. A. J. Marcus, *J. Lipid Res.* **19** (1978) 793.
102. K. Schrör: *Prostaglandine und verwandte Verbindungen,* G. Thieme, Stuttgart 1984.
103. B. J. R. Whittle, S. Moncada, *Br. Med. Bull.* **39** (1983) 232.
104. R. Ross, *New Engl. J. Med.* **314** (1986) 488.
105. H. N. Antoniades, L. T. Williams, *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **42** (1983) 2630.
106. R. Ross, *Arzneim. Forsch./Drug Res.* **33** (1983) 1399.
107. R. Ross, *J. Cardiovasc. Pharm.* **7 (Suppl. 3)** (1985) S 186.
108. J. H. Chesebro, I. P. Clements, V. Fuster, L. R. Elveback, H. C. Smith, W. T. Bardsley et al., *New Engl. J. Med.* **307** (1982) 73.
109. V. Fuster, M. K. Dewanjee, M. P. Kaye, M. Josa, M. P. Metke, J. H. Chesebro, *Circulation* **60** (1979) 1508.
110. K. K. Wu, J. C. Hoak, *Lancet* 1974, II, 924.
111. F. E. Preston, J. D. Ward, B. Marcola, N. R. Porter, W. R. Timparkey, B. C. O'Malley, *Lancet* 1978, I, 238.
112. J. W. Ten Cate, J. Vos, H. Oosterhuis, D. Prenger, C. S. P. Jenkins, *Thromb. Haemostasis* **39** (1978) 223.
113. H. J. Dougherty jr., D. E. Levy, B. B. Weksler, *Lancet* 1977, I, 821.
114. J. Mehta, P. Mehta, *Am. J. Cardiol.* **47** (1981) 331.
115. A. Maseri, A. L'Abbate, G. Baroldi, S. Chierchia, M. Marzili, A. M. Bellestra et al., *New Engl. J. Med.* **299** (1978) 1271.
116. L. A. Harker, J. L. Ritchie, *Circulation* **62** (1980) V-13.
117. A. M. Lefer in R. McConn (ed.): *Role of Chemical Mediators in the Pathophysiology of Acute Illness and Injury,* Raven Press, New York 1982, p. 101.
118. H. R. Baumgartner, *Schweiz. Med. Wochenschr.* **104** (1974) 109.
119. G. V. R. Born, *Nature (London)* **194** (1962) 927.
120. K. Breddin, H. Grun, H. J. Krzywanek, W. P. Schremmer, *Thromb. Haemostasis* **35** (1976) 669.
121. P. Gresele, C. Zoja, H. Deckmyn, J. Arnout, J. Vermylen, M. Verstraete, *Thromb. Haemost.* **50** (1983) 852.
122. R. T. Owen, *Med. Actual.* **15** (1979) 487.
123. J. G. Milton, M. M. Frojmovic, *J. Pharmacol. Methods* **9** (1983) 101.
124. H. Patscheke, *Haemostasis* **8** (1979) 65.
125. J. Kloeze, *Thromb. Diath. Haemorrh.* **23** (1970) 286.
126. M. Freund et al., *Thromb. Haemost.* **69** (1993) 515.
127. S. L. Cooper, D. J. Fabrizius, T. G. Grasel, *Ann. N.Y. Acad. Sci.* **516** (1987) 572.
128. R. B. Philp, *Methods Find. Exp. Clin. Pharmacol.* **1** (1979) 197.
129. K. Meng, K. O'Dea, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **283** (1974) 379.
130. K. Meng, F. Seuter, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **301** (1977) 115.
131. V. Fuster, M. K. Dewanjee, M. P. Kaye, M. Josa, M. P. Metke, J. H. Chesebro, *Circulation* **60** (1979) 1508.
132. K. Schrör, *Drugs* **50** (1995) 7.
133. G. J. Roth, P. W. Majerus in M. J. Silver, J. B. Smith, J. J. Kocsis (eds.): *Prostaglandins in Hematology,* Spectrum Publications Inc., New York 1977, p. 345.
134. R. Clarke, G. Mayo, P. Price, G. A. FitzGerald, *N. Engl. J. Med.* **325** (1991) 1137.
135. G. Masotti, G. Galanti, L. Poggesi, R. Abbate, G. G. Neri Serneri, *Lancet* 1979, II, 1213.
136. E. F. Ellis, K. F. Wright, P. S. Jones, D. W. Richardson, C. K. Ellis, *J. Cardiovasc. Pharmacol.* **2** (1980) 387.
137. C. Patrono, G. Ciabattoni, E. Pinca, F. Pugliese et al., *Thromb. Res.* **17** (1980) 317.
138. M. A. Peckham, J. F. Mustard in E. Margulies (ed.): *Myocardial Infarction and Cardiac Death,* Academic Press, New York 1983, p. 93.
139. J. C. Hoak, *Thromb. Res.,* Suppl. IV (1983) 47.
140. J. J. Thiessen, *Thromb. Res.,* Suppl. IV (1983) 105.
141. J. E. Willard, R. A. Lange, L. D. Hillis, *N. Engl. J. Med.* **327** (1992) 175.
142. D. B. Matchar et al., *Ann. Intern. Med.* **121** (1994) 41.
143. "ISIS-2," *Lancet* **2** (1988) 349.
144. L. J. Grimm, D. R. Knapp, D. Senator, P. V. Halushka, *Thromb. Res.* **24** (1981) 307.
145. B. Hoet, C. Falcon, S. de Reys, J. Arnout, H. Deckmyn, J. Vermylen, *Blood* **75** (1990) 646.

146. The RAPT Investigators, *Circulation* **89** (1994) 588.
147. Janssen Pharmaceutica, EP 221 601, 1986 (E. J. E. Freyne, A. H. M. Raeymaekers, V. Sipido, M. G. Venet).
148. B. J. R. Whittle, S. Moncada, J. R. Vane, *Prostaglandins* **16** (1978) 373.
149. H. S. Mueller, P. S. Rao, M. A. Greenberg et al., *Circulation* **72** (1985) 1336.
150. K. Jaschonek, K. R. Karsch, H. Weissenberger et al., *J. Am. Coll. Cardiol.* **8** (1986) 259.
151. H. Sinzinger, K. Silberbauer, A. K. Horsch, A. Gall, *Prostaglandins* **21** (1989) 49.
152. P. J. Lewis, C. T. Dollery, *Br. Med. Bull.* **39** (1983) 281.
153. A. H. Gershlick, D. Spriggins, S. W. Davies et al., *Br. Heart J.* **71** (1994) 7.
154. M. J. Cho, M. A. Allen, *Prostaglandins* **15** (1978) 943.
155. J. R. Vane, *Angew. Chem.* **95** (1983) 782; *Angew. Chem. Int. Ed. Engl.* **22** (1983) 741.
156. K. C. Nicolaou, W. E. Barnette, R. L. Magolda, *J. Chem. Res. Synop.* 1979, 202; *J. Chem. Res. Miniprint* 1979, 2444.
157. E. J. Corey, G. E. Keck, I. Székely, *J. Am. Chem. Soc.* **99** (1977) 2006.
158. K. C. Nicolaou, G. P. Gasic, W. E. Barnette, *Angew. Chem.* **90** (1978) 360; *Angew. Chem. Int. Ed. Engl.* **17** (1978) 293.
159. I. Tömösközi, G. Galambos, V. Simonidesz, G. Kovács, *Tetrahedron Lett.* 1977, no. 18, 2627.
160. N. Whittaker, *Tetrahedron Lett.* 1977, no. 18, 2805.
161. R. A. Johnson, F. H. Lincoln, E. G. Nidy, W. P. Schneider, J. L. Thompson, U. Axen, *J. Am. Chem. Soc.* **100** (1978) 7690.
162. B. J. R. Whittle, S. Moncada, F. Whiting, J. R. Vane, *Prostaglandins* **19** (1980) 605.
163. K. Schrör, H. Darius, R. Matzky, R. Ohlendorf, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **316** (1981) 252.
164. W. W. Nickols, F. A. Nicolini, S. Khan et al., *Am. Heart J.* **126** (1993) 285.
165. Schering, DE-OS 2 845 770, 1978 (W. Skuballa, B. Radüchel, H. Vorbrüggen, G. Mannesmann, W. Losert, J. Casals).
166. W. Skuballa, H. Vorbrüggen, *Angew. Chem.* **93** (1981) 1080; *Angew. Chem. Int. Ed. Engl.* **20** (1981) 1046.
167. W. Skuballa, H. Vorbrüggen, *Adv. Prostaglandin, Thromboxane, Leukotriene Res.* **11** (1983) 299.
168. P. R. Emmons, M. G. J. Harrison, A. J. Honour, J. R. A. Mitchell, *Nature (London)* **208** (1965) 255.
169. H. J. Weiss, *New Engl. J. Med.* **298** (1978) 1344.
170. P. Gresele, C. Zoja, H. Deckmyn, J. Arnout, J. Vermylen, M. Verstraete, *Thromb. Haemostasis* **50** (1983) 852.
171. L. A. Harker, R. A. Kadatz, *Thromb. Res.,* Suppl. IV (1983) 39.
172. J. F. Mustard, M. A. Packham in G. S. Avery (ed.): *Antithrombotic Drugs* (vol. 3 of Cardiovascular Drugs), ADIS Press, New York 1978, p. 1.
173. R. L. Kinlough-Rathbone, H. M. Groves, J. P. Cazenave, M. Richardson, J. F. Mustard, *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **37** (1978) 260.
174. M. J. Silver, C. Ingerman-Wojenski, A. W. Sedar, *Pharmacologist* **25** (1983) 134.
175. L. A. Harker, S. J. Slichter, *New Engl. J. Med.* **287** (1972) 999.
176. A. Dembinska-Kièć, W. Rücker, P. S. Schönhöfer, *Atherosclerosis* **33** (1979) 315.
177. Persantine-Aspirin Reinfarction Study Research Group, *Circulation* **62** (1980) 449.
178. H. Diener, L. Cuhla, C. Forbes, J. Sivenius, P. Smets, A. Lowenthal, *J. Neurolog. Sci.* **143** (1996) 1.
179. Thomae, DE 1 116 676, 1955 (G. Fischer, J. Roch, A. Kottler).
180. Yamanouchi, DE-AS 1 962 261, 1969 (M. Murakami et al.).
181. S. Ashida, Y. Abiko, *Thromb. Haemostasis* **41** (1979) 436.
182. J. J. Bruno, *Thromb. Res.,* Suppl. IV (1983) 59.
183. G. Di Minno, A. M. Cerbone, P. L. Mattioli, S. Turco, C. Iovine, M. Mancini, *J. Clin. Invest.* **75** (1985) 328.
184. A. H. Reece, P. L. Walton, *Thromb. Haemostasis* **42** (1979) 367.
185. J. R. O'Brien, M. D. Etherington, R. D. Shuttleworth, *Thromb. Res.* **13** (1978) 245.
186. J. L. David, F. Monfort, F. Herion, R. Raskinet, *Thromb. Res.* **14** (1979) 35.
187. M. Gent et al., the CATS Group, *Lancet* **1** (1989) 1215.
188. W. K. Hass et al., the Ticlopidine Aspirin Stroke Study Group, *N. Engl. J. Med.* **321** (1989) 501.
189. a) Centre d'Études pour l'Industrie Pharmaceutique, DE-AS 2 404 308, 1974. b) J.-P. Maffrand, F. Eloy, *Eur. J. Med. Chem. – Chim. Ther.* **9** (1974) 483.
190. a) Parcor, DE 2 530 516, 1975. b) J.-P. Maffrand, F. Eloy, *J. Heterocycl. Chem.* **13** (1976) 1347.
191. Sanofi, EP 522 956, 1992 (P. J. Harrington, I. H. Sanchez).

192. Sanofi, EP 465 358, 1991 (A. Bousquet, S. Calet, A. Heymes).
193. Parcor, US 4 127 580, 1977 (E. Braye); see also: Parcor, US 4 174 448, 1978 (A. Bousquet, E. Braye).
194. J.-P. Maffrand, R. Boigegrain, *Heterocycles* **12** (1979) 1479.
195. K. Satake, T. Imai, M. Kimura, S. Morosawa, *Heterocycles* **16** (1981) 1271.
196. S. K. Yao et al., *Circ. Res.* **70** (1992) 39.
197. A. J. Coukell, A. Markham, M. Gent, K. Schrör, *Drugs* **54** (1997) 745.
198. S. K. Yao et al., *Am. J. Physiol.* **267** (1994) H488.
199. G. Feuerstein, A. J. Nichols, R. R. Ruffolo jr., *Expert Opin. Invest. Drugs* **4/5** (1995) 425.
200. CAPRIE Steering Committee, *Lancet* **348** (1996) 1329.
201. Sanofi, EP 099 802, 1983 (D. Aubert, C. Ferrand, J.-P. Maffrand).
202. Sanofi, EP 281 459, 1988 (A. Badorc, D. Frehel).
203. Synthesis described in [192] ; for the starting material used in that reference, see FR 2 292 706, 1974 (G. Carraz).
204. Sanofi, EP 420 706, 1990 (M. Bouisset, J. Radisson).
205. Sanofi, EP 466 569, 1991 (M. Descamps, J. Radisson).
206. Synthesis described in [205] ; for the starting material used in that reference, see Bristol-Myers, GB 962 719, 1963.
207. Sanofi, EP 274 324, 1987 (J. Radisson, E. Braye).
208. F. Torbjörn, T. Klingstedt, *Synthesis* **1987,** 40.
209. R. H. Foster, L. R. Wiseman, *Drugs* **56** (1998) 629.
210. The EPIC Investigators, *N. Engl. J. Med.* **330** (1994) 956.
211. R. E. Jordan et al. in M. Horton (ed.): *Adhes. Recept. Ther. Targets,* CRC, Boca Raton, Fla 1996, p. 281.
212. A. Fleckenstein, *Annu. Rev. Pharmacol. Toxicol.* **17** (1977) 149.
213. H. Kreiskott, H. P. Hofmann, *Arzneim. Forsch./ Drug Res.* **23** (1973) 1555.
214. Y. Ikeda, M. Kikuchi, K. Toyama, K. Watanabe, Y. Ando, *Thromb. Haemostasis* **45** (1981) 158.
215. H. Johnsson, *Thromb. Res.* **21** (1981) 523.
216. A. Shinjo, Y. Sasaki, M. Inamasu, T. Morita, *Thromb. Res.* **13** (1978) 941.
217. E. Coeffier, J. Cerrina, E. Jouvin-Marche, J. Benveniste, *Thromb. Res.* **31** (1983) 565.
218. D. E. MacIntyre, A. M. Shaw, *Thromb. Res.* **31** (1983) 833.
219. P. Han, C. Boatwright, N. G. Ardlie, *Thromb. Haemostasis* **50** (1983) 513.
220. P. Gaehtgens, *Arzneim. Forsch./Drug Res.* **31** (1981) 1995.
221. H. Schmid-Schönbein, *Arzneim. Forsch./Drug Res.* **31** (1981) 1999.
222. J. A. Dormandy, E. Ernst in K. Messmer, B. Fagrell (eds.): *Mikrozirkulation und arterielle Verschluß-krankheiten,* Karger, Basel 1981, p. 59.
223. G. Driessen, H. Heidtmann, H. Schmid-Schönbein, *Pflügers Arch.* **380** (1979) 1.
224. R. E. Wells, R. Denton, E. W. Merril, *J. Lab. Clin. Med.* **57** (1961) 646.
225. H. Schmid-Schönbein, L. Heinich, *Res. Exp. Med.* **161** (1973) 49.
226. H. Schmid-Schönbein, J. v. Gosen, L. Heinich, H. J. Klose, E. Volger, *Microvasc. Res.* **6** (1973) 366.
227. H. Schmid-Schönbein, J. Weiss, H. Ludwig, *Blut* **26** (1973) 369.
228. H. Schmid-Schönbein, K. Messmer, H. Rieger (eds.): *Hemodilution and Flow Improvement,* Karger, Basel 1981.
229. W. R. Bell, *CRC Handb. Ser. Clin. Lab. Sci. Sect. I,* **3** (1980) 301.
230. K. Stocker: "Fibrinolytics and Antifibrinolytics", *Handb. Exp. Pharmakol.* **46** (1978) 451.
231. F. N. McKenzie, K.-E. Arfors, O. Tangen, *Thromb. Res.* **3** (1973) 565.
232. A. M. Ehrly, *Vasa* Suppl. 1 (1973) 1.
233. H. Böhme, *Zeitschr. Allgemeinmed.* **54** (1978) 1644.
234. The Ancrod Study Investigators, *Stroke* **25** (1994) 1755.
235. a) C. Nolan, L. S. Hall, G. H. Barlow, *Methods Enzymol.* **45** (1976) 205. b) J. P. Collins, J. G. Jones, *Eur. J. Biochem.* **42** (1974) 81.
236. National Research Development Corporation, US 3 657 416, 1964 (H. A. Reid, M. P. Esnouf, K. E. Chan).
237. a) Abbott Laboratories, US 3 819 605, 1973 (W. H. Holleman, L. J. Weiss). b) W. H. Holleman, L. J. Weiss, *J. Biol. Chem.* **251** (1976) 1663.
238. Pentapharm, US 3849 252, 1972 (E. E. Percs, K. F. Stocker, B. Blombäck, M. Blombäck, B. Hessel).
239. K. F. Stocker, G. H. Barlow: "Defibrinierung mit thrombinähnlichen Schlangengiftenzymen", *Aktuel. Probl. Angiol.* **26** (1975) 45.

240. K. Stocker, G. H. Barlow, *Methods Enzymol.* **45** (1976) 214.
241. H. Hess, L. Franke, M. Jaunch, *Fortschr. Med.* **91** (1973) 743.
242. P. E. M. Jarrett, M. Moreland, N. L. Browse, *Curr. Med. Res. Opin.* **4** (1977) 492.
243. K. V. Weithmann, *IRCS Med. Sci.* **8** (1980) 293.
244. D. J. Mehta, *Annu. Rep. Med. Chem.* **17** (1982) 99.
245. W. Mohler, A. Söder, *Arzneim. Forsch./Drug Res.* **21** (1971) 1159.
246. Albert, DE 1 235 320, 1964 (W. Mohler, M. Reiser, K. Popendiker).
247. Kohjin, JP-Kokai 79112893, 1978 (S. Ueno et al.).
248. Albert, DE-OS 2 234 202, 1972 (A. Söder).
249. Albert, DE 2 302 772, 1973 (G. Nesemann, A. Söder, H. Thurm).
250. Albert, DE-OS 2 330 741, 1973 (A. Söder).
251. L. Sunder-Plassmann, *Therapiewoche* **33** (1983) 93.
252. J. A. Dormandy, E. Ernest, *Angiology* **32** (1981) 714.
253. L. J. Wurzinger in K. Messmer, B. Fagrell (eds.): *Mikrozirkulation und arterielle Verschluß-krankheiten,* Karger, Basel 1981, p. 77.
254. B. Endrich, R. Schosser, J. Beyer, E. Martin, K. Messmer in K. Messmer, B. Fagrell (eds.): *Mikrozirkulation und arterielle Verschluß-krankheiten,* Karger, Basel 1981, p. 154.
255. G. Trübestein, K. Balzer, H. Bisler, N. Klüken, Y. Mahfoud, H. Müller-Wiefel, B. Unkel, W. Ziegler, *Dtsch. Med. Wochenschr.* **107** (1982) 1957.
256. Orsymonde, DE-OS 2 122 144, 1971 (L. Lafon).
257. Abbott Laboratories, US 4 326 083, 1980 (W. H. McKellin).

Carnauba Wax → Waxes
Carrageenan → Polysaccharides
Case Hardening → Metals, Surface Treatment
Casein → Milk and Dairy Products
Casting Technology → Foundry Technology
Castor Oil → Fats and Fatty Oils
Catechol → Phenol Derivatives

Heterogeneous Catalysis and Solid Catalysts

Organometallic Compounds and Homogeneous Catalysis is a separate Keyword

HELMUT KNÖZINGER, Department Chemie, Universität München, Butenandtstr. 5 – 13 (Haus E), 81377 München, Germany

KARL KOCHLOEFL, Schwarzenbergstr. 15, 83026 Rosenheim, Germany

1. Introduction 3
1.1. Types of Catalysis 3
1.2. Catalysis as a Scientific Discipline 4
1.3. Industrial Importance of Catalysis 6
1.4. History of Catalysis 6
2. Theoretical Aspects 7
2.1. Principles and Concepts 7
2.1.1. Sabatier's Principle 7
2.1.2. The Principle of Active Sites 9
2.1.3. Surface Coordination Chemistry . . 10
2.1.4. Modifiers and Promoters 11
2.1.5. Active Phase – Support Interactions 11
2.1.6. Spillover Phenomena 12
2.1.7. Phase-Cooperation and Site-Isolation Concepts 13
2.1.8. Shape-Selectivity Concept 13
2.1.9. Principles of the Catalytic Cycle . . 14
2.2. Kinetics of Heterogeneous Catalytic Reactions 15
2.2.1. Concepts of Reaction Kinetics (Microkinetics) 16
2.2.2. Application of Microkinetic Analysis 18
2.2.3. Langmuir – Hinshelwood – Hougen – Watson Kinetics 18
2.2.4. Activity and Selectivity 19
2.3. Determination of Catalytic Mechanisms 21
3. Classification of Solid Catalysts . 22
3.1. Unsupported (Bulk) Catalysts . . 23
3.1.1. Metal Oxides 23
3.1.1.1. Simple Binary Oxides 23
3.1.1.2. Complex Multicomponent Oxides . 25
3.1.2. Metals and Metal Alloys 31
3.1.3. Carbides and Nitrides 31
3.1.4. Carbons 32
3.1.5. Ion-Exchange Resins and Ionomers 32
3.1.6. Molecularly Imprinted Catalysts . . 33
3.1.7. Metal Salts 33
3.2. Supported Catalysts 33
3.2.1. Supports 33
3.2.2. Supported Metal Oxide Catalysts . 34
3.2.3. Surface-Modified Oxides 35
3.2.4. Supported Metal Catalysts 35
3.2.5. Supported Sulfide Catalysts 36
3.2.6. Hybrid Catalysts 36
3.2.7. Ship-in-a-Bottle Catalysts 38
3.2.8. Polymerization Catalysts 38
4. Production of Heterogeneous Catalysts 39
4.1. Unsupported Catalysts 40
4.2. Supported Catalysts 43
4.2.1. Supports 44
4.2.2. Preparation of Supported Catalysts 44
4.3. Skeletal Catalysts 45
4.4. Unit Operations in Catalyst Production 46
5. Characterization of Solid Catalysts 48
5.1. Physical Properties 48
5.1.1. Surface Area and Porosity 48
5.1.2. Particle Size and Dispersion 50
5.1.3. Structure and Morphology 50
5.1.4. Local Environment of Elements . . 52
5.2. Chemical Properties 54
5.2.1. Surface Chemical Composition . . . 54
5.2.2. Valence States and Redox Properties 55
5.2.3. Acidity and Basicity 58
5.3. Mechanical Properties 60
6. Reaction Networks and Mechanisms of Selected Catalytic Reactions on Solid Catalysts 60
6.1. Carbon Monoxide Oxidation . . . 61
6.2. Ammonia Synthesis 61
6.3. Acid – Base catalysis of Hydrocarbon Transformations 62
6.3.1. Introduction 62
6.3.2. Carbocations and Their Reactions . 64
6.3.2.1. Nature and Formation of Carbocations 64
6.3.2.2. Reactions of Carbenium Ions 65
6.3.3. Catalytic Reactions Involving Carbocation Intermediates 70
6.4. Metal Catalysis of Hydrocarbon Transformations 75
6.5. Selective Hydrocarbon Oxidation Reactions 78

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_313

6.5.1. Epoxidation of Ethylene and Propene 78
6.5.2. Ammoxidation of Alkenes 80
6.6. Hydroprocessing Reactions 81
7. Application of Catalysis in Industrial Chemistry 83
7.1. Important Inorganic Reactions . 83
7.2. Important Organic Reactions . . . 85
7.3. Petroleum Refining 88
7.4. Production of Synthetic Fuels . . 89
7.5. Pollution Control 89
7.5.1. Mobile Sources 89
7.5.2. Stationary Sources 90
7.5.3. New Processes in Environmental Catalysis 90
7.6. Energy Conversion 90
7.6.1. Fuel Cells 90
7.6.2. Catalytic Combustion 91
8. Technology of Catalytic Processes 91
8.1. Catalytic Reactors 91
8.1.1. Classification of Reactors 91
8.1.2. Laboratory Reactors 92
8.1.3. Industrial Reactors 94
8.1.4. Special Reactor Types and Processes 99
8.2. Filling Fixed-Bed Reactors 101
8.3. Reactor Start-Up 101
8.4. Diffusion, Mass and Heat Transfer Effects 101
8.4.1. Effectiveness Factor 102
8.4.2. Effects on Selectivity 103
8.4.3. Catalysts and Transfer Processes . . 104
9. Catalyst Deactivation and Regeneration 104
9.1. Different Types of Deactivation . 104
9.2. Catalyst Regeneration 105
9.3. Catalyst Reworking and Disposal 106
10. References 106

1. Introduction

Catalysis is a phenomenon by which chemical reactions are accelerated by small quantities of foreign substances, called *catalysts*. A suitable *catalyst* can enhance the rate of a thermodynamically feasible reaction but cannot change the position of the thermodynamic equilibrium. Most catalysts are solids or liquids, but they may also be gases.

The catalytic reaction is a cyclic process. According to a simplified model, the reactant or reactants form a complex with the catalyst, thereby opening a pathway for their transformation into the product or products. Afterwards the catalyst is released and the next cycle can proceed.

However, catalysts do not have infinite life. Products of side reactions or changes in the catalyst structure lead to catalyst deactivation. In practice spent catalysts must be reactivated or replaced (see Chapter 9).

1.1. Types of Catalysis

If the catalyst and reactants or their solution form a common physical phase, then the reaction is called *homogeneously catalyzed*. Metal salts of organic acids, organometallic complexes, and carbonyls of Co, Fe, and Rh are typical *homogeneous catalysts*. Examples of *homogeneously catalyzed* reactions are oxidation of toluene to benzoic acid in the presence of Co and Mn benzoates and hydroformylation of olefins to give the corresponding aldehydes. This reaction is catalyzed by carbonyls of Co or Rh.

Heterogeneous catalysis involves systems in the which catalyst and reactants form separate physical phases. Typical *heterogeneous catalysts* are inorganic solids such as metals, oxides, sulfides, and metal salts, but they may also be organic materials such as organic hydroperoxides, ion exchangers, and enzymes.

Examples of *heterogeneously catalyzed* reactions are ammonia synthesis from the elements over promoted iron catalysts in the gas phase and hydrogenation of edible oils on Ni – kieselguhr catalysts in the liquid phase, which are examples of *inorganic*and *organic catalysis*, respectively.

In *biocatalysis*, enzymes or microorganisms catalyze various biochemical reactions. The catalysts can be immobilized on various carriers such as porous glass, SiO_2, and organic polymers. Prominent examples of biochemical reactions are isomerization of glucose to fructose, important in the production of soft drinks, by using enzymes such as glucoamylase immobilized on SiO_2, and the conversion of acrylonitrile to acrylamide by cells of corynebacteria entrapped in a polyacrylamide gel.

The main aim of *environmental catalysis* is environmental protection. Examples are the re-

duction of NO_x in stack gases with NH_3 on $V_2O_5 - TiO_2$ catalysts and the removal of NO_x, CO, and hydrocarbons from automobile exhaust gases by using the so-called *three-way catalyst* consisting of $Rh - Pt - CeO_2 - Al_2O_3$ deposited on ceramic honeycombs.

The term *green catalytic processes* has been used frequently in recent years, implying that chemical processes may be made environmentally benign by taking advantage of the possible high yields and selectivities for the target products, with little or no unwanted side products and also often high energy efficiency.

The basic chemical principles of catalysis consist in the coordination of reactant molecules to central atoms, the ligands of which may be molecular species (homogeneous and biocatalysis) or neighboring atoms at the surface of the solid matrix (heterogeneous catalysis). A1though there are differences in the details of various types of catalysis (e.g., solvation effects in the liquid phase, which do not occur in solid – gas reactions), a closer and undoubtedly fruitful collaboration between the separate communities representing homogeneous, heterogeneous, and biocatalysis should be strongly supported. A statement by David Parker (ICI) during the 21st Irvine Lectures on 24 April 1998 at the University of St. Andrews should be mentioned in this connection, namely, that, "... at the molecular level, there is little to distinguish between homogeneous and heterogeneous catalysis, but there are clear distinctions at the industrial level" [1].

1.2. Catalysis as a Scientific Discipline

Catalysis is a well-established scientific discipline, dealing not only with fundamental principles or mechanisms of catalytic reactions but also with preparation, properties, and applications of various catalysts. A number of academic and industrial institutes or laboratories focus on the study of catalysis and catalytic processes as well as on the improvement of existing and development of new catalysts.

International journals specializing in catalysis include *Journal of Catalysis*, *Journal of Molecular Catalysis*(Series A: Chemical; Series B: Enzymatic), *Applied Catalysis* (Series A: General; Series B: Environmental), *Reaction Kinetics and Catalysis Letters*, *Catalysis Today*, *Catalysis Letters*, *Topics in Catalysis*, *Advances in Organometallic Catalysis*, etc.

Publications related to catalysis can also be found in *Journal of Physical Chemistry*, *Langmuir*, and *Physical Chemistry Chemical Physics*.

Well-known serials devoted to catalysis are *Handbuch der Katalyse* [edited by G.-M. Schwab, Springer, Wien, Vol. 1 (1941) - Vol. 7.2 (1943)], *Catalysis* [edited by P. H. Emmett, Reinhold Publ. Co., Vol. 1 (1954) - Vol. 7 (1960)], *Catalysis – Science and Technology* [edited by J. R. Anderson and M. Boudart, Springer, Vol. 1 (1981) - Vol. 11 (1996)], *Catalysis Reviews* (edited by A. T. Bell and J. J. Carberry, Marcel Dekker), *Advances in Catalysis* (edited by B. C. Gates and H. Knözinger, Academic Press), *Catalysis* (edited by J. J. Spivey, The Royal Society of Chemistry), *Studies in Surface Science and Catalysis* (edited by B. Delmon and J. T. Yates), etc.

Numerous aspects of catalysis were the subject of various books. Some, published since 1980, are mentioned here:

C. N. Satterfield, *Heterogeneous Catalysis in Practice*, McGraw Hill Book Comp., New York, 1980.

D. L. Trimm, *Design of Industrial Catalysts*, Elsevier, Amsterdam, 1980.

J. M. Thomas, R. M. Lambert (eds.), *Characterization of Heterogeneous Catalysts*, Wiley, Chichester, 1980.

R. Pearce, W. R. Patterson (eds.), *Catalysis and Chemical Processes*, John Wiley, New York, 1981.

B. L. Shapiro (ed.), *Heterogeneous Catalysis*, Texas A & M Press, College Station, 1984.

B. E. Leach (ed.), *Applied Industrial Catalysis*, Vol. 1, 2, 3, Academic Press, New York, 1983 – 1984.

M. Boudart, G. Djega-Mariadassou, *Kinetics of Heterogeneous Reactions*, Princeton University Press, Princeton, 1984.

F. Delannay (ed.), *Characterization of Heterogeneous Catalysts*, Marcel Dekker, New York, 1984.

R. Hughes, *Deactivation of Catalysts*, Academic Press, New York, 1984.

M. Graziani, M. Giongo (eds.), *Fundamental Research in Homogeneous Catalysis*, Wiley, New York, 1984.

H. Heinemann, G. A. Somorjai (eds.), *Catalysis and Surface Science*, Marcel Dekker, New York, 1985.

J. R. Jennings (ed.), *Selective Development in Catalysis*, Blackwell Scientific Publishing, London, 1985.

G. Parshall, *Homogeneous Catalysis*, Wiley, New York, 1985.

J. R. Anderson, K. C. Pratt, *Introduction to Characterization and Testing of Catalysts*, Academic Press, New York, 1985.

Y. Yermakov, V. Likholobov (eds), *Homogeneous and Heterogeneous Catalysis*, VNU Science Press, Utrecht, Netherlands, 1986.

J. F. Le Page, *Applied Heterogeneous Catalysis – Design, Manufacture, Use of Solid Catalysts*, Technip, Paris, 1987.

G. C. Bond, *Heterogeneous Catalysis*, 2nd ed., Clarendon Press, Oxford, 1987.

P. N. Rylander, *Hydrogenation Methods*, Academic Press, New York, 1988.

A. Mortreux, F. Petit (eds.), *Industrial Application of Homogeneous Catalysis*, Reidel, Dordrecht, 1988.

J. F. Liebman, A. Greenberg, *Mechanistic Principles of Enzyme Activity*, VCH, New York, 1988.

J. T. Richardson, *Principles of Catalytic Development*, Plenum Publishing Corp., New York, 1989.

M. V. Twigg (ed.), *Catalyst Handbook*, Wolfe Publishing, London, 1989.

J. L. G. Fierro (ed.), *Spectroscopic Characterization of Heterogeneous Catalysts*, Elsevier, Amsterdam, 1990.

R. Ugo (ed.), *Aspects of Homogeneous Catalysis*, Vols. 1 – 7, Kluwer Academic Publishers, Dordrecht, 1990.

W. Gerhartz (ed.), *Enzymes in Industry*, VCH, Weinheim, 1990.

R. A. van Santen, *Theoretical Heterogeneous Catalysis*, World Scientific, Singapore, 1991.

J. M. Thomas, K. I. Zamarev (eds.), *Perspectives in Catalysis*, Blackwell Scientific Publications, Oxford, 1992.

B. C. Gates, *Catalytic Chemistry*, Wiley, New York, 1992.

G. W. Parshall, S. D. Ittel, *Homogeneous Catalysis*, 2nd ed., Wiley, New York, 1992.

J. J. Ketta (ed.), *Chemical Processing Handbook*, Marcel Dekker, New York, 1993.

J. A. Moulijn, P. W. N. M. van Leeuwen, R. A. van Santen (eds.), Catalysis – An Integrated Approach to Homogeneous, Heterogeneous and Industrial Catalysis, Elsevier, Amsterdam, 1993.

J. W. Niemantsverdriet, *Spectroscopy in Catalysis*, VCH, Weinheim, 1993.

J. Reedijk (ed.), *Bioinorganic Catalysis*, M. Dekker, New York, 1993.

G. A. Somorjai, *Introduction to Surface Chemistry and Catalysis*, Wiley, New York, 1994.

J. M. Thomas, W. J. Thomas, *Principles and Practice of Heterogeneous Catalysis*, VCH, Weinheim, 1996.

R. J. Wijngarden, A. Kronberg, K. R. Westerterp, *Industrial Catalysis – Optimizing Catalysts and Processes*, Wiley-VCH, Weinheim, 1998.

G. Ertl, H. Knözinger, J. Weitkamp (eds.), *Environmental Catalysis*, Wiley-VCH, Weinheim, 1999.

G. Ertl, H. Knözinger, J. Weitkamp (eds.), *Preparation of Solid Catalysts*, Wiley-VCH, Weinheim, 1999.

B. Cornils, W. A. Herrmann, R. Schlögl, C.-H. Wong, *Catalysis from A – Z*, Wiley-VCH, Weinheim, 2000.

B. C. Gates, H. Knözinger (eds.), *Impact of Surface Science on Catalysis*, Academic, San Diego, 2000.

A comprehensive survey of the principles and applications: G. Ertl, H. Knözinger, J. Weitkamp (eds.), *Handbook of Heterogeneous Catalysis* with 5 Volumes and 2479 pages was published by VCH, Weinheim, 1997.

The first *International Congress on Catalysis* (ICC) took place in 1956 in Philadelphia and has since been held every four years in Paris (1960), Amsterdam (1964), Moscow (1968), Palm Beach (1972), London (1976), Tokyo (1980), Berlin (1984), Calgary (1988), Budapest (1992), Baltimore (1996), Granada (2000). The 13th Congress will be held in Paris in 2004. Presented papers and posters have been published in the *Proceedings* of the corresponding congresses. *The International Congress on Catalysis Council (ICC)* was renamed at the Council meeting in Baltimore 1996. The international organization is now called *International Association of Catalysis Societies (IACS)*.

In 1965 the *Catalysis Society of North America* was established and holds meetings in the USA every other year.

The European Federation of Catalysis Societies (EFCATS) was established in 1990. The *EUROPACAT* Conferences are organized under the auspices of *EFCATS*. The first conference took place in Montpellier (1993) followed by Maastricht (1995), Cracow (1997), Rimini (1999), and Limerick (2001).

Furthermore, every four years (in the even year between two International Congresses on Catalysis) an *International Symposium* focusing on *Scientific Basis for the Preparation of Heterogeneous Catalysts*is held in Louvain-La Neuve (Belgiüm).

Other international symposia or congresses devoted to catalysis are: *International Zeolite Conferences*, *International Symposium of Catalyst Deactivation*, *Natural Gas Conversion Symposium*, *Gordon Conference on Catalysis*, TOCAT (*Tokyo Conference on Advanced Catalytic Science and Technology*), *International Symposium of Acid-Base Catalysis*, the European conference series, namely the *Roermond, Sabatier- and Schwab-conference*, and the *Taylor Conference*.

1.3. Industrial Importance of Catalysis

Because most industrial chemical processes are catalytic, the importance and economical significance of catalysis is enormous. More than 80 % of the present industrial processes established since 1980 in the chemical, petrochemical, and biochemical industries, as well as in the production of polymers and in environmental protection, use catalysts.

More than 15 international companies have specialized in the production of numerous catalysts applied in several industrial branches. In 1996 the turnover in the catalysts world market was estimated to be about $\$ 8 \times 10^9$ (see Chapter 4).

1.4. History of Catalysis

The phenomenon of catalysis was first recognized by Berzelius [2], [3] in 1835. However, some catalytic reactions such as the production of alcoholic beverages by fermentation or the manufacture of vinegar by ethanol oxidation were practiced long before. Production of soap by fat hydrolysis and of diethyl ether by dehydration of ethanol belong to the catalytic reactions that were performed in the 16th and 17th centuries.

Besides Berzelius, Mitscherlich [3] was also involved at the same time in the study of catalytic reactions accelerated by solids. He introduced the term *contact catalysis*. This term for heterogeneous catalysis lasted for more than 100 years.

In 1895 Ostwald [3], [4] defined catalysis as the acceleration of chemical reactions by the presence of foreign substances which are not consumed. His fundamental work was recognized with the Nobel prize for chemistry in 1909.

Between 1830 and 1900 several practical processes were discovered, such as flameless combustion of CO on a hot platinum wire, and the oxidation of SO_2 to SO_3 and of NH_3 to NO, both over Pt catalysts.

In 1912 Sabatier [3], [5] received the Nobel prize for his work devoted mainly to the hydrogenation of ethylene and CO over Ni and Co catalysts.

The first major breakthrough in industrial catalysis was the synthesis of ammonia from the elements, discovered by Haber [3], [6], [7] in 1908, using osmium as catalyst. Laboratory recycle reactors for the testing of various ammonia catalysts which could be operated at high pressure and temperature were designed by Bosch [3]. The ammonia synthesis was commercialized at BASF (1913) as the *Haber – Bosch* [8] process. Mittasch [9] at BASF developed and produced iron catalysts for ammonia production.

In 1938 Bergius [3], [10] converted coal to liquid fuel by high-pressure hydrogenation in the presence of an Fe catalyst.

Other highlights of industrial catalysis were the synthesis of methanol from CO and H_2 over $ZnO - Cr_2O_3$ and the cracking of heavier petroleum fractions to gasoline using acid-activated clays, as demonstrated by Houdry [3], [6] in 1928.

The addition of isobutane to $C_3 - C_4$ olefins in the presence of $AlCl_3$, leading to branched $C_7 - C_8$ hydrocarbons, components of high-quality aviation gasoline, was first reported by

IPATIEFF et al. [3], [7] in 1932. This invention led to a commercial process of UOP (USA).

Of eminent importance for Germany, which possesses no natural petroleum resources, was the discovery by FISCHER and TROPSCH [11] of the synthesis of hydrocarbons and oxygenated compounds from CO and H_2 over an alkalized iron catalyst. The first plants for the production of hydrocarbons suitable as motor fuel started up in Germany 1938. After World War II, Fischer-Tropsch synthesis saw its resurrection in South Africa. Since 1955 Sasol Co. has operated two plants with a capacity close to 3×10^6 t/a.

One of the highlights of German industrial catalysis before World War II was the synthesis of aliphatic aldehydes by ROELEN [12] by the addition of CO and H_2 to olefins in the presence of Co carbonyls. This homogeneously catalyzed reaction was commercialized in 1942 by Ruhr-Chemie and is known as Oxo Synthesis.

During and after World War II (till 1970) numerous catalytic reactions were realized on an industrial scale (see also Chapter 7). Some important processes are compiled in Table 1.

Table 2 summarizes examples of catalytic processes representing the current status of the chemical, petrochemical and biochemical industry as well as the environmental protection (see also Chapter 7).

2. Theoretical Aspects

The classical definition of a catalyst states that "a catalyst is a substance that changes the rate but not the thermodynamics of a chemical reaction" and was originally formulated by OSTWALD [4]. Hence, catalysis is a dynamic phenomenon.

As emphasized by BOUDART [16], the conditions under which catalytic processes occur on solid materials vary drastically. The reaction temperature can be as low as 78 K and as high as 1500 K, and pressures can vary between 10^{-9} and 100 MPa. The reactants can be in the gas phase or in polar or nonpolar solvents. The reactions can occur thermally or with the assistance of photons, radiation, or electron transfer at electrodes. Pure metals and multicomponent and multiphase inorganic compounds can act as catalysts. Site-time yields (number of product molecules formed per site and unit time) as low as $10^{-5}\ s^{-1}$ (corresponding to one turnover per day) and as high as $10^9\ s^{-1}$ (gas kinetic collision rate at 1 MPa) are observed.

It is plausible that it is extremely difficult, if not impossible, to describe the catalytic phenomenon by a general theory which covers the entire range of reaction conditions and observed site-time yields (reaction rates). However, there are several general *principles* which are considered to be laws or rules of thumb that are useful in many situations. According to BOUDART [16], the value of a principle is directly related to ist generality. In contrast, *concepts* are more specialized and permit an interpretation of phenomena observed for special classes of catalysts or reactions under given reaction conditions.

In this chapter, important principles and concepts of heterogeneous catalysis are discussed, followed by a section on kinetics of heterogeneously catalyzed reactions. The chapter is concluded by a section on the determination of reaction mechanisms in heterogeneous catalysis.

2.1. Principles and Concepts

2.1.1. Sabatier's Principle

The Sabatier principle proposes the existance of an unstable intermediate compound formed between the catalyst surface and at least one of the reactants [5]. This intermediate must be stable enough to be formed in sufficient quantities and labile enough to decompose to yield the final product or products. The Sabatier principle is related to linear free energy relationships such as a Brønsted relation [16]. These relations deal with the heat of reaction q (thermodynamic quantity) and the activation barrier E (kinetic quantity) of an elementary step in the exothermic direction ($q > 0$). With an empirical parameter a ($0 < a < 1$) and neglecting entropy effects, a Brønsted relation can be written as

$$\Delta E = a\,\Delta q,$$

where ΔE is the decrease in activation energy corresponding to an increase Δq in the heat of reaction. Hence, an elementary step will have a high rate constant in the exothermic direction when its heat of reaction q increases. Since the activation barrier in the endothermic direction is equal to the sum of the activation energy E

Table 1. Important catalytic processes commercialized during and after World War II (until 1970) [3], [6]

Year of commercialization	Process	Catalyst	Products
1939 – 1945	dehydrogenation	$Pt - Al_2O_3$	toluene from methylcyclohexane
	dehydrogenation	$Cr_2O_3 - Al_2O_3$	butadiene from *n*-butane
	alkane isomerization	$AlCl_3$	*i*-$C_7 - C_8$ from *n*-alkanes
1946 – 1960	oxidation of aromatics	V_2O_5	phthalic anhydride from naphtalene and *o*-xylene
	hydrocracking	Ni – alumosilicate	fuels from high-boiling petroleum fractions
	polymerization (Ziegler – Natta)	$TiCl_4 - Al(C_2H_5)_3$	polyethylene from ethylene
	dehydrogenation	$Fe_2O_3 - Cr_2O_3 - KOH$	styrene from ethylbenzene
	oxidation (Wacker process)	$PdCl_2 - CuCl_2$	acetaldehyde from ethylene
1961 – 1970	steam reforming	$Ni - \alpha\text{-}Al_2O_3$	Co, (CO_2), and H_2 from methane
	ammoxidation	Bi phosphomolybdate	acrylonitrile from propene
	fluid catalytic cracking	H zeolites + aluminosilicates	fuels from high boiling fractions
	reforming	bimetallic catalysts (Pt, Sn, Re, Ir)	gasoline
	low-pressure methanol synthesis	$Cu - ZnO - Al_2O_3$	methanol from CO, H_2, CO_2
	isomerization	enzymes immobilized on SiO_2	fructose from glucose (production of soft drinks)
	distillate dewaxing	ZSM-5, mordenites	removal of *n*-alkanes from gasoline
	hydrorefining	Ni – , $CO - MoS_x$	hydrodesulfurization, hydrodenitrification

Table 2. Important catalytic processes commercialized after 1970 [6], [7], [13–15]

Year of commercialization	Process	Catalyst	Product
1971 – 1980	automobile emission control	$Pt - Rh - CeO_2 - Al_2O_3$ (three-way catalyst)	removal of NO_x, CO, CH_x
	carbonylation (Monsanto process)	organic Rh complex	acetic acid from methanol
	MTG (Mobil process)	zeolite (ZSM-5)	gasoline from methanol
1981 – 1985	alkylation (Mobil – Badger)	modified zeolite (ZSM-5)	ethylbenzene from ethylene
	selective catalytic reduction (SCR; stationary sources)	V Ti (Mo, W)oxides (monoliths)	reduction of NO_x with NH_3 to N_2
	eserification (MTBE synthesis)	ion-exchange resin	methyl-*tert*-butyl ether from isobutene + methanol
	oxidation (Sumitomo Chem., 2-step process)	1. Mo, Bi oxides 2. Mo, V, PO (heteropolyacids)	acrylic acid from propene
	oxidation (Monsanto)	vanadylphosphate	maleic anhydride from *n*-butane
	fluid-bed polymerization (Unipol)	Ziegler – Natta type	polyethylene and polypropylene
	hydrocarbon synthesis (Shell)	1. $Co - (Zr,Ti) - SiO_2$ 2. $Pt - SiO_2$	middle destillate from $CO + H_2$
	environmental control (combustion process)	$Pt - Al_2O_3$ (monoliths)	deodoration
1986 –	oxidation with H_2O_2 (Enichem)	Ti silicalite	hydroquinone and catechol from phenol
	hydration	enzymes	acrylamide from acrylonitrile
	ammoxidation (Montedipe)	Ti silicalite	cylohexanone oxime from cyclohexanone, NH_3, and H_2O_2
	dehydrogenation of C_3, C_4 alkanes (Star and Oleflex processes)	Pt(Sn) – zinc aluminate, $Pt - Al_2O_3$	C_3, C_4 olefins

and the heat of reaction, the rate constant will decrease with increasing q.

The Brønsted relationship represents a bridge between thermodynamics and kinetics and, together with the Sabatier principle, permits an interpretation of the so-called *volcano plots* first reported by Balandin [17]. These volcano curves result when a quantity correlated with the rate of reaction under consideration is plotted against a measure of the stability of the intermediate compound. The latter quantity can be the heat of adsorption of one of the reactants or the heat of formation of a bulk compound relative to the surface compound, or even the heat of formation of any bulk compound that can be correlated with the heat of adsorption, or sim-

ply the position of the catalytic material (metal) along a horizontal series in the Periodic Table [16].

As an example, Figure 1 shows the volcano plot for the decomposition of formic acid on transition metals [18]. The intermediate in this reaction was shown to be a surface formate. Therefore, the heats of formation ΔH_f of the bulk metal formates were chosen as the measure of the stability of the intermediate. At low values of ΔH_f, the reaction rate is low and corresponds to the rate of adsorption, which increases with increasing heat of formation of the bulk formates (representing the stability of the surface compound). At high values of ΔH_f the reaction rate is also low and corresponds to the desorption rate, which increases with decreasing ΔH_f. As a consequence, a maximum in the rate of reaction (decomposition of formic acid) is observed at intermediate ΔH_f values which is neither a pure rate of adsorption nor a pure rate of desorption but which depends on both.

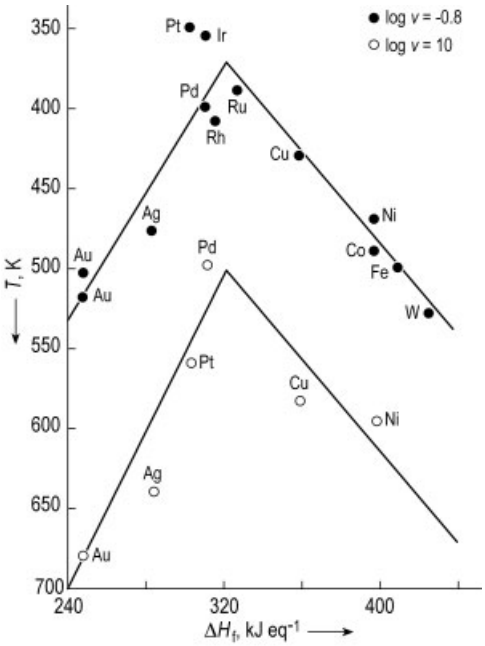

Figure 1. Volcano plot for the decomposition of formic acid. The temperature T at which the rate of decomposition v has a fixed value is plotted against the heat of formation ΔH_f of the metal formate (adopted from [28]).

2.1.2. The Principle of Active Sites

The Sabatier principle of an unstable surface intermediate requires chemical bonding of reactants to the catalyst surface, most likely between atoms or functional groups of reactant and surface atoms. This leads to the *principle of active sites*. When Langmuir formulated his model of chemisorption on metal surfaces [19], he assumed an array of sites which were energetically identical and noninteracting, and which would adsorb just one molecule from the gas phase in a localized mode. The Langmuir adsorption isotherm results from this model. The sites involved can be considered to be active sites.

Langmuir was already aware that the assumption of identical and noninteracting sites was an approximation which would not hold for real surfaces, when he wrote [20]: "Most finely divided catalysts must have structures of great complexity. In order to simplify our theoretical consideration of reactions at surfaces, let us confine our attention to reactions on plane surfaces. If the principles in this case are well understood, it should then be possible to extend the theory to the case of porous bodies. In general, we should look upon the surface as consisting of a checkerboard." Langmuir thus formulated the surface science approach to heterogeneous catalysis for the first time.

The heterogeneity of active sites on solid catalyst surfaces and its consequences were emphasized by Taylor [21], who recognized that "There will be all extremes between the case in which all atoms in the surface are active and that in which relatively few are so active." In other words, exposed faces of a solid catalyst will contain terraces, ledges, kinks, and vacancies with sites having different coordination numbers. Nanoscopic particles have edges and corners which expose atoms with different coordination numbers [22]. The variation of coordination numbers of surface atoms will lead to different reactivities and activities of the corresponding sites. In this context, Schwab's adlineation theory may be mentioned [23], which speculated that one-dimensional defects consisting of atomic steps are of essential importance. This view was later confirmed by surface science studies on stepped single-crystal metal surfaces [24].

In addition to variable coordination numbers of surface atoms in one-component solids, the surface composition may be different from that of the bulk and different for each crystallographic plane in multicomponent materials (*surface segregation* [25]). This would lead to a heterogeneity of the local environment of a surface atom and thus create nonequivalent sites.

Based on accurate kinetic measurements and on the Taylor principle of the existence of inequivalent active sites, BOUDART et al. [26] coined the terms *structure-sensitive* and *structure-insensitive reactions*. A truly structure-insensitive reaction is one in which all sites seem to exhibit equal activity on several planes of a single crystal. Surprisingly, many heterogeneously catalyzed reactions turned out to be structure-insensitive. Long before experimental evidence for this phenomenon was available and before a reliable interpretation was known, TAYLOR predicted it by writing [21]: "The amount of surface which is catalytically active is determined by the reaction catalyzed." In other words, the surface of a catalyst adapts itself to the reaction conditions for a particular reaction. The driving force for this reorganization of a catalyst surface is the minimization of the surface free energy, which may be achieved by *surface-reconstruction* [27], [28]. As a consequence, a meaningful characterization of active sites requires experiments under working (in situ) conditions of the catalytic system.

The principle of active sites is not limited to metals. Active sites include metal cations, anions, Lewis and Brønsted acids, acid – base pairs (acid and base acting simultaneously in chemisorption), organometallic compounds, and immobilized enzymes. Active sites may include more than one species (or atom) to form multiplets [17] or ensembles [29]. A mandatory requirement for these sites to be active is that they are accessible for chemisorption from the fluid phase. Hence, they must provide free coordination sites. Therefore, BURWELL et al. [30], [31] coined the term *coordinatively unsaturated sites* in analogy with homogeneous organometallic catalysts. Thus, active sites are to be considered as atoms or groups of atoms which are embedded in the surface of a matrix in which the neighboring atoms (or groups) act as ligands. *Ensemble*and *ligand effects* are discussed in detail by SACHTLER [32] and quantum chemical treatments of geometric ensemble and electronic ligand effects on metal alloy surfaces are discussed by HAMMER and NØRSKOV [33].

2.1.3. Surface Coordination Chemistry

The surface complexes formed by atoms or molecules are now known to usually resemble a local structure similar to molecular coordination complexes. The bonding in these surface complexes can well be described in a localized picture [34], [35]. Thus, important phenomena occuring at the surface of solid catalysts may be described in the framework of *surface coordination chemistry* or *surface organometallic chemistry* [36], [37].

This is at variance with the so-called *band theory* of catalysis, which attempted to correlate catalytic performace with bulk electronic properties [38–40]. The shortcomings of this theory in oxide catalysis are discussed by STONE [41].

2.1.4. Modifiers and Promoters

The performance of real industrial catalysts is often adjusted by *modifiers* (additives) [42], [43]. A modifier is called a *promoter* when it increases the catalyst activity in terms of reaction rate per site. Modifiers may also affect a catalyst's performance in an undesired manner. In this case the modifier acts as a catalyst *poison*. However, this simple distinction between promoters and poisons is less straightforward for reactions yielding more than one product in parallel or consecutive steps, of which only one is the desired product. In this case not only high activity but high selectivity is desired. The selectivity can be improved by adding substances that poison undesirable reactions. In exothermic reactions excessively high reaction rates may lead to a significant temperature increase (sometimes only locally: hot spots) which can yield undesirable products (e.g., CO and CO_2 in selective catalytic oxidation). A deterioration of the catalyst due to limited catalyst stability may also occur. Consequently, a modifier is required which decreases the reaction rate so that a steady-state temperature and reaction rate can be maintained. Although the modifier acts as a poison in these cases, it is in fact a promoter as far as selectivity and catalyst stability are concerned.

Modifiers can change the binding energy of an active site or its structure, or disrupt an ensemble of atoms, e.g., by alloying an active with an inactive metal. A molecular approach toward an understanding of promotion in heterogeneous catalysis was presented by HUTCHINGS [44].

As an example, the iron-based ammonia synthesis catalyst is promoted by Al_2O_3 and K_2O [45]. Alumina acts as a *textural* promoter, as it prevents the rapid sintering of pure iron metal. It may also stabilize more active sites on the iron surface (*structural promoter*). Potassium oxide appears to affect the adsorption kinetics and dissociation of dinitrogen and the binding energy of nitrogen on adjacent iron sites (*electronic promoter*).

The addition of Co to MoS_2-based catalysts supported on transitional aluminas has a positive effect on the rate of hydrodesulfuration of sulfur-containing compounds at Co/(Co + Mo) ratios below ca. 0.3 [46] (see Section 3.2.4). The active phase is proposed to be the so-called Co-MoS phase which consists of MoS_2 platelets, the edges of which are decorated by Co atoms. The latter may act as structural and electronic promoters simultaneously.

Another example concerns bifunctional catalysts for catalytic reforming [47], which consist of Pt supported on strongly acidic aluminas, the acid strength of which is enhanced by modification with chloride. Since these materials lose chlorine during the catalytic process, the feed contains CCl_4 as a precursor of the surface chloride promoter.

2.1.5. Active Phase – Support Interactions

Several concepts have proved valuable in interpreting phenomena which are pertinent to certain classes of catalysts. In supported catalysts, the active phase (metal, oxide, sulfide) undergoes active phase-support interactions [48–50]. These are largely determined by the surface free energies of the support and active phase materials and by the interfacial free energy between the two components [48–50]. Active transition metal oxides (e.g., V_2O_5, MoO_3, WO_3) have relatively low surface free energies as compared to typical oxidic support materials such as γ-Al_2O_3, TiO_2 (anatase), and SiO_2. Although the interfacial free energies between active phase and support are not known, the interaction between the two components appears to be favorable, with the exception of SiO_2-supported transition metal oxides. As a consequence spreading and wetting phenomena occur if the thermal treatment of the oxide mixtures is carried out at temperatures sufficiently high to induce mobility of the active oxide. As a rule of thumb, mobility of a solid typically occurs above the Tammann temperature, which is equal to half the melting point of the bulk solid. As a result, the active transition metal oxide tends to wet the support surface and forms a monolayer (monolayer-type catalysts).

Transition and noble metals typically have high surface free energies [49], and therefore, small particles or crystallites tend to agglomerate to reduce their surface area. Stabilization of nanosize metal particles therefore requires deposition on the surface of supports providing favorable *metal-support interactions* (MSI). The smaller the particle the more its physical properties and morphology can be affected by these interactions. Therefore, the nature of the support material for a given metal also critically influences the catalytic properties of the metal particle.

Supported metals are in a nonequilibrium state and therefore still tend to agglomerate at sufficiently high temperatures in reducing atmospheres. Hence, deactivation occurs because of the reduced metal surface area. Regeneration can typically be achieved by thermal treatment in an atmosphere in which the active metal is oxidized. The surface free energies of transition and noble metal oxides are significantly lower than those of the parent metals, so that their spreading on the support surface becomes more favorable. Subsequent reduction under sufficiently mild conditions can restore the high degree of metal dispersion [dispersion D is defined as the ratio of the number of metal atoms exposed at the particle surface (N_S) to the total number of metal atoms N_T in the particle ($D = N_S/N_T$)].

So-called *strong metal-support interactions* (SMSI) may occur, e.g., for Pt – TiO_2 and Rh – TiO_2 [48], [50], [51]. As shown experimentally, the adsorption capacity for H_2 and CO is drastically decreased when the precursor for the catalytically active metal on the support is reduced in H_2 at temperatures above ca. 770 K [48], [51]. Simultaneously, the oxide support

is slightly reduced. Although several explanations have been proposed for the SMSI effect, the most probable explanation is encapsulation of the metal particle by support oxide material. Encapsulation may occur when the support material becomes mobile. Although the electronic properties of the metal particle may be affected by the support oxide in the SMSI state, the decrease of the adsorption capacity appears to be largely due to a geometric effect, namely, the resulting inaccessibility of the metal surface.

The various possible morphologies and dispersions of supported metals are schematically shown in Figure 2. The metal precursor typically is well dispersed after impregnation of the support. Low-temperature calcination may lead to well-dispersed oxide overlayers, while direct low-temperature reduction leads to highly dispersed metal particles. This state can also be reached by low-temperature reduction of the dispersed oxide precursor, this step being reversible by low-temperature reoxidation. For the preparation of highly dispersed Ni catalysts, it is important to remove the water that is formed by hydrogen reduction of NiO. H_2 diluted with N_2 is used for this purpose. Surface compound formation may also occur by a solid-state reaction between the active metal precursor and the support at high calcination temperatures. Reduction at high temperatures may lead to particle agglomeration when cohesive forces are dominant, and to so-called pill-box morphologies when adhesive forces are dominant. In both cases, the metal must be mobile. In contrast, when the support is mobile, sintering of the support can occur, and the small metal particles are stabilized on the reduced surface area (cohesive forces). Alternatively, if adhesive forces are dominant encapsulation (SMSI effect) may occur.

2.1.6. Spillover Phenomena

In multiphase solid catalysts *spillover* may occur of an active species (spillover species) adsorbed or formed on one phase (donor phase) onto a second phase (acceptor) which does not form the active species under the same conditions [53–55]. A well-known example is hydrogen spillover from Pt, on which dihydrogen chemisorbs dissociatively, onto WO_3 with formation of a tungsten bronze [56]. According to SOMORJAI [57] the spillover phenomenon must be regarded as one of the "modern concepts in surface science and heterogeneous catalysis". Nevertheless, the exact physical nature of spillover processes has only rarely been verified experimentally. The term is typically used to explain nonlinear effects (synergistic effects) of the combination of chemically different components of a catalytic material on its performance.

Besides hydrogen spillover, oxygen spillover has been postulated to play an important role in oxidation reactions catalyzed by mixed oxides. For example, the addition of antimony oxide to selective oxidation catalysts enhances the catalytic activity at high levels of selectivity by a factor of up to five relative to the Sb-free system, although antimony oxide itself is completely inactive.

Observations of this kind motivated DELMON et al. [58], [59] to formulate the *remote-control concept* to explain the fact that all industrial catalysts used for the partial oxidation of hydrocarbons or in hydrotreatment are multiphasic and that particular phase compositions develop synergy effects. The remote-control concept is, however, not undisputed.

2.1.7. Phase-Cooperation and Site-Isolation Concepts

GRASSELLI [60] proposed the *phase-cooperation concept* for partial oxidation and ammoxidation reactions. It is suggested that two phases (e.g., α-$Bi_2Mo_3O_{12}$ and γ-Bi_2MoO_6) cooperate in the sense that one phase performs the actual catalytic function (α-phase) and the other (γ-phase) the reoxidation function. The concept could be verified for many other multiphase, multicomponent mixed metal oxide catalysts, such as multicomponent molybdates and multicomponent antimonates [60], [61].

Another concept most relevant for selective oxidation and ammoxidation is the *site-isolation concept* first formulated by CALLAHAN and GRASSELLI [62]. Site isolation refers to the separation of active sites from each other on the surface of a heterogeneous catalyst and is considered to be the prerequisite for obtaining the desired selective partial oxidation products. The concept states that reactive surface lattice oxygen atoms must be structurally isolated

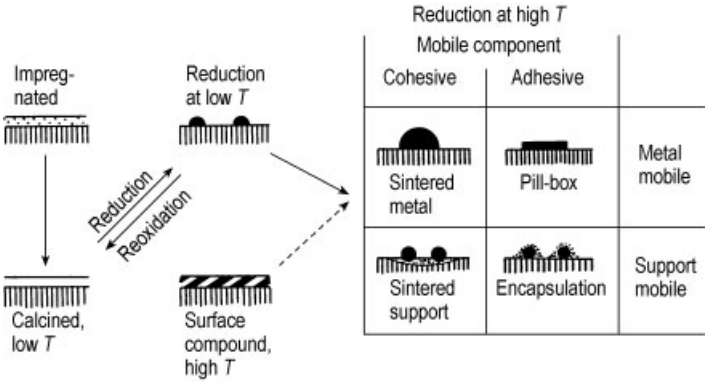

Figure 2. Schematic representation of metal-support interactions (adopted from [47])

from each other in defined groupings on a catalyst surface to achieve selectivity. The number of oxygen atoms in a given isolated grouping determines the reaction channel through the stoichiometry requirements imposed on the reaction by the availability of oxygen at the reaction site. It was postulated that two and up to five adjacent surface oxygen atoms would be required for the selective oxidation of propene to the desired product acrolein. Lattice groupings with more than five oxygen atoms would only produce total oxidation products (CO and CO_2), while completely isolated single oxygen atoms would be either inactive or could produce allyl radicals. The latter would couple in the vapor phase to give hexadiene and ultimately benzene. These scenarios are schematically shown in Figure 3.

2.1.8. Shape-Selectivity Concept

Zeolites and related materials have crystalline structure and contain regular micropores, the diameters of which are determined by the structure of the materials. The pore sizes are well defined and have dimensions similar to those of small organic molecules. This permits *shape-selective catalysis* to occur. The geometric constraints may act on the sorption of reactants, on the transition state of the catalyzed reaction, or on the desorption of products. Correspondingly, shape-selective effects have been classified as providing reactant shape selectivity, restricted transition state shape selectivity, and product shape selectivity [64], [65]. These scenarios are schematically illustrated in Figure 4 for the cracking of *n*-heptane and 1-methylhexane (reactant shape selectivity), for the transalkylation of *m*-xylene (transition state shape selectivity), and for the alkylation of toluene by methanol (product shape selectivity). In the first example, the kinetic diameter of *n*-heptane is smaller than that of 1-methylhexane. The latter is not able to enter micropores, so that shape-selective cracking of *n*-heptane takes place when both hydrocarbons are present in the feed. An example for shape-selective control of the transition state is the transalkylation of *m*-xylene. The reaction is bimolecular and the formation of 1,2,4-trimethylbenzene has a less bulky transition state than the formation of 1,3,5-trimethylbenzene. The latter product can thus not be formed if the pore size and geometry is carefully adapted to the transition state requirements. Finally, p-xylene can be selectively formed by methylation of toluene with methanol and zeolites whose pore openings only allow *p*-xylene to be released. The *o* and *m* isomers either accumulate in zeolite cages or are isomerized to *p*-xylene.

2.1.9. Principles of the Catalytic Cycle

The most fundamental principle in catalysis is that of the *catalytic cycle*, which may be based on a redefinition of a catalyst by Boudart [66]: "A catalyst is a substance that transforms reactants into products, through an *uninterrupted* and *repeated* cycle of elementary steps in which the catalyst is changed through a sequence of *reactive intermediates*, until the last step in the cycle regenerates the catalyst in its original form".

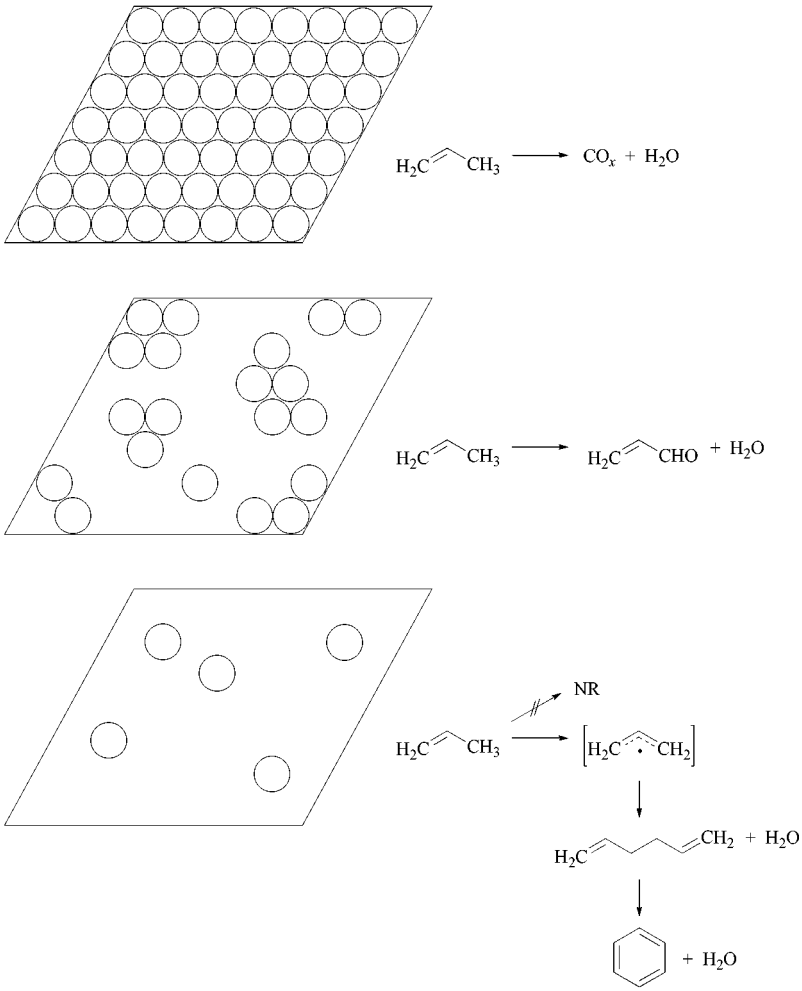

Figure 3. Site-isolation principle. Schematic of lattice oxygen arrangements on hypothetical surfaces. Anticipated reaction paths of propene upon contact with these surfaces (NR = no reaction; adopted from [60])

The catalytic substance or active sites may not be present originally, but may be formed by activation during the start-up phase of the catalytic reaction. The cycle must be uninterrupted and repeated since otherwise the reaction is stoichiometric rather than catalytic. The *number of turnovers*, a measure of catalyst life, must be greater than unity, since the catalyst would otherwise be a reagent. The total amount of catalyst (active sites) is typically small relative to the amounts of reactants and products involved (catalytic amounts). As a consequence, the reactive intermediates can be treated by the kinetic quasi-steady-state approximation of BODENSTEIN.

The *activity* of the catalyst is defined by the number of cycles per unit time or turnovers or turnover frequency (TOF; unit: s^{-1}). The *life* of the catalyst is defined by the number of cycles before it dies.

2.2. Kinetics of Heterogeneous Catalytic Reactions [66–75]

The catalytic cycle is the principle of catalytic action. The mechanism of a catalyzed reaction can be described by the sequence of elementary reaction steps of the cycle, including adsorption, surface diffusion, chemical transformations of

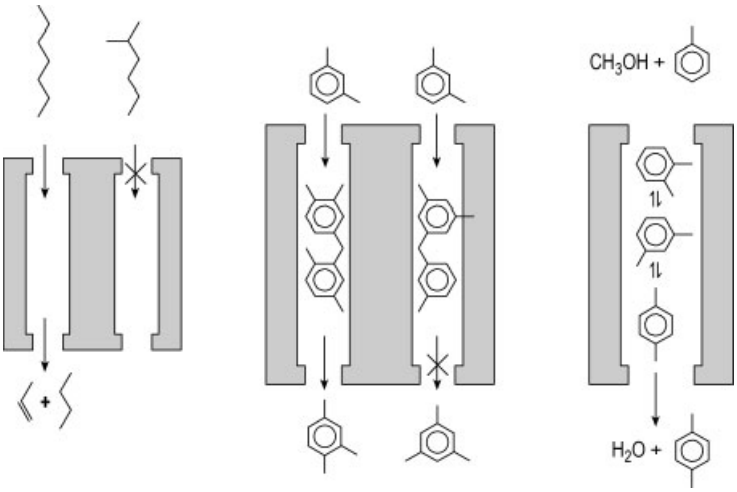

Figure 4. Classification of shape-selective effects

adsorbed species, and desorption, and it is the basis for deriving the kinetics of the reaction. It is assumed that for each individual elementary step the transition-state theory is valid. An early treatize of the kinetics of heterogeneously catalized reactions was published by Schwab [79].

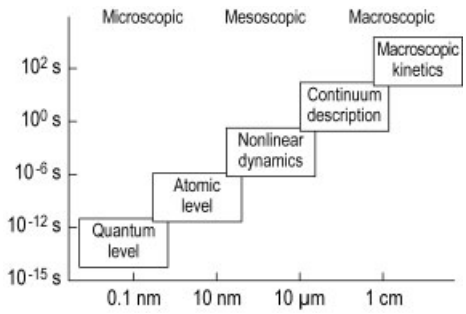

Figure 5. Schematic classification of the various aspects of the dynamics of surface reactions (adopted from [28])

The various aspects of the dynamics of surface reactions and catalysis have been classified by Ertl [28] into five categories in terms of time and length scales, as shown schematically in Figure 5. In the *macroscopic regime*, the rate of a catalytic reaction is modeled by fitting empirical equations, such as power laws, to experimental data, so as to describe its concentration and pressure dependence and to determine rate constants that depend exponentially on temperature. This approach was very useful in chemical engineering for reactor and process design. Assumptions about reaction schemes (kinetic models) provide correlations between the surface coverages of intermediates and the external variables, an approach that led to the Temkin equation [76] modeling the kinetics of ammonia synthesis.

Improved kinetic models could be developed when atomic processes on surfaces and the identification and charaterization of surface species became available. The progress of a catalytic reaction is then described by a *microkinetics* approach by modeling the macroscopic kinetics through correlating atomic processes with macroscopic parameters within the framework of a suitable continuum model. Continuum variables for the partial surface coverages are, to a first approximation, correlated to external parameters (partial pressures and temperature) by the Langmuir lattice model of a surface consisting of identical noninteracting adsorption sites.

The formulation of rate laws for the full sequence of elementary reactions will usually lead to a set of nonlinear coupled (ordinary) differential equations for the concentrations (coverages) of the various surface species involved. The temporal behavior of the reaction system under constant continuous-flow conditions may be nonstationary. In certain parameter ranges it may be oscillatory or even chaotic. Also, there may be local variations in surface coverages which lead to coupling of the reaction with transport processes (e.g., particle diffusion, heat transfer). The formation of spatiotemporal concentration profiles

on a *mesoscopic scale* is the consequence of these nonlinear dynamic phenomena.

Since the Langmuir lattice model is not valid in reality, the continuum model can describe the reaction kinetics only to a first approximation. Interactions between adsorbed species occur, and adsorbed particles occupy nonidentical sites, so that complications arise in the description of the reaction kinetics. Apart from the heterogeneity of adsorption sites, surfaces may undergo structural transformations. Surface science investigations provide information on these effects on an *atomic scale*.

As mentioned above, it is assumed that the transition-state theory is valid for description of the rates of individual elementary steps. This theory is based on the assumption that at all stages along the reaction coordinate thermal equilibrium is established. Temperature then is the only essential external macroscopic parameter. This assumption can only be valid if energy exchange between all degrees of motional freedom of the particles interacting with the solid acting as a heat bath is faster than the elementary step which induces nuclear motions. Energy transfer processes at the *quantum level* are the basic requirements for chemical transformations.

Nonlinear dynamics and the phenomena occuring at the atomic and quantum levels were reviewed by Ertl [28].

2.2.1. Concepts of Reaction Kinetics (Microkinetics)

The important concepts of (catalytic) reaction kinetics were reviewed by Boudart [66], [67], [77], [78], and by Cortright and Dumesic [73].

The term *microkinetics* was defined to denote reaction kinetics analyses that attempt to incorporate into the kinetic model the basic surface chemistry involved in the catalytic reaction at a molecular level [72], [73]. An important prerequisite for this approach is that reaction rates are measured in the absence of heat- and mass-transfer limitations. The kinetic model is based on a description of the catalytic process in terms of information and/or assumptions about active sites and the nature of elementary steps that make up the catalytic cycle. The ultimate goal of a kinetic analysis is the determination of preexponential factors and activation energies (cf. Arrhenius equation) for all elementary steps in forward and reverse direction. Usually there is not sufficient information available to extract the values of all kinetic parameters. However, it has been established that in many cases the observed kinetics are controlled by a limited number of kinetic parameters [72], [73]. Questions to be answered in this situation are: (1) how many kinetic parameters are required to calculate the overall rate from a reaction scheme? (2) What species are the *most abundant intermediates*on the catalyst surface under reaction conditions? (3) Does the reaction scheme include a *rate-determining step* for the kinetic parameters of interest under the reaction conditions? Generally, only a few parameters are kinetically significant, although it is difficult to predict which parameters control the overall rate of the catalytic process. Therefore, initial estimates require a larger set of parameters than are ultimately necessary for the kinetic description of the catalytic process of interest. Besides experimental values of kinetic parameters for individual elementary reactions (often resulting from surface science studies on single-crystal surfaces), quantum chemical calculations permit mechanistic investigations and predicitions of kinetic parameters [33–35].

Assume that a kinetic model has been established which consists of n elementary steps, each proceeding at a net rate

$$r_i = r_{\mathrm{f}i} - r_{\mathrm{r}i} \quad (\mathrm{i} = 1, 2, \ldots, n) \tag{1}$$

The subscripts f and r stand for "forward" and "reverse", respectively. As mentioned above, the validity of the Bodenstein steady-state concept can be assumed. The kinetic steady state is then defined by:

$$\sigma_i r = r_i \tag{2}$$

where r is the net rate $r_{\mathrm{f}} - r_{\mathrm{r}}$ of the overall catalytic reaction defined by a stoichiometric equation. σ_{i} is the stoichiometric number of the ith step, i.e., the number of times that this step must occur for the catalytic cycle to turnover once. If the transition-state theory is valid for each individual elementary step, the ratio of the forward rate $r_{\mathrm{f}i}$ to the reverse rate $r_{\mathrm{r}i}$ of step i is given by the De Donder relation [80], [81]:

$$r_{fi}/r_{ri} = \exp(A_i/RT) \tag{3}$$

where A_i is the affinity of step i:

$$A_i = [\partial G_i/\partial \xi_i]_{T,P} \tag{4}$$

where ξ_i is the extent of reaction of step i.

At steady state, the affinity for each step but one may be very small as compared to the affinity A of the overall reaction. Each step but one is then in quasi-equilibrium. The step that is not in quasi-equilibrium (subscript d) is called the *rate-determining step* (rds) as defined by HORIUTI [82]. As a consequence of this definition, the following inequalities are valid:

$$r_{fi} >> r_{fd}, \text{ and } r_{ri} >> r_{rd} \ (i \neq d).$$

If there is an rds, then the affinity $A_i = 0$ for all values of i except for the rds ($i \neq d$), i.e., all (or almost all) of the affinity for the catalytic cycle is dissipated in the rds, hence

$$A = \sigma_d A_d \tag{5}$$

It follows that

$$r_f/r_r = r_{fd}/r_{rd} \tag{6}$$

At steady state $\sigma_d(r_f - r_r) = r_{fd} - r_{rd}$. Hence

$$\sigma_d r_f = r_{fd} \text{ and } \sigma_d r_r = r_{rd}. \tag{7}$$

The stoichiometric equation for the overall reaction can always be written such that σ_d is equal to unity. It is then clear that the rds is appropriately and uniquely named as the step for which the forward and reverse rates are equal to the forward and reverse rates, respectively, of the overall reaction [66].

Clearly the rds (if there is one) is the only *kinetically significant step*. A kinetically significant step is one whose rate constants or equilibrium constant appear in the rate equation for the overall reaction. In some cases there is no rds in the Horiuti sense, but frequently only a few of the elementary steps in a catalytic cycle are kinetically significant. It is sometimes said that a rate-limiting step is the one having the smallest rate constant. However, rate constants can often not be compared because they have different dimensions.

The relative importance of rate constants of elementary steps in a catalytic cycle provides useful guidelines for the development of activity and selectivity. This can be achieved by *parametric sensitivity analysis* [83], which was first proposed by CAMPBELL [84] for analysis of kinetic parameters of catalytic reactions (see also ref. [73]). CAMPBELL [84] defined a *degree of rate control* for any rate constant k_i in a catalytic cycle turning over at a rate r

$$X_i = k_i/r \cdot \partial r/\partial k_i \tag{8}$$

where the equilibrium constant for step i and all other rate constants are held constant. The main advantage of this mathematical operation is its simplicity. It turns out that HORIUTI's rds, as the only kinetically significant step in a catalytic cycle, has a degree of rate control $X_i = 1$, whereas the X values for all other steps are equial to zero. Clearly, all intermediate values of X_i are possible, and probable in most cases.

As a catalytic cycle turns over at the quasi-steady state, the steady-state concentrations (coverages) of the reactive intermediates may be significantly different from the values that they would attain if they were at equilibrium with fluid reactants or products. The steady-state concentrations (coverages) of reactive intermediates may be lower or higher than the equilibrium values. The reason for this phenomenon is *kinetic coupling* between elementary steps at the steady state, where the net rate of each step is equal to the net rate of the overall reaction multiplied by the stoichiometric number of the step. With kinetic coupling, a reactive intermediate can accumulate as a reactant or be depleted as a product [67], [77], [80].

The principle of *microscopic reversibility* is strictly valid only for reactions at equilibrium. Away from equilibrium, it remains valid provided that transition-state theory is still applicable, which appears to be the case in heterogeneous catalysis [16]. Hence, the principle remains valid for any elementary step in a heterogeneous catalytic reaction. However, the principle must be applied with caution to a catalytic cycle, as opposed to a single elementary reaction. If valid, the principle of microscopic reversibility allows the calculation of a rate constant if the second rate constant and the equilibrium constant K_i of an elementary reaction i are known: $k_{fi}/k_{ri} = K_i$.

2.2.2. Application of Microkinetic Analysis

The ammonia synthesis reaction has been investigated extensively, both on real iron synthesis catalysts and on iron single crystals [45], [85]. Numerous data were therefore available, on the basis of which a reaction model could be established and a microkinetic analysis performed. Ammonia synthesis yields could be theoretically predicted from known kinetic parameters and compared with experimentally measured yields for an industrial iron catalyst. Figure 6 shows the correlation between predicted and experimental yields over a wide range of pressures [86]. The straight line convincingly demonstrates the power of the microkinetic approach.

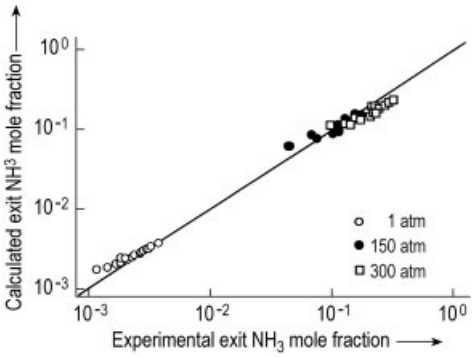

Figure 6. Comparison of calculated and measured ammonia production over a commercial iron-based catalyst for a broad range of temperatures, pressures, N/H ratios, and gas flows (adopted from [86]).

Additional examples for several other reactions are reviewed in [73].

2.2.3. Langmuir – Hinshelwood – Hougen – Watson Kinetics [63], [69], [71], [87]

The Langmuir – Hinshelwood – Hougen – Watson (LHHW) approach is based on the Langmuir model describing the surface of a catalyst as an array of equivalent sites which do not interact either before or after chemisorption. Further, for derivation of rate equations, it is assumed that both reactants and products are equilibrated with surface species that react on the surface in a rate-determining step. Surface coverages are correlated with partial pressures or concentrations in the fluid phase by means of Langmuir adsorption isotherms. It was mentioned above that the Langmuir model is unrealistic. Moreover, it was demonstrated in Section 2.2.1 that the surface coverages of adsorbed species are by no means identical to the equilibrium values predicted by the Langmuir adsorption isotherm for reaction systems in which kinetic coupling occurs, and rate-determining steps do not generally exist.

Despite these weaknesses, the LHHW kinetics approach has proved valuable for modeling heterogeneous catalytic reactions for reactor and process design. The kinetic parameters which are determined by fitting the rate equations to experimental data, however, do not have a straightforward physical meaning. As an alternative, simple power-law kinetics for straightforward reactions (e.g., $A \rightarrow B$) can be used for technical application.

Often it is difficult to discriminate between two or more kinetic models within the accuracy limits of the experimental data. Sophisticated mathematical procedures have therefore been developed for the discrimination of rival models [88].

As an example for a typical LHHW rate equation consider the reaction

$$A + B \rightleftharpoons C.$$

The form of rate equation is as follows [88]:

$$r = \frac{k_{\mathrm{rds}} N_{\mathrm{T}} K_i \left(P_{\mathrm{A}} P_{\mathrm{B}} - P_{\mathrm{C}}/K_{\mathrm{eq}}\right)}{\left(1 + K_{\mathrm{A}} P_{\mathrm{A}} + K_{\mathrm{B}} P_{\mathrm{B}} + K_{\mathrm{C}} P_{\mathrm{C}} + \sum_j K_{\mathrm{j}} P_{\mathrm{j}}\right)^{\mathrm{n}}} = \frac{\text{rate factor} \times \text{driving force}}{\text{inhibition term}} \tag{9}$$

The numerator is a product of the rate constant of the rds k_{rds}, the concentration of active sites N_{T}, adsorption equilibrium constants K_i, and the driving force for the reaction. The latter is a measure of how far the overall reaction is from thermodynamic equilibrium. The overall equilibrium constant K_{eq}, can be calculated from thermodynamics. The denominator is an inhibition term which takes into account the competitive adsorption of reactants and products.

A few examples of LHHW rate equations are summarized in Table 3. A collection of useful LHHW rate equations and kinetic data for almost 100 industrially important catalytic reactions is available in [89].

Table 3. General structure of Langmuir type rate equations [87]

	Reaction	Controlling step	Net rate	Kinetic constant	Driving force	Adsorption term
1.	$A \rightleftharpoons P$	a. adsorption of A	$k_A p_A(1-\Sigma\Theta) - k_A'\Theta_A$	k_A	$p_A - \frac{p_P}{K}$	$1+K_A p_A + K_P p_P$
		b. surface reaction single-site mechanism	$k_S\Theta_P - k_S'\Theta_P$	$k_S K_A$	$p_A - \frac{p_P}{K}$	$1+K_A p_A + K_P p_P$
		c. desorption of P	$k_P'\Theta_P - k_P p_P(1-\Sigma\Theta)$	$k_P' k_S K_A$	$p_A - \frac{p_P}{K}$	$1+K_A p_A + K_P p_P$
2.	$A \rightleftharpoons P+Q$	surface reaction, A (ads) reacts with vacant site	$k_S\Theta_A(1-\Sigma\Theta) - k_S\Theta_P$	$k_S K_A$	$p_A - \frac{p_P p_Q}{K}$	$(1+K_A p_A + K_P p_P + K_Q p_Q)^2$
3.	$A_2 \rightleftharpoons 2P$	a. dissociative adsorption of A_2	$k_A p_{A_2}(1-\Sigma\Theta)^2 - k_A'\Theta_A$	k_A	$p_{A_2} - \frac{p_P^2}{K}$	$(1+K_A p_{A_2} + K_P p_P)^2$
		b. surface reaction following dissociative adsorption of A_2	$k_S\Theta_A - k_S'\Theta_P$	$k_S K_A$	$p_{A_2} - \frac{p_P}{K}$	$1+K_A p_{A_2} + K_P p_P$
4.	$A+B \rightleftharpoons P$	a. adsorption of A	$k_A p_A(1-\Sigma\Theta) - k_A'\Theta_A$	k_A	$p_A - \frac{p_P}{K p_B}$	$1+K_A p_A + K_B p_B + K_P p_P$
		b. surface reaction	$k_S\Theta_A\Theta_B - k_S'\Theta_P(1-\Sigma\Theta)$	$k_S K_A K_B$	$p_A p_B - \frac{p_P}{K}$	$(1+K_A p_A + K_B p_B + K_P p_P)^2$
		c. desorption of p	$k_P'\Theta_P - k_P p_P(1-\Sigma\Theta)$	$k_P' K_S K_A K_B$	$p_A p_B - \frac{p_P}{K}$	$1+K_A p_A + K_B p_B + K_P p_P$
5.	$\frac{1}{2}A+B \rightleftharpoons P$	dissociative adsorption of A_2, only half of which react	$k_A p_{A_2}(1-\Sigma\Theta) - k_A'\Theta_A$	k_A	$p_{A_2} - \frac{p_P}{K p_B}$	$1+K_A p_{A_2} + K_B p_B + K_P p_P$
6.	$A+\frac{1}{2}B \rightleftharpoons P$	adsorption of A, which reacts with half of B produced from the dissociative adsorption of B_2	$k_A p_A(1-\Sigma\Theta) - k_A'\Theta_A$	k_A	$p_A - \frac{p_P}{K p_{B_2}}$	$1+K_A p_A + K_B p_{B_2} + K_P p_P$
7.	$A_2+2B \rightleftharpoons 2P$	a. dissociative adsorption of A_2	$k_A p_{A_2}(1-\Sigma\Theta)^2 - k_A'\Theta_{A^2}$	k_A	$p_{A_2} - \frac{p_P^2}{K p_{B_2}}$	$1+K_A p_A + K_B p_B + K_P p_P$
		b. surface reaction following the dissociative of adsorption A	$k_S\Theta_A\Theta_B - k_S' p(1-\Sigma\Theta)$	$k_S K_A K_B$	$p_{A_2} p_B - \frac{p_P}{K}$	$1+K_A p_{A_2} + K_B p_B + K_P p_P$

2.2.4. Activity and Selectivity

Catalytic activity is expressed in terms of reaction rates, preferably normalized to the surface area of the active phase (e.g., metal surface area for supported metal catalysts). These surface areas can be obtained by suitable chemisorption techniques (see Section 5.1). As an alternative to these *areal rates*, *specific rates* are also used which are normalized to catalyst weight. The best possible measure of catalytic activity, however, is the *turnover rate* or *turnover frequency*, since it is normalized to the number of active sites and represents the rate at which the catalytic cycle turns over. For comparison of rates reported by different research groups, the methodology for the determination of the number of active sites must be carefully reported. The hitherto unresolved problem is that the site densities measured prior to the catalytic reaction are not necessarily identical to those available under reaction conditions.

A readily available measure of catalytic activity is *space–time yield*, expressed in units of amount of product made in the reactor per unit time and unit reactor volume.

A considerable obstacle for the comparison of catalytic activities for a given reaction that were obtained in different laboratories for the same catalyst is the use of different reactors. For a series of catalysts, reasonable comparisons of activities or rates are possible when relative values are used.

Conversion data alone, or conversion versus time plots are not sufficient as a measure of catalytic activity.

Selectivity can be defined as the amount of desired product obtained per amount of consumed reactant. Selectivity values are only useful if the conversion is also reported. A simple measure of selectivity is the *yield* (yield = selectivity × conversion). Selectivities can also be used to indicate the relative rates of two or more competing reactions; competition may occur when several reactants form products in parallel (type I):

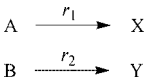

when one reactant transforms into several products in parallel (type II):

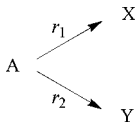

or in consecutive reactions (type III):

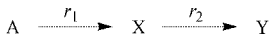

The selectivity is defined as the ratio of the rate of formation of the desired product to the rate of consumption of the starting material [90]. Thus, the selectivities for product X for the first-order reactions I and II is $r_1/(r_1 + r_2)$, whereas it is $(r_1 - r_2)/r_1$ for type III.

In the case of type I or II reactions, selectivity for X or Y is independent of the conversion of the starting material. In type III reactions, the selectivity for X is 100 % initially, decreases gradually with increasing conversion, and drops to zero at 100 % conversion. At an intermediate conversion, there is a maximum yield of X which depends on the ratio of the rate constants k_1 and k_2 of the rates r_1 and r_2. The integrated rate equations are:

$$[\mathrm{A}] = \exp(-k_1 t) \tag{10}$$

$$[\mathrm{X}] = k_1/(k_2 - k_1)\,[\exp(-k_1 t) - \exp(-k_2 t)] \tag{11}$$

where [A] is the concentration of unconverted A, [X] the concentration of A converted to product X, and t time. The maximum yield is reached at

$$t = (k_1 - k_2)^{-1} \ln(k_1/k_2) \tag{12}$$

2.3. Determination of Catalytic Mechanisms [31]

The best description of a catalytic mechanism is the corresponding catalytic cycle. As a first step, a detailed product analysis is required to differentiate between a single clean reaction and systems undergoing parallel and/or consecutive reactions. The microkinetic approach, as outlined in Sections 2.2.1 and 2.2.2, for the predicition of the overall rate of a catalytic reaction taking into account the surface chemistry of the catalyst and the elementary reactions involved, is the most promising procedure to predict a mechanism.

However, the results of microkinetic analyses may not always be unequivocal, and the discrimination between different kinetic models may not be straightforward. Therefore, additional information is necessary to prove or disprove the sequence of elementary steps (catalytic cycle) that represent the mechanism of a catalyzed reaction at the molecular level.

Quantum chemical calculations at various levels of sophistication and computer modeling procedures are available today [33–35], [91], [92] which permit structures of adsorbed intermediates and transition states to be elucidated and reaction energy diagrams to be computed.

Transient kinetics measurements can also provide quantitative values of kinetic parameters and the elucidation of individual reaction steps [93], [94]. Pulse reactors are one type of transient reactors. The TAP reactor (temporal analysis of products) is discussed in Section 8.1.2. In another kind of transient experiment, step changes in concentrations are effected, and the response of product concentration is measured as a function of time. The analysis of this response provides details of the course of reaction and permits kinetic parameters to be determined.

A powerful technique for the kinetic and mechanistic study of heterogeneous catalytic reactions is steady-state isotopic-transient kinetic analysis (SSITKA) [94], [95]. The technique is based on the detection of isotopic labels in the reactor effluent species versus time following a step change in the isotopic labeling of one of the reactants in the reactor feed. Reactant and product concentrations and flow rates remain undisturbed during the step change and – in the absence of isotopic mass effects – steady-state conditions are maintained unter isotopic-transient operation. In contrast to other transient experiments, the steady-state kinetic behavior of the catalyst surface can be studied. Steady-state kinetic and mechanistic information which can be obtained from SSITKA includes concentrations of different types of adsorbed reaction intermediates, coverages, surface lifetimes, site heterogeneity, activity distributions, and identification of possible mechanisms [94].

The use of isotopes can greatly aids the elucidation of catalytic mechanisms [96]. The most frequently used isotopes are ^{2}H, ^{13}C, ^{14}C, and ^{18}O. Deuterium-exchange reactions with organic reactants yield isotopic distribution patterns, i.e., the fractions of the isotopically exchanged molecules, D_1, D_2, ... , where D_i is the fraction of the compound containing *i* deuterium atoms. These isotopic distribution patterns are often specific enough to eliminate a number of conceivable mechanisms. When carried out in conjunction with structure variations, isotopic distribution patterns may be effective in narrowing the range of possible mechanisms even though such studies cannot give "the mechanism" [31]. Deuterium labeling is also used to determine which carbon atoms end up where or whether a reaction is inter- or intramolecular [31]. ^{13}C labeling can be used for the same purpose. Although nonradioactive labels are preferred, radioactive tracers such as ^{14}C have also been used [97]. ^{18}O labeling has been applied to elucidate the relative rates of CO and CO_2 in methanol synthesis [98].

Kinetic isotope effects [99], [100] are due to the different masses of a given element and its corresponding isotope. The resulting difference in zero-point energies may lead to an increase in activation energy of the labeled molecule and therefore a reduction of the rate. Whether a kinetic isotope effect occurs or not when isotopic labeling exists for an atom in a certain position or group (say hydrogen in an X – H bond replaced by X – D) for a catalytic reaction of interest provides information on whether the weakening or rupture of the X – H bond is involved in a kinetically significant elementary step.

Modification of organic molecules with suitable substituent groups may provide valuable information on reaction mechanisms from the stereochemistry of the reaction of interest [31], [102], [103]. Substituents generally also have electronic effects on the reactivity of a parent reactant (substituent effects). Resulting linear free-energy relationships for a series of substituents also assist the determination of kinetically significant reaction steps of a conceivable reaction mechanism [104–107], since the substituents directly affect the relative energy of the transition state and hence the activation barrier of a kinetically significant step.

Modification of molecules by substituents may also cause intra- or intermolecular steric effects [31] and steric interactions between adsorbate and catalyst surface can be studied. The latter studies provide almost the only way to directly probe the steric nature of active catalytic

sites without confusion with adsorption sites that are not catalytic sites [31].

Selective feeding and scavenging has been proposed for the characterization of reaction intermediates [31]. Suppose Q is a suspected intermediate for a particular reaction. This hypothesis can be tested by adding (feeding) a compound to the reaction feed which is supposedly adsorbed to form the suspected intermediate Q, and by testing whether the added compound is indeed converted to the expected product. In scavenging, a compound is added which should react with the intermediate Q to form another compound which is not normally a product.

The nature of catalytic sites can be tested and their density estimated by selective poisoning [108].

Of course, any spectroscopic technique which is surface-sensitive and which has sufficiently high sensitivity and can be applied under catalytic working conditions can provide valuable information on the nature of sufficiently long-lived intermediates. However, spectroscopy often detects spectator species rather than reaction intermediates. Therefore it is mandatory to demonstrate that true intermediates are in fact seen. This can be done by varying critical reaction parameters and monitoring the response of spectroscopic signal intensities as a function of time. Sufficiently high time resolution of the applied spectroscopic technique is therefore required (see Chapter 5).

3. Classification of Solid Catalysts

Solid catalysts are extremely important in large-scale processes [109–113] for the conversion of chemicals, fuels, and pollutants. Many solid materials (elements and compounds) including metals, metal oxides, and metal sulfides, are catalysts. Only a few catalytic materials used in industry are simple in composition, e.g., pure metals (e.g., Ni) or binary oxides (such as γ-Al_2O_3, TiO_2). Typical industrial catalysts, however, consist of several components and phases. This complexity often makes it difficult to assess the catalytic material's structure.

In the following a variety of families of existing catalysts are described, and selected examples are given. These families include (1) unsupported (bulk) catalysts; (2) supported catalysts; (3) confined catalysts (ship-in-a-bottle catalysts); (4) hybrid catalysts; (5) polymerization catalysts, and several others. The selected examples not only include materials which are in use in industry, but also materials which are not yet mature for technological application but which have promising potential.

3.1. Unsupported (Bulk) Catalysts

3.1.1. Metal Oxides

Oxides are compounds of oxygen in which the O atom is the more strongly electronegative component. Oxides of metals are usually solids. Their bulk properties largely depend on the bonding character between metal and oxygen. Metal oxides have widely varying electronic properties and include insulators (e.g., Al_2O_3, SiO_2), semiconductors (e.g., TiO_2, NiO, ZnO), metallic conductors (typically reduced transition metal oxides such as TiO, NbO, and tungsten bronzes), superconductors (e.g., $BaPb_{1-x}Bi_xO_3$), and high-T_c superconductors (e.g., $YBa_2Cu_3O_{7-x}$).

Metal oxides make up a large and important class of catalytically active materials, their surface properties and chemistry being determined by their composition and structure, the bonding character, and the coordination of surface atoms and hydroxyl groups in exposed terminating crystallographic faces. They can develop acid-base and redox properties. Metal oxides can have simple composition, like binary oxides, but many technologically important oxide catalysts are complex multicomponent materials.

3.1.1.1. Simple Binary Oxides

Simple binary oxides of base metals may behave as solid acids or bases or amphoteric materials [114]. These properties are closely related to their dissolution behavior in contact with aqueous solutions. Amphoteric oxides (e.g., Al_2O_3, ZnO) form cations in acidic and anions in basic milieu. Acidic oxides (e.g., SiO_2) dissolve with formation of acids or anions. Transition metal oxides in their highest oxidation state (e.g., V_2O_5, CrO_3) behave analogously. Basic oxides (e.g., MgO, lanthanide oxides) form hydroxides

or dissolve by forming bases or cations. These dissolution properties must be considered when such oxides are used as supports and impregnated from aqueous solutions of the active phase precursor [115], [116]. The dissolution properties also are closely related to the surface properties of the oxides in contact with a gas phase, where the degree of hydration/hydroxylation of the surface is a critical parameter. Silica, alumina and magnesia are commonly used catalysts and catalyst supports representative for a wide range of surface acid – base properties.

Aluminas are amphoteric oxides, which form a variety of different phases depending on the nature of the hydroxide or oxidehydroxide precursor and the conditions of their thermal decomposition. Bayerite, nordstrandite, boehmite, and gibbsite can be used as starting materials. The thermal evolution of the various poorly crystalline transitional phases (namely η-, Θ-, γ-, χ-, and κ-Al_2O_3) and of the final crystalline, thermodynamically stable α-Al_2O_3 phase (corundum) is shown in Figure 7. The structures of these oxides can be described as close-packed layers of oxo anions with Al^{3+} cations distributed between tetrahedral and octahedral vacancy positions. Stacking variations of the oxo anions result in the different crystallographic forms of alumina. The most commonly used transitional phases are η- and γ-Al_2O_3, which are often described as defect spinel structures [117] that incorporate Al^{3+} cations in both tetrahedral and octahedral sites. The Al sublattice is highly disordered, and irregular occupation of the tetrahedral interstices results in a tetragonal distortion of the spinel structure. There is a higher occupancy of tetrahedral cation positions in γ-Al_2O_3, and a higher density of stacking faults in the oxygen sublattice of η-Al_2O_3. Crystallites are preferentially terminated by anion layers, and these layers are occupied by hydroxyl groups for energetic reasons [118].

Acidic and basic sites and acid-base pair sites have been identified on the surfaces of aluminas [119]. Thermal treatment of hydroxylated oxides leads to partial dehydroxylation with formation of coordinatively unsaturated O^{2-} ions (basic sites) and an adjacent anion vacancy which exposes 3- or 5-coordinate Al^{3+} cations (Lewis acid sites). The remaining hydroxyl groups can be terminal or doubly or triply bridging with the participation of Al^{3+} in tetrahedral and/or octahedral positions. The properties of the resulting OH species range from very weakly Brønsted acidic to rather strongly basic and nucleophilic [118], [119]. As a result of this complexity, alumina surfaces develop a rich surface chemistry and specific catalytic properties [120].

Besides their intrinsic catalytic properties and their use as catalysts in their own right (e.g., for elimination reactions, alkene isomerization [120], and the Claus process [121]), aluminas are frequently used as catalyst supports for oxides and metals. The surface area and particle size of aluminas can be controlled by the preparation conditions, and their redox and thermal stability give the supported active phases high stability and ensure a long catalyst lifetime.

Silicas are weakly Brønsted acidic oxides which occur in a variety of structures such as quartz, tridymite and cristobalite (→ Silica) [122], [123]. The most commonly used silica in catalysis is amorphous silica. The building blocks of silica are linked SiO_4 tetrahedra, with each O atom bridging two Si atoms. Bonding within the solid is covalent. At the fully hydrated surface, the bulk structure is terminated by hydroxyl (silanol) groups, SiOH [119], [122], [123]. Two types of these groups are usually distinguished: isolated groups and hydrogen-bonded vicinal groups. Fully hydrated samples, calcined at temperatures below 473 K, may contain geminal groups $Si(OH)_2$ [119], [122], [123]. Heating in vacuum removes the vicinal groups by dehydroxylation, i.e., condensation to form H_2O and Si – O – Si linkages (siloxane bridges). Complete removal of the hydroxyl groups occurs at temperatures well above 973 K in vacuo and is believed to result in significant changes in surface morphology.

The surface hydroxyl groups are only weakly Brønsted acidic and therefore hardly develop any catalytic activity. They are, however, amenable to hydrogen-bonding [124] and they are usually regarded as the most reactive native surface species, which are available for functionalization of silicas. The siloxane bridges are (at least after heating at elevated temperatures) essentially unreactive. For this reason and because of the low acidity of silanol groups, silicas are not used as active catalysts, but they play an important role as oxide supports and for the

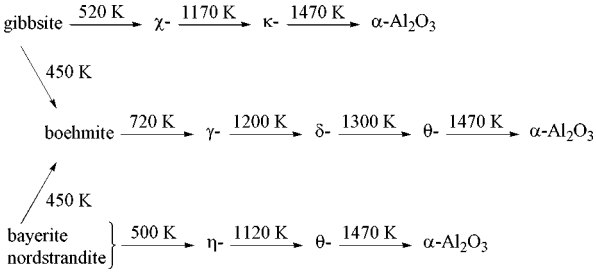

Figure 7. The dehydration sequences of the aluminum trihydroxides in air (adopted from [117])

synthesis of functionalized oxide supports (see Section 3.2.5).

Tailored silicas can be synthesized by controlling the preparation conditions [122], [123]. Thus, surface area, particle size and morphology, porosity and mechanical stability can be varied by modification of the synthesis parameters.

In addition to amorphous silicas, the crystalline microporous silica silicalite I can be obtained by hydrothermal synthesis [125]. This material has MFI structure and can be considered as the parent siliceous extreme of zeolite ZSM-5.

Large-pore mesoporous structures, the so-called porosils, have also been reported [125–127]. The dimensions of their linear and parallel pores can be varied from 2 to 10 nm in a regular fashion. These pores can therefore accomodate bulky molecules and functional groups.

The incorporation of foreign elements such as Al^{3+} substituting for Si^{4+} induces Brønsted acidity and creates activity for acid catalysis.

Magnesium oxide is a basic solid. It has the simple rock salt structure, with octahedral coordination of magnesium and oxygen. Ab inito molecular orbital calculations indicated that the electronic structure is highly ionic, with the $Mg^{2+}O^{2-}$ formalism being an accurate representation of both bulk and surface structures [128]. The lattice is commonly envisaged to terminate in (100) planes incorporating five-coordinate (5c) Mg^{2+} and O^{2-} ions [129] (see Figure 8). This model appears to be physically accurate for MgO smoke, which may be regarded as a model crystalline metal oxide support [130]. Although the (100) plane is electrically neutral, hydroxyl groups are present on the surfaces of polycrystalline MgO. These groups and the O^{2-} anions are responsible for the basic properties, coordinatively unsaturated Mg^{2+} ions being only weak Lewis acid sites. The hydroxyl groups are also highly nucleophilic.

These properties dominate the surface chemistry of MgO. Organic Brønsted acids have been shown to chemisorb dissociatively to form surface-bound carbanions and surface hydroxyl groups [131]. Even the heterolytic dissociative adsorption of dihydrogen on polycristalline MgO has been reported. $Mg^{2+}O^{2-}$ pairs with Mg^{2+} and O^{2-} in 4- or 3-coordination seem to play a crucial role.

The presence of low-coordinate Mg^{2+} and O^{2-} ions (see Figure 8) on the MgO surface after activation at high temperatures has been demonstrated [129], [132], [133], and the unique reactivity of 3c centers has been discussed [134].

MgO has also been used as a host matrix for transition metal ions (**solid solutions**) [135]. These materials permit the properties of isolated transition metal ions to be studied.

Transition metal oxides [136–138] can be struturally described as more or less dense packings of oxide anions, the interstices of which are occupied by cations. The bonding, however, is never purely ionic, but rather mixed ionic-covalent, sometimes also developing metallic character (e.g., bronzes). The surface of these oxides is often partially occupied by hydroxyl groups, so they possess some acidic character. However, it is the variability in oxidation states and the possibility of forming mixed-valence and nonstoichiometric compounds that are responsible for their important redox catalytic properties. The most frequently used transition metal oxides are those of the early transition metals (mostly suboxides). Fields of application are particularly selective oxidation and dehydrogenation reactions.

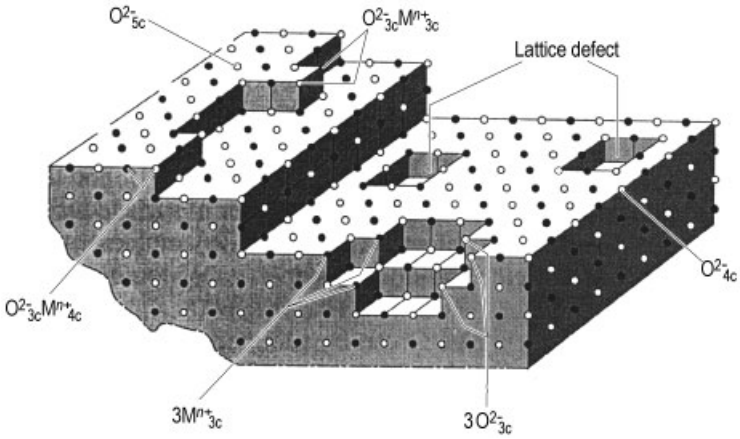

Figure 8. Representation of a surface plane (100) of MgO showing surface imperfections such as steps, kinks, and corners wich provide sites for ions of low coordination (adopted from [129]).

Titania TiO_2 exists in two major crystallographic forms: anatase and rutile. Anatase is the more frequently used modification since it develops a larger surface area, although it is a metastable phase and may undergo slow transformation into the thermodynamically stable rutile phase above ca. 900 K. Vanadium impurities seem to accelerate the rutilization above 820 K. Other impurities such as surface sulfate and phosphate seem to stabilize the anatase phase. The anatase $\rightarrow$ rutile phase transition must be sensitively controlled for supported VO_x/TiO_2, which plays a significant role in selective oxidation and NO_x reduction catalysis.

Titania is a semiconductor with a wide band gap and as such is an important material for photocatalysis [139], [140].

Zirconia has attracted significant interest in the recent past as a catalyst support and as a base material for the preparation of strong solid acids by surface modification with sulfate or tungstate groups [141]. The most important crystallographic phases of ZrO_2 for catalytic applications are tetragonal and monoclinic. The latter is the thermodynamically stable phase. Higher surface areas, however, are developed by the metastable tetragonal phase, which is stabilized at low temperatures by sulfate impurities or intentional addition of sulfate or tungstate.

ZrO_2 is the base material for the solid-state electrolyte sensor for the measurement of oxygen partial pressure in, e.g., car exhaust gases [142]. The solid electrolyte shows high bulk conductivity for O^{2-} ions.

Other transition metal oxides are used as supported catalysts or as constituents of complex multicomponent catalysts.

Only a few examples are reported on the application of the unsupported binary oxides as catalysts. Iron oxide Fe_2O_3 and chromium oxide Cr_2O_3 catalyze the oxidative dehydrogenation of butenes to butadiene. Fe_2O_3-based catalysts are used in the high-temperature water gas shift reaction [143] and in the dehydrogenation of ethylbenzene [144]. Vanadium pentoxide V_2O_5 is active for the selective oxidation of alkenes to saturated aldehydes [145]. Acidic transition metal oxides such as vanadium pentoxide and molybdenum trioxide MoO_3 can be used for the synthesis of formaldehyde by oxidative dehydrogenation of methanol, while the more basic iron oxide Fe_2O_3 leads to total oxidation [146]. Zinc oxide ZnO is used as a catalyst for the oxidation of cyclohexanol to cyclohexanone.

3.1.1.2. Complex Multicomponent Oxides

Complex multicomponent oxides play a major role as catalytic materials.

Aluminium silicates are among the most important ternary oxides. Four-valent Si atoms are isomorphously substituted by trivalent Al atoms in these materials. This substitution cre-

ates a negatively charged framework of interconnected tetrahedra. Exchangeable cations are required for charge compensation when protons are incorporated as charge-compensating cations, OH groups bridging Si and Al atoms are created which act as Brønsted acidic sites.

Amorphous silica – alumina can be prepared by precipitation from solution. This mixed oxide is a constituent of hydrocarbon cracking catalysts.

Zeolites. Hydrothermal synthesis can be used for preparation of a large family of crystalline aluminosilicates, known as *zeolites* (→ Zeolites), which are microporous solids with pore sizes ranging from ca. 3 to 7 Å [125], [147], [148]. Characteristic properties of these structurally well-defined solids are selective sorption of small molecules (molecular sieves), ion exchange, and large surface areas. Zeolites possess a framework structure of corner-linked SiO_4^{4-} and AlO_4^{5-}-tetrahedra with two-coordinate oxygen atoms that bridge two tetrahedral centers (so-called T atoms). Zeolite frameworks are open and contain channels (straight or sinusoidal) or cages of spherical or other shapes. These cages are typically interconnected by channels. The evolution of several zeolite structures from the primary tetrahedra via secondary building blocks is demonstrated in Figure 9 [149]. The diameter of the channels is determined by the number n of T atoms surrounding the opening of the channels as n-membered rings. Small-pore zeolites contain 6- or 8-membered rings (diameter d: $2.8 < d < 4$ Å), medium-pore zeolites contain 10-membered rings ($5 < d < 6$ Å) and the openings of large-pore zeolites are constructed of 12-membered rings ($d > 7$ Å). Examples of small-pore zeolites are sodalite and zeolite A, of medium-pore zeolites ZSM-type zeolites (see Figure 9), while large-pore zeolites include faujasites and zeolites X and Y (see Figure 9).

The H forms of zeolites develop strong Brønsted acidity and play a major role in large-scale industrial processes such as catalytic cracking, the Mobil MTG (methanol-to-gasoline) process and several others.

Besides Si and Al as T atoms P atoms can also be incorporated in zeolite structures. In addition, transition metal atoms such as Ti, V, and Cr can substitute for Si, which leads to oxidation catalysts of which titanium-silicalite-1 (TS1) is the most outstanding catalyst for oxidation, hydroxylation, and ammoxidation with aqueous H_2O_2 [150].

Basic properties can be created in zeolites by ion-exchange with large alkali metal ions such as Cs^+ and additional loading with CsO [151].

Aluminum phosphates (AlPO) [152], [153] are another family of materials whose structures are similar to those of zeolites. They can be regarded as zeolites in which the T atoms are Si and Al. More recently they have been named zeotypes, the T atoms of which are Al and P. In contrast to aluminosilicate zeolites, AlPOs typically have a Al/P atomic ratio of 1/1, so that the framework composition $[AlPO_4]$ is neutral. Therefore, these solids are nonacidic and have hardly any application as catalysts. However, acidity can be introduced by substituting Al^{3+} by divalent atoms, which yields *metal aluminophosphates* (MAPOs), e.g., MnAPO or CoAPO, or by partial substitution of formally pentavalent P by Si^{4+} to give *silicoaluminophosphates* (SAPO). The AlPO family contains members with many different topologies which span a wider range of pore diameters than aluminosilicate zeolites.

Mesoporous solid acids with well-defined pore structures can be obtained by replacing a certain amount of Si atoms in MCM-type oxides by Al atoms.

Clays (→ Clays) are aluminosilicate minerals (montmorillonite, phyllosilicates (smectites), bentonites, and others). Montmorillonite is an aluminohydroxysilicate and is the main constituent of most clay minerals. It is a 2 : 1 clay, i.e., one octahedral AlO_6 layer is sandwiched between two tetrahedral SiO_4 layers. Montmorillonites are reversibly swellable and possess ion-exchange capacity. They can be used as catalyst supports. The structural layers can be linked together by introducing inorganic pillars which prevent the layers from collapsing at higher temperatures when the swelling agent is evaporated (pillared clays) [154]. A bimodal micro-/mesoporous pore size distribution can thus be obtained. Pillaring can be achieved with a wide variety of reagents including hydroxylaluminum polymers, zirconia hydroxypolymers, silica, and silicate pillars. Cat-

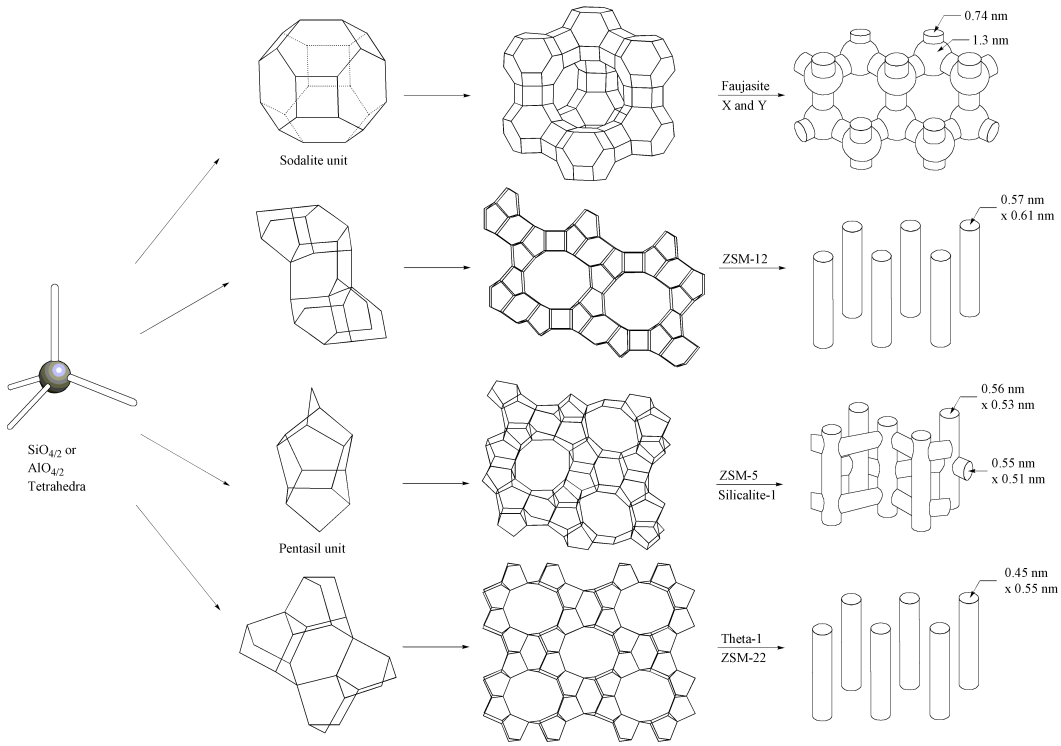

Figure 9. Structures of four representative zeolites and their micropore systems and dimensions [149]

alytically active components may be built in by the pillaring material, e.g., transition metal oxide pillars.

Table 4. Examples of mixed metal oxide catalysts and their applications∗

Catalyst	Active phases	Industrial processes
Copper chromite	$CuCr_2O_4$, CuO	low-temperature CO conversion, oxidations, hydrogenation
Zinc chromite	$ZnCr_2O_4$, ZnO	methanol synthesis (high pressure)
Copper/zinc chromite	$Cu_xZn_{1-x}Cr_2O_4$, CuO	methanol synthesis (low pressure)
Iron molybdate	$Fe(MoO_4)_3$, MoO_3	methanol to formaldehyde
Zinc ferrite	$ZnFe_2O_4$	oxidative dehydrogenation
Chromia – alumina	$Cr_xAl_{2-x}O_3$	dehydrogenation of light alkanes

∗ Adapted from [155]

Mixed metal oxides are multimetal multiphase oxides which typically contain one or more transition metal oxide and exhibit significant chemical and structural complexity [155], [156]. Their detailed characterization is therefore extremely difficult, and structure-property relationships can only be established in exceptional cases. Bulk mixed metal oxide catalysts are widely applied in selective oxidation, oxydehydrogenation, ammoxidation and other redox reactions. Several examples of mixed metal oxides and their application in industrial processes are summarized in Table 4.

Vanadium phosphates (e.g., $VOHPO_4 \cdot 0.5\,H_2O$) are precursors for the so-called *VPO catalysts*, which catalyze ammoxidation reactions and the selective oxidation of *n*-butane to maleic anhydride. It is proposed that the crystalline vanadyl pyrophosphate phase $(VO)_2P_2O_7$ is responsible for the catalytic properties of the VPO system. The vanadium phosphate precursor undergoes transformations in re-

ducing and oxidizing atmospheres, as shown in the following scheme [157]:

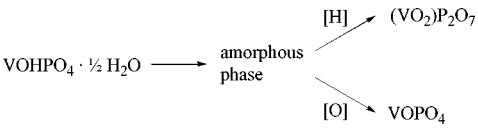

As discussed by GRASSELLI [61] effective ammoxidation (and oxidation) catalysts are multifunctional and need several key properties, including active sites which are composed of at least two vicinal oxide species of optimal metal – oxygen bond strengths. Both species must be readily reducible and reoxidizable.

The individual active sites must be spatially isolated from each other (site-isolation concept) to achieve the desired product selectivities. They should either be able to dissociate dioxygen and to incorporate the oxygen atoms into the lattice, or they must be located close to auxiliary reoxidation sites which contain metals having a facile redox couple. These sites are generally distinct from each other. They must, however, be able to communicate with each other electronically and spatially so that electrons, lattice oxygen, and anion vacancies can readily move between them. The lattice must be able to tolerate a certain density of anion vacancies without structural collapse [61]. It is clear that these complex requirements can only be achieved by multicomponent materials.

GRASSELLI [61], [158] has listed three key catalytic functionalities required for effective ammoxidation/oxidation catalysts:

1) An α-H-abstracting component, which may be Bi^{3+}, Sb^{3+}, Te^{4+}, or Se^{4+}.
2) A component that chemisorbs alkene/ammonia and inserts oxygen/nitrogen (Mo^{6+}, Sb^{5+}).
3) A redox couple such as Fe^{2+}/Fe^{3+}, Ce^{3+}/Ce^{4+}, or U^{5+}/U^{6+} to facilitate lattice oxygen transfer between bulk and surface of the solid catalyst.

An empirical correlation was found between the electron configurations of the various metal cations and their respective functionalities [61], [158] as shown in Table 5. This correlation can be used to design efficient catalysts.

Bismuth molybdates are among the most important catalysts for selective oxidation and ammoxidation of hydrocarbons [157], [61]. The phase diagram shown in Figure 10 demonstrates the structural complexity of this class of ternary oxides [159]. The catalytically most important phases lie in the compositional range Bi/Mo atomic ratio between 2/3 and 2/1 and are α-$Bi_2Mo_3O_{12}$, β-$Bi_2Mo_2O_9$, and γ-Bi_2MoO_6. An industrially used Bi molybdate catalyst was optimized in several steps and has the empirical formula $(K,Cs)_a(Ni,Co,Mn)_{9.5}(Fe,Cr)_{2.5}BiMo_{12}O_x$ [61]. This material is supported on 50 % SiO_2 and was subsequently optimized further to give a catalyst with the empirical formula $(K,Cs)_a(Ni,Mg,Mn)_{7.5}(Fe,Cr)_{2.3}Bi_{0.5}Mo_{12}O_x$.

Antimonites are a second important class of ammoxidation catalysts [61], the most important of which are those containing at least one of the elements U, Fe, Sn, Mn, or Ce, which all have multiple oxidation states. Many formulations of catalysts have been proposed over the years. Those of current commercial interest have extremely complex compositions, e.g., $Na_{0-3}(Cu,Mg,Zn,Ni)_{0-4}(V,W)_{0.05-1}Mo_{0.1-2.5}$-$Te_{0.2-5}Fe_{10}Sb_{13-20}O_x$ [61], [160].

Scheelites. Numerous multicomponent oxides adopting the scheelite ($CaWO_4$) structure with the general formula ABO_4 are known [161]. This structure tolerates cation replacements irrespective of valency provided that A is a larger cation than B and that there is charge balance. An additional property of the scheelite structure is that it is often stable with 30 % or more vacancies in the A cation sublattice. As an example, $Pb^{2+}_{1-3x}Bi^{3+}_{2x}\square_x Mo^{6+}O^{2-}_4$, where $\square$ indicates a vacancy in the Bi^{3+} (A cation) sublattice, possesses scheelite structure. The materials are active for selective oxidation of C_3 and C_4 alkenes, which involves formation of allyl species followed by extraction of O^{2-} from the lattice. Replenishment of the created vacancies occurs by oxygen chemisorption at other sites and diffusion of O^{2-} ions within the solid. The introduction of A cation vacancies has a significant effect on allyl formation, and the more open structure which prevails when cation vacancies are present facilitates O^{2-} transport.

Table 5. Electronic structure of some catalytically active elements and their functionalities [158]

α-H abstraction	Alkene chemisorption/O insertion	Redox couple	Example
Bi^{3+} $5d^{10}6s^2 6p^0$	Mo^{6+} $4d^0 5s^0$	Ce^{3+}/Ce^{4+} Fe^{2+}/Fe^{3+}	$Bi_2O_3 \cdot nMoO_3$ $M_a^{2+}M_b^{3+}Bi_xMo_yO_z$
Te^{4+} $4d^{10}5s^2 5p^0$	Mo^{6+} $4d^0 5s^0$	Ce^{3+}/Ce^{4+}	Te_2MoO_7 $(Te_aCe_bMo_v)O_z$
Sb^{3+} $4d^{10}5s^2 5p^0$ U^{5+} $5f^1 6d^0 7s^0$ Se^{4+} $3d^{10}4s^2 4p^0$	Sb^{5+} $5d^0 5s^0$ Sb^{5+} $5s^0 5p^0$ Te^{6+} $5s^0 5p^0$	Fe^{2+}/Fe^{3+} U^{5+}/U^{6+} Fe^{2+}/Fe^{3+}	$Fe_xSb_yO_z$ USb_3O_{10} $Fe_aSe_b^{4+}Te_c^{6+}O_x$

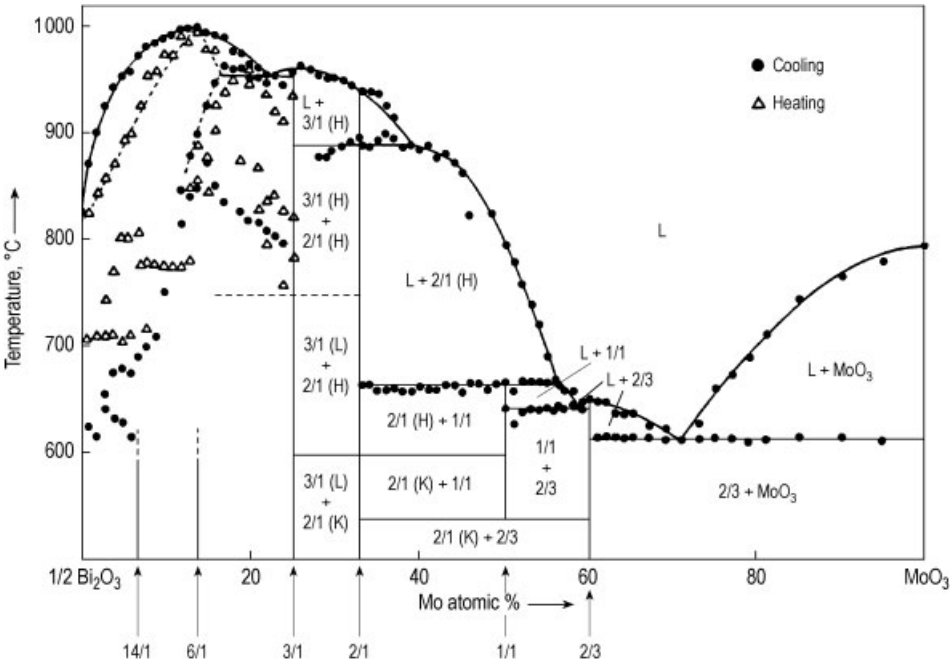

Figure 10. Phase compositions
2/3: $Bi_2O_3 \cdot 3\,MoO_3$; 1/1: $Bi_2O_3 \cdot 2\,MoO_3$; 2/1 (K): $Bi_2O_3 \cdot MoO_3$ (koechlinite); 2/1 (H): $Bi_2O_3 \cdot MoO_3$ (high-temperature form); 3/1 (L): $3\,Bi_2O_3 \cdot 2\,MoO_3$ (low-temperature form); 3/1 (H): $3\,Bi_2O_3 \cdot 2\,MoO_3$ (high-temperature form) [62]

Perovskite is a mineral ($CaTiO_3$) which is the parent solid for a whole family of multicomponent oxides with the general formula ABO_3 [136], [162]. The common feature, which also resembles that of the scheelite-type oxides, is the simultaneous presence of a small, often highly charged, B cation and a large cation A, often having a low charge. The structure also tolerates a wide variety of compositions. As an example, $La^{3+}_{1-x}Sr^{2+}_xY^{3+}O^{2-}_{3-1/2x}$ is active for methane coupling. Other applications of perovskite-type oxides in catalysis are in fuel cells, as catalysts for combustion and for $DeNO_x$ reactions.

Hydrotalcites are another family of solids which tolerate rather flexible compositions [155], [163], [164]. Hydrotalcite is a clay mineral. It is a hydroxycarbonate of Mg and Al of general formula $[Mg_6Al_2(OH)_{16}]CO_3 \cdot 4H_2O$. The compositional flexibility of the hydrotalcite lattice permits the incorporation of many different metal cations and anions to yield solids with the general formula $[M^{2+}_{1-x}M^{3+}_x(OH)]^{x+}(A^{n-})_{x/n} \cdot nH_2O$ ($M^{2+} = Mg^{2+}$, Ni^{2+}, Zn^{2+}, etc.; $M^{3+} = Al^{3+}$, Fe^{3+}, Cr^{3+}, etc.; $A^{n-} = CO_3^{2-}$, SO_4^{2-}, NO_3^-, etc.). Hydrotalcites develop large surface areas and basic properties. They have consequently been applied as solid catalysts for base-catalyzed

reactions for fine-chemicals synthesis, polymerization of alkene oxides, aldol condensation, etc. Hydrotalcite-type phases (and also malachite (rosasite)- and copper zinc hydroxycarbonate (aurichalcite)-type phases) can also be used as precursors for the synthesis of mixed oxides by thermal decomposition, for example, Cu – Zn and Cu – Zn – Cr catalysts [155].

Heteropolyanions are polymeric oxoanions (*polyoxometallates*) formed by condensation of more than two kinds of oxoanions [165], [166]. The amphoteric metals of Groups 5 (V, Nb, Ta) and 6 (Cr, Mo, W) in the +5 and +6 oxidation states, respectively, form weak acids which readily condense to form anions containing several molecules of the acid anhydride. Isopolyacids and their salts contain only one type of acid anhydride. Condensation can also occur with other acids (e.g., phosphoric or silicic) to form heteropolyacids and salts. About 70 elements can act as central heteroatoms in heteropolyanions. The structures of heteropolyanions are classified into several families according to similarities of composition and structure such as, e.g.: Keggin type $XM_{12}O_{40}^{n-}$; Dawson type $X_2M_{18}O_{62}^{n-}$; Anderson type $XM_6O_{24}^{n-}$, where X stands for the heteroatom. The most common structural feature is the Keggin anion, for which the catalytic properties have been studied extensively. Typically the M atoms in catalytic applications are either Mo or W. Heteropoly compounds can be applied as heterogeneous catalysts in their solid state. Their catalytic performance is determined by the primary structure (polyanion), the secondary structure (three-dimensional arrangement of polyanions, counter cations, and water of crystallization, etc.), and the tertiary structure (particle size, pore structure, etc.) [167], [168]. In contrast to conventional heterogeneous catalysts, on which reactions occur at the surface, the reactants are accomodated in the bulk of the secondary structure of heteropoly compounds. Certain heteropolyacids are flexible, and polar molecules are easily absorbed in interstitial positions of the bulk solid, where they form a pseudoliquid phase [167], [168].

Heteropoly compounds develop acidic and oxidizing functions, so that they can be used for acid and redox catalysis. In addition, polyanions are well-defined oxide clusters. Catalyst design is therefore possible at the molecular level. The pseudoliquid provides a unique reaction environment.

Some solid heteropolyacids have high thermal stability and are therefore suitable for vapor-phase reactions at elevated temperatures. The thermal stability of several heteropolyacids decreases in the sequence $H_3PW_{12}O_{40} > H_4SiW_{12}O_{40} > H_3PMo_{12}O_{40} > H_4SiMo_{12}O_{40}$ [167], [168]. It can be enhanced by formation of the appropriate salts [169], [170].

Because of their multifunctionality, heteropolyacids catalyze a wide variety of reactions including hydration and dehydration, condensation, reduction, oxidation, and carbonylation chemistry with Keggin-type anions of V, Mo [168], [169], [171–173]. A commercially important process, the oxidation of methacrolein, is catalyzed by a Cs salt of $H_4PVMo_{11}O_{40}$. Heteropoly salts with extremely complex compositions have been proposed, e.g., for the oxydehydrogenation of ethane. A Keggin-type molybdophosphoric salt with formula $K_2P_{1.2}MO_{10}W_1Sb_1Fe_1Cr_{0.5}Ce_{0.75}O_n$ was found to be the most efficient among the tested solids in terms of activity, selectivity, and stability [174].

3.1.2. Metals and Metal Alloys

Metals and metal alloys are used as bulk, unsupported catalysts in only a few cases. *Metal gauzes* or *grids* are used as bulk catalysts in strongly exothermic reactions which require catalyst beds of small height. A typical example are platinum – rhodium grids used for ammonia oxidation in the nitric acid process [175] and silver grids for the dehydrogenation of methane to formaldehyde.

Skeletal (*Raney-type*) *catalysts*, particularly skeletal nickel catalysts, are technologically important materials [176] which are specifically applied in hydrogenation reactions. However, their application is limited to liquid-phase reactions. They are used in particular for the production of fine chemicals and pharmaceuticals. Skeletal catalysts are prepared by the selective removal of aluminum from Ni – Al alloy particles by leaching with aqueous sodium hydroxide [176]. Besides skeletal Ni, cobalt, copper, platinum, ruthenium, and palladium catalysts have

been prepared, with surface areas between 30 and 100 $m^2 g^{-1}$. One of the advantages of skeletal metal catalysts is that they can be stored in the form of the active metal and therefore require no prereduction prior to use, unlike conventional catalysts, the precursors of which are oxides of the active metal supported on a carrier.

Fused catalysts are particularly used as alloy catalysts. The synthesis from a homogeneous melt by rapid cooling may yield metastable materials with compositions that can otherwise not be achieved [177]. Amorphous metal alloys have also been prepared (metallic glasses) [177], [178].

Oxide materials can also be fused for catalytic applications [177]. Such oxides exhibit a complex and reactive internal interface structure which may be useful either for direct catalytic application in oxidation reactions or in predetermining the micromorphology of resulting catalytic materials when the oxide is the catalyst precursor. The prototype of such a catalyst is the multiply promoted iron oxide precursor of catalysts used for ammonia synthesis [45], [179].

3.1.3. Carbides and Nitrides [45], [180]

Monometallic carbides and nitrides of early transition metals often adopt simple crystal structures in which the metal atoms are arranged in cubic close-packed (ccp), hexagonal close-packed (hcp), or simple hexagonal (hex) arrays. C and N atoms occupy interstitial positions between metal atoms (interstitial alloys). The materials have unique properties in terms of melting point (> 3300 K), hardness (> 2000 kg mm^{-2}), and strength ($> 3 \times 10^5$ MPa). Their physical properties resemble those of ceramic materials, although their electronic and magnetic properties are typical of metals. Carbon in the carbides donates electrons to the d band of the metal, thus making the electronic characteristics of, e.g., tungsten and molybdenum resemble more closely those of the platinum group metals.

Bulk carbides and nitrides, e.g., of tungsten and molybdenum, can be prepared with surface areas between 100 and 400 $m^2 g^{-1}$ by advanced synthetic procedures [180], so that they can be applied as bulk catalysts. They catalyze a variety of reactions for which noble metals are still preferentially used. Carbides and nitrides are exceptionally good hydrogenation catalysts, and they are active in hydrazine decomposition. Carbides of tungsten and molybdenum are also highly active for methane reforming, Fischer – Tropsch synthesis of hydrocarbons and alcohols, and hydrodesulfurization, and the nitrides are active for ammonia synthesis and hydrodenitrogenation [179]. The catalytic properties of carbides can be fine tuned by treatment with oxygen, which leads to the formation of oxycarbides [184]. While clean molybdenum carbide is an excellent catalyst for C – N bond cleavage (cracking of hydrocarbons), molybdenum oxidecarbide is selective for skeletal isomerization [184].

In conclusion, carbides and nitrides, especially those of tungsten and molybdenum, may well be considered as future substitutes for platinum and other metals of Groups 8 – 10 as catalysts.

3.1.4. Carbons [185]

Although carbons are frequently used as catalyst supports, they may also be used as catalysts in their own right [186], [187]. Carbons exist in a variety of thermodynamic phases (allotropes of carbon) and metastable structures, which are often ill-defined (see also → Carbon). The surface chemistry of carbons is rather complex [119], [185]. Carbon surfaces may contain a variety of functional groups, particularly those containing oxygen, depending on the provenience and pretreatment of the carbon. At a single adsorption site several chemically inequivalent types of heteroatom bonds may form. Strong interactions between surface functional groups further complicate the variety of surface chemical structures derived for the most important carbon – oxygen system. Two functions of the carbon surface act simultaneously during a catalytic reaction. Firstly, the reactants are chemisorbed selectively on the carbon surface by ion exchange via oxygen functional groups or directly by dispersion forces involving the graphite valence-electron system. The second function is the production of atomic oxygen occurring on the graphene basal faces of all sp^2 carbon materials [185].

Carbon can already be catalytically active under ambient conditions and in aqueous media. Therefore efforts have been made to apply carbons as catalysts in condensed phases. Its appli-

cation in the gas phase under oxidizing conditions is severely limited by its tendency to irreversible oxidation.

Catalytic applications of carbons include the oxidation of sulfurous to sulfuric acid, the selective oxidation of hydrogen sulfide to sulfur with oxygen in the gas phase at ca. 400 K, the reaction between phosgene and formaldehyde, and the selective oxidation of creatinine by air in physiological environments.

A potential technological application of carbon catalysts involves the catalytic removal of NO by carbon [185].

Recently, carbon nanotubes [185] find significant interest as catalysts and catalyst supports.

3.1.5. Ion-Exchange Resins and Ionomers

Ion-exchange resins (→ Ion Exchangers) are strongly acidic organic polymers which are produced by suspension copolymerization of styrene with divinylbenzene and subsequent sulfonation of the cross-linked polymer matrix [188]. This matrix is insoluble in water and organic solvents. Suspension polymerization yields spherical beads which have different diameters in the range 0.3 – 1.25 mm. The Gaussian size distribution of the beads can be influenced by the polymerization parameters.

A network of micropores is produced during the copolymerization reaction. The pore size is inversely proportional to the amount of cross-linking agent. In the presence of inert solvents such as isoalkanes during the polymerization, which dissolve the reactive monomers and precipitate the resulting polymers, beads with an open spongelike structure and freely accessible inner surface are obtained. The matrix is then a conglomerate of microspheres which are interconnected by cavities or macropores. Macroporous resins are characterized by micropores of 0.5 – 2 nm and macropores of 20 – 60 nm, depending on the degree of cross-linking.

Strongly acidic polymeric resins are thermally stable at temperatures below 390 – 400 K. Above 400 K, sulfonic acid groups are split off and a decrease in catalytic activity results.

Industrially, acidic resins are used in the production of methyl *tert*-butyl ether [189].

The ionomer *Nafion* is a perfluorinated polymer containing pendant sulfonic acid groups which is considered to develop superacidic properties. It can be used as a solid acid catalyst for reactions such as alkylation, isomerization, and acylation [190].

3.1.6. Molecularly Imprinted Catalysts [191]

Molecular imprinting permits heterogeneous *supramolecular catalysis* to be performed on surfaces of organic or inorganic materials with substrate recognition. Heterogeneous catalysts with substrate specificity based on molecular recognition require a material having a shape- and size-selective footprint on the surface or in the bulk. The stabilization of transition states by imprinting their features into cavities or adsorption sites by using stable transition-state analogues as templates is of particular interest.

Imprinted materials can be prepared on the basis of Al^{3+}-doped silica gel [192] and of cross-linked polymers [193], [194]. Chiral molecular footprint cavities have also been designed and imprinted on the surface of Al^{3+}-doped silica gel by using chiral template molecules.

When transition-state or reaction-intermediate analogues are used as templates for molecular imprinting, specific adsorption sites are created. Such molecular footprints on silica gel consist of a Lewis site and structures complementary to the template molecules. These structures can stabilize a reacting species in the transition state and lower the activation energy of the reaction, thus mimicking active sites of natural enzymes and catalytic antibodies.

Although this approach seems to have a high potential for heterogeneous catalysis, the real application of imprinted materials as catalysts still remains to be demonstrated.

3.1.7. Metal Salts

Although salts can be environmentally harmful, they are still used as catalysts in some technologically important processes. $FeCl_3 - CuCl_2$ is a catalyst for chlorobenzene production, and $AlCl_3$ is still used for ethylbenzene synthesis and *n*-butane isomerization.

3.2. Supported Catalysts

Supported catalysts play a significant role in many industrial processes. The support provides high surface area and stabilizes the dispersion of the active component (e.g., metals supported on oxides). Active phase – support interactions, which are dictated by the surface chemistry of the support for a given active phase, are responsible for the dispersion and the chemical state of the latter. Although supports are often considered to be inert, this is not generally the case. Supports may actively interfere with the catalytic process. Typical examples for the active interplay between support and active phase are bifunctional catalysts such as highly dispersed noble metals supported on the surface of an acidic carrier.

To achieve the high surface areas and stabilize the highly disperse active phase, supports are typically porous materials having high thermostability. For application in industrial processes they must also be stable towards the feed and they must have a sufficient mechanical strength.

3.2.1. Supports

Many of the bulk materials described in Section 3.1 may also function as supports. The most frequently used supports are *binary oxides* including transitional aluminas, α-Al_2O_3, SiO_2, MCM-41, TiO_2 (anatase), ZrO_2 (tetragonal), MgO etc., and ternary oxides including amorphous $SiO_2 - Al_2O_3$ and zeolites. Additional potential catalyst supports are aluminophosphates, mullite, kieselguhr, bauxite, and calcium aluminate. *Carbons* in various forms (charcoal, activated carbon) can be applied as supports unless oxygen is required in the feed at high temperatures. Table 6 summarizes important properties of typical oxide and carbon supports.

Silicon carbide, SiC, can also be used as a catalyst support with high thermal stability and mechanical strength [182]. SiC can be prepared with porous structure and high surface area by biotemplating [195]. This procedure yields ceramic composite materials with biomorphic microstructures. Biological carbon preforms are derived from different wood structures by high-temperature pyrolysis (1100 – 2100 K) and used as templates for infiltration with gaseous or liquid Si to form SiC and SiSiC ceramics. During high-temperature processing the microstructural details of the bioorganic preforms are retained, and cellular ceramic composites with unidirectional porous morphologies and anisotropic mechanical properties can be obtained. These materials show low density, high specific strength, and excellent high-temperature stability. Although they have not yet found application in catalysis, the low-weight materials may well be advantageous supports for high-temperature catalysis processes.

Table 6. Properties of typical catalyst supports

Support	Crystallographic phases	Properties/applications
Al_2O_3	mostly α and γ	SA up to 400; thermally stable three-way cat., steam reforming and many other cats.
SiO_2	amorphous	SA up to 1000; thermally stable; hydrogenation and other cats.
Carbon	amorphous	SA up to 1000; unstable in oxid. environm., hydrogenation cats.
TiO_2	anatase, rutile	SA up to 150; limited thermal stability; SCR cats.
MgO	fcc	SA up to 200; rehydration may be problematic; steam reforming cat.
Zeolites	various (faujasites, ZSM-5)	Highly defined pore system; shape selective; bifunctional cats.
Silica/alumina	amorphous	SA up to 800; medium strong acid sites; dehydrogenation cats.; bifunctional catalysts.

SA = surface area in m^2/g

Monolithic supports with unidirectional macrochannels are used in automotive emission control catalysts (→ Automotive Exhaust Control) where the pressure drop has to be minimized [196]. The channel walls are nonporous or may contain macropores. For the above application the monoliths must have high mechanical strength and low thermal expansion coefficients to give sufficient thermal shock resistance. Therefore, the preferred materials of

monolith structures are ceramics (cordierite) or high-quality corrosion-resistant steel. Cordierite is a natural aluminosilicate ($2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$). The accessible surface area of these materials corresponds closely to the geometric surface area of the channels. High surface area is created by depositing a layer of a mixture of up to 20 different inorganic oxides, which include transitional aluminas as a common constituent. This so-called *washcoat* develops internal surface areas of 50 to 300 m^2/g [197], [198].

Silica, *MCM-41*, and *polymers* can be functionalized for application as supports for the preparation of immobilized or hybrid catalysts [119], [122], [199–205]. The functional groups may serve as anchoring sites (surface bound ligands) for complexes and organometallic compounds. Chiral groups can be introduced for the preparation of enantioselective catalysts (see Section 3.2.6).

3.2.2. Supported Metal Oxide Catalysts

Supported metal oxide catalysts consist of at least one active metal oxide component dispersed on the surface of an oxide support [49], [206], [207]. The active oxides are often transition metal oxides, while the support oxides typically include transitional aluminas (preferentially γ-Al_2O_3), SiO_2, TiO_2 (anatase), ZrO_2 (tetragonal), and carbons.

Supported vanadia catalysts are extremely versatile oxidation catalysts. V_2O_5/TiO_2 is used for the selective oxidation of *o*-xylene to phthalic anhydride [208], [209] and for the ammoxidation of alkyl aromatics to aromatic nitriles [209]. The latter reaction is also catalyzed by V_2O_5/Al_2O_3 [209]. The selective catalytic reduction (SCR) of NO_x emissions with NH_3 in tail gas from stationary power plants is a major application of $V_2O_5-MoO_3-TiO_2$ and $V_2O_5-WO_3-TiO_2$ [210], [211]. $MoO_3-Al_2O_3$ and $WO_3-Al_2O_3$ (promoted by oxides of cobalt or nickel) are the oxide precursors for sulfided catalysts (see Section 3.2.5) for hydrotreating of petroleum (hydrodesulfurization, hydrodenitrogenation, hydrocracking) [46, 212, 213]. WO_3-ZrO_2 develops acidic and redox properties [214], [215]. When promoted with Fe_2O_3 and Pt it can be applied as a highly selective catalyst for the low-temperature isomerization of n-alkanes to isoalkanes [216]. $Re_2O_7-Al_2O_3$ is an efficient metathesis catalyst [217]. $Cr_2O_3-Al_2O_3$ and $Cr_2O_3-ZrO_2$ are catalysts for alkane dehydrogenation and for dehydrocyclization of, e.g., *n*-heptane to toluene [218].

The above-mentioned transition metal oxides have lower surface free energies than the typical support materials [49], [219]. Therefore, they tend to spread out on the support surfaces and form highly dispersed active oxide overlayers. These supported oxide catalysts are thus frequently called *monolayer catalysts*, although the support surface is usually not completely covered, even at loadings equal to or exceeding the theoretical monolayer coverage. This is because most of the active transition metal oxides (particularly those of V, Mo, and W) form three-dimensional islands on the support surface which have structures analogous to molecular polyoxo compounds [49], [206].

3.2.3. Surface-Modified Oxides

The surface properties, that is acidity and basicity, of oxides can be significantly altered by deposition of modifiers. The acid strength of aluminas is strongly enhanced by incorporation of Cl^- into or on the surface. This may occur during impregnation with solutions containing chloride precursors of an active component [115] or by deposition of $AlCl_3$. Chlorinated aluminas are also obtained by surface reaction with CCl_4 [119]. The presence of chlorine plays an important role in catalytic reforming with $Pt-Al_2O_3$ catalysts [47].

Strongly basic materials are obtained by supporting alkali metal compounds on the surface of alumina [220]. Possible catalysts include KNO_3, $KHCO_3$, K_2CO_3, and the hydroxides of the alkali metals supported on alumina.

Sulfation of several oxides, particularly tetragonal ZrO_2, yields strong solid acids, which were originally considered to develop superacidic properties [221–223], because, like tungstated ZrO_2 (see Section 3.2.2), they also catalyze the isomerization of *n*-alkanes to isoalkanes at low temperature.

3.2.4. Supported Metal Catalysts

Metals typically have high surface free energies [219] and therefore a pronounced tendency to reduce their surface areas by particle growth. Therefore, for applications as catalysts they are generally dispersed on high surface area supports, preferentially oxides such as transitional aluminas, with the aim of stabilizing small, nanosized particles under reaction conditions [115], [224]. This can be achieved by some kind of interaction between a metal nanoparticle and the support surface (*metal – support interaction*: MSI), which may influence the electronic properties of the particles relative to the bulk metal. This becomes particularly significant for raft-like particles of monatomic thickness, for which all atoms are surface atoms. Furthermore, the small particles expose increasing numbers of low-coordinate surface metal atoms. Both electronic and geometric effects may influence the catalytic performance of a supported metal catalyst (*particle-size effect*). Aggregation of the nanoparticles leads to deactivation.

Model supported metal catalysts having uniform particle size and structure can be prepared by anchoring molecular carbonyl clusters on support surfaces, followed by decarbonylation [225], [226]. Examples are Ir_4 and Ir_6 clusters on MgO and in zeolite cages.

Bimetallic supported catalysts contain two different metals, which may either be miscible or immiscible as macroscopic bulk alloys. The combination of an active and an inactive metal [e.g., Ni and Cu (miscible) or Os and Cu (immiscible)] dilutes the active metal on the particle surface. Therefore, the catalytic performance of reactions requiring ensembles of several active metal atoms rather than single isolated atoms is influenced [227], [228]. Selectivities of catalytic processes can thus be optimized. Typically, the surface composition of binary alloys differs from that of the bulk. The component having the lower surface free energy is enriched in the surface layer. For example, Cu is largely enriched at the surface of Cu – Ni alloys, even at the lowest concentration. Also, surface compositions of binary alloys may be altered by the reaction atmosphere.

In industrial application, supported metal catalysts are generally used as macroscopic spheres or cylindrical extrudates. By special impregnation procedures, metal concentration profiles within the pellet can be created in a controlled way. Examples are schematically shown in Figure 11 [115]. The choice of the appropriate concentration profile may be crucial for the selectivity of a process because of the interplay between transport and reaction in the porous mass of the pellet. For example for the selective hydrogenation of ethyne impurities to ethene in a feed of ethene, eggshell profiles are preferred.

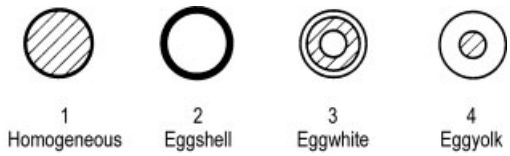

Figure 11. The four main categories of macroscopic distribution of a metal within a support

Applications of supported metal catalysts, such as noble (Pt, Pd, Rh) or nonnoble (Ru, Ni, Fe, Co) metals supported on Al_2O_3, SiO_2, or active carbon include hydrogenation and dehydrogenation reactions. Ag on Al_2O_3 is used for ethene epoxidation. Supported Au catalysts are active for low-temperature Co oxidation. Multimetal catalysts Pt – Rh – Pd on Al_2O_3 modified by CeO_2 as oxygen storage component are used on a large scale in three-way car exhaust catalysts [198]. Pt supported on chlorinated Al_2O_3 is the bifunctional catalyst used for catalytic reforming, isomerization of petroleum fractions, etc.

Modification of supported Pt catalysts by cinchona additives yields catalysts for the enantioselective hydrogenation of α-ketoesters [229].

3.2.5. Supported Sulfide Catalysts

Sulfided catalysts of Mo and W supported on γ-Al_2O_3 or active carbons are obtained by sulfidation of oxide precursors (supported MoO_3 or WO_3; see Section 3.2.2) in a stream of H_2/H_2S. They are typically promoted with Co or Ni and serve (in large tonnage) for hydrotreating processes of crude oil, including hydrodesulfurization (HDS) [46], [212], [213], hydrodenitrogenation HDN [213], and hydrodemetalation HDM [230]. Currently, the CoMoS and NiMoS models are most accepted for describing the active phase. These phases consist of a single MoS_2 layer or stacks of MoS_2 layers in which

the promotor atoms are coordinated to edges [46], [213], as shown in Figure 12. This figure also indicates that Co may simultaneously be present as Co_9S_8 and as a solid solution in the Al_2O_3 support matrix. It is inferred that the catalytic activity of the MoS_2 layers is related to the creation of sulfur vacancies at the edges of MoS_2 platelets. These vacancies have recently been visualized on MoS_2 crystallites by scanning tunneling microscopy (STM) [231].

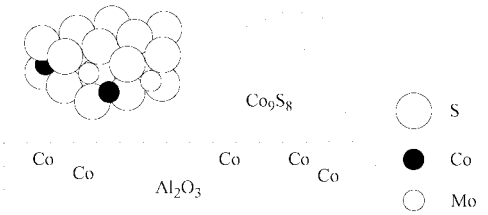

Figure 12. Three forms of Co present in sulfided $Co-Mo/Al_2O_3$ catalysts: as sites on the MoS_2 edges (the so-called Co – Mo – S phase), as segregated Co_9S_8, and as Co^{2+} ions in the support lattice.

3.2.6. Hybrid Catalysts [199], [201], [202], [204], [205]

Hybrid catalysts combine homogeneous and heterogeneous catalytic transformations. The goal of the approach is to combine the positive aspects of homogeneous catalysts or enzymes in terms of activity, selectivity, and variability of steric and electronic properties by, e.g., the appropriate choice of ligands (including chiral ligands [232]) with the advantages of heterogeneous catalysts such as ease of separation and recovery of the catalyst. This can be achieved by *immobilization* (*heterogenization*) of active metal complexes, organometallic compounds, or enzymes on a solid support.

There are several routes for the synthesis of immobilized homogeneous catalysts:

1) Anchoring the catalytically active species via covalent bonds on the surface of suitable inorganic or organic supports such as SiO_2, mesoporous MCM-41, zeolites, polystyrenes, and styrene – divinylbenzene copolymers [199–201], [205]. The polymerization or copolymerization of suitably functionalized monomeric metal complexes is also known.
2) Chemical fixation by ionic bonding using ion exchange.
3) Deposition of active species on surfaces of porous materials by chemi- or physisorption, or chemical vapor deposition (CVD). The "ship-in-bottle" principle belongs to this synthetic route, but is treated separately in Section 3.2.7.
4) Moleculary defined surface organometallic chemistry may also yield immobilized active organometallic species.

The reagents for covalent bonding on siliceous materials (SiO_2, MCM-41) are often alkoxy- or chlorosilanes which are anchored to the surface by condensation reactions with surface hydroxyl groups [199–201], [205]. Functional groups thus created on the surface can include phosphines, amines etc., which serve as anchored ligands for active species that undergo ligand-exchange reactions. Careful control of the density of functional groups leads to spatial separation of active complexes (site isolation) and thus helps to avoid undesired side reactions [233].

Immobilized enzymes (→ Immobilized Biocatalysts) are frequently used in biocatalysis and in organic synthesis. The synthesis and catalytic performance of this class of heterogenized materials is discussed in several review articles [205], [234].

Dendrimers [235] which are functionalized at the ends of the dendritic arms can be used for immobilization of metal complexes. A catalytic effect is thus generated at the periphery of the dendrimer. Dendrimers with core functionalities have also been synthesized. The resemblance of the produced structures to prosthetic groups in enzymes led to the introduction of the word *dendrizymes* [236]. Dendrimers have found application, e.g., in membrane reactors.

Immobilized homogeneous catalysts are used for selective oxidation reactions, for hydrogenation, and for C – C coupling reactions. They have proved very efficient in asymmetric synthesis [201], [204], [237].

Special processes with immobilized catalysts are *supported (solid) liquid-phase catalysis* (SLPC) [238] and *supported (solid) aqueous-phase catalysis* (SAPC) [239]. In SLPC a solution of the homogeneous catalyst in a high-boiling solvent is introduced into the pore vol-

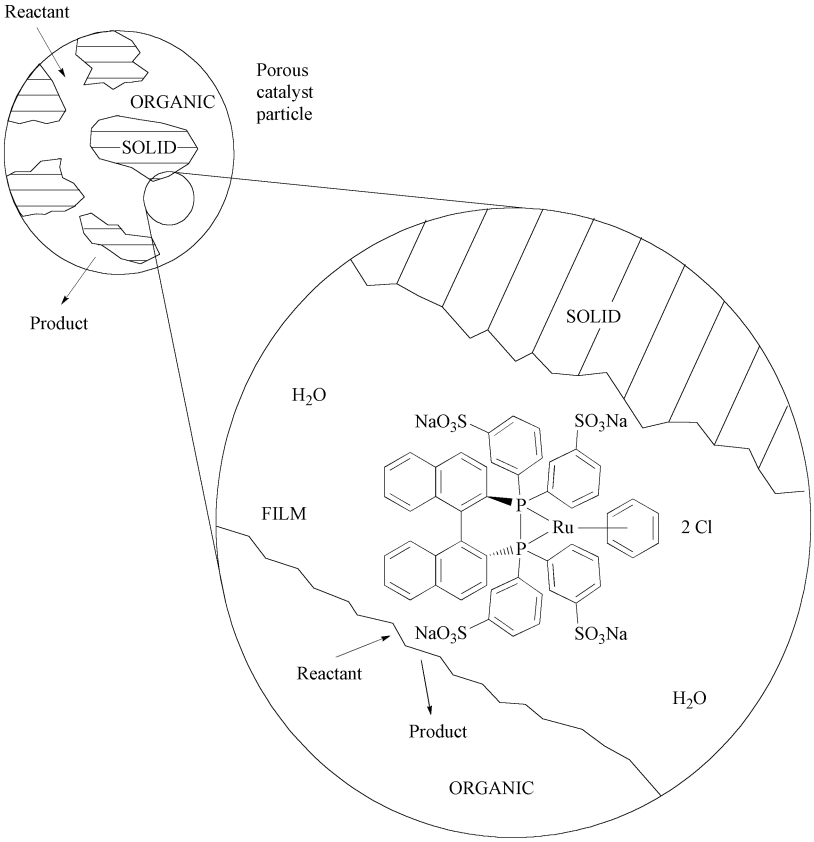

Figure 13. Schematic representation of SAP catalysis

ume of a porous support by capillary forces, and the reactants pass the catalyst in the gas phase. For example, the active phase – a mixture of vanadium pentoxide with alkali metal sulfates or pyrosulfates – is present as a melt in the pores of the SiO_2 support under the working conditions of the oxidation of SO_2 [240]. In SAPC hydrophobic organic reactants are converted in the liquid phase. The catalyst consists of a thin film of water on the surface of a support (e.g., porous SiO_2) and contains an active hydrophilic organometallic complex [239]. The reaction takes place at the interface between the water film and an organic liquid phase containing the hydrophobic reactant. The nature of these catalyst systems is schematically shown in Figure 13.

3.2.7. Ship-in-a-Bottle Catalysts [241]

Metal complexes which are physically entrapped in the confined spaces of zeolite cages (*confined catalysts*) are known as *ship-in-a-bottle catalysts* or *tea-bag catalysts*. The entrapped complexes are assumed to retain many of their solution properties. The catalytic performance can be modified in a synergistic manner by shape selectivity, the electrostatic environment, and the acid-base properties of the zeolite host. Ligands for metal centers in the zeolite cages include ethylenediamine, dimethylglyoxime, various Schiff bases, phthalocyanines, and porphyrins [241], [242]. The entrapped complexes can be obtained via three principal routers [241]:

1) Reaction of the preformed flexible ligand with transition metal previously introduced into the zeolite cages (flexible ligand method). The synthesis of a zeolite entrapped metal-salen complex is schematically shown in Figure 14.
2) Assembling the ligand from smaller species inside the zeolite cavities (ship-in-a-bottle

technique). For example, the synthesis of a zeolite-entrapped metal phthalocyanine is schematically shown in Figure 15.

3) Synthesis of the zeolite structure around the preformed transition metal complex (zeolite synthesis technique).

The success of ship-in-a-bottle catalysts in catalytic processes has still to be demonstrated.

Zeolite-encapsulated complexes have also been suggested as model compounds for mimicking enzymes. These zeolite-based enzyme mimics are called *zeozymes* to describe a catalytic system, in which the zeolite replaces the protein mantle of the enzyme, and the entrapped metal complex mimics the active site of the enzyme (e.g., an iron porphyrin) [243].

Host – guest supramolecular compounds may also be mentioned in this context [191], [244].

3.2.8. Polymerization Catalysts

Ziegler – Natta (→ Polyolefins, Chap. 1.3.2.2.) catalysts are mixtures of solid and liquid compounds containing a transition metal such as Ti or V [245], [246]. $TiCl_4$ combined with $Al(C_2H_5)_3$ or other alkylaluminum compounds were found to be active for olefin polymerization. More active catalysts were produced commercially by supporting the $TiCl_4$ on solid $MgCl_2$, SiO_2 or Al_2O_3 to increase the amount of active titanium. Currently, Ziegler – Natta catalysts are produced by ball milling $MgCl_2$ with about 5 % of $TiCl_4$, and the cocatalyst is $Al(C_2H_5)_3$.

The *Phillips catalyst* (→ Polyolefins, Chap. 1.3.2.1.) consists of hexavalent surface chromate on high surface area silicate supports. Cr^{6+} is reduced by ethylene or other hydrocarbons, probably to Cr^{2+} and Cr^{3+}, the catalytically active species [245], [246].

More recently, so-called single-site catalysts using metallocenes as active species were developed (→ Metallocenes, → Polyolefins, Chap. 1.3.2.3.) [246], [247]. The activity of these materials is dramatically enhanced by activation with methylaluminoxane (MAO), obtained by incomplete hydrolysis of $Al(CH_3)_3$, the catalytic performance of which is significantly more versatile than that of the classical Ziegler – Natta or Phillips catalysts. Activities and the nature of the polymeric product can be taylored by the choice of the metal and ligands.

4. Production of Heterogeneous Catalysts

The develompent of heterogeneously catalyzed reactions for the production of chemicals initiated the preparation of the required catalysts on a technical scale. Up to the end of World War II, solid catalysts were produced predominantly in process companies such as IG Farben and BASF in Germany, and Standard Oil Company and UOP in the USA [248], [249]. About ten years later some independent catalyst producing companies were founded in the USA, Western Europe, and Japan [248]. At present more than 15 international companies [248], [250] are producing solid catalysts on multitonne scale; for example:

- Engelhard Corp. (incl. Harshaw Catalyst)
- ICI Catalysts and ICI Catalco
- Davison Chemicals and Grace
- SÜD-Chemie Catalyst Group (incl. UCI, Houdry, Prototec in the USA, NGC, CCIFE in Japan, UCIL in India and AFCAT, SYNCAT in South Africa)
- UOP and Katalystiks
- BASF
- Monsanto
- Shell and Criterion Catalysts
- Akzo Chemicals
- Johnson Matthey
- Haldor Topsøe
- Calsicat
- Degussa
- Nippon Shokubai
- Nikki Chemical

In 1991 the catalyst world market achieved a turnover of about $\$ 6 \times 10^9$ [248], [249], [251], grew to $\$ 7.3 \times 10^9$ in 1994, and reached $\$ (8-9) \times 10^9$ in 1996.

Approximately 24 – 28 % of produced catalysts were sold to the chemical industry and 38 – 42 % to petrochemical companies including refineries. 28 – 32 % of solid catalysts were used in environmental protection, and 3 – 5 % in the production of pharmaceuticals [248], [251].

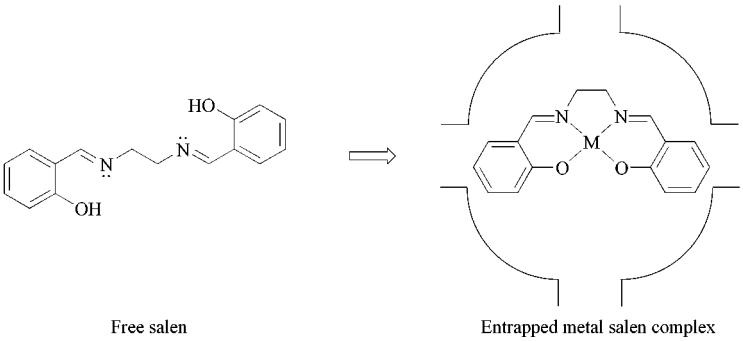

Figure 14. Synthesis of zeolite-entrapped metal salen complexes by the flexible ligand method

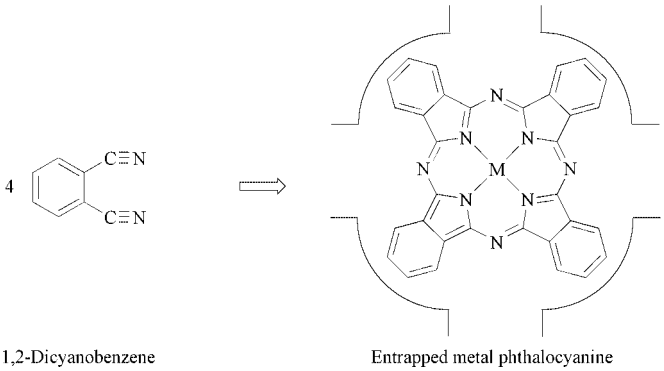

Figure 15. Ship-in-a-bottle synthesis of zeolite-encaged metal phthalocyanines

The catalytic properties of solid catalysts are strongly affected by the preparation method, production conditions, and quality of source materials. Therefore, it is necessary to control each production step and the physical or mechanical properties of all intermediates. To attain a better reproducibility of catalyst production, batch procedures were mainly replaced by continuous operations, such as precipitation, filtration, drying, calcination, and forming.

Automatization of various operations and computer control of different equipments were installed in catalyst production lines [248]. Recently, SPC (statistic process control) and QA (quality assurance) were integrated into the catalyst production process. Some companies, especially in Western Europe and in the USA, produce solid catalysts according to ISO Standard which guarantees a standard catalyst quality to the customer [248], [252].

Catalysts applied in several industrial processes can be subdivided into the following categories:

- Unsupported (bulk) catalysts
- Supported catalysts
- Skeletal catalysts

4.1. Unsupported Catalysts

Unsupported catalysts represent a large category and are applied in numerous industrial processes. Various preparation methods were adopted in the past decades in the commercial production of unsupported catalyst, such as mechanical treatment or fusion of catalyst components, precipitation, coprecipitation, flame hydrolysis, and hydrothermal synthesis [177], [250], [253–257].

Mechanical treatment, for example, mixing, milling, or kneading of catalytic active materials or their precursors with promoters, structure stabilizers, or pore-forming agents, is one of the simplest preparation methods [250], [253–255]. In some cases, however, the required intimate contact of catalyst components could not

be achieved and therefore the activity, selectivity, or thermal stability of catalysts prepared in this way was lower than of those prepared by other methods. However, recent improvements in the efficiency of various aggregates for the mechanical treatment of solids resulted in activity enhancement. An important advantage of these methods is that formation of wastewater is avoided.

Industrial catalysts produced by mechanical treatment are summarized in Table 7.

Table 7. Unsupported catalysts prepared by mechanical treatment (MT) or by fusion (F)

Catalyst∗	Preparation method	Application
Fe_2O_3(K, Cr, Ce, Mo)	MT	ethylbenzene dehydrogenation (styrene production)
Fe_2O_3(K)	MT	Fischer – Tropsch synthesis
$ZnO - Cr_2O_3$	MT	hydrogenation of carbonyl compounds
Fe_3O_4(K, Al, Ca, Mg)	F	NH_3 synthesis
$V_2O_5 - K_2S_2O_7$	F	SO_2 oxidation (H_2SO_4 production)
Pt/Rh grid	F	NH_3 oxidation (HNO_3 production)

∗ Elements in parentheses are promoters.

Fusion of components or precursors is used for the production a small group of unsupported catalysts. The fusion process [177] permits the synthesis of alloys consisting of elements which do not mix in solution or in the solid state. However, preparation of unsupported catalysts by fusion is an energy-consuming and quite expensive process.

The most important application of this method is the production of ammonia synthesis catalysts based on the fusion of magnetite (Fe_3O_4) with promoters such as oxides of K, Al, Ca, and Mg [177]. Another example is the preparation of SO_2 oxidation catalysts by fusion of V_2O_5 with K pyrosulfate ($K_2S_2O_7$) [177]. Some producers incorporate Cs oxide as an activity promoter in this catalyst.

Quite recently, amorphous alloys composed, e.g., of Pd and Zr, so-called *metallic glasses* were found to be active in catalytic oxidations [177], [178].

Industrial catalysts produced by the fusion process are listed in Table 7.

Precipitation and coprecipitation are the most frequently applied methods for the preparation of unsupported catalysts or catalyst supports [250], [253–257]. However, both methods have the major disadvantage of forming large volumes of salt-containing solutions in the precipitation stage and in washing the precipitate.

Source materials are mainly metal salts, such as sulfates, chlorides, and nitrates. Acetates, formates, or oxalates are used in some cases. In industrial practice nitrates or sulfates are preferred. Basic precipitation agents on an industrial scale are hydroxides, carbonates, and hydroxocarbonates of sodium, potassium, or ammonium.

Then alkali metal nitrates or sulfates formed as precipitation by-products must be washed out of the precipitate. Thermally decomposable anions, e.g., carbonates, carboxylates, and cations such as NH_4^+, are especially favored in catalyst production.

Coprecipitation of two or more metal cations is a suitable operation for the homogeneous dispersion of the corresponding oxides, especially if the catalyst precursors have a defined crystalline structure, for example, $Cu(OH)NH_4CrO_4$ or $Ni_6Al_2(OH)_{16}CO_3$. After thermal treatment, binary oxides such as $CuO - Cr_2O_3$ and $NiO - Al_2O_3$ are formed [250], [253–257].

Precipitation and coprecipitation can be carried out in batch or continuous operations.

If the metal salt solution is placed in the precipitation vessel and the precipitating agent is added, then the pH changes continuously during the precipitation. Coprecipitation should be carried out in the reverse manner (addition of the metal salt solution to the precipitation agent) to avoid sequential precipitation of two or three metal species.

If the metal salt solution and the precipitating agent are simultaneously introduced into the precipitation vessel, then it is possible to keep the pH constant. However, the residence time of the precipitate in the vessel changes continuously.

Finally, if the metal salt solution and the precipitating agent are continuously introduced in the precipitation vessel, and the reaction products are removed continuously, then pH and res-

idence time can be kept constant [250], [253–257].

Besides pH and residence time, other precipitation parameters, such as temperature, agitation, and concentration of starting solutions, affect the properties of the precipitate. The choice of anions, the purity of raw material, and the use of various additives also play an important role [250], [253–257].

In general, highly concentrated solutions, low temperatures and short ageing times result in finely crystalline or amorphous materials which are difficult to wash and filter. Lower concentrations of the solutions, higher temperatures, and extended ageing provide coarse crystalline precipitates which are easier to purify and separate [250], [253–257].

The industrial production of precipitated catalysts usually involves the following steps:

- Preparation of metal salt solution and of precipitating agent (dissolution, filtration)
- Precipitation
- Ageing of the precipitate
- Washing of the precipitate by decantation
- Filtration
- Washing of the filtercake (spray drying)
- Drying
- Calcining
- Shaping
- Activation

Operations such as filtration, drying, calcination etc. are discussed in Section 4.4.

Typical unsupported industrial catalysts produced by precipitation or coprecipitation are compiled in Table 8.

The sol – gel process [258] involves the formation of a sol, followed by the creation of a gel. A sol (liquid suspension of solid particles smaller than 1 µm) is obtained by the hydrolysis and partial condensation of an inorganic salt or a metal alkoxide. Further condensation of sol particles into a three-dimensional network results in the formation of a gel. The porosity and the strength of the gel are strongly affected by conditions of its formation. For example, slow coagulation, elevated temperature, or hydrothermal posttreatment increase the crystalline fraction of the gel.

Alumina and silica can be produced from sodium aluminate or sodium silicate by treatment with nitric, hydrochloric, or sulfuric acid. In this process, first sols and then gels are formed. Washing the sodium from the gels is essential [255], [256], [258].

Table 8. Catalysts (their precursors) or supports prepared by precipitation or coprecipitation

Catalysts (precursors) or supports	Source materials	Application
Alumina	Na aluminate, HNO_3	support, dehydration, Claus process
Silica	Na silicate (water glass), H_2SO_4	support
Fe_2O_3	$Fe(NO_3)_3$, NH_4OH	ethylbenzene dehydrogenation (styrene production)
TiO_2	Fe titanate, titanyl sulfate, NaOH	support, Claus process, NO_x reduction
$CuO-ZnO-(Al_2O_3)$	Cu, Zn, (Al) nitrates, Na_2CO_3	LTS, methanol synthesis
Fe molybdate	$Fe(NO_3)_3$, $(NH_4)_2MoO_4$, NH_4OH	methanol oxidation to formaldehyde
Vanadyl phosphate	vanadyl sulfate, $NaHPO_4$	butane oxidation to maleic acid anhydride
$NiO-Al_2O_3$	Ni, Al nitrates, Na_2CO_3	hydrogenation of aromatics
$NiO-SiO_2$	Ni nitrate, Na silicate, Na_2CO_3	hydrogenation of aromatics

Spherical silica or silica-alumina gels are produced directly by injecting drops of a gelling mixture into oil at a proper rate to allow setting of the gel. To avoid bursting during drying, the gel beads are washed to reduce the salt content ($NaNO_3$, NaCl, or Na_2SO_4). Finally, the beads are dried and calcined [255], [256].

Based on the sol – gel process, high-purity materials such as alumina, TiO_2, ZrO_2 are produced on an industrial scale. Raw materials are the corresponding metal alkoxides, e.g., $Al(C_{12}-C_{18}$ alkoxide$)_3$ and $Ti(n\text{-}C_4H_9O)_4$ [258].

Flame Hydrolysis. In flame hydrolysis [259] a mixture of the catalyst or support precursor, hydrogen, and air is fed into the flame of a continuously operating reactor. Precursors (mainly chlorides such as $AlCl_3$, $SiCl_4$, $TiCl_4$ or $SnCl_4$) are hydrolyzed by steam (formed by H_2 oxidation). The products of flame hydro1ysis are oxides. More than 100 000 t/a of so-called fumed silica, alumina, or titania are produced

by Degussa, Wacker (both Germany), and Cabot (USA).

Thermal decomposition of metal – inorganic or metal – organic catalyst precursors is sometimes used in industrial catalyst production. For example, mixtures of Cu- and $Zn(NH_3)_4(HCO_3)_2$ decompose at 370 K to form binary Cu – Zn carbonates, which are transformed during calcination into the corresponding binary oxides, used as low temperature water gas shift catalysts [255].

Industrial production of Cu – Cr oxides (copper chromites), used in the hydrogenation of carbonyl compounds, is based on the thermal decomposition of a basic copper ammonium chromate [$CuNH_4(OH)CrO_4$] at 620 – 670 K [255], [256].

Highly active Ni catalysts for the hydrogenation of fats and oils are obtained by the thermal decomposition of Ni formate [$(HCOO)_2Ni$] at 390 – 420 K. The decomposition is usually carried out in hard fat, which protects Ni against oxidation [255].

Catalysts or supports produced by flame hydrolysis or thermal decomposition of inorganic complexes are summarized in Table 9.

Hydrothermal synthesis [125], [161] is a very important preparation method for zeolites and other molecular sieves.

Currently, the importance of zeolites in industrial catalysis is still increasing. They are used as catalysts or supports not only in petrochemical operations but also in the production of fine chemicals.

In hydrothermal synthesis (see also → Zeolites, Chap. 6.1.) a mixture of silicon and aluminium compounds containing alkali metal cations, water, and in some cases organic compounds (so-called templates) is converted into microporous, crystalline aluminosilicates [256], [260].

Common sources of silicon are colloidal silica, water glass, fumed silica, and silicon alkoxides. Aluminum can be introduced as aluminium hydroxide, metahydroxide, or aluminate salts. Common templates are tetrapropyl- or tetraethylammonium bromides or hydroxides [256], [260].

Hydrothermal synthesis is a complex process consisting of three basic steps: achievement of supersaturation, nucleation, and crystal growth. It is affected by the hydrogel molar composition, a1kalinity, temperature, and time [256], [260].

In general, synthesis is carried out at 360 – 450 K under atmospheric or autogenous water pressure (0.5 – 1 MPa) with residence times of 1 – 6 d. After synthesis, the crystalline product is separated by filtration or centrifugation, washed, dried, and calcined. Sodium-containing zeolites, which are the products of the hydrothermal synthesis, are converted into acidic forms by exchange of sodium ions with the ammonium, followed by thermal treatment [256], [260].

Zeolites used in the industrial catalysis are above all Y zeolite, mordenite, ZSM-5, ZSM-11, and zeolite β [125], [160].

To improve the thermal stability of zeolites, especially of Y zeolite, Al ions are extracted from the lattice by steaming or acid treatment. For example, fluid-cracking catalysts (amorphous aluminosilicates) contain 10 – 50 % of ultrastable Y zeolites [256], [260].

Related to zeolites are other molecular sieves such as aluminum phosphates (AlPO) and silicoaluminum phosphates (SAPO), the importance of which in industrial catalysis is growing. SAPO-11 was applied recently in the isomerization of cyclohexanone oxime to ε-caprolactam, instead of sulfuric acid in a demonstration unit [125].

Other preparation methods include condensation of more than two kinds of oxoanions, such as MoO_4^{2-}, WO_4^{2-}, HPO_4^{2-}, etc. to give heteropolyacids such as $H_3PW_{12}O_{40}$ or $H_3PMo_{12}O_{40}$ [167].

In industrial practice, the source materials are Na_2HPO_4, Na_2WO_4, or Na_2MoO_4 solutions. Hydrolysis and subsequent condensation are carried out with HCl. The heteropolyacids are extracted with organic so1vents. Heteropolyacids are very strong acids with Hammett acidity function $H_o < -8$. They have found industrial application in acid-catalyzed reactions conducted in the liquid phase, such as hydration, esterification, and alkylation. Their activity is evidently higher than that of inorganic acids [167]. Their K or Cs salts are used as catalysts in the selective vapor phase oxidation of propene to acrolein or isobutene to methacrolein [167].

Table 9. Catalysts (their precursors) or supports prepared by sol – gel process (SG), flame hydrolysis (FH), or thermal decomposition (TD) of inorganic metal complexes

Catalyst (precursor) or support	Source material	Preparation method	Application
Alumina (high-purity)	Al alkoxides	hydrolysis, SG	support for noble metals
Alumina (acidic, low bulk density)	$AlCl_3$	FH	support or additive
Silica (low bulk density)	$SiCl_4$	FH	support or additive
TiO_2	$Ti(n\text{-}C_4H_9O)_4$	hydrolysis, SG	support
TiO_2 (low bulk density)	$TiCl_4$	FH	support or additive
CuO – ZnO	Cu, Zn $(NH_3)_4HCO_3$	TD	low-temperature shift, methanol
CuO – Cr_2O_3	$Cu(NH_4)OHCrO_4$	TD	hydrogenation of carbonyl compounds
Ni (kieselguhr)	Ni formate	TD	hydrogenation of fats and oils

Another preparation method is based on the treatment of alumina or aluminosilicate with gases such as HF, HCl, BF_3, $AlCl_3$ to create highly acidic centers. Such catalysts are active in the skeletal isomerization of hydrocarbons, e.g., n-C_4 or n-C_5 [256].

4.2. Supported Catalysts

The main feature of supported catalysts is that the active material forms only a minor part and is deposited on the surface of the support [254–256], [261].

In some cases, the support is more or less inert, e.g., α-alumina, kieselguhr, porous glass, ceramics. In other cases the support takes part in the catalytic reaction, as in the case of bifunctional catalytic systems, e.g., alumina, aluminosilicate, zeolites, etc. [254–256], [261].

Additionally, some supports can alter the catalytic properties of the active phase. This so-called strong metal – support interaction (SMSI) can decrease, for example, the chemisorption capacity of supported metals (Pt – TiO_2) or can hinder the reduction of supported metal oxides (Ni silicate, Ni and Cu aluminates, etc.) [254–256].

4.2.1. Supports

Currently, various industrial supports are available in a multitonne quantities possessing a wide range of surface areas, porosities, shapes, and sizes.

Widely used supports include alumina, silica, kieselguhr, porous glass, aluminosilicates, molecular sieves, activated carbon, titania, zinc oxide, silicates such as cordierite ($2\,MgO \cdot Al_2O_3 \cdot 5\,SiO_2$) and mullite ($3\,Al_2O_3 \cdot 2\,SiO_2$), and Zn and Mg aluminates [254–256].

The supports are produced by specialized producers or directly by catalyst producers. Well known support manufacturers are:

- Davison-Grace (USA, UK, Germany)
- Bayer (Germany)
- Alcoa (USA)
- Kaiser Chem. (USA)
- RWE-DEA, CONDEA (Germany)
- Filtrol-Engelhard (USA)
- Norton (USA)
- Degussa (Germany)
- Cabot Corp. (USA)

In the past, mostly natural supports, e.g., bauxite, pumice and kieselguhr, were used in catalyst production. At present (with the exception of kieselguhr) mainly synthetic supports with "tailored" physical properties are preferred in industrial catalysis.

Because the majority of supports also have catalytic properties (e.g., alumina, aluminosilicates, zeolites) their production methods are described in Section 4.1.

4.2.2. Preparation of Supported Catalysts

The broad application of supported catalysts in industrial catalysis led to the development of numerous preparation methods applicable on a technical scale. Some of these methods are identical with those used in the production of unsupported catalysts, e.g., mechanical treatment, precipitation, thermal decomposition of metal – inorganic or metal – organic complexes and therefore they will be discussed here very briefly.

Mechanical treatment, e.g., kneading of a catalyst precursor with a support is applied, e.g., in the production of kieselguhr-supported Ni (precursor $NiCO_3$) [253], [255], [256]. Also, MoO_3 supported on Al_2O_3 is sometimes produced by this process. However, the distribution of the active phase on the support is in some cases not sufficient [253].

Better results are obtained by the combination of mechanical and thermal treatment which, results in spreading [49] of, e.g., MoO_3 on Al_2O_3 or of V_2O_5 on Al_2O_3 or TiO_2.

Impregnation by pore filling of a carrier with an active phase is a frequently used production method for supported catalysts [115], [253]. The object of this method is to fill pores of the support with a solution of the catalyst precursor, e.g., a metal salt of sufficient concentration to achieve the desired loading. If higher loadings with active phases are required, it is mostly necessary to repeat the impregnation after drying or calcination of the intermediate. Examples of catalysts prepared by the pore filling method are Ni or Co on $Al_2O_3 - MoO_3$, MoO_3 on alumosilicates including zeolites, Ni or Ag on α-alumina, noble metals on active carbon, etc. [250], [253–256].

Adsorption is a very good method to achieve uniform deposition of small amounts of active component on a support. Powders or particles exposed to metal salt solutions adsorb equilibrium quantities of salt ions, in accordance with adsorption isotherms. Adsorption may be either cationic or anionic, depending on the properties of the carrier surface. For example, alumina (depending on the adsorption conditions, mainly on the pH of the solution) adsorbs both cations and anions. Silica weakly adsorbs cations, while magnesia strongly adsorbs anions [250].

Adsorption of $PdCl_2$ from aqueous solution on different aluminas is very fast, and a high equilibrium concentration (ca. 2 wt %) can be obtained. The Pd deposition takes place mainly in an outer shell (egg shell profile) of shaped particles [250], [256].

With $H_2[PtCl_4]$ only 1 wt % Pt loading on alumina is possible owing to from the flat adsorption isotherm [250], [256].

The addition of oxalic, tartaric, and citric acid to the metal salt solution changes the profiles of active component on the carrier. In general, with increasing acid strength the metal ions are forced deeper into the support particles [250].

Ion exchange [115], [250] is very similar to ionic adsorption but involves exchange of ions other than protons. Lower valence ions, such as Na^+ or NH_4^+ can be exchanged with higher valence ions, for example, Ni^{2+} or Pt^{4+}. This method is used mainly in the preparation of metal-containing zeolites, e.g., Ni- or Pd-containing Y zeolites or mordenites used in petroleum-refining processes.

Thermal decomposition of inorganic or organic complexes in the presence of a support. This method is identical with that used in the preparation of unsupported catalysts by thermal decomposition of precursors. Supports can be used either as powder or preshaped. For example, Ni or Co deposited on kieselguhr or silica is produced by this method [255].

Precipitation onto the support is carried out in a similar way as in the case of unsupported catalysts [250], [253], [261]. Supports, mainly as powders, are slurried in the salt solution, and alkali is added. Rapid mixing is essential to avoid precipitation in the bulk.

Uniform precipitation can be achieved by using urea rather than conventional alkalis [261]. An appropriate amount of urea is added to the metal salt – support slurry and the mixture is heated while stirring. At 360 K urea decomposes slowly to NH_3 and CO_2, and precipitation takes place homogeneously over the surface and in pores of the support. This method is called *deposition – precipitation* [261] and is used especially in the production of highly active $Ni - SiO_2$ or $Ni - Al_2O_3$ catalysts.

Coating of nonporous supports [198] with a thin layer of high surface area material is an important step in the production of catalysts for pollution control. Oxides such as Al_2O_3, CeO_2, and ZrO_2 (washcoat) are deposited on monoliths with a honeycomb like structure by dipping into an aqueous slurry containing primary particles (about 20 nm in diameter) of these materials.

Reductive deposition is a preparation method in which especially precious metals are

deposited on the carrier surface by reduction of aqueous metal salts, mainly chlorides or nitrates, with agents such as H_2, Na formate, formaldehyde, and hydrazine. Examples of commercial catalysts produced by this method are precious metals on active carbon, SiO_2 or α-Al_2O_3. Reductive deposition is preferred especially in the case of bimetallic supported catalysts such as Pt – Re or Pd – Rh [262].

Heterogenization of homogeneous catalysts is based on the binding of metal complexes to the surface or entrapment in pores of the inorganic or organic support [205]. Such catalysts are used mainly in stereospecific hydrogenations in the production of fine chemicals or pharmaceuticals.

Enzymes [205] can also be heterogenized. They found industrial application in biochemical processes. A prominent example is the isomerization of glucose to fructose in the production of soft drinks.

4.3. Skeletal Catalysts

Skeletal catalysts, also called porous metals, consist of the metal skeleton remaining after the less noble component of an alloy was removed by leaching with alkali, preferentially NaOH. Skeletal catalysts were discovered 1925 and introduced into chemistry by Raney, and therefore some bear his name, e.g., Raney Ni or Co.

The group of skeletal catalysts [176] includes Ni, Co, Cu, Fe, Pt, Ru, and Pd. The second alloy component can be Al, Si, Zn, or Mg, with Al being used preferentially.

Small amounts of a third metal such as Cr or Mo have been added to the binary alloy as activity promoter.

Ni – Al, Cu – Al, and Co – Al alloys with different grain sizes are commercially available and can be leached out before use. Davison-Grace and Degussa provide finished extremely pyrophoric Ni or Co skeletal catalysts protected by water in commercial quantities. Skeletal Ni found technical application in the hydrogenation of aliphatic or aromatic nitro compounds and nitriles.

4.4. Unit Operations in Catalyst Production

As in other branches of the chemical industry, unit operations have also been established in catalyst production in the past few decades, for example, in operations such as filtration, drying, calcination, reduction and catalyst forming [250], [252], [256]. Continuous operations are favored because of larger throughputs, lower operating costs, and better quality control. Additionally, factors such as environmental pollution and hazards to human health can be minimized more easily.

Filtration, Washing. The main purpose of these operations is to separate precipitates and to remove byproducts and possible impurities. In batch operations mainly plate-and-frame filter presses are used [250], [256]. The washing of the filter cake proceeds in countercurrent to the direction of the filtration.

Continuous vacuum rotary filters are widely used. By changing the speed of the filter drum (covered with filter cloth), the quantity of slurry filtered and the thickness of the filter cake can be varied over a broad range. As the drum rotates, there is a washing phase in which water is sprayed against the moving filter cake. Finally, the filter cake is scraped or blown to remove it from the drum [256].

Another filtration equipment is the centrifuge. However, its application is possible only when the filtered material is grainy or crystalline, e.g., zeolites. Washing can be carried out by introducing washing water into the centrifuge. Centrifuges can operate either in a discontinuous or continuous manner [256].

Drying. Because drying conditions such as rate, temperature, duration, or gas flow rate can change the physical properties of the resulting material, it is important to measure and keep these parameters constant. For example the porosity of precipitated catalysts depends on the drying procedure. The drying of impregnated supports can change the distribution of active components. Their uniform distribution can be obtained only if all the liquid is evaporated spontaneously [250], [253], [256]. Usually, drying proceeds up to 400 K.

For the drying of filter cake, various tools are used, e.g., box furnaces with trays, drum dryers, rotary kilns, and spray dryers [250], [253], [256].

The main problems with drums and rotary dryers are the feeding of the wet filter cake and removal of adhering material from the walls. Because lumps are usually formed in the drying process, the resulting material must pass a granulator equipped with a sieve [256].

Spray dryers provide microspherical materials with a narrow particle-size range. Spray dryers are equipped with a nozz1e or a rotating disk to disperse the watery slurry of the filter cake in a stream of hot air [250], [256], [263].

All the above drying equipment operates in continuous mode.

Small batches of catalyst precursors are dried in box furnaces with trays [253].

For the drying of extrudates, continuously operating belt dryers have found technical application [256].

Calcination The main object of calcination (thermal treatment in oxidizing atmosphere) is to stabilize physical and chemical properties of the catalyst or its precursor. During calcination, thermally unstable compounds (carbonates, hydroxides, or organic compounds) decompose and are usually converted to oxides. During calcination new compounds may be formed, especially if the thermal treatment is carried out at higher temperatures [207]. For example, in the thermal decomposition of Cu or Ni nitrate deposited on alumina, not only CuO or NiO but also Cu or Ni aluminate is additionally formed [207].

Furthermore amorphous material can become crystalline. Various crystalline modifications can undergo reversible or irreversible changes.

Physical and mechanical properties and pore structures can also change. The calcination temperature is usually slightly higher than that of the catalyst operating temperature [250], [253], [256], [263].

For the calcination of powder or granulate, rotary kilns are preferably used [253], [256]. Smaller batches of powdered catalysts are calcined in box or muffle furnaces with trays, as in the case of drying. The gases that are mainly used for heating are in direct contact with the material being calcined [250], [253], [256].

Pellets or extrudates are calcined in belt or tunnel furnaces. The tunnel-type calciner can operate at substantially higher temperatures (close to 1270 – 1470 K) than the belt type (870 – 1070 K). The tunnel calciner is especially suitable for firing ceramic carriers. The material being calcined is taken in boxes or carts which are recycled to the entrance via a continuous chain or belt [256].

Reduction, Activation, and Protection. Reduction, activation, or passivation, is in several cases the last step in catalyst production. These operations are performed by the catalyst producer or in the plant of the client.

For example, in the production of Ni catalysts for the liquid-phase hydrogenation of fats and oils, the reduction of NiO deposited on kieselguhr is carried out exclusively by the catalyst producer. The reduction of powders (50 – 500 μm particle size) is performed on an industrial scale in fluid-bed reactors. The reduced material is pyrophoric and must be protected with a hard fat such as tallow, to make its handling easy and safe. The finished catalyst is supplied in the form of flakes or pastilles [250], [256], [264].

The reduction of metal oxides such as NiO, CuO, CoO, or Fe_2O_3 is carried out with H_2 at elevated temperature (> 470 K) and has two steps. In the first step metal nuclei are formed. In the second, nuclei accumulate to form metal crystallites. The rates of both processes depend on temperature and on the nature of the substrate [250]. Reduction at lower temperatures (< 570 K) provides a narrow distribution of small metal crystallites. Reduction at higher temperatures (> 670 K) gives a broader distribution and larger metal crystallites [250].

Reduction of some oxides, such as those of Cu and Fe, is exothermic and needs to be carried out carefully with H_2 diluted with N_2.

Water, the reduction product, has negative effects on the rate and on the extent of reduction [250]. In industrial practice, where H_2 is recycled, the removal of water by freezing out (below 270 K) and by adsorption on molecular sieve is essential.

To achieve optimal activity, partial reduction of oxidic catalysts is common [250], [265]. For

example, Ni catalysts for fat and oil hydrogenation contain about 50 – 60 wt % of metallic Ni, 45 – 35 wt % NiO, and about 5 wt % Ni silicate.

When the reduction of shaped oxidic catalysts is conducted by the catalyst producer, then the active material is protected either with a high-boiling liquid such as higher aliphatic alcohols or $C_{14} - C_{18}$ paraffins [265] or it is passivated. In this procedure, chemisorbed hydrogen is removed in a gas stream composed of N_2 and about 0.1 – 1.0 vol % of O_2 at ambient temperature. After this treatment catalysts can be handled in air without any precautions [264]. The activity is restored in the client's plant by treatment with H_2 [250].

The activation of hydrotreating catalysts composed of Ni- or Co-promoted $MoO_3 - Al_2O_3$ is carried out with H_2 containing 10 vol % of H_2S [207]. In the past, this activation was performed exclusively in the hydrodesulfurization plants. However, presulfiding at catalyst producers is becoming more common [250].

Mainly electrical or gas-heated shaft reactors are used for the reduction of extrudates, spheres, or pellets in plants of catalyst producers.

Catalyst Forming. The size and shape of catalyst particles depend on the nature of the reaction and on the type of applied reactor.

Reactions in the *liquid phase* require small particles or even powders (50 – 200 µm). Such materials are made by grinding of a dried or calcined precursor, e.g., filter cake, using granulators equipped with sieves to give uniform particle size [250], [256], [263], [265].

Catalysts for *fluidized bed reactors* (0.05 – 0.25 mm) are usually made by spray drying or by cooling molten material droplets (V_2O_5) in an air stream [250], [256], [263].

Spheres consiting of Al_2O_3, SiO_2, or aluminosilicate with 3 – 9 mm diameter are used preferentially as a support for catalysts in *moving-bed* or *ebullating-bed reactors*. They are produced by the so-called oil-drop method (see Section 4.1). Spheres prepared in this way possess sufficient abrasion resistance [250], [256], [263].

Another method for producing spheres is based on agglomeration of powder by moistening on a rotating disk (spherudizer) [250], [256], [263]. As the spheres reach the desired diameter they are removed automatically and transported to the dryer and calciner. Such spheres are suitable for *fixed bed reactors*.

Other methods for forming spherical particles include tumbling short, freshly extruded cylinders in a rotating drum [256].

In the briquetting technique, the plastic mixture of catalyst powder with a binder is fed between two rotating rolls provided with hollowed-out hemispheres [256].

Extrusion of pastes containing catalyst powder, binders and lubricants is a frequently used industrial shaping method [250], [256], [263]. Depending on the properties of the paste, press or screw extruders are applied. Press extruders are principally suitable for viscous pastes. Screw extruders are preferred for thixotropic masses. In both cases, pastes are forced through a die, and the extruded material is cut with a special device to a desired length and falls onto a moving belt that transports it through a drier or calciner [250], [256], [263]. Poly(vinyl alcohol), powdered stearine, and Al stearate are used as lubricants. If the mass being extruded contains alumina, then peptizing agents such as nitric acid are added mainly to improve the mechanical strength [250], [255], [256].

Another type of binder is calcium aluminate cement, which sets up by treatment with steam [250], [256].

The extruded material can have different shapes, such as cylinders (noodles), hollow cylinders (macaroni), or ribbed cylinders. The sizes depend on the shape and are in the range of 1.5 – 15 mm diameter [256], [265].

Added organic lubricants and pore-forming agents can be removed by calcination in a stream of air.

Special extrusion techniques and equipment are necessary to produce honeycombs.

Extruding is less expensive than pelletizing, but extrudates have less resistance to abrasion than pellets. Extrudates are suitable for different types of *fixed-bed reactors* operating in the gas or trickle phase.

Pelletizing is a very common method for catalyst forming. It is based on compression of a certain volume of powder in a die between two moving punchers, one of which also serves to eject the formed pellet [250], [256], [263]. Depending on the size and the shape of the prepared pellets, the material being pelletized must be crushed and forced through a correspond-

ing sieve [256]. Furthermore, lubricants such as graphite, Al stearate, pol(vinyl alcohol), kaolin, and bentonite are added before the material enters the tabletting machine. The fluidity of the material is required to assure homogeneous filling of the die [250], [256], [263]. As in the case of extrudates, organic lubricants can be removed by calcination of the pellets.

Industrial pelleting machines are equipped with around thirty dies and produce about 10 liter of pellets per hour or more, depending on their shape and size. Pressures in the range 10 – 100 MPa in the pelleting machine are common [250], [256], [263].

Commercially, cylindrical pellets with sizes such as 3×3, 4.5×4.5, 5×5, or 6×6 mm are offered [250], [256], [263]. Production of 3×3 mm pellets is more expensive than that of larger sizes. Besides cylindrical pellets, various companies provide rings, cog wheels, spoked wheels, multi-hole pellets, etc. [264], [265].

Pellets of different shapes and sizes are suitable for various types of fixed-bed reactors.

5. Characterization of Solid Catalysts

Catalytic activity and selectivity critically depend on the morphology and texture, surface chemical composition, phase composition, and structure of solid catalysts. Therefore, many physical and chemical methods are used in catalysis research to characterize solid catalysts and to search for correlations between structure and performance of catalysts. These methods include classical procedures [266] as well as techniques developed more recently for the study of the chemistry and physics of surfaces [266].

5.1. Physical Properties

5.1.1. Surface Area and Porosity [267], [268]

The specific surface area of a catalyst or support (in m^2/g) is determined by measuring the volume of gas, usually N_2, needed to provide a monomolecular layer according to the Brunauer – Emmett – Teller (BET) method.

In this approach, the determination of the monolayer capacity is based on the physisorption of the test gas. The volume adsorbed at a given equilibrium pressure can be measured by static methods, namely, volumetric or gravimetric measurements. Flow or dynamic techniques are also applied.

The *total surface area* of a porous material is given by the sum of the *internal* and *external* surface areas. Pores are classified as *micropores* (pore width < 2 nm), *mesopores* (pore width 2 – 50 nm), and *macropores* (pore width > 50 nm) according to IUPAC definitions [269].

The *specific pore volume*, *pore widths*, and *pore-size distributions* for micro- and mesopores are determined by gas adsorption. For mesopores, the method is based on the dependence of the pressure of capillary condensation on the radius of a pore in which condensation takes place, which is given by the Kelvin equation:

$$\ln (p/p_o) = \frac{V}{RT} \left(\frac{2\sigma \cos\theta}{r_\kappa} \right) \tag{13}$$

where p/p_o = pressure/saturation pressure, V is the molar volume, σ the surface tension of the liquid adsorbate, Θ the contact angle between adsorbate and adsorption layer on pore walls (hence, $\Theta = 0$ is a good approximation), and r_K is Kelvin radius of a pore assuming cylindrical shape. Since an adsorption layer is typically formed before capillary condensation occurs, the geometric radius r_p of a pore is given by the sum of the Kelvin radius r_K and the thickness of the adsorption layer t: $r_p = r_K + t$. *Mesopore size distributions* can be calculated when adsorption and desorption isotherms are available in the full pressure range up to $p/p_o = 0.95$. The mesopore volume V_p is assumed to be completely filled at this relative pressure, which corresponds to $r_p \approx 20$ nm.

In the *micropore* range (pore width < 2 nm), where the pore dimensions are comparable to molecular dimensions, *pore filling* occurs rather than condensation [268]. The Dubinin – Radushkevich and the Dubinin – Stoeckli theories then permit the estimation of pore dimensions from physisorption data. In addition, several empirical methods exist, such as the t-method [270] and the α_s-method [271]. In the original t-method the amount of nitrogen adsorbed at 77 K was plotted against t, the corresponding multilayer thickness

calculated from an universal N_2 isotherm, while in the α_s-method the multilayer thickness t is replaced by the reduced adsorption α_s. Here, α_s is defined as the dimensionless adsorption n^a/n_x^a such that $\alpha_s = 1$ at $p/p_o = 0.4$, and n^a is the adsorbed amount in moles of the adsorbate (e.g., N_2) at a given relative pressure and n_x^a is the amount adsorbed (in moles) at a relative pressure of 0.4.

For meso- and macroporous materials (pore width > 2 nm), the pore size distribution is determined by measuring the volume of mercury (or another nonwetting liquid) forced into the pores under pressure [250]. The measurement, carried out with a *mercury pressure porosimeter*, depends on the following relation:

$$P = \frac{2\pi\sigma\cos\alpha}{r_p} \tag{14}$$

where P is pressure, σ is surface tension of mercury, and α is the contact angle of mercury with solid. At pressures of 0.1 – 200 MPa, pore size distributions in the range of 3.75 – 7500 nm can be measured. Because the actual shape of the pores is not exactly cylindrical as assumed in the derivation of the above equation, the calculated pore sizes and distributions can deviate appreciably from the actual values shown by electron microscopy.

For pore systems with narrow pore size distributions, the *average pore radius* can be approximated by using

$$r_p = 2\, V_p/S_p \tag{15}$$

where V_p is pore volume, and S_p is the surface area. The pore volume of a catalyst or support is given by

$$V_p = 1/\varrho_p - 1/\varrho \tag{16}$$

where ϱ_p and ϱ are the *particle* and *true densities*, respectively. The former is determined by a pycnometer using a nonpenetrating liquid, such as mercury, whereas the true density is obtained by measuring the volume of the solid part of a weighed sample by helium displacement.

In certain instances, pore dimensions can be determined by high-resolution electron microscopy (HREM) [272].

5.1.2. Particle Size and Dispersion [273]

The surface area of active metals dispersed on a support deserves particular consideration since the metal surface area and particle size (which are interrelated quantities) determine the catalytic properties of supported metal catalysts. The metal *dispersion D* is given by $D = N_S/N_T$, where N_S is the number of metal atoms exposed at the surface and N_T is the total number of metal atoms in a given amount of catalyst. The fraction of surface atoms D can be determined if N_S is experimentally available. It can be determined by chemisorption measurements with adsorptives that strongly bind to the metal but which interact negligibly with the support at the chosen temperatures and pressures. H_2, CO, NO, and N_2O have been used for this purpose at or above room temperature [273], and static, dynamic, and desorption methods have been applied. Saturation values of the chemisorbed amounts permit N_S to be calculated if the chemisorption stoichiometries are known.

Dispersion is directly related to particle size and particle size distribution. Assuming reasonable model shapes for the metal particles, average particle sizes can be calculated from the chemisorption data.

Average crystallite size distributions can be determined independently from X-ray diffraction line broadening [273–275], and small-angle X-ray scattering (SAXS) permits the determination of particle sizes and particle size distributions, but also of the specific surface area of the metal and of the support [273–275].

Electron microscopy offers the unique opportunity to observe catalyst morphologies over the entire range of relevant particle sizes [272], [273], [276–278]. Particle shapes and sizes of the support or active phase and their size distributions can be extracted from micrographs, but structural information can be also obtained by electron-diffraction and lattice-imaging techniques [272].

5.1.3. Structure and Morphology

X-ray powder diffraction (see also → Structure Analysis by Diffraction, Chap. 2.6.) (XRD) is a routine technique for the identification of phases present in a catalyst [274], [279]. It is based on the comparison of the observed set

of reflections of the catalyst sample with those of pure reference phases, or with a database (Powder Diffraction File (PDF) distributed by ICDD, the International Centre for Diffraction Data). XRD studies can now be carried out in situ on the working catalyst [279], and the use of synchrotron radiation permits dynamic experiments in real time [280]. Time-resolved studies on a timescale of seconds are now becoming possible. Quantification of phase compositions can also be performed.

More sophisticated analysis of the diffraction patterns of crystalline materials provides detailed information on their atomic structure. The Rietveld method is used for structure refinements. Perhaps more importantly for catalytic materials, the local atomic arrangement of amorphous catalysts is based on the Debye equation, which gives the intensity scattered by a collection of randomly distributed atoms. The Fourier transform of the Debye equation gives the radial distribution function (RDF) of electrons, from which the number of atoms (electrons) located in the volume between two spheres of radius r and $r + \mathrm{d}r$ around a central atom, i.e., the radial density of atoms, can be obtained [279]. This approach has been applied for the structural analysis of amorphous or poorly crystalline catalyst materials and of small metal particles.

X-ray Absorption Spectroscopy (XAS) [281–283], is the method of choice where the applicability of XRD for structure analyses ceases to be possible. Because of their high photon flux, synchrotron facilities are the preferred sources for XAS experiments. The physical principle of XAS is the ejection of a photoelectron from a core level of an atom by absorption of an X-ray photon. The position of the absorption edge gives the binding energy of the electron in the particular core level and is thus characteristic of the respective element and its chemical state (see also → Surface and Thin-Film Analysis, Chap. 2.1.) and the shape of the absorption edge provides information on the distribution of the local density of states (LDOS).

The ejected photoelectron wave is backscattered at neighboring atoms, and the scattered wave interferes with the outgoing primary wave. This interference results in a modulation of the absorption coefficient at energies between 50 and 1000 eV beyond the absorption edge (extended X-ray absorption fine structure, EXFAS). Analysis of these oscillations provides information on the chemical nature of atoms at well-defined distances from the central (ionized) atom and gives coordination numbers. Qualitative information on coordination of the central atom may also be obtained from the observation of pre-edge peaks. Information on dynamic and static disorder can also be extracted from the EXAFS. Hence, a detailed microscopic picture of the structure of a catalyst can be derived. XAS is particularly attractive for studies of catalysts under working conditions, although there are limitations regarding temperature [281]. The combined application of XAS and XRD on the same sample using synchrotron radiation for in situ studies is an ideal tool in catalysis research [284].

Electron Microscopy and Diffraction [272], [276–278], (see also → Microscopy, Chap. 2.). When electrons penetrate through matter in an electron microscope, contrast is formed by differential absorption (amplitude contrast) or by diffraction phenomena (phase contrast). Electron micrographs of catalyst materials can provide for identification of phases, images of surfaces and their morphologies, and elemental compositions and distributions. Image interpretations are often not straightforward and need expert analysis. Several variants of electron microscopy use different electron optics and working principles and therefore have to be chosen according to the problem to be solved.

Conventional transmission electron microscopy (CTEM) operates in the 100 – 200 kV range of electron energies, and imaging is based on amplitude contrast in the bright-field mode. Point resolutions of 0.2 – 0.3 nm can be achieved in favorable cases. A typical application of CTEM in catalysis research is the examination of metal particle sizes and their distributions in supported catalysts.

Dark-field images are produced when the directly transmitted electron beam is excluded by the objective aperture, and only diffracted electrons are used for imaging. This mode of operation selectively detects crystallites with crystallographic spacings within a narrow range.

High-resolution electron microscopy (HREM) can be performed in CTEM instruments by modifying the mode of imaging, or

in dedicated instruments operating at electron energies of 0.5 – 1.0 MeV. HREM images can be directly related to the atomic structure of the material [285]. Lattice fringes can be resolved, and the determination of the spacings of atomic planes is enabled. Support particles can thus be identified, and the crystal structure of heavy metal particles having sizes in the range down to 1 nm can be investigated.

In dedicated *scanning transmission electron microscopes* (STEM) an annular detector provides the image formed from diffracted beams, while the central transmitted beam can be further analyzed by using an electron spectrometer to simultaneously provide elemental analysis. The intensity distribution of electrons scattered at high angles (40 – 150 mrad) depends on the square of the atomic number Z according to the Rutherford scattering cross section. The STEM images are therefore also called Z-contrast images, and they are particularly useful for the study of catalysts containing small metal particles [272].

In *scanning electron microscopy* (SEM) the image is produced by scanning a finely focused probe beam in a raster pattern across the specimen surface. Emitted signals such as backscattered and secondary electrons are detected and used for image formation. Secondary electrons are most commonly used. The best resolutions that can be achieved with current generation SEM instruments are approximately 1 nm. SEM is most useful for studying sample topographies, and it can be applied with a significant background pressure of a reactive gas while the sample is observed (environmental SEM or ESEM).

Selected area electron diffraction (SAED) provides information on phase compositions and structures at a microscopic level. The combination of microdiffraction patterns and bright-field images enables the determination of shapes and exposed facets in dispersed phases in solid catalysts.

Analytical electron microscopy (AEM) permits the determination of the elemental composition of a solid catalyst at the microscopic level by energy dispersive detection of the electron-induced X-ray emission. Energy dispersive spectroscopy (EDS) is sensitive for elements with atomic numbers $Z > 11$. For lighter elements ($Z < 11$), electron energy loss spectroscopy EELS is applied.

Controlled atmosphere electron microscopy (CAEM) [272], [286] is arousing considerable interest as it will permit the observation of changes in the catalyst structure and morphology under reaction conditions.

Vibrational Spectroscopy [287] (→ Infrared and Raman Spectroscopy). Vibrational spectroscopy is one of the most promising and most widely used methods for catalyst characterization, since it provides detailed structural information on the solid catalyst material and on surface groups and adsorbates. Several vibrational spectroscopic methods can be applied in situ, and they can be successfully used for studies on ill-defined high surface area porous materials. In situations where X-ray diffraction techniques are not applicable, vibrational spectroscopy can often provide information on phase transitions and changes in compositions of bulk catalyst materials, on their crystallinity, and on the nature of surface functional groups. Most vibrational spectroscopic methods are not surface-sensitive, but they become surface-sensitive when vibrational spectra are recorded for groups or adsorbates that are present exclusively at the material's surface. Representative examples for the structural characterization of solid catalysts by vibrational spectroscopy are bulk oxides (including simple binary oxides, multicomponent materials such as oxidation catalysts, and zeolites and molecular sieves), and supported oxides (e.g., monolayer-type catalysts), and sulfides. The vibrational analysis of surface groups, particularly of hydroxyl groups, can also be addressed. In many cases surface hydroxyl groups (e.g., on oxides) are simply formed by dissociative chemisorption of water molecules, which reduces the surface free energy. Hydroxyl groups can also be constituents of the solid-state structure, for example as in zeolites.

There are several methods and techniques of vibrational spectroscopy which are particularly suitable in catalysis research. *Infrared transmission – absorption spectroscopy* is the most commonly used technique. The KBr disk technique is routine for transmission spectroscopy of powder samples. However, for in situ investigations pressed self-supporting wafers have to be used. Samples which exhibit only weak bulk absorption, and the av-

erage particle size d of which is smaller than the wavelength of the infrared radiation ($d < \lambda$) are optimally suited for the transmission mode. Transmission – absorption infrared spectroscopy has been particularly successful in elucidating the structure of hydroxyl groups [119], [287]. More strongly absorbing materials, and particularly those having average particle sizes greater than the wavelength of the infrared radiation, which therefore cause significant scattering losses in transmission, may preferentially be studied by *diffuse reflectance spectroscopy* (DRS, DRIFT). A powerful technique for structural studies on catalytic materials under extreme temperature and pressure conditions is *laser Raman spectroscopy* (LRS), although laser heating and laser-induced fluorescence may cause serious problems. One way, among others [287], to avoid fluorescence is to use of UV light instead of visible radiation for spectral excitation [288]. LRS has been successfully applied for the structural characterization of complex oxides, zeolites, and supported oxides and sulfides [287]. *Surface-enhanced Raman spectroscopy* (SERS) has found some application in studies of finely divided metal catalysts, particularly silver [289]. *Second harmonic generation* (SHG) and *sum frequency generation* (SFG) [290], [291] are nonlinear optical techniques with high surface sensitivity which will probably find increasing application in studies relevant to catalysis.

Neutron techniques [292] include *neutron diffraction* and *inelastic neutron scattering* (INS). Both techniques are particularly sensitive to light elements (such as H or D) and provide complimentary structural information to XRD.

5.1.4. Local Environment of Elements

Nuclear spectroscopic methods provide information on the local environment of several selected elements.

Mössbauer spectroscopy and time differential perturbed angular correlation (TDPAC) belong to the class of techniques which detect solid-state properties mediated by hyperfine interactions via nuclear spectroscopy [293]. Both techniques are γ spectroscopies; they are bulk techniques and can be applied under in situ conditions, although Mössbauer spectroscopy requires low temperatures.

Mössbauer spectroscopy (Mössbauer Spectroscopy) [293], [294] provides information on oxidation states, phases, lattice symmetry, and lattice vibrations. Its application is limited to elements which exhibit the Mössbauer effect, such as iron, cobalt, tin, iridium, ruthenium, antimony, and platinum. Particularly valuable information on catalyst structures has been obtained for iron catalysts for Fischer-Tropsch and ammonia synthesis, and for cobalt-molybdenum hydrodesulfurization catalysts.

The time differential observation of the perturbed angular correlation of γ rays emitted from radioactive nuclei (TDPAC) [293], [295], [296] is a γ-spectroscopic technique which also allows the determination of hyperfine interactions such as nuclear electric quadrupole interactions (NQI). The NQI parameters enable local structural information around the γ emitter to be extracted. The technique has been successfully applied in studies on molybdenum-containing catalysts, and its application to tungsten seems promising.

Solid State Nuclear Magnetic Resonance [297–299] (see also → Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy, Chap. 4.). NMR spectroscopy in heterogeneous catalysis principally allows the characterization of the chemical and structural environment of atoms in the catalysts (or in species adsorbed on catalyst surfaces). NMR studies on catalysts can be carried out over a wide range of temperatures and pressures, as well as in the presence of gases and liquids. Information can therefore be derived about the structures of catalysts and their thermal or chemical transformations. In addition, specific adsorbent – adsorbate interactions, the nature of chemically bonded surface species, and chemical reactions occurring at the catalyst surface can be studied. Most elements of interest in catalysis have isotopes that can be studied with modern NMR spectrometers. Isotope enrichments may be desirable or even necessary for certain elements, for example, ^{17}O.

NMR spectra of solids are often complex since structure-dependent interactions such as dipolar interactions, chemical shift interactions,

quadrupolar interactions (for nuclei with spin $I > 1/2$) contribute strongly to the shape and position of NMR lines. Because of their structure-dependence these interactions are the main source of information on the structural environment of the nucleus in question. The selective determination of the related interaction parameters of structurally unequivalent nuclei is the major goal of an NMR experiment. In well-crystallized samples, the interaction parameters adopt unique values, while in poorly crystallized or amorphous powders they must be described by distribution functions.

The anisotropy of the above-mentioned interactions results in line broadening, and the spectra of polycrystalline samples consist of a broad superposition of signals arising from different orientations of the crystallites relative to the direction of the external magnetic field B_o, weighted by the statistical probability with which each orientation occurs (powder patterns). Special techniques have been developed which remove or at least reduce substantially these line-broadening effects and permit highly resolved NMR spectra of powders with individual lines for inequivalent nuclei to be recorded. The most important of these techniques are *dipolar decoupling*, *magic-angle spinning* (MAS), and *double oriented rotation* (DOR). *Cross-polarization* (CP) improves the sensitivity for nuclei with low natural abundance and allows the spatial proximity of nuclei to be monitored.

Typical examples for structural characterizations by solid-state NMR [297] are studies on zeolites using ^{27}Al and ^{29}Si NMR. Information on the distribution of Al in the environment of Si atoms and on the possible presence of nonframework Al species has been obtained. The location of exchangeable alkali metal ions has been studied by ^{23}Na and ^{133}Cs NMR. Vanadium- and molybdenum-based catalysts have successfully been characterized by ^{51}V and ^{95}Mo NMR.

5.2. Chemical Properties

5.2.1. Surface Chemical Composition

The atomic composition of a catalyst surface plays a decisive role for the catalytic properties. Electron and ion spectroscopies [300] are surface-sensitive analytical tools which provide information on the atomic composition within the topmost atomic layers. The information depth, i.e., the number of atomic layers contributing to the measured signal, depends on the method. Concentration profiles can be obtained by sputter etching of the surface by ion bombardment. The application of these particle spectroscopies requires ultrahigh-vacuum (UHV) conditions.

The basis for the identification of atoms on surfaces of solid materials by electron spectroscopies, such as *Auger electron spectroscopy* (AES) and *X-ray photoelectron spectroscopy* (XPS) are the electronic binding energies. With ion spectroscopies, such as *low-energy ion scattering* (LEIS) and *Rutherford backscattering*(RDS), surface atoms are identified by their nuclear masses. Ion bombardment of surfaces is accompanied by sputtering processes (surface etching) which lead to the removal of secondary ionic and neutral particles. These are analysed by mass spectroscopic techniques, such as *secondary ion mass spectroscopy* (SIMS), and *secondary neutral mass spectroscopy* (SNMS). Less frequently used is *laser microprobe mass analysis* (LAMMA). Relevant information on the properties of the various surface analytical techniques is summarised in Table 10.

The physical principles of the various techniques have been discussed in several articles and monographs [300], [301].

Electron Spectroscopy (AES, XPS) (see also → Surface and Thin-film Analysis, Chap. 2.3.). These techniques use electrons as information carriers. The electrons can be produced by the absorption of photons resulting in photoemission. In XPS, X-ray photons are used to ionize core levels, and the kinetic energy E_k of the emitted photoelectrons is measured. The energy balance is given by:

$$E_k = h\nu - E_b - \Phi \quad (17)$$

This equation permits the electron binding energy E_b (relative to the Fermi level) to be measured when the photon energy $h\nu$ and the work function Φ of the spectrometer are known. The binding energies are characteristic for a particular element.

As a result of the photoionization a singly ionized atom is formed, which can also be produced

Table 10. Characteristics of surface analytical techniques in standard applications (adapted from ref. [300])

Information	Technique					
	AES	XPS	LEIS	RBS	SIMS	SNMS
Surface sensitivity (monolayers)	2–5	5–10	1–2	20–50	2–4	2–4
Detection limits (monolayers)	$10^{-2}-10^{-3}$	10^{-2}	10^{-3}	10^{-3}	10^{-6}	10^{-6}
Quantification ∗	+	++	+	+++	–	+
Chemical information ∗	(+)	+	–	–	+	–
Structural information ∗	–	(+)	(+)	+	(+)	–

∗ Increasing number of positive signs indicates better capabilities; parentheses indication of restriction to special conditions.

by electron impact. The core hole (e.g., in the K shell) can be filled by an electron from a higher shell (e.g., the L_1 shell) and the energy of this de-excitation process can be released by emission of an X-ray photon (*X-ray fluorescence*, XRF) or can be transferred to another electron (e.g., in the L_2 shell) which is then emitted with a well-defined kinetic energy (Auger process). This kinetic energy is determined by the orbital energies E_K, E_{L_1}, and E_{L_2} of the three orbitals involved. The Auger energy E_{KLL} is then given by:

$$E_{KLL} = E_K - E_{L_1} - E_{L_3} - \delta E - \Phi \quad (18)$$

where δE is a relaxation energy, and Φ the spectrometer work function. Clearly, E_{KLL} is characteristic for an element and independent of the initial ionization process. Thus, both techniques permit the elemental constituents of a surface to be identified.

The information depth of both electron spectroscopies is determined by the mean free path of the emitted electrons, which depends on the kinetic energy of the electron in the solid matrix. This dependence is known [300–302]. The electron mean free path is typically larger in oxides than in metals at equal energy, and it is particularly large for zeolites because of their low density. Together with reported ionization cross sections and, in the case of AES, Auger decay probabilities, quantitative surface analysis is possible. The ratios of integral peak areas are proportional to concentration ratios. These can be analyzed as a function of preparation and treatment conditions of a given catalyst system (e.g., supported metal, oxide, or sulfide catalysts) and compared with model calculations [301]. Information on the elemental distributions and on dispersions of active components thus becomes available.

Ion-scattering Spectroscopies [300] (see also → Surface and Thin-Film Analysis, Chap. 3.3.). In ion-scattering spectroscopies solid surfaces are bombarded with monoenergetic ions, which are scattered on the top atomic layer (ion energies of about 0.5 – 5 keV, *low-energy ion scattering* (LEIS) [303], [304]) or within near-surface regions (ion energies of about 0.1 – 23 MeV, *Rutherford backscattering* (RBS) [305], [306]). In both cases the collision kinematics can be described as simple binary collisions, so that the kinetic energy of the backscattered ion is directly dependent on the ratio of the masses of the projectile and the scattering target atom and on the scattering angle. The mass of the projectile is known and the scattering angle is fixed and determined by the geometry of the spectrometer. Thus, the mass, and hence the identity, of the scattering target atoms can be determined unequivocally.

The LEIS technique provides information on the nature of the atomic constituents of the topmost atomic layers. Quantitative analysis, however, is difficult since neutralization probability, which makes the technique surface sensitive, is not easily available. Only a few percent of the primary ions are backscattered as ions in the case of noble gas ion (e.g., He^+). The technique can be applied for the characterization of real catalyst surfaces, although surface roughness reduces the signal intensity.

In contrast, in the energy regime of RBS the scattering cross sections can be calculated exactly. As a consequence, quantitative analysis is possible by RBS, but the surface sensitivity is lower than for LEIS. In optimal cases an information depth of 1 – 5 nm can be achieved. A combined application of LEIS, RBS, and perhaps XPS is often most informative [300].

Secondary Particles. Ion bombardment of a surface leads to ion etching with the release of atoms and molecular fragments with varying charges (anions, cations, and neutrals) and excitation states. The mass analysis of secondary ions by mass spectrometry [*secondary ion mass spectroscopy* (SIMS)] has been developed as a highly sensitive and powerful surface analytical method (→ Surface and Thin-Film Analysis, Chap. 3.1.) [307], [308]. Although destructive because of the need for sputtering, the sputtering rate can be kept low in the so-called static mode (low primary-ion current density) so that the surface remains essentially unchanged. Since the sputtered particles are preferentially released from the first two atomic layers, the SIMS technique is surface-sensitive. In contrast to the ion-scattering techniques, not only atomic constituents of a surface can be detected but information on the local environment of an atom in the surface can be obtained by analysis of molecular fragments. The detection of light elements, particularly hydrogen, is also possible. Quantification of the method is difficult, although not entirely impossible.

A high percentage of the sputtered secondary particles are neutral and must be postionized for mass spectroscopic analysis [*secondary neutral mass spectroscopy* (SNMS), → Surface and Thin-Film Analysis, Chap. 3.2.] [309], [310]. Post-ionization can be achieved by electron impact in a plasma or by an electron beam. Alternatively, resonant and nonresonant laser ionization can be applied. Applications of SNMS for catalyst characterization have still not been reported.

5.2.2. Valence States and Redox Properties

Electron Spectroscopies. *X-ray photoelectron spectroscopy* (XPS) and *Auger electron spectroscopy* (AES), in addition to elemental analysis, permit information to be obtained on the valence and bonding states of a given element. This is due to the fact that the core-level binding energies and Auger kinetic energies are dependent on the chemical state, which leads to characteristic chemical shifts. In solids, the Madelung potential also plays an important role [301], [311], [312]. In addition, the ionization of an atom leads to relaxation phenomena which provide a relaxation energy that is carried on by the emitted photoelectron. The binding energy of a core electron in level C of an atom is given by:

$$E_{\mathrm{b}}\,(\mathrm{C}) = \mathrm{const.} + \sum_j (q_j/R_j) + kq - R_{\mathrm{ea}} \qquad (19)$$

where q_j are the charges of all other atoms and R_j their distances from the core-ionized atom. The constant is the Hartree-Fock energy of the core electron in the atomic level C for the free atom, k is the change of the core potential resulting from the removal of an electron, and R_{ea} represents the extra-atomic polarization energy.

For an Auger transition involving the atomic levels C, C′, and C″, the kinetic energy E_{k}(C, C′, C″) of the Auger electron is related to the photoelectron binding energy E_{b}(C) in good approximation by:

$$\alpha' = E_{\mathrm{b}}\,(\mathrm{C}) + E_{\mathrm{k}}\,(\mathrm{C,C',C''}) = \mathrm{const.} + R_{\mathrm{ea}} \qquad (20)$$

where α' is the so-called Auger parameter.

Chemical shifts and the Auger parameter provide detailed information on the chemical state of an element as regards its oxidation state and its local environment. The latter is reflected in the Auger parameter, which is dependent on the extra-atomic polarization energy and hence, on the structural and bonding characteristics of the atom under consideration. So-called Wagner plots [312], [313] in which the XP binding energy, the Auger energy, and the Auger parameter are correlated for families of compounds often permit the analysis of a compound of unknown structural and bonding characteristics.

Optical Spectroscopy and Electron Paramagnetic Resonance. Optical excitations in the UV, VIS, and NIR regions and electron paramagnetic resonance (EPR) are classical techniques which provide information on the electron configuration (oxidation state) of a metal center and on the symmetry of the ligand sphere [314–317]. While optical spectroscopy is applicable to practically all systems, EPR is limited to paramagnetic species, i.e. those which contain one or more unpaired electrons.

UV–VIS–NIR spectroscopy covers a wide range of energies (typically 0.5 – 6 eV or 4000 to 50 000 cm^{-1} or wavelengths (2500 to 200 nm) as shown in Figure 16. Several types of transitions occur in this range, namely, chargetransfer

(CT) and $d-d$ transitions, as also indicated in Figure 16. The first class of excitations involves two adjacent atoms, one of which is typically a metal center and the other a ligand or another metal atom. Electromagnetic radiation can promote charge transfer from the ligand (L) to the metal (M), from the metal (M) to the ligand (L) or from one metal center to another. These transitions are therefore called ligand-to-metal CT (LMCT), metal-to-ligand CT (MLCT), and metal-to-metal CT (MMCT), respectively. Such transitions occur in molecular complexes and in nonmolecular solids, such as metal oxides. The energy of CT transitions depends on the symmetry and oxidation state of the metal center and on the nature of the ligand or of the second metal atom [318]. Hence, information on these properties can be extracted from CT spectra. These spectra are relatively intense since they are dipole-allowed. In contrast, metal-centered or intra-atomic transitions in transition metal atoms or ions (ligand-field or $d-d$ transitions) are of moderate or weak intensity because they are forbidden by the Laporte (orbital) selection rule (and also by the spin selection rule) unless the selection rules are relaxed by vibronic or spin–orbit coupling. The $d-d$ transitions also provide information on the electron configuration and on the symmetry of a complex (or local environment in a solid). Typical systems that have been studied by optical spectroscopy are transition metal and base metal oxides, transition metal ions or complexes grafted on support surfaces, and transition metal ion-exchanged zeolites.

While spectra of liquid samples can be recorded in the transmission mode, catalyst powders must be studied by the diffuse reflectance technique (DRS) [315], [316], [319]. The measured diffuse reflectance can be converted into the so-called Schuster–Kubelka–Munk (SKM) function, which is directly proportional to the absorption coefficient. The wavelength dependence of the SKM function is thus equivalent to an absorption spectrum, provided the scattering coefficient is independent of the wavelength. This condition is often fulfilled in the UV and VIS spectral regions. Simple quartz cells for in situ treatments can be designed, to which an EPR tube can also be connected for simultaneous optical spectroscopy and EPR on the same sample [314].

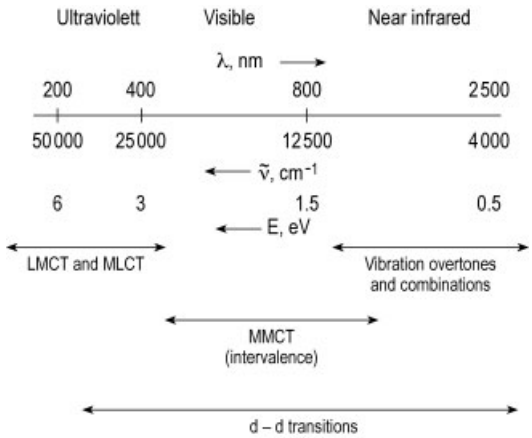

Figure 16. Energy ranges of different types of electronic transitions (adopted from [314])

Luminescence spectroscopy [320], [321] has proved to be a valuable addition to the spectroscopy techniques for characterization of solid catalysts under well-defined conditions.

As mentioned above, *electron paramagnetic resonance* (EPR) is used to study paramagnetic species in catalytic materials. Besides the simple qualitative (and quantitative) detection of the presence of paramagnetic sites, the spin density distribution at the paramagnetic center and on the neighboring atoms can be deduced from the spectra. Simulations of EPR spectra are often useful for full interpretation. The extremely high sensitivity of the EPR technique can be an advantage but also a drawback because the importance of minority radical species may be overemphasized. EPR is not surface-sensitive. However, radical species in the surface can easily be identified by exposing the sample to paramagnetic O_2, which leads to significant broadening or disappearance of signals of surface species because of dipole–dipole interactions.

Several reviews on the use of EPR in catalyst characterisation have been published [314], [322], [323]. Typical applications of EPR are the detection of paramagnetic states of transition metal ions and analysis of the symmetry of their ligand sphere and/or their coordination, redox properties of catalytic materials and their surfaces, and surface anion or cation radicals deliberately produced by organic molecules as probes for the redox properties of the solid catalyst. Radical species (e.g., in connection with

coke formation) formed during catalytic reactions have also been detected.

Thermal Analysis. Thermoanalytical techniques such as *differential thermal analysis* (DTA), *thermogravimetry* (TG), and *differential scanning calorimetry* (DSC) are well-established methods (→ Thermal Analysis and Calorimetry) in solid-state chemistry [324], [325] which have successfully been applied to investigating the genesis of solid catalytic materials. They can also be used to follow reduction and oxidation processes by measuring either thermal effects and/or weight changes. When combined with an on-line mass spectrometer, changes in the gas-phase composition occurring during chemical transformations of the solid sample can be monitored simultaneously.

In temperature-programmed reduction (TPR), as first described by Robertson et al. [326], a stream of inert gas (N_2 or Ar) containing ca. 5 vol % H_2 is passed through the catalyst bed of a flow reactor containing a reducible solid catalyst [327]. By monitoring continuously the H_2 concentration in the gas stream and its eventual consumption with a thermal conductivity detector while heating the sample with a linear temperature ramp of ca. 10 K/min, the rates of reduction are obtained as a function of time (or temperature). The total amount of H_2 consumed determines the reduction equivalents present in the catalyst, and detailed analysis of the experiment permits the kinetic parameters of the reduction process to be determined and provides information on reduction mechanisms. Characteristic numbers which depend on the experimental parameters (amount of reducible species present, H_2 concentration, flow rate, and temperature ramp) have been defined [328], [329]. These numbers must be kept in certain ranges for optimal performance of the experiment.

TPR experiments have been used to investigate the reduction behavior of bulk and supported reducible species, solid solutions, promoted metal catalysts, metals in zeolites, and of supported sulfides and of nitrides [327].

Temperature-programmed oxidation (TPO) is an equally valuable technique for investigating the oxidation kinetics and mechanisms of reduced materials [327]. Cyclic application of TPR and TPO provides information on the redox behavior of catalytic materials, e.g., of catalysts for selective catalytic oxidations.

5.2.3. Acidity and Basicity

Acid-catalyzed reactions are among the industrially most important hydrocarbon conversions. Acid sites can be classified as Lewis acidic sites, such as coordinatively unsaturated cations (e.g., Al^{3+} on the surface of partially dehydroxylated alumina, and Brønsted acidic sites, which are typically surface OH groups as, e.g., in H forms of zeolites. Carbenium and carbonium ions are thought to be formed by protonation of hydrocarbons on these groups. Surface oxygen ions may function as Lewis basic centers, and if strong enough they may abstract protons from hydrocarbon molecules to form carbanion intermediates. A typical solid base is MgO.

For characterisation of acid and base properties, the nature (Lewis or Brønsted) of the sites, the acid or base strength, and the number of sites per unit surface area of a solid catalyst must be determined. Brønsted acidity is almost certainly required for all acid-catalysed reactions. However, the mechanistic details on surfaces are significantly different from the well-known carbenium and carbonium ion chemistry in solution, because of the lack of the stabilizing effect of solvation in heterogeneously catalysed gas-phase reactions. As shown by Kazansky [330], the electronic ground state of surface acidic OH groups of oxides and in H forms of zeolites is essentially covalent. The main differences in their acid strength are thought to be due to the energetic positions of their electronically excited heterolytic terms. Similarly, the interaction of acid groups with alkenes does not result in the formation of adsorbed carbenium ions but rather in the formation of more stable covalent alkoxides. Basic sites (surface O^{2-} ions) in the vicinity of OH groups could be involved in this process. Carbenium ions (and even more so carbonium ions) are therefore not considered to be reaction intermediates in solid acid catalysis, but rather excited unstable ion pairs or transition states resulting from electronic excitation of covalent surface alkoxy species. Because of the proposed bifunctional nature of active acid sites in heterogeneous acid catalysis, it is necessary to characterise both the acidic and basic properties of solid catalysts.

Many different methods have been developed for the characterisation of acidity, but only little is known about the basic character, particularly of materials that are typically considered to be acid catalysts.

Chemical Characterisation [331]. Titration methods in aqueous medium are not very informative, because H_2O tends to strongly modify surface properties by molecular or dissociative chemisorption. Therefore, nonaqueous methods have been proposed, in which the solvent (e.g., benzene or isooctane) does not or only weakly interact with the catalyst surface. Hammett indicators were used to determine the acid strength in terms of the Hammett – Deyrup H_0 function:

$$H_0 = -\log a_{H^+}(f_B/f_{BH^+}) \quad (21)$$

where a_{H^+} is the proton activity and f_B and f_{BH^+} are the activity coefficients of the basic probe and its protonated form, respectively. A series of Hammett indicators covers the range of $-18 < H_0 < 4$, where $H_0 = -12$ corresponds to 100 % H_2SO_4.

Site densities and acid strength distributions were determined by the *n*-butylamine titration method [332]. As these titration and indicator methods can yield erroneous results [333], they are not frequently used today.

Isosteric heats of adsorption of strong bases (e.g., pyridine) may be considered as measures of acid strength. However, a discrimination between Lewis and Brønsted sites is not possible.

Temperature-programmed desorption [also called *thermal desorption spectroscopy* (TDS)] of basic probe molecules has been developed as a powerful tool for the characterisation of solid acids [334], [335]. In this method, a strong base is isothermally pre-adsorbed on an acidic catalyst and then exposed to a stream of inert gas (e.g., He). Heating by a temperature ramp (ca. 10 K/min) leads to desorption of the base. The integral area of the desorption peak gives the total acid site density, and the position of the peak maximum provides the activation energy of desorption (which may be close or identical to the heat of adsorption), which can be considered to be a measure of the acid strength. This approach has been applied for investigations of H forms of zeolites using ammonia as the probe [336], [337]. However, discrimination between Lewis and Brønsted acid sites is again only possible with the assistance of, e.g., vibrational spectroscopy.

Microcalorimetry [334]. Differential heats of adsorption of probe molecules can be measured with high accuracy by heat-flow calorimetry and differential scanning calorimetry. These data provide information on the acid (or base) strength distribution. Ammonia and other amines have been used as probes for acid sites on oxides [338] and in H forms of zeolites [339], [340], and carbon dioxide and sulfur dioxide were adsorbed as acidic probes on several oxides [338].

Vibrational spectroscopy [341–346]. *Transmission infrared spectroscopy* is the most frequently applied technique for investigations into acidic and basic properties of solid catalysts. Surface hydroxyl groups can easily be detected since they function as dipolar oscillators. However, the stretching frequency of unperturbed OH groups can not be taken as a measure of the acid strength. Lewis acidic and basic centers can only be detected by vibrational frequencies with the adsorption of suitable probe molecules. Criteria for the selection of optimal probe molecules have been defined by Knözinger et al. [341], [346].

The use of basic probe molecules permits a discrimination between Brønsted and Lewis acid sites. When a base B is adsorbed on an acidic OH group, hydrogen bonding followed eventually by protonation of the base may occur (Eq. 20):

$$OH + B \rightleftharpoons OH \cdots B \rightleftharpoons O^- \cdots H^+B \quad (22)$$

The strength of the hydrogen bond and the ability of the OH group to protonate the base is determined by the acid strength of the surface OH group and by the base strength (or proton affinity) of B. When hydrogen bonding occurs, the induced frequency shift of the O – H stretching mode $\Delta\tilde{\nu}_{OH}$ is a measure of the strength of the hydrogen bond ΔH_B (Eq. 23) [347] and hence, of the acidity of the OH group.

$$|\Delta\tilde{\nu}_{OH}|^{1/2} \sim \Delta H_B \quad (23)$$

Simultaneously, internal molecular modes of the base B are modified, particularly when protonation occurs. These changes can also be used

for the interpretation of the bonding type of the probe. Strong bases such as the traditional probe molecules ammonia and pyridine are protonated by even very weak Brønsted sites which may not be at all relevant in acid catalysis. Weaker bases such as nitriles, carbon monoxide, and even dinitrogen and dihydrogen only undergo hydrogen bonding (hydrogen bonding method [342]), but due to their weak interactions they are very specific and can provide very detailed information on the properties of acidic surfaces.

The same bases can be used for the detection of Lewis acid sites L, with which they form surface coordination compounds:

$$L + B \rightleftharpoons L \leftarrow B \tag{24}$$

The frequency shifts of the internal B modes are a measure of the nature and strength of the coordinative bond. For example, carbon monoxide when coordinated to L sites undergoes very typical shifts of the C – O stretching frequency which provide information on the nature of the element and of the coordinative bond, on the oxidation state, and on the coordination of the L site [341], [346].

The investigation of basic sites O^{2-} by acidic probe molecules AH with analysis of the vibrational spectra is much less advanced than that of acid sites. The hydrogen-bond method can in principle be applied:

$$O^{2-} + H\text{–}A \rightleftharpoons O^{2-} \cdots H\text{–}A \tag{25}$$

Here the shift of the H–A stretching mode is a measure of the hydrogen-bond strength and hence, the basic character (proton affinity) of the surface O^{2-} site. Recently, CH compounds such as trichloromethane [342], acetylene and substituted acetylenes [346], [348] and even methane [350] were proposed and successfully tested as acidic probe molecules. Pyrrole [345] and several Lewis acids [341] have also been used.

Surface chemical transformations of, e.g., CO_2, alcohols, ketones, acetonitrile, and pyridine gave detailed information on the bifunctional acid – base pair character of several oxides, particularly of alumina [341], [351].

Nuclear Magnetic Resonance [352–354]. Solid-state ^1H magic-angle spinning (MAS) NMR spectroscopy measures proton chemical shifts, which were thought to reflect the deprotonation energies of surface OH groups. However, proton chemical shifts are also very sensitive to hydrogen bonding. Therefore, changes in proton chemical shifts induced by hydrogen bonding of probe molecules can also be used for the characterization of protonic acidity.

Dissociative chemisorption of CH_3I was proposed for characterization of surface basicity by ^{13}C NMR spectroscopy [356].

5.3. Mechanical Properties [357]

Catalyst particles are exposed to diverse mechanical strains during transportation, charging to the reactor, and operation. In fixed bed reactors catalyst particles must withstand pressure caused by the mass of the catalyst charge and erosion by high-velocity gas streams. In fluidized- and moving-bed reactors, the particles must resist attrition from rubbing against each other and from colliding with the walls of the reactor system. The technical performance of catalysts depends on their mechanical strength to maintain integrity for a reasonable time in spite of these strains.

There are three types of methods for determining the strength of catalysts used under static conditions [358]. For pellets and rings with no areas of distortion (preferably sized 1 cm or larger), the *crushed* (or crush) *strength* is determined by exerting pressure on the specimen placed between two horizontal plates of a hydraulic press. The upper plate moves down until the specimen is crushed, at which point the pressure is recorded. The test is repeated for several particles, and the values are averaged. In the knife edge hardness test, the upper plate of the press is replaced by a knife with a 0.3 mm edge. A mass of 1 kg is applied to the knife and the percentage of broken samples is recorded. The mass is then raised in increments of 1 kg, and the test is repeated until 100 % of the particles are broken or until a mass of 10 kg is reached. Catalyst particles of irregular form are tested in a cylinder provided with a ram. After a definite pressure is applied, the sample is discharged, and the weight percent of fines formed during the test is determined by screening.

Tests for impact strength and resistance against abrasion or attrition are carried out un-

der dynamic conditions. For *impact testing* of very strong catalysts (e.g., ammonia synthesis catalysts), a mass of 500 – 1000 g is dropped on the particle from a standard height and the percentage of unbroken, split, and broken samples out of 20 or more is recorded. *Abrasion tests* on tableted and extruded catalysts are carried out in a rotating horizontal steel cylinder provided with one baffle. The percent of fines (based on the mass of the catalyst tested) formed after 1 h is reported as attrition loss. The *attrition loss* of fluid catalysts is measured by exposing the catalyst particles to a high-velocity air stream in a glass pipe. The fines formed during the test (prevented from escaping by a filter) are reported as attrition loss expressed as the percentage of the sample charged [359].

6. Reaction Networks and Mechanisms of Selected Catalytic Reactions on Solid Catalysts

As discussed in Section 2.3, a catalytic mechanism can be described by a catalytic cycle taking all elementary reaction steps and surface intermediates into account. For a complete description of a mechanism, exact knowledge of the reaction dynamics and energy-transfer processes is required. Investigations into these details are emerging only recently [28]. The maximum knowledge that can presently be obtained (with a few exceptions) is a quantitative analysis of experimentally determined kinetic parameters such as preexponential factors and activation energies for forward and backward reactions of elementary steps. A kinetic model can thus be established and reaction rates predicted. The necessary information is often obtained from surface science studies on model systems such as metal single crystals.

Frequently, reaction schemes are derived from product and intermediate distributions that can be observed in the fluid phase, often without any information on the nature of surface intermediates. These reaction schemes should be considered as reaction networks rather than mechanisms.

In the following, CO oxidation and NH_3 synthesis are discussed as reactions for which the mechanisms are known in significant detail at the atomic level. The discussion of reaction schemes ("mechanisms" and reaction networks) for several organic reactions follows.

6.1. Carbon Monoxide Oxidation

CO oxidation on noble metals (Pt, Pd, etc.)

$$CO + 1/2\,O_2 \longrightarrow CO_2 \tag{24}$$

occurs in catalytic reactors in automobiles and helps reduce air pollution. The reaction is relatively well understood, based on surface science studies. On Pd, dioxygen is chemisorbed dissociatively, while CO binds associatively [360], [361]. Molecular CO then reacts with atomic oxygen in the adsorbed state:

$$O_2 + 2* \longrightarrow O_{2,ads} \longrightarrow 2\,O_{ads} \tag{25}$$

$$CO + * \longrightarrow CO_{ads} \tag{26}$$

$$CO_{ads} + O_{ads} \longrightarrow CO_2 + 2* \tag{27}$$

Here $*$ denotes a free surface site and the subscript 'ads' indicates an adsorbed species. The reaction steps (25) – (26) suggest that CO oxidation is a Langmuir – Hinshelwood process in which both reacting species are adsorbed on the catalyst surface. The reverse of reaction (25), i.e., the recombination of two oxygen atoms is kinetically insignificant at temperatures below ca. 600 K. Possible Eley – Rideal steps (28) and (29), in which a gas-phase molecule reacts with an adsorbed species were found to be unlikely. Once CO_2 is formed according to step (26) or (29) it desorbs instantaneously.

$$CO + O_{ads} \longrightarrow CO_2 + * \tag{28}$$

$$2\,CO_{ads} + O_2 \longrightarrow 2\,CO_2 + 2* \tag{29}$$

Quantitative experiments led to a schematic one-dimensional potential energy diagram characterizing the elementary steps on the Pd(111) surface (Fig. 17). Most of the energy is liberated upon adsorption of the reactants, and the activation barrier for the combination of the adsorbed intermediates is relatively small; this step is only weakly exothermic, and the heat of adsorption (activation energy for desorption) of CO_2 is very low.

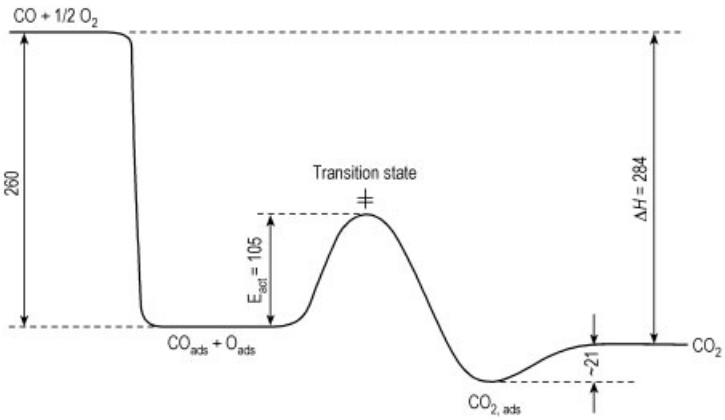

Figure 17. Schematic one-dimensional potential energy diagram characterizing the $CO + O_2$ reaction on Pd(111) [361]

The sequence of elementary steps (25) – (27) is quite simple. The overall kinetics, however, is not. This is due to the nonuniformity of the surface and the segregation of the reactants into surface domains at higher coverages. As a consequence, the reaction between the surface species CO_{ads} and O_{ads} can only occur at the boundaries between these domains. A simple Langmuir-Hinshelwood treatment of the kinetics is therefore ruled out, except for the special case of low surface coverages by CO_{ads} and O_{ads}, when these are randomly distributed and might be considered to a first approximation as being part of an ideal surface.

6.2. Ammonia Synthesis [45], [85], [179]

The ammonia synthesis reaction

$$N_2 + 3\,H_2 \rightleftharpoons 2\,NH_3 \qquad (30)$$

is exothermic and accompanied by a decrease in the number of moles. The reaction toward the target product NH_3 is therefore thermodynamically favored at low temperature and high pressure. The reaction is catalyzed by iron-based catalysts. Model experiments have been extensively carried out to elucidate the reaction mechanism by the surface science approach using iron single crystals.

The adsorption of N_2 on iron is slow and is characterized by a very low sticking coefficient (ca. 10^{-6}) and high activation energy [85]. Single-crystal surfaces of iron are reconstructed upon adsorption of nitrogen. Dinitrogen is dissociated above 630 K [363] and forms complex surface structures. These have been inferred to be surface nitrides with depths of several atomic layers [85]. Their composition is roughly Fe_4N. The corresponding bulk compound is thermodynamically unstable under conditions for which the surface structure is stable. The rate of dissociative adsorption of dinitrogen is structure-sensitive, the Fe(111) face being by far the most active, since the activation energy is the smallest and the rate of adsorption the highest [364]. The same crystal face is also the catalytically most active. These observations are consistent with the earlier suggestion [365] that dinitrogen adsorption is an activated process and that it is the rate-determining step in the catalytic cycle.

In contrast, the adsorption of dihydrogen on iron is fast and characterized by a high sticking coefficient (ca. 10^{-1}) and an extremely small activation barrier. The chemisorption is dissociative yielding covalently bonded H atoms which have high mobility on the iron surface.

Atomic nitrogen was shown to be the most stable and predominant chemisorbed species on Fe(111) after evacuation [360], [361], and it was inferred to be an intermediate in the catalytic reaction. Adsorbed dinitrogen could be excluded as a reactive intermediate. The involvement of adsorbed N atoms in the rate-determining step was also demonstrated by kinetics experiments [366]. Other less stable and less abundant surface intermediates include NH and NH_2 species.

Based on these results the following sequence of elementary steps was formulated (∗ denotes a surface site):

$$N_2 + 2* \longrightarrow 2\,N_{ads} \quad (31)$$

$$H_2 + 2* \longrightarrow 2\,H_{ads} \quad (32)$$

$$N_{ads} + H_{ads} \longrightarrow NH_{ads} + * \quad (33)$$

$$NH_{ads} + H_{ads} \longrightarrow NH_{2ads} + * \quad (34)$$

$$NH_{2ads} + H_{ads} \longrightarrow NH_3 + 2* \quad (35)$$

A schematic potential energy diagram for the catalytic cycle is shown in Figure 18. Decomposition of N_2 is exothermic, whereas the steps involved in the successive hydrogenation yielding NH_x species are endothermic. The addition of the first H atom is the most difficult step.

The promotion of the iron catalyst with potassium lowers the activation energy for dissociative N_2 chemisorption [367].

The sequence of elementary steps (31) – (35) was the basis for the microkinetic analysis discussed in Section 2.2.2, which permitted the prediction of overall reaction rates based on the kinetic parameters characterizing the individual elementary steps.

6.3. Acid – Base catalysis of Hydrocarbon Transformations

(reprinted from D. S. Santilli and B. C. Gates [368])

6.3.1. Introduction

The technological importance of surface-catalyzed hydrocarbon reactions has motivated years of effort to understand their mechanisms. But because of the difficulty of determining surface reaction intermediates, understanding of surface reaction mechanisms lags far behind that of solution reaction mechanisms, and what is known about the former is fragmentary and often largely based on presumed analogies with the latter, combined with results such as those from tracer experiments, kinetics experiments, and theoretical chemistry.

Catalysis by acids and bases generally proceeds via cycles involving hydrogen transfer reactions. In solution, solvent molecules such as water often play a role, e.g., by interacting strongly with reaction intermediates and/or by being donors or acceptors of protons. Reactions catalyzed by soluble acids or bases are also catalyzed by solid acids or bases, but the solid-catalyzed reactions usually take place in the absence of solvents and at much higher temperatures than are practical with solutions. When solid catalysts incorporate functional groups similar to those of soluble catalysts, there are close analogies between the two. For example, sulfonic acid groups in ion-exchange resins (used to catalyze the synthesis of methyl *tert*-butyl ether from isobutylene and methanol) act catalytically much like soluble toluenesulfonic acid.

When proton-donor groups are dissociated and hydrated so that (hydrated) H_3O^+ ions are present, then these are the catalytically active species, and the term *specific acid catalysis* is applied. Such catalysis is virtually the same in solution and near the surface of a hydrated solid acid such as an ion-exchange resin. Catalysis by OH^- is called *specific base catalysis*. On the other hand, when undissociated acid (or base) groups are the catalytic sites, the catalysis is called *general acid (or base) catalysis*. General acid catalysis, whether in solution or on a surface, is usually complex because a variety of unidentified species may be simultaneously present and catalytically active.

General acid catalysis predominates in solid acid catalysis. Because typical solid acids such as metal oxides incorporate both a spectrum of proton-donor groups such as OH groups (Brønsted acid sites) and a spectrum of electron-pair acceptor groups (Lewis acid sites) with various strengths, the catalytic sites often remain unidentified. Thus although there may be simple relationships between catalytic activity and easily measured properties characterizing acidity, they often do not provide much insight into the nature of the catalytic sites or the reaction mechanism. The same holds true for basicity and catalysis by bases. Thus the analogies between solution and surface acid (or base) catalysis are for the most part rather weak, and the mechanistic interpretations summarized here are simplified.

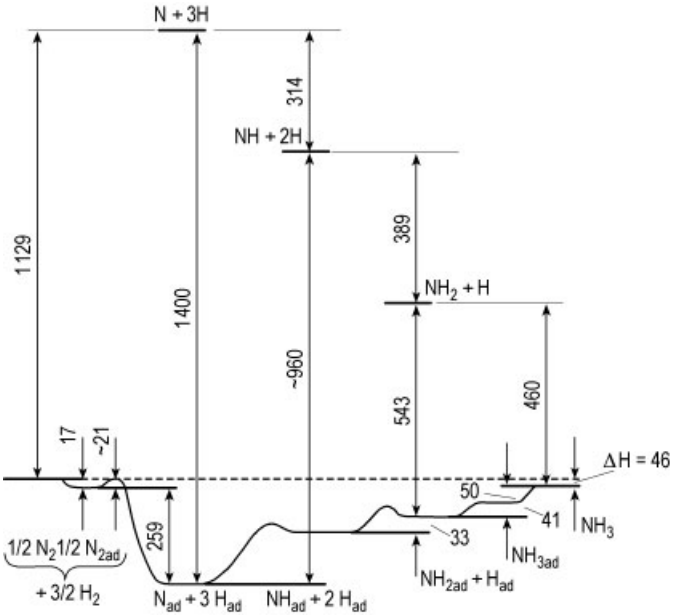

Figure 18. Potential energy diagram for the sequence of elementary steps of the ammonia synthesis reaction (energies in kJ mol^{-1}) [179]

6.3.2. Carbocations and Their Reactions

The key reaction intermediates in acid-catalyzed hydrocarbon reactions are carbocations [109], [370–375], the reactions of which are summarized below. Reactions proceeding by radical intermediates are ignored here, although they are important in hydrocarbon reactions catalyzed by acids at temperatures higher than those of the processes mentioned here.

6.3.2.1. Nature and Formation of Carbocations

Carbenium Ions. Hydrocarbons are weak bases, and the acid catalysts required for their conversion are relatively strong acids. The most important reactant classes include alkenes, aromatic hydrocarbons, and alkanes. Aromatics and alkanes are present in crude petroleum, and alkenes are formed as products of petroleum cracking. Solid acids are important catalysts for many large-scale conversions of these compounds. Alkenes and aromatics in the presence of a strong acid are converted in part into carbenium ions, which are intermediates in a variety of acid-catalyzed reactions, summarized below. In solution, carbenium ions are usually solvated or present in ion pairs, but, for simplicity, the carbenium ion structure alone is often written.

Carbenium ions form from alkenes and from aromatics by proton additions (Eq. 36) and from alkanes via carbonium ion intermediates (Eq. 37).

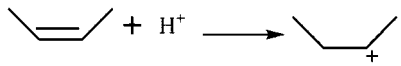

(36)

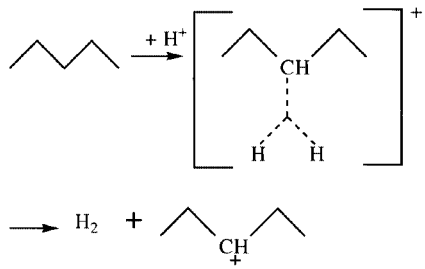

(37)

Much has been learned about the reactions of carbenium ions from their investigation in superacid solutions [373]. Superacids are combinations of Lewis and Brønsted acids, such as $HF + SbF_5$ or $FSO_3H + SbF_5$. In aprotic solvents, they are so strongly acidic that they are

capable of protonating alkenes almost entirely to give carbenium ions; the following reaction (Eq. 38) generating a carbenium ion is driven almost completely to the right.

$$R{-}CH{=}CH{-}R' + H^+ \longrightarrow R{-}\overset{+}{C}H{-}CH_2{-}R' \qquad (38)$$

Superacids have thus allowed investigations of the reactivities of carbenium ions in the near absence of alkenes. Complete conversion of alkenes to carbenium ions greatly simplifies matters because alkenes react with carbenium ions. Investigations with proton and ^{13}C NMR spectroscopy have provided a wealth of quantitative information about carbenium ion reactivity and, consequently, about mechanisms of catalytic conversion of hydrocarbons [109], [369]. This chemistry is a foundation for what follows.

The great proton-donor strengths of superacids are associated with reactions that stabilize protonated forms of the Brønsted acids [373]. For example, when FSO_3H is in the presence of SbF_5, $[FSO_3H_2]^+$ is formed; when HF is used with SbF_5, the superacidic proton donor is H_2F^+. These species can form because of the great stability of the anions $SbF_5(SO_3F)^-$ and SbF_6^-, with which they may be ion-paired.

Tertiary carbenium ions are much more stable than secondary carbenium ions, which are much more stable than primary carbenium ions. Calorimetric measurements have shown that the difference in stability between tertiary and secondary carbenium ions having the same carbon skeleton is about 54 kJ mol^{-1} [376]. The difference in stability between secondary and primary carbenium ions having the same carbon skeleton is greater than about 71 kJ mol^{-1}. Because entropy effects are expected to be negligible, the free energy differences correspond to the enthalpy differences. There are no significant differences in stabilizing effects attributed to different alkyl groups.

Quantum mechanical calculations have been done to help visualize how an alkene molecule interacts with acidic OH groups such as those present on the surfaces of metal oxides or zeolites [377], [378]. Basically, the π-electron cloud of the hydrocarbon interacts with the proton-donor group, and then the proton bound to the catalyst surface forms a bond to one carbon atom in a C=C moiety. The surface O–H bond is broken, and another oxygen atom of the surface forms a bond with the other carbon atom of the C=C moiety. Regardless of how this actually occurs, the result is a covalently bonded surface alkoxy species, which is regarded as an intermediate in numerous catalytic reactions (Scheme 1).

It is presumed that excited states of such alkoxy species are carbenium ion-like in behavior and are the transition states for catalytic reactions for which the alkoxy species are intermediates (Fig. 19) [379], [380]. The reactions of the carbenium ions are discussed below.

An alkane forms a carbenium ion through carbonium ion-like reactions, which are discussed below. This can result from interaction of an alkane with a proton to give a carbenium ion plus H_2 or a smaller carbenium ion plus an alkane. Altematively, a carbenium ion can form as a result of an interaction of an alkane with a carbenium ion:

1) Hydrogen transfer from an alkane to a carbenium ion, giving another alkane and another carbenium ion
2) Disproportionation of a carbenium ion to give a carbenium ion of a different carbon number

These are discussed below.

Carbonium Ions The interaction of an alkane with a superacidic proton donor can be written as in Scheme 2, whereby addition can occur across a C–C or a C–H bond. The product, a carbonium ion, has a three-center, two-electron bond [373]. Such a species must be strongly coordinated to the catalyst surface, and calculations have shown the permissibility of surface transition states such as those shown in Scheme 3 [381–383]. It is believed that concerted bond-breaking and bond-forming steps make the energetics favorable relative to the energetics of stepwise reactions.

6.3.2.2. Reactions of Carbenium Ions

The following sections describe the elementary steps undergone by carbenium ions, whatever their exact nature.

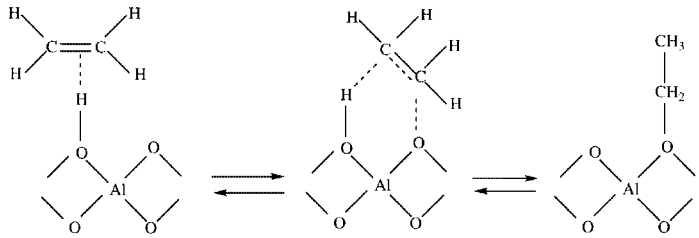

Scheme 1.

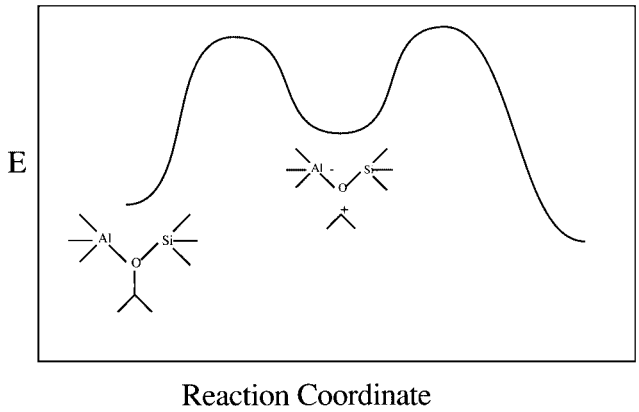

Figure 19. Energy diagram for conversion of alkoxy species into carbenium ion-like species

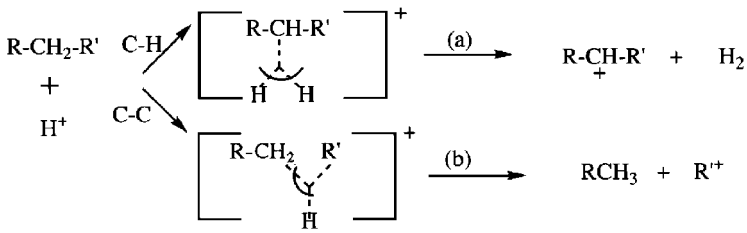

Scheme 2.

Intramolecular Reactions.

Hydride and Alkyl Shifts. The important reactions of carbenium ions include isomerizations proceeding by 1,2-hydride and 1,2-alkyl shifts, symbolized by H~ and CH_3 ~, respectively. The hydride shift is probably the most facile reaction of carbenium ions. Formally, it occurs as shown in Equation (39). Mechanistically, it may involve a three-membered (or larger) ring intermediate and a more or less symmetrical transition state involving concerted bond making and breaking (Scheme 4).

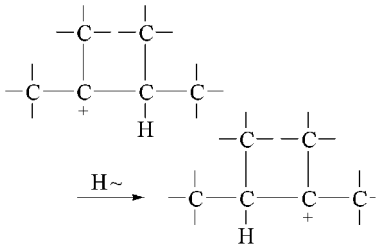

(39)

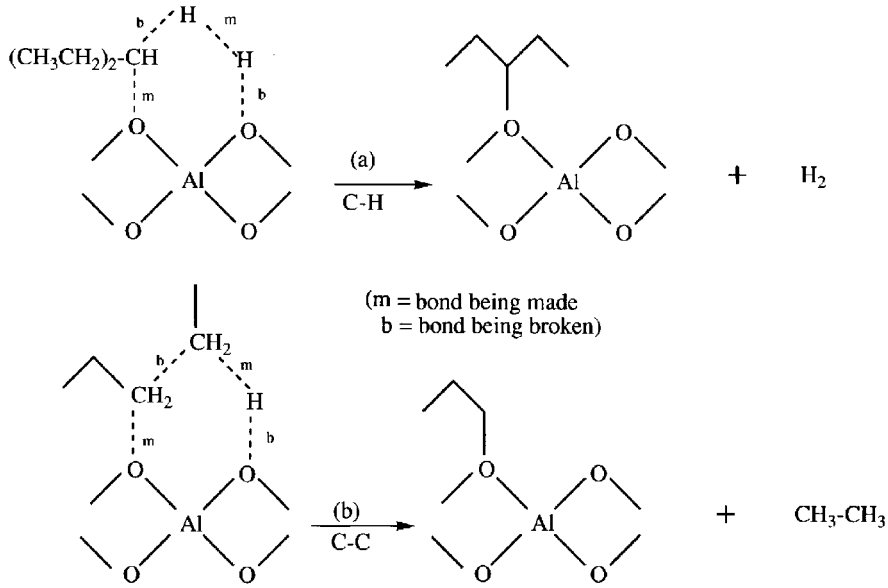

Scheme 3.

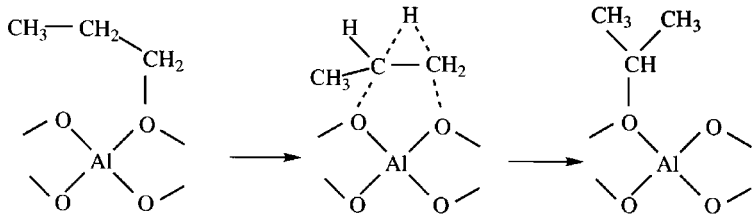

Scheme 4.

Carbenium ions equilibrate rapidly to give a distribution of isomers favoring the more stable highly substituted species.

Alkyl shifts involve the migration of alkyl groups, as shown formally in Equation (40) and Scheme 5.

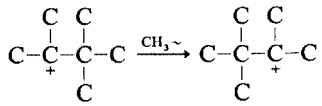

(40)

The distribution of products is affected by the stabilities of the transition states by which they are formed, with tertiary being favored over secondary being favored over primary carbenium ions; vinyl and benzyl carbenium ions are also relatively stable. However, the formation of stable highly substituted species can be sterically inhibited, for example, by the confining pores of zeolites, as discussed below. The rates of alkyl shifts are less than the rates of hydride shifts, and they may be competitive with hydride transfer and β-scission, depending on the nature of the carbenium ions and other interacting species.

The alkyl shift mechanisms probably involve protonated cyclopropyl (or larger-ring) intermediates, as shown in Scheme 6 [384–386].

Combinations of these elementary steps account for reactions such as skeletal isomerization of a tertiary octyl cation (Eq. 41), which proceeds via a hydride shift followed by a methyl shift (Eq. 42). The intermediate is a secondary carbenium ion, which, in solution, is about 54 kJ mol^{-1} less stable than the tertiary reactant ion. Because the activation energy barrier for a

hydride or methyl shift is roughly 5 kJ mol^{-1}, the activation energy for the overall reaction in solution is about 60 kJ mol^{-1}.

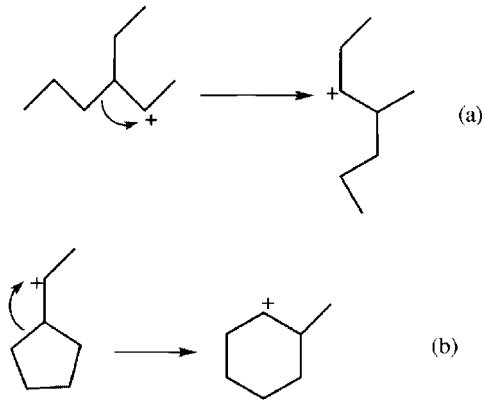

Scheme 5.

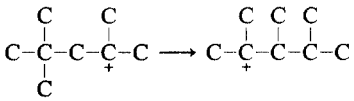

(41)

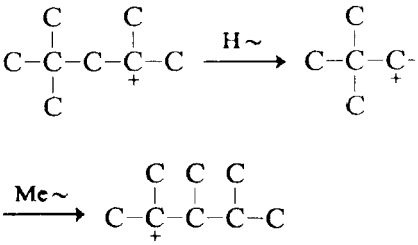

(42)

Carbenium ion rearrangements leading to a change in the degree of branching are not accounted for by simple hydride and alkyl shifts. For example, the isomerization shown in Equation (43), which leads from a carbenium ion with a single branch to one with two branches, does not proceed via a sequence of hydride and methyl shifts, because it would involve a highly unstable primary carbenium ion. The accepted mechanism for a branching rearrangement involves an intermediate incorporating a protonated cyclopropane ring, which is much lower in energy than a primary carbenium ion. Carbenium ion rearrangements leading to changes in the degree of branching are believed generally to proceed via such intermediates [371], [384–386].

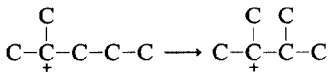

(43)

β-Elimination. When an adsorbed carbenium ion reacts to give back the proton to the surface with desorption of an alkene, the reaction is the reverse of protonation of an alkene by the surface. This reverse reaction is a β-elimination, which may proceed as in Equation (44).

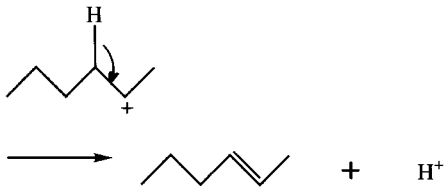

(44)

This reaction is favored at high temperatures (≥ 800 K) and competes with β-scission (described below). At lower temperatures, intermolecular reactions (described below) can intervene.

β-Scission. An important reaction of carbenium ions is β-scission, the breaking of a carbon – carbon bond located β to the charged carbon atom, which occurs in catalytic cracking. The reaction occurs formally as in Scheme 7 (it may occur in concert with a hydride shift to form more stable carbenium ions). Again, a concerted

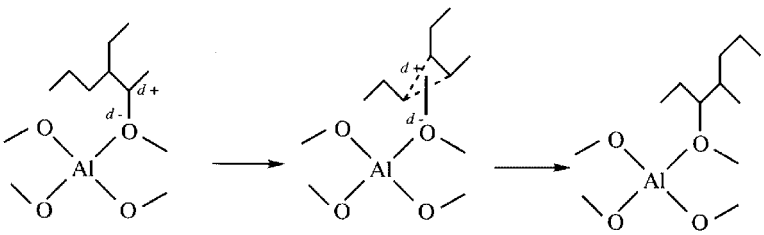

Scheme 6.

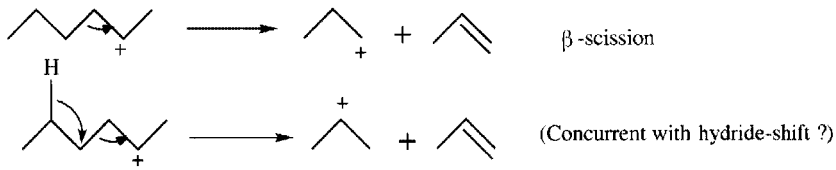

Scheme 7.

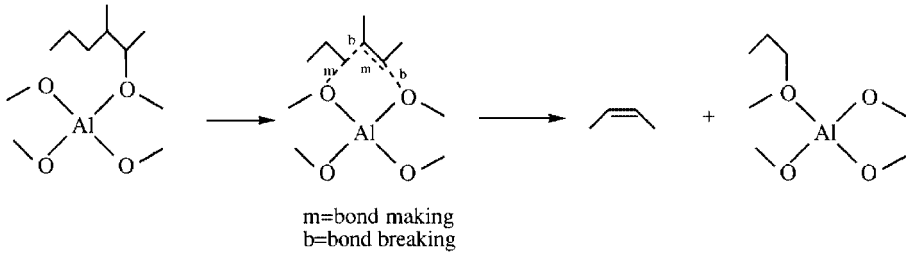

Scheme 8.

reaction can be written for a possible mechanism (Scheme 8).

The reaction generates an alkene and another carbenium ion. The rate depends on the relative stabilities of the reactant and product carbenium ions. When both are tertiary, the reaction is fast, faster than a branching rearrangement but slower than a nonbranching rearrangement [371]. The β-scission step may be as fast as some intermolecular reactions such as hydride transfer (described below), but with branched alkanes in the reactant mixture and with carbenium ions with < 7 carbon atoms, hydride transfer can be much faster.

Cyclization. Formally, the reaction is written as in Equation (45). Products formed by cyclization reactions include carbonaceous deposits (coke).

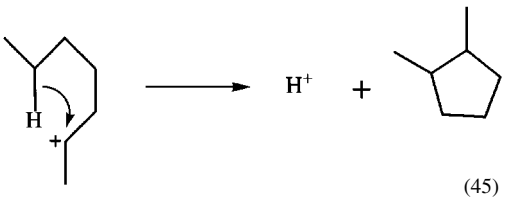

(45)

Intermolecular Reactions. The reactions here involve carbenium ions as Lewis acids reacting with alkanes, alkenes, and aromatics.

Reactions with Alkanes: Hydride Transfer. Carbenium ions are Lewis acids and capable of abstracting hydride ions, shown formally as in Equation (46). The alkane is converted into a carbenium ion and vice versa. The bimolecular reaction is extremely rapid when both reactant and product cations are tertiary and much less rapid when a tertiary cation reacts to give a secondary cation. The mechanism is not known, but it might proceed as shown in Equation (47).

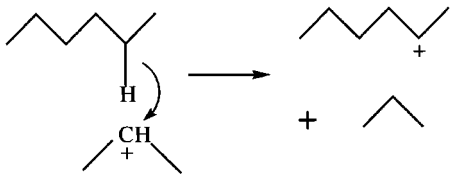

(46)

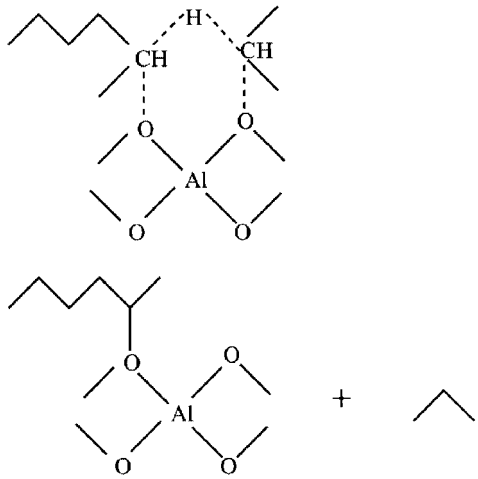

(47)

Alternatively, it has been proposed [387] that this reaction may also proceed through carbonium ion-like intermediates **1**, perhaps similar to those proposed for the alkane-proton reaction. However, free carbonium ions are unlikely to form, and if this reaction proceeds through carbonium ion-like species, Equation (46) is more plausible.

In principle, hydride transfer involving alkenes can also take place much as hydride transfer involving alkanes takes place.

Disproportionation. Formally, this occurs as shown in Equation (47), in which fragments of the carbon skeletons of the interacting alkane and carbenium ion are exchanged. Mechanistically, what might take place is shown in structure (48).

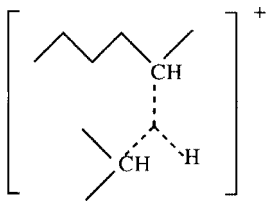

(48)

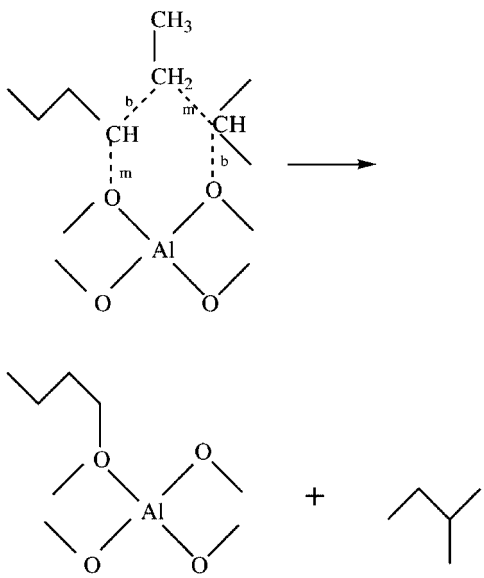

(49)

Reactions with Alkenes: Dimerization. The alkene adds to the carbenium ion to form a larger carbenium ion (Eq. 49). This is the reverse of β-scission and takes place in alkylation reactions. It is thermodynamically favored at low temperatures (alkylation is carried out at subambient temperatures), whereas β-scission is favored at high temperatures (cracking is carried out at temperatures of roughly 800 K).

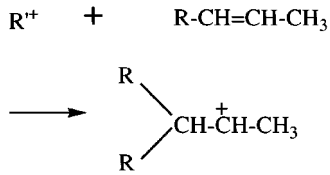

(50)

Reactions with Aromatics. Addition of aromatics to carbenium ions is illustrated by Equations (51) and (52), using the reaction of Equation (50).

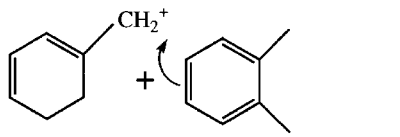

(51)

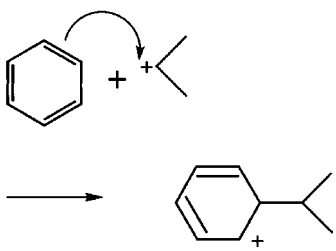

(52)

6.3.3. Catalytic Reactions Involving Carbocation Intermediates (see also [388])

Alkene Oligomerization and Polymerization [389]. After formation of a carbenium ion, repeated addition reactions can occur, forming longer carbenium ions (Eq. 49). These carbenium ions can undergo β-elimination to give

desorbed alkenes as products (Eq. 44), or they can undergo hydride transfer with hydrocarbons to form alkanes (Eq. 45).

Isomerization [390]. Double bond migration in alkenes results from protonation to form a carbenium ion, one or more hydride shifts, and then β-elimination to give the desorbed alkene. As in solution, *cis – trans* isomerization can also occur, as a result of bond rotation in the carbenium ion or alkoxy species formed from the alkene, followed by β-elimination.

Skeletal isomerization (e.g., *n*-pentane $\rightarrow$ isopentane; *p*-xylene $\rightarrow$ *o*-xylene) involves formation of the carbenium ion (Scheme 3 (a) or Eq. 45 for alkanes; Eq. 36 for unsaturates) followed by alkyl shifts (Eq. 40), perhaps concomitant with hydride shifts (Eq. 39) to avoid formation of primary carbenium ions, and finally hydride transfer (Eq. 45), leading to the desorbed alkane. (Primary carbenium ion formation can also be avoided if the reactions proceed via protonated cyclopropyl intermediates.) The process is represented by Scheme 9.

Disproportionation reactions can also lead to isomerization (Eq. 53) [391], [392].

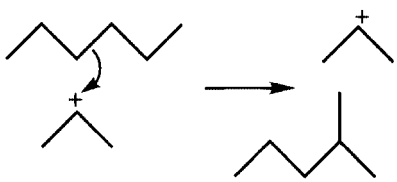

(53)

Alkene isomerization can take place by protonation to give a carbenium ion, followed by alkyl shifts and β-elimination to give the rearranged alkene. Aromatic isomerization can take place by carbenium ion formation (Eq. 36) followed by hydride (Eq. 39) and alkyl (Eq. 40) shifts leading to ring expansion or contraction, as shown in Scheme 10. The reactions lead to changes in the type of alkyl group or the location of the alkyl group on the ring. Product formation results from β-elimination (Eq. 44). Alternatively, aromatics can isomerize intermolecularly via transalkylation, as described below.

Ring expansion or contraction of cycloalkanes takes place as a carbenium ion is formed (Scheme 3 (a) or Eq. 45), followed by alkyl shifts, as shown in Scheme 5 (b). However, the mechanism may be more complex than what is shown here [393]. For example, concerted hydride shifts may occur to avoid formation of primary carbenium ions (Scheme 9). Furthermore, protonated pseudo-cyclopropyl rings can form (Scheme 6).

Cyclization of alkenes takes place by carbenium ion formation followed by carbon-carbon bond formation (Eq. 44). Once an alkane is converted into a carbenium ion (Scheme 3 (a); Eq. 45), the same steps can follow.

Cracking [394]. Mechanisms of many hydrocarbon reactions catalyzed by acidic surfaces are interpreted within the framework of the chemistry stated in the preceding sections. One of the most important of these is catalytic cracking, whereby hydrocarbons are converted into smaller hydrocarbons in the presence of zeolite-containing acid catalysts at temperatures of roughly 800 K. Besides gas oil cracking and residuum cracking, processes in this class include shape-selective dewaxing. Reactants include alkanes, alkylaromatics, and products of cracking, including alkenes.

Alkanes and cycloalkanes crack via a variety of combinations of steps. Alkanes are activated as stated above. They undergo both bimolecular (i.e., carbenium ion or classical) cracking [395] and unimolecular (i.e., carbonium ion or Haag-Dessau) cracking [395–398], with the former dominating at temperatures less than about 700 K. Unimolecular cracking of an alkane is initiated as the alkane is protonated by the catalyst. Some solid acids (superacids) may even be able to protonate alkanes at temperatures much less than those of cracking processes.

If addition of the proton is across a C–C bond, the resultant carbonium ion can crack to give an alkane and a carbenium ion, and the latter can undergo a β-elimination to give an alkene. This mechanism predicts equimolar yields of the alkane and alkene products, which have been observed, for example, for cracking of *n*-butane catalyzed by HZSM-5 at low conversions [397]. If addition of the proton is across a C–H bond, the resultant carbonium ion splits out H_2 and a carbenium ion, which can undergo β-elimination to give an alkene. In cracking of *n*-butane catalyzed by HZSM-5, H_2 and butenes were observed in equimolar yields at low conversions, consistent with this mechanism [399].

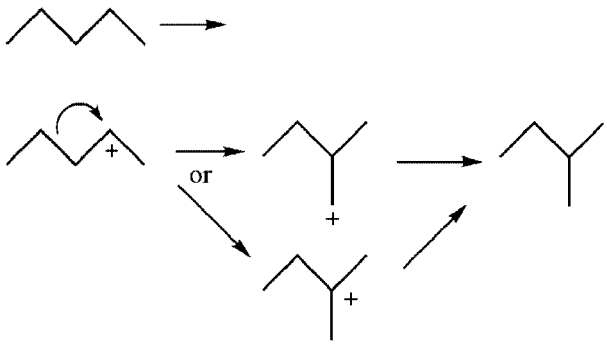

Scheme 9.

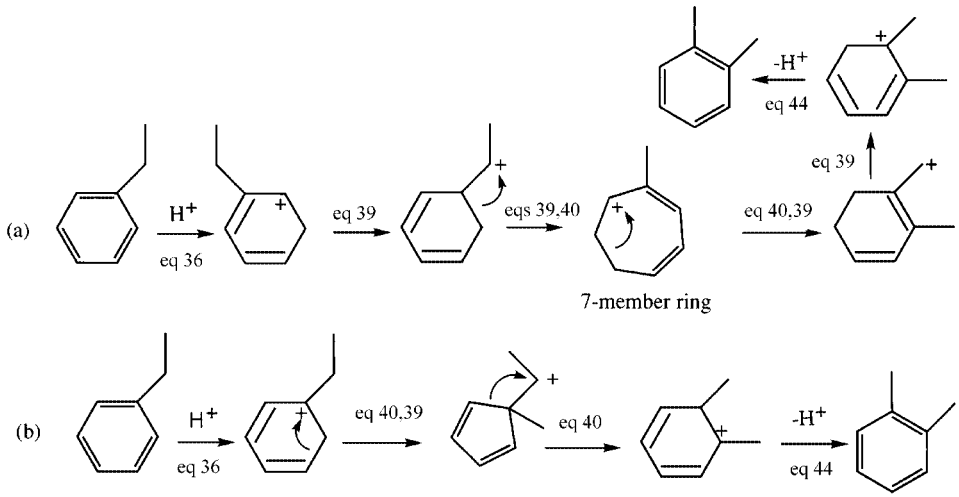

Scheme 10.

Instead of β-elimination, the carbenium ion formed by cracking of the carbonium ion may undergo β-scission in either of these mechanisms (Scheme 7).

The alkenes produced by unimolecular cracking, especially in zeolites, are readily protonated by the catalyst to give carbenium ions. This may be a negligible side reaction at very low conversions, but otherwise it can become a dominant pathway of cracking, whereby carbenium ions undergo bimolecular reactions with alkanes. This leads to hydride transfer and disproportionation reactions, as shown in Equations (46) and (48). Disproportionation can lead directly to cracked products (Eq. 48).

In this bimolecular cracking mechanism, a C_6 molecule interacting with a C_3 carbenium ion can crack to give C_4 and C_5 fragments without giving C_1 and C_2 fragments (Eq. 54). In contrast, in cracking by the unimolecular mechanism, in which a proton is added across a C–C bond, all these products are formed [Scheme 2 (b)].

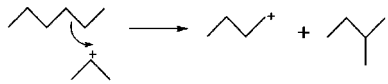

(6 + 3 → 4 + 5) (54)

In bimolecular cracking, the final carbenium ion fragment shown leaving as an alkene in Equation (44) can instead undergo hydride transfer with an alkane (Eq. 46) and leave as an alkane. This

step activates another feed alkane molecule; the carbenium ion is thus a chain carrier, and the process is cyclic (catalytic), as shown in Figure 20. Thus bimolecular cracking becomes an important route at conversions greater than a few percent; it is especially favored at lower temperatures with zeolite catalysts at conversions that are not very low.

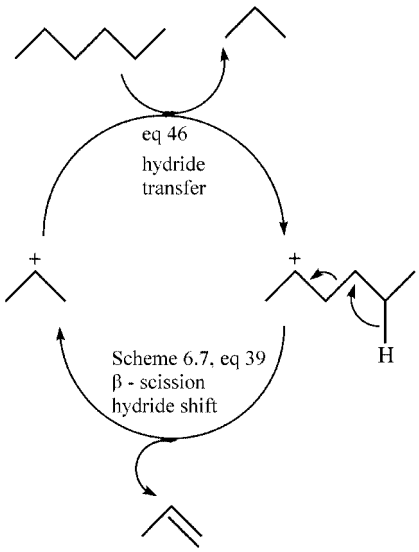

Figure 20. Schematic catalytic cycle for the bimolecular cracking of alkanes

Note that alkane cracking is a very complex process. The alkane interacts with protons and a wide variety of carbenium ions (which form from the alkane). Any of the hydrocarbon bonds can interact with the acidic site. This makes it difficult to unravel any mechanism. Furthermore, many other reactions accompany cracking, including isomerization, disproportionation, and C-C bond forming. A number of recent investigations have led to a fairly complete statement of the complex product distributions and relative rates of the elementary reactions [395–404]. The most incisive cracking experiments have been done with zeolite catalysts having low concentrations of proton-donor groups, such a HZSM-5; catalyst deactivation is minimal with this zeolite.

Alkylation [405], [406]. Aromatic alkylation [407] takes place by addition of a carbenium ion to the aromatic ring, as shown in Equation (52), followed by β-elimination (Eq. 44). Alternatively, one could envisage addition of an alkene to a carbenium ion formed from the aromatic hydrocarbon, but in practice this is unimportant because carbenium ions are formed much more readily from alkenes than from aromatics because aromatics lose the stabilization of resonance energy upon addition of a proton, whereas alkenes do not.

Alkane alkylation, for example, the industrially important formation of octanes (among other products) from the reactions of 1-butene with isobutane occurs via formation of the carbenium ion by protonation of the alkene, activation of the alkane by hydride transfer (Eq. 46), addition of the alkene to the carbenium ion (Eq. 50), hydride transfer from the isobutane to release the dimer and form another carbenium ion, and so forth [407]. The steps are shown in the schematic catalytic cycle of Figure 21.

Practical alkylation is catalyzed by liquid HF or H_2SO_4, and research with solid alkylation catalysts has been motivated by environmental concerns associated with the liquid acids. So far, there is no good solid acid alternative to the liquid acids; the solids undergo rapid deactivation.

Transalkylation of Aromatics. This industrially applied reaction [408] leads to the transfer of alkyl groups between aromatic rings; one application is transalkylation of toluene to give xylenes. The transalkylation of xylene to give toluene and trimethylbenzene is shown in Scheme 11.

The aromatic is activated to give a carbenium ion **2** and, after hydride shifts, this adds to another aromatic. Hydride shifts (Eq. 39) and β-scission (Scheme 7) combine to produce the transfer. However, steric constraints can force this reaction to proceed by an otherwise higher energy route. This involves formation of the carbenium ion **3**, β-scission to release the aromatic (Scheme 7), addition of the carbenium ion R^+ that is thereby formed to another aromatic molecule, and so on (Scheme 12).

These mechanisms can also lead to aromatic isomers, for example, if the toluene product undergoes transalkylation with xylene in Scheme 11.

Aromatization of Alkanes and Alkenes [409]. A carbenium ion can cyclize (Eq. 45)

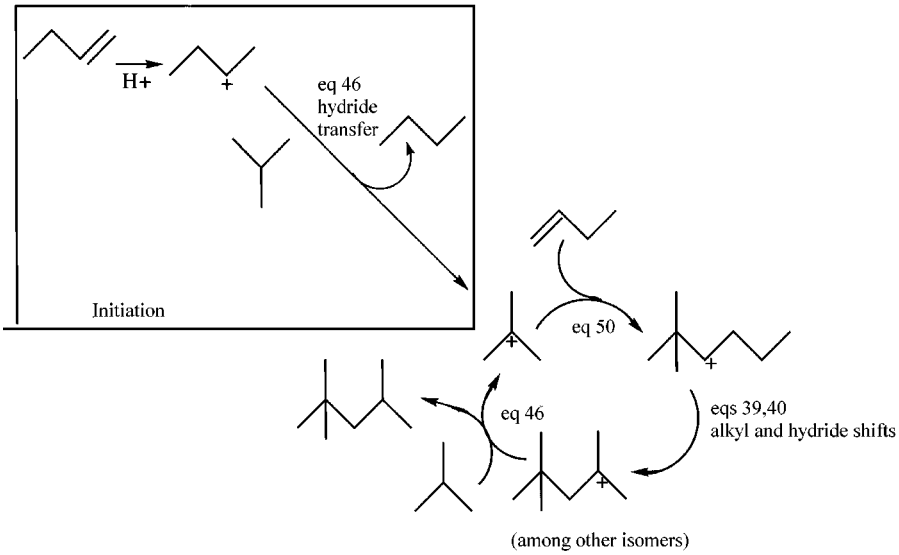

Figure 21. Schematic representation of the mechanism of 1-butene – isobutane alkylation, represented as a catalytic cycle. Numerous side reactions are omitted for simplicity.

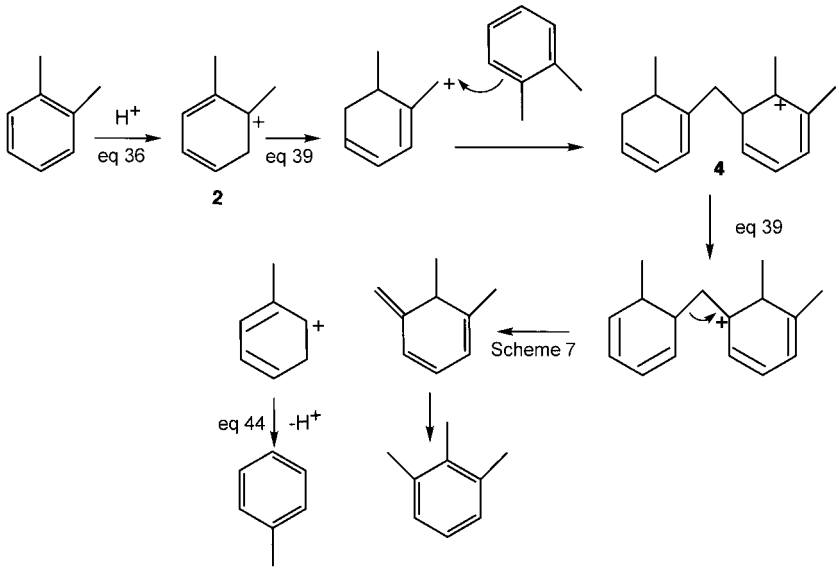

Scheme 11.

followed by β-elimination (Eq. 44) and/or hydride transfer reactions (Eq. 46) to form an aromatic compound, as shown in Scheme 13. The R^+ species come from alkenes adding to proton acidic sites. Thus the net reaction converts alkenes into alkanes and converts naphthenes (cyclic alkanes) into aromatics. This is an important class of reaction that is catalyzed efficiently by zeolites in catalytic cracking processes.

Coke Formation [410]. The term coke refers to high molecular mass hydrocarbon residues that form on surfaces of catalysts and contribute to catalyst activity loss [370], [411]. The mechanism of coke formation is complex

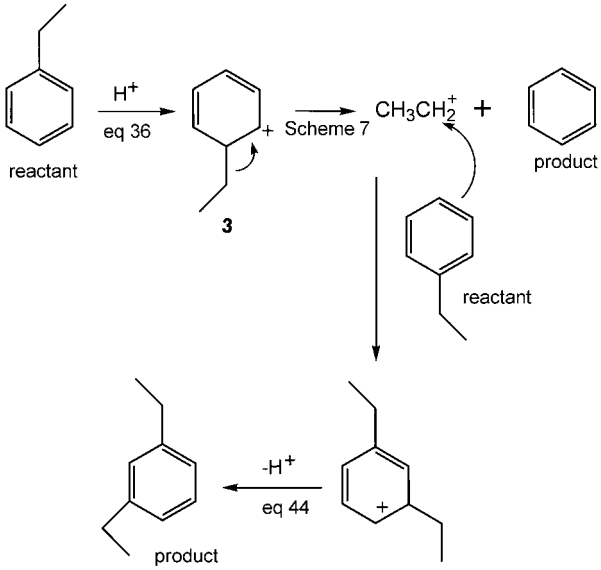

Scheme 12.

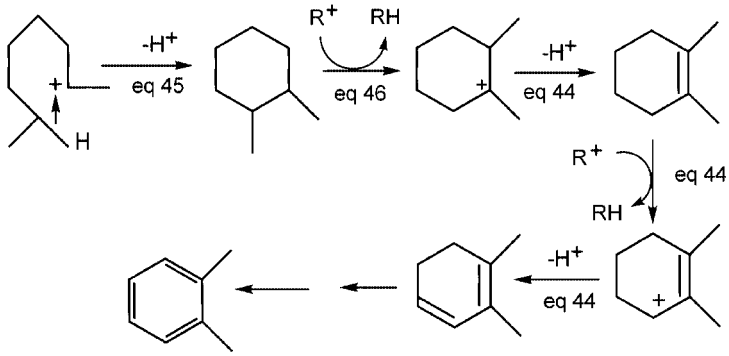

Scheme 13.

and only partially understood, and continuing research has the goal of understanding of the nature of coke and the mechanisms of its formation. Different reactants can form different types of coke on the surface of a given catalyst and, on different catalysts, the same reactants give different cokes.

Coke generation may proceed via carbenium ion formation followed by repeated addition reactions with unsaturates to form large species (Eqs. 50 and 52), which can block acidic sites and fill pores. Cyclization, reactions along with β-elimination and hydride transfer can lead to polymers, polycyclics, and graphitic species.

Steric Influence of the Catalyst. Catalysts such as zeolites, with pore diameters about equal to molecular dimensions, can exert strong steric influences in catalysis; the phenomena are referred to as shape-selective catalysis [412–418]. These phenomena are discussed in Section 2.1.8.

6.4. Metal Catalysis of Hydrocarbon Transformations

(reprinted from D. S. Santilli and B. C. Gates, ref. [368], p. 1132)

A great variety of hydrocarbon reactions proceed on surfaces of metal catalysts, and exten-

sive research has been done with the goal of understanding their mechanisms. Some of the most thoroughly investigated reactions are occurring in naphtha reforming and catalyzed by transition metals such as platinum. Hydrocarbon reactions catalyzed by surfaces of transition metals include alkene and arene hydrogenation, alkene isomerization, homologation, and hydrogenolysis, and alkane isomerization, dehydrogenation, hydrogenolysis, and dehydrocyclization. Recent reviews provide summaries of mechanistic interpretations of these metal-catalyzed hydrocarbon reactions [419–427].

This subject is less than well developed and summarized here in only the briefest form. Many of the published mechanisms are no more than postulated reaction pathways. The subject is so fragmented that systematic statements of what is known are largely lacking. Mechanistic information has been determined from results of tracer experiments and kinetics as well as from the effects of changes in product distribution resulting from changes in reactant structure, catalytic metal, structure of the surface of a given metal, and metal particle size and composition, among others. For the most part, structures of proposed reaction intermediates are no more than hypothetical, with details of bonding between reactant and surface typically being unknown. Theoretical chemistry has not yet had much impact on this subject. Researchers have often drawn on assumed parallels with organometallic chemistry to formulate postulates of mechanisms of metal-surface-catalyzed reactions.

Skeletal Isomerization of Hydrocarbons. The literature of metal-catalyzed skeletal isomerization reactions was summarized by GAULT [419], who used ^{13}C-labeled molecules extensively in attempts to elucidate reaction pathways and infer mechanisms. Numerous such reactions are believed to proceed via cyclic intermediates that are postulated to be metallacycles, although other intermediates are important also. Cyclizations take place, for example, by 1,5-ring closure; subsequent breaking of a C–C bond may give an isomerized acyclic product [427]. When the hydrocarbon structure does not allow formation of a C_5 cyclic intermediate, rearrangements may take place by a so-called bond shift mechanism. Alternatively, 1,6-ring closure may also take place, and when followed by subsequent dehydrogenations, it leads to aromatic products.

Many suggestions have been offered as to the nature of the postulated intermediates and their bonding to catalyst surfaces, but none is well justified. Many of the postulated surface intermediates are dehydrogenated.

Hydrogenation of Alkenes [428]. Hydrogenation of alkenes, especially ethylene, has often been investigated in attempts to understand the reaction mechanism. Because of the simplicity of the reactant molecules, this reaction was long considered to be a good candidate as a reaction offering a relatively simple and easily elucidated mechanism. This early expectation has not been borne out, however, and successive series of experiments have led to recognition of greater and greater degrees of subtlety in the mechanism. We cite here a recent investigation that provides details of the mechanism of ethylene hydrogenation catalyzed by a Pt(111) single crystal [430]. The work demonstrates the advantages of a new surface-science technique, infrared-visible sum frequency generation spectroscopy.

Vibrational spectra were measured with the catalyst in the working state; the method allows determination of spectra of surface species that are not influenced by the composition of the gas phase. The spectra indicate the formation of several surface species formed from ethylene [430]: These are ethylidyne **1**, di-σ-bonded ethylene **2**, and a species inferred to be π-bonded ethylene, presumed to interact with a single surface platinum atom **3**.

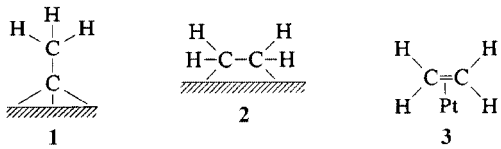

During catalytic reaction at 295 K and pressures of a fraction of 1 bar, the spectra indicated the presence on the Pt surface of all three species. Only the π-bonded ethylene was involved as a catalytic intermediate under these conditions. The other two species were only spectators on the surface (although they could be hydrogenated under more severe conditions). The di-σ-bonded species could be converted

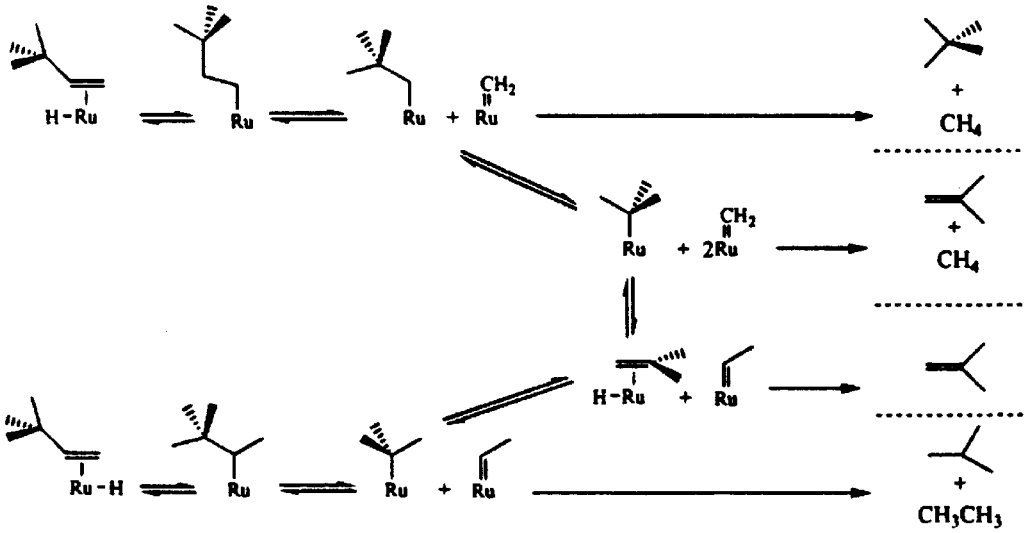

Scheme 14.

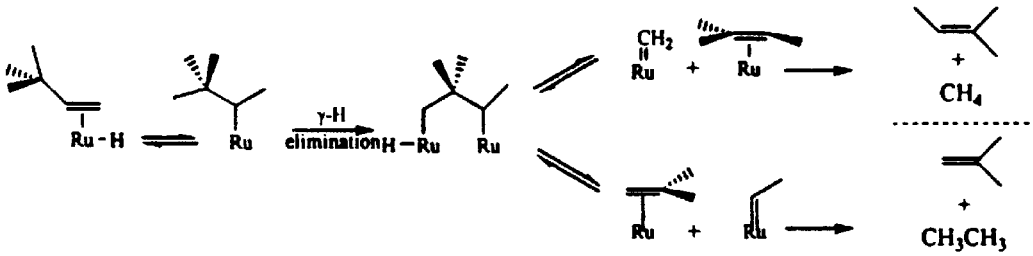

Scheme 15.

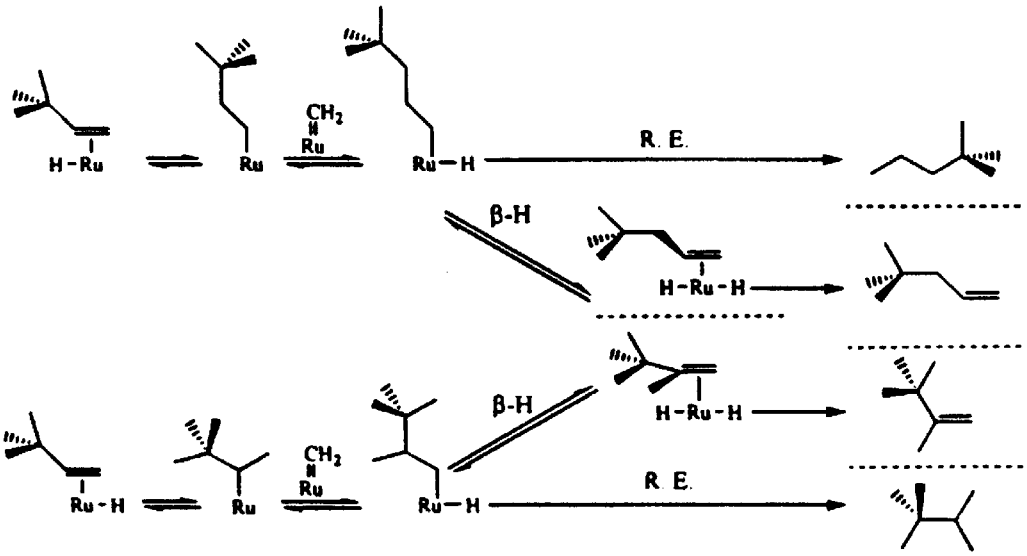

Scheme 16.

into the alkylidyne, and the two competed for surface sites. However, the π-bonded ethylene did not compete for these sites; it was present on the surface at various concentrations depending on the reaction conditions. The authors [430] concluded that the π-bonded ethylene was the only significant hydrocarbon reaction intermediate under these conditions. Because π-bonded ethylene is presumed to be bonded to a single Pt atom of the surface, its identification as a reaction intermediate is consistent with the observation that ethylene hydrogenation is a structure-insensitive reaction [430]. If it is assumed that there is an analogy between the surface-catalyzed reaction and the Wilkinson hydrogenation catalyzed by mononuclear organorhodium phosphine complexes, then it is expected that the π-bonded intermediate would be converted into a σ-bonded ethyl group and then into ethane; however, there is no evidence of these postulated intermediates on the working catalyst surface.

Hydrogenolysis and Homologation of Alkenes. A recent investigation of the reactions of 3,3-dimethyl-1-butene catalyzed by silica-supported ruthenium at temperatures ≥ 373 K [431] illustrates some current mechanistic ideas and the importance of organo-metallic chemistry as a foundation for interpreting mechanisms of surface-catalyzed reactions. The predominant reaction of 3,3-dimethyl-1-butene was hydrogenation. Slower accompanying reactions were homologation and hydrogenolysis, which proceeded at approximately equal rates. The homologation products were 2,2-dimethylpentane and 2,2,3-trimethylbutane in a 10 : 1 ratio, as well as traces of 4,4-dimethyl-1-pentene. The hydrogenolysis products were neopentane, methane, isobutylene, and isobutane, as well as traces of other alkenes and alkanes. In addition, slow skeletal isomerization occurred, giving the nonprimary products 2,3-dimethyl-1-butene and 2,3-dimethyl-2-butene at 373 K.

The product distributions at low conversions were used to infer mechanisms of the homologation and hydrogenolysis reactions. The most plausible mechanism of the hydrogenolysis of 3,3-dimethyl-1-butene was suggested [429] to be one in which the C–C cleavage step is a deinsertion of a methylene fragment from a Ru-neopentyl moiety (Scheme 14). An alternative, proceeding through a metallacyclic intermediate, would involve a γ-hydrogen elimination, which was impossible because the γ-carbon atom in the alkyl chain has no hydrogen atom. The absence of significant formation of 2-methyl-2-butene is consistent with the absence of reaction according to Scheme 15.

Scheme 14 also accounts for isobutylene product. The same mechanism also accounts for the principal homologation products (Scheme 16: R.E. = reductive elimination; β-H = β-hydride elimination) [431].

These mechanisms are no more than formalisms that are consistent with known organometallic chemistry and account for the observed products. The bonding of the intermediates to the ruthenium surfaces is not well understood.

6.5. Selective Hydrocarbon Oxidation Reactions

Selective hydrocarbon oxidation reactions include several important classes of heterogeneously catalyzed reactions, which find large-scale industrial application for the synthesis of bulk chemicals. Reviews on the mechanisms of selective hydrocarbon oxidation [145], oxidative dehydrogenation of alkanes [429], ammoxidation of alkenes, aromatics and alkanes [61], and epoxidation of alkenes [432] are available. Here, some mechanistic aspects of the epoxidation of alkenes and of the ammoxidation of alkenes are discussed.

6.5.1. Epoxidation of Ethylene and Propene [432], [433]

The epoxidation of ethylene by dioxygen is catalyzed by silver metal and yields ethylene oxide (→ Ethylene Oxide), an important intermediate for the synthesis of glycols and polyols.

Total oxidation of the reactant and the target product limit the selectivity of the process. Scheme 17 shows the three competing reactions.

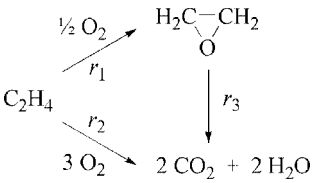

Scheme 17.

The catalyst therefore must be tuned such that the optimal selectivity for ethylene oxide is achieved. The active phase consists of large Ag particles supported on low surface area α-Al_2O_3 promoted by alkali metal salts. A beneficial effect is also obtained by adding chlorine-containing compounds such as vinyl chloride to the reaction feed. Unter reaction conditions this additive is readily combusted on silver, and chlorine is adsorbed on the metal surface.

Oxygen can be adsorbed on transition metals in general and on silver in particular in three different states: (1) molecular dioxygen, (2) adsorbed atomic oxygen, and (3) subsurface atomic oxygen [433]. Molecular oxygen is stable on an Ag(111) surface at temperatures below ca. 220 K. It dissociates at higher temperatures. Oxygen dissociation occurs at high-coordination sites, since at least two neighboring metal atoms must be available. It has been shown that ensembles with a minimum of five silver atoms are required [434], [435]. Oxygen atoms adsorbed originally on the external silver metal surface may move to subsurface lattice positions. Subsurface oxygen atoms have been proved to form on transition metals including Rh, Pd, and Ag [436]. The maximum oxygen coverage on silver surfaces is one oxygen atom per silver atom, corresponding to the composition of AgO [433].

The presence of subsurface oxygen atoms reduces the electron density on adjacent silver atoms. Hence, oxygen atoms adsorbed on the external surface which share bonds to silver surface atoms with subsurface oxygen atoms become highly polarizable. When exposed to ethylene, the interaction of the surface oxygen atoms with the π electrons of ethylene leads to a flow of electron density from the surface oxygen atom to the positively charged surface silver atom [437]. The surface oxygen atoms behave chemically as electrophilic oxygen atoms, which preferentially react with the part of the reactant molecule having the highest electron density. This situation is most likely at high oxygen coverages, consistent with the experimental observation that the epoxidation selectivity is dramatically enhanced by increasing oxygen coverage [438]. Scheme 18 illustrates this scenario [432]. At low oxygen coverages the density of subsurface oxygen atoms is also reduced so that the polarizability of oxygen atoms adsorbed on the external surface is reduced. Consequently, these oxygen atoms behave as nucleophilic oxygen atoms and tend to interact preferentially with hydrogen atoms of the ethylene molecule, thus leading to total oxidation. This situation is schematically shown in Scheme 19 [432]. Therefore, epoxidation selectivity must decrease with decreasing oxygen coverage. The fact that vacant silver sites exist in the vicinity of an adsorbed oxygen atom at low coverage (see Scheme 19), is also detrimental.

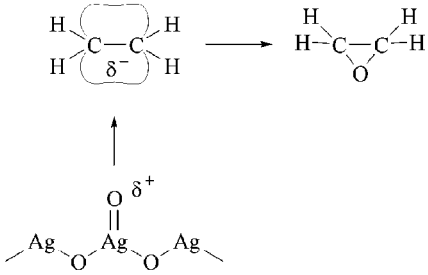

Scheme 18.

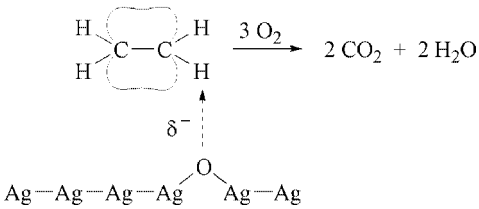

Scheme 19.

The influence of alkali metal and chlorine modifiers is complex. The effect of chlorine is twofold: (1) it suppresses vacant silver sites, and (2) it enhances the electron deficiency of silver. The latter effect is due to the ability of chlorine to also occupy subsurface positions [439] and thus to adopt the role of subsurface oxygen as illustrated in Scheme 20 [432]. These effects

improve the initial selectivity r_1/r_2 (r_i denotes a reaction rate, see Scheme 17). The overall selectivity is also reduced by subsequent combustion of the epoxide (r_3 in Scheme 17), particularly at high conversions. The combustion of the epoxide is induced by the residual acidity of the α-Al_2O_3 support. The presence of alkali metal reduces the density of acid sites and thus has a beneficial effect on selectivity by blocking reaction step r_3.

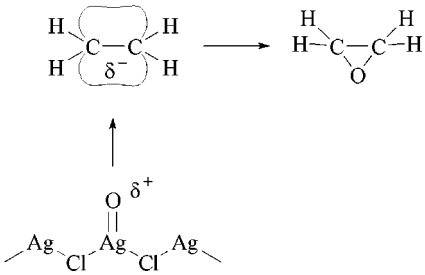

Scheme 20.

The rate-limiting step of the epoxidation reaction is the dissociative chemisorption of dioxygen. Alkali metal compounds enhance the dissociation rate of dioxygen by reducing the activation barrier, and consequently the alkali metal modifier enhances the epoxidation rate as its coverage increases [432]. Interestingly, when a chlorine-modified catalyst is promoted by alkali metal compounds, the reaction rate decreases, and this is suggestive of an enhancement of the steady-state concentration of adsorbed chlorine, which leads to site blocking. Therefore, there is a very subtle interplay between the two additives which must be carefully controlled to optimize conversion and selectivity of the ethylene epoxidation reaction.

The epoxidation of *propene* with dioxygen is unfavorable because of the enhanced reactivity of the methyl group for nucleophilic attack. Activation of the methyl group leads to the allyl or combustion of the propylene epoxide. Alternative oxidants are hydrogen peroxide or hydroperoxide (→ Propylene Oxide, Chap. 4.2.). The reaction of propene with hydrogen peroxide yields the target product propylene epoxide and water (Eq. 55).

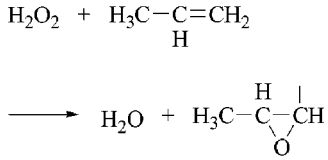

(55)

The preferred catalyst for this reaction is titanium-silicalite-1 (TS-1) (see Section 3.1.1), in which four coordinate Ti^{4+} plays the decisive role [440]. Although the exact nature of the reaction intermediate is not known yet, hydrogen peroxide may coordinate nondissociatively to a Lewis acidic tetrahedral Ti^{4+} site as shown in Scheme 21. This induces electron deficiency on the oxygen atoms of the peroxides, which is favorable for epoxidation. An analogous reaction path has been proposed for the homogeneous epoxidation of propene by peroxides [432].

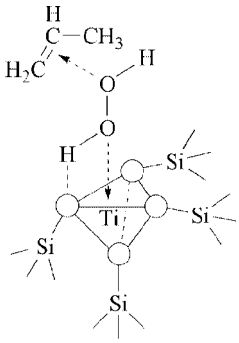

Scheme 21.

6.5.2. Ammoxidation of Alkenes [61], [441]

In ammoxidation, ammonia reacts with a reducible organic molecule, most frequently an alkene, alkane, or aromatic, in the presence of dioxygen to yield nitriles (e.g., Eq. 56).

$$2\,CH_2{=}CRCH_3 + 2\,NH_3 + 3\,O_2 \longrightarrow 2\,CH_2{=}CRCN + 6\,H_2O \qquad (56)$$

The ammoxidation of an alkene is a six-electron oxidation that produces an unsaturated nitrile and water. The reaction is related to the four-electron oxidation of alkenes (Eq. 56) [74] producing unsaturated aldehydes and water, and to

the two-electron oxydehydrogenation of alkenes to dialkenes and water (Eq. 57) [74].

$$CH_2{=}CRCH_3 + O_2 \longrightarrow CH_2{=}CRCHO + H_2O$$

$$2\,CH_2{=}CHCH_2CH_2R + O_2 \longrightarrow 2\,CH_2{=}CHCH{=}CHR + 2\,H_2O \quad (57)$$

Catalysts for these reactions are complex mixed metal oxides containing variable-valence elements (see Section 3.1.1), the ammoxidation catalysts typically being the most complex. These materials possess redox properties, i.e., they can readily be reduced by ammonia and reoxidized by dioxygen present in the gas phase. It is the lattice oxygen of the catalyst which reacts with ammonia and the hydrocarbon, and the reduced solid is reoxidized by gas-phase oxygen (Mars – van Krevelen mechanism [442], see also [145]).

The most important alkene ammoxidation is that of propene to acrylonitrile (Sohio Acrylonitrile Process, Eq. 58; see also → Acrylonitrile, Chap. 5.) [429]

$$2\,CH_2{=}CH{-}CH_3 + 2\,NH_3 + 3\,O_2 \longrightarrow 2\,CH_2{=}CHCN + 6\,H_2O \quad (58)$$

Molybdates and antimonates can be used as catalysts for this reaction. The active sites are thought to have bifunctional nature [61], [443], [444]. A generalized catalytic cycle for alkene ammoxidation is shown in Figure 22 [61]. Ammonia is proposed to interact first with the bifunctional active site generating an ammoxidation site. The alkene coordinates to this site to form an allylic intermediate. After several rearrangements and oxidation steps, the surface intermediate is transformed into the nitrile, which subsequently desorbs. A reduced surface site is thus formed, which is restored to its original fully oxidized state by lattice oxygen O^{2-}, which is provided by adjacent reoxidation sites. These sites then dissociate dioxygen to lattice oxygen. The newly formed lattice oxygen then diffuses to the oxygen-deficient reduced surface site, from where vacancies simultaneously penetrate through the lattice of the solid to the reoxidation sites. Clearly, these sites must communicate with each other via a common solid-state lattice which is capable of facile transport of electron, anion vacancies, and lattice oxygen [61]. As an example, the proposed bifunctional active site of Bi_2MoO_6 (see Section 3.1.1) is schematically illustrated in Figure 23 [158]. The various functionalities were assigned to specific elements and to specific lattice oxygen positions. Bridging oxygen atoms Bi–O–Mo are considered to be responsible for α-hydrogen abstraction from the alkene, while oxygen atoms associated with Mo are responsible for oxygen (Mo=O) and nitrogen (Mo=NH) insertion into an allylic intermediate. The oxygen dissociation and its reduction to lattice oxygen is assumed to occur in the region of high electron density generated by the two lone pair electron orbitals of Bi–O–Bi sites. More easily reducible elements than Bi are Fe, Ce, U, and Cu, which are components of more complex, multicomponent catalysts (see Section 3.1.1) [61]. As an illustration of the mechanisms of ammoxidation and selective oxidation of propene, Figure 24 shows the proposed catalytic cycles for the two reactions [445].

6.6. Hydroprocessing Reactions [46], [212], [213], [447], [448], [449]

(see also → Oil Refining, Chap. 4.)

Hydroprocessing treatment, including hydrodesulfurization (HDS), hydrodenitrogenation (HDN), hydrodeoxygenation (HDO), hydrometalation (HDM), hydrogenation, and hydrocracking, are among the largest industrial processes in terms of catalyst consumption. Crude petroleum contains particularly organosulfur and organonitrogen compounds, which are most abundant in heavy petroleum fractions. These contaminants must be removed for environmental reasons. The reactions take place in the presence of H_2 at high temperatures (ca. 600 – 700 K) and pressures of 500 kPa to 1 MPa. Because of the lower reactivity of organonitrogen compounds as compared to organosulfur compounds, the reaction conditions are more severe for HDN than for HDS.

Catalysts for hydroprocessing are highly dispersed metal sulfides (mainly MoS_2, but also WS_2) supported on γ-Al_2O_3. The materials are promoted by cobalt or nickel, depending on application.

Although the detailed mechanisms have not yet been elucidated, significant progress in understanding the chemistry of the various hydroprocessing reactions at a molecular level has

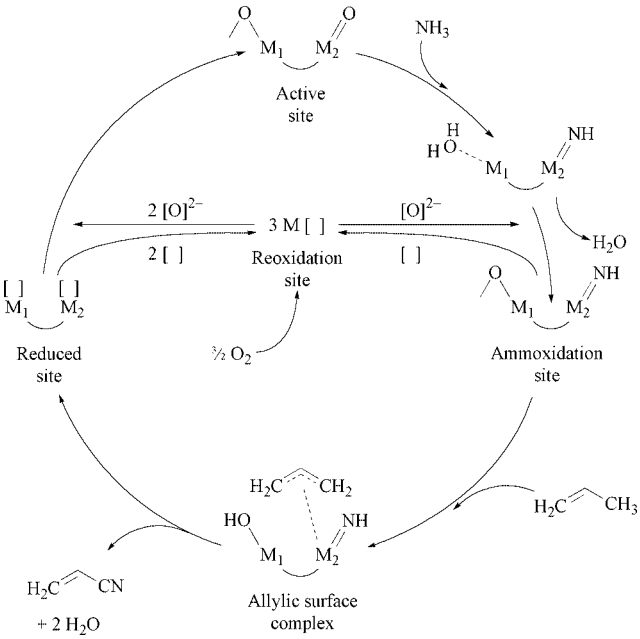

Figure 22. Generalized mechanistic cycle for alkene ammoxidation

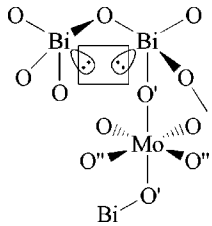

Figure 23. Schematic representation of the active site of Bi_2MoO_6 [446]
O′ = Oxygen responsible for α-H abstraction; O″ = Oxygen associated with Mo; reponsible for oxygen insertion into the allylic intermediate; □ = Proposed center for O_2 reduction and dissociative chemisorption.

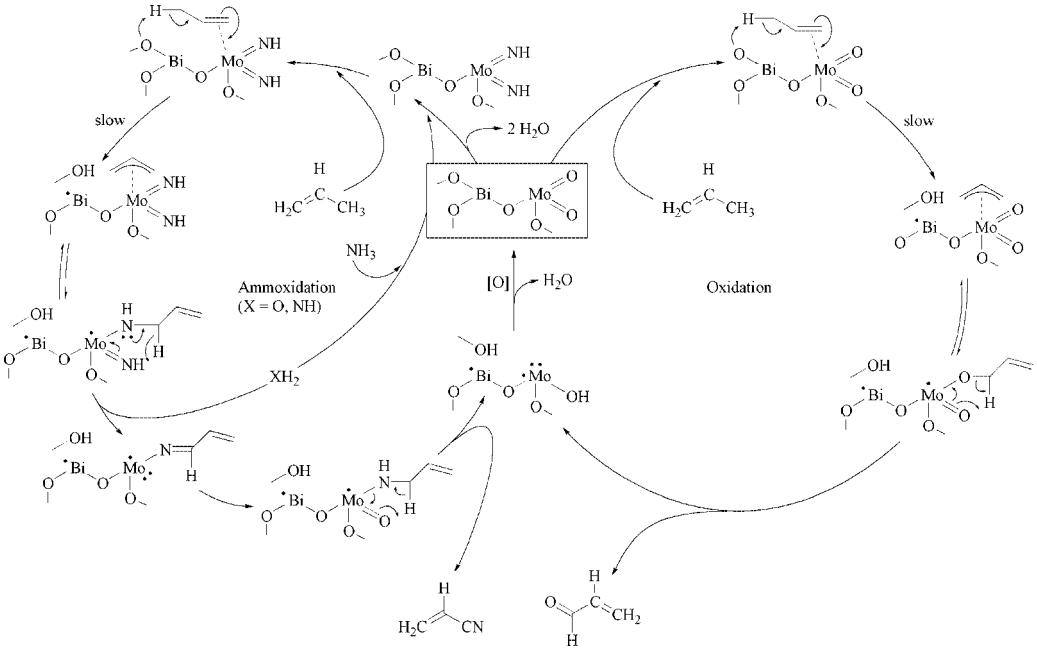

Figure 24. Mechanism of selective ammoxidation and oxidation of propene over bismuth molybdate catalysts [445]

been made [46], [212], [213], [448], [449]. In the following, however, the focus is on reaction networks of several hydroprocessing reactions with pseudo-first-order rate constants for individual reaction steps.

The organosulfur compounds in petroleum include sulfides, disulfides, and aromatics (including thiophene, benzothiophene, dibenzothiophene, and related compounds). Benzo- and dibenzothiophene are predominant in heavy fuels. The reaction network for hydrodesulfurization of dibenzothiophene, a representative member of organosulfur contaminants in fuel, is shown in Figure 25 [450]. Hydrogenation and hydrogenolysis take place in parallel. The latter reaction is essentially irreversible and leads to the formation of H_2S and biphenyl. At low H_2S concentrations in the feed, the sulfide catalysts are highly selective for hydrogenolysis. The selectivity, however, drops sharply as the H_2S concentration in the feed increases.

Hydroprocessing reactions accompanying hydrogenation and hydrodesulfurization include hydrodenitrogenation, whereby organonitrogen compounds in the feed react with H_2 to give NH_3 and hydrocarbons. As an example, a reaction network for the hydroprocessing of quinoline is shown in Figure 26 [447]. The supported metal sulfide catalysts are much less selective for nitrogen removal than for sulfur removal.

7. Application of Catalysis in Industrial Chemistry

Modern industrial chemistry is predominantly based on catalytic processes, of which 80 – 85 % are heterogenously catalyzed. The processes include the production of inorganic and organic chemicals, petroleum refining, production of synthetic fuels, pollution control, and energy conversion.

7.1. Important Inorganic Reactions

Inorganic catalytic reactions represent different key steps in the production of more than a million metric tons of inorganic chemicals.

Production of hydrogen (→ Hydrogen, Chap. 4.1.) is based on the water gas shift (WGS) reaction:

$$CO + H_2O \rightleftharpoons H_2 + CO_2$$

The WGS reaction is reversible and exothermic. To increase the hydrogen content in the product it is thus desirable to perform the WGS reaction at low temperature. However, for technical reasons the reaction is performed industrially in two steps. The first step proceeds at high temperature (HTS) and partly converts CO to CO_2 and H_2. In the second step the remaining of CO reacts at low temperature (LTS) [451].

One source of carbon monoxide is methane or naphtha steam reforming performed on α-alumina supported Ni.

$$CH_4 + H_2O \longrightarrow CO + 3\,H_2$$

Ammonia synthesis (→ Ammonia) from hydrogen and nitrogen:

$$N_2 + 3\,H_2 \rightleftharpoons 2\,NH_3$$

The production of ammonia is one of the largest catalytic processes in the world and has a strategic position as the key to nitrogen chemistry. The synthesis reaction is exothermic and reversible. Low temperature and high pressure facilitate ammonia formation [45], [452].

Nitric acid production (→ Nitric Acid, Nitrous Acid, and Nitrogen Oxides) is based on the catalytic oxidation of ammonia with atmospheric oxygen:

$$NH_3 + 1.25\,O_2 \longrightarrow NO + 1.5\,H_2O$$

The key reaction is exothermic and irreversible at 1070 – 1170 K. It is performed at slightly elevated pressure on a suitable catalyst. Subsequent production steps do not require a solid catalyst:

$$NO + O_2 \longrightarrow 2\,NO_2$$

$$3\,NO_2 + H_2O \longrightarrow 2\,HNO_3 + NO$$

Nitric acid is one of the most important industrial chemicals in terms of tonnage [183].

Production of sulfuric acid (→ Sulfuric Acid and Sulfur Trioxide) includes heterogenously catalyzed oxidation of SO_2:

$$SO_2 + 1/2\,O_2 \rightleftharpoons SO_3$$

$$SO_3 + H_2O \longrightarrow H_2SO_4$$

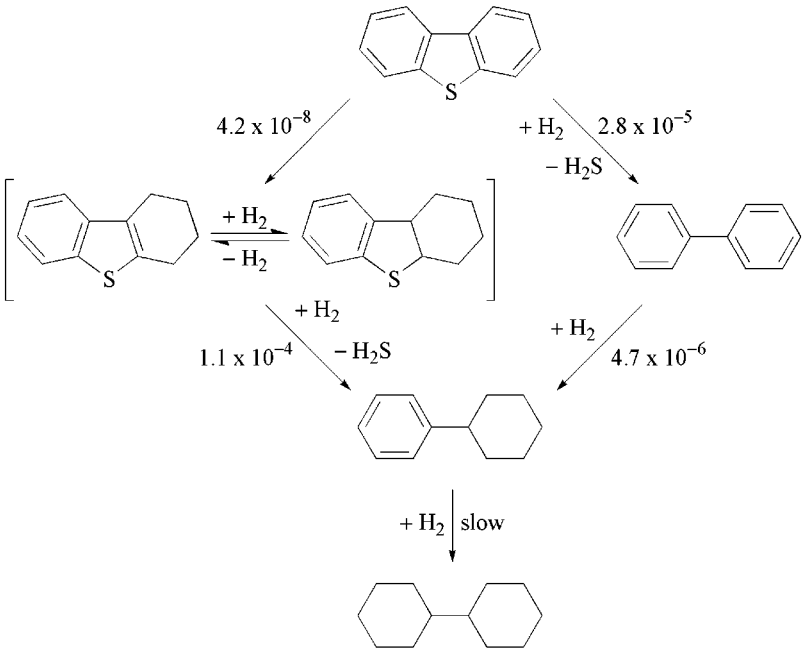

Figure 25. Reaction network for hydrodesulfurization and hydrogenation of dibenzothiophene catalyzed by sulfided $Co-Mo/Al_2O_3$ at 570 K and 10 MPa [450]
Numbers next to the arrows represent the pseudo-first-order rate constants in units of L/(g of catalyst · s) when the H_2S concentration is very small. Addition of H_2S markedly decreases the selectivity for hydrodesulfurization.

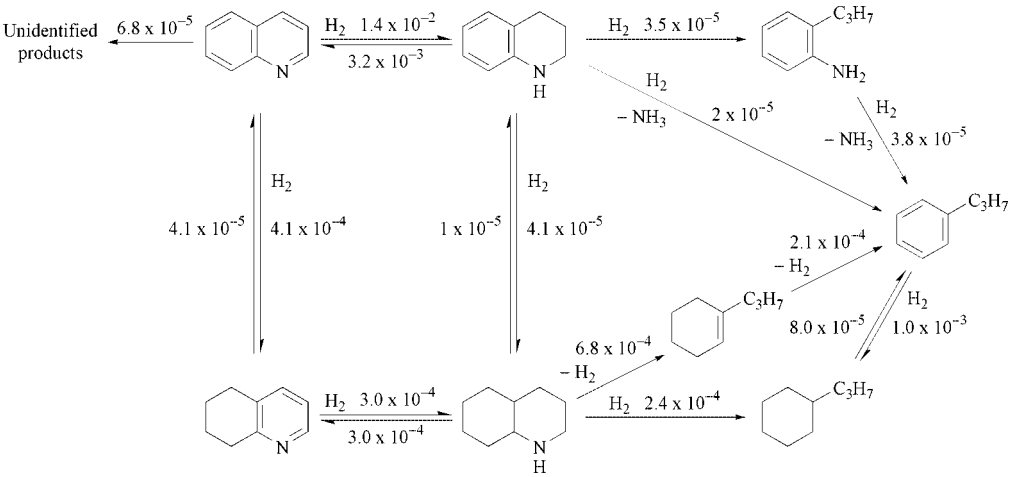

Figure 26. Reaction network for hydrogenation and the hydrogenolysis of quinoline catalyzed by sulfided $Ni-Mo/\gamma\text{-}Al_2O_3$ at 620 K and 3.5 MPa [447]
Numbers next to the arrows represent the pseudo-first-order rate constants in units of L/(g of catalyst · s) when the H_2S concentration is small but sufficient to maintain the catalysts in the sulfided forms.

SO_2 oxidation is a reversible and exothermic reaction. On an industrial scale it is carried out at 670 – 820 K. The worldwide production amounts to 150×10^6 t/a [240].

Sulfur recovery (Claus process) according to the following reactions (→ Sulfur, Chap. 4.):

$$2\,H_2S + O_2 \longrightarrow S_2 + 2\,H_2O$$

$$2\,H_2S + SO_2 \rightleftharpoons 3/2\,S_2 + 2\,H_2O$$

The second reaction is reversible and is performed in practice between 470 – 570 K. The Claus process is used by petroleum refineries to recover sulfur from H_2S obtained by hydrodesulfurization. Hundreds of Claus plants are in operation worldwide with individual capacities of 2000 t/d or more [121].

Solid catalysts applied in the above-mentioned processes are summarized in Table 11.

Table 11. Catalysts applied in some processes in inorganic reactions [451]

Process	Catalyst	Ref.
Hydrogen production by WGS		[451]
HTS	$Cr_2O_3 - Fe_2O_3$	
LTS	CuO – ZnO	
Ammonia synthesis	Fe_3O_4 (K, Ca, Mg, Al oxides)	[45], [452]
Ammonia oxidation	Pt/Rh wire gauze	[183]
SO_2 oxidation	V_2O_5 (K_2SO_4)	[240]
Claus process	Al_2O_3 or TiO_2	[121]

7.2. Important Organic Reactions

While the application of solid catalysts in inorganic reactions is 1imited to a few large-scale processes, their use in the production of organic chemicals is very broad. More than two hundred different catalysts are applied currently in the manufacture of organic intermediates, fine chemicals, and pharmaceuticals [453].

Important catalytic processes for the production of organic chemicals include the following reaction classes:

Hydrogenations (→ Hydrogenation and Dehydrogenation) are typical reversible and exothermic reactions. They are performed mainly at lower temperatures in the liquid or in the vapor phase at atmospheric or elevated hydrogen pressure in the presence of supported metallic catalysts containing, e.g., Ni, Co, Cu, Pt, Pd. Some examples of catalysts used in total hydrogenations [454] are collected in Table 12.

Table 12. Catalysts applied in total hydrogenations [454]

Process	Catalyst
Olefins → paraffins (olefinic gasoline or diesel)	Ni – or Pt – alumina
Benzene → cyclohexane	Ni – alumina, Ni – kieselguhr, Pt – alumina
Naphthalene → tetra- and decahydronaphthalenes	Ni – alumina, Ni – kieselguhr, Pt – alumina
Phenol → cyclohexanol	Ni – kieselguhr, Pd – kieselguhr
Aldehydes and ketones → alcohols	Ni – kieselguhr, Ni (Cu,Co) – kieselguhr, Cu (Ba,Mn) – chromites
Nitro compounds → amines	Pd – kieselguhr, Co – kieselguhr
Nitriles → amines	Co – kieselguhr, Cu – chromites
Carboxylic acid esters → alcohols (Henkel, Lurgi)	Cu (Ba,Mn) – chromites

Of major industrial importance is the hydrogenation of olefins in gasoline and diesel fractions.

The hydrogenation of benzene to cyclohexane (→ Cyclohexane) is the first step in the manufacture of adipic acid (→ Adipic Acid, Chap. 5.) and finally of synthetic fibers, e.g., Nylon 66. In a standard process, a Ni catalyst is used. In a new technology, realized by IFP, a homogeneous Ni catalyst [$Ni^{2+}/Al(Et)_3$] is used [455].

The manufacture of synthetic fibers (Nylon 6), based on caprolactam starts with the hydrogenation of phenol to cyclohexanol.

The hydrogenation of carboxylic esters is carried out under severe conditions (470 – 570 K, 200 – 300 bar hydrogen pressure).

A special class of reactions are selective hydrogenations of unsaturated or functionalized hydrocarbons. This process plays an important role in the removal of alkyne compounds from $C_2 - C_5$ cuts produced by naphtha pyrolysis [428], [456].

The direct hydrogenation of benzene to cyclohexene is a new process, commercialized in 1995 [457].

The partial saturation of double bonds in fats and oils provides the possibility to obtain variable products in the food industry. The world ca-

pacity of fat-hardening plants in 1993 was about 66×10^6 t [459].

Table 13 summarizes selective hydrogenation reactions together with the catalyst used in each particular case.

Table 13. Catalysts applied in selective hydrogenations [453], [456], [458], [459]

Process	Catalyst
Acetylene → ethylene	Pd – alumina, Pd(Ag) – alumina
Diolefins → olefins (1,3-butadiene, isoprene, cyclopentadiene)	Pd – alumina, Pd – silica
Benzene → cyclohexene	Ru – alumina
Fats and edible oils (fat hardening)	Ni – kieselguhr, Ni(Zr) – kieselguhr
Unsaturated aldehydes → saturated aldehydes	Cu – alumina, Raney Pd
Unsaturated aldehydes → unsaturated alcohols	Rh(Sn) – silica, Pt(Fe,Co) – silica

Another special class of organic reactions involves the hydrogenation of CO. The examples presented in Table 14 show the large variety of reactions and products formed in CO hydrogenation.

Table 14. Catalysts applied in CO hydrogenations and related reactions [460], [462], [464]

Process	Catalyst
CO methanation $CO + 3\,H_2 \rightleftharpoons CH_4 + H_2O$	Ni – alumina, Ru – alumina
Methanol synthesis $CO + 2\,H_2 \rightleftharpoons CH_3OH$	$CuO - ZnO - Al_2O_3$
Hydroformylation (Oxo synthesis) $CO + H_2 + \text{olefin} \rightarrow \text{aldehyde}$	
(Aldox, Shell)	$[HCo(CO)_3PBu_3]$
(Lp oxo, Davy McKee, Union Carbide)	$[HRh(CO)(PPh_3)_3]$
Fischer – Tropsch synthesis	Fe oxides (K)
Methanol carbonylation $CH_3OH + CO \rightarrow CH_3COOH$ (Monsanto)	$[Rh(CO)_2I_2]^-$

One possibility to remove CO impurities from hydrogen, used in ammonia synthesis, is its conversion to methane. Above 500 K and at a pressure of 10 MPa the application of Ni-based catalysts is common. Below 470 K, Ni catalysts cannot be used because of $[Ni(CO)_4]$ formation. Ru-containing materials are the low-temperature catalysts of choice [265].

The world production capacity of methanol (→ Methanol) reached more than 26×10^6 metric tons in 1995. At present plants using low-pressure technology (5 – 7 MPa) operate at up to 490 – 530 K with Cu – Zn – Al catalysts. Older high-pressure plants using $ZnO - Cr_2O_3$ catalysts have mainly ceased operation [460].

Three Fischer – Tropsch synthesis plants (2×10^6 t/a) are currently operating in South Africa. With fixed-bed technology, the main products are waxes (→ Waxes, Chap. 5.). In fluidized-bed reactors mainly gasoline and a variety of oxygen-containing compounds are produced (→ Coal Liquefaction, Chap. 2.2.) [461].

The production of aldehydes from olefins (→ Oxo Synthesis) is one of the most important processes of homogeneous catalysis. The older oxo process, realized by Shell, is using Co-catalyst and operates at 100 bar pressure. The newer LP-oxo process based on Rh catalysts performs at only 2 MPa [462], [463].

The carbonylation of methanol to acetic acid [464] currently dominates all other processes for acetic acid production (→ Acetic Acid).

Dehydrogenation (→ Hydrogenation and Dehydrogenation) is the reverse reaction of hydrogenation. Dehydrogenations are endothermic and are facilitated at high temperature and low pressure. From the industrial view point important dehydrogenations are those of $C_3 - C_4$ aliphatic hydrocarbons to produce olefins or dienes [218]. The dehydrogenation of ethylbenzene (→ Styrene) [465] and of some aliphatic alcohols [466] are also industrially important. Catalysts applied in dehydrogenations are summarized in Table 15.

Table 15. Catalysts applied in industrial dehydrogenations [218], [465], [466]

Process	Catalyst
$C_3 - C_4$ (iso-)Alkanes → olefins	
Catofin (Houdry)	$Cr_2O_3 - Al_2O_3$ (K), $Cr_2O_3 - ZrO_2$
Oleflex (UOP)	Pt(Sn) – alumina
STAR (Phillips)	$Pt(Sn) - ZnAl_2O_4$, $MgAl_2O_4$
Butane → butadiene (Catadiene, Houdry)	$Cr_2O_3 - Al_2O_3$ (K)
$C_{10} - C_{14}$ Alkanes → linear olefins	
(Pacol-Olex)	Pt(Sn) – alumina
Ethylbenzene → styrene	Fe oxides (K, Cr, Mo, Ce)
2-Butanol → methylethylketone	ZnO, Cu – kieselguhr
Cyclohexanol → cyclohexanone	ZnO – CaO, Cu – kieselguhr

Table 16. Catalysts applied in some industrially important elimination and addition reactions [189], [389], [405], [467], [469]

Process	Catalyst
Hydration	
Propene → 2-isopropanol	H_3PO_4 – kieselguhr
Dehydration	
2-Phenylethanol → styrene	alumina
Methanol → dimethylether	alumina
Alkylation	
Benzene → Ethylbenzene	
Monsanto	$AlCl_3$ (liq. phase)
Alkar (UOP)	BF_3 – alumina
Mobil-Badger	Zeolites (ZSM-5)
Benzene → isopropylbenzene (UOP)	H_3PO_4 – kieselguhr
Dealkylation	
Diethylbenzene → ethylbenzene, benzene	alumosilicate
Disopropylbenzene → isopropylbenzene, benzene	alumosilicate
Additions	
Isobutene + CH_3OH → methyl *tert*-butyl ether (MTBE)	Acid ion exchanger (Amberlyst)
Isopentene + CH_3OH → methyl *tert*-amyl ether	Acid ion exchanger (Amberlyst)
Oligomerization	
Propene → C_3H_6 (dimers), C_6H_{12} (trimers), $C_{12}H_{24}$ (tetramers)	H_3PO_4 – kieselguhr, zeolites (ZSM-5, mordenite), Amberlyst

The highly endothermic dehydrogenation of lower alkanes can be facilitated by the introduction of oxygen to burn the formed hydrogen and produce the required heat. These so-called *oxidative dehydrogenations* are exothermic and can be realized at lower temperatures. On an industrial scale, the oxidative dehydrogenation of *n*-butane to butenes on supported Sn, Bi, Co, and Ni oxides and of butenes to butadiene (→ Butadiene) over Cu and Ni phosphates were realized [218].

Elimination and Addition Reactions. There are various reversible organic reactions such as hydration – dehydration, alkylation – dealkylation, etc. Catalysts applied in these reactions have mainly acidic properties, for example solid phosphoric acid, aluminosilicates, zeolites, etc. [467]. The industrial importance of dehydrations and hydrations has decreased. An exception is methanol dehydration, the first step in the Mobil fuel process.

The production of ethylbenzene (→ Ethylbenzene) and isopropylbenzenes by alkylation and the addition of methanol to isobutene or isopentene are multimillion-ton processes. Some examples of elimination and addition reactions [189], [389], [405], [468], [469] are presented in Table 16.

Oxidations. Catalytic oxidations of hydrocarbons occupy a prominent place in the modern chemical industry, because they represent the easiest route for the functionalization of hydrocarbon molecules [145].

Oxidation of organic compounds in the liquid phase is a traditional area of homogeneous catalysis [470], [471]. The common feature of these reactions is a high selectivity. Some established industrial processes and applied catalysts are summarized in Table 17.

Table 17. Catalysts applied in liquid-phase oxidation [470], [471]

Process	Catalyst
n-Butane → acetic acid	Cr, Mn, Co, Ni acetates, naphthenates
Cyclohexane → cyclohexanol/one	Co naphthenate
Toluene → benzoic acid	Co, Mn acetates
Benzoic acid → phenol	Cu, Mn benzoates
p-Xylene → terephtalic acid	Mn acetate
Acetaldehyde → acetic acid anhydride	Co acetate
Ethylene → acetaldehyde (Wacker process)	$PdCl_2$ – $CuCl_2$
Propene → acetone	$PdCl_2$ – $CuCl_2$
Ethylene + CH_3COOH → vinyl acetate	$PdCl_2$ – $CuCl_2$

Table 18. Solid catalysts applied in vapor-phase oxidation [145], [146], [432], [472], [473]

Process	Catalyst
Ethylene → ethylene oxide	Ag(Cs) – alumina
Propylene + H_2O_2 → propylene oxide (Enichem)	Ti silicalite
Propene → acrolein	Mo, Bi, Fe oxides
Acrolein → acrylic acid (Mitsubishi Ryon, Nippon Kayku)	Mo, V oxides
Isobutene → methacrolein	Mo, Bi oxides
Methacrolein → methacrylic acid (Nippon Kayku, Nippon Shokubai)	Mo, V, heteropoly acids
o-Xylene, naphthalene → phthalic anhydrate	
(Alusuisse)	V_2O_5 – Al_2O_3
(Nippon Shokubai)	V_2O_5 – TiO_2
n-Butane → maleic anhydrate (Lummus Crest)	$(VO)_2P_2O_7$ (Fe)
Benzene → maleic anhydrate	V_2O_5 – MoO_3 (P, Ni)
Methanol → formaldehyde	Fe molybdate + MoO_3

In the past 15 years several oxidation processes utilizing solid catalysts and operating in the vapor phase were installed in the chemical industry, e.g., oxidation of propene or isobutene to unsaturated aldehydes or acids and the production of phtalic and maleic anhydrates based on the oxidation of hydrocarbons etc. [145], [146], [432], [472], [473]. Table 18 lists some important processes.

Ammoxidations. The interaction of ammonia with hydrocarbons and oxygen in the presence of a suitable catalyst results in the formation of the corresponding nitrile. Acrylonitrile, benzonitrile and nitriles obtained in the reaction with *o*-, *m*- and *p*-xylenes achieved a high industrial importance [61].

In the reaction of cyclohexanone with NH_3 and H_2O_2, cyclohexanone oxime is formed [474]. Catalysts applied in various ammoxidations are described in Table 19.

Table 19. Catalysts applied in the ammoxidation of hydrocarbons [61], [474]

Process	Catalyst
Propene → acrylonitrile (Sohio)	Bi – molybdate (Fe,Ni,Co)
Isobutene → methacrylonitrile	Bi – Sb molybdate (Fe)
Toluene → benzonitrile	Bi – P molybdate (Fe,Ni,Co) on SiO_2
o-, *m*-, *p*-Xylene → phthalonitriles	Sb, V, Bi, Fe oxide on TiO_2
Cyclohexanone + H_2O_2 → cyclohexanone oxime (Enichem)	Ti silicate

Dimerizations, Metathesis, Polymerizations. The high demand of various higher olefins, especially in the range $C_6 - C_{12}$, initiated intensive research in the field of dimerization. IFP developed the dimersol process for converting propene into hexenes or 1-butene into octenes [475]. In the SHOP process, $C_{10} - C_{18}$ α-olefins are synthesized from ethylene [217].

Alkene metathesis is a catalytic reaction in which olefins are converted into a new product by rupture and reformation of the carbon – carbon bonds. For example, isopentene is produced on industrial scale from isobutene and *n*-butene [217].

Major achievements in the area of catalytic polymerization in the past led to the development of a new branch of the chemical industry [245], [476]. In 1995 the production of polyethylene and polypropylene amounted to 44×10^6 t (→ Polyolefins). Catalysts applied in these reactions are listed in Table 20.

Table 20. Catalysts applied in dimerizations, metathesis, and polymerization [217], [245], [475], [476]

Process	Catalyst
Propene → hexenes	
1-Butene → octenes (Dimersol, IFP)	$Ni^{2+} - AlEtCl_2$
Ethylene → $C_{10} - C_{18}$ 1-olefins (SHOP, Shell)	bis(cyclooctadienyl)Ni
Isobutene + 2-butene → isopentene + propene	
$(CH_3)_2C{=}CH_2 + 2\text{-}C_4H_8 \rightarrow (CH_3)_2C{=}CHCH_3 + C_3H_8$ (Phillips)	$MoO_3 - SiO_2$, $WO_3 - SiO_2$, $Re_2O_7 - Al_2O_3$
Cyclohexene + ethylene → 1,7-cyclooctadiene	$Re_2O_7 - Al_2O_3$
Ethylene → polyethylene (HDPE, Phillips)	
370 K, 70 bar	$CrO_3 - SiO_2$
Ethylene → polyethylene (LDPE, Ziegler – Natta, Solvay)	$TiCl_4 - Al(Et)_3$
2 – 10 bar	$TiCl_3 - AlEt_2Cl - MgCl_2$
Propene → polypropylene	$TiCl_3 - AlEt_2Cl - MgCl_2$

7.3. Petroleum Refining (see also → Oil Refining)

The integration of chemical, and especially of catalytic processes into crude oil upgrading, created a new industrial branch called petroleum refining. The main task of this industry is to ensure high yield and good quality of liquefied petroleum gas (LPG), gasoline, diesel fuel, jet fuel, etc. The major processes in catalytic refining are catalytic cracking, hydrotreating, reforming, and hydrocracking [477].

The primary objective of catalytic cracking, currently performed exclusively with fluidized-bed technology (FCC), is to increase the yield of high-octane gasoline from crude oil. FCC process conditions are 750 – 810 K and atmospheric or slightly reduced pressure [394].

The hydrotreating processes include hydrodesulfurization (HDS), hydrodenitrogenation (HDN), hydrodeoxygenation (HDO), and demetalation. In these processes N, S, O, and metals such as V and Ni are removed from petroleum fractions mainly at 570 – 670 K and 2 – 4 MPa hydrogen pressure [46], [212], [230].

The main purpose of reforming is to increase the octane number of gasoline by isomerization of alkanes and naphthenes and above all by the conversion of these hydrocarbons into aromatics. It is usually performed at 720 – 820 K and slightly elevated pressure [47].

Hydrocracking was developed for the conversion of relatively heavy oil feedstocks into lighter transport fuel products at 570 – 720 K and 5 – 10 MPa hydrogen pressure. It is performed in one or two stages [478].

Other processes run by modern refineries are alkylation [406], isomerization [386], dewaxing [478], and etherification (MTBE and TAME production) [189].

Table 21 lists typical catalysts used in petroleum refining.

Table 21. Catalysts applied in some processes in petroleum refining [47], [394], [477]

Process	Catalysts
Fluid catalytic cracking (FCC)	ultrastable (US) Y zeolite in aluminosilicate matrix, US Y zeolite + ZSM-5 + aluminosilicate
Hydrotreating (HDS, HDN, HDO)	$Co-MoO_3-Al_2O_3$, $Ni-WO_3-Al_2O_3$
Catalytic reforming	
UOP	Pt(Sn) – alumina
Chevron	Pt(Re) – alumina
EXXON	Pt(Ir) – alumina
Hydrocracking	
single-stage	$Ni-MoO_3-Y$ zeolite – alumina, $Ni-WO_3$ – alumina
two-stage	$Ni-MoO_3-Y$ zeolite – alumina, Pd – or Pt – mordenite – alumina
Dewaxing	$Ni-WO_3$ – alumina (HF), Ni-ZSM-5, Pt – zeolite β
Alkylation of isobutane with C_3-C_4 olefins (commercial process, UOP, etc.)	HF/BF_3, H_2SO_4
Chevron (pilot stage)	SbF_5-SiO_3
Isomerization of C_4-C_6 alkanes	
Penex (UOP)	Pt – alumina (HCl-activated)
Hysomer (Shell)	Pt – H mordenite

7.4. Production of Synthetic Fuels

The energy crisis in the early 1970s initiated intensive research into direct and indirect synthesis of fuels.

An example of a direct route is coal liquefaction based on catalytic coal hydrogenation. The H-Coal process (→ Coal Liquefaction, Chap. 3.9.1.) developed by Hydrocarbon Research is carried out at 670 – 720 K and 14 – 28 MPa hydrogen pressure in the presence of $Co-MoO_3$ catalysts [479], [480].

In the Exxon H-donor solvent process (→ Coal Liquefaction, Chap. 3.9.4.) the hydrogen transfer from solvent to coal proceeds at 670 K and 1 – 4 MPa. Solvent recovered from the liquefied coal is saturated with hydrogen in the presence of an $Ni-MoO_3$ catalyst and recycled.

The indirect route is represented by Fischer – Tropsch (FT) synthesis and above all by the Mobil process for converting methanol to gasoline (MTG) on a shape selective zeolite (ZSM-5).

The FT synthesis was discussed in Section 7.2.

Analogous to the FT synthesis is the SMDS process (Shell Middle Distillate Synthesis) established quite recently in an operation in Malaysia. The process consists of two basic steps: Heavy paraffins synthesis from CO and H_2 over a Zr-promoted Co catalyst and selective hydrocracking of heavy paraffins to middle distillate over a zeolitic catalyst [481].

The Mobil MTG process is based on the catalytic conversion of methanol to hydrocarbons over a shape-selective zeolite (ZSM-5) [482]. The conversion process consists of a series of consecutive reactions: first dimethyl ether is formed, followed by the formation of alkenes, alkanes, and aromatics. In fixed-bed reactors, methanol conversion takes place at 620 – 685 K and about 0.2 MPa over alumina and ZSM-5. The MTG process was commercialized in New Zealand, where it is used to convert offshore natural gas via methanol into high-octane gasoline. In the fluid-bed MTO process (methanol to olefins) the main products at 740 – 780 K are olefins, which can be oligomerized over ZSM-5 to jet fuel.

7.5. Pollution Control

7.5.1. Mobile Sources

Road traffic is one of the major sources of CO, hydrocarbons, and NO_x emissions. In 1995 more than 600 million vehicles were on the roads

world-wide, mainly in North America, Europe, and Japan.

Since 1975 – 1980 catalytic converters have been standard equipment of gasoline driven engines in the USA, Western Europe, and Japan (→ Automotive Exhaust Control).

At present, predominantly three-way catalysts are installed in cars. The heart of the catalytic converter is a cylindrical cordierite monolith having 62 channels per square centimeter. This structure is coated with a thin layer of washcoat consisting of a mixture of inorganic oxides such as Al_2O_3, CeO_2, LaO_2 etc. In the washcoat precious metals such as Pt and Rh (5 – 20 : 1) are dispersed. Instead of cordierite monoliths, metall supports are sometimes applied [198].

The emissions are removed by the following catalytic reactions taking place in the temperature range 570 – 770 K:

$$C_mH_n + (m + 0.25\,n)\,O_2 \longrightarrow m\,CO_2 + 0.5\,n\,H_2O$$

$$2m\,H_2O + C_mH_n \longrightarrow m\,CO_2 + (2m + 0.5\,n)\,H_2$$

$$CO + H_2O \longrightarrow CO + H_2$$

$$CO + NO \longrightarrow 0.5\,N_2 + CO_2$$

$$NO + H_2 \longrightarrow 0.5\,N_2 + H_2O$$

$$CO + 0.5\,O_2 \longrightarrow CO_2$$

Since 1991, catalytic converters have been also installed in passenger cars with diesel engines in the European Community. Emissions such as CO, aliphatic and aromatic hydrocarbons, aldehydes, and soot present in the exhaust gas are oxidized on a catalyst to CO_2 and H_2O. The catalyst consists of a cordierite monolith covered by a oxide washcoat and precious metals such as Pt or Pd.

The removal of NO_x from the exhaust gases of diesel or lean burn engines is difficult because of the high concentration of oxygen. The consumption of reducing agents, e.g., H_2, CO, or olefins, would be too high. However, catalytic NO decomposition to N_2 and O_2 is technically feasible. Suitable catalysts are Cu-containing ZSM-5 or oxides such as Co_3O_4, CuO, NiO, Fe_2O_3, and ZrO_2 [483].

7.5.2. Stationary Sources

Fossil fuels such as coal, oil, gas, and others are burnt or gasified for energy conversion. In Western European countries and Japan measures have been implemented since 1980 for reducing emissions, especially of NO_x from power plants.

One method to remove NO_x from exhaust gases of power plants, industrial boilers, and gas turbines is based on selective catalytic reduction (SCR) with ammonia in the presence of oxygen:

$$4\,NO + 2\,NH_3 + O_2 \longrightarrow 4\,N_2 + 6\,H_2O$$

Catalysts applied are mainly monoliths with 70 – 100 cm length, a cross section of 15 × 15 cm, and various numbers of channels. Industrial catalysts consist mainly of V_2O_5 supported on TiO_2. They operate in the presence of SO_2 at 470 – 670 K [210].

7.5.3. New Processes in Environmental Catalysis

The cleaning of the air from road tunnels and car parks requires CO oxidation at ambient temperature. Au on Co_3O_4 is highly active [484].

The high concentration of ozone in the recycle air of silicon chip factories is diminished by O_3 decomposition on Pd alumina catalysts at 420 K [485].

The removal of various odors from exhaust air, produced by bakeries, fish-smoking factories, coffee roasters, etc. is carried out on monolith Pt-alumina catalysts at 470 – 520 K [485].

7.6. Energy Conversion

7.6.1. Fuel Cells

Chemical energy is converted to electrical energy in fuel cells (→ Fuel Cells) by using solid catalysts on both the anode and the cathode. Fuel cells were developed and successfully applied, e.g., as a source of power in spacecraft and in some specical road vehicles. Till now various cells were tested, but only H_2/O_2 cells in the form of PAFC (phosphoric acid fuel cells) and PEMC (proton exchange membrane cells) are of technical relevance, especially modules producing 40 – 1200 kW [486].

The reaction of hydrogen on the anode

$$H_2 \rightleftharpoons 2\,H^+ + 2\,e^-$$

is catalyzed by Pt/Pd or Pt/Ru alloys. The latter is less sensitive to CO, while hydrogen for fuel cells is mainly produced by methane steam reforming:

$$CH_4 + H_2O \rightleftharpoons 3\,H_2 + CO$$

$$CO + H_2O \rightleftharpoons H_2 + CO_2$$

Catalysts for the oxygen reduction proceeding on the cathode:

$$0.5\,O_2 + 2\,H^+ + 2\,e^- \longrightarrow H_2O$$

are Pt or Pt – V alloy finely dispersed on a carbon support. PAF cells operate at 290 – 370 K and 0.1 – 0.6 MPa pressure [486].

Recent developments have shown that methanol fuel cells can be an attractive and realistic alternative.

7.6.2. Catalytic Combustion

Flameless combustion is a process in which the fuel reacts with oxygen on a catalyst surface providing CO_2, H_2O, and thermal energy. This process proceeds at much lower temperatures (1270 – 1570 K) than conventional combustion, and hence less NO_x (< 5 ppm) is formed [487]. Conventional catalysts consist of ceramic or metal monoliths covered by alumina stabilized with barium oxide and bearing finely dispersed precious metals, particularly Pt and Pd [487].

8. Technology of Catalytic Processes

8.1. Catalytic Reactors [488–492]

The development of a catalytic process involves the search for the catalyst and the appropriate reactor, and typically occurs in a sequence of steps at different levels. Small-scale reactors are used for screening to determine the optimal catalyst formulation. Since catalyst development and sequential screening is a slow and cost-intensive process, *fast-throughput techniques* [493–496] which permit parallel testing of small amounts of catalyst in microreactor systems are presently attracting great interest and are being developed with significant emphasis. The preparation of new or improved catalyst formulations is mainly empirical but can be aided by knowledge-based (expert) systems or molecular design [497–499].

Small-scale reactors are also used for establishing reaction networks and for determination of the reaction macrokinetics and selectivities, which are required for the optimization of the overall process. These studies are preferentially carried out under ideal conditions. The requirements for such *ideal reactors* are (1) good contact between reactants and catalyst; (2) no limitations caused by heat and mass transport inside and outside the catalyst particles; and (3) good description of reactor characteristics. These conditions can be achieved by plug flow or well-stirred tank reactors (see Section 8.1.1) which must guarantee well-defined residence times and residence-time distributions under isothermal conditions. Channeling of the fluid partially bypassing catalyst particles and uneven flow distributions must be avoided. In three-phase reactors, even distributions of gaseous and liquid reactants over the solid catalyst must be ensured.

Once a promising catalyst is found and reaction networks and macrokinetics of the target reaction are known, testing typically is continued in larger scale (bench and/or pilot) reactors for determination of catalyst lifetime and of the influence of process variables such as temperature, pressure, composition, and impurities in the reactant feed. Ultimately the scale-up to commercial plant dimensions will be performed [499].

Approximate dimensions of tubular reactors (see Section 8.1.1) for various applications on different scale are summarized in Table 22.

8.1.1. Classification of Reactors [113], [488], [490]

Catalytic reactors can be classified by their mode of operation under steady-state or transient conditions or on their mode of contacting/mixing of reactants and solid catalyst.

Typical *steady-state reactors* are packed-bed tubular reactors under continuous flow conditions, either plug flow or mixed flow. In an ideal tubular reactor ideal mixing takes place in the

Table 22. Approximate dimensions of tubular catalytic reactors

Scale	Diameter	Length	Mass of catalyst
Laboratory Microreactors	0.5 cm	0.5 – 1 cm	0.1 – 1 g
Laboratory Bench Scale Reactors	1.0 – 3.0 cm	10 – 30 cm	10 – 200 g
Pilot	5 – 7.5 cm	60 – 100 cm	50 – 100 kg
Plant	0.5 – 5 m	5 – 20 m	10 – 100 tons

radial direction, but there is no mixing in the axial direction. Plug flow is attained in this case. *Plug flow reactors* can be operated either in integral or differential mode. In the latter case, single-pass experiments in small-scale reactors provide the data for differential conditions required for analysis of the reaction kinetics. As an alternative, the effluent from the differentially operating reactor can also be recycled externally or internally, thus approaching a well-mixed reactor system, the continuous-flow *stirred tank reactor* (CSTR). Without inlet and outlet feed a continuous recycle flow results, characteristic of a *batch reactor* in which the feed composition changes with time (transient conditions as opposed to steady-state conditions). The catalyst must not necessarily be kept in a packed bed but can be suspended in the liquid or gaseous fluid reactant mixture. In the fluidized-bed mode, the solid catalyst consisting of fine powder (particle diameter 10 – 200 µm) is kept in motion by an upward gas flow (*fluidized-bed reactor*, see Section 8.1.3). If the fluid is a liquid the catalyst can be suspended easily in a CSTR by efficient stirring (*slurry reactor*, see Section 8.1.3). In so-called *riser reactors*, catalyst material is continuously introduced into and removed from the reactor with the reactant and product feeds.

Batch and semi-batch reactors operate under transient conditions. Discontinuous step or pulse operation necessarily results in transient conditions.

Several catalytic reactor types are schematically shown in Figure 27. *Tubular fixed bed reactors* (A) have an inlet flow $\dot{n}_0$ and an outlet flow $\dot{n}$ of the reactant-product mixture. The *adiabatic fixed-bed reactor* is shown in Figure 27 B. *Multitube fixed-bed reactors* (C) are used for highly exothermic reactions such as the oxidation of *o*-xylene to phthalic anhydride. The principle of a CSTR is demonstrated as D. A *fluidized-bed reactor* with catalyst recirculation is sketched in E. A *slurry CSTR reactor*is schematically shown in F. An alternative for three-phase reactions (gas – liquid – solid) is the *packed bubble column* or *slurry reactor* (G). Discontinuous batch reactors with internal and external recycling operating under transient conditions are depicted as H and I, respectively.

8.1.2. Laboratory Reactors [488], [489]

A compilation of laboratory reactors and their operation mode is given in Figure 28.

As indicated in Table 22, laboratory reactors are small-scale reactors. Steady-state fixed-bed tubular flow reactors are most frequently used for catalyst testing and determination of the reaction kinetics on the laboratory scale. These reactors are favored because of the small amounts of catalyst required, the ease of operation, and the low cost. Stirred tank reactors (e.g., with fixed or rotating basket) and batch suspension reactors are less frequently used. Provided that heat- and mass-transfer limitations can be neglected under the operating conditions in a tubular flow reactor, the catalyst bed is isothermal, and the pressure drop across the catalyst bed is negligible, the reaction rate can be determined from the mass balance for a component *i*. This is usually verified in microreactors operating under differential conditions. The *reaction rate* per unit mass r_W (mol s^{-1} kg^{-1}) is then given by Equation (59):

$$\frac{dx_i}{d\left(\frac{W}{F_i^0}\right)} = -\nu_i r_W, \tag{59}$$

where x_i is the conversion of component *i*, ν_i its stoichiometric coefficient, *W* the catalyst mass (kg), and $F_i{}^o$ the molar rate of component *i* at

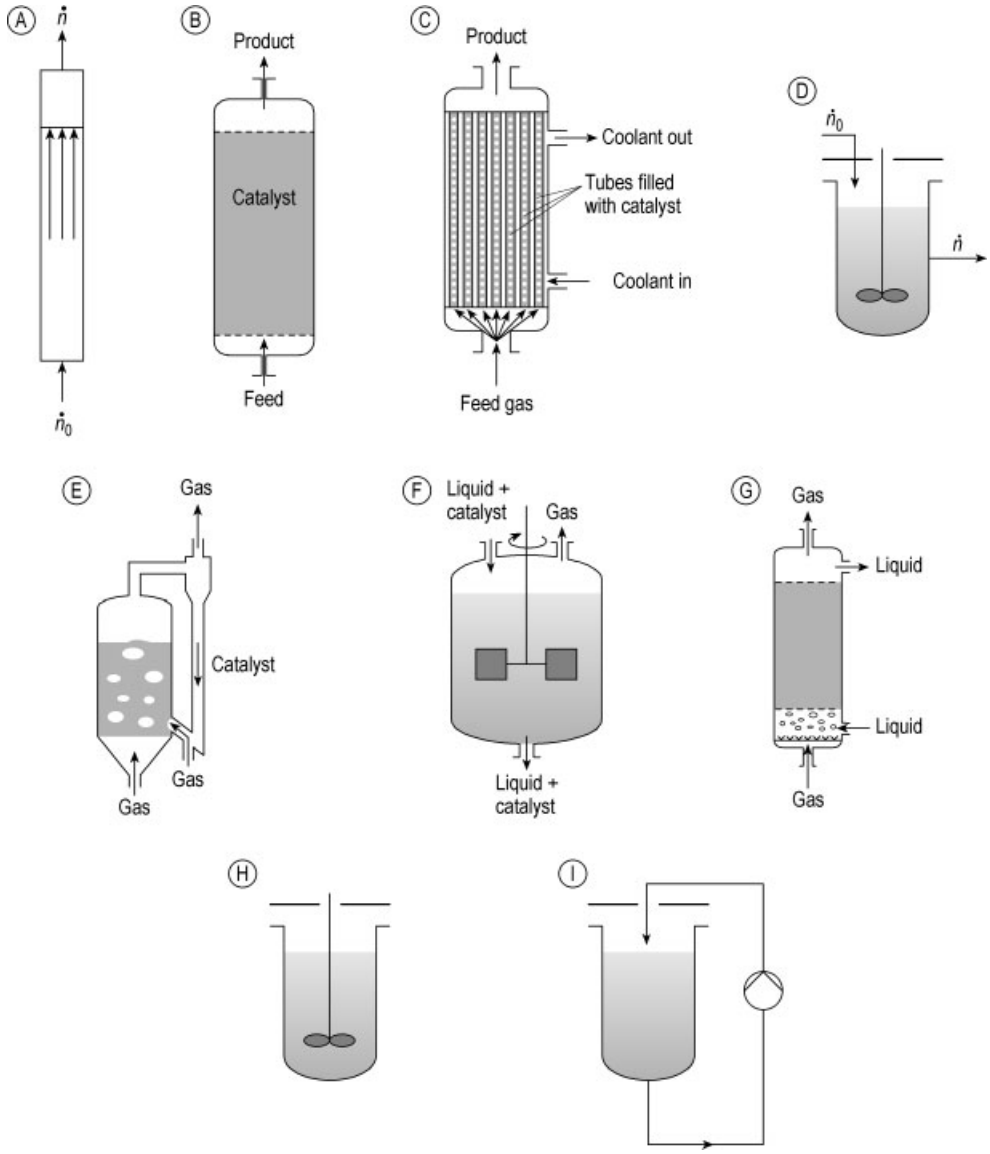

Figure 27. Schematic representation of several types of catalytic reactors
A) Tubular fixed bed reactor; B) Adiabatic fixed-bed reactor; C) Multitube fixed-bed reactor; D) CSTR; E) Fluidized-bed reactor with catalyst recirculation; F) Slurry CSTR reactor; G) Packed bubble column or slurry reactor; H), I) Discontinuous batch reactors

the reactor inlet (mol s^{-1}). The ratio W/F_i° is the space – time.

Ancillary techniques in laboratory units for catalyst testing such as generation of feed streams, product sampling are discussed in [501].

A valuable laboratory pulse reactor (transient operation) is the *TAP reactor* (TAP = *temporal analysis of products*). Pulses containing small amounts of reactants (10^{13} – 10^{17} molecules/pulse) are injected into the evacuated reactor containing the catalyst bed. The reactant/product molecules leaving the reactor (response signal) are analyzed by mass spectroscopy with a time resolution $< 100\,\mu$s. This approach permits the study of surface pro-

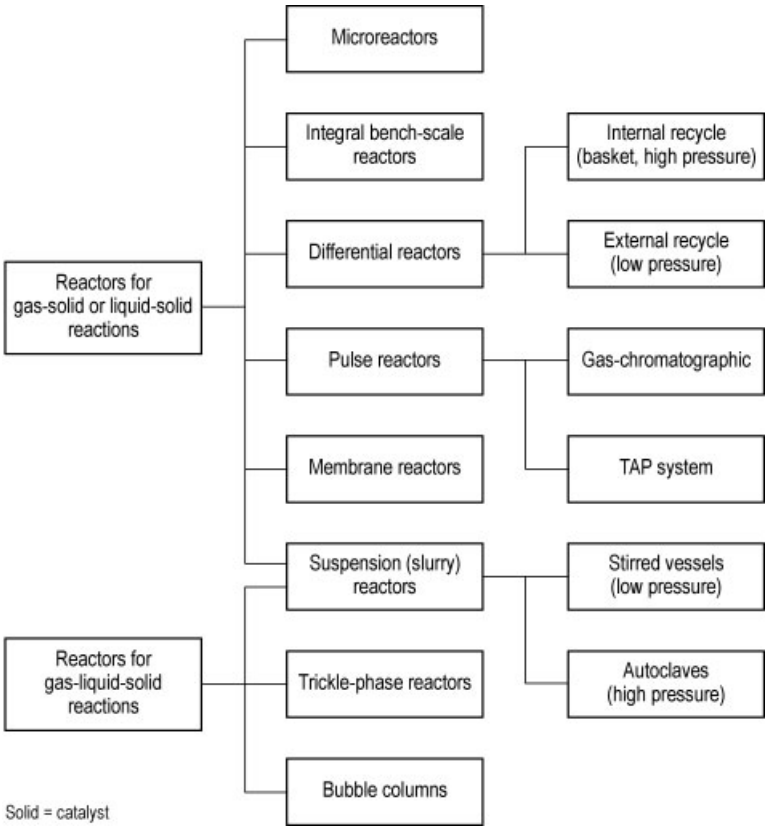

Figure 28. Classification of laboratory reactors

cesses on solid catalysts such as adsorption, reaction, and desorption, and reaction mechanisms and kinetic models can be established [502–504].

8.1.3. Industrial Reactors [113], [490]

Various types of industrial reactors and their operation mode are listed in Figure 29.

Catalytic Fixed-Bed Reactors [505–507]. In the chemical industry fixed-bed reactors are the standard type of reactors for heterogeneously catalyzed gas-phase reactions (two-phase reactors). Fixed catalyst beds can be realized in various ways. Randomly packed beds (deep beds) require catalyst particles having different shapes such as spheres, cylinders, rings, flat disk pellets, or crushed material of a defined sieve fraction. The geometries and dimensions of the catalyst particles are dictated by pressure drop and heat- and mass-transfer considerations. The use of monolithic catalysts significantly reduces the pressure drop across the catalyst bed. This type of catalyst is employed for example for automobile exhaust gas purification and for the removal of nitrogen oxides from tail gases of power stations.

Fixed-bed reactors can be operated under adiabatic or nonadiabatic conditions. Adiabatic reactors can be applied for reactions with low heats of reaction such as gas purification. They consist of a cylindrical tube in which the catalyst is packed on a screen and is traversed in axial direction (Fig. 27 B). This design is particularly suitable when short residence times and high temperatures are required. In this case a fixed bed of large diameter and small height (5 – 20 mm) is used (shallow bed). As an example, for ammonia oxidation in nitric acid plants the fixed bed consists of several layers of platinum wire gauze

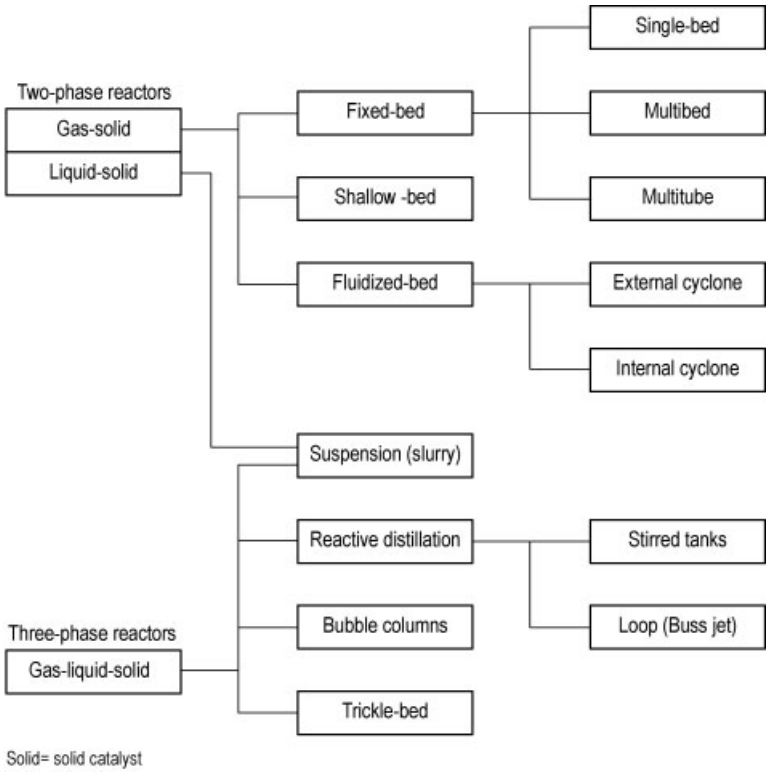

Figure 29. Classification of industrial reactors

with bed diameters up to several meters. This type of reactor is limited to small catalyst volumes. The radial flow concept is preferred when large amounts of catalyst are required [505]. In this reactor type, the catalyst is charged in the annular space around an axially located tube. The reactants are traversing radially, either from the inside or from the outside of perforated plate rings. Because of the low pressure drop, smaller catalyst pellets (4×4 or 3×3 mm) can be used in this reactor type.

Only limited conversions can be achieved by adiabatic reactors because of the necessary control of the adiabatic temperature change. Multistage reactors consisting of several sequential adiabatic stages which are separated by interstage heat exchangers have therefore been introduced.

Nonadiabatic operation can be achieved with fixed-bed reactors which are cooled or heated through the reactor walls. Efficient heat exchange results in so-called isothermal reactors. A typical example is the multitubular reactor schematically shown in Fig. 27 C, which is used for highly exothermic and temperature sensitive reactions (e.g., *o*-xylene oxidation) with, e.g., salt melts as heat-transfer media.

Autothermal reactors can favorably be applied for exothermic and temperature sensitive reaction systems. The conventional reactor design consists of an adiabatic packed-bed reactor coupled with a countercurrent heat exchanger in which the cold reactant feed is brought to reaction temperature.

Multifunctional reactors are being developed [508] with the goal of improving operation conditions which are not necessarily optimally determined in standard fixed-bed reactor configurations.

The following industrial processes are performed in various types of fixed-bed reactors:

i) "deep-bed" adiabatic system:
   – Isomerization of $C_4 - C_6$ alkanes or of light gasoline at 620 – 770 K, 20 – 40 bar on Pt-alumina (HCI activated) or on Pt-H-mordenite.

- Catalytic reforming of heavy gasoline using a cascade of single bed reactors at 700 – 820 K, 20 – 25 bar on K-promoted Cr_2O_3-Al_2O_3 catalyst.
- Hydrocracking of vacuum gas oil at 670 – 770 K, 20 – 40 bar, using single or two stage processes on Ni-MoO_3-Y-zeolite-alumina and Pt-mordenite-alumina, respectively.

ii) "multi-bed" adiabatic system:
- Ammonia synthesis at 670 – 770 K, 200 – 300 bar on K-, Mg-, Al-promoted iron catalysts.
- Oxidation of SO_2 in the sulfuric acid production at 720 – 770 K, atmospheric pressure on K_2SO_4-V_2O_5 catalysts. Reactor with externally located heat exchangers are in operation.

iii) "radial flow" system:
- Water gas shift (HTS) at 620 – 670 K, 25 – 50 bar on Cr_2O_3-Fe_2O_3 catalysts.

iv) "shallow-bed" system:
- Methanol dehydrogenation to formaldehyde at 870 K, atmospheric pressure on metallic Ag (granulate).
- Ammonia oxidation to NO_x at 1170 K atmospheric pressure on Pt/Rh-grids.

v) "quench" system:
- Methanol synthesis at 490 – 520 K, 50 – 100 bar on Cu-ZnO-Al_2O_3 catalysts.

vi) "multi-tube" system:
- Methanol synthesis at 490 – 520 K, 50 – 100 bar on Cu-ZnO-Al_2O_3 catalysts.
- Oxidation of ethylene to ethylene oxide at 470 – 520 K, atmospheric pressure on Ag-α-alumina.
- Oxidation of *o*-xylene to phthalic anhydride at 640 – 680 K, at atmospheric pressure on V_2O_5-TiO_2 catalysts.
- Hydrogenation of benzene to cyclohexane at 470 – 520 K, 35 bar on Ni-SiO_2 catalysts.
- Dehydrogenation of ethylbenzene to styrene at 770 – 870 K, atmospheric or reduced pressure on promoted (K, Ce, Mo) Fe-oxide.

Fluidized Bed Reactors [509–512].
Fluidized-bed reactors are preferred over fixed-bed reactors if rapid catalyst deactivation occurs or operation in the explosive regime is required. In this type of reactor, an initially stationary bed of catalyst is brought to a fluidized-state by an upward stream of gas or liquid when the volume flow rate of the fluid exceeds a certain limiting value, the minimum fluidization volume flow rate. The catalyst particles are held suspended in the fluid stream at this or higher flow rates. The pressure drop of fluid passing through the fluidized bed is equal to the difference between the weight of the solid catalyst particles and the buoyancy divided by the cross-sectional area of the bed. Major advantages of fluidized-bed reactors are excellent gas – solid contact, good gas – solid heat and mass transfer, and high bed-wall heat transfer coefficients.

The gas distribution in the fluidized bed is performed industrially by, for example, perforated plates, nozzles, or bubble caps which are mounted at the reactor bottom.

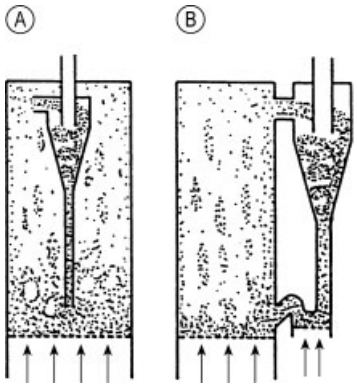

Figure 30. Schematic representation of two forms of gas – solid fluidized beds under turbulent flow conditions

Depending on the volume flow rate of the fluid different types of fluidized beds form. Fluidization with a liquid feed leads to a uniform expansion of the bed. In contrast, solid-free bubbles form when fluidization is carried out in a gas stream. These bubbles move upwards and tend to coalesce to larger bubbles as they reach increasing heights in the bed. At high gas volume flow rates, solid particles are carried out of the bed. To maintain steady-state operation of such a turbulent fluidized bed, the solid catalyst particles entrained in the fluidizing gas must be collected and transported back to the reactor bed. This can be achieved most easily with an integrated cyclone, as schematically shown in Figure 30 A. A circulating fluidized bed is finally formed at still higher gas volume rates. An efficient external

recycle system, as shown in Figure 30 B, is required for such operating conditions because of the high solids entrainment.

Catalytic cracking is carried out in fluidized-bed reactors because the solid acid catalysts are rapidly deactivated by coke deposition. The catalyst must therefore continuously be discharged from the reactor and regenerated in an air-fluidized regenerator bed where the coke is burned off. The regenerated catalyst is then returned to the fluidized-bed reactor. The heat of combustion of the coke can be used for preheating of the reactant feed.

The main advantages of *fluidized-bed* reactors are:

- Uniform temperature distribution (due to intensive solid mixing),
- large solid-gas exchange area and
- high heat-transfer coefficient between bed and immersed heating or cooling surfaces.

However, these reactors have also some disadvantages, e.g.:

- expensive catalyst separation and purification of reaction products (installation of cyclones and filters),
- undesired bypass of reactants due to bubble development,
- catalyst attrition and
- erosion of internals resulting from high solids velocities.

Fluidized-bed reactors found very broad application in the petroleum refining and production of chemicals, for example:

- Catalytic cracking of vacuum gas oil to gasoline at 720 – 820 K on alumosilicates containing ultrastable Y-zeolites.
- Fischer – Tropsch synthesis (Synthol process) from CO und H_2 at 620 – 670 K and 15 – 30 bar on promoted Fe-oxide catalysts.
- Ammoxidation of propylene to acrylonitrile (SOHIO-process) at 670 – 770 K, 1 – 2 bar on promoted Bi-MoO_x catalysts.
- Oxidation of naphthalene or *o*-xylene to phtalic anhydride at 620 – 650 K, at atmospheric pressure on V_2O_5–SiO_2 catalysts.

A special type of the fluidized-bed reactor is the so-called "riser reactor". This reactor consists of a vertical tube in which the reaction takes place in the presence of the entrained catalyst. Catalyst coming from the riser tube is collected in the vessel, before passing through the stripper to the regenerator (fluidized-bed type). The "riser reactor" is mainly used in the catalytic cracking of heavy oils on highly active zeolitic catalysts.

The *moving-bed* reactor [509] operates with spherical catalyst particles larger (2 – 6 mm) than those used in the fluid-bed system. In the standard arrangement, catalyst particles are moving slowly through the agitated bed. Catalyst reaching the reactor top is transported into the regenerator. Using mechanical or pneumatic conveyer the regenerated catalyst is returning to the bottom of the reactor.

The main advantage of the moving-bed reactor is a lower catalyst attrition than in the fluidized-bed system. The disadvantage is a poor heat transfer and therefore this reactor is not suitable for exothermic reactions.

The moving-bed reactor found application mainly in petroleum cracking.

Slurry Reactors [513]. The aim of slurry reactors is intimate contact between a gas phase component (which is to be dissolved in a liquid-phase component) and a finely dispersed solid catalyst (three-phase reactors). The particle size of the solid catalyst is kept sufficiently small ($< 200\,\mu m$) that it remains suspended by the turbulence of the liquid in the slurry reactor. This is in contrast to three-phase fluidized-bed reactors, in which an upward liquid flow is required to suspend the larger catalyst particles.

On the basis of the contacting pattern of the phases and the mechanical devices that achieve contact and the mass transfer, nine types of slurry reactors can be distinguished [513]:

1) Slurry bubble column
2) Countercurrent column
3) Cocurrent upflow
4) Cocurrent downflow
5) Stirred vessel
6) Draft-tube reactor
7) Tray column
8) Rotating-disk or multiagitator column
9) Three-phase spray column

Slurry reactors are industrially applied for a multitude of processes [514], many of which are heterogeneously catalyzed processes for hydrogenation of edible oils. A new development is

the continuous Fischer – Tropsch slurry synthesis process of SASOL in South Africa [515].

Three-phase reactors are classified as fixed-bed or suspension reactors depending on the catalyst arrangement and shape:

Fixed-bed reactors operate either in the "trickle-bed" or in the "bubble-flow" mode [513]. In the first case, liquid reactants or reactants dissolved in a solvent are flowing downward through the catalyst bed and the gaseous reactants are conducted in the countercurrent or concurrent direction.

In the "bubble-flow" reactors, liquid and gaseous reactants are fed into the bottom of the column and are flowing upward through the catalyst fixed-bed.

The "trickle-bed" arrangement has some advantages such as:

- Fast diffusion of gases through the liquid film to the catalyst surface,
- lower back-mixing,
- no problems with catalyst separation,
- selective removal of catalytic poison in the entrance zone of the bed and
- simple catalyst regeneration.

Drawbacks of trickle-bed reactors are:

- Poor heat transfer,
- partial utilization of the catalyst in case of the incomplete wetting and
- possibility of so-called "brooks" formation.

The successful performance of the trickle-phase reactors depends on the suitable diameter/length ratio, catalyst shape and size, and the liquid flow distribution through the catalyst bed. The catalyst particle size is limited by the allowed pressure drop. Larger sizes (6 – 10 mm diameter) are therefore preferred, which, however, can bring diffusion problems.

The "bubble flow" version is favored in particular for reactions with a low space velocity. A good heat transfer and no problems with an incomplete catalyst wetting are the main advantages of the "bubble flow" reactors.

Both types of the three-phase reactors have found numerous industrial applications, e.g.:

- Hydrotreating of petroleum fractions at 570 – 620 K, 30 – 60 bar on Ni-MoO_3-Al_2O_3 catalysts.
- Hydrocracking of high boiling distillates at 570 – 670 K, 200 – 220 bar on Ni-MoO_3-Y-zeolite-alumina catalysts.
- Selective hydogenation of C_4-fractions (removal of dienes and alkines) at 300 – 325 K, 5 – 20 bar on Pd-Al_2O_3 catalysts.
- Hydrogenation of aliphatic carbonyl compounds to alcohols at 370 – 420 K, 30 bar on CuO-Cr_2O_3 catalysts.

Suspension reactors [513], [514] are operated successfully in the chemical industry because of their good heat transfer, temperature control, catalyst utilization, and simple design.

Because of the small catalyst particle size, there are no problems with internal diffusion. *Suspension reactors* operate either in the discontinuous or in the continuos mode.

One serious disadvantage is the difficult catalyst separation, especially if fine particles have to be removed from the viscous liquid.

Currently in use are two types of suspension reactors: "stirred vessels" and "three-phase bubble columns".

In the case of "stirred vessels" the catalyst particles (mainly below 200 μ) are suspended in the liquid reactant or solutions of reactants, whereas gaseous reactants are introduced at the bottom of the vessel through perforated tubes, plates or nozzles. The vessels are equipped with different types of stirrers or turbines keeping the catalyst suspended. Cooling and heating coils as well as the gas recycle system belong to the standard equipment. "Stirred vessels" operate mostly discontinously. However, if continuous operation is favored, then "stirred vessels" are arranged in a cascade to complete the required conversion.

The "bubble column" represents mainly continuously operating three-phase reactors. The gas is introduced at the bottom of the column through nozzles, perforated plates or tubes. In the standard arrangement the liquid reactant flows in the cocurrent direction with the gas. In some cases stirrers are installed to keep powdered catalysts in the suspension.

The gaseous reactants can be recycled by an external loop or by an internal system, such as Ventury jet tube. This equipment is driven by a recycle of the slurry using a simple pump. The Ventury tube is sucking the gas from the free board above the reactor back into the slurry. Heat exchangers can be installed in the loop in both cases.

The main advantages of "bubble columns" are simple and low priced construction, good heat transfer, and temperature control.

Suspension reactors are used predominantly for fat and oil hydrogenation 420 – 470 K, at 5 – 15 bar using various Ni-kieselguhr catalysts.

Also, hydrogenolysis of fatty acid methylesters to fatty alcohols is performed in suspension reactors at 450 – 490 K, 200 – 300 bar on promoted copper chromites ($CuO–Cr_2O_3$).

Further applications are: syntheses of acetaldehyde and of vinylacetate according to the Wacker process.

8.1.4. Special Reactor Types and Processes

Microreactors [516–518]. *Microreactors* (→ Microreators) are used for laboratory testing of catalysts and for fast-throughput techniques, as mentioned in Section 8.1. In addition, microreactor technology can also be applied in synthetic chemistry. A microreactor can be defined as a series of interconnecting channels having diameters between 10 and 300 μm which are formed in a planar surface in which small quantities of reagents are manipulated. Among the advantages of microreactors over conventional catalytic reactors are high heat-transfer coefficients, increased surface to volume ratios of 10 000 to 50 000 $m^2\ m^{-3}$, as opposed to 1000 $m^2\ m^{-3}$ for conventional catalytic laboratory reactors, shorter mixing times, and localized control of concentration gradients. The small scales reduce exposure to toxic or hazardous materials, and the enclosed nature of the microreactors permits greater ease of containment in the event of a runaway reaction. Furthermore, the rapid heat transfer from the area of reaction is fast enough to prevent explosion even when reactions are carried out beyond the explosion limit [518].

Despite their small size, microreactors can be used for synthetic chemistry [519], since as few as 1000 microreactors operating continuously could produce 1 kg of product per day. The first chemical synthesizers are already commercially available. Further developments are still required to make microreactors mature for industrial application, although their use in the production of speciality chemicals and pharmaceuticals seems to be feasible.

Unsteady-State Reactor Operation [520]. Forced unsteady-state reactor operation has been applied to continuous catalytic processes in fixed and f1uidized-bed reactors. This operating mode can lead to improved reactor performance. Nonlinearity of chemical reaction rates and complexity of reaction systems are responsible for conversion or selectivity improvements under forced unsteady-state conditions [521].

An unsteady-state in a fixed-bed reactor can be created by oscillations in the inlet composition or temperature (control function) such as schematically shown in Figure 31. The most widely applied technique in a fixed-bed reactor is periodic flow reversal. Examples of this operation mode in industrial applications are SO_2-oxidation, NO_x-reduction by NH_3, and oxidation of volatile organic compounds (VOC). In fluidized-beds for exothermic reactions, favorable unsteady-state conditions of the catalyst can be achieved by catalyst circulation inside the reactor. The unsteady-state operation in the fluidized-bed is applied for example for the partial oxidation of *n*-butane to maleic anhydride on vanadyl pyrophospate catalysts developed by DuPont. In the first step, *n*-butane diluted with an inert gas is contacted in the riser reactor (residence time 10 – 30 sec) with spherical catalyst particles (100 μ). The partial oxidation of n-butane proceeds on account of the lattice oxygen in the surface layer of vanadyl pyrophosphate. In the second step, the partially reduced catalyst is transported into the fluidized-bed reactor where the catalyst reoxidation takes place. The obtained selectivity was about 10 % higher than that in the multitubular fixed-bed reactor, operating under steady-state conditions. A further group of forced unsteady-state processes uses the combination of a chemical reaction with the separation of products (chromatographic reactor). Systems applied till today operating on the principle of chromatographic columns are filled with a catalyst possessing suitable adsorption properties, such as $Pt-Al_2O_3$. Pulses of reactant are periodically injected into this reactor which is purged by carrier gas during periods between the pulses. This operation can provide a higher conversion for reversible reactions if one of the reaction products is adsorbed on the

catalyst more strongly than the other one. The feasibility of this principle was tested in the dehydrogenation of cyclohexane to benzene on pilot scale.

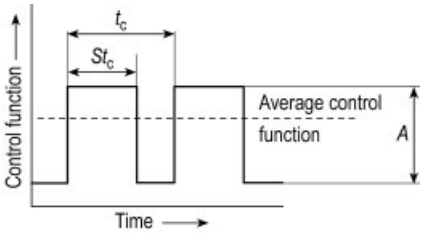

Figure 31. Stepwise variation of inlet parameters to create unsteady state conditions

Membrane Reactors [522–524]. In membrane reactors (→ Membrane Reactors) catalytic conversion is coupled with a separation effect provided by the integrated membrane. High conversions can be achieved for equilibrium-restricted reactions when one of the reaction products can be removed from the reaction mixture by diffusion through a permselective membrane. Hydrocarbon dehydrogenation reactions are the best example for this application. In addition, catalytic membrane reactors have been proposed for improved control of the selectivity of some catalytic reactions. For example, controlled introduction of a reactant into the reactor by selective or preferential permeation may limit possible secondary reactions of the target product.

Two types of membranes can be distinguished: dense membranes and porous membranes. Dense metallic membranes consist of thin metal foils. Palladium and palladium alloys (PdAg, PdRu) are specific for hydrogen permeation. Dense oxide membranes are usually solid electrolytes such as ZrO_2 or CeO_2, which are permeable for O_2 [524]. Porous membranes are typically made of oxides, although carbon membranes have also been used [525]. Ceramic membranes consist of several layers of material with progressively decreasing pore size. The top layer with the smallest pore size controls the separation. Most of these membranes are produced by sol – gel techniques. Intrinsic catalytic properties can be introduced into these membranes, which can be produced as cylindrical tubes forming the basis for tubular reactors. Zeolite membranes can also be prepared, and the structure of the zeolite should permit high selectivity in separation processes.

Despite the potential of membrane reactors, their development is still not mature for industrial application.

Reactive Distillation [526–528]. In reactive distillation (→ Reactive Distillation) fractional distillation and chemical reaction are performed simultaneously, e.g., for a reaction of the type

$$A + B \longrightarrow C + D$$

in which at least one of the products has a volatility which is different form those of the other compounds. The most attractive features of reactive distillation are:

1) The separation of at least one of the products by distillation drives reactions to completion which are otherwise equilibrium-limited.
2) Reactions in which high concentrations of a product or one of the reactants lead to undesired side reaction, can be carried out.

Despite these advantages, several chemical and physical limitations exist in practice for its use in chemical processes, so that reactive distillation has found application for only a few important reactions.

Industrially important processes are various etherifications, esterifications, alkylations or isomerizations.

IFP, Mobil, Neste Oy, Snamprogetti, Texaco, and UOP are providing licenses for plants to produce t-butyl- or t-amyl ethers from corresponding olefins and methanol or ethanol using strong acidic resins as solid catalyst.

Reactions under Supercritical Conditions [529–532]. Only in the last decade have reactions under supercritical conditions gained importance in chemical technology because of the relatively high pressures and the rigorous demands on the equipment usually required when working in supercritical fluids. The major advantage of these media as solvents in catalytic reactions is the fact that carbon dioxide and water can be used as environmentally benign solvents. When in their supercritical state, these nontoxic compounds are good solvents for many organic compounds. A multicomponent system under supercritical conditions may behave like a

single gaslike phase with advantageous physical properties. Under reaction conditions this leads to [529]:

1) Higher reactant concentrations
2) Elimination of contact problems and diffusion limitations in multiphase reaction systems
3) Easier separations and downstream processing
4) In situ extraction of coke precursors
5) Strongly pressure and temperature dependent solvent properties near the critical point
6) Higher diffusivities than in liquid solvents
7) Better heat transfer than in gases
8) Use of clustering to alter selectivities

Hence, supercritical fluids offer strategies for more economical and environmentally benign process design, mainly because of enhanced reaction rates, prolongation of catalyst lifetime, and simplification of downstream processing.

8.2. Filling Fixed-Bed Reactors

In filling the reactor, catalyst distribution and particle size must be uniform to avoid channeling, leading to poor heat transfer and impairment of catalyst performance. In a *multitube reactor*, the pressure drop must be the same in each tube so that the gas flow is distributed uniformly over the tubes. Small changes in the packing density in the tubes can cause uneven heat transfer and, in the case of highly exothermic reactions, hot spots and selectivity loss. For these reasons multitube reactors are filled by special equipment that charges each tube with the same amount of catalyst at a definite rate. After filling, the tubes are checked for pressure drop, and if necessary, the charge is adjusted. Satisfactory heat transfer and uniform gas flow can only be accomplished if the reactor tube diameter is approximately 8 – 10 times, and the reactor length 50 – 200 times, as large as the diameter of the catalyst particle. For reactions occurring at atmospheric pressure, the catalyst size should not exceed 1 cm to avoid channeling and to ensure full utilization of catalyst activity.

Pyrophoric catalysts are charged from special containers filled with inert gas and provided with an adapter for displacing air from the reactor or reactor tubes. Corresponding safety measures are also taken on discharging air-sensitive catalysts from reactors. However, passivated catalysts and catalysts embedded in, or coated with, protective, high-binding material (e.g., hardened edible oil for hydrogenation catalysts) can be handled without precautionary measures.

8.3. Reactor Start-Up

During the initial days of operation of many heterogeneous catalytic processes, especially in selective oxidation reactions, a fresh catalyst undergoes minor chemical and physical changes (oxidation, reduction, sintering, etc.) before it reaches a steady state. During this time the reactor is operated according to a prescribed program at reduced throughput to avoid damage to the catalyst by heat evolution in case unexpectedly high activity should develop in the start-up period. When using nickel-containing catalysts below 420 K, carbon monoxide must be completely excluded from the reactants to prevent the formation of highly toxic nickel carbonyl and the destruction of the catalyst.

8.4. Diffusion, Mass and Heat Transfer Effects [63], [533], [534]

Heterogeneous catalytic reactions take place on the external and internal surfaces (pores) of the catalyst. *External and internal concentration* and *temperature gradients* can build up in the fluid (gas or liquid) boundary layer around the catalyst particles and inside the pores if the mass and heat transfer between the bulk of the fluid and the active surfaces is not sufficiently fast. Such gradients tend to exist (1) in fixed-bed reactors charged with large, porous catalysts; (2) in operation at low-mass flow velocities (mass flow rate per unit cross-sectional area); or (3) in the case of highly exothermic reactions. On the other hand, gradient effects usually are absent when nonporous catalysts are used in fluidized-bed reactors or in fixed-bed reactors at high mass velocities.

Because such internal and external gradients can substantially reduce the activity and the selectivity of the catalyst, conditions have been delineated under which their adverse effects can

be minimized [63], [533], [534]. By carefully matching operating conditions, catalyst, and reactor, optimum catalyst performance can be ensured.

Mass and heat transfer in heterogeneous catalytic reactions occurs in two ways. *External transport* to the external surface involves diffusion through the more or less stationary hydrodynamic boundary layer that surrounds the catalyst particle. The thickness of this layer depends on the characteristics of the fluid and its flow rate past the particle, and affects the rate of mass and heat transfer. The *internal transport* to the stationary fluid in the pores of catalyst particle is controlled by diffusion alone.

Depending on the relative rates of the transport processes and of the catalytic reaction, three or four types of regimes can be distinguished. In the *kinetic regime* the rates of external and internal mass transport are much higher than the rate of the chemical reaction. Therefore, concentration and temperature gradients between the fluid and the center of a catalyst particle are negligible, and the catalyst is fully utilized.

In the *internal diffusion regime*, mass transport in the catalyst pores is about as fast as, or slower than, the chemical reaction. In this case there are considerable concentration gradients along the length of the pores, the effectiveness of the catalyst is impaired, and the apparent energy of activation is lower than that observed in the kinetic regime.

At a still lower ratio of rates of transport and conversion, there is an *intermediate regime* in which the reaction takes place only on the external surface of the catalyst particles while the internal surface of the pore is inactive. Because of the limited heat transfer in this regime, exothermic reactions can overheat the catalyst, resulting in a higher activity than that corresponding to the temperature of the fluid.

Finally, on further decrease of the ratio of the rates of transport and conversion (e.g., by raising the temperature of the fluid), the *external mass transfer regime* is reached in which the reaction rate is controlled by mass transfer, and the concentration of the reactants at the surface of the catalyst particles drops. Raising the reactor temperature in this regime has little effect on the reaction rate, and the apparent activation energy drops.

8.4.1. Effectiveness Factor [63], [533], [534]

The *effectiveness factor* η is the ratio of the actual reaction rate observed on a porous catalyst particle to the rate that would be obtained if the inside of the particle were exposed to the temperature and reactant concentrations of the fluid. Mathematical analysis [63], [535–544] of mass transfer in porous particles of different shapes has shown that effectiveness is a function of a dimensionless quantity, called the *Thiele diffusion modulus* φ [535]: for a sphere of radius R and for a plate sealed on one side and on the edges the thickness of which is L, φ is defined by the following equations:

$$\text{sphere}: \varphi_{\mathrm{s}} = R\left(\frac{k_{\mathrm{v}}\, c_{\mathrm{s}}^{m-1}}{D_{\mathrm{eff}}}\right)^{1/2} \tag{60}$$

$$\text{plate}: \varphi_{\mathrm{L}} = L\left(\frac{k_{\mathrm{v}}\, c_{\mathrm{s}}^{m-1}}{D_{\mathrm{eff}}}\right)^{1/2}$$

where k_{v} is the reaction rate constant per unit of gross catalyst volume, c_{s} is the concentration on the surface, m is the reaction order, and D_{eff} the effective diffusion coefficient, given by

$$D_{\mathrm{eff}} = D\frac{\Theta}{\tau} \tag{61}$$

where D is the diffusion coefficient for a pair of fluids, Θ the void fraction of the porous mass, and τ a factor allowing for tortuosity and varying cross-sections of the pores.

For first-order reactions ($m = 1$), the effectiveness factors are as follows (tan is hyperbolic tangent):

$$\text{sphere}: \eta = \frac{3}{\varphi}\left(\frac{1}{\tan\varphi_{\mathrm{s}}} - \frac{1}{\varphi_{\mathrm{s}}}\right) \tag{62}$$

$$\text{plate}: \eta = \frac{\tan\varphi_{\mathrm{L}}}{\varphi_{\mathrm{L}}}$$

Correlation between the effectiveness factor and the Thiele modulus for nonexothermic reactions is shown in Figure 32 [540]. The effectiveness factor is ca. unity for $\varphi < 1$ and inversely proportional to φ for $\varphi > 3$.

If the intrinsic velocity constant k_{v} (Eqs. 61 and 62) cannot be determined directly, another dimensionless modulus Φ has been derived [538–540]. For first-order reactions occurring in a sphere, it is defined by

$$\Phi \equiv \varphi^2 \eta = \left(\frac{R^2}{D_{\text{eff}} V_{\text{c}} C_{\text{s}}}\right)\left(\frac{\text{d}n}{\text{d}t}\right) \tag{63}$$

where dn/dt is the conversion rate in moles per second of the reactant in the catalyst volume V_{c}. The effectiveness factor as a function of Φ is shown in Figure 33 for a moderate energy of activation ($E = 10\,RT$; first-order reaction in a spherical particle) and variable enthalpy change ΔH (λ is the thermal conductivity of the catalyst). For exothermic reactions ($\beta > 0$), the effectiveness factor goes through a maximum value exceeding unity because of the interaction of two opposing effects. Poor mass transfer lowers the efficiency of the catalyst, whereas insufficient heat transfer raises catalyst temperature and reaction rate.

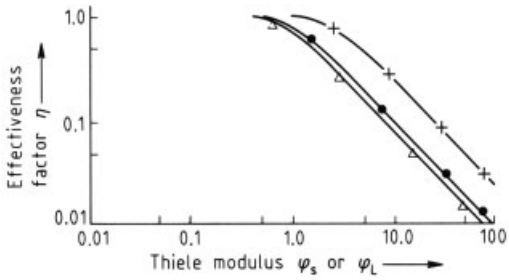

Figure 32. Effectiveness factor φ as a function of the Thiele modulus φ_{s} or σ_{L}
• Flat plate sealed on one side and on edges, first-order reaction; Δ Same, second-order reaction; + Spherical particle, first-order reaction
∗ Reproduced with permission [540]

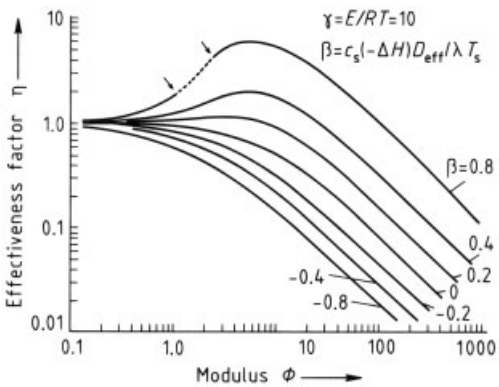

Figure 33. Effectiveness factor as a function of modulus (Eq. 4)
— Unstable region
∗ Reproduced with permission [545]

8.4.2. Effects on Selectivity [63], [533], [534]

The effect of mass and heat transport processes on the selectivity of reactions yielding more than one product depends on the *selectivity type*. In a type **I** reaction at low-effectiveness factors, the observed selectivity factor changes from k_1/k_2 to $(k_1\ D_{\text{Aeff}}\ k_2^{-1}\ D_{\text{Beff}}^{-1})^{1/2}$, provided the order of the two reactions (A → X and B → Y) is the same. If, as usual, the ratio of the effective diffusivities is smaller than k_1/k_2, the selectivity will drop at low effectiveness factors. However, if the diffusivity ratio is high, the selectivity can increase in a porous catalyst. This is the situation in the so-called shape selective zeolites [63]. If the reaction orders are different, the reaction with the lower order is favored in the porous catalyst at low η values.

In reaction type **II**, the effectiveness factor has no influence on the selectivity if the orders of the two reactions are equal. Otherwise, the effect of different orders is the same as in type **II** reactions [534], [545].

For an isothermal first-order reaction of type **III** occurring at low effectiveness factor on a porous plate, the observed selectivity is approximately the square root of the intrinsic rate constant ratio k_1/k_2 [546]. Figure 34 shows comparative conversions of A to X for type **III** at effectiveness factors $\eta = 1$ and $\eta < 0.3$, for $k_1/k_2 = 4$; the maximum yield and selectivity both drop by ca. 50 % at the low effectiveness factor [546].

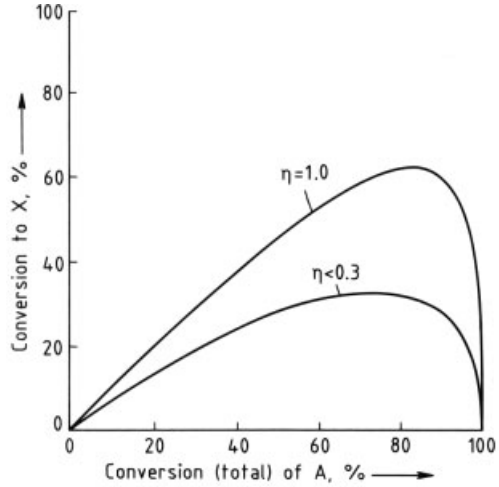

Figure 34. Effect of the effectiveness factor on catalyst selectivity
∗ Reproduced with permission [546]

Temperature gradients caused by exothermic reactions favor reactions with the higher apparent *energies of activation*. Because such undesired side reactions as decomposition and oxidative degradation generally have high energies of activation, large catalysts having narrow pore structure can have an unfavorable effect on the selectivity and product yield.

In the regime of external mass transfer, resistance of the boundary layer to *diffusion* has similar effects on the selectivity of parallel and consecutive reactions as does diffusion in pores.

8.4.3. Catalysts and Transfer Processes

Industrial processes that occur in the kinetic regime include reactions conducted in fluidized-bed reactors using catalysts 0.05 – 0.25 mm in size [63], [534]. Because diffusion coefficients in liquids are smaller by several orders of magnitude than those in the gaseous phase, liquid-phase operation in the kinetic regime requires a finely powdered catalyst.

Because the resistance to flow of catalysts increases steeply with decreasing size, use of catalysts smaller than 2 – 3 mm in fixed-bed reactors is restricted to radial flow, slurry, and riser reactions. Ring-shaped and tableted catalysts show relatively favorable pressure drop and diffusion characteristics.

In various partial oxidation processes (e.g., naphthalene or *o*-xylene to phthalic anhydride), good results have been obtained with so-called *egg shell catalysts* in which the active catalyst mass is applied in a thin layer of a few tenths of a millimeter to the external surface of an inert, nonporous support.

In the regime of *internal diffusion*, it is often advantageous to use catalysts with a bimodal pore system in which micropores (< 2 nm) are connected by macropores (> 50 nm) to the external surface of the catalyst particle. The micropores provide the needed high surface, whereas the macropores facilitate mass transport to and from the micropores. Such catalysts are especially suitable for operation at process pressures below 3 MPa (30 bar). At these pressures mass transport into pores smaller than 100 – 1000 nm occurs increasingly through *Knudsen diffusion* involving molecular collisions with the pore wall rather than between molecules [63], [534].

In fast processes occurring in the regime of *external mass transfer*, the overall conversion rate is independent of the porosity of the catalyst and is essentially proportional to its external surface area [541]. Examples of such reactions are oxidation of ammonia to nitric acid and ammoxidation of methane to hydrogen cyanide carried out at 1070 – 1270 K on stacks of fine mesh Pt – Rh gauze.

9. Catalyst Deactivation and Regeneration

9.1. Different Types of Deactivation

As has been observed in the laboratory and in industrial application, heterogeneous catalysts are deactivated during time on stream. For example, in fluid-bed catalytic cracking and propene ammoxidation the catalyst life is limited to a few seconds or minutes, while in other reactions, such as NH_3 and CO oxidation the catalyst remains active for several years. Not only loss of activity but also a decrease in selectivity is usually caused by catalyst deactivation [410], [547], [548].

The activity loss can be compensated within certain limits by increasing the reaction temperature. However, if such compensation is not efficient enough, the catalyst must be regenerated or replaced.

The main reasons for catalyst deactivation are:

1) Poisoning
2) Fouling
3) Thermal degradation
4) Volatilization of active components

Some types of catalyst deactivation are reversible, e.g., catalyst fouling and some special types of poisoning. Other types of deactivation are mostly irreversible [410], [547], [549].

Catalyst Poisoning. The blocking of active sites by certain elements or compounds accompanied by chemisorption or formation of surface complexes are the main causes of catalyst deactivation [410], [547], [549]. If the chemisorption is weak, reactivation may occur; if it is strong, deactivation results. Chemical species often considered as poisons can be divided in five classes:

1) Group 15 and 16 elements such as As, P, S, Se and Te
2) Metals and ions, e.g., Pb, Hg, Sb, Cd
3) Molecules with free electron pairs that are strongly chemisorbed, e.g., CO, HCN, NO
4) NH_3, H_2O and organic bases, e.g., aliphatic or aromatic amines, pyridine, and quinoline
5) Various compounds which can react with different active sites, e.g., NO, SO_2, SO_3, CO_2.

Elements of the class (1) and their compounds, for example, H_2S, mercaptans, PH_3, and AsH_3, are very strong poisons for metallic catalysts, especially for those containing Ni, Co, Cu, Fe, and noble metals [410], [547], [549].

Elements of the class (2) can form alloys with active metals and deactivate various systems in this way [410], [550].

CO is chemisorbed strongly on Ni or Co and blocks active sites. Below 450 K and at elevated pressure the formation of volatile metal carbonyls is possible [410], [547], and catalyst activity is strongly reduced.

Ammonia, amines, alcohols, and water are well known poisons for acidic catalysts, especially for those based on zeolites [410], [551].

Catalysts or carriers containing alkali metals are sensitive to CO_2, SO_2, and SO_3.

Catalyst Fouling. Because most catalysts and supports are porous, blockage of pores, especially of micropores, by polymeric compounds is a frequent cause of catalyst deactivation. At elevated temperatures (>770 K) such polymers are transformed to black carbonaceous materials generally called *coke* [410], [552]. Catalysts possessing acidic or hydrogenating – dehydrogenating functions are especially sensitive to coking.

There are different types of coke, such as C_α, C_β, carbidic or graphitic coke, and whisker carbon [410]. C_α is atomic carbon formed as a result of hydrocarbon cracking on nickel surfaces above 870 K. C_α carbon can be transformed at higher temperatures to polymeric carbon (C_β) which has a strongly deactivating effect. C_α carbon can also dissolve in metals and forms metal carbides, and it may precipitate at grain boundaries. Metal-dissolved carbon may also initiate the growth of carbon whiskers, which can bear metal particles at their tops.

Coke formation can be minimized, for example, in methane steam reforming by sufficiently high steam/methane ratio or/and by the alkalization of the carrier.

Thermal Degradation. One type of thermal degradation is the agglomeration of small metal crystallites below the melting point, called *sintering* [410]. The rate of sintering increases with increasing temperature. The presence of steam in the feed can accelerate the sintering of metal crystallites.

Another type of thermal degradation are solid-solid reactions occurring especially at higher temperatures (above 970 K). Examples are reactions between metals, such as Cu, Ni, Co, and alumina carriers which result in the formation of *inactive* metal aluminates [410], [553].

Also, phase changes belong to the category of thermal degradations. A prominent example is the reduction of the surface area of alumina from $250\,m^2\,g^{-1}$ (γ phase) to $1-2\,m^2\,g^{-1}$ (α phase) by thermal treatment between 870 and 1270 K.

Alternating oxidation and reduction of the system, as well as temperature fluctuations, are often accompanied by activity losses and can cause mechanical strain in catalyst pellets. Therefore, mechanical strength of catalyst particles has major industrial importance.

Volatization of Active Components. Some catalytic systems containing P_2O_5, MoO_3, Bi_2O_3, etc. lose their activity on heating close to the sublimation point. Cu, Ni, Fe and noble metals can escape from catalysts after conversion to volatile chlorides if traces of chlorine are present in the feed.

9.2. Catalyst Regeneration

The regeneration of metallic catalysts poisoned by Group 16 elements is generally rather difficult. For example, oxidation of sulfur-poisoned metallic catalysts converts metal sulfides to SO_3, which desorbs from metals. However, if the catalyst or carrier contains Al_2O_3, ZnO, MgO then SO_3 forms the corresponding sulfates. When the catalyst is subsequently brought on-line under reducing conditions, then H_2S is formed from sulfates and the catalyst will be repoisoned [410], [554].

Therefore it is necessary to remove poisons from the feed as completely as possible. Further prevention is the installation of a guard-bed containing effective poison adsorbents in front of the reactor.

Ni catalysts poisoned with CO or HCN can be regenerated by H_2 treatment at temperatures that allow formation of methane and NH_3, respectively.

The original activity of acidic catalysts poisoned partially by H_2O, alcohols, NH_3, and amines can be restored by thermal treatment at sufficiently high temperatures.

From the industrial point of view, the regeneration of coked catalysts is very important. The removal of coke depends on its structure and on the catalyst composition. Alkali metals, especially potassium, accelerate coke gasification. Oxidation is the fastest gasification reaction, but it is highly exothermic [410], [555]. To maintain the temperature within allowed limits, mixtures of O_2, steam, and N_2 are mainly used to remove the coke.

Catalysts deactivated by thermal degradation are very difficult to regenerate. Certain $Pt-Al_2O_3$ catalysts, deactivated as a result of thermal Pt sintering, can be partly regenerated by chlorine treatment at elevated temperatures, which makes Pt redistribution possible.

9.3. Catalyst Reworking and Disposal

The leaching of precious metals from spent catalysts is widely practiced. Producers of noble metal catalysts, such as Engelhard Corp., Johnson Matthey, and Degussa have plants for this purpose. Yields of recovered precious metals are 90 – 98 % depending on the original metal content and on the nature of the support used.

Besides precious metals, Ni from spent Ni – kieselguhr catalysts used in fat and oil hydrogenations are reworked to a large extent. Before Ni leaching, fats must be removed.

Hydrotreating catalysts containing Mo, Ni or Co are also reworked. Before leaching of the metals, coke and sulfur are removed by roasting [410], [547].

In various cases, however, problems arise when the price of recovered components does not cover the costs of catalyst reworking. The final alternative is the disposal of the spent catalyst. In general, catalysts containing Al, Si, Fe can be disposed of without any special precautions or can be used in construction materials. However, if such catalysts contain Ni or V accumulated during their use, then removal of these elements below legislative limits is necessary [410].

Spent Cu- and Cr-containing catalysts are sometimes accepted by metallurgical plants.

Before disposal, spent catalysts containing various contaminants need to be encapsulated to avoid their release into water. Materials used for encapsulation are, e.g., bitumen, cement, wax, and polyethylene [410]. Nevertheless, the disposal of encapsulated catalysts is not only expensive but is becoming increasingly difficult.

10. References

1. *Appl. Catal. A, General* **173** (1998) N3.
2. J. J. Berzelius, *Ann. Chim. Phys.* **61** (1836) 146.
3. B. H. Davis, "Development of the Science of Catalysis" in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 13.
4. W. Ostwald, *Nature* **65** (1902) 522.
5. P. Sabatier, J. B. Senderens, *C. R. Acad. Sci.* **134** (1902) 514.
6. H. Heinemann, "Development of Industrial Catalysis" in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 35.
7. S. A. Topham in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 7,** Springer, Berlin, 1987, p. 1.
8. F. Aftalion: *A History of the International Chemical Industry,* University of Pennsylvania Press, Philadelphia, 1991.
9. A. Mittasch, *Adv. Catal.* **2** (1950) 81.
10. DE 301 231, 1919 (F. Bergius, J. Billwiller).
11. F. Fischer, H. Tropsch, *Brennstoff-Chem.* **4** (1923) 276.
12. DE 103 362, 1943 (O. Roelen).
13. J. N. Armor, *Appl. Catal.* **78** (1991) 141.
14. A. Chauvel, B. Delmon, W. F. Hoelderich, *Appl. Catal. A* **115** (1994) 173.
15. M. Misono, N. Nojiri, *Appl. Catal.* **64** (1990) 1.
16. M. Boudart in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* Wiley-VCH, Weinheim, 1997, p. 1.
17. A. Balandin, *Adv. Catal. Rel. Subj.* **19** (1969) 1.

18. W. J. M. Rootsaert, W. M. H. Sachtler, *Z. Phys. Chem.* **26** (1960) 16.
19. I. Langmuir, *J. Am. Chem. Soc.* **37** (1915) 1139.
20. I. Langmuir, *Trans. Faraday Soc.* **17** (1922) 607.
21. H. S. Taylor, *Proc. Roy. Soc. (London)* **A 108** (1925) 105.
22. M. Che, C. O. Bennett, *Adv. Catal.* **36** (1989) 55.
23. G.-M. Schwab, E. Pietsch, *Z. Phys. Chem.* **131** (1929) 385.
24. S. M. Davis, G. A. Somorjai in D. A. King, D. P. Woodruff (eds.): *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis,* **Vol. 4,** Elsevier, New York 1982, p. 217.
25. P. G. Menon, T. S. R. Prasada Rao, *Catal. Rev.-Sci. Eng.* **20** (1979) 97.
26. M. Boudart, A. Aldag, J. E. Benson, N. A. Dougharty, C. G. Harkins, *J. Catal.* **6** (1966) 92; M. Boudart, A. Aldag, L. D. Ptak, J. E. Benson, *J. Catal.* **11** (1968) 35.
27. G. A. Somorjai, *Ann. Rev. Phys. Chem.* **45** (1994) 721.
28. G. Ertl, *Adv. Catal.* **45** (2000) 1.
29. V. Ponec, W. M. H. Sachtler in G. C. Bond, P. B. Wells, F. C. Tompkins (eds.): *Proc. 5th Intern. Congr. Catal.,* **Vol. 1,** The Chemical Society, London 1976, p. 645.
30. R. L. Burwell, G. L. Haller, K. C. Taylor, J. F. Read, *Adv. Catal.* **20** (1969) 1.
31. R. L. Burwell in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 9,** Springer, Heidelberg 1991, p. 1.
32. W. M. H. Sachtler in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 1077.
33. B. Hammer, J. K. Nørskov, *Adv. Catal.* **45** (2000) 71.
34. R. A. van Santen: *Theoretical Heterogeneous Catalysis,* World Scientific, Singapore 1991.
35. R. A. van Santen, M. Neurock in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 991.
36. J. M. Basset, J. P. Candy, A. Choplin, B. Didillon, F. Quignard, A. Theolier in J. M. Thomas, K. I. Zamaraev (eds.): *Perspectives in Catalysis,* Blackwell Scientific Publications, London 1992, p. 125.
37. J. M. Basset, B. C. Gates, J. P. Candy, A. Choplin, M. Leconte, F. Quignard, C. Santini (eds.): *Surface Organometallic Chemistry: Molecular Approaches to Surface Catalysis,* Kluwer Academic Publ., Dordrecht, 1988.
38. K. Hauffe, H.-J. Engell, *Z. Elektrochem.* **56** (1952) 366.
39. K. Hauffe, *Adv. Catal. Rel. Subj.* **7** (1955) 213.
40. Th. Wolkenstein, *Adv. Catal. Rel. Subj.* **9** (1957) 807, 818.
41. F. S. Stone, *J. Solid State Chem.* **12** (1975) 271.
42. M. P. Kiskinova, *Poisoning and Promotion in Catalysis, Stud. Surf. Sci. Catal.* **70** (1992).
43. Z. Paal, G. A. Somorjai in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 1084.
44. G. J. Hutchings, *Catal. Lett.* **75** (2001) 1.
45. R. Schlögl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim 1997, p. 1697.
46. H. Topsøe, B. S. Clausen, F. E. Massoth in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 11,** Springer, Berlin, 1996.
47. J. H. Sinfelt in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim 1997, p. 1939.
48. S. A. Stevenson, J. A. Dumesic, R. T. K. Baker, E. Ruckenstein: *Metal-Support Interactions in Catalysis, Sintering and Redispersion,* Van Nostrand Reinhold, New York, 1987.
49. H. Knözinger, E. Taglauer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim 1997, p. 216.
50. G. C. Bond in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim 1997, p. 752.
51. S. J. Tauster, *Acc. Chem. Res.* **20** (1987) 389.
52. F. C. M. J. M. van Delft, A. D. van Langeveld, B. E. Nieuwenhuys, *Solid State Ionics* **16** (1985) 233.
53. G. M. Pajonk in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 1064.
54. W. C. Conner, G. M. Pajonk, S. J. Teichner, *Adv. Catal.* **34** (1986) 1.
55. W. C. Conner, Jr., J. L. Falconer, *Chem. Rev.* **95** (1995) 759.
56. E. Keren, A. Soffer, *J. Catal.* **50** (1977) 43.
57. G. A. Somorjai, *J. Phys. Chem.* **94** (1990) 1013.
58. B. Delmon, *Stud. Surf. Sci. Catal.* **77** (1993) 1.

59. L. T. Weng, P. Ruiz, B. Delmon, *Stud. Surf. Sci. Catal.* **72** (1982) 399.
60. R. K. Grasselli, *Top. Catal.* **15** (2001) 93.
61. R. K. Grasselli in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim 1997, p. 2302.
62. J. L. Callahan, R. K. Grasselli, *AIChE J.* **9** (1963) 755.
63. G. Emig, R. Dittmeyer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 1209.
64. P. B. Weisz, V. J. Frilette, R. W. Maatman, E. B. Mower, *J. Catal.* **1** (1962) 307.
65. S. M. Csicsery, *J. Catal.* **19** (1970) 394.
66. M. Boudart in J. M. Thomas, K. I. Zamaraev (eds.): *Perspectives in Catalysis,* Blackwell Scientific Publications, London, 1992, p. 183.
67. M. Boudart in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim 1997, p. 958.
68. M. Boudart, *Catal. Rev.-Sci. Eng.* **23** (1981) 1.
69. G. F. Froment, L. Hosten in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 2,** Springer, Heidelberg, 1980, p. 97.
70. M. Boudart, G. Djéga-Mariadassou: *Kinetics of Heterogeneous Catalytic Reactions,* Princeton University Press, Princeton, 1984.
71. O. A. Hougen, K. M. Watson: *Chemical Process Principles. Part Three: Kinetics and Catalysis,* Wiley, New York, 1947.
72. J. A. Dumesic, D. F. Rudd, L. M. Aparicio, J. E. Rekoske, A. A. Treviño: *The Microkinetics of Heterogeneous Catalysis,* American Chemical Society, Washington, 1993.
73. R. D. Cortright, J. A. Dumesic, *Adv. Catal.* **46** (2001) 161.
74. R. A. van Santen, J. W. Niemantsverdriet: *Chemical Kinetics and Catalysis,* Plenum Press, New York, 1995.
75. *Catalysis Today,* 1999, Vol. 3, No. 2.
76. M. I. Temkin, V. Pyzhev, *Acta Physicochim. URSS* **12** (1940) 217.
77. M. Boudart in *Catalysis,* **Vol. 14,** The Royal Society of Chemistry, Cambdrige, p. 93.
78. M. Boudart, *Topics Catal.* **14** (2001) 181.
79. G. M. Schwab: *Katalyse vom Standpunkt der chemischen Kinetik,* Springer, Berlin, 1931.
80. M. Boudart, *I & EC Fundamentals* **25** (1986) 70.
81. R. D. Cortright, J. A. Dumesic, R. J. Madon, *Topics Catal.* **4** (1997) 15.
82. J. Horiuti, *J. Res. Inst. Catal. (Hokkaido Univ.)* **5** (1957) 1.
83. A. Varma, M. Morbidelli, H. Wu: *Parametric Sensitivity in Chemical Systems,* Cambridge Univ. Press, Cambridge, 1999.
84. C. T. Campbell, *Topics Catal.* **1** (1994) 364.
85. G. Ertl in J. R. Jennings (ed.): *Catalytic Ammonia Synthesis: Fundamentals and Practice, Fundamental and Applied Catalysis,* Plenum Press, New York, 1991, p. 109.
86. P. Stoltze, J. K. Nørskov, *Phys. Rev. Lett.* **55** (1985) 2502.
87. W. J. Thomas in J. M. Thomas, K. I. Zamaraev (eds.): *Perspectives in Catalysis,* Blackwell Scientific Publications, London, 1992, p. 251.
88. F. Kapteijn, J. A. Moulijn in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol.** Wiley-VCH, Weinheim 1997, p. 1189.
89. R. Mezaki, H. Inoue: *Rate Equations of Solid Catalyzed Reactions,* University of Tokyo Press, Tokyo, 1991.
90. A. Wheeler, *Adv. Catal.* **2** (1951) 250.
91. E. Shustorovich (ed.): *Metal-Surface Reaction Energetics,* Wiley-VCH, Weinheim, 1991.
92. J. K. Nørskov, P. Stoltze in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim 1997, p. 984.
93. C. O. Bennett, *Adv. Catal.* **44** (1999) 329.
94. K. Tamaru in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 9,** Springer, Berlin, 1991, p. 87.
95. S. L. Shannon, J. G. Goodwin, Jr., *Chem. Rev.* **95** (1995) 677.
96. A. Ozaki: *Isotopic Studies of Heterogeneous Catalysis,* Kodansha, Tokyo and Academic Press, New York, 1977.
97. G. F. Berndt in *Catalysis,* **Vol. 6,** The Royal Society of Chemistry, London, 1983, p. 144.
98. G. Liu, D. Willcox, M. Garland, H. H. Kung, *J. Catal.* **96** (1985) 251.
99. R. P. Bell, *Chem. Soc. Rev.* **3** (1974) 513.
100. L. Melander, W. H. Sauder, Jr.: *Reaction Rates of Isotopic Molecules,* Wiley, New York, 1980.
101. K. Tamaru, S. Naito in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1005.
102. S. Siegel, *Adv. Catal.* **16** (1966) 124.
103. J. K. A. Clarke, J. J. Rooney, *Adv. Catal.* **25** (1976) 125.
104. M. Kraus, *Adv. Catal.* **29** (1980) 151.

105. M. Kraus in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1051.
106. A. J. Gellman, *Curr. Opin. Solid State Mater. Sci.* **5** (2001) 85.
107. A. J. Gellman, *Acc. Chem. Res.* **33** (2000) 19.
108. H. Knözinger, *Adv. Catal.* **25** (1976) 184.
109. B. C. Gates: *Catalytic Chemistry,* Wiley, New York, 1992.
110. C. N. Satterfield: *Heterogeneous Catalysis in Industrial Practice,* McGraw-Hill, New York, 1991.
111. R. J. Farrauto, C. H. Bartholomew: *Fundamentals of Industrial Catalytic Processes,* Blackie Academic and Professional, London, 1997.
112. G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vols. 1 – 5,** Wiley-VCH, Weinheim, 1997.
113. J. Hagen: *Industrial Catalysis, A Practical Approach,* Wiley-VCH, Weinheim, 1999.
114. K. Tanabe: *Solid Acids and Bases,* Kodansha, Tokyo, Academic Press, New York, London, 1970.
115. M. Che, O. Clause, Ch. Marcilly in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 191.
116. J. P. Brunelle, *Pure Appl. Chem.* **50** (1978) 1211.
117. B. G. Linsen (ed.): *Physical and Chemical Aspects of Adsorbents and Catalysts,* Academic Press, New York, 1970.
118. H. Knözinger, P. Ratnasamy, *Catal. Rev.-Sci. Eng.* **17** (1978) 31.
119. H. P. Boehm, H. Knözinger in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 4,** Springer, Berlin, 1983, p. 39.
120. C. S. John, M. S. Scurrell: *Catalysis, The Chemical Society, London,* **1** (1977) 136.
121. J. Wagner, W. Nehb in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1761.
122. E. F. Vansant, P. Van Der Voort, K. C. Vrancken (eds.): *Stud. Surf. Sci. Catal.* **93** (1995).
123. A. E. Legrand (ed.): *The Surface Properties of Silicas,* Wiley, Chichester, New York, Weinheim, Brisbane, Singapore, Toronto,1998.
124. H. Knözinger in P. Schuster, G. Zundel, C. Sandorfy (eds.): *The Hydrogen Bond,* **Vol. 3,** North Holland, Amsterdam, 1976, p. 1263.
125. J. M. Thomas, R. G. Bell, C. R. A. Catlow in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 286.
126. C. T. Kresge, M. E. Leonowicz, W. J. Roth, J. C. Vartuli, J. S. Beck, *Nature* **359** (1992) 710.
127. J. S. Beck, J. C. Vartuli, W. J. Roth, M. E. Leonowicz, C. T. Kresge, K. D. Schmidt, C. T. W. Chu, D. H. Olson, E. W. Sheppard, S. B. McCullan, J. B. Higgins, J. L. Schlenker, *J. Am. Chem. Soc.* **114** (1992) 10834.
128. E. A. Colbourn, W. C. Mackrodt, *Surf. Sci.* **143** (1984) 391.
129. S. Coluccia, A. J. Tench: *Proc. 7th Intern. Congr. Catal., Tokyo, 1980,* Kodansha, Tokyo, Elsevier, Amsterdam, 1981, p. 1154.
130. A. Zecchina, D. Scarano, S. Bordiga, G. Spoto, C. Lamberti, *Adv. Catal.* **46** (2001) 265.
131. R. N. Spitz, J. E. Barton, M. A. Barteau, R. H. Staley, A. W. Sleight, *J. Phys. Chem.* **90** (1986) 4067.
132. A. Zecchina, M. G. Lofthouse, F. S. Stone, *J. Chem. Soc. Faraday Trans. 1* **71** (1975) 1476.
133. S. Coluccia, A. M. Deane, A. J. Tench, *J. Chem. Soc. Faraday Trans. 1* **74** (1978) 2913.
134. H. Knözinger, *Science* **287** (2000) 1407.
135. A. Cimino, F. S. Stone in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 845.
136. C. N. R. Rao, B. Raveau: *Transition Metal Oxides,* VCH, Weinheim, 1995.
137. P. A. Cox: *Transition Metal Oxides,* Clarendon Press, Oxford, 1995.
138. H. H. Kung: *Transition Metal Oxides: Surface Chemistry and Catalysis,* Elsevier, Amsterdam, 1995.
139. P. Pichat in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2111.
140. A. L. Linsebigler, G. Lu, J. T. Yates, Jr., *Chem. Rev.* **95** (1995) 735.
141. K. Arata, *Adv. Catal.* **37** (1990) 165.
142. W. Göpel, K. D. Schierbaum in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1284.
143. K. Kochloefl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1831.
144. K. Kochloefl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2151.

145. J. Haber in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2253.
146. M. Muhler in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2274.
147. R. M. Barrer: *Hydrothermal Chemistry of Zeolites,* Academic Press, London, 1982.
148. W. M. Meier, D. H. Olson, Ch. Baerlocher: *Atlas of Zeolite Structure Types,* 4th Ed., Butterworth-Heinemann, London, 1996.
149. J. Weitkamp, *Solid State Ionics* **131** (2000) 175.
150. G. Bellussi, V. Fattore in P. A. Jacobs, N. Jaeger, L. Kubelkova, B. Wichterlova (eds.): *Zeolite Chemistry and Catalysis,* Elsevier, Amsterdam, 1991, p. 79.
151. D. Barthomeuf, *Catal. Rev.-Sci. Eng.* **38** (1996) 521.
152. J. M. Thomas, R. G. Bell, C. R. A. Catlow, G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 286.
153. T. Ishihara, H. Takita, *Catalysis* **12** (1996) 21.
154. J. J. Fripiat in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 387.
155. P. G. Menon, B. Delmon in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 100.
156. R. K. Grasselli, J. F. Brazdil: *Solid State Chemistry in Catalysis*, ACS Symposium Series, Amer. Chem. Soc., Washington, **279** (1985).
157. J. Haber, G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2253.
158. R. K. Grasselli, *J. Chem. Ed.* **63** (1986) 216.
159. M. Egasihra, K. Matsuo, S. Kawaga, T. Seiyama, *J. Catal.* **58** (1979) 409.
160. US 4 370 279, 1983 (Y. Sasaki, T. Nakamura, Y. Nakamura, K. Moriya, H. Utsumi, S. Saito).
161. A. W. Sleight in J. J. Burton, R. L. Garten (eds.): *Advanced Materials in Catalysis,* Academic Press, New York, 1977, p. 181.
162. L. G. Tejuca, J. L. G. Fierro: *Properties and Applications of Perovskite-type Oxides,* M. Dekker, New York, 1993.
163. P. Cavani, F. Trifiró, A. Vaccari, *Catal. Today* **11** (1991) 173.
164. F. Basile, M. Campanati, E. Serwicka, A. Vaccari, *Appl. Clay Sci.* **18** (2001) 1.
165. P. Gouzerh, A. Proust, *Chem. Rev.* **98** (1998) 77.
166. E. Coronado. C. J. Gómez-García, *Chem. Rev.* **98** (1998) 273.
167. K.-Y. Lee, M. Misono in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 118.
168. M. Misono, *Catal. Rev.-Sci. Eng.* **30** (1988) 339.
169. J. B. Moffat: *Metal-Oxygen Clusters,* Kluwer Academic/Plenum Publishers, New York, 2001.
170. J. A. Gamelas, F. A. S. Couto, M. C. N. Trovao, A. M. V. Cavaleiro, J. A. S. Cavaleiro, M. Guelton, *Thermochim. Acta* **326** (1999) 165.
171. T. Okuhara, M. Yamashita, K. Na, M. Misono, *Chem. Lett.* (1994) 1450.
172. Y. Izumi, M. Ogawa, W. Nohara, K. Krabe, *Chem. Lett.* (1992) 1987.
173. G. Centi, J. L. Nieto, C. Iapalucci, K. Brückman, E. M. Serwicka, *Appl. Catal.* **46** (1989) 197.
174. F. Cavani, M. Koutyrev, F. Trifiró, *Catal. Today* **28** (1996) 319.
175. E. Wagner, T. Fetzer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1748.
176. M. S. Wainright in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 64.
177. R. Schlögl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 54.
178. A. Baiker in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 803.
179. J. R. Jennings (ed.): *Catalytic Ammonia Synthesis: Fundamentals and Practice,* Plenum Press, New York, 1991.
180. S. T. Oyama in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 132. in ref. [4], p. 132.
181. A. York, *Chem. Br.* (1999) 25.
182. M. J. Ledoux, C. Pham-Huu, *CATTECH* **5** (2001) 226.
183. S. Kaskel, K. Schlichte, *J. Catal.* **201** (2001) 270.

184. C. Bouchy, C. Pham-Huu, B. Heinrich, C. Chaumot, M. J. Ledoux, *J. Catal.* **190** (2000) 92.
185. R. Schlögl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 138.
186. H. Jüntgen, H. Kühl, *Chem. Phys. Carbon* **22** (1990) 145.
187. L. R. Radovic, F. Rodriguez-Reinoso, *Chem. Phys. Carbon* **35** (1997) 243.
188. F. de Dardel, T. V. Arden: *Ullmann's Encyclopedia of Industrial Chemistry,* 5th ed., **Vol. A 14,** VCH Verlagsgesellschaft, Weinheim, 1987, p. 393.
189. A. Stüwe, C.-P. Hälsig, H. Tschorn in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1986.
190. G. A. Olah, A. Molnar: *Hydrocarbon Chemistry,* Wiley, New York, 1995.
191. J.-P. Vigneron in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 888.
192. K. Morihara, S. Doi, M. Takiguchi, T. Shimada, *Bull. Chem. Soc. Jpn.* **66** (1993) 2977.
193. G. Wulff in C. G. Gebelein (ed.): *Biomimetic Polymers,* Plenum Press, New York, 1990, p. 1.
194. K. Morihara, M. Takaguchi, T. Shimada, *Bull. Chem. Soc. Jpn.* **67** (1994) 1078.
195. H. Sieber, C. Hoffmann, A. Kaindl, P. Greil, *Adv. Eng. Mater.* **2** (2000) 105.
196. J. L. Williams, *Catal. Today* **69** (2001) 3.
197. M. Valentini, G. Groppi, C. Cristiani, M. Levi, E. Tronconi, P. Forzatti, *Catal. Today* **69** (2001) 307.
198. E. S. J. Lox, B. H. Engler in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1569.
199. J. M. Thomas, *Angew. Chem.* **111** (1999) 380; *Angew. Chem. Int. Ed.* **38** (1999) 3588.
200. A. Stein, B. J. Melde, R. C. Schroden, *Adv. Mater.* **12** (2000) 1403.
201. D. E. De Vos, B. F. Sels, P. A. Jacobs, *Adv. Catal.* **46** (2001) 2.
202. D. C. Sherrington, A. P. Kybett (eds.): *Supported Catalysts and their Applications,* The Royal Society of Chemistry, Cambridge, 2001.
203. B. K. Hodnett, A. Kybett, J. H. Clark, K. Smith: *Supported Reagents and Catalysts in Chemistry,* The Royal Society of Chemistry, Cambridge, 1998.
204. B. Clapham, T. . Reger, K. D. Janda, *Tetrahedron* **57** (2001) 4637.
205. W. Keim, B. Driessen-Hölscher in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 231.
206. I. E. Wachs: *Catalysis,* **Vol. 13,** The Royal Society of Chemistry, Cambridge, 1997, p. 37.
207. B. Delmon in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 264.
208. G. C. Bond, J. C. Vedrine, *Catal. Today* **20** (1994) 1.
209. G. Centi, *Appl. Catal. A: General* **147** (1996) 267.
210. F. J. Janssen in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1633.
211. J. F. Armor, *Chem. Mat.* **6** (1994) 730.
212. R. Prins in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1908.
213. R. Prins, *Adv. Catal.* **46** (2001) 399.
214. M. Hino, K. Arata: *Chem. Commun.* 1988, 1259.
215. S. Kuba, P. Concepción, R. K. Grasselli, B. C. Gates, M. Che, H. Knözinger, *Phys. Chem. Chem. Phys.* **3** (2001) 146.
216. S. Kuba, B. C. Gates, P. Vijayanand, R. K. Grasselli, H. Knözinger: *Chem. Commun.* 2001, 57.
217. J. C. Mol in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2387.
218. F. Buonomo, D. Sanfilippo, F. Trifiró in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2140.
219. S. H. Overbury, P. A. Bertrand, G. A. Somorja, *Chem. Rev.* **75** (1975) 547.
220. Y. Ono, T. Baba: *Catalysis,* **Vol. 15,** The Royal Society of Chemistry, Cambridge, 2000, p. 1.
221. K. Tanabe, H. Hattori in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 404.
222. X. Song, A. Sayari, *Catal. Rev.-Sci. Eng.* **38** (1996) 329.
223. *Topics in Catalysis* **6** (1998).
224. K. Foger in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 6,** Springer, Berlin, 1984, p. 228.

225. B. C. Gates in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 793.
226. B. C. Gates, *Chem. Rev.* **95** (1995) 511.
227. J. H. Sinfelt: *Bimetallic Catalysts,* Wiley, New York, 1983.
228. C. T. Campbell in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 814.
229. A. Baiker, *Curr. Opin. Sol. State Mater. Sci.* **3** (1998) 86.
230. J. Wei in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1928.
231. S. Helveg, J. V. Lauritsen, E. Lægsgaard, I. Stensgaard, J. K. Nørskov, B. S. Clausen, H. Topsøe, F. Besenbacher, *Phys. Rev. Lett.* **84** (2000) 951.
232. D. E. De Vos, I. F. J. Vankelecom, P. A. Jacobs: *Chiral Catalyst Immobilization and Recycling,* Wiley-VCH, Weinheim, 2000.
233. D. E. De Vos, S. de Wildman, B. F. Sels, P. J. Grobet, P. A. Jacobs, *Angew. Chem.* **111** (1999) 1033; *Angew. Chem. Int. Ed.* **38** (1999) 980.
234. K. Dranz, H. Waldmann: *Enzyme Catalysis in Organic Synthesis,* VCH Verlagsgesellschaft, Weinheim, 1994.
235. A. W. Bosman, H. M. Janssen, E. W. Meijer, *Chem. Rev.* **99** (1999) 1665.
236. H. Brunner, *J. Organomet. Chem.* **500** (1995) 39.
237. A. Kirschning, H. Monenschein, R. Wittenberg, *Angew. Chem. Int. Ed.* **40** (2001) 650.
238. B. Cornils, W. A. Herrmann: *Applied Homogeneous Catalysis with Organometallic Compounds,* Wiley-VCH, Weinheim, 1996, p. 619.
239. J. P. Arhancet, M. E. Davis, J. S. Merola, B. E. Hanson, *Nature* **339** (1989) 454; K. T. Wan, M. E. Davis, *Nature* **370** (1994) 449.
240. J. Adlkofer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1774.
241. G. Schulz-Ekloff, S. Ernst in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 374.
242. D. E. De Vos, P. P. Knops-Gerrits, R. F. Parton, B. M. Weckhuysen, P. A. Jacobs, R. A. Schoonheydt, *J. Incl. Phnom.* **21** (1995) 185.
243. R. Parton, D. E. De Vos, P. A. Jacobs in E. G. Derouane, F. Lemos, C. Naccache, F. Ramoa Ribeiro (eds.): *Zeolite Microporous Solids: Synthesis, Structure and Reactivity,* Kluwer Academic Publ., Dordrecht, 1995, p. 555.
244. J.-M. Lehn: *Supramolecular Chemistry,* VCH Verlagsgesellschaft, Weinheim, 1995.
245. M. P. McDaniel in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2400.
246. W. Kaminsky, *Adv. Catal.* **46** (2001) 89.
247. W. Kaminsky, *Macromol. Chem. Phys.* **197** (1996) 3907.
248. K. Kochloefl, *Quo vadis heterogene Katalyse,* Dechema Tagung, XXVI Jahrestreffen deutscher Katalytiker, Schloß Reinhardsbrunn, Germany, 1993.
249. H. Heinemann: "Development of Industrial Catalysis" in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 35.
250. J. T. Richardson in M. V. Twigg, M. S. Spencer (eds.): *Principle of Catalyst Development,* Plenum Press, New York, 1989, p. 95.
251. J. A. Cusumano in J. M. Thomas, K. I. Zamaraev (eds.): *Perspectives in Catalysis,* Blackwell Scient. Publ., Oxford, 1991, p. 1.
252. K. Fouhy, G. Samdani, S. Moore, *Chem. Eng.,* October (1992) 47.
253. J. M. Fulton, *Chem. Eng.* **7** (1986) 59.
254. P. Courty, C. Marcilly in G. Poncelet, P. Grange, P. Jacobs (eds.): *Preparation of Catalysts III,* Elsevier, Amsterdam, 1983, p. 485.
255. M. Sitting: *Handbook of Catalyst Manufacture,* Noyes Data Corp., Park Ridge, 1971.
256. B. Stiles, T. A Koch, *Catalyst Manufacture,* 2nd. ed., M. Dekker, NewYork, 1995.
257. F. Schüth, K. Unger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 72.
258. E. I. Ko in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 86.
259. H. Jacobsen, P. Kleinschmit in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 94.
260. E. J. P. Feijen, J. A. Martens, P. A. Jacobs in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 311.

261. J. W. Geus, J. van Dillen: "Preparation of Supported Catalysts by Deposition – Precipitation" inG. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 240.
262. J. Barbier in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 257.
263. J. F. Le Page in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 412.
264. J. W. Fulton, *Chem Eng.,* May 12 (1986) 97.
265. *General Catalogue,* Süd-Chemie AG, Catalyst Division, Munich, Germany.
266. G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997.
267. K. S. W. Sing, D. H. Everett, R. A. W. Haul, L. Moscou, R. A. Pierotti, J. Rouquerol, T. Siemieniewska, *Pure Appl. Chem.* **57** (1985) 603.
268. K. S. W. Sing, J. Rouquerol in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 427.
269. J. Rouquerol, D. Avnir, C. W. Fairbridge, D. H. Everett, J. M. Haynes, N. Pernicone, J. D. F. Ramsay, K. S. W. Sing, K. K. Unger, *Pure Appl. Chem.* **66** (1994) 1739.
270. B. C. Lippens, J. H. de Boer, *J. Catal.* **4** (1965) 319.
271. K. S. W. Sing, D. H. Everett, R. H. Ottewill (eds.): *Surface Area Determination,* Butterworths, London, 1970, p. 25.
272. K. Datye in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 493.
273. G. Bergeret, P. Gallezot in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 439.
274. P. Gallezot in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 5,** Springer, Berlin, 1984, p. 221.
275. R. J. Matyi, L. R. Schwartz, J. B. Butt, *Catal. Rev.-Sci. Eng.* **29** (1987) 41.
276. A. K. Datye, D. J. Smith, *Catal. Rev.-Sci. Eng.* **34** (1992) 129.
277. H. Poppa, *Catal. Rev.-Sci. Eng.* **35** (1993) 359.
278. M. J. Yacaman, G. Diaz, A. Gomez, *Catal. Today* **23** (1995) 161.
279. G. Bergeret in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 464.
280. B. S. Clausen, G. Steffensen, B. Fabius, J. Villadsen, L. R. Feidenhaus, H. Topsøe, *J. Catal.* **132** (1991) 524.
281. M. Vaarkamp, D. C. Konigsberger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 475.
282. J. H. Sinfelt, G. D. Meitzner, *Acc. Chem. Res.* **26** (1993) 1.
283. J. C. Conesa, P. Esteban, H. Dexpert, D. Bazin, *Stud. Surf. Sci. Catal.* **57** (1990) 225.
284. J. M. Thomas, *Chem. Eur. J.* **3** (1997) 1557.
285. D. J. Smith, M. R. McCartney, J. K. Weiss, *Ultramicroscopy* **52** (1993) 591.
286. R. T. K. Baker, *Catal. Rev.-Sci. Eng.* **19** (1979) 161; R. T. K. Baker, N. M. Rodriguez, *Energy and Fuels* **8** (1994) 330.
287. G. Mestl, H. Knözinger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 539.
288. C. Li, P. C. Stair, *Catal. Lett.* **36** (1995) 119.
289. H. Knözinger, *Catal. Today* **32** (1996) 71.
290. Y. R. Shen, *Surface Sci.* **299/300** (1994) 551.
291. K. B. Eisenthal, *Chem. Rev.* **96** (1996) 1343.
292. H. Jobic in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 574.
293. J. W. Niemantsverdriet, T. Butz in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 512.
294. A. M. van der Kraan, J. W. Niemantsverdriet in G. J. Lang, J. G. Stevens (eds.): *Industrial Applications of the Mössbauer Effect,* Plenum Press, New York, 1985, p. 609.
295. A. Lerf, T. Butz, *Angew. Chem.* **99** (1987) 113.
296. P. Mottner, T. Butz, A. Lerf, G. Ledezma, H. Knözinger, *J. Phys. Chem.* **99** (1995) 8260.
297. G. Engelhardt in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 525.
298. G. Engelhardt, D. Michel: *High-Resolution Solid-State NMR of Silicates and Zeolites,* Wiley, Chichester, 1987.
299. A. T. Bell, A. Pines (eds.): *NMR Techniques in Catalysis,* M. Dekker, New York, 1994.
300. E. Taglauer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 614.

301. J. W. Niemandsverdriet: *Spectroscopy in Catalysis,* VCH Verlagsgesellschaft, Weinheim, 1995.
302. D. Briggs, M. P. Seah: *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy,* Wiley, New York, 1985.
303. E. Taglauer in A. W. Czanderna, D. M. Hercules (eds.): *Ion Spectroscopies for Surface Analysis,* Plenum Press, New York, 1991, p. 363.
304. H. Niehus, W. Heiland, E. Taglauer, *Surf. Sci. Rep.* **17** (1993) 217.
305. W. K. Chu, J. W. Mayes, M. A. Nicolet: *Backscattering Spectrometry,* Academic Press, New York, 1978.
306. L. C. Feldman in A. W. Czanderna, D. M. Hercules (eds.): *Ion Spectroscopies for Surface Analysis,* Plenum Press, New York, 1991, p. 311.
307. A. Benninghoven, F. G. Rüdenauer, H. W. Werner: *Secondary Ion Mass Spectrometry,* Wiley, New York, 1987.
308. H. J. Borg, J. W. Niemandsverdriet in J. J. Spivey, S. Agarwal (eds.): *Catalysis,* **Vol. 11,** The Royal Society of Chemistry, Cambridge, 1994, p. 1.
309. J. C. Vickerman, A. Swift in J. C. Vickerman (ed.): *Surface Analysis – The Principal Techniques,* Wiley, Chichester, 1997,
310. H. Oechsner, *Scanning Microsc.* **2** (1988) 9.
311. J. C. Vickerman (ed.): *Surface Analysis – The Principal Techniques,* Wiley, Chichester, 1997, p. 135.
312. G. Moretti in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 632.
313. C. D. Wagner, A. Joshi, *J. Electron Spectrosc.* **47** (1988) 283.
314. M. Che, F. Bozon-Verduraz in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 641.
315. W. N. Delgass, G. L. Haller, R. Kellerman, J. H. Lunsford: *Spectroscopy in Heterogeneous Catalysis,* Academic Press, New York, 1979.
316. R. A. Schonheydt in F. Delannay (ed.): *Characterization of Heterogeneous Catalysts,* Dekker, New York, 1984, p. 125.
317. F. Stone in J. P. Bonella, B. Delmon, E. G. Deronane (eds.): *Surface Properties and Catalysis by Non-Metals,* Reidel, Boston, 1983, p. 237.
318. M. Gerlach, E. C. Constable: *Transition Metal Chemistry,* VCH Verlagsgesellschaft, Weinheim, 1994.
319. G. Kortüm: *Reflexionsspektroskopie,* Springer, Berlin, 1969.
320. M. Anpo in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 664.
321. M. Anpo, M. Che, *Adc. Catal.* **44** (1999) 119.
322. J. H. Lunsford in J. R. Anderson, M. Bondart (eds.): *Catalysis – Science and Technology,* **Vol. 8,** 1997, p. 227.
323. K. Dyrek, M. Che, *Chem. Rev.* **97** (1997) 305.
324. P. D. Garn: *Thermoanalytical Methods of Investigation,* Academic Press, New York, 1965.
325. R. C. Mackenzie: *Differential Thermal Analysis,* Academic Press, London, New York, 1972.
326. S. D. Robertson, B. D. McNicol, J. H. de Bass, S. C. Kloet, J. W. Jenkins, *J. Catal.* **37** (1975) 424.
327. H. Knözinger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 676.
328. D. A. M. Monti, A. Baiker, *J. Catal.* **83** (1983) 323.
329. P. Malet, A. Caballero, *J. Chem. Soc. Faraday Trans. I* **84** (1988) 2369.
330. V. B. Kazansky in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 740.
331. W. K. Hall in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 692.
332. H. A. Benesi, *J. Phys. Chem.* **61** (1957) 970.
333. M. Deeba, W. K. Hall, *J. Catal.* **60** (1979) 417.
334. B. E. Spiewak, R. D. Cartright, J. A. Dumesic in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 698.
335. J. L. Falconer, J. A. Schwarz, *Catal. Rev.-Sci. Eng.* **25** (1983) 414.
336. H. Karge, V. Dondur, *J. Phys. Chem.* **94** (1990) 765.
337. S. Chatterjee, H. L. Greene, Y. J. Park, *J. Catal.* **138** (1992) 179.
338. A. Auroux, A. Gervasini, *J. Phys. Chem.* **94** (1990) 6371.
339. D. T. Chen, L. Zhang, C. Yi, J. A. Dumesic, *J. Catal.* **146** (1994) 257.
340. W. E. Farneth, R. J. Gorte, *Chem. Rev.* **95** (1995) 615.

341. H. Knözinger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 707.
342. E. A. Paukshtis, E. N. Yurchenko, *Russ. Chem. Rev.* **52** (1983) 42.
343. J. C. Lavalley, *Trends Phys. Chem.* **2** (1991) 305.
344. J. A. Lercher, C. Gründling, G. Eder-Mirth, *Catal. Today* **27** (1996) 353.
345. J. C. Lavalley, *Catal. Today* **27** (1996) 377.
346. H. Knözinger, S. Huber, *J. Chem. Soc. Faraday Trans.* **94** (1998) 2047.
347. G. C. Pimentel, A. L. McClellan: *The Hydrogen Bond,* Freeman, San Francisco, London, 1960.
348. S. Huber, H. Knözinger, *J. Mol. Catal.* **141** (1999) 117.
349. D. Mordenti, P. Grotz, H. Knözinger, *Catal. Today* **70** (2001) 83.
350. A. M. Ferrari, S. Huber, H. Knözinger, K. M. Neyman, N. Rösch, *J. Phys. Chem. B* **102** (1998) 4548.
351. H. Knözinger, H. Krietenbrink, H. D. Müller, W. Schulz: *Proceedings of the 6th International Congress on Catalysis, London, 1976,* The Chemical Society, London, 1977, p. 183.
352. H. Pfeifer in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 2,** Wiley-VCH, Weinheim, 1997, p. 732.
353. V. M. Mastikhin, I. L. Mundrakovsky, A. V. Nosov, *Progress NMR Spectrosc.* **23** (1991) 259.
354. M. Hunger, *Solid State Nucl. Magn. Res.* **6** (1996) 1.
355. E. Brunner, *Catal. Today* **38** (1997) 361.
356. V. Bosácek, *J. Phys. Chem.* **97** (1993) 10732; and *Z. Phys. Chem.* **189** (1995) 241.
357. J. F. Le Page: *Applied Heterogeneous Catalysis ⟶ Design, Manufacture, Use of Solid Catalysts,* Editions Technip, Paris, 1987.
358. J. C. Dart, *Chem. Eng. Prog.* **71** (1975) 46; and E. R. Beaver, *Chem. Eng. Prog.* **71** (1975) 44.
359. W. L. Forsythe, W. R. Hertwig, *Ind. Eng. Chem.* **41** (1949) 1200.
360. T. Engel, G. Ertl, *Adv. Catal.* **28** (1979) 1.
361. T. Engel, G. Ertl in D. P. Woodruff (ed.): *The Chemical Physics of Solid Surfaces and Heterogeneous Catalysis,* **Vol. 4,** Elsevier, Amsterdam, 1982, p. 73.
362. A. Nielsen: *Ammonia Synthesis,* Springer, New York, 1995.
363. K. Aika, A. Ozaki, *J. Catal.* **16** (1970) 97.
364. F. Bozso, G. Ertl, M. Weiss, *J. Catal.* **50** (1977) 519.
365. P. H. Emmett, S. Brunauer, *J. Am. Chem. Soc.* **59** (1937) 310.
366. M. Boudart, *Catal. Rev.-Sci. Eng.* **23** (1981) 1.
367. G. Ertl, M. Weiss, S. B. Lee, *Chem. Phys. Lett.* **60** (1979) 391.
368. D. S. Santilli, B. C. Gates in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1123.
369. G. A. Olah: *Friedel-Crafts Chemistry,* Wiley Interscience, New York, 1973.
370. H. Pines: *The Chemistry of Catalytic Hydrocarbon Conversions,* Academic Press, New York, 1981.
371. D. M. Brouwer, R. Prins, G. C. A. Schuit (eds.): *Chemistry and Chemical Engineering of Catalytic Processes,* Sijthoff and Nordhoff, Alphen an den Rijn, The Netherlands, 1980, pp. 137 – 160.
372. G. A. Olah, *Angew. Chem. Int. Ed. Engl.* **12** (1973) 173.
373. G. A. Olah, G. K. S. Prakash, J. Sommer: *Superacids,* Wiley, New York, 1985.
374. D. M. Brouwer, H. Hogeveen, *Prog. Phys. Org. Chem.* **9** (1972) 179.
375. B. C. Gates: *Catalytic Chemistry,* Wiley, New York, 1992, pp. 40 – 60.
376. D. M. Brouwer, J. A. Van Doorn, *Rec. Trav. Chim.* **91** (1972) 903.
377. V. B. Kazansky, I. N. Senchenya, *Catal. Lett.* **8** (1991) 317.
378. V. B. Kazansky, I. N. Senchenya, *J. Catal.* **119** (1989) 108.
379. V. B. Kazansky in J. C. Jansen, M. Stöcker, H. G. Karge, J. Weitkamp (eds.): *Advanced Zeolite Science and Applications,* Elsevier, Amsterdam, 1994, p. 251.
380. M. T. Aronson, R. J. Gorte, W. E. Farneth, D. White, *J. Am. Chem. Soc.* **111** (1989) 840.
381. V. B. Kazansky, I. N. Senchenya, M. Frash, R. A. van Santen, *Catal. Lett.* **27** (1994) 345.
382. V. B. Kazansky, M. Frash, R. A. van Santen, *Catal. Lett.* **28** (1994) 211.
383. J. A. Lercher, R. A. van Santen, H. Vinek, *Catal. Lett.* **27** (1994) 91.
384. J. A. Martens, P. A. Jacobs, *J. Catal.* **124** (1990) 357.
385. A. Corma, P. J. Miguel, A. V. Orchilles, *Appl. Catal. A* **117** (1994) 29.
386. S. T. Sie, *Ind. Eng. Chem. Res.* **32** (1993) 397.
387. D. S. Santilli, *Appl. Catal.* **60** (1990) 137.

388. J. A. Martens, P. A. Jacobs in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1137.
389. C. T. O'Connor in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2380.
390. S. T. Sie in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1998.
391. M. Guisnet, F. Avendano, C. Bearez, F. Chevalier, *J. Chem. Soc., Chem. Commun.* (1985) 336.
392. C. Bearez, F. Avendano, M. Guisnet, *Bull. Soc. Chim. France* **3** (1985) 346.
393. J. Abbot, *J. Catal.* **123** (1990) 383.
394. R. von Ballmoos, D. H. Harris, J. S. Magee in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1955.
395. C. T. Thomas, *Ind. Eng. Chem.* **41** (1949) 2564.
396. W. O. Haag, R. M. Dessau in *Proceedings of the 6th International Congress on Catalysis,* **Vol. 2,** Berlin, 1984, p. 305.
397. E. A. Lombardo, R. Pierantozzi, W. K. Hall, *J. Catal.* **110** (1988) 171.
398. S. Kotrel, H. Knözinger, B. C. Gates, *Microporous Mesoporous Mater.* **35/36** (2000) 11.
399. H. Krannila, W. O. Haag, B. C. Gates, *J. Catal.* **135** (1992) 115.
400. E. A. Lombardo, W. K. Hall, *J. Catal.* **112** (1988) 565.
401. D. B. Lukyanov, V. I. Shtral, S. N. Khadzhiev, *J. Catal.* **146** (1994) 87.
402. B. R. Bamwenda, Y. X. Zhao, B. W. Wojciechowski, *J. Catal.* **148** (1994) 595.
403. A. Corma, B. J. Miguel, A. V. Orchilles, *J. Catal.* **145** (1994) 171.
404. W. Wojciechowski, M. M. Bassir, *J. Catal.* **147** (1994) 352.
405. J. S. Beck, W. O. Haag in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2123.
406. J. Weitkamp, Y. Traa in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2039.
407. L. Schmerling in G. A. Olah (ed.): *Friedel-Crafts and Related Reactions,* Wiley Interscience, New York, 1964, pp. 363 – 407.
408. J. S. Beck, W. O. Haag in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2136.
409. C. T. O'Connor in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2069.
410. D. L. Trimm in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1263.
411. J. L. White, N. D. Lazo, B. R. Richardson, J. F. Haw, *J. Catal.* **125** (1990) 260.
412. S. M. Csicsery, *Zeolites* **4** (1984) 202.
413. N. Y. Chen, W. E. Garwood, *J. Catal.* **52** (1978) 453.
414. P. A. Jacobs: *Carboniogenic Activity of Zeolites,* Elsevier, Amsterdam, 1977, pp. 119 – 121.
415. D. S. Santilli, *J. Catal.* **99** (1986) 327.
416. W. O. Haag, R. M. Lago, P. B. Weisz, *Faraday Disc. Chem. Soc.* **72** (1981) 317.
417. D. S. Santilli, *J. Catal.* **99** (1986) 335.
418. D. S. Santilli, T. V. Harris, S. I. Zones, *Microporous Materials* **1** (1993) 329.
419. F. G. Gault, *Adv. Catal.* **30** (1981) 1.
420. Z. Paal, *Adv. Catal.* **29** (1980) 273.
421. J. K. A. Clarke, J. J. Rooney, *Adv. Catal.* **25** (1976) 125.
422. J. R. Anderson, *Adv. Catal.* **23** (1973) 1.
423. J. H. Sinfelt, *Adv. Catal.* **23** (1973) 91.
424. E. van Broekhoven, V. Ponec, *Prog. Surf. Sci.* **19** (1985) 351.
425. Z. Paal, P. Tetenyi in G. C. Bond, G. Webb (eds.): *Catalysis Specialist Periodical Reports,* **Vol. 5,** Royal Society of Chemistry, London, 1982, p. 80.
426. M. Leconte, *J. Mol. Catal.* **86** (1994) 205.
427. Z. Paal in G. J. Antos, A. M. Aitani, J. M. Parera (eds.): *Catalytic Naphtha Reforming,* M. Dekker, New York, 1995, p. 19.
428. H. Arnold, F. Döbert, J. Gaube in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2165.
429. R. K. Grasselli in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2302.
430. P. S. Cremer, X. Su, Y. R. Shen, G. A. Somorjai, *J. Am. Chem. Soc.* **118** (1996) 2942.
431. J. Toyir, M. Leconte, G. P. Niccolai, J.-M. Basset, *J. Catal.* **152** (1995) 306.
432. R. A. van Santen G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2244.

433. R. A. van Santen, H. C. P. E. Kuipers, *Adv. Catal.* **35** (1988) 265.
434. M. C. Zonnevylle, J. J. C. Geerlings, R. A. van Santen, *J. Catal.* **148** (1994) 417.
435. M. Neurock, R. A. van Santen, W. Biemolt, A. P. J. Jansen, *J. Am. Chem. Soc.* **116** (1994) 6860.
436. J. W. He, U. Memmert, K. Griffiths, P. R. Norton, *J. Chem. Phys.* **90** (1989) 5082.
437. P. J. van den Hoek, E. J. Baerends, R. A. van Santen, *J. Phys. Chem.* **93** (1989) 6469.
438. L. M. Akella, H. N. Lee, *J. Catal.* **86** (1984) 465.
439. M. Bowker, R. C. Waugh, *Surf. Sci.* **155** (1984) 1.
440. G. Bellussi, M. S. Rigutto, *Stud. Surf. Sci. Catal.* **85** (1991) 177.
441. P. N. Rylander in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* **Vol. 4,** Springer, Berlin, 1983, p. 1.
442. P. Mars, D. W. van Krevelen, *Chem. Eng. Sci. Special Suppl.* **3** (1954) 41.
443. R. K. Grasselli, J. D. Burrington, *I & EC Res. & Dev.* **23** (1984) 393.
444. T. P. Snyder, C. G. Hill, *Catal. Rev.-Sci. Eng.* **31** (1989) 43.
445. J. D. Burrington, C. T. Kartisek, R. K. Grasselli, *J. Catal.* **87** (1984) 363.
446. R. K. Grasselli, *J. Chem. Ed.* **63** (1986) 216.
447. B. C. Gates: *Catalytic Chemistry,* Wiley, New York, 1992, p. 406.
448. R. Prius in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1908.
449. D. D. Whitehurst, T. Isoda, I. Mochida, *Adv. Catal.* **42** (1998) 345.
450. M. Houalla, N. K. Nag, A. V. Sapre, D. H. Broderick, B. C. Gates, *AIChE J.* **24** (1978) 1015.
451. K. Kochloefl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1831.
452. A. Nielsen: *Ammonia Synthesis,* Springer, New York, 1995.
453. W. F. Hölderich in L. Guczi, F. Solymosi, P. Tétényi (eds.): *Proc. 10th Intern. Congress on Catal., Akademia Kiado,* Budapest, 1993, p. 127.
454. P. Baumeister, M. Studer, F. Roesler in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2186.
455. R. J. Peterson: *Hydrogenation Catalysts,* Noyes Data Corp., New Jersey, USA, 1977.
456. H. D. Neubauer et al.: *Selective Hydrogenations and Dehydrogenations,* Kassel, 1993, p. 67.
457. Eur. pat. 552 809, 1993 (K. Yamashita et al.).
458. P. Gallezot in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2209.
459. E. W. Gritz in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2221.
460. J. B. Hansen in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1856.
461. M. E. Dry in J. R. Anderson, M. Boudart (eds.): *Catalysis: Science and Technology,* Springer, New York, 1982, p. 159.
462. V. A. Likholobov, B. L. Moroz in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2231.
463. R. Whyman, "Industrial Applications of Homogeneous Catalysts" in Selected Developments in Catalysis, Critical Reports on Appl. Chem., **Vol. 12,** Blackwell Scientific Publ., Oxford, 1985.
464. R. T. Eby, T. C. Singleton in Appl. Ind. Catal., **Vol. 1,** Academic Press, New York, 1983.
465. K Kochloefl in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2151.
466. M. Kraus in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2159.
467. M. Kraus in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2370.
468. C. T. O'Connor: "Oligomerization," in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2381.
469. C. T. O'Connor, M. Kojima, *Catalysis Today* **6** (1989) 329.
470. J. E. Lyons, "Oxidation of Hydrocarbons in the Liquid Phase" in Appl. Ind. Catal., Vol. 3, Academic Press, NewYork, 1984.
471. W. Keim: *Industrial Aspects of Selectivity Applying Homogeneous Catalysis,* Kluwer Acad. Publ., New York, 1991.

472. A. Chauvel, B. Delmon, W. Hölderich, *Appl. Catal.* **115** (1994) 173.
473. M. Misono, N. Nojiri, *Appl. Catal.* **64** (1990) 1; **93** (1993) 103.
474. R. K. Grasselli in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2326.
475. W. Keim, *Angew. Chem.* **102** (1990) 251.
476. W. Kaminsky, M. Arndt in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 5,** Wiley-VCH, Weinheim, 1997, p. 2405.
477. G. Martino, P. Courty, C. Marcilly in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1801.
478. I. E. Maxwell, J. K. Minderhoud, W. H. Stork, J. A. R. van Veen in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2017.
479. D. D. Whitehurst, M. L. Gorbaty in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2078.
480. EP 188 304 A1, 1986 (D. Bode, S. T. Sie).
481. V. Ponec in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1876.
482. C. D. Chang in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1894.
483. M. Iwamoto, H. Yahiro, N. Mizuno, W.-X. Zhang, Y. Mine, H. Furukawa, S. Kagawa, *J. Phys. Chem.* **96** (1992) 9360.
484. J. N. Armor, *Appl. Catal. B: Environmental* **1** (1992) 221.
485. N. Nojiri, M. Misono, *Appl. Catal.* **93** (1993) 103.
486. W. Vielstich, T. Iwasita in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 2090.
487. R. L. Garten, R. A. Dalla Betta, J. C. Schlatter in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 4,** Wiley-VCH, Weinheim, 1997, p. 1668.
488. F. Kapteijn, J. A. Moulijn in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1359.
489. K. C. Pratt in J. R. Anderson, M. Boudart (eds.): *Catalysis – Science and Technology,* **Vol. 8,** Springer, Berlin, 1987, p. 173.
490. R. J. Farrauto, C. H. Bartholomew: *Fundamentals of Industrial Catalytic Processes,* Blackie Academic & Professional, London, 1997, p. 199.
491. J. M. Berty, *Plant Oper. Progr.* **3** (1984) 163.
492. L. K. Doraiswamy, D. G. Tjabl, *Cat. Rev.-Sci. Eng.* **10** (1974) 177.
493. S. Senkan, *Angew. Chem.* **113** (2001) 322; *Angew. Chem. Int. Ed.* **40** (2001) 312.
494. B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, H. W. Weinberg, *Angew. Chem.* **111** (1999) 2649; *Angew. Chem. Int. Ed.* **38** (1999) 2494.
495. J. N. Cawse, *Acc. Chem. Res.* **34** (2001) 313.
496. C. Hoffmann, H.-W. Schmidt, F. Schüth, *J. Catal.* **198** (2001) 348.
497. M. Baerns, E. Körting in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 1,** Wiley-VCH, Weinheim, 1997, p. 419.
498. D. Wolf, O. V. Buyevskaya, M. Baerns, *Appl. Catal. A: General* **200** (2000) 63.
499. K. Kochloefl, *Chem. Eng. Technol.* **24** (2001) 3.
500. B. Cornils, W. A. Herrmann, R. Schlögl, C.-H. Wong: *Catalysis from A to Z,* Wiley-VCH, Weinheim, 2000, p. 105.
501. J. Weitkamp in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1376.
502. J. T. Gleaves, J. R. Ebner, T. C. Kuechler, *Catal. Rev.-Sci. Eng.* **30** (1988) 49.
503. O. V. Buyevskaya, M. Rothaemel, H. W. Zanthoff, M. Baerns, *J. Catal.* **150** (1994) 71.
504. G. Creten, D. S. Lafyatis, G. F. Froment, *J. Catal.* **154** (1995) 151.
505. G. Eigenberger in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1399.
506. G. F. Froment, K. B. Bischoff: *Chemical Reactor Analysis and Design,* Wiley, New York, 1990.
507. K. R. Westerterp, W. P. M. van Swaaj, A. A. C. M. Beenackers: *Chemical Reactor Design and Operation,* Wiley, New York, 1984.
508. K. R. Westerterp, *Chem. Eng. Sci.* **47** (1992) 2195.
509. J. Werther, H. Schoenfelder in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1426.

510. D. Geldart (ed.): *Gas Fluidization Technology,* Wiley, Chichester, 1986.
511. J. F. Davidson, R. Clift, D. Harrison: *Fluidization,* Academic Press, London, 1985.
512. M. Pell: *Gas Fluidization,* Elsevier, Amsterdam, 1990.
513. A. A. C. M. Beenackers in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1444.
514. L. K. Doraiswamy, M. M. Sharma: *Heterogeneous Reactions,* **Vol. 2,** Wiley, New York, 1984, p. 9.
515. B. Jager, R. Espinoza, *Catal. Today* **23** (1995) 17.
516. W. Ehrfeld, V. Hessel, H. Löwe: *Microreactors,* Wiley-VCH, Weinheim, 2000.
517. K.-F. Jensen, *Chem. Eng. Sci.* **56** (2001) 293.
518. S. J. Haswell, R. J. Middleton, B. O'Sullivan, V. Skelton, P. Watts, P. Styring, *J. Chem. Soc. Chem. Commun.* 2001, 391.
519. P. D. I. Fletcher, S. J. Haswell, *Chem. Br.* **35** (1999) 38.
520. Y. Sh. Matros, G. A. Bunimovich in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1464.
521. A. Renken, *Int. Chem. Eng.* **33** (1993) 61.
522. J.-A. Dalmon in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1387.
523. H. P. Hsieh, *Catal. Rev.-Sci. Eng.* **33** (1991) 1.
524. G. Sarraco, V. Specchia, *Catal. Rev.-Sci. Eng.* **36** (1994) 305.
525. R. Soria, *Catal. Today* **25** (1995) 285.
526. G. Donati, N. Habashi, I. Miracca, D. Sanfilippo in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1479.
527. D. B. Keyes, *Ind. Eng. Chem.* **24** (1932) 1096.
528. D. F. Othmer, *Ind. Eng. Chem.* **33** (1941) 1106.
529. P. E. Sauvage in G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* **Vol. 3,** Wiley-VCH, Weinheim, 1997, p. 1339.
530. A. Baiker, *Chem. Rev.* **99** (1999) 453.
531. P. G. Jessop, W. Leitner (eds.): *Chemical Synthesis Using Supercritical Fluids,* Wiley-VCH, Weinheim, 1999.
532. R. Wandeler, A. Baiker, *Cattech* **4** (2000) 128.
533. C. N. Satterfield, T. K. Sherwood: *Role of Diffusion in Catalysis,* Addison-Wesley, Reading, Mass., 1963, p. 56.
534. C. N. Satterfield: *Mass Transfer in Heterogeneous Catalysis,* MIT Press, Cambridge, Mass., 1970, p. 129.
535. E. W. Thiele, *Ind. Eng. Chem.* **31** (1939) 916.
536. G. Damköhler, *Chem. Ing.* **3** (1939) 430.
537. Y. B. Zeldowitch, *Acta Physicochim. USSR* **10** (1939) 582.
538. P. B. Weisz, *Adv. Catal.* **13** (1962) 137.
539. C. D. Prater, *Chem. Eng. Sci.* **8** (1958) 284.
540. P. B. Weisz, C. D. Prater, *Adv. Catal.* **6** (1954) 143.
541. E. Wicke, *Angew. Chem.* **19** (1947) 57.
542. E. Wicke, *Z. Elektrochem.* **60** (1956) 774.
543. R. Aris: *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts,* **Vols. 1 and 2,** Clarendon Press, Oxford, 1975.
544. J. J. Carberry in J. R. Anderson, M. Boudart (eds.): *Catalysis – Science and Technology,* **Vol. 8,** Springer, Berlin, 1987, p. 131.
545. P. B. Weisz, J. S. Hicks, *Chem. Eng. Sci.* **17** (1962) 265.
546. A. Wheeler, *Adv. Catal.* **2** (1951) 250.
547. J. R. Rostrup-Nielsen in C. H. Bartholomew, J. B. Butt (eds.): *Catalyst Deactivation,* Elsevier Science, Amsterdam, 1991.
548. J. Barbier in J. Oudar, H. Wise (eds.): *Deactivation and Poisoning of Catalysts,* M. Dekker, New York, 1985.
549. C. H. Bartholomew, P. K. Agrawal, J. R. Katzer, *Adv. Catal.* **31** (1982) 135.
550. P. Dufresne, A. Quesada, S. Miguarel in D. L. Trimm, S. Akasheh, M. Absi-Halabi, A. Bishara (eds.): *Catalysis in Petroleum Refining,* Elsevier Science, Amsterdam, 1990.
551. K. Tanabe, M. Misono, Y. Ono, H. Hattori: *New Solid Acids and Bases,* Elsevier, Amsterdam, 1989.
552. D. L. Trimm, *Chem. Eng. Process* **18** (1984) 137.
553. D. C. McCulloch in B. Leach (ed.): *Applied Industrial Catalysis,* **Vol. 1,** Academic Press, New York, 1983, p. 103.
554. Y. Huang, N. W. Cant, J. Guerbios, D. L. Trimm, A. Walpole: *Proc. Third Intern. Congress on Catal. and Automotive Pollution Control, Brussels,* Elsevier, Amsterdam, 1995, p. 56.
555. C. A. Bernardo, D. L. Trimm, *Carbon* **17** (1979) 115.

Hexachlorocyclohexane → Insect Control

Cationic Dyes

HORST BERNETH, Bayer AG, Leverkusen, Federal Republic of Germany

1. Definition and Classification 1
1.1. Dyes with Delocalized Charge 1
1.2. Dyes with Localized Charge 3
2. Applications 4
2.1. Dyeing of Synthetic Fibers 4
2.2. Dyeing of Paper, Leather, and Other Substrates 5
3. Economic Aspects 6
4. References 6

1. Definition and Classification

Cationic dyes carry a positive charge in their molecule. The salt-forming counterion is, in most cases, the colorless anion of a low molecular mass inorganic or organic acid. Many of these dyes can be converted into water-insoluble dye bases by addition of alkali. For this reason, they were formerly called *basic dyes*; although still in use today, the term should be abandoned.

The positive charge of cationic dyes may be either localized or delocalized. In (**1**), the positive charge is localized at an ammonium group:

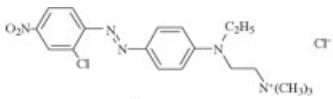

1

In the diazahemicyanine dye **2**, the positive charge is delocalized across the dye cation.

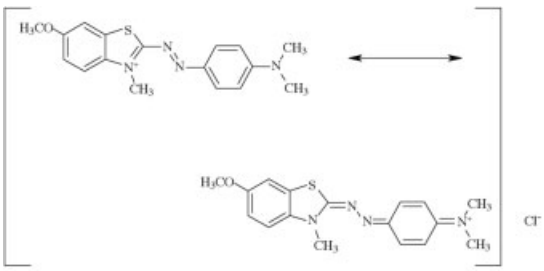

2

As in these examples, the charge-carrying atom is usually nitrogen, but in some dyes this function is adopted by an oxygen, sulfur, or phosphorus atom.

This article gives a short survey on cationic dyes and shows typical representatives of each class of cationic dye. In general, dyes are classified according to their chemical structure in this encyclopedia. For further detail the reader is referred to the following individual keywords:

Anthraquinone Dyes and Intermediates, Azine Dyes, Azo Dyes, Benzoquinone and Naphthoquinone Dyes, Fluorescent Dyes, Methine Dyes and Pigments, Naphthalimide Dyes and Pigments, Nitro and Nitroso Dyes, Optical Brighteners, Phthalocyanines, Triarylmethane and Diarylmethane Dyes.

1.1. Dyes with Delocalized Charge

Cationic dyes with delocalized charge are classified with the *methine dyes*. They may be viewed as vinylogous amidinium salts (**3**) [1]:

3

All of these dyes may be represented by various resonance formulas. They are characterized by a high color strength.

Dyes in which both charged terminal atoms are not part of a heterocyclic ring are called *streptocyanine dyes*. An example is (**4**), which dyes paper in a greenish yellow shade [2]:

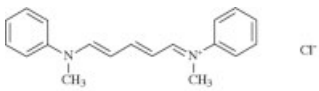

4

In the *hemicyanine dyes*, one of the charged terminal atoms is part of a heterocyclic ring, and the other is a nitrogen atom directly linked to the methine chain (**5**):

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_369

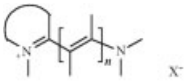

5

An example of this type is the *enamine dye* (**6**), which confers a yellow color to polyacrylonitrile fibers [3]:

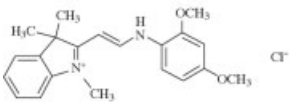

6

Insertion of an aryl residue between the second nitrogen atom and the methine chain leads to the *styryl dyes*, which dye polyacrylonitrile in red and purple shades, e.g., (**7**) [4]:

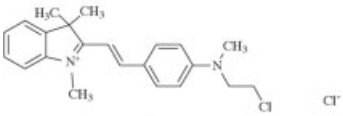

7

In a special form of this type, the chain may be completely missing. This leads to *zeromethine dyes* such as (**8**) [5], which dye polyacrylonitrile in very lightfast violet and blue shades.

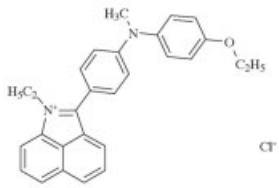

8

If tazhe second charged terminal atom is also part of a heterocyclic ring, the *cyanine dyes* are obtained (**9**).

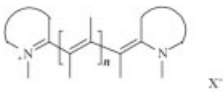

9

These dyes are used predominantly as photographic spectral sensitizers. Compound (**10**) dyes paper in brilliant pink shades [6]:

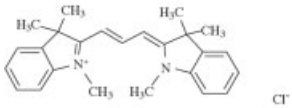

10

One or several of the carbon atoms in the chain may be replaced by nitrogen, as is indicated by the prefix "aza." Depending on the number of nitrogen atoms introduced into the chain, one may obtain either *mono-* or *diazadimethine dyes*. *Tri-* and *tetraazamethine* dyes are also known. An example is dye (**11**), which gives polyacrylonitrile fibers a golden yellow tint [7]:

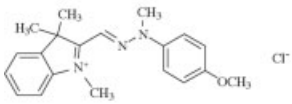

11

Another example is the above-mentioned dye **2**, which dyes polyacrylonitrile in lightfast blue shades [8].

Diphenylmethane and *triphenylmethane dyes* are monomethine dyes with two or three terminal aryl systems; of these at least one but preferably two or three are substituted by a donor group *para* to the methine carbon atom. The most important donor is the amino group. Important dyes belonging to this class include the well-known >malachite green (**12**) [9] and crystal violet (**13**)[10], which are some of the oldest synthetic cationic dyes:

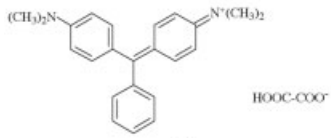

12

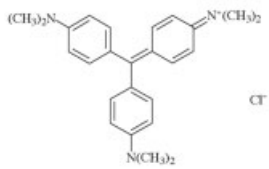

13

Introduction of an oxygen bridge into the triphenylmethane dye molecule leads to the *xanthene dyes.* The shade is shifted from blue to red. The restricted rotation of the phenyl residues inhibits radiationless de-excitation and gives rise to very strong fluorescence. Rhodamine B (**14**) is used for dyeing paper [11] and as a laser dye:

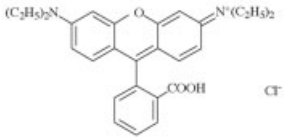

14

Azine dyes, which have the general formula (**15**) are ring-closed azadiphenylmethane or indamine dyes. Examples are *phenazine dyes, oxazine dyes,* and *thiazine dyes*.

15

where Y = -O-,-S-, or -NR-
R = hydrogen, alkyl, aryl

The phenazine dye mauveine (**16**), discovered by PERKIN [12], was the first synthetic dye produced on an industrial scale.

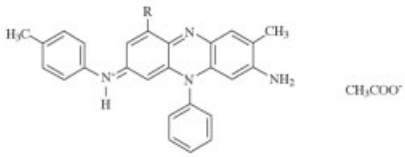

mixture with R = H, CH_3

16

Of current importance are the oxazine dye (**17**)[13] and the thiazine dye (**18**) (methylene blue) [14]:

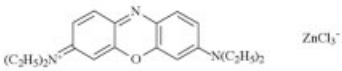

17

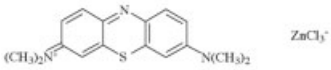

18

1.2. Dyes with Localized Charge

In contrast to dyes with delocalized charge, in which the charge is an essential part of the chromophore, in dyes with localized charge the charged atom in most cases is isolated from the chromophore by a nonconjugated group. Therefore, any chromophore can be used, and this chromophore is not or only slightly influenced by the charged atom.

Representatives of cationic dyes with localized charge are found in the following classes of dyes:

1) *Azo Dyes.* The localized charge in cationic azo dyes may reside either in the diazo moiety (formula **19**) [15] or in the coupling component (formula **20**) [16]:

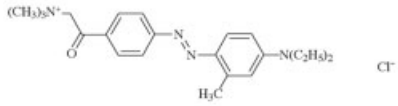

19

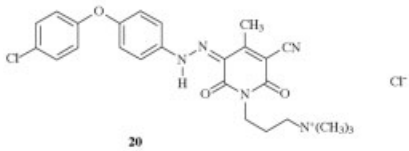

20

2) *Anthraquinone dyes* [17]:

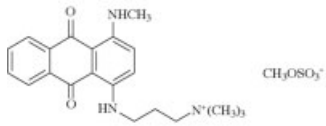

21

3) *Phthalocyanine dyes* [18]:

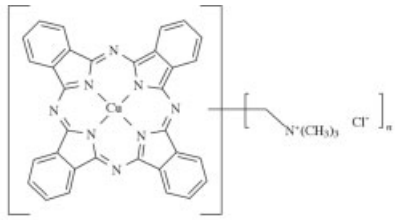

22

where n = 2-8

4) *Perinone dyes* [19]:

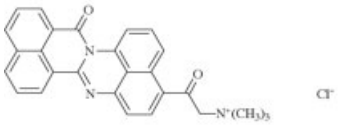

23

5) *Naphthalimide dyes* [20]:

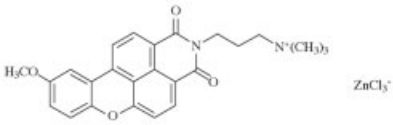

24

6) *Quinophthalone dyes* [21]:

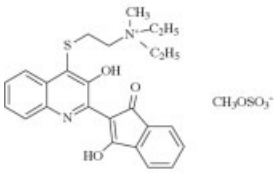

25

7) *Neutrocyanine dyes* [22]:

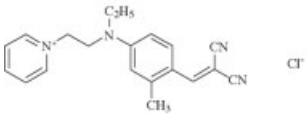

26

8) *Nitro dyes* [23]:

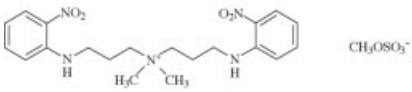

27

2. Applications

Cationic dyes were used initially for dyeing of silk, leather, paper, and cotton mordanted with tannin, as well as for the production of ink and copying paper in the office supplies industry. Their high brilliance and color strength partly compensated for their poor lightfastness on these materials. With the development of synthetic fibers their most suitable substrates were found, and their importance for dyeing textiles increased greatly. Cationic dyes were employed initially for printing of cellulose acetate but gained much greater importance for dyeing of polyacrylonitrile, acid-modified polyester, and polyamide fibers.

2.1. Dyeing of Synthetic Fibers

GLENZ and BECKMANN [24] were the first to investigate color balance and surface potentials in polyacrylonitrile fibers. On the basis of their results, they proposed the following dyeing mechanism: The cationic chromophore is first absorbed by the negatively charged fiber surface and then diffuses from there, at elevated temperature, into the interior of the fiber; there it binds to active acid groups, the number of which is limited and the accessibility of which depends on temperature and fiber constitution. Therefore, the dyeing characteristics of a cationic dye are determined by affinity and diffusibility.

The *affinity* of cationic dyes is enhanced by increasing the size of the molecule and especially by introducing aromatic residues. The phenyl-substituted dye (**28**) is absorbed onto polyacrylonitrile materials at a higher rate than the dye (**11**) [25].

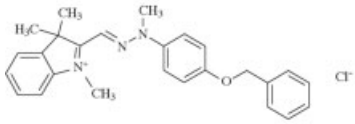

28

Residues which make the substance more hydrophilic, such as carboxyl and hydroxyethyl groups, lower the affinity of cationic dyes for polyacrylonitrile. Therefore, dye (**29**) is absorbed onto polyacrylonitrile materials much more slowly than dye **2** [26].

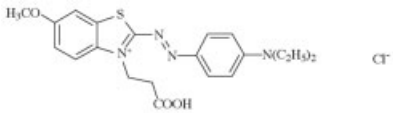

29

High affinity may, however, lead to problems in leveling because these dyes are bound quickly

and nearly irreversibly to the acid groups of the fiber.

Diffusibility increases with decreasing molecular mass. Dyes with a cation mass of less than 275 migrate well and are suitable for the production of level dyeings, especially in the lighter shades [27].

Cationic dyes are also used to dye polyacrylonitrile fibers during the spinning process either from dimethylformamide solution or in the gel stage subsequent to aqueous spinning processes.

Du Pont started production of polyacrylonitrile fibers in the United States in 1944; ten years later Bayer started production in Western Europe. Total fiber production amounted to 10 000 t in 1950, increased to 129 000 t by 1960, and reached 1 005 000 t in 1970. Production doubled within the next 10 years and reached 2 577 000 t in 1996. Worldwide capacities amount to 3 148 000 t. They are mainly located in Asia (1 584 000 t) and Western Europe (751 000 t). North America, South America, and Eastern Europe (including the former Soviet Union) each have capacities in the range of 230 000 to 300 000 t. The largest producing company of polyacrylonitrile fibers is Enichem/Montefibre in Italy and Spain (320 000 t). Japan has the largest capacity (444 000 t), followed by Germany, Italy, and China with 255 000 to 350 000 t. Acrylic fibers represent 6.5 % of world textile fiber usage.

In Western Europe, the bulk of acrylic fibers is targeted for the clothing industry. As much as two-thirds of the acrylic fiber production is used in this sector, whereas only one-third is used for home furnishings and in technical applications.

2.2. Dyeing of Paper, Leather, and Other Substrates

Cationic dyes exhibit good affinity for negatively charged wood pulp and unbleached pulp grades. Anionic fixing agents improve their wetfastness. Cationic dyes are chosen because of their brilliance and intensity, which makes them especially valuable for recycled paper grades. Their poor lightfastness is not a maior problem because of the low light stability of the substrate itself. They suffer from poor affinity for bleached, lignin-free pulp grades, which are used, for instance, in the manufacture of table cloths and napkins. Therefore, dyes have been developed for this application that bear several cationic charges linked by substantive moieties, such as urea or triazine [28], e.g., **30**.

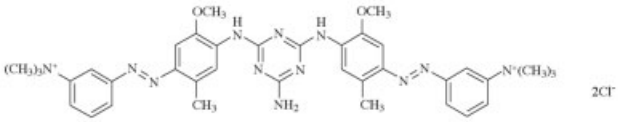

30

Cationic dyes were the first synthetic organic dyes employed for dyeing leather, initially for dyeing of vegetable tanned leather. Brilliant cationic dyes are used for enhancing the brilliance of the shades obtained with 1/2 metal complex dyes.

Because of their coloring strength, cationic dyes are used in the office supplies industry for the manufacture of ink, typewriter ribbon, and copying paper.

Colorless derivatives of the carbinol bases of cationic dyes are used for the manufacture of carbonless copy paper and for special paper used for thermal printers. Examples of such color formers are the crystal violet derivative **31** [29], which gives a blue color development, and the black-developing fluorane lactone **32** [30] and benzoxazine **33** [31].

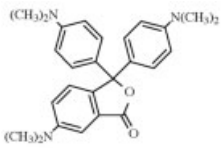

31

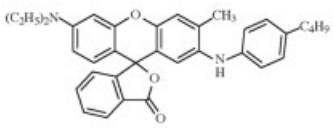

32

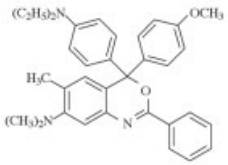

33

Cationic dyes which absorb in the near infrared part of the spectrum are used in optical data storage disks, e.g., **34** [32]:

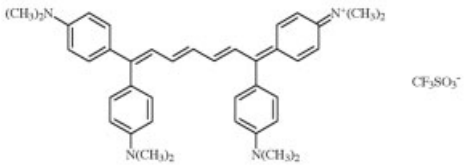

34

3. Economic Aspects

In assessing the economic importance of cationic dyes, it is not possible to distinguish between dyes bearing localized and delocalized charges. Worldwide, 20 000 t of cationic dyes was used in 1995 to dye polyacrylonitrile. Some additional quantities, which cannot exactly be quantified, are used in China, India, and the former Soviet Union. Dominant regions are Western Europe including Turkey, the United States, Canada, and Mexico. In Japan usage is decreasing. Usage increased most rapidly in developing nations. Market leaders are DyStar, Ciba, BASF, and Yorkshire, who have a total market share of 58 %.

More than one-third of the worldwide tonnage consists of the following dyes: Basic Yellow 28 (Astrazon Golden Yellow GL [*36528-10-6*]: 2200 t, Basic Red 18 (Astrazon Red GTL) [*25198-22-5*], and Basic Red 54 (Maxilon Red 2GL) [*36984-04-6*]: 600 t, Basic Red 46 (Maxilon Red GRL) [*12270-28-9*], [*29508-47-2*]: 1100 t, Basic Blue 41(Maxilon Blue GRL) [*26850-47-5*]: 1000 t, BasicBlue 3 (Basacryl Blue X-5G) [*4444-00-2*]: 300 t, Basic Green 4 (malachite green) [*569-64-2*]: 2000 t. An additional 5000 t of yellow, red, and blue dyes are used in mixturesto yield black shades. Dyes with special coloristic properties such as brilliance, fluorescence, and dye bath stability, make up another 7800 t.

In 1996, 16 000 t of cationic dyes with a total value of DM 210 $\times$ 10^6 were used for dyeing paper worldwide. One-half of this are yellow dyes. The other half is split up as follows: orange: 1000 t, red: 2000 t, blue: 1200 t, violet: 2200 t, green: 1000 t. Market leaders are BASF, Zeneca, and Bayer, who share 72 % of the market. There is no trend for a shift to developing countries.

In 1996, 6500 to 7000 t of color formers were used worldwide, 60 % in the production of carbonless copy paper and 40 % in the production of paper for thermal printers, e.g., in fax machines, tills, and ticket printers. Market leaders are Ciba, Yamada, Yamamoto, and Hilton-Davis, who share 90 % of the market.

4. References

1. H. A. Staab, *Einführung in die Theoretische Organische Chemie,* 4th ed., Verlag Chemie, Weinheim, 1964, pp. 236 – 336, and references cited therein. S. Dähne *Z. Chem.* **11,** (1965) 441.
2. Bayer, FR 1 490 350, 1965.
3. IG-Farbenind., DE 686 198, 1935 (K. Winter, N. Roh, P. Wolff, G. Schäfer).
4. IG-Farbenind./Bayer, DE 891 120, 1935 (W. Müller, C. Gerres).
5. Bayer, DE 1 190 126, 1962 (A. Brack, D. Kutzschbach).
6. Bayer, DE 410 487, 1922 (W. König); DE 415 534, 1923 (A. Blömer); IG Farbenind., DE 641 799, 1928 (K. Schmidt).
7. Bayer, DE 1 083 000, 1957 (R. Raue).
8. Geigy, DE 1 044 023, 1954, (W. Bossard, J. Voltz, F. Favre); DE 1 050 940, 1954 (W. Bossard, J. Voltz, F. Favre).
9. O. Fischer, *Ber. Dtsch. Chem. Ges.* **10** (1877) 1625; **11** (1878) 950; **14** (1881) 2520.
10. BASF, DE 26 016, 27 032, 27 789, 1883 (Kern, Caro) Meister, Lucius und Brüning (today Hoechst Aktiengesellschaft), DE 34 463, 1884.
11. BASF, DE 44 002, 1887 (M. Ceresole). Meister, Lucius und Brüning (today Hoechst Aktiengesellschaft), DE 48 367, 1888.
12. W. H. Perkin, *J. Chem. Soc.* **69** (1896) 596, and references cited therein; O. Meth-Cohn, M. Smith, *J. Chem. Soc. Perkin Trans. 1* **1994,** 5.

13. Hoechst, DE 1 569 604, 1966 (G. Schäfer, N. Ottawa).
14. BASF, DE 1886, 1877 (Caro).
15. Du Pont, DE 1 054 616, 1954 (S. N. Boyd, Jr.).
16. Sandoz, DE 2 118 536, 1970 (R. Entschel, C. Müller, W. Steinemann); BE 756 820, 1970; CH 549 629, 1971 (R. Entschel, C. Müller, W. Steinemann).
17. Bayer, DE 1 150 652, 1955 (G. Gehrke, L. Nüssler).
18. Allied Chem. Corp., US 3 565 570, 1968 (C. K. Dien).
19. Ciba, FR 1 489 639, 1965.
20. Hoechst, DE 2 122 975, 1971 (H. Tröster, K. Löhe, E. Mundlos, R. Mohr).
21. Sandoz, DE 2 055 918, 1969 (B. Gertisser, F. Müller, U. Zirngibl).
22. Eastman Kodak, DE 1 250 947, 1963 (J. G. Fisher, D. J. Wallace, J. M. Straley).
23. Hoechst, DE 1 941 376, 1969 (K. Lohe).
24. O. Glenz, W. Beckmann, *Melliand Textilber.* **38** (1957) 296.
25. Bayer, DE 2 040 872, 1970 (H.-P. Kühlthau, R. Raue).
26. Sandoz, FR 2 013 732, 1969 (R. Entschel, V. Käppeli, C. Müller).
27. Ciba-Geigy, DE 2 548 009, 1974 (J. Koller, M. Motter, P. Moser, U. Horn, P. Galafassi).
28. Bayer, DE 2 933 031, 1979 (H. Nickel, F. Müller, P. Mummenhoff).
29. BASF, DE 1 962 881, 1969 (D. Leuchs, V. Schabacker).
30. Nisso Kako, JP 50 82 126, 1971; JP 48 62 506, 1971; DE 2 242 693, 1971 (M. Yagahi, T. Suzuki, T. Igaki, S. Kawagoe, S. Horiuchi). Yamamoto Kagaku, JP 51 15 445, 1971; JP 5 10 6561, 1974; DE 2 202 315, 1971 (Y. Hatano, K. Yamamoto). Nat. Cash Reg., DE 2 155 987, 1970 (Ch.-H. Lin). Fuji, DE 2 025 171, 1969 (S. Kimura, T. Kobayashi, S. Ishige, M. Kiritani).
31. Bayer, EP 0 187 329, 1985 (H. Berneth); EP 0 254 858, 1986 (H. Berneth, G. Jabs).
32. Minnesota Mining and Manufacturing, WO 83/2 027, 1982 (V. L. Bell, I. A. Ferguson, M. J. Weatherley).

Caulking Materials → Sealing Materials
Caustic Soda → Sodium Hydroxide
Cellophane → Films
Cells, Electric → Batteries
Celluloid → Cellulose Esters

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Cellulose

Standard Article

Hans Krässig[1], Josef Schurz[2], Robert G. Steadman[3], Karl Schliefer[4], Wilhelm Albrecht[5], Marc Mohring[6], Harald Schlosser[6]

[1]Seewalchen, Austria
[2]University of Graz, Graz, Austria
[3]La Trobe University, Bundoora, Victoria, Australia
[4]Textilforschungsanstalt Krefeld, Krefeld, Federal Republic of Germany
[5]Wuppertal, Federal Republic of Germany
[6]J. Rettenmaier & Söhne GmbH + Co, Rosenberg, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a05_375.pub2
Article Online Posting Date: April 30, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Cellulose
1.1. Properties
1.1.1. Molecular Structure
1.1.2. Supermolecular Structure (Texture)
1.1.3. Physical Properties
1.1.4. Chemical Properties
1.2. Occurrence
1.3. Production
1.4. Quality Testing
1.5. Applications
2. Natural Cellulosic Fibers
2.1. Cotton
2.1.1. Molecular Arrangement, Morphology, and Fine Structure
2.1.2. Properties
2.1.3. Production
2.1.3.1. Harvesting
2.1.3.2. Ginning
2.1.3.3. Byproducts
2.1.3.4. Processing
2.1.3.5. Finishing
2.1.3.6. Special Finishes
2.2. Bast (Soft) Fibers
2.3. Jute
2.4. Flax
2.5. Hemp
2.6. Leaf (Hard) Fibers
2.7. Coir
2.8. Ramie
2.9. Economic Aspects
2.10. Occupational Health
3. Regenerate Cellulose
3.1. Viscose Fibers
3.1.1. Principle of the Viscose Process

3.1.2. Viscose Preparation
3.1.3. Viscose Fiber Spinning
3.1.4. Fiber Types
3.1.5. Modified Viscose Fibers
3.1.6. Fiber Properties
3.1.7. Uses
3.1.8. Economic Aspects
3.2. Lyocell Fibers
3.2.1. Principles of the Lyocell Process
3.2.2. Process Description
3.2.3. Fiber Properties
3.3. Cuprammonium Fibers
3.4. Tentative Cellulose Fiber Production by Other Processes - Outlook
4. d Cellulose/Powdered Cellulose
4.1. Production
4.2. Properties
4.3. Uses

[Top of Page]

1. Cellulose

Hans Krässig and Josef Schurz

Cellulose [9004-34-6] deserves a special position among the industrially used raw materials for two general reasons. First, cellulose belongs to the natural products which, when used carefully, are inexhaustible since it is regularly regenerated by nature in relatively short time periods. As long as we ensure that the primary sources of cellulose, forests and cotton plantations, are not damaged by destructive lumbering or overcropping, we can expect regular and significant natural annual reproduction.

According to reference [8], the annual yield of cellulosic matter resulting from photoinitiated biosynthesis amounts to approximately 1.3×10^9 metric tons. A tree produces an average of 13.7 g of cellulose daily. If they were lined up, the cellulose chain molecules formed each day would result in a string of 2.62×10^{10} km in length, or 175 times the distance between the sun and the earth.

In wood, cellulose is part of an ingeniously constructed fiber-reinforced composite in which long, stiff cellulose chain molecules organized in thin fibrils constitute the plant reticulum material held together and protected by hydrophobic lignin acting as binder and encasement.

To isolate cellulose from wood for industrial applications, the wooden composite must be broken up by so-called pulping processes. In these treatments, other wood constituents, such as lignin and hemicelluloses, are to a large extent degraded and dissolved. Thus far, these byproducts have found only limited use. In most cases, wood pulp manufacturers concentrate the waste pulping liquors to concentrates consisting of ca. 50 % solids. The organic matter is used as fuel to produce steam and electric power, while the inorganic pulping chemicals (soda, magnesium, or ammonium base and sulfur dioxide) are simultaneously recovered. These recovery processes have practically solved the long-standing environmental problems of the wood pulp industry.

Both cellulose and lignin are biologically degradable and, thus, ecologically beneficial. They will decompose in the open. Cellulose products, such as paper or cellulosic textiles, will decompose and eventually form valuable humus. In industrial use, environmental problems are not caused by cellulose or lignin but by the chemicals used in the isolation or in subsequent chemical processing and transformation into cellulose derivatives, films, or fibers. Therefore, the long-term task of modern cellulose research will be the development of novel processes which yield no or only a few ecologically harmful byproducts. If these efforts are successful, cellulose will surely maintain and strengthen its position as a renewable and environmentally beneficial, industrially important raw material competing with synthetically produced polymers.

1.1. Properties

1.1.1. *Molecular Structure*

Cellulose is an isotactic β-1,4-polyacetal of cellobiose (4-*O*- β-D-glucopyranosyl-D-glucose). The actual base unit, the cellobiose, consists of two molecules of glucose. For this reason, cellulose can also be considered as a (syndiotactic) polyacetal of glucose.

Basic Structure. The basic chemical formula of cellulose is the following:

$$C_{6P}H_{10P+2}O_{5P+1} \approx (C_6H_{10}O_5)_P \text{ or } (C_6H_{10}O_5)_n$$

where *P* = the degree of polymerization; *n* = the number of units in the chain.

The elemental composition of 44.4 % C, 6.2 % H, and 49.4 % O was already known to PAYEN in 1842 [9]. The molecular

mass of the glucose base unit is m_0 = 162, and the molecular mass of the cellulose polymer is

$$M_r = m_0 P + 18 \approx 162 P$$

Constitutional Formula. HAWORTH [10] first discovered the covalent bonds inside and between the glucose units while STAUDINGER [11] found the final proof for the macromolecular nature of the cellulose molecule.

Conformational Formula. The glucopyranosic ring adopts a $^4c_1$ chair conformation, as revealed by X-ray crystallography and nuclear magnetic resonance studies [12], [13] with glucose. The chair formation in comparison to the tray conformation exhibits a free stabilization enthalpy of G_s = 20.05 kJ/mol [14]. In this conformation, the three hydroxyl groups are positioned in the ring plane while the hydrogen atoms are in a vertical position. It seems only natural to assume that the same conformation also exists in the cellulose molecule.

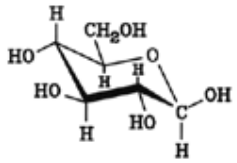

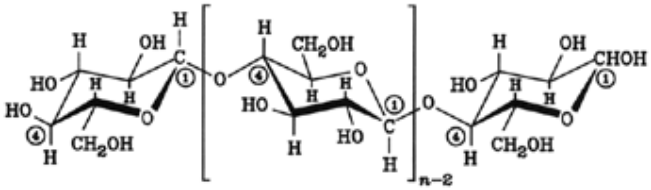

Structural Anomalies. As a naturally occurring polymer, cellulose always contains small amounts of other constituents in addition to glucose (over 99 %). These may already be partially built into or onto the cellulose molecules during biosynthesis, such as lignin – cellulose complexes [15]. Most of the changes in the molecular structure, however, result from secondary reactions, i.e., hydrolysis or oxidation, during isolation from natural sources. For morphological reasons, such chemical changes occur preferably in the accessible interlinking regions between the crystallites of the elementary fibrils or their aggregations. The glucosidic links in these accessible areas, especially if oxidized sites are also present, split 1 000 – 5 000 times faster than glucosidic linkages inside the well-ordered crystallites. The existence of weak links, as proposed in reference [16], is hard to determine. In homogeneous acid hydrolysis, all glucosidic linkages split at the same rate [17].

Cellulose always contains carboxyl groups: In wood pulp, one –COOH group per 100 – 1000 anhydroglucose units (AHG) exists; in cotton, one –COOH group per 100 – 500 AHG units.

Molecular Size. The molecular size of a polymer can be defined by its average molecular mass ($\bar{M}_r$) or its average degree of polymerization ($\bar{P}$); whereby $\bar{M}_r = \bar{P}\, m_0$ (m_0 = molecular mass of the base unit, i.e., of glucose in the case of cellulose).

By investigation of certain physical properties of cellulose or polyhomologous cellulose derivative solutions, the average degree of polymerization can be determined. Table 1 lists the number average degree of polymerization of a number of celluloses of various origin.

Table 1. Degree of polymerization of celluloses of different origin [18]

Type of cellulose	$\bar{P}_n$
Cotton, raw	7 000
Cotton, raw (according to Russian work)	14 000
Cotton, purified	1 500 – 300
Cotton linters	6 500
Flax	8 000
Ramie	6 500
α-Cellulose (isolated from wood fibers)	1 100 – 800
Spruce, pulped	3 300
Beech, pulped	3 050
Aspen	2 500
Fir	2 500
Bacterial cellulose	2 700
Acetobacter cellulose	600

For such physical investigations, cellulose solutions in aqueous copper(II)tetrammonium hydroxide (Schweitzer's reagent; Cuoxam), copper(II)ethylenediamine hydroxide (Cuen), alkaline solutions of the ethylenediamine complexes of cadmium or

nickel can be used. Cellulose trinitrate (CTN) or cellulose tricarbanilate (CTC) solutions in appropriate solvents are also suitable for such studies (see Table 2). In the latter, it should be kept in mind that chain degradation often occurs in substitution reactions performed under unfavorable conditions.

Light scattering studies performed on dilute solutions of cellulose or cellulose derivatives will yield the weight average ($\bar{M}_w$) and osmotic measurements the number average ($\bar{M}_n$) of the molecular mass (or the corresponding average degrees of polymerization: $\bar{P}_w$ or $\bar{P}_n$). Sedimentation experiments in an ultracentrifuge enable the determination of a higher order average molecular mass, the so-called "Z-average" ($\bar{M}_z$). These various quantities are defined as follows:

$$\bar{M}_w = \frac{\sum N_i \cdot M_i}{\sum N_i} \quad \text{or} \quad P_w = \frac{\sum N_i \cdot P_i}{\sum N_i} \quad \text{(weight average)}$$

$$\bar{M}_n = \frac{\sum N_i \cdot M_i^2}{\sum N_i \cdot M_i} \quad \text{or} \quad P_n = \frac{\sum N_i \cdot P_i^2}{\sum N_i \cdot P_i} \quad \text{(number average)}$$

$$\bar{M}_z = \frac{\sum N_i \cdot M_i^3}{\sum N_i \cdot M_i^2} \quad \text{or} \quad P_z = \frac{\sum N_i \cdot P_i^3}{\sum N_i \cdot P_i^2} \quad \text{(Z-average)},$$

whereby $M = 162\,P$; i = fraction $1, 2, 3 \ldots i$; N = number of molecules with $M_1, M_2 \ldots M_i$ or $P_1, P_2 \ldots P_i$.

The simplest and most widely applied practical method for the determination of the degree of polymerization is based on measuring the "intrinsic viscosity η" (Staudinger index). The intrinsic viscosity expresses the reduced viscosity of a solution at an infinitely small concentration. The latter can be derived from the relative viscosity, which is the ratio of the flow time of the dilute polymer solution of a given concentration (t_{ps}) and that of the solvent (t_s) in a capillary viscometer:

$$\eta_{rel} = t_{ps}/t_s$$

$$\eta_{red} = \frac{\eta_{rel} - 1}{c}$$

wherein c = concentration of the cellulose or its derivative in the solution.

The degree of polymerization can be calculated from [η] by using the formula:

$$[\eta] = \lim_{c \to 0} \eta_{red} = K_p P_v^a \text{ (or } K_m M_v^a)$$

The definition of $\bar{P}_v$ (viscosity average of the degree of polymerization) is as follows

$$P_v = \left\{ \frac{w_i P_i^{1+a}}{w_i P_i} \right\}^{1/a}$$

wherein w_i = weight fraction of a molecularly uniform fraction with a degree of polymerization of P_i; $\bar{P}_v$ = viscosity average of the degree of polymerization which for cellulose or cellulose derivative solutions closely resembles the weight average $\bar{P}_w$; and $\bar{M}_v$ = viscosity average of the molecular mass.

Table 2 summarizes some of the more important a, K_m, and K_p values. These values have been obtained by calibration to one of the above-mentioned absolute methods for the determination of the degree of polymerization.

Table 2. Values for the constants *a*, K_m, and K_p [19], [20] (concentration in g/mL)

	Solvent system	*T*, °C	*a*	K_m, ×10^5	K_p, ×10^3
Cellulose	cuoxam	20	0.9	5.54	5.40
Cellulose	cuen	25	0.9	12.5	11.07
Cellulose	cadoxen	20	1.0	3.14	5.09
Cellulose	iron – sodium tartrate	20	1.0	4.08	6.61
CTN	acetone	25	1.0	2.8	4.54
CTN	acetone	25	0.93	7.1	8.06
CTN	butyl acetate	25	1.0	2.8	4.54
CTN	ethyl acetate	25	0.76	37.1	60.10
CTN	ethylene chlorohydrine	25	0.83	15.5	10.57
CTC [20]	acetone	25	0.91	143	146.6
CTC [20]	dioxane	25	0.97	81.3	113.1

Quite often one speaks simply of the "average degree of polymerization" (DP), which is misleading, however, unless the method of determination is properly stated ($\bar{P}_n$, $\bar{P}_w$, $\bar{P}_z$, or $\bar{P}_v$).

Polymolecularity. Cellulose isolated from its native sources is always polydisperse; i.e., it consists, as do all polymers, of a mixture of molecules with the same basic composition and chemical constitution but differing widely in their chain length or degree of polymerization, respectively. The relative amounts of molecules of various lengths present in a given cellulose substrate can be specified by the so-called differential mass distribution curves. Unlike most synthetic polymers, cellulose substrates have complicated mass distribution functions. Figure 1 shows typical mass distributions of various cellulose samples.

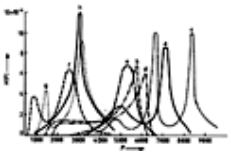

Figure 1. Differential mass distribution curves of various celluloses [18]

a) Cotton; b) Cotton; c) China grass (ramie); d) Flax; e) Ramie; f) Balsam; g) White fir; h) Birch

Knowledge of the molecular mass distribution is important for many applications. However, most of the methods applied in the past, such as fractionation by precipitation or selected dissolution, consume a great deal of time and are often subject to objection. In recent years, gelpermeation chromatography, especially in combination with low-angle laser light scattering, was developed to such a state as to provide a fast and highly reproducible method for this purpose [21].

As a convenient measure of the broadness of the mass distribution, the so-called nonuniformity factor (NU) is often used:

$$\mathrm{NU} = P_w / P_n - 1$$

A polymer with a normal (most probable) molecular mass distribution will have a nonuniformity factor of 1. In most cases, cellulose substrates show much higher molecular nonuniformities.

Secondary Structure. In solution, the cellulose molecules exist in form of largely expanded coils. In addition to isolated and solvated molecules, cellulose and cellulose derivative solutions frequently also contain supermolecular gel particles, so-called micelles [22].

In solid cellulose, high-order microcrystalline structures (“crystalline regions”) alternate with those of a distinctly lower order (“amorphous regions”). Cellulose is polymorph; i.e., depending on the origin or the conditions during isolation or conversion, cellulose will have or can adopt various crystal lattice structures.

Lattice Structure. Native celluloses all show the so-called cellulose I lattice structure. Each unit cell houses two countercurrently arranged cellulose molecules. The lattice of cellulose I is of the monoclinic sphenolitic type. The cellulose chains are in line with the *b* axis of the unit cell. Figure 2 shows a schematic of the unit cell of the cellulose I modification.

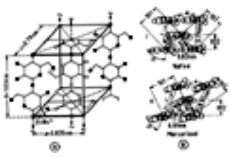

Figure 2. A) Unit cell of the crystal lattice of cellulose I

B) Comparison of the unit cell cross-sections of the native cellulose I and mercerized cellulose II

The cellulose molecules are aligned in the fibrillar axis and form the *b* axis of the unit cell. The length of the *b* axis is 1.03 nm (10.3 Å), somewhat shorter than the extended length of a cellobiose unit, which suggests a slight helical twist in the cellulose chains along the *b* axis [23-25]. This twist is caused by intramolecular hydrogen bonding primarily between the hydroxyl groups on the carbon atom C–3 of one glucose unit and the pyranose ring oxygen in the adjacent glucose unit of the same chain molecule. This intramolecular secondary valence bond is also responsible for the relative rigidity of the cellulose molecule [26]. Some authors [27] suggest a second intramolecular hydrogen bond involving the hydroxyl groups on C–6 and C–2 of adjacent glucose units in the same molecule.

In more recent years, some researchers suggested a unit cell for cellulose I in which the *a* and *c* axis of the Meyer–Mark–Misch model are doubled [28], [29]. However, these newer interpretations of X-ray and electron diffraction results are more or less closely related to the Meyer–Mark–Misch lattice structure. In reference [30], it is claimed that there is a closer agreement with the observed diffraction intensity data for parallel arrangement of the cellulose molecules in the unit cell. However, this is still controversial [31].

The internal cohesion of the cellulose molecules in the unit cells and crystalline domains is due to intermolecular secondary valences — partly hydrogen bonds and partly van der Waal's forces. These bonds can act either between molecules situated in the same crystal lattice plane (intraplanar bonds) or between molecules located in neighboring lattice planes (interplanar bonds). The intraplanar hydrogen bonds are formed primarily between adjacent cellulose molecules in the same 002 lattice planes giving a sheetlike structure. The 002 sheets are then bonded to one another by hydrogen bridges involving the hydroxyl groups on C–6 and the glucosidic ring oxygen atoms of cellulose molecules favorably located in neighboring 002 planes, or by van der Waal's forces acting between neighboring glucopyranose rings.

The unit cells of the other polymorphic structures of cellulose — the most important one being the so-called cellulose II — differ basically in the lengths of their *a* and *c* axis and the angle of inclination β. The cellulose II modification is formed as the thermodynamically most stable polymorph when cellulose fibers are treated with concentrated sodium hydroxide solution (> 14 %) or precipitated (regenerated) from solution.

In addition to the cellulose I and cellulose II modifications, two other polymorphic lattice structure are known, the cellulose III and cellulose IV crystal modifications. The cellulose III structure is formed when the reaction product of native cellulose fibers is decomposed with liquid ammonia. This modification has a lattice structure closely related to that of cellulose II. The cellulose IV modification is obtained by treating regenerated cellulose fibers in hot baths under stretch. The lattice of this polymorph is closely related to that of cellulose I. Some distinct differences in their infrared absorption spectra seem to indicate their existence. Some researchers however doubt their actual existence [32].

Table 3 lists the lattice parameters of the unit cells of these four polymorphic crystal structures.

Table 3. Lattice parameters of the unit cells of the cellulose polymorphs

Type	Source	Dimensions, nm			
		a	b	c	β, degree
Cellulose I	cotton	0.821	1.030	0.790	83.3
Cellulose II	cotton, mercerized	0.802	1.036	0.903	62.8
	viscose fiber	0.801	1.036	0.904	62.9
Cellulose III		0.774	1.030	0.990	58.0
Cellulose IV		0.812	1.030	0.799	90.0

Crystallites. The ability of hydroxyl groups to form secondary valence hydrogen bonds is – together with the stiff and straight chain nature of the cellulose molecule – the cause for the high tendency to organize into crystallites in parallel arrangement and crystallite strands (elementary fibrils), the basic elements of the supermolecular structure of cellulose fibers.

The dimensions of the elementary crystallites differ only slightly for native or regenerated cellulose fibers. Their length ranges between 12 and 20 nm (= 24 – 40 glucose units) and their width between 2.5 and 4.0 nm. The often observed larger “micro- or macrofibrils” (or fragments thereof) are aggregations of elementary fibrils.

Two questions concerning the crystal structure are still under dispute. The first deals with the antiparallel or parallel arrangement of the cellulose molecules in the crystal lattice as previously mentioned. The second question (still open) concerns the existence or nonexistence of folded chains in the lattice [33-35]. While a folded cellulose chain position in the lattice seems unlikely to most experts, the parallel molecule arrangement in the cellulose I lattice is principally acceptable, under the condition that two cellulose II lattice structures exist, one for heterogeneously mercerized native celluloses with parallel arrangement of the molecules and the other for regenerated cellulose substrates with antiparallel molecule arrangement.

1.1.2. *Supermolecular Structure (Texture)*

The basic structural element of cellulose fibers is the so-called elementary fibril. It can be seen with the electron microscope, as illustrated in Figure 3.

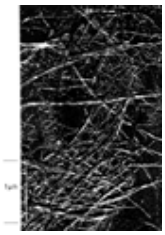

Figure 3. Electron micrograph of the fibrillar nature of cellulose fibers

The cross-dimensions of the elementary fibrils correspond with those of the elementary crystallites. The elementary fibril is a strand of elementary crystals linked together by segments of long cellulose molecules. The lateral order in the interlinking regions is distinctly less pronounced (amorphous). This structure is schematically shown in Figure 4 [36-39].

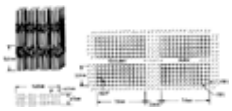

Figure 4. The architecture of elementary fibrils and microfibrils of native celluloses

Several elementary fibrils associate to form larger aggregations of so-called microfibrils and macrofibrils, which can also be seen with a light microscope.

The elementary fibrils and their aggregations are determined by nature in such native fibers as cotton or wood pulp fibers and are laid down in various cell wall layers in a typical manner [18], [19]. Figure 5 shows the structural organisation of wood pulp and cotton fibers.

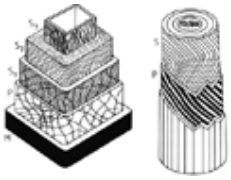

Figure 5. Positioning of the cellulose fibrils in wood (left) and cotton fibers (right)

Wood fibers: M) Middle lamella (lignin and hemicelluloses); P) Primary wall (fibril position unarranged); S_1) Secondary wall I (two or more fibrillar layers crossing one another and positioned spirally along the fiber axis); S_2) Secondary wall II (fibrils wound spirally around the fiber axis); S_3) Secondary wall III (fibrils tightly interlaced)

Cotton fibers: P) Primary wall (interlaced fibrils); S) Secondary wall (fibrils wound spirally around the fiber axis; in distinct distances along the fiber axis the spiral reverses direction)

Synthetic cellulose fibers, such as viscose, do not have a native morphology. Their supermolecular structure can be described as a network of elementary fibrils and their more or less random associations. This is called a “fringe fibrillar” structure [40], which is shown in Figure 6.

Figure 6. Fringe fibrillar model of fiber structure

Structure Characterization. The methods used to characterize the molecular and fine structure of native and synthetic cellulose fibers include the following [41]:

1. determination of the average degree of polymerization ($\bar{P}_n$) by the osmotic method;
2. determination of the average crystallite length by meridional X-ray low-angle scattering on slightly hydrolyzed fiber samples or by measurement of the band width of the meridional 040 X-ray wide-angle reflection at half-maximum intensity;
3. determination of the degree of order (“crystallinity,” CrI) with a method for separating overlapping equatorial X-ray diffractions [42] and deriving from the band width at half-maximum intensity the average cross-dimensions of the crystalline regions; furthermore, this analysis yields information on the lattice structure, polymorphic composition, and accessibility;
4. determination of the degree of orientation by measuring the azimuthal intensity distribution of major equatorial X-ray diffraction arcs or by IR dichroism.

Structure and Properties. Physicomechanical properties of cellulose fibers such as tenacities, elongations, or moduli in the conditioned or wet state are determined by the following structural parameters [43]: 1) the average length of the fiber-forming molecules ($\bar{P}_n$); 2) the average length of the elementary crystallites ($\bar{P}_{nL}$ = number average “limiting” degree of polymerization); 3) the degree of lateral order (crystallinity, CrI); 4) the degree of orientation (f_r) with respect to the fiber axis; and 5) the presence of heterogeneities (natural defects, incorporated gel or sand particles, etc.).

This may be illustrated by the following examples: As shown by Figure 7, the tenacity of the conditioned fibers is determined by the length of the molecules in relation to the length of the elementary crystallites building the elementary fibrils ($1/\bar{P}_{nL} - 1/\bar{P}_n$), by the degree of order (CrI), and by the degree of orientation (f_r).

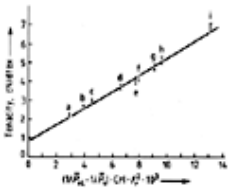

Figure 7. Relation between structure parameters and tenacities of regenerated cellulose fibers

a) Regular viscose fiber; b) Medium-strength viscose fiber; c) High-strength viscose fiber; d) High-wet-modulus viscose fiber; e) Polynosic type viscose fiber; f) Medium-strength viscose tire cord; g) High-strength viscose tire cord; h) Meryl fiber; i) Fortisan fiber

The elongation at break in the conditioned state is mainly dependent on the degree of orientation. Simple geometric considerations give the parameter ($1/\cos\alpha - 1$) in which the angle α derived from the orientation factor (f_r) is the mean angle of deviation of the basic structure elements from the fiber axis. Figure 8 illustrates the relation of this parameter to the breaking elongation of a number of cellulosic fibers.

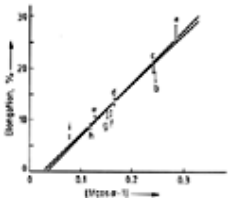

Figure 8. Relation between the degree of orientation and breaking elongation of regenerated cellulose fibers (α = average deviation of structure units from the fiber axis in degree)

a) Regular viscose fiber; b) Medium-strength viscose fiber; c) High-strength viscose fiber; d) High-wet-modulus viscose fiber; e) Polynosic type viscose fiber; f) Medium-strength viscose tire cord; g) High-strength viscose tire cord; h) Meryl fiber; i) Fortisan fiber

The wet moduli of the various fibers show a close relation to the product of the length of the elementary crystallites ($\bar{P}_{nL}$), the

degree of order (CrI), and the square of the orientation factor (f_r^2). This is demonstrated in Figure 9.

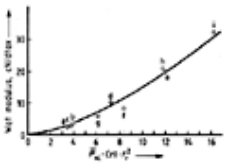

Figure 9. Relation between structure parameters and wet modulus of regenerated cellulose fibers

a) Regular viscose fiber; b) Medium-strength viscose fiber; c) High-strength viscose fiber; d) High-wet-modulus viscose fiber; e) Polynosic type viscose fiber; f) Medium-strength viscose tire cord; g) High-strength viscose tire cord; h) Meryl fiber; i) Fortisan fiber

1.1.3. *Physical Properties*

Cellulose is relatively hygroscopic. Under normal atmospheric conditions (20 °C, 60 % relative humidity), it adsorbs ca. 8 – 14 % water. Cellulose swells in water (see Table 4). It is, however, insoluble in water or dilute acids. In concentrated acids, solution can be achieved under severe degradation. Caustic solutions cause extensive swelling and dissolution of low molecular mass portions ($P \leqq 200$). Solvents for cellulose are listed in Table 2).

Table 4. Swelling of various celluloses in water

Type	Swelling, %
Cotton	18
Viscose, continuous filament	74
Cuprammonium rayon	86
Cellulose triacetate	10
Cellulose tripropionate	2.5
Cellulose tributyrate	1.8
Cellulose trivalerate	1.6
Cellulose tristearate	1.0

Cellulose is nonmelting; thermal decomposition starts at 180 °C; the ignition point is > 290 °C. With chlorine and zinc iodide, cellulose takes on a red-violet to blue color; with phloroglucinol–hydrochloric acid, pure cellulose should not take on a red color (test for residual lignin). Additional data:

Density: 1.52 – 1.59 g/cm^3
Refractive index:
1.62 parallel to the fiber axis
1.54 perpendicular to the fiber axis
Dielectric constant: 2.2 – 7.2 (at 50 Hz) Highly dependent on humidity conditions
Insulation resistance: 10^{14} – 10^{17} Ω cm Highly dependent on humidity conditions
Electric strength: 500 kV/cm
Heat of combustion: 17.46 J/g
Heat of crystallization: 18.7 – 21.8 kJ/mol of glucose
Specific heat: 1.00 – 1.21 J g^{-1} K^{-1}
Coefficient of thermal conductivity:
0.255 kJ m^{-1} h^{-1} K^{-1} (loosely packed) to
0.920 kJ m^{-1} h^{-1} K^{-1} (compressed)
Specific internal surface: 10 – 200 m^2/g

1.1.4. *Chemical Properties*

The chemical reactivity of cellulose is determined to a large extent by the supermolecular structure of its solid state. Most of the reactions on cellulose fibers are heterogeneous in nature. The reaction medium acts on a two-phase solid system: (a) the less-ordered (amorphous) regions which are mainly located on the surface of the elementary fibrils or their aggregations and in the interlinking regions between the elementary crystallites in the fibrils, and (b) the well-ordered elementary crystallites or fused associations of the elementary fibrils. Any reaction will first start on the less-ordered surface of the elementary fibrils or their aggregations (topochemical reaction) and then, under favorable conditions, proceed into the interlinking regions between the elementary crystallites to penetrate from both ends into the crystallites. Therefore, as long as the reaction is limited to the accessible surface of the fibrils or fibrillar aggregations and the regions interlinking the elementary crystallites (i.e., up to degrees of substitution (DS) of 1.3 – 1.7), there is no visible effect in the crystalline structure. At increased degrees of substitution (to ca. DS = 2.5), the X-ray diffractogram shows overlapping diffraction bands of the original cellulose I structure and the cellulose derivative. At still higher degrees of substitution, the pure diffraction pattern of the derivative will finally result. This course of reaction implies that partially substituted cellulose derivatives are always a mixture of completely substituted cellulose, partially substituted portions (of block-polymer nature), and unsubstituted cellulose.

A quasi-homogeneous reaction can be achieved when the fiber structure is loosened by swelling treatments to such an extent that all cellulose molecules can react simultaneously. A real homogeneous reaction can, however, only be achieved

by bringing the cellulose into a molecularly dispersed solution.

The reactivity of cellulose substrates can be greatly enhanced by activation treatments, such as swelling, solvent exchange, inclusion of structure-loosening additives, degradation, or mechanical grinding, which enlarge accessible surfaces by opening fibrillar aggregations. These treatments also restore, in most cases, the loss of reactivity due to so-called hornification, which occurs when water is removed from cellulose by drying under severe conditions.

Swelling with water or other polar liquids is the most frequently applied activation treatment. It exclusively opens the interfibrillar interstices and swells the less-ordered surface and interlinking regions of the fibrillar elements. The solvent exchange technique is a special kind of activation from the water-swollen state. It allows the introduction of media being inert in subsequent reactions that are unable to swell the cellulose substrate thus maintaining the reactive water-swollen state. An interesting variation of the solvent exchange treatment is the so-called inclusion technique [44]. Inert liquids, such as cyclohexane or benzene, are introduced into the cellulose substrate by solvent exchange from the water-swollen state. During drying, they are permanently incorporated into the interfibrillar interstices or voids, thus preventing fusion of fibrils, i.e., the hornification. Such inclusion celluloses are very reactive, as shown in Table 5.

Table 5. Acetylation of vacuum-dried native and mercerized cotton after inclusion of benzene using acetic anhydride – sulfuric acid at 60°C

Time of acetylation	Native		Mercerized	
	P 2500	*P* 340	*P* 2200	*P* 290
	% Acetyl			
Acetylated after vacuum drying:				
24 h	2.0	5.6	0.9	1.3
48 h	3.8	8.3	1.9	2.9
96 h	6.8	12.5	2.5	3.6
192 h	14.2	21.9	3.0	4.8
408 h	37.8	39.9	11.0	14.5
Acetylated after benzene inclusion, benzene wet:				
1 h	19.4	19.1	25.7	24.3
24 h	42.5	44.1	45.0	44.7
Acetylated after benzene inclusion and drying:				
1 h	17.3	16.8	23.9	23.0
24 h	41.5	42.1	43.2	43.1

Another very effective way of activating cellulose fibers is to enhance the accessibility of fibrillar surfaces and to open the less-ordered regions interlinking the crystallites in the fibrils by treatment with systems causing not only interfibrillar, but also intracrystalline, swelling. Some inorganic acids, various salt solutions, and especially certain inorganic and organic bases achieve this at distinct concentrations. They apparently penetrate the fiber through existing capillaries and pores by opening the fibrillar interstices and entering the interlinking regions between the crystallites. From there, they enter the elementary crystallites from both ends and force them open. At suitable concentrations and temperatures, they ultimately cause crystal lattice transfer, particularly with respect to opening the 101-plane distances (see Table 6).

Table 6. Effect of swelling agents on the (101) interplanar spacing of cellulose [45]

Swelling agent	Observed (101) distance, nm
Liquid ammonia	1.03 – 1.06
Methylamine	1.467
Ethylamine	1.572
n-Propylamine	1.848
n-Butylamine	1.973
n-Amylamine	2.192
n-Hexylamine	2.485
n-Heptylamine	2.874
Hydrazine	1.03

Ethylenediamine	1.226
Tetramethylenediamine	1.465
Tetramethylammonium hydroxide	1.30
Ethyltrimethylammonium hydroxide	1.30
Benzyltrimethylammonium hydroxide	1.65
Dibenzyldimethylammonium hydroxide	1.65

So-called mercerization is a frequently used practical method of activation, i.e., the treatment of native cellulose substrates with 10 – 20 % sodium hydroxide solutions at moderate temperature (< 20 °C). In this treatment, sodium cellulose I is formed in which the 101-plane distance is increased from 0.61 nm (6.1 Å) in native cellulose to 1.22 nm (12.2 Å). In this lattice transition, the glucopyranose rings are dislocated and aligned into the 101 lattice plane. The hydroxyl groups on the C–2 and C–6 carbon atoms are thus freely exposed and jut into the widened space between the 101 lattice planes, making them accessible for reactions.

The chemical character of cellulose is determined by the sensitivity of the β-glucosidic linkages between the glucose repeating units to hydrolytic attack and the presence of one primary and two secondary reactive hydroxyl groups in each of the glucopyranose units. These reactive hydroxyl groups are able to undergo exchange, oxidation, and substitution reactions, such as esterification and etherification.

Sorption and Exchange Reactions. Cellulose undergoes sorption and exchange reactions with water and deuterium oxide. These reactions are of special interest since they give a good indication of the accessibility of the cellulose substrate.

The cross-dimensions of the well-ordered regions can be derived from the width of the equatorial X-ray wide-angle diffractions. If it is assumed that the crystallites are surrounded by one layer of disturbed ("amorphous") unit cells and also that the molecules in the next outer layer of the well-ordered crystalline core of the crystallites are in addition accessible for reactants such as water or deuterium oxide, an explanation for the extent of water adsorption and deuterium exchange should be possible, see Figure 10 [46].

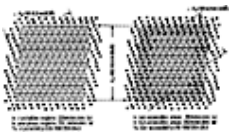

Figure 10. Model of crystallinity (left) and accessibility (right) of fibrils in a high-wet-modulus viscose staple fiber [46]

Table 7 compares the experimentally obtained water sorption and deuterium exchange values with those predicted from structure investigations according to the above outlined concept.

Table 7. Crystallinity and accessibility of cellulose substrates from X-ray diffraction in relation to H_2O adsorption and deuterium exchange

Sample	Relative dimensions [a]			% molecules in:			Deuterium exchange, %	H_2O adsorption	
	101	101	002	Ordered regions	Amorphous regions	Accessible regions		Theoretical [b], %	Experimental, %
Cotton	200	155	230	83	17	47		7.8	8.1
Fortisan	140	100	160	80	20	54	53.5	9.0	9.6
HWM-rayon	80	70	85	65	35	69	72.0	11.5	12.4
Polynosic	95	80	100	70	30	60	61.5	10.0	10.9
Regular rayon	90	75	95	67	33	65	66.5	10.8	11.6

[a] Derived from X-ray wide-angle diffractograms corrected for the band-widening effects of sample size and beam dimension.
[b] Calculated by assuming that accessible glucose units bind 1.5 mol of H_2O each.

The internal surface data calculated from crystallite cross-dimensions are also in good agreement with those determined by gas adsorption.

Degradation by Acid Hydrolysis. Degradation in acidic medium is based on the hydrolysis of the β-glucosidic linkages between the glucose base units. The reaction depends strongly on pH and already proceeds at a remarkable rate at low pH and temperatures well under 100 °C. Initially the acetal oxygen of the glucosidic linkage is protonated. Through heterolysis, an intermediate carbonium ion is formed, causing chain-splitting. The carbonium ion finally reacts with water, which reforms

the proton. The following reaction scheme illustrates the course of this reaction.

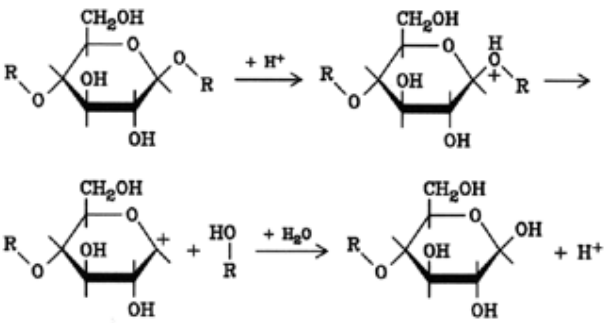

Homogeneous and heterogeneous hydrolysis of cellulose are both first-order reactions. The reaction speed is strongly dependent on the acid and the cellulosic material. The course of homogeneous and heterogeneous hydrolysis also reveals a basic difference. In homogeneous hydrolysis degradation proceeds at a constant rate until all of the cellulose is degraded to cellobiose or glucose, respectively. In heterogeneous hydrolysis, the rate decreases continuously and degradation stops almost completely when the number-average degree of polymerization reaches 25 – 100. This corresponds to the length of the elementary crystallites ("level-off degree of polymerization"). Hydrolysis proceeds thereafter at a very slow rate. It is interesting to note that during the course of homogeneous hydrolysis, the degraded and isolated residue adopts an increasingly normal molecular mass distribution as indicated when the ratio of its weight to the number average degree of polymerization approaches a value of 2. This is indicative of a statistical degradation (see Table 8).

Table 8. Comparison of the changes in molecular non-uniformity during the course of homogeneous and heterogeneous degradation of cellulose [47], [48]

Time, min	$\bar{P}_w$	$\bar{P}_n$	$\bar{P}_w/\bar{P}_n$
Homogeneous hydrolysis in H_3PO_4 at 25 °C			
0	1600	1245	1.29
60	1055	785	1.34
210	610	410	1.50
750	260	150	1.73
2100	100	55	1.82
Heterogeneous hydrolysis with 1 N HCl at 50 °C			
0	1750	1200	1.46
180	800	560	1.43
540	440	310	1.41
1620	250	180	1.38
3240	120	95	1.26

In contrast, the residue in heterogeneous hydrolysis tends toward a value of 1 for the ratio of its weight to number-average molecular mass. This is a strong indication that the course of heterogeneous degradation is not determined solely by the sensitivity of the β-glucosidic linkages, but primarily by morphological aspects [49]. Hydrolytic attack is almost completely limited to the molecules situated on the surface of the fibrillar strands and the accessible molecule segments connecting the crystallites. It is also important to note that mass loss in heterogeneous hydrolysis performed under moderate conditions is relatively small in the initial fast reaction.

Degradation in Alkaline Media. Hydrolysis of the β-glucosidic linkages in alkaline media occurs at a significant rate only at temperatures above 150 °C. It is most probable that chain-splitting proceeds by way of the 1,2-anhydro configuration [50].

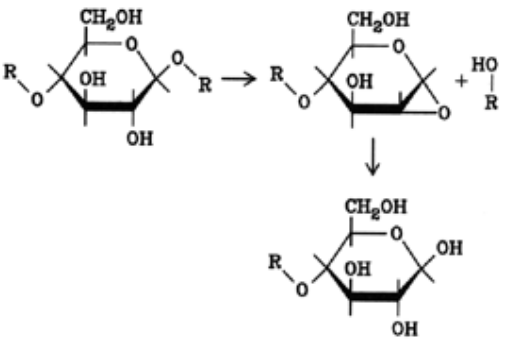

Acidic as well as alkaline hydrolysis of the glucosidic bond is remarkably enhanced (β-elimination) by oxidative changes at the C–2, C–3, or C–6 carbons leading to carbonyl groups. An example of this chain-splitting reaction, which can even occur at moderate temperatures, is [51]:

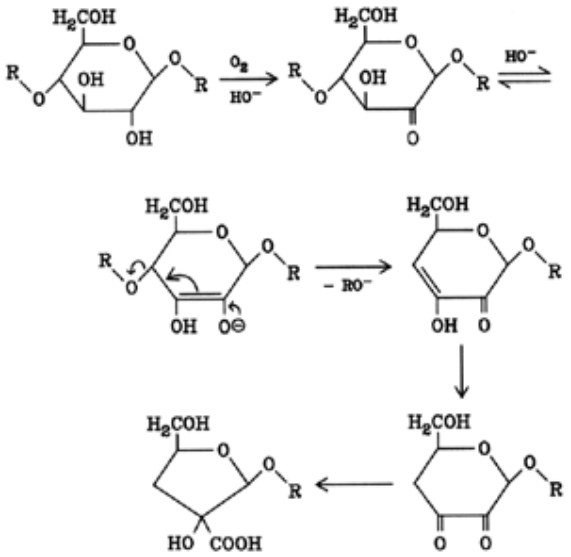

An additional degradation reaction taking place in alkaline media is the so-called peeling reaction. In the course of this reaction, which even takes place at temperatures well below 100 °C, the cellulose chain molecules are degraded step-by-step, beginning at the reducing end and proceeding in a “zipper-like” reaction. The terminal glucose unit is first transformed into the 1,2-enediol, which isomerizes to the corresponding ketose and splits off the chain. The ketose is transformed further into the alkali-stable isosaccharinic acid [52].

The newly formed aldehydic end of the cellulosic chain will repeat the same reaction. When ca. 50 – 60 glucose units are split off under these conditions, the reaction normally stops due to the interference of a chain-stopping reaction. This termination reaction leads by way of the 2,3-enediol to an alkali-stable metasaccharinic acid end group, which stabilizes the cellulose molecules against further degradation.

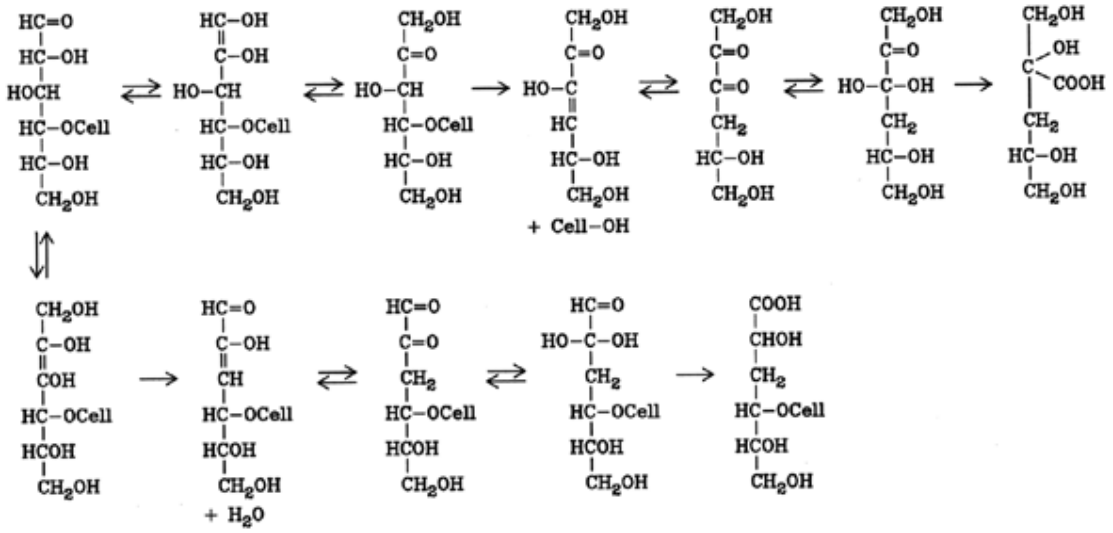

The degradation in acidic as well as in alkaline media is of great importance in the manufacture of pulp from wood and other plants and in the processing of cellulose derivatives, regenerated fibers, and films.

The microbiological degradation of cellulose should also be mentioned in this connection. This degradation takes place through enzymatic hydrolysis of the β-glucosidic linkages and is of interest in connection with the use of plant biomass [53].

Oxidation Reactions. The hydroxyl groups and the aldehydic end groups participate in the oxidation reactions of cellulose. These reactions form aldehyde, ketone, and carboxyl groups. Extensively oxidized and degraded products are designated as oxycelluloses. Some oxidizing agents show specific action. They attack only specific functional groups, forming defined oxidation products. Other oxidants react nonspecifically with all types of oxidizable groups in the cellulose molecules. Under special conditions hypoiodite and chlorite attack only the aldehyde end group on C–1, oxidizing it to form a carboxylic group. Another oxidant with specific action is periodate, which attacks the glycol configuration on the carbon atoms C–2 and C–3, thus causing ring-splitting and forming a dialdehyde structure.

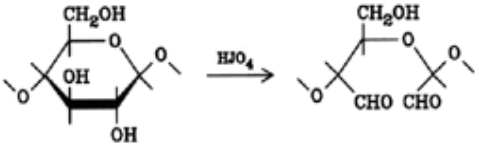

Nitrogen dioxide (dinitrogen tetroxide) reacts not quite as specifically. It oxidizes, however, with a certain preference the hydroxyl group on C–6 to a carboxyl group and to a lesser extent hydroxyl groups on C–2 and C–3 to ketone groups.

Nonspecific oxidants are chlorine, hypochlorite and chromic acid. They also oxidize all accessible hydroxyl groups to aldehyde, ketone, and carboxyl groups. The oxidative action of chlorine and hypochlorite is extensively used as a bleaching agent in the pulp industry. However, one must keep in mind that the introduction of carbonyl groups on C–2, C–3, and C–6 causes alkali instability of glucosidic linkages in the β-position which initiates degradation under alkaline conditions.

Esterification and Etherification. Both of these substitution reactions are used in industry for the manufacture of widely used products (⟶ Cellulose Esters, ⟶ Cellulose Ethers). The acetate and nitrate esters and the methyl and carboxymethyl ethers of cellulose, and to a lesser extent the ethyl and hydroxyethyl ethers, have acquired practical significance.

To perform the esterification or the etherification reaction properly, the hydroxyl groups in the cellulose substrate must be made accessible for the reaction. The supermolecular structure of the cellulose must be activated before or in the course of the substitution reaction. For this purpose the cellulose substrate is treated with strong acids or alkali. In these treatments addition compounds are formed between the cellulose, acid, or alkali and the water present in the system. During the opening of hydrogen bonds between the molecules in the cellulose substrate, more or less defined addition compounds are formed in the crystalline regions. Their formation is often accompanied by changes in the lattice structure, resulting mostly in an increased distance between the 101 planes. This exposes the hydroxyl groups on C–2 and C–6, making them accessible. Hydrated cations or anions of the reactant, respectively, are incorporated into the widened interplanar space, where they initiate and facilitate the reactions. The addition compounds thus formed are only stable in equilibrium with the reactant and will decompose readily when the system is diluted with water.

The basic principle of the esterification and etherification reaction is quite similar. The first step is a nucleophilic substitution or addition and the formation of an oxonium ion on the carbon atom carrying the reactive hydroxyl group. A surplus of esterification or etherification reactant will lead to the formation of the corresponding ester or ether:

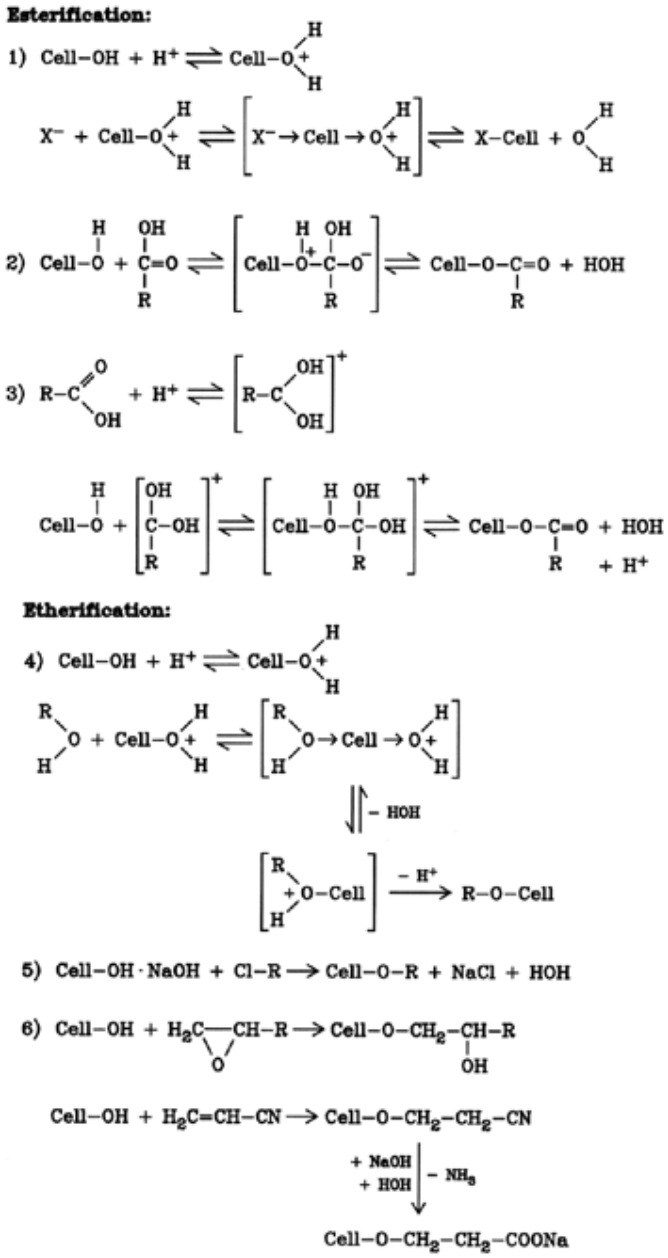

1. General esterification mechanism (formation of an oxonium ion; inorganic acid esterification)
2. Esterification with organic acids (nucleophilic addition)
3. Acid-catalyzed esterification
4. General etherification mechanism (formation of an oxonium ion; alcohol excess leads to ether formation)
5. Etherification of cellulose with alkali consumption
6. Etherification of cellulose without alkali consumption

The esterification reaction is promoted by water-binding catalysts. In etherification, prior swelling of the cellulose substrate with alkali or the transfer to alkaline cellulose is essential for the reaction.

The presence of three hydroxyl groups in each glucose unit allows the formation of mono-, di-, and triesters or ethers, respectively. Contrary to the reaction behavior of primary and secondary hydroxyl groups in low molecular mass alcohols, where the primary hydroxyl group always shows a higher reactivity, the secondary hydroxyl group on C–2 quite often shows preferred reactivity in heterogeneous esterification or etherification of cellulose. Table 9 illustrates this for the relative reaction rate of the hydroxyl groups on C–2, C–3, and C–6 in various etherification reactions [54].

Table 9. Relative etherification velocities at the OH-groups on C–2, C–3, and C–6

Derivative	K_2	K_3	K_6
Methylcellulose	5	1	2
Ethylcellulose	4.5	1	2
Carboxymethylcellulose	2	1	2.5
Hydroxyethylcellulose	3	1	10
Cyanoethylcellulose	3	1	3

The normally expected preferred reactivity of primary hydroxyl groups on C–6 can be observed only in substitution reactions on cellulose in solution. The fact that most substitution reactions on cellulose are performed in heterogeneous systems also has the consequence that most partially substituted cellulose derivatives are actually mixtures of fully substituted, irregularly substituted (block substitution), and unsubstituted cellulose molecules.

The extent of substitution in cellulose derivatives is described by the so-called “degree of substitution” (DS). This value states the average number of substituents linked to each glucose base unit in the cellulose molecules. Since substitution normally occurs irregularly along the cellulose chains, the DS can assume any value between 0 and 3.

Industrially produced cellulose esters and ethers find practical use as fibers, films, lacquers, explosives, adhesives, and as auxiliaries in paper, textile, and food industries. Their properties and, consequently, their applications are primarily dependent on the nature of the substituents, the degree of substitution, the distribution of substituents along the cellulose molecules and from molecule to molecule, and their degree of polymerization.

In particular, the xanthation of cellulose is of great economic importance, i.e., the esterification with dithiocarbonic acid. This reaction is the basis for the so-called viscose process, which is widely used for the manufacture of regenerated cellulose fibers and films (see Chap. Regenerate Cellulose). The cellulose xanthate is obtained in the reaction of alkali cellulose with carbon disulfide and is soluble even in dilute sodium hydroxide solutions at relatively low degrees of substitution (DS < 1).

Graft Copolymerization. During the last two decades, another way of modifying cellulose, the grafting reaction, has found substantial interest. This method allows the attachment of chemically different side-chains to a given polymer molecule. The reaction mechanisms are principally the same as in the synthesis of polymers [55], [56]. The most frequently used method is to initiate grafting with radical catalysts. To achieve graft copolymerization, the cellulose substrate must be in the presence of peroxides or redox systems (i.e., hydrogen peroxide, hydrogen peroxide – iron(III) ions, Ce(IV) ions) and as such is able to undergo polymerization with such compounds as styrene, acrylic acid, acrylic ester, etc. The introduction of substituents that are able to enhance radical transfer, such as thiol or xanthate groups, promotes the grafting reaction. Industrial applications for graft-modified cellulose fibers are thus far very limited.

1.2. Occurrence

Cellulose is one of the main cell wall constituents of all major plants. It occurs there in varying amounts. Cellulose is found in nonlignified (such as cotton) and lignified (such as wood) secondary plant cells and constitutes as such the major portion of all chemical cell components. Table 10 lists the cellulose content of some wood species.

Table 10. Cellulose content of various wood species

Type of wood	Cellulose content, %
Spruce (*Picea abies*)	43
Pine (*Pinus silvestris*)	44
Birch (*Betula verrucosa*)	40
Beech (*Fagus silvatica*)	43
Poplar (*Populus tremuloides*)	53

The cell walls of green algae also contain cellulose. Furthermore, cellulose is also found in the membranes of most fungi and in the cell walls of some flagellates. So-called bacterial cellulose is synthesized by *Acetobacter xylinum* on nutrient media containing glucose. In the animal kingdom one finds cellulose in the tunics of the Tunicatae.

The raw material sources for industrially used cellulose are almost exclusively the fiber cells of more fully developed plants, especially cotton, bast, and leaf fibers for the textile industry, and wood and some other graminaceous plant fibers (i.e., grain straws, bamboo canes, and sugar cane wastes) for the paper industry. Wood consists of up to approx. 40 – 50 % cellulose. Additional wood constituents are lignin, hemicelluloses (i.e., lower molecular mass polysaccharides containing other sugars as monomer units in addition to glucose), extracts (such as resins, gums, fats, waxes, terpenes, etc.), and numerous other native organic compounds in smaller amounts.

1.3. Production

Cotton fibers and wood are the primary raw materials for the production of industrially used cellulose.

Raw cotton fibers and linters contain small amounts of proteins, waxes, pectins, and inorganic impurities in addition to approximately 95 % pure cellulose. These constituents can be removed if they interfere with the intended further processing, with relatively simple procedures. Treatments with hot alkali and subsequent bleaching with hypochlorite, peroxides, or chlorine dioxide are in use.

A sequence of processing steps is necessary for the isolation of relatively pure cellulose (pulp) from wood. To remove most of the lignin and of some of the other wood constituents, a pulping procedure has to be performed. This can either be done with sulfurous acid containing solutions of hydrogen sulfites (i.e., magnesium, calcium, sodium, or ammonium hydrogen sulfite in an excess of sulfurous acid) or with alkaline sodium hydroxide – sodium sulfate solutions at ca. 130 – 180 °C. These

treatments reduce wood chips to fibers by using a slight mechanical action. To further purify and brighten the raw pulp stock for the manufacture of paper and for chemical applications, a multistage refining procedure must follow. These refining steps consist of alkali treatments in the presence of oxidizing agents (chlorine, chlorine dioxide, hydrogen peroxide, oxygen, or air) to remove residual lignin followed by bleach sequences with hypochlorite or hydrogen peroxide. In the production of dissolving pulps for the manufacture of cellulose films, fibers, or cellulose derivatives, extractions with cold or hot alkali are used in addition to remove low molecular mass polysaccharides (hemicelluloses) and to increase the pure cellulose (α-cellulose) content. Depending on the number of refining steps and conditions, wood pulps containing up to ca. 99 % pure cellulose can be obtained. The yield from wood, however, will be ca. 70 % for the production of low cellulose content paper pulps and only 30 % for the production of pulps with high cellulose content. Cellulose content higher than 99 % is practically unachievable since the cellulose would undergo extensive degradation in such an attempt and thus be of no further use. Table 11 lists the cellulose contents (α-cellulose) of some major paper and dissolving pulps.

Table 11. α-Cellulose content of various paper and dissolving pulps

Type	α-Cellulose content, %
Paper pulps, bleached:	
Spruce, sulfite	89
Beech, sulfite	89
Spruce, sulfate	82
Birch, sulfate	72
Sulfite dissolving pulps for:	
Regular rayon staple	89 – 91
High-wet-modulus rayon staple	91 – 93
Acetate filament yarn	94 – 95
Sulfate dissolving pulps for:	
Regular rayon staple	93 – 95
High-wet-modulus rayon staple	95 – 98
Rayon tire cord (super II – III)	96 – 98
Acetate filament yarn	98

1.4. Quality Testing

The determination of the residual impurities and the extent of molecular degradation or other chemical changes of the cellulose is, in most cases, not essential for cotton fibers in the textile field or wood pulps in the paper field.

However, celluloses (cotton linters or wood pulp) which will be used either for the manufacture of regenerated cellulose and cellulose acetate fibers or for cellulose derivatives must undergo careful quality testing.

Because the first step in the further processing quite often consists of a treatment with sodium hydroxide solutions, the testing for alkali solubility is very important. The solubility in 5, 10, and 18 % sodium hydroxide solution can give a preliminary indication of the content of noncellulosic low molecular mass polysaccharides (also including low molecular mass cellulose). A more detailed determination of cellulose and noncellulosic polysaccharide content is possible after complete hydrolytic degradation to the corresponding monomeric sugars and the use of chromatographic methods.

One of the most important characteristics for the chemical processing of cellulose is the degree of polymerization (see Molecular Size.). For industrial testing practice, viscosimetric methods using standardized flow or ball-fall viscometers are quite satisfactory.

Further test methods for dissolving pulps are the determination of residual lignin, of extractables soluble in organic solvents, and of the composition of inorganic ingredients (ash). Standardized methods for all of these tests are described by ISO (International Standards Organization), ASTM (American Society for Testing and Materials), TAPPI (Technical Association of Pulp and Paper Industry, United States), DIN (German Industrial Norms), or Zellcheming (German Association of Pulp and Paper Chemists and Engineers), etc.

1.5. Applications

The greatest portion of industrially used wood cellulose by far (ca. 150×10^6 t/a) is used after partial removal of the noncellulosic constituents in its original fiber form for the production of paper, board, and nonwovens. The same applies to most of the long-haired cotton fibers (ca. 15×10^6 t/a) and their use in the manufacture of textiles. Only a minor portion (ca. 7×10^6 t/a) is used in form of high- α-cellulose wood pulps or cotton linters as starting material for the production of synthetic

cellulosic fibers, primarily viscose and acetate fibers, or regenerated cellulose films, and cellulose derivatives, especially esters and ethers. Small amounts of pure cellulose are further used in the form of microcrystalline cellulose as powder or colloidal suspensions, additives in low-calorie foodstuffs, binder for forming pharmaceuticals into tablets, and for a number of industrial uses.

[Top of Page]

2. Natural Cellulosic Fibers

Robert G. Steadman and Karl Schliefer

All plants produce cellulose, a natural polymer of the plant sugar glucose, as they mature from the “sweet” to the “stringy” phase. This stringy component, located in various parts of the plant, has long been exploited for textile and related uses, with more than 2 000 species having been spun into some kind of yarn or cord at some time or place.

Only a few of these plants are presently cultivated solely for their fibers, cotton being predominant. Few plant fibers lend themselves to home consumption. They are sold mostly as cash crops. All are obtained in impure form and require on-site processing to remove other plant components even before being shipped to spinning mills for further cleaning.

The breakdown of dead plants by bacteria, fungi, and fire is an essential stage in the life cycle of natural flora. Cellulose is thus susceptible to rot, mildew, and fire, three major shortcomings of these fibers in textiles. These problems can be overcome by the application of topical finishes.

Although cellulose predominates, other constituents also influence the properties of these fibers (Table 12). Polymers of plant sugars other than glucose are collectively termed hemicellulose. Lignins appear in most fibers except cotton. Lignified fibers are identified by the phloroglucinol test (1,3,5-benzenetriol [108-73-6]) that is used to distinguish between mechanical and chemical wood pulp (⟶ Paper and Pulp).

Table 12. Composition (in %) of natural cellulose fibers at 10 % moisture content [57]

	Cotton	Jute	Flax	Hemp	Ramie	Sisal	Abaca
Cellulose	92.7	64.4	62.1	67.0	68.8	65.8	63.2
Hemicellulose	5.7	12.0	16.7	16.1	13.1	12.0	19.6
Pectin	0	0.2	1.8	0.8	1.9	0.8	0.5
Lignin	0	11.8	2.0	3.3	0.6	9.9	5.1
Water solubles	1.0	1.1	3.9	2.1	5.5	1.2	1.4
Wax	0.6	0.5	1.5	0.7	0.3	0.3	0.2

Useful fibers may be obtained from various plant parts:

- the stringy bast component of the stem — jute, flax, hemp, etc.
- leaves — agaves, notably sisal
- leaf petioles — abaca (manila)
- fruit — coconut fiber (coir)
- seed — cotton, kapok

Although few plants have hairy seeds, one genus, *Gossypium,* has been cultivated and bred for its fiber with such success that cotton now accounts for 90 % by value of all plant fibers and, if regenerated fibers are included, 78 % of all cellulosic fibers [58]. Figure 11 gives an idea of the worldwide cultivation of cotton and its economic importance relative to that of other plant fibers.

Figure 11. World production of natural cellulosic fibers, 1995

c) Cotton; j) Jute and kenaf; s) Sisal; f) Flax; a) Combined abaca and ramie

Each letter represents $ 400×10^6 production

2.1. Cotton

Table 13 describes the properties of the primary commercial species of cotton, of which American Upland (*Gossypium hirsutum*) is the most important, while Sea Island types account for only a few hundred bales annually. The species *G. hirsutum* has gained importance only in the last 200 years because of the lint's characteristically firm connection to the seed. This lint needed to be separated by hand until the invention of the saw-type gin — generally ascribed to ELI WHITNEY — improved the daily output of lint from 0.5 to 20 kg per person in 1 800, a rate now greatly exceeded.

Table 13. Properties of chief species of cotton

	G. barbadense	*G. hirsutum*	*G. herbaceum*

Common names	Egyptian American Pima	American Upland	Indian
Number of haploid chromosomes, designation	26, “new-world”	26, “new-world”	13, “old-world”
Usual ginning	roller	saw	roller or saw
Typical properties:			
Staple length, mm	35	26	16
Bundle strength, 3 mm gauge, g/tex*	30	24	24
Linear density, mtex*	120	200	350
Micronaire	3	4	7
Color	buff	white	tan

* 1 mtex = 1 mg/km

2.1.1. *Molecular Arrangement, Morphology, and Fine Structure*

Ginned cotton lint consists of ca. 91 % cellulose (dry mass). The remainder is made up of fibrous impurities, chiefly nitrogen compounds (proteins), wax, and pectins, which are found primarily on the fiber surface. (For the molecular structure and behavior of cellulose, see Chap. Cellulose.)

The unicellular cotton fibers grow on the seeds, whereby a tubular outer layer, the cuticle, is initially formed. When a normal fiber is fully developed, an empty cell channel, the lumen, which contains the residual dead protoplasm, remains.

The originally round fiber takes on the form of a flat or ribbon-like tube twisted in such a way as to have alternating S- and Z-shaped twists irregularly arranged along the length of the fiber. The cell wall of normal mature cotton is 2 – 5 µm thick. The fiber diameter is between 12 and 22 µm and is greatest (between 20 and 40 µm) at the base or seed end. Immature cotton has a thin wall, is only slightly twisted, and has a lumen filled with protoplasm. Prematurely withered (“dead”) fibers have an extremely thin cell wall (only ca. 0.5 µm thick) with conspicuous folds.

The cuticle is ca. 0.5 µm thick and covered with a layer of wax and pectins which form wrinkled overlapping rings when they are made to swell, e.g., with cuprammonium hydroxide. The primary wall, 0.5 – 1 µm thick, consisting of ca. 50 % cellulose with pectins, wax, and albumin, follows. The cuticle and primary wall account for ca. 3 % of the fiber's mass. The secondary wall consists of ca. 95 % cellulose and is arranged in numerous concentric layers. The layers denoted by S_1, S_2, and S_3 in Figure 5, consist of tightly packed parallel bundles of fibrils with helical coils along the fiber axis.

The fibrils are considerably less densely packed on the concave side, i.e., the side pushed into the interior of the fiber, than on the convex side, which has a more regular arrangement and a greater packing density [59]. This results from the collapsing of the fiber during its initial drying in the field. As a result, there are internal stresses, which determine the fiber's mechanical properties and, because of differing accessibility, its chemical properties.

The fibrils can be isolated by means of mechanical or chemical degradation (hydrolysis) of the fibers and can be identified with an electron microscope [60]. The major units are the macrofibrils with prevalent diameters of 300, 120, and 60 nm. They consist of microfibrils with a diameter of 10 – 40 nm and a length of 50 – 60 nm. The smallest fibril units are the elementary fibrils consisting of densely packed bundles of single linear polymeric cellulose chains with a diameter of 3.5 – 10 nm [61]. In the longitudinal direction the bundles of fibrils, connected by hydrogen bonds, alternate between highly organized (crystalline) and less-organized regions. This periodicity in the arrangement of the elementary fibrils reappears in the microfibrils in intervals of ca. 50 nm. Figure 4, shows a diagrammatic structural model of the microfibrils, which are constructed of elementary fibrils.

The review of ZAHN [62] summarized progress in elucidating the complex structure of the cotton fiber. Japanese research on the cultured algae *Valonia macrophysa* showed that the algae have regular cell walls. This indicates that the microfibrils are crystals containing 1 200 to 1 400 perfectly arranged cellulose chains [63], see Figure 12. Some of the earlier stress on the “amorphous” parts of the fiber or even of the molecular chains has shifted to assessing the pore-size distribution [64], because pores between 1 and 3 nm are most important for the uptake of dyes and many other finishes. The authors showed an increase in the proportion of pores in that range when cotton is boiled off, and a further increase when mercerized.

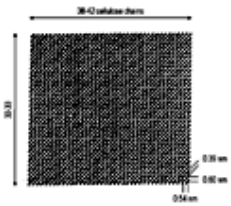

Figure 12. Suggested model for the ac projection of a cellulose crystal from the cell wall of *Valonia macrophysa*

No crystalline substructures exist, one single microfibril being one single crystal containing 1 200 – 1 400 perfectly arranged cellulose chains.

Attempts in recent decades to explain fiber strength, extensibility, and maturity in terms of molecular structure, crystallinity

and the distribution of degree of polymerization, whether using X-ray, near infrared, chemical, or other techniques that do not discriminate between primary and secondary walls, have commonly yielded only modest correlations, of limited practical value. There is increasing recognition of the difference between the properties of the primary wall, which forms first, and the secondary wall, which usually accounts for most of the fiber's mass [65].

One study of the secondary wall confirmed its close relationship with maturity. The development of the secondary wall was manipulated by changing temperature and light intensity to obtain vast differences in cell wall growth and secondary-wall development [66]. Comparative examination showed that the primary and secondary walls polymerize according to different mechanisms, with higher molecular mass, crystallinity, and crystallite size in the secondary (inner) wall [67].

The transverse dimensions of the cotton fiber are less obvious than in most other fibers, because most commercial measurements confound linear density with maturity and because of the peculiar and variable shapes; even within one boll on the plant, there is great variation in fiber development. Attempts to analyze measured fineness into linear density and maturity, by a variety of techniques, are becoming increasingly succesful. Automated image analysis, commonly longitudinal, but preferably of cross-sections, provides several measures that shed light on corresponding commercial properties:

Fiber thickness (the inappropriate term diameter is often used) is estimated from the Martin radius, such as the average of four diameters through the cross-section's center of gravity.

Linear density , in millitex or micrograms per meter, is directly proportional to the cell (i.e., total minus lumen) cross-sectional area, though calibration against known linear densities is needed at present.

Maturity is expressed as degree of thickening, which may be defined as the ratio of the above cross-sectional area to that of a circle having the same perimeter as the fiber.

Varietal fineness (UK: *Standard hair weight*) has defied reproducible measurement until recently. Each variety and species of cotton has a potential linear density reached when “fully” mature; experience with image analysis suggests that the average fiber perimeter is likely to distinguish best between varieties and to provide breeders with the best means of monitoring intrinsic fineness.

Convolutions with some sharp folds along the cotton fiber are shown longitudinally in Figure 13. Figure 14 is a cross-sectional view at the same magnification (1 100×), showing the lumens. These two figures indicate why cotton fibers are cohesive and easily spun in spite of their relative shortness.

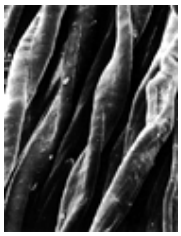

Figure 13. Cotton fiber, longitudinal view, magnification 1 100×

Photomicrograph: Shirley Institute, Manchester, England

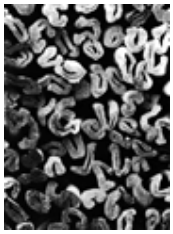

Figure 14. Cotton fiber, cross-sectional view, magnification 1 100×

Surface waxes are not strictly part of fiber structure, as they can easily be stripped off. Along with convolutions, they play a major part in cotton's processability, and may be regarded as a natural spin finish and the reason why cotton is the only fiber normally carded and spun without the addition of finishes to alter friction and cohesion.

2.1.2. *Properties*

Spinning Properties. The chief requirements for spinning textile fibers, and the extent to which they are met by cotton, are the following:

Fiber Length. Because a longer fiber can be spun into a better yarn with less twist, mills pay a substantial premium for “staple length.” This term is important, yet poorly defined. It describes a fiber that is longer than the mean but shorter than the longest fiber. Because of this premium, displacement of cotton by synthetic fibers has been greater at the long end of the cotton spectrum. This is because the price of synthetic staple fibers, which are cut or broken from continuous tow, is independent of fiber length. Cotton fibers vary considerably in length, even on individual seeds. Uniform lengths are coming into increasing demand.

Length Uniformity. Excessive numbers of short fibers increase processing waste and impair spinning. For the production of fine yarn, they must be removed by an expensive combing process.

Impurities and Grade. Cotton lint contains visible particles of leaf, seed, bark, and boll, known collectively as “trash.” Although grade is measured optically, either by the classifier or by colorimetry, it is largely a measure of the absence of trash in the cotton. Table 14 illustrates the correlation between grades and other characteristics. Examination of the last two columns shows that the price of cotton increases at ca. 3 times the rate that would be expected from its clean content. This demonstrates the importance of removing trash and dust early.

Table 14. Grades of American Upland cotton, white 1-in. staple length, 1984

Grade number	Traditional grade	Laboratory nonlint content, %	Cotton content after carding, %	Approximate commercial value, US cents/kg
1	good middling			170
2	strict middling	1.9	93.6	169
3	middling	2.3	93.0	167
4	strict low middling	3.1	91.8	163
5	low middling	4.3	90.2	156
6	strict good ordinary	5.5	89.0	135
7	good ordinary	7.8	85.0	121

The steep rise in value as impurities are removed shows, in common commercial grades, a price sensitivity of 3 : 1, i.e., each percentage increase in clean cotton content raises the unit price by 3 %. This provides an incentive for ginners to remove trash in rural areas before bales are shipped to the mills, where waste removal is more expensive. The controversial practice of lint cleaning degrades most fiber properties while maximizing the cotton grower's total revenue [68].

Traditional assessment of grade by classers includes a close look at “leaf”, because of the nuisance that it causes, especially to the extent that it is shattered into pepper trash in ginning and in the opening and carding processes at the mill. Colorimeters that earlier made assessment more reproducible failed to distinguish between lint that is generally discolored and white lint that contains abundant trash. This difficulty has been resolved by image analysis using digital trash meters, calibration of which enables reliable “gravimetric” estimation of trash content, measurement of particle size distribution, and conversion between laboratory and classers' measurements ([69] and Table 14).

Table 15. Effect of visible trash content on leaf grade (USDA, 1993)

Classer's leaf grade	Trash measurement, mass %
1	0.08
2	0.12
3	0.18
4	0.34
5	0.55
6	0.86
7	1.56

Further instrumental developments are approaching the classer's skill in distinguishing between different forms of trash, with bark and seed-coat fragments being the worst because, in different ways, they behave in processing much like discrete fibers, and are more likely to cause yarn breaks and unsightly fabric defects.

Color. Discoloration due to weather, excessive heat in drying at the gin, bacetria or insects may indicate sugary secretions or general degradation, and usually causes strength loss which results in price discounts. Colorimetry shows that grade and color correspond two-for-two with regard to reflectance and blueness/yellowness, but colorimetry also shows that greenness/redness has no influence on classification. Depending on the degree of yellowness appropriate for each grade, lint is classified as “white,” “light spotted,” “spotted,” “tinged,” or “yellow stained.” Excessively “white” cotton may exhibit bacterial damage within the fiber and is discounted as “gray.” In Egyptian cotton, a creamy tint is natural and considered desirable.

Micronaire. A complex measure of maturity (ca. one-third) and coarseness (ca. two-thirds), micronaire is determined by airflow instruments in micrograms per inch. The linear density or hair weight is, in fact, usually higher than this reading would indicate. Whatever the theoretical limitations of micronaire, it has proved to be a valid guide for assessing processing performance when the species and variety of the cotton — hence, the full potential of the micronaire — is known. American cottons are discounted if the micronaire is below 3.5, suggesting immaturity, or above 4.9, suggesting coarseness. Low maturity implies insufficient development of the secondary wall, with a tendency to form neps, more yarn breaks, and difficulty in dyeing.

Strength. Fiber strength is rarely a factor in the pricing of a bale of cotton. Not only is a stronger fiber likely to make a

stronger yarn and cause fewer processing problems, but as suggested by the trend in fabric weights since 1950, stronger fiber can be used more sparingly, thus making a lighter-weight fabric. Development of hybrid cottons, with strengths almost double that of Upland varieties, promises to widen the scope of cotton's applications.

Elongation and Stress – Strain Behavior. Cotton fiber stretches only ca. 7 % at break, less than most competitive fibers, but well above the 2 – 3 % of most of the other vegetable fibers described in this article. Cotton stretches in response to load in much the same way as polyester. This compatibility of mechanical properties enables the two fibers to be successfully blended and helps to explain why polyester/cotton is the most common fiber blend.

Cohesiveness. The spinning of yarn from discontinuous fibers requires a degree of surface cohesion. This is all the more important if, as in cotton, the fibers are short. Broken cotton yarns usually show evidence that only about half of the fibers have ruptured at the breaking point, while the other half have slipped. In cotton much of the cohesion is due to the natural convolutions in the ribbon-like fiber (see Chap. Cellulose), as can be shown by mercerization, which makes the fibers become rodlike and incapable of being spun. As yet, no commercial method for measuring this property in cotton is available.

Other Properties. *Moisture Uptake.* Cotton has a moisture regain of 8 % at 65 % relative humidity and of 15 % at saturation. Although this is exceeded by wool, jute, and rayon, the lack of a thick cuticle or sheath on the cotton fiber enables the fabric, when freed of natural waxes during finishing, to absorb moisture rapidly. Thus, cotton accounts for 94 % of the United States towel and diaper market [70], and although most printing is now done with pigments, cotton's rapid uptake of water-based dyes is an advantage to the finisher and printer. On the other hand, drying of cotton garments made of such water-retentive fibers after they have been washed adds to energy costs in comparison to those of the less-absorbent noncellulose fibers [71].

Secondary effects of moisture uptake are conspicuous in cotton because it is associated not only with the heat of sorption but with considerable *swelling.* In densely woven fabrics, such as canvas, this swelling blocks large pores, making the fabric impermeable to rain. At the same time, fabric thickening forces the tightly interlaced yarns into a more crimped configuration, which makes the fabric shrink. When the cotton dries, the restoring force is much weaker than the shrinking force, which leaves relaxation shrinkage.

2.1.3. *Production*

Cotton was originally a tropical perennial, but is now cultivated as a summer annual and with a ratoon crop in some tropical countries. It requires a longer growing season than any other annual crop. As an annual, it can be grown in cold-winter regions provided that there are at least ca. 1 200 degree-days above 15 °C during the growing season (one degree-day is recorded for each degree by which each daily mean temperature exceeds 15 °C). Thus, cotton is cultivated in latitudes as high as 45° N, with about 70 % of the world's cotton now being grown in temperate zones such as Uzbekistan, Northern China, and much of the American cotton belt (Fig. 11). The winter in these areas is cold enough to break the continuity of disease and insect problems.

Particularly in North America, the Mexican boll weevil has influenced the geography of cotton growing. Insects such as the heliothis, including the pink boll worm, and spider mite cause expensive problems. Honeydew that is secreted by aphids, whitefly, and other insects produces sticky cotton, which causes processing difficulties, especially when the fiber is passed through drawing rolls. In analysis, this “entomological” sugar must be distinguished from “metabolic” sugar produced by the maturing plant.

2.1.3.1. *Harvesting*

In developing countries, seed cotton is harvested by hand. It is pulled from the bolls, usually in two harvests, starting near the bottom of the plants as they ripen and open. Because strong winds may cause losses, stormproof varieties are planted in windy regions. These are short plants which resist lodging and have bolls that only partially open.

In high-wage countries, mechanical harvesting is predominant. Since this method is less discriminating than human fingers, the plants must first be defoliated either by frost or chemically. The seed cotton is still more contaminated than that obtained by hand. There is normally a single passage with the spindle harvester, as the bolls ripen. Some cottons, including all stormproof types, are harvested in a single operation by stripping the entire bolls, which gives a less-expensive, lower-grade lint. Stripping tends to remove strips of bark, which, because they are fibrous, are less easily removed than other impurities.

2.1.3.2. *Ginning*

Ginning is essentially the separation of lint from seed. Upland cotton is separated by rows of circular saws, which pass through slots that admit lint but not seed. Modern gins perform such functions as drying, blending, and cleaning seed cotton, lint cleaning, packing into bales of 220 kg and 0.5 m^3, invoicing, or marketing. Cottonseed, which accounts for ca. 65 % by weight and 15 % by value of the harvest, is sold separately.

When machine harvesting replaced hand harvesting, a more impure material was delivered to the gins, which resulted in the introduction of lint cleaners. Two passages through lint cleaners, at much higher speeds than in textile mills, restore the lint to the degree of purity that prevailed in the days of hand harvesting and maximize the growers' return. Lint cleaning is particularly useful for stripped cotton. However, it causes many fibers to break and invariably leads to more end breaks in spinning and, thus, a weaker yarn [72]. This weakness can usually not be detected at the fiber-testing stage.

2.1.3.3. *Byproducts*

Cottonseed directly from the gin is covered with short, coarse fibers called linters. These are cut off in one or two stages and used in wadding or as a raw material for cellulose plastics, such as ethyl cellulose. They were formerly an important raw material in rayon manufacture. (Seed for planting may be completely delinted by acid.)

The seed is then stripped of its hull, leaving a valuable “meat.” This is crushed, often with the help of solvent extraction, to

obtain the oil which may be purified for human and animal consumption. With an iodine value of 106, it is classified as a semidrying oil. Primary components are linoleic, oleic, and palmitic acid. The remaining meal, with a shelf life too short for retail trade, is used in commercial baking and as animal feed. The cotton seed contains gossypol, which is poisonous to nonruminants and must be removed for nonindustrial uses by hexane extraction.

2.1.3.4. *Processing*

On arrival at the mill, bales are selected for the desired yarn properties and blended, sometimes with synthetic fibers, in lots of about 30 bales. The first stages of processing greatly increase the bulk and remove much of the trash and dust. Further cleaning is achieved through carding and blending through drawing.

Generally, a roving is spun, which is then drawn and twisted into yarn by a ring and traveler during spinning. The yarn may be plied or finished before being made into a fabric. The newer process of rotor spinning, an open-end method, is excessively expensive for longer fibers, but cotton lends itself well to this technique, especially since rotor spinning is less sensitive to short or immature fibers than ring spinning.

Even though cotton yarn is easily knitted or woven, little is tufted into carpets. To withstand repeated abrasion during weaving, cotton warp yarn that has not been plied is treated with size. Increased speeds and stresses in weaving require yarn of uniform strength, which is largely true for rotor yarn.

2.1.3.5. *Finishing*

Because of its excellent stability in cool alkali and most bleaches, cotton is a particularly washable fiber, provided that shrinkage is controlled. Wet processing normally begins only after the yarn has been spun, usually after making the “gray” (unbleached) fabric.

Natural waxes inhibit the uptake of dye. The first stages are called “preparation,” i.e., for dyeing and printing, and include the following:

1. *Scouring.* Treatment in a hot detergent solution to remove natural impurities from the fiber surface as well as any oil and dirt from processing.
2. *Bleaching.* Removal of any remaining colored impurities through immersion into an oxidizing agent, usually cold hydrogen peroxide; this destroys trash particles or renders them invisible.
3. *Desizing.* Removal of sizes that were applied to the warp before weaving, either by acid hydrolysis or by the use of enzymes to remove natural starches.
4. *Mercerizing.* Originally a specialized process, mercerization is routinely used to increase dye uptake in cotton and polyester/cotton blends. It includes treatment with a cool, concentrated sodium hydroxide or ammonium hydroxide solution, which alters the crystalline structure from cellulose I to cellulose II, thereby improving the accessibility to water and, hence, to water-based dyes. If the yarns are under tension, the luster is enhanced and yarn strength increases by ca. 20 %. This process was formerly used to give cotton fabrics the appearance of the much more expensive silk, as well as to increase strength.

The fabric can then be dyed, usually by direct, vat, or reactive — occasionally basic, sulfur, or azoic — dyes. Pigment printing is popular, partly because of its excellent colorfastness, and because it can be applied to produce a solid shade in imitation of dyeing.

2.1.3.6. *Special Finishes*

Permanent Press. Because interchain bonds in cellulose are weak, cotton is easily deformed. Crease-resistant finishes, originally developed to reduce the severe wrinkling of rayon, are now widely applied to cotton. Several cross-linking agents, notably dimethylolethylenurea (1,3-bis(hydroxymethyl)-2-imidazolidinone [140-95-4]), are in use. These are applied to the fabric, dried, and then reacted by using heat (curing). The harsh texture that is thus imparted is mitigated by applying softeners. The introduction of polyester/treated cotton blends in 1964 proved successful, as the strength of the unaffected polyester enabled the application of resin to provide improved wrinkle resistance.

Flameproofing. Cellulose burns readily in air.

$$C_6H_{12}O_6 + 6\,O_2 \rightarrow 6\,CO_2 + 6\,H_2O + \Delta H$$

Carbon monoxide is produced in confined spaces:

$$C_6H_{12}O_6 + 3\,O_2 \rightarrow 6\,CO + 6\,H_2O + \Delta H$$

In the presence of a Lewis acid, only charring occurs:

$$C_6H_{12}O_6 \rightarrow 6\,C + 6\,H_2O$$

Most flame retardants inhibit the first two reactions. They include simple soluble substances such as borax or alum and insoluble chemicals formed in the fiber by reaction, such as stannous oxychloride or antimony oxide.

Most introduce problems of handle, or “feel,” toxicity, or poor durability and have been largely replaced by a variety of organophosphorus compounds or by substitution with synthetic fibers in which an insoluble flame retardant is dispersed before the fiber emerges from the spinneret.

Shrinkproofing. In bottomweight fabrics (those > 140 g/m^2 and used for trousers, workwear, etc.), compressive shrinkage is used. A rubber blanket is moved in opposition to the fabric through the finishing machinery, so that the filling yarns are forced

together and the progressive shrinkage that would otherwise occur during repeated laundering is forestalled.

In topweight fabrics, such as shirtings and knits, satisfactory results are usually obtained as a side effect of the permanent-press resin.

Chemical Stonewashing. In many cotton-rich fabrics, especially in woven outerwear, a strong demand for moleskin and stonewashed finishes exists. Mild treatments include emersing, in which moderate abrasion of this fine fiber produces a soft finish. Increasingly, and especially when a “stonewashed” appearence is desired on denim, literal stonewashing is being superseded by controlled enzyme treatment with selected cellulases. Despite a weight loss of about 3 %, the finish does less damage to the fabric, and may be combined with other finishes such as crushing. It is applied to solvent-spun rayons of the Tencel type, the only type of regenerated cellulose fiber whose consumption is not in decline. Some celluloses are beginning to be used for the reduction or bleaching stage of vat dyeing, especially in denim.

2.2. Bast (Soft) Fibers

Plants with a stem containing a fibrous bast component may be commercially processed if the fiber yield and the ease of separation warrant commercial use [73]. The cross-section of the stem, shown in Figure 15, shows several undesired components which must be removed mechanically or chemically, a process that is frequently so odorous that the retting process must be carried out in less-populated regions.

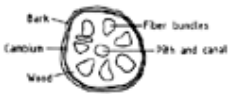

Figure 15. Cross-section of plant stem with bast fiber

Although the process used varies according to the fiber, region, and available machinery, all bast fibers are generally subjected to the following processes after harvesting [73]:

mechanical or manual removal of seeds and leaves (rippling)
mechanical removal of bark (decortication)
biological removal of woody components through decomposition (retting)

At this stage the fiber, with some residual impurities, is usually baled and shipped to the mill for further processing.

The bast consists of long fiber bundles, which may be as long as the stem. These are made up of ultimate fibers containing lumens and are held together by woody gums; unlike most textile fibers, bast and leaf fibers are bundled. The elementary fibers, even though suitable for pulping, are too short for spinning, so that processing requires mechanical rupture of the bundles into manageable lengths. This is often achieved with a lubricant which softens the gums and allows the elementary particles to separate gently.

In addition to the fibers described here, other bast fibers of commercial importance include urena, sunn hemp, and kenaf.

2.3. Jute

Jute ranks second only to cotton among the natural cellulosic fibers. The net production of jute is greater than that of wool. Declining prices, about 40 ¢/kg in 1995, have reduced its importance since 1970. Since the early 19th century, jute has been intensively grown in the Ganges delta, and except for the substantial increase in jute cultivation in China since 1950, it has seldom been successfully cultivated anywhere else. Most jute-growing countries process jute to produce cloth locally, with only Bangladesh being a major exporter of the fiber [74].

Two varieties are used in commercial production. *Corchorus capsularis* (white jute) is tolerant of waterlogging and is grown on small delta farms subject to flooding during the summer monsoon. Harvesting of the 4-m stems with knives is sometimes done by divers. Good grades, such as Bangla (formerly Dacca) White A and B, are sometimes used for inexpensive apparel. Much bleached white jute in apparel weights appears in world markets as “hemp”. The two fibers are not easily distinguished, but jute is identified by turning yellow in sulfuric acid and iodine, not blue.

C. olitorius (tossa or upland jute) is grown on the wet slopes of India, Bangladesh, and Nepal. It is brown and contains noncellulosic impurities which cannot be completely bleached because of resultant excessive fiber damage. Some lowland areas produce “daisee” jute, which is gray or black because of iron salts in the soil.

The stems are retted in rivers, sometimes ponds, giving a fiber yield of only ca. 6 % (i.e., 18 t of “green” jute must be harvested to obtain 1 t of retted fiber). If the prices are high, the stumps (jute butts) are salvaged, though the middle of the stem produces the best fiber.

At the mill, the fiber undergoes blending and spreading with addition of ca. 5 % processing oil before carding. Because the fiber bundles are uneven in both length and diameter, individual yarns are irregular, giving the cloth its characteristic appearance. It is variously known as burlap, hessian, gunny, or sackcloth.

Jute has the shortcomings of cotton, with an even greater tendency to rot, but unlike cotton, it is usually too heavy and cheap to justify the cost of finishing treatments. It is normally not even scoured, formerly leaving the characteristic smell of old vegetable or whale oils on the cloth. Used mostly in carpet backing, sacks and bale wrapping, jute has been largely replaced in the developed world by bulk loading and the lighter, inexpensive polypropylene made from plastic film. Even organizations that advocate the consumption of their own natural fibers have urged the replacement of jute bales with synthetics. In primary carpet backing, polypropylene has also for the most part replaced jute, but it has the drawback of being thermoplastic. A roll

of carpet that has just been stored in a cold warehouse may, therefore, be too stiff to lay. Partly for this reason, jute is still common in the secondary backings which are applied to the underside, where its low extensibility helps stabilize the carpet.

2.4. Flax

Unlike cotton, flax (*Linum usitatissimum*) is grown less commonly for its fiber than for the seed, which is crushed to yield linseed oil. Over half of the world's flax is grown in countries that used to be centrally planned. With the sudden transition from subsidized farming to direct exposure to world markets, flax production for fiber has declined, e.g., from 2 000 to 800 t in two years in Estonia, but it remains a fashion fiber and is widely imitated.

The plant is much smaller than jute and is harvested by pulling out of the ground. It is retted on the ground ("dew" retting) or, increasingly, under controlled conditions in tanks. With a fiber yield of only ca. 11 %, an experimental process for retting in the field before harvest is receiving great attention.

At the mill, the fiber is processed more thoroughly than jute, often with "hackling," similar to combing, to separate long ("line") from short ("tow") fibers. About half of the linen yarns are moistened during spinning to lubricate the fiber ultimates which at about 10 mm length are twice as long as jute. Wet-spun flax yarn may be very fine and strong, but its low elasticity limits its ability to blend with other fibers.

The fact that flax has long been the standard against which other fibers have been measured, is expressed in terms such as "bed linen" and "table linen," even though flax fabrics are less durable than those of cotton or blends. The high absorbency of these fabrics destines it for use in towels but this is also combined with severe wrinkling and requires large amounts of crease-resistant resins.

2.5. Hemp

Hemp (Indian hemp) is rarely produced outside of the former Communist countries. The plant (*Cannabis sativa*) is widely cultivated for its leaves, known as marijuana, but the potential for salvaging the fiber in the stems has hardly been exploited. Yields per acre are lower than those for jute, which has similar properties. When grown for fiber, the plants are more closely spaced than for the drug crop, as leaf development is not needed. To distinguish the two uses and to encourage government support for hemp growing, many people have renamed the fiber "industrial hemp".

Because of restrictions on the drug, governments have discouraged production of hemp. World production of true hemp, insufficient to appear on the map (Fig. 11), is probably even lower than official statistics would indicate, because of widespread use of the misnomers "sisal hemp" and "manila hemp."

2.6. Leaf (Hard) Fibers

Fibers obtained from the leaves of plants such as *Yucca filamentosa* and agaves are extracted from the harvested, dried leaves — usually mechanically, since the leaves have spikes at the end and, in some species, along the sides.

These fibers are coarser, but more resistant to rot than the soft fibers and find application in cordage [75]. They have met with increasing competition from synthetic fibers. Common species are *Agave sisalana* (sisal) and *A. fourcroydes* (henequen).

Some banana trees, notably *Musa textilis,* have very fibrous leaf petioles. When separated from the rest of the leaf (tuxying), this abaca fiber can be used for high-quality ropes, often called manila. The natural tan color is sometimes imitated in other fibers through dyeing. Almost as elastic as cotton, manila ropes absorb shock loads better than most vegetable fibers.

2.7. Coir

The fibrous outer layer of the coconut, called coir, may be considered a seed fiber, like cotton, but is usually classified as a fruit fiber. The coarse fibers are commonly processed in India, Sri Lanka, and the West Indies and are woven into coverings for floors and playing surfaces.

2.8. Ramie

Ramie is a bast fiber taken from the stingless nettle, grown in the Philippines and the People's Republic of China. It rivals cotton in purity, whiteness, and strength. A minor fiber, it was omitted from the GATT multifiber agreement. Then it appeared abundantly, often in blends with cotton or polyester, in international trade, unhindered by quotas.

2.9. Economic Aspects

All of the natural cellulose fibers have been affected by the introduction of synthetic substitutes, first as regenerated cellulose (whose total worldwide production is less than one seventh that of cotton), later as fibers of fully synthetic polymers. Because rayon shares many of the problems of natural cellulose — and has a few of its own as well — most of the inroads into markets for plant fibers have been made by such fibers as polyester, nylon, and polypropylene. The replacement of jute and the leaf fibers by polypropylene has been particularly harsh. These newer fibers offer a higher strength : weight ratio than cellulose, as well as greater resistance to rot, mildew, burning, and wrinkling. On the other hand, some plant fibers, especially cotton, have maintained their position, world cotton production being at record levels. Of all the fibers used in apparel, cotton's share in all textiles has remained almost unchanged at near 50 % since 1974. Table 16 shows production trends.

Table 16. Proportions (in %) of the chief apparel-type fibers worldwide [76], [77]

Year	Cotton	Wool	Regenerated cellulose	Noncellulosic synthetics

1930	82	14	3	0
1940	75	12	12	0
1950	71	11	17	1
1960	67	10	17	5
1971	52	7	14	27
1984	49	5	9	37
1995	46	4	6	44

A good annual yield in most parts of the world is considered to be about 500 kg/ha (approximately, 1 kg/ha = 1 lb/acre) of cotton or 2 000 kg/ha of jute. Due to a growing world population, food production has been intensified and has caused some land to be diverted from the growth of cotton and jute to food crops. Table 17 shows production of the chief natural cellulose fibers. In interpreting these data, the increasing proportion of leaf fibers being made into pulp, rather than ropes or other textiles, should be kept in mind.

Table 17. World production of natural cellulose fibers (×10^3 t) [78], [79]

Year	Cotton	Jute	Flax	Hemp	Sisal	Abaca
1950	6 600	1 417	768	948		
1960	10 100	2 420	650	340		720
1971	11 600	3 446	670	269		821
1983	14 700	4 057	669	230	384	230
1994	18 900	3 530	610	120	380	76

Many of the fibers in this group, notably cotton and flax, are stronger when wet. Conversely, they are weak when warm and dry, a factor that played a part in reducing cotton's once-dominant position in tire cord.

2.10. Occupational Health

These fibers contain endotoxins which are shaken loose in processing if, as is usually done, the fibers are worked dry. Exposure to these airborne particles over decades may cause byssinosis in textile workers, especially in smokers. Regulations governing this problem [80] have caused profound changes in mills for processing cotton and have led to the redesign of machinery to reduce airborne dust. These changes are responsible for important secondary improvements in the processing behavior of cotton.

[Top of Page]

3. Regenerate Cellulose

Wilhelm Albrecht

Terms and Definitions. Man-made fibers are divided into two groups — cellulosics, and synthetics.

In both cases, the fibers are spun under conditions that are oriented to the end use in the form of continuous filaments. These filaments can be used as they are, textured, combined to heteroyarns, or cut into a staple of defined length.

In Germany, cellulose-based man-made fibers are produced according to DIN 60 001 from cellulose [81]. At present, the following three methods are used:

the viscose process — by far the most important (see Section Viscose Fibers)
the Lyocell process (see Section Lyocell Fibers)
the cuprammonium process (see Section Cuprammonium Fibers)
the acetate process (see ⟶ Cellulose Esters – Cellulose Xanthate)

Other production methods are discussed in Section Tentative Cellulose Fiber Production by Other Processes – Outlook.

While the first three of these processes result in regenerated cellulose fibers, the fourth gives cellulose ester fibers. The viscose process has been widely modified over time. One special kind of regenerated cellulose fibers is the so-called Modal fiber, which is available in two subgroups: the polynosics (= polymer non synthétique), and HWM fiber (= high wet modulus). In the Lyocell process, cellulose is regenerated from a solution in an organic solvent. Cellulose ester fibers are sub-divided into acetate and triacetate fibers (see ⟶ Cellulose Esters – Cellulose Acetate Fibers). Table 18 gives a listing of fiber designations according to ISO/TC 38.

Table 18. Fiber designation according to ISO/TC 38

Fiber	Generic name Code

Cuprammonium	cupro	CUP
Lyocell	lyocell	CLY
Modal	modal	CMD
Viscose	viscose	CV
Acetate	acetate	CA
Triacetate	triacetate	CTA

In addition, the main processing and application properties of the fibers are indicated by figures, descriptions, and short references such as:

number of single filaments
staple cut in mm
titer in dtex (1 dtex = 1 g/10 000 m)
bright (no dulling agent added)
dull/superdull (containing 0.3 – 4.0 % of a dulling agent, generally anatase)
hollow (hollow fibers)
ribbon (tapiform fibers)
crimped/high-crimped fibers (90 – 140 waves/ 100 mm)

Other descriptions may describe color, moisture retention, flammability, degree of optical brightening, suitability for medical applications, X-ray contrast, recommended uses, etc.

History. Regenerated cellulose fibers are the first artificial fibers ever made. Processes capable of dissolving the cellulose derived from wood or cotton linters were first discovered by SCHÖNBEIN (1845, nitrocellulose soluble in organic solvents), SCHWEIZER (1857, cellulose in cuprammonium solution), CROSS, BEVAN, BEADLE (1885, cellulose sulfidized in sodium hydroxide; 1894, cellulose triacetate in chloroform). The production of threads from cellulose (derivative) solutions goes back to AUDEMARS (1855), who drew filaments from cellulose nitrate submerged in alcohol ether, and SWAN (1883), who transformed nitrocellulose filaments into hydrated cellulose. Members of the Swan family used the threads, which were intended for the manufacture of light bulbs, for their craft work as well. SWAN called his discovery “artificial silk.” This early developmental work also set the trend for subsequent uses with regenerated cellulose fibers. In 1891, Count HILAIRE DE CHARDONNET built the first industrial plant for the production of artificial silk according to the nitro process in Besançon, France (daily output ca. 110 kg), and MAX FREMERY established the “Rheinische Glühlampenfabrik” for the production of cellulosic fibers suitable for carbon filament bulbs in Oberbruch (which developed into the Akzo Fibres Group). The spinning techniques were gradually improved: H. PAULY (1897) was granted a patent on a process in which “solutions of cellulose in copper oxide ammoniac are passed through fine holes into a liquid capable of decomposing these solutions” and published for the first time operable instructions. Other patents include TOPHAM's first spinning centrifuge (1900) and spin pump (1901); THIELE's stretch-spinning method for cuprammonium silk (1901); MÜLLER's spin bath for the viscose process (1905); and BOOS' continuous production of artificial silk (1906). A new phase was initiated in 1925, when H. STAUDINGER began describing macromolecules. The introduction of modifiers by COX in 1950 and the development of high-wet-strength fibers initiated by TACHIKAWA in 1951 again increased the variety of cellulosic man-mades. Commercial production of acetate fibers started in 1919; in 1955 they were joined by triacetate (see ⟶ Cellulose Esters). When during the 1960s synthetic fibers swept the textile market and water/air-pollution problems began darkening the image of cellulosic man-mades, they seemed doomed. However, after years of fierce competition, cellulosic fibers still maintain their characteristic place among fibers, doubtlessly aided by new developments in raw materials and in fiber manufacture [82]. The search for more ecological and more productive production methods led to the development of the Lyocell process in 1972 and in the course of time to a better adjustment of the raw material to improve the viscose process.

3.1. Viscose Fibers

3.1.1. *Principle of the Viscose Process*

The starting material for the production of filaments and staple fibers is pulp, mainly produced from wood and linters, but also from annual plants (⟶ Paper and Pulp). Since pulp fibers in their natural state are much too short to be spun into yarns, they are dissolved and regenerated in filament form. Part of these filaments is cut into fibers of the required staple length.

The spinnable solution is prepared by dissolving the raw material “pulp” in sodium hydroxide to obtain alkaline cellulose (first step), which is, if necessary, allowed to age (preripen) until a certain degree of depolymerization is reached (second step). Carbon disulfide is then added (third step). The resulting xanthate is dissolved in dilute sodium hydroxide (fourth step). With modern equipment, these four steps are combined into one.

Subsequent ripening (maturing) or a previously set degree of polymerization (DP) guarantee that the fiber can be perfectly spun and meet the requirements of the textile end use. During the process, the viscose is filtered, deaerated, and finally pressed through the holes of the spinnerets into the regeneration bath, where filaments are formed and drawn off at high speed (Fig. 16).

Figure 16. Flow diagram of viscose fiber production

a) Pulp; b) NaOH treatment and pressing; c) Shredding, aging, if necessary; d) Carbon disulfide treatment, xanthate crumbs; e) Dissolving xanthate in NaOH, viscose maturing; deaeration, filtration; f) Filament yarn; g) Staple

Fiber properties can be widely varied by changing the processing conditions, such as the compositions of the viscose and the regeneration bath, chemical additives, and modified spinning technology (spinning bath temperature and spinning speed, drawing ratio and geometry, number and nature of the baths, etc.) [86-90].

3.1.2. *Viscose Preparation*

For textile filaments and staple, the preferred raw material is bleached wood pulp with an α-cellulose content of 89 – 93 %; it comes in sheets, rolls, or crumbs. High-tenacity filament yarns and modal fibers require even higher grades of pulp (≧ 94 % α-cellulose). Cotton linters are used mainly in cotton-producing countries.

The cellulose must fulfill certain requirements with regard to processibility (unit weight, swelling behavior, hemicellulose content, inorganic substances, degree of polymerization (DP)) and the properties of the resulting fibers. Many efforts were made to “activate” cellulose by thermal treatment, radiation, introduction of spacers, and surfactants [82]. It has been found that a suitable irradiation with β-rays is the most useful method [83].

Alkalization. When exposed to sodium hydroxide of 20 % or more, the pulp absorbs water and alkali, releases heat (ca. 12 kJ/mol anhydroglucose), and loosens the cellulose lattice. This results in sodium cellulose I (Section Chemical Properties), most of which changes into cellulose II during regeneration.

Although alkaline cellulose is an important intermediate product in cotton mercerization and in the preparation of viscose as well as various cellulose ethers, its nature has not yet been fully explored. Cotton mercerization is primarily intended to change the interfibril structure, which is also possible with other swelling agents. Alkaline treatment during viscose preparation is, apart from opening the molecular structure, aimed at transforming as much cellulose as possible into sodium cellulose I. This explains why several different definitions are suggested for the alkaline phase of the cellulose, such as “pseudostoichiometric combination $(C_6H_{10}O_5) \cdot (NaOH) \cdot (H_2O)_{3-5}$,” or “a more or less loose addition of water molecules and hydroxyl groups of the cellulose to an Na^+OH^- dipole.” Besides, some authors mention mesomerizable onium compounds with a cation of limited mobility as well as the fact that some of the cellulose has the form of a hydrated alcoholate Cell-$O^-Na^+ \cdot H^+OH^-$ attached to it [97-99].

Commercial treatment of the pulp starts with the addition of 18 – 22 % sodium hydroxide of 17 – 45 °C. Sometimes depolymerization catalysts such as manganese salts are added. An alkaline cellulose content of ca. 30 % is obtained by pressing. If necessary, the alkaline cellulose is shredded and left to age (generally a continuous process). At this stage, it is important to control the temperature and the surface of the crumbs must not be allowed to dry out. This is particularly important in shortened ripening methods. The degree of polymerization for normal textile fibers is set at ca. 300 and for high-tenacity yarns at ca. 450 and higher.

It has long been known that during the third stage of the viscose process, hemicellulose readily attracts the carbon disulfide, and numerous efforts have been made to reduce the hemicellulose content of the alkaline cellulose. One possibility would be tandem steeping, i.e., two alkaline treatments. The first serves to wash out part of the hemicellulose that is already present in pulp and is formed during swelling and alkaline cellulose formation. After the second alkaline treatment the alkaline cellulose has a low hemicellulose content and distinctly less carbon disulfide is needed for xanthation (20 – 30 % less).

The composition and condition of the alkaline cellulose are governing factors for its processing properties (grinding, preaging, xanthation, filtration). Therefore, it is necessary to check the temperature and the composition of the alkaline solution (hemicellulose, sodium carbonate, alkaline- earth metals, and heavy-metal ions). Sodium carbonate is generally present in percentages between 0.09 and 0.3 %. Higher concentrations may create difficulties in viscose filtration. Varying heavy-metal ion concentrations cause differences in the depolymerization reactions during ripening [100], [101].

Shredding. If the alkaline treatment starts with pulp sheets or rolls, it is followed by shredding. At the same time, the alkaline viscose is adjusted to the ripening temperature. Higher temperatures (up to 45 °C) accelerate the process, lower temperatures (around 20 °C) prolong it. Shredded alkaline cellulose consists of white crumbs. The density by volume is held constant.

Preripening. The pulp and, hence, the alkaline cellulose now usually have a degree of polymerization (DP) which is much higher than desired for fiber formation. The delicate viscose process requires a viscose with an optimal combination of concentration and viscosity. Therefore, the DP must be reduced to ca. 300 (for regular fibers) or 450 or slightly higher (for high-tenacity fibers). The preferred method is to store the alkaline cellulose for a certain period of time at 20 – 45 °C. Care must be taken to avoid drying of the crumbs, which would result in carbonate formation, particularly at higher temperatures. For this reason, the alkaline cellulose is stored in closed containers or is constantly agitated (in continuous operations). Depolymerization is an exothermal process and can be accelerated by suitable catalysts. The results depend on such factors as time, temperature, the nature of the pulp, and the composition of the alkaline cellulose (NaOH, H_2O, hemicellulose, and heavy metals) [102]. This ripening process is either discontinuous (i.e., in containers) or continuous (i.e., in special cabinets

provided with conveyor belts or in rotating drums).

It is generally believed that cellulose fiber producers should themselves adjust the preferred DP by ripening the alkaline cellulose. Currently, this is the most economical method. This principle is questionable, since new studies have been published and attractive uses for byproducts from pulp production are thinkable. Preripening can be avoided if pulp is used as feed material, whose degree of polymerization has been adjusted to the desired level by irradiation with β-rays. This treatment also leads to a favorable leveling of the DP.

Xanthation. This term means treatment with carbon disulfide with which the alkaline cellulose reacts and generates heat. The reaction is very rapid and begins on the outer surface of the crumbs — hemicellulose included — and slowly proceeds into the crystalline structure. This homogenating process continues until the viscose is ready for spinning. In commercial production, based on unirradiated pulp an average substitution degree of 0.5 – 0.6 is obtained, reflecting a ϒ-value (number of xanthate groups per 100 glucose units) of 50 – 60. ϒ-Values of ca. 50 are sufficient for a readily filterable viscose free from fibrous material. High-tenacity yarns require higher ϒ-values. Under special conditions, it is even possible to obtain ϒ-values of up to 300. Xanthation is not uniformly distributed over the cellulose chain, and even with a substitution degree of 1, not all of the glucose groups will have one substituent, let alone one in a specific place. Alkaline cellulose made of irradiated pulp can be reacted with distinctly less carbon disulfide to yield a good soluble xanthate. In practical operation ϒ-values of 30 – 35 for textile viscose fibers are used.

Table 19 shows the distribution of the xanthate groups over C-2, C-3, and C-6. There are indications that the speed with which the primary and secondary xanthates are formed and destroyed must be different. The hydroxyl groups in position C-2 are the fastest to react with xanthate, but are less stable than those attached to primary C-6 hydroxyl groups.

Table 19. Distribution of xanthate groups over C-2, C-3, and C-6, in % [103-105]

Hydroxyl groups occupied	Acc. to [103]	Acc. to [104]	Acc. to [105]	
			fresh	ripened viscose
			ϒ = 60	ϒ ≪ 49
C-2	43	38	a	b
C-3	20	28	a	b
C-6	37	34	31	60 – 100

[a] ∑ (C-2 + C-3) = 69 %;
[b] ∑ (C-2 + C-3) = 0 – 40 %

Xanthation speed depends on temperature, the condition and composition of the alkaline cellulose, the amount of carbon disulfide added, and the pressure of the carbon disulfide vapors formed during the process [103-106]. Pulp that has been adjusted to a desired DP by β-irradiation shows a more favorable behavior during xanthation than untreated pulp because of its narrower DP distribution. A much lesser amount of carbon disulfide is required for reaction and the reaction time is reduced. In addition, the spinning properties of the viscose obtained are even improved, when certain limiting values are complied with.

Side Reactions. Xanthation of alkaline cellulose is accompanied by a great number of side reactions. They are responsible for the characteristic orange color of the xanthate and the viscose made from it.

Carbon Disulfide. Depending on the type of fiber and the alkaline process used, between 150 and 400 kg of carbon disulfide is needed to produce 1000 kg of fiber. CS_2 is a clear, colorless liquid that is to distilled at 46 – 47 °C up to 98.5 %; it should contain < 5 ppm of hydrogen sulfide and be free from carbonyl sulfide and acid. The nonvolatile residue content should be less than 0.01 %, and the density should be 1.26 – 1.27 g/cm^3.

Process Techniques. Older equipment may still contain hexagonal drums (heated jacket, teflon coated, vacuum tight). They are filled with alkaline cellulose with adjusted DP and evacuated. The required amount of CS_2 is added via the perforated axle of the drum (15 – 55 %, in the classic process 28 – 35 %). Pressure develops and replaces the vacuum; gradual normalization is aided by cooling. Xanthation takes 1.5 – 2.5 h, after which nitrogen is injected; the remaining carbon disulfide is removed and the orange xanthate crumbs are dumped into the dissolving tank. This early process was work-intensive and exposed the workers to noxious carbon disulfide vapors. During the 1950s, deep-reaching changes took place. Kneaders were introduced which held much larger quantities of alkaline cellulose for xanthation and allowed the first phase of xanthate dissolution to take place. The kneaders were followed by automatic sulfidizing and dissolving equipment.

Processes involving carbon disulfide are dangerous. Sparks may trigger explosions. Safety instructions must be carefully observed.

Viscose Solution. The xanthate crumbs are dissolved in sodium hydroxide either in batches or continuously to give a sticky substance; hence, the name viscose. Dissolution is accelerated by stirring and trituration. The honey-colored viscose contains 7 – 12 % cellulose and 5 – 8 % caustic, as well as various sulfur compounds and carbonate. Viscosity varies from 3.5 to over 10 Pa · s and depends on the concentration of the xanthate, the degree of polymerization of the cellulose, the degree of dissolution, which in turn is a function of dispersion, the degree of substitution and its distribution, and on the purity

and concentration of the dissolving sodium hydroxide.

To avoid such common interferences as short fibers and gels, certain chemical auxiliaries are added directly either to the pulp or to the viscose. These auxiliaries are surfactants such as fatty amines or polyglycol compounds in fractions of 1 ppm of viscose.

Maturing. Maturing is another delay period, during which sulfidation continues and substitution becomes more uniform. At the same time, the ϒ-value is reduced and the amount of byproducts increased, while the ability of the viscose to coagulate also increases (chemocolloidal maturity).

Chemocolloidal maturity can be described by the Hottenroth index (°Ho). It is determined by thorough mixing of 20 g of viscose 3 times with 10 mL of distilled water. A 10 % solution of ammonium chloride is added at 20 °C. When “maturity titration” is completed, the viscose changes into a semisolid gel which should adhere to an horizontal stick for at least 20 s. The Hottenroth index corresponds with the number of cubic centimeters of ammonium chloride required to obtain that result. At the start of the maturing process, the Hottenroth index reaches a maximum [107].

This manual maturity test can be easily automated. With the help of suitable apparatus, the maturing process of the viscose can be made much more uniform [108].

The decomposition speed of the sodium xanthate depends on the temperature and concentration of the sodium hydroxide, cellulose, and carbon disulfide [109]. The nature of the reaction has not yet been fully determined.

With older equipment, the maturing process at ca. 18 °C takes several days (50 – 80 h). It can be adjusted to the type of fiber to be spun and the spinning method to be used. Maturing is accompanied by filtration and deaeration.

These three operations have been accelerated and automated by increasing the viscose temperature, the development of various kinds of filters, and the introduction of vacuum evaporation.

3.1.3. *Viscose Fiber Spinning*

Fiber formation, i.e., regeneration of the cellulose from the viscose with the help of the spin bath and treatment outside the spin bath, includes coagulation of the viscose (sol – gel), deswelling of the filaments, regeneration of the cellulose, neutralization, and stretching of the yarn. The chronological sequence of these steps (which overlap and run parallel) determines the structure and, hence, the properties of the resulting fibers.

Composition and Effect of the Spin Bath. The formation of the thread in the sulfuric acid bath involves a number of chemical reactions. The most important is the decomposition of xanthate into cellulose and carbon disulfide. Besides, the thiocarbonates formed in side reactions during sulfidation and maturation disintegrate. These two basic reactions can be described in a simplified form as follows:

$$2\,\text{Cellulose} - \text{O} - \text{CSSNa} + \text{H}_2\text{SO}_4 \rightarrow 2\,\text{Cellulose} - \text{OH} + 2\,\text{CS}_2 + \text{Na}_2\text{SO}_4$$

$$\text{Na}_2\text{CS}_3{}^* + \text{H}_2\text{SO}_4 \rightarrow \text{Na}_2\text{SO}_4 + \text{H}_2\text{S} + \text{CS}_2$$

* Representative for a number of inorganic thiocompounds.

In addition, carbon dioxide is formed when the caustic lye reacts with the carbon disulfide, and finally, colloidal sulfur is released.

Sulfates — mainly Na_2SO_4, but also $MgSO_4$ and rarely $(NH_4)_2SO_4$ — allow the viscose to coagulate, i.e., macromolecules and associates join into larger aggregates; acids alone initiate rapid disintegration of the viscose, and sulfuric acid of higher concentrations than 55 % gives high-tenacity, low-elongation Lilienfeld Silk at low bath temperatures.

The composition of the spin bath is adjusted to the composition of the viscose and the type of fiber to be spun (i.e., titer, tenacity, and crimp). Spin bath temperature, spinning speed, the amount of viscose injected, and the construction of the spinneret are further factors. In general, a spin bath will contain 7 – 12 % sulfuric acid, 12 – 24 % sodium sulfate, and 0.5 – 3 % zinc sulfate. In exceptional cases, part of the sodium and zinc sulfate can be replaced by magnesium sulfate (ca. 5 %) so that the swelling of the filament is reduced before winding.

In practical spinning operations, the matured viscose from the tank is injected by precision pumps through fabric filters inside the spinneret support and then through the holes of the spinneret. The body of the spinnerets is made of gold or platinum alloys, with or without rhodium, or of platinum/iridium alloys. Some tantalum and glass spinnerets can still be found. However, since it became clear that the spinnerets are a very important machine part in viscose fiber production, more attention is being paid to high quality. This also includes the finish and form of the spinning holes. Seen in the flow direction, these holes have a conical inlet followed by a cylindrical channel. The cone is about double the length of the channel, but must be adjusted to the thickness of the plate. Depending on the intended fibers and their thickness, the channels have a diameter between 50 and 250 µm. In filament yarn spinning, the number of holes is equal to the desired number of single filaments in the yarn (for example from 18 to 1 000); in staple fiber production, it may surpass 90 000. Spinnerets of this kind are composed of several smaller spinnerets combined into one spinning point. For relatively wide ribbon fibers spinnerets with oblong holes are employed.

When the viscose emerges from the submerged spinneret and enters the spin bath, it begins to solidify. Special “marching-off” devices ensure regeneration of the cellulose in filament form and completion of cellulose regeneration. Spinning and pumping speed are precisely adjusted. Filament yarns of 100 dtex are wound at up to 160 m/min; filament tow for staple production is spun at between 35 and 120 m/min, depending on the fiber type and the number of filaments.

The regenerated cellulose fibers gain more tenacity and less elongation with a higher final degree of orientation. The process is supported by keeping sulfuric acid in the spin bath at a minimum, by adding certain chemicals, and by increasing the zinc content to decelerate the decomposition of the xanthate. Polynosic fibers are spun into a low-temperature bath at lower speeds.

Depending on the composition of the viscose, the spin bath and the flow diagram of the viscose leaving the cylindrical part of the holes starts filament formation in differently structured layers. This is due to the differences in the diffusion speeds of the hydrogen and zinc ions and the reversal of the acceleration diagram of the flowing viscose. On their way out, the degradation products that result from reaction with the faster hydrogen ions meet the slower zinc ions, with which they combine and precipitate ($ZnCS_3$ and ZnS_2). At the same time, two distinct layers are formed in the fiber; the outer one being less-swollen zinc xanthate and the inner one highly swollen sodium xanthate (skin/core). Both layers are differently structured, e.g., regarding the size and total volume of cavities [110-115]. The outer layer differs from the core in the following respects:

smaller crystallites (sometimes cellulose IV residues)
increased strength and elongation
greater resistance to fatigue
less lateral order and density by volume
lower refractory index
lower wet modulus
less accessible to large low-polar molecules
lower defibrillating tendency

The relation between skin and core is about 1 : 3 in standard viscose fibers and 1 : 0.5 or less in high-tenacity fibers.

The cross-sectional shape of the viscose fibers (see Fig. 20) depends on spinning conditions (deswelling, shrinking, drawing, and drying). Addition of varying amounts of zinc sulfate, adjustment of viscose and spin bath composition, variation in maturity, and high byproduct sulfur content may increase the thickness of the skin to almost full fiber thickness (see Section Fiber Types). The amount of crystalline matter (28 – 50 %) and degree of orientation, lateral order and size of the crystallites are determined by stretching the filaments at different formation points [90]. In standard viscose yarn production, stretching occurs late in the decomposition phase; high-tenacity yarns are stretched in a less decomposed condition. Stretching of the plastic gel filaments has no measurable effect.

Stretching is generally done in several steps. Little stretch is applied between the spinneret and marching-off device, while a higher stretch is exerted between the drawing rolls on the spinning machine. High-tenacity fibers are drawn on separate roller arrangements. Continuous stretching between rolls mounted at an angle or on godets does not necessarily seem more advantageous than sectional drawing. The total amount of stretch applied ranges between 20 % for standard fibers and up to 150 % for high-tenacity fibers [116-120].

The sulfuric acid for the coagulation bath (ca. 1 kg of 96 % H_2SO_4/kg of fiber) should be highly concentrated (iron storing tanks), of high purity (colorless and free of odor), and free from arsenic. The Fe content should be < 100 ppm, and Cu content < 1 ppm.

The zinc sulfate should not present dispensing problems (hydrated crystals or zinc oxide). Purity requirements are as follows: Cu < 10 ppm, Pb < 30 ppm, Fe/Mn/Ca < 100 ppm each, Mg < 500 ppm, chloride < 0.3 %.

During the spinning process, sodium sulfate is produced in the form of Glauber's salt ($Na_2SO_4 \cdot 10\ H_2O$, ca. 1.2 kg/kg of fiber). It is generally calcined and sold to the glass and detergent industries.

Processing water should be soft and low in heavy metals (total and carbonate hardness ≈ 0, Fe < 0.05 mg/L, Mn < 0.03 mg/L, permanganate index ca. 2.5 mg/L).

Surfactant auxiliaries (modifiers, additions to the spin bath, aftertreatment agents, and finishes) are used at various stages of the viscose process. They must be selected for compatibility. Ethylene oxide additives with or without nitrogen and polyalkylene glycols will, for example, promote xanthation, dispersion, and alkalization, hence, the filter properties and spinnability. Surfactants may be added to the pulp or the alkaline cellulose. For high-tenacity fibers, special additives are used as modifiers [121], [122]. Spin bath additions make for a clearer spin bath and reduce deposits on the spinnerets.

Recovery of Chemicals. The steeping sodium hydroxide pressed from the alkaline cellulose is partially cleaned by dialysis and reused. It dissolves the cellulose xanthate and serves as a washing bath for the regenerated fibers.

Spin bath solution is filtered, vacuum-deaerated (removal of H_2S), reduced, and reused. The Glauber's salt generated in the process is removed and calcined.

The carbon disulfide released during the fiber-making process is collected, freed from hydrogen sulfide (if any), and adsorbed on activated carbon. In this way, up to 70 % of the initial amount can be recovered and distilled for reuse.

Part of the hydrogen sulfide is changed into sodium sulfide (for desulfation baths) and oxidized to pure sulfur or sulfuric acid.

3.1.4. *Fiber Types*

Regular Fibers. Classic spinnable xanthate solution (7 – 10 % cellulose, 5 – 7 % sodium hydroxide, 25 – 35 % carbon disulfide, maturity 10 – 15 °Ho) is spun into an acid salt bath (≧ 80 g/L sulfuric acid, 150 – 300 g/L sodium sulfate, 10 – 20 g/L zinc sulfate) at 45 – 55 °C. The first regeneration trough is 20 – 80 cm long and the fibers are drawn off at a speed of

80 – 160 m/min. The prestretched filaments are washed with cold and hot water and collected on rolls, spools, or cakes for deacidification. Other treatments have the purpose of removing sulfur, bleaching, application of lubricant, and drying.

The process can also be arranged for continuous operation by using wide machines for collective treatment of yarn sheets or single filament treatment. Collective treatment gives larger throughput, more uniform quality, and larger bobbins. Insulation of the machines is, however, more difficult than in single filament treatment, which is less productive, supplies smaller bobbins, and can result in differences within the yarns (between spinning points, as well as over the length of the threads). The threads are normally spun by immersing the spinnerets into the spin bath and drawing off the filaments upwards. This involves the formation of a “jacket” of varying size between the filaments and around the fiber bundle. The size of this jacket depends on the take-off speed. This jacket, which is stripped off, limits the production speed. That is why more recent production systems involve spinning from top to bottom, similar to the cupro process. In principle, a production speed of ca. 350 m/min is feasible for fine single filaments, this speed is currently limited by the drying operation.

Regular viscose fibers have a rather low wet modulus; wet strength is about 60 % of dry strength, and moisture absorption is about 100 %, this is 2.5 times higher than that of cotton. These facts needed to be considered in textile processing and led to the search for fibers with a higher wet modulus, higher wet strength, and reduced moisture absorption. The successful result of these efforts was the so-called modal fibers (staple) and high-tenacity (industrial) filaments.

Crimped viscose fibers are turned into functional fabrics either alone or in blends with wool, polyester, or acrylics. They are based on a relatively “unripe” grade of viscose (15 – 20 °Ho). Xanthation uses the same amount of carbon disulfide as previously described, but spinning conditions are different. The spin bath is less acid, whereas the second bath has a higher acid content (up to 30 g/L sulfuric acid). When the filament passes the second bath, it is stretched to a higher degree; the initial skin breaks up under the tension, and a new irregular skin is formed. This in turn creates tensions over the cross-section of the yarn which deform the fiber into waves and crimps (Fig. 17). Their number per 100 mm can be widely varied, but processing requirements set certain limits (in a 3-dtex staple fiber, the crimp varies with fiber type from 100 to 125 waves/100 mm) [123].

Figure 17. Skin – core structure of a crimped viscose fiber

In cellulose fibers, the crimp durability is inferior to the crimp in synthetic fibers, particularly in the wet state.

Industrial (High-Tenacity) Yarns. When filaments are spun from top-quality viscose (ca. 7.5 % cellulose from high-quality pulp, 7 % sodium hydroxide, 38 – 40 % carbon disulfide) and highly stretched during regeneration (up to 100 %), the result is a *high-tenacity filament* or *cord yarn.*

Supercord yarns, i.e., fibers with a homogeneous full-skin cross-sectional structure, require a starting pulp with a high α-cellulose content and a uniformly high degree of polymerization. Viscose ripening is lowered, and carbon disulfide content is increased. The spin bath consists of less sulfuric acid, but a larger amount of zinc sulfate. Spinning speed is reduced, no stretch is applied in the regeneration bath, but very high stretch ($\gg$ 100 %) is applied in the second bath. Modifiers are added to the viscose and bath, e.g., monoamines, polyethylene glycols, oxyethylated ammonium derivatives, and polyglycols condensed with various organic radicals.

Extra-high-modulus fibers (EHM) are obtained by adding formaldehyde to the spin bath, sometimes also to the viscose itself. This will further increase fiber strength and reduce elongation to < 10 %. The fibers show a certain tendency to split into fibrids.

Medium- and high-wet-strength viscose staple is made under much the same conditions as cord and supercord yarns. Compared to the viscose used as regular staple, the viscose for higher tenacity fibers contains more sodium hydroxide, has a lower degree of maturity, and is spun into a regeneration bath containing up to 40 g/L zinc sulfate and more. Spinning speed is lower, but the drawing ratio higher. To ensure optimal uniformity of coagulation and decomposition of the many thousands of single filaments, it is not unusual to use a tube arrangement as in supercord production. This also allows a certain reduction in the amount of spin bath required per spinning point.

Modal Fibers. Modal fibers are also spun from viscose and come in two varieties: *polynosics* and *HWM* (*high wet modulus*). Both are based on higher quality viscose than regular staple (6 – 8 % cellulose, 6.5 – 8.5 % sodium hydroxide, 40 – 50 % carbon disulfide; small amounts of a modifier and some zinc may also be added; the viscose for polynosics can also contain traces of formaldehyde). Modal fibers — they are always cut to staple lengths — are spun into a slightly acid bath of low temperature and with a strong coagulating effect (20 – 30 g/L sulfuric acid, $\geqq$ 25 g/L sodium sulfate, < 10 g/L zinc sulfate), which is sometimes difficult to recover. The stretch applied in the second bath may exceed 150 %.

HWM fibers can generally be spun on conventional viscose spinning equipment, while polynosics require essential machine modifications. If formaldehyde is added, special precautions in spin bath treatment are necessary.

Cellulose xanthate for polynosics contains relatively long molecules and is spun at lower speeds than regular viscose. Due to the lower acidity of the bath, the number of nuclei formed when the mixture enters the bath is reduced. They grow into long, tapiform crystallites which lend themselves to easy orientation. The result is a network of long, solid crystalline bundles in whose inner regions even the noncrystalline matter shows a certain degree of orientation and, hence, resistance to penetration. As regeneration speed is also reduced, the nascent filaments remain deformable for a longer time, which allows

absorption at higher drawing rates than with HWM fibers. This gives almost complete orientation of the cellulose molecules in the direction of the fiber axis.

In polynosic fibers, the cellulose crystallites are made up of alkali-resistant cellulose II. This is the only type of fiber made from cellulose xanthate which displays the fine fibrils seen in cotton.

In HWM fibers spun in the presence of modifiers, the length of the crystallites and their orientation are somewhat less than those with the polynosics. Moreover, the crystallites contain not only cellulose II, but also orthorhombic cellulose IV [99, p. 90]. Sodium lye of concentrations greater than 4 % causes the lattice of cellulose IV to change into cellulose II. If not subjected to tension during this process, the cellulose II remains unoriented and acts as an instable disturbance within the pattern of ordered macromolecules [124], [125]. This in turn jeopardizes the mercerizing properties and gives a higher elongation than found in polynosics. Practical experience has shown that dry elongation of more than 15 % can result in fiber damage during mercerization.

Addition of formaldehyde or formaldehyde-releasing compounds in the course of the polynosic fiber spinning process will promote gel orientation. Methylene chains will attach to the cellulose chains and strengthen the network to an extent that sometimes allows drawing ratios of 500 – 600 %, a mixed blessing.

3.1.5. *Modified Viscose Fibers*

Using weakly acidic baths (ca. 15 g/L sulfuric acid) containing more than 100 g/L ammonium sulfate, it is possible to spin so-called coagulation fibers (see Section Fiber Properties) from ripe viscose (< 10 °Ho). They have a round cross-section and a scarred surface, are easy to dye, and offer good dimensional stability. Their structure results in reduced strength and, thus, makes production of fine fibers (< 3 dtex) difficult.

With the help of inorganic pigments which are added to the viscose, colorfast spun-dyed fibers are produced. Addition of some 40 % barium sulfate (based on the weight of the fiber substance) gives the fibers X-ray contrast mainly required for medical textiles.

Optical brighteners added to the viscose give a brilliant white. These fibers, however, should not be used for medical purposes. It is also possible to include flame retardants (halogen, phosphorus, and nitrogen compounds). Certain sequestering agents may be used to modify the dye affinity of the fiber so that it accepts acid dyes. Blends of regular and acid-dyeing fibers then allow attractive differential dyeing.

The wet spinning process is well-suited for the addition of substances to the spinning solution. In connection with the viscose process, the additives must, however, be selected for resistance to the strongly alkaline and acid process conditions.

Another fiber modification involves *grafting.* Grafting can be carried out during fiber production as well as in textile finishing. In process grafting, the method should not require extremely drastic changes in the normal schedule and avoid reduction of the spinning speed, which is usually very difficult to manage [126-129].

Depending on the type of monomer or polymer to be grafted (e.g., styrene or acrylonitrile) the grafting process can be initiated by ions ($Ce^{4+/3+}$, $Fe^{3+/2+}$) or free radicals. It is possible to produce radicals on the cellulose by UV radiation in the presence of photosensitive agents [130]. In addition, high-energy rays from γ-ray sources and electron accelerators are used [131-133]. These processes differ in the chronological order of the monomer bath, drying, and source of radiation. The choice of the method is a function of the desired effect, process economics, and the properties of the monomer and polymer. The radiation dose reaches an approximate maximum of 10 J/kg. Yield is affected by the monomer concentration, type of solvent, swelling degree of the textile, and the radiation dose [134]. From an economic standpoint, low-energy radiation, e.g., argon plasma, which as a highly ionized gas emits positive ions and electrons, may also have importance in the future [135].

Grafted fibers are still in a very early phase of commercial use in spite of years of strenuous research efforts. On a pilot scale, some carpet fibers (CV/PAC) as well as flame-retardant, bacteria-resistant, water- and oil-repellent fibers have been spun in the Russia; also included in this research program were acid-protective, ion- and electron-exchanging, complex-forming, antimicrobic, and hemostatic fibers [129].

For more information on fiber modification, ⟶ Textile Auxiliaries.

3.1.6. *Fiber Properties*

Structure and Properties. As with cotton, the physical properties and the reaction of the fibers to chemical treatment is governed by the fiber buildup – the macrostructure (morphology) and the fine structure. Compared with cotton, the structure of regenerated cellulose fibers is much simpler [136], [137]. Unlike natural cellulose fibers with an advanced fibrid system in their secondary walls, made up of bundles of chain molecules (elementary fibrids and crystallites) and higher configurations (micro- and macrofibrids), the regenerated fibers rarely show such bundles. Rather, they have orderly and noncrystalline portions which, however, have flowing boundaries. In regular viscose fibers, formation of fibrids is negligible; in modifed fibers intentionally suppressed. The polynosics, on the other hand, have a distinct fibrid structure in the form of relatively stiff ropes 10 – 20 µm across. With them, the primary and secondary bundles are comparable with natural cellulose fibers.

The crystallinity, size of the morphological units, and degree of orientation in regenerated fibers are functions of the coagulation and regeneration conditions and subsequent drawing [90]. Regarding the crosswise order, it is assumed, that straight and folded chains exist side by side, with the number of straightened chains increasing with increased stretch.

Physical Properties. The structural characteristics listed in Table 20 are the outcome of process conditions and result in certain physical properties as shown in Table 21 for three types of regenerated cellulose fiber. The regular viscose fiber data in Table 20 indicate high swelling, low wet modulus, and the typical shrinking behavior of this fiber.

Table 20. Structural data of regular viscose staple, polynosic fibers, and cotton [138] (Measuring units and methods in parentheses)

	Regular viscose staple	Polynosic fibers	Cotton
Crystallinity, % (X-ray diffraction)	33 – 36	40 – 47	50 – 52
Length of crystalline portions, DP (marginal DP)	60 – 80	110 – 140	60
Thickness of crystalline portions, nm (wide-angle X-ray)	5 – 7	8 – 10	10
Accessibility of OH groups, % (H-D exchange)	60 – 70	50 – 55	50
Chain orientation, % (X-ray diffraction)	70 – 80	80 – 90	
Total chain length, DP (viscosity)	300 – 450	300 – 500	2000
Porosity, cm^3/g [139]	0.016	0.07	0.087

Table 21. Textile data of various viscose fibers

	Regular viscose staple	Modal fibers: Polynosic fibers	Modal fibers: HWM fibers
Titer, dtex	0.9 – 100	0.9 – 4.2	1.3 – 3.0
Breaking strength, conditioned, cN/dtex	1.8 – 3.3	4.0 – 6.5	4.0 – 5.5
Breaking strength, wet (5 % NaOH, 20 °C), cN/dtex	1.3	3.2	2.5
Elongation at break, conditioned, %	18 – 40	6 – 12	14 – 18
Density, g/cm^3 (65 % R.H., 20 °C)	1.50 – 1.52	1.50 – 1.52	1.50 – 1.52
Water retention, %	90 – 120	55 – 70	66 – 90
Moisture regain, % (65 % R.H., 20 °C)	13	11.5 – 12.5	11 – 12.5
Relaxation shrinkage, %	12 – 15	3 – 5	5 – 7
Wet modulus, cN/dtex	5 – 10	14 – 18	10 – 12
Mercerization	not possible	very good	under certain conditions

Table 21 indicates that the modal fibers are superior to regular viscose staple with regard to dimensional stability, i.e., the interaction of water retention, relaxation, and wet modulus. The decisive characteristic is the improved wet modulus. Figure

18 illustrates the characteristic steepness of the stress – strain curve of both wet and dry polynosics. This gives increased dimensional stability to woven fabrics even after repeated washing, as indicated in Figure 19. Their utility in certain end uses did not help the polynosics achieve economic importance as had been expected, possibly due to their fibrillating tendency.

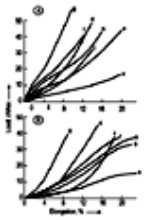

Figure 18. Stress – strain diagrams of various fibers [140]

A) Conditioned; B) Wet

a) Viscose textile grade; b) Viscose industrial grade; c) Texas cotton; d) Egyptian cotton; e) High-wet fiber; f) High-wet modulus fiber; g) Polynosic

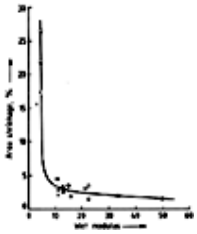

Figure 19. Changes in woven fabric dimensions from 1st to 50th washing [124]

+ Regular viscose staple

□ High-tenacity viscose staple

• HWM fiber, commercial

HWM fiber, experimental

Polynosic, commercial

Polynosic, experimental

Polynosic with formaldehyde

× American cotton I

* Egyptian cotton

Figure 20 shows the lengthwise sections and cross-sections of some regenerated cellulose fibers.

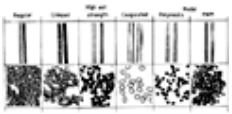

Figure 20. Lengthwise sections and cross-sections of some regenerated cellulose fibers [141]

The main properties of these cellulose fibers are [141]:

Viscose fibers
- Regular viscose:
  - skin – core structure
  - medium tenacity
  - medium elongation
  - 1.3 – 50 dtex (standard program)
  - bright, dull, extra dull
  - optically bleached, spun dyed
- Crimped fibers:
  - thick/thin skin – core structure
  - medium tenacity
  - medium elongation
  - good crimp (dry)
  - 2.2 – 25 dtex (standard program)
  - bright, dull, spun dyed
- High-wet-strength fibers:
  - all-skin structure
  - high tenacity
  - high elongation
  - 1.3 – 3.1 dtex (standard program)
  - bright
- Coagulated fibers:
  - all-core structure
  - low tenacity
  - higher elongation
  - irregular good crimp
  - 3.6 – 25 dtex
  - bright, dull, spun dyed

Modal fibers
- Polynosics:

high tenacity
low elongation
dimensional stability
1.3 – 3.6 dtex
bright
suitable for mercerization

HWM:

high tenacity
low elongation (but higher than polynosics)
fair dimensional stability
1.3 – 3.6 dtex
bright, optically bleached
mercerizable under special conditions

Figure 20 illustrates the data shown in Tables 20 and 21. The dark areas of the cross-sections reflect areas of higher orientation and density. Increased thickness of the skin — up to an all-skin structure — is coupled with distinct increases in fiber tenacity.

Chemical Properties. In several respects, viscose fibers behave in the same way as cotton. All these fibers are attacked by acids and oxidizing agents. If properly executed, hypochlorite and peroxide bleaching is possible. Different fibers react differently to dissolution in sodium hydroxide. Regular fibers show a dissolving maximum in ca. 10 % sodium hydroxide, even quicker as the temperature decreases, and will completely dissolve at – 5 °C. The all-core fibers are more easily dissolved in alkali than skin–core fibers.

For these reasons, polynosic and HWM fibers show different strength losses in sodium hydroxide treatments (Table 22). The hydroxyl groups can be utilized for numerous conversions.

Table 22. Strength loss (in %) of modal fibers in sodium hydroxide of various concentrations (relaxed)

	5 % NaOH	23.5 % NaOH
HWM fibers	40	60
Polynosics	15	30

Dyeing Properties. Similar to natural cellulosics, regenerated fibers can be dyed with vat dyes, Naphtol-AS, sulfuric, direct, and reactive dyes (⟶ Textile Dyeing). Particularly good fastness is obtained with reactive dyes.

Regenerated cellulose fibers absorb much more dye than natural cellulosics. The relationship between production conditions and dyeing properties of regenerated cellulose fibers is extensively described in [88]. It is said that the dye affinity of the regenerated fibers depends on the accessibility, i.e., crystallinity and degree of orientation. HWM fibers are very similar to regular viscose, while the fiber structure of the polynosics puts them nearer to cotton in this respect.

The possibility of making spun-dyed fibers offers some advantages (fastness, cross-dyeing, and low water pollution), but also some disadvantages (minimum order size, variety, and sales disposition).

Similar to dyes also other chemicals can be included in the fiber during spinning if they are resistant to the alkaline and acid processing conditions.

3.1.7. *Uses*

The varied range of natural and man-made fibers is nowadays used by the textile industry with a keener eye on the specific fiber properties. In this way, the cellulosic man-mades have acquired a leading position in certain applications and are blended or combined with other fibers to add their advantageous properties to the textiles.

Viscose fibers are used in many ways for many textile purposes: clothing, home textiles, and industrial textiles. Regular viscose staple is primarily used on its own for outerwear and nonwoven fabrics. Crimped fibers are used alone or in blends with polyester, wool, or acrylics for woven and knitted outerwear. The most successful use of HWM fibers is in blends with cotton and/or polyester in cotton-type fabrics for shirts, blouses, dresses, raincoats, bed linen, and the like. In blends with cotton, the HWM fibers act as a spin carrier and will sustain mercerization under suitable conditions. In polyester blends, they take the place of cotton; their comparatively high cross-strength makes for fabrics with good wearing properties. High-wet-strength fibers have high work-absorbing capacity (high strength at relatively high elongation) and, therefore, lend themselves as industrial textiles and nonwoven interlinings.

Viscose filament yarns are used mostly in lining fabrics as well as in ladies' outerwear and home furnishings. High-tenacity yarns continue to play an important role in industrial textiles such as automobile tires, conveyor belts, and coated fabrics.

All end uses profit by the following viscose properties: easy processing in the textile and finishing stages, hydrophilic behavior, and attractive cost/benefit ratio. Disadvantages derive from the following: relatively high density in comparison with synthetic fibers and low strength, particularly in the wet state, as well as moderate dimensional stability, which, due to the hydrophilic nature of the fiber, may be unsatisfactory for certain uses. The cost/strength ratio is also rather low. The lack in dimensional stability can often be made up by suitable after-treatment of the fabrics. Care must, however, be taken not to excessively damage the physical properties of the textile.

Viscose fibers hold a secure niche in the textile field. Since raw materials for viscose fiber production are secure, viscose can be expected to continue on an upward trend.

Outlook. Regenerated cellulose fibers hold a firm place among our textile raw materials. Solution of production problems would, therefore, be a sensible proposition. One of the aims should be an optimal cost/benefit ratio.

Regenerated cellulose fibers are popular materials for clothing, home textiles, and industrial textiles for a number of reasons. This also applies to nonwovens for household, personal care, and the wide field of medical applications, where absorption and cleanliness are primary considerations. Possible substitution by fluff has not yet been fully explored.

In the area of industrial fibers the unfavorable cost/tenacity ratio of cellulosic in comparison with synthetic fibers will result in a shift in the quantities produced in favor of synthetic fibers. This is already discernible from the figures given in Table 23. The ecological problems connected with the production of viscose fibers promote investigations in the improvement of process conditions and the search for sulfur-free processes (see Sections Cuprammonium Fibers and Tentative Cellulose Fiber Production by Other Processes – Outlook).

3.1.8. *Economic Aspects*

Growth of the world population and higher standards of living also caused an increase in the demand for textile fibers as illustrated in Table 23. It also indicates that the demand for cellulosic man-made fibers has markedly decreased and lost importance relative to the synthetics.

Table 23. World production of man-made fibers 1970 – 1996 [160]

	1970		1980		1990		1996	
	10^3 t	%	10^3 t	%	10^3 t	%	10^3 t	%
Cellulosics	3 585	43	3 581	25	3 145	16	2 900	12
Synthetics	4 809	57	10 703	75	16 006	84	21 200	88
Man-mades total	8 394	100	14 284	100	19 151	100	24 100	100

The fact that production of staple fibers kept growing slightly, while filament yarns lost more and more ground is not shown. This was a consequence of development in the nonwoven field, in garment making, and in the tire yarn industry.

Regional differences in cellulosic man-made fiber production are also interesting. Approximately 22 % is made in Western Europe, ca. 8 % in the United States, 10 % in Japan, and the remainder of more than 60 % is distributed among the other regions. Production sites are being shifted to the Far East to a certain extent (see Table 24). Of the production methods, the viscose process is by far the most important.

Table 24 gives only general trends because of differences in the computing methods and includes cigarette filter tow (acetate) as well as material for medical use (Cupro).

Table 24. Production development of cellulosic fibers in different parts of the world (in 10^3 t) [160]

	1985	1990	1995
World	3 217.6	3 146.3	3 040.4
Western Europe	709.0	700.1	684.0
Eastern Europe	994.4	833.0	272.6
USA and Canada	287.7	268.9	240.1
Japan	387.7	275.8	297.1
People's Republic of China	180.0	191.7	462.4
India	142.0	216.7	262.2
Indonesia	40.0	60.0	177.0

Depending on fiber type (titer, textile, or industrial grade), viscose filament production units should have a capacity of 10 – 30 000 t. Staple fiber units, on the other hand, should be of an even larger size, i.e., ≧ 50 000 t and more. The lower limits are governed by economic considerations and the upper limits by ecological aspects.

The future prospects of cellulose fiber production are not clear. The industry is faced with increasingly stricter environmental regulations; most production equipment is old and the market seems convinced of the advantages of synthetics. Positive

aspects of cellulosics are their hydrophilic properties, quantitative importance and the long-term availability of raw materials. These are important reasons for finding solutions to the problems so that adequate textile supplies may be secured for many years to come. Thus, in the long run, the importance of the regenerated cellulose fibers will undoubtedly increase once again.

3.2. Lyocell Fibers

In view of the problems associated with the viscose process, attempts have been made for decades to find a “shorter, simpler, sulfur-free process” for the production of cellulosic man-made fibers (see also Section Tentative Cellulose Fiber Production by Other Processes – Outlook). Of the various possible approaches, the NMMO process has now been developed to the commercial stage. The process involves dissolution of cellulose in aqueous *N*-methyl morpholine *N*-oxide (NMMO) solutions and subsequent regeneration from this solvent. This process is distinguished from other methods of producing cellulosic man-made fibers by the use of an organic solvent. This also determines the specific property profile of the fibers which justifies that they are classified as separate category within the group of cellulosic man-made fibers.

The basic principles of this process were developed around 1976 by AEC (American Enka Corp., United States) and ARLO (Akzo Research Lab. Obernburg, Federal Republic of Germany) and patented in 1980 [91]. After having discontinued the production of cellulosic staple fibers in 1983, Akzo Fibres granted licences to Lenzing in 1987 and to Courtaulds in 1990 and concentrated their development activities on filament yarns [92].

Courtaulds produced some 70 000 t/a of staple fibers under the trade name Tencel in Mobile, United States and Grimsby, United Kingdom. Lenzing has started production of staple fibers with a production volume of 12 000 t/a in Heiligenkreuz, Austria in 1997, while Akzo Nobel Fibres are preparing for commercial production of filament yarn, following successful conclusion of a market survey based on products from their Obernburg pilot plant. The technology is now also being offered to interested patries by the Thüringisches Institut für Textil- und Kunststoff-Forschung (TITK, Thuringia Institute for Textile and Plastics Research) in Rudolstadt, Federal Republic of Germany.

3.2.1. *Principles of the Lyocell Process*

The cellulose briefly described in Section Principle of the Viscose Process, NMMO, and water are mixed, after adding a stabilizer, to give a homogeneous solution [85], [92-94]. The solution is extruded into a regeneration bath consisting of an aqueous NMMO solution to obtain the desired cellulosic man-made fibers (or films). The NMMO used can be recovered at > 99 % and recycled.

A schematic of the process is depicted in Figure 21. As Figure 21 shows, the process — like the production process for synthetic fibers — is short and the regeneration bath can be recycled almost completely. As in the viscose process, certain substances can be added to the solution to impart specific properties to the fibers.

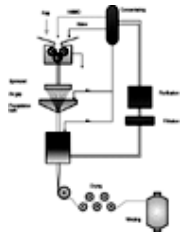

Figure 21. Lyocell process

3.2.2. *Process Description*

The cellulose is introduced into a mixture of NMMO and water. The slurry (which should be as homogeneous as possible) of, e.g., 13 % cellulose, 20 % water, 67 % NMMO, and some stabilizer is then adjusted to 14 % cellulose, 10 % water, 76 % NMMO, and stabilizer by hydroextraction, to dissolve the cellulose (see phase diagram, Fig. 22). The solution of this composition is extruded at temperatures slightly higher than 100 °C into an aqueous NMMO bath whose composition is beyond the solution range for cellulose. The fibers or films thus obtained are washed and dried. The regeneration bath is purified, concentrated and reused [92].

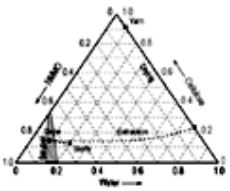

Figure 22. Phase diagram

3.2.3. *Fiber Properties*

Although lyocell fibers also consist of regenerated cellulose II, their properties differ substantially from those of other cellulosic man-made fibers. This difference can best be demonstrated on finished textiles. That is why Table 25 compares the requirements for four textiles typically made from viscose filament yarn with the performance of traditional viscose and lyocell yarns. Table 25 shows that the requirement profiles of the textiles are met by conventional yarns, except for three types of embroidery yarns. It is also evident from Table 25 that with lyocell fibers comparable but lighter textiles with equal good use properties can be obtained which correspond to fashion trends and customers' wishes [92].

Table 25. Comparison of requirement for four typical viscose yarns with performance properties of classical viscose and lyocell yarns, according to [92]

	Requirement for ladies' outerwear					Performance	
	General requirements	L	S	C	E	Enka Viscose (bobbin)	NewCell (current performance)
Total count , dtex	50 – 600	=	=	=	=	67 – 660	55 – 110
Single filament count, dtex	1.5 – 8	=	=	=	=	2.7 – 7.5	1.1 – 2.2
Cross section		=	=	=	=	lobal	round
Uniformity	high	=	=	=	=	+	+
Degree of luster	bright-extra dull	=	=	=	=	+	bright-dull
Tenacity							
dry, cN/tex	≥ 16	=	=	=	>	16 – 20	34 – 40
wet, cN/tex	= 10	=	=	=	=	7 – 10	22 – 27
Elongation							
dry, %		=	=	=	=	16 – 21	6 – 12
wet, %		=	=	=	=	20 – 26	8 – 14
Knot strength, cN/tex	13	=	=	=	>	13 – 16	23 – 27
Loop strength, cN/tex	13	=	=	=	>	13 – 16	18 – 23
Shrinkage at the boil, %	≦ 2	=	=	=	=	0.5 – 2.4	0.5 – 2.0
Moisture gain, %	> 10	=	=	=	=	11 – 14	11 – 13
Water inhibition, %		=	=	=	=	95	60 – 70
Natural shade (whiteness)	natural white	=	=	=	=	+	(+)

L = Lining fabrics; S = Ladies' outerwear from flat yarn; C = Ladies' outerwear from crêpe yarn; E = Embroidering yarn; NewCell = Trade name of lyocell filament yarn from Akzo Nobel; Enka Viscose = Trade name of viscose filament yarns from Akzo Nobel.
= Can be produced without problems, no discernible differences in properties for consumer; > Production advantages; + Better than reqired.

However, lyocell fibers have a special property that must be taken into account in processing and use. When subjected to a rubbing treatment, they will — especially in the wet state — form fibrils influencing the feel and appearance of the fabric. This is of advantage for producing a soft handle and/or a peach-skin fabric surface. If this is not desired, various possibilities exist to eliminate this property partially or totally (either during fiber production or aftertreatment of the textiles). These examples and the consequences for fabric design demonstrate that lyocell fibers are no substitutes for viscose fibers, but a fiber group of its own which, when put to appropriate use, opens up new possibilities for textile design without causing the kind of problems associated with viscose fibers [95], [96].

The characteristic of lyocell fibers can also be advantageously utilized in hetero-yarns, specialty textiles, and nonwoven fabrics with specific properties.

3.3. Cuprammonium Fibers

Present Situation. Production of cellulosic fibers after the copper oxide ammonia process (cuoxam process) began to decline in the early 1960s. The reasons were the outdated production equipment, the limited possibilities to increase the variety of fiber types available, and the sweeping success of the synthetics. At that time, virtually no one would have predicted the comeback of cuprammonium fibers. Moreover, elimination of the Cu ions from wastewaters seemed an unsurmountable obstacle (hazard for biological water reprocessing). In addition, the fibers were suitable for only a limited range of textiles that was waning under pressure from the synthetics. On the other hand, the largest producer of this type of fiber, Bemberg AG (now Akzo Nobel Fibres AG), diverted its research and development efforts from 1965 on to membranes and used the cuoxam solution for the manufacture of flat and tubular membranes and hollow fibers, currently the standard material worldwide for blood dialysis (Cuprophan). In artificial kidneys, they serve hundreds of thousands of kidney patients as the first artificial organ ever made [142-145]. Membrana — a business unit of Akzo Nobel Fibres — is the only producer of these hollow fibers in the Federal Republic of Germany. Cuprammonium fibers are also made in other countries: in Italy by Bemberg S.p.A., Gozzano (Novara) (filament); in Russia by Chemical Fiber Combine Kalinin (staple); and in Japan by Asahi, Nobeoka (filament and staple).

In view of the decreased importance of cuprammonium filaments and staple, we shall limit ourselves to a short description of the process principles [146-148].

Production. There are no essential differences between filament and staple production. Linters or cellulose pulp, or a

mixture of them is opened and mixed with copper hydroxide or a basic copper salt in the presence of a highly concentrated solution of ammonia in water. This results in a viscous spinning solution, which is filtered and deaerated. The method is simpler than the viscose process. The filaments are spun after Thiele's spin-drawing method by extrusion through relatively large spinning holes into a vertical water stream and high stretching. This is followed by copper extraction, neutralization, finish application, and drying. Currently, the process is continuous from start to bobbin/section beam, but some “historical” apparatus is still in use for separate spinning/winding, scouring, drying, and spooling.

Raw Materials. Starting material is generally bleached linters, but a mixture of wood pulp and linters is also occasionally used. It is customary to prepare the linters in a 1 % cuoxam solution at 20 °C to a 25 mPa · s viscosity; wood pulp is adjusted to 16 – 24 mPa · s viscosity at 20 °C, reflecting a polymerization degree (DP) of 1 000 – 1 200 and 800 – 1 000, respectively. Peculiar in cuprammonium fiber spinning from pulp is the fact that the cellulose is not treated with sodium hydroxide, with the result that none of the hemicellulose is washed out. For this reason, the pulp should have a low hemicellulose content (< 4 %).

The cellulose is dissolved in basic copper sulfate — obtained from recycling — in complex compounds with ammonia.

Spinning Solution. As a rule, the cellulose is moistened and immersed in the dissolving liquid made up of ammonium hydroxide and copper salt (basic copper sulfate or copper chloride). Since heat is released during the reaction (ca. 40 kJ/kg spinning solution), cooling is recommended so that the temperature does not exceed 25 °C. Heating causes discoloration due to the reaction of copper hydroxide with ammonia, cellulose with copper(II) aminehydroxide hydrate, and the basic copper salt with sodium hydroxide giving ammonium copper(II) hydroxide hydrate. Starting as a highly viscous mass, agitation requires very strong straight or helical mixing blades. The spinning solution is then filtered through presses equipped with noncorroding metal fabric packs (nickel or stainless steel), mesh size 40 – 70 µm. Finally, the filtered spinning solution is deaerated to remove air bubbles; at the same time some of the excess ammonia is eliminated. Depending on the spinning conditions, the solution will then contain 4 – 11 % cellulose, 4 – 6 % Cu, and 6 – 10 % NH_3, and have a viscosity of ca. 200 Pa s at 20 °C. When spun-dyed or delustered fibers are desired, inorganic and/or organic pigment dyes are added.

The solution is very stable. Stored in air-tight vessels, its composition will not change and the Υ-value remains constant, ensuring unchanged coagulating properties.

Spinning Methods [148], [149]. Cuoxam cellulose solutions are always spun according to the wet process (⟶ Fibers, 3. General Production Technology), with one or two bath passages. Strong coagulants are, for example, concentrated sodium hydroxide [150] or 3 – 30 % sulfuric acid. Single-bath wet spinning in NaOH is used for coarse filaments such as upholstery stuffing and hollow fibers [151], [152]. In this case, a bath like that in viscose fiber production is used. A commercial single-bath process using acid and salt is under consideration.

Thus far, however, the Thiele spin-drawing method in a virtually neutral bath is the rule. Vertical spin-draw funnels of glass or transparent plastic are used [153]. The spinning speed of about 150 m/min is very slow compared with the speeds in synthetic fiber spinning. Trials were made to adapt the Thiele funnel spinning method for higher speeds as well.

In several patent applications, Asahi Kasei Kogyo [154] describes methods in which the threads coming from conventional spinning funnels are led over longer drawing and coagulation distances to be finally deposited in a random layer on a perforated conveyor belt for transportation through the subsequent production stages (copper extraction, washing, finish application, and drying). This method is claimed to allow spinning speeds of up to 800 m/min.

Aftertreatment. The plastic and, therefore, mechanically sensitive yarns made up of 20 – 230 single filaments are freed from copper and aftertreated in one continuous operation.

The sheet of parallel threads spun by batteries of several hundred spinning points in staggered rows (thread spacing = 2 – 4 mm) meanders through a bath of 2 – 6 % sulfuric acid and is rinsed in two or three water baths to remove the acid. Finally, the yarn sheet is immersed for finishing and dried on heated drums or in drying chambers to a residual moisture content of ca. 11 % for winding onto beams or bobbins.

American Bemberg Corp. (Beaunit Mills Inc.) introduced a continuous spinning method, in which the funnels are arranged at an angle in a staggered row and the freshly spun yarns are led as a sheet on a straight path through shallow troughs. After the yarns have dried, they are wound onto bobbins [155]. This method is still used in Italy and Japan.

Cuprammonium Nonwovens. In [156], an apparatus for the production of nonwovens from cuprammonium filaments is described. It follows the funnel spinning and drawing principle, although with a rectangular spinneret arrangement from which the threads (ca. 90 000) cascade onto a partially permeable conveyor belt which traverses and causes crosswise entanglement of the yarns. As cuprammonium filaments will securely bake together in an ammoniacal medium, a nonwoven fabric is formed whose thickness and stability can be varied via crossing points and fusion. Production speed is claimed to be ca. 35 m/min.

Cuprammonium Staple. The spinning solution is almost the same as in filament spinning, except that a 1 – 2 % lower ammonia concentration is used to reduce the tendency of the filaments to stick together. Sticking may help in filament spinning, but is disturbing in staple production.

The dimensions of the funnels are much larger. The spinnerets are 80 – 100 mm across and have 2 000 – 3 000 holes of 0.5 – 1 mm across. The filaments emerging from 50 – 100 funnels are collected into thick tows. Aftertreatment takes place in the form of filament tow or cut to staple lengths.

For conversion into fiber laps, 5 – 10 tows are laid parallel and led through troughs of 50 –100 mm width and 10 – 20 m length and are rinsed in the same way as filament yarns. After drying, the moisture content is readjusted to 11 %. Alternating

application of strain and relaxation during wet treatment gives the filaments a regular crimp [157].

Staple fibers with a coarse irregular crimp and a rough surface are obtained if the freshly spun plastic threads are cut into 40 – 150 mm lengths on special machinery [158] and deposited on screen belts for tensionless aftertreatment [146, p. 273], [159].

Recovery of Chemicals. The copper from the ammonia baths is bound by ion exchangers. The “spinning acids” obtained during acid removal and regeneration of the “blue threads” are used to regenerate the ion exchangers. In this way, the acids concentrate around the copper. Neutralization with alkaline water gives basic copper sulfate, which is filtered off and reused to set the spinning solution.

Lately, electrolytic methods have been devised. Copper recovery is $\geqq$ 99 %. In addition, part of the ammonia — some 40 – 50 % — is recovered by distillation of the spin bath in special columns.

From the ecological point of view, the cuoxam process no longer poses any problems. At present, it is superior to the viscose process in this respect.

Properties. The most striking properties of cuprammonium fibers are fineness of the filaments, silklike appearance, and silky touch. In this respect, they are far superior to viscose yarns. Moreover, their inner structure makes hollow fibers particularly suitable for blood dialysis. In other respects, they are similar to viscose filaments without, however, reaching the high strength level of the industrial yarns in commercial production and high crimp properties.

3.4. Tentative Cellulose Fiber Production by Other Processes – Outlook

As mentioned earlier, the image of the cellulosics had come under pressure. This first resulted in resignation, as may be inferred from the fewer publications, which are a measure for research activity. After the first oil crisis in November 1973, public thinking changed. In the cellulose fiber field, various activities were initiated. Work was resumed to improve the viscose process with the intention of reducing air and water pollution; other investigators looked for ways of eliminating sulfur from the dissolving process. Assistance came from the plant breeding side in the form of higher cellulose yield per acre.

Air and water pollution was attacked from three sides: first, the reduction of chemicals to limit the problem; second, collection of chemicals where they are generated, concentration where possible, and transportation to ancillary treatment equipment; and third, improvements in the waste treatment plants. For details on waste air treatment see [82]. Suitable custom-tailored water treatment plants are available.

Besides, it must be remembered that the viscose process has set standards of cost, quality, and variety, which any new process must at least equal, if not surpass. Research which shows the possibility that cellulosics may be produced by more efficient methods gives reason for hope [82].

Successful research using NMMO as solvent for cellulose led to the start of staple fiber production (Courtaulds in the United States and the United Kingdom and Lenzing in Austria; see Section Lyocell Fibers). It is anticipated that also filament yarns will be produced by this process (Akzo Nobel Fibres). The limited variations in the process limit the variety of fiber types produced and their area of application. Experiments with cellulose carbamate [82], [84], zinc chloride – water, sodium hydroxide – water [85], and DMA/LiCl, DMF/LiCl, or DMSO/LiCl [82] as solvents have not been successful thus far. The fibers were either strong enough but too low in maximum elongation, or they were unsatisfactory as far as all physical properties were concerned. In addition, problems arose with optimizing certain production parameters (cellulose concentration, temperature, production conditions, utilization of byproducts, recycling of auxiliaries, etc.).

The insight gained from these experiments and the decades of experience with viscose processing, combined with progress in the petrochemical field, will hopefully lead to useful regenerated cellulose fibers without the help of sulfur compounds. It was intended that the new sulfur-free process should also produce a large variety of types of cellulosic fibers. According to the experience gained, this variety can probably best be obtained by modifications in the regeneration process of dissolved cellulose derivatives.

[Top of Page]

4. Ground Cellulose/Powdered Cellulose

Marc Mohring and Harald Schlosser

4.1. Production

Powdered cellulose is obtained from pulp cellulose by milling and fractionation (Figure 23). Because cellulose, hemicellulose, pectin, and lignin are all incorporated into pulp, chemical reprocessing (chemical pulping) is required for the preparation of pure cellulose. The Kraft (sulfate) or sulfite process is usually used (⟶ Paper and Pulp – Pulp). The preferred raw material for the economical production of cellulose is wood followed by annual plants such as straw.

Figure 23. Production of powdered cellulose

Powdered cellulose (Figure 24, top) is obtained from cellulose in several milling and fractionating steps. Size reduction can be performed with jet, pinned-disk, beater, cutting mill, or hammer mills, or in pan grinders. Fibrillated powdered cellulose (Figure 24, middle) is obtained by wet grinding of the cellulose. Products with a defined particle-size distribution (Figure 25) are obtained by sieving and classifying.

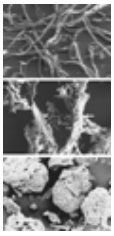

Figure 24. SEM images (×300) of powdered cellulose and microcrystalline cellulose

Top: powdered cellulose; Middle: powdered cellulose, highly fibrillated; Bottom: microcrystalline cellulose

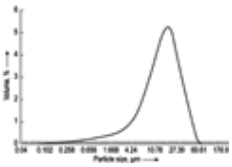

Figure 25. Particle-size distribution of a powdered cellulose (Arbocel BE 600-10 TG)

To minimize dust formation, improve flowability, and reduce the volume for transport, processes for compacting, granulating, pelletizing, and extrusion of cellulose have been developed.

Microcrystalline cellulose (MCC; Figure 24, bottom) is produced from highly pure cellulose. The more easily accessible amorphous components are released by partial hydrolysis with dilute mineral acids. In this way the crystalline fraction is increased, while the degree of polymerization is reduced by acid treatment to an average DP value (degree of polymerization) of 300. In contrast, native cellulose has a DP of ca. 3000.

4.2. Properties

The physical properties of cellulose are determined by the raw material and the milling technique used. The milling process not only regulates the fiber length and fiber thickness, it also influences fiber properties such as water-binding capacity.

Selected properties of some powdered celluloses (Arbocel) and of a microcrystalline cellulose (Vivapur) are listed in Table 26.

Table 26. Selected properties of some powdered celluloses (Arbocel) and a microcrystalline cellulose (Vivapur)

Parameter	Arbocel BE 600–10 TG	Arbocel BE 00	Arbocel BC 1000	Arbocel FIF 400	Vivapur 101
Fiber length, µm	18	120	700	2000	50
Fiber diameter, µm	15	20	20	35	50
Whiteness, %	85 ± 5	86 ± 5	86 ± 5	82 ± 5	86 ± 3
Shape	fiber	fiber	fiber	fiber	particle
Bulk density, g/L	230 – 300	150 – 180	30 – 45	10 – 25	320
Ash content, %	0.3	0.3	0.3	0.3	0.05
pH value	6 ± 1	6 ± 1	6 ± 1	6.5 ± 1	6.5 ± 1
Water-binding capacity, %	400	520	1200	1350	300

Fiber lengths from 18 to 2200 µm can be produced. The individual fractions have a production-dependent particle-size distribution.

The *fiber diameter* is set by nature and differs corresponding to the type of cellulose.

Whiteness is determined with a spectrometer at 461 nm. Highly pure cellulose has a degree of whiteness of over 90 %.

The *bulk density* is dependent on the fiber length. By using a compacting process, the bulk density can be increased up to 600 g/L. Bulk density is determined according to DIN 53468.

For determination of the *ash content*, the sample is heated at 850 °C in a muffle furnace. Highly pure cellulose has an ash content of max. 0.3 %.

The *pH value* is determined according to NF XVII.

The *water-binding capacity* is determined according to the AACC method. The water-binding capacity depends on the fiber length and the degree of fibrillation. Cellulose fiber can bind up to 15 times its own weight of water.

4.3. Uses

The large range of applications for powdered cellulose can be divided into two main categories:

1) Cellulose functions as a process fiber during the production of semi-finished and end products; it helps to speed up processing and to reduce manufacturing costs.
2) The properties of intermediate or end products can be improved by the use of cellulose: increasing viscosity or improving flowability of dustlike products by converting to resoluble granules.

Cellulose has been used since the mid-1990s as a *disintegration agent* for laundry detergent tablets. Laundry detergent tablets are mainly used in Europe. They have a market share of ca. 10%. Cellulose fiber is also used as a disintegration agent for pigment granulates. Swelling pressure is built up by the capillaries of the cellulose. The fibers increase their volume several times and thus lead to a quick and complete disintegration of granulates, pellets, and tablets.

In spice mixes and grated cheese, cellulose fiber is used as a *flow agent* and *anticaking aid*. It acts as spacers between the individual particles and prevents clumping of the particles by its water-binding capacity.

Cellulose acts as a *transport medium for liquids*: it extracts water and dries but can also bring liquids into a product. Because of their coiled shape, the fibers forms drainage channels and thus speed up the transport of liquids. This transport step is not directionally fixed, but rather is determined by a concentration gradient. During drying water must be transported from the interior to the surface. Thus, for example, in the manufacture of cardboard articles, the drying time can be shortened and energy costs reduced. When manufacturing synthetic leather, cellulose is used to speed up the rinsing out of the organic solvents and thus optimize the precipitation process.

Cellulose fibers are used as a *filter aid*. They form a filter cake in suspensions with only a small amount of solids. Slimy solids form dense, compact filter cakes during separation which are difficult for the filtrate to penetrate. The fibers ensure a loose, porous filter cake with good permeability. Some examples of applications are filtration of titanium dioxide (sulfate process), beverages, dextrin (starch gum), and glucose.

In pharmaceuticals and in technical applications such as dishwasher detergent tablets, cellulose is used as a *tabletting aid*. Cellulose makes it possible to reduce the pressing power but maintain the same tablet hardness, increasing the running efficiency and output of the tabletting press.

Due to the high length-to-width ratio (up to 100:1) and the high degree of fibrillation, the fibers form a three-dimensional network. This reinforcing function of the fibers can be used to lend food products an improved texture. In the construction industry (e.g., cement tile adhesive, concrete spacers) the fibers provide improved structural stability.

[Top of Page]

References

General References

1. E. Treiber (ed.): *Die Chemie der Pflanzenzellwand*, Springer Verlag, Berlin-Göttingen-Heidelberg 1957.
2. A. Frey-Wyssling: *Die pflanzliche Zellwand*, Springer Verlag, Berlin-Göttingen-Heidelberg 1959.
3. S. A. Rydholm: *Pulping Processes*, Interscience Publ., New York 1965.
4. N. I. Nikitin: *The Chemistry of Cellulose and Wood*, Academy of Science of the USSR, Institute of High Molecular Compounds (Engl. Translation: Israel Program of Scientific Translation, Jerusalem 1966).
5. E. Ott, H. M. Spurlin: *Cellulose and Cellulose Derivatives*, vols. I – III, Interscience Publ., New York 1954 – 1955.
6. N. M. Bikales, L. Segal: *Cellulose and Cellulose Derivatives*, vols. IV and V, Wiley-Interscience, New York 1971.
7. J. C. Arthur (ed.): “Cellulose Chemistry and Technology,” *ACS Symp. Ser.* 1977, no. 48.

Specific References

8. W. Sanderman, *Holz Roh Werkst.* **31** (1973) 11.
9. A. Payen: *Troisième mémoire sur le développement des Végétaux*, Extrait des Mémoires de l'Academie Royale des Sciences, Tome III des Savants Etrangères, Imprimerie Royale, Paris 1842.
10. W. N. Haworth, *Helv. Chim. Acta* **11** (1928) 534; Links *Ber. Dtsch. Chem. Ges. A* **65** (1932) 43.
11. H. Staudinger: *Die hochmolekularen organischen Verbindungen – Kautschuk und Cellulose*, Springer Verlag, 1932 (reprinted 1960).
12. A. J. Michell, H. G. Higgins, *Tetrahedron* **21** (1965) 1109. Links
13. O. Ellefsen, B. A. Tonnesen in N. M. Bikales, L. Segal (eds.): *Cellulose and Cellulose Derivatives*, Interscience, New York 1971, part IV, p. 151.
14. V. S. R. Rao, P. R. Sundararajan, C. Ramakrishnan, G. N. Ramachandran in G. N. Ramachandran (ed.): *Conformation of Biopolymers*, Academic Press, London-New York 1957, p. 721.
15. D. Fengel, G. Wegener: *Wood – Chemistry, Ultrastructure, Reactions*, Chap. 6.5, De Gruyter, Berlin-New York 1984, p. 167.
16. G. V. Schulz, E. Husemann, *Z. Naturforsch.* **1** (1946) 268.
17. K. Freudenberg, C. Blomquist, *Ber. Dtsch. Chem. Ges B* **68** (1935) 2070.
18. E. Treiber (ed.): *Die Chemie der Pflanzenzellwand*, Springer Verlag, Berlin-Göttingen-Heidelberg 1957, p. 142.
19. J. Schurz: “Theoretische Grundlagen der Viskoseverfahren,” in K. Götze (ed.): *Chemiefasern,* vol. I, Springer Verlag, Berlin-Heidelberg-New York 1957.
20. V. P. Shanbhag, *Ark. Kemi* **29** (1968) 1.
21. H. Krässig, *Papier (Darmstadt)* **26** (1971) no. 12, 841.

22. E. Gruber: “Microgel Particles in Solutions of Cellulose and Cellulose Derivatives,” *Cellul. Chem. Technol.* **13** (1979) 259.
23. K. H. Meyer, H. F. Mark, *Z. Phys. Chem. Abt. B* **2** (1929) 115.
24. K. H. Meyer, L. Misch, *Ber. Dtsch. Chem. Ges. B* **70** (1937) 266; *Helv. Chim. Acta* **20** (1937) 232. Links
25. A. Viswanathan, S. G. Shenouda, *J. Appl. Polym. Sci.* **11** (1967) 659. Links
26. R. H. Marchessault, C. Y. Liang, *J. Polym. Sci.* **43** (1960) 71.
27. J. Blackwell, F. J. Kolpak, K. H. Gardner: “Structure of Native and Regenerated Celluloses,” in J. C. Arthur (ed.): “Cellulose Chemistry and Technology,” *ACS Symp. Ser.* **1977,** no. 48, 42.
28. G. Honje, W. Watanabe, *Nature (London)* **181** (1958) 326.
29. K. C. Ellis, J. O. Warwicker, *J. Polym. Sci.* **56** (1962) 339.
30. K. H. Gardner, J. Blackwell, *Biopolymers* **13** (1974) 1975. Links
31. A. Viswanathan, S. G. Shenouda, *J. Appl. Polym. Sci.* **15** (1971) 519. Links
32. R. H. Atalla in T. E. Timell (ed.): “Proc. VIIth Cellulose Conference, Syracuse 1975,” *Polym. Sci. Symp.* 1976, no. 28, 659.
33. S. Watanabe, J. Hayashi, T. Akahori, *J. Polym. Sci.* **12** (1974) 1065.
34. A. Nissan, according to D. W. Jones in N. M. Bikales, L. Segal (eds.): *Cellulose and Cellulose Derivatives*, Wiley-Interscience, New York-London-Sidney-Toronto 1971, p. 118.
35. R. S. J. Manley, *J. Polym. Sci. Polym. Phys. Ed.* **9** (1971) 1025.
36. K. Hess, H. Mahl, E. Gütter, *Kolloid Z.* **155** (1957) 1; **158** (1958) 115.
37. O. Kratky, G. Porod in H. A. Stuart (ed.): *Physik der Hochpolymeren,* vol. 3, Springer Verlag, Berlin-Göttingen-Heidelberg 1955.
38. A. Frey-Wyssling: *Submikroskopische Morphologie des Protoplasmas und seiner Derivate*, Gebr. Bornträger, Berlin 1938.
39. A. Frey-Wyssling: *Submicroscopic Morphology of Protoplasma*, Elsevier, Amsterdam 1949.
40. J. W. S. Hearle in J. W. S. Hearle, R. H. Peters (eds.): *Fiber Structure*, Butterworth – The Textile Institute, London-Manchester 1963.
41. H. Krässig, *Colloid Polym. Sci.* **259** (1981) 1.
42. H. Krässig, *Papier (Darmstadt)* **23** (1969) 881.
43. H. Krässig, *Appl. Polym. Symp.* 1976, no. 28, 777.
44. H. Staudinger, K.-H. In den Birken, M. Staudinger, *Makromol. Chem* **9** (1953) 148.
45. J. A. Howsman, W. A. Sisson in E. Ott, H. W. Spurlin, M. W. Grafflin (eds.): *Cellulose and Cellulose Derivatives*, Interscience Publ., New York-London 1954, part I, Chap. IVB, p. 326.
46. H. Krässig, *Tappi* **61** (1978) no. 3, 94.
47. L. Jörgensen: *Studies on the Partial Hydrolysis of Cellulose*, E. Moestue, Oslo 1950.
48. H. Krässig, *Papier (Darmstadt)* **38** (1984) 571.
49. H. Krässig, *Makromol. Chem.* **26** (1958) 17.
50. O. Samuelson: “Some Undesirable Carbohydrate Reactions During Alkaline Cooking and Bleaching,” in: *The Ekman Days 1981,* vol. 2, Int. Symp. Wood Pulp Chem., Stockholm 1981, p. 78.
51. Y. Z. Lai: “Kinetics of Base Catalyzed Cleavage of Glucosidic Linkages,” in: *The Ekman Days 1981,* vol. 2, Int. Symp. Wood Pulp Chem., Stockholm 1981, p. 26.
52. B. Lindberg, *Sven. Papperstidn.* **59** (1956) 531.
53. H. Esterbauer, M. Hayn, G. Jungschaffer, E. Taufratzhofer, J. Schurz: “Enzymatic Conversion of Lignocellulose Materials to Sugars,” *J. Wood Chem. Technol.* **3** (1983) no. 3, 261. Links
54. I. Jullander, *Papier (Darmstadt)* **19** (1965) 166, 224.
55. H. Krässig, V. Stannett, *Fortschr. Hochpolym. Forsch.* **4** (1965) 111.
56. J. Schurz: “Chemical Combinations of Natural and Synthetic Polymers,” in A. Varmavuori (ed.): *Proc. IUPAC-27th Int. Congress of Pure and Applied Chemistry,* Pergamon Press, Oxford-New York 1980, p. 307.
57. A. J. Turner, *J. Text. Inst.* **40** (1949) 973.
58. Food and Agriculture Organization: *1982 Handbook of Trade*, United Nations, Rome 1983.
59. P. Kassenbeck, *Text. Res. J.* **40** (1970) 330. Links
60. A. Frey-Wyssling, K. Muhlethaler, R. W. G. Wyckhoff, *Experientia* **4** (1948) 475.
61. H. U. H. Dolmetsch, *Text. Res. J.* **39** (1969) 568. Links
62. H. Zahn, “Latest Findings on the Micro-Structure of Cotton,” *Internat. Cotton Conf.*, Bremen 1988, p. 7.
63. J. Sugiyama, H. Harada, Y. Fujiyoshi, N. Uyeda, *Planta* **166** (1985) 161 – 168.
64. A. A. Saafan, A. M. Habib, *Melliand Textilber.* **68** (1987) 687 – 680.
65. R. G. Steadman: *Cotton Testing*, Textile Inst., Manchester 1996, pp. 73.
66. W. R. Goynes, B. F. Ingber, B. A. Triplett, “Cotton Fiber Secondary Wall Development — Time versus Thickness,” *Tex. Res. J.* **65** (1995) 489 – 494.
67. E. K. Boylston, “The Primary Wall of Cotton Fibers,” *Tex. Res. J.* **65** (1995) 429 – 431.
68. R. G. Steadman: *Cotton Testing*, vol. **27 (1)** Textile Inst., Manchester 1997, pp. 63.
69. United States Department of Agriculture, Agricultural Marketing Service (ed.): *The Classification of Cotton,* Agricultural Handbook no. 566, Memphis TN, 1993, pp. 24.

70. National Cotton Council: *Cotton Counts its Customers*, N.C.C., Memphis 1983. International Institute for Cotton: “Cotton,” *Textiles* **11** (1982) 58 – 64.
71. T. L. Van Winkle, J. Edeleanu, E. A. Prosser, C. A. Walker: “Cotton versus Polyester,” *Am. Sci.* **66** (1978) 280.
72. J. B. Cocke, R. A. Wesley, I. W. Kirk, *USDA Market Res. Rep.* **1065** (1977) 16 pp.
73. J. E. Ford: “Jute and Other Vegetable Fibers,” *Textiles* **4** (1975) 58 – 62.
74. R. H. Kirby: *Vegetable Fibers*, Leonard Hill, London 1963.
75. R. Himmefarb: *Technology of Cordage Fibers and Rope*, Interscience, New York 1957.
76. M. L. Joseph: *Introductory Textile Science*, 4th ed., Holt, Rinehart & Winston, New York 1980.
77. *Textile Organon* **55** (1984) 62.
78. H. L. Roder, *Text. Ind. (Mönchen-Gladbach, Ger.)* **74** (1972) no. 22, 383.
79. Food and Agriculture Organization: *1982 Handbook of Production*, United Nations, Rome 1983.
80. U.S. Department of Labor, Occupational Safety and Health Administration, “Occupational Exposure to Cotton Dust,” *Fed. Reg.* **43** (1978) no. 122, 27 349 –27 418.
81. DIN 60 001, Textile Faserstoffe, August 1970.
82. Meeting report “Cellulosefaserforschung,” Dornbirn 1984, *Lenzinger Ber.* 1985, no. 59.
83. W. Albrecht, M. Reintjes, B. Wulfhorst, Fiber Table Appendix 1997, *Chemical Fibers International* **47** (1997) 282.
84. A. Urbanowski, *CTJ* **46** (1996) 260.
85. W. Berger, *CTJ* **44** (1994) 747.
86. K. Götze: *Chemiefasern nach dem Viskoseverfahren*, 3rd ed., Springer Verlag, Berlin 1967.
87. H. F. Mark et al.: *Man-Made Fibres,* vol. 2, Interscience, New York 1968.
88. W. Albrecht, *Melliand Textilber. Int.* **51** (1970) 1487.
89. R. W. Moncrieff: *Man-Made Fibres*, Heywood Books, London 1970, p. 152.
90. C. V. Nikonovich *et al.*, *J. Polym. Sci. Polym. Symp.* 1973, no. 42, 1625.
91. American Enka, US 4 145 523, 1979, US 4 196 282, 1980, US 4 246 221, 1981. Akzo Fibres, DE 30 34 685 C2, 1984.
92. R. Krüger, *CTJ* **44** (1994) 24 – 27.
93. H. Frigo, M. Eibl, D. Eichinger, Lenzinger Berichte (company brochure), vol. 75, 1996, pp. 47.
94. I. Marini, F. Brauneis, *Textilveredlung* **31** (1996) 182 – 187.
95. F. Brauneis, Lenzinger Berichte (company brochure), vol. 75, 1996, p. 105 – 111.
96. H. Nemec, Lenzinger Berichte (company brochure), vol. 9, 1994, pp. 69.
97. J. Chedin, A. Marsaudon, *Makromol. Chem.* **15** (1955) 115.
98. A. Dietl *et al.*, *Papier (Darmstadt)* **20** (1966) 609.
99. J. O. Warwicker *et al.*: *A Review of the Literature on the Effect of Caustic Soda and Other Swelling Agents on the Fine Structure of Cotton*, Shirley Inst., Manchester 1966, Pamphlet no. 93, p. 75.
100. E. Treiber, *Faserforsch. Textiltech.* **22** (1971) 62.
101. E. Treiber in J. Schurz, *Papier (Darmstadt)* **20** (1966) 66.
102. P. Barthel, B. Philipp, *Faserforsch. Textiltech.* **18** (1967) 266.
103. J. J. Willard, E. Pacsu, *J. Am. Chem. Soc.* **82** (1960) 4350. Links
104. C. Y. Chen, R. E. Montana, C. S. Grove, *Tappi* **34** (1951) 420.
105. B. Philipp, Ke Tsing Liu, *Faserforsch. Textiltech.* **10** (1959) 555.
106. B. Philipp, *Faserforsch. Textiltech.* **8** (1957) 45.
107. A. Lyselius, O. Samuelson, *Sven. Papperstidn.* **64** (1961) 145.
108. M. E. Schwab, R. Kloss, lecture 23. Int. Chem. Fasertagung, Dornbirn 1984.
109. B. Philipp *et al.*, *Faserforsch. Textiltech.* **21** (1970) 279.
110. A. T. Sorkov *et al.*, *Khim. Volokna* 1971, no. 4, 32; Sowj. Beitrag *Faserforsch. Textiltech.* **8** (1971) 581.
111. H. Klare, A. Gröbe, *Faserforsch. Textiltech.* **8** (1957) 310; **9** (1958) 262; **10** (1959) 155.
112. A. Gröbe, H. J. Gensrich, *Faserforsch. Textiltech.* **21** (1970) 470.
113. P. H. Hermans: *Physics and Chemistry of Cellulose Fibres*, Elsevier, Amsterdam 1948, pp. 172, 365.
114. J. Schurz in [5], pp. 544 – 545.
115. A. Künschner, *Chemiefasern* **15** (1965) 662 – 668, 783 – 788.
116. W. Bandel in [5], p. 671.
117. A. Gröbe, B. Philipp, H. Klare, *Chemiefasern* **15** (1965) 502.
118. D. Vermaas, *Text. Res. J.* **32** (1962) 353. Links
119. P. H. Hermans: *Physics and Chemistry of Cellulose*, Elsevier, Amsterdam 1949.
120. R. J. E. Cumberbirch, *Rep. Prog. Appl. Chem.* **46** (1961) 233.
121. H. Klare, A. Gröbe, *Österr. Chem. Ztg.* **65** (1964) 218.
122. I. C. Witkamp, W. R. Saxton, *Tappi* **45** (1962) 650.
123. H. Hampe, B. Philipp, *Cellul. Chem. Technol.* **6** (1972) 447.
124. L. Szegö, *Faserforsch. Textiltech.* **21** (1970) 422.
125. V. Jan ča řik, L. Kuniak, *Faserforsch. Textiltech.* **20** (1969) 491.
126. H. Krässig, *Papier (Darmstadt)* **24** (1970) 926.

127. V. Stannett, H. Hopfenberg, *High Polym.* **5** (1971) 907.
128. S. A. Rogowin, *Lenzinger Ber.* 1970, no. 30, 16.
129. Z. A. Rogowin, L. S. Galbraich: *Die chemische Behandlung und Modifizierung der Zellulose*, Thieme Verlag, Stuttgart 1983.
130. M. LeGall, *Bull. Sci. Inst. Text. Fr.* **2** (1973) no. 6, 77.
131. L. Wiesner, *Melliand Textilber.* **49** (1968) 99.
132. A. S. Hoffmann, *Isot. Radiat. Technol.* **8** (1970) 84.
133. A. Heger, *Dtsch. Textiltech.* **17** (1967) 307; *Textiltechnik (Leipzig)* **23** (1973) 665.
134. W. Bobeth *et al.*: *Faserforsch. Textiltech.* **24** (1973) 412; *Lenzinger Ber.* 1975, no. 38.
135. Surface Activation Corp. (SAC), US 3 600 122, 1966.
136. H. Dolmetsch, *Melliand Textilber.* **45** (1964) 12.
137. H. Dolmetsch, *Melliand Textilber.* **51** (1970) 182.
138. H. Mark, *Chemiefasern* **15** (1965) 422.
139. D. Paul, D. Bartsch, *Faserforsch. Textiltech.* **23** (1972) 187.
140. P. Weber, *Melliand Textilber.* **50** (1969) 372.
141. W. Albrecht, *Chemiefasern* **63** (1982) 790.
142. W. Bandel: “Entwicklungen auf dem Cuprophan-, Membran- und Hohlfasergebiet”, *Mitt. Klin. Nephrologie* **4** (1975) 29.
143. G. Seyfart, W. Henne, G. Marx, P. Schroeder, H. J. Gurland: “Die Kapillarniere”, *Biomedizinische Technik* **19** (1974) 174.
144. W. Henne, G. Dünweg, W. Bandel: “A New Cellulose Membrane Generation for Hemodialysis and Hemofiltration,” *Proc. 2nd Meeting Int. Soc. Artif. Org.* (ISAO) 3 (Suppl.) (1979) 466.
145. N. A. Hoenich, T. Frost, D. N. S. Kerr in W. Drukker, F. M. Parsons, J. F. Maher (eds.): *“Replacement of Renal Function by Dialysis*,” Martinus Nijhoff Publ., Den Haag-Boston-London 1979, p. 80.
146. T. Malkomes, A. Reichle in Ullmann, 3rd ed., vol. 11, pp. 260 – 278 (published 1960, numerous references).
147. R. Jährling: *Die Herstellung der Zellwolle und Kunstseide*, Fachbuchverlag, Leipzig 1957.
148. Z. A. Rogowin in: *Chemiefasern*, Thieme Verlag, Stuttgart 1982, pp. 172 – 181.
149. A. Reichle, *Reyon Zellwolle Andere Chem. Fasern* **32** (1954) 133 – 139.
150. H. A. Schlichter: “Untersuchungen über Gleichgewichtsverhältnisse im System Cuoxam-NaOH-Cellulose,” *Thesis*, TH Aachen 1956.
151. Bayer, DE 961 287, 1957.
152. Asahi Kasei Kogyo KK, DE 2 328 583, 1973.
153. Bayer, DE 1 099 690, 1961.
154. Asahi Kasei Kogyo KK, US 3 049 755, 1960; US 3 131 429, 1961; DE 1 660 144, 1965; DE 2 059 177, 1970.
155. American Bemberg Corp., GB 633 108, 1946.
156. Asahi Kasei Kogyo KK, DE 242 964, 1972.
157. Bayer, DE 886 770, 1953.
158. Bayer, FR 1 116 671, 1954.
159. Fiat Final Report 50.
160. CIRFS statistics 1996.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Cellulose Esters

Klaus Balser, Wolff Walsrode AG, Walsrode, Federal Republic of Germany (Chap. 1)

Lutz Hoppe, Wolff Walsrode AG, Walsrode, Federal Republic of Germany (Chap. 1)

Theo Eicher, Stuttgart, Federal Republic of Germany (Chaps. 2.1 and 2.2)

Martin Wandel, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 2.2 and 2.4)

Hans-Joachim Astheimer, Rhodia AG, Freiburg, Federal Republic of Germany (Chap. 2.3)

Hans Steinmeier, Rhodia Acetow AG, Freiburg, Federal Republic of Germany (Chap. 2.3)

1. Inorganic Cellulose Esters 2
1.1. Esterification 3
1.2. Cellulose Nitrate 4
1.2.1. Physical Properties 4
1.2.2. Chemical Properties 5
1.2.3. Raw Materials 7
1.2.4. Production 8
1.2.4.1. Cellulose Preparation 8
1.2.4.2. Nitration 10
1.2.4.3. Stabilization and Viscosity Adjustment 11
1.2.4.4. Displacement and Gelatinization . . 12
1.2.4.5. Acid Disposal and Environmental Problems 12
1.2.4.6. Other Nitrating Systems 13
1.2.5. Commercial Types and Grades . . . 14
1.2.6. Analysis and Quality Control 15
1.2.7. Uses 16
1.2.8. Legal Provisions 17
1.3. Other Inorganic Cellulose Esters 19
1.3.1. Cellulose Sulfates 19
1.3.2. Cellulose Phosphate and Cellulose Phosphite 19
1.3.3. Cellulose Halogenides 20
1.3.4. Cellulose Borates 20
1.3.5. Cellulose Titanate 20
1.3.6. Cellulose Nitrite 20
1.3.7. Cellulose Xanthate 21
2. Organic Esters 21
2.1. Cellulose Acetate 22
2.1.1. Chemistry of Cellulose Esterification 22
2.1.2. Raw Materials 23
2.1.3. Industrial Processes 23
2.1.3.1. Pretreatment 23
2.1.3.2. Esterification 24
2.1.3.3. Hydrolysis 26
2.1.3.4. Precipitation and Processing 26
2.1.4. Recovery of Reactants 27
2.1.5. Properties of Cellulose Acetate . . . 27
2.1.6. Analysis and Quality Control 28
2.2. Cellulose Mixed Esters 29
2.2.1. Production 29
2.2.2. Composition 29
2.2.3. Properties 29
2.2.4. Other Organic Cellulose Mixed Esters 30
2.2.5. Uses 30
2.3. Cellulose Acetate Fibers 30
2.3.1. Properties 30
2.3.2. Raw Materials 31
2.3.3. Production 32
2.3.4. Uses 32
2.3.5. Economic Aspects 32
2.4. Plastic Molding Compounds from Cellulose Esters 33
2.4.1. Physical Properties 33
2.4.2. Polymer-Modified Cellulose Mixed Esters 37
2.4.3. Chemical Properties 38
2.4.4. Raw Materials 39
2.4.5. Production 40
2.4.6. Trade Names 40
2.4.7. Quality Requirements and Quality Testing 40
2.4.8. Storage and Transportation 40
2.4.9. Uses 41
2.4.10. Toxicology and Occupational Health 42
3. References 42

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_419.pub2

1. Inorganic Cellulose Esters

Definition. Cellulose esters are cellulose derivatives which result by the esterification of the free hydroxyl groups of the cellulose with one or more acids, whereby cellulose reacts as a trivalent polymeric alcohol. Esterification can be carried out by using mineral acids as well as organic acids or their anhydrides with the aid of dehydrating substances. Cellulose nitrate [*9004-70-0*] is the most important and only industrially produced inorganic cellulose ester (abbreviation CN, according to DIN 7728, T 1, 1978). A comprehensive bibliography on inorganic cellulose esters may be found in [12–19].

Historical Aspects [20]. The nitric acid ester of cellulose is the oldest known cellulose derivative and is still the most important inorganic cellulose ester. The term "nitrocellulose" is still used, but it is not the precise scientific term for cellulose nitrate. Cellulose esters were first described and industrially used at a time when the structure of esters was unknown and information on the polymeric primary material cellulose was not yet available.

The nitration of polysaccharides with concentrated nitric acid had already been described in 1832. H. Braconnot obtained a white and easily inflammable powder when he transformed starch with nitric acid. The product obtained was xyloidine. Th.-J. Pelouze treated paper with nitric acid and obtained an insoluble product containing ca. 6 % nitrogen which he called pyroxiline and which he provided for military use.

C. F. Schönbein and R. Böttger are considered to be the inventors of so-called gun cotton (1845). They transformed cotton with a mixture of nitric and sulfuric acid into a highly nitrated product that could serve as a substitute for black powder. Production on an industrial scale was stopped in 1847 because its extremely rapid catalytic decomposition was the cause of numerous plant explosions. Production was legally prohibited in 1865.

The use of cellulose nitrate as an explosive brought new momentum to its further industrial and scientific development, as well as to its economic significance. F. Abel made a basic breakthrough in 1865 when he succeeded in developing a safe method of handling. He was able to achieve a better washing of the adhering nitrating acid and a hydrolytic decomposition of the unstable sulfuric acid ester by grinding the nitrated fibers in water. This process allowed this product to attain military importance for its use as gunpowder.

In 1875, A. Nobel phlegmatized nitroglycerine by mixing with cellulose nitrate and discovered blasting gelatin. In the 1880s, smokeless gunpowders were developed. Vieille developed Poudre B (blanche) and Nobel developed Ballistit, the first dibasic gunpowder from cellulose nitrate and nitrogycerol. Abel and Dewar developed a similar gunpowder called Cordit.

The discovery that fibrous products could be modified by, for example, dissolution in an alcohol/ether mixture (film for wound protection) or by gelatinization with softeners brought additional uses. Films made from camphor and castor oil were used in collodium photography as carriers for light-sensitive materials (Archer, 1851). Nitrofilms found increasing use in photography and cinematography until they were replaced by nonflammable cellulose acetate films.

The year 1869 is considered to be the beginning of the age of plastics. J. W. Hyatt discovered celluloid, the first thermoplastic synthetic material. It was originally used as a substitute for ivory in the production of billiard balls.

The practical use of cellulose nitrate as a raw material for lacquers began in 1882, when Stevens suggested amyl acetate as a highly volatile solvent (Zaponlack). The nitro lacquers achieved importance at first after World War I (Flaherty, 1921), when new applications were being sought as a result of the sharply declining demand for gunpowder. Only after the possibility of depolymerization of cellulose nitrate by pressure boiling during production had become known during the 1930s it was possible to use cellulose nitrate in protective and pigmented lacquers. Thus, it became possible to use nitro lacquers for painting automobiles on the assembly line.

Cellulose xanthate [*9032-37-5*] is a cellulose ester obtained with the inorganic acid dithiocarbonic acid. Ch. F. Cross discovered this important alkali-soluble cellulose ester in 1891 while he was reacting cellulose with alkali and carbon disulfide. It represents the base of the viscose process introduced in 1894 by Bevan and Bea-

DLE for producing man-made cellulosic fibers (rayon, rayon staple) and cellophane.

Other cellulose esters with inorganic acids are presently only of theoretical interest and have not attained any industrial or economic importance.

Present Significance. Cellulose nitrate is still important, 150 years after its discovery. It is industrially produced in large quantities for diversified applications. The reasons for this are the relatively simple production process with high yields, its solubility in organic solvent systems and its excellent film-forming properties from such solutions (collodion cotton as a raw material for lacquers), compatibility and gelatinability with softeners and other polymers (thermoplastics), as well as inflammability (gun cotton for explosives). Cellulose nitrate has maintained its importance as a raw material for the manufacture of protective and coating lacquers as well as blasting agents and explosives.

Densified products, colored or pigmented chips kneaded with softeners, as well as aqueous dispersion systems with a low solvent content, are available today. They facilitate transport and processing, secure existing application forms, open up new ones, and are becoming increasingly nonpolluting.

The viscose process (see →Cellulose, Chap. 3.1.1.) with its essential intermediate cellulose xanthate will remain, because of the availability of a constantly regrowing raw material supply, an important source of textile fibers for years to come. Alternative processes are intensively being sought to reduce pollution by sulfurous decomposition gases resulting during manufacturing. Cellulose nitrite, the cellulose ester of nitrous acid, is the most prominent example.

Cellulose esters with other inorganic acids have been frequently described and investigated. Cellulose sulfate [*9032-43-3*] was of some interest because of its solubility in water, but never achieved any practical importance. Cellulose phosphate [*9015-14-9*], borate, and titanate show interesting properties such as fire retardation, but are not yet of any industrial significance.

1.1. Esterification

Mechanism. The alcoholic hydroxyl groups of cellulose are polar and can be substituted by nucleophilic groups in strongly acid solutions. The mechanism of esterification assumes the formation of a cellulose oxonium ion followed by the nucleophilic substitution of an acid residue and the splitting off of water. Esterification is in equilibrium with the reverse reaction; saponification can be inhibited largely by binding the resulting water.

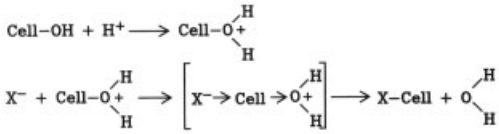

Course of Reaction. The three functional hydroxyl groups on each anhydroglucose unit of cellulose are blocked by intermolecular and intramolecular hydrogen bonds and, therefore, are not freely accessible for the reaction partners. The supermolecular arrangement and microstructure within the cellulose fiber, whose intensity depends on the origin and previous history of the cellulose material, is determined by these hydrogen bonds. The accessibility to the reaction partners and the reactivity of the alcohol groups also depend on this structure.

Due to the fact that cellulose is insoluble in all common solvents, reactions to form derivatives are usually carried out in heterogeneous systems. As the reaction proceeds, new reactive centers are created so that ultimately almost all parts of the cellulose fibers are included and in special cases yield soluble derivatives which react to completion in a homogeneous phase.

Little information is available on the esterification process. The following two reaction types are under discussion:

– An intermicellar reaction, which initially consists of the penetration of the reaction partner into the so-called amorphous regions between the highly organized cellulose micelles and proceeds during the course of esterification from the surface to the innermost regions of the micelles. The reaction speed is determined by diffusion.

– An intramicellar or permutoid reaction, in which the reagent penetrates all regions including the micelles so that practically all cellulose molecules react almost simultaneously. The reaction speed is specified by adjustment of the esterification equilibrium.

Arguments for both mechanisms are based on X-ray analyses. The possibility exists that both reaction types occur and ultimately merge. This depends on the reaction conditions, especially the esterification mixture and the temperature.

The hydrogen bonds between the cellulose molecules are almost completely broken down during esterification. The introduction of ester groups separates the cellulose chains so completely that the fiber structure is either altered or completely destroyed. Whether the cellulose ester is soluble in a solvent or in water depends on the types of substituents added.

Substitution. The esterification reactions do not necessarily proceed stoichiometrically because of equilibrium adjustment. The maximal attainable substitution with a mean degree of substitution (DS) of 3 is generally not reached. A triester can only be obtained under carefully controlled conditions. The primary hydroxyl group on the C-6 atom reacts most readily, while the neighboring hydroxyl groups on the C-2 and C-3 atoms of the anhydroglucose ring react considerably slower due to steric hindrance.

Basically, esterification is possible with all inorganic acids. Limiting factors are the type and the size of the acid residue as well as the varying degree of acid-catalyzed hydrolysis, which can lead to a complete cleavage of the cellulose molecule as the result of statistical chain splitting.

1.2. Cellulose Nitrate

Summary monographs on cellulose nitrate in addition to those in the Reference list can be found in [21] and [22].

1.2.1. Physical Properties

Cellulose nitrate (CN) is a white, odorless, and tasteless substance. Its characteristics are dependent on the degree of substitution.

Density. The density of cellulose nitrate is dependent on its nitrogen content and, therefore, on the degree of substitution (Table 1).

Table 1. The density of cellulose nitrate in relation to its nitrogen content (degree of substitution)

Nitrogen content, %	Degree of substitution (DS)	Density at 20 °C, g/cm^3
11.5	2.1	1.54
12.6	2.45	1.65
13.3	2.7	1.71

The bulk density of commercially available CN types is between 0.25 and 0.60 kg/L for moistened CN cotton, 0.15 – 0.40 kg/L when converted to dry mass.

Cellulose nitrate chips, which contain at least 18 % dibutyl phthalate in addition to cellulose nitrate, have a density of 1.45 g/cm^3 (measured at 20 °C in an air-comparison pycnometer). The bulk density is 0.3 – 0.65 kg/L.

Specific Surface. The laboratory apparatus described by S. Rossin [23] is best suited for the determination of the specific surface of cellulose nitrate, which is 1 850 – 4 700 cm^2/g, depending on the fineness of the cellulose nitrate.

The determination of the inner surface according to the BET method showed dependence on the molar mass (i.e., an inner surface area of 1.44 m^2/g would correspond to a molar mass of 180 000 g and a surface area of 2.41 m^2/g would correspond to a molar mass of 400 000 g).

It must, however, be noted that the degassing temperature was lowered from the usual 200 °C to 60 °C due to the fact that cellulose nitrate deflagrates at 180 °C. It is possible that complete desorption did not take place under these conditions.

Thermodynamic Properties , see [24, pp. 137 –154]. The most important thermodynamic properties are listed in Table 2.

Table 2. Thermodynamic properties of some cellulose nitrates

Heat of formation	trinitrate	– 2.19 kJ/g
	dinitrate	– 2.99 kJ/g
	cellulose	– 5.95 kJ/g
Heat of combustion	trinitrate	– 9.13 kJ/g
	dinitrate	– 10.91 kJ/g
	cellulose	– 17.43 kJ/g

Table 2. continued

Specific heat	celluloid film (70 % CN and 30 % camphor)	1.26 – 1.76 J g^{-1} K^{-1}
Thermal conductivity	celluloid film (70 % CN and 30 % camphor)	0.84 kJ m^{-1} h^{-1} K^{-1}
Heat of solution in acetone	CN with 11.5 % N content	– 73.25 J/g
	CN with 14.0 % N content	– 81.64 J/g

Electrical Properties [24, p. 136]. The following electrical properties were measured on cellulose nitrate containing 30 wt % camphor (celluloid):

Dielectric constant
at 50 – 60 Hz7.0 – 7.5
10^6 Hz6.0 – 6.5
Dissipation factor (tan δ)
at 50 – 60 Hz0.09 – 0.12
10^6 Hz0.06 – 0.09
Specific resistance $10^{11} - 10^{12}$ $\Omega \cdot$ cm

Mechanical Properties [25]. The stress – strain diagram of cellulose nitrate films shows the elongation and tensile strength to be dependent on the size of the molecule (expressed as a term of viscosity).

The higher the molecular mass of a CN, the more elastic is the film made from it. Films become more brittle and their tensile strength declines with decreasing molecular mass (see Table 3).

Table 3. Mechanical properties of CN lacquer films

Type∗	Elongation, %	Tensile strength, N/mm^2
E 4	24 – 30	98 – 103
E 6	23 – 28	98 – 103
E 9	23 – 28	88 – 98
E 13	20 – 25	88 – 98
E 15	18 – 23	78 – 98
E 21	12 – 18	78 – 88
E 22	10 – 15	74 – 84
E 24	8 – 12	69 – 78
E 27	5 – 10	59 – 69
E 32	<5	39 – 49
E 34	<3	29 – 49

∗ According to DIN 53179: The E-type designation specifies the CN concentration (% in dry condition) in acetone which gives a viscosity of 400 ± 25 mPa · s.

Optical Properties. Cellulose nitrate films are optically anisotropic because of their microcrystalline structure. The colors change in polarized light in relation to the nitrogen content of the CN:

11.4 % N weakly red
11.5 – 11.8 % N yellow
12.0 – 12.6 % N blue to green

The index of refraction is 1.51, and the maximal light transmission is achieved at 313 nm.

Light Stability. Exposure to sunlight, and especially to ultraviolet light, has a detrimental effect on cellulose nitrate film by causing it to become yellowish and brittle. Solvents, softeners, and resins can either promote or hinder yellowing.

1.2.2. Chemical Properties

The three hydroxyl groups of cellulose can be completely or partially esterified by nitrating acid. The varying degrees of nitration can be related to the following theoretical nitrogen contents:

Cellulose mononitrate, $C_6H_7O_2(OH)_2(ONO_2)$:
6.75 % N
Cellulose dinitrate, $C_6H_7O_2(OH)(ONO_2)_2$:
11.11 % N
Cellulose trinitrate, $C_6H_7O_2(ONO_2)_3$:
14.14 % N

Cellulose nitrate with a nitrogen content between 10.8 and 12.6 % is a suitable raw material for lacquers, and CN with > 12.3 % N is suitable for explosives exclusively.

Degree of Substitution – Nitrogen Content – Solubility. The degree of substitution can be calculated from the nitrogen content of the various CN types (Fig. 1). The degree of substitution determines the solubility of cellulose nitrate in organic solvents. CN for lacquers can be classified according to its solubility in organic solvents as follows:

alcohol-soluble CN (A types)
nitrogen content: approx. 10.9 – 11.3 %
readily soluble in alcohols, esters, and ketones
moderately soluble CN (AM types)
nitrogen content: approx. 11.4 – 11.7 %
soluble in esters, ketones, and glycol ethers with excellent blendability or compatibility with alcohol
CN soluble in esters (E types)

nitrogen content: 11.8 – 12.2 % for lacquer cotton, up to 13.7 % for guncotton
readily soluble in esters, ketones, and glycol ethers

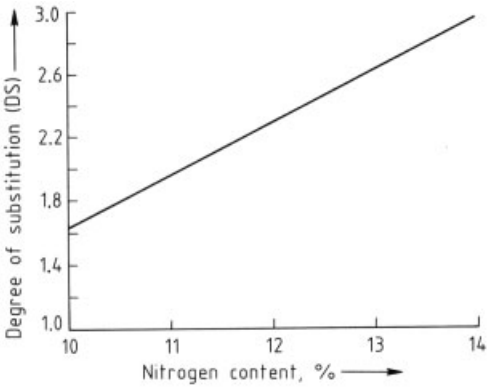

Figure 1. Variation of the degree of substitution with the nitrogen content of cellulose nitrate

Intrinsic Viscosity – Degree of Polymerization [24, pp. 85 – 121]. The mean number of anhydroglucose units in cellulose nitrate molecules is designated as the mean degree of polymerization (DP). The viscosity of the solution (at the same concentration in the same solvent) is generally considered to be a relative measure of the molecular mass. The molecular mass can be mathematically expressed as a function of the intrinsic viscosity (Staudinger – Mark – Houwink equation). For further information → Plastics, Analysis.

Distribution of the Molecular Mass. The starting material of cellulose nitrate is natural cellulose, the quality of which is subjected to annual growth cycles. It is, therefore, of great importance to have polymolecular data, such as the mean degree of polymerization and the distribution of the molecular mass, available in addition to viscosity, solubility behavior, and nitrogen values. These values are important, for example, in assessing the mechanical properties and aging processes of polymer products.

The isolation of the polymers according to their molecular mass can be achieved elegantly by gel permeation chromatography (GPC).

Chemical Compatibility. An everyday use of cellulose nitrate is in nitro lacquers, where it is dissolved in organic solvents. In this solution, cellulose nitrate is extremely compatible with essential substances in the lacquer formulation such as alkyd resins, maleic resins, ketone resins, urea resins, and polyacrylates. A large number of softeners, such as adipates, phthalates, phosphates, and raw and saturated vegetable oils are compatible with cellulose nitrate.

Chemical and Thermal Stability. Cellulose nitrate, as a solid or in solution, should not be brought into contact with strong acids (degradation), bases (denitration), or organic amines (decomposition) since they all induce a destruction of cellulose nitrate. This may proceed very rapidly and lead to deflagration of the cellulose nitrate.

The ester bonds of cellulose nitrate, which can be broken by saponifying agents or by catalysis, are responsible for its physicochemical instability. This substance-specific property is dependent on the temperature, the specimen, and whether catalytically active decomposition products remain or are removed from the sample.

Another basic instability of cellulose nitrate is observed during the production process. Mixed sulfuric acid esters transmit a chemical instability to the nitrocellulose molecule. These mixed esters are destroyed in weakly acid water during the stabilization phase of production. The long reaction time required by this procedure can be considerably shortened by increasing the reaction temperature. The time required can be reduced to only a few hours by raising the temperature to 60 – 110 °C. Under these conditions the nitrate ester remains stable; the glucosidic bond of cellulose nitrate, on the other hand, is attacked. This property is used to advantage to specifically reduce the degree of polymerization of the cellulose nitrate.

Thermogravimetry, IR spectroscopy, and electron spectroscopy (ESCA) [26], [27] have been used to determine the extent of thermally induced and light-induced decomposition of cellulose nitrate. The reaction proceeds as follows:

$$\text{Cell}-\text{O}-\text{NO}_2 \longrightarrow \text{Cell}-\text{O}\cdot + \cdot\text{NO}_2$$

It is proceeded by a series of extremely exothermic oxidation reactions triggered by the NO_2 radical, which often leads to spontaneous deflagration. NO_2 is reduced to NO and in the presence of air NO_2 is reformed, thus initiating an autocatalytic chain reaction, at the end of which the gaseous reaction products CO_x, NO_x, N_2, H_2O, and HCHO are found.

By adding stabilizers such as weak organic bases (diphenylamine) or acids (phosphoric acid, citric acid, or tartaric acid), intermediary nitric oxides can be bound and the autocatalytic decomposition prevented.

Thermal decomposition does not occur at temperatures below 100 °C. The temperature (according to [28]) at which cellulose nitrate spontaneously deflagrates is used as a measure of its thermal stability. A well-stabilized lacquer cotton has a deflagration temperature of ≥ 180 °C. The deflagration temperature of plasticized cellulose nitrate chips with at least 18 wt % softener (i.e., dibutyl or dioctyl phthalate) is ≥ 170 °C.

The Bergmann – Junk test [29] and the warm storage test are additional methods for determining the stability of cellulose nitrate.

1.2.3. Raw Materials

Cellulose. Until the beginning of World War I, the only raw material available for nitration was cellulose obtained from cotton in the form of bleached linters (as flakes or crape). This was due to the high degree of purity (α-cellulose > 98 %), which allowed a high yield and products with good clarity and little yellowing.

Especially in times when linters were scarce, it was possible to produce gunpowder from wood celluloses, even unbleached, other cellulose fibers (annual plants), and even from wood if attention was given to the adequate disintegration of the raw materials. Lacquer types obtained from wood celluloses, especially from hardwood, gave dull and mat films and lacquers with inferior mechanical properties. This is due to the high content of pentosans, which is also nitrated but is easily split by hydrolysis in conventional nitrating acid systems and thus becomes insoluble.

The development of highly purified chemical-grade wood pulp by refinement with hot and cold alkali having R_{18} values of 92 – 95 % (see Table 16) allows this type of raw material to be used in the same manner as were linters, which currently are used only for the production of special and highly viscous CN types. The highly refined prehydrolyzed sulfate pulps with R_{18} values of above 96 % are especially well-suited for nitration. The viscosity range of CN products can be adjusted in advance by choosing an initial cellulose with an adequate DP. A low ash content, and above all a low calcium content, of the cellulose is important in preventing calcium sulfate precipitation during industrial nitration.

A comparative study on the nitrating behavior of linters and wood pulps [30] shows the morphological factors of the fibers (fiber length and distribution, cross-section form and thickness of the secondary wall, and fine structure including packing density, degree of crystallization, and lateral arrangement of the fibrils), the chemical composition of the cellulose (DP and polydispersibility) as well as the type, quantity, and topographic distribution of the accompanying hemicelluloses and lignin to be responsible for the nitrating capability of celluloses. These factors determine the swelling properties and thereby the uniformity of nitration, as well as the compressibility and the relaxation capacity of the fibers, which in turn influence the retention capacity of the fiber mass. Linters with a lower acid retention capacity of 110 – 130 % are definitely superior to wood pulp (acid retention capacity of prehydrolyzed sulfate pulps up to 230 %, of sulfite pulps up to 300 %) in this respect. The suitability of a raw material for nitration can be tested by a specially developed machine that measures the compression and relaxation characteristics of a cellulose fiber pile.

Approximately 150 000 t (3.4 %) of the annual worldwide production of 4.4×10^6 t of chemical-grade pulps are used for the production of cellulose nitrate.

Industrial Nitrating Agents. The so-called nitrating acid as developed by SCHÖNBEIN, the nitric acid/sulfuric acid/water system, is still the nitrating agent of choice for industrial purposes. The highest attainable degree of substitution using this system is at DS $2.7 \triangleq 13.4$ % N. This is achieved only when the nitric acid used is not hydrated and the molar ratio of nitric acid to sulfuric acid monohydrate is 1 : 2. The optimal nitrating mixture is as follows:

	HNO_3	H_2SO_4	H_2O
molar ratio	1	2	2
wt %	21.36	66.44	12.20

Water plays a special role as far as the attainable degree of nitration is concerned. Below 12 % water there is no increase in substitution, but a higher water content results in a drastic decline in the degree of nitration (Fig. 2).

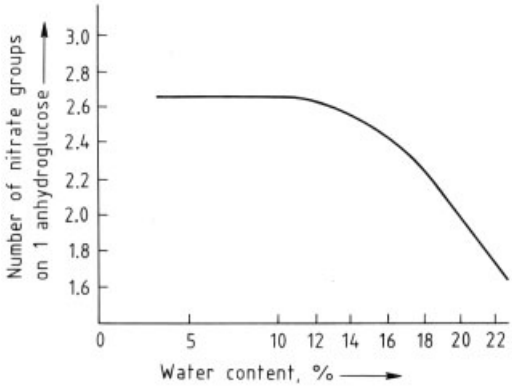

Figure 2. Dependence of the degree of esterification (DS) on the water content of the optimal nitrating mixture ($HNO_3 : H_2SO_4 = 1 : 2$)

It is assumed that increasingly hydrated nitric acid causes increased swelling and gelatinization of the cellulose so that the nitrating acid is no longer able to penetrate into the inner structures of the micelles.

The desired degrees of esterification can be adjusted by varying the nitrating acid mixture according to the CN types (Table 4), whereby in industrial processes the nitric acid content is kept nearly constant at 25 – 26 %.

The ternary system $HNO_3/H_2SO_4/H_2O$ has been extensively investigated. The results are summarized in Figure 3.

The curves of the same degrees of substitution (% N) in relation to the nitrating acid composition are presented here. The cross-hatched band identifies those areas in which the cellulose material is extremely swollen and gelatinized. Three zones can therefore be differentiated in the phase diagram:

1) Area of technical nitration:
   nitric acid 15 – 100 %
   sulfuric acid 0 – 80 %
   water 0 – 20 %
   In this range, nitric acid is present in a nonhydrated form and induces true nitration (N content 10 %). The industrially used range with 20 – 30 % HNO_3, 55 – 65 % H_2SO_4, and 8 – 20 % water is also included in this area.
2) Area of solution:
   nitric acid 0 – 10 %
   sulfuric acid 60 – 100 %
   water 0 – 40 %
   Little or no nitration takes place in this range. Cellulose is degraded to the point of complete dissolution in concentrated sulfuric acid.
3) Area of swelling:
   Nitric acid is increasingly hydrated in this range of increasing water content. Nitration decreases rapidly.

A process developed in the United States, but less important, uses magnesium nitrate instead of sulfuric acid as a dehydrating agent [31]. Magnesium nitrate can bind water as its hexahydrate. The nitrating mixture consists of 45 – 94 % nitric acid, 3.3 – 34 % magnesium nitrate, and 2.7 – 21 % water; the ratio of magnesium nitrate : water is 1.2 – 2.2 : 1. A cellulose nitrate with an N content of 11.9 % was obtained, for example, with 64.5 % HNO_3, 19.5 % $Mg(NO_3)_2$, and 16 % H_2O. This nitrating system is appropriate for a continuous process, in which waste and washing acid are reprocessed in ion exchangers and the magnesium nitrate is recycled. Thus, acid and sulfate no longer pose a waste disposal problem.

1.2.4. Production

The flow diagram (Fig. 4) shows the industrial production of CN according to the mixed acid process. The viscosity of the end product is determined by the choice of the initial cellulose, and the degree of nitration is determined by the composition of the mixing acid. The final viscosity adjustment follows during the pressure boiling step (see Section 1.2.4.3).

1.2.4.1. Cellulose Preparation

Cotton linters with a moisture content of up to 7 % are mechanically disintegrated homogeneously. Pressed pulp sheets must be appropriately shredded to obtain rapid and uniform nitration. Spruce or beech celluloses, preactivated with 20 % sodium hydroxide (mercerization),

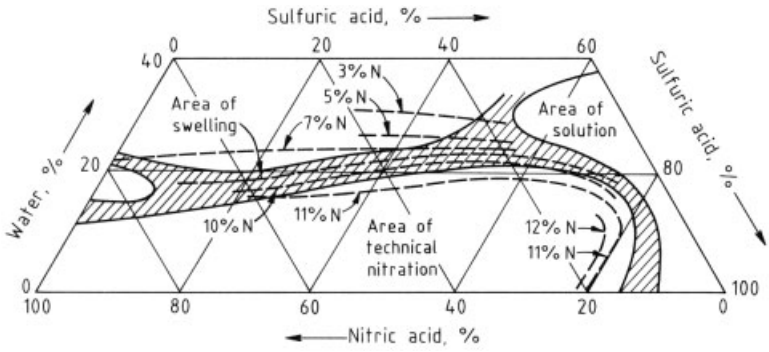

Figure 3. Composition of the nitrating mixtures and attainable N contents of cellulose nitrates

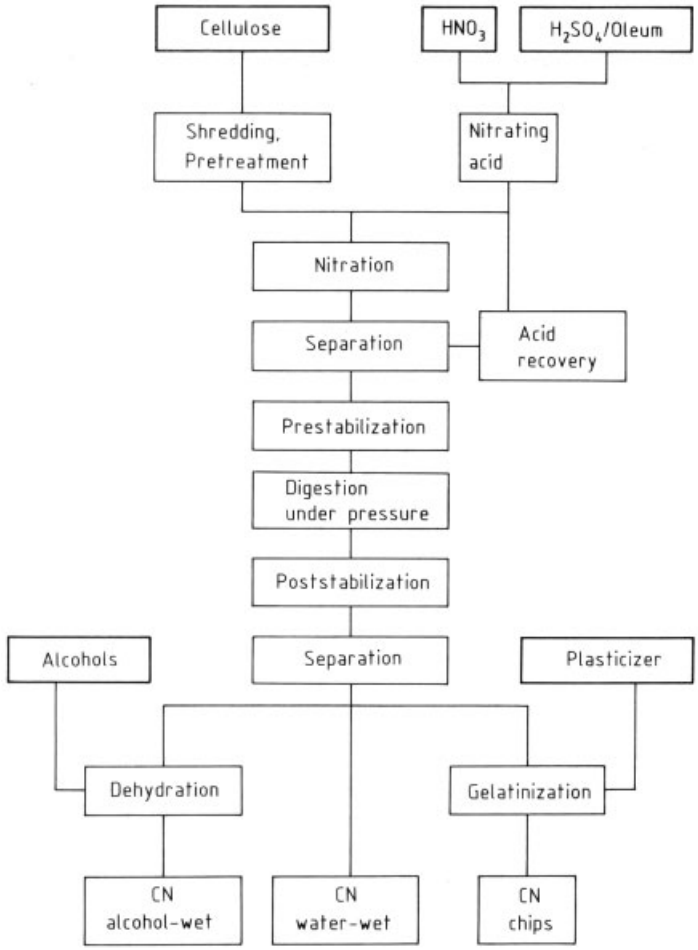

Figure 4. Flow diagram of cellulose nitrate production

Table 4. Industrially used nitrating acid solutions

CN type	Nitrating acid			N content,	DS
	% HNO_3	% H_2SO_4	% H_2O	%	
Lacquer cotton A	25	55.7	19.3	10.75	1.90
Celluloid cotton	25	55.8	19.2	10.90	1.95
Lacquer cotton AM	25	56.6	18.4	11.30	2.05
Dynamite cotton	25	59.0	16.0	12.10	2.30
Lacquer cotton E	25	59.5	15.5	12.30	2.35
Powder cotton	25	59.8	15.2	12.60	2.45
Gun cotton	25	66.5	8.5	13.40	2.70

were formerly used for this purpose in the form of crape papers with a mass per unit area of ca. 25 g/m^3. To avoid the costly transformation of the cellulose to paper sheets, a process was attempted to obtain a loose product resembling linters by direct disintegration of pulps to fibers. A moisture content of 50 % proved to be optimal for nitration and washing out the acid. The required drying of the cellulose flakes before nitration proved to be disadvantageous.

The Stern shredder [32], in which the pulp sheets are torn rather than being cut into small elongated shreds to avoid compression at the edges, was a definite improvement. Currently, cellulose for nitration is used in the form of fluff, shreds, or chips. The packing density and compression behavior of the cellulose fibers in the fiber pile are decisive factors for the swelling and nitration kinetics, as well as the acid retention capability [30].

1.2.4.2. Nitration

Nitration on an industrial scale is still frequently carried out according to a batch process that was developed from a process described by DuPont in 1922. The equipment is constructed of stainless steel. The adjusted and preheated nitrating acid reaches the stirring reactor that is charged with cellulose by means of a measuring system; a large excess of acid (1 : 20 to 1 : 50) is added to retain the ability of the reaction mixture to be stirred and to ensure that heat is carried off. The nitrating temperature is between 10 °C (dynamite type) and 36 °C (celluloid type). The total heat of reaction is estimated to be over 200 kJ per kg of CN, of which the enthalpy of formation of CN is about one-third.

Even though the reaction is nearly complete after ca. 5 min, the mixture remains in the reactor for about 30 min. The temperature must remain constant (cooling), since hydrolytic degradation processes that lead to considerable losses in yield begin at temperatures as low as 40 °C.

The theoretical yield of commonly used industrial types with a DS of 1.8 – 2.7 ($\hat{=}$ 10.4 – 13.4 % N) is between 150 and 176 % with respect to cellulose. The practical yield, however, is up to 15 % lower and depends on the type and purity of the cellulose, as well as on the temperature and duration of nitration. Losses arise from the inevitable complete decomposition of cellulose to oxalic acid by way of oligo- and monosaccharides, whereby the nitric acid is reduced to nitrogen oxides, NO_x. In addition, mechanical losses during the subsequent separation process, due particularly to short fibers (cellulose from hard wood), must also be taken into consideration.

The reaction mixture is drained from the reactor into the centrifuge, where the excess acid is separated and removed at high speed and reprocessed for recycling. The mixture must remain moist so that it does not ignite and deflagrate.

The degree with which the product retains acid after separation is of economic importance because of the acid loss and the expense of the ensuing washings. Linters with an acid retention of 100 – 130 % clearly surpass wood celluloses in this respect which, depending on the wood and cellulose type as well as its processing, can retain up to 3 times more adhering acid relative to CN [30].

The still acid-moist product is immediately placed into a great excess of water (consistency 1 %) so that the adhering acid is displaced as rapidly as possible and the saponification of the CN is prevented.

Continuous nitrating processes, which are more economical, were developed in the 1960s

[33], [34]; they ensure a more uniform product quality and are safer to handle. The nitrating system consists of two or more consecutively arranged straight-run vats or tube systems containing conveyers (screw conveyer or turbulence stirring apparatus) which forward the reaction mixture. The prepared cellulose is directed into this cascading equipment from storage bunkers over automatic weighing scales and continuously mixed with the added nitrating acid. It is important that the cellulose is rapidly added and immediately covered with acid. There it remains for 30 – 55 min. A newer process using a continuous loop-formed pressure reactor [35] requires the cellulose to remain only for 6 – 12 min. The reactant is then sent into a continuously operating special centrifuge, where the excess acid is separated and simultaneously taken up with water. The fact that the reactant remains only a few seconds reduces the risk of spontaneous deflagration and saponification.

Figure 5 shows schematically the continuous process according to Hercules [31].

The broken-up and preconditioned cellulose (a) is brought by way of the automatic scales (b) to the continuous reactor (c). The reaction product is centrifuged in a washing zone (d) and simultaneously washed by zones with water in a countercurrent. The product leaves the centrifuge almost free of acid, and the washed out acid can be recycled and reused almost without loss [36].

1.2.4.3. Stabilization and Viscosity Adjustment

The prestabilization step following the prewashing further purifies the product by means of repeated washing and boiling with water that contains 0.5 – 1 % acid residue. The batch method requires large amounts of space, water, and energy; the required boiling time varies between 6 (celluloid type) and 40 h (guncotton). Automatic continuous processes have been developed in this case as well [37].

Most of the remaining sulfuric acid is removed during prestabilization, since it would promote the catalytic self-decomposition of CN. The sulfuric acid is bound by adsorption and esterified. A total sulfate content of 1 – 3 % was found in weakly nitrated CN, of which 70 – 85 % is in the form of the acidic sulfuric acid semiester, while highly nitrated CN contains only 0.2 –0.5 % total sulfate, of which 15 – 40 % is thought to exist as an ester. Semiesters can be easily saponified and washed out by boiling with water. It is not yet certain whether the so-called resistant sulfate content exists in the form of the neutral sulfuric acid ester or the physically adsorbed sulfuric acid.

The desired final viscosity of the CN is adjusted in the following process step, which is pressure boiling (digestion under pressure) in a consistency of 6 – 8 % at 130 – 150 °C, by means of specific degradation of the degree of polymerization. The remaining extremely low sulfuric acid content induces hydrolytic decomposition at this temperature and under pressure. The viscosity can, for example, be reduced to 1/10 of the initial viscosity within 3 h at 132 °C by using this process. This process made the development of high solid coating and protective nitro lacquers possible. The stabilization process of guncottons is accelerated by pressure boiling; dynamite wools are usually not pressure boiled.

Further product losses are due to chain degradation ranging from soluble cleavage products to oxalic acid. Nitrous gases (NO_x) are released by the reduction of nitric acid, which must be continuously drawn off to avoid decomposition of CN.

Pressure boiling can be achieved batchwise in autoclaves, as well as continuously in a tube reactor of 1 500 m in length and 100 mm in diameter, e.g., with direct steam. A one-pot process in which prestabilization, pressure boiling, and poststabilization are carried out in one operation is described in [38].

During the stabilization process, the remaining sulfuric acid is almost completely removed by additional washing and boiling. While celluloid and lacquer types are finished in flaky, fibrous form, guncotton must be ground. This is done in grinding hollander engines at 12 – 15 % consistency or continuously in a series of cone refiners, whereby the material is gradually concentrated from 3 to 10 % between the various grinding steps. Sorting steps are inserted by hydrocyclones during the final washing processes. The last acid remnants in the fiber capillaries are removed during the grinding process by means of diffusion against water. Weak bases, sodium carbonate, or chalk are used to main-

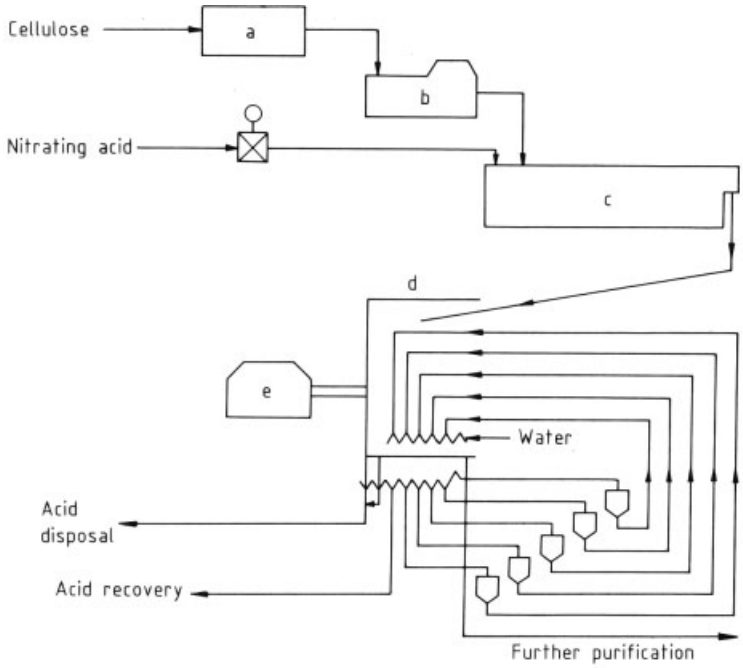

Figure 5. Continuous cellulose nitrate production according to Hercules
a) Preconditioning; b) Auto-matic scale; c) Reactor; d) Washing zone; e) Centrifuge

tain a pH of 7. Stabilizers (organic acids) may be added during this step.

1.2.4.4. Displacement and Gelatinization

A water-wet CN cotton with a water content of 25 – 35 % remains in the centrifuge after the final separation and is then packed into drums or PE sacks.

Water contained in celluloid and lacquer types is displaced by alcohols specified by the processors (ethanol, 2-propanol, *n*-butanol) in displacement presses or displacement centrifuges. Continuous processes prevail here also [39]. The resulting aqueous alcohols must be distilled to remove the water.

The water-wet CN cotton can be gelatinized with softeners such as phthalates in kneading aggregates and dried on drum or band driers for the production of CN chips [40], [41]. Colored chips are obtained by adding carbon black or pigments from which colored enamels can be produced without the use of ball mills or roller mills.

1.2.4.5. Acid Disposal and Environmental Problems

The nitrous gases formed during the nitrating and stabilizing side reactions are drawn off and washed out in trickling towers. The lower nitrogen oxides are regained after oxidation as 50 – 60 % nitric acid.

The waste acids resulting from the first separation contain 2 – 3 % more water and 3 – 4 % less nitric acid than the initial mixture. They are circulated in a closed system and constantly regenerated with nitric acid and oleum. The acid that adheres to the product must also be replaced.

The proportion of adhering acid depends on the initial cellulose and the CN type. It ranges between 80 % (guncotton) and up to 200 – 300 % (lacquer types) with regard to CN and is removed with the water used for washing and boiling.

Aside from the economic aspects of acid loss in wastewater, environmental considerations are beginning to play an increasingly important role. While older manufacturing facilities using simple centrifuging to remove waste acid produced 300 m^3 of water per ton of CN, containing 0.5 %

acid and with a pH of 1, it was possible to reduce the volume of wastewaters to a fraction of its previous volume by almost completely closing the cycles.

Before proceeding into the draining ditch, the wastewater must be separated from the hardly decomposable sludge consisting of cellulose and CN, and then be neutralized. The sulfate proportion can be reduced by calcium sulfate precipitation, while the nitrate proportion remains completely in the wastewater. Organic matter of communal sewage, for example, can be biologically decomposed without additional oxygen, whereby nitrates disappear almost completely as a result of biological denitrification.

The salt/acid process with magnesium nitrate (see Section 1.2.4.2) is more favorable with regard to the wastewater problem. Sulfates are completely absent, and magnesium nitrate is recycled and, therefore, causes no water pollution problems. The amount of wastewater can be reduced by 80 % and the nitric acid requirement by 83 % in comparison to the formerly used discontinuous processes.

1.2.4.6. Other Nitrating Systems

Numerous attempts have been made to improve nitration by the introduction of other nitrating systems, or at least to increase the degree of substitution. Further details may be obtained from [12], [14], [17], and [21]. Table 5 gives a summary of alternative nitrating systems, none of which was able to displace the ternary system $HNO_3/H_2SO_4/H_2O$ for industrial nitration.

Nitration with pure nitric acid is possible in principle. Esterification is not possible with acid concentrations below 75 %. Acid concentrations less than 75 % cause the formation of the unstable so-called Knecht compound, which has been described as either a molecular complex or an oxonium salt of the nitric acid. Cellulose nitrates with 5 – 8 % N, which dissolve in excess acid, are formed at acid concentrations of 78 – 85 %. Nitrogen contents of 8 – 10 % are attained at concentrations between 85 and 90 % HNO_3; these products have a strong tendency to gelatinize. Heterogeneous nitration without apparent swelling takes place at a HNO_3 concentration above 89 %, and 13.3 % N can be achieved with 100 % HNO_3. Nitration can be increased to 13.9 % N with 100 % HNO_3 by addition of inorganic salts such as sulfates, acid phosphates, and particularly nitrates, preferably in a 15 % concentration.

Table 5. Nitrating systems

Nitrating system	Max. N content, %	Comments
$HNO_3/H_2SO_4/H_2O$	13.4	Industrial nitration
$HNO_3 < 75$ %		"Knecht compound," unstable
78 – 85 %	8	Dissolution in the nitrating acid
85 – 89 %	10	Gelatinization
90 – 100 %	13.3	No swelling
HNO_3 + nitrates, sulfates, phosphates	13.9	
HNO_3 vapor	13.75	Slow reaction, stable nitrate
HNO_3 vapor + nitrogen oxides	13.8	
N_2O_5	14.12	
N_2O_5 in CCl_4	14.14	Trinitrate
HNO_3 in CH_2Cl_2	14.0	
HNO_3 in nitromethane	14.0	Homogeneous reaction
$HNO_3 + H_3PO_4/P_2O_5$	14.04	Rapid reaction without decomposition (polymer analogue)
	14.12	After extraction with methanol
HNO_3 + acetic acid/acetic anhydride	14.08	Great stability
	14.14	After extraction with ethanol
HNO_3 + propionic acid/butyric acid	14.0	

The nitric acid/phosphoric acid system is of special interest in a 1 : 1 ratio with 2.5 % phosphorus pentoxide added, with which an almost completely nitrated product of great stability was achieved. The nitric acid/acetic acid/acetic anhydride system in a ratio of 2 : 1 : 1 gives highly nitrated and highly stable products in which the fiber structure remains intact largely. After extraction of these nitrates with water or alcohol, the theoretical degree of substitution of the trinitrate may be attained.

Nitrating systems which achieve a high degree of nitration without degradation of the cellulose chain are of special scientific interest. This process is known as polymer analogous nitration. After a critical examination of all known nitrating mixtures, the nitric acid/acetic acid/acetic

anhydride system in a ratio of 43 : 32 : 25 at 0 °C [42] was recommended for determining the molecular mass of native celluloses of such solutions by using absolute methods and the intrinsic viscosity number [43]. The system anhydrous nitric acid in dichloromethane also allows the application of such polymer analogous reactions at temperatures between 0 and −30 °C [44]. Other authors [45] prefer the system nitric acid/phosphoric acid/phosphorus pentoxide.

Nitration in the Laboratory. Preparative cellulose nitration with HNO_3/H_2SO_4 nitrating acid to products with whatever N content up to 13.65 % is desired, stabilization and stabilization tests, nitration with the nitric acid/phosphoric acid (< 13.9 % N) and nitric acid acetic anhydride systems up to the trinitrate, denitration with hydrogen sulfide to cellulose II, the analytic determination of the N and sulfate content, and the solution of the CN and the viscosity determination of the solution are extensively described in [46].

1.2.5. Commercial Types and Grades

Cellulose nitrates receive, because of their fluffy structure and cottonlike appearance, the additional designation "cotton."

Two parameters are decisive for the industrial use of cellulose nitrate:

Nitrogen content (including the resulting solubility properties)
Viscosity

As seen in Table 6, cellulose nitrates with differing nitrogen contents have various applications. Cellulose nitrates for lacquers are available in numerous viscosities. It is possible to categorize all stages of viscosity according to the European norm (DIN 53179), but the viscosity of cellulose nitrates is primarily categorized by using the Cochius method and the British or American ball drop method (ASTM D 1343-69).

In addition to the so-called cotton types densified CN types are available. These may be obtained by either nitrating compressed cellulose or by subsequently compressing the fluffy cellulose nitrate. It is possible to almost double the bulk density by compression.

For safety reasons, the commercially available CN cotton types must be wetted with at least 25 wt % water or aliphatic alcohols. In addition to water, ethanol, *n*-butanol, and 2-propanol may also be used as wetting agent.

Table 6. Cellulose nitrate types

Type	N content, %	Degree of substitution (DS)
Celluloid cotton	10.5 – 11.0	1.82 – 1.97
Alcohol-soluble >lacquer cotton	10.9 – 11.3	1.94 – 2.06
Lacquer cotton moderately soluble in alcohol	11.4 – 11.7	2.08 – 2.17
Ester-soluble >lacquer cotton	11.8 – 12.2	2.20 – 2.32
Powder cotton	12.3 – 12.9	2.55 – 2.57
Gun cotton	13.0 – 13.6	2.58 – 2.76

The largest manufacturers of cellulose nitrates are the following:

Hercules Inc.	USA
Wolff Walsrode AG	FR Germany
Hagedorn	FR Germany
WNC Nitrochemie GmbH	FR Germany
Société Nationale des Poudres et Explosifs (SNPE)	France
Imperial Chemical Industries (ICI)	Great Britain
S.I.P.E. Nobel S.p.A.	Italy
Unión de Explosivos Río Tinto S.A.	Spain
Bofors	Sweden
Asahi	Japan
Daicel Chemical Industries, Ltd.	Japan

Many countries in South America, Asia, and Eastern Europe maintain small CN production facilities. The total world capacity may be estimated to 150 000 t/a of dry cellulose nitrate.

Other Commercial Types. Also available, in addition to cellulose nitrate cotton types, are so-called cellulose nitrate chips, made from cellulose nitrate plasticized by gelatinizing softeners. For safety reasons, the softener content has been established at a minimum of 18 wt %. Chips are preferred in processes where alcohols interfere in the formulation of lacquers.

The dispersions of cellulose nitrate with softeners or resins in water manufactured by the Wolff Walsrode AG are other available forms. The solvent-free or low-solvent dispersions are not polluting and may be used in all areas in which cellulose nitrate lacquers also are used [55].

1.2.6. Analysis and Quality Control

The most important analytical characteristics relate to the determination of the N content and, thereby, the average degree of substitution (DS), as well as the viscosity of the solution as a measure of the average molecular mass or chain-length.

Analytic Tests. The most commonly used analytic procedures are summarized in [25], [46], [47], and [48].

Dry content is determined by careful drying of a small, thinly layered alcohol or water-wet sample at room temperature for 12 – 16 h, in a weighing glass at 100 – 105 °C for 1 h, or with compressed warm air at 60 – 65 °C for 0.5 – 1 h.

Ash content is determined by decomposing a dried sample with HNO_3 and incinerating the residue. Specifications require that the ash content should not be above 0.3 %.

N-content is determined by reducing nitrates according to the following reaction (Schulze-Thiemann):

$$NO_3^- + 3\,FeCl_2 + 4\,HCl \longrightarrow 3\,FeCl_3 + Cl^- + 2\,H_2O + NO$$

or by the following reaction:

$$2\,NO_3^- + 4\,H_2SO_4 + 3\,Hg \longrightarrow 3\,HgSO_4 + SO_4^{2-} + 4\,H_2O + 2\,NO$$

The resulting NO is collected in a Du Pont nitrometer.

Stability Tests [25]. *Deflagration Temperatures:* Well-stabilized CN deflagrates at temperatures above 180 °C.

Bergmann – Junk Test [29]: A quantity of 2 g of dried CN is kept at a temperature of 32 °C for 2 h in a special apparatus for the elimination reaction, after which time the amount of the developed nitrous gases (after reduction to NO) is determined. CN is stable according to this test if no more than 2.5 cm^3 of NO per gram is measured.

Warm Storage Test: A quantity of 5 g of dried CN is stored in a glass-stoppered tube at 75 °C. Note is then made when the first nitric oxide (red-brown gas) becomes visible. Well-stabilized CN can be stored at 75 °C for at least 10 days.

ASTM Stability Test [48]: After storage at 134.5 ± 0.5 °C the time is noted in which the nitrous gases discolor methyl violet test paper.

Viscosity. *Viscosity according to DIN 53179:* If CN is dissolved in acetone in the appropriate concentration, CN solutions meeting this requirement show a apparent dynamic viscosity of 400 ± 25 mPa · s in the ball drop viscometer according to Höppler (ball no. 4) at 20 °C (Table 7).

Cochius Viscosity [25]: The viscosity of the various cellulose types is measured in commonly used solvent mixtures:

A and AM types: butanol/ethylene glycol/toluene/ethanol in the following proportions 1 : 2 : 3 : 4

E types: butanol/butyl acetate/toluene in the following proportions 3 : 4 : 5

Dried CN is dissolved in varying concentrations depending on the type and the time which an air bubble requires to rise 500 mm between two calibrations in a 7 mm Cochius tube at 18 °C is measured in seconds. The Cochius seconds are converted to absolute viscosity units mPa · s by multiplying with the factor 3.64 mPa.

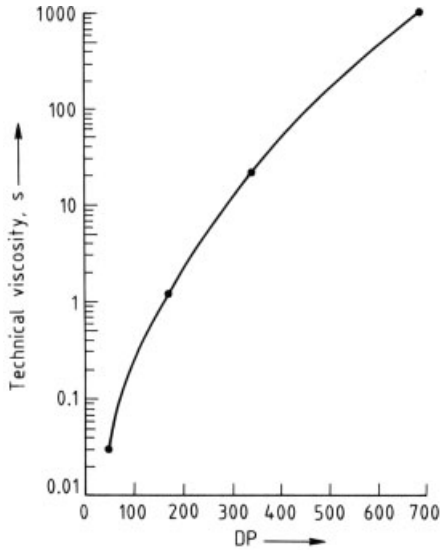

Figure 6. Degree of polymerization (DP) and technical viscosity ("ball drop" in seconds of a 12.2 % CN solution with 12 % N in acetone). DP = 170 × viscosity

Ball drop method according to ASTM [49]: Dried CN is dissolved according to its viscosity stage in 12.5, 20.0, or 25.0 % ethanol/toluene/ethyl acetate according to [48]. The

Table 7. Characterization of cellulose nitrates according to DIN 53179

A types	CN concentration, % (absol. dry)	AM types	CN concentration, % (absol. dry)	E types	CN concentration, % (absol. dry)
				E 1440	4
				E 1160	7
				E 950	9
				E 840	12
		AM 760	14		
		AM 750	15		
				E 730	15
		AM 700	17		
				E 620	21
				E 560	22
				E 510	24
A 500	27	AM 500	27		
A 400	30			E 400	27
				E 375	32
		AM 330	36	E 330	34

drop time of the balls with a diameter between 1/4 and 1/16 in. at 25 °C is given in seconds or converted into Pa · s. Figure 6 shows the relationship between the degree of polymerization and the technical viscosity (fall velocity of the balls in a 17.2 % CN solution in acetone).

Comparative viscosity charts for converting the various viscosities and comparing the various types are found in [25].

Solubility and Color. The color and cloudiness of solutions produced according to [48] are tested visually. Consistency, appearance, and depth of color can be controlled according to [50].

Dilution with Toluene. Toluene is added to a 12.2 % CN solution in butyl acetate at 25 °C until CN continuously precipitates. The dilution factor is noted. The dilution ratio of CN solutions with other solvents and blending agents is determined according to [51].

Film Test. The solutions made according to [48] are diluted with an equal volume of butyl acetate and poured as a film onto a glass plate. The dried films are examined for undissolved particles, surface structure, transparency, and gloss.

1.2.7. Uses

Explosives. Explosives may be categorized according to their use:

blasting agents
propellants and shooting agents
detonating agents
igniting agents
pyrotechnical agents

Cellulose nitrates are used primarily as propellants and gun powder, whereby the following distinctions can be made: *monobasic powder,* which is based solely on cellulose nitrate; *dibasic powder,* which contains further energy carriers such as, for example, nitroglycerin or diglycol dinitrate in addition to cellulose nitrate; *tribasic powder,* which contains in addition to the components of the dibasic powder a third agent such as nitroguanidine.

The selection of the cellulose nitrate is of special importance. The types of cellulose nitrates that differ in the degree of nitration were standardized as follows:

CP I (Collodium powder) also known as guncotton, nitrogen content: 13.3 – 13.5 %
CP II (Collodium) nitrogen content: 12.0 – 12.7 %, mostly 12.6 %
PE (Powder standard) nitrogen content: 11.5 – 12.0 %, mostly 11.5 %

Aromatic amines, such as diphenylamine, are added to gunpowder as stabilizers. They are capable of binding the nitrous gases generated during the decomposition of the nitric acid ester. A mixture of ca. 80 % highly nitrated gunpowder (13.4 % N) and ca. 20 % less-nitrated collodium cotton (12.5 % N) is used for the production of the monobasic propellant powder. Since cellulose nitrate granules are easy to charge electrostatically, they are made conductive with a fine graphite coating.

The multibasic powders usually contain cellulose nitrate CP II. Mixtures of 40 % PE cotton and 60 % CP I are also used because they have the same energy content as CP II with 12.6 % N.

The introduction of a third component to tribasic powder results in a lower heat of combustion in comparison to dibasic powder, thereby lengthening the life of the gun barrel.

Gunpowder is used in small-arms ammunition as well as large-caliber guns and tanks. (For further details → Explosives.)

Lacquers. Cellulose nitrate lacquers are characterized by the outstanding film-forming properties of the physically drying cellulose nitrate. Moreover, cellulose nitrate is compatible with many other raw materials used in lacquers and can be used advantageously in combination with resins, softeners, pigments, and additives.

In addition to the nonvolatile lacquer components, the composition of the solvent mixture is decisive for the formation of a film.

The most important uses for nitro lacquers are as follows: wood lacquers (especially furniture lacquers), metal lacquers, paper lacquers, foil lacquers (also as hot sealing lacquers, e.g., cellophane, plastic, and metal foils), leather lacquers, adhesive cements, putties, and printing ink (for flexo and gravure printing).

The processes used for applying cellulose nitrate lacquers to substrates are as follows: spraying (compressed-air, airless, and electrostatic spraying), casting (for example, with a curtain coater), rolling (especially for the application of small amounts of lacquer), doctor knife coating, and dipping.

The casting and rolling processes are used for lacquering large, even areas. Irregularly shaped objects are sprayed. The choice of a suitable type of cellulose nitrate (e.g., completely or moderately soluble in alcohol, soluble in esters, degree of viscosity) is dependent on the lacquer type. A highly viscous cellulose nitrate type is used if elastic and thin applications are desired (e.g., leather). However, if hard and thick layers are desired, low-viscosity types are preferred.

The concentration or the degree of viscosity of the cellulose nitrate determine the viscosity of the lacquer solution. However, the formulation of the lacquer must be taken into consideration when the mode of application is chosen. For example, a highly viscous dipping lacquer cannot be sprayed or casted.

Furthermore, the striking differences between ester-soluble and alcohol-soluble types should be taken into consideration when nitro lacquers are formulated (Table 8).

For further information on the formulation of cellulose nitrate lacquers, see [52], [53], and also → Paints and Coatings.

Dispersions. Conventional cellulose nitrate lacquers contain between 60 and 90 % organic solvents, which are released during drying. For economic and environmental reasons, it is desirable to substitute organic solvents by water. Aqueous cellulose nitrate/softener dispersions (e.g., Isoderm, Bayer AG; Coreal, BASF; Waloran N, Wolff Walsrode AG) are available for such absorbing substrates as leather [54]. Other aqueous cellulose nitrate dispersions for use on wood, foil, and metal have also been developed (Waloran N, special-types, Wolff Walsrode AG) [55]. The film forming process of water-insoluble cellulose nitrate requires a small amount of coalescents in the dispersion systems.

Celluloid. A special use of cellulose nitrate is in the production of celluloid [56]. Cellulose nitrate with a nitrogen content of 10.5 – 11.0 % is mixed in a kneader with softeners, particularly camphor, and solvents (alcohols).

Normal celluloid contains ca. 25 – 30 % camphor and 70 – 75 % cellulose nitrate. Celluloid that contains 10 – 15 % solvent can be formed into the desired articles in heated piston or screw presses (e.g., tubes and round and profile rods).

In the past decades, celluloid has been widely replaced by synthetic materials and thermoplastics. Celluloid is still of economic importance in the following areas: combs and hair ornaments, toilet articles, office supplies (drafting and measuring instruments), ping-pong balls, and various special uses.

1.2.8. Legal Provisions

Toxicology and Industrial Safety. Concentrated sulfuric acid, nitric acid, and nitrous gases formed during the production of cellulose nitrate are considered hazardous chemical products [57]:

Table 8. CN lacquer cottons

Ester-soluble type	Alcohol-soluble type
Possible use of alcohol in the lacquer formulation	Use of alcohol, especially ethanol, in any desired amount as a solvent
Good dilutability with aliphatic and aromatic hydrocarbons	Good dilutability with aromatic hydrocarbons
Very rapid solvent release	Rapid solvent release
Formation of hard films	Formation of films with thermoplastic properties
Attainment of good mechanical properties as far as the cold-check test, stretch, hardness, and tensile strength are concerned	Attainment of good mechanical properties; some special problems of lacquer production may be solved such as: Lacquers which can be diluted with ethanol in any desired manner (wood polishes) Odorless lacquers (printing inks) Gel dipping lacquers Hot sealing waxes (cellophane lacquers and aluminum foil lacquers)

1) Sulfuric acid
   5 – 15 % EC-No. 016-020-01-5
   above 15 % EC-No. 016-020-00-8
2) Nitric acid
   20 – 70 % EC-No. 007-004-01-9
   above 70 % EC-No. 007-004-00-1
3) Nitrous gases
   EC-No. 007-002-00-0

They are subjected to the Arbeitsstoffverordnung (working substance regulation) [58] and must, therefore, be adequately labeled.

Concentrated nitric acid and mixed nitrating acids are oxidizing when brought into contact with organic materials [59]. The MAK values (maximum working place concentration) are as follows:

nitric acid vapors: 10 mL/m^3 (ppm); $\triangleq$ 25 mg/m^3

nitrogen oxides (NO_2): 5 mL/m^3 (ppm); $\triangleq$ 9 mg/m^3

Employees should be examined regularly for obstructive respiratory tract illnesses.

Cellulose nitrate is neither toxic nor hazardous to health [60]. Damping agents in CN and nitrous gases which may be formed during combustion and smoldering processes are potentially hazardous to health if inhaled.

Commercially available phlegmatized cellulose nitrate for the production of lacquer with less than 12.6 % N contains at least 18 % of a gelatinizing softener. According to the first paragraph in [58], cellulose nitrate is a hazardous substance and must be packaged and labeled accordingly. EEC regulations (1982) are similar.

Damping agents such as ethanol and 2-propanol are not subjected to these regulations; butanol belongs to category II d, but is not considered to be hazardous to health in a damped mixture of a maximum 35 % concentration.

Storage and Shipping. Cellulose nitrate, especially guncotton, burns in air with a yellow flame and deflagrates if present in larger quantities, especially after rapid heating. An explosion can be caused by friction or a sharp impact. Dry CN has electrostatic charge. Friction, particularly on metals but also on plastics, can cause sparks which lead to a deflagration. Therefore, cellulose nitrate should be stored in a moist and cool place [60], [61]. Rooms in which cellulose nitrate is processed must be adequately protected according to the guidelines for protection from explosions.

Cellulose nitrate is subjected to the regulations governing explosives [62]. The transportation of phlegmatized cellulose nitrates proceeds according to the most recent versions of the hazardous materials regulation; see [25]. Wetted cellulose nitrate is shipped in thick-walled, galvanized, tightly closing iron or fiber drums which are adequately labeled.

Dried cellulose nitrate may not be shipped under any circumstances.

For further information on the properties, handling, storage, and transportation of hazardous goods, see also [63].

1.3. Other Inorganic Cellulose Esters

Summaries on the esterification products of cellulose with other inorganic acids may be found in [12–19]. For publications on the modification of cellulose, including esterification, see [64].

1.3.1. Cellulose Sulfates

Cellulose sulfates [*9032-43-3*] are the most frequently investigated of all other inorganic cellulose esters. The ability of concentrated sulfuric acid to dissolve cellulose, particularly in concentrations between 70 and 75 %, has been known since 1819. After precipitation immediately following dissolution, the cellulose contains little or no bound sulfate. An almost homogeneous esterification takes place only if the cellulose is left in a sulfuric acid solution over a longer period of time. However, the ester yield is very poor. The major portions of the reaction products consist of hydrolytically split decomposition products with a maximum degree of substitution of 1.5.

In their free acidic form, cellulose sulfates are fairly unstable and easily saponified. A semiester was developed in 1953 in the United States [65] in an esterification mixture consisting of 1 mol of cellulose with 20 – 30 % water, 3.5 – 15 mol of sulfuric acid, 0.3 – 1.0 mol of a primary or secondary C_3–C_5 alcohol, and an inert volatile organic solvent such as toluene or carbon tetrachloride (reaction temperatures between − 5 and − 10 °C). The product was soluble in hot or cold water, yielded relatively stable, clear, and highly viscous solutions, and was recommended for use as a thickener for aqueous systems (emulsion paints and printing inks, printing pastes for textiles, and food products), as well as for fat- and oil-proof finish, and as paper glue. This product, however, is of no economic importance.

Numerous attempts have been made to find improved preparation methods for water-soluble cellulose sulfates stable to saponification. Known reaction systems are summarized in a tabular overview [65]. The reaction of cellulose with sulfuric acid in organic solvents, especially in lower-mass aliphatic alcohols, gives by way of a heterogeneous reaction fibrous and water-soluble cellulose sulfates with a maximum DS value of 1. More highly substituted products are obtained by reaction with sulfuric acid/acetic acid anhydride (up to a DS value of 2.8) or esterification with chlorosulfuric acid in pyridine or formamide. The reaction with SO_3 only or in various organic systems yields trisubstituted products. The reaction mechanism may be described as the addition of the strongly electrophilic SO_3 to the hydroxyl groups with the succeeding disintegration of the intermediately formed oxonium ion.

$$\text{Cell–OH} + SO_3 \longrightarrow \left[\text{Cell–}\overset{\overset{H}{|}}{\underset{+}{O}}\text{–}SO_3^-\right] \longrightarrow \text{Cell–O–}SO_3^- + H^+$$

Completely water-soluble, highly viscous sodium cellulose sulfate semiesters are obtained in homogeneous systems by the reaction of cellulose nitrite [67]. The intermediate that is formed and dissolved, cellulose nitrite, is obtained in the N_2O_4/dimethylformamide system and is at the same time transesterified by the SO_3/DMF complex. Uniformly substituted cellulose sulfate with a range of DS values between 0.3 and 2.0 and solution viscosities up to 7 000 mPa · s (in 1 % solution) can be obtained by using this process [68]. Such transesterified products can be cross-linked by metal ions to form highly effective thickening agents for aqueous media [69].

Such processes have been further developed [66] and make interesting novel fields of application accessible as a result of the rheological and gel-forming properties of the Na cellulose sulfate semiester.

Mixed esters such as cellulose acetate sulfates, cellulose acetate butyrate sulfates [70], cellulose acetate propionate sulfate [71], and ethyl cellulose sulfates [70], [72] are described in the patent literature.

Being polyelectrolytes, cellulose sulfates form salts and have ion-exchanging properties; thus, they have been recommended for use as cation exchangers [64, p. 65], [73], [74].

1.3.2. Cellulose Phosphate and Cellulose Phosphite

Reaction of cellulose with aqueous phosphoric acid gives the following unstable addition compound: $3\,C_6H_{10}O_5 \cdot H_3PO_4$, from which the cellulose can be regenerated unchanged by reaction

with water. Cellulose phosphates [*9015-14-9*] with a low phosphorus content are obtained by reacting cellulose or linters with phosphoric acid in an urea melt [75]. Higher phosphorus contents and a lower degradation rate of the cellulose may be obtained with excess urea at reduced reaction time (ca. 15 min) and at high temperature (ca. 140 °C). Water-soluble cellulose phosphate with a high degree of substitution may be obtained from a mixture of phosphoric acid and phosphorus pentoxide in an alcoholic medium [76].

Phosphorylated cellulose fibers show increased swelling after partial hydrolytic degradation and transfer into the alkali salt form and were, therefore, suggested for use as adsorbents [77].

Cellulose phosphates with a 17 % phosphorus content (this represents about 3/4 of the maximal possible substitution of triphosphate with 23 % phosphorus) were already produced in 1933 by reacting cellulose with a mixture consisting of concentrated sulfuric acid and phosphoric acid in the presence of a weakly acidic catalyst [78].

Cellulose phosphites [*37264-91-8*] and cellulose phosphonates may be prepared by transesterification with alkyl phosphites. All cellulose esters containing phosphorus have fire-retarding properties [78] and have attracted some interest due to their ion-exchanging effect [74], [79], but are not yet industrially used.

1.3.3. Cellulose Halogenides

Various preparative methods are suitable for the synthesis of halogenated cellulose derivatives [64, p. 64]. Halogenation can be carried out by transesterification of such cellulose esters as tosylate, nitrate, and sulfate with hydrohalic acids [80]. Nucleophilic substitution proceeds considerably faster in homogeneous systems than in heterogeneous aqueous systems.

The Finkelstein transesterification process of cellulose nitrate with sodium iodide in anhydrous acetone leads to deoxyiodo cellulose. The reaction of cellulose with thionyl chloride, $SOCl_2$, in the presence of pyridine produced a monosubstituted, but strongly decomposed and unstable, hydrogen chloride ester.

Halogenation of cellulose improves its water-resistant and flame-resistant properties. Slight fluorination increases oil resistancy and lowers the soiling potential of cellulose textiles [64]. Commercial applications are not yet known.

1.3.4. Cellulose Borates

The preparation of cellulose borate succeeded by means of transesterification of methyl and *n*-propyl borate with cellulose [64, p. 7]. The products with a maximum DS value of 2.88 are, however, extremely sensitive to hydrolysis and alcoholysis.

1.3.5. Cellulose Titanate

Cellulose can be reacted to cellulose titanates in a heterogeneous reaction system by reacting it with titanium tetrachloride in DMF or with chlorinated anhydrides, chlorinated ester anhydrides, and esters of the hypothetical orthotitanic acid $Ti(OH)_4$ [81]. Ethyl trichlorotitanate has been shown to be the most reactive. Esters with 16 % titanium content are possible.

Cellulose esters with a titanium content between 3 and 5 % do not burn or smolder. They possess considerable hydrolytic stability in neutral and weakly alkaline media, but are easily hydrolyzed at a low pH.

1.3.6. Cellulose Nitrite

The nitrite of cellulose came of scientific and possibly practical interest as a cellulose derivative in 1974 [67]. It is obtained by reacting cellulose with nitrosyl compounds such as dinitrogen tetroxide, N_2O_4 (corresponding to nitrosyl nitrate), or nitrosyl chloride, NOCl, in dimethylformamide or dimethylacetamide as a proton acceptor and solvent for the resulting ester. The reaction proceeds in a homogeneous phase to the trinitrite.

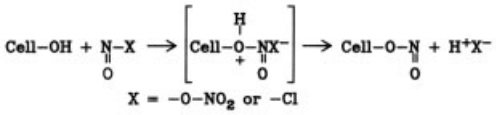

Cellulose nitrite is extremely sensitive to hydrolysis. Chain degradation to a DP of 200

(level-off DP) was observed to take place within 3 h in the presence of water. The scientific and preparative importance of cellulose nitrite is based on its high reactivity, which may be used to produce many other cellulose esters, also mixed esters, by transesterification in a homogeneous phase [82]. Transesterification to stable cellulose sulfates has already been mentioned [67]. In this manner, water-soluble cellulose nitrates with a DS value of 0.5 – 0.6 may also be obtained [83].

Cellulose solutions produced under cold conditions (up to ca. 5 °C) in a N_2O_4/DMF system are relatively stable to degradation and can be produced, depending on the DP of the cellulose, up to a concentration of 14 %. The cellulose can be regenerated in an unaltered form to cellulose II, with the result that this process has already been considered as an alternative to the environmentally detrimental viscose process [84]. Not only an attempt was made to achieve good mechanical textile properties from the regenerated fibers, but also to recycle the expensive solvent. An economic solution to the competition with the viscose process has not yet been found.

1.3.7. Cellulose Xanthate

Cellulose xanthate [*9032-37-5*], an important intermediary molecule for the production of regenerated cellulose according to the viscose process (see Cellulose, Chap. 3.1.1.), must also be considered as an ester of an inorganic acid, namely the nonexistent thiol – thion carbonic acid.

The O ester of this compound with organic residues is the xanthic acid and the appropriate salts. Sodium cellulose xanthate is obtained by reacting alkali cellulose with carbon disulfide, which dissolves in dilute sodium hydroxide to an orange-yellowish, highly viscous solution, the so-called viscose.

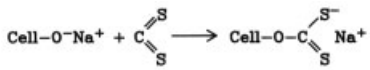

The regenerated cellulose is precipitated in the form of fibers (rayon, cord, and rayon staple), foils (cellophane), or tubes in precipitation baths containing sulfuric acid and salts (see → Cellulose, Chap. 3.1.3.). About 4×10^6 t of regenerated cellulose is presently produced worldwide by using the viscose process.

For further information, see → Cellulose, Chap. 3.

2. Organic Esters

Cellulose can theoretically form an unlimited number of organic acid esters because of its anhydroglucose units with three reactive hydroxyl groups each. Industrial possibilities are, however, drastically limited by the complex nature of the cellulose molecule. Highly esterified organic esters are, therefore, only produced from a few aliphatic fatty acids with up to four carbon atoms.

Cellulose acetate [*9004-35-7*], cellulose acetate propionate [*9004-39-1*], and cellulose acetate butyrate [*9004-36-8*], which have been known for quite some time and are in large-scale production, are especially important esters. Formic acid esters are, because of their instability, of no industrial importance.

Only cellulose acetate phthalate [*9004-38-0*] has found limited use as a tablet coating; all the other described cellulose esters such as cellulose palmitate, cellulose stearate, esters of unsaturated acids such as crotonic acid, or esters of dicarboxylic acids are not used industrially.

Historical Aspects. Cellulose acetate was first synthesized by P. Schutzenberger in 1865 by heating cellulose and acetic acid under pressure, whereby a product of very low molecularity was obtained. In 1879, A. P. N. Franchimont added sulfuric acid to the esterification process, which remains to this day the most frequently used catalyst. The limited solubility of cellulose acetate in less-expensive solvents and poor compatibility with the then-known softeners was a considerable obstacle for its industrial use. The problem was solved in 1904 when F. D. Miles and A. Eichengrün simultaneously succeeded in synthesizing an acetone-soluble secondary acetate by partially hydrolyzing a primary triacetate.

During World War I, the less-flammable airplane paints based on cellulose secondary acetate reached considerable importance as a re-

placement for nitrocellulose. At almost the same time, the manufacture of foils, films, synthetic silk, and plastic masses developed.

An especially high number of publications and patents were achieved between 1920 and 1935. Ultimately, only a few processes proved to be industrially useful, most of which are still used today.

Even though the technology of cellulose ester production is considered for the most part to be complete, research continues in the fields of rationalization and improvement of production methods, leading to products with greater uniformity and improved properties, as well as in the development of new fields of application, especially with the introduction of cellulose acetate propionate and cellulose acetate butyrate.

2.1. Cellulose Acetate

2.1.1. Chemistry of Cellulose Esterification

The esterification reaction of the primary and secondary hydroxyl groups of cellulose does not basically differ from that of other alcohols. The peculiarities lie in the macromolecular structure of the cellulose molecule. The splitting of the molecule chain competes with the catalyzed esterification, but can be fairly well controlled under appropriate conditions. The speed and completeness of the reaction is dependent on the quality of the cellulose, whereas the different reactivities of the primary and secondary hydroxyl groups [83], [86] have little influence on industrial processes.

Acids, acid chlorides, and acid anhydrides are possible *esterification reagents* for the three hydroxyl groups in each glucose unit.

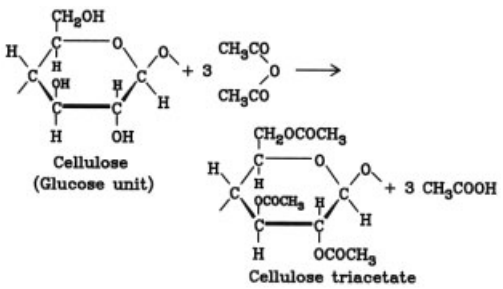

Esterification with free acids, with the exception of formic acid [87], whose esters are not stable, requires such high temperatures and catalyst concentrations that only low molecular mass products are obtained. Acid chlorides in pyridine, however, were suggested for use in the production of esters from higher fatty acids (lauric acid, stearic acid, and palmitic acid), without ever having attained any industrial significance. Attempt was made to manufacture cellulose acetate industrially with acetyl chloride and catalysts, but the process proved to be useless.

All industrial processes in current use, therefore, are based on acetic acid anhydride as a reactant, whereby theoretically 3 mol of anhydride per unit of glucose are used and 3 mol of acetic acid are formed.

Attempts to use ketene, which could be directly accumulated without incurring acetic acid, which must again be processed, did not lead to any results [88].

Numerous *catalysts* were suggested to accelerate the reaction. Only sulfuric acid and perchloric acid are of any practical importance. Zinc chloride, which is required in large amounts of 0.5 – 1 part per part of cellulose, is no longer used today because of the high reprocessing costs. Other mineral acids, however, are not sufficiently acidic in the water-free acetic acid – acetic acid anhydride system. The catalytic effect of sulfuric acid is primarily in the rapid and quantitative formation of acidic cellulose – sulfuric acid esters (sulfoesters) [89], which are substituted by acetyl groups as the reaction progresses and the temperature rises.

Due to the topochemical character of the reaction, soluble cellulose esters with low acetic acid content can be obtained only from the triester stage by way of hydrolysis. Incompletely esterized and insoluble derivatives are found in addition to the triester before completion of the reaction.

Cellulose esters with low acetic acid content are produced by subsequent *hydrolysis* in a homogeneous system by adding water or dilute acetic acid to destroy the excess anhydride and possibly by adding sulfuric acid for acceleration and then again split off under controlled conditions (temperature, water content, and time) a certain number of acetyl groups without further breakdown of the cellulose chain.

2.1.2. Raw Materials

Cellulose. The production of high-quality cellulose esters requires that special attention be paid to the selection of the starting materials. The cellulose bases generally consist of highly purified cotton linters with an α-cellulose content of over 99 % and celluloses from wood pulp which contain between 90 and 97 % α-cellulose.

After the long layered spinnable cotton has been freed of the cotton seed by ginning, the remaining shorter fibers on the seed pod are usually removed with two cuts before the seeds go to the oil presses for further processing. The first cut gives about 4 % longer linters relative to the entire cotton flower, which are preferentially processed to medicinal cotton, felt, paper, etc. The second cut gives about 8 % shorter layered linters, which are best suited for further chemical processing.

The raw linters undergo mechanical cleaning by means of screening, pressure boiling in a 3 – 5 % sodium hydroxide solution, and finally acid – alkaline bleaching. Special care should be taken during drying, since local overdrying of cellulose, the water content of which should lie between 3 and 8 %, impairs the reactivity considerably. Table 9 shows analytical values of good linters [90].

Table 9. Analytical values obtained from bleached linters according to [90]

α-Cellulose	99.7 %
β-Cellulose	0.2 %
γ-Cellulose	0.1 %
Carboxyl groups	<0.02 %
Total ash	0.02 %
Degree of polymerization	1000 – 7000

For a long time, cellulose from wood pulp could only be used for the manufacture of lower-quality cellulose esters because of the 90 – 95 % α-cellulose content. Celluloses with an α-cellulose content of 96 % have been available since about the 1950s. Due to special processing techniques, they give cellulose esters comparable to those produced from linters as far as tensile strength, color, clarity of the solutions, and light stability as well as thermal stability are concerned.

Acetic Acid Anhydride. Most manufacturers of cellulose acetate convert the resulting acetic acid to anhydride directly on the premises and adjust the concentrations as required for their process, generally between 90 and 95 %.

2.1.3. Industrial Processes

Only a few of the proposed industrial processes for the manufacture of cellulose esters have attained industrial significance. Even when the fact that no two manufacturers use identical processes is taken into consideration, the following categories can be distinguished:

1) Acetylation in a homogeneous system (solution acetate)
   - Use of glacial acetic acid as a solvent (glacial acetic acid process)
   - Use of methylene chloride as a solvent (methylene chloride process)
2) Acetylation in a heterogeneous system (fiber acetate)

Whereas the triester that is formed during esterification according to the solution acetate process goes into solution and can subsequently be hydrolyzed to secondary acetate, fiber acetate is formed in the presence of nonsolvents, similar to the nitration of cellulose. This method does not permit hydrolysis.

A flow diagram of the entire process is shown in Figure 7.

2.1.3.1. Pretreatment

The cellulose is first dried to obtain a moisture content of 4 – 7 %. Too little moisture would lower the reactivity of the cellulose, and too much moisture would lead to a higher acetic acid consumption and an extremely violent reaction start.

Acetic acid is generally used as a pretreatment reagent, which for some processes can contain small amounts of sulfuric acid to further improve the diffusion of the acetylating reagents. It is believed that the pretreatment process causes a partial swelling of the cellulose by splitting the hydrogen bonds so that the acetylating reagents can enter the fibers more rapidly.

A ratio of 30 – 100 parts of glacial acetic acid per 100 parts of cellulose, which is sprayed into the thoroughly mixed cellulose at temperatures

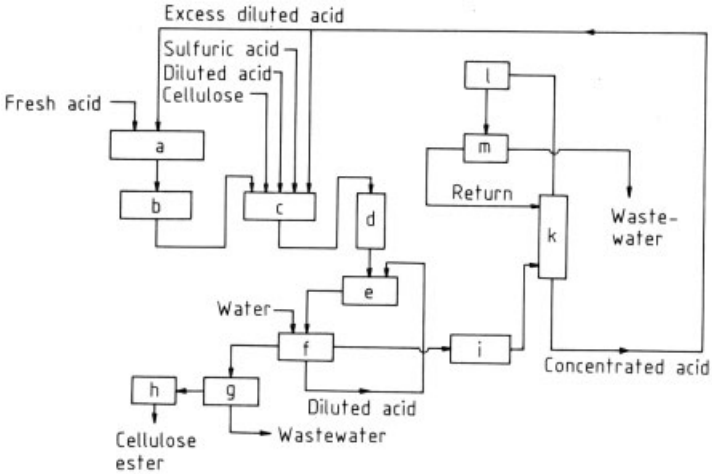

Figure 7. Flow chart for the production of cellulose esters according to [91]
a) Acid reconditioning; b) Acidanhydride; c) Esterification; d) Hydrolysis; e) Precipitation; f) Washing; g) Centrifuge; h) Drier; i) Evaporator; k) Azeotropic distillation; l) Cooler; m) Decanter

of up to 50 °C in appropriate equipment such as mixing vats, is usually sufficient for the pretreatment process. Pretreatment, depending on the process and temperatures, takes from one to several hours.

2.1.3.2. Esterification

Acetic Acid Process. The acetylation mixture consists of glacial acetic acid as a solvent for triacetate, an excess of 10 – 40 % glacial acetic acid anhydride and, depending on the process, 2 – 15 % sulfuric acid with respect to the cellulose.

Esterification begins after the initial spontaneous reaction of the water contained in the cellulose with part of the anhydride; the semiliquid mass uniformly saturated with acetylation mixture forms a fiber pulp while developing temperatures of up to 50 °C and ultimately a highly viscous solution.

A decomposition of the cellulose chain takes place parallel to the strongly exothermic esterification. The chain degradation can be controlled by controlling the reaction temperature, e.g., by adding cooled solution portionwise and cooling the reaction vessel down to the desired temperature range. The use of glacial acetic acid as a solvent, in which the highly esterified triacetate is poorly dissolved, results in gel formation and precipitation of the triacetate. This is due to the formation of sulfate acetate, which is formed as an intermediate during the reaction. It is reesterified to pure acetic acid esters especially at higher temperatures.

After the reaction solution is free of fibers, the degradation of the cellulose ester molecule can continue until the desired viscosity is attained. The reaction is then stopped by adding water or dilute acetic acid, which destroys the excess anhydride.

Cooled kneaders are suitable reaction vessels in that they allow a rapid and uniform mixture and catalyst distribution through intensive mixing. This is important for controlling the reaction (Fig. 8).

Methylene Chloride Process. Using methylene chloride (*bp* 41 °C) as a solvent presents several advantages over acetic acid. Due to the fact that methylene chloride is such an excellent solvent for primary triacetate, lower catalyst concentrations (1 % sulfuric acid) are required at higher esterification temperatures. Furthermore, due to its low boiling point, the heat of reaction can be removed by means of vaporization and return of the cooled methylene chloride. The reaction of highly viscous solutions can, thus, be better controlled. Finally, only a half to a third as much dilute acetic acid must

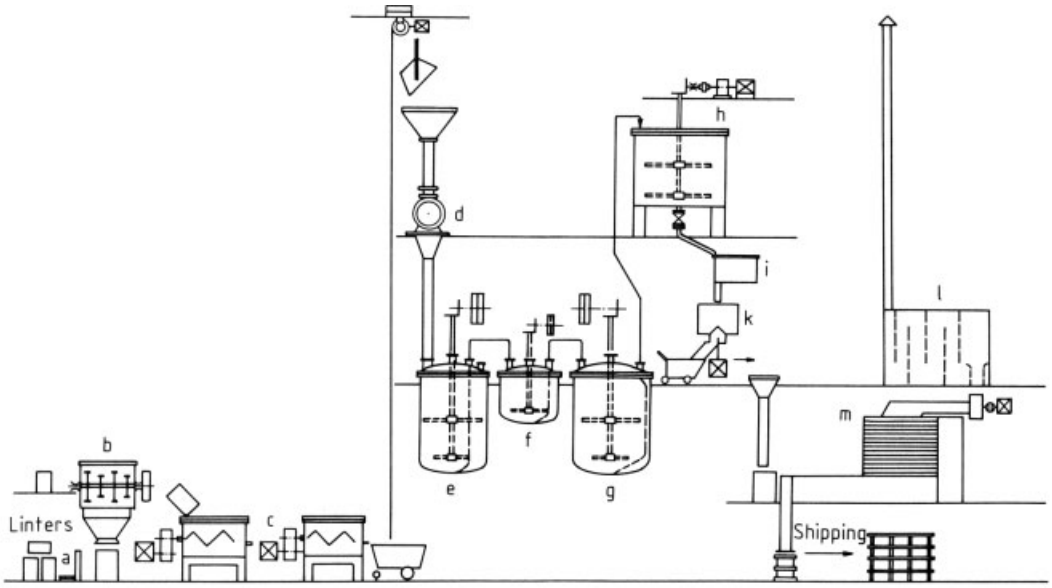

Figure 8. Cellulose acetate production by the kneader method according to [92]
a) Weighing scale; b) Sprinkling vat; c) Kneader; d) Mill; e) Rinsing vat; f) Stabilizing vat; g) Bleaching vat; h) Floater; i) Stock pan; k) Centrifuge; l) Dust chamber; m) Drier

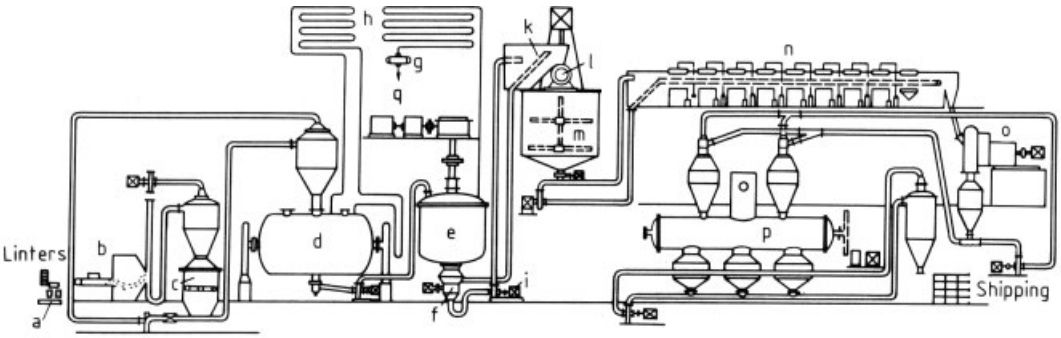

Figure 9. Cellulose acetate production according to the methylene chloride process [93]
a) Weighing scale; b) Bale opener; c) Sprinkling vat; d) Acetylator; e) Precipitating vat; f) Prebreaker; g) Vacuum vessel; h) Pipe cooler; i) Pump for viscous substances; k) Filter bath; l) Mill; m) Floater; n) Sprinkling line; o) Centrifuge; p) Vacuum shovel drier; q) To reprocessing of methylene chloride

be recycled compared with the glacial acetic acid process.

Table 10. Acetylating preparations according to the glacial acetic acid process and the methylene chloride process [92]

	Acetic acid process	Methylene chloride process
Cellulose	700 kg	3 500 kg
Pretreatment	700 kg glacial acetic acid	1 200 kg glacial acetic acid
Acetylation	1 900 kg anhydride	10 500 kg anhydride
	4800 kg glacial acetic acid	14000 kg methylene chloride
	50 kg sulfuric acid	35 kg sulfuric acid

Table 10 shows typical examples for acetylation according to the glacial acetic acid and methylene chloride processes. Figure 9 shows a sceme for the production of cellulose acetate according to the methylene chloride process.

Acetylation according to the methylene chloride process is carried out largely in rotating drums (roll vats) or in horizontal containers with shovel-like stirrers on both sides. The problem of corrosion, which arises during esterification and especially during hydrolysis, has only been

partially solved by using equipment constructed of bronze, high-alloy steels, or plates containing metals such as silver, titanium, or tantalum.

Fiber Acetate. Cellulose can be esterified maintaining its fiber structure by adding sufficient amounts of nonsolvents to the triacetate during acetylation. Carbon tetrachloride, benzene, or toluene can be used as nonsolvents [94], [95].

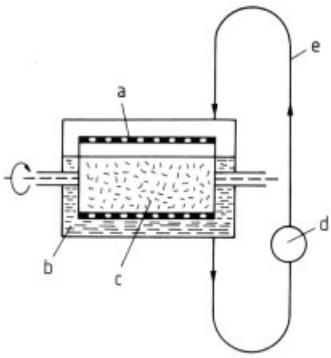

Figure 10. Acetylation equipment for cellulose triacetate fibers [92]
a) Perforated drum; b) Reaction solution; c) Cellulose fibers; d) Cooler for acetylating liquid; e) Acetylating liquid circulation

Temperature and catalyst concentrations are similar to those required for the solvation process. A large amount of liquid is required to keep the loose voluminous cellulose in suspension. Perchloric acid is preferred as a catalyst because of the great difficulty with which the sulfuric acid ester is split in this system.

The hydrolysis of fiber acetate to acetone-soluble esters is not possible in a heterogeneous system. Its utilization is limited to special applications, such as the manufacture of foils and films from triacetate.

As shown in Figure 10, fiber acetate is produced by rotation in various directions and at various speeds in a perforated drum enclosed in a metal casing. The shaft of the drum is hollow so that liquid may be added during rotation [96].

Continuous Processes. In spite of numerous attempts, continuous acetylation and hydrolysis processes have achieved only limited industrial applications and are not used in the production of high-quality esters for, among others, plastic masses. Due to the varying reactivities of the different celluloses, considerable variations between batches are possible so that the development of a continuous process is seriously impeded.

2.1.3.3. Hydrolysis

After the esterification is completed, the process is interrupted by adding water or dilute acetic acid. Sufficient water must be added to decompose the excess anhydride and adjust the water content of the solution to 5 – 10 % so that, aside from hydrolysis, no further decomposition of the molecular mass occurs, but at the same time the bound sulfuric acid is almost completely split off. The speed of hydrolysis depends on the temperature, which – depending on the process – ranges between 40 and 80 °C, and on the amounts of sulfuric acid and water.

The course of hydrolysis is constantly monitored by checking the solubility of the secondary acetate. The reaction is terminated at the desired degree of substitution by neutralizing the catalyst, preferably with sodium acetate or magnesium acetate. The proportion of free hydroxyl groups in the ester should be maintained as reproducibly precise as possible because it determines the properties and the utilization of the ester.

2.1.3.4. Precipitation and Processing

The product precipitates in the form of flakes or powder during intense stirring after water or dilute acetic acid has been gradually added. The methylene chloride process requires that the methylene chloride be completely distilled before precipitation occurs. A part of the precipitating acid is added to the solution to just before the point of precipitation when the main portion of the acid is added. The process must be carefully monitored to achieve a product with an open structure which can be easily rinsed out. The precipitate can be broken down and thoroughly rinsed, whereby the dilute acid is brought back into the cycle. Rinsing is done primarily by using continuous methods based on the countercurrent principle (Fig. 9).

High-quality products for plastic masses are stabilized and bleached. Bound sulfuric acid

remnants are removed by either boiling under pressure or heating in 1 % mineral acids during stabilization.

After further rinsing and removal of excess water by suctioning or by thrust extraction, the product is carefully dried, preferably in a vacuum shovel drier, to a water content of < 1 – 3 %.

The cellulose acetate yield from a good process is at least 95 % of the theoretical yield.

Major manufacturers of cellulose acetate are the following (1971) [103]: Bayer AG, Leverkusen; British Celanese Ltd., London; Celanese Corp. of America, New York; Courtaulds Ltd., Manchester; Daicel Ltd., Osaka, Japan; E. I. DuPont de Nemours & Co., Inc., Wilmington; Eastman Kodak Comp., Kingsport; UCB Fabelta, Tubize, Belgium; Gevaert-Agfa, Antwerp; Hercules Powder Comp. Ltd., London; Rhodiaceta, Lyon; Soc. Rhodiatoce SpA, Milan; Rhône-Poulenc, Paris; VEB Orbitaplast, Eilenburg, GDR.

A number of products are further processed to fibers, films, or injection-molding compounds in the manufacturers' own facilities and never reach the market as a raw material.

2.1.4. Recovery of Reactants

The recovery of the incurred large amounts of acetic acid is a decisive factor for the profitability of a process.

Recovery of Acetic Acid. Depending on the process, 2 – 6 parts of 15 – 25 % dilute acetic acid per part of cellulose accumulate. They must be reprocessed to glacial acetic acid and acetic acid anhydride.

Only continuous processes consisting of a combination of extraction and azeotropic distillation are of practical importance. The dilute acid is, for example, extracted with ethyl acetate in a countercurrent and is subsequently distilled, so that 99.8 % pure glacial acetic acid can be removed from the bottom of the column while the azeotropic ethyl acetate water is removed from the top.

Recovery of Acetic Acid Anhydride. Since only a portion of the accumulated glacial acetic acid is required for the acetylation process, the remainder must be converted to glacial acetic acid anhydride.

The ketene process developed by the Wacker Co. (→ Acetic Anhydride, Chap. 1.3.1.1.) is in general use [97]: Pure, almost anhydrous glacial acetic acid is continuously vaporized under a vacuum and is split to ketene in the presence of small amounts of the catalyst triethyl phosphate; ketene then reacts with glacial acetic acid to form the anhydride (→ Acetic Anhydride, Chap. 1.3.1.2.).

Recovery of Methylene Chloride. Methylene chloride can be recovered inexpensively because of its insolubility in water, which allows its recovery from the raw solution without further processing steps in almost pure form.

2.1.5. Properties of Cellulose Acetate

Table 11. Physical characteristics of cellulose acetate [90], [92], [98]

Characteristic	Triacetate	Secondary acetate
Density, g/cm^3	1.27 – 1.29	1.28 – 1.32
Thermal stability, °C	>240	ca. 230
Tensile strength of fibers, kg/mm^2	14 – 25	16 – 18
Tensile strength of foils		
longitudinal, kg/mm^2	12 – 14	8.5 – 10
transverse, kg/mm^2	10 – 12	8.5 – 10
Refractive index of fibers toward the fiber axis		
longitudinal	1.469	1.478
transverse	1.472	1.473
Double refraction	−0.003	+0.005
Dielectric constant ε		
50 – 60 Hz	3.0 – 4.5	4.5 – 6.5
10^6 Hz		4.0 – 5.5
Dielectric loss factor tan δ		
50 – 60 Hz	0.01 – 0.02	0.007
10^6 Hz		0.026
Specific resistance, $\Omega \cdot$ cm	$10^{13} - 10^{15}$	$10^{11} - 10^{13}$
Specific heat, J g^{-1} K^{-1}		1.46 – 1.88
Thermal conduction, J m^{-1} h^{-1} K^{-1}		0.63 – 1.25

Cellulose acetate and the other fatty acid esters are white, amorphous products that are commercially available as a powder or flakes. They are nontoxic, odorless, tasteless, and less

flammable than nitrocellulose. They are resistant to weak acids and are largely stable to mineral and fatty oils as well as petroleum.

Some physical characteristics are given in Table 11.

Properties and applications of cellulose acetates are primarily determined by the following:

1) viscosity of their solution
2) degree of esterification or the amount of bound acetic acid

*Viscosity*as an indicator for the degree of polymerization influences to a great extent the mechanical properties of the resulting fibers, films, or plastic masses, as well as their workability.

The *degree of esterification* primarily determines the solubility and compatibility with softeners, resin, varnish, etc., and ultimately also influences the mechanical properties.

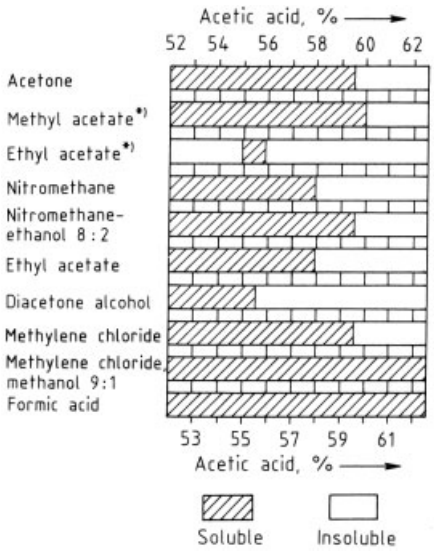

Figure 11. Solubility of cellulose acetate in various solvents (abridged according to [99])
∗ Technical grade

The wide span of solubility properties of hydrolyzed cellulose acetate is shown in Table 12.

The compatibility of the plasticizer and the solubility in polar solvents increase with decreasing acetic acid content while the solubility in nonpolar solvents decreases. Moreover, a correlation between the incompatibility with nonsolvents such as water, alcohol, benzene, or toluene and a decreasing degree of esterification exists. Furthermore, a number of solvent combinations are known which are able to dissolve the cellulose acetate although each of the components is a nonsolvent.

Figure 11 shows a selection of solvents for the industrially interesting range of esterification (52 – 62 % bound acetic acid). Detailed information on solubility and softener selection can be found in the literature and the manufacturers' information brochures.

2.1.6. Analysis and Quality Control

The *viscosity* is determined in practice by the usual methods. Along with the relative viscosity, the measurement of 15 – 20 % solutions according to the ball drop method corresponding to ASTM 871-56 has been generally accepted.

The *acetic acid content* is generally determined by saponification of the ester with 0.5 N-potassium hydroxide and back-titration of the excess. The free acids from mixed esters are separated from the residue after saponification by means of distillation or extraction and by extraction or azeotropic distillation isolated and titrated. Gas-chromatographic methods have found increasing popularity. A comprehensive presentation can be found in [100], [101].

Determination of the *free hydroxyl groups* in pure cellulose acetate is not necessary, since a precise analysis of bound acetic acid is possible. It is primarily used to characterize the mixed esters and can be carried out according to various methods. Complete esterification with acetic acid anhydride in pyridine and back-titration of the excess is a proven method [102]. The value is mostly given as a percentage of the hydroxyl or as the hydroxyl value (mg of KOH/g).

Additional quality control methods for unprocessed cellulose esters are the following: determination of temperature stability by heating to 220 – 240 °C and evaluation of discoloration and melting behavior, the determination of free acid as an indicator for the efficacy of the rinsing process, and determination of the ash content as well as clarity, color, and filterability of the solution.

Table 12. Solubility of cellulose acetate at various degrees of esterification [91]

Degree of esterification	Bound acetic acid	Chloroform	Acetone	2-Methoxyethanol	Water
2.8 – 3.0	60 – 62.5 %	soluble			
2.2 – 2.7	51 – 59 %		soluble		
1.2 – 1.8	31 – 45 %			soluble	
0.6 – 0.9	18 – 26 %				soluble
<0.6	<18 %				

2.2. Cellulose Mixed Esters

Apart from the long-established cellulose acetate, only cellulose mixed esters of acetic and propionic acid, cellulose acetate propionate [*9004-39-1*], or acetic and butyric acid, cellulose acetate butyrate [*9004-36-8*], have attained any notable importance.

Mention should also be made of cellulose acetate phthalate [*9004-38-0*], which is used in certain special fields of application.

Pure cellulose propionates [*9004-48-2*] and cellulose butyrates [*9015-12-7*] are difficult to produce and – like formates, palmitates and stearates, and esters of unsaturated acids and dicarboxylic acids – have attained no industrial importance [4].

2.2.1. Production

As far as the chemistry of the esterification reaction and the subsequent partial saponification are concerned, the basic description given in the chapter on cellulose acetate is also valid for mixed acids to a large extent.

The raw materials used are the same as for cellulose acetate, i.e., cotton linters or cellulose produced by special processes.

Pretreatment of the cellulose raw materials is similar to that used in the production of cellulose acetate. In practice, esterification takes place only in a homogeneous system and not, as is sometimes the case with cellulose acetate, in a heterogeneous system. The esterification mix consists of a mixture of anhydrides of acetic acid and propionic acid or of acetic acid and butyric acid. The reactivity of the aliphatic fatty acids decreases very rapidly as the chain-length increases.

Mixed esters consisting of propionic acid and butyric acid or of acetic acid, propionic acid, and butyric acid are not produced on an industrial scale.

2.2.2. Composition

The properties of the mixed esters are determined not only by their viscosity, but also in particular by the ratio of the two bound acids and by the content of free hydroxyl groups.

At present, cellulose acetate propionate and cellulose acetate butyrate flake is only produced by two manufacturers: Bayer AG, Leverkusen, in the Federal Republic of Germany, and the Eastman Kodak Co., Kingsport, Tennessee, in the United States.

Whereas pure cellulose acetates are clearly characterized by their viscosity and content of bound acetic acid, with mixed esters, data on the individual acids and possibly the free hydroxyl groups are also required.

2.2.3. Properties

As the degree of hydrolysis changes, the properties of cellulose mixed esters vary over a wide range from pure acetates to pure butyrates, with the propionates occupying a property-profile position between the cellulose mixed esters and pure acetates.

In the case of cellulose acetate butyrate, for example, if one considers the mixing range from pure cellulose acetate through the various mixing ratios to pure cellulose butyrate (the degree of esterification is adjusted in such a way that the esters in pure cellulose acetate contain 50 – 60 % acetic acid, and those in pure butyrate 60 – 70 % butyric acid), then the density varies from ca. 1.32 (cellulose acetate) to 1.16 (cellulose butyrate). The melting point is between ca. 300 °C (cellulose acetate) and 160 °C (cellulose butyrate), while the water absorption at 90 % relative humidity varies from ca. 12 % (cellulose acetate) to 1.5 % (cellulose butyrate).

Solubility in various solvents (like acetone, methyl ethyl ketone, methyl isobutyl ketone, and various phthalate plasticizers) also varies over a wide range [104].

Typical examples of cellulose acetate propionates, whose acetic acid and propionic acid levels are practically identical, are given in Table 13.

With cellulose acetate propionates, the acetic acid content varies from ca. 3 to 8 %, and the propionic acid content from ca. 55 to 62 %. With cellulose acetate butyrates, acetic acid contents of 19 – 23 % and butyric acid contents between 43 and 47 % are normal.

There are many more grades of cellulose acetate butyrate available than of cellulose acetate propionate.

Examples of cellulose butyrate grades are given in Table 14.

2.2.4. Other Organic Cellulose Mixed Esters

Cellulose acetate phthalates are cellulose mixed esters of minor industrial importance. They are produced from hydrolyzed cellulose acetate by reaction with an excess of phthalic anhydride in acetone or dioxane [106].

This produces esters of phthalic acid with a free carboxyl group. The products are used on a small scale as water- or alkali-soluble textile auxiliaries, tablet coatings, and antistatic agents in film coating.

2.2.5. Uses

The wide scope of variation for cellulose esters has led to the development of special grades for different fields of application.

Mixed esters based on cellulose acetate butyrate and cellulose acetate propionate are chiefly used in the production of molding plastics (see Section 2.4). The applications described in the following are of lesser importance.

Films. Triacetate is more widely used than cellulose acetate butyrate as an electrical insulating film, which is mainly produced by casting. Cellulose acetate butyrates with an acetic acid/butyric acid ratio between 2 : 1 and 1 : 1 are preferred.

The importance of cellulose mixed esters as film substrates in the photographic industry has greatly diminished as a result of the increasing use of polyester film.

Surface coatings. Cellulose ester lacquers, with their excellent lightfastness, gloss, low combustibility, and good thermal stability, coupled with their indifference to hydrocarbons, oils, and greases, became very quickly established in numerous fields of application.

One significant step forward was the introduction of cellulose acetate butyrates and cellulose acetate propionates. Both are particularly characterized by lower water absorption and good compatibility with extenders and, in the case of the low-viscosity grades, also allow the production of very high-solid lacquers. Cellulose acetate butyrate has become particularly important as an extender for metal effect finishes [105].

2.3. Cellulose Acetate Fibers

Cellulose acetate is the most important cellulose ester. It is primarily used for textile yarn and cigarette filter tow. The cellulose acetate is usually dissolved in a suitable organic solvent and spun by dry spinning (→ Fibers, 3. General Production Technology). Secondary (2.5) acetate with an acetic acid content of 54 – 56 % is normally used, whereas only a small amount of cellulose triacetate is normally produced.

2.3.1. Properties

The viscosity and the filterability of the spinning solution (spinning dope) are particularly important in the production of cellulose acetate fibers. The spinning dope has a high viscosity, which depends on the degree of polymerization. The strength and stretch properties of the fibers also depend on the concentration and the degree of polymerization as well as on the distribution of the acetate groups along the cellulose chain.

Because the fibers are produced by extruding the spinning dope through minute spinnerette holes, insoluble particles must first be removed

Table 13. Characteristic data of cellulose acetate propionate (Cellit PR) [105]

	PR 900	PR 800	PR 500
Acetyl content, %	3.5	3.5	3.5
Propionyl content, %	45	45	45
Hydroxyl content, %	1.6	1.6	1.6
Viscosity (DIN 53 015), mPa · s	4700 – 7800	2200 – 3800	150 – 240
Melting range, °C	200 – 220	190 – 210	180 – 200

Table 14. Characteristic data of cellulose acetate butyrate (Cellit BP) [105]

	BP 300	BP 500	BP 700/25	BP 700/40	BP 900
Acetyl content, %	14	14	15	14	15
Butyryl content, %	37	37	37	37	37
Hydroxyl content, %	1.2	1.2	0.8	1.2	0.8
Viscosity (DIN 53015), mPa · s	30 – 60	150 – 240	750 – 1500	750 – 1500	5000 – 8000
Melting range, °C	160 – 180	170 – 190	170 – 190	170 – 190	180 – 200

from the spinning dope by filtration. These particles are primarily composed of very small, incompletely acetylated cellulose fibers or gels, which can obstruct the spinnerett holes.

Secondary acetate and triacetate fibers have similar physical properties (Table 15). Their densities are lower than that of viscose rayon fibers and equal to that of wool. For textile yarns, the fibers should be as free of color as possible.

Table 15. Physical properties of acetate fibers and tow [107]

	Secondary acetate	Triacetate
Strength, cN/dtex	1.0 – 1.5	1.0 – 1.5
Stretch, %	25 – 30	25 – 30
Density, g/cm^3	1.33	1.30
Moisture uptake, % (65 % relative humidity, 20 °C)	6 – 6.5	4 – 4.5
Water retention capability, %	25 – 28	16 – 17
Melting point, °C	225 – 250	decomposition at 310 – 315
DP	300	300

The chemical reactions of cellulose acetate are similar to those of organic esters. Cellulose acetate is hydrolyzed by strong acids and alkali; it is sensitive to strong oxidizing agents but not affected by hypochlorite or peroxide solutions.

Acetate fibers cannot be dyed under the same conditions as viscose rayon fibers because their swelling properties are different. Acetate fibers can only be dyed with water-disperse dyes at the boiling point of the medium usually in the presence of carriers (→ Textile Auxiliaries). The carriers promote fiber swelling and enhance dye uptake by the fibers. The dyeing process coupled with the textile spinning operation assures color fastness. Triacetate fibers have better wash-and-wear properties than secondary acetate because of better dimensional stability and higher crease resistance.

2.3.2. Raw Materials

Table 16. Typical properties of acetate wood pulps

Characteristic*	Sulfite softwood pulp (conifer)	Sulfate hard-wood pulp (deciduous)
R_{10}, %	95	96
R_{18}, %	97	98
Ash, %	0.08	0.08
Silica, %	0.001	0.003
Calcium, %	0.006	0.008
Pentosans, %	1.2	1.2
Moisture content, %	6.5	6.5
Apparent density, g/cm^3	0.45	0.5
DP	2300	1700

* R_{10} and R_{18} are residues in 10 or 18 % sodium hydroxide at 20 °C [108]

Wood pulp produced from various softwood (conifer) or hardwood (deciduous) species is the cellulose source for the production of cellulose acetate fibers. The wood pulps are produced by the sulfite pulping process with hot alkali extraction or by the prehyrolized sulfate (Kraft) process with cold caustic extraction (→ Paper and Pulp). The lignins and hemicelluloses are removed from the wood to give wood pulps with an α-cellulose content of over 96 % (Table 16). High-purity cotton linters are no longer used in

the production of cellulose acetate fibers for economic reasons.

For the production of high-quality cellulose acetate fibers the wood pulp must have good swelling properties for uniform accessibility of the cellulose to the catalyst and the acetylation agent and a uniform reactivity. In addition, it must produce a spinning solution without fibers and gels which can easily be filtered.

2.3.3. Production

The general points discussed in Section 2.3.1 for the production of cellulose acetate also apply here. The sulfuric acid catalyst initially forms the cellulose sulfate ester. The sulfate groups are then replaced by acetyl groups as the acetylation proceeds. The sulfate ester contents is further reduced in the hydrolysis stage. However, any sulfate ester groups remaining at the end of the hydrolysis stage must be neutralized with an appropriate stabilizer, e.g., magnesium salts [109], [110]. Any "free" sulfate ester groups will affect the stability of the acetate because under the influence of heat and humidity they splitt off as sulfuric acid and degrade the fiber [111].

For secondary acetate spinning, acetone is used as the solvent. For triacetate, the solvent is 90 % dichloromethane and 10 % methanol or acetic acid (wet-spinning process). The viscosity of the spinning solution with a cellulose acetate concentration of 20 – 30 % is between 300 and 500 Pa s at 45 – 55 °C. The spinning dope is filtered in one or more steps and is then deaerated in large vessels.

Dry spinning is used almost exclusively; wet spinning is occasionally used for triacetate only. The spinneretts for textile filament have between 20 and 100 holes and those for tow up to 1 000. The fibers are formed by evaporating the solvent with a countercurrent of air at 80 – 100 °C in a 4- to 6-m spinning column. The fibers are then stretched while still plastic to increase their strength. Melt spinning of cellulose acetate or triacetate has no commercial importance due to the limited heat stability at the melting point.

A core-skin structure is formed in triacetate fibers. The acetyl groups are distributed very regularly in cellulose triacetate compared to secondary acetate; therefore, crystallization occurs when triacetate fibers are heated at 180 – 200 °C (heat setting) [107], [111], [112]. This heat treatment, which enhances the wash-and-wear properties of triacetate textiles, requires several minutes at 180 °C or several seconds at 220 °C. Heating for shorter periods is not effective and longer heating periods lead to deterioration of the mechanical properties of the textile. Heat-setting reduces water retention to 10 % and water absorption to 2.5 %.

2.3.4. Uses

By blending and twisting of cellulose acetate or triacetate fibers with nylon or polyester a combination of properties is achieved that make them suitable for different end uses in linings. In this way the weaker physical properties of acetate fibers can be compensated for while maintaining the positive characteristics, for instance, the high moisture absorption and the silk-like softness.

Due to the unique hydrophobic – hydrophilic properties, semipermeable membranes made from cellulose acetate fibers have a remarkable potential in desalination (reverse osmosis) of water.

Cellulose acetate hollow fibers are also suitable for gas separation and hemodialysis [113].

For cellulose acetate is non-toxic, biodegradable and the raw material is a renewable natural polymer, it is expected to find application for other uses in the future.

2.3.5. Economic Aspects

Secondary acetate and triacetate fibers for textiles and filter cigarette tow accout for 80 % of all cellulose ester production. The balance is used for plastics and film. Secondary acetate and triacetate textile fibers have a small share (about 1 %) of all textile fiber production.

In the late 1990s synthetic fiber production continued to expand (Table 17), whereas the acetate production was stable at about 850 000 t/a. The acetate fiber production decreased slightly compensated by a slight increase in filter tow production. The five largest manufactures of filter tow are Hoechst-Celanese and Eastman Chemicals in the United States, Rhodia Acetow

in Germany, Daicel in Japan, and Courtaulds in the United Kingdom.

Table 17. Worldwide production of textile fibers (1000 t) [114]

Fiber	1993	1995
Man-made fibers ∗	19781	21741
Synthetics	16652	18471
Cellulosics	3129	3270
Cotton	18494	18602
Wool	1687	1767
Silk	68	92

∗ Excl. polyolefin fibers, textile glass fibers, and acetate cigarette filter tow.

2.4. Plastic Molding Compounds from Cellulose Esters

In the category "plastics made from natural materials," thermoplastics based on cellulose esters or cellulose mixed esters are still the most important [115].

As early as 1920, A. Eichengrün developed thermoplastic cellulose ester molding compounds as a spraying and molding powder. Cellulose acetate and mixed esters are used in injection molding and extrusion; mixed esters are also used for fluidized-bed dip coating and rotational molding.

The use of inorganic cellulose esters (see Section 1.2.7) is continually decreasing in the plastics sector because of their high flammability.

2.4.1. Physical Properties

Like cellulose acetate, cellulose mixed esters can be plasticized at elevated temperatures by using plasticizers. This results in a large number of grades with property combinations found in no other type of thermoplastic.

Thermoplastic cellulose ester molding compounds are generally characterized by good transparency, high mechanical strength, and toughness; one particularly noteworthy feature is that the material reacts to mechanical stresses by exhibiting cold flow, so that the otherwise problematic insert molding of metal parts presents no problems and there is no risk of stress cracking. Light-stable, transparent material is available in a wide range of transparent, translucent, and opaque color shades. High surface gloss coupled with antistatic properties (i.e., electrical charges disperse rapidly and no annoying dust patterns form) ensures that moldings retain their attractive appearance for years. High surface elasticity ensures a good "natural feel" and imparts a "repolishing" effect to the material: this means that scratches disappear as the object is being used. The relatively low modulus of elasticity gives excellent damping of vibrations, so that the acoustic behavior is not affected by annoying resonance or ambient noise.

The individual cellulose esters generally differ in their mechanical properties and in their compatibility with plasticizers. As a rule, cellulose mixed esters contain higher-boiling plasticizers in amounts ranging from 3 to 25 % whereas cellulose acetate contains 15 – 35 % plasticizers. Heat distortion temperature increases as the plasticizer content is reduced. Mixed esters absorb considerably less water than cellulose acetates, with the result that parts produced from mixed esters retain their dimensional stability even in humid climates. Finally, cellulose acetate butyrates and (with certain restrictions) cellulose acetate propionates can also be treated with UV inhibitors to ensure serviceability of the moldings even during years of outdoor exposure [116]. In principle, it is possible to reinforce thermoplastic cellulose ester molding compounds with glass fibers [117].

Figures 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22 show the physical properties as a function of the plasticizer content; the property levels may vary by as much as 15 – 20 % in either direction, depending on the type of plasticizer and the relative viscosity of the cellulose ester. In all of the diagrams and tables featured here, the abbreviations used are as follows:

CA = Cellulose acetate molding compound (acetic acid content > 55 %)
CP = Cellulose acetate propionate molding compound
CP∗ = Polymer-modified cellulose acetate propionate molding compound
CAB = Cellulose acetate butyrate molding compound
CAB∗ = Polymer-modified cellulose acetate butyrate molding compound

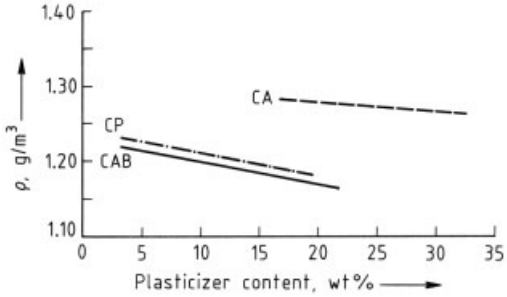

Figure 12. Density of cellulose acetate (CA), cellulose acetate propionate (CP), and cellulose acetate butyrate (CAB) as a function of the plasticizer content (determined in accordance with DIN 53479 or ISO/R 1183)

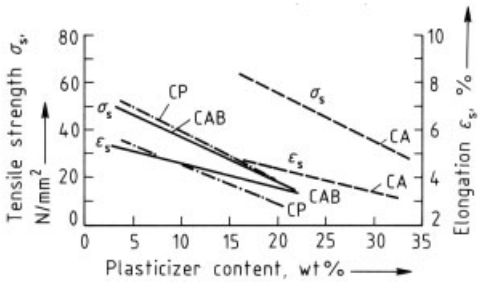

Figure 13. Tensile strength at yield σ_s and elongation ε_s of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53455 or ISO/R 527; specimen no. 3, rate of deformation 25 mm/min)

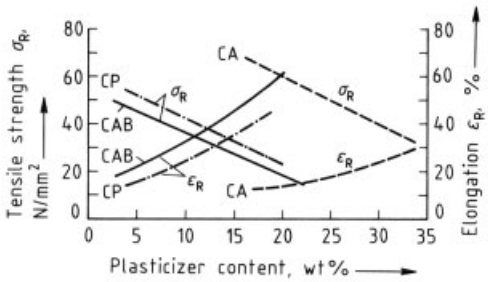

Figure 14. Tensile strength at break σ_R and elongation ε_R of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53455 or ISO/R 527; specimen no. 3, rate of deformation 25 mm/min)

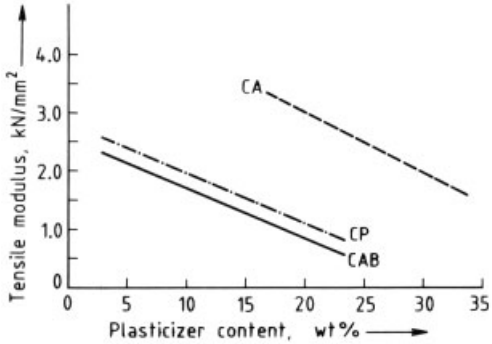

Figure 15. Tensile modulus of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53455 or ISO/R 527)

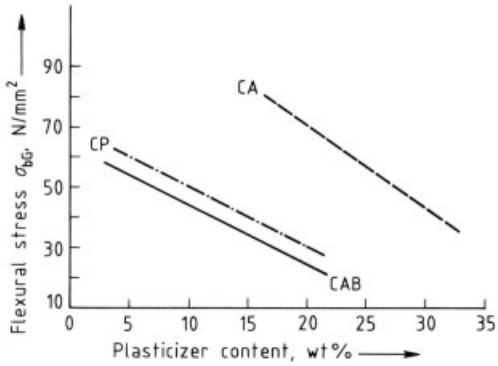

Figure 16. Flexural stress at a given strain σ_{bG} of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53452 or ISO/R 178; test specimen $4 \times 10 \times 80$ mm, rate of deformation 2 mm/min)

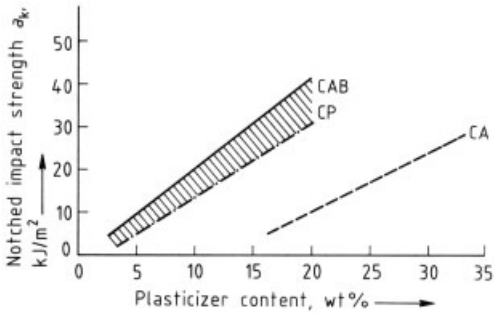

Figure 17. Notched impact strength a_k of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53453 or ISO/R 179; specimen no. 2)

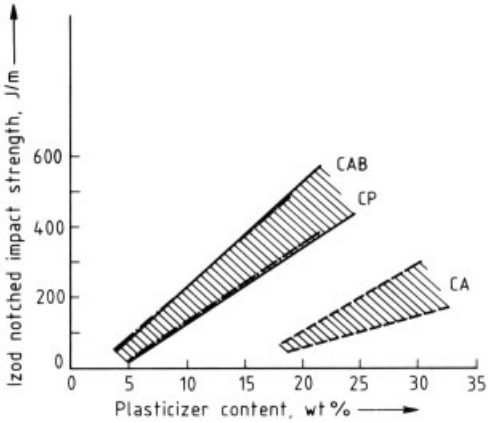

Figure 18. Izod notched impact strength of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with ASTM D 256, Method A, or ISO/R 180; test specimen 63.5×12.7×3.2 mm)

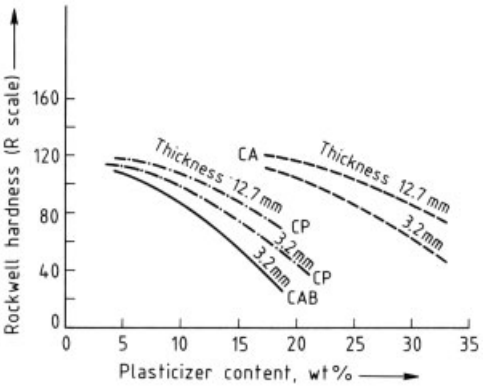

Figure 19. Rockwell hardness (R scale) of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with ASTM D 785)

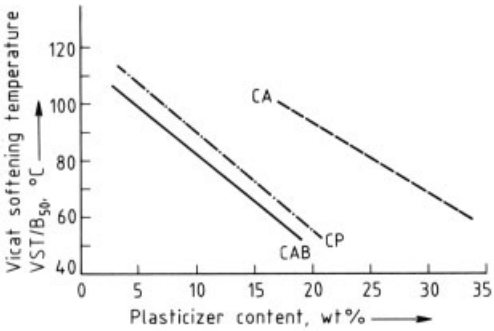

Figure 20. Vicat softening temperature VST/B_{50} of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53460/B or ISO/R 306; sheet 10×10×4 mm)

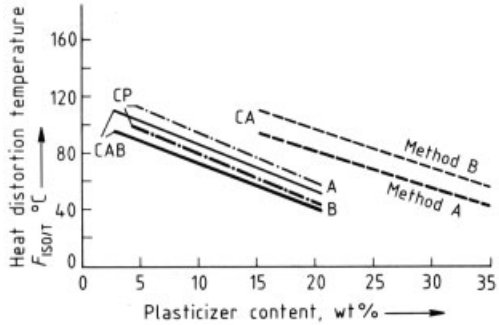

Figure 21. Heat distortion temperature F_{ISO} of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with ASTM D 648, ISO/R 75, or DIN 53461; test specimen 12.7×12.7×120 mm)

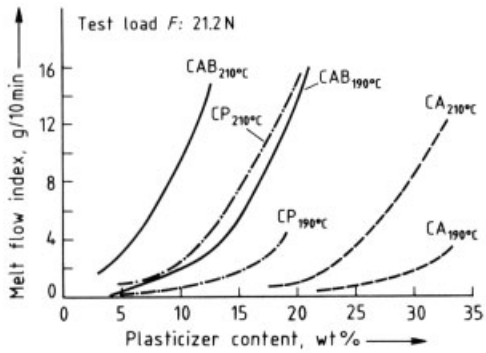

Figure 22. Melt flow index of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content (determined in accordance with DIN 53735 or ISO/R 1133)

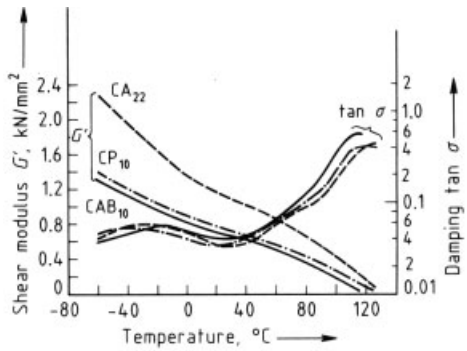

Figure 23. Shear modulus G' and damping tan σ of cellulose acetate_{22}, cellulose acetate propionate_{10}, and cellulose $\text{acetate butyrate}_{10}$ (determined in accordance with DIN 53445 or ISO/R 537)

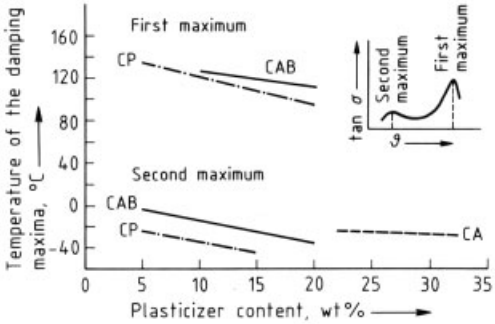

Figure 24. Temperature of the damping maxima of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate as a function of the plasticizer content

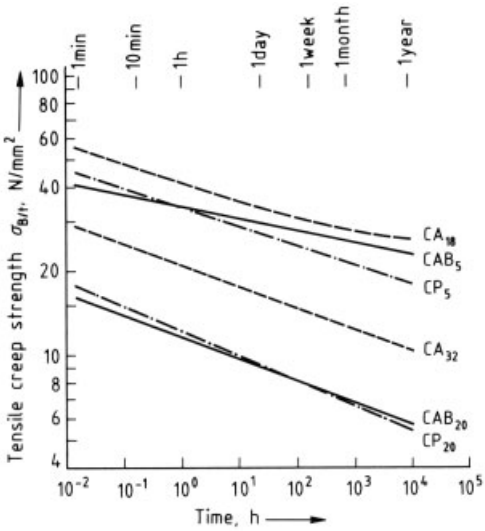

Figure 25. Tensile creep strength $\sigma_{\mathrm{B/t}}$ of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate (determined in accordance with DIN 53444 or ISO/R 899; test specimen no. 3)

The indices (e.g., CAB_{10}) give the plasticizer content in percent by weight.

Table 18 shows the electrical properties of medium-hardness cellulose ester molding compounds; their shear moduli and damping properties are given in Figure 23. Figure 24 shows the position of the damping maxima as a function of the plasticizer content.

Long-term properties derived from the tensile creep test are shown in Figures 25, 26, 27, 28, 29. Time-to-failure curves, modulus of creep curves and isochronous stress – strain curves of slightly and highly plasticized grades of cellulose acetates, cellulose acetate propionates, and cellulose acetate butyrates are given here.

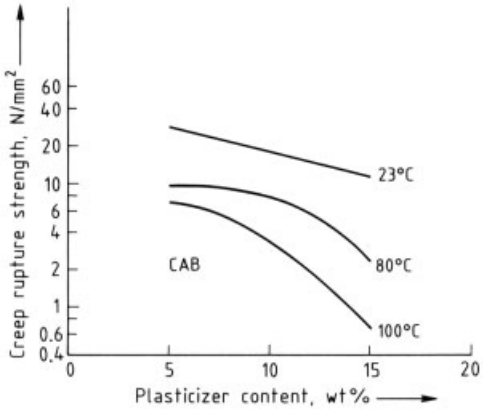

Figure 26. Creep rupture strength $\sigma_{\mathrm{B10_3}}$ of cellulose acetate butyrate at 23 °C, 80 °C, and 100 °C as a function of the plasticizer content (determined in accordance with DIN 53444 or ISO/R 899)

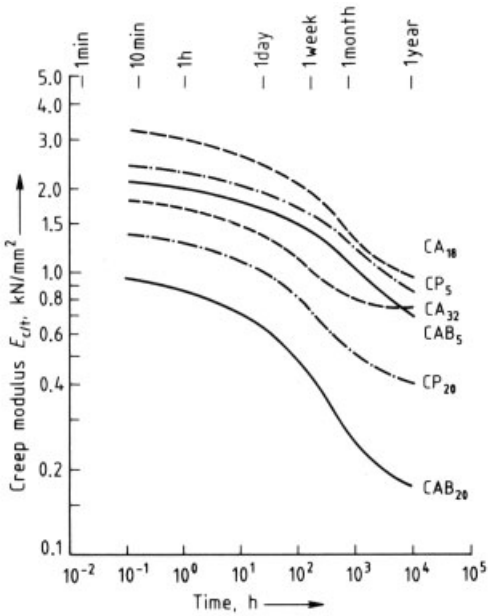

Figure 27. Creep modulus $E_{\mathrm{c/t}}$ of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate (determined in accordance with DIN 53444 or ISO/R 899)

Figure 30 shows results of the dynamic fatigue test in the tensile pulsating range on a medium-plasticity cellulose acetate, a slightly plasticized cellulose acetate propionate, and a medium-plasticity cellulose acetate butyrate.

Figure 31 shows results of the alternating bending test on slightly and highly plasticized cellulose acetate, cellulose acetate propionate,

Table 18. Electrical properties of organic cellulose ester molding compounds

Type of test	Unit	Test specification	Specimen	Cellulose acetate [130]	Cellulose acetate propionate [119]	Cellulose acetate butyrate [119]
Dielectric strength Ed (50 Hz, 0.5 kV/s)						
dry		VDE 0303	Circular	315	355	350
4 days at 80 % rel. humidity	kV/cm	Pt. 2,	discs 95 mm Ø	290	330	330
24 h water immersion		DIN 53481	×1 mm	280	330	330
Surface resistance R_0						
dry		VDE 0303		8×10^{13}	2×10^{14}	9×10^{13}
4 days at 80 % rel. humidity	Ω	Pt. 3,	150×15×4 mm	3×10^{12}	1×10^{13}	9×10^{12}
24 h water immersion		DIN 53482		4×10^{11}	5×10^{12}	9×10^{12}
Insulation resistance R_a						
dry		VDE 0303		5×10^{15}	5×10^{15}	5×10^{15}
4 days at 80 % rel. humidity	Ω	Pt. 3,	150×5×4 mm	1×10^{13}	6×10^{13}	5×10^{13}
24 h water immersion		DIN 53482		7×10^{11}	2×10^{13}	2×10^{13}
Volume resistivity ϱ_D						
dry	Ω/cm	VDE 0303	Circular	2×10^{15}	1×10^{16}	4×10^{15}
4 days at 80 % rel. humidity		Pt. 3,	discs 95 mm Ø	2×10^{12}	5×10^{13}	6×10^{13}
24 h water immersion		DIN 53482	×1 mm	2×10^{11}	1×10^{13}	2×10^{13}
Relative permittivity ε_r, dry						
at 50 Hz		VDE 0303	Circular	5.1	4.1	4.0
at 800 Hz		Pt. 4,	discs 95 and	4.0	3.9	3.8
at 1 MHz		DIN 53483	80 mm Ø×1 mm	4.1	3.6	3.4
Dissipation factor tan δ, dry						
at 50 Hz		VDE 0303	Circular	0.009	0.005	0.006
at 800 Hz		Pt. 4	discs 95 and	0.019	0.011	0.012
at 1 MHz		DIN 53483	80 mm Ø×1 mm	0.050	0.026	0.028
Tracking resistance		VDE 030				
KB method		Pt. 1/9.64				
Test solution A		DIN 53480/6	20×15 mm	>600	>600	>600

and cellulose acetate butyrate molding compounds.

2.4.2. Polymer-Modified Cellulose Mixed Esters

In 1977, cellulose acetate butyrate molding compounds modified with ethylene – vinyl acetate without monomolecular plasticizers were introduced for the first time [118]. They have since become firmly established, particularly in the extrusion sector (automotive decorative trim, etc.). As far as the property pattern was concerned, the decisive factor was the matching of the ethylene – vinyl acetate component to the cellulose acetate butyrate used.

A polymer modification of cellulose acetate propionate without monomolecular plasticizers has also been successfully carried out [119]. In this case, however, a modifier of complex structure, based on an ethylene – vinyl acetate graft polymer, is necessary [120].

The main advantages of these polymer-modified cellulose mixed ester molding compounds over the former plasticized systems are that the existing combination of characteristic properties is retained, while the values for heat distortion temperature, creep behavior, and stiffness are considerably improved.

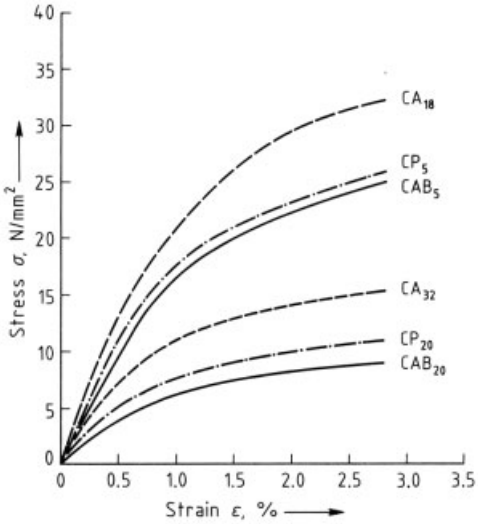

Figure 28. Isochronous stress – strain curves of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate for 1 h (determined in accordance with DIN 53444 or ISO/R 899)

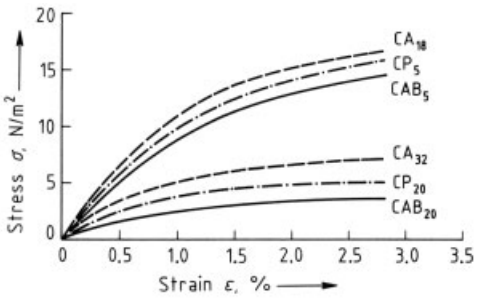

Figure 29. Isochronous stress – strain curves of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate for 1000 h (determined in accordance with DIN 53444 or ISO/R 899)

Another feature of these molding compounds is their freedom from plasticizer migration.

The polymer modifiers that are incorporated also delay the occurrence of crazing during long-term outdoor exposure, a fact which significantly increases the service life of, for example, external automotive decorative trim made from cellulose acetate butyrate molding compounds [118].

Some striking examples of the different properties of monomolecular-plasticized and polymer-modified cellulose acetate propionate and cellulose acetate butyrate are shown in Figures 27, 28, and 31.

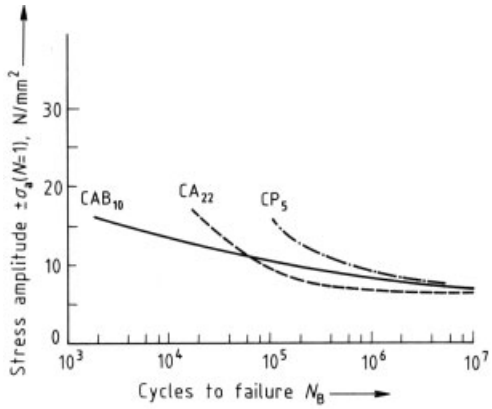

Figure 30. Dynamic fatigue test in the range of pulsating tensile stresses (number of load cycles) of cellulose acetate$_{22}$, cellulose acetate propionate$_5$, and cellulose acetate butyrate$_{10}$ (determined in accordance with DIN 50100; stress amplitude $\pm\sigma_a$ ($N = 1$) means stress amplitude under initial loading)

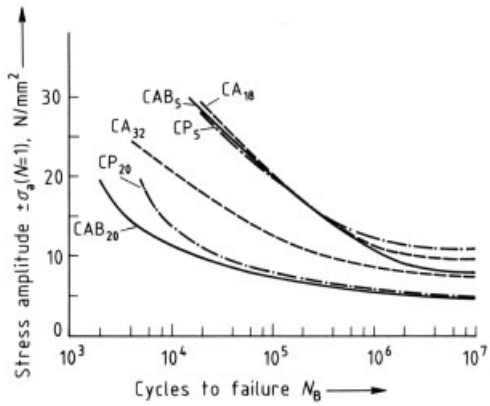

Figure 31. Dynamic fatigue test in the range of alternating flexural stresses (number of load cycles) of cellulose acetate, cellulose acetate propionate, and cellulose acetate butyrate (determined in accordance with DIN 50100; stress amplitude $\pm\sigma_a$ ($N = 1$) means stress amplitude under initial loading)

Polymer modification has a similar effect on the tensile test, shear modulus, creep behavior, and hardness [119].

2.4.3. Chemical Properties

Thermoplastic cellulose ester molding compounds are resistant to white spirits, oil, and grease. Table 19 gives guide values for resistance to a range of substances, but thorough practical tests are recommended in each case.

Table 19. Typical values∗ for the chemical resistance of organic cellulose ester molding compounds

Solvent	Cellulose acetate (< 55 % acetic acid)	Cellulose acetate (> 55 % acetic acid)	Cellulose acetate propionate	Cellulose acetate butyrate
Water	+	+	+	+
Alcohols	− −	− −	− −	− −
Ethyl acetate	− −	0	0	0
Methylene chloride	− −	0	0	0
Acetone	0	0	0	0
Carbon tetrachloride	+	+	+ −	+ −
Trichloroethylene	+	+ −	− −	− −
Perchloroethylene	+	+	+ −	+ −
Benzene	+	+ −	− −	− −
Xylene	+	+	− −	− −
Petroleum spirit	+	+	+	+
Motor fuel mixture (high octane)	+	+	+ −	+ −
Mineral oil (paraffin)	+	+	+	+
Linseed oil	+	+	+	+
Turpentine oil	+	+	+ −	+ −
Lavender oil	+	+	− −	− −
Ether	+	+	+ −	+ −
Formalin	− −	− −	+ −	+ −
2-Chlorophenol	0	0	0	0
Sulfuric acid, conc.	−	−	−	−
Sulfuric acid, 10 %	+ −	+ −	+ −	+
Hydrochloric acid, conc.	−	−	−	−
Hydrochloric acid, 10 %	−	−	−	−
Nitric acid, conc.	−	−	−	−
Nitric acid, 10 %	−	−	−	−
Caustic potash solution, 50 %	−	−	−	−
Caustic potash solution, 10 %	−	−	+ −	+ −

∗ Key to symbols: + = resistant; + − = resistant, but swells; − = not resistant; − − = not resistant, swells; 0 = soluble.

2.4.4. Raw Materials

The following cellulose esters are used for the production of cellulose ester molding compounds:

Cellulose acetate propionate molding compounds,
55 – 62 % propionic acid content
3 – 8 % acetic acid content
Cellulose acetate butyrate molding compounds,
43 – 47 % butyric acid content
19 – 23 % acetic acid content
Cellulose acetate molding compounds,
51.6 – 56.3 % acetic acid content
51.5 – 53.5 % acetic acid content (for block acetate only)

Of the large number of plasticizers that are compatible with cellulose esters [6], the following have acquired industrial significance, either alone or in combination with one another:

For cellulose acetate propionates and cellulose acetate butyrates:
di-2-ethylhexyl phthalate, dibutyl adipate, di-2-ethylhexyl adipate, dibutyl azelate and dibutyl sebacate, dioctyl azelate, dioctyl sebacate, palmitates, stearates, etc.
For cellulose acetates:
dimethyl, diethyl, dibutyl, di-2-ethylhexyl, and di-2-methoxyethyl phthalate; triphenyl and trichloroethyl phosphate.

Ethylene – vinyl acetate copolymers have proven to be particularly suitable as polymer modifiers for cellulose acetate butyrate molding compounds, while graft polymers based on ethylene – vinyl acetate are preferred for cellulose acetate propionate molding compounds.

The *stabilizers* and *antioxidants* used for cellulose ester molding compounds include: alkali salts and alkaline-earth salts of sulfuric, acetic, and carbonic acid, tartaric acid, oxalic acid, citric acid, higher molecular mass epoxides, and phenolic antioxidants. In special cases these stabilizers and antioxidants can be complemented by others [121].

From the range of *ultraviolet absorbers* available, various benzophenones, benzotriazoles,

salicylates, and benzoates are recommended for organic cellulose ester molding compounds [122].

Processing auxiliaries for cellulose ester molding compounds include zinc stearate, butyl stearate, and paraffin oil.

Numerous combinations of *dyes* can be used for coloring cellulose ester molding compounds [123]. The following groups of dyes have proven successful in practice: alkaline, acid, and substantive dyes (provided they are sufficiently soluble in the solvent); Zapon, Sudan, and Ceres dyes (provided they are sufficiently resistant to sublimation); and organic and inorganic pigments.

2.4.5. Production

Like cellulose acetates, cellulose acetate propionates and cellulose acetate butyrates are also thoroughly mixed at room temperature with plasticizers, stabilizers, antioxidants, dyes, and sometimes ultraviolet absorbers and processing aids. Plastification and homogenization are carried out at higher temperatures (between 150 and 210 °C, depending on the type and the degree of plasticization) in single- or twin-screw kneaders or roll mills. Depending on the type of equipment used, this results in granules in the form of pellets (bulk density 500 – 620 g/L) or cubes (bulk density 400 – 470 g/L).

There are no wastewater or waste gas problems associated with the production of thermoplastic cellulose ester molding compounds. The inevitable plasticizer vapors that occur during processing should be removed by exhaust ventilation.

2.4.6. Trade Names

The most important trade names of thermoplastic cellulose mixed ester molding compounds are as follows:

Cellulose acetate propionates: Cellidor CP (Bayer AG, Leverkusen, Federal Republic of Germany), Tenite Propionate (Eastman Chemical Products, Inc., Kingsport, United States).

Cellulose acetate butyrates: Cellidor B (Bayer AG, Leverkusen, Federal Republic of Germany), Tenite Butyrate (Eastman Chemical Products, Inc., Kingsport, United States).

Cellulose acetate: Acety (Daicel, Osaka, Japan), Cellidor S (Bayer AG, Leverkusen, FRG), Dexel (Courtaulds Chem. and Plastics, GB), Saxetat (VEB Eilenburg, GDR), Setilithe (Tubize Plastics, Belgium), Tenite Acetate (Eastman Chemical Products, Inc., Kingsport, United States).

2.4.7. Quality Requirements and Quality Testing

As has already been stated, the mechanical properties of thermoplastic cellulose ester molding compounds are dependent on the chain-length of the molecules, on the plasticizer or polymer modifier, the combination of plasticizers, and the content of plasticizer or polymer modifier.

With respect to cellulose acetate, determination of the viscosity and the viscosity ratio in a dilute solution [124], determination of the insoluble constituents [125], viscosity loss during molding [126], light absorption before and after heating [127], and the determination of constituents which are extractable with ethyl ether [128] are all standardized. The methods described for cellulose acetate are similarly applicable to cellulose mixed esters.

The manufacturers also carry out numerous in-house tests during the course of their quality control programs. These include determination of mechanical data, testing of purity, checking thermal stability, colorimetry, determination of flow properties (melt index, Brabender, extrusiometer), etc.

In the Federal Republic of Germany, cellulose ester molding compounds (CA, CP, and CAB) are standardized in accordance with DIN 7742, Parts 1 and 2, and in the United States in accordance with ASTM D 706 (cellulose acetate), D 707 (cellulose acetate butyrate), and D 1562 (cellulose acetate propionate).

2.4.8. Storage and Transportation

The transportation of thermoplastic cellulose ester molding compounds is not governed by the GGVS/ADR, GGVE/RID, GGVSee/IMDG code or DGR/ICAO regulations for the transportation of hazardous goods. The storage of

thermoplastic cellulose ester molding compounds presents no problems. Even after 10 years in storage, no changes in composition have been detected. It is, however, recommended that thermoplastic cellulose ester molding compounds should be predried in accordance with the particular manufacturer's guidelines before processing.

2.4.9. Uses [129]

In the field of injection-molded eyeglass frames, cellulose acetate propionates have become increasingly important. This is due mainly to the fact that their dimensional stability is better than that of cellulose acetate, as a result of lower moisture absorption and greater stiffness and dimensional stability under heat [130].

Cellulose acetate propionate is also used for high-quality frames for sunglasses, protective goggles for industry, and sports goggles [130].

Cellulose acetate still dominates as sheet material for making ophthalmic frames. For this purpose, uni- or multicolored plates are either cast or made by extrusion.

Due to its high transparency, good impact resistance, and low level of light scattering, cellulose acetate propionate has become increasingly popular as a glazing for visors (for skiers, drivers, and workers in industry) and for sunglasses and sports goggles. These applications have been made possible by surface saponification of the plastic, which ensures permanent antifogging properties (the surface has very good wetting properties and excellent water absorption) [130].

Cellulose acetate propionates with special infrared/ultraviolet-absorbing characteristics have become more important for welding goggles and certain types of sunglasses [130].

Where greater demands are placed on impact resistance, cellulose mixed esters as well as cellulose acetate are used (e.g., for the production of tool handles, hammer heads, and covers for handles of pliers, wrenches, etc.). Particularly important factors in this field of application are toughness, transparency, lightfastness, the "natural feel," the repolishing effect, and the absence of stress cracking. No problems occur when metal parts are insertion molded (the blades can even be driven "cold" into the handles), which brings obvious economic advantages.

With their minimal plasticizer migration, cellulose mixed esters are preferred over cellulose acetate for the packaging of toiletries. Other reasons for their use in this sector include the brilliance and depth of color, as well as the ability to produce special color effects.

Resistance to stress cracking and good toughness properties make cellulose mixed esters an excellent material for brushes, particularly toothbrushes, for which new designs require close spacing of the drill holes and a high bristle density. Cellulose mixed esters also satisfy the requirement for high tuft pull-out strength.

Cellulose acetate butyrate sheet is used for illuminated advertising signs, machine hoods, lamp covers, and dome lights. High light transmission, practically unrestricted choice of colors, antistatic characteristics, easy accurate processing, ease of joining (by simply gluing), good printing and coating properties, the absence of stress cracking, and finally, excellent mechanical strength are the main factors influencing the choice of this material. Cellulose acetate propionate is used in place of cellulose acetate for transparent, large-size seat shells.

Transparent, thin-section, and large-area hoods, lids, and covers with excellent toughness and good weathering resistance are injection-molded from specially modified cellulose acetate propionate.

Decorative trim made of cellulose acetate butyrate combined with aluminum foil [131] has been firmly established in industry for years. An aluminum foil is coated with the cellulose mixed ester and shaped in the crosshead die of an extruder [132]. With its practically unlimited scope for metal and wood effects, its elasticity, resistance to detergents, and simple fixing, this combination of materials has been used with great success in the automotive industry [133] in particular, as well as in the electrical, audio, and domestic appliance sectors [134].

Further applications for cellulose esters include lamp covers, toilet seats, writing and drawing instruments, tap handles, casings for disposable hypodermic syringes, transparent mouthwash spray attachments, handles, high-quality toys, instrument panel covers (glazing), and knife handle grips.

Economic Facts. With an apparently guaranteed supply of raw materials and a tremendous scope for variation of cellulose ester molding compounds and of their property combinations, this class of plastics should maintain its market significance in special areas of application for years to come.

2.4.10. Toxicology and Occupational Health

Cellulose acetate and cellulose propionate molding compounds comply with Recommendation XXVI of the Federal Health Authorities of the Federal Republic of Germany [135].

There are also various cellulose ester molding compounds on the market which satisfy the requirements of the U.S. Food and Drug Administration for plastics [136].

3. References

General References

1. V. Stannet: *Cellulose Acetate Plastics,* Temple Press Ltd., London 1950.
2. K. Thinius: *Analytische Chemie der Plaste,* Springer Verlag, Berlin-Göttingen-Heidelberg 1952.
3. *Kirk-Othmer,* vol. 3, pp. 357 ff.
4. E. Ott, H. M. Spurlin, M. W. Grafflin: *Cellulose and Cellulose Derivatives,* 2nd ed., **vol. V, part II,** Interscience Publ., New York-London 1954.
5. R. Houwink, A. J. Staverman: *Chemie und Technologie der Kunststoffe,* 4th ed., vol. II/2, Akademische Verlagsgesellschaft Geest & Porting KG, Leipzig 1963.
6. K. Thinius: *Chemie, Physik und Technologie der Weichmacher,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1963.
7. V. E. Yarsley, W. Flarell, P. S. Adamson, N. G. Perkins: *Cellulose Plastics,* Iliffe Book Ltd., London 1964.
8. *Encyclopedia of Polymer Science and Technology,* vol. 3, J. Wiley & Sons, New York 1965.
9. R. Vieweg, E. Becker: *Kunstoff-Handbuch,* vol. III: "Abgewandelte Naturstoffe," Hanser Verlag, München 1965.
10. P. B. Koslov, G. I. Braginski: *Chemistry and Technology of Polymers,* Iskustvo, Moscow 1965 (Russ.).
11. H. Temming, H. Grunert: *Temming-Linters,* 2nd ed., P. Temming, Glückstadt 1972.

Specific References

12. J. Barsha: "Inorganic Esters," in E. Ott, H. M. Spurlin, M. W. Grafflin (eds.): *Cellulose and Cellulose Derivatives,* 2nd ed., part II, Interscience Publ., New York-London 1954, pp. 713 – 762.
13. E. D. Klug: "Cellulose Derivatives," in *Kirk-Othmer,* 2nd ed., vol. 4, J. Wiley & Sons, New York-London-Sydney-Toronto 1964, pp. 616 – 652.
14. G. D. Hiatt, W. J. Rebell: "Esters," in N. M. Bikales, L. Segal (eds.): *Cellulose and Cellulose Derivatives,* part V, Wiley-Interscience, New York-London-Sydney-Toronto 1971, pp. 741 – 777.
15. J. Honeyman: *Recent Advances in the Chemistry of Cellulose and Starch,* Heywood & Comp., London 1959.
16. E. Sjöström: *Wood Chemistry, Fundamentals and Applications,* Academic Press, New York-London 1981.
17. D. Fengel, G. Wegener: *Wood – Chemistry, Ultrastructure, Reactions,* De Gruyter, Berlin-New York 1983.
18. K. Balser: "Derivate der Cellulose," in W. Burchard (ed.): *Polysaccharide,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1985, pp. 84 – 110.
19. L. C. Wadsworth, D. Daponte: "Cellulose Esters," in T. P. Nevell, S. H. Zeronian (eds.): *Cellulose Chemistry and its Applications,* Ellis Horwood Ltd., Chichester 1985, pp. 344 – 362. A. Revely: "A Review of Cellulose Derivatives and their Industrial Applications," in *Cellulose Chemistry and its Applications,* pp. 211 – 225.
20. Company Publication of Wolff Walsrode AG: *100 Jahre Collodiumwolle* (1979).
21. K. Fabel: *Nitrocellulose,* Enke Verlag, Stuttgart 1950.
22. F. D. Miles: *Cellulose Nitrate,* Oliver & Boyd, London 1955.
23. S. Rossin, *Mémorial des Poudres* **40** (1958) 457 – 471.
24. H. Temming, H. Grunert: *Temming-Linters,* 2nd ed. (engl.), Peter Temming AG, Glückstadt 1973.
25. Company Publication of Wolff Walsrode AG: *Walsroder Nitrocellulose,* no. 500/203/007 (1984).

26. J. Isler, D. Flegier: "The Self-Ignition Mechanism in Nitrocellulose," in J. F. Kennedy, G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Cellulose and its Derivatives,* Ellis Horwood, Chichester 1985, pp. 329 – 336.
27. A. H. K. Fowler, H. S. Munro: "Some Surface Aspects of the Thermal and X-Ray Induced Degradation of Cellulose Nitrates as Studied by ESCA," in [26] pp. 245 – 253.
28. Enclosure I of the International Rail Transport Regulations (RID), Appendix 1.
29. E. Berl, *Angew. Chem.* 1904, 982, 1018, 1074. E. Berl-Lunge: *Chemisch-technische Untersuchungsmethoden,* 8th ed., **vol. III,** Julius Springer, Berlin 1932, p. 1294.
30. P. Kassenbeck: "Faktoren, die das Nitrierverhalten von Baumwoll-Linters und Holzzellstoffen beeinflussen," report 3/83, ICT-Fraunhofer-Institut für Treib- und Explosivstoffe, Pfinztal-Berghausen 1983.
31. Hercules Powder, US 2776965, 1957; US 3063981, 1962.
32. Hercules Powder, US 2028080, 1936.
33. Hercules Powder, US 2950278, 1960.
34. Société Nationale des Poudres et Explosifs, FR 1394779, 1964; DE-OS 1246487, 1965; FR 1566688, 1968; DE-OS 1914673, 1969.
35. Société Nationale des Poudres et Explosifs, DE-OS 2813730, 1977.
36. Hercules Powder, US 2776944, 1957; US 2776964, 1957.
37. Wolff Walsrode, DE-OS 2727553, 1977; DE-OS 2727554, 1977.
38. Wasag Chemie, DE-OS 1771006, 1968.
39. Wolff & Co. Walsrode, DE-OS 1153663, 1961.
40. Wolff & Co. Walsrode, DE-OS 1203652, 1964.
41. Société Nationale des Poudres et Explosifs, DE-OS 2338852, 1973.
42. C. F. Bennett, T. E. Timell, *Sven. Papperstidn.* **58** (1955) 281 – 286.
43. D. A. J. Goring, T. E. Timell, *Tappi* **45** (1962) 454 – 460.
44. K. Thinius, W. Thümmler, *Makromol. Chem.* **99** (1966) 117 – 125.
45. M. Marx-Figini, *Makromol. Chem.* **50** (1961) 196 – 219.
46. J. W. Green in R. L. Whistler, J. W. Green, J. N. BeMiller (eds.): "Cellulose,"*Methods in Carbohydrate Chemistry,* **vol. III,**Academic Press, New York-London 1963, pp. 213 – 237.
47. W. J. Alexander, G. C. Gaul: "Cellulose Derivatives," in *Encyclopedia of Industrial Chemical Analysis,* vol. 9, J. Wiley & Sons, New York-London-Sydney-Toronto 1970, pp. 59 – 94.
48. ASTM D 301-72 (1983): Soluble Cellulose Nitrate.
49. ASTM D 1343-69 (1979): Viscosity of Cellulose Derivatives by Ball-Drop Method.
50. ASTM D 365-79: Soluble Nitrocellulose Base Solutions.
51. ASTM D 1720-79: Dilution Ratio in Cellulose Nitrate Solutions for Active Solvents, Hydrocarbons, Diluents and Cellulose Nitrates.
52. A. Kraus: *Handbuch der Nitrocelluloselacke,* Westliche Berliner Verlagsgesellschaft Heenemann KG, Berlin, vol. 1: 1955, vol. 2: 1963, vol. 3: 1961, vol. 4: 1966.
53. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* vols. 1 – 8, Heenemann GmbH, Stuttgart-Berlin 1971 – 1980.
54. Bayer, DE-OS 2853578, 1978.
55. Wolff Walsrode, DE-OS 3139840, 1981; DE-OS 3407932, 1984.
56. *Ullmann,* 4th ed., **9** (1975) 179 – 183.
57. Leaflet M 051 by the German Employers' Liability Insurance Association of the Chemical Industry: *Dangerous Chemical Substances,* Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg 1984.
58. *Verordnung über gefährliche Arbeitsstoffe* 11 Feb. 82, BGBl I, Carl Heymanns, Köln 1982, p. 144.
59. Leaflet M 014 by the German Employers' Liability Insurance Association of the Chemical Industry: *Nitric Acid, Nitrogen Oxides, Nitrous Gases,* Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg 1985.
60. Leaflet M 037 by the German Employers' Liability Insurance Association of the Chemical Industry: *Nitrocellulose,* Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg 1984.
61. E. v. Schwartz: *Handbuch der Feuer- und Explosionsgefahr,* 5th ed., Feuerschutz-Verlag P. L. Jung, München 1958, p. 303.
62. German Explosives Law, publ. 13 Sept. 1976, BGBl I, p. 2737; amended 10 Apr. 1981, BGBl I, Carl Heymanns, Köln 1981, p. 388.
63. H. Dorias: *Gefährliche Güter,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1984.
64. Z. A. Rogovin, L. S. Galbraich, W. Albrecht (eds.): *Die chemische Behandlung und Modifizierung der Cellulose,* Thieme Verlag, Stuttgart-New York 1983.
65. Hercules Powder Comp., US 2753337, 1953.
66. B. Philipp, W. Wagenknecht: *Cellul. Chem. Technol.* **17** (1983) 443.

67. R. G. Schweiger, *Tappi* **57** (1974) 86; *ACS Symp. Ser.* **77** (1978) 163; US 4138535, 1977.
68. R. G. Schweiger, *Carbohydr. Res.* **70** (1979) 185.
69. R. G. Schweiger, DE-OS 3025094, 1980.
70. Eastman Kodak, US 3075962, 1963; US 3075963, 1963; US 3075964, 1963.
71. Eastman Kodak, US 3068007, 1963.
72. W. D. Slowig, M. E. Rowley, *Text. Res. J.* **38** (1968) 879.
73. J. Pastyr, L. Kuniak, *Cellul. Chem. Technol.* **6** (1972) 249.
74. J. B. Lawton, G. O. Phillips, *Text. Res. J.* **45** (1975) 4.
75. K. Katsuura, T. Fujinami, *Kogyo Kagaku Zasshi* **71** (1968) 771.
76. Eastman Kodak, US 2759924, 1956.
77. Kimberly-Clark, US 3658790, 1970.
78. National Chemical & Manufacturing, US 1896725, 1933.
79. R. A. A. Muzzarelli, G. Marcotrigiano, C.-S. Liu, A. Frêche, *Anal. Chem.* **39** (1967) 1792.
80. L. S. Sletkina, A. J. Poljakov, Z. A. Rogovin, *Vysokomol. Soedin* **7** (1965) 199; *Faserforsch. Textiltech.* **2** (1965) 299.
81. D. A. Predvoditelev, M. S. Bakseeva; *Zh. Prikl. Khim.* (*Leningrad*) **45** (1972) 857; *Faserforsch. Textiltech.* **9** (1972) 361.
82. D. C. Johnson: "Solvents for Cellulose," in *Cellulose Chemistry and its Applications,* Ellis Horwood, Chichester 1985, pp. 181 – 201.
83. L. P. Clermont, F. Bender, *J. Polym. Sci. Part A-1* **10** (1972) 1669.
84. R. B. Hammer, A. F. Turbak, *ACS Symp. Ser.* **58** (1977) 40. H. L. Hergert, *Tappi* **61** (1978) 63. ITT Rayonier, US 4056675, 1977.
85. C. J. Malm, L. J. Tanghe, B. C. Laird, *J. Am. Chem. Soc.* **72** (1950) 2674.
86. C. J. Malm, L. J. Tanghe, B. C. Laird, G. D. Smith, *J. Am. Chem. Soc.* **75** (1953) 80.
87. J. Sakurada, *Kogyo Kagaku Zasshi* **35** (1932) B 123.
88. Du Pont, US 1990483, 1935.
89. C. J. Malm, L. J. Tanghe, B. C. Laird, *Ind. Eng. Chem.* **38** (1946) 77.
90. H. Temming, H. Grunert: *Temming Linters,* 2nd ed., P. Temming, Glückstadt 1972.
91. B. P. Rousse, *Encycl. Polym. Sci. Technol. 1964 – 1977,* **3.**
92. *Ullmann,* 3rd ed., **5**, 187 ff.
93. Bios Final Report No. 1600, Item No. 22, H. M. Stationary Office, London 1948.
94. L. Ledderer, DE 200916, 1905.
95. Boehringer, GB 363700, 1930.
96. Bios Final Report No. 1850, Item No. 21, H. M. Stationary Office, London 1948.
97. Cons. f. Elektrochem. Ind., DE 687065, 1933; US 2108829, 1934.
98. Company Publication of Bayer AG, Leverkusen.
99. Company Publication of Hercules Powder Comp., Wilmington, USA.
100. E. Yarsley, W. Flavell, P. S. Adamson, N. G. Perkins: *Cellulosic Plastics,* Iliffe Book Ltd., London 1964.
101. K. Thinius: *Analytische Chemie der Plaste,* Springer Verlag, Berlin-Göttingen-Heidelberg 1952.
102. C. J. Malm, L. B. Genung, R. F. Williams, *Ind. Eng. Chem. Anal. Act.* **14** (1942) 935.
103. H. Saechtling, W. Zebrowski: *Kunststoff-Taschenbuch,* 18th ed., C. Hanser, München 1971.
104. C. J. Malm, C. R. Pordyce, H. A. Tanner, *Ind. Eng. Chem.* **34**(1942)430. E. Ott, H. M. Spurlin, M. W. Grafflin: *Cellulose and Cellulose Derivatives,* vol. V, Part II, 2nd edn., Interscience Publ. Inc., New York-London 1954. W. Ballas, Internal communication, Bayer AG, Leverkusen.
105. Bayer AG: *Cellit, Cellit BP and PR,* Brochures nos. KL 44172 and KL 44170, Leverkusen 1982.
106. Eastman Kodak Comp.: Brochure on Cellulose Acetate Phthalate, Rochester.
107. J. Corbiere, *Faserforsch. Textiltech.* **22** (1971) 71.
108. Technical Association of Paper and Pulp Industry (TAPPI), DIN or ISO-standards.
109. Celanese Corp. of America, US 2597156, 1952 (M. E. Martin, L. G. Reed).
110. Eastman Kodak Co., US 2652340, 1953 (G. D. Hiatt, R. F. Willmans).
111. E. Heim, *Chemiefasern* **16** (1966) 618.
112. *Ullmann,* 4th ed., **9,** 213.
113. W. Pusch: "Membranes from Cellulose and Cellulose Derivatives," in *Wood and Cellulosics,* Ellis Horwood Limited, Chichester 1985, pp. 475 – 482.
114. *Man-Made Fiber Yearbook,* Fiber Organon, Washington D.C. 1996, p. 12.
115. W. Fischer, *Kunststoffe* **62** (1972) 653.
116. Eastman Chem. Prod.: Specifications, T 58-3289, Kingsport, Tennessee 1958.
117. W. Fischer, *Kunststoffe* **63** (1973) 292.
118. W. Fischer, C. Leuschke, H. P. Baasch, *Kunststoffe* **67** (1977) no. 6, 348 – 252.
119. C. Leuschke, M. Wandel, *Plastverarbeiter* **33** (1982) no. 9, 1095 – 1098.

120. Bayer, DE-OS 2951800, 1979; DE-OS 2951747, 1979; DE-OS 2951748, 1979. K. Thinius: *Chemie, Physik und Technologie der Weichmacher,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1963.
121. *Mod. Plast. Encycl.* 1970 – 1971, 840, 870.
122. *Mod. Plast. Encycl.* 1970 – 1971, 876.
123. *Mod. Plast. Encycl.* 1970 – 1971, 850.
124. ISO/R 1157-1970.
125. ISO/R 1598-1970.
126. ISO/R 1599-1970.
127. ISO/R 1600-1970.
128. ISO/R 1875-1971.
129. M. Wandel, C. Leuschke, *Kunststoffe* **74** (1984) no. 10, 589 – 592.
130. Bayer AG: *Cellidor für die optische Industrie,* Company Publication KU 40029, Leverkusen 1984.
131. J. Göller, H. Peters, W. Fischer, *Kunststoffe* **65** (1975) no. 6, 326 – 332.
132. Glas Lab, DE-AS 1226291, 1955 (A. Shanok, V. Shanok, J. Shanok).
133. Bayer AG: *Zierleisten und Profile in der Autoindustrie,* Company Publication KL 40025, Leverkusen 1979.
134. Bayer AG: *Zierleisten in der Elektro- und Möbelindustrie,* Company Publication KL 40011, Leverkusen 1975.
135. *Bundesgesundheitsblatt,* 19th ed. 1976, No. 6, Carl Heymanns Verlag, Köln-Berlin-Bonn-München.
136. Eastman Kodak Comp.: Brochures on Tenite Propionate and Tenite Butyrate, nos. MB-33 C and MB-35 E, Rochester 1981 and 1983.

Cellulose Ethers

Heiko Thielking, Wolff Cellulosics GmbH & Co. KG, Walsrode, Germany
Marc Schmidt, Wolff Cellulosics GmbH & Co. KG, Walsrode, Germany

1. Introduction 1
2. Properties 1
2.1. Swelling and Dissolving Behavior . 2
2.2. Solution State 2
2.3. Thermally Induced Coagulation . . 4
2.4. Degree of Substitution 4
2.5. Molar Mass Distribution 5
3. Production 5
4. Product Groups 7
4.1. Carboxymethyl Cellulose (CMC) . . 7
4.1.1. Applications / Market 7
4.1.2. Synthesis 8
4.1.3. Processes 8
4.1.4. Characterization 8
4.2. Methyl and Hydroxyalkyl Methyl Celluloses 10
4.2.1. Applications / Market 10
4.2.2. Synthesis 10
4.2.3. Processes 11
4.2.4. Characterization 11
4.3. Hydroxyethyl Cellulose (HEC) . . . 13
4.3.1. Applications / Market 13
4.3.2. Synthesis 13
4.3.3. Processes 14
4.3.4. Characterization 14
4.4. Hydroxypropyl Cellulose (HPC) . . 15
4.4.1. Applications/Market 15
4.4.2. Synthesis 15
4.4.3. Processes 15
4.4.4. Characterization 15
5. Uses 16
6. Economic Aspects 16
7. Toxicology and Occupational Health 16
8. References 17

1. Introduction

Cellulose ethers are nontoxic, usually water-soluble, white to yellowish powders or granules. There are some cellulose ethers that are not soluble in water, but their share of sales is insignificant as compared with the water-soluble substances. Cellulose ethers are produced by a polymer-analogous reaction of cellulose with low-molecular alkoxylating agents, which can support further functional groups.

The manufacture of cellulose ethers was first published in an article by W. Suida [1] in 1905 and the first patents [2] relating to their industrial production were issued as early as 1918. In the decade between 1920 and 1930 carboxymethyl cellulose was the first cellulose ether to gain economic significance, followed by methyl celluloses and hydroxyethyl cellulose some ten years later. These three product categories still dominate the market today.

During the time cellulose ethers have been used, continuous intensive research focusing on production processes and applications has been carried out by both industry and academia. Research was and is imperative to enable this group of substances to be utilized in ever new applications. Despite major progress in the areas of regioselective synthesis and analytics over recent years, the potential of these substances has by no means been fully exploited [3, 4].

Cellulose ethers can be sorted by their economic significance. The most important products in terms of sales are carboxymethylcelluloses (approx. 230 000 tons per annum), methyl- and hydroxyalkylmethylcelluloses (approx. 120 000 t/a), hydroxyethylcelluloses (approx. 60 000 t/a) and hydroxypropylcellulose (less than 10 000 t/a). The demand is served by a small number of manufacturers, which employ proprietory processes and have developed their own areas of expertise.

2. Properties

Cellulose ethers are nontoxic, usually water-soluble, white to yellowish powders or granules.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_461.pub2

There are some cellulose ethers, that are not soluble in water, but their share of sales is insignificant as compared with the water-soluble substances.

The bulk density of cellulose ethers is between 300 and 600 g/L. They are lightfast and their storage life is claimed by the suppliers to be between 18 and 36 months. Cellulose ethers turn brown at elevated temperatures (around 200 °C) and show signs of decomposition above 250 °C. Their biodegradability can be influenced by the degree of substitution.

Cellulose ethers are capable of swelling or are colloidally soluble; they increase the viscosity of the solvent and develop a specific rheological profile. Depending on the substitution, cellulose ethers in solution are surface-active (surface tension: water 72 mN/m, carboxymethyl cellulose 70 mN/m, hydroxyethyl cellulose 63 mN/m, methyl cellulose 54 mN/m). The nonionic derivatives are film-forming substances, resulting in transparent films with a high elasticity. Solutions and films obtained from cellulose ethers are compatible with other hydrocolloids.

In many applications, however, where certain properties of an initially dry mixture such as dissolving and solvent binding behavior (water retention) are required, cellulose ethers are not used in ready-made solutions, but are blended with other solids,.

These three characteristics – dissolving behavior, the resultant solution structure, and the cellulose ether's capability of binding solvents – depend on molecular characteristics such as the type, number and distribution of the substituent and the molar mass distribution.

In technical specifications generally only the degree of purity, the viscosity in aqueous solution and in some cases the degree of substitution are stated.

2.1. Swelling and Dissolving Behavior

The different swelling and dissolving behavior of the various cellulose ethers originates from the interplay of the hydrogen bonds between free OH groups on the one hand and hydrophilic and hydrophobic substituent groups on the other. This interplay can be imagined as being the gradual introduction of hydrophobic substituents into native cellulose. Starting from a crystal structure based on hydrogen bridges, the partial imperfection of this structure caused by substituents results in free hydroxyl functions, hence first the substance becomes soluble in alkalies and subsequently in water. Further etherification with hydrophobic substituents results in solubility in organic solvents and ultimately thermoplasticity (intrinsic solubility).

Alongside the hydrophilic-hydrophobic interaction which culminates in "thermodynamic solubility", the particle-size distribution has a crucial impact on the obtainable state of solution. Particle size determines the kinetics of the dissolution process. If the particles are too large, they may not dissolve fast enough; they will swell and release some of their outer layer into the solution; due to the rising viscosity the mobility within the solution will decrease, leaving undissolved particle residues which will no longer dissolve in a reasonable time. Particles, that are too small, on the other hand, may cause formation of lumps during the dissolving process. Again this will ultimately result in undissolved particle residues preventing formation of a homogeneous solution.

To counteract these problems, most suppliers issue recommendations as to which particle size is best for a particular application. Some suppliers also offer specially treated products for enhanced dispersibility. These are mostly powder grades, whose initial dissolving rate is greatly reduced ("delayed or retarded solubility") to ensure that the substance is homogeneously distributed in the solvent before starting to dissolve. This delay can be achieved, for instance, by a reversible cross-linkage with glyoxal. The rate of dissolution is thus pH-dependent.

2.2. Solution State

The state of solution obtained, i.e., the sum of the interaction between the cellulose ether and the solvent and the intramolecular and intermolecular interaction of the cellulose ether molecules, determines the rheological profile of a solution or of an end product.

The standard rheological models of polymer chains dissolved in coil formation can be used to describe the characteristics of cellulose ether solutions, depending on the intrinsic viscosity

(also referred to as Staudinger index) and concentration. All known polymeric solution states from the diluted particle solution to the concentrated network solution [5] can thus be described. However, on account of their inclination to form aggregates, cellulose ethers often set scientists impossible tasks to solve.

The viscosity of a cellulose ether solution is generally stated as a characteristic feature. Measured in a 2 wt% aqueous solution, it is sometimes included in the product grade designation for orientation purposes. At first glance this seems to be useful, because many application-related properties are dependent on the same molecular parameters as viscosity. On closer, more discriminating examination, however, this turns out to be misleading, first because the dissolving and measuring methods differ from one supplier to another, which can result in a similar product having twice the viscosity, and second because – although giving a rough guide – the viscosity does not describe the solution structure and seldom the product's behavior in the final application system.

In the range relevant to the applications the viscosity of cellulose ether solutions rises disproportionately with increasing concentration and intrinsic viscosity:

$$\text{viscosity}_{\text{solution}} \propto \text{viscosity}_{\text{solvent}} \bullet (\text{concentration} \bullet \text{intrinsic viscosity})^{\frac{3.4}{a}}$$

($0.5 < a < 1.0$; a: exponent of the Mark-Houwink equation)

Most solutions show a pseudoplastic behavior, which is very important in practical terms.

Many cellulose ether solutions manifest time-dependent rheological functions, so the measurements depend on how the sample has been pretreated. The elasticity of the solutions also increases with rising concentration and intrinsic viscosity, but the exponents are different from those of the viscosity function; therefore the cellulose ether selected can modify the ratio of viscosity to elasticity within certain limits [6]. Some application systems exhibit a solid-like behavior at low shear stress and viscous flow only starts at stress levels above the so-called yield stress (plastic flow behavior).

The intrinsic viscosity describes the volume "filled" by the free molecule. The intrinsic viscosity is often used as a measure of the molar mass or chain length; the correlation is shown by the Mark-Houwink equation. When the intrinsic viscosity is gauged, not only the molar mass is determined but also the quality or performance of the solvent. Therefore effects attributable to pH, temperature and low-molecular salts, which have a substantial impact on the solution structure, are also part of the analysis. Many cellulose ethers can be salted out by "removing" the hydration shell of the molecules. With the ionic products, polyelectrolytic effects such as "repulsion" or "cross-linking" due to the charging of the macromolecules among themselves or with the charges of salts present in the solution also need to be observed.

The maximum possible thickening effect is predetermined by the used cellulose. The actual viscosity achieved by a cellulose ether depends on how well chain scissions can be prevented during the production process and on the type and quantity of substituents introduced into the cellulose molecule. Each substituent represents an increase in molar mass without an increase in the chain length. This reduces the thickening effect. The substituents are, of course, indispensable for solubility and other properties; however, derivatives which predominantly carry substituents with a low molar mass and where the degree of substitution is not too high are particularly effective thickeners. The substituent group with the lowest molar mass is the methyl group which simultaneously cancels out the intermolecular hydrogen bonds very effectively. As pure methyl cellulose manifests a specific temperature-dependent dissolving behavior, which restricts its universal use, mixed cellulose ethers with a low degree of substitution with hydroxyethyl or hydroxypropyl groups are generally used as high-viscosity thickeners, for instance in the construction sector.

The amount of cellulose ether used or its concentration in an aqueous solution has a similarly great effect as that of the chain length. When used as a thickener, the concentration range employed is usually in the range of a moderately concentrated network solution in which the polymer chains are not isolated but looped, forming a temporary physical suspension network.

The temperature dependence (below the flocculation point) of the flow behavior of aqueous cellulose ether solutions follows the principle

of temperature-time superposition. At a higher temperature all the molecules move faster. If a flow curve (shear stress as a function of the shear rate) is plotted at a reference temperature, the flow curve determined at a higher temperature follows the same shear stress path, except that it is merely shifted to higher shear rates. Accordingly, at the higher temperature the viscosity values are on a lower level.

Changes in flow behavior are thus predeterminable (as long as there is sufficient distance from the flocculation point). Therefore, when using the cellulose ether as a thickener in aqueous systems, the rheological behavior can be very reliably estimated above and below the reference temperature, if the flow activation energy is known. In the case of many cellulose ethers in aqueous solution the flow activation energy is in the proximity of 40 kJ/mol. Additional substances dissolved in the water may cause a shift in this value, but there is no fundamental change in the correlations.

2.3. Thermally Induced Coagulation

The solubility of cellulose ethers with hydrophobic substituents like methoxy groups decreases as the temperature rises, i.e., the performance of the solvent deteriorates. Cellulose ether dissolved in a cold solution can be made to coagulate by heating the solution. The process is reversible: when the solution cools down, the cellulose ether dissolves again. Depending on the selected type and degree of substitution, the flocculation point of industrial used products is varied between 30 and 100 °C.

This is an important characteristic for industrial production of cellulose ethers because in this way byproducts and salts can be washed out with hot water instead of with mixtures of water and alcohol.

To assess the flocculation point with mixed substituent groups, suitable flocculation diagrams have been published [7].

2.4. Degree of Substitution

The type of substituent and the degree of substitution determine most of the properties of a cellulose ether. With substituent groups where no secondary substitution is possible (methyl, carboxymethyl), the molar substitution (MS), i.e., the absolute number of substituents per anhydroglucose unit, is identical to the degree of substitution (DS), i.e., the number of substituents directly linked to the anhydroglucose unit. In the case of substituents containing hydroxyl groups (hydroxyethyl, hydroxypropyl), for example, further (secondary) reactions with the substituent can occur, i.e., there is no decrease in the number of reactive centers. With these products a distinction has to be made between MS and DS. Comparing MS and DS shows the proportion of side-chain substitution. On account of the simpler analytics involved, frequently only the MS is stated.

To examine the characteristic features more precisely, it is necessary to analyze the substituent distribution. Primary substitution (i.e., prior to any secondary reactions with the substituent) already involves three kinds of distribution:

- Substituent distribution within an anhydroglucose unit: depending on the reagent, the distribution can be between C-2, C-3, and C-6. This distribution can be analyzed by means of nuclear resonance spectroscopy or hydrolysis and subsequent chromatography [8].
- Substituent distribution along the backbone chain: in contrast to the random distribution normally found, a strictly uniform substituent sequence along the chain or a block structure formation are conceivable as theoretical borderline cases, which show up the different effects on solubility and therefore on the properties in general. To gain an impression of the distribution along the chain, various breakdown methods and subsequent liquid chromatography or derivatization and gas chromatography coupled with mass spectroscopy are used to obtain the proportions of non-, mono-, di- and trisubstituted anhydroglucose units. To assess the distribution of these units along the chain, the results of analysis of fragments of various lengths can be compared. Appropriate methods have so far only been developed for particular derivatives and only few research scientists have a good command of them [9, 10].
- Substituent distribution between the individual chains: the substituent density can also

vary between the individual chains of the cellulose ether, one known example being the nonsubstituted cellulose fibers in otherwise clear-soluble products. With ionic products the interchain distribution can be detected by polarographic analysis and an appropriate method for other derivatives is chromatography after specifically marking the free hydroxyl groups. The problem with both methods is that the molar mass distribution has to be known in order to carry out the assessment [11].

In the case of derivatives where a reaction with the substituent is possible, the potential number of different anhydroglucose units and their distribution increases randomly, and the structure becomes enormously complex. At the same time, however, the possibilities for adapting the hydrophilic-hydrophobic interaction to suit individual applications increase as well.

2.5. Molar Mass Distribution

Knowing the molar mass or the molar mass distribution makes it possible in theory to access various material functions. An absolute determination of these variables has so far only been achieved in isolated cases and with an enormous research input. Due to the "poor solubility" and the propensity of most derivatives toward association, producing time-stable molecularly disperse solutions is only rarely possible. Frequently it is a mixture of molecules, associates and aggregates that is analyzed, and the results are often impossible to interpret reliably.

It is precisely because of this complexity that so many attempts to make a breakthrough have been launched in the past, starting with "batch methods" such as static and dynamic light scattering in many different solvents, and "classifying" methods such as ultracentrifugation, size-exclusion chromatography and flow-field-flow fractionation, and also combined methods involving chromatography and various detection systems. All these approaches have in common that, although they give an interesting insight into the solution structure, the results often depend on how the samples have been prepared and how they are influenced by the measuring procedure itself. Taking the measuring conditions into account, some vital assertions can be made relating to existing systems. However, only in rare cases do the results stand up to comparison (between various methods) in respect of an "absolute" molar mass distribution.

3. Production

Cellulose consists of chains of anhydroglucose units (AGUs) linked together glycosidically (β-1-4-bonds) (→ Cellulose). Accordingly, each AGU still has three free hydroxyl groups which vary in their reactivity and which can be converted to ether groups in polymer-analogous reactions.

The free hydroxyl groups form intermolecular and intramolecular hydrogen bonds, so before the cellulose can be etherified it needs to be 'activated'. For this purpose the cellulose is chemically converted, usually with sodium hydroxide solution (caustic soda), in order to open up its natural structure and to obtain a sufficiently swollen material that is accessible to further reagents. The outcome is known as alkali cellulose or soda cellulose. The degree of swelling is controlled by the ratio of sodium hydroxide solution to cellulose and by the method of adding suspending agents or swelling agents. A minimum quantity of alkali (approx. 0.8 mol per mole of AGU) is required to swell the cellulose and for the reagents. During this activation process several properties of the end product are predetermined, because the degree of swelling and the uniformness of the activation influence the distribution of the substituents introduced in the second stage of cellulose ether production.

After the cellulose has been activated, the etherifying reagent is added. In principle, any etherification reaction known in the context of low-molecular alcohols could be applied, many of which have been studied [3]. Two basic types have become established in the industrial production of cellulose ethers – reactions where sodium hydroxide solution is consumed (e.g., Williamson ether synthesis) and reactions where sodium hydroxide only acts as a catalyst (e.g., alkoxylation). The supernatant caustic solution is neutralized after the reaction.

The kind of cellulose and its quality are crucial with regard to the ultimate properties of the cellulose ether. Therefore, many different types

of cellulose obtained from wood pulp or cotton linters are used. The type of cellulose used determines, among others, the maximum chain length, hence the viscosity yield of the final cellulose ether. The chain length is reduced on account of the chemical and thermal stresses imposed during the process. The incidence of chain scissions is inherent in the production process but can be selectively increased. Oxidative, acid and radiation-induced decomposition are some of the known degradation processes which, depending on the product category, are used at various points in the process to obtain a desired low viscosity.

Processes. The processes depend on the synthesis-related reaction stages which can take place in one set of equipment or in several sets arranged in sequence.

A distinction is made between reactions at ambient pressure and reactions carried out at up to 30 bar, depending on the different reagents and on the required weight percentage of cellulose or cellulose ether in the reaction medium (mass fraction). Heterogeneous syntheses are used for industrial-scale production, whereas homogeneous syntheses are so far only found in research. Heterogeneous systems are classified by mass fraction into suspension processes (mass fraction below 10 wt%) and slurry processes (mass fraction approx. 10 – 50 wt%). There is also the option of a gas-phase or vapor-phase process, that works without an inert liquid. The process technology and engineering is adapted to meet the requirements of the mass fraction involved.

The cellulose is usually supplied as rolls of cellulose webs or as cellulose sheets. In the early days of cellulose ether production, the cellulose was treated in an alkali bath (approx. 15 – 20 wt% caustic soda) with the disadvantage that the cellulose had not been ground and was therefore not fully accessible to the activating agent. Consequently, the activation process took 4 h and the ripening time 8 h or more [13]. During the ripening process the alkali cellulose was kept in intermediate tanks where the reaction was allowed to complete at a specific temperature. Prior to the ripening stage any excess activating agent was pressed out and the alkali cellulose was then shredded. Storing it in ripening tanks made handling more difficult and reduced its processability.

To make the cellulose more accessible to the chemicals involved, in modern processes the cellulose is generally ground or milled before the activating and reaction stages. The cellulose is then alkalized with highly concentrated (approx. 30 – 70 wt%) aqueous caustic soda solution. Alkalization usually takes place in mixers or stirred reaction vessels, depending on the mass fraction and the reaction system. The mass fraction can be selected to influence the water balance which affects side reactions or reagent yield and also the quality of the cellulose ethers. The way the chemical reaction is engineered during the alkalization stage also has a significant effect on the quality of the cellulose ethers (see above).

The reaction can be carried out either batchwise, semibatchwise (fed batch) or continuously, and the reaction temperatures vary from room temperature up to 110 °C, depending on the type of etherification. Here again, mixers and stirred vessels are the established equipment. There are some isolated patents describing continuous processes in different equipment setups [14 – 16]. The starting materials are fed as solids or liquids under pressure. In some older processes for the production of methyl or hydroxyalkylmethyl celluloses gaseous reagents are used.

The equipment assembled downstream of the reaction depends on the required degree of purity. In general a distinction is made between technical and nontechnical cellulose ethers which differ in their content of byproducts and neutralization products – predominantly salts. Depending on the washing agents and the type of wash (displacement or dilution-type wash) the equipment used for solid-liquid separation includes suction filters, belt filters, drum filters, centrifuges and decanters, operating either continuously or batchwise. The washing processes are carried out in aqueous or alcoholic media, depending on the product group.

If the product is still wet with solvent after the washing sequence, the inert dissolving or washing agents can be recovered by displacing them with water prior to the drying stage. Suitable equipment for this are, for example, continuous mixers. Auxiliary facilities are used for regenerating the reagents and washing media and for extracting byproducts.

At the end of the process line the products are converted in forming stages from their fibrous form into the required supply form, usually powder grades with a defined particle size distribution and containing small amounts of residual moisture (in most cases < 10 wt%). Various drying and comminuting equipment is used in a number of different combinations. This process step is also utilized for modifying the cellulose ethers so that they offer optimal properties for the individual fields of application. The particle size distribution ranges from granular grades with an average grain diameter of approx. 0.5 mm to very fine powders where the percentage of particles with a diameter of less than 63 µm is higher than 50 wt%.

4. Product Groups

4.1. Carboxymethyl Cellulose (CMC)

In terms of sales volume carboxymethyl celluloses are the largest product group. They are available in various levels of purity from completely "unpurified" to "highly purified" for food-grade applications. The CMC group also includes mixed ethers with hydroxyalkyl substituents, that have a predominantly ionic character. The structure of CMC is shown in Figure 1.

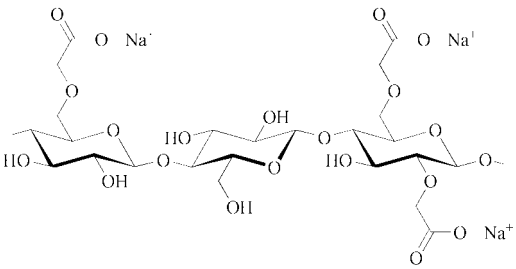

Figure 1. Structure of CMC

4.1.1. Applications / Market

The market can be divided into two segments: one for "purified" and another for "unpurified" CMC. However, the terms "purified" and "unpurified" are subject to regional and supplier-specific differences in definition. Products with an active content of 55 – 85 % are considered to be "unpurified"; 85 – 95 %: "semi-purified"; around 98 %: "purified"; and products with an active content greater than 99.5 % are classed as "highly purified" (high-purity grades).

Annual sales worldwide amount to approx. 230 000 t, of which purified CMC accounts for 130 000 t and unpurified CMC totals 100 000 t. It is difficult to accurately apportion figures to the two segments, not only because of varying definitions but also due to the fact that in some applications both purified and unpurified grades of CMC can be used, and many suppliers have both purified and unpurified products in their portfolios.

CMC can be produced with less technological input than the other cellulose ethers, which explains why a comparatively large number of manufacturers (more than 20) exists. The leading producer is CP Kelko with a market share of around 33 %. The largest regional market and the region with the largest production capacity is currently Europe, which still accounts for some 60 % of global capacity.

The initial application fields of CMC were detergents, where it acts as a soil carrier, and in deep-well drilling for oil and water, where it is used as a flotation aid in drilling mud. Unpurified grades were and in some cases still are used in these applications. Purified products are used in surface coatings and various other technical areas such as in the paper industry for coatings and pulp sizing to improve fiber retention, filler/pigment/dye yield as well as strength. CMC also improves the printability and smoothness of paper. Together with gelatin CMC is used as a coacervate to encapsulate ink in the production of noncarbon copy papers. Users of high-purity grades include the cosmetics and pharmaceutical industries, where CMC is used for instance as a fat-free ointment base or as a tablet disintegrant. In food and pet food production CMC's simplest function is to improve consistency and to provide a low-calorie replacement for starch and proteins, or to control more complex functions such as the freeze-thaw stability of deep-frozen products or the creaminess of ice cream.

4.1.2. Synthesis

CMC is produced in a Williamson ether synthesis from alkali cellulose with sodium chloroacetate or with chloroacetic acid itself, which reacts in-situ with caustic soda to form the salt. The process takes place in an aqueous or aqueous-alcoholic medium (slurry). The alcohols used are ethanol, isopropanol, *tert*-butanol and mixtures thereof.

$$\text{Cell}-\text{OH}\cdot\text{NaOH}+\text{ClCH}_2\text{COONa}\rightarrow$$
$$\text{Cell}-\text{O}-\text{CH}_2\text{COONa}+\text{NaCl}+\text{H}_2\text{O}$$

Alkalyzation is carried out at room temperature; the reaction takes place between 50 °C and the boiling point of the slurry medium under appropriate system pressure. The reaction is exothermic. With an activation energy of 87.9 kJ/mol it is highly temperature-dependent. The yield from this reaction in relation to chloroacetic acid is between 65 % and 80 %. Hydrolysis of chloroacetic acid occurs as a side reaction, forming glycolate. Commercial CMCs are produced with a degree of substitution between 0.2 and 1.5. Carboxymethylation takes place with a slight preference toward C2- followed by C6- and C3-substitution of the anhydroglucose unit [17].

4.1.3. Processes

In the early days of CMC production the cellulose was activated in alkaline steeping tanks [18]. After the excess caustic soda solution had been pressed out and the alkali cellulose had been shredded to fibers, the reaction was carried out with the addition of sodium chloroacetate in a kneader, where temperatures reached 30-35°C. The reaction was completed in a tumbling drum at 50 °C. The material was then washed with a mixture of methanol and water, filtered, ground and finally dried.

The first continuous process was used by Wyandotte Chem. Corp. (1947), but there was no washing stage, so it was only suitable for unpurified CMC [19]. Sodium hydroxide solution and monochloroacetic acid were added to the ground cellulose in a rotating drum. After a dwell time of 3 h the product was allowed to "ripen" for 10 h in drums.

In contemporary CMC production processes the ground cellulose is processed batchwise either as a suspension in the case of a low mass fraction of cellulose or, if the mass fraction is high, in mixers in the presence of "inert" media (see Fig. 2). Both processes use one or more short-chain alcohol(s) as a mass-transfer and heat-exchanging medium or as a suspending agent. For processes with a high mass fraction ethanol is usually employed a slurry medium, whereas isopropanol is used in suspension-type processes [20, 21]. By placing several sets of the same kind of equipment in parallel or in series it is possible to implement a quasi-continuous operation in the activation and reaction stages.

The starting materials and suspension or slurry media are either solid or liquid at normal room and reaction temperatures. Activation and reaction can therefore be carried out at ambient pressure.

Depending on the required product purity, the product is washed with alcohol-water mixtures, preferably with the same alcohol used in the reaction. Washing with water is not possible because of the water-solubility of the products. To meet product purity requirements, it is necessary to remove the alcohols from the product after the washing stage and prior to the grinding and drying processes.

Suspension agents and washing liquids are collected and reprocessed either by distillation and extraction or membrane processes.

4.1.4. Characterization

Carboxymethyl cellulose becomes water-swellable or alkali-soluble even at a low degree of substitution. The water-solubility of industrially produced CMC begins at a DS of approx. 0.6. The solubility increases up to a rather theoretical DS value of 3.0. Research work has shown that products with a lower degree of substitution (DS < 0.6) can manifest good solubility, the reason being different substituent distribution along the backbone [22].

Aqueous carboxymethyl cellulose solutions have very complex solution structures, because boundary conditions such as pH or salts simultaneously affect various characteristics. The solution state achieved depends, as with other cellulose ethers, on the nature of the dissolution pro-

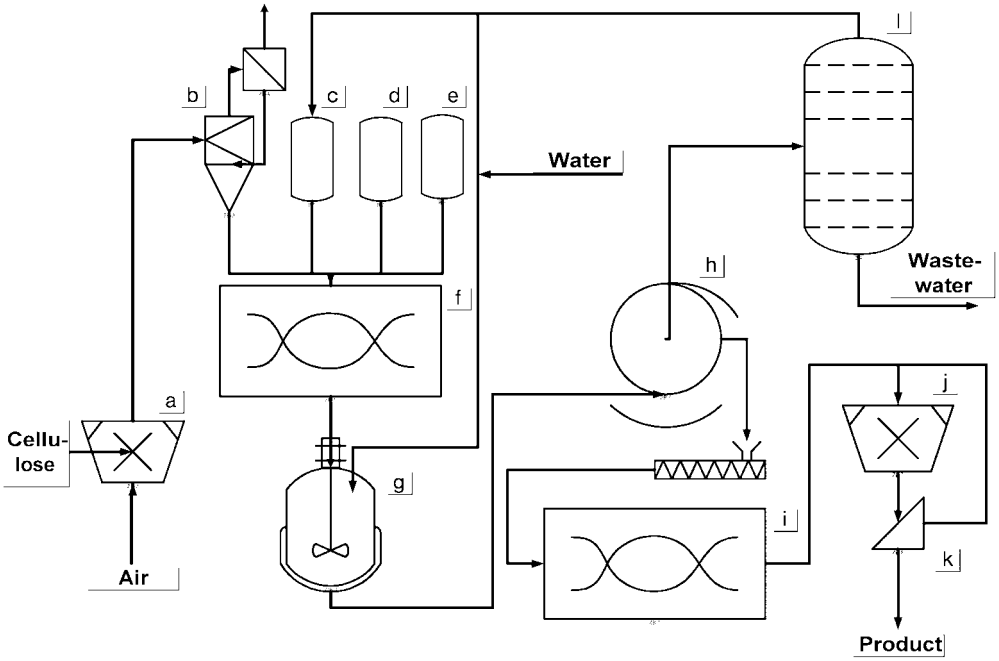

Figure 2. Reaction mixer processes for purified CMC production
a) Cellulose shredding; b) Cellulose buffer; c) Slurry medium dosing vessel; d) Caustic soda dosing vessel; e) Chloroacetic acid dosing vessel; f) Reaction mixer; g) Washing vessel; h) Filtration equipment; i) Dryer; j) Comminuting; k) Screening equipment; l) Slurry medium recycling

cess and the performance of the solvent. The parameter "solvent performance" covers all thermodynamic effects of protons, salts and other low-molecular substances on the interaction between the solvent and the macromolecule.

In addition CMC is a weak carboxylic acid. The proton concentration in the solution not only affects the solvent's performance but simultaneously alters the number of free carboxylate groups. CMC's equivalence point is at a pH of 8.25; its pK_a value is between 4 and 5. In its acid form CMC is insoluble in water. Depending on the degree of substitution CMC can be precipitated from the solution at low pH values. Deprotonation increases with rising pH values and the character of the polyelectrolyte becomes apparent: The repulsion among the individual substituents becomes more distinct and the salt sensitivity increases. When CMC comes into contact with selected divalent or trivalent salts, this leads to complexing and in some cases to interlinking and precipitation of the CMC. With divalent salts these effects depend largely on the ratio of substituent to salt [23].

The many different possibilities of microscopic interaction are also reflected in the macroscopic properties of CMC. From a rheological point of view CMC solutions are inclined to form superstructures. Viscosities of 2 wt% solutions ranging from 3 mPa·s to 100 000 mPa·s are available on the market. Most of the products exhibit pseudoplasticity, some of them a thixotropic behavior. Depending on the selected degree of substitution and the product group, the solutions are either smooth-flowing or highly gelatinous.

Commercial grades are not only categorized according to their purity, their morphology and their mean viscosity, but also according to their degree of substitution. Water-soluble CMC commodities are available with a DS of around 0.7, 0.9 and 1.2. These three product groups do not only differ in their DS, but also in their solution structure, hence their rheological behavior.

4.2. Methyl and Hydroxyalkyl Methyl Celluloses

Pure methyl cellulose (MC) and hydroxyalkyl methyl celluloses (HAMC) are collectively known as methyl cellulose. Pure methylcellulose only accounts for a minor share of the market. The mixed ethers hydroxyethyl methyl cellulose (HEMC, see Fig. 3) and hydroxypropyl methyl cellulose (HPMC) have established themselves in particular for technical applications.

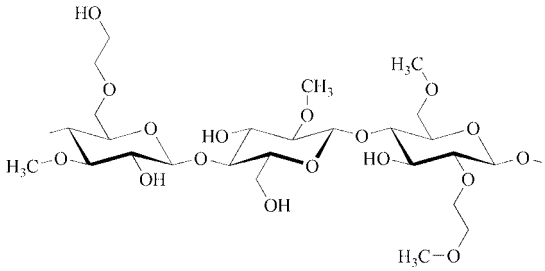

Figure 3. Structure of HEMC

4.2.1. Applications / Market

The market for methyl cellulose and hydroxyalkyl methyl celluloses was in the range of 120 000 t in 2003, the market leader being Dow Chemical Company, with Shin-Etsu Chemical and Wolff Cellulosics having slightly lower market shares. High-purity grades are marketed for applications in the food, pharmaceutical and cosmetics sectors and purified products are used in building materials and industrial applications. The building material segment is by far the largest field of application.

The applications and therefore also the types of product used in the building materials sector vary from region to region due to different construction methods and traditions (→ Dry Mortars). In the USA joint compounds account for the largest segment; in Europe the main applications of MC and HAMC are plasters, renders, and tile adhesives. Efficient modern application methods would not be possible without the use of methyl celluloses. Properties such as water retention, open time, wet adhesion, initial thickening and setting behavior are regulated using only low additive contents of between 0.01 and 2 wt% (in relation to the building material system). Accordingly, the cellulose ether has to meet high quality requirements.

In the life-science sectors only nonhydroxyalkylated methyl cellulose (modified vegetable gum) and hydroxypropyl methyl cellulose (carbohydrate gum) are used. In the pharmaceutical industry methyl celluloses are employed as tablet base and in coatings of controlled-release drugs. In the food and cosmetics industries it is the thickening and emulsifying properties that are exploited to give the products the desired consistency and texture. The most noteworthy industrial application is the use of highly methylated hydroxypropyl methyl cellulose as a protective colloid in the polymerization of vinyl chloride.

4.2.2. Synthesis

The synthesis of hydroxyalkyl methyl cellulose is a combination of both principles of synthesis in industrial cellulose ether production: methylation is along the lines of Williamson ether synthesis where a stoichiometric amount of NaOH solution is consumed, whereas alkylation only requires catalytic amounts of alkaline solution.

Pure MC is produced from alkali cellulose by a reaction with gaseous or liquid methyl chloride. In side reactions the methyl chloride is hydrolyzed to form methanol, and subsequently the methanol is etherified by methyl chloride, forming dimethyl ether. The reactions take place between 70 and 120 °C. Methylation is an exothermic reaction, the activation energy being 80 kJ/mol.

$$\mathrm{Cell{-}OH{\cdot}NaOH + ClCH_3 \rightarrow Cell{-}O{-}CH_3 + NaCl + H_2O}$$

The grades produced have a degree of substitution of 1.7 and 2.3. In case of pure methylation the substituent is found to prefer C2-, followed by C6- and C3-positions of the anhydroglucose unit [25, 26].

In the course of hydroxyalkyl methyl cellulose production alkoxylation of the cellulose takes place prior to or parallel to its methylation. As both the cellulose and the alkoxy groups formed can be methylated and alkoxylated, the outcome is a potentially indefinite number of

different products, allowing the range of properties to be tailored to the requirements of specific applications.

4.2.3. Processes

The development of processes to produce methyl cellulose and hydroxyalkyl methyl celluloses began with methods at ambient pressure, i.e. gas circulation processes, where the cellulose is activated (for example by a steeping process), shredded and subsequently brought into contact with the reagents in a mixer. Nonreacted gaseous feed materials and byproducts are continuously extracted, the byproducts are condensed out and the gas flow is returned to the reactor. The product is then washed, filtered, compressed, ground and dried.

The slurry process (see Fig. 4), operated at pressures of up to 30 bar, is the process in most widespread use today for producing HAMC batchwise or semibatchwise in a mixer. Processes patented in the 1950s suggest the use of pressure-resistant reactors without any additional inert slurry media [27]. More recent patents propose adding inert slurry media, mostly ethers [28]. The slurry medium acts as a heat-transfer medium, which evaporates and is condensed in a dome-shaped condenser.

The reagents are added to the ground cellulose in any sequence, depending on the targeted product characteristics. The slurry medium has to be added prior to etherification on account of the heat generated during the exothermic reactions. Once the reaction has been completed, low-boiling slurry media are extracted by reducing the pressure and are collected for reuse. Byproducts are discharged. High-boiling slurry media are separated during the washing stage and subsequently reprocessed.

Suspension-type processes are mentioned in several patents. Their advantages are their simple process engineering, whereas disadvantages can be seen in the more complex preparation of the suspension agent.

The flocculation point of MC and HAMC allows the products to be washed with water above this flocculation point. To do so, the contents of the reactor are led into a stirred tank filled with hot water or they are suspended in the reaction mixer; the salts formed during the reaction are dissolved and any nonreacted activation reagents are neutralized with acid. The suspension is then separated in a solid-liquid separating unit. Further washing stages may be required, depending on the product specifications, in order to comply reliably with restrictions in content of byproducts.

Again depending on the type of product, the HAMC is then conditioned for the grinding and drying stage. Drying and grinding can be carried out in succession or in combined milling and drying units. The required fineness of grind, i.e., the particle size, which in some cases is less than 63 μm in more than 50 wt% of the product, can often only be achieved in screening and sifting stages, after which the coarse-grained material is recycled.

4.2.4. Characterization

With a degree of substitution of around 1, methyl celluloses exhibit good water solubility at room temperature. The higher the degree of substitution, the better the solubility in water. However, it decreases again with further methylation and changes to solubility in organic solvents at a DS of approx. 2.6, because the material becomes increasingly hydrophobic.

Available on the market are pure methyl celluloses with a degree of substitution between 1.7 and 2.3, hydroxyethyl methyl celluloses with methylation levels between 1.3 and 2.0 and ethoxylation levels between 0.1 and 0.4, and hydroxypropyl methyl celluloses with degrees of methylation between 1.3 and 2.1 and of propoxylation between 0.1 and 1.0. The limits of substitution are more or less predetermined by the products' flocculation point, because washing with water is less expensive than with alcohol.

The flocculation point, i.e., the gelation and subsequent coagulation as the temperature rises, is a characteristic found only in nonionic cellulose ethers. By varying the methyl and/or alkyl content of hydroxyalkyl methyl celluloses, it is possible to modify the macromolecule's affinity to its water shell and therefore the flocculation point: with increasing methyl substitution the number of hydrophilic groups is reduced, and the flocculation point drops. Accessibility to the hydrophilic groups increases in proportion to the

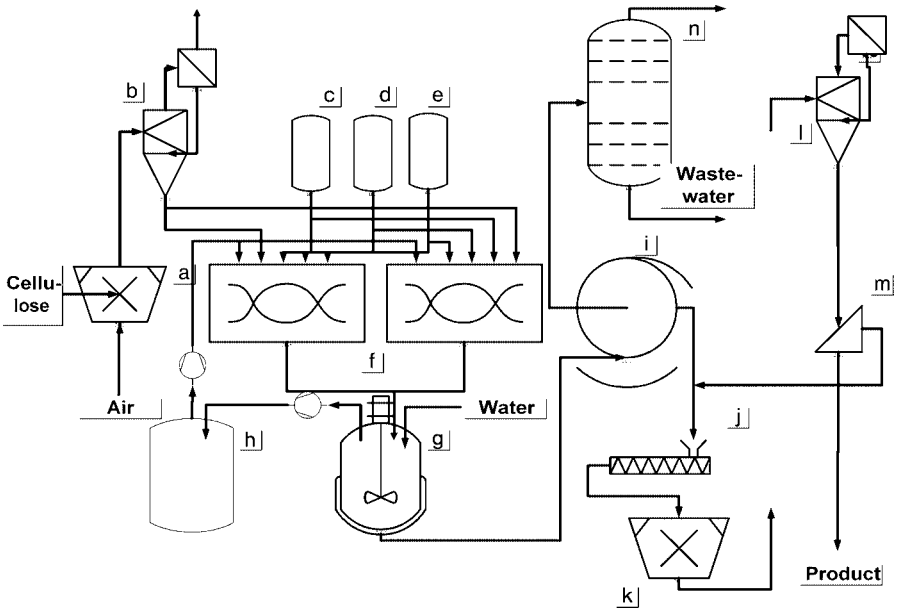

Figure 4. Modern HAMC production process
a) Cellulose shredding; b) Cellulose buffer; c) Caustic soda dosing vessel; d) Alkoxylation agent dosing vessel; e) Methyl chloride dosing vessel; f) Reaction mixer; g) Washing vessel; h) Gas buffer; i) Filtration equipment; j) Conditioning equipment; k) Comminuting; l) Stream dryer; m) Screening equipment; n) Slurry medium recycling

rise in hydroxyethylation, and the flocculation point shifts to higher temperatures. (The introduction of hydroxypropyl substituents results in opposite characteristics where, depending on the degree of substitution, the flocculation point either rises or drops.) Especially for applications where there is a high ambient temperature, e.g. on building sites in the summer, a methyl cellulose (MHEC or MHPC) with a sufficiently high flocculation point should be used to eliminate any undesired changes in characteristics.

The rheological properties of methyl celluloses and hydroxyalkyl methyl celluloses vary within a broad range, exhibiting some very different features, because there is a wide viscosity range in the grade spectrum available and because the products can have very differing chemical structures.

Products with a particularly low viscosity, such as HPMC for tablet coatings with a typical viscosity of between 3 and 6 mPa·s at a concentration of 2 wt% in water, still show a purely Newtonian (ideal) flow behavior in aqueous solutions with a HPMC concentration of 10 – 20 wt% at room temperature, and the viscosity of these solutions can be as high as 10 000 mPa·s. The Newtonian behavior is crucial for spraying highly concentrated solutions and for good leveling and film formation.

HEMC or HPMC grades with a higher viscosity for use as thickeners, however, exhibit a non-Newtonian flow behavior even with at low concentrations. In building materials applications, such as tile adhesives, the high static shear viscosity guarantees standing strength and resistance to demixing as well as high water retention, while the non-Newtonian properties ensure rapid and easy workability. The thickening characteristics of nonionic cellulose ethers are scarcely affected by the presence of dissolved (even polyvalent) salts or by changes in the pH value, so they are also ideal for use in cement-based building materials in which, after water is added, highly alkaline conditions prevail.

The most effective way to achieve good thickening at a low shear rate is to use small input quantities of high-viscosity grades. In these systems the viscosity soon drops considerably as the shear rate increases. A high thickening effect at high or at least at medium shear rates is desirable for improved workability in certain high-quality building material systems, and it can only be obtained with higher levels of methyl cellulose, which also provide additional benefits, such as

long open times and improved adhesion in the case of tile adhesives.

4.3. Hydroxyethyl Cellulose (HEC) [29]

Hydroxyethyl celluloses are the third largest product group in the marketplace. They are nonionic and soluble in both cold and hot water. Their manufacture was first described in 1920 in a patent granted to Farbenfabriken Bayer. The structure of HEC is shown in Figure 5.

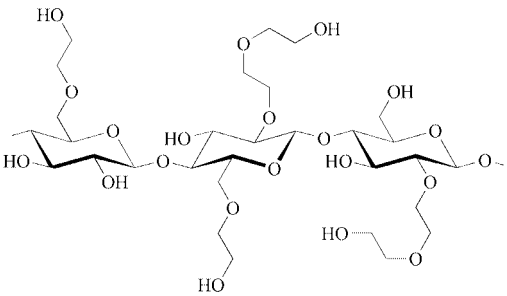

Figure 5. Structure of HEC

4.3.1. Applications / Market

The hydroxyethyl cellulose market is served mainly by Hercules Incorporated/Aqualon and Dow Chemical Company. Annual sales are around 60 000 t, most of which are used in the construction industry, and specifically in surface coatings. Hydroxyethyl cellulose is also found in cosmetic applications where it competes against HPMC derivatives. Other applications are in drilling fluids and as a protective colloid in emulsion polymerization.

In the paint industry HEC is the cellulose ether product group with the largest market share. Although products based on HEMC, HPMC or CMC have smaller shares, they are firmly established in these applications for technical and cost-related reasons. Beside pure HEC this product group also includes mixed cellulose ethers similar to HEC: ethyl hydroxyethyl cellulose (EHEC) and hydrophobically modified (hm) variants of HEC. The additional hydrophobic properties are imparted to hm-HEC by minor secondary substitution, for instance with long-chain alkyl residues (C_{12} through C_{24}) [30]. The hydrophobically modified grades are, like purely synthetic associative thickeners, capable of providing additional effects by interacting with particle surfaces, in particular those of the binder in latex paints, and by forming micelles and mixed micelles with other surfactants contained in the paint. These options are exploited especially in high-grade paint formulations in order to reduce spattering when applying the paint with a roller, and to obtain improved brush resistance and leveling.

4.3.2. Synthesis

Hydroxyethyl cellulose is formed by reaction of cellulose with ethylene oxide. This ethoxylation only requires catalytic quantities of alkaline solution.

$$\text{Cell}-\text{OH}+\text{OH}^- \rightarrow \text{Cell}-\text{O}^- + \text{H}_2\text{O}$$
$$\text{Cell}-\text{O}^- + \text{CH}_2\text{CH}_2\text{O} \rightarrow \text{Cell}-\text{O}-\text{CH}_2\text{CH}_2-\text{O}^-$$
$$\text{Cell}-\text{O}-\text{CH}_2\text{CH}_2-\text{O}^- + \text{H}_2\text{O} \rightarrow \text{Cell}-\text{O}-\text{CH}_2\text{CH}_2-\text{OH}+\text{OH}^-$$

Just as the cellulose alcoholate reacts with the ethylene oxide, also the hydroxyl ions present can react with the ethylene oxide or with the glycolates formed in the process. Large amounts of alkali reduce the yield of the main reaction, but to produce HEC a minimum quantity of caustic soda solution is essential to break down the cellulose.

Ethoxylation takes place in the presence of an "inert" solvent (isopropanol, *tert*-butanol, 1,2-dimethoxyethane or acetone); it starts at a very low temperature ($\approx 30\,^{\circ}\text{C}$) and is highly exothermic. HECs are produced with a molar substitution of 1.5 and 3.5, these degrees being controlled via the quantity of ethylene oxide added. The yield is between 40 and 75 % in relation to ethylene oxide. The reaction of ethylene oxide takes place in the following order of preference (from high to low): alkoxy groups, then C6 positions, then (followed by some distance) C2 and C3 positions [31]. The relationship of the reaction rates can be influenced by varying the alkali concentration, e.g., by partial neutralization [32]. After the reaction has taken place, the alkali quantity added needs to be neutralized.

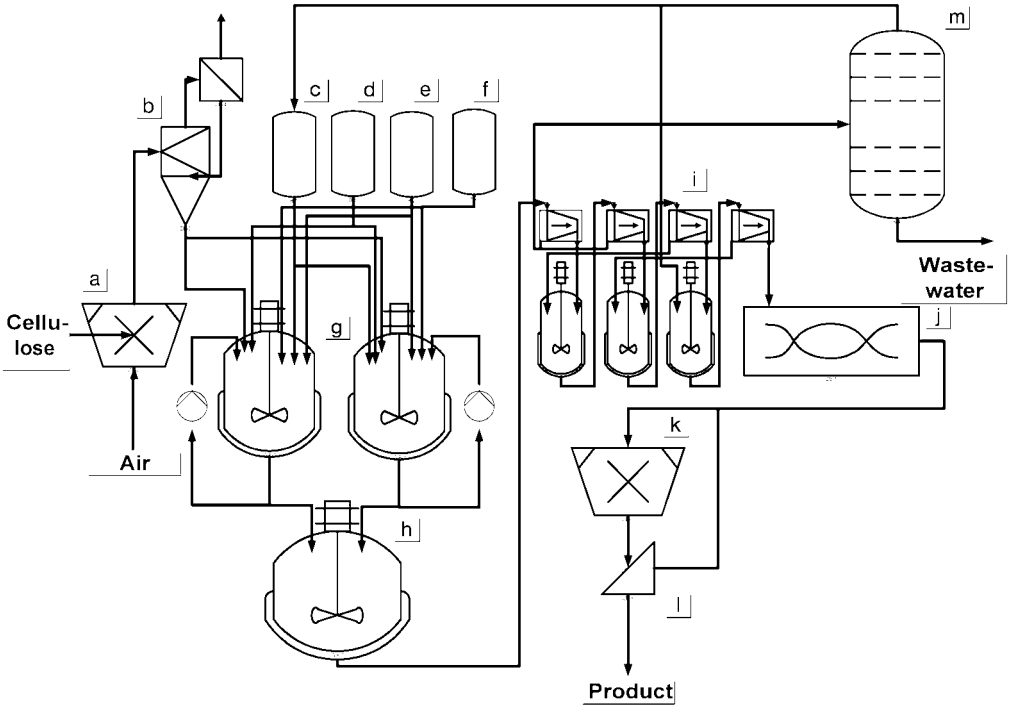

Figure 6. Process for the production of HEC/HPC with stirred tanks
a) Cellulose shredding; b) Cellulose buffer; c) Slurry medium dosing vessel; d) Caustic soda dosing vessel; e) Acid dosing vessel; f) Alkoxylation agent dosing vessel; g) Stirred reaction tank; h) Washing vessel; i) Washing and filtration process; j) Stream dryer; k) Comminuting; l) Screening equipment; m) Slurry medium recycling

4.3.3. Processes

For the production of HEC, suspension processes above ambient pressure have become the established methods [33]. Stirred pressure vessels are used for the reaction (see Fig. 6).

The ground cellulose is suspended in the suspending medium and mixed with the alkalizing agent. After alkalization, liquid ethylene oxide is added to the suspension and allowed to react for about 2 h.

Salts and byproducts have to be washed out of the products. As with other cellulose ethers, this involves appropriate equipment for washing and solid-liquid separation either in series or in a combined unit. As this cellulose ether does not have a flocculation point, mixtures of water and an organic solvent are used, followed by drying and grinding.

Parallel to this, the suspension medium and any extracting agents have to be regenerated and recycled. Distillation is the method usually selected for this.

4.3.4. Characterization

Hydroxyethyl celluloses are available on the market with viscosities between 10 mPa·s and 100 000 mPa·s and molar substitutions of 1.5 and 3.0 (up to 4.5 in the case of hydrophobically modified grades). They are water-soluble and have no thermal flocculation point. Hydroxyethyl celluloses are also more difficult to salt out than methylated derivatives and are not displaced from water even by larger amounts of organic solvent. The degree of substitution varies in the region of 1 in industrially produced grades, i.e., higher hydroxyethylation leads chiefly to secondary, tertiary, and quaternary substitution of the hydroxyalkyl substituent (formation of ethylene oxide side chains).

The rheological behavior of HECs is comparable with that of carboxymethyl celluloses. By introducing a small number of hydrophobic side chains, grades can be obtained, whose rheological profile in solution is highly modified,

particularly when mixed with surfactants. The hydrophobic residues form part of the surfactant micelles and this increases the viscosity, especially at elevated shear rates.

4.4. Hydroxypropyl Cellulose (HPC)

In relation to the quantities of cellulose ethers used worldwide, hydroxypropyl celluloses are the smallest product group. Their properties range from simple swellability, through solubility in cold water to enhanced thermoplasticity, which enables them to be processed in melt extruders. Hydroxypropyl cellulose was first launched on the market in the USA at the end of the 1960s. The structure of hydroxypropyl cellulose is shown in Figure 7.

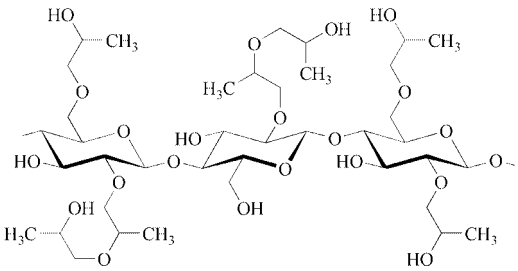

Figure 7. Structure of HPC

4.4.1. Applications/Market

The annual global market for hydroxypropyl cellulose is far below 10 000 t. The main fields of application are in the pharmaceutical and cosmetics industries [34]. High-purity grades are available for food and pharmaceutical applications, and purified grades are supplied to the cosmetics industry and for technical applications.

Hydroxypropyl cellulose is supplied by Hercules Incorporated/Aqualon, by Shin-Etsu Chemical and by Nippon Soda.

4.4.2. Synthesis

The synthesis of hydroxypropyl cellulose does not differ from that of hydroxyethyl cellulose. Prior to the reaction with propylene oxide, the cellulose is activated with caustic soda solution. The reaction has a somewhat higher activation energy and requires a higher reaction temperature than that with ethylene oxide. The byproducts are propylene glycol and polypropylene glycols which, like the salts from the neutralization stage, need to be removed in one or several washing stage(s).

4.4.3. Processes

The production processes for HPC are largely the same as for HEC. The reaction is generally carried out in a slurry. The slurry media described in relevant literature are hexane, toluene, tetrahydrofuran and dioxane, and also the common alcohols. Apart from the slurry process, the gas-phase method — which produces high-quality grades due to the high degree of substitution — is also applied. Since highly substituted HPC has a thermal flocculation point, however, washing can be carried out using hot water, which is not possible with HEC [35].

4.4.4. Characterization

Highly substituted hydroxypropyl cellulose is readily soluble in cold water and also soluble in polar organic solvents. The HPCs available on the market have a high molar substitution (3.0 – 4.5) and mean viscosities between 10 and 30 000 mPa·s. HPC solutions are generally smooth-flowing and exhibit scarcely any tendency toward thixotropic behavior. Gelation and coagulation begins at relatively low temperatures (40 – 50 °C). Hydroxypropyl cellulose is surface-active and tends to foam. Highly concentrated HPC solutions can develop liquid-crystalline phases. With a sufficiently high degree of substitution hydroxypropyl cellulose is thermoplastic and can be processed at approx. 150 °C. The extruded moulded parts are water soluble yet, due to their low propensity to absorb water, they are not sticky even when the humidity is high.

HPC with a low molar substitution of between 0.2 and 0.4, on the other hand, swells in water, but is not water-soluble. These grades are used as binders or disintegrants in the manufacture of tablets.

5. Uses

Cellulose ethers are used in a wide variety of fields ranging from oil drilling to industrial applications such as polymerization, from surface coatings and construction materials to the health care sector, cosmetics, food and pharmaceuticals. Not all types of cellulose ethers are equally suitable in all fields of application. Table 1 provides an overview of the great variety of applications and the appropriate types of cellulose ether. A listing of this kind can only be incomplete, but indicates the versatility and above all the adaptability of the products to the various fields in which they are to perform.

In many cases the amount of cellulose ether required is only between 0.02 and 2 %. This small level is, however, crucial for properties such as water-binding capacity, film formation and thickening as well as for controlling other rheological properties. Many modern products and processing techniques would not be possible without the use of cellulose ethers.

Each field of application requires specific solutions which in many cases involve modifying the cellulose ether product. This means that, once a suitable product category has been selected, all the parameters relating to the appropriate type of substitution and viscosity range have to be fine-tuned. Besides their chemical structure, characteristics such as product purity and forming, i.e., the physical state in which the products are made available, play a substantial part. Fine-tuning requires a great deal of experience in respect of the properties of these products as well as with regard to the systems and methods of application. Most suppliers provide this expertise and offer additional technical service.

6. Economic Aspects

In 2003 the sales volume of the cellulose ether market worldwide was approximately \$ 2×10^9. The largest market is Europe, followed by the USA and Japan. Annual consumption is ca. 420 000 t.

Sorted in the order of their economic significance (by sales volume), carboxymethylcelluloses rank first with approx. 230 000 t/a, while methyl- and hydroxyalkylmethylcelluloses are second with approx. 120 000 t/a, followed by hydroxyethylcelluloses on third place with approx. 60 000 t/a and last by hydroxypropylcellulose with less than 10 000 t/a.

Taking all market segments together, the cellulose ether market has grown continuously in recent years. However, there has been and still is a noticeable shift among the individual markets. As opposed to vigorous growth in the threshold countries and moderate growth in Europe and the USA, the Japanese market is dwindling.

The market is served overall by a small number of suppliers of cellulose derivatives. Most of the manufacturers produce a variety of different products for several market segments. The carboxymethyl cellulose market is spearheaded by CP Kelko, a company with a market share of 33 %. The remaining market is shared by 10 to 15 small to medium-size suppliers. The methyl and hydroxyalkyl methyl cellulose market is served by a few major producers [Dow Chemical Company, Wolff Cellulosics, and Shin-Etsu Chemical (incl. SE Tylose, the former cellulose ether division of Clariant AG)] with a market share totaling 80 %. Hydroxyethyl cellulose is supplied more or less exclusively by Hercules and Dow Chemical Company, whereas HPC is mostly supplied by Hercules and Shin-Etsu Chemical.

7. Toxicology and Occupational Health

Toxicology. Cellulose ethers are generally nontoxic. High-purity grades of most commercial products are approved as food additives and for use in cosmetic compositions.

Handling. Fine powders of cellulose ethers form explosive dusts in air as do natural polysaccharides or sawdust. Dry nonionic ethers undergo electrostatic charging similar to that of other organic polymers. When cellulose ethers are stored and handled, the general precautions concerning powdered organic polymers must be observed. The flammability is similar to that of cellulose. Spilling of solutions forms very slippery films that are difficult to remove.

Ecology. The biodegradation of cellulose ethers by cellulase-producing microorganisms also occurs in wastewater and therefore prevents the accumulation of cellulose ethers. Glucose, glucose ethers, and ether oligomers result from

Table 1. Major fields of application in which the individual product groups are used

Carboxymethyl cellulose	Methyl cellulose, hydroxyalkyl methyl cellulose	Hydroxyethyl cellulose	Hydroxypropyl cellulose
Paper	tile adhesives	latex paints	adhesives
Detergents	plasters/renders	adhesives	ceramics
Drilling for oil and gas	pharma/cosmetics	building materials	cosmetics
Pharma	joint compounds	cosmetics	encapsulation
Cosmetics	wallpaper paste	drilling for oil & gas	food
Textile industry	polymerization	agriculture	household goods
Food	food	paper	printing inks
Coatings	latex paints	synthetic resins	polymerization
Encapsulation	cement extrusion	textile industry	films

enzymatic hydrolysis; they are further degraded to carbon dioxide and water in slow bioreactions. Biostable or toxic metabolites are not known.

Cellulose ethers have no fish toxicity and are poor nutrients for most microorganisms. Nevertheless, wastewater bacteria may adapt to enhanced cellulose ether degradation after some exposure.

The suppliers provide Safety Data Sheets containing detailed data with regard to handling and the toxicological properties of the respective products.

8. References

1. W. Suida, *Monatshefte Chemie* **26** (1905) 413–427.
2. E. Jansen, DE 332203, 1918.
3. D. Klemm, et al.: *Comprehensive Cellulose Chemistry*, Vol. **2**, WILEY-VCH Verlag GmbH, Weinheim 1998.
4. J. Schurz, et al.: “Methodische Fortschritte in der Instrumentalanalytik”, *Das Papier* **12** (1999) 712–764.
5. W. W. Graessley, *Polymer* **21** (1980) 258.
6. W.-M. Kulicke, et al.: “Characterization of aqueous carboxymethylcellulose solutions in terms of their molecular structure and its influence on rheological behaviour”, *Polymer* **37/13** (1996) 2723–2731.
7. R. Dönges, *Das Papier* **12** (1997) 653–670.
8. P. Käuper, et al.: “Development and evaluation of methods for determining the pattern of functionalization in sodium carboxymethylcelluloses”, *Angew. Makromol. Chem.* **260** (1998) 53–63.
9. P. Mischnick, et al.: “Structure Analysis of 1,4-Glucan Derivatives”, *Macromol. Chem. Phys.* **201** (2000) 1985–1995.
10. P. W. Arisz: “Substituent distribution along the cellulose backbone in O-methylcelluloses using GC and FAB-MS for monomer and oligomer analyses”, *Carbohydr. Res.* **271** (1995) 1–14.
11. B. Reiche, H. Jehring, H. Dautzenberg, B. Philipp: “Elektrosorptionsuntersuchungen an Carboxymethylcellulosen”, *Faserforschung und Textiltechnik* **29/5** (1978) 324–328.
12. “Cellulose”, in: *Ullmann*, 6th ed., vol. **6**, WILEY-VCH Verlag GmbH, Weinheim 2002, pp. 593–645.
13. O. Wurz: *Celluloseäther - Herstellung und Anwendung*, Roether, Darmstadt 1961.
14. Kalle AG, DE 1543136, 1969 (F, Eichenseer, S. Janocha, H. Macholdt).
15. Dow Chemical Company, US 4015067, 1975 (G.Y.T Liu, C.P. Stange).
16. Henkel KgaA, DE 2929011, 1981 (W. Willi, H. Leischner, W. Rähse, F.-J. Carduck, N. Kühne).
17. H. M. Spurlin, *J. Am. Chem. Soc.* **61** (1939) 2222.
18. J. Voss: “Celluloseäther”, in *Kunststoffhandbuch*, vol. **3**, Hanser, München 1965 pp. 349–397.
19. R. N. Halder, W. F. Waldeck. F. W. Smith, *Ind. Eng. Chem.* **44** (1952) 2803.
20. V. Stigsson, G. Kloow, U. Germgård, *PaperAsia* **17/10** (2001) 16–21.
21. R. Dönges, *Das Papier* **12** (1997) 653–670.
22. T. Heinze, T. Liebert, P. Klüfers, F. Meister: “Carboxymethylation of cellulose in unconventional media”, *Cellulose* **6** (1999) 153–165.
23. P. Käuper: “Natriumcarboxymethylcellulose: Bestimmung der chemischen Struktur, der Lösungsstruktur und der Elektrolytwechselwirkungen”, *Berichte aus der Chemie* (1998), Shaker, Aachen.

24. "Dry Mortars", in: *Ullmann*, 6th ed., vol. **11**, WILEY-VCH Verlag GmbH, Weinheim 2002, pp. 83–108..
25. R. Dönges, *Br. Polym. J.* **23** (1990) 315–326.
26. K. G. Rosell,; *J. Carbohydr. Chem.* **7** (1988) 525–536.
27. Henkel & Cie GmbH, GB 754876, 1953.
28. Britisch Celanese Ltd., GB 909039, 1960 (E. W. Hitchin, H. Bates).
29. G. M. Powell: "Hydroxyethylcellulose", in *Water-Soluble Gums Resins*, McGraw-Hill, New York 1980.
30. L.-M. Zhang, *Carbohydr. Polym.* **45** (2001) 1–10.
31. M. G. Wirick, *J. Polym. Sci.* **6** (1968) 1705–1718.
32. P. W. Arisz, H. T. T. Thai, J. J. Boon, W. G. Salomon, *Cellulose* **3** (1996) 45–61.
33. Hercules Powder Company, US 688486, 1950 (E.D. Dlug, H.G. Tennent).
34. R. W. Butler, E. D. Klug: "Hydroxypropylcellulose", in *Water-Soluble Gums Resins*, McGraw-Hill, New York 1980.
35. Hercules Powder Company, US 3479190, 1969 (A. J. Ganz).

Cement and Concrete

Siegbert Sprung, Forschungsinstitut der Zementindustrie, Düsseldorf, Federal Republic of Germany (Chap. 1)

Jörg Kropp, Institute for Building Materials, Hochschule Bremen – University of Applied Sciences, Bremen, Federal Republic of Germany (Chap. 2)

1. Cement . . . 3
1.1. History . . . 3
1.2. Cement Constituents . . . 5
1.2.1. Portland Cement Clinker . . . 5
1.2.2. Granulated Blast-Furnace Slag (S) . 7
1.2.3. Pozzolanic Materials (P, Q) . . . 8
1.2.4. Fly Ash (V, W) . . . 9
1.2.5. Burnt Oil Shale (T) . . . 9
1.2.6. Limestone (LL, L) . . . 10
1.2.7. Silica Fume (D) . . . 10
1.2.8. Other Constituents . . . 10
1.3. Types of Cement . . . 11
1.3.1. Standardized Common Cements . . 11
1.3.2. Standardized Special Cements . . . 11
1.3.3. Other Cements . . . 13
1.4. Clinker Production Process . . . 13
1.4.1. Raw Materials . . . 13
1.4.2. Mining of Raw Materials . . . 14
1.4.3. Grinding, Mixing, and Homogenization of Raw Materials . . . 14
1.4.4. Fuels . . . 15
1.4.5. Burning and Cooling . . . 16
1.4.5.1. Technology . . . 16
1.4.5.2. Energy Consumption . . . 19
1.4.5.3. Thermochemistry of the Burning Process . . . 20
1.4.6. Characterization of Cement Clinker 21
1.5. Cement Production Process . . . 24
1.5.1. Storage . . . 24
1.5.2. Grinding and Blending . . . 24
1.6. Hardening of Cement . . . 26
1.6.1. Hydration Products . . . 26
1.6.2. Course of Hydration . . . 27
1.6.2.1. Water Demand . . . 27
1.6.2.2. Setting . . . 28
1.6.2.3. Hardening . . . 29
1.6.3. Heat of Hydration . . . 32
1.7. Hardened Cement Paste . . . 32
1.7.1. Binding of Water . . . 32
1.7.2. Structure . . . 33
1.7.3. Composition of Pore Solution . . . 34
1.7.4. Trace Elements and Environmental Compatibility . . . 35
1.7.5. Strength . . . 35
1.7.6. Deformation . . . 37
1.8. Testing of Cement . . . 39
1.9. Environmental Protection . . . 41
1.9.1. Particulate Emissions . . . 41
1.9.2. Gaseous Emissions . . . 42
1.9.3. Noise and Vibration . . . 44
1.10. Toxicology and Occupational Health . . . 44
1.11. Economic Aspects . . . 45
2. Concrete . . . 46
2.1. Introduction . . . 46
2.2. Materials . . . 46
2.2.1. Cement . . . 46
2.2.2. Aggregates . . . 47
2.2.3. Water . . . 48
2.2.4. Additives . . . 48
2.2.4.1. Setting and Hardening Additives . . 48
2.2.4.2. Workability Additives . . . 48
2.2.4.3. Porosity Additives . . . 48
2.2.4.4. Other Additives . . . 49
2.2.4.5. Fine-grained Material and Polymers 49
2.3. Production . . . 49
2.3.1. Definition of Concrete Properties . 49
2.3.2. Proportioning of Materials . . . 50
2.3.3. Production Process . . . 51
2.4. Fresh Concrete . . . 52
2.4.1. Workability . . . 52
2.4.2. Placement, Consolidation, and Finish . . . 53
2.4.3. Curing . . . 53
2.4.4. Special Methods . . . 54
2.5. Mechanical Properties . . . 55
2.5.1. Crack Development . . . 55
2.5.2. Strength . . . 55
2.5.2.1. Compressive Strength . . . 55
2.5.2.2. Tensile Strength . . . 56
2.5.2.3. Fracture Mechanics . . . 57
2.5.3. Deformation Characteristics . . . 57
2.5.3.1. Relationship of Stress to Strain . . . 57
2.5.3.2. Modulus of Elasticity and Poisson's Ratio . . . 58
2.5.3.3. Creep . . . 58

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_489

2.5.3.4. Shrinkage and Swelling 59
2.6. Physical Properties 59
2.6.1. Density 59
2.6.2. Porosity 59
2.6.3. Thermal Conductivity 60
2.6.4. Electrical Conductivity 60
2.6.5. Permeability 60
2.6.6. Thermal Expansion 61
2.6.7. Shielding Properties 61
2.7. Durability 61
2.7.1. Chemical Attack 62
2.7.2. Physical Attack 62
2.7.3. Corrosion of the Reinforcement . . 63
2.8. Special Concretes 63
2.8.1. Lightweight Concrete 64
2.8.2. Heavyweight Concrete 64
2.8.3. Massive Concrete 64
2.8.4. Fiber-Reinforced Concrete 64
2.8.5. Polymer Concrete 65
2.9. Mortar 65
3. References 65

1. Cement

Cement is a hydraulic binder. It is a finely ground, nonmetallic, inorganic material. When mixed with water it forms a paste that sets and hardens by hydration. After hardening, it retains its soundness, strength, and stability, even under water. The hydraulic hardening is primarily due to the formation of calcium silicate hydrates. Accordingly, cements consist of components which react within a short period of time with the mixing water to form such calcium silicate hydrates in due quantity to guarantee strength and durability. Besides calcium silicates, smaller amounts of other components, for example, calcium aluminates, take part in setting and hardening reactions. Calcium aluminate cements harden hydraulically by forming mainly calcium aluminate hydrates.

1.1. History

The name cement in the sense of building materials originates from the Romans, who called a concrete-like brickwork, made from stones and a burnt lime binder, *opus caementitium*. Later, hydraulic binders which were obtained from a mixture of additives such as ground ceramic bricks, volcanic tuff, and burnt lime were given the names cementum, cimentum, caement, and cement. The notation cement, known from metallurgical processes, is not linked with building materials.

The importance of clay for the hardening properties of hydraulic lime that is burnt from a natural mixture of limestone and clay was discovered by John Smeaton (1724 – 1792). This hydraulic lime was used to produce a seawater-resistant mortar for the Edystone lighthouse near Plymouth. The later cement chemist W. Michaelis called this discovery the basis of modern knowledge on hydraulic mortars. In a similar way in 1796 James Parker burnt a hydraulic lime from natural London marl and called it Roman cement.

Louis-Joseph Vicat (1786 – 1861) in France and Johann Friedrich John (1782 – 1847) independently discovered that artificial mixtures of natural limestone and 25 – 30 % clay were most suitable for the production of hydraulic lime. Even the binder produced in 1824 by Joseph Aspdin by burning an artificial mixture of limestone and clay corresponded formally and according to modern definitions to Roman lime in its composition and properties because it had not been burnt to sintering. He called this binder Portland cement, since the artificial stones made from Portland cement were similar in color to Portland stone, an oolitic limestone that was quarried on the channel coast of the Portland peninsula in Dorsetshire. William Aspdin (1815 – 1864), the son of Joseph Aspdin, started the production of Portland cement in 1843 in his newly established plant in Rotherhithe, near London. His cement proved to be far superior to Roman cement, as was demonstrated in the construction of the Houses of Parliament in London. The reason for this improvement was a higher burning temperature, which led to partial sintering of the raw mix. William Aspdin is therefore regarded as the first producer of real Portland cement.

The importance of the sintering process was pointed out for the first time on a more scientific basis by Isaac Charles Johnson (1811 – 1911) in 1844. The first German Portland cement, based on the Aspdin patents,

was produced in Buxdehude near Hamburg in 1850. The foundations for the production of Portland cement in Germany on an industrial scale were laid by HERMANN BLEIBTREU (1824 – 1881), who built two cement plants, in Züllchow near Szczecin (1855) and in Oberkassel near Bonn (1858). The production of Portland cement in France started in 1850 based on the finding that the sintered residue obtained when burnt lime was hydrated resulted in a slow-setting binder after grinding.

In the United States, DAVID SAYLOR first produced sintered cement clinker in 1870. He homogenized the raw material by grinding and then formed bricks from the resulting powder which were burnt.

WILHELM MICHAELIS (1840 – 1911) played an important part in further developments. His book *The Hydraulic Mortar*, published in 1868, described for the first time details of the optimal composition of the raw material mixture and the first valuable information on the highest possible content of calcium oxide in the raw material mixture, that is, the amount which is able to bind SiO_2, Al_2O_3, and Fe_2O_3 at sintering temperatures (hydraulic modulus). Further knowledge on reactions during burning and cooling of cement clinker resulted among others from the studies of S. B. and W. B. NEWBERRY [12], E. WETZEL [13], E. SPOHN [14], F. M. LEA and T. W. PARKER [15], and H. KÜHL [16].

In the early stages, only the simple, discontinuously operating shaft kiln was available, but later the annular brick kiln was also employed to burn cement clinker. The term *cement clinker* originates from this time, because the material to be burnt in the annular kiln was formed into bricks, which were placed and burnt like ceramic bricks. The first rotary cement clinker kilns were operated in the United States in 1895 and in Germany in 1896. Later, the first grate preheater kiln was introduced in Germany by Polysius in 1929. The first cyclone or suspension preheater kiln of industrial scale was built in 1953 by KHD, also in Germany, based on a Danish patent from 1932.

The latent hydraulic properties of granulated, that is, rapidly cooled and largely glassy, blast-furnace slag were discovered by EMIL LANGEN in 1862, who found out that a mixture of ground granulated blast-furnace slag and burnt lime can attain considerable strength. In 1882 GODHARD PRÜSSING (1828 – 1903) was the first to add ground granulated blast-furnace slag to Portland cement. Cement with a lower slag content was designated ″iron Portland cement″ in Germany in 1901, and the cement with a higher content of slag has been called ″blast-furnace slag cement″ since 1907. In 1908, H. KÜHL discovered the sulfate activation of granulated blast-furnace slag, which was the basis of the production of ″supersulfated cement″. The first cement characterized by a higher early strength was produced in an Austrian cement plant in Lorüns (Vorarlberg) in 1912. This was a very finely ground Portland cement from clinker, obtained by burning a lime-rich mixture of raw materials at a higher temperature.

The first Portland cement with a high sulfate resistance was the ″Erzzement″ patented by Krupp-Grusonwerk in Magdeburg in 1901 and produced in the cement plant in Hemmoor near Hamburg. It contained a smaller amount of aluminum oxide and a large amount of iron oxide, and this resulted in an alumina modulus of only 0.3. The properties of this cement are similar to those of the so-called *Ferrari cement*, a Portland cement with an alumina modulus of 0.64, which was first produced in Italy in 1919. The importance of a high slag content for the sulfate resistance of blast-furnace cements was established between 1920 and 1930. White Portland cement was produced in small amounts in the years after 1880 in the Portland cement plant near Heidelberg, Germany. Later, other cement plants also started the production of this kind of cement.

The development of *oil-well cement* started about 1930, when deep wells became necessary for the recovery of crude oil. These types of cement, which set and harden very gradually, even at high temperatures and pressures, are required to line the oil wells. The first description of *expanding cement* appeared in 1920 [17]. The expanding cements produced today are based on the work of V. V. MIKHAILOV and A. KLEIN [18]. The Portland Cement Association of the United States developed *regulated-set cement* [19]. A similar cement is the *jet cement* produced in Japan [20]. *Calcium aluminate cement* was first produced in France during World War I, based on a patent awarded to the French chemist J. BIED in 1908. He discovered that crystallized melts with the composition of monocalcium aluminate can

harden hydraulically and attain especially high early strength.

The first cement standards were introduced in Germany in 1878. In France, the use of cement for the construction of government buildings was permitted in 1885. The cement standards in the United States and Great Britain were established in 1904. The European Standard for Common Cements will be introduced in 2001/2002.

1.2. Cement Constituents

The production of cement includes extraction and preparation of the raw materials, burning of the raw materials mixture to give cement clinker, preparation of the different main components of the cement, and grinding of the cement components together with calcium sulfates to regulate setting. These continuous processes demand the monitoring of large material flows with respect to amount and required composition. This procedure is necessary to compensate variations of the raw materials composition or deviations caused by the process of clinker burning and cement grinding. Quality management systems according to DIN-ISO 9000 and DIN-ISO 14 000 support the effort to achieve a high level of product quality and uniformity, low energy consumption, and minimized emissions.

1.2.1. Portland Cement Clinker

Portland cement clinker consists mainly of tricalcium silicate, dicalcium silicate, tricalcium aluminate, and calcium aluminoferrite. It is made from a mixture of raw materials which normally contains limestone, clay, sand, iron ore, and other components in definite proportions. When this mixture is heated to the sintering temperature range, new compounds form, the clinker phases. Table 1 shows their proportions.

The terms alite and belite were introduced by A. E. Törnebohm, who examined clinker microscopically in 1897 [21]. He used the first letters of the alphabet to designate the main components because he did not know their composition. These names are still used to distinguish between pure silicates and silicate phases in clinker, which always incorporate small amounts of aluminum, iron, magnesium, alkali metals, and trace elements.

Tricalcium silicate (C_3S) (see Fig. 1) is the compound with the highest CaO content in the binary system $CaO-SiO_2$. C_3S decisively determines most of the properties of cement. It hardens quickly and attains a very high strength when it is finely ground and mixed with water to give a paste. It is formed by the solid-state reaction between calcium oxide and silica and can be produced from pure limestone and quartz. The starting materials must be very finely ground and burnt at extremely high temperatures. The reaction, however, proceeds more rapidly in the presence of a melt consisting of calcium, aluminum, and iron(III) oxides at about 1450 °C. For this reason, raw materials used in the production of Portland cement clinker contain predominantly calcium oxide and silica and additionally the necessary amounts of aluminum and iron(III) oxide to form a melt proportion of about 20 – 25 %.

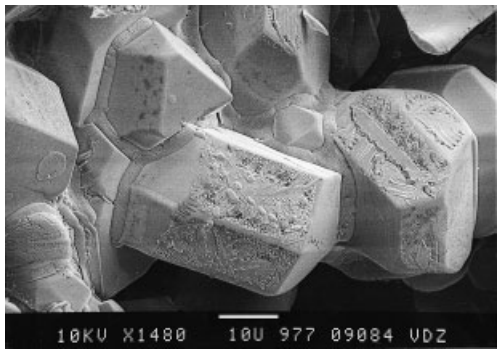

Figure 1. Scanning electron micrograph (SEM) of idiomorphic alite crystals embedded in an undifferentiated matrix of Portland cement clinker

Pure tricalcium silicate melts incongruently at 2070 °C, releasing calcium oxide [22]. Tricalcium silicate is metastable below 1250 °C and decomposes to calcium oxide and dicalcium silicate. However, decomposition occurs only at very slow cooling rates and/or when Fe^{2+} instead of Fe^{3+} is present in the lattice [23]. Therefore, cement clinker that contains iron oxide should be burnt and cooled under oxidizing conditions. In the metastable temperature region below 1250 °C, six polymorphic modifications of tricalcium silicate have been identified [24].

Table 1. Phase composition of Portland cement clinker [11]

Constituent	Formula (abbreviation)	Content, wt % Range	Average
Tricalcium silicate	$3\,CaO \cdot SiO_2$ (C_3S)	46 – 79	61
Dicalcium silicate	$2\,CaO \cdot SiO_2$ (C_2S)	5 – 30	15
Calcium aluminoferrite	$2\,CaO \cdot (Al_2O_3, Fe_2O_3)$ [$C_2(A, F)$]	4 – 16	8
Tricalcium aluminate	$3\,CaO \cdot Al_2O_3$ (C_3A)	6 – 18	12
Free calcium oxide	CaO (C)	0.1 – 4	
Free magnesium oxide	MgO (M)	0.7 – 1.5	1.5

These are stabilized by solid solutions of such ions as Mg^{2+}, Zn^{2+}, Al^{3+}, and Fe^{3+}.

Dicalcium silicate occurs when cement clinker is not completely saturated with calcium oxide. Pure dicalcium silicate melts congruently at 2130 °C [22]. In lime-rich raw mixes, C_2S forms already at temperatures below 1000 °C by solid-state reaction. During the sintering process in the presence of a melt above 1250 °C, C_2S incorporates additional Ca^{2+}, resulting in the formation of C_3S. It crystallizes in four polymorphic modifications which are known under the designations α, α', β, and γ [24]. The β modification is metastable at all temperatures. It is converted to the less dense γ modification at below 500 °C. This conversion is the reason why a cement clinker rich in C_2S may disintegrate at low cooling rates. In industrial clinkers, however, burnt above 1200 °C, disintegration generally does not take place because the different high temperature modifications [24] of C_2S are stabilized by incorporation of, e.g., Al^{3+}, Mg^{2+}, P^{3+}, Na^{+}, K^{+} etc. The type of modification present in industrial clinkers additionally depends on the cooling rate. Commonly, C_2S exists in its β modification besides the α and α' modifications.

The high-temperature modifications of C_2S have hydraulic properties, in contrast to γ-C_2S. In cement, however, the hydration reaction of C_2S is slowed down significantly as compared to C_3S.

Calcium aluminoferrite [C_2(A,F)] contains most of the iron and part of the aluminum present in cement clinker. It is a phase in the limited solid solution series $2\,CaO \cdot (Al_2O_3, Fe_2O_3)$, with $2\,CaO \cdot Fe_2O_3$ (dicalcium ferrite) and $2\,CaO \cdot (0.69\,Al_2O_3, 0.31\,Fe_2O_3)$ as its limits. In common cement clinker, the composition of calcium aluminoferrite roughly corresponds to the formula $4\,CaO \cdot Al_2O_3 \cdot Fe_2O_3$. It can incorporate up to 2 wt % of magnesium oxide in its crystal lattice. This causes a change in color from brown to gray, the color of common Portland cement [25]. In clinker which is burnt under reducing conditions and which contains 0.2 – 0.3 % Fe^{2+} as a result, the remaining C_2(A, F) phase shows a brown color. Possibly, Fe^{2+} is bound in the lattice instead of Mg^{2+} [167]. Calcium aluminoferrite contributes little to the hydraulic hardening process. The hydraulic reactivity decreases with an increasing Fe_2O_3 content of C_2(A, F).

Tricalcium aluminate (C_3A) is the compound with the highest CaO content among the aluminate phases. C_3A contains the aluminum oxide that is not combined in calcium aluminoferrite. It melts incongruently, releasing calcium oxide. Pure C_3A crystallizes in cubic form. The lattice can incorporate different ions in solid solution, e.g., Fe^{3+}, Mg^{2+}, Si^{4+}, or K^{+} and Na^{+}. Due to the incorporation of alkali metals, the symmetry of crystallization changes from cubic to orthorhombic and monoclinic, and this results in a higher hydraulic reactivity as compared to cubic C_3A. Consequently, the technical properties of Portland cement clinker are changed (Section 1.6). Altogether, C_3A reacts rapidly with water, but its hydraulic properties are not very pronounced. However, together with silicates, it improves the initial strength of cement.

Free Lime (CaO) and Periclase (MgO). Minor components of cement clinker are free calcium oxide (free lime) and free magnesium oxide (Periclase). They react with water to form calcium hydroxide and magnesium hydroxide, which occupy more space than the original oxides. Therefore, free calcium oxide and free

magnesium oxide may cause expansion when they are present in larger amounts as coarsly crystalline components. In this case the reaction with water proceeds very slowly and still continues when the paste has already started to harden. Hence, all cement standards contain appropriate specifications to limit CaO and MgO contents so that damaging expansion does not occur and mortar and concrete keep their soundness.

Alkali metal compounds, such as alkali metal sulfates and alkali-metal-containing calcium aluminate, are also minor components of cement clinker which originate mainly from raw materials and fuel ashes. Clinker contains up to 2.0 wt % alkali metal oxides ($Na_2O + K_2O$) and up to 2.0 wt % sulfate (SO_4^{2-}). Molten alkali metal sulfate forms at sintering temperatures. It does not mix with the aluminoferrite melt in the clinker. Solid solutions of alkali metal sulfate crystallize during cooling with a composition that depends on the K_2O/Na_2O ratio. Components such as $2\,CaSO_4 \cdot K_2SO_4$ can also be formed. The molar ratio of alkali metals and sulfate in the clinker is designated as the degree of sulfatization. If an excess of alkali metal is present, alkali-containing calcium aluminate forms with a composition and crystalline structure similar to that of tricalcium aluminate [26]. The total amount of alkali metals and the nature of their binding in the clinker can significantly affect the setting and hardening of cement.

Other compounds which may occur in common Portland cement clinker generally do not have a significant influence on cement properties. Special cements may contain, e.g., calcium aluminates of different composition which determine the properties of the product in a characteristic way (see Section 1.3.3) and [164].

1.2.2. Granulated Blast-Furnace Slag (S)

Processing. Blast-furnace slag is obtained from nonferrous components of iron ore, coke ash, limestone, and other additives during the smelting of iron ore in a blast furnace. It is removed from the furnace separately or together with molten pig iron at temperatures of approx. 1350 – 1550 °C. It solidifies at about 1250 °C. Glassy blast-furnace slag is produced by rapidly cooling the molten slag with water and/or compressed air to get a finely dispersed material (granulation). The glass content, which is decisive for the hydraulic properties, must be kept as high as possible. Normally, the glass content is higher than 90 %. Standards require a glass content of at least two thirds of the mass [197]. Figure 2 shows the typical morphology of glassy grains.

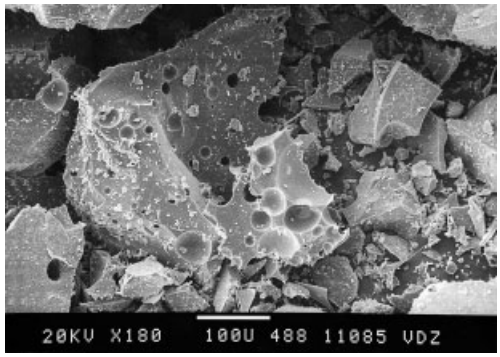

Figure 2. Particles of rapidly cooled glassy blast furnace slag (SEM)

The moist granulated slag is stored for a limited time in large drained stockpiles to lower the moisture content and to homogenize different furnace charges. The granulated slag retains at least 7 – 20 % residual moisture and is dried in different kinds of drying equipment before grinding. In some cases it is dried while grinding cement by using the heat produced during grinding and/or hot gas. Altogether, the grinding of cements containing blast-furnace slag requires more energy than the grinding of Portland cement [165] to achieve a cement of a similar high efficiency as Portland cement.

Properties. In contrast to slowly cooled crystalline slag, glassy blast-furnace slag is a latent hydraulic material. It must be activated if it is to harden hydraulically in a commercially acceptable period of time. The activator in the case of Portland slag cement and blast-furnace cement is Portland cement clinker (clinker activation). In the case of supersulfated cement, it is anhydrite or gypsum (sulfate activation).

The hydraulic properties of glassy blast-furnace slag depend on the microscopically determined glass content and on its chemical composition [33]. Table 2 gives the chemical composition of blast-furnace slags used in Europe in

Table 2. Concentrations of main and minor chemical components in granulated blast-furnace slag, natural pozzolana, siliceous fly ash, and burnt oil shale [167]

Component	Granulated blast-furnace slag	Natural pozzolana	Siliceous fly ash (Germany)	Burnt oil shale
SiO_2	30 – 39	45 – 89	35 – 53	12 – 51
Al_2O_3	9 – 18	3 – 20	21 – 30	5 – 16
TiO_2	0.3 – 1.4		0.7 – 1.5	
P_2O_5	0.1 – 0.4		0.1 – 0.8	
Fe_2O_3	0.1 – 1.0 ∗∗	1 – 10	6 – 12	6 – 7
Mn_2O_3	0.2 – 3.0 ∗∗		0.1 – 0.4	
CaO	33 – 48	2 – 9	0.5 – 10	18 – 60
MgO	2 – 13	1 – 5	2 – 5	1 – 4
Na_2O	0.2 – 1.2	1 – 7	0.4 – 2.0	1 – 2
K_2O	0.4 – 1.3	1 – 11	1.0 – 5.0	1 – 2
SO_3	0.0 – 0.2	0.0 – 0.8	0.3 – 1.5	5 – 10
S^{2-}	0.5 – 1.8			

∗ The components are usually given as simple oxides. ∗∗ Calculated from FeO and MnO.

the production of cement [167]. In other countries the chemical composition may differ to a certain extent.

In general, the hydraulic reactivity of glassy blast-furnace slag is proportional to the amounts of calcium oxide and magnesium oxide. Whereas CaO increases the reactivity, MgO diminishes the hydraulic properties because of the formation of crystalline spinel ($MgO \cdot Al_2O_3$). A higher content of aluminum oxide especially increases the initial strength. The sulfide content promotes glass formation because the larger sulfide ions are incorporated into the glass structure rather than the smaller oxide ions.

Simple formulas, based on chemical composition, are used to evaluate the hydraulic properties of glassy slag. The European cement standard, for example, requires the ratio of $(CaO + MgO)/SiO_2$ to exceed 1.0. However, these formulas can only be applied to slag of the same origin. Moreover, the activating effect of the clinker and the curing conditions must be considered. Therefore, strength achieved by mortar test specimens made from slag-containing cements, considering the fineness, the particle size distribution, the proportion of the components, and the SO_3 content [167] is at present the only reliable means of evaluating the hydraulic properties of glassy blast-furnace slag. Additionally, the strength of mixtures of Portland cement with granulated blast-furnace slag is evaluated by comparison with mixtures containing an inert material of the same fineness, e.g., finely ground quartz [34], [35].

1.2.3. Pozzolanic Materials (P, Q)

Pozzolanas are naturally occurring siliceous or silico-aluminous fine-grained materials which at normal temperatures can only harden in the presence of water and calcium hydroxide. Major components of pozzolana are more or less reactive silica and various aluminosilicates. The proportion of reactive CaO is negligible. Products of the pozzolanic reaction are calcium silicate hydrates as strength-giving constituents. Additionally, calcium aluminate hydrates, obtained from the reaction of calcium hydroxide solution with reactive aluminum oxide, can influence the hydraulic hardening and the product properties. According to the European standard specifications [197], the reactive silica content must not be less than 25 %.

These properties are primarily exhibited by volcanic tuff which originates from volcanic ash deposits. The natural pozzolana (P) is named after the tuff deposit in Pozzuoli, near Naples, Italy, which has been used since ancient times. Even today, the volcanic tuff found in Campagna and Latium is of considerable importance to the Italian cement industry. The Rheinish trass found in the Neuwied basin near Koblenz, Germany, is also volcanic tuff. The Suevit or Bavarian trass occuring in the Nördlinger Ries was originally sedimentary rock, transformed by the impact of a meteorite.

Sedimentary rocks containing larger amounts of reactive silica also belong to the group of natural pozzolanic materials. They only need to be dried before they can be used as main constituent

in the production of cement. Rocks with a high content of free silica of organic origin, such as the diatomites found in northern Germany and Denmark and French gaize, also belong to this category.

Natural calcined pozzolanas (Q) are of volcanic or of sedimentary origin (e.g., clay and shale). They are heated to 400 – 800 °C to increase their reactivity. This applies, e.g., to phonolite, an alkali-rich volcanic rock. The zeolite present in phonolite is converted to reactive silica when it is heated. Table 2 lists the chemical compositions of natural pozzolanas.

The reactivity of pozzolanic materials can be determined, as in the case of blast-furnace slag, by their contribution to the strength development of mortar test specimens [34–36]. The European cement standard requires a special test for pozzolanic cement, a procedure which has long been part of the Italian cement standard. This test shows that the pozzolanic portion of cement has the ability to bind a certain minimum amount of the calcium hydroxide released during hydration of the clinker portion. The lower the pozzolanic reactivity, the higher the content of pozzolanas must be to meet the standard requirements. However, the proportion in the cement is limited to meet the minimum requirements on strength and durability.

Rocks, generally also including tuff, are quarried by blasting and are then crushed to the size of gravel. Sedimentary rocks are directly removed from the quarry front. The materials are dried and, if necessary, heated in a kiln. They are then ground with Portland cement clinker and calcium sulfate to give pozzolanic cement.

1.2.4. Fly Ash (V, W)

Fly ash is a fine-grained residue from the combustion of pulverized coal. It is precipitated in the electrical or mechanical dust-collecting equipment of power plants. Fly ash may be of siliceous or calcareous nature. The former has only pozzolanic properties. The latter may additionally have hydraulic properties. The unburnt residue is normally limited to 5 wt % as loss on ignition. The fineness is in the range of 2100 – 6200 cm^2/g (Blaine). Siliceous fly ash (V) consists mainly of fine-grained, partly glassy particles, which are usually spherical but can also have an irregular shape and a porous structure. Fly ashes from wet-bottom boilers contain about 85 wt % and those from dry-bottom boilers less than 61 – 76 wt % of glass, with available Al_2O_3-contents of 14 – 25 wt %. Crystalline components are mullite, quartz, and magnetite [167], as well as small amounts of alkali metal chlorides and sulfates. Table 2 lists the chemical composition of siliceous fly ashes, and Figure 3 gives an example of its morphology. The properties of fly ash, especially the amount of unburnt or partially burnt coal, of quartz, and other crystalline components are mainly determined by the origin of the coal, the ash content and also by the firing and combustion conditions.

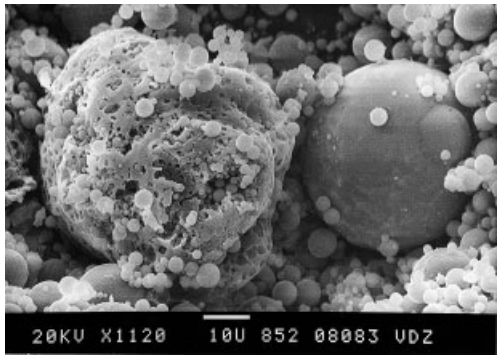

Figure 3. Morphology of siliceous fly ash (SEM)

Calcareous fly ash (W) having pozzolanic and/or hydraulic properties consists essentially of reactive CaO, SiO_2, and Al_2O_3. As defined by EN 197-1, the portion of CaO must not be less than 10 wt %, in contrast to siliceous fly ash. The portion of hydraulic compounds may increase with increasing CaO content. Therefore, calcareous fly ashes with more than 15 wt % of reactive CaO have to exhibit a compressive strength development in mortar (EN 196-1) of at least 10 MPa after 28 d. Calcareous fly ashes with CaO contents between 8 and 40 wt % which result from burning lignite from the Rhine area of Germany contain large amounts of free lime and anhydrite ($CaSO_4$) and therefore do not easily meet the requirements regarding soundness.

1.2.5. Burnt Oil Shale (T)

Burnt shale, specifically burnt oil shale, is another main constituent for the manufacture of

cement. Its use is dependent on the composition. In Germany, such material is found in the area northwest of the Swabian Mountains and is used for this purpose. This oil shale is a lime-containing bituminous shale with ca. 11 wt % of organic substances, 41 wt % calcium carbonate, 27 wt % clay minerals, and 12 wt % quartz [38]. It is burnt at 800 °C in a fluidized-bed furnace. The energy released when oil shale is burnt is utilized to generate electricity, and the ash is ground with cement clinker to give Portland/burnt (oil) shale CEM II/A/B-T cement and other binders. Oil shale can also be used as a raw material component in the clinker burning process.

The chemical composition of burnt oil shale is characterized by the figures listed in Table 2. Burnt oil shale itself is able to harden hydraulically according to the requirements of EN 197-1. This is due to its content of dicalcium silicate and various calcium aluminates. Mainly the high content of reactive silica additionally shows pronounced pozzolanic properties.

1.2.6. Limestone (LL, L)

Limestone, under practical conditions, acts as a filler in cement. A filler is defined as a cement component which does not participate, or only slightly participates, in the hardening reactions of the hydraulically active components. However, as a result of its physical properties, especially its particle size distribution, it improves the properties of fresh and hardened cement paste, mortar, or concrete. This effect proves to be the more significant, the steeper the particle size distribution of the Portland cement component. It is based on the filling out of hollow spaces between the cement particles. As a result, the particle size distribution of the filler-containing cement in the RRSB grid becomes broader. The water demand of the paste, mortar, and concrete is reduced, and the workability improved [168], [169]. This leads to a denser structure, and the filler can increase the strength to a limited extent without taking part in the hardening reactions [39]. For compensation of the diluting effect of amounts higher than, for example, 10 wt % of limestone filler, the Portland cement clinker portion in the cement should be ground to a higher fineness. The amount of filler should not exceed 20 wt %. The filler must not impair other properties; for example, it should not decrease the frost resistance of concrete made from this kind of cement. Therefore, the limestone employed has to meet requirements on $CaCO_3$ and TOC contents and on methylene blue adsorption [197].

1.2.7. Silica Fume (D)

Silica fume is a byproduct of the production of silicon and ferrosilicon alloys. It consists of very fine and mainly spherical particles with a BET specific surface area of at least 15 m^2/g but normally higher than 20 m^2/g. Consequently, silica fume has a low bulk density. This is the reason why this material is compacted or pelletized before transport or handled as slurry when used as a concrete additive.

The content of SiO_2 must not be less than 85 wt % [197]. Other components may be Al_2O_3, Fe_2O_3, MgO, and CaO, depending on the kind of production process. The SiO_2, especially that from ferroalloy production, has an amorphous structure. Therefore, silica fume commonly has a marked pozzolanic reactivity. It reacts with calcium hydroxide to form calcium silicate hydrates. Besides its use as a main constituent in cement, silica fume is also used as an additive in high-strength concrete. The extreme fineness of silica fume increases the water demand of paste, mortar, and concrete. Therefore, the employment of water-reducing admixtures is necessary.

1.2.8. Other Constituents

Minor additional constituents are defined as specially selected, inorganic, natural and mineral materials, derived from the clinker production process [197]. After appropriate preparation and on account of their particle size distribution they improve the physical properties of cement, such as water retention and workability. Their addition in limited to 5 wt %.

Additives for the purpose of improving the manufacture or the properties of cement must not exceed a total quantity of 1 wt % [197].

Calcium sulfate is added to cement and interground to control setting. Sulfates also act as activators for strength development. For reasons

of soundness the total sulfate content (SO_3) of cements must be less than 3.5 (4.0) wt %. Added calcium sulfate can be natural gypsum ($CaSO_4 \cdot 2\,H_2O$), natural anhydrite ($CaSO_4$), or a mixture thereof. Gypsum is partly dehydrated at mill temperatures above 80 °C to form hemihydrate ($CaSO_4 \cdot 0.5\,H_2O$). Calcium sulfate is also available as a byproduct of some industrial processes. Gypsum from flue gas desulfurization of power plants has proved suitable. For the reaction of water-soluble sulfates with calcium aluminate hydrates to adjust the setting time of cement and the workability (optimization), see Section 1.6.

1.3. Types of Cement

Different economic and industrial developments, sources of raw materials, and climatic conditions have led to different developments in building methods and construction materials and, therefore, to different types of cement. Hence, there have always been substantial differences in the national standards for composition, strength development, and other properties. With the elaboration of a European cement standard an attempt was made to adopt all cements described as traditional by the respective national standardization bodies, which have been produced and used for decades and which have proven durability performance in practice. Due to the large number of different cements involved it was considered necessary to separate "common cements" from "cements with special properties". Although the European standard for common cements has not yet been formally introduced by the national standardization bodies, the basic principles of the standard have already been adopted in most of the existing national standards.

1.3.1. Standardized Common Cements

The European cement standard EN 197-1/2 [197] exclusively contains cements which can be used for general purposes. Cements with special properties such as low-heat cements, sulfate-resistant cements, and calcium aluminate cements will be specified separately. On a national basis separate standards for different cement types as well as one standard for all cement types do exist. The standard EN 197-1 distinguishes between five main types of common cement:

- CEM I: Portland cement
- CEM II: Portland composite cement
- CEM III: Blast-furnace cement
- CEM IV: Pozzolanic cement
- CEM V: Composite cement

These five main types are subdivided into 27 different products (types of common cement; Table 3).

CEM I is a Portland cement with a clinker content of at least 95 wt %. The main type CEM II includes all composite cements which contain 6 – 20 (A) and 21 – 35 wt % (B) of one or more main constituents (see Section 1.2) besides clinker. The content of silica fume is limited to 10 wt %. CEM III designates all granulated cements containing blast-furnace slags with a slag content of 36 – 65 (A), 66 – 80 (B), and 81 – 95 wt % (C). CEM IV are Pozzolana cements with pozzolan (D, P, Q, V, W) contents of 11 – 35 (A) and 36 – 55 wt % (B). CEM V are composite cements containing granulated blast-furnace slag (S) and pozzolana (P, Q, V) in a total portion of 36 – 40 (A) and 62 – 80 wt % (B) as main constituents besides clinker.

1.3.2. Standardized Special Cements

European standards for cements with special properties are not yet available. Therefore, national regulations are still applicable.

Sulfate-resistant cements (SRC) are to be used for concrete structures which are exposed to the attack of soluble sulfates in ground waters, surface waters, and soils. Portland cements (CEM I) with a low C_3A content of less than 3 wt % and blast-furnace cements with at least 65 wt % of granulated blast-furnace slag are well tried in Europe and other parts of the world and meet the requirements for chemical resistance. Additionally, on a national basis, several cements of the main types CEM I – V with common composition are standardized as sulfate-resistant cements, seawater-resistant cements, or as medium sulfate-resistant cements. The suitability of these cements, however, depends on

regional environmental conditions. Without further experience based on long-term investigations, general adoption into a European regulation does not seem to be practicable.

Table 3. Cement types and composition specified in EN 197-1

Main cement types	Ordinary cement types	
CEM I	Portland cement	CEM I[a]
CEM II	Portland slag cement	CEM II/A-S[a] CEM II/B-S[a]
	Portland silica fume cement	CEM II/A-D
	Portland pozzolanic cement	CEM II/A-P[a] CEM II/B-P[a] CEM II/A-Q CEM II/B-Q
	Portland fly ash cement	CEM II/A-V[a] CEM II/B-V CEM II/A-W CEM II/B-W
	Portland shale cement	CEM II/A-T[a] CEM II/B-T[a]
	Portland limestone cement	CEM II/A-L CEM II/A-LL[a], [b] CEM II/B-L CEM II/B-LL
	Portland composite cement	CEM II/A-M CEM II/B-M[c]
CEM III	Blast-furnace cement	CEM III/A[a] CEM III/B[a] CEM III/C
CEM IV	Pozzolanic cement	CEM IV/A CEM IV/B
CEM V	Composite cement	CEM V/A CEM V/B

[a] Cement types standardized in DIN 1164-1. [b] Corresponds to the CEM II/A-L Portland limestone cement currently standardized in Germany with a limestone TOC content of ≤ 0.20 wt % by weight [c] Only CEM II/B-SV Portland fly ash slag cement is currently standardized in DIN 1164-1.

Low-heat cements (LH), used in massive concrete structures, are characterized by limitated exothermic heat release during hydration. The heat of hydration leads to an increase of concrete temperature which may cause mechanical tension and formation of cracks.

The heat of hydration is mainly estimated according to two methods, standardized on a European basis. The heat of hydration is measured by solution calorimetry [170] after hardening of the paste under strictly isothermal conditions. The heat development of a mortar is measured continuously in a calorimeter under semiadiabatic conditions [171]. Low-heat cements are required to release a heat of hydration of not more than 270 J/g of cement [166]. Further subdivisions of the LH standard are under discussion. Characteristic for the LH cements is a decreased early strength development. LH CEM cements will be grouped in the five main cement types CEM I to CEM V.

Low-alkali cements (NA) are used in the production of concrete which contains aggregates with alkali-reactive constituents and which is used under moist atmospheric conditions. Low-alkali cements are defined worldwide as cements with a maximum total alkali content of 0.60 wt %, calculated as Na_2O (equivalent), and in Germany as cements containing granulated blast-furnace slag (Table 4).

Table 4. Specification of low-alkali cements (DIN 1164, 1999)

Cement type	Slag content, wt %)	Na_2O (equiv), wt %
All		≤ 0.60
CEM II/B-S	21 – 35	≤ 0.70
CEM III/A	36 – 49	≤ 0.95
CEM III/A	50 – 65	≤ 1.10
CEM III/B	66 – 80	≤ 2.00

Calcium aluminate cement with an Al_2O_3 content between 36 and 55 wt % is a hydraulic binder which sets and hardens by formation of aluminate hydrates. The main constituent is monocalcium aluminate ($CaO \cdot Al_2O_3$, CA). Other mineral phases as minor constituents include calcium aluminate ferrite $2\,CaO \cdot (Al_2O_3,\ Fe_2O_3)$ ($C_2(A,F)$), dicalcium silicate $2\,CaO \cdot SiO_2$ (C_2S), and Gehlenite $2CaO \cdot Al_2O_3 \cdot SiO_2$ (C_2AS). The hydraulic properties are predominantly based on the hydration of CA. High early strength development of more than 20 N/mm^2 after 6 h and more than 40 N/mm^2 after 24 h together with normal setting behavior ($>$ 60 min) and a high chemical resistance of dense mortars and concretes are the main characteristics of this cement. Increasing Al_2O_3 contents result in a rising refractory resistance. The kind of hydration products and the structure formed are dependent on the temperature during hydration. Hydration below

25 °C leads to metastable hexahydrates CAH_{10}, C_2AH_8, and Gibbsite (AH_3). In course of time CAH_{10} and C_2AH_8 transform into stable cubic hydrogarnet C_3AH_6 with an increase of porosity. Therefore, future specifications [172] will require a hydration temperature of more than 40 °C to achieve the direct formation of stable hydrogarnet and to avoid the porosity increase by conversion.

1.3.3. Other Cements

There are many different cement types with individual properties which are often used for special purposes only. In most cases they are not standardized but may be introduced into the market on the basis of national approval. Some examples are given below.

Approved cements for the production of *shotcrete* are mainly fine-ground Portland cements with or without addition of small amounts of calcium sulfate. The initial setting time after water addition is required to be less than one minute and may be regulated by calcium sulfate additions. Water-repellent additives are used to avoid prehydration and strength loss of the cement caused by moist aggregates. The use of soluble alkali metal silicates (water glass) as accelerator is questionable with regard to a damaging alkali/silica reaction.

Supersulfated cement consists of blast-furnace slag, at least 5 % anhydrite ($CaSO_4$), and a maximum of 5 wt % Portland cement clinker.

The use of *water-repellent cement* is advantageous, for example, in soil compaction. Water-repellent cement is a common Portland cement which contains small amounts of water-repellent substances added during grinding at elevated temperatures.

Oil-well cements are used for cementing the steel casing of gas and oil wells to the walls of the bore-hole and to seal porous formations. They are Portland and pozzolanic cements which set even at high temperatures and high pressure because of their composition.

Regulated-set cement, known in Japan as *Jet cement* and as *Schnellzement* in Germany, sets quickly and hardens very rapidly because of its high content of the aluminate compound $11\,CaO \cdot 7\,Al_2O_3 \cdot CaF_2$.

Expanding cement generally contains large amounts of aluminates and sulfates. The cement expands during hardening, by formation of ettringite.

Masonry cements are factory-produced binders for brickwork and plasterwork mortar with an adequate hydraulic strength development. Masonry cement is usually a mixture of Portland cement clinker, crushed rocks, and air entraining agents. These binders will be specified in a European Standard.

1.4. Clinker Production Process

1.4.1. Raw Materials

Portland cement clinker is the product which arises during the sintering process from a mixture of different raw material components and fuel ash. Both are fed to the kiln to give a mixture with the desired content of the main constituents CaO, SiO_2, Al_2O_3, Fe_2O_3, and MgO in the clinker. Attention also must be paid to minor constituents such as alkali metals, sulfates, chlorides, and trace elements.

The essential CaO-containing natural raw materials are limestone and chalk. The components SiO_2, Al_2O_3, and Fe_2O_3 are provided by addition of clay and quartz sand. The raw mix for cement clinker should contain 75 – 79 wt % $CaCO_3$ and 21 – 25 wt % clay. A natural mixture of the two is marl. Quality reasons allow only a small deviation in the total composition of the raw materials, including fuel ash. The target raw meal composition is calculated according to a multicomponent equation system using the hydraulic, siliceous, and aluminous deviations from given composition requirements (see Section 1.4.6).

Secondary raw materials can be employed if they meet the requirements on composition and environmental compatibility of the product, on other products properties, and on emission limits. Provisions must be made especially for the input of heavy metals. Therefore, quality assurance systems for the materials and the production process are advantageous. Examples for applicable secondary raw materials, grouped according to their main element, are given in the following [173]:

1) Ca group
   - Industrial lime
   - Lime sludge

2) Si group
   Used foundry sand
3) Fe group
   Roasted pyrite
   Synthetic hematite
   Red mud
4) Si – Al – Ca group
   Fly ashes
   Slags
   Crushed sand
5) S group
   Gypsum from flue gas desulfurization
   Chemical gypsum
6) F group
   CaF_2-containing filter mud

In the combined production of Portland cement and sulfuric acid (Müller – Kühne process), the mixture of raw materials may consist of anhydrite ($CaSO_4$), gypsum ($CaSO_4 \cdot 2\,H_2O$), clay, coke, other sulfite-, sulfate-, SiO_2-, Al_2O_3-, and Fe_2O_3-containing secondary raw materials, and fuels. The sulfur dioxide in the exhaust gas is converted into sulfuric acid by the usual processes (→ Sulfuric Acid and Sulfur Trioxide).

1.4.2. Mining of Raw Materials

Limestone, marl, and argillaceous limestone are usually quarried by blasting. The rocks are loaded onto heavy trucks by high-power excavators or bucket loaders and are transported to crushers of different construction. They are crushed in one or two stages to a particle size of < 30 mm. Mobile crushers that can follow the excavators along the working front are also employed. Soil type materials are torn away without blasting by caterpillars with bucket teeth and pushed toward the excavators. Chalk, marl, and clay are directly removed from the quarry site by using bucket-wheel or chain-bucket excavators.

The main prerequisite for the quality and uniformity of cement is that the composition of the raw materials before entering the kiln system meets the demands only with minor deviation over time. If the chemical composition of the different kinds of raw material varies greatly, then the broken raw material is often prehomogenized in rectangular or circular blending beds [5]. The blending beds generally contain a one-week supply of raw material, that is, 20 000 – 80 000 t. The material is subsequently reclaimed in the transverse direction of the pile. Variations in the deposits can be compensated by this measure to a blending efficiency of 70 – 90 %. By means of samples from the exploration of the deposit, samples taken during drilling, and samples taken during the build-up of the blending bed, the required average chemical composition can be regulated, following schemes which are optimized for each plant.

1.4.3. Grinding, Mixing, and Homogenization of Raw Materials

The raw materials are further processed to form a dry powder or a slurry.

Dry processing requires the fine grinding of the raw material constituents. The streams of raw materials are regulated by various kinds of dosage equipment. If necessary, corrective components such as quartz sand and iron ore can be added to obtain the desired chemical composition of the raw meal. During the grinding process, the mill feed is dried by hot gas, which is drawn through the mill in parallel flow. The waste heat of the exhaust gases from the kiln and/or exhaust air from the cooler is used for this purpose. Depending on the construction of the mill, raw material containing 8 – 12 wt % moisture can be dried. Supplementary firing is required in the case of a higher moisture content. Extremely moist raw materials are dried in a rotary dryer before they are ground. The residual moisture content in the raw meal must be less than 0.5 wt %.

Tube mills, high-pressure grinding rolls in combination with a deglomerator or a tube mill, and roller mills are employed to grind the raw materials. In tube mills, the material is ground by impact and friction with steel balls, in grinding rolls by high pressure, and in roller mills it is comminuted by pressure on a rotating pan by rollers. All grinding systems are equipped with separators and operate in a closed circuit. The separated coarse portions of the raw materials are fed back into the circuit. Roller mills require less energy and have a larger drying capacity than tube mills, but they are of limited use for very abrasive materials. Tube mills for raw meal have a diameter of up to 6 m and a capacity of up to 400 t/h. Their specific energy requirements

are around 12 – 16 kW h/t. Roller mills and high-pressure grinding rolls are built with throughput capacities of up to 400 t/h and specific energy requirements of about 10 – 14 kW h/t. High-pressure grinding rolls need a feed with a relatively low moisture content.

Wet processing is especially effective when the raw materials contain more that 20 wt % of water. The mixture of raw materials is ground in a tube mill with water addition to form a slurry.

The finished raw meal or slurry is continuously analyzed. Variations in chemical composition are compensated by adjusting the proportions of the constituents with correcting materials. Any remaining variations are equalized in homogenizing silos, which can hold the 8 – 10 h throughput of a mill. Homogenization is carried out by fluidizing the raw material powder with air and circulating it in a discontinuously operated fluidized bed. Alternatively the raw meal can be mechanically circulated through several silos or through a continuous blending silo. Raw materials slurry is homogenized by mechanical agitation or by using compressed air. Fluidizing and circulation of raw meal in discontinuous blending silos is the most effective method of homogenization, however, the power consumption is relatively high.

1.4.4. Fuels

Fuels for burning cement clinker are primarily hard coal, lignite, and petroleum coke, besides secondary fuels. Other fuels have only minor importance. The proportion of the different kinds of fuel is shown in Figure 4 [174]. The proportion, however, may vary from country to country, depending on the industrial and economic situation.

The substitution of fossil fuels by secondary fuels in general shows an increasing tendency. Suitable secondary fuels include used tires, waste oil, plastics, scrap wood, and burnable fractions from municipal waste and other materials. In general the energy content of common and secondary fuels is used in the clinker burning process in the sense of a materials-conversion process. As in the case of natural fuels, the mineral ashes are incorporated completely and become a part of the cement clinker. Therefore, cement clinker production ensures that both the energy and mineral content of waste materials are recovered in the rotary kiln. The simultaneous recovery of energy and ash distinguishes the clinker-burning process from combustion processes with the aim of energy conversion.

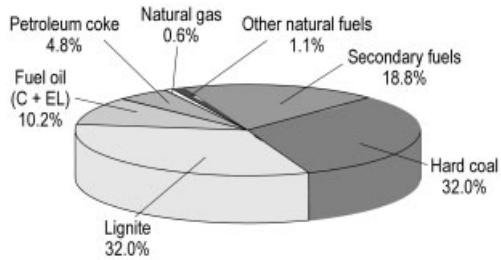

Figure 4. Proportions of the different kinds of fuels used for Portland cement clinker burning in 1998 (total consumption of thermal energy: 100.7×10^6 GJ) [C: Bunker C viscosity fuel oil; EL: Super light fuel oil]

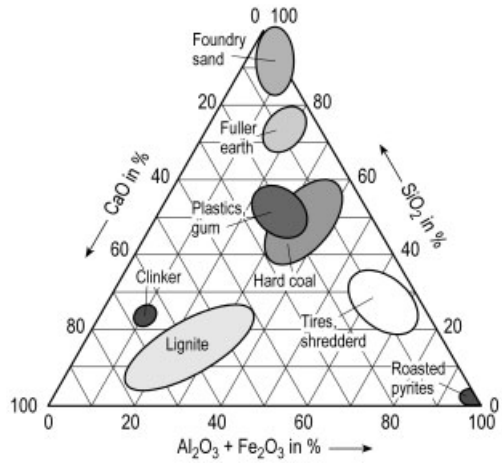

Figure 5. Fuel-ash and clinker composition in the system $CaO - (Al_2O_3 + Fe_2O_3) - SiO_2$

The differences between clinker and ash composition are demonstrated in Figure 5 [175]. The calorific value and the ash composition of natural and secondary fuels vary in a wide range [164]. Depending on the kind, composition and portion of the ash component the raw feed composition has to be corrected properly to achieve unchanged clinker composition and hydraulic properties of the cements. Ashes rich in SiO_2 and Al_2O_3 need the existence and use of limestone deposits with $CaCO_3$ contents of more than 85 wt %.

1.4.5. Burning and Cooling

1.4.5.1. Technology

The production of cement belongs to the group of energy intensive industrial production processes. About 70 % of the total energy consumption is needed as thermal energy and 30 % as electrical energy (Fig. 6). Following the official federal statistics, in Germany the portion of energy costs of the net production cost value amounted to 24 % in 1998 [167]. Therefore, it was, and still is, necessary to decrease the energy consumption. This economic pressure has led to a marked change of technology in cement production, especially in the clinker burning process (see Section 1.4.5.2).

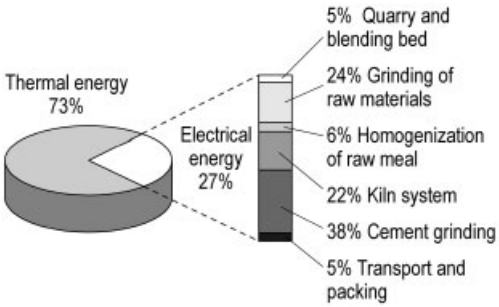

Figure 6. Proportion of thermal and electrical energy consumption in the cement production process (example)

In Europe, cement clinker is predominantly burnt in rotary kilns; shaft kilns are rarely used. *Rotary kilns* are refractory-lined tubes with a diameter (D) of up to about 6 m and of different length (L). They are inclined at an angle of 3 – 4° and rotate at 1.2 – 2 rpm. As a result of the inclination and rotation of the tube, the feed moves from the top of the kiln down the tube toward the coal-dust, oil, or gas flame. Near the flame, in the sintering zone of the rotary kiln with a gas temperature of 1800 – 2000 °C, the material reaches the temperatures of 1350 – 1500 °C necessary for the formation of clinker phases.

The energy released from fuel combustion in the rotary kiln is transferred to the material stream by radiation and convection. To ensure effective heat transfer in the temperature range below the clinkering temperature, the rotary kiln must be sufficiently long. In modern kiln systems, process steps such as drying and preheating of raw materials and precalcination are performed separately; hence the kiln can be shorter (L/D = 10 – 12). For comparison, the L/D ratio of long rotary kilns is 32 – 35.

The largest kilns have lengths of more than 200 m. Long kilns, without grate or suspension preheaters and with clinker capacities of more than 3000 t/d, are used in both the wet and dry process. They are known for their simplicity and their operating reliability. The upper part of the long kilns is equipped with chain curtains and fixed installations to improve the heat transfer for drying and preheating the raw material. Whereas the wet process requires 5000 – 6000 kJ/kg of clinker, the corresponding figure for the dry process is 3400 – 5000 kJ/kg. This is because besides increased energy losses (waste gas and kiln wall) additional energy is needed for water evaporation.

Rotary kilns equipped with preheaters have an L/D ratio of 10 – 17. With 3100 – 4200 kJ/kg clinker, they need a considerably lower input of thermal energy. According to CEMBUREAU (Association of the European Cement Industry) statistics, the proportion of clinker produced in dry-process kilns equipped with preheater and calcinator is increasing. Figure 7 shows the European status as of 1992 [176]. In Germany the proportion of energy-saving suspension preheater kiln systems increased to 84 % in 1998. The grate preheater system dropped to 14 % and long wet-process kilns were shut down completely.

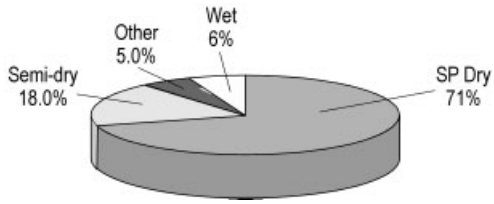

Figure 7. Portland cement clinker production in Europe broken down according to different process technologies (CEMBUREAU statistics 1992) [SP: Suspension typ preheater kilns, dry process]

Two types of preheater exist: grate and suspension preheaters. Grate preheaters are fed with pellets or briquettes. Pellets are formed from raw meal, wetted with about 12 wt % water in a pelletizer (semidry process). The cylinder-shaped briquettes are prepared from a raw material slurry, which is dewatered in a mechanical filter press and subsequently extruded in a screen compactor (semi-wet process).

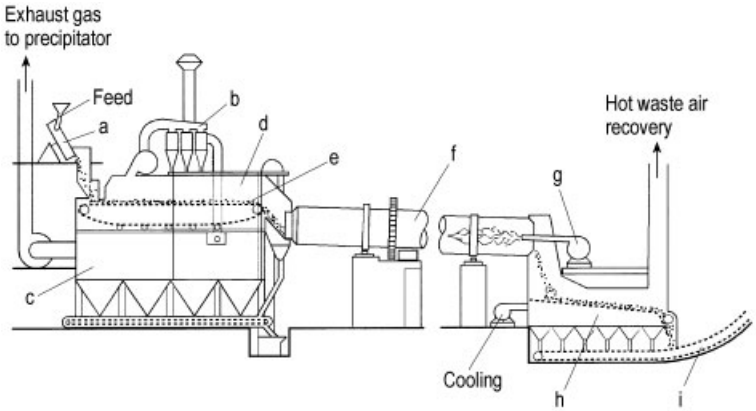

Figure 8. Schematic of clinker burning in a grate-preheater kiln system with grate cooler [177]
a) Pelletizer; b) Intermediate dust collector; c) Drying chamber; d) Hot chamber; e) Grate; f) Rotary kiln; g) Burner; h) Grate cooler; i) Clinker belt

In the grate-preheater kiln (Fig. 8), pellets or briquettes of the feed material are placed on a grate. The compartment is separated into a hot gas chamber and a drying chamber by a partition wall with an opening for the grate. The first fan draws the exhaust gas from the rotary kiln into the top of the preheater, through the pellet layer in the hot gas chamber, and then through the cyclones of the intermediate dust collector. In these cyclones, large dust particles have to be removed to avoid damage to the fan. The second fan then draws the gas into the top of the drying chamber, through the layer of moist pellets, and finally into the dust collector.

The process gas leaves the rotary kiln with a temperature of 1000 – 1200 °C. On passing through the hot gas chamber, the process gas is cooled to 250 – 300 °C, and it leaves the drying chamber at 90 – 150 °C. The low temperature (60 – 90 K above the water dew point) is due to the energy consumption for pellet drying. As a result of its intensive contact with the process gas, the feed reaches a temperature of about 150 °C in the drying chamber and about 800 °C in the heating chamber. The largest grate preheater kilns have a clinker capacity of more than 3000 t/d.

The suspension preheater kilns show a great variety in construction. In principle, the preheater consists of four cyclone stages which are arranged one above the other in a 50 – 70 m high tower (Fig. 9). New kilns also can be equipped with five or six stages, depending on the heat demand for raw material drying. The uppermost stage comprises two parallel cyclones for better dust separation. The process gases from the rotary kiln move through the cyclone stages in countercurrent flow to the raw material feed. The dry raw material is added to the exhaust gas in the riser duct before the uppermost cyclone stage. It is separated from the gas in the cyclones and re-mixed with the exhaust gas from the next cyclone stage. This procedure is repeated until the material, preheated to about 800 – 850 °C, is fed to the calcinator and/or the kiln. This procedure leads to an intensive heat transfer.

Further development of the preheater kiln system led to the introduction of the precalcination technique around 1970. In this process calcination of the raw feed has been shifted from the upper end of the rotary kiln to the calciner/preheater. Hence, the total thermal energy input had to be split up into two burning locations: the primary burner in the sintering zone and a secondary burning system (calciner) between rotary kiln and the lowest preheater stage. The energy input by the secondary burning system can be high enough to calcine 70 – 90 % of the $CaCO_3$ in the raw feed before it enters the rotary kiln. In conventional systems without a secondary burner the thermal energy for the dissociation of $CaCO_3$ is released completely by the primary burner, and the raw feed is calcined by only 40 – 50 % before it enters the kiln.

The major advantage of the precalcination process is that the majority of the energy re-

quired for $CaCO_3$ dissociation is released in the calciner by flameless combustion at about 900 °C. For this reason, low-grade materials such as ash-rich coal and waste material can be used as secondary fuels, provided the ash content can be compensated in the raw feed composition. For a given kiln dimension, precalcination increases the clinker production capacity. The largest clinker capacities of the suspension preheater kilns with precalcination amount to 8000 – 10 000 t/d.

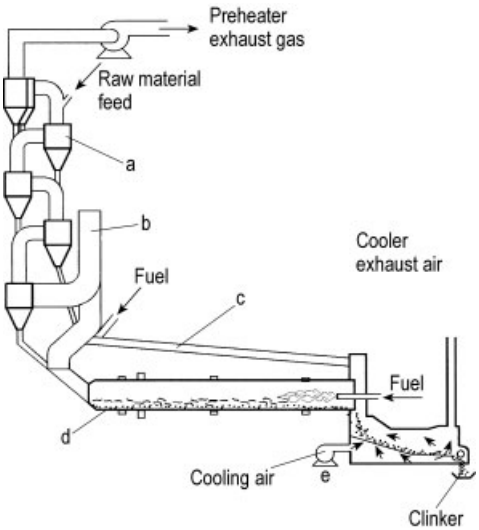

Figure 9. Schematic of clinker burning in a four-stage suspension preheater kiln system with grate cooler [177]
a) Preheater; b) Calciner; c) Tertiary air duct; d) Rotary kiln; e) Grate cooler

The temperature of the exhaust gas leaving a 4-stage preheater is as high as 380 °C, and it can be used to dry raw materials with a moisture content of about 8 – 10 wt %, as well as coal. In the case of raw materials with a lower moisture content than about 5 wt %, it may be economic to partly recover the heat of the preheater gas, for example by installing a waste-heat boiler [174].

Calciners exist in different designs [178], [179]. Besides precalcination the equipment additionally offers the possibility for a stepped fuel and/or raw material input. Staged combustion is already successfully being used to reduce the thermal NO_x load in the kiln exhaust gas by using sub-stoichiometric burning conditions in the first calciner step. The combustion air which is necessary to burn the secondary fuel is supplied by transport through the rotary kiln and/or by a separate tertiary air duct. Figure 10 shows a calciner combustion air supply with and without a tertiary air duct [177]. After the reduction phase the exhaust gas flow is mixed with tertiary air to burn out, for example, carbon monoxide [167].

The clinker leaves the rotary kiln after being sintered at about 1150 – 1350 °C and enters a grate, tube, or planetary cooler, in which it is cooled in countercurrent or cross/countercurrent flow. Figure 11 shows the different cooler types. In the grate coolers of different construction, the clinker is transported on a moving or on a combination of stationary and moving grates. In the grate, tube, and planetary or satellite cooler (which consists of 10 or 11 tube coolers, attached to the circumference of the rotary kiln), the hot parts are equipped with refractories. Additionally, tube and satellite coolers are constructed with air-lifting systems to improve the heat exchange. The heated cooling air is used completely or to a great extent as so-called secondary or tertiary air, which is fed back into the sintering zone of the kiln or into the calcinator as preheated combustion air at 700 – 1000 °C. Grate coolers require more cooling air than is necessary for combustion. The excess air volume leaves the system as waste air at about 250 – 350 °C. After cooling, the clinker has a temperature of 80 – 200 °C. In modern kiln systems most of the waste gas heat and the heat of the clinker can be recovered. The efficiency of energy use has thus increased by more than 70 %.

The formation of alkali metal, sulfur, chloride, and heavy metal cycles can be diminished by a bypass arrangement, if necessary. For this purpose, a partial volume of dust- and vapors-containing kiln exhaust gas is separated from the main stream before it enters the calciner and/or the preheater [180]. However, a bypass increases the energy loss of a kiln system, it needs special waste-gas cleaning equipment, and leads to waste materials, which in most cases must be disposed of.

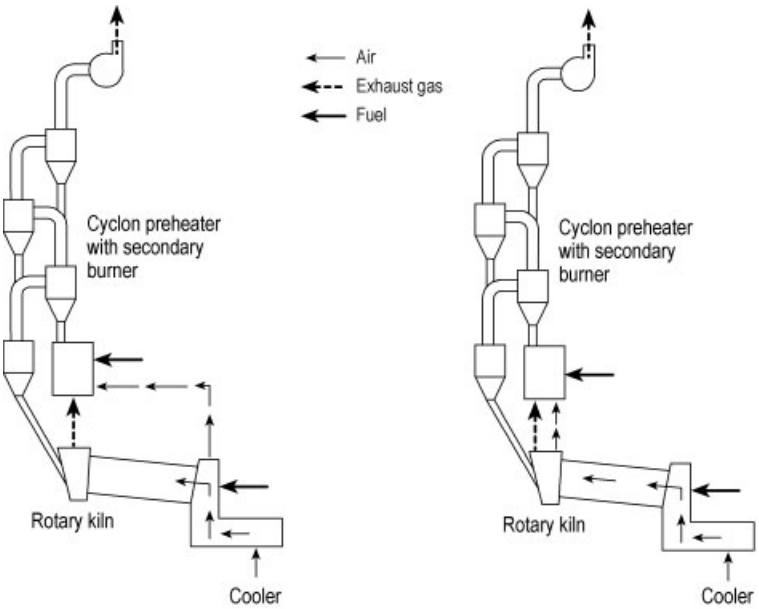

Figure 10. Suspension preheater kiln system with a precalcining stage and different kinds of combustion air supply [177]

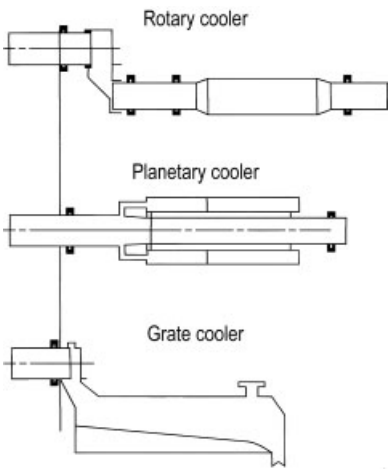

Figure 11. Different cooler types for cooling of Portland cement clinker [177]

1.4.5.2. Energy Consumption

Thermal energy is spent more or less exclusively for burning clinker. The change from processes with a high specific energy demand to energy-saving processes has led to a remarkable decrease in thermal energy consumption. Figure 12 shows the development in Germany since 1950. The average demand of about 3500 kJ/kg clinker in 1998 is the result of continuous and long-term improvements in process technology and energy efficiency which additionally need a permanent capital investment [174]. The benefits of process conversion are also obvious for other European countries [176].

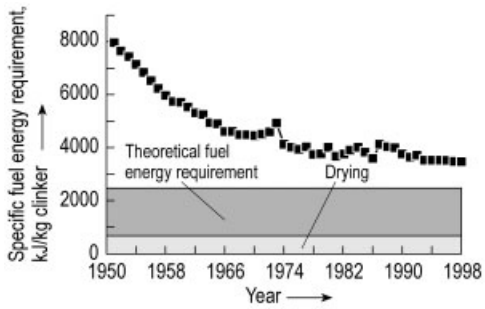

Figure 12. Development of the specific thermal energy consumption in Germany since 1950 [174]

The approximately constant expenditure of thermal energy for the drying of raw materials (average moisture content 5 – 8 wt %) and the energy demand for chemical reactions (1700 kJ/kg clinker) are unavoidable for natural raw material components. Experience has shown that technical measures to improve the burning process in many cases are exhausted, so that the fur-

ther energy saving potential in general seems to be small [176]. The utilization of electrical energy saving processes in the comminution of raw materials, coal, and especially cement has the favorable effect that the total specific consumption at present is not increasing in a marked way. Due to growing requirements on the performance of cements, e.g., higher fineness and strength, on a higher proportion of cements with other main constituents besides clinker, or on measures which continue the decrease of particulate and gaseous emissions, a rising consumption of electrical energy must be expected. This applies also to technologies which are based on the production of low-energy (C_2S) clinker or on burning processes with mineralizers.

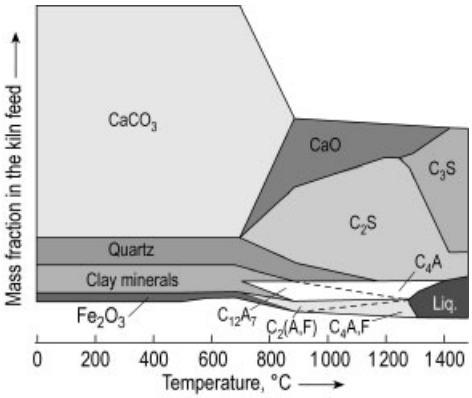

Figure 13. Decomposition of raw material components and formation of clinker phases with increasing temperature [181]

The energy-saving potential of the cement production process is summarized in the following:

1) Optimization of the burning process
   – Dry process (precalcination, grate cooler)
   – Waste gas utilization
   – Process control
   – Burnability (LSF, low energy clinker, mineralizers)
   – Secondary raw materials and fuels
2) Optimization of the grinding process
   – Primary crushing
   – High-pressure comminution (roller mill, high-pressure grinding rolls)
   – High-capacity separator
3) Utilization of different main constituents in cement

1.4.5.3. Thermochemistry of the Burning Process

Several chemical reactions occur during the heating of raw materials to produce cement clinker. Carbon dioxide and water vapor, released from limestone and clay, enter the waste gas stream derived from the burning of fuel and affect its composition. Figure 13 shows schematically the reactions taking place with increasing temperature and the reaction products [181].

Accordingly, the clinker-burning process can be subdivided into six temperature-dependent steps:

1) Starting composition (20 °C): Calcite ($CaCO_3$), quartz (SiO_2), clay minerals ($SiO_2 - Al_2O_3 - H_2O$), iron-containing components (Fe_2O_3)
2) Temperature range 20 – 700 °C (preheater): liquid water and adsorbed water escape during initial heating, followed by dehydration of clay minerals, and activation of oxides by decomposition of clay minerals.
3) Temperature range 760 – 900 °C (preheater and calciner): Calcination of $CaCO_3$, binding of activated Al_2O_3, SiO_2, and Fe_2O_3; first formation of clinker phases such as C_2S, $C_2(A,F)$, $C_{12}A_7$, (CA).
4) Temperature range 900 – 1250 °C (kiln): completion of C_2S formation, initial formation of melt (C_3A).
5) Temperature range 1250 – 1450 °C (kiln): starting at 1250 °C and activated by the melt, C_3S is formed from CaO and C_2S.
6) Cooling step (not shown in Fig. 13): cooling of the clinker commences in the precooling zone of the kiln. The length of this zone and the cooling rate depend on several factors, for example, on the primary burner position and the type of cooler. The clinker leaves the kiln at a temperature of about 1350 °C (sintering zone close to the kiln exit) to 1150 °C (long cooling zone for satellite coolers). The optimization of cooler efficiency by new techniques improves the efficiency of the whole kiln system [182]. Additionally, the clinker and cement properties can be influenced positively by a high cooling rate, especially in the precooling area of the kiln [183].

Table 5. Energy demand of various clinker burning processes [167]

	Suspension preheater kiln	Grate preheater kiln	Long wet process kiln	Long dry process kiln
Percentage of total consumption				
Clinker formation	52.5	47.2	30.5	34.6
Evaporation of water	0.5	16.1	41.5	0.4
Waste gas (after preheater)	25.5	7.8	13.2	47.8
Exhausted cooler air	8.8	5.8	1.8	1.7
Residual heat of the clinker	2.4	1.8	1.0	1.2
Radiation and other wall losses	7.5	11.9	9.5	11.8
Others	2.8	9.4	2.5	2.5
Total thermal energy consumption, kJ/kg clinker	3000 – 3800	3100 – 3800	5000 – 6000	5100
Electrical energy demand (kiln system), kW h/t clinker	10 – 20	12 – 20	10 – 20	17

The total amount of energy required to burn cement clinker comprises several components. The theoretical energy demand of clinker formation is the difference in enthalpy between the product (clinker) and the reacting raw materials [29]. The complete balance of the thermal energy consumption additionally involves the fraction of energy losses. Both the theoretical energy demand and the losses are equivalent to the fuel demand. Energy losses are caused by evaporation of water, exhaust gas, wall losses of the preheater, kiln, and cooler, and residual heat of the hot clinker. Table 5 shows characteristic examples of energy balances and the distribution of losses for different burning processes [167].

The dehydration and decomposition of the clay minerals and the calcination of calcium and magnesium carbonate are endothermic reactions. The formation of clinker phases from calcium oxide and the decomposition products of the clay components involves exothermic reactions. Hence, the theoretical energy requirements depend on the chemical and mineral composition of the raw material mixture and amount to 1600 – 1900 kJ/kg clinker.

The theoretical energy demand for clinker burning is mainly dependent on the amount of energy required for $CaCO_3$ calcination (1780 kJ/kg). A change in the raw meal composition therefore influences the theoretical energy demand. The value can be significantly decreased by utilization of raw feed with a lower $CaCO_3$ content or with a standard lime factor of less than 92. As a consequence of these measures, the thermal energy consumption, the sintering temperature, and also the formation of thermal NO_x decrease. However, the increasing C_2S content (Belite clinker), the higher electrical energy demand for grinding, and the loss in strength under industrial production conditions are unfavorable [167]. The utilization of siliceous or calcareous fly ashes and granulated blast-furnace slag as constituents of the raw mix can decrease the theoretical energy demand.

The kinetics of raw meal decomposition and of the formation of clinker phases depend not only on the composition and burning temperature but also on the partial pressure of CO_2, the fineness and crystallinity of $CaCO_3$, the degree of homogeneity in the distribution of quartz among clay minerals and $CaCO_3$, on the fineness of quartz, the heating velocity, and on mineralizers and fluxes, which influence the amount and viscosity of the clinker melt. Figure 14 shows the particle size distribution of constituents in raw meal of different geological origin [184]. Size distribution and the homogeneity of the distribution are of technical importance. They are often summarized under the term ″burnability″ of a cement clinker raw mix [180].

1.4.6. Characterization of Cement Clinker

Portland cement clinker can be characterized by chemical and mineralogical analysis. Whereas the chemical composition is, at present, analyzed mainly by X-ray fluorescence techniques (XRF), the analysis of clinker phases is deduced from calculations according to Bogue [31] and by microscopic [30], X-ray diffraction (XDA), or scanning electron microscopic (SEM) methods. In modern cement plants, XRF is employed for regular inspection of the homogeneity and constancy of material flows including raw materials, fuels, secondary materials, and clinker. Many other analytical techniques [e.g., backscattered

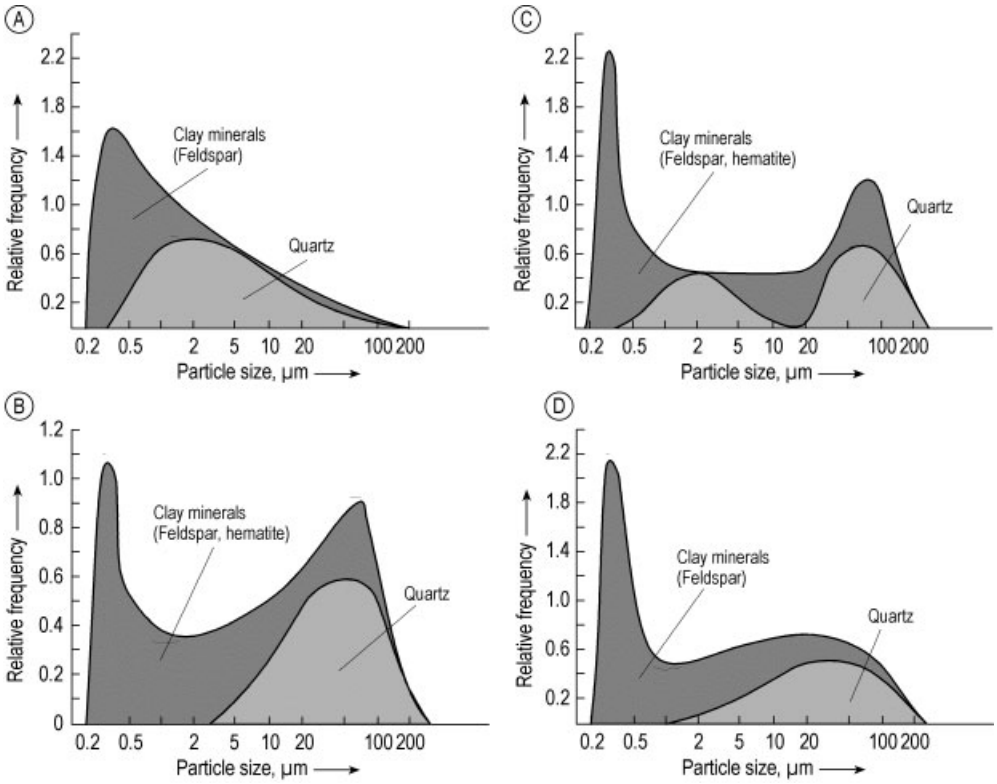

Figure 14. Particle size distribution of constituents in raw meal [184]
a) Raw meal I (Upper Cretaceous); b) Raw meal II (Devonian); c) Raw meal III (White Jurassic); d) Raw meal IV (Lower Muschelkalk)

imaging (BEI), electron probe microanalysis (EPMA), nuclear magnetic resonance (NMR), etc.] are predominantly used for scientific purposes [10].

The content of main and minor chemical components in the clinker may vary over a wide range. An example is given in the following (in wt % with average content in parentheses; the calculated phase composition is given in Table 1) [167]:

SiO_2	19.7 – 24.3	(21.4)
Al_2O_3	3.8 – 6.8	(5.5)
TiO_2	0.2 – 0.5	(0.3)
P_2O_5	0.02 – 0.3	(0.1)
Fe_2O_3	1.3 – 4.6	(2.7)
Mn_2O_3	0.03 – 0.7	(0.1)
CaO	63.8 – 70.1	(66.6)
MgO	0.0 – 4.5	(1.5)
SO_3	0.2 – 2.1	(0.8)
K_2O	0.3 – 1.8	(0.9)
Na_2O	0.03 – 0.3	(0.1)
CO_2	0.03 – 0.8	(0.1)
H_2O	0.04 – 1.1	(0.3)

The calculation of the phase composition from a chemical analysis [31] requires the fulfilment of some prerequisites. Accordingly, the actual composition of the clinker phases should correspond to their formulas. Furthermore, the clinker melt should be in continuous equilibrium with the solid phases, not only at the sintering temperature, but also during crystallization, when it is cooled. Additionally, a strictly oxidizing atmosphere in the kiln is required. This means that tricalcium silicate is resorbed in an amount corresponding to the tricalcium aluminate content.

In practice, neither the composition of phases corresponds to the formula nor are the equilibrium conditions met because of the relatively high cooling velocity of the clinker when it leaves the sintering zone of the rotary kiln. Consequently, the calculated clinker composition values are often too low for tricalcium silicate and too high for dicalcium silicate. Addition-

ally, due to the composition of the feed, the melt at the sintering temperature does not only vary in the Al_2O_3/Fe_2O_3 ratio, but the melt fraction may also contain too little calcium. The quantity of Ca^{2+} is therefore not high enough to form the theoretically possible quantities of C_3A and $C_2(A,F)$ during the cooling process [32]. Consequently, the C_3S, which is richer in lime, would be resorbed. The extent of these reactions and the exchange of other constituents, such as Fe_2O_3, Al_2O_3, and MgO, depend on the chemical composition and the cooling velocity. Clinker with C_3A contents of less than 11 wt % show a fairly low C_3S corrosion and only a small migration of Ca, Al, and Fe from C_3S into the melt. However, higher amounts of C_3A than 11 wt % lead to a higher corrosion rate of C_3S. Whereas the Ca migrates from C_3S into the melt, the C_3S absorbs Fe, Al, and Mg. The higher the C_3A content in the clinker and the more slowly the clinker was precooled in the kiln, the greater is the exchange of material [182], [183]. Therefore, a calculation of the phase composition under the known preconditions only gives a rough figure as compared to the true clinker composition. Moreover, the corresponding change of the clinker properties with respect to setting and strength development cannot be predicted with the necessary precision.

Besides various expensive microscopic methods, quantitative X-ray diffraction analyses by the Rietveld method [185], [186] is sufficiently accurate to fulfil the requirements for clinker-phase analyses. In principle, a complete powder pattern of a pure phase or of a multiphase mixture is computed and subsequently compared with an actually measured diffractogram. The computed data files contain complete information on the chemical structure of the relevant crystalline phases. They are the result of long-term structural research and will gradually be completed, also for industrial products. This method of standard free phase analyses will be further developed for plant application; it may possibly be used as an important element in the quality assurance system.

The focal point of microscopic examinations of clinker is the qualitative specification of type, formation, and distribution of the phases in the clinker. The type of phases produced depends primarily on the chemical composition of the burnt material. The structure, that is, the formation, distribution, and intergrowth of the compounds, depends on the preparation of the raw material mixture and on the conditions of clinker burning and cooling.

Examination under a light microscope is carried out on a polished, etched section of clinker in reflected light. In general, distilled water, potassium hydroxide, alcoholic nitric acid, or dimethylammonium citrate is used for etching [30]. The individual constituents of the clinker can be identified by their light reflection and by their behavior during etching. Studies on thin sections and on powder preparations with transmitted light are more difficult. Figure 15 shows a microphotograph of a polished clinker section that has been etched with water. The heavily etched, and hence dark, idiomorphic alite and the weakly etched spherical belite can be distinguished. The matrix surrounding these particles consists of solidified and crystallized clinker melt; it shows dark tricalcium aluminate and light calcium aluminoferrite. For SEM studies a freshly fractured surface of a clinker piece is used. Metal atoms, generally gold, are sputtered onto this surface under vacuum.

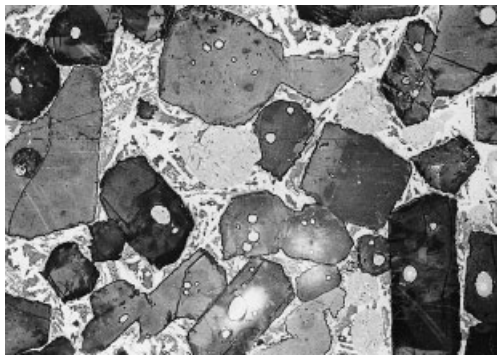

Figure 15. Portland cement clinker, polished section, etched with water, as observed under the light microscope

Another common method to characterize the clinker composition is the calculation of the potential composition by using the lime saturation factor (LSF), the silica modulus, and the alumina or iron modulus. The lime saturation factor (standard lime, KST) is defined as the ratio between the amount of calcium oxide actually present in the raw material mixture or clinker, and the maximum amount of calcium oxide which could be bound by the actual content of

silica, aluminum oxide, and iron(III) oxide under industrial burning and cooling conditions [15].

$$\mathrm{LSF} = \frac{100 \cdot \mathrm{CaO}}{2.80 \cdot \mathrm{SiO_2} + 1.18\mathrm{Al_2O_3} + 0.65\mathrm{Fe_2O_3}}$$

The LSF corresponds to KST II. KST III takes MgO contents of ≤ 2.0 or ≥ 2.0 wt % MgO [32] into consideration.

The silica modulus (SM) is the ratio of the amount of silica and the sum of aluminum and iron oxide [8]. At the sintering temperature, silica is primarily bound in the solid phases tricalcium silicate and dicalcium silicate; aluminum and iron oxides are contained in the melt. For this reason, the SM defines the solid-to-liquid ratio in the clinker at sintering temperature. The silica modulus lies generally between 1.8 – 3.0, but the most frequent and favorable value is 2.3 – 2.8.

The alumina modulus is the ratio of the amount of aluminum oxide and the amount of iron oxide present [8]. It gives information about the ratio of calcium aluminate to calcium aluminoferrite and therefore about the composition and viscosity of the clinker melt. Clinker with a normal composition has an alumina modulus of 1.5 – 4.0. With an alumina modulus lower than 0.638, arithmetically the entire aluminum oxide contained in the clinker is bound as calcium aluminoferrite $C_2(A,F)$. The KST, the silica modulus, and the alumina modulus are used to control the raw mix composition by a calculation method based on the hydraulic, siliceous, and aluminous deviation [8].

The amount of free lime present, determined by X-ray diffraction or analytically after extraction with organic solvents indicates the burning degree of cement clinker at LSF values below 100.

1.5. Cement Production Process

1.5.1. Storage

In the production process the storage of clinker is needed to compensate the difference between production and dispatch. Such differences may be due among others on plant interruptions and on fluctuations in cement demand. For environmental and quality reasons cement clinker is predominantly stored in closed silos or halls. Because of a relatively low bulk density of about 1.5 t/m^3 the dimension has to be adjusted to the storage capacity needed. Guidelines are capacity for four- to six-weeks' production. A kiln with a production of 3000 t/d therefore needs a clinker storage capacity of about 80 000 – 120 000 t [187]. Storage facilities additionally have a homogenizing effect on the clinker mass stream. The other cement constituents need a suitable storage volume as well.

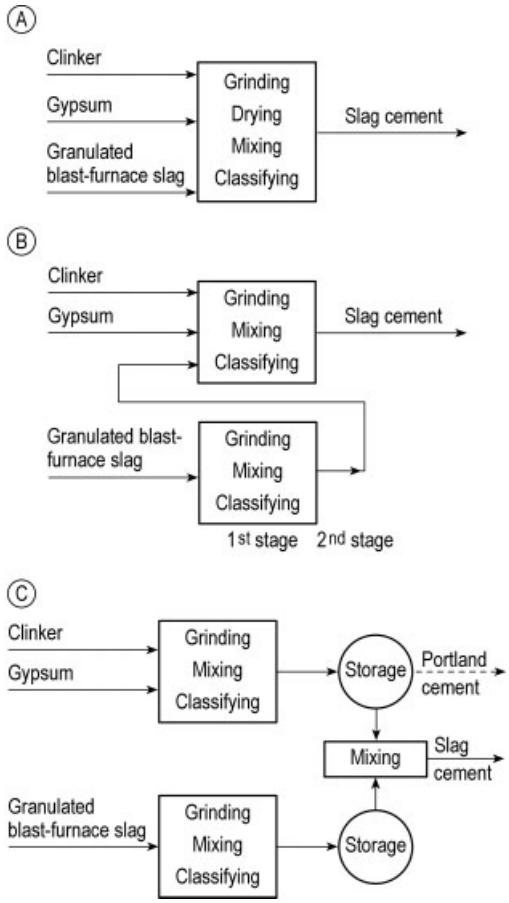

Figure 16. Principle process technology for cement grinding [167]
A) Compound grinding; B) Two-stage grinding; C) Separate grinding and mixing

1.5.2. Grinding and Blending

Cement is traditionally produced by a fine grinding of either clinker alone or with blast-furnace slag, pozzolana, fly ash, or filler. Gypsum or anhydrite is added to the material being ground

to control both setting and strength development. Another technology is based on separate grinding and blending of the constituents. The process stages also include drying of wet constituents like granulated blast-furnace slag and classifying of the mill output by air separators. All steps must be combined in a suitable manner to achieve the necessary product and production efficiency.

In principle there are three different well-tried plant systems which are employed in practice. Figure 16 shows the plant systems for compound grinding, two-stage grinding, and separate grinding and blending [167], [188] for the production of blast-furnace slag cements.

In the compound grinding process, drying, grinding, and homogenization take place simultaneously in one mill. Sometimes a separate air-heating system is needed to support drying, especially in the case of higher moisture contents of the slag and to avoid quality losses caused by prehydration in the mill. In the case of CEM II and CEM III production, the grinding plant is run in closed circuit with an air separator. In general, open-circuit grinding is mainly used for the production of CEM I cement with a fineness of less than 3500 cm^2/g (Blaine). During simultaneous grinding of several main constituents the achievable fineness and particle size distribution of the cement depend on the grindability of the individual constituents and the kind and duration of grinding.

The two-stage grinding processes are characterized by a combination of two grinding plants configurated in a different way. This system offers the possibility of grinding and drying moist constituents separately from clinker. Additionally, one of the constituents can be preground in the first stage to a required fineness and size distribution, dependent on grindability, which may be easily optimized during finish grinding in the second stage.

The third plant system is characterized by separate grinding of the main constituents, storage of the fine components, and a blending stage which allows the production of a finished cement with the required formulation with respect to, for example, different hydraulic or pozzolanic reactivities, fineness, and particle size distribution.

The grinding systems can be equipped with ball mills with tube diameters of up to 6 m, tube lengths of up to 20 m, drive ratings of as high as 10 000 kW, and a throughput of 300 t/a. The mills mostly operate in closed circuits with adjustable air separators of various designs. High-pressure grinding rolls are integrated into new or existing grinding systems in various configurations with the aim of improving the cement output and the energy efficiency. They can work as a pregrinder but also in hybrid and semifinishing configuration. The coarse feed is comminuted by high compressive stress (50 – 350 MPa) in the gap between two counter-rotating grinding rollers with a maximum diameter of up to 2 m and a length of up to 1.4 m. The compacted cake must be disagglomerated and contains up to 40 wt % of comminuted materials smaller than 90 µm. Whereas the power consumption is rather low, about 50 – 60 % of that of a ball mill, the particle size distribution in the RRSB grid generally has slope steeper than 1.0, and this leads to a significantly higher water demand of the cement. Therefore, the particle size distribution of the cement has to be adjusted by finish grinding to give a shallower slope, for example, in a subsequent grinding step with a ball mill. By finish grinding, for example, from 2200 cm^2/g to 3500 cm^2/g the slope can be decreased to about 0.9, and the workability can be improved. In addition, also the calcium sulfate optimization (see Section 1.6.2) can favorably be corrected.

Besides these two grinding systems, vertical and horizontal roller mills are also employed for cement grinding [188]. Compared to ball-mill grinding (100 %), the high-pressure grinding rolls system needs 50 – 65 %, the vertical roller mill system 70 – 75 %, and the horizontal roller mill system 65 – 70 % of the electrical power consumption. Altogether, the power consumption depends on the grindability of the cement components and the required fineness of the cement. The specific power consumption for grinding common cements in ball mill systems is given in the following figures for different ranges of fineness [164]:

Specific surface area, cm^2/g	Specific electrical power demand, kW h/t
3000	25 – 35
4000	35 – 55
5000	50 – 80

However, quality requirements and many other factors such as drying capacity, service and

maintenance, capacity, and capital expenditure make it necessary to find the best solution for each separate case.

Fineness, particle size distribution, and composition of cements are continuously monitored for reasons of a constant quality level. Simulation models of the grinding circuit make it possible to minimize the influence of disturbances, short-term stoppage, and grindability fluctuation on the uniformity of cement properties and to adjust hydraulic capacity, fineness, and particle size distribution to individual requirements on cement properties for concrete production, especially for self-compacting concrete with increased contents of size adjusted fillers.

Grinding aids are added in small amounts of 0.02 – 0.04 wt % to the mill feed. They improve the energy efficiency of cement comminution in ball mills by decreasing the agglomeration within the feed and they help to minimize coating on the balls or the lining plates. Grinding aids also fluidize the cement, but they may lower the efficiency of dust filters. The mill output is increased, especially when grinding cements with a specific surface area of more than 4000 cm^2/g. Common grinding aids are glycols and ethanolamines; their effect is based on a chemisorption at the surface. Water vapor has a similar effect. Grinding aids do not impair the water demand, setting, and strength development of cement. Because of the high sorption capacity of cement the environmental impact is negligible.

1.6. Hardening of Cement

Setting and hardening of cement is based on the formation of water-containing hydrate phases which develop during the reaction of cement components with mixing water. The amount of mixing water is related to the mass of dry cement powder. The ratio of water to cement is called water/cement ratio (w/c). For technical purposes the hydration of cement commonly proceeds with a w/c of 0.3 – 0.7. The reactions take place in a very thin film of water surrounding each cement particle, as Figure 17 shows schematically [189]. Surface and size distribution of the cement particles influence to a large extent the w/c which is necessary to gain workability, strength development, and a dense structure.

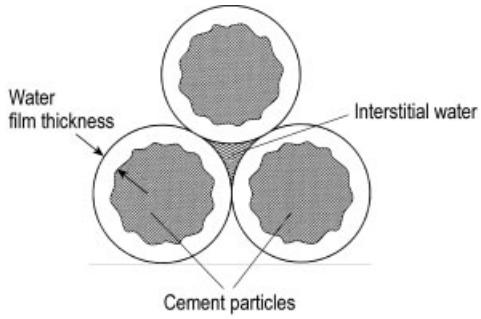

Figure 17. Water film at the surface of cement particles and interstitial water between the particles as influencing factors for the workability of cement paste and strength development [188]

The plastic mixture of cement and water is called cement paste, and the hardened mixture is known as hardened cement paste or cement stone. Mixtures of cement with sand (0 – 4 mm) are designated as mortar. If coarser aggregates are additionally used, following the standardized size distribution curves, concrete is obtained. The technically most important criteria for the properties of mortar and concrete are strength, density of the structure, and durability. They are dependent on the packing density of aggregates and the hydrated cement.

1.6.1. Hydration Products

The calcium silicate hydrates, $m\,CaO \cdot SiO_2 \cdot n\,H_2O$ (CSH), are the most important water-containing compounds formed during hydration of all cements, except for calcium aluminate cement. They are responsible for the strength of the hardened cement paste. The molar ratio of calcium oxide to silica m varies over a wide range between 10 and 30 °C, following the equilibrium in solution [164]. The ratio of Ca/Si in CSH is 1.5 in equilibrium with $Ca(OH)_2$ and a solution containing 1170 mg CaO/L. In solutions which are supersaturated with $Ca(OH)_2$, CSH with a Ca/Si ratio of about 1.7 – 2 is formed, similar to reactions in the hydrated paste of Portland cement. During the hydration of cement the phases CSH(I) and CSH(II) mainly develop. Their Ca/Si ratios are ca. 0.8 – 1.5 or 1.0 – 2.0. With increasing Ca/Si ratio, the crystallinity decreases. CSH phases also develop due to hydrothermal reactions during hardening of,

Table 6. Natural calcium silicate hydrates [191]

No.	Name	Chemical Composition	C/S
1	okenite	CS_2H_2	0.50
1	nekoite	$C_3S_6H_8$	0.50
2	truscottite	$C_6S_{10}H_3$	0.60
3	gyrolite	$C_2S_3H_2$	0.67
4a	plombierite(14 Å tobermorite)	$C_5S_6H_9$	0.83
4b	11 Å tobermorite	$C_5S_6H_5$	0.83
4c	riversideite (9 Å tobermorite)	$C_5S_6H_{0-2}$	0.83
5	xonotlite	C_6S_6H	1.00
6	afwillite	$C_3S_2H_3$	1.50
7	foshagite	C_4S_3H	1.67
8	hillebrandite	C_2SH	2.00

for example, sand – lime bricks or silicate concrete. In these building materials, CSH phases with a low lime content predominate which show a composition and structure similar to that of tobermorite. In the relevant literature at least 16 crystalline CSH phases with Ca/Si ratios of 0.5 – 2.0 are mentioned. Eight of them also occur as natural calcium silicate hydrates. Their composition is shown in Table 6 [164], [190], [191].

The degree of crystallization of CSH hydrates is low. Electron microscopic studies show that calcium silicate hydrates occur in pore spaces of hardened cement as fibers, needles, or sheets, depending on the Ca/Si ratio, but probably also on the available microscopic techniques [164], [190]. Future research with ESEM techniques promises to lead to a realistic picture of the morphology. Generally, CSH phases have a particle size of less than 0.1 µm. This is one of the reasons why XRD techniques cannot be used for direct quantitative measurements. Therefore, the proportion of hydrate phases in hydrated cement may be only estimated from the difference between the original and the remaining content of clinker phases [40].

C_3A has the highest hydration rate and forms the tabular aluminate hydrate, $4\,CaO \cdot Al_2O_3 \cdot 19\,H_2O$. This is irreversibly converted to the compound $4\,CaO \cdot Al_2O_3 \cdot 13\,H_2O$ when it is air-dried and no carbonatization takes place. The unstable C_4AH_{13} converts to the stable compound C_3AH_6, preferably at temperatures above 20 °C. Calcium aluminoferrite yields the same compounds during hydration, with the exception that aluminum oxide is partially replaced by iron(III) oxide.

In the presence of sulfate, calcium aluminates are preferably converted to two calcium aluminate sulfate hydrates: the needle-shaped prismatic trisulfate $3\,CaO \cdot Al_2O_3 \cdot 3\,CaSO_4 \cdot 32\,H_2O$ (ettringite) is formed in sulfate-rich solutions, whereas in solutions deficient in sulfate but rich in lime, the tabular monosulfate $3\,CaO \cdot Al_2O_3 \cdot CaSO_4 \cdot 12\,H_2O$ is formed. Fe^{3+}, Mn^{3+} and Cr^{3+} can replace the aluminum oxide in these compounds. Cr(VI) is bound like SO_4^2 in ettringite. Atmospheric carbon dioxide converts the monosulfate to the corresponding monocarbonate $3\,CaO \cdot Al_2O_3 \cdot CaCO_3 \cdot 11\,H_2O$, and chloride solutions can lead to the formation of the monochloride $3\,CaO \cdot Al_2O_3 \cdot CaCl_2 \cdot 10\,H_2O$ (Friedel's salt), which has a low solubility in water. The partial release of sulfate as a consequence of carbonatization and/or attack of chloride (e.g., by thawing salts) may cause secondary ettringite formation in the hardened structure of concrete [190].

Ettringite and thaumasite ($CaSiO_3 \cdot CaCO_3 \cdot CaSO_4 \cdot 16\,H_2O$) form an uninterrupted series of mixed crystals. Both components may lead to concrete damage caused by chemical attack of sulfates.

1.6.2. Course of Hydration

1.6.2.1. Water Demand

Setting and hardening of cement starts with the addition of water. The water is needed for both the chemical hydration reactions and physical wetting. The physical fraction builds up the water layer around the cement particles and fills the space between the particles to give a plastic consistency of the paste. The water proportion

needed for standard consistency is designated as water demand. The consistency of the paste in fresh mortar and concrete is decisive for their workability. In normal practice, the water addition for mortar and concrete deviates between w/c ratios of about 0.3 and 0.8. In general, for workability reasons, the water addition is higher than the amount needed for complete hydration of the cement. The workability may be improved and the w/c ratio decreased by employment of additives (see Section 2.2.4).

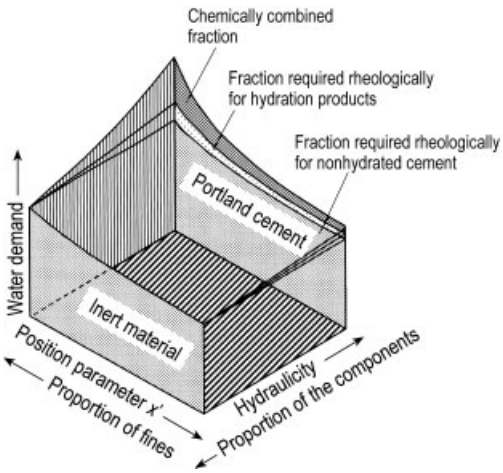

Figure 18. Water demand of cements, divided between chemically bound water and the water required for rheological reasons [168]

For a given chemical and mineralogical composition, reactivity of clinker, and optimal dosage of calcium sulfates [40], the grinding fineness and particle size distribution of the cement largely influence the water demand (Fig. 18) [168], [169]. It is possibly also influenced by the particle shape and particle roughness. The portion of chemically and instantly bound water increases only slightly at a constant slope of the size distribution with increasing fraction of fine grains, characterized by x' (position parameter of the RRSB particle size function). The majority is used for wetting the surface and filling of the void space between the clinker particles (interstitial water). The physical water demand corresponds to a w/c ratio of ca. 0.2. The diagram additionally shows that, for a given particle size distribution, with increasing fractions of those main constituents in cement which show lesser hydraulic properties compared with clinker, the water demand decreases. Hence, the decrease in water demand results from a diluting effect. For coarse-ground cement, the influence of constituents with minor hydraulic reactivity is hardly noticeable, but it is clearly noticeable in finely ground cement.

With an increasing slope n of the RRSB particle distribution, the physical water demand of Portland cements rises at constant fineness x'. This correlation was measured for 180 Portland cements of three strength classes, which were ground in different plants from different clinkers [169]. The influence of the slope is pronounced for the Portland cements with high grinding fineness ($x' < 15\,\mu m$) and leads to a water demand of more than 0.30. When the slope increases, the void fraction extends and causes the higher water demand of the cement. It can be reduced most efficiently by filling the void fraction with fines of defined quantity and adjusted particle size distribution. Suitable fillers are inert materials such as limestone or fine-grained siliceous fly ashes with low porosity and water absorption. The controlled minimization of the water demand can only be realized most effectively in grinding plants of cement works with effective control systems.

1.6.2.2. Setting

Chemical reactions start immediately after mixing cement with water. The alkali metal sulfate in the clinker, the calcium sulfates added to control setting, and minor parts of the tricalcium aluminate and tricalcium silicate dissolve very rapidly. Thus, a saturated solution of calcium and alkali metal hydroxides forms quickly. The pH of the solution is at least 12.5. Alkali metal sulfates react with dissolved aluminate to form alkali metal hydroxide and aluminate sulfate phases. As the content of alkali metal hydroxide increases, the content of calcium hydroxide decreases, and the pH of the solution exceeds 12.5.

These reactions result in a stiffening of the cement paste that increases with time. The setting point is defined as the time interval in which the loss of consistency has proceeded to a certain extent (initial setting). After final setting the setting period is followed by hardening. The rate of reaction and conversion of C_3A and the reaction with calcium sulfate are decisive for setting.

The hydration of calcium silicate is important for hardening [40].

Immediately after mixing, about 0.2 – 3 wt % of the C_3A, depending on its reactivity, dissolve and react with dissolved calcium sulfate to form trisulfate. This reaction stops after a few minutes and starts again after the dormant or induction period. This period may last several hours and is practically free of chemical reactions. Trisulfate forms a thin layer of very fine, prismatic crystals on the surface of the clinker particle depending on the pH of the mixture. The crystals are too small to bridge the space between the particles and to impair the consistency (Fig. 19). According to cement standards, setting of cement paste starts 1 – 3 h after mixing, within the dormant or induction period. Because of the absence of further measurable C_3A conversion, the loss of consistency may be explained in terms of recrystallization of the already formed ettringite [40]. At later stages of the hydration process, when the SO_4^{2-} concentration decreases in the pore solution, ettringite becomes unstable, and monosulfate is formed in the presence of calcium hydroxide and aluminate hydrate [190].

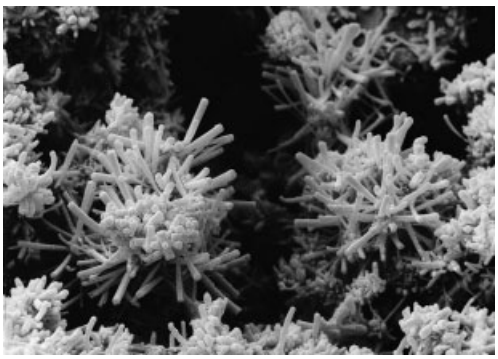

Figure 19. Structure of cement paste during the dormant period (SEM)

Controlled setting requires the addition of calcium sulfates during grinding and control of the grinding temperature. If finely ground cement clinker without interground calcium sulfates is mixed with water, the mixture generally sets immediately. In solutions with only a low sulfate concentration, large, flat crystals of tetracalcium aluminate hydrate are formed from the fraction of C_3A that is hydrated after addition of the mixing water until the dormant or induction period starts (ΔC_3A). These hexagonal hydrate crystals grow together to form a bulky structure in the space between the cement particles that is filled with mixing water, thus impairing the consistency and causing rapid or instant setting. A similar effect occurs when the portion of soluble sulfates is too low compared to ΔC_3A, and large flat crystals of monosulfate may form instead of ettringite. This effect, however, can be used in principle to employ the cement in shotcrete.

Another kind of rapid setting can be caused by crystallization of secondary gypsum. This happens if a surplus of soluble calcium sulfate exists compared to reactive C_3A (ΔC_3A).

To avoid rapid setting and a reduced workability period of paste in mortar or concrete, the type and proportion of instantly soluble calcium sulfates also must be adjusted to the amount of ΔC_3A in each cement (″optimization″). Figure 20 shows schematically examples of normal and rapid setting. Besides the type, amount, and solubility of calcium sulfates, the milling conditions must be considered, too [40].

The calcium sulfate added to control setting should be completely spent within 24 h to avoid expansion by late reactions between aluminates and sulfates in the already hardened paste. This is the reason why the total amount of sulfate is limited in cement standards.

1.6.2.3. Hardening

The hardening of Portland cement is primarily due to the hydration of fast-reacting tricalcium silicate and of slower reacting dicalcium silicate. The reactions result in fiberlike calcium silicate hydrate which has a lower lime content than the unhydrated clinker phases. As a result, calcium hydroxide is released, forming larger crystals in the hardened cement. Any tricalcium aluminate that has not reacted with sulfate forms tetracalcium aluminate hydrate. The slower reacting calcium aluminoferrite also forms tetracalcium aluminate hydrate, whereby iron(III) oxide replaces part of the aluminum oxide.

These hydration products do not develop simultaneously. They form at the reacting surface of the cement particle and grow into the water-filled space between the grains. X-ray diffraction and electron microscopic studies show different hydration stages; three or five hydration stages can be distinguished [164], [190]. Fig-

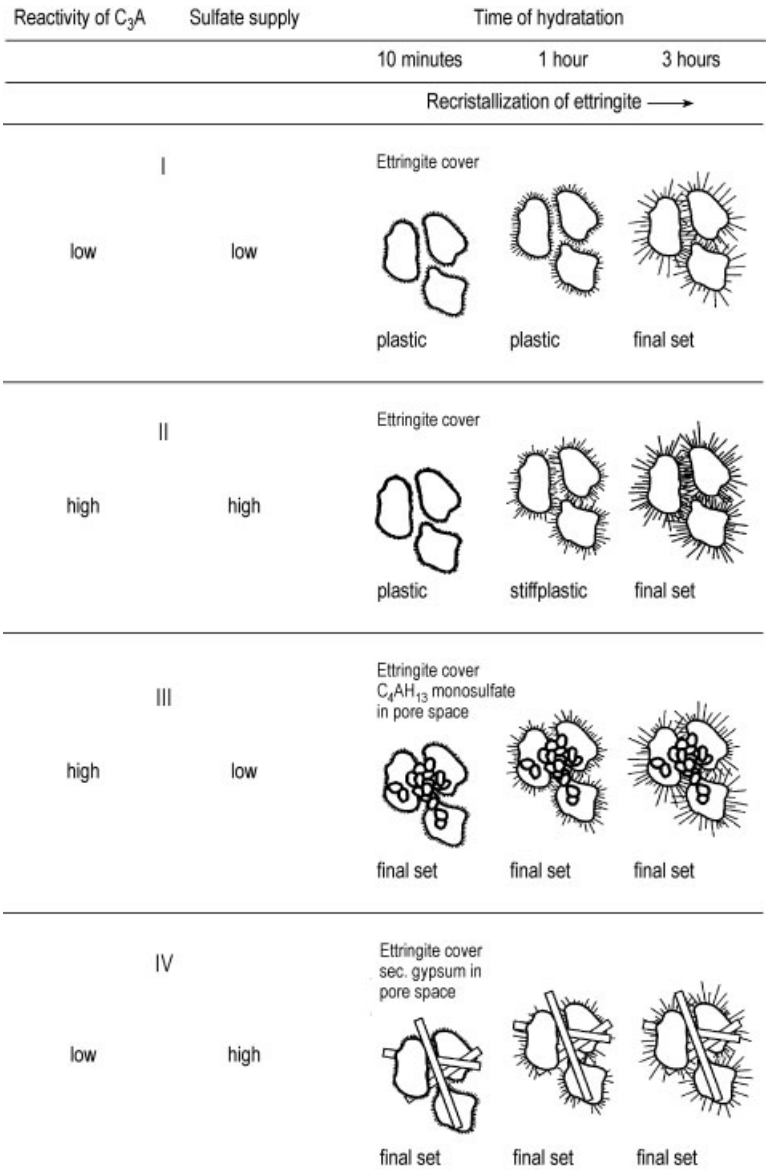

Figure 20. Normal and rapid setting of Portland cement [194]

ure 21 shows schematically the time-dependent formation of hydrate phases [11]. Only small amounts of calcium hydroxide and calcium aluminate trisulfate are formed in the first stage after addition of mixing water (paste). The pH increases to > 12.5. This chemical reaction is followed after a few minutes by a dormant or induction period, during which no detectable hydration reactions occur. The length of this time period may be different for C_3S and C_3A, it is independent of the w/c ratio, and in the case of C_3A it is also independent of the amount of soluble sulfate. The hydration of C_3S starts again after about 2 h in the presence of sulfate, but significantly later in sulfate-free pastes. In the course of the dormant or induction period, stiffening and setting occur. The reasons for the course of the induction period are not finally clear; there-

fore, different explanations for the mechanism still exist [164]. At the end of this period the hydration reactions are accelerated, and larger amounts of CSH, ettringite, and calcium hydroxide are formed. During this hydration stage hardening starts in combination with the formation of early strength. Further hardening is characterized by declining hydration velocities because of diffusion-controlled reactions. The reaction products fill out the pore space, the structure densifies, and strength increases.

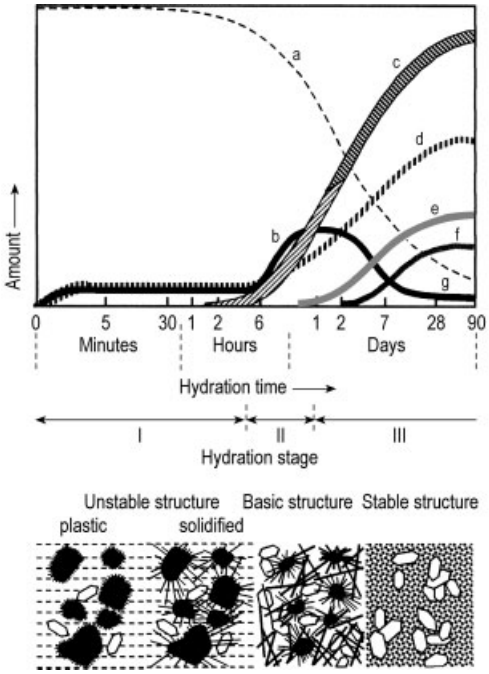

Figure 21. Formation of hydrate phases and structure development in a hardening Portland cement paste [11]
a) Porosity; b) Calcium silcate hydrate, long fibers; c) Calcium silcate hydrate, short fibers; d) Calcium hydroxide; e) Calcium aluminate hydrate containing iron(III) oxide; f) Monosulfate; g) Trisulfate

The remaining pore volume is determined by the original w/c ratio of the paste, as Figure 22 shows schematically for hydrated paste with different w/c ratios. All reactions are interrupted if the paste in mortar or concrete dries out. Therefore, curing conditions must be observed thoroughly to achieve the required properties of mortar and concrete.

Investigations proved that these hydration reactions in principle occur in all different types of cement containing other main constituents besides clinker. The difference is whether the other main constituents act as latent hydraulic or pozzolanic materials and take part in the process in due time or only function as fillers. The evaluation of their hydraulic reactivity is therefore of practical importance and should be compared to the high hydraulic activity of Portland cement, taking physical aspects, such as particle fineness, into consideration.

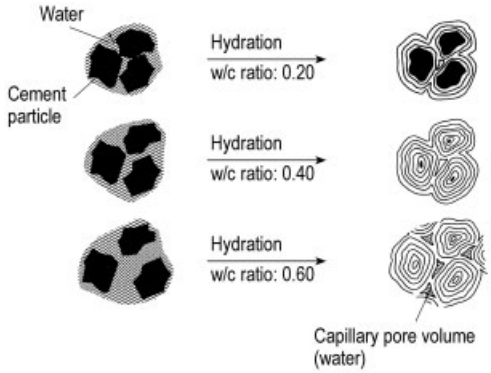

Figure 22. Influence of w/c ratio on the structure of hardened cement paste [11]
w/c ratio 0.20:incomplete hydration; w/c ratio 0.40: complete hydration; w/c ratio 0.60: complete hydration and water-filled capillary pores

Calcium hydroxide in the structure of hardened cement paste does not contribute to strength but is responsible for keeping the pH at the necessary high level. Together with a dense structure it secures the passivation of reinforcement in concrete. Other main constituents besides clinker consume $Ca(OH)_2$ to form hydration products in equal or similar amounts to those of the clinker hydration. For this reason it is necessary to limit the addition of pozzolanic materials to cement and especially concrete to keep the $Ca(OH)_2$ reserve at a sufficiently high level [205].

If the hardening process is delayed, e.g., by low temperature or by retarding admixtures, the second hydration stage is generally extended. As a consequence, more of the long, fiberlike calcium silicate hydrate may form, and higher final strength than under normal hardening conditions is attained. If the hydration is accelerated, the initial strength is higher, but the final strength is usually lower. This effect may not only be caused by admixtures but also by changes in the composition of the mixing water solution. The solubility of alkali metals and sulfates and their amount in solution (available alkali metal

content in cement and degree of sulfatization), as well as changes in the equilibrium of alkali metal and calcium hydroxides may lead to similar effects.

1.6.3. Heat of Hydration

The hydration of cement is an exothermic process. The released heat increases the temperature of concrete, especially if the heat loss of the structure is small. This fact must be considered in the construction of bulky foundations and in the construction of tunnels to avoid cracks caused by internal stress and residual stress [204]. Additionally, the heat release accelerates the hydration process under semiadiabatic and adiabatic conditions. This may improve the progress of construction but can also impair the transport time and workability of concrete.

Heat of hydration measurements are standardized for cements [170], [171], with procedures for estimating the heat of hydration under isothermal and semiadiabatic conditions.

The heat developed during complete hydration under isothermal conditions depends on the composition of the cement. The ranges of isothermal heat development (in J/g) after complete hydration for different types of cement are listed in the following:

CEM I	375 – 525
CEM II/B-S, CEM III	355 – 440
CEM II/A-B/P, CEM V	315 – 420
High-alumina cement	545 – 585

The contributions of the main clinker phases to the total heat of hydration (in J/g) of Portland cement at a degree of hydration of 100 %, calculated from studies with cements of different composition [43], are as follows:

Tricalcium silicate	500
Dicalcium silicate	250
Tricalcium aluminate	865
Calcium aluminoferrite	420
Free calcium oxide	1160
Free magnesium oxide	850

The relationship between the heat release and its contribution to strength varies with the cement component. Tricalcium aluminate and tricalcium silicate contribute more heat per unit of strength than dicalcium silicate. The more reactive a cement is, the faster it hardens and the quicker it releases heat of hydration. The hardening rate increases with the fineness of the different cement constituents and, in the case of Portland cement, with the contents of tricalcium silicate and tricalcium aluminate.

1.7. Hardened Cement Paste

1.7.1. Binding of Water

Immediately after adding the mixing water the hydraulic cement particles react at the surface with water, forming initial hydration products. The hydrated proportion is relatively small until the dormant or induction period starts (Section 1.6.2). Apart from a small amount of water which is chemically bound, the remaining water is adsorbed on the surface of the hydration products and forms a film which surrounds all cement particles and fills the gaps between the particles. Depending on the water/cement ratio, the particles commonly remain mobile in the paste, which shows a distinct consistency. In contrast to cement clinker particles, the other main constituents of cement (e.g., blast-furnace slag, siliceous fly ash, and filler) practically do not bind any water chemically at this first stage of the hydration process. The structure is loose and hollow. It densifies with time and proceeding hydration reactions.

A clear distinction between the chemically bound water and water which is absorbed or contained in pores cannot be made. A differentiation is, however, possible between ″evaporable″ and ″nonevaporable″ water [44]. Evaporable water is defined as that portion which escapes when the hydrated cement is dried at 20 °C at the partial water vapor pressure of ice at − 79 °C (temperature of a cooling trap operated with dry ice and acetone, D drying) [45]. Nonevaporable water escapes completely only at 1000 °C. It is chemically bound and therefore a part of the hardened cement. Part of the evaporable water can also be considered as chemically bound. This especially applies to chemically bound water in trisulfate and monosulfate phases. The loss is partly reversible [192]. Similar results are known from CSH. The proportion of this water, which can be regarded as part of the solid material, is estimated at ca. 0.05 – 0.15 g H_2O

per gram of cement [164]. After complete hydration, the nonevaporable water amounts to about 0.21 – 0.24 g H_2O per gram of cement (free of water), on average 0.23 g H_2O/g. Because the nonevaporable portion increases with proceeding hydration, the degree of hydration of hardened paste can be estimated and calculated, based on complete hydration. The content of nonevaporable water of Portland cement is nearly independent of its composition.

The volume of the hydration products of Portland cement is less than the volume of the original cement and the volume of the chemically bound portion of the mixing water. The water bound in the hydration products occupies less space than liquid water. Accordingly, a volume decrease of 6 cm^3 per 100 g of original cement is observed up to complete hydration. Therefore, during hydration a volume contraction may occur, designated as chemical or intrinsic shrinkage. Dimensional changes or cracks can be avoided if the hardening paste can absorb a corresponding amount of water under wet curing conditions.

1.7.2. Structure

The structure of hardened cement paste is developed from hydration products and from residual nonhydrated constituents of the cement. The hydrated fraction of Portland cement contains calcium silicate hydrates (CSH) in combination with adsorbed water in gel pores, calcium hydroxide (portlandite), calcium aluminate (ferrite) hydrates, calcium aluminate (ferrite) sulfate hydrates (ettringite, monosulfate) and pore space, subdivided into capillary pores, pores from chemical shrinkage, and pores from incomplete compaction [44–46]. The structure and quantity of hardened cement paste is responsible for strength and other properties of mortar and concrete. The structure development and the factors which decisively influence it are not yet fully understood. This is partly due to the extremely fine-sized hydration products, especially CSH. Improved structural concepts could be derived from the results of further investigations on mechanical properties of the paste in mortar and concrete. At present the structure model of Powers [45], [47], [50] is widely used to explain the properties of hardened cement paste. However, the model has since been improved and extended [46], [146], [193].

In detail, the structure of hardened cement paste depends on the grain size and shape of the hydration products, and the volume and size distribution of the pores. The BET specific surface area is an indirect indication of the particle size of the hydrated phases. A value of ca. 250 m^2/g is measured for completely hydrated Portland cement with water vapor, and a value of about 150 m^2/g with nitrogen or argon as adsorbate. The higher value obtained with water vapor may be explained by the assumption that hardened cement possesses very small pores which are only permeable to water molecules but not to nitrogen or argon. Another explanation could be that a part of the chemically bound water, removed under vacuum before the BET measurement starts, is reabsorbed during the measurement and results in an unrealistically high specific surface area value.

The average particle size of the hydration products can be calculated from the specific surface area. Assuming that the particles are spherical, an average diameter of ca. 15 nm is obtained, based on a surface area of 150 m^2/g (nitrogen or argon measurement). Because of their extreme fineness, the hydration products of cement are designated as cement gel. The hydration products in the cement gel are 1000 times smaller than cement particles. They cannot completely fill the space, even when densely packed. The remaining space is designated as gel pores. The average diameter of a gel pore corresponds to the particle size of the gel.

During hydration, which is a diffusion-controlled reaction, the hydration products grow into the water space surrounding the individual cement particles. After complete hydration, which can take days or months, depending on the size of the cement particle, the gel occupies more than twice the space occupied by the original cement particle. The space available to the cement gel is dependent on the original volume of cement and mixing water. The amount of gel formed at complete hydration totally fills the space available when the amount of mixing water corresponds to a w/c ratio of 0.35 – 0.40 (Fig. 22). At a lower w/c ratio, the available space is completely filled with cement gel before the cement has been hydrated as a whole. The hardened cement then contains cement par-

ticles with a nonhydrated core. If the w/c ratio exceeds 0.4, the cement gel cannot fill the space between the cement particles even after complete hydration of the cement. Water-filled hollow space remains as capillary pores, which are 1000 times larger than gel pores.

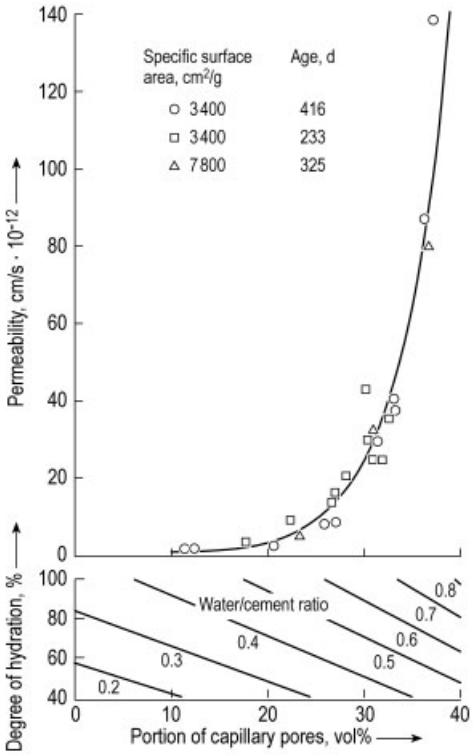

Figure 23. Permeability of hardened cement paste as a function of capillary porosity at different degrees of hydration and w/c ratios [44]

Therefore, the w/c ratio determines the magnitude of the capillary porosity of the hardened cement at a degree of hydration of 100 %. At a given w/c ratio which allows a 100 % hydration of the cement, the capillary porosity decreases when hydration is able to proceed under wet-curing conditions. It is therefore absolutely necessary to keep the hardening cement moist for a sufficiently long period to minimize the capillary porosity. This requirement should especially be met when employing cement with main constituents other than clinker or when replacing cement by pozzolanas (e.g., fly ash) in concrete and thus decreasing the hydraulic capacity of the binder.

The capillary porosity strongly determines the technical properties of hardened cement in concrete and mortar. Its permeability decreases with decreasing capillary porosity, while its strength, physical resistance against chemical attack, and frost resistance increase (Fig. 23) [50].

1.7.3. Composition of Pore Solution

When the mixing water is added, the alkali metal sulfates and parts of the free lime of the clinker, the soluble calcium sulfates of the cement, and the initial hydration products mainly determine the composition of the solution. In a similar way admixtures and additions in concrete may influence its composition. The aluminates are precipitated by formation of tri- and/or monosulfate phases with negligible solubility. As a consequence, at the beginning of hydration the solution contains practically only alkali metal, calcium, hydroxyl, and sulfate ions, which lead to a pH of > 12.5. The equilibrium in the solution remains constant during the first few hours. Later, the SO_4^{2-} concentration decreases due to precipitation of tri- and/or monosulfates, and, under certain conditions, of syngenite ($K_2SO_4 \cdot CaSO_4 \cdot H_2O$). After a longer hydration period the pore solution mainly contains high alkali metal and OH^- but low Ca^{2+} concentrations, and the pH therefore increases. In the individual case, the equilibrium concentration and its influence on hydration and strength development depend on the actual alkali metal contents and their degree of sulfatization in the clinker, on the type of calcium sulfate addition, and on the hydraulic reactivity of C_3A in clinker and cement [194]. The concentration of alkalis decreases in the presence of alkali-reactive components in cement, mortar and concrete.

The high OH^- concentration and pH values above 12 in the pore solution protect steel reinforcement in concrete from corrosion by formation of a passive layer on the steel surface. The alkalinity is diminished by carbonatization, which occurs when atmospheric carbon dioxide reacts with the hardened cement at relative humidities between 50 and 80 % and neutralizes the alkalinity. The rate of this reaction can be minimized by a dense structure of the hardened cement in mortar and concrete. Additionally, a sufficiently high capacity of alkalinity must be

guaranteed by the composition of cement, mortar, and concrete [205].

Chloride, bromide, nitrate, hypochlorite, thiocyanite, and cyanide ions in the solution may cause pitting corrosion on the steel surface even at high pH values. For this reason their concentrations are limited. The use of additives and admixtures containing such ions is prohibited in many standards. The Cl^- content of cements is limited to 0.1 wt %. Chloride only causes corrosion if it is dissolved in the pore solution. It can be bound in hydration products such as Friedel's salt ($3\,CaO \cdot Al_2O_3 \cdot CaCl_2 \cdot 10\,H_2O$). The remaining low Cl^- concentration follows the equilibrium between solution and hydrate phase. Carbonatization, however, decomposes chloride-containing compounds and mobilizes the chloride. Steel corrosion is not expected if the Cl^-/OH^- mass ratio in the pore solution does not exceed 1.2 [164]. Penetration of chloride from the outside, for example, from thawing salts, is diffusion-controlled and therefore dependent on the density of the structure of the hardened cement in concrete.

1.7.4. Trace Elements and Environmental Compatibility

Raw materials and fuels employed for the clinker burning process, as well as the initial materials for cement grinding and the production of concrete and mortar, contain heavy metals in trace concentrations of less than 0.1 wt %, or 0.1 g/kg [173], [175], [180]. A definition of main, minor, and trace elements is given in Figure 24 [196].

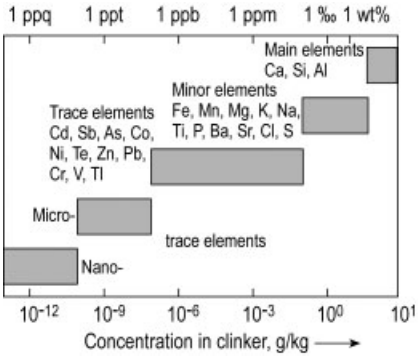

Figure 24. Concentration range for main, minor, and trace elements in clinker and cement [196]

Table 7 shows the input of trace elements into concrete by cement, aggregate, (gravel, sand), and additions (e.g., siliceous fly ash) as maximum values. The input via admixtures or mixing water is usually negligible. The comparison of the maximum heavy metal contents in concrete reveals that they do not differ significantly from the corresponding contents in natural rock.

The hardened cement paste is characterized by a high binding capacity so that a release of heavy metals is prevented, as leaching tests have shown. Elements such as Zn, Pb, and Cd in this concentration range have extremely low solubility in pore solutions of $pH > 12$. Additionally, the leachability from dense concrete is diffusion controlled [195], [203]. Therefore, the heavy metal concentration in leaching solutions lies considerably below limit values (e.g., for drinking water), even when the leaching time is prolonged, the leaching water is enriched with chemically aggressive carbonic acid, or the concrete is artificially doped with significantly higher quantities of heavy metals. Leaching tests were also carried out on concrete which was damaged in its structure by crushing, as an example for recycled concrete as an aggregate. Table 8 shows that heavy metals are largely fixed in the solid hardened cement paste matrix and that in many cases the concentration in the leaching water fell just below the detection limit.

1.7.5. Strength

Hardened cement paste, its original water addition (w/c ratio), and the degree of hydration establish the basis of strength. The paste forms a continuous framework of very fine calcium silicate and calcium aluminate/sulfate hydrate particles, in which coarser calcium hydroxide crystals and unhydrated grains of clinker or other main constituents together with voids of different size and shape are embedded. The strength of mortar and concrete additionally depends on the strength of the single aggregate grain, the adhesive pull strength between aggregate and hardened cement matrix, and the packing density of cement and aggregate grains according to their particle size distribution. The term "cement" includes clinker and other main constituents as well as calcium sulfates. Crystals of $Ca(OH)_2$ in the matrix, especially in the rim or transition

Table 7. Maximum values for heavy metal contents in concrete and basic materials for concrete in ppm (g/t) [175]

Element	Cement		Gravel, sand	Siliceous fly ash	Concrete
	Range	Max.			
Arsenic	2 – 15	–			
Beryllium	< 0.7 – 1.1	–			
Lead	5 – 105	260	20	800	100
Cadmium	0.01 – 1.5	6	1	4	3
Chromium	10 – 90	130	70	330	100
Nickel	10 – 50	100	10	300	50
Mercury	< 0.01	0.2	0.1	1	0.2
Thallium	< 0.01	4	1	4	
Vanadium	20 – 100				
Zinc	4 – 350	680	50	910	200

zone between aggregate grain and the matrix of normal paste, decrease the strength. The transition zone has a depth of 25 – 100 µm.

The strength of the hardened cement paste is generally higher with a higher proportion of long, fiberlike calcium silicate hydrate and a lower calcium hydroxide content. When cement hardens slowly at low temperatures or under the influence of retarding admixtures, the proportion of long calcium silicate hydrate crystals increases. The amount of calcium hydroxide present in the completely hardened cement decreases with decreasing alite and increasing belite contents in Portland cement clinker and with increasing amounts of other main constituents. Additions of siliceous fly ash and/or silica dust in the concrete also diminish the $Ca(OH)_2$ content. However, these cements harden more slowly, so that the potential higher final strength only becomes apparent after prolonged moist curing.

Table 8. Leaching (DEV-S4) of crushed concrete [175]

Element	Content in cement, mg/kg	Content in leaching water, mg/L
As	5	< 0.0002
Be	1	< 0.0002
Cd	< 0.5	< 0.0001
Co	10	< 0.0002
Cr	58	0.003
Cu	22	0.0008
Hg	< 0.05	< 0.0002
Tl	< 0.5	< 0.0002
Mn	500	0.002
Mo	1	0.0004
Ni	28	0.002
Pb	17	0.001
S	1	< 0.0002
Sn	3	< 0.0002
V	66	< 0.0002
Zn	310	0.001

After hardening, dry cement stone achieves a higher strength than wet cement stone. Soaking of liquids other than water, e.g., mineral oil, also reduces strength. The porosity has the greatest influence on the strength of hardened cement. The strength increases even faster than the porosity decreases, as shown in Figure 25, which depicts the results of different investigations [164]. Although the effect varies with the type and size of pores, all pores participate in the reduction of strength.

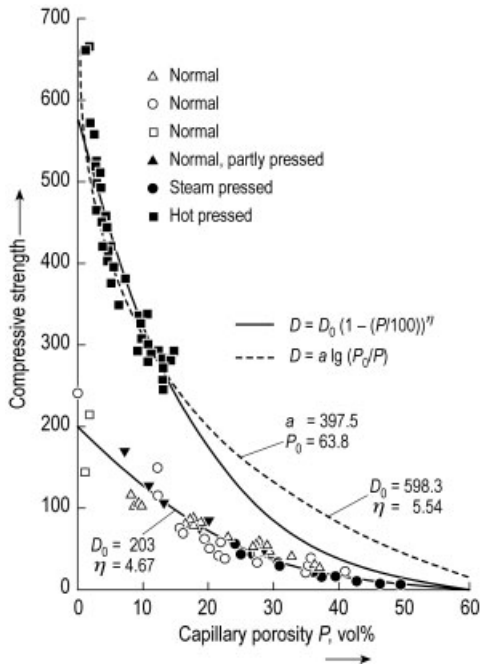

Figure 25. Correlation between capillary porosity and compressive strength of hardened cement paste [164]

The influence on the strength of hardened cement paste exerted by the porosity can be described by the following formula:

$$D = D_0\,(1 - P)^n$$

where D is the compressive strength of hardened cement paste, D_0 is the compressive strength of pore-free hardened paste, P is the fraction of the volume occupied by the pores, and n is a constant. The compressive strength of Portland cement depends mainly on capillary porosity, which increases with an increasing w/c ratio and decreases with an increasing degree of hydration. The flexural and tensile strength of hardened cement paste with a very low w/c ratio is about 10 % of the compressive strength. As the w/c ratio increases, and with it the porosity, the flexural strength diminishes more rapidly than the compressive strength. At high w/c ratios, the flexural strength diminishes less rapidly than the compressive strength. The adhesive pull strength of the transition zone is estimated to be 70 % of the tensile strength of hardened cement. Hence, this zone is a weak point in the structure of concrete.

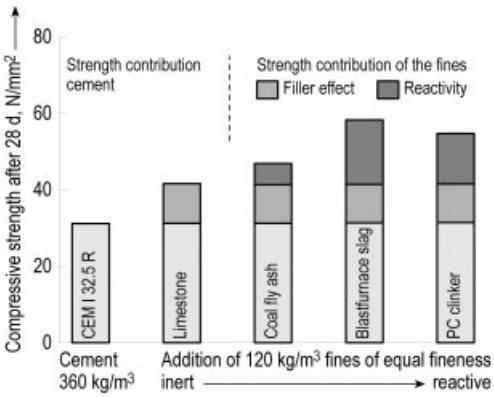

Figure 26. Strength development of mortar with increased binder content by utilization of finely sized filler, pozzolanic, and latent hydraulic materials and clinker [167]

There are in principle the same relations for cements with main constituents other than clinker. However, the effect of these constituents as inert fillers and as latent hydraulic or pozzolanic components must additionally be taken into account. Moreover, in case of high grinding fineness all these materials increase the packing density in cement, mortar, and concrete, and higher strength is achieved. Unlike inert constituents, siliceous fly ash and blast-furnace slag lead to a further increase in strength (Fig. 26) [167]. In the case of blast-furnace slag this increase is similar to that of using clinker as a fine addition. To achieve a performance of cements with several main constituents (CEM II-B/S, CEM II-B/LL) in mortar or concrete which is comparable to that of Portland cement, the cement content per cubic meter of concrete or the hydraulic activity must be increased by further fine grinding of the clinker portion and/or the other constituents.

1.7.6. Deformation

Modulus of Elasticity. Hardened cement paste is a viscoelastic material, that is, deformation caused by stress is partly elastic and reversible and partly viscous and irreversible. The modulus of elasticity E is a measure of the relationship between stress and the elastic part of deformation. Elastic deformation increases more than proportionally with stress. Hence, Hooke's law does not apply to cement stone. For reasons of comparison, the elastic deformation used to calculate E is one that reaches one third of the fracture stress after repeated loading and unloading.

The E value is influenced by the same factors as the compressive strength. It increases with decreasing porosity:

$$E = E_o\,(1 - P)^n$$

Using experimental capillary porosity data gives an E_o value of 10 000 – 30 000 N/mm^2.

When the total evaporable water is used to determine the porosity, E_o is about 70 000 – 80 000 N/mm^2, and n about 2.5 – 3.5 [164].

Shrinkage and Expansion. Shrinkage and swelling are volume changes that occur when hardened cement paste is dried or wetted. This behavior is due either to volume changes of the calcium silicate hydrate layers, caused by the reversible diffusion of water into and out of interlayer sites [46], or to a change of the force on solid particles as a result of the surface tension of water [47].

The amount of shrinkage after the first drying of cement stone can be as much as 10 mm/m (1 %). In mortar, shrinkage amounts to about 2.5 mm/m, and in concrete of normal composition to about 0.5 mm/m. The subsequent changes in length occurring on re-drying and re-wetting are reversible and cover a range of 3 – 5 mm/m [48].

Shrinkage caused by the first drying increases with increasing capillary porosity, that is, higher w/c ratios. Shrinkage is a property of hardened cement paste. Thus, higher shrinkage figures occur when larger amounts of hydrated paste are present. Finely ground cements tend to shrink more than coarser cements when they need more mixing water; the composition of the cement is of minor importance. Expansion of cement stone stored constantly under water is much less than shrinkage, only about 1 mm/m. The exchange of water by drying and re-wetting takes a certain time, its actual duration depending on the dimensions of mortar and concrete specimens. Therefore, due to different moisture contents, tensile stress may occur, leading to cracks.

Creep. Creep is a remaining deformation occurring under constant load in addition to elastic deformation and drying shrinkage. Creep proceeds rapidly at first and slows down after one to two years. After the load has been removed, the deformation only partially reverses. Creep is roughly proportional to the load and generally lower at higher strength or concrete age. It also depends on the moisture content in concrete. If loading and drying takes place simultaneously, creep is larger than in the case of drying prior to loading. If the change in the moisture content occurs before the load is applied, creep decreases increased contents of evaporable water. Because drying takes months, even in the case of concrete, substantial creep must always be expected, even when the load is applied later after the manufacture of the concrete construction unit.

After relief of the load on a concrete specimen under laboratory conditions, the total deformation value partly decreases for the proportion of elastic deformation and for the retarded elastic deformation. The remaining deformation value is irreversible and corresponds to a viscous flow. The ratio of creep deformation to load (tension) is designated as specific creep figure. The creep figure is given by the ratio of creep deformation to elastic deformation. The reasons for creep deformation are partly still unknown.

Thermal Expansion. The thermal expansion of hardened cement paste depends largely on its moisture content. It is the lowest in the case of completely dried cement stone. The coefficient of thermal expansion is then $10-12 \times 10^{-6}\,K^{-1}$, the maximum at a relative humidity of 70 % is $24 \times 10^{-6}\,K^{-1}$ [49]. For concrete, the coefficients of the hardened cement paste, the different kinds of aggregates, their proportion and the humidity content are decisive [164].

Permeability. The water permeability of hardened cement paste depends exclusively on capillary porosity. The permeability is very small, almost zero, when the proportion of capillary pores is less than 20 vol % [50]. This low capillary porosity is obtained at a w/c ratio of ca. 0.5 and on complete hydration. Hardened cement paste with a w/c ratio of 0.7 has so many capillary pores that it remains permeable to water even after complete hydration. The correlation between water permeability and capillary porosity, w/c and degree of hydration of hardened cement paste is given in [50], [164] (Fig. 23). Dense aggregates have a permeability similar to completely hardened paste ($w/c < 0.7$). The transition zone between hardened cement paste and aggregate may raise the permeability and diffusion under certain conditions. Gaseous and dissolved organic and inorganic components can also penetrate concrete or mortar. They can cause physical or chemical changes which impair the durability. Besides diffusion-controlled permeation (differences in concentration or partial pressure), permeation may also be promoted by pressure differences.

Resistance to Chemical Attack. The resistance of mortar and concrete to chemical attack by the environment is based on physical and chemical preconditions. Physical requirements are a low permeability and a high resistance to diffusion of the hardened cement paste, as well as of the concrete microstructure. A minimized reaction potential against attacking constituents in water, soil and air is the decisive chemical requirement. Two kinds of attack are differentiated: dissolving attack and attack leading to expansion. In both cases the concrete or mortar may be damaged. Dissolving attack starts at the surface of the concrete. The hydrate phases of the hardened cement paste are decomposed, and the reaction products dissolved and carried away with the water flow. An expanding attack develops when reactive components penetrate

the concrete and react with constituents of the hardened cement paste, so that expanding compounds are formed as reaction products. In this respect efflorescence products at the concrete surface can, in a broad sense, also be regarded as products of chemical reactions which are transported to the concrete/mortar surface.

A slight dissolving attack occurs in water with a hardness of less than 5°d, and more severe attack in water containing CO_2 or dissolved salts with exchangeable cations, such as NH_4^+ and Mg^{2+}, and acids (pH). Sulfate penetrates the concrete or mortar structure by diffusion and reacts with calcium aluminate hydrate to form ettringite, or reacts with calcium hydroxide to form secondary gypsum. The reaction may cause damaging expansions. Therefore, the concentration of attacking components in water, soil, and gas of natural origin is limited by various standards [154]. The aggressiveness of sea water is evaluated in different ways, depending on regional and climatic conditions.

For the prevention of attack the aggressiveness is subdivided into three or more groups with ascending concentration ranges for salts, and with descending ranges for the pH. According to this subdivision, physical and chemical measures have to be taken, for example, measures to improve the density of the hardened cement paste or concrete structure and the employment of sulfate-resistant cements (Section 1.3.2). If siliceous fly ash or natural pozzolana is used as a main constituent in the cement or as an addition in concrete, the release of alumina and the formation of aluminates as a consequence of pozzolanic reactions should be considered.

Gases can also penetrate the hardened cement paste and react with its components in the presence of moisture. Hydrogen sulfide and sulfur dioxide are oxidized to sulfuric acid, which dissolves the hydrate phases in the first stage and forms expanding compounds in the second. Carbon dioxide reacts with calcium hydroxide, calcium silicate hydrates, and calcium aluminate hydrates of the hardened paste to form calcium carbonate. The strength of the cement stone may generally be increased by this process. Carbonation, however, can impair the protection of the steel reinforcement in concrete.

1.8. Testing of Cement

The requirements on the composition and properties of cement are stipulated in the various national and European standards [51], [70], [71], [197]. The most important standardized properties of cement are strength, setting and dimensional stability (soundness), pozzolanicity, and fineness. Depending on the type of cement, limits for the contents of blast-furnace slag, pozzolana, fly ashes, silica fume, burnt oil shale, and fillers are established. In addition, the cement standards specify certain characteristics, such as limits for the insoluble residue, pozzolanicity, the loss on ignition, and the contents of carbon dioxide, magnesium oxide, sulfate, and chloride. Cements with special properties must meet additional requirements concerning composition or heat of hydration. Methods of cement testing are described in DIN-EN 196, parts 1 – 9:

Part 1: Determination of strength
Part 2: Chemical analysis of cement
Part 3: Determination of setting time and soundness (expansion)
Part 4: Quantitative determination of constituents
Part 5: Pozzolanicity test for pozzolanic cements
Part 6: Determination of fineness
Part 7: Methods for taking and preparing samples of cement
Part 8: Hydration heat-solution method
Part 9: Hydration heat-semi-adiabatic method

Other test producers are described in [70], [71].

In many countries conformity with the standard requirements has to be checked; in the EU member states this follows the principles of the Construction Products Directive [198]. A precondition for an attestation of conformity by a certification body is not only a factory production control by the manufacturer but also third-party inspection by an inspection body. The factory production control includes an auto-control testing on samples of finished cement and an internal quality control with regard to manufacturers' operational techniques and activities to fulfil the requirements of the product standard. The third-party inspection comprises the inspection of the factory production control and audit

testing according to test methods of the EN 196 series [199].

Setting. The standard test on setting (DIN-EN 196-3) evaluates the workability of cement at standard water demand at initial setting time. In practice, the final set is not important. According to the various national cement standards, the initial setting time must not be less than 0.75 – 1 h, and the final setting time not more than 6 – 12 h, depending on the type of cement. The test is carried out by using the cylindrical Vicat needle. The method differentiates between normal, quick, and false set as a property of the cement. For the assessment of the cement behavior in mortar and concrete additional tests are necessary.

Strength. The strength development of cement is tested on mortar samples. The dimensions of the specimen, the composition of the mortar, the kind of curing, the time of hardening and the test procedures are defined in various cement standards. Many countries have adopted the ISO-RILEM-CEM procedure, which has been incorporated into the European cement standard.

According to DIN-EN 196-1 the compressive strength is tested on mortar prisms, $4 \times 4 \times 16$ cm, consisting of 1 part by weight of cement, 3 parts by weight of standard sand mixture, and 0.5 parts by weight of mixing water. The strength of the prisms is tested after 2 or 7 and 28 d. The strength classes 32.5, 42.5, and 52.5 are designated according to the minimum requirements of compressive strength after 28 d of wet curing. Each strength class, with the exception of class 52.5, is additionally limited by a maximum strength characteristic value. The range between minimum and maximum strength is 20 N/mm^2. The objective of this additional limitation is, for safety reasons, to obtain similar strength and comparable properties of cements of different origin within one class. Additional requirements on early strength (2 d) are indicated by the letter R after the strength class designation.

Soundness. If hardened cement paste is dimensionally stable, it is designated as sound. Expansion is observed only in those cements that contain larger amounts of free lime or free magnesium oxide due to the formation of expanding hydroxides after reaction with water. Sulfate expansion may occur if more calcium sulfate is added as retarder than can be consumed for the reaction with calcium aluminates during the initial period of time in which the paste is still deformable.

For evaluation of the volume stability the Le Chatelier test is used. The test is carried out by using cylindrical test molds filled with cement paste of standard consistency. After 24 h of moist curing, the specimen is boiled for 3 h in the mold. The mold is a flexible strip of brass with 150 mm long needles at each end. The distance between the needles is a criterion for the expansion. Cement is designated as dimensionally stable if the expansion does not exceed the specified limit of 10 mm.

This method only assesses the expansion by free lime. Expansion due to free magnesium oxide or sulfate is prevented by limiting the contents of magnesium oxide in clinker and of sulfur in cement. According to EN 197-1 the MgO content in Portland cement clinker must not exceed 5.0 wt %. The requirements limit the sulfate content (as SO_3) to 3.5 and 4.0 wt %, respectively, depending on the cement type and cement fineness. Expansion caused by free lime and free magnesium oxide can also be determined by using the autoclave method specified in ASTM standards.

Fineness. The specific surface area according to Blaine [37] is taken as a measure for the fineness of cement. It is calculated from the air permeability of a compacted cement bed, its porosity, the cement density, and the viscosity of the air. The range of specific surface areas which can be estimated with this method lies between 2000 and 6000 cm^2/g. More advantageous and practical is the employment of a (wet) laser granulometer to estimate the particle size distribution and to calculate the surface.

Heat of Hydration. The heat of hydration of cement can be estimated by using the classic solution method or the more recently standardized semi-adiabatic method. Both methods are described in the European cement standards EN 196-8 and EN 196-9.

The solution method consists of calculating the heat of hydration from the difference bet-

ween the heats of solution of an anhydrous cement and of a cement paste (w/c = 0.4) hydrated under isothermal conditions at 20 °C, e.g., for 7 d, in a sealed tube.

The semiadiabatic method consists of determining the quantity of heat emitted from a sample of fresh mortar in the calorimeter by measuring the temperature rise. At a given time the heat of hydration of the cement contained in the sample is equal to the sum of heat accumulated in the calorimeter and the heat loss during the period of testing. The heat loss is estimated by comparison with the temperature loss of an inert sample in a reference calorimeter. The heat of hydration of cement estimated under isothermic conditions at 7 d and under semiadiabatic conditions at 41 h can be taken as equivalent.

1.9. Environmental Protection

Clinker production is a material conversion process. The energy input is used to bring the material to reaction temperature and to mect the demand of endothermic reaction enthalpie. It takes place in direct contact with the product, leading to immediate reactions between fuel and raw materials. Therefore, the resulting emissions are dependent on the interaction between the two. Environmental protection includes all parts of a factory. Besides emissions from the kiln system, which comprise particulate and gaseous components, emissions of noise and vibrations, caused by the operation of machines, traffic, and blasting, must also be considered. Finally, also the effort for recultivation or renaturalization of old quarries belong to the important protection targets. The erection and operation of plants for the production of cement are subject to the Federal Immission Law (*Bundes-Immissionsschutzgesetz*) in Germany and to several EU Directives. These regulations also contain limit values for the emission of particulate matter, gaseous components, noise, and vibration.

1.9.1. Particulate Emissions

In the course of cement production about 2.6 – 2.8 t of raw materials, fuels, cement clinker, and other main constituents of cement are handled as dusty materials to produce 1 t of cement. The dust, dispersed in air and gas streams, is nearly completely precipitated in dust filters and fed back into the material streams of the production process. The main sources of dust emissions are the kiln waste-gas stream and the waste air of the raw material and cement grinding systems. Additionally, the emission of diffuse sources, e.g., from drilling and blasting of raw materials or transport systems of different kind, has to be considered.

The dust is mainly precipitated in electric and bag filters. The cleaned waste gas (1.5 – 2.5 m^3 (STP)/kg clinker) contains about 10 – 30 mg dust/m^3 (STP) on average per day. Construction size and electrical power input of electrical precipitators increase exponentially with decreasing dust concentration. Therefore, besides the investment costs, requirements on lower dust concentration in the waste gas stream of a kiln may increase the specific electric power consumption significantly. Since 1950 the efficiency of dust precipitation has been improved continuously.

The composition of dust differs. It depends on the source, on the place where the sample was taken, and on process conditions. Precipitated cement kiln dust contains predominantly raw material components and smaller amounts of alkali metal sulfates and halides. Heavy metals such as As, Pb, Cd, Cr, Co, Ni, Tl, and Hg, are present but usually only in trace concentrations [53]. The occurrence of alkali metal sulfates and halides, predominantly chlorides, as well as of heavy metal compounds and their enrichment is a consequence of the establishment of cycles within the kiln system. Volatile components evaporate in hot zones and condense in cooler zones of the kiln system, partly after further reaction, e.g., in the calcinator or preheater. They are transported back to the kiln and build up so-called internal cycles. External cycles develop, when volatile components are transported into the drying/grinding mill area or in the dust precipitator, and when the materials are recycled (Fig. 27). Calcinator, preheater, and the drying/grinding mill have a high retention capacity, especially for volatile components. Cycles are interrupted, when the components are bound in the clinker or contained in gas or other material streams which leave the kiln system. If a common outlet does not exist, or if the input is

larger than the release, the components within an internal or external cycle accumulate. The consequence of significant alkali metal sulfate and chloride enrichments are impacts on the process by coating and ring formation. The growth of cycles can artificially be interrupted by cost intensive bypass systems [54–58]. Emissions of trace elements may take place predominantly when they are bound in the particulate matter. Under certain conditions emissions of volatile components as vapor are possible. Balance measurements allow the estimation of the magnitude of such cycles and the evaluation and prognosis of possible emissions, dependent on the individual input of minor and trace elements into a kiln system. Emissions are limited according to the regulations of immission laws and directives.

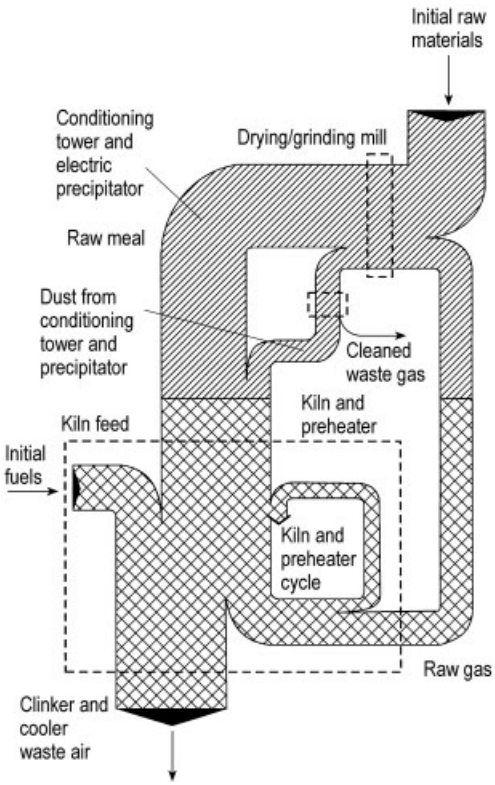

Figure 27. Internal and external material circuits in the process of clinker burning [55]

The volatilities of the heavy metal compounds vary considerably at high temperature and in the alkaline and oxidizing atmosphere of the cement kiln. The rate of retention in the clinker is defined as the ratio of the amount removed from the kiln with the clinker to the total amount introduced into the kiln. It characterizes the volatility of elements. Table 9 lists the order of volatility of some elements [55], [167]. Under normal process conditions, nonvolatile elements and their compounds are predominantly bound in the clinker. A proportion of < 0.1 % of the balance input may be emitted in a dust-bound form. Measurements have proved that the actual emission figures of cement kilns are far below the legal limit values at a high efficiency of the dust precipitators.

Table 9. Trace elements in cement kiln cycles according to their volatility [55], [167]

Volatility	Elements	Condensation temperature, °C
Nonvolatile	Zn, V, Be, As, Co, Ni,Cr, Cu, Mn, Sb,Sn	
Slightly volatile	Cd, Pb	700 – 900
Volatile	Tl	450 – 550
Highly volatile	Hg	< 250

The environmental compatibility of the cement production process can be evaluated by considering pollution levels in the vicinity of plants. Such investigations have shown that cement plants are no longer the main source. The dust deposition in rural areas where cement is produced is as low as in areas without cement production. Based on the example of heavy metals, Table 10 shows that the forecast of 30-years input into the soil in the vicinity of a cement plant, complying with the legal emission limit values, lies significantly below the respective soil guide values for unrestricted use. If the calculations were based on actual emission concentrations instead, the input into the soil would be even less [175].

1.9.2. Gaseous Emissions

Gaseous emissions of the cement production process are mainly based on the burning process and emissions from raw materials. The exhaust gas from the cement kiln consists primarily of nitrogen, carbon dioxide, oxygen, and water vapor. The CO_2 concentration deviates between 22 and 35 vol %, depending on the process (see Section 1.4.5). In addition, the proportions of sulfur dioxide, nitrogen oxides, carbon monoxide, and organic carbon compounds have to be considered [52], [167].

Table 10. Environmental relevance of heavy metals emitted by a rotary kiln system in a cement works at maximum permissible concentrations (limit value) by total accumulation over 30 years [175]

Metal	Guide values∗, mg/kg	Forecast 30 year, µg/kg	Fraction, %
Arsenic	20	11	0.06
Lead	100	182	0.18
Cadmium	1	4	0.40
Chromium	50	56	0.11
Cobalt	30	11	0.04
Copper	50	34	0.07
Nickel	40	56	0.14
Thallium	0.5	0.7	0.15
Mercury	0.5	14	2.7
Zinc	150	20	0.01
Tin	50	11	0.02

∗ Guide values for soils subjected to multifunctional use according to Eickmann – Kloke.

Under oxidizing burning conditions, the sulfur input of the cement kiln with the raw materials and fuel is converted into gaseous sulfur dioxide. Most of the sulfur dioxide reacts with additional oxygen and vaporized alkali metal compounds to form alkali metal sulfate [56], [57]. Alkali metal sulfates build up cycles. The input equals the output with clinker and dust. In the presence of excess alkali metal, the sulfur dioxide emission is below 0.4 g/m^3 (STP). Higher SO_2 emissions may occur if, e.g., reducing conditions in the calcinator exist, or the raw material contains oxidable FeS_2 (pyrite, markasite) releasing SO_2 at temperatures below 700 °C in the preheater system. Secondary measures, like $Ca(OH)_2$ additions (dry additive or dry sorption method), support a reduction of SO_2 concentrations in the uncleaned waste gas stream. The additives bearing the absorbed sulfates may, however, cause disturbing coating and ring formation when passing through the kiln system.

The high temperature and the oxidizing conditions of the clinker burning process, required to obtain high-quality clinker, lead to the formation of nitrogen oxides. Nitrogen monoxide and nitrogen dioxide are commonly present in a volume ratio of 95:5. Process-dependent NO_x emissions without reducing measures amount to 1.2 – 1.8 g/m^3 (STP). Primary (e.g., burner optimization [200]) and secondary reduction measures (e.g., staged precalcination and selective, noncatalytic reduction (SNCR) with NH_3 [201]) make it possible, after optimization, to meet NO_x emission limits of 0.8 g/m^3 (STP) in existing plants and of 0.5 g/m^3 (STP) in new plants [52]. Catalytic NO_x reduction (SCR) still belongs to the emerging technologies within the Portland cement clinker production process, and full-scale industrial experience could not yet be gathered.

Gaseous fluorine as HF [58] and chlorine compounds as HCl [57] are not emitted.

The counter flow of kiln feed and gas, the high gas temperature of > 2000 °C in the sintering zone, as well as a residence time of 3 – 4 s at gas temperatures of > 1200 °C and a sufficiently high oxygen content offer extremely favorable conditions for complete destruction of organic compounds in common and secondary fuels or in replacement materials. This also applies to toxic compounds like dioxins, furans, polychlorinated biphenyls (PCBs), and polynuclear aromatic hydrocarbons (PAHs). Fresh raw meals fed to the kiln contain up to 0.5 wt % of total organic components (TOC) as kerogens and bitumens. These materials volatilize and decompose at temperatures of 200 – 700 °C, forming mainly carbon monoxide, carbon dioxide, and organic constituents, measurable e.g., as total C. Therefore, the clean gas of kiln systems may contain higher concentrations of CO (< 1.5 g/m^3 (STP)) and total C (< 0.015 g/m^3 (STP)) and exceed existing limit values. This is, however, mainly due to TOC in the raw materials and cannot be taken as an indicator for incomplete combustion [52], [167].

Grinding aids such as ethylene glycol are added during the grinding of cement. At least 85 % thereof are bound to the cement by chemisorption. The rest is emitted as vapor [59].

1.9.3. Noise and Vibration

Noisy parts of a cement plant, e.g., tube mills with the revolving charge of grinding balls, are either accommodated in sound-absorbing buildings, covered by a soundproof capsule, or enclosed within soundproof walls. Disturbing noises produced by fans at the intake and exhaust outlets are eliminated by the installation of silencers. In addition, the working area of the plant operators is soundproofed. Vibrations produced by blasting can be significantly reduced by using appropriate blasting methods, e.g., sequential blasting at millisecond intervals.

1.10. Toxicology and Occupational Health

Portland cement clinker is produced from cement raw meal as a physical mixture of inorganic mineral constituents, each having its characteristic properties. The clinker, burnt from this mixture, can be designated as a preparation, consisting of different mineral phases. Cements are a mixture of different materials in form of a preparation, both in the sense of the German Chemicals Act (Table 11).

Table 11. Definition of cement as a preparation

Raw meal	Portland cement-clinker	Cements as defined in DIN 1164
Raw material mixture	Preparation	Preparation
Limestone: $CaCO_3$	$3\,CaO \cdot SiO_2$ $2\,CaO \cdot SiO_2$	Initial constituents: – Portland cement-clinker
Clay: Kaolinite	$3\,CaO \cdot Al_2O_3$ $2\,CaO \cdot (Al_2O_3, Fe_2O_3)$	– $CaSO_4 \cdot n\,H_2O$ – Granulated blast-furnace slag (S)
$(Al_2O_3 \cdot 2\,SiO_2 \cdot 2\,H_2O)$	CaO	– Siliceous fly ash (V)
	MgO $(K, Na)_2SO_4$	– Burnt oil shale (T) – Limestone (L)
Illite $[(K_2O, H_2O) \cdot 3\,Al_2O_3 \cdot 6\,SiO_2 \cdot 2\,H_2O]$		– Trass (P) Cements:
Montmorillonite $(Al_2O_3 \cdot 4\,SiO_2 \cdot H_2O \cdot n\,H_2O)$ Sand: SiO_2		– CEM I – CEM II A/B, with S, V, T, L, P – CEM III A/B, with S

After addition of mixing water and during hydration of cement, an alkaline solution forms with a pH of > 12.5. Toxicological and dermatological investigations, following the requirements of the German Chemicals Act and the European Directive (67/548 EWG) on classification, packaging, and labeling of dangerous substances have shown that cement must be classified as Irritant X_i, with the following risk

R 38: Irritant for skin
R 41: Danger of severe eye damage
R 43: Sensitization possible by skin contact

and safety phrases [202]:

S 22: Do not breath dust
S 24: Avoid contact with skin
S 25: Avoid contact with eyes
S 26: In case of contact with eyes, immediately wash eyes and contact a physician
S 36: Use appropriate protective clothing

The safety data sheet summarizes the characteristic chemical, physical, and health data for each cement and gives recommendations, for a safe storage, handling and transport, following the EC-Directive 91/155/EWG.

The assessment of environmental compatibility of cement in concrete or mortar has to consider all the phases of its life cycle. This includes both the manufacture of cement and its use. Whereas the compatibility of the production process is characterized by the pollution levels in the vicinity of production plants (see Section 1.9), the environmental compatibility of fresh and hardened mortar and concrete has to be evaluated during the time period of their manufacture, construction, and employment after hardening.

Heavy metals are present in varying concentrations in all starting materials, including aggregates, additions, additives, mixing water, and cement, not exceeding concentrations of 100 g/t [175]. As an example, Table 7 summarizes the maximum heavy metal contents of starting materials and their maximum content in concrete.

In the solution which forms instantly after addition of mixing water, hexavalent chromium is the major health-relevant constituent and is present in concentrations of about < 2 to 20 ppm. This is equivalent to about 10 – 20 % of the total amount of Cr(VI) in cement which, above all, originates from the raw materials [60]. Only in the case of continuous skin contact and after 10 to 20 years of employment brick- and

floorlayers may become sensitized. The sensitization is the precondition for the formation of allergic skin eczema. The risk of skin sensitization can be decreased by chemical reduction of Cr(VI) to concentrations below 2 ppm (e.g., by intergrinding or addition of iron(II) sulfate) and, most effectively, by skin protection with special gloves made of synthetic materials, e.g., nitrile-soaked cotton gloves. The risk of an allergic skin reaction caused by concentrations below 2 ppm of soluble Cr(VI) or any other element has not yet been investigated and justifies a risk-declaration in safety data sheets. The preventive measures are common practice in several countries, e.g., in Germany for all types of bagged cement, cement-containing building materials, and mortars. As early as 2 – 3 h after adding the mixing water, especially in the later stages of hydration, Cr(VI) is firmly bound in the hydration products of cement (ettringite) during hardening.

The high chemical and physicochemical binding capacity of hardened cement in addition to a dense concrete or mortar structure prevents the release of heavy metals. The leaching of concrete is a diffusion-controlled process. Therefore, the values measured in leaching tests lie considerably below the limiting values for drinking water quality [175], [203]. This applies to dense, undamaged concrete structures as well as to crushed concrete from recycling plants.

The MAK value for Portland cement dust is 5 mg/m^3 as inhalable (total) dust. The detection of this limit value is uncertain, however, because the Portland cement portion must be separately estimated from the total amount of precipitated dust by chemical analysis.

1.11. Economic Aspects

Within the European CEN member states 27 different cement types are produced which are standardized in the European cement standard EN 197, parts 1 and 2. Besides the traditional Portland cement an increasing proportion of cements is produced which contain as main constituents, besides clinker, latent hydraulic (e.g., granulated blast-furnace slag), pozzolanic (e.g., natural pozzolanas, fly ashes), and inert (e.g., limestone filler) materials as well as burnt oil shale with several of these properties. The percentage market share of these cements is different, as a comparison of the situation in Germany and in other CEMBUREAU countries (Fig. 28) shows. In 1996 the total European production amounted to about 207×10^6 t, and world production was about 1423×10^6 t [61]. Essential reasons for the development in Europe are market requirements within the EU which differ to a large extent due to climatic conditions and tradition. This particularly applies when other main constituents, like granulated blast-furnace slag or pozzolanas, are utilized. Cement is regarded as a product for the economic manufacture of concrete of high durability. Besides the classical performance criteria, which are workability and strength development, additional requirements on frost resistance, carbonatization and protection of reinforcement as well as on a high quality level and on the homogeneity of properties have to be met. Cement with several main constituents is more and more regarded as a product with properties which can individually be adjusted to the requirements of the customer, using the technical and quality control capacity of a cement plant. This will be obtained within the framework of the cement standard or by technical approval and serves to produce concrete under controlled conditions.

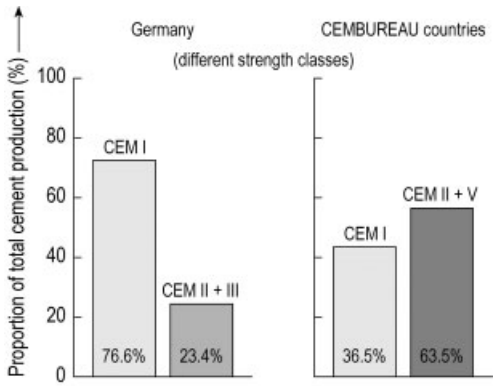

Figure 28. Market share of different cement types according to prEN 197-1 & 2 in 1996 in Europe (EC and EFTA countries, Turkey, Poland, Czech Republic, and Slovakia)

2. Concrete

2.1. Introduction

Concrete is an artificial stone in which natural stones of well-graded sizes are bound together by a cement matrix, thus providing strength. Although this principle of manufacturing building materials can be traced back at least to the Roman Empire, it was later neglected for centuries. Concrete regained importance in the 1800s, when methods for large-scale production of modern cement were developed [62], [63].

Concrete is a brittle material, but reinforcing with steel bars (reinforced concrete) or with prestressed tendons (prestressed concrete) has greatly expanded its applicability. It is now one of the most important building materials because of its structural, physical, and chemical properties and its economics. Furthermore, it can be considered an environmentally friendly material [206]. Fresh concrete can easily be shaped, and most concrete is poured in place on the construction site. However, there is increasing use of concrete elements or concrete products that have been prefabricated in plants.

In 1998, ca. 1530×10^6 t of cement was produced worldwide to manufacture ca. 5×10^9 m^3 of concrete for the construction industry. Ten countries, namely, The People's Republic of China (34.6 %), India (5.7 %), Japan (5.6 %), the United States (5.4 %), South Korea (3.2 %), Brasil (2.6 %), Turkey (2.5 %), Italy (2.4 %), Germany (2.2 %) and Spain (2.2 %), consumed almost 70 % of the world production of cement and concrete [64].

Concrete is one of the less energy-intensive building materials. Concrete based on portland cement has a total energy content of only ca. 1250 MJ/m^3 [65], and even less when waste materials or industrial byproducts such as fly ash, silica fume, or ground granulated slags are incorporated. Table 12 gives a comparison of total energy content and total energy content per unit of strength of different building materials [65–67].

The supply of natural concrete aggregates of good quality is not guaranteed worldwide, and because of continuous consumption, a shortage of suitable material is observed in some locations. Aside from more sophisticated methods of recovery or treatment, the use of low-quality aggregates for special applications as well as recycling of old concrete, other building materials, or waste products must be considered in the future [68]. Corresponding regulations are available on a national basis, e.g., [207] or are under preparation by CEN TC 154 on a European level. Recommendations have been established by RILEM [208].

Aside from its composition, the properties of a structural concrete depend to a great extent on its age and curing history as the controlling factors of cement hydration. Therefore, physical and chemical properties must be tested by standardized procedures both before and during the construction process. National standards, as well as international standards or recommendations given by the International Organization for Standardization (ISO), Comité Européen de Normalisation (CEN), Comité Européen du Béton (CEB) – now Fédération International du Béton (fib) after merging with Fédération Internationale de la Précontrainte (FIP) – , or Réunion Internationale des Laboratoires d'Essais et de Recherche sur les Matériaux et les Constructions (RILEM), cover quality control of concrete raw materials, as well as that of the concrete itself. Additional guidelines may be specified by the purchaser.

2.2. Materials

Basic materials for the production of concrete are cement, aggregates, and water. Up to certain limits, additional compounds – concrete additives – may be used to control some properties of the fresh or hardened concrete. These materials must meet specifications set forth in standards or in approval documents. The compatibility of all of the compounds used must also be considered.

2.2.1. Cement

The hydraulic binder cement reacts with water to form a highly dispersed cement paste matrix in which the aggregates are embedded. This cement matrix normally constitutes ca. 25 % of the volume of concrete. Different types of cement can be distinguished according to cement standard compressive strength, chemical composition, kinetics and heat of hydration, or resistance

Table 12. Total energy content of some building materials

Material	Mean strength, MPa	Total energy content, MJ/m^3	Total energy content/ strength
Steel	370 ∗	237 000	640
Aluminum	320 ∗	730 000	2280
Clay brick	25 ∗∗	4200	168
Sand lime brick	25 ∗∗	1220	49
Polyethylene		92 800	
Concrete	35 ∗∗	1242	35.5

∗ Tensile strength. ∗∗ Compressive strength.

to chemical or physical attack [69–72]. Economic considerations may influence the choice of the type of cement, depending on the local supply of raw materials for cement production or concrete aggregates.

2.2.2. Aggregates

The remaining 75 % of the volume of concrete is made up by aggregates, which act as inert fillers and give concrete the properties needed in a structural building material (e.g., volume stability, limited crack development, modulus of elasticity, wear resistance, ductility). The mechanical and physical properties of the aggregates may differ considerably from those of the cement paste. They are reflected in the properties of the hardened concrete but not necessarily in proportion to the concentrations of the aggregates. Concrete aggregates can be categorized according to their origin as natural or artificial aggregates; according to their density as light, normal, or heavyweight concrete aggregates; or according to their composition as mineral, metallic, or organic materials. The last two types are used only for special concretes.

Grading of Aggregates. Aggregates used in producing high-quality concrete have a particle-size distribution in the range 0.01 – 100 mm. The finer fraction is classified as sand, which comprises particles up to 4 – 5 mm in diameter; larger particles are called coarse aggregates. As a general rule, the size distribution of aggregates is selected to maximize their concentration in the concrete while maintaining sufficient workability of fresh concrete with a low water content. Therefore, the total surface area of the aggregates should be kept as small as possible [73].

The particle-size distribution of aggregates is determined by sieve analysis, which subdivides a sample into fractions. Within each fraction, particles of sizes between a lower and an upper limit are present, depending on the mesh sizes of two consecutive sieves. For two consecutive sieves, these mesh openings are normally increased by a factor of two. Specifications for sieves, as well as procedures for sieve analyses, are given in various national and international standards or recommendations [74–76].

After subdivision of the aggregates into fractions, grading is based on the specific surface area, water requirement factor, average particle size, fineness modulus, etc. The grading curve is a graphic presentation of the cumulative percentage retained on or passing through successive sieves.

In general, specifications of limiting grading curves for the aggregates as a whole, or for the sand only, are commonly used [77], [78]. The maximum aggregate size is usually specified. Concrete aggregates must not contain compounds that inhibit the hardening of the cement paste or that are vulnerable to chemical or physical degradation. Therefore, the permissible contents of organic or expansive compounds, soluble sulfates, silt, and clay are limited. Aggregates to be used in reinforced or prestressed concrete must not contain soluble compounds, such as chlorides, that attack steel [74], [78]. Aggregates that contain reactive silica, e.g., opal, demand special, low-alkali cements to prevent expansive alkali – silica reactions in the hardened concrete [79], [80], [209–211].

Other characteristics of aggregates, such as modulus of elasticity, hardness, density, porosity, thermal conductivity, thermal expansion, and resistance to wear and frost action, must be considered in many applications.

2.2.3. Water

The mixing water added in the preparation of fresh concrete serves to hydrate the cement and assures the workability of the fresh concrete. Although only 20 – 25 % of the mass of cement must be water for the chemical reaction with cement, additional water is needed to achieve complete hydration, as well as good workability.

Impurities in the mixing water, such as sugar, organic compounds, oil, carbonic acid, sulfates, or other salts, may affect the hardening of the concrete or reduce the concrete strength. The concentration of dissolved chlorides is restricted for reinforced and prestressed concrete. Doubtful cases should be subjected to chemical analysis, but normally hints of excessive amounts of impurities can be derived from the color, odor, taste, turbidity, or formation of bubbles or foam. Guidelines for permissible impurities can be found in [73], [81], and a European Standard pr EN 1008 is in preparation by CEN TC 104. In concrete plants wastewater from cleaning of production equipment from remains of fresh concrete may be re-used with certain restrictions [212].

As a general rule, potable water is suitable as mixing water, but also spring water and water from rivers or lakes if not contaminated. Sea water must not be used for reinforced or prestressed concrete because of the dissolved chlorides, which may induce the corrosion of the reinforcement. In case of doubt, a trial mix with the water should be prepared and the development of concrete strength compared to that of a control mix. A trial mix is also important because the effect of impurities on concrete properties may depend on the type and amount of cement used.

2.2.4. Additives

Compounds may be added to control specific properties of the fresh or hardened concrete. These additives are either water-soluble (mainly organic) compounds in low concentrations or finely divided mineral material in higher concentrations. The classification of additives with respect to their effect is arbitrary, because in most cases several concrete properties are influenced simultaneously; guidelines can be found in [77], [82], [83], [213], [214]. The majority of additives function to control setting and hardening, workability, or concrete porosity.

2.2.4.1. Setting and Hardening Additives

In some instances, the setting and hardening development of concrete must be accelerated or retarded to meet the needs of a construction job. Set-retarding agents are used to prolong the period of workability during hot weather, to avoid discontinuities in subsequent placements of fresh concrete, or to reduce crack formation due to deformations in an early stage. Accelerators are used when concrete is placed under cold weather conditions or when there is need to strip the formwork earlier than normal. Use of high concentrations of accelerators in shotcrete (see Section 2.4.4) leads to rapid development of strength only seconds after mixing.

2.2.4.2. Workability Additives

Many types of additives exert a significant influence on concrete workability while also affecting other properties such as the setting time. The main objective of using such additives is to improve concrete workability. These plasticizers or superplasticizers reduce the amount of water needed and are the most commonly used additives, because they facilitate the placing and compacting of the concrete. Because they permit the use of lower water-to-cement (w/c) ratios, the strength and durability of the concrete is improved [84].

2.2.4.3. Porosity Additives

The resistance of concrete to frost and deicing salt can be improved substantially by the use of air-entraining additives that form spherical pores with diameters considerably less than 0.5 mm [63]. To prevent frost damage, these pores must be closely spaced and uniformly distributed in the cement paste matrix of the concrete. The necessary volume of entrained air depends on the mix proportion of the concrete and is normally 2 – 6 % by volume of the concrete [77], [85]. Foam-forming additives are used to make concrete with fine aggregates and a cellular structure. Such special lightweight concretes are primarily used for thermal insulation.

Air-removing additives are also available, and are often added to plasticizers.

2.2.4.4. Other Additives

Other additives are used for special applications, for example, expansion-producing agents; corrosion inhibitors; fungicidal, insecticidal, or germicidal agents; damp-proofing and permeability-reducing additives; bonding agents; or pigments. Their total consumption, however, is low [73], [83].

Much experience and careful control of concrete properties is required when additives are used because their effects may depend on many parameters, such as concrete composition, type of cement, and temperature. The simultaneous use of more than one additive may lead to problems because of interferences and to undesired concrete properties. In such situations a trial batch is mandatory [77].

2.2.4.5. Fine-grained Material and Polymers

Fine-Grained Material. The main groups of fine-grained mineral solids used as additives are powdered unreactive rock, such as limestone or quartz or finely divided pozzolanas, which are either natural or industrial byproducts. Unreactive rock material is an inert filler that can improve the workability of the fresh concrete, especially when the available natural aggregates lack fines. Pozzolanas may increase the strength and reduce the permeability of the hardened concrete because of chemical reactions with hydration products of the cement. These reactions consume the calcium hydroxide formed during the hydration of portland cement. Accordingly, excessive amounts of pozzolanas may lower the alkaline reserve of the concrete.

Reactive Minerals. Natural pozzolanas include volcanic ashes, opaline shales, and cherts that contain reactive compounds of silicon or aluminum. Specifications for natural pozzolanas are given in [86], [87]. Among the synthetic pozzolanas, pulverized fuel ash (fly ash) is the most important additive and is used in large quantities. Its requirements as a concrete additive are specified in standards [86], [215–217]. Aside from improving certain concrete properties, the use of fly ash can reduce the amount of energy-expensive cement and save natural resources by recycling waste material [88].

Silica fume – a pozzolanic waste material in silicon and ferrosilicon industries – has opened new areas of application for advanced cement-based materials. Owing to its small particle size (ca. 0.1 μm) silica fume acts as a microfiller in the cement paste, thus reducing the total porosity considerably. Compressive strengths exceeding 250 MPa have been reported for concretes containing silica fume. For special applications, the new material can be substituted for cast iron, other metals, ceramics, or polymers [89]. Requirements for the properties of silica fume are given in [218].

Metal Oxide Coloring Agents. Coloring additives are generally alkali-resistant oxides of iron, titanium, or chromium. A uniform coloring effect is best achieved when such additives are used in combination with white cement.

Polymers. A considerable increase in concrete strength and durability under severe conditions is achieved by the addition of certain polymers to concrete (polymer cement concrete, PCC). However, the properties of polymer-modified concrete may deteriorate on continuous exposure to high humidity [90].

2.3. Production

2.3.1. Definition of Concrete Properties

Advances in concrete technology have established the basic rules for making concrete with desired properties in both the fresh and hardened states. Although the desired properties of the fresh concrete must be deduced from the working conditions and equipment at a particular construction site, the requirements for hardened concrete are determined by the expected service conditions. Strength, physical properties, and durability must be considered. A careful analysis of all required concrete properties is necessary before a selection of materials and their proportions is made.

In many instances, the need for concrete durability may lead to more stringent requirements for materials selection and proportions than those for the desired strength. Therefore, in national and international standards and guidelines or recommendations on the production and use of concrete and reinforced or prestressed concrete, the expected service conditions are sumarized in exposure classes and requirements on the concrete composition are set, e.g., on maximum water-to-cement (w/c) ratios, type of cement, and minimum cement content or the use of particular aggregates or additives [77], [219–221]. Expected chemical attack of the hardened concrete, e.g., by sulfates, may require the selection of sulfate-resistant cements. Anticipated exposure to severe weather and deicing salts requires the preparation of air-entrained concrete. Alkali-sensitive aggregates should be used only in combination with low-alkali cements [77], [91]. In all cases where external attack is expected, a concrete with a dense pore structure and low permeability should be used. In reinforced concrete, corrosion protection of the steel reinforcement must be assured by a low w/c ratio. Therefore, for moderate service conditions, the w/c ratio is limited to a maximum of 0.65, whereas severe conditions may require w/c ratios not above 0.4 [77], [91], [92], [219], [221]. Other uses may require similar restrictions for the w/c ratio, type of cement, or cement content, as well as the selection of suitable aggregates and beneficial additives.

2.3.2. Proportioning of Materials

The selection of the concrete-making materials and their proportioning may be specified by the purchaser of the concrete or the owner of the structure to be built, or it is in the responsibility of the concrete producer or contractor [219]. The basic concrete materials and additives must be proportioned in such a manner that all desired properties of the concrete are achieved.

There is no general approach, but rather empirical methods for determining the proper mix proportions, based on either volume or density. The volume or mass of the individual concrete constituents is determined in consecutive steps, which in general lead to a first trial mix for evaluation.

Only for low-grade concrete with a characteristic strength $f_{ck} < 25$ MPa are suitable mix proportions presented in the tables given in [77]. High-grade concrete or concrete containing additives require a thoroughly detailed mix design for the first trial batch. The mix design may be changed after the properties of the trial mix have been tested. Guidelines and methods for proportioning of concrete mixes can be found in [62], [63], [73], [92].

One of several possible approaches to obtain a first estimate of the mix proportions is presented below. The concrete quality is specified by its characteristic compressive strength, which normally is determined after 28 days on cubes or cylinders [93–96].

Because of variations in the strength of concrete, the mean strength f_{cm} of a set of test samples must exceed the required characteristic strength f_{ck} by a certain amount, which can be derived by statistical methods [97], [98], [219], [221].

For normal aggregates and a given type of cement, the strength of completely compacted concrete is a function of the w/c ratio. Approximate values of this ratio for achieving different strength levels are given in Table 13 for non-air-entrained and air-entrained concretes [92]; these values can be obtained from diagrams for different types of cement [99], such as Figure 29 for non-air-entrained concrete.

Table 13. Relationship between w/c ratio and compressive strength of concrete cylinders [92]

Compressive	Water : cement ratio	
strength, MPa	Non-air-entrained	Air-entrained
45	0.38	
40	0.43	
35	0.48	0.40
30	0.55	0.46
25	0.62	0.53
20	0.70	0.61
15	0.80	0.71

The content of mixing water required to achieve a certain consistency of the fresh concrete depends mainly on the maximum size and the grading of the aggregates. As a general rule, more sand or fine particles and smaller maximum aggregate size require more mixing water. For angular coarse aggregates, Table 14 gives maximum quantities of mixing water as a function of aggregate size and consistency as de-

Table 14. Approximate water content (in kg/m^3) of fresh concrete [92]

Slump,	Maximum diameter of aggregate, mm				
mm	10	12.5	20	25	40
30 – 50	205	200	185	180	160
80 – 100	225	215	200	195	175
150 – 180	240	230	210	205	185

scribed by the slump (see Section 2.4.1) of the fresh concrete.

The amount of water required for a concrete of given consistency may be estimated from the parameter k, the water requirement factor, which is defined as the sum of the percentages of aggregate retained on 9 standard sieves (0.25, 0.5, 1.0, 2.0, 4.0, 8.0, 16.0, 31.5, and 63 mm mesh), divided by 100:

$$k = 1/100 \sum R_{\mathrm{i}}$$

where R_{i} is the percentage of material retained on each of nine sieves, each of which is charged with the complete range of aggregate sizes. In Figure 30 the water content of concrete with rounded aggregates is shown as a function of k for various ranges of consistency as described by the compaction factor v (Section 2.4.1).

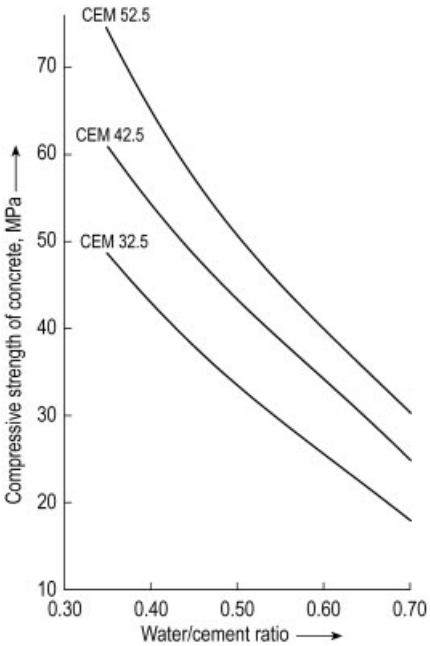

Figure 29. Characteristic compressive strength f_{ck} of concrete cylinders for cements with different standard compressive strengths as a function of the w/c ratio

The amount of water added to the mix may need to be corrected for the amount of water adsorbed on the aggregate surfaces, particularly for the sand fraction. After the water content is estimated, the cement content of the mix is determined from the w/c ratio. The amount of aggregates can be calculated by considering the absolute volume of the various concrete constituents contained in 1 m^3 of compacted fresh concrete.

In this example, no restrictions on the w/c ratio or cement content are considered, which may be necessary to achieve a sufficient durability or workability. Furthermore, the results obtained by any method of mix design will give only a first estimate of the concrete composition. Necessary corrections to the individual parameters are determined by trial batches. Trial batches are mandatory whenever additives are used because the effects of the additives cannot be reliably predicted and may depend on the temperature, intensity of mixing, and composition of the concrete.

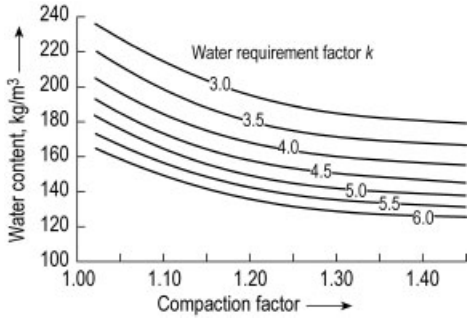

Figure 30. Water content of fresh concrete as a function of the compaction factor for different water requirement factors [100]

2.3.3. Production Process

For the production of uniform concrete of high quality, adequate storage facilities for the individual components must be provided. The ce-

ment should be protected from moisture to prevent partial prehydration. Fine and coarse aggregates stored in separate fractions must be handled carefully to avoid segregation. Sand, if not stored in protected bins, should be drained to avoid excessive water adsorption. Liquid additives must be protected from freezing and thoroughly mixed before use.

All solid materials are measured by mass; liquids can be measured by volume. All measuring devices should be calibrated regularly to operate accurately within 1 – 3 % [77], [101]. On large construction sites or in ready-mixed concrete plants, measurement of the components may be automated at the bin discharge. They are carried to the mixing station by conveyor belts. Automatic sensing of the surface moisture of the aggregates allows for the correction of the amount of water to be added to the mix. Solid additives are added to the mix with the cement, whereas liquid additives are preferably mixed with the water. Some types of superplasticizers are effective for only a limited period of time. Therefore, they are added to the fresh concrete in a second mixing step immediately before placing the concrete.

To achieve a thoroughly mixed and homogeneous fresh concrete, the materials are charged into mixers which generally consist of revolving drums equipped with blades, stationary or revolving pans in which blades rotate on vertical shafts, or horizontal drums in which spiral blades operate on a rotating horizontal shaft. The necessary mixing time depends on the intensity of mixing, as well as on the batch size, and normally ranges from 30 to 120 s. Truck-mixed concrete may require up to 100 revolutions of the truck-mounted drum rotating at mixing speed [63], [77], [100], [101].

2.4. Fresh Concrete

2.4.1. Workability

The term workability describes several characteristics of fresh concrete which are of particular importance for its placement, consolidation, and finish. These characteristics include flowability, resistance to plastic deformation, compactibility, cohesiveness, and tendency to bleed or segregate. There is no single test method for evaluating all aspects of the workability of fresh concrete. Accordingly, test methods commonly determine either the flowability, expressed in terms of consistency, or the compactibility, which may be described by different compaction factors.

A widely used method for evaluating the consistency of fresh concrete is the slump test, which is specified (with minor differences) in various standards [102–104]. In principle, in this test a frustum of a cone which serves as a mold is filled with concrete in a standard manner. After the mold is removed, the subsidence of the concrete cone is measured to determine the slump (Fig. 31).

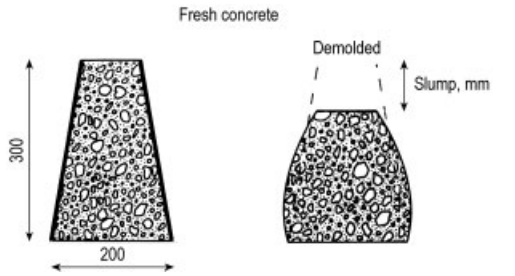

Figure 31. Test setup for the determination of the slump of fresh concrete [103]

In [93], the spread of a cone of fresh concrete is determined on a flow table, which is dropped in a standard manner (15 times within 15 seconds one side is lifted 40 mm and dropped) to agitate the fresh concrete. In the Vebe test, a cone of fresh concrete is remolded to a cylindrical shape by standard vibration. The time required to achieve complete remolding describes the workability of the fresh concrete in terms of Vebe seconds [105], [106].

Tests for the compactibility of fresh concrete are specified in [93], [107]. In general, a mold is filled with concrete in such a manner that any uncontrolled compaction is avoided. After the mold is filled, the fresh concrete is consolidated by vibration. Compaction factors are then derived from density measurements or a comparison of the original volume (height) of the fresh concrete to the volume (height) after compaction. The test setups are shown schematically in Figure 32.

There is no strict correlation of the results obtained from the different methods for evaluating the workability of fresh concrete, and no single test method is universally applicable. Although

plastic concrete mixes can be characterized by the slump or spread, the compacting tests are more appropriate for stiff and dry mixes. The different methods of testing workability are correlated in Figure 33 [100].

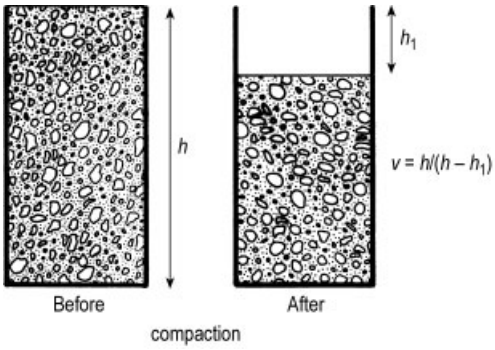

Figure 32. Test setup for the determination of the compaction factor v of fresh concrete [93]

2.4.2. Placement, Consolidation, and Finish

After mixing, the concrete must be placed in its final position, consolidated, and eventually surface-finished within 1 – 2 h and before the initial set. The method of placing concrete into forms depends on its consistency and the equipment available at the site. Separate placement of single batches by buckets or buggies, as well as continuous placement by concrete pumps and hoses or conveyor belts, is successful if care is taken to avoid separation of the concrete constituents. To prevent cold joints, subsequent layers must be placed while the first layer is still plastic and while knitting of the two layers can be achieved by vibration.

Consolidation of the fresh concrete after casting removes entrapped air, allowing for a monolithic and densely packed concrete structure. The reinforcement must be embedded thoroughly to provide a good bond with the concrete and to be protected from corrosion.

Fresh concrete is consolidated by vibration, which is usually provided by vibrators inserted into the concrete. Vibrators mounted on vertical formworks of thin concrete cross-sections or surface vibrators are also used. The vibration energy necessary for complete compaction decreases as the workability of the fresh concrete increases. Excess vibration can cause segregation [100], [108].

Plasticizing additives may substantially reduce the consolidation work, and concrete containing superplasticizers and an appropriate fraction of mineral fines may not require any consolidation. Self compacting concrete may increase the quality of the concrete placement and, at the same time facilitates the work and increases the efficiency on the site [222], [223]. Concrete may benefit from a second vibration after the initial setting to repair cracks that have developed due to settlement, early shrinkage, or deformations of the formwork.

The top surface of fresh concrete can be troweled or screeded. The durability of horizontal concrete surfaces, such as pavements and parking garage decks, is increased by vacuum processing of the fresh concrete surface. Partial withdrawal of the mixing water results in lower porosity and higher strength of the surface region [63], [100], [108].

2.4.3. Curing

The development of concrete strength and impermeability requires favorable conditions for hydration of the cement. In addition, crack formation should be avoided. Curing of concrete, therefore, comprises all measures undertaken to protect the hardening concrete from drying and to control the temperature across a section [109].

Early loss of water from the young concrete may prevent further hydration of the cement, especially in the surface region, resulting in insufficient strength development and high permeability. Furthermore, premature drying causes extensive shrinkage, which will lead to random cracks in the surface layer. Protection against moisture loss is normally achieved by covering the concrete surface with water-absorbent materials such as burlap or straw, which are kept wet under damp-proof plastic sheets, or by the application of membrane-forming curing compounds. The last-named compounds contain wax or resins and are sprayed or brushed on the concrete surface; however, they do not provide sufficient protection for hot weather concrete curing [110], [111].

Control of concrete temperature may be necessary when concrete is placed at low ambient temperature. Because the hydration of cement

proceeds only slowly or even ceases at temperatures < 5 °C, thermal insulation of the concrete surface is used to prevent rapid loss of the heat of hydration. Thermal insulation also prevents excessive temperature gradients in a concrete cross section and protects the young concrete from freezing [112]. Thermal insulation is particularly important for massive concrete sections.

The required curing period depends primarily on the ambient climatic conditions and on the hydration kinetics of the cement used; it may vary from two days for rapidly hardening cement under moderate climatic conditions up to 14 days for massive concrete structures. Guidelines can be found in [109], [111], [219].

Interior sections of a concrete member are less affected by curing than the surface regions. Adequate curing is, therefore, essential for achieving the required durability of a concrete structure.

The accelerating effect of elevated temperature on the hardening of concrete is used in the prefabrication of concrete elements; however, curing temperatures exceeding ca. 60 °C may reduce the final strength of the concrete. Steam curing at atmospheric or high pressure is applied in the preparation of certain concrete products such as pavement units, masonry units, and precast elements [113], [114].

2.4.4. Special Methods

A variety of special methods have been developed to simplify or increase the efficiency of preparing, transporting, placing, consolidating, and curing concrete.

For repair work, for strengthening existing structures, or for special applications, shotcrete may be superior to conventional concrete placing practice. In shotcrete, concrete is transferred pneumatically in a hose and projected through a nozzle onto a rigid surface at high velocity; consolidation is readily achieved by impact [115–119]. Dry and lean concrete mixes for road subbases or bearing courses as well as concrete in dams are placed as roller-compacted concrete (RCC) [224].

In construction of large concrete elements such as dams or for concrete placed under water, grout may be injected into the voids of preplaced coarse aggregates to produce concrete [101]. Similar methods are applied to pre-packed steel fibers to produce fiber-reinforced concrete with high fiber concentrations. [225].

Special procedures also exist for the production of concrete products such as masonry units, concrete pavements, or concrete tubes [113], [114].

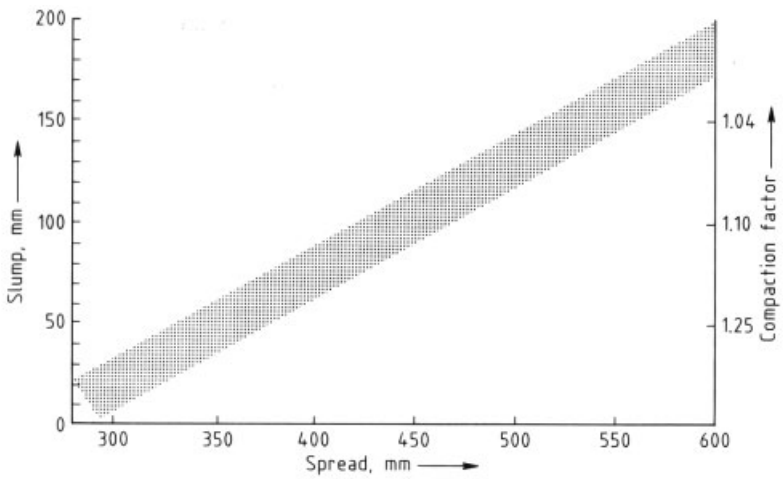

Figure 33. Approximate correlation of the slump, spread, and compaction factor of fresh concrete [100]

2.5. Mechanical Properties

The most important design characteristics of concrete needed for structural analysis are the compressive strength and the modulus of elasticity of the hardened concrete; in many instances the tensile strength is not a design parameter. For concrete sections subjected to tensile stresses, a fracture mechanics approach is recommended because the tensile strength of concrete is controlled by such material defects as voids, pores, or microcracks. Structural design also requires information on dimensional changes due to creep, shrinkage as a result of moisture migration, or temperature variation.

2.5.1. Crack Development

The mechanical behavior of concrete is controlled by defects or microcracks, which are caused by the nature of the hydrated cement paste or which are generated during the hardening of concrete. The cement hydrates continuously, thereby dissicating the matrix and causing it to shrink. Additional shrinkage is due to evaporation of water from the concrete surface. Because the stiff aggregates restrain the shrinkage of the cement paste, microscopic cracks are formed in the interfaces between mortar and coarse aggregates. Crack propagation continues from these defects when the concrete is subjected to stress. Tensile stresses and compressive stresses cause different types of crack propagation [120].

When concrete is loaded in compression, microcracks start to propagate when the stress exceeds ca. 40 % of the maximum stress. With an increasing stress level, crack growth proceeds mainly along the interfaces between coarse aggregates and the matrix. At ca. 80 % of the maximum stress, the cracks start to extend into the mortar matrix. Frequently cracks are arrested at aggregates, which results in branch cracking. As a consequence, the nonlinearity of the stress – strain diagram increases substantially. The predominant orientation of the cracks is then parallel to the axis of loading. At high stress, microcracks join to form longer cracks. Thus, the cracks reach a critical length, and unstable crack growth can only be avoided by reduction of the stress. As long as crack propagation remains stable, the strains continue to increase and the stress – strain relationship shows a descending part (stress σ = force/area, strain ε = change in length/original length).

If the load is kept constant at a strain greater than the strain at maximum stress, a continuous fracture plane is formed, causing the complete collapse of the concrete [121].

Concrete loaded in tension behaves like a brittle material. Initiation of crack growth takes place at preexisting microcracks, but multiple crack growth and branch cracking are much less pronounced than when under compression. Crack propagation is usually limited to one section. Nevertheless, a *process zone* is formed in front of a propagating crack, thus increasing the energy dissipation. At tensile failure stress, unstable crack growth forms a fracture surface perpendicular to the axis of applied tensile stress. Sophisticated testing procedures can detect some stable crack propagation, which is more pronounced for larger amounts and sizes of aggregates [122].

2.5.2. Strength

2.5.2.1. Compressive Strength

The compressive strength of normal concrete is in the range 10 – 100 MPa (N/mm^2), depending on the composition and preparation procedures. High-strength concrete (HSC) can be produced with selected types and grades of aggregates, mineral fines, and superplastizers to allow very low w/c ratios, and compressive strengths exceeding 250 MPa have been reported [226], [227]. The compressive strength of most concrete placed, however, is in the range 25 – 50 MPa (N/mm^2). The most important parameter controlling concrete strength is the w/c ratio [123]. Other factors are the type of cement, curing conditions, moisture content, degree of hydration of the cement, and the grading and shape of aggregates. If sufficient water is provided, the hydration of cement will continue over a period of several years, and concrete strength may increase continuously during this time [62], [63]. It is customary to classify concrete according to the compressive strength attained after a standardized curing of 28 d. In certain cases, however, the compressive strength after shorter or longer curing periods may be specified.

Because of inevitable differences in the properties of concrete components, preparation procedures, handling, consolidation, and curing, the strength of any particular type of concrete varies within a range. For the classification of concrete strength, therefore, a characteristic strength f_{ck} is defined, below which only a certain percentage, usually 5 %, of the entire population of data falls. This value of f_{ck} defines a 5 % defective concrete [93], [98]. Assuming a normal frequency distribution of the individual strength values, the mean strength f_{cm} required to reach a certain 5 % defective for a given standard deviation s can be estimated.

$$f_{cm} = f_{ck} + 1.64s$$

When good quality control is maintained, the standard deviation s is ca. 5 MPa. If a sufficient number of individual strength data are available, the standard deviation can be computed from the observed frequency distribution of individual strength results [93], [97], [124].

In general, the strength of concrete is determined by testing separately cast companion specimens, which are either cured under standard laboratory conditions or are exposed to the same curing conditions as the ultimate structure. Because the test results depend on the size and slenderness (i.e., height/thickness) of the concrete specimens tested, standard test specimens must be used. The most common specimens are cubes with a side length of 150 or 200 mm [93], [95] or cylinders with a diameter of 150 mm and 300 mm high [94], [95]. Specimen preparation and testing procedures are specified in the corresponding standards [93], [125].

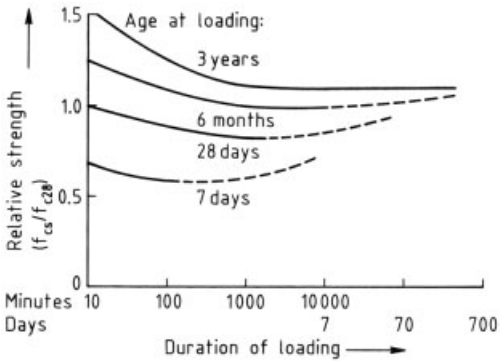

Figure 34. Relative sustained load strength of concrete f_{cs}/f_{c28} for different durations of loading and ages at the time of loading [129]

The use of drilled cores for destructive laboratory testing is the most reliable method for testing the strength of an existing concrete structure. Nondestructive methods such as ultrasonic pulse velocity, rebound, or penetration resistance require calibration [93], [126–128].

The sustained load strength of concrete is controlled by two counteracting mechanisms. While a continuous increase of strain under a sustained load may weaken the integrity of the concrete, a strengthening effect is exerted by the continuing hydration of the cement, which is more pronounced for younger concrete than for older concrete. The sustained load strength of concrete loaded at an age of 28 d is ca. 80 % of its short-time strength, which is less influenced by creep (Fig. 34) [129], [221].

The fatigue strength of concrete decreases with an increasing number of stress cycles and an increasing stress range. The relations between the maximum stress and the average number of cycles leading to failure for various ratios of minimum to maximum stress are given in Figure 35. Concrete does not show a fatigue limit up to 10^7 load repetitions [130]. Analytical procedures for the prediction of the fatigue strength of concrete are presented in [221].

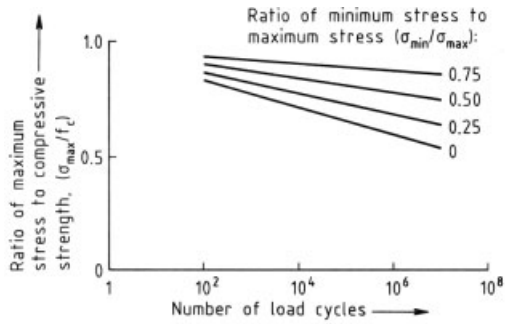

Figure 35. Fatigue strength of concrete for different stress levels

2.5.2.2. Tensile Strength

The tensile strength of normal concrete is approximately one-tenth the compressive strength. Detailed analyses by CEB [221] predict that the tensile strength should not increase in proportion to the compressive strength. In general, the tensile strength of concrete is determined by testing beams subjected to flexure or cylinders

that are subjected to diametrically opposed axial line loads, giving the so-called tensile splitting strength [93], [132–136]. The direct tensile strength can be determined by axial tension according to [137]. Different test methods yield different results because of differences in stress state, stress distribution, and probability of failure.

2.5.2.3. Fracture Mechanics

The application of linear elastic fracture mechanics to concrete is limited because of the heterogeneity of the material and because of the development of a process zone, i.e., a region of microcracks in front of a major crack.

Accordingly, specimens with minimum dimensions at least 20 times the maximum aggregate size should be used when the conventional fracture mechanics characteristics of a concrete are determined.

The fracture toughness K_{IC} of cement paste, mortar, and concrete are as follows:

hydrated cement paste $K_{IC} = 5 - 15\ \mathrm{N/mm^{3/2}}$
mortar $K_{IC} = 10 - 20\ \mathrm{N/mm^{3/2}}$
concrete $K_{IC} = 15 - 50\ \mathrm{N/mm^{3/2}}$

K_{IC} increases with increasing degree of hydration and decreasing w/c ratio. Because aggregates act as crack arresters and cause multiple crack growth, K_{IC} increases with increasing content and size of aggregates.

Fracture mechanics concepts can be applied successfully in concrete technology and design, e.g., for the description of fracture processes, for estimating notch sensitivity, and for the design of large unreinforced concrete members [138].

A further parameter is the fracture energy G_F, which corresponds to the total energy required to separate a concrete section loaded in tension. Since G_F is less dependent on size than K_{IC}, it can be used to analyze smaller concrete members [139]. According to [221] the fracture energy G_F of a given concrete can be estimated from its compressive strength and the maximum aggregate size:

$$G_F = G_{F0}\,(f_{cm}/f_{cm0})^{0.7}$$

where f_{cm} is the the mean cylinder compressive strength in $\mathrm{N/mm^2}$ and $f_{cm0} = \mathrm{const.} = 10\ \mathrm{N/mm^2}$. G_F increases from 0.25 to 0.38 as the maximum aggregate size increases from 8 to 32 mm.

2.5.3. Deformation Characteristics

The deformations of concrete under load can be separated into elastic and time-dependent deformations. In structural analysis, time-dependent strains are of considerable importance in estimating the serviceability of constrained members or prestressed structures in which a loss of prestress may occur. All time-dependent deformations are markedly influenced by the concrete composition and by the prevailing climatic conditions. Thermal expansion is discussed in Section 2.6.6.

2.5.3.1. Relationship of Stress to Strain

Stress – strain diagrams of various concretes with different strengths are presented for compression and tension in Figure 36. The diagrams were recorded at a constant strain rate. Under compression, these diagrams are nonlinear even at low stress levels. With increasing stress, the nonlinearity increases. The strain at maximum compressive stress is almost independent of the strength of the concrete and amounts to ca. 2×10^{-3}. For higher strength concretes the increasing brittleness of the material is reflected in a steep descending branch of the stress – strain curve. Analytical models for the description of stress strain relationships as functions of concrete strength are given in [221] The stress – strain relationships are also influenced by the strain rate, as shown in Figure 37. A reduction in strain rate causes a slight reduction in strength and a pronounced increase in the ductility of the concrete.

The stress – strain relationship for concrete loaded in tension is almost linear up to failure, which is normally brittle. Failure strains range from ca. 0.1×10^{-3} for a concrete loaded in axial tension to 0.3×10^{-3} for concrete subjected to flexure [63], [120].

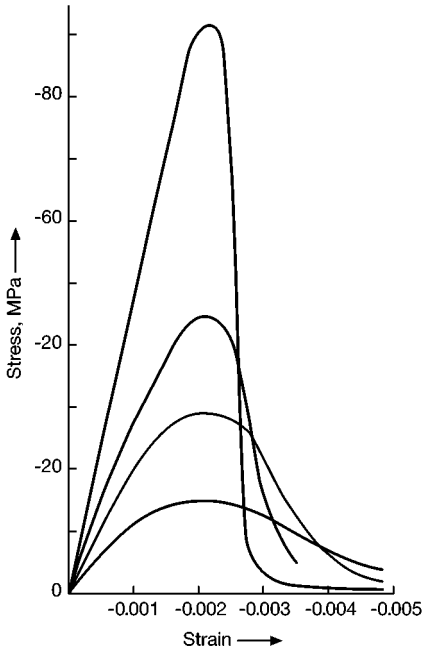

Figure 36. Stress – strain relationship for concretes of different strengths

2.5.3.2. Modulus of Elasticity and Poisson's Ratio

The curves in Figure 36 show that there is no constant modulus of elasticity for concrete in any part of the stress – strain curve. It is, therefore, customary to define an initial tangent modulus given by the slope of the stress – strain relationship at the origin, a tangent modulus at a given stress, or a secant modulus over a certain range of the diagram at ca. $0 < \sigma \leq 0.5 f_c$. The lower and upper limits of this range vary in different standards and recommendations [93], [140], [141]. The modulus of elasticity E_c for normal concrete is in the range 20 000 – 40 000 MPa. A higher compressive strength of the concrete leads to a higher modulus of elasticity. This correlation [221] is expressed empirically as follows:

$$E_c = \alpha_E E_{c0} \left(f_{cm}/f_{cm0}\right)^{1/3}$$

where E_c = modulus of elasticity in N/mm^2, α_E = factor to consider the stiffness of the aggregates; for natural aggragates $0.7 < \alpha_E < 1.2$, f_{cm} = mean cylinder compressive strength in N/mm^2, and f_{cm0} = const. = 10 N/mm^2. The modulus of elasticity can also be deduced from the natural frequency of concrete specimens subjected to longitudinal, transverse, or torsional vibration as the dynamic modulus of elasticity [63]. These nondestructive tests yield higher values of E_c.

Poisson's ratio ν of concrete is about 0.2 as long as the stress level σ does not exceed $0.8 f_c$. At higher stress levels, values of Poisson's ratio exceeding 0.5 are observed due to multiple crack formation [120].

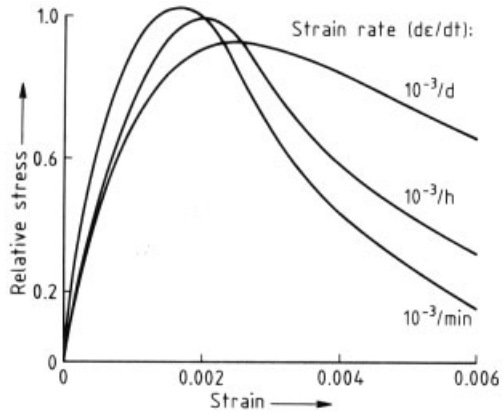

Figure 37. Influence of strain rate on the stress – strain relationship of concrete

2.5.3.3. Creep

Creep is defined as the time-dependent increase of load-dependent strain under constant stress. It is often expressed in terms of the creep coefficient φ_c, which is the ratio of creep strain ε_{cc} to initial (elastic) strain ε_{ce}. For the range of working loads, the creep coefficient of concrete ranges from ca. 1 to 4. Creep is partially reversible. The part of creep that can be recovered on unloading is referred to as *delayed elastic strain,* whereas the irreversible part is generally described as *flow.*

The water content of concrete exerts an important influence on creep deformation. *Basic creep* is defined as the creep of concrete in a state of moisture equilibrium. Basic creep becomes smaller as the water content of the concrete prior to load application decreases. *Drying creep* is defined as the difference between basic creep and the creep of concrete that is allowed

to dry under load. Drying creep is thus proportional to moisture loss. Creep deformation increases because of microcracking, particularly at higher stress levels [142].

Creep of concrete increases with an increase in the amount of cement paste, of the water content of the concrete prior to loading, and of the w/c ratio. Creep increases at low relative humidity and with decreasing member size. A higher degree of hydration at the time of loading reduces creep.

There are several mathematical methods for estimating the time development of creep. In these formulas, which may incorporate viscoelastic models, terms or functions represent characteristics of the concrete, geometry of the member, and nature of the environment [142], [143], [221], [228], [219].

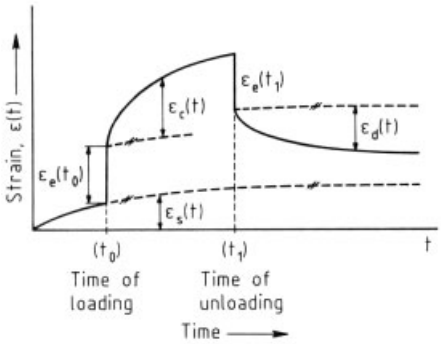

Figure 38. Total deformation of concrete loaded at t_0 and unloaded at t_1 as a function of time.
ε_e is elastic strain, ε_s is shrinkage strain, ε_c is total creep strain, ε_d is delayed elastic strain

2.5.3.4. Shrinkage and Swelling

Shrinkage and swelling are load- and temperature-independent volume changes essentially caused by changes in the moisture content of the concrete. As a first approximation, shrinkage is proportional to the moisture loss. When the concrete is rewetted, swelling occurs, although some irreversible shrinkage may remain.

The total shrinkage of hydrated cement paste may amount to several percent. Aggregates restrain shrinkage so that shrinkage of concrete is considerably less and is ca. 0.1×10^{-3} to 1×10^{-3}. Rapid drying of concrete generates shrinkage-induced tensile stresses in the surface zones that may lead to cracks [142], [143]. Analytical models for the prediction of shrinkage as a function of concrete composition, geometry and exposure of the concrete member are given for example in [220], [221], [228].

The total deformation of concrete is shown as a function of time in Figure 38.

2.6. Physical Properties

Most of the physical properties of concrete are controlled by its composition, mainly by the type and amount of aggregates used and the water content of the fresh or hardened concrete.

2.6.1. Density

Concrete is classified according to its density:

lightweight concrete $\varrho \leq 1850\,\mathrm{kg/m^3}$ [144]
$\varrho \leq 2000\,\mathrm{kg/m^3}$ [145]
normal-weight concrete $\varrho \leq 2800\,\mathrm{kg/m^3}$
heavyweight concrete $\varrho > 2800\,\mathrm{kg/m^3}$
[77], [92]

Structural lightweight concrete can be reinforced or prestressed like normal concrete. Heavyweight concrete is used mainly as a radiation-shielding material.

2.6.2. Porosity

The total porosity of concrete comprises gel pores and capillary pores in the hydrated cement paste, the porosity of aggregates and aggregate interfaces, and voids due to incomplete consolidation. Additionally, a protective pore system may be introduced into the matrix by air-entraining additives or air-entraining cements. The size of pores ranges from small gel pores with diameters of ca. 10^{-6} mm up to several millimeters for the voids [146]. In normal concrete with dense aggregates, the total porosity is 8 – 15 % by volume. Air-entrained concrete contains an additional 2 – 6 % pores. Lightweight concrete with porous aggregates, lightweight concrete with a honeycomb structure, or cellular lightweight concrete may have a total porosity of 60 – 90 % by volume. Such concretes have low strength and are used only for thermal insulation [63].

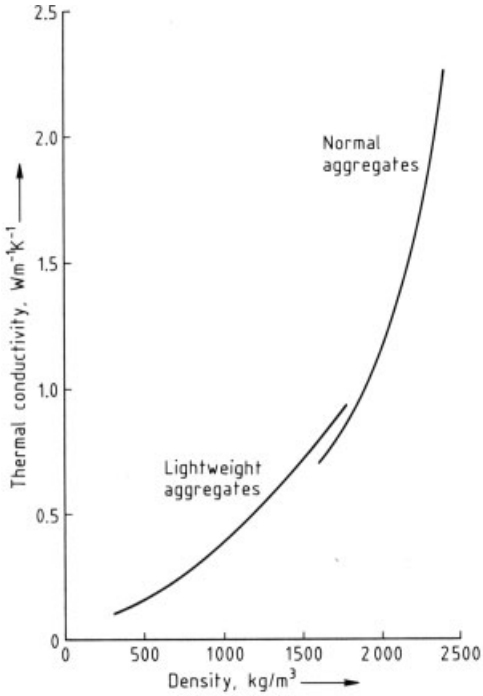

Figure 39. Thermal conductivity of concrete as a function of concrete density

2.6.3. Thermal Conductivity

The thermal conductivity of concrete increases with decreasing porosity and increasing moisture content. Crystalline aggregates give higher conductivity than amorphous aggregates. For normal concrete the thermal conductivity is $1.5-3.7\,\mathrm{W\,m^{-1}\,K^{-1}}$. The thermal conductivity is shown as a function of the density of dry concrete in Figure 39 [63].

2.6.4. Electrical Conductivity

Dissolved ions present in the water in the pores of hardened concrete are primarily responsible for the electrical conductivity. For dry concrete, specific ohmic resistance is $10^{11}\,\Omega\cdot\mathrm{cm}$, whereas for concrete saturated with water it is ca. $10^{4}\,\Omega\cdot\mathrm{cm}$. Concrete contaminated with soluble salts has even lower resistance. Because the water in the pores is an electrolyte, conductivity increases with increasing w/c ratio [63].

2.6.5. Permeability

An important parameter for concrete durability is its permeability, which describes the ability of gases or liquids to penetrate into and through concrete sections. The movement of gases or liquids in concrete primarily occurs in the capillary pores of the cement paste matrix. Because the volume of capillary pores increases with an increasing w/c ratio, permeability increases correspondingly. High values of permeability can be expected when the capillaries form an interconnected continuous network caused by a low degree of hydration. For w/c ratios exceeding ca. 0.7, continuity of capillaries must be assumed even after complete hydration [63].

The saturated capillary flow of water under the influence of hydraulic pressure is usually described by an empirical relation according to Darcy [229]. Neglecting the viscosity and the density of water, a specific water permeability k_{w} is derived, values of which ranging from $k_{\mathrm{w}} = 10^{-10}$ to 10^{-14} m/s have been reported for normal concretes [147], [230].

For the flow of gases, the viscosity and compressibility of the fluid must be considered. For normal concrete mixes the coefficient of gas permeability K_{g} is in the range of 10^{-14} to $10^{-18}\,\mathrm{m^2/s}$. Since the flow of gases depends on empty capillary pores, the coefficient of gas permeability is strongly influenced by the moisture content of the concrete [231]. The gas permeability of concrete is closely related to the ingress of deleterious substances into concrete and thus may serve as an indication of its durability [230].

For barriers or retaining structures in the chemical and petrochemical industries, the penetration of hazardous fluids by capillary action must be controlled to protect the soil, ground water, and other environmental compartments from contamination. Recommendations and requirements are given in [232], [233]

Moisture movement is important in the drying of concrete. The diffusion of water molecules in the vapor phase or in condensed films depends strongly on the local moisture concentration and can be 10^{-8} to $10^{-11}\,\mathrm{m^2/s}$ for concretes of normal compositions at relative humidity of 0 – 100 % [234].

2.6.6. Thermal Expansion

The thermal expansion of concrete is controlled by the thermal expansion of the cement paste and the aggregates. Therefore, different aggregates lead to different coefficients of thermal expansion of the concrete. Restraint of the matrix by coarse aggregates or the formation of microcracks, however, causes a total deformation of concrete that is not equal to the sum of the individual deformations of matrix and aggregates weighted by their volume concentrations. The coefficient of thermal expansion for concrete made with quartzitic aggregates is ca. $12 \times 10^{-6}\,\mathrm{K}^{-1}$. For concrete made with calcareous aggregates, it is ca. $6 \times 10^{-6}\,\mathrm{K}^{-1}$ [63], [148].

Changing temperatures always cause a moisture movement in the matrix, which may lead to shrinkage or swelling. These deformations are added to the thermal dimensional changes.

2.6.7. Shielding Properties

Concrete is an economical material for shielding against high-energy X-rays, gamma rays, or neutrons because it combines structural properties with good shielding characteristics. Because of its heterogeneity, concrete contains a variety of different elements for attenuation or absorption of the different types of radiation. Although these elements exist in normal concrete, special designs can improve the shielding properties.

Neutrons are classified according to their energy as fast, intermediate, and slow (thermal) neutrons. The attenuation of fast and intermediate neutrons by elastic scattering by light elements, such as hydrogen or boron, and inelastic scattering by heavy elements lead to slow neutrons that are finally absorbed by heavy elements. Gamma rays are emitted in the steps of neutron attenuation [63], [149].

Because the attenuation of X-rays and gamma rays is proportional to the density of the material penetrated, heavy elements are required for effective shielding.

Special aggregates for shielding concrete can be natural heavy-element aggregates like baryte, iron ores, or granulated iron. Light elements can be added by using aggregates that contain large amounts of crystal water such as serpentine or by the introduction of insoluble boron compounds [150].

In the structural design of a biological shield, one should keep in mind that the attenuation of radiation generates heat that causes stress. Furthermore, high-energy radiation causes crystal defects in the aggregates that may reduce the strength of the concrete [151].

2.7. Durability

Structural concrete is subjected to a variety of physical and physicochemical influences that result from environmental conditions or other types of attack. In most instances, concrete needs no particular protective measures. In reinforced or prestressed concrete structures, the concrete must prevent corrosion of the reinforcement by providing a high alkalinity to passivate the steel surface.

Concrete aggregates are inert to most types of external attack. However, the hydrated cement paste may interact with the environment because of its interconnected *pore structure*. Water, ionic solutions, or gases may penetrate into the concrete through the capillary pores or microcracks to react chemically with constituents of the cement paste, thus causing deterioration. A durable concrete, therefore, should have a dense structure with a low total porosity and consequently, a low permeability. Low w/c ratios and sufficient curing of the concrete are mandatory. By following the principles in the design of high-strength concrete, which basically involve a very dense cement paste and an improved aggregate/paste interfacial zone, the resulting material also exhibits high resistance against most corrosive agents, because the penetration of aggressive compounds into the concrete is prevented. In many applications of these advanced materials, the enhanced durability is more important than the high strength. In more general terms the material is referred to as high-performance concrete (HPC).

An important factor in durability is the *moisture content* of the concrete. Because most damage mechanisms require a certain amount of water, dry concrete exhibits a high durability. Continuous wetting of concrete may increase the depth of penetration of dissolved compounds, and repeated drying – wetting cycles can accelerate the deterioration [152].

2.7.1. Chemical Attack

Chemical attack on hydrated cement paste involves either partial or complete dissolution of certain hydration products or formation of new compounds, accompanied by an increase in solid volume or in expansive gels.

Water that does not contain excessive amounts of ions or industrial waste pollutants does not dissolve cement paste compounds to a significant degree. However, continuing exposure to distilled water may leach the calcium hydroxide from the paste. Natural waters containing considerable amounts of *carbonic acid* will attack the cement paste and calcareous aggregates, forming soluble calcium hydrogencarbonate. Efflorescence of $CaCO_3$ is often observed.

The compounds of hydrated cement paste are dissolved by *organic and inorganic acids* and the rate of dissolution increases as the pH of the environment decreases. Moderate attack can be seen when the pH of the surrounding solution drops below 6.5 – 6.8.

The formation of expansive phases in interior concrete sections can be detected by the formation of a network of surface cracks, through which white efflorescence or penetration of a gellike material can occur.

Sulfates penetrating into the matrix will react with either calcium hydroxide or calcium aluminates to form gypsum ($CaSO_4 \cdot 2\,H_2O$) or ettringite ($3\,CaO \cdot Al_2O_3 \cdot 3\,CaSO_4 \cdot 32\,H_2O$). The disruptive effect of ettringite formation is more pronounced because it involves a considerable increase in solid volume [153].

Magnesium compounds, either contained in the aggregates or penetrating from outside, can be decomposed to form an expansive gel of magnesium hydroxide.

Alkali – silica reactions occur between the alkaline materials of the cement and amorphous or poorly crystallized silica of some siliceous aggregates. When moisture is available, an expansive gel is formed that causes a continuous deterioration of the concrete. Questionable aggregates can be tested for the presence of reactive silica [80]. Their use as concrete aggregates may require special cements with low effective alkali contents.

The activity of natural waters and soils is evaluated in standards, where recommendations are given for the composition of concrete to be used in contact with them. Concrete that is subject to strong chemical attack may require protective coatings of organic materials on the surface [77], [153–155]. A comprehensive list of chemical compounds and their effect on concrete is given in [154], [221].

2.7.2. Physical Attack

Frost. The mechanism of freezing of water in concrete is rather complex and not entirely clear. Thorough discussions of this subject may be found in [146]. The following description is highly simplified.

Water contained in the larger capillary pores of concrete starts to freeze at temperatures below ca. – 8 °C. If the concrete is critically saturated, i.e., more than ca. 90 % of its pores are water-filled, the expansion of ice may have a disruptive effect if free expansion of ice or migration of the remaining liquid water into free spaces cannot take place. Thus, important parameters for controlling frost damage are the amount of freezable water in the concrete and the rate of ice formation.

The application of deicing salts to a frozen concrete surface may cause further damage. While the ice on the concrete surface will melt, a sudden drop in the temperature of the underlying concrete layer occurs. Furthermore, deicing salts can increase the degree of saturation and the number of freeze – thaw cycles. Additional damaging effects due to osmotic pressure or pressure from salt crystallization have not been clarified [146].

Preparing concrete with high resistance to freezing and deicing agents requires aggregates with a high resistance to freezing and a dense pore structure of the cement paste that is provided by low w/c ratios and sufficient curing. Furthermore, when critical saturation is likely to occur, a protective system of small, spherical, closely spaced pores must be introduced into the mortar matrix by air-entraining additives or air-entraining cements. Because these pores do not fill with water under normal conditions, they provide space for the expansion of ice or water upon freezing.

Elevated Temperature. The resistance of concrete to elevated temperatures depends on the

mineral composition of the aggregates and the water content of the concrete. The mortar matrix in a concrete that is allowed to dry while hot will exhibit large shrinkage deformations, whereas the aggregates expand as the temperature increases. These opposing deformations lead to the formation of cracks, thus reducing concrete strength. Individual compounds of the cement paste start to decompose above 250 °C, and the dehydration continues up to the sintering of particles above 1000 °C [156], [157].

Concrete exposed to temperatures above 100 °C in a moist state undergoes phase transformations in the hydrated cement paste. In the absence of finely divided siliceous material, these phase transformations increase the total porosity of the paste and cause a significant reduction in the concrete strength. If fine siliceous aggregates or siliceous additives are present, new compounds are formed that can compensate for the loss of strength [156].

Radiation. Concrete subjected to high-energy radiation may lose strength. While there is no damaging effect of radiation on the hydrated cement paste, lattice defects can be formed in some types of crystalline rocks used as aggregates. Heterogeneous rock material with crystalline compounds such as granites undergoes a substantial loss in strength [151].

Abrasion and Wear. The surface regions of concrete structures can be subjected to abrasion and wear due to traffic loads, sliding, or scraping of hard objects, or repeated impact. Damage is often encountered on pavements, walls of silos or bunkers, and in pipes and channels for water that carries solids (erosion). Water flowing at a high velocity can cause cavitation. Concrete with a high resistance to wear requires a low w/c ratio and careful curing. Vacuum processing of the fresh concrete and the use of hard aggregates or additives can substantially improve the resistance to wear.

2.7.3. Corrosion of the Reinforcement

The surface zones of reinforced concrete structures can be damaged by corrosion of the embedded reinforcement. The formation of voluminous iron hydroxides generates tensile stresses that cause spalling of the concrete cover. The most common causes of corrosion of the reinforcement are carbonation of the concrete or intrusion of corrosive ions such as chlorides.

Carbon dioxide in the atmosphere penetrates the surface zones of concrete and reacts with the alkaline hydration products of the cement. In this diffusion-controlled reaction, insoluble calcium carbonates are formed and the pH of the pore solution of the matrix drops from more than 12 to ca. 8.7. In this environment, an embedded steel reinforcement is no longer protected by passivation because the passive layers become unstable, and anodic dissolution of the iron occurs when oxygen and moisture are available [158].

The passive layer of steel in concrete can also be destroyed by chloride ions. The major sources of chloride contamination of concrete are deicing agents, seawater, and the thermal decomposition of polymers such as PVC by fire.

Although some chloride can be immobilized chemically or physically by the hydration products of the cement, additional chloride ion greater than a threshold concentration will destroy the passive layers of the steel reinforcement locally, and corrosion occurs [91], [147]. Structures exposed to seawater may, therefore, require an increased concrete cover of the reinforcement [63].

Carbonation and intrusion of active ions are controlled by diffusion. A dense pore structure of the cement paste matrix, especially in the surface zones, reduces the penetration rate and thus improves durability. Under very severe conditions, additional protective measures such as surface coating of the concrete or the steel reinforcement may be required.

2.8. Special Concretes

The basic principles of concrete technology for normal concrete are generally valid for other types of concrete, which may differ from normal concrete with respect to composition or properties tailored for special applications. Preparation, handling, and testing of such concretes are often specified in separate standards or guidelines. The most important types of concrete that differ from normal concrete are lightweight and heavyweight concrete, mass concrete, fiber-reinforced concrete, and polymer concrete.

2.8.1. Lightweight Concrete

Lightweight concrete has a lower bulk density than normal concrete, which can be achieved in several ways. Structural lightweight concrete with a dense structure is prepared with lightweight aggregates, such as tuff, expanded shales, or calcined waste products. Structural lightweight concrete may be reinforced or prestressed like normal concrete.

Lightweight concrete with a honeycomb structure is prepared with only coarse aggregates, which are enveloped by a thin shell of fine mortar, thus providing a large volume of interstitial voids. Such concretes can be used for masonry units or drainage purposes. Cellular lightweight concrete and aerated autoclaved concrete are prepared by introducing foam-forming additives into a mortar mix. Because of their low compressive strength, these cellular lightweight concretes are used only for making masonry units or for thermal insulation.

The basic relations between the w/c ratio and concrete strength are valid for structural lightweight concrete, although with limitations because the strength of the matrix may be higher than the strength of the aggregates. Furthermore, the definition of a w/c ratio is uncertain because the porous aggregates can absorb various amounts of the water added to the mix. Preparing lightweight concrete, therefore, always requires trial mixes. In general, compressive strength increases with increasing bulk density. For densities of 1.0 – 2.0 kg/dm^3, the compressive strength is 10 – 60 MPa. Compared with normal concrete, lightweight concrete has a lower compressive strength, lower modulus of elasticity, lower thermal conductivity, and lower thermal expansion, but higher creep and shrinkage deformations [63].

2.8.2. Heavyweight Concrete

The major applications of heavyweight concrete are as shielding against nuclear radiation, as ballast, and as anchor bodies. Heavyweight aggregates like baryte, ores with densities of 3.6 – 4.6 kg/dm^3, slags, or iron can be used to achieve bulk densities of the concrete exceeding 2.8 kg/dm^3. The mix design for heavyweight concrete follows the principles for normal concrete except that additional care must be taken to ensure the specified concrete density. Segregation must be avoided in handling heavyweight concrete. Therefore, comparatively dry mixes are preferred [92], [100].

2.8.3. Massive Concrete

The term massive concrete refers not to a particular composition but rather to a concrete structure of such size that the heat of hydration of the cement must be considered. In interior parts of large concrete cross-sections, the heat of hydration may generate temperatures $> 80\,^{\circ}$C, causing thermal stresses that can lead to crack formation in the concrete during early hydration. The temperature increase in the concrete member can be kept low by using cements with a low heat of hydration, addition of finely divided pozzolanas, and reduction of the cement content of the mix. The size of the aggregates may be as great as 125 – 150 mm. The curing of massive concrete may require thermal insulation of the surfaces to avoid steep temperature gradients over the cross-section. Otherwise cracking of the concrete is likely to occur [159], [160].

2.8.4. Fiber-Reinforced Concrete

The mortar matrix of concrete can be reinforced by short fibers randomly oriented and homogeneously distributed. The diameter of these fibers may vary from several micrometers to ca. 1 mm and the length from a few millimeters to 50 or 60 mm. High-strength steel fibers are used in steel fiber reinforced concrete placed as shotcrete or poured as normal concrete. High fiber concentrations lead to poor workability, which can be overcome with special placing procedures [225]. Alkali-resistant glass fibers were formerly sometimes used in thin shells or cladding elements. New types of organic, ceramic, and carbon fibers are being tested as substitutes for asbestos.

Fiber-reinforced concrete has good ductility in both compression and tension, and fiber reinforcement is an effective tool for crack width control of concrete [161]. Problems in manufacturing fiber-reinforced concrete are the limited workability of the fresh concrete and the diminished effectiveness of the fibers because of insufficient bonding between fiber and matrix.

2.8.5. Polymer Concrete

In polymer concrete, organic resins form an essential part of the binding material. Polymer concrete has high strength in tension and compression, and exhibits a superior resistance to chemical attack, freezing, and thawing as compared to normal concrete. Disadvantages may result from the high cost of the resins and reduced thermal stability compared to that of normal concrete. The polymers may be added to the fresh mix as the only binding material, e.g., epoxy concrete: polymer concrete (PC); they can be combined with water and cement, as in polymer cement concrete (PCC); or an already hardened concrete can be impregnated with monomers for subsequent polymerization as in polymer-impregnated concrete (PIC) [90].

Typical applications for polymer-modified cementitious materials are protective surface layers in aggressive environments or the repair and strengthening of damaged or deteriorated reinforced concrete structures [235], [236].

2.9. Mortar

Mortar is distinguished from concrete by the maximum size of its aggregates. In general, mortar contains only sand with a maximum grain size of 2 – 4 mm. Other binding materials such as hydraulic lime can be used in addition to cement.

In concrete construction, cement mortars are used for connecting precast elements, as injection mortars in ducts of posttensioned structures with bonded tendons, and for the repair of old concrete surfaces with areas of deterioration.

Because of the small maximum size of aggregates, mortars contain higher w/c ratios than are normally encountered in concrete technology. Therefore, they exhibit lower compressive strength and volume stability. These disadvantages can be offset by the use of water-reducing and expansion-producing additives [63], [100], [120].

The composition and strength requirements for masonry mortars are specified in national standards [162], [163] and international documents, e.g., [237]. Lightweight aggregate mortars with reduced thermal conductivity have been developed to improve the thermal insulation properties of a masonry structure. Polymer-modified cement-based mortars with fine sand as the aggregate make possible bed joints only ca. 3 mm thick (thin-layer mortar). Centrally mixed mortars may contain retarding additives to allow placement of the mortar up to 48 h after mixing.

A variety of polymer-modified mortars designed for the repair of concrete structures with deteriorated surface zones are available. Basic requirements of mortar for such applications are the ability to form strong bonds with old concrete and a low shrinkage strain. Furthermore, they must resist the prevailing environmental attack and prevent the penetration of aggressive compounds[235], [236].

3. References

General References

1. R. H. Bogue: *The Chemistry of Portland Cement,* 2nd ed., Reinhold Publ., New York 1955.
2. P. Barnes: *Structure and Performance of Cements,* Applied Science Publishers, London-New York 1983.
3. G. C. Bye: *Portland Cement – Composition, Production and Properties,* Pergamon Press, Oxford1983.
4. W. Czernin: *Zementchemie für Bauingenieure,* 3rd ed., Bauverlag, Wiesbaden-Berlin 1977.
5. W. H. Duda: *Cement Data Book,* 2nd ed., Bauverlag, Wiesbaden-Berlin 1977.
6. F. Keil: *Hochofenschlacke,* 2nd ed., Verlag Stahleisen, Düsseldorf 1963.
7. F. Keil: *Zement – Herstellung und Eigenschaften,* Springer Verlag, Berlin-Heidelberg-New York 1971.
8. H. Kühl: *Zement-Chemie,* 3rd. ed., **vol. 2, 3** VEB Verlag Technik, Berlin 1961.
9. F. M. Lea: *The Chemistry of Cement and Concrete,* 3rd ed., Edward Arnold, London 1970.
10. H. F. W. Taylor: *Cement Chemistry,* (2nd ed.) Telford Services, London 1997.
11. Verein Deutscher Zementwerke e.V.: *Zement-Taschenbuch.* 48th ed., Bauverlag, Wiesbaden-Berlin 1984, and 49th ed., Verlag Bau + Technik GmbH, Düsseldorf 2000.

Specific References

12. S. B. Newberry, W. B. Newberry: "The constitution of hydraulic cements," *J. Soc. Chem. Ind. London* **16** (1897) 887 – 893.

13. E. Wetzel: "Bericht über den Stand der auf Antrag des Vereins Deutscher Portland-Cement-Fabrikanten im Kgl. Materialprüfungsamt ausgeführten Arbeiten über die Konstitution des Portlandzements," *Protokolle des Vereins deutscher Portland-Zement-Fabrikanten* (1911) 281 – 306; *Protokolle des Vereins deutscher Portland-Zement-Fabrikanten* (1912) 217 –249; *Protokolle des Vereins deutscher Portland-Zement-Fabrikanten* (1913) 347 – 358; *Protokolle des Vereins deutscher Portland-Zement-Fabrikanten* (1914) 145 – 160358.
14. E. Spohn: "Die Kalkgrenze des Portlandzementes und die technischen Eigenschaften seiner Klinkermineralien," *Zement* **21** (1932) 702 – 706, 717 – 723, 731 – 736.
15. F. M. Lea, T. W. Parker: "The quaternary system $CaO - Al_2O_3 - SiO_2 - Fe_2O_3$ in relation to cement technology," *Build. Res. Techn. Paper* nr. 16, published by His Majesty's Stationery Office, London (1935).
16. H. Kühl: "Gelöste und ungelöste Aufgaben der Zementforschung," *Protokolle des Vereins deutscher Portland-Zement-Fabrikanten* (1936) 196 – 216.
17. A. Guttmann: "Der Einfluß von Gips- und Chlorkalziumzusätzen zum Zement auf sein Schwinden," *Zement* **9** (1920) 310 – 313, 429 – 432. DE 330 784,1920(A. Guttmann).
18. *Klein Symposium on Expansive Cement Concretes,* Publication SP-38, Am. Concrete Assoc., 1973.
19. Portland Cement Assoc., DE-OS 1 929 684, 1969.
20. Onoda Cement Co., DE-OS 2 165 434, 1971.
21. A. E. Törnebohm: "Die Petrographie des Portlandzements," *Tonind. Ztg.* **21** (1897) 1148 – 1151, 1157 – 1159.
22. J. H. Welch, W. Gutt: "Tricalcium silicate and its stability within the system $CaO-SiO_2$," *J. Am. Ceram. Soc.* **42** (1959) 11 – 15.
23. E. Woermann: "Decomposition of alite in technical portland cement clinker," *Proc. Int. Sympos. Chem. Cem., 4th* (1960) vol. I, 119 – 128.
24. A. Guinier, M. Regourd: "Structure of portland cement minerals," *Proc. Int. Sympos. Chem. Cem., 5th* (1968) vol. I, 1 – 32.
25. H. E. Schwiete, H. zur Strassen: "Über den Einfluß des Magnesiagehalts im Portlandzementklinker auf das Tetracalciumaluminatferrit," *Zement* **23** (1934) 511 – 514.
26. M. Regourd, A. Guinier: "The crystal chemistry of the constituents of portland cement clinker," *Proc. Int. Sympos. Chem. Cem. 6th* (1974) Principal Paper I – 2.
27. M. A. Swayze: "A report on studies of 1. the ternary system $CaO - C_5A_3 - C_2F$, 2. the quaternary system $CaO - C_5A_3 - C_2F - C_2S$, 3. the quaternary system as modified by 5 % magnesia," *Am. J. Sci.* **244** (1946) 1 – 30, 65 – 94.
28. H. Rechmeier: "Der fünfstufige Wärmetauscherofen zum Brennen von Klinker aus Kalkstein und Ölschiefer," *Zement-Kalk-Gips* **23** (1970) 249 – 253.
29. H. zur Strassen: "Der theoretische Wärmebedarf des Zementbrandes," *Zement-Kalk-Gips* **10** (1957) 1 – 12.
30. Verein Deutscher Zementwerke e.V.: *Mikroskopie des Zementklinkers,* Beton-Verlag, Düsseldorf 1965.
31. R. H. Bogue: "Calculation of the compounds in portland cement", *Ind. Eng. Chem., Analyt. Ed.* **1** (1929) 192 – 197.
32. F. W. Locher: "Berechnung der Klinkerphasen," *Zement-Kalk-Gips* **14** (1961) 573 – 580; *Schriftenr. Zementind.* **29** (1962) 7 – 19.
33. H. G. Smolczyk: "Zum Einfluß der Chemie des Hüttensands auf die Festigkeiten von Hochofenzementen," *Zement-Kalk-Gips* **31** (1978) 294 – 296.
34. F. Keil: "Zur Bewertung der Zementschlacken," *Zement* **33** (1944) 90 – 93.
35. G. Haegermann: "Über die Bewertung der Erhärtungsfähigkeit hydraulischer Zusatzstoffe in Mischungen mit Portlandzement," *Zement* **33** (1944) 93 – 97.
36. W. Wittekindt: "Zur Qualitätsbeurteilung von Hochofenschlacken und Puzzolanen," *Zement-Kalk-Gips* **16** (1963) 314 – 320.
37. R. L. Blaine: "A simplified air permeability fineness apparatus," *ASTM Bull.* **123** (1943) 51 – 55.
38. R. Rohrbach: "Herstellung von Ölschieferzement und Gewinnung elektrischer Energie aus Ölschiefer nach dem Rohrbach-Lurgi-Verfahren," *Zement-Kalk-Gips* **22** (1969) 293 – 296.
39. G. Wischers, W. Richartz: *Einfluß der Bestandteile und der Granulometrie des Zements auf das Gefüge des Zementsteins,* Beton-Verlag, Düsseldorf 1982.

40. F. W. Locher, W. Richartz, S. Sprung: "Erstarren von Zement," Part I: "Reaktion und Gefügeentwicklung," *Zement-Kalk-Gips* **29** (1976) 435 – 442; Part II: "Einfluß des Calciumsulfatzusatzes," *Zement-Kalk-Gips* **33** (1980) 271 – 277.
41. W. Richartz: "Einfluß von Zusätzen auf das Erstarrungsverhalten von Zement," *Beton* **33** (1983) 425 – 429, 465 – 471.
42. W. Richartz: "Über die Gefüge- und Festigkeitsentwicklung des Zementsteins," *Beton* **19** (1969) 203 – 206, 245 – 248.
43. L. E. Copeland, D. L. Kantro, G. Verbeck: "Chemistry of hydration of portland cement," *Proc. Int. Sympos. Chem. Cem. 4th* (1960) vol. 1, 429 – 465.
44. T. C. Powers, T. L. Brownyard: "Studies of the physical properties of hardened portland cement paste," Part 2: "Studies of water fixation," *Proc. Am. Concr. Inst.* **43** (1946) 249 – 336; *Res. Lab. Portl. Cem. Assoc., Bull.* 22 (1947).
45. T. C. Powers: "The nonevaporable water content of hardened portland cement paste – Its significance for concrete research and its method of determination," *ASTM Bull.* **158** (1949) 68 – 76; *Res. Lab. Portl. Cem. Assoc., Bull.* 29 (1949).
46. R. F. Feldman, P. Sereda: "A model for hydrated portland cement paste as deduced from sorption-length change and mechanical poperties," *Matér. Constr.* **1** (1968) 509 – 520.
47. T. C. Powers: "The thermodynamics of volume change and creep," *Matér. Constr.* **1** (1968) 487 – 507.
48. R. A. Helmuth, D. H. Turk: "The reversible and irreversible drying shrinkage of hardened portland cement and tricalcium silicate pastes," *J. Res. Portl. Cem. Assoc. Res. Dev. Lab.* **9** (1967) no. 2, 8 – 21; *Res. Dev. Lab. Portl. Cem. Assoc. Bull.* 215 (1967).
49. G. Wischers: "Physikalische Eigenschaften des Zementsteins," *Beton* **11** (1961) 481 – 486.
50. T. C. Powers: "Structure and physical properties of hardened portland cement paste," *J. Am. Ceram. Soc.* **41** (1958) 1 – 6.
51. S. Sprung, G. Thielen, F. Sybertz, "DIN 1164", *Beton* **45** (1995) no. 7, 490 – 497.
52. Verein Deutscher Ingenieure: VDI-Richtlinie 2094 *"Emissionsminderung Zementwerke;"* Draft, 1999.
53. S. Sprung: "Die chemische und mineralogische Zusammensetzung von Zementofenstaub," *Tonind. Ztg.* **90** (1966) 441 – 449.
54. F. W. Locher: "Stoffkreisläufe und Emissionen beim Brennen von Zementklinker," *Fortschr. Mineral.* **60** (1982) 215 – 234.
55. S. Sprung: "Technologische Probleme beim Brennen des Zementklinkers, Ursache und Lösung," *Schriftenr. Zementind.* no. 43 (1982).
56. S. Sprung: "Das Verhalten des Schwefels beim Brennen von Zementklinker," *Schriftenr. Zementind.* no. 31 (1964).
57. F. W. Locher, S. Sprung, D. Opitz: "Reaktionen im Bereich der Ofengase," *Zement-Kalk-Gips* **25** (1972) 1 – 12.
58. S. Sprung, H.-M. v. Seebach: "Fluorhaushalt und Fluoremission von Zementöfen," *Zement-Kalk-Gips* **21** (1968) 1 – 8.
59. W. Rechenberg: "Über das Verhalten von Mahlhilfen beim Mahlen von Zement," *Zement-Kalk-Gips* **39** (1986) no. 10, 577 – 580.
60. H. Pisters: "Chrom im Zement und Chromatekzem," *Zement-Kalk-Gips* **19** (1966) 467 – 472.
61. Cembureau: *European Annual Review* no. 18, Brussels 1997.
62. S. Mindess, J. F. Young: *Concrete,* Prentice-Hall, Inc., Englewood Cliffs 1981.
63. A. M. Neville: *Properties of Concrete,* Pitman Publ., London 1973.
64. *Zahlen und Daten,* Bundesverband der Deutschen Zementindustrie e.V., Köln 1999.
65. W. Marmé, J. Seeberger: "Der Primärenergieinhalt von Baustoffen," *Bauphysik* **4** (1982) no. 5, 155 – 165; **4** (1982) no. 6, 208 – 214.
66. W. A. Gutteridge, C. D. Pommery: "Cement in its Conventional Uses: Problems and Possibilities," *Technology in the 1990s: Developments in Hydraulic Cement,* The Royal Society, London 1983, pp. 7 – 15.
67. J. D. Birchall: "Cement in the context of new materials for an energy-expensive future," *Technology in the 1990s: Developments in Hydraulic Cements,* The Royal Society, London 1983, pp. 31 – 42.
68. H. K. Hilsdorf, J. Kropp: "Entwicklungstendenzen der Baustoffe," *Beratende Ingenieure* **1** (1982) 18 – 23.
69. DIN 1164: Teil 1: Zement; Zusammensetzung, Anforderungen (10/1994). DIN 1164: Teil 2: Zement; Übereinstimmungsnachweis (1996).
70. ASTM C 150-84: Specification for Portland Cement.
71. ASTM C 595-83: Specification for Blended Hydraulic Cements.
72. DIN VEN V 197: Definitions and Specifications for Cements (1984).

73. S. Popovics: *Concrete-Making Materials,* Hemisphere Publ. Corp., Washington 1979.
74. DIN 4226: Zuschlag für Beton (1983).
75. ASTM C 136-84: Method for Sieve Analysis of Fine and Coarse Aggregates.
76. ISO 6274: Concrete – Sieve Analysis of Aggregates.
77. DIN 1045: Beton und Stahlbeton – Bemessung und Ausführung (1978).
78. ASTM C 33-84: Standard Specification for Concrete Aggregates.
79. G. M. Idorn, S. Rostam: "Alkalis in Concrete – Research and Practice," *Proc. 6th Int. Conf.,* Copenhagen, 22 – 25 June 1983.
80. ASTM C 227-81: Potential Alkali Reactivity of Cement-Aggregate Combinations.
81. "Zugabewasser für Beton. Merkblatt für die Vorabprüfung und die Beurteilung vor Baubeginn sowie die Prüfungswiederholung während der Bauausführung, Jan. 1982," *Beton Stahlbetonbau* **77** (1982) no. 5, 137 – 140.
82. ASTM C 494-82: Specification for Chemical Admixtures for Concrete.
83. M. R. Rixom, *Chemical Admixtures for Concrete,* J. Wiley & Sons, Inc., New York 1978.
84. *Superplasticizers in Concrete,* American Concrete Institute, Detroit 1979, Publication SP – 62.
85. ASTM C 457-82 a: Standard Practice for Microscopical Determination of Air-Void Content and Parameters of the Air-Void System in Hardened Concrete.
86. ASTM C 618-84: Standard Specification for Fly Ash or Calcined Pozzolan for Use as a Mineral Admixture in Portland Cement Concrete.
87. DIN 51043: Trass, Anforderungen, Prüfung (1979).
88. *Fly Ash, Silica Fume, Slag and Other Mineral By-Products in Concrete,* **vol. I,** II, American Concrete Institute, Detroit 1983, Publication SP 79.
89. L. Hjorth: "Development and Application of High-Density Cement-Based Materials," *Technology in the 1990s: Developments in Hydraulic Cements,* The Royal Society, London 1983, pp. 167 – 174.
90. *Polymers in Concrete,* American Concrete Society, Detroit 1978, Spec. Publ. SP 58.
91. "Guide to Durable Concrete," ACI Committee 201, *ACI Manual of Concrete Practice,* Part 1, American Concrete Institute, Detroit 1979.
92. "Recommended Practice for Selecting Proportions for Normal and Heavyweight Concrete," ACI Committee 211, *ACI Manual of Concrete Practice,* Part 1, American Concrete Institute, Detroit 1979.
93. DIN 1048: Prüfverfahren für Beton (1978).
94. ASTM C 39-83 b: Test Method for Compressive Strength of Cylindrical Concrete Specimens.
95. ISO 4012: Concrete – Determination of Compressive Strength of Test Specimens.
96. RILEM Recommendation CPC 4: *Compression Test.*
97. DIN 1084: Überwachung im Beton- und Stahlbetonbau (1978).
98. ISO 3893: Concrete – Classification by Compressive Strength.
99. K. Walz: *Herstellung von Beton nach DIN 1045,* Bauverlag GmbH, Düsseldorf 1972.
100. Deutscher Betonverein e. V.: *Betonhandbuch,* Bauverlag, Wiesbaden 1984.
101. "Recommended Practice for Measuring, Mixing Transporting and Placing Concrete," ACI Committee 304, *ACI Manual of Concrete Practice,* Part I, American Concrete Institute, Detroit 1979.
102. ASTM C 143-78: Test Method for Slump of Portland Cement Concrete.
103. RILEM Recommendation CPC 2.1: *Slump Test* .
104. ISO 4109: Fresh Concrete – Determination of the Consistency – Slump Test.
105. ISO 4110: Fresh Concrete – Determination of the Consistency – Vebe Test.
106. RILEM Recommendation CPC 2.2: Vebe Test.
107. ISO 4111: Fresh Concrete – Determination of the Consistency – Degree of Compactibility.
108. ACI Standard 309-72:*Recommended Practice for Consolidation of Concrete.*
109. "Recommended Practice for Curing Concrete," ACI Committee 308, *ACI Manual of Concrete Practice,* Part I, American Concrete Institute, Detroit 1979.
110. "Hot Weather Concreting," ACI Committee 305, *ACI Manual of Concrete Practice,* Part I, American Concrete Institute, Detroit 1979.
111. *Empfehlungen zur Nachbehandlung von Beton,* ed. 1983, Deutscher Ausschuß für Stahlbeton, Berlin 1984.
112. ACI Standard 306: *Recommended Practice for Cold Weather Concreting.*
113. "High Pressure Steam Curing: Modern Practice, and Properties of Autoclaved Products," ACI Committee 516, *ACI Manual of Concrete Practice,* Part III, American Concrete Institute, Detroit 1978.

114. ACI Standard 517:*Recommended Practice for Atmospheric Pressure Steam Curing of Concrete.*
115. ACI Standard 506: *Recommended practice for shotcreting.*
116. ACI Standard 506.2:Specification for Materials, Proportioning, and Application of Shotcrete.
117. DIN 18551: Spritzbeton; Herstellung und Prüfung.
118. *Richtlinie für die Ausbesserung und Verstärkung von Betonbauteilen mit Spritzbeton,* Deutscher Ausschuß für Stahlbeton, Berlin 1976.
119. "Stahlfaserspritzbeton," *Beton Stahlbeton* **79** (1984) no. 5, 134 – 136.
120. K. Wesche: *Baustoffe für tragende Bauteile,* vol. 2, Bauverlag GmbH, Wiesbaden 1981.
121. J. Eibl, G. Ivanyi: *Studie zum Trag- und Verformungsverhalten von Stahlbeton,* no. 260, Deutscher Ausschuß für Stahlbeton, Verlag Ernst & Sohn, Berlin 1976.
122. P.-E. Petersson: "Crack Development and Fracture Zones in Plain Concrete and Similar Materials," *Report TVMB – 1006,* Lund, Sweden, 1981.
123. T. C. Powers, T. L. Brownyard: "Studies of the Physical Properties of Hardened Cement Paste," *Res. Dev. Lab. Portland Cem. Assoc. Res. Dep. Bull.* 1948, no. 22.
124. ACI Standard 214:Recommended Practice for Evaluation of Strength Test Results of Concrete.
125. ASTM C 192-81: Method of Making and Curing Concrete Test Specimens in the Laboratory.
126. ASTM C 597-83: Test Method for Pulse Velocity Through Concrete.
127. ASTM C 803-82: Test Method for Penetration Resistance of Hardened Concrete.
128. ASTM C 805-79: Test Method for Rebound Number of Hardened Concrete.
129. H. Rüsch, R. Sell, C. Rasch, E. Grasser, A. Hummel, K. Wesche, H. Flatten: *Festigkeit und Verformung von unbewehrtem Beton unter konstanter Dauerlast,* no. 198, Deutscher Ausschuß für Stahlbeton, Verlag Ernst & Sohn, Berlin 1968.
130. RILEM Report 36 RDL: *Long Term Random Dynamic Loading of Concrete Structures,* Feb.1980.
131. *Model Code for Concrete Structures,* CEB-FIP International Recommendations, 3rd ed., Comité Euro-International du Béton (CEB), 1978.
132. ASTM C 78-84: Test Method for Flexural Strength of Concrete (Third Point Loading).
133. ASTM C 293-79: Test Method for Flexural Strength of Concrete (Center Point Loading).
134. ASTM C 496-71: Test Method for Splitting Tensile Strength of Cylindrical Concrete Specimens.
135. RILEM Recommendation CPC 5: *Flexural Test.*
136. RILEM Recommendation CPC 6: *Tension by Splitting.*
137. RILEM Recommendation CPC 7: *Direct Tension.*
138. F. H. Wittman (ed.): *Fracture mechanics of concrete,* Elsevier Science Publ., Amsterdam 1983.
139. H. K. Hilsdorf, W. Brameshuber: *Size Effects in the Experimental Determination of Fracture Mechanics Parameters,* Nato Adv. Res. Workshop, Evanston, Illinois, USA, 1984.
140. ASTM C 469-83: Test Method for Static Modulus of Elasticity and Poisson's Ratio of Concrete in Compression.
141. RILEM Recommendation CPC 8: *Modulus of Elasticity of Concrete in Compression* .
142. H. Rüsch: *Stahlbeton-Spannbeton,* vol. 1: "Werkstoffeigenschaften und Bemessungsverfahren," Werner-Verlag, Düsseldorf 1972.
143. H. Rüsch, D. Jungwirth, H. K. Hilsdorf: *Creep and Shrinkage,* Springer Verlag, New York 1983.
144. ACI Standard 211.2: *Recommended Practice for Selecting Proportions for Structural Lightweight Concrete* .
145. DIN 4219: Leichtbeton und Stahlleichtbeton mit geschlossenem Gefüge (1979).
146. M. J. Setzer: *Einfluß des Wassergehaltes auf die Eigenschaften des erhärteten Betons,* no. 280, Deutscher Ausschuß für Stahlbeton, Verlag Ernst & Sohn, Berlin 1977.
147. "Durability of Concrete Structures Under Normal Outdoor Exposure," *Proceedings of the RILEM Seminar,* Hannover, 26th – 29th March 1984.
148. S. Ziegeldorf, K. Kleiser, H. K. Hilsdorf: *Vorherbestimmung und Kontrolle des thermischen Ausdehnungskoeffizienten von Beton,* no. 305, Deutscher Ausschuß für Stahlbeton, Verlag Ernst & Sohn, Berlin 1979.
149. T. Jaeger, *Grundzüge der Strahlenschutztechnik,* Springer Verlag, Berlin 1960.

150. R. G. Jaeger: *Engineering Compendium on Radiation Shielding,* vol. II: "Shielding materials," Springer Verlag, Berlin 1975.
151. J. Seeberger, H. K. Hilsdorf: *Einfluß von radioaktiver Strahlung auf die Festigkeit und Struktur von Beton,* Institut für Massivbau und Baustofftechnologie, Universität Karlsruhe 1982.
152. S. Rostam: "Durability of concrete structures," *CEB-RILEM International Workshop,* Copenhagen, 18 – 20 May 1983.
153. I. Biczok: *Betonkorrosion, Betonschutz,* Bauverlag GmbH, Wiesbaden 1968.
154. DIN 4030: Beurteilung betonangreifender Wässer, Böden und Gase; Grundlagen und Grenzwerte (6/1991).
155. "Guide for the Protection of Concrete Against Chemical Attack by Means of Coatings and Other Corrosion-Resistant Materials," ACI Committee 515, *ACI Manual of Concrete Practice,* Part 3, American Concrete Institute, Detroit 1978.
156. J. Seeberger, J. Kropp, H. K. Hilsdorf: *Festigkeitsverhalten und Strukturänderungen von Beton bei Temperaturbeanspruchung bis 250 °C,* no. 360, Deutscher Ausschuß für Stahlbeton, Verlag für Architektur und technische Wissenschaften, Berlin 1985.
157. U. Schneider: *Verhalten von Beton bei hohen Temperaturen – Behavior of Concrete at High Temperatures,* no. 337, Deutscher Ausschuß für Stahlbeton, Verlag Ernst & Sohn, Berlin 1982.
158. K. Wesche: *Baustoffe für tragende Bauteile,* vol. 3, Bauverlag GmbH, Wiesbaden 1985.
159. "Mass Concrete for Dams and Other Massive Structures," ACI Committee 207, *ACI Manual of Concrete Practice,* Part 1, American Concrete Institute, Detroit 1979.
160. *Sachstandbericht Massenbeton,* Deutscher Ausschuß für Stahlbeton, no. 329, Verlag Ernst & Sohn, Berlin 1982.
161. A. Neville: *Fibre Reinforced Cement and Concrete,* The Construction Press, Hornby, Lancaster, 1975.
162. ASTM C 270-82: Specification for Mortar for Unit Masonry.
163. DIN 1053: Mauerwerk – Berechnung und Ausführung (1974).
164. F. W. Locher: *Zement – Grundlagen der Herstellung und Verwendung, Cement – Basic principles of the production and use,* Verlag Bau + Technik GmbH, Erkrath, (in preparation).
165. H.-G. Ellerbrock, H. Mathiak: "Comminution technology and energy management," *Zement-Kalk-Gips* **47** (1994) no. 9, 524 – 534.
166. DIN 1164-1: Zement – Zusammensetzung, Anforderungen (1994).
167. Verein Deutscher Zementwerke e.V.: Activity Report 1996 – 1999, Düsseldorf 1999.
168. H.-G. Ellerbrock, S. Sprung, K. Kuhlmann: "Korngrößenverteilung und Eigenschaften von Zement – Teil 3," *Zement-Kalk-Gips* **43** (1990) no. 1, 13 – 19.
169. S. Sprung, K. Kuhlmann, H.-G. Ellerbrock: "Korngrößenverteilung und Eigenschaften von Zement – Teil 2," *Zement-Kalk-Gips* **36** (1985) no. 9, 528 – 534.
170. prEN 196-8: Methods of testing cement – Part 8: Determination of heat of hydration – solution method.
171. prEN 198-9: Methods of testing cement – Part 9: Determination of heat of hydration – semiadiabatic method.
172. prENV 197-10: Calcium aluminate cement: Composition, specifications and conformity criteria Draft August 2000.
173. S. Sprung: "Umweltentlastung durch Verwertung von Sekundärstoffen," *Zement-Kalk-Gips* **45** (1992) no. 5, 213 – 221.
174. Verein Deutscher Zementwerke e.V.: *5th Monitoring Report (1998) on the reduction of CO_2 emissions,* Düsseldorf 1998.
175. Verein Deutscher Zementwerke e.V.: *Environmental compatibility of cement and concrete. Manufacture, application and use of alternative materials. Information from the German cement industry,* Düsseldorf 1996.
176. S. Sprung: *"Energy Saving Potential in the Cement Sector,"* Cembureau General Assembly 1995, Istanbul.
177. Verein Deutscher Zementwerke e.V.: *"Execution and evaluation of rotary kiln investigations,"* Internal Report of the VDZ-Committee of Process Technology, Düsseldorf 1992.
178. Spenner Zement GmbH & Co. KG, Erwitte, Werk Diamant: *Investitionen zur Vermeidung von Umweltbelastungen* – Programm des Bundesministers für Umwelt, Naturschutz und Reaktorsicherheit, Abschlußbericht 3071-5/ 207 v Dez. 1995.
179. D. Kupper, K. Adler: "Multi-Stage Combustion Minimises NO_x Emissions," *Int. Cement Rev.* (1993) no. 6, 61 – 69.
180. S. Sprung: *Technologische Probleme beim Brennen des Zementklinkers, Ursache und*

Lösung, Schriftenreihe der Zementindustrie, Heft 43/1982, VDZ, Düsseldorf 1982.
181. A. Wolter: "Einfluß des Ofensystems auf die Klinkereigenschaften," *Zement-Kalk-Gips* **38** (1985) no. 10, 612 – 614.
182. V. Hoenig, H.-M. Sylla: "Industrial clinker cooling with due regard to the cement properties," *Zement-Kalk-Gips* **51** (1998) no. 6, 318 – 333.
183. H.-M. Sylla: *"Einfluß der Klinkerzusammensetzung und Klinkerkühlung auf die Zementeigenschaften,"* Verfahrenstechnik der Zementherstellung, VDZ-Kongress '93, Bauerverlag, Wiesbaden 1995, pp. 135 – 145.
184. Verein Deutscher Zementwerke e.V.: Activity report 1969 – 71, Düsseldorf 1971, p. 65.
185. H. M. Rietveld: "Lime Profiles of neutron powder-diffraction peaks for structure refinement," *Acta Cryst.* **22** (1967) 151 – 152.
186. H. Möller, "Standardless quantitative phase analysis of Portland cement clinkers," *World Cement* (1995) no. 9, 75/84.
187. Verein Deutscher Zementwerke: *Clinker Storage – Guidance Paper of the VDZ committee on Process Technology,* Düsseldorf 1984.
188. H. Rosemann, H.-G. Ellerbrock: "Grinding technology for cement production," *Zement-Kalk-Gips* **51** (1998) no. 2, 51 – 62.
189. J. Krell: *"Die Konsistenz von Zementleim, Mörtel und Beton und ihre zeitliche Veränderung,"* Schriftenreihe der Zementindustrie Heft 46/1985, VDZ, Düsseldorf 1985.
190. J. Stark, B. Wicht: *Anorganische Bindemittel,* Schriften der Bauhaus-Universität Weimar, Heft 109, 1998.
191. H. Gundlach: *Dampfgehärtete Baustoffe,* Bauverlag GmbH, Wiesbaden – Berlin 1973.
192. H. Ludwig: *Wasserbindung und Raumänderung von Zementstein,* Dissertation TU Clausthal, 1985.
193. F. H. Wittmann: *Grundlagen eines Modells zur Beschreibung charakteristischer Eigenschaften des Betons,* DafStb, Heft 290, Verlag W. Ernst u. Sohn, Berlin 1977.
194. F. W. Locher, W. Richartz, S. Sprung, W. Rechenberg: "Erstarren von Zement – Teil IV: Einfluß der Lösungszusammensetzung," *Zement-Kalk-Gips* **36** (1983) no. 4, 224 – 231.
195. S. Sprung, W. Rechenberg, G. Bachmann: "Environmental compatibility of cement," *Zement-Kalk-Gips* **47** (1994) no. 8, 456 – 461 (englisch: Nr. 10, E 262 – 267).
196. S. Sprung, W. Rechenberg: "Levels of heavy metals in clinker and cement," *Zement-Kalk-Gips* **47** (1994) no. 5, 258 – 263 (englisch: Nr. 7, E 183 – 188).
197. Harmonized European Standard EN 197: Part 1: Cement – Composition, specification and conformity criteria for common cements - June 2000. Part 2: Cement – Conformity evaluation - June 2000.
198. *COUNCIL DIRECTIVE of 21 December 1988 on the approximation of laws, regulations and administrative provisions of the Member States relating to construction products* (89/106/EEC), Official Journal of the European Communities, (1989) no. L40, 12 – 26.
199. G. Thielen, F. Sybertz: "Attestation of conformity (quality inspection) of cement," *Beton* **45** (1995) no. 8, 562 – 568.
200. S. Gajewski: *Theoretische und betriebliche Untersuchung feuerungstechnischer Maßnahmen zur NO_x-Minderung an Drehofenanlagen der Zementindustrie,* Schriftenreihe der Zementindustrie, Heft 60/1999, VDZ, Düsseldorf 1999.
201. A. Scheuer: "Non-catalytic reduction of NO with NH_3 in the cement burning process," *Zement-Kalk-Gips* **43** (1990) no. 3, E 50 – 54.
202. M. Kietzmann, W. Bäumer, E. Bien, D. Lubach: "Anmerkungen zu hautirritierenden Wirkungen von Zement," *Dermatosen in Beruf und Umwelt* **47** (1999) no. 5, 184 – 189.
203. S. Sprung, W. Rechenberg, G. Bachmann: *Environmental compatibility of cement and concrete. Environmental aspects of construction with waste materials,* Elsevier, Amsterdam 1994, pp. 368 – 386.
204. H. Grube, W. Hintzen: Test method for predicting the temperature rise in concrete by the heat of hydration of the cement, *Beton* **43** (1993) no. 5, 230 – 234, no. 6, 292 – 295.
205. W. Manns: "Gemeinsame Anwendung von Silicastaub und Steinkohlenflugasche als Betonzusatzstoff – Eine technologische Betrachtung," *Beton* **47** (1997) no. 12, 716 – 720.
206. I. Hohberg, C. Müller, P. Schießl, G. Volland: Umweltverträglichkeit zementgebundener Baustoffe – Sachstandsbericht, Deutscher Ausschuss für Stahlbeton, Heft 458/1996, Verlag Ernst & Sohn, Berlin 1996.
207. DAfStb Richtlinie Beton mit rezykliertem Zuschlag, Ausgabe August 1998.

208. "Rilem Technical Recommendation Specifications for concrete with recycled aggregates," *Materials Structures* **27** (1994) 557 – 559.
209. "Rilem Test Method for Alkali-Silica Reactivity of aggregate combinations using concrete prisms," *Materials Structures* **29** (1996) 326 – 334.
210. P. Nixon, I. Sims: "Testing aggregates for alkali-reactivity," *Materials Structures* **29** (1996) 323 – 326.
211. DAfStb-Richtlinie Vorbeugende Maßnahmen gegen schädigende Alkalireaktionen im Beton, 1997.
212. DAfStb-Richtlinie zur Herstellung von Beton unter Verwendung von Restwasser, Restbeton und Restmörtel, 1995.
213. Richtlinien für die Erteilung von Zulassungen für Betonzusatmittel, Ausgabe 12-1996, Mitteilungen Deutsches Institut für Bautechnik 28 (1997) no. 5.
214. prEN 934 des CEN TC 104 Admixtures for Concrete.
215. DIN EN 450 Flugasche für Beton: Definitionen, Anforderungen und Güteüberwachung, 1994.
216. DIN EN 451-1 Prüfverfahren für Flugasche – Bestimmung des freien Calciumoxides, 1994.
217. EN 451-2 Prüfverfahren für Flugasche – Bestimmung der Feinheit durch Nasssiebung, 1994.
218. DAfStb-Richtlinie Verwendung von Flugasche nach DIN EN 450 im Betonbau, 1996.
219. DIN V EN V 206 Beton: Eigenschaften, Herstellung, Verarbeitung und Gütenachweis, Ausgabe 10-1990.
220. DIN V EN V 1992 Eurocode 2: Planung von Stahlbeton- und Spannbetontragwerken, Teil 1, Ausgabe 6.92.
221. CEB FIP Model Code 1990. Comité Euro-International du Béton, Bulletin d'Information No. 213/214, 1993, Lausanne.
222. DAfStb-Richtlinie für Fließbeton: Herstellung, Verarbeitung und Prüfung, Ausgabe 1995.
223. A. Skarendahl, Ö. Peterssen (eds.): *"Self compacting concrete,"* Proceedings of the First RILEM International Symposium, Stockholm 1999; RILEM Proceedings PRO 7, Cachan, France.
224. K. D. Hansen, R. W. Reinhardt: *Roller Compacted Concrete Dams,* Mc Graw Hill, New York 1991.
225. D. R. Lankard: "Slurry infiltrated fiber concrete (SIFCON) – Properties and applications," *Proc. Mater. Res. Soc. Symp.* **42** (1985).
226. I. Schrage: Hochfester Beton – Sachstandsbericht Teil I des Deutschen Ausschusses für Stahlbeton, Heft 438, Beuth Verlag, Berlin 1994.
227. High Strength Concrete – State of the art report, CEB Bulletin d'Information No. 197, Lausanne, 1990.
228. DIN 4227 Spannbeton: Bauteile aus Normalbeton mit beschränkter oder voller Vorspannung, 1998.
229. H. P. G. Darcy: *Les fontaines de la ville de Dijon,* Dalmont, Paris 1856.
230. J. Kropp, H. K. Hilsdorf (eds.): Performance criteria for concrete durability, RILEM Technical Report No. 12, F&N Spon, London, 1995.
231. J. Kropp: "Concrete Durability – An approach towards performance testing, Final report of RILEM Technical Committee 116-PCD," *Materials Structures* **32** (1999) 163 – 173.
232. H. W. Reinhardt (ed.): *Penetration and Permeability of concrete: Barriers to organic and contaminating liquid,* RILEM Technical Report No. 16, F&N Spon, London 1997.
233. DAfStb-Richtlinie für Betonbau beim Umgang mit wassergefährdenden Stoffen, Ausgabe 1996.
234. J. Kropp: "Struktur und Eigenschaften karbonatisierter Betonrandzonen," *Bautenschutz Bautensanierung* **9** (1986) 33 – 38 .
235. DAfStb-Richtlinie für Schutz und Instandsetzung von Betonbauteilen; Teil 1 Allgemeine Regeln und Planungsgrundsätze, 1990; Teil 2 Bauplanung und Bauausführung, 1990; Teil 3 Qualitätssicherung der Bauausführung, 1991; Teil 4 Qualitätssicherung der Bauprodukte 1992.
236. ENV 1504: Products and systems for the protection and repair of concrete structures; Definitions, requirements, quality control and evaluation of conformity, 1997.
237. DIN V EN V 1996: Eurocode 6 – Allgemeine Regeln für den Hochbau – Regeln für bewehrtes und unbewehrtes Mauerwerk, Teil 1, Ausgabe 12-1996.

Cemented Carbides → Hard Materials

Cements, Chemically Resistant

Jürgen Fenner, Keramchemie GmbH, Siershahn/Westerwald, Federal Republic of Germany

1. Introduction 1
2. Types of Cement 2
2.1. Cements Containing Inorganic Binders 2
2.1.1. Silicate Cements 2
2.1.2. Sulfur Cements 2
2.1.3. Hydraulic Cement Mortars 2
2.2. Cements Containing Organic Binders 3
2.2.1. Phenol–Formaldehyde Resin Cements 3
2.2.2. Furan Resin Cements 3
2.2.3. Epoxy Resin Cements 3
2.2.4. Unsaturated Polyester Resin Cements 4
2.2.5. Vinylester Resin Cements 4
2.2.6. Bituminous Material Cements 5
3. Specifications and Testing 5
4. Storage and Transportation 6
5. Toxicology 6
6. References 6

1. Introduction

Chemically resistant cements are materials used in chemically resistant tiling or brick lining to join together such chemically resistant, nonmetallic units as tiles, bricks, stones, blocks, etc., and to the membrane (Figs. 1 and 2) [9], [10].

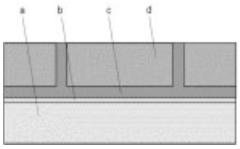

Figure 1. Fully bedded and jointed bricks
a) Substrate (concrete or steel); b) Membrane; c) Cement; d) Tiles or bricks

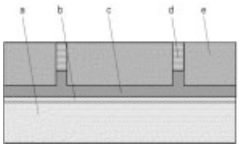

Figure 2. Brick lining with open joints
a) Substrate (concrete or steel); b) Membrane; c) Bedding cement; d) Jointing cement; e) Tiles or bricks

Chemically resistant units are, for example, ceramic tiles and bricks, refractory tiles and bricks, carbon bricks, parts made from porcelain, cast basalt, graphite, silicon carbide, granite, etc., alone or in combination with other materials. Chemically resistant bricklining is used, for instance, to protect the floors of production and storage buildings, pits, trenches, emergency containments, and production and storage tanks against chemicals that are produced or stored in these places.

Chemically resistant brick lining can be self-supporting, e.g., in chimneys, or it can serve as a mechanical, thermal, and, to some extent, a chemical barrier on top of a liquid-tight membrane [1–3].

Acid-proof, or rather chemically resistant, cement or mortar always consists of an inorganic or organic binder and inorganic fillers. Most of these materials are similar to mortar at normal temperatures, and they harden when the binder undergoes a chemical reaction. Some cements, however, can also be worked at higher temperatures; these harden as a result of physical setting or crystallize when cooled.

The cement of choice depends primarily on the desired chemical resistance and physical properties. Further factors to be considered are the compatibility with other components of the masonry, working time, setting time, stability during storage, transportation requirements, and price [4].

Stress caused by shrinkage and by changes in volume that occur during the hardening process must also be considered [5].

The oldest chemically resistant cements are the silicate cements, which date from about 1920. Even bitumen-bound materials gained early importance and were used especially for flooring. In the 1930s, cements based on phenol–formaldehyde resins were developed. Furan resins were introduced in the next decade, fol-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_539

lowed in the 1950s by unsaturated polyester and epoxy resins [2] and vinylester resins.

2. Types of Cement

Chemically resistant mortar is classified according to the binder it contains, and is further classified according to the type of filler it contains. Even if the binder and filler are the same, the properties of cements from different producers can differ with respect to mixing and application.

The following specifications, outlined according to the respective ASTM standards [6], are characteristic of each binder, but do not consider properties typical of any particular cement manufacturer.

2.1. Cements Containing Inorganic Binders

2.1.1. Silicate Cements

Soluble silicate cements contain potassium silicate or sodium silicate as the binder. The filler is quartz or another inert material. Sodium fluorosilicate, potassium fluorosilicate, aluminum phosphate, organic acid esters, amines, or formamide may be used as the hardener, which is usually mixed with the filler. Soluble silicate cements are delivered as one liquid component and one powder. They are used according to the instructions given in ASTM C397. The mixture is placed at ambient temperature and hardens without heating as a result of the coagulation of silicon dioxide caused by a change in pH. The hardened cement can withstand temperatures up to 850 °C.

Silicate cements are used to produce such self-supporting, chemically resistant constructions as the linings of chimneys, foundations, and other containers and equipment subject to high temperatures or exposed to acids. Because they are sensitive to water, they are not recommended for use in flooring.

Trade name: Keranol WG 300 (Keramchemie, FRG), Hoechst Acidproof Cement HB (Permatex/Hoechst, FRG), Acidproof Cement S 50 HF (Steuler Industriewerke, FRG).

2.1.2. Sulfur Cements

Quartz or carbon is the filler used in cements with a sulfur binder. The powdered cement is used in accordance with ASTM C386. It is heated until the sulfur melts and becomes free flowing (138 – 149 °C). The molten cement is poured into the open joints between spaced blocks, where it hardens upon cooling. The hardened cement can resist temperatures up to 88 °C.

Sulfur cements are used, primarily in the United States, for the linings of containers used in the pickling of steel with mixtures of nitric acid and hydrofluoric acid. Their use is limited because they are difficult to handle.

Trade name: Corobond (Ceilcote, USA), Basolit no. 600, Basolit no. 610 Sauereisen, USA).

2.1.3. Hydraulic Cement Mortars

Hydraulic cement mortars contain blast-furnace cement or portland cement which binds a quartz sand filler. They are delivered in the form of a cement powder accompanied by dry or moist sand and are used in accordance with ASTM C398. The mortar is worked using routine construction industry methods. When hydrated, the mortar hardens to a crystalline structure that resists temperatures up to 400 °C.

Hydraulic cement mortars are used for ceramic flooring that is laid with open joints which are subsequently filled with synthetic resin cements. Other uses include ceramic container linings (especially those used in the cellulose industry) and self-supporting constructions.

2.2. Cements Containing Organic Binders

2.2.1. Phenol–Formaldehyde Resin Cements

Phenol–formaldehyde resins or modified phenol–formaldehyde resins are used as the binder in some cements. The latter often contain furfural [*98-01-1*] to increase the chemical resistance.Today it is possible to produce furfural-free cements with adequate chemicalstabilities. The filler is quarz, another inert material, or carbon and is usually mixed with a hardener such as *p*-toluenesulfonic acid, naphthalenesulfonic acid, or sulfonyl chloride. The cement is delivered as a liquid accompanied by a powder and is used in accordance with ASTM C399. The mixture is worked at ambient temperature and hardens (without heating) by the polycondensation of

phenol–formaldehyde resin. The resulting bond is resistant to temperatures up to 180 °C.

Steel and concrete should not come into direct contact with phenolic resin.based mortars because of corrosive active hardeners.

Phenol–formaldehyde resin cements are used in laying and jointing of flooring made from nonmetallic inorganic tiles or bricks. They are used to produce chemically resistant masonry. Containers and equipment subject to high pressure and high temperature can be lined with compressive prestressed bricklining by using a mortar made with a modified phenol–formaldehyde resin cement [1], [3].

Trade name: Asplit CN (Permatex/Hoechst, FRG), Asplit CN 916 (Permatex/Hoechst, FRG).

2.2.2. Furan Resin Cements

Furan resin cements employ the polycondensation products of furfuryl alcohol [*98-00-0*] as the binding agent. Furfural [*98-01-1*] is usually included as a reactive thinner. The filler is quartz, another inert material, or carbon and is usually mixed with the hardener. The hardener can be an acid, such as amidosulfuric acid [*5329-14-6*] or *p*-toluenesulfonic acid [*104-15-4*], an acid chloride, a urea salt, or an amine salt. The cement is deliveredas a liquid accompanied by a powder and is used in accordance with ASTM C399. The mixture is worked at ambient temperature and hardens (without heating) due to polycondensation of the furan resin to form a bond capable of withstanding temperatures up to 200 °C. With special formulations temperatures even up to 250 °C are possible. Steel and concrete should not come into direct contact with furan resin-based mortars because of the corrosive active hardeners (Fig.3).

Furan resin cements containing quartz or another inert material are widely used because of their resistance to a broad spectrum of chemicals and their excellent storage properties. They are used to lay and joint flooring and to line trenches, pits, and collecting tanks. The cements that contain a carbon filler are used to solve special problems.

Trade name: Keranol FU 310 (Keramchemie, FRG), Keranol FU 315 (furan-free; Keramchemie, FRG), Asplit FQ (Permatex/Hoechst, FRG), Furadur Mortar (Steuler Industriewerke, FRG).

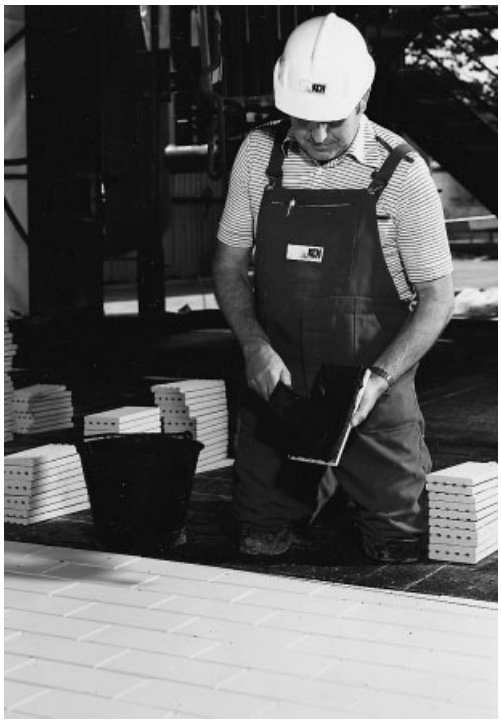

Figure 3. Application of ceramic tiles in furan resin cement (applied with open joints)

2.2.3. Epoxy Resin Cements

Epoxy resin cements use the product of condensation of bisphenol A [*80-05-7*] with epichlorohydrin as their binder, if necessary with a reactivethinner. The filler is quartz, another inert material, or (infrequently) carbon, and the hardener is a polyamine, modified polyamine, or polyamidoamine. The cement is delivered either as two liquids and a powder or as a liquid and a paste; it is used in accordance with ASTM C399. The mixture is worked at ambient temperature and hardens (without heating) due to polyaddition of the epoxy resin to the polyamine to form a bond capable of resisting temperatures up to 100 °C (Fig. 4).

Epoxy resin cements are widely used to bond light-colored ceramic surfaces, such as those used in the food and beverage industry, the production of drinking water, and the treatment of cellulose. They are also used to lay and joint flooring.

Figure 4. Floor tiling with chemical resistant tiles bedded and jointed with epoxy resin cement

Trade names: Keranol EP 310 (Keramchemie, FRG), Keranol EP 110 (Keramchemie, FRG) Asplit ET (Permatex/Hoechst, FRG), Alkadur K 75 (Steuler Industriewerke, FRG).

2.2.4. Unsaturated Polyester Resin Cements

The products of the condensation of unsaturated polycarboxylic acids with polyalcohols dissolved in styrene serve to bind unsaturated polyester resin cements. Products formed by the reaction of unsaturated polyester resins with polyisocyanate or modified polyisocyanate dissolved in styrene are also added occasionally. The filler is quartz, another inert material, or carbon. The hardener is an organic peroxide, e.g., benzoyl peroxide; it is mixed with the filler unless the filler is carbon. Unsaturated polyester resin cements that do not contain carbon as the filler are delivered as one or two liquids along with a powder. Those with a carbon filler are delivered as two liquids, one powder, and one paste. These cements are used in accordance with ASTM C397. The mixture is worked at ambient temperature and hardens (without heating) by the polymerization of the unsaturated polyester resin with styrene to form a bond that resists temperatures up to 100 °C.

Unsaturated polyester resin cements are used primarily where oxidation reactions are performed, e.g., bleaching in the cellulose industry. They are also employed for light-colored jointing in the food and beverage industry, in flooring made from nonmetallic inorganic tiles, and in lining foundations. Cements containing carbon are used to prepare the brick linings of containers used for pickling steel with nitric acid and hydrofluoric acid. Their relatively large shrinkage upon hardening should be kept in mind.

Trade names: Asplit OC (Permatex/Hoechst, FRG), Keranol UP 311 (Keramchemie, FRG), Keranol UP 320 (Keramchemie, FRG), Oxydur A Mortar (Steuler Industriewerke, FRG).

2.2.5. Vinylester Resin Cements

The vinylester resin based on bisphenol-A-acrylate and novolak – acrylate contains styrene as reactive thinner. The filler is quarz, another inert material or carbon. The hardener is a peroxide. The mixture is worked at ambient temperature and hardeners (without heating) by polymerization of the vinylester resin with styrene to form a bond that resists temperatures up to 140 °C (Fig. 5).

Vinylester resin are widely used because of their resistance to a broad spectrum of chemicals, especially against oxidizing chemicals like nitric acid, chromic acid and sodium hypochlorite. They are used to lay and joint flooring made from acidproof ceramic tiles and cements containing carbon are used to prepare the brick linings of containers used for pickling steel with nitric acid and hydrofluoric acid. Vinylester resin cements are especially in textile industry, bleaching in the cellulose industry and for electroplating plants.

Trade names: Asplit VE-series (Permatex/Hoechst, FRG), Keranol VE 310 (Keramchemie, FRG), Keranol VE 311 (Keramchemie, FRG), Oxydur VE Mortar (Steuler Industriewerke, FRG).

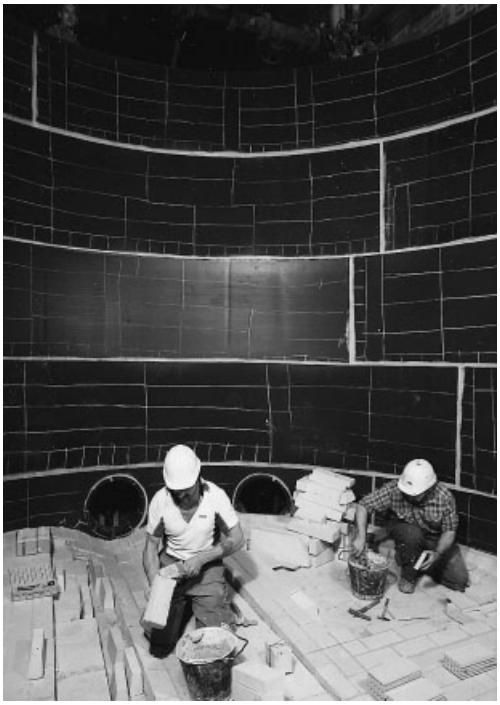

Figure 5. Application of brick lining with vinylester resin cement at the bottom of a rubber lined tank

2.2.6. Bituminous Material Cements

Blown bitumen is the binder in some cements. Quartz, another inert material, or carbon is used as the filler. The cement is delivered as a powder or in blocks and is heated to 200 – 220 °C before it is applied hot by means of a trowel. The material hardens as it cools to form a bond resistant to temperatures up to 80 °C.

Cements based on bituminous materials are used to produce chemically resistant flooring; to line trenches, pits, and emergency containments; and to repair damaged containers. Bituminous materials are inexpensive and easy to apply. They are resistant to a wide spectrum of chemicals. As a result of high labor costs, however, the use of these materials is declining in Europe and is almost unknown in the United States.

3. Specifications and Testing

The American Society for Testing and Materials (1916 Race Street, Philadelphia, PA 19103, USA) has published a series of standards for testing the properties and physical characteristics of cements (applied in accordance with ASTM C308 and C414 [6]). Standard values are also available in the DECHEMA Richtlinien [7] and in DIN 28062 [8] (Table 1).

Table 1. Tests and Specifications of Physical Properties

Property	US Standard	DIN Standard
Tensile strength	ASTM C307	DIN 53455
Flexural strength	ASTM C580	DIN 53452
Compressive strength	ASTM C579	DIN 51067
Bond strength	ASTM C321	
Thermal expansion	ASTM C531	
Shrinkage	ASTM C531	
Absorption	ASTM C413	DIN 51056

Table 2 presents the materials specifications for chemically resistant cements with quartz or carbon fillers. These specifications are always greatly surpassed by all major products.

Chemical Resistance. Tests of chemical resistance are carried out according to such standards as ASTM C267 and DIN ISO 175.

Table 3 provides information on the resistance of different types of cement to a variety of groups of chemicals [1], [4], [11].

4. Storage and Transportation

Beside phenol – formaldehyde, unsaturated polyester and vinylester resins, the individual components, sealed in their original packing,

Table 2. Properties of chemically resistant cements

	Silicate cements	Sulfur cements	Hyraulic cement mortars	Bituminous material cements	Phenolic resin cements	Furan resin cements	Epoxy resin cements	Unsaturated polyester resin cements	Vinylester resin cements
Binder	Potassium or sodium silicates	Sulfur	Blast-furnace or portland cement	Blown bitumen	Phenol–formaldehyde resins	Furan resins	Epoxy resins	Unsaturated polyester resins	Vinylester resins
Filler	Quartz or other inert material	Quartz or carbon	Quartz	Quartz, carbon, kaolin, baryte	Quartz, carbon, baryte	Quartz, carbon, baryte	Quartz, carbon, baryte	Quartz, carbon, baryte	Quartz, carbon, baryte
Hardener	Neutralization agent		Water		Organic acid	Organic acid	Polyamine	Organic peroxide	Organic peroxide
Processing aids								Organic accelerator	Organic accelerator
Hardening reaction	Coagulation	Solidification	Hydration	Solidification	Polycondensation	Polycondensation	Polyaddition	Polymerization	Polymerization
Pot life (at 20 °C)	$\geq$ 0.5 h $\leq$ 2h		0.5 to several hour		$\geq$ 0.5 h $\leq$ 1 h	$\geq$ 0.5 h $\leq$ 1 h	$\geq$ 0.5 h $\leq$ 1 h	0.5 h	0.5 h
Fit for exposure to stress	Several days	After cooling	Several days	After cooling	2 to 7 days	2 to 7 days	2 to 7 days	1 to 7 days	1 to 7 days
Adhesion to:									
Carbon	not to apply	O	not to apply	+	++	++	++	+	+
Ceramic	+	O	+	+	++	++	++	+	+
Steel	+	O	+	+	(+) [b)]	(+) [b)]	++	+	+
Concrete	+ [a)]	O	+	+	(+) [b)]	(+) [b)]	++	+	+
Rubber	–	O	–/+ [a)]	+	+	+	++	+ [a)]	+ [a)]
Thermoplasts	–/+ [a)]	O	–/+ [a)]	+	– to +	– to +	– to +	– to +	– to +
Resin coatings	–/+ [a)]	O	–/+ [a)]		O to ++	O to ++	O to ++	O to ++	O to ++
Used for:									
Flooring and walls	O	–	++	+	++	++	++	++	++
Tanks and apparatus	++	++	+	+	++	++	+	++	++
Chemical loads	++ (pH $<$ 5)	+	O (pH $>$ 4)	+	++	++	+	+	+
Mechanical loads	+	O	+	O	++	++	++	++	++
Thermal loads	++	O	++	–	+	++	O	+	+

[a)] Primed and sanded off; [b)] In combination with membrane; ++ Very good and/or very suitable; + Good and/or suitable; O Conditional suitable; – Insufficient and/or unsuitable

Table 3. Resistance of cements to various chemicals

Chemical	Silicate cement	Sulfur cement	Hydraulic cement mortar	Bituminous material cement	Phenolic resin cement	Furan resin cement	Epoxy resin cement	Unsaturated polyester resin cement	Vinylester resin cement
Non-oxidizing acids	+	+	–	+	+	+	O	+	+
Oxidizing acids	+	+	–	O	O	O	O	+	+
Silica-dissolving acids	–	+	–	+	+	+	+	+	+
Bases	–	+	O	+	O	+	+	O	+
Oxidizing bases	–	–	O	O	–	O	O	+	+
Salts	+	+/–[a)]	O	+	+	+	+	+	+
Water	–	+	+	+	+	+	+	+	+
Organic acids	+	O	–	O	+	+	O	+	+
Aliphatic compounds	+		O	–	+	+	+	+	+
Aromatic compounds	+		O	–	+	+	O	–	+
Alcohols	+	+	+	O	+	+	O	+	+
Ketones, esters	+	–	–	–	+	+	O	–	O
Aliphatic chlorinated hydrocarbons	+	–	O	–	+	+	–	–	O
Aromatic chlorinated hydrocarbons	+	–	O	–	+	+	–	–	O
Aldehydes	+	–	O	O	+	+	O	O	O
Aliphatic amines	+		+	–	+	+	–	O	+
Aromatic amines	+		+	–	+	+	–	–	O
Phenoles	O	–	–	–	+	+	–	–	O
Fats and oils	+	–	O	–	+	+	O	+	+

[a)] resistant with pH ≤ 7; + resistant; – not resistant; O resistant under special conditions

can generally be stored for 12 months or longer at room temperature and under dry conditions.

Special care must be taken in the transportation of the binders because of their reactivity and toxicity.

5. Toxicology

The national guidelines for the handling of the binder must be carefully observed. Indeed, the binding agents can be toxic, irritating, and detrimental to health. Some powders can also be physiologically active because of the hardener they contain.

6. References

1. F. K. Falcke: *Kleines Handbuch des Säureschutzbaues,* Verlag Chemie, Weinheim 1966.
2. W. L. Sheppard: *Handbook of Chemically Resistant Masonry,* C.C.R.M. Inc., Havertown, Pennsylvania, 1977.
3. F. K. Falcke, G. Lorentz (eds.): *Handbook of Acid-Proof Construction,* VCH Verlagsgesellschaft, Weinheim 1985.
4. E. Schacht: "Ausmauerungen in chemisch beanspruchten Behältern und Apparaten," *Z. Werkstofftech.* **5** (1974) 297 – 307.
5. J. Dück: "Schwund- und Quellverhalten reaktionshärtender Kunstharzkitte für den Säureschutzbau (Shrinking and Swelling Properties of Chemically Cured Resin Mortars for Chemical Resistant Linings)," *Z. Werkstofftech.* **12** (1981) 73 – 83.
6. *ASTM-Standards* 1977, Annual Book of ASTM Standards, vol. 04.05.1997. Chemical Resistant Materials; Vitrified Clay; Fiber-Cement Products; Mortars; Masonry, American Society for Testing and Materials, 1916 Race St., Philadelphia, Pa. 19103, USA.

7. *DECHEMA Richtlinie: Bestimmung physikalischer, insbesondere mechanischer Kennwerte von Kitten für den Säureschutzbau,* Dechema, Frankfurt.
8. DIN 28062, Chemische Apparate; Bau-und Werkstoffe für Ausmauerungen; Einteilung – Eigenschaften – Prüfung; 1978.
9. DIN 28052–5, Chemische Apparate; Oberflächenschutz mit nichtmetallischen Werkstoffen für Bauteile aus Beton in verfahrenstechnischen Anlagen; Kombinierte Beläge, 1997.
10. AGI Arbeitsblatt S10 Part 3, Schutz von Baukonstruktionen mit Plattenbelägen gegen chemische Angriffe–Plattenlagen, part 3, Vincentz Verlag, Hannover.
11. W. A. Kuenning: "Guide for the Protection of Concrete against Chemical Attack by Means of Coatings and other Corrosion-Resistant Materials," *Proc. Amer. Concr. Inst.* **63** (1966) 1305 – 1391.

Ceramic Colorants

Richard A. Eppler, Eppler Associates, Cheshire, Connecticut 06410, United States

1. **Introduction** 1
2. **Classification** 1
3. **Pigment Structures** 3
4. **Colorant Systems** 3
5. **Industrial Production of Ceramic Pigments** 10
6. **Use of Ceramic Pigments in Glazes** . . 11
7. **Ceramic Glazes** 11
8. **Application Media** 12
9. **Quality Control** 12
10. **References** 12

1. Introduction

Color is an important characteristic of many ceramic products. In fact, any product for which aesthetics is a consideration for purchase will be enhanced by proper use of color. Thus, ceramic products which use color include dinnerware, tile, porcelain enamel, sanitary ware, and some glasses and structural clay products.

Ceramic material can be colored in three general ways [1]. First, the ceramic material itself may contain transition metal ions that are colored. This method is rarely used, except for colored glass, because adequate tinting strength and purity of color cannot be obtained this way. A second method to obtain color in ceramics is to induce precipitation of a suitable crystalline phase during processing. For example, certain oxides, such as zirconium oxide and titanium dioxide, dissolve in vitreous material at high temperature. When the temperature is reduced, the solubility is also reduced and precipitation occurs. This method is used for opacification, that is, production of an opaque white color. For oxide colors other than white, this method lacks the control necessary for reproducible results. It is, however, often used for production of nonoxide colorants, such as gold, copper, and cadmium–selenium crystals, dispersed in a vitreous matrix.

The third and most common method of obtaining color in a ceramic material is to disperse within that material a colored crystalline phase that is insoluble in the matrix. This crystalline phase, commonly called a pigment, imparts its color to the matrix.

The materials chosen for this purpose must possess certain properties beyond the tinctorial strength required of any pigment [2]. To be used in a ceramic material, pigments must be resistant to the high temperature and corrosive environment encountered in the firing process. Their rates of dissolution or reaction in the ceramic or molten glaze at high temperature must be low. This problem is aggravated by the fine particle sizes ($1-10\,\mu m$) that are required for uniform dispersion of the pigment in the ceramic matrix.

2. Classification

Most of the materials used in ceramics as pigments are oxides because of their greater stability in oxygen-containing ceramic systems [1]. The only important exceptions are the cadmium sulfoselenide pigments. These nonoxide-containing materials are used because a bright red color cannot be obtained any other way. However, these pigments are very difficult to use in ceramics. Great care is required to prevent their being oxidized during firing of the ceramic body.

The various oxide pigments are classified according to their crystallographic structure (see Table 1) [3]. These structures vary significantly in the range of concentration of constituent oxides allowed. A wide range of cation substitution may be permitted if the ionic sizes of the substituted ions are similar. Different formulations are used to vary the color obtained. Thus, some of the crystal classes have individual pigments with various primary colors. This effect is particularly evident in the spinel class, wherein are found blue, green, pink, brown, and black pigments.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a05_545

Table 1. Classification of mixed-metal oxide inorganic pigments∗

Crystal class, name (i.e., category)	CAS registry number	Basic chemical formula	DCMA number
Baddeleyite			
– Zirconium vanadium yellow baddeleyite	[*68187-01-9*]	$(Zr, V)O_2$	1-01-4
Borate			
– Cobalt magnesium red-blue borate	[*68608-93-5*]	$(Co, Mg)_2B_2O_5$	2-02-1
Corundum-hematite			
– Chrome alumina pink corundum	[*68187-27-9*]	$(Al, Cr)_2O_3$	3-03-5
– Manganese alumina pink corundum	[*68186-99-2*]	$(Al, Mn)_2O_3$	3-04-5
– Chromium green-black hematite	[*68909-79-5*]	Cr_2O_3	3-05-3
– Iron brown hematite	[*68187-35-9*]	Fe_2O_3	3-06-7
Garnet			
– Victoria green garnet	[*68553-01-5*]	$3\,CaO \cdot Cr_2O_3 \cdot 3\,SiO_2$	4-07-3
Olivine			
– Cobalt silicate blue olivine	[*68187-40-6*]	Co_2SiO_4	5-08-2
– Nickel silicate green olivine	[*68515-84-4*]	Ni_2SiO_4	5-45-3
Periclase			
– Cobalt nickel gray periclase	[*68186-89-0*]	(Co, Ni)O	6-09-8
Phenacite			
– Cobalt zinc silicate blue phenacite	[*68412-74-8*]	$(Co, Zn)_2SiO_4$	7-10-2
Phosphate			
– Cobalt violet phosphate	[*13455-36-2*]	$Co_3(PO_4)_2$	8-11-1
– Cobalt lithium violet phosphate	[*68610-13-9*]	$CoLiPO_4$	8-12-1
Priderite			
– Nickel barium titanium primrose priderite	[*68610-24-2*]	$2\,NiO \cdot 3\,BaO \cdot 17\,TiO_2$	9-13-4
Pyrochlore			
– Lead antimonate yellow pyrochlore	[*68187-20-2*]	$Pb_2Sb_2O_7$	10-14-4
Rutile-cassiterite			
– Nickel antimony titanium yellow rutile	[*71077-18-4*]	$(Ti, Ni, Sb)O_2$	11-15-4
– Nickel niobium titanium yellow rutile	[*68611-43-8*]	$(Ti, Ni, Nb)O_2$	11-16-4
– Chrome antimony titanium buff rutile	[*68186-90-3*]	$(Ti, Cr, Sb)O_2$	11-17-6
– Chrome niobium titanium buff rutile	[*68611-42-7*]	$(Ti, Cr, Nb)O_2$	11-18-6
– Chrome tungsten titanium buff rutile	[*68186-92-5*]	$(Ti, Cr, W)O_2$	11-19-6
– Manganese antimony titanium buff rutile	[*68412-38-4*]	$(Ti, Mn, Sb)O_2$	11-20-6
– Titanium vanadium antimony gray rutile	[*68187-00-8*]	$(Ti, V, Sb)O_2$	11-21-8
– Tin vanadium yellow cassiterite	[*68186-93-6*]	$(Sn, V)O_2$	11-22-4
– Chrome tin orchid cassiterite	[*68187-53-1*]	$(Sn, Cr)O_2$	11-23-5
– Tin antimony gray cassiterite	[*68187-54-2*]	$(Sn, Sb)O_2$	11-24-8
– Manganese chrome antimony titanium brown	[*69991-68-0*]	$(Ti, Mn, Cr, Sb)O_2$	11-46-7
– Manganese niobium titanium brown rutile	[*70248-09-8*]	$(Ti, Mn, Nb)O_2$	11-47-7
Sphene			
– Chrome tin pink sphene	[*68187-12-2*]	$CaO \cdot SnO_2 \cdot SiO_2 \cdot Cr_2O_3$	12-25-5
Spinel			
– Cobalt aluminate blue spinel	[*68186-86-7*]	$CoAl_2O_4$	13-26-2

Color variations within a given structural class are of practical importance. Normally, pigments within a given class have excellent chemical and physical compatibility. Thus, it is possible to mix them to obtain intermediate shades.

3. Pigment Structures

The structures of inorganic solid-state materials are governed by several principles [4]. In the first place, the nearest-neighbor cation–anion distances almost totally determine the lattice en-

Table 1. (Continued)

Crystal class, name (i.e., category)	CAS registry number	Basic chemical formula	DCMA number
– Cobalt tin blue-gray spinel	[*68187-05-3*]	Co_2SnO_2	13-27-2
– Cobalt zinc aluminate blue spinel	[*68186-87-8*]	(Co, Zn) Al_2O_4	13-28-2
– Cobalt chromite blue-green spinel	[*68187-11-1*]	Co(Al, Cr)O_4	13-29-2
– Cobalt chromite green spinel	[*68187-49-5*]	$CoCr_2O_4$	13-30-3
– Cobalt titanate green spinel	[*68186-85-6*]	Co_2TiO_4	13-31-3
– Chrome alumina pink spinel	[*68201-65-0*]	Zn(Al, Cr)$_2O_4$	13-32-5
– Iron chromite brown spinel	[*68187-09-7*]	Fe(Fe, Cr)$_2O_4$	13-33-7
– Iron titanium brown spinel	[*68187-02-0*]	Fe_2TiO_4	13-34-7
– Nickel ferrite brown spinel	[*68187-10-0*]	$NiFe_2O_4$	13-35-7
– Zinc ferrite brown spinel	[*68187-51-9*]	(Zn, Fe)Fe_2O_4	13-36-7
– Zinc iron chromite brown spinel	[*68186-88-9*]	(Zn, Fe)(Fe, Cr)$_2O_4$	13-37-7
– Copper chromite black spinel	[*68186-91-4*]	$CuCr_2O_4$	13-38-9
– Iron cobalt black spinel	[*68187-50-8*]	(Fe, Co)Fe_2O_4	13-39-9
– Iron cobalt chromite black spinel	[*68186-97-0*]	(Co, Fe)(Fe, Cr)$_2O_4$	13-40-9
– Manganese ferrite black spinel	[*68186-94-7*]	(Fe, Mn)(Fe, Mn)$_2O_4$	13-41-9
– Chrome iron manganese brown spinel	[*68555-06-6*]	(Fe, Mn)(Fe, Cr, Mn)$_2O_4$	13-48-7
– Cobalt tin alumina blue spinel	[*71750-83-9*]	$CoAl_2O_4$/Co_2SnO_4	13-49-2
– Chromium iron nickel black spinel	[*71631-15-7*]	(Ni, Fe)(Cr, Fe)$_2O_4$	13-50-7
– Chromium manganese zinc brown spinel	[*71750-83-9*]	(Zn, Mn)Cr_2O_4	13-51-7
Zircon			
– Zirconium vanadium blue zircon	[*68186-95-8*]	(Zr, V)SiO_4	14-42-2
– Zirconium praseodymium yellow zircon	[*68187-15-5*]	(Zr, Pr)SiO_4	14-43-4
– Zirconium iron pink zircon	[*68187-13-3*]	(Zr, Fe)SiO_4	14-44-5

∗ Reprinted with permission [3].

ergy of an ionic solid material at room temperature. Moreover, the preferred coordination polyhedra of the anions surrounding each cation in ionic solid phases are determined almost solely by the ratio of the ionic size of the cation to that of the anion. Furthermore, the structures that can be built from any combination of cations and anions are subject to the rules of electrostatic neutrality. Thus, in a stable ionic structure, the valence of each anion, with changed sign, is exactly or nearly exactly equal to the sum of the strength of the electrostatic bonds to it from the adjacent cations. Lastly, the coordination of a cation is larger when its charge is less, and is smaller as its field strength increases. The result of these principles is that for any given stoichiometry and ionic size of the materials to be used, only one or two structures exist that will accommodate them.

For example, the spinel class, which contains 19 pigments, is restricted to those materials of A_2BX_4 stoichiometry with:

$0.06\,\text{nm} < r_A < 0.100\,\text{nm}$

$0.055\,\text{nm} < r_B < 0.100\,\text{nm}$

where r_A and r_B are the ionic radii of the respective cations, and X is oxygen. Similar restrictions apply to the other structures found in Table 1.

4. Colorant Systems

The pigments given in Table 1 are the alternatives available for coloring ceramics. Their specific applications are described in this chapter.

Opacifiers. Whiteness or opacity is introduced into transparent ceramic materials, such as glazes, by the addition of substances that will disperse in those materials as discrete particles. These particles scatter and reflect some of the incident light. To do this, the dispersed substance must have a *refractive index* that differs appreciably from that of the clear ceramic material. The refractive index (n_D^{20}) of most ceramic materials is 1.5 – 1.6 and, therefore, the refractive indexes of opacifiers must be either greater or less than this. In practice, opacifiers of higher refractive index are used. Some examples of opacifiers are tin(II) oxide ($n_D^{20} = 2.04$), zirconia ($n_D^{20} = 2.40$), zircon ($n_D^{20} = 1.85$), and titania ($n_D^{20} = 2.5$ for anatase or 2.7 for rutile).

In glazes and other ceramic coatings fired at temperatures in excess of 1000 °C, *zircon* [*1490-68-2*] is the opacifier of choice [5]. Its solubilityin many ceramic glazes is about 5 %

at high temperature and 2 – 3 % at room temperature. A customary mill addition would be 8 – 10 % zircon. Consequently, most opacified glazes contain both zircon that was placed in the mill and went through the firing process unchanged and zircon which dissolved in the molten glaze during firing, but which recrystallized on cooling.

The effectiveness of a zircon opacifier is a function of particle size. The finer the particle size, the greater is the opacity of the pigment. Because this greater fineness is achieved by milling, the finer zircons are also the most expensive. On the other hand, zircon for smelting into frit is best when of an intermediate size. Therefore, the effectiveness of zircon opacifiers can be improved in partially or fully fritted glazes by smelting some of the zircon into the frit.

Zirconia is rarely used as an opacifier because in the vast majority of ceramic glazes it reacts with the silica in the glaze to produce zircon [6]. Therefore, because zircon is less expensive than zirconia, zircon is preferred. Tin oxide is a more effective opacifying agent than any of the other possibilities because its solubility is lower. However, the price of tin oxide is so high that it is no longer an economic solution. Its use is restricted to those special cases, such as chrome – tin pinks, where it also enhances the effectiveness of the coloring pigment.

In porcelain enamels and in glazes where the firing temperature is less than 1000 °C, *titania* [*13463-67-7*]in the anatase crystal phase is the opacifying agent normally used [7], [8]. It is the most effective opacifying agent because it has the highest index of refraction. However, at ca. 850 °C, anatase inverts to rutile in ceramic systems. Once inverted to rutile, titania crystals are able to grow rapidly to sizes that are no longer effective for opacification. Moreover, as the rutile particles grow, their absorption band extends into visible wavelengths, leading to a pronounced cream color. Thus, while titania is an effective opacifier at lower temperatures, it cannot be used above 1000 °C.

The solubility of titania in molten silicates is ca. 8 – 10 % in most cases, and at room temperature is reduced to ca. 5 %. Thus, titanium dioxide is customarily used at concentrations of ca. 15 %.

When a *pastel color* is required, an opacifier plus a pigment is added to the coating. If this is done, the overall coating–pigment–opacifier system must be considered in selecting materials. The opacifier and the pigment must be compatible. For example, zircon opacifiers should be used with zircon or zirconia pigments. Chrome–tin pinks are stronger if some tin oxide opacifier is used. Titanium-based pigments are used in enamels when titania is the opacifier.

Black Pigments. Black ceramic pigments are formed by calcination of several oxides to develop the *spinel structure* [9], [10]. The formulation of black pigments is an excellent illustration of the wide flexibility of this structure for incorporating various chemical elements. Table 1 lists five different black spinel pigments: copper chromite black spinel, iron cobalt black spinel, iron cobalt chromite black spinel, manganese ferrite black spinel, and chromium iron nickel black spinel.

Selection of a particular black pigment depends somewhat on the specific material with which the pigment is to be used. If care is not taken, the pigment may show a green, blue, or brown tint after firing. A particularly important problem is the tendency of some ceramics to attack the pigment and release any cobalt that may be present. Thus, in some cases, using a cobalt-free pigment is desirable. The relatively high price of cobalt oxide also encourages use of cobalt-free pigments.

One black pigment that is not a spinel is *chromium black hematite.* This pigment is limited to zinc-free systems, as it reacts with zinc oxide to yield a brown spinel.

The basic black pigment is *iron cobalt chromite black spinel.* In some systems, however, it has a slightly greenish tint. In zinc-containing bases, therefore, iron cobalt black spinel is recommended. For a black with a slightly bluish tint, iron cobalt chromite black spinel with some manganese and a higher concentration of cobalt is used, and for a black with a brownish tint, manganese ferrite black spinel is the pigment of choice. In those instances where a cobalt-free system is desirable, copper chromite black spinel can be considered for use in systems fired at < 1000 °C, such as glass colors or porcelain enamels. Chromium black hematite can be used in zinc-free systems. This is the least expensive black on the market for use in materials fired at > 1000 °C. However, the presence of zinc ox-

ide in the material results in a chemical reaction that alters the color. For systems containing zinc and fired at > 1000 °C, chrome iron nickel black spinel can be considered. This pigment can be used with most glaze systems and at all firing temperatures from cone 06 up to sanitary ware firing temperatures of cone 11 or 12 (for cone temperature equivalents, see Table 2).

Table 2. Cone temperature equivalents

Cone number	Orton standard pyrometric cones ∗, °C		Seger cones, °C
	Large cones, 150 °C∗∗	Small cones, 300 °C∗∗	(used in Europe)
010			
	894	919	900
09			
	923	955	920
08			
	955	983	940
07		1 008	
	984		960
06		1 023	
	999		980
05	1 046	1 062	1 000
04	1 060	1 098	1 020
03	1 101	1 131	1 040
02	1 120	1 148	1 060
01	1 137	1 178	1 080
1	1 154	1 179	1 100
2	1 162	1 179	1 120
3	1 168	1 196	1 140
4	1 186	1 209	1 160
5	1 196	1 221	1 180
6	1 222	1 255	1 200
7	1 240	12 64	1 230
8	1 263	1 300	1 250
9	1 280	1 317	1 280
10	1 305	1 330	1 300
11	1 315	1 336	1 320
12	1 326	1 335	1 350

∗ From the Edward Orton, Jr., Ceramic Foundation, Columbus, Ohio.
∗∗ Temperature rise per hour.

Gray Pigments. The simplest way to obtain a gray pigment is to dilute a black pigment with a white opacifier. This *dilution* must be done with great care to provide an even color, without specking. Therefore, in most cases, use of a compound that has been formulated to give a gray color is preferred [10].

More uniform results are obtained when a *calcined pigment,* such as cobalt nickel gray periclase, is used. For certain special effects in underglaze decorations, a beautiful deep gray can be prepared by using tin antimony gray cassiterite. The limitation on the use of this material is the high cost of the tin oxide base material.

An important point to note with gray is the many subtle shade variations that are possible. With appropriate blending of three or four carefully chosen pigments, many different shades are possible. On the other hand, uniformity of color in this area requires careful quality control.

Blue Pigments. The traditional way to obtain blue in a ceramic material is with *cobalt,* which has been used as a solution color since antiquity [11]. Today, cobalt is reacted with aluminum oxide to produce the spinel $CoAl_2O_4$ or with silica to produce the olivine Co_2SiO_4. Some formulations are mixtures of these two materials. Cobalt silicate involves the use of a higher percentage of cobalt oxide than does aluminate spinel. However, the color is only modestly more intense.

In the spinel system, the shade can be adjusted toward turquoise or green by additions of chromium oxide replacing alumina and zinc oxide replacing cobalt (see section on Green Pigments).

At the lower temperatures encountered in porcelain enamels and glass colors, pigments based on cobalt continue to be fully satisfactory both for stability and for tinting strength, which is quite high. At the higher temperatures encountered with ceramic glazes, however, difficulties arise from partial dissolution of the pigment. The cobalt oxide diffuses into the glaze, giving a defect commonly called *cobalt bleeding.* Thus, in glazes, cobalt pigments have been largely replaced by pigments based on vanadium-doped zircon [2], [12], [13]. These pigments are less intense than cobalt pigments and tend toward turquoise. Therefore, they are not applicable in all cases. However, when they are applicable, they give vastly improved stability.

The *zircon–vanadium blue pigment* is made by calcining a mixture (in the stoichiometry of zircon) of zirconia, silica, and ammonium metavanadate in the presence of a mineralizer [2]. The latter materials, which are selected from various halides and silicohalides, facilitate transport of silica during the reaction forming the pigment. Although there is extensive literature on this subject making many claims with respect to composition, the fact is that for development of a strong blue color, the stoichiometry of zircon must be

retained and such mineralizers used as will simultaneously optimize the various transport processes and incorporate the optimum amount of vanadium into the zircon structure when it is formed.

With these pigments, use of zircon for opacification is generally desirable. In addition, at least some zirconium oxide in the glaze is preferred to stabilize the pigment.

Green Pigments. Five of the more important methods to obtain green pigmentation in a ceramic material are discussed in this section [14].

Historically, the basis of most green pigmentation was the chromium ion. Although *chromium oxide* itself may be used to produce a green color, this procedure has a number of limitations. First, pure chromium oxide has some tendency to fume or volatilize during the firing of the ceramic coating, which leads to absorption into the refractory of the furnace used. Second, if tin-containing white pigments or pastel colors containing tin are also in the furnace, the chrome will react with the tin to form a pink coloration. Finally, the ceramic material into which chromium oxide is placed must meet particular requirements. It must not contain zinc oxide, which produces an undesirable dirty brown color. As already mentioned, no tin oxide may be used as opacifier or as a constituent of the glaze.

More satisfactory results are obtained if chromium oxide is used as a constituent in a calcined ceramic pigment. One such system is *cobalt chromite blue-green spinel.* In these spinels, varying amounts of cobalt and zinc appear in tetrahedral sites and varying amounts of alumina and chromium oxide appear in octahedral sites. Greener pigments are obtained by using a higher concentration of chromium oxide and a lower concentration of cobalt oxide. Conversely, shades from blue-green to blue result from lowering the amount of chromium oxide and raising the amount of cobalt oxide. These pigments should not be used in low concentration because they give an undesirable dirty gray color.

The final type of chromium oxide containing green is *Victoria green garnet.* This material is prepared by calcining silica and a dichromate (sodium or potassium) with calcium carbonate to form the garnet $3\,CaO \cdot Cr_2O_3 \cdot 3\,SiO_2$. This pigment gives a beautiful bright green color but is transparent. When the color is applied thinly, it has a tendency to blacken. Victoria green garnet is not satisfactory for opaque glazes or pastel shades because the tone always has a gray cast and lacks brilliance. It can be used only in zinc-free coatings with high calcium content. In the presence of zinc, the stability of the garnet structure is inadequate. In addition, because this is a difficult pigment to manufacture correctly, the price is high, reflecting the care required.

Because of all of the difficulties mentioned in the use of chromium-containing pigments, and also because there is a definite limitation on the brilliance of green pigments made with chromium, most ceramic glazes use pigments in the *zircon system* [2]. Originally, pigments in the zirconia–vanadia–silica system were recommended. However, because these pigments are, in fact, in-place mixtures of a zircon–vanadium blue pigment and a zirconia–vanadium yellow pigment, superior products can be obtained by preparing the pigments separately, using the optimum preparative conditions for each one. Moreover, because the zircon–praseodymium yellows are the strongest in their color family, their use as the yellow constituent of a green blend gives even better results. Therefore, the cleanest, brightest, most stable greens are obtained by blending a zircon–vanadium blue and a zircon–praseodymium yellow. The bright green shades are obtained from a mixture of ca. two parts of the yellow pigment to each part of the blue pigment.

Finally, *copper compounds* are used in certain low-temperature firing applications [14]. The use of copper is of little interest to the majority of industrial manufacturers, but the colors obtained from it are of great interest to art potters because of the many subtle shades that can be obtained. This variety arises because the pH of the glaze affects the color obtained from copper. If the glaze is alkaline, a turquoise blue color results; if the glaze is acidic, a beautiful green color develops. Copper oxide dissolves in the glaze composition, and is, therefore, a transparent color. Because copper oxide volatilizes quite readily, it should not be used above 1000 °C.

Another limitation on the use of copper colors is the fact that copper oxide renders many lead-containing glazes unsafe for contact with food or

drink. Therefore, copper pigments should never be used on such articles.

Yellow Pigments. Although a number of systems exist for preparing yellow ceramic colors, there are technical and economic reasons for the use of a particular yellow pigment. The pigments of greatest tinting strength, the lead antimonate yellows and the chrome–titania maples, do not have adequate resistance to molten ceramic coatings. Therefore, other systems must be used if the firing temperature exceeds ca. 1000 °C. Three of these higher temperature systems are considered in this section.

Zirconia–vanadium yellows are prepared by calcining zirconium oxide with small amounts of ammonium metavanadate [13], [15]. Titanium dioxide or iron oxide may be used to alter the shade. In the absence of these latter materials, lemon-yellow is obtained; in their presence, orange-yellow results. In ceramic coatings, zirconia–vanadium yellows are usually weaker than tin–vanadium yellows and muddier than praseodymium–zircon yellows. However, they are economical stains for use with either zinc-containing or zinc-free coatings. They are stronger and brighter in low-lead, low-boron glazes. Zirconium silicate is the preferred opacifier.

Tin–vanadium yellows are prepared by introducing small amounts of a vanadium oxide into the cassiterite structure of tin oxide [16]. The shade may be varied by addition of titanium dioxide or iron oxide. In the absence of these materials, a lemon-yellow shade is obtained. A stronger yellow may be made by adding titanium dioxide to the color batch, and the increased strength of this modified yellow is accompanied by an increase in the apparent redness of the pigment when iron oxide is added.

Tin–vanadium yellow pigments develop a yellow color in all ceramic materials, although the actual shade may be influenced by the nature of the substrate material. These are opaque pigments, which need minimum amounts of opacifier. However, these pigments are sensitive to reducing conditions. Moreover, any blends with chrome-bearing pigments should be avoided. The reason is that tin oxide and chrome oxide combine easily to form a compound with a color similar to that of chrome–tin pink, which shows up in the ceramic material as a brown discoloration. Finally, grinding the tin–vanadium yellow pigment should be minimized because it tends to weaken the pigment.

The primary deterrent to the use of tin–vanadium yellows, however, is not any technical deficiency. Rather, it is the high cost of the tin oxide that is the major component. The result of this high cost, together with the quality of the praseodymium–zircon pigments, has been a decline in the use of tin–vanadium yellows.

Praseodymium–zircon pigments are formed by calcination of ca. 5 % praseodymium oxide with a stoichiometric mixture of zirconium oxide and silica in the presence of mineralizers to yield a bright yellow pigment [2], [16], [17]. This pigment is quite analogous to zircon–vanadium blue pigments in that the crystal structure is that of zircon. Praseodymium–zircon pigments have excellent tinting strength in high-temperature coatings. They can be used in almost any ceramic coating, although preferably with zircon opacifiers. They blend well with other pigments, particularly with other zircon and zirconia pigments. These pigments are being increasingly used for all applications in which the firing temperature exceeds 1000 °C.

For lower temperature applications, the tinting strength of the *lead antimonate pigments* is unsurpassed, except by cadmium sulfoselenides [18]. Lead antimonate pigments, which have traditionally been called Naples yellow, are exceptionally clean and bright and have good covering power, requiring little or no opacifier.

The primary limitation is their instability in ceramic coatings above ca. 1000 °C, which leads to volatilization of the antimony oxide. Substitutions of cerium oxide, alumina, or tin oxide are sometimes made for a portion of the antimony oxide to improve its stability. Thus, although these materials have limited usage in ceramic glazes, they are the pigment of choice in porcelain enamel.

For the brightest, low-temperature applications, *cadmium sulfoselenide yellow* can be considered [19]. The pure cadmium sulfoselenide colors are produced in a range from primrose yellow through yellow to orange and red. Cadmium sulfide itself is yellow to orange, depending on details of its manufacture and the ratio of the alpha to beta forms of the crystal. The primrose yellow and light yellow shades are made

by precipitating small amounts of zinc sulfide along with the cadmium sulfide.

One final orange-yellow pigment needs to be considered. This is the pigment formed when chromium oxide is added with antimony oxide to titanium dioxide to form a *doped rutile* [18]. This material gives an orange-yellow or maple shade, and is useful in lower melting ceramic coatings. Like the lead antimonate yellows, it begins to decompose at ca. 1000 °C. Although it is of limited use in high-firing ceramics, it is one of the largest volume pigments used in porcelain enamel, where it forms the basis for some of the high-volume appliance colors.

Brown Pigments. By far the most important brown pigment used in ceramics is *zinc iron chromite brown spinel* [20]. This family produces a wide palette of tan and brown shades and can be controlled with reasonable care to produce uniformity within the production variables existing in commercial plants. Within the spinel structure, the zinc oxide is found on the tetrahedral sites and the chromium oxide on the octahedral sites. The iron oxide is distributed in such a way as to fulfill the requirements of the structure. Consequently, adjustment of the formula does result in alteration of the shade. For example, a substantial increase in chrome and decrease in zinc results in greener shades in zinc-free coatings and yellow to gray shades in zinc-containing coatings. Minor addition of manganese to this system results in yellowish and grayish shades, whereas addition of minor amounts of nickel oxide results in a much darker brown. Because they are comparatively inexpensive, these pigments are the brown selected for most applications. However, two systems closely related to the zinc iron chromite brown spinel have been developed to improve the firing range and stability of brown pigments.

The first of these is the *addition of alumina to the zinc iron chromite brown spinel.* This creates a pigment that is a hybrid of the zinc iron chromite brown spinel and the chrome alumina pink spinel. It produces warm, orange-brown shades with improved firing stability. This pigment is used in coatings that are high in zinc and alumina and low in calcium oxide. The alumina–to–zinc ratio is kept as high as practical to improve the brightness and cleanliness of the pigment.

Another related pigment is a *tin-containing iron chromite brown spinel,* which is sometimes called a tin tan. As produced, this material is a mixture of chromium oxide, tin oxide, and iron aluminate. It is always used in a zinc-containing coating to obtain optimum brown shades. Most likely, this is because the pigment reacts with zinc from the coating during the firing process to produce a zinc iron chromite brown spinel pigment. In coatings that are free of zinc, this pigment produces shades of gray to dark mahogany. The pigment has excellent stability at low concentration. Therefore, it makes an excellent toner for some tan and beige shades in blends with various pink pigments.

The final brown pigment to be considered is *chrome iron manganese brown spinel.* Manganese is well-known as the colorant in amethyst-stained glass and, with iron oxide, it has been responsible for the deep brown glazes associated with electrical porcelain insulators, artware, and bean pots. It is used, therefore, where a deep brown shade is needed. However, in producing medium to light shades of brown, the presence of manganese often causes poor surface and unstable color with tendencies to volatilization. Therefore, the use of this pigment is rather restricted.

Pink and Purple Pigments. Only a short step in the color spectrum separates brown and pink. This is reflected in the *chrome alumina pink spinels,* which are similar in crystal structure and behavior to the zinc iron chromite brown pigments except for the absence of iron oxide [21]. Chrome alumina pinks are combinations of zinc oxide, aluminum oxide, and chromium oxide. Depending on the concentration of zinc, the crystal structure may be either spinel (zinc aluminate–chromite) or corundum (solid solution of chromium in aluminum oxide). The latter is analogous to the composition of a ruby.

In general, a ceramic coating formulated for chrome alumina pink spinels should be free of calcium oxide, with low concentrations of lead oxide and boric oxide, and with a surplus of zinc oxide and alumina. Using an improper glaze results in a brown pigment in place of the desired pink. Sufficient zinc oxide must be in the coating to prevent the glaze from attacking the pigment and removing zinc from it. A surplus of alumina

prevents the molten coating from dissolving the pigment.

A related, but somewhat stronger, pink pigment is *manganese alumina pink corundum.* This pigment is formulated by addition of magnesium oxide and phosphate to aluminum oxide. A pure, clean pigment is obtained. The use of a proper formulation, however, is important. A zinc-free system with a high concentration of alumina is required. Unfortunately, the manufacture of this pigment involves serious pollution problems. As a result, several companies have stopped manufacturing it and there is question as to its continued availability.

The most stable pink pigment is the *iron–zircon* system [2], [22], [23]. It is made by calcining a mixture of zirconium oxide, silica, and iron oxide, using a stoichiometry that will produce zircon. This pigment is sensitive to minor variations in the production process, so that one manufacturer's pigment may not duplicate another's [24]. Shades extend from coral to pink. The pigment is stable in all coating formulations, but those without zinc oxide are bluer in shade.

The final pink system, and the only one to produce purple and maroon shades as well as pinks, is *chrome–tin pink.* These are pigments produced by calcining mixtures of small amounts of chromium oxide with substantial amounts of tin oxide. In addition, most such materials have large quantities of silica and calcium oxide in the formulation. The chemistry of these materials is complex and only recently has their chemical composition been determined [25]. Mixing ca. 90 % tin oxide with small amounts of chromium oxide and either calcium oxide or cerium oxide, together with boric oxide as a mineralizer, gives chrome tin orchid cassiterite. This material is a solid solution of chromium oxide in tin oxide. Although this is not the crystal structure of most chrome–tin pinks, residual amounts are present in almost all cases. It is this residual amount of chromium-doped tin oxide that gives most chrome–tin pinks a somewhat gray or purple overtone.

For most chrome–tin pinks, addition of substantial amounts of calcium oxide and silicon oxide is required to make chrome tin pink sphene. Only in the presence of these materials can pink, red, or maroon shades be obtained. In this case the crystal structure is tin sphene ($CaO \cdot SnO_2 \cdot SiO_2$) in which chromium oxide is dissolved as an impurity.

The color of this pigment depends to a great extent on the ratio of the concentration of chromium oxide to that of tin oxide. Generally speaking, when this ratio is 1 : 5, the resulting color is green; 1 : 15, purple; 1 : 17 – 20, red or maroon; and 1 : 25, pink.

These pigments are calcined at 1260 – 1320 °C and, under the right conditions, are stable at these temperatures. They can be used in coating materials that are low in zinc and high in calcium oxide. Either tin oxide or zirconium silicate opacifiers can be used, but tin oxide as a mill addition improves the strength and stability of the pigment.

Gold purple, commonly called Purple of Cassius, is an old pigment consisting of tin oxide gel colored by finely divided gold. This pigment can be used in low-temperature materials, such as porcelain enamels, where it has good coverage and brilliance. It is, however, an expensive pigment. This is due not only to the high price of gold, but also to the difficult methods of preparing the pigment.

Red Pigments. There are no oxide systems which can be used to produce a true red pigment that is stable in ceramic systems. Therefore, orange, red, and dark red pigments are obtained by the use of *cadmium sulfoselenide pigments* [19], [26]. The specific shade results from varying the ratio of the concentration of cadmium sulfide to that of cadmium selenide. An orange pigment is obtained at a ratio of ca. 4 : 1, a red pigment at 1.7 : 1, and a deeper red at 1.3 : 1.

These pigments are prepared by one of several chemical processes involving wet precipitation of suitable raw materials, such as cadmium carbonate and elemental sulfur or selenium, followed by calcination at 500 – 600 °C under an inert atmosphere.

Cadmium sulfoselenide pigments require the use of a glaze specially designed for this purpose. This glaze contains only small amounts of lead oxide because high-lead flux materials react with selenium in a cadmium sulfoselenide pigment to form lead selenide, which is black. The glaze is a low-alkaline borosilicate type. It contains a few percent of cadmium oxide, which reduces the potential for dissolution of the pigment in the glaze during firing. It is free of vig-

orous oxidizing agents, such as nitrates, which oxidize the pigment, completely destroying the color.

These pigments are temperature sensitive. Therefore, although they can be used in glass colors, in porcelain enamels, and in low-temperature glazes fired up to ca. 1000 °C, they cannot be used in higher temperature applications.

In order to extend the range of these colors, an inclusion pigment system has recently been developed [2], [27]. In this system, cadmium sulfoselenide is incorporated in a clear zircon lattice during manufacture. In this way, the superior stability of zircon is imparted to the pigment. However, these pigments are difficult to make and not all shades have been made successfully. The color palette extends from yellow through orange to red. Dark reds are not yet available.

Precious Metal Compositions. Although bright gold, burnished gold, and the corresponding silver and platinum preparations are not ceramic pigments in the strict sense, they still play a considerable role in the decoration of ceramic materials [28]. The production of these materials consists essentially of the reaction of pinene with sulfur or hydrogen sulfide to give a pinene thiol, which then reacts with tetrachlorauric acid to yield an auric sulforesinate. This gold resinate is dissolved in an organic solvent and then reacted with various additives, in the form of metal organic compounds, to affect the color tone and to achieve the necessary adherence. These materials are applied to the outside surface of the ceramic object and fired at 500 – 850 °C.

Lusters. The pigments that are designed as lusters are closely related to the precious metal compositions because they are also usually prepared and applied as organic compounds of metals. After firing, they precipitate as a thin, often irridescent layer on the substrate. They are not a coherent metallic layer as with precious metal coatings, but rather are an oxide layer. Colored lusters result when compounds of transition metals are included in the formulation. Iron lusters give light brown-red to golden coatings; cobalt lusters in high dilution, a chocolate brown; copper lusters, a reddish brown; nickel lusters, a light brown; and manganese lusters, a gray brown.

5. Industrial Production of Ceramic Pigments

Ceramic pigments are customarily prepared by solid-state reactions. For that reason, rapid, uniform, and reproducible conversion requires intimate mixing of the raw materials, which must be of optimum particle size for the given reaction. In most pigment preparations, additives called mineralizers are included to increase the rate of reaction and make the mixture more uniform. For zircon pigments, these mineralizers are usually alkali and alkaline-earth halides [13], [15]. Boric acid is often used as a mineralizer for spinel pigments. Much of the art of ceramic pigment manufacture is concerned with selection of particular mineralizers for a given reaction.

After careful mixing, the pigments are *calcined* in either batch kilns or continuous calciners. In this operation, careful attention must be paid to the control of the temperature of the kiln. The advantages of batch calcination procedures lie in greater production flexibility and the ability to prepare smaller quantities when required. The continuous tunnel kiln provides greater product quality and greater consistency in calcining conditions. However, its use requires a minimum production level.

After calcination, hard clumps of calcined, sintered products are broken in *jaw or roll crushers* and then ground to the necessary fineness in *mills.* Depending on the particular pigment and the particle size required, either wet or dry ball mills may be used. Wet ball milling yields a finer product, but it is considerably more expensive than dry ball milling.

Some pigments must be *washed* to remove soluble constituents that would otherwise cause difficulty in the final application. Washing is particularly required for pigments containing vanadium oxide. The ground, suspended particles are washed in filter presses or decanters. The water is removed mechanically and the remaining pigment slip or filter cake dried.

The final production step involves careful control of color tone by adjustment with toners. *Toners* are formulations at various extremes of the color spectrum covered by a given pigment family and are used to adjust the color of products to specifications.

6. Use of Ceramic Pigments in Glazes

There are five ways in which ceramic colors can be applied to a glazable ceramic article: as a body stain, as an engobe, as an underglaze color, as a colored glaze, and as an overglaze or glass color. The use of a *body stain* refers to a pigment added to the body formulation itself. The technique of using *engobes* may be described as that of applying a ceramic pigment to a raw body. *Underglaze* decorating is the application of color to a bisque body. In *colored glazes,* the pigment is dispersed in the glaze itself. Finally, *overglazes or glass colors* are applied to the already formed and fired glaze as an overcoat.

The *selection of a technique* depends on the requirements of the particular application. For example, if an engobe or body stain is used, it must be stable to the bisque as well as to the glost fires. An underglaze color or a colored glaze need be stable only to the glost fire. On the other hand, an overglaze or glass color need not be stable to either of these firings.

Moreover, the *range of colors and effects* which can be obtained is directly related to the stability. Some colors, such as bright red, can only be obtained in overglaze decoration or in specially formulated glazes. On the other hand, the *durability* of a decoration or color in service depends largely on the distance between the outer surface of the ware and the pigment. Therefore, overglaze decorations are distinctly inferior to other techniques of application with respect to durability in service.

7. Ceramic Glazes

The ceramic pigment and application method must be compatible with the ceramic glaze that is to be used. Glazes having a wide range of firing temperatures are available, from hobby glazes firing at cone 010 to sanitary ware glazes from cone 7 to cone 12 [29–31].

Low-firing glazes, which mature at cone 010 to 01, are used primarily by hobbyists and artware potters. The reason is that it is only at these temperatures that a full palette of colors can be used in-glaze as well as overglaze. In particular, the cadmium sulfoselenide red pigments can be used only in this firing range and only with a specially designed glaze formulation. In addition, a wide range of special effects can be used at these firing temperatures. These effects include crackle glazes, high-calcium–alumina matte glazes, and crystalline glazes containing materials such as rutile.

The glazes that fire out at cone 01 to cone 6 consist of most of the dinnerware and tile glazes and a minority of high-temperature artware. The color palette at these temperatures is somewhat reduced, but still quite extensive. Spinel blacks, blues, blue-greens, and browns are suitable, as are all the pigments based on zirconia and zircon: zircon–vanadium blues, zircon–praseodymium yellows, zirconia–vanadium yellows, zircon–iron pinks, and zirconia grays. Tin–vanadium yellows are stable as are chrome–tin pinks. Chrome–alumina pigments are stable with proper glaze formulation.

The colors used in glazes up to cone 12 are much the same as those used at cone 01 to cone 6. Tin–vanadium and zirconia–vanadium yellows are quite stable. Chrome–tin pigments are acceptably stable if a suitable formulation is used. The same is true of chrome–alumina pinks. Chrome–iron–zinc browns are satisfactory with suitable formulations. The zircon pigments are all stable to cone 12, although zircon–praseodymium yellow begins to lose strength at higher temperature. Colors containing chromium oxide are stable to cone 12 in a zinc-free glaze.

8. Application Media

For application methods other than in-glaze, the pigment is mixed with an organic medium, which serves only to apply the decorative color to the ceramic substrate. This medium must then burn or evaporate completely in the firing of the ceramic. The materials that are used are divided into two types: hydrophilic media and lyophilic media.

Hydrophilic media are always needed if the decorating color is to be applied to a wet substrate, or if aqueous suspensions are to be applied over the decoration in a subsequent operating step. A typical example would be underglaze colors applied to bisque-fired pieces that are subsequently to be glazed. Hydrophilic preparations can be based on glycerol, ethylene

glycol, poly-(ethylene oxide), or polyglycols as binding agents, and water or other alcohols as solvents.

Lyophilic media are used primarily in overglaze decoration. In one application, they are a component in the decalcomania that are extensively used in overglaze decoration. In terms of volume, however, the largest amount of lyophilic media is used in silk screening. These materials generally consist of a methacrylate-based binding agent, which will depolymerize on firing of the ceramic decoration, and hence, completely evaporate. Pine oil or turpentine usually serves as the solvent.

9. Quality Control

Quality control in the production of ceramic pigments is primarily a matter of controlling the color of the pigment after application [32]. The human eye is an extremely sensitive measuring device for color, particularly on a comparative basis [33]. Thus, control of color solely by electronic procedures is difficult [34]. In many cases, electronic quality control techniques must be backed up with preparation of trial glazes for visual comparison. Moreover, it is usually necessary to adjust the color of each lot of pigment to ensure adequate reproducibility from lot to lot. In addition, for some applications, the particle size of the pigment is critical. If vanadium oxide is used in the formulation, residual free vanadium must be controlled.

10. References

1. A. Burgyan, R. A. Eppler, *Am. Ceram. Soc., Bull.* **62** (1983) 1001 – 1003.
2. R. A. Eppler, *Am. Ceram. Soc., Bull.* **56** (1977) 213 – 216.
3. *DCMA Classification and Chemical Description of the Mixed-Metal Oxide Inorganic Colored Pigments,* 2nd ed., Dry Color Manufacturers' Assoc., Alexandria, Va., 1982.
4. R. A. Eppler, *J. Am. Ceram. Soc.* **66** (1983) 794 – 801.
5. F. T. Booth, G. N. Peel, *Trans. Brit. Ceram. Soc.* **58** (1959) 532 – 564.
6. C. W. F. Jacobs, *J. Am. Ceram. Soc.* **37** (1954) 216 – 220.
7. R. D. Shannon, A. L. Friedberg, *Univ. Ill. Eng. Exp. Sta. Bull.* **456** (1960) 1 – 49.
8. R. A. Eppler, *J. Am. Ceram. Soc.* **52** (1969) 89 – 99.
9. R. A. Eppler, *Am. Ceram. Soc., Bull.* **60** (1981) 562 – 565.
10. W. F. Votava, *Am. Ceram. Soc., Bull.* **40** (1961) 17 – 18.
11. R. K. Mason, *Am. Ceram. Soc., Bull.* **40** (1961) 5 – 6.
12. Harshaw Chemical Co., US 2 441 447, 1948; US 3 025 178, 1962 (C. A. Seabright).
13. C. A. Seabright, H. C. Draker, *Am. Ceram. Soc., Bull.* **40** (1961) 1 – 4.
14. P. Henry, *Am. Ceram. Soc., Bull.* **40** (1961) 9 – 10.
15. F. T. Booth, G. N. Peel, *Trans. Brit. Ceram. Soc.* **61** (1962) 359 – 400.
16. E. H. Ray, T. D. Carnahan, R. M. Sullivan, *Am. Ceram. Soc., Bull.* **40** (1961) 13 – 16.
17. R. A. Eppler, *Ind. Eng. Chem. Prod. Res. Dev.* **10** (1971) 352 – 355.
18. *The Colour Index,* 3rd ed., Soc. Dyers & Colourists, Bradford-London 1971.
19. R. A. Eppler, D. S. Carr: *Proc. 3rd Int. Cadmium Conference,* International Lead Zinc Research Organization (ILZRO), New York 1982.
20. J. E. Marquis, R. E. Carpenter, *Am. Ceram. Soc., Bull.* **40** (1961) 19 – 24.
21. R. L. Hawks, *Am. Ceram. Soc., Bull.* **40** (1961) 7 – 8.
22. Glidden Co., US 3 189 475, 1965 (J. E. Marquis, R. E. Carpenter).
23. Harshaw Chemical Co., US 3 166 430, 1965 (C. A. Seabright).
24. R. A. Eppler, *J. Am. Ceram. Soc.* **53** (1970) 457 – 462.
25. R. A. Eppler, *J. Am. Ceram. Soc.* **59** (1976) 455.
26. Glidden Co., US 2 643 196, 1953; US 2 777 778, 1957 (B. W. Allan, F. O. Rummery). Fabriques de Produits Chimiques de Thann et de Mulhouse, US 3 528 834, US 3 528 835 , 1970 (J. Gascon).
27. Fabriques de Produits Chimiques de Thann et de Mulhouse, US 3 445 199, 1969 (B. H. P. Fehr, J. Gascon). H. D. DeAhna, *Ceram. Eng. Sci. Proc.* **1** (1980) 860 – 862.
28. Du Pont, US 2 924 540, 1960 (J. B. D'Andrea).
29. R. A. Eppler in D. R. Uhlmann, N. J. Kreidl (eds.): *Glass Science and Technology,* **vol. 1,** Academic Press, New York 1983, pp. 301 – 338.

30. C. W. Parmalee, C. G. Harman: *Ceramic Glazes,* 3rd ed., Cahners Publ., Boston, Mass., 1973.
31. F. Singer, W. L. German: *Ceramic Glazes,* Borax Consolidated, London 1964.
32. K. Shaw: *Ceramic Colors and Pottery Decoration,* MacLaren & Sons, London 1962.
33. D. B. Judd, G. Wyszecki: *Color in Business, Science and Industry,* Wiley-Interscience, New York 1963.
34. D. A. Klimas, A. Canonico, *Am. Ceram. Soc., Bull.* **63** (1984) 445.

Ceramics, General Survey

Ceramic Colorants; Ceramics, Advanced Structural Products; Ceramics, Ceramic – Metal Systems; Ceramics, Electronic; Construction Ceramics; Refractory Ceramics; and Whitewares are separate keywords.

GIRARD W. PHELPS, Department of Ceramics, Rutgers – The State University, Piscataway, New Jersey 08854, United States

JOHN B. WACHTMAN, JR., Center for Ceramics Research, Rutgers – The State University, Piscataway, New Jersey 08854, United States

1. Traditional and Advanced Ceramics 2
1.1. Traditional Ceramics 3
1.2. Advanced Ceramics 4
1.2.1. Advanced Structural Ceramics 5
1.2.2. Electronic Ceramics 7
1.2.3. Other Advanced Ceramics 8
1.3. Characterization of Ceramic Materials 9
2. Raw Materials for Traditional Ceramics 10
2.1. The Structure of Clays and Nonplastics 11
2.2. Clay – Water System 12
2.3. Commercial Ceramic Clays 13
2.4. Commercial Nonplastics for Ceramics 16
3. Raw Materials for Advanced Ceramics 19
3.1. Metal Oxides and Carbonates 20
3.2. Borides, Carbides, and Nitrides 20
4. Processing Ceramic Ware 20
4.1. Preparation of Clay-based Forming Systems 20
4.2. Preparation of Advanced Ceramic Systems 24
4.3. Forming Ceramic Articles 26
4.4. Drying and Finishing 28
4.5. Firing Ceramic Products 30
4.5.1. Firing Traditional Ceramics 31
4.5.2. Densification of Advanced Ceramic Products 31
4.6. Kilns and Firing Conditions 32
4.6.1. Modern Periodic Kilns 33
4.6.2. Tunnel Kilns 33
4.6.3. Advanced Ceramics Furnaces 33
4.6.4. Kiln Atmosphere 33
4.6.5. Fired Ware Finishing 34
5. Glazes and Glazing 34
5.1. The Nature of Glazes 34
5.2. Preparation of Glazes 36
5.3. Glaze Application 36
6. Glass 36
7. Refractories 37
8. Abrasives 38
9. Cement 38
10. Properties of Ceramic Materials and Products 38
11. Testing Ceramic Raw Materials and Products 39
11.1. Raw Material and Product Tests 39
11.2. Simplified Testing of Clay Body Materials 40
11.3. Quality Control of Advanced Ceramics 41
12. Economic Aspects 41
13. References 43

1. Traditional and Advanced Ceramics

This general survey covers the fields of traditional ceramics and advanced (or high-technology) ceramics, touching on the materials employed, processing and forming, firing and finishing, and the use of products. Advantages and disadvantages of various types of ceramic ware are discussed.

The word ceramic is a "general term applied to the art or technique of producing articles by a ceramic process, or to articles so produced" [24]. In general, it applies to any of a class of inorganic, nonmetallic products subjected to high temperature during manufacture or use. "High temperature" means any temperature above red heat, ca. 540 °C [13].

Typically, although not exclusively, a ceramic item is a metal oxide, boride, carbide, or nitride, or a compound of such materials. Thus, a *ce-*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_001

ramic article is "a glazed or unglazed object of crystalline or partly crystalline structure (or of glass), produced from essentially inorganic, nonmetallic substances; such objects are made from either a molten mass which solidifies upon cooling or which is formed and matured simultaneously or subsequently by action of heat" [1, p. 197].

The noun ceramic is derived from the Greek *keramos* meaning "burned earth." *Traditional ceramics* refers to ware prepared from an unrefined clay or to combinations of one or more refined clays in combination with one or more powdered or granulated nonplastic minerals or prereacted ceramic compositions. Traditional ceramics also refers to ware or products made from compositions or naturally occurring materials in which clay mineral substance exceeds 20 %. Traditional ceramics and clay ceramics are synonymous expressions.

The past 50 years have seen an increasing interest in ceramic items made from highly refined natural or synthetic compositions designed to provide special properties [19, pp. 150 – 153]. These objects are termed *advanced,* new, or (in Japan) fine *ceramic products,* and find use as key components in such high-technology fields as electronics, computers, optical communication, cutting tools, metal forming dies, wear-resistant parts, high-temperature reactors, high-temperature engine parts, medical implants, and many other special purpose applications. Advanced ceramics must be considered as an enabling technology – one essential to competitive or functional performance of larger systems.

Advanced new roles for ceramics depend on properties inherent in basic structure and composition. Recognition of special capabilities of ceramics is largely due to progress over the past 30 years in relating physical to compositional and structural features [25].

Historical Aspects. The qualities of plasticity, dried strength, and fired hardness of clays were discovered and used possibly as long ago as 10 000 B.C. [26], and certainly by 5500 B.C. [27].

The earliest societies that give reasonable evidence of a ceramic industry seem to have been in the Near East, where a pottery tradition dates back ca. 7000 – 8000 years [28]. In the Far East, Neolithic villagers at Banpo in Shaanxi Province of China were making fine red, gray, black, and painted pottery at least as early as 4500 B.C. [8, pp. 135 –162].

Urban planners at Mohenjo-Daro in the Indus Valley were using fired clay bricks and tile for public building, water-supply conduits, and an advanced sewer system 4000 years ago [29]. A cuneiform tablet of the 17th century B.C. describes the making of a copper – lead glaze [30].

By the Shang Dynasty (1500 – 1025 B.C.) the Chinese had changed Neolithic earthenware to a fine-grained white stoneware [8, pp. 135 – 162]. A primitive feldspathic glaze appeared during the Zhou Dynasty (1000 – 771 B.C.), followed by a soft green to brown lead silicate glaze in the Han Period (206 B.C. – 220 A.D.). In the Six Dynasties time period (265 – 907 A.D.), marked developments occurred in art pottery. Ceramics of the Song Dynasties (960 – 1127 and 1127 – 1279 A.D.) and into the Ming Dynasty (1368 –1644 A.D.) showed increasing use of hard-paste porcelain formulas. An egg-shaped kiln with a special stack designed for high draft was developed for firing Ming porcelains; the fuel was pine wood [31].

Beginning in the sixth century B.C., Attic vases of ancient Greece represented a ceramic art milestone. Use was made of a local illitic clay [32] to prepare a levigated glaze that was black under reducing conditions and red under oxidizing conditions. Rome seems to have had no ceramic tradition, but drew on ware and workmen from various parts of its empire. The famous Arretine ware was made near what is now Florence by Oriental Greeks, and terra sigillata, sometimes thought to have been invented by Romans, originated on the island of Samos [33].

Chinese ceramic technology is thought to have been transferred to the Near and Middle East by cultural contacts and by Chinese prisoners following defeat of a Chinese army by Persian Abbasids in 751 A.D. Islamic influence was felt late in Italy and France [34], and later still in Germany, the Netherlands, and England [35].

Salt-glazed stoneware was developed in High Germany near the end of the fourteenth century [36], and soft-paste porcelain was being manufactured in Meissen as early as 1730 and in Sèvres, France, by 1751 – 1754. In England, COOKWORTHY developed and patented a hard-fire porcelain (1768), and sometime before 1750, calcined bone was being used in making chinaware [37].

Stoneware and hard porcelain served as prototypes for development in the United States, England, and Europe of mechanically strong, vitreous bodies for use in manufacture of sanitary ware, high-tension electrical porcelain, and impact-resistant dinnerware [11, p. 4]. The advent of steam power in the eighteenth century permitted WEDGWOOD and others in England to mechanize preparation and forming operations. Although acceptance of new ideas in the way of processing has been slow over the centuries, within the past two generations "felt wants" in industry have stimulated research in the areas of advanced ceramics.

1.1. Traditional Ceramics

Clay is the oldest ceramic material. The earliest ceramic ware was most likely made from natural clay, selected by the potter for its forming properties. However, at very early times, it was customary to add some other nonclay materials. A sticky, high-shrinkage clay might be modified by addition of crushed stone, sand, or crushed shell to reduce shrinkage and cracking. The major nonclay materials used in making clay-based ceramic items are silica powder and certain alkali-containing minerals added as fluxes. Traditional ceramics can be regarded as ware made from formulations in which clay provides the plastic and dry bonding properties required for shaping and handling. Analyses of natural clay bodies show that the actual clay mineral content is 25 – 40 %.

Pottery is sometimes used as a generic term for all fired ceramic wares that contain clay in their compositions, except technical, structural, and refractory products [1, p. 201].

The term *whiteware* was originally applied to white tableware and artware [11, p. 4], but has been broadened to include ware that is ivory colored or has a light gray appearance in the fired state. Fine ceramic whitewares are conveniently divided into two classes: (1) formulas consisting primarily of clay minerals, feldspathics, and quartz; and (2) nontriaxial bodies made entirely or predominantly of other materials. For purposes of this discussion, ceramic whiteware is placed into five categories, namely, (1) earthenware, (2) stoneware, (3) chinaware, (4) porcelain, and (5) technical ceramics.

Earthenware is defined as glazed or unglazed nonvitreous (porous) clay-based ceramic ware. NORTON subdivides earthenware into four categories: (1) natural clay body, (2) refined clay body, (3) talc body, and (4) semivitreous triaxial body [11, p. 4]. Fired absorptions may range from 4 – 5 % for semivitreous ware to 20 % for the high-talc formulas. Fired color may range from red for high iron oxide bodies to white for the talc and triaxial formulas.

Stoneware is a vitreous or semivitreous ceramic ware of fine texture, made primarily from nonrefractory fireclay or some combination of clays, fluxes, and silica that matches the forming and fired properties of a natural stoneware. Thus, stoneware may be made either from a clay or may be a synthesized stoneware. Synthesized stoneware can range from highly refined, zero-absorption chemical stoneware to less demanding dinnerware and artware formulas.

Chinaware is vitreous ware of zero or low-fired absorption used for nontechnical applications. It can be either glazed or unglazed. The expression soft-paste porcelain has the same meaning [11, p. 4]. Formulas can be simple clay – flux – silica triaxial bodies or bodies containing significant percentages of alumina, bone ash, frit, or low-expansion cordierite or lithium mineral powders. Fired absorptions range from 0 to 5 % for ovenware.

Porcelain is defined as glazed or unglazed vitreous ceramic ware used primarily for technical purposes. Formulations are generally of the triaxial type although some or all of the silica can be replaced by calcined alumina to increase mechanical strength. Firing of ware may be bisque (unglazed) at low temperature with glazing at high temperature or by single-firing at high temperature.

Technical ceramics include vitreous (i.e., nonporous) ceramic whiteware used for such products as electrical insulation, chemical ware, mechanical and structural items, and thermal ware.

The clays used for making *common brick* are usually of low grade and in most cases red-burning. The main requirements are that they are easy to form and fire hard at as low a temperature as possible, with a minimum loss from cracking and warping. An average of analyses of a number of brick clays from sources in New Jersey [38] showed approximately 67 % SiO_2,

18 % Al_2O_3, 3 % Fe_2O_3, 2 % alkaline-earth oxides, and 4 % alkalies, with an ignition loss of about 4 % (→ Construction Ceramics).

Bodies [11, p. 2] can be classified as being either *fine* (having particles not larger than ca. 0.2 mm) or *coarse* (having the largest particle ca. 8 mm). These can, in turn, be subdivided into bodies fired to a *porous* state and those with a fired absorption [1, p. 197] not exceeding 5 %, i.e., a *dense* state. The classes of fine clay ceramics and product uses are arranged in Table 1 to show the percent absorptions and body colors. The classes of coarse clay ceramics and their fired porosities are given in Table 2.

1.2. Advanced Ceramics

Advanced ceramics are generally used as components in processing equipment by virtue of such ceramic properties as special electromagnetic qualities, relative chemical inertness, hardness and strength, and temperature capabilities, sometimes in combination.

A systematic classification of advanced ceramics based on function is presented in Table 3, and examples of materials and uses are shown. A broader system classifies all applications into structural, electronic, and other. *Structural* applications are mechanical, but do include chemical aspects where these are required to carry out the mechanical function. The *electronic* category covers electric, magnetic, and optical functions plus chemical functions that involve direct use of electronic properties. The *other* classification includes strictly chemical functions, for example, catalysis, as well as biological functions.

Of course, any classification is likely to be inexact because many applications involve simultaneous use of several functions. However, a functional classification system does point to the fact that, in contrast to metals, ceramics can be made to embody a wide variety of electronic functions while also having desirable chemical and mechanical properties.

Ceramics are already widely used in process industries, especially where corrosion, wear, and heat resistance are important. Excellent examples are found in metallurgical refractories [40], an area already feeling the effect of new developments in ceramics and the new demands of advanced metallurgical processing.

1.2.1. Advanced Structural Ceramics

(→ Ceramics, Advanced Structural Products)

The prominent families of advanced structural ceramics and structural materials involving ceramics include

alumina
silicon carbide
silicon nitride
partially stabilized zirconia
transformation-toughened alumina
lithium aluminosilicates
ceramic – ceramic composites
ceramic-coated materials

These materials are widely used in diesel, turbocharger, and gas-turbine engines; in high-temperature furnaces; and in the machines and equipment needed for manufacturing.

Although *alumina* [*1344-28-1*] denotes pure Al_2O_3, the term is commonly applied to any ceramic whose major constituent is alumina, even if the ceramic contains other components. Commercial alumina microelectronic substrates with strengths above 350 MPa are obtained by conventional sintering. Hot-pressing techniques result in strengths of ca. 750 MPa, although parts are expensive with limited size and geometries. A variation on conventional sintering [41] produces a glass-bonded alumina with strengths of ca. 700 MPa. Although the glassy phase limits applications to moderate temperatures, this new alumina ceramic should compete with other more expensive, advanced ceramic items.

Fibrous alumina is employed as a reinforcing agent in metal matrix composites and offers promise for filtration of hot gases and as high-temperature insulation. Alumina is used with SiO_2 in making such fibers [42]. Pure Al_2O_3 fibers are made by a variety of solution processes to produce fibers with strength of 1400 MPa.

Silicon carbide [*409-21-2*], a synthetic product, has good wear and erosion resistance and can be produced in either cubic or hexagonal crystal structure. Unfortunately, SiC is inherently unstable in oxygen so that long life under oxidizing conditions requires a surface coating of protective oxide.

Table 1. Fine ceramic products

Type of product	Earthenware		Stoneware		Chinaware		Porcelain	
	% Absorption	Color	% Absorption	Color	% Absorption	Color	% Absorption	Color
Artware	10 – 20	red-white	0 – 5	red-white	0 – 1	white	0 – 0.5	white
Ballmill balls							0 – 0.2	white
Ballmill liners							0 – 0.2	white
Chemical ware			0 – 0.2	gray-white			0 – 0.2	white
Cookware			0 – 5	gray-white				
Drainpipe			0 – 5	gray				
Insulators							0 – 0.2	white
Kitchenware	10 – 15	white	0 – 5	gray				
Ovenware	10 – 20	white			1 – 5	tan		
Sanitary ware					0.1 – 0.3	tan-white		
Tableware	5 – 20	white	0 – 5	white	0 – 1	white	0 – 0.5	white
Tile	10 – 20	white	0 – 5	red-white				

Table 2. Coarse ceramic products

Porous (> 5 % (+ 5 %) absorption)		Dense (< 5 % (− 5 %) absorption)	
Building materials	Refractory	Chemical	Structural
bricks	flue linings	acid-resistant bricks	quarry tile
terra cotta	fireclay bricks		sewer pipe
roofing tile	insulating bricks		fireclay sanitary ware
drain tile			

Silicon nitride [*12033-89-5*], Si_3N_4, is likewise a synthetic product, existing in two phases, alpha and beta, each having hexagonal crystal structures. Silicon nitride ceramics include hot-pressed, reaction-bonded, and sintered products. The SiAlON family is a solid solution of Al_2O_3 and/or other metal oxides in the β-Si_3N_4 structure [43]. Reaction-bonded Si_3N_4 is made by nitriding cast or cold-pressed shapes of silicon powder, whereas hot-pressed Si_3N_4 is made from silicon nitride powder as a sintered Si_3N_4 powder product. Reaction-bonded Si_3N_4 retains its strength at high temperature if it is protected from oxidation [44]. Hot-pressed Si_3N_4 has high short-term strength and better oxidation resistance, but needs additives to facilitate compaction [45].

The advanced cutting tool industry is dominated by cemented carbides [46]. Ceramic vapor-deposited coatings have extended tool life. Efforts are under way to increase tool use by basing tools on Si_3N_4 and SiAlON to reduce dependence on strategic W, Ta, and Co [47].

Silicon nitride possesses many interesting properties that suggest use in bearings [48]. Tests showed an estimated life for Si_3N_4 bearings of 8 times that of steel bearings. The economics of machining and finishing is the biggest obstacle to widespread use of Si_3N_4 bearings.

Zirconia [*1314-23-4*], ZrO_2, finds widespread use in a stabilized cubic form as an oxygen sensor in process industries and the automobile industries [49]. The destructive transformation of ZrO_2 at 1100 °C from monoclinic to cubic form has been overcome by keeping unstabilized particle size of ZrO_2 grains below 1 µm diameter. Then an alumina matrix toughens the Al_2O_3 ceramic [50]. Hot pressing was initially used, but slip-cast forming and sintering has been found to be feasible [51], [52].

Cordierite [*12182-53-5*], $2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$, has a thermal expansion of $(8-12) \times 10^{-7}$ over the range 20 – 1000 °C and is widely used as a catalyst support for automobile emission control units. Similar materials are used as heat exchangers in automotive gas-turbine prototypes and can be considered candidates for other heat-exchanger applications where good thermal shock resistance and moderate crushing strength are required [53]. Silicon carbide and silicon nitride also find application in heat exchangers [54].

Ceramic – ceramic composites and *ceramic–metal composites* (→ Ceramics, Ceramic – Metal Systems) are receiving increasing attention. Sil-

Table 3. Classification of high-technology ceramics by function [39]

Function	Material	Uses
Electric functions	insulation materials (Al_2O_3, BeO, MgO)	IC circuit substrate, package, wiring substrate, resistor substrate, electronics interconnection substrate
	ferroelectric material ($BaTiO_3$, $SrTiO_3$)	ceramic capacitor
	piezoelectric materials (PZT)	vibrator, oscillator, filter transducer, ultrasonic humidifier, piezoelectric spark generator
	semiconductor materials ($BaTiO_3$, SiC, $ZnO-Bi_2O_3$, V_2O_5, and other transition-metal oxides)	NTC thermistor: temperature sensor, temperature compensation PTC thermistor: heater element, switch, temperature compensation CTR thermistor: heat sensor element thick-film thermistor: infrared sensor varistor: noise elimination, surge current absorber, lighting arrestor sintered CdS material: solar cell SiC heater: electric furnace heater, miniature heater
	ion-conducting materials (β-Al_2O_3, ZrO_2)	solid electrolyte for sodium battery ZrO_2 ceramics: oxygen sensor, pH meter, fuel cells
Magnetic functions	soft ferrite	magnetic recording head, temperature sensor
	hard ferrite	ferrite magnet, fractional-horsepower motor
Optical functions	translucent alumina	high-pressure sodium vapor lamp
	translucent magnesia, mullite	lighting tube, special purpose lamp, infrared transmission window
	translucent $Y_2O_3-ThO_2$ ceramics	laser material
	PLZT ceramics	light memory element, video display and storage system, light modulation element, light shutter, light valve
Chemical functions	gas sensor (ZnO, Fe_2O_3, SnO_2)	gas leakage alarm, automatic ventilation fan, hydrocarbon detector, fluorocarbon detector
	humidity sensor ($MgCr_2O_4-TiO_2$)	cooking control element in microwave oven
	catalyst carrier (cordierite)	catalyst carrier for emission control
	organic catalyst	enzyme carrier, zeolite
	electrodes (titanates, sulfides, borides)	electrowinning aluminum, photochemical processes, chlorine production
Thermal functions	ZrO_2, TiO_2 ceramics	infrared radiator
Mechanical functions	cutting tools (Al_2O_3, TiC, TiN)	ceramic tool, sintered SBN, cermet tool, artificial diamond, nitride tool
	wear-resistant materials (Al_2O_3, ZrO_2)	mechanical seal, ceramic liner, bearings, thread guide, pressure sensor
	heat-resistant materials (SiC, Al_2O_3, SiN_4)	ceramic engine, turbine blade, heat exchangers, welding-burner nozzle, high-frequency combustion crucible
Biological functions	alumina ceramics implantation hydroxyapatite bioglass	artificial tooth root, bone, and joint
Nuclear functions	nuclear fuels (UO_2, UO_2-PuO_2)	
	cladding material (C, SiC, B_4C) shielding material (SiC, Al_2O_3, C, B_4C)	

(Courtesy of the American Ceramic Society)

icon carbide fibers in glass – ceramic matrices have shown toughness values up to 24 MPa $m^{0.5}$ at 1000 °C with cross-plied and unidirectional strengths of 500 and 900 MPa [55]. The reinforcing action of 60 % alumina fibers in aluminum gave a tensile strength of 690 MPa up to 316 °C [56]. Use of as little as 3 % of pure Al_2O_3 particles in aluminum increased strength and wear resistance [57].

A thickness of 10 – 15 mils (25 – 38 mm) of plasma-sprayed porous *ceramic coating* such as ZrO_2 can reduce the temperature of the metal surface under the coating by 160 °C [58]. Such coatings are used on aircraft burners and aircraft afterburners, but not in critical parts of aircraft gas turbines. Pore-free coatings applied by chemical vapor deposition, sputtering, or reactive evaporation are 70 – 80 times as resistant to wear and erosion as porous coatings. High-temperature lubrication may make use of solid ceramic lubricants.

1.2.2. Electronic Ceramics (→ Ceramics, Electronic)

Ceramics are involved in electronics as discrete units; however, as component sizes become progressively smaller, they are increasingly integrated into overall electronic assemblies. FISHER [59] has classified discrete ceramic parts into three categories: insulators, magnetic ceramics, and transducers.

Insulators represent a complex category including integrated circuit packages, insulating substrates, and a variety of special tube circuits. Electrical insulation materials are, in a sense, descended from traditional electrical porcelains, but property requirements plus the complex nature of integrated circuits make them a new family. Aluminum oxide is the dominant advanced ceramic insulator [11, pp. 426 – 429]. Tape-cast alumina ceramics dominate in uses requiring high heat dissipation and hermeticity. Alumina ceramics also compete with polymers and coated metals as supports for electronic chips. As excellent as alumina is for this purpose, alternative materials are being studied in an effort to lower the dielectric constant, permit higher frequency operation, and provide a closer match to silicon thermal expansion. Multiphase ceramics in the Al_2O_3– SiO_2– MgO family may be the second generation of ceramics, with Si_3N_4 as the third generation.

Several trends are apparent in the development of later-generation ceramic substrates. One line of development seeks to use low-firing, glass-bonded aluminas that can be cofired with copper, silver, or gold electrodes. A second line of development seeks to exploit the high thermal conductivity of AlN [*24304-00-5*]. Another candidate is BeO-doped SiC. A third line of development is concerned with finding lower-loss materials for microwave applications.

Ferroelectric ceramics, primarily high dielectric constant $BaTiO_3$ [*12047-27-7*] and related materials, find use in capacitors, which are indispensable in electronics. The use of cheaper metals as electrodes may lower unit costs [11, pp. 415 – 417].

Piezoelectrics are crystals whose charge centers are offset: a mechanical stress alters the polarization of the crystal just as an electrical field would. Piezoelectric crystals are widely used for voltage – pressure *transducers.* Piezoelectric ceramics, such as lead zirconate titanate [*12626-81-2*], are used in a wide variety of devices to convert motion into electrical signals and vice versa. Vibrators, oscillators, filters, loudspeakers, all using piezoelectric devices, are essential parts of many industrial and consumer products [19, pp. 55, 287].

Certain ceramics are termed semiconductors, electrical conduction occurring only if external energy is applied to fill energy gaps between filled and empty electron bands. An increase in temperature can also provide the required energy. Ceramic semiconductor materials include titanates, SiC, ZnO, NiO, and Fe_2O_3. In some instances, they are used as thermistors for temperature control. They may be used as voltage-sensitive resistors (varistors) to protect against voltage surges, as chemical sensors, or as mini-heaters [19, pp. 47 – 50].

Ion-conducting ceramics, such as β-alumina and stabilized ZrO_2, are employed as oxygen sensors in automobiles and as electrolytes in fuel cells [60], [61].

Ceramic materials having magnetic properties are commonly termed ferrites. *Magnetic ceramics,* such as ferrites of Fe_2O_3 in combination with one or more of the oxides of Ba, Pb, Sr, Mn, Ni, and Zn, can be made into either hard or soft magnets. These are widely used in loudspeak-

ers, motors, transformers, recording heads, and the like [11, pp. 417 – 421], (→ Magnetic Materials).

The optical properties of a material include absorption, transparency, refractive index, color, and phosphorescence. Optical transparency is often important. Glass and various ionic ceramics are transparent to visible light, and there are many applications for windows, lenses, prisms, and the like. Fiber optics offer enormous potential for communication; small fiber bundles transmitting coherent laser light can carry many times the information carried by wire cables. Magnesium oxide, Al_2O_3, and fused SiO_2 are transparent in the ultraviolet and a portion of the infrared and radar wavelengths. Magnesium fluoride, ZnS, ZnSe, and CdTe are transparent to infrared and radar wavelengths [19, p. 59].

Special pore-free Al_2O_3 is widely used as the inner envelope of high-pressure sodium vapor lights. Lead zirconate titanate ceramics are finding increasing use in light modulation and displays. Translucent Y_2O_3– ThO_2 ceramics are also useful optical materials.

Ceramic sensors can use bulk grain phenomena (such as piezoelectric effects, oxygen-ion conductivity, or negative temperature coefficient of resistivity), grain boundary phenomena (such as positive temperature coefficient of resistivity, voltage-dependent resistivity, or gas absorption), or controlled pore structure (moisture absorption). Occasionally all three microstructural features come into play, with different levels of importance. A broad class of sensors is based on optical fibers [62]. New types of optical sensors using optical fibers can measure temperature, pressure, sound, rotation, current, and voltage. A blood oxygen meter using optical fibers measures light transmission at eight different wavelengths, thus permitting blood oxygen determination.

1.2.3. Other Advanced Ceramics

One of the oldest uses of ceramics is as a thermal insulator at high temperature, and this role is continued in modern form, e.g., as super insulators such as the silica tile used on the U.S. space shuttle. Modern ceramics such as silicon carbide and silicon nitride are increasingly attractive as heat exchangers, as are low-expansion ceramics such as cordierite.

A potentially important market for new ceramics is as implants to replace teeth, bone, and joints.

Ceramics have long been used in the nuclear field as a fuel, cladding material, and shielding material. They are leading candidates as matrices to contain radioactive wastes for long-term storage.

1.3. Characterization of Ceramic Materials

The technology of ceramic manufacturing rests on measurement of the structural and chemical properties of the raw materials used in ceramic forming systems. The need for adequate test procedures is being met by continuing advances in materials science. Many sophisticated instruments and equally sophisticated techniques are available for evaluation of formula ingredients and of forming systems at various stages of manufacture [5, chap. 1].

Purity of ingredients has a profound influence on high-temperature properties of advanced ceramics, including strength, stress rupture life, and oxidation resistance. The presence of Ca^{2+} is known to sharply decrease the creep resistance of Si_3N_4 hot pressed with MgO sintering aid [63], but seems to have little effect on Si_3N_4 hot pressed with Y_2O_3 densifying aid [64]. Electrical, magnetic, and optical properties must be carefully tailored by additions of a dopant; slight variations in distribution or concentration can alter final properties significantly. Ceramic materials can occur in different geometries. As an example α-Si_3N_4 is preferred over β-Si_3N_4 for hot pressing or ordinary sintering.

In recognition of the importance of consistent properties of raw materials and synthetic powders used for advanced ceramic items, an ad hoc committee appointed by the Materials Advisory Board of the National Research Council (United States) gave the term *characterization* a special, restrictive meaning in the following definition [65]: "Characterization describes those features of the composition and structure (including defects) of a material that are significant for a particular preparation, study of properties, and suffice for the reproduction of the mate-

rial." True characterization involves a direct correlation between test results and properties. The mere taking of data is not characterization unless the test procedure serves a particular function in predicting properties of the material under test.

Although this definition was designed as an aid in establishing significant features for advanced ceramic products and their constituents, the concept has been successfully applied in the field of traditional ceramics. The many properties encountered in forming and firing are found to be consequences of the interaction of two or more of a limited list of fundamental characterizing features [66]. Table 4 provides a listing of significant, interacting features for traditional clay-based ceramics, with a partial list of the more important consequential properties encountered in forming and firing. An exhaustive survey of pertinent literature, in addition to a continuing review of plant and laboratory results, has shown no exceptions to the list of characterizing features of Table 4.

Table 4. Characterizing features and ceramic properties

Characterizing features	
chemical composition	
mineral composition	
particle-size distribution	
specific surface	
colloid modifiers	
Ceramic properties	
Unfired/forming	Firing/fired
slip viscosity	vitrification
water of plasticity	shrinkage
workability	pyroplasticity
shrinkage	absorption
strength	strength
slip dispersion	color
casting rheology	thermal behavior
casting rate	microstructure

Characterization, itself rapidly developing as a discipline, has suggested ways whereby selected properties of materials or a body can be used in development and control of clay bodies. Sanitary ware and vitreous chinaware are typical clay-based traditional ceramic products. The chemical, mineral, particulate, and surface data of Table 5 constitute complete characterizing descriptions of examples of formulas used in making these products.

Table 5. Characterization of two clay-based bodies

Properties	Vitreous sanitary ware	Vitreous china
Chemical, wt %		
SiO_2	65.0 ^{x}	69.4 ^{x}
Al_2O_3	23.1 ^{x}	19.5 ^{x}
Fe_2O_3	0.44	0.30 ^{x}
TiO_2	0.28	0.14 ^{x}
CaO	0.33	1.33 ^{x}
MgO	0.13	0.11
K_2O	2.68 ^{x}	1.45 ^{x}
Na_2O	2.41 ^{x}	1.14 ^{x}
Ignition loss	5.67	6.46
Mole of flux	0.0766 ^{x}	0.0604 ^{x}
Minerals, wt %		
Smectite	3.7	3.0
Kaolin group	32.7	33.3
Mica	8.8 ^{x}	5.8 ^{x}
Free quartz	23.7 ^{x}	39.6 ^{x}
Organic	0.46 ^{x}	0.23 ^{x}
Auxiliary flux		2.0
Particle size		
% < 20 µm	76	76
% < 5 µm	47	45
% < 2 µm	33	36
% < 1 µm	25 ^{x}	28 ^{x}
% < 0.5 µm	19	21
Surface		
MBI, meq/100 g	3.3 ^{x}	2.7 ^{x}

^{x} Key indicators.

Two terms require definition. The *mole of flux* is the sum of the percentages of CaO, MgO, K_2O, and Na_2O divided by their respective molecular masses. The *MBI* (methylene blue index) is the milliequivalents of methylene blue cation (chlorine salt) absorbed per 100 g of clay and is a measure of surface area [67].

Reproducibility of desired forming, firing, and fired properties is ensured by maintaining these characterizing features within prescribed limits [66].

Experience has shown that when any or all of the ingredients of a clay body must be replaced the 20-odd characterizing values of a full description may be reduced to 8 – 10 key indicators. A key indicator is a feature that is critical to controlling a particular property. The superscript *x*'s of Table 5 label the key indicators for the two examples.

Because fired body color is much more critical for vitreous chinaware than for sanitary ware, the coloring effect of Fe_2O_3 and TiO_2 must be taken into account when, for example, vitreous chinaware is reformulated [68]. The presence of

mica in sanitary ware slip-casting significantly improves the casting rate and the quality of cast [69]. The presence of colloidal organic matter can increase response to deflocculants and result in significant increases in dry bonding power [70]. The rheology of clay-based forming systems can be altered adversely by apparently minor changes in subsieve particle-size distribution [71]: the percentage finer than 1 µm equivalent spherical diameter is an excellent indicator of any change [72]. The methylene blue indices (MBI) correlate with plastic forming properties and dry strength of unfired ware, both of which are functions of specific surface [67].

2. Raw Materials for Traditional Ceramics

Clay-based ceramics are predominant among ceramic products. Clay formulas (or bodies) may consist of a single clay or one or more clays mixed with mineral modifiers such as powdered quartz and feldspar. The special properties of the clay minerals that permit preparation of high-solids fluid systems and plastic forming masses are critical in the shaping of ware.

In developed countries, ceramic manufacturers and raw material suppliers usually work together in establishing standards [2]. The supplier assumes responsibility for continuity of material quality and works closely with the manufacturer in solving material-related plant problems.

However, in less developed countries, manufacturers may need to depend on suppliers who lack facilities and expertise for maintaining material uniformity. An alternative is that the manufacturer may be forced to mine and refine his own materials. In either case, the potter must be prepared to cope with variation in material properties, either by active supervision of supplier mining or through in-plant beneficiation prior to use. The characterization concept (Section 1.3) has permitted development of objective, simple test procedures for use in mining and beneficiation control [17].

2.1. The Structure of Clays and Nonplastics

The atomic structures of the common clay minerals are based on Pauling's generalizations for the structure of the micas and related minerals [73]. Two structural units are involved in most clay mineral lattices. One is the *silica sheet,* formed of tetrahedra consisting of a Si^{4+} surrounded by four oxygen ions. These tetrahedra are arranged to form a hexagonal network repeated to make a sheet of composition $Si_2O_5^{2-}$. The tetrahedral apex oxygens all point in the same direction with pyramid bases in the same plane.

The other structural unit is the *aluminum hydroxide, or gibbsite, sheet,* consisting of octahedra in which an Al^{3+} ion is surrounded by six hydroxyl groups. These octahedra make up a sheet, owing to sharing of edges: two layers of hydroxyls have cations embedded in octahedral coordination, equidistant from six hydroxyls. These octahedral sheets condense with silica sheets to form important clay minerals.

Kaolinite [*1318-74-7*] is the main mineral of kaolins, with usually tabular particles made up from units resulting from the interaction of gibbsite and silica sheets:

$$Al_2(OH)_6 + (Si_2O_5)^{2-} \longrightarrow Al_2(OH)_4(Si_2O_5) + 2\,OH^-$$

The kaolinite platelets have negative charges on their faces (or basal planes) due to an occasional Al^{3+} ion missing from the octahedral (gibbsite) layer or an Si^{4+} from the tetrahedral (silica) layer.

Disordered kaolinite is a variant of kaolinite in which Fe^{2+} and Mg^{2+} are thought to replace some Al^{3+} in the octahedral layer [23, p. 59]:

$$Al^{3+}_{1.8}Ca^{2+}_{0.1}Fe^{2+}_{0.1}(Si_2O_5)(OH)_4 \cdots M^{2+}_{0.1}$$

The M^{2+}, usually Ca^{2+}, is a balancing exchangeable cation. Hydrogen bonds between gibbsite and the silica layers can be weakened by changes in the octahedral dimensions caused by replacement of the small Al^{3+} (ionic radius of 0.051 nm) by the larger Fe^{2+} (0.074 nm) and Mg^{2+} (0.066 nm) ions. This produces the smaller grain size of disordered kaolinite found in some sedimentary kaolin and ball clay deposits.

Kaolinite crystals consist of a large number of two-layer units held together by hydrogen bonds acting between OH groups of the gibbsite structural layer of one unit and oxygens of adjacent silica structural layers. Unit layers are displaced regularly with respect to one another along the *a* axis. In the case of *halloysite,* the unit layers are stacked along both *a* and *b* axes in random fashion; because of less hydrogen bonding, water can penetrate between successive layers, thereby forming a hydrated variety of kaolinite,

$$Al_2(OH)_4(Si_2O_5) \cdots 2\,H_2O$$

According to KELLER [74], halloysite can exist as spheres, tubular elongates, or polygonal tubes; thus, kaolin occurs in a number of morphologies ranging from worms through stacks, irregular platelets, to euhedral kaolinite crystals. Particle morphology can have significant effects on ceramic forming systems [17].

The montmorillonites result from isomorphous replacements of portions of Al^{3+} or Si^{4+} in the three-layer mineral *pyrophyllite* [*12269-78-2*], which is formed by fusion of two silica sheets with one gibbsite sheet [75]:

$$Al_2(OH)_6 + 2\,(Si_2O_5)^{2-} \longrightarrow$$
$$Al_2(OH)_2 \cdot 2\,(Si_2O_5) + 4\,OH^-$$

When Mg^{2+} replaces some of the Al^{3+} in the octahedral layer, the result is *montmorillonite* [*1318-93-0*] (smectite),

$$Al_{1.67}Mg_{0.33}(OH)_2 \cdot 2\,(Si_2O_5) \cdots M^+_{0.33}$$

M^+ lying between two adjacent three-layer units as an exchangeable cation, offsetting the excess basal-plane negative charge. Because the SiO_2 of adjacent unit layers are held together only by weak van der Waals attraction, montmorillonite particles are thin and small.

If one-quarter of the Si^{4+} ions of the tetrahedral layers of pyrophyllite are replaced by Al^{3+}, a charge of sufficient magnitude is produced to bind univalent cations in regular 12-fold coordination. If the cation is K^+, the result is *muscovite mica* [*1318-94-1*] [23, p. 23]:

$$KAl_2(OH)_2 \cdot 2\,(Si_{1.5}Al_{0.5}O_5)$$

If the cation is Na^+, the result is *paragonrite mica* [*12026-53-8*]*:*

$$NaAl_2(OH)_2 \cdot 2\,(Si_{1.5}Al_{0.5}O_5)$$

Many natural clays contain a micaceous mineral, resembling muscovite but containing less M^+ and more combined water than normal muscovite. This *illite* [*12173-60-3*] occurs in sedimentary clays sometimes associated with montmorillonite and kaolinite. Analyses of illites from various localities show K_2O contents of 3 – 7.5 %; SiO_2 of 38 – 53 %; and Al_2O_3 of 9 – 32 %. Knowledge of illite is as yet incomplete [23, pp. 24 – 25].

Table 6 shows the names and chemical compositions of plastic clay minerals and nonplastic layered aluminum and alkaline-earth silicate minerals commonly encountered in ceramic clays.

2.2. Clay – Water System

When a clay is dispersed in water, its balancing *exchangeable cations* retreat to a distance from the clay determined by their size and charge, forming an electrical double layer. If the water contains cations of a different kind and charge, an exchange of solution cations for clay-held cations may occur. Some cations are attracted more strongly to the clay than others. Cations can be arranged in a lyotropic (Hofmeister) series [22, p. 24]; hydrogen is held most strongly and lithium least:

H Al Ba Sr Ca Mg NH_4 K Na Li

The capacity of a clay for absorbed cations is termed its cation exchange capacity (c.e.c.) and is a function of clay specific surface [76]. The usual measure of the cation exchange capacity is the MBI (see Section 1.3).

The stability of a suspension of clay particles in water depends on the *degree of deflocculation* of the particles. Deflocculation depends on the character of an electrical double layer made up of the following parts [22, pp. 92 – 110]:

1) Negative surface charge consisting of the inherent negative planar surface charge plus absorbed OH on normally positively charged edges
2) Absorbed layer of cations at the negative surface, the Stern layer

Table 6. Layer lattice minerals

Mineral	Composition
Plastic	
Kaolinite	$Al_2O_3 \cdot 2\,SiO_2 \cdot 2\,H_2O$
Fireclay	$(Al_{1.8} \cdot Fe_{0.1} \cdot Mg_{0.1})O_3 \cdot 2\,SiO_2 \cdot 2\,H_2O \cdots Ca_{0.05}$
Montmorillonite	$(Al_{1.67} \cdot Mg_{0.33})O_3 \cdot 4\,SiO_2 \cdot 2\,H_2O \cdots Ca_{0.165}$
Illite group	muscovite $\longrightarrow$ illites $\longrightarrow$ montmorillonite
Halloysite [*12244-16-5*]	$Al_2O_3 \cdot 2\,SiO_2 \cdot 4\,H_2O$
Nonplastic	
Muscovite	$K_2O \cdot 3\,Al_2O_3 \cdot 6\,SiO_2 \cdot 4\,H_2O$
Pyrophyllite [*12269-78-2*]	$Al_2O_3 \cdot 4\,SiO_2 \cdot H_2O$
Talc [*14807-96-6*]	$3\,MgO \cdot 4\,SiO_2 \cdot H_2O$
Tremolite [*14567-73-8*]	$5\,MgO \cdot 2\,CaO \cdot 8\,SiO_2 \cdot H_2O$
Chlorite [*14998-27-7*]	$5\,MgO \cdot Al_2O_3 \cdot 3\,SiO_2 \cdot 4\,H_2O$

3) Diffuse cloud of cations that extends to a distance from the charged particle that is determined by the
   a) concentration of ions in the bulk solution away from diffuse cation cloud
   b) size and charge of the cations

The thickness of the electrical double layer is a maximum when the concentration of hydroxides or hydrolyzable salts of the monovalent cations of the Hofmeister series is the minimum needed to fully charge the clay surface. Excess deflocculant reduces the extent of the diffuse layer.

In the absence of a double layer, the bringing together of two clay particles by Brownian motion results in formation of a doublet. Attraction between platelets is either by edge – face attraction or by van der Waals force, or both. Where the normally positive edge has been neutralized or made negative, there is only van der Waals attraction. Particles provided with diffuse, extended counterion clouds cannot approach one another closely enough to allow the inherent van der Waals forces to function fully [21, pp. 183 – 212], so deflocculation or reduced flocculation is the result.

The very polar *water molecules* are attracted strongly to negative faces or positive edges of clay particles. The adsorbed water molecules, in turn, attract other water molecules, and these, in turn, attract yet other water molecules. Thus, a water structure is built on the surfaces of clay platelets or rods. The extension of the water envelope from the particle surface is thought [77] to depend on the size and valence of the cations present in the water. Exchangeable cations can adsorb water molecules and build up a structure whose extension from the clay surface depends on the amount and kind of cations present. Where large singly charged cations are present, a loose, wide extension occurs; for small multiply charged cations, the counterion cloud is compact and less extended [77]. Water of plasticity and plastic qualities are functions of surface area, particle geometry, and exchangeable cations.

However, if a clay is allowed to absorb *organic colloids,* such as tannic acid or humic acid colloids derived from soil organic matter or lignites, the attraction between clay particles is greatly reduced, water of plasticity drops significantly, response to deflocculants is enhanced, and dry strength rises [70]. Apparently the absorbed organic particles with their absorbed water layers neutralize positive edges and provide a measure of steric hindrance to the close approach of particles.

Deflocculation is, thus, a neutralization reaction between acidic groups of absorbed organic colloids and the monovalent cations and hydroxyl groups provided by the deflocculating compound, rather than a reaction between clay and the deflocculant. Some functional groups are more responsive than others; as a consequence, organic-bearing ball clays vary in their forming properties.

The hydroxyl ion is necessary in the deflocculation of clays [70]. The presence of any soluble sulfate or chloride salts in the clay – water system reduces the formation of OH^- and lessens the deflocculating effect of a given quantity of deflocculant.

2.3. Commercial Ceramic Clays

In the United States and the United Kingdom, the major classes of ceramic clays are termed kaolin (or china clay) and ball clay. *Kaolin* may occur at its point of origin in primary deposits or in sedimentary deposits composed of clay particles washed from the point of formation by stream action and laid down in quiet water. Kaolin deposits are widely distributed in the temperate zone. However, in the tropics alteration may be rapid, resulting in bauxite [78].

The term *ball clay* has no technological significance; it is derived from older mining practice in England, whereby cubes of moist, plastic clay were cut from the working face with a special tool, rolled down the clay face, assuming a vaguely spherical shape, and loaded onto wagons by women workers (ball maidens). A general definition of ball clay would be sedimentary clay of fine to very fine grain size, consisting mainly of ordered and disordered kaolinite with varying percentages of illite, mica, montmorillonite, free quartz, and organic matter.

Clays classified as ball clays are widely used in North and South America, England, and to an increasing extent, in Asia. Ball clay is far less used in Europe. The use of ball clays in clay-based forming systems is designed to improve plasticity, reduce water of plasticity, increase unfired strength, improve casting slip properties, and in some cases, improve firing and fired properties. The unfired functions of a ball clay can sometimes be matched by treating fine-grained kaolins with colloidal organic substances [70].

Table 7 characterizes representative china and ball clays from major producing areas in England and the United States. The china and ball clays from Thailand provide examples of ceramic clays available in less-developed nations. The mineral constituents of the clays of Table 7 were calculated from the chemical analyses with a procedure suggested by Holdridge [79].

The primary kaolins of the china clay deposits of England and Thailand contain more mica than the sedimentary kaolins of Georgia (United States), as demonstrated by their higher K_2O contents [66]. English ball clays are much higher in mica than their U.S. analogues [79]. Mica has favorable effects in slip casting and provides a measure of fluxing.

The flow diagrams of Figure 1 are representative of mining and refining practices in ball clay producing areas of Dorsetshire and Devonshire and china clay deposits of Cornwall in England. The ball clay deposits are very thick with relatively thin overlaying soil. The china clay deposits are kaolinized granite and consist largely of mixtures of kaolinite, muscovite mica, quartz, and small amounts of accessory minerals. Over the past 40 years, the clay producers of England have raised mining and refining of their materials to a very high level of technology. As a consequence of already desirable clay properties, coupled with close control and technical competence, a large export trade has been developed.

The thin overburden and thick deposits of English ball clay permit both open-pit mining and underground mining. Open-pit operations are of two types: (1) excavating of uniform seams with backhoes and (2) selective mining of some clays with a spade-carrying version of the pneumatic jackhammer. Air-spaded clay is lifted from the pit with a boom, placed in a truck, and transported to a processing center. Backhoe-dug clay is placed directly into the truck for transport to a processing center. Underground ball clay mining is done either with air-spading for selective mining or by a rotating cutter that loads the clay directly into the mine car for transport to a processing, storage, and refining center.

English ball clays are stored in accordance with types determined by characterizing feature tests. Clays are sliced (shredded) into thumb-size pieces and often blended with one or more other selections to provide controlled, specified properties. Such blends may be extruded in the form of pellets for bulk shipment or dried and subjected to grinding to a refined powder in an air-elutriation grinding mill. Air-floated clay is usually bagged for shipment.

English china clay is recovered by subjecting the parent ore to "hydraulicking" (high-pressure jet of water). The clay and fine muscovite mica are separated from the ore and transported by the resulting stream to a classifier for removal of the coarser mica and quartz. Further nonclay impurities are removed with a hydrocyclone. The low-solids slip is thickened, characterized, and stored as a 20 % solids slurry. Two or more slurried china clay selections may be blended to give desired, controlled properties before filter press-

Table 7. Characterizations of typical clay

Properties	Clay∗									
	A	B	C	D	E	F	G	H	J	K
Chemical, wt %										
SiO_2	45.7	46.7	50.5	60.4	46.6	47.2	48.5	59.8	49.0	58.5
Al_2O_3	38.3	38.2	28.7	27.0	38.1	37.6	32.3	26.4	34.6	24.4
Fe_2O_3	0.41	0.60	0.91	0.93	0.69	0.50	0.98	1.00	0.71	1.26
TiO_2	1.55	1.42	1.48	1.62	0.07	0.05	1.16	1.39	0.02	0.92
CaO	0.08	0.12	0.40	0.28	0.19	0.20	0.18	0.20	0.35	0.05
MgO	0.06	0.20	0.30	0.26	0.20	0.08	0.21	0.51	0.34	1.05
K_2O	0.06	0.15	0.89	1.70	1.47	1.35	1.89	2.42	2.52	2.36
Na_2O	0.14	0.03	0.18	0.50	0.08	0.07	0.19	0.38	0.48	0.12
Ignition loss	13.65	13.79	16.58	7.59	12.66	12.62	14.78	7.88	10.66	8.64
Minerals, wt %										
Montmorillonite	nil	3	8	7	6	2	6	14	9	29
Kaolin group	96	93	58	44	80	82	60	34	60	23
Mica	2	2	10	21	12	11	18	25	27	22
Free quartz	trace	1	14	26	1	1	9	23	3	19
Organic	trace	trace	8	0.5	trace	trace	5	2	nil	3
Particle size										
% < 20 µm	95	99	99	98	100	88	96	97	77	95
% < 5 µm	69	88	95	79	97	74	91	85	52	87
% < 2 µm	52	72	82	61	64	30	83	81	36	72
% < 1 µm	35	56	69	43	57	23	77	75	19	56
% < 0.5 µm	28	41	51	29	35	15	62	65	16	43
Surface MBI, meq/100 g	1.6	10.5	12.1	5.6	5.4	2.4	8.7	12.8	3.4	16.5

∗ Key to designations:
A) Coarse kaolin, sedimentary, Washington County, Georgia, United States
B) Fine kaolin, sedimentary, Wilkinson County, Georgia, United States
C) Dark fine ball, Graves County, Kentucky, United States
D) Coarse light ball, Weakley County, Tennessee, United States
E) Fine china clay, Cornwall, England, United Kingdom
F) Coarse china clay, Cornwall, England, United Kingdom
G) Dark ball, Devonshire, England, United Kingdom
H) Light ball, Dorset, England, United Kingdom
J) China clay, primary, Thailand
K) Ball clay, Thailand

ing and drying. The dried clay may be shipped in bulk pellet form or passed through a pulverizer and shipped in bags.

Figure 2 provides flow diagrams representative of mining and refining methods employed in U.S. sedimentary kaolin deposits of Georgia and South Carolina and in ball clay deposits of Tennessee and Kentucky. Overburden is usually no more than 8 – 10 m thick. Neither ball clay nor kaolin deposits exceed ≈ 15 m. All mining is open pit.

Selective mining based on drill hole and working face characterization tests is done with dragline or power shovel. Transport from pit to processing and storage sites is by trucks carrying 5 – 10 t up to 10 km. Storage is in the form of shredded clay.

Kaolins are blended to specification and either dry-ground for bulk or bagged shipment or subjected to wet processing. High-solids slurries (70 %) are prepared for tank-car shipment to ceramic plants using slip-cast manufacturing. Low-solids slurries are subjected to centrifugal fractionation with subsequent thickening, filtration, and drying. The dried filter cake may be shipped in bulk, air-floated and sent in hopper cars as bulk, or pulverized for bagged shipment.

The ball clays are blended to specification and shipped as is, as high-solids slurries, or dried.

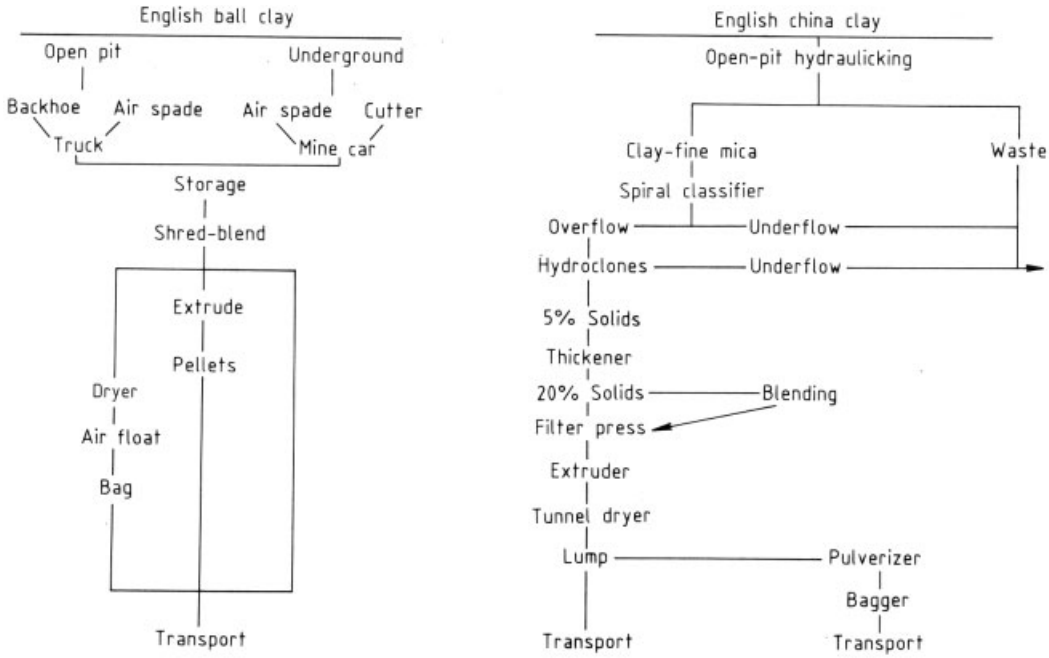

Figure 1. Flow diagrams showing mining and processing methods for the English ball clay deposits of Devonshire and Dorsetshire and the English china clay deposits of Cornwall

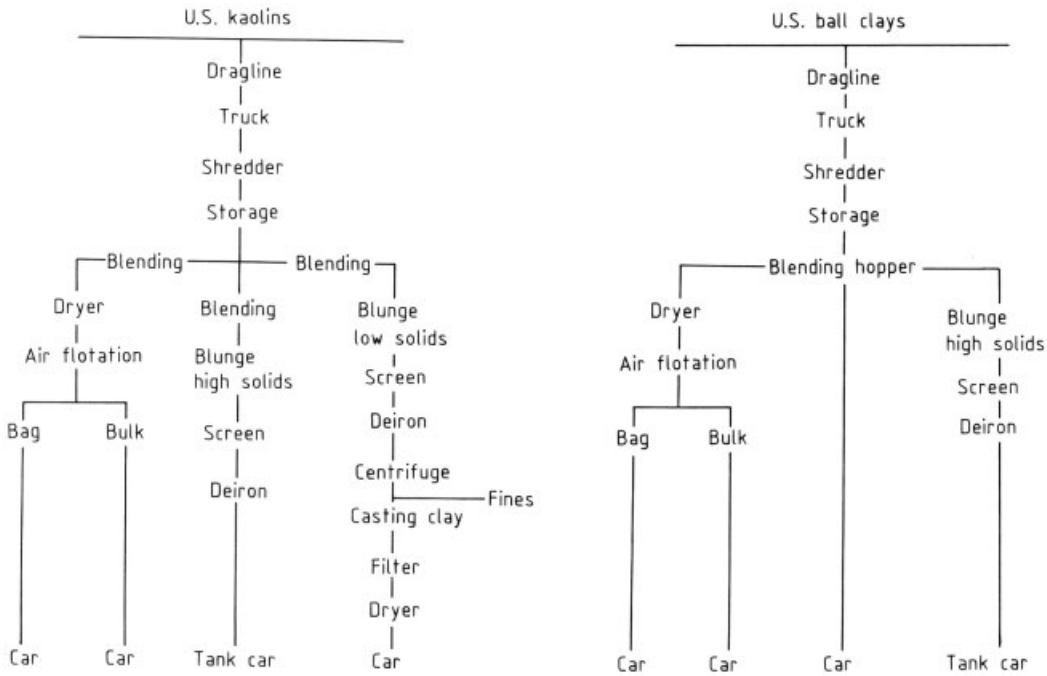

Figure 2. Flow diagrams showing mining and refining methods for the sedimentary kaolins of South Carolina and Georgia and the ball clay deposits of Kentucky and Tennessee

The processing of clays for use in ceramics is also described under → Clays.

2.4. Commercial Nonplastics for Ceramics

A large proportion of ceramic ware is made from clay-based formulas whose major constituents are clay minerals, powdered silica, and powdered feldspar or a related feldspathoid. Such bodies are termed triaxial [11, pp. 178 – 183]. The fluxing feldspathoids and silica minerals are termed nonplastics. The term *flint* is properly used only with reference to powdered flint pebbles.

The *feldspar group* of minerals is the most important source of fluxing oxides for clay bodies. All are framework aluminosilicates based on an SiO_2 structure. Replacement of Si^{4+} by Al^{3+} results in charge deficits that are balanced by K^+, Na^+, or Ca^{2+} lying in framework voids. The smaller Na^+ and Ca^{2+} ions confer a different crystal structure than the larger K^+ ion. Albite [*12244-10-9*] ($NaAlSi_3O_8$) and anorthite [*1302-54-1*] ($CaAl_2Si_2O_8$) are isomorphous and form the plagioclase solid-solution series. Albite and anorthite are triclinic, whereas microcline [*12251-43-3*] ($KAlSi_3O_8$) is monoclinic. Nepheline syenite is a type of rock consisting of nepheline [*12251-27-3*] ($K_2O \cdot 3\,Na_2O \cdot 4\,Al_2O_3 \cdot 9\,SiO_2$) mixed with microcline and albite.

An old saying, attributed to the Chinese [11, p. 92], says in effect that *silica* [*7631-86-9*] is the skeleton and clay the flesh of a ceramic body. There is a tendency to regard silica as an inert substance in the body. However, this is far from the case: the silica can have profound effects both in forming and firing.

Table 8 provides examples of fluxing feldspathoids and silicas used in clay-based ceramic formulations. The mineral constituents of the feldspars and silicas of Table 8 were calculated from the chemical analyses with a method by Koenig [80]. Feldspar A is a froth-floated feldspar recovered from North Carolina alaskite granite. Material C is dry-ground, selectively mined nepheline syenite from Ontario, Canada. Material E is wet-ground feldspar from Thailand. All are successfully used in clay-based ceramic formulations.

In addition to the feldspathics and silica, some clay-based bodies contain calcined Al_2O_3 to increase fired strength; ground limestone and/or dolomite as auxiliary flux; talc for special heatshock bodies and wall tile; chlorite to lower the maturing temperature of slip-cast porcelains; or wollastonite, a wall-tile body constituent.

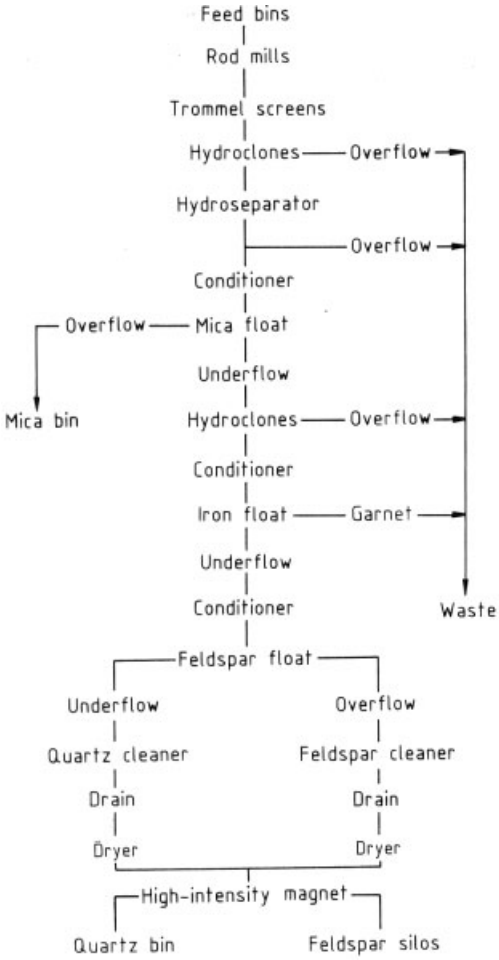

Figure 3. Flow diagram showing typical froth flotation recovery of muscovite mica, feldspar, and quartz from a coarse granite found in western North Carolina

The principal sources of pottery and glass grade feldspar in the United States are deposits in Connecticut, North Carolina, South Carolina, Oklahoma, and California [81]. Nepheline syenite, also widely used in ceramic formulations and in glass batches, is produced from deposits in Methuen Township, Ontario, Canada [82].

Table 8. Characterizations of typical nonplastics

Properties	Feldspathic/Feldspathoid∗					Flint/Quartz∗				
	A	B	C	D	E	F	G	H	J	K
Chemical, wt %										
SiO_2	66.8	68.5	60.7	79.5	71.1	96.6	98.5	97.9	99.5	95.7
Al_2O_3	19.6	17.5	23.3	12.0	16.0	0.2	0.9	0.5	0.2	2.1
Fe_2O_3	0.04	0.08	0.07	0.08	0.26	0.10	0.09	0.40	0.06	0.03
TiO_2				0.01	0.34	0.01	0.06			0.09
CaO	1.70	0.30	0.70	0.20	1.54	0.20	0.02			0.20
MgO	trace	trace	0.10	0.09	0.37	0.09	0.03			0.20
K_2O	4.80	10.40	4.60	3.80	0.06	0.39	0.05			0.30
Na_2O	6.90	3.00	9.80	3.90	8.64	0.15	0.04			0.03
Ignition loss	0.20	0.30	0.70	0.45	0.42	1.58	0.17	0.20	0.12	1.60
Mole of flux	0.1728	0.1644	0.2220	0.1092	0.1768	0.0124	0.0028			0.0123
Minerals, wt %										
Feldspars	92	83	75	48	81					
Nepheline			24							
Mica	4	7			trace					2
Quartz	4	9		41	18	96	97	97	99	91
Clay				3		1	1		trace	6
Organic						trace				trace
Other			1	8	1	4	2	3	1	1
Particle size										
% < 20 µm	67	64	68	60	53	56	57	75	58	53
% < 5 µm	26	26	22	23	20	16	18	26	15	12
% < 2 µm	11	12	9	11	10	5	7	9	5	5
% < 1 µm	9	9	3	6	5	1	2	3	1	3
% < 0.5 µm	trace	6	trace	2	2			1		2

∗ Key to designations:
A) Flotation feldspar, Mitchell County, North Carolina, United States
B) Block feldspar, Custer County, South Dakota, United States
C) Nepheline syenite, Ontario, Canada
D) Cornish stone, Cornwall, England, United Kingdom
E) Feldspar, Thailand
F) Flint, France
G) Quartzite, Pennsylvania, United States
H) Quartzite, Venezuela
J) Silica sand, California, United States
K) Silica sand, Philippines

Prior to 1940 all feldspar mined in the United States was selectively quarried, crushed, and hand-cobbed on picking belts before being ground. Just after World War II a froth floating procedure began to be applied to mixed-mineral rocks containing feldspar. At the present time over 80 % of the feldspar produced in the United States is recovered by froth flotation from a variety of ores, including alaskite granite, pegmatite, graphic granite, beach sand, and weathered granite. The remaining feldspar, mainly high K_2O feldspar, is block mined, hand-cobbed, and processed dry. Nepheline syenite is also selectively mined and subjected to dry processing.

Figure 3 provides a generalized flow diagram for froth flotation recovery of feldspar from coarse granites.

After (normally) thin overburden has been removed from the ore, the granite is blasted and transported to a processing plant. The large pieces are passed through, successively, a jaw crusher and cone crusher to prepare rodmill feed.

From the feed bins the thumb-sized pieces of ore pass through rodmills where they are reduced to millimeter-sized grains. The rodmilled pulp then goes onto rotating screens to remove oversize, which is returned to the rodmill for further grinding. Passage of screened pulp suspended in water through a hydroseparator removes most of the fines that might interfere with the chemistry of flotation processes. The sized, de-slimed pulp is then sent to a chemical conditioner where the mica particles are treated to promote bubble adherence. The underflow (feldspar, quartz, and

garnet) is conditioned chemically to allow only the iron-containing garnet to be attracted to bubbles and so removed in the froth overflow. Next comes separation of feldspar from the quartz by adjusting the reagents to cause feldspar particles to adhere to the froth and the quartz to be rejected.

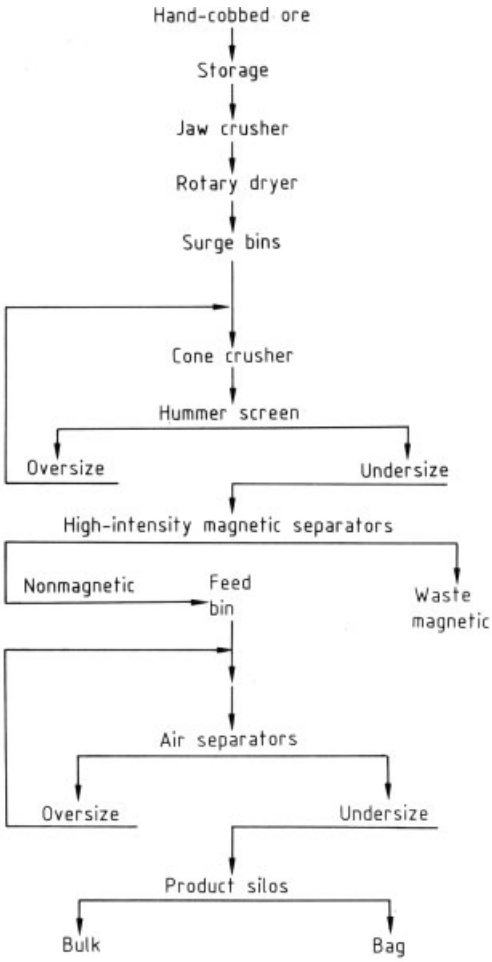

Figure 4. Flow diagram showing a typical cobbing (hand selection) method for mining and processing a pegmatite feldspar in southwestern South Dakota

Final steps involve draining, rewashing to remove reagents and draining of the cleaned products, passage of drained material through a dryer and through a magnetic field, and finally storage. Pottery uses require fine grinding; glass grade requires no grinding of the granular feldspar or quartz.

Where a deposit is sufficiently pure, block feldspar may be processed as shown by the diagram of Figure 4.

The blasted block material is passed through a jaw crusher prior to passage through a rotary dryer into a surge bin. The crushed, dried product passes through a cone crusher onto a 2.4-mm vibrating screen, with any oversize being returned for further crushing.

The minus-2.4-mm product goes into a surge bin that feeds a high-intensity magnetic separator; magnetic particles pass to waste, while nonmagnetics go to mill feed bins. Milling is by pebble mills; ground product goes to air classifiers, with any oversize returned for further milling. Undersize passes to storage silos for bagged or bulk-loaded shipment. Nepheline syenite is processed in much the same manner, with an additional step designed to produce a minus-1.0-mm size for glass batching.

3. Raw Materials for AdvancedCeramics

Although traditional ceramics are composed of natural raw materials that are physically separated and reduced in size, advanced ceramics require chemical conversion of raw materials into intermediate compounds. These intermediates lend themselves to purification and eventual chemical conversion into a final desired form.

Oxides and carbonates available in powder form include those of Al, Sb, Ba, Be, Bi, Co, Mn, Mg, Ni, Si, Th, Ti, and Zr. Also available are carbides of Si, Ti, and W and the nitrides of Al, B, Hf, Si, and Zr. However, needs exist for specialized powders for some advanced ceramics, and a variety of chemical routes can be used to synthesize these powders. Chemical routes, such as sol – gel processing, can bypass the powder stage.

Requirements for high strength and smooth finishes, particularly of small parts, necessitate fine-grained powders. Thus, one line of advanced ceramic research aims at producing very fine, essentially spherical, monosize particle powders. These are typically made by colloidal chemistry for oxides. Nitrides and carbides involve controlled nucleation and growth in gas-phase reactions. However, most high-technology ceramics are still made from pow-

ders with broad size distributions in the submicrometer (under 1 μm) range.

3.1. Metal Oxides and Carbonates

Alumina [*1344-28-1*] is derived from bauxite by selective leaching with NaOH, precipitation of purified $Al(OH)_3$, and thermal conversion of the resulting fine-size precipitate to Al_2O_3 powder (→ Aluminum Oxide) for use in polycrystalline Al_2O_3-based ceramics. Antimony [*7440-36-0*] is derived from Sb_2S_3 (stibnite) by reduction with iron scrap, and antimony trioxide [*1314-60-9*] is formed by burning antimony in air.

Barium oxide [*1304-28-5*] is obtained by decomposition of $BaCO_3$ at high temperature; the carbonate itself is made by reaction of Na_2CO_3 with BaS. Beryllium oxide [*1304-56-9*] is prepared by heating $Be(NO_3)_2$ or $Be(OH)_2$. Bismuth oxide [*1332-64-5*] is obtained by heating $Bi(NO_3)_3$ in air.

Cobalt compounds are derived from ore concentrates by roasting and leaching with acid or ammonia; the oxide [*1307-96-6*] is formed by calcination of the carbonate or sulfate. Magnesium oxide [*1309-48-4*] is readily available as the 99.5 % pure grade powder, but greater purity may require calcining of high-purity salt solutions. Manganese oxide [*1344-43-0*] can be prepared by calcination of manganous nitrate.

Nickel ores are either sulfidic or oxidic. Sulfides are flotation-separated and roasted to sintered oxide. Oxides are treated by hydrometallurgical leaching with ammonia. Nickel oxide [*1313-99-1*] is then prepared by gentle heating of $Ni(NO_3)_2 \cdot 6\,H_2O$. Strontium carbonate is formed by boiling celestite, $SrSO_4$, in a solution of $(NH_4)_2CO_3$; SrO [*1314-11-0*] is formed by decomposition of the resulting $SrCO_3$.

Vanadium pentoxide [*1314-62-1*] is prepared by ignition of alkali solutions from vanadium minerals. Zinc carbonate [*3486-35-9*] is prepared by action of sodium bicarbonate on a zinc salt, such as zinc chloride. Zirconia [*1314-23-4*], ZrO_2, is derived from $Zr(OH)_4$ or $Zr(CO_3)_2$ by heating.

3.2. Borides, Carbides, and Nitrides

Boron and carbon can be made into B_4C [*12069-32-8*] by heating B_2O_3 and carbon in an electric furnace. Boron nitride [*10043-11-5*] is made by heating B_2O_3 and tricalcium phosphate in an ammonia atmosphere in an electric furnace (→ Boron Carbides, Boron Nitride, and Metal Borides).

Boron, carbon, and nitrogen can be made into other synthetic compounds with refractory and wear properties. Examples are silicon carbide (SiC), silicon nitride (Si_3N_4), tungsten carbide (WC), titanium carbide (TiC), titanium nitride [*25583-20-4*] (TiN), tungsten boride (WB_2). A translucent AlN has been developed that is 5 times as thermally conductive as Al_2O_3 ceramics.

4. Processing Ceramic Ware

Traditional and advanced ceramic industries use many techniques for processing their products. The exact process is governed by the nature of the forming system, the size and geometry of the piece, product specification, and practices in various areas of the ceramic industry.

Most ceramic manufacturing processes start with formulas consisting of one or more particulate materials. These formulas are used for shaping products that are further processed by firing and by finishing of the fired items.

In many cases products have complex shapes made by use of one or another of such forming techniques as dry or isostatic pressing, plastic shaping, extrusion, slip casting, injection molding, tape casting, and green finishing.

Forming systems employed in making traditional and advanced ceramic ware are (1) liquid suspensions, (2) plastic masses, or (3) more or less dry granulated or powdered formulations.

4.1. Preparation of Clay-based Forming Systems

The clay bodies of traditional ceramics are normally mixtures of clays and powdered nonclay minerals or else natural mixtures of clay substances and nonclay particulate materials. Most clays occur as aggregates of clay particles. When contacted with water, such aggregates tend to

break apart or slake. The development of a water structure on the surfaces of the particles results in plasticity (see Section 2.2. Clay – Water System). If sufficient water is added to the clay and the mixture is agitated, a dispersion forms. Because the powdered nonplastics, i.e., the nonclays, do not develop any great degree of plasticity when moistened with water, the various ceramic systems of traditional ceramics depend on the plastic component (usually but not always clay) to provide (1) the workability required in plastic forming or dry pressing, (2) the deflocculant response of fluid systems in slip casting, and (3) the green and dry strength of unfired ware.

Figure 5 shows the moisture-content variation and forming-pressure ranges for soft plastic shaping, extrusion, dry pressing, dust pressing, isostatic pressing, and slip casting of clay-based bodies.

Because the ingredients used by any given plant may range from highly purified to as-mined lump materials, the body preparation process must vary with the particular circumstances. However, the main objectives of processing are always (1) to arrive at as intimate a mixture of clay and nonplastic particles as possible, (2) to provide uniformity of shaping properties from lot to lot, and (3) to maintain uniformity of firing and fired properties from lot to lot.

Preparation processes for these forming systems can be divided into two general classes: (1) wet processing and (2) dry processing.

Wet Processing. Wet processing is usually employed whenever one or more of the ingredients needs initial or supplementary beneficiation. General practice in the United States and the United Kingdom subjects dinnerware bodies (Table 5, Vitreous china) to wet processing to ensure adequate dispersion of clay constituents, permit sieving for removal of oversize, and allow magnetic treatment to remove iron particles. Such a process uses relatively unrefined shredded or lump ball clays and filter cake or coarsely pulverized china clay.

Third-world ceramic manufacturers may have access to producer-beneficiated materials but often must depend upon their own mines for at least a portion of their raw materials. In some instances beneficiation of local materials becomes an integral part of body preparation. In the People's Republic of China and Thailand, for example, the silica and fluxing feldspars may be received in block form and ground during the body preparation process.

Because grinding is readily accomplished by dry crushing, followed by wet ball milling, one approach is to wet-grind the nonplastics along with a small, fixed percentage of suspending fine clay. The nonplastic slop (suspension) is then sieved, deironed by magnets, and stored in agitators. Clays are wet-dispersed as suspensions, sieved and deironed, and then blended by formula with nonplastic slop in agitator tanks.

A modification of this method is to simply weigh all formula ingredients as a unit, transfer the batch to a ball mill with the required water, and mill to a specified sieve residue percentage.

For plastic or dry-press forming systems, wet processing is done in the flocculated state. Indeed, flocculation is often enhanced in preparation of high-tension electrical porcelain bodies by addition of $AlCl_3$ or $MgSO_4$. Sufficient water must be used to allow sieving and passage through a magnetic separator.

Consistency of plastic masses is controlled by four major factors: (1) specific surface of the body, (2) modifying inorganic ions such as Ca^{2+}, Mg^{2+}, Al^{3+}, SO_4^{2-}, and Cl^-, (3) the amount and kinds of organic colloid present, and (4) the proportion and temperature of water present.

The water content of plastic forming systems is reduced to a working level by filter-pressing and, when necessary, by further drying of the filter cake. Air is removed from the filter cake by passage through a vacuum pug mill.

Pressing dust can be made from filter cake by drying the cake and passing the dried cake through a granulating hammer mill. Otherwise, sieved and deironed slip can be diverted to a spray drier and formed into pressing-size granules.

Dry Processing. The flow diagrams of Figures 1234 (Chap. 2) indicate that finely ground, deironed clays and nonplastics can be obtained in both North America and the United Kingdom. Such materials make it possible to prepare both plastic forming systems and pressing dust without a slip stage. Dry pressing dusts are prepared by dry blending the ingredients with ribbon blenders or rotating cone mixers and then incorporating the required moisture with a muller

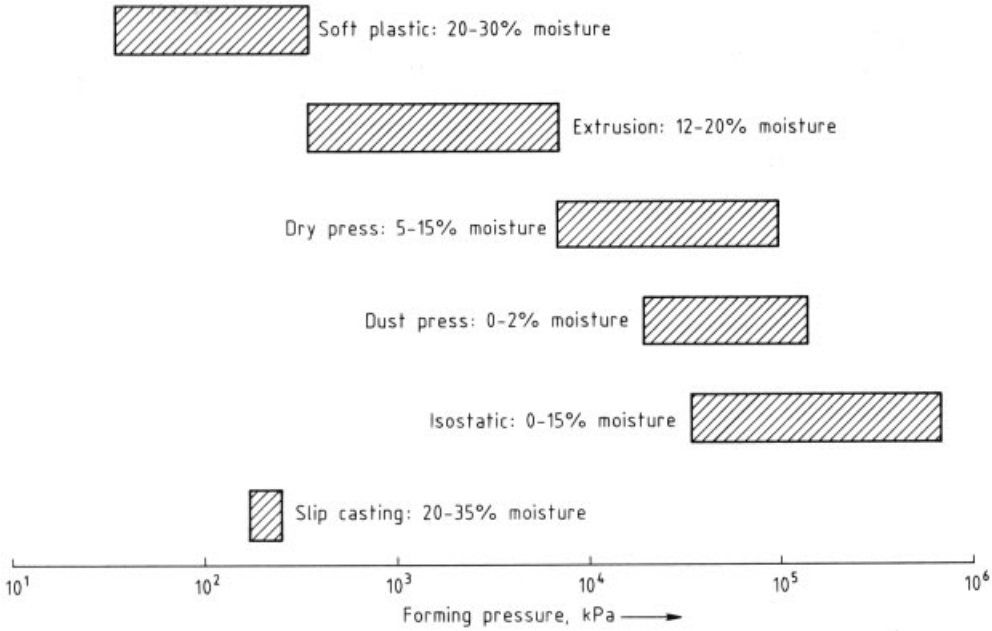

Figure 5. Moisture content and pressure ranges required for shaping clay-based forming systems

mixer. If sufficient water is mulled into the mix and the resulting plastic mass is passed through a vacuum pug mill, the resulting forming system can be used for plastic forming.

Refractories and heavy clay products are usually made from combinations of clay and coarse nonplastics by crushing them in a wet pan (heavy rollers revolving in a pan) and adding water plus other modifiers. By variation of the moisture, the mulled mixes can be made into pressing dusts by granulation or into plastic systems by a deairing operation.

Casting Slip. Although filter-cake clay body is sometimes made into casting slip by addition of deflocculating agents, by far most casting slips are made by direct wet methods.

Clay-based casting slips must be made to cast to a firmly plastic state within a prescribed time range. Casting properties, such as rate, amount of retained water, and plastic quality of casts, are each in some way related to freshly stirred consistency of the slip and its tendency to thicken on standing. Common practice in industry is to control casting properties by maintaining a constant *solids concentration* by measuring slip specific gravity and adjusting *slip rheology* to targeted freshly stirred viscosity and thixotropic gelling. Unfortunately, the mere meeting of a targeted rheology is no guarantee of constant casting performance.

Variation in *slip temperature* can alter slip viscosity and casting rate significantly [83]. Thus, it is possible for two slip batches at different temperatures to have identical viscosities and thixotropies, yet to cast in decidedly different ways. Ryan and Worrall [84] found that the nature of exchangeable cations in casting slip governs the rate of cast under constant temperature and rheological conditions. The custom in sanitary ware slip control is to buffer the effect of deflocculant-enhancing organic colloid by addition of divalent alkaline-earth-metal carbonates or sulfates to control the rate and structure of the cast [85].

The rheology and casting properties of casting slips are strongly influenced by apparently minor changes in the distribution of *particle sizes* in the subsieve region. Brociner and Bailey [86] have shown that the coarse kaolin component of a casting slip can be made variably finer as the input of energy imparted in blunging or ball milling is varied: the mixing or milling operation must be very carefully controlled, and

both equipment and time of mixing should be kept constant.

In direct preparation of casting slips, on occasion a standard sequence and timing of additive and raw material introduction into the mixer is not followed. If, for example, a light ball clay is added first with the Na_2CO_3, followed by an organic-bearing ball clay, the amount of adjusting sodium silicate required is significantly greater than if the reverse order were used, and the resulting slip requires a longer aging period. If deflocculation is initiated with sodium silicate and the Na_2CO_3 is added later, the aging time [87] is greatly extended. When slips prepared by using differing sequences of addition are adjusted to the same viscosity and thixotropy, their casting rates and cast structures are also likely to differ significantly.

Equipment. Those plants that grind their own nonplastics use *ball milling,* either continuous dry grinding in an air-swept conical ball mill or batch wet grinding in a cylindrical ball mill. Dry grinding demands that the feed material be dry to avoid packing and to allow air sweeping of fines to a collector. Wet grinding is claimed to require less power than dry grinding, but dry grinding produces less wear on the mill lining and grinding media [88, sect. 6, pp. 20 – 25].

Ball mills belong to a class of grinding devices termed tumbling mills. The rotating container is a cylinder mounted with its axis horizontal. The grinding action is due to the tumbling of the grinding media, which are cast iron or steel balls, hard rock (e.g., flint pebbles), or some nonmetallic material such as high-alumina porcelain [88, sect. 6, pp. 20 – 25].

Blunging refers to the agitation or blending of ceramic materials in a mixing tank equipped with an impeller to stir the suspension and baffles to direct the suspension to the impeller. Impellers may be simple paddles or specially designed shapes for increased efficiency of dispersion [89].

Screening (or sieving) of fluid dispersions is termed wet screening. Two general types are employed in the sieving of blunged or ball-milled slips: (1) an inclined rectangular panel of wire mesh having the proper openings and (2) circular screens. The *inclined rectangular panels* are subjected to vibration that agitates and separates the coarser particles during transit of the slip. Vibration can be by shaking or electromagnetic pulse [88, sect. 7, pp. 34 – 37].

Circular vibratory screens can [90] effectively separate particles as fine as 44 μm in diameter. The basic arrangement consists of a motor plus interchangeable frames that hold screening wire cloth and discharge ports. The frame is held rigidly to a main screen assembly. The motor has a vertical upward and downward extended shaft fitted with eccentric weights. The main screen assembly is mounted on a circular base by springs that permit the assembly to vibrate freely, while preventing vibration of the floor. A number of three-dimensional patterns of the suspension on the screen can be developed by varying the angle between top and bottom weights. This type of screen is widely used in the United States and the United Kingdom.

Screens used for pressing-dust sizing are relatively coarsely meshed (2.0 – 3.0 mm), whereas those used for plastic body systems and casting slips are much finer (0.20 – 0.05 mm).

To remove magnetic particles, granular nonclay ball-mill feed can be subjected to a *magnetic separator,* passage either through a magnetic field or over a magnetic pulley prior to the grinding operation. Suspensions of clays or nonclay powders can be passed through the grid of an electromagnetic purifier prior to the dry or pugging operations. High-gradient magnetism is capable of removing such colorants as TiO_2 from kaolin slurries; this can transform the high-TiO_2 Georgia and South Carolina kaolins into very white-firing fine-china constituents [91].

Dewatering of slips for preparation of plastic forming systems or pressing dusts is usually by *filter pressing.* The basic concept of filter presses involves feeding the slurry under pressure into the space between square, round, or rectangular plates. This space is created by frames that alternate with the plates. Plates are hollow and normally covered with filter cloth. As the space fills with suspension under pressure, the liquid is forced through the cloth and drains away as the solids form a cake [92].

Pugging is the process of blending clays and water by manual or mechanical means. A pug mill is an open trough with a lengthwise shaft on which are mounted blades that blend the clay and water to a plastic forming system of the desired consistency. Filter-cake bodies are subjected to a combination of pugging in an auger trough,

coupled with passage through a vacuum chamber, followed by extrusion. Vacuum pug milling (or deairing) makes the plastic mass more workable and cohesive by elimination of the air from the system [13, p. 267].

A fluid suspension of particulate material can be dried and formed into pressing dusts or granules by *spray drying.* A spray dryer consists essentially of a drying chamber. A downward flow of heated air is introduced at the top of this chamber. A flow of suspension is transformed into an upward flowing cloud of droplets by a nozzle atomizer. The droplets are dewatered and fall to a product outlet at the bottom of the drying chamber. An attached exhaust removes excessively fine particles to a cyclone collector. Relatively uniform spheres are formed, and moisture content is also uniform [93].

When fully purified clays and nonclay powders are available, *dry blending and tempering* are employed. A shell or ribbon mixer may be used as an initial step, followed by addition of water in a mixing muller. Otherwise, dry blending and tempering may be done in stages in a mixing muller. Mixing mullers normally have heavy wheels, under which the moistened body rotates. First, a smearing action occurs, and second, a rotating plow scrapes the compressed body up and turns it under the mullers for additional mulling [94]. The tempered body then goes to dust mill/sieving operation for lower-moisture-dust processing or dry pressing or to a deairing pug mill for plastic systems.

4.2. Preparation of Advanced Ceramic Systems

Traditional ceramic forming systems are nearly always polydisperse, with particle size ranging more or less continuously from an upper to a lower limit. As a distribution ranging between definite limits approaches linearity on an arithmetic plot, optimum packing results in minimum voids [95]. The more extended the range between upper and lower limiting sizes, the lower the void volume for a given distribution [96]. However, the more extended the distribution is, the more sensitive it is, with respect to void volume, to deficits or excesses of intermediate particle sizes. This finding has been related to differences in calcined alumina slip occasioned by altering particle-size distribution size limits and intermediate size distribution [97].

Although a controlled optimum particle size distribution is needed for maximum, reproducible strength, sometimes a mono-size distribution must be approached to avoid growth of larger particles at the expense of the smaller: very fine particles are much more reactive than larger particles, and quite porous initial compacts can be sintered at high temperature to nearly theoretical density. Transparent polycrystalline Al_2O_3 is an example. The finer the powder, the more rapid the sintering and the lower the densification temperature, thereby reducing grain growth and increasing fired strength.

Sizing of Advanced Ceramic Materials. Because particle size and distribution are so important for controlling properties of advanced ceramic products, the manufacturer must often further refine an already refined as-received material to meet his specifications. A variety of techniques are used for modifying particle size and distribution:

screening
air elutriation
ball milling
attrition milling
vibratory milling
fluid energy milling
liquid elutriation
precipitation
freeze drying
laser
plasma
calcining
sol – gel

Dry *screening* is used for sizing particles down to 44 µm, whereas wet, slurry screening is often employed for subsieve sizes. *Air elutriation* (or classification) is used to separate coarse and fine fractions. Special air classifiers are available for separating minus-20-µm particles, but care must be exercised to avoid contamination. *Liquid elutriation* can be used to separate a single specific material into fractions or to separate materials having different specific gravities.

Ball milling [16, pp. 410 – 438] consists in placing either a dry or a suspension charge in a

closed container with appropriate grinding media and rotating the container to create a cascading action of the media. Media selection is important. Higher density pebbles or cylinders will grind more quickly than lower density media. Wear of media creates contamination that can be controlled by careful selection of wear-resistant mill lining and hard grinding media. Wet ball milling requires removal of water from the powder. Dry ball milling requires additional grinding aids such as a lubricating stearate. A very small amount of moisture has been found to prevent packing of high-alumina prereacted body during dry grinding.

Attrition milling is similar to ball milling, but the container is held in a fixed vertical position and the grinding media agitated by arms attached to a rotating shaft. The attrition mill can be used for dry grinding or wet grinding with vacuum or various controlled atmospheres [16, pp. 439 – 443].

Vibratory milling uses fixed containers typically lined with polyurethane or rubber. Suitable grinding media are placed in the container with the material to be ground, and a vibration is transmitted through the bottom center. The resulting cascading – mixing action leads to shear and impact breaking of particles between grinding media [16, pp. 410 – 438].

Fluid energy milling functions by causing particles of the material to be ground to impact one another. They are carried at high velocity in a fluid – air, water, superheated steam. Jet mills are lined with wear-resistant materials [98].

Precipitation of soluble salts and pyrolysis to the oxide has been used to provide controlled particle size and high purity. Calcined alumina has been made by precipitating alumina trihydrate from solution by changing pH and using seed crystals. The very fine, reactive alumina greatly extends the uses for alumina [19, p. 165].

Freeze drying involves forming drops from solutions of metal salts, freezing them rapidly, removing water by sublimation under vacuum, and calcining the crystallized salts [99]. Another method for preparing pressing granules is by dispersing the powder and additives as a slurry and drying by spraying the slurry or solution into a chamber where the drops fall through hot gases. Surface tension holds the drops in spherical form. These drops, when dry, flow readily into a die [100].

Slip-cast advanced ceramic forming systems require a particle distribution [97] that provides maximum packing. Often sizing is accomplished by blending several narrow distributions [95], or the material may be ball-milled with binder, wetting agents, deflocculants, and densification aids. Disk mills [16, pp. 468 – 488] are especially effective in dispersing agglomerated powder. The liquid phase normally used in mold casting is water, whereas in tape casting the liquid is usually nonaqueous [101]. In each instance, all air bubbles must be removed from slips by vacuum treatment prior to use [102].

A number of glasses have been prepared in the laboratory by hot pressing or sintering gels of single oxides or combinations of two or more oxides, such as SiO_2, Al_2O_3, and TiO_2. Carefully controlled processing makes monolithic objects possible [103]. Commercial uses of *sol – gel* are fibers, powders, bulk shapes, and oxide coatings of films [104]. Of these uses, film or oxide coatings are regarded as very important.

Processing of a sol – gel starts with a metal alkoxide: $Si(OC_2H_5)_4$, $Ti(OC_2H_5)_4$, as well as $Al(OC_2H_5)_3$ are examples. Alcohol and distilled water are hydrating reagents. A wide variety of silicate and aluminosilicate systems have been made with other cations, such as those of Li, Na, K, Rb, Mg, Ca, Sr, Ba, Pb, Ga, Fe, Ln, Ti, Zr, and Th, as well as ternary or quaternary compositions with two or more of these elements [105].

The basic procedure for making SiO_2 and metal oxide gels is to dissolve $Si(OC_2H_5)_4$ in ethyl alcohol and add alcohol or water solutions of the desired metal nitrate. Hydrolysis is effected with an excess of distilled water. At 60 °C the SiO_2 precipitates as a stiff gel.

Preconsolidation of Advanced Ceramics. Preparation of a pressing dust sometimes involves addition of a binder, a lubricant, possibly a sintering aid, and finally, development of a free-flowing powder by granulation. This may be done by blending the fine, low-bulk-density powder with binder solution and lubricant, and then compacting the mass into blocks that are chopped, crushed, or coarsely pulverized. The resulting granules are screened to obtain proper size for die filling.

4.3. Forming Ceramic Articles

Forming systems used to make traditional and advanced ceramic ware include slip casting, soft plastic, stiff plastic, dust pressing, dry casting, and a number of modified or special systems for advanced ceramics (see Fig. 5).

Soft Plastic Forming. The simplest method of forming plastic masses is by hand molding. This requires a soft plastic system. Soft plastic forming systems are used in the production of soft mud bricks; pottery by throwing; jiggered or roller-formed tableware; hot-plunge insulators; and ram process products. In *soft mud brick making,* the selected clays are prepared by wet panning and passed through a pug mill that forces the plastic clay through a die into wooden molds. *Throwing* on the wheel is a soft plastic method for making vases and the like, used in simple cultures and by art potters. The wheel is a disk on top of a shaft turned by a weighed kickwheel or by a motor.

Jiggering was developed from throwing. A measured slug of soft plastic body is placed on a plaster form that revolves on a wheel head. A template tool is brought down onto the moist bat, pressing it down onto the plaster mold and so forming the upper part of the piece. At the same time, the template tool scrapes away excess body from the moist piece with the aid of a spray of water. Automation requires carefully controlled, deaired forming masses [106].

The *roller-head method* for soft plastic forming is an alternative to the jigger, especially for less plastic formulations such as bone china and hard porcelain. Instead of a scraping template blade, a polished (and sometimes heated) contoured metal roller is brought down and rolls out the plastic body onto the plaster form. In this case, the form remains stationary [107].

Hot-plunging or jollying of plastic body articles involves the placing of a measured slug of body in a plaster mold and having a heated, revolving polished metal tool press down and form hollow objects, such as pin insulators or cups. The term *hot pressing* is sometimes applied to the hot-plunging operation, but hot pressing is more generally used for special, advanced ceramics processed by application of high pressure to fine-grained oxides in refractory molds held at high temperature [108].

The *ram process* involves pressing a lump of soft plastic body between two hard plaster molds and squeezing them together to form a plate, ash tray, or similar object. In the pressing stage, water is squeezed out of the piece and a vacuum pulls moisture into the molds. In the removal step, the vacuum is maintained on the upper mold and pressure is applied to the lower mold to release the piece. The upper mold then lifts the piece free, and pressure is applied to free the object from the upper mold. Pressure is also applied to blow moisture from the mold halves before another cycle starts [109].

Stiff Plastic Forming. Stiff plastic systems are extruded through a die, either by auger extrusion or piston extrusion. Auger extrusion is a continuous operation, whereas piston extrusion is necessarily intermittent. *Piston extrusion* is used for extruding fine-grained refractories, cermets, and electronic bodies. A preformed, deaired slug is placed in the cylinder and forced through a die at pressures up to 35 MPa. Pieces as small as 1 mm in diameter with a half-dozen 0.1-mm-diameter holes can be made. Large sewer pipes are piston-extruded with a vertical piston extruder [11, p. 147].

Auger extrusion finds use in extruding bricks and hollow tile on a continuous basis. Short sections are cut off at desired lengths. The auger device consists of a pugging trough that feeds a screw, which in turn pushes the clay through a shredder into a vacuum chamber. The deaired shreds are recompacted with a screw and pushed through the die. High-tension insulator blanks of up to 1 m in diameter [12, pp. 111 – 112] are extruded with auger deairing pug mills and are used in lathe-turning segments of very large electrical insulators.

Dust Pressing. The term applied to forming of damp, granulated body batches containing 5 – 15 % moisture that are formed at high pressure in a steel die is dust pressing. All wall tile, floor tile, some quarry tile, and most low-tension electrical porcelain is formed by dust pressing. More than 85 % of all fireclay brick and nearly all silica brick and basic brick are formed by dust pressing. Hydraulic presses and hydraulic toggle presses are used [11, pp. 149 – 151].

Dry Pressing. Dry pressing is similar to dust pressing, but the moisture content is < 2.0 %, so that a binder and internal lubricant must be employed. Dry pressing is employed for advanced ceramic products in two ways [11, pp. 149 – 151]. Small shapes are pressed by *uniaxial compaction* [11, pp. 151 – 152], wherein the pressure is applied (usually) in a downward, vertical direction, thereby producing pressure variations due to wall friction and particle – particle friction. This results in nonuniform density. *Isostatic compaction* involves application of pressure equally to all sides of the charge. An isostatic press consists of a thick-walled pressure vessel. Powder is enclosed in a liquid-proof rubber mold that is immersed in a noncompressible fluid. The fluid is pressurized and transmits pressure equally to all sides of the mold; pressures can range from 35 MPa to as much as 1400 MPa, but usually ca. 210 MPa [110].

Slip Casting. Deflocculated liquid systems are made into ware by slip casting [17, chap. 10]. Formation is accomplished by consolidation of the particles into a semirigid state through removal of a portion of the liquid phase by an absorbent, porous mold. The most common mold material is the hemihydrate of gypsum, $CaSO_4 \cdot 0.5\,H_2O$ [*10034-76-1*], which when mixed with water rehydrates and forms needles of gypsum crystals as an interlocked mass [111], thus forming continuous capillary pores (→ Calcium Sulfate). The size and liquid-carrying capacity of plaster molds is controlled by varying the plaster : water ratio.

Walker [112] observed that as water rises from the stoichiometric 18.5 kg of water to the region of 60 – 100 kg (depending on the processing and ore source of the plaster) per 100 kg of $CaSO_4 \cdot 0.5\,H_2O$, the rate of casting increases to a maximum and then decreases with greater water ratio. Both the specific gravity and the compressive strength decrease with further increases in water.

Although the suction pressure of porous plaster decreases as the amount of water increases, a larger pore size allows freer passage of moisture from the developing cast and provides a larger reservoir for liquid as it is removed from the slip [112]. The loss of moisture from the exterior of the mold by evaporation is a significant controlling factor in governing the rate and condition of casts; high external humidity reduces, and low external humidity raises, the rate of cast and time of setup.

Slip casting takes two general forms. In the first, slip is poured into the mold where water is absorbed, leaving a semirigid layer of particles next to the mold wall. After a sufficiently thick layer has developed, the excess slip is poured out. The cast wall continues to pass moisture into the mold, thus reducing the moisture gradient from drain to wall, and allowing the cast to assume the firmly plastic state needed for cast removal. This is *drain casting,* which is used for hollow items. In the second, a slip at a somewhat higher solids concentration (55 vol% against 50 vol% for the drain-cast slip) and a greater thixotropy (reversible thickening) is poured into the mold and allowed to cast solid. This is termed *solid casting.* On occasion, solid casting and drain casting are used on the same piece. The character of the cast and its rate of buildup are controlled by manipulation of the particle size [72] and colloid modifiers [17, chap. 1].

Special Systems for Advanced Ceramics. Advanced ceramics can be consolidated and formed by the following methods:

1) Pressing
   - uniaxial pressing
   - isostatic pressing
   - hot pressing
   - hot isostatic pressing
2) Casting
   - slip casting
   - soluble-mold casting
   - thixotropic casting
3) Plastic forming
   - extrusion
   - injection molding
   - transfer molding
   - transfer molding
   - compression molding
4) Others
   - tape forming
   - flame spraying
   - green machining

To this point discussion has focused on the shaping methods originally used for the less-demanding clay-based formulas, but which

have been refined for use in making small, more-demanding advanced ceramics. Certain advanced ceramic products require very thin sheets. A method for making such products makes use of casting or spreading a specially prepared slip or slurry onto a moving carrier surface and controlling its thickness with a doctor blade [113]. In such cases, the system resembles an oil-base paint. The powder is dispersed in a volatile solvent (nonaqueous organic liquid) with unsaturated organic acids of 18 – 20 carbons, and a polymer binder and plasticizer are added. Drying consists primarily in removal of the volatile solvent, which leaves a thin flexible tape.

An interesting and useful modification of slip casting also involves an adaptation of investment casting. First, a water-soluble wax is injection molded to make a pattern. The pattern is then coated with a water-insoluble wax, and the water-soluble part is dissolved away. The wax mold is fastened to an absorbent plaster block and is filled with slip. Once casting is completed, the water-insoluble wax is dissolved from the cast with an organic solvent, and the cast is dried, machined as needed, and fired to the proper temperature for densification [19, pp. 197 – 199].

Injection molding makes use of the techniques for molding plastic combs and the like, the difference being that the polymer, either thermosetting or thermoplastic, serves only to disperse the ceramic powder and to provide lubrication [114]. A sized powder is milled dry with organic binders and made plastic by preheating. The plastic mass may require as little as 24 % or as much as 50 % binder by volume, depending on particle size and particle-size distribution. Complex shapes can be made [19, pp. 200 – 203].

4.4. Drying and Finishing

Drying of ceramic products is one of the more critical processing operations. The moisture must be removed as rapidly as possible without generating stresses great enough to cause cracking or distortion.

A plastic ceramic piece contains liquid in three forms: (1) adsorbed liquid on the colloidal particles; (2) liquid films on particles of noncolloidal dimensions; and (3) free liquid held in pores between the particles. Liquid must leave the system in three distinct stages [20, pp. 82 – 84]:

1) By evaporation from the surface of the piece, bringing the particles closer together, decreasing the volume of the piece proportionately, and eventually allowing the particles to come into contact, at which time shrinkage ceases
2) By removal of the remaining free moisture
3) By removal of the adsorbed moisture

As moisture leaves the piece, a gradient is established between the surface and interior of the ware. Because of the shrinkage factor, this gradient must not be too great; otherwise, excess shrinkage at the surface will cause cracking.

Moisture Stress. Many mechanisms affect the behavior of clay-based ceramic forming systems during dewatering processes such as slip casting, filtration, and drying. Some of the mechanisms involved are capillarity, adsorption, osmotic pressure, the electrical double-layer, and pore water structure [115]. The moisture changes in unfired ceramic bodies can be studied by measuring the specific energy of the water as the fundamental parameter. This energy concept is termed moisture stress [116], [117].

Moisture stress is defined as the work done per unit mass of water when a small amount of the water is moved from the clay – water system to a free water surface at the same temperature. The SI unit is J/kg. Expressed in simpler terms, moisture stress is a measure of the affinity of a porous system, such as a moist clay body or liquid slip, for its moisture.

A number of methods have been proposed for measuring moisture stress. Packard [117] employed a direct suction device in which moist clay was placed on a water-saturated, fritted glass plate that was in contact with water. A tube connected to the water vessel could be raised and lowered to increase or decrease suction, thus altering the moisture content of the clay. Coleman and Marsh [118] used a pressure-membrane apparatus for very high moisture stress. Packard [117] proposed use of a series of evacuated closed containers where moist clay was suspended over solutions of known relative humidity. Samudio [119] was able to assign moisture stress (pF) values to casts made from

slips containing small percentages of various inorganic salts by using a pressure-membrane method.

Moisture stress ranges from nearly 0 for dilute suspensions to ca. 10^6 J/kg for oven-dried clay bodies. Soft plastic bodies have moisture stresses in the region of 40 – 50 J/kg. Leather-hardness occurs at moisture stress of in the region of 5×10^3 J/kg, whereas air-dried clay bodies have moisture stresses of ca. 10^5 J/kg.

Rate of Drying. The moisture stress concept implies that for a body of a given particle-size distribution, the rate of drying depends on a structure imparted by the interaction of many factors. A simple example is the control of permeability and water holding in slip-cast pieces by interaction of organic matter, deflocculants, and flocculating salts [120]. Organic colloids are known to reduce the water of plasticity of fine-grained clays [17, p. 45], yet that same organic matter has a strong affinity for water and can retard moisture loss [121]. Similarly, the presence of fine-grained muscovite mica in plastic clay bodies or moist cast appears to slow drying [69].

If a granular solid is involved, moisture loss proceeds at about the same rate as from a pan of water under the same conditions [11, p. 161]. However, from a plastic mass of fine clay, the rate of moisture loss may be less than that of the coarser system [122]. Ceramic clay bodies would be expected to show loss rates between those for granular masses and clay masses. In any one of these circumstances, the drying rate is constant until particles touch and shrinkage stops. At this point the continuum of the pore water ceases and flow of water from the interior of the piece cannot maintain the surface film needed for rapid evaporation. The rate of drying then falls.

The shrinkage properties of clay bodies are useful in setting up an efficient drying schedule. Natural clays give shrinkage traces that vary with clay fineness and packing characteristics of the noncolloidal particles [122]. Contrary to other opinion [4, p. 554], Norton stated that the rate of water removal from the surface of a plastic clay piece is approximately one half that from a free water surface under the same conditions. Accordingly, evaporation rate from a moist granular solid is much greater than that from a plastic fine-grained clay [123]. The rate of flow of liquid from the interior can be increased by decreasing the viscosity of the liquid and this is accomplished by raising its temperature.

Speed of drying is also governed by the moisture capacity of the air surrounding the piece, i.e., the relative humidity, and the volume of air passing over the ware. Because the moist piece is a porous system, a balance must be struck between loss of moisture at the surface and movement of moisture through the particles from the interior to the surface. If liquid loss at or near the surface exceeds liquid movement from interior pores, differential shrinkage can result in cracking or warping.

Defects. Drying defects can originate wherever there are discontinuities in the formed piece. *Cracks* can develop at these points during shrinkage associated with drying. Clay particles tend to orient with their long dimension normal to the direction of pressure. Because shrinkage is least in the direction parallel to particle orientation and greatest in the normal direction, solid-cast test bars made in an open mold tend to *warp* on the exposed face away from the oriented mold-face layer. Similarly, frictional forces in extrusion force moisture into the interior and cause orientation of clay particles parallel to the direction of extrusion. Differential shrinkage and excessive moisture gradient are leading causes for cracking and warping.

Even though a moist object has a uniform distribution of moisture, it warps unless it is evenly dried. A wall tile, for example, placed on a plate so that one face is protected from air flow and evaporation warps. During initial drying, an originally uniformly moist piece can develop a moisture gradient through (1) loss of moisture at the surface or (2) uneven heating of the pore water, which lowers its viscosity at the warmer surface but not in the cool interior.

Some operations subsequent to slip casting can cause livering (dilatant consolidation) in one part of the cast and not in adjacent areas, resulting in a lower moisture-release rate and lower shrinkage in the livered area than in the adjoining portion.

Even after defect sources originating in the forming operation have been eliminated, the problem of removing the moisture without rupturing or warping the ware remains. This objec-

tive is achieved through the techniques of *humidity drying* that (1) lower the viscosity of the water, (2) uniformly warm the pore water without causing differential shrinkage, and (3) re-move the water economically with respect to both time and fuel consumption.

The principle of humidity drying involves (1) heating greenware all the way through in a saturated atmosphere, (2) reducing the humidity as fast as possible without stressing the ware, and (3) once shrinkage ceases, raising the temperature and reducing the humidity to zero relative humidity [20, pp. 82 – 84].

Drying Methods. Drying methods fall into two classes: (1) convection and (2) radiation [124], [125]. *Convection methods* circulate warm air around the ware being dried, the warm air serving the dual purpose of heating the pieces and removing moisture by convection circulation.

A *simple tunnel dryer* requires passage of cars loaded with ware through the dryer, while heat is supplied by steam coils underneath, hot air from a heater, or waste heat from kilns. Drying is likely to be uneven from top to bottom of the load in tunnel dryers, but cross-circulation of heated air by fans or jets improves uniformity in such dryers.

Controlled humidity can also be attained in a tunnel dryer by introducing the moist ware at one end into hot moist air sent in originally as warm dry air at the exit end, the *counterflow method.* The dry air picks up moisture in its passage over the loaded cars, thus becoming saturated with moisture. Sometimes an auxiliary heating unit is located at the ware entrance end. In this way the moist ware is heated uniformly with no initial loss of moisture or shrinkage. As the ware moves toward the exit end, the surrounding air becomes progressively cooler and drier. The exterior parts of the pieces are reduced in temperature, owing to evaporative cooling, while the interior remains warmer: in a clay mass with a temperature gradient moisture moves toward the cooler part [126].

A refinement of the counterflow method has the *tunnel divided into sections,* each with its own independent heat and humidity controls. The ware enters a hot, saturated zone and is warmed without moisture loss. The car then passes through progressively drier zones, each held at constant but higher temperature.

A *chamber humidity dryer* operates by stages on a stationary load. Hot, saturated air warms the ware. When the chamber is at a uniformly high temperature, drier and slightly cooler air is passed through the load and evaporation occurs. Ware can be dried more quickly if air is directed at right angles than if it is blown parallel to a surface [127]. For small, simple shapes, a method termed *jet drying* is sometimes used: air is blown in a definite pattern at right angles to the surface of the piece, thereby saving fuel, space, and time [11, p. 167].

Radiation methods involve transfer of heat to the moist ware as infrared or high-frequency radio waves. The principle of heat transfer by radiation methods is being employed increasingly in the drying of ceramic ware. Over 40 years ago, infrared lamps were used to dry large slip-cast units in open settings while ordinary air movement was used to remove the moisture. Drying times were reduced from 14 days to 12 h [128]. Hotel china from automatic jiggers can be dried sufficiently for removal from the bat in 10 – 15 min [125]. The infrared drying is followed by hot-air jet drying [127].

Finishing. Nearly all ceramic ware, however formed, must be subjected to finishing operations. These may be as simple as removal of casting spares, mold seams, and fins. However, the operation may involve the turning of a foot on a leather-hard cup or an elaborate turning of a high-tension insulator from a 1-m-diameter extruded blank. The term *trimming* means the shaving away of seams from a cast piece or cutting off the casting spare. *Fettling* refers to removal of fins, mold seams, and rough edges from dry, or nearly dry, ware.

4.5. Firing Ceramic Products

The terms used to describe the densifying processes that occur during heat treatment of ceramic items can be confusing. The expression *sintering* is used [13, p. 232] to describe a process by which a substance is bonded together, stabilized, or agglomerated by being heated to a point close to, but below, the melting point.

Vitrification is defined as a progressive reduction of the pores of a ceramic piece as a result of heat treatment [1, p. 202], but says nothing about formation of a liquid, glassy phase. However, the dictionary definitions [129] dealing with the terms vitreous, vitrify, vitrification, and vitrifiable all center on some aspect of glass. Burke [130] notes that the word sintering is generally used in referring to processes that assume no liquid is formed during heat treatment.

Here the term *densification* [19, chap. 7] is applied to processes where removal of pores from a ceramic product by heat treatment can take place either by formation of a glassy phase or by solid-state material transport, or both. However, traditional ceramics and advanced ceramics are described separately because clay-based products invariably involve development of some glass, whereas advanced ceramic products nearly always involve solid-state reactions.

4.5.1. Firing Traditional Ceramics

The main reactions occurring in the course of heating a clay-based product to maturity are summarized in Table 9 [11, p. 267]. As shown by the expansion curves in Figure 6 [131], in the initial stages of firing clay-based bodies, there is an expansion to a peak at ca. 600 °C with a small inflection (the quartz inversion) at 573 °C, followed by a gradual drop up to ca. 775 °C. After remaining level to ca. 850 °C, a rise in expansion follows to a peak at ca. 900 °C. Shortly thereafter, the bodies begin a decided contraction. Above 1300 °C (not shown in Figure 6) irreversible thermal expansion occurs in all types of clay-based bodies [132], and this expansion must be taken into account in devising firing schedules [131]. The expansion on heatup of body A, which contains pyrophyllitic South American clays, is much greater than that of body B, a U.S. clay body containing no pyrophyllite.

Shrinkage and porosity changes with increasing temperature vary, depending on the body composition and the porosity at maturity. Vitrification/shrinkage curves can be shifted downward in the maturing range while their configuration and the forming properties of the bodies can be maintained by the judicious selection of auxiliary fluxing constituents [133]. An understanding of those factors affecting the nature of the glassy phase permits ready control of the pyroplastic deformation characteristic of large clay body units [134] without altering the unfired forming properties of the body.

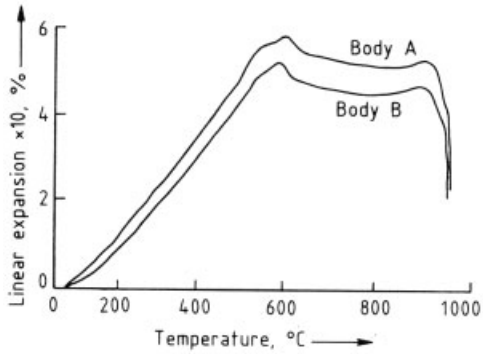

Figure 6. Irreversible heatup thermal-expansion traces for vitreous sanitary ware bodies containing pyrophyllite and sericite (A) and no pyrophyllite and little mica (B)

4.5.2. Densification of Advanced Ceramic Products

The densification of formed advanced ceramic items is generally referred to as sintering. Sintering is, in simple terms, removal of voids (pores) of the formed piece, accompanied by shrinkage. Criteria to be met before sintering can take place are (1) an available means for material transport and (2) an energy source to initiate and promote material transport. Diffusion and viscous flow are transport mechanisms. Heat is the primary source of energy, functioning with particle – particle contact and surface tension to produce energy gradients [19, pp. 217 – 223].

Vapor-Pressure Sintering. Difference in vapor pressure as a function of surface curvature provides the driving energy. Material goes from particle surfaces with a positive curvature radius and high vapor pressure to contact regions with negative curvature radius and much lower vapor pressure. The smaller the particles, the greater the driving force. While vapor-phase sintering bonds particles, it does not eliminate pores.

Diffusion Sintering. Diffusion can be the movement of atoms or vacancies along a surface or grain boundary through the body. Only

Table 9. Reactions occurring in firing clay bodies

Temperature, °C	Reactions in the course of firing
$\leq$ 100	free moisture removed from the piece
100 – 200	loss of adsorbed moisture
200 – 400	gradual loss of H_2O from halloysite and montmorillonite, pyrophyllite and fine sericitic mica begin decided expansion
400 – 700	organic matter oxidized, breakup of clay mineral structures, pyrophyllite starts sharp expansion
573	quartz inversion
700 – 950	pyrophyllite attains maximum expansion, spinel forms in clays
950 – 1000	muscovite structure destroyed, γ-Al_2O_3 or mullite forms
1000 – 1100	mullite, $3\,Al_2O_3 \cdot 2\,SiO_2$, forms from clay
1100 – 1200	feldspars melt, clay and cristobalite dissolve, porosity decreases, shrinkage increases rapidly
>1300	glass increases, ware expands, absorption increases, strength decreases

grain boundary or body diffusion results in sintering. The driving force is differential free energy or chemical potential between free surfaces and contact points of adjacent particles. Finer particles sinter more rapidly and at lower temperatures than coarser particles. Uniformity of particle shape, size, and distribution governs the uniformity of the final product.

Liquid-Phase Sintering. The main densification mechanism for most silicate systems is liquid-phase sintering. The sintering occurs best where the liquid phase thoroughly wets the solid grains at sintering heat. Capillary pressure in narrow pores between particles may be ≥ 7 MPa. Because small particles have higher surface energy and form smaller pores, there is more densification driving energy than for compacts of larger particles. Temperature strongly affects sintering; generally, small increases in temperature cause significant increases in the amount of liquid. In some cases, this is desirable, but in others excessive grain growth and fire distortion occur. The amount of liquid at a given temperature can be predicted from phase equilibrium diagrams.

Hot Pressing. Hot pressing resembles sintering except that temperature and pressure are applied at the same time [135]. Pressure speeds densification by increasing particle packing and by stressing points of contact. The densification energy can be increased 20-fold by applying pressure. *Hot isostatic pressing* can be done with special heat-treating equipment [136] and provides results superior to those attained with conventional hot pressing. Hot pressing injection-molded items gives results superior to those obtained with simple isopressing/sintering, slipcasting/sintering, or injection molding/sintering. The Weibull probability plot of Figure 7 shows the superior uniformity and higher strength of hot-pressed, injection-molded products.

4.6. Kilns and Firing Conditions

The furnaces in which ceramic products are heat-treated are usually termed *kilns.* Kilns, depending on the manner of operation, can be termed periodic (intermittent) or tunnel (continuous):

1) Intermittant firing kilns
   - stationary periodic
   - lifting charge
   - lifting kiln
   - moving charge
   - moving kiln
2) Continuous firing kilns
   - chamber
   - conveyor belt
   - roller slab
   - muffled tunnel
   - direct-fire tunnel

Periodic kilns are heated and cooled in accordance with prescribed schedules that differ with the kind of product. A *tunnel kiln* has temperature zones held at specific temperatures through which kiln cars (or other supports) are passed to provide the specified time – temperature cycle. Tunnel kilns are adapted to firing one type of body in long runs, whereas periodic kilns can be adapted to a variety of products.

Periodic kilns can be heated by electrical elements or fired with gas or oil. Traditionally,

where hot gases are involved, heating is accomplished by having the combustion products pass through the load of ware, either upward (updraft kiln) or downward (downdraft kiln), before going out in a flue. The disadvantages of such kilns is that (1) they must be loaded and unloaded by hand, (2) there is a long cooling period, and (3) the entire kiln must be reheated in the next firing.

4.6.1. Modern Periodic Kilns

Elevator kilns are of three types: (1) ware to be fired is placed on a refractory protected car that is pushed in position under a suspended kiln that is then lowered over the car (a top-hat kiln); (2) the car with its ware is elevated into the kiln that is permanently fixed; and (3) an elevated kiln is moved horizontally over a series of cars and placed over any one of them, as desired.

Shuttle kilns are positioned permanently. One end has a movable door. Ware is loaded onto a car that is then run on rails into the kiln, the door is closed, and the ware is fired. A variant is the envelope kiln, which is rolled over and encloses ware placed on a permanent hearth.

4.6.2. Tunnel Kilns

Tunnel kilns are refractory chambers, sometimes 90 – 100 m long, through which ware is moved to achieve gradual heating and cooling. The entry section is the preheat zone, the middle section is the firing zone, and the exit portion is the cooling zone.

Cooling air is blown into the cooling zone, is heated, and moves into the firing zone, where it improves combustion and preserves an oxidizing atmosphere. Combustion gases from the firing zone are conveyed into the preheat zone to heat and dry the ware.

Refractory-topped cars riding on insulated rails carry the ware into and through tunnel kilns. Pushing is done on a prescribed schedule, expressed in terms of "cars per 24 hours." Some smaller tunnel kilns have positively rotated refractory (alumina) rollers, on which refractory slabs carry the ware. Other tunnel kilns use sled hearths, which are intermediate between cars and roller slabs, for smaller fast-fired products.

Advances in ceramic-fiber technology [136], [137] have provided alumina, silica, and kaolin fiber products for use in place of high-density castables for kiln cars and for insulating replacements for higher-density brick. This has made possible kiln designs that greatly reduce fuel consumption and permit faster firing of ware [138].

4.6.3. Advanced Ceramics Furnaces

Because advanced ceramic products often have special sintering requirements, the furnaces differ from those of ordinary ceramic kilns and require new furnace technology [139]. A common requirement is the need for total control of kiln atmosphere, as well as control of temperature and time scheduling. Atmospheric control can be achieved by sintering in a vacuum furnace or an autoclave. Initial air can be pumped out so that contaminants are vaporized and evacuated. The atmosphere around the ware can then be controlled with respect to composition and pressure.

Separate furnaces may be needed to eliminate lubricants and volatile binders. However, delubing can be done at low temperature, along with degassing, and the ware can be fired in an inert atmosphere on a controlled schedule to a required temperature.

4.6.4. Kiln Atmosphere

The atmosphere has a profound influence on the fired properties of clay-based ceramics. If there is enough O_2 to permit the piece to absorb some, the atmosphere is regarded as oxidizing. An oxidizing atmosphere helps eliminate carbon and converts salts to oxides. A low-oxygen atmosphere reduces multivalent ions to their lowest positive state, thus causing color and other changes.

Sulfur-bearing fuels provide SO_2, which is harmful to body and especially to glazes of clay-based ceramics. The ware must be protected by saggers (refractory boxes) or by keeping the combustion gases away from the work with a refractory wall (muffles), through which heat is radiated.

Electric kilns of all types [11, pp. 305 – 306], which avoid contamination from burned fuels,

are used widely in Europe. In the United States, such kilns are mainly used for decorating, special ceramic products, and wall tile. The heating elements are nichrome, kanthal, or silicon carbide. Nichrome (a nickel – iron – chromium alloy) elements are used for decorative materials firing, kanthal (iron) for intermediate-temperature kilns, and silicon carbide for high-fire kilns.

4.6.5. Fired Ware Finishing

Postfiring processes fall in the category of finishing. Finishing may include grinding to size and removing kiln marks (grains of kiln dirt). Technical ceramics must be examined for flaws.

5. Glazes and Glazing

Glazes are applied to clay-based ceramic products to provide a shiny, generally smooth surface that seals the body. The surface may be either matt or bright. Glazes resemble glass in structure and texture, but have greater viscosity in the molten state. Glaze adheres strongly and uniformly to the ware. Application of glaze suspension to ware is by spraying or dipping.

Glazes can be made for maturing from ca. 600 °C up to ca. 1500 °C, depending on the items to which they are applied. If necessary, surfaces can be made resistant to various corrosive liquids and gases. Semiconducting glazes can be prepared for electrical porcelains.

5.1. The Nature of Glazes

Glasses and glazes used by ceramists are normally combinations of oxides (Table 10). Oxides that form glasses by themselves are termed *network formers;* by Zachariasen's rules [11, p. 130], SiO_2, B_2O_3, and P_2O_5 should, and do, form glasses. The holes in the network are filled by network *modifiers,* which weaken the bonds. Such modifiers are usually Na_2O, K_2O, CaO, and MgO. Generally speaking, the more modifier present, the lower the glass viscosity and chemical resistance. Oxides such as Al_2O_3, PbO, ZnO, ZrO_2, and CdO can enter the network by replacing some Si^{4+} or B^{3+}. These are the *intermediate glass formers.* The relative single-bond strengths of the oxides correlate with glass formation: network formers have high bond strengths, modifiers have low bond strengths, and intermediates fall in between.

Table 10. Oxide glasses [140]

Oxide	Oxidation state	Coordination number	Bond strength, relative
Network formers			
SiO_2	4	4	106
B_2O_3	3	3	119
Intermediates			
Al_2O_3	3	4	90
Al_2O_3	3	6	60
ZnO	2	2	72
ZnO	2	4	36
PbO	2	2	73
PbO_2	4	6	39
Modifiers			
Na_2O	1	6	20
CaO	2	8	32
Substitutions for Na_2O			
K_2O	1	9	13
Rb_2O	1	10	12
Cs_2O	1	12	10
Li_2O	1	4	36
Substitutions for CaO			
MgO	2	6	37
BaO	2	8	33
SrO	2	8	32

The conventional representation of glaze formulas is the Seger convention, by which the formula is expressed with the fluxing oxides R_2O and RO, including PbO and ZnO, adding up to unity on a molar basis. The Al_2O_3 and SiO_2 plus B_2O_3 are listed as separate items, also on a molar basis, in the following manner:

$$1.0\,(R_2O + RO) \cdot x\,Al_2O_3 \cdot y\,(SiO_2 + B_2O_3)$$

Glazes can be classified as raw glazes and fritted glazes. In the case of *raw glazes,* the oxides are introduced in the form of compounds or minerals, such as feldspar, which melt readily and act as solvents for the other ingredients. A *frit* is a prereacted glass containing ingredients, such as Na_2CO_3, which are soluble in water. Fritting fixes the desired oxide in a relatively insoluble form. Fritted glazes are used for whitewares. Frits may form only a part of the whiteware glaze formula, or the glaze may be composed entirely of frit [11, p. 190].

Raw Glazes. Raw glazes include (1) porcelain glazes, (2) Bristol glazes, (3) raw lead glazes, (4) raw leadless glazes, and (5) slip glazes. According to the Seger convention the basic *porcelain glaze* is expressed as follows [141]:

0.3 K_2O 0.4 Al_2O_3 4.0 SiO_2
0.7 CaO

It matures as a bright glaze in the region of Orton pyrometric cones 8 – 10 (1236 – 1285 °C). When Al_2O_3 is raised to 0.5 and SiO_2 to 5.0, a bright glaze is obtained at cone 12 (1306 °C). At SiO_2 7.0 and Al_2O_3 0.5, the maturity of a bright glaze rises to cone 14 (1388 °C), whereas at SiO_2 8.0 and Al_2O_3 1.0, maturity occurs at cone 16 (1455 °C). Increasing the Al_2O_3 and SiO_2 content produces semimatt and matt low-gloss surfaces.

Bristol glazes are raw glazes containing ZnO, which are used on terra-cotta clayware and sometimes on stoneware items [142]. A typical Bristol glaze falls near the following molar composition:

0.40 K_2O 0.5 Al_2O_3 3.4 SiO_2
0.35 CaO
0.25 ZnO

Maturity occurs at cones 4 – 8 (1168 – 1236 °C).

Raw lead glazes [143] are used only on artware, never on commercial ware, owing to health hazards from soluble lead. A typical bright glaze maturing at ca. cone 05 (1031 °C) has roughly the following composition:

0.55 PbO 0.23 Al_2O_3 1.55 SiO_2
0.36 CaO
0.09 Na_2O

Lead-free glazes, or leadless glazes, are designed to provide lower maturity than true porcelain glazes without the use of lead oxide. A typical glaze has the following composition [144]:

0.2 K_2O 0.3 Al_2O_3 3.0 SiO_2
0.3 SrO
0.1 CaO
0.4 BaO

Glazes built around this composition have an excellent maturing range as low as cone 03 and up to cone 9 (1086 – 1260 °C), but lack the covering power and brightness of lead glazes.

Slip glazes [145] are natural clays having the following approximate composition:

0.20 K_2O 0.60 Al_2O_3 4.00 SiO_2
0.45 CaO 0.08 Fe_2O_3
0.35 MgO

Such clays are used for artware glazing and often for high-tension electrical porcelain insulators. Slip clay glazes have a maturing range from ca. 1200 °C to 1300 °C.

Frits. Because B_2O_3 and most borates are soluble in water, the B_2O_3 must be added in a frit. A typical lead-free borate frit has the following composition:

0.69 CaO 0.37 Al_2O_3 2.17 SiO_2
0.19 Na_2O 1.16 B_2O_3
0.12 K_2O

Although lead bisilicate [*11120-22-2*] ($PbO \cdot 2\,SiO_2$ or $PbSi_2O_5$) is relatively insoluble in water, lead is normally introduced into commercial glazes in fritted form. A more complex frit might have a composition such as the following:

0.50 PbO 0.10 Al_2O_3 2.70 SiO_2
0.30 Na_2O
0.20 K_2O

Boric oxide and lead oxide often appear together in low-temperature glaze. A typical composition is the following:

0.50 PbO 0.10 Al_2O_3 2.70 SiO_2
0.10 Na_2O 0.60 B_2O_3
0.20 K_2O
0.30 CaO

Special Glazes. Low-expansion glazes are required by zircon, cordierite, and low-expansion lithium silicates [146]. Recommended glazes for zircon porcelain in the range of cones 10 – 14 (1285 – 1400 °C) approximate the following composition:

1.00 RO 0.6 – 0.7 Al_2O_3 9 – 11 SiO_2

Semiconducting glazes are used to remove charges from surfaces of electrical insulators. This property can be produced with a high concentration of Fe_2O_3 in the glaze, crystals of ZnC_2O_4, or activated SnO_2 [11, p. 197].

Glaze opacity is obtained by mill additions of zirconium-type opacifiers [147].

Lithium oxide additions increase the hardness of commercial glazes [148]. High-compression glazes resist scratching. Crystalline glazes grow large crystals in the firing and cooling of the glaze. NORTON [149] demonstrated that crystals grow when temperatures of nucleation and crystal growth do not overlap. For example, proper control of the heating schedule allowed the growth of large willemite crystals.

Salt glazing is an old method for glazing stoneware. The glaze is formed by throwing damp common salt into the kiln during the sintering stage of firing. The NaCl decomposes to form Na_2O and HCl, the Na_2O combining with the Al_2O_3 and SiO_2 of the body to form complex silicates. BARRINGER [150] found the limits of the $Al_2O_3 : SiO_2$ ratio within which it is commercially possible to produce good salt glaze to be 1 : 4.6 – 1 : 12.5.

5.2. Preparation of Glazes

Prepared frit, clay, and materials not incorporated in the frit are ground generally in ball mills with water. Grinding is followed by sieving and magnetic treatment.

The ball mills are large steel drums lined with quartzite or ceramic blocks. Flint pebbles of various size grades are frequently used. High-density (usually high-alumina) balls or cylindrical rods have found favor for this operation. The higher-density media reduce grinding times and lessen contamination from pebble wear. The media, batch, and water occupy ca. 60 % of mill volume. Media weight is around 3 times that of batch. Water content runs 30 – 50 %, depending on the material being processed. Mill speed varies with mill size: the smaller the mill, the faster the rotation.

The practice of grinding to a stated percentage of residue remaining on a test sieve does not take into account differences in particle-size distribution brought about by changes in media size and size distribution and linear wear. PHELPS [151] has shown that apparently minor differences in glaze and enamel slip particle-size distribution can cause marked differences in slip rheology.

While fine grinding improves glaze brightness, hardness, and chemical stability, grinding too fine can result in crawling (parting of the glaze, which leaves bare spots) or peeling (flaking away) of the glaze. Changes in the soluble materials content of the water can lead to difficulties in consistency and fineness of glazes.

5.3. Glaze Application

The prepared glaze slip is adjusted with additives designed to control consistency and adhesion. Hand dipping of prefired (bisque) ware in glaze slip was the general practice prior to 1920. However, spraying then became common. Automatic spraying on a conveyor line is now used. Tiles are glazed by spraying or by passing them unter a falling sheet of glaze slip.

Aside from grinding-induced differences, the nature of the materials can result in glaze settling. Suspending agents, such as bentonite or organic agents, are then used. Flocculants, such as $MgSO_4$ or $CaCl_2$, can be used.

6. Glass (→ Glass)

Glass has been defined in simple terms as a fusion product of an inorganic material that has been cooled to a reasonably rigid, noncrystalline state [2, p. 42]. Objects made of glass are simply called glass, although specific kinds of glass are qualified by types, such as flint, barium, lead, container, or window glass.

As shown by Table 10 glasses consist of the following types of oxides: (1) network formers (form glasses by themselves); (2) network modifiers (alkali-metal ions and alkaline-earth-metal ions); (3) intermediate glass formers (partial substitutes for network formers).

A fairly high PbO content (10 – 45 %) is characteristic of *flint glasses.* Low-expansion, chemically stable glasses contain significant levels of B_2O_3 (6 – 12 %). Optical glasses contain variable percentages of lead and barium oxides.

Soda – lime – silica glass represents the major proportion of commercial glass [12, chap. 13]. It is made by melting more or less pure silica sand by fluxing with soda (Na_2O) and stabilizing with CaO or $CaO \cdot MgO$.

Container glass is a typical soda – lime – silica glass containing smaller percentages of materials having special functions. Alumina (from feldspar or nepheline syenite) helps chemical

stability, sodium nitrate functions as an oxidizing agent, and arsenic is a fining agent (for elimination of bubbles and undissolved gases). Clear glass requires low Fe_2O_3 content, whereas high Fe_2O_3 materials are used in colored glasses [152]. Manganese blanks out the green color of iron.

Container glass production is usually large scale. Batches are weighed automatically, blended, and conveyed to glass melting tanks (refractory containers for melting). Tanks hold 180 – 275 t of glass. Batches are fed into tanks as layers 15 – 20 cm thick and melted by heat from side burners fired with oil or gas. Exhaust heat passes into checker chambers below the burner ports; flow is reversed 3 – 4 times per hour. Combustion air passes through the heated checkerwork.

Large refractory blocks made of fusion-cast Al_2O_3 or Al_2O_3– ZrO_2– SiO_2 mixtures function as side walls and end walls, while the bottom is $ZrSiO_4$ and the roof is SiO_2. A container glass tank may have melting areas of 90 – 150 m^2 with glass depths of 1 – 1.5 m. A campaign (working life before major repair) of a tank may be 4 – 5 years.

Sheet glass tanks, holding 1200 – 1500 t, are much larger than container glass tanks. Such a tank can supply 180 – 275 t every 24 h.

Optical glass tanks are much smaller, with outputs of perhaps 40 – 200 kg every 24 h.

Forming is by pressing, vacuum, or blowing (using air pressure to transform the gob, i.e., mass of molten glass, into a hollow piece). Otherwise, a sheet is formed by pulling (drawing) a continuous sheet of molten glass from the tank and passing it through a flattener or roller. Float glass involves a process in which the sheet floats on a bath of molten tin with heaters above and in the bath; the glass settles to an even ribbon and is allowed to cool slowly. Sheet glass requires polishing and grinding for use as plate glass, whereas float glass does not.

Fiberglass takes two forms: (a) continuous thread for textiles or (b) discontinuous fiber for insulation, filtering, or reinforced fiberglass. Glass marbles are fed continuously into a melting chamber, and filaments are pulled through platinum spinnerets. Discontinuous fibers are blown by striking a molten stream of glass with a high-velocity steam jet. Such glass generally has a lower viscosity than textile fiberglass. Fiberglass, with its high surface area, also must be chemically stable.

Certain glasses [153] and glass-bonded ceramics [154] have application as refractory substances. Included are vitreous silica, high-silica glasses, aluminosilicate glasses, aluminate glasses, mullite glass, and barium feldspar glass ceramics.

7. Refractories (→ Refractory Ceramics)

Refractory materials are essential to the manufacture of all forms of ceramic products, including refractories themselves. Table 11 [1, pp. 13 – 15] gives classes and types of refractory brick. Insulating firebrick are rated in eight progressively more refractory groups where reheat shrinkages are not more than 2 % at testing temperatures of 845 °C (group 16), 1065 °C (group 20), 1230 °C (group 23), 1400 °C (group 26), 1510 °C (group 28), 1620 °C (group 30), 1730 °C (group 32), and 1790 °C (group 33) [1, p. 104].

Table 11. Classes and types of refractories

Class	Type	PCE∗	MOR∗∗, MPa
Fireclay	Super duty	33	4.14
	High duty	31.5	3.45
	Medium duty	29	3.45
	Low duty	15	4.14
High Al_2O_3	50 % Al_2O_3	34	
	60 % Al_2O_3	36	
	70 % Al_2O_3	36	
	80 % Al_2O_3	37	
	90 % Al_2O_3		
	99 % Al_2O_3		

∗ PCE pyrometric cone equivalent.
∗∗ MOR modulus of rupture.

Special refractories include zircon ($ZrSiO_4$), zirconia (ZrO_2), silicon carbide (SiC), chromic oxide (Cr_2O_3), and graphite (C). Refractory specialties include nonformed products such as mortars, castables, plastics, and ramming mixes. Basic refractories include the chrome brick, chrome – magnesite brick, and magnesite brick used in basic oxygen steelmaking [155].

8. Abrasives (→ Abrasives)

Several natural minerals are employed as abrasives for cutting, grinding, and polishing. These include quartz, garnet, corundum, emery, and diamond. Manufactured abrasives include boron carbide, boron nitride, diamond, fused alumina, silicon carbide, titanium carbide, tungsten carbide, and zirconium silicate. Abrasive products include loose grains, wheels, coated abrasives, and grinding pebbles.

9. Cement (→ Cement and Concrete),

Cement is a synthetic mineral mixture (clinker) that when ground to a powder and mixed with water forms a stonelike mass, and is thus a ceramic product [156]. A primary requirement for cement manufacture is a source of CaO; this can be limestone, oyster shell, slag, etc. Also necessary is a source of Al_2O_3 and SiO_2, most commonly clay and, where needed, a quartz rock or sand.

Processing involves (1) grinding of rock material, (2) blending of ground materials to a desired chemical composition or slurry blending of the powders, (3) burning the blended material to form a clinker, (4) blending the clinker and gypsum, and (5) grinding the gypsum – clinker. Grinding can be accomplished by dry grinding – ball milling, rod milling, roller, race, tube mill – with air classification. Oversize is recycled.

The American Society for Testing and Materials has listed specifications for eight types of cement in accordance with chemical composition and physical requirements [157].

10. Properties of Ceramic Materials and Products

Ceramic products have relatively high strength associated with brittle fracture, high thermal stability, and low electrical conductivity. These properties are related to structure and depend on the size and arrangement of multiphase polycrystalline constituents and the glassy phase. Size, type, and distribution of pores must be considered because pores affect strength, thermal expansion, heat insulation, corrosion and weathering resistance, and electrical properties.

Table 12 shows mechanical properties of a number of representative ceramic products. In all cases, there is a characteristic direct transition from a small elastic deformation, with no or small plastic deformation, to fracture. Irreversible deformations from above the elastic region up to fracture may be due to viscous processes within the particle structure.

Table 12. Mechanical properties of ceramic materials

Type of material	Compressive strength, MPa	Flexural strength, MPa	Modulus of elasticity, GPa
Solid brick	10 – 25	5 – 10	5 – 20
Roof tile	10 – 25	8 – 15	5 – 20
Steatite	850 – 1000	140 – 160	1 – 3
Silica refractories, 96 – 97 % SiO_2	15 – 40	30 – 80	8 – 14
Fireclay refractories, 10 – 44 % Al_2O_3	10 – 80	5 – 15	20 – 45
Corundum refractories, 75 – 90 % Al_2O_3	40 – 200	10 – 150	30 – 120
Forsterite refractories	20 – 40	5 – 10	25 – 30
Magnesia refractories	40 – 100	8 – 200	30 – 35
Zircon refractories	30 – 60	80 – 200	35 – 40
Whiteware	30 – 40	20 – 25	10 – 20
Stoneware	40 – 100	20 – 40	30 – 70
Electrical porcelain	350 – 850	90 – 145	55 – 100
Capacitor ceramics	300 – 1000	90 – 160	

Important ceramic oxides have high melting points (°C):

Al_2O_3	2050	Fe_3O_4	1600
MgO	2800	FeO	1360
CaO	2600	$MgO \cdot Al_2O_3$	2135
SiO_2	1780	$2\,MgO \cdot SiO_2$	1890
ZrO_2	2700	$3\,Al_2O_3 \cdot 2\,SiO_2$	1810
Cr_2O_3	2265	$ZrO_2 \cdot SiO_2$	1775

Table 13. Thermal expansion coefficients, $10^{-6}\,K^{-1}$, of components of ceramic materials

Component	Temperature ranges, °C		
	20 – 300	20 – 900	20 – 1400
Silica	36.5	15.5	10.0
Magnesite	10	12.7	14.2
Chrome magnesite	8.3	9.4	10.5
Chromite	8.3	9.1	9.5
Corundum 99	7.3	7.2	*
Corundum 90	4.3	5.2	6.5
Zircon	2.7	3.8	**
Sillimanite	3.3	4.4	4.8
Silicon carbide	1.6	3.5	4.4

* 7.9 over the range 20 – 1200 °C.
** 4.5 over the range 20 – 1200 °C.

Table 13 gives thermal expansion coefficients for a number of ceramic product constituents. Because high melting points generally correlate with low thermal expansion, these materials generally have low coefficients of thermal expansion. The anisotropic structure results from polycrystalline mixed phases and varying amounts of glassy phases, which explains the relatively poor thermal-shock resistance.

Most ceramic products have thermal conductivities lower than platinum, for example, but higher than, for example, insulating firebrick or organic polymers. Table 14 shows thermal conductivity coefficients for a number of refractory brick products [4, p. 942], and Table 15 provides thermal conductivity ranges for a number of electrical porcelain types [19, p. 47].

Table 14. Thermal conductivity of refractory brick

Material	% Porosity	Thermal conductivity, $W\,m^{-1}\,K^{-1}$	
		371 °C	1000 °C
28 % Al_2O_3	22	0.84	1.72
42 % Al_2O_3	19	1.21	1.42
72 % Al_2O_3	22	1.55	1.42
99 % Al_2O_3	24	3.77	2.47
Silica	23	1.34	1.76
Mullite	23	0.92	1.76
94 % MgO	20	6.86	2.76
Chrome magnesite	22	1.72	1.80
Zircon	17	2.76	2.38

Table 15. Thermal conductivity of electrical ceramics at room temperature

Body type	Thermal conductivity, $W\,m^{-1}\,K^{-1}$
Electrical porcelain	0.8 – 1.7
Steatite porcelain	2.1 – 2.5
Cordierite	1.3 – 2.1
Zircon porcelain	4.6 – 5.0
Titania porcelain	2.9 – 4.2
Titanate	3.3 – 4.2

Table 16 gives resistivities for a number of ceramic products that serve as electrical insulators: their resistivities are of the order of 10^{12} – 10^{13} Ω cm – several orders of magnitude higher than for metals [19, p. 47].

Table 17 provides a tabulation of mechanical and thermal properties of materials employed in making advanced ceramics articles. The temperature for which these values are valid is room temperature or somewhat above, unless specified otherwise.

Table 16. Resistivities of some metals, ceramic insulators, and semiconductors at room temperature

Ceramic material	Resistivity, Ω cm
Insulators	
Low-voltage porcelain	10^{12} – 10^{14}
Steatite porcelain	10^{14}
Mullite porcelain	10^{13}
Cordierite porcelain	10^{13}
Zircon porcelain	10^{14}
Alumina porcelain	10^{16}
Silica	10^{19}
Semiconductors	
Silicon carbide	10
Boron carbide	0.5
Ferric oxide	10^{-2}

Chemical stability of nonporous ceramic products in the presence of acids or alkalies is adequate although it decreases as the temperature is increased. Nonporous ceramics can withstand atmospheric effects up to their melting points.

11. Testing Ceramic Raw Materials and Products

A distinction is made between tests made to determine the suitability of raw materials or ceramic products for particular applications and quality control test procedures. Tests for suitability can be very involved, whereas acceptance tests agreed upon between the raw materials supplier and the user or employed by the supplier in mining and refining and the manufacturer in processing can be simple. The characterization concept is considered the basis for determining the suitability of raw materials and products [2], [17, pp. 195 – 249], [65].

11.1. Raw Material and Product Tests

Various ceramic manufacturers and suppliers of raw materials for ceramic products have joined together in "the development of standards on characteristics and performance of materials, products, systems, and services; and the promotion of related knowledge" [2, p. iii].

Table 17. Properties of advanced ceramics [158]

Material	Crystal structure	Theoretical density, Mg/m^3	Knoop or Vickers hardness, GPa	Transverse rupture strength, MPa	Fracture toughness K_{IC}, $MPa \cdot m^{0.5}$	Young's modulus, GPa	Thermal expansion, $10^{-6}\ K^{-1}$	Thermal conductivity $W\,m^{-1}\,K^{-1}$ 400 K	1400 K	Specific heat, $J\,kg^{-1}\,K^{-1}$
Alumina, Al_2O_3	hexagonal	3.97	18–23	276–1034	2.7–4.2	380	7.2–8.6	27.2	5.8	1088
Mullite, $3\,Al_2O_3 \cdot SiO_2$	orthorhombic	2.8		185	2.2	145	5.7	5.2	3.3	1046
Partially stabilized zirconia, ZrO_2	cubic, monoclinic, tetragonal	5.70–5.75	10–11	600–700	8–9	205	8.9–10.6	1.8–2.2		400
Titanium dioxide, TiO_2	tetragonal (rutile)	4.25	7–11	69–103	2.5	283	9.4	8.8	3.3	799
Silicon carbide, SiC	hexagonal (α) cubic (β)	3.21	20–30	230–825 (hot pressed)	4.8–6.1	207–483	4.3–5.6	63–155	21–33	628–1046
Silicon nitride, Si_3N_4	hexagonal (α) hexagonal (β)	3.18–3.19	8–19	700–1000 (hot pressed)	3.6–6.0	304	3.0	9–30		400–1600
Titanium nitride, TiN	cubic	5.43–5.44	16–20			251	8.0	24	67.8*	628

* At 1773 K.

Table 18 lists the volume identifications and subject areas of interest to the several areas of ceramic endeavor [2, pp. vi – vii].

Table 18. Listing of ASTM book of standards (1985) for ceramic-related test compilations

Volume identification	Subject area
03.05	chemical analysis of metals and metal-bearing ores
03.06	emission spectroscopy: surface analysis
04.06	thermal insulation
10.01	electrical insulation, solids, composites and coatings
10.02	electrical insulation (II), wire and cable, heating and electrical tests
10.04	electronics (I)
10.05	electronics (II)
11.01	water (I)
11.02	water (II)
11.03	atmospheric analysis: occupational health and safety
12.01	nuclear energy (I)
12.02	nuclear energy (II), solar and geothermal
14.01	analytical methods – spectroscopy; chromatography; temperature; computerized
14.02	general test methods – nonmetal; lab apparatus; statistical methods; durability
15.01	refractories, carbon and graphite
15.02	glass; ceramic whitewares, porcelain enamels

11.2. Simplified Testing of Clay Body Materials

The tests listed in Table 19 were derived from the characterization concept to serve as control tests by miners and refiners of clay body materials and by manufacturers in acceptance testing and for plant control. Table 19 lists suggested control tests for clays and nonplastics. These tests are designed as surrogate procedures for the methods that require complex, expensive equipment and highly trained personnel. Chemical analysis, particle-size distribution, and mineral constituents govern fired properties and glaze fit of clay-based bodies. A simple test, using two clear glazes (one fitted to a standard kaolin and the other to a standard ball clay) make possible detection of variation in free silica. The presence of specking impurities is shown more strongly by clear glazes than by bisque clay. Deflocculation tests indicate changes in particle size and soluble salts. The solubles test also relates to deflocculation and the rate of cast. Nonplastics should

Table 19. Suggested control tests

China clay	Ball clay	Feldspar	Silica
residue: 100, 200, 325 mesh	residue: 100, 200, 325 mesh	residue: 100, 200, 325 mesh	residue: 100, 200, 325 mesh
solubles: hardness, Cl^-, SO_4^{2-}	solubles: hardness, Cl^-, SO_4^{2-}	particle-size analysis	particle-size analysis
methylene blue index	methylene blue index	body deflocculation	body deflocculation
particle size analysis	particle size analysis	casting cups and bars	casting cups and bars
fired shrinkage and absorption	fired shrinkage and absorption	firing shrinkage and sag bars	firing shrinkage and sag bars
glaze fit	glaze fit	glaze fit and absorption	glaze fit and absorption
fired color, specking	fired color, specking	fusion	
chemical: TiO_2 and Fe_2O_3	ignition loss		
neat clay deflocculation	clay – flint deflocculation		
casting	casting		

be evaluated in a standard body for deflocculation, rheology, and fired properties. A fusion test, compared against standard specimens [159], is indicative of changes in mineral composition and particle size.

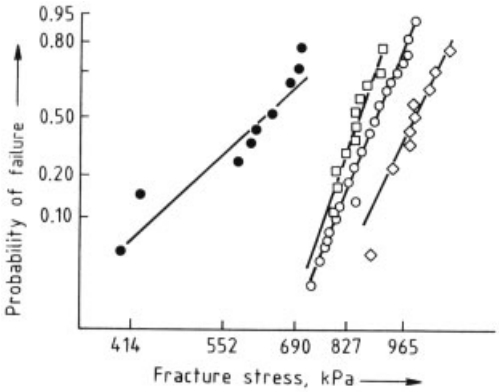

Figure 7. Weibull probability plot showing the effect of forming methods for silicon nitride on the range and level of stress fracture
● Isopressed/sintered; □ Slipcast/sintered; ○ Injection molded/sintered; ◇ Injection molded/hot isopressed theory

11.3. Quality Control of Advanced Ceramics

The degree of quality control needed in a ceramic manufacturing depends on the critical requirements of the application. In advanced ceramic manufacture, most end uses need a specified manufacturing procedure in writing in addition to certification that this procedure has been followed. The more demanding applications must have proof tests, destructive sample tests, and nondestructive inspections of various kinds [19, chap. 9].

A popular method for characterizing flaw distribution is the use of the Weibull approach [160], based on the weakest link theory. This assumes that a given volume of a ceramic material under uniform stress will fail at the worst flaw. Data is shown as a probability of failure F (a function of stress σ and volume V or area A under stress) plotted against σ. The probability F can be estimated from $Fn/(N+1)$, where n = rank of sample and N = total number of samples. When plotted on a log normal grid, as in Figure 7, the result is a straight line.

The data of Figure 7 show the effect of forming technique on the uniformity of an advanced ceramic material, with respect to stress fracture. A vertical plot at a particular fracture stress indicates absolute uniformity. Slip cast and injection-molded parts are much more uniform than simple isopressed items after sintering. Slip-cast pieces are more uniform than injection-molded parts. Hot isopressing of injection-molded pieces gives a significant increase in fracture stress over sintered injection molding, but at about the same level of uniformity.

12. Economic Aspects

The cost of a ceramic product at its point of manufacture depends on such factors as (1) cost of raw materials; (2) cost of energy for processing, forming, firing, and finishing; (3) capital cost and maintenance; and (4) cost of labor. The impact of the various factors necessarily vary with the product being made.

Raw material cost involves the cost of mining, refining, and transporting a given commodity to a point of ceramic manufacture. Equipment cost depends on the country where it is manufactured. However, countries with highly developed, technologically advanced manufacturing capability, coupled with moderate wage scales, can compete in domestic markets of countries having comparable manufacturing ability but higher wage levels. Countries with skilled workers and low wage scales can export products that require detailed handwork and can undersell domestic manufacturers with high wage costs. For example, countries of the Far East, having highly skilled but low-wage labor, make and deliver high-quality tableware to European and American markets at prices lower than domestic potteries can [161].

Other high-quality, less labor-intensive products, such as vitreous plumbing ware, are less vulnerable to imports. However, as developing countries gain expertise and improve quality, it is possible that sanitary ware imports may take over some of the sanitary ware markets in developed countries.

In terms of unit weight, the cost of making and delivering a ceramic product depends on the nature of the product and can vary widely. The fine ceramic products of Table 1, in general, cost more to make per unit weight than the coarse ceramic products of Table 2. Furthermore, a vitreous china plumbing fixture costs far less per unit weight than a highly decorated fine china platter. By the same token, a building brick costs far less per unit weight than a high-alumina refractory brick of comparable volume.

Physical and economic geographical factors are likely to limit exports of relatively heavy products from countries where mountains make road or rail transport difficult. The distance from market is a factor in the cost of transport. In the United States, where rail, water, and road transport are reasonably good, the cost of transporting desirable raw materials from sources in the southeast to points on the West Coast can exceed the FOB cost of the raw material. This has encouraged use of local, less desirable raw materials. In glass manufacturing, plants will seek out closest possible sources of silica sand and glass feldspar or nepheline syenite. In smaller countries, transport costs are of less consequence.

Specialty products, such as high-tension electrical porcelain insulators, have stringent mechanical and dielectric strength requirements. Careful processing of controlled compositions is the key to superior quality. Those countries with lower wage scales and good technical capabilities can often meet or exceed these requirements at a lower or equivalent cost than in higher wage North American markets.

Manufacture of structural clay products, such as brick, clay pipe, and tile, has been historically located as close as possible to a given market area, consistent with access to cheap raw material, fuel, and labor. However, structural clay product companies are increasing in size and becoming correspondingly more conscious of the necessity for controlling raw materials and servicing an expanding market [162]. Research is being done on forming methods to reduce losses and improve quality [163]; automated plants are increasingly more common.

The economic health of *traditional, clay-based ceramic* manufacture is closely tied to the state of the economy, especially the level of building [164–166]. Traditional ceramics continue to evolve from an art to a science as more use is made of the findings of materials science in better control over raw materials, body preparation, and automated forming. As noted earlier, improvements in kilns and body composition will permit faster firing, lower fuel and refractory costs, and lowered losses.

By way of contrast, *advanced ceramics* is a rapidly developing field of large, although somewhat undefined, potential for growth. Market forecasts for high-technology ceramics vary from one to another with respect to absolute size; although starting bases differ, growth rates are more consistent.

Bowen [167] estimated sales rising from 1.5×10^9 to 7.0×10^9 dollars in the period 1980 – 1995 at an annual rate of 11 % in the United States. For Japan, the figures are 1.9×10^9 to 9.0×10^9 dollars, also at an annual growth rate of 11 %. For the world market, the values are 4.1×10^9 to 17.0×10^9 dollars at an annual growth of 10 %. Toshiba [168] forecasts a rise in the Japanese market from 1.26×10^9 to 11.4×10^9 dollars from 1982 to 2000, with an overall growth of 32 %. The U.S. Department of Commerce [169] predicts estimated sales for advanced ceramics rising from 0.60×10^9 to

5.9×10^9 dollars over the period 1980 – 2000 at an annual growth rate of 12 – 15 %.

The Western world in general and the United States in particular depend for economic strength on abundant, readily available natural resources. The countries of Western Europe, the United Kingdom, and North America are either self-sufficient in traditional ceramic materials or have ready access to such materials. This does not appear to be the case for many materials required for advanced ceramics. Of 27 basic industrial minerals or metals listed by the U.S. Department of the Interior, 18 are imported at levels above 50 %, including cobalt, manganese, and chromium [170].

13. References

General References

1. Annu. Book ASTM Stand. **1972,** part 13, Refractories, Glass, and Other Ceramic Materials. Terms Relating to Ceramic Whitewares and Related Products, C242 – C272.
2. Annu. Book ASTM Stand. **1985,** vol. 15.02, Glass; Ceramic Whitewares; Porcelain Enamels.
3. R. J. Charleston (ed.): *World Ceramics,* Paul Hamlyn, London1968.
4. R. W. Grimshaw: *The Chemistry and Physics of Clays and Allied Ceramic Materials,* 4th ed., J. Wiley & Sons, New York 1980.
5. L. L. Hench, R. W. Gould: *Characterization of Ceramics,* Dekker, New York 1971.
6. G. Herdan et al.: *Small Particle Statistics,* Butterworth, London 1960.
7. W. D. Kingery (ed.): *Ceramic Fabrication Processes,* J. Wiley & Sons, New York 1958.
8. W. D. Kingery (ed.): *Ceramics and Civilization,* vol. 1, *Ancient Technology to Modern Science,* American Ceramic Society, Columbus, Ohio, 1984.
9. C. Klein, C. S. Hurlbut, Jr. (after J. D. Dana): *Manual of Mineralogy,* 20th ed., J. Wiley & Sons, New York 1985.
10. S. J. Lefond (ed.): *Industrial Minerals and Rocks,* 5th ed., 2 vols., Soc. Min. Eng. of AIMME, New York 1983.
11. F. H. Norton: *Fine Ceramics: Technology and Applications,* McGraw-Hill, New York 1970.
12. F. H. Norton: *Elements of Ceramics,* 2nd ed., Addison-Wesley, Reading, Mass., 1974.
13. L. S. O'Bannon (ed.): *Dictionary of Ceramic Science and Engineering,* Plenum Press, New York 1984.
14. G. Y. Onoda, L. L. Hench (eds.): *Ceramic Processing before Firing,* J. Wiley & Sons, New York 1978.
15. H. Palmour et al. (eds.): *Processing of Crystalline Ceramics,* Plenum Press, New York 1978.
16. T. C. Patton: *Paint Flow and Pigment Dispersion,* 2nd ed., J. Wiley & Sons, New York 1979.
17. G. W. Phelps et al: *Rheology and Rheometry of Clay-Water Systems,* Cyprus Industrial Minerals, Sandersville, Ga., 1982.
18. P. Rado: *An Introduction to the Technology of Pottery,* Pergamon Press, Oxford 1969.
19. D. W. Richerson: *Modern Ceramic Engineering,* Dekker, New York 1982.
20. H. Salmang: *Ceramics: Physical and Chemical Fundamentals,* 4th ed., Butterworth, London 1961.
21. D. J. Shaw: *Introduction to Colloid and Surface Chemistry,* 3rd ed., Butterworth, London 1980.
22. H. van Olphen: *An Introduction to Clay Colloid Chemistry,* 2nd ed., J. Wiley & Sons, New York 1977.
23. W. E. Worrall: *Clays and Ceramic Raw Materials,* 2nd ed., Pergamon, Oxford 1982.

Specific References

24. W. W. Perkins (ed.): *Ceramic Glossary – 1984,* American Ceramic Society, Columbus, Ohio, 1984.
25. J. R. H. Black et al., *Am. Ceram. Soc. Bull.* **64** (1985) 39 – 41, 50.
26. W. G. Solheim, *Sci. Am.* **226** (1972) 34 – 41.
27. T. A. Wertime, *Am. Sci.* **61** (1973) 670 – 682.
28. J. G. Ayers, in [3]chap. 2
29. J. B. Hennessy, in [3] chap. 1
30. D. J. Hamlin: *The First Cities,* Time-Life Books, New York 1973, chap. 6.
31. Liu Zhen, Hu Youzhi, Zhen Naizhang, *J. Jingdezhen Ceram. Inst.* **5** (1984) no. 2, 17 – 36.
32. M. Farnsworth, *Am. J. Archaeol.* **68** (1964) 221 – 231.
33. G. Fehervari, in [3]chap. 3.
34. A. S. H. Megaw in [3]pp. 100 – 109.
35. F. A. Drier in [3]pp. 127 – 134.
36. S. Ducret in [3] pp. 216 – 224.
37. A. Ray in [3] pp. 246 – 254.

38. H. Ries, H. B. Kummel, G. N. Knapp: *The Clays and Clay Industry of New Jersey,* N.J. State Geological Survey, Trenton, N.J., 1901, chap. 11.
39. G. B. Kenney, H. K. Bowen, *Am. Ceram. Soc. Bull.* **62** (1983) 591.
40. E. Ruh: "Metallurgical Refractories," in M. G. Berer (ed.): *Encyclopedia of Materials Science and Engineering,* Pergamon Press, Oxford 1985.
41. R. A. Haber, V. A. Greenhut, E. J. Smoke, U.S. patent application, 30 April 1984.
42. J. D. Birchall, *Trans. Br. Ceram. Soc.* **82** (1983) 143 – 145.
43. R. A. Katz, *Science (Washington, D.C.)* **208** (1980) 841 – 847.
44. M. E. Washburn, H. R. Baumgartner, *Second Annual Army Materials Conference on Ceramics for High Performance Applications,* Hyannis, Mass., 1973.
45. G. Q. Weaver, J. W. Luckek, *Am. Ceram. Soc. Bull.* **58** (1978) 1131 – 1135.
46. J. Friberg, B. Aronsson in S. Somiya (ed.): *Ceramic Science at the Present and in the Future,* Uchido Rokakuyo Publ. Co., Tokyo 1981, pp. 109 – 130.
47. B. North, *Materials and Society* **8** (1984) 271 – 281.
48. H. R. Baumgartner: "Evaluation of Roller Bearings Containing Hot-Pressed Silicon Nitride Rolling Elements," in *Second Annual Army Materials Conference on Ceramics for High Performance Applications,* Hyannis, Mass., 1973.
49. J. F. Baumard, B. Cales, A. M. Anthony in S. Somiya (ed.): *Ceramic Science and Technology at the Present and in the Future,* Uchido Rokakuyo Publ. Co., Toyko 1981, pp. 161 – 191.
50. P. A. Janeway, *Ceram. Ind. (Chicago)* **122** (1984) 40 – 45.
51. P. F. Becher, *J. Am. Ceram. Soc.* **64** (1981) 37 – 39.
52. W. R. Cannon, K. Wilfinger, personal communication (1984).
53. I. M. Lachmann, R. N. McNally, *Ceram. Eng. Sci. Proc.* **2** (1981) 337 – 351.
54. D. W. Roy, K. E. Green, *Ceram. Eng. Sci. Proc.* **4** (1983) 510 – 519.
55. K. M. Prewo, J. J. Brennan, *J. Mater. Sci.* **15** (1980) 463 – 468.
56. A. K. Dhingra, *Philos. Trans. R. Soc. London* **A 294** (1980) 559 – 564.
57. M. K. Surappa, P. K. Rohatgi, *J. Mater. Sci.* **16** (1981) 983 – 993.
58. W. J. Lackey et al.: "Ceramic Coatings for Heat in Engine Materials – Status and Future Needs," in *Proc. Int. Symp. Ceramic Components for Heat Engines,* Hakone, Japan, 1981.
59. G. Fisher, *Am. Ceram. Soc. Bull.* **63** (1984) 569 – 571.
60. J. T. Kummer, N. Weber, *SAE J.* **76** (1968) 1003 – 1007.
61. W. D. Kingery et al., *J. Am. Ceram. Soc.* **42** (1959) 393 – 398.
62. J. Hecht, *High Technol.* **3** (1983) July/Aug., 49 – 56.
63. D. W. Richerson, M. E. Washburn, US 3 836 374, 1974.
64. G. E. Gazza, *Am. Ceram. Soc. Bull.* **54** (1975) 778 – 781.
65. *Characterization of Materials,* Materials Adv. Board, Div. Eng., Nat. Research Council, Publ. MAB-229-M, Nat. Acad. Sci. – Nat. Acad. Eng., Washington, D.C., 1967.
66. G. W. Phelps, *Am. Ceram. Soc. Bull.* **55** (1976) 528 – 529, 532.
67. G. W. Phelps, D. L. Harris, *Am. Ceram. Soc. Bull.* **47** (1968) 1146 – 1150.
68. W. A. Weyl, N. A. Terhune, *Ceram. Age* **62** (1953) no. 23, 40 – 41.
69. D. Arayaphong, M. G. McLaren, G. W. Phelps, *Am. Ceram. Soc. Bull.* **62** (1984) 1181 – 1185.
70. G. W. Phelps, *The Role of Naturally Occurring Organic Matter in Clay Casting Slips,* Univ. Microfilms, Ann Arbor, Mich., 1963, pp. 144 – 180.
71. G. W. Phelps, M. G. McLaren in [14] pp. 211 – 225.
72. G. A. Loomis, *J. Am. Ceram. Soc.* **23** (1940) 159 – 162.
73. L. Pauling: *The Nature of the Chemical Bond,* 3rd ed., Cornell University Press, Ithaca, N.Y., 1960, p. 544.
74. W. D. Keller, *Clays Clay Miner.* **33** (1985) 161 – 172.
75. R. E. Grim, *Clay Mineralogy,* 2nd ed., McGraw-Hill, New York 1968, pp. 77 – 92.
76. A. L. Johnson, W. G. Lawrence, *J. Am. Ceram. Soc.* **25** (1942) 344 – 346.
77. W. G. Lawrence, R. R. West, *Ceramic Science for the Potter,* 2nd ed., Chilton Book Co., Radnor, Pa., 1982, pp. 45 – 55.
78. J. W. Shaffer in [10] pp. 506 – 508 in vol. 1
79. D. A. Holdridge, *Trans. Br. Ceram. Soc.* **62** (1963) 857 – 875.
80. E. W. Koenig, *J. Am. Ceram. Soc.* **25** (1942) 420 – 422.

81. C. P. Rogers, J. P. Neal, K. H. Teague (rev.) in [10]pp.709 – 722 in vol. 1
82. D. G. Minnes et al. in [10]pp. 931 – 960 in vol. 2
83. J. van Wunnik, J. S. Dennis, G. W. Phelps, *J. Can. Ceram. Soc.* **30** (1961) 1 – 7.
84. W. Ryan, W. E. Worrall, *Trans. Br. Ceram. Soc.* **60** (1961) 540 – 555.
85. G. W. Phelps, *Am. Ceram. Soc. Bull.* **38** (1959) 411 – 414.
86. R. E. Brociner, R. T. Bailey: "Mechanical Treatment of Ceramic Bodies," in *Trans. Int. Ceram. Congr.* **9** (1966).
87. V. S. Schory, *J. Am. Ceram. Soc.* **3** (1920) 286 – 295.
88. A. F. Taggart (ed.): *Handbook of Mineral Dressing,* J. Wiley & Sons, New York 1945.
89. N. H. Parker, *Chem. Eng. (N.Y.)* **71** (1964) 8 June, 165 – 220.
90. L. H. Stone, *Chem. Eng. (N.Y.)* **86** (1979) 15 Jan., 125 – 130.
91. R. Remirez, *Chem. Eng. (N.Y.)* **85** (1978) 4 Dec., 72, 74.
92. *Chem. Eng. (N.Y.)* **92** (1985) 5 Aug., 47.
93. Bowen Engineering, *Spray Dryers for Ceramics,* North Branch, N.J., Bulletin 42 – 2.
94. J. T. Jones, M. F. Berard: *Ceramics: Industrial Processing and Testing,* Iowa State University Press, Ames, Iowa, 1972, pp. 30 – 35.
95. M. K. Bo et al., *Trans. Inst. Chem. Eng.* **43** (1965) T228 – T232.
96. C. C. Furnas, *Ind. Eng. Chem.* **23** (1931) 1052 – 1058.
97. G. W. Phelps et al., *Am. Ceram. Soc. Bull.* **50** (1971) 720 – 722.
98. P. M. Rockwell, A. J. Gitter, *Am. Ceram. Soc. Bull.* **44** (1965) 497 – 499.
99. J. G. M. deLau, *Am. Ceram. Soc. Bull.* **49** (1970) 572 – 574.
100. D. W. Johnson, F. J. Schettler, *J. Am. Ceram. Soc.* **53** (1970) 440 – 444.
101. R. E. Mistler et al. in [14]pp. 411 – 438.
102. R. Russell et al., *J. Am. Ceram. Soc.* **32** (1949) 105 – 113.
103. L. L. Hench et al., *Ceram. Eng. Sci. Proc.* **3** (1982) 477 – 483.
104. L. C. Klein, *Ceram. Eng. Sci. Proc.* **5** (1984) 379 – 384.
105. R. Roy, *J. Am. Ceram. Soc.* **52** (1969) 344.
106. R. E. Gould, R. W. Cline, *Am. Ceram. Soc. Bull.* **29** (1950) 291 – 292.
107. R. C. P. Cubbon, *Br. Ceram. Trans. J.* **83** (1984) 121 – 124.
108. R. C. Rossi, R. M. Fulrath, *J. Am. Ceram. Soc.* **48** (1965) 558 – 564.
109. A. R. Blackburn, *Am. Ceram. Soc. Bull.* **29** (1950) 230 – 234.
110. O. J. Whittemore, Jr. in [14]pp. 343 – 355.
111. B. W. Nies, C. M. Lambe, *Am. Ceram. Soc. Bull.* **35** (1956) 319 – 323.
112. E. G. Walker, *Trans. Br. Ceram. Soc.* **64** (1965) 233 – 248.
113. D. J. Shanefield, R. E. Mistler, *Am. Ceram. Soc. Bull.* **53** (1974) 416 – 420.
114. J. A. Mangels, W. Trela in J. A. Mangels, G. L. Messing (eds.): *Forming of Ceramics,* Amer. Ceram. Soc., Columbus, Ohio, 1984, vol. 9, pp. 220 – 233.
115. A. Swineford (ed.): "Symposium on the Engineering Aspects of the Physico-Chemical Properties of Clays" in *Proc. 9th Nat. Conf. on Clays and Clay Min.,* Pergamon Press, New York 1962.
116. R. K. Schofield, *Trans. Int. Congr. Soil Sci. 3rd, 1935,* **vol. 2,** pp. 37 – 48.
117. R. Q. Packard, *J. Am. Ceram. Soc.* **50** (1967) 223 – 229.
118. A. D. Coleman, J. D. Marsh, *J. Soil Sci.* **12** (1961) 342 – 361.
119. F. Samudio, unpublished research, Rutgers University, 1981.
120. G. W. Phelps in [15]pp. 57 – 65.
121. G. W. Phelps: *Proceedings of the Materials & Equipment/Whitewares Divisions,* American Ceramic Society, Bedford, Pa., Sept. 1970, pp. 9 – 14.
122. F. H. Norton, *J. Am. Ceram. Soc.* **16** (1933) 88 – 92.
123. W. R. Morgan, R. K. Hursh, *J. Am. Ceram. Soc.* **22** (1939) 271 – 278.
124. D. Woo et al., *J. Am. Ceram. Soc.* **38** (1955) 383 – 388.
125. R. P. Allaire, *Ceram. Ind. (Chicago)* **86** (1966) no. 3, 38.
126. B. Vassiliou, J. White, *Trans. Br. Ceram. Soc.* **52** (1953) 329 – 385.
127. G. W. Bird, A. J. Dale, *Trans. Br. Ceram. Soc.* **51** (1952) 559 – 573.
128. L. H. Hepner, *Am. Ceram. Soc. Bull.* **24** (1945) 415 – 417.
129. *Webster's Ninth New Collegiate Dictionary,* Merriam-Webster, Springfield, Mass., 1983.
130. J. E. Burke in W. D. Kingery (ed.): *Ancient Technology to Modern Science,* American Ceramic Society, Columbus, Ohio, 1984, pp. 315 – 333.
131. J. G. Weinstein, C. Chanyavanich, unpublished report, Rutgers University, Oct. 1980.
132. J. R. Schorr, R. Russell, *Am. Ceram. Soc. Bull.* **49** (1970) 1042 – 1051.

133. C. R. Moebus et al., *Ceram. Eng. Sci. Proc.* **4** (1983) 935 – 945.
134. J. G. Weinstein et al., *Ceram. Eng. Sci. Proc.* **3** (1982) 879 – 887.
135. T. Vasilos, R. M. Spriggs, *Proc. Br. Ceram. Soc.* **3** (1967) 195 – 221.
136. R. M. Lonero, *Am. Ceram. Soc. Bull.* **62** (1983) 1000, 1009.
137. N. M. Hintz, *Ceram. Eng. Sci. Proc.* **4** (1983) 1014 – 1022.
138. C. G. Harmon, Jr., *Ceram. Eng. Sci. Proc.* **2** (1981) 908 – 916.
139. S. W. Kennedy, K. W. Doak, *Ceram. Eng. Sci. Proc.* **5** (1984) 1012 – 1024.
140. E. C. Bloor, *Trans. Br. Ceram. Soc.* **55** (1956) 631 – 660.
141. H. H. Sortwell, *J. Am. Ceram. Soc.* **4** (1921) 718 – 730.
142. A. S. Watts, *Trans. Am. Ceram. Soc.* **19** (1917) 301 – 302.
143. F. Singer, *Trans. Br. Ceram. Soc.* **53** (1954) 398 – 421.
144. C. G. Harmon, H. R. Swift, *J. Am. Ceram. Soc.* **28** (1945) 48 – 52.
145. R. P. Isaacs, *Am. Ceram. Soc. Bull.* **45** (1966) 714 – 715.
146. C. B. Lutrell, *J. Am. Ceram. Soc.* **32** (1949) 327 – 332.
147. C. W. F. Jacobs, *J. Am. Ceram. Soc.* **37** (1954) 216 – 220.
148. W. J. Koch et al., *J. Am. Ceram. Soc.* **33** (1950) 1 – 8.
149. F. H. Norton, *J. Am. Ceram. Soc.* **20** (1937) 217 – 224.
150. L. E. Barringer, *Trans. Am. Ceram. Soc.* **4** (1902) 211 – 229.
151. G. W. Phelps, *Proc. Porcelain Enamel Inst. Tech. Forum* **38** (1976) 246 – 250.
152. H. N. Mills in [10] pp. 339 – 347 in vol. 1.
153. W. H. Dumbaugh, J. W. Malmendier in A. M. Alper (ed.): *High Temperature Oxides,* part 4, Refractory Glasses, Glass-Ceramics, and Ceramics, Academic Press, New York 1971, pp. 1 – 14.
154. G. H. Beall in A. M. Alper (ed.): *High Temperature Oxides,* part 4, Refractory Glasses, Glass-Ceramics, and Ceramics, Academic Press, New York 1971, pp. 15 – 36.
155. J. A. Crookston, W. D. Fitzpatrick in [10]pp. 373 – 385 in vol. 1.
156. J. A. Ames, W. E. Cutliffe in [10]pp. 133 – 159 in vol. 1.
157. *Annu. Book ASTM Stand.* 1980, part 14, *Concrete and Mineral Aggregates* .
158. W. J. Lackey, D. P. Stinton, G. A. Cerny, L. L. Fehrenbacher, A. C. Schaffhauser: "Ceramic Coatings for Heat Engine Materials – Status and Future Needs," *Proc. Int. Symp. Ceram. Components for Heat Engines,* 1983; ORNL/TM-8959.
159. H. B. Dubois, *J. Am. Ceram. Soc.* **15** (1932) 144 – 148.
160. W. Weibull, *J. Appl. Mech.* **18** (1951) 293 – 297.
161. R. J. Beals, *Am. Ceram. Soc. Bull.* **64** (1985) 47 – 50.
162. J. H. Belger, *Am. Ceram. Soc. Bull.* **61** (1982) 1285 – 1286.
163. J. J. Walsh, *Am. Ceram. Soc. Bull.* **61** (1982) 1284 – 1285.
164. "U.S. Industry Trends," *Am. Ceram. Soc. Bull.* **62** (1983) 547.
165. "U.S. Industry Trends," *Am. Ceram. Soc. Bull.* **63** (1984) 547.
166. "U.S. Industry Trends," *Am. Ceram. Soc. Bull.* **64** (1985) 11.
167. H. K. Bowen, personal communication to J. B. Wachtman, Jr., 1985.
168. J. B. Wachtman, Jr., *Ceram. Ind. (Chicago)* **121** (1983) 24 – 33.
169. U.S. Dept. Commerce: *A Competitive Assessment of the U.S. Advanced Ceramics Industry,* NTIS Access. No. PB84-162288 (1984).
170. P. C. Maxwell, *Am. Ceram. Soc. Bull.* **59** (1980) 1158 – 1159.

Ceramics, Advanced Structural Products

W. Roger Cannon, Department of Ceramics, Rutgers – The State University, Piscataway, New Jersey 08854, United States

Ernst Gugel, CFI – Ceramics for Industry GmbH & Co. KG, Roedental, Federal Republic of Germany

Gerhard Leimer, CFI – Ceramics for Industry GmbH & Co. KG, Roedental, Federal Republic of Germany

Gerhard Woetting, CFI – Ceramics for Industry GmbH & Co. KG, Roedental, Federal Republic of Germany

1. Introduction 1
2. Mechanical Properties 2
2.1. Failure of Brittle Materials – Flaws . 2
2.2. The Small Strain to Failure 3
2.3. Effects of Microstructure 4
2.4. Effects of Temperature 4
2.5. Wear Resistance and Hardness 6
3. Individual Ceramic Materials 6
3.1. Alumina 6
3.2. Zirconia 7
3.3. Silicon Carbide 8
3.4. Silicon Nitride 9
3.5. Ceramic – Ceramic Composites 9
3.6. Other Materials 10
4. Manufacturing Methods 11
5. Uses 11
6. References 15

1. Introduction

The viability and lifetime of chemical engineering facilities depend on the quality of the materials from which the critical components are made. Wear caused by abrasion and/or corrosion results in economic losses of many million marks. Hence, materials which lead to a longer lifetime of industrial structures or even to an improved process, are of key importance. In this respect, the increasing interest in ceramic materials for structural applications is not surprising, since modern ceramic materials have reached a high standard of quality and reliability by intensive developments of materials and production technology.

Ceramic materials are well-known to the chemical industry; they have been used on a large scale for many years, for example, as linings of large containers, packings, catalyst supports, filters, etc. Furthermore, in the early years of the chemical industry, ceramics in the form of chemical stoneware [1] made a major contribution to their success. Containers with a volume of 2000 L, columns several meters high, pipelines and heat exchangers, and, in particular, pumps made of the classical ceramic stoneware were notable early forerunners of modern engineered ceramics.

Further developments in chemical stoneware [1] led to corundum stoneware with its higher strength, and to cordierite and Li – Al stoneware, which show improved thermal shock resistance, but unsatisfactory chemical resistance. All these desirable properties are met by beryllium stoneware, which did not reach application because of difficulty in processing large components, but more especially because of its toxicity in production.

Stoneware is made from natural raw materials. Due to impurities and inhomogeneities, its microstructure is rather coarse, contains considerable glass and microcracks, and is inhomogeneous, resulting in only limited mechanical properties. The decisive step from low-strength (classical) ceramics, such as bricks, dishes, and wall and floor tiles, to high mechanical strength ceramics, resulted from progress in materials science and production technology knowledge in the course of the last few decades.

A new generation of ceramic materials is being developed with specifically controlled properties, based on fine-grained or whiskerlike pure powders of defined composition, which have been developed by the chemical industry and are now available in large quantities. With these powders, ceramics with a controlled microstructure, which are characterized by small crack and

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_043

defect sizes can be produced, provided that their production is carefully controlled. Defects become increasingly critical if the powder of the starting material is very fine grained. Though an extremely fine microstructure may have some advantages, too much should not be expected of nanotechnology, at least in the near future, as its realization for large components with reliable properties at acceptable costs is problematic.

The same applies to very tough ceramics. Although considerable progress has been made with the incorporation into ceramics of particles, whiskers, or fibers which inhibit crack growth and therefore lead to an increase in fracture toughness and a certain crack tolerance, a general elimination of the brittleness to give "ductile" ceramics cannot be expected.

Regardless of these possibilities for the future, the quality of currently available materials is sufficiently high that even high tensile stress applications are possible, provided the components are properly designed. However, even the modern, highly dense ceramics are brittle, and the component designer needs to learn how to design components to eliminate high tensile stress concentrations. With increasing familiarity, designers will not only look to the metallic and polymeric materials with which they are already familiar, but also to ceramic materials for structural design work.

Brittleness also causes a higher sensitivity to material defects. This is reflected in the variability in strength, which can be described by Weibull statistics (see Section 2.2). Apart from the mean value, the statistical value for fracture probability, expressed as the Weibull modulus, is required for reliable design with ceramics. It even allows data gained from small samples to be transferred to complex components. A desirable low variation range is indicated by a high Weibull modulus; a value of greater than 15 is sufficient in most cases.

Development of modern ceramic materials is certainly not yet finished, however. The stage has been set for applications which could not be imagined until a short time ago. With regard to the future, it is not only important to increase the strength level. The critical point is that the properties achieved in the laboratory must be realized in mass production. This requires homogeneous properties within components and from part to part, which means minimizing the variation range and therefore elimination of problems resulting from the brittleness by even tighter control. Such control is absolutely necessary for high-reliability applications.

2. Mechanical Properties

2.1. Failure of Brittle Materials – Flaws

The key to understanding the mechanical properties of ceramics is the Griffith theory for failure of brittle materials [3]. Unlike most metals and polymers, ceramics are extremely brittle and, therefore, extremely sensitive to the presence of any sort of flaw on the surface or in the bulk of the ceramic. Flaws of some size are always present in a material. These may arise during manufacture or in subsequent handling of the material. Since polycrystalline ceramics are usually manufactured by sintering a powder compact at high temperature, the flaws may arise from nonuniformities in the packing of the powder or accidental inclusions in the compact.

The Griffith criterion states that if the flaw is larger than a critical size, then the rate of release of elastic strain energy is greater than the rate of gain of surface energy and the flaw will grow (crack extension). The growth is catastrophic because as the crack grows longer, the driving force becomes greater. The Griffith criterion is a necessary but not sufficient condition for crack growth. In addition, the stress at the tip of the crack must be large enough to break the ionic or covalent bonds. Orowan estimated this stress and found that for a sharp crack the Griffith stress was sufficient to break the bonds [4]. The *Griffith equation* is

$$\sigma_{\mathrm{f}} = \sqrt{\frac{2E\gamma}{\pi C}}$$

where σ_{f} is the fracture stress, the applied stress at which the preexisting flaw becomes critical, E is the elastic modulus (Young's modulus), γ is the effective surface energy, and C is the length of a surface flaw or $2\,C$ is the length of an interior flaw. One assumption is that only the increasing surface area of the crack acts to resist crack growth. However, other features besides surface energy resist the propagation of the crack: plastic deformation near the tip of the crack, crack

bridging, and crack deflection. Commonly these contributions are included in γ, which is then renamed the fracture energy, γ_f.

The Griffith equation is sometimes written in terms of the critical stress intensity factor defined as

$$K_C \equiv Y \sigma_f \sqrt{C}$$

where Y is a dimensionless constant that depends on the geometry of loading and the crack configuration. The fracture stress is then written as follows:

$$\sigma_f = \frac{1}{Y} \sqrt{2E\gamma/C} = \frac{K_C}{Y\sqrt{C}} \quad (1)$$

This second form of the Griffith equation is useful because K_C can be measured by various fracture-mechanics techniques. (K_{IC} is often used instead of K_C, the subscript I indicating the mode of cracking.) Table 1 shows how the fracture stress and the critical flaw size are related for a promising structural ceramic such as hot-pressed silicon carbide (SiC). The values chosen for this example are a K_C of 4 MPa m$^{0.5}$ and a Young's modulus E of 400 GPa. This table indicates that the flaw size for the high-strength ceramics is very small and that careful manufacturing techniques are necessary to avoid accidentally producing such small flaws in polycrystalline materials.

Table 1. Fracture strength vs. flaw size

σ_f, MPa	C, µm
1000	5
500	20
250	80
100	500
50	2025

Even if great care is taken to manufacture a material so that the largest flaw is on the order of 1 µm, there is no guarantee that new flaws will not be introduced during operation by scratching or particle impact. As a result, most research in developing structural ceramics is now concerned with increasing K_C. Equation (1) shows that acceptable flaw size is proportional to the square of the K_C value, e.g., doubling K_C allows a flaw size 4 times larger.

The K_C value is really an indicator of how brittle or tough the material is. Various methods have been used to increase K_C to as high as 10 – 15 MPa m$^{0.5}$. These high fracture toughnesses have been achieved with ceramic – ceramic composites. Continuous fiber composites, usually fabricated from woven ceramic cloth, are almost completely insensitive to flaws; however, they are strongly dependent on the mean strength of the fibers.

2.2. The Small Strain to Failure

A second aspect of brittleness in ceramics is the small strain to failure. Ceramics do not undergo plastic deformation at room temperature as do metals. Plastic deformation in both metals and ceramics results from microscopic defects called dislocations. These dislocations move easily in the presence of a stress field in metals, but they do not move easily in ceramics. The stress at which plastic deformation could take place in ceramics is much higher than the fracture stress. The opposite is true for metals. As a consequence, the strain to failure in ceramics is usually on the order of only 0.1 %, and ceramics can only be used when the stress is safely below the fracture stress.

The strong sensitivity to flaws and small strains to failure, however, may be compensated by other outstanding properties of ceramics, e.g., wear resistance or corrosion resistance.

Safe design of ceramics in tensile or bending load applications requires knowledge of the necessary safety factor. To estimate the safety factor properly, information on the statistics of failure is necessary. The most widely used statistics are Weibull statistics. The *Weibull equation* is as follows:

$$\ln\left(1/\left(1-P_f\right)\right) = \ln V + m \ln\left(\sigma_f - \sigma_\mu\right) - m \ln \sigma_0$$

where P_f is the probability of failure of a part subject to the stress, V is the volume of the sample under the tensile stress, σ_μ is the stress of zero probability of failure, usually taken as zero, σ_0 is a normalizing parameter of no physical significance, and m is a constant designated as the Weibull modulus. This equation indicates that the probability of failure at a certain stress is related to the volume of the material under stress (or may be related to the area under stress if failure is from surface flaws). Such a relationship between volume and stress has been verified experimentally. The equation may be linearized by

plotting $\ln\ln(1/(1-P_f))$ vs. $\ln\sigma_f$, allowing the probability of failure at any given stress to be estimated by extrapolating a straight line. This is termed a *Weibull plot* (Fig. 1).

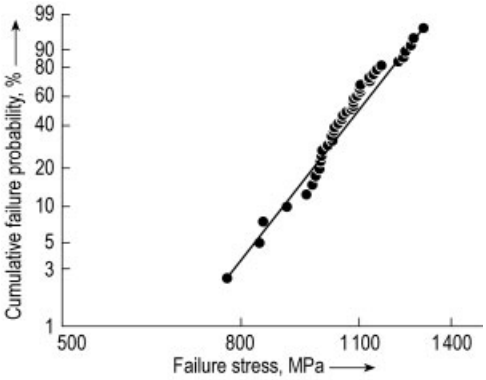

Figure 1. Failure probability vs. fracture stress, a Weibull plot [5]

An additional important concern in design is slow crack growth. The crack grows under stress in the presence of atmospheric moisture. The velocity v of crack growth is given by the equation

$$v = \alpha K^N$$

where K is the stress intensity factor, α is a constant, and N is the velocity exponent. Such an equation can be incorporated into Weibull statistics for a more accurate treatment. This subject is reviewed in [6].

2.3. Effects of Microstructure

Microstructure of polycrystalline materials has an important effect on the strength of the ceramic. Generally, strength is improved by achieving a fine grain size. If the flaw size is much larger than the grain size, however, grain size has little effect. For materials having very anisotropic thermal coefficients of expansion and large grain size, *microcracking* develops when the ceramic is cooled to room temperature after sintering. This can be detrimental to strength. In addition, microcracking may occur near second phases having different coefficients of expansion than the continuous phase. Finally, microcracking may occur if the second phase undergoes a phase transformation involving a large change in volume, for instance, the phase transformation of quartz at 573 °C.

Large grains and second phases are not always bad. Large grains if properly designed may bridge cracks, so that higher stresses are required to propagate the crack, and therefore, toughen the ceramic and in many cases also increase the fracture strength. A new class of self-reinforced ceramics (primarily silicon nitride) has been developed. In situ reinforced silicon nitride mimics whisker-reinforced ceramics, except that the whiskerlike grains are grown from β-silicon nitride seed grains during the sintering process. Careful control of the additives, temperature, and the starting powder is necessary to achieve high toughness and strength. A strength of 1000 GPa and a toughness of $10\,\mathrm{MPa}\cdot\mathrm{m}^{0.5}$ can be achieved by careful control of the number and size of the whiskerlike grains in a fine-grained silicon nitride matrix [7]. This innovation was necessary for the commercialization of silicon nitride roller bearings.

2.4. Effects of Temperature

The effect of temperature on the strength of ceramics is illustrated in Figure 2. There is a decrease in strength at high temperatures, where some plastic deformation begins to take place. The temperature at which this decrease in strength begins depends on the material. In materials with a cubic crystal symmetry, such as MgO or UO_2, the temperature at which the strength begins to drop is generally low, but in Al_2O_3 the strength drops off little up to 1000 °C. At high temperatures the strength depends on the rate at which the stress is applied.

Above one-half the melting temperature, creep (slow time-dependent deformation) becomes appreciable. Since many of the important applications of ceramics are high-temperature applications, creep is important. Creep of ceramic materials has been reviewed by Cannon and Langdon [8]. Most fine-grained polycrystalline ceramics under moderately low applied stresses deform according to one of the following two equations: the Nabarro – Herring equation [9],

$$\dot{\varepsilon} = \frac{10\Omega\sigma D_v}{kTd^2} \tag{2}$$

or the Coble equation [10],

$$\dot{\varepsilon} = \frac{50\Omega\sigma w D_{\mathrm{b}}}{kTd^3}$$

where $\dot{\varepsilon}$ is the creep rate, Ω is the volume of a vacancy, σ is the applied stress, D_{v} is volume self-diffusion coefficient, d is grain diameter, k is Boltzmann's constant, T is temperature (K), w is the effective width of the grain boundary, and D_{b} is grain boundary diffusion coefficient. The two equations indicate that the most highly creep resistant materials are large-grained ceramics. The equations also indicate that the creep resistance is related to the diffusion coefficient. The diffusion coefficient is generally very low in ceramics having a high melting temperature, and therefore ceramics are generally creep resistant materials at high temperatures ($> 1000\,^{\circ}$C). If the ceramic contains a glassy grain boundary phase, then creep rates are somewhat higher than those given in these equations. If high creep resistance is desired, the glassy phase should be eliminated or high-temperature viscous glass should be used.

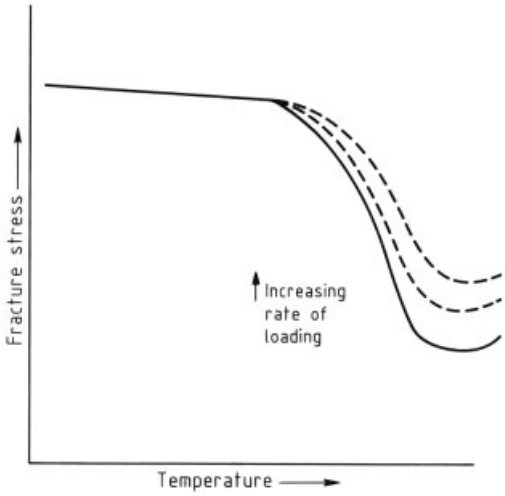

Figure 2. Typical fracture strength vs. temperature curves

Ceramics are sometimes limited as a high-temperature material by the inability to withstand *thermal shock,* i.e., rapid drop in temperature. Their lack of thermal shock resistance is primarily due to their inability to accommodate the strain caused by a rapidly cooling surface adjacent to a hot interior. The low thermal conductivity of ceramics contributes to this lack of thermal shock resistance. The critical temperature drop ΔT_{c} through which a material may be shocked without exceeding the fracture stress at the surface is given by the equation

$$\Delta T_{\mathrm{c}} = \frac{\sigma_{\mathrm{f}}\,(1-\nu)}{\psi E\alpha}$$

where α is the thermal coefficient of expansion, ν is Poisson's ratio, and ψ is a factor depending on thermal conductivity, size of piece, and heat transfer coefficient. The equation contains several materials parameters generally available, and therefore may be used to estimate the relative thermal shock resistance of various materials. Table 2 compares the thermal shock resistance of various advanced structural ceramic materials for both rapid thermal shock $(R = \sigma_{\mathrm{f}}\,(1-\nu)\,/E\alpha)$ and slow thermal shock $(R' = \sigma_{\mathrm{f}} k\,(1-\nu)\,/E\alpha)$, where k is the thermal conductivity).

2.5. Wear Resistance and Hardness

Metallic parts that tend to wear out, e.g., in engines, may be replaced by ceramic parts for longer life. Ceramics exhibit good wear resistance and erosion resistance because of their hardness, i.e., they do not plastically deform easily. Those with the greatest resistance to plastic deformation are the covalently bonded materials and the ones of greatest hardness. However, resistance to *crack propagation* also has some influence on the hardness and wear resistance since cracks form under concentrated point loading. These point loadings may be due to dust particles, grit, or other types of concentrated loads. Some of these cracks propagate parallel to the surface at first and then up toward the surface, resulting in small chips dislodged from the surface. After many of these chips have been dislodged, the surface is rougher and wears even faster. Thus, under severe load conditions ceramics may not have better wear resistance than much softer materials. Improving fracture toughness can improve abrasion resistance.

3. Individual Ceramic Materials

Table 3 lists a number of ceramics that are considered advanced structural ceramics. The most important advanced structural materials being

Table 2. Factors affecting the thermal shock behavior

Material	Bending strength, MPa	Modulus of elasticity E, GPa	Poissons ratio ν	Coefficient of thermal expansion α, 273 – 1273 K, 10^{-6} K^{-1}	Thermal conductivity k at 773 K, W m^{-1} K^{-1}	R, K	R', kW m^{-1}
Hot-pressed Si_3N_4	850	310	0.27	3.2	17	625	11
Reaction-bonded Si_3N_4	240	220	0.27	3.2	15	250	3.7
Reaction-bonded SiC	500	410	0.24	4.3	84	215	18
Hot-pressed Al_2O_3	500	400	0.27	9.0	8	100	0.8

developed at this time are the oxides zirconia and alumina and the covalent materials silicon carbide and silicon nitride. In addition, cordierite is important because of its use in heat exchangers. Table 4 lists the important mechanical properties of several of these materials. Flexure strengths are included though they depend strongly on the flawsize and therefore the care in fabrication. The range of values of K_{IC} is included, although these values depend strongly on the method of testing. (For the general production methods and economic aspects of advanced structural products → Ceramics, General Survey.)

Table 3. Some advanced structural ceramics and potential materials with their chemical formulas

Generic name	Chemical formula
Alumina [*1344-28-1*]	Al_2O_3
Zirconia [*1314-23-4*]	ZrO_2
Zircon [*1490-68-2*]	$ZrO_2 \cdot SiO_2$
Spinel [*1302-67-6*]	$MgO \cdot Al_2O_3$
Mullite [*55964-99-3*]	$3\,Al_2O_3 \cdot SiO_2$
Cordierite [*12182-53-5*]	$2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$
Silicon carbide [*409-21-2*]	SiC
Silicon nitride [*12033-89-5*]	Si_3N_4
SiAlON	$Si_{6-z}Al_zO_zN_{8-z}$
Boron carbide [*12069-32-8*]	B_4C
Aluminum nitride [*24304-00-5*]	AlN
Glass ceramics	a common composition is Al_2O_3, Li_2O, and SiO_2 with TiO_2 or ZrO_2 nucleating agents

3.1. Alumina

Of the advanced structural materials polycrystalline alumina is perhaps the easiest to manufacture, is relatively inexpensive, has high strength, and is the most widely used. It is widely used for crucibles, tubes, and rods for high temperature and for a large number of wear-resistant and corrosion-resistant specialized items. Perhaps the most important single products are spark plugs and the optically translucent polycrystalline alumina lamp hulls for high-temperature sodium vapor street lamps.

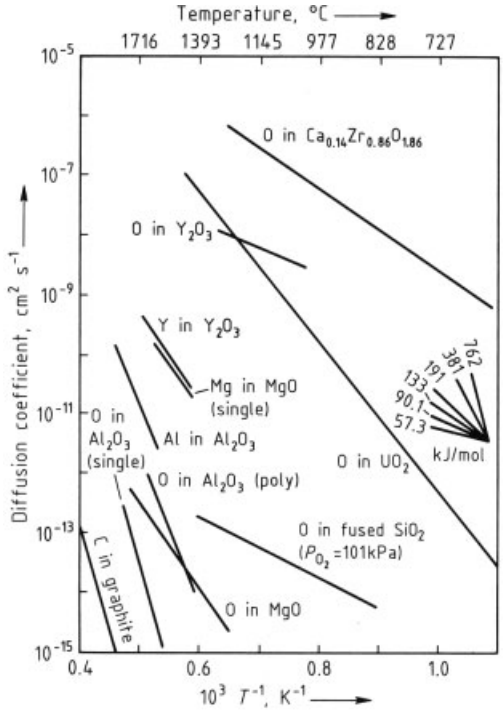

Figure 3. Ionic self-diffusion coefficients in ceramic oxides [11]

Several grades of sintered alumina are available, the grades being distinguished by the purity. The chief impurity is silica, which forms glass at the grain boundary. This glass acts as a sintering aid, lowering the sintering temperature by several hundred kelvin, but as a result these aluminas have lower strength and hardness. The proper glass compositions, however, can improve the low-temperature strength. High-purity alumina is necessary for high-temperature creep resistance, for otherwise the glass in the grain boundary begins to flow, producing grain bound-

Table 4. Engineering ceramic materials and their basic properties

Property	Si_3N_4	SiC	ZrO_2	Al_2O_3 (+ZrO_2)	Al_2TiO_5
Density, g/cm^3	3.2	3.2	6	3 – 4	3
Bending strength, MPa	200 – 1000	500	500 to > 1000	300 – 600	40
Bending strength at 1400 °C, MPa	200 – 600	200 – 400	100	100 – 400	50
Hardness, GN/m^2	14 – 17	25 – 30	12	18 – 23	
Fracture toughness K_{IC}, MPa m$^{0.5}$	3 – 10	3 – 5	6 – 15	3 (8)	1
Young's modulus	200 – 300	400	200	400 (300)	20
Coefficient of thermal expansion 20 – 1200 °C, (10^{-6} K^{-1})	3	4.5	10	8	2
Thermal conductivity, W m^{-1} K^{-1}	10 – 40	100 – 140	2	30	2
Thermal shock resistance	high	high	medium	low (medium)	very high
Abrasion resistance	very good	very good	good	good	no

ary sliding. Magnesium oxide is used as a sintering aid to sinter high-purity, high-density alumina.

Figure 3 is a summary of bulk diffusion coefficients of various ions in ceramic oxides. Equations (2) and (3) predict that creep rates should be proportional to the diffusion coefficients, and a comparison of diffusion coefficients is a good way of comparing the fundamental resistances to creep. As shown in Figure 3, both the Al ion and the O ion in alumina have two of the lowest diffusion coefficients, accounting for the low creep rate of high-purity alumina. However, the creep resistance of alumina is exceeded by that of silicon carbide, and thus for many very creep resistant applications, silicon carbide is being increasingly used.

New developments in improving the creep resistance of alumina are noteworthy. Ohji et al. [12] reported that alumina containing 17 % silicon carbide nanocrystalline particulates have a creep rate three orders of magnitude slower and creep lifetimes ten times longer than monolithic alumina polycrystals. French et al. [13] reported alumina with only 1000 ppm of yttria had a creep rate one to two orders of magnitude slower than monolithic alumina polycrystals. The former development, though somewhat difficult to manufacture, would allow alumina to approach competitiveness with silicon carbide for creep resistance without the oxidation problems of the latter.

3.2. Zirconia

Because of its high melting temperature (2764 °C) zirconia can be used for structural applications at higher temperatures than alumina. It is also used widely as a refractory because it does not react with glass. A few percent of MgO, CaO, or Y_2O_3 are added to zirconia to stabilize the high-temperature cubic phase; otherwise, zirconia undergoes a destructive phase transformation from tetragonal to monoclinic when cooled from the sintering temperature. However, *partially stabilized zirconia,* i.e., that containing less additive than necessary to establish 100 % cubic, was determined as early as 1947 to be more resistant to thermal shock than the fully stabilized form [14].

The thermal shock resistance of the partially stabilized zirconia may result from a very fine tetragonal precipitate present even at room temperature. The tetragonal precipitate is prevented from transforming to monoclinic by the constraint of the matrix, since the precipitate must necessarily expand 3 – 5 % and shear 8 % during the transformation to monoclinic. The toughening effect of the precipitate arises from the transformation that takes place as a crack approaches the precipitate [15]. The stress field of the crack interacts with the precipitate, allowing the transformation to occur. As the crack passes the precipitate, the transformation of the precipitate then places a compressive stress on the crack, thus reducing its tendency to propagate further. This type of transformation, which takes place in response to stress rather than a change in temperature, is termed a martensitic transformation.

If less than 4 % yttria is added to zirconia and a fine grain structure is maintained, the material can be made 100 % tetragonal phase. These materials, called *tetragonal zirconia*, are stronger than partially stabilized zirconia. They can be

made even stronger by adding alumina precipitates [16].

Transformation-toughened ceramics [17], as these materials are called, are thus far limited to low-temperature use. The strength drops off linearly as the temperature approaches 1200 °C, where the tetragonal-to-monoclinic transformation normally occurs, as shown in Figure 4. Thermal fatigue is also important. If the materials soak at a high temperature, the microstructure changes slightly, reducing the strength. The high-temperature loss of strength and thermal fatigue characteristics are of concern since one potential use is for the cylinder lining in an "adiabatic" diesel engine. The high strength and good insulating ability of transformation-toughened ceramics make them a strong candidate for this use.

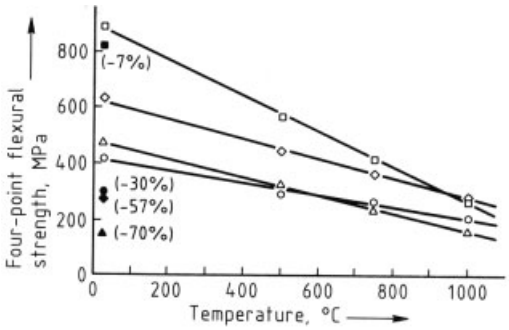

Figure 4. Bending strength of several commercial partially stabilized zirconia materials [18]
The companies and stabilizers: □■ NGK, Y_2O_3; ◇◆ Nilsen, MgO; △▲ Coors, 3 % MgO; and ○● Feldmühle, MgO. Filled symbols are room-temperature strength after 1000-h exposure at 1000 °C

3.3. Silicon Carbide

Silicon carbide [19], like most other covalent materials, is not easily sintered from a powdered compact. One process for producing a dense silicon carbide structure is to bond silicon carbide grains, produced by the Acheson process, together with fired clay, glass, silicon nitride, or by other proprietary means. These materials are quite creep resistant and thus can be used for furnace elements and for high-temperature structural use.

The first self-bonding silicon carbide was originally known as Refel silicon carbide, but is now generally known as *reaction-sintered silicon carbide.* It is made by forming a compact of silicon carbide grains with excess carbon and then at a high temperature immersing the compact into a molten silicon bath. The silicon and carbon are then allowed to react at some high soak temperature to form a silicon carbide reaction layer around the original grains. A variation in the process is to introduce the silicon in the vapor phase. Another variation is to introduce the excess carbon as a polymer binder during injection molding. Although the properties are usually considerably better than those of the glass-bonded or clay-bonded types, they are somewhat degraded by the residual unchanged silicon nearly always present in reaction-sintered silicon carbide.

In 1973 Prochazka of the General Electric Co. demonstrated that very fine grained SiC could be *pressureless sintered* with only a small amount of sintering aids, typically 0.5 % boron and 1 % excess carbon. The sintering temperature is 1900 – 2100 °C [19]. Later a similar process was patented to sinter α-SiC.

Both reaction-sintered silicon carbide and sintered silicon carbide have structural properties superior to those of the earlier bonded silicon carbides. The sintered silicon carbide, generally, is slightly superior to the reaction-sintered silicon carbide. Even better mechanical properties may be achieved by hot pressing silicon carbide to full density with boron and carbon or alumina sintering aids. Since hot pressing is expensive, this material is not widely produced.

3.4. Silicon Nitride

Silicon nitride [20], also a covalent material, is likewise not sinterable from powder without additives. Unlike silicon carbide, however, no sintering aids have been found that allow silicon nitride to be sintered to full density without very high pressures except for intentionally added 5 – 15 % glass phase. The sintering additives commonly used are Al_2O_3, Y_2O_3, MgO and rare earth additives. These sintering additives combine with silica present on the surface of the powder to form a glass that allows the material to densify during the sintering operation.

A sufficient amount of glass of low viscosity is necessary to achieve high densities; how-

ever, large amounts of glass of low viscosity degrade the important high-temperature properties of the material. Hence, an optimum is sought between high density, which results in good low-temperature strength, and additives that produce good high-temperature strength. Ideally, the additive after sintering would enter into solid solution with silicon nitride, but thus far only limited success has been achieved in finding such alloys.

A compound related to silicon nitride that has found successful commercialization is *SiAlON*, an aluminum silicon oxide nitride. The most common form has the formula $Si_{6-z}Al_zO_zN_{8-z}$, where z ranges from 0 to 4. Compounds with z near zero have good high-temperature strength, creep resistance, and oxidation resistance, whereas those with z approaching 4 have good low-temperature properties, i.e., strength, toughness, and abrasion resistance.

The properties of silicon nitride are strongly dependent on the additive. Table 5 compares properties of silicon nitride hot pressed with MgO additive, silicon nitride sintered with Y_2O_3 additive, reaction bonded silicon nitride, and SiAlON. Generally higher strengths are achieved both at low and high temperatures by hot pressing; however, this is not an easily commercialized process for mass production. Reaction-bonded silicon nitride, made by exposing a silicon powder compact to nitrogen at high temperatures, has poor low-temperature strength but good high-temperature strength.

In situ reinforced silicon nitrides containing high-temperature second phases [22] have increased creep resistance almost comparable to that of sintered α-silicon carbide [23]. They have, however, two advantages over silicon carbide. The thermal expansion coefficient of silicon nitride is lower than that of silicon carbide, and therefore the rapid thermal shock resistance of silicon nitride is better. Second, the strength of silicon nitride, even at high temperature, is quite high and generally much better than that of silicon carbide.

3.5. Ceramic – Ceramic Composites

There has been much activity in developing various ceramic – ceramic composites [24]. These include continuous fiber composites, whisker composites, layered composites, and particulate composites. Most of the activity has centered on two-dimensional woven continuous fiber composites made either by chemical vapor infiltration (CVI), or ceramic precursor polymer infiltration into woven ceramic cloth layed-up into two-dimensional layers. The most popular fibers are silicon carbide and more recently alumina. The advantages are extremely good flaw insensitivity, high thermal shock resistance and high strains to failure. Figure 5 illustrates an extreme example of the increased strength and strain to failure gained by adding carbon fibers to glass.

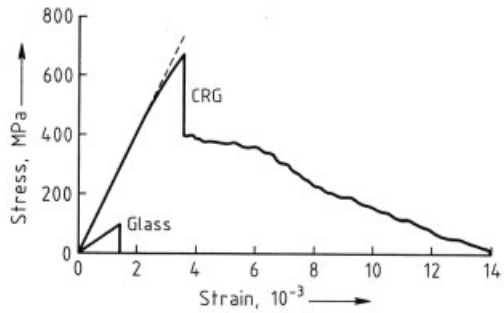

Figure 5. Stress – strain data for glass and carbon fiber reinforced glass (CRG) tested in bending [25]

At the point where the dotted line deviates from the solid line, the matrix, but not the fibers, begins to crack. At higher strain the fibers begin to crack and pull out of the matrix, but the composite still supports some load until all the fibers have fractured. As shown in the figure, the strain can be greater than 1 % before complete failure, a contrast to the normal 0.1 % strain before failure. This improvement in strain to failure is the most appealing attribute of fiber composites.

Most composites do not exhibit such high strains, but they do have strains two to ten times higher than those of monolithic ceramics. The second major advantage of the fiber composite is its insensitivity to flaws. Flaws introduced during machining or impact by high velocity grit have little effect on the fracture strength. The key to achieving high strength and large strains to failure is to have strong fibers and a weak interface between the fibers and the matrix. When a crack impinges on a fiber, the matrix/fiber interface debonds, allowing the crack to pass by without fracturing the fiber. Various coatings are applied to fibers to allow them to debond, but the most popular are carbon and boron nitride.

Table 5. The effects of additives on the properties of silicon nitride [21]

Material	Bending strength, four point, MPa			Modulus of elasticity, GPa	Coefficient of thermal expansion, 10^{-6} K^{-1}	Thermal conductivity, $W m^{-1} K^{-1}$
	RT∗	373 K	1648 K			
Hot pressed (MgO additive)	690	620	330	317	3.0	30 – 15
Sintered (Y_2O_3 additive)	655	585	275	276	3.2	28 – 12
Reaction bonded (2.45 g/cm^3)	210	345	380	165	2.8	6 – 3
SiAlON (sintered)	485	485	275	297	3.2	22

∗ RT = room temperature.

Application of high temperature may burn off the carbon and boron nitride coatings, limiting the fiber composites to generally about 1000 °C. Much research is focused on oxide coatings, which will readily debond but are not burned off at high temperatures. There are already a few applications where these materials have found a niche, but generally their cost is prohibitive. The fibers themselves cost from a few hundred to thousands of dollars per kilogram.

Whisker composites are much less expensive and easier fabricated into a variety of shapes. Alumina/silicon carbide whisker composites are already used widely for cutting tools. The advantage again is flaw insensitivity. When silicon carbide whiskers are added to alumina, the toughness increases from 4 MPa $m^{0.5}$ to 9 MPa $m^{0.5}$. This is not so large a flaw insensitivity as those of continuous-fiber composites, but it is sufficient to make them very useful. The largest difficulty facing whisker-reinforced ceramics are the environmental health and safety issues during manufacturing. These issues have discouraged some companies from introducing ceramic whisker composite products.

A third technology is layered composites. The layers are usually produced by tape casting and stacking, similar to multilayered microelectronic packages. The key to toughness is a weak interface between the layers. These composites are only in the research stage.

3.6. Other Materials

Several other materials listed in Table 3 have a wide range of uses. Glass ceramic has a wide range of compositions and uses. Cordierite has a low coefficient of thermal expansion and is widely used where good thermal shock resistance is required: cordierite is used for substrates in catalytic converters (see → Automobile Exhaust Control) and in the regenerator core of heat exchangers. Mullite is used much like alumina but at lower temperature. Its thermal shock resistance is better than that of alumina, and mullite of the proper composition and grain structure can be very resistant to creep. Zircon also has an excellent thermal shock resistance. Aluminum nitride has a high thermal conductivity.

4. Manufacturing Methods

The ability to produce specific components in a desired shape and size depends on the ceramic shaping and sintering processes (see Fig. 6).

The starting powders are mixed homogeneously with appropriate organic additives, solvents or water, and disperants, depending on the forming technique. In the next step, this mixture is shaped into a green body. This green body can be machined mechanically in an intermediate step, if it is not possible to obtain the desired shape by the ceramic shaping process itself. After pyrolysis of the temporary binders etc. that are necessary for shaping, the sintering step follows under temperature, time, pressure, and atmospheric conditions that depend on the material. During the sintering process the body suffers a 10 – 20 % shrinkage, making exact tolerances difficult. For this reason and because of the possible low-quality surface finish, costly mechanical machining with diamond tools is often required. These machining processes have also been considerably improved resently.

In special cases the piece that is to be sintered is subjected to hot isostatic pressure (HIP) under N_2 or Ar gas to optimize density and strength.

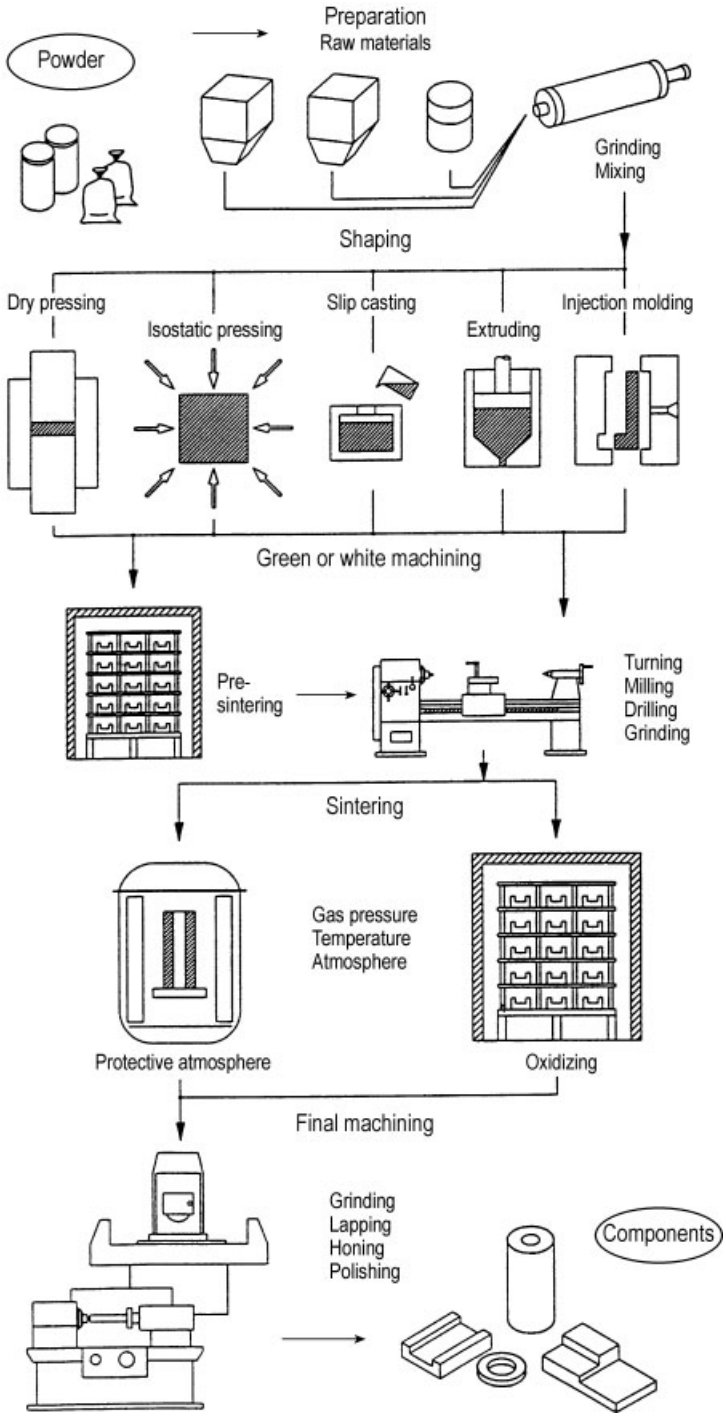

Figure 6. Production of engineering ceramic components

This is used for the production of ball bearings made of silicon nitride.

To avoid defects, nondestructive testing methods, such as ultrasonic testing, microfocus radiography, and proof tests, are applied. Non destructive testing has been developed to a high level and proven to be reliable, for example, for the quality control of silicon nitride ball bearings for space flight, and can be easily integrated into mass production.

Several engineering ceramic components are already produced in bulk quantities. In many cases the quantities are still rather small, but an impressive variety of shapes can be fabricated. The fact that ceramics are not easily machined for prototype parts has been a disadvantage; however, solid free form fabrication (SFF) techniques, which are CAD/CAM computer designed, can fabricat prototypes with a rapid turnaround time and are likely to be widely used in the future [26].

5. Uses

The application of engineered ceramics is widely spread over many fields of technical processes for which use is made of their unique properties. Furthermore, the high potential extends well beyond already established applications.

Compared with other structural materials, metals or polymers, ceramic materials will not play in future such a big role with respect to quantity. Their significance can be seen more in the fact that they will be found in key applications. They are applied in essential parts that are crucial for the functionality of the whole system, i.e., enabling technologies.

Some applications and possibilities for ceramics in chemical engineering are summarized in the following:

Fittings	
– Valves (cones, seats, relief nozzles)	Si_3N_4 (ZrO_2, SiC)
– Cocks (ball, cylinder, rotating-disks)	Al_2O_3, SiC, ZrO_2, Si_3N_4
– Seal disks	Al_2O_3, SiC
– Flow meters	Al_2O_3, ZrO_2
Pumps	
– Face seals	SiC, Al_2O_3
– Separation cans	Si_3N_4, ZrO_2
– Plungers	Al_2O_3, ZrO_2
– Impellers, cases	Al_2O_3, SiC
Fittings	
Bearings	
– Friction bearings, tube protectors	SiC
– Roller bearings	Si_3N_4
Heat exchangers	
– Tubes	SiC
Ventilators	
– Rotors, turbines	SiC, Si_3N_4
Wear protection	
– Linings	Al_2O_3
– Nozzles	Al_2O_3, B_4C, SiC, Si_3N_4
– Metal forming	Si_3N_4, ZrO_2
Thermocouple protection tubes	Al_2O_3, ZrO_2, Si_3N_4, SiC

Thread Guides. Ceramic thread guides have been used for many decades now and are a good example illustrating the reliability of ceramics. Generally, Al_2O_3 is the material of choice, but in special cases also TiO_2 or Si_3N_4. Due to the application of ceramics, durability of thread guides in textile machines has been improved considerably. About 3×10^6 km of textile thread at a speed of 6000 m/min runs through a thread guide. The costs for the ceramic material is less than DM 1 per tonne of processed textile threads.

Fittings. In the broad field of fittings for the measurement and control technology, such as cocks, valves, pushers, and gates for gaseous, liquid, and powder media, ceramics also have attained an increased significance, especially, due to their outstanding wear resistance. As an example aluminum oxide seal disks have become standard equipment. Furthermore, fittings equipped with engineering ceramics can be found in mining, steel and metallurgical engineering, oil and gas industry, power stations, chemical and petrochemical plant, paper and pulp industry flue gas desulfurization, and incineration plants for refuse, chemicals, and toxic media [27], [28]. Although the price for fittings equipped with ceramics is slightly higher, cost reduction results from the broader cost/profit reduction, in the crucial improvements in functionality, lower servicing requirements, and greater durability. For example, inductive mass flow meters are now made of aluminum oxide or zirconium oxide [29].

Mechanical Face Seals. A good example of successful substitution of metals by ceramics is the ring and its counterpart in modern face seals.

The requirement for resistance to wear, corrosion, and heat under increasingly severe conditions, cannot be met by other materials. Hard metals and materials based on carbon have from the beginning substituted for metallic seal materials; however, these are being increasingly replaced first by aluminum oxide and then later by silicon carbide [30]. The main applications for pumps are chemical and engineering technology, petrochemistry, energy generation, water supply, and environmental technology [31]. As a result of increased production rates, costs could be reduced to a level such that silicon carbide can even replace aluminum oxide in the seals of the automotive cooling-water pumps.

Friction Bearings. A steadily increasing application is magnetically coupled, seal-less pumps which enable an absolutely leakage-free and therfore emission-free and environmentally harmless transportation of media [31].

The impeller of the pump has a bearing which is lubricated by the pumped medium itself. Because of their high wear resistance, especially radial and axial friction, ceramic bearings have proved to be an excellent application. However, in some cases insufficient running quality has been noticed; here the application of roller bearings is recommended.

Separation Can. In the magnetically driven fluid pump described above, a separation can is applied as a separator between the media and the environment. Using a ceramic material, such as zirconium oxide or silicon nitride, is advisable here due to its corrosion resistance [30]. Because of the required low wall thickness, production of this component is not simple but is feasible.

Pump Cases and Wheels. The classical materials for pumps, cast iron and cast steel, are sometimes unsatisfactory because of corrosion and erosion; polymers are often unsuitable due to insufficient thermal resistance. In such cases, only ceramic materials can meet the requirements (Tables 4 and 6) [33].

Formerly stoneware was primarily used but increasingly has been replaced by polymers and metals. However, where these do not work entirely satisfactorily for technical or cost reasons, it is possible to use ceramics, especially aluminum oxide and silicon carbide. These complex components are a real challenge for the ceramic producer.

Silcon carbide pumps are excellently well suited for titanium dioxide production by the sulfate process, where metal pumps usually fail within a few months [34].

Roller Bearings. Silicon nitride has proved to be an ideal material for roller bearings. Hybrid bearings with ceramic rolling elements and metallic bearing races can be found in many applications [35]. In chemical environments, mainly full ceramic bearings are used [36]. Properties, such as low weight, low thermal expansion, high hardness, low wear, high strength and temperature stability, high chemical resistance, nonmagnetism, and electrically insulating behavior are important advantages. These result, for example, in higher bearing stiffness, lower friction and therefore lower rate of heating, as well as the possibility of operating at higher speeds.

Finally, ceramic ball bearings can be used in various media and still show excellent running properties, especially in case of deficient lubrication or even lack of lubrication. For high-speed bearings centrifugal force is lower because of the lower weight, leading to better kinematic properties. These result in a more favorable roll-off behavior as well as a higher lifetime.

Silicon nitride roller bearings are also applied as stiff and high-speed, high-precision spindle bearings in machine manufacturing, aeronautics, and space technology. But also in common machine manufacturing, in automotive and technical engineering, in chemical engineering, in instrument technology as well as in measurement and control technology, silicon nitride roller bearings have found more and more in roads and are increasingly being considered for wider use. Hybrid bearings in vacuum pumps, for example, have become standard [35], as they lead to an increased lifetime with a higher reliability. Furthermore, hybrid bearings are also increasingly applied to high-speed drilling machines in dental technology, especially due to better sterilizability. For full ceramic bearings a wide field of application is opening up in media-lubricated pumps for chemical environments [36]. Unlike friction bearings based on silicon carbide, they do not have dry-running problems during start-up.

Table 6. Comparison of materials used in pumps

	Stress ∗					
	Mechanical	Temperature	Corrosion	Abrasion erosion	Tribological	Cavitation
Cast iron	+	+	–	+	+	+/–
High-/low-alloy iron/cast steel	++	+	++/+	–	–	+/–
Polymers	+/–	–	+/–	+	+/–	+/–
Ceramics	–	++	++/+	++	++	+/–

∗ ++ Very high resistance, + high resistance, – limited resistance.

Production of ceramic ball bearings, rollers, and rings is a great challenge, in view of fabrication processes and final grinding technology. Due to the local stress concentration during use, microstructures need to be very homogeneous with a grain size on the order of 10 µm, i.e., without any larger grain, pores, and/or inclusions. Not long ago, this was only possible by hot isostatic pressing at up to 1000 bar. Now a suitable quality can be achieved by use of cost-effective gas-pressure sintering and careful processing [37]. Mechanical finishing to required dimensions and surface quality require special processes that need to be further optimized. Experience developed with metals can only be partially utilized. Structural design of ceramic bearings is also quite different from those made of metals [38].

High-Temperature Heat Exchangers and Ventilators. For heat exchangers, silicon carbide is highly suitable because of high thermal conductivity as well as its dimensional reproducibility [39]. However, construction is quite difficult and costly; for this reason a wider application has not yet been achieved.

At present, the movement of gas at high temperatures is still a problem to be solved. Of great interest is a ventilator equipped with ceramic elements which is beginning to show promising results [40].

Uses in Engines. A list of applications for advanced structural ceramics is shown in Table 7.

In a *reciprocal engine,* a number of ceramic parts may replace their current metal equivalents. Parts that must be wear resistant, such as the rocker arm chip and the tappet face, may be replaced with ceramics. Ceramics are also being considered for the cylinder lining and the piston caps, both for wear resistance (ceramics may not require lubrication) and heat resistance: the cylinder may be allowed to get much hotter, improving the efficiency of the engine and requiring little or no cooling. These engines are often called *adiabatic* engines, although they are not adiabatic in the strictest sense. In a *turbocharged* engine a ceramic rotor is a good candidate because its moment of inertia is lower than those of metal rotors.

Table 7. Uses of advanced structural ceramics in engines

Diesel	Turbocharger	Gas turbine
bearings	bearings	blades
cylinder liners	heat shields	heat exchangers
glow plugs	housing	manifolds
manifolds	turbocharger rotors	rotor shaft
piston caps		shrouds
piston rings		stators
prechamber		
seals		
tappets		
valves		
valve seats		

A second type of engine for which ceramics have great potential is the *gas turbine engine*. The primary purpose of ceramics would be to improve the thermodynamic efficiency by allowing the engine to operate at a higher temperature. The expected temperature of the turbine blades is 1200 – 1350 °C. Not only are high-temperature strength and impact resistance important, thermal shock resistance is also necessary. The candidates for these uses are silicon carbide and silicon nitride.

Other Wear Resistance Applications. Important applications for high-performance ceramics can be found in almost every field of chemical production, where processes can be disrupted by mechanical and/or chemical wear

which leads to significant economic loss or even impact on the environment. In addition, as previously discussed, ceramics can be introduced for highly wear related parts, such as for the milling and pouring of loose material or abrasive and corrosive slurries, liquids and gases. In pipelines, cyclones, locks, etc., ceramic wear-protected parts can be found, particularly aluminum oxide [41], [42]. Ceramics have become the standard material, for example, for welding nozzles which are made of aluminum oxide; sand-blasting nozzles which are made of silicon carbide or the even harder boron carbide and aluminum titanate, which are used in the atomization of metals.

Applications also cover tools for metal working (cold and melt processing) as well as linings, flowing, sealing, and dosing components used in metallurgy. All these examples indicate a clear advantage in applying ceramic materials rather than the competing metallic materials. However, it has to be pointed out that in all these cases ceramic materials can only be applied successfully, if designed properly to avoid high tensile stress concentrations.

6. References

1. H. Vogel: *Steinzeug,* 4th ed., **vol. 22,** Verl. Ullmann 1982, pp. 219 – 240.
2. Coors Porcelain Co.: *AD-995 Alumina.*
3. A. A. Griffith, *Philos. Trans. R. Soc. London* **A221** (1920) 163.
4. E. Orowan, *Rep. Prog. Phys.* **12** (1949) 185.
5. R. W. Davidge: *Mechanical Behavior of Ceramics,* Cambridge University Press, Cambridge 1979.
6. J. B. Wachtman, *Mechanical Properties of Ceramics,* J. Wiley and Sons, New York 1996.
7. P. F. Becher et al.: "Microstructural Design of Silicon Nitride with Improved Fracture Toughness: I, Effects of Grain Shape and Size," *J. Am. Ceram. Soc.* **81** (1998) 2821 – 2830.
8. W. Roger Cannon, T. G. Langdon, *J. Mater. Sci.* **18** (1983) 1 – 50.
9. F. R. N. Nabarro: *Report of a Conference on Strength of Solids,* The Physical Society, London 1948, p. 75. C. Herring, *J. Appl. Phys.* **21** (1950) 437.
10. R. L. Coble, *J. Appl. Phys.* **34** (1963) 1679.
11. W. D. Kingery, H. K. Bowen, D. R. Uhlmann: *Introduction to Ceramics,* J. Wiley & Sons, New York 1976, p. 240.
12. T. Ohji, A. Nakahira, T. Hirano, K. Niihara, *J. Am. Ceram. Soc.* **77** (1994) 3259 – 3262.
13. J. D. French, J. Zhao, M. P. Harmer, H. M. Chan, G. A. Miller, *J. Am. Ceram. Soc.* **77** (1994) 2857 – 2865.
14. C. E. Curtis, *J. Am. Ceram. Soc.* **30** (1947) 180.
15. R. C. Garvie, R. H. Hannink, R. T. Pascoe, *Nature (London)* **258** (1975) 703.
16. K. Tsukuma, K. Ueda, K. Matsushita, M. Shimada, *J. Am. Ceram. Soc.* **68** (1985) C-56.
17. W. R. Cannon: "Transformation Toughened Ceramics for Structural Uses," in J. B. Wachtman, Jr., (ed.): *Structural Ceramics,* vol. 29, Academic Press, Boston 1989.
18. D. C. Larsen, J. W. Adams: "Long-Term Stability and Properties of Partially Stabilized Zirconia," presented at 22nd DOE Contractors Coordination Meeting, Dearborn, Mich., Nov. 1984.
19. S. Prochazka in J. J. Burke, A. E. Gorum, R. N. Katz (eds.): *Ceramics for High Performance Applications,* Brock Hill Publ. Co., Chestnut Hill, Mass., 1947, pp. 239 – 252.
20. F. F. Lange, *Int. Met. Rev.* **25** (1980) 1.
21. R. N. Katz, *Science (Washington, D.C.)* **208** (1980) 841.
22. C. J. Gasdaska, *J. Am. Ceram. Soc.* **77** (1994) 2408 – 2418.
23. J. E. Lane, C. H. Carter, R. F. Davis, *J. Am. Ceram. Soc.* **71** (1988) 281 – 295.
24. K. K. Chawla: *Ceramic Matrix Composites,* Chapman & Hall, London 1993.
25. D. H. Bowen, D. C. Phillips, R. A. J. Sambell, A. Briggs, *Mech. Behav. Mater. Proc. Int. Conf., 1st 1971,* 1972, vol. 5, pp. 123 – 134.
26. Naval Research Reviews, Office of Naval Research **3,** 1998 Volume L.
27. R. Berchem: "Erfahrungen mit Ingenieurkeramik," *VDI Z.* **130** (1988) 69 – 73.
28. G. T. Robertson, R. H. Owens: "Valves and actiators," *Proc. 3rd. Int. Conf. STI* (1990) 49 – 58.
29. D. Lupton, A. Reckziegel, U. Stevens: "Magnetisch induktive Durchflußmeßgeräte," *Symp. Mat. Forschung,* PLR-Forschungszentrum Jülich, Jülich 1994, p. 865.
30. D. Zeus: "Ingenieurkeramik in Gleitringdichtungen und Magnetkupplungen. Verschleiß und Korrosion," *DGK: Verschleiß u. Korrosion* (1990) 20 – 49.

31. D. Zeus, *J. Soc. Trib. Lubr. Eng.* (1992) 73 – 77.
32. I. Veness, W. Wiesenrieder, *Oesterr. Chem. Z.* **93** (1992) no. 10, 283 – 285.
33. W. Prechtl, H. Tischner: "Einsatz von ingenieurkeramischen Werkstoffen in Pumpen und Armaturen. Verschleiß und Korrosion," *DGK: Verschleiß u. Korrosion* (1990) 141 – 158.
34. H. Holdschick: "Siliciumcarbid als verschleißfester und korrosionsbeständiger Werkstoff für Kreiselpumpen. Verschleiß und Korrosion," *DGK: Verschleiß u. Korrosion* (1990) 63 – 74.
35. Firma Leybold AG, brochure 175.2801.
36. M. Popp: "Innovative Lagerungskonzepte," *Chem. Anlagen Verfahren* **27** (1994) no. 5, 186 – 188.
37. G. Wötting, H. J. Böhmer, G. Martin: "Entwicklung kostengünstiger Herstellungs- und Prüfverfahren für Si_3N_4-Wälzlager hoher Zuverlässigkeit," *Symp. Mat. Forschung,* PLR-Forschungszentrum Jülich, Jülich 1994, p. 866.
38. Firma Wemhöhner & Popp, Herzogenaurach: Cerobear User-Workshop 1995.
39. R. N. Katz: "High performance structural ceramics prepare for the 1990s," *Industrial Ceram.* **11** (1991) no. 4, 199 – 201.
40. Sonderforschungsbereich TU Clausthal.
41. N. A. Stone, A. W. Peterson: "Review on the status of advanced ceramic technology in USA with emphasis on wear resistant ceramics in the mining industry," *Mat. Sci. Forum* **34 – 36** (1988) 51 – 57.
42. H. Hoppert: "Verschleißfeste keramische Werkstoffe, ein wirkungsvoller Schutz für Anlagen in der Aufbereitungs- und Fördertechnik. Verschleiß und Korrosion", *DGK: Verschleiß u. Korrosion* (1990) 85 – 94.

Ceramics, Ceramic – Metal Systems

VICTOR A. GREENHUT, Rutgers – The State University, Piscataway, New Jersey 08854, United States

RICHARD A. HABER, Rutgers – The State University, Piscataway, New Jersey 08854, United States

1. Introduction 1
2. Fundamentals of Bonding 2
2.1. Wetting 2
2.2. Bonding and Adherence Mechanisms 3
2.3. Stresses in Ceramic – Metal Systems 4
3. Glass-to-Metal Joining 5
3.1. Glass-to-Metal Seals 6
3.2. Enamels 7
3.2.1. The Enameling Process 7
3.2.2. Theory of Adhesion 10
4. Ceramic-to-Metal Joining 12
4.1. Sintered Metal Powder Process (SMPP) 12
4.2. Metal Powder – Glass Frit Method . 13
4.3. Active Metal Process 13
4.4. Gas – Metal Eutectic Process 13
4.5. Pressed Diffusion Joins 14
4.6. Vapor-Phase Metallizing 14
4.7. Liquid-Phase Metallizing 14
4.8. Electroforming 14
4.9. Graded Powder Process 14
4.10. Nonmetallic Fusion Process 15
5. Thin Films and Coatings 15
5.1. Uses 15
5.2. Deposition Techniques 16
5.2.1. Chemical Vapor Deposition (CVD) . 16
5.2.2. Evaporation 17
5.2.3. Sputtering 19
5.2.4. Plasma Spraying 20
5.2.5. Sol – Gel Processing 20
5.2.6. Ion Implantation 21
6. Ceramic – Metal Composites 21
6.1. Cermets and Cemented Carbides . 22
6.2. Metal Matrix Composites 23
7. References 25

1. Introduction

The use of ceramics and metals in conjunction allows for a combination of properties. Ceramics, here to include glasses, typically show strong covalent/ionic bonding with chemical inertness, high elastic modulus (stiffness) and compressive strength, low electrical and thermal conductivity, low friction and wear behavior, and transparency to electromagnetic radiation. Metallic bonds result in complementary properties, for mechanical ductility is not usually exhibited by ceramic materials. Ceramic – metal systems have been used since prehistoric times in such forms as enamels on metal and metallic decoration on ceramics. The contributions provided by both the ceramic and the metal make such systems important in traditional and advanced technologies.

The ceramic and metal may be coupled as macroscopic pieces or engineering structures. These find use in such components as semiconductor substrates, turbines, electronic equipment, and lamps. Coatings of ceramic or glass can be placed on metal or vice versa. Enamels used as corrosion protection and decoration are an example. A microscopic combination of materials can be produced in the form of a ceramic – metal composite, often termed a cermet when the ceramic is the principal phase. Such materials are used for carbide cutting tools, high-temperature jet orifices, and turbine parts.

In all cases it is necessary to bring the ceramic and metal into intimate contact and create a bond between materials. The distinct differences between ceramic and metal in terms of bonding and properties make it necessary to consider the major factors required to join the two classes of materials and satisfy the physical, chemical, and mechanical properties required of the system.

2. Fundamentals of Bonding

Metals and ceramics may be joined in the solid state or with one component deposited as a liquid or vapor. Intimate contact occurs between the materials by physical means. Bonding may occur by surface or interfacial interaction. Dependent on the chemical nature of the materials,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_055

solution and redox reactions may proceed, thus promoting adherence. Quite commonly a liquid phase is present during fabrication of a joint, coating, or composite. In such a case, the wetting behavior of the liquid phase on the solid is quite important. Solution and/or redox reactions are likely, which can affect the wetting behavior and adherence. Wetting is usually required for adherence between ceramic and metal.

A further factor that may affect the bond at a ceramic-to-metal interface is stress between the ceramic and metal. Such stresses may arise in the presence of an applied stress because of differences in elastic constant. As the material cools from fabrication at elevated temperature, differences in thermal expansion can give rise to residual stresses. Changes in temperature during use can have a similar effect. Forces at the ceramic – metal interface can exceed the strength of the bond between materials, preventing a macroscopic join, causing delamination of a coating, or affecting composite properties.

2.1. Wetting

Commonly a bond between metal and ceramic is created by either melting the metal or softening a glassy ceramic material. The success of the bond depends on the wetting or capillary action [1], [2]. To join two pieces of material, the liquid must spread across the surface or interface. In a composite material, a phenomenon called sweating occurs in a nonwetting situation such that liquid is exuded at the surface of the material. This may cause nonuniformity or porosity, or it may make the composite impossible to produce.

A number of methods have been developed to determine the wetting response of the liquid phase. These include the geometry of liquid in a capillary or a sessile drop on a flat substrate, the rise of liquid in a capillary, and wicking in a porous material [2–4]. Perhaps the simplest and most common approach is the sessile drop method, which is shown schematically in Figure 1. A drop of liquid, shown in cross section, is observed at elevated temperature, and the *contact angle* θ is measured. The relationship is as follows:

$$\gamma_{lv}\cos\theta = \gamma_{sv} - \gamma_{sl}$$

where γ_{lv}, γ_{sv}, and γ_{sl} are the interfacial energies between liquid – vapor, solid – vapor, and solid – liquid, respectively. For $\theta > 90°$, a nonwetting situation occurs, while if the angle is $<90°$, the liquid wets the surface. If $\theta = 0°$, the liquid spreads over the surface as a thin layer. The surface energy can be obtained from this experiment when equilibrium is achieved [3–6], providing an estimate of interfacial energy and theoretical bond strength if extrapolated to room temperature.

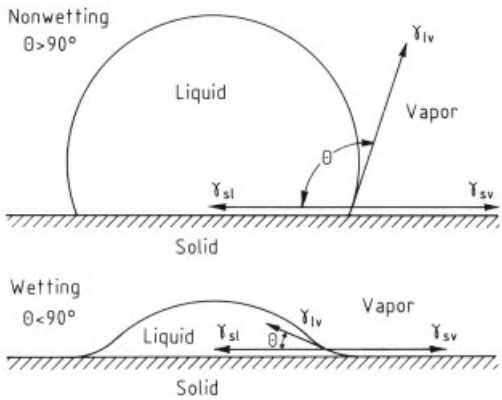

Figure 1. Sessile drop configuration

The wetting angle itself is valuable in evaluating the system. The contact angle may change with small changes in the composition of the atmosphere, liquid, or solid phases. Temperature and time also play an important role. In industrial processes, the liquid is seldom brought to equilibrium as it reacts with the solid oxide films, and contaminants on the surface, as well as the atmosphere of the furnace. While interfacial energy cannot be obtained in this case, the wetting information in itself indicates the degree to which the liquid spreads or infiltrates the solid. Although a strict relationship cannot be drawn between wetting and bonding, usually a good wetting response and a degree of bonding coincide. However, a reasonable bond strength may develop in a nonwetting case, and adequate bond strength may not develop in a wetting case. The wetting response should, therefore, be regarded as a necessary but not a sufficient condition for good ceramic-to-metal bonding.

For ceramic – metal sealing, a transition from metal oxide to metal layers may be used to promote wetting. Such a layer may be ap-

plied as a metallizing layer in the form of a moly1 – manganese paste (see Chap. 4). Such a material may provide a bond to the ceramic because of the oxide ceramic and glass phases present while providing a low interfacial energy with a molten metal because of a metal-rich surface structure [7].

In glass – metal joining, the metal is commonly preoxidized [3], [5], [8]. The oxide layer may be more compatible with the oxide bond of the glass, thereby lowering the interfacial energy between liquid and solid [3], [5] and possibly promoting wetting. Glass-to-metal and ceramic-to-metal bonding often involves a forming gas, which may affect wetting by oxidizing or reducing the metal or ceramic, altering interfacial energies, or diffusing a particular species into solid or liquid. These techniques suggest that some transition between metallic and ionic/covalent bonding enhances the wetting response.

2.2. Bonding and Adherence Mechanisms

The mechanisms of ceramic – metal bonding have been classed in various ways, but perhaps a most general view is that physical, mechanical, and chemical bonding mechanisms can occur. These effects do not occur independently, and all may be involved in a ceramic – metal system.

Physical bonding may be regarded as the bonding effect that occurs when a pair of flat surfaces are brought within atomic-interaction distances. The work of adhesion is given by the specific energy of the metal and ceramic minus that of the metal – ceramic interface. This work of adhesion can yield a theoretical breaking stress near that of the metal or ceramic [4], [9]. Nevertheless, actual interfaces can fail at stresses many orders of magnitude smaller as the result of flaws in the bond.

Mechanical bonding relates to the interlocking structure of rough surfaces, which gives rise to either frictional effects in shear or interlocking of microstructure, which provides mechanical anchors in tension. When ceramic and metal are joined, liquid metal or glass can penetrate pores and cavities, a ductile metal solid can conform to the ceramic surface, or a vapor phase may deposit in surface asperities. A roughened surface, which promotes mechanical bonding, also increases the surface area for physical bonding. However, a mechanically interlocked structure could in itself lead to a bond.

Chemical bonding occurs in a transitional oxide film between the metal and the enamel and is responsible for adherence, the oxide film actually bonding the enamel to the metal. Many complex reactions take place in the transition zone. One accepted model hypothesizes that the adherence is the result of metal-to-metal bonds between the atoms of the base metal and some metallic ions in the enamel. The enamel at the interface must be saturated with the lowest oxidation state oxide of the base metal [5].

Chemical reaction can produce a transition structure that gradually varies from metallic to ceramic. Often such a structure is developed by design. Metals may be preoxidized to provide an oxide that can react with molten glass, saturating the glass near the interface in the metal species. Alternatively, the oxide may adhere to the metal, providing a ceramic surface for the bonding glass. A transition structure may be provided by a mixed oxide and metallic material applied to the ceramic, such as in the sintered powder process, which provides a ceramic-like interface at the ceramic and a sufficiently metallic one at the metal bond [1–9].

In practice, complex combinations of these mechanisms may occur. For example, in porcelain enameling, a preoxidized metal surface may undergo galvanic attack, roughening and pitting the metal surface locally, changing the local chemistry of the glass, and redepositing dendritic anchors of metal at the interface. Features of physical, mechanical, and chemical contributions to bonding may all be identified, but their relative contributions may be difficult to assess. Chemical interaction can produce material properties rather different from those of either original material (see also Section 3.2.2).

2.3. Stresses in Ceramic – Metal Systems

Thermally Induced Residual Stresses. The thermal expansion of a metal is generally considerably greater than those of ceramics and glasses, as can be seen in Figure 2. An unconstrained metal may contract significantly more than the ceramic, but if the metal is constrained, the differential expansion can result in

residual stresses in the ceramic and the metal. These stresses add to those resulting from applied forces, thermal gradients, phase transformations, and rapid cooling. This residual stress in itself or in combination with external stress may cause failure of the bond.

The greater shrinkage of a metal in a flat sheet causes net compressive stresses in the ceramic and tensile stresses in the metal. Such a situation is favorable because a ceramic typically shows far greater compressive than tensile strength: a residual stress can enhance properties such as fracture strength and wear resistance. However, if the stress is too great, failure may occur in the interface or possibly in either material.

Several approaches may be used to limit the thermally induced residual stresses. The ceramic and metal can be chosen to minimize the mismatch in thermal expansion. For example, molybdenum and tungsten reasonably match several ceramic and glass compositions and can be used where their poor oxidation resistance is not a limitation. For platinum and chromium, several glasses and ceramic compositions show similar expansion, but expense and useful bonding technology, respectively, can limit application. A common alloy in glass sealing, ASTM F15, is a set of iron, 28 – 29 wt % nickel, and 17 – 18 wt % cobalt alloys. These alloys show expansion behavior similar to that of borosilicate glasses, some showing an extraordinary match over both low and high temperature ranges. Such alloys are often referred to by the trade name *Kovar* (Carpenter Technologies). These alloys also show ductility, do not usually embrittle, and can be soldered, brazed, and welded. Some glasses and ceramics have been designed to match the expansions of particular alloys, e.g., lead silicate and other enamel compositions applied to steels.

Another approach to matched expansion is the *Housekeeper seal*, in which a very thin piece of metal is used. The slight size, low yield strength, and ductility of the metal allow it to accommodate the induced stresses [3], [8].

Continuum mechanics can predict the magnitude of residual stresses that result from thermal expansion differences [3], [8], [10] and can be used to design a system configuration or applied loads such that critical stresses are not exceeded. One approach is to design a *compression seal* so that the ceramic is placed in compression and the metal is placed in tension. A compressive seal can be used to obtain a purely mechanical bond between solid metal and ceramic during fabrication. One way is to use the differing thermal expansions to produce a "shrink fit" during cooling. A pair of concentric cylinders with the metal as the outer sleeve is an example, the greater contraction of the metal on cooling resulting in a strong, tight fit. However, different stress components can be in tension or compression for the same location in a material. Caution must be taken not to exceed the fracture stress for any tensile stress component in the ceramic or the interface.

Elastic Modulus Effect. The difference in the elastic modulus or stiffness of the ceramic and the metal components of a system or composite can result in material or interfacial failure under applied and residual stresses. This situation is shown in Figure 3. A ceramic is typically stiffer than a metal, as shown by the greater slope of the ceramic curve. For this example the ceramic has been chosen to be stronger in tension than the metal. If the metal is the massive member in the system and is loaded to a stress σ_{m} at the ceramic – metal interface, it will experience a strain $\varepsilon^{\mathrm{c}}_{\mathrm{m},}$. The deformation or strain must be compatible across the interface, thereby imposing a stress σ^{c} on the ceramic or interface because of the higher modulus of the material. Even when the ceramic is much stronger than the metal, the failure stress may be exceeded, as shown by the x, which indicates failure stress.

This effect can lead to failure of a ceramic – metal joint or coating. For composite material, the situation may be more complex as sharing of load may be desired for properties such as elastic modulus. However, when a failure crack propagates, failure at the ceramic – metal interface may be desirable so that the crack is diverted and fracture energy is expended in the process. Such considerations may be complex, involving the properties of ceramic and metal, chemical interaction zones between materials, and microstructural and geometrical factors. Predictive theoretical treatments for such materials are being developed.

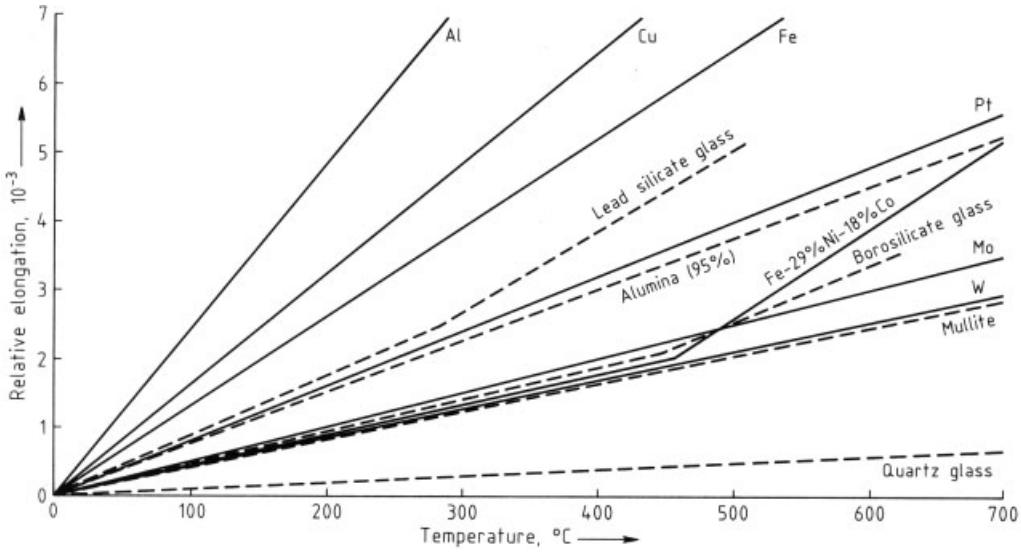

Figure 2. Thermal expansion of typical ceramics (dashed curves) and metals (solid curves)

3. Glass-to-Metal Joining

The bonding of glass and metal can be divided into two general categories, glass-to-metal seals and enamels. These two areas are distinguished by use, glass-to-metal seals being used primarily in electronic applications such as hermetic sealing of lamps and devices, isolation of electrical feed throughs, packaging, and vacuum-tube technology. Enamels are used principally as corrosion protection and decorative coatings on such products as domestic appliances, bathroom fixtures, and cookware.

Enamels and glass-to-metal seals are also distinguished by the base metal. Most enamels are applied to low-carbon steels and cast iron, while glass-to-metal seals are commonly used with nonferrous metals, stainless steels, and the iron – nickel – cobalt Kovar alloys.

A third distinguishing feature is how the carbon and hydrogen are treated. These species may be either in the base metal or on the surface as hydrocarbons. These elements combine with oxygen at elevated temperature and can form gaseous bubbles in the molten glass. In the case of glass-to-metal sealing, the base metal is cleaned of hydrocarbons and then decarburized or the hydrogen baked out. In enameling, higher levels of carbon and hydrogen are permissible, and in many cases these elements are not removed. The formation of bubbles in the enamel is considered undesirable if it weakens the bond with the base metal, results in a large bubble that exposes base metal, or causes aesthetic problems in surface roughness or color. Instead of preventing bubbles from forming, a controlled fine-sized bubble structure is introduced to act as a getter for evolving gases, thus avoiding large, objectionable bubbles in the glass.

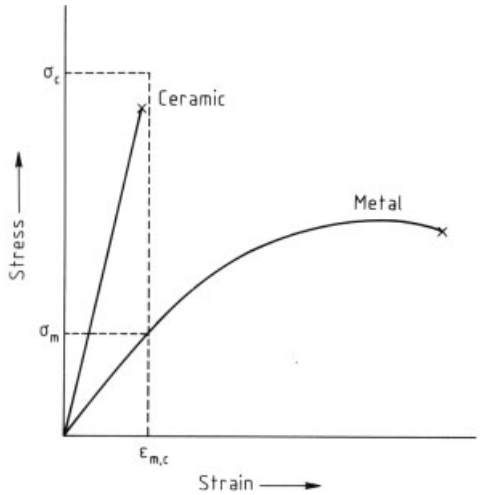

Figure 3. Difference in stresses for a bonded ceramic and metal under applied stress

3.1. Glass-to-Metal Seals

A good glass-to-metal seal requires a glass and metal with thermal expansions that match adequately over the temperature range of fabrication and application. The base metal is generally cleaned and then decarburized and/or the hydrogen baked out. Other products, such as nitride in the case of titanium alloys, may need to be removed. The metal generally is oxidized to provide a tightly adherent layer of oxide that remains adherent during forming of the glass-to-metal join. The oxide typically is wet by the glass and dissolves in the glass to yield a structure with sufficient chemical bonding [5]. Glass-to-metal sealing has been reviewed [3], [8], [11].

Metals commonly used for glass-to-metal joining include tungsten, molybdenum, titanium, tantalum, nickel – iron alloys, iron – chromium alloys, iron – nickel – cobalt alloys, platinum, copper, gold, and copper-clad nickel or iron – nickel Dumet alloys. A large range of borosilicate and lead silicate glasses are commonly used in glass-to-metal seals; borosilicate glasses are most commonly used. Where copper and copper-clad alloys or iron – chromium alloys such as stainless steels are used, lead silicate glasses are more common.

The glass is usually applied as a frit, bead, or paste. A solid fabricated glass shape may be heated and softened to the required viscosity by electrical, flame, or radio-frequency heating, and then pressed into contact with the preoxidized metal to bond. Frequently a protective or special forming atmosphere is required to promote wetting and bonding. Careful cooling or annealing avoids residual stress because of excessively rapid cooling.

A typical sequence of operations is that of the specialty Kovar alloy ASTM F15 (1983) (Fe – 29 % Ni – 17 % Co) glass-to-metal sealed with borosilicate glass. This system is common because of the excellent thermal expansion match between materials, good wetting response, and strong bond. It is used for microelectronics packages, electrical and high-frequency feed throughs, laser cavity-to-window bonding, and other diverse applications.

The Kovar is first degreased, chemically cleaned, and decarburized in a hydrogen, hydrogen/nitrogen, or dissociated ammonia atmosphere with various water levels at 900 – 1100 °C. The temperature should be higher than final glass-to-metal sealing temperature. A typical cycle is 1100 °C for 10 – 15 min in a hydrogen/nitrogen/water vapor atmosphere, typically 1 vol % water and as little as 10 vol% hydrogen.

The next step is to preoxidize the metal in a controlled fashion. While air and natural gas can be used, the variability of components such as water in these gases makes process control difficult. Pure oxygen gives a thick, flaky coating that does not penetrate metal grain boundaries. An atmosphere of nitrogen/water/hydrogen at 800 – 1050 °C is used: an atmosphere of nitrogen/1 vol% water/0.4 vol% hydrogen at 1000 °C for 10 min produces good oxide structure with optimum thickness and grain boundary penetration.

The final step is the glass sealing. The glass preform or frit is assembled or applied. Firing is performed in a nitrogen-based atmosphere, such as that used for controlled preoxidation, or an exothermic (natural) gas atmosphere is employed. Sealing is done at ca. 1000 °C, the precise time and temperature depending on part geometry, glass viscosity, and wetting response [12].

3.2. Enamels

Enamel is a fused, vitreous, superficial coating on metals or a decorative pattern on a glass surface. Enamels are similar in properties to ceramic glazes.

History. The origin of enameling metals goes back to the Egyptians, centuries before Christ. Enamels were used to provide elaborate designs in jewelry. The materials enameled were typically gold and silver. Later more metals were enameled; bronze and copper jewelry inserts were coated with enamels. By variations in the composition of the enamel the glassy coating could be made to hide defects in the metal.

The application of enamels to iron and steel made enamels available for uses other than jewelry. Enamels were sought to provide decorative coatings on a wide variety of cast-iron shapes. In the earliest operations, cast iron was heated to redness, the enamel was dusted upon the metal as a fine dry powder, and as the powdered glass

stuck to the hot metal it softened and adhered to the surface. This operation required a number of successive coating steps to obtain a layer thick enough to provide the desired effect. Many problems were related to the firing: enamels blistered, turned black, lacked the desired color, and exhibited poor adhesion.

Advances corrected or improved these problems. Adherence was related to the surface properties of the metal. Surface pretreatments, such as sand blasting, cleaning, and pickling, evolved. New, cheaper sources of raw materials were found, e.g., borax, allowing lower firing temperatures and producing cheaper enamels. The application of enamels as a suspension, rather than dry powder, was a great improvement. Greater control of the chemistry was found to provide newer, more stable enamels.

3.2.1. The Enameling Process

The enameling process can be simplified into three basic steps: (1) metal preparation, (2) ground-coat application, and (3) cover-coat application. Figures 4 and 5 provide process flow diagrams for three commonly enameled metals – cold-rolled steel, low-carbon steel, and cast (enameling) iron [13].

Metal Preparation. The metal surfaces must be free from grease, machining oils, dirt, oxidation products, and rust. A roughened or suitably etched surface is desirable for most metals. The exact metal preparation depends on the composition of the base metal and the type of enamel to be applied.

No pretreatment is required for gold or silver, provided the metal surfaces are clean. For copper and bronze the cleaning step is often followed with a 2 – 4 % nitric acid pickle and a clean water rinse. Aluminum and its alloys require special surface pretreatment, including cleaning, chromatizing, and preoxidation.

Stainless steel and other metal substrates for high-temperature end use are usually prepared by sand or grit blasting. Otherwise, the metals are processed typically. The pickling acids may be quite reactive; i.e., HF is typically incorporated into the pickle. Cast iron and heavy-gauge sheet steel are prepared by sand or grit blasting and then typical chemical processing.

Metal preparation can be described as the following processes:

1) Annealing. The metal to be enameled is heated to a temperature sufficient to remove organic surface impurities. Such impurities can act as a source of gases that become entrapped in the enamel during the firing.
2) Grit or sand blasting. This acts as part of the cleaning process where heavier rust, scale, and dirt are removed. More importantly, sand blasting creates a rough surface for the enamel to anchor mechanically.
3) Solvent cleaning. Cleaning removes mill oils and the soaps, oils, and rust preservatives added during forming.
4) Acid pickling. Pickling acts to remove metal oxidation products and scale. It also acts to create a rougher surface and increase enamel adherence.
5) Nickel flash. The nickel deposit follows the acid pickle and subsequent acid rinse. Its function is to promote adherence between ground coat and enamel and reduce the occurrence of defects in the ground coat resulting from noncontrolled metal oxidation.
6) Neutralization. This functions to increase the pH of the metal surface.

The type of surface pretreatment is governed not only by the metal, but also by the processing. The three types of process systems are batchtype immersion, continuous or automatic immersion, and continuous spray. For example, the steps used for sheet steel include the following:

1) Physical and chemical cleaning. Heavily oxidized steel is sand blasted prior to chemical cleaning. Alkaline cleaners, typically baths containing caustic soda at 40 – 60 °C, are used to remove organic soils.
2) Acid pickling. The removal of rust and scale from the metal surface by pickling is carried out in sulfuric acid (6 – 9 wt % at 65 – 75 °C) or in hydrochloric (muriatic) acid (10 – 12 wt % acid at room temperature). Phosphoric acid or an acid salt can also be used. Sulfuric acid is most commonly used because of its speed and lower cost. The sulfuric acid should be pure, and inhibitors must be avoided since they interfere with metal removal and etching, resulting in faulty adherence. After pickling, the metal should be

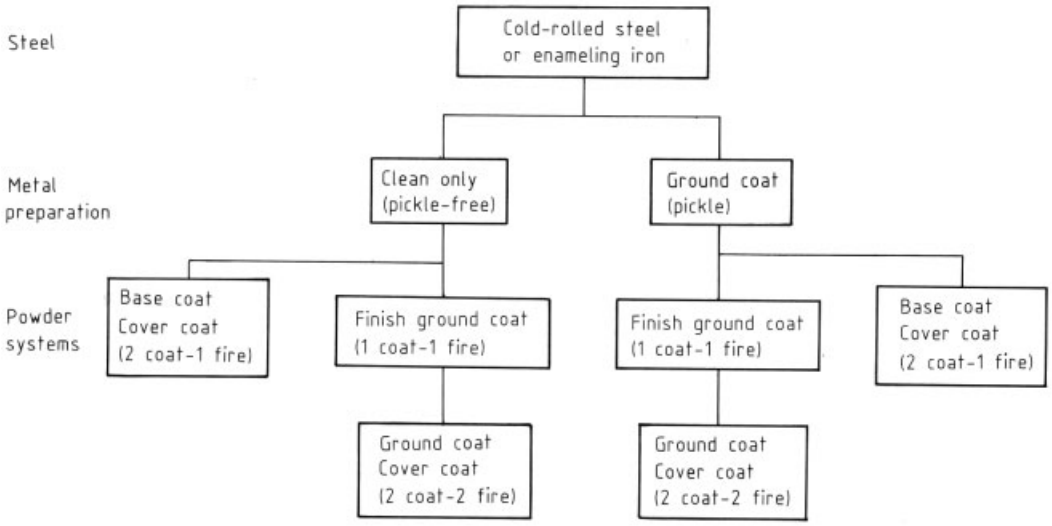

Figure 4. Flow diagram for the enameling of cold-rolled steel or enameling iron

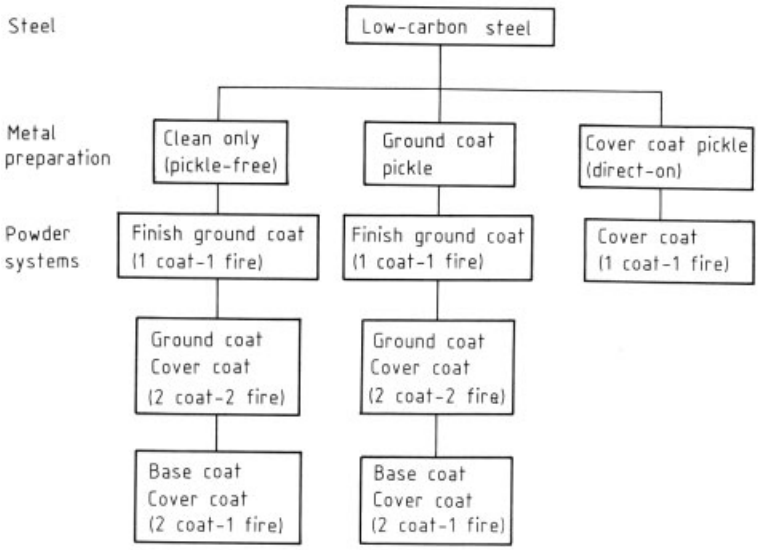

Figure 5. Flow diagram for the enameling of low-carbon steel

rinsed in cold running water, usually acidified to pH 3 – 3.5 to prevent hydrolysis of iron salts.

3) Nickel deposition. Time of treatment, typically 5 – 15 min, should be sufficient to provide the desired coating of nickel. The aqueous nickel solutions are maintained at concentrations of 0.25 – 3 wt % nickel sulfate at 70 – 80 °C. The pH should be between 2.6 and 3.2. The temperature, pH, and nickel concentration are closely monitored. For conventional enameling, the nickel deposited should range from 0.27 to 1.6 g/m^2.
4) Neutralization. The final step is neutralizing acid residuals on the metal surfaces, and also providing rust protection and minimizing the presence of undesirable surface chemicals. Typically, either alkaline soda ash or borax

solutions equivalent to 0.02 – 0.3 wt % Na_2O are used.

Table 1. Chemical oxide composition of ground coats for steel, parts by weight

Component	I	II	III	IV	V
KNa_2O *	17.8	18.5	21.0	18.0	19.4
B_2O_3	16.0	15.2	11.8	14.1	13.3
Al_2O_3	7.7	8.8	8.7	5.4	8.0
SiO_2	51.1	52.1	44.4	44.1	56.6
CaF_2	5.5	3.8	8.7	6.8	
BaO			2.0		
CoO	0.5	0.4	0.4		
Mn_2O_3	0.9	1.2	0.9		
NiO	0.5		0.2		
CuO			0.7		
CaO			1.2		
Sb_2O_3				4.8	1.4
ZrO_2				3.7	6.2

Ground coats I, II, and III are typical blue ground coats, whereas coats IV and V are white ground coats for use with white enamel cover coats.
* KNa_2O appears often in the enameling literature, indicating that the proportions of K_2O and Na_2O in the oxide raw materials do vary. The formulation K,Na_2O also appears in the literature.

Ground Coats. Porcelain enamel ground coats for sheet steel are basically alkali-metal aluminoborosilicate glasses containing adherence-promoting oxides of such metals as cobalt, nickel, and copper. These coatings serve three purposes. The first is to provide a satisfactory bond or adherence to the base metal by complex reactions with the steel during firing. The second is to provide a protective layer or coating that minimizes surface defects caused by the metal or metal preparation, smoothing the surface for the cover coat. The last is to provide a resistant coating, in some cases decorative, thermally resistant, or chemically resistant. While increased adherence is the primary function of the ground coat, the other functions should not be underemphasized. For some steels satisfactory adherence may be obtained without the use of ground coats; however, special surface treatments must be employed if a ground coat is not used.

Ground coats are most typically composed of *ground-glass powders* or *frits.* In most instances the base ground coat may contain a number of frits, all having specific properties that make them useful. For example, one frit may provide sag resistance to the molten glass when the ground coat is heated. A second frit may fuse at a lower temperature, thus sealing the metal surface and promoting glass – metal bonding at an earlier stage. A third may be used as a decorative or colored base to hide defects prior to the final coat.

Ground coats are applied in the form of a slurry to the treated metal by spraying or dipping, and are then allowed to dry. Typical firing temperatures range from 780 to 870 °C. Table 1 shows typical chemical oxide compositions of ground coats for steel.

Cover Coats. Enamel cover coats are designed to provide specific color and appearance characteristics coupled with resistance to atmospheric and liquid corrosion, abrasion resistance, and thermal shock resistance, to name a few properties. Cover coats are applied either as a suspension, followed by drying and firing, or by spreading a dry powder over the hot metal. Cover coats range in thickness from 0.08 to 0.64 mm, depending on application. The firing temperatures for the cover coat are lower than those for the ground coat to prevent excessive flow in the ground coat.

The corrosion properties of the enameled metal are most often those of the cover coat. For instance, titania-based enamels are used for acid resistance, while zirconia-based cover coats are used for alkali resistance. Table 2 shows typical steel cover coats.

Table 2. Chemical oxide composition of cover coats used for cast iron, parts by weight

Component	I	II	III	IV	V
KNa_2O	37.1	21.8	32.0	37.6	32.4
B_2O_3	11.8	13.4	9.1	11.8	14.9
PbO	10.4	13.4	16.0		
Al_2O_3	6.4	4.6	5.4	6.5	5.3
SiO_2	39.0	2.9	24.0	44.2	14.9
CaF_2	6.2	5.6	7.0	6.2	4.9
CaO	3.9	6.7	2.3	4.0	10.5
ZnO		7.5	7.9	5.0	9.2
Sb_2O_3	3.7		4.2	3.7	
ZrO_2		4.7			5.9
BaO		12.0	8.0		
MgO					5.4

3.2.2. Theory of Adhesion

The theory of adherence can be divided into two main branches: the mechanical theory and the chemical or oxide theory [14–16]. Each is characterized by numerous hypotheses that have evolved to explain adhesion. In addition, there

is electrolytic theory and dendritic theory. However, no one theory can adequately describe adhesion for all systems.

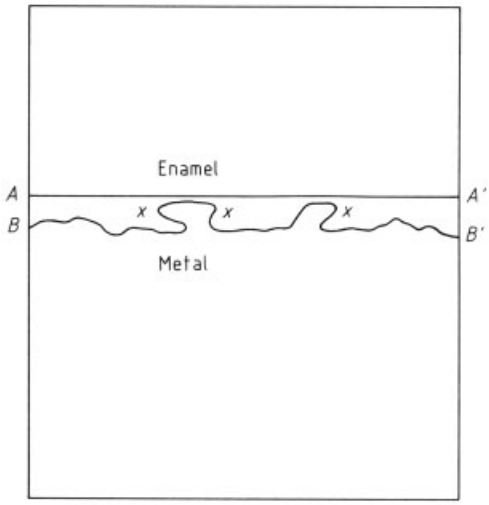

Figure 6. Schematic section of the enamel–metal interface comparing measures of roughness

Mechanical Theory. The mechanical theory states that adherence is due to mechanical gripping of the roughened base metal. The enamel glass is held to the metal by enclosing projections and by filling the depressions occurring on the metal surface.

The application of porcelain enamels to cast iron was the first commercial application of ceramic coatings. Although the surface of the cast iron was inherently rough, further roughening, usually by sand or grit blasting, improved the adherence. This practice suggested the mechanical theory of adherence. Other metals could be coated with ceramics if their surface was roughened. Treatment included combinations of sandblasting, scouring, electrolytic corosion, pickling, etc. This extension to other metals increased acceptance of the mechanical theory.

To better understand the relationship between the roughness of the ceramic – metal interface and adherence, a model contrasting two measures of roughness was proposed:

1) undercuts/distance in centimeters (or anchor points/centimeter)
2) interface ratio

Both are measured on microscopic polished sections of the ceramic – metal interface. Figure 6, a schematic section of the enamel – metal interface, shows the two measures of roughness. The undercuts or anchor points (each marked by an x) are counted and expressed as the number per centimeter, in this case, number of x's/AA'. The interface ratio is calculated by measuring the length of the interface line with a map measure and dividing it by the length of the line parallel to the interface, in this case BB'/AA'. These two measures are then correlated to adherences. Adherence is determined by the standard test of deforming the metal and measuring the amount of metal exposed.

If adherence is due to the gripping action of the rough ceramic – metal interface (mechanical theory), then the best correlation should be between adherence and anchor points per centimeter. If adherence is due to chemical bonding (chemical theory), the interface ratio should give a better correlation between adherence and roughness. For enameled cast iron:

1) Positive correlation is found between the adherence of a porcelain enamel ground coat and the roughness of the interface.
2) In general, adherence correlates better with anchor points/centimeter than with the interface ratio.
3) The method of metal preparation has a marked effect on the relationship between roughness and adherence, pickled iron generally producing better adherence than sandblasted iron for the same degree of roughness.
4) Most of the roughness associated with good adherence develops during firing.
5) Roughness of interface is a necessary but not a sufficient condition for the development of good adherence.
6) Factors other than roughness of interface influence adherence.

These conclusions tend to confirm the mechanical theory of adherence. However, some adherence phenomena cannot be explained by the mechanical theory:

1) Excellent adherence can be obtained on extremely smooth surfaces.
2) A roughened surface exhibiting many anchor points does not ensure adherence significantly better than chemical adherence alone.

3) The measured values for the strength of the bond necessitate assignment of extremely high values for the strength of iron because the total cross section of the anchor points is always a relatively small portion of the total interface area.

Chemical or Oxide Layer Theory. Chemical reactions include dissolution, redox, interdiffusion, and precipitation, which change the nature of the interface and may result in a transition zone between the metal and the ceramic. The bond in metals and ceramics is fundamentally different, and this may prevent a strong physical bond between bare metal and ceramic. A theory developed for glass-to-metal bonds and verified experimentally may be applied more generally to all ceramic-to-metal bonds. The theory states that chemical bonding occurs across an interface when chemical equilibrium relative to the lowest valence oxide of the metal phase is present [5]. The glass and the metal are saturated in the metal oxide at the interface, and therefore the activity of the lowest valence metal oxide is one. Then there is an idealized molecular monolayer of the metal oxide at the interface, which is part of and compatible with both metal and glass. Bond energies should then balance, and the electronic structure should be continuous across the interface. If oxide activity is less than one, complete continuity is not realized. Reaction of metal and glass with chemical interdiffusion is driven by the lack of chemical equilibrium relative to the lowest valence metal oxide. A chemical bond occurs when equilibrium in terms of chemical potentials or activities is reached at the interface. In glass-to-metal bonding this occurs when the glass near the interface is saturated in the metallic species.

Electrolytic Theory. According to the electrolytic theory, metal oxides dissolved in enamels are reduced during firing to the metallic state by the base metal. The adherence-promoting metal in the enamel is believed to form shorted galvanic cells that strongly corrode the base metal. This concept is illustrated in Figure 7. The current flows from the iron through the melt to the cobalt and back to the iron. These local cells are not exhausted during firing because anodic iron and diffusing oxygen are abundant. Iron goes continuously into solution, the surface becomes roughened, and the glass anchors itself into the depressions produced. Final adhesion is basically mechanical. The electrolytic theory has two main faults:

1) Some metals that exhibit a greater potential difference with iron than does cobalt do not cause greater corrosion.
2) The theory is still based on mechanical theory, which was previously discussed and partially discredited.

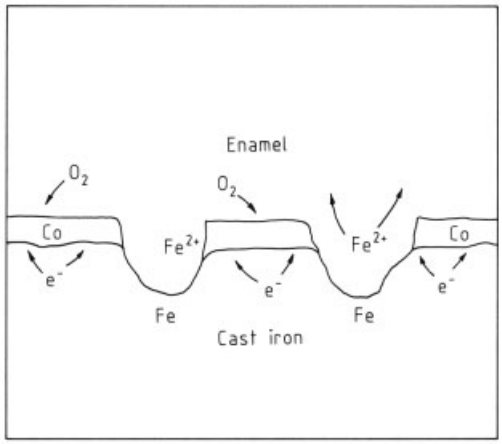

Figure 7. Electrolytic cells between molten enamel and cast iron

Dendritic Theory. Dendrites are the form in which the metal normally crystallizes from the melt, aqueous solutions, or any nonmetallic liquid in which compounds of the metal dissolve. They are branched crystals that occur periodically, giving fernlike, treelike, or toothlike appearance.

The dendrite theory claims that the α-iron dendritic precipitates that occur in the boundary layer between enamel and base metal are responsible for adherence. The dendrites grow from the base metal into the enamel, and after cooling they tend to mechanically attach the enamel to the base metal. This differs from the mechanism of the electrolytic theory, in which the metals and dendrites in the interfacial zone are Co, Ni, or other adherence-promoting metals.

Although a possible adherence mechanism, dendritic theory correlates poorly with adhesion.

4. Ceramic-to-Metal Joining

A ceramic component and a metal component may be joined in various ways, including fasteners, adhesives, and direct joining of ceramic and metal. For durable, integral, hermetic joins capable of extended service, particularly at extreme temperatures, the direct join is used. The successful join usually requires a preparation step to achieve a transition structure, good wetting, and chemical bonding between ceramic and metallic materials. The following are common joining processes:

sintered metal powder (moly – manganese paste)
metal powder – glass frit
active metal
gas – metal eutectic
pressed diffusion joins
vapor-phase metallizing
liquid-phase metallizing
electroforming
graded powder
nonmetallic fusion

These processes are listed in approximate order of frequency of use.

The various joining techniques are often categorized as liquid phase, vapor phase, and solid phase. Ceramic-to-metal joining is discussed comprehensively in [3], [11], [17–19].

4.1. Sintered Metal Powder Process (SMPP)

The sintered metal powder process, also referred to as SMPP or moly – manganese paste process, uses a paste composed of molybdenum and manganese metals and their oxides suspended in an organic vehicle. Compositions containing tungsten, tantalum, iron, rhenium, and titanium with their oxides can also be used. Thus, moly – manganese pastes may contain neither molybdenum nor manganese.

A slurry of a composition such as

60 g of molybdenum powder (0.2 μm)
15 g of manganese powder (0.2 μm)
0.2 g of nitrocellulose
30 mL of butyl acetate

is applied by brushing, spraying, roller coating, transfer tape, silk screening, or dipping to a thickness of ≈ 20 μm. For single-phase ceramics such as high alumina classes without a glassy silicate phase, the usual pastes do not bond well. Compositions containing approximately equal amounts of the metal powders and 13 Al_2O_3–52 MnO – 35 SiO_2 or 41 Al_2O_3–54 CaO – 5 MgO, for example, are used. (The numbers are the weight percents.) After application, the paste is dried in air, often under heat lamps. It is then fired in a hydrogen atmosphere (dew point ≈ 25 °C) at 1200 – 1600 °C for ca. 0.5 h. Manganese or titanium additions lower the firing temperature by 100 – 200 °C.

The process gives rise to a transition coating with a glassy or crystalline ceramic phase that has migrated in the metal – metal oxide powder or that comes from the ceramic added to the paste in the case of single-phase ceramics [1], [2], [7], [20]. The structure of the liquid-phase sintered product consists of a glassy phase with metallic and oxide particles. This structure provides the necessary bridging phases between ceramic and metal. The resulting coating can be directly brazed to metal, but more commonly it is electroplated or electroless plated with a 5 – 8 μm thick nickel layer. Alternatively, copper can be plated, or either metal can be applied as a powder and fired to bond by diffusion in a controlled atmosphere. The resulting nickel surface is suitable for conventional brazing materials and methods.

While SMPP involves several steps, the procedures are relatively straightforward, and the results are reliable, SMPP producing some of the strongest ceramic – metal joins. It is the most common method for producing ceramic-to-metal joins and metallization in industry.

4.2. Metal Powder – Glass Frit Method

For the metallization of hybrid microelectronic circuitry, a finely divided noble metal or alloy (platinum, palladium, gold, silver) mixed in a glass frit and suspended in an organic binder is commonly used. The required pattern is usually screen-printed on an electronic substrate material such as alumina or on a multilayer capacitor material such as barium titanate. These systems permit air firing of the material. The structure of the bridging glassy phase providing a bond bet-

ween the ceramic and a metal-rich free surface is akin to that produced by SMPP. The metal powder – glass frit method may be regarded as a liquid-phase process.

The incorporation of active metal components such as titanium, zirconium, nickel, or copper can lead to formation of a mixed oxide phase by reaction with the ceramic. These phases, like devitrifying phases, aid in bonding and may be fired in a reducing atmosphere [21].

Copper and nickel have been used as a less expensive alternative to the noble metals. The glass frit is composed of oxides that are stable at low oxygen partial pressures (Al_2O_3, B_2O_3, CaO, MgO, SiO_2, ZnO). Firing is done in low nitrogen with a low partial pressure of oxygen (10 ppm) to achieve slight, but controlled, oxidation of the metal so that it can be wet by the glass. The glass must show sufficiently low viscosity for good flow at firing temperatures [21], [22].

4.3. Active Metal Process

The active metal techniques employ an active constituent such as titanium, zirconium, or hafnium in the braze metal to promote wetting and bonding to the ceramic. The active metal may be alloyed with a metal such as nickel or copper to reduce its reactivity and provide a brazing temperature as low as 800 – 1200 °C. The braze materials are used in the form of powders, shims, or wires. Brazing is typically performed under vacuum or in an inert atmosphere with careful temperature control. The hydride of the active metal may be applied separately as a layer on the ceramic and fired, a process resembling SMPP [23].

In the liquid braze the active metal is transported to the ceramic surface where it can react [24]. Wetting is enhanced to the point that caution must be taken to prevent braze metal from flowing into unwanted areas. The joints formed by this method cannot withstand elevated temperatures, even in inert gas atmospheres, because of continued reaction of the bonding phase with the ceramic.

4.4. Gas – Metal Eutectic Process

In the gas – metal eutectic technique, a copper sheet is bonded directly to a ceramic substrate by liquid-phase reaction. A carefully controlled oxygen atmosphere and temperature are required. A eutectic copper – copper oxide melt is formed to wet the ceramic. A reaction layer containing $AlCuO_2$ can form and aids in bonding. Strong bonds have been reported for alumina, beryllia, silica, and spinels [25], [26].

4.5. Pressed Diffusion Joins

Joining by solid-phase diffusion and reaction can also be used to create a ceramic-to-metal bond. Such bonds are typically termed pressure diffusion, crunch, or ram joins. A low-yieldstress metal such as copper is pressed to the ceramic, or the ceramic is forced into a cavity in the metal, so that there is good contact between metal and ceramic and there is an applied or residual force at their interface. The parts should be as clean as possible before contact. The parts are held, often with great external force applied to the interface, at as high a temperature as possible above one-half the melting point of the metal in kelvin. A variety of metals and ceramics can be bonded by the resulting diffusion and chemical reaction [27], [28].

4.6. Vapor-Phase Metallizing

A thin film of metal, which may be base, noble, or reactive metal, is applied by various vapor-phase deposition methods, including ion implantation, evaporation, sputtering, reactive sputtering, plasma methods, and chemical vapor deposition (CVD). Often the initial bonding layers are reactive metals such as Ti, Zr, and Cr to promote bonding, followed by base or noble metals such as Cu, Ni, Au, and Pt. Because platinum shows good ductility and its expansion matches that of alumina, it is often directly deposited. Final layers are usually oxidation-resistant metals. Layers up to several micrometers thick can be built up in this manner, although usually the metallization is quite thin.

These layers may be built up by more conventional solder or braze methods for bonding with a metal part. Brazing is usually done in hydrogen, under vacuum, or in cracked ammonia. Strong hermetic seals can be made on various

polyphase, single-phase, and single-crystal ceramics. Furthermore, glasses and other ceramics not tolerant of brazing can be soldered at lower temperature.

4.7. Liquid-Phase Metallizing

An alternative metallizing method involves thermal reduction of a salt solution of a precious metal (platinum, gold, silver, or palladium). A thin, conductive coating is obtained and may be electroplated or soldered directly. Molten salts have been used in conjunction with reactive metals (Ti, Zr, Hf, or U) to metallize ceramics [3].

4.8. Electroforming

Electroforming can be used to avoid high-temperature brazing after a ceramic part has been metallized by another method. The metallized ceramic is placed in contact with a metal such as copper under slight pressure. The assembly is then placed in an electroplating bath for several days, and the electrodeposition of metal in and around the gap between metal and metallized ceramic creates the bond. Large, complex parts have been joined in this way. This technique has been coupled with pressure diffusion joining to produce an enhanced bond, but in such a case high temperatures are not avoided.

4.9. Graded Powder Process

The graded powder process produces a join by first pressing a powder compact of metal and metal oxide either to the base ceramic or between ceramic and metal. Often this process is similar to SMPP or metal powder – glass frit methods. A gradual change in the powder composition from ceramic-rich to metal-rich is accomplished by producing the preform in layers. This changing composition provided by mechanically mixing powders provides a ceramic-rich layer and a metal-rich layer for the respective materials to be joined, with graded composition and properties in between. The powder is sintered at elevated temperature similar to SMPP. Bond strengths are not especially greater than those of other methods, and the preform is expensive. Only limited commercialization has taken place.

4.10. Nonmetallic Fusion Process

Nonmetallic fusion bonds employ a glass frit to produce a bonding glass between ceramic and metal components. This is a liquid-phase method.

Ceramic glazes and glass frits can be used to join ceramics to each other or to metals if used in layers sufficiently thin to be incorporated into the ceramic and metal on cooling. The incorporation of the glass in the structure may allow use at significantly higher temperatures than the fabrication temperature or softening temperature of the glass suggests.

An oxide composed of constituents such as MnO_2, SiO_2, and Al_2O_3 may be applied to the join interface as a suspension or preform. The assembly is heated in an inert gas atmosphere for up to 15 min at 1200 – 1500 °C and then cooled relatively rapidly, 100 °C/min, so that the glass does not devitrify. The composition of the frit can be adjusted to match the expansion, wetting, and bonding characteristics of ceramic and metal. The bond may be heated to temperatures substantially higher, ca. 800 °C, than SMPP joins. Success has be obtained with high-purity alumina and sapphire. The high temperatures and cooling rates can lead to defects or interfacial failure.

5. Thin Films and Coatings

A coating may be defined as a near-surface region of a material with properties different from those of the bulk. The fabrication of coatings varies widely, depending on application, substrate, coating materials, and economic considerations. Coating techniques may be categorized as follows: (1) those processes that produce a layer of material in the surface region of the bulk material that has properties differing from those of that bulk material and (2) those processes that modify the immediate surface of the bulk material, causing that region to have different properties [29].

Ceramic – metal coatings can be divided into two categories: (1) coatings containing both ceramic and metal phases and (2) ceramic films

and coatings on metals. Ceramic – metal coatings are a unique type of coating in that the properties of the coating are strongly controlled by the microstructure of the metallic phase. These composite films exhibit properties unavailable from either the ceramic or metal phases alone (see Section 6.1).

Ceramic films and coatings are a special class of ceramics in the sense of their special form and the special preparation techniques required. However, their uses are diverse and exploit the wide range of unique and desirable properties that various bulk ceramics possess. In this sense, ceramic films and coatings should be viewed as a large family of materials with quite diverse compositions and properties [30].

5.1. Uses

Table 3 lists uses of ceramic films and coatings and gives typical examples. The table is illustrative rather than exhaustive, but does show the wide range of ceramic films and coatings, both in terms of uses and in terms of compositions.

Table 3. Uses of ceramic films and coatings

Use	Typical ceramic materials
Wear reduction	Al_2O_3, B_4C, Cr_3C_2, CrB_2, $CrSi_2$, Cr_3Si_2, DLC ∗, Mo_2C, $MoSi_2$, SiC, TiB_2, TiC, TiN, WC
Friction reduction	MoS_2, BN, BaF_2–CaF_2
Corrosion reduction	Cr_2O_3, Al_2O_3, Si_3N_4, SiO_2
Thermal protection	Ca_2Si_4, $MgAl_2O_4$, MgO, ZrO_2 (stabilized with Mg or Ca)
Electrical conductivity	In_2O_3–SnO_2
Semiconductors	GaAs, Si
Electrical insulation	SiO_2
Ferroelectricity	$Bi_4Ti_3O_{12}$
Electromechanical	AlN
Selective optical transmission and reflectivity	BaF_2–ZnS, CeO_2, CdS, CuO – Cu_2O, Ge – ZnS, SnO_2
Optical wave guides	SiO_2
Optical processing (electrooptic, etc.)	GaAs, InSb
Sensors	SiO_2, SnO_2, ZrO_2

∗ DLC = diamond-like carbon.

Resistance to wear is generally correlated with hardness. Of special interest is the use of ceramics for wear resistance at elevated temperatures. Table 4 gives the microhardness and oxidation temperatures for some hard coat materials. Although some hard-coating materials have good wear resistance, they generally are not low-friction materials. When low friction is required in addition to wear resistance, softer materials may be used. Calcium fluoride and composites including calcium fluoride are among the most promising solid lubricants for service at ca. 650 °C [31].

Table 4. Bulk properties of some hard coat materials

Material	Microhardness, kg/mm^2	Oxidation temperature, °C ∗
Al_2O_3	2350	
B_4C	4200	1090
Cr_3C_2	2650	1370
Cr_2O_3	1800	
SiC	2900	1650
Si_3N_4	2000	1400
TiC	3200	540
TiN	1950	540
WC	2050	540

∗ Temperature of appreciable detrimental oxidation. Passivating oxide films form at lower temperatures.

Corrosion and oxidation resistance at high temperatures constitute a complex field of chemical reactions and protective mechanisms provided by coatings. Often the protective mechanism involves the production of a dense, adherent, diffusion-resistant oxide film by in situ oxidation, e.g., chromia and alumina on high-temperature superalloys and silica on silicon carbide. This requires that the base alloy or ceramic has a composition that is self-passivating. Much work has been directed toward developing pinhole-free coatings with sufficient adherence to use as corrosion protection mechanisms. Another corrosion-related problem involves water attack, especially the stress-assisted water attack involved in slow crack growth in many ceramics. Silicon nitride films are promising barriers for protection of glass, including glass fibers for optical communication [32].

Thermal protection coatings, consisting of a thin layer of ceramic applied over a metallic bond coat, are used in gas turbine components, such as combustion chambers, and are under development for additional uses. Bond coat oxidation and stress development appear to be limiting aspects of present coatings [33].

Thin films play a central role in electronic and optical devices, especially as the dimensions of these devices are reduced. Coating applications include thin-film resistors and photothermal converts for solar energy. Indeed, modern

electronics would not be possible without thin-film technology [34].

5.2. Deposition Techniques

Table 5 lists coating techniques that have been used for ceramics. Table 5 includes four subcategories of processes: (1) atomic deposition, (2) particulate deposition, (3) bulk coating, and (4) surface modified coatings. Ceramic coating processes could also be divided into those that directly apply the ceramic compound or a precursor and those that react metallic species with gases to form the ceramic coating compound [35].

With the exception of CVD-coated cutting tools, traditional ceramic coatings were limited to those materials that could be applied to a surface as a powder and subsequently fused, such as glazes (→ Ceramics, General Survey, Chap. 5.) and enamels (see Section 3.2), which were either dipped or sprayed. Studies have shown that nontraditional ceramic materials, such as nitrides, carbides, and oxides, could be applied to a variety of materials ranging from polymers to metals to other ceramics. The techniques that are most used include chemical vapor deposition, evaporation, sputtering, plasma spraying, sol–gel processing, and ion implantation [36].

5.2.1. Chemical Vapor Deposition (CVD)

Chemical vapor deposition is the reaction to form the desired compound and the condensation of the compound or compounds from the gas phase onto a substrate where reaction occurs to produce a solid deposit. The compound bearing the deposit material, if not already a gas, is vaporized by either a pressure differential or the action of a carrier gas and is transported to the substrate. Since the vapor condenses on any relatively cool surface that it contacts, all parts of the deposition system must be at least as hot as the vapor source. The reaction portion of the system is generally much hotter than the vapor source but considerably colder than the melting temperature of the deposit.

Generally the deposition process is either thermal decomposition (Fig. 8) or chemical reduction. Both thermal decomposition and chemical reduction most frequently involve organometallic compounds, but halides and other simple inorganic compounds can be used. CVD processes are dependent on both the thermodynamics and kinetics of the reaction. Even though the formation of a material is thermodynamically feasible, its growth rate must not be too slow. When reaction occurs on the substrate, the stages required to obtain a deposit are

1) Formation of the gaseous phase
2) Gas transfer to the substrate
3) Adsorption
4) Decomposition of the adsorbed phase
5) Compound deposition and desorption of the other decomposition products
6) Removal of these products

Table 6 lists ceramics that can be deposited by CVD. The CVD process finds its greatest application in the preparation of coatings that are not easily applied by physical vapor deposition (PVD) techniques.

It has become common to apply thin coatings of carbide and oxide ceramics to cemented carbides and cermets to increase performance. Such coatings are commonly applied by chemical vapor deposition in micrometer thicknesses and can provide abrasion resistance (Al_2O_3, TiC), oxidation resistance (Al_2O_3), friction reduction (TiN), or a thermal or diffusion barrier (Al_2O_3, TiN, TaC). Often many different coatings are used to obtain matched thermal expansions and better adherence properties. In the area of cemented carbide cutting tools, cutting speeds can be increased by three times and tool lifetimes may be two to ten times greater than those of an uncoated tool. In the past decade such coated tools have been introduced and now dominate the marketplace for industrial machining of steel [37].

The advantages of CVD are the following:

1) Various kinds of coatings possible – carbides, nitrides, and oxides
2) Crystal growth controlled by controlling the reactive gas concentrations
3) High deposition rates
4) Possibility to coat complex bulky shapes

Its disadvantages include the following:

1) Rather high temperatures required to produce certain compounds

Table 5. Categorization of ceramic coating techniques

Atomic deposition	Particulate deposition	Bulk coatings	Surface modification
chemical vapor environment	fusion coatings	diffusion	chemical (liquid)
chemical vapor deposition	sol–gel	diffusion bonding	oxidation
reduction	thermal spraying	hot isostatic pressing	chemical (vapor)
decomposition	plasma spraying	wetting processes	thermal
plasma enhanced	low-pressure plasma spraying	dipping	plasma
plasma environment	laser-assisted plasma spraying	enameling	ion implantation
sputter deposition	electric arc spraying	spraying	sputtering
diode			
triode			
reactive			
evaporation			
direct			
activated reactive			
vacuum environment			
vacuum evaporation			
ion implantation			

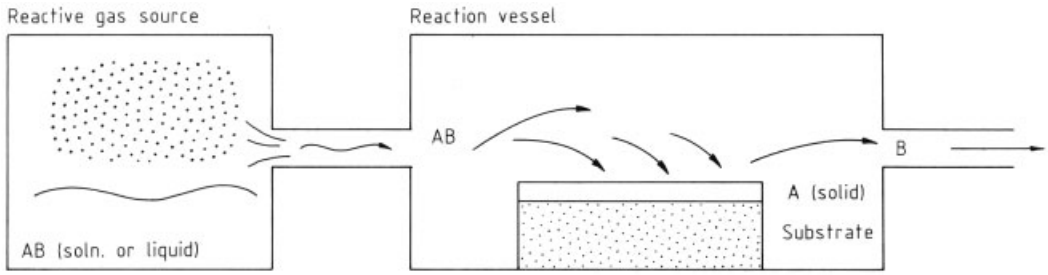

Figure 8. Schematic of CVD reactor

2) Low pressures restrict the types of substrate that can be coated
3) Uniform heating of the substrate at times difficult
4) Reactor vessel design is critical to coating complex shapes uniformly [38].

5.2.2. Evaporation

In the evaporation technique, the coating material is placed in a vacuum chamber opposite the substrate. A relatively high vacuum is used, typically 10^{-6} Torr ($\approx 10^{-4}$ Pa), because the coating material must be evaporated either by heating it to some large fraction of its melting point or by using an electron beam to locally heat it. When it is evaporated, its atoms are ejected in straight lines from the solid and impinge and condense on all exposed surfaces, including the substrate to be coated. Because there is no driving force to accelerate the freed coating atoms, they impinge with ca. 0.2 eV.

Evaporation processes for the deposition of refractory compounds are subdivided into two types:

1) Direct evaporation – the evaporant is the refractory compound itself.
2) Reactive evaporation – a metal or compound in a low valence state is evaporated in the presence of a reactive gas to form a compound deposit.

Direct evaporation of ceramics can occur with or without dissociation of the compound into fragments. However, the observed vapor species show that very few compounds evaporate without dissociation. Examples are MgF_2, B_2O_3, CaF_2, SiO, GeO, and SnO. In the more general cases, ceramic materials are not transformed to the vapor state with the same molecular composition as the solid. Subsequently the fragments must then recombine to reconstitute the compound. Therefore, the stoichiometry of the deposit depends on several factors, which can lead to cation- or anion-deficient structures.

Reactive evaporation broadens the range of ceramics for coatings – the ceramic need not form directly or recombine from a vapor. Rather, a metal or suboxide is evaporated in the presence of a reactive gas to produce a ceramic deposit by reaction.

Figure 9 illustrates the evaporation process with an electron-beam source. The entire system is within a vacuum chamber. If reactive gas is allowed to enter the vacuum chamber, evaporation can be reactive rather than direct. However, the reaction must be both thermodynamically and kinetically feasible.

Table 6. Chemical vapor deposition of ceramics

Materials	Reactive gas	Temperature, °C
Nitrides		
BN	$BCl_3 + NH_3$	1000 – 2000
	thermal decomp. $B_3N_3H_3Cl_3$	1000 – 2000
HfN	$HfCl_x + N_2 + H_2$	950 – 1300
Si_3N_4	$SiH_4 + NH_3$	950 – 1050
	$SiCl_4 + NH_3$	1000 – 1500
TaN	$TaCl_5 + N_2 + H_2$	2100 – 2300
TiN	$TiCl_4 + N_2 + H_2$	650 – 1700
VN	$VCl_2 + N_2 + H_2$	1100 – 1300
ZrN	$ZrCl_4 + N_2 + H_2$	2000 – 2500
Oxides		
Al_2O_3	$AlCl_3 + CO_2 + H_2$	800 – 1300
SiO_2	$SiH_4 + O_2$	300 – 450
	thermal decomp. $Si(OC_2H_5)_4$	800 – 1000
Silicon oxide nitride	$SiH_4 + H_2 + CO_2 + NH_3$	900 – 1000
SnO_2	$SnCl_4 + H_2O$	
TiO_2	$TiCl_4 + O_2$ + hydrocarbon (flame)	
Silicides		
V_3Si	$SiCl_4 + VCl_4 + H_2$	
MoSi	$SiCl_2$ + Mo (substrate)	800 – 1100
Borides		
AlB_2	$AlCl_3 + BCl_3$	≈1000
HfB_x	$HfCl_4 + BX_3$ (X = Br, Cl)	1900 – 2700
SiB_x	$SiCl_4 + BCl_3$	1000 – 1300
TiB_2	$TiCl_4 + BX_3$ (X = Br, Cl)	1000 – 1300
VB_2	$VCl_4 + BX_3$ (X = Br, Cl)	1900 – 2300
ZrB_2	$ZrCl_4 + BBr_3$	1700 – 2500
Carbides		
B_4C	$BCl_3 + CO + H_2$	1200 – 1800
	$B_2H_6 + CH_4$	
	thermal decomp. $(CH_3)_3B$	≈ 550
Cr_7C_3	$CrCl_2 + H_2$	≈1000
Cr_3C_2	$Cr(CO)_5 + H_2$	300 – 650
HfC	$HfCl_4 + H_2 + C_6H_5CH_3$	2100 – 2500
	$HfCl_4 + H_2 + CH_4$	1000 – 1300
Mo_2C	$Mo(CO)_6$	350 – 475
	Mo + hydrocarbon	1200 – 1800
SiC	$SiCl_4 + C_6H_5CH_3$	1500 – 1800
	$MeSiCl_3 + H_2$	≈1000
TiC	$TiC_4 + H_2 + CH_4$	980 – 1400
W_2C	thermal decomp. $W(CO)_6$	300 – 500
	$WF_6 + C_6H_6 + H_2$	400 – 900
VC	$VCl_2 + H_2$	≈1000

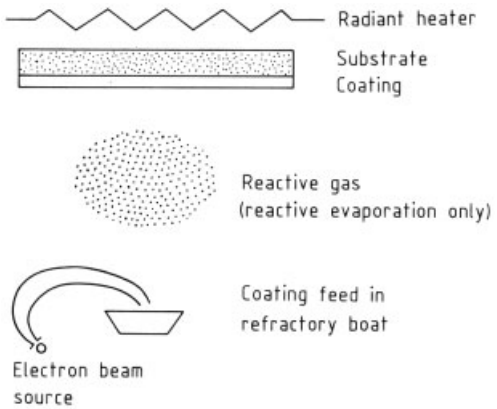

Figure 9. Schematic of evaporation (reactive if reactive gas present)

Compounds that have been synthesized by the reactive evaporation process include α-Al_2O_3, γ-Al_2O_3, Y_2O_3, TiO_2, In_2O_3, In_2O_3–SnO_2 (ITO), SnO_2, BeO, SiO_2, TiC, ZrC, NbC, Ta_2C, VC, W_2C, HfC, VC–TiC, TiC–Ni, Ti_2N, TiN, MoN, $Cu_xMo_6S_8$, and Cu_xS [39].

5.2.3. Sputtering

Sputtering is a coating process that involves the transport of almost any material from a source, or target, to a substrate of almost any other material. The ejection of the source material is accomplished by the bombardment of the surface of the target with gas ions accelerated by high voltage. Particles of atomic dimensions from the target are ejected as a result of momentum transfer between incident ions and the target. The target-ejected particles traverse the vacuum chamber and are subsequently deposited on a substrate as a thin film. Sputtering can be direct or reactive. It can be accomplished with diode, triode, or magnetron sources. Figures 10 and 11 illustrate diode and triode assemblies.

Diode sputtering uses a simple two-element electrode configuration. The expendable target is attached to the cathode and an anode platen supports the substrates. The system can be modified to allow for reactive sputtering. Diode sputtering may be operated in either the d.c. or RF (radiofrequency) excitation modes, with RF being preferred for ceramics because it permits lower pressures and thus generally purer films. In addition, RF sputtering reduces charge

buildup, permitting direct sputtering of ceramics.

In *triode sputtering* a hot electron-emitting filament is added between the anode and cathode and serves to control the plasma. The ion current is significantly increased, and lower gas pressures can be used. These are important advantages for the sputtering rate: production throughput is increased, while potential radiation damage to sensitive electronic devices is minimized.

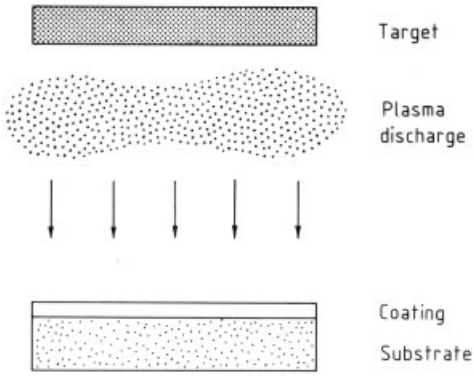

Figure 10. Schematic of a diode sputtering assembly [9]

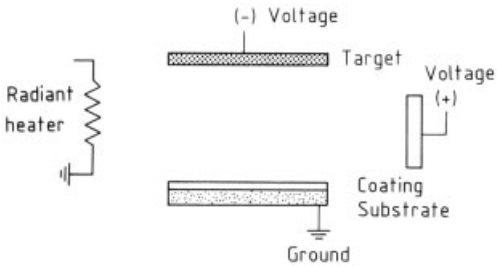

Figure 11. Schematic of a triode sputtering assembly

In *magnetron sputtering* the cathode incorporates a magnetic field aligned perpendicular to the electric excitation field. The magnetic field serves to impart a cycloidal motion to the free electrons being liberated at the cathode surface both by the electric field and by the ionized argon. The great length of the cycloidal path taken by the electrons in response to the presence of the magnetic field greatly enhances the yield of argon ionization for a given current density and voltage. The greater flux of argon ions impinging on the cathode causes rapid disintegration of the cathode and a sputtering rate for a given voltage and current density much higher than that for diode or triode sputtering.

The driving force for sputtering is electrical rather than thermal (as in evaporation). Because of the high kinetic energy transferred, the sputtered atoms impinge on the substrate with ca. 5 eV. Because of this relatively high kinetic energy of impingement, adhesion to the substrate is usually stronger than for evaporation.

Materials having refractive indices ranging between 1.45 and 4.08 have been investigated for optical applications. Indium tin oxide, zinc oxide, and titanium and tantalum oxides have been sputtered for window films, solar cells, and antireflective coatings.

5.2.4. Plasma Spraying

The plasma spraying process is a technique that combines particle melting, quenching, and consolidation into a single process. The arc plasma spraying process involves injection of powder particles into a plasma jet stream created by heating an inert gas in an electric arc. The particles injected into the plasma jet melt rapidly and at the same time are accelerated toward the workpiece surface. Rapid quenching of the molten particles occurs when the droplets impact the surface. Cooling rates are typically 10^5 – 10^6 K/s, and the resulting microstructures are fine grained but may retain a substantial amount of porosity.

The torch itself consists of two concentric electrodes that are cooled by water and electrically isolated from each other. Inert or nonoxidizing gases, or a mixture, are passed through the annular space formed by these electrodes, and a d.c. arc is initiated. With proper operating procedure, the arc is continually sustained, forming a high-temperature plasma that issues from the nozzle front. Figure 12 illustrates the configuration of a typical arc plasma torch system.

Plasma-sprayed coatings offer several advantages over other coating processes:

1) High particle velocities result in higher bond strength coupled with higher coating density.
2) The heat source is more efficient because the plasma is in a high-energy state, providing efficient particle heating.
3) The heating source is inert, minimizing oxidation.

4) High plasma temperatures permit the spraying of materials with high melting points.

Disadvantages include the following:

1) Consistency and compositional uniformity of a coating over a large area at times can vary.
2) Spraying tends to produce porous coatings, especially at high velocities, because of gas entrapment.
3) Reactive materials cannot be sprayed in open air; however, low-pressure plasma spraying, LPPS, can be used in air.
4) The deposits contain oxidation products, together with some porosity because of incomplete melting, wetting, and fusing of deposited particles.

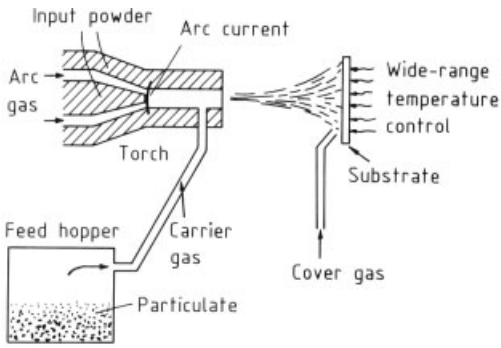

Figure 12. Schematic of an arc plasma spraying assembly

In LPPS the reduction of the amount of air present in the torch region is accomplished by maintaining the plasma system under vacuum with only 30 – 60 Torr (4 – 8 kPa) of inert gas. This greatly reduces the chances of included porosity; therefore, denser deposits and more uniform deposits are possible.

5.2.5. Sol – Gel Processing

Sol – gel coatings are based on a process of applying sols, stable dispersions of hydrous oxides or hydroxides, to a substrate material. Once these liquid dispersions are applied to a surface, they undergo a catalyzed transition to form a gel. The sol – gel coating is subsequently dried to remove excess water and densified. The specific properties of the films, such as pore volume and pore-size distribution, can be controlled during densification. Typical applications of these films include dipping and spin coating [40].

Advantages of sol – gel-derived films include

1) Low densification temperatures
2) High degree of chemical homogeneity
3) Multicomponent films possible

Disadvantages include

1) High shrinkages due to high volume of solvent present
2) Difficulty in making physically stable, thick coatings
3) Coating and substrate densified by heating

5.2.6. Ion Implantation

In ion implantation, ions are formed by collisions with electrons ejected from a heated filament and accelerated in an electric field. In the instances where the element to be ionized is a gaseous species, such as argon, nitrogen, oxygen, or carbon dioxide, which is typical for implanted ceramic coatings, the gas is introduced directly into the ionization chamber, where the high-velocity electrons strip away other electrons, creating a plasma. An electric field then directs the plasma from the chamber to a magnet, which is able to select ions of the desired mass. These ions are then focused in an electric field and accelerated to the final energy to bombard the specimen. The penetration depth depends on the type and energy of the ions and material to be implanted. Figure 13 illustrates an ion implantation system [41].

Sputtered films of carbides, nitrides, and borides exhibited increased adhesion and microhardness when bombarded with heavy inert ions, such as Kr and Xe. The formation of nitride layer on the surface of a Ti – Al – V alloy by bombardment with nitrogen ions has been found to slow the wear rate of the metal by a factor of ten [42].

6. Ceramic – Metal Composites

Ceramic – metal composites may be defined as the heterogeneous combination of one or more ceramic phases with one or more metals or alloys. Systems with very small amounts of

metal or ceramic are usually not considered as composite materials. Also not generally viewed as ceramic – metal composites are materials in which metal particles may be in a continuous ceramic – colloidal and other metal particle colorants in glasses, glazes, and enamels, and thin-film materials such as sputtered and CVD resistors and microwave devices for hybrid microelectronics.

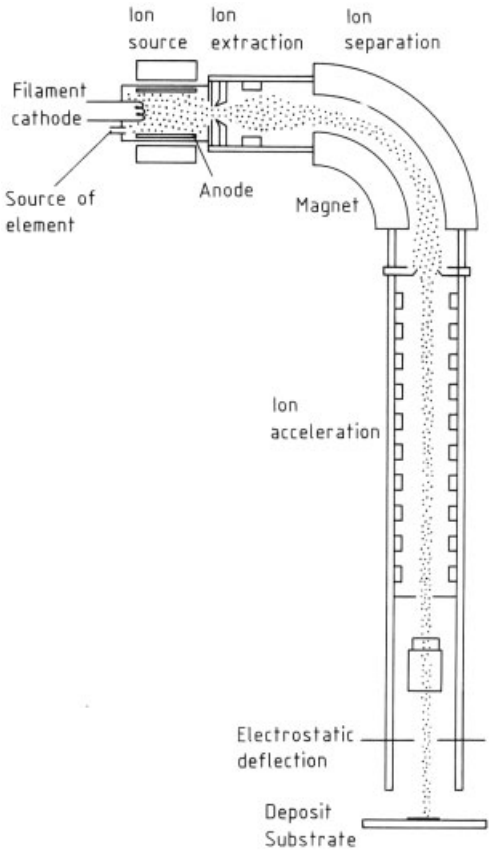

Figure 13. Schematic of an ion implantation system

The distribution of phases in ceramic – metal composites is typically at a fine structural level, commonly microscopic. The metal and ceramic phases are usually mixed physically, often with a binder, formed, and fired at elevated temperatures. For the usual production times and temperature, there is relatively little or only incomplete solution between phases.

A nomenclature problem exists for the ceramic – metal composite materials. Two classes of materials have evolved: cermets and metal matrix composites (MMC). Neither strict definition nor common technological usage provides a clear distinction. Indeed, some use the two terms interchangeably. Usually, however, a cermet is understood to consist of continuous or partially continuous metal phase with ceramic particles having similar size in all dimensions. The metal matrix composites may be distinguished by the high aspect ratio of the ceramic constituent: the metal surrounds long continuous or chopped fibers or elongated single-crystal whiskers of ceramic composition.

6.1. Cermets and Cemented Carbides

Within the cermet subcategory of ceramic – metal composites, there is a nomenclature difficulty. Tungsten carbide – cobalt is termed a cemented carbide, while other carbide – metal materials, such as titanium carbide – nickel, chromium carbide – nickel, as well as oxide ceramic – metal systems, such as alumina – chromium and thoria – tungsten, are termed cermets. This distinction is largely historical. The cermets were developed during and after World War II as materials with better hardness and high-temperature properties than the cemented carbides. However, the microstructures and properties of the cemented carbides and the cermets, particularly the carbide cermets, are closely related. The structure consists of a continuous metal or alloy cement bonding grains (particles) of carbide and/or oxide ceramic.

Cermets and cemented carbides derive their usefulness from the combination of compressive strength, hardness, and thermal resistance provided by the ceramic with ductility, lubricity, and adhesion provided by the metal. The cermets and cemented carbides are used for their strength, hardness, and abrasion resistance, particularly at elevated temperatures, which may arise from the environment or frictional heating. Uses include metal and rock cutting or grinding tools, friction and glide devices, turbine parts, flame nozzles, magnetron tube cathodes, seals, high-temperature containers and pouring spouts, and the ball for ballpoint pens.

Cermets are generally prepared by mixing carbide and/or oxide powder with metal powder in the presence of a binder. The chemistry of

each component is often quite complex. Tungsten carbide typically contains purposeful additions of tantalum carbide and titanium carbide. Molybdenum is added to nickel to promote wetting of titanium carbide. Some cermet is unidirectionally or isostatically pressed and heated to a temperature at which the metal liquifies, coats, and bonds the ceramic. Alternative preparation methods may include slip casting, liquid-metal infiltration, or hot pressing of parts. The factors that dictate the properties of the final part are the following [43–45]:

1) Bonding between carbide grains and matrix
2) Composition and structure of the matrix
3) Shape and size of the carbide or oxide grain
4) Dispersion of the carbide grain in the matrix

A good bond between metal binder and ceramic grain is required to transfer load properly and to provide a homogeneous microstructure for uniform load bearing and deformation. Furthermore, since a wetting angle of near zero degrees is usually desired to have uniform spreading of binder, the degree of bonding is usually quite high. The wetting angle is affected by the composition of the metal binder and ceramic grain, as well as by the firing atmosphere [43].

Phase composition, distribution, and stability have major effects on properties. For example, the tungsten carbide – cobalt system is stable only over a narrow carbon content. At 16 wt % cobalt, WC and Co are the sole phases in the compositional range from 6.00 to 6.12 wt % C – an allowable composition range of 0.12 %. Graphite forms above 6.12 % C and eta phase (η-WC_{1-x}) below 6.00 % (see Carbides, Chap. 2.1.). Both phases embrittle the cemented carbide, causing a loss of one-quarter of the optimum strength for compositions about 0.1 % outside the acceptable range [44], [45].

The microstructure of a cemented carbide or cermet has an important influence on its properties. A uniform dispersion of carbide or oxide ceramic grains in a matrix of metal binder gives the best mechanical properties. Rounded grains, which can result either during powder preparation or from partial dissolution by molten metal during firing, improve the fracture strength. The relative amount of binder phase strongly affects the mechanical properties, as can be seen in Figure 14. Fracture strength increases with the amount of binder because the metal ductility provides fracture toughness. Hardness decreases with increased metal binder because the cermet is less resistant to deformation [43], [44]. The size of the ceramic grain also affects hardness: for a 16 wt % Co cemented carbide, the hardness decreases almost linearly from 89 to 85 R_A as the grain size decreases from 1 to 6 µm. Fine submicrometer grain size increases hardness and resistance to deformation. However, for a fixed amount of binder metal, a decrease in the grain size and the corresponding increase in grain surface area may result in inadequate metal to wet the grain, leading to loss of fracture strength.

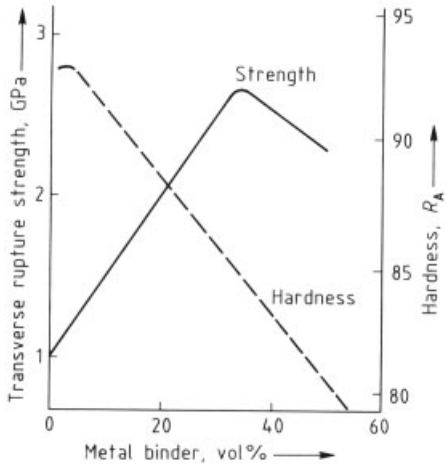

Figure 14. Effect of increased binder on mechanical properties of TiC – Ni – Mo cermet with an average TiC grain size of 1 µm

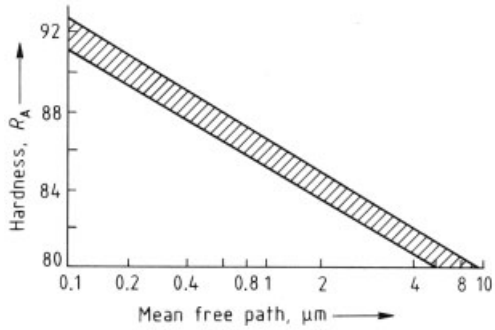

Figure 15. Effect of mean free path on hardness for a cemented carbide

A useful way of examining the properties of cermets and cemented carbides, as well as other composites, is provided by the concept of mean free path, the average distance between neighbor ceramic grains, i.e., the average thickness of

the binder phase. The mean free path depends on the size and shape of the ceramic grain and the amount of metal available. The hardness (as well as a variety of other mechanical properties) depends, within a narrow band, only on this mean free path (see Fig. 15), provided there is adequate metal binder phase to cover the grains uniformly. For the tungsten carbide – cobalt example shown, a mean free path of less than ca. 0.1 µm corresponds to incomplete binder and poorer properties.

6.2. Metal Matrix Composites

The concept of reinforcing a metal with ceramic is not new. In the 1950s the need for jet engine components that could withstand high temperatures for long periods of time began the search for materials that had properties better than those that could be attained by the metal alone.

The advantages that metal matrix composites offer over conventional metals include (1) taking advantage of the anisotropic character of the composite in the efficient design and fabrication of components, (2) manufacturing an engineered material with a specific set of strength and toughness requirements, and (3) increasing the strength, modulus, creep resistance, and oxidation resistance of common engineering alloys such as aluminum, titanium, and nickel. In addition, a major advantage of metal matrix composites arises from the reduction in the strength : weight ratio. For instance, an aluminum alloy reinforced with graphite exhibits the same strength as steel while exhibiting a 45 % increase in the strength : density ratio [46], [47].

In comparison to other composite matrices, the greatest potential advantages of the metal matrix are its resistance to severe environments, toughness, and strength at high temperatures. The temperatures where metal matrix composites exhibit the greatest potential advantages over rival metallic alloys and polymer matrix composites are above 800 °C (Fig. 16). In fiber- or filament-reinforced metals the environmental stability of the matrix at elevated temperatures can be emphasized because the required mechanical strength and stiffness are obtained from the reinforcement. The strength requirements of the matrix are nominal, the matrix only serving to transfer load into the fiber. Up to 540 °C, several potential metal composite systems have been identified. For higher temperatures, at which oxidation presents a major problem, stable reinforcements such as alumina and silicon carbide fibers have considerable potential. Below 320 °C, the potential is less clear, but the advantages of metal matrix composites over polymer matrix composites include improved wear and erosion resistance, greater shear, compressive and transverse strength, increased toughness and hardness, and greater strain prior to failure. The disadvantages include higher density, more difficult fabrication, metal – fiber reactions, and higher cost.

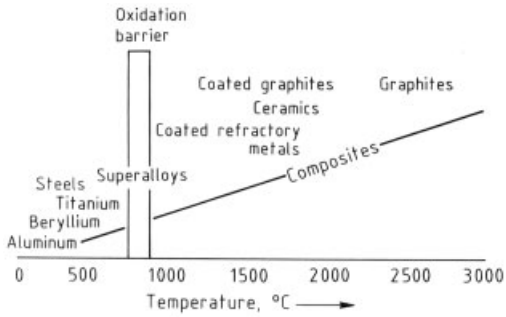

Figure 16. Temperature regimes of usefulness for various materials

Reinforcement Materials. The most appropriate reinforcing material depends on the requirements put upon the metal matrix composite. A wide variety of materials of high elastic modulus, high strength, low density, and high-temperature stability are available. Table 7 lists a number of ceramic materials along with typical mechanical properties; in addition, the properties of a few glasses and metals available as fibers are shown for comparison. In many cases the metal or the fabrication technique results in composites with less than theoretical properties. In these instances such things as process defects and matrix – filler incompatibility affect properties. Matrix – filler compatibility can include the following factors: (1) wetting, (2) chemical bonding, (3) filler surface character, i.e., shape, roughness, and porosity, (4) filler volume fraction and size, i.e., particle size or fiber length and diameter, (5) filler orientation, and (6) relative thermal expansion and elastic modulus match. The last can give rise to residual stresses that can affect the mechanical properties.

Table 7. Filament properties for a variety of reinforcing materials

Material	Tensile strength, (ksi) *	Elastic modulus, GPa	Density, g/cm^3	Strength : density ratio, 10^6
Fibers				
Steel	600	200	7.9	2.1
Tungsten	575	407	19.3	0.83
Molybdenum	320	331	10.2	0.88
S-glass	600	86	2.5	6.52
E-glass	500	72	2.5	5.43
Silica	500	72	2.2	5.43
Boron	450	400	2.6	4.68
Silicon carbide	350	379	3.4	2.84
Alumina	350	483	3.98	2.43
Graphite	285	345	1.66	4.75
Whiskers				
Alumina	2000	483	3.98	13.9
Silicon carbide	1500	483	3.2	13.0

* ksi, a non-SI unit, equals 6.89 MPa.

Fabrication Process. Numerous techniques have been employed for the consolidation of ceramic fibers, whiskers, and particles into metal matrices. These techniques vary widely, from typical powder-metallurgical techniques such as slip casting to techniques such as diffusion bonding and plasma spray bonding. The most widely used consolidation techniques are the following:

diffusion bonding
plasma spray bonding
electroforming
liquid-metal infiltration
high-energy rate forming
explosive welding
hot roll bonding
hot drawing
cold rolling
cold drawing
chemical vapor deposition
powder slip casting
hot extrusion

Each technique employed to consolidate a composite requires specific methods for dealing with the anisotropic and fragile reinforcing fillers. The ceramic fiber may be used in oriented or random configurations as individual fiber, bundles, or weaves. Each process must meet to some reasonable degree each of the following objectives:

avoid filament or fiber breakage
consolidate composite with minimal filler degradation
establish and maintain fiber alignment
be able to vary filler loading
achieve a high-density matrix
establish filler–matrix interfacial bond sufficient to transmit applied load from matrix to filler
allow a degree of matrix alloying
allow for postfabrication heat treatment
provide flexibility in filament spacing
allow for cross and angle lay-ups in some instances
minimize product variability
be amenable to scale-up

Compatibility. Filler–matrix compatibility includes the chemical reactions between matrix and filler during consolidation, secondary processing, and use. Some degree of chemical interaction should take place. In a successful composite, the metal completely wets the filler and chemically bonds to it. However, the exact degree of chemical interaction needed to produce optimum properties is not yet known. If the metal and filler are able to bond chemically during consolidation, there is a likelihood that this chemical interaction continues during use, degrading the filler as a result. Knowledge of the thermodynamics for the combination of metal and filler is an important composite design criterion and a suitable starting point for selecting stable systems.

7. References

1. J. T. Klomp: "Interfacial Reactions between Metals and Oxides during Sealing," *Bull. Am. Ceram. Soc.* **59** (1980) 794 – 800.
2. M. E. Twentyman: "High Temperature Metallizing," *J. Mater. Sci.* **10** (1975) 765 – 766.
3. W. H. Kohl: *Handbook of Materials and Techniques for Vacuum Devices,* Reinhold Publ. Co., New York 1967.
4. J. R. Tinklepaugh, W. B. Crandall (eds.): *Cermets,* Reinhold Publ. Co., New York 1960.
5. J. A. Pask, R. M. Fulrath: "Fundamentals of Glass-to-Metal Bonding: VIII. Nature of Wetting and Adherence," *J. Am. Ceram. Soc.* **45** (1962) 592 – 596.
6. F. Bashforth, S. C. Adams: *An Attempt to Test Theories of Capillarity,* Cambridge University Press, London 1883.

7. A. G. Pincus: "Mechanism of Ceramic-to-Metal Adherence, Adherence of Molybdenum to Alumina Ceramics," *Ceram. Age* **63** (1954) 16 – 20, 30 – 32.
8. J. D. Partridge: *Glass-to-Metal Seals,* Society of Glass Technology, Sheffield, England, 1949.
9. D. Tabor: "Interaction between Surfaces: Adhesion and Friction," in *Surface Physics of Materials,* vol. 11, Academic Press, New York 1975, pp. 475 –529.
10. S. S. Rekhson: "Annealing of Glass-to-Metal and Glass-to-Ceramic Seals," *Glass Technol.* **20** (1979) 27 – 35.
11. H. E. Pattee, R. E. Evans, R. E. Monroe: *Joining Ceramics and Graphite to Other Materials,* WASH SP-5052, NASA, Washington, D.C., 1968.
12. J. J. Schmidt, J. L. Carter: "Using Nitrogen Based Atmosphere for Glass-to-Metal Sealing," *Met. Prog.* **128** (1985) 29 – 34.
13. A. I. Andrews: *Enamels: The Preparation, Application and Properties of Vitreous Enamels,* Twin City Printing, Champaign, Ill., 1945.
14. M. Borom, J. Pask: "Role of Adherence Oxides in the Development of Chemical Bonding at Glass-Metal Interfaces," *J. Am. Ceram. Soc.* **49** (1966) 1 – 6.
15. B. W. King, W. H. Duckworth: "Nature of Adherence of Porcelain Enamels to Metals," *J. Am. Ceram. Soc.* **42** (1959) 504 – 525.
16. L. S. O'Bannon: *"The Adherence of Enamels, Glass, and Ceramic Coatings to Metals – A Review,"* Batelle Memorial Institute, Columbus, Ohio, 1959.
17. G. R. Van Houten: "A Survey of Ceramic-to-Metal Bonding," *Ceram. Bull.* **38** (1959) 301 – 307.
18. G. P. Chu: "Some Aspects on Glass-Ceramics to Metal Sealing," *Int. Cong. Glass, 10th,* Ceramic Soc. Japan, 1967, pp. 94 – 102.
19. C. I. Helgesson: *Ceramic-to-Metal Bonding,* Boston Technical Publ., Cambridge 1968.
20. S. S. Cole, G. Sommer: "Glass Migration Mechanism of Ceramic to Metal Seal Adherence," *J. Am. Ceram. Soc.* **44** (1961) 265 – 271.
21. R. G. Loasby, N. Davey, H. Barlow: "Enhanced Property Thick Film Pastes," *Solid State Technol.* **15** (1972) no. 5, 46 – 72.
22. M. Monneraye: "Les Encres Sengraphiables en Microélectronique Hybride – les Matériaux et leur Comportement," *Acta Electron.* **21** (1978) 263 – 281.
23. C. S. Pearsall: "New Brazing Methods for Joining Non-Metallic Material to Metals," *Mater. Methods* **30** (1949) 61 – 62.
24. J. E. McDonald, A. Eberhart: "Adhesion in Metal Oxide Systems," *Trans. Metall. Soc. AIME* **233** (1964) 512 – 517.
25. J. F. Burgess, C. A. Neugebauer, G. Flanagan: "The Direct Bonding of Metals to Ceramics by the Gas–Metal Eutectic Method," *J. Electrochem. Soc. Solid-State Science and Technol.* May, 1975.
26. J. T. Klomp, P. J. Vrugt: "Interfaces between Metals and Ceramics in Surfaces and Interfaces," *Materials Science Research* **14** (1980) 97 – 105.
27. W. Dawihl, E. Klinger: "Mechanical and Thermal Properties of Welded Joints between Alumina and Metals," *Ber. Dtsch. Keram. Ges.* **46** (1969) 12 – 18.
28. H. J. de Bruin, A. F. Moodie, C. E. Warble: "Reaction Welding of Ceramics Using Transition Metal Intermediates," *J. Aust. Ceram. Soc.* **7** (1971) no. 2, 57 – 58.
29. J. B. Wachtmann, W. H. Rhodes: "Ceramics – Advanced Forms Find Essential Uses as an Enabling Technology," *Science (Washington, D.C. 1979)* **6** (1985) May, 6 – 18.
30. J. B. Wachtman Jr., R. A. Haber: "Ceramic Films and Coatings," *Proc. Chem. Eng. Prog.* **82** (1986) no. 1, 39 – 46.
31. W. A. Brainard: "The Friction and Wear Properties of Sputtered Hard Refractory Compounds," NASA Technical Memorandum TM-78 895, 1978.
32. E. Lang (ed.): *Coatings for High Temperature Applications,* Applied Science Publishers, New York 1983.
33. I. Kvernes, P. Fartum: "Use of Corrosion Resistant Plasma Sprayed Coatings in Diesel Engines," *Thin Solid Films* **2** (1978) 119 – 128.
34. W. A. Pliskin: "Comparison of Properties of Dielectric Films Deposited by Various Methods," *J. Vac. Sci. Technol.* **14** (1977) 1064 – 1081.
35. R. F. Bunshah: *Deposition Technologies of Films and Coatings,* Noyes Press, Park Ridge, N.J., 1982.
36. D. M. Mattox: "Commercial Applications of Overlay Coating Techniques," *Thin Solid Films* **84** (1981) 361 – 365.
37. B. M. Kramer: "Requirements for Wear-Resistant Coatings," *Thin Solid Films* **2** (1980) 117 – 128.

38. W. A. Bryant: "Review: The Fundamentals of Chemical Vapor Deposition," *J. Mater. Sci.* **12** (1977) 1285 – 1306.
39. R. F. Bunshah, A. C. Raghuram: "Activated Reactive Evaporation Process for High Rate Deposition of Compounds," *J. Vac. Sci. Technol* **9** (1972) 1385 – 1389.
40. C. J. Brinker, S. P. Mukherjee: "Comparison of Sol-Gel Derived Thin Films with Monoliths in a Multicomponent Silicate Glass System," *Thin Solid Films* **77** (1981) 141 – 148.
41. T. S. Picraux, P. S. Peercey: "Ion Implantation of Surfaces," *Sci. Am.* **252** (1985) no. 3, 102 – 110.
42. B. H. Kear, J. W. Mayer, J. M. Poate, P. R. Strutt: "Surface Treatments Using Laser, Electron and Ion Beam Processing Methods" *Metall. Treatises 1981,* 321 – 342.
43. J. R. Tinkelpaugh, W. B. Crandall (eds.): *Cermets,* Reinhold Publ. Co., New York 1960.
44. J. Gurland, P. Bardzil: "Relation of Strength, Composition, and Grain Size of Sintered WC–Co Alloys," *J. Metals* **7 (Trans. Metall. Soc. AIME 203)** (1955) 311 – 315.
45. J. Gurland, J. T. Norton: "Role of the Binder Phase in Cemented Tungsten Carbide-Cobalt Alloy," *J. Met.* **4** (1952) 1051 – 1056.
46. C. T. Lynch, J. P. Kershaw: *Metal Matrix Composites,* CRC Press, Cleveland, Ohio, 1969.
47. R. W. Rice: "Ceramic Matrix Composites," *Proc. Am. Ceram. Soc.* 1980, 661 – 689.

Ceramics, Electronic

JOHN B. BLUM, Norton Company, High Performance Ceramics Division, Northboro, Massachusetts 01532, United States

1. **Introduction** 1
2. **Linear Dielectrics** 1
3. **Nonlinear Dielectrics** 4
4. **Semiconducting Ceramics** 12
5. **Sensors** 13
6. **References** 14

1. Introduction

The use of ceramics for electronic applications is widespread and expanding rapidly. Ceramic materials are used for electrical circuit elements, e.g., capacitors, as substrates for circuits, as sensors and transducers, and in a wide variety of magnetic applications. Table 1 classifies electronic ceramics.

A brief history of the development of electronic ceramics prior to 1945 was given by THURNAUER [1]. In 1964 KOENIG presented an overview [2]; and in 1969 HENRY wrote an introductory book in the field of electronic ceramics [3].

The field of electronic ceramics is very broad, as can be seen from Table 1. In this article, important representative materials, technologies, uses, and background are discussed. Substrates are discussed under the linear dielectrics, while some of the semiconducting ceramics are discussed as sensor materials. Magnetic ceramics are not discussed (→ Magnetic Materials).

2. Linear Dielectrics

Linear dielectrics are materials in which the variation of electric displacement D is linear with respect to electric field E. An important class of these materials is used as electrical insulators. In such applications it is important to isolate portions of an electric circuit from one another.

Al_2O_3. Aluminum oxide [*1344-28-1*], or alumina, is a versatile ceramic material (→ Aluminum Oxide). WEFERS and BELL have reviewed the properties of the oxides and hydroxides of aluminum [4]. In this section only aspects relevant to the electronic use of alumina are discussed.

An important use of alumina is in packaging microelectronic circuitry [5]. This use requires a material with a high electrical resistivity, low dielectric constant, good mechanical strength, good thermal conductivity, and a thermal expansion that matches that of the integrated circuit material, which is usually silicon. Alumina is a good compromise.

The material used is polycrystalline corundum with purity generally ranging from 92 to 99 wt % Al_2O_3. The other 1 – 8 wt % is predominantly MgO and SiO_2. These materials are added to affect the firing behavior. As would be expected, different compositions have different properties, as shown in Table 2. The properties also depend on the manufacturer (Table 2), because the various suppliers use different additives in the same nominal alumina composition (e.g., 96 %). The alumina is processed with careful considerations to purity and is generally formed into substrates by the tape casting, or doctor blading, process (see petit section in page 6) [6].

In *multilayer alumina substrates* [7], thinner alumina layers are used than those used as single-layer substrates. The layers are laminated together. In between the individual ceramic layers are metallic lines, as well as holes, or vias, filled with metal. These metal lines and vias act as a three-dimensional conducting path interconnecting individual parts of a circuit. As of 1985, IBM has the most advanced such design in their Thermal Conduction Module (TCM) [7], [8]. In this module is a 90×90×15 mm ceramic substrate made up of 33 layers of 92 % alumina. The substrate supports and, through the three-dimensional conducting network, interconnects,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_079

Table 1. Classification of electronic ceramics*

	Dielectric	Magnetic	Semiconductor	Substrates/composites	Optical
Types	linear/nonlinear low/high loss para/ferro/antiferroelectrics piezoelectrics	ferri/ferro/antiferro-magnetics magnetostrictives	luminescents thermoelectrics thermoionics photovoltics conductive ceramics	IC packages films/coatings laminates fiber composites bonded/impregnated composites	lasers windows electro-optical materials fiber optics nonlinear optical materials
Pertinent properties	density resistivity dielectric constant dielectric losses dissipation factor electric strength hysteresis loop properties frequency dependence thermal expansion temperature coefficients Curie temperature piezoelectric constants and coupling	density resistivity permeability coercive force losses residual flux density saturation magnetization resonant frequency hysteresis loop properties frequency dependence thermal expansion coupling coefficient mechanical strength temperature coefficients thermal conductivity Curie point	density resistivity carrier (n, p) properties conductivity resistance vs. temperature Hall coefficient Seebeck coefficient thermal conductivity photoconductivity photovoltage/current electroluminescence thermoelectricity mechanical strength	density resistivity dielectric properties hardness mechanical strength elastic moduli thermal expansion thermal conductivity hermeticity green properties compressibility screenability platability cosinterability single-crystal properties vacuum deposition properties	density optical transmission/transparency dispersion refractive index ferroelectricity bond levels/transitions mechanical strength melting point Curie temperature dielectric constant piezoelectric modulus fabrication process crystal structure single-crystal properties polarization

Table 1. (Continued)

	Dielectric	Magnetic	Semiconductor	Substrates/composites	Optical
Uses	insulators capacitors transducers nonlinear elements attenuators	magnets cores transducers read/write heads microwave devices coils yokes flyback transformers	phosphors sensors infrared detectors thermoelectric devices resistors nonlinear thermistors varistors magnetoresistors	electronic packages hermetic seals electrodes thick-film resistors printed circuit materials dielectric films bubble memory devices IC passivation multilayer ceramics	holographic memories optical switches lasers radomes Kerr cells optical doublers windows fiber optics lamp envelopes
	Dielectric	Magnetic	Semiconductor	Substrates/composites	Optical
			lumistors cathodes oxide batteries		
Materials	steatites titanates glasses bonded micas aluminas	ferrites – hard/soft garnets magnetoplumbites	elements oxides chalcogenides carbides silicides borides β-aluminas phosphides arsenides	oxides glass – ceramics glass – metals porcelain – metal oxide films garnets cermets	oxides glasses halides gallium arsenide zinc selenide silica gadolinium molybdate

∗ Prepared by Wayne Young (IBM).

Table 2. Properties of commercial alumina substrates

Product no.	Alumina content, wt %	Volume resistivity, Ω cm	Dielectric constant at 1 MHz	Linear expansion [d], 25 – 800 °C, 10^{-6} K^{-1}	Thermal conductivity, $W m^{-1} K^{-1}$	Bending strength, MPa
ADO-90 [a]	92	$> 10^{14}$	22	8.1	12.6	365
HA-92 [b]	92	$> 10^{14}$	8.5	7.5	16.8	314
AD-94 [a]	94	$> 10^{14}$	8.9	7.6	18.0	352
AD-96 [a]	96	$> 10^{14}$	9.0	8.0	24.7	358
HA-96 [b]	96	$> 10^{15}$	9.3	7.7	20.1	324
AT-396 [c]	96	$> 10^{14}$	9.6	7.7	33.6	314
HA-995 [b]	99.5	$> 10^{15}$	9.7	8.1	29.4	392
AT-3995 [c]	99.5	$> 10^{14}$	10.6	7.6	37.8	314
AT-3997 [c]	99.7	$> 10^{14}$	10.5	7.8	37.8	392

[a] Product of Coors Ceramics, United States.
[b] Product of NTK Technical Ceramics, Japan.
[c] Product of Narumi Technical Ceramics, Japan.
[d] For comparison, the linear expansion of silicon is $\approx 4\times10^{-6}$ K^{-1}.

supplies power to, and transmits data to and from up to 118 integrated circuit chips mounted on the substrate surface.

New Substrate Materials. There are several properties desirable in a substrate for integrated circuitry, and alumina is a good compromise. However, for the increasingly demanding requirements of new integrated circuit technology, materials with properties better than those of alumina are needed. For very high speed integrated circuits (VHSIC), a low-dielectric-constant material is needed to minimize delays in signal propagation. For very large scale integrated circuits (VLSI), the density of circuit elements is so great that the substrate material must have a high thermal conductivity to dissipate the heat. For circuits based on silicon [*7440-21-3*], a substrate with a close matchin thermal expansion to that of Si (4×10^{-6} K^{-1}) is preferred. No single material has yet been found that incorporates all of these properties, as can be seen in Table 3, which lists some newer materials being evaluated for use as microelectronic substrates.

3. Nonlinear Dielectrics

Nonlinear dielectrics are materials in which the variation of electric displacement D is a nonlinear function of the electric field E. This class of materials includes ferroelectrics and piezoelectrics. Ceramics displaying these effects are useful in capacitors and transducers. There are books that deal with ferroelectric ceramics [13] and piezoelectric ceramics [14].

Piezoelectricity is defined as electric polarization produced by mechanical strain, the polarization being proportional to the strain and changing sign with it [15]. Ferroelectricity is the spontaneous alignment of electric dipoles by their mutual interaction. The conspicuous property of ferroelectrics is the reversibility of their permanent polarization by an electric field [16]. For a polycrystalline body to exhibit piezoelectricity, the material must also be a ferroelectric. Thus, for ceramics ferroelectricity and piezoelectricity are interrelated.

Ferroelectricity. Ferroelectricity is the dielectric analog of ferromagnetism. The term ferroelectricity is derived from this analogy, but does not imply that iron is important in the phenomenon. Ferroelectricity in a material is a consequence of permanent electric dipoles that result from its crystal structure. These dipoles are aligned because of their mutual interaction. In a ferroelectric material, the dipoles are aligned in parallel. Within a single crystal, all of the dipoles do not point in the same direction; rather, there are small areas within each single crystalline region in which all of these permanent dipoles are aligned in parallel for a net nonzero electric polarization. These areas are called *domains.* In a related class of materials, *antiferroelectrics,* the permanent electric dipoles are aligned antiparallel, which cancels out the individual dipole moments for a net polarization of zero.

Table 3. Properties of newer electronic substrate materials

Material	Electrical resistivity, Ω cm	Dielectric constant	Thermal conductivity, $W\,m^{-1}\,K^{-1}$	Linear expansion, $10^{-6}\,K^{-1}$
SiC [a]	$> 10^{13}$	42	267	
AlN [b]	$> 10^{13}$	8.8	80	4.5
BeO [c]	10^{17}	6.7	260	6.5
Glass – ceramic [d]	$> 10^{14}$	7.5	4.2	4.2

[a] Hitachiceram, product of Hitachi Corp., Japan [9].
[b] Toshiba Corp., Japan [10].
[c] 99.5 wt % BeO, product of National Beryllia Corp., United States [11].
[d] Lead borosilicate – alumina, NEC Corp., Japan [12].

If the crystal structure of a material does not yield permanent dipoles, the material does not exhibit ferroelectricity. Therefore, as any temperature or composition variations that lead to phase changes occur in ferroelectric materials, ferroelectric behavior is observed or not, depending on the crystallographic phase present. The temperature below which ferroelectricity is observed is called the *Curie point* of the material. Above the Curie point, the material is nonferroelectric or paraelectric.

The domains in a ferroelectric possess nonzero polarization; however, in the as-formed material the domains are randomly oriented, which leads to an overall polarization of zero in the sample, as seen in Figure 1. The domain structure (i.e., number, size, and direction) in a given sample is the result of several competing factors and, thus, depends on the specific conditions for the sample. The dipole moments in the various domains can be realigned by an externally applied electric field, yielding a nonzero polarization in the sample. The process by which domains change orientation is nucleation and growth; i.e., domains of the new orientation nucleate within the domain and grow preferentially because of the external electric field. The resultant polarization depends on electrical history: if the electrical field is increased and then decreased back to its original value, the final polarization differs. In Figure 2 this *hysteresis* is shown. The hysteresis loop is a consequence of the domains that make up the ferroelectric.

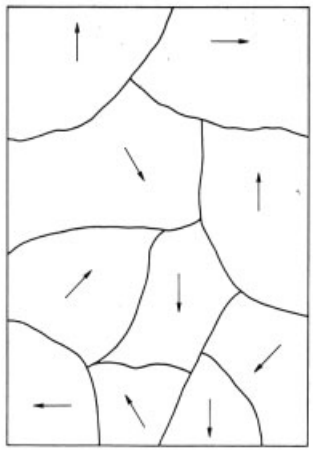

Figure 1. Domain pattern in an unpolarized ferroelectric (schematic), the arrows representing the polarization within each domain

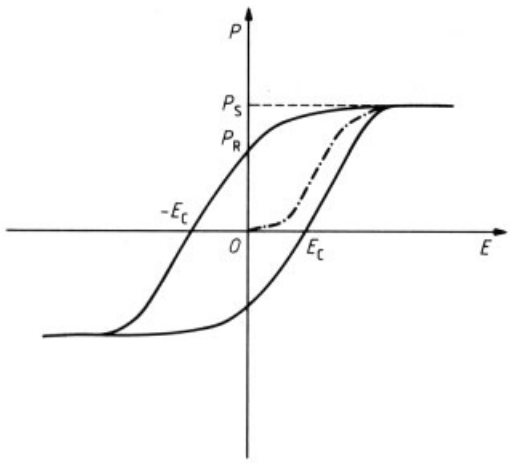

Figure 2. Hysteresis loop for a ferroelectric (schematic)

Consider the following experiment on a ferroelectric material:

1) Increase the electric field to the value E_1
2) Decrease the field past zero to the same value but of the opposite direction, $-E_1$
3) Cyclically, repeat steps 1 and 2

As shown in Figure 1, for the starting material the random orientation of the domains results in a net polarization of zero. When not initially subjected to an electric field, the material has no polarization, represented at the origin O of Figure 2. As the electric field is increased, the resulting polarization for low fields is reversible and nearly linear with E. At high fields the domains begin to orient under the influence of the applied electrical field, so the polarization increases more rapidly. At some high value of E, all of the domains that could reorient with the electrical field have done so: no further increase in polarization can occur. This is known as polarization saturation, and extrapolation of this portion of the curve back to the ordinate gives the saturation polarization P_S. The ferroelectric is then said to be *polarized.*

Consider what happens when the field is decreased to 0. The inability of oriented domains to return to their original random ($P = 0$) positions, because of the irreversibility of the nucleation and growth process, results in a nonzero polarization, the remnant polarization P_R. The domains must be forced into a configuration that again results in a net polarization of zero. The magnitude of the field needed to obtain $P = 0$ is called the coercive field E_C. Further decrease in $E < -E_C$ causes domain reorientation, but in the direction opposite to that discussed above. The resulting polarization behavior is symmetrical to that discussed above.

Once the ferroelectric has been polarized, only the outer portion of Figure 2, the hysteresis loop, is followed. The curve in Figure 2 from the origin to saturation is followed only by the unpolarized material. However, a sample can be depolarized by heating it above its Curie point. When the sample is cooled, the domains in the sample again orient randomly.

Piezoelectricity. Piezoelectric materials convert mechanical energy to electrical energy, and vice versa. The conversion of an applied mechanical stress (force) to an electrical polarization is called the direct piezoelectric effect. Conversion of an applied electrical voltage to a mechanical strain (deformation) is called the converse piezoelectric effect. Any material that displays one effect displays the other. An important parameter of a piezoelectric is its electromechanical coupling factor k, a measure of its conversion efficiency.

As with ferroelectricity, piezoelectricity is a consequence of the crystal structure. The piezoelectric effect requires a noncentrosymmetric crystal structure. In materials possessing these structures, the application of a stress creates electrical dipoles. In different grains of a polycrystalline piezoelectric material, these dipoles point in different directions, thus canceling out. For this reason, until the 1940s polycrystalline piezoelectrics were thought to be impossibilities. However, if a ferroelectric material is subjected to a high electric field, aligning the domains, a polycrystalline piezoelectric can be obtained. This process is called *poling.*

The first commercial piezoelectric ceramic was $BaTiO_3$. Although it is a versatile piezoelectric and many uses were found for it, an important advance was the discovery of the piezoelectric properties of the $PbTiO_3$–$PbZrO_3$ solid-solution series. This solid-solution series, often designated PZT for lead zirconate titanate, has better piezoelectric properties than $BaTiO_3$ and can be used at a higher temperature, 300 °C vs. 120 °C, than $BaTiO_3$ (see page 9). However, $BaTiO_3$ is still a useful ferroelectric material in capacitors.

$BaTiO_3$. In the 1930s and 1940s, work at the Titanium Alloy Manufacturing Co. (United States) led to the discovery of high dielectric constant titanate compositions, among them barium titanate [*12047-27-7*], $BaTiO_3$ [17]. Independently, researchers in Russia and elsewhere had been conducting similar research [18]. In a series of articles, McQuarrie has given an overview of the work and theory of ferroelectric $BaTiO_3$ up to 1955 [19].

In an early report on the dielectric properties of $BaTiO_3$, peaks in the dielectric constant were observed at ca. − 70, + 10, and 120 °C [20]. These correlate to crystallographic phase transformations. $BaTiO_3$ has a Curie temperature of ca. 125 °C. Above this temperature, the crystal structure is cubic (perovskite), while at room temperature there is ca. 1 % distortion of one

of the cubic axes, yielding a tetragonal structure. At room temperature the lattice constants of $BaTiO_3$ are $a = 0.39860$ nm, $c = 0.40263$ nm, for a *c/a* ratio of 1.0101 [21]. Below 0 °C the symmetry is orthorhombic; and below − 80 °C the symmetry is rhombohedral [22].

$BaTiO_3$ is usually prepared from chemically treated barium and titanium precursors. In an early report, the reaction of $BaCO_3$ and TiO_2 in various atmospheres was studied [23]. More recently, the process patented by PECHINI [24] was used for the study of controlled stoichiometry materials [25]. In this procedure tetraisopropyl orthotitanate and barium carbonate are reacted. Barium titanium oxalate [26] and other barium and titanium compounds have also been used to prepare high-purity $BaTiO_3$. Small, uniform powders have been fabricated in an effort to produce more reactive powders [27–29].

There is a large increase in the dielectric constant of $BaTiO_3$ as the temperature approaches the Curie point. To moderate this effect and to alter the dielectric properties, *additives* are usually incorporated into commercial $BaTiO_3$ formulations. If the amount of SiO_2 [*7631-86-9*] is increased from 0.5 to2 wt %, there is a progressive lowering of the room-temperature dielectricconstant and a suppression of the maximum dielectric constant at the Curie point. Al_2O_3 additions have the same effects but to a lesser degree; i.e., a 0.5 % SiO_2 addition has the same effect as 2 % Al_2O_3. Variations in the Ba : Ti ratio only slightly affect the dielectric constant and the Curie temperature, but dramatically affect the dielectric loss (dissipation factor). Additions of $PbTiO_3$ increase the Curie point, while additions of $CaSnO_3$ lower it [30]. In the $CaTiO_3$–$BaTiO_3$ series, increasing the percentage of $CaTiO_3$ does not significantly affect the Curie point, but it does lower the dielectric constant at room temperature. Addition of $MgZrO_3$ broadens the sharp temperature dependence near the Curie point [31].

Additions of more than 0.25 % La_2O_3 decrease both the dielectric constant and the Curie temperature of coarse-grained materials [32]. Additions of up to 0.5 % CeO_2 increase the dielectric constant, but further additions lower it [33]. Many other additives also affect the dielectric properties of $BaTiO_3$ formulations.

The crystallite, or grain, size in polycrystalline $BaTiO_3$ also affects the dielectric properties. The dielectric constant increases to a maximum as grain size decreases to ca. 1 µm [34], [35], phenomenologically explained as being the effect of anomalous variations of permittivity on internal stresses in fine-grained material due to the absence of twinning [36].

In addition to affecting the dielectric properties, additives are also incorporated into $BaTiO_3$ formulations to affect the densification, or sintering behavior. The firing temperature and atmosphere [37], as well as microstructure, affect the dielectric properties of $BaTiO_3$ formulations. To obtain a given set of properties, a particular firing condition or microstructure may be necessary, therefore requiring sintering additives.

Additions of $PbTiO_3$ and of CdO, ZnO, and CuO in amounts above their solid-solubility limits at 1200 °C or additions of more than 3 % of NiO lead to the formation of a reactive liquid phase, which leads to rapid densification. Additions of Bi_2O_5, MgO, $CaTiO_3$, and NiO above their solubility limits at 1200 °C lead to precipitates at the grain boundaries that also enhance densification. Additions of Cr_2O_3 and SiO_2 do not enhance densification [38].

A $0.74\,Bi_2O_3$–$0.26\,B_2O_3$ glass is another liquid-phase former that enhances densification [39]. Lithium fluoride, LiF, is also an effective liquid-phase sintering aid [40–42], as is $BaLiF_3$ [43].

To increase the capacitance in a given volume or to decrease the volume for a given capacitance, a multilayer capacitor design is often used [44]. This design takes advantage of the increase in capacitance with decreased dielectric thickness and parallel connection of individual capacitor layers. In this design, thin layers of $BaTiO_3$-based material are used.

Tape Casting Process. The tape casting process is used to achieve thin, stress-free sheets of dielectric. This process is also known as doctor blading, sheet casting, and knife coating [45]. The tape casting process reliably and efficiently can produce layers 25 – 1250 µm (1 – 50 mils) thick [46]. Tape casting has its origins in the paint industry [47] and was first applied to ceramics by HOWATT [48], [49] during World War II to form thin titanate sheets to replace mica in capacitors. Since that time, it has been used extensively for a wide variety of ceramic materials [6], [50–55].

A slurry that contains the ceramic powder, solvents, and organic compounds is formulated. The organic compounds serve several purposes. To adjust the viscosity of the slurry prior to casting, organics are used to disperse the ceramic particles and to plasticize the mix. Organics are also used to bind the ceramic particles together after casting to provide suitable strength of the cast tape for handling. Table 4 lists one representative formulation [56].

Table 4. Batch formulation for tape casting slurry [56]

Ingredient	Function	Parts, by weight
$BaTiO_3$ [a]	ceramic	100
Methyl ethyl ketone/ethanol (50/50 mixture)	solvent	20
Menhaden fish oil [b]	dispersant	1
Santicizer 160 [c] (butyl benzyl phthalate)	plasticizer	4
Carbowax 400 (poly(ethylene glycol)	plasticizer	4
Cyclohexanone	homogenizer	0.7
Acryloid B-7 [d] (methyl ethyl ketone solution, 30 %)	binder	13.32

[a] TAM Ceramic Co., Niagara Falls, New York.
[b] Defloc Z-3, Spencer Kellogg Inc., Buffalo, New York.
[c] Monsanto Inc., St. Louis, Missouri.
[d] Rohm and Haas Co., Philadelphia, Pennsylvania.

Most commonly a moving carrier film passes under a reservoir containing the slurry. Steel and Mylar are among the materials used as a carrier film. As the film moves beyond the reservoir, a thin layer of the slurry is carried on the film. To maintain a consistent, controllable thickness of the layer, a "doctor blade" is positioned at a predetermined height above the carrier film. This doctor blade trims off excess slurry so that the carrier film beyond the doctor blade has a constant thickness. Variations in design have incorporated the reservoir and doctor blade into an entity that moves over a fixed surface.

$PbTiO_3$. In 1950, indirect evidence suggested that lead titanate [*12060-00-3*] was ferroelectric [57]. Not until 1970was ferroelectricity in $PbTiO_3$ actually verified [58]. Lead titanate has the cubic perovskite crystal structure above its Curie point, ca. 490 °C. Below the Curie point the structure is tetragonal. At room temperature, the lattice constants are $a = 0.3895$ nm and $c = 0.4146$ nm, for a c/a ratio of 1.064 [59]. This relatively large tetragonal distortion of the cubic perovskite structure leads to a large amount of strain in cooling through the cubic – tetragonal transition temperature. This spontaneous strain and a large thermal expansion anisotropy is thought to break up polycrystalline bodies of pure $PbTiO_3$ during cooling [60]. For these reasons polycrystalline $PbTiO_3$ must be formed in the presence of additives [61].

$PbTiO_3$ is usually formed by the solid-state reaction of PbO and TiO_2. The compound can be formed at temperatures as low as 375 °C [62]. Recently, however, a new technique for forming $PbTiO_3$ has been reported [63]. This technique involves preparation of a porous $PbTiO_3$ body from organometallic precursors at 70 °C and, because it lowers the densification temperature so dramatically, may lead to fully dense, pure $PbTiO_3$.

Because of the difficulty in preparing pure polycrystalline lead titanate, there are few reports on the properties of the material. Banno, however, has found that pure $PbTiO_3$ is useful in piezoelectric – polymer composites [64]. There are reports on the properties of modified $PbTiO_3$ [60], [65]. However, the commercial application of these materials is still quite small.

$PbZrO_3$. Lead zirconate [*12060-01-4*] was shown to have unusual dielectric propertiesin 1950 [66]. Because of its high dielectric constantand temperature dependence, the material was designated a ferroelectric. Actually $PbZrO_3$ is antiferroelectric [67]. Therefore, by itself lead zirconate cannot form a polycrystalline piezoelectric. Below ca. 230 °C, $PbZrO_3$ is orthorhombic with lattice constants $a = 0.587$ nm, $b = 1.174$ nm, and $c = 0.820$ nm. Above 230 °C the material is cubic.

In the original work $PbZrO_3$ was prepared by mixing ZrO_2 and PbO and then calcining in a platinum crucible for 1 h at 1050 °C. As with the forming and densification of many lead-containing compounds, the volatility of PbO must be controlled. This is usually done by firing in an atmosphere that contains PbO (see *volatility of PbO* in page 9).

The properties of $PbZrO_3$ can be altered by modifying the composition. Some of the additions can change the crystallographic phases present as well as alter the properties [68].

$Pb(Zr,Ti)O_3$. Lead zirconate and lead titanate form a complete series of solid solutions [69]. These materials are referred to generically as $Pb(Zr,Ti)O_3$ [*12626-81-2*], or *PZT*. It is importantto realize that PZT refers to a wide range of compositions rather than an individual chemical compound.

The discovery of the piezoelectric properties of PZT ceramics was a result of research around the world. In 1952, SHIRANE and SUZUKI, in Japan, investigated the crystal structure of the solid solution $PbZrO_3$–$PbTiO_3$ [70]. This work was followed in 1953 by SAWAGUCHI's comprehensive report on his investigation of ferroelectricity in the PZT solid solution [71]. In 1954, JAFFE et al., in the United States, published the first report on the piezoelectric properties of the $PbZrO_3$–$PbTiO_3$ solid solution [72]. The same authors summarized the work on PZT and expanded it in a more encompassing report in 1955 [73]. The piezoelectric properties of PZT were found to be better than those of the most popular ceramic piezoelectric at that time, $BaTiO_3$, as shown in Table 5.

Table 5. Typical properties for piezoelectric ceramics

Property	Barium titanate based	PZT based
Specific gravity	5.60	7.70
Dielectric constant, 1 MHz, 25 °C	1200	1300
Dielectric strength, MV/m	3.5	3.9
Modulus of elasticity, GPa	117	82
Piezoelectric constants		
d_{31}, C/N	-60×10^{-12}	-120×10^{-12}
d_{33}, C/N	150×10^{-12}	290×10^{-12}
g_{31}, V/N	-5×10^{-3}	-10×10^{-3}
g_{33}, V/N	14×10^{-3}	25×10^{-3}
Coupling coefficients		
K_{31}	0.2	0.3
K_{33}	0.5	0.7
K_P	0.3	0.6
Mechanical Q	400	500
Curie temperature, °C	120	345

As Zr replaces Ti in $PbTiO_3$, the amount of tetragonal distortion (*c/a* ratio) decreases. At ca. 55 mol% $PbZrO_3$, an abrupt change in the crystal structure of the solid solution is observed. This sort of abrupt change in structure with compositional change is known as a morphotropic phase transformation. The *morphotropic phase boundary* (MPB) in PZT is nearly independent of temperature. Compositions near the MPB yield materials with desirable piezoelectric properties over a wide temperature range. Therefore, the most useful commercial PZT compositions are near the MPB. A commonly used composition has a Zr : Ti molar ratio of 53 : 47 [73].

Solid solutions higher in $PbZrO_3$ content than 55 mol% have rhombohedral crystal structures. Above 90 mol% $PbZrO_3$, the crystal structure is orthorhombic. The tetragonal and rhombohedral phases are ferroelectric, while the orthorhombic phase is antiferroelectric.

PZT is commonly formed from the solid-state reaction of PbO, TiO_2, and ZrO_2. There have been many investigations into the calcination reactions. When the three oxide powders are mixed and heated, $PbTiO_3$ first forms exothermically between 450 and 600 °C, and then PZT forms above 700 °C. No measurable amount of $PbZrO_3$ was observed in the mixture. A multistep sequence was proposed to explain the reaction kinetics [74].

As with most technical ceramics, the processing of PZT articles is exacting and critical to their performance. The problems involved in balancing densification with grain growth are compounded in PZT by the *volatility of PbO*. Densification, grain growth, and evaporation rates all increase with temperature. To fully densify a PZT piece while minimizing grain growth and eliminating lead loss is not possible solely by varying the temperature during firing. To minimize lead loss, pieces are usually fired in a closed crucible containing a PbO-rich atmosphere, usually created by packing the PZT pieces in a powder containing lead oxide. This packing powder can be of the same composition as the pieces, or it can be PbO, $PbZrO_3$, etc. The packing powder affects the microstructure and properties of PZT ceramics [75]: the ideal arrangement seems to be to use a closed system with a packing powder of $PbZrO_3$ saturated with PbO.

A major effort through the years has been devoted to preparing PZT ceramics with a good microstructure and good electrical properties. Hot pressing has been used to obtain pieces with a dense, uniform microstructure. However, high-density PZT can be processed without hot pressing by careful control of composition, impurities, calcining, forming, and sintering [76]. An excess of PbO in the PZT composition acts as a liquid-phase densification aid, and is commonly used. However, while a PbO-rich liquid phase

increases the initial densification rate, it lowers final density [77]. Additions of V_2O_5 [78] and of LiF [42] also lower the sintering temperature.

Soon after the discovery of the useful piezoelectric properties of PZT, work was directed at improving these properties via *chemical modifications.* Strontium is especially effective [79]. When Sr is substituted for Pb in PZT, the dielectric constant and the electromechanical coupling factor are greater near the MPB. Strontium substitution was also found to lower the Curie point and to shift the MPB to a composition slightly more rich in Zr.

Near the MPB, additions of some +3 or +5 ions (La, Nd, Nb, and Ta) increase the dielectric constant and give the PZT better aging and stability characteristics [82]. Modifying PZT with tungsten and thorium gives results similar to those for PZT containing Nb or La [81]. Additions of Fe_2O_3 to PZT compositions containing 45 – 60 mol% $PbZrO_3$ (i.e., both sides of the MPB) increase the mechanical and dielectric stiffnesses, while the mechanical and dielectric losses decrease [80]. The effects were greater for the rhombohedral $PbTiO_3$-rich compositions than for the tetragonal compositions.

Solid solutions of PZT with other perovskite ferroelectrics have also been studied. Several of these involved a material $Pb(Nb_{2/3}X_{1/3})O_3$, with X being Mg [83], Co [84], or Ni [85]. These systems contain pseudocubic phases as well as tetragonal and rhombohedral phases. The materials that were produced had high dielectric constants and electromechanical coupling coefficients and good temperature stability.

Uses. There are many possible uses for devices based on piezoelectric materials to convert mechanical to electrical energy, and vice versa. The PZT family of materials is the most commercially important ceramic piezoelectric. By modification of the composition and control of the microstructure through processing variations, materials can be tailored to a specific use. Table 6 lists some uses. A monograph describing the uses of PZT ceramics, giving operating characteristics and a variety of specific applications, is entitled *Piezoelectric Ceramics* [86].

Tungsten-modified PZT is suitable for piezoelectric ignition of gases, W – Mn-modified PZT is useful as ceramic wave filters, and Nb – Mnmodified PZT is suitable for high-power ultrasonic transducers [87].

Table 6. Uses of PZT ceramics

Use	Devices
High-voltage generators cigarette lighter smallgasoline-motorspark generator static remover	igniter for gas appliances
Flexural element liquid-level sensor fine-motion controller strain gage phonograph pickup	buzzer
Resonators electrical resonators	electrical filters
Sound/ultrasound generators intruder alarm tone generator echo sounder hydrophone ultrasonic cleaners	remote control of appliances
Electro-optics flash shield	displays

PZT can be used to generate high voltages. The voltages produced are high enough to cause sparking (5 – 20 kV). The spark can ignite gas in an appliance such as a stove or clothes dryer. A piezoelectric igniter makes pilot flames unnecessary, reducing the amount of gas used by the appliance. The mechanical force can be applied to the PZT piece by a spring-loaded hammer or by a lever system. In a related device, air in the vicinity of the tip of a wire connected to a PZT block is ionized to remove static from phonograph records.

Flexural elements are also commonly made from PZT ceramics. The amount of motion generated by a single PZT element is often too small, e.g., 4 μm for an applied 10 kV [86]. However, a laminated cantilever structure, a *bimorph,* can be used. A bimorph consists of two thin layers of PZT laminated together. Often a metallic layer is used between the PZT layers to increase the durability of the bimorph and the electromechanical coupling. The way the bimorph is mounted tailors the device to the application.

Mechanical resonances possible in PZT have been exploited in electrical resonators and filters. The dimensions of the PZT disk and the mode of oscillation (e.g., radial or thickness) determine the frequency.

PZT transducers can be used to generate sound and ultrasound in air. The same materi-

als can be used as receivers. Remote control of slide projectors and televisions has been developed on the basis of these systems. Other uses are intruder alarms, telephone microphones, and tone generators. As with flexural elements, laminated structures are commonly used for ultrasound generators or receivers.

Boats commonly have echo-sounding equipment to determine water depth, and in some cases to locate fish. The systems in small boats usually require a PZT transducer with resonant frequency in the range 150 – 200 kHz. This transducer is also capable of receiving the signal. A related device is a hydrophone, which is used for underwater detection by passively listening rather than generating and receiving a signal. Composites made from PZT and epoxy offer some unique properties that can be exploited in hydrophone as well as other applications [88].

When high-intensity ultrasonic signals must be generated, such as in ultrasonic cleaning, a sandwich or composite transducer is used. The intensity of the signal necessitates a large-volume transducer.

When doped with rare-earth or a few other additives, PZT was found to act as an electrooptic material [89]; the optical properties could be electrically controlled (see the section on PLZT). An early composition contained 2 mol% Bi and a Zr : Ti molar ratio of 65 : 35 [90].

PLZT. Doped PZT was found to be an electro-optic material. Lanthanum doping produces PZT ceramics with good electro-optical properties and a high optical transparency [91]. This family of materials is designated *PLZT* for lead lanthanum zirconate titanate. A useful electro-optic composition has a Zr : Ti molar ratio of 65 : 35 and 8 – 9 mol % La. The specific compositions are often designated by a three-term ratio of "mole percents" in the form La/Zr/Ti, e.g., 9/65/35 PLZT.

The phase relations in the X/65/35 PLZT system, $0 \leq X \leq 15$, was investigated [92]. At room temperature there are two ferroelectric rhombohedral phases, with the transition occurring near $X = 1$. For $9 \leq X \leq 12$, there is a pseudocubic phase, designated antiferroelectric on the basis of its dielectric properties. For $X > 12$, there is a cubic phase.

In the initial work on PLZT by HAERTLING and LAND [91], the samples were *hot-pressed* to reach optical transparency. The calcined powder was initially cold-pressed at 24 GPa in an alumina hot-press mold [93]. The mold assembly was then placed into the hot-pressing furnace, and pressure was applied from both ends. Typical hot-pressing conditions were 1100 °C for 16 h at 13.8 GPa [91]. Above 1100 °C, ZrO_2 powder acts as a setter to prevent reaction between the alumina mold and the PLZT. Hot pressing is still used to fabricate PLZT pieces.

Transparent PLZT ceramics can be fabricated without hot pressing [94]: cold-pressed 9/65/35 PLZT samples are first sintered in O_2 in platinum crucibles for 45 min at 1180 °C and then heat treated in air for 60 h at 1200 °C in Al_2O_3 crucibles containing $PbZrO_3$ powder.

Uses. Displays made from PLZT can image, store, and display information. PLZT displays have several advantages over other types of electronic displays [95]. They can be selectively erased and do not need to be refreshed: part or all of the information can be erased, and once information is input, no power is required to maintain it. The memory is *nonvolatile.*

The thin PLZT plate used in displays is transparent, initially optically isotropic. Poling reduces the symmetry to uniaxial. A uniaxial material is birefringent; that is, the index of refraction is different for different polarizations of light. By selective reorientation of the domains in the ceramic, the birefringence can be electrically controlled and an image stored.

Electro-optic PLZT has also been used for flash-protection goggles: when a light sensor is triggered by a high-intensity flash, a signal is sent to the transparent PLZT, which reorients the domains and darkens the material, blocking the transmission of the light flash.

$PbNb_2O_6$. Lead metaniobate [*12034-88-7*], $PbNb_2O_6$, was the first nonperovskite oxide ferroelectric discovered. After this discoveryby GOODMAN [96], other lead-based nonperovskite ferroelectrics, not discussed here, were identified [97].

The Curie point of $PbNb_2O_6$ is 570 °C [96]. The paraelectric phase has a tetragonal crystal structure similar to that of some alkali-metal tungsten bronzes, such as $K_{0.57}WO_3$ [98], [99]. Below the Curie point, two phases have been observed, a rhombohedral nonferroelectric phase and an orthorhombic ferroelectric phase that

is metastable at room temperature [100]. The lattice constants of the ferroelectric form are $a = 1.765$ nm, $b = 1.791$ nm, $c = 0.7736$ nm [98].

In the original work, lead metaniobate was prepared in a two-step process. The first step is a calcination of Nb_2O_5 and $PbSO_4$ in the form of pressed pellets at 1275 °C for 1 h in air, to form $PbNb_2O_6$. The calcined pellets are crushed and repressed. The final firing is carried out in platinum vessels at 1250 °C. Lead volatility is less a problem than for PZT.

Modified lead niobate has been used in transducers. Because of its high Curie point, $PbNb_2O_6$ can be used at higher temperatures than PZT, but the relatively high electrical conductivity of lead metaniobate limits its use at higher temperatures [101].

4. Semiconducting Ceramics

The broad definition that ceramics are nonmetallic inorganic solids classifies the common semiconductors (silicon, germanium, gallium arsenide, etc.) as ceramic materials. However, they are not normally considered ceramics and, therefore, are not discussed here. Electronically, ceramic materials can be regarded as wide-band-gap semiconductors and as such can have useful electronic properties. Semiconducting ceramic materials are commonly used as phosphors, resistors, and sensors, as was shown in Table 1, and can be used as active semiconductors in some applications.

SiC. Silicon carbide [*409-21-2*] is a semiconductingmaterial that could be useful in high-frequency and high-operating-temperatureuses. In addition, SiC-based devices could be useful as blue-light-emitting diodes and high-power – high-frequency devices [102], [103]. Among the devices fabricated from SiC are bipolar transistors [104] and MOS diodes [105].

Silicon carbide exists in a number of crystallographic polytypes, The two polytypes used most commonly for devices are 3 *C*, which is cubic, and 6 *H*, which is hexagonal. In the following discussion, the two polytypes will be differentiated where possible.

For high-frequency uses there are different figures of merit in the literature [106], [107]. Silicon is predicted to outperform gallium arsenide, or vice versa, depending upon which figure of merit is used. However, silicon carbide is predicted to outperform both materials in high-frequency applications. This potential advantage is described in the SiC literature [102], [108], and even in the gallium arsenide literature [109].

In choosing a material to be used as a semiconductor at high temperature, several factors must be considered. A wide electronic band gap is necessary so that the thermally generated intrinsic carriers (e.g., electrons excited across the band gap) do not outnumber the extrinsic carriers added intentionally via doping. The carrier mobility should be high at these elevated temperatures so that the carriers can conduct. Any operating solid-state device will generate heat; thus, the material should have good thermal conductivity. In addition, the material must be stable at the operational temperature.

These considerations have been discussed [110] in a comparison of the properties of silicon, gallium arsenide, gallium phosphide, silicon carbide, and diamond. Of these, gallium phosphide [*12063-98-8*], GaP, and silicon carbide have been actively investigated.While GaP and cubic SiC have similar band gaps and mobilities, the inferior physical stability of GaP at elevated temperature makes it a less useful material than SiC.

To produce SiC-based devices, several technologies must be fairly well developed:

single-crystal growth
doping methods
junction fabrication
contacting

Crystal Growth. Many of the simpler crystal growth techniques developed for silicon and compounds like GaP are inapplicable to SiC because the material sublimes rather than melts. Thus, Czochralski growth and zone refining cannot be used with SiC. In addition, the many polytypes make producing crystals of a desired polytype difficult.

A technique developed at NASA-Lewis (United States) to produce thick single-crystalline layers of cubic SiC grows a buffer layer on single-crystal Si [111]. The SiC layer is then grown on this buffer layer via chemical vapor deposition (CVD).

Research at Siemens in Germany has led to a process for growing relatively large crystals of 6 *H* SiC [112]. This process uses a hollow growth tube in which polycrystalline SiC is sublimed at one end and deposited on a seed crystal at the other end. Single crystals up to 24 mm long and 20 mm in diameter have been grown.

Doping and Device Fabrication. Solid-state device technology depends on being able to control precisely the local impurity levels within the host semiconductor. The semiconductor should start with as few impurities as possible, and the desired dopants can then be incorporated where desired with little damage to the rest of the circuit. Nitrogen is an electrically active *n*-type dopant in SiC and is usually present in the atmosphere when crystals are grown. For these reasons, most as-grown SiC crystals start off *n*-type, although overcompensation for the nitrogen impurities by adding a sufficient number of *p*-type dopants to the crystal during the growth process is possible. Therefore, both *n*- and *p*-type starting crystals are available. In silicon-based device technology, two of the more common doping methods are diffusion and ion implantation.

Diffusion of impurities into SiC requires temperatures above 1800 °C. At these high temperatures SiC decomposes; therefore, a protective atmosphere must be created. Light-emitting diodes can be fabricated by diffusing boron and aluminum into 6 *H* SiC crystals surrounded by a protective atmosphere [113]. The diffusion is accomplished at 2150 – 2250 °C for 2 – 5 h. More recently, the diffusion of boron into *p*-type SiC [114] was studied to determine the effect of temperature (1500 – 2550 °C) and acceptor dopant concentrations on the diffusivity of boron.

Ion implantation has been accomplished in SiC since the late 1960s. Implantation of nitrogen, boron, and phosphorus was studied [115]. For example, 1-MeV N^+ ions can be implanted to a depth of 1 μm.

To fabricate a device, *p* – *n* junctions must be fabricated. This can be done by starting with an *n*- or *p*-type material and selectively counterdoping by diffusion or implantation. A more useful technique has proven to be epitaxial growth of a junction [110]. In this process the dopant source is changed from *n*- to *p*-type (or vice versa) during crystal growth. Thus, a *p* – *n* junction is grown directly. A problem is that little postgrowth control is possible.

Device definition is possible by using oxidation and etching processes [102]: an oxide layer of 400 nm thickness was grown at 1070 °C in wet oxygen. Conventional photolithographic techniques were used, and the SiC was selectively etched in an $Ar - Cl_2 - O_2$ mixture at 1050 °C.

A solid-state device must have ohmic (nonrectifying) contacts to the *n*- and *p*-type regions. A thin nickel layer for ohmic contact to *n*-type material and an Al – Si eutectic contact for the *p*-type material have been used in 6 *H* SiC [102]. On 3 *C* SiC, Ni – SiC mixed layers may provide ohmic contact [116].

5. Sensors

Semiconducting ceramics are useful in a wide variety of sensors. The electrical properties of these materials are affected by changes in ambient temperature, atmosphere, and electrical parameters, e.g., voltage fluctuations. Through careful control of the chemical composition and microstructure of the device, inexpensive sensors can be made. The field of ceramic sensors is expanding rapidly and has been the subject of reviews [117], [118] and conferences.

ZrO_2. Zirconium dioxide [*1314-23-4*] is used for oxygensensing [119]. An electrochemical cell is made inwhich there is a known fugacity of oxygen on one side of the sensor and the unknown fugacity on the other side. This sets up an electrochemical potential drop, generating a voltage across the device that is related to the fugacity.

$BaTiO_3$. Semiconducting $BaTiO_3$ has been used as a positive temperature coefficient (PTC) resistor [120]. In most semiconductors, as temperature is increased, the resistivity decreases, which corresponds to a negative temperature coefficient (NTC). However, in appropriately doped $BaTiO_3$, the resistivity increases six orders of magnitude near the Curie point. At first this effect was observed only in donor-doped polycrystalline $BaTiO_3$ that had been heat treated in an oxidizing atmosphere [120]. For such materials the PTC effect was observed across a single-grain boundary but not within

a single grain [121], showing that this phenomenon is associated with grain boundaries. The PTC effect has also been observed in porous $BaTiO_3$ [122]. There have been several explanations of this effect [123–125].

ZnO. Devices based on ZnO [*1314-13-2*] are useful as varistors.The resistance is voltage dependent, i.e., nonohmic. Additives are used toaffect the microstructure, nonohmic behavior, and stability. The microstructure is important to the nonohmic behavior. A typical composition is 96.5 wt % ZnO, 0.5 % Bi_2O_3, 1 % CoO, 0.5 % MnO, 1 % Sb_2O_3, and 0.5 % Cr_2O_3 [126].

WONG has found that the nonohmic behavior is due to semiconducting ZnO grains and a thin insulating grain-boundary phase [127]. However, CLARKE found no discrete grain-boundary phase. Instead, he observed a bismuth-rich region along the grain boundaries in a commercial ZnO varistor [128]. Several explanations for these effects have been presented [129–131].

6. References

1. H. Thurnauer, *Am. Ceram. Soc. Bull.* **56** (1977) 219 – 220, 224.
2. J. H. Koenig: *1964 Edgar Marburg Lecture,* Am. Soc. Test. Mater., Philadelphia 1965.
3. E. C. Henry: *Electronic Ceramics,* Anchor Books, Doubleday & Co., New York 1969.
4. K. Wefers, G. M. Bell: *Tech. Paper No. 19,* Alcoa Research Laboratories, Pittsburgh 1972.
5. B. Schwartz, *Am. Ceram. Soc. Bull.* **63** (1984) 577 – 581.
6. D. J. Shanefield, R. E. Mistler, *Am. Ceram. Soc. Bull.* **53** (1974) 416 – 420.
7. B. Schwartz, *J. Phys. Chem. Solids* **45** (1984) 1051 – 1068.
8. A. J. Blodgett, D. R. Barbour, *IBM J. Res. Dev.* **26** (1982) no. 1, 30 – 36.
9. Y. Takeda et al., *Adv. Ceram.* **7** (1983) 253 – 259, 260 – 268.
10. N. Iwase, A. Tsuge, Y. Sugiura, *Int. J. Hybrid Microelectron.* **7** (1984) no. 4, 49 – 53.
11. P. L. Fleischner, *Solid State Technol.* **20** (1977) no. 1, 25 – 30.
12. Y. Shimada, K. Utsumi, M. Suzuki, H. Takamizawa et al., *NEC Res. Dev.* **75** (1984) 8 – 15.
13. M. Deri: *Ferroelectric Ceramics,* Gordon and Breach Co., New York 1969.
14. B. Jaffe, W. R. Cook, Jr., H. Jaffe: *Piezoelectric Ceramics,* Academic Press, New York 1971.
15. W. G. Cady: *Piezoelectricity,* McGraw-Hill, New York 1946, p. 4.
16. W. Kanzig: *Ferroelectrics and Antiferroelectrics,* Academic Press, New York 1957, p. 5.
17. E. Wannier, *Trans. Electrochem. Soc.* **89** (1946) 47 – 71.
18. H. Hausner, *Ceram. Age* **50** (1947) 162 – 164, 190.
19. M. McQuarrie, *Am. Ceram. Soc. Bull.* **34** (1955) 169 – 172, 225 – 230, 256 – 260, 295 – 298, 328 – 331.
20. A. de Bretteville Jr., *J. Am. Ceram. Soc.* **29** (1946) 303 – 307.
21. H. Megaw, *Nature (London)* **155** (1945) 484 – 485.
22. E. Sawaguchi, M. L. Charters, *Phys. Rev.* **117** (1960) 465 – 469.
23. L. K. Templeton, J. A. Pask, *J. Am. Ceram. Soc.* **42** (1959) 212 – 216.
24. M. Pechini, US 3 330 697, 1967.
25. R. K. Sharma, N.-H. Chan, D. M. Smyth, *J. Am. Ceram. Soc.* **64** (1981) 448 – 451.
26. P. K. Gallagher, F. Schrey, *J. Am. Ceram. Soc.* **46** (1963) 567 – 573.
27. K. Kiss, J. Magder, M. S. Vukasovich, R. J. Lockhart, *J. Am. Ceram. Soc.* **49** (1966) 291 – 295.
28. K. S. Mazdiyasni, R. T. Dolloff, J. S. Smith II, *J. Am. Ceram. Soc.* **52** (1969) 523 – 526.
29. Y. Enomoto, A. Yamaji, *Am. Ceram. Soc. Bull.* **60** (1981) 566 – 570.
30. E. G. Graf, *Ceram. Age* **58** (1951) no. 6, 16 – 19.
31. E. G. Graf, *Am. Ceram. Soc. Bull.* **31** (1952) 279 – 282.
32. J. B. MacChesney, P. K. Gallagher, F. V. DiMarcello, *J. Am. Ceram. Soc.* **46** (1963) 197 – 202.
33. M. A. A. Issa, N. M. Molokhia, Z. H. Dughaish, *J. Phys. D* **16** (1983) 1109 – 1114.
34. L. Egerton, S. E. Koonce, *J. Am. Ceram. Soc.* **38** (1955) 412 – 418.
35. K. Kinoshita, A. Yamaji, *J. Appl. Phys.* **47** (1976) 371 – 373.
36. W. R. Buessem, L. E. Cross, A. K. Goswami, *J. Am. Ceram. Soc.* **49** (1966) 33 – 39.
37. K. Traub, C. A. Best, W. J. Baldwin, *Ceram. Age* **65** (1955) no. 1, 9 – 14.
38. M. N. Swilam, A. M. Gadalla, *Trans. J. Br. Ceram. Soc.* **74** (1975) no. 5, 165 – 169.
39. K. Ramesh Chowdary, E. C. Subbarao, *Ferroelectrics* **37** (1981) 689 – 692.

40. B. E. Walker, Jr., R. W. Rice, R. C. Pohanka, J. R. Spann, *Am. Ceram. Soc. Bull.* **55** (1976) 274 – 276, 284 – 285.
41. J. M. Haussonne, G. Desgardin, P. Bajolet, B. Raveau, *J. Am. Ceram. Soc.* **66** (1983) 801 – 807.
42. S. L. Fu, C. C. Wei, S. Y. Cheng, T. P. Yeh, *Int. J. Hybrid. Microelectron.* **8** (1985) no. 1, 1 – 5.
43. G. Desgardin, I. Mey, B. Raveau, J. M. Haussonne, *Am. Ceram. Soc. Bull.* **64** (1985) 564 – 570.
44. D. W. Hamer, *Ceram. Ind. (Chicago)* **93** (1969) no. 1, 49 – 56, 68 – 70.
45. J. C. Williams: "Ceramic Fabrication Processes," in F. F. Y. Wang (ed.): *Treatise on Materials Science and Engineering*, vol. 9, Academic Press, New York 1979, pp. 173 – 198.
46. R. E. Mistler, R. B. Runk, D. J. Shanefield in G. Y. Onoda, Jr., L. L. Hench (eds.): *Ceramic Fabrication Processing Before Firing,* John Wiley & Sons, New York 1978, pp. 411 – 448.
47. H. A. Gardner: *Physical and Chemical Examination of Paints, Varnishes, and Colors,* Institute of Paint and Varnish Research, Washington, D.C., 1925.
48. G. N. Howatt, US 2 582 993, 1952.
49. G. N. Howatt, R. G. Breckenridge, J. M. Brownlow, *J. Am. Ceram. Soc.* **30** (1947) 237 – 242.
50. J. J. Thompson, *Am. Ceram. Soc. Bull.* **42** (1963) 480 – 481.
51. C. Wentworth, G. W. Taylor, *Am. Ceram. Soc. Bull.* **46** (1967) 1186 – 1193.
52. G. J. Asher, *Ceram. Age* **87** (1971) no. 9, 28 – 31.
53. D. J. Shanefield, R. E. Mistler, *West. Electr. Eng.* **15** (1971) no. 2, 26 – 31.
54. J. V. Biggers, T. R. Shrout, W. A. Schulze, *Am. Ceram. Soc. Bull.* **58** (1979) 516 – 518, 521.
55. J. B. Blum, W. R. Cannon, *Mater. Res. Soc. Symp. Proc.* **40** (1985) 77 – 82.
56. R. J. MacKinnon, J. B. Blum, *Adv. Ceram.* **9** (1984) 150 – 157.
57. G. Shirane, S. Hoshino, K. Suzuki, *J. Phys. Soc. Jpn.* **5** (1950) 453 – 455.
58. J. P. Remeika, A. M. Glass, *Mater. Res. Bull.* **5** (1970) 37 – 45.
59. T. Y. Tien, E. C. Subbarao, J. Hrizo, *J. Am. Ceram. Soc.* **45** (1962) 572 – 575.
60. T. Y. Tien, W. G. Carlson, *J. Am. Ceram. Soc.* **45** (1962) 567 – 571.
61. E. C. Subbarao, *J. Am. Ceram. Soc.* **43** (1960) 119 – 122.
62. S. S. Cole, H. Espenschied, *J. Phys. Chem.* **41** (1937) 445 – 451.
63. S. Gurkovich, J. B. Blum in L. L. Hench, D. R. Ulrich (eds.): *Ultrastructure Processing of Ceramics, Glasses and Composites,* John Wiley & Sons, New York 1984, pp. 152 – 160.
64. H. Banno, *Ferroelectrics* **50** (1983) 3 – 12.
65. T. Yamamoto, H. Igarashi, K. Okazaki, *J. Am. Ceram. Soc.* **66** (1983) 363 – 366.
66. S. R. Roberts, *J. Am. Ceram. Soc.* **33** (1950) 63 – 66.
67. E. Sawaguchi, H. Maniwa, S. Hoshino, *Phys. Rev.* **83** (1951) 1078.
68. Jaffe, Cook, Jaffe [14] pp. 123 – 131.
69. S. Fushimi, T. Ikeda, *J. Am. Ceram. Soc.* **50** (1967) 129 – 132.
70. G. Shirane, K. Suzuki, *J. Phys. Soc. Jpn.* **7** (1952) 333.
71. E. Sawaguchi, *J. Phys. Soc. Jpn.* **8** (1953) 615 –629.
72. B. Jaffe, R. S. Roth, S. Marzullo, *J. Appl. Phys.* **25** (1954) 809 – 810.
73. B. Jaffe, R. S. Roth, S. Marzullo, *J. Res. Natl. Bur. Stand. U.S.* **55** (1955) 239 – 254.
74. S. S. Chandratreya, R. M. Fulrath, J. A. Pask, *J. Am. Ceram. Soc.* **64** (1981) 422 – 425.
75. S. S. Chiang, M. Nishioka, R. M. Fulrath, J. A. Pask, *Am. Ceram. Soc. Bull.* **60** (1981) 484 – 489.
76. Y. S. Kim, R. J. Hart in H. Palmour III, R. F. Davis, T. M. Hare (eds.): *Processing of Crystalline Ceramics,* Plenum Publ., New York 1978, pp. 323 – 333.
77. A. I. Kingon, J. B. Clark, *J. Am. Ceram. Soc.* **66** (1983) 256 – 260.
78. D. E. Wittmer, R. C. Buchanan, *J. Am. Ceram. Soc.* **64** (1981) 485 – 490.
79. F. Kulcsar, *J. Am. Ceram. Soc.* **42** (1959) 49 – 51.
80. F. Kulcsar, *J. Am. Ceram. Soc.* **42** (1959) 343 – 349.
81. F. Kulcsar, *J. Am. Ceram. Soc.* **48** (1965) 54.
82. T. B. Weston, A. H. Webster, V. M. McNamara, *J. Am. Ceram. Soc.* **52** (1969) 253 – 257.
83. H. Ouchi, K. Nagano, S. Hayakawa, *J. Am. Ceram. Soc.* **48** (1965) 630 – 635.
84. T. Kudo, T. Yazaki, F. Naito, S. Sugaya, *J. Am. Ceram. Soc.* **53** (1970) 326 – 328.
85. H. Banno, T. Tsunooka, I. Shimano, *Proc. Meet. on Ferroelectric Mater. and Their Applications 1975, 1st,* 339 – 344.
86. J. van Randeraat, R. E. Setterington (eds.): *Piezoelectric Ceramics,* Mullard, London 1974.

87. H. Banno, paper no. 2-FS-80P, presented at the Pacific Coast Regional Meeting – Am. Ceram. Soc., San Francisco, October 27, 1980; abstract, *Am.Ceram. Soc. Bull.* **59** (1980) 824.
88. R. E. Newnham, L. J. Bowen, K. A. Klicker, L. E. Cross, *Mater. Eng. (Reigate, U.K.)* **2** (1980) 93 – 106.
89. A. H. Meitzler, J. R. Maldonado, D. B. Fraser, *Bell Syst. Tech. J.* **49** (1970) 953 – 967.
90. C. E. Land, P. D. Thatcher, *Proc. IEEE* **57** (1969) 751 – 768.
91. G. H. Haertling, C. E. Land, *J. Am. Ceram. Soc.* **54** (1971) 1 – 11.
92. H. M. O'Bryan, Jr., *J. Am. Ceram. Soc.* **56** (1973) 385 – 388.
93. G. Haertling, *J. Am. Ceram. Soc.* **49** (1966) 113 – 118.
94. G. S. Snow, *J. Am. Ceram. Soc.* **56** (1973) 91 – 96, 479 – 480.
95. A. H. Meitzler, J. R. Maldonado, *Electronics* 1971, Feb. 1, 34 – 39.
96. G. Goodman, *J. Am. Ceram. Soc.* **36** (1953) 368 – 372.
97. Jaffe, Cook, Jaffe [14] pp. 217 – 222.
98. M. H. Francombe, B. Lewis, *Acta Crystallogr.* **11** (1958) 696 – 703.
99. R. S. Roth, *Acta Crystallogr.* **10** (1957) 437.
100. M. H. Francombe, *Acta Crystallogr.* **9** (1956) 683.
101. Jaffe, Cook, Jaffe [14] pp. 215 – 216.
102. E. Pettenpaul, W. von Munch, G. Ziegler, *Conf. Ser. Inst. Phys.* **53** (1980) 21 – 35.
103. J. Feitknecht, *Springer Tracts Mod. Phys.* **58** (1971) 48 – 118.
104. W. von Munch, P. Hoeck, *Solid-State Electron.* **21** (1978) 479 – 480.
105. A. Suzuki, K. Mameno, N. Furui, H. Matasunami, *Appl. Phys. Lett.* **39** (1981) 89 – 90.
106. E. O. Johnson, *RCA Rev.* **26** (1965) 163 – 177.
107. R. W. Keyes, *Proc. IEEE* **60** (1972) 225.
108. R. W. Keyes, *Silicon Carbide Proc. Int. Conf. 3rd 1973* , 1974, 534 – 541.
109. M. Nowogrodzki (ed.): *Advanced III – V Semiconductor Materials Technology Assessment,* Noyes Publications, Park Ridge, NJ, 1984, p. 178.
110. J. A. Powell, *NASA Tech. Memo.* **83514** (1983) 1 – 5.
111. S. Nishino, J. A. Powell, H. A. Will, *Appl. Phys. Lett.* **42** (1983) 460 – 462.
112. G. Ziegler, P. Lanig, D. Theis, C. Weyrich, *IEEE Trans. Electron Devices* **ED-30** (1983) 277 – 281.
113. J. M. Blank, *Silicon Carbide Proc. Int. Conf. 2nd 1968,* 1969, 179 – 186.
114. E. N. Mokhov, E. E. Goncharov, G. G. Ryabora, *Sov. Phys. Semincond. (Engl. Transl.)* **18** (1984) 27 – 30.
115. D. E. Davies, J. J. Comer, *Silicon Carbide Proc. Int. Conf. 3rd 1973,* 1974, 640 – 644.
116. D. Fathy, J. Narayan, O. W. Holland, B. R. Appleton et al., *Mater. Lett.* **2** (1984) 324 – 327.
117. P. McGeehin, D. E. Williams in H. Krockel, M. Merz, O. van der Biest (eds.): *Ceramics in Advanced Energy Technologies,* D. Reidel, Boston 1984, pp. 422 – 468.
118. B. M. Kulwicki, *J. Phys. Chem. Solids* **45** (1984) 1015 – 1031.
119. E. M. Logothetis, *Adv. Ceram.* **3** (1981) 388 – 405.
120. B. M. Kulwicki, *Adv. Ceram.* **1** (1981) 138 – 154.
121. H. Nemoto, I. Oda, *J. Am. Ceram. Soc.* **63** (1980) 398 – 401.
122. M. Kuwabara, *J. Am. Ceram. Soc.* **64** (1981) 639 – 644.
123. W. Heywang, *J. Am. Ceram. Soc.* **47** (1964) 484 – 490.
124. J. Daniels, K. H. Hardtl, R. Wernicke, *Philips Tech. Rev.* **38** (1978/1979) no. 3, 73 – 82.
125. M. Kuwabara, *Adv. Ceram.* **7** (1984) 128 – 136.
126. M. Matsuoka, *Adv. Ceram.* **1** (1981) 290 – 308.
127. J. Wong, *J. Am. Ceram. Soc.* **57** (1974) 357 – 359.
128. D. R. Clarke, *J. Appl. Phys.* **50** (1979) 6829 – 6832.
129. G. D. Mahan, L. M. Levinson, H. R. Philipp, *J. Appl. Phys.* **50** (1979) 2799 – 2812.
130. F. A. Selim, T. K. Gupta, P. L. Hower, W. G. Carlson, *J. Appl. Phys.* **51** (1980) 765 – 768.
131. P. Williams, D. L. Krivanek, G. Thomas, M. Yodogawa, *J. Appl. Phys.* **51** (1980) 3930 – 3934.

Cereals

Wilfried Seibel, Detmold, Germany

Okkyung Kim Chung, USDA-ARS, Grain Marketing and Production Research Center, Manhattan, Kansas, USA

Dorian Weipert, Detmold, Germany

Seok-Ho Park, USDA-ARS, Grain Marketing and Production Research Center, Manhattan, Kansas, USA

1. Introduction . . . 1
2. Production . . . 2
2.1. Breeding . . . 2
2.2. Cultivation . . . 5
2.3. Handling . . . 7
3. Nutritional and Technological Properties . . . 8
3.1. Structure of Cereal Grains . . . 8
3.2. Composition of Cereals-Primary Constituents . . . 9
3.2.1. Starch (→ Starch) . . . 9
3.2.2. Proteins . . . 10
3.2.3. Pentosans . . . 10
3.2.4. Lipids . . . 11
3.2.5. Secondary Constituents . . . 12
3.2.5.1. Cellulose (→ Cellulose) . . . 12
3.2.5.2. β-Glucan . . . 12
3.2.5.3. Lignins (→ Lignin) . . . 12
3.2.5.4. Lignans . . . 13
3.2.5.5. Vitamins . . . 13
3.2.5.6. Mineral Matters and Phytate . . . 13
3.2.5.7. Tannins . . . 14
4. Standards in Grain Trade . . . 14
4.1. Test Weight (TW) . . . 14
4.2. 1000 Kernel Weight (TKW) . . . 15
4.3. Besatz . . . 15
4.4. Dockage . . . 15
4.5. Foreign Material . . . 15
4.6. Sprout Damage . . . 15
4.7. Chemical Characteristics that Determine Wheat Quality . . . 15
5. Uses . . . 16
5.1. Human Food and Animal Feed . . . 16
5.1.1. Wheat . . . 16
5.1.2. Rice . . . 19
5.1.3. Maize (Corn) . . . 19
5.1.4. Other Cereals . . . 20
5.2. Importance of Whole Cereals, Refined Products, and Fortification for Nutrition, Health, and Requirement . . . 21
6. Processing . . . 22
6.1. Dry-Milling Process . . . 22
6.1.1. Wheat . . . 22
6.1.2. Rye . . . 27
6.1.3. Maize . . . 27
6.2. Dehulling . . . 27
6.3. Wet-Milling Process . . . 28
6.4. Extrusion Process . . . 28
7. Specific Properties and Use of Grains . . . 29
7.1. Wheat . . . 29
7.2. Rye . . . 31
7.3. Triticale . . . 32
7.4. Barley . . . 32
7.5. Oats . . . 33
7.6. Rice . . . 34
7.7. Sorghum and Millets . . . 35
7.8. Maize (Corn) . . . 35
7.9. Pseudocereals . . . 36
7.9.1. Buckwheat . . . 36
7.9.2. Amaranth . . . 37
7.9.3. Quinoa . . . 37
8. Economic Aspects and Outlook . . . 38
9. References . . . 38

1. Introduction

Cereals are the oldest cultivated plants and have been grown since prehistoric times. The population of each continent prefers to cultivate cereal varieties adapted to the climate and these again formed the eating habits of the people living there. Wheat and rye originated in the Near East and have spread throughout Europe from where they were taken to North America. Asia is the cradle of rice and barley, Africa is the na-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_093.pub2

tive land of millet and sorghum, and South and Central America gave birth to maize. Cereals also have made possible permanent settlements of humans through the production and storage of food instead of hunting and gathering. Among all edible plants, the cereals provide the most nutritive and, therefore, the most important food sources for people over longer periods of time. Both the ability of cereals to adapt to a variety of soils and climate conditions and the selective work of humans over the centuries have established cereals as the basis for living in many regions of the earth.

2. Production

The production of cereals has increased over the centuries, but the most significant increase has occurred since the 1950s. This increase is doubtless a result of interaction of acreage, breeding, and tilling but the breeding seems to have been the strongest factor in this success, given the high production figures. For example, the yield of wheat in Germany in the last 50 years has increased by 1.5 dt/ha (decitons per hectare) annually. Selecting and releasing new high yielding cultivars is the main contribution to the increased productivity.

The spread of cereals worldwide, along with the amounts produced is shown in Table 1. On about a third of the worldwide arable land over 2×10^9 t of cereals were produced annually between 1999 and 2001 [1]. This is more than three times the total production of roots and tubers, the second most important food for humans and more than the sixfold tonnage of potatoes produced. Besides this other agricultural products are low in dry matter and protein content which makes them less efficient food sources. Pulses also play an important role, but in comparison to cereals their importance is marginal at best. These figures underscore the significance of cereals in nourishment and in the fight against hunger. Maize, rice, and wheat are the most-cultivated cereals, in almost equal parts in the world, and the remaining 13 % of produced cereals is shared among barley, oats, and the small-grain cereals, sorghum and millet. In this context, rice can be cropped twice or even three times a year as the climatic conditions and the technique of producing layers of developed rice plants allow. Over time and because of the increase in living standards, rye seems to have completely lost its relevance; growing rye for traditional bread is limited to a few countries in Europe. In some traditional rice-growing countries, rice is losing importance due to an increase in wheat consumption.

Cereals production in total and in different parts of the world has also changed dramatically since the 1950s. North America was an important producer and exporter of wheat, as was South America. After the Second World War, cereal production, particularly of wheat, has generally increased rapidly, but producer countries have shifted. Asia was once a buyer of wheat on the world market, but has since became a considerable producer, of about half of the world's cereal grains. Europe was also a buyer of American and Canadian wheat, particularly because of its excellent quality, but today Europe and the European Union (EU; initially 15, now 25 states) alone produces more high-quality wheat than North America. The ample supply in the world cereal market means prices are rather low, and trade with cereals is difficult and not very profitable. The prices are given in international dollars ($), a hypothetical currency unit, that has the same purchasing power as the US dollar.

In addition to the caloric and nutritive value from cereals derived from starch and protein, other constituents have been recognized to provide health benefits. The contents of dietary fiber, minerals, vitamins, and phytohormones, which are preventative against certain types of cancer, has improved the image of cereal grains. The use of whole kernel flours and meals in bakery products and of various grains in breakfast cereals is recommended.

In contrast to wheat, a significant portion of rice and maize, barley and oats, all highly estimated foods, are used as animal feed. In this context, cereals alone do not provide a complete diet and must be supplemented by the consumption of proteins from animal or plant (e.g., soybean) sources.

2.1. Breeding

From the beginning of unconscious selection by early farmers to the advanced targeted selection by today's breeders, humans have participated

Table 1. Cereal production worldwide in 1000 t (average of 1999 – 2001 [1])

Region	Wheat	Rice	Maize	Barley	Rye	Oats	Millet	Sorghum	Total cereals
Africa	15 953	17 034	42 987	2949	30	265	13 169	19 342	113 850
N. America	86 811	11 479	274 016	19 207	515	5319	336	19 892	418 230
S. America	20 489	20 852	56 903	1408	128	1204	42	4812	105 906
Asia	253 368	547 621	159 621	19 673	986	1155	9882	12 141	1 010 120
Europe	186 265	3185	70 736	85 185	19 279	16 621	903	724	369 945
EU (of 25)	105 654	2568	65 419	47 882	10 947	8435	14	629	234 289
EU (of 15)	95 035	1382	38 454	43 184	5691	6307	–	620	191 057
Oceania	23 991	1438	563	6741	21	1195	53	2042	36 745
Total	**586 908**	**601 608**	**604 560**	**135 162**	**20 959**	**25 758**	**28 009**	**57 943**	**2 081 878**
% Total cereals	28.5	29.2	29.3	6.6	1.0	1.3	1.4	2.8	
Price, $/t	144.8	335.9	128.1	111.2	85.3	112.8	–	–	

in the alteration of cereal genomes for over 10 000 years. Breeding has focused upon increasing grain yield which remains the main selected trait to this day. To obtain 1 g of dehulled kernel, more than 500 kernels of wheat ancestors such as sweet grass (*Glyceria fluitans*) or more than 120 kernels of rye brome (*Bromus secalinus*) were needed. Today 1 g of cultivated barley or wheat comes from not more than 35 kernels. Since in most parts of Europe and North America hunger has been overcome, breeders have focused on selection of traits that improve technological quality. Cultivars with improved milling and baking properties, as well as improved behavior in processing to food and nonfood products have been created. Particularly in nonfood applications, cereal products are used today to replace mineral oil and other more precious materials from nonrenewable sources. These improvements came from increasing the protein content in grain by intensifying the nitrogen uptake from soil, by changing the composition of proteins and starch, and also by changing the structure or hardness of kernels. Monoculture, i.e., growing of one cereal cultivar over large areas, has led to dissemination of pests of bacterial, fungal, and insect origin. Therefore, crop protection was necessary for both the yield and the sanitation of stored grains. Breeders have developed cereal cultivars resistant to pests, and abiotic stress from drought, freezing (winter hardness), and preharvest sprouting. These inbred resistances have reduced production costs and have ensured a high crop yield free from chemical pesticides. However, resistance does not protect from new races of pests, which seem to be more virulent and do more harm to plants and grain. For this reason breeding against pests will never end.

Resistance to preharvest sprouting is essential in regions with an abundance of rain that can damage grain before harvest. This resistance can be achieved in two ways, by breeding to achieve a high level of abscisic acid (ABA, a phytohormone controlling the synthesis of enzymes initiating germination) and low levels of gibberellic acid (GA, a hormone that promotes germination), also by breeding to improve shape and form of ears, which will reduce water uptake during grain ripening. Besides this, the stability and strength of stems and whole plants can prevent lodging, thus protecting from undesired water uptake in ears. ABA governs dormancy, a state in which the grain is viable, but kept from germinating. Dormancy is also caused by presence of lipids and flavonoids, substances that affect the grain color. Influenced by time, light, and temperature, dormancy will break and its duration can be shortened. In general, rye has virtually no dormancy, wheat dormancy is not longer than a few weeks after harvest, and dormancy of plants with a high lipid content lasts several months. In arid regions, cultivars with elevated enzyme activity ensure the production of baked products without the addition of enzymes.

In selection work the breeders use different techniques. The oldest technique is crossing two or more cultivars as parents to obtain a new cultivar with desired properties. This is a long process without guarantee of success, because quite a number of genetic combinations result and the crosses must be screened for the desired results, propagated and multiplied to obtain seed for continued planting, and material to be tested

for quality. A new cultivar requires 10 or more years before it can be released on the market.

Over decades of selection breeders have developed new methods, which have shortened the process while improving the success of new varieties. Using the technique of producing hybrids, two proven parents of hybrids pollinate free (male or father plants) in the field to produce seed (on a female or mother plant) with desired and controlled properties. The male flowers containing pollen from mother plants must be removed before becoming ripe to prevent pollination. The new property is observed in the yield, which is higher than of the separate parents. This effect is high in the first year of growing but decreases if the seed from the crop is used in subsequent years, until yield becomes equal to that of the parents. Therefore, it is necessary to synthesize new hybrid seed each year. This technique is costly; the seed is more expensive than the conventionally bred seed, but promises 20 % higher yield, which more than covers the cost of seed. This technique has been used for decades to create many maize hybrids with a variety of beneficial agricultural and processing properties and has made maize production possible in regions where maize has not been grown due to adverse climatic conditions (length of daylight hours, temperature, and length of vegetation period). This technique applied to maize is fairly simple because the male and female flowers on a maize plant are situated far apart from each other and removing male flowers from female plants is easy. Producing wheat and rye hybrids is more difficult, because each flower contains anthers (male part with pollen sacs) and stigma (female part with the ovary); to ensure and to control the desired pollination, the anthers of the mother plants must be removed. This procedure is not possible to produce saleable seed, so plants with sterile pollen were found for mother plants. The father plant has to contribute a restorer gene, which should restore fertile pollination; otherwise, the hybrid plants in the field would not be fructified or produce a decent yield. Moreover retarded flowering opens the door to ergot infection. Wheat and rye hybrid cultivar production is a well understood technique, providing seed that not only increases yield, but also gives desired properties in processing.

In order to force further advances in breeding, the techniques of artificially production of mutants by X-ray or other irradiation has been applied, disturbing the chromosome splitting during nuclear division. Such techniques are still used to increase the set of rye chromosomes to a higher number that matches wheat chromosomes, necessary in breeding new triticale cultivars (see Section 7.3). Molecular biologists have contributed much to breeding success using molecular marker technique. Molecular markers are linked to the gene of interest, and thus make possible the recognition of a specific gene (establishing a property) in early generations [2].

A very useful tool for accelerating results is a rather new technique that allows reproducing a plant from a single generative cell (e.g., pollen kernel) by growing it on a substratum containing everything necessary for the growth of a cell to a new organism. This cloning, a specific kind of vegetative propagation of precious breeding material from conventionally bred or genetically modified plants, saves time and production costs in comparison to conventional breeding in the field.

Very recently, a new method of producing transgenic plants was introduced as a promising step in breeding pest-resistant plants. Molecular biologists succeeded in transferring genes from noncereal plants (e.g., grapevine, *Vitis vinifera*) and fungi (e.g., *Aspergillus giganteus*) into the ovary cells of cereal plants [3]. The resulting plants were resistant to pests (e.g., milstem) and insects (e.g., maize borer) by producing chitinase, an enzyme that degrades chitin, the main component of the insect exoskeleton. With this technology the use of chemical pesticides can be markedly decreased, which lowers the production cost and grain contamination. Currently more than 100 genetically modified plants (maize, soybean, rapeseed, potato, tomato, rice) and about 40 transgenic plants (among them soybean, barley, maize) are registered and present on the seed market [4]. Growing and using such cultivars is controlled by strict security regulations, and their presence in food must be labeled [5].

In spite of the benefits in agricultural and processing terms, producing hybrids and transgenic plants is still questioned. Although it is clear that the grain and seeds of genetically modified plants present no threat to consumers [4, 6, 7], apprehension is present that these techniques might produce something not control-

lable in the future, something that can harm consumers. The idea of protecting consumers from genetically modified and transgenic plants and from conventionally (nonorganic) produced food and feed was initiated and is supported by the green movement worldwide. Moreover, the ethical and political background that forms consumer opinions differs from country to country [7]. Transgenic cultivars in general are protected from free production and selling by patent law. These new promising techniques are expensive and affordable only by research institutions such as universities and by wealthy enterprises, which in return expect to earn money to cover the expenses of developing the cultivars and to bring profit to the owner. The techniques are, regardless of the vehement and boisterous discussion, promising, providing progress in breeding for better agrotechnical and processing qualities of cereals [7].

As a result of breeding, a new cultivar can be approved and released. The new cultivar must meet the needs of agriculture and quality in processing, and must be clearly characterized and recognizable from other cultivars. Preferences for a cultivar rely on genetically linked properties, which can also be influenced by the environment to a certain extent. The yield and the quality of grains is an interaction of genetic potential which can be increased or decreased by environmental conditions, including the technical measures applied by farmers.

2.2. Cultivation

Cereals are grown to obtain grain to use for food and feed. High quality and quantity of the grain are vital. Quality is defined by the different end uses but the quantity is defined only by dt/ha (i.e., 100 kg/ha) yields. Yield is affected by cultivar and tillage (fertilizing), which is in the hands of farmers, and of soil condition and climate, which are not. Cereals can be grown either as winter or spring forms or exclusively as spring forms. Winter forms have the advantage of growing through the spring, a longer vegetation period, and thus a higher yield than spring forms; however, winter forms need a high level of winter hardiness. In development and growth, cereal plants must pass through several phenological phases in the vegetative period from sprouting until ripeness. Among these phases three are crucial for both yield and grain quality. Early in tillering, the number of ears is fixed, and with this, the first factor determining yield. Stem elongation is the phase in which the plant passes from the vegetative into the generative stage. These phases are induced by the cold in the winter months and initiated by longer light periods in the spring. The number of spikelets in the ear is an additional basis for yield, and nitrogen is the limiting factor for the number of spikelets and flowers. Finally, in the phase of ear emergence and flowering, starch and proteins, and other minor constituents are stored in the growing kernel. Storage protein, built from nitrogen which is taken up through the roots, benefits quality, both technologically in breadmaking and nutritionally in food and feed. Thus farmers provide the cereal plants with nutrients, especially with nitrogen to ensure the best results in yield and in quality [8]. The quantities of applied nutrients can be considerable. Potassium and phosphorus are applied at planting with nitrogen; they influence the stability and fertility of the cereal plants. Additional nitrogen may be as high as 100 or even 150 kg per hectare pure N (500 to 750 kg/ha of N-fertilizer), particularly during the flowering, to obtain the expected yield. More nitrogen assists in the synthesis of protein, sufficient for processing, food, and feed quality.

The growing plant is exposed to quite a number of pests attacking the stem, leafs, and ears. The stem is especially vulnerable to pests attacking stability while pests that diminish the active area of leaves for photosynthesis reduce yield and quality of grain. The main pests are take-all fungi of cereals (Ophiobolus, Cercosporela), black and stem rust (Puccinia), net bloth (Ryncosporium; Septoria), powdery mildew (Erysiphe) smut and bunt (Tiletia, Ustilago) and some others. Ergot does not affect only rye, but also other cereals and weeds, and creates a secondary source of infection. Recently, an increase in head blight (*Fusarium culmorum*) fungus has been observed, which produces mycotoxins responsible for causing cancer. Further serious pests in maize include the corn borer (*Pyrausta nubilalis*), the wheat blossom midge (*Contarinia tritici, Sitodiplosis mosellana*), and the straw fly (*Chlorops pumilionis*) [9]. Protecting plants in the fields and stored grains in bins

Table 2. Grain yield in dt/ha (average of 1999 – 2001 [1])

Region	Wheat	Rice	Maize	Barley	Rye	Oats	Millet	Sorghum	Total cereals
Africa	18.75	22.19	16.72	7.07	5.43	11.19	6.65	8.74	12.50
North and Central America	26.85	58.72	70.08	29.89	19.57	23.59	16.09	35.03	47.24
South America	23.12	37.13	32.66	18.30	13.78	15.36	15.69	31.77	30.04
Asia	26.11	39.66	36.87	15.82	14.29	16.07	8.91	10.09	31.52
Europe	33.98	54.92	54.03	29.20	25.51	20.03	9.02	37.32	33.49
EU (of 25)	46.64	59.60	78.45	38.39	28.35	32.58	15.41	39.61	40.47
EU (of 15)	51.73	64.25	87.08	43.28	35.01	36.14	-	35.32	50.40
Oceania	19.21	86.75	58.61	20.95	6.36	18.14	13.95	30.66	14.46
World average	27.43	39.04	43.63	24.90	21.62	20.06	7.68	13.97	30.95

is therefore indispensable. The protection may require direct application of effective chemical agents or by avoiding conditions conducive to development and spread of pests (bacteria, fungi, or insects), and for lodging or preharvest sprouting. The choice of cultivars, some of which are resistant to these menaces may help ensure a safe grain crop.

Generally, cereal grains are grown using conventional methods which include choosing prolific cultivars, providing the best growing and prevention conditions, applying chemical fertilizers to replace the nutrients used by the previous crops, applying pesticides when necessary, avoiding the risks of delayed cropping and caring for the harvested grain by drying, cleaning, and controlling conditions during storage. Fertilizing techniques have been studied for years and are well understood [8]. Too much fertilizer harms the plants and increases production costs. However, this conventional growing of cereals is charged with producing a "contaminated and unhealthy" food by organic producers. Production costs and material input are relatively high and give low prices on cereal market. Thus, high production ensures a farmer's living. Because of high production costs, farmers strive to apply the lowest necessary treatments. Moreover the amount of pesticide in and on the grain can be reduced by technological measures such as cleaning, peeling, and dehulling. Cereals are still a healthy reliable food.

Organic farmers in alternative agriculture grow the old cultivars, applying only manure as fertilizer once before sowing, which provides nutrients at the stages needed, without additional applications. In this type of agriculture, chemicals are avoided. Herbal extracts are used as natural pesticides and repellents to protect plants. The input is low, but so are the yields. Thus, the prices for organic products are high. Protein content of organically produced cereals is low, and in many cases, the technological quality is also poor. In organic agriculture, several organization control standards to ensure the quality of the product. However, pollution cannot be eliminated. Moreover, in many cases the cropped grain is contaminated not by chemicals, but by mycotoxins resulting from infection with bacterial or fungal microorganisms.

Such organically produced cereals are on the market but at minimal levels. Many consumers cannot buy them or do not believe in the advantages of organic over conventionally grown cereals. In highly developed countries organic agriculture is a philosophy, a strategy to place food with a high ideological value on the market. In regions where agriculture is not developed due to poor soils and lack of water or where the environment does not favor production, yields are low and food remains scarce.

Cereal yield worldwide depends on genus and on the regions of cultivation. Generally, maize (hybrid cultivars) and rice (up to three crops a year) give the highest yield, followed by wheat (see Table 2) [1]. Moderate climate, such as in parts of Europe, allows higher yields, whereby the dry regions of Africa allow only low yields. The interaction of favorable climate and intensive agriculture can be seen in the figures

from the EU (15 states). Optimal fertilizing at the crucial stages in plant development make yield as high as 100 dt/ha (10 000 kg/ha) possible with wheat and rye. Such extremely high yield is very close to the natural limits of the genus and cannot be increased much more. European countries such as Belgium, Denmark, Germany, and the UK have the highest yield (> 70 dt/ha) while Estonia and Portugal have the lowest (> 20 dt/ha). In comparing the productivity of former EU (of 15 states) and the present EU (of 25 states), it is apparent that applying modern intensive agriculture will help increase cereal production in virtually all European states making them independent of imports of agricultural products.

2.3. Handling

After being cropped, grains must be dried to a safe level of water content, not higher than 14 – 15 %, at which microbiological activity and the activity of enzymes inside the grain is reduced to a minimum. The drying temperature should not be higher than 45 °C in grain to prevent denaturation of amino acids in proteins and the gelatinization of starch. The dryer should remove no more than 4 % moisture at one pass through the dryer. Sudden and harsh drying changes the structure of the grain. This increases breakage, which is detrimental to processing rice, breakfast cereals, and in wet-milling of maize for starch. Breakage is also detrimental for brewing and in distilling alcohol, as well as in baking bread and other baked goods. The heating regime should not harm the viability of grains; viability is one way to test the grain for heat damage. Among the many dryers, direct heating should be avoided because they may contaminate the grain with carcinogenic organic compounds such as benzopyrene.

Cleaning is an important procedure before grain is stored, transported, and particularly processed. Intensive cleaning and treating of the grain surface is actually a first step in cereal processing. The aim is to remove "dockage", "Besatz" (see Sections 4.3 and 4.4) and all impurities, such as chaff and straw, stones and clods of soil, but also weed seed, and if necessary and possible, the unsound grain. Therefore, sieves, aspiration (air stream), paddy machines, disk separators, and differently calibrated trieurs are used to sort grain by differences in specific weight, shape and form of sound grain and impurities. Screening for color using photometric devices is used to sort out ergot-infected as well as other infected grains. Transporting vehicles and bins are often sources of other contaminants that contribute to microbiological deterioration. Intensive cleaning is the basis of safe storage and processing and must be given scrupulous attention.

The storage of grain has a long tradition making human culture and civilization possible. Stored grain should not suffer a diminution in grade and value, i.e., the quality of the harvested grain must be preserved. The loss of quality is caused by both abiotic and biotic factors. Having a relatively low content of water, grain is easy to store, and the biochemical processes that take place in stored grain are well understood. Temperature should be controlled constantly. In some regions under poor storage conditions caused by biotic factors such as microorganisms (e.g. bacterial and fungal contamination), infestation by insects, rodents and birds can lead to total postharvest losses. The respiration and gradual deterioration of viability, nutritive, and processing quality may cause somewhat lesser, but still significant losses occur even under good professional storage conditions. The deterioration of processing quality of stored grain is obvious for all grains regardless of end-use. Therefore, preservation of the quality of stored grains and preventing of quality losses is critical in the fight against the loss of precious food resources.

Moisture, temperature and aeration are the factors controlling the respiration of stored grains. Only a moisture level of about 14 % water content allows safe storage over a long period without deterioration [9]. However, not only moisture, but also relative atmospheric humidity is crucial for safe storage. Water binding capacity and hygroscopicity of specific grains are different. Critical grain moisture at which the grain is in equilibrium with the relative atmospheric humidity of about 75 % differs between the grains. Fungal spores will germinate at greater than 75 % relative atmospheric humidity leading to the rapid spread of deterioration of grain. A relative atmospheric humidity less than 65 % is safe for long-term storage. Controlling temperature, moisture, and humid-

ity in grain is important in preserving quality. For dangerous changes in storage conditions, remedies such as aeration and cooling may be performed promptly. The relative atmospheric humidity can be measured and expressed easily as the ERH (equilibrium relative humidity) and A_w-value (water activity).

Increased respiration leads to increase of temperature and the release of free water, which again intensifies respiration. The microorganisms propagate in such conditions with the result of increased metabolic activity of grains and of microorganisms, deterioration results in loss of viability, changes in color, odor and taste, and formation of mycotoxins. The main microorganisms attacking grain are fungi belonging to genera *Aspergillus, Penicillium, Alternaria, Cladosporium, Helminthosporium*, and *Fusarium*, although the last named is more a field fungus, however, its activity continues in storage. Among the insects brought in with cereals and their products only a few cause serious damage. Insect pests include beetles such as *Sithophilus, Tribolium, Trogoderma, Tenebroides*, moths like *Sitotroga, Plodia, Anagasta*, and mites, i.e., *Acarus* and *Tyrophagus*. They do not harm the grain directly by destroying the endosperm or denaturing the grains, but may cause allergic reactions in consumers. Some of the insects spend their entire development from the egg into adults inside a grain and are therefore invisible to the casual observer.

Pests are either primary pests, infesting sound grain, or secondary pests, which can attack only broken or damaged grains. Pest control begins with a preventive disinfection of empty storage containers and storing grain in pest-free conditions [10]. Bulk grain can also be treated with chemical pesticides and insecticides of inorganic or organic origin as fumigants (e.g., phosphine), which develop toxic gases under moist conditions. Hermetically closed containers are advantageous. Chemical insecticides may be hazardous to the personnel and may leave residue in grains after treatment if applied inappropriately.

Physical methods of pest control include heating to more than 40°C or cooling below 0°C. In both cases the immature forms and to some extent the eggs of the pests will be killed. Infected grain can be successfully treated by ionizing radiation, in case of which even the immature insects and eggs will be destroyed. Light traps also help to control pests. Pheromone traps and food attractants are biological ways to monitor, identify, and protect grain from pests.

The grain can be stored in flat warehouses or in concrete or steel bins. The flat warehouses are less expensive to build, but vertical bins hold more, allow simple mechanical loading and unloading, and provide better control and prevention of pests. Storage under controlled atmosphere has proved to be best for long-term preservation of quality in grains.

3. Nutritional and Technological Properties

3.1. Structure of Cereal Grains

Cereals belong to the grass family known as the *Gramineae*. Cereals produce a one-seeded fruit called a caryopse or, commonly, kernel or grain. The dry fruit consists of seed and pericarp (fruit coat), and the seed is composed of three parts: seed coat, endosperm, and embryo (germ). The pericarp tightly covers the seed by attaching to the seed coat. The largest part of the seed is the endosperm, which contains the basic human nutrition resources (carbohydrate and protein), whereas the embryo, the smallest part of the seed, serves as an oil source. Kernel structure is important because it affects the process of grain harvesting, drying, handling, storage, and milling, and as a result gives different nutritional value.

The structure of the wheat kernel is shown in Figure 1. The dorsal side of the wheat grain is rounded, while the ventral side has a deep groove or crease along the entire longitudinal axis. At the apex or small end (stigmatic end) of the grain is a cluster of short, fine hairs known as brush hairs. The pericarp consists of four layers: the epidermis, hypodermis, cross cells, and tube cells. The remaining tissues of the grain are the inner bran (seed coat and nucellar tissue), endosperm, and embryo (germ). The aleurone layer consists of large rectangular, heavy-walled, starch-free cells. Botanically, the aleurone is the outer layer of the endosperm, but because it tends to remain attached to the outer coats during wheat milling, it is shown in the diagram as the innermost bran layer.

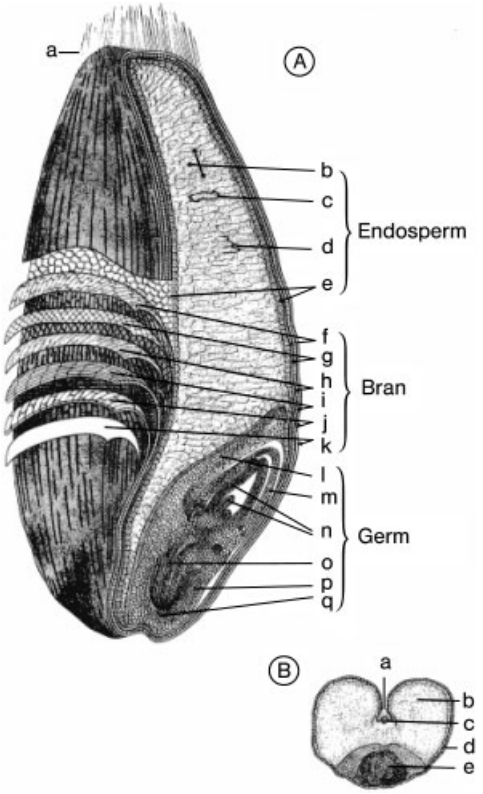

Figure 1. A wheat kernel (A) Longitudinal section (enlarged ca. 12×): a) Hairs of brush; b) Endosperm; c) Cell filled with starch granules in protein matrix; d) Cellulose walls of cells; e) Aleurone cell layer (part of endosperm but separated with bran); f) Nucellar tissue; g) Seed coat (testa); h) Tube cells; i) Cross cells; j) Hypodermis; k) Epidermis; l) Scutellum; m) Sheath of shoot; n) Rudimentary shoot; o) Rudimentary primary root; p) Root sheath; q) Root cap (B) Cross section: a) Crease; b) Endosperm; c) Pigment strand; d) Bran; e) Germ

The embryo consists of the plumule and radicule, which are connected by the mesocotyl. The scutellum serves as an organ for food storage. The outer layer of the scutellum, the epithelium, may function as either a secretory or an absorption organ. In a well-filled wheat kernel, the germ comprises 2 – 3 % of the kernel, the bran 13 – 17 %, and the endosperm the remainder. The inner bran layers (the aleurone) are high in protein, whereas the outer bran (pericarp, seed coats, and nucellus) is high in cellulose, hemicelluloses, and minerals. Biologically, the outer bran functions as a protective coating and remains practically intact when the seed germinates. The germ is high in proteins, lipids, sugars, and minerals; the endosperm consists largely of starch granules embedded in a protein matrix. Grains of rice, barley, oats, rye, and triticale are similar in structure to wheat.

Hardness of grain, or kernel texture, is an important factor that affects milling characteristics such as yield, ash content, flour particle size distribution, and finally properties of end-use products [11].

3.2. Composition of Cereals-Primary Constituents

3.2.1. Starch (→ Starch)

Starch is a major source of energy for both human and animals and is produced by plant photosynthesis. Starch has been recognized as a mixture of the essentially linear polymer amylose (M_r 250 000 – 1 900 000) and the highly branched amylopectin (M_r $10^6 - 10^8$). Amylose is made up of links (predominantly α-1,4-glycosidic linkage) of more than 2000 individual glucose units, whereas amylopectin in addition has several hundreds of branches linked by α-1,6-glycosidic linkage. The composition of amylose and amylopectin is affected mainly by genotype. Low-amylose (called waxy) and high-amylose varieties are found in diploid cereals such as maize, barley, rice, and sorghum. There are some wheat varieties (hexaploid) containing waxy starch.

Starch is deposited in the form of granules that range between 1 and 100 µm and have varying shapes depending on the botanical source. Particle size distribution is also different among cereals. Of all cereal and noncereal starches, wheat, rye, barley, and triticale are the only cereals having a bimodal starch granule size distribution, with large A type and small B type granules. Some researchers have reported a trimodal distribution in wheat starch that has C type or very small starch granules. Among the two types of wheat starch granules, the large A-granules have a diameter of 10 – 35 µm and are formed first in the developing endosperm, whereas the small B-granules have a diameter of 1 – 10 µm and are formed late in kernel development. In addition, the proportion of A and B granules deposited in an endosperm varies by genotype and environment. Generally, large A granules comprise

around 70 % of total starch weight with less than 10 % of total number of starch granules in the endosperm and vice versa.

The length of the starch chain is responsible for the physical properties of starch such as water binding capacity, pasting and gelatinization, and degradability by enzymes present in the kernel or imported by microorganisms.

3.2.2. Proteins

On the basis of solubility, the original sequential procedure for separation, cereal proteins are divided into four groups: albumins, soluble in water; globulins, soluble in diluted salt solution; prolamin, soluble in aqueous ethanol; and glutelins, soluble or dispersible in diluted acid or alkali [13]. The proportion of each solubility fraction is different among cereal grains. A reciprocal relation between the prolamin contents and the biological value of cereal proteins was estimated. This is because the cereal prolamin fraction generally has a poor amino acid balance (low level of the essential amino acids lysine and threonine). Rice and oats, with a low prolamin content have a high protein utilization by rats, whereas corn and sorghum, with a high prolamin content provides only a low protein uptake [14]. Barley and wheat take an intermediate position. Amino acid compositions of cereals are shown in Table 3.

More simply albumins and globulins (and some enzymes) are referred to as soluble proteins, and gliadins and glutenins are insoluble gluten proteins. Also, wheat flour proteins often have been classified into mainly two size groups, either polymeric proteins or monomeric proteins. This size-based classification is now more extensively used for interpreting the relation of functional properties to composition. Polymeric proteins include glutenins and high-molecular-weight (HMW) albumins and globulins, whereas monomeric proteins are gliadins and low-molecular-weight (LMW) albumins and globulins.

Protein is not only a nutritional source but also has functional properties in the end products. Among the cereal proteins, wheat protein is unique because it gives viscoelastic properties to mixed dough and finally gives distinct characteristics to various end products (→ Bread and Other Baked Products, Chap. 5.1.2). Barley and rye have similar physical dough properties but to a lesser extent. The major parts of wheat flour to contribute to the rheological properties of doughs are the *gluten proteins*. All gluten proteins have a relatively similar amino acid composition, but with a broad and continuous molecular weight distribution from about 30 000 to 20×10^6. Gluten proteins are usually separated into two groups, gliadins and glutenins, based on their different solubility properties.

Gliadins are a highly heterogeneous mixture of proteins that are soluble in 70 % ethanol. Gliadins contain mainly single polypeptide chains (monomers) that are stabilized by hydrogen bonding and hydrophobic interactions, and intramolecular disulfide bonds [15] with molecular masses of 30 000 – 70 000.

Glutenin proteins account for 40 – 50 % of the total protein present in wheat flour. Glutenins are insoluble in a salt solution or neutral 70 % ethanol, but soluble in diluted acid or alkali. Molecular masses of glutenin were reported ranging from 100 000 to 20×10^6.

Even though there are somewhat controversial reports on the role of gliadins in breadmaking, it is generally accepted that both fractions seemed to be essential for dough plasticity (gliadins) and for dough strength and elasticity (glutenins) necessary for good dough handling, fermenting, and bread properties.

3.2.3. Pentosans

Nonstarch polysaccharides are the main constituents of cell wall material of plants. These polysaccharides are not digestible by humans and are referred to as dietary fiber. Nonstarch polysaccharides can be divided into three groups: cellulose, β-glucans, and pentosans. Pentosans, or alternatively hemicelluloses, are made up of pentoses (monosaccharides with five carbon atoms), mostly D-xylose and L-arabinose. Wheat kernels contain about 7 % pentosans whereas wheat flour has 1.5 – 2.5 % pentosans, of which 20 – 30 % is water soluble. The pentosan content of rye (9 % or higher) and thus of rye flours is higher than that of wheat. Pentosans affect the rheological properties of dough because of their high molecular mass and gel-forming ability. Pentosans

Table 3. Amino acid composition of 100 g of edible portion

Amino acid	Rice, brown	Wheat, HRW*	Corn, yellow	Sorghum	Millet, raw	Barley	Oats	Rye	Buck-wheat
Tryptophan	0.096	0.160	0.067	0.124	0.119	0.208	0.234	0.154	0.192
Threonine	0.275	0.365	0.354	0.346	0.353	0.424	0.575	0.532	0.506
Isoleucine	0.318	0.458	0.337	0.433	0.465	0.456	0.694	0.549	0.498
Leucine	0.620	0.854	1.155	1.491	1.400	0.848	1.284	0.980	0.832
Lysine	0.286	0.335	0.265	0.229	0.212	0.465	0.701	0.605	0.672
Methionine	0.169	0.201	0.197	0.169	0.221	0.240	0.312	0.248	0.172
Cystine	0.091	0.322	0.170	0.127	0.212	0.276	0.408	0.329	0.229
Phenylalanine	0.387	0.592	0.463	0.546	0.580	0.700	0.895	0.674	0.520
Tyrosine	0.281	0.387	0.383	0.321	0.340	0.358	0.573	0.339	0.241
Valine	0.440	0.556	0.477	0.561	0.578	0.612	0.937	0.747	0.678
Arginine	0.569	0.595	0.470	0.355	0.382	0.625	1.192	0.813	0.982
Histidine	0.190	0.285	0.287	0.246	0.236	0.281	0.405	0.367	0.309
Alanine	0.437	0.450	0.705	1.033	0.986	0.486	0.881	0.711	0.748
Aspartic acid	0.702	0.640	0.655	0.742	0.726	0.779	1.448	1.177	1.133
Glutamic acid	1.528	3.998	1.768	2.439	2.396	3.261	3.712	3.661	2.046
Glycine	0.369	0.528	0.386	0.346	0.287	0.452	0.841	0.701	1.031
Proline	0.352	1.289	0.822	0.852	0.877	1.484	0.934	1.491	0.507
Serine	0.338	0.586	0.447	0.462	0.644	0.527	0.750	0.681	0.685
Total protein, %	7.50	12.61	9.42	11.30	11.02	12.48	16.89	14.76	13.25

* Hard red winter.
Source: U.S. Department of Agriculture, Agricultural Research Service. 2004. USDA National Nutrient Database for Standard Reference, Release 17, Nutrient Data Laboratory.

have a high water-holding capacity because of their highly branched structure. The hydroxyl groups of pentosans interact with water and absorb about ten times their weight. Flour reconstituted with gluten, water solubles, and tailings starch, which has a higher content of water insoluble pentosans, shows drastic differences in water absorption, mixing behavior, and final baked product texture compared with flours reconstituted with prime starch. The addition of water-insoluble pentosans during dough mixing increases water absorption more than added water-soluble pentosans. Pentosans also have gelling capacity. Highly viscous solutions can be obtained when water-soluble pentosans dissolved in water (1 %) are treated with oxidizing agents.

3.2.4. Lipids

Lipids are relatively minor constituents in cereal grains. Lipid content and composition depend largely on extraction and purification procedures (extractant, extraction time and temperature, apparatus, the solvent to solute, purifying methods, etc.) and to a lesser extent on the condition of samples (particle size, moisture content, varietal differences, growing conditions, etc.). Moreover, lipid contents are significantly different depending on the anatomical parts of cereal. Germs are the richest source of lipids among all cereal grain fractions, even though they are a relatively small part of grain kernels. In corn kernels, 73 – 85 % of lipids are distributed in the germ fractions, whereas in rye, triticale, and wheat kernel 34 – 42 % of the lipids are in the germ fraction. By simple definition, lipids are materials soluble in organic solvents, but they are classified into many subcategories depending on the following: (1) location of lipids (nonstarch lipids, starch surface lipid, and starch lipids; (2) extraction and solubility (free lipids, bound lipid, total nonstarch lipids, and hydrolyzate lipids; and (3) biochemical structure (nonpolar or neutral lipids: triglyceride, diglyceride, monoglyceride, free fatty acid, and sterol esters, and polar lipids: glycolipids and phospholipids).

Wheat flour has a total lipid content of 2.5 – 3.0 % flour weight. There are two main fractions: about one-third of total lipids are starch lipids present inside the starch granules with mostly lyso-phospholipids; two-thirds are nonstarch lipids consisting of free lipids (0.8 – 1.0 % flour weight), and bound lipids (0.6 – 1.0 %). The free lipids are easily extractable from defatted flour with nonpolar solvents such as pe-

troleum ether or hexane, and bound lipids are extractable with polar solvents such as an aqueous mixture of alcohol at ambient temperature. The starch lipids are only extractable with polar solvent at an elevated temperature.

Two-thirds of free lipids are nonpolar lipids. The remaining free lipids are polar lipids of which two-thirds are glycolipids and the remaining one-third is phospholipids. Two-thirds of bound lipids are polar lipids, which consist of 45 – 50 % glycolipids and 50 – 55 % phospholipids, while one-third is nonpolar lipids [16].

Wheat flour free lipids, a minor component, have a significant effect on breadmaking [17]. When reconstituting defatted flours with extracted lipids to their original levels, polar lipids fractions of free lipids completely restored loaf volume and crumb grain but not nonpolar lipids [18]. Among polar lipids fractions, glycolipids improved loaf volume better than phospholipids, and among glycolipids, digalactosyl diglycerides was the best [16]. Polar lipids in flour have benefits for dough handling and baking [19, 20]. Significant effects of free lipids on breadmaking parameters among wheat varieties were shown by statistical analyses [21]. A significant correlation of loaf volume with polar lipids and lipid carbohydrate contents, and the nonpolar lipids/polar lipids ratios of free lipids from hard red winter wheats and flours was reported.

3.2.5. Secondary Constituents

3.2.5.1. Cellulose (→ Cellulose)

Cellulose is the main component of plant cell walls and the basic building block for foods, textiles, and paper. It constitutes 20 – 50 % in dry matter of vegetables and cereals. Cellulose is a straight-chain polysaccharide linked by β-D-1,4-glucosidic bond. The length of the chain varies greatly from a few hundred glucose units in wood pulp to more than 6000 for cotton. The hard stable crystalline regions of the chain give the plant its structural stability and strength, whereas noncrystalline (amorphous) regions, which can readily absorb water, make the plant flexible and less brittle. Cellulose is the most abundant form of biomass because it is found in large amounts in nearly all plants. However, it is not digestible by humans because they lack the enzymes necessary to cleave the bonds between the sugars in cellulose. Because of these properties, crystallite cellulose is used in foods as dietary fiber. Also, some cellulose derivatives such as sodium carboxymethyl cellulose (CMC), which is an anionic, linear, and water-soluble polymer, are added to some foods to control the rheological properties, foam stabilization, ice crystal formation and growth, and water-binding capacity.

3.2.5.2. β-Glucan

Most cereals have β-glucans in the cell wall. Barley and oats contain higher amount (3 – 7 %) of β-glucans followed by rye (2 %), wheat (1 %), and rice (< 1 %). The structure is, like that of other polysaccharides, a linear unbranched polymer of β-(1,3), (1,4)-D-glucopyranose units. β-Glucans form cylindrical molecules containing up to 250 000 glucose residues and have several distinct properties because of their unique molecular structure. High-molecular-weight β-glucans are viscous because of their constant cooperative associations, whereas low-molecular-weight β-glucans can form soft gels because chains can easily be rearranged to maximum linkages. The viscosity of a 1 % solution of barley β-glucan is 20 – 30 times higher than that of water at 20 °C, providing texturizing effects like fat substitutes. Health benefits have been reported; β-glucans, as a dietary fiber, contribute to reducing the glycemic response and cholesterol level in hypercholesterolemic individuals [22].

3.2.5.3. Lignins (→ Lignin)

Lignin is a constituent of plant cell walls and the most abundant biosource next to cellulose. Lignin is formed by the irreversible removal of water from sugars to create aromatic structures through the phenylpropanoid pathway. The large lignin molecules are three-dimensional and highly cross-linked networks built by *p*-hydroxyphenyl, guaiacyl, and syringyl units [23]. Sometimes lignin is present as a brown powder, but more often it exists as a gummy mixture of lignins with a wide range of molecular masses. Lignin is naturally resistant to attack by

most microorganisms and to chemical degradation as well. The strength of wood is a result of lignin which makes up to one-third of the mass of dry wood. Studies have shown that lignin provides health benefits, for example, by binding bile acids and cholesterol, inhibiting colonic carcinogenesis, and enhancing absorption of vitamin A.

3.2.5.4. Lignans

In the late 1990s the importance of some specific secondary plant substances for human health was discovered and discussed in a number of publications [22, 24 – 28]. These substances were claimed to be effective in prevention of carcinomas, in lowering the cholesterol and glucose levels in blood, in initiating the immunosystems and thus contributing to the health of the consumers. Out of about ten of these substances, such as polyphenols, phytic acid, protease inhibitors and phytoestrogens are present in cereal grains in significant amounts. Particularly isoflavonoids (isoflavone) and lignans (secoisolariciresinol and matairesinol) proved to be very effective in prevention and control of carcinomas at any stage of development of the disease. Rye contains significant amounts of lignans — since the aleurone and pericarp are rich in these substances, it is recommended to consume whole grain-meal bakery items or by bran enriched bakery items [26]. Despite the fact that the amounts of these healthy substances in rye kernel are not satisfactorily high enough, the whole kernel bread is a valuable supplement to the healthy food from vegetables having a high content on lignans.

3.2.5.5. Vitamins

Two types of vitamins exist in cereal grains: fat-soluble vitamins such as A, D, E, and K and water-soluble vitamins such as vitamin B, including thiamine, riboflavin, niacin, and pyridoxine (vitamin B_6) [29]. Generally, cereals have low levels of fat-soluble vitamins because the fat content in the kernel is low. Most lipids exist in the germ; corn is an exception with high levels of vitamin A and E. Corn and wheat germ oils have high level of vitamin E (50 – 300 mg/100 g), whereas whole wheat flour contains a small amount (0.5 – 5 mg/100 g). Cereal products are good sources of vitamin B including thiamine, riboflavin, and niacin. Wheat flour is an even better source of B vitamins because flours are usually enriched with those vitamins. Wheat bran is a good source of folate (60 – 80 mg/100 g), and barley, corn, oats, and wheat are considered moderate sources (20 – 60 mg/100 g). Cereals, however, contain no vitamin C. Whole cereal products are recommended to be consumed rather than milled products (no germ and bran), because vitamins and other good nutrients are generally concentrated in the germ and bran.

3.2.5.6. Mineral Matters and Phytate

Mineral matters comprise the inorganic compounds in cereal kernels, which after "burning" at high temperatures in an oven will be reduced to "ashes". This residue consists mostly of potassium and phosphorus, with small amounts of magnesium, calcium, sodium, iron, and trace elements. The contents of mineral matters or ash content vary according to cereal variety and kernel size. Maize and sorghum have the lowest (about 1.5 %), and unhulled barley, oats, and rice the highest (2.6 – 4.7 %) ash contents [29]. The bread cereals, wheat and rye, as well as dehulled barley are intermediate (1.8 – 2.2 %). Small kernels are richer in mineral matters than the bigger ones filled with starch. In addition, the distribution of mineral matter within a kernel is not uniform. The lowest ash content is inside the endosperm (0.4 %); ash content increases toward the outer parts of the kernel, with the aleurone layer having the highest content of mineral matter (8 – 10 %). The nonuniform distribution of mineral matters in kernels means that flour extraction when milling wheat and rye can be estimated, and the flour types based on ash content are established. Again, the whole kernel food is a significant supplier of mineral matters in human nourishment. Phosphorus as phytate is important due to specific properties among the mineral matters in cereal kernels.

Phytate (myoinositol hexaphosphoric acid) is a naturally occurring component of plant fiber comprising 1 – 3 % of all nuts, cereals, legumes, and oil seeds. Phytate has important physiological functions in the kernel. It has antioxidant

properties during dormancy and acts as a storage site of phosphorus, cations, and cell wall precursors. Nutritional functions of phytate in foods are controversial. Phytate has six binding sites to immobilize minerals such as calcium, magnesium, iron, and zinc, keeping them from being absorbed in the human body. This arouses nutritional concerns of decreasing mineral bioavailability. However, it was debated that the health benefit of dietary phytate overweighs the potential risk of absorbing fewer minerals. Phytate could act as an antioxidant and anticancer agent [30]. Also, phytate may have beneficial effects on blood sugar level, like many dietary fibers [31]. Although some animals are able to digest and/or absorb phytate it is not known whether humans can.

3.2.5.7. Tannins

Tannins are oligomeric compounds with multiple structure units with free phenolic groups. They are generally soluble in water except for some high molecular weight structures and can bind proteins and form insoluble or soluble tannin-protein complexes. The molecular masses of tannins range from 500 to 20 000. Tannins exist widely in the plant kingdom. Of cereal grains, only sorghum and millet are a significant source of tannins. In addition, only a small percentage (1 %) of sorghum varieties produced in U.S. contain high tannin contents. Sorghum tannins are found in the pigmented testa, a layer between the outer integument and the aleurone layer. They are involved in maintaining plant dormancy and have allelopathic and bactericidal properties. Tannins act to protect plants against pathogens, herbivores, and hostile environmental conditions. When tannins are consumed, they induce instantaneous negative responses like astringency, bitterness, or unpleasant taste. Also, tannins have several negative effects including reducing feed intake and digestibility and toxicity to rumen microorganisms. As a result, animals fed with high tannins (over 6 %) in the diet experience depression in growth and production rate, and even death in poultry and some other animals. Some studies, however, suggest that low or moderate levels of tannins (less than 4 %) in forage legumes can have beneficial effects in ruminants, resulting in higher growth rates and milk yield [32]. These effects vary depending on the content and type of tannin ingested and on the animal's tolerance. There is some evidence that low levels of tannins may be beneficial to humans. Tannins in brown sorghum bran can be beneficial to health because they have higher antioxidant activity (2000 – 3000 µmol TE/g, ORAC, Oxygen Radical Absorbance Capacity) even compared with plums (500 µmol TE/g) and blueberries (800 µmol TE/g). Tannins are stable to heat and form complexes with many molecules such as carbohydrates, proteins, polysaccharides, bacterial cell membranes, and enzymes involved in digestion.

4. Standards in Grain Trade

Grain buyers and sellers have used various systems over the years to come to an agreement on the price to be paid for grain on the basis of its quality. However, the criteria of wheat quality, for example, are as varied as its uses and have resulted in the need for inspectors to judge wheat quality to minimize disputes between marketers. Many countries use grading systems in which specifications based on both objective and subjective evaluations define the grade. There are many different standards, depending on country and the specific type of cereal grain. Tables 4 and 5 show grain grade standards in the USA and Europe.

4.1. Test Weight (TW)

Test weight is the first and most widely used (except for rice) simple determination to grade the quality of cereal grains for trading and milling. In the United States, the unit of test weight is pound per Winchester bushel (2150.42 in^3), whereas the Imperial bushel (2219.36 in^3) is used in Canada. Many other countries use kilograms per hectoliter as unit. Test weight is affected by kernel density and packing characteristics. Packing characteristics are influenced by kernel characteristics (shape, uniformity of shape and size, moisture content, and surface condition) and foreign materials. Environment and genetics influence the test weight. Also, frequent handling and moving increases the test

weight by polishing the bran coat. Some studies suggested that test weight might be useful to predict flour yield.

4.2. 1000 Kernel Weight (TKW)

Kernel weight is generally expressed as grams per 1000 kernels. Grains with bigger and denser kernels give heavier weights so that kernel weight is generally considered as a better gauge of the physical condition of grains than test weight. Kernel size used to be measured according to average cross-sectional area showing that kernel size is highly correlated to milling yield. Today, kernel weight and size (diameter) are easily determined using a Single Kernel Characterization System (SKCS). Kernel weight is influenced by both environment and heritable characteristics. Also, kernel weight is significantly affected by the rate of grain filling during maturation.

4.3. Besatz

Besatz is defined by the International Association for Cereal Science and Technology (ICC) as all the components of a grain sample which differ from the normal basic variety. ICC subdivided Besatz into four categories: grain dockage (broken grains, shriveled grains, other grains or grains of contrasting varieties, sprouted grains, insect-damaged grains, frost-damaged grains, and grains with discolored germs); black dockage (toxic or nontoxic weed seeds, ergot, unsound grains, smutty grains, and impurities and chaff); total Besatz (the sum of grain dockage and black dockage); and animal impurities or impurities of animal origin and insect fragments [33].

4.4. Dockage

According to the official U.S. Grain Standards, dockage is defined as all matter other than grains that can be removed from the original sample by use of an approved device (Carter Dockage Tester) in accordance with procedures prescribed in Federal Grain Inspection Service (FGIS) instructions as well as underdeveloped, shriveled, and small pieces of wheat kernels removed in properly separating the material other than grain and that cannot be recovered by properly rescreening or recleaning. The main difference between Besatz and dockage is that Besatz includes unsound grains.

4.5. Foreign Material

Foreign material is all matter other than grains that remains in the sample after the removal of dockage and shrunken and broken kernels.

4.6. Sprout Damage

Sprout damage can occur on wheat when there is rain before harvest. Even though there is no visible sprout damage on kernels, such kernels usually have highly elevated α-amylase activity. Baking characteristics of flours milled from sprout damaged wheat are poor. Wheats grown in North America generally have fewer problems than those grown in Europe where weather is often wet during harvest. The α-amylase activity is conveniently determined by Hagberg – Perten falling number and Brabender Amylograph procedures or newer method with a Newport Scientific Rapid-Visco-Analyser [33, 34].

4.7. Chemical Characteristics that Determine Wheat Quality

Moisture and protein contents are important factors in trading wheat and other grains. Even though *moisture content* is no longer a determinant of wheat grade in the U.S. Standards, it should be reported on all samples for two reasons. Firstly, it is directly related with money because higher water content means less dry matter. Secondly, excessive moisture content in grain prompts deterioration by molds, insects, and self-heating during storage and transportation. Other conditions also contribute the deterioration of grain, but moisture content of many grains should be less than 14 % for extended storage at 18 °C and less than 13 % at 25 °C or higher. The Australian Wheat Board Receival Standard allows maximum moisture content of 12 %.

Table 4. Official U.S. grade and grade requirement for wheat

Grading factors	Grades U.S. Nos. 1	2	3	4	5
	Minimum pound limits of:				
Test weight per bushel					
Hard Red Spring wheat or White Chib wheat	58.0	57.0	55.0	53.0	50.0
All other classes and subclasses	60.0	58.0	56.0	54.0	51.0
	Maximum percent limits of:				
Defects:					
Damaged kernels					
Heat (part of total)	0.2	0.2	0.5	1.0	3.0
Total	2.0	4.0	7.0	10.0	15.0
Foreign material	0.4	0.7	1.3	3.0	5.0
Shrunken and broken kernels	3.0	5.0	8.0	12.0	20.0
Total[1]	3.0	5.0	8.0	12.0	20.0
Wheat of other classes:[2]					
Contrasting classes	1.0	2.0	3.0	10.0	10.0
Total[3]	3.0	5.0	10.0	10.0	10.0
Stones	0.1	0.1	0.1	0.1	0.1
	Maximum count limits of:				
Other material:					
Annual filth	1	1	1	1	1
Castor beans	1	1	1	1	1
Crotalana seeds	2	2	2	2	2
Glass	0	0	0	0	0
Stones	3	3	3	3	3
Unknown foreign substances	3	3	3	3	3
Total[4]	4	4	4	4	4
Insect-damaged kernels in 100 grams	31	31	31	31	31

U.S. Sample grade is Wheat that:
(a) Does not meet the requirements for U.S. Nos. 1, 2, 3, 4, or 5; or
(b) Has a musty, sour, or commercially objectionable foreign odor (except smut or garlic odor) or
(c) Is heating or of distinctly low quality.

[1] Includes damaged kernels (total), foreign material, shrunken and broken kernels.
[2] Unclassed wheat of any grade any contain not more than 10.0 percent of wheat of other classes.
[3] Includes contrasting classes.
[4] Includes any combination of animal filth, castor beans, crotalaria seeds, glass, or stones, or unknown foreign substance.

Protein content is also important in wheat trading. Wheat is usually blended from different varieties and lots before trade to tailor the buyer's request. As a result, wheat with high protein content is in demand for delivering high-protein composite wheat (11 – 13 %) that is suitable for producing pan-bread. Protein quality is another factor that affects the final properties of end-use products. For wheat trading, however, it is not listed as wheat grade determinant because of the difficulties of determining protein quality for specific purposes.

5. Uses

5.1. Human Food and Animal Feed

5.1.1. Wheat

Wheat, a major world cereal grain, is consumed mainly by humans in nearly all countries around the world, irrespective of wheat-exporting or wheat-importing. The type of food use depends on kernel hardness (the texture of the endosperm) which affects milling process quality and protein content/quality, which determines suitability for various products. To a lesser

Table 5. European grain standards [35]

	Durum wheat	Common wheat	Rye	Barley	Maize	Sorghum
A. Maximum moisture content	14.5 %	14.5 %	14.5 %	14.5 %	14.5 %	14.5 %
B. Maximum percentage of matter which is not basic cereal of unimpaired quality:	12 %	12 %	12 %	12 %	12 %	12 %
1. Broken grains	6 %	5 %	5 %	5 %	10 %	10 %
2. Impurities consisting of grains (other than indicated at 3) of which:	5 %	7 %	5 %	12 %	5 %	5 %
(a) shrivelled grains					–	–
(b) other cereals	3 %					
(c) grains damaged by pests				5 %		
(d) grains in which the germ is discoloured			–	–	–	–
(e) grains overheated during drying	0.50 %	0.50 %	1.5 %	3 %	3 %	3 %
3. Mottled grains and/or grains affected with fusariosis of which:	5 %	–	–	–	–	–
– grains affected with fusariosis	1.5 %	–	–	–	–	–
4. Sprouted grains	4 %	4 %	4 %	6 %	6 %	6 %
5. Miscellaneous impurities (Schwarzbesatz) of which:	3 %	3 %	3 %	3 %	3 %	3 %
(a) extraneous seeds						
– noxious	0.10 %	0.10 %	0.10 %	0.10 %	0.10 %	0.10 %
– other						
(b) damaged grains:						
– grains damaged by spontaneous heating or too extreme heating during drying	0.05 %	0.05 %				
– other						
(c) extraneous matter						
(d) husks						
(e) ergot	0.05 %	0.05 %	0.05 %	–	–	–
(f) decayed grains			–	–	–	–
(g) dead insects and fragments of insects						
C. Maximum percentage of wholly or partially piebald grains	27 %	–	–	–	–	–
D. Maximum tannin content[1]	–	–	–	–	–	1 %
E. Minimum specific weight (kg/hL)	78	73	70	62	–	–
F. Minimum protein content[1]:						
– 2000/2001 marketing year	11.5 %	10 %	–	–	–	–
– 2001/2002 marketing year	11.5 %	10.3 %	–	–	–	–
– 2002/2003 onwards	11.5 %	10.5 %				
G. Hagberg falling number (s)	220	220	120			
H. Minimum Zeleny index (mL)	–	22	–	–	–	–

[1] As % of dry matter.

extent, types of food depend on the starch, lipid, or other minor constituent composition and content.

The average annual per capita consumption of wheat from 1997 to 1999 was 70.8 kg globally, ranging from 15.9 kg in Sub-Saharan Africa and 16.3 kg in Bangladesh to 131.7 kg in Russia and 97.6 kg in Western Europe (Table 6). The U.S. total annual food usage, excluding alcoholic beverage production, was 25.2×10^6 t of which 97.4 % was consumed as flour products, including white, whole wheat, and semolina flour, with 2.6 % for cereals.

Nearly all foods derived from wheat use milled flours or whole wheat ground meals of "common wheat" for bread, biscuits, and other food production, and semolina and farina milled from durum wheat for pasta production. A small portion of wheat grain is directly used for manufacturing ready-to-eat breakfast cereals, such as flaked, puffed, or shredded wheat.

Industrial applications of wheat are in the manufacture of malt, potable spirits, starch,

Table 6. Annual per-capita consumption (kg per person) of cereal grains, regional and selected countries (1997 – 1999 average)

Area	Wheat	Milled rice	Corn	Sorghum	Millet	Barley	Rye
North and Central America	70.9	10.8	40.1	1.2	0	0.4	0.2
United States	86.8	8.6	13.8	1.2	0	0.5	0.3
Central America	37.1	9.4	112.1	1.8	0	0.1	0
South America	55.5	31.8	21.8	0	0	0.5	0
Brazil	47.4	39.5	18.0	0	0	0	0
Western Europe	97.6	4.8	5.8	0	0	0.6	3.9
Russia	131.7	4.9	0.3	0	2.9	0.9	9.2
Africa	46.3	17.8	41.4	19.5	12.9	4.1	0
Sub-Saharan Africa	15.9	17.5	38.9	24.9	16.9	1.5	0.1
Asia	69.9	86.4	13.9	2.8	3	0.9	0.1
China	82.6	91.6	19.7	1.1	0.8	0.6	0.2
India	57.3	75.8	8.8	8	9.1	1.2	0
Indonesia	16.3	151.0	34.4	0	0	0	0
Bangladesh	19.0	161.1	0.3	0	0.4	0	0
Oceania	66.9	15.2	3.4	0.6	0	0.4	0.6
World average	70.8	57.8	19.0	4.3	3.5	1.3	1.1

[36]: Source from FAO (2001).

gluten, pastes, and core binders. In the brewing and distilling industries only small amounts of wheat malt are employed. More malt is used by the flour milling industries to increase the α-amylase activity of the different types of flours. A certain amount of wheat flour is utilized to manufacture wheat starch as a byproduct of vital gluten (its functional properties make it preferable for breadmaking). Gluten supplements flour proteins for specialty baked goods. Flour is also a raw material for manufacturing monosodium glutamate, which is used to flavor food. Some low-grade flours are used to make pastes for book-binding or paper hanging as well as to make core-binders for preparing molds for iron castings. Especially in Australia, starch is a byproduct of wheat gluten manufacture. High yields of wheat in Western Europe make producing starch and gluten attractive because both products can be marketed economically.

Flours from *hard* "*common*" *wheats* are used for bread (leavened or unleavened) including dinner rolls, hamburger buns, hot dog rolls, bagels, pretzels, pizza crust, hearth breads, English muffins, croissants, sweet yeast goods, doughnuts, various types of fermented Arabic flat breads or unleavened chapatti, chemically leavened *Barbari* bread, and Swedish crisp bread or Chinese steamed bread, noodles (Asian alkaline types), bulgur, and ingredients for wide range of foods. Flours from soft "common" wheats are used for groups of products including biscuits, cookies (rotary molded, wire-cut, cutting-machine, soft, and trolley goods), crackers, wafers, and pretzels; cakes of all types and sizes from sponges and high-ratio cakes to heavily fruited cakes and from cup cakes to wedding cakes, including the wide range of prepared cake mixes; pastry products, from piecrust, puff pastries to sweet Danish pastries; waffles, pancakes, and doughnuts, and refrigerated/frozen doughs for a chemically leavened breakfast biscuit; and Asian noodles (Japanese "Udon" type). Soft wheat flour is also used as a thickener for soups and soup mixes, in the manufacture of crumbs for coating fish and meat products, and also as the basis for ready-to-eat breakfast cereals (see also → Bread and Other Baked Products).

Durum wheat is the raw material of choice for the production of pasta products, macaroni, spaghetti, vermicelli, and noodles that are manufactured primarily from semolina, durum granulars, and flour produced from the milling of durum wheat. The 10-year (1976–1985) average world durum wheat production was 24.24 $\times 10^6$ t: 28 % of world production originated from Near East Asia, 23.4 % from North America, 19.2 % from Western Europe, 12.1 % from North Africa, 1.1 % from South America, and 16.2 % from other regions. Durum wheat products are (a) pasta products, including pasta and

couscous, and (b) nonpasta products, including *bulgur* and *frekeh*. The use of pasta dates back many centuries. Although such products as pasta were made popular by the Italians, most certainly they were a part of earlier cultures in China and later cultures in many other countries.

Animal Feeds. In animal feeds relatively less wheat than other coarse grains are used. The use of wheat for animal feeds depends on the quality of wheat, whether it is unsuitable for making foods, or on the cultural practices of regions and countries. The USA and the UK consume about 6×10^6 t each for animal feed, representing only 10 % and 25 % of total consumption, respectively. The UK produces approximately 15×10^6 t of wheat each year, and 25 % is exported. About 40 % of the produced wheat is used in animal feed rations going to chickens, cows and pigs with the remaining 35 % consumed by humans with thousands of products and the daily production of 10 million loaves of bread.

5.1.2. Rice

Rice is a primarily food crop, like wheat, with an average global annual production (1999 – 2001) of around 602×10^6 t on a rough (paddy) basis, approaching maize production and exceeding wheat production (Table 1). Most rice is used in its producing countries with relatively little international trade (5 % of total world production). The most important rice exporter is Thailand ($> 6 \times 10^6$ t milled rice export in 2000), followed by the USA ,with an average annual amount of 4.4×10^6 t of rough rice or 3.1×10^6 t of milled rice, and Vietnam ($> 3 \times 10^6$ t milled rice in 2000) (Table 6) [36]. Rice is a dietary staple for more than half of the world's population. The average annual per capita consumption of milled rice globally was 57.8 kg per person, ranging from 4.8 kg in Western Europe and 4.9 kg per person in Russia to 161.1 kg in Bangladesh and 151.0 kg per person in Indonesia (Table 6). In the USA, 97 % of domestic use is for foods and industry, and 3 % seed use, and almost no feed use. The U.S. annual per capita consumption during 1997 – 1999 was 8.6 kg of milled rice (Table 6), which is small compared to total cereal grain consumption in the USA. However, the world average per capita consumption of total cereal grains shows that milled rice constitutes 36.3 % of grain consumption, with a range of 3.2 % in Russia to 89.1 % of total cereal grains consumed in Bangladesh.

Unlike other cereals, rice is consumed as a whole grain, after the bran layers are removed. For human consumption, rough (paddy) rice is de-hulled and milled to remove the bran layers, leaving the lustrous, white grain. Brown rice retains the bran layers, which contain many nutrients. Short-grain varieties (*indica* types) become usually sticky on cooking and are very well preferred in many parts of northern Asia. However, most western countries prefer the drier, flaky quality provided by the long-grain varieties (*japonica* type). Milled rice is consumed usually as a cooked grain, although there are many food products made from rice, including parboiled rice, rice crackers, rice noodles, rice cakes and snack foods, rice flour, and fermented drinks, i.e., sake and rice wine.

Animal Feeds. Rice is used throughout the world as a food commodity and, with the exception of rice bran and germ, is rarely used in animal feeding. In countries where more rice is grown and human food uses require the entire grain production, ruminants are fed rice straw [36]. In the USA, the largest category for rice use in the processed foods is the pet food industry, which used 3.13×10^5 t in 1999 – 2000, with sales of $\$11.8 \times 10^9$ in 2000, a 6 % increase from the year before [37]. Broken or brewers' rice is most commonly used in pet food because of the lower cost, but flour, bran, and whole long- or medium-grain rice are also commonly used.

5.1.3. Maize (Corn)

Maize, also known as corn in the USA, and indigenous to the Americas, is produced worldwide (Table 1) with the highest world average yield (43.63 dt/ha, Table 2). Maize is consumed in various forms after different processing methods in many different countries and cultural regions (Table 6). Most maize consumed worldwide is fresh or immature (green) whole grain maize on the cob, popmaize (popcorn), and maize bread. Unfermented breads like tortillas and arepas are the most popular maize food in Central America, Venezuela, and Colombia. In

Mexico, the average annual per-capita consumption of maize in the form of tortillas (nixtamalized, i.e., lime-cooked) was 120 kg per person. Porridge is another popular form of maize food that is prepared in many ways in Central and South America, Africa, Southwestern USA, and Southern Europe. Maize is also used for making alcoholic beverages in Central and South America and Africa.

Maize is processed by three industrial methods: wet-milling, dry-milling, and nixtamalization. Wet-milling (Section 6.3) is mainly used by industries to separate out the corn components efficiently to obtain valuable coproducts such as starch, protein, fiber, and germ. More than 90 % of commercially available starch is from maize, and most starch is further processed into maize syrups by acid hydrolysis or enzymatic conversion (→ Glucose and Glucose-Containing Syrups). Maize syrups (sweeteners) are extensively used by the soft drink industry because of their superior characteristics such as high sweetness, ease of use, and functionality with other ingredients. Oil mills produce maize oil from germ, which is widely used as cooking oil because it is very digestible and provides high levels of essential fatty acids, vitamin E, and polyunsaturated fatty acids. Dry-milling produces endosperm fractions (grits, meal, and flour), germ, and dietary fiber. Products from dry-milling are mainly used for snacks, breakfast cereals, and alcohol production. Nixtamalization is the traditional Mexican method for processing maize, produces tortilla, tortilla chips, maize chips, and other snacks, which are gaining popularity also in other parts of the world. Production of fuel alcohol consumed the most maize (1168×10^6 bushels, 46.4 %), and high-fructose maize syrup was second (530×10^6 bushels, 21.1 %). Starch, glucose, cereals and other products, and beverage alcohol were produced from 271 (10.8 %), 228 (9.1 %), 187 (7.4 %), and 130 (5.2 %) $\times 10^6$ bushels of maize, respectively.

Animal Feeds. Maize refiners produce four major feed products: condensed fermented maize extractives (steep liquor), gluten feed, gluten meal, and maize germ meal. During the wet-milling process, steep liquor comes out of the steeps. Steep liquor contains 5 – 10 % solids consisting of about 44 % protein and other micronutrients [38]. Steep liquor is used as nutritious binder mixed with maize fiber. The mixture of steep liquor and maize fiber, called gluten feed, is processed into pellets to increase density by 25 – 30 %, thus improving handling costs, and fed to cattle. Dried maize gluten meal supplies energy, vitamins, and minerals to poultry. It is important for poultry feeding that maize gluten meal should keep natural golden pigmentation (xanthophyll) because it provides desirable yellow egg yolk color and carcass skin. Maize germ meal is the remaining residue after oil is extracted. Germ meal has a high content of protein (about 24 %, dry basis), vitamins, and minerals, so that it is often combined with gluten feed.

5.1.4. Other Cereals

Sorghum. The major sorghum producing centers are in Africa and Asia (54.4 % of total production), sorghum is the third major cereal used in the USA. Most (87 %) is used as feed. Recently, sorghum has gained more interest because of its unique agricultural and nutritional characteristics [39]. Sorghum is drought-resistant and heat-tolerant, so it can often be grown in areas where other cereals cannot. This may become more important as the world population continues to rise, and water shortages occur. Recent research has discovered many new health benefits of sorghum, due to its high antioxidant levels, related to its content of polyphenolic compounds and sorghum wax, containing polyicosanols, which may be important in cardiac health. In addition, sorghum is often recommended as a safe food for celiac patients who are intolerant to gluten (cereal protein from wheat, barley, and rye) because it is more closely related to maize than to wheat, barley, and rye. Sorghum might, therefore, provide a good basis for gluten-free breads and other baked products such as pasta, cookies, and snacks.

Barley. Most barley produced worldwide is used for feed (75 %) and malt for alcoholic beverages (20 %) as well as human food. However, in the USA, barley is used more for human consumption, seed, or industry uses (64 %) than as feed (36 %) because the production of barley in the USA has sharply decreased from 375×10^6 bushel (1994) to 276×10^6 bushel (2003). Barley is used in many different products, such

Table 7. Nutrient composition of whole grain flour of wheat, rice, and maize and of their refined products (100 g flour and normal moisture content basis)

Cereal/ Product	Wheat flour, whole grain	Wheat flour, white, bread, unenriched	Wheat flour, white, bread, enriched	Rice flour, brown	Rice flour, white	Maize flour, whole grain, yellow	Maize flour, degermed, yellow
Energy, kcal	339	361	361	363	366	361	375
Protein, g	13.7	11.98	11.98	7.23	5.95	6.93	5.59
Fat, g	1.87	1.66	1.66	2.78	1.42	3.86	1.39
Carbohydrate, g	72.57	72.53	72.53	76.48	80.13	76.85	82.75
Dietary fiber, g	12.2	2.4	2.4	4.6	2.4	13.4	1.9
Thiamin, mg	0.447	0.08	0.812	0.443	0.138	0.246	0.074
Riboflavin, mg	0.215	0.06	0.512	0.080	0.021	0.080	0.058
Niacin, mg	6.365	1.00	7.554	6.340	2.590	1.900	2.656
Pantothenic acid, mg	1.008	0.438	0.438	1.591	0.819	0.658	0.052
Vitamin B6, mg	0.341	0.037	0.037	0.763	0.436	0.370	0.097
Folate, µg	44	33	183	16	4	25	48
Calcium, mg	34	15	15	11	10	7	2
Iron, mg	3.88	0.9	4.41	1.98	0.35	2.38	0.91
Magnesium, mg	138	25	25	112	35	93	18
Phosphorus, mg	346	97	97	337	98	272	60
Potassium, mg	405	100	100	289	76	315	90
Zinc, mg	2.93	0.85	0.85	2.45	0.8	1.73	0.37

Source: U.S. Department of Agriculture, Agricultural Research Service. 2004. USDA National Nutrient Database for Standard Reference, Release 17, Nutrient Data Laboratory.

as pearl barley, barley flour, flakes and grits, malt flour supplements, sweetener, malted milk, syrup, rice extenders, breakfast cereals, soups, pilaf mixes, breads, cookies, crackers and snack bars. Barley is known to have a high content of soluble fiber (β-glucan). β-Glucan is effective in lowering the blood cholesterol levels of chicken, rats, swine, and humans [22]. Barley is uniquely suited for malting as opposed to other grains. In brewing, barley malt provides hydrolytic enzymes and a fermentable substrate. Because the quality of the barley malt is critical in brewing, malting barley is subject to very strict quality requirements and traded as a specific cultivar (→ Beer, Chap. 2.1).

Oats. Most oats produced in the USA are used as feed (69 %), the rest is as food and in industry. As feed, oats are good for ruminants such as dairy and beef cattle and for sheep. Oats are also preferred feed for horses because of high palatability, digestibility, and nutritional value. The demand for oats for human consumption has increased lately because of their many health benefits. Scientific evidence suggests that oats can flatten the postprandial blood glucose and insulin rise, reduce LDL (bad) cholesterol, and reduce the risk of cancer because oats are rich in β-glucan and antioxidants (e.g., phenolic acids, avenanthramides, flavonoids, sterols, and phytic acid) [22]. Oats are a minor crop but have advantages nutritionally because they are usually eaten as whole grain, providing better nutrition than other cereals that are usually consumed after milling or refining which removes much of the fiber, vitamins, and minerals from cereals. Most common oats products are porridge or oatmeal, breakfast cereals, granola bars, and cookies. Industrial usages of oats are for adhesives, paper, coating agents for tablet, and cleansers (soap, shampoo, etc.).

5.2. Importance of Whole Cereals, Refined Products, and Fortification for Nutrition, Health, and Requirement

Cereals usually go through several milling processes before they are consumed. During this process, a significant part of the nutrients, such as proteins, vitamins, and minerals, are removed from the grain because many cereals tend to have valuable nutritive constituents in the outer layers (bran and aleurone layer) and the germ. Table 7 shows the nutrient composition of whole grain flours and of their refined products. Unenriched white flour contains far less dietary fiber, vitamins, and minerals than whole grain flour. Especially high losses occur for dietary fiber (80 % reduction), thiamine (82 %), niacin (84 %), vita-

min B6 (89 %), and magnesium (82 %). In enriched flour most vitamins are replaced, some even to higher levels, but dietary fiber and most minerals remain same as those of unenriched flour. White rice and degermed maize flour also have significantly lower nutritional values than brown rice and whole grain maize flour. Much information on the benefits of consuming whole grain has been published, in both epidemiological and clinical studies. Whole grain is an excellent source of high fiber, vitamins, minerals, disease-fighting phytochemicals, and antioxidants [22, 26, 29]. Many health benefits are associated with consuming whole grains: reducing the risks of cancer, heart disease (high cholesterol), stroke, high blood pressure, obesity, type II diabetes, and bowel disorders. In 2005, the U.S. Department Health and Human Services and the Department of Agriculture revised the "Dietary Guidelines for Americans" dramatically from the previous one [40]. One key to the newly added recommendations is whole grain. These guidelines recommend at least three servings (1 serving: 16 g of whole grain flour = 1 ounce or one slice of whole grain bread) of whole grains per day. However, Americans eat less than one serving of whole grains per day, and more than 30 % of Americans never eat whole grains. Other countries gave the similar situation. Whole grain consumption is generally low except in Scandinavian countries, where whole grain products are regularly consumed. Whole grain consumption in the UK was known as less than one serving per day. Considering all the benefits, increasing consumption of whole grain products and reducing consumption of refined flour products is highly recommended.

In 1942, the U.S. government passed the Enrichment Act, requiring the replacement of niacin, thiamine, iron, and riboflavin (more than is present in whole grain) in refined white flour. A deficiency of vitamin B (thiamine) causes beriberi (great weakness), which makes people suffer weakness in legs and loss of feeling in the feet, finally they die from heart failure. Iron deficiency, which is more common in people whose staple food is rice, causes anaemia. More recently, in 1998, the U.S. Food and Drug Administration mandated folic acid supplementation in white flour to reduce the risk of neural tube birth defect. Wheat flour fortification is now widely accepted in many countries. Other fortifications of cereal and cereal-based foods are performed voluntarily by food manufacturers. The most common nutrients used in fortification are vitamins A, B_1, B_2, niacin (B_3), pantothenic acid (B_5), B_6, B_{12}, C, E, and folic acid, and of the minerals calcium, copper, iodine, iron, magnesium, manganese, phosphorus, and zinc. Generally, fortification of cereals improves the nutritional profile of foods, resulting in health benefits to consumers, but government and/or manufacturers need to closely control fortification programs to prevent exposure of consumers to the risk of overintakes or imbalance of nutrients. Moreover, cereals are not a complete food or feed, so other foods, vegetables and proteins from animal or dairy products, are needed to be consumed to meet the requirement of balanced daily diet.

6. Processing

6.1. Dry-Milling Process

6.1.1. Wheat

The primary use of wheat in food production is flour (see Section 5.1). Some wheat is also used in breakfast cereals. Much of the flour is sold, not in its original form, but as blended and prepared flours for restaurants, cafeterias, schools, and all-purpose flours for private households.

Wheat and Flour Quality. In cereal technology, the term quality denotes the suitability of the material for a specific end use. The term flour quality has no relation to nutritional attributes. For example, high-protein wheat flour is a good quality for breadmaking. Low-protein wheat flour is good for chemically leavened products such as biscuits, cakes, and pastry. The miller wants wheat that gives a high yield of flour. Therefore, wheat kernels should be plump and uniformly large for easy and simple separation of foreign material. The wheat should produce a high yield of flour with maximum and clean separation from the bran and germ and without excessive consumption of power. Because the endosperm is denser than the bran and the germ, semihard wheat produces more flour. The production of bread flours requires that the difference in protein content between wheat and flour be no more than 1 – 1.5 %.

Gluten is particularly important for food production. Gluten imparts physical properties to dough made from wheat flour that differ from those of dough made from other cereal grains. Gluten is responsible for gas retention, which makes producing well-leavened baked goods possible. Usually, bread volume increases with an increase in gluten protein content [41].

In addition to gluten proteins, lipids are also essential for maximum performance. Bread baked from defatted flour or gluten-starch mixtures is consistently lower in volume than bread baked from normal flour with comparable protein content.

Bread of satisfactory quality cannot be baked without wheat starch, which is a main flour component (→ Bread and Other Baked Products, Chap. 5.1.1). Starches from wheat, rye, barley, and triticale have comparable granule sizes and gelatinization characteristics and can be used in bread making in combination with wheat gluten.

Bread flour should have adequate α-amylase activity for starch modification. Optimum α-amylase levels are maintained by careful wheat selection and by adding small amounts of malted wheat or malted barley flour or α-amylases. To improve the baking behavior, oxidizing agents such as potassium bromate (in the U.S.) or ascorbic acid (in Europe) are used. In addition to this flour treatment, consumers in North America want a relatively white crumb; therefore, bleaching agents are used to convert the carotenoide pigments of the flour to colorless compounds. Some flours are enriched with certain vitamins or minerals, particularly in the USA.

Dry-Milling of Wheat. Milling grain as food for humans has been traced back more than 8000 years. Milling started with hand mills, followed by horse-driven mills. Later, water- and wind-driven mills were developed. Today a sophisticated and automated milling industry exists [42 – 45]. In producing white flour, the objective is to separate the starchy endosperm of the grain from the bran and the germ. Then, the separated endosperm is pulverized. This separation of the starchy endosperm is possible because its physical properties differ from those of the fibrous pericarp and oily germ. The bran is tough because of its high fiber content whereas the starchy endosperm is friable. The germ has high oil content and is flaked between smooth rolls. The differences in toughness of the bran and the friability of the starch endosperm are enhanced by wheat conditioning, which involves adding water before milling and using different tempering times. Water toughens the bran and mellows the endosperm. Wheat flour production involves wheat selection and blending; cleaning and conditioning; breaking, sieving, and reduction; and special treatments (bleaching, enrichment, oxidizing agents, etc.).

Selection and Blending. To produce wheat flour of definite characteristics, wheat with different baking strengths must be separated and blended. In some mills, blending to a certain quality level is done before cleaning. In larger milling units, this mostly occurs after cleaning. A uniform mixture after blending must meet stringent specifications. The uniformity of the mixture is usually controlled by a near infrared (NIR) technique (analytical control).

Cleaning. The cleaning process is very important in milling because poor cleaning directly negatively influences the result [42, 44]. Wheat contains many damaged kernels (broken grains, shriveled grains, other cereals like rye), sprouted grains, grains with discolored germ, insect-damaged grains, frost-damaged grains, and heat-damaged grains. Also, weed seeds (toxic, nontoxic), ergot, unsound grains, smutty grains, impurities, and chaff must be removed [33].

Usually cleaning starts with a magnetic separator to remove any free metal pieces in the wheat stream (Fig. 2). A second separator removes Besatz fractions according to size and specific weight. Stones are removed in a dry stoner. Disk separators or trieur cylinders can separate wheat by kernel shape. An intensive surface cleaning is done by a scourer. This process removes spores and mycotoxins-polluted kernels from the wheat grains, as well as heavy metals such as lead.

Conditioning. In the cleaning process, water is added to the grain and allowed to stand for different times to ensure maximum toughening of the bran with optimum mellowing of the starchy endosperm. The tempering time depends on the hardness of the wheat grains. Hard wheats are tempered to 16 – 18 % moisture and soft wheats to 14 – 16 % moisture content. After conditioning, metal pieces are again separated, followed by an air classification to improve the cleaning effect. The cleaned wheat is stored in a mill bin

and weighed exactly before it is transported to the first break in the mill.

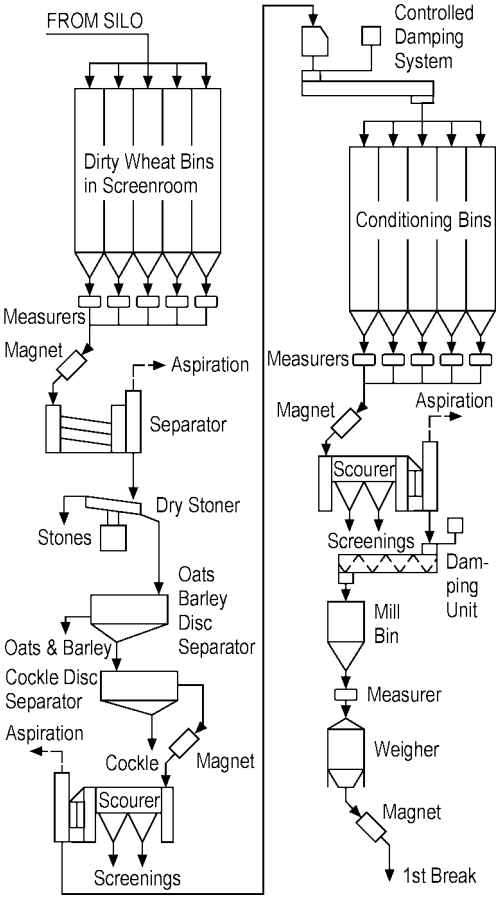

Figure 2. Flow sheet for a conventional cleaning house

Grinding. The grinding process is the most important step in the milling system. The first part of grinding is carried out on corrugated rolls (break rolls) (Fig. 3). The manner in which the kernel is broken by two pairs of rolls that rotate in opposite directions at different speeds affects the subsequent sifting and purifying operations, both in granulation (particle size distribution) and in the amount of fine bran present in the ground material, in the reduction and sizing systems. The grinding process is divided in four systems [42, 44]:

– break system; separates the endosperm from the bran and germ,
– sizing system; separates the small bran pieces attached to the large pieces of endosperm,
– reduction system; reduces the endosperm to flour, and
– tailing system; separates the fiber from the endosperm, removed from the other three systems.

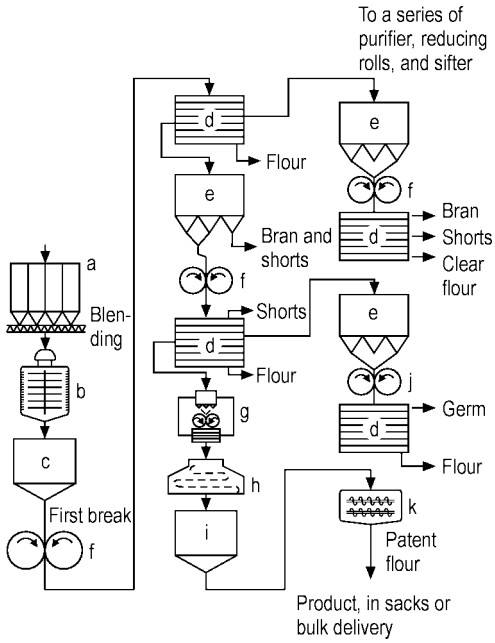

Figure 3. Flow sheet for milling process a) Tempering bins; b) Entoleter; c) Grinding bin; d) Sifter; e) Purifier; f) Reducing rolls; g) Reducing rolls and sifters; h) Bleaching; i) Bulk storage; j) Germ rolls; k) Flour treatment/enriching

The principal forces of grinding are compression, shear, friction-abrasion, and impact. Most grinders operate on a combination of these principles. The most common grinding machines are roller, attrition, impact, ball, or cutter mills, bran dusters, and pearlers. The oldest attrition mill for grinding wheat was the stone mill applying the forces of compression, shear, and abrasion.

The first break rolls pull the wheat-stock down with the rolls into the gap between the rolls or into the grinding zone. The wheat is crushed slightly and only a small quantity of white flour is produced. After sieving, the coarsest material is conveyed to the second break rolls. These rolls are set a little closer together than the first break rolls. Therefore, the material is crushed finer, and more endosperm particles are released.

Usually the process of grinding and sifting is repeated five to six times.

Roller mills have been made with different roll configurations such as two or four pairs of rolls in one stand. The roll surface can be smooth or corrugated, with grooves (flutes) running the length of the surface of the roll in a spiral to produce and support cutting or shear action. The grinding action of a pair of rolls is influenced by manufacture (diameter, length and surface of the rolls, profile differential direction, number of corrugations, and revolutions per minute, rpm).

Sieving. After each grinding step the ground material is conveyed to a plane sifter with a series of sloping sieves. The sifter gyrates, so the finer particles pass through the sieves from the head (top) to the tail (bottom). The plane sifter has multiple sections containing up to 30 sieves that gyrate horizontally to move the stock. During sieving, three classes of material are separated:

- coarse fractions, which are fed to the next break until only bran remains,
- flour, or fine particles, passing through the finest (flour) sieve, and
- intermediate granular particles (middlings)

Purification. Purification separates bran particles and germ from the endosperm. The process is the gradual reduction of middlings from the endosperm and the flattening of the germ and bran particles so that they can be removed in the sifter. The middlings are fragments of endosperm and small pieces of bran and the germ. Since the late 1990s the role of purifiers in the mill flow diagram has diminished as millers have become more aware of cost. The trend is to produce baker's flour with a little higher ash content because there is less demand for low ash patent flour. This tendency reduces the need for intensive purification.

Reduction. The purified and classified middlings are continuously pulverized to flour between smooth reduction rolls, which revolve at a differential of about 1.5:1. The flour particle size depends on the space between the rolls. Each reduction step is followed by sieving.

The germs are released largely by the break system and appear as yellow particles in some of the coarser middling streams. These streams are called sizings. The germs are flattened as the sizings are reduced and are separated as flakes during sieving.

Flour Handling and Blending. Each grinding and sieving operation produces flour. The flour blending department has facilities for special treatments of the flour: infestation control, enrichment and additives, flour stabilization, flour drying, fine grinding, and air classification. A process computer can be used for fully automatic operation of flour blending, final packing, and loading for shipment to the customer.

Different countries have different flour grades. In the USA the average flour yield is about 72 %, with 14 % bran and 14 % shorts. Moreover, patent flour (± 45 % extraction), short patent flour (65 – 68 % extraction), and straight flour (72 % extraction) exist.

In most European countries the flour grades are based on ash content. The German flour types are as follows:

Wheat flour	Content of mineral matters (d.m.)
Wheat flour	Content of mineral matters (d.m.)
Type 405	0.38 – 0.47 %
Type 550	0.49 – 0.58 %
Type 630	0.60 – 0.70 %
Type 812	0.75 – 0.87 %
Type 1050	1.00 – 1.15 %
Type 1600	1.55 – 1.75 %

The average total flour yield in Germany is about 80 %. Some other countries use a flour grading system based on flour color, not ash content.

To improve baking strength of wheat flour, some compounds are added in very small amounts. Flour bleaching was introduced as early as 1879 in Great Britain and around 1900 in North America. Currently, many bread flours and practically all cake flours in the USA are bleached. Furthermore, maturing agents are used to obtain maximum baking performance. In North America, potassium bromate is allowed to improve the baking strength of wheat flour. In the EU only ascorbic acid may be used legally as a flour improver. The bleaching agent benzoyl peroxide is not allowed in the EU. The quantitative requirement for oxidation of flour depends on several factors. Flour with higher protein content requires an increase in oxidation, but naturally time-aged wheat and flour need less

oxidation. The tolerance of flour to overtreatment varies with the improvers. Flour can tolerate overtreatment with ascorbic acid much better than with other improvers such as potassium bromate. Ascorbic acid is added at a level of about 10 – 20 ppm in some European countries and of about 25 – 75 ppm in the U.S. baking industry.

Infestation and Pest Control. Pests in flour mills are insects, rodents, birds, and microbes. Removing pests can be done both nonchemically or chemically. Nonchemical means are, of course, preferred because fumigants may leave a residue. In the past, insects were controlled with fumigants but now physical control is more common, especially heat. The air in the mill buildings is heated to about 50 °C for 24 h; these conditions kill all insects and also rodents [44].

Air Classification and Fine Grinding. Wheat flour produced by conventional roller mills contains particles from 1 – 200 µm. The flour can be ground (pin-milled) to fine particles in which the protein is separated from the starch. This fine flour can then pass through an air classifier. Particles smaller than 20 µm have up to twice the protein content than the unfractionated flour. The fraction with particles sizes between 20 and 40 µm has the normal protein content, and the fraction with particle sizes more than 40 µm has lower protein content. The high-protein flour is used to fortify low-protein bread flours or for enrichment in the production of specialty baked goods. Air classification has generated considerable interest in the milling industry, especially in producing flours with specific characteristics [44].

Quality Control and Analysis. To obtain an accurate analysis of the flour and provide information for blending the components into a final product, an exact sampling system is important. There are on-line automatic samplers, which can sample a flour stream before it enters the storage bin. Sampling systems provide representative samples of mill products for the laboratory or for on-line instrumentation. In the laboratory, routine testing includes the determination of ash, protein, and moisture contents, starch damage, and particle size distribution. Also, a baking test and a physical, i.e., rheological test on dough are performed. On-line quality control of mill products becomes a necessity in large capacity milling units. NIR analysis is used for on-line determination of parameters such as moisture, protein, and ash contents. Flour color is also determined on-line in the modern mill [33].

Whole Wheat Products. Whole wheat products are produced from the whole wheat kernel without any separation, and thus include endosperm, bran, and germ. The grinding of the wheat (or rye) can be done with stones or rolls with specified granulation [46, 47]. Using a roller mill, the system includes a small sifter which leads the oversized ground meal particles back to the grinding system for further reduction. The specifications for whole wheat products should be the same in both the USA and EU. Because the product is made from whole wheat meals, the raw material must be extremely well cleaned, including an intensive surface cleaning of the grains.

Whole wheat products are produced with different particle sizes: coarse, medium, and fine whole meals. Whole meal flour should pass to 80 % through 200 µm sieve openings. Under these conditions, the endosperm is milled to the same particle size as the normal flour. Therefore, a normal dough development, good gas retention, and an optimum bread volume are obtained. With coarse whole wheat products, the gluten-forming during dough mixing is hindered and aggravated. Therefore, gas retention and bread volume are much lower. Whole-meal products like whole-meal bread, whole-meal buns, and whole-meal cakes are advertised as having the highest dietary fiber content of all baked products. Therefore, consuming whole-meal products can add to the daily dietary fiber intake. Whole-meal pasta has a lower dietary fiber content than whole-meal bread because the moisture content after cooking is higher than 80 %.

Durum Wheat Milling. The objective of durum milling is to produce a maximum yield of highly purified semolina without specks. Therefore, the design of the milling system differs somewhat from the normal wheat mill. In semolina manufacture, impurities must be removed by cleaning and purifying systems. The reduction system is shorter in durum mills because of the particle size of the semolina (250 – 400 µm). Durum wheat with good milling quality yields 62 – 65 % semolina. In the 1990s the particle size distribution of semolina was very much reduced in combination with modern mixing systems in macaroni production lines [48].

6.1.2. Rye

In some countries in Europe (Denmark, Germany, Poland, and the Russian Republic), rye is still an important raw material for bread [49 – 52]. But even in these countries consumption of rye and rye-wheat mixed breads is decreasing. The ratio of rye to wheat in bread consumption in Germany has declined to only about 18 % rye, with 82 % wheat products. Rye flour has no gluten, and therefore, the gas retention capacity is lower than in wheat doughs. Rye bread has usually half the bread volume of wheat bread, so most rye flour is used in a blend with wheat flour for bread.

Rye is milled to flour in a process similar to that described for wheat. Very often ergot, a poisonous contaminant, is in the raw material and must be removed to a maximum level of 0.05 %. The bran of the rye kernel adheres tenaciously to the endosperm. Therefore, it is not practical to produce clean middlings or purify them by aspiration. The objective of rye milling is to produce flour during the breaking process. The break rolls are set relatively close and have finer corrugations than those used in milling wheat. Rye grain requires little tempering because it is very difficult to separate the bran from the endosperm. Rye flours have higher ash contents than wheat flours at similar extraction. Moreover, bread with a high percentage of rye flour in the recipe can only be produced by adding acids or using sourdough formula.

6.1.3. Maize

Maize is dry-milled for human consumption using two general methods: mill stones (old milling process) or steel rolls (modern milling process).

6.2. Dehulling

Dehulling is necessary for all so-called "covered cereals" such as rice, spelt wheat, oats, millet, sorghum, and some pseudocereals such as buckwheat. The kernels are enclosed with a tough siliceous hull that is unsuitable for human consumption [42, 53]. After removing the hulls, the kernel is obtained, consisting of the inner and outer bran, the endosperm, and the germ. Before dehulling, the grain must be intensively cleaned. Many machines can separate the hulls from the kernels.

Conventional shellers consist of two steel blades mounted horizontally. The inner surfaces are coated with a mixture of cement and carborundum. One blade is stationary and the other rotates. When the blade revolves, pressure on the ends of the upturned grains disengages the hulls. The hulls are removed by intense aspiration, and the remaining hulled and unhulled grains are separated in a paddy machine. In this machine, the unhulled grains move upwards and the heavier, hulled grains move down. After the dehulling process, the grains are considered as whole grain, because they still have their bran, endosperm and germ.

After dehulling, grains like *rice* are sent to machines with grooved tapering cylinders that revolve rapidly in stationary, uniformly perforated cylinders. The outside bran layers and germ are removed by the scouring action of the rice grains moving against themselves near the surface of the perforated cylinder. After passing through a succession of hullers, rice is nearly free of germ and outer bran. Scouring is usually completed by polishing in a brush machine. The polished rice contains whole endosperms and broken kernel particles of various sizes. The broken particles are usually separated by grading reels or disc separators.

For *barley*, the hulls are gradually removed by a pearling process; the outer portion of the barley kernel is removed by abrasion. These dehulled whole barley kernels are ready to be used as human food. Barley flour and barley grits are made from dehulled barley kernels by gradual reduction in a roller milling similar to the process of milling wheat.

The *oat milling* process has the following aims:

- dehulling the oat kernel,
- improving the shelf-life of the final product,
- increasing digestibility, and
- improving sensory characteristics.

In the cleaning process, the small oat kernels must be removed because their hull content makes up to 60 % of the whole kernel. The hull is separated with dehulling stones or impact milling. The dehulling stones consist of a

lower stone, which is flat and stationary, and an upper stone, which is slightly conical and rotates rapidly. The hulls are separated from the kernels by aspiration. The dehulled kernels are given a hydrothermal treatment to inactivate enzymes in the high fat content of the oat kernel. Using steam, the temperature increases within a few minutes up to 87 °C, and moisture increases up to 20 % [54]. Then, contact heating increases the temperature up to 100 °C for about 15 min. After this, the kernels are cooled to 8°C, and the moisture content is reduced to 12 %. The end product of oat processing is the oat flake. The flake is produced from steam-treated oat kernels or oat grits using a flake roll. It is also possible to produce oat bran with dietary fiber content up to 20 % by grinding and sieving steam-treated oat kernels.

6.3. Wet-Milling Process

The wet-milling industry uses about 80 % of processed *maize* to produce maize starch, sweeteners, maize oil, and various feed byproducts. The remainder is used in various food products using a dry-milling process as well as alcoholic beverages [55, 56]. Maize is processed for human food in the following ways:

- to be eaten as whole kernels;
- ground to make hominy;
- milled for maize grits, meal, or flour; and
- converted to a variety of breakfast foods.

Maize or wheat starch is produced from maize or wheat flour by the wet-milling process. The starch kernels are separated from the raw material using a multiple wet-sieving process. For this sieving process, special equipment is used, i.e., hydrocyclones, multicyclones, three-face dye separators, etc. After wet-sieving, the moisture content of the starch is reduced by mechanical separation and then by drying. Very high drying temperatures cannot be used to reduce the moisture content because starch granules are thermolabile and the gelatinization characteristics can deteriorate.

The following end products can be produced from 100 kg maize: 61 kg starch, 7 kg germs, 7 kg maize bran, and 5 kg gluten.

Sorghum is also wet-milled, but finer mash screens are necessary. Sorghum starch is similar to maize starch but is blander and does not become rancid.

6.4. Extrusion Process

Cereals can be processed into food by extrusion (Fig. 4). *High-temperature short-time (HTST) extrusion cooking* is used extensively to produce different products for infant foods, snacks, convenience foods, expanded pet food, feed, and cereals for industry [57].

The basic components of an HTST extrusion cooking system are as follows:

- continuous and uniform feeding of the raw material to the extruder;
- preconditioning the materials with steam or water at a controlled temperature (moisture: 20 to 30 % and heat: up to 80 %);
- controlling temperature during the whole process;
- controlling residence time in the extruder to optimize temperature, shear, and agitation;
- controlling extrudate shape and size; and
- drying, cooling, sizing, or heating the product by adding flavors, vitamins, or fats, etc.

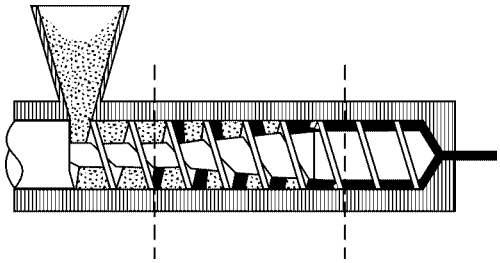

Figure 4. Extrusion cylinder

Pasta products are also processed by extrusion but in a *cold-extrusion process*. The first mechanical process to extrude pasta products rather than cut them from sheeted dough dates back to the early 1800s. Mixers, kneaders, hydraulic presses, and drying cabinets have been also used since the turn of 20th century. Continuous extrusion dates from the 1920s and fully automated processing, weighing, and packaging were introduced only in the 1950s.

High-quality pasta products are processed from semolina milled from durum wheat. The

advantage of durum wheat is its high protein content and the yellow color of the endosperm. There are pasta products processed only from semolina, water, and salt while other types have eggs as an important part of the recipe. In different countries, egg pasta should have a defined amount of egg content in the recipe.

Alimentary pastas are formed by extrusion on large automatic machines. Water and other ingredients are added to make stiff dough with a water content of about 30 %. The dough is then forced under pressure through dies of an extrusion machine. The dies for long goods, e.g., spaghetti, consist of a metal plate with drilled holes. The extruded spaghetti is folded on a rack, cut to length and dried on hanging rods under carefully controlled conditions. The temperature program during that drying process is very important for the cooking quality of the pasta. In former times, the drying temperature was about 40 to 50 °C, which did not improve the microbiological status. Most modern drying equipment uses high-temperature drying at 90 to 100 °C. Accelerated drying has enhanced both production efficiency and also product quality. Pasta products marketed in Europe and the western hemisphere have the following sizes:

- Spaghetti – small diameter, solid rods;
- Macaroni – hollow tubes of different sizes;
- Noodles – flat strips or extruded oval strips; and
- Variety products – cut by rotating or a blade cutters.

Some pasta companies have tried to introduce pasta made from whole durum meal flour to introduce dietary fiber into consumer diets, but customers did not accept this whole meal pasta because of color and the taste.

7. Specific Properties and Use of Grains

7.1. Wheat

Wheat (*Triticum* ssp.) is often referred to as the main and only bread cereal. Even though wheat production worldwide is likely the same as that of rice and maize (Table 1.), wheat appears to be the most important cereal and is constantly gaining importance. Originating in the Arabian Peninsula and Near East, wheat has spread to Europe, but also to eastern parts of Asia and to North Africa and was brought by immigrants to America and Australia. As a bread cereal, kernel hardness and structure affect processing properties in milling, and the proteins that form gluten, a network of strands and membranes providing fermenting dough with the gas retention necessary for leavening, affect baking properties [41, 58 – 61].

Botanically or genetically, wheat is a cereal differing much in number and type of genomes and in shape and characteristics of the spikes. The characteristics of spikes of wild, primitive wheat and of naked, cultivated forms are still present. Primitive wheat does not thresh well, breaking into spikelets covered closely with chaff, which must be removed from the kernel by dehulling before milling. In contrast, the spikes of cultivated naked wheat thresh well, freeing the kernel easily and ready for milling. According to the number of chromosomes and genomes, wheat is either einkorn (*T. monococcum*) with the genome AA, or emmer (*T. dicoccum*) with the tetraploid genome AABB, or finally common wheat (*T. aestivum* or *vulgare*) with the hexaploid genome AABBDD. Wheat with all the three types of genomes can be used for producing bread, but common wheat with the AABBDD genome has proved to process best, given bread volume and crumb properties (chewability) of breads and other bakery items. The tetraploid durum wheat (AABB) is mainly used for producing pasta from semolina, but in some countries of Southern Europe, Near East Asia, and North Africa the durum wheat is also used for bread making.

Wheat hardness can be separated largely into three classes including soft, hard, and extra hard (durum) wheat. Soft wheat flour has low protein content (7 – 9 %) and is good for making cake and cookies with extraction rates of 40 – 50 %, whereas hard wheat flour has higher protein contents (10 – 14 %) and is preferred for breadmaking with extraction rates of 70 %. Durum flour has the highest protein contents (12 – 16 %) and is mainly used for making pasta. Hardness is governed by interaction (adhesion) between starch granules and protein matrix covering starch granules. Soft wheat has a special protein (M_r 15 000, friabilin), which hard wheat

does not have, on the surface of starch granules. This protein prevents adhesion between starch granules and the protein matrix. The friabilin is known to be composed of two major proteins called puroindolines-a and -b. Either the absence of puroindoline-a or the presence of a glycine to serine mutation in puroindoline-b is required for grain hardness [12].

The applicability of milled wheat in bread making depends on the ability of wheat proteins to build gluten when mixed and kneaded with water and other ingredients into dough. The developed gluten is a cohesive viscoelastic mass of strands and membranes, which provide the dough with moldability and gas retention. Molded, shaped, and formed dough should preserve its loaf form through the process of breadmaking and retain the gas produced by yeast. Thus, the dough must be kneaded to a desired consistency, and the added water plays an important role as a plasticizer and softener.

The glutenin and gliadin ratio in gluten mass determines the physical and chemical properties of the proteins in flour, i.e., the viscoelastic properties of dough, its stretchability, and its firmness [60, 62]. The physical and chemical properties of wheat proteins are genetically linked and so far, are characteristic of a cultivar, although they are also influenced by the environment. A rather dry climate and soil deficient in nitrogen produce more glutenin protein in wheat with low protein content, and thus a firm and short dough resulting in low volume bread. On the other hand, a moderate climate and particularly soil with plenty of nitrogen promote the synthesis of gliadin protein, which softens the physical dough properties and ensures leavening, and consequently, gives a high bread volume and palatable crumb.

Optimal stretchability and firmness of the dough gives a bread of good quality and appearance, which will sell well and taste good. A second crucial quality characteristic of flour in the bread making process is the activity of α-amylase. The quality of wheat is, therefore, also characterized by α-amylase activity. Only a well-balanced quality of gluten and α-amylase activity in flour guarantees the best quality bakery items. Both properties, including gluten quality and sprout resistance, are intrinsic, i.e., genetically controlled, but also depend on environment and thus are influenced by climate and tillage. Physiologically, wheat produced using organic or alternative agriculture (applying low manure) has a higher ratio of essential amino acids, e.g., lysine, expressed as a percentage of the protein than that of wheat produced by applying high nitrogen fertilization. However, the wheat produced under high fertilization has a significantly higher protein content, the lysine content in a slice of bread, expressed in grams, is distinctly higher than that in bread produced by alternative or organic agriculture.

Most wheat is used to produce food, but wheat is also used in feed and in industry. This variability in properties and versatility of use make wheat indeed unique [59]. Processing for food starts in a mill that produces a variety of milled products. The milled products (flours of different grades, farina, semolina, and meals) are used to produce bread and other small bakery items of different types, pasta, confectionery, and breakfast cereals. The requirements of physical and chemical properties of flour, as a raw material for a variety of end-products, make it necessary rather to speak about the suitability of a specific flour for a specific purpose but not about quality in general. A specific product requires specific properties of flour, a degree of fines (particle distribution), contents and properties of proteins, and enzyme activities. These can be achieved by selecting wheat, choosing an adequate milling diagram, and adjusting the flour by blending, enriching and supplementing. Each baker can obtain from the mill the flour tailored for specific purposes.

Through long years of growing and selecting, wheat yields have been increased and wheat can be processed more easily but has lost the intrinsic robustness of a wild grass. Grown intensively across large areas, wheat needs intensive agricultural control, fertilization, and protection to achieve both high yield and quality. The genetic and spike characteristics determine the end use and thus determine the spread of wheat forms. Out of the known forms, only the highly cultivated and selected tetraploid durum and the hexaploid common wheat are grown on a large scale. Because of the green movement in some countries, the farming of einkorn, emmer, and particularly of spelt wheat (*T. spelta*), a primitive hexaploid wheat, had for a short time some success, but interest in this ancient wheat has waned. *Spelt wheat* was considered a healthy

field product, protected by chaff from air pollution, and containing substances relevant for wellness [63, 64]. The belief that spelt is good for celiac patients, has not been proved, because spelt has the same genome as common wheat, producing gluten proteins and seriously harming the intestinal walls of celiac patients. However, spelt is a welcome form of wheat, grown in adverse climatic conditions in mountain regions and ecological niches that should be offered to the consumers who appreciate spelt.

Kamut wheat and *selenium-rich wheat* are other sources of healthy products such as bread or breakfast cereals. Kamut has a higher content of protein and of essential amino acids than common wheat and has a specific nutty taste. Selenium is an important micronutrient that appears to be deficient in some regions. Grown on soils containing selenium, wheat and its products may contain more of this element beyond the threshold estimated to be necessary for health.

Some wheat is used in animal feed. Although the protein levels and the amino acid composition of all cereals are not sufficient for an effective feed, the cereals and wheat have sufficient starch content to supply energy [61]. Fed to poultry and pigs in ratios of more than 50 %, wheat provides animals with metabolizable energy, but for their growth and healthy development, the feed must be supplemented with other sources of plant or animal proteins and minerals. The nonstarch polysaccharides, cellulose and pentosans (arabinoxylans and xylans), may have antinutritive effects, particularly with poultry. Adding enzymes to degrade the pentosans has proved effective in increasing uptake and digestibility of proteins and fats from feed containing wheat and barley. Microbial and fungal contaminations, particularly with *Fusarium graminearum* and smut spores (*Tiletia* sp.), affect voluntary feed-intakes. Thus, the feed to gain performance of animal meat production is affected.

Surpluses of cereals beyond what is needed for food and feed have found other possible uses. Cereals in general and wheat in particular have been used as a raw material in nonfood industry [59]. Wheat can be used as a whole plant or as milled products, milled to a requested degree of extraction. Wheat straw can be used in the paper industry, and wheat starch and gluten, separated by wet milling in a starch mill, are already in use as sizing and coating material in paper manufacturing.

The separation of starch and gluten by wet-milling is a traditional use of wheat flours, and the starch and starch derivatives have very different properties. Both starch and gluten are used for both food and nonfood industries. Starch is used as a binder in soups, sauces, sausages, in pudding-jellies, and as a component in paper and wood adhesives, but also in medicine and pharmacy as carrier of medicaments (→ Starch). Gluten, as a byproduct of starch production, is used as an additive to feed and as an adhesive (→ Proteins, Chap. 2.1.1). Specific gluten, not damaged in drying, is called vital gluten and is used to fortify flours, as a protein-fortifier, and as an improver in flours for specific bakery items.

Wheat and wheat starch is also a source for chemicals. Fermented and distilled alcohol as a beverage or for industrial purposes is only one of chemical products. Starch is also a source of polyols, of biodegradable plastics, and of rubber additives. A promising and interesting branch of starch products are the microbial polysaccharides that transform starch to new polymers of useful properties. These uses, and many others, of starch and gluten have been discussed, opening new possibilities in production and use of starch and gluten [56].

7.2. Rye

Rye (*Secale cereale*) is the main crop in regions with adverse growing conditions for other cereals [49, 51, 52]. As a modest plant, it can be grown on poor soils, providing a higher yield than wheat. In addition, rye is resistant to cold, drought, pests, and diseases, and its yield potential on good soils is comparable to that of wheat. The hybrid cultivars are particularly high yielding. However, rye is susceptible to lodging, to preharvest sprouting, and to ergot, which in years with high rain fall obviates its use as food and feed. Nevertheless, rye in former times has made growing of wheat possible, because rye and wheat were sown together expecting to obtain a yield from rye when the wheat, under bad conditions, failed.

Rye is milled to integral flour and meal, and highly extracted to dark flour with a rather high mineral content. It is used in sour-dough bread,

pure or blended with wheat flour. Unlike wheat, the allotment of water-soluble proteins of rye (up to 40 %) is higher than for wheat, and therefore, rye does not form a firm gluten mass. The baking behavior of milled rye products is affected by its pentosan (arabinoxylan) content [65, 66]. The pentosan content may vary between 7 to 10 % in kernels; it is low in flour (2 – 4 %) and high in whole kernel meal and particularly in bran. Rye is a good source of minerals, vitamins, dietary fiber, β-glucan, and of lignan (phytohormones such as secoisolariciresinol and matairesinol), that can help prevent cancer [22, 24 – 26, 28, 51, 67].

Due to its high content of the essential amino acids, particularly of lysine, rye was at one time considered an effective feed for pigs. However, the use of rye as feed is limited by its content of alkylresorcinol and antinutritive substances such as pentosan, which are indigestible and also presents starch and proteins from being digested. Rye is still a cheap component in feed and pet food and is also used for various industrial purposes [68].

7.3. Triticale

The first report on triticale, a genus bastard of wheat and rye, was recorded at the end of 19th century. Triticale was a spontaneous cross between the two genera, but the incompatibility of chromosome number meant it was sterile. At the beginning of the 20th century, a fertile product of the crossing between wheat and rye was achieved; since then quite a number of intensive breeding projects have been started. Triticale is the first man-made cereal, and today a great number of adapted cultivars are grown around the world, particularly where soil conditions are adverse. The main producers of new cultivars and mercantile lots are Canada, Mexico, Germany, and Poland, but Australia has also made efforts to breed and grow triticale. The produced triticale has not been yet registered by FAO statistics.

The general idea was to combine the modesty and resistance of the rye plants with the milling and baking properties of wheat. The general problem with this combination is how to obtain fertile types and cultivars of triticale starting with parents with different numbers of genomes. The rye genome of 7 chromosomes must be increased to 14 chromosomes to be compatible with hexaploid wheat (*T. durum*) or to 21 chromosomes for octaploid type of wheat (*T. aestivum* or *vulgare*). This can be obtained only by disrupting chromosome division with toxic colchicines or with ionizing irradiation. Triticale cultivars are either primary (rye × wheat) or secondary (triticale × triticale).

As a product of different parental properties, the properties of triticale cultivars lie intermediate between wheat and rye, with different allotments of each parent. In growth habit and in processing, some triticale cultivars are more like wheat and others more like rye, but the main aim, to crop wheat from rye plants, has not yet been reached. Triticale yield has not reached yield potentials of wheat, nor are its baking properties (weak gluten, susceptibility to sprouting) equal to wheat [65, 69]. Nevertheless, many bakeries produce triticale bread with an adapted formula. In general, triticale is susceptible to preharvest sprouting requiring special care at harvest and as it is processed to bread. Triticale seems to be rich in protein and in essential amino acids, like rye, which is good for animal feed. Thus, triticale is a feed grain, mostly used on the home farm.

Several wheat cultivars with high-baking quality do have rye chromosomes introduced for resistances, which have not been detrimental to their baking properties. This promises that triticale cultivars of good processing value and baking properties can be created. Cereal breeders and processors would be interested in such cultivars and offer a bakery item to consumers or consumers would ask for it.

Triticale can be used as any other cereals for various industrial uses, as a source of chemicals and as a source of energy. Attempts have been made to breed triticale cultivars with a high-yield potential and a high total mass of dry matters, which can be burned to produce heat or energy as a renewable biomass resource.

7.4. Barley

Barley (*Hordeum vulgarum*) is probably the oldest and the most widely grown cereal in the world. Its winter hardness and resistance to climate and soil adversity mean that barley is grown

farther north and higher in the mountains than any other cereal [70]. Barley is grown worldwide as two main types, two-rowed and six-rowed barley, and also as winter and spring barley. Most produced barley is used as feed, with small portions as malt and food. And yet the image of barley is closely connected with brewing of beer. The six-rowed winter barley is used for feed in Europe and the two-rowed for malting, in America the six-rowed barley is preferred for malting.

As food, milled barley was used for bread-making, later as pearled barley in porridge and soups, and ground to semolina as couscous. Recently, barley flakes were introduced and used in bakeries and in breakfast cereals. In any case, barley grain must be dehulled prior to milling. Other types of barley are hull-less, thus avoiding the disadvantages of dehulling, and have been promoted. The rediscovered hull-less barley and the new high-lysine barley found acceptance for both food and feed. For this and for its high content of soluble β-glucan, barley products are considered healthy, reducing blood cholesterol and heart disease, and helping prevent colon cancer. These all might have increased the barley consumption, but barley will never regain its former importance as food [70].

Barley is fed to ruminants (cattle, sheep), monogastric animals (predominantly pigs), and poultry; as a starch-rich product, it contributes to the animals' energy supply. Barley grown for feed has a higher protein level than that grown for malting. For better digestion, the barley kernel should be rolled or flaked for ruminants or ground by a hammer mill for monogastric animals. Barley glucans seem to reduce availability of starch and protein fed to poultry and to produce sticky faeces. Therefore, adding enzymes with β-glucanase is recommended.

Barley is also a cheap ingredient for alcoholic beverages. To produce malt for both beer and distilling, whole plump barley with unharmed germ, a high potential of enzymatic activities, and low protein content is required. Even so, different distilled alcoholic beverages, such as malt whisky or grain whisky, have different requirements. In the malt whisky prepared only from barley malt, the malt should have a high level of starch and only enough enzymes to degrade the endogenous starch in the malt. Grain whisky, on the other hand, contains a limited amount of barley malt to produce enzymes that degrade starch from adjunct starch from other cereals. The difference lies in the enzymatic activities of the malt types. Barley for brewing should have a good proportion of active starch-degrading enzymes and low soluble protein content. It should be free of undesirable compounds such as β-glucan and polyphenols. Different types of malt have different aromas, flavors, and colors and hence give special character to the final brew. Although beer consumption is decreasing, it remains the most popular beverage worldwide.

7.5. Oats

Oat is a genus comprising many wild, weedy, and cultivated species found and grown in all agricultural areas. Of the cultivated species, white oats (*Avena sativa*) is the most wide spread. It is grown as a spring cereal; its inflorescence is a panicle bearing hulled kernels. Besides the hulled type, a naked type of oats (*Avena sativa* var. nuda, former *Avena nuda*) has been pushed by breeders but did not find wide acceptance [71]. Oat kernels are unique, with high levels of dietary fiber, good quality proteins, and high levels of lipids, all making oats a healthy cereal. Indeed, its healing effects and medical value have been described early in the history of mankind [22, 72].

Oats are used both for food and feed. Dehulled and milled into groats or flaked oats, the cereal is highly appreciated, prepared as porridge and oatcake, flour for baby food, and breakfast cereals. A white kernel, unspotted by bacteria, is required to produce clean, white products. Nude oats with their unprotected kernels often do not meet these requirements. In the flaking process, oats are heat-treated, giving the product a pleasant nutty taste and preventing the flakes from deteriorating by enzymatic oxidation of lipids (oats peroxidase). This treatment prolongs the shelf life of oat products in comparison to untreated oats. The commercial products are rolled oats, steel-cut groats, and quick oats produced from them, baby oats, instant oat flakes, oat flour, and oat bran. In the hulls and in the outer parts of the oat kernel also an anticarcinogenic factor has been found, increasing the health value of oat products. Some oat products are used in bread making, providing bak-

ery items with nutritional benefits, typical flavor, and moisture retention, and thus a longer shelf life. Bread making with oats is only possible in combination with wheat flour.

Oats have been used mostly as feed, especially for horses. Even today, oats are preferred feed for horses because of their palatability, digestibility, and nutritional quality, although low-cost alternative feed containing maize, soy meal, cottonseed meal, and canola meal has been brought on the market. Compared to hulled oats, such products have higher energy density and are more effective in feed. The dietary fiber, β-glucan, seems to reduce feed performance in young animals. The fact that nude oats also have high energy density suggests that oats may regain its position, at least for racing horses.

In the nonfood field, oat hulls are an important milling byproduct as a source for production of furfural. Novel products such as oat starch are used as adhesives and in cosmetics. In another novel process, oats are treated with α-amylase, the water-soluble dietary fiber separated from insoluble materials and used as a fat replacer for bakery products. This and the cholesterol reducing properties of oat products may open the market to new oat products.

7.6. Rice

The rice grain, a covered caryopsis, slightly smaller than wheat, is flattened laterally, and has a small point at the end distal from the germ. The kernel is enclosed in a tough siliceous hull that is unsuitable for human consumption. Rice (*Oryza sativa* L.) is one of the leading food crops in the world and the staple food for more than half the world's population. It is considered a semi-aquatic, annual grass plant. Cultivars of the two cultivated species, *O. sativa* L. and *O. glaberrima* Steud., can grow in a wide range of water-soil regimes including over 100 countries on every continent except Antarctica. Annual production (601.61×10^6 t) of rough rice, paddy, approaches that of corn (604.560×10^6 t) and exceeds that of wheat (586.91×10^6 t), as shown in Table 1.

Rice and wheat are the two most important food grains, because maize is used mainly as feed except in Latin America and parts of Asia. Rice is the staple food of East, Southeast, and South Asia. This region produces 91 % of total world production (Table 1). When the hull is removed from the rough rice, or paddy, the kernel or caryopsis, comprising the pericarp and the seed proper is known as brown rice. When brown rice is milled further, the bran, aleurone layer, and germ are removed, and the purified endosperms are marketed as white rice, polished rice, or milled rice.

Among cereal grains, paddy, or rough rice, is unique in that it is milled as whole grain, not flour. It is also unique in that milled rice can take two forms, white rice (raw rice, i.e., nonparboiled) or parboiled rice, obtained by treating paddy rice with water and heat before it is dried and milled. Parboiled rice likely originated in ancient India, but the exact origin is unknown. Parboiled rice is the major staple food throughout the Indian subcontinent, where more than 90 % of the world's parboiled rice is produced and consumed. It is estimated that approximately 20 % of the world's rice is parboiled [73, 74].

The uses of rice components and coproducts include (1) stabilized rice bran for health foods, cereals, baked goods, fat replacer, and dietary supplements as a source of fiber as well as oryzanol, a phytosterol esterified to ferulic acid, demonstrated to reduce serum cholesterols in human; (2) rice bran oil, a high-quality cooking oil with an excellent balance of fatty acids and excellent frying oil with low formation of peroxides, foam, free fatty acids, and polymers; (3) rice starch as a cosmetic dusting powder, laundry stiffening agent, paper and photograph paper powder, sugar coating in confectionery, excipient for pharmaceutical tablets, and fat replacers, due to the smallest granule size and very white color of rice bran; and (4) rice protein concentrates with good amino acid balances. Uses for rice hulls include roughage for cattle feed; chicken litter; ammoniated for fertilizers; filter aids; burnt hulls for floor-sweepings; fuel; abrasives; binder for pelleted feeds; insulating material; filler for building materials; binders and absorbents for pesticides, explosives, and other materials; and also in the commercial manufacture of furfural, although their pentosan content is lower than that of oat hulls or maize cobs.

7.7. Sorghum and Millets

Sorghum (*Sorghum bicolor* [L.] Moench) is an ancient grain first domesticated in Africa and spread to other parts of the world around 3000 B.C. Sorghum is a drought-resistant cereal grain, often growing in semiarid conditions where other cereal grains cannot. Sorghum is a tropical cereal grass like rice and maize. It is a diploid plant with $2n = 20$ chromosomes like maize. However, it has a relatively small genome, only one-third the size of maize [75]. Among cereal grains, sorghum ranks fifth, with 57.9×10^6 t harvested between 1999 and 2001 (Table 1). The major sorghum producing centers are in Africa and Asia (54.4 % of total production), with 34 % of the world supply of sorghum grown in North and Central America. About 30 % of the world production is provided by the USA, mainly from Kansas, Texas, Nebraska, and Missouri [76].

The relative proportions of pericarp, endosperm, and germ in the sorghum seed account for 6, 84, and 10 % of kernel, respectively. Although the thousand kernel weight (TKW) of sorghum, a naked caryopsis, ranges from 3 to 80 g, grains of commercial sorghum hybrids have a flattened-spherical shape (4 mm long, 2 mm wide, and 2.5 mm thick) with a TKW of 25 – 35 g, test weight (TW) of 55 – 62 pounds per bushel (lb/bu), and grain density of 1.38 g/cm^3 [76]. Sorghum is broadly classified in two classes, i.e., tannin sorghums and nontannin sorghum, irrespective of the grain color. The color of the grains depends on the color of the pericarp, whereas the presence or absence of tannins depends on the presence or absence of testa in the inner pericarp. All sorghums contain phenolic compounds, but not tannins. About 99 % of US sorghums are of the nontannin type.

Sorghum is a major staple food crop in many African countries and in India where agricultural and environmental conditions do not favor the production of other grain crops. Thus, it is critical for food security in the semiarid tropics of Africa, Asia, and Latin America [77]. In the Western Hemisphere, where more than 40 % of sorghum is produced, it is used primarily as an ingredient in livestock feeds. Based on the types of end-uses, sorghums referred to as (a) grain sorghums - short, mechanically harvested varieties used in food and feed; (b) forage sorghums – tall, yielding fodder and grain for livestock and foods; (c) sweet sorghums – juicy leaves and stem used for sugar, syrup, and alcohol production, and (d) grassy sorghum – tall stems for grazing silage and hay [75].

As shown in 2004 this ancient grain may have unique health benefits, because of high antioxidant levels related to the content of polyphenolic compounds and sorghum wax, containing polyicosanols, which may be important in cardiac health. Therefore, the ancient grain sorghum has drawn new interest and promises to become an important cereal grain in the 21st century as a potential staple food for people with celiac disease; because of high antioxidant and wax levels for nutritional and health claims; and because of its drought-resistance and good production with limited water [39].

Millet is a generic term implying "small seeded grain" and thus, the variations in millets are quite large. Millets and sorghum account for only 4.13 % of the total world cereal grain production with approximately 28 and 57.9×10^6 t, respectively (Table 1). Moreover, among all cereal grains, the average yield is the lowest with 7.68 dt/ha (768 kg/ha) for millet, followed by 13.97 dt (1397 kg)/ha for sorghum (Table 2). Both are, however, extremely important staple foods in semiarid and sub-humid zones. Two broad categories of millet species include pearl millet (*Pennisetum americanum* [L.] Leeke) and the small and/or minor millet. Pearl millet, also known as bulrush millet, is the world's sixth most important cereals. Pearl millet has the largest seed and is the most widely grown of all the millets. Because of its highest drought tolerance, it is grown primarily in the dry Sahel zone across Africa and in the semiarid areas in India.

7.8. Maize (Corn)

Maize (*Zea Mays* L. ssp. *mays*), a diploid plant with 10 chromosomes ($x = 10$, $2n = 20$) is native to the Americas and is their most important cereal crop. Maize originated in Mexico, evolving from the wild grass Teosinte (*Zea Mays* sp. *Mexicana*). Maize was domesticated and grown as a crop by Native Americans as early as 5000 B.C. in Mexico's valley of Tehuacan. Maize became a symbol of religion and prosperity in the Mayan civilization. Maize spread

northward to the USA and Canada and southward to Argentina. In 1492, the explorer Christopher Columbus discovered maize, known as "mahyz", this staple food of the Native Americans in the New World. Maize was transplanted to Spain following Columbus's discovery and quickly spread across Europe and to Africa and Asia from Spain [78].

In addition to being an economically important crop plant, maize is a model plant for studying genetics, because one of the major contributions of maize genetics was to the field of cytogenetics: maize cytogeneticists were the first to demonstrate that chromosomes are individually recognizable by their lengths, arm ratios, and other physical features and that physical exchange of chromosome segments accompanies genetic recombination. Through various breeding and genetics biotechnology and/or genetical modification techniques, maize has been selected for improved genotypes, kernel modification, and wide variations, tailored to the quality parameters and quantitative traits desired by consumers, including not only producers but also end-users. Some varieties include high-protein/high-lysine "Opaque-2" and "floury-2" maize; "waxy maize" with over 99 % amylopectin, which grew on 0.22×10^6 ha in the U.S. in 2000; "high-amylose maize or amylomaize" with 80 % amylose; "high-oil hybrids" with general accompaniment of reduction in yield, but increased oil content and also alteration of fatty acid balance.

Maize is a unique crop to be able to provide so many diverse products by the dry-milling, with maize grain fractions mainly for food uses, but some industrial uses, including wallpaper paste with fine grits; floor wax and hand-soap with maize meals; dusting agents and hand-soap with cones; fermentation media, explosives, wallboard, and briquetting with maize flour; and foundry binders, label adhesives, and edge paste with specialty products. The estimated uses of maize dry-milling products are 36.2 % for animal feeds, 54.3 % for food, and 9.5 % for industrial (nonfood) uses [78]. Typical wet-milling industry products are maize sweeteners, starch, maize oil, maize gluten meal (60 % protein) with byproducts including maize gluten feeds (fiber, germ cake, and steep liquor solubles to yield 21 % protein). The estimated utilization of maize starch is 59 % unmodified starch, 11 % acid-modified starches, 5 % each of dextrins and oxidized starches, 3 % each of cationic starches and pregelatinized starches, and 14 % others. The zein (maize proteins) fraction, recovered from maize gluten meal is highly resistant to water and forms tough films and fibers, unlike most other commercially available proteins. Thus, zein has a number of industrial nonfood applications, such as in paper sizings, paper and paperboard adhesive, additives to oil cloth and linoleum, felt stiffener, pharmaceuticals (binding and seal coating tablets), varnish substitutes, printing inks, cork binding, spun textile fibers, floor coatings, and paper coatings.

Since the 1980s, maize processing in the U.S. nearly tripled from 16.7×10^6 t (0.659 $\times 10^9$ bushel) in 1980 to 50×10^6 t (1.968 $\times 10^9$ bushel) in 2000; the biggest increase was for fuel alcohol production, which rose nearly 20fold from 0.889×10^6 t (35×10^9 bushel) in 1980 to 15.93×10^6 t (627×10^9 bushel) in 2000 and 17.53×10^6 t (690×10^9 bushel) in 2001. Use of ethanol for blending with gasoline has been encouraged by reduced federal and state tax rates on gasoline blends. The future prospects of maize products to replace those currently derived from nonrenewable resources, such as petroleum and minerals, and to possess environmentally-friendly properties will increase demand for maize in nonfood and nonfeed industrial products such as plastics as building materials[78].

7.9. Pseudocereals

7.9.1. Buckwheat

Buckwheat (*Fagopyrum esculentum*) is the best known pseudocereal. It belongs to the knotweed (*Polygonaceae*) family, of which several genus and cultivars are grown. *Fagopyrum tataricum* is probably the most widely planted cultivar. Originating in China, buckwheat came in the late Middle Ages to Europe and has spread to Russia and to Japan. It was always a poor people's food, grown on poor soils, but recently, buckwheat has gained interest as a healthy food and has been used in ways other than the traditional dishes [79]. As estimated by FAO, buckwheat production worldwide was 3.6×10^6 t in 1990,

has reached as much as 4.4×10^6 t in 1992, and has fallen down to about 2.6×10^6 t in 1999 [79].

Dehulled and milled to flour or groats buckwheat is considered a nutritious food, rich in high-quality proteins of high biological value with a well-balanced amino acid composition rich in lysine. It also contains considerable amounts of alanine and fagonin (a specific amino acid), and a relatively high level of dietary fiber. The proteins consist mostly of globulins, but the prolamin fraction seems to be absent. This makes buckwheat suitable for celiac patients.

The buckwheat seed contains flavonoids, which are reduced in dehulling and milling, but both are used as therapeuticals containing rutin [79 – 81]. Almost all parts of the buckwheat plant are used, green plants and leafs are fed to cattle (but with care because the rutin is light-sensitive and may cause colored skin spots), and for extraction of the pharmaceutical rutin. As food, buckwheat is used mainly in breakfast cereals, as mush-porridge, in soups, as replacement for rice, for pasta products, but also prepared in various traditional and gluten-free dishes. There are quite a number of recipes for bread and cookies from both pure buckwheat flour and when added to wheat or maize flour [80]. Even the hulls can be used to fill pillows or as a packing aid.

7.9.2. Amaranth

In pre-Columbian times, the amaranth grain (*Amaranthus caudatus*) was, with maize and beans, a basic food in the New World. Quite a number of other names in Spanish and Indian dialects exist, including sangoracha, achita, and kiwicha. Because of its use in pagan rituals and human sacrifice, the Spanish conquistadors and missionaries forbade the cultivation and the use of amaranth [82]. Other than this, amaranth was used as pot herbs, as medical plants, and as vegetable in China, India, and in parts of Africa and Europe. About 30 years ago amaranth started to be an attractive crop again but has not yet achieved the popularity it had. There are at present no exact data on production of amaranth, except that amaranth has been grown on more than 1000 000 ha in Russia and China, and that the produced amounts are not enough to meet the demands of potential consumers. In many industrial countries, amaranth is grown as an ecological or biological product, and the taste of different amaranth cultivars seem not to meet the taste expectations of many of consumers; both are obstacles in gaining expected popularity on the market [82].

In addition to *A. caudatus*, about 60 species are known, such as *A. cruentus* and *A. hypocondriacusas* grain amaranth, and other as vegetable and weed amaranth. Amaranth grain is a small seed with a diameter of about 1 mm, coated partly with a ring-shaped thin coat, and thus it is not necessary to dehull it before use as food or feed. Although amaranth seed is considered rich in starch, it is also higher in protein (14 – 16 %) than cereals in general. The amino acid composition of amaranth proteins is nutritionally very beneficial because its lysine content is comparable with that of soybeans. The fat content (6 – 8 %) is also higher than in cereals, which limits storage time and interferes with the process of extrusion cooking.

Amaranth is usually boiled to mush-porridge and lime-cooked (to replace the maize in tortillas), popped by short and dry heat, but can also be blended with wheat for bread and pasta products (not higher than 10 – 25 %) [83]. The extruder-cooked amaranth is used as breakfast cereal or granola bars. Further uses are in fermented foods (comparable to soy sauce), protein concentrate production, starch and starch hydrolysis products as well as oil products.

7.9.3. Quinoa

Quinoa (*Chenopodium quinoa*) is originally from the higher Andean region of South America but is also found in Mexico and far away in Himalaya [84]. There is an evidence that quinoa has been cultivated for more than 7000 years. The worldwide 5-year average annual production is estimated as about 18 000 t from 1970 to 1974 and over 45 000 t from 1995 to 1999 [84].

The grains are small (diameter 1 – 2.6 mm), but enveloped in a dry two-layered pericarp that requires dehulling before use as food. The grains are protein- (about 14 %) and fat- (about 9 %) rich, and the content of total dietary fiber (about 13 %) is also high. The somewhat high content of saponins and other antinutrients reduces its

use as food, and processing to food includes techniques to remove saponin. The removal of saponins is traditionally performed by washing or by physical abrasion. A technique based on both processes has proved to be most effective, whereby loss of nutrients is minimized.

Dry-milling makes it possible to obtain whole kernel flour as well as two fractions of milling products: the protein- and fat-rich bran fraction and a starch-rich flour fraction. To achieve starch gelatinization, flour has to be boiled, or dry-cooked by high-temperature short-time extrusion. The flour is mostly used as porridge or instant infant porridge because of its high energy and of high protein quality in food. Quinoa flour can be used also to produce bread and cookies [83] and is gluten-free, which makes it appropriate for celiac patients. It is used in malting and fermentation, and as a raw material for starch production. Washed-out saponins are useful in beer brewing or the hygiene industry, where surface active substances are required as foamers [84].

8. Economic Aspects and Outlook

Cereals are traditionally considered food and feed grains. However, during the last decades, special market situations have encouraged exploring industrial uses of cereals [68]. Cereal grains, milled products, and byproducts of milling are finding increasing uses in a variety of nonfood applications.

First of all, the straw as well as the grain can be used as renewable biomass. Residues and byproducts of cereals are excellent sources of furfural (→ Furan and Derivatives, Chap. 3), a basic raw material in many industrial technologies. The biomass can be pyrolyzed to produce sugars, olefins, charcoal, and gaseous fuel. Carbohydrates can be hydrolyzed to sugars, which can be fermented to make ethanol, a potential source for ethylene and butadiene. Biomass can be digested by anaerobic bacteria to produce methane.

Cereal polymers (carbohydrates and proteins) can be converted to monomers, which are basic raw materials. Equally promising are novel uses of undegraded polymers. Examples are the modification of starch and graft polymerization to give new polymers. A small amount of dry-milled corn products are used as pregelled corn flours in foundry cores and molds, in ore refining, in building materials, as binders, in fermentation, in oil-well drilling mud, as adhesives, as binders in paper, paperboard, and corrugated products. Wheat or low-grade wheat flours have also a broad application in industrial uses: in paper sizing and coating, in the plywood industry and in detergent formulations. Gluten is used in paper manufacture; in surface active agents; and in adhesives, monosodium glutamate and glutamic acid, coatings, and gums. Wheat germs are raw material for the production of antibiotics, vitamins, pharmaceuticals, and skin conditioners.

Finally cereals are a very important source of energy. The production of alcohol as a source of fuel has been a subject of many investigations. In the USA, South Africa, Australia, and especially Europe, several companies are engaged in producing alcohol from cereals. The average world cereal production is now more than 2×10^9 t/a. Cereals are the most important stable food in the world. Only cereals have the potential to support the world human population with the needed amounts of energy and proteins. In some countries, up to 80 % of food calories are provided by cereals or cereal products. In other countries, this percentage has diminished, as meat production has grown, down to 20 % of caloric intake. Worldwide cereal production, primarily due to increased yields, has caught up with the world population. Taking into consideration the very large differences in yields (dt/ha) that exist in different areas, further yield increases depending on economic circumstances can be expected. Genetically modified organism (GMO)-breeding could be an important factor in increasing yields by reducing crop losses caused by insects and special growing conditions.

9. References

1. FAO: Bulletin of Statistics, Vol. 3, No. 2, 2002. Cereals, Rome 2003, pp. 17–39.
2. C. W. Wrigley, C. F. Morris: "Breeding cereals for quality improvement," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, pp. 321–370.
3. K. Oldach, G. Leckeband, D. Becker, H. Lörz: "Transformation and Expression of Antifungal

Genes in Wheat and Barley to Increase Their Fungal Resistance - Two Strategies," in D. Weipert (ed.): *Eighth International Symposium on Pre-Harvest Sprouting-Cereals*, Association of Cereal Research, Detmold, 1999, pp. 307–315.
4. B. Kniel, M. Moser: "Gentechnik und Ernährung. Die Bedeutung mittels Gentechnik gewonnener Zutaten für die Backwarenherstellung und deren analytischer Nachweis", in *Handbuch Backwaren*, 11. Akt.-Lfg. 12/00, Behr´s Verlag, Hamburg 2000.
5. R. E. Evenson, V. Santaniello (eds.): *The Regulation of Agricultural Biotechnology*, CABI Publishing, Wallingford 2004.
6. A. H. Meyer: *Gen Food, Novel Food - Recht neuartiger Lebensmittel*, H. C. Beck, München 2002.
7. R. E. Evenson, V. Santaniello (eds.): *Consumer Acceptance of Genetically Modified Foods*, CABI Publishing, Wallingford 2004.
8. H. Knittel, E. Albert (eds.): *Praxishandbuch Dünger und Düngung*, AgriMedia, Bergen/Dumme 2003.
9. J. T. Mills: "Quality of Stored Cereals," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, pp. 441–478.
10. F. J. Baur: *Insect Management for Food Storage and Processing*, AACC, St Paul, MN 1984.
11. Y. Pomeranz, D. B. Bechtel: "Structure of Cereal Grains as Related to End-Use Properties," in Y. Pomeranz (ed.): *Cereals '78: Better Nutrition for the World's Millions*, Am. Assoc. Cereal Chem., St. Paul, MN 1978, pp. 85–104.
12. M. J. Giroux, C. J. Morris: "Wheat Grain Hardness Results from Highly Conserved Mutations in the Friabilin Components of Puroindoline-a and b," *Proc. Nat. Acad. Sci.* **95** (1998) 6262–6266.
13. T. B. Osborne: *The Proteins of Wheat Kernel*, Carnegie Institute, Washington, D.C., Publ. No. 84, 1907.
14. O. K. Chung, Y. Pomeranz: "Functional and nutritional characteristics of cereal proteins," in J. W. Finley, D. T. Hopkins (eds.): *Digestibility and Amino Acid Availability in Cereals and Oilseeds*, Am. Assoc. Cereal Chem., St. Paul, MN 1985, pp. 169–231.
15. P. R. Shewry, A. S. Tatham, J. Forde, M. Kreis, B. J. Miflin: "The Classification and Nomenclature of Wheat Gluten Protein: A Reassessment", *J. Cereal Sci.* **4** (1986) 97–106.
16. O. K. Chung, Y. Pomeranz: "Recent Research on Wheat Lipids", *Bakers Dig.* **55** (1981) no. 5, 38–50, 55, 96, 97.
17. O. K. Chung: "Functional Significance of Wheat Lipids," in Y. Pomeranz (ed.): *Wheat is Unique*, AACC, St. Paul, MN 1989, 341–368.
18. O. K. Chung, Y. Pomeranz, K. F. Finney, M. D. Shogren: "Surfactants as Replacements for Natural Lipids in Bread Baked from Defatted Wheat Flour", *J. Am. Oil Chem. Soc.* **55** (1977) 635–641.
19. F. MacRitchie, P. W. Gras: "The Role of Flour Lipids in Baking," *Cereal Chem.* **50** (1973) 292–302.
20. R. A. Graybosch, C. J. Peterson, K. J. Moore, M. Stearns, D. L. Grant: "Comparative Effects of Wheat Flour Protein, Lipid, and Pentosan Composition in Relation to Baking and Milling Quality," *Cereal Chem.* **70** (1993) 95–101.
21. O. K. Chung, Y. Pomeranz, K. F. Finney: "Relation of Polar Lipid Content to Mixing Requirement and Loaf Volume Potential of Hard Red Winter Wheat Flour," *Cereal Chem.* **59** (1982) 14–20.
22. L. Marquart, J. L. Slavin, R. G. Fulcher (eds.): *Whole-grain Foods in Health and Disease*, AACC, St. Paul, MN, second printing 2005.
23. O. Theander, E. Westerlund, P. Amend: "Structure and Components of Dietary Fiber," *Cereal Foods World* **38** (1993), 135–141.
24. H. Adlercreutz: Phytoestrogens: Epidemiology and a possible role in cancer protection, *Environ Health Perspect* **103** (1995) 103–112.
25. W. Andlauer, P. Fürst: "Krebsprävention durch Getreide," *Getreide Mehl und Brot* **52** (1998) no. 6, 355–357.
26. D. Weipert, J. M. Brümmer: "Health and Technical Aspects of Whole Rye Bread," in Proceedings of the International Conference: Cereals for Human Health and Preventive Nutrition, Brno 1998, pp. 70–74.
27. B. Fretzdorff: "Bioaktive sekundäre Pflanzenstoffe – Protektive Wirkungen in einer gesundheitsfördernden Ernährung," *Bericht 13. Detmolder Studientage*, Granum Verlag, Detmold 1999, pp. 16–20.
28. M.-G. Lindhauer: "Trends und Möglichkeiten bei der Züchtung von Getreide mit gesteigertem Gesundheitswert," *Getreide Mehl und Brot* **53** (1999) 67–69.
29. W. Souci, W. Fachmann, H. Kraut: *Food Composition and Nutrition Tables, Cereal and*

Cereal Products, medpharm, Stuttgart 1994, pp. 499–595.
30. A. M. Shamsuddin, I. Vucenik, K. E. Cole: "IP-6: A Novel Anticancer Agent," *Life Sci*. **61** (1997) 343–54.
31. J. H. Yoon, L. U. Thompson, D. J. A. Jenkins: "The Effect of Phytic Acid on in Vitro Rate of Starch Digestibility and Blood Glucose Response," *Am. J. Clin. Nutr.* **38** (1983) 835–842.
32. J. D. Reed: "Nutritional Toxicology of Tannins and Related Polyphenols in Forage Legumes," *J. Anim. Sci*. **73** (1995) 1516–1528.
33. International Association of Cereal Science and Technology/ICC: *ICC-Standards*, Verlag Moritz Schäfer, Detmold 1998.
34. American Association of Cereal Chemists/AACC: *Approved Methods*, 9th ed., AACC, St. Paul, MN 2000.
35. Official Journal of the European Community 2000.
36. N. W. Childs: "Production and Utilization of Rice," in E. T. Champagne (ed.): *Rice Chemistry and Technology*, 3rd ed., AACC, St. Paul, MN 2004, pp. 1–24.
37. H. C. Wilkinson, E. T. Champagne: "Value-Added Rice Products," in E. T. Champagne, (ed.): *Rice Chemistry and Technology*, 3rd ed., AACC, St. Paul, MN 2004, pp. 473–494.
38. L. A. Johnson, J. B. May: Wet Milling: "The Basis for Corn Bio refineries," in P. J. White, L. A. Johnson (eds.): *Corn: Chemistry and Technology*, 2nd ed., Am. Assoc. Cereal Chem., St. Paul, MN 2003, pp. 449–494.
39. O. K. Chung, S. R. Bean, S. H. Park: "Sorghum foods: New health benefits from an ancient grain," *Food Science* (Chinese Journal, ISSN 1002–6630) **25** (2004) 431–437.
40. Dietary Guidelines for Americans, U.S. Department Health and Human Services and Department of Agriculture, Washington 2005.
41. R. W. Klingler: *Grundlagen der Getreidetechnologie*, Behrs Verlag, Hamburg 1995.
42. P. Erling (ed.): *Handbuch Mehl- und Schälmüllerei*, Agri Media, Bergen/Dumme 2003.
43. J. F. Lockwood: *Flour milling*, Northern Publishing Co., Ltd., Liverpool 1960.
44. E. S. Posner: *Wheat flour milling*, AACC, St. Paul, MN 1997.
45. J. H. Scott: *Flour milling process*, Chapman and Hall Ltd., London 1951.
46. W. Seibel (ed): *Bio-Lebensmittel aus Getreide*, Behr´s Verlag, Hamburg 1990.
47. H. Zwingelberg, W. Seibel, H. Stephan: "Einfluss der Zerkleinerung von Weizen und Roggen auf das Backergebnis von Vollkornmahlerzeugnissen," *Getreide Mehl und Brot* **39** (1985) no. 1, 3–12.
48. W. Klabunde: "Durummühle – Beispiel einher fortschrittlichen Neubaukonzeption," *Mühle und Mischfuttertechnik* **129** (1992) no. 51/52, 736–739.
49. W. Seibel, W. Steller (eds.): *Roggen, Anbau Verarbeitung Markt*, Behr´s, Verlag, Hamburg, 1988.
50. H. Zwingelberg: "Untersuchungen zur Verkürzung der Roggenvermahlung," *Mühle und Mischfuttertechnik* **129** (1992) no. 9, 103–107.
51. W. Bushuk (ed.): *Rye: Production, Chemistry, and Technology*, 2nd ed., AACC, St. Paul, MN, 2001.
52. W. Seibel: "Roggen," in *Handbuch Backwaren*, Behr's Verlag, Hamburg 2002.
53. F. O. Görlitz, J. Oeding: "Ein neues Konzept zur Trennung von Hafer- und Reis- und Schälgemischen," *Mühle und Mischfuttertechnik* **135** (1998) no. 6, 183–184.
54. W. Rohde: "Automation in der Hafermühle," *Mühle und Mischfuttertechnik* **126** (1989) no. 10, 136–138.
55. G. Tegge: *Stärke und Stärkederivate*, Behr´s Verlag, Hamburg 2004.
56. F. Meuser, D. Manners, W. Seibel (eds.): *Progress in Plant Polymeric Carbohydrate Research*, Behr's Verlag, Hamburg 1995.
57. T. Eerikäien, P. Linko, W. Seibel, K. Seiler: "Application des modeles boite noire, des systemes experts et des reseaux neuronaux de la couisson-extrusion," in P. Colonna, G. Della Valle (eds.): *La Cuisson-extrusion Technique et Documentation*, Lavoisier, Paris 1994.
58. Y. Pomeranz (ed.): *Wheat, Chemistry and Technology*, 3rd ed., AACC, St. Paul, MN 1988.
59. Y. Pomeranz (ed.): *Wheat is Unique: Structure, Composition, Processing, End-Use Properties and Products*, AACC, St. Paul, MN, 1989.
60. R. C. Hoseney: *Principles of Cereal Science and Technology*, 2nd ed., AACC, St. Paul, MN, 1994.
61. C. F. Morris, S. P. Rose: "Quality Requirements of Cereal Users – Wheat," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, 3–54.

62. A. C. Eliasson, K. Larsson: "Physicochemical Behaviour of the Components of Wheat Flour," in *Cereals in Breadmaking*, Marcel Dekker, Inc., New York, 1993, pp. 31–159.
63. R. Cubadda, E. Marconi: "Spelt Wheat," in P. S. Belton, J. Taylor (eds.): *Pseudocereals and Less Common Cereals*, Springer, Berlin 2002, pp. 153–173.
64. D. Weipert: "Dinkelweizen," in *Handbuch Backwaren Kap. 3.2.1 Getreiderohstoffe*, Behr's Verlag, Hamburg 2002, 16. Akt.-Lfg. 08/02 1–33.
65. D. Weipert: "Rye and Triticale," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, pp. 205–224.
66. W. Seibel, D. Weipert: "Bread Baking and other Food Uses around the World," in W. Bushuk (ed.): *Rye, Production Chemistry and Technology*, 2nd ed., AACC St. Paul MN 2001, pp. 147–211.
67. M. Nilsson, P. Aaman, H. Härkönnen, G. Hallmans, K. E. B. Knudsen, W. Mazur, H. Adlercreutz: "Content of nutrients and lignans in roller milled fraction of rye," *J. Sci. Food Agric.* **73** (1997), no. 2, 143–148.
68. W. Seibel, D. Weipert: "Animal Feed and Industrial Uses," in W. Bushuk (ed.): *Rye, Production, Chemistry, and Technology*, 2nd ed., AACC, St. Paul, MN, 2001, pp. 213–233.
69. D. Weipert: "Triticale," in *Handbuch Backwaren*, 10. Akt.-Lfg. 7/00, 1–19, Behr's Verlag, Hamburg 2000.
70. M. J. Edney: "Barley," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, pp. 113–151.
71. F. H. Webster: "Oats," in R. J. Henry, P. S. Kettlewell (eds.): *Cereal Grain Quality*, Chapman and Hall, London 1996, pp. 179–203.
72. W. Gansmann in Proceedings of the International Conference: Cereals for Human Health and Preventive Nutrition, Brno 1998, pp. 54–62.
73. N. S. Hettiarchchy, Z. Y. Ju, T. Siebenmorgen, R. N. Sharp: "Rice: Production, Processing, and Utilization," in K. Kulp, J. G. Ponte, Jr. (eds.): *Handbook of Cereal Science and Technology*, 2nd ed., Marcel Dekker, Inc., New York, 2000, pp. 203–221.
74. K. R. Bhattacharya: "Parboiling of Rice," in E. T. Champagne (ed.): *Rice Chemistry and Technology*, 3rd ed., AACC, St. Paul, MN 2004, 329–404.
75. J. R. N. Taylor, P. S. Belton: "Sorghum," in P. Belton, J. Taylor (eds.): *Pseudocereals and Less Common Cereals*, Springer, Berlin 2002, pp. 25–91.
76. L. W. Rooney, S. O. Serna-Saldivar: "Sorghum," in K. Kulp, J. G. Ponte, Jr. (eds.): *Handbook of Cereal Science and Technology*, 2nd ed., Marcel Dekker, Inc., New York 2000, pp. 149–175.
77. S. M. Kaup, C. E. Walker: "Couscous in North Africa," *Cereal Foods World* **31** (1986) 179–182.
78. L. A. Johnson: "Corn: The Major Cereal of the Americas," in K. Kulp, J. G. Ponte, Jr. (eds.): *Handbook of Cereal Science and Technology*, 2nd ed., Marcel Dekker, Inc., New York 2000, pp. 31–80.
79. P. Biacs, E. Aubrecht, J. Leder, J. Lajos: "Buckwheat," in P. Belton, J. Taylor (eds.): *Pseudocereals and Less Common Cereals*, Springer, Berlin 2002, pp. 123–151.
80. I. Kreft: "Buchweizen als Lebensmittel," in H. Weis (ed.): *Das Buchweizen Buch*, Islek ohne Grenzen, Verlag der Sankt Paulus-Druckerei, Luxemburg 1999.
81. W. Seibel: "Buchweizen," in *Handbuch Backwaren*, Behr's Verlag, Hamburg 2002.
82. E. Berghofer, R. Schoenlechner: "Grain Amaranth," in P. Belton, J. Taylor (eds.): *Pseudocereals and Less Common Cereals*, Springer, Berlin 2002, pp. 219–260.
83. J.-M. Brümmer, G. Morgenstern: "Backeigenschaften der Pseudo-Cerealien Amaranth und Quinoa," *Getreide Mehl und Brot* **46** (1992) 78–84.
84. J. R. N. Taylor, M. L. Parker: "Quinoa," in P. Belton, J. Taylor (eds.): *Pseudocereals and Less Common Cereals*, Springer, Berlin 2002, pp. 93–122.

Related Articles
Bread and Other Baked Products and Cereals are separate keywords

Cerium Mischmetal, Cerium Alloys, and Cerium Compounds

Klaus Reinhardt, Th. Goldschmidt AG, Essen, Federal Republic of Germany

Herwig Winkler, Treibacher Chemische Werke AG, Treibach, Austria

1. Introduction 1
2. Cerium Mischmetal 2
2.1. Preparation of the Raw Materials . 2
2.1.1. Wet Chemical Attack 2
2.1.2. High-Temperature Chlorination 3
2.2. Fused-Salt Electrolysis 4
2.2.1. Iron and Graphite Cells 4
2.2.2. Ceramic Cells 5
2.2.3. Oxide Process 6
2.3. Properties 6
2.4. Use of Cerium and Cerium Mischmetal 7
3. Cerium Alloys 8
3.1. Cerium – Iron Alloy – Flint Alloy . . 9
3.2. Cerium – Cobalt Alloys – Permanent Magnets 10
3.3. Cerium – Nickel Alloys – Hydrogen Storage 11
4. Cerium Compounds 12
4.1. Production 12
4.2. Uses 12
5. Analysis 14
6. Economic Aspects 14
7. Toxicology and OccupationalHealth 14
8. References 15

The authors thank I. S. Hirschhorn of Lanthanide Research Corporation, West Orange, New Jersey, United States, previously with Ronson Metals Corporation, for reading the English manuscript and for his helpful suggestions.

1. Introduction

Cerium is the most abundant rare-earth element and exceeds in abundance such well-known elements as tin, cobalt, and lead. The following article treats pure cerium and rare-earth mixtures in which cerium content either exceeds the naturally occurring composition or determines the overall properties or economics. Other rare-earth elements are treated under a separate keyword (→ Rare Earth Elements).

The term *rare-earth elements* includes the elements scandium, yttrium, and the *lanthanides* lanthanum, cerium, praseodymium, neodymium, promethium, samarium, europium, gadolinium, terbium, dysprosium, holmium, erbium, thulium, ytterbium, and lutetium, all of which belong to group 3 of the periodic table (definition of lanthanides → Rare Earth Elements). In this article the symbol R is used for any rare-earth element. The term cerium group, or light rare-earth elements, includes lanthanum through europium inclusively.

The presence of the rare-earth elements lanthanum to lutetium at the position of lanthanum (atomic number 57) in the periodic table is due to the fact that, starting with cerium (Xe $4f^1 5d^1 6s^2$), 14 electrons are able to enter the inner $4f$ orbital. Lanthanum, which has an empty $4f$ level ($4f^0$), represents the stable electronic configuration. This configuration is also attained by tetrapositive cerium, which explains why cerium has not only the normal rare-earth +3 oxidation state, but also a +4 state. This exceptional oxidation state simplifies separation of cerium from the other rare-earth elements.

Cerium was discovered in 1814 by Berzelius and named after the asteroid Ceres, which had been discovered in 1800/1801.

2. Cerium Mischmetal

*Cerium mischmetal*or more simply *mischmetal* (MM) [*8049-20-5*] is a mixture of rare-earth metals of the ceriumgroup with cerium as the major constituent. Mischmetal is the lowest priced rare-earth metal because no expensive

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_139

chemical separation is needed to produce it. It is produced predominantly by fused-salt electrolysis of rare-earth chlorides. Pure rare-earth metals of the cerium group are produced by fused-salt electrolysis of a mixture of chlorides and fluorides. The electrolytic reduction of rare-earth oxides of the cerium group dissolved in a fluoride electrolyte has been put into production. Metallothermic reduction of rare-earth chlorides of fluorides is also employed (→ Rare Earth Elements).

Mischmetal was first produced industrially in 1908 by AUER VON WELSBACH, who succeeded in finding an outlet for surplus rare earth in the production of lighter flints. At that time, monazite was used exclusively as the source of thorium needed for the manufacture of incandescent mantles.

2.1. Preparation of the Raw Materials

Rare-earth minerals in which cerium predominates over lanthanum, praseodymium, and neodymium (the other elements in the cerium group) are *monazite* [*1306-41-8*], $(R,Th)PO_4$, from beachplacers of India, Southeast Asia, Australia, South Africa, and Brazil, associatedwith cassiterite, rutile, ilmenite, and zircon and *bastnaesite* [*1299-84-9*], [*12172-82-6*], $RFCO_3$, from the carbonatitic deposit of the Mountain Pass mine in California andfrom the iron ore deposit of Bayan Obo in Inner Mongolia (People's Republic of China).

Monazite and bastnaesite are generally concentrated by physical methods such as gravity concentration, flotation, and magnetic separation [1, B 1, pp. 60 – 67]. The rare-earth concentrates are converted into rare-earth chloride hydrates by wet chemical methods or into anhydrous rare-earth chlorides by high-temperature chlorination.

2.1.1. Wet Chemical Attack

After being ground (50 µm – 1 mm), the concentrates are attacked by acids or caustic soda [1, B 1, pp. 67 – 104].

Monazite. At temperatures between 140 and 170 °C, concentrated sodium hydroxide solution (50 – 70 wt %) reacts with the phosphates:

$$RPO_4\,(s) + 3\,NaOH \rightleftharpoons R\,(OH)_3\,(s) + Na_3PO_4$$

$$Th_3\,(PO_4)_4\,(s) + 12\,NaOH \rightleftharpoons 3\,Th\,(OH)_4\,(s) + 4\,Na_3PO_4$$

The attack is carried out in an autoclave [13]. More concentrated sodium hydroxide solutions (> 70 wt %) react at lower temperature [14]. The rare-earth hydroxides are dissolved in hydrochloric acid. The thorium does not dissolve if the pH is kept close to 4. Trisodium phosphate [*7601-54-9*] is a saleablebyproduct after recrystallization.

Bastnaesite. The carbonates are attacked by hydrochloric acid. The residue of rare-earth fluorides is converted by alkaline treatment into rare-earth hydroxides, which can be used to neutralize the acid solution resulting from the hydrochloric acid attack [15]. Concentrated sodium hydroxide solution attacks bastnaesite to form rare-earth hydroxides, which are dissolved in hydrochloric acid [16].

$$3\,RFCO_3\,(s) + 6\,HCl \longrightarrow 2\,RCl_3 + RF_3\,(s) + 3\,CO_2 + 3\,H_2O$$

$$RF_3\,(s) + 3\,NaOH \rightleftharpoons R\,(OH)_3\,(s) + 3\,NaF$$

$$RFCO_3\,(s) + 3\,NaOH \rightleftharpoons R\,(OH)_3\,(s) + NaF + Na_2CO_3$$

$$R\,(OH)_3\,(s) + 3\,HCl \longrightarrow RCl_3 + 3\,H_2O$$

The rare-earth chloride solution obtained by these various processes is evaporated to form the hexahydrates (*mp* 120 – 130 °C), which solidify on cooling. The rare-earth chloride hexahydrates must be dehydrated for fused-salt electrolysis. Dehydration is carried out under conditions that avoid excessive formation of rare-earth oxide chlorides by hydrolysis:

$$RCl_3 \cdot 6\,H_2O \rightleftharpoons ROCl + 2\,HCl + 5\,H_2O$$

The dehydration is carried out either in vacuo or by adding salts like $CaCl_2$, NaCl, or NH_4Cl that reduce hydrolysis.

Oxide chlorides increase the viscosity of the electrolyte and cause reoxidation of the rare-earth metals that are formed. However, rare-earth oxide chloride content near 5 wt % is not detrimental, perhaps even desirable, for one special type of electrolytic cell, the ceramic cell. Dehydration can then be carried out in metal ovens or on heated rotating steel plates, at temperatures between 170 and 600 °C, depending on the process. In modern plants, spray drying

and fluidized-bed drying are employed for the dehydration process. Special materials like titanium and Hastelloy can be used to avoid corrosion by the hydrochloric acid generated in the off-gas during drying.

2.1.2. High-Temperature Chlorination

The chlorination of rare-earth ores [17] in a shaft furnace at 900 – 1100 °C produces anhydrous rare-earth chlorides that can be used directly for fused-salt electrolysis. The reaction proceeds as follows:

$$3\,RFCO_3 + 6\,C + 3\,Cl_2 \longrightarrow 2\,RCl_3 + RF_3 + 9\,CO$$
$$RPO_4 + 3\,C + 3\,Cl_2 \longrightarrow RCl_3 + POCl_3 + 3\,CO$$

Bastnaesite yields a mixture of rare-earth chlorides and fluorides; monazite yields the chlorides. In the case of monazite, the radioactivity, caused mainly by radium, a disintegration product of thorium, remains with the rare-earth chlorides. In addition, $ThCl_4$ must be recovered from the off-gas to avoid environmental pollution. Therefore, thorium-free monazite should be used if available. The recovery of $POCl_3$ [*10025-87-3*] is an economic necessity.

Typical analyses of rare-earth chlorides from various sources are listed in Table 1.

2.2. Fused-Salt Electrolysis

For proprietary reasons, there are no modern detailed publications on the process technology for production of mischmetal by fused-salt electrolysis. The know-how of the producers covers optimization of electrolytic cell construction for continuous operations, special refractories resistant to the aggressive salt and metal melt, and maximization of current efficiency by the choice of the composition of the electrolyte, with a view to the viscosity, temperature, oxygen content, and solubility of deposited metal in the melt.

The kind of electrolytic-cell refractory characterizes the two main processes for the production of mischmetal: (1) iron/graphite and (2) ceramic [18], [19].

Independent of the type of cell, the amperage, voltage, electrode distance, and current density are chosen to reach the reduction potentials of the rare-earth metals and to avoid an excessive electrolyte temperature, which may cause vaporization of the electrolyte, attack of the refractory lining by the electrolyte and metal, and formation of metal fogs.

Cerium deposits at a lower voltage than praseodymium, neodymium, and especially lanthanum. Samarium and europium are reduced at the cathode only to the +2 oxidation state. Stirring by the chlorine liberated at the anode or diffusion allows both elements to migrate to the anode, where they are reoxidized [20]. Metal fogs reduce current efficiency similarly.

Table 1. Typical analyses of rare-earth chlorides from various sources

Element, as oxide	$RCl_3 \cdot 6\,H_2O$ from wet chemical treatment, wt % *						RCl_3 from high-temperature chlorination, wt % *
Brazil	China	France	India	Malaysia	United States		Germany, Fed. Rep.
R_2O_3	45.8	45.8	47	46	46.1	46	65 – 67
RF_3							6 – 10
CeO_2/TRO **	47.2	50.0	49	46	50.3	50	48 – 50
Na_2O				1	0.66		< 0.2
MgO				1		0.2	< 0.3
CaO	0.63		0.5	1	0.59	0.4	< 2
BaO						< 0.1	< 3
Fe_2O_3	0.03			0.005		0.02	< 0.1
PbO	100 ppm	22 ppm	230 ppm	10 ppm	15 ppm		< 50 ppm
P_2O_5	trace	trace	0.02	0.01			< 0.2
SO_3	0.1	0.06	0.14	0.05			< 0.1
SiO_2				0.05			< 0.1

* Unless otherwise stated.
** The cerium(IV) oxide content divided by the total rare-earth oxide content (see 5).

Added rare-earth fluorides and alkali and alkaline-earth chlorides and fluorides having decomposition voltages higher than those of the rare-earth chlorides lower viscosity, increase the conductivity, and improve the metal yield.

2.2.1. Iron and Graphite Cells

The iron or graphite crucible or pot serves directly as the cathode. Iron crucibles made of spheroidal cast iron are directly connected as the cathode, whereas graphite crucibles, which are protected by an iron housing against air oxidation, are connected via this iron housing.

Graphite-lined iron pots are connected directly. Crucibles up to an electrolyte volume of 50 L are round; larger crucibles are rectangular. The schematic design of a rectangular graphite-lined iron cell is shown in Figure 1. The anodes consist of graphite or special carbon and can be moved vertically to adjust the amperage and current density.

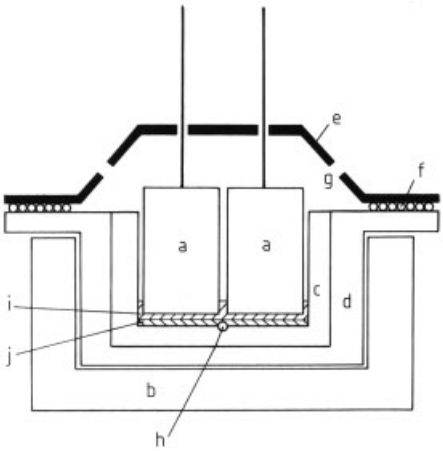

Figure 1. Schematic design of a rectangular graphite-lined iron cell (Th. Goldschmidt AG)
a) Graphite anodes; b) Insulated housing; c) Graphite lining (cathode); d) Iron crucible; e) Hood; f) Seal; g) Feeding chute and chlorine suction; h) Bottom tap; i) Electrolyte melt; j) Mischmetal deposit

The electrolysis is started via short-circuit heating by melting rare-earth chlorides and mischmetal down until a sufficient amount of melt is obtained for electrolysis. Already molten chloride, for example, from the chlorination process, also may be used. Fresh rare-earth chlorides are fed as solids.

As a result of the higher heat losses by the iron and graphite cathode, the so-called cooled cathode, the mischmetal is electrodeposited in the form of dendritic sponge or droplets, together with solidified electrolyte, at the bottom of the cell. The deposition of solid mischmetal diminishes the damage to the refractory lining.

Toward the end of the electrolysis, the metal is melted by increasing voltage and amperage or by external heating via gas or resistance heating. Before tapping, the temperature of the metal must be 850 – 950 °C. The temperature of the electrolyte at the surface may be 100 °C higher. Addition of iron scrap lowers the melting point, making the melt easier to handle. The total content of the cell is cast into molds by tipping or bottom pouring; reoxidation of the molten metal is prevented by the residual electrolyte. The mischmetal settles to the bottom of the mold and is covered by the electrolyte, which can be removed after solidification with an air chisel. The separation of the residual electrolyte from the liquid metal can be accomplished via a mold that acts as part of a siphon: the metal collects at the bottom, while the electrolyte overflows. Mischmetal may be recast more cleanly.

Immediately after tapping, the electrolysis is restarted by feeding fresh rare-earth chlorides, which melt instantly in the hot cell.

The residual electrolyte, in which europium and samarium are enriched, especially when rare-earth fluorides are involved, may be recycled for further enrichment and later used for winning those two elements [21].

The cells are covered with a hood. The chlorine formed at the anode is removed from the exhaust gases by alkaline scrubbing: the resulting hypochlorite is either reduced to chloride or sold. Recovery of the chlorine itself is normally uneconomical because of dilution by air.

In modern plants, direct current is supplied by thyristor-controlled silicon rectifiers; in older plants, it is supplied by transformer-controlled selenium rectifiers. The rectifiers are connected in parallel, series, or individually to the cells. Cell voltage ranges between 6 and 15 V, depending on the radiative loss of heat. The amperage may be as high as 50 kA. The current density depends on the cell design: in large rectangular cells it is $\leq 2\ A/cm^2$, in smaller ones it is much higher. The current efficiency can reach 75 %.

From 2 to 2.5 kg of rare-earth chlorides are consumed to produce 1 kg of mischmetal. The metal yield is between 85 and 95 %, depending on the extent the residues are recycled.

For a 50-kA cell, the daily production is nearly 1 t. The d.c. power consumption per kilogram of mischmetal produced is 10 – 15 kW h.

2.2.2. Ceramic Cells

Figure 2 shows the construction of a ceramic cell [19]. Electrolysis takes place in a ceramic crucible. The crucibles of small, simple cells have a diameter of ca. 30 cm and a height of ca. 50 cm. Further development resulted in the bigger ceramic cells now in operation.

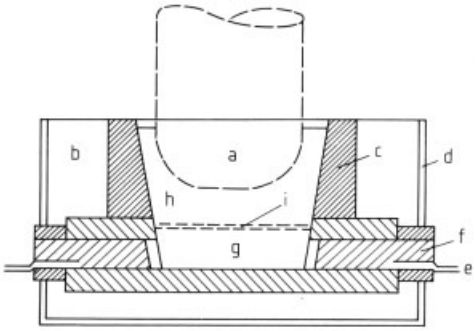

Figure 2. Ceramic cell [19]
a) Graphite anode; b) Insulation; c) Ceramic crucible; d) Iron housing; e) Current supply for cathode; f) Cathode; g) Mischmetal; h) Electrolyte melt; i) Rare-earth oxide chloride diaphragm

A block of graphite or iron at the bottom of the crucible serves as the cathode. If iron cathodes are used, contamination of mischmetal with iron cannot be avoided. Mischmetal at the electrolysis temperature of ca. 900 °C attacks most of the construction materials in the electrolytic cell. The working life of the ceramic cells is limited to a maximum of 1 year.

The current density and voltage are controlled by raising or lowering an anode of amorphous carbon. Service life of the anode is between 30 and 50 d.

The electrolytic process in ceramic cells is started by short-circuit heating. The cells are filled with anhydrous rare-earth chlorides and short-circuited with an iron or carbon rod. If molten electrolyte from other cells is available, mischmetal lumps are placed in the melting pots and covered with melt. Electrolysis can be started immediately. The ceramic cells operate continuously. Anhydrous rare-earth chlorides are added as required.

Between metal at the bottom of the cell and the electrolyte is built up a diaphragm of rare-earth oxide chlorides, which operates as an electrical resistance, generating the heat that keeps the deposited metal in liquid form.

One or two times a day, molten mischmetal is tapped, sucked, or ladled from the bottom of the crucible.

During the process the anode is consumed by oxidation. For this reason, the electrode must be adjusted from time to time. Oxidation results from contact with atmospheric oxygen or the chlorine produced or from reaction with oxide chlorides in the electrolyte. Therefore, a high oxide chloride content in the rare-earth chlorides causes considerable anode consumption.

The amount of molten electrolyte increases during operation. Enrichment of europium and samarium takes place. To maintain a constant level of electrolyte, a part, the so-called slag, is ladled out at regular intervals. This slag can be treated chemically to recover Eu_2O_3 and Sm_2O_3. A typical composition of slag is presented in Table 2.

Table 2. Analyses of mischmetal slag, Treibacher Chemische Werke AG, Austria

Element, as oxide	Relative content, wt %∗
La_2O_3	26.6
CeO_2	31.9
Pr_6O_{11}	4.1
Nd_2O_3	19.8
Sm_2O_3	14.3
Eu_2O_3	0.4
Total	*97.1*

Total rare-earth oxide content (TRO) is 41.9 wt %

∗ The element oxide content divided by total rare-earth oxide content (see chap. 5).

2.2.3. Oxide Process

On the basis of a process developed by the U.S. Bureau of Mines [22], [23] and further improved by Santoku Metal Industries, Japan, Santoku now produces mischmetal from rare-earth oxides. The process is similar to aluminum electrolysis (Hall – Héroult process). It avoids emission of chlorine and the consequent expensive purification of off-gas.

Rare-earth oxides are dissolved in an electrolyte consisting of alkali fluorides (to improve conductivity), alkaline-earth fluorides (to reduce melting point), and rare-earth fluorides (to improve the solubility of rare-earth oxides) and are reduced electrolytically to the rare-earth metals. The electrolysis cell (Fig. 3) consists of a graphite crucible with graphite anode and molybdenum cathode, working under an inert-gas atmosphere to prevent oxidation of the construction materials. The electrolyte is initially melted by resistance heating. Electrolyte and electrodeposited rare-earth metals are kept liquid by the joule heat. A cell produces ca. 500 kg of metal per day. If bastnaesite is a raw material, it must be specially purified prior to electrolysis.

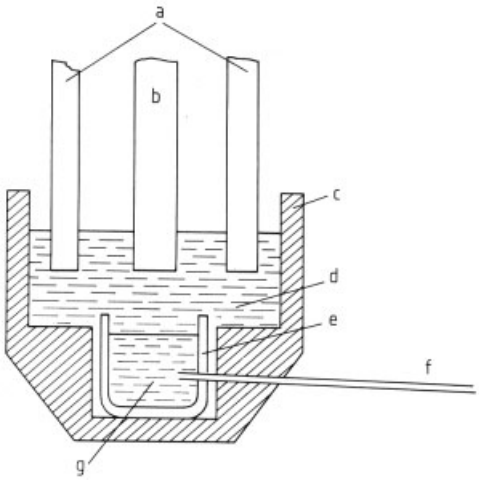

Figure 3. Electrolysis cell for the production of mischmetal from oxides [23]
a) Graphite anodes; b) Molybdenum cathode; c) Graphite crucible; d) Liquid electrolyte; e) Molybdenum crucible; f) Tapping pipe; g) Molten mischmetal

2.3. Properties

Table 3 shows typical analyses of mischmetal from various producers.

The physical and chemical properties of mischmetal are determined by the properties of its four main constituents – cerium, lanthanum, neodymium, and praseodymium [1], [24]. There can be no exact values for mischmetal because the composition depends on the type of raw material and treatment. However, the values correspond most closely to those of the main constituent, cerium. Table 4 lists several properties for the four main rare-earth metals.

Mischmetal is ductile. The freshly cut surface has a metallic gray appearance. In air, the surface oxidizes to form yellow to greenish-gray rare-earth hydroxide carbonates or oxide hydrates. Some types of oil may protect the surface against corrosion, but only for a limited period of time. Alloying with 1 – 2 wt % of magnesium increases corrosion resistance. Massive

Table 3. Typical analyses of mischmetal from various producers, weight percent unless otherwise stated

Metal	Producer						
	Treibacher (Austria)	Ronson (United States)	Santoku (Japan)	Corona (Brazil)	Fluminense (Brazil)	Former (USSR)	Minmetals (China)
Total rare-earth metals	99.5	> 95	97.8	97.4	96.0	97	98.6
Relative content∗							
La	23	18 – 28	22.9	17.5	17.9	25.1	19.6
Ce	49.8	50 – 55	51.9	54.2	55.3	53.2	54.9
Pr	5.6	4 – 6	5.4	5.6	7.3	5.2	6.1
Nd	18.6	12 – 18	16.2	16.0	17.4	14.0	17.1
Sm	0.2	< 0.1					0.4
Others	2.8	< 2	3.6	6.7	4.8	2.8	1.9
Al	0.05		0.14			0.05	0.1
Pb	50 ppm		343 ppm		90 ppm	227 ppm	4600 ppm
Fe	0.05	< 5	1	2.0	2.3	1.7	0.28
Si	0.20			0.1	0.54	0.1	0.05
Mg	0.03	< 0.5	0.20	0.20	0.2	0.033	0.35
Ca		< 0.1		0.02	0.1		0.013

∗ The content of rare-earth metal divided by total rare-earth metals.

Table 4. Properties of lanthanum, cerium, praseodymium, and neodymium [1], [2], [4]

Property	Unit	La	Ce	Pr	Nd
Atomic number z		57	58	59	60
Atomic mass A_r		138.9055	140.115	140.90765	144.24
Crystal structure at 293 K		dhcp ∗	fcc	dhcp ∗	dhcp∗
a_0	nm	0.3774	0.51612	0.36725	0.36579
c_0	nm	1.2159		1.18354	1.17992
Density ϱ	g/cm^{-3}	6.146	6.770	6.773	7.008
Melting point *mp*	°C	918	798	931	1021
Heat of fusion ΔH	kJ/mol^{-1}	6.2	5.5	6.9	7.1
Boiling point *bp*	°C	3464	3433	3520	3074
Temperature at a vapor pressure of					
133 Pa	°C	2208	2174	1968	1741
0.133 Pa		1581	1581	1360	1197
Specific heat at 298 K	$J\,mol^{-1}\,K^{-1}$	26.2	27.0	27.0	27.5
Coefficient of linear thermal expansion at 298 K	$10^{-6}\,K^{-1}$	4.9	8.5	6.79	9.98
Thermal conductivity at 300 K	$J\,s^{-1}\,cm^{-1}\,K^{-1}$	0.135	0.114	0.125	0.165
Electrical resistivity at 273 K	μΩ cm	61	77	71	64

∗ Double hexagonal closest packed in the direction of the *c* axis.

metal burns above 150 °C in pure oxygen; however, chips, turnings, and powder burn at this temperature even in air. Mischmetal dissolves in dilute mineral acids with evolution of hydrogen.

The ingots may be easily machined. However, the pyrophoric character of mischmetal may cause autoignition of the turnings. Mischmetal may be easily extruded at temperatures just below its melting point. Rollability is adversely effected by oxygen content. Nearly all of the alloying elements cause brittleness.

2.4. Use of Cerium and Cerium Mischmetal

The most important uses of mischmetal or cerium are metallurgical. The metallurgical importance of rare-earth metals is based on reactions to form solids with oxygen, hydrogen, nitrogen, sulfur, arsenic, bismuth, and antimony, reducing the effects of these elements on the properties of the metals [25]. To avoid the formation of harmful rare-earth oxide inclusions by secondary reactions of the mischmetal with refractories, slag, and atmospheric oxygen, the mischmetal is plunged into the molten metal or added under an inert-gas atmosphere. Metallurgical mischmetal can also be clad with aluminum or steel. Mischmetal is added as lumps, rods, doughnuts, or wire [26].

Iron and Steel [27]. Rare-earth sulfides and oxide sulfides are formed in liquid steel and precipitate as globular particles uniformly dispersed, which, unlike manganese sulfides, are not deformable during rolling and do not form stringers. This is called sulfide shape control and is used in *microalloyed* or *HSLA* (high-strength low-alloy) *steels* to reduce anisotropy in toughness, notch toughness, and bend formability. This is especially important for pipeline steels used at subzero temperature in the Arctic.

Mischmetal may entrap hydrogen and diminish hydrogen-induced cracking.

Hot shortness in *stainless steels* can be reduced by removal of tramp elements (As, Bi, Sb) together with deoxidation and sulfide shape control. Heat and oxidation resistance can be increased by mischmetal, which forms protective surface layers of rare-earth oxides together with oxides of the steel components, for example, Cr_2O_3, that are resistant to scaling.

Mischmetal or cerium-containing master alloys are added to *cast iron* to improve ductility, toughness, and the microstructure. Cerium allows graphite to form nodules, causing nucleation in spheroidal and vermicular cast iron, and neutralizes the harmful effect of the tramp elements [28].

The addition rate depends on the application and preparation of the steel or iron melt. For sulfide shape control and nodularization of graphite, up to 1 kg of mischmetal is added per tonne. Other effects require even larger additions, up to 8 kg per tonne.

Nonferrous Metals [29]. Addition of mischmetal to copper alloys improves tensile strength and deep-drawing properties. The heat resistance and ductility of aluminum conductor cables are improved without any significant decrease in electrical conductivity. Titanium alloys show a higher grade of grain refinement, better mechanical properties, and improved corrosion resistance as a result of mischmetal additions.

The need for improved galvanizing compositions (increased corrosion resistance, fluidity, wettability, and freedom from intergranular corrosion) without affecting formability, weldability, and paintability led to the development of Galfan, which is the classical zinc – aluminum eutectic alloy (95 wt % Zn, 5 wt % Al) with 0.05 wt % mischmetal [30].

In nickel- and cobalt-based superalloys for turbine engines, cerium (yttrium is even better) increases oxidation and sulfidation resistance at high temperature. Similar effects are achieved in chromium-based alloys, and by removing gases cerium prevents embrittlement of niobium- and tantalum-based alloys.

Other Uses. Cerium and mischmetal are used as getters to absorb traces of gases in evacuated devices.

Mischmetal is said to increase the efficiency of fuel consumption and to decrease CO and NO_x contents in the exhaust of internal combustion engines if steam is passed through mischmetal spirals before injection into the carburetor [31].

In some of these uses, mischmetal (= MM) can be replaced by CeSi [*12014-84-5*], MMSi, or MMFeSi alloys produced directly frombastnaesite by reduction in an arc furnace.

The pyrophoric character of mischmetal alloyed with iron and magnesium is used for flints (see Section 3.1) and pyrotechnics.

3. Cerium Alloys

The atomic radii of lanthanum, cerium, praseodymium, and neodymium differ only slightly so that miscibility is complete in the liquid state and close to the solidus curve. The liquidus curves correspond to those of ideal mixtures.

In the molten state the main constituents of mischmetal are completely miscible with nearly all non-rare-earth metals. Their large atomic radii and their low electronegativities are the reasons why there are only a few cases of their solid solubility in other metals. One such case, however, is the limited solid solubility of mischmetal in magnesium [32]. As solids they do not dissolve other metals; instead, they form intermetallic phases. With elements of the groups 4, 5, and 6 they do not form intermetallic phases [32], [33].

The intermetallic phases, with a degree of heteropolar bonding, may be brittle. Therefore, rare-earth metals and alloys are unsuitable as construction materials. Cerium phase diagrams are generally used in place of mischmetal phase diagrams [32], [34], [35].

Depending on the purity desired, alloying is done in crucibles made of tantalum, molybdenum, boron nitride, graphite, or clay graphite. An inert-gas atmosphere, vacuum, or inert slag cover are necessary.

3.1. Cerium – Iron Alloy – Flint Alloy

The cerium – iron phase diagram is shown in Figure 4.

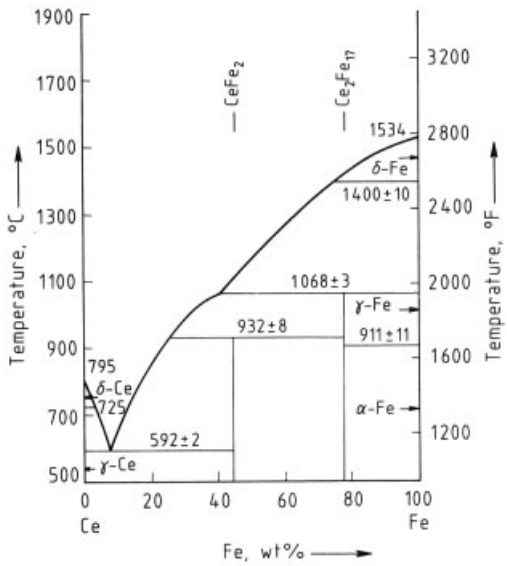

Figure 4. Ce – Fe phase diagram [35]

Pyrophoric Properties. None of the numerous developments of cerium-free pyrophoric alloys of the last decades has been able to replace the cerium – iron alloy invented and patented by Auer von Welsbach in 1903. Lighter flint alloy consists basically of mischmetal and iron. Some other metals are added in small amounts to modify the pyrophoric properties and to improve processing. Typical analyses of lighter flint alloys are shown in Table 5.

The frictional pyrophoric properties of cerium – iron are based on a combination of microstructure and mechanical and chemical properties of the alloy. The typical crystal structure of commercial alloy consists of tough, brittle primary crystals of Ce_2Fe_{17} enclosed by a peritectic layer of $CeFe_2$ embedded in a soft matrix of $CeFe_2$ and Ce (Fig. 4).

Mechanical friction generates primary cracks in the intermetallic compound, initially leading small particles with adherent matrix to break away. Residual frictional and deformation heat in the small particles heats them adiabatically to ignition temperature. On ignition, these particles burn totally. Heat is emitted predominantly by radiation, particle temperature reaching ca. 2000 °C. The size of these particles, which depends on the hardness of the lighter flint, influences the ignition behavior. Smaller particles ignite rapidly, resulting in quick adiabatic heating, whereas larger particles emit a higher local output of reaction heat.

Magnesium and cerium form an intermetallic compound that ensures a high heat of combustion. The temperature of the spark is also increased by alloying with aluminum, which in addition produces a spiky spark. Zinc and copper improve fluidity during extrusion.

Manufacture of Lighter Flints. Lighter flint alloy is produced by melting mischmetal, iron, and the other metals in clay graphite or other suitable crucibles heated in induction furnaces at 1000 – 1200 °C under a protective salt layer ($CaCl_2$, $BaCl_2$). After removal of the salt layer, the alloy is cast into preheated molds.

For better shape control of the $CeFe_2$ crystallites, the melt is cooled slowly. The as-cast products are billets of 3 – 6 cm diameter. Average yield of alloy, based on raw material input, is 95 – 98 %. Losses arise from air oxidation and reaction with the crucible material.

After preheating at 400 – 500 °C, the flint billets are extruded to rods on horizontal or vertical extrusion presses at pressures between 20 and 50 MPa. Alloying with aluminum, magnesium, zirconium, titanium, and copper lowers the amount of pressure required. Extrusion pressure is also affected by impurities in the mischmetal, the rare-earth distribution, and the iron content of the alloy. During extrusion, the eutectic matrix liquifies. After extrusion, the rods are cooled rapidly in air or cooling oil. The diameter of the rods usually ranges between 2 and 6 mm.

Table 5. Analyses of lighter flints of different producers, weight percent

Metal	Producer			
	Treibacher (Austria)	Santoku (Japan)	Electro Centre (France)	Ronson (United States)
Total rare-earth metals	76.2	78.3	76.7	77.0
Relative content∗				
Ce	51	52.3	53.3	51.5
La	22.7	24.7	19.2	23.1
Nd	16.6	14.3	16.5	16.5
Pr	5.4	5.0	6.4	5.8
Sm	0.3	0.4	0.4	0.3
Fe	20.6	19.3	20.4	20.0
Mg	2.2	1.8	2.1	1.5
Al	0.09	0.28	0.04	0.05
Cu		0.02	0.03	
Si	0.22	0.29	0.22	0.2
Zn	0.54	0.03	0.34	0.5
C	0.02	0.08	0.06	0.01

∗ The content of rare-earth metal divided by total rare-earth metals.

These rods are cut on horizontal or vertical dies to the length of the lighter flint, usually between 4 and 12 mm. A subsequent heat treatment between 350 and 450 °C improves the pyrophoric properties considerably [36] by conversion of Ce – Fe mixed crystals, formed through rapid solidification during extrusion, into $CeFe_2$ and Ce.

The lighter flints are sized on sorting machines, generally with sorting sieves. These mechanically presorted flints are checked carefully by mechanical and optical sorting machines. Some producers make a final inspection by eye: flints are moved over a conveyor belt and imperfect flints are picked out by hand. Flints are checked for length, diameter, and straightness, as well as to see if the area of cut is circular as desired and if cavities or burrs are absent from the cut surface.

Durability of lighter flints is limited to 4 – 5 years normally. Humid air and a warm climate, such as prevails in the tropics, cause quick decomposition of the flint, visible as a dusting at the beginning of the decomposition process. For protection of the flint surface against corrosion, flints are normally coated and colored with lacquer.

Generally, there is a distinction between the normal lighter flints (most commonly 2.4 mm in diameter and 4 mm in length) and the lighter flints for disposable lighters. Flints for disposable lighters are manufactured in a length from 7 to 12 mm and a diameter from 2.3 to 2.5 mm, depending on the amount of gas filling in the lighter. Such flints should ignite at least 1000 times before 50 % of the flint is used up.

Accuracy of size is important for flints for disposable lighters because filling is carried out by machines and each faulty flint stops production. Maximum tolerances of ± 0.1 mm in length and in diameter are acceptable. One kilogram of lighter flints contains from 3000 to 10 000 flints.

3.2. Cerium – Cobalt Alloys – Permanent Magnets (→ Magnetic Materials)

The cerium – cobalt phase diagram is shown in Figure 5.

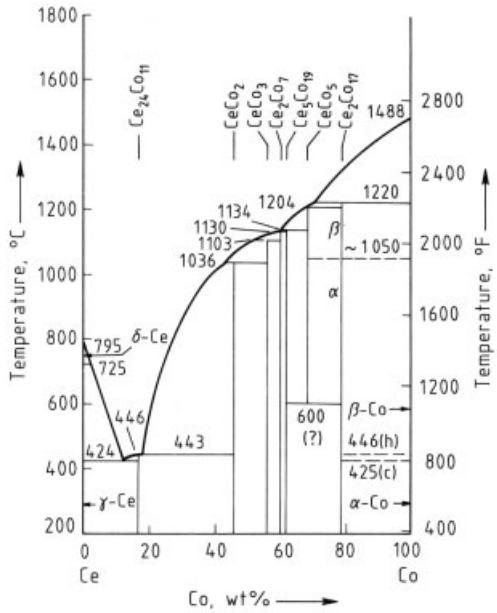

Figure 5. Ce – Co phase diagram [35]

The classic RCo_5 [*12017-68-4*] permanent magnet alloys are based on samarium, which gives the best hard magnetic properties. Replacement of samarium, an expensive rare-earth metal, by cerium or even by inexpensive mischmetal reduces the hard magnetic properties, especially because of the lower crystal anisotropy field [37]. However, adding small amounts of samarium to the mischmetal and/or changing the composition of the mischmetal [38] or partial replacement of cobalt [39], [40] makes magnets with sufficient coercivity and energy product for less sophisticated uses and increases the range of uses because of the lower costs.

These alloys are manufactured by melting the metals or by calciothermic reduction of oxide and metal-powder mixtures [41].

These magnets are used for electronic watches, microwave amplifier tubes, loudspeakers, headphones and microphones, magnetic coupling, bearings, sensors, clamp systems, switches, d.c. motors, servomotors, and step motors, mainly to achieve miniaturization.

Some magnetic properties are listed in Table 6.

Table 6. Intrinsic magnetic properties of RCo_5 compounds at room temperature [40]

Property	$SmCo_5$	$CeCo_5$	$MMCo_5$ *
Saturation magnetization B_s, T	1.07	0.77	0.89
Anisotropy constant K, 10^6 J/m^3	9.6	5.6	6.6
Anisotropy field H_A, MA/m	20	15.1	14.7
Curie temperature T_C, °C	724	347	520
Theoretical energy product $B_s^2/4$, kJ/m^3	228	118	158

* The values for $MMCO_5$ are only approximate, for composition varies.

3.3. Cerium – Nickel Alloys – Hydrogen Storage

The cerium – nickel phase diagram is shown in Figure 6.

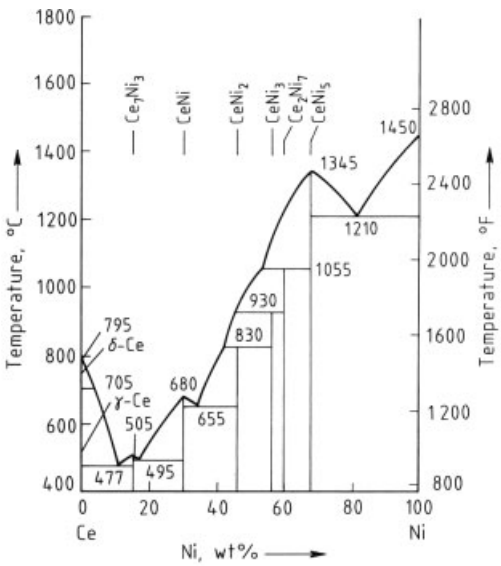

Figure 6. Ce – Ni phase diagram [35]

$LaNi_5$ [*12196-72-4*] has a high storage capacityfor hydrogen gas [42], up to the formula $LaNi_5H_{6.7}$. To decrease costs, lanthanum can be replaced by mischmetal [43]. However, to lower the higher hydrogen equilibrium pressure and form hydrides up to the formula $MMNi_5H_{6.3}$, the nickel must be partially replaced [44] by aluminum [45], calcium [46], cobalt [47], chromium [48], copper [49], iron [50], or manganese [51]. Pressure – composition isotherms are shown in Figure 7. Alloys of the type $CeMg_{11}M$ (M = Ni, Cu, Zn) also absorb hydrogen [52].

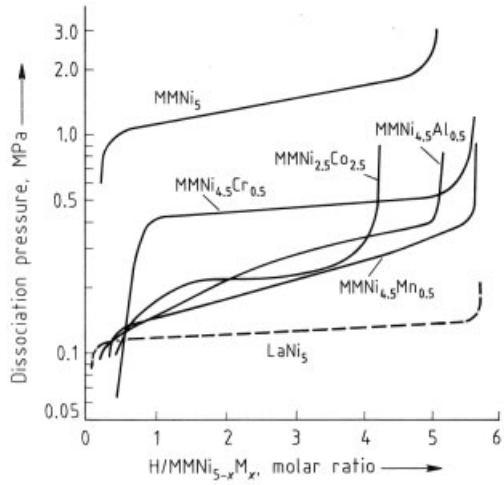

Figure 7. Pressure – composition isotherms for desorption in the $LaNi_5$–H, $MMNi_5$–H, and modified $MMNi_{5-x}M_x$ – H systems at 293 K [44] MM represents mischmetal, and M represents a metal (here Al, Co, Cr, or Mn).

These systems may be used for hydrogen transportation and storage; vehicular and stationary fuel supply, an ideal fuel from the ecological point of view; air conditioning; refrigeration; heat pumps; storing and regaining waste heat; and hydrogen-storage electrodes in alkaline batteries. These systems may also be used for purification [53] and isolation [54] of hydrogen and H_2/D_2 separation.

4. Cerium Compounds

Cerium is the most abundant rare-earth element. Nearly 50 % of the available rare-earth raw material consists of cerium oxide. Cerium is also the only rare-earth element that can be easily separated from a mixture of rare-earth elements by simple chemical methods, for cerium has a tetravalent state.

4.1. Production

Cerium is separated from other rare-earth elements by oxidation of solutions resulting from attack of bastnaesite or monazite. The oxidizing agent can be H_2O_2, hypochlorite, or atmospheric oxygen, or anodic oxidation can be used. Cerium precipitates as cerium(IV) oxide hydrate [*63394-44-5*], [*67285-52-3*].

For example, cerium(III) rare-earth oxide hydrate is dried and oxidized by air to cerium(IV) rare-earth oxide hydrate. The oxidized hydrate is dissolved in nitric acid, and the solutions are neutralized slowly. Cerium(IV) oxide hydrate is collected by filtration. Initially, of the total rare-earth oxides 45 – 50 wt % is CeO_2. This is increased to 95 wt % in the product.

In a process of Molycorp [55], ground bastnaesite ore concentrates are roasted, and cerium is oxidized with atmospheric oxygen. Roasted material is treated with hydrochloric acid to dissolve lanthanum, neodymium, praseodymium, etc. The pH is adjusted to 4, and a cerium-rich residue ($\approx$ 90 wt % CeO_2) is collected by filtration. This residue has high fluoride content and poor solubility. To get soluble cerium salts out of this material a subsequent digestion with sulfuric acid or caustic soda is necessary.

Cerium salts are produced by liquid – liquid extraction from rare-earth –cerium-containing solutions (→ Rare Earth Elements). Cerium can be extracted out of cerium nitrate – nitric acid solutions in a few steps in the form of a cerium(IV) nitrate complex in tributyl phosphate and therefore separated from the accompanying trivalent rare-earth elements, which form less stable nitrate complexes. Purities of 99.99 % and better can be achieved easily.

4.2. Uses

The principal uses for cerium compounds are as polishing agents and as a component in glass.

Cerium Oxide Polishing Compounds. Cerium-(IV) oxide [*1306-38-3*], CeO_2, has now replaced other polishing oxides like iron oxide (red rouge), silica(white rouge), and zirconium dioxide almost completely. The special merits of cerium oxide are its 100 % faster polishing speed and its cleanliness.

Cerium oxide is used for polishing glass mirrors, plate glass, television tubes, ophthalmic lenses, and precision optics. However, the advent of the Pilkington float process (→ Glass) in the early 1970s significantly reduced the use of cerium oxide in the manufacture of plate glass.

Cerium oxide polishing powder is produced by calcining oxidic cerium mineral concentrates in rotary kilns or other furnaces at temperatures of ca. 1000 °C. Calcined concentrates are milled and cleaned by sieving or sifting afterwards. The impurities (CaO, SrO, non-rare-earth elements) must not exceed a certain level in the raw material; otherwise, cerium oxide will agglomerate during calcining and lose polishing capacity.

Polishing powders production can also start from cerium salt solutions. Cerium carbonates or hydroxides are precipitated and then calcined in rotary kilns or muffle furnaces to oxide at temperatures of ca. 1000 °C. The calcined oxides are milled and screened or sifted to get the desired grain size distribution and to remove scratching impurities.

Polishing powders made from mineral concentrates polish more slowly and less cleanly than precipitated products. The former are cheaper but are not used for high-performance polishing processes (precision optics).

Additives improve suspension properties of polishing powders in aqueous solutions and increase polishing rate, for instance, preventing foaming (antifoaming) and settling of cerium oxide (anticaking) in tanks and pipes.

Pure cerium oxide is yellowish white. Small amounts of Pr_6O_{11} in combination with other rare-earth oxides give a brown oxide. The color does not appear to affect the polishing properties for fixed CeO_2 content. Standard concentrations are 50, 70, 90, and 100 %. The polishing rate increases with CeO_2 content. The average grain size of polishing powders is normally 0.5 – 5 µm. A narrow grain size distribution is advantageous.

There are many theories about the mechanism of the polishing process [56]. According to these theories, both a chemical and a physical component, among others, are effective during polishing [57]. The glass surface is hydrolyzed by reaction with water (chemical theory) and the silica gel layer formed is removed mechanically by the polishing compound. Another theory states that glass is removed mechanically

(wear theory) and to some extent by chemical reaction [58].

Cerium Oxide as a Glass Constituent. Cerium oxide can be used to decolorize soda – lime glasses for bottles, jars, etc. [59]. The Ce^{4+} oxidizes Fe^{2+} impurities, which are always present in the raw materials and therefore in glasses, to Fe^{3+}. A change from the blue-green color of Fe^{2+} to the 10 times weaker yellow Fe^{3+} takes place. Arsenic, zinc selenite, or manganese are used for the same purpose. The combination of these materials with cerium reduces costs. Presently, cerium oxide has been replaced by less expensive decolorizers.

Cerium protects glass against solarization and browning, a discoloration caused by irradiation. It is therefore a constituent of glass for the faceplates of television screens, which are under constant bombardment by an electron beam [60]. Radiation-shielding windows for nuclear and radiochemical uses consist of lead glass stabilized by CeO_2. Since maximum transmission is necessary for these thick windows, only 99.99 % cerium oxide (15 ppm Fe max.) can be used.

For phototropic glasses, as in phototropic eye glasses, windshields, and window glasses that darken in the sunlight and lighten in the shade, cerium is a sensitizer.

Other Uses. Cerium rare-earth fluorides improve the brightness of carbon arcs. Cerium fluoride oxide mixtures are mixed with carbon for electrodes. These electrodes are used on movie sets and as military searchlights to increase brightness as much as tenfold [61]. Cerium oxide (99.9 %) is used as white pigment in enamels for tiles.

Basic oil-soluble cerium salts of organic acids, such as cerium alkylsulfonates, alkyl sulfates, and alkyl phosphates, as well as cerium octoates, serve as driers in paints and varnishes. Cerium compounds that are soluble in organic liquids find use as combustion additives in fuels [62]. Particle emission in exhaust gases is reduced by 60 %. Diesel oil savings in the range 2 – 3 % are achieved. Cerium compounds added to silicones increase the thermal stability.

Cerium oxide is used in self-cleaning ovens as a catalyst [63].

In zeolitic cracking catalysts, cerium is a thermal and hydrothermal stabilizer to extend the life and increase the activity of the catalysts [64]. The following uses of cerium in noncracking catalysts are listed by PETERS and KIM [65]:

ammonia synthesis
hydrogenation
dehydrogenation
polymerization
isomerization
oxidation
automobile emissions control

5. Analysis

Mischmetal and cerium oxide can be analyzed for total rare-earth content (TRO) and for individual rare-earth elements [6], [66].

The determination of the *total rare-earth content* is carried out by dissolving the rare-earth-containing material with acids or by alkaline fusion. The rare-earth elements are precipitated as hydroxides by adding NH_3, dissolved in HCl or HNO_3, and reprecipitated as oxalates, which are ignited to the oxides, R_2O_3, with the exception of cerium, which forms CeO_2.

The *individual rare-earth elements* are determined by X-ray fluorescence (La), atomic absorption spectrometry (Y), and spectrophotometry (Nd, Pr, Sm). Analysis by means of optical plasma emission spectroscopy is a new, efficient method. Cerium can be determined by titration because it can be oxidized to the tetrapositive state.

6. Economic Aspects

Total rare-earth production in the Western world is 20 000 – 25 000 t of rare-earth oxides per year in 1986, of which 45 – 50 % is cerium. Raw material prices have stabilized for bastnaesite at $ 1 per pound TRO FAS Los Angeles, for monazite at A$ 450 per tonne FOB Australia, and rare-earth chlorides ($RCl_3 \cdot 6\,H_2O$) at ca. $ 0.70 per pound delivered.

The growing production of rare-earth raw material and products from the People's Republic of China should only stabilize the market and not sensationally affect prices.

Rhône-Poulenc (France, compounds) and Molycorp-Union Oil (United States, bastnaesite and compounds) are the leading producers; other producers are Treibacher (Austria, metals, alloys, and compounds), Santoku (Japan, metals, alloys, and compounds), Ronson and Reactive Metals (United States, metals and alloys), Indian Rare Earths (India, compounds), and Corona and Fluminense (Brazil, metals and alloys). In Germany, Goldschmidt specializes in magnet alloys. Further magnet alloy producers are Hitachi and Research Chemicals (United States) and Sumitomo and Shin-Etsu (Japan).

Important producers of lighter flints in the Western world are Treibacher, Electro Centre (France), Santoku, and Ronson. Total world market for lighter flints is estimated at ca. 700 t/a. The share of flints for disposable lighters is more than 60 %.

Modern steel technology has decreased demand for mischmetal substantially. The metallurgical uses of rare-earth metals consume less than 15 % of the total rare-earth production. The production capacity for mischmetal is 4 000 – 5 000 t/a, but less than 50 % is in operation (1986). Mischmetal is sold at $ 4 – 5 per pound, depending on quantity, shape, and size.

Total world demand for cerium-based polishing powders, the biggest consumer of cerium oxide at present, is estimated to be 3 0000 – 4 0000 t/a. The price for cerium polishing compounds depends on the cerium content and ranges between $ 2 per pound for 50 % CeO_2 and $ 5 per pound for 99.9 % CeO_2.

The market for more sophisticated uses – electronics, magnets, etc. – is expected to grow.

The production of rare earths is profitable only if the market for all the rare-earth elements, which are produced in the relative amounts occurring in their minerals, is reasonably balanced.

7. Toxicology and OccupationalHealth

The rare-earth elements and compounds are considered to be only slightly toxic. No toxic effects during production or use have been reported in the rare-earth industry.

Progressive lung retention was observed after inhalation of dust containing rare-earth oxides and fluorides from carbon arc electrodes. The damage to the lung that was assessed by X-ray did not seem to be attributable to the rare earths, but to thorium and its disintegration products. Continual progress in rare-earth processing has reduced the radioactive impurities in rare-earth products substantially, so they are practically free of radioactivity today. Intravenous injections of rare-earth salts damage the liver; oral administration has no pathological effect on animals [1, B 2, pp. 282 – 283], [67].

8. References

General References

1. *Gmelin,* system no. 39, Seltene Erden, A 1 (1938); A 2 – A 5, A 7, A 8, B 1 – B 7, C 1 – C 9, D 1 – D 6 (1973 to date).
2. K. A. Gscheidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 1 (1978), vols. 2 – 4 (1979); vol. 5 (1982), vols. 6 and 7 (1984), vol. 8 (1986)** North Holland Publ., Amsterdam.
3. L. Eyring (ed.): *Progress in the Science and Technology of the Rare Earths,* **vol. 1 (1964), vol. 2 (1966), vol. 3 (1968),** Pergamon Press, Oxford.
4. J. A. Gibson, G. S. Harvey: *Properties of the Rare Earth Metals and Compounds,* Technical Report AFML-TR-65-430, 1966.
5. F. H. Spedding, A. H. Daane: *The Rare Earths,* J. Wiley & Sons, New York 1961.
6. F. Trombe, J. Loriers, P. Gaume-Mahn, C. Henry La Blanchetais: "Scandium-Yttrium-Éléments des Terres Rares-Actinium," in P. Pascal (ed.): *Nouveau Traité de Chimie Minérale,* vol. VII, Fascicule 1 + 2, Masson, Paris 1959.
7. J. Helgorsky, A. Lévêque, T. Petzel, K. Reinhardt: "Seltene Erden," in *Ullmann,* 4th ed., **21,** 235 – 271.
8. E. Greinacher, K. Reinhardt: "Seltene Erden," in *Winnacker-Küchler,* 4th ed., vol. 2, pp. 678 – 707.
9. K. A. Gschneidner, Jr. (ed.): *Industrial Application of Rare Earth Elements,* ACS Ser. no. 164, Washington, D.C., 1981.
10. I. H. Jolly, R. P. Smith: "Availability of Rare Earths to Industry," in *Proc. Rare Earth Res. Conf., 12th, 1976*.
11. C. R. Neary, D. E. Highley: "The Economic Importance of the Rare Earth Elements," *Dev. Geochem.* **2** (1984) 423 – 466.

12. J. Griffiths, *Ind. Min. (London)* **199** (1984) April, 19 – 37.

Specific References

13. Soc. Prod. Chim. Terres Rares, FR 995 112, 1949 (C. de Rhoden, M. Peltier).
14. A. E. Bearse, G. D. Calkins, J. W. Glegg, R. B. Filbert, *Chem. Eng. Prog.* **50** (1954) 235 – 239.
15. P. R. Kruesi, G. Duker, *J. Met.* **17** (1965) 847 – 849.
16. D. J. Bauer, V. E. Shaw, *Rep. Invest. U.S. Bur. Mines* no. 6381, Washington, D.C., 1964.
17. W. Brugger, E. Greinacher, *J. Met.* **19** (1967) 32 –35. Th. Goldschmidt AG, DE 956 993, 1954 (W. Brugger).
18. I. S. Hirschhorn, *J. Met.* **20** (1968) 19 – 22. BIOS Final Rep. no. 22 (file no. XXXIII-31), 1947, p. 133.
19. BIOS Final Rep. no. 400 (item no. 21), 1945, p. 36.
20. D. Tang, R. Song, S. Du. M. Zao, Y. Minshou, W. Yu, *Proc. Int. Symp. Molten Salt Chem. Technol., 1st,* 1983, 103 – 106.
21. Th. Goldschmidt AG, DE 2 254 245, 1974 (E. Greinacher).
22. T. A. Henrie, *J. Met.* **16** (1964) 978 – 981.
23. E. S. Shedd, J. D. Marchant, M. M. Wong, *Rep. Invest. U.S. Bur. Mines* no. 7398, Washington, D.C., 1970.
24. B. J. Beaudry, K. A. Gschneidner, Jr. L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 1** North Holland Publ., Amsterdam 1978, pp. 173 – 232 D. C. Koskenmaki, K. A. Gschneidner, Jr. L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 1** North Holland Publ., Amsterdam 1978, pp. 337 – 378.
25. K. A. Gschneidner, Jr., N. Kippenhan: *Thermochemistry of the Rare Earth Carbides, Nitrides, and Sulfides,* IS-RIC-5, Rare Earth Information Center, Iowa State University, Ames, Iowa, 1972. K. A. Gschneidner, Jr., N. Kippenhan, O. D. McMasters: *Thermochemistry of the Rare Earths,* IS-RIC-6, Rare Earth Information Center, Iowa State University, Ames, Iowa, 1973.
26. K. Reinhardt: "Addition Techniques of Rare Earth Metals for the Treatment of Special Steels with Rare Earths," in *Proc. Rare Earth Res. Conf., 12th, 1976*.
27. A. Raman, *Z. Metallkd.* **67** (1976) 780 – 789. L. A. Luyckx: "The Rare Earth Metals in Steel," [9] pp. 43 – 78.
28. B. Zao, E. W. Langer, *Scand. J. Metall* . 13 (1984) no. 1, 15 – 22, 23 – 31. H. F. Linebarger, T. K. McLuhan: "The Role of the Rare Earth Elements in the Production of Nodular Iron," [9] pp. 19 – 42.
29. A. Raman, *Z. Metallkd.* **68** (1977) 163 – 172.
30. S. F. Radtke, D. C. Herrschaft, *J. Less-Common Met.* **93** (1983) 253 – 259. F. E. Goodwin, S. F. Radtke: *Galfan Galvanizing Alloy and Technology,* 2nd ed., Int. Lead and Zinc Res. Org., ILZRO, New York 1983.
31. Manassa, DE-OS 2 551 350, 1975.
32. K. A. Gschneidner, Jr.: *Rare Earth Alloys,* Van Nostrand, Princeton-London 1961.
33. H. R. Kirchmayr, *Z. Metallkd.* **60** (1969) 699 – 708, 778 – 784. W. E. Wallace: *Rare Earth Intermetallics,* Academic Press, New York-London 1973.
34. K. H. J. Buschow, *Rep. Prog. Phys.* **42** (1979) 1373 – 1477. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* 2nd ed., McGraw-Hill, New York 1958. R. P. Elliot: *Constitution of Binary Alloys, First Supplement,* 1965. T. A. Shunk: *Constitution of Binary Alloys, Second Supplement,* 1969. W. G. Moffart: *The Handbook of Binary Phase Diagrams,* General Electric, Schenectady, N.Y., 1978. K. A. Gschneidner, Jr., F. W. Calderwood: *Critical Evaluation of Binary Rare Earth Phase Diagrams,* **IS-RIC-PR 1 – 10 (to be continued),** Rare Earth Information Center, Iowa State University, Ames, Iowa, 1981 on.
35. K. A. Gschneidner, Jr., M. E. Verkade: *Selected Cerium Phase Diagrams,* IS-RIC-7, Rare Earth Information Center, Iowa State University, Ames, Iowa, 1974.
36. Treibacher Chemische Werke AG, AT 342 323, 1977 (H. Zeiringer).
37. H. R. Kirchmayr, C. A. Poldy, in [2] vol. 2, pp. 55 – 230, especially pp. 197 – 230. *Proc. Int. Workshop Rare Earth-Cobalt Perm. Magnets Their Appl., 5th, 1981. Proc. Int. Workshop Rare Earth-Cobalt Perm. Magnets Their Appl., 6th, 1982. Proc. Int. Symp. Magn. Anisotropy Coercivity in Rare Earth-Transition Met. Alloys, 3rd, 1982. Proc. Int. Workshop Rare Earth-Cobalt Perm. Magnets Their Appl., 7th, 1983.*
38. K. Bachmann, *J. Magn. Magn. Mater*. **4** (1977) 8 – 12. Ugimag Recoma AG, DE 2 449 361, 1976 (S. Gaiffi, A. Menth, H. Nagel). Ugimag Recoma AG, DE 2 543 916, 1977 (S. Gaiffi, A. Menth, H. Nagel).

39. G. Schaefer, H. H. Wiegand, *Metall (Berlin)* **29** (1975) no. 2, 113 – 121. J. W. Walkiewicz, E. Morrice, M. M. Wong, *IEEE Trans. Magn.* **MAG-19** (1983) 2053 – 2055.
40. M. G. H. Wells, K. S. V. L. Narasimhan: "Developments in Mischmetal-Cobalt$_5$ and Low-Samarium Magnet Materials," *"Goldschmidt informiert..."* no. 48 (1979) pp. 15 – 22.
41. Th. Goldschmidt AG, DE 2 303 697, 1974 (H.-G. Domazer).
42. E. L. Huston, J. J. Sheridan III: "Applications for Rechargeable Metal Hydrides," [9] pp. 223 – 250. K. H. J. Buschow, [2] vol. 6, pp. 1 – 111.
43. J. Liu, E. L. Huston, *J. Less-Common Met.* **90** (1983) 11 – 20.
44. Y. Osumi, H. Suzuki, A. Kato, K. Oguro, M. Nakane, *J. Less-Common Met.* **74** (1980) 271 – 277.
45. G. D. Sandrock, US 4 152 145, 1979. Y. Osumi, A. Kato, H. Suzuki, M. Nakane, Y. Miyake, *J. Less-Common Met.* **66** (1979) 67 – 75. V. K. Sinha, W. E. Wallace, *J. Phys. Chem.* **88** (1984) 102 – 105.
46. G. D. Sandrock, *Proc. World Conf. Hydrogen Energy Systems. 2nd*, Zuerich 1978, vol. 3, pp. 1625 – 1656. International Nickel Co., US 4 161 402, 1979 (G. D. Sandrock).
47. R. A. Guidotti, G. B. Atkinson, M. M. Wong, *J. Less-Common Met.* **52** (1977) 13 – 28. Agency of Industrial Science and Technology, US 4 147 536, 1979 (Y. Osumi, H. Suzuki, A. Kato, M. Nakane, Y. Miyake).
48. Agency of Industrial Science and Technology, US 4 347 082, 1982 (Y. Osumi, H. Suzuki, A. Kato, M. Nakane). Y. Osumi, H. Suzuki, A. Kato, K. Oguro, M. Nakane, *J. Less-Common Met.* **79** (1981) 207 – 214.
49. F. Pourarian, W. E. Wallace, *J. Less-Common Met.* **87** (1982) 275 – 281.
50. International Nickel Co., US 4 249 940, 1981 (G. D. Sandrock, S. L. Keresztes).
51. C. E. Lundin, F. E. Lynch, *Altern. Energy Sources Proc. Miami Int. Conf. 1977,* **8** (1978) 3803 – 3820.
52. B. Darriet, A. Hbika, M. Pezat, *J. Less-Common Met.* **75** (1980) 43 – 50.
53. Shin-Etsu Chemical, DE-OS 2 623 213, 1976 (Z. Hagiwara, S. Matsui, S. Sakaguchi, Y. Yamanaka).
54. U.S. Atomic Energy Commission, US 3 825 418, 1973 (J. J. Reilly, R. H. Wiswall, Jr.).
55. H. W. Harrah: *Deco Trefoil,* Denver Equipment Co., Denver, Colorado, 1967, p. 9.
56. R. V. Horrigan: "Rare Earth Polishing Compounds," [9] pp. 95 – 100.
57. A. Kaller, *Silikattechnik* **31** (1980) 208 – 214.
58. T. Izumitani, S. Harada, *Wiss. Z. Friedrich-SchillerUniv., Jena, Math.-Naturwiss. Reihe* **28** (1979) no. 2 – 3, 389 – 413.
59. A. P. Herring, R. W. Dean, J. L. Drobnick: "Use of Cerium Concentrate for Decolorizing Soda-Lime Glasses," *Glass Ind.* **51** (1970) 316 – 322, 350 – 356, 394 – 399.
60. L. W. Riker: "The Use of Rare Earths in Glass Compositions," [9] pp. 81 – 94.
61. Molycorp-Union Oil: "Cerium," *Overview of Application Information and Possibilities,* no. 58.
62. Gulf Research & Development Co., US 4 264 335, 1981 (C. Bello, R. J. Hartle, G. M. Singerman).
63. E. I. Du Pont de Nemours & Co., US 3 266 477, 1966; US 3 271 322, 1966 (A. B. Stiles).
64. D. N. Wallace: "The Use of Rare Earth Elements in Zeolite Cracking Catalysts," [9] pp. 101 – 116.
65. A. W. Peters, G. Kim: "Rare Earths in Noncracking Catalysts," [9] pp. 117 – 131.
66. A. Brusdeylins, *Chem. Ztg.* **97** (1973) 343 – 347. O. B. Michelsen (ed.): *Analysis and Application of Rare Earth Materials,* Universitetsforlaget, Oslo 1973. Cheng Jai-Kai, *Inorg. Chim. Acta* **94** (1984) 249 – 258. P. Melard: "Quality Control on an Industrial Scale at the LaRochelle Rare Earth Plant," *Rare Earths Mod. Sci. Technol.* **2** (1980) 517 – 526.
67. T. J. Healey, [2] vol. 4, pp. 553 – 585. P. Arvela: "Toxicity of Rare Earths," *Prog. Pharmacol.* **2** (1979) no. 3, 69 – 114.

Cesium and Cesium Compounds

MANFRED BICK, Chemetall GmbH, Frankfurt/Main, Federal Republic of Germany

HORST PRINZ, Chemetall GmbH, Frankfurt/Main, Federal Republic of Germany

1. **Physical and Chemical Properties** . . . 1
2. **Resources** 2
3. **Production** 2
4. **Cesium Compounds** 3
5. **Quality Specifications** 3
6. **Chemical Analysis** 3
7. **Storage and Transportation** 3
8. **Uses** 3
9. **Economic Aspects** 4
10. **Toxicology and Occupational Health** . 4
11. **References** 5

1. Physical and Chemical Properties

Cesium [*7440-46-2*] (from the Latin word *caesius*, gray blue, which characterizes the pale blue color of the emissionof thermally excited cesium atoms) is a soft, gold-colored metal, which melts slightly above room temperature. Table 1 summarizes the physical properties of elemental cesium.

Table 1. Physical properties of cesium [4], [5]

Property	Value
Atomic number	55
Relative atomic mass A_r	132.90543
Melting point *mp*	28.7 °C
Boiling point *bp*	685 °C
Atomic radius ∗	0.274 nm
Ionic radius ∗∗	0.165 nm
Density ϱ at 20 °C	1.873 g/cm^3
Specific heat, liquid	0.236 J g^{-1} K^{-1}
Heat of fusion	2.13 kJ/mol
Heat of vaporization at 0.1 MPa	65.9 J/mol
Ionization potential	3.87 eV
Standard electrode potential	−2.923 V
Work function	1.91 eV
Electrical conductivity	
solid, 25 °C	$4.9 \times 10^4\ \Omega^{-1}\ \mathrm{cm}^{-1}$
vapor, 1250 °C	$2.0 \times 10^2\ \Omega^{-1}\ \mathrm{cm}^{-1}$
Vapor pressure *P*, kPa	
$\log P = -0.2185\frac{A}{T} + B$	

T, K	*A*	*B*
200 – 350	17543.0	6.0739
279 – 690	17070.7	5.8889

∗ 12-Coordinate metal.
∗∗ 6-Coordinate salt, empirical.

Cesium belongs to the alkali-metal group. The metal reacts with most elements, to form alloys when reacted with metals or to form typical salts when reacted with halogens. In the salts, cesium is strictly monovalent, and its chemistry is typical of an alkali metal. Because of its extremely low ionization potential, cesium is usually far more reactive than lithium, sodium, or potassium and still pronouncedly more reactive than rubidium. When cesium is exposed to air, an explosion-like oxidation to form cesium superoxide, CsO_2, occurs; contact with water results in a vigorous reaction to form cesium hydroxide and hydrogen gas, which may ignite spontaneously.

Naturally occurring cesium consists only of the stable isotope ^{133}Cs. Radioactive cesium isotopes, such as ^{137}Cs (half-life of 30.1 a), are generated in fuel rods during operation of nuclear power plants.

2. Resources

Cesium is the 40th most abundant among the elements, occurring about as frequently as germanium. Resources can be categorized into two groups. One, of no commercial importance, consists of the diffuse occurrence of the few grams of cesium per ton contained in potassium salt deposits, sedimentary rocks, and seawater [3].

The only sources of cesium of commercial importance originated during the solidification of residual melts of silicate magmas. After the initial formation of huge granite masses, the remaining melt rich in rare elements like lithium, rubidium, cesium, tantalum, niobium, beryllium, and tin then crystallized to form a type of ore body well-known under the name *pegmatite.* Under favorable conditions, which

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_153

seem to have existed only in relatively few cases, these pegmatites differentiated into separate bodies. Sodium, potassium, and rubidium formed feldspar-type minerals of considerable mutual solubility, while lithium and cesium, because of the considerable difference in ionic radii from the other alkali metals, formed separate minerals. In the case of cesium, this seems to have occurred in far fewer cases than for lithium, yielding the only commercial cesium mineral, *pollucite* [*1308-53-8*] [6], [7].

Pollucite is a cesium aluminum silicate, which typically contains 18 – 26 % cesium oxide. The theoretical content is 42 %, but usually pollucite contains considerable quantities of quartz. Its appearance is also very similar to that of quartz. For this reason and because the demand for cesium has not been great enough to result in any systematic exploration for cesium minerals, possibly further deposits may be discovered. Well known are the large deposits at Bernic Lake in Canada, at Bikita in Zimbabwe, and in South-West Africa. In the former Soviet Union a new mineral, cestibtantite, a mixed cesium – antimony – tantalum oxide, has been reported [8], and pollucite has been found in some deposits of lithium, niobium, and tantalum minerals [9], [10]. Apparently, pollucite also occurs in Afghanistan [11]. Small concentrations of cesium, less than one percent, are found in lepidolite, a lithium mineral.

3. Production

For any of the production methods described, pollucite must be powdered first. Production processes can be categorized as (1) decomposition with bases and (2) acid digestion. The second category includes the group of processes used industrially.

Alkaline decomposition can be carried out either by mixing ore with lime and calcium chloride and heating to 800 – 900 °C followed by leaching of the residue [12] or by heating pollucite with sodium chloride and soda ash to 600 – 800 °C followed by leaching [13]. In both cases, solutions of impure cesium chloride result. Hydrothermal alkaline leaching with $Ca(OH)_2$ at 200 – 280 °C under high pressure [14] has been developed as an industrial process to yield a low-purity cesium hydroxide directly from pollucite [15], [16].

Acid digestion can be carried out with hydrochloric acid, sulfuric acid, or hydrobromic acid. Treatment of pollucite with hydrochloric acid at elevated temperature produces a solution of chlorides of cesium, aluminum, and alkali metals, which is separated from the silica residue. The cesium is precipitated as mixed chloride with lead, antimony, or tin. Hydrolysis precipitates the auxiliary metal [17]. Alternatives are precipitation with hydrogen sulfide or recovery of cesium by solvent extraction from the leach liquor [18] or ion exchange from cesium chloride solution.

Treatment of pollucite with sulfuric acid [12, p. 5] yields sparingly soluble cesium alum, cesium aluminum sulfate [*7784-17-0*], which is roasted with carbon to convert the aluminumto alumina and the sulfate sulfur to SO_2. The residue is leached to obtain a cesium sulfate solution, which can be converted to the desired salts by ion exchange, treatment with ammonia or lime (to precipitate aluminum), or solvent extraction [19]. Hydrobromic acid converts pollucite into a solution of bromides of cesium, aluminum, and impurity metals, from which cesium can be precipitated by addition of alcohol. Leaching the precipitate with bromine selectively extracts cesium as $CsBr_3$ [20].

Cesium metal can be produced as an amalgam by electrolyzing concentrated cesium salt solutions on a mercury cathode [21], but reduction of solid cesium salts, especially the halides, with calcium or barium at elevated temperature and removal of cesium by vacuum distillation is the usual method [22].

4. Cesium Compounds

Cesium carbonate [*534-17-8*], Cs_2CO_3, M_r 325.82, 81.58 % Cs, is a colorless, hygroscopic powder, decomp. at 610 °C, ϱ 4.07 g/cm^3, solubility of 261.5 g in 100 g of water, basic solution.

Cesium hydrogen carbonate [*15519-28-5*], $CsHCO_3$, M_r 193.92, 68.54 % Cs, colorless crystalline powder, decomp. at 175 °C, solubility of 209 g in 100 g of water, basic solution.

Cesium chloride [*7647-17-8*], CsCl, M_r 168.36, 78.9 % Cs, colorless, crystalline,

hygroscopic powder, *mp* 642 °C, ϱ 3.983 g/cm^3, solubility of 186 g in 100 g of water.

Cesium hydroxide [*21351-79-1*], CsOH, M_r149.91, 88.66 % Cs, anhydrous, colorless, lumpy solid, *mp* 272 °C, ϱ 3.68 g/cm^3, solubility of ca. 400 g in 100 g of water. The solution is a strong base and very caustic.

Cesium hydroxide monohydrate [*35103-79-8*], $CsOH \cdot H_2O$, M_r167.93, 79.14 % Cs, colorless, crystalline, hygroscopic powder, *mp* 205 – 208 °C, ϱ 3.5 g/cm^3, solubility of ca. 860 g in 100 g of water. The solution is a strong base and very caustic.

Cesium iodide [*7789-17-5*], CsI, M_r259.81, 51.2 % Cs, colorless, slightly hygroscopic powder, *mp* 621 °C, ϱ 4.51 g/cm^3, solubility of 74 g in 100 g of water.

Cesium nitrate [*7789-18-6*], $CsNO_3$, M_r194.91, 68.19 % Cs, colorless crystalline powder, *mp* 414 °C, ϱ 3.69 g/cm^3, oxidant.

Cesium sulfate [*10294-54-9*], Cs_2SO_4, M_r361.87, 73.46 % Cs, colorless hygroscopic powder, *mp* 1010 °C, ϱ 4.243 g/cm^3.

5. Quality Specifications

Cesium metal and cesium salts are marketed in purities from 99 % for technical grades to 99.999 % for high-purity compounds. For some applications, a crude form with ca. 85 % purity can also be produced.

6. Chemical Analysis

Assays and purities of commercial products are derived by subtracting the sum of analyzed impurity levels from unity. Alkali-metal impurities are analyzed by emission spectroscopy, whereas alkaline-earth metals are determined by atomic absorption. Other metals and anions, such as phosphate and sulfate, can be determined by photometric methods; chloride is established argentometrically.

7. Storage and Transportation

Special precautions are necessary for cesium metal, which is shipped in evacuated glass ampules or steel containers with an appropriate outer packing to ensure that the metal is kept from air and moisture under any conditions.

Many cesium salts, especially halides, are hygroscopic and must be stored dry. Transportation regulations, where they exist, are governed by the anion, i.e., the hydroxide being caustic, the nitrate being an oxidant, because nothing inherent in the cesium cation calls for special precautions.

8. Uses

Cesium *metal* is used in frequency standards, especially the time standards known as atomic clocks, which lock in on one frequency of excited gas-phase cesium. The metal is also used in the production of various types of vacuum tubes, where it acts as a scavenger to reduce residual gaseous impurities after the tubes have been sealed. The cesium is normally generated in situ by firing the tube to convert a pellet of cesium chromate mixed with a metal powder to elemental cesium vapor.

In ion thrusters, cesium metal is the propellant. It is ionized in a vacuum chamber, and cesium ions are then accelerated through a field and ejected through a nozzle, producing a high specific impulse because of their high atomic mass. Devices of this type are used in satellites for orientation control.

Recognition that the reserves of fossil fuels are limited has focused considerable research on conversion of solar energy into electricity. One approach to achieve this uses thermionic converters, which require cesium. Direct conversion of heat from nuclear reactors into electricity by generating a voltage difference across a cesium vapor-filled vacuum tube between hot and cold electrodes has been proposed [23]. Another potential use of cesium metal for energy generation is its use as a plasma seeding agent in closed-cycle magnetohydrodynamic generators that use high-temperature nuclear reactors as their heat source.

Of the *compounds,* the halides (especially the chloride), the trifluoroacetate, and the sulfate are used in ultracentrifuges, where aqueous solutions of high-purity grades are a medium for separation and purification of nucleic acids (DNA and RNA) for biochemical research. At

high rates of rotation, these solutions form a density gradient that separates nucleic acids according to their densities [24].

Various catalysts can be doped with cesium salts as activators, much like the corresponding potassium salts. High-purity cesium halides are transparent to infrared radiation; therefore, they are used for cuvettes, prisms, and windows for spectroscopic equipment. Cesium iodide can be doped to make it a scintillator [25]; single crystals are used in scintillation counting equipment. Cesium fluoride is used for fluorination in organic chemistry.

Open-cycle magnetohydrodynamic generators could offer a considerable potential for cesium compounds as a plasma seeding agent. These devices are under development, especially in North America and the former Soviet Union, with the hope that they can boost the overall efficiency of power plants that depend on fossil fuels from ca. 35 to ca. 45 %. Hot combustion gas is seeded with potassium or cesium carbonate to make a highly conductive plasma, which is passed through a magnetic field. At right angles to both the field and the direction of plasma flow, there is a voltage difference [26]. Nevertheless, the higher seeding efficiency of cesium compounds must compete with the lower price of potassium compounds.

The very low toxicity of the cesium cation and the pronounced physical and chemical properties of cesium formate to form clear solutions of high density (up to 2.3 t/m^3), have led to the suggestion to use these solutions as brines in oil and gas well drilling [27], alone or in mixtures with sodium or potassium formate.

As an environmentally friendly alternative, cesium formate (and the other alkali metal formates) could potentially replace large quantities of the toxic, hazardous, corrosive heavy metal compound zinc bromide in this application. This is most likely in areas where environmental regulations are strict, for example, in the off-shore North Sea fields.

9. Economic Aspects

The producers do not publish production or consumption figures. The U.S. Bureau of Mines estimated world consumption in 1978 at about 20 t of cesium, as metal and in compounds. At that time prices for cesium salts were 64 – 81 $/kg for technical grades and 147 – 170 $/kg for high-purity products.

The price levels in 1996 were 35 – 85 $/kg for technical and pure grades and 85 – 150 $/kg for high-purity salts.

10. Toxicology and Occupational Health

The *cesium ion* itself is only very slightly toxic, more toxic than the sodium ion, but less toxic than the potassium ion. Typical LD_{50} values of cesium salts are 1400 mg/kg (rat, intraperitoneal) and 1000 mg/kg (rat, oral) [28]. Exceptions are caused by the toxicity of the particular anion. Cesium hydroxide is strongly caustic, cesium nitrate is an oxidant, and cesium fluoride exhibits the typical toxicity of fluoride.

Special precautions are necessary when handling the metal because exposure to this substance results in severe caustic burns.

11. References

General References

1. *Gmelin,* system no. 25, Caesium (1938).
2. *Kirk-Othmer,* 3rd ed., vol. 5, pp. 327 – 338.
3. C. A. Hampel: *Rare Metals Handbook,* 2nd ed., Reinhold Publ. Co., New York 1969.

Specific References

4. W. Peek, *Chem.-Ztg.* **87** (1963) 315 – 318.
5. *Handbook of Chemistry and Physics,* 53rd ed., CRC Press, Cleveland 1972 – 1973.
6. V. M. Goldschmidt: *Geochemistry,* Clarendon Press, Oxford 1954.
7. V. V. Gordiyenko, *Int. Geol. Rev.* **13** (1970) no. 2, 134.
8. P. Černý: "Mineralogy of Rubidium and Cesium," *Short Course Handb. Mineral. Assoc. Can.* **8** 1982 , 149 – 161.
9. V. F. Durnev, G. B. Melentyev, V. A. Sokolov, Y. N. Pokrovskiy, G. A. Cherepivskaya, *Dokl. Akad. Nauk SSSR* **213** (1973) no. 1, 180 – 183.
10. L. G. Kuznetsova, B. M. Shmakin, *Zap. Vses. Mineral. O-va.* **110** (1981) no. 1, 59 – 70.
11. L. N. Rossovskii, *Dokl. Akad. Nauk SSSR* **236** (1977) no. 1, 216 – 219.
12. K. C. Dean, P. H. Johnson, I. L. Nichols, *Rep. Invest. U.S. Bur. Mines* **6387** (1964) .

13. W. D. Arnold, D. J. Crouse, K. B. Brown, *Ind. Eng. Chem.* **4** (1965) 249.
14. Y. G. Goroshchenko, E. B. Panasenko, V. A. Roi, V. P. Izotov, *Tsvetnye Metally* **2** (1961) no. 5, 57 – 59.
15. Metallgesellschaft AG, DE 42 37 954 C1 (M. Wegner, K. Köbele, H. Hoffmann, H. Prinz).
16. Metallgesellschaft AG, DE 43 39 062.5-41 (K. Schade, K. Köbele, H. Hoffmann, H. Prinz).
17. J. C. Bailor, Jr.: *Inorganic Syntheses,* vol. 4, McGraw-Hill, New York 1965.
18. *Chem. Eng. News* **41** (1963) no. 51 (Dec. 23), 35.
19. H. W. Parsons, J. A. Vezina, R. Simard, H. W. Smith, *Can. Dept. Mines, Mineral Branch, Tech. Bull.* **50** (1963) , reprint of *Can. Metall. Q.* **2** (1963) 1 – 13; *Chem. Abstr.* **58** (1963) 12 199c.
20. V. A. Stenger, US 2 481 455, 1949.
21. R. E. Davis: "Electrowinning of Rubidium and Cesium," *Encyclopedia of Electrochemistry,* Reinhold Publ. Co., New York 1964.
22. L. Hackspill, *Helv. Chim. Acta* **11** (1928) 1003.
23. J. Raloff, *Sci. News (Washington D.C.)* **113** (1978) no. 13, 202.
24. J. Vinograd, J. E. Hearst: *Equilibrium Sedimentation of Macromolecules and Viruses in a Density Gradient,* Springer-Verlag, Wien 1962.
25. P. Brinckmann, US 3 446 745, 1966.
26. J. Melcher, *Min. Eng. (Littleton, Colo.)* **29** (1977) no. 12, 34.
27. Shell Internationale Research Maatschappij B.V., EP-A 0 259 939 A3 (A. J. Clarke-Sturman, P. L. Sturla).
28. K. W. Cochran et al., *Arch. Ind. Hyg. Occup. Med.* **1** (1950) 637.

Cetyl Alcohol → Fatty Alcohols
Chalks → Drawing and Writing Materials

Charcoal

See also Carbon, Chap. 5.

Horst-Günter Brocksiepe, Degussa AG, Werk Bodenfelde, Federal Republic of Germany

1. Introduction 1
2. Physical and Chemical Properties . . 1
3. Raw Materials 2
4. Production 3
4.1. Wood Pyrolysis and Byproducts 3
4.2. Carbonization in Large-Volume Retorts 4
4.3. New Carbonization Processes 4
5. Environmental Protection 5
6. Storage and Transportation 6
7. Uses 6
8. References 6

1. Introduction

Charcoal is produced in kilns and retorts. The thermal decomposition of wood in a kiln with a controlled air supply produces charcoal. The gaseous products and some of the wood are burned. Retort charcoal is generated in the absence of air. In addition to charcoal, gaseous and liquid byproducts form: pyrolysis gas (wood gas), crude pyroligneous liquor, and pyrolysis oil (bio oil). The charcoal yield is 50 % higher in a retort than in a kiln.

History. Charcoal has been produced since prehistoric times, first in pits and later in kilns. Not before the end of the 18th century were masonry and iron retorts used. With retorts the acetic acid and methanol byproducts could be recovered in greater yields. However, the recovery of the byproducts became uneconomical when these products could be synthesized.

Originally, charcoal was used as a smokeless, odorless fuel for heating and cooking, and it is still used today for these purposes by people in well-wooded underdeveloped countries. In industry, charcoal is used metallurgically. In addition, charcoal is used for the production of activated carbon and for barbecuing foods.

The worldwide production of charcoal was estimated to be 18×10^6 t in 1984:

United States and Canada	1.3×10^6 t
South America	5.5×10^6 t
Asia, without former Soviet Union and China	5.3×10^6 t
Africa	1.8×10^6 t
Europe, without East and Southeast Europe	0.2×10^6 t

The use of charcoal for particular purposes differs from country to country. In the United States, 95 % of the production is used for barbecuing, with only 5 % needed for industrial purposes. In South America, especially Brazil, nearly all of the charcoal production is consumed by iron works. In Europe and Japan, the amounts of charcoal used for industrial purposes and barbecuing are about the same.

2. Physical and Chemical Properties

The charcoal quality depends mainly on the kind of wood used to make the charcoal. Charcoal produced from hardwood (angiospermous) trees, especially beech, is dense and firm. In contrast, charcoal from softwood (coniferous) trees is soft and porous. The reason for this is the differing wood structures, which do not change much during carbonization. The annual rings remain completely visible. The density of the charcoal depends on the density of the porous dry wood.

The apparent density of beech charcoal is 0.45 g/mL and that of pine charcoal 0.28 g/mL. The density of pore-free charcoal ranges between 1.38 and 1.46 g/mL, depending on the kind of wood. The porosity of beech charcoal is ca. 70 %. The surface area of charcoal is enormous, ca. 50 m^2/g. The bulk density of charcoal, which depends on the kind of wood and the size of the pieces, ranges from 180 to 220 kg/m^3. The pressure resistance of beech charcoal is 26 MPa in the fiber direction, 6 MPa radial to the annual rings, and 2 MPa tangential to them.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_157

The ignition point of charcoal ranges between 200 and 250 °C, depending on the content of volatile compounds. The heating value ranges between 29 and 33 kJ/g.

In the Federal Republic of Germany the quality of charcoal is specified in DIN 51 749. Excellent charcoal is black and shiny. It is free of cracks and makes a metallic sound when struck. The carbon content of retort charcoal is 78 – 90 wt %. This carbon is a finely crystalline type of graphite [*7782-42-5*]. It is pureand practically free of sulfur. In addition to carbon, charcoal contains between10 % and 18 % hydrogen, oxygen, and nitrogen, which escape at 875 °C. The inorganic components comprise 1 – 2 %. The water content of charcoal depends on the temperature and the relative humidity of the air. It varies between 5 % and 8 %.

Analysis. In practice, the fixed carbon content is determined by heating a sample to 900 ± 10 °C under air-tight conditions, which drives off the volatile components. The residue is then ashed at 710 ± 15 °C. The fixed carbon content is calculated as follows:
Fixed carbon, % = 100 – (%volatiles + %ash)

3. Raw Materials

The natural raw material for the charcoal production is *wood*, both hardwood and softwood. In addition, sawdust, wood shavings, fruit stones, nuts, nutshells, corn cobs, bark, cotton seeds, and similar products can be employed. Wood is used directly from the forest or from industrial waste, which is available in the form of outside planks, edges, and pieces.

For the carbonization process in a kiln, wood can be in any shape. For carbonization in retorts, equal-sized pieces are better. Special machines have been developed to shred wood, e.g., choppers. Small pieces of wood can be handled like bulk goods: they can be stored easily and can be transported by shovel loaders and belt conveyors.

Different units are used to measure the amount of wood: industrial wastes are measured in tons, whereas wood from the forest is measured in cubic meters or theoretical cubic meters.

Wood that has been very recently cut contains 50 wt % water. This green wood is not suitable for charcoal production. The high content of water makes the duration of the carbonization too long and the demand for heat too great. The disadvantages become even more evident if the liquid byproducts are recovered. Therefore, the wood that is used should be as dry as possible. By no means should the water content be higher than 5 – 20 wt %. This percentage of moisture in wood is achieved by storing split wood in stacks for 1 – 2 years. This degree of drying can also be reached in a few hours by artificial drying. Small pieces of wood or wastes are dried in a rotary dryer.

In cooperation with Lambiotte & Cie., Brussels, Degussa has developed a special vertical dryer for the charcoal industry (Fig. 1).

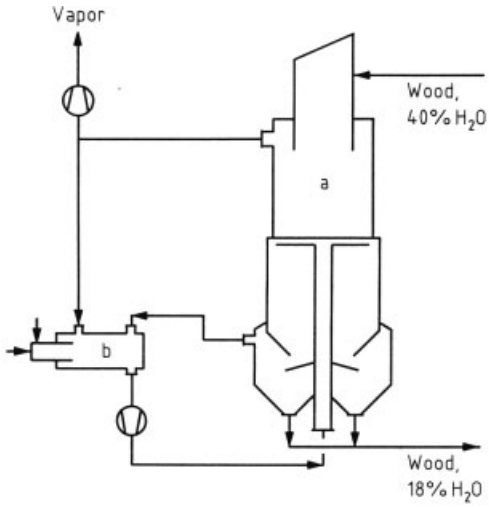

Figure 1. Wood dryer
a) Drying shaft; b) Flue gas heater

The dryer consists of a drying shaft, which is always filled with wood. Charge and discharge are controlled automatically. The wood is dried by circulation of hot flue gas. The temperature must not exceed 170 °C because of the danger of self-ignition. If waste heat is not available, flue gas can be generated by burning wood gas, pyrolysis oil, or another source of energy. To reduce the moisture content from 40 % to 18 %, a retention time of 16 – 18 h is necessary at a drying rate of 3000 t per month calculated as absolutely dry wood. The dryer volume is 300 m^3. Because drying produces a volume reduction of ca. 17 %, the wood passes the dryer without problem.

4. Production

The commercial production of charcoal involves the carbonization of wood and the formation of byproducts. In Europe, large-volume retorts are used, which produce charcoal and allow the byproducts to be saved. Elsewhere, recently developed processes produce charcoal without saving the byproducts.

4.1. Wood Pyrolysis and Byproducts

A series of decomposition reactions take place if wood is heated in the absence of atmospheric oxygen above 100 °C. These reactions are complex. Mainly, macromolecular chains of the wood components hemicellulose, cellulose, and lignin depolymerize to a variety of smaller molecules. Up to 270 °C, primary reactions, such as hydrolysis and isomerizations, take place predominantly. Above 270 °C, secondary reactions, in which molecules are split and recombine by condensation and polymerization, prevail [5], [6].

There are three typical ranges of temperature. From 100 – 170 °C, the free and bound water is evaporated. At the same temperatures small amounts of carbon dioxide, carbon monoxide, formic acid, and acetic acid are formed. These compounds indicate the first decomposition of hemicellulose [*9025-56-3*] and cellulose [*9004-34-6*], thermallythe least stable wood components. This phase of decomposition is neither especiallyendothermic nor especially exothermic.

The temperature range between 200 and 350 °C is characterized by exothermic reactions. Reaction is most intense between 270 and 280 °C. Increasing amounts of reaction water, methanol, and acetic acid form. In the range of this exothermic maximum, the amounts of pyrolysis products increase spontaneously: pyrolysis oil [7], hydrogen, methane, and ethylene appear as decomposition products of lignin [*9005-53-2*]. Charcoal is the residue.

Above 350 °C the decomposition is no longer exothermic. In practice, the exothermic reaction is complete at 380 °C. To obtain the desired carbon content, which depends on the final temperature, postheating is necessary. During the entire carbonization process, 1600 J/g are released, 8 – 10 % of the total heat content [8].

The yields of products obtained by the carbonization of wood depend on the type of wood, the kind of process, the duration of the carbonization, and the final temperature. The highest charcoal yields are obtained from wood up to 10 cm in diameter, by slow carbonization, and with a final temperature of ca. 400 °C. The yields achieved in a Degussa laboratory test (10-L retort, indirect heating) are shown in Table 1. The yields of acetic acid and pyroligneous spirit are lower from softwood than from hardwood.

The carbonization of wood with a moisture content of 10 – 15 wt % produces a distillate of crude pyroligneous liquor of the following composition:

12 % acetic acid and homologues
3 % pyroligneous spirit (mostly methanol)
10 % dissolved pyrolysis oil
75 % water

Whether recovering or burning the byproducts is more economical depends on several factors: the demand in the country for the byproducts, the environmental regulations, especially concerning wastewater, etc. The following byproducts can be recovered: acetic acid, propionic acid, pyroligneous spirit (methanol), pyrolysis oil, wood creosote, liquid smoke, pitch, flotation agents, 3-hydroxy-2-methyl-4-pyrone, and 2-hydroxy-3-methyl-2-cyclopenten-1-one. For a detailed description of the recovery, properties, and use of the byproducts, see [5], [9].

4.2. Carbonization in Large-Volume Retorts

In Europe, charcoal is produced mainly in large-volume retorts operating discontinuously or continuously. The Degussa process and the process of Lambiotte & Cie. predominate [10–16]. For a description of older processes, all obsolete today, see [5], [9].

The *Degussa retort* [11–15] is different from most of the others used in charcoal production (Fig. 2). It differs not only in size, but also in the type of heating. The retort is heated directly with hot wood gas. In this way less heat is required. If the moisture content of wood is less than 20 wt %, no additional fuel is necessary.

Table 1. Experimental yields obtained from various kinds of wood (wt % based on absolutely dry wood)

Product	Type of wood							
	Maple	Bongossi	Beech	Oak	Alder	Eucalyptus	Spruce	Pine
Charcoal ∗	35.0	41.4	34.0	35.7	35.5	36.5	34.2	32.8
Acetic acid	6.6	3.1	7.7	5.6	6.5	4.1	3.6	3.9
Pyroligneous spirit	1.8	2.4	2.1	1.6	1.9	2.1	1.7	1.5
Pyrolysis oil	15.5	11.0	14.0	13.6	16.2	12.3	15.6	18.9
Gases	15.5	17.2	16.0	14.9	16.8	16.3	15.2	15.4

∗ The charcoal yield is based on charcoal containing 80 wt % C.

The rate of carbonization is determined by the amount and the temperature of the wood gases, which means that the carbon content of the charcoal can be achieved within 1 % accuracy in the range 78 – 90 %.

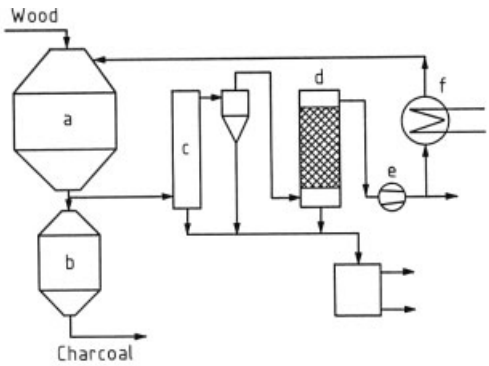

Figure 2. Carbonizing plant using the Degussa process
a) Retort; b) Cooling bunker; c) Cooler; d) Scrubber; e) Blower; f) Gas heater

The retort has the following size: 100 m^3 volume, 5 m diameter, 8.5 m height. The retort is fed by a belt conveyer. The wood gas is heated to 450 – 550 °C in a heat exchanger and passes through the retort from top to bottom. The process of carbonization starts in the upper part of the retort and proceeds downward zone by zone.

The circulating gas is wood gas from which the condensable products (i.e., crude pyroligneous liquor and pyrolysis oil) have been removed in coolers and scrubbers. The purified wood gas is fed into the gas heat exchanger by a blower to heat it to the necessary temperature. Excess gas is discharged and used for heating the wood dryer and the gas heater.

The finished charcoal is allowed to fall into enclosed air-tight cooling bunkers. After cooling and saturation with atmospheric oxygen, the charcoal is transferred on belts to a separator. The lump charcoal is shipped in bulk or in bags. The fines are processed to granules or dust, or are pressed to briquettes. The capacity of a large-volume retort is 300 t of charcoal per month. Normally a charcoal plant has six or seven of such units.

For the production of 100 kg of charcoal, including the drying of the wood, 250 MJ of heat and 27 MJ of electrical power are needed, as well as 5 m^3 of water.

The continuous retort [16] was developed from experience gathered using the discontinuous charcoal process.

4.3. New Carbonization Processes

In North and South America and in Japan, charcoal is produced without recovering the byproducts. To carbonize sawdust, barks, and other wastes, Nichols (New York) and BSP (Brisbane, Calif.) have developed a *multiple-hearth furnace* (Herreshoff furnace, Fig. 3) [17]. This furnace consists of 4 – 12 hearths, with capacities of 700 – 2800 t per month.

The raw material is transported by a screw conveyer or a rotary valve to the uppermost hearth. The rakes of the air-cooled vertical hollow shaft move the material spirally across each hearth. The hearths alternately contain drop holes either near the shaft or close to the outer shell. The solid material drops through these openings from hearth to hearth. The charcoal is discharged by a screw conveyer in which water is sprayed.

The reaction gases are burned on the hearths. The air supply is controlled automatically, depending on the temperature. The gases move up countercurrent to the solid material, leaving the furnace at the top through two openings and being completely burned in a chimney. The heat

can be used for steam generation or for predrying the raw material.

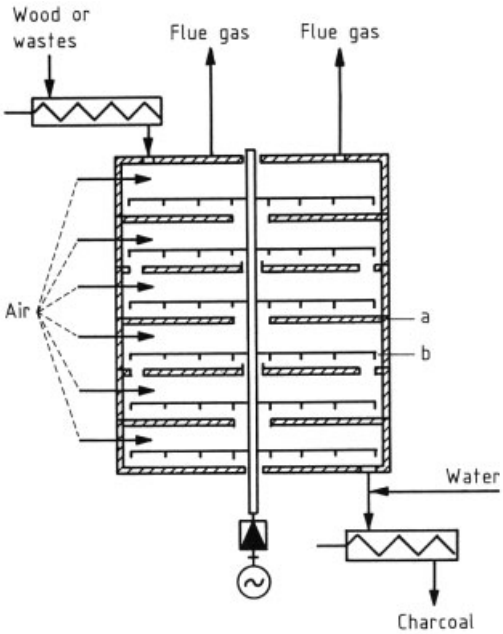

Figure 3. Multiple-hearth furnace
a) Hearth; b) Rakes

The Herreshoff furnace operates like a pile or kiln. The reaction gases and a part of the raw material are burned to generate heat. Therefore, the charcoal yield is only 20 – 24 % on the basis of absolutely dry wood. Since the wet charcoal fines are not directly usable, they are briquetted with starch binder and dried. Complete briquetting units with a capacity of 1 – 6 t/h are built by Aeroglide (Raleigh, N.C.).

Charcoal production in Japan was described by S. Kishimoto and G. Sygiura [18].

In the United States and South America, much charcoal is produced in masonry or concrete kilns [5], [9]. Well-known are the beehives and the Missouri kilns [19].

A series of suggestions has been made about carbonization of wood and vegetable wastes in continuously operating processes, which are self-sufficient in heat [20–22]. For the carbonization of small pieces of raw material, the pieces must be kept moving. This can be done by a fluidized bed, a rotary kiln, stirring arms, or screw conveyer transporters.

A number of processes that produce charcoal or torrefied wood have been described in the last ten years: the Wienhaus plant [23], the Erco reactor [24], the Enerco pyrolysis system [25], the Tech-Air process [26], and a process newly developed in France to produce torrefied wood [27]. Finally, there are pyrolysis processes carried out in rotary kilns, e.g., that of Angelo Industries (Jonesboro, Ark.) and the rotary furnace of Industrial Boiler (Thomasville, Ga.).

5. Environmental Protection

Producing charcoal in a traditional kiln allows the gaseous products to escape into the air and the liquid components to ooze into the ground, resulting in pollution of the environment. Therefore, concrete kilns are preferred for charcoal production. This type of kiln (Missouri kiln) allows the reaction products to be sucked off and burned.

The retort process does not pollute the environment because it is a completely closed system. After separation of the pyrolysis oil, the byproducts are used to generate heat. If the byproducts are recovered, the wastewater, with a BOD of 30 g/L, must be processed in a biological treatment plant. The apparatus and tanks for recovery and storing of liquid byproducts must be made of copper or stainless steel.

Toxicology. In a period of more than 100 years of charcoal production in retorts and of byproduct recovery, neither toxicological damage nor danger to health has been observed.

6. Storage and Transportation

Charcoal can be stored in bulk or in silos. It is shipped in bulk or in containers, in big bags, or in paper bags on pallets. Charcoal cannot be shipped immediately after production, because a certain time is necessary to saturate it with atmospheric oxygen. This reaction is vigorous and has a tendency to self-ignite. In no case should the charcoal to be shipped have a higher temperature than the environment. The German “Regulation for Dangerous Goods” states that lump charcoal can be shipped 4 d after production, and charcoal dust or granulated charcoal can be shipped 8 d after production.

7. Uses

Because of its adsorptive properties, charcoal is a convenient raw material for the production of activated carbon (→ Carbon, Chap. 5.), e.g., DeguSorb (Degussa). Charcoal granules are used as a food additive for domestic animals, e.g., Lorch (Degussa). Because of its reactivity and purity, charcoal is employed for the production of carbon disulfide and alkali-metal cyanide. In addition, charcoal is used for the production of black powder.

Metallurgical uses of charcoal include production of ferrosilicon, pure silicon, ferromanganese, and refining copper. Molded charcoal is used as a carburizing granulate, e.g., Durferrit (Degussa). Charcoal granules are used for tempering. Charcoal is used in foundries as well as for descaling of metal sheets and wires. In countries without deposits of coal, charcoal is used in blast-furnace processes.

Charcoal fines are mixed with the binder starch and pressed to briquettes for use in industry and the household, and for barbecuing, e.g., Grillis (Degussa), Kingsford, and Match Light (Kingsford). Charcoal briquettes burn longer than lump charcoal. The briquettes burn without flame or odor, generating a constant supply of heat. There is a constantly increasing demand for charcoal for barbecuing.

8. References

General References

1. M. Klar: *Technologie der Holzverkohlung,* 2nd ed., Springer Verlag, Berlin 1923.
2. H. M. Bunbury: *The Destructive Distillation of Wood,* Benn Brothers, London 1923.
3. L. F. Hawley: *Wood Distillation,* Chemical Catalog Company, New York 1923.
4. W. Emrich: *Handbook of Charcoal Making*, D. Reidel Publ. Co., Dordrecht-Boston-Lancaster 1985.

Specific References

5. *Winnacker-Küchler,* 3rd ed., **vol. 3,** pp. 417 – 435.
6. N. I. Nikitin: *Chemie des Holzes,* Akademie-Verlag, Berlin (GDR) 1955, pp. 416 – 453.
7. A. W. Goos, A. A. Reiter, *Ind. Eng. Chem.* **38** (1946) 132.
8. A. F. Roberts, *Combust. Flame* **17** (1971) 79 – 86.
9. *Ullmann,* 3rd ed., **8,** 579 – 590.
10. Etablissement Lambiotte Frères, FR 760 593, 1933.
11. Degussa, DE 592 119, 1930 (Th. Reichert).
12. Degussa, DE 666 387, 1932 (Th. Reichert).
13. Degussa, DE 712 552, 1937 (Th. Reichert).
14. Degussa, DE 713 290, 1937 (Th. Reichert).
15. Degussa, DE 744 135, 1937 (Th. Reichert).
16. Etablissement Lambiotte Frères, FR 859 741, 1939.
17. Nichols, BE 709 363, 1968.
18. *Wood Fuels in Japan,* Wood Carbonization Laboratory, Forestry and Forest Products Research Institute, Tokyo 1981.
19. J. P. Jarvis, *Bull. Univ. Mo. Eng. Exp. Stn. Ser.* no. 48 (1960).
20. W. G. Nelson, *Ind. Eng. Chem.* **22** (1930) 312.
21. L. C. Burt, US 3 201 329, 1965.
22. H. L. Barnebey, US 3 525 674, 1970.
23. O. Wienhaus, *Holztechnologie* **20** (1979) 144 – 148.
24. M. S. Lin, R. Sevenius, *For. Prod. J.* **26** (1976) no. 9, 56 – 59.
25. White et al., US 4 165 216, 1979.
26. M. D. Bowen et al., *ACS Symp. Ser.* no. 76 (1978).
27. Armines, FR 8 116 461, 1982.

Cheese, Processed Cheese, and Whey

Guido Burkhalter, Swiss Cheese Union, Bern, Switzerland (Chaps. 1, 2, and 8)

Christian Steffen, Swiss Federal Institute for Dairy Research, Liebefeld-Bern, Switzerland (Chaps. 3 – 5)

Zdenko Puhan, Institute of Food Science, Swiss Federal Institute of Technology, Zürich, Switzerland (Chaps. 6 and 7)

1. Introduction 1
2. Classification 2
3. Production 2
3.1. Raw Materials 2
3.2. Cheesemaking Technology 4
3.3. Ripening 5
4. Analysis 6
5. Flavor Components 6
6. Processed Cheese 6
7. Whey 7
7.1. Composition and Treatment of Whey 7
7.2. Use of Whey 7
8. Economic Aspects 8
9. References 9

1. Introduction

The names cheese, kaas, Käse, and queso, as well as fromage and formaggio, are derived from the Latin words *caseus* (cheese) and *coagulum formatum* (shaped curds) [10], [11].

Man probably already had knowledge of the production of cheese from reindeer milk well before 10 000 B.C. Several thousand years ago, herdsmen of the Alps and Orient began to intentionally separate milk solids (milk fat, enriched casein) from milk liquids (whey). They began with the milk of sheep and later used that of goats and cows. The two phases were produced by the spontaneous acid curdling of the milk in vessels and rennin curdling in the stomachs of mammals. These vessels were used as containers, and the fat curd was subsequently transferred to a basket or vessel with holes. The result was an acid- or rennet-fresh cheese. The first historical evidence of cheese production (written evidence from about 5000 B.C., illustrated from about 3000 or 4000 B.C.) was discovered at Ur (Mesopotamia), in Egypt, and in India. Reports are available on highly developed cheese manufacture in ancient Persia and Greece (Homer, Aristotle). Curdling of milk by the use of rennin or fig leaves and subsequent heating, shaping, pressing, salting, spicing, smoking, and ripening produced a durable cheese that was marketed over considerable distances. In some cases it was suitable for grating. Cheese made from mare's milk was reported in the Black Sea area and the eastern Mediterranean area (Bible, Talmud). Ancient Rome was supplied with cheese from Gaul (caseus gallicus) and from the Celtic area of the Alps (caseus helveticus, caseus alpinus). The latter cheeses were independently developed by their ancestral peoples (Illyrians, Raetians).

During the first century A.D., the production of cheese was introduced to the Roman area of Germany. The cloisters spread the knowledge and skills, and from about 1000 A.D. the development of larger cities brought an increase in cheese production. This period witnessed the emergence of many important cheeses (Roquefort, Gruyère, Edam, Gouda, Parmesan, Emmentaler, Cheddar). These cheeses were frequently marketed over considerable distances from central market places and were copied in other countries. With population growth and the development of transportation, cheesemaking equipment, and scientific knowledge, cheese has gained a major share in world trade and in the nutrition of developed nations [12], [13].

Cheese is the fresh or matured solid or semisolid product obtained by one of the following methods:

a) coagulating milk, skimmed milk, partly skimmed milk, cream, whey cream, or buttermilk, or any combination of these materials, through the action of rennet or other suit-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_163

able coagulating agents, and partially draining the whey that results from such coagulation;

b) processing techniques involving coagulation of milk or materials obtained from milk which give an end product that has the same essential physical, chemical, and organoleptic characteristics as the product defined in (a).

Cured or ripened cheese is cheese which is not ready for consumption shortly after manufacture but which must be held for such a time, at such a temperature, and under such other conditions as will result in the necessary biochemical and physical changes characteristic of the cheese.

Mold-cured or mold-ripened cheese is a cured cheese in which the curing has been accomplished primarily by the development of characteristic mold growth throughout the interior or on the surface of the cheese.

Uncured, unripened, or fresh cheese is cheese which is ready for consumption shortly after manufacture [14].

2. Classification

Because of the extraordinarily diverse characteristics of the various cheeses, there exists no complete classification system [10], [11], [15]. The individual cheeses are ascribed according to important classification criteria:

1) *Source of the milk*: cow, sheep, goat, buffalo, or mare
2) *Kind of milk coagulation*: precipitation of casein by rennet, lactic acid, or a mixture consisting of rennet and lactic acid at ca. 30 – 35 °C; precipitation of casein and whey protein by lactic acid at ca. 90 – 92 °C (Ziger cheese)
3) *Consistency*: extra-hard = below 51 %, hard = 49 – 56 %, semihard = 54 – 63 %, semisoft = 61 – 69 %, soft = over 67 % moisture on a fat-free basis [14]
4) *Fat content*: high fat = over 60 %, full fat = 45 – 60 %, medium fat = 25 – 45 %, low fat = 10 – 25 %, skim = below 10 % fat on the dry basis [14]
5) *Ripening*: uncured cheese (fresh cheese, quarg); cured cheese with or without rind with a smeared surface, with or without white or blue mold on the rind or in the cheese
6) *Potential use*: grating, planing, scratching, cutting, or spreading; meltable cheese (cheese fondue, raclette); cheese as raw material for cheese products (cheese preparations, processed cheese)

The characteristics of some of the best-known cheeses are presented in Table 1 [13], [15].

3. Production

3.1. Raw Materials

The raw material for cheese manufacture is fresh raw milk as produced by regular and thorough milking of healthy, normally fed dairy animals. The milk of several mammals can be used, but the major sources are cows, sheep, goats, buffaloes, and mares. Cheese is also made from cow's milk mixed with milk from any of the other sources.

Cheese milk must comply with general food hygiene requirements for consumer milk (i.e., it must be free from microorganisms pathogenic to humans, toxins, hazardous foreign matter, and undesirable flavors). Additionally, it must meet certain standards regarding its chemical composition, coagulation properties, and microbial condition. It must contain no antimicrobial residues, particularly antibiotics [16–18].

The cheese yield depends primarily on the fat and protein contents of the milk. The salt (particularly calcium salts) and protein concentrations, as well as the acidity, are very important for coagulation.

Possible individual peculiarities and physiological variations in the composition of milk from certain animals or herds may be largely compensated for by using mixed milk from different animals and herds for cheesemaking.

The microbial requirements for cheese milk differ for pasteurized milk and raw milk. If milk is pasteurized before processing, the types of bacteria present in the milk are important because some of them, particularly the heatresistant spore-formers, e.g., the butyric acid bacteria (*Clostridium tyrobutyricum*), are detrimental to cheesemaking (late blowing of the cheese

Table 1. Characteristics of some of the best-known cheeses [13], [15]

Designation	Country of origin	Characteristics
Parmesan	Italy	made from cow's milk extrahard, medium fat, very small holes or none, cured; used for slicing
Emmentaler	Switzerland	made from cow's milk hard, full fat, cured, big round holes; used for slicing, grating, and for fondue
Gruyère	Switzerland	made from cow's milk hard, full fat, cured, small round holes; used for slicing, grating, and for fondue
Cheshire	England	made from cow's milk hard, full fat, cured, irregular small holes, used for slicing
Cheddar	England	made from cow's milk hard, full fat, cured, no holes; used for slicing
Fontina	Italy	made from cow's milk semihard, full fat, cured, irregular small holes; used for slicing
Edam	Netherlands	made from cow's milk semihard, medium fat, cured, small round holes; used for slicing
Gouda	Netherlands	made from cow's milk semihard, full fat, cured, small round holes; used for slicing
Tilsiter	Germany ∗	made from cow's milk semihard, medium to high fat, cured, small slit-shaped or round holes, rind with smear; used for slicing
Saint Paulin	France	made from cow's milk semihard, medium fat, cured, very small slit-shaped holes; used for slicing
Gorgonzola	Italy	made from cow's milk semisoft, full fat, cured, blue moldy veins; used for slicing and spreading
Danish blue	Denmark	made from cow's milk semisoft, full to high fat, cured, blue-green moldy veins; used for slicing and spreading
Brie	France	made from cow's milk soft, medium to full fat, cured, rind with white mold; used for slicing
Camembert	France	made from cow's milk soft, medium to full fat, cured, rind with white mold; used for slicing
Romadur	Germany	made from cow's milk soft, low to high fat, cured, rind with smear; used for slicing
Limburger	Belgium	made from cow's milk soft, medium to full fat, cured, rind with smear; used for slicing
Cottage cheese	United States	made from cow's milk soft, skim, uncured (fresh), used for spreading
Mozzarella	Italy	made from cow's or buffalo's milk or both semihard to semisoft, medium to full fat, cured; used for slicing and for pizza
Roquefort	France	made from sheep's milk semihard to semisoft, full fat, cured, blue-green moldy veins, irregular holes; used for slicing and spreading
White cheese (fresh cheese, quark)	unknown	made from cow's, sheep's, goat's, or buffalo's milk, or any combination of these soft, skim to high fat, uncured, no holes, used for spreading
Feta	Greece	made from sheep's, goat's, or cow's milk, or any combination of these soft, full to high fat, cured in salt-brine, used for slicing
Brinza	Romania	made from sheep's milk soft, full fat, cured, used for cutting

∗ According to the name of the Lithuanian city.

loaves). If untreated raw milk is processed, both the type of bacteria and the bacterial count are essential to the quality of the product.

The quantitative and qualitative compositions of the microbial milk flora depend on the hygienic conditions of milk production and on what the animals were fed, e.g., silage containing *Clostridium tyrobutyricum.* Milk produced under hygienic conditions contains less than 20 000 germs/mL. It generally contains less than 1000 germs/mL upon leaving the teats of a healthy udder. The other germs enter the milk from the surfaces of the milking equipment and the surrounding air during milking.

A normal microbial milk flora consists predominantly of gram-positive acid-producing *Streptococci* and *Lactobacilli.* It also includes psychrotrophic and thermophilic bacteria, e.g., *Enterobacteriaceae, Pseudomonas* spp., *Achromobacter* or *Micrococci,* spore-forming bacilli and *Clostridia.* To comply with the microbial standards of cheese milk, the growth of these microorganisms must be prevented by appropriate cooling and prompt processing of the milk.

Milk from animals suffering from febrile diseases or clinical or subclinical mastitis does not fulfill the requirements for cheese milk. Its chemical, biochemical, and microbial compositions are changed in a way that hinders rennet coagulation or fermentation.

Inappropriate administration of veterinary drugs, particularly antibiotics, or nonobservance of withholding times for milk use following medication favor the presence of substances in milk which may impair the fermentation processes essential for cheesemaking and ripening.

Though cheese milk should contain no microorganisms pathogenic for humans, such pathogens cannot always be excluded entirely. There have been a few reports of diseases provoked by the consumption of cheese made from such milk. Owing to the inhibitory effect of lactic acid bacteria, most pathogens do not survive the conditions of cheese making and ripening [19].

3.2. Cheesemaking Technology

Most production techniques are intended to selectively increase the casein, fat, and dry matter contents in cheese by expelling whey. Moreover, they create appropriate conditions that permit lactic fermentation and ripening to develop in a way that is specific to each type of cheese. The principal cheese quality criteria are odor, taste, body texture, shape, appearance, and in some cases eye formation.

Milk Treatment. The treatment of milk before it is processed into cheese depends on the microbiological quality of the milk, the temperatures applied during cheesemaking, and the duration of ripening. Milk is treated primarily to eliminate undesirable microorganisms. With a high-quality raw milk and use of appropriate technology, it is possible to use untreated raw milk for cheesemaking. The removal of the bacterial flora from milk can be achieved by heat treatments such as thermization (62 – 65 °C for 10 – 20 min) or pasteurization, bactofugation (elimination of microorganisms by centrifugation), and peroxide – catalase treatment. The addition of nitrate removes certain bacterial groups, e.g., *Clostridia.* Cheese milk can be standardized by centrifugation (adjustment of the fat content), addition of skim milk powder or whey protein, or ultrafiltration.

Cheese quality depends largely on the *starter cultures* added to the milk to induce lactic fermentation and ripening before processing. The starters used for lactic fermentation consist of lactic acid bacteria and are introduced in amounts ranging from 0.05 to 5 wt %. Some types of cheese require additional admixtures of mold spores (mold-ripened cheese) or propionic acid bacteria, e.g., Emmentaler.

Coagulation. The cheesemaking process begins with coagulation of the milk in a vessel (kettle, vat, or cheesemaking machine). Rennet enzymes, acid, or combinations of enzymes and acid coagulate the casein, and the resulting casein micelles form a three-dimensional network enclosing the other milk components. This is called the coagulum. The enzymes used are chymosin and pepsin from calf or other bovine stomachs or enzymes from microorganisms. Generally, the coagulation takes place at a temperature of 28 – 35 °C and lasts 10 – 50 min. The acid produced by lactic acid bacteria before rennet is added lowers the pH and consequently accelerates coagulation. Some types of cheese are made by using solely this acid coagulation (fresh or unripened cheese), in some cases with the addition of an organic acid (cottage cheese).

Curd Making. As soon as the coagulum has attained the desired consistency, it is cut into cubes. In this curd-making process, water exudes from the rennet coagulum along with dissolved lactose, whey protein, and salts. The solid curds are thereby separated from the liquid whey. The diameter of the curd particles varies between 1 mm and 3 cm. Under the influence of the rennet enzymes and the acid produced by the lactic acid bacteria, the curd particles shrink and become firmer. This process is called *syneresis.* Whey expulsion is subsequently regulated according to the final water content desired in the ripe cheese. Large curd surface area (small curd particles), long and vigorous stirring, strong acid development, and cooking of the whey – curd mixture favor curd drying. For some hard cheese varieties, the cooking temperature is raised to 60 °C and stirring is continued at this temperature for some time.

Curd Separation. When the curd is firm enough, it is separated from the whey by a process called *dipping* or *hooping*. There are three different methods. The settled curd mass can be transferred in a cloth to the hoop, the whole curd – whey mixture can be transferred to a perforated metal or plastic mold, or the curd can be scooped out into the mold. In the mold, the curd particles fuse together, and the cheese obtains its proper shape. The mass of the hooped cheese varies from 100 g for soft cheese to over 100 kg for Emmentaler.

Whey Drainage. The next stage is whey drainage. This depends on the development of lactic fermentation, which in most cheese types is most rapid in the hours following dipping. Pressing and turning of the cheese in the mold enhances the expulsion of whey.

Salting. All cheese types are salted, except fresh cheese. Salt absorption influences the flavor, keeping quality, body characteristics, and bacterial flora of the cheese, and hence, cheese ripening. Normally, the young cheese is put into a brine bath. Sometimes the salt is added to the curd (e.g., Cheddar) or applied to the cheese surface in the form of a solution.

In *Cheddar manufacture,* the whey is removed and replaced by warm water. The acidification takes place within blocks or in the curd mass. When lactic fermentation is sufficiently advanced, the curd is ground and salted (cheddaring) and subsequently pressed.

In *pasta filata cheese* (e.g., Provolone), lactic fermentation also takes place in the mass of fused curd particles. This mass is treated with hot water or hot whey (at >60 °C), and the curd is then worked by twisting and stretching it until it forms a moldable stringy mass, and finally it is given the desired shape.

3.3. Ripening

After salting, the cheese is submitted to further curing and ripening procedures, depending on the specificity of each variety. Ripening takes place in special rooms, mostly cellars, where temperature, relative air humidity, air motion, and treatment of the cheese are essential factors.

The following processes, more or less intense according to the cheese type, determine the characteristics of the end product.

Lactic Acid Breakdown. In most cheese varieties, part of the lactic acid produced by lactic acid bacteria is decomposed during ripening by the activity of surface molds, surface smear flora, molds inside the loaves, and propionic acid or butyric acid bacteria. This produces further specific fermentation products.

Proteolysis. All cheese types undergo protein breakdown or proteolysis. Due to their proteolytic enzymes, microorganisms inside the cheese loaf decompose the casein to form low-molecular-mass products (peptides and amino acids).

Lipolysis. Some microorganisms (e.g., *Penicillium roqueforti*), because of the enzymatic activity of the lipase, split milk fat into different fatty acids.

According to the curing and ripening procedures used, cheese can be divided according to several characteristics.

Dry-Rind Cheese. In dry-rind cheese, the microbial growth on the cheese surface is suppressed and ripening is stimulated mainly by the microorganisms inside the loaf, e.g., Emmentaler.

Smear-Ripened Cheese. Microbial growth on the cheese surface of smear-ripened cheese is influenced by regular wet treatment and high air humidity (>90 %) e.g., Gruyère, Tilsiter. The principal smear producers are yeasts and *Brevibacterium linens*.

Mold-Ripened Cheese. In numerous cheese varieties, ripening proceeds predominantly from the surface to the center of the loaf due to the growth of white molds (*Penicillium candidum, Penicillium camemberti*, *Geotrichum candidum*), e.g., Camembert.

Blue-Mold Cheese. Blue mold (*Penicillium roqueforti*) develops along air channels pierced into the cheese. Roquefort and Gorgonzola are typical blue-veined cheeses.

Eye Formation. Some cheeses have eyes resulting from gas production by specific bacteria that develop in the cheese mass. In Emmentaler cheese, eye formation comes from propionic fermentation (lactic acid breakdown into propionic acid, acetic acid, and carbon dioxide).

4. Analysis

Certain chemical analyses of *cheese that is less than 1 day old* can give important information on the success of cheesemaking. The most important parameter is the complete transformation of lactose into lactic acid [20]. Other important parameters at this stage are the pH [21] and water content [22]. The determination of the leucine aminopeptidase activity [23] in fresh cheese predicts proteolysis at a later stage.

Proteolysis plays a decisive role in *cheese ripening.* Determinations of the water-soluble nitrogen fraction [24], [25], the nonprotein fraction [24], peptides [25], [26], free amino acids [27], amines [28], and ammonia [29] indicate the type and extent of proteolysis. In some cheeses such as Emmentaler, where important quantities of lactic acid are at a later stage transformed into acetic acid, propionic acid, and carbon dioxide, determination of their relative concentrations is relevant [30].

Specifications for the Marketplace. *Marketable cheese,* prior to being sold, must comply with legal standards for fat [31] and water contents [22]. In most countries these concentrations must be declared. Residue concentrations of pesticides [32] and heavy metals [33] are limited by law.

The types and amounts of additives used during cheesemaking are also prescribed in food regulations. Some countries forbid the use of certain additives, whereas others set limits. The most frequently used additives are sodium chloride [34], calcium chloride [35], sodium nitrate or potassium nitrate [36], sodium sorbate or potassium sorbate [37], and antibiotics, such as nisin. Sorbate and nisin are used to prevent undesirable mold growth on cheese rind.

5. Flavor Components

The flavor of cheese, which is strongly dependent on the texture, is due to the presence of a complex mixture of volatile and nonvolatile odorous and sapid components. These compounds result from numerous chemical and biochemical (enzymatic) reactions that take place during the making and ripening of cheese. Some substances originate from the milk used for cheesemaking and others from the various additives, such as coagulating agents, salts, and spices. The result is that both the qualitative and quantitative compositions of the flavor vary considerably depending on the bacterial cultures and technology used. These compounds come from the catabolism of the carbohydrate, protein, and lipid components of the cheese during ripening. Commonly found are fatty acids (mainly short chain); α-hydroxycarboxylic acids, primary and secondary alcohols, carbonyl compounds (especially methyl ketones and aldehydes), esters, lactones, pyrazines (and other nitrogen- and sulfur-containing compounds), sesquiterpenes, and traces of aromatic components. Peptides, amino acids, and other casein decomposition products as well as mono-, di-, and triglycerides also contribute to the taste of cheese. Many metabolic pathways have been described for both the desired and undesired flavors [38–52].

6. Processed Cheese

Processed cheese is a secondary product made from the primary product cheese by adding emulsifying salts and applying heat treatment. In 1911, two Swiss cheesemakers developed a method for producing a long-life processed cheese, suitable for export, that can be stored for several months in tropical climates without undergoing changes. A large variety of products with varying consistencies and flavors are produced.

Basically, the production of processed cheese consists of transforming the partially watersoluble casein into a soluble form by adding emulsifying salts and water and by applying heat. All constituents, such as fat, protein, and other ingredients, must be distributed homogeneously in the product. A total of 2 – 3 wt % of emulsifying salts is added to the cheese to be processed. Citrates and ortho-, pyro-, and polyphosphates of sodium and potassium are used. The pH of the

cheese mass can vary from 5.4 to more than 6.0, depending on the required consistency.

The following physicochemical steps can be distinguished in the production of processed cheese:

1) Replacement of calcium in the casein by sodium or potassium
2) Peptization and dispersion
3) Hydration (water binding)
4) Stabilization of pH
5) Structure formation during cooling

A mixture of different cheeses, such as Emmentaler, Gouda, Tilsiter, Camembert, is chosen as the starting material. Milk, whey powder, cream, butter, herbs, meat, vegetables, or fruit can be added. The consistency of the product can be spreadable or firm, depending on the water content and the nature of the cheeses processed. Processed cheese is made in the following steps:

1) Derinding and cleaning
2) Coarse cutting and mincing
3) Adding emulsifying salts and ingredients
4) Cooking batchwise with continuous agitation and direct steam injection at 75 – 96 °C (up to 140 °C in some continuous cookers)
5) Pouring the hot processed cheese into molds, blocks, or cans
6) Cooling to 20 °C or lower for a period that depends on the type of processed cheese (the softer the shorter)

High-quality processed cheese requires starting materials of good quality.

7. Whey

Whey is the byproduct of cheese and casein production that remains after the selective coagulation of the casein. Whey contains all of the water-soluble constituents of the milk along with whey proteins and some fat. The concentrations of the constituents in the whey depend on the quantity of water added to the milk prior to coagulation or during cheesemaking.

7.1. Composition and Treatment of Whey

Typical rennet whey contains 6.5 – 6.8 wt % dry matter. The remainder is water. The dry matter constituents are 5.0 – 5.2 wt % lactose, 0.4 – 0.5 wt % ash, 0.2 – 0.5 wt % fat, 0.4 – 0.5 wt % protein, and 0.15 – 0.20 wt % other nitrogenous substances [53]. The content of water-soluble vitamins is equal to that of milk.

Whey proteins are its most important nutritional constituents [54]. Varying amounts of the whey proteins are denatured and remain in the cheese, depending on the heat treatment of the milk prior to the cheesemaking. Whey separated from cheese made with raw milk contains all of the natural whey proteins. Up to 10 % of the whey proteins can be denatured during the pasteurization normally applied in the cheesemaking (70 – 75 °C for 20 – 60 s) [55].

In milk, 95 % of the nitrogen exists in the form of proteins, but proteins in whey account for only 70 – 75 % of the nitrogen. The remaining nitrogen has been split from the casein during rennet coagulation and is not protein [53], [56].

7.2. Use of Whey

Whey can be used in many ways and is by no means a waste product. Whey has always been used as feed for pigs. In Switzerland, the largest part of whey is still used for this purpose. In some parts of the country, a pig fattening unit is directly attached to the cheese factory to eliminate transportation. However, the transport of whey with only 6 – 7 wt % dry matter from the cheese factory to the user may not be cost effective [57].

Whey is used as a raw material for numerous products in human and animal nutrition, in the pharmaceutical and fermentation industries, etc. [58], [59]. Basically, there is a difference between products made from whey containing total dry matter and those using only some of its constituents. Regardless of the product to be manufactured, the whey must be cleared of fat and cheese dust (coagulated casein) by centrifugation prior to processing.

The simplest process concentrates the total dry matter by dehydration. The concentrate obtained by evaporation or hyperfiltration can contain up to 50 wt % dry matter. Further dehydration by spray drying yields whey powder with a water content of no more than 4 – 5 wt %.

The dry matter in turn consists of 71 – 75 wt % lactose, 12 – 14 wt % protein (N %×6.38), and 8 – 9 wt % mineral salts. Whey concentrate and whey powder are used in the production of baby food, ice cream, and other products [60], [61]. Large quantities of whey powder are also incorporated into animal feeds. The nutritive value of 1 kg of whey dry matter corresponds approximately to that of 1 kg of barley.

The content of mineral salts in whey can be reduced by electrodialysis. The more expensive demineralized whey powder is more suitable for the manufacture of baby food [62].

An important group of products from whey are the whey proteins. The whey proteins are concentrated by acid – heat precipitation in which the whey proteins are denatured, lose their solubility, and separate from the liquid. The precipitated protein is then washed and dried. If processed carefully, the high biological value of the whey proteins is not affected. The product is known as lactalbumin; it contains more than 90 wt % protein and is used primarily for baby food [63]. Great efforts have been made to produce whey protein concentrates containing only slightly denatured whey proteins on an industrial scale. Separation processes such as ultrafiltration allow the separation of whey proteins from whey without changing their properties. Such concentrates contain 10 – 20 wt % protein. Careful drying allows further concentration without substantially damaging the whey proteins. Whey proteins obtained by ultrafiltration can be used in numerous foods for protein enrichment and because of their functional properties, such as solubility, viscosity, foaming, gelling, and emulsification [58], [59], [64], [65].

Deproteinized whey is used to make lactose, the milk-specific dissaccharide consisting of glucose and galactose. Deproteinized whey containing 5 wt % lactose is first concentrated to 60 –70 wt % dry matter, whereupon lactose crystallizes under controlled conditions. The final product contains 98 – 99.5 wt % lactose, depending on the quality requirements [66]. It is used in the food production and the pharmaceutical industry. To increase its sweetness and solubility, the lactose can be split into the monosaccharides, glucose and galactose, by acid or enzymatic hydrolysis (β-galactosidase). Hydrolyzed, deproteinized whey is concentrated to form *lactose syrup* with up to 80 wt % dry solids. This is used in food and candy making [67–70].

Moreover, whey is most suitable for the production of beverages. It is fermented with lactic acid bacteria and then clarified. The best known product is Rivella, produced in Switzerland. In comparison with other beverages, the products made from whey contain lactic acid, mineral salts, and soluble vitamins from milk [71].

In some countries, a whey cheese is produced by evaporating whey alone or together with milk or cream. Because of their specific tastes, these products have only regional significance [13].

Whey is a suitable substrate for the production of biomass [72], [73], ethanol [74], products like lactic acid [75], lactosyl urea [76], and other organic chemicals [76].

An excellent annotated bibliography on "Utilization of by-products from whey" is issued periodically by Commonwealth Agricultural Bureaux in England. The most recent (D20B), covering the period 1982 – 1984, was issued in June, 1985.

8. Economic Aspects

The approximate world production of cheese in 1984 was 12.3 Mt, representing an approximate world market value of $ 25 000 000 000 [77] Table 2. The world cheese production has been steadily growing and was in 1998 at 14.8 Mt [78].

Table 2. Some major cheese-producing regions

	Production, Mt	Production share, %
Western Europe	4.4	36
Canada and the United States	2.3	19
FormerSoviet Union	0.8	7
Australia and New Zealand	0.3	2

Cheddar cheese with approximately 1.7 Mt represented the largest share of production, followed by Emmentaler with about 0.6 Mt [77].

With 20 kg per capita, France has the highest annual consumption of cheese [77].

9. References

General References

1. J. G. Davis: *Cheese, Basic Technology,* vol. I, J. & A. Churchill Ltd., London 1965.
2. J. G. Davis: *Cheese, Manufacturing Methods,* vol. III, Churchill Livingstone, Edinburgh-London-New York 1976.
3. J. G. Davis: *Cheese, Annotated Bibliography with Subject Index,* vol. IV, Churchill Livingstone, Edingburgh-London-New York 1975.
4. H. Mair-Waldburg: *Handbuch der Käse, Käse der Welt von A – Z,* Volkswirtschaftlicher Verlag, Kempten 1974.
5. G. W. Reinbold: *Swiss Cheese Varieties, Pfizer Cheese Monography,* vol. V, Pfizer Inc., New York 1972.
6. M. E. Schwartz: *Cheese-Making Technology, Food Technology Review no. 7,* Noyes Data Corp., London 1973.
7. M. G. van den Berg: "The Thermization of Milk," *IDF Bull.* no. 182 (1984).
8. R. Scott: *Cheesemaking Practice,* Appl. Science Publ., London 1981.
9. A. Eck: *Le Fromage,* Technique et Documentation, Paris 1984.

Specific References

10. M. E. Schulz, E. Voss: *Das große Molkerei-Lexikon,* Volkswirtschaftlicher Verlag, Kempten 1965.
11. J. G. Davis: *Cheese,* J. & A. Churchill Ltd., London 1965.
12. F. Kosikowski: *Cheese and Fermented Milk Foods,* Edwards Brothers, Inc., Ann Arbor, Michigan, 1966.
13. H. Mair-Waldburg: *Handbuch der Käse,* Volkswirtschaftlicher Verlag, Kempten 1974.
14. *Recommended General Standard for Cheese A-6, Code of Principles concerning Milk and Milk Products,* Joint FAO/WHO Food Standards Programme, Rome 1978
15. G. Burkhalter: *IDF-Catalogue of Cheeses,* Document 141, International Dairy Federation, Brussels 1981.
16. Bundesanstalt für Milchforschung: "Die hygienische Wertigkeit der Rohmilch und Empfehlungen zu ihrer weiteren Verbesserung," *Kiel. Milchwirtsch. Forschungsber.* **21** (1969) no. 3, 201.
17. H. O. Gravert: *Die Milch – Erzeugung, Gewinnung, Qualität,* E. Ulmer Verlag, Stuttgart 1983.
18. G. H. Schmitt, L. D. van Vleck: *Principles of Dairy Science,* W. H. Freeman Comp., San Francisco 1974. A. Töpel: *Chemie und Physik der Milch,* 2nd ed., VEB Fachbuchverlag, Leipzig 1981.
19. "Behaviour of Pathogens in Cheese," *IDF Bull.* 1980 no. 122.
20. C. Steffen, B. Nick, B. Blanc, *Schweiz. Milchw. Forsch.* **4** (1975) 13.
21. F. Kiermeier, E. Lechner: *Milch und Milcherzeugnisse,* P. Parey, Berlin-Hamburg 1973, p. 401.
22. International Standard FIL-IDF 4 A: 1982, International Dairy Federation, Square Vergote 41, 1040 Brussels (Belgium).
23. C. Steffen, B. Nick, H. Fiechter, *Schweiz. Milchztg.* **101** (1975) 123.
24. G. Steiger, E. Flückiger, *Schweiz. Milchw. Forsch.* **8** (1979) 39.
25. C. N. Kuchroo, P. F. Fox, *Milchwissenschaft* **37** (1982) 331.
26. F. Lottspeich, A. Henschen, K.-P. Hupe: *High Performance Liquid Chromatography in Protein and Peptide Chemistry,* Proceedings of Int. Symposium, Max-Planck-Institut, Martinsried-Munich 1981.
27. A. A. Ismaïl, K. Hansen, *Milchwissenschaft* **27** (1972) 556.
28. P. Lavanchy, Ch. Bühlmann, G. Steiger, *Schweiz. Milchw. Forsch.* **14** (1985) 3.
29. M. Mrowetz, *Milchwissenschaft* **34** (1979) 593.
30. J. F. Horwood, G. T. Lloyd, *Aust. J. Dairy Technol.* **35** (1980) 30.
31. International Standard FIL-IDF 5 A: 1969, International Dairy Federation, Square Vergote 41, 1040 Brussels (Belgium).
32. International Standard FIL-IDF 75 B: 1983, International Dairy Federation, Square Vergote 41, 1040 Brussels (Belgium).
33. W. Horwitz: *Official Methods of Analysis of the AOAC, Chap. 25: "Metals and Other Elements at Trace Levels in Foods,"* 14th ed., Association of Official Analytical Chemists, Washington, DC, 1984.
34. International Standard FIL-IDF 88: 1979, International Dairy Federation, Square Vergote 41, 1040 Brussels (Belgium).
35. K. N. Pearce, *N. Z. J. Dairy Sci. Technol.* **12** (1977) 113.
36. International Standard FIL-IDF 84: 1978, International Dairy Federation, Square Vergote 41, 1040 Brussels (Belgium).
37. W. Horwitz: *Official Methods of Analysis of the AOAC, Methods 20.115 – 20.118,* 14th ed.,

Association of Official Analytical Chemists, Washington, DC, 1984.
38. J. Adda: "Formation de la Flaveur," in A. Eck (ed.): *Le Fromage,* Technique et Documentation (Lavoisier), Paris 1984, pp. 330 – 340.
39. J. Adda, J. C. Gripon, L. Vassal: "The Chemistry of Flavour and Texture Generation in Cheese," *Food Chem.* **9** (1982) 115 – 129.
40. J.-P. Dumont, J. Adda: "Flavour Formation in Dairy Products," in D. G. Land, H. E. Nursten (eds.): *Progress in Flavour Research,* Applied Science Publ. Ltd., London 1979, pp. 245 – 262.
41. J. Adda, S. Roger, J.-P. Dumont: "Some Recent Advances in the Knowledge of Cheese Flavor," in G. Charalambous, G. E. Inglett (eds.): *Flavor of Foods and Beverages, Chemistry and Technology,* Academic Press, New York-San Francisco-London 1978, pp.65 – 74.
42. K. H. Ney: "Recent Advances in Cheese Flavor Research," in G. Charalambous, G. Inglett (eds.): *The Quality of Foods and Beverages, Chemistry and Technology,* vol. 1, Academic Press, New York-London-Toronto-Sydney-San Francisco 1981, pp. 389 – 435.
43. U. Behnke: "Zur Biogenese des Käsearomas," *Nahrung* **24** (1980) no. 1, 71 – 83.
44. G. J. Moskowitz: "Flavor Development in Cheese," in G. Charalambous (ed.): *The Analysis and Control of Less Desirable Flavors in Foods and Beverages,* Academic Press, New York-London-Toronto-Sydney-San Francisco 1980, pp. 53 – 70.
45. D. Lamparsky, I. Klimes: "Cheddar Cheese Flavour – Its Formation in the Light of New Analytical Results," in P. Schreier (ed.): *Flavour 81,* De Gruyter, Berlin-New York 1981, pp. 557 – 588.
46. B. A. Law: "The Formation of Aroma and Flavour Compounds in Fermented Dairy Products," *Dairy Sci. Abstr.* **43** (1981) no. 3, 143 – 154, Review Article no. 180.
47. H. T. Badings, R. Neeter: "Recent Advances in the Study of Aroma Compounds of Milk and Dairy Products." *Neth. Milk Dairy J.* **34** (1980) 9 – 30.
48. D. A. Forss: "Review of the Progress of Dairy Science: Mechanisms of Formation of Aroma Compounds in Milk and Milk Products," J. Dairy Res. **46** (1979) 691 – 706.
49. I. D. Morton, A. J. Macleod: *Food Flavours, Part A: Introduction,* Elsevier Scientific Publ. Comp., Amsterdam-Oxford-New York 1982.
50. J. J. Panouse, J. D. Masson, Truonc Thanh Tong: "Les Arômes de Fromage," *Ind. Aliment. Agric.* **89** 133 – 140 (1972).
51. R. Liardon, J. O. Bosset, B. Blanc: "The Aroma Composition of Swiss Gruyère Cheese, Part I: The Alkaline Volatile Components," *Lebensm.-Wiss. Technol.* **15** 143 – 147 (1982).
52. J. O. Bosset, R. Liardon: "The Aroma Composition of Swiss Gruyère Cheese, Part II: The Neutral Volatile Components," *Lebensm.-Wiss. Technol.* **17** (1984) 359 – 362; "Part III: Relative Changes in the Content of Alkaline and Neutral Volatile Components During Ripening," *Lebensm.-Wiss. Technol.* **18** (1985) 178 – 185; "Part IV: The Acidic Volatile Components and Their Changes in Content During Ripening," (in preparation).
53. E. Cousin, M. Collomb, P. Lavanchy, R. Sieber, G. Steiger, B. Blanc, *Schweiz. Milchw. Forsch.* **12** (1983) no. 4, 73 – 76.
54. E. Renner: *Milch und Milchprodukte in der Ernährung des Menschen,* 4th ed., Volkswirtschaftlicher Verlag, München 1982.
55. Z. Puhan, O. Flüeler, *Milchwissenschaft* **29** (1974) 148.
56. P. Walstra, R. Jenness: *Dairy Chemistry and Physics,* J. Wiley & Sons, New York 1984.
57. B. Blanc, E. Flückinger, C. Steiner, *Schweiz. Milchztg.* **103** (1977) no. 96, 796.
58. N. Kosaric, Y. Asher, *Conserv. Recycling* **5** (1982) no. 1, 23.
59. M. T. Gillies: *Whey Processing and Utilization – Economic and Technical Aspects,* Noyes Data Corp., London 1974.
60. B. H. Webb, E. O. Whittier: *Byproducts from Milk,* AVI Publ. Comp. Inc., Westport, Connecticut, 1970.
61. E. J. Mann, *Dairy Ind. Int.* **47** (1982) no. 2, 22.
62. W. Ulrich, *Schweiz. Milchw. Forsch.* **5** (1976) 99.
63. F. M. Mann, *N.Z. J. Dairy Sci. Technol.* **12** (1977) 93.
64. T. E. Galesloot, B. J. Tinberger: *Milk Proteins '84,* Pudoc, Wageningen 1985.
65. J. N. de Wit, *Neth. Milk Dairy J.* **35** (1981) no. 1, 47.
66. R. Landré, *Nordeuropaeisk Mejeri Tidsskrift* **41** (1975) no. 4, 116.
67. R. Boerde, T. Robbertson, *Neth. Milk Dairy J.* **35** (1981) 95.
68. R. Boerde, D. J. Romiju, J. Straatsma, *Neth. Milk Dairy J.* **36** (1982) 317.
69. R. Dicker, *Dairy Ind. Int.* **47** (1982) no. 4, 19.
70. E. J. Mann, *Dairy Ind. Int.* **48** (1983) no. 1, 11.
71. R. K. Robinson, A. Y. Tamine, *Dairy Ind. Int.* **43** (1978) no. 3, 14.

72. J. Meyrath, K. Bayer, *Econ. Microbiol.* **4** (1979) 207.
73. Z. Puhan, N. Halter, *Nordeuropaeisk Mejeri Tidsskrift* **50** (1984) no. 2, 33.
74. L. Reesen, R. Strube, *Process Biochem.* **13** (1978) no. 11, 21.
75. G. C. Cox, R. D. Mac Bean, *Res. Proj. Ser, Victoria Dep. Agric.* **29** (1977) .
76. G. Coton, *Int. Dairy Fed. Doc.* **126** (1980) 23.
77. ZMP, Zentrale Markt- und Preisberichtstelle für Erzeugnisse der Land-, Forst- und Ernährungswirtschaft GmbH, Bonn.
78. Food and Aqriculture Organization of United Nations (FAO), Nov. 1999.

Chelating Agents → Ethylenediaminetetraacetic Acid and Related Chelating Agents
Chelating Agents → Nitrilotriacetic Acid

Chemotherapeutics

Anthelmintics, Antibiotics, Disinfectants, Biocides, and Veterinary Drugs are separate keywords.

Paul Actor, Smith Kline & French Laboratories , Philadelphia, Pennsylvania 19101, United States (Chaps. 1, 3)

Alfred W. Chow, Smith Kline & French Laboratories , Philadelphia, Pennsylvania 19101, United States (Chaps. 1, 3)

Frank J. Dutko, Sterling-Winthrop Research Institute, Rensselaer, New York 12144, United States (Chap. 4)

Mark A. McKinlay, Sterling-Winthrop Research Institute , Rensselaer, New York 12144, United States (Chap. 4)

1. Introduction 3
2. Chemotherapy of Bacterial Infections 3
2.1. Classification of Bacteria Causing Disease 3
2.2. Emergence of New Bacterial Pathogens 5
2.3. Antimicrobial Resistance 5
2.4. Pathogenesis and Virulence Factors 6
2.5. Selection of an Appropriate Antimicrobial Agent 6
2.6. Chemoprophylaxis 7
2.7. Chemotherapy 7
2.7.1. Biochemical Targets for Antimicrobial Agents 7
2.7.2. Quinolone Antibacterial Agents . 8
2.7.2.1. Structure Function 8
2.7.2.2. Mechanism of Action 9
2.7.2.3. Nalidixic Acid and First-Generation Quinolones . . . 9
2.7.2.4. Second-Generation Fluoroquinolones 12
2.7.3. Sulfa Drugs 16
2.7.3.1. Biological Activity and Medical Uses 17
2.7.3.2. Mechanism of Action and Antimicrobial Resistance 17
2.7.3.3. Structure – Function Relationships 18
2.7.3.4. Pharmacokinetics 18
2.7.3.5. Toxicity and Drug Interactions . . 18
2.7.3.6. Combination Therapy 19
2.7.3.7. Rapidly Absorbed Short- and Medium-Acting Sulfa Drugs . . . 19
2.7.3.8. Long-Acting Sulfonamides 21
2.7.3.9. Sulfonamides for Use in the Gastrointestinal Tract 22
2.7.4. Agents for Treating Mycobacterial Infections 22
2.7.4.1. Antituberculosis Agents 23
2.7.4.2. Antileprosy Agents 26
2.7.5. Miscellaneous Nitroheterocycles Used to Treat Bacterial Infection 27
3. Chemotherapy of Protozoan Infections 28
3.1. Classification of Pathogenic Protozoa 28
3.2. Flagellates 28
3.2.1. Hemoflagellates 28
3.2.1.1. African Trypanosomiasis 29
3.2.1.1.1. Biology and Epidemiology 29
3.2.1.1.2. Chemotherapy 29
3.2.1.2. American Trypanosomiasis 31
3.2.1.2.1. Biology and Epidemiology 31
3.2.1.2.2. Chemotherapy 32
3.2.1.3. Leishmaniasis 32
3.2.1.3.1. Biology and Epidemiology 32
3.2.1.3.2. Chemotherapy 33
3.2.2. Intestinal and Urogenital Flagellates 33
3.2.2.1. Trichomonas Vaginalis 34
3.2.2.1.1. Biology and Epidemiology 34
3.2.2.1.2. Chemotherapy 34
3.2.2.2. *Giardia Lamblia* 35
3.2.2.2.1. Biology and Epidemiology 35
3.2.2.2.2. Chemotherapy 35
3.3. Sporozoans 36
3.3.1. Plasmodia 36
3.3.1.1. Biology and Epidemiology 36
3.3.1.2. Chemotherapy 37
3.3.2. Babesia 41
3.3.3. Isosporiasis 41
3.3.4. Toxoplasmosis 41
3.3.5. Cryptosporidium 42
3.3.6. Pneumocystis Carinii 42
3.4. Ciliates 42

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_173

3.5. **Amebas** 43
3.5.1. Biology and Epidemiology 43
3.5.2. Chemotherapy 43
4. **Chemotherapy of Viral Infections** 45
4.1. **Physical and Biological Characteristics** 45
4.2. **Classification of Viruses** 46
4.3. **Assessment of Antiviral Activity in Cell Culture** 46
4.4. **Animal Models of Virus Infection** 47
4.5. **Rationale for Chemotherapy of Viral Infections** 47
4.6. **Chemotherapeutic Agents** . . . 48
4.6.1. Nucleoside Analogues 48
4.6.2. Phosphonoacetate and Phosphonoformate 52
4.6.3. Amantadine and Rimantadine . . 52
4.6.4. Enviroxime 53
4.6.5. 4′,6-Dichloroflavan 54
4.6.6. Chalcone Ro 09 – 0410 54
4.6.7. Arildone and Disoxaril 54
4.6.8. 3′-Azidothymidine 55
4.6.9. Suramin 55
4.6.10. HPA 23 55
5. **References** 55

1. Introduction

This article discusses the synthetic agents that are effective against pathogenic bacteria, protozoa, and viruses. Although a large percentage of the agents employed to treat these infections, especially bacterial infections, are natural products, i.e., antibiotics, many synthetic compounds continue to provide useful alternatives, and some are the agents of choice for the treatment of specific clinical entities.

2. Chemotherapy of Bacterial Infections

2.1. Classification of Bacteria Causing Disease

A broad variety of organisms are primary human pathogens. No universal agreement exists for the classification of these bacteria, and it is beyond the scope of this chapter to discuss the arguments inherent in attempting to create a universally acceptable and useful system. The definitive work in bacterial classification is Bergy's *Manual of Determinative Bacteriology* Bacterial species still are arbitrarily defined by descriptive features. In the approximate sequence of importance, some of the major features employed include the following:

1) *morphological appearance of the cell,* including shape, size, flagellar pattern if motile, capsule occurrence, colonial morphology, and pigmentation
2) *staining properties,* including the gram capsule, spore, and acid fast
3) *metabolic patterns*
4) *macromolecular composition* and *structure*
5) *ecological habitat*
6) ability to be *pathogenic*

The major families of bacteria causing human infection are shown in Table 1. However, many bacteria that are not primary pathogens are capable of causing clinical infections in immunocompromised hosts. Furthermore, transfer of genetic material between bacterial species, e.g., via plasmids, allows saprophytic bacteria to acquire virulence factors and become pathogenic.

2.2. Emergence of New Bacterial Pathogens

The emergence and recognition of organisms as important human pathogens are strongly influenced by three major factors: (1) the transfer of virulence factors and antibiotic resistance between bacterial species by extrachromosomal elements, (2) opportunistic infection of immunocompromised hosts by saprophytic bacteria, and (3) improvement and widespread use in the microbiological clinical laboratory of new diagnostic procedures for isolating and identifying pathogens.

Opportunistic infections, particularly in hospitalized individuals, have become major infectious disease problems. This problem is likely to continue because of the increasing use of instrumentation, antibiotics, and drugs that either

Table 1. The major groups of pathogenic bacteria

Morphological type	Family	Genus	Gram stain	Oxygen utilization ∗
Cocci	Micrococcaceae	*Micrococcus*	+	A
		Staphylococcus	+	A/F – AN
	Streptococcaceae	*Streptococcus*	+	F – AN
	Peptococcaceae	*Peptococcus*	+	AN
		Peptostreptococcus	+	AN
	Neisseriaceae	*Neisseria*	–	A
	Veillonellaceae	*Veillonella*	–	AN
Rods	Enterobacteriaceae	*Escherichia*	–	F – AN
		Shigella	–	F – AN
		Salmonella	–	F – AN
		Citrobacter	–	F – AN
		Klebsiella	–	F – AN
		Enterobacter	–	F – AN
		Erwinia	–	F – AN
		Serratia	–	F – AN
		Hafnia	–	F – AN
		Edwardsiella	–	F – AN
		Proteus	–	F – AN
		Providencia	–	F – AN
		Morganella	–	F – AN
		Yersinia	–	F – AN
	Vibrionaceae	*Vibrio*	–	F – AN
		Aeromonas	–	F – AN
	Pasteurellaceae	*Pasteurella*	–	F – AN
		Haemophilus	–	F – AN
	Pseudomonadaceae	*Pseudomonas*	–	A
	Legionellaceae	*Legionella*	–	A
	Neisseriaceae	*Moraxella*	–	A
		Acinetobacter	–	A
		Brucella	–	A
		Bordetella	–	A
	Bacteroidaceae	*Bacteroides*	–	AN
		Fusobacterium	–	AN
		Leptotricha	–	AN
	Bacillaceae	*Bacillus*	+	AN
		Clostridium	+	AN
		Listeria	+	A
		Erysipelothrix	+	A
Actinomycetes andrelated organisms		*Corynebacterium*	+	A/F – AN
	Propionibacteriaceae	*Propionibacterium*	+	AN
		Eubacterium	+	AN
	Actinomycetaceae	*Actinomyces*	+	FA
	Mycobacteriaceae	*Mycobacterium*	+	A
	Nocardiaceae	*Nocardia*	+	A
Rickettsias and Chlamydias	Rickettsiaceae	*Rickettsia*	–	P
		Coxiella	–	P
	Bartonellaceae	*Bartonella*	–	P
	Anaplasmataceae	*Grahamella*	–	P
		Anaplasma	NR∗∗	P
		Haemobartonella	–	P
		Eperythrozoan	–	P

bypass or reduce the level of natural resistance or the specific immune mechanisms of the host. Thus, microorganisms previously considered as innocuous commensals or contaminants are now able to invade and cause disease. Some of the more important bacterial pathogens that have created significant problems include *Staphylococcus epidermidis, Bacteroides fragilis* and other *Bacteroides* species, *Clostridium difficile,* the gram-negative rods including *Acinetobacter* species, *Serratia, Citrobacter* species, *Yersinia, Moraxella,* and the atypical mycobacteria.

Since the mid-1960s, the role of *S. epidermidis* has changed from a bothersome contami-

Table 1. (Continued)

Morphological type	Family	Genus	Gram stain	Oxygen utilization ∗
	Chlamydiaceae	*Chlamydia*	–	P
Mycoplasmas	Mycoplasmataceae	*Mycoplasma*	–	FA
		Ureaplasma	–	FA
	Acholeplasmataceae	*Acholoplasma*	NR∗∗	FA
Spirochetes	Spirochaetaceae	*Treponema*	–	AN
		Borrelia	–	AN
		Leptospira		A

∗ A= Aerobic; AN = Anaerobic; F – AN = Facultative anaerobic; A/F – AN = Aerobic or facultative anaerobic growth; P = Parasitic usually require host cells for growth.
∗∗ NR = Gram stain not revealed.

nant to a major pathogen causing infections associated with foreign bodies (prosthetic valvular endocarditis and infections of CNS shunts and of joint and vascular prostheses). In addition, this organism can cause infections not associated with foreign bodies, such as endocarditis and urinary tract infections. *Bacteroides fragilis,* the preeminent anaerobic human pathogen, has evolved as a problem. Information as to the importance and widespread incidence of this organism only became available when techniques were established for its facile isolation and rapid diagnosis. Another important anaerobe is *Clostridium difficile,* a component of the normal gut flora, which overgrows to large populations in the presence of antibiotics. Toxigenic stains can cause pseudomembranous colitis, a severe necrotizing disease of the large intestine. With the waning of *M. tuberculosis* infections in the United States, other mycobacterioses, the so-called atypical mycobacteria, have become increasingly important.

2.3. Antimicrobial Resistance

Mechanisms of Resistance. The widespread clinical use of antimicrobics has resulted in the emergence of many strains of bacteria resistant to one or more of these agents. In most cases in which adequate studies have been done, the role of antimicrobial agents apparently is to exert selective pressure, resulting in the emergence of resistant organisms. In some instances, the organisms are *naturally resistant* to the antibiotic used. In other cases, the resistant bacteria may have acquired *R-factors* or *plasmids.* These extrachromosomal agents may provide the organism with the ability to synthesize enzymes that modify or inactivate the antimicrobial agent. The plasmids may also cause changes in the organism's ability to accumulate the antimicrobial agent, or they may stimulate the cell to produce or overproduce metabolic enzymes resistant to inhibition by the antimicrobial agent. Additionally, alterations in the permeability of the bacterial cell envelope could result in drug resistance.

Chromosomal resistance develops as a result of spontaneous mutation in a locus that controls susceptibility to a given antimicrobial agent. The presence of the drug serves as a selection mechanism for drug-resistant mutants. Chromosomal mutants are most commonly resistant by virtue of an alteration in a structural receptor for the drug. Cross resistance is frequently observed between chemically related drugs showing a similar mechanism of action, but may also exist for unrelated chemicals.

Clinical Implications of Resistance. Bacterial resistance problems have resulted in the continuing need for new antimicrobics or modification in the ways in which we treat patients with the available antimicrobials. Important examples with the synthetic antimicrobials include the sulfonamides, which were only active against a small percentage of gonococcal strains 6 years after they were first successfully employed for the treatment of gonorrhea. The sulfonamides also have lost their usefulness in the prevention and treatment of meningococcal infection because of drug resistance. Drug-resistant mutants have arisen in tuberculosis, and naturally resistant species of mycobacteria have become clin-

ically important especially in immunocompromised hosts. The emergence of drug-resistant bacteria in the hospital setting has led to restricted use of certain valuable drugs in hospitals. Other strategies that have been employed to minimize the drug resistance problem include treatment regimes to maintain high drug levels in tissues and the simultaneous administration of two drugs, each of which delays emergence of resistance to the other drug (e.g., sulfonamides and trimethoprim).

2.4. Pathogenesis and Virulence Factors

All organ systems are subject to the pathogenic effects of bacteria, and the resulting infections range from the trivial to the fatal. The major organ systems involved are skin and soft tissues, urinary and reproductive tissues, the respiratory tract, and the central nervous, digestive, and cardiovascular systems. Infections accompanying other medical problems that result in a breach of the anatomical and immunological barriers to infection are particularly difficult to treat, e.g., *Pseudomonas* infection in burn patients or those with cystic fibrosis. Mixed infections involving more than one organism are common.

Although humans are continually exposed to many different environmental microorganisms, only a small percentage of these have the capacity to produce disease. The production of *virulence factors* by microorganisms is the important determinant in disease. Virulence factors are those components of the microbe that are essential for the establishment of infection and the development of disease in the host. Usually more than one factor is involved in the disease process. The adherence of microorganisms to host tissues is determined by highly specific host receptors for bacterial surface components and is a necessary prerequisite for infection. Some bacteria, e.g., *Vibrio cholerae* and *Escherichia coli*, colonize mucosal surfaces and cause damage by elaboration of a toxin. Other organisms, e.g., *Salmonella* and *Shigella,* can invade following attachment and either enter host cells or disseminate throughout the body. The invasion is facilitated by the production of enzymatic substances that circumvent anatomical barriers. To survive, these organisms may have specialized virulence factors that enable them to avoid or disarm host defenses. Survival and continued proliferation of the invading organism are often accompanied by the production of *toxins.* These toxins are proteinaceous substances capable of producing adverse biological effects.

2.5. Selection of an Appropriate Antimicrobial Agent

The selection of a specific agent for treating infection involves consideration of the infecting organism, the status of the host, and the specific attributes of the antimicrobic. The key factors to be considered are the drug's antimicrobial spectrum and potency, its physical characteristics (solubility and stability in body fluids), safety profile, pharmacokinetics, compatibility with other drugs, and cost. The choice of an appropriate agent involves identifying the infecting organism, obtaining information as to its susceptibility, and assessing various host factors.

Identification of the Microorganism. Rapid identification can often be made by *microscopic examination* of gram-stained specimens. A number of immunologic procedures can be employed, such as *latex agglutination tests* or the *enzyme-linked immunoabsorbent assay* (ELISA). Much progress has been made with assays employing *DNA probes* but these are not generally available to the clinical laboratory. *Cultural techniques* are the major means for identification of microbial pathogens. A number of automated and semiautomated tests and kits facilitate this identification and may combine the capability of antimicrobial susceptibility testing.

Determination of Antimicrobial Susceptibility. Several methods for determination of antimicrobial susceptibility are available. The *disk-diffusion method* is widely employed and gives semiquantitative information that is clinically useful. Methods employing *serial dilution* of the drug in culture broth or agar give quantitative data. This is usually expressed as the minimal inhibitory concentration (MIC). A subculture onto an antibiotic-free medium of broth cultures from the MIC tube showing no visible growth can allow for the determination of the minimal bactericidal concentration (MBC), which is usually

defined as the concentration that causes decline in colony count of 99.9 % or more.

Host Factors. The clinician must consider a number of host factors in selecting the drug. A history of adverse reactions to antimicrobials, the patient's age, genetic or metabolic abnormalities, pregnancy, renal or hepatic function, and the site of infection all influence this choice.

Combination Therapy. At times combination therapy offers advantages over treatment with a single antimicrobial agent; however, most infections can be treated with one drug. Combination therapy may be valuable in prevention of the emergence of resistance or for polymicrobial infections. When the initial diagnosis is unclear, a combination of antimicrobial agents may be needed.

2.6. Chemoprophylaxis

Chemoprophylaxis is the administration of drugs prior to or shortly after exposure to an infectious agent to prevent infection. For example, isoniazid may be given to individuals who show recent change from negative to positive skin tests for tuberculosis. The risk of infection must be balanced against toxicity, cost, efficacy, and inconvenience of taking the proposed chemoprophylaxis.

Antibacterial chemoprophylaxis is an accepted clinical procedure for preventing group A *Streptococcus* and *Meningococcus,* and plague infections. Prophylactic drugs have also been used to prevent infective endocarditis, postcoital cystitis, and exacerbations in chronic bronchitis in high-risk individuals. Immunosuppressed patients are often given antimicrobials to prevent the complications resulting from the spread of endogenous bacteria. In some surgical procedures, prophylactic administration of antimicrobials is valuable in preventing infection.

2.7. Chemotherapy

2.7.1. Biochemical Targets for Antimicrobial Agents

A chemical affects cell growth if it hinders the cell's endogenous biosynthetic processes by restricting the availability of building materials or catalytic enzymes or interferes with the supply of usable energy. Considerable advances have been made in identifying several general biochemical targets (i.e., peptidoglycan synthesis, biosynthetic enzymes, and ribosomal function), although in many cases the final molecular site of action is uncertain. At the cellular and subcellular level, most known antimicrobial agents function in one of four major ways:

1) inhibition of bacterial cell wall synthesis
2) alteration of cell membrane permeability or inhibition of active transport across cell membranes
3) inhibition of nucleic acid synthesis
4) inhibition of protein synthesis

The specific mechanisms of action for the individual classes of antimicrobials are discussed in the following sections [2].

2.7.2. Quinolone Antibacterial Agents

Historical Background. Following the discovery of nalidixic acid, numerous antibacterial quinolones were synthesized [3]. These early quinolone carboxylic acids were absorbed fairly well, but suffered from poor pharmacokinetics and tissue penetration, which relegated them to use only in urinary tract infections caused by gram-negative bacteria. In addition to poor pharmacokinetics, these compounds also showed CNS side effects and rapid development of resistant microorganisms. Since the initial introduction of the early compounds, many second- and third-generation compounds have been discovered with increased potency and antibacterial spectrum against gram-negative bacteria, including *P. aeruginosa.* In addition, some of the newer compounds are active against gram-positive bacteria and anaerobes and appear to have fewer adverse effects, as well as improved pharmacokinetics and tissue penetration. Several have shown promise when administered parenterally and thus offer therapeutic potential outside the urinary tract. The early compounds have been reviewed comprehensively, especially with respect to synthetic methods, microbiology, and structure activity relationships [4]. The newer compounds have been reviewed [5–7].

2.7.2.1. Structure Function

Nalidixic acid, the prototype of the series, is a naphthyridine derivative; however, the newer, more potent congeners are not derived from the naphthyridine skeleton. The prototype compound of the quinoline class is oxolinic acid. Although all congeners of nalidixic acid can be named as derivatives of quinoline, such a system is cumbersome because the common skeleton is termed 4-oxo-1,4-dihydroquinoline. A simpler term, 4-quinolone (**1**), has been suggested as a generic name for the agents.

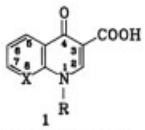

4-Quinolones

X = N, Naphthyridine

X = CH, Quinolone

The second- and third-generation 4-quinolones have been modified at C-6, C-7, and C-8 (**1**). The position 1 is vital to antimicrobial activity. The presence of a two-carbon fragment or a spatial equivalent is essential and almost all marketed compounds in this class have an ethyl group at position 1. Ciprofloxacin has a cyclopropyl group at this site, which fills the spatial requirements of the two-carbon chain. Very few modifications at C-2 have been studied, although cinoxacin, which substitutes a nitrogen for the carbon at this position, has improved absorption properties. Relative to its parent compound, oxolinic acid, cinoxacin is somewhat less active in vitro.

Positions C-3 and C-4 are the most structurally critical, and substitution of other groups at these positions results in loss of activity. Modification of C-5 also offers few advantages. Addition of a nitrogen within the ring may result in an improved use profile. Cinoxacin, mentioned previously, is an example of improved pharmacokinetics resulting from such a change at C-2; however, substitution of a ring nitrogen for C-5, C-6, C-7, and C-8 generally decreases or abolishes activity.

Addition of a fluorine atom to C-6 results in a dramatic increase in the antibacterial spectrum and potency of these compounds [8]. Consequently, all of the newer quinolones have C-6-substituted fluorine as a component of the molecule. These structures have been given the name fluoroquinolones [5].

An aromatic group at C-7 may tend to increase CNS side effects (e.g., rosoxacin), but the piperazine ring may not produce this enhanced effect. The piperazine ring at C-7 also seems to achieve the necessary balance of enhancing penetration capabilities and in vivo activity. The activity of the 4-quinolones against gram-positive and anaerobic bacteria is enhanced through substitutions at C-8 with short one-to-three-atom chains (e.g., ofloxacin). Increased activity against gram-positive bacteria is achieved by the addition of a fluorine atom at C-8 (e.g., CI – 934).

2.7.2.2. Mechanism of Action

The quinolone antibacterial agents act by inhibiting bacterial DNA-gyrase (topoisomerase II). This essential bacterial enzyme, discovered in 1976 in *Escherichia coli* [9], contains two A-subunits and two B-subunits. The *B-subunit,* which is the site of action of coumermycin – novobiocin type antibiotics, is responsible for supplying the energy necessary for the supercoiling of DNA. The *A-subunit,* the site of action for the quinolone type antibiotics, performs the physical act of supercoiling the bacterial DNA molecule in conjunction with the B-subunit [10].

2.7.2.3. Nalidixic Acid and First-Generation Quinolones

Nalidixic acid and the related first-generation 4-quinolones are orally administered agents that concentrate primarily in the urinary tract. Because effective plasma concentrations are not obtained with safe doses of these agents, they cannot be used for the treatment of systemic infections; thus, their use is for the most part limited to infections of the urinary tract.

Nalidixic acid (International Nonproprietary Name – INN, *United States Pharmacopeial – U.S.P.)* [*389-08-2*], 1-ethyl-1,4-dihydro-7-methyl-4-oxo-1,8-naphthyridine-3-carboxylic acid, $C_{12}H_{12}N_2O_3$, M_r 232.24, *mp* 229 – 230 °C, is a pale buff crystalline powder.

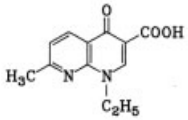

Nalidixic acid is more active against gram-negative bacteria than against gram-positive bacteria. It is active against most of the members of the Enterobacteriaceae, including ca. 99 % of the strains of *Escherichia coli,* 98 % of *Proteus mirabilis,* 75 – 97 % of other *Proteus* species, 92 % of *Klebsiella* and *Enterobacter* species, and 80 % of other coliform bacteria.

These organisms are susceptible to a urinary concentration of 16 µg/mL or less [3]. Some strains of *Salmonella* and *Shigella* are also sensitive. *Pseudomonas* species are resistant, as are most of the important species of gram-positive clinical pathogens, including *Staphylococcus* species, *Streptococcus pneumoniae,* and *Streptococcus faecalis.* Resistance can be induced by in vitro passage [11]; however, surveys of resistance in clinically isolated uropathogens have shown that the incidence of isolates resistant to nalidixic acid remained surprisingly low despite extensive clinical use [12].

Nalidixic acid administered orally is 96 % absorbed from the gastrointestinal tract [13]. It is rapidly metabolized in the liver. Plasma levels of 20 – 50 µg/mL may be obtained 2 h after a 1-g dose [14]; however, the drug does not accumulate in the tissues even after prolonged administration, and the kidney is the only organ in which tissue concentrations may exceed plasma levels [13]. Approximately 85 % of the drug excreted in the urine is the inactive conjugated form; the remainder is the hydroxynalidixic acid metabolite, which is 16 times more active than the parent compound [15]. Urinary concentrations range from 25 – 250 µg/mL following a single oral dose of 0.5 – 1 g and remain between 100 and 500 µg/mL with a 1-g dose administered every 6 h [14], [16].

Oral nalidixic acid is generally well tolerated; however, various *adverse reactions* have been reported. Gastrointestinal side effects include nausea, vomiting, diarrhea, and abdominal pain. Dermatological reactions and photosensitivity have also been reported. Additionally, a range of reversible central nervous system reactions are observed. The drug is contraindicated in infants and during early pregnancy; it should not be used in children.

Synthesis: Condensation of 2-amino-6-methylpyridine with diethyl ethoxymethylenemalonate and thermal cyclization gives 4-hydroxy-7-methyl-1,8-naphthyridine-3-carboxylate. Alkaline saponification of the ester group and alkylation of the nitrogen atom in one step yields nalidixic acid [17].

Trade names: Cybis (Sterling-Winthrop), NegGram, and Wintomylon (Sterling-Winthrop).

Oxolinic acid (INN) [*14698-29-4*], 5-ethyl-5,8-dihydro-8-oxo-1,3-dioxolo[4.5-*g*]quinoline-7-carboxylic acid, $C_{13}H_{11}NO_5$, M_r 261.23, *mp* 314 – 316 °C (decomp.), occurs as crystals.

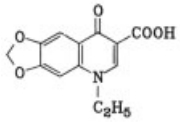

Oxolinic acid shares a similar antibacterial spectrum of activity with nalidixic acid, i.e., it is active against gram-negative rods with the exception of *Pseudomonas* species. Its in vitro potency, however, is significantly greater and it is more active in vivo [18]. Except for *S. aureus* strains that are inhibited at concentrations of 6.25 µg/mL, oxolinic acid does not inhibit gram-positive bacteria. Cross resistance with nalidixic acid is observed [19]. As with nalidixic acid, emergence of resistance during treatment of patients with bacteriuria has been reported.

Oxolinic acid achieves good urinary concentrations within 4 h following oral administration. An oral dosage of 2 g/d produces an average urinary concentration range of 16 – 64 µg/mL in 24-h urinary collections. Elimination of the drug is via urine and feces with low or borderline plasma levels against susceptible bacteria.

Adverse Effects: Central nervous system toxicity has been observed with this drug, and the potential is increased in elderly patients, in which CNS side effects are more common than with nalidixic acid.

Synthesis: Condensation of 3,4-methylenedioxyaniline with diethyl ethoxymethylenemalonate in boiling diphenyl ether and

alkylation of the resultant 1,4-dihydroxy-6,7-methylenedioxy-4-oxoquinoline-3-carboxylate with ethyl iodide in the presence of caustic soda – DMF (*N,N*-dimethylformamide) gives the precursor ester. Saponification with dilute sodium hydroxide and neutralization yields oxolinic acid [20].

Trade name: Utibid (Parke-Davis).

Cinoxacin (INN, U.S.P.) [*28657-80-9*], 1-ethyl-1,4-dihydro-4-oxo[1,3]dioxolo[4,5-*g*]cinnoline-3-carboxylic acid, $C_{12}H_{10}N_2O_5$, M_r 262.22, *mp* 261 – 262 °C (decomp.), occurs as tan crystals.

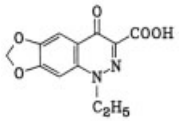

The antibacterial activity spectrum of cinoxacin resembles that of nalidixic acid, although it is more potent against selected species [21]. Generally its potency lies between that of nalidixic acid and oxolinic acid. Cinoxacin is most active against *E. coli*, of which > 90 % of the strains are inhibited at 16 µg/mL. Its activity is mainly against gram-negative rods; a broad spectrum of these organisms are susceptible to cinoxacin [21], [22]. Strains that are resistant to cinoxacin are cross-resistant with nalidixic acid and oxolinic acid [22], [23]. Similar to observations with the other quinolones, no evidence exists that plasmids play a role in resistance.

Excellent absorption is observed on oral administration. Peak plasma levels occur at 2 – 3 h after administration of a 250- or 500-mg dose. Serum binding is ca. 70 % and the serum half-life is ca. 1 h. Food may delay and reduce absorption. Peak urine concentrations range from 88 to 925 µg/mL within 4 – 6 h after dosing. Cinoxacin concentrates in renal tissue, where levels may exceed serum levels [24]. Approximately 60 % of the parent compound is excreted in the urine along with at least four microbiologically inactive metabolites [25].

Low frequencies of *adverse effects* (4.4 %) are reported on oral administration of cinoxacin with gastrointestinal reactions being the most common.Central nervous system side effects are < 1 %.

Synthesis: 2-Nitro-4,5-methylenedioxyacetophenone is catalytically reduced to the corresponding amine, which is cyclized by treatment with $NaNO_2$ and HCl in water to give 6,7-methylenedioxycinnolin-4-ol. Bromination followed by treatment with cuprous cyanide in refluxing DMF gives 4-hydroxy-6,7-methylenedioxycinnolin-3-carbonitrile; hydrolysis with HCl in refluxing acetic acid gives cinoxacin [26].

Trade name: Cinubac (Eli Lilly).

Pipemidic Acid and Piromidic Acid. Both pipemidic and piromidic acid have been marketed outside of the United States as urinary tract antiseptics. These compounds previously have been reviewed in detail [27], [28]. Generally they offer no significant advantages over the other available 4-quinolones and are cross-resistant with these quinolones. They are of historical interest because they are the *forerunners of the second-generation fluoroquinolones.* The addition of a piperazine ring at C-7 of the 4-quinolone skeleton (**1**) in pipemidic acid seems to enhance in vivo performance through improved pharmacokinetics. Pipemidic acid has been reported to be orally effective against experimental *P. aeruginosa* infections in mice; however, its clinical utility against this organism is not established.

Pipemidic acid (INN, dénomination commune Francaise, French approved name – DCF, *Merck Index* – MI) [*51940-44-4*], 8-ethyl-5,8-dihydro-5-oxo-2-(1-piperazinyl)pyrido[2,3-*d*]pyrimidine-6-carboxylic acid, $C_{14}H_{17}N_5O_3$, M_r 303.32, *mp* 253 – 255 °C, is a yellowish-white, odorless, bitter-tasting crystal.

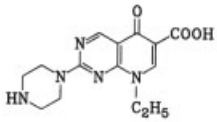

Synthesis: Ethyl 8-ethyl-5,8-dihydro-2-methylthio-5-oxopyrido[2,3-*d*]pyrimidine-6-carboxylate is heated with excess piperazine at 90 – 110 °C in DMSO (dimethyl sulfoxide) for 3 h [29].

Piromidic acid (INN, National Formulary Name – NFN, MI 9) [*19562-30-2*], 8-ethyl-5,8-dihydro-5-oxo-2-(1-pyrrolidinyl)pyrido[2,3-*d*]py-rimidine-6-carboxylic acid, $C_{14}H_{16}N_4O_3$,

M_r 288.31, *mp* 314 – 316 °C, is crystalline in form.

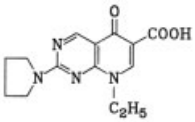

Synthesis: Ethyl 8-ethyl-5,8-dihydro-2-methylthio-5-oxopyrido[2,3-*d*]pyrimidine-6-carboxylate is heated in a pressure tube with pyrrolidine at 95 °C in ethanol for 6 h to yield piromidic acid [30].

Rosoxacin (INN) [*40034-42-2*], 1-ethyl-1,4-dihydro-4-oxo-7-(4-pyridyl)-3-quinoline-carboxylic acid, $C_{17}H_{14}N_2O_3$, M_r 294.31, *mp* 290 °C, forms yellow crystals, is stable in dry heat at 70 °C, and is sensitive to light.

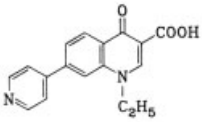

Rosoxacin is a potent first-generation quinolone with minimum inhibitory concentrations (MIC) ranging from 0.1 µg/mL for the most susceptible *E. coli* to 1.6 µg/mL for *Providencia* and *Klebsiella* species [31]. Some strains of *Pseudomonas* are susceptible. *Neisseria gonorrhoeae, Neisseria meningitidis,* and *Haemophilus influenzae* are particularly susceptible to the action of rosoxacin (MIC, 0.02 µg/mL); low plasma levels are achieved in humans after oral administration. A dose of 250 mg produces a peak plasma level of 6.4 µg/mL 2 h after administration.

This quinolone has been marketed as an antigonorrheal agent. Single oral doses are reported to produce high cure rates in patients with acute uncomplicated gonorrhea. *Adverse effects* include dizziness and itching.

Synthesis: A five-step synthesis starting with 3-nitrobenzaldehyde yields 4-(3-aminophenyl)-pyridine, which is condensed with diethyl ethoxymethylenemalonate and then cyclized thermally to yield ethyl 1,4-dihydro-4-oxo-7-(4-pyridyl)quinoline-3-carboxylate. Ethylation followed by saponification and neutralization yields rosoxacin [32].

Trade name: Roxadyl (Sterling-Winthrop).

Miloxacin (INN) [*37065-29-5*], 5,8-dihydro-5-methoxy-8-oxo-1,3-dioxolo[4,5-*g*]quinoline-7-carboxylic acid, $C_{12}H_9NO_6$, M_r 263.21, *mp* 264 °C (decomp.), occurs as colorless prisms from DMF.

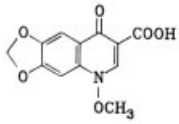

Miloxacin, marketed in Japan, has an antimicrobial spectrum and potency similar to that of oxolinic acid. Clinically, it does not seem to have any advantages over oxolinic acid; however, it appears to offer somewhat superior absorption and higher urinary excretion over oxolinic acid [33], [34]. A peak serum concentration of 7.7 µg/mL is achieved in humans 1 – 2 h following a 500-mg oral dose [35]. The drug is highly metabolized (87 %) to a biologically inactive glucuronide form, and only 5.4 % is found in the urine 8 h after dosing.

Synthesis: [36].

2.7.2.4. Second-Generation Fluoroquinolones

The second-generation fluoroquinolones are distinguished by their excellent and broad antibacterial activity, their relatively fewer adverse effects when compared with nalidixic acid, and their low propensity for inducing bacterial resistance. Some of these compounds can be administered parenterally and are effective in systemic infections outside of the urinary tract.

Their *activity* seems scarcely affected by inoculum size, type of medium, or presence of serum [37–40]. Some of the compounds lose activity in the presence of urine and when the pH of the medium is < 5.0 [38–41]. Generally these drugs are highly active against the enteric gram-negative bacilli and cocci. They are also active against other gram-negative bacteria, including *P. aeruginosa* (but less so against other *Pseudomonas* species), *Aeromonas hydrophila, H. influenzae,* and *Legionella pneumophilia.* They have excellent activity against pathogens of the gastrointestinal tract, including *E. coli*, *Salmonella* species, *Shigella* species, *Yersinia*

enterocolitica, Camplobacter jejuni, and *Vibro* species. Their activity against gram-positive species is poorer than against gram-negative ones, but it is still within a potential therapeutic range especially for ofloxacin and ciprofloxacin. Activity against anaerobic bacteria is marginal. Cross resistance among the various quinolones is observed [41], [42]. The fluoroquinolones are active in vitro against multiple antibiotic-resistant bacteria. They are rapidly bactericidal. Other organisms reported to be susceptible to one or more of these compounds include *Chlamydia trachomatis, Ureaplasma urealyticum, Mycoplasma hominis, Mycoplasma pneumoniae,* and *Mycobacterium tuberculosis.* Fungi are not susceptible to the fluoroquinolones.

Although differences in potency between the fluoroquinolones exist for the various pathogens, their clinical usefulness will be predicted best in conjunction with key parameters, such as pharmacokinetic properties and toxicities of individual drugs [5].

Norfloxacin (INN, BAN – British Approved Name) [*70458-96-7*], 1-ethyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid, $C_{16}H_{18}FN_3O_3$, M_r 319.34, *mp* 227 –228 °C, is crystalline in form.

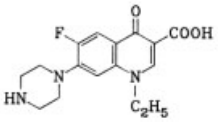

Norfloxacin has an in vitro spectrum of antibacterial activity that includes most gram-negative organisms with MIC_{90} values in the range of 1 µg/mL or less [43], [44]. Norfloxacin is about 100 times more active than nalidixic acid, and its spectrum of activity includes enterococci and staphylococci as well as *Pseudomonas* [43]. Norfloxacin is also active against *H. influenzae* at a concentration of 0.12 µg/mL and *N. gonorrhoeae* at 0.016 µg/mL. The MIC_{90} for *S. aureus* strains is 1.6 µg/mL, with the streptococci being even less sensitive [44]. Members of the *Bacteroides fragilis* group of anaerobes are relatively resistant to norfloxacin (MIC 8 – 128 µg/mL), as are other anaerobic bacteria. Norfloxacin is more potent against sexually transmitted diseases than available quinolone chemotherapeutics [45]. It is less active than other fluoroquinolones against *Ureaplasma* and shows only moderate activity against *Chlamydia trachomatis* [46].

Absorption of norfloxacin is poorer than that of enoxacin or ciprofloxacin [47–50]. Prostatic tissue levels in humans exceed serum levels, whereas blister fluid levels are 70 % of serum levels [48], [51]. Approximately 30 % of the administered dose is excreted in urine; 80 % of the excreted dose is parent compound [52]. Oral administration of 400 mg BID (twice daily) for 10 d was as effective as trimethoprim – sulfamethoxazole in treatment of patients with urinary tract infections or acute pyelonephritis [53]. *Adverse effects,* which included dizziness and nausea, were low. Norfloxacin administered in a single oral dose of 600 mg gave 100 % cure of gonococcal infections in men [45].

Synthesis: by condensation of 1-ethyl-6-fluoro-7-chloro-1,4-dihydro-4-oxo-3-quinolinecarboxylic acid with piperazine at 170 °C in water in a pressure tube [54].

Trade (or code) *names:* Noroxin, Baccidal, MK – 0366, and AM – 715 (Kyorin, Merck).

Ciprofloxacin (INN, BAN) [*85721-33-1*], 1-cyclopropyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-3-quinolinecarboxylic acid, $C_{17}H_{18}FN_3O_5$, M_r 331.35, *mp* 255 – 257 °C, is crystalline in form.

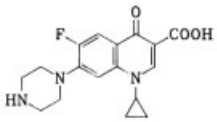

Ciprofloxacin has the most potent in vitro activity of the fluoroquinolones reported to date. It is active against most bacterial strains that cause urinary tract infections at concentrations well in excess of those observed in the urine [55–57]. The minimum inhibitory concentrations of ciprofloxacin that inhibit 90 % of clinical isolates (MIC_{90}), in µg/mL, are as follows.

Pathogen	MIC_{90}, µg/mL
Escherichia coli	0.06
Klebsiella species	0.25
Salmonella species	0.015
Shigella species	0.008
Citrobacter diversus	0.03
Citrobacter freundii	0.125
Enterobacter cloacae	0.03
Enterobacter aerogenes	0.06
Proteus mirabilis	0.06
Proteus vulgaris	0.06
Morganella morganii	0.016
Providencia stuartii	0.50
Pseudomonas aeruginosa	0.25
Pseudomonas maltophilia	4.0
Pseudomonas cepacia	8.0
Serratia marcescens	0.13
Acinetobacter calcoaceticus	0.5
Yersinia enterocoliticia	0.06
Seromonas hydrophilia	0.008
Pasteurella multocida	0.016
Haemophilus influenzae	0.015
Neisseria gonorrhoeae	0.004
Bacteroides fragilis	4.00
Staphylococcus aureus	0.50
Staphylococcus epidermidis	0.25
Streptococcus pyogenes	0.20
Streptococcus agalactiae	1.00
Streptococcus faecalis	2.00
Streptococcus pneumoniae	2.00
Streptococcus viridans	4.00
Listeria monocyogenes	1.00
Chlamydia trachomatis	1.00

As can be seen from these values, most organisms are highly susceptible to ciprofloxacin, although the gram-positive bacteria are much less susceptible than the gram-negative organisms.

Oral doses of ciprofloxacin are rapidly absorbed and peak plasma levels of 2 – 3 µg/mL are observed 1 – 1.5 h after a 500-mg dose. The serum half-life is 3.9 – 4.9 h [50], [58]. Approximately 20 % of the administered dose can be recovered in the urine as active drug during the first 4 h, with a total of 30 – 40 % recovered in 24 h. Ciprofloxacin can also be recovered from blister fluid, where it achieves 57 % of the serum concentration [50]. A parenteral formulation is available. An intravenous dose results in a serum half-life of 4 h, with ca. 76 % of the dose recovered in the urine [49]. Clinical trials have shown the drug to be efficacious for the treatment of urinary tract infections as well as systemic bacterial infections.

Synthesis: Condensation of 2,4-dichloro-5-fluorobenzoylchloride with diethyl malonate by means of magnesium ethoxide in ether gives diethyl 2,4-dichloro-5-fluorobenzoylmalonate, which is partially hydrolyzed and decarboxylated with *p*-toluenesulfonic acid – water, yielding ethyl 2,4-dichloro-5-fluorobenzoylacetate. Condensation of this with triethyl orthoformate in refluxing acetic anhydride affords ethyl 2-(2,4-dichloro-5-fluorobenzoyl)-3-ethoxyacrylate, which is treated with cyclopropylamine in ethanol to give ethyl 2-(2,4-dichloro-5-fluorobenzoyl)-3-cyclopropylaminoacrylate. Cyclization with NaH in refluxing dioxane then yields 7-chloro-1-cyclopropyl-6-fluoro-1,4-dihydro-4-oxoquinoline-3-carboxylic acid, which is finally condensed with piperazine in hot DMSO to yield ciprofloxacin [59].

Trade name: Bay 09867 (Bayer).

Enoxacin (INN) [*74011-58-8*], 1-ethyl-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-1,8-naphthyridine-3-carboxylic acid, $C_{15}H_{17}FN_4O_3$, M_r 320.32, *mp* 220 – 224 °C, occurs as crystals; HCl salt *mp* 300 °C.

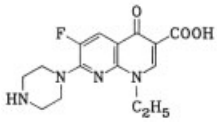

Enoxacin has a spectrum of activity similar to that of norfloxacin, but it appears to be somewhat less potent [41]. The antibiotic is well-absorbed in humans and animals [47], [49], [60]. In experimental animals, enoxacin produced greater tissue concentrations than did norfloxacin when administered orally. In humans, a 600-mg oral dose of enoxacin shows superior oral absorption to norfloxacin, a higher peak serum concentration c_{max} (3.7 µg/mL), a longer serum half-life (6.2 h), greater urinary recovery (67 % in 0 – 24 h), and greater blister fluid penetration [49]. This quinolone is effective in treating urinary tract and respiratory infections in humans, with a low incidence of side effects [49]. Clinical success has also been reported in the treatment of uncomplicated gonorrhea and systemic *P. aeruginosa* infections [61], [62]. A parenteral formulation is available and is currently being tested clinically.

Synthesis: Condensation of 2,3-difluoro-6-nitrophenol with chloroacetone yields 2-acetonyloxy-3,4-difluorobenzene, which is cyclized (H_2/Raney Ni) to give 7,8-difluoro-2,3-

dihydro-3-methyl-4*H*-benzoxazine. Condensation with diethyl ethoxymethylenemalonate followed by cyclization (ethyl polyphosphate, 145 °C) yields 9,10-difluoro-3-methyl-7-oxo-2,3-dihydro-7*H*-pyrido[1,2,3-*d*]1,4-benzoxazine-6-carboxylate. Hydrolysis with HCl gives enoxacin [63].

Trade (or code) *names:* CI – 919, AT – 2266, Flumark, and PD – 107779 (Dainippon).

Pefloxacin (INN) [*70458-92-3*], 1-ethyl-6-fluoro-7-(4-methyl-1-piperazinyl)-4-oxo-1,4-di-hydroquinoline-3-carboxylic acid, $C_{17}H_{20}FN_3O_3$, M_r 333.36, *mp* 270 – 272 °C. The mesylate salt, *mp* 284 – 286 °C (decomp.), is a yellowish-white crystalline powder.

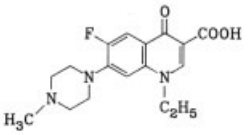

Pefloxacin is marketed in France. The in vitro spectrum and potency of this fluoroquinolone are similar to norfloxacin but pefloxacin appears to have improved activity against gram-positive bacteria [64]. In experimental immunocompromised animals with *Pseudomonas* infection, pefloxacin showed better activity than did ciprofloxacin. In humans, renal plasma levels were 3.8 µg/mL following a 400-mg oral dose, and the plasma half-life (9 – 10 h) exceeded that for most quinolones currently under development [65]. Approximately 10 % is recovered in the urine after an 800-mg oral dose. One of the active metabolites of this compound is norfloxacin. Pefloxacin has been reported to be efficacious in treating meningitis in humans [66].

An intravenous preparation is available, and significant bone levels are achieved after i.v. dosing. Pefloxacin has been reported active clinically in the treatment of staphylococcal osteomyelitis when administered along with rifampicin [67].

Synthesis: 3-Chloro-4-fluoroaniline is condensed with diethyl ethoxymethylenemalonate and then cyclized thermally to give ethyl 6-fluoro-7-chloro-4-hydroxyquinoline-3-carboxylate. Subsequent alkylation with ethyl iodide leads to the ester, which is then saponified and neutralized to give pefloxacin [68].

Trade (or code) *names:* 1589 – RB, Eu – 5306 (Rhône – Poulenc).

Ofloxacin (INN) [*83380-47-6*], (±)9-fluoro-3-methyl-10-(4-methyl-1-piperazinyl)-7-oxo-2,3-dihydro-7*H*-pyrido[1,2,3-*de*]-1,4-benzoxazine-6-carboxylic acid, $C_{18}H_{20}FN_3O_4$, M_r 361.16, *mp* 250 – 257 °C (decomp.), occurs as colorless needles.

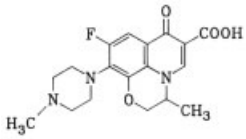

Ofloxacin has no spectrum or potency advantages over the other newer quinolones. Its in vitro microbiological activity is comparable to that of norfloxacin, but it has somewhat better activity against gram-positive cocci [37], [64]. The major attribute of this fluoroquinolone lies in its superior pharmacokinetics. A peak serum level of 10.7 µg/mL at 1.2 h after a 600-mg oral dose was achieved in human volunteers [69]. A long serum half-life (7.0 h) and excellent urinary recovery (73 %) make this compound one of the more interesting members of the group of fluoroquinolones under worldwide development. Ofloxacin has been reported to be efficacious in a variety of clinical indications after oral dosing of 100 – 600 mg/d [60].

Synthesis: prepared from 2,3,4-trifluoronitrobenzene by a seven-step synthesis [70].

Trade (or code) *names:* DL – 8280, Hoe – 280, Ru – 43 – 280, and Tarivid (Daiichi Seiyaku).

Amifloxacin (INN) [*86393-37-5*], 6-fluoro-1,4-dihydro-1-(methylamino)-7-(4-methyl-1-piperazinyl)-4-oxo-3-quinolinecarboxylic acid, $C_{16}H_{19}FN_4O_3$, M_r 334.35, *mp* 299 – 301 °C (decomp.), occurs as crystals.

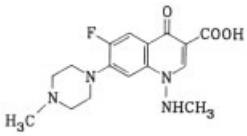

Amifloxacin has broad spectrum in vitro activity similar to that of the other quinolones, but with poorer potency than ciprofloxacin and lit-

tle activity against anaerobic bacteria [71], [72]. The MIC_{90} for various Enterobacteriaceae, including *E. coli, Klebsiella, Proteus, Enterobacter, Citrobacter, Salmonella,* and *Shigella,* was comparable to that of norfloxacin and enoxacin, but was 2- to 8-fold less than that of ciprofloxacin [73]. Amifloxacin was more active than norfloxacin and enoxacin against *P. aeruginosa,* but less active than ciprofloxacin.

Amifloxacin is well-absorbed on oral administration and is excreted at high urinary levels in experimental animals [74]. A parenteral dosage form is available for humans, and early phase 1 studies have been initiated.

Synthesis: prepared from ethyl 6-fluoro-7-chloro-1,4-dihydro-4-oxo-3-quinolonecarboxylate through a five-step synthesis [75].

Trade (or code) *name:* WIN 49375 (Sterling-Winthrop).

Other Fluoroquinolones. A number of interesting fluoroquinolones are currently under study in various countries. They include flumequine (R – 802), A – 56620, AM – 833, and CI – 934. The most advanced of these is flumequine, which is available for human and animal use in Europe.

Compound A – 56620 (difloxacin), like the other fluoroquinolones, has potent antibacterial activity, but also has good activity against staphylococci, streptococci, and *B. fragilis* [76], [77]. Its potency is not appreciably affected by inoculum size. In dogs, the half-life is extended (8.2 h), and the compound undergoes enterohepatic circulation. The compound is in phase I development.

Fluoroquinolone AM – 833 has in vitro microbiological activity similar to that of norfloxacin, but improved activity against staphylococci. It is more active in animal infections when administered orally than is norfloxacin, probably because of superior pharmacokinetics [78]. Peak serum levels and serum half-life in dogs are clearly superior to the values obtained with norfloxacin [79].

Compound CI – 934 has broad gram-negative activity, but is less potent against these organisms than is norfloxacin or ciprofloxacin. It does have superior activity against gram-positive cocci when compared with the other quinolones [80–82].

Flumequine (INN, BAN) [*42835-25-6*], 9-fluoro-6,7-dihydroxy-5-methyl-1-oxo-1*H*,5*H*-benzo[*i,j*]quinolizine-2-carboxylic acid, $C_{14}H_{12}FNO_3$, M_r 261.15, *mp* 253 – 255 °C, is a white crystalline powder.

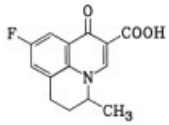

Synthesis: Condensation of 5-fluoro-2-methyltetrahydroquinoline with diethyl ethoxymethylenemalonate followed by thermal cyclization gives ethyl 6,7-dihydro-9-fluoro-5-methyl-1-oxo-1*H,5H*-benzo[*i,j*]quinolizine-2-carboxylate,which is saponified with sodium hydroxide to give flumequine [83].

A – 56620 [*98105-99-8*], 1-(*p*-fluorophenyl)-6-fluoro-1,4-dihydro-4-oxo-7-(1-piperazinyl)-quinoline-3-carboxylic acid hydrochloride, $C_{20}H_{17}F_2N_3O_3 \cdot HCl$, M_r 421.83, *mp* 275 °C.

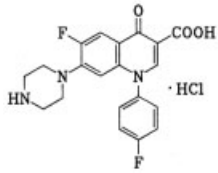

Synthesis: Condensation of 7-chloro-1-(*p*-fluorophenyl)-6-fluoro-1,4-dihydro-4-oxoquinoline-3-carboxylic acid with *N*-carboethoxypiperazine in hot 1-methyl-2-pyrrolidinone yields the *N*-carboethoxy derivative of A – 56620. Hydrolysis of the derivative with sodium hydroxide in aqueous ethanol followed by treatment with dilute hydrochloric acid gives A –56620 [84].

AM – 833 [*79660-72-3*], 6,8-difluoro-1-(2-fluoroethyl)-1,4-dihydro-7-(4-methyl-1-piperazinyl)-4-oxo-3-quinolinecarboxylic acid, $C_{17}H_{18}F_3N_3O_3$, M_r 369.34; HCl salt [*79660-53-0*], *mp* 269 –271 °C.

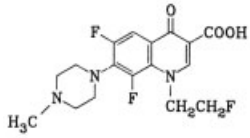

Synthesis: The reaction of 6,7,8-trifluoro-1,4-dihydro-4-oxoquinoline-3-carboxylic acid with 1-bromo-2-fluoroethane by means of NaI in DMF gives 6,7,8-trifluoro-1-(2-fluoroethyl)-1,4-dihydro-4-oxoquinoline-3-carboxylicacid, which is then condensed with *N*-methylpiperazine in refluxing pyridine to give AM – 833.

Manufacturer: Kyorin.

CI – 934 [*91188-00-0*], 1-ethyl-7-{3-(ethylamino)methyl-1-pyrrolidinyl}-6,8-difluoro-1,4-dihydro-4-oxo-3-quinolinecarboxylic acid, $C_{19}H_{23}F_2N_3O_3$, M_r 379.41, *mp* 208 – 210 °C.

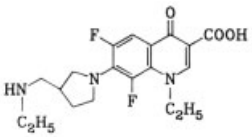

Synthesis: Refluxing 1-ethyl-1,4-dihydro-4-oxo-6,7,8-trifluoro-3-quinolinecarboxylic acid and *N*-ethyl-3-pyrrolidinemethanamine and 1,8-diazabicyclo[5.4.0]undec-7-ene in acetonitrile gives CI – 934 [85].

2.7.3. Sulfa Drugs

The sulfonamides continue to maintain a niche in human and animal medicine even 50 years after their discovery. More than 5000 sulfonamides have been synthesized with a broad diversity of antimicrobial activity and pharmacokinetic profiles. Approximately 13 compounds continue to be used in human medicine; however, their use is limited by the emergence of resistant organisms. The sulfonamide prontosil, one of a group of dyes synthesized in 1935, was the first clinically effective compound. Prontosil protected mice and rabbits against bacterial infections, but had no in vitro activity. Subsequently, prontosil was found to be, in effect, a prodrug, which on metabolic cleavage produced 4-aminobenzenesulfonamide (sulfonalamide) [86]. Because of their ease of synthesis, low cost, relative safety, and broad efficacy in humans, many research organizations embarked on broad synthesis programs designed to produce compounds with improved potency, spectrum of activity, and pharmacokinetic properties. Interest shifted from the sulfonamides in 1945 after the introduction of penicillin, but was revived again in the late 1950s.

2.7.3.1. Biological Activity and Medical Uses

Sulfonamides have broad antimicrobial activity that includes many species of bacteria and protozoa, such as the following:

Gram-negative bacteria
Escherichia coli
Proteus species
Klebsiella species
Salmonella species
Shigella
Vibrio cholerae
Neisseria species
Haemophilus species
Pseudomonas (some strains)
Calymmatobacterium granulomatis
Legionella pneumophila

Gram-positive bacteria	Others
Bacillus anthracis (some strains)	*Chlamydia trachomatis*
Staphylococcus aureus	Trachoma virus
Streptococcus pyogenes	Lymphogranuloma venereum virus
Clostridium welchii	*Plasmodium falciparum*
Clostridium tetani	*Plasmodium malariae*
	Toxoplasma
	Nocardia species

The common gram-positive and gram-negative bacteria are susceptible, as are the protozoan organisms that cause malaria and toxoplasmosis. The lymphogranuloma venereum and trachoma viruses are reported to be susceptible as well. A high percentage of bacterial organisms have developed resistance to sulfonamides; and thus, these drugs are indicated only in a few diseases, including urinary tract infections, chancroid, inclusion conjunctivitis, trachoma, and nocardia infections.

The sulfonamides have been classified on the basis of their serum half-lifes in humans and their topical or gastrointestinal uses. Those compounds with a half-life of < 10 h are termed short-acting, and those with a half-life between 10 and 24 h are medium-acting; long-acting sulfonamides have a half-life > 24 h.

2.7.3.2. Mechanism of Action and Antimicrobial Resistance

4-Aminobenzoic acid (**3**) is an essential metabolite for many microorganisms.

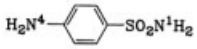

2

4-Aminobenzenesulfonamide
[*63-74-1*]

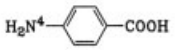

3

4-Aminobenzoic acid
[*150-13-0*]

This compound is used by bacteria as a precursor in the synthesis of folic acid [*59-30-3*], which, in turn, serves as an important intermediate in nucleic acid synthesis. *Sulfonamides,* as structural analogues of **3**, interfere with microbial folic acid synthesis by inhibiting the adenosine triphosphate dependent condensation of a pteridine with **3** to yield dihydropteroic acid, which is subsequently converted to folic acid [87–89]. As a result, nonfunctional analogues of folic acid are formed that do not allow bacterial cells to grow. Because mammalian cells cannot synthesize folic acid, the activity of sulfonamides is selective for bacteria that are capable of producing it. Tubercule bacilli are poorly inhibited by sulfonamides, but their growth is inhibited by 4-aminosalicylic acid [*65-49-6*], whereas most sulfonamide-sensitive bacteria are resistant to this acid. Thus, the receptor site for 4-aminobenzoic acid apparently differs in different types of organisms.

Trimethoprim [*738-70-5*] inhibits dihydrofolic acid reductase 10 000 times more efficiently in bacteria than in mammalian cells. This enzyme is important in the folic acid pathway leading to the synthesis of purines and ultimately of DNA. The combination of sulfonamides with trimethoprim acts synergistically to inhibit bacteria by sequential blockage (see Section 2.7.3.6).

Pyrimethamine [*58-14-0*] (Daraprim) also inhibits dihydrofolate reductase, but it is more active against the enzyme in mammalian cells, and, thus more toxic than trimethoprim. Toxoplasma and other protozoal infections can be successfully treated with pyrimethamine – sulfonamide combinations.

The usefulness of sulfa drugs has been limited by the development of *bacterial resistance.* At least three mechanisms are well documented: (1) alteration of cell wall permeability, (2) increased production of an essential metabolite, and (3) a mechanism involving dihydropteroic acid synthetase. Two enzymes have been reported for *E. coli* with less affinity for sulfonamides than enzymes from susceptible strains, but with no change in 4-aminobenzoic acid affinity. Resistant bacteria, e.g., staphylococci, synthesize excess 4-aminobenzoic acid, which can antagonize sulfa drugs. Plasmids may code for drug-resistant enzymes [90] or decreased bacterial cell permeability to sulfonamides [91]. Multiple resistance mechanisms in the same organism are possible [92]. Sulfonamide resistance is found in 20 – 40 % of community and nosocomial strains of bacteria, including *Staphylococcus,* Enterobacteriaceae, *Neisseria meningitidis,* and *Pseudomonas* species [93].

2.7.3.3. Structure – Function Relationships

All clinically useful sulfa drugs are derived from sulfonalamide (**2**), an analogue of 4-aminobenzoic acid (**3**). The synthesis of thousands of derivatives of **2** has allowed for excellent structure – function determinations, which have been reviewed in detail [94]. A free amino group at N-4 is essential for microbiological activity. Acyl substitution, such as phthalyl or succinyl, at N-4 is acceptable, provided the substituent is hydrolyzed in the body to the free amine (e.g., phthalylsulfathiazole). Replacement of the benzene ring by a heterocycle leads to loss in activity; thus, only the acid portion of the molecule is amenable to substitution. Substitution at the sulfonyl radical can result in marked changes in pharmacologic properties, such as absorption, solubility, pharmacokinetics, and gastrointestinal tolerance.

Monosubstitution at N-1, especially when the substituent is a heterocycle, has led to most of the clinically useful sulfa drugs. Derivatives containing five- or six-membered heterocyclic rings, including oxazole, thiazole, pyridine, pyrimidine, and many others, have been most successful.

2.7.3.4. Pharmacokinetics

Sulfa drugs are commonly administered orally; absorption is generally rapid. Systemic deriva-

tives, such as sulfisoxazole and sulfadiazine, are available, but are rarely used in clinical medicine. An intravenous formulation of trimethoprim and sulfamethoxazole is used to treat *Pneumocystis carinii* pneumonitis, shigellosis, and urinary tract infections caused by susceptible organisms. In addition various topical preparations are available for treating ophthalmic and vaginal infections and burn patients.

Generally the available sulfa drugs are rapidly absorbed from the gastrointestinal tract. Their excretion rates vary widely, depending on their physical and chemical properties. Several derivatives of **2** substituted at N-4 are poorly absorbed and are employed for diseases of the gastrointestinal tract. Peak blood levels following oral absorption are usually observed at 2 – 4 h. A wide variation in half-lifes from 2.5 to 150 h for the commercially available agents has been reported, but most agents are excreted over a 24-h period.

Sulfonamides are usually well-distributed in the body and achieve tissue levels of ca. 80 % of those observed in the serum [95]. The degrees of protein binding and lipid solubility play a major role in determining serum half-life and distribution in the body. The primary route of metabolism is via the liver, where acetylation or glucuronidation occur. Excretion is mainly via the kidney, where most drugs are removed by glomerular filtration, although partial reabsorption and active tubular secretion also are involved. Sulfa drugs bind reversibly to serum proteins, primarily albumin, which tends to slow metabolism by the liver [96].

2.7.3.5. Toxicity and Drug Interactions

Significant hypersensitivity reactions have been associated with the sulfonamides; the long-acting derivatives, especially, have caused fatal hypersensitivity reactions. Other hypersensitivity reactions include erythema nodosum and erythema multiforme, including Stevens – Johnson syndrome. Photosensitization, itching, and rash usually require cessation of therapy. With earlier drugs, crystalluria and related renal damage occurred, primarily with the less water-soluble compounds, such as sulfadiazine, sulfamerizine, and sulfapyridine. Other adverse effects observed less commonly are nausea, vomiting, diarrhea, hepatic toxicity, and a syndrome that resembles serum sickness. Serious adverse reactions include hemolytic anemia, aplastic anemia, agranulocytosis, thrombocytopenia, and leukopenia.

Sulfonamides interact with other drugs and compete with drugs, such as warfarin and methotrexate, for albumin-binding sites resulting in increased toxicity of the displaced drugs. Concomitant administration of probenecid, which results in decreased renal tubular secretion, prolongs sulfa drug levels and may increase toxicity.

2.7.3.6. Combination Therapy

Both oral and intravenous formulations of sulfamethoxazole in combination with trimethoprim are available. This synergistic mixture is used orally to treat urinary tract and certain other infections. In the United States, trimethoprim is also available as a single agent for the treatment of urinary tract infections. Combinations of sulfa drugs and various antibiotics for both oral and topical use are also available.

Systemic formulations in combination with 2,6-diamino-3-phenylazopyridine · hydrochloride, an azo dye that is a mild local analgesic, have been employed to reduce pain of inflammation that results from urinary tract infections.

2.7.3.7. Rapidly Absorbed Short- and Medium-Acting Sulfa Drugs

Sulfamethoxazole (INN, U.S.P.) [*723-46-6*], N^1-(5-methyl-3-isoxazolyl)sulfanilamide, $C_{10}H_{11}N_3O_3S$, M_r 253.28, *mp* 170 °C, is an odorless and colorless crystalline powder.

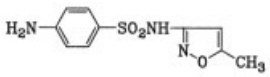

Sulfamethoxazole is rapidly absorbed from the gastrointestinal tract and has a moderate elimination rate. Peak blood levels of 80 – 100 µg/mL are observed 1 – 4 h following a 2-g oral dose. The plasma half-life is 9 – 11 h and the drug is 70 % protein-bound; 80 – 90 % of the unbound compound remains unacylated and biologically active in the blood. About 60 % of the drug is excreted in the urine, and half of this

is N-4 acylated or conjugated. Urine levels are 3 times blood levels.

The major indication for sulfamethoxazole is treatment of urinary tract infections caused by susceptible strains of *E. coli, Klebsiella, Proteus* species, and staphylococci. Although the compound has broader microbiological activity, it generally is not used for infections outside of the urinary tract. Sulfamethoxazole is also indicated in the treatment of nocardiosis, inclusion conjunctivitis, trachoma, chancroid, and malaria caused by chloroquine-resistant strains, and is used in combination with pyrimethamine for toxoplasmosis.

The usual dose for mild infections is 2 g followed by 1 g every 12 h for 4 days. For severe infections, the drug is administered 3 times a day at twice this dose. Pediatric dose is 50 – 60 mg/kg followed by 12-h doses of 25 – 30 mg/kg.

Synthesis: prepared from ethyl 5-methylisoxazole-3-carbamate [97].

Trimethoprim (INN, U.S.P.) [*738-70-5*], 2,4-diamino-5-(3,4,5-trimethoxybenzyl)pyrimidine, $C_{14}H_{18}N_4O_3$, M_r 290.32, *mp* 199 – 203 °C, is a white-to-cream-colored bitter, crystalline powder.

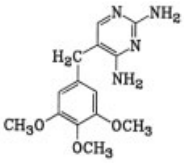

Synthesis: prepared from guanidine and β-ethoxy-3,4,5-trimethoxybenzylbenzalnitrile [98].

Trade names: Cofrim (Lemmon), Bactrim (Hoffman-La Roche), Septa (Burroughs Wellcome), Proloprim (Burroughs Wellcome), and Trimpex (Hoffman-La Roche, trimethoprim only).

Sulfamethoxazole – Trimethoprim. Sulfamethoxazole is usually administered along with trimethoprim in a fixed ratio of 5 : 1. This synergistic combination results in a sequential blockage of folic acid synthesis. Synergistic antibacterial activity has been demonstrated in both in vitro and in vivo systems [99], [100].

The pharmacokinetic profiles of sulfamethoxazole and trimethoprim fit well, both compounds showing peak blood levels 1 – 4 h after oral administration. With administration at 12-h intervals, a 20 : 1 ratio of unbound sulfa drug: trimethoprim is achieved in the plasma, an optimal ratio for antibacterial activity. Trimethoprim is 44 % protein bound and has a serum half-life of 8 – 10 h. Both drugs are metabolized by the liver and excreted by the kidneys.

Trimethoprim has a broad spectrum of antimicrobial activity and extends the spectrum of sulfamethoxazole to include many urinary and systemic pathogens untreatable by this sulfa drug alone. The major use of the drug combination is for urinary tract infections, in which case the drugs are administered orally for 10 – 14 days every 12 h. Additionally, the combination is indicated for chronic bronchitis in adults and otitis media in children for susceptible strains of *Streptococcus pneumoniae* and *Haemophilus influenzae.* Bacillary dysentery caused by *Shigella* species also responds to oral treatment.

An intravenous preparation is available for treatment of severe infections, as well as pneumonitis caused by the protozoan, *Pneumocystis carinii.* For treatment of *Pneumocystis* pneumonia, the recommended i.v. dose is 15 – 20 mg/kg, based on the trimethoprim component, given in 3 or 4 equally divided doses every 6 – 8 h for up to 14 days [101]. For severe urinary tract infections and shigellosis, the total daily dose is 8 – 10 mg/kg, based on the trimethoprim component, given in 2 – 4 equally divided doses every 6, 8, or 12 h for up to 14 days for urinary tract infections and 5 days for shigellosis.

Sulfisoxazole [*127-69-5*], sulfafurazole (U.S.P., INN), N^1-(3,4-dimethyl-5-isoxazolyl)sulfanilamide, $C_{11}H_{13}N_3O_3S$, M_r 267.30, *mp* 192 °C, occurs as colorless and odorless prisms.

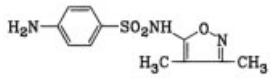

Sulfisoxazole has antimicrobial activity somewhat less than that of sulfadiazine; however, it is more easily tolerated because of its high solubility and rapid excretion. It is clinically useful in urinary tract infections and is especially potent against *E. coli* and *Prot. vulgaris.* Later derivatives of this compound, i.e.,

sulfamethoxazole, have superior solubility properties. At 2 h after oral administration of a 2-g dose, peak serum levels of ca. 150 µg/mL are achieved. The drug is 85 % protein-bound, and about 30 % of the unbound portion is in the acetylated form.

Excretion is primarily by the kidneys with ca. 95 % of a single oral dose eliminated within 24 h; 70 % of the excreted material is in the active form. The acetyl sulfisoxazole is metabolized in the gastrointestinal tract to release sulfisoxazole, which results in delayed absorption.

Although this sulfa drug is still used clinically, its effectiveness is limited mainly because of high protein binding. The usual adult dose is 2 – 4 g in a loading dose followed by 4 – 8 g/d divided in 4 – 6 doses. The compound is also formulated in vaginal preparations for the treatment of vaginitis caused by *Haemophilus vaginalis.* This sulfa drug is also available as the water soluble diolamine salt in 40 % aqueous solution for dilution with sterile water as a parenteral preparation.

Synthesis: p-Acetaminobenzenesulfonyl chloride is reacted with 3,4-dimethyl-5-aminoisoxazole followed by deacetylation [102].

Trade names: Gantrisin (Hoffmann-La Roche), SK-Soxazole (Smith Kline & French).

Sulfadiazine (INN, U.S.P.) [*68-35-9*], N^1-2-pyrimidinylsulfanilamide, $C_{10}H_{10}N_4O_2S$, M_r 250.27, *mp* 252 – 256 °C, is a white or pale yellow crystalline powder.

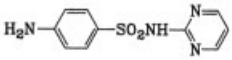

Sulfadiazine is classified as a rapidly absorbed agent with a moderate rate of excretion. It is the sulfa drug of choice for treating CNS infections caused by antibiotic-resistant organisms and for treating nocardia infections. Sulfadiazine is also used to treat urinary tract infections; however, crystalluria is a problem because of its poor water solubility. The usual precautions of urine alkalinization and forced fluid intake are recommended for patients taking sulfadiazine.

Sulfadiazine is absorbed rapidly from the gastrointestinal tract after oral administration. The peak blood level of free drug following a 2-g dose is 30 – 60 µg/mL with about 45 % protein-binding. The drug achieves excellent cerebrospinal fluid levels, which are 5 – 13 % of the blood levels. The plasma half-life is 17 h. About 20 % of an oral dose is excreted in the urine as the free form. The major metabolite is the N-4-acetylated compound.

This sulfa drug can be administered parenterally; however, because its alkaline solution is irritating, the i.v. route is preferred. The usual oral dose for adults is 2 – 4 g initially, followed by 2 – 4 g/d in 3 – 6 divided doses. When given intravenously for adults and children over 2 months of age, the drug is administered at a loading dose of 50 mg/kg or 1.25 g/m^2 followed by a total daily dose of 100 mg/kg or 2.25 g/m^2 administered 4 times a day.

A silver salt of sulfadiazine is available for topical treatment of burn patients. In this case the sulfonamide acts primarily as a vehicle for release of silver ions, which exert an antibacterial effect. Resistant organisms have been reported in burn units.

Synthesis: prepared by condensing 2-aminopyrimidine with acetylsulfaninyl chloride followed by hydrolysis with NaOH [103].

Trade name: Suladyne (Reid-Provident).

Sulfacytine (INN) [*17784-12-2*], N^1-(1-ethyl-1,2-dihydro-2-oxo-4-pyrimidinyl)sulfanilamide, $C_{12}H_{14}N_4O_3S$, M_r 294.33, *mp* 166.5 – 168 °C, is crystalline in form.

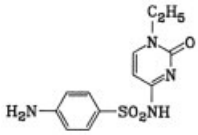

Sulfacytine is absorbed rapidly from the GI (gastrointestinal) tract with a high rate of urinary excretion. Its major use is for treating urinary tract infections. It is active against the common urinary tract pathogens, such as *E. coli, Klebsiella* species, *Proteus* species, as well as *S. aureus.* Urinary levels far in excess of the MIC are found following a daily dose of 1 g. The usual dose is 500 mg initially followed by 250 mg 4 times a day for 10 days. Peak blood levels occur within 2 – 3 h after administration, with ca. 90 % of the oral dose being recovered in the urine as the free form. High urinary concentrations are

observed, but the drug is bound 86 % to serum proteins. The serum half-life is 4.5 h.

Synthesis: [104].

Trade Name: Renoquid (Parke-Davis).

Sulfamethizole (INN, U.S.P.) [*144-82-1*], N^1-(5-methyl-1,3,4-thiadiazol-2-yl)sulfanilamide, $C_9H_{10}N_4O_2S_2$, M_r 270.32, *mp* 208 °C, is a colorless crystal.

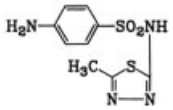

Sulfamethizole is well-absorbed and rapidly excreted. Its major indications are confined to therapy of urinary tract infections caused by susceptible organisms. It has the same therapeutic indications as sulfisoxazole; however, it is less bound to serum proteins than is sulfisoxazole. The usual adult dose is 2 – 4 g followed by 2 – 4 g/d administered in 6 divided doses.

Synthesis: prepared by the reaction of acetaldehyde thiosemicarbazone with *p*-acetylaminobenzenesulfonyl chloride in pyridine [105].

Trade names: Thiosnifil-A (Ayerst), Proklar (O'Neal, Jones & Feldman), component of Azotrex (Bristol), and component of Suladyne (Reid-Provident).

2.7.3.8. Long-Acting Sulfonamides

Sulfadoxine (INN, U.S.P.) [*2447-57-6*], *N'*-(5,6-dimethoxy-4-pyrimidinyl)sulfanilamide, $C_{12}H_{14}N_4O_4S$, M_r 310.33, *mp* 201 – 202 °C, forms colorless crystals.

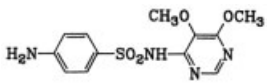

Sulfadoxine is the only long-acting agent available for clinical use in the United States. All of the long-acting sulfonamides are associated with serious hypersensitivity reactions, such as Stevens – Johnson syndrome. Sulfadoxine, originally known as sulfamethoxine, is combined with pyrimethamine (Fansidar). It has a half-life of 200 – 230 h and a peak serum level of 51 – 76 µg/mL at 2.5 – 6 h following an oral dose of 0.5 g. The major indication is for the treatment and prophylaxis of chloroquine-resistant *Plasmodium falciparum* malaria.

Synthesis: [106].

Trade name: Fanasil (Hoffmann-La Roche International).

2.7.3.9. Sulfonamides for Use in the Gastrointestinal Tract

Sulfasalazine (INN, U.S.P.) [*599-79-1*], 5-{[*p*-(2-pyridylsulfamoyl)phenyl]azo}salicylic acid, $C_{18}H_{14}N_4O_5S$, M_r 398.39, *mp* 240 – 245 °C, is a brownish-yellow powder.

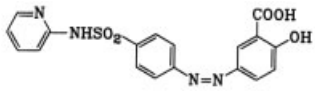

Sulfasalazine, used for the treatment of ulcerative colitis, is a prodrug of sulfapyridine. The compound is a combination of sulfapyridine and 5-aminosalicylic acid joined by an azo link, which is cleaved in the lower intestinal tract by bacterial enzymes. The release of 5-aminosalicylic acid is believed to produce a local antiinflammatory action in the colon; this is the mechanism by which the drug is effective in ulcerative colitis.

Sulfasalazine is absorbed only 10 – 15 % from the upper gastrointestinal tract and gives peak blood levels of 14 µg/mL 2 – 4 h after a 2-g oral dose. The drug is excreted in the bile and reabsorbed from the intestine as its metabolite, sulfapyridine. Peak blood levels of sulfapyridine have been reported at 13 µg/mL 6 – 24 h after dosing. The major urinary metabolite of sulfasalazine is sulfapyridine and its metabolites (60 %); however, small amounts of the unchanged drug (15 %) and 5-aminosalicylic acid and its metabolites (20 – 33 %) are also observed.

Synthesis: prepared by coupling diazotized 2-sulfanilamidopyridine with salicylic acid [107].

Sulfaguanidine, Sulfasuxidine, Sulfathalidine. Sulfaguanidine, sulfasuxidine, and sulfathalidine are all absorbed poorly from the GI tract following an oral dose. They have been employed as prophylactic agents prior to bowel surgery.

Sulfaguanidine (INN, *National Formulary* – NF **11,** MI) [*57-67-0*], N^1-(diaminomethylene)-sulfanilamide, $C_7H_{10}N_4O_2S$, M_r 214.24, *mp* 190 – 193 °C, occurs as colorless crystals.

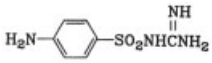

Synthesis: prepared by fusing N^4-acetylsulfanilamide and dicyanodiamide [108].

Sulfasuxidine, succinylsulfathiazole (U.S.P. **18,** MI) [*116-43-8*], 4′-(2-thiazolylsulfamoyl)-succinanilic acid monohydrate, $C_{13}H_{13}N_3O_5S_2 \cdot H_2O$, M_r 373.41, *mp* reported as 184 – 186 °C and as 192 – 195 °C, occurs as crystals.

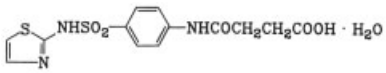

Synthesis: prepared by refluxing sulfathiazole with a slight excess of succinic anhydride in alcohol [109].

Sulfathalidine [*85-73-4*], phthalylsulfathiazole (INN, U.S.P.), 4′-(2-thiazolylsulfamoyl)-phthalanilic acid, $C_{17}H_{13}N_3O_5S_2$, M_r 403.43, *mp* 270 °C, is a white crystalline powder.

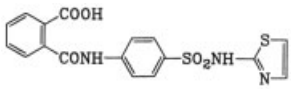

Synthesis: prepared by refluxing sulfathiazole with a slight excess of succinic anhydride in alcohol [110].

2.7.4. Agents for Treating Mycobacterial Infections

The mycobacteria are a group of rod-shaped bacteria that include many pathogenic species [111–113]. They do not stain readily, but once stained, they resist decoloration by acid or alcohol; therefore, they are called acid-fast bacilli. They cause chronic diseases producing granulomatous lesions. The major organisms causing human disease are *Mycobacterium tuberculosis, M. leprae,* and the so-called atypical mycobacteria.

Mycobacterium tuberculosis and *M. bovis,* the causative agents of tuberculosis, produce no known toxins. The organisms are usually taken in via the respiratory tract, where they establish pulmonary infections. Resistance and hypersensitivity of the host generally influence the development of the disease. Once established in the tissues, they reside in phagocytic cells, where their intracellular location makes chemotherapy difficult. The organisms spread in the host by direct extension, through the lymphatics and bloodstream, and via bronchi and the gastrointestinal tract. Every organ system can be involved. Chronic pulmonary infections are usually established; however, meningitis or urinary tract involvement can occur in the absence of other symptoms of tuberculosis. Dissemination via the bloodstream leads to miliary tuberculosis involving many organs and a high fatality rate.

Treatment of *M. tuberculosis* infections usually involves combinations of chemotherapeutic agents. The most widely used antituberculosis drugs are isoniazid, pyrazinamide, ethambutol, and the antibiotics rifampin and streptomycin (→ Antibiotics). Other drugs, such as ethionamide, 4-aminosalicylic acid, viomycin, and cycloserine, have severe adverse effects and are employed less frequently. Treatment is for long periods of time and clinical cure can be achieved in 6 – 18 months. Drug-resistant strains to all of these agents emerge rapidly; thus, greater success is achieved when the drugs are administered concomitantly.

There has been a increase in the incidence of atypical mycobacterial infection usually observed in patients whose immunity has been compromised by other factors. Many of these species of *Mycobacterium* cause severe disease closely resembling tuberculosis. These atypical mycobacteria, e.g., *M. avium – intracellularis* group, are generally much less sensitive to the available antituberculosis chemotherapeutics, especially streptomycin and isoniazid. Other organisms respond to the antitubercular chemotherapeutic agents, whereas many are sensitive to antibiotics employed for treating other bacterial infections (aminoglycosides and tetracyclines). A summary of the chemotherapeutic agents used to treat atypical mycobacterial infections is shown in Table 2 [114].

Mycobacterium leprae, the etiological agent of *leprosy*, is a disease involving the cooler tissues of the body, i.e., skin, superficial nerves, nose, pharynx, larynx, eyes, and testicles. In un-

Table 2. Antimicrobials used to treat atypical mycobacteria ∗

Mycobacterial species	Primary drug	Secondary drug
M. kansasii	isoniazid with rifampin, with or without ethambutol	ethambutol, streptomycin, pyrazinamide, PAS, cycloserine, ethionamide, kanamycin, and capreomycin
M. fortuitum complex	amikacin and doxycycline	cefoxitin, rifampin, erythromycin, and a sulfonamide
M. avium – intracellularis – scrofulaceum	isoniazid, rifampin, ethambutol, and streptomycin	clofazimine, capreomycin, ethionamide, cycloserine, ansamycin, imipenem, and amikacin
M. marinum (balnei)	minocycline	trimethoprim – sulfamethoxazole, rifampin, and cycloserine

∗ Adapted from [114].

treated cases, severe disfiguration is observed because of skin infiltration and nerve involvement. There are two distinct types of leprosy, lepromatous and tuberculoid. The more severe, lepromatous type is progressive and leads to lesions involving skin and nerves and to bacteremia. The tuberculoid type is benign and nonprogressive. Treatment usually is lengthy and requires administration of one of the sulfones and the semisynthetic antibiotic, rifampin. Drug resistance to the sulfones has emerged.

2.7.4.1. Antituberculosis Agents

Isoniazid (INN, U.S.P.) [*54-85-3*], isonicotinic acid hydrazide, $C_6H_7N_3O$, M_r 137.14, *mp* 171.4 °C, occurs as crystals (alc).

In 1952 isoniazid was demonstrated to be valuable in treating tuberculosis [115]. It kills *M. tuberculosis* organisms by inhibiting mycolic acid synthesis [116]. At higher concentrations, it may have another mechanism of action against atypical bacteria. The MIC of isoniazid for susceptible *M. tuberculosis* strains ranges from 0.025 to 0.05 µg/mL. Emergence of resistance is observed when the drug is administered alone [115].

Isoniazid is well-absorbed after oral or intramuscular administration and is subsequently distributed throughout the body. The drug is acetylated in the liver, with the rate of acetylation being separated into genetic populations of slow and normal acetylators [116].

Isoniazid is generally well-tolerated. An infrequent major toxicity is hepatitis [117]. Elevations in serum glutamic-oxaloacetic transaminase (SGOT), which are observed in 15 % of patients, disappear with continued therapy. Peripheral neuropathy occurs in 17 % of people receiving 6 mg kg^{-1} d^{-1} of isoniazid. Manifestations of central nervous system toxicity have been reported, as have hypersensitivity reactions. Isoniazid potentiates dilantin toxicity, especially in slow acetylators.

Isoniazid is indicated for all forms of tuberculosis. It is administered in combination with one or more other antitubercular agents, the usual dose being 5 – 10 mg kg^{-1} d^{-1}. Higher doses of the drug (15 mg/kg, orally) can be administered twice weekly along with other agents. This twice-weekly regime can be started after an initial period of daily drug administration and serves to encourage compliance and reduce cost [118].

Synthesis: prepared by condensing pyridine-4-carboxyethylate and hydrazine [119].

Trade names: Cotinazin (Pfizer), Dinacrin (Sterling-Winthrop), INH (Ciba-Geigy), and Nydrazid (Squibb).

Ethambutol hydrochloride (INN, U.S.P.) [*1070-11-7*], [*74-55-5*], (+)-2,2′-(ethylenediimino)-di-1-butanol dihydrochloride, $C_{10}H_{24}N_2O_2 \cdot 2\,HCl$, M_r 277.23, *mp* 194.5 °C.

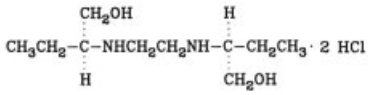

Ethambutol is a tuberculostatic agent active against most human strains of tubercle bacilli at 8 µg/mL, with ca. 75 % of the strains susceptible at 1 µg/mL. It is not cross-resistant with isoniazid-resistant strains. Resistance develops stepwise when the drug is administered alone. The major use of the drug is in combination with other antitubercular agents to limit the emergence of resistance. Ethambutol has replaced 4-aminosalicylic acid as the companion drug for isoniazid because of its greater potency and lower incidence of adverse effects.

The usual dose is 15 – 25 mg kg^{-1} d^{-1} initially, followed after 60 days by 15 mg kg^{-1} d^{-1} as a single dose. Ethambutol is 75 – 80 % absorbed after an oral dose with peak plasma levels of 5 µg/mL after a 25-mg/kg dose. It is well-distributed in the body and reached CNS levels of 10 – 50 % of serum levels in patients with meningeal inflammation. The drug is excreted mainly unchanged in the urine with ca. 15 % being converted to inactive metabolites following absorption. Neuropathic toxicities are the major *adverse effects* of this drug, especially retrobulbar neuritis, which is common at a dose of 50 mg kg^{-1} d^{-1}.

Synthesis: prepared by warming (+)-2-amino-1-butanol with ethylene bromide at 110 – 115 °C, liberation of the base with potassium hydroxide, and conversion to the dihydrochloride [120].

Trade name: Myambutol (Lederle).

Pyrazinamide (INN, U.S.P.) [*98-96-4*], pyrazinecarboxamide, $C_5H_5N_3O$, M_r 123.11, *mp* 189 – 191 °C, occurs as crystals.

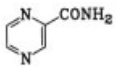

Pyrazinamide is a bactericidal antitubercular agent that is used in combination with other antitubercular agents, especially in developing nations [121]. It is a nicotinamide analogue with an unknown mechanism of action. When used alone, resistance develops rapidly. Because of its toxicity, it is used for short-course therapy regimes. Treatment with pyrazinamide can result in hepatotoxicity, especially in the extended dose regimes employed in the early clinical trials, i.e., 40 mg kg^{-1} d^{-1}, but a recommended dose of 20 – 35 mg kg^{-1} d^{-1} appears to be safe [122].

In the United States, pyrazinamide is generally reserved for combination use in patients infected with drug-resistant strains. Mainly because of low cost and convenient once-weekly administration, its use is favored in developing nations, especially in areas of high incidence of drug resistance [123].

Pyrazinamide is well-absorbed when administered orally and reaches levels in body fluids in excess of the MIC for susceptible organisms. It is metabolized by the liver and excreted in the urine. The usual total daily dose is 1.5 – 2.0 g divided into 2 – 4 dosing intervals. Higher levels administered once or twice weekly appear to be clinically effective with no observable liver toxicity.

Synthesis: prepared by ammonolysis of methyl pyrazinoate (from quinoxaline) [124].

Manufacturer: Lederle.

Ethionamide (INN, U.S.P.) [*536-33-4*], 2-ethylthioisonicotinamide, $C_8H_{10}N_2S$, M_r 166.24, *mp* 164 – 166 °C (decomp.), occurs as yellow crystals.

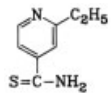

This nicotinic acid analogue is considered a secondary agent for treating drug-resistant tuberculosis. It is tuberculostatic for most susceptible strains at 0.6 – 2.5 µg/mL. The usual dose is 1 g/d, starting at 250 mg in divided doses and increasing by 125 mg kg^{-1} d^{-1} until the desired dose is reached. A frequent *adverse effect* is gastrointestinal irritation accompanied by nausea and vomiting. Central nervous system side effects, including psychiatric disturbances and peripheral neuropathy, have been reported. In about 5 % of the patients, a reversible hepatotoxicity is observed. Ethionamide is absorbed well from the gastrointestinal tract, giving plasma concentrations of 20 µg/mL. It penetrates both normal and inflamed meninges, giving high CNS

levels. Ethionamide metabolism is in the liver, and the metabolites are excreted in the urine.

Synthesis: prepared by addition of hydrogen sulfide to 2-ethylnicotinonitrile in the presence of triethanolamine [125].

Trade name: Trecator – SC (Ives).

***p*-Aminosalicylic acid** [*65-49-6*], PAS (U.S.P.), 4-aminosalicylic acid, $C_7H_7NO_3$, M_r 153.14, *mp* 150 – 151 °C (effervescence), is in the form of minute crystals (alcohol).

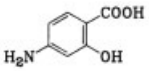

4-Aminosalicylic acid is an inhibitor of mycobacterial growth by impairing of folate synthesis. It has largely been replaced by ethambutol in the standard combination therapy for *M. tuberculosis.* The compound is poorly absorbed when administered orally; thus, a 4-g dose results in peak plasma concentrations of only 7 – 8 µg/mL. The metabolites of PAS are excreted in the urine. The major *adverse effect* is gastrointestinal intolerance, which frequently results in poor patient compliance. 4-Aminosalicylic acid has been reported to produce a lupus-like syndrome. Hypersensitivity reactions are frequent, occurring at a 5 – 10 % rate. The usual adult dose is 10 – 12 g/d in three or four divided doses and 200 – 300 mg kg^{-1} d^{-1} in divided doses for children.

Synthesis: prepared by carboxylation of 3-aminophenol with ammonium carbonate solution at 110 °C under pressure [126].

Trade name: PAS – Heyl (Heyl).

Amithiozone [*104-06-3*], thiacetazone (INN, MI), 4′-formylacetanilide thiosemicarbazone, $C_{10}H_{12}N_4OS$, M_r 236.29, *mp* 225 – 230 °C (decomp.), occurs as minute, pale yellow crystals that darken on exposure to light.

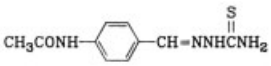

Amithiozone is a second-line drug for the treatment of tuberculosis. This thiosemicarbazone is active against *M. tuberculosis,* inhibiting most strains at 1 µg/mL [127]. The drug is administered orally at 150 mg/d or 450 mg twice weekly. Peak serum concentrations are 1 – 2 µg/mL. Amithiozone should be used in combination with other drugs because resistance develops readily when the drug is used alone. The drug is not available in the United States and has only limited use in Europe because of gastrointestinal irritation and bone-marrow suppression. It is, however, used as a first-line drug in East Africa, where apparent lower incidence of adverse effects in Africans and low cost favor its use.

Synthesis: prepared by treating 4-acetamidobenzaldehyde with thiosemicarbazide in alcohol [128].

2.7.4.2. Antileprosy Agents

Dapsone (U.S.P.) [*80-08-0*], 4,4′-sulfonyldianiline, $C_{12}H_{12}N_2O_2S$, M_r 248.30, *mp* 175 – 176 °C (also recorded 180.5 °C), occurs as crystals (alc).

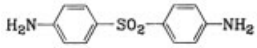

Dapsone is the basic therapeutic agent for the treatment of leprosy. It is used either alone or more frequently as a component of multidrug programs. Dapsone has been employed, primarily as a single chemotherapeutic agent, since the mid-1940s with satisfactory clinical efficacy. Both secondary and primary drug resistance have emerged and are of worldwide concern; thus, combination therapy with the antibiotic rifampin or with clofazimine is recommended. Many dapsone derivatives, all of which share many pharmacological properties, have been synthesized; however, dapsone remains the most clinically useful agent for the treatment of leprosy.

Sulfones share a mechanism of action similar to that of the sulfonamides and are bactericidal for *M. leprae* [129]. Dapsone is slowly and completely absorbed from the gastrointestinal tract with peak levels achieved in 1 – 3 h. The half-life varies over a broad range with a mean of 28 h. The drug is 50 % bound to plasma proteins [130]. Tissue distribution is broad and reabsorption of bile-secreted drug from the gastrointestinal tract tends to extend the residence time of the drug in the circulation. The drug is excreted in the urine.

Daily therapy with 50 mg of dapsone has been successful in adults. The daily dose can

be increased to 100 mg if necessary, and twice-weekly doses of 100 – 400 mg have been clinically successful. Therapy is gradually increased to the effective level and should continue for at least 2 years, but may be necessary for the lifetime of the patient. The most common *adverse effect* is hemolysis, generally observed in patients dosed with levels greater than 100 mg/d. Anorexia, nausea, and vomiting may result from oral administration.

Synthesis: [131].

Trade name: Avlosulfon (Ayerst).

Sulfoxone, sodium (INN, U.S.P.) [*144-75-2*], [*144-76-3*] free acid, disodium sulfonylbis(*p*-phenyleneimino)dimethanesulfinate, $C_{14}H_{14}N_2Na_2O_6S_3$, M_r 448.43.

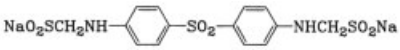

Sulfoxone has been substituted for dapsone in cases in which poor gastrointestinal tolerance restricts dapsone therapy. The maximum daily oral dose of sulfoxone is 660 mg. Sulfoxone is incompletely absorbed from the gastrointestinal tract and large amounts are excreted in the feces. Generally the distribution and excretion of sulfoxone in humans after oral absorption is similar to that observed for dapsone.

Synthesis: prepared by combining 4,4′-sulfonyldianiline with sodium formaldehyde sulfoxylate in acetic acid or alcohol [132].

Trade name: Diasone Sodium Enterab (Abbott).

Acedapsone (INN) [*77-46-3*], 4,4′-sulfonylbis(acetanilide), $C_{16}H_{16}N_2O_4S$, M_r 332.37, *mp* 289 – 292 °C, is a crystalline solid.

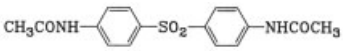

Acedapsone is a long-acting injectable repository derivative of dapsone. It is slowly absorbed when administered intramuscularly. Peak serum concentrations occur 3 – 5 weeks after administration. Acedapsone is metabolized to dapsone in the body, where it has a serum half-life exceeding 40 d [133]. Injections administered 5 times yearly have been employed with promising results [134].

Synthesis: [135].

Clofazimine (INN) [*2030-63-9*], 3-(*p*-chloroanilino)-10-(*p*-chlorophenyl)-2,10-dihydro-2-(isopropylimino)phenazine, $C_{27}H_{22}Cl_2N_4$, M_r 473.40, *mp* 210 – 212 °C, occurs as dark red crystals.

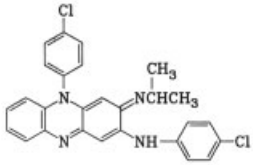

Clofazimine is a phenazine dye with weak bactericidal activity against *M. leprae* [136]. Its main use has been for treating sulfone-resistant infections and people who cannot tolerate sulfones. It also exerts an antiinflammatory effect and prevents the development of erythema nodosum leprosum. The drug is absorbed from the gastrointestinal tract and appears to accumulate in the tissues. Appreciable clinical effects are not observed until 50 days after initiation of therapy. The dose of clofazimine used is 100 – 300 mg spaced at 2-week intervals [137–139]. The major *adverse effects* are skin pigmentation and mild gastrointestinal intolerance; other effects are negligible.

Synthesis: by heating the hydrochloride of the corresponding 2-imino compound with isopropylamine at 80 °C [140].

Trade name: Lampren (Ciba-Geigy).

2.7.5. Miscellaneous Nitroheterocycles Used to Treat Bacterial Infection

Metronidazole (INN, U.S.P.) [*443-48-1*], 2-methyl-5-nitroimidazole-1-ethanol, $C_6H_9N_3O_3$, M_r 171.16, *mp* 158 – 160 °C, occurs as cream-colored crystals.

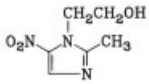

Metronidazole was originally introduced as an antitrichomonal agent and was later found to be useful in treating infections caused by several other protozoal organisms as well as by anaerobic bacteria.

Metronidazole inhibits most anaerobic bacteria at in vitro concentrations of 16 µg/mL or less; however, some organisms are less susceptible [141], [142].

This drug is useful clinically for treating the majority of anaerobic infections with the exception of actinomycosis. A major use of metronidazole is to treat infections caused by the *Bacteroides fragilis* group, the most common etiological agent of serious anaerobic infections. The emergence of resistance to clindamycin, an antibiotic useful in the treatment of anaerobic bacteria, makes metronidazole an even more important agent.

Metronidazole is absorbed rapidly after oral dosing and the serum levels obtained during the elimination phase are similar to those observed when an equivalent dose is administered intravenously. The drug is distributed widely in the body and diffuses well into all tissues including the central nervous system. Five major metabolites are found in the urine, with the hydroxy metabolite being the most important. The usual intravenous treatment regime for the susceptible anaerobic bacteria is a loading dose of 15 mg/kg followed by 7.5 mg/kg 4 times a day. The oral dose is 1 – 2 g/d in 2 – 4 doses at intervals of 6 or 12 h [143]. The maximum daily dosage is 4 g. Generally the drug is well-tolerated; however, major *adverse reactions* involving the central nervous system have been reported. In patients ingesting alcohol, metronidazole may cause reactions similar to those observed with disulfiram.

Synthesis: prepared by nitrating 2-methyl-5-nitroimidazole followed by alkylation with 2-chloroethyl alcohol [144].

Trade names: Flagyl (Searle), Metro I.V. (American McGaw), Satric (Savage), Metronid (Asher), Metryl (Lemmon), and SK-Metronidazol (Smith Kline & French).

Nitrofurantoin (INN, U.S.P.) [*67-20-9*], [*17140-81-7*], monohydrate, 1-[(5-nitrofurfurylidene)amino] hydantoin, $C_8H_6N_4O_5$, M_r 238.16, *mp* 270 – 272 °C (decomp.), occurs as orange-yellow needles (dilute acetic acid).

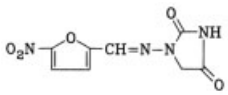

Nitrofurantoin is used to prevent and treat urinary tract infections caused by susceptible bacteria. The drug is bacteriostatic for most susceptible organisms at 32 µg/mL or less. Nitrofurantoin is active against many strains of *E. coli;* however, most *Proteus* and *Pseudomonas* species and many of the *Enterobacter* and *Klebsiella* species are resistant [145]. Other bacteria commonly found outside the urinary tract are susceptible in vitro to nitrofurantoin, but their susceptibility is of little practical significance [146]. Microorganisms initially sensitive to nitrofurantoin generally do not become resistant during therapy.

Nitrofurantoin is absorbed rapidly and completely after oral administration with therapeutically active concentrations found only in the urinary tract. Approximately one-third of the drug is rapidly excreted in the urine by both glomerular filtration and tubular secretion, with significant reabsorption when the urine is acid [147]. The macrocrystalline form of the drug is absorbed and excreted more slowly. The average dose gives urinary concentrations of ca. 200 µg/mL. The usual adult dose for treating acute and recurrent uncomplicated urinary tract infections is 50 or 100 mg 4 times per day for 2 weeks. The most common *adverse effects* are nausea, vomiting, and diarrhea. Various hypersensitivity reactions have been reported. Other rare, but serious side effects involving various organ systems occur.

Synthesis: prepared by reaction of 5-nitrofurfural diacetate with 1-aminohydantoin in acid solution [148].

Trade names: Furadantin (Norwich Eaton), Macrodantin (Norwich Eaton), and the sodium salt is Dantrirem (Norwich Eaton).

3. Chemotherapy of Protozoan Infections

3.1. Classification of Pathogenic Protozoa

The human parasites that are members of the phylum Protozoa are a diverse group of eukaryotic single-cell organisms. They can be divided into four principal groups:

1) flagellates (Mastigophora)

2) sporozoans (Sporozoa),
3) ciliates (Ciliata)
4) amebas (Sarcodina)

These organisms continue to be among the leading causes of infectious diseases in underdeveloped countries and are now encountered more frequently throughout the world. The ease of travel and the increased use of immunosuppressive drugs have increased the incidence of protozoal infections in the temperate climates. Table 3 lists the major parasitic protozoal human diseases and their causative agents. Excellent reviews and textbooks, which detail the epidemiology, pathogenesis, and chemotherapy of protozoal infections, are available [149–163].

Table 3. Major protozoan organisms causing human disease

Disease	Causative organism
	Flagellates
African trypanosomiasis (sleeping sickness)	*Trypanosoma gambiense*
	Trypanosoma rhodesiense
American trypanosomiasis (Chagas' disease)	*Trypanosoma cruzi*
Visceral leishmaniasis	*Leishmania donovani*
Mucocutaneous leishmaniasis	*Leishmania braziliensis*
Cutaneous leishmaniasis	*Leishmania tropica*
	Leishmania mexicana
Trichomoniasis	*Trichomonas vaginalis*
Lambliasis	*Giardia lamblia*
	Sporozoans
Malaria	*Plasmodium falciparum*
	Plasmodium vivax
	Plasmodium malariae
	Plasmodium ovale
Babesiosis	*Babesia* species
Isosporiasis (human coccidiosis)	*Isospora belli*
Toxoplasmosis	*Toxoplasma gondii*
Cryptosporidium	*Cryptosporidium* species
Pneumocystosis	*Pneumocystis carinii*
	Ciliates
Balantidiasis	*Balantidium coli*
	Amebas
Amebiasis	*Entamoeba histolytica*
	Entamoeba polechi
	Dientamoeba fragilis

3.2. Flagellates

3.2.1. Hemoflagellates

The hemoflagellates of humans include the genera *Trypanosoma* and *Leishmania* (Table 3). There are two distinct types of human trypanosomiasis: (1) African (or sleeping sickness), caused by *T. rhodesiense* and *T. gambiense*, and transmitted by tsetse flies; and (2) American, caused by *T. cruzi,* which is the agent of Chagas' disease and is transmitted by cone-nosed bugs. The *Leishmania* include four species and several subspecies, all of which are transmitted by sand flies.

Trypanosomes appear in the blood as trypanomastigotes, an elongated form with a free flagellum that is an extension of a lateral undulating membrane. Other developmental forms among the hemoflagellates include the amastigote, a leishmanial rounded intracellular stage; the promastigote (formerly called a leptomonad), an elongated flagellated extracellular form without an undulating membrane; and an epimastigote (formerly called the crithidial stage), a flagellated form with a short undulating membrane.

In *Leishmania* only the amastigote and promastigote stages are found, the latter being restricted to the insect. In African trypanosomes only the trypanomastigote is found in humans, whereas the other flagellated stages appear in the tsetse fly. In *Trypanosoma cruzi* all three stages appear in humans, whereas only the trypanomastigote and epimastigote stages are found in the insect.

3.2.1.1. African Trypanosomiasis

3.2.1.1.1. Biology and Epidemiology

The two human parasites causing African trypanosomiasis, i.e., *Trypanosoma rhodesiense* and *T. gambiense,* are believed to have evolved from *T. brucei,* a pathogen of livestock and game animals, and are considered by some to be subspecies of this animal trypanosome. The three forms are indistinguishable morphologically but differ ecologically and epidemiologically.

These parasites are introduced through the bite of the tsetse fly (*Glossina*) and the disease is generally restricted to tsetse fly areas. *Trypanosoma gambiense,* transmitted primarily by *Glossina palpalis,* extends from west to central Africa and produces a relatively chronic infection with progressive central nervous system

disease. *Glossina morsitans* transmits *T. rhodesiense* in more restricted geographical areas, primarily to the south and west of Lake Tanganyika (Africa). This disease is generally more acute, leading to death in a matter of weeks to months.

After introduction of the parasite by the tsetse fly bite, a primary lesion is formed; the parasite may spread to lymph nodes and the bloodstream and in terminal stages to the central nervous system, where it produces the syndrome of sleeping sickness. Trypanosomes appear in the blood and tissues as trypanomastigotes. When trypanosomes are ingested with a blood-meal, they undergo a developmental cycle in the tsetse fly, producing metacyclic forms in the salivary glands that are reintroduced with the next blood-meal.

3.2.1.1.2. Chemotherapy

The chemotherapy of African trypanosomiasis centers on three key drugs: (1) pentamidine for chemoprophylaxis, (2) suramin for treatment of the early stages of the disease, and (3) melarsoprol for treatment of the late stages when trypanosomes are present in the central nervous system [164]. A polyamine biosynthesis inhibitor, α-difluoromethylornithine, has been reported to be effective clinically in the treatment of African trypanosomiasis [165]. Detailed reviews have been published on the chemotherapy of the African trypanosomes [166–172].

Drug resistance is not a serious clinical problem despite the fact that producing resistance to the available agents in the laboratory is relatively easy [166–169]. Suramin and pentamidine do not penetrate the central nervous system, a serious defect that limits the clinician to the more toxic arsenic compounds for treating late-stage disease. In addition lack of oral activity, adverse effects, and lack of efficacy against all stages of disease and all strains and species of pathogens limit the available agents.

Suramin (see Section 4.6.9). Suramin is a sulfated naphthylamine that is the drug of choice for the treatment of early stage hemolymphatic African trypanosomiasis. The usual adult dose is a test dose of 100 – 200 mg followed by 5 intravenous doses of 1 g administered on days 1, 3, 7, 14, and 21. Initial plasma concentrations fall rapidly and the drug, tightly bound to serum proteins, remains in low concentration for up to 3 months. Suramin does not penetrate the central nervous system and, thus, does not cure infections of the central nervous system. A transient albuminuria is often observed during treatment, as are shock, febrile reactions, skin lesions, and other toxic effects.

Trade names: Antrypol, Bayer 205, Belganyl, Fourneau 309, Germanin, Moranyl, Naganol, and Naphuride.

Pentamidine (INN, BAN, DCF, NFN) [*100-33-4*], 4,4′-(pentamethylenedioxy)dibenzamidine, $C_{19}H_{24}N_4O_2$, M_r 340.42, *mp* 186 °C (decomp.); pentamidine isethionate (M & B 800) [*140-64-7*], $C_{23}H_{36}N_4O_{10}S_2$, M_r 592.28, *mp* ca. 180 °C, occurs as very bitter crystals.

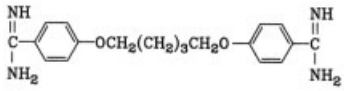

Two salts of this aromatic diamidine are in use: pentamidine isethionate and pentamidine methanesulfonate (lomidine). The doses and adverse effects of the two salts are similar and are discussed together. Treatment usually involves an intramuscular dose of 4 mg kg^{-1} d^{-1} for 10 days. For chemoprophylaxis, the usual dose is 4 mg/kg given intramuscularly every 3 – 6 months. Generally, this drug is more effective against the Gambian disease than the Rhodesian disease. Abdominal cramping may occur during treatment. Pentamidine can cause hypoglycemia and renal damage.

Pentamidine has been shown to be effective in the treatment of pneumonia due to *Pneumocystiscarinii,* an opportunistic protozoal pathogen, and as a secondary drug for treating visceral leishmaniasis.

Synthesis: Saturating an anhydrous alcoholic solution of 4,4′-dicyanodiphenoxypentane with dry hydrogen chloride and allowing it to stand gives pentamidine [173].

Trade names: Lomidine, dimethanesulfonate salt (Specia, France); pentamidine isethionate B.p., diisethionate salt (May & Baker, United Kingdom).

Melarsoprol (INN, BAN, DCF, NFN, MI 9) [*494-79-1*], 2-[4-(4,6-diamino-1,3,5-triazin-2-

yl-amino)phenyl]-1,3,2-dithiarsolane-4-methanol, $C_{12}H_{15}AsN_6OS_2$, M_r 398.33.

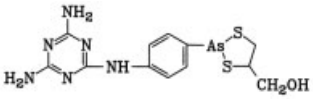

Melarsoprol and the other arsenic compounds discussed in this section are rarely used because of their extreme toxicity. The use of these arsenic compounds had been restricted to patients with overt central nervous system symptoms; however, it is now realized that invasion of the central nervous system occurs early in the course of infection and that late relapses after treatment with drugs other than the arsenic compounds are due to failure to kill organisms present in the central nervous system. Thus, a regime of intravenous suramin followed by one or more courses of intramuscular injections of melarsoprol has been proposed to prevent relapse [155]. The recommended treatment for central nervous system infections is a daily intravenous dose of 2 – 3.6 mg/kg for 3 days followed after a week with 3.6 mg/kg for 3 days and repeated after 10 – 21 days.

Adverse effects with melarsoprol and other arsenic compounds are common and can be quite severe. These include various neuropathies, including the optic nerve, skin rashes, and a syndrome that resembles acute encephalitis and is believed to be due to a reaction that destroys parasites in the brain, which occurs in up to 10 % of the patients. Treatment with arsenic compounds is also complicated by the emergence of drug-resistant strains.

Synthesis: [174].

Tryparsamide (INN; U.S.P. **17,** MI 9) [*554-72-3*], monosodium *N*-(carbamoylmethyl)arsanilate, $C_8H_{10}AsN_2NaO_4 \cdot 1/2\,H_2O$, M_r 305.10, occurs as hemihydrate, platelets, is slowly affected by light and stable to air.

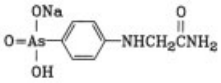

Tryparsamide is a pentavalent arsenic compound that has no in vitro activity, but is most probably metabolized to the trivalent arsenic form after injection. This drug is active only against *T. gambiense,* which limits its use. It crosses the blood – brain barrier, and, thus can be used to treat late-stage trypanosomiasis; however, adverse effects are such that melarsoprol is preferred. Tryparsamide has been employed as an alternate drug of choice in tandem with suramin for treating late-stage disease with central nervous system involvement. The regime employed for tryparsamide is intravenous administration of 30 mg/kg given every 5 days for a total of 12 injections. The regime may be repeated after 1 month. Suramin should be administered intravenously at 10 mg/kg at the same dose regime employed for tryparsamide.

Synthesis: reaction of arsanilic acid with chloracetamide in the presence of sodium hydroxide and sodium carbonate [175].

Trade names: Tryparsam, Tryparsamidium, Glyphenarsine, and Tryparsone.

Berenil [*908-54-3*], diminazene aceturate (INN, BAN, NFN), 4,4′-(1-triazene-1,3-diyl)bis(benzenecarboximidamide)bis(*N*-acetylglycinate), $C_{22}H_{29}N_9O_6$, M_r 515.54, *mp* 217 °C (decomp.), is a yellow solid; free base [*536-71-0*].

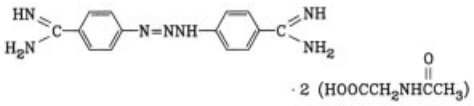

This aromatic diamidine, closely related to pentamidine, was developed originally as a cattle trypanocide. It has been employed effectively in an intravenous dose to treat both Gambian and Rhodesian forms of trypanosomiasis. Recently, the drug was also found to be active on oral administration, with the effective dose being 5 mg/kg taken at 2-day intervals for a total of 3 doses [164]. The drug appears to be relatively well tolerated; however, a persistent albuminuria has been observed in humans following therapy.

Synthesis: [176].

Mel W [*13355-00-5*], melarsonyl potassium (INN, BAN, DCF), melarsenoxide potassium dimercaptosuccinate, dipotassium 2{4-[(4,6-diamino-1,3,5-triazin-2-yl)amino]phenyl}-1,3,2-dithiarsolane-4,5-dicarboxylate, $C_{13}H_{11}AsK_2N_6O_4S_2$, M_r 532.51; free acid [*37526-80-0*].

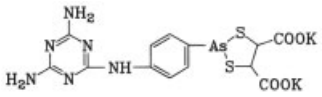

Mel W is an arsenic compound that was developed as a possible replacement for melarsoprol. Similar to other arsenic compound, it crosses the blood – brain barrier and thus can be employed to treat late-stage Gambian and Rhodesian trypanosomiasis. Although it is more water soluble than melarsoprol, Mel W has similar adverse effects to those of melarsoprol. It is almost never employed clinically because of lack of obvious efficacy and lack of safety advantages.

3.2.1.2. American Trypanosomiasis

3.2.1.2.1. Biology and Epidemiology

Trypanosoma cruzi, which causes Chagas' disease, is a parasite of humans and many small animals. This disease is found mainly in Central and South America, but several cases have been reported as far north as the southern United States. This organism is transmitted by triatomid bugs of the family Reduviidae, of which at least 36 species have been found to be naturally infected.

The trypanosomes are ingested with a bloodmeal and undergo development in the intestine of the bug. They eventually give rise to infective metacyclic trypanosomes, which resemble those in vertebrate blood. Transmission usually results from contamination of mucous membranes or skin with infected insect excreta.

After the organisms gain access to the host, they usually multiply in lymphoid tissue. From there, all cells of the body may be infected, but the reticuloendothelial system, the central nervous system, and cardiac muscle are most likely involved. The digestive tract also may be infected, especially the esophagus and colon, leading to development of megaesophagus or megacolon.

The amastigote (leishmania) form is found intracellularly where it multiplies, leading to the formation of pseudocysts, which rupture and release trypanosomes capable of invading other cells or circulating in the blood. Inflammatory reactions resulting from cell damage and infected cells produce many of the disease symptoms.

Acute Chagas' disease occurs mainly in children and only in a small percentage of those people who become infected in endemic areas. Chronic disease develops slowly, with cardiac involvement observed in 20 – 40 % of patients.

3.2.1.2.2. Chemotherapy

A great need exists for agents to treat Chagas' disease. Only two drugs have been reported to have any significant efficacy for treatment (nifurtimox and benznidazole) and none are available for chemoprophylaxis [169], [177–181].

Nifurtimox [*23256-30-6*], 4-[(5-nitrofurfurylidene)amino]-3-methyl-4-thiomorpholine-1,1-dioxide, $C_{10}H_{13}N_3O_5S$, M_r 287.29, *mp* 180 – 182 °C, occurs as orange red crystals.

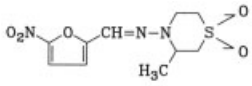

This nitrofuran has been available for 10 years for treating the acute stage of the disease. Its efficacy in chronic Chagas' disease is open to question. Apparently there are variations in response by various strains of the parasite. The usual dose is 8 – 10 mg kg^{-1} d^{-1} in 4 divided oral doses given for 120 days. It is considered by some to be the only available agent for the treatment of Chagas' disease. Nifurtimox is not well tolerated by patients over the 120-day recommended treatment period. *Adverse effects*, such as nausea, weight loss, and memory and sleep disorders, are common, and few patients complete the full treatment period.

Synthesis: prepared from 5-nitrofurfural and 5-amino-3-methyltetrahydro-1,4-thiazine-1,1-dioxide [182].

Trade names: Bayer 2502, Lampit.

Benznidazole [*22994-85-0*], *N*-benzyl-2-nitroimidazole-1-acetamide, $C_{12}H_{12}N_4O_3$, M_r 260.26, *mp* 188.5 – 190 °C, is crystalline in form.

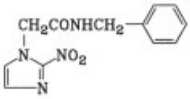

This orally administered nitroimidazole is claimed to be effective for both acute and chronic Chagas' disease. The recommended dose is $5\,mg\,kg^{-1}\,d^{-1}$ for 60 days. No evidence has shown that this drug has any significant advantages over nifurtimox. In doses employed in the initial clinical trials in excess of $5\,mg\,kg^{-1}\,d^{-1}$, serious *adverse effects*, such as polyneuropathy, were observed.

Synthesis: [183].

Trade names: R07-1051, Radanil and Rochagan.

3.2.1.3. Leishmaniasis

3.2.1.3.1. Biology and Epidemiology

Leishmania infections are caused by four species of *Leishmania* (*L. donovani, L. tropica, L. braziliensis,* and *L. mexicana*). These are obligate intracellular parasites during their amastigote (nonflagellar) stage in the human host. One species, *L. donovani,* mainly invades the internal organs whereas the others invade the skin. The Central and South American forms of leishmaniasis, caused by *L. braziliensis* and *L. mexicana* involve the mucous membranes of the nose, mouth, and pharynx, in addition to the skin. *Leishmania* are transmitted by sand flies of the genus *Phlebotomus.* The sand fly acquires the parasite with a blood-meal or from infected skin. In the sand fly, the organism is transformed into the flagellated promastigote stage, which multiplies and is introduced into the skin of a host when the fly feeds.

A broad spectrum of pathology is observed, partly dependent on the site of invasion of the specific parasite and the host's inflammatory response to the organism. The visceral form of the disease, kala-azar, caused by *L. donovani,* produces a progressive disease mainly of the reticuloendothelial system (spleen, liver, and bone marrow). The disease usually results in death if untreated. The cutaneous forms vary from self-healing sores to diffuse progressive disfiguring lesions that can lead to a disseminated disease in hosts with defective cellular immunity.

3.2.1.3.2. Chemotherapy

Two groups of compounds are used to treat visceral and cutaneous forms of leishmaniasis: pentavalent organic antimony compounds and aromatic diamidines. The antifungal antibiotic, amphotericin B, has been shown to be effective in selected situations.

Pentavalent antimony compounds have been used to treat kala-azar since the 1920s. Their introduction was preceded, for a few years, by the use of trivalent antimony compounds, especially tartar emetic. With correct use, cure rates exceeding 90 % have been obtained with pentavalent antimony compounds [184]. Response to therapy for visceral leishmaniasis varies in different geographical locations, with more resistance to therapy being observed in the Sudan than in India.

Pentamidine, a drug used for treating African trypanosomiasis, has been employed for the therapy of visceral leishmaniasis after initial treatment with pentavalent antimony compounds has failed. Although this compound, which now is the only readily available diamidine for use in humans, is effective, the patients may have a high relapse rate. The recommended dose of pentamidine is 2 – 4 mg/kg intramuscular or slow intravenous injections administered weekly or, at most, 3 times weekly, until a clinical and parasitological cure is achieved, usually involving many months of therapy.

Sodium stibogluconate (INN, BAN, DCF) [*16037-91-5*], antimony(V) derivative of sodium gluconate, 2′,4′-*O*-(oxydistibylidyne)bis(D-gluconic acid)-2,4-Sb,Sb′-dioxide trisodium salt · nonahydrate, $C_{12}H_{17}Na_3O_{17}Sb_2 \cdot 9\,H_2O$, M_r 907.6, water-soluble amorphous powder.

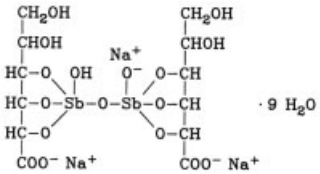

Sodium stibogluconate is the drug of first choice for the treatment of all forms of leishma-

niasis. Some evidence suggests that the antiprotozoal activity of this compound may depend on its reduction to the trivalent antimony compound after treatment. The drug can be administered by either the intravenous or the intramuscular routes.

A large percentage of patients are cured with a single course of therapy consisting of 20 mg kg^{-1} d^{-1} up to a maximum of 800 mg/d. Sodium stibogluconate is tolerated relatively well; however, a disturbing aspect of therapy is the occurrence of sudden death, although it is difficult to determine whether the deaths are related to the clinical disease, the drug therapy, or the interaction of disease and drug.

Synthesis: heating gluconic acid with freshly made antimonic acid paste until it is completely dissolved and then neutralized with sodium hydroxide.

Trade name: Pentostam (Burroughs Wellcome, United Kingdom).

Glucantime [*133-51-7*], meglumine antimonate, 1-deoxy-1-(methylamino)-D-glucitol trioxoantimonate, $C_7H_{18}NO_8Sb$, M_r 365.91.

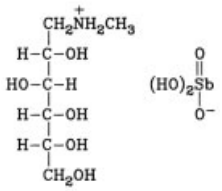

Glucantime is a pentavalent antimony compound related to pentostam. The former drug has no obvious advantage over the latter one and is used interchangeably with pentostam as the drug of choice in some countries. It can be administered intramuscularly or intravenously. The usual dose is 10 or 20 mL of a 30 % solution on alternate days for a total of 200 – 250 mL. As has been observed with pentostam, viable organisms can be recovered from treated patients both during and even after therapy.

Trade names: Protosib, 2168-RP (Rhône-Poulenc).

3.2.2. Intestinal and Urogenital Flagellates

Although the human alimentary and urogenital tracts are colonized by seven species of flagellate protozoa, only two, *Trichomonas vaginalis* and *Giardia lamblia*, are generally considered to be pathogens. Reviews covering broad biological data for these two protozoans are available [185–191].

3.2.2.1. Trichomonas Vaginalis

3.2.2.1.1. Biology and Epidemiology

Of the three species of *Trichomonas* that infect humans, only *T. vaginalis* is pathogenic, causing trichomoniasis. *Trichomonas vaginalis* is a common pathogen in the female genitourinary tract, where it exists only as the trophozoite form. In the female, the infection is normally limited to the vulva, vagina, and cervix. In the male, where infection occurs less often, the prostate, seminal vesicles, and urethra may be involved. Signs and symptoms in the female include profuse vaginal discharge in addition to local tenderness, vulval pruritus, and burning. About 10 % of infected males have a thick, white urethral discharge.

Trichomonas infection is typically transmitted during sexual intercouse, although nonvenereal routes of transmission cannot be excluded. Infection rates vary greatly, but may be quite high in some populations, e.g., 38 – 56 % of symptomatic women attending a venereal disease clinic. Assessing the prevalence of trichomoniasis in men is difficult because most infections in men are asymptomatic. An estimated 3×10^6 American women contract trichomoniasis every year.

3.2.2.1.2. Chemotherapy

Successful treatment of vaginal infections requires the destruction of the trichomonads. Numerous topical preparations are available for treatment; however, they suffer from failure to completely eliminate the parasite in female infections and cannot be applied to infected males. Systemically active drugs are, therefore, the normal treatment of choice. Treatment is necessary both in symptomatic patients and simultaneously in asymptomatic sexual consorts to prevent reinfection or spread of infection.

The effective agents for the systemic treatment of trichomoniasis are all related to the 5-nitroimidazole, metronidazole (see Section 2.7.5). The drugs include tinidazole, nimorazole, ornidazole, secnidazole, and carnidazole. None of these compounds are approved for use in the treatment of trichomoniasis in the United States, and not all are available in every country. No clear clinical advantage has been shown for any of these 5-nitroimidazoles over the others. Cure rates for trichomoniasis are ca. 85 – 95 %, with symptoms usually relieved within a few days. The original recommended regimes for the various compounds for therapy called for oral doses of up to 1 g/d in divided doses for 4 – 10 days; e.g., the recommended dose for metronidazole is 250 mg given 3 times each day for 7 days. Patients now generally receive a shortened regime or a single large oral dose of 2 g. A single dose is as efficacious as multiple-dose treatment and may be superior with some imidazoles. Treatment failures with the 5-imidazoles may be associated with poor serum and local tissue levels or resistant isolates. Cross resistance with metronidazole and the other nitroimidazoles, has been reported. Generally, the 5-nitroimidazoles are free of acute side effects. The most common *adverse effects* are gastrointestinal, and these generally are transient and mild. This group of compounds is mutagenic in bacteria, and tumors have been observed in rodents fed high doses of metronidazole for long periods.

Metronidazole: see Section 2.7.5.

Tinidazole (INN) [*19387-91-8*], 1-[2-(ethylsulfonyl)ethyl]-2-methyl-5-nitroimidazole, $C_8H_{13}N_3O_4S$, M_r 247.27, *mp* 127 – 128 °C, occurs as colorless crystals (benzene).

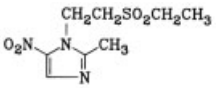

Synthesis: heating 2-methyl-5-nitroimidazole with 2-ethylsulfonylethyl-4-toluenesulfonate for 4 h [192].

Trade names: Fasigyn (Pfizer, USA), Simplotan (Pfizer, FRG).

Nimorazole (INN, BAN, NFN, MI) [*6506-37-2*], 4-[2-(5-nitroimidazolyl-1-yl)-ethyl]morpholine, $C_9H_{14}N_4O_3$, M_r 226.23, *mp* 110 – 111 °C, occurs as crystals (water).

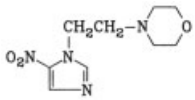

Synthesis: condensing the sodium salt of 5-nitroimidazole with 4-(2-chloroethyl)morpholine in boiling acetone [193].

Trade names: Naxogin (Carlo Erba, Italy).

Ornidazole (INN) [*16773-42-5*], α-(chloromethyl)-2-methyl-5-nitro-1*H*-imidazole-1-ethanol, $C_7H_{10}ClN_3O_3$, M_r 219.63, *mp* 77 – 78 °C, is crystalline in form.

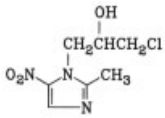

Synthesis: [194].

Secnidazole (INN, DCF, NFN, BAN) [*3366-95-8*], α-2-dimethyl-5-nitro-1*H*-imidazole-1-ethanol, $C_7H_{11}N_3O_3$, M_r 185.18, *mp* 76 °C, occurs as crystals.

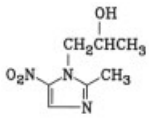

Synthesis: [195].

Code names: PM 185184, RP 14539.

Carnidazole (INN) [*42116-76-7*], *O*-methyl [2-(2-methyl-5-nitro-1*H*-imidazol-1-yl)ethyl]thio-carbamate,

$C_8H_{12}N_4O_3S$, M_r 244.27.

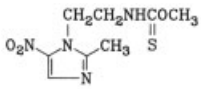

Synthesis: [196].

Code names: R28096 (as hydrochloride), R25831 (as the free base).

3.2.2.2. *Giardia Lamblia*

3.2.2.2.1. Biology and Epidemiology

Giardia lamblia is a flagellated protozoan residing in humans in the upper portion of the small intestine. The trophozoite form of the parasite is converted to the infective cyst stage on its passage to the colon. Ingestion of cysts and the subsequent formation of motile trophozoite forms in the upper small intestine complete the life cycle. This parasite, which was long considered a harmless commensal, is now regarded as the most common intestinal parasite in the United States and United Kingdom.

Endemic infections with *Giardia* are found in every country of the world, but the incidence varies from one country to another. Giardiasis occurs more commonly in children. Transmission is either by direct fecal – oral contamination or by indirect transfer of cysts in food or water. Diarrhea is the most frequent symptom. The disease may be acute or chronic. Anorexia, abdominal cramps, bloating, and weight loss are common symptoms. Most infected individuals do not present evidence of the disease.

3.2.2.2.2. Chemotherapy

Several reviews describe in detail the treatment regimes for *Giardia* infection [191], [197], [198]. Although there is not general agreement as to the optimal chemotherapeutic agent, all three of the following chemical compounds have been reported to give cure rates $\geqq 90$ % with oral administration: metronidazole (and other 5-nitroimidazoles), quinacrine, and furazolidone. In the United States, quinacrine is the most commonly used drug, whereas in many other countries the 5-nitroimidazoles are preferred. Furazolidone is not widely used except as a suspension for children. This agent is generally considered less efficacious than the nitroimidazoles.

Metronidazole has been used to treat giardiasis since 1961, but the original dose regime has been modified. The recommended oral dose for metronidazole is 250 mg 3 times a day for 5 days, although shortened dosing regimes similar to those employed in the treatment of trichomoniasis have been reported to be successful. Other 5-nitroimidazoles that are employed include tinidazole, nimorazole, and ornidazole. Similar to metronidazole, these compounds are generally used with good clinical success in the same dose regimes as have been employed for treating trichomoniasis.

Metronidazole: see Section 3.2.2.1.2.

Tinidazole: see Section 3.2.2.1.2.

Nimorazole: see Section 3.2.2.1.2.

Ornidazole: see Section 3.2.2.1.2.

Quinacrine hydrochloride (U.S.P.) [*6151-30-0*], 6-chloro-9-{[4-(diethylamino)-1-methylbutyl]-amino}-2-methoxyacridine dihydrochloride dihydrate, $C_{23}H_{30}ClN_3O \cdot 2\,HCl \cdot 2\,H_2O$, M_r 508.91, *mp* 248 – 250 °C (*mp* poorly discernable), is a bitter, bright yellow crystal; anhydrous quinacrine [*83-89-6*], atabrine, mepacrine.

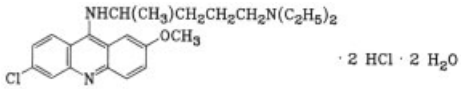

Quinacrine is an acridine derivative previously employed to treat malaria and tapeworm infections. The usual adult dose is 100 mg 3 time a day given for 5 days. The pediatric dose is 2 mg/kg 3 times a day for 5 days with a maximum of 300 mg/d. Gastrointestinal *adverse effects* as well as malaise and headache, occur with administration of this compound. Quinacrine is excreted in the urine, imparting a deep yellow color. Approximately 4 – 5 % of the patients develop yellow skin staining at the dose used for giardiasis. Quinacrine can induce red blood cell hemolysis in G-6-PD deficient patients. Prolonged high-dose therapy with quinacrine has caused rare instances of retinopathy similar to that observed with chloroquine.

Synthesis: condensing 1-(diethylamino)-4-aminopentane with 3,9-dichloro-7-methoxyacridine [199].

Furazolidone (INN, BAN, U.S.P.) [*67-45-8*], furoxone, 3-[(5-nitrofurfurylidene)amino]-2-oxazolidinone, $C_8H_7N_3O_5$, M_r 225.16, *mp* 275 °C (decomp.), occurs as yellow crystals.

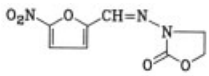

Furazolidone is a nitrofuran that is active against *G. lamblia* and various gram-negative bacteria, including *Salmonella, Shigella,* and *Vibrio cholerae.* The recommended adult oral dose is 100 mg 4 times a day for 7 – 10 days. The pediatric dose is 1.25 mg/kg 4 times a day for 7 – 10 days. Compared with other agents used for treating *Giardia* infections, furazolidone is somewhat less active. Furazolidone treatment can result in gastrointestinal and central nervous system *adverse effects.* Furazolidone turns the urine brown and can cause red blood cell hemolysis and mild, reversible anemia in individuals with G-6-PD deficiency. Furazolidone, similar to other nitrofurans, has carcinogenic potential. Adenocarcinoma of the lung in mice and of the mammary gland in female rats has been reported with long-term administration of furazolidone.

Synthesis: [200].

Trade names: Furox (Smith Kline Beckman), Furoxone (Norwich Eaton).

3.3. Sporozoans

Sporozoans have a complex life cycle that often involves two hosts (e.g., arthropod and human). The Coccidia, a subclass of essentially intestinal Sporozoa, and the hemosporidians, which include the malaria parasite, are animal and human parasites. *Toxoplasma,* a parasite of cats, is a common human parasite. *Babesia,* a tick-borne protozoan and a common animal parasite, is a rare human pathogen (these organisms are no longer considered sporozoans, but will be covered in this section). *Pneumocystis carinii* and *Cryptosporidium* have emerged as important human pathogens, especially in immunocompromised patients.

3.3.1. Plasmodia

3.3.1.1. Biology and Epidemiology

Malaria infections are distributed widely in countries of Africa, Asia, and Latin America, with an estimated worldwide prevalence of 100×10^6 cases associated with ca. 1×10^6 deaths each year. At least five species of *Plasmodium* can infect humans: *P. vivax, P. ovale, P. malaria, P. falciparum,* and *P. knowlesi* (only in Malaysia). In addition, at least two species of nonhuman primate plasmodia are transmissible to humans experimentally. Diagnosis of infection is made on the basis of morphological characteristics of specific species. The identification of the infecting organism is important in determination of the chemotherapeutic approach. Equally important in the therapy and prophylaxis of infection is an understanding of the life cycle and course of infection of the malaria parasites.

Transmission to humans is by the blood-sucking bite of various species of mosquitoes of the genus *Anopheles,* in which the sexual or sporogonic cycle of development occurs. The asexual cycle (schizogany) takes place in humans. The mosquito introduces the infective sporozoites, which quickly invade the liver parenchymal cells (the preerythrocytic cycle). After further development, numerous asexual progeny (the merozoites) enter the bloodstream and invade the erythrocytes. Multiplication in the red blood cell is characteristic for each species, resulting in synchronous destruction of host cells (the erythrocytic cycle). Successive production of merozoites occurs every 48 h (*P. vivax, P. ovale,* and *P. falciparum*) or at 72-h intervals (*P. malariae*). The *P. falciparum* infections are generally confined to the red blood cells after the first liver cycle; thus, untreated infections will terminate spontaneously, usually in 6 – 8 months, or end fatally. The other three species continue to multiply in liver cells, and persistence of the parasite in these cells long after the parasites have disappeared from the bloodstream is observed. During the erythrocytic cycle, certain merozoites become differentiated as male or female gametocytes. Thus, the sexual cycle begins in humans, but for continuation of the cycle, the gametocytes must be ingested by mosquitoes. Development time of the various stages in the mosquito's stomach wall ranges from 8 to 14 days, depending on the malaria species. The infective stage is the sporozoite, which migrates to the salivary gland of the mosquito and is injected with a blood-meal.

The *pathogenic mechanisms* resulting in clinical illness in humans can be divided into four

processes: (1) fever, (2) anemia, (3) tissue hypoxia, and (4) immunopathologic events. Cyclic fever and its physiological consequences involve rupture of erythrocytes and release of schizonts. Anemia is a common complication of malaria and mainly results from the rupture of red blood cells during schizogony, but other factors, such as autoimmune mechanisms, may contribute to the anemia. Tissue hypoxia, resulting from alterations in microcirculation and anemia, may cause serious complications including renal failure, pulmonary edema, and cerebral dysfunction in *P. falciparum* infections. Immunologic response to malaria infection can result in clinical disease. Immune-complex glomerulonephritis and greatly enlarged spleens are two examples of this phenomenon.

3.3.1.2. Chemotherapy

Antimalarial drugs can be divided into two groups on the basis of their mechanism of action: (1) The aminoquinolines, such as chloroquine, apparently exert their effect by intercalation into parasite DNA. This is not the only mechanism of action of these compounds, because mefloquine does not intercalate DNA. (2) The other group of compounds inhibits the synthesis of folic acid from 4-aminobenzoic acid. This group includes chloroguanide, pyrimethamine, and their derivatives, as well as the sulfonamides and sulfones. Quinine, a natural product and a mainstay of therapy for more than 60 years, still remains an important chemotherapeutic agent, especially for treating malaria parasites that are resistant to the newer synthetic agents.

Antimalarial drugs have been grouped in at least six different categories according to use, including (1) causal prophylaxis, (2) suppressive treatment, (3) clinical cure or treatment of the acute attack, (4) radical cure, (5) suppressive cure, and (6) gametocidal therapy [201]. True *causal prophylactic agents* should be capable of killing sporozoites prior to their entry into red blood cells. Although no such agent is available, several agents, e.g., primaquine and chloroguanide, have activity against the preerythrocytic stages of *P. falciparum*. *Suppressive therapy* inhibits development of the erythrocytic stages, thus preventing clinical symptoms. Chloroquine, chloroguanide, and pyrimethamine are suppressive; however, insensitivity or drug resistance in malarial strains in certain localities has created problems. A *clinical cure* can be achieved by agents that interrupt the development of the intracellular schizont in the red blood cell. The 4-aminoquinoline derivatives, chloroquine and amodiaquin, are the major drugs in this category, although the slower-acting chloroguanide and pyrimethamine are also highly active schizonticides. *Radical cure* involves elimination of both erythrocytic and exoerythrocytic parasites. Vivax malaria can be treated with the 8-aminoquinoline derivatives, of which only primaquine is used currently. Radical cure of falciparum malaria is relatively easy to achieve by continuation of treatment. A *suppressive cure* involves complete elimination of the malaria parasites by treatment that exceeds the life span of the parasite. *Drugs that kill gametocytes* directly are not available, but chloroguanide and pyrimethamine prevent the development of gametocytes in mosquitoes.

Drug resistance to all of the synthetic agents used in the treatment and prophylaxis of malaria has been observed in many areas of the world. *Plasmodium falciparum* is resistant to amodiaquine and chloroquine and to the combination of sulfadoxine and pyrimethamine. This therapeutic problem extends to South America, 13 African nations, south of the Equator, and to the South Asian continent, from Pakistan to eastern India. When used alone, quinine has lost some of its activity against falciparum malaria in those areas where it has been used indiscriminately or with poor compliance. Treatment of the other malarial species with 4-aminoquinoline derivatives, such as amodiaquine and chloroquine, still remains effective.

Chloroquine (INN, U.S.P.) [*54-05-7*], 7-chloro-4-(4-diethylamino-1-methylbutylamino)-quinoline; $C_{18}H_{26}ClN_3$, M_r 319.88, *mp* 87 °C; diphosphate, *mp* 193 – 195 °C or 215 – 218 °C (two modifications); sulfate, *mp* ca. 207 °C.

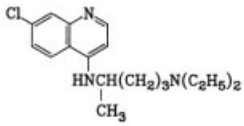

Chloroquine, a 4-aminoquinoline, is highly effective against the asexual erythrocytic forms of *P. vivax* and *P. falciparum,* and gametocytes of *P. vivax.* In human vivax malarias, chloroquine has no prophylactic or radically curative value; however, it is effective in terminating acute attacks of vivax malaria. When administered continuously for long periods, it acts as a suppressive agent. It is highly effective in controlling acute attacks of falciparum malaria and generally cures the disease. In certain parts of the world, drug-resistant strains of *P. falciparum* limit the use of chloroquine.

Chloroquine is well absorbed when administered orally. It is highly bound to tissues, especially liver, spleen, kidney, and lung; about 55 % of the drug is bound to plasma proteins. The drug is excreted slowly in the urine, with ca. 70 % being recovered as parent compound. Because of high tissue binding, a loading dose is necessary to achieve adequate plasma concentration. After a single dose, the half-life of the drug in plasma is 3 days, with a longer half-life achieved on multiple doses. Chloroquine at the dose employed for prophylaxis of acute malarial attacks causes relatively few *adverse effects*, mainly pruritus and gastrointestinal discomfort.

The oral dose for suppression of sensitive malarial strains in adults is 300 mg of the free base weekly, starting 2 weeks prior to and while present in a malarious area and for 6 weeks after leaving the area. For oral treatment of acute malaria in adults, 600 mg of base is given initially, followed by 300 mg 6 h later and 300 mg at 24 and 48 h. Primaquine should be given after chloroquine doses when treating malaria due to *P. vivax* and *P. ovale* or for radical cure following exposure.

Chloroquine is of use in treating extraintestinal amebiasis, and is also reported to be of some value in the treatment of *Giardia* and *Babesia* infections.

Synthesis: condensation of 4,7-dichloroquinoline with 1-(diethylamino)-4-aminopentane [202].

Trade names: Aralen hydrochloride, Aralen phosphate (Sterling Winthrop), Resochin (diphosphate, Bayer), Nivaquine (sulfate, Specia, France).

Amodiaquine (INN, U.S.P.) [*86-42-0*], 4-[(7-chloro-4-quinolyl)amino]-α-(diethylamino)-*o*-cresol, $C_{20}H_{22}ClN_3O$, M_r 358.87, *mp* 208 °C, is crystalline in form; dihydrochloride dihydrate, *mp* 150 – 160 °C (decomp.), is a yellow, bitter crystal.

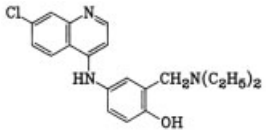

Amodiaquine, a congener of chloroquine, is claimed to have superior activity against some strains of *P. falciparum* that are partially resistant to chloroquine. It is used for treating overt malaria attacks and for suppression. The observed *adverse effects* and the frequency of these effects are similar to chloroquine, i.e., diarrhea, vomiting, and vertigo. Dosing also is similar to that employed for chloroquine. For suppressive therapy, the unit dose is 400 mg of the free base (520 mg of dihydrochloride). For treatment of acute attacks, 600 mg of the base is given initially with subsequent daily doses of 400 mg for 2 days.

Synthesis: Condensation of 4,7-dichloroquinoline and 4-acetamido-α-diethylamino-*o*-cresol gives amodiaquin dihydrochloride dihydrate [203].

Trade name: Camoquin hydrochloride (Parke-Davis).

Hydroxychloroquine sulfate (U.S.P.), hydroxyquinine (INN) [*747-36-4*], 2-({4-[(7-chloro-4-quinolyl)amino]pentyl}ethylamino)ethanolsulfate (1 : 1 salt), $C_{18}H_{26}ClN_3O \cdot H_2SO_4$, M_r 439.95, *mp* (usual form) ca. 240 °C, *mp* (other form) ca. 198 °C, is a white crystalline odorless powder with a bitter taste; hydroxychloroquine [*118-42-3*].

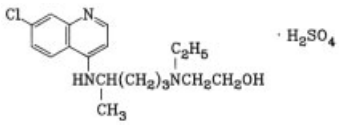

Hydroxychloroquine is an *N*-ethyl-β-hydroxylated chloroquine. This compound has an activity and safety profile that is similar to chloroquine. A dose of 400 mg of hydroxychloroquine sulfate is equivalent to 500 mg of chloroquine phosphate.

Synthesis: prepared by reacting a mixture of 4,7-dichloroquinoline, phenol, and N'-ethyl-N'-β-hydroxyethyl-1,4-pentadiamine at 125 – 130 °C [204].

Trade name: Plaquenil sulfate (Sterling-Winthrop, USA).

Primaquine (INN) [*90-34-6*]; primaquine phosphate (U.S.P.) [*63-45-6*], 8-[(4-amino-1-methylbutyl)amino]-6-methoxyquinoline phosphate (1 : 2 salt), $C_{15}H_{21}N_3O \cdot 2\,H_3PO_4$, M_r 455.34, *mp* 197 – 198 °C, occurs as yellow crystals (90 % ethanol).

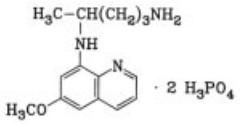

Primaquine is an α-aminoquinoline used for prevention of malarial relapses and for radical cure of *P. vivax* and *P. ovale* malaria by acting on the exoerythrocytic stage of the parasite. It also destroys the gametocytes of these species. It has no significant activity against the asexual blood forms of *P. falciparum.*

Primaquine, similar to all of the other α-aminoquinolines, is absorbed rapidly when administered orally and is rapidly metabolized. Plasma levels peak at 6 h. Minor *adverse effects* include nausea, abdominal discomfort, and headache. Severe hemolytic reactions in people with a glucose-6-phosphate dehydrogenase deficiency of the type found in those of Mediterranean ancestry have been observed. Most blacks have this enzyme deficiency, but it is usually confined to older erythrocytes and thus the hemolytic reaction is less severe.

Primaquine is supplied as the phosphate salt with 26.3 mg being equivalent to 15 mg of free base. For prevention of attack after departure from areas where *P. vivax* and *P. ovale* are endemic, a daily oral dose of 15 mg of base along with the last 2 weeks of chloroquine prophylaxis is recommended. For prevention of relapses to vivax and ovale malaria, a daily dose of 15 mg/d for 14 days or a weekly dose of 45 mg of base for 8 weeks is employed.

Synthesis: 6-Methoxy-8-aminoquinoline is condensed with 1-phthalimido-4-bromopentane, and the phthalyl group is cleaved by heating with hydrazine in ethanol and hydrochloric acid [205].

Pyrimethamine (INN, U.S.P.) [*58-14-0*], 2,4-diamino-5-(4-chlorophenyl)-6-ethylpyrimidine, $C_{12}H_{13}ClN_4$, M_r 248.71, *mp* 233 – 234 °C, is crystalline in form.

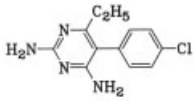

Pyrimethamine is a 2,4-diaminopyrimidine that has high affinity for the enzyme dihydrofolate reductase from the malaria parasite and thus interferes with folate synthesis. Trimethoprim, a related compound with good antibacterial activity, also has antimalarial activity. Pyrimethamine by itself has little value in treating a primary attack of malaria. However, combined with sulfadoxine (Fansidar), malaria has been prevented in subjects in areas where there is drug-resistant falciparum malaria. The combination oral product of 25 mg of pyrimethamine and 500 mg of sulfadoxine is given once weekly during exposure. It is also employed in combination with quinine sulfate for treatment of chloroquine-resistant falciparum malaria. The dose employed for the combination is 650 mg of quinine 3 times daily and 25 mg of pyrimethamine twice daily for 3 days. Pyrimethamine is well absorbed when administered orally. It is eliminated slowly with a plasma half-life of 4 days. The dose of 25 mg weekly produces few adverse effects. Excessive doses may result in a reversible megaloblastic anemia resembling folic acid deficiency.

Synthesis: prepared by condensing 3-isobutoxy-2-(4-chlorophenyl)pent-2-enonitrile with guanidine nitrate in the presence of sodium methylate [206].

Trade name: Daraprim (Burroughs Wellcome).

Chloroguanide [*500-92-5*], chloroguanide hydrochloride (U.S.P. **14**) [*637-32-1*], proguanil (INN), 1-(4-chlorophenyl)-5-isopropylbiguanide hydrochloride, $C_{11}H_{16}ClN_5 \cdot HCl$, M_r 290.20, *mp* 243 – 244 °C, is a white powder with a bitter taste.

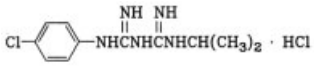

Chloroguanide is used to treat overt clinical vivax and falciparum malaria; however, response to treatment is slower than that observed with most other antimalarial agents. Chloroguanide is not recommended for treatment of an acute attack of falciparum malaria. Although it is of use in acute vivax malaria, it offers no advantages over other available agents. It is a causal prophylactic, suppressive, and radical cure agent in falciparum malaria. Chloroguanide is active against developing preerythrocytic stages of some malarias and is reported to sterilize gametocytes. Resistance to chloroguanide greatly compromises its usefulness.

Chloroguanide is converted metabolically to a triazine derivative that inhibits the enzyme dihydrofolate reductase. It is slowly absorbed from the gastrointestinal tract with peak serum concentrations achieved at 2 – 4 h after oral administration. About 50 % of the drug is excreted in the urine and 10 % directly into the feces. The prophylactic dose is 100 – 200 mg daily for nonimmune subjects. Less drug (300 mg weekly) is employed for partly immune individuals. For treatment of acute attack of vivax malaria, an initial dose of 300 – 600 mg is followed by a daily dose of 300 mg, usually for 5 – 10 days. Chloroguanide is a relatively safe drug at the doses employed for malaria treatment. Large daily doses of 1 g may cause vomiting, abdominal pain, and diarrhea.

Synthesis: prepared by heating 4-chlorophenyldicyandiamide with isopropylamine [207].

Trade name: Paludrine (ICI, United Kingdom).

Mefloquine (INN) [*53230-10-7*], (DL-*erythro*-α-2-piperidyl-2,8-bis(trifluoromethyl)-4-quinolinemethanol, $C_{17}H_{16}F_6N_2O$, M_r 378.32, *mp* 178 – 178.5 °C; mefloquine · HCl [*51773-92-3*], *mp* 259 – 260 °C.

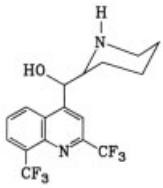

Mefloquine is a 4-quinolinemethanol that is currently well along into clinical trials, especially in Asia and South America, where multiple drug-resistant *P. falciparum* is a major problem [208, 209]. This compound has an unusually long half-life in humans (17 days) and may be efficacious for use as a chemoprophylactic when given weekly. It has proved highly effective in curing infections caused by *P. falciparum* when administered orally. Liver concentrations are high and prolonged, and the drug is excreted mainly in the feces. A quinate salt is available with a markedly increased water solubility and an improved pharmacokinetic profile.

Synthesis: [210].

Trade name: Mefloquine Quinate, $C_{17}H_{16}F_6N_{20} \cdot C_7H_{12}O_6$ (Smith Kline Beckman).

Halofantrine [*69756-53-2*], [±]-halofantrine [*66051-63-6*]; halofantrine hydrochloride (INN) [*36167-63-2*], 1,3-dichloro-α-[2-(dibutylamino)-ethyl]-6-(trifluoromethyl)-9-phenanthrenemethanol hydrochloride, $C_{26}H_{30}Cl_2F_3NO \cdot HCl$, M_r 536.89, *mp* 93 – 96 °C, *mp* 203 – 204 °C, occurs as crystals.

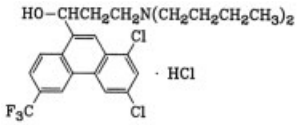

Halofantrine is a 9-phenanthrene methanol currently under development for the treatment of drug-resistant falciparum malaria. It is curative with minimal adverse effects in nonimmune subjects infected with Vietnam or Cambodian strains of *P. falciparum* or Chesson strain of *P. vivax*. The oral dose employed in these studies was 250 mg every 6 h for 12 doses [211]. The compound is rapidly but incompletely absorbed when administered orally.

Synthesis: [212].

Trade name: Halofantrine-β-glycerophosphate, $C_{26}H_{30}Cl_2F_3NO \cdot C_3H_9O_6P$ (Smith Kline Beckman).

3.3.2. Babesia

Babesia are intraerythrocytic protozoan parasites that are transmitted by ticks. They are mainly parasites of domestic and wild animals, although on occasion they can be transmitted to humans. Infections tend to be self-limiting and are characterized by fever, hemolytic anemia, and hemoglobinuria. No effective therapy exists for these infections.

3.3.3. Isosporiasis

The sporozoan order Coccidia contains a number of parasites that invade the intestinal mucosa at one stage in their life cycle. *Isospora belli,* usually acquired by fecal contamination of food or water, is a member of this group. Although a rare human pathogen, *T. belli* can cause serious gastrointestinal disease. Treatment is usually with antifolates, such as pyrimethamine and sulfadiazine. The drug of choice is trimethoprim – sulfamethoxazole. The usual oral dose of this combination is 160 mg of trimethoprim and 800 mg of sulfamethoxazole four times daily for 10 days then twice a day for 3 weeks. *Isospora hominis,* now called *Sarcocystis hominis,* is a related parasite that can cause intestinal infection in humans. This organism, similar to toxoplasma, has an asexual stage in the muscles of many mammals. Human infection is acquired from eating improperly cooked beef or pork containing *Sarcocystis.* Treatment is similar to that employed for *I. belli.*

3.3.4. Toxoplasmosis

Toxoplasma gondii is a sporozoan of the order Coccidia. The parasite is an obligate intracellular organism that exists in three forms including the tachyzoites (formerly trophozoites), tissue cysts, and oocysts. *Toxoplasma gondii* is a ubiquitous parasite with a worldwide distribution and a capability of infecting a wide range of animals and birds. The final hosts are members of the cat family. Infection is by ingestion of either tissue cysts or oocysts, releasing viable organisms that invade the epithelial cells of the intestine, where they undergo an asexual cycle and a sexual cycle in the cat. Many oocysts are shed in the feces and sporulate outside the cat.

Ingestion of primarily lamb or pork containing tissue cysts or other food products contaminated with oocysts is the major means of transmission. Active infection during gestation can result in congenital infection. Toxoplasma infection can be acute or chronic, symptomatic or asymptomatic. Most infected adults show no symptoms. Chronic infections can be latent and later exacerbate in immunocompromised individuals, resulting in severe disease, such as encephalitis, myocarditis, and pneumonitis. Congenital infection may lead to subsequent disease, such as impaired vision, hearing loss, or neurologic disorders.

The most effective treatment for toxoplasmosis is a combination of pyrimethamine and sulfadiazine. Although the tachyzoites are susceptible to these agents, the tissue cyst is resistant to available agents. A loading regime of pyrimethamine at 100 mg $kg^{-1}\,d^{-1}$ twice a day for 2 days followed by 25 mg every other day is the recommended treatment for adults with significant infection. A synergistic sulfa drug, such as sulfadiazine, sulfamethazine, or sulfamerazine, is usually given in combination with pyrimethamine. The antibiotics spiramycin and clindamycin or trimethoprim – sulfamethoxazole are less active than the combination of pyrimethamine and a sulfa drug.

3.3.5. Cryptosporidium

This coccidian protozoa is a significant cause of death in immunocompromised patients. The parasite, long known in domestic animals, is the same organism that causes disease in humans. Infection is acquired by ingestion of oocysts, which excyst and release sporozoites that invade and replicate in the intestinal microvilli. Male and female gametes are produced, initiating the sexual cycle. Oocysts are then formed and can be passed in the feces or can release sporozoites to cause autoinfection. In individuals with normal host defenses, the organism may

not cause symptoms or may cause a transient diarrhea and gastrointestinal disease that is generally self-limiting. The disease is more serious in immunocompromised individuals and is a major contribution to death in patients with acquired immune deficiency syndrome (AIDS). No effective chemotherapy exists for cryptosporidiosis.

3.3.6. Pneumocystis Carinii

This protozoan causes infection in the immunocompromised host and is the most common pathogen of AIDS patients, accounting for 43 % of reported opportunistic diseases. The disease is usually limited to the lungs, where it causes a diffuse pneumonitis. The parasite has been found in the lungs of a wide variety of animals and is distributed globally. Subclinical infections in normal individuals are probably common. The untreated infection in the immunocompromised patient is 90 – 100 % fatal. The drug of choice for treating pneumocystis infection is the combination of trimethoprim and sulfamethoxazole. The recommended dose is 20 mg kg^{-1} d^{-1} of trimethoprim with 100 mg kg^{-1} d^{-1} of sulfamethoxazole given orally or intravenously in 4 doses for 14 days. This combination therapy results in a high incidence of rash, neutropenia, fever, and diarrhea in AIDS patients. Pentamidine isethionate is an alternative choice, but is associated with significant adverse effects. The recommended dose is 4 mg kg^{-1} d^{-1} administered intramuscularly for 12 – 13 days.

3.4. Ciliates

Infections in humans caused by ciliated protozoans are rare. Only one parasite, *Balantidium coli*, causes disease in humans. The ciliated form of this parasite penetrates the mucosa of the colon and multiplies in the submucosal tissues. This results in colitis, with mucus and blood in the feces. The parasite can form cysts when conditions are not ideal for penetration and it is the cyst that allows for transmission of this disease, usually from animals (swine) to humans. The disease is treated with the antibiotics tetracycline or paromomycin or with metronidazole (750 mg twice a day for 5 days).

3.5. Amebas

3.5.1. Biology and Epidemiology

The disease amebiasis is worldwide in distribution and is almost always associated with poor sanitary conditions. The etiological agent, *Entamoeba histolytica*, occurs in three stages: (1) the inactive cyst, (2) the intermediate precyst, and (3) the ameboid trophozoite, which is the only stage found in the tissues. The ameboid form is also found in liquid feces during amebic dysentery. Multiplication among trophozoites occurs by binary fission. After ingestion of infective cysts in food or water contaminated with feces, the cysts are activated in the stomach, and development takes place during passage to the large intestine.

A population of lumen-dwelling trophozoites, capable of invading the intestinal epithelium, emerge from the cyst. Most infections are without symptoms, although when tissue invasion occurs (in ca. 10 % of the cases), there is disease associated with the infection. Asymptomatic infected persons harbor lumen-dwelling amebas that produce cysts, which are passed in the feces. Diseased individuals usually pass trophozoites and cysts in the feces. Extraintestinal amebiasis is observed in ca. 4 % or more of clinical infections, usually taking the form of amebic hepatitis or liver abscess. Abscesses may also occur rarely in other areas, including the lungs, brain, and spleen.

Entamoeba histolytica must be distinguished from four other amebas that are also intestinal parasites of humans. These include the common ameba *Ent. coli; Dientamoeba fragilis,* the only intestinal ameba other than *Ent. histolytica* suspected of causing diarrhea, but not of invading the tissues; *Iodamoeba butschlii;* and *Endolimax nana.*

Primary amebic meningoencephalitis has been reported in less than 100 cases with amebas living free in soil and in water. Two genera of ameba, *Naegleria* and *Acanthamoeba,* have been associated with pathogenicity. Most cases of infection have developed in children who were swimming in contaminated outdoor pools. *Naegleria* meningoencephalitis infection is rapidly fatal and does not respond to the available amebicides. A few patients have been re-

ported to respond to intravenous and intrathecally administered amphotericin B.

3.5.2. Chemotherapy

The approach to treating amebiasis is in part based on the location of the amebic organism and includes treatment of the asymptomatic individual who is passing cysts. The criterion for cure in intestinal amebiasis is the elimination of the organism. This is usually achieved with tetracycline and an 8-hydroxyquinoline derivative or diloxanide furoate. The antibiotic paromomycin also may be employed. For those infections involving only the bowel wall and resulting in acute amebic dysentery, a variety of agents are employed. Symptomatic amebiasis involving the intestine can be treated with a nitroimidazole, such as metronidazole or tinidazole, a course of therapy usually resulting in 90 % cure. Dehydroemetine, a semisynthetic derivative of the alkaloid emetine, can be used for the rapid relief of symptoms in severely ill patients. As an alternative to metronidazole, combination therapy of dehydroemetine with tetracycline or paromomycin has been employed. For extraintestinal infections, use of metronidazole is recommended, usually in combination with diiodohydroxyquin or diloxanide furoate, to prevent continued intraluminal infection. In seriously ill patients with complicated amebic infections, parenteral emetine in combination with iodoquinol may be employed.

8-Hydroxyquinolines. A number of 8-hydroxyquinolines are available for treating amebiasis. Iodoquinol, the most widely used drug, is described in this chapter; however, other derivatives have been applied in specific areas of the world. These include clioquinol (iodochlorhydroxyquin), broxyquinoline, chlorquinaldol, and chiniofon.

Iodoquinol [*83-73-8*], diiodohydroxyquinoline (INN, U.S.P.), 5,7-diiodo-8-quinolinol, $C_9H_5I_2NO$, M_r 396.95, *mp* 200 – 215 °C (extensive decomp.), occurs as crystals or yellowish brown powder.

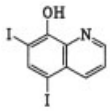

Iodoquinol is an 8-hydroxyquinoline that is directly amebicidal. It is active only on amebas in the intestinal tract and, like the other members of this class of compounds, is ineffective against amebic abscess and hepatitis. The drug is partly absorbed when it is administered orally, with ca. 25 % of the drug recovered in the urine as the glucuronide. Although the 8-hydroxyquinolines were originally thought to be of low toxicity, a number of *toxic reactions* are now known to result from their use, the most significant being a subacute myelooptic neuropathy, which is particularly observed with clioquinol. The usual dose of iodoquinol for treating asymptomatic carriers of *Ent. histolytica* is 650 mg 3 times a day for 20 days. This same dose may be employed in combination therapy with metronidazole in treating frank intestinal disease or in hepatic abscess, where it serves to limit reoccurrence of the intestinal forms.

Synthesis: prepared by the action of iodine monochloride on 8-hydroxyquinoline [213].

Trade name: Yodoxin (Glenwood).

Clioquinol [*130-26-7*], iodochlorhydroxyquin (INN, U.S.P., BAN), 5-chloro-8-hydroxy-7-iodoquinoline, C_9H_5ClINO, M_r 305.50, *mp* ca. 178 – 179 °C (decomp.), is a brownish-yellow bulky powder.

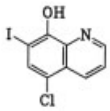

Synthesis: [214].

Trade names: Vioform (Ciba-Geigy), Rheaform, veterinary (Squibb).

Broxyquinoline (INN, DCF, MI) [*521-74-4*], 5,7-dibromo-8-hydroxyquinoline, $C_9H_5Br_2NO$, M_r 302.95, *mp* 196 °C, occurs as monoclinic needles.

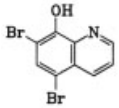

Synthesis: prepared by bromination of 8-quinolinol [215].

Chiniofon (INN, DCF, NFN, NF **11**) [*8002-90-2*], mixture of four parts (by mass) 8-hydroxy-7-iodo-5-quinolinesulfonic acid and one part sodium hydrogencarbonate.

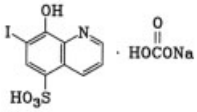

Synthesis: of 8-hydroxy-7-iodo-5-quinolinesulfonic acid [216].

Diloxanide furoate [*3736-81-0*], 4-[(dichloroacetyl)methylamino]phenyl-2-furancarboxylate, $C_{14}H_{11}Cl_2NO_4$, M_r 328.15, *mp* 112.5 – 114 °C.

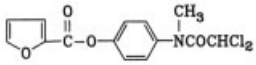

The furoate ester of diloxanide is one of the agents of choice in the treatment of persons who are asymptomatic passers of cysts or for invasive and extraintestinal amebiasis (administered with other appropriate drugs). The recommended oral dose is 500 mg 3 times daily given for 10 days; a second course may be necessary. When administered orally, the ester is hydrolyzed in the lumen or mucosa of the intestine, resulting in diloxanide and furoic acid. Most of the oral dose is excreted in the urine within 48 h, and peak drug concentrations appear in the plasma in 1 h. There are few *adverse effects* reported with this compound, the most common being flatulence or mild abdominal discomfort.

Diloxanide (INN, BAN, DCF, NFN, MI) [*579-38-4*], 2,2-dichloro-4′-hydroxy-*N*-methylacetanilide, $C_9H_9Cl_2NO_2$, M_r 234.08, *mp* 175 °C, is crystalline in form.

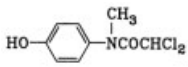

Synthesis: Diloxanide is prepared from 3,3-dianisyl-4-hexanone [217]. Diloxanide furoate is obtained by reacting diloxanide with 2-furoyl chloride in pyridine.

4. Chemotherapy of Viral Infections

4.1. Physical and Biological Characteristics

Obligate Intracellular Parasites. Viruses are among the smallest of all life forms. Unlike bacteria, viruses are obligate intracellular parasites that are metabolically inert in the extracellular state. Even though viruses depend to varying degrees on the host cell's metabolic machinery, the differences existing between the metabolic processes of the host cell and those specified by viruses have been exploited in the search for antiviral drugs that selectively inhibit virus replication.

Nucleic Acid. Unlike other microorganisms, viruses contain only one form (double-stranded or single-stranded) of only one type (RNA or DNA) of nucleic acid as their genetic material. The polarity of the genetic material determines whether transcriptase enzymes (enzymes not found in eukaryotic cells) must be present within virions. For example, a negative-stranded RNA genome is complementary to messenger RNA (mRNA) and must be transcribed by a virion transcriptase enzyme to produce viral mRNA necessary for initiation of the infection process. Other virion transcriptases, such as the reverse transcriptase present in retroviruses, catalyze the formation of DNA from viral RNA. The unique reverse transcriptase enzyme of the human T-cell lymphotrophic virus type III, lymphadenopathy associated virus, AIDS-related virus, or human immunodeficiency virus (HIV), which are various names for the virus implicated in AIDS (acquired immune deficiency syndrome), is the target of several antiviral drugs.

Viral Proteins. The functions of the structural virion proteins are to protect the genetic material from degradation and also to determine what types of host cells can be infected by a particular virus (host range). The latter function

is determined by envelope proteins, which may be glycosylated, because these proteins are the first to contact the host cell. In viruses that lack an envelope, capsid proteins protect the genetic material and determine host range. The capsid is the protein coat that is complexed with the nucleic acid. Nonstructural proteins, such as the virion transcriptases, are also present within the virion.

Lipid Envelope. The lipids present in enveloped viruses resemble, in composition, those lipids of the host-cell membranes. Certain viral proteins may be embedded in the lipid bilayer of virions.

Laboratory Diagnosis of Viral Infections. Rapid detection of viral infections is critical for rational chemotherapeutic intervention [218]. Detection methods include serological tests, as well as direct examination of clinical specimens for virus. Serological tests measure the rise in serum antibody late in the disease course (convalescent) vs. serum antibody levels found early in the course of the disease (acute). The presence of antibodies specific for a particular virus may be difficult to assess if antibodies to related viruses are present or if the virus infection represents reactivation or reinfection of the host with the same virus. The immunoglobulin antibody of class IgM, which is produced early in infections, can indicate the presence of active virus replication. This IgM antibody can be detected by using IgM capture techniques in which serum interacts with anti-IgM on an immobilized solid surface. Other serological techniques use enzyme- (peroxidase), radioactive-, or fluorescent-labeled antibodies.

Direct observation using an electron microscope can rapidly detect some virus infections. This technique works well for examining fecal contents in acute gastroenteritis caused by rotaviruses, caliciviruses, astroviruses, and the Norwalk agent, in which sufficiently large numbers of morphologically distinct virions are present. In these cases, the viruses do not grow well in cell culture.

The classic method, and still the standard, is to isolate and identify viruses by growing them from clinical specimens in a cell culture. Several different cell cultures must be used because no one culture can support the replication of all viruses. Preliminary virus identification can be obtained by examination of infected cells for the cytopathic effect (CPE) that is characteristic for

Table 4. Virus families

Virus family	Distinguishing characteristics *	Human pathogens *
Poxviridae	DS DNA, enveloped	smallpox, vaccinia
Herpesviridae	DS DNA, enveloped	HSV, VZV, CMV, EBV
Adenoviridae	DS DNA, nonenveloped	human adenoviruses
Papovaviridae	DS DNA, nonenveloped	papilloma, BK and JC
Hepadnaviridae **	DS DNA, nonenveloped	hepatitis B virus
Parvoviridae	SS DNA, nonenveloped	parvoviruses, Norwalk (?)
Reoviridae	DS RNA, nonenveloped	reoviruses, rotaviruses
Togaviridae	SS RNA, enveloped, no DNA step, (+) sense	EEE, WEE, VEE, rubella, yellow fever, dengue
Coronaviridae	SS RNA, enveloped, no DNA step, (+) sense	coronaviruses
Paramyxoviridae	SS RNA, enveloped, no DNA step, (−) sense, nonsegmented	parainfluenza, measles, mumps, RSV
Rhabdoviridae	SS RNA, enveloped, no DNA step, (−) sense, nonsegmented	VSV, rabies
Filoviridae **	SS RNA, enveloped, no DNA step, (−) sense, nonsegmented	Marburg, Ebola
Orthomyxoviridae	SS RNA, enveloped, no DNA step, (−) sense, segmented	influenza A and B
Bunyaviridae	SS RNA, enveloped, no DNA step, segmented, ambisense	Bunyamwera Rift Valley fever
Arenaviridae	SS RNA, enveloped, no DNA step, segmented, ambisense	Lassa, Machupo, Junin, LCMV
Retroviridae	SS RNA, enveloped, DNA step	human T-lymphotrophic viruses
Picornaviridae	SS RNA, nonenveloped	poliovirus, Coxsackievirus, echovirus, hepatitis A, rhinovirus
Caliciviridae	SS RNA, nonenveloped	Norwalk (?), caliciviruses

* DS = double-stranded; SS = single-stranded; HSV = herpes simplex virus; VZV = varicella – zoster virus; CMV = cytomegalovirus; EBV = Epstein – Barr virus; EEE = eastern equine encephalitis virus; WEE = western equine encephalitis virus; VEE = Venezuelan equine encephalitis virus; RSV = respiratory syncytial virus; VSV = vesicular stomatitis virus; LCMV = lymphocytic choriomeningitis virus; (+) sense = genomes containing nucleotide sequences translated by ribosomes; (−) sense = genomes composed of nucleotide sequences complementary to (+) sense; ambisense = genome composed of (−) sense covalently attached to (+) sense.
** Not yet approved by ICTV.

a virus group. In the absence of a characteristic CPE, serological tests are used for final identification of viruses.

Many viruses have been cloned by using recombinant DNA techniques. Probes produced from these cloned virus DNAs can be used to detect certain viruses in clinical samples. Eventually, nonradioactive probes that use biotin or fluorescent tags may be widely used in clinical laboratories.

4.2. Classification of Viruses

The International Committee on Taxonomy of Viruses (ICTV) published a report in 1982 [219] that is summarized in Table 4. A total of 16 of the 55 virus families recognized by the ICTV contain viruses that are human pathogens. The distinguishing characteristics are the type of nucleic acid, the presence or absence of an envelope, the replication strategy of the genome, whether it is a positive-, negative-, or ambisense genome, and genome segmentation.

4.3. Assessment of Antiviral Activity in Cell Culture

In vitro cell culture systems for determining the antiviral effect of compounds have been useful, even though in vitro activity does not always predict in vivo or clinical efficacy. Plaque reduction [220], [221], yield reduction, cell growth, and macromolecular synthesis assays are among of the methods used to quantify the effects of antiviral candidates on virus replication and host cell metabolism. In the plaque reduction assay, the concentration of compound that reduces the number of plaques to a defined end point (usually 50 % or 90 %) is determined. In the yield reduction assay, the concentration of compound that reduces the yield of infectious virus by 50 % or 90 % is determined. In the cell growth or macromolecular synthesis assay, the effect of a compound on the increase in cell number or the incorporation of radioactive precursors into DNA, RNA, and protein, respectively, is determined. A compound is a viral-specific inhibitor if the antiviral concentration in the plaque reduction or virus yield assay does not affect cell growth or macromolecular synthesis.

4.4. Animal Models of Virus Infection

In addition to the in vitro cell culture systems described in Section 4.3, animal models of human virus infections are used to assess the systemic efficacy of candidate antiviral compounds. A compound that is a potent inhibitor of virus replication in a cell culture is commonly completely ineffective in preventing disease when administered to a virus-infected animal. These results are usually explained by poor absorption or rapid metabolism of the compound, resulting in a concentration of the compound that is too low to inhibit virus replication in the critical tissues.

Results in animal models of human viral infections often differ significantly from the natural course of the human infection, even when the identical virus is used. For example, mice infected intracerebrally with a mouse-adapted human poliovirus develop flaccid limb paralysis and succumb to the infection within 24 h of the onset of paralysis [222]. The severity of the poliovirus infection in mice is greater than that usually observed in humans, presumably because the route of inoculation in mice does not allow the animal to develop a protective immune response prior to the onset of paralysis. In humans, poliovirus replicates first in the Peyer's patches in the gut, which stimulates an immune response prior to the spread of virus to the central nervous system.

Mouse models of herpes simplex virus (HSV) infections are used to assess the efficacy of antiherpetic compounds; however, the disease in mice is more severe than the acute infection in humans. Mice generally succumb to encephalitis following intraperitoneal or intravaginal inoculation of virus. The severity of the disease in the animal model used in evaluating an antiviral compound should be considered when assessing potency and projecting dose levels in humans.

4.5. Rationale for Chemotherapy of Viral Infections

The rationale for the discovery of specific chemotherapeutic agents effective against

viruses is based on the knowledge that certain virus functions are distinct from cellular functions (for reviews of antiviral chemotherapeutics, see [223]). A number of unique virus-coded enzymes are known to be critical for virus replication. These enzymes (e.g., deoxypyrimidine kinase, also known as thymidine kinase, and ribonucleotide reductase from HSV) have different substrate specificities compared to cellular enzymes and, therefore, can be used as targets for antiviral chemotherapy. The antiherpetic drug, acyclovir (see Section 4.6.1), is an excellent example of the exploitation of the differences between the viral and cellular enzyme. Acyclovir is activated by deoxypyrimidine kinase to eventually form acyclo-GTP, which inhibits the viral DNA polymerase. This process does not occur to a significant extent in uninfected cells. A second example is the reverse transcriptase enzyme (from the retroviruses implicated in AIDS), which presumably does not have a normal cell counterpart. This enzyme has been the target of several agents that specifically inhibit the replication of the AIDS viruses.

Virus replication can also be inhibited by agents, such as rhodanine or arildone, that specifically bind with high affinity to the picornaviral capsid proteins and stabilize virion capsid conformation. This stabilization prevents cell-induced uncoating of the virion nucleic acid, thereby halting the virus infection at the uncoating step.

The host's immune response to a viral infection works in concert with antiviral drugs to clear the infection. The virus infection of a host can be viewed as a race between replication of the virus and the mounting immune response to the virus. If the extent of virus replication and resultant tissue destruction exceeds a threshold value, disease occurs. If the extent of virus replication is limited by the immune response or by chemotherapeutic intervention, then the disease course will be milder or asymptomatic. Thus, the role of antiviral chemotherapy in the treatment of viral disease is to inhibit virus replication sufficiently to enable the host's immune system to overcome the virus infection.

4.6. Chemotherapeutic Agents

4.6.1. Nucleoside Analogues

Acyclovir (INN) [*59277-89-3*], 9-[(2-hydroxyethoxy)methyl]guanine, $C_8H_{11}N_5O_3 \cdot 1/2\,H_2O$, M_r 234.2, $mp > 200\,°C$.

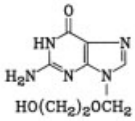

The antiherpetic activity of acyclovir, the acyclic analogue of guanosine, was first reported along with its synthesis in 1978 [224]. The rationale for this synthesis was the previous observation that the acyclic analogue of adenosine was a substrate for adenosine deaminase [225]. Acyclovir has potent activity against several herpesviruses (herpes simplex virus types 1 and 2, Epstein – Barr virus, and varicella – zoster virus), but is not as potent against human cytomegalovirus.

Acyclovir is clinically effective in the treatment of herpesvirus infections in humans. Acyclovir was first reported to be clinically efficacious in the topical treatment of herpetic keratitis [226]. Acyclovir was subsequently shown to be as effective as vidarabine [227] or as trifluorothymidine [228] against herpetic keratitis. Topical acyclovir cream was found to be effective against herpes labialis [229] and primary genital herpes [230], but was not effective against recurrent genital herpes [231]. Unlike the topical cream, orally administered acyclovir did reduce the duration of viral shedding, as well as the time to healing in recurrent genital herpes. Oral acyclovir, however, had no significant effect on symptoms of recurrent genital herpes [232–234]. Treatment protocols using daily administration of prophylactic oral acyclovir for 4-month periods have been described for patients with frequent recurrences [235], [236]. During treatment with acyclovir, the patients had significantly fewer recurrences than those receiving placebo. However, after treatment with acyclovir was stopped, the patients' recurrence rates returned to pretreatment frequencies. While on treatment with acyclovir, many patients reported the symptoms of a pro-

drome that did not progress to a complete lesion [237].

Adverse Effects. Several problems exist with the chronic use of acyclovir in patients. The major question is one of long-term safety. First, because acyclovir is a nucleoside analogue, there is concern it may be mutagenic. Thus far, however, acyclovir has not been shown to cause cancer in laboratory animals. Furthermore, it is not known whether long-term prophylactic use of acyclovir is more likely to result in development of resistant viruses. Other minor concerns with acyclovir include reversible toxicity to the bone marrow, reversible kidney damage due to acyclovir crystals, and diarrhea. Because only 15 – 20 % of acyclovir is absorbed after oral administration, a prodrug of acyclovir activated by xanthine oxidase has recently been developed [238].

Synthesis: Reaction of 2,6-dichloro-9-(2-benzoyloxyethoxymethyl) purine with methanolic ammonia, followed by treatment with nitrous acid and then with methanolic ammonia yields acyclovir [224].

Trade name: Zovirax (Burroughs Wellcome).

Idoxuidine [*54-42-2*], IDU (INN, U.S.P.), 2′-deoxy-5-iodouridine, $C_9H_{11}IN_2O_5$, M_r 354.12, *mp* 160 °C (decomp.).

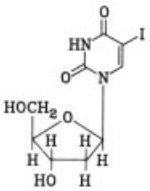

Idoxuidine was the first clinically effective nucleoside analogue. It was synthesized in 1959 [239] as part of a program aimed at developing synthetic nucleosides as DNA synthesis inhibitors to treat cancer. In effect, antiviral chemotherapy became a reality in 1962 when IDU was reported to cure herpes simplex keratitis in humans [240]. Although IDU was effective topically against herpes simplex keratitis, it was not active when administered systemically.

Adverse Effects. Because of the insolubility and toxicity of IDU, other drugs, such as trifluorothymidine, are used for the topical treatment of ocular herpetic infections [241].

Synthesis: obtained by refluxing uracildeoxyriboside, iodine, chloroform, and HNO_3 [242].

Trade names: Stoxil (Smith Kline).

Trifluridine (INN) [*70-00-8*], 5′-trifluoromethyl-2′-deoxyuridine, trifluorothymidine, TFT, $C_{10}H_{11}F_3N_2O_5$, M_r 296.2, *mp* 186 – 189 °C.

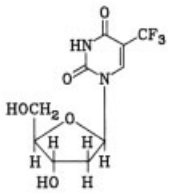

Trifluridine was first synthesized in 1964 [243]. Its use is limited to the topical treatment of herpetic keratitis. It was more effective than IDU in patients with dendritic or ameboid herpetic keratitis [244].

The *antiviral activity* and the *mechanisms of action* of TFT have been reviewed [245], [246]. It is converted to the active triphosphate form by cellular and viral thymidine kinases. The triphosphate form of TFT is then preferentially incorporated into viral DNA, which results in inhibition of the transcription of late viral mRNAs.

Adverse Effects. Because TFT can also be incorporated into cellular DNA, TFT is toxic to uninfected cells; thus, TFT cannot be used systemically.

Synthesis: prepared by treatment of 5-trifluoromethyluracil with a bacterial enzyme [243].

Trade name: Viroptic (Burroughs Wellcome).

Vidarabine [*24356-66-9*], ara-A (INN, U.S.P.), 9-β-D-arabinofuranosyladenine, $C_{10}H_{13}N_5O_4 \cdot 4\,H_2O$, M_r 339.36, *mp* 257 °C.

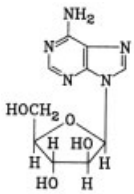

Ara-A was initially synthesized in the 1960s as a potential anticancer agent [247], [248]. Ara-A was reported to inhibit herpesvirus, vaccinia virus, cytomegalovirus, and varicella – zoster viruses [249], [250]. Topical ara-A was as effective against herpetic keratitis as IDU [251]. Intravenous ara-A was also effective against herpes zoster in immunosuppressed patients [252]. Ara-A was sufficiently nontoxic to uninfected cells to permit systemic administration to patients with herpes encephalitis. Ara-A reduced mortality from 70 to 28 % in a study of 28 patients with herpes encephalitis [253]. Recently, acyclovir was shown to be superior to ara-A for herpes encephalitis [254]. Topical ara-A and ara-A monophosphate have no effect in patients with genital or labial herpes [255], [256].

Mechanism of Action. Clearly, ara-A has multiple sites of inhibition of HSV. Ara-A is phosphorylated to give ara-A triphosphate (ara-ATP). Ara-ATP is a selective inhibitor of ribonucleoside diphosphate reductase and HSV DNA polymerase, as well as of the addition of poly-A to viral mRNA. Ara-A also slows viral DNA elongation, inhibits terminal deoxynucleotidyl transferase, and inhibits *S*-adenosylmethionine-dependent methylation (capping) of viral mRNA [257].

Synthesis: Ara-A is prepared by treating 9-(3′,5′-*O*-isopropylidene-β-D-xylofuranosyl)adenine with methanesulfonyl chloride. The crystalline 9-(3′,5′-*O*-isopropylidene-2′-*O*-methanesulfonyl-β-D-xylofuranosyl)adenine is exposed to 90 % aqueous acetic acid at 100 °C for 5 h. The epoxide is formed with the use of methanolic sodium hydroxide. The epoxide is converted to the arabinoside by reaction with sodium benzoate or sodium acetate in 95 % aqueous *N,N*-dimethylformamide [248].

Trade name: Vira-A (Parke-Davis).

Bromovinyldeoxuridine [*73110-56-2*], BVDU, (*E*)-5-(2-bromoethenyl)-2′-deoxyuridine, $C_{11}H_{13}BrN_2O_5$, M_r 331.6.

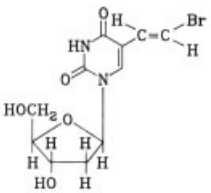

The antiherpetic activity of BVDU was first reported in 1979 [258]. This pyrimidine derivative is about 100 times more active against HSV – 1 in cell culture than against HSV – 2 and is also active against varicella – zoster virus.

The clinical efficacy of BVDU has been demonstrated in several uncontrolled (open) trials. First, because BVDU was well-absorbed when given orally, the progression of varicella – zoster virus in immunocompromised patients was retarded by BVDU [259]. Second, topical administration of BVDU was effective against herpetic keratitis [260]. Placebo-controlled, double-blinded clinical trials have been initiated.

The *mechanism of action* of BVDU is similar to that of other selective nucleoside analogues having antiherpetic activity. It is converted to the triphosphate by the viral and cellular thymidine kinases. The 5′-triphosphate is then incorporated into viral DNA in virally infected cells.

Adverse Effects. There are several problems with the use of BVDU. First, BVDU is rapidly degraded by thymidine phosphorylase. Second, herpesviruses can readily develop resistance to BVDU by lowering the virus content of thymidine kinase activity. Finally, the development of BVDU by G. D. Searle has been halted because of increased incidence of tumors in animals dosed for long periods [261].

Synthesis: Condensation of the trimethylsilyl derivative of (*E*)-5-(2-bromovinyl)uracil with 2′-deoxy-3,5-di-*O*-*p*-toluoyl-α-D-erythro-pentofuranosyl chloride gives a mixture of the α- and β-anomers of the protected nucleoside. The *p*-toluoyl protecting groups are removed by treatment with sodium methoxide in methanol. The β-anomer, but not the α-anomer, is biologically active [258].

FIAC [*69124-05-6*], 1-(2′-fluoro-2′-deoxy-β′-D-arabinofuranosyl)-5-iodocytosine hydrochloride, $C_9H_{11}FIN_3O_4 \cdot HCl$, M_r 407.58, *mp* 177 – 181 °C.

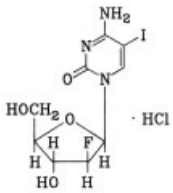

The synthesis and antiherpetic activity of a series of 2′-fluoro-5-substituted arabinofuranosylcytosines and uracils have been recently reported [262], [263]. Several compounds in this series are potent inhibitors of HSV – 1, HSV – 2, varicella – zoster virus, cytomegalovirus, and Epstein – Barr virus. For example, 1-(2′-fluoro-2′-deoxy-β-D-arabinofuranosyl)-5-iodocytosine (FIAC) was active systemically after i.v. administration to immunocompromised patients infected with varicella – zoster virus [264].

The mode of action of these 2′-fluoro-5-substituted nucleoside analogues is similar to that of acyclovir. The analogues are substrates for the herpesvirus-coded thymidine kinase (deoxypyrimidine kinase) [265]. Viral DNA polymerase is more sensitive to inhibition by the 5′-triphosphate of FIAC than is the host cell DNA polymerase [266].

A related compound, 1-(2′-fluoro-2′-deoxy-β′-D-arabinofuranosyl)-5-iodouracil (FIAU), has recently been reported to have some toxic effects in humans. FIAU caused cardiac fibrosis at a dose of 150 mg/kg in a long-term study [267].

Synthesis: FIAC is prepared by condensation of 3-*O*-acetyl-5-*O*-benzoyl-2′-deoxy-2′-fluoro-D-arabinofuranosyl bromide with trimethylsilylated cytosines to yield the blocked nucleosides. After deprotection by saponification to the 2′-fluoroarabinofuranosylcytosine nucleoside, the 5-iodo analogue (FIAC) is obtained by iodination [262].

Ribavirin [*36791-04-5*], Virazole (INN, British Approved Name – BAN), 1-β-D-ribofuranosyl-1*H*-1,2,4-triazole-3-carboxamide, $C_8H_{12}N_4O_5$, M_r 244.2, *mp* 166 – 168 °C.

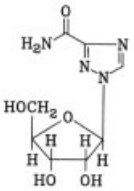

Ribavirin was initially described in 1972 as the result of a program to synthesize broad spectrum antiviral agents [268]. Ribavirin is active in cell culture against most DNA and RNA viruses [269].

Ribavirin probably has several *mechanisms of action.* Ribavirin is phosphorylated by cellular enzymes, and both the monophosphate (RMP) and the triphosphate (RTP) have antiviral activity. The monophosphate inhibits synthesis of guanosine 5′-monophosphate (GMP) [270]. Thus, the supply of guanosine triphosphate is depleted and nucleic acid synthesis is inhibited. In addition RTP inhibits the virus-specific mRNA capping enzymes of vaccinia virus [271]. The ability of RTP to inhibit the capping of viral-specific mRNA could explain how ribavirin could be active against both DNA and RNA viruses. An interesting finding is that ribavirin was active in cell culture against the virus which causes AIDS [272].

Clinical trials with ribavirin have been performed with respiratory syncytial virus (RSV), Lassa virus, and influenza viruses. Aerosolized ribavirin was effective in reducing fever and symptoms in adults experimentally infected with RSV [273]. Similar findings were observed in infants treated with aerosolized ribavirin after natural RSV infection [274]. Intravenous ribavirin reduced mortality in comatose patients infected with Lassa virus [275]. Conflicting results have been found with influenza virus and oral ribavirin. Several studies demonstrated decreased severity of illness and decreased viral shedding with oral ribavirin [276], [277]. Other studies demonstrated no therapeutic effect of oral ribavirin against influenza virus [278], [279]. However, aerosolized ribavirin has been shown to reduce fever and systemic illness in patients infected with influenza virus [280], [281].

Synthesis: Acid-catalyzed fusion of methyl 1,2,4-triazole-3-carboxylate and 1,2,3,5-tetra-*O*-acetyl-β-D-ribofuranose or 1-*O*-acetyl-2,3,5-tri-*O*-benzoyl-β-D-ribofuranose yields blocked methyl ester nucleosides. Ribavirin is obtained after treatment of the blocked nucleoside with methanolic NH_3 [268].

Trade name: Virazole (ICN).

Cyclaradine [*69979-46-0*], carbocyclic arabinofuranosyladenine, [±]-9-[2α,3β-dihydroxy-4α-(hydroxymethyl)cyclopent-1α-yl]adenine, $C_{11}H_{15}N_5O_3 \cdot H_2O$, M_r 283.33, *mp* 253 – 255 °C.

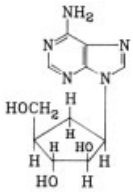

Cyclaradine is a carbocyclic derivative of ara-A. Cyclaradine was synthesized in 1977 in an effort to obtain a derivative of ara-A that is resistant to adenosine deaminase [282]. The deamination product of ara-A is 9-β-D-arabinofuranosylhypoxanthine, which is considerably less active than ara-A against herpesviruses. Cyclaradine is resistant to adenosine deaminase.

Cyclaradine is active in cell culture against herpes simplex virus type 1 [282]. Clinical studies in humans have not yet been conducted with cyclaradine. However, cyclaradine is effective against HSV – 1 induced encephalitis in mice [283] and HSV – 2-induced genital infections in guinea pigs [284].

Synthesis: Reaction of 5-amino-4-*N*-[2α,3β-dihydroxy-4α-(hydroxymethyl)cyclopent-1α-yl]-amino-6-chloropyrimidine with diethoxymethyl acetate gives the 6-chloropurine. Treatment of the 6-chloropurine with liquid ammonia gives cyclaradine [282].

DHPG [*82410-32-0*], 2-amino-1,9-dihydro{[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl}-6*H*-purin-6-one, BIOLF – 62 [285], BW B759U, 2′-nor-2′-deoxyguanosine (2′-NDG) [286], $C_9H_{13}N_5O_4$, M_r 255.3, *mp* > 300 °C.

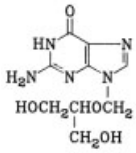

DHPG is an analogue of acyclovir; its potency is equal to or greater than that of acyclovir against some herpesviruses. In comparison with acyclovir, DHPG is at least 10-fold more active against CMV and Epstein – Barr virus in cell culture and about 50-fold more active than acyclovir in animal models of herpesvirus infection [287].

Several preliminary clinical studies with DHPG have been reported. It was effective against CMV pneumonia in transplant recipients [288] and against CMV retinitis [289]. DHPG has also been used for CMV infections in AIDS patients [290].

Synthesis: 2-*O*-(Acetoxymethyl)-1,3-di-*O*-benzylglycerol is condensed with N^2,9-diacetylguanine in the presence of a catalytic amount of *p*-toluenesulfonic acid in sulfolane. The desired isomer is crystallized from toluene, debenzylated over 20 % palladium hydroxide, and deacetylated with concentrated NH_4OH/methanol [291].

Buciclovir [*86304-28-1*], (*R*)-9-(3,4-dihydroxybutyl)guanine, (*R*)-DHBG, (*R*)-2-amino-9-(3,4-dihydroxybutyl)-1,9-dihydro-6*H*-purin-6-one, $C_9H_{13}N_5O_3$, M_r 239.23, *mp* 260 – 261 °C (decomp.).

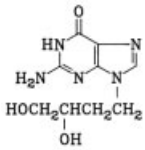

(*RS*)-9-(3,4-Dihydroxybutyl)guanine, which is a close analogue of acyclovir and DHPG, has recently been shown to inhibit HSV – 1 and HSV – 2 both in cell culture and in animal models of virus infection [292]. The (R)-enantiomer of DHBG is about 5-fold more active against HSV – 2 in plaque reduction tests than the (S)-enantiomer. The mode of action of DHBG is similar to that of acyclovir [293], [294]. DHPG is phosphorylated by the HSV-induced thymidine kinase; presumably it is phosphorylated by cellular enzymes to give the 5′-triphosphate derivative of DHPG, which selectively inhibits viral DNA synthesis.

No clinical data on DHBG are available at the present time.

Synthesis: Condensation of 2-amino-6-chloropurine with 4-bromo-2-hydroxybutyrate and subsequent hydrolysis yields 4-(9-guanyl)-2-hydroxybutyric acid, which is esterified and reduced to buciclovir [295].

4.6.2. Phosphonoacetate and Phosphonoformate

Phosphonoacetic acid (PAA) [*4408-78-0*], $C_2H_5O_5P$, M_r 136.01, was discovered by routine screening methods in 1973 to inhibit HSV

[296]. The sodium salt is called phosphonoacetate. PAA has been shown to inhibit the replication in cell culture of HSV, cytomegalovirus, Epstein–Barr virus, varicella–zoster virus, vaccinia virus, avian herpesvirus, and equine abortion virus. PAA inhibits the replication of herpesvirus by interacting directly with the virus-induced DNA polymerase at the pyrophosphate binding site [297]. Host cell polymerases are less sensitive to PAA [298]. Preclinical studies in rats showed that intravenous PAA was deposited in bone [299]. This result prevented clinical studies from being performed with PAA. Structure–activity studies have been reported for 100 analogues of PAA, but all analogues were less active than PAA [300].

Phosphonoformate [*63585-09-1*], foscarnet sodium, phosphonoformate sodium, PFA (INN), phosphonoformic acid (trisodium salt), foscarnet, CNa_3O_5P, M_r 191.95, *mp* >250 °C.

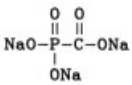

Phosphonoformate is closely related to PAA and also directly inhibits the virus-induced DNA polymerase [301]. Phosphonoformate is deposited in bone, but to a lesser extent than PAA [302]. Clinical trials have indicated that topical treatment of patients with recurrent genital herpesvirus infection was effective in shortening the time to healing, but this effect was only observed in male patients [303]. Intravenous phosphonoformate was effective against CMV in allograft recipients [304].

Phosphonoformate has also shown activity in cell culture against HTLV–III (the virus implicated in AIDS) [305].

Synthesis: Phosphonoformate is prepared by saponification of the triethyl ester of phosphonoformate with NaOH. Trisodium phosphonoformate hexahydrate is recrystallized from H_2O [301].

Trade name: Foscarnet (Astra).

4.6.3. Amantadine and Rimantadine

Amantadine is presently the only chemotherapeutic agent accepted for clinical use against influenza virus. Amantadine was discovered in 1964 and was found to be active against influenza A viruses [306]. Administered orally, subcutaneously, or intraperitoneally, amantadine reduced the mortality in mice infected intranasally with influenza A virus. The precise *mechanism of action* of amantadine is not defined although most studies indicate that amantadine inhibits uncoating (the release of infectious nucleic acid into the cell's cytoplasm) [307]. The first clinical trials in volunteers lacking antibody showed that prophylactic amantadine reduced the incidence of influenza [308]. These results were confirmed in a natural outbreak of influenza virus [309]. The use of amantadine did not become widespread because until September, 1976; the FDA would not approve its use against any strain of influenza except for the original A2/Asian strain. Thus, amantadine could not be used against the major Hong Kong influenza virus epidemic in the late 1960s.

In comparison studies with amantadine, rimantadine is equally effective prophylactically with fewer *adverse effects* (nervousness, difficulty in concentrating, and insomnia) [310]. Rimantadine has similar therapeutic efficacy as amantadine with respect to faster resolution of symptoms in patients with influenza [311]. A randomized, placebo-controlled, double-blind trial demonstrated the therapeutic efficacy of oral rimantadine given once daily to patients with influenza A virus infection [312].

Amantadine (INN) [*665-66-7*], 1-adamantanamine hydrochloride, $C_{10}H_{17}N \cdot HCl$, M_r 187.74, *mp* 206–208 °C.

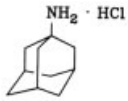

Synthesis: 1-Bromoadamantane is reacted with acetonitrile in the presence of concentrated sulfuric acid. The resulting 1-acetaminoadamantane is hydrolyzed with NaOH in boiling diethylene glycol.

Trade name: Symmetrel (Du Pont).

Rimantadine (INN) [*13392-28-4*], [*1501-84-4*], α-methyl-tricyclo[3.3.1.1^{3,7}]decane-1-methan-amine, $C_{12}H_{21}N$, M_r 179.31, *mp* 373–375 °C, an analogue of amantadine,

was also described in 1964 [306]; HCl salt [*1501-84-4*].

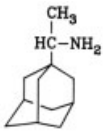

4.6.4. Enviroxime

Enviroxime (INN) [*72301-79-2*], (*E*)-2-amino-6-benzoyl-1-(isopropylsulfonyl)benzimidazoleoxime, 6-{[(hydroxyimino)phenyl]methyl}-1-[(1-methylethyl)sulfonyl]-1*H* -benzimidazol-2-amine, $C_{17}H_{18}N_4O_3S$, anti-isomer, M_r 358.45, *mp* 198 – 199 °C.

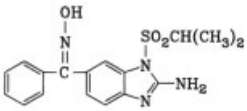

Enviroxime is an extremely potent inhibitor of human rhinoviruses in cell culture. A total of 15 different rhinovirus serotypes were inhibited in plaque reduction assays by enviroxime concentrations ranging from 0.04 µg/mL to < 0.01 µg/mL [313], [314]. However, the results of five clinical trials showed that enviroxime had only a modest, clinically significant, beneficial effect on rhinovirus infections. Prophylactic intranasal enviroxime administered prior to virus challenge in volunteers resulted in a statistically significant reduction in rhinorrhea, with no significant difference detected in infection rate or in the quantity of secreted virus. In addition, oral administration of enviroxime caused nausea and vomiting [315–317]. Therapeutically, intranasal administration of enviroxime lessened symptoms, but only on the fifth (last) day of treatment [318]. An open-field trial of intranasally applied enviroxime against natural rhinovirus infections showed no statistically significant differences compared to placebo [319]. Eli Lilly has terminated all studies on enviroxime because of the marginal efficacy observed in the clinic.

Synthesis: Reaction of cyanogen bromide and 3,4-diaminobenzophenone gives 2-amino-1-(isopropylsulfonyl)-6-benzimidazolyl phenylketone. The benzimidazole is sulfonated by using isopropylsulfonyl chloride and sodium hydride in dimethoxyethane. The anti- and syn-isomers of enviroximes are separated by high-performance liquid chromatography or by fractional crystallization [313].

4.6.5. 4′,6-Dichloroflavan

4′,6-Dichloroflavan [*73110-56-2*], BW-683C, 6-chloro-2-(4-chlorophenyl)-3,4-dihydro-4′,6-di-chloroflavan, $C_{15}H_{12}Cl_2O$, M_r 279.18.

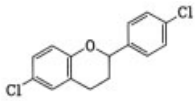

Dichloroflavan, like enviroxime, is a potent inhibitor of the replication of human rhinoviruses in cell culture [320]. Dichloroflavan appears to bind to virion capsid proteins, thereby inhibiting uncoating of the viral RNA [321], [322]. Only one clinical trial with oral dichloroflavan has been reported. In this double-blind, placebo-controlled trial, orally administered dichloroflavan was given to volunteers before and after challenge with rhinovirus 9. Dichloroflavan had no effect on the incidence or course of rhinovirus infection in this study [323].

Chemical synthetic details are not available.

4.6.6. Chalcone Ro 09 – 0410

Chalcone Ro 09 – 0410 [*76554-66-0*], 1-(4-ethoxy-2-hydroxy-6-methoxyphenyl)-3-(4-methoxyphenyl)-2-propen-1-one, 4′-ethoxy-2′-hydroxy-4,6′-dimethoxychalcone, Ro 09 – 0410, $C_{19}H_{20}O_5$, M_r 328.37.

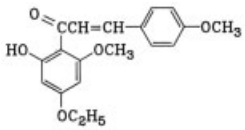

Chalcone Ro 09 – 0410 is a potent inhibitor of human rhinovirus in cell culture. Only 0.002 µg of Ro 09 – 0410 per mL is sufficient to inhibit some serotypes of human rhinovirus. Other serotypes of human rhinovirus require as much as 2 µg of the chalcone per mL for inhibition to

occur [324]. The mode of action of Ro 09 – 0410 is similar to dichloroflavan in that it prevents uncoating by irreversible binding to viral capsid proteins. The chalcone could, however, be removed from virus particles by extraction with a nonpolar solvent [325].

A phosphate ester prodrug of the chalcone has been tested in volunteers because the chalcone itself was not well-absorbed orally and was irritating intranasally. The prodrug did not have any effect on the course of the rhinovirus infection, probably because even though the prodrug was well-absorbed, the chalcone did not reach the nasal mucosa [326].

Chemical synthetic details are not available.

4.6.7. Arildone and Disoxaril

Arildone [327] and disoxaril [328] have been shown to be active in cell culture, as well as in experimental animals infected with human picornaviruses [329–331]. *Mechanism of action* studies indicated that arildone blocked poliovirus uncoating, but did not inhibit adsorption or penetration [332]. Further work has shown that both arildone and disoxaril interact directly with picornavirion capsid proteins to prevent uncoating of virions.

Arildone (INN) [*56219-57-9*], 4-[6-(2-chloro-4-methoxy)phenoxy]hexyl-3,5-heptanedione, $C_{20}H_{29}ClO_4$, M_r 368.94, *bp* 180 °C.

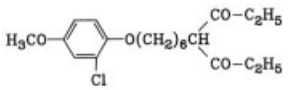

Synthesis: Treatment of 1-(2-chloro-4-methoxyphenoxy)-6-iodohexane and lithio-3,5-heptanedione in dimethylformamide for 48 h at 60 °C gives arildone [327].

Disoxaril (INN), Win 51711 [*87495-31-6*], 5-{7-[4-(4,5-dihydro-2-oxazolyl)phenoxy]heptyl}-3-methylisoxazole, $C_{20}H_{26}N_2O_3$, M_r 342.42, *mp* 86 – 89 °C.

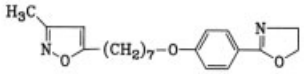

4.6.8. 3′-Azidothymidine

3′-Azidothymidine [*30516-87-1*], AZT, 3′-azido-3′-deoxythymidine, $C_{10}H_{13}N_5O_4$, M_r 267.3, *mp* 119 – 121 °C.

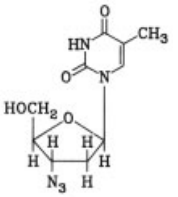

3′-Azidothymidine was initially synthesized in 1979 [333]. This compound is active against the viruses responsible for AIDS (HTLV – III); it blocks the cytopathic effects of HTLV – III in cell culture [334]. 3′-Azidothymidine entered phase II clinical trials against AIDS in January, 1986 [335]; it has shown some positive results in preventing mortality because of the AIDS virus in patients treated with AZT for a few months.

Synthesis: prepared by detritylation of the 3′-azidoderivative of 1-(2′-deoxy-5′-*O*-trityl-*β*-D-lyxosyl)thymine [333], [336].

4.6.9. Suramin

Suramin (INN, U.S.P. **12,** DCF) [*145-63-1*], 8,8′{carbonylbis[imino-3,1-phenylenecarbonyl-imino(4-methyl-3,1-phenylene)-carbonylimino]}-bis(1,3,5-naphthalene)trisulfonic acid; suramin sodium (hexasodium salt) [*129-46-4*], $C_{51}H_{34}N_6Na_6O_{23}S_6$, M_r 1429.15.

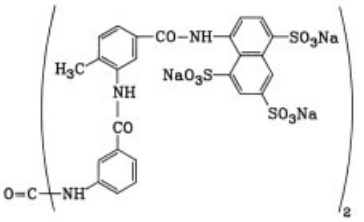

Suramin sodium is the hexasodium salt of 8,8′-{carbonylbis[imino-3,1-phenylenecarbonyl-imino(4-methyl-3,1-phenylene)carbonylimino]}-bis(1,3,5-naphthalene)trisulfonic acid. It is the drug of choice for the therapy of early African trypanosomiasis

[337], and has been the subject of renewed interest as an inhibitor of the reverse transcriptase of the virus that causes AIDS [338], [339]. Preliminary findings in a few AIDS patients showed that suramin sodium was beneficial after short-term intravenous treatment [340].

Synthesis: Suramin sodium is prepared by condensing 1-naphthylamine-4,6,8-trisulfonic acid with 3-nitro-4-methylbenzoyl chloride, reducing the product, condensing with 3-nitrobenzyl chloride, reducing again, and then treating with carbonyl chloride and neutralizing with sodium hydroxide [341].

4.6.10. HPA 23

HPA 23 [*89899-81-0*], ammonium 21-tungsto-9-antimonate, 5′-tungsto-2-antimonate, $(NH_4)_{17}Na/NaW_{21}Sb_9O_{86} \cdot 14\,H_2O$, M_r 6937.39.

HPA 23 is a tungsto-antimonate compound that is a competitive inhibitor of the reverse transcriptase of murine oncornavirus [342] and of the AIDS viruses. In a preliminary uncontrolled clinical trial, HPA 23 was able to inhibit AIDS virus replication in patients. The virus, however, reappeared when therapy was discontinued [343].

Synthesis: Addition of a hot hydrochloric acid solution of antimony trioxide to an aqueous solution of sodium tungstate and NH_3 gives HPA 23 [344].

5. References

1. R. E. Buchanan, N. E. Gibbons, (eds.): *Bergey's Manual of Determinative Bacteriology,* 8th ed., The Williams & Wilkins Co., Baltimore 1974.
2. E. F. Gale, E. Cundliffe, P. E. Reynolds, M. H. Richmond, M. J. Waring in *The Molecular Basis of Antibiotic Actions,* 2nd ed., J. Wiley & Sons, London 1972.
3. G. Y. Lesher, E. J. Froelich, M. D. Gruett, J. H. Bailey et al., *J. Med. Pharm. Chem.* **5** (1962) 1063.
4. R. Albrecht, *Prog. Drug Res.* **21** (1977) 9.
5. J. S. Wolfson, D. C. Hooper, *Antimicrob. Agents Chemother.* **28** (1985) 581 – 585.
6. M. P. Wentland, J. B. Cornett in D. M. Bailey (ed.): *Annual Reports in Medicinal Chemistry,* **vol. 20,** Academic Press, New York 1985, pp. 145 – 153.
7. D. C. Hooper, J. S. Wolfson, *Antimicrob. Agents Chemother.* **29** (1986) in press.
8. H. Koga, A. Itoh, S. Murayama, S. Suzae, T. Irikura, *J. Med. Chem.* **23** (1980) 1358 – 1364.
9. M. Gellert, K. Mizuuchi, M. H. O'Dea, H. A. Nash, *Proc. Natl. Acad. Sci.* **73** (1976) 3872 – 3878.
10. A. Sugino, C. L. Peebles, K. N. Kruezer, N. R. Cozzar, *Proc. Natl. Acad. Sci.* **47** (1977) 4767 – 4771.
11. A. M. Barlow, *Br. Med. J.* **2** (1963) 1308.
12. T. A. McAllister, A. Percival, J. G. Alexander, J. G. Boyce et al., *Postgrad. Med. J.* **47** (1971, Sept.) Suppl., 7.
13. E. W. McChesney, E. J. Froelich, G. Y. Lesher et al., *Toxicol. Appl. Pharmacol.* **6** (1964) 292.
14. T. A. Stamey, N. J. Nemoy, M. Higgins, *Invest. Urol.* **6** (1969) 582.
15. G. A. Portmann, E. W. McChesney, H. Stander et al., *J. Pharm. Sci.* **55** (1966) 72.
16. M. Buchbinder, J. C. Webb, L. Anderson, W. R. McCabe, *Antimicrob. Agents Chemother.* 1962, 308.
17. Sterling Drug, US 3 149 104, 1964 (G. Y. Lesher, M. D. Gruett).
18. F. J. Turner, S. M. Ringel, J. F. Martin et al., *Antimicrob. Agents Chemother.* 1967, 475.
19. S. M. Ringel, F. J. Turner, F. L. Lindo et al., *Antimicrob. Agents Chemother.* 1967, 480.
20. D. Kaminsky, R. I. Meltzer, *J. Med. Chem.* **11**(1968) 160.
21. W. E. Wick, D. S. Preston, W. A. White et al., *Antimicrob. Agents Chemother.* **4** (1973) 415.
22. R. G. Gordon, L. I. Stevens, C. E. Edmiston et al., *Antimicrob. Agents Chemother.* **10**(1976) 918.
23. W. A. Goss, W. H. Dietz, T. M. Book, *J. Bacteriol.* **10** (1965) 918.
24. R. A. P. Burt, T. Morgan, J. P. Payne et al., *Br. J. Urol.* **49** (1977) 147.
25. H. R. Black, K. S. Israel, R. L. Wolen et al., *Antimicrob. Agents Chemother.* **15** (1979) 165.
26. Eli Lilly, DE 2 065 719, 1975 (W. A. White).
27. J. R. Prous (ed.), *Ann. Drug Data Rep.* **2** (1979/1980)177 – 180.
28. J. R. Prous (ed.), *Ann. Drug Data Rep.* **2** (1979/1980)181 – 184.
29. J. Matsumoto, S. Minami, *J. Med. Chem.* **18** (1975) 74.
30. Dainippon, JP 25 912, 1967 (Minami et al.).

31. J. R. O'Connor, R. A. Dobson, P. E. Came, R. B. Wagner in J. D. Nelson, C. Grassi (eds.): "Current Chem. and Infect. Dis.," *Proc. 11th ICC and 19th ICAAC,* vol. 1, The American Society for Microbiology, Washington, D.C., 1980, pp. 440 – 442.
32. Sterling Drug, US 3 753 993, 1973 (G. Y. Lesher, P. M. Carabateas).
33. A. Izawa, Y. Kisaki, K. Irie, Y. Eda et al., *Antimicrob. Agents Chemother.* **18** (1980) 37 – 40.
34. A. Izawa, A. Yoshitake, T. Komatsu, *Antimicrob. Agents Chemother.* **18** (1980) 41 – 44.
35. A. Yoshitake, K. Kawahara, F. Shono, I. Umeda et al., *Antimicrob. Agents Chemother.* **18** (1980) 45 – 49.
36. Sumitomo, US 3 799 930, 1974 (T. Nakagome et al.).
37. K. Sato, Y. Matsuura, M. Inoue, T. Une et al., *Antimicrob. Agents Chemother.* **22** (1982) 548 – 553.
38. N. X. Chin, H. C. Neu, *Antimicrob. Agents Chemother.* **25** (1984) 319 – 326.
39. D. S. Reeves, M. J. Bywater, H. A. Holt, L. O. White, *J. Antimicrob. Chemother.* **13** (1984) 333 – 346.
40. J. B. Cornett, R. B. Wagner, R. A. Dobson, M. P. Wentland et al., *Antimicrob. Agents Chemother.* **27** (1985) 4 – 10.
41. N. X. Chin, H. C. Neu, *Antimicrob. Agents Chemother.* **24** (1983) 754 – 763.
42. A. L. Barry, R. N. Jones, *Antimicrob. Agents Chemother.* **25** (1984) 775 – 777.
43. A. Ito, K. Hirai, M. Inoue et al., *Antimicrob. Agents Chemother.* **17** (1980) 103.
44. H. C. Neu, P. Labthavikul, *Antimicrob. Agents Chemother.* **22** (1982) 23.
45. S. R. Crider, S. D. Colby, L. K. Miller, W. O. Harrison et al., *N. Eng. J. Med.* **311** (1984) 137.
46. P. Cantet, H. Ranaudin, C. Quentin, C. Babear, *Pathol. Biol. (Paris)* **31** (1983)501.
47. S. Nakamura, N. Kurobe, S. Kashimoto, T. Ohue et al., *Antimicrob. Agents Chemother.* **24** (1983) 54.
48. Z. N. Adhami, R. Wise, D. Wetson, B. Crump, *Antimicrob. Chemother.* **13** (1984) 87.
49. R. Wise, R. Lockley, J. Dent, M. Webberly, *Antimicrob. Agents Chemother.* **26** (1984) 17.
50. B. Crump, R. Wise, J. Dent, *Antimicrob. Agents Chemother.* **24** (1983) 784.
51. M. Bologna, L. Vaggii, C. M. Forchetti, E. Martine, *Lancet* **2** (1983) 280.
52. B. N. Swanson, V. K. Boppana, P. H. Vlasses, H. M. Rotmensch et al., *Antimicrob. Agents Chemother.* **23** (1983) 284.
53. G. Panichi, A. Pantosti, G. P. Testore, *J. Antimicrob. Chemother.* **11** (1983) 589.
54. Kyorin, BE 870 917, 1979.
55. R. Wise, J. M. Andrew, L. J. Edwards, *Antimicrob. Agents Chemother.* **23** (1983)559.
56. H. L. Muytjens, J. van der Ros-van de Repe, G. van Veldhuizen, *Antimicrob. Agents Chemother.* **24** (1983) 302.
57. R. J. Fass, *Antimicrob. Agents Chemother.* **24** (1983) 568.
58. W. Brumfitt, I. Franklin, D. Grady, J. M. T. Hamilton-Miller et al., *Antimicrob. Agents Chemother.* **26** (1984) 757.
59. Bayer, EP-A 78 362, 1983 (K. Grohe, H. J. Zeiler, K. G. Metzger).
60. S. Nakamura, K. Nakata, H. Katae, A. Minami et al., *Antimicrob. Agents Chemother.* **23** (1983) 742.
61. A. Notowicz, E. Stolz, B. van Klingeren, *J. Antimicrob. Chemother.* **14** (1984) Suppl. C, 91.
62. J. M. Hubrechts, R. Vanhoof, J. Servais, R. Toen et al., *Lancet* **1** (1984) 860.
63. Dainippon, EP-A 9 425, 1980 (J. I. Matsumoto, Y. Takase, Y. Nishimura).
64. D. L. van Caekenberghe, S. R. Pattyn, *Antimicrob. Agents Chemother.* **25** (1984) 518.
65. G. Montay, Y. Gouffon, R. Roquet, *Antimicrob. Agents Chemother.* **25** (1984) 463.
66. M. Wolff, B. Regnier, C. Daldoss, M. Nkam et al., *Antimicrob. Agents Chemother.* **26** (1984) 289.
67. N. Desplaces, L. Gutmann, J. F. Acar, *24th ICAAC,* Washington, D.C., 1984, article no. 279.
68. F. B. McGillion, *Drugs of the Future* **7** (1982) 946.
69. M. R. Lockley, R. Wise, J. Dent, *Antimicrob. Agents Chemother.* **14** (1984) 647.
70. Daiichi Seiyaku, EP-A 47 005, 1982 (I. Hayakawa, T. D. S. Hiramitsu, Y. Tanaka).
71. N. V. Jacobus, F. P. Tally, M. Barza, *Antimicrob. Agents Chemother.* **26** (1984) 104 – 107.
72. I. Garcia, G. P. Bodey, V. Fainstein, D. Hsitto et al., *Antimicrob. Agents Chemother.* **26** (1984) 421 – 423.
73. H. C. Neu, P. Labthavikul, *24th ICAAC,* Washington, D.C., 1984, article no. 402.
74. R. B. Wagner, R. A. Dobson, M. P. Wentland, D.M. Bailey et al., *23rd ICAAC,* Las Vegas, Nev., 1983, article no. 378.

75. Sterling Drug, EP-A 90 424, 1983 (M. P. Wentland, D. M. Bailey).
76. P. B. Fernandes, D. Chu, R. Bowen, N. Shipkowitz et al., *24th ICAAC,* Washington, D.C., 1984, article no. 78.
77. P. B. Fernandes, N. Shipkowitz, D. Chu, L. Doen et al., *24th ICAAC,* Washington, D.C., 1984, article no. 79.
78. N. X. Chin, D. C. Brittain, H. C. Neu, *24th ICAAC,* Washington, D.C., 1984, article no. 399.
79. K. Takagi, M. Hosaka, H. Kusajima, Y. Oomori et al., *23rd ICAAC,* Las Vegas, Nev., 1983, article no. 659.
80. M. A. Cohen, P. A. Bien, T. J. Griffin, C. L. Heifetz, *24th ICAAC,* Washington, D.C., 1984, article no. 81.
81. J. C. Sesnie, M. A. Shapiro, C. L. Heifetz, T. F. Mich, *24th ICAAC,* Washington, D.C., 1984, article no. 82.
82. J. M. Domagala, J. B. Nichols, C. L. Heifetz, T. F. Mich, *24th ICAAC,* Washington, D.C., 1984, article no. 80.
83. Riker Lab, US 3 896 131, 1975 (J. F. Gerster).
84. D. T. W. Chu, G. R. Granneman, P. B. Fernandes, *Drugs of the Future* **10** (1985) 546.
85. Warner Lambert, AU 8 318 698, 1984 (T. P. Culbertson, J. M. Domatgala, T. F. Michh, J. B. Nickols).
86. J. Trefouël, J. Trefouël, F. Nitti, D. Bovet, *C. R. Seances Soc. Biol. Ses Fil.* **120**(1935) 756.
87. D. D. Woods, *Br. J. Exp. Pathol.* **21** (1940)74.
88. P. Fildes, *Lancet* **1** (1940) 955.
89. A. K. Miller, P. Bruno, R. M. Berglund, *J. Bacteriol.* **54** (1947) 9.
90. O. Sköld, *Antimicrob. Agents Chemother.* **9** (1976)49.
91. S. A. Kabins, M. V. Panse, S. Cohen, *J. Infect. Dis.* **123** (1971) 158.
92. R. L. Then, *Rev. Infect. Dis.* **4** (1982) 261.
93. J. M. J. Hamilton-Miller, *J. Antimicrob. Chemother.* **5 Suppl. B** (1979) 61.
94. J. K. Seydel, *J. Pharm. Sci.* **57** (1968) 1455.
95. N. Anand: "Sulfonamides and Sulfones," in M. E. Wolff (ed.): *Burger's Medicinal Chemistry,* 4th ed., J. Wiley & Sons, New York 1979, Part II, Chap. 13.
96. T. Fujita, C. Hanch, *J. Med. Chem.* **10** (1967) 991.
97. H. Kano, US 2 888 455, 1959.
98. B. Roth et al., *J. Med. Chem.* **23** (1980) 535.
99. S. R. M. Bushby, *J. Infect. Dis.* **128 Suppl.** (1973) S443.
100. E. Grunberg, *J. Infect. Dis.* **128 Suppl.** (1973) S478.
101. D. J. Winston, W. K. Lau, R. P. Gale, L. S. Young, *Ann. Intern. Med.* **92** (1980) 762.
102. J. Wüst, *Antimicrob. Agents Chemother.* **11** (1977) 631.
103. R. O. Roblin, *J. Am. Chem. Soc.* **62** (1940) 2002.
104. L. Doub, *J. Med. Chem.* **13** (1970) 242.
105. H. Iundbeck & Co., US 2 447 702, 1948 (O. Hübner).
106. H. Bretschneider, US 3 132 139, 1962.
107. Aktiebolaget Pharmacia, US 2 396 145, 1946 (A. Askëlof).
108. ICI, GB 551 513, 1943 (E. Haworth).
109. M. L. Moore, *J. Am. Chem. Soc.* **64** (1942) 1572.
110. Sharpe & Dohme, US 2 324 013 – 15, 1943 (M. L. Moore).
111. R. M. Des Prez, R. A. Goodwin, Jr. in G. L. Mandell, R. G. Douglas, Jr., J. E. Bennett (eds.): *Principles and Practice of Infectious Diseases,* 2nd ed., J. Wiley & Sons, New York 1985, p. 1383.
112. W. E. Bullock, [111], p. 1406.
113. W. E. Sanders, Jr., [111], p. 1413.
114. *Med. Lett.* **28** (1986, Mar. 28) 710.
115. Tuberculosis Chemotherapy Trials Committee: Interim Report to the Medical Research Council, "The Treatment of Pulmonary Tuberculosis with Isoniazid," *Br. Med. J.* **2** (1952) 735.
116. W. B. Pratt in *Chemotherapy of Infection,* Oxford University Press, New York 1977, p. 231.
117. D. E. Kopanoff, D. E. Spider, G. J. Caras, *Am. Rev. Respir. Dis.* **117** (1978) 991.
118. American Thoracic Society, *Am. Rev. Respir. Dis.* **110** (1974) 374.
119. Distillers, US 2 280 994, 1958.
120. R. G. Wilkinson et al., *J. Am. Chem. Soc.* **83** (1961) 2212.
121. W. W. Stead, A. K. Dutt, *Am. Rev. Respir. Dis.* **125 Suppl.** (1982) no. 3, 94.
122. D. J. Girling, *Tubercule* **59** (1978) 13.
123. A. K. Dutt, W. W. Stead, *J. Infect. Dis.* **146** (1982) 698.
124. Merck, US 2 149 279, 1939 (O. Dalmer).
125. D. Lieberman, *Bull. Soc. Chim. Fr.* (1958) 687.
126. J. T. Sheehan, *J. Am. Chem. Soc.* **70** (1948) 1665.
127. G. L. Mandell, M. A. Sande in A. G. Gilman, L. S. Goodman, A. Gilman (eds.): *Drugs Used in the Chemotherapy of Tuberculosis and Leprosy, The Pharmacological Basis of Therapeutics,* 6th ed., Macmillan Publ. Co., New York 1980, p. 1200.

128. G. Domagk et al., *Naturwissenschaften* **33** (1946) 315.
129. C. C. Shepard, L. Levy, P. Fasal, *Am. J. Trop. Med. Hyg.* **18** (1969) 258.
130. R. W. Riley, L. Levy, *Proc. Soc. Exp. Biol. Med.* **142** (1973) 1168.
131. I.G. Farbenind, FR 829 926, 1938.
132. H. Bauer, *J. Am. Chem. Soc.* **61** (1939) 617.
133. J. H. Peters, J. F. Murray, G. R. Gordon et al., *Am. J. Trop. Med. Hyg.* **26** (1977) 127.
134. D. A. Russell, C. C. Shepard, D. H. McRae et al., *Am. J. Trop. Med. Hyg.* **24** (1975) 485.
135. E. Eslarger et al., *J. Med. Chem.* **12** (1969) 357
136. L. Levy, *Am. J. Trop. Med. Hyg.* **23** (1974) 1097.
137. J. Convit, S. G. Browne, J. Languillon et al., *WHO Bull.* **42** (1970) 667.
138. C. C. Shepard, L. L. Walker, R. M. Van Lindingham et al., *Proc. Soc. Exp. Biol. Med.* **137** (1971) 725.
139. C. C. Shepard, L. L. Walker, R. M. Van Lindingham et al., *Proc. Soc. Exp. Biol. Med.* **137** (1971) 728.
140. V. C. Barry, *J. Chem. Soc.* (1958) 859.
141. V. L. Sutter in S. M. Finegold, W. L. George, R. D. Rolfe (eds.): *First United States Metronidazole Conference, Proc. Symp.,* Tarpon Springs, Fla., Biomedical Information Corp., New York 1982, p. 61.
142. Hoffmann La Roche, US 2 430 094, 1947 (H. M. Wuest).
143. S. M. Finegold, [111], p. 220.
144. R. M. Jacob, Rhône-Poulenc, US 2 944 061, 1960.
145. M. Turck, A. R. Ronald, R. G. Petersdorf, *Antimicrob. Agents Chemother.* 1966, 446.
146. A. Kucers, N. M. Bennett in A. Kucers, N. M. Bennett (eds.): *The Use of Antibiotics,* J. B. Lippincott Co., Philadelphia 1979, p. 749.
147. H. K. Reckendorf, R. G. Castringius, H. K. Spingler, *Antimicrob. Agents Chemother.* 1962, 531.
148. Eaton Labs., US 2 610 181, 1952 (K. J. Hayes).
149. R. A. Marcial-Rojas (ed.): *Pathology of Protozoal and Helminth Diseases, With Clinical Correlation,* Williams & Wilkins, New York 1971.
150. H. Van de Bossche, *Nature London* **273** (1978) 626.
151. T. C. Jones, [111], pp. 1505, 1513, 1560.
152. D. J. Wyler, [111], p. 1514.
153. D. R. Hill, [111], p. 1552.
154. J. J. Marr, [111], p. 286.
155. [111], p. 1537.
156. J. I. Ravdin, T. C. Jones, [111], p. 1506.
157. W. T. Hughes, [111], p. 1549.
158. L. V. Kirchoff, F. Neva, [111], p. 1531.
159. M. R. Rein, [111], p. 1556.
160. T. K. Ruebush, II, [111], p. 1559.
161. H. W. Brown, *Basic Clinical Parasitology,* 4th ed., Appleton-Century-Crofts, New York 1976.
162. R. D. Pearson, A. de Queiroz Sousa, [111], p. 1522.
163. R. E. McCabe, J. E. Remington, [111], p. 1540.
164. W. E. Gutteridge, *Br. Med. Bull.* **41** (1985) 162.
165. A. Sjoerdsma, J. A. Golden, P. J. Schechter et al., *Clin. Res.* **32** (1984) 559 A.
166. J. Williamson, *Exp. Parasitol.* **12** (1962) 274.
167. J. Williamson in H. W. Mulligan (ed.): *The African Trypanosomiases,* Allen & Unwin, London 1970, p. 125.
168. J. Williamson, *Trop. Dis. Bull.* **73** (1976) 531.
169. F. Hawking in R. J. Schnitzer, F. Hawking (eds.): *Experimental Chemotherapy,* **vol. 1,** Academic Press, New York 1963, p. 129.
170. B. B. Waddy, [167], p. 711.
171. F. I. C. Apted, *Pharmacol. Ther.* **11** (1980) 391.
172. D. A. Evans, *Antibiot. Chemother.* **30** (1981)272.
173. May & Baker, US 2 410 796, 1946 (G. Neuberry).
174. E. Friedheim, *Chem. Abstr.* **47** (1953) 144.
175. W. A. Jacobs, *J. Am. Chem. Soc.* (1919) 1590.
176. Hoechst, US 2 838 485, 1958 (R. Brodersten).
177. Z. Brenner, *Adv. Pharmacol. Chemother.* **13** (1975)1.
178. Z. Brenner, *Pharmacol. Ther.* **7** (1979) 71.
179. W. E. Gutteridge, *Trop. Dis. Bull.* **73** (1976) 699.
180. W. E. Gutteridge in J. R. Baker (ed.): "Perspectives in Trypanosomiasis Research," *Proceedings of the Twenty-first Trypanosomiasis Seminar,* Research Studies Press, Cinchester 1982, p. 47.
181. R. Ribeiro-dos-Santos, A. Rassi, F. Koberle, *Antibiot. Chemother.* **30** (1980) 115.
182. Bayer, US 3 262 930, 1964 (H. Herlinger).
183. Hoffmann-La Roche, *Chem. Abstr.* **71** (1969) 3383.
184. P. H. Rees, P. A. Kager, T. Ogada, J. K. M. Eeftinck Schattenkerk, *Trop. Geogr. Med.* **37** (1985) 37.
185. L. L. Branborg, R. Owen, R. Fogel, H. Goldberg et al., *Gastroenterology* **78** (1980) 1602.
186. B. M. Honigberg in L. Kreiser, *Parasitic Protozoa,* **vol. 2,** Academic Press, New York 1978, p. 275.

187. O. Jirovec, M. Petru, *Adv. Parasitol.* **6** (1968) 117.
188. R. Knight, *Clin. Gastroenterol.* **7** (1978) 47.
189. E. A. Meyer, S. Radulescu, *Adv. Parasitol.* **17** (1979) 1.
190. J. W. Smith, M. S. Wolfe, *Ann. Rev. Med.* **31** (1980) 373.
191. J. G. Meingassner, P. G. Heyworth, *Antibiot. Chemother.* **30** (1981) 163.
192. K. Butler, *Chem. Abstr.* **71** (1964) 3384 e.
193. Carlo Erba, US 3 399 193, 1965 (P. N. Giraldi).
194. Hoffmann La Roche, US 3 435 049, 1966 (M. Hoffer).
195. Rhône-Poulenc, *Chem. Abstr.* **63** (1965) 11 571 d.
196. J. Heeres, *Eur. J. Med. Chem.-Chim. Ther. (Paris)* **11** (1976) 237.
197. S. J. Lerman, R. A. Walker, *Clin. Pediatr. (Phila.)* **21** (1982) 409.
198. M. S. Wolfe, *J. Am. Med. Assoc.* **233** (1975)1362.
199. B. R. Brown, *J. Chem. Soc.* 1948, 99.
200. G. D. Drake, *Chem. Abstr.* **51** (1957) 2051 e.
201. I. M. Rollo in A. F. Goodman, L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th ed.,Macmillan Publ. Co., New York 1980, p. 1061.
202. A. R. Surrey, *J. Am. Chem. Soc.* **68** (1946) 113.
203. J. H. Burckhalter, *J. Am. Chem. Soc.* **70** (1948) 1363.
204. A. R. Surrey, *J. Am. Chem. Soc.* **72** (1950) 1814.
205. R. C. Elderfield, *J. Am. Chem. Soc.* **68** (1946) 1525.
206. P. B. Russel, *J. Am. Chem. Soc.* **73** (1951) 3763.
207. J. Souza, *WHO Bull.* **61** (1983) 815.
208. F. H. S. Curd, *J. Chem. Soc.* (1948) 1630.
209. T. Harinasuta, D. Bunnag, W. H. Wernsdorfer, *WHO Bull.* **61** (1983) 299.
210. F. I. Carroll, *J. Med. Chem.* **17** (1974) 210.
211. J. Rinehart, *Am. J. Trop. Med. Hyg.* **25** (1976) 769.
212. K. Killer, *Drugs of the Future* **5** (1980) 547.
213. V. Papesch, *J. Am. Chem. Soc.* **58** (1936) 1314.
214. A. Das, *J. Org. Chem.* **22** (1957) 1111.
215. A. Luis, *Chem. Abstr.* **47** (1953) 10 533 d.
216. I.G. Farbenind, DE 545 915, 1930 (K. Schranz).
217. Boots Pure Drug, US 2 912 438, 1959 (P. Oxley).
218. T. H. Flewett, *Br. Med. Bull.* **41** (1985) 315 – 321.
219. R. E. F. Matthews, *Intervirology* **17** (1982) 1 – 199.
220. B. Rada, D. Blaskovic, F. Sorm, J. Skoda, *Experientia* **16** (1960) 487.
221. E. C. Herrmann, *Proc. Soc. Exp. Biol. Med.* **107** (1961) 142 – 145.
222. B. Jubelt, O. Narajan, R. T. Johnson, *J. Neuropathol. Exp. Neurol.* **39** (1980) 149 – 159.
223. B. Clement, *Pharm. Unserer Zeit* **15** (1986) 72 – 84.
224. H. J. Schaeffer, L. Beauchamp, P. deMiranda, G. B. Elion et al., *Nature (London)* **272** (1978) 583 – 585.
225. H. J. Schaeffer, S. Gurwara, R. Vince, S. Bittner, *J. Med. Chem.* **14** (1971) 367 – 369.
226. B. R. Jones, D. J. Coster, P. N. Fison, G. M. Thompson et al., *Lancet* **1** (1979) 243 – 244.
227. P. R. Laibson, D. Pavan-Langston, W. R. Yeakley, J. Lass, *Am. J. Med.* **73 A** (1982) 281 – 285.
228. C. La Lau, J. A. Oosterhuis, J. Versteeg, G. Van Rij et al., *Am. J. Med.* **73 A** (1982) 305 – 306.
229. J. M. Yeo, A. P. Fiddian, *J. Antimicrob. Chemother.* **12** (1983) Suppl. B, 95 – 103.
230. L. Corey, A. J. Nahmias, M. E. Guinan, J. K. Benedetti et al., *New. Engl. J. Med.* **306** (1982) 1313 – 1319.
231. L. Corey, J. K. Benedetti, C. W. Critchlow, M. R. Remington et al., *Am. J. Med.* **73** (1982) 326 – 334.
232. A. E. Nilsen, T. Aasen, A. M. Halsos, B. R. Kinge et al., *Lancet* **2** (1982) 571 – 573.
233. O. P. Salo, A. Lassus, T. Hovi, A. P. Fiddian, *Eur. J. Sex. Trans. Dis.* **1**(1983) 95 – 98.
234. R. C. Reichman, G. J. Badger, G. J. Mertz, L. Corey et al., *J. Am. Med. Assoc.* **251** (1984) 2103 – 2107.
235. J. M. Douglas, C. Critchlow, J. Benedetti, G. J. Mertz et al., *New. Engl. J. Med.* **310** (1984) 1551 – 1556.
236. S. E. Straus, H. E. Takiff, M. Seidlin, S. Bachrach et al., *New Engl. J. Med.* **310** (1984) 1545 – 1550.
237. D. Brigden, *Br. Med. Bull.* **41** (1985) 357 – 360.
238. T. A. Krenitsky, W. W. Hall, P. deMiranda, L. M. Beauchamp et al., *Proc. Natl. Acad. Sci. U.S.A.* **81** (1984) 3209 – 3213.
239. G. R. Revankar, J. H. Huffman, L. B. Allen, R. W. Sidwell et al., *J. Med. Chem.* **18** (1975) 721 – 726.
240. H. E. Kaufman, *Proc. Soc. Exp. Biol. Med.* **109** (1962) 251 – 252.
241. H. E. Kaufman, E. D. Varnell, Y. M. CentifantoFitzgerald, J. G. Sanitato, *Antiviral Res.* **4** (1984) 333 – 338.

242. W. H. Prusoff, *Biochim. Biophys. Acta* **32** (1959) 295 – 296.
243. P. C. Heidelberger, D. G. Parsons, D. C. Remy, *J. Med. Chem.* **7** (1964) 1 – 5.
244. C. Wellings, P. N. Awdry, F. H. Bors, B. R. Jones et al., *Am. J. Ophthalmol.* **73** (1972) 932 – 942.
245. P. C. Heidelberger, D. King, *Pharmacol. Ther.* **6** (1979) 427 – 442.
246. P. C. Heidelberger, *Ann. N. Y. Acad. Sci.* **255** (1975) 317 – 325.
247. W. W. Lee, A. Benitez, L. Goodman, B. R. Baker, *J. Am. Chem. Soc.* **82** (1960) 2648 – 2649.
248. E. J. Reist, A. Benitez, L. Goodman, B. R. Baker et al., *J. Org. Chem.* **27**(1962) 3274 – 3279.
249. J. DeRudder, M. Privat de Garlihe, *Antimicrob. Agents Chemother.* **5** (1965) 578 – 584.
250. F. M. Schabel, *Chemotherapy* **13** (1968) 321 – 338.
251. D. Pavan-Langston, C. H. Dohiman, *Am. J. Ophthalmol.* **74** (1972) 81 – 88.
252. R. J. Whitley, L. T. Ch'ien, R. Dolin, G. J. Galasso et al., *New. Engl. J. Med.* **294** (1976) 1193 – 1199.
253. R. J. Whitley, S.-J. Soong, R. Dolin, G. J. Galasso et al., *New. Engl. J. Med.* **297** (1977) 289 – 294.
254. B. Sköldenberg, M. Forsgren, K. Alestig, T. Bergström et al., *Lancet* **2** (1984) 707 – 711.
255. H. G. Adams, E. A. Benson, E. R. Alexander, L. A. Vontver et al., *J. Infect. Dis.* **133** (1976) Suppl. A, 151 – 159.
256. S. L. Spruance, C. S. Crumpacker, H. Haines, C. Bader et al., *New. Engl. J. Med.* **300** (1979) 1180 – 1184.
257. W. H. Prusoff, M. Zucker, W. R. Mancini, M. J. Otto et al., *Antiviral Res.* (1985) Suppl. 1, 1 – 10.
258. E. DeClercq, J. Descamps, P. DeSomer, P. J. Barr et al., *Proc. Natl. Acad. Sci.* **76** (1979) 2947 – 2951.
259. E. DeClercq, H. Degreef, J. Wildiers, G. DeJonge et al., *Br. Med. J.* **281** (1980)1178.
260. P. C. Maudgal, E. De Clercq, L. Missotten, *Antiviral Res.* **4** (1984) 281 – 291.
261. R. K. Robins, *Chem. Eng. News* **64** (1986) 28 – 40.
262. K. A. Watanabe, U. Reichman, K. Hirota, C. Lopez et al., *J. Med. Chem.* **22** (1979) 21 – 24.
263. K. A. Watanabe, T.-L. Su, R. S. Klein, C. K. Chu, *J. Med. Chem.* **26** (1983) 152 – 156.
264. C. W. Young, R. Schneider, B. Leyland-Jones, D. Armstrong et al., *Cancer Res.* **43** (1983) 5006 –5009.
265. Y.-C. Cheng, G. Dutschman, J. J. Fox, K. A. Watanabe et al., *Antimicrob. Agents Chemother.* **20** (1981) 420 – 423.
266. H. S. Allaudeen, J. Descamps, R. K. Sehgal, J. J. Fox, *J. Biol. Chem.* **257** (1982) 11 879 – 11 882.
267. C. McLaren, M. S. Chen, R. H. Barbhaiya, R. A. Buroker et al. in R. Kono (ed.): *Herpesviruses and Virus Chemotherapy,* Elsevier Science Publ., New York 1985, pp. 57 – 61.
268. J. T. Witkowski, R. K. Robins, R. W. Sidwell, L. N. Simon, *J. Med. Chem.* **15** (1972) 1150 – 1154.
269. R. W. Sidwell, R. K. Robins, I. W. Hillyard, *Pharmacol. Ther.* **6** (1979) 123 – 146.
270. D. G. Streeter, J. T. Witkowski, G. P. Khare, R. W. Sidwell et al., *Proc. Natl. Acad. Sci.* **70** (1973) 1174 – 1178.
271. B. B. Goswami, E. Borek, O. K. Sharma, J. Fujitaki et al., *Biochem. Biophys. Res. Commun.* **89** (1979) 830 – 836.
272. J. B. McCormick, J. P. Getchell, S. W. Mitchell, D. R. Hicks, *Lancet* **2** (1984) 1367 – 1369.
273. C. B. Hall, J. T. McBride, E. E. Walsh, D. M. Bell et al., *New Engl. J. Med.* **308** (1983) 1443 – 1447.
274. C. B. Hall, J. T. McBride, C. L. Gala, S. W. Hildreth et al., *J. Am. Med. Assoc.* **254** (1985) 3047 – 3051.
275. J. B. McCormick, I. J. King, P. A. Webb, C. L. Scribner et al., *New Engl. J. Med.* **314** (1986) 20 – 26.
276. C. R. Magnussen, J. R. Douglas, R. F. Betts, F. K. Roth et al., *Antimicrob. Agents Chemother.* **12** (1977) 498 – 502.
277. F. Salido-Rengell, H. Nasser-Quinones, B. Briseno Garcia, *Ann. N.Y. Acad. Sci.* **284** (1977) 272 –277.
278. C. B. Smith, R. P. Charette, J. P. Fox, M. K. Cooney et al. in R. A. Smith, W. Kirkpatrick (eds.): *Ribavirin: A Broad Spectrum Antiviral Agent,* Academic Press, New York 1980, pp. 147 – 164.
279. Y. Togo, E. A. McCracken, *J. Infect. Dis.* **133** (1976) Suppl. A, 109 – 113.
280. V. Knight, S. Wilson, J. M. Quarles, S. E. Greggs et al., *Lancet* **2** (1981) 945 – 949.
281. H. W. McClung, V. Knight, B. E. Gilbert, S. Z. Wilson et al., *J. Am. Med. Assoc.* **249** (1983) 2671 – 2674.
282. R. Vince, S. Daluge, *J. Med. Chem.* **20** (1977) 612 – 613.

283. W. M. Shannon, L. Westbrook, G. Arnett, S. Daluge et al., *Antimicrob. Agents Chemother.* **24** (1983) 538 – 543.
284. R. Vince, S. Daluge, H. Lee, W. M. Shannon et al., *Science* **221** (1983) 1405 – 1406.
285. K. K. Ogilvie, U. O. Cheriyan, B. K. Radatus, K. O. Smith et al., *Can. J. Chem.* **60** (1982) 3005 – 3010.
286. W. T. Ashton, J. D. Karkas, A. K. Field, R. L. Tolman, *Biochem. Biophys. Res. Commun.* **108** (1982) 1716 – 1721.
287. A. K. Field, M. E. Davies, C. DeWitt, H. C. Perry et al., *Proc. Natl. Acad. Sci.* **80** (1983) 4139 – 4143.
288. D. H. Shepp, P. S. Dandliker, P. deMiranda, T. C. Burnette et al., *Ann. Intern. Med.* **103** (1985) 368 – 373.
289. D. Felsenstein, D. J. D'Amico, M. S. Hirsch, D. A. Neumeyer et al., *Ann. Intern. Med.* **103** (1985) 377 – 380.
290. M. C. Bach, S. P. Bagwell, N. P. Knapp, K. M. Davis et al., *Ann. Intern. Med.* **103** (1985) 381 – 382.
291. J. C. Martin, C. A. Dvorak, D. F. Smee, T. R. Matthews et al., *J. Med. Chem.* **26** (1983) 759 – 761.
292. A. Larsson, B. Öberg, S. Alenius, C.-E. Hagberg et al., *Antimicrob. Agents Chemother.* **23** (1983) 664 – 670.
293. A. Larsson, P.-Z. Tao, *Antimicrob. Agents Chemother.* **25** (1984) 524 – 526.
294. A. Larsson, S. Alenius, N.-G. Johansson, B. Öberg, *Antiviral Res.* **3** (1983) 77 – 86.
295. Astra, EP-A 55 239, 1982 (C. E. Hagberg, K. N. G. Johansson, Z. M. I. Kovacs, G. B. Stening).
296. N. L. Shipkowitz, R. R. Bower, R. N. Appell, C. W. Nordeen et al., *Appl. Microbiol.* **26** (1973) 264 – 267.
297. J. C.-H. Mao, E. E. Robishaw, *Biochemistry* **14** (1975) 5475 – 5479.
298. C. L. K. Sabourin, J. M. Reno, J. A. Boezi, *Arch. Biochem. Biophys.* **187** (1978) 96 – 101.
299. B. A. Bopp, C. B. Estep, D. J. Anderson, *Fed. Proc.* **36** (1977) 939.
300. J. C.-H. Mao, E. R. Otis, A. M. von Esch, T. R. Herrin et al., *Antimicrob. Agents Chemother.* **27** (1985) 197 – 202.
301. J. M. Reno, L. F. Lee, J. A. Boezi, *Antimicrob. Agents Chemother.* **13** (1978) 188 – 192.
302. E. Helgstrand, H. Flodh, J.-O. Lernestedt, J. Lundstrom et al. in L. H. Collier, J. Oxford (eds.): *Developments in Antiviral Therapy,* Academic Press, London 1980, pp. 63 – 83.
303. J. Wallin, J.-O. Lernestedt, S. Ogenstad, E. Lycke, *Scand. J. Infect. Dis.* **17** (1985) 165 – 172.
304. G. Klintmalm, B. Lönngvist, B. Öberg, G. Gahrton et al., *Scand. J. Infect. Dis.* **17** (1985) 157 – 163.
305. E. G. Sandstrom, J. C. Kaplan, R. E. Byington, M. S. Hirsch, *Lancet* **1** (1985) 1480 – 1482.
306. W. L. Davies, R. R. Grunert, R. F. Haff, J. W. McGahen et al., *Science* **144** (1964) 862 – 863.
307. N. Kato, H. J. Eggers, *Virology* **37** (1969) 632 – 641.
308. G. G. Jackson, R. L. Muldoon, L. W. Akers, *Antimicrob. Agents Chemother.* **3** (1963) 703 – 707.
309. A. W. Galbraith, J. S. Oxford, G. C. Schild, G. I. Watson, *Lancet* **2** (1969) 1026 – 1028.
310. R. Dolin, R. C. Reichman, H. P. Madore, R. Maynard et al., *New Engl. J. Med.* **307** (1982) 580 – 584.
311. L. P. VanVoris, R. F. Betts, F. G. Hayden, W. A. Christmas et al., *J. Am. Med. Assoc.* **245** (1981) 1128 – 1131.
312. F. G. Hayden, A. S. Monto, *Antimicrob. Agents Chemother.* **29** (1986) 339 – 341.
313. J. H. Wikel, C. J. Paget, D. C. DeLong, J. D. Nelson et al., *J. Med. Chem.* **23** (1980) 368 – 372.
314. D. C. DeLong, S. E. Reed, *J. Infect. Dis.* **141** (1980) 87 – 91.
315. R. J. Phillpotts, D. C. DeLong, J. Wallace, R. W. Jones et al., *Lancet* **1** (1981) 1342 – 1344.
316. F. G. Hayden, J. M. Gwaltney, *Antimicrob. Agents Chemother.* **21** (1982) 892 – 897.
317. R. A. Levandowski, C. T. Pachucki, M. Rubenis, G. G. Jackson, *Antimicrob. Agents Chemother.* **22** (1982) 1004 – 1007.
318. R. J. Phillpotts, J. Wallace, D. A. J. Tyrrell, V. B. Tagart, *Antimicrob. Agents Chemother.* **23** (1983) 671 – 675.
319. F. D. Miller, A. S. Monto, D. C. DeLong, A. Exelby et al., *Antimicrob. Agents Chemother.* **27** (1985) 102 – 106.
320. D. J. Bauer, J. W. T. Selway, J. F. Batchelor, M. Tisdale et al., *Nature (London)* **292** (1981) 369 – 370.
321. M. Tisdale, J. W. T. Selway, *J. Gen. Virol.* **64** (1983) 795 – 803.
322. M. Tisdale, J. W. T. Selway, *J. Antimicrob. Chem.* **14** (1984) Suppl. A, 97 – 105.
323. R. J. Phillpotts, J. Wallace, D. A. J. Tyrrell, D. S. Freestone et al., *Arch. Virol.* **75** (1983) 115 – 121.

324. H. Ishitsuka, Y. T. Ninomiya, C. Ohsawa, M. Fujiu et al., *Antimicrob. Agents Chemother.* **22** (1982) 617 – 621.
325. Y. Ninomiya, C. Ohsawa, M. Aoyama, I. Umeda et al., *Virology* **134** (1984) 269 – 276.
326. R. J. Phillpotts, P. G. Higgins, J. S. Willman, D. A. J. Tyrrell et al., *J. Antimicrob. Chemother.* **14** (1984) 403 – 409.
327. G. D. Diana, U. J. Salvador, E. G. Zalay, R. E. Johnson et al., *J. Med. Chem.* **20** (1977) 750 – 756.
328. G. D. Diana, M. A. McKinlay, M. J. Otto, V. Akullian et al., *J. Med. Chem.* **28** (1985) 1906 – 1910.
329. M. A. McKinlay, J. V. Miralles, C. J. Brisson, F. Pancic, *Antimicrob. Agents Chemother.* **22** (1982) 1022 – 1025.
330. M. J. Otto, M. P. Fox, M. J. Fancher, M. F. Kuhrt et al., *Antimicrob. Agents Chemother.* **27** (1985) 883 – 886.
331. M. A. McKinlay, B. A. Steinberg, *Antimicrob. Agents Chemother.* **29** (1986) 30 – 32.
332. J. J. McSharry, L. A. Caliguiri, H. J. Eggers, *Virology* **97** (1979) 307 – 315.
333. T.-S. Lin, W. H. Prusoff, *J. Med. Chem.* **21** (1978) 109 – 112.
334. C. Norman, *Science* **230** (1985) 1355 – 1358.
335. *FDC Reports,* Mar. 3, 1986.
336. J. P. Horwitz, J. Chua, M. Noel, *J. Org. Chem.* **29** (1964) 2076 – 2078.
337. R. S. Goldsmith in B. G. Katzung (ed.): *Basic Clinical Pharmacology,* 2nd ed., Lange Medical Publications, Los Altos, Calif., 1984, pp. 648 – 675.
338. E. DeClercq, *Cancer Lett.* **8** (1979) 9 – 22.
339. H. Mitsuya, M. Popovic, R. Yarchoan, S. Matsushita et al., *Science* **226** (1984) 172 – 174.
340. D. Rouvroy, J. Bogaerts, J.-B. Habyarimana, D. Nzaramba et al., *Lancet* **1** (1985) 878 – 879.
341. M. Windholz, S. Budavari, R. F. Blumetti, E. S. Otterbein (eds.): *The Merck Index,* 10th ed., Merck & Co., Rahway, N.J., 1983, p. 1294.
342. J. C. Chermann, F. C. Sinoussi, C. Jasmin, *Biochem. Biophys. Res. Commun.* **65** (1975) 1229 – 1236.
343. W. Rozenbaum, D. Dormont, B. Spire, E. Vilmer et al., *Lancet* **1** (1985) 450 – 451.
344. M. Michelon, G. Herve, *C. R. Acad. Sci. Paris* **274** (1972) 209 – 212.

Chiral Compounds → Optically Active Compounds

Chitin and Chitosan

SHIGEHIRO HIRANO, Tottori University, Tottori, Japan

1. Introduction 1
2. Molecular Structure and Conformation 1
2.1. Chitin 2
2.2. Chitosan 2
3. Raw Materials and Production . . . 2
3.1. Isolation of Chitin from Crab and Shrimp Shells 2
3.2. Preparation of Chitosan from Chitin 4
4. Metabolism and Biosynthesis 5
5. Chemical Properties 5
5.1. Reactions on the Amino Group . . . 5
5.1.1. *N*-Acylation 5
5.1.2. Formation of *N*-Alkylidene and *N*-Arylidene Derivatives 5
5.1.3. *N*-Alkylation and *N*-Arylation 5
5.2. Reactions at the Hydroxyl Group . 6
5.3. Reactions at C-6 7
5.4. Graft Polymerization on Chitin and Chitosan 7
6. Application Forms and Formulations 7
7. Uses 8
8. Economic Aspects 9
9. Toxicology and Environmental Aspects 9
10. References 10

1. Introduction

In 1811 BRACONNOT [1] isolated from a fungus a compound which he called "Fungin". In 1823 ODIER discovered that this compound is a constituent of the exoskeleton of insects and gave it the name "chitin", which means envelope in Greek [2].

In 1859 ROUGET [3] prepared a compound from chitin by treatment with concentrate alkalis. This compound was named "chitosan" by HOPPE-SEILER [4] in 1894.

Both chitin and chitosan are naturally occurring cationic biopolymers, that are obtained from crab and shrimp shells, being waste materials of marine food processing companies. Chitin and chitosan are biodegradable, biocompatible with animal and plant cells and tissues, and almost nontoxic. Both can be processed into various products including hydrogels, beads, membranes, and sponges.

2. Molecular Structure and Conformation

Chitin is a linear (1→4)-linked 2-acetamido-2-deoxy-β-*d*-glucopyranan (*N*-acetyl-β-*d*-glucosaminan) in the chair $^4C_1$ conformation without any branching, and chitosan is a linear (1→4)-linked 2-amino-2-deoxy-β-*d*-glucopyranan (β-*d*-glucosaminan) (see below). Between chitin and chitosan, a number of derivatives of partially *N*-deacylated chitins and partially *N*-acetylated chitosans exist. Chitins are insoluble in aqueous organic acids (e.g., acetic acid), although chitosans are soluble. Their physical properties vary with the molecular mass, the degree of substitution (d.s.) and the distribution of free amino group on the chain.

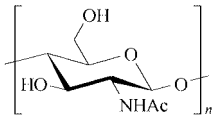

Chitin

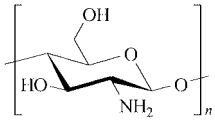

Chitosan

2.1. Chitin

On the basis of X-ray diffraction patterns, chitin occurs in two polymorphs (α- and β-chitin) in the solid state, owing to the orientation of the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_231

extended chains and inter- and intra-molecular hydrogen bonds (Fig. 1). In α-chitin the chains have an antiparallel, in β-chitin a parallel orientation [5], [6]. The α-orientation is more stable and more widely distributed in the nature than the β-orientation which is found only in squid pens. The β-orientation transforms into the α-orientation in the solid state by treatment with acids, but the reverse reaction does no occur under these conditions.

2.2. Chitosan

Chitosan occurs mainly in two molecular conformations: as extended two-fold helix and as extended eight-fold helix (Fig. 2) [7]. The eight-fold helix conformation transforms into the two-fold helix under conditions of high humidity [8]. No ordered conformation is present in the aqueous acidic solution of chitosans. The molecular flexibility increases with increasing *N*-deacetylation, with increasing ionic strength in the solutions, and increasing temperature.

3. Raw Materials and Production

It is estimated that about 100×10^9 t of chitin are produced every year on the earth. As shown in Table 1, chitin is distributed widely in the exoskeleton of crustaceans (crab, shrimp etc.), in the cartilages of mollusks (krill, squid etc.), in the cuticles of insects (cockroach, beetle etc.), and in the cell walls of micro-organisms (fungi, yeasts, etc.) [9]. At present the major industrial source of chitin and chitosan are the shell wastes of crabs and shrimps. The shells mainly consist (on a dry basis) of chitin (20 – 30 %), proteins (30 – 40 %), calcium carbonate (30 – 50 %), lipids, and astaxanthin (less than 1 %).

Table 1. Chitin content of some organisms

Organism	Chitin content, %
Crustacea	
Carcinus (crab)	0.4 – 3.3 [a]
Callinectes (blue crab)	14 [a]
Paralithodes (king crab)	10.4 [a], 35 [b]
Pleuroncodes (red crab)	1.3 – 1.8 [b]
Crangon (shrimp)	5.8 [b]
Insects	
Blatella (cockroach)	10 [b]
Coleoptera (beetle)	5 – 15 [b]
Grasshopper	2 – 4 [a]
May beetle	16 [b]
Mollusks	
Clam shell	6.1 [b]
Krill	40.2 [b]
Oyster shell	3.6 [b]
Squid pen	41 [b]
Microorganisms	
Aspergillus niger	42 [c]
Lactarius vellereus (mushroom)	19.0 [c]
Mucor rouxii	44.5 [c]
Penicillium notatum	18.5 [c]
Saccharomyces cerevisiae (baker's yeast)	2.9 [c]

[a] Wet body weight. [b] Dry body weight. [c] Dry weight of the cell wall.

3.1. Isolation of Chitin from Crab and Shrimp Shells

Chitin is isolated from crab and shrimp shells as follows: (1) Ca_2CO_3 in the shell flakes is dissolved out by treatment with dilute HCl (demineralization), (2) astaxanthin pigments and lipids are extracted with organic solvents, e.g., acetone and ethanol, (decolorization), and (3) proteins are extracted with dilute NaOH or digested enzymatically by proteases or microorganisms [10] (deproteinization). Chitin is obtained as residue in the form of flakes.

Astaxanthin is extracted more efficiently from fresh wet shells than from dry shells [11]. The pigment is easily decomposed by exposing to sunlight in the air-drying step, because the conjugated double bonds in the pigment molecule are very sensitive to ultraviolet light. An oxidative bleaching treatment with H_2O_2 or NaOCl is used at some companies in order to obtain a white-colored product, but this process partially cleaves the glycosidic linkage at the *N*-deacetylated portions, resulting in the formation of low molecular mass products having a 2,5-anhydro-*d*-mannose residue at the reducing end group.

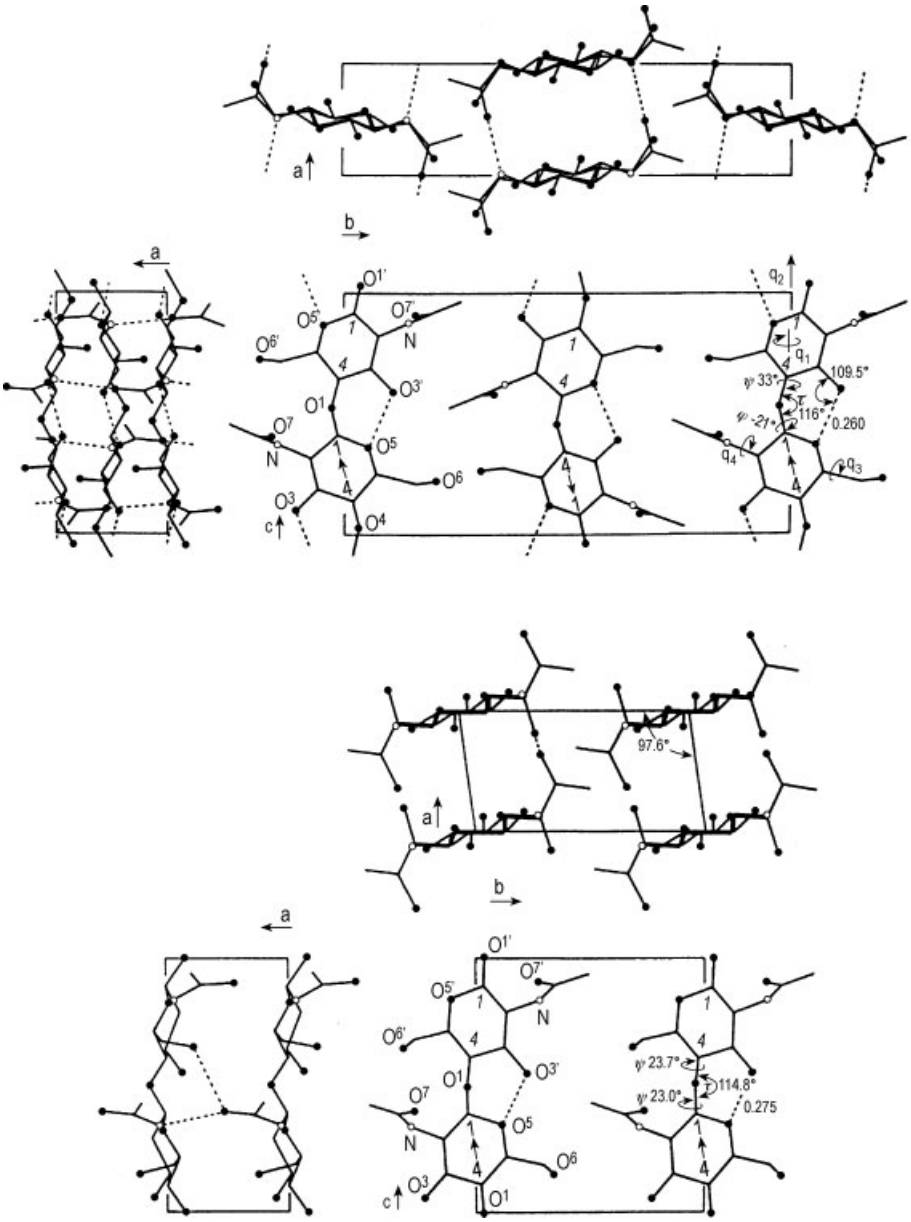

Figure 1. Molecular conformations of chitin in the sold state β-chitin in the parallel conformation (above) and α-chitin in the antiparallel conformation (below)

3.2. Preparation of Chitosan from Chitin

For preparation of chitosan, chitin flakes are treated in suspension with aqueous 30 – 60 % NaOH at 80 – 120 °C with stirring for 4 – 6 h, and this treatment is repeated once or more times for obtaining highly *N*-deacetylated products. However, repeated treatment is generally accompanied with depolymerization. Alternatively, the clear homogeneous solution of the sodium salt of chitin (alkaline chitin) in aqueous 14 % NaOH is treated at 25 °C to give rise to the partially *N*-deacetylated derivatives (d.s.

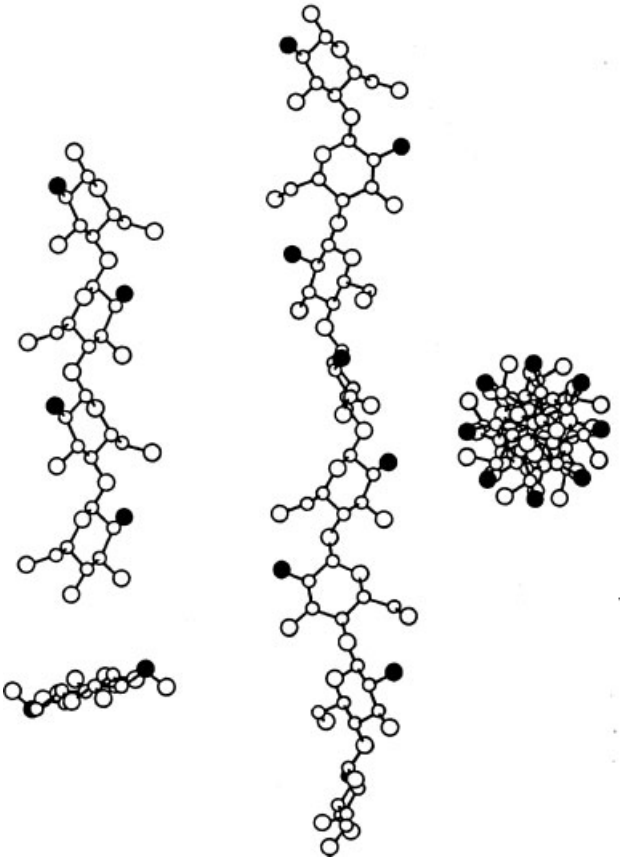

Figure 2. Molecular conformations of chitosan at the solid state. The two-fold helix conformation is shown on the left: a side view (above) and a sectional view (below). The eight-fold helix conformation is shown on the right: a side view (left) and a sectional view (right).

about 0.5 for *N*Ac), which are soluble in water [12]. A random distribution of *N*-acetyl group is found in these products [13].

Chitosan can be prepared from chitin also enzymatically [14], [15]. A powdered sample of chitin is treated with *N*-deacetylase (EC 3.5.1.41) (see Chap. 4) or with microbes which secret *N*-deacetylase. The enzymatic method yields chitosan with low degree of N-deacetylation and low degree of depolymerization.

The degree of acetylation in chitosan may be determined (1) by the C : N ratio in the elemental analyses, by, ^{15}N NMR, or ^{13}C NMR spectroscopy [16–18], by IR spectroscopy [19] by colloidal titration [20], and by pyrolysis-gas chromatography [21].

4. Metabolism and Biosynthesis

Chitinase (EC 3.2.1.14) and lysozyme (EC 3.2.1.17) hydrolyze the internal β-(1→4)-glycosidic linkages of chitin to give chitin oligosaccharides, which are further hydrolyzed by β-*d*-glucosaminidase (EC 3.2.1.30) to afford *N*-acetyl *d*-glucosamine. Chitin is converted to chitosan by *N*-deacetylase (EC 3.5.1.41) [22], and chitosan is hydrolyzed by chitosanase (EC 3.2.1.132) to give chitosan oligosaccharides, which are hydrolyzed by β-*d*-glucosaminidase to give rise to *d*-glucosamine (Fig. 3). Chitinase is widely distributed in plants, insects, fish and microorganisms, and lysozyme is found in animals [23].

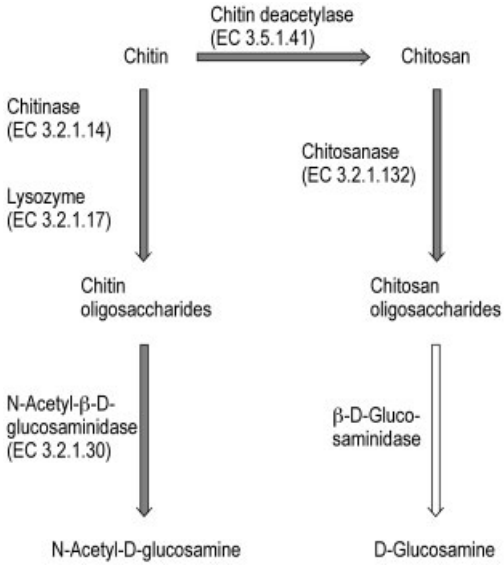

Figure 3. Enzymes for the hydrolyses of chitin and chitosan

5. Chemical Properties

Chitin has two hydroxyl groups at C3 and C6 of the repeating *N*-acetyl-*d*-glucosamine moiety. The OH group at C6 is more reactive then that at C3. Chitosan has one reactive amino group at C2 and two OH groups at C3 and C6. The chemical reactions on these groups give rise to various derivatives (Table 2).

5.1. Reactions on the Amino Group

5.1.1. *N*-Acylation

Chitosan is regioselectively *N*-acylated by treatment with carboxylic anhydrides [24], [25]. The reactions with low fatty acid anhydrides ($C_2 - C_{10}$) yield *N*-acyl derivatives having a d.s. of 1.0, the reactions with high fatty acid anhydrides ($C_{16} - C_{18}$) give rise to the products having d.s. < 1.0.

Chitosan is also *N*-acetylated by reactions with acyl halides [26], with carboxylic acids in the presence of dehydration agents [e.g., dicyclohexylcarbodiimide (DCC)] [27], and with halocarboxylic anhydrides [28]. Under these conditions not only the amino group but also the hydroxyl groups are acylated.

5.1.2. Formation of *N*-Alkylidene and *N*-Arylidene Derivatives

The amino group of chitosan reacts with a series of aldehydes and ketones [29–37], to yield the *N*-alkylidene, *N*-arylidene, and ketamine (Schiff's base) derivatives. The reactions of chitosan with long-chain alkyl mono-aldehydes give rise to the corresponding derivative [38]. Reactions with dialdehydes (e.g., glutaraldehyde and glyoxal) [33], [37], [39], [40] give the cross-linked derivatives.

5.1.3. *N*-Alkylation and *N*-Arylation

N-Alkyl and *N*-aryl derivatives of chitosan are produced (1) by direct *N*-alkylation and *N*-arylation, and (2) by the reduction of the *N*-alkylidene and *N*-arylidene derivatives.

In direct *N*-alkylation and *N*-arylation, the amino group of chitosan reacts with alkyl halides [41], [42] in the presence of a base to form the *N*-mono and *N*-dialkyl derivatives. The *N*-dialkyl derivatives react further with alkyl halides to form the *N*-trialkyl quaternary ammonium slats [36], [41], [43].

The *N*-alkylidene and *N*-arylidene derivatives are reduced by treatment with $NaBH_3CN$ or $NaBH_4$ to yield the corresponding *N*-alkyl and *N*-aryl derivatives [31], [34], [42], [44–48].

5.2. Reactions at the Hydroxyl Group

Chitin, *N*-acylchitosans, and *N*-alkylidene and *N*-arylidene chitosans are *O*-acylated by treatment with acyl halides [49], [50] and with carboxylic anhydrides in pyridine [51], [52] or methanesulfonic acid [53], [54]. *O*-Acyl chitosan derivatives are produced from the *O*-acyl *N*-arylidene- and *N*-alkylidene derivatives by acid hydrolysis of the *N*-arylidene- and *N*-alkylidene groups [55].

The *O*-alkyl and *O*-aryl derivatives are produced from the sodium salt of chitin by treatment with an alkylating agent (e.g., dialkyl sulfate or an alkyl halide).

3-*O*-(1-carboxyethyl)chitosan (muraminan) is prepared as follows (see below):

Table 2. Chemical structures of chitin and chitosan derivatives ∗

Chitosan derivatives	R^1	R^2	R^3
Reactions at the amino group			
1) Acyl	–NHC(=O)–R	–OH	$–CH_2OH$
2) Alkylidene + arylidene	–N=CH–R	–OH	$–CH_2OH$
3) Alkyl + aryl	$–NH–CH_2R$	–OH	$–CH_2OH$
4) Deacyl	$–NH_2$	–OH	$–CH_2OH$
5) Imido	$–N(C=O)_2=R$	–OH	$–CH_2OH$
6) Metal chelate	–NH2Me	–OH	$–CH_2OH$
7) Nitro	–NHNO2	–OH	$–CH_2OH$
8) Salts	$–NH_3^{+-}R$	–OH	$–CH_2OH$
9) Sulfate	$–NHS(=O)_2ONa$	–OH	$–CH_2OH$
10) Sulfonyl	$–NHS(=O)_2R$	–OH	$–CH_2OH$
11) (Thiol)thiocarbonyl	–NH(C=S)SH	–OH	$–CH_2OH$
Reactions at the hydroxyl group			
1) Acyl	–NHAc	–OC(=O)–R	$–CH_2O(C=O)–R$
2) Alkyl + aryl	–NHAc	–O–R	$–CH_2O–R$
3) Metal alcoholate	–NHAc	–ONa	$–CH_2ONa$
4) Nitro	–NHAc	$–ONO_2$	$–CH_2ONO_2$
5) Phosphate	–NHAc	$–OP(=O)(ONa)_2$	$–CH_2OP(=O)(ONa)2$
6) Sulfate	–NHAc	$–OS(=O)_2ONa$	$–CH_2OS(=O)_2ONa$
7) Sulfonyl	–NHAc	$–O–S(=O)_2–R$	$–CH_2O–S(=O)_2–R$
8) Siryl	–NHAc	$–O–Si(R)_3$	$–CH_2–Si(R)_3$
9) (Thiol)thiocarbonyl	–NHAc	–OC(=S)SR	–OC(=S)SR
Reaction at the carbon chain			
1) 3,6-Anhydro	$–NHC(=O)CH_3$	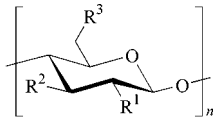	
2) Azido	–NHAc	–OH	$–CH_2N_3$
3) Halo	–NHAc	–OH	$–CH_2X$
4) Mercapto	–NHAc	–OH	$–CH_2SH$
5) Oxidation	–NHAc	–OH	–COOH
6) Oxidative-deamination	–O–C5	–OH	$–CH_2OH$
7) Reduction	–NHAc or $–NH_2$	–OH	$–CH_3$

∗ Ac, acetyl; R, alkyl or aryl group; X, halogen.

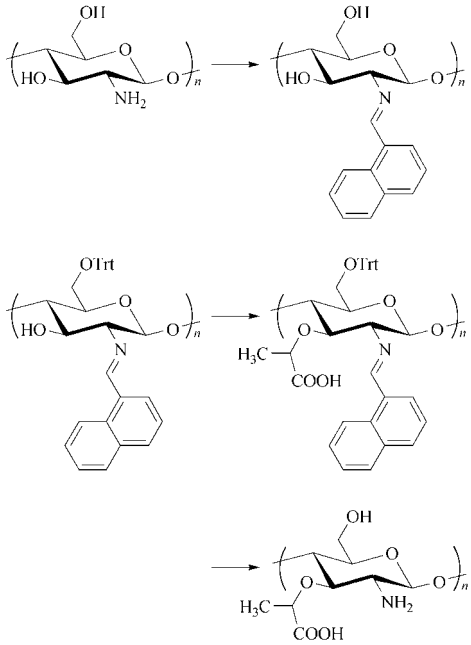

N-1-naphthylmethylenechitosan, which is produced from chitosan by treatment with 1-naphthaldehyde, is reacted with trityl chloride to yield the *N*-(1-naphthylmethylene)-6-*O*-trityl derivative. The compound reacts with 2-chloropropionic acid in *N*-*N*-dimethyl acetamide to form the 3-*O*-(1-carboxyethyl)-*N*-(1-naphthylmethylene)-6-trityl derivative. Subsequently, both, the Schiff's base and the trityl groups, are removed by treatment with aqueous 1 N HCl to give rise to the muraminan [56].

5.3. Reactions at C-6

6-Carboxychitosan is produced from chitosan by treatment with chromium trioxide in acetic acid [57]. 6-Carboxychitin is prepared from chitin by the regioselective oxidation at C6 with NaOCl at 25 °C in aqueous solution [58].

5.4. Graft Polymerization on Chitin and Chitosan

Chitin-polystyrene copolymers are produced by cationic and radical graft copolymerizations of styrene onto 6-iodo-6-deoxychitin [59] and 6-mercapto-6-deoxychitin [60]. Chitin-poly(*N*-acetylethyleneimine) copolymers are produced by graft copolymerization of 2-methyl-2-oxazoline onto both the 6-tosyl and 6-iodo-6-deoxy derivatives [61].

Onto the amino group of chitosan, the lactams of amino acids (e.g., *l*-alanine *N*-carboxylic anhydride) are graft-copolymerized to give rise to chitosan-poly(amino acids) and chitosan-poly(peptides)copolymers [62], [63]. The grafting of poly(4-vinylpyridine) [64] and poly(2-methyl-2-oxazoline) [65] and poly(isobutyl vinyl ether) [66] on chitosan are also performed.

6. Application Forms and Formulations

Hydrogels. Various hydrogels are produced from chitin and chitosan. Chitosan oxalate hydrogel is prepared by dissolution of chitosan in aqueous oxalic acid by heating and subsequent cooling. The hydrogel melts on heating at 80 – 90 °C and solidifies on cooling at room temperature. Chitosan hydrogel can be obtained from the oxalate hydrogel by treatment with aqueous 1 N NaOH [67], [68].

Hydrogels can also be produced from the *N*-acyl, *O*-acyl-*N*-acyl and the alkylidene/arylidene derivatives of chitosan. Chitosan-Ca alginate and *N*-acylchitosan-collagen hydrogels are also known.

Chitosan polyelectrolytes are formed by dissolving chitosan in aqueous organic acids (e.g., acetic, propionic and lactic acids), and reaction of the produced cationic chitosan with polyanions (e.g., chondroitin sulfate, hyaluronate [69], heparin [70], other acidic polysaccharides [71], [72], DNA [73], and some acidic proteins [74]).

Hard composites of chitosan and metals are produced from the corresponding hydrogels (see page 7) by soaking in an aqueous $NaCO_3$, K_2CO_3 or Na_2HCO_3 solution at room temperature for a few days or weeks, [75–81].

Membranes. Chitin and chitosan membranes are prepared as follows:

1) Dissolution of the polymers in formic acid [12], [82–85], aqueous acids [86], [87], or organic solvents [26], [44]. The solutions are spread as a thin layer on a glass plate. After removing the solvent the corresponding membrane is formed
2) Air drying of a thin-sliced piece of *N*-acylchitosan hydrogels or their blend hydrogels (see page 7) [88], [89].
3) Chemically modified membranes are produced by chemical modification of chitin and chitosan membranes in the solid state [90].

N-Acylchitosan membranes are stable in both aqueous acid and aqueous alkaline solutions, but chitosan membranes and *N*-alkylidene- and *N*-arylidene-chitosan membranes are unstable in aqueous acid solutions. Flow rates of water through *N*-acylchitosan membranes are in the range of $10.0 - 23.6 \times 10^{-3}$ m/cm^2 min under a pressure of 29.4×10^4 Pa, and are almost independent on the membrane thickness (12 – 60 µm). The flow rate of water through chitosan membrane (30 – 35 µm in thickness) is 7.1×10^{-4} mL/cm^2 min, and the rate decreases with an increase of the membrane thickness [92].

7. Uses

Chitin and chitosan are utilized as ecologically harmless materials in various fields. These include biotechnology, water-treatment, medicine and veterinary medicine, textiles, membranes, cosmetics, agriculture, and food industry.

Chitin and chitosan hydrogels are used as media for affinity chromatography of enzymes (e.g., chitinase and lysozyme) [91] and lectins (e.g., wheat germ agglutinin, WGA); as media for gel-permeation chromatography [37], [67], [92]; for the isolation of bovine serum albumin [93]; as materials for the preparation of membranes, sponges and sponge sheets; as immobilization media of enzymes [94], [95] and cells [96]; and as wound-dressing materials [97].

Polyelectrolytes are used in paper sizing; as textiles auxiliaries [98]; for immobilization of enzymes [99], [100]; and for wastewater treatment [101], [102].

Metal composites of chitin and chitosan are used as solid electrolytes for secondary lithium cells (chitosan-lithium trifluorate) [103]; and media for the removal of uranium ions [104], [105] and harmful radioisotopes from water [106].

Chitin and chitosan fibers are used as antibacterial, antithrombogenic, and thrombogenic (hemostatic), deodorizing, moisture-controlling, and nonallergenic fibers. Staple fibers and nonwoven fabrics are used as materials for bandages, for wound-dressing in plant and animals, as perfume-releasing fibers, and as textile materials for underwear, sportswear, and socks [107].

Chitin filament (about 5 μm in diameter), bundled up into 16 – 20 filaments, finds application as surgical suture, which has an enough strength for clinical uses [108]. The suture is digestible in the tissues by lysozyme and chitinase, and its digestion period is controlled by the structure of *N*-acyl groups and by their d.s.

Chitosan beads (> 1 mm in diameter) and microspheres (0.1 – 10 μm in diameter) chitin beads, and composite beads are used as media for anion-exchange and affinity chromatography [109]; as controlled release carriers of drugs and agrochemicals [110], [111]; as encapsulating materials for mammalian cells, microbes, and drugs [112–125]; and for immobilization of enzymes [126].

Membranes find application as active transport membranes for halogen and organic ions; in protein purification [127], [128]; as affinity membranes for purification of lysozyme [129]; in dialysis, as edible chitosan-gelatin films [130], [131]; electrochemical luminescence sensors [132]; as enzyme-immobilized membranes [133], [134], anion-exchange, [135–137]; pervaporation [138], [139], potentiometric response membranes [140]; in reverse osmosis and ultrafiltration [92].

Colloidal chitin and powders of chitin and chitoson are used as cosmetic ingredients, solid substrates for chitinase, lysozyme and chitosanase, and emulsifiers in food processing [141].

Aqueous acidic solutions of chitosan, and of *O*-carboxymethylchitin, *O*-hydroxyethylchitin (neutral) and 3,6-methoxychitin in water are highly viscous, and have moisture retaining, antielectrostatic, hair-protecting, odor-absorbing, and anti-bacterial properties. They are used as cosmetic ingredients [142], [143]; and coating agents for lumber surface.

Chitosan and chitin have an *antimicrobial action*, [144], [145], which is even more promnounced in the corresponding quaternary ammonium salts (e.g., *N*-trimethylchitosan hydrochloride). The cationic amino groups of chitosan bind to anionic groups on microbe cell walls, resulting in their growth inhibition. Chitin and chitosan also have an antiviral action in vitro [146]. On supplying chitosan as a feed ingredient in animals, the proportion of the useful microbes (e.g., *Bifidobacteria*) in the intestine increases and the proportion of the harmful microbes decrease [147].

Chitin and chitosan are applied as additives for nonfermented pickles, as anti-microbial agents; antimicrobial food processing additives, and soil amendments in agriculture [148].

Chitosan *N*-acyl derivatives are used as artificial blood vessels, and materials for contact lenses, chitosan is applied as sealing material for arterial puncture sites [149].

Treatment with chitin and chitosan, induces production of various extracellular bioactive compounds that enhance the self-defense function of plants against pathogen infections and diseases resulting in the increase of plant tissue production [150]. Chitin and chitosan activate mammalian cells [151]. As the result, immune reactions [152], [153] are stimulated by producing lysozyme, interleukins [154], leukotriene B4 and prostaglandin E2. Chitin and chitosan thus find application as agricultural materials; coating materials for plant seeds; medical, and veterinary materials [155].

Because of its *hypocholesterolemic function* [156–160] chitosan is used as an additive for feeds and foods; and as ingredient for health foods.

The coating or spraying of an aqueous solution of chitosan over the surface of fruit and vegetables slows down their respiration of CO_2 and ethylene gases. As the results, the ripening of fruits is delayed, the storage life of fruit and

vegetables are prolonged, and their freshness is kept for a long period [161–163].

8. Economic Aspects

The industrial production of chitin and chitosan from crab shells was started in Japan in 1987. At that time, a relatively large amount of crab shells was produced as waste in shellfish processing companies, and easy to collect without any costs. The first industrial application of chitosan was as a flocculating and dewatering agent in municipal wastewater treatment. Since then, the utilization of chitin and chitosan has been expanded into wide fields. The estimated consumption of chitin and chitosan in 1998 in Japan was about 832 t as calculated as chitosan. About 30 % of the total consumption was for wastewater treatment, about 24 % as foods and feeds additives, and about 14 % in agriculture. In the future, the usage in textile, medical, and veterinary materials will increase, that in the wastewater treatment will decrease.

9. Toxicology and Environmental Aspects

Chitin and chitins are almost nontoxic (LD_{50} 16 g per kilogram body weight for rats). No abnormal symptoms are observed with several animals after the oral administration of chitosan for 8 months at a daily dose of 0.7 – 0.8 g per kilogram body weight, and after the intravenous injection of low molecular mass chitosan (M_r ca. 3000) and chitosan oligosaccharides (M_r 304 – 1162) for 11 days at a daily dose of 4.5 mg per kilogram body weight.

Chitin and chitosan are degradable in the biosphere, in the agricultural soil, and in the hydrosphere to produce oligosaccharides. The biodegradation rate varies with seasons in the soil and hydrosphere, and is controlled by chemical modification [164–170].

10. References

1. H. Braconnot, *Ann. Chim. Phys.* **79** (1811) 269.
2. A. Odier, *Mem. Soc. Hist. Nat. Paris* **1** (1823) 29.
3. C. Rouget, *Comp. Rend.* **48** (1859) 792.
4. F. Hoppe-Seiler, *Ber.* **27** (1894) 3329.
5. R. Minke, J. Blackwell, *J. Mol. Biol.* **120** (1978) 167.
6. Y, Saito, T. Okano, H. Chanzy, J. Sugiyama, *J. Struc. Biol.* **114** (1995) 218.
7. K. Ogawa, *Agric. Biol. Chem.* **55** (1991) 375.
8. P. Cairn, *Carbohydr. Res.* **235** (1992) 23.
9. H. M. Cauchiel, G. Murugan, J. P. Thome, H. J. Dumont, *Hydrobiologia* **359** (1997) 23.
10. S. L. Wang, S. H. Chio, *Enzyme Micorb. Technol.* **22** (1998) 629.
11. Jpn. Kokai Tokkyo Koho, JP 99 49 972, 1999 (M. Matsumara, J. Sato).
12. T. Sannan, K. Kurita, Y. Iwakura, *Makromol. Chem.* **176** (1975) 1191.
13. S. Aiba, *Int. J. Biol. Macromol.* **13** (1991) 40.
14. US Patent 5 739 015, 1998 (V. R. Srinivasan).
15. K. Tokuyasu, M. Ohnishi-Kameyama, K. Hayashi, *Biosci. Biotech. Biochem.* **60** (1996) 1598.
16. K. M. Varum, M. W. Anthonsen, H. Grasdalen, O. Smidsrod, *Carbohydr. Res.* **211** (1991) 17.
17. G. Yu, F. G. Morin, G. A. R. Nobes, R. H. Marchessault, *Macromolecules* **32** (1999) 518.
18. K. M. Varum, M. W. Anthonsen, H. Grasdalen, O. Smidsrod, *Carbohydr. Res.* **217** (1991) 19.
19. J. G. Domszy, G. A. F. Roberts, *Macromol. Chem.* **186** (1985) 1671.
20. H. Terayama, *J. Polym. Sci.* **8** (1952) 243.
21. H. Sato, S. Mizutani, S. Tsuge, H. Ohtani, K. Aoi, A. Takasu, M. Okada, S. Kobayashi, T. Kiyosada, S. Shoda, *Anal. Chem.* **70** (1998) 7.
22. A. Martinou, U. Bowiotis, B. Stokke, K. M. Varumd, *Carbohydr. Res.* **311** (1998) 71.
23. M. Matsumiya, K. Miyauchi, A. Mochizuki, *Fish Sci.* **64** (1998) 166.
24. S. Hirano, Y. Ohe, H. Ono, *Carbohydr. Res.* **47** (1976) 315.
25. L. Vachoud, N. Zydowicz, A. Domard, *Carbohydr. Res.* **302** (1997) 169.
26. S. Fuji-i, H. Kumagai, M. Noda, *Carbohydr. Res.* **83** (1980) 389.

27. K. Kurita, T. Sannan, Y. Iwakura, *Makromol. Chem.* **178** (1977) 2595.
28. S. Hirano, Y. Kondo, *Nippon Kagaku Kaishi* (1982) 1622.
29. J. Dutkiewicz, *J. Macromol. Sci. Chem.* **20** (1983) 877.
30. S. Hirano, N. Matsuda, O. Miura, H. Iwaki, *Carbohydr. Res.* **71** (1979) 339.
31. G. K. Moore, G. A. F. Roberts, *Int. J. Biol. Macromol.* **3** (1981) 337.
32. S. Hirano, T. Osaka, *Agric. Biol. Chem.* **47** (1983) 1389.
33. R. A. A. Muzzarelli, F. Tanfani, M. Emanuelli, *Carbohydr. Res.* **107** (1982) 199.
34. P. Tong, Y. Baba, Y. Adachi, K. Kawazu, *Chem. Lett.* (1991) 1529.
35. R. A. A. Muzzarelli, A. Zattoni, *Int. J. Biol. Macromol.* **8** (1986) 137.
36. L. A. Nud'ga, E. A. Plisko, S. N. Danilov, *Zh. Obshch. Khim.* **43** (1973) 339. *Chem. Abst.* **80** 96251.
37. S. Hirano, N. Matsuda, T. Tanaka, *Carbohydr. Res.* **71** (1979) 344.
38. K. Kurita, M. Ishiguro, T. Kitajima, *Int. J. Biol. Macromol.* **10** (1988) 124.
39. R. A. A. Muzzarelli, G. Barontini, R. Rocchetti, *Biotechnol. Bioeng.* **18** (1976) 1445.
40. R. A. A. Muzzarelli, F. Tanfani, S. Mariotti, E. Emanuelli, *Carbohydr. Res.* **104** (1982) 235.
41. E. Kokufuta, Y. Hirai, I. Nakamura, *Makromol. Chem.* **182** (1981) 1714.
42. I. D. Hall, M. Yalpani, *Biopolymers* **20** (1981) 1413.
43. A. Domard, R. Rinaudo, C. Terassin, *Int. J. Biol. Macromol.* **8** (1986) 105.
44. R. A. A. Muzzarelli, F. Tanfani, M. Emanuelli, S. Mariotti, *J. Membr. Sci.* **16** (1983) 295.
45. R. A. A. Muzzarelli, F. Tanfani, *Carbohydr. Res.* **5** (1985) 297.
46. R. A. A. Muzzarelli, F. Tanfani, S. Mariotti, E. Emanuelli, *Carbohydr. Polym.* **2** (1982) 145.
47. M. Yalpani, I. D. Hall, *Can. J. Chem.* **59** (1981) 2934.
48. J. M. Harris, E. Struck, M. Case, M. S. Paley, M. Yalpani, J. Vanalstine, D. Brooks, *J. Polym. Sci. Part A: Polym. Chem.* **22** (1984) 341.
49. S. Hirano, S. Kondo, Y. Ohe, *Polymer* **16** (1975) 622.
50. K. Kaifu, N. Nishi, T. Komai, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1981) 2361.
51. S. Hirano, Y. Koide, *Carbohydr. Res.* **65** (1978) 166.
52. K. Kurita, S. Ishi, K. Tomita, S. Nishimura, K. Shimoda, *J. Polym. Sci. Part A: Polym. Chem.* **32** (1994) 1027.
53. N. Nishi, J. Noguchi, S. Tokura, H. Shiota, *Polym. J.* **11** (1979) 27.
54. N. Nishi, H. Ohmura, S. Nishimura, O. Somorin, S. Tokura, *Polym. J.* **14** (1982) 919.
55. G. K. Moore, G. A. F. Roberts, *Int. J. Biol. Macromol.* **4** (1982) 246.
56. S. Hirano, Y. Kondo, H. Inui, F. Hirano, K. Nagamura, T. Yoshizumi, *Carbohydr. Polym.* **31** (1996) 29.
57. D. Horton, E. K. Just, *Carbohydr. Res.* **29** (1973) 173.
58. R. A. A. Muzzarelli, C. Muzarelli, A. Cosani, M. Terbojevich, *Carbohydr. Polym.* in press.
59. K. Kurita, H. Yoshino, S. Inoue, K. Yamamura, S. Ishi, S. Nishimura, *Macromolecules* **25** (1992) 3791.
60. K. Kurita, S. Hashimoto, H. Yoshino, S. Ishii, S. Nishimura, *Macromolecules* **29** (1996) 1939.
61. K. Kurita, S. Hashimoto, S. Ishii, S. Nishimura, *Polym. J.* **28** (1996) 686.
62. K. Kurita, A. Yoshida, Y. Koyama, *Macromolecules* **21** (1988) 1579.
63. K. Kurita, S. Iwaki, S. Ishi, S. Nishimura, *J. Polym. Sci. Part A: Polym. Chem.* **30** (1992) 685.
64. H. Caner, H. Hasipoglu, O. Yilmaz, E. Yilmaz, *Eur. Poly. J.* **34** (1998) 493.
65. K. Naka, R. Yamashita, T. Nakamura, I. Ohki, S. Maeda, K. Aoi, A. Takasu, M. Okada, *Int. J. Biol. Macromol.* **23** (1998) 259.
66. S. Yoshikawa, T. Takayama, N. Tsubokawa, *J. Appl. Polym. Sci.* **68** (1998) 1883.
67. S. Hirano, R. Yamaguchi, N. Fukui, M. Iwata, *Carbohydr. Res.* **201** (1990) 145.
68. S. Hirano, R. Yamaguchi, N. Fukui, M. Iwata in C. G. Gebelein (ed.): *Biotechnology and Polymers,* Plenum, New York 1991, pp. 181 – 188.
69. A. Denuziere, D. Ferrier, O. Damour, A. Domard, *Biomaterials* **19** (1998) 1275.
70. Y. Kikuchi, *Makromol. Chem.* **157** (1974) 2209.
71. Y. Kikuchi, H. Fukuda, *Makromol. Chem.* **178** (1977) 2895.
72. S. Hirano, C. Mizutani, R. Yamaguchi, O. Miura, *Biopolymers* **17** (1978) 805.
73. S. C. W. Richardson, V. H. J. Koll, R. Duncan, *Int. J. Pharm.* **178** (1999) 231.
74. R. Yamaguchi, A. Arai, S. Hirano, T. Ito, *Agric. Biol. Chem.* **42** (1978) 1297.
75. S. Hirano, O. Miura, R. Yamaguchi, *Agric. Biol. Chem.* **41** (1977) 1755.

76. S. Hirano, K. Yamamoto, H. Inui, K. I. Draget, K. M. Varum, O. Smidsrod, *Adv. Chitin Sci.* **2** (1997) 1.
77. S. Hirano, K. Yamamoto, H. Inui, K. I. Draget, K. M. Varum, O. Smidsrod, *Studies on Surface and Catalysis* **114** (1998) 621.
78. C. H. Chu, H. Kumagai, T. Sakiyama, S. Ikeda, K. Nakamura, *Biosci. Biotech. Biochem.* **60** (1996) 1627.
79. S. Hirano, Y. Yamada, K. Yamamoto, H. Inui, M. Ji, *Advan. Chitin Sci.* **1** (1996) 137.
80. S. Hirano, K. Yamamoto, H. Inui, M. Ji, M. Zhang, *Macromol. Symp.* **105** (1996) 149.
81. S. Hirano, K. Yamamoto, H. Inui, *Energy Convers. Manage.* **38** (1997) S517.
82. S. Tokura, N. Nishi, J. Noguchi, *Polym. J.* **11** (1979) 781.
83. T. Uragami, Y. Ohsumi, M. Sugihara, *Polymer* **22** (1981) 1155.
84. S. Aiba, M. Izume, N. Minoura, Y. Fujiwara, *British Polymer* **17** (1985) 38.
85. S. Aiba, M. Izume, N. Minoura, Y. Fujiwara, *Carbohydr. Polym.* **5** (1985) 285.
86. S. Hirano, K. Tobetto, Y. Noishii, *J. Biomed. Mater. Res.* **15** (1981) 903.
87. A. L. Andrady, P. Xu, *J. Polym. Sci. Part B: Polym. Phys.* **35** (1997) 517.
88. S. Hirano, *Agric. Biol. Chem.* **42** (1978) 1939.
89. O. W. Achaw, M. F. A. Goosen, *Polym. Prepr. (ACS)* **35** (1994) 75.
90. E. Khor, A. Wan, C. A. Tee, F. Chae, G. W. Hastings, *J. Polym. Sci. Part A: Polym. Chem.* **35** (1997) 2049.
91. S. Hirano, H. Kaneko, M. Kitagawa, *Agric. Biol. Chem.* **55** (1991) 1683.
92. S. Hirano, K. Tobetto, M. Hasegawa, N. Matsuda, *J. Biomed. Mater. Res.* **14** (1980) 477.
93. M. F. A. Goosen, O. W. Achaw in M. F. A. Goosen (ed.): *Applications of Chitin and Chitosan,* Technomic, Lancaster 1997, pp. 233 – 254.
94. H. Tanibe in M. Yabuki (ed.): *Applications of chitin and chitosan,* Gihodo Publishing Co., Tokyo 1990, pp. 123 – 154.
95. T. Y. Hsien, G. C. Rorrer, *Ind. Eng. Chem. Res.* **36** (1997) 3631.
96. S. Kirkwood, *Ann. Rev. Biochem.* **43** (1974) 401.
97. S. L. Wang, S. H. Chio, *Enzyme Microb. Technol.* **22** (1998) 634.
98. M. R. Julia, J. M. Canal, D. Jocic, *Rev. Quim. Text.* **140** (1988) 58, 62. *Chem. Abst.* **130** 326239.
99. W. S. Ngah, L. K. H. Wan, *Ind. Eng. Chem. Res.* **38** (1999) 1411.
100. H. Minamisawa, H. Iwanami, N. Arai, T. Okutani, *Anal. Chim. Acta.* **378** (1999) 279.
101. L. Guerrero, F. Omil, R. Mendez, J. M. Lema, *Bioresour. Technol.* **63** (1998) 221.
102. M. Hashimoto, Y. Ishii, Y. Ohi, 55th Annual Conf. Water Pollution Control Federation, St. Louis, Mo. USA, 1982.
103. N. M. Morni, A. K. Arot, *J. Power Sources* **77** (1999) 42.
104. T. Sakaguchi, T. Horikoshi, A. Nakajima, *Nippon Nogei Kagaku Kaishi* **53** (1979) 211.
105. S. Hirano, Y. Kondo, Y. Nakazawa, *Carbohydr. Res.* **100** (1982) 431.
106. R. A. A. Muzzarelli, B. Spalla, *J. Radiational Chem.* **10** (1972) 27.
107. M. Yoshikawa, T. Midorikawa, T. Otsuki, T. Terashi, *Eur. Pat. Appl. EP* **794** (1997) 223.
108. K. M. Nakajima, K. Atsumi, K. Kifune in J. P. Zikakis (ed.): *Chitin, Chitosan and Related Enzymes,* Academic Press, London 1984, pp. 407 – 410.
109. A. Kristane, A. Nysaeter, H. Grasdalen, K. M. Varum, *Carbohydr. Poly.* **38** (1999) 23.
110. C. Aval, J. Akbuga, *Int. J. Pharm.* **168** (1998) 9.
111. F. Bugamelli, M. A. Raggi, I. Orieti, V. Zecchi, *Arch. Pharm. (Weinheim, Ger.)* **331** (1998) 133.
112. P. Giunchedi, L. Genta, B. Conti, R. A. A. Muzzarelli, U. Conte, *Biomaterials* **19** (1998) 157.
113. R. H. Chen, M. L. Tsaih, *J. Appl. Polym. Sci.* **66B** (1997) 161.
114. M. J. B. Miguez, B. C. Rodrigues, M. N. M. Sandrez, M. C. M. Laranjeira, *J. Encapsulation* **14** (1997) 639.
115. M. L. Huguet, E. Dellacherie, *Process Biochem. (Oxford)* **31** (1996) 745.
116. K. Nilsson, W. Scheirer, O. Merten, L. Ostberg, E. Liehl, H. Katinger, K. Mosbach, *Nature* **302** (1983) 629.
117. J. Wang, P. Li, H. Shi, Y. Qian, *J. Environ. Sci. (China)* **9** (1997) 283.
118. A. Javis, T. Springs, W. Chipuru, M. Sullivan, G. Koch, *In Vitro* **18** (1982) 267.
119. L. L. McCormick, D. K. Lichatowich, *J. Polym. Sci., Polym. Lett. Ed.* **17** (1979) 479.
120. Y. Kawashima, T. Handa, A. Kosai, H. Takenaka, S. Lin, Y. Ando, *J. Pharm. Sci.* **74** (1985) 264.
121. T. Yoshioka, R. Hirano, T. Shioya, M. Kako, *Biotechnol. Bioeng.* **35** (1990) 66.
122. K. Aiedeh, E. Gianasi, I. Orienti, V. Zecchi, *J. Microencapsulation* **14** (1997) 567.

123. F. L. Mi, T. B. Wong, S. S. Shyu, *J. Microencapsulation* **14** (1997) 577.
124. N. Tarimci, D. Ermis, *Int. J. Pharm.* **147** (1997) 71.
125. P. Calvo, J. L. Vila-Jato, M. J. Alonso, *Int. J. Pharm.* **153** (1997) 41.
126. A. Illanes, L. Wilson, C. Altamirano, A. Aillapan, *Prog. Biotechnol.* **15** (1998) 27.
127. X. Zeng, E. Ruckenstein, *J. Membr. Sci.* **148** (1998) 195.
128. E. Ruckenstein, X. Zeng, *J. Membr. Sci.* **142** (1998) 13.
129. R. Eli, Z. Xianfnag, *Biotechnol. Bioeng.* **58** (1998) 117.
130. I. S. Arvanitoyannis, A. Nakayama, S. Aiba, *Carbohydr. Polym.* **37** (1998) 371.
131. C. Caner, P. J. Vergano, J. L. Wiles, *J. Food Sci.* **63** (1998) 1049.
132. C. Zhao, N. Egashira, Y. Kurauchi, K. Ohga, *Anal. Sci.* **14** (1998) 439.
133. G. Cho, I. S. Moon, J. S. Lee, *Chem. Lett.* (1997) 577.
134. H. S. Liu, W. H. Chen, J. T. Lai, *Appl. Biochem. Biotechnol.* **66** (1997) 57.
135. J. T. Shieh, R. Y. M. Hung, *J. Membr. Sci.* **127** (1997) 185.
136. A. Ito, M. Sato, T. Anma, *Angew. Makromol. Chem.* **248** (1997) 85.
137. Y. M. Lee, S. Y. Nam, D. J. Woo, *J. Membr. Sci.* **133** (1997) 103.
138. T. Uragami, M. Saito, K. Takigawa, *Makromol. Chem. Rapid Commun.* **9** (1988) 361.
139. J. Jegal, K. H. Lee, *J. Appl. Polym. Sci.* **71** (1999) 671.
140. B. H. Choi, Y. J. Yun, *Anal. Sci. Technol.* **11** (1998) 235.
141. P. C. Schulz, M. S. Rodriguz, L. F. DelBlanco, M. Pistonesi, E. Agulls, *Colloid. Polym. Sci.* **276** (1998) 1159.
142. S. Hirano, K. Hirochi, K. Hayashi, T. Mikami, H. Tachibana in C. G. Gebelein (ed.): *Cosmetic and Pharmaceutical Polymers,* 1991, pp. 95 – 104.
143. S. Hirano, Y. Akiyama, M. Ogura, Y. Ayaki in S. Tokura, I. Azuma (eds.): *Chitin Derivatives in Life Science,* Jap. Soc. Chitin/Chitosan, Sapporo 1992, pp. 115 – 120.
144. S. Uchida: "Antimicrobial function of chitin and chitosan, and its applications", in M. Yabuki (ed.): *Applications of Chitin and Chitosan,* Ed. Japn Chitin/Chitosan Soc., Gihodo, Tokyo 1990, pp. 71 – 98.
145. A. Tokoro, M. Kobayashi, N. Tatewaki, K. Suzuki, Y. Okawa, T. Mikami, S. Suzuki, M. Suzuki, *Microbiol. Immun.* **33** (1989) 357.
146. J. Iida, T. Une, C. Ishihara, K. Nishimura, S. Tokura, N. Mizukoshi, I. Azuma, *Vaccine* **5** (1987) 270.
147. K. Suzuki, Y. Okawa, K. Hashimoto, S. Suzuki, M. Suzuki, *Microbiol. Immunol.* **28** (1984) 930.
148. K. Ohta, A. Taniguchi, N. Konishi, T. Hosoki, *Fort. Science* **34** (1999) 233.
149. A. Hoekstra, H. Struszczyk, O. Kivekas, *Biomaterials* **19** (1998) 1467.
150. H. Yamamoto, T. Koga, S. Hayakawa, Y. Ohgata, T. Kurasaki, K. Tohyama, *Nihon Sakumotsugakkai Kiji* **67** (1998) 452.
151. K. Nishimura, S. Nishimura, H. Seno, N. Nishi, I. Saiki, S. Tokura, I. Azuma, *Vaccine* **5** (1987) 136.
152. Y. Shibata, L. A. Foster, M. Kurimoto, H. Okamura, R. Nakamura, K. Kwawajiri, J. P. Justice, M. R. Van Scott, N. Myrvik, *J. Immunol.* **161** (1998) 4283.
153. S. Minami, Y. Suzuki, Y. Okamoto, T. Fujinaga, Y. Shigemasa, *Carbohydr. Polym.* **36** (1998) 151.
154. T. Mori, Y. Irie, S. Nishimura, S. Tokura, M. Matsuura, M. Okumura, T. Kadosawam, T. Fujinaga, *J. Biomed. Mater. Res.* **43** (1998) 469.
155. K. Suzuki, T. Mikami, Y. Okawa, A. Tokoro, S. Suzuki, M. Suzuki, *Carbohydr. Res.* **151** (1986) 403.
156. I. Furda, *A. C. S. Symposium Series* **214** (1983) 105.
157. D. J. Ormrod, C. C. Holmes, T. E. Miller, *Atheroseclerosis* **138** (1998) 329.
158. S. Hirano, H. Seino, Y. Akiyama, I. Nonaka in C. G. Gebelein, R. L. Dunn (eds.): *Progress in Biomedical Polymers,* Plenum, New York 1990, pp. 283 – 290.
159. S. Hirano, Y. Akiyama, *J. Sci. Food Agric.* **69** (1995) 91.
160. P. T. Mathew, K. G. R. Nair, *Fish Technol.* **35** (1998) 46.
161. A. E. Ghaouth, J. Arul, R. Ponnampalam, M. Boulet, *J. Food Sci.* **56** (1991) 1618.
162. A. E. Ghaouth, R. Pnnampalm, F. Castaigne, J. Arul, *Host. Sci.* **27** (1992) 1016.
163. A. E. Ghaouth, J. Arul, C. Wilson, N. Benhamou, *Physiol. Mol. Plant Pathol.* **44** (1994) 427.
164. S. Hirano, Y. Yagi, *Carbohydr. Res.* **83** (1980) 103.
165. K. Tomihata, Y. Ikada, Y., *Biomaterials* **18** (1997) 567.
166. S. Hirano, Y. Yagi, *Agric. Biol. Chem.* **44** (1980) 963.

167. S. Hirano, H. Tsuchida, N. Nagao, *Biomaterials* **10** (1989) 574.
168. N. Hutadilok, T. Mochimasu, H. Hisamori, K. Hayashi, H. Tachibana, T. Ishii, S. Hirano in R. A. A. Muzzarelli (ed.): *European Chitin Soc.,* Ancona 1993, pp. 289 – 294.
169. N. Hutadilok, T. Mochimasu, H. Hisamori, K. Hayashi, H. Tachibana, T. Ishii, S. Hirano, *Carbohydr. Res.* **268** (1995) 143.
170. S. Hirano, Y. Yagi, *Carbohydr. Res.* **92** (1981) 319.

Chloral → Chloroacetaldehydes
Chlor-Alkali Electrolysis → Chlorine
Chloramine T → Chloroamines
Chloramphenicol → Antibiotics
Chloranil → Benzoquinone
Chlorates → Chlorine Oxides and Chlorine Oxygen Acids
Chloric Acid → Chlorine Oxides and Chlorine Oxygen Acids
Chlorinated Biphenyls → Chlorinated Hydrocarbons

Chlorinated Hydrocarbons

Manfred Rossberg, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany
Wilhelm Lendle, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany
Gerhard Pfleiderer, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany
Adolf Tögel, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany
Eberhard-Ludwig Dreher, Dow Chemical GmbH, Stade, Federal Republic of Germany
Ernst Langer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
Heinz Rassaerts, Chemische Werke Hüls AG, Marl, Federal Republic of Germany
Peter Kleinschmidt, Bayer AG, Dormagen, Federal Republic of Germany
Heinz Strack, formerly Dynamit Nobel AG,
Richard Cook, ICI Chemicals and Polymers, Runcorn, United Kingdom
Uwe Beck, Bayer AG, Leverkusen, Federal Republic of Germany
Karl-August Lipper, Bayer AG, Krefeld, Federal Republic of Germany
Theodore R. Torkelson, Dow Chemical, Midland, Michigan, United States
Eckhard Löser, Bayer AG, Wuppertal, Federal Republic of Germany
Klaus K. Beutel, Dow Chemical Europe, Horgen, Switzerland
Trevor Mann, INEOS Chlor Limited, Runcorn, United Kingdom (Chap. 7)

1. Chloromethanes 3
1.1. Physical Properties 4
1.2. Chemical Properties 7
1.3. Production 9
1.3.1. Theoretical Bases 9
1.3.2. Production of Monochloromethane 12
1.3.3. Production of Dichloromethane and Trichloromethane 13
1.3.4. Production of Tetrachloromethane . 18
1.4. Quality Specifications 21
1.4.1. Purity of the Commercial Products and Their Stabilization 21
1.4.2. Analysis 22
1.5. Storage, Transport, and Handling 22
1.6. Behavior of Chloromethanes in the Environment 23
1.6.1. Presence in the Atmosphere 24
1.6.2. Presence in Water Sources 24
1.7. Applications of the Chloromethanes and Economic Data 25
2. Chloroethanes 26
2.1. Monochloroethane 29
2.1.1. Physical Properties 29
2.1.2. Chemical Properties 29
2.1.3. Production 30
2.1.4. Uses and Economic Aspects 32
2.2. 1,1-Dichloroethane 32
2.2.1. Physical Properties 32
2.2.2. Chemical Properties 33
2.2.3. Production 33
2.2.4. Uses and Economic Aspects 34
2.3. 1,2-Dichloroethane 34
2.3.1. Physical Properties 34
2.3.2. Chemical Properties 35
2.3.3. Production 35
2.3.4. Uses and Economic Aspects 42
2.4. 1,1,1-Trichloroethane 42
2.4.1. Physical Properties 42
2.4.2. Chemical Properties 42
2.4.3. Production 43
2.4.4. Uses and Economic Aspects 46
2.5. 1,1,2-Trichloroethane 47
2.5.1. Physical Properties 47
2.5.2. Chemical Properties 47
2.5.3. Production 47
2.5.4. Uses and Economic Aspects. 49
2.6. 1,1,1,2-Tetrachloroethane 49
2.6.1. Physical Properties 49
2.6.2. Chemical Properties 49

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_233.pub2

2.6.3. Production 49
2.7. 1,1,2,2-Tetrachloroethane 50
2.7.1. Physical Properties 50
2.7.2. Chemical Properties 50
2.7.3. Production 51
2.7.4. Uses and Economic Aspects 52
2.8. Pentachloroethane 52
2.8.1. Physical Properties 52
2.8.2. Chemical Properties 53
2.8.3. Production 53
2.8.4. Uses and Economic Aspects 53
2.9. Hexachloroethane 53
2.9.1. Physical Properties 54
2.9.2. Chemical Properties 54
2.9.3. Production 54
2.9.4. Uses and Economic Aspects 54
3. Chloroethylenes 54
3.1. Vinyl Chloride (VCM) 55
3.1.1. Physical Properties 55
3.1.2. Chemical Properties 56
3.1.3. Production 56
3.1.3.1. Vinyl Chloride from Acetylene . . . 57
3.1.3.2. Vinyl Chloride from 1,2-Dichloroethane 59
3.1.3.3. Vinyl Chloride from Ethylene by Direct Routes 63
3.1.3.4. Vinyl Chloride from Ethane 64
3.1.3.5. Vinyl Chloride by Other Routes . . 66
3.1.4. Uses and Economic Aspects 66
3.2. 1,1-Dichloroethylene (Vinylidene Chloride, VDC) 67
3.2.1. Physical Properties 67
3.2.2. Chemical Properties 67
3.2.3. Production 67
3.2.4. Uses and Economic Aspects 69
3.3. 1,2-Dichloroethylene 70
3.3.1. Physical Properties 70
3.3.2. Chemical Properties 70
3.3.3. Production 71
3.3.4. Uses and Economic Aspects 71
3.4. Trichloroethylene 71
3.4.1. Physical Properties 71
3.4.2. Chemical Properties 72
3.4.3. Production 72
3.4.4. Uses and Economic Aspects 74
3.5. Tetrachloroethylene 75
3.5.1. Physical Properties 75
3.5.2. Chemical Properties 75
3.5.3. Production 75
3.5.4. Uses and Economic Aspects 79
3.6. Analysis and Quality Control of Chloroethanes and Chloroethylenes 80
3.7. Storage and Transportation of Chloroethanes and Chloroethylenes 80
3.8. Environmental Aspects in the Production of Chloroethanes and Chloroethylenes 81
4. Chloropropanes 82
4.1. 2-Chloropropane 82
4.2. 1,2-Dichloropropane 83
4.3. 1,2,3-Trichloropropane 84
5. Chlorobutanes 85
5.1. 1-Chlorobutane 85
5.2. *tert*-Butyl Chloride 86
5.3. 1,4-Dichlorobutane 86
6. Chlorobutenes 87
6.1. 1,4-Dichloro-2-butene 87
6.2. 3,4-Dichloro-1-butene 87
6.3. 2,3,4-Trichloro-1-butene 88
6.4. 2-Chloro-1,3-butadiene 88
6.4.1. Physical Properties 88
6.4.2. Chemical Properties 89
6.4.3. Production 89
6.4.3.1. Chloroprene from Butadiene 89
6.4.3.2. Chloroprene from Acetylene 90
6.4.3.3. Other Processes 91
6.4.4. Economic Importance 91
6.5. Dichlorobutadiene 91
6.5.1. 2,3-Dichloro-1,3-butadiene 91
6.5.2. Other Dichlorobutadienes 92
6.6. 3-Chloro-2-methyl-1-propene . . . 92
6.6.1. Physical Properties 92
6.6.2. Chemical Properties 92
6.6.3. Production 93
6.6.4. Quality Specifications and Chemical Analysis 94
6.6.5. Storage and Shipment 95
6.6.6. Uses 95
6.7. Hexachlorobutadiene 95
7. Chlorinated Paraffins 96
7.1. Physical Properties 97
7.2. Chemical Properties and Structure 97
7.3. Production 99
7.4. Analysis and Quality Control . . . 101
7.5. Storage and Transportation 101
7.6. Toxicology, Environmental Impact and Regulation 102
7.7. Uses 103
7.8. Summary 104
8. Nucleus-Chlorinated Aromatic Hydrocarbons 104
8.1. Chlorinated Benzenes 105
8.1.1. Physical Properties 105
8.1.2. Chemical Properties 105

8.1.3. Production 109
8.1.3.1. Monochlorobenzene 112
8.1.3.2. Dichlorobenzenes 113
8.1.3.3. Trichlorobenzenes 114
8.1.3.4. Tetrachlorobenzenes 114
8.1.3.5. Pentachlorobenzene 115
8.1.3.6. Hexachlorobenzene 115
8.1.4. Quality and Analysis 115
8.1.5. Storage and Transportation 115
8.1.6. Uses 116
8.2. Chlorinated Toluenes 116
8.2.1. Physical Properties 116
8.2.2. Chemical Properties 117
8.2.3. Production 117
8.2.3.1. Monochlorotoluenes 121
8.2.3.2. Dichlorotoluenes 121
8.2.3.3. Trichlorotoluenes 122
8.2.3.4. Tetrachlorotoluenes 122
8.2.3.5. Pentachlorotoluene 123
8.2.4. Quality and Analysis 123
8.2.5. Storage and Transportation 123
8.2.6. Uses 124
8.3. Chlorinated Biphenyls 124
8.3.1. Physical and Chemical Properties . 125
8.3.2. Disposal 125
8.3.3. Analysis 126
8.3.4. Storage and Transportation 126
8.3.5. Uses 127
8.4. Chlorinated Naphthalenes 127
8.4.1. Physical Properties 127
8.4.2. Chemical Properties 128
8.4.3. Production 129
8.4.4. Quality and Analysis 129
8.4.5. Storage and Transportation 129
8.4.6. Use 130
8.5. Environmental Protection 130
8.6. Economic Facts 131
9. Side-Chain Chlorinated Aromatic Hydrocarbons 132
9.1. Benzyl Chloride 132
9.1.1. Physical Properties 132
9.1.2. Chemical Properties 133
9.1.3. Production 134
9.1.4. Quality Specifications and Analysis 136
9.1.5. Storage and Transportation 137
9.1.6. Uses 137
9.2. Benzal Chloride 137
9.2.1. Physical Properties 137
9.2.2. Chemical Properties 138
9.2.3. Production 138
9.2.4. Quality Specifications and Analysis 138
9.2.5. Storage and Transportation 138
9.2.6. Uses 139
9.3. Benzotrichloride 139
9.3.1. Physical Properties 139
9.3.2. Chemical Properties 139
9.3.3. Production 140
9.3.4. Quality Specifications and Analysis 140
9.3.5. Storage and Transportation 141
9.3.6. Uses 141
9.4. Side-Chain Chlorinated Xylenes . 141
9.4.1. Physical and Chemical Properties . 141
9.4.2. Production 142
9.4.3. Storage and Transportation 143
9.4.4. Uses 143
9.5. Ring-Chlorinated Derivatives . . . 143
9.6. Economic Aspects 145
10. Toxicology and Occupational Health 145
10.1. Aliphatic Chlorinated Hydrocarbons 145
10.1.1. Chloromethanes 146
10.1.2. Chlorinated C_2 Hydrocarbons . . . 148
10.1.3. Chloropropanes and Chloropropenes 152
10.1.4. Chlorobutadienes 152
10.1.5. Ecotoxicology and Environmental Degradation 152
10.2. Chlorinated Aromatic Hydrocarbons 154
10.2.1. Chlorinated Benzenes 154
10.2.2. Chlorotoluenes 154
10.2.3. Polychlorinated Biphenyls 155
10.2.4. Chlorinated Naphthalenes 156
10.2.5. Benzyl Chloride 156
10.2.6. Benzoyl Chloride 157
10.2.7. Benzotrichloride 157
10.2.8. Side-Chain Chlorinated Xylenes . . 157
11. References 157

1. Chloromethanes

Among the halogenated hydrocarbons, the chlorine derivatives of methane monochloromethane (methyl chloride) [*74-87-3*], dichloromethane (methylene chloride) [*75-09-2*], trichloromethane (chloroform) [*67-66-3*], and tetrachloromethane (carbon tetrachloride) [*56-23-5*] play an important role from both industrial and economic standpoints. These products find broad application not only as important chemical intermediates, but also as solvents.

Historical Development. *Monochloromethane* was produced for the first time in 1835 by J. Dumas and E. Peligot by the reaction of sodium chloride with methanol in the presence of sulfuric acid. M. Berthelot isolated it in 1858 from the chlorination of marsh gas (methane), as did C. Groves in 1874 from the reaction of hydrogen chloride with methanol in the presence of zinc chloride. For a time, monochloromethane was produced commercially from betaine hydrochloride obtained in the course of beet sugar manufacture. The earliest attempts to produce methyl chloride by the chlorination of methane occurred before World War I, with the intent of hydrolyzing it to methanol. A commercial methane chlorination facility was first put into operation by the former Farbwerke Hoechst in 1923. In the meantime, however, a high-pressure methanol synthesis based on carbon monoxide and hydrogen had been developed, as a result of which the opposite process became practical — synthesis of methyl chloride from methanol.

Dichloromethane was prepared for the first time in 1840 by V. Regnault, who successfully chlorinated methyl chloride. It was for a time produced by the reduction of trichloromethane (chloroform) with zinc and hydrochloric acid in alcohol, but the compound first acquired significance as a solvent after it was successfully prepared commercially by chlorination of methane and monochloromethane (Hoechst AG, Dow Chemical Co., and Stauffer Chemical Co.).

Trichloromethane was synthesized independently by two groups in 1831: J. von Liebig successfully carried out the alkaline cleavage of chloral, whereas M. E. Soubeirain obtained the compound by the action of chlorine bleach on both ethanol and acetone. In 1835, J. Dumas showed that trichloromethane contained only a single hydrogen atom and prepared the substance by the alkaline cleavage of trichloroacetic acid and other compounds containing a terminal CCl_3 group, such as β-trichloroacetoacrylic acid. In analogy to the synthetic method of M. E. Soubeirain, the use of hypochlorites was extended to include other compounds containing acetyl groups, particularly acetaldehyde. V. Regnault prepared trichloromethane by chlorination of monochloromethane. Already by the middle of the last century, chloroform was being produced on a commercial basis by using the J. von Liebig procedure, a method which retained its importance until ca. the 1960s in places where the preferred starting materials methane and monochloromethane were in short supply. Today, trichloromethane — along with dichloromethane — is prepared exclusively and on a massive scale by the chlorination of methane and/or monochloromethane. Trichloromethane was introduced into the field of medicine in 1847 by J. Y. Simpson, who employed it as an inhaled anaesthetic. As a result of its toxicologic properties, however, it has since been totally replaced by other compounds (e.g., Halothane).

Tetrachloromethane was first prepared in 1839 by V. Regnault by the chlorination of trichloromethane. Shortly thereafter, J. Dumas succeeded in synthesizing it by the chlorination of marsh gas. H. Kolbe isolated tetrachloromethane in 1843 when he treated carbon disulfide with chlorine in the gas phase. The corresponding liquid phase reaction in the presence of a catalyst, giving CCl_4 and S_2Cl_2, was developed a short time later. The key to economical practicality of this approach was the discovery in 1893 by Müller and Dubois of the reaction of S_2Cl_2 with CS_2 to give sulfur and tetrachloromethane, thereby avoiding the production of S_2Cl_2.

Tetrachloromethane is produced on an industrial scale by one of two general approaches. The first is the methane chlorination process, using methane or mono-chloromethane as starting materials. The other involves either perchlorination or chlorinolysis. Starting materials in this case include C_1 to C_3 hydrocarbons and their chlorinated derivatives as well as Cl-containing residues obtained in other chlorination processes (vinyl chloride, propylene oxide, etc.).

Originally, tetrachloromethane played a role only in the dry cleaning industry and as a fire extinguishing agent. Its production increased dramatically, however, with the introduction of chlorofluoromethane compounds 50 years ago, these finding wide application as non-toxic refrigerants, as propellants for aerosols, as foam-blowing agents, and as specialty solvents.

1.1. Physical Properties

The most important physical properties of the four chloro derivatives of methane are presented in Table 1; Figure 1 illustrates the vapor pressure curves of the four chlorinated methanes.

Table 1. Physical properties of chloromethanes

	Unit	Monochloromethane	Dichloromethane	Trichloromethane	Tetrachloromethane
Formula		CH_3Cl	CH_2Cl_2	$CHCl_3$	CCl_4
M_r		50.49	84.94	119.39	153.84
Melting point	°C	− 97.7	− 96.7	− 63.8	− 22.8
Boiling point at 0.1 MPa	°C	− 23.9	40.2	61.3	76.7
Vapor pressure at 20 °C	kPa	489	47.3	21.27	11.94
Density of liquid at 20 °C	kg/m^3	920 (0.5 MPa)	1328.3	1489	1594.7
Density of vapor at *bp*	kg/m^3	2.558	3.406	4.372	5.508
Enthalpy of formation ΔH^0_{298}	kJ/mol	− 86.0	− 124.7	− 132.0	− 138.1
Specific heat capacity of liquid at 20 °C	kJ kg^{-1} K^{-1}	1.595	1.156	0.980	0.867
Enthalpy of vaporization at *bp*	kJ/mol	21.65	28.06	29.7	30.0
Critical temperature	K	416.3	510.1	535.6	556.4
Critical pressure	MPa	6.68	6.17	5.45	4.55
Cubic expansion coeff. of liquid (0–40 °C)	K^{-1}	0.0022	0.00137	0.001399	0.00116
Thermal conductivity at 20 °C	W K^{-1} m^{-1}	0.1570	0.159	0.1454	0.1070
Surface tension at 20 °C	N/m	16.2×10^{-3}	28.76×10^{-3}	27.14×10^{-3}	26.7×10^{-3}
Viscosity of liquid at 20 °C	Pa · s	2.7×10^{-4} (0.5 MPa)	4.37×10^{-4}	5.7×10^{-4}	13.5×10^{-4}
Refractive index n_D^{20}			1.4244	1.4467	1.4604
Ignition temperature	°C	618	605	–	–
Limits of ignition in air, lower	vol%	8.1	12	–	–
Limits of ignition in air, upper	vol%	17.2	22	–	–
Partition coefficient air/water at 20 °C	$\frac{mg/L(air)}{mg/L(water)}$	0.3	0.12	0.12	0.91

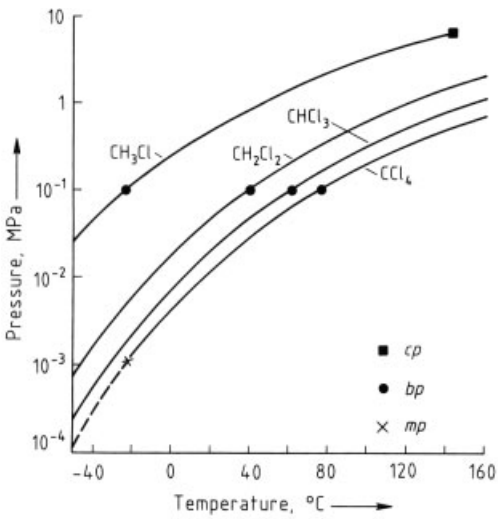

Figure 1. Vapor pressure curves of chloromethanes

The following sections summarize additional important physical properties of the individual compounds making up the chloromethane series.

Monochloromethane is a colorless, flammable gas with a faintly sweet odor. Its solubility in water follows Henry's law; the temperature dependence of the solubility at 0.1 MPa (1 bar) is:

t, °C	15	30	45	60
g of CH_3Cl/kg of H_2O	9.0	6.52	4.36	2.64

Monochloromethane at 20 °C and 0.1 MPa (1 bar) is soluble to the extent of 4.723 cm^3 in 100 cm^3 of benzene, 3.756 cm^3 in 100 cm^3 of tetrachloromethane, 3.679 cm^3 in 100 cm^3 of acetic acid, and 3.740 cm^3 in 100 cm^3 of ethanol. It forms azeotropic mixtures with dimethyl ether, 2-methylpropane, and dichlorodifluoromethane (CFC 12).

Dichloromethane is a colorless, highly volatile, neutral liquid with a slightly sweet smell, similar to that of trichloromethane. The solubility of water in dichloromethane is:

t, °C	− 30	0	+ 25
g of H_2O/kg of CH_2Cl_2	0.16	0.8	1.98

The solubility of dichloromethane in water and in aqueous hydrochloric acid is presented in Table 2.

Dichloromethane forms azeotropic mixtures with a number of substances (Table 3).

Table 2. Solubility of dichloromethane in water and aqueous hydrochloric acid (in wt %)

Solvent	Temperature, °C			
	15	30	45	60
Water	2.50	1.56	0.88	0.53
10 % HCl	2.94	1.85	1.25	0.60
20 % HCl	–	2.45	1.20	0.65

Table 3. Azeotropic mixtures of dichloromethane

wt %	Compound	Azeotropic boiling point, in °C, at 101.3 kPa
30.0	acetone	57.6
11.5	ethanol	54.6
94.8	1,3-butadiene	−5.0
6.0	*tert*-butanol	57.1
30.0	cyclopentane	38.0
55.0	diethylamine	52.0
30.0	diethyl ether	40.80
8.0	2-propanol	56.6
7.3	methanol	37.8
51.0	pentane	35.5
23.0	propylene oxide	40.6
39.0	carbon disulfide	37.0
1.5	water	38.1

Dichloromethane is virtually nonflammable in air, as shown in Figure 2, which illustrates the range of flammable mixtures with oxygen–nitrogen combinations [1, 2]. Dichloromethane thereby constitutes the only nonflammable commercial solvent with a low boiling point. The substance possesses no flash point according to the definitions established in DIN 51 755 and ASTM 56–70 as well as DIN 51 758 and ASTM D 93–73. Thus, it is not subject to the regulations governing flammable liquids. As a result of the existing limits of flammability (CH_2Cl_2 vapor/air), it is assigned to explosion category G 1 (VDE 0165). The addition of small amounts of dichloromethane to flammable liquids (e.g., gasoline, esters, benzene, etc.) raises their flash points; addition of 10–30 % dichloromethane can render such mixtures nonflammable.

Trichloromethane is a colorless, highly volatile, neutral liquid with a characteristic sweet odor. Trichloromethane vapors form no explosive mixtures with air [2]. Trichloromethane has excellent solvent properties for many organic materials, including alkaloids, fats, oils, resins, waxes, gums, rubber, paraffins, etc. As a result of its toxicity, it is increasingly being replaced as a solvent by dichloromethane, whose properties in this general context are otherwise similar. In addition, trichloromethane is a good solvent for iodine and sulfur, and it is completely miscible with many organic solvents. The solubility of trichloromethane in water at 25 °C is 3.81 g/kg of H_2O, whereas 0.8 g of H_2O is soluble in 1 kg of $CHCl_3$.

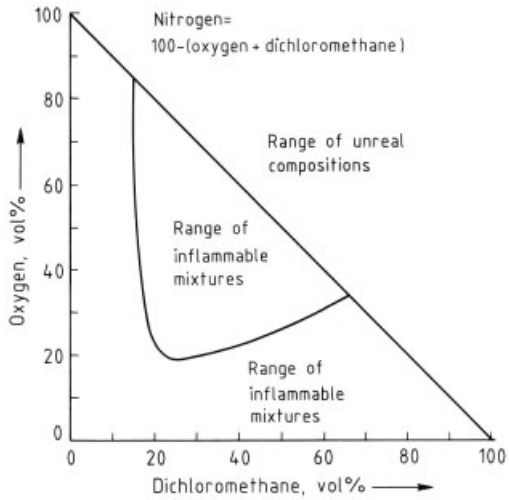

Figure 2. Range of flammability of mixtures of CH_2Cl_2 with O_2 and N_2 [1]

Important azeotropic mixtures of chloroform with other compounds are listed in Table 4.

Table 4. Azeotropic mixtures of trichloromethane

wt %	Compound	Azeotropic boiling point, in °C, at 101.3 kPa
15.0	formic acid	59.2
20.5	acetone	64.5
6.8	ethanol	59.3
13.0	ethyl formate	62.7
96.0	2-butanone	79.7
2.8	*n*-hexane	60.0
4.5	2-propanol	60.8
12.5	methanol	53.4
23.0	methyl acetate	64.8
2.8	water	56.1

Ternary azeotropes also exist between trichloromethane and ethanol – water (boiling point 55.5 °C, 4 mol% ethanol + 3.5 mol% H_2O), methanol – acetone, and methanol – hexane.

Tetrachloromethane is a colorless neutral liquid with a high refractive index and a strong, bitter odor. It possesses good solubility properties for many organic substances, but due to its high toxicity it is no longer employed (e.g., as a spot remover or in the dry cleaning of textiles). It should be noted that it does continue to find application as a solvent for chlorine in certain industrial processes.

Tetrachloromethane is soluble in water at 25 °C to the extent of 0.8 g of CCl_4/kg of H_2O, the solubility of water in tetrachloromethane being 0.13 g of H_2O/kg of CCl_4.

Tetrachloromethane forms constant-boiling azeotropic mixtures with a variety of substances; corresponding data are given in Table 5.

Table 5. Azeotropic mixtures of tetrachloromethane

wt %	Compound	Azeotropic boiling point, in °C, at 101.3 kPa
88.5	acetone	56.4
17.0	acetonitrile	71.0
11.5	allyl alcohol	72.3
81.5	formic acid	66.65
43.0	ethyl acetate	74.8
15.85	ethanol	61.1
71.0	2-butanone	73.8
2.5	butanol	76.6
21.0	1,2-dichloroethane	75.6
12.0	2-propanol	69.0
20.56	methanol	55.7
11.5	propanol	73.1
4.1	water	66.0

1.2. Chemical Properties

Monochloromethane as compared to other aliphatic chlorine compounds, is thermally quite stable. Thermal decomposition is observed only at temperatures in excess of 400 °C, even in the presence of metals (excluding the alkali and alkaline-earth metals). The principal products of photooxidation of monochloromethane are carbon dioxide and phosgene.

Monochloromethane forms with water or water vapor a snowlike gas hydrate with the composition $CH_3Cl \cdot 6\ H_2O$, the latter decomposing into its components at + 7.5 °C and 0.1 MPa (1 bar). To the extent that monochloromethane still finds application in the refrigeration industry, its water content must be kept below 50 ppm. This specification is necessary to prevent potential failure of refrigeration equipment pressure release valves caused by hydrate formation.

Monochloromethane is hydrolyzed by water at an elevated temperature. The hydrolysis (to methanol and the corresponding chloride) is greatly accelerated by the presence of alkali. Mineral acids show no influence on the compound's hydrolytic tendencies.

Monochloromethane is converted in the presence of alkali or alkaline-earth metals, as well as by zinc and aluminum, into the corresponding organometallic compounds (e.g., CH_3MgCl, $Al(CH_3)_3 \cdot AlCl_3$). These have come to play a role both in preparative organic chemistry and as catalysts in the production of plastics.

Reaction of monochloromethane with a sodium – lead amalgam leads to tetramethyllead, an antiknocking additive to gasoline intended for use in internal combustion engines. The use of the compound is declining, however, as a result of ecological considerations.

A very significant reaction is that between monochloromethane and silicon to produce the corresponding methylchlorosilanes (the Rochow synthesis), e.g.:

$$2CH_3Cl + Si \rightarrow SiCl_2(CH_3)_2$$

The latter, through their subsequent conversion to siloxanes, serve as important starting points for the production of silicones.

Monochloromethane is employed as a component in the Wurtz-Fittig reaction; it is also used in Friedel-Crafts reactions for the production of alkylbenzenes.

Monochloromethane has acquired particularly great significance as a methylating agent: examples include its reaction with hydroxyl groups to give the corresponding ethers (methylcellulose from cellulose, various methyl ethers from phenolates), and its use in the preparation of methyl-substituted amino compounds (quaternary methylammonium compounds for tensides). All of the various methylamines result from its reaction with ammonia. Treatment of CH_3Cl with sodium hydrogensulfide under pressure and at elevated temperature gives methyl mercaptan.

Dichloromethane is thermally stable to temperatures above 140 °C and stable in the presence of oxygen to 120 °C. Its photooxidation produces carbon dioxide, hydrogen chloride, and a small amount of phosgene [3]. Thermal reaction with nitrogen dioxide gives carbon monoxide, nitrogen monoxide, and hydrogen chloride [4]. In respect to most industrial metals (e.g., iron, copper, tin), dichloromethane is stable, exceptions being aluminum, magnesium, and their alloys; traces of phosgene first arise above 80 °C.

Dichloromethane forms a hydrate with water, $CH_2Cl_2 \cdot 17\ H_2O$, which decomposes at 1.6 °C and 21.3 kPa (213 mbar).

No detectable hydrolysis occurs during the evaporation of dichloromethane from extracts or extraction residues. Only on prolonged action of steam at 140 – 170 °C under pressure are formaldehyde and hydrogen chloride produced.

Dichloromethane can be further chlorinated either thermally or photochemically. Halogen exchange leading to chlorobromomethane or dibromomethane can be carried out by using bromine and aluminum or aluminum bromide. In the presence of aluminum at 220 °C and 90 MPa (900 bar), it reacts with carbon monoxide to give chloroacetyl chloride [5]. Warming to 125 °C with alcoholic ammonia solution produces hexamethylenetetramine. Reaction with phenolates leads to the same products as are obtained in the reaction of formaldehyde and phenols.

Trichloromethane is nonflammable, although it does decompose in a flame or in contact with hot surfaces to produce phosgene. In the presence of oxygen, it is cleaved photochemically by way of peroxides to phosgene and hydrogen chloride [6, 7]. The oxidation is catalyzed in the dark by iron [8]. The autoxidation and acid generation can be slowed or prevented by stabilizers such as methanol, ethanol, or amylene. Trichloromethane forms a hydrate, $CHCl_3 \cdot 17\ H_2O$, whose critical decomposition point is + 1.6 °C and 8.0 kPa (80 mbar).

Upon heating with aqueous alkali, trichloromethane is hydrolyzed to formic acid, orthoformate esters being formed with alcoholates. With primary amines in an alkaline medium the isonitrile reaction occurs, a result which also finds use in analytical determinations. The interaction of trichloromethane with phenolates to give salicylaldehydes is well-known as the Reimer-Thiemann reaction. Treatment with benzene under Friedel-Crafts conditions results in triphenylmethane.

The most important reaction of trichloromethane is that with hydrogen fluoride in the presence of antimony pentahalides to give monochlorodifluoromethane (CFC 22), a precursor in the production of polytetrafluoroethylene (Teflon, Hostaflon, PTFE).

When treated with salicylic anhydride, trichloromethane produces a crystalline addition compound containing 2 mol of trichloromethane. This result finds application in the preparation of trichloromethane of the highest purity. Under certain conditions, explosive and shock-sensitive products can result from the combination of trichloromethane with alkali metals and certain other light metals [9].

Tetrachloromethane is nonflammable and relatively stable even in the presence of light and air at room temperature. When heated in air in the presence of metals (iron), phosgene is produced in large quantities, the reaction starting at ca. 300 °C [10]. Photochemical oxidation also leads to phosgene. Hydrolysis to carbon dioxide and hydrogen chloride is the principal result in a moist atmosphere [11]. Liquid tetrachloromethane has only a very minimal tendency to hydrolyze in water at room temperature (half-life ca. 70 000 years) [12].

Thermal decomposition of dry tetrachloromethane occurs relatively slowly at 400 °C even in the presence of the common industrial metals (with the exception of aluminum and other light metals). Above 500 – 600 °C an equilibrium reaction sets in which is shifted significantly to the right above 700 °C and 0.1 MPa (1 bar) pressure. At 900 °C and 0.1 MPa (1 bar), the equilibrium conversion of CCl_4 is $> 70\,\%$ (see Chaps. 3.5, cf. Fig. 6).

Tetrachloromethane forms shock-sensitive, explosive mixtures with the alkali and alkaline-earth metals. With water it forms a hydratelike addition compound which decomposes at + 1.45 °C.

The telomerization of ethylene and vinyl derivatives with tetrachloromethane under pressure and in the presence of peroxides has

acquired a certain preparative significance [13 – 15]:

$$CH_2 = CH_2 + CCl_4 \rightarrow CCl_3 - CH_2 - CH_2Cl$$

The most important industrial reactions of tetrachloromethane are its liquid-phase conversion with anhydrous hydrogen fluoride in the presence of antimony (III/V) fluorides or its gas-phase reaction over aluminum or chromium fluoride catalysts, both of which give the widely used and important compounds trichloromonofluoromethane (CFC 11), dichlorodifluoromethane (CFC 12), and monochlorotrifluoromethane (CFC 13).

1.3. Production

1.3.1. Theoretical Bases

The industrial preparation of chloromethane derivatives is based almost exclusively on the treatment of methane and/or monochloromethane with chlorine, whereby the chlorination products are obtained as a mixture of the individual stages of chlorination:

Thermodynamic equilibrium lies entirely on the side of the chlorination products, so that the distribution of the individual products is essentially determined by kinetic parameters.

Monochloromethane can be used in place of methane as the starting material, where this in turn can be prepared from methanol by using hydrogen chloride generated in the previous processes. The corresponding reaction is:

In this way, the unavoidable accumulation of hydrogen chloride (hydrochloric acid) can be substantially reduced and the overall process can be flexibly tailored to favor the production of individual chlorination products. Moreover, given the ease with which it can be transported and stored, methanol is a better starting material for the chloro derivatives than methane, a substance whose availability is tied to natural gas resources or appropriate petrochemical facilities. There has been a distinct trend in recent years toward replacing methane as a carbon base with methanol.

Methane Chlorination. The chlorination of methane and monochloromethane is carried out industrially by using thermal, photochemical, or catalytic methods [16]. The thermal chlorination method is preferred, and it is also the one on which the most theoretical and scientific investigations have been carried out.

Thermal chlorination of methane and its chlorine derivatives is a radical chain reaction initiated by chlorine atoms. These result from thermal dissociation at 300 – 350 °C, and they lead to successive substitution of the four hydrogen atoms of methane:

$$Cl_2 \longrightarrow 2\,Cl^{\cdot} \quad \text{(initiation step)} \tag{7}$$

$$CH_4 + Cl^{\cdot} \longrightarrow CH_3^{\cdot} + HCl \tag{8a}$$

$$CH_3^{\cdot} + Cl_2 \longrightarrow CH_3Cl + Cl^{\cdot} \tag{8b}$$

chain propagation

$$2\,Cl^{\cdot} + M \longrightarrow Cl_2 + M \quad \text{chain termination} \tag{9}$$

(where M = walls, impurities, O_2)

The conversion to the higher stages of chlorination follows the same scheme [17 – 21]. The thermal reaction of methane and its chlorination products has been determined to be a second-order process:

$$dn\,(Cl_2)\,/dt = k \cdot p\,(Cl_2) \cdot p\,(CH_4)$$

It has further been shown that traces of oxygen strongly inhibit the reaction. Controlling the high heat of reaction in the gas phase (which averages ca. 4200 kJ per m^3 of converted chlorine) at STP is a decisive factor in successfully carrying out the process. In industrial reactors, chlorine conversion first becomes apparent above 250 to 270 °C, but it increases exponentially with increasing temperature [22], and in the region of commercial interest — 350 to 550 °C — the reaction proceeds very rapidly. As a result, it is necessary to initiate the process at a temperature which permits the reaction to proceed by itself, but also to maintain the reaction under adiabatic conditions at the requisite temperature level of 320 – 550 °C dictated by both chemical and technical considerations. If a certain critical temperature is exceeded in the reaction mixture (ca. 550 – 700 °C, dependent both on the residence time in the hot zone and on the materials making up the reactor), decomposition of the metastable methane chlorination products occurs. In that event, the chlorination leads to formation of undesirable byproducts, including highly chlorinated or high molecular mass compounds (tetrachloroethene, hexachloroethane, etc.). Alternatively, the reaction with chlorine can get com-

pletely out of control, leading to the separation of soot and evolution of HCl (thermodynamically the most stable end product). Once such carbon formation begins it acts autocatalytically, resulting in a progressively heavier buildup of soot, which can only be halted by immediate shutdown of the reaction.

Proper temperature control of this virtually adiabatic chlorination is achieved by working with a high methane : chlorine ratio in the range of 6 – 4 : 1. Thus, a recycling system is employed in which a certain percentage of inert gas is maintained (nitrogen, recycled HCl, or even materials such as monochloromethane or tetrachloromethane derived from methane chlorination). In this way, the explosive limits of methane and chlorine are moved into a more favorable region and it becomes possible to prepare the more highly substituted chloromethanes with lower $CH_4 : Cl_2$ ratios.

Figure 3 shows the explosion range of methane and chlorine and how it can be limited through the use of diluents, using the examples of nitrogen, hydrogen chloride, and tetrachloromethane.

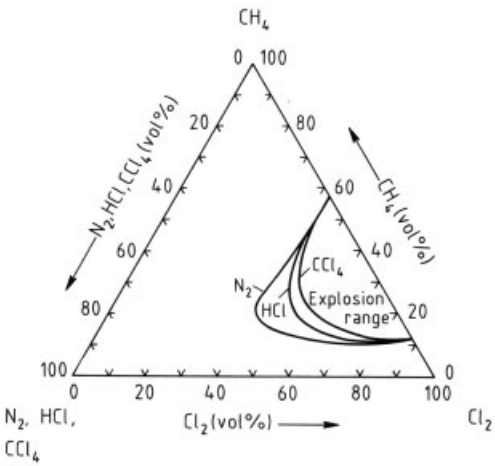

Figure 3. Explosive range of $CH_4 - Cl_2$ mixtures containing N_2, HCl, and CCl_4 Test conditions: pressure 100 kPa; temperature 50 °C; ignition by 1-mm spark

The composition and distribution of the products resulting from chlorination is a definite function of the starting ratio of chlorine to methane, as can be seen from Figure 4 and Figure 5.

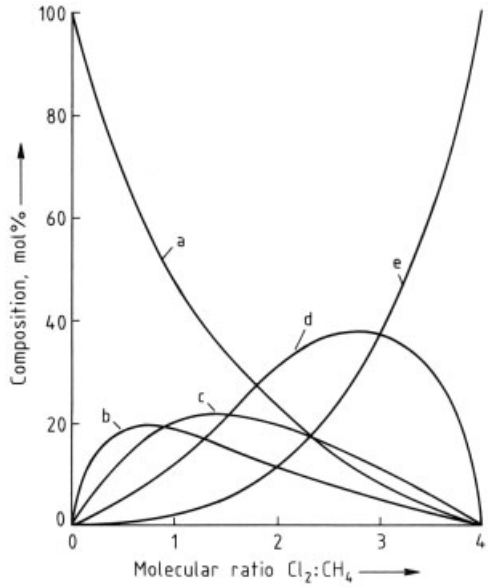

Figure 4. Product distribution in methane chlorination, plug stream reactor
a) Methane; b) Monochloromethane; c) Dichloromethane; d) Trichloromethane; e) Tetrachloromethane

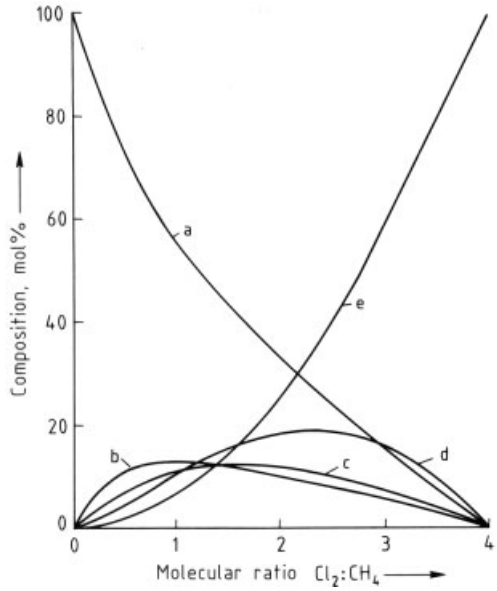

Figure 5. Product distribution in methane chlorination, ideal mixing reactor
a) Methane; b) Monochloromethane; c) Dichloromethane; d) Trichloromethane; e) Tetrachloromethane

These relationships have been investigated frequently [23, 24]. The composition of the reaction product has been shown to be in excellent agreement with that predicted by calculations employing experimental relative reaction rate constants [25 – 28]. The products arising from thermal chlorination of monochloromethane and from the pyrolysis of primary products can also be predicted quantitatively [29]. The relationships among the rate constants are nearly independent of temperature in the region of technical interest. If one designates as k_1 through k_4 the successive rate constants in the chlorination process, then the following values can be assigned to the relative constants for the individual stages:

$k_1 = 1$ (methane)
$k_2 = 2.91$ (monochloromethane)
$k_3 = 2.0$ (dichloromethane)
$k_4 = 0.72$ (trichloromethane)

With this set of values, the selectivity of the chlorination can be effectively established with respect to optimal product distribution for reactors of various residence time (stream type or mixing type, cf. Fig. 4 and Fig. 5). Additional recycling into the reaction of partially chlorinated products (e.g., monochloromethane) permits further control over the ratios of the individual components [30, 31].

It has been recognized that the yield of partially chlorinated products (e.g., dichloromethane and trichloromethane) is diminished by recycling. This factor has to be taken into account in the design of reactors for those methane chlorinations which are intended to lead exclusively to these products. If the emphasis is to lie more on the side of trichloro- and tetrachloromethane, then mixing within the reactor plays virtually no role, particularly since less-chlorinated materials can always be partially or wholly recycled. Details of reactor construction will be discussed below in the context of each of the various processes.

Chlorinolysis. The technique for the production of tetrachloromethane is based on what is known as perchlorination, a method in which an excess of chlorine is used and C_1- to C_3-hydrocarbons and their chlorinated derivatives are employed as carbon sources. In this process, tetrachloroethene is generated along with tetrachloromethane, the relationship between the two being consistent with Eq. 1 in page 13 and dependent on pressure and temperature (cf. also Fig. 6).

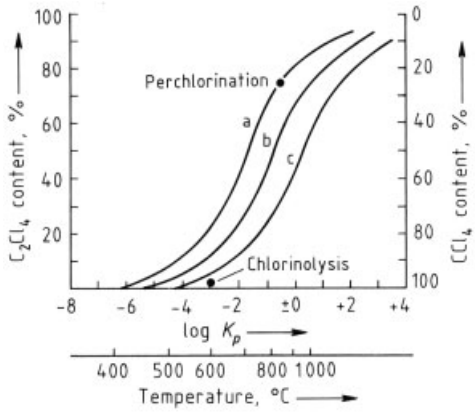

Figure 6. Thermodynamic equilibrium $2\ CCl_4 \rightleftharpoons C_2Cl_4 + 2\ Cl_2$
a) 0.1 MPa; b) 1 MPa; c) 10 MPa

It will be noted that at low pressure (0.1 to 1 MPa, 1 to 10 bar) and temperatures above 700 °C, conditions under which the reaction takes place at an acceptable rate, a significant amount of tetrachloroethene arises. For additional details see Chap. 3.5. Under conditions of high pressure — greater than 10 MPa (100 bar) — the reaction occurs at a temperature as low as 600 °C. As a result of the influence of pressure and by the use of a larger excess of chlorine, the equilibrium can be shifted essentially 100 % to the side of tetrachloromethane. These circumstances are utilized in the Hoechst high-pressure chlorinolysis procedure (see below) [32, 33].

Methanol Hydrochlorination. Studies have been conducted for purposes of reactor design [34] on the kinetics of the gas-phase reaction of hydrogen chloride with methanol in the presence of aluminum oxide as catalyst to give monochloromethane. Aging of the catalyst has also been investigated. The reaction is first order in respect to hydrogen chloride, but nearly independent of the partial pressure of methanol. The rate constant is proportional to the specific surface of the catalyst, whereby at higher temperatures (350 – 400 °C) an inhibition due to pore diffusion becomes apparent.

1.3.2. Production of Monochloromethane

Monochloromethane is produced commercially by two methods: by the hydrochlorination (esterification) of methanol using hydrogen chloride, and by chlorination of methane. Methanol hydrochlorination has become increasingly important in recent years, whereas methane chlorination as the route to monochloromethane as final product has declined. The former approach has the advantage that it utilizes, rather than generating, hydrogen chloride, a product whose disposal — generally as hydrochloric acid — has become increasingly difficult for chlorinated hydrocarbon producers. Moreover, this method leads to a single target product, monochloromethane, in contrast to methane chlorination (cf. Figs. 4 and 5). As a result of the ready and low-cost availability of methanol (via the low pressure methanol synthesis technique) and its facile transport and storage, the method also offers the advantage of avoiding the need for placing production facilities in the vicinity of a methane supply.

Since in the chlorination of methane each substitution of a chlorine atom leads to generation of an equimolar amount of hydrogen chloride — cf. Eqs. 2 – 5 in page 7 — a combination of the two methods permits a mixture of chlorinated methanes to be produced without creating large amounts of hydrogen chloride at the same time; cf. Eq. 6.

Monochloromethane production from methanol and hydrogen chloride is carried out catalytically in the gas phase at 0.3 – 0.6 MPa (3 – 6 bar) and temperatures of 280 – 350 °C. The usual catalyst is activated aluminum oxide. Excess hydrogen chloride is introduced in order to provide a more favorable equilibrium point (located 96 – 99 % on the side of products at 280 – 350 °C) and to reduce the formation of dimethyl ether as a side product (0.2 to 1 %).

The raw materials must be of high purity in order to prolong catalyst life as much as possible. Technically pure (99.9 %) methanol is employed, along with very clean hydrogen chloride. In the event that the latter is obtained from hydrochloric acid, it must be subjected to special purification (stripping) in order to remove interfering chlorinated hydrocarbons.

Process Description. In a typical production plant (Fig. 7), the two raw material streams, hydrogen chloride and methanol, are warmed over heat exchangers and led, after mixing and additional preheating, into the reactor, where conversion takes place at 280 – 350 °C and ca. 0.5 MPa (5 bar).

The reactor itself consists of a large number of relatively thin nickel tubes bundled together and filled with aluminum oxide. Removal of heat generated by the reaction (33 kJ/mol) is accomplished by using a heat conduction system. A hot spot forms in the catalyst layer as a result of the exothermic nature of the reaction, and this migrates through the catalyst packing, reaching the end as the latter's useful life expires.

The reaction products exiting the reactor are cooled with recycled hydrochloric acid (> 30 %) in a subsequent quench system, resulting in separation of byproduct water, removed as ca. 20 % hydrochloric acid containing small amounts of methanol. Passage through a heat exchanger effects further cooling and condensation of more water, as well as removal of most of the excess HCl. The quenching fluid is recovered and subsequently returned to the quench circulation system. The gaseous crude product is led from the separator into a 96 % sulfuric acid column, where dimethyl ether and residual water (present in a quantity reflective of its partial vapor pressure) are removed, the concentration of the acid diminishing to ca. 80 % during its passage through the column. In this step, dimethyl ether reacts with sulfuric acid to form "onium salts" and methyl sulfate. It can be driven out later by further dilution with water. It is advantageous to use the recovered sulfuric acid in the production of fertilizers (superphosphates) or to direct it to a sulfuric acid cleavage facility.

Dry, crude monochloromethane is subsequently condensed and worked up in a high-pressure (2 MPa, 20 bar) distillation column to give pure liquid monochloromethane. The gaseous product emerging from the head of this column ($CH_3Cl + HCl$), along with the liquid distillation residue — together making up ca. 5 – 15 % of the monochloromethane product mixture — can be recovered for introduction into an associated methane chlorination facility. The overall yield of the process, calculated on the basis of methanol, is ca. 99 %.

The commonly used catalyst for vapor-phase hydrochlorination of methanol is γ-aluminum oxide with an active surface area of ca. 200 m^2/g.

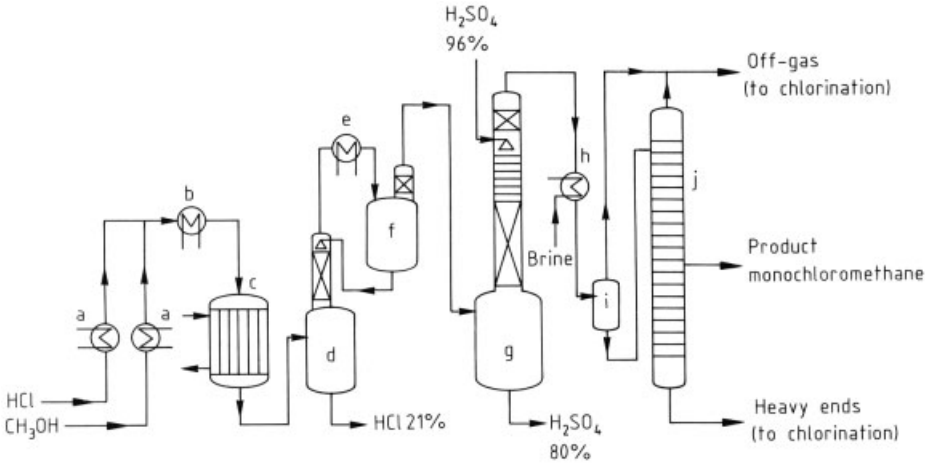

Figure 7. Production of monochloromethane by methanol hydrochlorination
a) Heat exchangers; b) Heater; c) Multiple-tube reactor; d) Quench system; e) Quench gas cooler; f) Quenching fluid tank; g) Sulfuric acid column; h) CH_3Cl condensation; i) Intermediate tank; j) CH_3Cl distillation column

Catalysts based on silicates have not achieved any technical significance. Catalyst aging can be ascribed largely to carbon deposition. Byproduct formation can be minimized and catalyst life considerably prolonged by doping the catalyst with various components and by introduction of specific gases (O_2) into the reaction components [35]. The life of the catalyst in a production facility ranges from about 1 to 2 years.

Liquid-Phase Hydrochlorination. The once common liquid-phase hydrochlorination of methanol using 70 % zinc chloride solution at 130 – 150 °C and modest pressure is currently of lesser significance. Instead, new production techniques involving treatment of methanol with hydrogen chloride in the liquid phase without the addition of catalysts are becoming preeminent. The advantage of these methods, apart from circumventing the need to handle the troublesome zinc chloride solutions, is that they utilize aqueous hydrochloric acid, thus obviating the need for an energy-intensive hydrochloric acid distillation. The disadvantage of the process, which is conducted at 120 – 160 °C, is its relatively low yield on a space – time basis, resulting in the need for large reaction volumes [36 – 38].

Other Processes. Other techniques for producing monochloromethane are of theoretical significance, but are not applied commercially.

Monochloromethane is formed when a mixture of methane and oxygen is passed into the electrolytes of an alkali chloride electrolysis [39]. Treatment of dimethyl sulfate with aluminum chloride [40] or sodium chloride [41] results in the formation of monochloromethane. Methane reacts with phosgene at 400 °C to give CH_3Cl [42]. The methyl acetate – methanol mixture that arises during polyvinyl alcohol synthesis can be converted to monochloromethane with HCl at 100 °C in the presence of catalysts [43]. It has also been suggested that monochloromethane could be made by the reaction of methanol with the ammonium chloride that arises during sodium carbonate production [44].

The dimethyl ether which results from methylcellulose manufacture can be reacted with hydrochloric acid to give monochloromethane [45]. The process is carried out at 80 – 240 °C under sufficient pressure so that water remains as a liquid. Similarly, cleavage of dimethyl ether with antimony trichloride also leads to monochloromethane [46].

In methanolysis reactions for the manufacture of silicones, monochloromethane is recovered and then reintroduced into the process of silane formation [47]:

$$Si + 2CH_3Cl \rightarrow SiCl_2(CH_3)_2 \quad (11)$$

1.3.3. Production of Dichloromethane and Trichloromethane

The industrial synthesis of dichloromethane also leads to trichloromethane and small amounts of

tetrachloromethane, as shown in Figure 4 and Figure 5. Consequently, di- and trichloromethane are prepared commercially in the same facilities. In order to achieve an optimal yield of these products and to ensure reliable temperature control, it is necessary to work with a large methane and/or monochloromethane excess relative to chlorine. Conducting the process in this way also enables the residual concentration of chlorine to be kept in the fully reacted product at an exceptionally low level (< 0.01 vol%), which in turn simplifies workup. Because of the large excess of carbon-containing components, the operation is customarily accomplished in a recycle mode.

Process Description. One of the oldest production methods is that of Hoechst, a recycle chlorination which was introduced as early as 1923 and which, apart from modifications reflecting state-of-the-art technology, continues essentially unchanged, retaining its original importance. The process is shown in Figure 8.

The gas which is circulated consists of a mixture of methane and monochloromethane. To this is added fresh methane and, as appropriate, monochloromethane obtained from methanol hydrochlorination. Chlorine is then introduced and the mixture is passed into the reactor. The latter is a loop reactor coated with nickel or high-alloy steel in which internal gas circulation is constantly maintained by means of a coaxial inlet tube and a valve system. The reaction is conducted adiabatically, the necessary temperature of 350 – 450 °C being achieved and maintained by proper choice of the chlorine to starting material ($CH_4 + CH_3Cl$) ratio and/or by prewarming the mixture [48]. The fully reacted gas mixture is cooled in a heat exchanger and passed through an absorber cascade in which dilute hydrochloric acid and water wash out the resulting hydrogen chloride in the form of 31 % hydrochloric acid. The last traces of acid and chlorine are removed by washing with sodium hydroxide, after which the gases are compressed, dried, and cooled and the reaction products largely condensed. Any uncondensed gas — methane and to some extent monochloromethane — is returned to the reactor. The liquified condensate is separated by distillation under pressure into its pure components, monochloromethane, dichloromethane, trichloromethane (the latter two being the principal products), and small amounts of tetrachloromethane. The product composition is approximately 70 wt % dichloromethane, is approximately 27 wt % trichloromethane, and 3 wt % tetrachloromethane.

Methane chlorination is carried out in a similar way by Chemische Werke Hüls AG, whose work-up process employs prior separation of hydrogen chloride by means of an adiabatic absorption system. After the product gas has been washed to neutrality with sodium hydroxide, it is dried with sulfuric acid and compressed to ca. 0.8 MPa (8 bar), whereby the majority of the resulting chloromethanes can be condensed with relatively little cooling (at approximately − 12 to − 15 °C). Monochloromethane is recycled to the chlorination reactor. The subsequent workup to pure products is essentially analogous to that employed in the Hoechst process.

Other techniques, e.g., those of Montecatini and Asahi Glass, function similarly with respect to drying and distillation of the products.

The loop reactor used by these and other manufacturers (e.g., Stauffer Chem. Co.) [49] has been found to give safe and trouble-free service, primarily because the internal circulation in the reactor causes the inlet gases to be brought quickly to the initiation temperature, thereby excluding the possibility of formation of explosive mixtures. This benefit is achieved at the expense of reduced selectivity in the conduct of the reaction, however (cf. Figs. 4 and 5). In contrast, the use of an empty tube reactor with minimal axial mixing has unquestionable advantages for the selective preparation of dichloromethane [50, 51]. The operation of such a reactor is considerably more complex, however, especially from the standpoint of measurement and control technology, since the starting gases need to be brought up separately to the ignition temperature and then, after onset of the reaction with its high enthalpy, heat must be removed by means of a cooling system. By contrast, maintenance of constant temperature in a loop reactor is relatively simple because of the high rate of gas circulation. A system operated by Frontier Chem. Co. employs a tube reactor incorporating recycled tetrachloromethane for the purpose of temperature control [52].

Reactor Design. Various types of reactors are in use, with characteristics ranging between

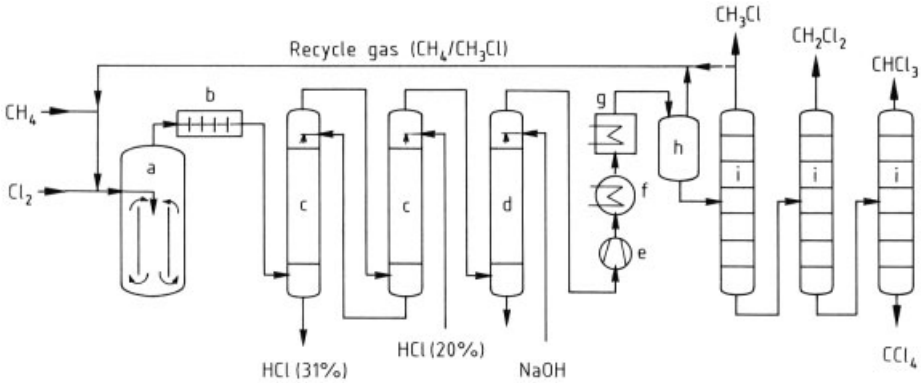

Figure 8. Methane chlorination by the Hoechst method (production of dichloromethane and trichloromethane)
a) Loop reactor; b) Process gas cooler; c) HCl absorption; d) Neutralization system; e) Compressor; f) First condensation step (water); g) Gas drying system; h) Second condensation system and crude product storage vessel (brine); i) Distillation columns for CH_3Cl, CH_2Cl_2, and $CHCl_3$

those of fully mixing reactors (e.g., the loop reactor) and tubular reactors. Chem. Werke Hüls operates a reactor that permits partial mixing, thereby allowing continuous operation with little or no preheating.

Instead of having the gas circulation take place within the reactor, an external loop can also be used for temperature control, as, e.g., in the process described by Montecatini [53] and used in a facility operated by Allied Chemical Corp. In this case, chlorine is added to the reacted gases outside of the chlorination reactor, necessary preheating is undertaken, and only then is the gas mixture led into the reactor.

The space – time yield and the selectivity of the chlorination reaction can be increased by operating two reactors in series, these being separated by a condensation unit to remove high-boiling chloromethanes [54].

Solvay [55] has described an alternative means of optimizing the process in respect to selectivity, whereby methane and monochloromethane are separately chlorinated in reactors driven in parallel. The monochloromethane produced in the methane chlorination reactor is isolated and introduced into the reactor for chlorination of monochloromethane, which is also supplied with raw material from a methanol hydrochlorination system. The reaction is carried out at a pressure of 1.5 MPa (15 bar) in order to simplify the workup and separation of products.

Because of its effective heat exchange characteristics, a fluidized-bed reactor is used by Asahi Glass Co. for methane chlorination [56]. The reaction system consists of two reactors connected in series. After separation of higher boiling components, the low-boiling materials from the first reactor, including hydrogen chloride, are further treated with chlorine in a second reactor. Reactors of this kind must be constructed of special materials with high resistance to both erosion and corrosion. Special steps are required (e.g., washing with liquid chloromethanes) to remove from the reaction gas dust derived from the fluidized-bed solids.

Raw Materials. Very high purity standards must be applied to methane which is to be chlorinated. Some of this methane is derived from petrochemical facilities in the course of naphtha cleavage to ethylene and propene, whereas some comes from low-temperature distillation of natural gas (the Linde process). Components such as ethane, ethylene, and higher hydrocarbons must be reduced to a minimum. Otherwise, these would also react under the conditions of methane chlorination to give the corresponding chlorinated hydrocarbons, which would in turn cause major problems in the purification of the chloromethanes. For this reason, every effort is made to maintain the level of higher hydrocarbons below 100 mL/m^3. Inert gases such as nitrogen and carbon dioxide (but excluding oxygen) have no significant detrimental effect on the thermal chlorination reaction, apart from the fact that their presence in excessive amounts results in the need to eliminate considerable quantities of off-gas from the recycling system, thus caus-

ing a reduction in product yield calculated on the basis of methane introduced.

Chlorine with a purity of ca. 97 % (residue: hydrogen, carbon dioxide, and oxygen) is compressed and utilized just as it emerges from electrolysis. Newer chlorination procedures are designed to utilize gaseous chlorine of higher purity, obtained by evaporization of previously liquified material.

Similarly, monochloromethane destined for further chlorination is a highly purified product of methanol hydrochlorination, special procedures being used to reduce the dimethyl ether content, for example, to less than 50 mL/m^3.

Depending on the level of impurities present in the starting materials, commercial processes incorporating recycling can lead to product yields of 95 – 99 % based on chlorine or 70 – 85 % based on methane. The relatively low methane-based yield is a consequence of the need for removal of inert gases, although the majority of this exhaust gas can be subjected to further recovery measures in the context of some associated facility.

Off-Gas Workup. The workup of off-gas from thermal methane chlorination is relatively complicated as a consequence of the methane excess employed. Older technologies accomplished the separation of the hydrogen chloride produced in the reaction through its absorption in water or azeotropic hydrochloric acid, leading to ordinary commercial 30 – 31 % hydrochloric acid. This kind of workup requires a major outlay for materials of various sorts: on the one hand, coatings must be acid-resistant but at the same time, materials which are stable against attack by chlorinated hydrocarbons are required.

A further disadvantage frequently plagues these "wet" processes is the need to find a use for the inevitable concentrated hydrochloric acid, particularly given that the market for hydrochloric acid is in many cases limited. Hydrogen chloride can be recovered from the aqueous hydrochloric acid by distillation under pressure, permitting its use in methanol hydrochlorination; alternatively, it can be utilized for oxychlorination of ethylene to 1,2-dichloroethane. Disadvantages of this approach, however, are the relatively high energy requirement and the fact that the hydrogen chloride can only be isolated by distillation to the point of azeotrope formation (20 % HCl).

Newer technologies have as their goal workup of the chlorination off-gas by dry methods. These permit use of less complicated construction materials. Apart from the reactors, in which nickel and nickel alloys are normally used, all other apparatus and components can be constructed of either ordinary steel or stainless steel.

Hydrogen chloride can be removed from the off-gas by an absorption – desorption system developed by Hoechst AG and utilizing a wash with monochloromethane, in which hydrogen chloride is very soluble [57]. A similar procedure involving HCl removal by a wash with trichloromethane and tetrachloromethane has been described by Solvay [55].

Other Processes. The relatively complicated removal of hydrogen chloride from methane can be avoided by adopting processes that begin with methanol as raw material. An integrated chlorination/hydrochlorination facility (Fig. 9) has been developed for this purpose and brought on stream on a commercial scale by Stauffer Chem. Co. [58].

Monochloromethane is caused to react with chlorine under a pressure of 0.8 – 1.5 MPa (8 – 15 bar) at elevated temperature (350 – 400 °C) with subsequent cooling occurring outside of the reactor. The crude reaction products are separated in a multistage condensation unit and then worked up by distillation to give the individual pure components. Monochloromethane is returned to the reactor. After condensation, gaseous hydrogen chloride containing small amounts of monochloromethane is reacted with methanol in a hydrochlorination system corresponding to that illustrated in Figure 7 for the production of monochloromethane. Following its compression, monochloromethane is returned to the chlorination reactor. This process is distinguished by the fact that only a minimal amount of the hydrogen chloride evolved during the synthesis of dichloromethane and trichloromethane is recovered in the form of aqueous hydrochloric acid.

As a substitute for thermal chlorination at high temperature, processes have also been developed which occur by a photochemically-initiated radical pathway. According to one

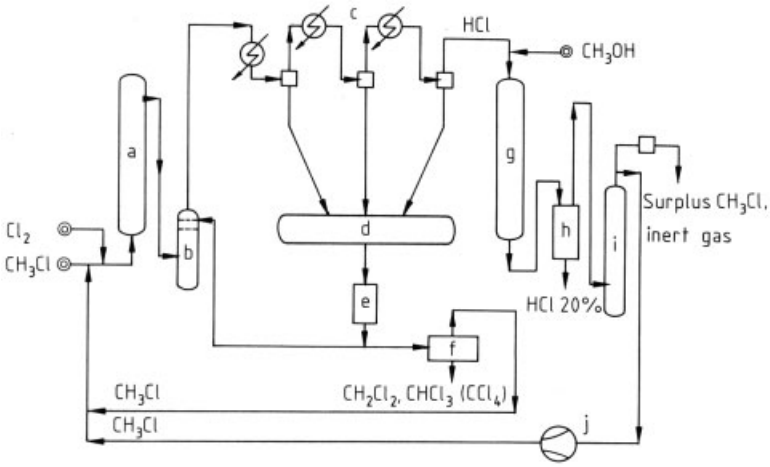

Figure 9. Chlorination of monochloromethane by the Stauffer process [59]
a) Chlorination reactor; b) Quench system; c) Multistage condensation; d) Crude product storage vessel; e) Drying; f) Distillation and purification of CH_2Cl_2 and $CHCl_3$; g) Hydrochlorination reactor; h) Quench system; i) H_2SO_4 drying column; j) Compressor

patent [59], monochloromethane can be chlorinated selectively to dichloromethane at − 20 °C by irradiation with a UV lamp, the trichloromethane content being only 2 – 3 %. A corresponding reaction with methane is not possible.

Liquid-phase chlorination of monochloromethane in the presence of radical-producing agents such as azodiisobutyronitrile has been achieved by the Tokuyama Soda Co. The reaction occurs at 60 – 100 °C and high pressure [60]. The advantage of this low-temperature reaction is that it avoids the buildup of side products common in thermal chlorination (e.g., chlorinated C_2-compounds such as 1,1-dichloroethane, 1,2-dichloroethene, and trichloroethene). Heat generated in the reaction is removed by evaporation of the liquid phase, which is subsequently condensed. Hydrogen chloride produced during the chlorination is used for gas-phase hydrochlorination of methanol to give monochloromethane, which is in turn recycled for chlorination.

It is tempting to try to avoid the inevitable production of hydrogen chloride by carrying out the reaction in the presence of oxygen, as in the oxychlorination of ethylene or ethane. Despite intensive investigations into the prospects, however, no commercially feasible applications have resulted. The low reactivity of methane requires the use of a high reaction temperature, but this in turn leads to undesirable side products and an unacceptably high loss of methane through combustion.

In this context, the "Transcat" process of the Lummus Co. is of commercial interest [61]. In this process, methane is chlorinated and oxychlorinated in two steps in a molten salt mixture comprised of copper(II) chloride and potassium chloride. The starting materials are chlorine, air, and methane. The process leaves virtually no residue since all of its byproducts can be recycled.

Experiments involving treatment of methane with other chlorinating agents (e.g., phosgene, nitrosyl chloride, or sulfuryl chloride) have failed to yield useful results. The fluidized-bed reaction of methane with tetrachloromethane at 350 to 450 °C has also been suggested [62].

The classical synthetic route to trichloromethane proceeded from the reaction of chlorine with ethanol or acetaldehyde to give chloral, which can be cleaved with calcium hydroxide to trichloromethane and calcium formate [63]. Trichloromethane and calcium acetate can also be produced from acetone using an aqueous solution of chlorine bleach at 60 – 65 °C. A description of these archaic processes can be found in [64].

1.3.4. Production of Tetrachloromethane

Chlorination of Carbon Disulfide. The chlorination of carbon disulfide was, until the late 1950s, the principal means of producing tetrachloromethane, according to the following overall reaction:

$$CS_2 + 2Cl_2 \rightarrow CCl_4 + 2S \quad (12)$$

The resulting sulfur is recycled to a reactor for conversion with coal or methane (natural gas) to carbon disulfide. A detailed look at the reaction shows that it proceeds in stages corresponding to the following equations:

$$2CS_2 + 6Cl_2 \rightarrow 2CCl_4 + 2S_2Cl_2 \quad (13)$$

$$CS_2 + 2S_2Cl_2 \rightleftharpoons CCl_4 + 6S \quad (14)$$

The process developed at the Bitterfeld plant of I.G. Farben before World War II was improved by a number of firms in the United States, including FMC and the Stauffer Chem. Co. [65 – 67], particularly with respect to purification of the tetrachloromethane and the resulting sulfur.

In a first step, carbon disulfide dissolved in tetrachloromethane is induced to react with chlorine at temperatures of 30 – 100 °C. Either iron or iron(III) chloride is added as catalyst. The conversion of carbon disulfide exceeds 99 % in this step. In a subsequent distillation, crude tetrachloromethane is separated at the still head. The disulfur dichloride recovered from the still pot is transferred to a second stage of the process where it is consumed by reaction with excess carbon disulfide at ca. 60 °C. The resulting sulfur is separated (with cooling) as a solid, which has the effect of shifting the equilibrium in the reactions largely to the side of tetrachloromethane. Tetrachloromethane and excess carbon disulfide are withdrawn at the head of a distillation apparatus and returned to the chlorination unit. A considerable effort is required to purify the tetrachloromethane and sulfur, entailing hydrolysis of sulfur compounds with dilute alkali and subsequent azeotropic drying and removal from the molten sulfur by air stripping of residual disulfur dichloride. Yields lie near 90 % of the theoretical value based on carbon disulfide and about 80 % based on chlorine. The losses, which must be recovered in appropriate cleanup facilities, result from gaseous emissions from the chlorination reaction, from the purification systems (hydrolysis), and from the molten sulfur processing.

The carbon disulfide method is still employed in isolated plants in the United States, Italy, and Spain. Its advantage is that, in contrast to chlorine substitution on methane or chlorinating cleavage reactions, no accumulation of hydrogen chloride or hydrochloric acid byproduct occurs.

Perchlorination (Chlorinolysis). Early in the 1950s commercial production of tetrachloromethane based on high-temperature chlorination of methane and chlorinating cleavage reactions of hydrocarbons ($\leq C_3$) and their chlorinated derivatives was introduced. In processes of this sort, known as perchlorinations or chlorinolyses, substitution reactions are accompanied by rupture of C – C bonds. Starting materials, in addition to ethylene, include propane, propene, dichloroethane, and dichloropropane. Increasing use has been made of chlorine-containing byproducts and the residues from other chlorination processes, such as those derived from methane chlorination, vinyl chloride production (via either direct chlorination or oxychlorination of ethylene), allyl chloride preparation, etc. The course of the reaction is governed by the position of equilibrium between tetrachloromethane and tetrachloroethene, as illustrated earlier in Figure 6, whereby the latter always arises as a byproduct. In general, these processes are employed for the production of tetrachloroethylene (see Section 3.5.3 and [68]), in which case tetrachloromethane is the byproduct. Most production facilities are sufficiently flexible such that up to 70 wt % tetrachloromethane can be achieved in the final product [69]. The product yield can be largely forced to the side of tetrachloromethane by recycling tetrachloroethylene into the chlorination reaction, although the required energy expenditure is significant. Higher pressure [70] and the use of hydrocarbons containing an odd number of carbon atoms increases the yield of tetrachloromethane. When the reaction is carried out on an industrial scale, a temperature of 500 to 700 °C and an excess of chlorine are used. The corresponding reactors either can be of the tube type, operated adiabatically by using a recycled

coolant (N_2, HCl, CCl_4, or C_2Cl_4) [71 – 73], or else they can be fluidized-bed systems operated isothermally [74, 75]. Byproducts under these reaction conditions include ca. 1 – 7 % perchlorinated compounds (hexachloroethane, hexachlorobutadiene, hexachlorobenzene), the removal of which requires an additional expenditure of effort.

Pyrolytic introduction of chlorine into chlorinated hydrocarbons has become increasingly important due to its potential for consuming chlorinated hydrocarbon wastes and residues from other processes. Even the relatively high production of hydrogen chloride can be tolerated, provided that reactors are used which operate at high pressure and which can be coupled with other processes that consume hydrogen chloride. Another advantage of the method is that it can be used for making both tetrachloromethane and tetrachloroethylene. The decrease in demand for tetrachloromethane in the late 1970s and early 1980s, a consequence of restrictions (related to the ozone hypothesis) on the use of chlorofluorocarbons prepared from it, has led to stagnation in the development of new production capacity.

Hoechst High-Pressure Chlorinolysis. The high-pressure chlorinolysis method developed and put in operation by Hoechst AG has the same goals as the process just described. It can be seen in Figure 6 that under the reaction conditions of this process — 620 °C and 10 to 15 MPa (100 to 150 bar) — the equilibrium

$$2CCl_4 \rightleftharpoons C_2Cl_4 + 2Cl_2$$

lies almost exclusively on the side of tetrachloromethane, especially in the presence of an excess of chlorine [32, 33, 76]. This method utilizes chlorine-containing residues from other processes (e.g., methane chlorination and vinyl chloride) as raw material, although these must be free of sulfur and cannot contain solid or polymerized components.

The conversion of these materials is carried out in a specially constructed high-pressure tube reactor which is equipped with a pure nickel liner to prevent corrosion. Chlorine is introduced in excess in order to prevent the formation of byproducts and in order maintain the final reaction temperature (620 °C) of this adiabatically conducted reaction. If hydrogen-deficient starting materials are to be employed, hydrogen-rich components must be added to increase the enthalpy of the reaction. In this way, even chlorine-containing residues containing modest amounts of aromatics can be utilized. Hexachlorobenzene, for example, can be converted (albeit relatively slowly) at the usual temperature of this process and in the presence of excess chlorine to tetrachloromethane according to the equilibrium reaction:

$$C_6Cl_6 + 9Cl_2 \rightleftharpoons 6CCl_4$$

The mixture exiting the reactor is comprised of tetrachloromethane, the excess chlorine, hydrogen chloride, and small amounts of hexachlorobenzene, the latter being recycled. This mixture is quenched with cold tetrachloromethane, its pressure is reduced, and it is subsequently separated into crude tetrachloromethane and chlorine and hydrogen chloride. The crude product is purified by distillation to give tetrachloromethane meeting the required specifications. This process is advantageous in those situations in which chlorine-containing residues accumulate which would otherwise be difficult to deal with (e.g., hexachloroethane from methane chlorination facilities and high-boiling residues from vinyl chloride production).

A number of serious technical problems had to be overcome in the development of this process, including perfection of the nickel-lined highpressure reactor, which required the design of special flange connections and armatures.

Multistep Chlorination Process. Despite the fact that its stoichiometry results in high yields of hydrogen chloride or hydrochloric acid, thermal chlorination of methane to tetrachloromethane has retained its decisive importance. Recent developments have assured that the resulting hydrogen chloride can be fed into other processes which utilize it. In principle, tetrachloromethane can be obtained as the major product simply by repeatedly returning all of the lower boiling chloromethanes to the reactor. It is not possible to employ a 1 : 4 mixture of the reactants methane and chlorine at the outset. This is true not only because of the risk of explosion, but also because of the impossibility of dealing with the extremely high heat of reaction. Unfortunately, the simple recycling approach is also uneconomical because it necessitates the availability of a very large workup facility. Therefore, it is most advantageous to employ several

reactors coupled in series, the exit gases of each being cooled, enriched with more chlorine, and then passed into the next reactor [77]. Processes employing supplementary circulation of an inert gas (e.g., nitrogen) have also been suggested [78].

The stepwise chlorination of methane and/or monochloromethane to tetrachloromethane is based on a process developed in the late 1950s and still used by Hoechst AG (Fig. 10) [79].

The first reactor in a six-stage reactor cascade is charged with the full amount of methane and/or monochloromethane required for the entire production batch. Nearly quantitative chlorine conversion is achieved in the first reactor at 400 °C, using only a portion of the necessary overall amount of chlorine. The gas mixture leaving the first reactor is cooled and introduced into the second reactor along with additional chlorine, the mixture again being cooled after all of the added chlorine has been consumed. This stepwise addition of chlorine with intermittent cooling is continued until in the last reactor the component ratio $CH_4 : Cl_2 = 1 : 4$ is reached. The reactors themselves are loop reactors with internal circulation, a design which, because of its efficient mixing, effectively shifts the product distribution toward more highly chlorinated materials. The gas mixture leaving the reactors is cooled in two stages to − 20 °C, in the course of which the majority of the tetrachloromethane is liquified, along with the less chlorinated methane derivatives (amounting to ca. 3 % of the tetrachloromethane content). This liquid mixture is then accumulated in a crude product storage vessel.

The residual gas stream is comprised largely of hydrogen chloride but contains small amounts of less highly chlorinated materials. This is subjected to adiabatic absorption of HCl using either water or azeotropic (20 %) hydrochloric acid, whereby technical grade 31 % hydrochloric acid is produced. Alternatively, dry hydrogen chloride can be withdrawn prior to the absorption step, which makes it available for use in other processes which consume hydrogen chloride (e.g., methanol hydrochlorination). The steam which arises during the adiabatic absorption is withdrawn from the head of the absorption column and condensed in a quench system. The majority of the chloromethanes contained in this outflow can be separated by subsequent cooling and phase separation. Wastewater exiting from the quench system is directed to a stripping column where it is purified prior to being discarded. Residual off-gas is largely freed from remaining traces of halogen compounds by low-temperature cooling and are subsequently passed through an off-gas purification system (activated charcoal) before being released into the atmosphere, by which point the gas consists mainly of nitrogen along with traces of methane.

The liquids which have been collected in the crude product containment vessel are freed of gaseous components — Cl_2, HCl, CH_3Cl — by passage through a degassing/dehydrating column, traces of water being removed by distillation. Volatile components are returned to the reaction system prior to HCl absorption. The crude product is then worked up to pure carbon tetrachloride in a multistage distillation facility. Foreruns (light ends) removed in the first column are returned to the appropriate stage of the reactor cascade. The residue in the final column (heavy ends), which constitutes 2 – 3 wt % of the tetrachloromethane production, is made up of hexachloroethane, tetrachloroethylene, trichloroethylene, etc. This material can be converted advantageously to tetrachloromethane in a high-pressure chlorinolysis unit.

Overall yields in the process are ca. 95 % based on methane and > 98 % based on chlorine.

Other Processes. Oxychlorination as a way of producing tetrachloromethane (as well as partially chlorinated compounds) has repeatedly been the subject of patent documents [80 – 82], particularly since it leads to complete utilization of chlorine without any HCl byproduct. Pilot-plant studies using fluidized-bed technology have not succeeded in solving the problem of the high rate of combustion of methane. On the other hand the Transcat process, a two-stage approach mentioned in page 13 and embodying fused copper salts, can be viewed more positively.

Direct chlorination of carbon to tetrachloromethane is thermodynamically possible at atmospheric pressure below 1100 K, but the rate of the reaction is very low because of the high activation energy (lattice energy of graphite). Sulfur compounds have been introduced as catalysts in these experiments. Charcoal can be chlorinated

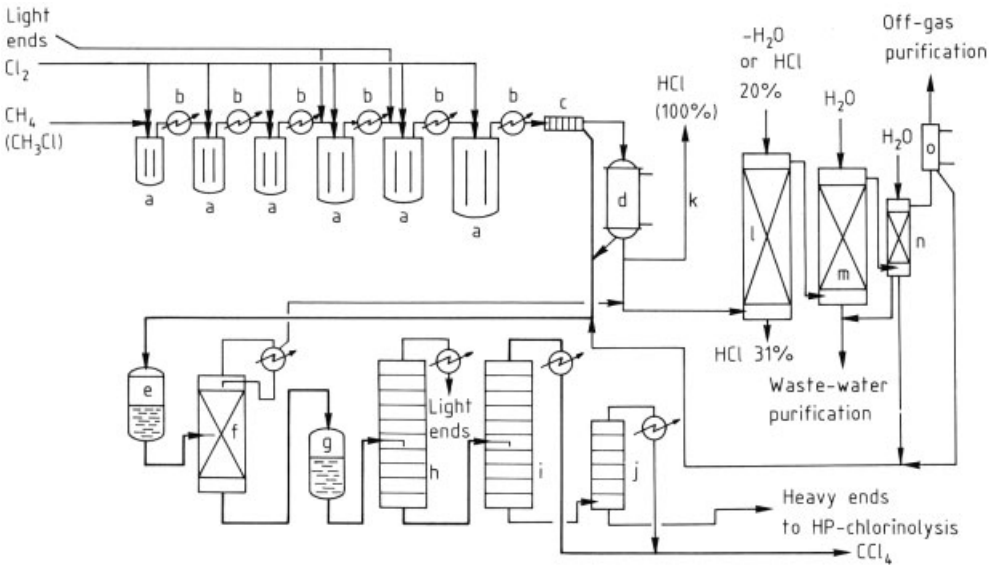

Figure 10. Production of tetrachloromethane by stepwise chlorination of methane (Hoechst process)
a) Reactor; b) Cooling; c) First condensation (air); d) Second condensation (brine); e) Crude product storage vessel; f) Degassing/dewatering column; g) Intermediate tank; h) Light-end column; i) Column for pure CCl_4; j) Heavy-end column; k) HCl stream for hydrochlorination; l) Adiabatic HCl absorption; m) Vapor condensation; n) Cooling and phase separation; o) Off-gas cooler

to tetrachloromethane in the absence of catalyst with a yield of 17 % in one pass at 900 to 1100 K and 0.3 – 2.0 MPa (3 – 20 bar) pressure. None of these suggested processes has been successfully introduced on an industrial scale. A review of direct chlorination of carbon is found in [83].

In this context it is worth mentioning the dismutation of phosgene

$$2COCl_2 \rightarrow CCl_4 + CO_2$$

another approach which avoids the formation of hydrogen chloride. This reaction has been studied by Hoechst [84] and occurs in the presence of 10 mol% tungsten hexachloride and activated charcoal at 370 to 430 °C and a pressure of 0.8 MPa. The process has not acquired commercial significance because the recovery of the WCl_6 is very expensive.

1.4. Quality Specifications

1.4.1. Purity of the Commercial Products and Their Stabilization

The standard commercial grades of all of the chloromethanes are distinguished by their high purity (> 99.9 wt %). Dichloromethane, the solvent with the broadest spectrum of applications, is also distributed in an especially pure form (> 99.99 wt %) for such special applications as the extraction of natural products.

Monochloromethane and tetrachloromethane do not require the presence of any stabilizer. Dichloromethane and trichloromethane, on the other hand, are normally protected from adverse influences of air and moisture by the addition of small amounts of efficient stabilizers. The following substances in the listed concentration ranges are the preferred additives:

Ethanol	0.1 – 0.2 wt %
Methanol	0.1 – 0.2 wt %
Cyclohexane	0.01 – 0.03 wt %
Amylene	0.001 – 0.01 wt %

Other substances have also been described as being effective stabilizers, including phenols, amines, nitroalkanes, aliphatic and cyclic ethers, epoxides, esters, and nitriles.

Trichloromethane of a quality corresponding to that specified in the Deutsche Arzneibuch, 8th edition (D.A.B. 8), is stabilized with 0.6 – 1 wt % ethanol, the same specifications as appear in the British Pharmacopoeia (B.P. 80). Tri-

chloromethane is no longer included as a substance in the U.S. Pharmacopoeia, it being listed only in the reagent index and there without any specifications.

1.4.2. Analysis

Table 6 lists those classical methods for testing the purity and identity of the chloromethanes that are most important to both producers and consumers. Since the majority of these are methods with universal applicability, the corresponding Deutsche Industrie Norm (DIN) and American Society for the Testing of Materials (ASTM) recommendations are also cited in the Table.

Table 6. Analytical testing methods for chloromethanes

Parameter	Method	
	DIN	ASTM
Boiling range	51 751	D 1078
Density	51 757	D 2111
Refraction index	53 491	D 1218
Evaporation residue	53 172	D 2109
Color index (Hazen)	53 409	D 1209
Water content (K. Fischer)	51 777	D 1744
pH value in aqueous extract	–	D 2110

Apart from these test methods, gas chromatography is also employed for quality control both in the production and shipment of chloromethanes. Gas chromatography is especially applicable to chloromethanes due to their low boiling point. Even a relatively simple chromatograph equipped only with a thermal conductivity (TC) detector can be highly effective at detecting impurities, usually with a sensitivity limit of a few parts per million (mg/kg).

1.5. Storage, Transport, and Handling

Dry *monochloromethane* is inert with respect to most metals, thus permitting their presence during its handling. Exceptions to this generalization, however, are aluminum, zinc, and magnesium, as well as their alloys, rendering these unsuitable for use. Thus most vessels for the storage and transport of monochloromethane are preferentially constructed of iron and steel.

Since it is normally handled as a compressed gas, monochloromethane must, in the Federal Republic of Germany, be stored in accord with Accident Prevention Regulation (Unfallverhütungsvorschrift, UVV) numbers 61 and 62 bearing the title "Gases Which Are Compressed, Liquified, or Dissolved Under Pressure" ("Verdichtete, verflüssigte, oder unter Druck gelöste Gase") and issued by the Trade Federation of the Chemical Industry (Verband der Berufsgenossenschaften der chemischen Industrie). Additional guidelines are provided by general regulations governing high-pressure storage containers. Stored quantities in excess of 500 t also fall within the jurisdiction of the Emergency Regulations (Störfallverordnung) of the German Federal law governing emission protection.

Gas cylinders with a capacity of 40, 60, 300, or 700 kg are suitable for the transport of smaller quantities of monochloromethane. Shut-off valves on such cylinders must be left-threaded. Larger quantities are shipped in containers, railroad tank cars, and tank trucks, these generally being licensed for a working pressure of 1.3 MPa (13 bar).

The three *liquid chloromethanes* are also normally stored and transported in vessels constructed of iron or steel. The most suitable material for use with products of very high purity is stainless steel (material no. 1.4 571). The use in storage and transport vessels of aluminum and other light metals or their alloys is prevented by virtue of their reactivity with respect to the chloromethanes.

Storage vessels must be protected against the incursion of moisture. This can be accomplished by incorporating in their pressure release systems containers filled with drying agents such as silica gel, aluminum oxide, or calcium chloride. Alternatively, the liquids can be stored under a dry, inert gas. Because of its very low boiling point, dichloromethane is sometimes stored in containers provided either with external water cooling or with internal cooling units installed in their pressure release systems.

Strict specifications with respect to safety considerations are applied to the storage and transfer of chlorinated hydrocarbons in order to prevent spillage and overfilling. Illustrative is the document entitled "Rules Governing Facilities for the Storage, Transfer, and Preparation for

Shipment of Materials Hazardous to Water Supplies" ("Verordnung für Anlagen zum Lagern, Abfüllen und Umschlagen wassergefährdender Stoffe", VAwS). Facilities for this purpose must be equipped with the means for safely recovering and disposing of any material which escapes [94].

Shipment of solvents normally entails the use of one-way containers (drums, barrels) made of steel and if necessary coated with protective paint. Where product quality standards are unusually high, especially as regards minimal residue on evaporation, stainless steel is the material of choice.

Larger quantities are shipped in containers, railroad tank cars, tank trucks, and tankers of both the transoceanic and inland-waterway variety. So that product specifications may be met for material long in transit, it is important during initial transfer to ensure high standards of purity and the absence of moisture.

Rules for transport by all of the various standard modes have been established on an international basis in the form of the following agreements: RID, ADR, GGVSee, GGVBinSch, IATA-DGR. The appropriate identification numbers and warning symbols for labeling as hazardous substances are collected in Table 7.

Table 7. Identification number and hazard symbols of chloromethanes

Product	Identification number	Hazard symbol
Monochloromethane	UN 1063	H (harmful) IG (inflammable gas)
Dichloromethane	UN 1593	H (harmful)
Trichloromethane	UN 1888	H (harmful)
Tetrachloromethane	UN 1846	P (poison)

The use and handling of chloromethanes — both by producers and by consumers of the substances and mixtures containing them — are governed in the Federal Republic of Germany by regulations collected in the February 11, 1982 version of the "Rules Respecting Working Materials" ("Arbeitsstoff-Verordnung"). To some extent, at least, these have their analogy in other European countries as well. Included are stipulations regarding the labeling of the pure substances themselves as well as of preparations containing chloromethane solvents. The central authorities of the various industrial trade organizations issue informational and safety brochures for chlorinated hydrocarbons, and these should be studied with care.

The standard guidelines for handling monochloromethane as a compressed gas are the "Pressure Vessel Regulation" ("Druckbehälter-Verordnung") of February 27, 1980, with the related "Technical Rules for Gases" ("Technische Regeln Gase", TRG) and the "Technical Rules for Containers" ("Technische Regeln Behälter", TRB), as well as "Accident Prevention Guideline 29 — Gases" ("Unfallverhütungsvorschrift [UVV] 29, Gase").

For MAK values, TLV values, and considerations concerning the toxicology see Chap. 10. The ecology and the ecotoxicology of the chloromethanes are described in Chapter 10.1.5.

1.6. Behavior of Chloromethanes in the Environment

Chloromethanes are introduced into the environment from both natural and anthropogenic sources. They are found in the lower atmosphere, and tetrachloromethane can even reach into the stratosphere. Trichloromethane and tetrachloromethane can be detected in many water supplies.

The chloromethanes, like other halogenated hydrocarbons, are viewed as water contaminants. Thus, they are found in both national and international guidelines related to water quality protection [85, 86].

There are fundamental reasons for needing to restrict chlorocarbon emissions to an absolute minimum. Proven methods for removal of chloromethanes from wastewater, off-gas, and residues are

Vapor stripping with recycling
Adsorption on activated charcoal and recycling
Recovery by distillation
Reintroduction into chlorination processes [87]
Combustion in facilities equipped with offgas cleanup

Table 8. Atmospheric concentration of chloromethanes (in 10^{-10}vol.%) [90]

Compound	Continents	Oceans	Urban areas
CH_3Cl	530 . . . 1040	1140 . . . 1260	834
CH_2Cl_2	36	35	<20 . . . 144
$CHCl_3$	9 . . . 25	8 . . . 40	6 . . . 15 000
CCl_4	20 . . . 133	111 . . . 128	120 . . . 18 000

Table 9. Velocity of decomposition of chloromethanes in the atmosphere [88]

Compound	Reaction velocity with OH radicals $k_{OH} \times 10^{12}$ cm^3 molecule^{-1} s^{-1}	Half-life, weeks
CH_3Cl	0.14	12
CH_2Cl_2	0.1	15
$CHCl_3$	0.1	15
CCl_4	<0.001	>1000

1.6.1. Presence in the Atmosphere

All four chloromethanes are emitted to the atmosphere from anthropogenic sources. In addition, large quantities of monochloromethane are released into the atmosphere by the combustion of plant residues and through the action of sunlight on algae in the oceans. Estimates of the extent of nonindustrial generation of monochloromethane range from 5×10^6 t/a [88] to 28×10^6t/a [89].

Natural sources have also been considered for trichloromethane [90] and tetrachloromethane [91] on the basis of concentration measurements in the air and in seawater (Table 8).

The emission of chloromethanes from industry is the subject of legal restrictions in many countries. The applicable regulations in the Federal Republic of Germany are those of the TA Luft [92].

The most important sink for many volatile organic compounds is their reaction in the lower atmosphere with photochemically generated OH radicals. The reactivity of monochloromethane, dichloromethane, and trichloromethane with OH radicals is so high that in the troposphere these substances are relatively rapidly destroyed.

By contrast, the residence time in the troposphere of tetrachloromethane is very long, with the result that it can pass into the stratosphere, where it is subjected to photolysis from hard ultraviolet radiation. The Cl atoms released in this process play a role in the ozone degradation which is presumed to occur in the stratosphere (Table 9) [93].

1.6.2. Presence in Water Sources

Seawater has been found to contain relatively high concentrations of monochloromethane (5.9 – 21×10^{-9} mL of gas/mL of water) [89], in addition to both trichloromethane (8.3 – 14×10^{-9} g/L) [90] and tetrachloromethane (0.17 – 0.72×10^{-9} g/L) [90]. Dichloromethane, on the other hand, could not be detected [88].

Chloromethanes can penetrate both surface and groundwater through the occurrence of accidents or as a result of improper handling during production, transportation, storage, or use (Table 10). Groundwater contamination by rain which has washed chlorinated hydrocarbons out of the air is not thought to be significant on the basis of current knowledge. One frequent additional cause of diffuse groundwater contamination that can be cited is defective equipment (especially leaky tanks and wastewater lines) [94].

The chloromethanes are relatively resistant to hydrolysis. Only in the case of monochloromethane in seawater is abiotic degradation of significance, this compound being subject in weakly alkaline medium to cleavage with the elimination of HCl.

The microbiological degradability of dichloromethane has been established [97 – 103]. This is understood to be the reason for the absence or only very low concentrations of dichloromethane in the aquatic environment [94].

Since trichloromethane and tetrachloromethane are stable compounds with respect to both biotic and abiotic processes, their disappearance is thought to be largely a consequence of transfer into the atmosphere by natural stripping phenomena.

Treatment with chlorine is a widespread technique for disinfecting drinking water. In the process, trihalomethanes result, largely trichloromethane as a result of the reaction of chlo-

Table 10. Chloromethane concentration in the Rhine river (µg/L) [95, 96]

Compound	Date	Mean value	Max. value
CH_2Cl_2	1980	not detected	
$CHCl_3$	1980	4.5	
$CHCl_3$	1982	0.4 . . . 12.5	50.0
CCl_4	1982	<0.1 . . . 3.3	44.4

rine with traces of organic material. A level of 25 µg/L of trihalomethanes is regarded in the Federal Republic of Germany as the maximum acceptable annual median concentration in drinking water [104].

1.7. Applications of the Chloromethanes and Economic Data

As a result of very incomplete statistical records detailing production and foreign trade by individual countries, it is very difficult to describe precisely the world market for chloromethanes. The information which follows is based largely on systematic evaluation of the estimates of experts, coupled with data found in the secondary literature, as well as personal investigations and calculations.

The Western World includes about 40 producers who produce at least one of the chlorinated C_1 hydrocarbons. No authoritative information is available concerning either the production capacity or the extent of its utilization in the Comecon nations or in the People's Republic of China. It can be assumed, however, that a large part of the domestic requirements in these countries is met by imports. In reference to production capacity, see [105].

In comparing the reported individual capacities it is important to realize that a great many facilities are also capable of producing other chlorinated hydrocarbons. This situation is a result of the opportunities for flexibility both in the product spectrum (cf. Sect. 1.3.1) and in the various manufacturing techniques (e.g., tetrachloromethane/tetrachloroethene, cf. Sect. 1.3.4). If one ignores captive use for further chlorination (especially of monochloromethane), it can be concluded that the largest portion of the world use of chloromethanes (ca. 34 %) can be attributed to *tetrachloromethane*. The most important market, accounting for over 90 % of the material produced, is that associated with the production of the fluorochlorocarbons CFC 11 (trichloromonofluoromethane) and CFC 12 (dichlorodifluoromethane). These fluorochlorocarbons possess outstanding properties, such as nonflammability and toxicological safety, and are employed as refrigerants, foaming agents, aerosol propellants, and special solvents.

The production level of tetrachloromethane is directly determined by the market for its fluorinated reaction products CFC 11 and CFC 12. The appearance of the so-called ozone theory, which asserts that the ozone layer in the stratosphere is affected by these compounds, has resulted since 1976 in a trend toward reduced production of tetrachloromethane. This has been especially true since certain countries (United States, Canada, Sweden) have imposed a ban on aerosol use of fully halogenated fluorochlorocarbons. However, since 1982/1983 there has been a weak recovery in demand for tetrachloromethane in the production of fluorine-containing compounds.

Outside Europe, a smaller amount of tetrachloromethane finds use as a disinfectant and as a fungicide for grain.

Monochloromethane and *dichloromethane* each account for about 25 % of the world market for chloromethanes (Table 11). The demand for monochloromethane can be attributed largely (60 – 80 %) to the production of silicones. Its use as a starting material for the production of the gasoline anti-knock additive tetramethyllead is in steep decline.

The most important use of dichloromethane, representing ca. 40 – 45 % of the total market, is as a cleaning agent and paint remover. An additional 20 – 25 % finds application as a pressure mediator in aerosols. One further use of dichloromethane is in extraction technology (decaffeination of coffee, extraction of hops, paraffin extraction, and the recovery of specialty pharmaceuticals).

In all of these applications, especially those related to the food and drug industries, the purity level requirements for dichloromethane are exceedingly high (> 99.99 wt %).

Table 11. Production capacities of chloromethanes 1000 t/year [115]

	Western Europe (FRG)		United States		Japan	
	1981	1993	1981	1993	1981	1993
Monochloromethane	265	295 (100)	300	274	70	106
Dichloromethane	410	237 (170)	370	161	65	86
Trichloromethane	140	247 (60)	210	226	65	53
Tetrachloromethane	250	182 (150)	380	140	70	40

Trichloromethane holds the smallest market share of the chloromethane family: 16 %. Its principal application, amounting to more than 90 % of the total production, is in the production of monochlorodifluoromethane (CFC 22), a compound important on the one hand as a refrigerant, but also a key intermediate in the preparation of tetrafluoroethene. The latter can be polymerized to give materials with exceptional thermal and chemical properties, including PTFE, Hostaflon, Teflon, etc.

Chloroform is still used to a limited extent as an extractant for pharmaceutical products. Due to its toxicological properties, its use as an inhalatory anaesthetic is no longer significant. Small amounts are employed in the synthesis of orthoformic esters.

Table 12 provides an overview of the structure of the markets for the various chlorinated C_1 compounds, subdivided according to region.

2. Chloroethanes

The class of chloroethanes comprises:

Monochloroethane (ethyl chloride)
1,1-Dichloroethane
1,2-Dichloroethane (ethylene dichloride, EDC)
1,1,1-Trichloroethane
1,1,2-Trichloroethane
1,1,1,2-Tetrachloroethane (asymmetric Tetra)
1,1,2,2-Tetrachloroethane (symmetric Tetra)
Pentachloroethane
Hexachloroethane

As with other chlorinated hydrocarbons, a common characteristic of these compounds is their low solubility in water and excellent solubility in most organic solvents, their volatility with water vapor, and their tendency to form azeotropic mixtures.

Boiling point, density, viscosity, and surface tension increase with increasing chlorine substitution, whereas heat of formation, solubility, and inflammability decrease. Substitution of vicinal positions generally has a greater impact than geminal substitution, as shown in Table 13. With the exception of gaseous monochloroethane (ethyl chloride) and solid hexachloroethane, the chloroethanes are colorless liquids at ambient conditions. Figure 11 shows the vapor pressure as a function of temperature. If light and oxygen are absent, most chlorinated ethanes are fairly stable. At higher temperatures (> 300 °C), they are susceptible to the elimination of hydrogen chloride. In the presence of light and oxygen, oxidation occurs yielding phosgene, carbon oxides, and acetyl or chloroacetyl chlorides. The latter easily hydrolyze with traces of moisture forming the corresponding chloroacetic acids, which are well-known as strongly corrosive agents. To prevent this unwanted decomposition, most industrially used chlorinated hydrocarbons are stabilized with acid acceptors such as amines, unsaturated hydrocarbons, ethers, epoxides or phenols, antioxidants, and other compounds able to inhibit free radical chain reactions. Longer storage periods and use without appreciable effect on tanks and equipment is then possible.

Of all chlorinated ethanes, approximately half are of industrial importance. Monochloroethane (ethyl chloride) is an intermediate in the production of tetraethyllead and is widely used as an ethylating agent. 1,2-Dichloroethane has by far the highest production rates. It is an intermediate for the production of 1,1,1-trichloroethane and vinyl chloride (see page 43 and 3.1.3.2), but it is also used in syn-

Table 12. Demand and use pattern of chloromethanes (1983)

	Western Europe	United States	Japan
Monochloromethane	230 000 t	250 000 t	50 000 t
Silicone	52 %	60 %	83 %
Tetramethyllead	12 %	15 %	–
Methylcellulose	15 %	5 %	1 %
Other methylation reactions, e.g., tensides, pharmaceuticals	ca. 21 %	ca. 20 %	ca. 16 %
Dichloromethane	210 000 t	270 000 t	35 000 t
Degreasing and paint remover	46 %	47 %	54 %
Aerosols	18 %	24 %	19 %
Foam-blowing agent	9 %	4 %	11 %
Extraction and other uses	27 %	25 %	16 %
Trichloromethane	90 000 t	190 000 t	45 000 t
CFC 22 production	78 %	90 %	90 %
Other uses, e.g., pharmaceuticals, intermediate	22 %	10 %	10 %
Tetrachloromethane	250 000 t	250 000 t	75 000 t
CFC 11/12 production	94 %	92 %	90 %
Special solvent for chemical reactions	6 %	8 %	10 %

thetic applications (e.g., polyfunctional amines) and as a fuel additive (lead scavenger).

Table 13. Physical properties of chlorinated ethanes

Compound	Boiling point (at 101 kPa), °C	Relative density, d_4^{20}
Monochloroethane	12.3	0.9240
1,1-Dichloroethane	57.3	1.1760
1,2-Dichloroethane	83.7	1.2349
1,1,1-Trichloroethane	74.1	1.3290
1,1,2-Trichloroethane	113.5	1.4432
1,1,1,2-Tetrachloroethane	130.5	1.5468
1,1,2,2-Tetrachloroethane	146.5	1.5958
Pentachloroethane	162.0	1.6780
Hexachloroethane	*mp* 186 – 187	2.0940

1,1,1-Trichloroethane, trichloroethylene, (see Section 3.4) and tetrachloroethylene (see Section 3.5) are important solvents widely used in dry cleaning, degreasing, and extraction processes.

The other chlorinated ethanes have no important end uses. They are produced as intermediates (e.g., 1,1-dichloroethane) or are formed as unwanted byproducts. Their economical conversion into useful end products is achieved either by cracking — tetrachloroethanes yield trichloroethylene — or more commonly by chlorinolysis, which converts them into carbon tetrachloride and tetrachloroethylene (see page 76).

Basic feedstocks for the production of chlorinated ethanes and ethylenes (see Chap. 3) are

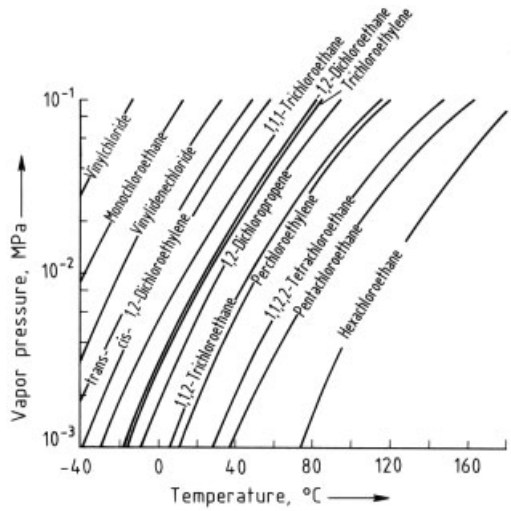

Figure 11. Vapor pressure as a function of temperature for chlorinated hydrocarbons

ethane or ethylene and chlorine (Fig. 12). The availability of ethylene from naphtha feedstocks has shifted the production of chlorinated C_2 hydrocarbons during the past three decades in the Western World from the old carbide–acetylene – vinyl chloride route toward the ethylene route. With the dramatic increase of naphtha prices during the past decade, the old carbide route has regained some of its attractiveness [106]. Even though a change cannot be justified presently in

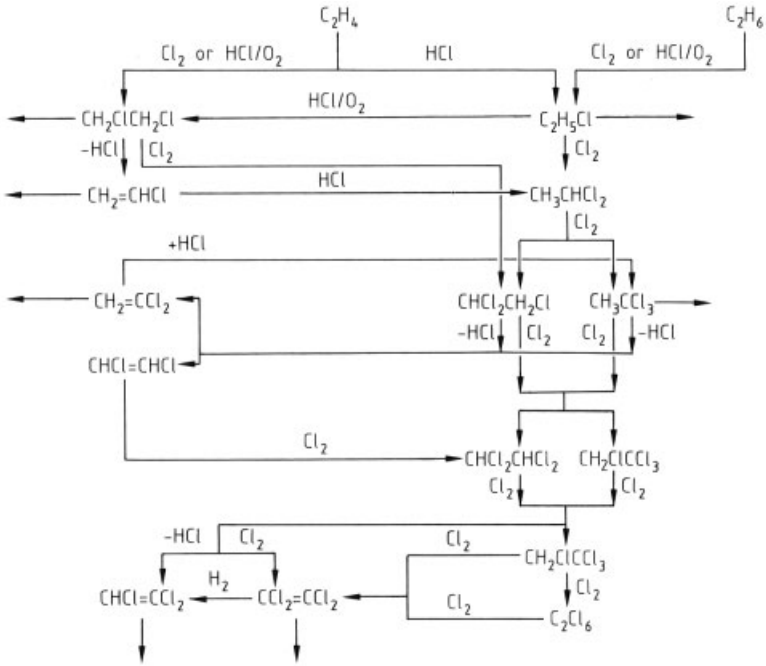

Figure 12. Chlorinated hydrocarbons from ethane and ethylene (simplified)

most countries, it could offer an alternative for countries where cheap coal is readily available. The use of ethanol derived from biomass as a starting material could likewise also be considered [107, 108].

In a few cases, ethane is used directly as a hydrocarbon feedstock. This 'direct' ethane route could offer an attractive alternative in some cases, because of the substantial cost differences between ethane and ethylene. It becomes evident why numerous patents on ethane-based processes have been filed. However, the major cost advantage of such processes is the reduced capital investment for cracker capacity. The direct ethane route must certainly be considered for future grass-root-plants, but at present, the conversions and selectivities obtained seem not to justify the conversion of existing plants if cracker capacity is available.

Less is known about the situation in Eastern block countries. The available information indicates, however, that in some Eastern European countries the acetylene route is still used.

Because chlorine is needed as a second feedstock, most plants producing chlorinated hydrocarbons are connected to a chlor-alkali electrolysis unit. The hydrocarbon feedstock is either supplied from a nearby cracker, — typical for U.S. gulf coast, — or via pipelines and bulk ship transports. The chlorine value of the hydrogen chloride produced as a byproduct in most chlorination processes can be recovered by oxychlorination techniques, hydrochlorination reactions (for synthesis of methyl and ethyl chloride) or, — less economically — by aqueous HCl electrolysis. A minor but highly valuable outlet is ultrapure-grade anhydrous HCl used for etching in the electronic industry.

Although most unwanted byproducts can be used as feed for the chlorinolysis process [109] (see page 76), the byproducts of this process, mostly hexachloroethane, hexachlorobutadiene, and hexachlorobenzene together with residual tars from spent catalysts and vinyl chloride production, represent a major disposal problem. The optimal ecological solution is the incineration of these residues at a temperature above 1200 °C, which guarantees almost complete degradation. Presently, incineration is performed at sea on special ships [110] without HCl scrubbing or on site with subsequent HCl or chlorine recovery. The aqueous HCl recovered can then be used for

pH adjustment in biological effluent treatment or brine electrolysis.

Due to their unique properties, the market for chlorinated C_2 hydrocarbons has shown excellent growth over the past 30 years and reached its maximum in the late 1970s. With increasing environmental consciousness, the production rate of some chlorinated hydrocarbons such as ethyl chloride, trichloroethylene (see page 73), and tetrachloroethylene (see 3.5.4) will in the long run decrease due to the use of unleaded gasoline, solvent recovery systems, and partial replacement by other solvent and extraction chemicals. However, new formulations for growing markets such as the electronic industry, the availability of ecologically safe handling systems, know-how in residue incineration, and the difficulty in finding superior replacements — causing fewer problems — guarantee chlorinated ethanes and ethylenes a long-term and at least constant market share.

2.1. Monochloroethane

Monochloroethane (ethyl chloride) [*75-00-3*] is thought to be the first synthesized chlorinated hydrocarbon. It was produced in 1440 by VALENTINE by reacting ethanol with hydrochloric acid. GLAUBER obtained it in 1648 by reacting ethanol (spirit of wine) with zinc chloride. Because of the growing automotive industry in the early 1920s, monochloroethane became an important bulk chemical. Its use as a starting material for the production of tetraethyl-lead (→ Lead Compounds) initiated a significant increase in ethyl chloride production and is still its major consumer. The trend toward unleaded gasoline in most countries, however, will in the long run lead to a significant decrease in production.

2.1.1. Physical Properties

M_r	64.52
mp	− 138.3 °C
bp at 101.3 kPa	12.3 °C
ϱ of the liquid at 0 °C	0.924 g/cm^3
ϱ of the vapor at 20 °C	2.76 kg/m^3
n_D^{20}	1.3798
Vapor pressure at	
− 50 °C	4.480 kPa
− 20 °C	25.090 kPa
− 10 °C	40.350 kPa
0 °C	62.330 kPa
+ 10 °C	92.940 kPa
+ 20 °C	134.200 kPa
+ 30 °C	188.700 kPa
+ 60 °C	456.660 kPa
+ 80 °C	761.100 kPa
Heat of formation (liquid) ΔH_{298}^0	− 133.94 kJ/mol
Specific heat at 0 °C	1.57 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	24.7 kJ/mol
Critical temperature	456 K
Critical pressure	5270 kPa
Viscosity (liquid, 10 °C)	2.79×10^{-4} Pa s
Viscosity (vapor, *bp*)	9.3×10^{-5} Pa s
Thermal conductivity (vapor)	1.09×10^{-3} W m^{-1} K^{-1}
Surface tension (air, 5 °C)	21.18×10^{-3} N/m
Dielectric constant (vapor, 23.5 °C)	1.0129
Flash point (open cup)	− 43 °C
Ignition temperature	519 °C
Explosive limits in air	3.16 – 15 vol% monochloroethane
Solubility in water at 0 °C	0.455 wt %
Solubility of water in monochloroethane at 0 °C	0.07 wt %

At ambient temperature, monochloroethane is a gas with an etheral odor.

Monochloroethane burns with a green-edged flame.

Combustion products are hydrogen chloride, carbon dioxide, and water.

Binary azeotropic mixtures of monochloroethane have been reported [111]. The data, however, have not been validated.

2.1.2. Chemical Properties

Monochloroethane has considerable thermal stability. Only at temperatures above 400 °C, considerable amounts of ethylene and hydrogen chloride are formed due to dehydrochlorination [111]a. This decomposition can be catalyzed by a variety of transition metals (e.g. Pt), transition-metal salts, and high-surface area oxides such as alumina and silica. Catalyzed decomposition is complete at temperatures slightly above 300 °C according to the thermodynamic equilibrium.

At ambient atmospheric conditions, both, hydrolysis (to ethanol) and oxidation (to acetaldehyde) are moderate.

At temperatures up to 100 °C, monochloroethane shows no detrimental effect on most structural materials if kept dry. Contact with aluminum, however, should be avoided under all circumstances for safety reasons.

Monochloroethane has the highest reactivity of all chlorinated ethanes. It is mainly used as an ethylating agent in Grignard- and Friedel-Crafts-type reactions, for ether, thioether, and amine synthesis. Halogene exchange [111]b and fluorination is also possible [111]c.

2.1.3. Production

Monochloroethane can be produced by a variety of reactions. Only two are of industrial importance: the hydrochlorination of ethylene and the thermal chlorination of ethane.

Hydrochlorination of Ethylene. Exothermic hydrochlorination of ethylene can be carried out in either the liquid or gas phase.

$$C_2H_4 + \mathrm{HCl} \rightarrow C_2H_5-\mathrm{Cl}\,\Delta H = -98\,\mathrm{kJ/mol}$$

The liquid-phase reaction is carried out mostly at near ambient temperatures (10 – 50 °C) and moderate pressure (0.1 – 0.5 MPa) in a boiling-bed type reactor. The heat of reaction is used to vaporize part of the monochloroethane formed, which in turn is then cooled down, purified, or partially recycled. The reactor temperature is controlled by the recycle ratio and the feed rate of the reactants. Unconverted ethylene and hydrogen chloride from reflux condensers and overhead light end columns are recycled back to the reactor. Sufficient mixing and catalyst contact time is achieved through recirculation of the reactor sump phase. Aluminum chloride in a 0.5 – 5 wt % concentration is mostly used as a catalyst. A part of it is continuously or intermittently removed via a recirculation slip stream, together with unwanted high boiling impurities consisting mostly of low molecular mass ethylene oligomers formed in a Ziegler-type reaction of the catalyst with the ethylene feed. New catalyst is added to the system either by a hopper as a solid or preferably as a solution after premixing with monochloroethane or monochloroethane/ethylene. A gaseous feed of vaporized $AlCl_3$ has also been suggested [112]. A simplified process diagram is shown in Figure 13; an optimized process has been patented [113]. In other process variations, the formed monochloroethane (sump phase) is washed with diluted NaOH to remove catalyst and acid and then dried and distilled. Excess ethylene is recycled.

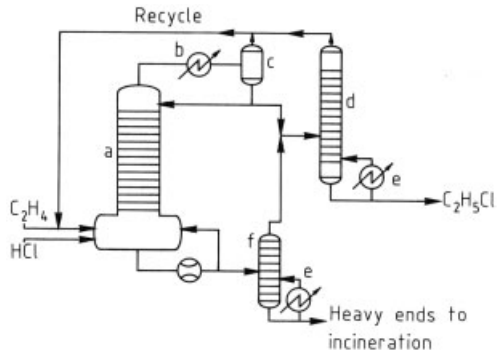

Figure 13. Schematic diagram (simplified) of an ethylene hydrochlorination process
a) Reactor; b) Cooler; c) Knock-out drum; d) Light-end columns; e) Reboiler; f) Stripper column (heavy ends)

Ethylene and HCl yields for hydrochlorination are almost quantitative; selectivities of 98 – 99 % have been reported. In addition to $AlCl_3$, other Lewis-acid catalysts, such as $FeCl_3$ [114], $BiCl_3$ [116], and $GaCl_3$ [117], have been patented. Suggestions to perform the reaction in benzene or higher boiling hydrocarbons [118], in 1,1,2-trichloroethane [119] or to complex $AlCl_3$ by nitrobenzene [120] have not found industrial acceptance.

The troublesome handling of the catalyst is minimized when ethylene and hydrogen chloride are reacted in the gas phase. Although the reaction equilibrium becomes unfavorable at a temperature above 200 °C, the process is carried out at temperatures of 250 – 450 °C in order to achieve sufficient conversion. Ethylene and HCl are preheated, mixed, and sent across the catalyst, which can be used as fixed or fluidized bed. The chloroethane formed is separated and purified. Unreacted ethylene and HCl are recycled. Selectivities are comparable to those of the liquid-phase process, conversion per pass, however, may not exeed 50 %, so that relatively high recycle rates are necessary. Because high pressure favors the formation of monochloroethane, the reaction is preferably carried out at 0.5 – 1.5 MPa.

Thorium oxychloride on silica [121], platinium on alumina [122], and rare-earth oxides on alumina and silica [123] have been patented as catalysts.

Chlorination of Ethane. Thermal chlorination of ethane for the production of monochloroethane can be used industrially in a tandem process developed by the Shell Oil Company (Fig. 14) [124]. This process was especially designed for a plant in which sufficient ethylene feedstock could only be supplied by increasing the cracker capacity. Ethane and chlorine were available, but not hydrogen chloride. For this feedstock constellation, the tandem process seems advantageous.

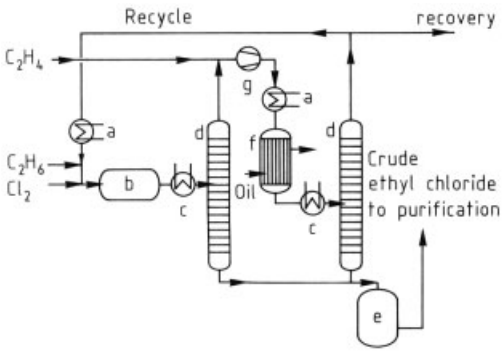

Figure 14. Production of monochloroethane by the Shell process [124]
a) Preheater; b) Ethane chlorinator; c) Cooler; d) Lightend tower; e) Crude chloroethane storage; f) Hydrochlorinator; g) Compressor

In the first stage, ethane and chlorine are reacted noncatalytically after sufficient preheating at 400 – 450 °C in an adiabatic reactor. The reaction gases are separated after cooling in a first monochloroethane distillation tower. The heavy bottoms of this tower containing chloroethane and more higly chlorinated products (mostly 1,1-dichloroethane and 1,2-dichloroethane) are sent to the purification stage. The overheads consisting mainly of unconverted ethane, hydrogen chloride, and ethylene are sent to a second isothermal fixed-bed reactor. Before entering this reactor, fresh ethylene is added to achieve a 1 : 1 ethylene to HCl feed ratio. Even though the type of catalyst used in the isothermal section is not described, any of the catalysts mentioned for gas-phase hydrochlorination in the previous section can be used. Conversion at this stage is 50 – 80 %. The products are then separated in a second tower. Unconverted ethane, ethylene, and hydrogen chloride are recycled to the first reactor. The monochloroethane formed by hydrochlorination is drawn off and purified together with the stream from the first tower.

Even though the recycled ethylene from the hydrochlorination step is present during thermal chlorination, the formation of 1,2-dichloroethane is insignificant. Because the first reaction is carried out at high temperatures, chlorine addition to the ethylene double bond is suppressed.

The process is balanced by the overall reaction equation:

$$C_2H_6 + Cl_2 \rightarrow C_2H_5Cl + HCl$$

$$HCl + C_2H_4 \rightarrow C_2H_5Cl$$

A 90 % overall yield for ethane and ethylene and a 95 % chlorine yield to monochloroethane are reported.

Monochlorination of ethane is favored because ethane chlorination is four times faster than the consecutive chlorination of monochloroethane to dichloroethanes.

Major byproducts from the chlorination step are 1,1-, 1,2-dichloroethane and vinyl chloride. To achieve a high selectivity for monochloroethane, a high ethane surplus — preferably a 3 – 5-fold excess over chlorine [125, 126] — and good mixing is required. Insufficient heat dissipation may enhance cracking and coking. A thermal chlorination reactor providing thorough premixing and optimal heat transfer by means of a fluidized bed has been described in [126]. Other patents claim contact of the reaction gases with metal chlorides [127] or graphite [128].

The photochemical chlorination of ethane described in several patents [129] is less important, because it is difficult to implement in large volume plants and offers no major advantages over the thermal process.

Monochloroethane as a Byproduct of the Oxy-EDC Process. Monochloroethane is a major byproduct in the Oxy-EDC process (see page 35), in which it is formed by direct hydrochlorination of ethylene. It can be condensed or scrubbed from the light vent gases and recovered after further purification.

Monochloroethane from Ethanol. The esterification of ethanol with HCl is possible in the liquid phase by using $ZnCl_2$ or similar Lewis-acid catalysts at 110 – 140 °C [130]. Similar to the production of monochloromethane (see 1.3.2), the reaction can also be carried out in the gas phase by using γ-alumina [131], $ZnCl_2$ and rare earth chlorides on carbon [132] or zeolites [133] as catalysts. At the present ethanol prices, these procedures are prohibitive. With some modification, however, they can offer outlets for surplus byproducts such as ethyl acetate from PVA production which can be converted to monochloroethane by HCl using a $ZnCl_2$/silica catalyst [134].

Other Synthetic Routes to Monochloroethane. Non-commercial routes to monochloroethane consist of electrolytic chlorination of ethane in melts [135], reactions with diethyl sulfates [136], metathesis of 1,2-dichloroethane [137], hydrogenation of vinyl chloride [138], and conversion of diethyl ether [139]. The oxychlorination of ethane is discussed later in this Chapter.

Small amounts of monochloroethane are formed during the reaction of synthesis gas – chlorine mixtures over Pt/alumina [140] and methane – chlorine mixtures in the presence of cation-exchange resins complexed with TaF_5 [141].

2.1.4. Uses and Economic Aspects

Monochloroethane became industrially significant as a result of the developing automotive industry. It is the starting material for tetraethyllead, the most commonly used octane booster. In the United States, about 80 – 90 % and in Europe ca. 60 % of the monochloroethane production is used for the production of tetraethyl lead.

Production has already been cut significantly due to the increased use of unleaded fuel for environmental reasons. U.S. projections indicate an average annual decline of ca. 10 % per year. With some delay, the same trend can also be predicted for Western Europe.

Minor areas of use for monochloroethane are the production of ethyl cellulose, ethylating processes for fine chemical production, use as a blowing agent and solvent for extraction processes for the isolation of sensitive natural fragrances.

Production in 1984 in the Western World was about 300 000 t. Almost all processes in use at present are ethylene based.

2.2. 1,1-Dichloroethane

1,1-Dichloroethane [*75-34-3*] is the less important of the two dichloroethane isomers.

It occurs — often as an unwanted byproduct — in many chlorination and oxychlorination processes of C_2 hydrocarbons.

The most important role of 1,1-dichloroethane is as an intermediate in the production of 1,1,1-trichloroethane.

Other uses are negligible.

2.2.1. Physical Properties

M_r	98.97
mp	− 96.6 °C
bp at 101.3 kPa	57.3 °C
ϱ at 20 °C	1.176 g/cm^3
n_D^{20}	1.4164
Vapor pressure at	
0 °C	9.340 kPa
10 °C	15.370 kPa
20 °C	24.270 kPa
30 °C	36.950 kPa
Heat of formation (liquid) ΔH_{298}^0	− 160.0 kJ/mol
Specific heat at 20 °C	1.38 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	30.8 kJ/mol
Critical temperature	523 K
Critical pressure	5070 kPa
Viscosity at 20 °C	0.38×10^{-3} Pa s
Surface tension at 20 °C	23.5×10^{-3} N/m
Dielectric constant at 20 °C	10.9
Flash point (closed cup)	− 12 °C
Ignition temperature	458 °C
Explosive limits in air at 25 °C	5.4 – 11.4 vol%
	1,1-dichloroethane
Solubility in water at 20 °C	0.55 wt %
Solubility of water in 1,1-dichloroethane at 20 °C	0.97 wt %

1,1-Dichloroethane is a colorless liquid. It is readily soluble in all liquid chlorinated hydrocarbons and in a large variety of other organic solvents (ethers, alcohols).

Binary azeotropes are formed with water and ethanol: with 1.9 % water, *bp* 53.3 °C (97 kPa) and with 11.5 % ethanol, *bp* 54.6 °C (101 kPa).

2.2.2. Chemical Properties

At room temperature, 1,1-dichloroethane is adequately stable. Cracking to vinyl chloride and hydrogen chloride takes place at elevated temperatures. However, compared to other chlorinated C_2 hydrocarbons, the observed cracking rates are moderate. This reaction can be promoted by traces of chlorine and iron [142]. 2,3-Dichlorobutane is often found as a dimeric byproduct of decomposition.

1,1-Dichloroethane was also found to enhance 1,2-dichloroethane cracking when added in lower concentrations ($\leq$ 10 wt %) [143].

Corrosion rates for dry 1,1-dichloroethane are marginal, increase however, with water content and temperature. Aluminum is easily attacked.

In the presence of water or in alkaline solution, acetaldehyde is formed by hydrolysis.

2.2.3. Production

Theoretically 1,1-dichloroethane can be produced by three routes:

1) Addition of HCl to acetylene:

$$2\,HCl + C_2H_2 \xrightarrow{\text{Cat.}} CH_3{-}CHCl_2$$

2) Thermal or photochemical chlorination of monochloroethane:

$$C_2H_5{-}Cl + Cl_2 \xrightarrow[\text{or heat}]{h\nu} CH_3{-}CHCl_2 + HCl$$

3) Addition of HCl to vinyl chloride:

$$CH_2 = CHCl + HCl \xrightarrow{\text{Cat.}} CH_3{-}CHCl_2$$

For the synthesis of 1,1-dichloroethane as an intermediate in the production of 1,1,1-trichloroethane only the latter route is important and industrially used.

1,1-Dichloroethane via the 1,2-Dichloroethane – Vinyl Chloride Route. Hydrogen chloride and vinyl chloride obtained from 1,2-dichloroethane cracking see page 58) are reacted in a boiling-bed-type reactor [144] in the presence of a Friedel-Crafts catalyst, preferably ferric chloride ($FeCl_3$). 1,1-Dichloroethane is used as solvent and the temperature ranges from 30 to 70 °C.

Depending on the process design, hydrogen chloride can be used in excess to achieve complete conversion of the vinyl chloride. The heat of reaction, which differs only slightly from the heat required for 1,2-dichloroethane cracking, can be used to distill the 1,1-dichloroethane and recover part of the energy input. Downstream hydrogen chloride and unconverted vinyl chloride are separated and recycled. If necessary, the 1,1-dichloroethane can then be further purified by distillation. Due to the formation of heavy byproducts (vinyl chloride polymers) and deactivation of the catalyst, a slipstream from the reactor bottom must be withdrawn and new catalyst added.

Improved processes use column-type reactors with optimized height [145] (hydrostatic pressure to avoid flashing of vinyl chloride!) and recycled 1,1-dichloroethane with intermittent cooling stages. In this case, the stoichiometric ratio of hydrogen chloride to vinyl chloride, as obtained from 1,2-dichloroethane cracking, can often be used. In such a process, the downstream distillation equipment can be less complex and expensive, because almost complete conversion is achieved and because no excess hydrogen chloride or the entrained vinyl chloride must be separated. However, the energy requirements may be higher because most of the heat of formation must be dissipated by cooling.

Both process variations yield between ca. 95 and 98 %. Yield losses result through polymerization of vinyl chloride. The concentration as well as the nature of the catalyst determine this side reaction. Zinc chloride ($ZnCl_2$) and aluminum chloride ($AlCl_3$), which also can be used as catalysts, promote the formation of high molecular mass byproducts more than ferric chloride ($FeCl_3$) [120, 146]. The removed spent catalyst can be burned together with the heavy byproducts in an incinerator, if the vent gases are subsequently scrubbed and the wash liquor appropriately treated. Environmental problems

caused by the residues are thereby almost eliminated.

1,1-Dichloroethane via the Acetylene Route. As with the synthesis of vinyl chloride (see 3.1.3.1), 1,1-dichloroethane can be produced from acetylene by adding 2 mol of hydrogen chloride. For the first reaction sequence — the formation of vinyl chloride — mercury catalyst is required [147].

Because ethylene has become the major feedstock for chlorinated C_2 hydrocarbons, this process has lost its importance.

1,1-Dichloroethane from Ethane. 1,1-Dichloroethane may also be obtained by ethane or chloroethane chlorination. This chlorination can be carried out as thermal chlorination [148], photochlorination, or oxychlorination [149]. These processes, however, are impaired by a lack of selectivity and are not used industrially.

2.2.4. Uses and Economic Aspects

As mentioned earlier, 1,1-dichloroethane is primarily used as a feedstock for the production of 1,1,1-trichloroethane.

Although several other applications have been patented [150], currently 1,1-dichloroethane is rarely used for extraction purposes or as a solvent.

Based on estimated production figures of 1,1,1-trichloroethane and disregarding other uses, the total Western World production of 1,1-dichloroethane is estimated at 200 000 – 250 000 t for 1985.

2.3. 1,2-Dichloroethane

The first synthesis of 1,2-dichloroethane (ethylene dichloride, EDC) [*107-06-2*] was achieved in 1795.

Presently, 1,2-dichloroethane belongs to those chemicals with the highest production rates. Average annual growth rates of $> 10\,\%$ were achieved during the past 20 years.

Although these growth rates declined during the past several years, in the long run 1,2-dichloroethane will maintain its leading position among the chlorinated organic chemicals due to its use as starting material for the production of poly(vinyl chloride) (→ Poly(Vinyl Chloride)).

2.3.1. Physical Properties

M_r	98.97
mp	− 35.3 °C
bp at 101.3 kPa	83.7 °C
ϱ at 20 °C	1.253 g/cm^3
n_D^{20}	1.4449
Vapor pressure at	
0 °C	3.330 kPa
20 °C	8.530 kPa
30 °C	13.300 kPa
50 °C	32.000 kPa
70 °C	66.650 kPa
80 °C	93.310 kPa
Heat of formation (liquid) ΔH°_{298}	− 157.3 kJ/mol
Specific heat (liquid, at 20 °C)	1.288 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	34.7 kJ/mol
Critical temperature	563 K
Critical pressure	5360 kPa
Viscosity at 20 °C	0.84×10^{-3} Pa s
Surface tension at 20 °C	31.4×10^{-3} N/m
Coefficient of cubical expansion (0 – 30 °C)	0.00116 K^{-1}
Dielectric constant	10.5
Flash point (closed cup)	17 °C
Flash point (open cup)	21 °C
Ignition temperature (air)	413 °C
Explosive limits in air at 25 °C	6.2 – 15.6 vol% 1,2-dichloroethane
Solubility in water at 20 °C	0.86 wt %
Solubility of water in 1,2-dichloroethane at 20 °C	0.16 wt %

1,2-Dichloroethane is a clear liquid at ambient temperature, which is readily soluble in all chlorinated hydrocarbons and in most common organic solvents.

Binary azeotropes with 1,2-dichloroethane are listed in Table 14.

Table 14. Binary azeotropes formed by 1,2-dichloroethane

wt %	Component	Azeotrope boiling point (101.3 kPa), °C
18.0	2-propen-1-ol	79.9
38.0	formic acid	77.4
37.0	ethanol	70.3
19.5	1,1-dichloroethane	72.0
43.5	2-propanol	74.7
32.0	methanol	61.0
19.0	1-propanol	80.7
79.0	tetrachloromethane	75.6
18.0	trichloroethylene	82.9
8.2	water	70.5

2.3.2. Chemical Properties

Pure 1,2-dichloroethane is sufficiently stable even at elevated temperatures and in the presence of iron. Above 340 °C, decomposition begins, yielding vinyl chloride, hydrogen chloride, and trace amounts of acetylene [111]a, [151]. This decomposition is catalyzed by halogens and more highly substituted chlorinated hydrocarbons [152].

Long-term decomposition at ambient temperature caused by humidity and UV light can be suppressed by addition of stabilizers, mostly amine derivatives. Oxygen deficient burning and pyrolytic and photooxidative processes convert 1,2-dichloroethane to hydrogen chloride, carbon monoxide, and phosgene.

Both chlorine atoms of 1,2-dichloroethane can undergo nucleophilic substitution reactions, which opens routes to a variety of bifunctional compounds such as glycol (by hydrolysis or reaction with alkali), succinic acid dinitrile (by reaction with cyanide), or ethylene glycol diacetate (by reaction with sodium acetate). The reaction with ammonia to ethylenediamine and use of 1,2-dichloroethane for the production of polysulfides is of industrial importance.

Iron and zinc do not corrode when dry 1,2-dichloroethane is used, whereas aluminum shows strong dissolution. Increased water content leads to increased corrosion of iron and zinc; aluminum, however, corrodes less [153].

2.3.3. Production

1,2-Dichloroethane is industrially produced by chlorination of ethylene.

This chlorination can either be carried out by using chlorine (direct chlorination) or hydrogen chloride (oxychlorination) as a chlorinating agent.

In practice, both processes are carried out together and in parallel because most EDC plants are connected to vinyl chloride (VCM) units and the oxychlorination process is used to balance the hydrogen chloride from VCM production (see page 62 and Fig. 24). Depending on the EDC/VCM production ratio of the integrated plants, additional surplus hydrogen chloride from other processes such as chlorinolysis (perchloroethylene and tetrachloromethane production, see page 18 and Section 3.5.3) or 1,1,1-trichloroethane (see page 43) can be fed to the oxychlorination stage for proper balancing and chlorine recovery.

The use of ethane as a starting material, although the subject of numerous patent claims, is still in the experimental stage. It could offer economic advantages if the problems related to catalyst selectivity, turnover, and long-term performance are solved.

Direct Chlorination in the Ethylene Liquid Phase.. In the direct chlorination process, ethylene and chlorine are most commonly reacted in the liquid phase (1,2-dichloroethane for temperature control) and in the presence of a Lewisacid catalyst, primarily iron(III) chloride:

$$CH_2 = CH_2 + Cl_2 \xrightarrow{\text{Cat.}} CH_2Cl-CH_2Cl \qquad \Delta H^\circ_{298} = 220\ \text{kJ/mol}$$

To avoid problems in product purification, the use of high-purity ethylene is recommended. Especially its propane/propene content must be controlled in order to minimize the formation of chloropropanes and chloropropenes, which are difficult to separate from 1,2-dichloroethane by distillation. Purified liquid chlorine is used to avoid brominated byproducts. Oxygen or air is often added to the reactants, because oxygen was found to inhibit substitution chlorination, yielding particularly 1,1,2-trichloroethane and its more highly chlorinated derivatives [154, 155]. Through this and an optimized reactor design, the use of excess ethylene is no longer required to control byproduct formation. In most cases, the reactants are added in the stoichiometric chlorine/ethylene ratio or with a slight excess of chlorine. This simplifies the processing equipment because an excess of ethylene, which was often used in the past [156], requires complicated condensor and post reactor equipment to avoid the loss of expensive ethylene in the off-gas [155, 157].

Although several other Lewis-acid catalysts with higher selectivities such as antimony, copper, bismuth, tin, and tellurium chlorides [158] have been patented, iron chloride is widely used. Because the reaction selectivities are not dependent on the catalyst concentration, it is used in a diluted concentration between ca. 100 mg/kg

and 0.5 wt %. Some processes use iron filler bodies in the reactor to improve mass and heat transfer or use iron as a construction material. This equipment generates sufficient $FeCl_3$ in situ [159].

In the liquid-phase reaction, ethylene absorption was found to be the rate-controlling step [160].

In addition to the distinct process modifications with which each producer of 1,2-dichloroethane has improved his process during the past years, two fundamental process variations can be characterized:

1) low-temperature chlorination (LTC) and
2) high-temperature chlorination (HTC)

In the *LTC process*, ethylene and chlorine react in 1,2-dichloroethane as a solvent at temperatures (ca. 20 – 70 °C) below the boiling point of 1,2-dichloroethane.

The heat of reaction is transferred by external cooling either by means of heat exchangers inside the reactor or by circulation through exterior heat exchangers [161].

This process has the advantage that due to the low temperature, byproduct formation is low. The energy requirements, however, are considerably higher in comparison to the HTC process, because steam is required for the rectification of 1,2-dichloroethane in the purification section. Conversions up to 100 % with chlorine and ethylene selectivities of 99 % are possible.

In the *HTC process*, the chlorination reaction is carried out at a temperature between 85 and 200 °C, mostly, however, at about 100 °C. The heat of reaction is used to distill the EDC. In addition, EDC from the Oxy-EDC process or unconverted EDC from the vinyl chloride section can be added, since the heat of formation equals the heat required for vaporization by a factor of ca. 6.

By sophisticated reactor design and thorough mixing conversion, and yields comparable to the LTC process may be obtained with considerably lower energy consumption for an integrated DC-Oxy-VCM process [162].

Description of the HTC Process (Fig. 15). Gaseous chlorine and ethylene are fed thoroughly mixed into a reaction tower which is also supplied with dry EDC from oxychlorination or recycled EDC from the VCM section.

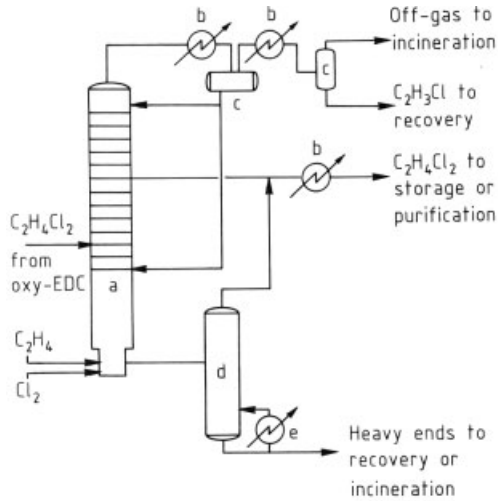

Figure 15. Simplified DC – HTC process
a) Reactor; b) Cooler; c) Knock-out drum; d) Heavy-end tower; e) Reboiler

The light ends are drawn off from the head section, and ethylene is condensed and recycled. In the following condensation section, vinyl chloride is separated and can then be processed with vinyl chloride from EDC cracking (see page 58). The remaining vent gas is incinerated. Pure EDC is taken from an appropriate section and condensed. In order to maintain a constant composition in the reactor sump phase, a slipstream is continuously withdrawn, from which the heavy byproducts are separated by rectification and sent to a recovery stage or incinerated. In some designs, the reactor is separated from the distillation tower [164]. In others, two towers are used for light ends/EDC separation. Solid adsorption has been patented for iron chloride removal [165].

For optimal heat recovery, cross exchange can be used for chlorine feed evaporation [166]. Due to the relatively low temperatures and anhydrous conditions, carbon steel equipment can be used [167].

Process developments using cracking gases instead of highly purified ethylene [168] and the use of nitrosyl chloride [169] as a chlorinating agent have not found any industrial importance.

Direct Chlorination in the Gas Phase. A catalytic gasphase process was patented by the Société Belge de l'Azote [170]. Because of the highly exothermic reaction, adequate dilution is

necessary. Several catalysts have been patented [171].

The noncatalytic chlorine addition reaction has been thoroughly studied [172], but is not industrially used, as is the case for the catalytic gas-phase chlorination of chloroethane [173].

Oxychlorination of Ethylene in the Gas Phase. In the oxychlorination process, ethylene and hy-drogen chloride are reacted with oxygen in the presence of an ambivalent metal catalyst. In most cases, copper salts are used at a temperature above 200 °C. The overall reaction can be formulated as

$$C_2H_4 + 2\,HCl + 1/2\,O_2 \xrightarrow{\text{Cat.}} C_2H_4Cl_2 + H_2O \qquad \Delta H^{\circ}_{298} = -295\ \text{kJ/mol}$$

The reaction sequence is similar to that of the Deacon process (see → Chlorine Chap. 10.2.1), although mechanistic studies indicate that ethylene is involved in the early reaction stages and the process may differ greatly from the classic Deacon process (e.g., oxidation of HCl to chlorine with subsequent addition of chlorine to ethylene by which chlorine is withdrawn from the Deacon equilibrium a high HCl conversion is achieved). The reaction sequence probably proceeds via chlorination of ethylene by cupric chloride. The copper salt is then regenerated by HCl and oxygen:

$$\begin{array}{l} C_2H_4 + 2\,CuCl_2 \longrightarrow C_2H_4Cl_2 + Cu_2Cl_2 \\ Cu_2Cl_2 + 2\,HCl + 1/2\,O_2 \longrightarrow 2\,CuCl_2 + H_2O \\ \hline C_2H_4 + 2\,HCl + 1/2\,O_2 \longrightarrow C_2H_4Cl_2 + H_2O \end{array}$$

Several investigations on the reaction mechanisms have been performed [174].

Ethylene oxychlorination has attained commercial importance since about 1960, when VCM producers began to pursue the ethylene route and HCl from EDC cracking had to be recovered. Due to this historic development, several process variations are presently used by the major EDC/VCM producers and will be discussed in detail later [175 – 179]. A common characteristic of all these processes is the catalytic gas-phase oxychlorination at temperatures between 200 and 300 °C and pressures of 0.1 to 1.0 MPa, usually at 0.4 – 0.6 MPa. HCl and ethylene conversions of 93 – 97 % are achieved at contact times between ca. 0.5 – 40 s with selectivities to EDC of 91 – 96 % [180].

Byproducts of ethylene oxychlorination are monochloroethane, formed by direct HCl addition to ethylene, VCM from the cracking of EDC, 1,1,2-trichloroethane formed by substitution chlorination of EDC or chlorine addition to VCM, 1,1-dichloroethane formed by the addition of HCl to VCM, and other crack or substitution products such as 1,1-dichloroethylene, cis- and trans-1,2-dichloroethylene, trichloroethylene, and tetrachloroethanes. Because oxygen is present, additional oxidation products such as acetaldehyde and its chlorinated derivatives, primarily trichloroacetaldehyde (chloral), are found in the reactor effluent. Oxirane (ethylene oxide) and glycols may also be formed. The ethylene feed is partially consumed, especially at higher temperatures, by deep oxidation to yield carbon oxides (CO, CO_2) and formic acid.

In some plants, major byproducts such as chloroethane and 1,1,2-trichloroethane are recovered and sold or used as feedstock for other chlorinated hydrocarbon (CHC) processes such as 1,1-dichloroethylene production (see Section 3.2.3) and chlorinolysis (see page 76).

Reactor Feed. Polymerization grade ethylene is used to minimize byproduct formation and purification problems.

In most cases, HCl from the EDC cracking section (see page 58) is used as a source of chlorine. The acetylene content (derived from VCM cracking) of this hydrogen chloride may be critical and should be controlled, because acetylene tends to form more highly chlorinated byproducts and tars, which can lead to catalyst deactivation by coking (pore plugging) and may also influence down-stream operations. Selective hydrogenation to ethylene is often used to remove acetylene from this HCl [181]. An other method proposes catalytic hydrochlorination with subsequent adsorption of the vinyl chloride formed [182]. In addition to HCl from EDC cracking, HCl from other CHC processes like 1,1,1-trichloroethane-, tri-, and tetrachloroethylene production can be used without problems, if kept free of such well-known catalyst poisons as fluorine and sulfur compounds.

In most processes, air is used as an oxygen source. Microfiltration prior to compression is employed to exclude particulate matter. In oxygen-based processes, pure oxygen is supplied by a nearby air liquefaction and separation

process and is used without additional processing.

Catalyst. Copper(II) salts, usually cupric chlorides, are used as standard catalysts [183 – 185]. In many cases, alkali, alkaline earth or aluminum chloride are added to reduce volatilization of the cupric salt. These salts form eutectic mixtures, which reduce the melting point. The reduction of the melting point, on the other hand, seems to be beneficial to the reaction rates. Furthermore, the addition of alkali salts suppresses direct addition reactions such as monochloroethane formation.

Some patents claim rare-earth salts (didymium salts) as promoters [185, 186] or use sodium/ammonium hydrogen sulfates [187] or tellurium salts [188].

High-surface-area alumina (150 – 300 m^2/g) is preferred as a support, because its production process allows the control of such important parameters as surface area, pore volume, and pore size distribution. Its high attrition resistance makes it very suitable for fluidized-bed reactors. Other support materials like graphite, silica, pumice, or kieselguhr are of minor importance.

For fluidized-bed reactors, alumina powder or microspheres (ca. 10 – 200 µm diameter) are used [189], whereas for fixed-bed reactors, catalyst tablets, extrudates, or spheres with a narrow size distribution (ca. 1/8 – 1/4″ diameter) are applied.

The catalyst is prepared to the support by the imbibition method using aqueous solutions of the catalyst salts followed by drying steps, or by special spray techniques [177].

Cupric chloride is usually added in concentrations of 3 – 12 wt % (of the total catalyst). Alkali salts are added in nearly double amounts to obtain molar alkali/copper ratios of 2 : 1 [190] and rare-earth salts in concentrations of 1 –10 wt %.

The fine adjustment of the catalyst composition as well as the selection of the appropriate support material and preparation procedure is a well kept secret of the individual technology and closely related to the reactor design.

Reactor Design. Theoretically, two basic reactor designs are in service:

1) fixed-bed reactors
2) fluidized-bed reactors

Due to the highly exothermic oxychlorination reaction, temperature control is a problem in *fixed-bed systems*. It is achieved by proper dilution of the catalyst with inactive diluents such as undoped alumina [184], graphite [184, 191], silicon carbide [184, 192], or nickel [193]. Thus, catalyst activity at the reactor inlet is normally low and increases to its maximum at the outlet.

In order to make catalyst charging not too complicated, several blends of active catalysts and inactive diluent are prepared and sequentially charged. One patent [191] claims to use four different catalyst zones containing from the reactor inlet to the outlet active catalyst concentrations of 7, 15, 40, and 100 vol%, respectively. The active catalyst consists of 8.5 wt % $CuCl_2$ on alumina. The inert diluent is made of graphite. In another process [194], highly concentrated active catalyst is deposited at the top to start the reaction followed by an inactive zone for temperature control. In the following zones, the active catalyst concentration increases and reaches 100 % in the last zone near the outlet.

Catalyst dilution requires exact mixing techniques and appropriate charging procedures in order to avoid demixing, i.e., segregation of diluent from active catalyst when different materials are used. This can lead to a rapid pressure drop buildup across the reactor.

Another approach is to vary catalyst activity through the catalyst particle size, which is not very practical, however [195]. More recent developments [194] favor staged catalysts, consisting of three to four different catalysts with varying amounts of $CuCl_2$ and KCl. The use of such catalyst systems, as offered by some manufacturers, does not require mixing and may offer advantages in some cases.

Fixed-bed technology is used by Dow Chemical, Stauffer, Toyo Soda and Vulcan. The size of the tubular reactors varies from 2 to 5 m in diameter and 4 to $>$ 10 m in length. They may comprise several thousand tubes for the catalyst with diameters up to 2″. Dow Chemical and Vulcan usually use one reactor, whereas Stauffer and Toyo Soda prefer successive oxychlorination systems with up to three reactors and split addition of oxygen. This latter method allows the formation of explosive mixtures at the reactor inlet to be more easily avoided, and it is claimed that fewer oxidation products are formed.

Nickel alloys are used for the construction of the tube section. Because of hot-spot formation, Alloy 200 may be prone to intergranular embrittlement, so that higher resistance may be obtained with Alloy 201 with a lower carbon content [167]. The tube sheet and the reactor head are lined with nickel on steel. For the reactor shell, carbon steel is primarily used.

Proper heat tracting for the interconnecting piping to the quench or absorber system is required to avoid corrosion.

The equipment for further processing such as the absorber – stripper – phase separator is lined with either bricks (towers) or teflon (pipes) to withstand the corrosion caused by aqueous HCl.

The heat of reaction is either used to generate steam at the side of the reactor shell or is transferred to a hot oil system, which may supply other plants.

Fluidized-bed reactors have the advantage of improved heat transfer and almost isothermal operation. However, backmixing, which influences conversion and selectivity, cannot be avoided. Nevertheless, HCl conversions of > 98 % have been reported [196]. This is achieved by feeding stoichiometric excesses of air or oxygen (10 – 80 %) and ethylene (up to 60 %) [197]. The temperature range between ca. 220 – 240 °C is somewhat lower than in fixed beds. Elevated pressure (0.2 – 0.5 MPa) is used to increase conversion. High-surface alumina powder (ca. 200 m^2/g) [189] or fuller's earth [198] are the preferred catalyst supports. The particle size distribution for representative samples shows a maximum at about 40 – 80 µm diameter [189]. Cupric chloride concentration on the catalyst varies from ca. 7 – 20 wt % $CuCl_2$. Higher concentrations are of no advantage, because the reaction rate will not improve and the catalyst will cake in the reactor.

Because of the lower temperature range, the reactor can be made of stainless steel if condensation (formation of aqueous HCl) can be avoided by means of proper shutdown procedures. Sparging equipment at the entrance of the reactor requires pipes, nozzles, and fittings of nickel alloys (Alloy 600 and 825) because they are more resistant to chloride stress corrosion [167].

Heat from the reaction is used to generate steam or is transferred to a hot oil system by internal cooling coils positioned in the fluidized bed.

One major advantage of the fluidized-bed reactor is that the reaction can be carried out within the explosive limit, which makes feed control less critical.

Reactor-integrated cyclones are used at the outlet to retain catalyst fines and to return them to the reaction zone.

Time – space yields may average 150 – 200 kg of EDC m^{-3} h^{-1} [199].

Fluidized-bed reactors are more widely used than fixed-bed systems. Companies using fluidized-bed technology are B. F. Goodrich, Hoechst, Pittsburgh Plate Glass (PPG), Ethyl Corp., Solvay, ICI, and Mitsui Toatsu Chemical.

Tokoyama Soda [178] and Pechiney [200] have combined the advantages of both processes by first reacting the gases in an isothermal fluidized bed and then passing them across a fixed bed for optimal yields and conversions.

Process Description (Fig. 16, 17 and 18). Ethylene and hydrogen chloride are preheated and fed with air or oxygen to the reactor. The hot reaction gases are quenched in a brick-lined tower and the resulting aqueous HCl is either treated together with the combined wastewaters or cleaned separately by stripping for further use, e.g., in a chlor-alkali process. The gases leaving the quench tower are cooled in a heat exchanger, and the organic phase is washed with dilute NaOH in order to remove chloral [201].

The off-gas is either vented after additional condensation and/or scrubbing or adsorption steps (for air-based systems) or compressed and recycled (if pure oxygen is used).

In some process modifications, heat exchangers and separators f and g are placed behind the NaOH wash. In other processes, the quench step is performed without addition of water, and a NaOH wash tower is not always required.

The wet EDC is dried by azeotropic destillation.

The bottoms from the azeotropic distillation are sent to the DC section for final purification. The light head products are submitted for further treatment together with the azeotrope for product recovery (ethylene, monochloroethane, EDC, chlorinated methanes) or incinerated. Care must be taken to remain outside of the flammable range during all process steps [202].

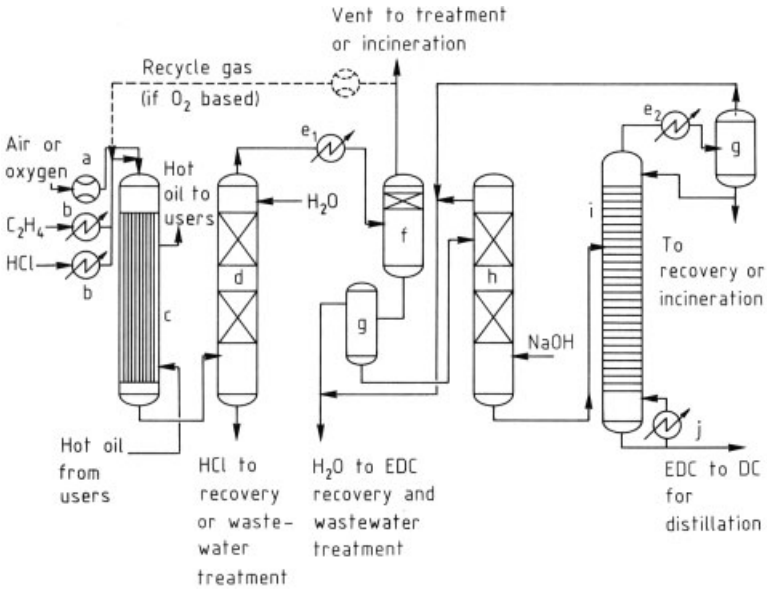

Figure 16. Oxy – EDC process (fixed bed, simplified)
a) Compressor; b) Preheater; c) Fixed-bed reactor; d) Quench tower; e) Cooler; f) Degasser; g) Separator; h) Wash tower; i) Azeotropic drying tower; j) Reboiler

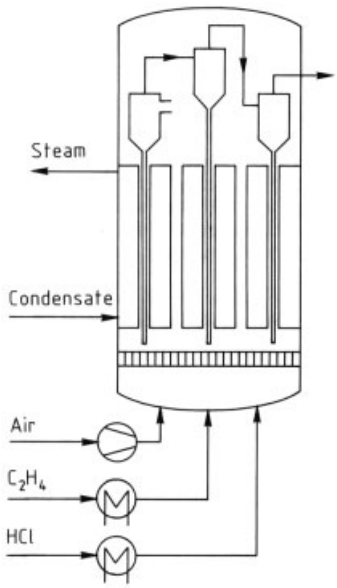

Figure 17. Oxy – EDC fluidized-bed reactor

Oxygen-based Oxychlorination. Vent gas from air-based oxychlorination processes is one of the major emission sources of CHC plants. In spite of intensive cooling and sophisticated absorber – stripper, adsorber – desorber, and post-reaction systems [203], the restrictions on many plants have become very stringent. The large amounts of nitrogen in the vent gas, however, makes the final treatment by incineration prohibitively expensive.

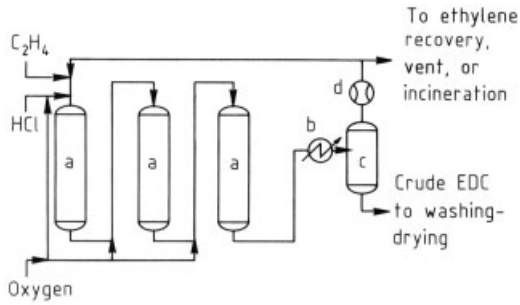

Figure 18. Stauffer oxygen-based Oxy-EDC process
a) Reactor; b) Cooler; c) Separator; d) Compressor

If oxygen is used instead of air [204], the vent stream becomes 20 – 100 times smaller, allowing vent incineration or catalytic oxidation [205].

Airbased systems are more manageable than others because the nitrogen from the air acts as diluent and removes heat.

In an oxygen-based system, this function is achieved with an excess of ethylene [206], which

is then recycled. Only a small quantity of the recycled stream must be drawn off in order to control the concentration of carbon oxides and other low-boiling byproducts. This slipstream is either burned or fed to the DC process to recover ethylene.

Since the heat capacity of ethylene in comparison to nitrogen is considerably higher, oxygen-based systems can be operated at lower temperatures or at higher throughput rates. This capacity increase together with the considerable savings for incineration [207] may offset the higher costs for oxygen and recycle compression energy.

During the past years, the conversion of many existing air-based facilities has proven to be feasible.

Oxychlorination of Ethylene in the Liquid Phase. An aqueous liquid-phase process for oxychlorination has been developed by the Kellog Co. [179, 208]. Ethylene, oxygen, and hydrogen chloride are fed to an aqueous solution of copper(II) salts (5 – 10 M) at 170 – 185 °C and 1.7 – 1.9 MPa. The 1,2-dichloroethane formed is stripped together with the steam generated by the heat of reaction. The gaseous products are quenched with water and further treated in a manner similar to gas-phase processes.

Although time – space yields and selectivities are comparable to the gas-phase process and feed impurities can be tolerated, the liquid-phase process is not industrially used. The main reason may be the troublesome handling of highly corrosive aqueous solutions at an elevated temperature and high pressure, even though similar problems have not been restrictive for the liquid-phase hydrochlorination of methanol (see page 13). Wastewater treatment may also pose more problems compared to gas-phase processes, because heavy metal contamination occurs. More information on homogeneously catalyzed oxychlorination may be found in the literature [209].

1,2-Dichloroethane from Ethane. The substantial cost margin between ethane and ethylene has prompted considerable research on direct ethane oxychlorination. This oxychlorination reaction is theoretically possible and proceeds via the sequence ethane – monochloroethane – dichloroethane.

Several processes comprising ethane and chloroethane oxychlorination or reacting mixtures of both components have been patented [149, 210].

A process developed by the Monsanto Company [211] comprises the direct thermal chlorination of ethane, yielding monochloroethane. In the next step, the reaction gases are oxychlorinated to 1,2-dichloroethane. In another variation [212], ethane is oxidized by oxygen in the presence of HCl at 400 – 600 °C to give ethylene. The resulting mixture is again oxychlorinated in a conventional manner.

This process has similarities with the autothermic cracking process [213], where ethylene, chlorine, and oxygen are reacted at 850 – 950 °C to form mainly ethylene and hydrogen chloride. In a further process step, the gases are oxychlorinated to 1,2-dichloroethane.

In both cases, chlorine balancing (HCl) in an integrated VCM process seems feasible. However, none of these processes have been implemented on an industrial scale. Compared to the oxychlorination of ethylene, ethanebased processes are frequently affected by poor conversion and selectivity. This necessitates high recycling rates, thereby increasing costs. The lack of selectivity also requires additional outlets for major byproducts (1,1-dichloroethane and trichloroethane), which may not always exist.

The cost advantage of ethane based processes further diminishes if cracker capacity for the ethylene supply and the infrastructure (loading stations, storage tanks, etc.) are already available.

Research is directed toward the development of more specific catalysts, e.g, zeolites, and a direct route for producing VCM from ethane without isolating EDC followed by cracking (see Section 3.1.3.4). Such processes may offer true cost advantages for plants with easy access to ethane (U.S. gulf coast) and good integration into other CHC plants for economical byproduct recovery.

Other Processes. The production of 1,2-dichloroethane from ethanol [214] is not industrially used. It might be of interest if the cost for ethanol derived from biomass were to become competitive [107, 108].

1,2-Dichloroethane is also a byproduct of oxirane (ethylene oxide) production via the old chlorohydrine route. The EDC yield can be improved up to 50 % by process modifications [215].

Because oxirane (ethylene oxide) is mostly produced by direct oxidation, this process is not important for EDC production.

2.3.4. Uses and Economic Aspects

Based on U.S. figures for 1981, ca. 85 % of the total EDC production is used for the production of vinyl chloride. 10 % is used in the production of chlorinated solvents such as 1,1,1-trichloroethane and tri- and tetrachloroethylene. The rest goes into various processes mainly for the synthesis of ethylenediamines. Its use as a solvent (dewaxing, deparaffinizing petroleum fractions, and coating remover) is marginal. EDC is further used in leaded gasoline as a lead scavenger. With the increasing trend toward unleaded fuel, however, this market will decline in future.

In Europe, market figures seem to be comparable to those in the US, if not shifted even more toward vinyl chloride production, because almost all European EDC plants are forward integrated to VCM units.

Production in 1985 is estimated at ca. 7×10^6 t for US, 8×10^6 t in Europe, and 2.5×10^6 t in Japan.

Installed capacity is ca. 10×10^6 t in North America, 10×10^6 t in Western Europe, and 3.5×10^6 t in Japan.

The future average growth rate is difficult to predict, since EDC production depends heavily on the big PVC consumers, the automotive industry and the construction business, and has, therefore, been subjected to severe fluctuations in the past. The growth rate may be estimated at ca. 2 – 5 % for the decade 1985 – 1995.

New EDC plants in construction or in the planning phase will preferentially be located in developing countries to increase autonomy from imports and in oil-producing countries to forward integrate refineries and basic chemicals already produced.

2.4. 1,1,1-Trichloroethane

1,1,1-Trichloroethane [*71-55-6*] was first synthesized in the mid-19th century. It was not used industrially for more than 100 years and was frequently found as an unwanted byproduct in chlorinated hydrocarbon processes.

The Dow Chemical Co. began commercial production in the early 1950's. With the development of effective stabilizer systems, 1,1,1-trichloroethane has become one of the major solvents for cold and vapor degreasing as well as several other applications. 1,1,1-Trichloroethane is in strong competition with trichloroethylene (see Section 3.4.4) and has replaced this solvent in many fields.

2.4.1. Physical Properties

M_r	133.41
mp	− 33 °C
bp at 101.325 kPa	74.1 °C
ϱ at 20 °C	1.325 g/cm^3
n_D^{20}	1.4377
Vapor pressure at	
0 °C	4.900 kPa
20 °C	13.300 kPa
40 °C	32.000 kPa
60 °C	62.700 kPa
70 °C	88.000 kPa
80 °C	120.000 kPa
Heat of formation (liquid) ΔH_{298}^0	− 170 kJ/mol
Specific heat (liquid 20 °C)	1.004 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	32 kJ/mol
Critical temperature	585 K
Critical pressure	4500 kPa
Viscosity at 20 °C	0.86×10^{-3} Pa s
Surface tension at 20 °C	25.6×10^{-3} N/m
Coefficient of cubical expansion (0 – 30 °C)	0.0013 K^{-1}
Dielectric constant at 20 °C	7.5
Flash point (closed cup)	none
Ignition temperature (air)	537 °C
Explosive limits in air at 25 °C	8.0 – 10.5 vol% 1,1,1-trichloroethane
Solubility in water at 20 °C	0.095 wt %
Solubility of water in 1,1,1-trichloroethane at 20 °C	0.034 wt %

1,1,1-Trichloroethane is a clear liquid at ambient temperature with a characteristic ethereal odor. It is soluble in all common organic solvents and is a very good solvent for fats, paraffins, and other organic compounds.

Some binary azeotropes are shown in Table 15.

2.4.2. Chemical Properties

Pure 1,1,1-trichloroethane is very unstable and tends to undergo dehydrochlorination. Noncatalyzed pyrolytic decomposition is almost

complete at 300–400 °C [216]. When catalyzed by metal salts [217], aluminum fluoride [218], alumina [219], or others [220], this reaction proceeds at considerably lower temperatures. Dehydrochlorination yields dichloroethylenes and hydrogen chloride. High molecular mass products may also be obtained by polymerization of dichloroethylene [219]. Phosgene formation at elevated temperature in the presence of air is marginal [221]. Compared to olefinic chlorination solvents (tri- and tetrachloroethylene) used in similar applications, 1,1,1-trichloroethane shows better stability against oxidation [222]. Photochemical oxidation yields phosgene, carbon monoxide, and hydrogen chloride [223]. Hydrolysis with water and aqueous acid yields acetyl chloride and acetic acid [224]. Under normal conditions, this reaction proceeds slowly. Dehydrochlorination to 1,1-dichloroethylene takes place in alkaline solutions.

Table 15. Binary azeotropes formed by 1,1,1-trichloroethane

wt %	Component	Azeotrope boiling point (101.3 kPa), °C
4.3	water	65.0
23.0	methanol	55.5
17.4	ethanol	64.4
17.0	isopropanol	68.2
17.2	*tert*-butanol	70.2

1,1,1-Trichloroethane is extremely corrosive to aluminum. Inhibitors must be used inevitably. Dry 1,1,1-trichloroethane moderately corrodes iron and zinc. Corrosion, however, increases with the water content.

2.4.3. Production

For the industrial production of 1,1,1-trichloroethane, three different routes are in use:

1) From 1,1-dichloroethane by thermal or photochemical chlorination
2) From 1,1,2-trichloroethane via 1,1-dichloroethylene and consecutive hydrochlorination
3) From ethane by direct chlorination

In the United States, more than 70 % of the 1,1,1-trichloroethane is produced by the 1,1-dichloroethane process. An additional 20 % is based on the 1,1-dichloroethane route, and ca. 10 % is made by direct ethane chlorination. In Europe too, the 1,1-dichloroethane route is used by the largest producers.

Compared to the latter process, the production from 1,1-dichloroethylene has the disadvantage that more expensive chlorine is required, because one-fourth of the total chlorine required is lost as inorganic chloride. In addition, this route requires an aqueous system because dilute NaOH is used for the 1,1,2-trichloroethane dehydrochlorination (see page 68), which may cause environmental problems. Furthermore, 1,2-dichloroethane as feedstock for the 1,1-dichloroethane route is more readily available than 1,1,2-trichloroethane. The HCl generated by 1,1-dichloroethane chlorination can be used in other processes such as Oxy-EDC (see page 37) or methanol hydrochlorination (see page 11).

In other words, the first route is a source of HCl, whereas the second route consumes HCl. These aspects may have been also decisive for the implementation of the various processes in integrated CHC plants.

Even though photochemical reactions are rarely used for industrial purposes because reactor design and operation is somewhat troublesome, the photochemical reaction is preferred for 1,1,1-trichloroethane preparation from 1,1-dichloroethane because of its higher selectivity compared to thermal chlorination.

The direct chlorination of ethane is the least used route because of its lack of selectivity.

Besides the three routes mentioned above, several other processes have been proposed, but are not used on an industrial scale.

1,1,1-Trichloroethane from 1,1-Dichloroethane. This process uses 1,2-dichloroethane (EDC) as feedstock which is rearranged to 1,1-dichloroethane via cracking to vinyl chloride (see Section 3.1.3.2) followed by the addition of HCl in the presence of a catalyst.

During the final step, 1,1-dichloroethane is thermally or photochemically chlorinated.

$$CH_2Cl{-}CH_2Cl \xrightarrow{\Delta} CH_2 = CHCl + HCl$$

$$CH_2 = CHCl + HCl \xrightarrow{\text{Cat.}} CH_3CHCl_2$$

$$CH_3CHCl_2 + Cl_2 \xrightarrow{h\nu \text{ or } \Delta} CH_3{-}CCl_3 + HCl \qquad \Delta H^0_{298} = -103 \text{ kJ/mol}$$

Photochemical Chlorination. Photochemical chlorination is used mostly, because this reaction can be carried out at a lower temperature, which increases the selectivity toward 1,1,1-trichloroethane [225].

Preferred temperatures range between 80 and 160 °C and the reactor pressure may average 0.1 – 0.4 MPa. Reactor design for photochemical chlorination is a compromise: although selectivity is increased in a plug-flow reactor [226], the systems in use resemble back-mixed tank reactors (CSTR-characteristic) due to the need for sufficient actinic light radiation.

Typical byproducts are 1,1,2-trichloroethane, 1,1,1,2- and 1,1,2,2-tetrachloroethane, and pentachloroethane, which may represent up to 30 % of the yield.

To minimize the formation of tetra- and more highly chlorinated byproducts and to dissipate the heat of formation, excess 1,1-dichloroethane is fed (3 – 10 M) to chlorine.

In order to maintain the formation of unwanted 1,1,2-trichloroethane as low as possible, the photochemical reaction is preferably carried out in the vapor phase, because liquid-phase chlorination favors the synthesis of 1,1,2-trichloroethane. Maximum selectivity toward 1,1,1-trichloroethane is ca. 90 % [227], which may, however, not be achieved in industrial processes. Catalytic traces of iodine or iodine-containing compounds were also found to increase the selectivity [228].

After separation from 1,1,1-trichloroethane, 1,1,2-trichloroethane can be used for the production of 1,1-dichloroethylene (vinylidene chloride see Section 3.2). The tetrachlorinated products may be used either in the production of trichloroethylene (see Section 3.4.3) or, without separation together with pentachloroethane, as feed for the perchloroethylene process.

Care must be taken to exclude traces of iron, which is a well-known promoter for the formation of 1,1,2-trichloroethane through a cracking – addition sequence. Monel or other copper – nickel alloys are the preferred construction materials for reactor and distillation equipment.

Special distillation processes using stabilizers [229] or thin-film evaporations [230] have been patented to avoid decomposition of 1,1,1-trichloroethane during cleanup.

Double-well UV lamps with external cooling [231] are used to prevent coking and to keep the shutdown frequency tolerable.

Process Description (Fig. 19). Finely dispersed 1,1-dichloroethane is fed together with chlorine into an adiabatic reactor equipped with an actinic light source (300 – 550 nm). Dichloroethane is fed in excess up to 10 M to dissipate the heat of reaction through evaporation and to suppress consecutive chlorination reactions. Reactor temperature averages between 80 – 100 °C at pressures of 0.1 – 1.0 MPa.

The reaction products are separated by distillation. In the first step, 1,1-dichloroethane is distilled together with hydrogen chloride and unreacted chlorine. HCl is then separated in a second tower. Dichloroethane and chlorine are recycled to the reactor.

The high-boiling components of the sump phase from the first tower are separated in at least two more steps yielding the crude products, which are then further purified.

Thermal Chlorination [232 – 234]. Fluidized-bed reactors offer the best technical solution for thermal chlorination because their uniform temperature profile minimizes the cracking reactions of 1,1,1-trichloroethane and reactor coking. Sand or silica [234], which must be free of iron, is used as a bed material.

Preheating of the reactants is required to maintain a reaction temperature of 350 – 450 °C, which is necessary to start the radical chain reaction.

Excess 1,1-dichloroethane is again added in proper dilution to avoid product losses (heavies formation).

Equipment and processing is quite similar to the photochemical system.

Although 1,1,1-trichloroethane yields of up to 82 % have been reported [233], actual yields may be considerably lower or poor conversions with higher energy requirements for dichloroethane recycling must be accepted. As far as product yield, selectivity, and specific energy consumption are concerned, this process is inferior to photochemical chlorination.

Other Processes. Catalytic liquid-phase chlorination of 1,1-dichloroethane using phosphorus catalysts (PCl_5) [235] as well as the highly selective chlorination by chlorine monoxide [236] are not industrially used. Either selectivity (PCl_5) or

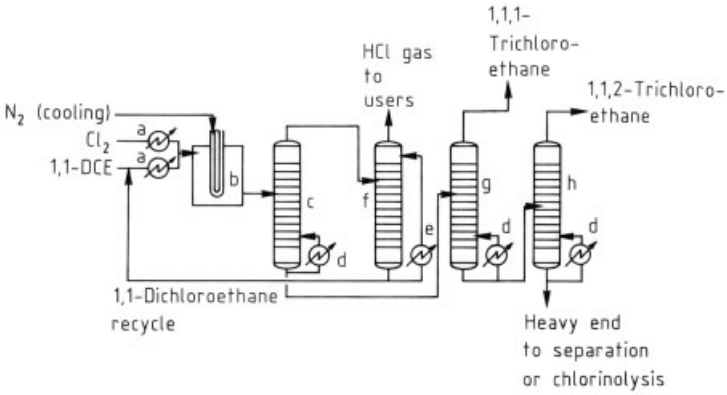

Figure 19. 1,1,1-Trichloroethane process (photochlorination)
a) Preheater; b) Photoreactor; c) Lights – heavies separation; d) Reboiler; e) Cooler; f) HCl tower; g) 1,1,1-Trichloroethane tower; h) 1,1,2-Trichloroethane tower

technical problems (Cl_2O) do not allow large-scale production.

1,1,1-Trichloroethane from 1,1,2-Trichloroethane. The overall reaction sequence again begins with 1,2-dichloroethane, which is chlorinated to form 1,1,2-trichloroethane. 1,1-Dichloroethylene is obtained by dehydrochlorination and is then hydrochlorinated to 1,1,1-trichloroethane:

$$CH_2Cl{-}CH_2Cl + Cl_2 \longrightarrow CHCl_2{-}CH_2Cl + HCl$$
$$CHCl_2{-}CH_2Cl + NaOH \longrightarrow CCl_2 = CH_2 + NaCl + H_2O$$
$$CCl_2 = CH_2 + HCl \xrightarrow{\text{Cat.}} CH_3{-}CCl_3$$
$$\Delta H^\circ_{298} = -80\ \text{kJ/mol}$$

As with the 1,1-dichloroethane route, an interim rearrangement via dehydrochlorination is again required. However, pyrolytic gas-phase dehydrochlorination of 1,1,2-trichloroethane does not have the required selectivity toward 1,1-dichloroethylene — the favored product is 1,2-dichloroethylene (see Section 3.3) — to be industrially attractive [237]. A yield increase to more than 90 % is only possible in aqueous systems using calcium hydroxide [238] or ammonia [239] and some of its derivatives [240]. Aqueous NaOH is primarily used for dehydrochlorination [241].

The NaCl-containing NaOH (8 – 10 wt % NaOH, 15 – 20 wt % NaCl) from diaphragm cells can be used directly without further evaporation. The use of an aqueous system however, results in loss of chlorine, which is discarded as a salt ($CaCl_2$ and NaCl).

The reaction is carried out at 80 – 120 °C in a packed tower or recirculation reactors. Because alkaline brine is used, nickel is applied as a construction material. Crude 1,1-dichloroethylene is withdrawn by live steam injection or flash evaporation and distilled.

In order to avoid polymerization of 1,1-dichloroethylene, all feed streams should be free of oxygen (< 1 mg/kg), or stabilizers (radical scavengers like phenols or amines) should be used.

The hydrochlorination of 1,1-dichloroethylene is carried out in a manner similar to the production of 1,1-dichloroethane from vinyl chloride at temperatures between 40 – 80 °C and in the presence of a Lewis acid catalyst ($FeCl_3$) [242]. Reactant ratios are almost stoichiometric or with a slight excess of HCl. The 1,1,1-trichloroethane formed can be used as solvent, but others such as 1,1,2-trichloroethane and perchloroethylene have also been mentioned [243]. Care must be taken to avoid entrainment of catalyst traces during purification. Remaining traces of catalyst and hydrogen chloride can be removed by addition of ammonia [244] and distillation or by careful filtration of the distilled 1,1,1-trichloroethane over partially deactivated NaOH flake beds or weak ion-exchange resins. Overall yields of more than 90 % are obtainable.

In a special process carried out by Atochem (France), 1,1-dichloroethylene is produced as byproduct from the high-temperature chlorination of ethylene for the production of vinyl chloride [245] (see Section 3.1.3.4). After separation

from other byproducts, it can be hydrochlorinated as usual to give 1,1,1-trichloroethane as a valuable byproduct.

A similar process has been patented by the FMC Corp. [243]. In this case ethylene is fed to the hydrochlorination stage forming monochloroethane, which is then subjected to high temperature chlorination to give 1,1-dichloroethylene.

1,1,1-Trichloroethane from Ethane. The direct synthesis of 1,1,1-trichloroethane has been patented by the Vulcan Materials Company and is mainly used in the United States:

$$CH_3-CH_3+3Cl_2 \rightarrow CH_3-CCl_3+3HCl$$
$$\Delta H^0_{298} = -330\,\mathrm{kJ/mol}$$

The highly exothermic reaction can be controlled by recycling the chloroethane byproducts (monochloroethane, 1,1-, and 1,2-dichloroethane), which consume some of the reaction heat by endothermic dehydrochlorination reactions, so that an adiabatic reactor can be used. Hot spot temperatures of ca. 440 °C are obtained. Mean residence times of 10 – 20 s at reactor pressures of 0.3 – 0.5 MPa are found in the patent literature [246].

Due to the rigorous reaction conditions and the long reaction sequence, numerous byproducts are formed which require extensive equipment for postprocessing after quenching of the reactor gas. Even extractive distillation steps have been considered [247].

Both 1,1- and 1,2-dichloroethane are recycled. Vinyl chloride and vinylidene chloride are hydrochlorinated. 1,1-Dichloroethane from vinyl chloride hydrochlorination is added to the recycle and 1,1,1-trichloroethane resulting from the vinylidene chloride is drawn off.

At optimal conditions, the reaction product contains 60 – 70 mol% 1,1,1-trichloroethane and 20 mol% vinylidene chloride. Overall ethane yields of 60 % and chlorine yields of 93 % are obtained.

The advantage of the cheaper raw material ethane is, at least partially, offset by the higher equipment costs and the lower overall yields.

Other Processes. Other patents use monochloroethane as feedstock for the thermal chlorination [231, 248] or monochloroethane is formed in situ by feeding additional ethylene to the reactor under mild conditions [249].

The photochemical chlorination of monochloroethane has also been described [250].

All of these processes have no distinct advantages over the two basic methods that begin with 1,2-dichloroethane because of a lack of selectivity. They are, therefore, currently of no importance for the industrial production of 1,1,1-trichloroethane.

2.4.4. Uses and Economic Aspects

1,1,1-Trichloroethane is used as a solvent in numerous industrial applications such as cold and hot cleaning and vapor degreasing. Formulations are used as solvent for adhesives and metal cutting fluids [251].

New applications have been found in textile processing and finishing and in dry cleaning, where 1,1,1-trichloroethane can replace the widely used perchloroethylene.

Special grades are used for the development of photoresists in the production of printed circuit boards. Because of its lower toxicity, 1,1,1-trichloroethane has replaced trichloroethylene in many fields, especially in the United States. In Europe, however, trichloroethylene has maintained its leading position.

Further advantages of 1,1,1-trichloroethane are its graduated solvency, which allows it to be used even in very sensitive areas, its good evaporation rate, and the fact that it has no fire or flash point.

Because it readily reacts with aluminum and other metals, inhibitors must be added prior to industrial use. The inhibitor systems, which may account for 3 – 8 % of the formulation, mainly comprise such acid acceptors as epoxides, ethers, amines, and alcohols as well as such metal stabilizers as nitro- and cyano-organo compounds, which build complexes, thereby deactivating metal surfaces or catalytic salt traces. Several formulations for proprietary grades of 1,1,1-trichloroethane have been filed and are in use [252].

U.S. producers are The Dow Chemical Co., Pittsburgh Plate Glass Inc., and the Vulcan Materials Co. In Europe, 1,1,1-trichloroethane is produced by The DOW Chemical Co., ICI Ltd., Atochem, and Solvay.

European and U.S. capacities together amount to ca. 600 000 t. Production in 1984 was

estimated to be ca. 450 000 t (Europe ca. 150 000 t). In Europe, 1,1,1-trichloroethane strongly competes with trichloroethylene and may replace this solvent in several applications, which could affect the future growth rate.

However, more stringent regulations in many industrialized countries requiring the reduction of losses from vapor degreasing units and other equipment may adversely affect future demand.

Present capacity seems to be sufficient to supply the market for the next decade.

2.5. 1,1,2-Trichloroethane

1,1,2-Trichloroethane [*79-00-5*] is primarily an unwanted byproduct of several chlorination processes such as the production of 1,2-dichloroethane and the chlorination of ethane or 1,1-dichloroethane to 1,1,1-trichloroethane.

It has a very high solvency, but the relatively high toxicity limits its uses.

1,1,2-Trichloroethane is only important as an intermediate in the production of 1,1-dichloroethylene and to some extent for the synthesis of tetrachloroethanes.

2.5.1. Physical Properties

M_r	133.41
mp	− 37 °C
bp at 101.325 kPa	113.5 °C
ϱ at 20 °C	1.4432 g/cm^3
n_D^{20}	1.4711
Vapor pressure at	
30 °C	4.800 kPa
90 °C	49.200 kPa
100 °C	67.300 kPa
110 °C	90.600 kPa
114 °C	101.800 kPa
Heat of formation (liquid) ΔH^0_{298}	− 188 kJ/mol
Specific heat (liquid, 20 °C)	1.113 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	39.1 kJ/mol^{-1}
Viscosity at 20 °C	1.20×10^{-3} Pa s
Surface tension at 20 °C	32.5×10^{-3} N/m
Coefficient of cubical expansion (0 – 25 °C)	0.001 K^{-1}
Autoignition temperature (air)	460 °C
Solubility in water at 20 °C	0.45 wt %
Solubility of water in 1,1,2-trichloroethane at 20 °C	0.05 wt %

1,1,2-Trichloroethane is a clear liquid at ambient temperature with a sweet smell. It is not flammable and easily miscible with most organic solvents.

Table 16. Binary azeotropes formed by 1,1,2-trichloroethane

wt %	Component	Azeotrope boiling point (101.3 kPa), °C
97	methanol	64.5
57	perchloroethylene	112.0
70	ethanol	77.8
15	water	85.3 (at 97.300 kPa)

For some binary azeotropes, see Table 16.

2.5.2. Chemical Properties

At an elevated temperature (400 – 500 °C), 1,1,2-trichloroethane is easily dehydrochlorinated to give a mixture of *cis*- and *trans*-1,2-dichloroethylene, 1,1-dichloroethylene, and hydrogen chloride [253].

Addition of alumina catalysts or an increase in temperature favors 1,2-dichloroethylene formation. In aqueous alkaline solution, 1,1,2-trichloroethane is selectively dehydrochlorinated to 1,1-dichloroethylene [254]. This reaction proceeds faster than that with 1,1,1-trichloroethane. In water, hydrolysis takes place, especially under reflux conditions. Hydrolysis, however, proceeds slower than with the 1,1,1-isomer.

1,1,2-Trichloroethane is highly corrosive to aluminum, iron, and zinc. Addition of water increases the rate of corrosion.

If chlorinated, a mixture of the isomeric tetrachloroethanes is formed.

2.5.3. Production

The industrial production of 1,1,2-trichloroethane proceeds by two routes:

1) Selective chlorination of 1,2-dichloroethane:

$$\begin{array}{c} CH_2-CH_2 \\ | \quad\quad | \\ Cl \quad\quad Cl \end{array} + Cl_2 \longrightarrow CHCl_2-CH_2Cl + HCl$$

$$\Delta H^0_{298} = -116 \text{ kJ/mol}$$

2) Addition of chlorine to vinyl chloride:

$$CH_2 = CHCl + Cl_2 \rightarrow CH_2Cl-CHCl_2$$

$$\Delta H^0_{298} = -224 \text{kJ/mol}$$

The liquid-phase 1,2-dichloroethane route is most often used in industrial processes when

1,1,2-trichloroethane is needed for the production of 1,1-dichloroethylene (see Section 3.2.3) and 1,1,1-trichloroethane (see Section 2.4.3).

The vinyl chloride route plays a minor role because of the more expensive feedstock and overall higher energy requirements.

A large portion of the current demand for 1,1,2-trichloroethane can be satisfied by the use of 1,1,2-trichloroethane obtained as a byproduct from the production of 1,1,1-trichloroethane. Because the latter is mostly produced by photochemical chlorination of 1,1-dichloroethane (see page 43), substantial amounts of 1,1,2-trichloroethane are obtained as a byproduct.

1,1,2-Trichloroethane is also one of the major byproducts in the production of 1,2-dichloroethane (see Section 2.3.3) and can be distilled from the heavy ends of these processes.

Thus, 1,1,2-trichloroethane is deliberately produced only when 1,1,1-tri- or 1,2-dichloroethane sources are unavailable or for the balancing of feedstocks.

1,1,2-Trichloroethane from 1,2-Dichloroethane. This process was patented by the Pittsburgh Plate Glass Co. [255] and by the Toa Gosei Chemical Industries [256]. It comprises the noncatalytic chlorination of 1,2-dichloroethane. The reaction is carried out in the liquid phase at temperatures between 100 – 140 °C. The addition of ethylene induces the reaction. The reaction mechanism has been intensively studied but is not yet clearly and fully understood [257]. Radical chlorination is very likely because metal salts, which support ionic reactions, must be excluded. The reactants should be free of radical-scavenging oxygen. Because consecutive chlorination to tetrachloroethanes and pentachloroethane takes place, the conversion per pass must be kept at 10 – 20 % for an optimum trichloroethane yield [256]. Since backmixing also favors the formation of higher chlorinated products, adiabatic plug-flow reactors are preferred.

Nickel alloys or nickel-clad steel is preferred as a structural material. As with most chlorination reactions, water must be removed to minimize corrosion.

Process Description (Fig. 20). Ethylene, chlorine, and fresh and recycled 1,2-dichloroethane are fed to a tubular reactor.

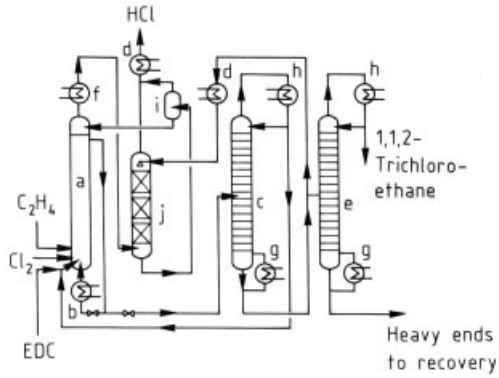

Figure 20. Process for the production of 1,1,2-trichloroethane by ethylene-induced liquid-phase chlorination [256]
a) Reactor; b) Preheater for priming; c) Light-ends tower (1,2-dichloroethane); d) Cooler (brine); e) 1,1,2-Trichloroethane finishing tower; f) Cooler (water); g) Reboiler; h) Condenser; i) Knock-out drum; j) 1,2-Dichloroethane wash tower

Even at low conversion rates, the heat of reaction is sufficient to maintain the reaction. Preheating of the recycled stream is, therefore, only required for startup.

The liquid phase from the reactor is first distilled to separate the excess 1,2-dichloroethane, which is then recycled to the reactor.

The crude 1,1,2-trichloroethane is then separated from the higher chlorinated byproducts, mainly tetrachloroethanes, by further distillation.

The gaseous reaction phase is cooled and high-boiling components are condensed. The remaining 1,2-dichloroethane in the HCl gas is washed out with crude 1,1,2-trichloroethane and recycled to the reactor.

Other catalysts such as azodiisobutyronitrile (ADIB) or peroxides [258], actinic light [159, 256, 259] (liquid-phase photochlorination of 1,2-dichloroethane favors 1,1,2-trichloroethane, whereas by gas-phase photochlorination, 1,1,1-trichloroethane is preferentially obtained and phosphoros chloride have been proposed instead of ethylene [260]. Gas-phase chlorination catalyzed by metal chlorides has also been patented [261].

All these processes, however, are commercially more demanding and offer no real advantage over the ethylene-induced reaction.

Instead of chlorinating 1,2-dichloroethane, ethylene can be used as a starting material. Either

liquid-phase chlorination with chlorine [262] (see DC – EDC, page 35) or catalytic gas-phase oxychlorination with HCl in fixed- or fluidized-bed reactors [263] (see Oxy – EDC, page 37) can be used. The selectivity toward 1,1,2-trichloroethane is generally lower with this process. Especially in gas-phase oxychlorination, substantial yield losses by "deep oxidation" (CO and CO_2 formation) may occur.

1,1,2-Trichloroethane from Vinyl Chloride. The reaction is carried out similar to the DC-EDC process (see page 35) in liquid phase (trichloroethane) at temperatures between 50 and 90 °C. As with other addition reactions, Lewis-acid catalysts like $FeCl_3$, $AlCl_3$ or $SbCl_5$ are used, but actinic light catalysis also seems possible [264]. Yields of more than 90 % are obtainable.

Catalytic oxychlorination of vinyl chloride with hydrogen chloride [265] or oxychlorination of mixtures of ethylene and vinyl chloride in fixed or fluidized beds [266] is also possible. Conditions are very similar to those for the Oxy – EDC process (see page 37). Instead of vinyl chloride, acetylene can also be used as a feedstock [267]. In this case, the catalyst must be doped with mercury salts in order to achieve adequate conversion.

Hydrochlorination of 1,2-dichloroethylene [268] may be of interest only for the recovery of an unusable byproduct.

2.5.4. Uses and Economic Aspects.

As previously mentioned, 1,1,2-trichloroethane only plays a role as an intermediate for the production of 1,1,1-trichloroethane and 1,1-dichloroethylene.

The relatively high toxicity, which is typical for all 1,2-substituted chloroethanes, does not allow general use as a solvent.

Based on production figures for 1,1,1-trichloroethanes and 1,1-dichloroethylene, Western World production of 1,1,2-trichloroethane is estimated to be at 200 000 –220 000 t/year for 1984.

In Western Europe, approx. 40 000 t was produced in 1984.

2.6. 1,1,1,2-Tetrachloroethane

1,1,1,2-Tetrachloroethane [*630-20-6*] was first synthesized by A. Mouneyrat in 1898. Today it is a byproduct in many industrial chlorination reactions of C_2 hydrocarbons. It is, however, not produced on an industrial scale.

If recovered from such industrial processes as the production of 1,1,1- and 1,1,2-trichloroethane, it can be used as feedstock for the production of trichloroethylene (see Section 3.4.3) and perchloroethylene (see page 76).

Because of its high toxicity, it is not used as a solvent.

2.6.1. Physical Properties

M_r	167.86
mp	− 68.7 °C
bp at 101.325 kPa	130.5 °C
ϱ at 20 °C	1.5468 g/cm^3
n_D^{20}	1.4822
Viscosity at 20 °C	1.50×10^{-3} Pa s
Surface tension at 20 °C	32.1×10^{-3} N/m
Solubility of water in 1,1,1,2-tetrachloroethane at 20 °C	0.06 wt %

1,1,1,2-Tetrachloroethane is a colorless, nonflammable heavy liquid.

2.6.2. Chemical Properties

In general, 1,1,1,2-tetrachloroethane is more stable than its symmetrically substituted isomer. Thermal decomposition at 500 – 600 °C yields trichloroethylene and hydrogen chloride. Tetrachloroethylene can be formed by disproportionation [269]. The thermal decomposition is catalyzed by numerous compounds [270], mainly Lewis acids such as $FeCl_3$ and $AlCl_3$.

Dichloroacetyl chloride is obtained through oxidation [271].

2.6.3. Production

1,1,1,2-Tetrachloroethane is not produced on an industrial scale. It is an undesired byproduct mainly from the production of 1,1,1-trichloroethane from 1,1-dichloroethane, 1,1,2-trichloroethane and 1,1,2,2-tetrachloroethane from 1,2-dichloroethane. The most economical

use is its conversion to tetrachloroethylene in the chlorinolysis process (see page 76).

It can be prepared in highly purified form by isomerization of 1,1,2,2-tetrachloroethane or by chlorination of 1,1-dichloroethylene at approx. 40 °C in the liquid phase. Aluminum chloride is used in both reactions as a Lewis-acid catalyst [272].

2.7. 1,1,2,2-Tetrachloroethane

1,1,2,2-Tetrachloroethane [*79-34-5*] was first synthesized by M. Berthelot and E. Jungfleisch in 1869. Based on experiments by A. Mouneyrat, the first industrial scale production process was developed by A. Wacker in 1903. Thus, 1,1,2,2-tetrachloroethane became the first chloroethane to be produced in large quantities. For almost 70 years this process, which consists of the catalytic chlorination of acetylene, was the basis for the production of such important solvents as trichloroethylene (Tri) and tetrachloroethylene (Per).

However, with the continuing replacement of trichloroethylene by 1,1,1-trichloroethane and the development of more economical processes for the production of perchloroethylene, 1,1,2,2-tetrachloroethane has become less important for the production of chlorinated solvents.

2.7.1. Physical Properties

M_r	167.86
mp	− 42.5 °C
bp at 101.325 kPa	146.5 °C
ϱ at 20 °C	1.5958 g/cm^3
n_D^{20}	1.4942
Vapor pressure at	
0 °C	0.180 kPa
20 °C	0.680 kPa
60 °C	5.330 kPa
91 °C	18.700 kPa
118 °C	46.700 kPa
138 °C	82.700 kPa
Heat of formation (liquid) ΔH^0_{298}	− 195 kJ/mol^{-1}
Specific heat (liquid, 20 °C)	1.122 kJ kg^{-1} K^{-1}
Specific heat (vapor, 146.5 °C)	0.92 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	45.2 kJ/mol^{-1}
Critical temperature	688 K
Critical pressure	4000 kPa
Viscosity at 20 °C	1.77×10^{-3} Pa s
Surface tension at 20 °C	35.0×10^{-3} N/m
Coefficient of cubical expansion	0.00098 K^{-1}
Dielectric constant at 20 °C	8.00
Solubility in water at 20 °C	0.3 wt %
Solubility of water in 1,1,2,2-tetrachloroethane at 20 °C	0.03 wt %

1,1,2,2-Tetrachloroethane is a clear heavy, nonflammable liquid with a sweetish odor. It is well miscible with all common organic solvents and exhibits the highest solvency of all aliphatic chlorohydrocarbons.

1,1,2,2-Tetrachloroethane does not form explosive mixtures with air. Some binary azeotropes are shown in Table 17.

Table 17. Binary azeotropes formed by 1,1,2,2-tetrachloroethane [273]

wt %	Component	Azeotrope boiling point (101.3 kPa), °C
68.0	formic acid	99.3
55.0	cyclohexanone	159.1
9.0	ethylene glycol	145.1
7.0	isobutyric acid	144.8
1.8	monochloroacetic acid	146.3
60.0	propionic acid	140.4
45.0	styrene	143.5
31.1	water	93.2 (at 97.300 kPa)

2.7.2. Chemical Properties

If moisture, air, and light are excluded, 1,1,2,2-tetrachloroethane is sufficiently stable and can be stored without adding stabilizers. At elevated temperatures (> 400 °C), it is cracked to trichloroethylene and hydrogen chloride. Tetrachloroethylene may also be formed via disproportionation [269, 270, 274]. The thermal cracking reaction can be promoted by a variety of catalysts [270].

To avoid cracking during distillation, vacuum distillation is recommended. One patent claims soft, nondecompositional evaporation by means of a fluidized-bed evaporator [275].

In weak alkali solutions, dehydrochlorination to trichloroethylene occurs. In strong alkali solutions, explosive dichloroacetylene is formed. Decomposition in the presence of air can lead to small quantities of phosgene.

Chlorination under mild conditions (eventually induced by UV light or catalysts) yields hexachloroethane via the pentasubstituted intermediate [276]. Under more rigorous conditions and more thermodynamic control (chlorinolysis reaction, see Section 3.5.3), tetrachloroethyl-

ene and tetrachloromethane are formed as main products. Chlorination at 400 °C in the presence of charcoal favors cleavage which gives primarily tetrachloromethane and hydrogen chloride [277]. Strong acids may hydrolyze 1,1,2,2-tetrachloroethane to glyoxal.

With hydrogen, hydrodechlorination to 1,2-dichloroethylenes occurs.

Oxidation in air yields dichloroacetyl chloride.

2.7.3. Production

Industrial processes for the production of 1,1,2,2-tetrachloroethane consist of two main routes:

1) Addition of chlorine to acetylene:

$$C_2H_2 + 2\,Cl_2 \xrightarrow{\text{Cat.}} CHCl_2-CHCl_2 \qquad \Delta H^0_{298} = -422\ \text{kJ/mol}$$

2) Liquid-phase chlorination of ethylene or 1,2-dichloroethane

$$C_2H_4 + 3Cl_2 \rightarrow CHCl_2-CHCl_2 + 2HCl \qquad \Delta H^0_{298} = -436\text{kJ/mol}$$

$$CH_2Cl-CH_2Cl + 2Cl_2 \rightarrow CHCl_2-CHCl_2 + 2HCl \qquad \Delta H^0_{298} = -216\text{kJ/mol}$$

The acetylene route was used primarily in the past. The ethylene-based process was developed during the late 1960s, when the hydrocarbon feedstocks shifted toward ethylene. However, the acetylene process is still in use — mainly in the Federal Republic of Germany — where acetylene is readily available as a byproduct form naphtha crackers. Furthermore, the acetylene route has the distinct advantage of preferentially yielding the 1,1,2,2-isomer, which can easily be cracked to trichloroethylene. The ethylene-based process produces both isomers in an approximately equimolar ratio because of its radical nature.

1,1,2,2-Tetrachloroethane is also an incidental byproduct of other production processes for chlorinated hydrocarbons, such as the production of 1,1,1- and 1,1,2-trichloroethane. If necessary, it is separated together with the unsymmetric isomer and used for the production of trichloroethylene.

In the Atochem process [245] (see Figure25), 1,1,2,2-tetrachloroethane is produced from 1,2-dichloroethylenes by chlorination as an intermediate for trichloroethylene synthesis.

Several other processes have been patented, but all of them are only of minor importance.

1,1,2,2-Tetrachloroethane from Acetylene. The technical principle of carrying out this reaction has not changed much since the development of this process [278]. It is similar to the DC – EDC process (see page 35) and other common liquid-phase chlorination or hydrochlorination reactions.

The reaction is carried out in the liquid phase (tetrachloroethane) at 60 – 90 °C. The reactor pressure is reduced to ca. 20 kPa in order to prevent the explosion of the chlorine – acetylene mixtures. Lewis-acid catalysts [278], primarily $FeCl_3$, are dissolved in the tetrachloroethane solvent from the reactor sump phase. Gaseous acetylene and chlorine are fed to the reactor sump. The highly exothermic reaction provides enough heat to distill the tetrachloroethane [279].

The upper part of the reactor is therefore a distillation tower. Tetrachloroethane is withdrawn from appropriate trays, and the overheads consisting mainly of 1,2-dichloroethylenes and acetylene can be recycled to the reactor sump. To control the heavies concentration in the sump phase, a slipstream is withdrawn. Partial evaporation for tetrachloroethane recovery followed by high-temperature incineration with subsequent flue gas scrubbing is the best treatment for this stream containing the spent catalyst.

Carbon steel can be used as a construction material. Water and moisture should be strictly avoided in order to minimize corrosion and rapid deactivation of the catalyst.

Chlorine and acetylene yields of 90 – 98 % have been reported.

A similar process, using crude crack gas instead of pure acetylene, has been patented [280]. Besides 1,1,2,2-tetrachloroethane, this process also produces 1,2-dichloroethane from the ethylene fraction of the crack gas feedstream.

1,1,2,2-Tetrachloroethane from Ethylene and from 1,2-Dichloroethane. Liquid-phase chlorination of ethylene or the ethylene-induced chlorination of 1,2-dichloroethane is the same process as that used for the production of 1,1,2-

trichloroethane (see page 43). By increasing the chlorine : ethylene or chlorine: dichloroethane ratios and optimizing the residence time, an almost equimolar mixture of 1,1,2,2- and 1,1,1,2-tetrachloroethane is obtained as the main product [256, 281].

Several kinetic studies have been performed to determine the individual relative rate constants and to optimize the yield [281, 282].

At a temperature between 80 and 130 °C, chlorine conversion as high as 100 % and maximum ethylene conversions of 95 – 98 % can be achieved. Low-substituted products such as 1,2-dichloroethane and 1,1,2-trichloroethane can be recycled [283], so that yield losses occur only through the formation of penta- and hexachloroethane. With some minor modifications, the process is carried out as described earlier (see page 43). Instead of 1,2-dichloroethane, 1,2-dichloroethylenes may be fed to the reactor, which favors the formation of the symmetric isomer [284].

Similarly, the liquid-phase chlorination of mixtures containing a variety of chloroethanes and chloroethylenes has been patented [285].

Other Processes. The liquid-phase chlorination of vinyl chloride or 1,1,2-trichloroethane in the presence of $AlCl_3$ as catalyst yields 1,1,2,2-tetrachloroethane with high selectivity [286].

Specific catalysts made from graphite-intercalated copper or iron salts, alumina, and organopolysiloxanes specifically yield 1,1,2,2-tetrachloroethane by gas-phase chlorination (ca. 200 °C, 0.1 – 1.0 MPa) of mixtures comprising monochloroethane, 1,1- and 1,2-dichloroethane, and 1,1,2-trichloroethane [287].

The catalytic gas phase oxychlorination of 1,2-dichloroethane, ethylene, vinyl chloride, and 1,2-dichloroethenes has also been described [288].

Other processes use gas phase chlorination of 1,2-dichloroethane in a fluidized-bed [289] or liquid-phase photochlorination of 1,2-dichloroethane [290] or 1,2-dichloroethylenes [291].

2.7.4. Uses and Economic Aspects

1,1,2,2-Tetrachloroethane is almost always used as an intermediate in the production of trichloroethylene. Although it has a high solvency (e.g., as solvent for the production of chlorinated PVS [292]), it is very rarely used as a solvent because of its high toxicity.

Production figures for 1,1,2,2-tetrachloroethane cannot be estimated.

2.8. Pentachloroethane

Pentachloroethane [*76-01-7*] was first synthesized by V. Regnault in 1839 – 1840 by chlorination of monochloroethane.

In the past, pentachloroethane was produced as an intermediate for the tetrachloroethylene process (pentachloroethane pyrolysis). However, it is an unwanted byproduct of many production processes for chlorinated hydrocarbons and is mostly converted to tetrachloroethylene and tetrachloromethanes by chlorinolysis because other uses have almost disappeared.

2.8.1. Physical Properties

M_r	202.31
mp	− 29.0 °C
bp at 101.325 kPa	162 °C
ϱ at 20 °C	1.678 g/cm^3
n_D^{20}	1.5035
Vapor pressure at	
20 °C	0.470 kPa
60 °C	3.470 kPa
80 °C	7.860 kPa
100 °C	17.330 kPa
120 °C	33.330 kPa
140 °C	60.000 kPa
Heat of formation (liquid) ΔH^0_{298}	− 188.4 kJ/mol
Density of vapor (162 °C, 101.325 kPa)	5.68 g/L
Specific heat (liquid, 20 °C)	0.9 kJ kg^{-1} K^{-1}
Heat of evaporation at 298 K	45.6 kJ/mol
Viscosity at 20 °C	2.49×10^{-3} Pa s
Surface tension at 20 °C	34.7×10^{-3} N/m
Coefficient of cubical expansion	0.0009 K^{-1}
Dielectric constant at 20 °C	3.6
Solubility in water at 20 °C	0.05 wt %
Solubility of water in pentachloroethane at 20 °C	0.03 wt %

Pentachloroethane is a colorless, heavy, nonflammable liquid with a sweetish odor. It is miscible in most common organic solvents and does not form explosive mixtures with air.

Some binary azeotropes are shown in Table 18.

Table 18. Binary azeotropes formed by pentachloroethane

wt %	Component	Azeotrope boiling point (101.3 kPa), °C
3	acetamide	160.5
36	cyclohexanol	157.9
28	cyclohexanone	165.4
15	glycol	154.5
43	isobutyric acid	152.9
9.5	phenol	160.9
43.4	water	95.1 (at 97.300 kPa)

2.8.2. Chemical Properties

If moisture and air are eliminated, pentachloroethane shows good stability even at elevated temperatures (> 100 °C). Pyrolysis at a temperature above 350 °C yields tetrachloroethylene and hydrogen chloride [293]. The dehydrochlorination reaction is catalyzed by Lewis acids and activated alumina. Dehydrochlorination also occurs in the presence of weak alkali solution. Chlorination in the liquid phase in the presence of a catalyst or induced by ethylene (see page 43) yields hexachloroethylene. Dry pentachloroethane does not corrode iron; if it is stored over longer periods of time, however, the addition of amine stabilizers is recommended.

Dichloroacetyl chloride is formed with fuming sulfuric acid. Air oxidation in the presence of UV light gives trichloroacetyl chloride.

In the presence of hydrogen fluoride and Lewis acids such as $SbCl_5$–chlorine, substitution occurs.

2.8.3. Production

Since tetrachloroethylene is more economically produced by the chlorinolysis process (see page 76), industrial production of pentachloroethane has become unimportant and is presently no longer used.

If required, the synthesis can be performed by two routes:

1) Chlorination of trichloroethylene:

$$CHCl=CCl_2 + Cl_2 \xrightarrow[\text{or } h\nu]{\text{Cat.}} CHCl_2-CCl_3$$

The reaction is best carried out in the liquid phase. $FeCl_3$ is used as a catalyst, but UV irradiation can also be used. Stepwise chlorination in a cascade system is also possible [294].

2) Ethylene-induced chlorination of 1,2-dichloroethane [256]:

$$\underset{Cl}{CH_2}-\underset{Cl}{CH_2} + 3\,Cl_2 \longrightarrow CHCl_2-CCl_3 + 3\,HCl$$

This reaction is similar to that of 1,1,2-trichloroethane (see page 43). Pentachloroethane is obtained with lighter chlorinated products, which can be rechlorinated. To avoid decomposition, pentachloroethane should be distilled at reduced pressure.

By oxychlorination of ethylene, 1,2-dichloroethane, or other chlorinated C_2 hydrocarbons, pentachloroethane is also obtained.

The main industrial source, however, is the photochemical production of 1,1,1-trichloroethane (see page 43) and the liquid-phase chlorination process for 1,1,2-trichloroethane production (see page 43). Pentachloroethane formed in these processes is frequently not isolated but fed together with the tetrachloroethanes to the chlorinolysis process.

2.8.4. Uses and Economic Aspects

Because of its low stability and toxicity, uses for pentachloroethane as a solvent (cellulose derivatives, rubbers, and resins) are insignificant.

About 10 000 – 20 000 t/a (1984) of pentachloroethane may be produced as a byproduct in the Western World. Most of it is used for the production of tetrachloroethylene and carbon tetrachloride.

2.9. Hexachloroethane

Hexachloroethane [*67-72-1*] is at ambient temperature the only solid compound of all chlorinated ethanes and ethylenes. Because it has specific properties, such as a tendency to sublime and a very high chlorine content, it has some specific applications, which are limited, however, for toxicological and ecological reasons.

2.9.1. Physical Properties

M_r	236.74
mp (closed capillary, as sublimes)	185 °C
bp at 101.325 kPa	185 °C
Crystal structure: rhombic	< 46 °C
triclinic	46 – 71 °C
cubic	> 71 °C
Specific density at 20 °C	2.094 g/cm^3
Vapor density at 185 °C	6.3 g/L
Vapor pressure at	
20 °C	0.290 kPa
40 °C	1.330 kPa
80 °C	2.400 kPa
120 °C	11.600 kPa
160 °C	45.320 kPa
180 °C	86.650 kPa
Heat of formation (liquid) ΔH^0_{298}	− 203.4 kJ/mol
Specific heat at 20 °C	0.615 kJ kg^{-1} K^{-1}
Heat of sublimation at 298 K	59 kJ/mol
Cryoscopic constant	5.6
Solubility in water at 22 °C	50 mg/kg

Hexachloroethane forms white crystals with a camphor-like odor. It is not flammable.

Some binary azeotropes of hexachloroethane are shown in Table 19.

Table 19. Binary azeotropes formed by hexachloroethane

wt %	Component	Azeotropic boiling point (101.3 kPa), °C
34	aniline	176.8
12	benzylic alcohol	182.0
25	monochloroacetic acid	171.2
28	*o*-cresol	181.3
30	phenol	173.7
15	trichloroacetic acid	181.0

2.9.2. Chemical Properties

Hexachloroethane is fairly stable and sublimes without decomposition. At temperatures above 250 °C, especially at 400 – 500 °C, it is cracked (disproportionated) [295]:

$$2C_2Cl_6 \rightarrow C_2Cl_4 + 2CCl_4$$

Small amounts of chlorine are also formed. With metals like iron, zinc, and aluminum, chlorination reactions start at a higher temperature, forming metal chlorides and tetrachloroethylene. This reaction can be used for the synthesis of pure metal chlorides and for ultrapurification of metals.

At moderate temperature hexachloroethane is stable against aqueous alkali and acids. At temperatures above 200 °C, hydrolysis to oxalic acid occurs.

2.9.3. Production

Limited industrial uses of hexachloroethane do not justify large-scale production processes.

A primary source for hexachloroethane is from the production of tetrachloroethylene and carbon tetrachloride by chlorinolysis of hydrocarbons and chlorinated hydrocarbon residues (see page 76). It can be separated from the residues by distillation and fractionated crystallization.

For the intentional production of hexachloroethane, tetrachloroethylene is chlorinated batchwise in presence of iron chloride. The hexachloroethane crystallizes from the mother liquor and is isolated. The mother liquor is recycled and again chlorinated [296].

The photochemical chlorination of tetrachloroethylene is performed similarily [297].

2.9.4. Uses and Economic Aspects

Industrial uses of hexachloroethane are diminishing, so that most hexachloroethane is either recycled or incinerated with HCl — or chlorine recovery depending on the individual technology applied.

Smaller quantities are used for the synthesis of metal chlorides or for the production of fluorocarbons.

Hexachloroethane is one of the more toxic chloroethanes. Its use in plasticizer or rubber formulations is, therefore, decreasing.

Since no significant applications exist, production figures cannot be estimated.

Analysis, quality control, storage, and transportation of the chloroethanes are treated with the chloroethylenes in Sections 3.6 and 3.7

3. Chloroethylenes

The class of chloroethylene comprises:

Vinyl chloride (monochloroethylene, VCM)
1,1-Dichloroethylene (vinylidene chloride)
cis- and *trans*-1,2-Dichloroethylene
Trichloroethylene (Tri)
Tetrachloroethylene
(perchloroethylene, Per, Perc).

3.1. Vinyl Chloride (VCM)

In addition to ethylene and NaOH, vinyl chloride [*75-01-4*] is one of the world's most important commodity chemicals. The 1984 worldwide consumption averaged 12 – 15 million t/a. About 25 % of the world's total chlorine production is required for its production.

The importance of vinyl chloride results from the widespread use of poly(vinyl chloride), one of the most important polymers.

The first synthesis of vinyl chloride dates back to 1830 – 1834 when V. Regnault obtained it by dehydrochlorinating 1,2-dichloroethane with alcoholic potash. In 1902, it was obtained by Biltz during thermal cracking of the same compound.

However, at that time, the state of the art in polymer science and technology was not very sophisticated, and this discovery did not lead to industrial or commercial consequences.

The basic work of F. Klatte [298] on the polymerization of vinylic compounds gave rise to the industrial production of vinyl chloride in the 1930s.

Vinyl chloride was obtained by Klatte in 1912 through catalytic hydrochlorination of acetylene [299]. This route was almost exclusively used for nearly 30 years. Because of the high energy requirements for acetylene production, its replacement by a cheaper substitute was a challenge for a long time.

From 1940 – 1950 on, acetylene could be partially replaced by ethylene, from which vinyl chloride was produced by direct chlorination to 1,2-dichloroethane and subsequent thermal cracking. The first large production units for this route were first constructed by Dow Chemical Co., Monsanto Chemical Co. and the Shell Oil Co. In these plants, the balance of HCl generated by dichloroethane cracking, however, was still achieved by acetylene hydrochlorination.

The complete changeover to the exclusive use of ethylene as a feedstock became possible when the large-scale oxychlorination of ethylene to 1,2-dichloroethane (see page 37) had been proven to be technically feasible (Dow Chemical, 1955 – 1958).

Since then, most plants use integrated, balanced DC – EDC – Oxy – EDC – VCM processes and more than 90 % of the vinyl chloride presently produced in the Western World is based exclusively on ethylene.

In addition to its use as an intermediate in the production of trichloroethane (1,1,1- and 1,1,2-trichloroethane), most vinyl chloride is used for polymerization to PVC.

With the use of plasticizers and because of its high energy efficiency, PVC has become one of the most important industrial polymers. Even though it is one of the oldest polymers, its ready availability, relatively inexpensive production by large plants, and the continuing development of new formulations [300] with widespread uses secure its attractiveness in the future.

Several VCM plants were under construction in 1986. Due to the feedstock and market situation, the new plants will be preferentially located either in oil-producing or in developing countries.

3.1.1. Physical Properties

M_r	62.5
mp	− 153.8 °C
bp at 101.325 kPa	− 13.4 °C
ϱ at − 14.2 °C	0.969 g/cm^3
at 20 °C	0.910 g/cm^3
n_D^{20}	1.445
Vapor pressure at	
− 30 °C	51.000 kPa
− 20 °C	78.000 kPa
− 10 °C	115.000 kPa
0 °C	165.000 kPa
10 °C	243.000 kPa
20 °C	333.000 kPa
30 °C	451.000 kPa
40 °C	600.000 kPa
50 °C	756.000 kPa
Heat of formation (gaseous) ΔH^0_{298}	+ 35.2 kJ/mol
Specific heat (liquid, 20 °C)	1.352 kJ kg^{-1} K^{-1}
(vapor, 20 °C)	0.86 kJ kg^{-1} K^{-1}
Heat of evaporation (259.8 K)	20.6 kJ/mol
Critical temperature	429.8 K
Critical pressure	5600 kPa
Viscosity at − 40 °C	0.34×10^{-3} Pa s
at − 10 °C	0.25×10^{-3} Pa s
at 20 °C	0.19×10^{-3} Pa s
Dielectric constant at 17.2 °C	6.26
Flash point (open cup)	− 78 °C
Autoignition temperature	472 °C
Explosive limits in air	4 – 22 vol%
Solubility in water at 20 °C	0.11 wt %
Solubility of water in vinylchloride at − 15 °C	300 mg/kg

Vinyl chloride is a colorless, flammable gas at ambient temperature with a sweetish odor. It

is soluble in most common organic liquids and solvents.

3.1.2. Chemical Properties

If oxygen and air are excluded, dry, purified vinyl chloride is highly stable and noncorrosive.

Above 450 °C, partial decomposition occurs yielding acetylene and hydrogen chloride. Trace amounts of 2-chloro-1,3-butadiene (chloroprene) may also be formed by acetylene-precursor dimerization.

Air combustion products of vinyl chloride are carbon dioxide and hydrogen chloride. Under oxygen deficient combustion, traces of phosgene may be formed. In oxidation reactions sensitized by chloride, monochloroacetaldehyde and carbon monoxide are obtainable from vinyl chloride [301].

In the presence of water, hydrochloric acid, which attacks most metals and alloys is formed. This hydrolysis most probably proceeds via a peroxide intermediate [302].

With air and oxygen, very explosive peroxides can be formed.

Because of the vinylic double bond, the most important reactions are polymerization reactions (co- and homopolymerization, → Poly(Vinyl Chloride)) and electrophilic or radicalic addition reactions — mainly chlorination or hydrochlorination — to yield 1,1,2-trichloroethane or 1,1-dichloroethane .

The substitution of the chlorine atom is more difficult to achieve. Vinyl anion addition reactions offering interesting synthetic routes are possible via the vinylmagnesium [303] and vinyllithium compounds [304].

Catalytic halogen exchange by hydrogen fluoride gives vinyl fluoride [305].

3.1.3. Production

The industrial production of vinyl chloride is based on only two reactions:

1) Hydrochlorination of acetylene:

$$C_2H_2 + \mathrm{HCl} \rightarrow \mathrm{CH_2 = CHCl}$$
$$\Delta H^0_{298} = -99.2\,\mathrm{kJ/mol}$$

2) Thermal cracking of 1,2-dichloroethane:

$$\mathrm{Cl{-}CH_2{-}CH_2{-}Cl \rightarrow CH_2 = CHCl + HCl}$$
$$\Delta H^0_{298} = -100.2\,\mathrm{kJ/mol}$$

All other reactions yielding vinyl chloride are industrially unimportant at present.

Acetylene hydrochlorination was mainly used in the past, when acetylene — produced via calcium carbide from coal — was one of the most important basic feedstocks for the chemical industry.

With the large-scale production of ethylene-derived polymers, such as polyethylene and polystyrene, and the general trend toward natural gas (United States), naphtha, and gas oil (Europe) as basic feedstocks, the cracker capacity increased substantially and ethylene became readily available at very competitive prices.

Besides the economic disadvantage of the higher priced hydrocarbon feed, the acetylene hydrochlorination has the drawback of not being balanced on the chlorine side because it requires only hydrogen chloride as a chlorine source.

With increasing demand for vinyl chloride and technical progress, the first balanced processes were established in the 1940s and 1950s, when acetylene was partially replaced by ethylene, which was converted to vinyl chloride by direct chlorination to 1,2-dichloroethane and subsequent thermal cracking. The hydrogen chloride from cracking could then be used for acetylene hydrochlorination:

$$\begin{array}{l}
\mathrm{C_2H_4 + Cl_2 \longrightarrow C_2H_4Cl_2} \\
\mathrm{C_2H_4Cl_2 \longrightarrow C_2H_3Cl + HCl} \\
\mathrm{C_2H_2 + HCl \longrightarrow C_2H_3Cl} \\
\hline
\mathrm{C_2H_2 + C_2H_4 + Cl_2 \longrightarrow 2\,C_2H_3Cl}
\end{array}$$

By direct use of crack gas, without separation of ethylene and acetylene, this process is still pursued with some modifications.

With the introduction of the first large-scale Oxy – EDC plant (see page 37) by The Dow Chemical Co. in 1958, a balanced process based only on inexpensive ethylene became available and found rapid acceptance within the chemical industry.

Using this balanced process, vinyl chloride is made only by thermal cracking of 1,2-dichloroethane, which in turn is produced by direct chlorination or oxychlorination of ethylene. The latter process balances hereby the hydrogen chloride formed during cracking:

$$CH_2 = CH_2 + Cl_2 \longrightarrow C_2H_4Cl_2 \quad \text{DC-EDC}$$
$$CH_2 = CH_2 + 2\,HCl + 1/2\,O_2 \longrightarrow C_2H_4Cl_2 + H_2O \quad \text{Oxy-EDC}$$
$$2\,CH_4Cl_2 \longrightarrow 2\,C_2H_3Cl + 2\,HCl \quad \text{Cracking}$$
$$2\,CH_2 = CH_2 + Cl_2 + 1/2\,O_2 \longrightarrow 2\,C_2H_3Cl + H_2O$$

Presently, more than 90 % of the vinyl chloride produced is based on this route.

Since ca. 1960, considerable efforts were undertaken to replace ethylene by ethane as the basic feedstock. All ethane-based processes developed so far, however, lack selectivity, which causes increased recycle rates and losses by side reactions and requires higher capital expenditures, so that the cost advantage even for grassroot plants is marginal.

The ethylene-based processes have been improved considerably by numerous process modifications which resulted in higher yields and lower energy requirements.

However, with the development and availability of new catalysts, the ethane route may become more attractive in the future.

Due to sharply increasing energy costs, chlorine has also become a very important cost factor in vinyl chloride production. This explains the research and development on an electrolysis-free route to vinyl chloride.

3.1.3.1. Vinyl Chloride from Acetylene

The catalytic hydrochlorination of acetylene is possible in either the gaseous or the liquid phase. The gas-phase reaction dominates in industrial processes.

In this process the gaseous reactants are brought into contact with the catalyst at slightly increased pressure (0.1 – 0.3 MPa) and 100 –250 °C (contact time 0.1 – 1 s) and then quenched and partially liquified. The reaction products are separated, recycled, or submitted to final purification.

The molar feed ratios, varying from almost equimolar to a 10-fold excess of HCl, depend heavily on catalyst performance. Acetylene conversions of 95 – 100 % at almost quantitative yields are achieved.

The acetylene fed to the reactor has to be free of common catalyst poisons such as sulfur, phosphorus and arsenic compounds [306]. Unsaturated hydrocarbons must also be minimized in the feed because they may clog and inactivate the catalyst upon polymerization [307]. The hydrogen chloride must be free of chlorine to avoid explosion and should not contain chlorinated hydrocarbons, which could also act as catalyst poisons.

Water must be entirely excluded to avoid corrosion on structural materials. Because of gas-phase reaction and anhydrous conditions, most equipment is made from carbon steel. If water is used for quenching or HCl absorption, either brick-lined or polymer-made equipment is used for these parts.

Although fluidized-bed reactors have been patented [308], fixed-bed, multitubular reactors (tube size 1″ – 4″ inside diameter, 10 – 20 ft) are almost exclusively used. To avoid volatilization of the catalyst (tube side), temperature control and near to isothermal operation of the reactor is achieved by external cooling using a hot oil system or water. The heat of reaction can be transferred and used in the reboilers of the downstream purification equipment [309]. A special reactor design has been patented for use of diluted hydrogen chloride from incineration processes [310]. Mercury(II) chloride on activated carbon is used primarily as a catalyst in concentrations of 2 – 10 wt %. Several other metallic catalysts [311] as well as 1 – 3 vol% chlorine [312] have been proposed or patented. However, the mercury salt had proven to be the most effective. The reaction rate is first order with respect to acetylene [311, 313].

Activated carbon with specific properties is preferentially used as support [314]. Several patents deal with appropriate pretreatment procedures, such as oxidation and thermal activation, to improve these properties [315].

Zeolites and molecular sieves were also found to be suitable support materials [316].

Because the volatility of mercury is a very limiting factor for reactor operation and throughput, additives such as cerium chloride [317], thorium [318] and copper chloride [319], as well as polymers [320], have been proposed to reduce volatility. Mercury chloride – graphite-intercalated catalyst is also thought to possess only a very low sublimation tendency [321]. Because volatization of the mercury catalyst cannot be completely avoided, the loss of catalyst activity by moving hot spots must be minimized by operational means e.g. reversal of the flow through

the reactor. Other possibilities are the adjustment of the reactor load depending on the catalyst activity [322], operating two reactors in series, using fresh catalyst in the second to complete conversion from the less active first reactor at reduced heat load [323], or the fine tuning of catalyst activity and heat transfer liquid flow in a two-reactor system [324].

With fixed-bed reactor systems, time – space yields up to 300 kg m^{-3} h^{-1} are possible. An average of 70 – 80 kg m^{-3} h^{-1} is achieved over the lifetime of the catalyst.

The mercury is removed from the spent catalyst by either thermal (pyrolytic) treatment [325] or steam desorption [326]. The mercury-free carbon can be incinerated or reactivated.

Process Description (Fig. 21). Acetylene and hydrogen chloride are mixed and fed with recycle gas to the reactor. The gases leaving the reactor are compressed and fed to a first tower, where most of the vinyl chloride is withdrawn as a liquid from the bottom. Most of the overhead product (HCl, C_2H_2 and C_2H_3Cl) is recycled to the reactor. For removal of inert matter, a small part of this recycle stream is drawn off and washed with heavies — preferably 1,1-dichloroethane formed by competitive addition of HCl to vinyl chloride — to recover vinyl chloride and acetylene.

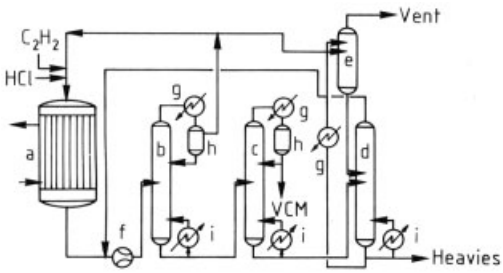

Figure 21. Production of vinyl chloride from acetylene and hydrogen chloride (schematic)
a) Reactor; b) Lights column; c) VCM column; d) Heavies stripper; e) Vent wash tower; g) Cooler; h) Knock-out drum; i) Reboiler

In the second tower, the crude vinyl chloride is purified and withdrawn at the head section. The heavy bottoms are submitted to a final stripping in the heavies column with the underflow of the washing tower and removed at the bottom for further use or for incineration. The overheads from the heavies treatment (acetylene and vinyl chloride) are recirculated to the compressor suction for optimal product recovery.

Other processes separate the heavy byproducts during the first distillation stage [327] or use water for HCl scrubbing [328, 329]. The resultant concentrated hydrochloric acid obtained can then be used for the final drying of the acetylene feed, which can contain significant amounts of moisture when produced by calcium carbide hydrolysis [329].

All process modifications attempt to achieve maximum recovery of the expensive acetylene and to avoid high pressures and temperatures, which may cause losses through polymerization [327, 330].

Vinyl Chloride from Crack Gases (Fig. 22). In the crack gas processes for vinyl chloride manufacture, unpurified acetylene produced by high-temperature cracking of naphtha [307, 331, 332] or methane [333, 334] is used. These processes are adventageous in that they do not require the cost-intensive separation of acetylene –ethylene mixtures [335].

The crack gas is fed directly to the hydrochlorinator, and the acetylene is converted to vinyl chloride which is then separated from the remaining constituents. Because all of the acetylene is consumed, the remaining ethylene is more easily separated or it can be introduced to a direct chlorination stage, where it is chlorinated to 1,2-dichloroethane, which is subsequently cracked to vinyl chloride [307, 331].

Since almost equimolar amounts of ethylene and acetylene can be achieved in the crack gas, the process can be balanced for chlorine.

Higher pressures (1.0 – 3.0 MPa) must be applied for the hydrochlorination stage in order to keep the reactor size reasonable. Because the acetylene is very diluted, hot spots are a smaller problem than with the pure acetylene process. For the chlorination and the cracking stage, standard technology can be used.

In a process developed by Solvay [332], hydrochlorination and chlorination are carried out together. The patent claims high yields without substantial formation of 1,1,2-trichloroethane, which can be formed by the addition of chlorine to vinyl chloride.

Another process modification [334] uses the quenching of the crack gases with chlorine at ca.

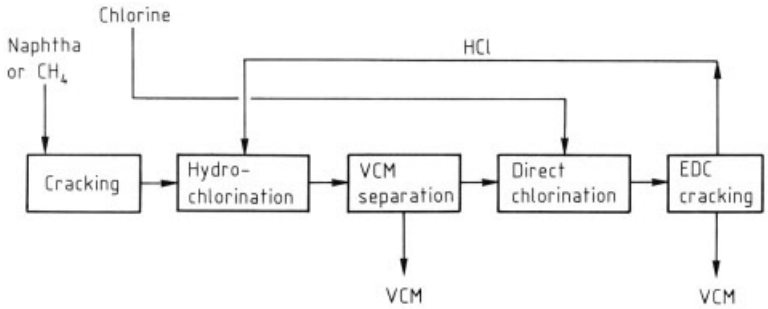

Figure 22. Balanced process for the production of vinyl chloride from crack gases

400 °C, which leads directly to vinyl chloride in yields up to 60 %. An acetylene-based process uses HCl generated from magnesium chloride hydrate by pyrolysis [336].

All acetylene-based processes, however, have the distinct drawback of using, at least partially, the more expensive hydrocarbon feed. Completely ethylene-based processes are economically superior and in only a few cases is the acetylene route still competitively pursued.

3.1.3.2. Vinyl Chloride from 1,2-Dichloroethane

The cracking reaction of 1,2-dichloroethane can be carried out in the liquid or gas phase.

The *liquid-phase* dehydrochlorination of 1,2-dichloroethane is industrially unimportant because expensive chlorine is lost as a salt when 1,2-dichloroethane is treated with alkaline solutions:

$$CH_2Cl{-}CH_2Cl + NaOH \rightarrow CH_2 = CHCl + NaCl + H_2O$$

In addition, the aqueous process stream to be discarded poses severe environmental problems or requires extensive pretreatment. Even though a good dehydrochlorination reaction can be achieved by using phase-transfer catalysts [337], this process is not suitable or economical for large-scale production.

The *gas-phase* dehydrochlorination is the most important route and industrially used for the production of vinyl chloride.

It can be carried out as a pure pyrolytic reaction or in the presence of catalysts.

The noncatalyzed process is used by the majority of the vinyl chloride producers (e.g., Dow Chemical, Ethyl, B. F. Goodrich, Hoechst, ICI, Mitsui Toatsu, Monsanto, Stauffer), whereas only a few producers (e.g., Wacker) use catalytic cracking.

Improved furnace designs for the noncatalytic reaction have made conversions and yields comparable to those obtained by catalytic cracking.

Because of the time-consuming catalyst removal, shutdown periods are considerably longer for catalytic furnaces and the catalyst is an additional cost factor, so that pure thermal cracking may be currently the more economical process.

Noncatalytic Gas-Phase Reaction. The reaction occurs via a first-order free radical chain mechanism [253, 338], which starts with the homolytic cleavage of a C – Cl bond

1)	$ClCH_2{-}CH_2Cl$	$\rightarrow ClCH_2{-}C{\cdot}H_2 + Cl\,\cdot$
2)	$Cl\,\cdot + ClCH_2{-}CH_2Cl$	$\rightarrow ClCH_2{-}C{\cdot}HCl + HCl$
3)	$ClCH_2{-}C{\cdot}HCl$	$\rightarrow CH_2{=}CHCl + Cl\,\cdot$
	$Cl\,\cdot + ClCH_2{-}CH_2Cl$	$\rightarrow ClCH_2{-}C{\cdot}HCl + HCl$
		etc.

The intermediate dichloroethane radical is stabilized by elimination of a chlorine radical, which propagates the chain.

The radical chain is terminated by recombination (reverse reaction to initiation) or wall collisions, as it is usual for this type of reaction.

Since chlorine or other radical species are important for the chain propagation, chlorine, [339 – 341] or chlorine delivering compounds such as tetrachloromethane [342] or hexachloroethane [340], as well as other radicals like oxygen [340] and nitrous oxide [340, 343] or other halogens (bromine and iodine) [341] can be added as initiators and promoters. The use of oxygen, however, is controversial because oxygen was also found to enhance coking of the

furnace walls [344]. Because chlorine is readily available in vinyl chloride plants and because of its minimal interference, chlorine is primarily used as a promoter. Promoter concentration in the 1,2-dichloroethane feed may vary between a few hundred mg/kg and up to 5 %. Good results were achieved when chlorine was fed at different points to the reaction zone [339], which may, however, be difficult to realize. When nitromethane was used as a promoter, high yields have been reported [345]. The addition of 1,1,2-trichloroethane was found to inhibit coke formation [346].

Even though 1,1-dichloroethane is more difficult to crack, good conversions are obtainable if the 1,1-dichloroethane concentration does not exceed 10 % in the feed [143]. The crack reaction is industrially carried out at temperatures between 400 and 650 °C, preferably, however, between 500 and 550 °C. Reactor pressure may vary from 0.1 to 4.0 MPa. However, high-pressure processes (2.0 – 3.0 MPa) are preferred because high pressure reduces furnace size, improves heat transfer, and makes the downstream separation easier, due to increased boiling points. Mean residence time is about 10 – 20 s. The 1,2-dichloroethane conversion is kept at 50 – 60 % per pass to control byproduct formation and coking, which significantly increases at higher conversion rates and causes yield losses. At these conversion rates, vinyl chloride yields of 95 – 99 % are obtainable. High-purity 1,2-dichloroethane should be used because most impure technical-grade dichloroethane reduces conversion.

The crack furnace has a plug-flow reactor design with one or more tubes (1 – 10″ diameter, up to 4000 feet long) being placed in the convection zone of the furnace. The furnace may be equipped with a single burner or have multiburner design. In most cases, natural gas is used as a burner feed; however, some plants use hydrogen-driven furnaces, using hydrogen from on-site chlor-alkali plants. Feed evaporation at ca. 200 °C and cracking at a much higher temperature is often carried out in the same furnace to make the best use of the fuel gas. Evaporation is preferably carried out in the upper, cooler part of the convection zone, whereas the cracking must take place in the lower, hotter part. Chromium –nickel alloys are the best construction materials [167].

Although several furnace designs have been patented [347] and the basic principles are quite similar, most vinyl chloride producers have developed their proprietary furnace technology for optimal yield and low shutdown frequency for pipe decoking. After leaving the reactors, the gases must be cooled down immediately to avoid yield losses by formation of heavy products. In most processes, this is achieved by a quench tower, where condensed and cooled 1,2-dichloroethane is recirculated at high rates. The heat withdrawn from the quench tower can be used for the reboilers of downstream distillation stages [348]. It is also possible to quench in two stages, first by indirect cooling (transfer line heat exchanger) and then by direct quenching [349]. Thus, substantial heat recovery is possible, which can be used by a hot oil system in other process stages. Only indirect cooling at the furnace tail pipe bears an increased risk of plugging the heat exchanger with coke and heavy byproducts.

For the downstream separation of the main constituents of the reaction gases, vinyl chloride, hydrogen chloride, and 1,2-dichloroethane, many processing possibilities have been patented [350]. However, a common principle is to first separate hydrogen chloride and then vinyl chloride from the reaction mixture by distillation. 1,2-Dichloroethane is then distilled from the remaining heavies or the whole stream is sent without separation to the DC – EDC section [351] (see page 35), where it can be economically purified.

The byproducts of 1,2-dichloroethane cracking can be theoretically divided into two groups:

1) *Volatile impurities* such as ethylene, acetylene, vinylacetylene, 1,3-butadiene, 2-chloro-1,3-butadiene, benzene, chlorobenzene, 1,2- and 1,1-dichloroethylene, 1,1-dichloroethane, 1,1,1- and 1,1,2-trichloroethane, methyl and methylene chloride, chloroform, and tetrachloromethane.
2) *Tars and coke*. To remove these, special filters are used [352].

2-Chloro-1,3-butadiene (chloroprene) forms tarry polymerization products, which plug the equipment when separated from 1,2-dichloroethane together with other light products. If this separation is not performed with the DC – EDC process, chlorination of the heavies is used to

improve separation and to avoid excessive plugging of the equipment [353, 354].

Other volatile impurities must also be removed by posttreatment because they cannot be completely separated from the main products by distillation.

Acetylene, which codistills with the hydrogen chloride, is either converted to vinyl chloride by catalytic hydrochlorination [355] or selectively hydrogenated to ethylene [356], which does not interfere when the hydrogen chloride is used in the Oxy-EDC-process (see page 37. 1,3-Butadiene, which mostly contaminates the vinyl chloride fraction, can be removed by polymerization during extended residence time [357] or with Lewis-acid catalysts [358] by chlorination [353, 359], hydrochlorination [360], hydrogenation [361], or by reaction with chlorosulfuric acid [362].

The vinyl chloride obtained by distillation is suitable for polymerization. If necessary, remaining impurities can be removed by extractive distillation with acetonitrile [363], distillation in the presence of alcohols [364], or orthoesters [365] as acid scavengers, or treatment with calcium oxide [366] or zinc [367]. Removal of ionic species by an electrostatic field was also proposed [368]. Remaining traces of 1,3-butadiene can be removed by clay adsorbents [369].

Process Description (Fig. 23). Pure 1,2-dichloroethane is fed to the evaporator in the upper part of the cracking furnace. The gas phase is separated from the remaining liquids and fed to the cracking zone. After having passed the cracking zone in the furnace, the gases are cooled and quenched.

Hydrogen chloride is removed from the reaction mixture in the first distillation tower and sent back to the Oxy – EDC process or used for other purposes (e.g., methanol hydrochlorination). Vinyl chloride is distilled in the second tower and drawn off as a head product. It can be washed with diluted caustic in order to remove the last traces of hydrogen chloride and 1,2-dichloroethane.

The bottoms of the vinyl chloride column are purified in two more distillation stages. First, the low-boiling impurities are removed in the light ends column, followed by 1,2-dichloroethane separation from the heavy ends in the last tower. The purified 1,2-dichloroethane is recycled to the cracking furnace.

In integrated processes (see page 62), the last two stages are more economically combined with the purification of 1,2-dichloroethane from the DC and Oxy train.

If high temperature DC processes are used (see page 35), 1,2-dichloroethane can be purified in the DC reactor.

Catalytic Gas-Phase Reaction. The catalytic gas-phase dehydrochlorination is only used by a minority of vinyl chloride producers [306, 307, 370]. Higher selectivities toward vinyl chloride and less formation of coke, which is mainly due to the lower temperatures (200 – 450 °C), are often claimed as advantages.

1,2-Dichloroethane conversions, however, are not much improved compared to the noncatalytic process. On the average, 60 – 80 % but mostly 60 – 70 % conversion per pass is obtained.

In addition to activated carbon [371], which can be doped with ammonium salt promoters [372], a variety of other materials has been patented as catalysts, consisting of silicates [373], metal-promoted alumina [374], sodium chloride [375], and zeolites [376]. The dehydrochlorination of 1,2-dichloroethane by melts containing copper or other metals has also been described [377] (see also Transcat process, Section 3.1.3.4).

With the development of improved noncatalytic gas-phase processes, the catalytic route has lost most of its economic attractiveness. The higher costs of catalytic processes for catalyst and extended shutdown periods no longer compensate the slightly higher energy requirements of modern yield- and energy-optimized noncatalytic processes.

Photochemically Induced Gas-Phase Dehydrochlorination. Considerable improvements in conversion and product quality were obtained by combining the thermal noncatalytic gas-phase reaction with a photochemical postreaction [378]. Either polychromatic actinic light from mercury, thallium, or tungsten lamps or, preferably, monochromatic light from suitable lasers is used as a light source for the excitation of 1,2-dichloroethane. The excited molecules then liberate chlorine atoms, which in turn start the free radical chain reaction:

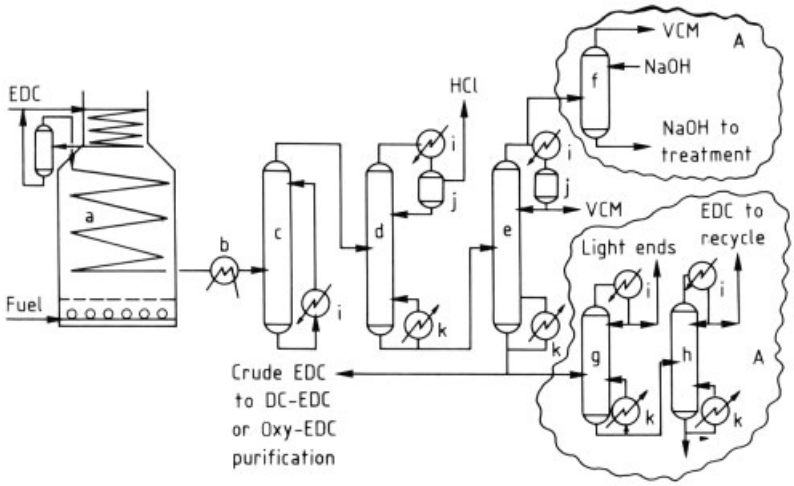

Figure 23. Schematic flow sheet for production of vinyl chloride by thermal cracking of 1,2-dichloroethane
a) Crack furnace; b) Transfer pipe heat exchanger; c) Quench tower; d) HCl distillation tower; e) VCM purification tower; f) VCM wash tower; g) Light-end tower; h) EDC – heavy end tower; i) Cooler; j) Knock-out drum; k) ReboilerA Process Modification

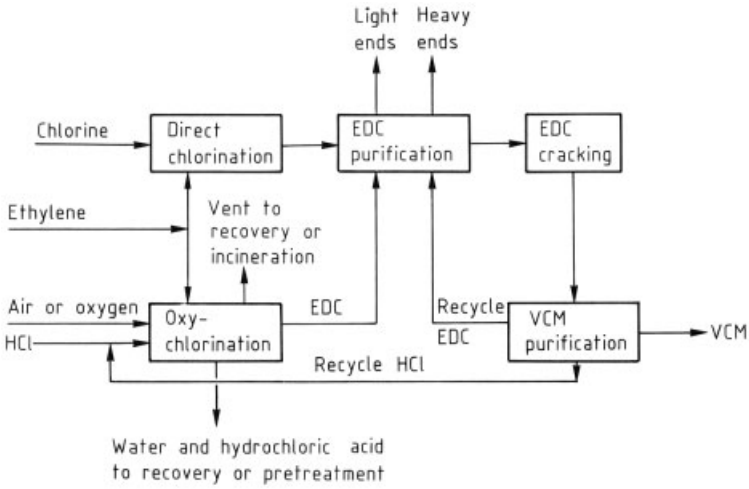

Figure 24. Ethylene-based integrated balanced process for the production of vinyl chloride

$C_2H_4Cl_2 + h\nu$	$\rightarrow (C_2H_4Cl_2)^*$	excited state
$(C_2H_4Cl_2)^*$	$\rightarrow C_2H_4Cl\cdot + Cl\cdot$	
$Cl\cdot + C_2H_4Cl_2$	$\rightarrow C_2H_3Cl_2\cdot + HCl$	
$C_2H_3Cl_2\cdot$	$\rightarrow C_2H_3Cl + Cl\cdot$	

A photochemical postreactor implemented in existing thermal processes allows higher conversion rates at increased selectivity and decreased energy consumption.

This, however, has not yet been proven on an industrial scale.

Combined Process (Fig. 24). Most of the vinyl chloride is presently produced in so-called integrated, balanced processes comprising three units [179, 379].

1) direct ethylene chlorination
2) ethylene oxychlorination
3) 1,2-dichloroethane cracking

Ethylene and chlorine are basic feedstocks which are reacted in the direct chlorination unit to yield 1,2-dichloroethane. Additional 1,2-dichloroethane is produced in the oxychlorination process. The combined streams are fed to the cracking train, where vinyl chloride is obtained. The hydrogen chloride formed during cracking

is recycled and consumed in the oxychlorination process. Thus, the process is balanced on hydrogen chloride. Integrated processes are not only advantageous due to lower energy requirements, due to the fact that energy-consuming steps can be combined with exothermic reactions, but they also allow variations in the chlorine distribution of manufacturing sites producing other chlorinated hydrocarbons and convert the largely unusable byproduct hydrogen chloride into a valuable product.

Differences in process technology used by the individual vinyl chloride producers are mainly due to the different technologies applied in the processes, such as high- or low-temperature direct chlorination, fixed- or fluidized-bed oxychlorination, which determine the needs for purification equipment and the energy requirements.

The most important processes presently used are described in detail elsewhere [175]. The major unit ratios for a balanced, air-based process are given in Table 20.

Table 20. Major unit ratios for an integrated, balanced vinyl chloride process [380] (oxychlorination is air-based)

Component	Unit ratio (kg/kg VCM)
Raw materials:	
Ethene	0.4656
Chlorine	0.5871
Air	0.7322
Water	0.0171
Byproducts:	
Lights	0.0029
Heavies	0.0023
Vents	0.6727 (0.5779 of this is N_2)
Aqueous streams	0.1218

New Developments. A new route was developed [381] which allows chlorine-free production of vinyl chloride by bringing ethylene in contact with an aqueous solution containing copper(II) chloride and iodine:

$$4CuCl_2+I_2+3C_2H_4 \rightarrow 3C_2H_4Cl_2+Cu_2Cl_2+Cu_2I_2$$

The copper salts are reoxidized after dichloroethane stripping by oxygen and an amine hydrochloride acting as a chlorine source:

$$3/2O_2+6(CH_3)_3N\cdot HCl+Cu_2Cl_2+Cu_2I_2 \rightarrow 4CuCl_2+I_2+6(CH_3)_3N+3H_2O$$

The amine hydrochloride is regenerated by sodium chloride and carbon dioxide.

$$(CH_3)_3N+CO_2+NaCl+H_2O \rightarrow (CH_3)_3N\cdot HCl+NaHCO_3$$

The driving force for this reaction is — quite similar to the Solvay soda ash process — the poor solubility of the hydrogen carbonate, which can be removed and calcined, whereby some carbon dioxide is recovered.

$$2NaHCO_3 \rightarrow Na_2CO_3+CO_2+H_2O$$

The overall process yields 1,2-dichloroethane and soda ash from ethylene, oxygen, carbon dioxide, and sodium chloride. The dichloroethane must then be conventionally converted to vinyl chloride by cracking:

$$C_2H_4+1/2O_2+2NaCl+CO_2 \rightarrow C_2H_4Cl_2+Na_2CO_3 C_2H_4Cl_2 \rightarrow C_2H_3Cl+HCl$$

Hydrogen chloride from the cracking reaction can be used to form aminohydrochloride so that balancing with sodium chloride or hydrogen chloride is possible.

$$(CH_3)_3N+HCl \rightarrow (CH_3)_3N\cdot HCl$$

Whether this process will be able to replace the present route will depend primarily on the feasibility of its overall industrial verification. Financial requirements for a grass root plant, however, may be considerably lower than for a conventional plant if the required chlorine capacity is added. The requirement for inexpensive energy instead of expensive electrical power is a further advantage. The byproduct soda ash, even though readily available from natural sources, should not be very limiting if one considers the fact that the conventional process also produces NaOH byproduct during brine electrolysis. However, plans by Akzo to built a commercial scale plant have been abandoned in 1986 [382].

3.1.3.3. Vinyl Chloride from Ethylene by Direct Routes

Because ethylene chlorination and oxychlorination are both highly exothermic reactions, numerous attempts have been made to combine one

or both reactions with the endothermic cracking reaction for 1,2-dichloroethane, i.e., not isolating the intermediate, but directly producing vinyl chloride by high-temperature chlorination or oxychlorination of ethylene. Several processes have been patented which claim direct synthesis of vinyl chloride from ethylene and chlorine or hydrogen chloride at temperatures between 300 and 600 °C.

In *direct chlorination processes* an excess of ethylene is often used to minimize byproduct formation [383]. Other processes use two reaction zones [384] or make use of an inert fluidized bed for heat transfer [384]. The hydrogen chloride formed can be consumed in a separate oxychlorination unit [385]. Additional processes have been proposed [386].

If the *oxychlorination of ethylene* is carried out at a temperature above 350 °C, substantial amounts of vinyl chloride are obtained. As with the Oxy-EDC process (see page 37), polyvalent metals are used as catalyst. However, low-surface area supports (e.g., α-alumina) are preferred [387, 388] because high-surface-area catalysts tend to ward rapid coking and deactivation by polymer formation at higher temperatures. The high temperature required can also cause considerable yield losses by "deep" oxidation of ethylene to CO and CO_2. Oxygen feed below the stoichiometric ratio may be required to control these unwanted side reactions [388].

Feeding the excess ethylene to the high-temperature oxychlorination reactor and converting the surplus to 1,2-dichloroethane in a second DC stage, which is then recycled and cracked in the Oxy reactor may be another possibility [389]. Further possibilities comprise contacting ethylene with melts containing copper(II) chloride. Vinyl chloride is formed and the reduced copper salt can be regenerated by chlorination or oxychlorination and then be recycled [390].

Common to all direct routes is the fact that the processes are difficult to control and operate and are characterized by poor selectivities because ethylene, vinyl chloride, chlorine, and hydrogen chloride undergo considerable addition and elimination reactions at elevated temperatures. Typical byproducts of direct high-temperature processes are dichloroethylenes and tri- and tetrachloroethylene. The low yield of vinyl chloride together with the need to dispose of high quantities of byproducts has considerably limited the industrial implementation of direct processes.

In a process, however, which is carried out on an industrial scale (approx. 150 000 t/a vinyl chloride) by Atochem, France, the byproducts are integrated and several other usable hydrocarbons in addition to vinyl chloride are intentionally produced (Fig. 25) [245, 391].

The process consists of high temperature chlorination of ethylene. The reaction products are separated, yielding vinyl chloride, dichloroethylenes, chloroethanes, and hydrogen chloride. Excess ethylene together with hydrogen chloride is oxychlorinated in a fluidized-bed reactor to give primarily 1,2-di- and 1,1,2-trichloroethane, which can be drawn off or recycled to the hot chlorination reactor.

Dichloroethylenes can be treated differently. 1,1-Dichloroethylene can be hydrochlorinated to give 1,1,1-trichloroethane (see page 46). Trichloroethylene is produced via noncatalytic cold chlorination of dichloroethylenes and subsequent cracking of the tetrachloroethanes obtained. Heavy byproducts accounting for ca. 3 % of the total production are incinerated and aqueous HCl is recovered. However, they can also be used for perchloroethylene synthesis. The process is said to allow considerable fluctuations in production ratios, ranging from 2.5 : 1 to 0.8 : 1 (VCM : chlorinated solvents) allowing good responsiveness to market demands. Overall carbon yields of 94.5 % are obtained.

3.1.3.4. Vinyl Chloride from Ethane

Numerous attempts have been made to convert ethane directly to vinyl chloride because this would save the processing costs for ethylene.

Ethane is readily available, particularly on the U.S. golf coast, and used as a feedstock for ethylene crackers. The direct feed of ethane to VCM plants could, thus, considerably decrease the raw material costs and make the plants less dependent on cracker capacity.

Conversion of ethane to vinyl chloride can be performed by various routes:

1) High-temperature chlorination:

$$C_2H_6 + 2Cl_2 \rightarrow C_2H_3Cl + 3HCl$$

2) High-temperature oxychlorination:

$$C_2H_6 + HCl + O_2 \rightarrow C_2H_3Cl + 2H_2O$$

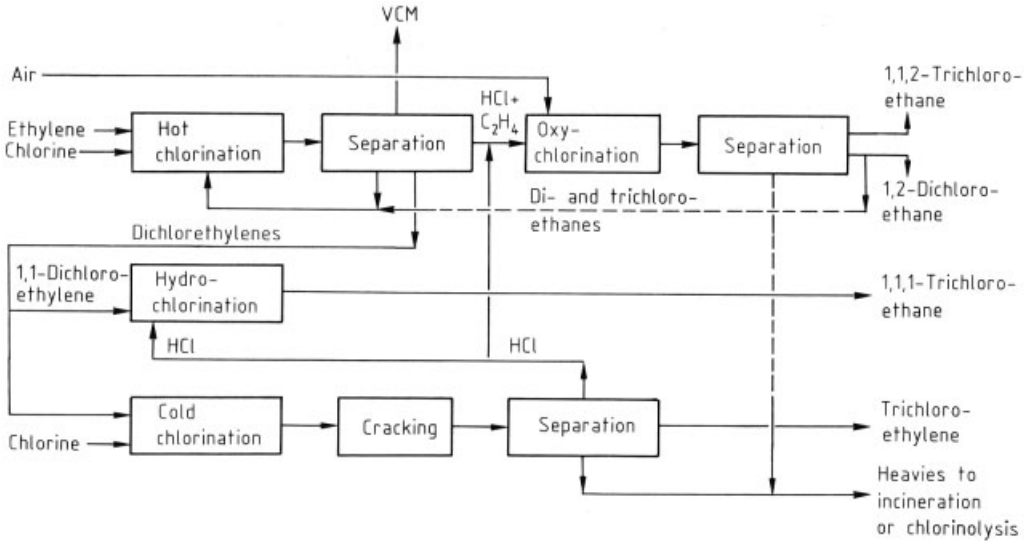

Figure 25. Schematic principle of the Atochem process for the production of vinyl chloride and other chlorinated hydrocyrbons

3) High-temperature oxidative chlorination (combines 1 + 2)

$$2C_2H_6 + 3/2O_2 + Cl_2 \rightarrow C_2H_3Cl + 3H_2O$$

A major drawback of ethane, however, is its lack of molecular functionality. In contrast to ethylene, which easily undergoes chlorine addition, ethane must first be functionalized by substitution reactions, which gives rise to a variety of consecutive and side-chain reactions (Fig. 26).

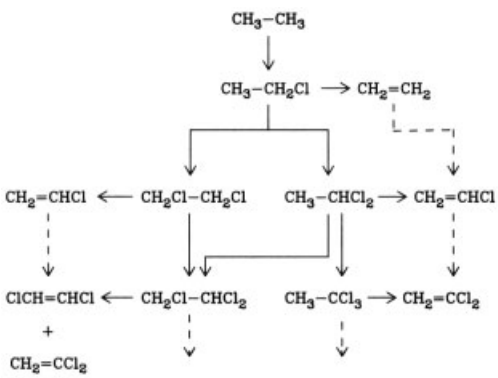

Figure 26. Ethane chlorination pathways

The reaction must, therefore, be kinetically controlled in order to obtain a maximal vinyl chloride yield. Conversion must be sacrificed because thermodynamic conditions would lead to stable products like tetrachloroethylene. Consequently, however, high recycle rates of unconverted ethane and byproducts such as ethyl chloride and dichloroethanes must be accepted. With special catalysts and at optimized conditions, however, ethane conversions of $> 96\,\%$ have been reported from oxychlorination reactions [392]. Vinyl chloride yields average 20 – 50 % per pass. Ethylene, ethyl chloride, and 1,2-dichloroethane are obtained as major byproducts. The formation of carbon oxides can be controlled with carbon yield losses of 3 – 10 %. The ethylene formed can either be recycled or oxychlorinated and cracked in a conventional manner. Some special processes have been suggested to purify and concentrate the aqueous hydrochloric acid obtained [393].

Another patent comprises the chlorination of ethylene – ethane mixtures with staged addition of chlorine to avoid an explosive reaction. Addition of ethylene is thought to suppress the formation of higher chlorinated byproducts [394]. Conversion and yields are comparable to the oxychlorination reactions mentioned above.

Some balanced ethane-based processes have been developed according to the following reaction schemes:

1) Hot ethane chlorination → VCM separation → oxychlorination of residual ethylene and chloroethane to yield additional VCM [395].

2) Thermal chlorination of ethane to ethyl chloride → oxychlorination without separation from hydrogen chloride to vinyl chloride [396].
3) Ethane chlorodehydrogenation to ethylene and hydrogen chloride → oxychlorination to 1,2-dichloroethane → thermal cracking to vinyl chloride [397].

Another balanced ethane-based process was developed by The Lummus Co. [163, 398]. In the final version of this so-called Transcat process, ethane and chlorine, as well as the recycle products ethylene, ethyl chloride, and hydrogen chloride are fed to a melt of copper(II) chloride and potassium chloride. Vinyl chloride is formed and separated. The reduced melt is transferred by an airlift system and regenerated with air, chlorine, or hydrogen chloride. It is then fed back together with the recycle products, which may also contain 1,2- and 1,1-dichloroethane, to the oxychlorinator. Even though the process has been operated on a pilot-plant scale, it has not been accepted by the vinyl chloride producers.

More ethane-based processes for chlorination [399] and oxychlorination [400] can be found in the literature. Because 1,1-dichloroethane is preferably formed by ethane chlorination or oxychlorination, its thermal cracking reaction has been intensively studied [401]. The photochemical chlorination of ethane at 250–400 °C yields ethyl chloride and the dichloroethanes (preferably the 1,1-isomer) as major products [402]. Only small amounts of vinyl chloride are formed.

3.1.3.5. Vinyl Chloride by Other Routes

Vinyl chloride can be obtained as a valuable byproduct in the synthesis of such important fluorocarbons as tetrafluoroethylene (F-1114) and chlorotrifluoroethylene (F-1113) when saturated chlorofluorocarbons are catalytically dechlorinated by ethylene [403].

Oxidative condensation of chloromethane derived from methane or methanol can also form vinyl chloride [404].

The catalytic dehydrochlorination of 1,1,2-trichloroethane [405] or its catalytic dechlorination with ethylene [406] both yield vinyl chloride. Although both processes are not suitable for large-scale production, they could be used to recover vinyl chloride from a major byproduct when there is no demand for 1,1,2-trichloroethane. The electrochemical dechlorination of 1,1,2-trichloroethane is also possible [407].

Some ethylene-based processes comprise the production of vinyl chloride during brine electrolysis in the presence of ethylene [408], ethylene oxychlorination by nitrosyl chloride (NOCl) [409], and a bromine-based process which converts vinyl bromide into vinyl chloride by reaction with hydrogen chloride [410]. Ethane sulfochlorination has been proposed as a very exotic route similar to oxychlorination, but using sulfur instead of oxygen [411].

3.1.4. Uses and Economic Aspects

About 95 % of the world production of vinyl chloride is used for the production of poly(vinyl chloride) (PVC). Thus, vinyl chloride is very dependent on the major processors of PVC as well as the housing and automotive industry with its frequent fluctuations. The rest of the vinyl chloride production goes into the production of chlorinated solvents, primarily 1,1,1-trichloroethane.

Table 21. World wide vinyl chloride capacity and production (1985, estimated) [412]

	Capacity, 10^3 t/a	Production, 10^3 t/a
North America	4 900	4 000
South America	500	400
Western Europe	5 700	5 000
Eastern Europe/USSR	2 100	1 500
Middle/Far East	3 000	2500
Rest	350	180
	16 550	**13 580**

Total world capacity of vinyl chloride is about 17 million t/a. As shown in Table 21, more than half of the world's total capacity (64 %) is concentrated in Western Europe and the United States. The annual growth rate is estimated between 1 and 5 %, depending on the economic situation. However, capacity utilization presently averages only 70–80 % [413], which may not be sufficient to make production of vinyl chloride profitable. Several new vinyl chloride plants are being planned or are under construction in Eastern Europe and in developing and oil-producing

countries (see Table 22)[414]. This significant increase in capacity outside of the traditional VCM –PVC countries and its consequences may in the longrun cause a geographical shift of VCM production.

Table 22. Vinyl chloride plants

Country	Capacity, 10^3 t/year	Technology source
Argentina	130	B. F. Goodrich
Brazil	150	B. F. Goodrich
China (P.R.)	200	Mitsui Toatsu
China (P.R.)	200	Mitsui Toatsu
Egypt	100	Mitsui Toatsu
Egypt	100	–
India	103	–
Iran	150	Toyo Soda
Nigeria	145	–
Poland	205	PPG Ind.
Portugal	110	Mitsui Toatsu
Saudi Arabia	300	B. F. Goodrich
Turkey	117	Solvay-ICI
USSR	270	Hoechst-B.F. Goodrich

3.2. 1,1-Dichloroethylene (Vinylidene Chloride, VDC)

1,1-Dichloroethylene [*75-35-4*] is important for upgrading 1,1,2-trichloroethane, which is very often an unwanted byproduct. Thus, 1,1-dichloroethylene is an intermediate in the production of 1,1,1-trichloroethane from 1,1,2-trichloroethane (see page 45). It is also used as a monomer for the production of poly(vinylidene chloride) (PVDC) and its copolymers, which are important barrier materials in the food packing industry. Of all important chloroethanes and -ethylenes, vinylidene chloride has presently the smallest sales volume. Because of its unique applications in polymers for food containers, longterm demand will grow, however.

3.2.1. Physical Properties

M_r	96.94
mp	− 122.6 °C
bp at 101.325 kPa	31.6 °C
ϱ at 20 °C	1.214 g/cm^3
n_D^{20}	1.42468
Vapor pressure at	
− 60 °C	0.782 kPa
− 40 °C	3.320 kPa
− 20 °C	10.850 kPa
0 °C	28.920 kPa
20 °C	66.340 kPa
Heat of formation (liquid) ΔH^0_{298}	− 24.5 kJ/mol
Specific heat (liquid, 25 °C)	1.15 kJ kg^{-1} K^{-1}
(gas, 25 °C)	0.69 kJ kg^{-1} K^{-1}
Heat of evaporation at 25 °C	26.5 kJ/mol
Heat of fusion at − 122.6 °C	6.51 kJ/mol
Heat of polymerization at 25 °C	− 75.4 kJ/mol
Critical temperature	494 K
Critical pressure	5200 kPa
Viscosity at 20 °C	3.3×10^{-4} Pa · s
Dielectric constant at 16 °C	4.67
Flash point (open cup)	− 30 °C
Autoignition temperature	+ 460 °C
Explosive limits in air	6 – 16 vol%
Solubility in water at 20 °C	2200 mg/kg
Solubility of water in vinylidene chloride at 20 °C	320 mg/kg

1,1-Dichloroethylene is a colorless clear liquid with a sweetish odor. It is soluble in most organic solvents.

3.2.2. Chemical Properties

Vinylidene chloride belongs to the less stable chloroethylenes because it is very susceptible to both oxidation and polymerization. To avoid these reactions, oxygen scavengers such as amino and sulfur compounds or phenol derivatives must be added as stabilizers [415]. Most stabilizers prevent autoxidative polymerization. They must not be removed, however, before vinylidene chloride is industrially polymerized. When pyrolyzed above 400 – 450 °C, chloroacetylene and hydrogen chloride are obtained.

With copper and other heavy metals or their salts, highly explosive acetylenes are formed. Therefore, copper and its alloys should not be used as a construction material if contact with vinylidene chloride is anticipated. Combustion with an excess of air yields carbon dioxide and hydrogen chloride. Traces of phosgene may be formed under oxygen deficient conditions.

Vinylidene chloride can be easily chlorinated at a slightly elevated temperature to give 1,1,1,2-tetrachloroethane.

The most important reaction, however, is hydrochlorination in the presence of a Lewis-acid catalyst for the production of 1,1,1-trichloroethane (see page 45).

3.2.3. Production

1,1-Dichloroethylene is almost exclusively produced from 1,1,2-trichloroethane. This allows

the recovery of valuable hydrocarbon and chlorine from a byproduct, which is obtained in large quantities during the production of 1,2-dichloroethane and 1,1,1-trichloroethane (see Sections 2.3.3 and 2.4.3).

1,1,2-Trichloroethane is converted to vinylidene chloride by dehydrochlorination, which can be carried out by two routes:

1) Liquid-phase dehydrochlorination in the presence of alkali, e.g., NaOH:

$$CHCl_2-CH_2Cl+NaOH\rightarrow$$
$$CCl_2 = CH_2+NaCl+H_2O$$

2) Pyrolytic gas-phase cracking at elevated temperatures:

$$2\ CHCl_2-CH_2Cl \xrightarrow{\Delta} CCl_2 = CH_2 + ClCH = CHCl + 2\ HCl$$

The latter route has the advantage that valuable chlorine is recovered as hydrogen chloride, which can be used again for oxychlorination processes. By this route, however, vinylidene chloride selectivity is low, since the formation of 1,2-dichloroethylenes is favored [416].

In the liquid phase reaction, vinylidene chloride selectivity is well above 90 %; however, hydrogen chloride is lost as a salt.

At present, the liquid-phase reaction is dominant. The development of new catalysts with increased selectivity and high stability could change this situation in the future.

Liquid-Phase Reaction. The liquid-phase reaction is carried out with aqueous solutions of alkali or alkaline earth hydroxides. As with the low solubility of alkaline earth hydroxides, the free concentration is small, NaOH (10 – 15 wt %) is widely preferred to increase the reaction rate. At optimum conditions [417], vinylidene chloride yields of 94 – 96 % are obtained. The higher alkali concentration, however, bears the risk of formation of chloroacetylenes, which tend toward explosive decomposition. Different methods have been patented and are used to minimize this reaction. These include thorough mixing [418], adjustment of proper feed ratios [419], the addition of amines [420], the addition of calcium and magnesium hydroxides as emulsifiers and buffers [421], and the use of sodium chloride containing NaOH [422] (cell effluent from brine electrolysis can be used directly). Monochloroacetylene in crude vinylidene chloride can also be removed by hydrochlorination in aqueous Cu_2Cl_2– HCl solutions [423].

1,1,2-Trichloroethane can be crude [419], even heavy ends from the Oxy – EDC process (see page 37) can be used [424, 425]. However, washing of the feed with water is beneficial to the finished product quality [426]. To avoid polymerization during purification, the feed streams should be free of oxygen. Stabilizers are added during distillation or even to the reaction mixture [417] to inhibit polymerization.

The reaction proceeds at temperatures between 80 and 100 °C at an acceptable rate. With promoters such as charcoal, alumina, and silica [428] or quaternary ammonium salts [425], high selectivity is obtainable at even lower temperatures ($\leq$ 40 °C).

This process is carried out continuously in packed-bed reactors to allow thorough contact. Life steam is injected to distill the vinylidene chloride. Unconverted 1,1,2-trichloroethane is recycled from the appropriate sections. A plug-flow type reactor with consecutive flash distillation has also been described [418].

Due to the alkaline conditions, nickel and some of its alloys are the best suited construction materials. This minimizes the risk of acetylide formation.

In order to obtain polymer grade vinylidene chloride, the stripped product is washed with alkali and water, dried, and fractionally distilled. Azeotropic distillation with methanol (6 wt %) and subsequent washing with water is also possible [429]. Even if purified, vinylidene chloride should be used directly and not stored for more than two days.

The use of aprotic polar solvents like dimethylsulfoxide and dimethylformamide instead of aqueous alkali has been patented [430]. The high cost of solvents, however, may not justify the large-scale production.

Gas-Phase Reaction. The noncatalyzed thermal decomposition yields *cis*- and *trans*-1,2-dichloroethylene together with 1,1-dichloroethylene in molar ratios of ca. 0.7 (1,1-/1,2-isomer). Radical chain as well as unimolecular mechanisms have been proposed for the decomposition reaction [416]. Radical chain sequences are very likely because the reaction rate can be increased

by chain initiators such as chlorine and chlorine-releasing compounds [431].

By photochemically induced reactions, the formation of the 1,1-disubstituted isomer is slightly improved [432].

For industrial purposes, two routes are used to overcome the problems related to the formation of unusable 1,2-substituted isomers.

The first route comprises pure, noncatalytic, thermal cracking of 1,1,2-trichloroethane with selectivity of ca. 30 – 40 % for vinylidene chloride. The 1,2-dichloroethylenes are separated and further chlorinated to tetrachloroethane, which can be recycled and cracked in the same reactor to yield trichloroethylene. This process can be performed by feeding 1,1,2-trichloroethane and chlorine together with the recycled 1,2-dichloroethylene [433]. It can be combined with the production of 1,1,1-trichloroethane by direct ethane chlorination [434] or with a direct chlorination processes for ethylene to yield vinyl chloride [245, 435], as is the case in the Atochem process (see Section 3.1.3.4).

The second route makes use of specific catalysts to increase selectivity. Besides simple catalysts like sodium chloride [375], barium chloride [436], and alumina or silica, which can be activated by steam treatment [437, 438], numerous other catalysts have been patented. These catalysts mainly consist of alkali metal salts [439 – 441], alkali metal hydroxides [442], metal fluorides [443], and nitrogen-containing compounds [444] on appropriate supports such as alumina and silica gels.

The use of alkaline catalysts seems to be important because basic centers on the catalysts are mandatory for high vinylidene chloride selectivity [442], whereas acidic centers favor the formation of the 1,2-isomers. Because the activity increases with the increasing atomic mass of the metal atom [442], cesium salts are preferred [439 – 441].

Additional doping with other metals may be beneficial for further selectivity and prolonged catalyst lifetime [440, 441]. The role of the supports pore structive has also been investigated [445].

For increased selectivity, the feed is frequently diluted with inert gases such as nitrogen. Industrially more important, however, is the possibility of using vinyl chloride as a diluent [441]. When methanol is added to trichloroethane, the hydrogen chloride is consumed by the alcohol and vinylidene chloride is obtained with methyl chloride [446].

Although excellent vinylidene chloride selectivities have been reported, catalytic gas-phase dehydrochlorination is still in the developmental stage. A major drawback is the marked tendency of vinylidene chloride to polymerize on catalyst surfaces [447] which requires frequent shut downs and catalyst turnarounds. This offsets its advantages over the noncatalytic gas phase or liquid-phase reaction. The latter two methods are both used in industrial scale processes.

Other Methods. If required, vinylidene chloride can be obtained from thermal cracking of 1,1,1-trichloroethane [448], which is, however, not always economical. Other routes use vinyl chloride oxychlorination [449] or tetrachloroethane dehydrochlorination [450] and high temperature reaction of methane with chlorinating agents [451]. All of these methods are presently of little interest, because the basic feedstock for the conventional route, 1,1,2-trichloroethane, is easily available.

3.2.4. Uses and Economic Aspects

Vinylidene chloride (VDC) is often captively used for the production of 1,1,1-trichloroethane. Apart from this, VDC is a basic material for poly(vinylidene chloride) (PVDC) or its copolymers with vinyl chloride, acrylonitrile, methacrylonitrile, and methacrylate. With these materials, barrier layers for food packaging are formed as well as laminated and polymer sandwich type films.

The annual production rate for the Western World amounts to about 150 000 – 200 000 t, of which ca. 120 000 t are used for PVDC and its copolymers. The rest is converted to 1,1,1-trichloroethane.

Because of the unique properties of PVDC, the long-term demand will probably increase. It could easily be satisfied because the 1,1,2-trichloroethane feedstock is available from other chlorinated hydrocarbon processes.

3.3. 1,2-Dichloroethylene

Dichloroethylenes (*cis*: [*156-59-2*]; *trans*: [*156-60-5*]) often occur as an isomeric mixture during the production of chlorinated hydrocarbons, where they are produced by sidereactions, e.g., by thermal decomposition of 1,1,2-trichloroethane or from acetylene by chlorine addition.

Because there are scarcely any industrial uses for these two compounds, they are often converted to trichloro- and tetrachloroethylene.

3.3.1. Physical Properties

trans-1,2-Dichloroethylene

M_r	96.94
mp	− 49.44 °C
bp at 101.325 kPa	48.5 °C
ϱ at 20 °C	1.260 g/cm^3
n_D^{20}	1.4462
Vapor pressure at	
− 20 °C	5.300 kPa
− 10 °C	8.500 kPa
0 °C	15.100 kPa
10 °C	24.700 kPa
20 °C	35.300 kPa
30 °C	54.700 kPa
40 °C	76.700 kPa
Heat of formation (liquid) ΔH_{298}^0	− 24.3 kJ/mol
Specific heat (liquid, 20 °C)	1.158 kJ kg^{-1} K^{-1}
Heat of evaporation (boiling point)	28.9 kJ/mol
Critical temperature	516.5 K
Critical pressure	5510 kPa
Viscosity at 20 °C	0.404×10^{-3} Pa · s
Surface tension at 20 °C	25×10^{-3} N/m
Coefficient of cubical expansion (15 – 45 °C)	0.00136 K^{-1}
Dielectric constant at 20 °C	2.15
Dipole moment	0 esu
Flash point	4 °C
Autoignition temperature	460 °C
Solubility in water at 25 °C	0.63 wt %
Solubility of water in *trans*-1,2-dichloroethylene at 25 °C	0.55 wt %

trans-1,2-Dichloroethylene is a colorless, light liquid with a sweetish odor. It forms explosive mixtures with air (9.7 – 12.8 vol% 1,2-dichloroethylene).

trans-1,2-Dichloroethylene forms azeotropic mixtures with ethanol (6 wt % ethanol, *bp* 46.5 °C) and water (1.9 wt % water, *bp* 45.3 °C). A ternary azeotrope of all three components (1.4 wt % ethanol, 1.1 wt % water) has a *bp* of 44.5 °C.

cis-1,2-Dichloroethylene

M_r	96.94
mp	− 81.47 °C
bp at 101.325 kPa	60.2 °C
ϱ at 20 °C	1.282 g/cm^3
n_D^{20}	1.4490
Vapor pressure at	
− 20 °C	2.700 kPa
− 10 °C	5.100 kPa
0 °C	8.700 kPa
10 °C	14.700 kPa
20 °C	24.000 kPa
30 °C	33.300 kPa
40 °C	46.700 kPa
Heat of formation (liquid) ΔH_{298}^0	− 26.8 kJ/mol
Specific heat (liquid, 20 °C)	1.176 kJ kg^{-1} K^{-1}
Heat of evaporation (boiling point)	30.2 kJ/mol
Critical temperature	544.2 K
Critical pressure	6030 kPa
Viscosity at 20 °C	0.467×10^{-3} Pa · s
Surface tension at 20 °C	28×10^{-3} N/m
Coefficient of cubical expansion (15 – 45 °C)	0.00127 K^{-1}
Dielectric constant at 20 °C	9.31
Dipole moment	0.185 esu
Flash point	6 °C
For autoignition temperature and explosive limits in air, see *trans*-1,2-dichloroethylene	
Solubility in water at 25 °C	0.35 wt %
Solubility of water in *cis*-1,2-dichloroethylene at 25 °C	0.55 wt %

cis-1,2-Dichloroethylene is a colorless, light liquid with a sweetish odor. It forms an azeotropic mixture with ethanol (9.8 wt % ethanol, *bp* 57.7 °C), methanol (13 wt % methanol, *bp* 51.5 °C) and water (3.35 wt % water, *bp* 55.3 °C). A ternary azeotrope with ethanol/water (6.55/ 2.85 wt %) has a *bp* of 53.8 °C.

The industrial product always contains both isomers and has a boiling range of 45 – 60 °C. If required, both isomers can be separated by fractional distillation.

3.3.2. Chemical Properties

Of the two isomers, the trans isomer is more reactive than the cis isomer. At higher temperatures and in the presence of bromine or alumina, isomerization is possible. Thermodynamically, the cis isomer is more stable.

If oxygen and moisture are excluded, 1,2-dichloroethylenes are sufficiently stable. With oxygen or peroxides, dimerization to tetrachlorobutene occurs. Upon oxidation, an intermediate epoxide is formed, which then undergoes rearrangement to give chloroacetyl chloride [452]. Combustion with air yields carbon

oxides and hydrogen chloride. Under oxygen deficient conditions, phosgene may be formed.

In the presence of water, hydrolysis occurs to yield hydrochloric acid. Corrosion of construction material can be avoided by such stabilizers as amines and epoxides.

With weak alkali, 1,2-dichloroethylene is not attacked; concentrated alkali, however, induces dehydrochlorination to explosive monochloroacetylene. With copper or its compounds, explosive acetylides can be formed.

In the presence of Lewis-acid catalysts, 1,2-dichloroethylene can be chlorinated to 1,1,2,2-tetrachloroethane or hydrochlorinated to 1,1,2-trichloroethane.

Polymerization is difficult because very high pressures are required. It is not carried out industrially.

3.3.3. Production

Because 1,2-dichloroethylenes are industrially unimportant, they are not deliberately produced in large quantities. They occur as byproducts in some processes, such as the production of vinyl chloride and trichloroethylene, and can be withdrawn and purified if required.

Synthetic routes are possible via

1) thermal cracking of 1,1,2-trichloroethane
2) chlorination of acetylene.

In the thermal dehydrochlorination of 1,1,2-trichloroethane, the 1,2-dichloroethylenes are obtained together with the 1,1-isomer. With increasing temperature, formation of the 1,2-isomers increases. The trans isomer is preferentially formed. With catalysts, the individual ratios (1,2/1,1 and trans/cis) can be varied to some extent.

The chlorination of acetylene with activated carbon catalyst yields almost exclusively the cis isomer. An excess of acetylene is required to suppress the formation of tetrachloroethane. Instead of carbon, mercury and iron salts can be used [453].

Other routes use liquid-phase acetylene oxychlorination [454] or synthesis from 1,1,2,2-tetrachloroethylene which can be dehydrochlorinated and dehydrochlorinated by steam and iron in one reaction.

3.3.4. Uses and Economic Aspects

The 1,2-dichloroethylenes are commercially unimportant, because they do not polymerize, have relatively low boiling points, and can form explosive mixtures with air.

In applications where dichloroethylenes could be used as solvents and for low temperature extraction processes, they have been replaced by methylene chloride, which has a higher solvency, is readily available, and is based on less expensive feedstocks.

1,2-Dichloroethylenes obtained as byproducts from manufacturing processes for other chlorinated hydrocarbons are often used as feed stock for the synthesis of tri- or perchloroethylene.

3.4. Trichloroethylene

E. Fischer first obtained trichloroethylene [*79-01-6*] in 1864 from hexachloroethane by reductive dehalogenation with hydrogen. An acetylene-based process was developed in Austria, and the first plant became operational in Jajce/Yugoslavia in 1908, a plant still producing tri- and tetrachloroethylene [455]. Because of its high solvency and a growing demand for degreasing solvents, trichloroethylene achieved rapid growth rates in the past. Since the late 1960s, however, the production rates have strongly declined as more stringent environmental regulations became effective. Trichloroethylene is also in strong competition with other solvents such as 1,1,1-trichloroethane.

The acetylene-based process has been partially replaced mainly in the United States by ethylene chlorination and oxychlorination routes. A considerable amount of trichloroethylene is still produced from acetylene, which, however, is not made from carbide, but is obtained from ethylene crackers as a byproduct.

3.4.1. Physical Properties

M_r	131.4
mp	− 87.1 °C
bp at 101.325 kPa	86.7 °C
ϱ at 20 °C	1.465 g/cm^3
n_D^{20}	1.4782

Vapor pressure at	
− 20 °C	0.720 kPa
0 °C	2.680 kPa
20 °C	5.780 kPa
40 °C	7.700 kPa
60 °C	42.500 kPa
80 °C	82.800 kPa
Heat of formation (liquid) ΔH^0_{298}	− 42.0 kJ/mol
Specific heat (liquid, 20 °C)	1.01 kJ kg^{-1} K^{-1}
Heat of evaporation (boiling point)	31.5 kJ/mol
Vapor density (boiling point)	4.45 g/L
Critical temperature	544.2 K
Critical pressure	5020 kPa
Thermal conductivity (liquid)	0.14 W m^{-1} K^{-1}
Surface tension (20 °C)	26.4×10^{-3} N/m
Viscosity (20 °C)	0.58×10^{-3} Pa · s
Coefficient of cubical expansion	
(0–40 °C)	0.001185 K^{-1}
Dielectric constant (20 °C)	3.41
Dipole moment	0.9×10^{-18} esu
Ignition temperature	410 °C
Explosive limits in air at 25 °C	7.9–10.5 vol%
at 100 °C	8.0–52 vol%
Solubility in water at 20 °C	0.107 wt %
Solubility of water in	
trichloroethylene at 20 °C	0.025 wt %

Trichloroethylene is a light, colorless liquid with a sweetish smell. It is miscible with most organic solvents and has a high solvency for natural and synthetic rubbers and various other polymers. Some binary and ternary azeotropes formed by trichloroethylene are shown in Table 23.

Table 23. Azeotropes formed by trichloroethylene

wt %	Component	Azeotropic boiling point (101.3 kPa), °C
18	1,2-dichloroethane	82.9
36	methanol	60.2
27	ethanol	70.9
17	1-propanol	81.8
30	2-propanol	75.5
2.5	1-butanol	86.9
33	*tert*-butanol	75.8
6.6	water	72.9
3.8	acetic acid	87.0

The ternary azeotropes contain 23.8 wt % ethanol and 6.8 wt % water, *bp* 67.4 °C; or 12 wt % propanol and 7 wt % water, *bp* 71.7 °C.

3.4.2. Chemical Properties

Trichloroethylene decomposes slowly to yield hydrogen chloride, carbon oxides, phosgene, and dichloroacetyl chloride. This decomposition is enhanced by elevated temperatures (> 100 °C), air or oxygen, sunlight, and moisture and causes corrosion on construction materials.

Trichloroethylene further reacts with aluminum to form pentachlorobutadiene and higher molecular mass polymers.

Atmospheric photooxidative degradation has also been studied [456]. Hydrolysis is less pronounced. With diluted hydroxides, glycolic acid is formed. Strong hydroxides eliminate hydrogen chloride to give highly explosive dichloroacetylene. Acidic hydrolysis with sulfuric acid gives monochloroacetic acid.

Trichloroethylene can be chlorinated to pentachloroethane or hydrochlorinated to give 1,1,2,2-tetrachloroethane [457].

Although trichloroethylene can be copolymerized with a variety of other monomers, it is used in commercial polymer applications only in the production of poly(vinyl chloride), where it allows the control of molecular mass distribution.

3.4.3. Production

For the production of trichloroethylene either acetylene or ethylene is used as a feedstock.

The *acetylene route* which is still used in Europe — the entire production of trichloroethylene in the Federal Republic of Germany is based on acetylene — comprises acetylene chlorination to 1,1,2,2-tetrachloroethane followed by dehydrochlorination to trichloroethylene:

$$C_2H_2 + 2\,Cl_2 \longrightarrow CHCl_2{-}CHCl_2$$
$$CHCl_2{-}CHCl_2 \xrightarrow{\Delta} CHCl = CCl_2 + HCl$$

In the *ethylene-based processes*, which are widely used in the United States and Japan, ethylene or ethylene based chlorohydrocarbons, preferably 1,2-dichloroethane, are chlorinated or oxychlorinated and dehydrochlorinated in the same reactor. Perchloroethylene is obtained as a byproduct in substantial amounts.

Instead of using pure starting materials, these processes can also be carried out very economically with residues from other chlorinated hydrocarbon processes, e.g., from the production of vinyl chloride.

Trichloroethylene from Tetrachloroethane. Because the chlorination of acetylene yields

1,1,2,2-tetrachloroethane, this isomer is preferably used in the production of trichloroethylene. It dehydrochlorinates also more easily than the 1,1,1,2-substituted isomer.

Dehydrochlorination can be carried out in the liquid and gas-phase.

The *liquid-phase process* uses diluted aqueous calcium hydroxide (10 – 20 %) for cracking [458]. The use of NaOH is not recommended because explosive dichloroacetylene could be formed. The heat of the highly exothermic reaction can be used for overhead distillation of the trichloroethylene as an aqueous azeotrope. The calcium chloride solution is continuously withdrawn from the bottom of the reactor and can be further purified from the remaining organics by steam or vacuum stripping. Although this process can be carried out with high selectivity, it is rarely used because hydrogen chloride is lost by salt formation. In carbide-derived acetylene processes, however, it offers an outlet for the calcium oxide obtained from carbide decomposition.

Gas-phase dehydrochlorination of 1,1,2,2-tetrachloroethane is an endothermic reaction.

$$\mathrm{CHCl_2{-}CHCl_2 \rightarrow CHCl = CCl_2 + HCl}$$
$$\Delta H^0_{298} = +61\,\mathrm{kJ/mol}$$

It can be carried out as a pure thermal reaction at temperatures between 300 – 600 °C in tubular reactors. However, because this reaction forms substantial amounts of heavy byproducts, catalytic dehydrochlorination is industrially preferred.

Since catalyst activated carbon silica or porcelain are used [459], barium chloride has been patented as promoter [460]. The feed material must be thoroughly cleaned from iron chloride traces (catalyst from acetylene chlorination) to avoid poisoning of the catalyst [461]. The reaction can be carried out in either fixed or fluidized bed reactors [462] at temperatures between 250 and 400 °C. The trichloroethylene yield ranges between 90 and 95 %. Catalytic traces of chloride were found to promote the reaction [463].

In addition to the acetylene chlorination –tetrachloroethane dehydrochlorination sequence, a direct synthesis by means of acetylene oxychlorination to trichloroethylene, is also possible [464].

A pure anhydrous liquid-phase process (170 – 200 °C), (0.3 – 0.6 MPa) for the dehydrochlorination of mixtures also containing the 1,1,1,2-tetrachloroethane isomer has been patented [465]. Iron chloride formed in situ from the construction material acts as a catalyst, but activated carbon can also be used [466]. The latter reaction, however, is important for ethylene-derived tetrachloroethanes, which are obtained by chlorination of 1,2-dichloroethane [467] (see page 43).

Trichloroethylene from Ethylene or 1,2-Dichloroethane. The synthesis of trichloroethylene from ethylene or 1,2-dichloroethane is possible by various routes, either by ethylene or 1,2-dichloroethane chlorination and subsequent dehydrochlorination or by oxychlorination of 1,2-dichloroethane.

The *chlorination – dehydrochlorination reaction* is either carried out in sequence [467, 468] or, preferably, performed in one reactor.

Although ethylene can be used as a starting material [468], 1,2-dichloroethane is the preferred feedstock because selectivities and yields can be increased.

The highly exothermic reaction is carried out at temperatures between 200 and 500 °C. Numerous catalysts such as activated carbon, silicates, graphite, and others have been patented [469]. For optimum temperature control, fluidized-bed reactors are used [469]. Even at the optimum chlorine : dichloroethane ratio of 2 : 1, substantial amounts of perchloroethylene are formed. This causes problems in the purification section because tetrachloromethane formed from perchloroethylene is difficult to separate from trichloroethylene. To solve this problem, a tandem process has been suggested [470]. However, the chlorination — dehydrochlorination process has the principal disadvantage of producing large amounts of hydrogen chloride, which may not fit into site balances.

Aside from ethylene and 1,2-dichloroethanes, other chlorinated ethane residues may also be used as feed [471].

The *oxychlorination process* for the production of trichloroethylene was developed by PPG Ind. [472]. It has the advantage of consuming hydrogen chloride formed during chlorination during the Deacon reaction, and only small amounts of aqueous hydrochloric acid are obtained.

In the oxychlorination process, ethylene, 1,2-dichloroethane, or chloroethane mixtures — which can be residues from other processes — are fed together with oxygen and chlorine to a fluidized-bed reactor.

The catalyst used contains potassium chloride and cupric chloride on fuller's or diatomaceous earth or silica. At reaction temperatures of 420 – 460 °C, the feed is converted by a series of substitution, crack, and Deacon reactions to trichloroethylene and tetrachloroethylene:

$$C_2H_4 + Cl_2 \rightarrow C_2H_4Cl_2$$

$$C_2H_4Cl_2 + 2Cl_2 \rightarrow C_2H_2Cl_4 + 2HCl$$

$$C_2H_2Cl_4 + Cl_2 \rightarrow C_2HCl_5 + HCl$$

$$C_2HCl_5 + Cl_2 \rightarrow C_2Cl_6 + HCl$$

$$C_2H_4Cl_2 \rightarrow C_2H_3Cl + HCl$$

$$C_2H_3Cl + Cl_2 \rightarrow C_2H_3Cl_3$$

$$C_2H_3Cl_3 \rightarrow C_2H_2Cl_2 + HCl$$

$$C_2H_2Cl_2 + Cl_2 \rightarrow C_2H_2Cl_4$$

$$C_2H_2Cl_4 \rightarrow C_2HCl_3 + HCl$$

$$C_2HCl_5 \rightarrow C_2Cl_4 + HCl$$

$$C_2Cl_6 \rightarrow C_2Cl_4 + Cl_2$$

$$4HCl + O_2 \rightarrow 2Cl_2 + 2H_2O$$

The chlorine yields average 90 – 98 %, carbon yields range between 85 – 90 %. Carbon losses occur by oxidation to carbon oxides and formation of tarry byproducts which cannot be recycled. Temperature control is very important because at a too low temperature (< 420 °C), the cracking reactions diminish whereas at a too high temperature (> 480 °C), the oxidation to carbon oxides increases.

The products are separated and purified by distillation and azeotropic distillation. Tri- and perchloroethylene are withdrawn, the light fractions and high boiling products are recycled to the reactor, and tarry byproducts can be incinerated.

Variation of the trichloroethylene : tetrachloroethylene ratio within a wide range (1.4 – 0.25) is possible by changing the feed ratios.

Instead of using a fluidized bed, the oxychlorination of C_2 residues in a melt of cupric iron and alkali metal chlorides has been patented [473].

Other Processes. Trichloroethylene is one of the major byproducts of the Atochem process (see Section 3.1.3.3), where it is obtained from dichloroethylene chlorination and subsequent cracking [245]. Other routes not industrially used are ethane chlorination [474], the pyrolysis of tri- and tetrachloromethane mixtures [475], and the hydrodehalogenation of tetrachloroethylene [476].

3.4.4. Uses and Economic Aspects

The major use for trichloroethylene is as a solvent for vapor degreasing in the metal industry. Because it can undergo hydrolysis, decomposition, and reaction with metals, it is stabilized with acid acceptors such as amines, alcohols [477], epoxides, and metal stabilizers.

Trichloroethylene is further used for degreasing in the textile industry, as an extraction solvent, in solvent formulations for rubbers, elastomers [478], paintstrippers, and industrial paints. In the production of poly(vinyl chloride), it serves as a chain-transfer agent to control the molecular mass distribution.

Since it was first produced on an industrial scale, trichloroethylene production rates have steadily increased with a peak in 1970, when 280 000 t was produced in the United States and 130 000 t in the Federal Republic of Germany. Since then, however, the production rate of trichloroethylene has declined not only because of reduced losses by improved degreasing systems, but also because of strong competition and replacement by 1,1,1-trichloroethane.

Production for 1984 is estimated at approx. 110 000 t for the United States, 80 000 t for Japan, and ca. 200 000 t for Western Europe (FRG: 30 000 t).

The annual decline of 5 – 7 % observed in 1983 and 1984 will probably continue because the more stringent environmental regulations in most countries will further reduce emissions from degreasing units and enforce reclaiming [479].

3.5. Tetrachloroethylene

Tetrachloroethylene [*127-18-4*] (perchloroethylene, Perc) was first obtained by M. FARADAY by the thermal decomposition of hexachloroethane.

Industrial acetylene-based production began during the first decade of this century. In the 1950s, perchloroethylene became the most important drycleaning solvent.

Most producers have replaced the old acetylene route by ethylene or 1,2-dichloroethane feedstocks or by the chlorinolysis process, which uses chlorinated hydrocarbon residues as starting material.

3.5.1. Physical Properties

M_r	165.8
mp	− 22.7 °C
bp at 101.325 kPa	121.2 °C
ϱ at 20 °C	1.623 g/cm^3
ϱ at 120 °C	1.448 g/cm^3
n_D^{20}	1.5055
Vapor pressure at	
0 °C	0.590 kPa
20 °C	1.900 kPa
40 °C	5.470 kPa
60 °C	13.870 kPa
80 °C	30.130 kPa
100 °C	58.500 kPa
120 °C	100.000 kPa
Heat of formation (liquid) ΔH_{298}^0	− 51.1 kJ/mol
Specific heat (20 °C)	0.86 kJ kg^{-1} K^{-1}
Heat of evaporation (boiling point)	34.7 kJ/mol
Vapor density (boiling point)	5.8 kg/m^3
Critical temperature	620.3 K
Critical pressure	9740 kPa
Thermal conductivity (liquid)	0.13 W K^{-1} m^{-1}
Surface tension at 20 °C	32.1×10^{-3} N/m
Viscosity at 20 °C	0.88×10^{-3} Pa · s
at 80 °C	0.54×10^{-3} Pa · s
Coefficient of cubical expansion (0–40 °C)	0.00102 K^{-1}
Dielectric constant at 20 °C	2.20
Solubility in water at 25 °C	150 mg/kg
Solubility of water in tetrachloroethylene at 25 °C	80 mg/kg

Tetrachloroethylene is a colorless heavy liquid with a mild odor. It is soluble with most organic solvents and exhibits high solvency for organic compounds. Tetrachloroethylene is neither flammable nor does it form explosive mixtures with air.

Some azeotropes formed by tetrachloroethylene are shown in Table 24.

Table 24. Azeotropes formed by tetrachloroethylene

wt %	Component	Azeotropic boiling point (101.3 kPa), °C
15.9	water	87.1
63.5	methanol	63.8
63.0	ethanol	76.8
48.0	1-propanol	94.1
70.0	2-propanol	81.7
29.0	1-butanol	109.0
40.0	2-butanol	103.1
50.0	formic acid	88.2
38.5	acetic acid	107.4
8.5	propionic acid	119.2
3.0	isobutyric acid	120.5
2.6	acetamide	120.5
19.5	pyrrole	113.4
43.0	1,1,2-trichloroethane	112.0
51.5	1-chloro-2,3-epoxypropane	110.1
6.0	glycol	119.1

3.5.2. Chemical Properties

Perchloroethylene is the most stable derivative of all chlorinated ethanes and ethylenes. It is stable against hydrolysis and corrosion on construction materials is less pronounced than with other chlorinated solvents.

Tetrachloroethylene reacts with oxygen or air and light to give trichloroacetyl chloride and phosgene. This autoxidation can be suppressed by such stabilizers as amines or phenols. Liquid-phase oxidation with oxygen, however, can be used for the deliberate synthesis of trichloroacetyl chloride [480].

Hexachloroethane is obtained on chlorination. The atmospheric degradation of tetrachloroethylene has been thoroughly investigated, since it is often found during air sampling [456, 481].

Chlorine substitution by fluorine has been studied [482]. Due to the deactivating effect of the chlorine atoms, perchloroethylene cannot be polymerized under normal conditions.

3.5.3. Production

The production of tetrachloroethylene is theoretically possible by high temperature chlorination of chlorinated lower molecular mass hydrocarbons.

For industrial purposes, three processes are important:

1) Production from acetylene via trichloroethylene:

$$C_2H_2+2Cl_2 \rightarrow C_2H_2Cl_4$$

$$C_2H_2Cl_4 \rightarrow C_2HCl_3+HCl$$

$$C_2HCl_3+Cl_2 \rightarrow C_2HCl_5$$

$$C_2HCl_5 \rightarrow C_2Cl_4+HCl$$

2) Production from ethylene or 1,2-dichloroethane through oxychlorination:

$$CH_2 = CH_2+CH_2Cl-CH_2Cl+2.5Cl_2+1.75O_2$$

$$\rightarrow CHCl = CCl_2+CCl_2 = CCl_2+3.5H_2O$$

3) Production from C_1-C_3 hydrocarbons or chlorinated hydrocarbons through high temperature chlorination

The synthesis from acetylene, which is similar to the production of trichloroethylene from acetylene, was for many years the most important production process.

With increasing prices for the acetylene feedstock, however, this route has become unimportant.

The first processes based on the high temperature chlorination of propene – propane mixtures were developed in the 1940s and early 1950s. These so-called chlorinolysis processes (*chlorin*ating pyr*olysis*) have been further developed and are currently the major source of tetrachloroethylene. Instead of propene – propane mixtures, ethane or C_1-C_3 chlorinated hydrocarbon residues are nowadays used as feed. The chlorinolysis process has become an important step in recovering hydrocarbons and valuable chlorine from residues of other processes (e.g., from vinyl chloride and 1,1,1-trichloroethane production).

With the development of oxychlorination techniques, ethylene or 1,2-dichloroethane oxychlorination has become the second most important route. This process also allows the use of residues instead of pure feed material.

The basic difference between both processes is that tri- and tetrachloroethylene are obtained primarily from oxychlorination, whereas by the chlorinolysis route, tetrachloromethane is generated as a byproduct. Furthermore, the oxychlorination process is the most balanced on hydrogen chloride. The chlorinolysis process is a net producer of hydrogen chloride, which must be consumed by other processes.

Depending on the individual site demands and on the proprietary technology of the producers, both processes presently play a key role in modern tetrachloroethylene production.

Tetrachloroethylene from Acetylene. Even though the direct chlorination of acetylene to tetrachloroethylene is possible [483], most industrial processes use trichloroethylene as an intermediate.

Chlorination of trichloroethylene in the liquid phase (70 – 110 °C) and in the presence of a Lewis-acid catalyst (0.1 – 1 wt % $FeCl_3$) gives pentachloroethane (pentachloroethane can also be obtained from ethylene induced liquid phase chlorination of 1,2-dichloroethane (see Section 2.8.3). Perchloroethylene is then produced from pentachloroethane by either liquid-phase (80 – 120 °C, $Ca(OH)_2$) or catalytic thermal cracking (170 – 330 °C, activated carbon). Overall yields (based on acetylene) of 90 – 94 % are possible.

Because of the long production sequence of four reaction steps and the higher costs for the starting material acetylene, this process has lost its importance during the past 20 years.

Tetrachloroethylene by Oxychlorination of Ethylene, 1,2-Dichloroethane, or Chlorinated C_2 Hydrocarbon Residues. The production of tetrachloroethylene by this route has been described earlier (see page 73). This process produces mainly tri- and tetrachloroethylene. Heavy byproducts such as hexachloroethane, hexachlorobutadiene, and chlorinated benzenes must be withdrawn and disposed of or incinerated. The light products can be recycled, which is important for tetrachloromethane, a major byproduct [484]. For further literature, see [485].

Tetrachloroethylene by Chlorination of Hydrocarbons and Chlorinated Hydrocarbons. Theoretically, three process modifications must be distinguished for this route:

1) High temperature chlorination of ethylene, 1,2-dichloroethane, or chlorinated C_2 hydrocarbons
2) Low pressure chlorinolysis
3) High pressure chlorinolysis

The *high temperature chlorination* based on ethylene or chlorinated C_2 hydrocarbons has been mainly developed by the Diamond Alkali Co. [486] and the Donau Chemie AG [487]. The feed is reacted with chlorine at an elevated temperature (200 – 550 °C) in either fluidized (Diamond Alkali) or fixed (Donau Chemie) catalyst beds. Silica and alumina for fluidized beds and activated carbon for fixed beds have been patented as catalysts.

After quenching, hydrogen chloride and tetrachloroethylene are withdrawn and purified by distillation. The light ends can be recycled to the reactor; heavies like hexachloroethane and hexachlorobenzene must be withdrawn. Major recyclable byproducts are dichloroethylenes, tetrachloroethanes, and trichloroethylene. Trichloroethylene can be converted to tetrachloroethylene by separate chlorination and recycled to the reactor, where the pentachloroethane formed is cracked [487, 488]. Because the pentachloroethane cracking is an endothermic reaction, the reactor temperature can be controlled by the addition of externally formed pentachloroethane. The carbon yield for tetrachloroethylene from the high temperature chlorination is about 90 – 92 %. Yield losses result from the formation of heavies. Chlorine conversions range between 95 and 98 %.

Because separation of tetrachloroethylene from 1,1,1,2-tetrachloroethane is difficult to achieve, ethylene derivatives may be added to the quench tower, which are more easily hydrochlorinated as trichloroethylene, the tetrachloroethane precursor [489].

In the *chlorinolysis process*, hydrocarbons or chlorinated hydrocarbons are chlorinated and pyrolyzed to give mainly tetrachloromethane and tetrachloroethylene.

Kinetically, the reaction consists of a whole series of radical crack and substitution reactions which lead to the most stable products.

Thermodynamically, the reaction is governed by two basic equilibria:

$$2CCl_4 \rightleftharpoons C_2Cl_4 + 2Cl_2$$

$$C_2Cl_6 \rightleftharpoons C_2Cl_4 + Cl_2$$

The thermodynamic equilibrium constants of this reaction are plotted as a function of the reciprocal temperature in Figure 27 [490]. The formation of tetrachloroethylene is favored by an increased temperature and reduced chlorine surplus and pressure. However, because industrial processes are very rarely thermodynamically controlled, the product mix can be widely varied in a range of ca. 5 : 1 (tetrachloroethylene : tetrachloromethane), depending on the feed products and ratios and on such physical conditions as temperature and pressure. Besides tetrachloroethylene and -methane, hexachloroethane, -butadiene, and -benzene are obtained because of their high stability. The latter three products may account for up to 10 % of the carbon yield. Hexachloroethane is almost exclusively and hexachlorobutadiene frequently recycled, whereas the hexachlorobenzene recycle is technically more difficult and not so often practiced. It is withdrawn with some hexachlorobutadiene and disposed or incinerated for the generation of hydrogen chloride or chlorine.

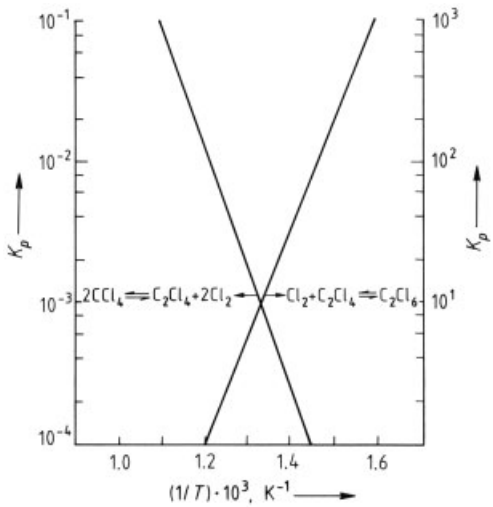

Figure 27. Thermodynamic equilibrium constants for the systems $CCl_4 - C_2Cl_4 + Cl_2$ and $C_2Cl_6 - C_2Cl_4 + Cl_2$ as function of temperature [490]

Presently, two modifications of the chlorinolysis process are in use: the low pressure chlorinolysis and the high pressure chlorinolysis.

The *low pressure chlorinolysis process* is used by most producers. Feedstock for this process are $C_1 - C_3$, preferably C_2 and C_3 hydrocarbons and chlorinated hydrocarbons.

Historically, this process dates back to the 1940s. It was first used by Dow Chemical [491] and somewhat later by Stauffer [492]. Originally

designed for substitution of acetylene by cheaper feedstocks such as ethane and propane, it was increasingly used for the conversion of unwanted byproducts, mainly from chlorinated hydrocarbon (vinyl chloride, allyl chloride, 1,1,1-trichloroethane) and chlorohydrin (propylene oxide, epichlorohydrin) processes, into more valuable products.

With the shift of the traditional feedstock from hydrocarbons toward such byproducts as 1,2-dichloropropane, tetrachloroethanes, pentachloroethane, dichloroethylenes, and chlorinated propanes and propenes, the chlorinolysis process fulfills an ecologically and economically important function for integrated chlorinated hydrocarbon sites. If the demand for perchloroethylene or carbontetrachloride exceeds the available residual feedstock capacity, the process can be carried out with ethylene, ethane, propene, and propane, of which the latter three products are preferred because of the cost advantage.

The reaction is carried out at a reactor temperatures between 600 and 800 °C and a pressure between 0.2 and 1.0 MPa. The slightly increased pressure makes the anhydrous purification of the formed hydrogen chloride easier. Adiabatic as well as isothermal reactors are used [491 – 493]. Most processes use empty tubular or backmixed tank reactors, but fluidized-bed reactors have also been patented [494]. In this reactor type, reaction temperatures are about 500 °C, which results in higher hexachloroethane formation and increased recycle. Mean residence times range between 1 and 10 s.

If chlorohydrin residues containing oxygenated compounds are used, these feedstreams must be pretreated (water wash) because most chlorinolysis processes are very sensitive to oxygen. Oxygen containing feed can lead to the formation of phosgene, carbon oxides, and water, which may contaminate the products or cause corrosion. A process designed and carried out by Chemische Werke Hüls, however, is capable of also handling oxygenated compounds in the feed [495].

Rapid quenching of the gases is important to avoid excessive formation of hexachloroethane. Modern processes avoid the aqueous quench systems used in old plants [496] because corrosion is difficult to control and complicated drying systems are required to recover excess chlorine. In most plants, therefore, quenching is achieved by a high recirculation rate of condensed reaction gases. The heat of reaction is removed by air coolers and heat exchangers or can be used for product distillation [497]. Excess chlorine is either removed by washing or absorption – desorption with tetrachloromethane [495] or it can be used for ethylene chlorination to 1,2-dichloroethane which is either recycled or consumed for the production of vinyl chloride [498].

Process Description (Fig. 28). Hydrocarbons (ethane, propylene) or chlorinated hydrocarbon residues are preheated and fed together with vaporized chlorine to the reactor (material: Ni-alloys or brick lined carbon steel). After the reaction, the hot reaction gases are quenched and chlorine and hydrogen chloride are distilled overhead. Hydrogen chloride is purified by fractional distillation, and the remaining chlorine is removed by absorption with carbon tetrachloride or other light ends and recycled to the reactor. If carbon tetrachloride is used for absorption, the chlorine can be stripped in a second tower to avoid recycling of the solvent [495]. After degassing, the quench bottoms are submitted to fractional distillation. Light ends and medium boilers are recycled, and perchloroethylene and carbon tetrachloride are withdrawn and can be further purified. The heavy byproducts are further treated to recover hexachloroethane and hexachlorobutadiene, both of which can be recycled to the reactor. Hexachlorobenzene is either disposed or incinerated to generate hydrogen chloride or chlorine.

Maximum yield may be as much as 95 % tetrachloroethylene [499]; in industrial scale processes, however, ca. 90 % is achieved [500].

In a process modification by Progil Electrochimie, a reactor cascade with two reactors is used instead of one reactor [501]. The reaction temperature is kept below 600 °C by external cooling.

Another modification consists of liquid phase chlorination at 160 – 200 °C followed by catalytic gas phase chlorinolysis at 450 – 600 °C. Molybdenum pentachloride was patented as a catalyst [502].

The *high pressure chlorinolysis process* was developed by Hoechst AG [503]. The chlorinolysis reaction is non-catalytic at a pressure up to 20.0 MPa and a temperature of about 600 °C. It

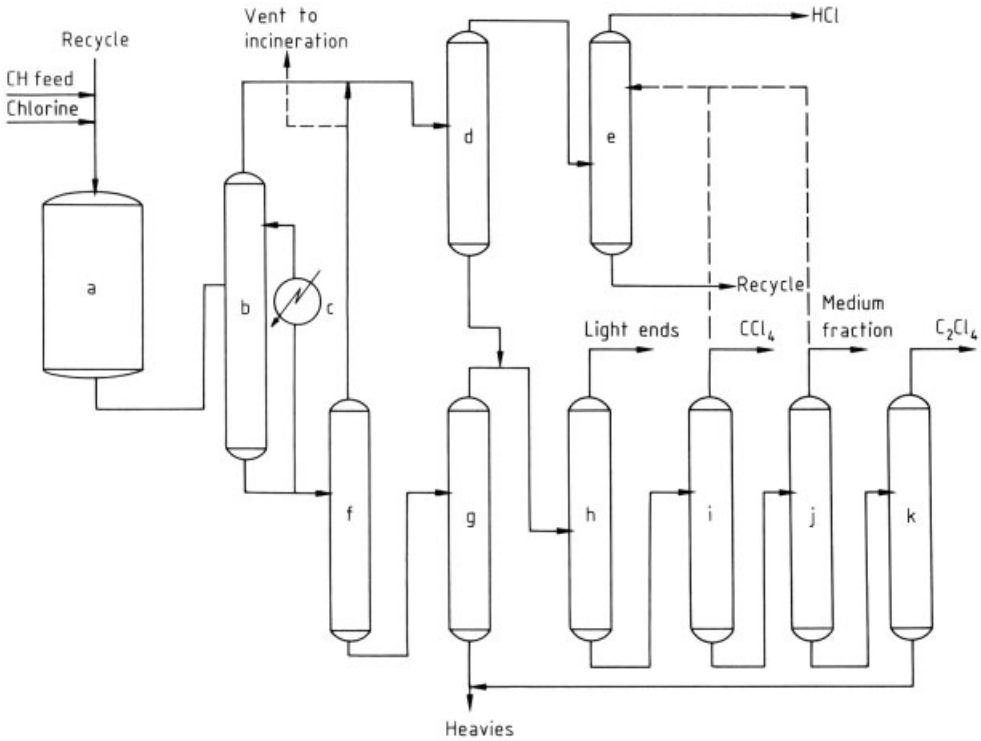

Figure 28. Schematic flow diagram of the chlorinolysis process
a) Reactor; b) Quench tower; c) Cooler; d) and e) HCl tower; f) Degasser; g) Heavies tower; h) Light-end tower; i) Carbon tetrachloride tower; j) Medium-boilers tower; k) Perchloroethylene tower

is claimed that this process can also use higher molecular mass feed, which may contain aromatic and alicyclic compounds. A nickel-plated steel reactor is used.

A 50 000 t/a plant was installed at the Frankfurt site.

However, apart from the tolerance against higher molecular mass feed and the easier separation of hydrogen chloride and chlorine due to the increased pressure, this process offers no major advantages over the low pressure process.

Other Processes. A process to produce tetrachloroethylene from carbon tetrachloride and carbon monoxide has been developed [504], but has not gained any importance. The conventional chlorinolysis process allows broad variations between tetrachloroethylene and carbon tetrachloride production to accommodate varying market demands without being limited by the second product.

3.5.4. Uses and Economic Aspects

The major use for tetrachloroethylene is as a solvent for dry cleaning (ca. 60 % of the total consumption). It has replaced almost all other solvents in this field because it is non-flammable and allows safe operation of drycleaning units without special precautions. Because tetrachloroethylene is very stable, it contains only low concentrations of stabilizers, preferably alkylamines and morpholine derivatives. Because of its high stability, it is also used in addition to trichloroethylene and 1,1,1-trichloroethane for metal degreasing. Particularly for aluminum parts, it is superior to other degreasing formulations.

Other uses are textile finishing and dyeing and extraction processes.

In smaller quantities, tetrachloroethylene is used as an intermediate for the production of trichloroacetic acid and some fluorocarbons.

Because more than half of the Western World's tetrachloroethylene production is based

on the chlorinolysis process, which coproduces carbon tetrachloride in varying ratios, capacity and output are difficult to estimate.

U.S. production capacity was estimated at 380 000 t/a for 1985. In Europe the installed capacity is ca. 450 000 t/a. Total capacity of the Western World may average 1 000 000 t/a.

Even though the consumption of tetrachloroethylene has been declining since the late seventies, the annual rate of decline is very moderate [505] compared to that of trichloroethylene because a replacement of tetrachloroethylene in its main use in dry cleaning is difficult to achieve without sacrificing safety. Thus, the reduced consumption is mainly due to improved dry-cleaning units with reduced solvent losses to the atmosphere. This trend will continue for many years because more stringent environmental regulations have been passed in most countries, which may cause an even steeper decline in the future. The production of tetrachloroethylene in 1993 is estimated to be 123 000 t in the United States and 74 000 t in the Federal Republic of Germany. Assuming a similar capacity utilization for the entire Western World, the 1993 production may be 600 000 – 700 000 t. With the poor capacity utilization and decreased consumption, new plants cannot be justified since the mid 1980s. The construction of a new unit however, was announced in Poland in 1985 [506].

3.6. Analysis and Quality Control of Chloroethanes and Chloroethylenes

Standard methods are used for the analysis of chlorinated ethanes and ethylenes. A typical analysis comprises:

- purity
- water content
- acidity/alkalinity
- free chlorine content
- nonvolatile residues
- physical parameters such as density, refractive index, boiling point, and color

Because of the high volatility of all chlorinated ethanes and ethylenes, gas chromatographic analysis is the method of choice for purity control. Capillary columns are widely used for high resolution. Excellent separation is achieved with non-polar to medium polar stationary phases such as OV-1, OV-101, OV-17, OV-1701, and FFAP. The flame ionization detector (FID) is used for detection in concentrated substances. For low concentrations, the electron capture detector (ECD), together with capillary columns and direct on-column injection, offers excellent sensitivity down to the pg-level.

Unequivocal identification of chlorinated compounds in mixtures is possible by GC/MS analysis. For very diluted aqueous samples, head space, purge-and-trap or closed loop stripping techniques may be used to further enhance the sensitivity.

Analysis of air samples can be achieved by adsorption on activated carbon or other convenient materials, followed by thermal or liquid (CS_2) desorption and consecutive analysis. To eliminate matrix effects, vapor distillation or hexane extraction has been proven to be versatile for aqueous or solid samples. The water content is best determined by the Karl-Fischer method.

For the determination of the acidity and alkalinity, titrations with methanolic or ethanolic sodium hydroxide or hydrogen chloride can be used.

Free chlorine can be detected by iodometric analysis.

Some typical values are given in Table 25.

For such typical solvents as 1,1,1-trichloroethane, trichloroethylene, and tetrachloroethylene, the specifications are standardized in some countries, and standard methods for analysis (ISO, ASTM) have been developed.

3.7. Storage and Transportation of Chloroethanes and Chloroethylenes

Before being stored or transported over longer periods of time, chlorinated ethanes and ethylenes should be carefully analyzed for water, free acid, and stabilizers because decomposition may lead to excessive corrosion.

Chlorinated ethanes and ethylenes should not be brought into contact with tanks, containers, valves, etc. made of aluminum.

Contact with copper should be avoided under all circumstances because dichloroethylenes could form explosive acetylides.

Table 25. Typical degree of purity of some chlorinated ethanes and ethylenes

Compound	Purity	Acidity as HCl	Water content	Free chlorine	Residues
Chloroethane C_2H_5Cl	>99 %	<200 ppm	< 15 ppm	n.d.*	<100 ppm
1,2-Dichloroethane $C_2H_4Cl_2$	>99 %	< 20 ppm	<100 ppm	n.d.*	< 50 ppm
1,1,1-Trichloroethane CH_3CCl_3	>99 %**	< 10 ppm	<100 ppm	n.d.*	< 50 ppm
Trichloroethylene C_2HCl_3	>99 %**	< 10 ppm	<100 ppm	n.d.*	< 50 ppm
Tetrachloroethylene C_2Cl_4	>99 %**	< 10 ppm	<100 ppm	n.d.*	< 50 ppm

* n.d. = not detectable; ** unstabilized

Table 26. Transportation codes and classification for some chlorinated ethanes and ethylenes [507]

Compound	UN-No.	HAZ CHEM	IMDG	RID	Pollution category [508]
Chloroethane	1037	3 W E	class 2 D 2154 E 2057	class 2 201 3bt	–
1,1-Dichloroethane	2362	2 Y E	class 3.2 D 3268 E 3069–1	class 3 301 3 b	B
1,2-Dichloroethane	1184	2 Y E	class 3.2 D 3303 E 3079	class 3 301 16b	B
1,1,1-Trichloroethane	2831	2 Z	class 6.1 D 6387 E 6178–2	class 6.1 601 15 C	B
1,1,2,2-Tetrachloroethane	1702	2 X E	class 6.1 D 6374 E 6173	class 6.1 601 15b	B
Pentachloroethane	1669	2 Z	class 6.1 D 6322 E 6143	class 6.1 601 15b	–
Vinyl chloride	1086	2 W E	class 2 D 2244 E 2125	class 2 201 3 ct	–
1,1-Dichloroethylene	1303	3 Y E	class 3.1 D 3170 E 3050	class 3 301 1a	B
1,2-Dichloroethylene	1150	3 YE	class 2 D 3269 E 3069	class 3 301 1a	D
Trichloroethylene	1710	2 Z	class 6.1 E 6179	class 6.1 601 12b	B
Tetrachloroethylene	1897	2 Z	class 6.1 D 6375 E 6173–1	class 6.1 601 15c	B

Since 1,1,1-trichloroethane decomposes very easily, tanks coated with highly solvent resistant phenolic resin should be used.

Care must be taken with such high solvency compounds as 1,1,1-trichloroethane and 1,1,2,2-tetrachloroethane because they may attack coated or painted equipment.

Some transportation regulations are shown in Table 26.

3.8. Environmental Aspects in the Production of Chloroethanes and Chloroethylenes

The environmental impact of production facilities can be caused by waste (byproducts), emissions to the air, and emissions to the water. In

modern, integrated plants, however, waste problems are minimized by proper balancing and a proper choice of production processes.

Hydrogen chloride generated by crack or substitution reactions is consumed by oxychlorination, hydrochlorination, or HCl electrolysis processes.

Organic byproducts can be used as feedstock for another process and converted into valuable products. Examples are the conversion of 1,1,2-trichloroethane to 1,1-dichloroethylene, the conversion of tetrachloroethanes to trichloroethylene, and the consumption of a wide range of residues in the production of tri- and tetrachloroethylene by oxychlorination or in the chlorinolysis process for the production of carbon tetrachloride and tetrachloroethylene. Recyclable products can be recovered from *tar residues* by especially designed equipment such as rotating double screw heat exchangers [509] or by vapor distillation [510]. Ferric chloride, often used as a catalyst, can be removed from heavies by extraction with hydrochloric acid [511]. The main disposal method for non-recyclable products is high temperature incineration [512] or catalytic incineration [513]. Either hydrochloric acid or hydrogen chloride is obtained, which can be recycled to oxychlorination or used for other purposes. If incineration is carried out with pure oxygen, chlorine is formed by the Deacon equilibrium and can be separated and reused [514]. With the highly developed equipment presently in use, incineration is carried out with very high destruction efficiencies and the remaining emissions do not cause environmental problems.

Other techniques which have been proposed for recovery of chlorinated hydrocarbon residues are destruction by a potassium chloride melt to give hydrogen chloride [515] and use for the production of polysulfides [516] and metal chlorides [517]. If appropriately scrubbed, it is also possible to incinerate mercury-containing catalysts from acetylene-based processes [518].

Air emissions by tank and process vents can be collected and incinerated. When incineration facilities are unavailable, removal of chlorinated hydrocarbon impurities in vent streams is possible by adsorption units, since they have often been described for the removal of vinyl chloride [519].

In the past, emissions by the Oxy – EDC vent posed the largest problems. With the conversion to oxygen, however (see page 37), this vent has become manageable and can be eliminated.

For the Lummus Transcat-Process (see Section 3.1.3.4), a vent treatment system comprising CO_2 absorption followed by ethylene stripping has been proposed [520].

Emissions into water were drastically reduced in the past since most modern processes for the production of chlorinated ethanes and ethylenes avoid the quenching of reaction gases with water to separate hydrogen chloride. By dry distillation and the use of refrigeration units, this potential source for contamination could be excluded by an improved process design.

In processes in which water is formed during the reaction (e.g., Oxy – EDC) or chlorinated hydrocarbons must be brought into contact with water for other reasons (e.g., production of vinylidene chloride from 1,1,2-trichloroethane), the wastewater in modern plants is stripped with air or steam, depending on the nature of contaminants. The gas can be either directly incinerated or condensed and separated. The aqueous condensate is recycled to the stripper, and the organics can be reused or incinerated. Thus, only minor concentrations of chlorinated hydrocarbons are discharged to the biological treatment facilities, from which the emissions are negligible, and excessive, environmentally problematic discharge of chlorinated hydrocarbons to rivers and to the sea is avoided.

4. Chloropropanes

1,2-Dichloropropane is used both as a chemical intermediate and as an industrial solvent. 2-Chloropropane and 1,2,3-trichloropropane are of lesser significance.

4.1. 2-Chloropropane

Physical Properties. 2-Chloropropane [*75-29-6*], isopropyl chloride, $CH_3–CHCl–CH_3$, M_r 78.54, is a colorless liquid.

mp	− 117.2 °C
bp	36.5 °C
Liquid density ϱ at 20 °C	0.8588 g/cm^3

Refractive index n_D^{20}	1.3776
Dynamic viscosity η at 20 °C	3.2×10^{-4} Pa s
Dielectric constant at 25 °C	9.52
Azeotropes with 1.2 wt % water, *bp*	33.6 °C
with 6.0 wt % methanol, *bp*	33.4 °C
with 43 wt % pentane, *bp*	31 °C

Chemical Properties. Prolonged exposure to heat and light can cause dehydrochlorination, especially in the presence of aluminum, zinc, and iron. Reaction with NaOH or amines occurs only at elevated temperature.

Production. 2-Chloropropane can be synthesized by addition of hydrogen chloride to propene using, e.g., alumina, aluminum chloride, calcium chloride, or tellurium compounds as catalysts [521 – 525]. 2-Chloropropane is also formed by reaction of 2-propanol with hydrogen chloride and either zinc chloride or tellurium compounds as catalysts [526, 527].

Uses. 2-Chloropropane, like other chlorinated hydrocarbons, is employed as a solvent.

4.2. 1,2-Dichloropropane

Physical Properties. 1,2-Dichloropropane [*78-87-5*], propylene dichloride, $CH_2Cl-CHCl-CH_3$, M_r 112.99, is a colorless liquid.

mp	− 100.4 °C
bp	96.5 °C
Refractive index n_D^{20}	1.4394
Specific heat c_p at 30 °C	1.38 kJ kg^{-1} K^{-1}
Heat of vaporization	34.8 kJ/mol
Critical temperature T_c	577 K
Critical pressure p_c	4.43 MPa (44.3 bar)
Surface tension at 20 °C	3.0×10^{-2} N/m
Dielectric constant	8.96
Solubility in water at 20 °C	0.25 wt %
Solubility of water in 1,2-dichloropropane at 20 °C	0.16 wt %
Azeotropes with 12 wt % water, *bp*	78 °C
with 53 wt % methanol, *bp*	62.9 °C
with 57.74 wt % ethanol, *bp*	74.7 °C
with 84 wt % cyclohexane, *bp*	80.4 °C
with 84 wt % tetrachloromethane, *bp*	76.6 °C
Flash point (Abel-Pensky, DIN 51 755)	13 °C
Autoignition temperature (DIN 51 794)	557 °C
Explosion limits in air:	
lower limit	3.4 vol%
upper limit	14.5 vol%

Table 27 shows data on this compound.

1,2-Dichloropropane is miscible with most organic solvents, such as alcohols, esters, and ketones, as well as with aromatic, aliphatic, and chlorinated hydrocarbons.

Chemical Properties. 1,2-Dichloropropane is stable at room temperature but is dehydrochlorinated by thermal or catalytic cracking to allyl chloride [*107-05-1*] and 1-chloro-1-propene [*590-21-6*] [528, 529]. It is incompatible with strong oxidizers and strong acids. It is dehydrochlorinated by NaOH to give mainly 1-chloro-1-propene (45 % cis and 55 % trans isomer) [530].

Production. 1,2-Dichloropropane is a byproduct in the synthesis by the chlorohydrin process of the important chemical intermediate propylene oxide (1,2-epoxypropane) [*75-56-9*].

1,2-Dichloropropane is obtained in smaller amounts as a byproduct in the industrial synthesis of allyl chloride. Direct synthesis (e.g., by the addition of chlorine to propene) is not currently carried out [531, 532].

Quality Specifications and Chemical Analysis. 1,2-Dichloropropane is used both crude and as a commercial grade product. The crude product is used mainly as a chemical intermediate for the production of perchlorohydrocarbons (see below). The commercial grade product has a boiling range of 95 – 99 °C at 101.3 kPa (1013 mbar) (DIN 53 171, ASTM D 1078) and a water content below 0.1 wt % (DIN 51 777, ASTM D 1744). Quality analysis is carried out by gas chromatography.

Storage and Transportation. The usual precautions for a highly flammable liquid must be observed with 1,2-dichloropropane. The compound can be stored for several months, but it should be protected from heat, light, moisture and air. Therefore, it is recommended that the product be blanketed with nitrogen. Carbon steel is a suitable material for storage containers provided that the acid and water concentration of the product is low. Rust may increase the color number; however, this is a problem which can be avoided by using stainless steel. Light metals, such as aluminum, magnesium, and their alloys, can react violently with 1,2-dichloropropane. As material for gaskets Teflon, Hostaflon, or IT 400-C (DIN 3754) can be used but materials like PVC, Perbunane, polyethylene, polypropylene, and rubber are not suitable.

Railway tankcars, tanktrucks, and barrels are used for transportation.

Table 27. Vapor pressure (p), liquid density (ϱ), and dynamic viscosity (η) as a function of temperature

	t, °C			
	0	20	50	80
p, kPa	1.8	5.1	19.8	59.9
ϱ, g/cm^3	1.1815	1.1554	1.1161	1.0751
η, Pa s $\times 10^{-4}$	12	8.5	5.8	4.4

The EC directive on dangerous substances must be observed in labeling any containers intended for transporting and handling 1,2-dichloropropane (July 14, 1976, No. 67/548/EC).

Hazard symbols F + Xn; R paragraphs 11 – 20; S paragraphs 9–16–29–33; EC no. 602–020–00–0. UN no. 1279; Rail, RID: 3.1a; Road, ADR: 3.1a; Code no. for railcars and tank cars: 33/1279; Sea, IMDG code: 3.2; Air, IATA article no.: 1507.

Legal Aspects in the Federal Republic of Germany and in the European Community. The "Verordnung über gefährliche Arbeitsstoffe (ArbStoffV) vom 11. 02. 1982, Anhang I Nr. 1," and the "EG-Richtlinie vom 10. 06. 1982" must be observed.

The MAK value (FRG, 1984) is 75 ppm (350 mg/m^3).

According to the "Verordnung über brennbare Flüssigkeiten (VbF)", 1,2-dichloropropane belongs in the "Gefahrenklasse AI". The legal aspects cited in the "Vorschriften des Grundwasserschutzes (VLwF)" must also be considered, as well as the "Unfallverhütungsvorschriften der Berufsgenossenschaften". For safety recommendations, see [533 – 537].

Uses. The most important use of 1,2-dichloropropane is as an intermediate in the synthesis of perchloroethylene [*127-18-4*] and tetrachloromethane [*56-23-5*] [529]. 1,2-Dichloropropane is a good solvent for fats, oils, resins, and lacquers. It is suitable for extraction, cleaning, degreasing, and dewaxing operations in the chemical and technical industry. Since it forms an azeotrope with water at 78 °C, it can be used for removing water from organic solutions.

1,2-Dichloropropane dissolves bitumen and tar asphalt. It is used to promote the adhesion of asphalt layers, and it is suitable for the production of roofing paper, insulation material, and shoe-polish. In combination with 1,3-dichloropropene [*542-75-6*] it can be used as a soil fumigant for nematodes.

1,2-Dichloropropane is applicable as a lead scavenger for antiknock fluids [537], and it is used in petroleum refineries in platforming processes to adjust the catalyst activity [538].

4.3. 1,2,3-Trichloropropane

Physical Properties. 1,2,3-Trichloropropane [*96-18-4*], trichlorohydrin, $CH_2Cl-CHCl-CH_2Cl$, M_r 147.44, is a colorless liquid.

mp	− 14.7 °C
bp	156 °C
Refractive index n_D^{20}	1.4834
Liquid density ϱ at 20 °C	1.388 g/cm^3
Dynamic viscosity η at 20 °C	2.5×10^{-4} Pa s
Surface tension at 20 °C	3.77×10^{-2} N/m
Azeotrope with 31 wt %	
cyclohexanol, *bp*	154.9 °C
with 39 wt % cylcohexanone, *bp*	160.0 °C
with ca. 14 wt % ethylene glycol, *bp*	ca. 152.2 °C
with 65 wt % propionic acid, *bp*	139.5 °C
with 7.5 wt % acetamide, *bp*	154.4 °C
Flash point	74.0 °C
Autoignition temperature	304 °C
Explosion limits in air:	
lower limit	3.2 vol%
upper limit [539]	12.6 vol%

Chemical Properties. 1,2,3-Trichloropropane forms glycerol upon treatment with steam at 550 – 850 °C, and heating it with aqueous caustic gives rise to 2,3-dichloropropene, 2-chloroallyl alcohol, and the ether of the latter [540]. Heating with NaOH may also produce dangerous chloropropynes. The MAK value of 1,2,3-trichloropropane is 50 ppm (300 mg/cm^3) (FRG 1984).

Production. 1,2,3-Trichloropropane is produced by addition of chlorine to allyl chloride [541].

The selectivity of the chlorination of 1,2-dichloropropane to 1,2,3-trichloropropane is not good [542, 543].

Uses. 1,2,3-Trichloropropane is a solvent for oils, fats, waxes, chlorinated rubber, and resins. The reaction with water at high temperature to

Table 28. Physical properties of industrially important chlorinated derivatives of butane

	1-Chlorobutane	1-Chloro-2-methylpropane	2-Chlorobutane (optically active form)	2-Chloro-2-methylpropane	1,4-Dichlorobutane
M_r	92.57	92.57	92.57	92.57	127.02
mp, °C	−123.1	−131.2	−140.5	−25.2	−37.3
bp at 101.3 kPa, °C	78.6	68.4	68.3	50.7	155.0
Density at 20 °C, g/cm^3	0.8865	0.8780	0.8721	0.8435	1.1408
n_D^{20}	1.4025	1.3982	1.3966	1.3856	1.4550
Vapor pressure, kPa	2.0 (− 10.5 °C)	16.8 (0 °C)	6.5 (0 °C)	2.7 (− 32 °C)	1.3 (44 °C)
	4.0 (1 °C)		10.7 (10 °C)	13.3 (− 1 °C)	2.3 (50 °C)
	10.7 (19.7 °C)		17.0 (20 °C)	53.2 (32.6 °C)	5.8 (72 °C)
	27.1 (41.1 °C)		25.9 (30 °C)	87.4 (46.6 °C)	
	66.5 (65.5 °C)		38.0 (40 °C)		
Heat of formation ΔH° of liquid at 298 K, kJ/mol	− 180	− 191	− 193	− 212	
Heat of vaporization ΔH_v° (298 K), kJ/mol	33.1	31.8	31.4	28.9	
Viscosity at 20 °C, 10^{-4} Pa s	4.5	4.6	3.6 (30 °C)	5.1	14.3
Flash point (closed cup), °C	− 12		− 20	− 27	52
Autoignition temperature, °C	460			570	

give glycerol is not employed commercially. 1,2,3-Trichloropropane is used for the synthesis of thiokol polysulfide elastomers if some branching in the polymer structure is required [544].

5. Chlorobutanes

Chlorobutanes can be prepared from butane by liquid-phase or thermal chlorination, from the butenes by addition of elemental chlorine or hydrogen chloride, and from butanol by esterification. The chlorination of butane with chlorine proceeds just as with the lower hydrocarbons, although currently the process has no commercial significance. The reaction has been found to yield primarily the 1- and 2-chlorobutanes along with the 1,3- and 1,4-dichlorobutanes (see Fig. 29) [545 – 547]. For the influence of initiators and inhibitors on the chlorination of *n*-butane, see [548, 549].

Photochemical chlorination in the gas phase at 15 – 20 °C leads to a product distribution similar to that obtained from the strictly thermal process [550].

Table 28 provides an overview of the physical properties of industrially significant chlorine derivatives of butane. 2-Chlorobutane and isobutyl chloride are used primarily as starting materials in Friedel-Crafts reactions. They are made exclusively by hydrochlorination of the corresponding alcohols, a process analogous to that used for the preparation of isopropyl chloride (see page 83).

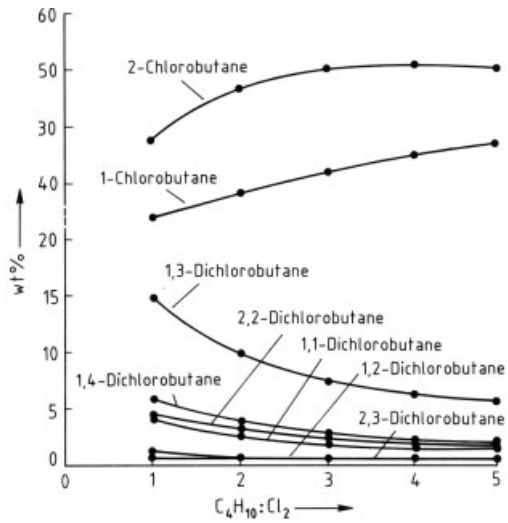

Figure 29. Product distribution in the chlorination of *n*-butane as a function of the *n*-butane : chlorine ratio (conditions: reaction temperature 390 °C, residence time 2 s, pressure 101.3 kPa)

5.1. 1-Chlorobutane

Physical Properties. 1-Chlorobutane, *n*-butyl chloride, $CH_3(CH_2)_3Cl$, is a colorless, flammable liquid with excellent solvent prop-

erties for fats, oils, and waxes. For additional physical properties, see Table 28.

Binary azeotropes are listed in Table 29.

Table 29. Azeotropes of 1-chlorobutane

wt %	Component	Boiling point of the azeotrope, °C (101.3 kPa)
25.0	formic acid	69.4
21.5	ethanol	66.2
8.0	2-butanol	77.7
28.5	methanol	57.2
38.0	methyl ethyl ketone	77.0
16.0	*n*-propanol	75.6
6.6	water	68.1

Chemical Properties. 1-Chlorobutane is stable if kept dry. In the presence of water, it will hydrolyze, but more slowly than secondary or tertiary chlorides. Of industrial significance is its reaction with magnesium to give the corresponding Grignard reagent.

Production. 1-Chlorobutane is obtained by esterification of *n*-butanol with hydrogen chloride or hydrochloric acid at 100 °C either without a catalyst [551] or utilizing the accelerating effect of zinc chloride [552, 553], tripentylamine hydrochloride [554], or phosphorus pentachloride [555]. *n*-Butyl chloride is also obtained, along with 2-chlorobutane, by the chlorination of butane over aluminum oxide at 200 °C [556].

Uses. *n*-Butyl chloride is used as a solvent and alkylating agent in reactions of the Friedel-Crafts type. It is also the starting material for the synthesis of bis (tributyltin) oxide, $(C_4H_9)_3Sn–O–Sn(C_4H_9)_3$ (TBTO), used as an antifouling agent in marine coatings and as a general fungicide.

5.2. *tert*-Butyl Chloride

Physical Properties. 2-Chloro-2-methylpropane, *tert*-butyl chloride, $(CH_3)_3CCl$, is a colorless liquid. It is miscible with common organic solvents and forms constant-boiling mixtures with a number of substances, including methanol, ethanol, and acetone.

Additional physical properties of 2-chloro-2-butane are listed in Table 28.

Chemical Properties. *tert*-Butyl chloride is very rapidly hydrolyzed to *tert*-butyl alcohol when warmed in aqueous alkaline medium. Complete decomposition to isobutylene and hydrogen chloride commences above 160 °C, particularly in the presence of such catalytic oxides as thorium oxide [557].

Production. *tert*-Butyl chloride is generally prepared by reaction of *tert*-butyl alcohol with hydrogen chloride. The hydrogen chloride is passed into the alcohol until the high acid content of the resulting aqueous phase indicates that reaction is complete. The aqueous phase is then removed and the residual product layer is washed to neutrality, dried, and, if necessary, distilled.

tert-Butyl chloride can also be prepared by gas-phase reaction of isobutylene with hydrogen chloride over catalytic chlorides and oxides such as Al_2O_3 at a temperature below 100 °C [558], as well as by passing hydrogen chloride into isobutylene at a low temperature [559, 560].

Uses. *tert*-Butyl chloride is used in Friedel-Crafts reactions (e.g., in the preparation of *tert*-butylbenzene or *tert*-butylphenol) as well as for the synthesis of 4-chloro-2,2-dimethylbutane (neohexyl chloride), used as a fragrance base.

5.3. 1,4-Dichlorobutane

Physical Properties. The compound is a liquid which is flammable when hot and miscible with numerous organic solvents. For additional physical properties of 1,4-dichlorobutane, $Cl(CH_2)_4Cl$, see Table 28.

Chemical Properties. The compound is converted to 1,3-butadiene either thermally at 600 – 700 °C or over alkaline catalysts at 350 °C. Both chlorine atoms are exchangeable. Reaction with sodium sulfide gives tetrahydrothiophene, and diacetyl peroxide converts 1,4-dichlorobutane to 1,4,5,8-tetrachlorooctane.

Production. 1,4-Dichlorobutane is obtained from butane-1,4-diol and hydrogen chloride in the presence of aqueous sulfuric acid at 165 °C [561].

It can be prepared in good yield by treatment of tetrahydrofuran either with hydrogen chloride or with aqueous hydrochloric acid at 100 – 200 °C and 1 – 2 MPa (10 – 20 bar) [562 – 565]; dehydrating agents such as sulfuric acid or zinc chloride favor the reaction [564, 566 – 568].

Other suitable reagents for chlorinating tetrahydrofuran include phosgene or phosphorus oxychloride in the presence of dimethylformamide [569].

Uses. 1,4-Dichlorobutane is used as a synthetic intermediate, as, for example, in the production of nylon.

6. Chlorobutenes

The chlorobutenes that have acquired substantial industrial and economic importance include 2-chloro-1,3-butadiene (chloroprene) and also 1,4-dichloro-2-butene, 3,4-dichloro-1-butene, 3-chloro-2-methyl-1-propene, 2,3-dichloro-1,3-butadiene, and hexachlorobutadiene.

6.1. 1,4-Dichloro-2-butene

$CH_2Cl-CH=CH-CH_2Cl$; Mr 124.96 1,4-Dichloro-2-butene occurs as a low boiling cis form [*1476-11-5*] and as a higher boiling trans form [*110-57-6*].

Physical Properties. *cis form:*

mp	− 48 °C
bp at 101.3 kPa	154.3 °C
d_4^{25}	1.188
n_D^{25}	1.4887
Vapor pressure at 20 °C	0.44 kPa (4.4 mbar)

trans form:

mp	+ 1 °C
bp at 101.3 kPa	156.8 °C
d_4^{25}	1.183
n_D^{25}	1.4861
Vapor pressure at 20 °C	0.31 kPa (3.1 mbar)

Chemical Properties. 1,4-Dichloro-2-butene is stable at room temperature. If prolonged storage is envisaged, it is advisable to arrange to do so under nitrogen.

Dehydrochlorination of the compound with NaOH in the presence of a phase transfer catalyst gives 1-chloro-1,3-butadiene (cis: [*10033-99-5*], trans: [*16503-25-6*]) [570], as does heating it with KOH to 90 °C.

Production. 1,4-Dichloro-2-butene is obtained in a yield of 93 % by chlorination of butadiene in the vapor phase at 300 – 350 °C, short residence times being possible [571, 572]. At lower temperatures about 40 % of the product is 3,4-dichloro-1-butene, and higher boiling products also are formed [573 – 578].

3,4-Dichloro-1-butene [*760-23-6*] can be converted to 1,4-dichloro-2-butene by isomerization in the presence of copper(I) chloride and/or zirconium phosphate [579 – 582].

The two dichlorobutene isomers are also formed in the oxychlorination of C_4 cracking fractions containing butadiene and isobutene [583 – 587].

Uses. 1,4-Dichloro-2-butene occurs as an intermediate in the production of chloroprene (see Section 6.4.3), as does 3,4-dichloro-1-butene.

1,4-Dichloro-2-butene is a starting material in the production of adiponitrile [*111-69-3*] (adipic acid dinitrile), butane-1,4-diol [*110-63-4*], and tetrahydrofuran [*109-99-9*]. Adipic acid and hexamethylenediamine (starting materials in the synthesis of nylon-6,6) are obtained from adiponitrile by saponification and hydrogenation, respectively.

Butane-1,4-diol and tetrahydrofuran can be produced according to a process developed by Toyo Soda [588] which entails hydrolysis of 1,4-dichloro-2-butene in the presence of sodium formate, followed by hydrogenation to butane-1,4-diol. Tetrahydrofuran is obtained by additional elimination of water. Tetrachlorobutane (R,R: [*14499-87-7*], R,S: [*28507-96-2*]) can also be produced from 1,4-dichloro-2-butene [589].

6.2. 3,4-Dichloro-1-butene

$CH_2Cl-CHCl-CH=CH_2$; M_r 124.96; [*760-23-6*]

Physical Properties.

bp at 101.3 kPa	118.6 °C
d_4^{25}	1.153
n_D^{20}	1.4630
Vapor pressure at 20 °C	1.74 kPa (17.4 mbar)

Chemical Properties. 3,4-Dichloro-1-butene is stable at room temperature. At elevated temperatures in the presence of glass or metals, however, it undergoes isomerization to 1,4-dichloro-2-butene. 3,4-Dichloro-1-butene is the starting point in the production of chloroprene (dehydrochlorination in the presence of aqueous sodium hydroxide solution (see Section 6.4.3).

Production. 3,4-Dichloro-1-butene is formed, together with 1,4-dichloro-2-butene (ratio 40 : 60), in the chlorination of butadiene [571 – 578]. It is also obtained by isomerization of 1,4-dichloro-2-butene in the presence of either copper naphthenate [590], a $PdCl_2$– benzonitrile or CuCl –adiponitrile complex [591], or a catalyst complex consisting of copper(I) chloride and an organic quaternary ammonium chloride [592].

Inhibitors such as nitriles [593, 594], thiols [595], acid amides, sulfoxides [596], and phosphines [597] are added to prevent isomerization and polymerization.

Uses. The only industrial use for 3,4-dichlorobutene is in the production of chloroprene.

6.3. 2,3,4-Trichloro-1-butene

$CH_2Cl-CHCl-CHCl=CH_2$; M_r 159.44 [*2431-50-7*].

Physical Properties.

bp at 101.3 kPa	155 °C
d_4^{20}	1.3430
n_D^{20}	1.4944
Vapor pressure at 20 °C	0.25 kPa (2.5 mbar)
Flash point	63 °C
Specific heat at 20 °C	1.088 J/g

Chemical Properties. 2,3,4-Trichloro-1-butene is a colorless, stable liquid at room temperature. Dehydrochlorination with NaOH in the presence of a phase transfer catalyst converts it to 2,3-dichloro-1,3-butadiene [*1653-19-6*] [598].

Production. 2,3,4-Trichloro-1-butene is produced from vinylacetylene or chloroprene by addition of HCl in the presence of a CuCl complex [599 – 603], followed by chlorination of the resulting 1,3-dichloro-2-butene (cis: [*10075-38-4*], trans: [*7415-31-8*]), during which HCl is eliminated [604].

Uses. 2,3,4-Trichloro-1-butene is used exclusively for the production of 2,3-dichloro-1,3-butadiene (see Section 6.5) [598, 605 – 608].

6.4. 2-Chloro-1,3-butadiene

$CH_2=CCl-CH=CH_2$; M_r 88.54,[*126-99-8*].

2-Chloro-1,3-butadiene, generally known as chloroprene, was discovered in 1930 by Carothers and Collins [609, 610] during work on the synthesis of vinylacetylene. Chloroprene is obtained from vinylacetylene by the addition of HCl (see Section 6.4.3.2).

The discovery of chloroprene by Carothers and Collins was based on the chemistry of acetylenes by Father J. Nieuwlands, in 1925.

Chloroprene exists in four isomeric forms: 1-Chloro-1,3-butadiene (α-chloroprene or 1-chloroprene; two isomers: cis [*10033-99-5*] and trans [*16503-25-6*]), 2-chloro-1,3-butadiene [*126-99-8*] (β-chloroprene, also known as 2-chloroprene or simply chloroprene), and 4-chloro-1,2-butadiene [*25790-55-0*] (isochloroprene). Only 2-chloroprene is economically important – as a monomer used to produce polychloroprene, also known as neoprene (the generic term for polymers of this type), e.g., Baypren, (Bayer) Butaclor (Distigugil), Denka chloroprene (Denki-Kagaku).

6.4.1. Physical Properties

Chloroprene is a colorless liquid with a characteristic ethereal odor. It is soluble in most organic solvents. Its solubility in water at 25 °C is 250 ppm.

2-Chloro-1,3-butadiene (β-chloroprene).

mp	− 130 °C
bp at 101.3 kPa	59.4 °C
d_{20}^{20}	0.9583
n_D^{20}	1.4583
Flash point	− 20 °C (ASTM)
Autoignition temperature	320 °C
Limits of inflammability in air	2.1 – 11.5 vol%
MAK value	10 ppm (36 mg/m^3)

Vapor pressure vs. temperature:

°C	0	10	20	30	40	50	59.4
kPa	10.4	16.3	25.0	37.3	53.3	73.9	101.3

Odor threshold	1 mg/L [611]
Specific heat at 20 °C	1.314 J g^{-1} K^{-1}
Latent heat of vaporization at 20 °C	325 J/g
at 60 °C	302 J/g
Reaction enthalpy (3,4-dichloro-1-butene → chloroprene)	− 83.7 kJ/mol
Heat of polymerization	− 711 J/g [612]
Reaction enthalpy of uninhibited chloroprene	− 900 J/g

1-Chloro-1,3-butadiene (α-chloroprene)

mp at 101.3 kPa	66 – 67 °C
d_4^{20}	0.954

4-Chloro-1,2-butadiene (isochloroprene)

bp at 101.3 kPa	88 °C
d_4^{20}	0.9891
n_D^{20}	1.4775

6.4.2. Chemical Properties

Chloroprene is a colorless, highly reactive liquid. It undergoes substantial dimerization, even at room temperature, from which the following compounds have been isolated:

1,2-Dichloro-1,2-divinylcyclobutane (cis: [*33817-64-0*], trans: [*33817-63-9*]), 1,4-dichloro-4-vinylcyclohexene [*65122-21-6*], 1-chloro-4-(1-chlorovinyl)cyclohexene [*13547-06-3*], and 2-chloro-4-(1-chlorovinyl)-cyclohexene [*28933-81-5*].

1,6-Dichloro-1,5-cyclooctadiene [*29480-42-0*] is formed from the *cis*-cyclobutane isomer through rearrangement at elevated temperatures.

In the presence of air, uninhibited chloroprene accumulates peroxides, which may cause it to undergo spontaneous ω-polymerization to give the so-called "popcorn polymer." This process is different from the well-known α-polymerization which leads to rubber-like products. α-Polymerization inhibitors include phenothiazione and *p-tert*-butylcatechol. ω-Polymerization can be inhibited with nitric oxide, *N*-nitrosodiphenylamine, *N*-nitrosophenylhydroxylamine, and other nitroso compounds [613 – 617].

Due to its polymerization tendency, chloroprene is stored and transported under inert gas at temperatures below − 10 °C, with further protection by the addition of sufficient amounts of inhibitor.

6.4.3. Production

6.4.3.1. Chloroprene from Butadiene

The production of chloroprene (2-chloro-1,3-butadiene) from butadiene comprises three steps:

1) Vapor phase chlorination of butadiene to a mixture of 3,4-dichloro-1-butene and 1,4-dichloro-2-butene:

$$\begin{aligned} Cl_2 + CH_2 &= CH - CH \\ &= CH_2 \rightarrow ClCH_2 - CHCl - CH \\ &= CH_2 + ClCH_2 - CH \\ &= CH - CH_2Cl \end{aligned}$$

2) Catalytic isomerization of 1,4-dichloro-2-butene to 3,4-dichloro-1-butene:

$$ClCH_2{-}CH = CH{-}CH_2Cl \overset{cat.}{\rightleftharpoons} ClCH_2{-}CHCl{-}CH = CH_2$$

3) Dehydrochlorination of 3,4-dichloro-1-butene with alkali to chloroprene (2-chloro-1,3-butadiene):

$$\begin{aligned} &ClCH_2{-}CHCl{-}CH = CH_2 + NaOH \\ &\rightarrow CH_2 = CCl{-}CH = CH_2 + NaCl + H_2O \end{aligned}$$

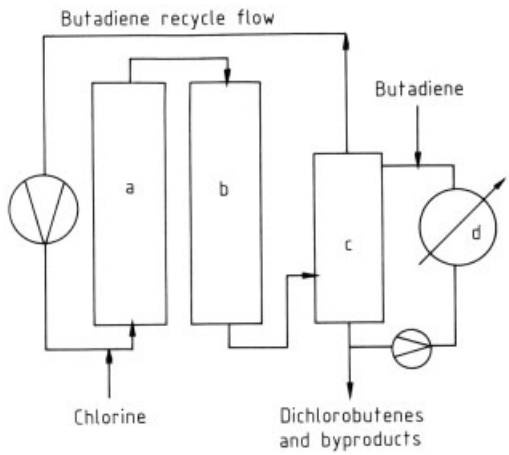

Figure 30. Flow diagram for the chlorination of butadiene a) and b) Chlorination reactors; c) Scrubber cooler; d) Heat exchanger for carrying away heat generated by the reaction

Chlorination of Butadiene. The reaction of chlorine with butadiene can be carried out in an adiabatically operating reactor (see Fig. 30) at temperatures not exceeding 250 °C and at pressures of 0.1 – 0.7 MPa (1 – 7 bar). Intensive and rapid mixing of the reactants is of decisive importance in preventing soot formation. In addition, temperature control is ensured by conducting the reaction in such a way that there is always

an excess of butadiene present (the molar ratio of Cl_2 to butadiene should be 1 : 5 – 1 : 50). The hot gaseous reaction mixture, consisting mainly of dichlorobutenes, butadiene, and highly chlorinated C_4–C_{12} products, is condensed with liquid dichlorobutene in a scrubber-cooler. The excess butadiene is removed at the top of the cooler and returned to the chlorinator together with fresh butadiene. The HCl gas which is formed must be removed in a scrubber. The liquid reaction products are removed at the bottom of the cooler and purified by distillation. The proportion of dichlorobutenes in the reaction mixture is about 92 % [578].

Chlorination processes which utilize high reactor temperatures (280 – 400 °C) and very short residence times but avoid the need for an excess of butadiene are known [571, 572].

Isomerization of the Dichlorobutenes. The distillate from the preceding chlorination step consists of a mixture of 1,4-dichloro-2-butene and 3,4-dichloro-1-butene. This mixture is subsequently isomerized to pure 3,4-dichloro-1-butene by heating to temperatures of 60 – 120 °C in the presence of a catalyst [590 – 592]. A mixture which is enriched in 3,4-dichloro-1-butene can be removed continuously from the reactor. The desired 3,4-dichloro-1-butene, whose boiling point is lower than that of 1,4-dichloro-2-butene, is separated from the latter in a subsequent distillation step. Unreacted 1,4-dichloro-2-butene is then returned to the isomerization reactor.

Dehydrochlorination of 3,4-Dichloro-1-butene to Chloroprene. The dehydrochlorination of 3,4-dichloro-1-butene with dilute NaOH in the presence of inhibitors gives 2-chloro-1,3-butadiene. 1-Chloro-1,3-butadiene is also formed as a byproduct [618].

The reaction is carried out at a temperature of 40 – 80 °C. Addition of phase transfer catalysts, such as quaternary phosphonium or ammonium compounds, accelerates the reaction considerably and gives higher yields per unit time, e.g., 99 % in 30 min [619 – 621].

The chloroprene and sodium chloride solution leaving the dehydrochlorination reactor is separated in a distillation column or by decantation [622]. The salt in the wastewater is either recovered by evaporation, or else the wastewater is recycled to a chlor – alkali electrolysis plant permitting recovery of chlorine and sodium hydroxide solution (thereby decreasing the consumption of these raw materials) [623].

Processes have also been described for the dehydrochlorination of 3,4-dichloro-1-butene using an aqueous mixture of alcohols and NaOH [624 – 627]. These processes have the advantage that the resulting sodium chloride is obtained as a solid, permitting its direct use in a chlor – alkali electrolysis plant.

Chloroprene can be dried by the freezing method: moist chloroprene is cooled to <0 °C and passed through a refrigerator, on the walls of which ice crystals form. These ice crystals are separated from the liquid chloroprene in a recrystallizer, the water content of the resulting chloroprene being < 150 ppm [628].

Purification. The crude chloroprene obtained at the end of the dehydrochlorination step is purified by distillation. Chloroprene should be made prior to polymerization, as free from 1-chloroprene as possible. Accomplishing this requires a column with a large number of plates. Sieve plate or packed columns are generally used.

To prevent polymerization in the column, inhibitors are added and the crude chloroprene is distilled under reduced pressure.

Inhibited chloroprene can be stored at a temperature of − 10 to − 20 °C under nitrogen for fairly long periods without risk.

The purity of chloroprene (target value 99.9 %) is checked by gas chromatography. Impurities can be identified by IR spectroscopy. The Fischer method can be used to determine water.

6.4.3.2. Chloroprene from Acetylene

Production of Monovinylacetylene. Acetylene dimerizes to monovinylacetylene in the presence of aqueous or anhydrous copper(I) chloride and a catalyst solution containing either an alkali metal salt or an ammonium salt.

$$2\,C_2H_2 \xrightarrow[70\,^\circ C]{Cu_2Cl_2/NH_4Cl} CH_2 = CH - C \equiv CH$$

The acetylene is led in under conditions that assure a short contact time, and the products are rapidly removed from the catalyst solution. Byproducts of the reaction are divinylacetylene,

acetaldehyde, and vinyl chloride [629]. Yields of monovinylacetylene range from 75 to 95 % depending on the catalyst system. After concentration and purification, monovinylacetylene is obtained as a colorless liquid boiling at 5.5 °C and 101.3 kPa. Because of its tendency to decompose, monovinylacetylene is diluted for handling, hydrocarbons commonly being used as the diluent.

Production of Chloroprene. The first step in the reaction of hydrogen chloride with monovinylacetylene is a 1,4-addition to give 4-chloro-1,2-butadiene [630].

$$CH_2 = CH - C \equiv CH + HCl \rightarrow ClCH_2 - CH = C = CH_2$$

The chlorine in this compound is very reactive and the substance rearranges in the presence of the Cu_2Cl_2-containing catalyst solution to give 2-chloro-1,3-butadiene:

$$ClCH_2 - CH = C = CH_2 \xrightarrow[60-70\,°C]{Cu^+} CH_2 = CCl - CH = CH_2$$

Under conditions where yields of 95 % are achieved the principal byproduct is 1,3-dichloro-2-butene, resulting from further HCl addition:

$$CH_2 = CCl - CH = CH_2 \xrightarrow{HCl} CH_3 - CCl = CH - CH_2Cl$$

The latter can be used as a starting material for 2,3-dichlorobutadiene, a comonomer of chloroprene.

Chlorides of mercury, magnesium, calcium, gold, and copper are all used in the production of chloroprene, as are ammonium chloride, ammonium bromide, pyridinium chloride, and methylammonium chloride. The most commonly used material, however, is the system Cu_2Cl_2/ NH_4Cl/HCl. Two reactors for chloroprene production have been described [631].

6.4.3.3. Other Processes

Other processes described in the literature for the production of chloroprene include oxychlorination of butenes [583 – 588], chlorination of butadiene in organic solvents [632 – 636] and electrochemical chlorination of butadiene [637, 638].

The direct chlorination of butadiene to chloroprene has been described in patents [639, 640].

To date, however, none of these processes has become commercially important.

6.4.4. Economic Importance

Chloroprene is the starting monomer for the specialized rubber known as polychloroprene.

The vulcanizates of polychloroprene have favorable physical properties and excellent resistance to weathering and ozone. Articles made with this rubber include electrical insulating and sheathing materials, hoses, conveyor belts, flexible bellows, transmission belts, sealing materials, diving suits, and other protective suits.

Adhesive grades of polychloroprene are used mainly in the footwear industry. Polychloroprene latexes have found application for dipped goods (balloons, gloves), latex foam, fiber binders, adhesives, and rug backing.

Chloroprene and polychloroprene are produced:

- in the United States: by Du Pont and Petrotex
- in Western Europe: by Bayer in the Federal Republic of Germany, Distugil S.A. in France, and Du Pont in Northern Ireland
- in Japan: by Showa-Denko, Denki Kagaku, and Toyo Soda
- and in the Soviet Union

World polychloroprene capacity, 1983, in 1000 t [641]:

United States	213
Federal Republic of Germany	60
France	40
United Kingdom	30
Japan	85
Central planned economy countries	220

6.5. Dichlorobutadiene

6.5.1. 2,3-Dichloro-1,3-butadiene

$CH_2 = CCl - CCl = CH_2$; M_r 122.95[*1653-19-6*]

Physical Properties.

bp at 101.3 kPa	98 °C
d_4^{20}	1.1829
n_D^{20}	1.4890

Chemical Properties. 2,3-Dichlorobutadiene is more reactive than chloroprene. Therefore, it is not known to have been produced industrially in the pure form. Inhibition with an inert gas containing 0.1 vol% of NO has been reported [642].

Production. 2,3-Dichlorobutadiene is produced by reacting butyne-1,4-diol with phosgene, followed by isomerization of the intermediate 1,4-dichloro-1-butyne [*62519-07-7*] in the presence of copper(I) chloride and amines [643 – 645]. The dehydrochlorination of both 2,3,4-trichlorobutene (see Section 6.3) and 1,2,3,4-tetrachlorobutane with sodium hydroxide solution [598, 605 – 608, 646 – 648] has acquired industrial importance.

2,3-Dichlorobutadiene is an important monomer for special grades of polychloroprene in which the tendency to crystallize has been reduced.

6.5.2. Other Dichlorobutadienes

1,1-Dichloro-1,3-butadiene [*6061-06-9*] and 1,3-dichloro-1,3-butadiene [*41601-60-9*] are obtained by chlorination of 1-chloro-1,3-butadiene, followed by dehydrochlorination [649, 650].

1,1-Dichloro-1,3-butadiene is also obtained by reduction of 4,4-dichloro-3-buten-2-one with $NaBH_4$, followed by dehydrogenation with Al_2O_3 [651]. This compound is used as a comonomer for rubber and as an insecticide.

1,4-Dichloro-1,3-butadiene (trans, trans: [*3588-12-3*]; trans, cis: [*3588-13-4*]; cis, cis: [*3588-11-2*]) is manufactured by dehydrochlorination of tetrachlorobutane with zinc powder (yield: 60 %) [652].

6.6. 3-Chloro-2-methyl-1-propene

3-Chloro-2-methyl-1-propene, [*563-47-3*] $CH_2=C(CH_3)-CH_2Cl$, methallyl chloride, was first prepared in 1884 by M. Sheshukov by the chlorination of isobutylene at room temperature. Investigation of the compound was taken up again in the late 1930s by the research group of A. P. A. Groll at the Shell Development Co. leading to its production at the pilot plant scale. It was this group which gave methallyl chloride its common name on the basis of the analogy of methacrylic acid. The IG Farben chemical works at Leuna in Germany prepared methallyl chloride on a semi-works scale a short time later (from 1939 to 1942) in order to carry out investigations in the polyamide field. Methallyl chloride production by chlorination of isobutylene resumed in Germany after World War II (about 1960) at Chemische Werke Hüls AG.

6.6.1. Physical Properties

M_r	90.5
mp	<-80 °C
bp at 101.3 kPa	72.0 °C
Density at 20 °C	0.9251 g/cm^3
n_D^{20}	1.4274

Vapor pressure vs. temperature:

°C	− 13	18.8	36	53
kPa	2.66	13.3	26.6	53.2

Heat of formation ΔH^0_{298} (liquid)	− 105.1 kJ/mol
Heat of vaporization (calculated from the vapor pressure curve)	31.8 kJ/mol
Viscosity at 20 °C	4.2×10^{-4} Pa s
Dielectric constant	7.0
Flash point	− 16 °C (Pensky-Martens)
Autoignition temperature	540 °C
Limits of inflammability in air at 20 °C	2.2 and 10.4 vol%

Methallyl chloride is a colorless liquid of low viscosity, irritating to the mucous membranes, and with a pungent odor. It is miscible with all of the common organic solvents.

Solubility in water at 20 °C is ca. 0.05 wt %; water solubility in methallyl chloride is 0.04 wt %

Methallyl chloride forms an azeotrope with water (5.8 wt %), which boils at 64.6 °C.

6.6.2. Chemical Properties

Methallyl chloride shows chemical behavior very similar to that of allyl chloride and undergoes essentially the same reactions as the latter. The presence of a β-methyl group increases its reactivity considerably, however. Under the influence of light and/or heat, the formation of low concentrations of dimeric methallyl chloride (2-methyl-4,4-bis(chloromethyl)-1-pentene) occurs. Moisture causes samples of

methallyl chloride to become markedly acidic as a result of hydrolysis. A similar phenomenon is induced by either iron or iron chloride through dehydrochlorination of dimeric and polymeric material.

The chlorine atom of methallyl chloride is readily exchanged, as occurs, for example, in its conversion to alcohols, ethers, amines, and esters. Furthermore, the presence of the methyl group activates the double bond with respect to reactions such as hydration, addition of halogen or halogen halides, sulfonation, dimerization, and polymerization. Of greatest commercial significance are its reaction with sodium sulfite to give sodium methallyl sulfonate and the production of 2-methylepichlorohydrin. For further chemical reactions of methallyl chloride, see [653 – 657].

6.6.3. Production

Methallyl chloride is produced exclusively by chlorination of pure isobutylene in the gas phase. Yields of 80 – 85 % are achieved at atmospheric pressure and a relatively low temperature (not exceeding 100 °C).

Chlorination Mechanism. The chlorination of isobutylene is somewhat unusual as compared to most olefin chlorinations. Treatment of olefins with elemental chlorine at a low temperature normally leads to addition as the dominant reaction. The competing substitution reaction, in which the double bond remains intact, increases in importance as the temperature is raised until a critical temperature is reached, above which one obtains almost exclusively the substitution product. The temperature range of this transition from addition to substitution lies highest for simple straight-chain olefins. Compounds with branching at the double bond, on the other hand, including isobutylene or 3-methyl-2-butene, react principally by substitution even at a very low temperature (– 40 °C), as shown in Table 30.

Presumably, chlorination of isobutylene follows an ionic mechanism, and solid metallic and liquid surfaces catalyze the reaction [659]. According to [660 – 662], the following mechanism is likely: isobutylene reacts with a chlorine molecule such that the equivalent of a positive chlorine adds to the terminal carbon of the double bond, thereby producing a tertiary carbonium ion. This tertiary carbonium ion can react in one of three ways: either it can eliminate a proton from C_1 or C_3, both of which possibilities lead to restoration of a double bond, or else it can react further by addition of the chloride anion. All three possible products are obtained in the chlorination of isobutylene: methallyl chloride (the favored product), the isomeric 1-chloro-2-methyl-1-propylene (isocrotyl chloride), and 1,2-dichloroisobutane.

Table 30. Critical temperature ranges for selected olefins for the transition from chlorine addition to chlorine substitution, based on [658]

Olefin	Critical temperature range, °C
Isobutylene (and other olefins branched at the double bond)	<-40
2-Pentene	125 – 200
2-Butene	150 – 225
Propylene	200 – 350
Ethylene	250 – 350

Process Description. In carrying out the reaction on the industrial scale, gaseous isobutylene (99 % pure) and chlorine are mixed at room temperature by using a dual-component jet, and the mixture is allowed to react in a cooled tube reactor at atmospheric pressure. The reaction rate is exceedingly high for such a low reaction temperature, all of the chlorine being converted within ca. 0.5 s. Proper mixing of the components by the inlet jet is a prerequisite to obtaining a high yield of methallyl chloride; reaction is so rapid that high local concentrations of chlorine cannot be dissipated, leading instead to the production of more highly chlorinated products.

In order to minimize further chlorination of methallyl chloride, an excess of isobutylene must be used, the amount necessary being dependent on the extent of reverse mixing of the reaction products in the reactor. An excess of isobutylene greater than ca. 25 % leads to no further significant increase in the yield of methallyl chloride with reactors in which minimal axial mixing occurs; greater excesses would simply increase the effort required to recover unreacted isobutylene. Low reaction temperatures are advantageous in respect to the product distribution: 100 °C is generally regarded as the limit below which the reaction should be kept.

The use of liquid isobutylene as starting material would seem a logical alternative, since one could then take advantage of the cooling effect resulting from the compound's heat of vaporization. It has been shown, however, that this approach increases the yield of high boiling products, presumably due to the decrease in effectiveness of mixing of the components relative to the gas phase reaction.

A typical analysis of a methallyl chloride production run is shown in Table 31.

The *tert*-butyl chloride which is reported as being formed is a secondary product from addition of hydrogen chloride to isobutylene. For information on the influence of other variables on the chlorination, see [663].

The work-up of the crude reaction product is, in principle, quite simple. The only difficulty would be removal by distillation of isocrotyl chloride, given its very similar boiling point (see Table 31). The vinylic placement of its chlorine, however, makes this substance very unreactive relative to methallyl chloride, so that its presence is not detrimental in many applications. For this reason, it is normally separated only incompletely, if at all.

The effluent gas stream is led out of the reactor into a water scrubber-cooler which absorbs the hydrogen chloride and causes the majority of the chlorinated hydrocarbons to condense. Remaining isobutylene exits from the top of the scrubber column. It is then collected, dried, compressed, and returned to the reactor. The crude methallyl chloride is separated from hydrochloric acid in separatory flasks, dried by azeotropic distillation, and freed by a preliminary distillation from more volatile byproducts (largely *tert*-butyl chloride and part of the isocrotyl chloride). Methallyl chloride and residual isocrotyl chloride are finally taken off at the head of a second column. The more highly chlorinated distillation residue is either worked up in a chlorinolysis unit or else it is burned under conditions permitting recovery of hydrogen chloride.

The use of glass apparatus is recommended both because of the fact that it is necessary to work to some extent with the crude product while it is still moist, and also because even the anhydrous crude product can evolve hydrogen chloride at elevated temperatures and in the presence of ferrous materials.

In contrast to other hydrocarbon chlorination procedures (those employed with methane or propylene, for example), an aqueous workup cannot in this case be avoided if isobutylene is to be recycled. This is because isobutylene reacts with hydrogen chloride to form *tert*-butyl chloride at the low temperature required for the separation of the chlorinated hydrocarbons. If one is willing to accept the presence of *tert*-butyl chloride as a byproduct, however, one could, perhaps with the aid of catalysts, avoid the complications of isobutylene recycling.

Other Production Methods. A thorough investigation has been made of the use in place of pure isobutylene in the methallyl chloride synthesis of a butane – butene (B – B) mixture remaining after extraction of butadiene from the C_4 fraction of petroleum cracking gases [664]. The C_4 fraction can also be chlorinated directly by using the liquid products as a reaction medium, in which case any straight-chain alkenes that are present react only to an insignificant extent [665]. As far as can be determined, however, these interesting possibilities have not yet seen useful commercial application.

Oxychlorination of isobutene is possible by using tellurium compounds as catalysts [666]. High yields of methallyl chloride are achieved on a small scale. Solution chlorination of isobutene in the presence of $TeCl_4$ or $SeCl_4$ is also known [667].

6.6.4. Quality Specifications and Chemical Analysis

The commercial product usually has a purity of 96 – 98 % and contains isocrotyl chloride as its chief impurity, along with traces of 1,2-dichloroisobutane. It is stabilized with substituted phenols or marketed without a stabilizer. The degree of purity is established by gas chromatography. In many of its applications — such as copolymerizations to produce synthetic fibers — even traces of iron are detrimental. The presence of iron is best determined spectroscopically, using *o*-phenanthroline, for example. Additional information which should accompany any shipment would include the material's color, water content, pH value, residue on evaporation, and boiling range.

Table 31. Product distribution from isobutene chlorination, with physical properties of the substances

Compound	Concentration range, wt %	M_r	*bp* at 101.3 kPa, °C	d_4^{20}	n_D^{20}
tert-Butyl chloride	2 – 4	92.5	50.8	0.8410	1.3860
Isocrotyl chloride	4 – 5	90.5	68.1	0.9186	1.4221
Methallyl chloride	83 – 86	90.5	72.0	0.9250	1.4274
1,2-Dichloroisobutane	6 – 8	127.0	108.0	1.089	1.436
3-Chloro-2-chloromethyl-1-propene		125	138.1	1.1782	
1-Chloro-2-chloromethyl-1-propene	2 – 3	125	132.0 (cis) 130.0 (trans)	1.1659	1.4702 (25 °C)
1,2,3-Trichloroisobutane		161.5	163.9		
2-Methyl-4,4-bis(chloromethyl)-1-pentene		181	84.4 (1.3 kPa)	1.0711 (d_{20}^{20})	1.4773

6.6.5. Storage and Shipment

Methallyl chloride is best stored under cool conditions in porcelain enameled vessels. Provided the material is kept very dry, it can also be stored in tightly-sealed stainless steel storage tanks. Porcelain enameled or baked enameled tankers are suitable for its transport, as are, for small shipments, special types of baked enameled drums.

6.6.6. Uses

The potential uses for methallyl chloride as a starting material in syntheses are, just as in the case of allyl chloride, extremely numerous [653]. Its use in pure form as a fumigant and disinfectant has been suggested, as has its application as a fumigating agent for seed grains [668, 669]. Methallyl chloride and other methallyl derivatives prepared from it are exceptionally well suited to copolymerization [670, 671]. The copolymerization of methallyl sulfonate with acrylonitrile has gained particular industrial significance [672]. Since 1970, 2-methylepichlorohydrin for use in the production of special epoxy resins has been manufactured from methallyl chloride at a production facility in Japan with an annual capacity of 6000 t. Methallyl chloride is used, together with allyl chloride, in the manufacture of Allethrin, a synthetic pyrethrum with applications as a pesticide.

6.7. Hexachlorobutadiene

$CCl_2=CCl-CCl=CCl_2$; M_r 260.8 [*87-68-3*].

Physical Properties.

mp	− 18 °C
bp at 101.3 kPa	212 °C
Density at 20 °C	1.680 g/cm^3
n_D^{20}	1.5663

Vapor pressure vs. temperature:

°C	20	34	60	64	111
kPa	0.036	0.097	0.497	0.623	5.87

Specific heat at 22 °C	0.85 kJ kg^{-1} K^{-1}
Heat of vaporization	48 kJ/mol
Coefficient of thermal conductivity	0.101 W K^{-1} m^{-1}
Viscosity at 15 °C	9.22×10^{-3} Pa s
at 21 °C	3.68×10^{-3} Pa s
at 50 °C	2.40×10^{-3} Pa s
Surface tension at 20 °C	3.14×10^{-2} N/m
Dielectric constant at 12 °C	2.56
Dielectric strength when freshly prepared	ca. 200 kV/cm

Hexachlorobutadiene is a colorless, oily liquid with a faint terpene-like odor. Its solubility in water at 20 °C is 4 mg/kg, that of water in hexachlorobutadiene at 20 °C 10 mg/kg.

Chemical Properties. Chemically, hexachlorobutadiene is very stable to acids and alkali and has no tendency to polymerize even under high pressure (10 MPa). Hexachlorobutadiene reacts with chlorine only under severe reaction conditions (e.g., under pressure in an autoclave at 230 – 250 °C), and then generally with cleavage of the carbon skeleton and formation of hexachloroethane and perchloroethylene [673]. Octachlorobutene and decachlorobutane are also produced in the temperature range 60 – 150 °C [674, 675]. For further information about the properties, reactions, and application possibilities of hexachlorobutadiene, see also the compilations [676].

Production. Hexachlorobutadiene occurs as a byproduct in all chlorinolysis processes for the production of perchloroethylene or carbon tetrachloride. Depending on the method employed, the crude product contains ca. 5 % or even more of hexachlorobutadiene, and this material can be recovered in pure form rather than being recycled into the reactor. Since chlorolysis plants are of large capacity, the demand for hexachlorobutadiene is generally met without the need for separate production facilities.

If hexachlorobutadiene is to be prepared directly, the preferred starting materials are chlorinated derivatives of butane. These are chlorinated at 400 – 500 °C and atmospheric pressure with a 4-fold excess of chlorine, giving a 75 % yield of hexachlorobutadiene [677].

The compound can also be prepared by chlorination of butadiene in a fluidized-bed reactor using a large excess of chlorine at 400 – 500 °C [678].

According to a method of the Consortium für Elektrochemische Industrie, hexachloro-1,3-butadiene can be made from 1,1,2,3,4,4-hexachlorobutane, which comprises about 60 % of the residue from the production of tetrachloroethane. The material is chlorinated repeatedly at 70 – 80 °C in the presence of iron(III) chloride and subsequently dehydrochlorinated at 170 °C.

Uses. As a consequence of its low vapor pressure at room temperature, hexachlorobutadiene is suitable as an absorbent for the removal of impurities in gases, e.g., removing carbon tetrachloride and other volatile compounds from hydrogen chloride [679]. A mixture composed of hexachlorobutadiene (50 – 70 %) and trichloroethylene has found use as a coolant in transformers [680]. The production of electrically conductive polymers with hexachlorobutadiene has been patented [681]. The compound's herbicidal properties have led to its use in the prevention of algal buildup in industrial water reservoirs, cooling towers, and cooling water systems [682]. Finally, it has been repeatedly suggested as a hydraulic fluid, as a synthetic lubricant, and as a nonflammable insulating oil.

7. Chlorinated Paraffins

"Chlorinated paraffins" is the collective name given to industrial products prepared by chlorination of straight-chain paraffins or wax fractions. The carbon chain length of commerical products is usually between $C_{10} - C_{30}$ and the chlorine content between 20–70 wt%. Four main types of chlorinated paraffins are in regular use today (see Table 32).

Table 32. Types of chlorinated paraffins

Chlorinated paraffin type	Common abbreviation	CAS registry number*
$C_{10} - C_{13}$ (short-chain) chlorinated paraffins	SCCP	[*085535-84-8*]
$C_{14} - C_{17}$ (medium-chain) chlorinated paraffins	MCCP	[*085535-85-9*]
$C_{18} - C_{20}$ (long-chain) chlorinated paraffins	LCCP	[*063449-39-8*]
$C_{20} - C_{30}$ chlorinated paraffin waxes (liquid and solid products)		[*063449-39-8*]

* These CAS numbers are appropriate for Europe and denote existing chemicals within the European EINECS inventory. However they do not necessarily appear on the inventories of other countries and it may be necessary to use alternative numbers in territories outside of Europe. For example [*085535-84-8*] and [*085535-85-9*] do not exist in the USA's TSCA inventory, so [*61788-76-9*] "alkanes, chloro; alkanes chlorinated" is often used to describe these types of product.

Compounds of this sort were first detected in the middle of the 19th century by J. B. A. Dumsa. It was he who showed that long-chain paraffins, known up to this date only as relatively inert substances, are in fact subject to certain chemical reactions, and that the influence of artificial light promotes the processes. The first systematic study on chlorinated paraffins was conducted some years later, between 1856 and 1858, by P. A. Bolley [683].

Chlorinated paraffins acquired industrial importance in the early 1930s and consumption grew rapidly, particularly during World War II, with increasing use of the substances as flameproofing and rot-preventing agents. Further rapid expansion subsequently took place during the 1960s when low-cost normal paraffin became freely available as feedstocks and chlorinated paraffins were recognized as effective plasticizers for PVC. Recent developments in this area have seen the introduction of a range of chlorinated α-olefins and aqueous based processes for producing solid chlorinated paraffin waxes [684, 685].

The total world production of chlorinated paraffins is estimated to be 500 000 t per year,

and they are regarded as important substances with a wide range of applications. The major markets for chlorinated paraffins are the USA, Europe, and Asia. The vast majority of the market comprises of liquid chlorinated paraffins with only limited quantities of solid chlorinated paraffin produced.

7.1. Physical Properties

Chlorinated paraffins are homogeneous, neutral, colorless to pale yellow liquids. Their viscosities, densities and refractive indices rise with increasing chlorine content for a given carbon chain length and also with chain length at constant chlorine content (see Fig. 31). Volatilities, on the other hand, decrease with an increase in chain length and degree of chlorination (see Fig. 32). Physical properties of selected commercially available chlorinated paraffins are given in Table 33.

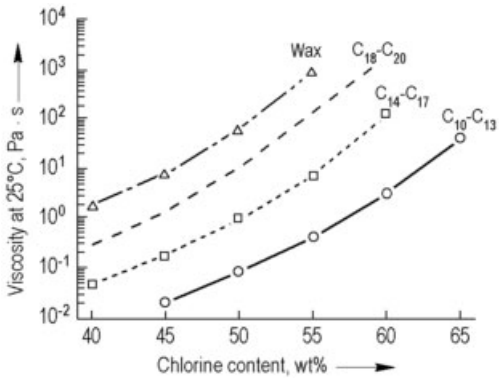

Figure 31. Viscosities of chlorinated paraffins

Chlorinated paraffins are practically insoluble in water and lower alcohols. They are soluble, however, in chlorinated aliphatic and aromatic hydrocarbons, esters, ethers, ketones, and mineral or vegetable oils. Their solubility in unchlorinated aliphatic and aromatic hydrocarbons is only moderate. Different types of chlorinated paraffins are completely miscible with one another.

At temperatures above 120 °C, some decomposition of chlorinated paraffins may occur accompanied by elimination of hydrogen chloride. The resulting polyalkenes once becoming conjugated cause a deepening of color.

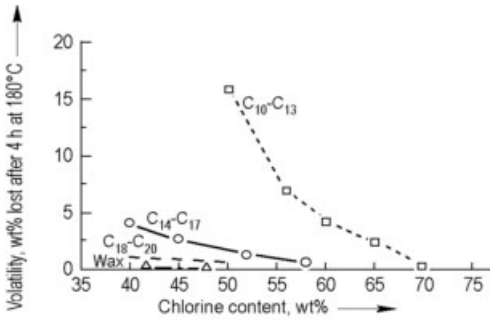

Figure 32. Densities of chlorinated paraffins

Accumulated hydrogen chloride operates as a catalyst for further dehydrochlorination.

Elimination of hydrogen chloride becomes significant at temperatures above 220 °C and for most applications this marks the upper limit of chlorinated paraffin use. For most commercially available grades volatility also becomes unacceptably high above this temperature.

The thermal stability of a chlorinated paraffin is defined by the extent to which it undergoes degradation at 175 °C over a 4 h period. The determination of the thermal stability requires use of a standardized piece of apparatus. There are a number of efficient stabilization systems, which increase the thermal stability of chlorinated paraffins, facilitating their use at the upper end of the temperature range noted above. Typically these involve the addition of small amounts of epoxide-containing compounds, antioxidants and organic phosphites. Likewise, stabilizers are known which prevent darkening of chlorinated paraffins caused by the adverse effects of ultraviolet light.

Apart from chlorine content and chain length, no convenient physical or physicochemical properties are available for the characterization of individual types of chlorinated paraffins.

7.2. Chemical Properties and Structure

Commercially available chlorinated paraffins are not simple well defined chemical compounds. Instead they are complex mixtures of many molecular species differing in the lengths

Table 33. Physical properties of selected commercial chlorinated paraffins

Trade name	Paraffin carbon chain length	Nominal chlorine content, wt %	Color hazen, APHA	Viscosity[a], mPas	Density[a], g/mL	Thermal stability[b], wt % HCl	Volatility[c], wt %	Refractive index
CERECLOR™ 50LV	$C_{10}-C_{13}$	50	100	80	1.19	0.15	16.0	1.493
CERECLOR™ 56L	$C_{10}-C_{13}$	56	100	800	1.30	0.15	7.0	1.508
CERECLOR™ 60L	$C_{10}-C_{13}$	60	135	3 500	1.36	0.15	4.4	1.516
CERECLOR™ 63L	$C_{10}-C_{13}$	63	125	11 000	1.41	0.15	4.3	1.522
CERECLOR™ 65L	$C_{10}-C_{13}$	65	150	30 000	1.44	0.20	2.5	1.525
CERECLOR™ 70L	$C_{10}-C_{13}$	70	200	800[d]	1.50[d]	0.20	0.5	1.537
CERECLOR™ S40	$C_{14}-C_{17}$	40	80	70	1.10	0.20	4.2	1.488
CERECLOR™ S45	$C_{14}-C_{17}$	45	80	200	1.16	0.20	2.8	1.498
CERECLOR™ S52	$C_{14}-C_{17}$	52	100	1 600	1.25	0.20	1.4	1.508
CERECLOR™ S58	$C_{14}-C_{17}$	58	150	40 000	1.36	0.20	0.7	1.522
CERECLOR™ M40	$C_{18}-C_{20}$	40	150	300	1.13	0.20	1.2	1.491
CERECLOR™ M47	$C_{18}-C_{20}$	47	150	1 700	1.21	0.20	0.8	1.506
CERECLOR™ M50	$C_{18}-C_{20}$	50	250	18 000	1.27	0.20	0.7	1.512
CERECLOR™ 42	$C_{>20}$	42	250	2 500	1.16	0.20	0.4	1.506
CERECLOR™ 48	$C_{>20}$	48	300	28 000	1.26	0.20	0.3	1.516
Solid chlorinated paraffin wax		70	100[e]		1.63	0.20	0.2	

[a] At 25 °C unless otherwise stated.
[b] Measured in a standard test for 4 h at 175 °C.
[c] Measured in a standard test for 4 h at 180 °C.
[d] At 50 °C.
[e] 10 g in 100 mL toluene.
[f] Solid, softening point 95 – 100 °C.
CERECLOR is a trade mark, the property of INEOS Chlor Limited.

of their carbon chains and in the number and relative positions of chlorine atoms present on each carbon chain. Their chemical formula can be stated as follows:

$$C_n H_{2n+2-m} \mathrm{Cl}_m$$

where n usually lies between 10 and 30 and m between 1 and 22. In the less chlorinated products, unchlorinated paraffin molecules may also be present. Studies have been conducted on the distribution of chlorine atoms within the molecule [686]. The results show that it is rare to find two chlorine atoms bound to one and the same carbon atom.

During paraffin chlorination, which is essentially a free radical process, tertiary carbon atoms react faster than those which are secondary; secondary carbon atoms in turn react faster than primary ones. The stability of chlorinated paraffins toward dehydrochlorination follows the reverse pattern, with chlorines attached to a tertiary carbon atom being the least stable and those linked to a primary one the most stable. Therefore, it can be concluded that most of the chlorine atoms present in the commercial product are attached to a secondary carbon atom with no more than about one in twelve primary-linked

hydrogens substituted by chlorine atoms. Any methyl side-chains are usually unchlorinated.

Because of their very limited reactivity, chlorinated paraffins are of little significance as chemical intermediates, although their conversion to alcohols [687] and the substitution of various chlorine atoms by sulfuric acid have been studied [688].

7.3. Production

Raw Materials. Because of the known instability of compounds containing tertiary-linked chlorine atoms, only straight-chain paraffins with a minimum content of branched isomers can be used as raw materials for the industrial production of useful chlorinated paraffins. The quality of the technical products underwent a significant increase when the petrochemical industry succeeded in producing paraffin fractions enriched in straight-chain components through urea adduction and later by treatment with molecular sieves.

The paraffin fractions used most frequently by producers of chlorinated paraffins are the three straight-chain mixtures $C_{10}-C_{13}$, $C_{14}-C_{17}$, $C_{18}-C_{20}$, which make short-chain chlorinated paraffins (SCCPs), medium-chain chlorinated paraffins (MCCPs), and long-chain chlorinated paraffins (LCCPs) respectively, together with waxes in the range of $C_{20}-C_{28}$. The feedstocks used most frequently to make chlorinated olefins are C_{14} or C_{16} linear α-olefins or mixtures of the two. The ranges employed can vary, depending on the regional availability of suitable raw materials.

Preparation. Chlorinated paraffins are prepared by reacting pure gaseous chlorine with the starting paraffins in the absence of any solvents at temperatures between 80 °C and 100 °C.

$$C_{1n}H_{2n+2}+m\mathrm{Cl}_2 \rightarrow C_nH_{2n+2-m}\mathrm{Cl}_m+m\mathrm{HCl}$$

Small amounts of oxygen are frequently used to catalyze the chlorination process. Temperatures above 120 °C must be avoided as these may cause dark or black products. Ultraviolet light is used by some manufacturers for initiating the reaction at relatively low temperatures. Once the reaction has started, the light source may be reduced in intensity or eliminated.

The reaction between chlorine and paraffin is exothermic with an enthalpy of reaction of ca. −150 kJ/mol; therefore, the reaction once initiated must be carefully controlled by cooling. Other important considerations are the need for efficient mixing to prevent localized overheating and control of the chlorine flow rate in order to minimize the amount of unreacted chlorine in the off gas. These requirements become increasingly important as the reaction proceeds and the viscosity of the product increases. For these reasons it is nearly impossible to prepare chlorinated paraffins containing more than 71 wt % chlorine by this method.

The reaction is terminated by stopping the chlorine flow once the required degree of chlorination is reached. The end point is assessed by a variety of methods including refractive index and viscosity. The product is then blown with nitrogen gas to remove any unreacted chlorine and residual HCl. In most case a small amount of a storage stabilizer, usually an epoxidized vegetable oil, is then added prior to sending the finished product for storage or drumming.

Other important considerations when producing chlorinated paraffins are the ability to deal effectively with HCl which is co-produced in large amounts.

Both batch and continuous processes are known for the preparation of chlorinated paraffins on an industrial scale. *Batch processes* are preferred, because of the large variety of special products synthesized. Figure 33 gives a general outline of the process employed in the commercial production of chlorinated paraffins.

The same basic principle is used in *continuous processes*. In this case, however, groups of three or four reactors are arranged in series.

The production of solid chlorinated paraffin wax generally employs a two-stage process, the first of which involves production of an intermediate liquid chlorinated paraffin wax with between 40–50 wt % chlorine. Traditionally, this was then dissolved in carbon tetrachloride solvent and further chlorinated to around 70 wt %, at which point the carbon tetrachloride was stripped off and the resulting cake ground to a fine powder. Since the beginning of the 1990s producers of the solid chlorinated paraffin wax have developed aqueous reaction systems to replace carbon tetrachloride [684] due to the restrictions on the use and emissions of car-

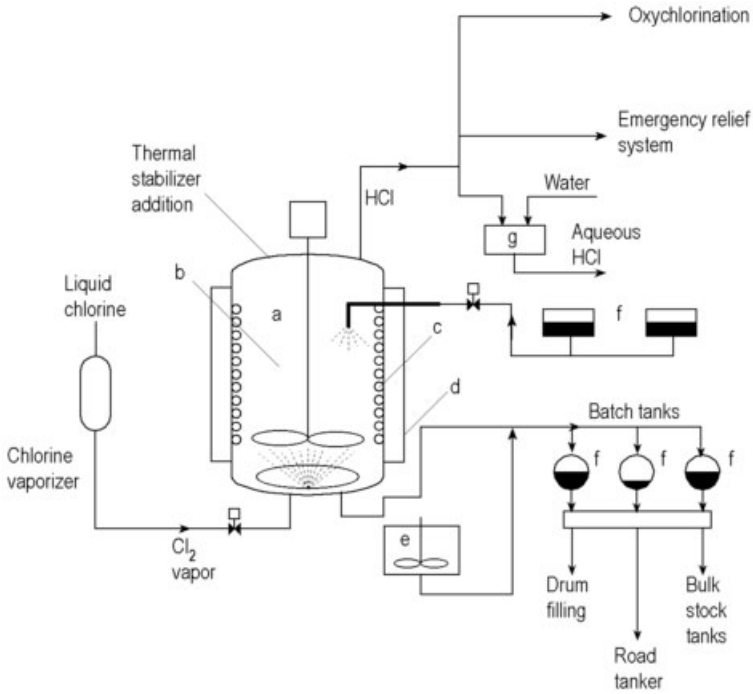

Figure 33. Commercial production of chlorinated paraffins
a) Reactor; b) Agitator; c) Heating/cooling coils; d) Jackets; e) Addition of special stabilizers; f) Batch tanks; g) Absorption of hydrogen chloride

bon tetrachloride, an ozone depleting substance, arising from the Montreal Protocol.

Construction Materials. Older reactors are generally lead-, ceramic- or occasionally silver-lined, whereas more modern reactors are normally constructed of glass-lined steel. Ferrous metal surfaces must be avoided to prevent discoloring or blackening of the finished product. Other equipment such as pumps, stirrers, pipework and valves should be constructed of corrosion-resistant materials and be resistant to hydrochloric acid.

Environmental Protection. The production process of chlorinated paraffin has minimal environmental impact because little contaminated water or other wastes are formed in the course of the reaction. Either after absorption or while it is still in the gaseous phase, the hydrogen chloride leaving the reactors as a byproduct must be cleaned in accordance with its future use. Hydrochloric acid is usually separated from any organic liquids and subsequently treated with activated charcoal. Purification of gaseous hydrogen chloride can be carried out after refrigeration by use of appropriate separators. Thus, the only wastes are those to be expected from cleaning procedures in the plant, in addition to activated charcoal. Such wastes and other byproducts should be burned in approved incinerators at a temperature above 1200 °C to prevent formation of cyclic or polycyclic chlorinated compounds.

Contamination of water caused by spillages must also be avoided. Asphalt surfaces of roads may be damaged by chlorinated paraffins, particularly when the product is at elevated temperatures.

Producers and Trade Names. Commercial grades of chlorinated paraffins are always mixtures of different alkanes chlorinated to varying degrees. Producers normally supply a range of materials intended to meet the differing needs of their customers.

In order to provide sufficient definition for individual products, numbers are often appended to the trade names. These typically indicate the approximate chlorine content of a given material and in some cases provide additional informa-

tion regarding the alkanes used as starting material (see Table 33).

Table 34 gives the major producers of chlorinated paraffins together with the corresponding trade names. There are also a number of smaller plants producing products to meet local market demands in China, India, and Russia. The largest single producer of chlorinated paraffins is INEOS, who operate three plants worldwide.

Table 34. Producers of chlorinated paraffins and trade names

Producer	Country	Trade name
INEOS	UK (Thailand, France)	CERECLOR
Caffaro	Italy	CLOPARIN, CLOPAROL
Leuna Tenside	Germany	HORDALUB, HORDAFLAM, HORDAFLEX
Dover	North America	PAROIL, CHLOROWAX
Navácke Chemické Závody	Slovakia	Chlorparaffin
Handy	Taiwan	PLASTOIL
NCP	South Africa	PLASTICLOR
Tosoh	Japan	TOYOPARAX
Asahi Denka	Japan	ADKciser

7.4. Analysis and Quality Control

The properties of chlorinated paraffins may be influenced by the process used for their preparation. Products prepared in a continuous process sometimes have a lower thermal stability and a higher volatility than products prepared in a batch process [689]. A less homogeneous distribution of the introduced chlorine seems to be responsible for this difference. Moreover, batch products are said to be more compatible with poly(vinyl chloride).

Even though chlorinated paraffins are well-known and very important substances with many applications, analysis for determining their composition or their presence in environmental samples is particularly problematic. Techniques employing gas chromatographic investigations (GC) are very difficult, both because chlorinated paraffins are complex mixtures and because dehydrochlorination may occur during separation in a GC column. Furthermore, the sensitivity of the usual GC detection systems is often insufficient. Capillary GC combined with negative ion chemical ionization mass spectrometry (MS) represents a suitable approach as does on-column reduction followed by either GC or combined GC-MS [690]. Resolution of groups of isomers and homologues has been obtained by a GC-MS method [691].

For routine quality control of chlorinated paraffins, usually the density, the viscosity, color, thermal stability and the chlorine content of specific products are determined. Approved standard methods are available for determination of densities, viscosities, and color. In order to determine their chlorine content, chlorinated paraffin samples are combusted under well-defined conditions, after which chloride is determined argentometrically, microcoulometrically, or by use of ion-selective electrodes. Good results have been obtained by the Wurtzschmitt method, in which a small sample of chlorinated paraffin is oxidized in a nickel bomb in the presence of an excess of sodium peroxide. Routine chlorine content determinations of production batches are usually carried out by direct correlation with other physical properties such as density or refractive index where well-established relationships exist for each feedstock.

7.5. Storage and Transportation

Chlorinated paraffins are non-corrosive substances at ordinary temperatures. Therefore, their storage for many months without deterioration of either products or containers is possible in mild steels drums. Nevertheless, temperatures of storage should not exceed 40 °C and prolonged storage should not be in direct sunlight, since this may cause discoloration. Reconditioned drums should be internally lacquered before use. In the case of bulk storage, stainless or mild steel tanks are recommended. If mild steel is used as the construction material, the tanks must be internally lined with epoxy or phenol – formaldehyde resins. Appropriate producers of resins should be contacted with respect to the suitability of their coatings. Many of the chlorinated paraffin types are very difficult to handle at low temperature because of their high viscosities. Installation of storage tanks in heated rooms is, therefore, advantageous in order to keep the products in a usable state. Outdoor tanks must be equipped with some sort of moderate warming system. Circulated warm water is recommended as the preferred heating agent, since it does not pro-

duce local hot spots where deterioration of the stored product might occur.

If steam is used for warming, its pressure must be reduced to ensure that the temperature will not exceed 105 °C at the inlet. It is also recommended that external pipe work is electrically trace-heated to prevent the formation of cold plugs of product. Chlorinated paraffins swell most types of rubber, therefore, any gaskets, etc. should be made of poly(tetrafluoroethylene).

Positive displacement type pumps, preferably ones equipped with an automatically operated security device, are preferred for material transfer at end use locations. It is particularly important that adequate flow rates be maintained on the suction side of such pumps.

Storage of both drums and bulk liquid in tanks should incorporate secondary containment such as bunding to prevent uncontrolled loss of product to the environment.

7.6. Toxicology, Environmental Impact and Regulation

The unreactive nature of chlorinated paraffins has led to them being generally regarded as low toxicity products. Studies have shown that they exhibit very low acute oral toxicity following a single exposure, with no signs of toxicity indicated in laboratory animals at doses of between 4 and 10 g per kilogram body weight [692]. Based on more limited data, chlorinated paraffins show similarly low acute toxicity for single exposures through dermal and inhalation routes. These results for acute toxicity are in accordance with the poor skin absorption and extremely low vapor pressures that characterize these substances.

Studies of longer-term exposure in a number of mammalian species have found target organ toxicity primarily in the liver, kidney, and thyroid gland [690]. An effect on postnatal survival in newborn rats, which is associated with an effect on Vitamin K levels, has also been described [693].

Chlorinated paraffins are considered to be nongenotoxic, supported by negative mutagenicity findings in bacterial [692] and animal bone marrow studies [690]. In 1985, however, the USA's National Toxicity Program (NTP) reported evidence of tumors in rats and mice as a result of studies of lifetime exposure to a C_{12} chlorinated paraffin with 58% chlorine [694]. This information has led some authorities to classify SCCPs ($C_{10}-C_{13}$) as carcinogens (e.g., IARC Group 2B: "possibly carcinogenic to humans"; European Union Category 3: "limited evidence of a carcinogenic effect"). Toxicological studies have elucidated the mechanism of the carcinogenic effect observed in rats and mice and have suggested that these effects are not relevant for humans [695, 696]. In general, the highest exposure levels under industrial conditions of use are likely to be far less than those that would elicit toxicological effects.

The water solubilities of chlorinated paraffins are low and decrease with increasing chain length. SCCPs are significantly more soluble (up to 150 μg/L) than the higher chain length materials. Laboratory studies have demonstrated that SCCPs exhibit toxicity effects towards fish and other varieties of aquatic life. Consequently these products are classified as "Dangerous For the Environment" by the European Union and as "Severe Marine Pollutants" under the UN International Maritime Organization regulations. MCCPs ($C_{14}-C_{17}$) show significantly lower toxicity to most aquatic species than the short-chain products, however, their recently confirmed toxicity towards *Daphnia magna*, an aquatic invertebrate species, has led some European producers to provisionally classify products based on MCCPs as "Dangerous For the Environment" in anticipation of a harmonized European Union classification expected to be implemented in the next adaptation of the EU's Dangerous Substances Directive. The very low solubilities and high molecular masses of the longer chain chlorinated paraffins means they show little or no toxicity towards aquatic species at or above their limits of solubility. In Germany chlorinated paraffins are given appropriate WGK classes under the criteria laid down by the Kommission zur Bewertung Wassergefährdender Stoffe. The classifications recommended by Eurochlor, representing the European producers of chlorinated paraffins, are as follows:

C_{10-13} Chlorinated paraffins	WGK class 3
C_{14-17} Chlorinated paraffins	WGK class 2
C_{18-20} Chlorinated paraffins (liquid)	WGK class 2

C_{18+} Chlorinated paraffins (solid)	WGK class 1
C_{20+} Chlorinated paraffins (liquid)	WGK class 1

Chlorinated paraffins biodegrade only slowly, the rate of biodegradation being higher for products with lower chlorine content and chain length. Bioconcentration factors (BCF) of 5300 for short-chain chlorinated paraffins, 1000 for medium chains and 50 for long-chain products have been measured [690, 697]. By comparison, the BCF for chlorinated paraffins are considerably lower than those for polychlorinated biphenyls (PCBs) and chlorine-containing pesticides, and there is no evidence of biomagnification of chlorinated paraffins by fish.

Recent regulatory activity has focussed on short-chain chlorinated paraffins in particular. The European Union undertook a Risk Assessment for SCCPs (under the EU Existing Substances Regulation 793/93), that was completed in 1999 and updated in 2005. The Risk Assessment has led to the implementation of a marketing and use restriction in Europe for SCCPs, prohibiting their use in metal-working fluids and leather treatment from January 2004. An EU Risk Assessment for medium-chain chlorinated paraffins is underway and nearing completion.

7.7. Uses

Chlorinated paraffins find widespread industrial use as plasticizers, flame-retardants, solvents, extreme-pressure additives and to lesser extent, as resin extenders. Choice of grade is frequently a compromise between the physical properties outlined above, in particular chlorine content, viscosity and volatility, together with a consideration of compatibility with the host polymer or base oil.

Plasticizers. This represents by far the largest use of chlorinated paraffins with the single largest application being the partial replacement of phthalate plasticizers in flexible PVC. The main reason for their use is reduction in formulation costs plus the additional benefits of flame retardancy, improved water and chemical resistance and better viscosity aging stability (i.e. a lesser increase in viscosity with time) of plastisols (→ Poly(Vinyl Chloride)). Typical end use applications include cables, flooring, wall covering and general extrusion and injection molding. Examples of the typical products most frequently used in these applications are CERECLOR S45 and S52 chlorinated paraffins (Table 33).

Chlorinated paraffins also act as plasticizers in polyurethane and liquid polysulfide sealants where they also replace phthalates. Again cost reduction is an important feature as is their very low solubility in water and stability towards biodegradation, hence their use in sealants for aggressive biological environments such as sewage treatment works. Very low volatility also makes them useful as plasticizers for insulating glass sealants. Typical grades used in these applications are CERECLOR 56L, 63L, S52 and M50 chlorinated paraffins (Table 33). Adhesive systems such as certain types of hot melt, pressure sensitive and poly(vinyl acetate) emulsion adhesives are also effectively plasticized and tackified by chlorinated paraffins.

Finally their very low volatility, inert nature and low water solubility makes them effective plasticizers for a range of paint systems. In general use is focused in heavy duty industrial applications with typical host paint resins being chlorinated rubbers, chlorosulfonyl polyethylene, styrene – butadiene rubbers, and modified acrylics. Typical grades used for paint applications are those based on wax feedstocks such as CERECLOR 42 chlorinated paraffin (Table 33). Solid 70 wt % chlorinated paraffin waxes also find use in this application where their low plasticizing action allows them to act as a paint resin extender as well as conferring additional fire retardancy.

Fire Retardants. Chlorinated paraffins are excellent cost effective flame-retardants [698], however their plasticizing action and limit on upper processing temperature can restrict their use in this application area. An important area of use is PVC where chlorinated paraffins act as a fire retardant plasticizer and are used to partially replace more expensive phosphate plasticizers in applications such as mine belting and safety flooring. Chlorinated paraffins are often used in combination with a synergist, such as antimony trioxide, to enhance the fire retardancy of phthalate-based PVC formulations, such as those found in fire retardant wire and cable. Other fire retardant applications include their use

in a range of rubbers including natural, nitrile, styrene – butadiene and chlorosulfonyl polyethylene rubbers. Also of importance is their use in polyurethanes, in particular rigid foams and one-component foams (OCF) and unsaturated polyesters. The actual grades used vary considerably depending on the host polymer. Solid 70 wt % chlorinated paraffin wax may be used as a flame retardant and can be used in polyethylene, polypropylene and high-impact polystyrene compounds in addition to those polymers already mentioned.

Extreme-Pressure Additives in Metal Working. Chlorinated paraffins are used as extreme-pressure additives to enhance lubrication and surface finish in demanding metal working and forming applications where hydrodynamic lubrication cannot be maintained. They function by providing a convenient source of chlorine that is liberated by frictional heat to form a chloride layer on the metal surface. This film has a lower shear strength than the metal itself, so the friction between the metals in sliding contact is reduced. Chlorinated paraffins are used predominantly in neat oils and to a lesser extent in soluble oil emulsions. They are frequently used in combination with other extreme-pressure additives including fatty acids, phosphorus- and sulfur-containing compounds, with which they display a synergistic action. Typical end use applications are stamping, forming, drawing and a range of cutting operations such as broaching. Historically, grades used in this application were based on $C_{10} - C_{13}$ paraffins, however, as a result of environmental concerns and the market restrictions in Europe described in Section 7.6, producers have developed highly effective medium chain grades for extreme pressure use, e.g., the CERECLOR E grade products of INEOS Chlor. In North America chlorinated $\alpha\alpha$olefins also find widespread use as extreme pressure additives.

Solvents. Chlorinated paraffins have found use as solvents for the color formers used in carbonless copy paper. The main benefits are good solvating power for the color formers, low volatility, stability in the encapsulation process and fast color development especially with blue color formers and clay-coated color fronts (CF). Chlorinated paraffins also tolerate higher levels of the diluents widely used by the industry compared to competitive products such as dialkyl naphthalenes.

7.8. Summary

Chlorinated paraffins are a versatile range of substances, which function as cost effective property-enhancing additives in a wide range of important end applications including secondary plasticizers and flame retardants in flexible PVC and other resins used in paints, sealants, foams and safety flooring, and as extreme-pressure additives in metal-working fluids. The thorough investigation of the varied toxicological profile and environmental impact of the different types of chlorinated paraffins allows for their continued safe use through risk assessment and the implementation of responsible working practices.

8. Nucleus-Chlorinated Aromatic Hydrocarbons

The term "nucleus-chlorinated aromatic compound" as used here refers to a substance containing a mesomeric π-electron system in a carbocyclic framework in which at least one of the ring carbons bears a chlorine substituent rather than hydrogen.

Laboratory work on the compounds making up this class began long ago. For example, A. Laurent reported in 1833 that he had obtained waxlike compounds in the course of chlorinating naphthalene. Nevertheless, their industrial manufacture and use was delayed until the first third of the 20th century.

Chlorinated aromatic hydrocarbons are of substantial economic significance. This is particularly true of the chlorinated benzenes, the most important being monochlorobenzene, and the chlorinated toluenes. The compounds are now recognized as important starting materials and additives in the production of high-quality insecticides, fungicides, herbicides, dyes, pharmaceuticals, disinfectants, rubbers, plastics, textiles, and electrical goods.

In general, the environmental degradability of heavily chlorinated organic compounds, whether by biotic or by abiotic mechanisms, is low. This persistence has led in recent years to

such drastic measures as prohibitions, restrictions on production and use, and legislation regulating waste disposal. Some highly chlorinated aromatics have been affected as well.

8.1. Chlorinated Benzenes

8.1.1. Physical Properties

Monochlorobenzene C_6H_5Cl, is a colorless liquid which is volatile with steam and is a good solvent. It is miscible with all commonly used organic solvents and forms many azeotropes [699, 700]. Monochlorobenzene is flammable and has an aromatic odor. Upon addition of 24 % benzene, it forms a eutectic mixture with a solidification point of − 60.5 °C.

Important physical data for monochlorobenzene are compiled in Table 35.

Dichlorobenzene $C_6H_4Cl_2$, occurs in three isomeric forms. 1,2-Dichlorobenzene is a colorless, mobile liquid which is miscible with the commonly used organic solvents. It is difficult to ignite and has an unpleasant odor. The properties of 1,3-dichlorobenzene are similar to those of 1,2-dichlorobenzene. 1,4-Dichlorobenzene is a white, volatile, crystalline compound soluble in many organic solvents. It has a strong camphor odor. 1,4-Dichlorobenzene occurs as a stable monoclinic α-modification, which is transformed into the triclinic β-form at 30.8 °C.

Eutectic mixtures of the dichlorobenzenes are: 86.0 % 1,2-dichlorobenzene, 14.0 % 1,4-dichlorobenzene, solidification point − 23.7 °C; 85.3 % 1,3-dichlorobenzene, 14.7 % 1,4-dichlorobenzene, solidification point − 30.8 °C. Further physical data are given in Table 35.

Trichlorobenzene $C_6H_3Cl_3$, occurs in three isomeric forms. 1,2,3-Trichlorobenzene forms colorless tabular crystals. 1,2,4-trichlorobenzene is a colorless liquid of low flammability, and 1,3,5-trichlorobenzene forms long, colorless needles.

The trichlorobenzenes are insoluble in water, slightly soluble in alcohol, and very soluble in benzene and solvents like petroleum ether, carbon disulfide, chlorinated aliphatic and aromatic hydrocarbons.

A combination of 34 % of 1,2,3-trichlorobenzene and 66 % of 1,2,4-trichlorobenzene forms a eutectic mixture with a solidification point of + 1.5 °C. Further physical data are given in Table 35.

Tetrachlorobenzene $C_6H_2Cl_4$, also occurs in three isomeric forms: 1,2,3,4-Tetrachlorobenzene and 1,2,3,5-tetrachlorobenzene crystallize as colorless needles. 1,2,4,5-Tetrachlorobenzene forms colorless, sublimable needles with a strong, unpleasant odor.

All of the tetrachlorobenzenes are insoluble in water, but they are soluble in many organic solvents, particularly at an elevated temperature. Further physical data are given in Table 36.

Pentachlorobenzene C_6HCl_5, forms colorless needles and is insoluble in water.

Hexachlorobenzene C_6Cl_6, forms colorless and sublimable prismatic crystals. It is insoluble in water, but soluble at an elevated temperature in several organic solvents (e.g., benzene, chloroform, and ether).

8.1.2. Chemical Properties

The chlorobenzenes are neutral, thermally stable compounds. Reactions may occur to replace hydrogen at unsubstituted positions on the ring (e.g., halogenations, sulfonations, alkylations, nitrations), by substitution of the chlorine (e.g., hydrolysis), and with de-aromatization (e.g., chlorine addition).

Monochlorobenzene. In this compound the chlorine is firmly bound to the aromatic ring and can only be substituted under energetic conditions. Chlorobenzene can be hydrolyzed to phenol with aqueous sodium hydroxide at 360 – 390 °C under high pressure [701 – 703] or with steam at 400 – 450 °C over calcium phosphate. Monochlorobenzene reacts with ammonium hydroxide at high temperature and in the presence of copper catalysts to give aniline [704].

In electrophilic substitution, e.g., nitration, the directing influence of the chlorine atom leads to the formation of derivatives in which the added substituent is found predominantly in the ortho or para position.

Light-induced addition of chlorine produces heptachlorocyclohexane [715].

Table 35. Physical properties of mono- to trichlorobenzenes

	Monochloro-benzene	1,2-Dichloro-benzene	1,3-Dichloro-benzene	1,4-Dichloro-benzene	1,2,3-Trichloro-benzene	1,2,4-Trichloro-benzene	1,3,5-Trichloro-benzene
				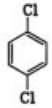		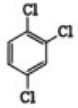	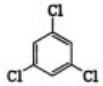
	[*108-90-7*]	[*95-50-1*] [*25321-22-6*]	[*541-73-1*]	[*106-46-7*]	[*87-61-6*] [*12002-48-1*]	[*120-82-1*]	[*108-70-3*]
M_r	112.56	147	147	147	181.45	181.45	181.45
Melting point, °C	−45.2	−16.7	−26.3	53.5 (α-mod.) 54 (β-mod.)	53.5	17.0	63.5
Solidification point, °C	−45.58	−17	−24.8	53.08	52.4	17.2	62.6
Boiling point at 101.3 kPa, °C	132.2	179.0	173	173.9	218.5	213.5	208.4
Heat of fusion, J/g	84.9	87.8	85.7	123.5	95.9	85.3	95.9
Heat of vaporization, J/g							
at 50 °C	355.5						
100 °C	329.5	308.1	295.2	299.0	291.0	285.5	278.8
150 °C	307.7	286.0	273.4	277.4	271.7	265.9	259.2
200 °C	276.3	262.5	251.2	253.3	251.6	246.0	239.1
bp at 101.3 kPa	315.0	273.0	262.9	264.6	242.4	241.2	236.1
Temperature °C, corresponding, to vapor pressure							
0.13 kPa	−13.0	20.0	12.1	16.0	40.0	38.4	37.2
0.67 kPa	+10.6	46.0	39.0		70.0	67.3	63.8
1.33 kPa	22.2	59.1	52.0	54.8	85.6	81.7	78.0
2.67 kPa	35.3	73.4	66.2	69.2	101.8	97.2	93.7
5.33 kPa	49.7	89.4	82.0	84.8	119.8	114.8	110.8
8.0 kPa	58.3	99.5	92.2	95.2	131.5	125.7	221.8
13.3 kPa	70.7	112.9	105.0	108.4	146.0	140.0	136.0
26.7 kPa	89.4	133.4	125.9	128.3	168.2	162.0	157.7
53.3 kPa	110.0	155.8	149.0	150.2	193.5	187.7	183.0
101.3 kPa	132.2	179.0	173.0	173.9	218.5	213.5	208.4
Heat capacity, $J\,g^{-1}\,K^{-1}$	1.338 (20 °C) 1.462 (100 °C)	1.131 + 0.0013 δ (− 37 to 104 °C)	1.131 + 0.0013 δ (− 38 to 104 °C)	0.992 + 0.0029 δ (− 78 to 52 °C) 1.088 (55 °C)		1.01	
Thermal conductivity, $W\,m^{-1}\,K^{-1}$							
at −40 °C	0.141						
−20 °C	0.137						
0 °C	0.133						
20 °C	0.128	0.124				0.114	
50 °C	0.124			0.105 (60 °C)			
100 °C	0.115						

Table 35. (continued)

Standard heat of formation, kJ/mol	-10.65 ± 0.8	-17.57 ± 0.7	-20.46	-42.34 ± 1.1		-47.7	
Density of the liquid, g/cm³							
at 20 °C	1.106	1.306	1.288			1.456	
25 °C	1.101	1.3	1.282	1.504			
30 °C	1.096	1.295	1.277	1.25 (55 °C)			
70 °C	1.052	1.249	1.234	1.231			
100 °C					1.381	1.358	1.356
Viscosity, mPa · s	0.806 (20 °C)	1.4 (20 °C)	1.11 (20 °C)	0.839 (55 °C)	1.68 (50 °C)	2.08 (20 °C)	0.841 (70 °C)
					0.91 (100 °C)	0.74 (100 °C)	0.642 (100 °C)
Heat of combustion at 25 °C, kJ/mol	−3100	−2962	−2955	−2934			
Surface tension at 20 °C, mN/m	33.5	36.6	36.2	29.9 (55 °C)		39.1	
Dielectric constant							
at 20 °C	5.641	9.82	4.9	2.67		3.98	
58 °C	5.02	8.5	4.63	2.62			
Refractive index, n_D^{20}	1.5248	1.5505	1.5464	1.5284 (55 °C)		1.5717 (25 °C)	
Coefficient of expansion of liquid, K⁻¹	95×10^{-5}	110×10^{-5}	111×10^{-5}	116×10^{-5}			
Critical temperature, °C	359.2	424.1	410.8	407.5	489.5	461.8	470.8
Critical pressure, mPa	4.52	4.1	3.78	4.11	3.01	3.99	3.01
Critical density, g/cm³	0.365	0.408	0.41	0.395			
Flash point, °C	28	67	65	65	113	110	107
Ignition temperature, °C	590	640	> 500	> 500	> 500	> 500	> 500
Classification of danger (Verordnung über brennbare Flüssigkeiten – Flammable Liquids Order of the FRG)	A II	A III	A III	A III			
Explosive limits in air							
lower: vol% (g/cm³)	1.3 (60)	2.2 (130)					
upper: vol% (g/cm³)	11 (520)	12 (750)		ca. 18			
Maximum vapor concentration, g/cm³							
at 20 °C	54	7.8		10		2.7	
30 °C	89	15		19		5.8	
50 °C	222	45		60		18	
MAK (FRG), ppm	50	50		75		5	
mg/m³	230	300		450		40	
TLV-TWA (USA), ppm	75	50		75		5	
mg/m³	350	300		450		40	
Solubility in water, g/100 g							
at 20 °C	0.05	0.015	0.11	0.007			
60 °C		0.023	0.20	0.016			

Table 36. Physical properties of tetra-, penta-, and hexachlorobenzenes

	1,2,3,4-Tetrachlorobenzene	1,2,3,5-Tetrachlorobenzene	1,2,4,5-Tetrachlorobenzene	Pentachlorobenzene	Hexachlorobenzene
	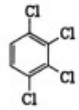	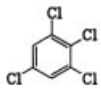	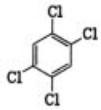	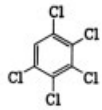	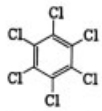
	[*634-66-2*] [*12408-10-5*]	[*634-90-2*]	[*95-94-3*]	[*608-93-5*]	[*118-74-1*]
M_r	215.90	215.90	215.90	250.34	284.78
Melting point, °C	47.5	51	141	85	229
Solidification point, °C	46.5	49.6	139.6	83.7	228.2
Boiling point at 101.3 kPa, °C	254	246	245	276	322
Heat of fusion, J/g				70.3	79.4
Heat of vaporization, J/g					
at 150 °C	254.6 (100 °C)				
200 °C	241.3		249		
bp at 101.3 kPa	216.6	217.9	224		191
Temperature °C, corresponding to vapor pressure,					20.0
0.31 kPa	68.5	58.2		98.6	130.0
0.67 kPa	99.6	89		129.7	
1.33 kPa	114.7	104.1		144.3	174.3
2.67 kPa	131.2	121.6		160.0	
5.33 kPa	149.2	140.0	146	178.5	209.5
8.0 kPa	160.0	152.0	157.7	190.1	
13.3 kPa	175.7	168.0	173.5	205.5	336.9
26.7 kPa	198.0	193.7	196	227.0	
53.3 kPa	225.3	220.0	220.3	251.6	290.5
101.3 kPa	254.0	246.0	245	276.0	322.0
Heat capacity, $J\,g^{-1}\,K^{-1}$					0.7 (25 °C)
Standard heat of formation, kJ/mol					− 131
Liquid density, g/cm³	1.539 (100 °C)	1.523 (100 °C)	1.454 (150 °C)	1.609 (84 °C) 1.625 (100 °C)	2.044 (23 °C) 1.462 (306 °C)
Heat of combustion at 25 °C, kJ/mol					− 2372
Critical pressure, MPa	3.38	2.8	3.38		
Critical temperature, °C	498.5	526.1	486.6		
Maximum vapor concentration at 20 °C, g/cm³			2.5		0.00018

Dichlorobenzenes. The reactivity to further substitution on the ring increases from 1,4- via 1,2- to 1,3-dichlorobenzene.

The position at which a third substituent is introduced into the ring depends on the directing influence of the two chlorine atoms. Thus, for example, no 1,3,5-derivatives can be formed in this way. Electrophilic substitution of 1,2-dichlorobenzene leads to 4-derivatives as main products and 3-derivatives as byproducts. Electrophilic substitution of 1,3-dichlorobenzene gives 4-derivatives as main product and 2-derivatives as byproduct, whereas 1,4-dichlorobenzene as starting material yields 2-derivatives.

The action of alkaline solutions or alcoholic ammonia solution on dichlorobenzene at 200 °C under pressure gives chlorophenols and dihydroxybenzenes, or chloroanilines and phenylenediamines, respectively. The addition of chlorine leads to octachlorocyclohexane [715].

Trichlorobenzenes. The reactivity of these compounds toward chlorine decreases in the order 1,3,5- > 1,2,3- > 1,2,4-trichlorobenzene. As expected, electrophilic substitution occurs preferentially at certain positions on the aromatic ring: With 1,2,3-trichlorobenzene an electrophilic substituent is led into the 4-position, with 1,2,4-trichlorobenzene into the 5-position

(main product) and 3-position (byproduct), and with 1,3,5-trichlorobenzene into the 2-position.

The trichlorobenzenes can be hydrolyzed to dichlorophenol.

Tetrachlorobenzenes. Like the lower chlorinated benzenes, the tetrachlorobenzenes can be chlorinated and nitrated. The reactivity toward chlorine decreases from 1,2,3,5- via 1,2,3,4- to 1,2,4,5-tetrachlorobenzene. At a temperature above 160 – 180 °C, the chlorine substituents can be hydrolyzed with sodium hydroxide in methanol. In the preparation of trichlorophenol from tetrachlorobenzene, the toxic polychlorinated dibenzo-*p*-dioxins – and also the extremely toxic compound 2,3,7,8-tetrachlorodibenzo-*p*-dioxin – may be formed if the very narrowly defined reaction conditions are not precisely maintained.

Pentachlorobenzene. Pentachlorobenzene can be chlorinated to hexachlorobenzene and nitrated to pentachloronitrobenzene. Hydrolysis to 2,3,4,5- and 2,3,5,6-tetrachlorophenol is also possible.

Hexachlorobenzene. Hexachlorobenzene, like the other polychlorobenzenes, can be dehalogenated to lower chlorinated benzenes with hydrogen or steam at a temperature above 500 °C in the presence of catalysts [716]. At a high temperature a mixture of hexachlorobenzene, chlorine, and ferric chloride gives carbon tetrachloride in high yield [717]. Reaction with sodium hydroxide and methanol leads to pentachlorophenol.

8.1.3. Production

Benzene Chlorination in the Liquid Phase. Chlorobenzenes are prepared industrially by reaction of liquid benzene with gaseous chlorine in the presence of a catalyst at moderate temperature and atmospheric pressure. Hydrogen chloride is formed as a byproduct. Generally, mixtures of isomers and compounds with varying degrees of chlorination are obtained, because any given chlorobenzene can be further chlorinated up to the stage of hexachlorobenzene. Because of the directing influence exerted by chlorine, the unfavored products 1,3-dichlorobenzene, 1,3,5-trichlorobenzene, and 1,2,3,5-tetrachlorobenzene are formed to only a small extent if at all. The velocity of chlorination for an individual chlorine compound depends on the compound's structure and, because of this, both the degree of chlorination and also the isomer ratio change continuously during the course of a reaction. Sets of data on the composition of products from different reactions are only comparable with one another if they refer to identical reaction conditions and materials having the same degree of chlorination. By altering the reaction conditions and changing the catalyst, one can vary the ratios of the different chlorinated products within certain limits. Lewis acids ($FeCl_3$, $AlCl_3$, $SbCl_3$, $MnCl_2$, $MoCl_3$, $SnCl_4$, $TiCl_4$) are used as principal catalysts (Table 37). Elevated temperatures in substitution reactions favor the introduction of a second chlorine atom in the ortho and meta positions, whereas para substitution is favored if cocatalysts and lower temperatures are used. As a further example of the influence of catalysts on the composition of the product, attention is drawn to the formation of 1,2,4,5-tetrachlorobenzene (Table 38). The optimal reaction temperature depends on the desired degree of chlorination. Mono- and dichlorination are carried out at 20 – 80 °C. In the production of hexachlorobenzene by chlorination of benzene, however, temperatures of about 250 °C are needed toward the end of the reaction.

The usual catalyst employed in large scale production is ferric chloride, with or without the addition of sulfur compounds. Ferric chloride complexed with 1 mol of water is claimed to have the best catalytic effect [732]. Benzene and chlorine of technical purity always contain some water, however, thus, it may be that this hydrate compound is always present in iron-catalyzed industrial reactions.

The ratio of resulting chlorobenzenes to one another is also influenced by the benzene : chlorine ratio. For this reason, the highest selectivity is achieved in batch processes. If the same monochlorobenzene : dichlorobenzene ratio expected from a batch reactor is to result from continuous operation in a single-stage reactor, then a far lower degree of benzene conversion must be accepted (as a consequence of a low benzene : chlorine starting ratio). The selectivity of a continuous reactor approaches that of a discontinuous reactor as the number of reaction stages is

Table 37. Influence of catalysts on the ratio 1,4-:1,2-dichlorobenzene

Catalyst	Proportion of 1,4-dichlorobenzene (in %) in the dichlorobenzene fraction	Ratio 1,4- : 1,2-dichlorobenzene	References
$MnCl_2 + H_2O$	ca. 50	1.03	[718]
$SbCl_5$		1.5	[719]
$FeCl_3$ or Fe	ca. 59	1.49 – 1.55	[720, 724, 727]
Metallosilicon organic compounds	61 – 74	1.56 – 2.8	[721]
$AlCl_3 - SnCl_4$		2.21	[722]
$AlCl_3 - TiCl_4$		2.25	[722]
Fe – S – PbO	ca. 70		[723]
$FeCl_3$ – diethyl ether		2.38	[724]
Aluminum silicate – hexamethylene-diamine		2.7	[725]
$FeCl_3 - S_2Cl_2$	ca. 76		[726]
$FeCl_3$ – divalent organic sulfur compounds	ca. 77	3.3	[727]
L-type zeolite	ca. 88	8.0	[728]
$TiCl_4$ (chlorinating agent is $FeCl_3$)		20 – 30	[729]

Table 38. Influence of catalysts on the ratio 1,2,4,5-:1,2,3,4-tetrachlorobenzene

Catalyst	Substrate	Ratio 1,2,4,5- : 1,2,3,4-tetrachlorobenzene	Ref.
$FeCl_3$	1,2,4-trichlorobenzene	2.3 – 2.4	[720], [727]
Lewis acids – divalent organic sulfur compounds	1,2,4-trichlorobenzene	3	[727]
$FeCl_3$ – aromatic iodine compounds	1,2,4-trichlorobenzene	2.7 – 4.6	[730]
$SbCl_3 - I_2$	benzene	10 – 17	[731]

increased [733]. Mathematical models analyzing and interpreting benzene chlorination will be found in [733 – 736] and elsewhere.

Solvents can also influence the chlorination rate as well as the selectivity of the reaction, although solvents are not used in industrial chlorination.

Continuous Chlorination. Benzene or a chlorobenzene derivative is treated with chlorine gas in a suitable reactor in the presence of dissolved ferric chloride. The reactants must be mixed as intensively as possible. The catalyst can be introduced along with the substrate or it can be allowed to form during the reaction on the surface of iron rings in the reactor. The reaction is highly exothermic.

$$C_6H_6 + Cl_2 \rightarrow C_6H_5Cl + HCl \qquad \Delta H = -131.5 \text{ kJ/mol}$$
$$C_6H_5Cl + Cl_2 \rightarrow C_6H_4Cl_2 + HCl \qquad \Delta H = -124.4 \text{ kJ/mol}$$
$$C_6H_4Cl_2 + Cl_2 \rightarrow C_6H_3Cl_3 + HCl \qquad \Delta H = -122.7 \text{ kJ/mol}$$
$$C_6H_3Cl_3 + Cl_2 \rightarrow C_6H_2Cl_4 + HCl \qquad \Delta H = -115.1 \text{ kJ/mol}$$

Unwanted heat of reaction can be dissipated either by circulating some of the reactor liquid through an external heat exchanger or by permitting evaporative cooling to occur at the boiling temperature. Circulation cooling has the advantage of enabling the reaction temperature to be varied in accordance with the requirements of a given situation. Evaporative cooling is more economical, however.

The reactor must be designed to ensure that the liquid within it has a suitable residence spectrum, since this favors high chlorination selectivity. As noted above, the quantity ratio of the chlorobenzenes to one another is determined by the benzene : chlorine starting ratio.

Almost quantitative conversion of chlorine is achieved in continuously operated plants for the manufacture of mono- and dichlorobenzenes under normal operating conditions.

Cast iron, steel, nickel, and glass-lined steel can be used as construction materials. However, all starting materials must be substantially free from water; otherwise, severe corrosion is caused by the hydrochloric acid formed. Intrusion of water is dangerous for another reason: water causes the ferric chloride catalyst to be inactivated. If this occurs, chlorine collects in the reactor and exhaust system, and local overheating may ensue, causing spontaneous combustion of the aromatic hydrocarbon with chlorine in a highly exothermic reaction to form carbon and hydrogen chloride (2 mol of HCl/mol of Cl_2). Furthermore, above 280 °C metallic iron begins to burn in the chlorine stream. It is, therefore, advisable to monitor the reaction continuously by observing the heat production rate and the chlorine content of the waste gas.

On leaving the reactor, the liquid and gas portions of the reaction mixture are separated. The waste gas contains hydrogen chloride and — in proportions corresponding to their vapor pressures at the temperature of the waste gas — benzene and chlorobenzenes. If the chlorine conversion is incomplete, chlorine may be present as well. Concerning the treatment of the waste gas, see Section 8.5.

The liquid phase contains benzene, chlorobenzenes, hydrogen chloride, and iron catalyst. Production processes exist in which the product mixture is neutralized with sodium hydroxide solution or soda before it is subjected to fractional distillation. In modern continuous distillation trains, the mixture of products can be distilled without preliminary treatment, however. The separated fractions consist of benzene, monochlorobenzene, dichlorobenzenes, trichlorobenzenes, and higher chlorobenzenes. Iron catalyst is removed along with the distillation residue, disposal of which is discussed in Section 8.5. Dissolved hydrogen chloride is removed during the benzene distillation and combined with the waste gas. Unreacted benzene is recycled to the reactor.

Whereas chlorobenzene and 1,2-dichlorobenzene are obtainable as pure distillates, the other fractions are mixtures of close-boiling polychlorobenzene isomers. A further separation, insofar as this is economically justifiable, would consist of combined crystallization/distillation processes.

Discontinuous Chlorination. Batch chlorination of liquid, molten, or dissolved aromatics is carried out industrially in agitator vessels equipped with external or internal cooling. The agitator must provide maximum exchange between the liquid and gas phases. Gaseous chlorine is introduced through a valve at the bottom of the vessel or through an ascension pipe beneath the agitator. The vessel may be constructed of glass-lined steel, cast iron, steel, or nickel. Absence of water must be ensured as a precaution against corrosion. The degree of chlorination can be ascertained from density determinations. The amount of chlorine which can be introduced in unit time depends on the heat output of the vessel and on the chlorine conversion rate. It is obviously desirable to keep the chlorine content of the off-gas as low as possible. It is also important that the reaction begins as soon as chlorine is introduced (the beginning of the reaction is indicated by a temperature increase and by the formation of hydrogen chloride). Chlorine is soluble in many hydrocarbons (see Table 39); therefore, if the onset of reaction is delayed, hydrogen chloride may be formed very rapidly, resulting in a substantial increase in the temperature of the reactants in the vessel.

Other Benzene Chlorination Processes. The following additional benzene chlorination processes are known:

1) Chlorination in the vapor phase with chlorine
2) Chlorination in the vapor or liquid phase with hydrogen chloride and air (oxychlorination)
3) Chlorination with chlorine-containing compounds
4) Electrolysis of benzene and hydrochloric acid

Apart from oxychlorination these processes are not industrially important. In oxychlorination, benzene vapor and a mixture of hydrogen chloride and air are reacted at about 240 °C in the presence of catalysts (e.g., $CuCl_2$–$FeCl_3/Al_2O_3$ or $CuO-CoO/Al_2O_3$). The main product is monochlorobenzene accompanied by 6 – 10 % dichlorobenzene [738 – 746].

$$C_6H_6 + HCl + 0.5O_2 \rightarrow C_6H_5Cl + H_2O$$

This process was developed in connection with the production of phenol from chlorobenzene (Raschig-Hooker process; → Phenol). The

Table 39. Solubility (in wt %) of chlorine in benzene and chlorobenzenes [737]

Solvent	Temperature			
	20 °C	50 °C	75 °C	100 °C
Benzene	27	6	0.8	
Monochlorobenzene	15	5		1.2
1,2-Dichlorobenzene	9	4		1.1
1,2,4-Trichlorobenzene	6	3.5		1.0

benzene conversion must be limited to 10 – 15 % in order to control the heat in the catalyst solid bed (222 kJ/mol). An excessively high reaction temperature favors the formation of dichlorobenzene and a side reaction, the highly exothermic oxidation of benzene to carbon dioxide and water (330 kJ/mol).

Concerning material balance and heat, see [747]. The high cost of energy and for a corrosion-resistant plant, as well as an insufficient shift of the ratio of monochlorobenzene to dichlorobenzene, makes the process uneconomical.

Gulf has developed a process for the oxychlorination of benzene in the liquid phase under pressure with aqueous hydrochloric acid, catalytic quantities of nitric acid, and air or oxygen. The process leads to a high rate of benzene conversion and good selectivity for monochlorobenzene [748], but it has not been reported whether commercial operation has commenced.

The reaction of benzene and chlorine at 400 to 500 °C in the vapor phase [749, 750] is likewise uneconomical. One interesting feature is associated with vapor phase chlorination catalyzed by non-metals, however: the isomer ratio at the dichlorobenzene stage is shifted in the direction of *m*-dichlorobenzene (10 % ortho, 66 % meta, 24 % para isomer). This effect is a consequence of the radical chain mechanism that is followed [751, 752].

Chlorine compounds are more selective for chlorination than chlorine itself, and they are thus used in the manufacture of special chlorobenzenes. The range of suitable chlorinating agents includes metal chlorides (e.g., of iron, antimony, titanium, and copper) of different valence stages [753 – 760], sulfuryl chloride [761 – 764], chlorine monoxide [765], and chlorosulfuric acid [766]. The latter three compounds are particularly suitable for the perchlorination of aromatics. In the chlorination of chlorobenzene with ferric chloride, high selectivity in favor of 1,4-dichlorobenzene (1,4 : 1,2-dichlorobenzene ratio till 25) is achieved [729, 759].

Chlorobenzenes can also be produced by electrolysis [767, 768], resulting in high selectivity for monochlorobenzene (98 %) and in the exceptional yield of 94 % [769].

Other Processes. The conversion of substituted aromatics to chlorobenzenes is of particular interest in connection with the preparation of not easily accessible isomers, such as 1,3-dichlorobenzene and 1,3,5-trichlorobenzene. The replacement of amino [770, 771], nitro [772 – 774], sulfonic acid or sulfonyl chloride [775 – 779], or acyl chloride groups [780] by chlorine leads to appropriate specialty chlorobenzenes.

8.1.3.1. Monochlorobenzene

Most monochlorobenzene is now produced from benzene and chlorine in continuously operated plants. Depending on the ratio of benzene to chlorine chosen, one can achieve either a low rate of benzene conversion and little dichlorobenzene formation, or almost complete conversion of the benzene with a higher degree of dichlorobenzene formation. Which of the two alternatives is favored depends on a profitability calculation, in which the distillation costs occasioned by the dichlorobenzenes need to be taken into account. The composition of a chlorination mixture containing the highest possible proportion of monochlorobenzene has been given as 4 – 5 % unreacted benzene, 73 % monochlorobenzene, and 22 – 23 % dichlorobenzene [727].

The production of monochlorobenzene from benzene, hydrogen chloride, and air has been described briefly in "Other Benzene Chlorination Processes", Section 8.1.3.

Only if special chlorinating agents are used benzene can be chlorinated to monochloro-

benzene without dichlorobenzene being formed simultaneously [757].

8.1.3.2. Dichlorobenzenes

Dichlorobenzenes are formed unavoidably in the production of monochlorobenzene. They arise as isomeric mixtures with a low content of the 1,3-isomer. A maximum dichlorobenzene concentration of 98 % is obtainable in a batch process in which 2 mol of chlorine is used per mole of benzene in the presence of ferric chloride and sulfur monochloride at mild reaction temperatures. The remainder of the product consists of mono- and trichlorobenzene. About 75 % 1,4-dichlorobenzene, 25 % 1,2-dichlorobenzene, and only 0.2 % 1,3-dichlorobenzene are obtained [781]. If the reaction is carried out with only ferric chloride as catalyst the maximum yield of dichlorobenzene is ca. 85 %. For further information about the effects of catalysts on isomer distribution, see Table 37.

If discontinuous separation of dichlorobenzene isomer mixtures is to be carried out, a distillation column with at least 60 practical plates is needed. At a reflux ratio of about 35 : 1, monochlorobenzene (if present), 1,3-dichlorobenzene – 1,4-dichlorobenzene mixture, and 1,2-dichlorobenzene are removed successively at the top of the column.

The intermediate fractions are recycled, their amounts being dependent on the separation capability of the column.

The 1,4-dichlorobenzene fraction is concentrated from the melt by crystallization – in one or more steps, depending on the impurities present. It can then be separated into pure 1,4-dichlorobenzene and the eutectic of 1,3-dichlorobenzene and 1,4-dichlorobenzene (14.7 % 1,4- and 85.3 % 1,3-dichlorobenzene). In industry (for economic reasons), the fraction is not cooled until the eutectic is obtained. Several different techniques are employed for the melt crystallization [782], e.g., the batch tube bundle crystallizer (a new development of this type is the Proabd-Raffineur), the semi-continuous falling film crystallizer from Sulzer-MWB, or the continuous purifiers from Brodie or Kureha Chem. Ind.

Small amounts of 1,2- and 1,3-dichlorobenzene that are present as contaminants in 1,4-dichlorobenzene can be removed chemically by exploiting the fact that they are considerably more reactive than 1,4-dichlorobenzene. In sulfonation [783, 784], bromination [785], chlorination [786], and other reactions first the 1,3- and then the 1,2-dichlorobenzene are consumed; this leaves pure 1,4-dichlorobenzene, which can be separated by some suitable means.

Pure *1,3-dichlorobenzene* can be obtained by working up the unavoidable 1,3-dichlorobenzene-containing mother liquor that results from *p*-dichlorobenzene crystallization. Combined crystallization/distillation processes, extractive distillation [787], or separation on zeolites [788] are feasible. However, there also exist synthetic processes which lead directly to 1,3-dichlorobenzene:

Chlorination of benzene in the gas phase at 200 – 500 °C in a tube or in molten salt [749, 750, 789].

Sandmeyer reaction of 3-chloroaniline [770, 790] or direct replacement of the amino group by chlorine [791].

Reaction of 1,3,5-trialkylbenzene with chlorine in the presence of I_2 or ferric chloride to give 2,6-dichloro-1,3,5-trialkylbenzene, the alkyl groups then being cleaved in the presence of aluminum chloride [792, 793].

Chlorination of 1,3-dinitrobenzene or chloronitrobenzene at a temperature above 200 °C [772 – 774].

Two-step chlorination of benzenesulfonyl chloride or diphenyl sulfone, with sulfur dioxide being eliminated [778, 779].

Cleavage of sulfur dioxide or carbon dioxide from the corresponding disulfonyl chlorides or dicarbonic acid chlorides at temperatures above 200/300 °C in the presence of catalysts [775, 777, 780].

Catalytic dechlorination of 1,2,4-trichlorobenzene with hydrogen in the vapor phase [794, 795].

Isomerization of 1,2- and 1,4-dichlorobenzene at a temperature of 150 – 200 °C in the presence of catalysts containing aluminum chloride [796 – 798]. The conversion to 1,3-dichlorobenzene is said to be quantitative if the isomerization is carried out in an antimony pentafluoride – hydrogen fluoride system at 20 °C [799].

Except in the case of the chlorinolysis of 1,3-dinitrobenzene, no details are available on the extent to which these processes are employed on an industrial scale.

8.1.3.3. Trichlorobenzenes

1,2,3-Trichlorobenzene and *1,2,4-trichlorobenzene* are formed in minor quantities in the production of monochlorobenzene and dichlorobenzene. Trichlorobenzenes become the main product if the chlorine input is increased to about 3 mol of chlorine per mole of benzene.

The batch reaction of benzene with a 2.8-fold molar quantity of chlorine in the presence of fer-ric chloride at temperatures increasing to 100 °C gives a chlorination mixture consisting of 26 % 1,4-dichlorobenzene, 4.5 % 1,2-dichlorobenzene, 48 % 1,2,4-trichlorobenzene, 8 % 1,2,3-trichlorobenzene, 8 % 1,2,3,4-tetrachlorobenzene, 5.5 % 1,2,4,5-tetrachlorobenzene, and less than 1 % pentachlorobenzene. The proportion of 1,4-dichlorobenzene, and thus also of 1,2,4-trichlorobenzene, can be raised by adding sulfur compounds as cocatalysts. After the chlorination mixture has been neutralized it can be separated by fractional distillation provided a column with more than 60 practical plates is used.

1,2,4-Trichlorobenzene can be obtained more directly by starting with pure 1,4-dichlorobenzene. Because of the directing influence of the chlorine substituents present (see Section 8.1.2), only 1,2,4-trichlorobenzene and (to an extent depending on the degree of conversion) higher chlorobenzenes are formed.

Another production method applicable to 1,2,4- and 1,2,3-trichlorobenzenes is based on the dehydrohalogenation of 1,2,3,4,5,6-hexachlorocyclohexane (stereoisomeric mixture), a byproduct of γ-hexachlorocyclohexane production (→ Insect Control). In the presence of aqueous alkali or alkaline earth solutions [800 – 803] or of ammonia [804], or directly through use of catalysts [805 – 811], hexachlorocyclohexane is converted at a temperature of 90 – 250 °C mainly to trichlorobenzene. The yield lies between 80 and 99 %, with the product mixture consisting of 70 – 85 % 1,2,4-trichlorobenzene and 13 – 30 % 1,2,3-trichlorobenzene.

It should be pointed out that polychlorinated dibenzofurans and dibenzodioxins are formed during the reaction.

References to further literature on the decomposition of benzene hexachloride will be found in [812].

1,3,5-Trichlorobenzene is not formed in the course of liquid phase chlorination of benzene (see Section 8.1.3). It can be produced, however, by Sandmeyer reaction on 3,5-dichloroaniline, by reaction of benzene-1,3,5-trisulfonic acid derivatives with phosgene [777], by vapor phase chlorination of 1,3-dichlorobenzene [813], and by chlorination of 3,5-dichloronitrobenzene [814] or 1-bromo-3,5-dichlorobenzene at 300 – 400 °C [815]. The proportion of the 1,3,5-isomer in a mixture can be raised by isomerization of the other trichlorobenzenes with aluminum chloride [816, 817] or by reacting tetrachlorobenzenes and higher chlorobenzenes with alkali metal amides [818]. Total isomerization is reported to occur when SbF_5 – HF is used as a catalyst system [799]. The three trichlorobenzenes can be separated by distillation and crystallization.

8.1.3.4. Tetrachlorobenzenes

1,2,3,4-Tetrachlorobenzene and *1,2,4,5-tetrachlorobenzene* are obtained by chlorination of benzene or of intermediate fractions obtained in the production of di- and trichlorobenzene. Antimony trichloride – iodine [819] is a catalyst combination that is particularly effective in giving 1,2,4,5-tetrachlorobenzene, the industrially preferred isomer (cf. Table 37). It is advisable to discontinue the chlorination before an appreciable amount of hexachlorobenzene has formed, since the low solubility of this compound makes separation of the chlorination mixture difficult. After the lower chlorinated benzenes have been removed by distillation, the tetrachlorobenzenes are separated from one another by crystallization with or without a solvent (the eutectic mixture consists of 12 % 1,2,4,5- and 88 % 1,2,3,4-tetrachlorobenzene and has a freezing point of 35 °C).

The 1,2,4,5-tetrachlorobenzene crystals which result are of high purity. The fraction containing 1,2,3,4-tetrachlorobenzene must be purified by distillation, however.

1,2,3,5-Tetrachlorobenzene which is not available by direct chlorination of benzene, can be obtained by chlorination of 1,3,5-trichlorobenzene.

8.1.3.5. Pentachlorobenzene

Pentachlorobenzene is of no economic significance. Chlorination of benzene gives a mixture of tetra-, penta-, and hexachlorobenzenes, from which pentachlorobenzene can be isolated by combined distillation and crystallization steps.

8.1.3.6. Hexachlorobenzene

Hexachlorobenzene can be produced by exhaustive chlorination of benzene, using chlorine in the presence of such catalysts as ferric chloride. The reaction is conducted at a temperature above 230 °C in the liquid or vapor phase [820 – 822]. Because much of the chlorine remains unreacted and because hexachlorobenzene sublimes, the off-gas from the process is passed through fresh starting material, in the course of which the chlorine is reacted and the sublimate washed out. The difficulties caused by sublimed hexachlorobenzene can be eliminated by batch chlorination with liquid chlorine in an autoclave [823].

Hexachlorobenzene can also be produced by decomposition of hexachlorocyclohexane in the presence of chlorine. The resulting lower chlorinated benzenes, mainly trichlorobenzenes, are converted to hexachlorobenzene without first needing to be isolated [824 – 826].

High yields of hexachlorobenzene are obtainable under very mild reaction conditions when chlorine-containing reagents such as chlorine monoxide [765] or chlorosulfuric acid/iodine are used [766].

8.1.4. Quality and Analysis

The usual separating processes employed in industry give chlorobenzenes of high purity. Those impurities which remain, as well as their concentrations, depend on the nature of the production process and the separating technique applied.

The method now used almost exclusively to determine the quantitative compositions of commercial chlorobenzenes is gas chromatography, an approach which is both rapid and reliable. Glass capillary and specially packed columns are used. For those chlorobenzenes with solidification points above 0 °C, the temperature of solidification can serve as a criterion of purity of the main component. Typical analyses of several commercial chlorobenzenes are as follows:

Monochlorobenzene > 99.9 %, < 0.02 % benzene, < 0.05 % dichlorobenzenes.

1,2-Dichlorobenzene. Technical grade: 70 – 85 % 1,2-dichlorobenzene, < 0.05 % chlorobenzene, < 0.5 % trichlorobenzene, remainder 1,4- and 1,3-dichlorobenzene

Pure grade: > 99.8 % 1,2-dichlorobenzene, < 0.05 % chlorobenzene, < 0.1 % trichlorobenzene, < 0.1 % 1,4-dichlorobenzene.

1,3-Dichlorobenzene 85 – 99 %, < 0.01 % chlorobenzene, < 0.1 % 1,2-dichlorobenzene, remainder 1,4-dichlorobenzene.

1,4-Dichlorobenzene > 99.8 %, < 0.05 % chlorobenzene and trichlorobenzene, < 0.1 % 1,2- and 1,3-dichlorobenzene, bulk density about 0.8 kg/L.

1,2,4-Trichlorobenzene > 99 %, < 0.5 % dichlorobenzenes, < 0.5 % 1,2,3-trichlorobenzene, < 0.5 % tetrachlorobenzenes.

1,2,4,5-Tetrachlorobenzene > 98 %, < 0.1 % trichlorobenzenes, < 2 % 1,2,3,4-tetrachlorobenzene, bulk density about 0.9 kg/L.

8.1.5. Storage and Transportation

The chlorobenzenes are all neutral, stable compounds which can be stored in the liquid state in steel vessels. The official regulations of various countries must be adhered with respect to the equipment of storage vessels, e.g., safety reservoir requirements, overflow prevention, and off-gas escape systems.

Chlorobenzenes that are liquid at ambient temperatures are shipped in drums, containers, or road/rail tankers. Solid compounds, such as 1,4-dichlorobenzene and 1,2,4,5-tetrachlorobenzene, can be transported in the molten state in heatable road/rail tankers or as granules or flakes in paper sacks and fiber drums. Steel containers are suitable. Any paper or fiber materials that are used must be impermeable to vapors arising from the product.

Liquid transfer must incorporate provisions for gas compensation, as well as protection against static charge. Chlorobenzene vapors form flammable mixtures with air. The compounds are regarded as potential water pollutants and must not be allowed to enter groundwater.

Spills must be collected (proper precautions being taken to safeguard the health of the workers involved) and burned in a suitable incinerator.

It should be noted that chlorobenzenes may decompose with the release of hydrogen chloride if they are exposed to severe heat.

Freight classifications are given in Table 40.

8.1.6. Uses

The chlorobenzenes, particularly mono-, 1,2-di-, and 1,2,4-trichlorobenzene, are widely used as solvents in chemical reactions and to dissolve such special materials as oils, waxes, resins, greases, and rubber. They are also employed in pesticide formulations (the highest consumption of monochlorobenzene in the United States).

Monochlorobenzene is nitrated in large quantities, the product subsequently being converted via such intermediates as nitrophenol, nitroanisole, nitrophenetole, chloroaniline, and phenylenediamine into dyes, crop protection products, pharmaceuticals, rubber chemicals, etc. The production of phenol, aniline, and DDT from monochlorobenzene, formerly carried out on a large scale, has been almost entirely discontinued due to the introduction of new processes and legislation forbidding the use of DDT.

1,2-Dichlorobenzene after conversion to 1,2-dichloro-4-nitrobenzene, is used mainly in the production of dyes and pesticides. It is also used to produce disinfectants and deodorants and on a small scale as a heat transfer fluid.

1,3-Dichlorobenzene is used in the production of various herbicides and insecticides. It is also important in the production of pharmaceuticals and dyes.

1,4-Dichlorobenzene is used mainly in the production of disinfectant blocks and room deodorants and as a moth control agent. After conversion into 2,5-dichloronitrobenzene, it finds application in the production of dyes. It is also used in the production of insecticides and, more recently, of polyphenylene-sulfide-based plastics, materials with excellent thermal stability [827, 828].

1,2,4-Trichlorobenzene is used as a dye carrier and (via 2,4,5-trichloronitrobenzene) in the production of dyes. Other uses are associated with textile auxiliaries and pesticide production (where 2,5-dichlorophenol serves as an intermediate). In the field of electrical engineering, it finds use as an additive for insulating and cooling fluids.

The hydrolysis of **1,2,4,5-tetrachlorobenzene** to 2,4,5-trichlorophenol, an intermediate for pesticides, has been almost entirely discontinued throughout the world due to the risk of formation of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD).

Pentachlorothiophenol, a mastication agent used in the rubber industry, is obtained from **hexachlorobenzene**. In the Federal Republic of Germany, the use of hexachlorobenzene as an active ingredient for pesticides has been prohibited since 1981.

Producers of chlorinated benzenes are: Anic SpA (Italy); Bayer AG (FRG); Hodogaya Chemical Co., Ltd. (Japan); Hoechst AG (FRG); Kureha Chemical Ind. Co., Ltd. (Japan); Mitsui Toatsu Chemicals, Inc. (Japan); Monsanto Chemical Corp. (USA); Nippon Kayaku Co., Ltd. (Japan); Produits Chimiques Ugine Kuhlmann SA (France); Rhône-Poulenc Chimie de Base (France); Standard Chlorine Chemical Co., Inc. (USA); Sumitomo Chemical Co., Ltd. (Japan).

8.2. Chlorinated Toluenes

8.2.1. Physical Properties

The chlorotoluenes occur in five chlorination stages, four of which have several isomers:

	Formula	Number of isomers	Table in which physical data are given
Monochlorotoluenes	C_7H_7Cl	3	41
Dichlorotoluenes	$C_7H_6Cl_2$	6	42
Trichlorotoluenes	$C_7H_5Cl_3$	6	43
Tetrachlorotoluenes	$C_7H_4Cl_4$	3	44
Pentachlorotoluene	$C_7H_3Cl_5$		44

Table 40. Freight classification

	GGVE/GGVS and RID/ADR		IMDG Code and IATA-DGR			US-DOT	
	class	number	class	UN No.	pack. group		
Monochlorobenzene	3	3	3.3	1134	II	UN 1134	Flamm. liquid
1,2-Dichlorobenzene	3	4	6.1	1591	III	ORM. A	UN 1591
1,4-Dichlorobenzene	–		6.1	1592	III	ORM. A	UN 1592
1,3-Dichlorobenzene	3	4	6.1	1591	III	ORM. A	UN 1591
Trichlorobenzene	6.1	62	6.1	2321	III	Poison B	UN 2810
Tetrachlorobenzene	–		6.1	2811	III	Poison B	UN 2811
Hexachlorobenzene	6.1	62	6.1	2729	III	Poison B	UN 2811

All of the *monochlorotoluenes* are colorless, mobile, flammable liquids with a faint odor, similar to that of benzene; they form binary azeotropes with many organic compounds [699, 700].

The *dichlorotoluenes*, apart from 3,5-dichlorotoluene, are liquid at room temperature, and they are likewise colorless and flammable. The *tri, tetra*-, and *pentachlorotoluenes* are colorless crystalline compounds.

The liquid chlorotoluenes are good solvents and are miscible with most organic solvents. All chlorotoluenes are insoluble in water. The polychlorotoluenes can be dissolved in many organic solvents, particularly at elevated temperatures.

8.2.2. Chemical Properties

The chlorotoluenes are neutral and stable compounds.

Chemical reactions may occur at unsubstituted positions on the aromatic ring (e.g. halogenation, nitration, or sulfonation), by replacement of the chlorine substituent (e.g, hydrolysis), and on the methyl group (e.g., side-chain chlorination or oxidation).

The influence of the methyl group of toluene leads to electrophilic substitution at positions 2 and 4, whereas in chlorotoluenes the directing influences of the methyl and chlorine groups overlap unpredictably. Introduction of a third substituent (e.g., $-Cl$, $-NO_2$, $-SO_3H$) into 2-chlorotoluene can lead to all four possible isomers being formed, though position 5 is occupied preferentially. In 4-chlorotoluene, position 2 is the most preferred position. The chlorine atoms in these compounds are bound very firmly to the aromatic ring and cannot be displaced except under forcing conditions. Nevertheless, the hydrolysis of monochlorotoluene with sodium hydroxide solution is possible at 350 – 400 °C and pressures up to 30 MPa (300 bar), the result being isomeric cresol mixtures.

Hydrogenation of chlorotoluenes over noble metal catalysts leads to dechlorination, just as in the case of chlorobenzenes.

Chlorine substituents can be exchanged for amino groups [829], but this reaction has no industrial application.

Under free-radical conditions at elevated temperatures, it is possible to replace sequentially the three hydrogen atoms of the methyl group by halogen, leading to ring-chlorinated benzyl-, benzal-, and benzotrihalides (see Chap. 9).

Oxidation of the methyl group leads to chlorinated benzaldehydes and benzoic acids [830, 831].

Catalytic ammonoxidation with oxygen and ammonia at 350 – 550 °C converts the methyl group into a nitrile group [832].

The polychlorinated toluenes are similar to one another in their chemical behavior. Their principal industrial use is in the manufacture of side-chain-halogenated products.

8.2.3. Production

Toluene Chlorination in the Liquid Phase. Monochlorotoluenes are produced on a large scale by reacting liquid toluene with gaseous chlorine at a moderate temperature and normal pressure in the presence of catalysts. Mixtures of isomers reflecting various chlorination stages are obtained. The chlorination conditions should be such as to give the highest possible yield of monochlorotoluene, because the dichlorotoluenes, all isomeric forms of which (with the exception of 3,5-dichlorotoluene)

Table 41. Physical properties of the monochlorotoluenes

	2-Chloro-toluene	3-Chloro-toluene	4-Chloro-toluene
			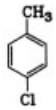
CAS reg. no.	[*95-49-8*]	[*108-41-8*]	[*106-43-4*]
M_r	126.59	126.59	126.59
Freezing point, °C	−35.59	−48.89	+7.6
Melting point, °C	−36.5	−47.8	+7.5
Boiling point at 101.3 kPa, °C	159.3	161.6	162.3
Density, g/cm³			
at 20 °C	1.0826	1.0722	1.0697
30 °C	1.0728	1.0625	1.0596
40 °C	1.0633	1.0530	1.0503
50 °C	1.0532	1.0433	1.0401
Viscosity, mPa s			
at 10 °C	1.188	1.0177	1.032
20 °C	1.022	0.877	0.892
60 °C	0.603	0.552	0.564
100 °C	0.439	0.392	0.397
Heat of fusion, J/g	79.99	82.95	102.46
Heat of vaporization, J/g			
at 100 °C	329.1		333.3
120 °C	319.5		323.6
140 °C	309.8		344.0
bp	299.6	327	303.5
Temperature corresponding to vapor pressure, °C			
0.13 kPa	5.4	4.8	5.5
0.67 kPa	30.6	30.3	31.0
1.33 kPa	43.2	43.3	43.8
2.67 kPa	56.9	57.4	57.8
5.33 kPa	72.0	73.0	73.5
8.00 kPa	81.0	83.2	83.3
13.3 kPa	94.7	96.3	96.6
26.7 kPa	115.0	116.6	117.1
53.3 kPa	137.1	139.7	139.8
101.3 kPa	159.3	161.6	162.3
Heat capacity, $J\,g^{-1}\,K^{-1}$	0.355		0.304
Heat of combustion at 18.8 °C, kJ/mol	−3747	−3749	−3754
Refractive index n_D^{20}	1.5267	1.5224	1.5209
Dielectric constant			
at 20 °C	4.45	5.55	6.08
58 °C	4.16	5.04	5.55
Critical temperature, °C	381.1		385.7
Flash point, °C	49	50	51
Ignition temperature, °C	> 500	> 500	> 500
Danger classification (VbF)	A II	A II	A II
Explosive limits in air			
lower, vol%	1.0		0.7
upper, vol%	12.6		12.2
TLV TWA (USA),			
ppm	50		
mg/m³	250		

result, cannot be separated economically. The relative proportions of the various chlorotoluenes obtained can be varied within wide limits by altering the reaction conditions and catalyst. With most catalyst systems, a high reaction temperature favors ortho and meta substitution as well as further chlorination. Reducing the temperature favors substitution in the para position and increases the total yield of monochlorotoluene. Because of the directing influence of the methyl group, the 3-chlorotoluene fraction of crude monochlorotoluene is limited to between 0.2 and 2 % depending on the catalyst. The influence of catalysts on the ratio of 2-chlorotoluene to 4-chlorotoluene is apparent from the data in Table 45. Selectivity is seen to be inversely proportional to the activity of the catalyst [833].

Solvents also, influence the isomer distribution. This fact is of little significance, however, because the toluene reactant in industrial chlorinations is not diluted with solvents.

The reaction is usually conducted at a temperature between 20 and 70 °C. Chlorination at temperatures below 20 °C is uneconomical because the rate of reaction is too low.

Toluene has a higher π-basicity than benzene, however, and it therefore, shows a substantially higher rate of chlorination than the latter. As a result, it can be chlorinated at relatively low temperatures (Table 46).

This fact permits efficient conversion of toluene to monochlorotoluenes without a substantial quantity of dichlorotoluenes also being formed (see Table 45).

At a moderate reaction temperature (below 100 °C), only traces of side-chain-chlorinated products are formed, provided activation by light is prevented and effective catalysts are used (ones for which amounts of no more than several tenths of a percent are required).

Both batch and continuous processes are used commercially, with the former being more selective with respect to particular stages of chlorination. For continuous processes, it is necessary to choose a toluene : chlorine starting ratio which gives a lower degree of toluene conversion to maximize the formation of monochlorotoluene.

Table 42. Physical properties of the dichlorotoluenes

	2,3-Dichlorotoluene	2,4-Dichlorotoluene	2,5-Dichlorotoluene	2,6-Dichlorotoluene	3,4-Dichlorotoluene	3,5-Dichlorotoluene
		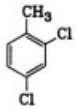	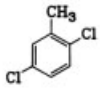		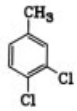	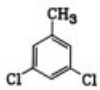
	[*32768-54-0*]	[*95-73-8*]	[*19398-61-9*]	[*118-69-4*]	[*95-75-0*]	[*25186-47-4*]
M_r	161.03	161.03	161.03	161.03	161.03	161.03
Freezing point, °C	+5.05	−13.35	+3.25	+2.6	−14.7	+25.1
Boiling point at 101.3 kPa, °C	208.1	201.1	201.8	200.6	209.0	202.4
Liquid density at 20 °C, g/cm³	1.266	1.25	1.254	1.266	1.254	
Refractive index n_D^{20}	1.551	1.548 (20 °C)	1.5449	1.5517	1.549 (22 °C)	1.5594
Flash point, °C	95	86	88	88	95	
Ignition temperature, °C	> 500	> 500	> 500	> 500	> 500	> 500
Explosive limits in air						
lower vol%		1.9				
upper vol%		4.5				
Temperature corresponding to vapor pressure, °C						
at 2.7 kPa	91.8	88.8	88.8	87.6	93.8	
8.0 kPa	120.2	115.6	115.9	114.7	121.6	
53.3 kPa	182.3	174.9	176.1	174.4	182.9	

Table 43. Physical properties of the trichlorotoluenes

	2,3,4-Trichlorotoluene	2,3,5-Trichlorotoluene	2,3,6-Trichlorotoluene	2,4,5-Trichlorotoluene	2,4,6-Trichlorotoluene	3,4,5-Trichlorotoluene
	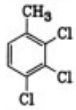	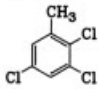	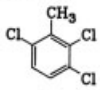	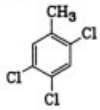	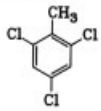	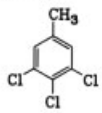
	[*7359-72-0*]	[*56961-86-5*]	[*2077-46-5*]	[*6639-30-1*]	[*23749-65-7*]	[*21472-86-6*]
M_r	195.48	195.48	195.48	195.48	195.48	195.48
Freezing point, °C	42.9	44.65	42.95	79.95	32.0	44.85
Boiling point at 101.3 kPa, °C	249.3	240.4	241.8	240.5	235.4	248.3
Liquid density at 100 °C, g/cm³	1.337	1.319	1.334	1.319	1.318	1.317

Table 44. Physical properties of tetra- and pentachlorotoluenes

	2,3,4,5-Tetrachlorotoluene	2,3,4,6-Tetrachlorotoluene	2,3,5,6-Tetrachlorotoluene	Pentachlorotoluene
	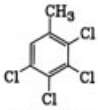	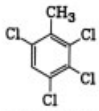	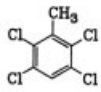	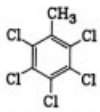
	[*1006-32-2*]	[*875-40-1*]	[*1006-31-1*]	[*877-11-2*]
M_r	229.92	229.92	229.92	264.37
Freezing point, °C	96.45	91.8	94.8	223.5
Boiling point at 101.3 kPa, °C	208.9	275.6	275.9	312
Liquid density at 100 °C, g/cm³	1.47	1.483	1.486	

Table 45. Influences of catalysts on the 2- : 4-chlorotoluene ratio

Catalyst	2- : 4-Chlorotoluene ratio	Dichlorotoluene in the chlorination mixture, %	Toluene conversion, %	Ref.
$TiCl_4$, $SnCl_4$, WCl_6 or $ZrCl_4$	3.3	1.5	~99	[834]
$[C_6H_5Si(OH)_2O]_4Sn$	2.2			[835]
$FeCl_3$	1.9	4.5	~75	[836]
$SbCl_3$ – diethylselenide	1.9		~50	[837]
$SbCl_3$	1.6			[727]
$AlCl_3$ – KCl	1.5	<1	~16	[838]
$SbCl_3$ – thioglycolic acid	1.2			[727]
$FeCl_3$ – S_2Cl_2	1.1	1.0	~99	[836]
Ferrocene – S_2Cl_2	1.06	0.2	~67	[839]
Lewis acids – thianthrene	0.91 – 1.1			[840]
$FeCl_3$ – diphenylselenide	0.93		~91	[837]
PtO_2	0.89	2	~96	[841]
$SbCl_3$ – phenoxathiin derivative	0.66 – 0.88	0.2 – 0.4	96 – 99	[842]
$SbCl_3$ – tetrachlorophenoxathiin	0.85 – 0.87			[843]
$SbCl_3$ – di- or tetrachlorothianthrene	0.7 – 0.9	0.1 – 0.2		[844]
Fe – polychlorothianthrene	0.76	1.2	~95	[845]

Table 46. Comparison of catalytic chlorination of benzene and toluene

Catalyst system	Ratio of the chlorination rates of toluene to benzene	Ref.
$FeCl_3$	16	[836]
$FeCl_3$ – S	200	[836]
Glacial acetic acid	345	[846]

The remarks in Section 8.1.3 on the continuous and batch chlorination of benzene apply correspondingly to the chlorination of toluene. Problems related to reactor design, materials, and dissipation of heat – 139 kJ is set free per mole of monochlorotoluene [847] – are also comparable. The ignition temperature of toluene in gaseous chlorine lies at 185 °C. The explosive limits of toluene in chlorine lie between 4 and 50 vol %. Conventional refinery toluene is sufficiently pure to serve as the starting material.

Traces of water that are entrained with technical chlorine and toluene have no influence on the composition of the product. They must not be so large as to reduce the efficiency of the catalyst, however.

One process for reducing the water content of the toluene below 30 ppm involves stripping with hydrogen chloride, and is described in a patent on the continuous production of monochlorotoluene [848].

As with the chlorobenzenes, the crude chlorotoluene mixture may be worked up by neutralization followed by separation in a distillation train. If the side-chain contains a detectable amount of chlorine, then this chlorine should be removed in order to prevent corrosion (e.g., pressure washing with aqueous alkaline solution may be carried out before the mixture enters the still).

Other Toluene Chlorination Processes. The following toluene chlorination processes are known in addition to liquid phase chlorination:

Chlorination in the vapor phase with chlorine [849]
Chlorination in the vapor or liquid phase with hydrogen chloride and air (oxychlorination)
Chlorination with chlorine-containing compounds
Electrolysis of toluene and hydrochloric acid

The oxidative chlorination of toluene with hydrogen chloride – air at a temperature of 150 –500 °C in the presence of cupric chloride as both catalyst and chlorinating agent gives mono- and dichlorotoluenes along with varying amounts of side-chain-chlorinated products [745, 850 – 852]. In the Gulf process, the toluene is chlorinated with hydrochloric acid – oxygen in the liquid phase in the presence of nitric acid and an additional strong acid at a temperature of 60 – 150 °C and pressures of 350 – 1000 kPa. By adding alkanes, e.g., *n*-octanes, the extent of side-chain attack can be reduced to 0.7 %. The addition of trialkylphenols is claimed to improve the selectivity so that only one chlorine atom reacts with a given aromatic nucleus [853]. Good monochlorotoluene yields are obtained at 50 °C if ferric chloride is used as a chlorinating agent in the presence of Lewis acids, such

as titanium chloride and aluminum chloride (*o*-chlorotoluene : *p*-chlorotoluene ratio 0.1 – 0.15) [729, 758, 854, 855]. Polymeric byproducts are also formed but the side-chains are said to be unattacked.

Sulfuryl chloride [761, 856] and chlorine monoxide [761] are other possible chlorinating agents.

Monochlorotoluene can also be manufactured electrochemically. Thus, toluene can be chlorinated electrochemically in methanol to which water and sodium, ammonium, or lithium chloride has been added. High yields of monochlorotoluene result without dichlorotoluene being formed [857].

Other Processes. Only a few of the isomeric chlorotoluenes can be produced economically by direct chlorination and separation of the reaction mixture. The other chlorotoluenes must be produced by indirect syntheses involving such steps as the replacement of amino substituents by chlorine [770, 791] or introduction of meta-directing sulfonic acid groups, which are then removed after chlorination on the ring [858, 859]. See the following Sections for additional information.

8.2.3.1. Monochlorotoluenes

Both 2- and 4-chlorotoluene are produced by chlorination of toluene, as described in Section 8.2.3. The economic importance of 4-chlorotoluene is presently greater than that of 2-chlorotoluene, a circumstance which has led to the development of new catalyst systems (see Table 45), which increase the output of the para isomer. The batch chlorination of toluene at 50 °C in the presence of ferric chloride (or of ferric chloride – sulfur monochloride at 40 °C the values for which are given in parentheses for comparison) gives at most 83 % (98 %) monochlorotoluene, with 52 % (52 %) of the product being *o*-chlorotoluene, 2 % (0.3 %) *m*-chlorotoluene, and 28 % (46 %) *p*-chlorotoluene, in addition to 7 % (1 %) unreacted toluene and 11 % (1 %) dichlorotoluene [836].

2-Chlorotoluene and 4-chlorotoluene are separated by fractional distillation. Continuous separation is possible with columns having more than 200 theoretical plates, leading to a pure 2-chlorotoluene distillate and a bottom product containing about 98 % 4-chlorotoluene. The boiling points of 3- and 4-chlorotoluene are so nearly alike that these two isomers cannot be separated by distillation. As in the production of pure 1,4-dichlorobenzene (see Section 8.1.3.2) the latter separation must be accomplished by crystallization from the melt. The eutectic mixture contains 77.5 % *m*-dichlorotoluene and 22.5 % *p*-dichlorotoluene and has a solidification point of − 63.5 °C. Consequently, crystallization from the melt must be conducted at low temperatures.

4-Chlorotoluene can also be separated from mixtures containing 2- and 3-chlorotoluene by adsorption on zeolites at 200 °C and 500 kPa [860].

3-Chlorotoluene can be produced from *m*-toluidine by a conventional Sandmeyer reaction [861] or by isomerization of 2-chlorotoluene on acid zeolites at 200 – 400 °C and 2000 – 4000 kPa, separation of the resulting mixture of 2- and 3-chlorotoluene being effected by distillation [860, 862].

8.2.3.2. Dichlorotoluenes

The isomers of dichlorotoluene have very similar chemical and physical properties, making it difficult to separate individual components from mixtures. Only 2,4-, 2,5-, and 3,4-dichlorotoluene can be produced economically by direct chlorination of the appropriate pure monochlorotoluene isomers.

2,3-Dichlorotoluene. In the chlorination of 2-chlorotoluene with ferric chloride, approximately 15 % of the product is 2,3-dichlorotoluene, accompanied by 2,4-, 2,5-, and 2,6-dichlorotoluene; the former can be separated by distillation. Its production by a Sandmeyer reaction on 3-amino-2-chlorotoluene is more economical, however.

2,4-Dichlorotoluene. Chlorination of 4-chlorotoluene in the presence of ring chlorination catalysts (e.g. chlorides of iron, antimony, or zirconium) leads to 2,4- and 3,4-dichlorotoluene in a ratio of ca. 4 : 1; the two can be separated by fractional distillation. Only traces of other dichlorotoluenes are formed [863, 864]. A low reaction temperature reduces the formation of trichlorotoluene. According to [856],

2-chlorotoluene is converted primarily into 2,4-dichlorotoluene when sulfuryl chloride is used as the chlorinating agent.

2,5-Dichlorotoluene. The proportion of 2,5-dichlorotoluene in the dichlorotoluene fraction can be raised to 60 % by chlorination of 2-chlorotoluene with sulfur compounds as catalysts or cocatalysts. The product is separated from the reaction mixture by crystallization [865, 866]. Chlorination with iodine catalysis is also possible [867]. In addition, 2,5-dichlorotoluene can be produced by a Sandmeyer reaction on either 2-amino-5-chlorotoluene or 2,5-diaminotoluene.

2,6-Dichlorotoluene. This isomer can be produced by chlorination of *p*-toluenesulfonyl chloride in the presence of antimonous chloride, followed by desulfonation [858]. Alternatively it can be made by chlorination of 4-*tert*-butyltoluene or 3,5-di-*tert*-butyltoluene with subsequent dealkylation [868, 869] or by means of a Sandmeyer reaction on 2-amino-6-chlorotoluene [870]. Distillatory separation of the 2,6-dichlorotoluene from a chlorination mixture (in which it constitutes about 30 % of the dichlorotoluene fraction) is prohibitively expensive. It is impossible at this time to assess the economics of its continuous separation by means of adsorption on faujasite-type zeolite [871].

3,4-Dichlorotoluene. The proportion of 3,4-dichlorotoluene in the product mixture from chlorination of 4-chlorotoluene can be raised to about 40 % if sulfur compounds are used as catalysts or cocatalysts. This increase is at the expense of the 2,4-dichlorotoluene [872].

3,5-Dichlorotoluene. This isomer is not formed in the direct chlorination of toluene. Its indirect synthesis is possible via 2-amino-3,5-dichlorotoluene [873]. 3,5-Dichlorotoluene can also be produced to an extent corresponding to its equilibrium ratio by isomerization of other dichlorotoluenes.

8.2.3.3. Trichlorotoluenes

When toluene is chlorinated with 3 mol chlorine per mole of toluene, four of the six possible trichlorotoluenes are formed in the following catalyst-dependent ratios (Table 47) [874].

2,3,4-Trichlorotoluene. Chlorination of 4-chlorotoluene in the presence of iron trichloride leads to a product in which 2,3,4-trichlorotoluene accounts for 22 % of the trichlorotoluene fraction. The compound can be separated from this mixture by distillation. It can also be produced from 3-amino-2,4-dichlorotoluene by using the Sandmeyer reaction [875].

2,3,6-Trichlorotoluene. The chlorination of 2-chlorotoluene with an iron powder catalyst gives a trichlorotoluene fraction containing 63 % 2,3,6-trichlorotoluene [876]. As 2,3,6- and 2,4,5-trichlorotoluene have similar boiling points, their separation by distillation is impractical. The concentration of 2,3,6-trichlorotoluene can be raised to 71 % by fractional crystallization in the absence of a solvent. This procedure gives a eutectic mixture containing 29 % 2,4,5-trichlorotoluene and having a solidification point of 21.5 °C.

Higher yields of 2,3,6-trichlorotoluene are possible from chlorination of 2,3-dichlorotoluene (75 %, in addition to 25 % 2,3,4-trichlorotoluene) or 2,6-dichlorotoluene (~ 99 %) [773]. 2,3,6-Trichlorotoluene can also be obtained by chlorination of *p*-toluenesulfonyl chloride or from 3-amino-2,6-dichlorotoluene by a Sandmeyer reaction [875].

2,4,5-Trichlorotoluene. Chlorination of 4-chlorotoluene at 20 – 50 °C gives > 80 % 2,4,5-trichlorotoluene if ferrous sulfide is used as a catalyst. The product is separated from the chlorination mixture by distillation [877].

2,4,6-Trichlorotoluene. The chlorination of 2,4-dichlorotoluene in the presence of ferric chloride gives a trichlorotoluene fraction containing 22 % of the 2,4,6-isomer, which can be isolated by fractional distillation.

8.2.3.4. Tetrachlorotoluenes

The three tetrachlorotoluene isomers are all formed when toluene is chlorinated (Table 48) [874].

2,3,4,5-Tetrachlorotoluene. The proportion of the 2,3,4,5-isomer in the tetrachlorotoluene fraction can be raised to 49 % by starting with 2,4,5-trichlorotoluene. Its separation by distillation is possible.

Table 47. Distribution of the isomeric trichlorotoluenes in the chlorination of toluene

Catalyst	2,3,4-Trichlorotoluene, %	2,3,6-Trichlorotoluene, %	2,4,5-Trichlorotoluene, %	2,4,6-Trichlorotoluene, %
Ferric chloride	15	46	35	4
Ferric chloride – sulfur monochloride	11	20	67	2

Table 48. Distribution of the isomeric tetrachlorotoluenes

Catalyst	2,3,4,5-Tetrachlorotoluene, %	2,3,4,6-Tetrachlorotoluene, %	2,3,5,6-Tetrachlorotoluene, %
Ferric chloride	20	44	36
Ferric chloride – sulfur	25	45	30

2,3,4,6-Tetrachlorotoluene. Yields of 51 % and 66 % of this isomer are obtainable by chlorination of 2,4,5-trichlorotoluene and 2,3,4-trichlorotoluene, respectively. 2,3,4,6-Tetrachlorotoluene is the exclusive product of chlorination of 2,4,6-trichlorotoluene [874].

2,3,5,6-Tetrachlorotoluene. This isomer is obtained by exhaustive chlorination of *p*-toluenesulfonyl chloride, followed by desulfonation [859].

The various tetrachlorotoluenes are also obtainable from the corresponding amino compounds by the Sandmeyer reaction.

8.2.3.5. Pentachlorotoluene

Exhaustive chlorination of toluene gives pentachlorotoluene. The reaction can be carried out with chlorine in either carbon tetrachloride or hexachlorobutadiene as solvent in the presence of iron powder and ferric chloride as catalysts [878], or with chlorine monoxide in carbon tetrachloride as solvent in the presence of an acid, e.g., sulfuric acid [765]. Sulfuryl chloride, used in the presence of sulfur monochloride and aluminum chloride catalysts, is also a suitable chlorinating agent, the method being that of O. Silberrad [761].

Further chlorination of pentachlorotoluene in the presence of a chlorination catalyst at a temperature above 350 °C leads to the formation of hexachlorobenzene and carbon tetrachloride [879].

8.2.4. Quality and Analysis

The industrially important chlorotoluenes can be produced such that they have a high degree of purity. No general agreement exists as to appropriate specifications. The following are typical analyses found for several chlorotoluenes:

2-Chlorotoluene	
Content	>99 %
toluene	< 0.1 %
4-chlorotoluene	< 0.9 %
3-chlorotoluene	< 0.01 %
4-Chlorotoluene	
Technical grade with	>98 %
3-chlorotoluene	< 1 %
2-chlorotoluene	< 0.5 %
dichlorotoluenes	< 0.5 %
Pure grade with	>99.5 %
3-chlorotoluene	< 0.2 %
2-chlorotoluene	< 0.2 %
dichlorotoluenes	< 0.1 %
2,4-Dichlorotoluene	
Content	>99 %
4-chlorotoluene	< 0.2 %
2,5-dichlorotoluene	< 0.3 %
2,6-dichlorotoluene	< 0.1 %
3,4-dichlorotoluene	< 0.4 %
3,4-Dichlorotoluene	
Content	>95 %
2,4-dichlorotoluene	< 5 %
trichlorotoluenes	< 0.5 %

Chlorotoluenes are preferably analyzed by gas chromatography. Packed columns are used to separate 2-, 3-, and 4-chlorotoluene. The more highly chlorinated toluenes can be separated by means of glass capillary columns.

8.2.5. Storage and Transportation

The mono- and dichlorotoluenes are stable, neutral liquids. They are shipped in drums, contain-

ers, and road or rail tankers. Steel is a suitable material for construction of containers.

For information on storage and handling, see Section 8.1.5.

Freight classifications are given in Table 49.

8.2.6. Uses

Isomeric mixtures of the monochlorotoluenes are hydrolyzed to cresol on a considerable scale. Chlorotoluenes are also used as solvents in reactions and to dissolve special products, e.g., dyes.

2-Chlorotoluene. 2-Chlorotoluene is a starting material in the production of 2-chlorobenzyl chloride, 2-chlorobenzaldehyde, 2-chlorobenzotrichloride, 2-chlorobenzoyl chloride and 2-chlorobenzoic acid, which are precursors for dyes, pharmaceuticals, optical brighteners, fungicides, and products of other types. 2-Chlorotoluene is also used in the production of dichlorotoluenes (chlorination), 3-chlorotoluene (isomerization), and *o*-chlorobenzonitrile (ammonoxidation).

4-Chlorotoluene. 4-Chlorotoluene is used mainly to produce *p*-chlorobenzotrichloride, from which is obtained *p*-chlorobenzotrifluoride, an important precursor of herbicides (e.g., trifluralin:α,α,α,-trifluoro-2,6-dinitro-*N,N*-dipropyl-*p*-toluidine). Other side-chain-chlorinated products or their derivatives are 4-chlorobenzyl chloride (for pharmaceuticals, rice herbicides, and pyrethrin insecticides), 4-chlorobenzaldehyde (for dyes and pharmaceuticals), 4-chlorobenzoyl chloride (for pharmaceuticals and peroxides), and 4-chlorobenzoic acid (for dyes). 4-Chlorotoluene is also a starting material in the synthesis of 2,4- and 3,4-dichlorotoluene and of 4-chlorobenzonitrile.

2,4-Dichlorotoluene. 2,4-Dichlorotoluene is used via its side-chain-chlorinated intermediates to produce fungicides, dyes, pharmaceuticals, preservatives, and peroxides (curing agents for silicones and polyesters).

2,6-Dichlorotoluene. 2,6-Dichlorotoluene is used to produce 2,6-dichlorobenzaldehyde, a dye precursor, and 2,6-dichlorobenzonitrile, a herbicide.

3,4-Dichlorotoluene. 3,4-Dichlorotoluene is used in small amounts in the production of 3,4-dichlorobenzyl chloride, 3,4-dichlorobenzaldehyde, 3,4-dichlorobenzotrichloride, and 3,4-dichlorobenzoic acid, from which disinfectants, crop protection products, and dyes are produced.

2,3,6-Trichlorotoluene. 2,3,6-Trichlorotoluene is used on a small scale, together with 2,4,5-trichlorotoluene, to produce 2,3,6-trichlorobenzoic acid, a herbicide precursor.

Producers of chlorinated toluenes are: Bayer AG (FRG); Enichem (Italy); Hodogaya Chemicals Co. Ltd. (Japan); Hoechst AG (FRG); Ihara Chemical Ind., Ltd. (Japan); Occidental Chemical Co., Ltd. (USA).

8.3. Chlorinated Biphenyls

Industrial use of the polychlorinated biphenyls first began in 1929 in the USA [880, 881]. The outstanding properties of these compounds, such as their high chemical and thermal stability, high dielectric constant, and the fact that they form only incombustible gases in an electric arc, made them appear ideally suited for use as insulating and cooling fluids for transformers and as dielectric impregnants for capacitors.

In subsequent years many other applications were found as well (see Section 8.3.5), particularly for isomeric mixtures containing two to six atoms of chlorine per mole of biphenyl.

In the mid-1960s, improved analytical methods revealed that polychlorinated biphenyls were accumulating in nature as a consequence of their extremely low rates of biological degradation (rates which decrease as the chlorine content rises). The compounds were detected in fresh water in all parts of the world, but also in many animals (e.g., birds, fish, and plankton). In the late 1970s, it was further discovered that at temperatures of 500 to 800 °C in the presence of oxygen, polychlorinated biphenyls can give rise to polychlorinated dibenzofurans and dibenzodioxins, including (although to a much smaller extent) the particularly toxic compound 2,3,7,8-tetrachlorodibenzodioxin [882 – 887].

In the meantime, all but a few of the well-known producers (see Section 8.3.5) discontinued the production of chlorinated biphenyls. Moreover, in many countries the production, sale, and use of polychlorinated biphenyls have

Table 49. Freight classification

	GGVE/GGVS and RID/ADR		IMDG Code and IATA-DGR			US-DOT
	class	number	class	UN No.	pack. group	
2- and 4-Chlorotoluene	3	3	3.3	2238	III	UN 1993 Flammable liquid
2,3-Dichlorotoluene	3	4		not restricted		
2,4-Dichlorotoluene	3	4		not restricted		
2,6-Dichlorotoluene	3	4		not restricted		
3,4-Dichlorotoluene	3	4		not restricted		

been restricted or entirely prohibited by legislation.

For many years, *o-, m-* and *p*-terphenyl mixtures were chlorinated and then used as plasticizers, flame retardants, and fillers in thermoplastic pattern and holding waxes. This application likewise has been substantially discontinued [888], particularly in view of the persistent nature of the compounds in question and their accumulation in the environment. Possible toxicological hazards are either unknown or have not been adequately investigated (Table 50).

Table 50. MAK and TLV values of chlorinated biphenyls

	MAK (FRG) mg/m^3	TLV (USA) mg/m^3
Chlorinated biphenyls (42 % Cl)	1 III B	1
Chlorinated biphenyls (54 % Cl)	0.5 III B	0.5

In view of the above, this article is devoted primarily to a review of recent patents covering methods for the disposal of polychlorinated biphenyls. For information regarding the production of these compounds and their specific physical and chemical properties, attention is directed to earlier surveys [882, 889, 890].

8.3.1. Physical and Chemical Properties

There are 209 possible chlorinated biphenyls. The mono- and dichlorobiphenyls [*27323-18-8*], [*25512-42-9*] are colorless crystalline compounds (the melting points of the pure isomers lie between 18 and 149 °C). When burned in air, they give rise to soot and hydrogen chloride.

The most important products are mixtures whose principal components are trichlorobiphenyl [*25323-68-6*], tetrachlorobiphenyl [*26914-33-0*], pentachlorobiphenyl [*25429-29-2*], or hexachlorobiphenyl [*26601-64-9*]. Such mixtures are liquid to viscous (pour points increase with chlorine content from − 22 to + 18 °C), and they are fire-resistant. Further chlorination gives soft to brittle thermoplastic waxes.

Chlorinated biphenyls are soluble in many organic solvents, particularly when heated, but are soluble in water only in the ppm range. Although they are chemically very stable, including to oxygen of the air, they can be hydrolyzed to oxybiphenyls under extreme conditions, e.g., with sodium hydroxide solution at 300 – 400 °C and under high pressure. Toxic polychlorodibenzofurans may be formed under these conditions.

The fact that the compounds may eliminate hydrogen chloride to a small extent at a high temperature explains why hydrogen chloride acceptors are often added to transformer fluids based on polychlorinated biphenyls.

The excellent electrical property data of polychlorinated biphenyls, such as high dielectric constant, low power factor, high resistivity, favorable dielectric loss factor, and high dielectric strength, have already been mentioned.

8.3.2. Disposal

Many products containing chlorinated biphenyls are still in use throughout the world, particularly in transformers, rectifiers, and capacitors with long service lives. Industry and national governments are now faced with the need to dispose of these products without causing additional pollution of the environment. Appropriate official regulations exist in many countries [888, 891 – 902].

Attention is drawn in the following survey to patents concerned with the removal of polychlorinated biphenyls from electrical devices and

with the disposal of these compounds. It is impossible to say which of the processes have actually reached maturity and which are already being used.

According to the present state of knowledge, polychlorinated biphenyls can be destroyed harmlessly by combustion at temperatures above 1000 °C and a residence time of 2 s, e.g., in a rotary burner equipped with a scrubbing tower for hydrogen chloride [888, 902 – 904]. Regulations in the Federal Republic of Germany specify a temperature of 1200 °C, a residence time of 0.2 s, and a residual oxygen content in the combustion gas of 6 % [897].

Removal of polychlorinated biphenyls from silicone- and hydrocarbon-based transformer fluids and heat transfer media is accomplished through the formation of a separable fraction rich in polychlorinated biphenyls [905], treatment with polyalkylene glycol and alkali metal hydroxide [906, 907], treatment with sodium naphthalenide [908], or heating with a sodium dispersion to 75 °C [909].

Polychlorinated biphenyls are removed from impregnated electrical parts by irradiation with microwaves, which causes gasification of the compounds [910], or by dry distillation at 500 – 1000 °C followed by addition of oxygen [911].

Destruction of polychlorinated biphenyls has been reported to be possible by the following methods: treatment with sodium naphthalenide in the presence of metallic sodium [912]; treatment at 145 °C with a dehalogenating reagent prepared from an alkali metal, polyethylene glycol, and oxygen [913];reaction with sulfur in the vapor phase [914];adsorption on paramagnetic or ferromagnetic material and subsequent irradiation with microwaves in the presence of oxygen [915]; and irradiation with light in aqueous solution in the presence of a catalyst [916].

Thermal decomposition of polychlorinated biphenyls occurs by pyrolysis under oxidative conditions (oxygen-enriched air) on molten alkali carbonates at 900 – 980 °C [917]; by use of a plasma burner at 3000 – 4000 °C [918]; by dissolution in kerosene, followed by combustion in air and introduction of the combustion gases into a special decomposition furnace [919]; or by evaporation with hydrogen as a carrier gas, followed by combustion in oxygen [920].

8.3.3. Analysis

The analytical methods most frequently used for detecting chlorinated biphenyls are capillary column gas chromatography coupled with mass spectrometry in the MID (Multiple Ion Detection) mode and capillary column gas chromatography with ECD (Electron Capture Detector). These methods are suitable for solution of even the most difficult problems. Clean-up steps are necessary when complex matrices are concerned, such as preliminary separation by column chromatography.

HPLC (High Pressure Liquid Chromatography) and infrared spectroscopy are applicable to a limited extent.

Summary polychlorinated biphenyl determinations are also possible, though not usual. These require either exhaustive chlorination and measurement of the decachlorobiphenyl content or else dechlorination and subsequent measurement of the biphenyl content.

For literature references on the subject of analysis, see [902, 921 – 927].

8.3.4. Storage and Transportation

At a normal temperature the commercially used polychlorinated biphenyls are liquid to viscous mixtures with a comparatively low vapor pressure (trichlorobiphenyl 6.5×10^{-5} kPa at 20 °C). Steel and aluminum are suitable as container materials. The storage and shipping of these compounds are subject to a variety of national regulations. Since these compounds accumulate in the environment they must be handled so that release cannot occur.

Exposure of polychlorinated biphenyls to fire may result in the formation of toxic chlorinated dibenzofurans and dibenzodioxins [882 – 887] and in the evolution of hydrogen chloride.

Classification of polychlorinated biphenyls are:

GGVE/GGVS and RID/ADR: Class 6.1, Number 23

IMDG-Code and IATA-DGR: Class 9, UN Number 2315, Packaging group II

US D.O.T.: ORM.E, UN Number 2315

8.3.5. Uses

Use of these compounds has fallen drastically [928] as a result of the extensive discontinuation of their production, voluntary renunciation of their application, and national restrictions. No details were available concerning products in which polychlorinated biphenyls are still used, nor concerning the scale of such use. The following list of important fields of application should be regarded as retrospective:

Cooling and insulating fluids for transformers
Dielectric impregnating agents for capacitors
Flame-retardant additives for resins and plastics used in the electrical industry
Alkali- and acid-resistant plasticizers for lacquers, plastics, adhesives, fillers and sealing compositions
Formulations for paints and printing inks
Water-repellent additives for surface coatings
Dye carriers for pressure-sensitive copying paper
Additives for thermally-stable lubricants and gear oils
Incombustible hydraulic fluids (particularly suitable for use in locations to which access is difficult, e.g., in mines)
Heat transfer fluids of high heat stability
Inert sealing fluids for vacuum pumps
Dust control agents for road construction

Mono- and dichlorobiphenyls have been used on a small scale as precursors for the corresponding oxybiphenyls.

Some registered trademarks are listed in Table 51. Most of the listed producers have discontinued production.

Table 51. Trade names of chlorinated biphenyls

Apirolio	Caffaro, Italy
Aroclor	Monsanto, USA, UK
Clophen	Bayer AG, FRG
Delor	Chemco, Czechoslovakia
Fenclor	Caffaro, Italy
Inerteen	Westinghouse, USA
Kanechlor	Kanegafuchi Chem. Co., Japan
Pyralene	Prodelec, France
Pyranol	Monsanto, UK
Pyroclor	Monsanto, USA
Sovtol	USSR

8.4. Chlorinated Naphthalenes

The first industrial applications of chlorinated naphthalenes took place at the beginning of the 20th century [929]. The compounds were used most extensively in the 1930s to 1950s, especially in cable and capacitor production, prompted by their dielectric, water-repellent, and flame-retardant properties.

More recently, most producers of polychlorinated naphthalenes have stopped their production, and output has been reduced drastically in all parts of the world. The reasons for this follow. First, connections have been established between highly chlorinated naphthalenes, especially pentachloronaphthalene and hexachloronaphthalene, and illness. Moreover, because of their high chemical and thermal stability, highly chlorinated naphthalenes are able to accumulate in the environment. Finally, new materials (polyesters and polycarbonate) have been introduced as substitutes for chlorinated naphthalenes in the capacitor and cable industries.

Monochloronaphthalenes, by contrast, are not considered to be problematic with regard to their effects on health and accumulation in the environment [930].

Official regulations relating to chlorinated naphthalenes differ considerably from country to country. In Japan, for example, polychlorinated naphthalenes are prohibited entirely. In the USA they may still be used without restriction, but changes in their production, importation, or use must be reported to the U.S. Environmental Protection Agency (EPA) [931], so that the effects of these changes on the environment may be monitored.

8.4.1. Physical Properties

Naphthalene has 75 chlorinated derivatives. To date, however, only a few have been synthesized and isolated in pure form. Only isomeric mixtures characterized according to their chlorine content are as a general rule commercially available. This situation arises because of the fact that the most important characteristics of the compounds are a function solely of their degree of chlorination, as a result of which there is little demand for the pure compounds. Moreover, precisely because the physical properties of the

various isomers are very similar, the cost of their separation is unrealistically high.

Except for 1-monochloronaphthalene, which is a liquid at room temperature, pure chlorinated naphthalenes are colorless, crystalline compounds. Mixtures of the compounds for industrial use have different degrees of chlorination, and their softening points lie considerably below the melting points of the pure components (generally between − 40 and + 190 °C). As the degree of chlorination increases, a transition occurs from liquids, via waxes, to hard solids, which causes the vapor pressures and water solubilities to fall. In contrast, the melting points, boiling points, and densities tend to rise, and any characteristics that are dependent on these properties become more pronounced. Mono- and dichloronaphthalenes are freely soluble in most organic solvents. Highly chlorinated naphthalenes are most soluble in chlorinated aliphatic and aromatic solvents and in petroleum naphthas.

Chlorinated naphthalenes have excellent dielectric properties. The tri- to hexachloronaphthalenes have dielectric constants of 4.5 – 5, a dissipation factor of 1×10^{-3} at 800 Hz and 20 °C, and a specific resistivity (100 V, 1 min) above 10^{14} Ω cm.

Chlorinated naphthalenes are compatible with many other commercial products, e.g., chlorinated paraffins, petroleum waxes, bitumen, various plasticizers (e.g., tricresyl phosphate), and polyisobutylene.

A selection of physical data obtained for pure monochloronaphthalenes and commercial mixtures is given in Tables 52 and 53 . For physical data on pure polychlorinated naphthalenes, see [932].

8.4.2. Chemical Properties

Reactions of the compounds may occur on the ring (electrophilic substitution), at the chlorine substituent (e.g., hydrolysis), or with dearomatization (chlorine addition).

1-Chloronaphthalene participates in electrophilic substitution reactions such as nitration [934], sulfonation, halogenation, and chloromethylation [935]. Reaction is especially favored at the para position relative to chlorine.

Hydrolysis with sodium hydroxide solution takes place at about 300 °C in the presence of copper catalysts [936] to give 1-naphthol and also 2-naphthol.

In the absence of a catalyst, additive chlorination of 1-chloronaphthalene yields pentachlorotrihydronaphthalene and hexachlorodihydronaphthalene [937].

Table 52. Physical data for monochloronaphthalenes*

	1-Chloronaphthalene [*90-13-1*]	2-Chloronaphthalene [*91-58-7*]
Melting point, °C	−2.3	95.5 – 60
Boiling point at 101.3 kPa, °C	260.2	258.6
Density, g/cm^3		
at 20 °C	1.194	1.178
80 °C	1.144	1.130
Temperature corresponding to vapor pressure , °C		
0.13 kPa	80.6	
0.67 kPa	104.6	
1.33 kPa	118.6	
2.67 kPa	134.4	
5.33 kPa	153.2	
8.00 kPa	165.6	161.2
13.30 kPa	180.4	
26.70 kPa	204.2	
53.30 kPa	230.8	
101.30 kPa	260.2	
Refractive index n_D^{20}	1.6326	
Viscosity at 25 °C, mPa s	2.94	
Flash point, °C	115	
Ignition temperature, °C	>500	

* A eutectic mixture of 75 % of 1-chloronaphthalene and 25 % of 2-chloronaphthalene has a solidification point of −17.5 °C.

The chemical and thermal stabilities of chlorinated naphthalenes increase with the number of chlorine substituents. Highly chlorinated naphthalenes withstand acids, caustic solutions, and oxidizing agents, even at elevated temperatures. An exception is concentrated nitric acid, which forms nitro derivatives relatively easily with polychlorinated naphthalenes and which at 90 °C oxidizes octachloronaphthalene to hexachloro-1,4-naphthoquinone and tetrachlorophthalic acid [938].

If naphthalene is chlorinated beyond the stage of octachloronaphthalene at a temperature exceeding 200 °C in the presence of ferric chloride catalyst, perchloroindane and carbon tetrachloride are formed. This phenomenon is a result of chlorine addition and subsequent ring constriction [939], decachlorodihydronaphthalene being an intermediate.

Table 53. Physical data for commercial chlorinated naphthalene mixtures [930, 932, 933]

	CAS reg. no.	Average chlorine content %	Softening point °C	Boiling point °C	Vapor pressure kPa	Density at 25 °C g/cm^3	Flash point °C	MAK (FRG) mg/m^3	TLV TWA (USA) mg/m^3
Monochloronaphthalene	[*25586-43-0*]	22	−25	250 – 260		1.2	135		
Mono-/dichloronaphthalene	[*28699-88-9*]	26	−33	250 – 290		1.22	130		
Tri-/tetrachloronaphthalene	[*1321-65-9*]	50	93	304 – 354	8×10^{-3}	1.58	200	5	5
Tetrachloronaphthalene	[*1335-88-2*]	52	115	312 – 360	10^{-5}	1.65	210		2
Tetra-/pentachloronaphthalene	[*1321-64-8*]	56	120	327 – 371		1.67	230	0.5	0.5
Penta-/hexachloronaphthalene	[*1335-87-1*]	62	137	343 – 384		1.78	250		0.2
Heptachloronaphthalene	[*32241-08-0*]								
Octachloronaphthalene	[*2234-13-1*]	70	185 – 197	440 (101.3 kPa) 246 (0.067 kPa)		2.0	>430		0.1

8.4.3. Production

The chlorination of naphthalene proceeds less rapidly than that of benzene or toluene. Consequently, chlorinated naphthalenes are produced batchwise in agitator vessels. Molten naphthalene, initially at 80 °C, is mixed with gaseous chlorine in the presence of ferric chloride or antimony pentachloride until the desired degree of chlorination has been reached. As the degree of chlorination increases, the reaction temperature must be raised to keep the mixture above its softening point.

The crude chlorination mixture is neutralized, e.g., with soda. The neutralizing agent is then separated and the crude product is fractionated by vacuum distillation.

Chlorine addition occurs if the chlorination is carried out in the absence of a catalyst, resulting in the unstable materials 1,2-dichloro-1,2-dihydro- and 1,2,3,4-tetrachloro-1,2,3,4-tetrahydronaphthalene.

Incomplete naphthalene conversion is accepted in the production of monochloronaphthalene in order to keep the proportion of polychlorinated naphthalenes small. The best possible result with ferric chloride as catalyst is 75 % monochloronaphthalene and 13 % polychlorinated naphthalenes in an isomer ratio of 90 – 94 % 1-chloro- and 6 – 10 % 2-chloronaphthalene. If peroxodisulfate and chloride ions are used as chlorinating agents, mainly monochlorinated derivatives are obtained [940]. Because of the large difference between their boiling points, the monochloronaphthalene isomer mixture is easily separated by distillation from the chlorination mixture, and it is of high purity. The isomers can be separated by crystallization [941]. Further information on the industrial chlorination of naphthalene will be found in the preceding edition of this encyclopedia [942].

The preparation of pure isomers representing the various chlorination stages is only of scientific interest. A review of the procedures is given in [932].

For information regarding the disposal of waste gas and chlorinated naphthalene wastes, see Section 8.5.

8.4.4. Quality and Analysis

The quality of chlorinated naphthalenes is monitored by wet analysis for chlorine content, as well as by gas chromatography employing glass capillary columns. Traces are determined by the same analytical methods that have been developed for polychlorinated biphenyls. For literature on this subject, see [932, 943 – 950].

One particular grade of monochloronaphthalene is used in wood preservatives and consists of

90 – 94 % 1-chloronaphthalene
6 – 10 % 2-chloronaphthalene
< 0.1 % dichloronaphthalene
< 0.01 % trichloronaphthalene and a chlorine content of 22 %.

8.4.5. Storage and Transportation

The chloronaphthalenes are stable, neutral substances. They can be stored in tanks as liquids, in which case adequate heating must be provided consistent with the melting point of the material. Steel containers are suitable, although stainless steels should be used for applications

in the electrical industry. Chlorinated naphthalenes that are liquid at ambient temperatures are shipped in drums, containers, or road tankers. Solid chloronaphthalenes are supplied as powder or flakes, either in fiber drums or in paper sacks. Existing national regulations must be complied with in connection with storage and transportation.

Polychlorinated naphthalenes are not expressly mentioned in the shipping regulations of either the EEC or the USA. Monochloronaphthalene has neither been allocated to a hazard class nor given a UN number.

8.4.6. Use

The use of chlorinated naphthalenes has diminished considerably during the last 30 years. Thus, except in special cases, chlorinated naphthalenes are no longer used in capacitors or electric cable coverings. Their use as lubricants has also been largely discontinued. Practice in individual countries varies, however. In the USA, for example, chlorinated naphthalenes are no longer used as wood preservatives. It is impossible to generalize concerning which of the following potential applications are permissible at present.

Monochloronaphthalenes: dye precursor; dye dispersant; fungicide and insecticide wood preservative; engine oil additive for dissolving sludges [951]; chemically and thermally stable sealing fluid; ingredient in special cleaning agents.

Polychlorinated naphthalenes: dielectric for impregnation of paper windings in automobile capacitors; insulating, waterproof, and flameretardant dipping and encapsulating compounds for special electrical parts; binder in the manufacture of ceramic elements for the electrical industry; paper coatings with water-repellent, flame-retardant, fungicidal, and insecticidal properties; plasticizers; electroplating stop-off compound.

Octachloronaphthalene: ingredient in the production of carbon elements by carbonization; additive for lubricants used under extreme conditions and for flame-retardant plastics.

Trade names for chlorinated naphthalenes. Most of the producers in the following list either have entirely discontinued the production of chlorinated naphthalenes, or else they have restricted their output and simultaneously reduced the content of highly chlorinated naphthalenes in their products.

Cerifal types, Caffaro, Italy
Clonacire types, Prodelec, France
Halowax types, Koppers Co., Inc., USA
Nibren Wax types, Bayer AG, FRG
Seekay wax types, ICI, Ltd., UK

8.5. Environmental Protection

In the production of chlorinated aromatics, organic compounds are contained in three different waste streams:

waste gas
wastewater
liquid or solid organic wastes

The correct disposal of these wastes results in no harm to the environment.

Waste Gas Treatment. It is necessary to distinguish between hydrogen chloride reaction gas from the chlorination process and substantially neutral waste gas, e.g. from distillation columns or storage containers. The second of these waste gas streams can be purified in an activated charcoal tower or incinerator. The hydrogen chloride reaction gas is processed in a complex manner to recover usable hydrochloric acid:

1) If the reaction gas contains chlorine, this is removed in a scrubbing tower (e.g., a bubble column [952]) containing an easily chlorinatable compound, preferably a raw material used in the chlorination process, and a chlorination catalyst.
2) In a second scrubbing tower, organic constituents of the reaction gas are washed out with a high-boiling solvent.
3) In addition to (or instead of) being passed through the second scrubbing tower, the hydrogen chloride gas is cooled to the lowest possible temperature in a cooler, in which further organic constituents are condensed out.
4) The hydrogen chloride is then absorbed in water in an adiabatic scrubber, from which it emerges as about 30 % hydrochloric acid with < 5 ppm of organically bound carbon.

Absorption in calcium chloride solution in the presence of calcium carbonate lumps has also been described [953]. In this case, organic compounds are removed with the escaping carbon dioxide, and a 33 % calcium chloride solution containing about 4 ppm of organic compounds remains.

Aqueous hydrochloric acid can be substantially freed from organic substances by extraction with dodecylbenzene [954].

The hydrogen chloride can also be liquefied in the absence of water and purified by distillation in a pressure column.

The pure hydrogen chloride and its aqueous solution are suitable for chemical processes. There are also techniques to recover the chlorine by electrolysis (e.g., the Hoechst-Uhde process) or oxidation (Shell-Deacon process or Kel Chlorine process).

Wastewater Treatment. The biological degradation rate of chlorinated aromatics decreases as their chlorine content increases. Only chlorinated aromatics with low degrees of chlorination are degradable in biological wastewater treatment plants, and then only if their concentration in the wastewater does not exceed certain levels. Therefore, wastewater streams containing chlorinated aromatics require preliminary purification. The following techniques [955] are suitable: stripping, extraction, and adsorption on activated carbon or polymeric resins [956].

Treatment of Wastes. These wastes may be distillation residues, useless fractions from separation processes, or industrial products containing chlorinated aromatics that are no longer suitable for use.

Normally these wastes are disposed of by incineration. In principle, it is also possible to convert chlorinated hydrocarbons into usable compounds by hydrodehalogenation or chlorinolysis. It is impossible to say whether these processes are already being used on an industrial scale.

Chlorine-containing aromatics are burned in special furnaces that provide reaction temperatures above 1000 °C and residence times of 1 – 2 s [904]. Only then is it certain that no polychlorodibenzodioxins are formed during combustion. This risk exists particularly with polychlorinated biphenyls (see Section 8.3). In an excess of oxygen, the chlorinated compounds are converted into hydrogen chloride, carbon dioxide and water. The hydrogen chloride is removed from the flue gas by water scrubbers.

Hydrodehalogenation is effected with hydrogen on palladium, platinum, or nickel catalysts at elevated temperature and high pressure. The nucleus-bound chlorine is substituted by hydrogen, and hydrogen chloride is thus formed. Mixtures of different chlorination stages down to the chlorine-free fundamental compound are obtained [706, 957 – 959].

Exhaustive chlorination of chlorinated hydrocarbons in the vapor phase at a temperature above 600 °C and pressure up to 200 bar splits the molecules, thus giving high yields of carbon tetrachloride [707, 708].

8.6. Economic Facts

Overall, it may be said that the use of chlorinated aromatic hydrocarbons is stagnant or declining worldwide (see Tab. 54).

Monochlorobenzene output in the USA fell from 275 100 t in 1960 to 115 700 t in 1982 [709] and increased in 1994 to 151 000 t. At the middle of 1984, Monsanto (at Muscatine, Iowa) discontinued the use of monochlorobenzene as a carrier for the herbicide Lasso [710]; this is expected to lead to a further drop in output.

As the production of monochloro-, 1,2-dichloro- and 1,4-dichlorobenzene is most economical in coproduction, it is becoming increasingly difficult to adapt the manufacturing conditions to the requirements of the market.

There are no output statistics for chlorinated toluenes. It is estimated that 30 000 t of 1- and 4-chlorotoluene was produced in Western Europe in 1983 [709].

The production of polychlorinated biphenyls has been almost completely discontinued throughout the world (cf. Section 8.3). Since the first industrial use of polychlorinated biphenyls at the beginning of the 1930s, about 1 million tons have been produced [712]; 40 % of this quantity is estimated to be still in use. Consumption reached its highest level in the 1960s.

The production of chlorinated naphthalenes has also fallen greatly, with a shift toward less highly chlorinated naphthalenes. Output in the

Table 54. Production of chlorobenzenes [709, 711] in 10^3 t/a

		Monochlorobenzene	1,2-Dichlorobenzene	1,4-Dichlorobenzene	Sum of all chlorobenzenes
USA	1970	220	30	32	
	1981	130	23	33	
	1988		11	42	
FRG	1970				127
	1981	97		34	138
Japan	1981	34	9	16	50

USA fell from 2270 tons in 1972 [713] to 320 tons in 1978; after the manufacture of chlorinated naphthalenes had been discontinued, the quantity imported into the USA settled at about 15 tons in 1980 and 1981 [714]. The output of the USA in 1972 consisted of 25 – 28 % of mono-/di-, 65 – 66 % of tri-/tetra-, and about 8 % of penta- to octanaphthalenes [713]. In the Federal Republic of Germany with an output of about 1000 t of chlorinated naphthalenes in 1972 [715], the production of polychlorinated naphthalenes has likewise been discontinued; only monochloronaphthalenes are now produced, output being of the order of 100 t/a.

9. Side-Chain Chlorinated Aromatic Hydrocarbons

Alkyl aromatics are somewhat unique in their behavior with respect to chlorination reactions.

The action of elemental chlorine can lead either to addition or substitution on the aromatic ring or it can cause substitution in the aliphatic side-chain, depending on the reaction conditions.

The side-chain chlorinated alkyl aromatics, particularly those based on toluene and xylene, have an exceptional place because of their role as chemical intermediates. Indeed, they are used in the manufacture of chemical products of almost all kinds, including dyes, plastics, pharmaceuticals, flavors and perfumes, pesticides, catalysts, inhibitors, and so forth.

Those side-chain-chlorinated alkyl aromatics which are of greatest importance in industrial chemistry are the toluene derivatives benzyl chloride, benzal chloride, and benzotrichloride.

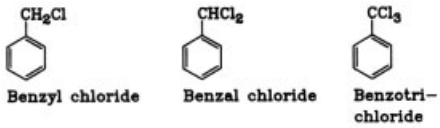

9.1. Benzyl Chloride

Benzyl chloride (chloromethylbenzene, α-chlorotoluene) [*100-44-7*] may be structurally the simplest side-chain chlorinated derivative of toluene, but economically it is the most important. Benzyl chloride is the starting material for a large number of industrial syntheses. The first preparation of it involved not the chlorination of toluene, however, but the reaction of benzyl alcohol with hydrochloric acid (S. Cannizzaro, 1853).

9.1.1. Physical Properties

Benzyl chloride is a colorless liquid which fumes in moist air. It has a pungent odor and is irritating to the mucous membranes and the eyes (i.e., it has a powerful lachrymatory effect).

M_r	126.58
bp at 101.3 kPa	179.4 °C
mp	− 39.2 °C
ϱ at 0 °C	1.1188 g/cm^3
10 °C	1.1081 g/cm^3
20 °C	1.1004 g/cm^3
30 °C	1.0870 g/cm^3
50 °C	1.072 g/cm^3
87 °C	1.037 g/cm^3
n_D^{20}	1.5389
Dynamic viscosity η at	
15 °C	1.501 mPa s
20 °C	1.38 mPa s
25 °C	1.289 mPa s
30 °C	1.175 mPa s

Surface tension σ at	
15 °C	38.43 mN/m
20 °C	37.80 mN/m
30 °C	36.63 mN/m
88 °C	29.15 mN/m
17 °C	19.5 mN/m
Specific heat at	
0 °C	178 J mol^{-1} K^{-1} (1403 J kg^{-1} K^{-1})
20 °C	181 J mol^{-1} K^{-1} (1432 J kg^{-1} K^{-1})
25 °C	183 J mol^{-1} K^{-1} (1444 J kg^{-1} K^{-1})
50 °C	189 J mol^{-1} K^{-1} (1495 J kg^{-1} K^{-1})
100 °C	212 J mol^{-1} K^{-1} (1675 J kg^{-1} K^{-1})
Heat of vaporization at 25 °C	50.1 kJ/mol (396 kJ/kg)
Heat of combustion at constant volume	3708 kJ/mol (29.29×10^3 kJ/kg)
Flash point	60 °C
Ignition temperature	585 °C
Explosive limits in air,	
lower	1.1 vol%
upper	14 vol%
Explosive limits in chlorine,	
lower	ca. 6 vol%
upper	ca. 60 vol%
Specific conductivity at 20 °C	1.5×10^{-8} S/cm
Vapor pressure at	
0 °C	0.025 kPa
10 °C	0.05 kPa
20 °C	0.12 kPa
30 °C	0.37 kPa
50 °C	0.99 kPa
100 °C	7.96 kPa
130 °C	23.40 kPa
179.4 °C	101.33 kPa

Numerous binary and ternary azeotropes containing benzyl chloride are known [970]. Examples are given in Table 55.

Table 55. Azeotropic mixtures with benzyl chloride

Component	Boiling point, °C	Benzyl chloride, wt %
Benzaldehyde	177.9	50
Hexanoic acid	178.7	95
Isovaleric acid	171.2	38
Valeric acid	175	25
Ethyl acetoacetate	175	35
Methyl acetoacetate	167	<80
1,3-Dichloro-2-propanol	168.9	57
2,3-Dichloro-2-propanol	171	40
Ethylene glycol	ca. 167	ca. 30

The solubility of benzyl chloride in water is 0.33 g/L at 4 °C, 0.49 g/L at 20 °C, and 0.55 g/L at 30 °C. Benzyl chloride is freely soluble in chloroform, acetone, acetic acid esters, diethyl ether, and ethyl alcohol. The solubility of chlorine in 100 g of benzyl chloride is 8.0 g at 30 °C, 5.4 g at 50 °C, and 2.1 g at 100 °C [971].

9.1.2. Chemical Properties

Benzyl chloride can serve as a starting point for the preparation of benzal chloride and benzotrichloride, both of which are accessible by side-chain chlorination. Nuclear chlorination, on the other hand, leads to chlorobenzyl chlorides. Oxidation with sodium dichromate gives sodium carbonate in aqueous solution benzaldehyde and benzoic acid.

Metals undergo a variety of reactions with benzyl chloride. For example, magnesium in ether gives benzyl magnesium chloride (Grignard), whereas copper powder or sodium gives 1,2-diphenylethane as the main product (Wurtz synthesis). The action of Friedel-Crafts catalysts such as $FeCl_3$, $AlCl_3$, and $ZnCl_2$ gives condensation products of the $(C_7H_6)_n$ type [972], but despite the fact that the degree of condensation can be controlled by changing the reaction conditions, these polymers have no commercial significance. If benzene or toluene is added to benzyl chloride in the presence of Friedel-Crafts catalysts, one obtains diphenylmethane or the isomeric benzyltoluenes, respectively.

The action of hydrogen sulfide and sulfides of the alkali metals leads to benzyl mercaptan and dibenzyl sulfide, respectively. Reactions with sodium salts of carboxylic acids produce the corresponding benzyl esters.

Hydrolysis with hot water is said to result in the formation of benzyl alcohol; nevertheless, this reaction has no industrial applications because the hydrochloric acid formed causes the re-formation of benzyl chloride from the product benzyl alcohol; it also catalyzes the formation of dibenzyl ether.

Hydrolysis in the presence of alkali does lead to benzyl alcohol, however.

Benzyl chloride reacts with sodium cyanide to give phenylacetonitrile (benzyl cyanide).

Reaction with ammonia or amines gives primary, secondary, tertiary amines, and quaternary ammonium salts.

The reaction with hexamethylenetetramine produces benzaldehyde (Sommelet reaction).

9.1.3. Production

Substitution of chlorine for hydrogen in the aliphatic side-chain occurs by way of a radical chain mechanism; by contrast, chlorine substitution on the aromatic ring takes place according to an electrophilic polar mechanism.

The steps in the radical chain process are as follows:

chain initiation	$Cl_2 + h\nu \rightarrow 2\ Cl\cdot$	(1)
chain propagation	$Cl\cdot + RH \rightarrow R\cdot + HCl$	(2)
	$R\cdot + Cl_2 \rightarrow RCl + Cl\cdot$	(3)
chain termination	$Cl\cdot + Cl\cdot \rightarrow Cl_2$	(4)
	$R\cdot + Cl\cdot \rightarrow RCl$	(5)
	$R\cdot + R\cdot \rightarrow RR$	(6)

This chlorination is highly exothermic (96 – 105 kJ/mol chlorine). In view of the high rate of chlorine radical formation and of hydrogen displacement, it is not surprising that, depending on the substrate and reaction conditions, radical chain lengths of 10^3 [973] to 10^6 [974] have been found. Because the mechanisms of side-chain chlorination and nuclear chlorination are fundamentally different, selectivity can be readily achieved. The prerequisites for high side-chain chlorination efficiency are as follows:

1) Achievement and maintenance of an optimal radical concentration
2) Elimination of components that might impart an electrophilic course to the reaction
3) Elimination of components capable of terminating the radical chains
4) Elimination of components conducive to other side reactions
5) In general, taking those precautions, which would encourage radical reactions or suppress electrophilic reactions

The discussion which follows considers each of these points separately.

1) Chlorine radical formation can be promoted by the addition of a radical-forming agent such as 2,2′-azobis(isobutyronitrile) (AIBN), benzoyl peroxide, or hexaphenylethane. Such compounds are consumed in the reaction, however, and thus have to be added repeatedly. For this reason, two other methods, singly or in combination, are of greater importance, particularly in industrial chlorinations: irradiation (ultraviolet light, *β-radiation) and the use of an elevated temperature (100* – 200 °C). In both cases, the effect is a result of excitation of chlorine molecules.
Irradiation is generally carried out with mercury vapor lamps that emit light in the wavelength range 300 – 500 nm, corresponding to the region in which the absorption bands of chlorine lie (hence, its yellowish-green color). The effects of varying the wavelength have been investigated in detail [975 – 977]. A solution containing 0.1 % chlorine absorbs 90 % of the available light energy over a distance of only 12 cm [978]. At a chlorine concentration of 3 – 4 mol/L, the corresponding distance is reduced to 0.01 cm [973]. Consequently, thorough mixing of the reactor contents is advantageous, as is the use of reactors of relatively small diameter. These considerations apply regardless of whether the sources of radiation are inside the reactor or external to it [979 – 981]. In the case of bubble column reactors, the height of the reactor is also important in view of the velocity of the rising gas bubbles, which is ca. 20 cm/s [982, 983].
2) Friedel-Crafts catalysts favor nuclear chlorination. For this reason, particular attention must be paid to ensure, for example, that the starting toluene is free of dissolved iron salts and that all rust particles have been filtered. By extension, steel, including stainless steel, is unsuitable as a reactor material. The preferred reactor materials are glass, enamel, and polytetrafluoroethylene.
Extensive purification of technical-grade starting materials is very expensive, however. For this reason, numerous suggestions have been made for additives that might inhibit the undesirable effects of heavy metals. Most such additives are assumed to complex with, e.g., iron cations. Some additives are reported to accelerate the side-chain chlorination, thereby inhibiting in a relative sense nuclear chlorination. The following are examples of additives which have been recommended at concentrations of 0.1 –2 %: pyridine and phenylpyridines [984], alkylene polyamines [985], hexamethylenetetramine [986], phosphoramides [987], acid amides, alkyl and dialkyl acid amides [988], ureas [989], phosphorus chlorides [990], alkyl phosphites and phosphines [991], phosphorus trichloride and trialkyl phosphates [992],

cyclic thioureas [993], lactams [994], red phosphorus [995], and aminoethanol [996].

3) Oxygen, a well-known radical scavenger, greatly reduces the consumption of chlorine; its presence is therefore undesirable. Toluene can be freed of oxygen, for example, puryng with an inert gas. The use of distilled chlorine, likewise, is an effective means of preventing oxygen from being present at significant concentrations [997 – 1001].
   It has also been reported, however, that even chlorine which contains up to 2 % oxygen is suitable for side-chain chlorination at relatively low temperatures (40 – 50 °C) provided that phosphorus trichloride (1 – 3 %) is present [1002].
4) The presence of water leads to the formation of aqueous hydrochloric acid (indicated by cloudiness) and possibly to the hydrolysis of the chlorinated toluene; in addition, hydrolysis products which arise may be further altered by chlorination. The hydrogen chloride that is inevitably formed during chlorination is thought to be responsible for the fact that nuclear chlorination can never be entirely suppressed [1003].
5) Various authors [998, 1001, 1004] have drawn attention to the adverse effects caused by excessive chlorine concentrations (especially nuclear chlorination, both addition and substitution). Besides ensuring that the level of irradiation is adequate for rapid reaction, it may be useful to introduce an inert gas (N_2, HCl) into the chlorine stream. Dividing the reactor into smaller units is apparently also advantageous, particularly in the case of a continuous process [980, 981, 1005, 1006]. Much attention has also been paid to the development of kinetic models, to the design of reactors, to the effects of agitators, to the distribution of radiation, and to the effects of inhibitors [1007 – 1009].

In the course of radical chlorination of toluene, all three hydrogen atoms of the side-chain are successively replaced by chlorine. As a result, mixtures of the three expected compounds are obtained: benzyl chloride, benzal chloride, and benzotrichloride.

$$C_6H_5CH_3 \xrightarrow[k_1]{Cl_2} C_6H_5CH_2Cl \xrightarrow[k_2]{Cl_2}$$
$$C_6H_5CHCl_2 \xrightarrow[k_3]{Cl_2} C_6H_5CCl_3$$

The change in the composition of the mixture as a function of the number of moles chlorine that have reacted is shown in Figure 34 [1010]. Similar results have been reported by other authors [975, 977, 1001, 1011].

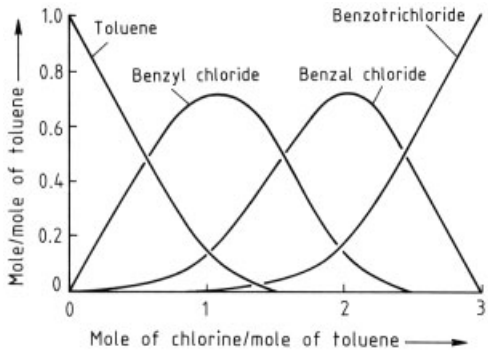

Figure 34. Progression of toluene chlorination [975, 977, 1001, 1010, 1011]

Extensive investigations [975] have shown that for batch operations, it is practically impossible to alter the shape of these curves, neither by changing the intensity and wavelength of the light, by adding radical-forming agents (e.g., peroxides), by changing the quantity of chlorine introduced in unit time, by using PCl_3 or PCl_5 or other catalysts, nor by other means.

Only the rate of the reaction and the nature of the observed side reactions can be shown to depend significantly on the reaction conditions. Little is known with certainty about the absolute reaction rates. One reason for the scarcity of information is the fact that the actual concentration of the dissolved chlorine is not accurately measurable. If the chlorine concentration is assumed to be constant, the reactions can be taken to be quasi first-order, leading to the following calculated ratios of rate constants at 100 °C:

$k_1 : k_2$	6.0 [975]	7.3 [999]	9.0 [1001]
$k_2 : k_3$	5.7 [975]	10.7 [999]	9.0 [1001]

where k_1, k_2, and k_3 are defined in the equations above.

The individual components of the mixture can be isolated in pure form by fractional distillation. To prevent the decomposition of ben-

zyl chloride during the distillation process, measures similar to those used to effect stabilization during storage and transportation (see Section 9.1.5) have been suggested [1012 – 1014].

Numerous attempts have been made to produce benzyl chloride that is as free as possible of the secondary products benzal chloride and benzotrichloride. One possibility is to restrict the chlorination to only 30 – 40 % of the toluene input and then to separate by distillation the resulting mixture, which still has a very high toluene content. Here, the principal disadvantage is the high cost of the distillation process.

According to [1015], pure benzyl chloride can be produced by chlorination in the vapor phase, provided the following equipment and procedure are used: The lower end of a packed column is provided with a bottom flask with a heater and an overflow. At the upper end is placed a condenser and a gas outlet. The feed point for fresh toluene and for toluene returning from the condenser is situated in the upper third of the column. The feed point for the chlorine (comprised of nozzles or frits) is located within the lower third. The temperature of the zone in which chlorination occurs is held at 125 °C during operation. This ensures that a temperature range is maintained within which benzyl chloride is a liquid whereas toluene is a gas. Under these conditions, benzyl chloride is rapidly removed from the chlorination section, flowing into the bottom flask; entrained toluene is returned from here to the chlorination section by evaporation. Hydrogen chloride formed during the chlorination, as well as toluene vapors, leave the column at its upper end. The toluene is liquefied in the condenser and returns to the chlorination section where it mixes with fresh toluene. The column is operated continuously. The bottom product is claimed to consist of 0.9 % toluene, 93.6 % benzyl chloride, and 5.5 % distillation residue. Several risks are inherent in this procedure. For example, the ignition temperature of toluene in chlorine gas is 185 °C. In addition, toluene –chlorine and benzyl chloride – chlorine mixtures are explosive over wide ranges (4 – 50 vol%, and 6 – 60 vol% respectively).

Monochlorination in the side-chain is said to be possible in the presence of alkyl or aryl sulfides or red phosphorus [995, 1016].

Today, most side-chain chlorination of toluene to produce benzyl chloride, together with benzal chloride (Fig. 34), is apparently carried out in a continuous fashion. The silver- or lead-lined reaction columns described in the past [1005] represent an obsolete technology since they are subject to considerable corrosion. For photochlorination, it is now customary to use reactors of enamel or glass, e.g., of the loop type [981]. The operating techniques have remained fundamentally unchanged, however.

Dibenzyl ether is formed as a byproduct in the alkaline hydrolysis of benzyl chloride to benzyl alcohol (see → Benzyl Alcohol). This ether can be re-converted to benzyl chloride by cleavage with hydrogen chloride at a temperature below 100 °C. Zinc chloride, iron chloride, and antimony chloride, are recommended as catalysts. Only a moderately high degree of conversion is achieved (60 – 70 %) [1017]. Higher yields are claimed to be obtained if the metal halogenide catalysts are replaced by pyridine, pyridine derivatives, or alkylbenzylammonium chloride [1018]. Another route to benzyl chloride is the chloromethylation of benzene, although this approach is of no commercial significance [1019].

Environmental Protection. Benzyl chloride, like chlorinated hydrocarbons in general, must be handled with special care. The threshold limit value (TLV) and the Federal Republic of Germany's MAK value are both 1 ppm (5 mg/m^3). Benzyl chloride has been allocated to Category III B of the MAK list (the category comprising substances for which there is reason to suspect carcinogenic potential). Therefore, special requirements must be met concerning the sealing of production equipment and ventilation of workrooms. Since benzyl chloride is a chlorinated hydrocarbon and homologue of benzene, regular medical inspection of affected personnel is required by law.

9.1.4. Quality Specifications and Analysis

Benzyl chloride is sold in two quality grades, identified as "benzyl chloride, pure" and "benzyl chloride, pure and stabilized". Their assay is in both cases > 99 %. Impurities include benzal chloride, toluene, chlorotoluene, chlorobenzene, and hydrogen chloride.

The most reliable analysis technique is gas chromatography, performed either with capillary or packed columns. The usual solid support in packed columns is Chromosorb AW-DMCS 80 – 100 mesh; recommended liquid phases include 4 % Silicone Fluid DC 550 and 4 % polyphenyl ether. Silicone resins are proven coating materials for capillary columns.

9.1.5. Storage and Transportation

As benzyl chloride is capable of reacting with heavy metals and their salts (Friedel-Crafts condensation reactions with the formation of HCl vapors), storage in enamel, glass, or lined vessels is essential. Suitable lining materials include bricks, lead, pure nickel, and stable synthetic resins. Drums with inserts of polyethylene or thick-walled polyethylene drums pigmented with graphite are suitable for transportation. Linings of lead, nickel, or special synthetic resins have proven to be suitable for tank cars and tank trucks. Many stabilizers have been proposed to make the storage and transportation of benzyl chloride safer. These act by neutralizing HCl and/or by forming complexes with heavy-metal ions. Examples are *N,N*-dimethylbenzylamine and *N,N*-diethylbenzylamine [1020]; pyridine and alkyl pyridines, quinoline, and bipyridyls (occasionally mixed with C_5–C_8 alcohols) [1013]; primary, secondary, and tertiary amines [1014, 1021, 1022]; phosphines [1023]; lactams [1024]; acid amides [1025]; ureas [1026]; and nitroalkanes [1027]. Aqueous sodium carbonate or sodium hydroxide solutions with a specific gravity identical with that of benzyl chloride were formerly used as stabilizers, but this practice has been largely discontinued. The emulsion these stabilizers produce lacks thermal stability and their presence makes it impossible to carry out reactions that require anhydrous benzyl chloride.

Legal Requirements. Benzyl chloride is a toxic chlorinated hydrocarbon and as such is subject to numerous regulations. The following regulations must be followed during its transport:

GGVS/ADR Class, 6.1 no. 15 b
GGVE/RID Class, 6.1 no. 15 b
GGV-See/IMDG Code Class 6.1;
UN no. 1738

Warning plate 6 (poison) must be displayed when benzyl chloride is transported on land. Primary label no. 6 (poison) and secondary label no. 8 (corrosive) are prescribed for marine transportation.

Benzyl chloride is additionally subject to the Arbeitsstoffverordnung of the Federal Republic of Germany and to the corresponding regulations of the European Community (EC compound no. 602–037–00–3).

9.1.6. Uses

Benzyl chloride is used mainly to produce plasticizers (e.g., benzyl butyl phthalate), benzyl alcohol, and phenylacetic acid via benzyl cyanide (used in the production of synthetic penicillin). On a smaller scale, it is used to produce quaternary ammonium salts (for disinfectants and phase-transfer catalysts), benzyl esters (benzyl benzoate and benzyl acetate for the flavors and perfumes industry), dyes of the triphenylmethane series, dibenzyl disulfide (antioxidant for lubricants), benzylphenol, and benzylamines.

9.2. Benzal Chloride

Benzal chloride (dichloromethylbenzene, α,α-dichlorotoluene, benzylidene chloride) [*98-87-3*] is produced exclusively by the side-chain chlorination of toluene. It was first synthesized in 1848 by A. Cahours, by using the reaction of PCl_5 with benzaldehyde. Almost the sole application of benzal chloride is in the production of benzaldehyde.

9.2.1. Physical Properties

Benzal chloride is a liquid which fumes in moist air and which has a pungent odor and a strong irritant effect on the mucous membranes and eyes.

M_r	161.03
bp at 101.3 kPa	205.2 °C
mp	− 16.2 °C
ϱ at 0 °C	1.2691 g/cm^3
20 °C	1.2536 g/cm^3
30 °C	1.2417 g/cm^3
57 °C	1.2122 g/cm^3
79 °C	1.1877 g/cm^3
135 °C	1.1257 g/cm^3
Vapor pressure at	
45.5 °C	0.6 kPa
75.0 °C	0.8 kPa
82.0 °C	1.3 kPa
89.5 °C	1.9 kPa
105 °C	4.0 kPa
118 °C	8.0 kPa
205.2 °C	101.3 kPa
n_D^{20}	1.5503
Surface tension σ at 20 °C	40.1 mN/m
at 100 °C	31.1 mN/m
Dynamic viscosity η at 20 °C	2.104 mPa s
at 50 °C	1.327 mPa s
Specific heat at 25 °C	222 J mol^{-1} K^{-1} (1377 J kg^{-1} K^{-1})
Heat of vaporization at 72 °C	50.4 kJ/mol (313.2 kJ/kg)
Flash point	93 °C
Ignition point	525 °C
Heat of combustion at constant pressure	3.852×10^3 kJ/mol (23.923×10^3 kJ/kg)
Explosive limits in air, lower	1.1 vol%
upper	11 vol%
Specific conductivity at 20 °C	3.4×10^{-9} S/cm

Benzal chloride is freely soluble in alcohol, ether, chloroform, and carbon tetrachloride, but only slightly soluble in water (0.05 g/L at 5 °C; 0.25 g/L at 39 °C). The solubility of chlorine in 100 g of benzal chloride is

6.2 g at 30 °C
4.3 g at 50 °C
1.5 g at 100 °C [971]

Several azeotropic mixtures are known of which benzal chloride is a component [970].

9.2.2. Chemical Properties

The action of chlorinating agents converts benzal chloride into benzotrichloride. In the presence of Lewis acids, the aromatic ring is chlorinated, with isomeric chlorobenzal chlorides being formed.

Hydrolysis under acid or alkaline conditions gives benzaldehyde.

Benzal chloride polymerizes in the presence of $AlCl_3$, $FeCl_3$, and similar compounds.

Metallic sodium converts benzal chloride into stilbene.

9.2.3. Production

Benzal chloride (together with benzyl chloride and benzotrichloride) is produced exclusively by side-chain chlorination of toluene. The preferred chlorination processes are those previously described under benzyl chloride. The pure compound is isolated by fractional distillation.

Environmental Protection. Benzal chloride is regarded as a toxic chlorinated hydrocarbon. Neither a TLV nor – in the Federal Republic of Germany – an MAK value has been established for benzal chloride. As a compound carrying a reasonable potential of being carcinogenic, benzal chloride has been allocated to Category III B of the MAK list. For this reason, stringent requirements must be met in its handling, including the sealing of production equipment and the ventilation of workrooms. Regular medical inspection is required of personnel coming in contact with the compound (chlorinated hydrocarbon and benzene homologues).

9.2.4. Quality Specifications and Analysis

The normal commercial form is "benzal chloride, pure," with an assay of > 99 %. The main impurities are benzyl chloride and benzotrichloride.

Benzal chloride is analyzed by the same methods described for benzyl chloride (see p. 359).

9.2.5. Storage and Transportation

Stabilization like that employed with benzyl chloride is not absolutely necessary. It may be advisable under some conditions, however, such as storage or transportation in the tropics. Compounds used to stabilize benzyl chloride are also effective for stabilizing benzal chloride. Enameled steel, lead, and stainless steel are suitable materials for the construction of storage tanks. Stainless steel tanks or drums coated with baked enamel are suitable for transportation.

Legal Requirements. Since benzal chloride is a toxic chlorinated hydrocarbon, it is subject to the following official regulations:

GGVS/ADR Class 6.1; 17 b
GGVE/RID Class 6.1; 17 b
GGV-See/IMDG Code Class 6.1;
UN no. 1886

Label no. 6 (poison) must be displayed. Benzal chloride is subject to the Arbeitsstoff-Verordnung of the Federal Republic of Germany and to the corresponding EC Directive (EC compound no. 602–058–00–8).

9.2.6. Uses

Benzal chloride is used almost exclusively to produce benzaldehyde.

Benzal chloride is hydrolyzed in the presence of water at a temperature above 100 °C by alkaline [1028] or acidic [1029 – 1032] agents. Friedel-Crafts catalysts or amines [1033] are recommended as catalysts. The latter are even recommended for mixtures of benzyl chloride and benzal chloride, whereby it is claimed that the benzyl chloride remains unchanged and that only benzaldehyde is formed. This process is unlikely to be of commercial interest; because benzyl chloride and benzaldehyde have almost identical boiling points their separation by fractional distillation would be very costly.

9.3. Benzotrichloride

Exhaustive chlorination of the side-chain of toluene leads to benzotrichloride (trichloromethyl benzene, α,α,α-trichlorotoluene, phenyl chloroform) [*98-07-7*]. The compound was first synthesized in 1858 by L. Schischkoff and A. Rosing, using the reaction of PCl_5 with benzoyl chloride. Benzotrichloride is now produced on a large scale, since it serves as an important intermediate in the preparation of acid chlorides (benzoyl chloride), dyes, herbicides, pesticides, and other products.

9.3.1. Physical Properties

Benzotrichloride is a colorless liquid with a pungent odor and is irritating to the eyes and mucous membranes. It fumes in moist air.

M_r	195.48
bp at 101.3 kPa	220.7 °C
mp	− 4.5 °C
ϱ at 15 °C	1.3777 g/cm^3
20 °C	1.3734 g/cm^3
30 °C	1.3624 g/cm^3
50 °C	1.342 g/cm^3
Vapor pressure at	
85 °C	1.1 kPa
95 °C	1.9 kPa
111 °C	3.1 kPa
121.5 °C	5.2 kPa
147 °C	13.3 kPa
220.7 °C	101.3 kPa
n_D^0	1.5677
n_D^{20}	1.5581
Dynamic viscosity η at 20 °C	2.40 mPa s
50 °C	1.517 mPa s
Surface tension σ at 20 °C	39.3 mN/m
100 °C	30.6 mN/m
Specific heat at 25 °C	235 J mol^{-1} K^{-1} (1206 kJ kg^{-1} K^{-1})
52 °C	248 J mol^{-1} K^{-1} (1269 kJ kg^{-1} K^{-1})
75 °C	249 J mol^{-1} K^{-1} (1273 kJ kg^{-1} K^{-1})
100 °C	250 J mol^{-1} K^{-1} (1281 kJ kg^{-1} K^{-1})
Heat of vaporization at 80 °C	52 kJ/mol (266 kJ/kg)
at 130 °C	47.5 kJ/mol (243 kJ/kg)
Flash point	108 °C
Ignition temperature	420 °C
Specific conductivity at 20 °C	6×10^{-9} S/cm
Heat of combustion at constant pressure	3684 kJ/mol (18878 kJ/kg)
Explosive limits in air, lower	2.1 vol%
upper	6.5 vol%

Benzotrichloride is freely soluble in alcohol, ether, and chloroform. It is only slightly soluble in water (0.05 g/L at 5 °C, 0.25 g/L at 39 °C). The solubility of chlorine in 100 g of benzotrichloride is

5.1 g at 30 °C
3.4 g at 50 °C
1.3 g at 100 °C [971]

Several azeotropic mixtures are known in which benzotrichloride is a component [970].

9.3.2. Chemical Properties

Acid or alkaline hydrolysis of benzotrichloride leads to benzoic acid. Partial hydrolysis gives benzoyl chloride.

Its reaction with carboxylic acids results in the corresponding acid chlorides and benzoyl chloride.

Condensation of benzotrichloride with benzene in the presence of $FeCl_3$, $AlCl_3$, or $ZnCl_2$ leads to diphenyl- and triphenylmethane.

All three chlorine atoms can be replaced by fluorine when benzotrichloride is treated with hydrofluoric acid or fluorides [1034, 1035].

Ortho-esters of benzoic acid can be prepared by reacting benzotrichloride with anhydrous alcohols.

9.3.3. Production

Exhaustive chlorination of the side-chain of toluene can be carried out in a manner analogous to that described under benzyl chloride. Photochemical chlorination in particular is widely applied for benzotrichloride production. Nevertheless, in order to prevent excessive chlorination and the appearance of ring-chlorinated materials, it is advisable, in continuous processes, to distribute the reaction over a cascade of six to ten reactors. Doing so makes it possible to introduce the chlorine at precisely the level appropriate to the progress of the reaction and results in benzotrichloride containing only a small amount of benzal chloride [980, 981].

A continuously operated plant for the production of benzotrichloride is illustrated in Figure 35 [981]. Fresh toluene flows into the first of a cascade of ten reactors. For reasons related to the removal of waste gases, the reactors can be regarded as being divided into three groups. Reactors 2 – 10 receive carefully metered amounts of chlorine. The off-gas from reactors 5 – 10 is rich in chlorine because the material in these reactors has already reached a high degree of chlorination; therefore, this gas is recycled to reactors 2 and 3.

Similarly, the off-gas from reactors 2 to 4 is introduced into reactor 1, which contains the highest proportion of toluene, so that the final traces of chlorine are removed. The off-gas from reactor 1 is thus free of chlorine.

With the condition that the chlorine and toluene are accurately metered, this technique is claimed to give practically complete conversion of toluene to benzotrichloride, and also to give a waste gas free of chlorine, i.e., consisting of pure hydrogen chloride.

Kinetic investigations of the formation of benzotrichloride have been published on several occasions [975, 999]. The yield and speed of the reaction are raised not only by exclusion of O_2 [999], but also by the use of high Reynolds numbers (35 000 – 160 000) [1036] or catalytic quantities of ammonium chloride [1037]. The chlorination of methylbenzenes using the corresponding trichlorides as solvents is claimed to give a high yield of very pure products [1038]. According to [1039] the use of bromine in the production of benzotrichlorides increases the reaction rate and the yield. One additional manufacturing process for benzotrichloride is based on the chlorination of dibenzyl ether [1040], which is formed as a byproduct in the conversion of benzyl chloride to benzyl alcohol. This particular chlorination leads to a mixture of benzotrichloride and benzoyl chloride, which can be worked in the usual way to give pure benzoyl chloride. Indirectly, this serves as a way to improve the economics of benzyl alcohol production.

Environmental Protection. Benzotrichloride is regarded as toxic. Neither a TLV nor – in the Federal Republic of Germany – an MAK value has been established for it. Benzotrichloride has been allocated to Category III B of the MAK list (this category comprises substances reasonably suspected of having carcinogenic potential). Therefore, special requirements must be met concerning the sealing of production equipment and the ventilation of workrooms. As with other chlorinated hydrocarbons and homologues of benzene, regular medical inspection of personnel is necessary.

9.3.4. Quality Specifications and Analysis

Benzotrichloride is sold in two quality grades, known as "benzotrichloride, technical" and "benzotrichloride, pure". The corresponding assays are $> 95\,\%$ and $> 98\,\%$ respectively. Impurities include chlorotoluenes, benzyl chloride, chlorobenzyl chlorides, benzal chloride, chlorobenzal chlorides, and chlorobenzotrichlorides.

Gas chromatography is the preferred method of analysis. The procedure is analogous to that used for benzyl chloride (see Section 9.1.4).

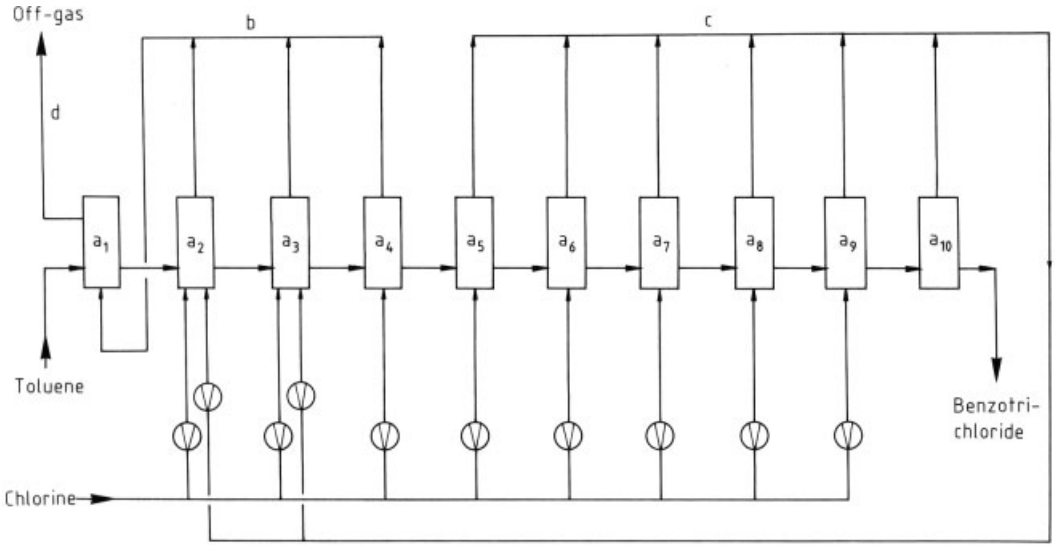

Figure 35. Continuous process for the manufacture of benzotrichloride [981] $a_1 - a_{10}$) Reactor cascade; b) Off-gas group 2, chlorine-containing; c) Off-gas group 3, high chlorine content; d) Off-gas group 1, chlorine-free

9.3.5. Storage and Transportation

Stabilization is unnecessary for storage purposes. Enameled, lead-lined, and stainless steel vessels are suitable for storage. Stainless steel tanks and drums coated with baked enamel are suitable for transportation.

Legal Requirements. Being a corrosive chlorinated hydrocarbon, benzotrichloride is subject to various regulations:

ADR/GGVS: Class 8, no. 66 b
RID/GGVE: Class 8, no. 66 b
GGV-See/IMDG Code: Class 6.1,
UN no. 2226

Label 8 (corrosive) must be displayed. Benzotrichloride is additionally subject to the Verordnung über gefährliche Arbeitsstoffe of the Federal Republic of Germany and to the corresponding directive of the European Community (EC compound no. 602–038–00–9).

9.3.6. Uses

Benzotrichloride is used mainly to produce benzoyl chloride, for which purpose it is either partially hydrolyzed with water or else reacted with benzoic acid. It is also of some significance in the production of pesticides (through transformation into benzotrifluoride), ultraviolet stabilizers, and dyes.

9.4. Side-Chain Chlorinated Xylenes

The side-chain-chlorinated xylenes play a less important role in the chemical industry than the corresponding toluene derivatives. In addition, substantial interest has been shown in only a few of the altogether 27 theoretically possible chloroxylenes, particularly the α-monochloro, α,α'-dichloro, and most notably, the $\alpha,\alpha,\alpha,\alpha',\alpha',\alpha'$-hexachloro derivatives.

9.4.1. Physical and Chemical Properties

A selection of chlorinated xylenes is listed in Table 56, together with certain physical data. The side-chain chlorinated xylenes are very similar in their chemical properties to the corresponding toluenes. They can therefore be made to undergo the same kinds of reactions as the latter. Thus, the hexachloroxylenes (*m*-, *p*-) are important in the production of carboxylic acid chlorides, and the α,α'-dichloroxylenes serve as sources of various bifunctional xylenes.

Table 56. Physical data of chlorinated xylenes

Compound	Formula	M_r [CAS No.]	*mp*, °C	*bp*, °C (kPa)	*ϱ*, g/cm³	n_D
α-Chloro-*o*-xylene 1-chloromethyl-2-methylbenzene	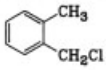	140.61 [*552-45-4*]		190 (101.3) 90 (2.7) 80 (1.6)	1.083 (21 °C)	1.5410 (20 °C)
α,α′-Dichloro-*o*-xylene 1,2-bis(chloromethyl)benzene	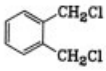	175.06 [*612-12-4*]	55	241 (101.3)	1.393	
α,α,α′-Trichloro-*o*-xylene 1-chloromethyl-2-dichloromethyl-benzene	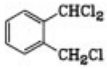	209.51 [*30293-58-4*]		139–141 (2.3)		
α,α,α′,α′-Tetrachloro-*o*-xylene 1,2-bis(dichloromethyl)benzene	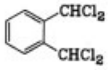	243.96 [*25641-99-0*]	90			
α,α,α,α′,α′-Pentachloro-*o*-xylene 1-dichloromethyl-2-trichloro-methylbenzene	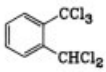	278.41 [*2741-57-3*]	53–55			
α-Chloro-*m*-xylene 1-chloromethyl-3-methylbenzene	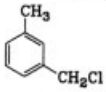	140.61 [*620-19-9*]		195–197 (101.3) 101–102 (4.0)		1.5345 (20 °C)
α,α′-Dichloro-*m*-xylene 1,3-bis(chloromethyl)benzene	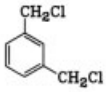	175.06 [*626-16-4*]	32–34	250 (101.3) 132 (2.1)		
α,α,α,α′,α′,α′-Hexachloro-*m*-xylene 1,3-bis(trichloromethyl)benzene		312.86 [*881-99-2*]	39–40	165–169 (1.6)		
α-Chloro-*p*-xylene 1-chloromethyl-4-methylbenzene	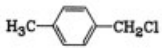	140.61 [*104-82-5*]	5	195 (101.3) 81 (2.0) 95 (2.7)	1.0512 (20 °C)	
α,α′-Dichloro-*p*-xylene 1,4-bis(chloromethyl)benzene		175.06 [*623-25-6*]	100.4	135 (2.1)		
α,α,α′-Trichloro-*p*-xylene 1-chloromethyl-4-dichloro-methylbenzene		209.51 [*7398-44-9*]	72	155–158 (3.6)		
α,α,α′,α′-Tetrachloro-*p*-xylene 1,4-bis(dichloromethyl)-benzene		243.96 [*7398-82-5*]	95			
α,α,α,α′,α′,α′-Hexachloro-*p*-xylene 1,4-bis(trichloromethyl)-benzene		312.86 [*68-36-0*]	111			

9.4.2. Production

The proven methods for the chlorination of toluene are basically suitable for the chlorination of xylene as well. Additives similar to those used in the chlorination of toluene are recommended to prevent nuclear chlorination [985 – 996]. Specialized additives include phosphoric acid esters together with sorbitol [1041, 1042] and a combination of boron trifluoride and ammonium chloride [1043]. Removal of air, moisture, and traces of metals by thorough purification of the chlorine and xylene feedstocks is also recommended [1044].

Chlorination in solvents, e.g., in carbon tetrachloride [988, 992, 1000, 1045] or hexachloroxylene [1038, 1046], has been described as particularly advantageous. Since in theory, there are up

to nine chlorinated derivatives of each of the xylene isomers, the course of the chlorination process is understandably very complex (Fig. 36) [1047].

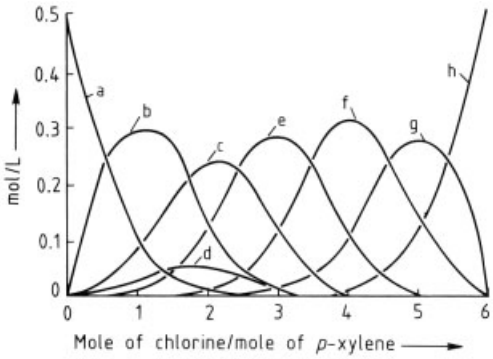

Figure 36. Progression of *p*-xylene chlorination [1047]
a) *p*-Xylene; b) α-Chloro-*p*-xylene; c) α,α'-Dichloro-*p*-xylene; d) α,α-Dichloro-*p*-xylene; e) α,α,α'-Trichloro-*p*-xylene; f) $\alpha,\alpha,\alpha',\alpha'$-Tetrachloro-*p*-xylene; g) $\alpha,\alpha,\alpha,\alpha',\alpha'$-Pentachloro-*p*-xylene; h) $\alpha,\alpha,\alpha,\alpha',\alpha',\alpha'$-Hexachloro-*p*-xylene

The corresponding kinetics have been investigated in detail [1047 – 1052].

Thus, in the manufacture of α-chloroxylene, the chlorination must be discontinued sufficiently early to ensure that only a small amount of dichloride is formed. The product is purified by distillation [1053].

p-α,α'-Dichloroxylene can be produced analogously, whereby any xylene and α-chloroxylene recovered at the distillation stage can be returned to the chlorination reactor [1054].

An alternative route to chloroxylenes involves the chloromethylation of toluene or benzyl chloride [1055, 1056]. This approach has the disadvantage, however, that it gives an isomer mixture, similar to that presumably formed by a double chloromethylation of benzene [1057]. This fact, together with the complexity entailed in a separation, makes its large-scale use less attractive. Isolation of various pure chloroxylenes is also possible (in some cases with high yields), not only by distillation, but also by direct crystallization from the reaction mixture [1058 – 1060].

The bis(trichloromethyl)benzenes are the chlorinated xylenes with the most commercial significance, and it is these whose manufacture has been investigated most thoroughly [992 – 994, 1002, 1006, 1038, 1044, 1046, 1049, 1061].

It is worth noting that in the case of *o*-xylene, the exhaustive chlorination of the side-chains leads only as far as $\alpha,\alpha,\alpha,\alpha',\alpha'$-pentachloro-*o*-xylene. Steric hindrance evidently makes the hexachloro stage inaccessible. The corresponding hexafluoro derivative is known, however.

9.4.3. Storage and Transportation

Individual chloroxylenes are not subject to special regulations. The relevant regulations concerning the handling of chlorinated hydrocarbons should be appropriately applied, however.

The same is true for transportation, where, depending on the properties of the compound concerned, allocation to existing hazard categories (assimilation) is necessary.

9.4.4. Uses

In terms of output quantity the *m*- and *p*-hexachloroxylenes are the most important side-chain chlorinated xylenes. These find application particularly in the production of isophthaloyl chloride and terephthaloyl chloride, important starting materials for polyester synthesis.

The α,α'-dichloroxylenes have been used together with diamines or glycols, bisphenols, or even amino alcohols in the production of polymers.

9.5. Ring-Chlorinated Derivatives

In comparison with the toluene and xylene derivatives that are chlorinated exclusively in the side-chain, those that are also chlorinated on the ring have achieved considerably less industrial importance.

Normally, such products are made from toluenes or xylenes whose rings already bear chlorine. These are then subjected to further chlorination under the conditions described above, thereby being converted into the desired derivatives. If products are desired in which all ring positions are chlorinated, it is often possible to chlorinate both the ring and the side-chains without purification of intermediates [1062].

Table 57. Physical data of chlorinated toluenes

Compound	Formula	M_r [CAS No.]	*mp*, °C	*bp*, °C (kPa)	ϱ, g/cm³	n_D
2-Chlorobenzyl chloride	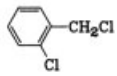	161.04 [*611-19-8*]		214 (101.3) 109 (3.4)	1.2743 (20 °C)	1.5592 (20 °C)
3-Chlorobenzyl chloride	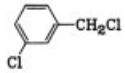	161.04 [*620-20-2*]		215 (100.4) 111 (3.4)	1.2695 (15 °C)	
4-Chlorobenzyl chloride		161.04 [*104-83-6*]	31	217 (102.9) 114 (3.3) 102 (2.1)	1.241 (40 °C)	1.5500 (40 °C)
2,6-Dichlorobenzyl chloride	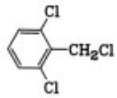	195.49 [*2014-83-7*]	39–40	88 (0.4)		
3,5-Dichlorobenzyl chloride	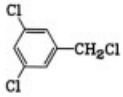	195.49 [*56961-85-4*]	36			
2,3,4,5,6-Pentachlorobenzyl chloride	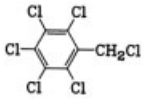	298.84 [*2136-78-9*]	103			
2-Chlorobenzal chloride	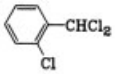	195.49 [*88-66-4*]		226–228 (99.3) 100 (1.3)	1.399 (15 °C)	1.5670 (16 °C)
3-Chlorobenzal chloride	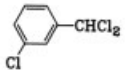	195.49 [*15145-69-4*]		235–237 (98.4) 105 (1.5)		
4-Chlorobenzal chloride	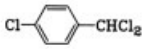	195.49 [*13940-94-8*]		236 (100.6) 108 (1.3)		
2,5-Dichlorobenzal chloride	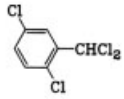	229.94 [*56961-83-2*]	42	325 (101.3)		
2,6-Dichlorobenzal chloride	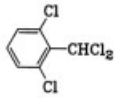	229.94 [*81-19-6*]		250 (101.3) 124–126 (2.1)		
2,3,4,5,6-Pentachlorobenzal chloride	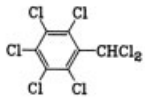	333.29 [*2136-95-0*]	117	199 (1.7)		
2-Chlorobenzotrichloride	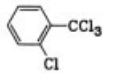	229.94 [*2136-89-2*]	30	260 (101.3) 129.5 (1.7)		
4-Chlorobenzotrichloride	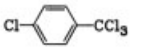	229.94 [*5216-25-1*]		108–112 (1.3)	1.4947 (30 °C)	1.5690 (30 °C)
2,4-Dichlorobenzotrichloride	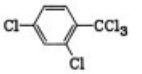	264.93 [*13014-18-1*]	48	147–151 (1.6) 155–159 (2.7)		

The uses of the ring-chlorinated compounds correspond to those of the parent series. A selection of such ring-chlorinated derivatives is compiled in Table 57.

9.6. Economic Aspects

Production capacities for the toluene derivatives discussed above were estimated to have been as follows in 1984 (Table 58):

Table 58. Capacities for chlorinated toluenes, in t/a

	Europe	World
Benzyl chloride	80 000	160 000
Benzal chloride	15 000	30 000
Benzotrichloride	30 000	60 000

It is not really possible to determine the extent of utilization of these capacities, since many companies produce the products for their further own use. It is likely, however, that ca. 60 % of the estimated capacity was utilized in 1984.

The 1984 price of benzyl chloride was ca. 0.90 $/kg; that of benzotrichloride was ca. 1.30 $/kg.

10. Toxicology and Occupational Health

10.1. Aliphatic Chlorinated Hydrocarbons

In this Section, chlorinated methanes, ethanes, ethylenes, propanes, and propenes of major commercial importance are discussed. A few other substances found as minor products, research chemicals, contaminants, or unwanted products are included when data are extensive or if they present an unusual or high toxicity.

The diversity of the toxic properties of chlorinated hydrocarbons is often inadequately appreciated, despite decades of use and studies proving great differences. They have all been mistakenly categorized generically as hepatotoxic, although large differences in their ability to injure the liver exist. Most of the compounds discussed are rather volatile and have a low potential for bioconcentration; nevertheless, the solvents and monomers have been incorrectly grouped with persistent chlorinated pesticides and other nonvolatile chlorinated materials.

While the substances discussed in this section show some common toxicological, chemical, and physical properties, exceptions are so common that categorization must be avoided. It is, therefore, imperative to examine the toxicity of each specific substance. Fortunately, the most common chlorinated solvents have been exhaustively studied. However, new toxicological data are being generated and the current literature and regulations should always be consulted.

Reviews may quickly become incomplete, but a few are listed in the references [1063 – 1065]. They will be most valuable for information about effects on the skin or eyes, as well as effects of single or short, repeated exposures by ingestion or inhalation. Information on carcinogenesis, mutagenesis, and birth effects are presented, but these are active research areas and current literature must be consulted. The manufacturer or supplier of a substance is responsible for acquiring and distributing such information, and most manufacturers excercise that responsibility. Table 59 shows some data on the acute toxicity of the aliphatic chlorinated hydrocarbons.

Abbreviations and Definitions.Certain terms, names, and organizations appear in this section. Most of the abbreviations are explained in the front matter, others are defined in the text. For detailed information on general toxicology, see the corresponding articles in the B series.

Carcinogenesis, Mutagenesis, and Teratogenesis.In the current regulatory climate, mutagenic changes, reproductive effects, and particularly cancer are of prime concern. This is appropriate, but excessive concern about cancer has distracted attention from other concerns that may be of equal or more importance.

Many, if not most, chlorinated substances can be made to produce an increase of tumors in certain laboratory animals, particularly in organs in which toxicity is exhibited and high tumor rates exist normally, e.g., in the livers of mice. However, no chlorinated hydrocarbon except vinyl chloride has yet been shown to have increased cancer in human populations. Many epidemiological studies lack power due to small populations and short duration. However, a number of studies are of adequate size, duration, and power

to demonstrate that cancer has not increased in the degree predicted from the studies in mice. Because of this inconsistency and a variety of other data, there has been considerable scientific discussion about interpreting animal studies in regard to human risk; see also [1066].

Table 59. Single-dose oral toxicity of common chlorinated C_1, C_2, C_3, and C_4 aliphatic hydrocarbons [1063]

species specified), mg/kg	LD_{50} (oral, rats; unless other of death[a]	Probably nature
Chloromethane (methyl chloride)	gas	–
Dichloromethane (methylene chloride)	2 000	A
Trichloromethane (chloroform)	2 000	A, LK
Tetrachloromethane (carbon tetrachloride)	3 000	A, LK
Monochloroethane (ethyl chloride)	gas	–
1,1-Dichloroethane (ethylidene dichloride)	>2 000[b]	–
1,2-Dichloroethane (ethylene dichloride)	700	LK
1,1,1-Trichloroethane (methyl chloroform)	10 000 – 12 000	A
1,1,2-Trichloroethane (vinyl trichloride)	100 – 200	LK
1,1,2,2-Tetrachloroethane (acetylene tetrachloride)	ca. 300 (dogs)	–
Pentachloroethane	≪1 750 (dogs)	–
Hexachloroethane	6 000	–
Monochloroethylene (vinyl chloride)	gas	–
1,1-Dichloroethylene (vinylidene chloride)	1 500	LK
1,2-Dichloroethylene (cis and trans)	1 000 – 2000	A
Trichloroethylene	4 900	A
Tetrachloroethylene (perchloroethylene)	2 000	A
Dichloroacetylene	–	–
2-Propyl chloride (isopropyl chloride)	>3 000 (guinea pigs)	A
1,2-Dichloropropane (propylene dichloride)	2 000	A
3-Chloropropene (allyl chloride)	450 – 700	LK
1,3-Dichloropropene	500 – 700	LK
2-Chloro-1,3-butadiene (chloroprene)	250	LK
Hexachlorobutadiene	200 – 350	LK

[a] A = Anesthesia, LK = Liver and kidney injury;
[b] Unpublished data, The Dow Chemical Company, Midland, Michigan, USA.

According to current scientific thought, certain substances, including the chlorinated hydrocarbons, increase cancer in certain organs (e.g., livers of mice) as a result of repeated stress and injury with subsequent increased cellular regeneration. Thus, preventing exposures that cause cellular changes (injury) should also prevent cancer. This nongenotoxic mechanism of induction is consistent with current human experience and other data related to most of the chlorinated hydrocarbons discussed herein except vinyl chloride.

Vinyl chloride appears to operate by a genetic mechanism and, although humans are much less responsive than rodents, the difference is quantitative and appears related to a lower rate of metabolism in humans. This indicates that exposure to all of these materials must be carefully controlled to avoid exposures that result in stress or injury. Close adherence to the occupational exposure limits (MAK or TLV) is recommended. Furthermore, good industrial hygiene practice requires that exposures to any substance be kept as low as reasonable and that careless operation should be prohibited regardless of whether the TLV or MAK is exceeded.

Reproductive effects do not appear to be of concern with any of the substances discussed in this Section, provided exposures are controlled to prevent injury to other organ systems of the mother during gestation. In other words, the reproductive system appears less sensitive than other systems [1067]. Likewise, mutagenic effects appear unlikely, based on the weight of the evidence from in vivo and in vitro studies.

Occupational Exposure Limits. Table 60 lists the 1985 TLV's and MAK's published by the American Conference of Governmental Industrial Hygienists (ACGIH) and the Deutsche Forschungsgemeinschaft (DFG), respectively [1068, 1069]. The definitions applied by these organizations must be understood in order to apply the values properly. For example, both organizations recommend that skin contact be limited if skin absorption is thought to influence the TLV or the significantly MAK. The reader must always consult the latest values published annually by these organizations and, further, must not assume that they are the legal standard. It is strongly recommended that supporting documentations be consulted when using the TLV's and MAK's.

10.1.1. Chloromethanes

Monochloromethane. Chloromethane [*74-87-3*], methyl chloride, is an odorless gas and, except for freezing the skin or eyes due to evaporation, inhalation is the only significant route

Table 60. Summary of TLV's and MAK's for common chlorinated C_1, C_2, C_3, and C_4 aliphatic hydrocarbons

	1985 ACGIH TLV		2001 DFG MAK	
	ppm	mg/m^3	ppm	mg/m^3
Chloromethane [74–87–3]	50	105	50 (3B)	100
Dichloromethane [75–09–2]	100	350	-(3A)	–(3A)
Trichloromethane [67–66–3]	10 (A2)	50 (A2)	0,5 (4)	2,5
Tetrachloromethane (skin) [56–23–5]	5 (A2)	30 (A2)	0,5 (4)	3,2
Monochloroethane [75–00–3]	1000	2600	– (3B)	– (3B)
1,1-Dichloroethane [75–34–3]	200	810	100	410
1,2-Dichloroethane [107–06–2]	10	40	– (2)	–
1,1,1-Trichloroethane [71–55–6]	350	1900	200	1100
1,1,2-Trichloroethane (skin) [79–00–5]	10	45	10 (3B)	55 (3B)
1,1,2,2-Tetrachloroethane (skin) [79–34–5]	1	7	1 (3 B)	7 (3 B)
Pentachloroethane [76–01–7]	–	–	5	42
Hexachloroethane [67–72–1]	10	100	1	9,8
Monochloroethylene (vinyl chloride) [75–01–4]	5 (A1a)	10 (A1a)	3 (2) (TRK) (1)	8 (5) (TRK) (1)
1,1-Dichloroethylene [75–35–4]	5	20	2 (3 B)	8 (3 B)
1,2-Dichloroethylene (cis and trans) [540–59–0]	200	790	200	800
Trichloroethylene [79–01–6]	50	270	– (1)	– (1)
Tetrachloroethylene [127–18–4]	50	335	– (3B)	– (3B)
Dichloroacetylene [7572–29–4]	C 0.1	C 0.4	– (2)	– (2)
1,2-Dichloropropane [78–87–5]	75	350	– (3B)	– (3B)
3-Chloropropene (allyl chloride) [107–05–1]	1	3	– (3B)	– (3B)
1,3-Dichloropropene (skin) (cis and trans, [542–75–6] and [126–99–8])	1	5	– (2)	– (2)
2-Chloro-1,3-butadiene (chloroprene) (skin)	10	36	– (2)	–
Hexachlorobutadiene (skin) [87–68–3]	0.02 (A2)	0.24 (A2)	– (3B)	– (3B)

A2 = 3 B = Suspected carcinogen; 2 = Carcinogen in animal experiments; A1 = Human carcinogen without gentoxic effects; C = Ceiling; TRK = Technical Guiding Concentration 3 ppm in existing facilities, 2 ppm in new facilities; skin = This designation is intended to suggest appropriate measures for the prevention of cutaneous absorption so that the threshold limit is not invalidated.

of exposure. It acts mainly on the central nervous system with well documented cases of excessive human exposure, leading to injury and even death [1063]. The symptoms of overexposure are similar to inebriation with alcohol (a shuffling gait, incoordination, disorientation, and change in personality), but last much longer, possibly permanent in severe exposures. According to experimental results, excessive exposure to methyl chloride was carcinogenic in mice and also affected the testes of male rats and fetuses of pregnant female rats [1070]. It is mutagenic in certain in vitro test systems. Available references indicate that methyl chloride may increase the rate of kidney tumors in mice in conjunction with repeated injury to this organ. The TLV and the MAK (1985) are both 50 ppm (105 mg/m^3).

Dichloromethane. Dichloromethane [*75-09-2*], methylene chloride, is the least toxic of the chlorinated methanes. It is moderate in toxicity by ingestion, but the liquid is quite painful to the eyes and skin, particularly if confined on the skin [1063 – 1065]. Absorption through the skin is probably of minor consequence if exposure is controlled to avoid irritation.

Inhalation is the major route of toxic exposure. The principal effects of exposure to high concentrations (greater than 1000 ppm) are anesthesia and incoordination. Exposure to methylene chloride results in the formation of carboxyhemoglobin (COHb) caused by its metabolism to carbon monoxide. This COHb is as toxic as that derived from carbon monoxide itself. However, at acceptable levels of exposure to methy-

lene chloride, any probable adverse effects of COHb will be limited to persons with pronounced cardiovascular or respiratory problems. Other possible toxic effects of carbon monoxide itself would not be expected.

Methylene chloride is not teratogenic in animals [1067] and has only limited mutagenic activity in *Salmonella* bacteria. It does not appear to be genotoxic in other species. Available reports of lifetime studies at high concentrations have produced inconsistent results in hamsters, rats, and mice. No tumors, benign or malignant, were increased in hamsters; rats developed only a dose related increase in commonly occurring nonmalignant mammary tumors; white mice, both sexes, had a large increase in cancers of the livers and lungs. Available epidemiological data do not indicate an increase in cancer in humans; they do indicate that the current occupational standards are protective of employee health [1068, 1069].

Trichloromethane. Trichloromethane [*67-66-3*], chloroform, is only moderately toxic from single exposure, but repeated exposure can result in rather severe effects [1063 – 1065]. Its use as a surgical anesthetic has become obsolete, primarily because of delayed liver toxicity and the development of anesthetics with a greater margin of safety.

Ingestion is not likely to be a problem unless large quantities are swallowed accidentally or deliberately. Chloroform has a definite solvent action on the skin and eyes and may be absorbed if exposure is excessive or repeated. Its recognized high chronic toxicity requires procedures and practices to control ingestion, skin, and eye contact, as well as inhalation exposure if liver and kidney injury, the most likely consequence of excessive exposure, is to be prevented.

In animals, chloroform is fetotoxic (toxic to the fetus of a pregnant animal) but only weakly teratogenic if at all [1067]. It does not appear to be mutagenic by common test procedures, but increases the tumor incidence in certain rats and mice. There is considerable evidence that the tumors in rat kidneys and mice livers are the result of repeated injury to these organs and that limiting exposure to levels that do not cause organ injury will also prevent cancer. It is, therefore, very important that human exposure be carefully controlled to prevent injury.

Tetrachloromethane. Tetrachloromethane [*56-23-5*], carbon tetrachloride, was once recommended as a "safety solvent." Misuse and its rather high liver toxicity, as well as the ready availability of alternate safe solvents, have eliminated its application as a solvent. Single exposures are not markedly injurious to the eyes and skin or toxic when small quantities are ingested. However, repeated exposure must be carefully controlled to avoid systemic toxicity, particularly to the liver and kidneys [1063 – 1065]. In humans, injury to the kidney appears to be the principal cause of death.

Inhalation can produce anesthesia at high concentrations, but transient liver as well as kidney injury result at much lower concentrations than those required to cause incoordination. There appears to be individual susceptibility to carbon tetrachloride, with some humans becoming nauseated at concentrations that others willingly tolerate. Ingestion of alcohol is reported to enhance the toxicity of carbon tetrachloride. Such responses should not occur, however, if exposures are properly controlled to the recommended occupational standards.

Carbon tetrachloride is not teratogenic in animals [1067] nor mutagenic in common test systems, but does increase liver tumors in mice, probably as a result of repeated injury to that organ. Therefore, it is very important that human exposure be carefully controlled to prevent liver injury.

10.1.2. Chlorinated C_2 Hydrocarbons

Monochloroethane. Chloroethane [*75-00-3*], ethyl chloride, has limited use as an industrial chemical and is most commonly recognized by its use as a local anesthetic that is sprayed on the skin for minor medical procedures. There are remarkably few published data on its toxicity, since there are very few people exposed in the few uses in industry. It appears to be low in toxicity by inhalation, the only likely route of toxic exposure. Evaporation of large quantities could freeze the skin or eyes [1063 – 1065]. The National Toxicology Program (NTP) has a cancer bioassay underway, and while it appears to be negative, complete results are not yet available.

1,1-Dichloroethane [*75-34-3*]. This flammable substance has limited use and only limited

toxicological data are available. It appears to be low in toxicity by all routes including oral, dermal, and inhalation. Animals tolerated repeated exposures 7 h/d for 9 months to either 500 or 1000 ppm with no adverse effect.

It was not teratogenic when inhaled by pregnant rats. The doses fed by gavage in a carcinogenesis study of the National Cancer Institute (NCI) were so high that mortality was increased. Although no tumor increases were reported, no conclusions could be drawn concerning the induction of cancer [1063 – 1065].

1,2-Dichloroethane. 1,2-Dichloroethane [*107-06-2*], ethylene dichloride, EDC, is one of the more toxic common chlorinated substances [1063 – 1065]. It can cause depression of the central nervous system, mental confusion, dizziness, nausea, and vomiting. Liver, kidney, and adrenal injury may result from both acute overexposure and repeated overexposure at levels significantly above the recommended occupational standards. It has moderate toxicity when swallowed and is often vomited. Skin and eye irritation generally occur only if the liquid is confined. Absorption through the skin is not likely a problem from single contact, but repeated exposure should be avoided. Studies to determine the carcinogenic properties have used excessive doses and produced mixed results. According to the available data, cancer in rodents is caused by repeated injury of the organs and is not likely to occur below occupational limits [1071]. Particular precautions should be taken to assure that skin and inhalation exposures are carefully and appropriately controlled.

1,1,1-Trichloroethane. 1,1,1-Trichloroethane [*71-55-6*], methyl chloroform, has consistently been shown to be among the least toxic chlorinated or nonchlorinated solvents from both acute and repeated exposure [1063 – 1065]. It has been shown repeatedly to have little effect on the liver. It is low in oral toxicity, and has a typical solvent (defatting) action on the skin and eyes; hence, liquid exposure should be minimized. Exposure to more than 1000 ppm of the vapors (significantly above the occupational standards) may cause incoordination with resulting lack of judgment and possible accidents. Misuse and carelessness have resulted in unnecessary deaths, primarily while working in confined spaces. Exposure concentrations under such conditions may attain 10 000 – 30 000 ppm or more. Death is due to anesthesia or possible sensitization of the heart to endogenous adrenalin. Recovery generally has been complete and uneventful if the victim is removed from the exposure before death occurs.

Injury to the liver or other internal organs is unlikely unless severe anesthetic effects have been observed. It has been found to cause no teratogenic or reproductive effects in animals, and extensive study indicates that the vapors are not carcinogenic in rats and mice. It is metabolized only to a very slight degree in animals and humans. It is probably not mutagenic. Human experience has been favorable, as have epidemiological studies on exposed workers. Exposure to high concentrations, particularly in confined spaces, must not be permitted.

1,1,2-Trichloroethane [*79-00-5*]. This substance is relatively toxic and must not be confused with the 1,1,1-isomer discussed previously [1063 – 1065]. It has little use and human experience is limited. According to data from animals, it can cause liver injury as a result of single or repeated exposure. It has a typical solvent effect on the skin and eyes; hence, exposure should be minimized. It apparently has not been studied for its teratogenic effects on animals; it is not mutagenic in common test systems, but it increases the number of liver tumors in mice, probably as a result of organ injury and repeated regeneration. Therefore, it is very important that human exposure be carefully controlled to prevent liver injury.

Other Chloroethanes. 1,1,2,2-tetrachloroethane [*79-34-5*], pentachloroethane [*76-01-7*], and hexachloroethane [*67-72-1*] have limited industrial use, partially because of their recognized high toxicity on repeated exposure [1063 – 1065]. Symmetrical *tetrachloroethane*, $CHCl_2CHCl_2$, at one time was used as a solvent, but liver injury was reported among overexposed workers. It is fetotoxic in rats and increased the tumor rate in the livers of mice. The results of the study of the National Cancer Institute (NCI) on rats was inconclusive. There is limited evidence that it is mutagenic. Much less data are available on the toxicity of *pentachloroethane*, but it is assumed to be highly toxic.

Older data on *hexachloroethane* indicate a much higher toxicity than more recent data

[1063]. This may be due to better purity of the new sample. The recent report indicates a low to moderate oral toxicity, only slight skin and eye irritation, but a moderate to high toxicity from repeated inhalation. It was not markedly hepatotoxic and, at 15 ppm, caused no injury in dogs, rats, guinea pigs, and quail exposed 6 h/d, 5 d/week for 6 weeks. Exposure to 48 ppm caused minimal injury, primarily in the eyes and respiratory tract of the animals. It was not teratogenic in rats or mutagenic in bacteria, but it increased the rate of liver tumors in mice. It caused no cancer in rats, but both rats and mice had evidence of kidney injury. The purity of the sample needs to be verified, however, because contamination with such other substances as tetra- or pentachloroethane may have caused the reported effect.

All three of these chlorinated ethanes must be handled to control exposure and possible liver injury.

Vinyl Chloride. Vinyl chloride [*75-01-4*] is the only chlorinated hydrocarbon that unquestionably has caused cancer (angiosarcoma of the liver) in humans. As a result, there are numerous regulations and laws with regard to its production and use that are intended to minimize exposure. These should effectively eliminate any other toxic effects of vinyl chloride as well as the possibility of cancer.

Because vinyl chloride is a gas, ingestion is not likely in an industrial setting [1063 – 1065]. Skin and eye contact appear to be of concern only from evaporative freezing. Even at adequate protection of the respiratory tract to prevent inhalation, some vinyl chloride may be absorbed through the skin, but the total contribution seems to be slight, even at high concentrations.

When inhaled in high concentrations (10 000 ppm), anesthetic effects can occur. At even higher concentrations, the effects increase and deaths have been reported from massive exposures. Odor provides little warning of excessive exposure.

Many other effects have been alleged to occur as a result of excessive exposure to vinyl chloride, but only a few are clearly the direct result of exposure. It is hepatotoxic, possibly mutagenic, and has caused angiosarcoma of the liver. A condition known as acroosteolysis with scleroderma has been associated with cleaning autoclaves used for polymerization. Whether the disease, called *kettle cleaner's disease*, is related to vinyl chloride itself or to some other substance is not known. Likewise, other cancers alleged to be caused by vinyl chloride may or may not be the result of vinyl chloride itself, since they are not consistent throughout the industry and appear to be found only in certain populations.

Careful controls to minimize exposures and adherence to regulatory requirements are essential.

Vinylidene Chloride. Vinylidene chloride [*75-35-4*], 1,1-dichloroethylene, is an anesthetic at high concentrations (several thousand ppm) [1063 – 1065]. Hepatotoxicity can result from rather low exposures; therefore, low TLVs and MAK's have been recommended. It has a solvent effect on the skin and eyes but its high volatility probably precludes absorption through the skin in most situations. Exposure should be carefully controlled. The effect on the liver is quite marked on repeated exposure of animals. Although not a teratogen, it does cause injury to embryos and fetuses of exposed animals at levels causing injury to the pregnant mothers. The metabolites of vinylidene chloride are at most weakly mutagenic in bacterial test systems, and tests in mammalian systems are negative. It is probably not carcinogenic based on the weight of evidence, for only one out of 14 tests has been marginally positive. It is important that human exposure be carefully controlled to prevent liver injury.

1,2-Dichloroethylene. Most toxicological testing has been on mixed isomers [1063]. It is not clear from available data on the isomers how they compare in toxicity. Most results indicate a moderate oral toxicity to rats. A typical solvent effect on the skin and eyes is expected although data are not available to verify this conclusion. Exposure should be controlled.

Most studies on animals indicate a rather low toxicity by inhalation, with little effect on the liver. Anesthesia occurs at higher levels, but the data are inconsistent as to the actual levels required.

The effects of repeated exposure are not clear either. One reference reported no adverse effect on 7-h daily exposure for six months to 500 or 1000 ppm, but a second reported rather marked

effects at 200 ppm after 14 weeks. No data were found on teratogenesis or carcinogenesis, but very limited data indicate no mutagenic effect.

Trichloroethylene [*79-01-6*]. There is a tremendous amount of literature on trichloroethylene because of its use as a degreasing solvent and even more so of its use as an anesthetic have resulted in considerable human exposure [1063 – 1065]. With such a vast literature, conflicting conclusions are possible. The toxicity is generally considered low to moderate. Liver and kidney injury do not appear to be a common response even after excessive exposures which cause anesthesia. Trichloroethylene has a typical solvent (defatting) action on the skin and eyes and exposure should be controlled. Absorption through the skin may occur but is not likely a significant source of exposure. When inhaled, it can have a pronounced anesthetic effect (depression of the central nervous system), which may become evident as incoordination at concentrations of 400 ppm or more. Visual disturbances, mental confusion, fatigue, and sometimes nausea and vomiting are observed at higher levels. The nausea is not nearly as marked as with carbon tetrachloride or ethylene dichloride. Sensitization of the heart to adrenalin may occur, but it does not appear to be significant unless markedly anesthetic concentrations are reached (several thousand ppm).

Deliberate sniffing has been a problem, although physical dependence does not appear to be involved. A peculiar vascular dilation of the face, neck and trunk, known as *"degreaser's flush,"* occasionally occurs when alcohol is consumed during or following exposure. Although upsetting, the flush does not appear to be serious.

Urinary metabolites, trichloroacetic acid and trichloroethanol, are measured in several countries to monitor workers' exposure, but this procedure has severe disadvantages. Urinary metabolites are more related to chronic (total) exposure than they are to acute (peak) exposure. Thus, the worker may have excessive peak exposures which could result in anesthesia, incoordination and possible accidents, and yet their urinary metabolites remain within accepted limits at the end of the day.

Trichloroethylene has not shown teratogenic effects in animals. It possibly is weakly mutagenic but much of the testing is suspect due to impurities (stabilizers) present in the samples.

Trichloroethylene appears to increase the number of liver tumors in certain mice given massive doses by gavage and lung tumor in one strain of mice by inhalation, but it has generally been negative in rats and other rodent studies. The significance to human cancer is not clear [1064]. Several rather small epidemiological studies have failed to show an increase in human liver cancer of exposed workers.

Tetrachloroethylene [*127-18-4*]. Although it is not a potent anesthetic, depression of the central nervous system (incoordination) is the most common response to tetrachloroethylene at concentrations above 200 ppm [1063 – 1065]. Liver and kidney injury does not appear to be a common response even after excessive exposure. It is moderate to low in oral toxicity, has a solvent action on the skin and eyes, and is poorly absorbed through the skin. Exposure of the skin and eyes should be carefully controlled. The odor begins to be objectionable at about 400 ppm for most people. When inhaled at high concentrations, it may cause nausea and gastrointestinal upset in addition to the anesthesia and incoordination. It is much weaker in producing nausea than carbon tetrachloride and ethylene dichloride.

Reports of human injury are uncommon despite its wide usage in dry cleaning and degreasing. Sensitization of the heart to adrenalin does not appear to be a likely consequence.

Tetrachloroethylene is not extensively metabolized and most of the absorbed dose is excreted unchanged in the expired air. Analysis of metabolites in urine is therefore of even less value than with trichloroethylene.

Tetrachloroethylene is not teratogenic in standard tests in animals; it does not appear to be significantly mutagenic and increases tumors in certain strains of mice. Like many of the chlorinated hydrocarbons, the significance of the mouse liver tumor to human cancer is questionable.

Dichloroacetylene [*7572-29-4*]. This substance is discussed only because it is a highly toxic and pyrophoric substance formed by dehydrochlorination of trichloroethylene [1063]. This has occurred when trichloroethylene vapors (or the liquid) are passed over soda – lime

or caustic soda. Much of the experience comes from the use of Hopcalite in rebreathing anesthesia machines or in closed environmental systems. Exposure to as little as 19 ppm is reported to cause the death of half of the mice (LC_{50}) in a 4-h exposure. Repeated exposure caused kidney injury and muscular paralysis. Dichloroacetylene is reported to cause headache, nausea, and nerve, liver, and kidney injury in humans or animals.

10.1.3. Chloropropanes and Chloropropenes

2-Chloropropane [*75-29-6*], isopropyl chloride, has had little use or study [1063]. Its flammability has probably discouraged its use, although it appears to be of low toxicity in animals. When fed by gavage to rats, a dose of 3 g/kg was survived. Ten repeated applications of the liquid on the skin of a rabbit was very slightly irritating if allowed to evaporate and slightly more irritating if confined under a bandage. Exposure of the skin and eyes should be prevented. Data on the effects of inhalation appear to be limited. According to one study, repeated exposure of rats to 250 ppm caused no effect, but 1000 ppm did. Another study claimed no effect in rats, mice, rabbits, guinea pigs, and monkeys exposed 7 h/d 5 d/week for 6 months to 500 or 1000 ppm. Limited data suggest that isopropyl chloride caused a mutagenic response in bacteria, but no data were found on teratogenesis or carcinogenesis.

1,2-Dichloropropane [*78-87-5*], propylene dichloride, appears to be low to moderate in toxicity on single exposure, but moderately toxic on repeated exposure [1063 – 1065]. It has a solvent effect (defatting action) on the eyes and skin. Exposure should be prevented. Absorption of the liquid through the skin may occur, particularly on repeated contact. When inhaled by mice, respiratory injury rather than liver toxicity appeared to be the primary cause of death following single exposure. There appears to be little recent data on the effect of repeated inhalation by animals; hence, the recommended TLV and MAK are based on old studies.

The National Toxicology Program (NTP) has a carcinogenic study under way, but no report has been issued. No data were found in regard to teratogenic effects, but minimal data indicate it may be mutagenic in bacterial test systems.

10.1.4. Chlorobutadienes

2-Chloro-1,3-butadiene [*126-99-8*], chloroprene, is a rather toxic, highly flammable monomer used to produce synthetic rubber [1063]. The toxicity of various samples appears to have been influenced by reaction products because chloroprene reacts with oxygen to form peroxides. It also dimerizes. It must be handled carefully to prevent these reactions. Liver injury is possible, as is hair loss. The hair loss appears to be caused by a reaction of chloroprene with the hair itself, since regrowth occurs when exposure ceases. The vapors cause respiratory irritation, as well as pain and irritation in the eyes. The pure material did not cause teratogenic effects. It is mutagenic in some bacterial test systems. Early reports of cancer among workers have not been confirmed in more carefully conducted studies, and animal studies have been negative [1064].

Hexachlorobutadiene [*87-68-3*] has been used as a pesticide, but it is a largely unwanted byproduct made during chlorination of hydrocarbons [1063]. It is highly toxic and, unlike the other substances discussed in this section, of low volatility. The literature on hexachlorobutadiene has been reviewed [1072], and a summary of the data is found in reference [1063]. There were no references to human exposure. In animals, hexachlorobutadiene has caused liver injury after single exposure and liver and kidney injury, as well as kidney tumors, after repeated ingestion. The kidneys are the primary target organ on repeated exposure.

Precautions must be taken to minimize exposure to hexachlorobutadiene by skin, ingestion, and inhalation.

10.1.5. Ecotoxicology and Environmental Degradation

Aquatic Toxicity. Most common aliphatic chlorinated hydrocarbons have been tested for their acute toxicity by both static and flow-through methods using vertebrates (fish) and invertebrates (water fleas). According to these studies the compounds tested were either nontoxic or slightly toxic as defined by the EPA (Table 61) [1076].

Table 61. Aquatic toxicity of aliphatic chlorinated hydrocarbons (LC_{50}, mg/L) [1073]

	Fathead minnow		*Daphnia* magna
	96-h static	96-h flow through	48-h static
Dichloromethane	310	193	224
Trichloromethane	131*	–	28.9
Tetrachloromethane	53.2**	43.1	35.2
1,1,1-Trichloroethane	105	52.8	>530
Trichloroethylene	66.8	40.7	85.2
Tetrachloroethylene	21.4	18.4	17.7

* [1074]
** [1075]

Presence in Water. According to analyses carried out in Western Europe, typical concentrations of 1,1,1-trichloroethane, trichloroethylene, and perchloroethylene in surface water range from 0.1 to 3 µg/L [1077]. Trace concentrations of the common chlorinated hydrocarbons have been found in drinking water. Dichloromethane has not been detected in natural waters, oceans, sediments, or fish [1078]. Trichloromethane is often formed at sub-ppm levels in the disinfection of drinking water because chlorine reacts with natural humic substances. The concentrations of common chlorinated solvents in drinking water are typically below 0.3 µg/L [1077].

Presence and Degradation in Air. During their use the common aliphatic chlorinated hydrocarbons escape into the atmosphere [1079]. Simple chlorinated hydrocarbons are destroyed in the troposphere, primarily by reaction of their hydrogen atoms or double bonds with hydroxyl radicals (HO ·) that are naturally present in the troposphere. However, carbon tetrachloride and some of the fluorochloroalkanes (CFC 11, CFC 12, and CFC 113) resist significant attack by hydroxyl radicals; therefore, these materials have much longer half-lifes than other common chlorinated solvents (Table 62) [1080]. The degradation of most commercially used chlorinated hydrocarbons in the atmosphere leads to HCl, chlorides, CO_2, and H_2O, which represent no significant threat to the general environment.

Table 62. Tropospheric residence time of aliphatic chlorinated hydrocarbons

Compound	Time, a
Dichloromethane	0.23
Trichloromethane	0.33
Tetrachloromethane	> 30.0
1,1,1-Trichloroethane	2.7
Trichloroethylene	0.01
Tetrachloroethylene	0.18

Atmospheric concentrations of chlorinated solvents are highly dependent on the site of measurement. In urban areas, the concentration of dichloromethane is $30-100\times10^{-6}$ mL/m^3 and the concentrations of 1,1,1-trichloroethane, trichloroethylene, or perchloroethylene are $100-600\times10^{-6}$ mL/m^3. In areas remote from industrial and populated regions, background concentrations are typically in the range of a few ppt (10^{-6} mL/m^3), except for 1,1,1-trichloroethane which has been reported to occur at $100-200\times10^{-6}$ mL/m^3 [1081].

Most of the traces of chlorinated hydrocarbons found in the atmosphere appear to be of anthropogenic origin. For methyl chloride, however, and possibly for some other halomethanes, the largest source is natural. Thus, methyl chloride results from forest fires, agricultural burning, chemical reactions in seawater, and possibly from marine plants [1082].

10.2. Chlorinated Aromatic Hydrocarbons

10.2.1. Chlorinated Benzenes

Acute Toxicity. Table 63 shows the LD_{50}'s of chlorinated benzenes. According to these data, the acute oral toxicity of chlorinated benzenes is low to moderate. Absorption through the skin seems to be minimal, but most of the compounds have some local irritant potency. In experimental animals, most chlorinated benzenes induce microsomal liver enzymes and cause porphyria, hypertrophy, and centrolobular necrosis of the liver. The chlorinated benzenes can induce kidney damage, changes in mucous membranes, and irritation of the upper respiratory tract, depending on the route and time of administration and on the dose applied. In addition, mono-, di-, and trichlorobenzenes are known to act as central nervous depressants, causing anesthesia and narcosis at higher doses [1063, 1064].

Table 63. Oral LD_{50}s of chlorobenzenes [1083]

Compound	LD_{50} (rat, oral), mg/kg
Monochlorobenzene	2 910
o-Dichlorobenzene	500
p-Dichlorobenzene	500
1,2,4-Trichlorobenzene	756
1,2,4,5-Tetrachlorobenzene	1 500
Pentachlorobenzene	1 080
Hexachlorobenzene	10 000

Chronic Effects. Hexachlorobenzene produced tumors in studies with mice and hamsters [1063]. There is no evidence for the carcinogenicity of the other chlorobenzenes. None of the compounds showed *mutagenic* activity in validated test systems. No data concerning *teratogenic* or *embryotoxic* effects of the chlorinated benzenes are available for most of the compounds. *p*-Dichlorobenzene has been tested in several species and produced no primary embryotoxic or teratogenic effects [1084].

Metabolism. Usually the chlorinated benzenes are partially hydroxylated to yield the corresponding phenols or are partially dechlorinated and then excreted in the urine and the feces. In contrast to rodents, sulfur-containing metabolites cannot be found in monkeys and humans [1063, 1085, 1086]. Only hexachlorobenzene has been shown to accumulate in animals and humans [1087, 1088].

Hexachlorobenzene has been suggested to produce *immunotoxic* effects in experimental animals, i.e., alterations of cell-mediated immune responses [1063].

Effects in Humans. Monochlorobenzene produced unconsciousness, vascular paralysis, and heart failure in a child after accidental oral uptake [1063]. *o*-Dichlorobenzene produced depression in conditioned reflex activity, demonstrating a cerebral – cortical effect. Erythropoiesis was significantly depressed. Symptoms of intoxication include headache, nausea, throat irritation, and stinging of the eyes. Skin irritation is also reported [1064].

No data are available on the effects of trichlorobenzenes in humans. Only minimal eye and throat irritation at 3 – 5 ppm in certain people are reported [1064]. An outbreak of cutanea tarda porphyria in Turkey in 1955 was attributed to the uptake of grain treated with hexachlorobenzene as fungicide [1063].

Regulations. The following occupational exposure limits have been established (MAK 2001, TLV 1985):

Monochlorobenzene:	MAK 10 ppm	TLV 75 ppm
o-Dichlorobenzene:	MAK 10 ppm	TLV 50 ppm
p-Dichlorobenzene:	Carcinogen in animal experiments	TLV 75 ppm
1,2,4-Trichlorobenzene:	Suspected carcinogen	

10.2.2. Chlorotoluenes

Little is known about the toxicity of chlorinated toluenes. The toxicity of *o*-chlorotoluene is considered to be relatively low and in the range of that of the chlorinated benzenes. The oral LD_{50} in rats is greater than or equal to 1600 mg/kg [1064]. Sublethal doses of *o*-chlorotoluene given orally to rats produce marked weakness; higher doses produce vasodilatation. Inhalation of 14 000 ppm for 6 h caused loss of coordination, vasodilatation, labored respiration, and narcosis in rats; 175 000 ppm was fatal to one of three rats [1064]. *o*-Chlorotoluene produces moderate skin irritation and conjunctival erythema in rabbits [1064].

p-Chlorotoluene induced no gene conversion in *Saccharomyces cerevisiae* [1089]. No data are

available on chronic effects or effects on reproduction caused by chlorinated toluenes.

In *humans*, no cases of poisoning or skin irritation caused by chlorinated toluenes have been reported [1064].

Regulations. o-Chlorotoluene TLV 50 ppm

10.2.3. Polychlorinated Biphenyls

Acute Toxicity. The acute toxicity of mixtures of polychlorinated biphenyls (PCB) seems to depend on the chlorine content. Table 64 demonstrates the influence of the chlorine content in mixed isomers of PCBs, in addition to their relatively low acute oral toxicity.

Table 64. Acute oral toxicity of PCBs [1090]

Chlorine content, wt %	LD_{50} (rat, oral), g/kg
21	3.98
32	4.47
42	8.65
48	11.0
62	11.3
68	10.9

The administration of acute or subacute doses results in liver enlargement, mainly due to enzyme induction; when the doses were increased, fatty degeneration and central atrophy of the liver occurred. In addition, hyperplasia and hemosiderosis of the spleen were also observed [1091, 1092]. Polychlorinated biphenyls are not likely to possess a substantial local irritating potential. Nevertheless, they seem to be readily absorbed through the skin, exerting such systemic effects as liver damage.

Chronic Effects. After oral application, severe liver damage (hypertrophy, fatty degeneration, and centrolobular necrosis) is most likely to be observed. The skin is also often affected (hyperplasia, hyperkeratosis, and cystic dilatation) [1069]. Polychlorinated biphenyls can interfere with heme metabolism as shown by an increased porphyrin content of the liver in rats [1069]. Hepatocellular tumors are produced in rats and mice after long-term oral application of PCBs [1066], vol. 20. However, the tumor formation is regarded as a response to tissue damage rather than triggered by a genotoxic mechanism.

The PCBs failed to show positive response in validated *mutagenicity* test systems. The interference of polychlorinated biphenyls with reproduction could be demonstrated in numerous studies with mammals. The compounds pass through the placental barrier and exhibit embryotoxic effects [1069].

Absorption, Metabolism, and Excretion. Polychlorinated biphenyls are readily absorbed from the gastrointestinal tract after ingestion or from the lung after inhalation. The rates of metabolism and excretion decrease and the storage in body fat increases with increasing chlorine content. The compounds are generally metabolized by selective hydroxylation. In primates, most of the metabolites are excreted as conjugates in the urine, whereas excretion of free metabolites in the feces is the major route in rodents [1093].

Other Effects. Immunosuppressive action of polychlorinated biphenyls in mammals could be evidenced by a decrease in infectious resistance with atrophy of the spleen, cortical thymus atrophy, and dose-dependent decreased in the production of specific antibodies [1063, 1069]. In hens, growth retardation, high mortality, and subcutaneous edema could be observed. These findings were accompanied by focal necrosis of the liver, hydroperitoneum, and epicardial as well as lung edema (chicken edema disease) [1063, 1069].

Experience in Humans. Accidental acute intoxications with PCBs are not reported [1069]. With workers handling these compounds, acneform dermatitis was observed, in addition to liver damage with necrosis [1063, 1069].

In 1968, a subacute intoxication of more than 1000 people in Japan by contaminated rice oil was reported (Yusho disease). Initial symptoms were, for instance, swelling of the eye lids, fatigue, and gastrointestinal disturbances. Later on, discoloration of the skin and mucous membranes, headache, signs of sensory nerve injury, diarrhea, and jaundice were found. Cases of influence on human fetuses have been attributed to this high PCB exposure [1063, 1069].

Polychlorinated biphenyls accumulate in fat and adipose tissue. They have been demonstrated in human milk. Because the PCB level was found to be higher in infant blood but lower in umbilical blood in comparison to maternal

blood, the transfer of PCBs via the milk is probably much more important than placental transfer [1063, 1094].

Regulations. The following exposure limits have been established:

Chlorine content 42 %:	MAK 0.1 ppm
Chlorine content 54 %:	MAK 0.05 ppm

Polychlorinated biphenyls are considered as possible teratogens [1095] and carcinogens [1069].

10.2.4. Chlorinated Naphthalenes

Monochlorinated naphthalenes are of low to moderate acute toxicity, as shown by their oral LD_{50} (Table 65). Subacute to subchronic uptake of mixtures of higher chlorinated naphthalenes (predominantly penta- and hexachloronaphthalene) resulted in liver injury [1063]. In general, the toxicity of chlorinated naphthalenes increases with the degree of chlorination [1064]. Chlorinated naphthalenes irritate the rabbit skin [1063]. Ingestion of lubricants containing chloronaphthalenes resulted in injury to farm animals (X disease). Marked hyperkeratosis of the skin, degenerations of the cells in pancreas, liver, and gall bladder, and damage of the renal cortex could be observed. Cattle poisoned with highly chlorinated naphthalenes show a rapid decline in vitamin A plasma levels [1063]. Octachloronaphthalene fed to rats also greatly enhances the loss of vitamin A from the liver [1096]. Mixtures of penta- and hexachloronaphthalenes can produce the so-called chicken edema disease, characterized by hydropericardium and ascites in chickens [1096].

Table 65. Oral LD_{50}s of monochlorinated naphthalenes [1083]

Compound	Species	LD_{50} (oral), mg/kg
1-Chloronaphthalene	rats	1540
	mice	1019
2-Chloronaphthalene	rats	2078
	mice	886

1-Chloronaphthalene and 1,2,3,4-tetrachloronaphthalene, when tested for point mutations in the *Salmonella* assay (*Ames test*), exhibited no positive results [1097, 1098]. No data are available on the effects of chlorinated naphthalenes on reproduction.

Metabolism. Chlorinated naphthalenes are readily absorbed. Metabolism occurs by conjugation or via hydroxylation to the respective naphthols or dihydrodiols. The metabolites are excreted with the urine or the feces [1063, 1096].

Effects in Humans. The main health problem arising from use and handling of chlorinated naphthalenes is chloracne, which usually occurs from long-term contact with the compounds or exposure to hot vapors. The penta- and hexachloro derivatives are suggested to have the greatest potential to generate acne [1063].

In accidental intoxications, liver damage occurred independently from chloracne. After loss of appetite, nausea, and edema of the face and hands, abdominal pain and vomiting followed; later on jaundice developed. Autopsy in cases of fatal intoxication revealed yellow atrophy of the liver [1096].

Regulations. The following exposure limits have been established:

Trichloronaphthalene:	MAK 5 mg/m^3,
	TLV 5 mg/m^3
Tetrachloronaphthalene:	TLV 2 mg/m^3
Hexachloronaphthalene (skin):	TLV 0.2 mg/m^3
Pentachloronaphthalene:	TLV 0.5 mg/m^3
Octachloronaphthalene (skin):	TLV 0.1 mg/m^3

10.2.5. Benzyl Chloride

The acute oral toxicity (LD_{50}) of benzyl chloride in rats is 1231 mg/kg and in mice 1624 mg/kg [1083]. The subcutaneous LD_{50} (in rats) of benzyl chloride in oil solution is 1000 mg/kg [1099]. Exposure of rats and mice to benzyl chloride concentrations of 100–1000 mg/m^3 for 2 h caused irritation of the mucous membranes and conjunctivitis [1066], vol. 11. Benzyl chloride is a strong skin-sensitizing agent for guinea pigs [1100]. Benzyl chloride acts weakly *mutagenic* in validated test systems [1101, 1102].

Subcutaneous injection of weekly doses of 80 mg/kg for 1 year followed by a post-observation period resulted in local sarcomas with lung

Table 66. Acute toxicity of side-chain chlorinated xylenes [1083]

Compound	Species	Route	LD_{50}, mg/kg
m-Xylylene dichloride	mice	intravenous	100
o-Xylylene dichloride	mice	intravenous	320
p-Xylylene dichloride	rats	oral	1780

metastases in rats. The mean induction time was 500 d [1099]. After dermal application of benzyl chloride, skin carcinomas were observed in mice [1103].

Metabolism. Benzyl chloride is readily absorbed from the lungs and gastrointestinal tract. The compound reacts with tissue proteins after subcutaneous injection and is metabolized into *N*-acetyl-*S*-benzylcysteine [1100]. After oral administration, mercapturic acid and benzoic acid (free or conjugated with glycine) are excreted in the urine [1104].

Effects in Humans. A concentration of 16 ppm of benzyl chloride in air is reported to be intolerable to humans within 1 min. The compound is a potent lachrymator, strongly irritating to the eyes, nose, and throat and capable of causing lung edema [1064].

Regulations. The exposure limits of benzyl chloride are: MAK 1 ppm; TLV 1 ppm. Benzyl chloride should be considered as a possible carcinogen [1069].

10.2.6. Benzoyl Chloride

Benzoyl chloride is of low acute oral toxicity in rats (LD_{50} 2529 mg/kg). It is more toxic by inhalation (LC_{50} 230 ppm, 4 h in male rats and 314 ppm, 4 h in female rats). The compound is irritating to skin, mucous membranes, eyes, and the respiratory tract [1105, 1106].

When benzoyl chloride or solutions of benzoyl chloride in benzene were applied to the skin of mice for up to 10 months irritation and keratinization resulted, and to some extent, ulceration and necrosis of the skin occurred. A few tumors (skin, lung) were observed in those mice [1103]. There is no clear evidence that benzoyl chloride is *mutagenic* [1101].

For *humans*, benzoyl chloride is classified as a lachrymator. It is irritating to the skin, eyes, and mucous membranes [1107]. The available data are inadequate to evaluate the carcinogenic potential of benzoyl chloride to humans [1066], vol. 29.

10.2.7. Benzotrichloride

The acute oral toxicity of benzotrichloride is 2180 mg/kg in male rats and 1590 mg/kg in female rats. The inhalative LC_{50}s are higher than 600 mg/m^3 in male rats and about 500 mg/m^3 in female rats after a 4-h exposure [1108]. Benzotrichloride irritates the skin and eyes [1109]. The compound proved to be *mutagenic* in bacterial test systems [1101]. Dermal application of benzotrichloride resulted in elevated tumor incidence in mice [1103].

In *humans*, benzotrichloride vapors are reported to be strongly irritating to the skin and mucous membranes [1107]. An increase in lung tumors has been reported in industrial plants that produce several chlorinated aromatic hydrocarbons [1110, 1111]. In the Federal Republic of Germany, and in Japan, benzotrichloride is considered as a possible carcinogen [1069, 1103].

10.2.8. Side-Chain Chlorinated Xylenes

Table 66 shows some acute toxicity data of side-chain-chlorinated xylenes. No data are available on other toxic effects in animals or humans.

11. References

General References

1. *Ullmann*, 4th ed., **9**, 404–420.
2. *Winnacker-Küchler*, 4th ed., vol. **6**; Organische Technologie II, pp. 2–11.
3. *Kirk-Othmer*, **5**, 668–714.
4. J. J. McKetta, W. Cunningham, *Encyl. Chem. Process. Des.* **8** (1979) 214–270.
5. C. R. Pearson: "C_1 and C_2-halocarbons," in: *The Handbook of Environmental Chemistry*, vol. **3**, Springer Verlag, Berlin 1982, pp. 69–88.

6. *Chloroform, Carbon Tetrachloride and other Halomethanes: An Environmental Assessment*, National Academy of Sciences Washington, D.C., 1978.
7. A. Wasselle: *C_1 Chlorinated Hydrocarbons*, Private report by the Process Economic Program SRI International, Report No. 126.
8. CEFIC-B.I.T. Solvants Chlorés "Methylene chloride–Use in industrial applications," 1983.
9. J. Schulze, M. Weiser: "Vermeidungs- und Verwertungsmöglichkeiten von Rückständen bei der Herstellung chloroorganischer Produkte," Umweltforschungsplan des Bundesministers des Innern–Abfallwirtschaft. Forschungsbericht 103 01 304, UBA-FB 82–128 (1985).

Specific References

10. I. Mellan: *Industrial Solvents Handbook*, Noyes Data Corp., Park Ridge, N.J., 1970, p. 73.
11. B. Kaesche-Krischer, *Chem. Ind. Tech.* **35** (1963) 856–860.
H.-J. Heinrich, *Chem. Ing. Tech.* **41** (1969) 655.
12. H. J. Schumacher, *Z. Elektrochem.* **42** (1936) 522.
13. D. V. E. George, J. H. Thomas, *Trans Faraday Soc.* **58** (1962) 262.
14. Du Pont, US 2 378 048, 1944 (C. W. Theobald).
15. H. J. Schumacher, *Z. Angew. Chem.* **49** (1936) 613.
16. A. T. Chapmann, *J. Am. Chem. Soc.* **57** (1935) 416.
17. R. Neu, *Pharmazie* **3** (1948) 251.
18. F. Lenze, L. Metz, *Chem. Ztg.* **56** (1932) 921.
19. W. B. Grummet, V. A. Stenger, *Ind. Eng. Chem.* **48** (1956) 434.
20. E. H. Lyons, R. G. Dickinson, *J. Am. Chem. Soc.* **57** (1935) 443.
21. R. Johns, Mitre Corp. Tech. Rep. MTR 7144 (1976) Mc Lean, Va.; Mitre Corp.
22. Celanese Corp., US 2 770 661, 1953 (T. Horlenko, F. B. Marcotte, O. V. Luke).
23. United States Rubber Co., GB 627 993, 1946.
24. R. M. Joyce, W. E. Hanford, *J. Am. Chem. Soc.* **70** (1948) 2529.
25. W. L. Faith, D. B. Keyes, R. L. Clark: *Industrial Chemicals*, 3rd ed., J. Wiley & Sons, New York 1965, pp. 507–513.
26. E. T. McBee et al., *Ind. Eng. Chem.* **41** (1949) 799–803.
27. R. N. Pease et al., *J. Am. Chem. Soc.* **53** (1931) 3728–3737.
28. W. E. Vaughan et al., *J. Org. Chem.* **5** (1940) 449–471.
29. S. Tustev et al., *J. Phys. Chem.* **39** (1935) 859.
30. F. Asinger: *Paraffins Chemistry and Technology*, Pergamon Press, New York 1968.
31. T. Arai et al., *Kogyo Kagaku Zasshi* **61** (1958) 1231–1233.
32. H. B. Hass, E. T. McBee et al., *Ind. Eng. Chem.* **34** (1942) 296–300.
33. M. J. Wilson, A. H. Howland, *Fuel* **28** (1949) 127.
34. G. Natta, E. Mantica, *J. Am. Chem. Soc.* **74** (1952) 3152.
35. A. Scipioni, E. Rapsiardi, *Chim. Ind. Milan* **43** (1961) 1286–1293.
36. S. Lippert, G. Vogel, *Chem. Tech. (Leipzig)* **21** (1969) 618–621.
37. W. Hirschkind, *Ind. Eng. Chem.* **41** (1949) 2749–2752.
38. B. E. Kurtz, *Ind. Eng. Chem. Process Des. Dev.* **11** (1972) 332–338.
39. D. Kobelt, U. Troltenier, *Chem. Ing. Tech.* **38** (1966) 134.
40. Stauffer Chem. Co., US 3 126 419, 1964 (W. M. Burks, R. P. Obrecht).
41. H. Krekeler, W. Riemenschneider, *Chem. Ing. Tech.* **45** (1973) 1019–1021.
42. D. Rebhahn, *Encycl. Chem. Process. Des.* **8** (1979) 206–214.
43. E.-G. Schlosser, M. Rossberg, W. Lendle, *Chem. Ing. Tech.* **42** (1970) 1215–1219.
44. Hoechst, EP 0 039 001, 1983 (W. Grünbein et al.).
45. Dow Chem. Co., US 3 981 938, 1976 (J. M. Steel et al.).
46. Dow Chem. Co., DE 2 640 852, 1978 (J. M. Steel et al.).
47. Shinetsu Chem. Co., DE 2 447 551, 1974 (K. Habata, S. Tanaka, H. Araki).
48. Sun Oil Co., US 3 236 754, 1962 (J. B. Bravo, R. Wynkoop).
49. A. A. Schamschurin, *Zh. Obshch. Khim.* **9** (1939) 2207.
50. R. F. Weinland, K. Schmidt, *Ber. Dtsch. Chem. Ges.* **38** (1905) 2327, 3696.
51. A. Hochstetter, DE 292 089, 1914.
52. Wacker Chemie, DE 1 211 622, 1964 (G. Künstle, H. Siegl).
53. Olin Mathieson Chem. Corp., US 2 755 311, 1951 (H. Spencer).
54. Hercules Powder Co., US 2 084 710, 1935 (H. M. Spurlin).

55. Th. Goldschmidt, DE 1 150 964, 1961 (E. Ruf).
56. Shinetsu Chem. Co., DE 3 146 526, 1981 (K. Habata, M. Shimizu, K. Ichikawa).
57. Hoechst, DE 946 891, 1953.
58. Stauffer Chem. Co., US 2 806 768, 1957 (H. Bender, R. P. Obrecht).
59. Chem. Werke Hüls, DE 1 101 382, 1957 (R. Marin, H. Sauer).
60. Hoechst, DE 2 137 499, 1973 (A. Bergdolt, H. Clasen, D. Houben).
61. *Hydrocarbon Process. Pet. Refiner* **42** (1963) no. 11, 157.
62. *Hydrocarbon Process. Pet. Refiner* **40** (1961) no. 11, 230–231.
63. Stauffer Chem. Co., DE 1 443 156, 1961 (R. P. Obrecht, W. M. Burks).
64. Solvay Cie., DE 2 106 886, 1971 (L. Forlano).
65. M. Nagamura, H. Seya, *Chem. Econ. Eng. Rev.* **2** (1970) 34–38.
66. Hoechst, DE 1 568 575, 1968 (M. Rossberg et al.).
67. Stauffer Chem. Co., US 3 968 178, 1976 (R. P. Obrecht, M. J. Benett).
68. IG Farbenind., DE 657 978, 1936 (K. Dachlauer, E. Schnitzler).
69. S. Akliyama, T. Hisamoto, S. Mochizuku, *Hydrocarbon Process.* **60** (1981) no. 3, 76.
Tokuyama Soda Co., JP Sho 55–43002, 1978 (S. Mochizuki, T. Hisamato, T. Hirashima).
Tokuyama Soda Co., JP OLS Sho 56–2922, 1979 (S. Mochizuki, T. Hisamato).
70. Lummus Co., US 3 949 010, 1976 (M. Sze).
71. Diamond Alkali Co., US 2 792 435, 1954 (J. J. Lukes, W. J. Lightfoot, R. N. Montgomery).
72. S. A. Miller, *Chem. Process. Eng.* **48** (1967) no. 4, 79–84.
73. *Ullmann*, 3rd ed., **5**, 409.
74. Food and Machinery Corp., US 3 109 866, 1963 (E. Saller et al.).
75. Food and Machinery Corp., US 3 081 359, 1963 (E. Saller, R. Timmermann, C. J. Wenzke).
76. Stauffer Chem. Co., US 3 859 425, 1975 (A. Wood, K. V. Darragh).
77. *Ullmann*, 4th ed., **9**, 461–463.
78. *Hydrocarbon Process. Pet. Refin.* **44** (1965) no. 11, 190.
79. Halcon Chem. Co., BE 663 858, 1965 (J. W. Colton).
80. Dow Chem. Corp., US 2 442 324, 1945 (R. G. Heitz, W. E. Brown).
81. Stauffer Chem. Co., US 2 857 438, 1957 (R. P. Obrecht, H. Bender).
82. Halcon Int. Co., GB 1 047 258, 1963.
83. Soc. d'Ugine, FR 1 147 756, 1956.
84. Solvay Cie., FR 1 373 709, 1963.
85. Hoechst, DE 1 668 030, 1967 (H. Krekeler, H. Meidert, W. Riemenschneider, L. Hörnig).
Hoechst, DE 1 668 074, 1968 (L. Hörnig, H. Meidert, W. Riemenschneider).
Hoechst, DE 1 918 270, 1969 (H. Krekeler, H. Meidert, W. Riemenschneider, L. Hörnig).
Hoechst, DE 2 218 414, 1972 (H. Gerstenberg, H. Krekeler, H. Osterbrink, W. Riemenschneider).
Hoechst, DE 3 040 851, 1980 (D. Rebhan, H. Schmitz).
86. Stauffer Chem. Co., US 3 126 419, 1960 (W. M. Burks, R. P. Obrecht).
87. *Hydrocarbon Process. Pet. Refin.* **40** (1961) no. 11, 230–231.
88. *Hydrocarbon Process. Pet. Refin.* **40** (1961) no. 11, 232.
89. Sicedison SpA, FR 1 386 023, 1964.
90. Pittsburgh Plate Glass Co., FR 1 355 886, 1963 (L. E. Bohl, R. M. Vancamp).
91. B. F. Goodrich Co., US 3 314 760, 1962 (L. E. Trapasso).
92. J. A. Thurlby, O. Stinay, *Chem. Anlagen Verfahren* **11** (1972) 65.
93. E. Dönges, R. Kohlhaas et al., *Chem. Ing. Tech.* **38** (1966) 65.
94. Gemeinsames Ministerialblatt 31/26, 430–452 (1980): Katalog wassergefährdender Stoffe.
95. Amtsblatt der Europäischen Gemeinschaften C 176, 7–10 (1982): Liste von Stoffen, die gegebenenfalls in die Liste I der Richtlinie des Rates 76/464/EWG aufzunehmen sind.
96. J. Schulze, M. Weiser, *Chem. Ind. (Düsseldorf)* **36** (1984) 347–353.
97. C. R. Pearson: "C1 and C2 Halocarbons," in D. Hutzinger (ed.): *The Handbook of Environmental Chemistry*, vol. **3, B**, Springer Verlag, Berlin 1982, pp. 69–88.
98. J. E. Lovelock, *Nature (London)* **256** (1975) 193–194.
99. NAS, *Chloroform, Carbon Tetrachloride and Other Halomethanes: An Environmental Assessment*, Washington, D.C., 1978.
100. J. E. Lovelock, R. J. Maggs, R. J. Wade, *Nature (London)* **241** (1973) 194–196.
101. Erste Allgemeine Verwaltungsvorschrift zur Reinerhaltung der Luft (TA-Luft) vom 28. 2. 1986 (GMBl. **37**, Nr. 7, S. 95–144). 2. Verordnung zur Durchführung des

Bundesimmissionsschutzgesetzes: Verordnung zur Begrenzung der Emission von leichtflüchtigen Halogenkohlenwasserstoffen vom 30. 4. 1986, Bundesgesetzblatt (1986) Nr. 17, S. 571–574.
102. NAS, *Halocarbons: Environmental Effects of Chlorofluoromethane Release*, Washington, D.C., 1976.
103. Ministerium für Ernährung, Landwirtschaft, Umwelt und Forsten, Baden-Württemberg: Leitfaden für die Beurteilung und Behandlung von Grundwasserverunreinigungen durch leichtflüchtige Chlorkohlenwasserstoffe, Heft **13** (1983) und **15** (1984) Stuttgart.
104. Mitteilung des Landesamtes für Wasser und Abfall/Nordrhein-Westfalen, Ergebnisse der Gewässergüte-überwachung, Düsseldorf 1981.
105. Internationale Arbeitsgemeinschaft der Wasserwerke im Rheineinzugsgebiet, Rheinbericht 1981/82, Amsterdam 1984.
106. B. E. Rittmann, P. McCarty, *Appl. Environ. Microbiol.* **39** (1980) 1225–1226.
107. W. Brunner, D. Staub, T. Leisinger, *Appl. Environ. Microbiol.* **40** (1980) 950–958.
108. G. M. Klećka, *Appl. Environ. Microbiol.* **44** (1982) 701.
109. G. Stucki et al., *Arch. Microbiol.* **130** (1981) 366–371.
110. H. H. Tabak et al., *J. Water Pollut. Control Fed.* **53** (1981) no. 10, 1503–1518.
111. J. Halbartschlager et al., *GWF Gas Wasserfach: Wasser/Abwasser* **125** (1984) no. 8, 380–386.
112. T. Leisinger, *Experientica* **39** (1983) 1183–1191.
113. *Bundesgesundheitsblatt* **22** (1979) 102.
114. *SRI International*, Report no. 126: C_1-Chlorinated Hydrocarbons.

General References

115. *Ullmann*, 4th ed., **9**, pp. 420–464.
116. *Kirk-Othmer*, **5**, pp. 714–762; **23**, pp. 764–798, 865–885.
117. L. Scheflan: *The Handbook of Solvents*, D. van Nostrand, New York 1953.
118. G. Hawley: *The Condensed Chemical Dictionary*, Van Nostrand, New York 1977.
119. S. Patai: *The Chemistry of the C-Halogen Bond*, Part 1–2, J. Wiley and Sons, New York 1973.
120. J. L. Blackford: "Ethylene dichloride, 1,1,1-Trichloroethane, Trichloroethylene" in Chemical Economics Handbook, Stanford Research Institute, Menlo Park California 1975.
121. Process Economics Program, Rep. Ser. 5; Supplements 5 A–5 C; *Vinyl Chloride*, Stanford Research Institute, Menlo Park, California 1965–1975.
122. J. S. Naworksi, E. S. Velez in B. E. Leach (ed.): "1,2-Dichloroethane," *Applied Industrial Catalysis*; Academic Press, New York 1983.
123. E. W. Flick: *Industrial Solvents Handbook*, 3rd ed., Noyes Data Corp., Park Ridge, N.J., 1985.
124. J. S. Sconce: *Chlorine*, its Manufacture, Properties and Uses, R. E. Krieger Publ. Co., Malabar, Fla., USA 1982.
125. Stanford Research Institute: *World Petrochemicals*, vol. **2**, Menlo Park, California, USA, 1984.
126. Stanford Research Institute: *Chem. Econ. Handbook, Marketing Research Report on PVC Resins*, Menlo Park, California, USA, 1982.
127. Stanford Research Institute: *Vinyl Chloride Report*, Menlo Park, California, USA, 1982.
128. Process Economic Program, *Chlorinated Solvents*, Rep.-No. 48; Stanford Research Institute, Menlo Park California 1969.
129. M. L. Neufeld et al., Market Input/Output Studies; *Vinylidene Chloride*, U.S. NTIS, PB-273205; Auerbach Corp., Philadelphia 1977.
130. J. A. Key et al., *Organic Chemical Manufacturing*, vol. **8**; U.S. EPA; EPA-450/3–80–028C; IT Envirosci. Inc., Knoxville 1980.
131. F. Asinger, *Die petrolchemische Industrie*, Akademie-Verlag, Berlin (GDR) 1971.

Specific References

132. K. Eisenächer, *Chem. Ing. Tech.* **55** (1983) 786.
133. H. Giesel, *Nach. Chem. Techn. Lab.* **32** (1984) 316.
134. W. Swodenk, *Chem. Ing. Tech.* **55** (1983) 683.
135. J. Schulze, M. Weiser, *Chem. Ind. (Berlin)* **36** (1984) 347.
136. J. D. Schmitt-Tegge, *Chem. Ing. Tech.* **55** (1983) 342.
137. L. H. Horsley, *Azeotropic Data* 1 and 2, no. 6, 35; Adv. Chem. Ser., ACS, Washington, D.C., 1952, 1962.
138. Ethyl Corp., US 2 818 447, 1953 (C. M. Neher).

139. Halcon International Inc., US 3 345 421, 1961 (D. Brown).
140. Shell Dev. Corp., US 2 742 511, 1950 (E. P. Franzen).
141. K. Weissermehl, H.—J. Arpe, *Industrielle Organische Chemie, 5. Auflage* Wiley-VCH, 1998.
142. Dow Chemical, US 2 031 288, 1935.
143. Ethyl Corp., US 2 522 687, 1948 (F. L. Padgitt, G. F. Kirby).
144. Dow Chemical, US 2 469 702, 1946 (C. C. Schwegler, F. M. Tennant).
145. Dow Chemical, US 2 140 927, 1936 (J. E. Pierce).
146. N. N. Lebedew, *Z. Obshch. Chim.* **24** (1954) 1959.
147. Pure Oil Co., CA 464 069, 1950 (D. C. Bond).
148. Marathon Oil Co., US 3 3240 902, 1966 (D. H. Olson, G. M. Bailey).
149. Hoechst, DE-OS 2 026 248, 1970 (H. Großpietsch, H.-J. Arpe, L. Hörnig).
150. Shell Development Co., US 2 246 082, 1939 (W. E. Vaughan, F. F. Rust).
A. W. Fleer et al., *Ind. Eng. Chem.* **47** (1955) 982.
Pet. Refiner **34** (1955) 149.
Hydrocarbon Process. Pet. Refiner **44** (1965) 201.
151. ICI, GB 667 185, 1949 (P. A. Hawkins, R. T. Foster).
152. Ethyl Co., US 2 838 579, 1958 (F. Conrad, M. L. Gould, C. M. Neher).
153. Dow Chemical, US 2 140 547, 1936 (J. H. Reilly).
154. I.G. Farbenindustrie AG, GB 483 051, 1936 (G. W. Johnson).
155. Standard Oil Develop. Co., US 2 393 509, 1946 (F. M. Archibald, H. O. Mottern).
Ethyl Co., US 2 589 698, 1952 (H. O. Mottern, J. P. Russel).
Air Reduction Co. Inc., US 3 506 553, 1970 (L. J. Governale, J. H. Huguet, C. M. Neher).
156. Dow Chemical, US 2 516 638, 1947 (J. L. McCurdy).
Société anon. des Matières Colorantes et Produits Chimiques de Saint-Denis, FR 858 724, 1940.
157. ICI, GB 1 134 116, 1966 (S. C. Carson).
158. H. Bremer et al., DD 85 068, 1971.
159. Socony Mobil Oil Co. Inc., US 3 499 941, 1970 (E. N. Givens, L. A. Hamilton).
160. Hoechst, DE-OS 1 919 725, 1970 (H. Fernholz, H. Wendt).
161. ICI, BE 654 985, 1965.
162. Distillers Co. Ltd., GB 566 174, 1942 (E. G. Galitzenstein).
A. P. Giraitis, *Erdoel Kohle* **9** (1951) 971.
163. Ethyl Corp., US 2 681 372, 1951 (P. W. Trotter).
164. I.G. Farbenindustrie AG, GB 470 817, 1936 (G. W. Johnson).
165. Hercules Powder Co., US 2 084 710, 1935 (H. M. Spurlin).
166. Exxon, US 4 041 087, 1977 (M. A. Vannice).
167. G. A. Olah, EP 73 673, 1983.
168. Dow Chemical, CA 561 327, 1954 (J. H. Brown, W. E. Larson).
169. BP Chemicals Ltd., GB 2 002 357, 1978 (D. A. Burch, E. J. Butler, C. W. Capp).
170. Ethyl Corp., DE 1 518 766, 1965 (A. O. Wikman, L. B. Reynolds, Jr.
171. Dynamit Nobel, DE 1 568 371, 1966 (R. Stephan, H. Richtzenhain).
172. R. N. Rothan, E. W. Sims, *Chem. Ind. (London)* 1970, 830.
173. Donau Chemie, AT 163 818, 1947 (O. Fruhwirth).
Donau Chemie, AT 170 262, 1950 (O. Fruhwirth).
Columbia Southern Chemical, US 2 945 897, 1958 (D. H. Eisenlohr).
174. Vulcan Materials Co., DE 1 668 842, 1963 (J. I. Jordan, H. S. Vierk).
175. Stauffer, US 3 987 118, 1976 (M. A. Kuck).
Union Carbide, US 3 427 359, 1969 (C. E. Rectenwald, G. E. Keller II, J. W. Clark).
176. Donau Chemie, AT 163 218, 1947 (G. Gorbach).
Chem. Eng. News **25** (1947) 1888.
Donau Chemie, AT 166 251, 1949 (O. Fruhwirth).
177. K. A. Holbrook et al., *J. Chem. Soc. B* 1971, 577.
178. S. Inokawa et al., *Kogyo Kagaku Zasshi* **67** (1964) 1540.
179. W. L. Archer, E. L. Simpson, *I and EC Prod. Res. Dev.* **16** (1977) 158.
180. Jefferson Chemical Co., US 2 601 322, 1952 (R. R. Reese).
181. Tokuyama Soda Co., JP 41 3168, 1963.
182. Bayer, DE 1 157 592, 1961 (H. Rathjen, H. Wolz).
183. Dynamit Nobel AG, DE-OS 2 156 190, 1971 (E. Feder, K. Deselaers).
184. Union Carbide, US 2 929 852, 1954 (D. B. Benedict).
Bayer, GB 960 083, 1962.
Wacker, DE-OS 1 668 850, 1967 (R. Sieber, A. Maier);
DE-OS 1 768 367, 1968 (O. Fruhwirth, L. Schmidhammer, E. Pichl).

185. Dynamit Nobel AG, DE-OS 2 128 329, 1974 (A. Hoelle).
Bayer, DE-OS 2 743 975, 1979 (S. Hartung); DE-OS 1 905 517, 1969 (R. Wesselmann, W. Eule, E. Köhler).
186. S. N. Balasubramanian, *Ind. Eng. Chem. Fundam.* **5** (1966) 184.
187. J. C. Vlugter et al., *Chim. Ind. (Milan)* **33** (1951) 613.
Hydrocarbon Process. **44** (1965) 198.
Knapsack AG, DE-OS 1 905 517, 1969.
188. Società Italiana Resine S.p.A., US 3 911 036, 1975 (L. Di Fiore, B. Calcagno).
Union Carbide, US 2 929 852, 1960 (D. B. Benedict).
189. Lummus Co., US 3 917 727, 1975 (U. Tsao); US 3 985 816, 1976 (U. Tsao).
Allied, US 3 941 568, 1976 (B. E. Kurtz, A. Omelian).
Stauffer, DE-OS 2 652 332, 1984 (R. G. Campbell, W. E. Knoshaug).
190. E. Lundberg, *Kem. Tidskr.* **10** (1984) 35.
Hoechst, EP 80 098, 1984 (J. Hundeck, H. Hennen).
191. B. F. Goodrich, GB 1 233 238, 1971 (R. C. Campbell).
Stauffer, US 4 000 205, 1976.
192. Dynamit Nobel AG, DE-OS 3 340 624, 1984 (H. Leuck, H.-J. Westermann).
193. C. M. Schillmoller, *Hydrocarbon Process.* **3** (1979) 77.
194. R. Remirez, *Chem. Eng. (N.Y.)* **75** (1968) no. 9, 142.
Kureha Chemical Ind., NL 6 504 088, 1965.
195. ICI, DE-OS 2 012 898, 1970 (N. Colebourne, P. R. Edwards).
196. F. F. Braconier, *Hydrocarbon Process.* **43** (1964) no. 11, 140.
Société Belge de l'Azote, GB 954 791, 1959 (F. F. A. Braconier, H. Le Bihan).
197. Shell Development Co., US 2 099 231, 1935 (J. D. Ruys, J. W. Edwards).
A. A. F. Maxwell, US 2 441 287, 1944.
198. N. N. Semenov: *Some Problems in Chemical Kinetics and Reactivity*, vol. **1**, Princeton University Press, Princeton 1958, p. 211.
L. F. Albright, *Chem. Eng. (N.Y.)* **74** (1967) no. 7, 123.
199. Montecatini, IT 755 867, 1967 (C. Renato, F. Gianfranco, C. Angelo).
200. R. V. Carrubba, *Thesis Columbia University*, 1968.
R. P. Arganbright, W. F. Yates, *J. Org. Chem.* **27** (1962) 1205.
H. Heinemann, *Chem. Tech. (Heidelberg)* **5** (1971) 287.
201. R. W. McPherson et al., *Hydrocarbon Process.* **3** (1979) 75.
202. J. A. Buckley, *Chem. Eng. (N.Y.)* **73** (1966) no. 29, 102.
P. H. Spitz, *Chem. Eng. Prog.* **64** (1968) no. 3, 19.
E. F. Edwards, F. Weaver, *Chem. Eng. Prog.* **61** (1965) no. 1, 21.
E. M. De Forest, S. E. Penner, *Chem. Eng. (N.Y.)* **79** (1972) no. 17, 54.
203. D. Burke, R. Miller, *Chem. Week* **5** (1964) no. 8, 93.
204. Y. Onoue, K. Sakurayama, *Chem. Eng. Rev.* **4** (1969) no. 11, 17.
205. L. F. Albright, *Chem. Eng. (N.Y.)* **74** (1967) no. 8, 219.
206. E.g. The Distillers Company Ltd., DE-OS 1 443 703, 1970 (A. F. Millidge, C. W. Capp, P. E. Waight).
207. Mitsui Toatsu Chemical, GB 1 189 815, 1970 (K. Miyauchi et al.).
208. VEB Chemische Werke Buna, DD 157 789, 1982 (J. Koppe et al.).
209. BASF, DE-OS 2 651 974, 1978 (P. R. Laurer, J. Langens, F. Gundel).
Società Italiana Resine S.p.A., DE-OS 2 543 918, 1976 (R. Canavesi, F. Ligorati, G. Aglietti).
Produits Chimiques Pechiney-Saint-Gobain, US 3 634 330, 1972 (M. Michel, G. Benaroya, R. Jacques).
Vulcan Materials Co., US 3 926 847, 1975 (W. Q. Beard, Jr., P. H. Moyer, S. E. Penner).
Diamond Shamrock Corp., BE 859 878, 1978 (F. C. Leitert, C. G. VinsonJr.,
M. W. Kellog Co., US 3 114 607, 1963 (T. H. Milliken).
B. F. Goodrich Co., GB 938 824, 1961.
Distillers Co., GB 932 130, 1961 (C. W. Capp, D. J. Hadley, P. E. Waight);
GB 971 996, 1962.
Chem. Werke Hüls AG, FR 1 421 903, 1965.
210. Shell Oil Co., US 3 892 816, 1975 (A. T. Kister).
211. FMC, US 3 360 483, 1963 (L. H. Diamond, W. Lobunez).
Shell Oil Co., US 3 210 431, 1965 (W. F. Engel).
Stauffer, US 3 657 367, 1972 (R. J. Blake, G. W. Roy).
212. Petro-Tex Chemical Corp., US 4 025 461, 1977 (L. J. Croce, L. Bajars, M. Gabliks); 4 046 821, 1977 (L. J. Croce, L. Bajars, M. Gabliks).

213. Toyo Soda Ltd., GB 1 016 094, 1963.
214. Hoechst, FR 1 440 450, 1965.
215. B. F. Goodrich Co., US 3 488 398, 1970 (J. W. Harpring et al.).
216. E. Gorin et al., *Ind. Eng. Chem.* **40** (1948) 2128.
217. Vulcan Materials Co., GB 980 983, 1963 (US 3 184 515, 1962) (S. E. Penner, E. M. De Forest).
218. Dow Chemical, US 2 866 830, 1956 (J. L. Dunn, Jr., B. Posey, Jr.
219. Pechiney, FR 1 286 839, 1961 (F. Lainé, G. Wetroff, C. Kaziz).
220. Stauffer, US 4 206 180, 1980 (R. G. Campbell et al.).
221. Toa Gosei Chemical Ind., US 3 699 178, 1968 (S. Yoshitaka, T. Atsushi, K. Hideo).
222. *Chem. Week* **99** (1966) no. 20, 56.
W. A. Holve et al., *Chem. Anlage + Verfahren* **11** (1972) 69.
223. B. F. Goodrich Co., US 4 226 798, 1980 (J. A. Cowfer et al.).
PPG, GB 1 220 394, 1968 (A. P. Muren, L. W. Piester, R. M. Vancamp).
224. PPG, US 3 679 373, 1972 (R. M. Vancamp, P. S. Minor, A. P. Muren, Jr.
225. B. F. Goodrich Co., DE-OS 1 518 930, 1965 (J. W. Harpring et al.);
1 518 933, 1965 (J. W. Harpring, A. E. van Antwerp, R. F. Sterbenz).
226. Pechiney, GB 959 244, 1962;
US 3 190 931, 1965 (F. Lainé, C. Kaziz, G. Wetroff).
227. Dow Chemical, US 3 966 300, 1976.
228. PPG, GB 2 119 802, 1983 (J. S. Helfand, T. G. Taylor).
229. Hoechst, DE-OS 2 300 844, 1974 (W. Kühn et al.).
Dow Chemical, FR 2 134 845, 1972;
CA 941 329, 1974 (J. E. Panzarella).
BASF, DE-OS 2 400 417, 1975 (D. Lausberg).
Hooker Chemical Co., NL 6 614 522, 1967.
230. Mitsui Toatsu Chemical, JP 45–32406, 1970;
JP 46–43367, 1971;
JP 46–33010, 1971;
GB 1 189 815, 1970 (K. Miyauchi et al.).
PPG, FR 1 220 394, 1971.
P. Reich, *Hydrocarbon Process.* **3** (1976) 85.
R. G. Markeloff, *Hydrocarbon Process.* **11** (1984) 91.
231. Mitsubishi Chemical, DE-OS 2 422 988, 1974 (Y. Kageyama).
232. Stauffer, US 3 892 816, 1972 (A. T. Kister).
233. W. E. Wimer, R. E. Feathers, *Hydrocarbon Process.* **3** (1976) 83.
234. *Chem. Eng. News* **44** (1966) 76.
Chem. Week **99** (1966) no. 13, 93.
235. M. L. Spektor et al., *Chem. Eng. News* **40** (1966) no. 44, 76;
Ind. Eng. Chem. Process Des. Dev. **6** (1967) no. 3, 327.
H. Heinemann, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) no. 6, 400.
F. Friend et al., *Adv. Chem. Ser.* **70** (1968) 168.
Chem. Ind. (Düsseldorf) **19** (1967) no. 3, 124.
236. Distillers Co., GB 958 458, 1962 (C. W. Capp, D. J. Hadley, P. E. Waight).
Union Carbide, BE 664 903, 1965 (J. W. Clark et al.).
PPG, NL 6 401 118, 1964.
National Distillers and Chemical Corp., US 3 720 723, 1973 (E. G. Pritchett).
Pullman Inc., US 3 159 455, 1964 (G. T. Skaperdas, W. C. Schreiner, S. C. Kurzius);
3 536 770, 1970 (G. T. Skaperdas, W. C. Schreiner).
Ethyl Corp., FR 1 398 254, 1964 (M. D. Roof).
237. Monsanto, NL 6 515 254, 1966.
238. Monsanto, NL 6 151 253, 1966.
239. Allied, DE-OS 2 449 563, 1975 (B. E. Kurtz et al.).
240. Shell Development Co., US 2 442 285, 1948 (H. A. Cheney).
241. N. Singer, DD 110 032, 1974.
242. D. H. R. Barton, P. F. Onyon, *J. Am. Chem. Soc.* **72** (1950) 988.
243. Ethyl Corp., US 2 765 350, 1955 (F. Conrad).
244. S. Okazakiand, M. Komata, *Nippon Kagaku Zaishi* 1973, no. 3, 459.
245. N. K. Taikova et al., *Zh. Org. Khim.* **4** (1968) 1880.
246. I. Mochida, Y. Yoneda, *J. Org. Chem.* **33** (1968) 2161.
247. W. B. Crummett, V. A. Stenger, *Ind. Eng. Chem.* **48** (1956) 434.
248. W. L. Archer, E. L. Simpson, *I and EC Prod. Res. and Dev.* **16** (1977) no. 2, 158.
249. L. Bertrand et al., *Int. J. Chem. Kinet.* **3** (1971) 89.
250. Dow Chemical, US 1 870 601, 1932 (E. C. Britton, W. R. Reed).
251. Ethyl Corp., US 3 019 175, 1959 (A. J. Haefner, F. Conrad).

Pechiney-Saint Gobain, FR 1 390 398, 1964 (A. Antonini, C. Kaziz, G. Wetroff);.
Dynamit Nobel, DE-OS 2 026 671, 1970 (H. Richtzenhain, R. Stephan).
252. Z. Přsil, *Radiochem. Radioanal. Lett.* **38** (1979) 103.
253. T. Migita et al., *Bull. Chem. Soc. Japan* **40** (1967) 920;
M. Kosugi et al., *Bull. Chem. Soc. Japan* **43** (1970) 1535;
T. N. Bell et al., *J. Phys. Chem.* **38** (1979) 2321.
254. Dow Chemical, GB 2 121 416, 1983 (J. C. Stevens, J. Perettie).
255. ICI, DE-OS 2 835 535, 1979 (C. S. Allen).
256. A. P. Mantulo et al., DE-OS 3 033 899, 1982; FR 8 020 077, 1980.
257. ICI, DE-OS 1 950 995, 1969 (A. Campbell, R. A. Carruthers).
258. Ethyl Corp., GB 843 179, 1963.
Monsanto, US 3 138 643, 1961 (K. M. Taylor, G. L. Wofford).
PPG, US 3 059 035, 1960 (F. E. Benner, D. H. Eisenlohr, D. A. Reich).
ICI, DE-OS 2 002 884, 1970 (A. Campbell, R. A. Whitelock).
259. Ethyl Corp., GB 1 106 533, 1965;
DE-OS 1 518 766, 1965 (A. O. Wikman, L. B. Reynolds).
260. ICI, DE-OS 1 950 996, 1969 (A. Campbell, R. A. Carruthers).
261. Montecatini Edison SpA, GB 1 170 149, 1967;
FR 1 524 759, 1967 (G. Pregaglia, B. Viviani, M. Agamennone).
262. Dow Chemical, US 3 971 730, 1976 (G. L. Kochanny, Jr., T. A. Chamberlin);
3 872 176, 1975 (G. L. Kochanny, Jr., T. A. Chamberlin).
263. Consortium f. Elektrochem. Ind., US 1 921 879, 1933 (W. O. Herrmann, E. Baum).
Saint Gobain, US 2 674 573, 1949 (M. J. L. Crauland).
Distillers Comp., US 2 378 859, 1942 (M. Mugdan, D. H. R. Barton).
264. I.G. Farbenindustrie, GB 349 872, 1930.
Bataafsche Petroleum M., GB 638 117, 1948.
265. Dow Chemical, US 2 610 214, 1949 (J. L. Amos).
266. Ethyl Corp., US 2 989 570, 1959 (F. Conrad, M. L. Gould).
267. PPG, DE-OS 1 443 033, 1961 (H. J. Vogt).
BASF, DE-OS 1 230 418, 1961 (H. Ostermayer, W. Schweter).
Dynamit Nobel, GB 997 357, 1962.
Feldmühle AG, GB 893 726, 1962.
268. Dow Chemical, US 2 209 000, 1937 (H. S. Nutting, M. E. Huscher).
Solvay, BE 569 355, 1961.
Dynamit Nobel, DE-OS 1 231 226, 1963 (R. Stephan, H. Richtzenhain).
269. FMC, US 3 776 969, 1973 (W. Lobunez).
270. Dynamit Nobel, DE-OS 1 235 878, 1963 (R. Stephan).
271. M. D. Rosenzweig, *Chem. Eng. (N.Y.)* **10** (1971) 105.
272. Vulcan Materials Co., DE-OS 1 518 166, 1963 (J. I. Jordan, Jr., H. S. Vierk);
US 3 304 337, 1967 (J. I. Jordan, Jr., H. S. Vierk).
273. Vulcan Materials Co., DE-OS 2 046 071, 1970 (K. F. Bursack, E. L. Johnston).
274. Ethyl Corp., US 3 012 081, 1960 (F. Conrad, A. J. Haefner).
Société d'Ugine, FR 1 514 963, 1966.
275. Detrex Chem. Ind. Inc., CA 116 460, 1971 (C. E. Kircher).
276. Ugine Kuhlmann, DE-OS 1 668 760, 1971 (A. Goeb, J. Vuillement).
277. W. G. Rollo, A. O'Grady, *Can. Paint and Finish.* **10** (1973) 15.
W. L. Archer, *Met. Prog.* **10** (1974) 133.
R. Monahan, *Met. Finish.* **11** (1977) 26.
P. Goerlich, *Ind.-Lackier-Betr.* **43** (1975) 383.
J. C. Blanchet, *Surfaces* **14** (1975) 51.
L. Skory et al., *Prod. Finish. (Cincinnati)* **38** (1974) .
H. A. Farber, G. P. Souther, *Am. Dyest. Rep.* **57** (1968) 934;
G. P. Souther, *Am. Dyest. Rep.* **59** (1970) 23.
J. J. Willard, *Text. Chem. Color.* **4** (1972) 62.
H. Hertel, *Kunststoffe* **71** (1981) 240.
278. Dow Chemical, US 2 838 458, 1955 (H. J. Bachtel);
2 923 747, 1958 (D. E. Rapp);
2 970 113, 1957 (H. J. Bachtel);
3 049 571, 1960 (W. E. Brown);
3 364 270, 1965 (M. J. Blankenship, R. McCarthy);
3 384 673, 1966 (M. J. Blankenship, R. McCarthy);
3 444 248, 1969 (W. L. Archer, E. L. Simpson, G. R. Graybill);
3 452 108, 1969 (W. L. Archer, E. L. Simpson, G. R. Graybill);
3 452 109, 1969 (W. L. Archer, E. L. Simpson, G. R. Graybill);

3 454 659, 1969 (W. L. Archer, E. L. Simpson);
3 467 722, 1967 (W. L. Archer, G. R. Graybill);
3 468 966, 1969 (W. L. Archer, E. L. Simpson);
3 472 903, 1969 (W. L. Archer, E. L. Simpson);
3 546 305, 1970 (W. L. Archer, E. L. Simpson);
3 681 469, 1972 (W. L. Archer, E. L. Simpson);
4 469 520, 1984 (N. Ishibe, W. F. Richey, M. S. Wing);
GB 916 129, 1961;
NL 6 919 176, 1969.
Ethyl Corp., US 3 002 028, 1958 (A. J. Haefner, L. L. Sims);
3 060 125, 1958 (L. L. Sims);
3 074 890, 1958 (G. N. Grammer);
3 189 552, 1963 (L. L. Sims);
3 238 137, 1961 (G. N. Grammer, P. W. Trotter);
3 629 128, 1968 (J. H. Rains).
Vulcan Materials Co., FR 1 369 267, 1963.
Solvay Cie., BE 743 324, 1969;
755 668, 1970.
PPG, US 3 000 978, 1959 (R. H. Fredenburg);
3 070 634, 1960 (D. E. Hardies, B. O. Pray);
3 128 315, 1961 (D. E. Hardies);
3 192 273, 1961 (W. E. Bissinger);
3 281 480, 1961 (D. E. Hardies);
BE 613 661, 1962 (D. E. Hardies).
ICI, FR 1 372 972, 1963;
1 372 973, 1963;
DE 1 243 662, 1962 (P. Rathbone, Ch. W. Suckling);
1 941 007, 1969 (A. Campbell, P. Robinson, J. W. Tipping);
US 3 336 234, 1964 (J. H. Speight);
GB 1 261 270, 1969 (G. Marsden, J. W. Tipping).
Dynamit Nobel, FR 1 371 679, 1967;
1 393 056, 1965;
DE 1 246 702, 1962 (R. Stephan, H. Richtzenhain).
PCUK, FR 1 555 883, 1967 (J. Vuillemenot).
Diamond Shamrock Corp., DE-OS 2 102 842, 1971 (N. L. Beckers, E. A. Rowe, Jr.

279. D. H. R. Barton, *J. Chem. Soc.* 1949, 148.
280. A. Suzuki et al., *Kogyo Kagaku Zasshi* **69** (1966) 1903.
281. PPG, US 3 344 197, 1967 (Ch. R. Reiche, J. M. Jackson).
282. Toa Gosei Chemical Ind. Co., DE-OS 1 944 212, 1969 (Y. Suzuki, R. Saito);
2 000 424, 1970 (T. Kawaguchi et al.).
283. J. C. Martin, E. H. Drew, *J. Am. Chem. Soc.* **83** (1961) 1232;
M. L. Poutsma, *J. Am. Chem. Soc.* **85** (1963) 3511.
Y. H. Chua et al., *Mech. Chem. Eng. Trans.* **7** (1971) 6.
284. ICI, GB 1 388 660, 1972 (J. S. Berrie, I. Campbell).
Solvay and Cie., DE-OS 2 505 055, 1984.
285. Bataafsche Petroleum, GB 627 119, 1947.
Shell Development Co., US 2 621 153, 1947 (R. H. Meyer, F. J. F. van der Plas).
Toa Gosei Chem. Ind., DE-OS 2 000 424, 1970.
Dow Chemical, US 2 174 737, 1963 (G. H. Coleman, G. V. Moore).
286. British Celanese Ltd., GB 599 288, 1945 (P. J. Thurman, J. Downing).
287. Dow Chemical, US 2 140 549, 1937 (J. H. Reilly).
288. PPG, US 3 065 280, 1960 (H. J. Vogt).
Hoechst, FR 1 318 225, 1962.
PPG, US 3 173 963, 1960 (Ch. R. Reiche, H. J. Vogt).
FMCI, FR 1 491 902, 1966 (S. Berkowitz, A. R. Morgan, Jr.
289. Produits Chimiques Péchiney-Saint-Gobain, FR 1 552 820, 1969 (G. Benaroya, M. Long, F. Lainé);
1 555 518, 1969 (A. Antonini, P. Joffre, F. Lainé).
Rhône-Poulenc Ind., US 4 057 592, 1977 (A. Antonini, P. Joffre, F. Lainé).
290. Consortium Elektrochem. Ind., FR 690 767, 1930.
I.G. Farben, FR 690 655, 1930;
DE 489 454, 1927.
British Celanese Ltd., GB 571 370, 1943 (P. J. Thurman, J. Downing).
Hoechst, FR 1 318 225, 1962.
291. Asahi Kasei Kogyo, FR 1 417 810, 1964.
Vulcan Materials Corp., GB 980 983, 1963.
Central Glass Co., JP 42–9924, 1965.
292. Produits Chimiques Péchiney-Saint-Gobain, FR 1 552 821, 1969 (A. Antonini, G. du Crest, G. Benaroya);
1 552 824, 1969 (A. Antonini, P. Joffre, C. Vrillon).
293. Donauchemie, DE 865 302, 1944 (O. Fruhwirth).
Du Pont, US 2 461 142, 1944 (O. W. Cass).

294. Knapsack-Griesheim AG, DE 939 324, 1954 (K. Sennewald, F. Pohl, H. Westphal).
295. D. H. R. Barton, K. E. Howlett, *J. Chem. Soc.* 1951, 2033.
296. Toa Gosei Chem. Ind. Ltd., FR 2 057 606, 1971;
US 3 732 3422, 1973 (T. Kawaguchi).
Asahi Glass Co., JP 75–34 0003, 1975.
Detrex Chemical Ind., US 3 304 336, 1967 (W. A. Callahan).
D. Gillotay, J. Olbregts, *Int. J. Chem. Kinet.* **8** (1976) 11.
FMC, DE-OS 1 928 199, 1969 (S. Berkowitz).
297. V. A. Poluektov et al., SU 195 445, 1976.
298. I. G. Farben DE 530 649 (1929).
299. L. E. Horsley, Azeotropic Data in *Adv. Chem. Ser.* **6**, Am. Chem. Soc., Washington, D.C., 1952.
300. S. Tsuda, *Chem. Eng. (N.Y.)* **74** (May 1970).
301. ICI, FR 2 003 816, 1969.
302. K. S. B. Prasa, L. K. Doraiswamy, *J. Catal.* **32** (1974) 384;
N. N. Lebedev et al., *Kinet. Katal.* **12** (1971) 560;
J. A. Pearce, *Can. J. Res.* **24 F** (1946) 369.
303. I.G. Farben AG, FR 836 979, 1939.
304. G. Brundit et al., Bios Report, 1056 (1947).
G. B. Carpenter, Fiat Final Report, 843 (1947).
305. Wacker, DE 733 750, 1940.
306. Hooker Chemical Co., NL 6 614 522, 1965.
307. T. Kawaguchi et al., *Ind. Eng. Chem.* **62** (1970) 36.
308. D. S. Caines et al., *Aust. J. Chem.* **22** (1969) 1177;
R. G. McIver, J. S. Ratcliffe, *Trans. Inst. Chem. Eng.* **51** (1973) 68.
309. Toa Gosei Chem. Ind. Co. Ltd., FR 2 057 605, 1971.
310. Toa Gosei Chem. Ind. Co. Ltd., FR 2 057 604, 1971.
311. Produits Chimiques, Péchiney-Saint-Gobain S.A., DE-OS 1 964 552, 1975 (J. C. Strini, Y. Correia);
1 964 551, 1975 (Y. Correia);
US 4 148 832, 1979 (Y. Correia);
DE-OS 1 817 193, 1975 (Y. Correia, J. C. Strini);
1 817 191, 1975 (Y. Correia, J. C. Strini).
312. Kanto Denka Kogyo Ltd., DE-OS 1 568 912, 1966 (A. Suzuki et al.).
313. Wacker, DE-OS 2 023 455, 1971 (L. Schmidhammer, O. Fruhwirth);
2 320 915, 1974 (L. Schmidhammer, D. Dempf, O. Sommer);
2 239 052, 1974 (S. Nitzsche, L. Schmidhammer).
314. Monsanto Chem. Co., US 2 846 484, 1954 (J. E. Fox).
Produits Chimiques Péchiney-Saint-Gobain, DE-OS 1 768 485, 1968 (G. du Crest, G. Benaroya, F. Lainé);
1 768 494, 1968 (A. Antonini, P. Joffre, C. Vrillon);
1 768 495, 1968 (A. Antonini, C. Kaziz, G. Wetroff);
FR 1 552 825, 1969 (A. Antonini, C. Kaziz, G. Wetroff);
1 552 826, 1969 (A. Antonini, C. Kaziz, G. Wetroff).
315. Solvay Cie., FR 1 380 970, 1964.
316. Japan Atomic Energy Research Inst., JP 42–13842, 1967.
317. Knapsack-Griesheim, GB 749 351, 1954.
318. G. Kalz, *Plaste Kautsch.* **18** (1971) 500.
319. T. J. Houser, R. B. Bernstein, *J. Am. Chem. Soc.* **80** (1958) 4439;
T. J. Houser, T. Cuzcano, *Int. J. Chem. Kinet.* **7** (1975) 331.
320. Wacker, DE 843 843, 1942 (W. Fritz, E. Schaeffer).
321. F. S. Dainton, K. J. Ivin, *Trans. Faraday Soc.* **46** (1950) 295;
J. Puyo et al., *Bull. Soc. Lorraine Sci.* **2** (1962) 75.
322. Kali-Chemie, DE 712 784, 1938 (F. Rüsberg, E. Gruner).
323. Du Pont, US 2 440 731, 1948 (W. H. Vining, O. W. Cass).
324. *Nachr. Chem. Tech. Lab.* **29** (1981) 10.
325. Chemische Fabrik Griesheim, DE 278 249, 1912.
326. D. Hardt et al., *Angew. Chem.* **94** (1982) 159; *Int. Ed.* **21** (1982) 174.
327. E. Sanhueza et al., *Chem. Rev.* **75** (1976) 801.
328. M. Lederer, *Angew. Chem.* **71** (1959) 162.
329. H. Normant, *Compt. Rend.* **239** (1954) 1510;
Bull. Soc. Chim. Fr. 1957, 728;
Adv. Org. Chem. **2** (1960) 1.
330. R. West, W. H. Glaze, *J. Org. Chem.* **26** (1961) 2096.
331. Dow Chemical, US 4 147 733, 1979 (T. R. Fiske, D. W. Baugh, Jr.
332. R. E. Lynn, K. A. Kobe, *Ind. Eng. Chem.* **46** (1954) 633;
Société Belge de l'Azote, US 2 779 804, 1954 (F. F. A. Braconier, J. A. R. O. L. Godart).

333. Société Belge de l'Azote, GB 954 791, 1959 (F. F. A. Braconier, H. Le Bihan).
334. Chemische Werke Hüls AG, GB 709 604, 1957;
BASF, GB 769 773, 1955;
Produits Chimiques de Péchiney-Saint-Gobain, FR 1 361 884, 1963 (F. Lainé, C. Kaziz, G. Wetroff).
335. Knapsack AG, DE-OS 2 053 337, 1972 (A. Lauke).
336. F. J. Gattys-Verfahrenstechnik GmbH, DE-OS 2 646 129, 1979 (F. J. Gattys).
337. D. H. R. Barton, M. Mugdan, *J. Soc. Chem. Ind.* **69** (1950) 75;
F. Patat, P. Weidlich, *Helv. Chim. Acta* **32** (1949) 783.
338. Yu. A. Pasderskii, SU 686 279, 1984.
339. K. Washimi, Y. Wakabayashi, *Kogyo Kagaku Zasshi* **68** (1965) 113;
R. D. Wesselhoft et al., *AIChEZ.* **5** (1959) 361.
340. VEB Chemische Werke Buna, DD 159 985, 1981 (J. Glietsch et al.).
341. VEB Chemiefaserwerk Friedrich Engels, DD 150 880, 1981 (H. E. Steglich et al.);
200 017, 1983 (H. E. Steglich et al.);
VEB Chemische Werke Buna, DD 126 454, 1977 (G. Henke);
132 711, 1978 (H. Stolze et al.);
149 212, 1981 (J. Glietsch, W. Linke).
342. Marathon Oil Co., GB 1 138 669, 1969.
343. Monsanto, GB 600 785, 1945;
757 661, 1954.
344. Air Reduction Co., US 2 448 110, 1946 (H. S. Miller).
345. Gevaert Photo Production N.V., GB 655 424, 1947.
346. Gevaert Photo Production N.V., GB 643 743, 1947.
347. Wacker, DE-OS 1 277 845, 1968 (O. Fruhwirth, H. Kainzmeier).
348. VEB Chemische Werke Buna, DD 139 976, 1980 (R. Adler et al.).
349. Hoechst, DE 2 558 871, 1984 (W. Gerhardt, H. Scholz).
350. VEB Chemische Werke Buna, DD 139 975, 1980 (K. Hartwig et al.);
143 367, 1980 (K. Hartwig et al.).
351. Chemische Werke Hüls AG, DE 1 205 705, 1964 (W. Knepper, G. Höckele);
DE-OS 2 054 102, 1970 (H. Maiwald, G. Höckele, H. Sauer).
352. H. Bremer et al., DD 84 182, 1971.
353. Institut für Chemieanlagen, FR 1 441 148, 1966;
DD 51 850, 1968 (K. Roland, C. Gerber, G. Voigt).
354. Institut für Chemieanlagen, GB 1 174 147, 1969;
FR 1 553 573, 1968;
DD 60 303, 1968 (K. Roland).
355. Institut für Chemieanlagen, DD 50 593, 1966 (K. Roland);
DE 1 260 416, 1968 (K. Roland).
356. Grupul Industrial de Petrochimie Borzesti, FR 7 202 582, 1972 (T. Has et al.);
Knapsack AG, DE-OS 1 254 143, 1967 (G. Rechmeyer, A. Jacobowsky).
357. Kureha Kagaku Kogyo Kabushiki Kaisha, GB 1 149 798, 1969;
FR 1 558 893, 1968.
Japan Gas-Chemical Co. Inc., FR 1 465 296, 1967.
358. Solvay and Cie., BE 698 555, 1967;
US 3 506 727 (J. Mulders).
359. N. L. Volodin et al., DE-OS 2 026 429, 1975;
US 4 014 947, 1977;
FR 2 045 822, 1971.
360. B. J. Pope, US 3 864 409, 1975.
361. S. Gomi, *Hydrocarbon Process.* **43** (1964) 165;
K. Washimi, M. A. Kura, *Chem. Eng. (N.Y.)* **73** (1966) no. 10, 133; no. 11, 121;
Y. Onoue, K. Sakurayama, *Chem. Econ. Eng. Rev.* **4** (1969) 17.
362. F. J. Gattys Ingenieurbüro, DE-OS 2 820 776, 1980 (F. J. Gattys);
DE 2 905 572, 1983 (F. J. Gattys).
363. H. J. Pettelkau, DE 3 007 634, 1982.
364. K. E. Howlett, *Trans. Faraday Soc.* **48** (1952) 25;
L. K. Doraiswamy et al., *Br. Chem. Eng.* **5** (1960) 618;
G. A. Kapralova, N. N. Semenov, *Zh. Fiz. Khim.* **37** (1963) 73;
P. G. Ashmore et al., *J. Chem. Soc., Faraday Trans. 1* 1982, 657.
365. Chemische Werke Hüls AG, DE-OS 2 130 297, 1975 (G. Scharein, J. Gaube).
366. E. V. Sonin et al., FR 2 082 004, 1971;
GB 1 225 210, 1969;
DE 1 953 240, 1984.
367. Allied, DE-OS 2 319 646, 1973 (B. E. Kurtz et al.).
368. BASF, DE-OS 2 349 838, 1974 (G. Krome).
369. BP Chemicals Intern. Ltd., GB 1 337 326, 1973 (N. F. Chisholm).
370. BP Chemicals Ltd., FR 2 099 466, 1972 (D. P. Young);
DE-OS 2 135 248, 1972 (D. P. Young);
US 3 896 182, 1975.

371. Mitsui Chem. Ind., JP 42 22921, 1967.
372. BP Chemicals Ltd., GB 1 494 797, 1977 (R. W. Rae, W. F. Fry).
373. Magyar Asvanyolaj es Földgaz Kiserleti Intezet, DE-OS 2 223 011, 1973 (L. Szepesy et al.);
2 225 656, 1973 (I. Vendel et al.);
Hoechst, NL 7 503 850, 1974;
Knapsack AG, DE-OS 2 313 037, 1974 (G. Rechmeier, W. Mittler, R. Wesselmann);
B. F. Goodrich Co., GB 938 824, 1961.
374. BASF, DE-OS 3 147 310, 1973 (W. Hebgen et al.).
375. Hoechst, DE-OS 2 907 066, 1980 (A. Czekay et al.);
3 013 017, 1981 (R. Krumböck et al.);
Hoechst and Uhde GmbH/Hoechst AG, EP 14 920, 1980 (A. Czekay et al.);
21 381, 1980 (G. Link et al.).
376. Halcon Intern. Inc., FR 1 505 735, 1967 (B. J. Ozero);
Knapsack AG, NL 6 612 668, 1967;
Knapsack AG, DE-OS 1 250 426, 1966 (H. Krekeler et al.);
1 910 854, 1972 (G. Rechmeier, A. Jacobowsky, P. Wirtz);
The Lummus Co., DE-OS 2 501 186, 1975 (R. Long, H. Unger);
Hoechst AG, DE 3 024 156, 1983 (A. Czekay et al.);
BASF, DE-OS 3 140 892, 1983 (W. Hebgen et al.).
377. Hoechst, GB 2 054 574, 1981 (J. Riedl, W. Fröhlich, E. Mittermaier).
378. PPG, NL 6 613 177, 1967;
BASF, DE 3 219 352, 1984 (E. Birnbaum, E. Palme).
379. Solvay and Cie., BE 746 270, 1970;
DE-OS 1 288 594, 1969 (G. Coppens);
2 101 464, 1971 (G. Coppens).
380. Knapsack AG, DE-OS 1 959 211, 1971 (P. Wirtz et al.);
Kanegafuchi Kagaku Kogyo K.K., DE-OS 2 426 514, 1975 (T. Ohishi, N. Yoshida, T. Hino);
Allied, NL 7 711 904, 1977;
Wacker, DE-OS 2 754 891, 1979 (L. Schmidhammer, H. Frey);
BASF, 2 307 376, 1974 (G. Krome);
Dynamit Nobel AG, DE 3 135 242, 1984; (R. Stephan et al.);
BASF, DE-OS 3 140 447, 1983 (W. Hebgen et al.).
381. ICI, GB 1 405 714, 1975;
Wacker, DE-OS 3 009 520, 1981 (L. Schmidhammer, R. Straßer);
Solvay and Cie., EP 101 127, 1983 (A. Closon).
382. Deutsche Gold- und Silberschmiedeanstalt, DE-OS 2 438 153, 1976 (G. Vollheim et al.).
383. Solvay and Cie, FR 1 602 522, 1970;
US 3 801 660, 1974 (G. Coppens).
384. Dow Chemical, US 3 723 550, 1973 (R. T. McFadden).
385. Monsanto, US 3 125 607, 1964 (H. M. Keating, P. D. Montgomery);
Hoechst, DE-OS 2 903 640, 1980 (G. Rechmeier, U. Roesnik, H. Scholz).
386. Monsanto, US 3 142 709, 1964 (E. H. Gause, P. D. Montgomery).
387. Monsanto, US 3 125 608, 1964 (D. W. McDonald).
388. Continental Oil Co., US 3 830 859, 1974 (R. Gordon et al.).
389. VEB Chemische Werke Buna, DD 143 368, 1980 (H. Hauthal et al.).
390. Rhône-Progil S.A., DE-OS 2 429 273, 1976 (Y. Correia, J.- C. Lanet).
391. Rhône-Progil, FR 2 241 519, 1973 (Y. Correia, J.-C. Lanet).
392. BASF, DE-OS 3 122 181, 1982 (E. Danz, G. Krome).
393. Goodyear Fire and Rubber Co., US 4 042 637, 1977 (E. J. Glazer, E. S. Smith).
394. M. Rätzsch et al., DD 112 603, 1975.
395. Continental Oil Co., US 3 917 728, 1975 (R. D. Gordon).
396. *Hydrocarbon Process.* **44** (1965) 290.
397. Wacker, DE 1 135 451, 1960 (O. Fruhwirth).
398. Wacker, DE-OS 2 156 943, 1973 (O. Fruhwirth, L. Schmidhammer, H. Kainzmeier).
399. Pullman Corp., GB 1 152 021, 1969;
Monsanto, FR 1 515 554, 1968 (R. L. Hartnett).
400. Wacker, DE-OS 2 239 051, 1975 (S. Nitzsche, L. Schmidhammer);
Shell Internat. Research, NL 6 610 116, 1968.
401. FMC, DE-OS 1 928 199, 1969 (S. Berkowitz).
402. B. F. Goodrich Co., EP 0 002 021, 1979 (A. J. Magistro).
403. The Lummus Co., DE-OS 1 806 547, 1976 (H. D. Schindler, H. Riegel);
2 502 335, 1975 (H. D. Schindler).
404. J. Wolfrum et al., DE 2 938 353, 1983;
EP 0 027 554, 1981.
K. Kleinnermanns, J. Wolfrum, *Laser Chem.* **2** (1983) 339.

405. Kanegafuchi Kagaku Kogyo Kabushiki Kaisha, FR 1 556 912, 1969.
406. M. Sittig: *Vinyl Chloride and PVC Manufacture*, Noyes Data Corp., Park Ridge, New York 1978.
407. Akzo, US 4 256 719, 1981 (E. van Andel); E. van Andel, *Chem. Ind. (London)* 1983, no. 2, 139.
408. *European Chem. News* **46** (1229) June 9, p. 4, 1986.
409. Montecatini S.G., IT 761 015, 1967 (M. Mugero, M. Boriningo).
410. Hitachi Zosen Kabushiki Kaisha, NL 6 604 833, 1966.
411. Badger Co. Inc, DE-OS 2 629 461, 1978 (H. R. Sheely, F. F. Oricchio, D. C. Ferrari).
Pullmann Inc., GB 1 177 971, 1970 (B. E. Firnhaber).
412. Ethyl Corp., US 2 681 372, 1951 (P. W. Trotter);
Hoechst, DE 1 003 701, 1954;
National Distillers, US 2 896 000, 1955 (F. D. Miller, D. P. Jenks);
Société Belge de l'Azote, GB 889 177, 1959 (P. J. Leroux, F. F. A. Braconier);
ESSO Research and Engin. Co., FR 1 344 322, 1962 (J.-M. Guilhaumou, M. Prillieux, P. Verrier);
S. Tsutsumi, GB 956 657, 1962;
Union Carbide, FR 1 385 179, 1964;
Toyo Koatsu Ind., BE 637 573, 1963;
658 457, 1965 (T. Takahashi et al.).
413. ESSO Research and Engin Co., US 3 670 037, 1972 (J. J. Dugan);
ICI, DE-OS 2 837 514, 1979 (J. D. Scott);
Shell Intern. Research, NL 6 611 699, 1968;
Allied, US 4 039 596, 1977 (W. M. Pieters, E. J. Carlson);
Du Pont, US 2 308 489, 1940 (O. W. Cass);
ICI, FR 1 330 367, 1962;
1 359 016, 1963;
Hoechst, BE 662 098, 1965;
Osaka Kinzoku Kogyo, US 3 267 161, 1963 (R. Ukaji et al.).
414. ICI, DE-OS 1 277 243, 1968 (B. Hancock, L. McGinty, I. McMillan).
415. Diamond Shamrock Corp., US 4 115 323, 1978 (M. F. Lemanski, F. C. Leitert, C. G. Vinson, Jr.
416. Hoechst, DE-OS 1 931 393, 1971 (H. Krekeler, H. Kuckertz);
British Petroleum Co. Ltd., DE-OS 1 907 764, 1972 (G. H. Ludwig);
GB 1 213 402, 1970 (G. H. Ludwig);
Marathon Oil Co., US 3 501 539, (D. H. Olson, G. M. Bailey).
417. Produits Chimiques Péchiney-Saint-Gobain, FR 1 552 849, 1969 (A. Antonini, G. Stahl, C. Vrillon).
418. British Petroleum Co. Ltd., DE-OS 2 540 067, 1976 (J. L. Barclay).
B. F. Goodrich Co., DE-OS 2 613 561, 1976 (W. J. Kroenke, P. P. Nicholas);
US 4 100 211, 1978 (A. J. Magistro);
4 102 935, 1978 (W. J. Kroenke, R. T. Carroll, A. J. Magistro);
4 102 936, 1978 (A. J. Magistro);
4 119 570, 1978 (W. J. Kroenke, P. P. Nicholas);
CA 1 096 406, 1981 (A. J. Magistro).
Monsanto, DE-OS 2 852 036, 1979 (T. P. Li);
CA 1 111 454, 1981 (T. P. Li).
ICI, GB 2 095 242, 1982 (D. R. Pyke, R. Reid);
2 095 245, 1982 (D. R. Pyke, R. Reid);
DE-OS 3 226 028, 1983 (D. R. Pyke, R. Reid).
419. Owens-Illinois Inc., US 4 042 639, 1977 (T. H. Gordon, H. F. Kummerle);
Princeton Chemical Research Inc., FR 1 595 619, 1970 (N. W. Frisch, R. I. Bergmann).
420. O. A. Zaidman et al., DE-OS 2 853 008, 1980;
GB 2 036 718, 1980.
421. Princeton Chemical Research Inc., DE-OS 1 806 036, 1969 (R. I. Bergmann);
Produits Chimiques Péchiney-Saint-Gobain, US 3 923 913, 1975 (A. Antonini et al.).
422. Monsanto, NL 6 515 252, 1966.
423. Ethyl Corp., US 3 658 933, 1972 (W. Q. Beard, Jr.
3 658 934, 1972 (W. Q. Beard, Jr.
3 629 354, 1971 (W. Q. Beard, Jr.
424. The Lummus Co., DE-OS 1 693 042, 1974 (H. Riegel);
1 812 993, 1974 (H. Riegel, H. Schindler);
1 952 780, 1970 (H. Riegel et al.);
2 230 259, 1973 (H. Riegel);
2 314 786, 1973 (U. Tsao);
2 335 949, 1974 (H. Riegel et al.);
2 336 497, 1974 (H. D. Schindler et al.);
2 509 966, 1975 (H. Riegel);
2 536 286, 1976 (U. Tsao);
US 3 920 764, 1975; (H. Riegel et al.);
3 935 288, 1976 (H. Riegel);
3 937 744, 1976 (H. Riegel);
3 796 641, 1974 (H. Riegel et al.);
FR 1 574 064, 1969 (M. C. Sze);

1 574 705, 1969 (H. Riegel);
1 576 909, 1969 (H. Riegel, H. D. Schindler);
1 577 105, 1969 (H. Riegel);
GB 1 258 750, 1971.
425. Ethyl Corp., US 2 838 579, 1954 (F. Conrad, M. L. Gould, C. M. Neher);
Monsanto, US 3 166 601, 1961 (K. M. Taylor);
Du Pont, US 3 234 295, 1961 (J. W. Sprauer, S. Heights).
426. PPG, GB 996 323, 1962 (W. K. Snead, R. H. Chandley);
998 689, 1965;
FR 1 341 711, 1962 (A. C. Ellsworth);
Princeton Chemical Research Inc., DE-OS 1 929 062, 1969 (N. W. Frisch);
The British Petroleum Co. Ltd., DE-OS 2 006 262, 1971 (G. H. Ludwig).
427. Ethyl Corp., US 2 765 349–352, 1955 (F. Conrad);
2 803 677 1955 (C. M. Neher, J. H. Dunn);
2 803 678–680 1955 (F. Conrad).
428. Air Reduction Co. Inc., US 3 506 552, 1970 (J. P. Russell).
429. Allied, US 4 155 941, 1979 (H. E. Nychka, R. E. Eibeck).
430. PCUK Produits Chimiques Ugine Kuhlmann, FR 2 529 883, 1982 (G. A. Olah).
431. ICI, EP 15 665, 1980 (C. S. Allen).
432. PPG, GB 2 121 793, 1982 (D. R. Nielsen).
433. Central Glass Co. Ltd., DE 2 818 066, 1982 (K. Yagii, H. Oshio).
434. Amer., US 3 558 453, 1971 (J. S. Mayell).
435. ICI, FR 2 064 406, 1971;
ZA 706 583, 1969 (C. Neville, P. R. Edwards, P. J. Craven).
436. Gulf R a D Co., US 3 577 471, 1971 (J. G. McNulty, W. L. Walsh).
437. Seymour C. Schuman, US 3 377 396, 1968.
438. SRI International, *Chem. Econ. Handbook*, Marketing Research Report on PVC Resins, Menlo Park, 1982.
SRI International, *Vinyl Chloride Report*, Menlo Park, 1982.
Eur. Chem. News **44** (1985) no. 1169.
439. *Eur. Chem. News* **45** (1985) no. 1194, 4.
Chem. Week (1985) 30.
440. *Hydrocarbon Process., HPI Construction Boxscore, October 1984; Eur. Chem. News* **44** (1985) no. 1167.
441. Dow Chemical, US 2 136 333, 1936 (G. H. Coleman, J. W. Zemba);
2 136 349, 1938 (R. M. Wiley);
2 160 944, 1938 (G. H. Coleman, J. W. Zemba);
Ethyl Corp., FR 1 308 101, 1961 (A. J. Haefner, E. D. Hornbaker).
442. R. J. Williams, *J. Chem. Soc.* 1953, 113;
K. Feramoto et al., *Kogyo Kagaku Zasshi* **67** (1964) 50.
443. J. Svoboda et al., *Petrochemia* **22** (1982) 21.
444. BASF, DE-OS 1 230 418, 1966 (H. Ostermayer, W. Schweter).
445. ICI, DE-OS 2 225 512, 1972 (G. A. Thompson, J. W. Tipping).
446. ICI, GB 1 453 509, 1976 (I. S. McColl, A. C. P. Pugh, G. A. Thompson).
447. Detrex Chemical Industries Inc., US 3 725 486, 1973 (W. L. McCracken et al.).
448. Toa Gosei Chem. Ind. Co., Ltd., JP 56/104 826, 1981.
449. PPG, DE-OS 2 651 901, 1980 (J. D. Mansell).
450. S. A. Devinter, FR 2 131 896, 1973.
451. Continental Oil Co., US 3 664 966, 1972 (R. D. Gordon).
452. Kanegafuchi Chemical Industry Co., Ltd., JP 48/92312, 1973.
453. Dow Chemical, US 3 984 489, 1976 (R. F. Mogford).
454. Continental Oil Co., US 3 869 520, 1975 (R. D. Gordon).
455. Dow Chemical, US 2 293 317, 1941 (F. L. Taylor, L. H. Horsley).
456. Asahi-Dow Ltd., JP 56/57722, 1981.
457. E. W. Sonin et al., DE-OS 1 952 770, 1973.
458. K. Nagai, M. Katayama, *Bull. Chem. Soc. Japan* **51** (1978) 1269.
459. Chemische Werke Hüls AG, DE-OS 2 135 908, 1974 (H. Rassaerts, G. Sticken, W. Knepper).
460. Dow Chemical, US 4 119 678, 1978 (T. S. Boozalis, J. B. Ivy).
461. Mitsui Toatsu Chemicals Co., Ltd., JP 47/18088, 1972.
462. Wah Young Lee et al., *Hwahak Konghak* **14** (1976) 169;
Chem. Abstr. **85** 142517 c.
463. I. Mochida et al., *Sekiyu Gakkai Shi* **21** (1978) 285;
Chem. Abstr. **90** 22226 p.
464. I. Mochida et al., JP 51/133207, 1976.
465. ICI, DE-OS 2 257 107, 1973 (M. H. Stacey, T. D. Tribbeck);
PPG, DE-OS 2 849 469, 1979 (W. H. Rideout);
US 4 144 192, 1979 (A. E. Reinhardt II).

466. Asahi-Dow Ltd., JP 55/87730, 1980; Toa Gosei Chemical Industry Co., Ltd., JP 52/31006, 1977.
467. Asahi-Dow Ltd., JP 55/87729, 1980.
468. A. P. Khardin et al., *Khim. Prom-st. Moscow* 1982, 208;
P. Y. Gokhberg et al., *Kinet. Katal.* **23** (1982) 50–53, 469–473;
A. P. Khardin et al., *Khim. Prom-st. (Moscow)* 1981, 16.
469. N. Yamasata, *Ibaraki Kogyo Koto Semmon Gakko Kenkyu Iho* **11** (1976) 213;
Chem. Abstr. **89** 214 828 n.
470. I. Mochida, *J. Mol. Catal.* **12** (1981) 359;
Asahi-Dow Ltd., JP 56/40622, 1981;
58/162537, 1983;
56/150028, 1981;
56/40622, 1981.
471. J. Kobayashi et al., *Kenkyu Hokoku–Asahi Garasu Kogyo Gijutsu Shoreikai* **37** (1980) 251;
Chem. Abstr. **95** 6055 e.
472. Kureha Chemical Industry Co., Ltd., JP 48/10130, 1973.
473. L. Espada et al., *Ion* **37** (1977) 495.
474. Wacker, DE-OS 1 925 568, 1974 (O. Fruhwirth, E. Pichl, L. Schmidhammer);
Kureha Kagaku Kogyo K.K., DE-OS 2 101 463, 1971 (K. Shinoda et al.);
2 135 445, 1973 (K. Shinoda, T. Nakamura).
475. Asahi-Dow Ltd., JP 56/57721, 1981.
476. ICI, DE-OS 2 850 807, 1979 (J. D. Wild).
477. Dow Chemical, US 3 726 932, 1973 (C. R. Mullin, D. J. Perettie).
478. Dow Chemical, US 3 654 358, 1977 (G. C. Jeffrey).
479. Donau-Chemie, AT 162 391, 1947.
480. Knapsack-Griesheim AG, DE 969 191, 1952 (A. Jacobowsky, K. Sennewald);
DE-OS 1 011 414, 1953 (A. Jacobowsky, K. Sennewald);
Péchiney, US 3 197 515, 1962 (P. Chassaing, G. Clerc).
481. B. Jakesevic, *Tekstil* **32** (1983) 321.
482. C. J. Howard, *J. Chem. Phys.* **65** (1976) 4771;
J. S. Chang, F. Kaufmann, *J. Chem. Phys.* **66** (1977) 4989.
483. L. Bertrand et al., *J. Phys. Chem.* **72** (1968) 3926;
J. Olbregts, *Int. J. Chem. Kinet.* **11** (1979) 117.
484. Wacker, DE 901 774, 1940 (W. Fritz, J. Rambausek).
485. Dynamit Nobel, DE 1 174 764, 1958 (E. E. Feder, K. Kienzle);
Detrex Chemical Ind., US 2 912 470, 1956 (C. E. Kircher, Jr., R. J. Jones);
M. Szczeszek, H. Chmielarska, *Przem. Chem.* **56** (1977) 255.
486. Wacker, DE 846 847, 1942 (W. Fritz, E. Schaeffer);
NL 6 600 884, 1966.
487. Du Pont, US 2 894 045, 1957 (E. G. Carley);
Hooker Chemical Co., US 3 100 233, 1960 (D. S. Rosenberg);
Donau-Chemie AG, AT 238 699 (F. Samhaber).
488. Donau-Chemie AG, AT 191 859, 1957 (L. Gavanda, A. F. Orlicek);
ICI, GB 697 482, 1951 (R. T. Foster, S. W. Frankish).
489. The Distillers Co., US 2 378 859, 1942 (M. Mugdan, D. H. R. Barton).
490. Du Pont, US 3 388 176, 1968 (J. L. Sheard);
3 388 177, 1968 (L. J. Todd).
491. Produits Chimiques Péchiney-Saint-Gobain, DE-OS 2 008 002, 1975 (R. Clair, Y. Correia);
Toa Gosei Chem. Ind., DE-OS 1 943 614, 1969 (T. Kawaguchi et al.).
492. Detrex Chemical Ind. Inc., DE-OS 1 643 872, 1967 (C. E. Kirchner, Jr., D. R. McAllister, D. L. Brothers).
493. S. Suda, *Chem. Eng. (N.Y.)* **77** (1970) 74.
T. Kawaguchi et al., *Ind. Eng. Chem.* **62** (1970) 36.
494. Detrex Chemical Ind. Inc., US 3 691 240, 1972 (C. E. Kirchner, D. R. McAllister, D. L. Brothers).
495. Dow Chemical, US 2 140 548, 1938 (J. H. Reilly);
Diamond Alkali Corp., GB 673 565, 1952;
Ethyl Corp., US 2 725 412, 1954 (F. Conrad);
Donau Chemie AG, NL 6 607 204, 1966.
496. Diamond Shamrock Corp., DE-OS 2 061 508, 1970 (J. J. Lukes, R. J. Koll).
497. Ruthner AG, AT 305 224, 1973 (F. Samhaber);
PPG, US 3 793 227, 1974 (W. K. Snead, F. Abraham).
498. D. Jaqueau et al., *Chem. Ing. Tech.* **43** (1971) 184;
J. F. Knoop, G. R. Neikirk, *Hydrocarbon Process.* **59** (1972) 109;
PPG, GB 913 040, 1961 (L. E. Bohl, R. M. Vancamp); 904 084, 1961 (L. E. Bohl);
904 405, 1961 (L. E. Bohl, A. P. Muren, R. M. Vancamp);

916 684, 1961 (R. E. McGreevy, J. E. Milam, W. E. Makris);
969 416, 1962 (A. C. Ellsworth;
1 012 423, 1964 (L. E. Bohl, R. M. Vancamp);
1 027 279, 1963 (L. E. Bohl, R. M. Vancamp);
1 104 396, 1965 (L. W. Piester, R. M. Vancamp);
Du Pont, GB 1 362 212, 1974,
US 3 232 889, 1966;
R. L. Espada, *Ion* **36** (1976) 595;
L. M. Kartashov et al., *Khim. Prom-st. (Moscow)* **1983**, 587;
Z. Trocsanyi, J. Bathory; *Kolor. Ert.* **17** (1975) 184; *Chem. Abstr.* **84** 31534 r;
L. Dubovoi et al., *Khim. Prom.-st. (Moscow)*) **1982**, 658.

499. Du Pont, US 3 697 608, 1972 (H. E. Bellis);
4 130 595, 1978 (H. E. Bellis);
Sumitomo Chemical Co., Ltd., DE-OS 2 413 148, 1977 (K. Iida, T. Takahashi, S. Kamata);
Marathon Oil Co., US 3 689 578, 1972 (D. H. Olson, G. M. Bailey).
500. Allied, BE 841 998, 1976.
501. Dow Chemical, US 4 105 702, 1978 (C. R. Mullin, D. J. Perettie).
502. Wacker, DE-OS 2 819 209, 1979 (W. Mack et al.).
503. Wacker, EP 59 251, 1984 (K. Blum, W. Mack, R. Strasser).
504. Dow Chemical, FR 1 508 863, 1967 (S. G. Levy).
505. *Eur. Chem. News* **44** (1985) no. 1161.
506. Dow Chemical, US 3 959 367, 1976 (G. C. Jeffrey).
507. G. Huybrechts et al., *Trans Faraday Soc.* **63** (1967) 1647;
H. B. Singh et al., *Environ. Lett.* **10** (1970) 253;
D. Lillian et al., *Environ. Sci. Technol.* **9** (1975) 1042;
U. Lahl et al., *Sci. Total Environ.* **20** (1981) 171;
E. P. Grimsrud, R. A. Rasmussen, *Atmos. Environ.* **9** (1975) 1014.
508. G. Kauschka, L. Kolditz, *Z. Chem.* **16** (1976) 377.
509. Wacker, DE 725 276, 1937 (G. Basel, E. Schäffer);
Donau-Chemie AG, DE 734 024, 1940 (O. Fruhwirth, H. Walla);
Société d'Ugine, FR 1 073 631, 1953.
510. FMC, ZA 715 781, 1970 (J. S. Sproul et al.).
511. Rhône-Progil, FR 2 260 551, 1974 (J.-C. Strini);
FMC, ZA 715 780, 1970 (J. S. Sproul, B. R. Marx).
512. Diamond Alkali Co., GB 701 244, 1951;
673 565, 1952;
Hydrocarbon Process. **46** (1967) no. 11, 210.
513. Donau-Chemie AG, DE-OS 1 277 245, 1965;
FR 1 525 811, 1966 (F. Samhaber);
NL 6 607 204, 1966.
514. Donau Chemie AG, GB 1 143 851, 1967 (F. Samhaber).
515. Diamond Shamrock Corp., US 3 860 666, 1976 (N. L. Beckers).
516. K. Shinoda et al., *Kogyo Kagaku Zasshi* **70** (1967) 1482;
N. A. Bhat et al., *Indian J. Technol.* **5** (1967) 255.
517. Dow Chemical, US 2 442 323, 1944 (C. W. Davis, P. H. Dirstine, W. E. Brown);
2 442 324, 1944 (R. G. Heitz, W. E. Brown);
2 577 388, 1945 (G. W. Warren).
518. Stauffer, US 2 857 438, 1957 (R. P. Obrecht, H. Bender).
519. Halcon International Inc., GB 1 047 258, 1963.
520. Société d'Ugine, FR 1 147 756, 1956;
Solvay and Cie, FR 1 373 709, 1963; J. Kraft et al.,
Chim Ind. (Paris) **83** (1960) 557.
521. Chemische Werke Hüls AG, DE 1 074 025, 1958 (F. Krüll, O. Nitzschke, W. Krumme);
F. Krüll, *Chem. Ing. Tech.* **33** (1961) 228.
522. C. H. Chilton, *Chem. Eng. (N.Y.)* **65** (1958) no. 9, 116.
523. V. N. Tychinin et al., *Khim. Prom-st (Moscow)* 1984, no. 4, 199.
524. Halcon International, Inc., FR 1 539 714, 1967 (I. E. Levine).
525. K. German, J. Rakoczy, *Przem. Chem.* **63** (1984) 93;
Chem. Abstr. **100** 17684 w.
526. Tokuyama Soda KK, JP 8 4038 932, 1984.
527. Progil-Electrochimie Cie., GB 819 987, 1957;
Hydrocarbon Process. **44** (1965) no. 11, 190.
528. Produits Chimiques Ugine Kuhlmann, US 4 211 728, 1980 (J. G. Guérin).
529. Hoechst, DE-OS 1 793 131, 1973 (H. Krekeler et al.);
2 100 079, 1973 (H. Krekeler, W. Riemenschneider);
2 150 400, 1974 (W. Riemenschneider);
2 231 049, 1974 (R. Walburg, H. Gerstenberg, H. Osterbrink).

530. FMC, US 3 364 272, 1968 (J. W. Ager).
531. *Chem. Week*, July 24, 1985.
532. *Eur. Chem. News* **44** (1985) no. 1167.
533. G. Hommel: *Handbuch der gefährlichen Güter*, Springer Verlag, Berlin 1985.
534. According to Annex II of Marpol 73/78, category regulation 3.
535. Chemische Werke Hüls AG, DE-OS 2 540 178, 1976 (R. Wickbold et al.); NL 7 610 030, 1977.
536. Solvay and Cie., EP 15 625, 1980 (R. Hembersin, R. Nicaise).
537. Stauffer, EP 94 527, 1983 (W. Burks, Jr. et al.).
538. BASF, DE-OS 2 827 761, 1980; Ezaki Shigeho, Chigasaki, Kanagawa, DE-OS 2 045 780, 1974; Knapsack AG, DE-OS 1 228 232, 1966; The Lummus Co., US 3 879 481, 1975; Nittetu Chemical Engineering Ltd., US 3 876 490, 1975.
539. The B. F. Goodrich Co., DE-OS 2 531 981, 1976; 2 532 027, 1976 (J. S. Eden); 2 532 043, 1976; 2 532 052, 1976; 2 532 075, 1976.
540. Chemische Werke Hüls AG, DE 1 246 701, 1961.
541. BASF, DE-OS 2 261 795, 1974.
542. Tessenderlo Chemie S.A., DE-OS 2 531 107, 1978.
543. Dow Chemical, US 4 435 379, 1984.
544. Hoechst, DE-OS 2 361 917, 1975 (H. Müller).
545. Grupul Industrial de Chimie Rimnicu, DE-OS 2 148 954, 1974; FR 2 156 496, 1973; GB 1 333 650, 1973; Shell International Research, GB 1 483 276, 1977; GAF, US 3 807 138, 1974; The Goodyear Tire a. Rubber Co., DE-OS 2 842 868, 1979; Air Products, DE-OS 2 704 065, 1977; BOC Ltd., GB 2 020 566, 1979; DE-OS 2 733 745, 1978.
546. The Lummus Co., DE-OS 2 604 239, 1976 (U. Tsao).
547. L. E. Swabb, H. E. Hoelscher, *Chem. Eng. Prog.* **48** (1952) 564–569.
548. N. N. Lebedev, *J. Gen. Chem. USSR (Engl. Transl.)* **24** (1954) 1925–1926.
549. R. Letterer, H. Noller, *Z. Phys. Chem. (Munich)* **67** (1969) 317–329.
550. M. F. Nagiev, Z. J. Kashkai, R. A. Makhumdzade, *Azerb. Khim. Zh.* 1969, no. 6, 27–32; *Chem. Abstr.* **74** (1971) 14 590 g.
551. Hoechst, DE 1 805 805, 1968 (H. Kuckertz, H. Grospietsch, L. Hörnig).
552. J. F. Norris, H. B. Taylor, *J. Am. Chem. Soc.* **46** (1924) 756.
553. Hoechst, DE 1 805 809, 1968 (H. Kuckertz, H. Grospietsch, L. Hörnig).
554. Shell Dev. Co., US 2 207 193, 1937 (H. P. A. Groll); *Chem. Abstr.* **34** (1940) 7934.
555. E. T. McBee, H. B. Hass, T. H. Chao, Z. D. Welch et al., *Ind. Eng. Chem.* **33** (1941) 176–181.
556. G. W. Hearne, T. W. Evans, H. L. Yale, M. C. Hoff, *J. Am. Chem. Soc.* **75** (1953) 1392–1394.
557. H. Gerding, H. G. Haring, *Recl. Trav. Chim. Pays-Bas* **74** (1955) 841–853.
558. H. P. A. Groll, G. Hearne, F. F. Rust, W. E. Vaughan, *Ind. Eng. Chem.* **31** (1939) 1239–1244.
559. R. Kühn, K. Birett: *Merkblätter Gefährliche Arbeitsstoffe*, paper no. D 26, Ecomed Verlagsgesellschaft, Landsberg/Lech.
560. G. Hommel: *Handbuch der gefährlichen Güter*, leaflet 170, Springer Verlag, Berlin-Heidelberg 1980.
561. N. I. Sax: *Dangerous Properties of Industrial Materials*, 6th ed., Van Nostrand Reinhold Co., New York 1984.
562. *Hazardous Chemicals Data Book*, Noyes Data Corp., Park Ridge, USA, 1980.
563. M. Sittig: *Handbook of Toxic and Harzardous Chemicals and Carcinogens*, 2nd ed., Noyes Publ., New Jersey (1985), 329.
564. *Ullmann*, 4th ed., **10**, 684.
565. R. Kühn, K. Birett: *Merkblätter Gefährliche Arbeitsstoffe*, paper no. T 24, Ecomed Verlagsgesellschaft, Landsberg/Lech.
566. A. L. Henne, F. W. Haeckl, *J. Am. Chem. Soc.* **63** (1941) 2692.
567. A. D. Herzfelder, *Ber. Dtsch. Chem. Ges.* **26** (1893) 2432–2438.
568. Du Pont, US 2 119 484, 1931 (A. A. Levine, O. W. Cass); *Chem. Abstr.* **32** (1938) 5413[4].
569. T. Kleinert, *Chem. Ztg.* **65** (1941) 217–219.
570. *Kirk-Othmer*, **18** 815.
571. H. B. Hass, E. T. McBee, P. Weber, *Ind. Eng. Chem.* **27** (1935) 1191.

572. Purdue Research Foundation, US 2 105 733, 1935.
573. A. Maillard et al., *Bull. Soc. Chim. Fr.* 1961, no. 5, 1640.
574. A. Pichler, *Peint. Pigm. Vernis* **41** (1965) no. 5, 293.
575. A. Pichler, G. Levy, *Bull. Soc. Chim. Fr.* **1964**, no. 9, 2815; **1966**, no. 11, 3656.
576. A. I. Gershenovich, V. M. Kostyuchenko, *Zh. Prikl. Khim.* **39** (1966) no. 5, 1160.
577. Usines de Melle, FR 1 352 211, 1962 (J. Mercier).
578. Sharpless Solvents Corp., US 2 122 110, 1937.
579. E. G. Bondarenko et al., SU 172 289, 1965; *Chem. Abstr.* **63** (1965) 16212 f.
580. Du Pont, US 2 570 495, 1946 (N. D. Scott).
581. J. Gerrad, M. Phillips, *Chem. Ind. (London)* 1952, 540.
582. G. Toptschijew, *J. Prakt. Chem.* **2** (1955) no. 4, 185.
583. F. Asinger: *Chemie und Technologie der Monoolefine*, Akademie Verlag, Berlin 1956, p. 248.
584. J. J. Leendertse et al., *Recl. Trav. Chim. Pays Bas* **52** (1933) 514–524; **53** (1934) 715.
585. C. C. Coffin, O. Maass, *Can. J. Res.* **3** (1930) 526.
586. C. C. Coffin, H. S. Sutherland, O. Maass, *Can. J. Res.* **2** (1930) 267.
587. BASF, DE 1 004 155, 1954 (J. Schmidt, W. Ritter).
588. BASF, DE 859 734, 1939.
589. O. W. Cass, *Ind. Eng. Chem.* **40** (1948) 216.
590. W. Reppe et al., *Justus Liebigs Ann. Chem.* **596** (1955) no. 1, 118.
591. Du Pont, US 2 889 380, 1956 (E. E. Hamel).
592. H. Jura, *J. Chem. Soc. Japan Ind. Chem. Sect.* **54** (1951) 433, C.A. 1238, 1954.
593. N. Shono, S. Hachihama, *Chem. High Polym. (Japan)* **8** (1951) 70; *Chem. Abstr.* **56** (1953) 7483.
594. G. Lutkowa, A. Kutsenko, *Zh. Prikl. Khim. (Leningrad)* **32** (1959) 2823.
595. Chem. Werke Hüls AG, DE 1 188 570, 1961.
596. E. M. Asatryan, G. S. Girgoryan, A. Malkhasyan, G. Martirosyan, *Arm. Khim. Zh.* **36** (1983) no. 8, 527–530.
597. R. F. Tylor, G. H. Morsey, *Ind. Eng. Chem.* **40** (1948) 432.
598. Distillers Co., DE 1 090 652, 1957 (C. W. Capp, H. P. Croker, F. J. Bellinger).
599. Du Pont, FR 1 504 112, 1967 (J. L. Hatten, K. W. Otto).
600. Knapsack, DE-OS 241 312, 1973 (K. Gehmann, A. Ghorodnik, U. Dettmeier, H. J. Berns).
601. ICI, GB 1 119 862, 1966 (R. L. Heath).
602. H. J. Zimmer, Verfahrenstechnik, DE 1 118 189, 1960 (N. W. Luft, K. Esser, H. Waider).
603. Sumitomo Chem. Co., DE-AS 1 961 721, 1969.
604. Du Pont, DE-OS 2 046 007, 1970 (O. K. Wayne, J. L. Hatten).
605. H. Stemmler et al., DD 14 502, 1958.
606. Showa K.K., JP 7 108 281, 1967.
607. ICI, (W. Costain, B. W. H. Terry).
608. I. G. Farben, US 2 242 084, 1938 (O. Nicodemus, W. Schmidt).
609. Showa Denko K.K., JP 7 601 682, 1976 (Ogawa, Masaro; Yoshinaga, Yasuo).
610. Agency of Ind. Science and Technology, JP 7 227 906, 1972; JP 7 334 570, 1973.
611. Monsanto Chem. Co., FR 1 326 120, 1962.
612. Shell Intern. Res., GB 1 007 077, 1964.
613. BASF, DE 1 115 236, 1961 (M. Minsinger).
614. K. Weissermel, H. J. Arpe: *Ind. Org. Chemistry*, Verlag Chemie, Weinheim 1978, p. 98.
615. Du Pont, US 294 987, 1972 (J. H. Richards, C. A. Stewart).
616. BP Chemicals Ltd., GB 1 296 482, GB 1 296 483, 1972 (P. J. N. Brown, C. W. Capp).
617. Du Pont, USA 1 918 067, 1968 (D. D. Wild).
618. Du Pont, DE-OS 2 248 668, 1972 (B. T. Nakata, E. D. Wilhoit).
619. Du Pont, GB 1 260 691, 1969 (D. D. Wild).
620. BP Chemicals Ltd., DE 2 107 467, DE 2 107 468, DE-OS 2 107 469, 1971 (C. W. E. Capp, P. J. N. Brown).
621. Distillers Co. Ltd., GB 877 586, 1958.
622. Electrochemical Ind., JP 6 809 729, 1967.
623. Toyo Soda Ind., JP 6 920 330, 1968.
624. Bayer, DE 3 208 796, 1982 (J. Heinrich, R. Casper, M. Beck).
625. W. H. Carothers et al., *J. Am. Chem. Soc.* **53** (1931) 4203; **54** (1932) 4066.
626. L. F. Hatch, S. G. Ballin, *J. Am. Chem. Soc.* **71** (1949) 1039.
627. Du Pont, US 1 950 431/32, 1934; 2 102 611, 1937; 2 178 737, 1937.
628. Distillers Co. Ltd., DE 1 193 936, 1963.
629. Bayer, DE 2 318 115, 1973 (F. Hagedorn, R. Mayer-Mader, K. F. Wedemeyer).

630. G. J. Beschet, W. H. Carothers, *J. Am. Chem. Soc.* **55** (1933) 2004.
631. Du Pont, US 1 965 369, 1934.
632. Knapsack, DE 1 149 001, 1961.
633. Bayer, DE 2 545 341, 1975 (H. J. Pettelkau).
634. Bayer, DE 2 717 672, 1977 (G. Scharfe, R. Wenzel, G. Rauleder).
635. Du Pont, US 1 950 431, 1934 (W. H. Carothers, A. M. Collins).
636. W. H. Carothers et al., *J. Am. Chem. Soc.* **53** (1931) 4203.
637. A. E. Nyström, *Acta Med. Scand. Suppl.* **132** (1948) 219.
638. *Chem. Abstr.* **44** (1950) 8758.
S. E. Ekegren et al., *Acta Chem. Scand.* **4** (1950) 126–139.
639. Knapsack, DE 1 148 230, 1959 (W. Vogt, H. Weiden, K. Gehrmann).
640. Distillers Co., Ltd., US 2 926 205, 1960 (F. J. Bellringer).
641. G. E. Ham: *High Polymers*, vol. **XVIII**, Interscience, New York 1964, p. 720.
642. Du Pont, US 2 395 649, 1946 (F. C. Wagner).
643. Bayer, GB 858 444, 1961 (A. R. Heinz, W. Graulich).
644. Distillers Co. Ltd., DE 1 222 912, 1963 (D. A. Tadworth, A. B. Sutton, A. Foord, E. S. Luxon).
645. Du Pont, DE 1 618 790, 1967 (J. B. Campbell, R. E. Tarney).
646. Du Pont, DE 1 909 952, 1969 (J. B. Campbell, J. W. Crary, C. A. Piaseczynski, J. Stanley).
647. Denka Chem. Corp., US 4 132 741, 1965 (A. J. Besozzi).
648. Du Pont, US 4 418 232, 1982 (J. L. Maurin).
649. Du Pont, US 4 215 078, 1979 (C. A. Hargreaves, A. T. Harris, R. A. Schulze).
650. Knapsack AG, DE 2 139 729, 1971 (A. Ghrodnik, U. Dettmeier, K. Gehrmann, H. J. Berns).
651. Bayer, DE 2 310 744, 1973 (R. Wenzel, G. Scharfe).
652. Bayer, DE 2 460 912, 1974;
2 533 429, 1975 (G. Scharfe, R. Wenzel).
653. BP Chem. Ltd., DE 2 707 073, 1976 (A. H. P. Hall, J. P. Merle).
654. Bayer, DE 529 387, 1975 (G. Beilstein, B. Ehrig, D. Grenner, K. Nöthen).
655. Du Pont, DE 588 283, DE 589 561, 1931.
656. W. H. Carothers, *J. Am. Chem. Soc.* **54** (1932) 4066.
657. Du Pont, US 2 207 784, 1937;
2 221 941, 1937.
658. Du Pont, US 2 949 988, 1972 (J. H. Richards C. A. J. Stewart).
659. Bayer, DE 2 234 571, 1972 (M. Weist, M. Leopold).
660. Bayer, DE 2 357 194, 1973 (M. Weist, M. Leopold).
661. Electrol-Chemical Ind. K.K., JP 011 135, 1977 (Ito, Akira; Watanabe, Seiichi).
662. M. Stojanowa-Antoszczyn, A. Z. Zielinski, E. Chojnicki, *Przem. Chem.* **58** (1979) no. 3, 160–162.
663. Y. Takasu, Y. Matsuda et al., *Chem. Lett.* **12** (1981) 1685–1686;
J. Electrochem. Soc. **131** (1984) no. 2, 349–351.
664. Toyo Soda, JP 81 139 686, 1981;
JP 82 161 076, 1982.
665. BASF, DE 1 115 236, 1961 (M. Minsinger).
666. Du Pont, US 3 406 215, 1968 (H. E. Holmquist).
667. K. Möbius, *Gummi Asbest Kunststoffe* **35** (1982) 6.
668. Toyo Soda Mfg Co Ltd., JP 7 525 048, 1975 (H. Kisaki, K. Tsuzuki, Shimizu).
669. Du Pont, GB 880 077, 1959.
670. Distillers Co. Ltd., GB 825 609, 1957 (H. P. Crocker).
671. *Chem. Eng. News* **41** (1963) 38.
672. Denki Kagaku Kogyo, JP 80 166 970, 1980 (A. Okuda, Y. Totake, H. Matsumura).
673. Denki Kagaku Kogyo, JP 7 700 926, 1977 (T. Kadowaki, T. Iwasaki, H. Matsumura).
674. Toyo Soda Mfg Co Ltd., JP 7 426 207, 1974 (S. Ootsuki, H. Ookado, Y. Tamai, T. Fujii).
675. G. M. Mkryan, R. K. Airapetyan et al., *Arm. Khim. Zh.* **34** (1981) 242–246.
676. BP Chemicals Ind. Ltd., FR 2 148 586, 1973 (C. W. Capp, P. J. N. Brown).
677. Bayer, DE 2 642 006, 1976,
DE 2 655 007, 1976 (R. Lantzsch, E. Kysela).
678. F. Asinger: *Petrolchemische Ind.*, vol. **II**, Akademie Verlag, Berlin (GDR) 1971, p. 1330.
679. E. Profft, H. Oberender, *J. Prakt. Chem.* **25** (1964) no. 4–6, 225–271.
680. M. Tamele et al., *Ind. Eng. Chem.* **33** (1941) 115.
681. J. Burgin et al., *Ind. Eng. Chem.* **33** (1941) 385.
682. G. Hearne et al., *Ind. Eng. Chem.* **33** (1941) 805 and 940.
683. L. H. Gale, *J. Org. Chem.* **31** (1966) 2475–2480.
684. *Houben-Weyl*, vol. **V/3**, 586.

685. T. D. Stewart, D. M. Smitt, *J. Amer. Chem. Soc.* **52** (1930) 2869.
686. W. Reeve, D. H. Chambers, *J. Amer. Chem. Soc.* **73** (1951) 4499.
687. W. Reeve, C. S. Prikett, *J. Amer. Chem. Soc.* **74** (1952) 5369.
688. I. V. Bodrikov et al., *Usp. Khim.* **3** (1966) no. 5, 853–880;
Chem. Abstr. **69** (1966) 657 003.
689. A. Striegler, *Chem. Tech.* **9** (1957) 523–529.
690. Chiyoda Kako Kensetsu Kabushiki Kaisha, DE-AS 1 281 430, 1966.
691. Z. S. Smoljan, I. V. Bodrikov, *Chim. Prom.* **43** (1967) no. 3, 192.
692. Hoechst, DE-AS 1 237 557, 1965;
DE-AS 1 237 555, 1965;
DE-AS 1 243 670, 1965.
Showa Denko KK, DE 1 817 281, 1969.
693. Mitsubishi Petrochemical, JP 117 576, 1977.
694. F. C. Hymas, *Food* **9** (1940) 254.
695. C. H. Richardson, H. H. Walkden, *J. Econ. Entomol.* **38** (1945) 471.
696. Pittsburgh Plate Glass Ind., US 2 356 871, 1940 (E. N. Moffett, R. E. Smith).
697. US 2 440 494, 1941.
United States Rubber Co., US 2 553 430, 1948;
2 553 431, 1948;
2 568 872, 1946;
2 569 959, 1948;
2 576 245, 1949;
2 592 211, 1948;
2 594 825, 1947;
2 597 202, 1948;
2 643 991, 1949.
698. Industrial Rayon Corp., US 2 601 256, 1950.
Du Pont, GB 823 345, 1957.
V. Gröbe, H. Reichert, W. Maschkin, *Faserforsch. Textiltech.* **15** (1964) no. 10, 463.
699. Kureha Chemical Ind. Co., JP 6 804 321, 1965.
700. Kureha Chemical Ind. Co., JP 6 804 326, 1965;
Chem. Abstr. **70** (1969) 11 103 g.
701. K. Shinoda, H. Watanabe, *Kogyo Kagaku Zasshi* **70** (1967) 2262;
Chem. Abstr. **68** (1968) 104 303 w.
702. L. M. Kogan, *Usp. Khim* **33** (1964) no. 4, 396–417.
K. Hisashi, K. Shigeaki, *Sekiyu To Sekiyu Kagaku* **16** (1972) no. 1, 73;
Chem. Abstr. **76** (1972) 129 588 w.
703. E. T. McBee, R. E. Hatton, *Ind. Eng. Chem.* **41** (1949) 809.
704. Y. G. Mamedaliev, M. M. Gusenov et al., SU 161 713, 1963.
705. Hooker Electrochemical Co., US 2 841 243, 1957.
706. Siemens AG, Wacker-Chemie, FR 1 564 230, 1968.
707. Matsushita Electric Industrial Co., Ltd., JP 7 010 150, 1970;
Chem. Abstr. **73** (1970) 360 899.
708. L. M. Kogan et al., SU 151 156, 1962.
709. P. A. Bolley, *Liebigs Ann. Chem.* **106** (1858) 230.
710. Daisoh Kabushiki, JP-Kokai 6 239 774, 1994 (Y. Itaya, H. Takakata).
711. Hoechst Aktiengesellschaft, EP-A 0 625 196, 1994 (D. Bewart, W. Freyer).
712. A. S. Bratolyubov, *Russ. Chem. Rev. Engl. Transl.* **30** (1961) 602.
713. A. M. Kuliev et al., *Pris. Smaz. Masl.* **116** (1967) .
714. T. T. Vasil'eva et al., *Isv. Akad. Nauk. SSSR, Ser. Khim.* (1973) 1068.
715. V. B. Shumyanskaya et al., *Sb. Tr. Vses. Nauchno Issled. Inst. Nov. Stroit. Mater.* **32** (1972) 46.
716. World Health Organisation, International Programme on Chemical Safety: Environmental Health Criteria 181, Chlorinated Paraffins, Geneva 1997
717. M. D. Müller, P. P. Schmid, *J. High Resolut. Chromatogr. and Chromatogr. Commun.* **7** (1984) 33.
718. R. D. N. Birtley et al., *Toxicol. Appl. Pharmacol.* **54** (1980) 514.
719. C. Elcombe, B. Elcombe, R. Powrie, S. Barton, D. Farrar: "Medium chain chlorinated paraffin (MCCP)-induced haemorrhagic lesions in neonatal Sprague-Dawley rats. A role for altered vitamin K disposition?", *The Toxicologist* (2005) Abstract No. 277.
720. International Agency for Research on Cancer: Chlorinated Paraffins, vol. 48, Lyon 1990, p. 55.
721. S. C. Hasmall, J. R. Foster, C. R. Elcombe, *Arch. Toxicol.* (1997).
722. I. Wyatt et al., *Arch. Toxicol.* (1997).
723. R. Thompson, M. Comber: "Current issues in the ecotoxicology of chlorinated paraffins", *Organohalogen Compounds* **47** (2000) 139–140.
724. R. T. Cook, *Speciality Chemicals* (1995) 159.
725. L. H. Hoorsley et al., *Azeotropic Data III*, Advances in Chemistry Series, No. 116, Am. Chem. Soc., Washington, D.C., 1973.

726. M. Lecat, *Tables azeotropiques*, 2nd ed., Chez l'Auteur, Brussels 1949.
727. Dow Chemical, US 1 607 618, 1926 (W. J. Hale, E. C. Britton).
728. J. W. Hale, E. C. Britton, *Ind. Eng. Chem.* **20** (1928) 114–124.
729. R. M. Crawford, *Chem. Eng. News* **25** (1947) no. 1, 235–236.
730. Dow Chemical, US 2 432 551, 1947; 2 432 552 (W. H. Williams et al.).
731. Oiwa et al., *Bl. Inst. Chem. Res. Kyoto* **29** (1952) 38–56.
732. M. Bonnet et al., *J. Org. Chem.* **48** (1983) 4396–4397.
733. Hoechst, EP 51 236, 1980 (D. Rebhan, H. Schmitz).
734. *Chem. Eng.*, March 31. 1975, 62.
735. *Chem.* Ind. **37**, p. 106, February 1985.
736. Chemical Marketing Reporter, vol. 226, no. 13, September 24, 1984.
737. *Japan Chem. Week*, April 15, 1982, 8.
738. U.S. Interdepartmental Task Force on PCBs, COM–72–10419, Washington 1972.
739. F. D. Kover, *Environmental Protection Agency*, Washington, D.C., 1975, EPA 560sol;–75–001, p. 8, p. 11.
740. U.S. Federal Register, vol. 48, no. 89, May 6, 1983, p. 20 668.
741. Figure from an unpublished survey conducted by Bayer (produced in the FRG for PCN only).
742. Dow Chemical, US 2 886 605, 1959 (H. H. McClure, J. S. Melbert, L. D. Hoblit).
743. Hoechst, BE 660 985, 1964.
744. PPG, US 4 017 551, 1977 (J. E. Milam, G. A. Carlson).
745. K. Shinoda, H. Watanabe, *Kogyo Kagaku Zasshi* **73** (1970) no. 11, 2549–2551; *Chem. Abstr.* **77** (1972) 5072.
746. S. Yamamoto et al., *Kagaku to Kogyo (Osaka)* **45** (1971) no. 9, 456–460; *Chem. Abstr.* **76** (1972) 99244.
747. Moscow Gubkin Petrochem., SU 514 860, 1976; SU 536 185, 1976; SU 540 657, 1976; SU 594 995, 1978 (J. M. Kolesnikow et al.).
748. Dow Chemical, US 3 636 171, 1972 (K. L. Krumel, J. R. Dewald).
749. T. Kametani et al., *Yakugaku Zasshi* **88** (1968) no. 8, 950–953; *Chem. Abstr.* **70** (1978) 19681.
750. PPG, US 4 235 825, 1980 (J. E. Milam).
751. Sergeev E. V., SU 650 985, 1979.
752. Nippon Kayaku Co., JP-Kokai 64/3821, 1961.
753. Union Carbide, US 3 226 447, 1965 (G. H. Bing, R. A. Krieger).
754. Ihara Chem. Ind. Co., Ltd., EP 118 851, 1984 (Y. Higuchi, T. Suzuki).
755. R. Commandeur et al., *Nouv. J. Chim.* **3** (1979) no. 6, 385–391; PCUK, DE-OS 3 344 870, 1983 (R. Commandeur, H. Mathais).
756. Dow Chemical, US 4 186 153, 1980; US 4 205 014, 1980 (J. W. Potts).
757. Sanford Chem. Co., US 3 557 227, 1971 (M. M. Fooladi).
758. Von den Berg et al., NL 72/1533, 1974; *Ind. Eng. Chem. Fundam.* **15** (1976) no. 3, 164–171.
759. R. B. McMullin, *Chem. Eng. Prog.* **44** (1948) no. 3, 183–188.
760. H. J. Hörig, D. Förster, *Chem. Tech. (Leipzig)* **22** (1970) no. 8, 460–461.
761. S. Sacharjan, *Chem. Tech. (Leipzig)* **31** (1979) no. 5, 229–231.
762. N. T. Tishchenko, V. M. Sheremet, A. I. Triklio, *Khim. Prom-st. (Moscow)* 1980, no. 7, 399–401.
763. *Ullmanns*, 4th ed., **9**, p. 506.
764. Raschig GmbH, DE 539 176, 1930; DE 575 765, 1931; US 1 963 761, 1934; US 1 964 768, 1933 (W. Prahl, W. Mathes).
765. Hooker Chem. Corp., DE 1 443 988, 1962; CH 432 485, 1962; US 3 644 542, 1972 (W. Prahl et al.).
766. Union Carbide, BE 636 194, 1962 (C. C. Diffteld, T. T. Szabo, C. A. Rivasi).
767. Vulcan Materials Co., DE 1 518 164, 1963 (E. S. Penner, A. L. Malone).
768. Hodogaya Chem. Ind. Co., JP-Kokai 65/12 613, 1962.
769. Asahi Chem. Ind. Co., JP-Kokai 67/3619, 1965.
770. W. H. Prahl, DE-OS 2 442 402, 1975.
771. Sumitomo, JP-Kokai 73/39 449, 1971; DE 2 232 301, 1972 (E. Ichiki et al.); *Chem. Abstr.* **79** (1973) 91 746.
772. C. A. Gorkowa, A. M. Potapow, C. R. Rafikow, SU 654 600, 1979.
773. U. Schönemann, *Chem. Ind, (Düsseldorf)* **5** (1953) no. 7, 529–538.
774. Gulf Res. a. Develop. Co., US 3 591 644; US 3 591 645, 1971 (V. A. Notaro, C. M. Selwitz).
775. ICI, GB 388 818, 1933 (T. S. Wheeler).
776. Dow Chemical, US 2 123 857, 1938 (J. P. Wibaut, L. van de Lande, G. Wallagh).

777. C. E. Kooyman, *Adv. Free Radical Chem.* **1** (1965) 137–153.
778. W. Dorrepaal, R. Louw, *Recl. Trav. Chim. Pays-Bas* **90** (1971) 700–704; *J. Chem. Soc. Perkin Trans* **2** (1976) 1815–1818; *Int. J. Chem. Kinet.* **10** (1978) no. 3, 249–275.
779. P. Kovacic, N. O. Brace, *J. Am. Chem. Soc.* **76** (1954) 5491–5494.
780. P. Kovacic, A. K. Sparks, *J. Am. Chem. Soc.* **82** (1960) 5740–5743; *J. Org. Chem.* **26** (1961) 1310–1213.
781. Pullman Inc., DE 1 468 141, 1963 (H. Heinemann, K. D. Miller).
782. National Distillers Chem. Corp., US 3 670 034, 1972 (R. E. Robinson).
783. J. S. Grossert, G. K. Chip, *Tetrahedron Lett.* **30** (1970) 2611–2612.
784. Nikkei Kako, Co., Ltd., Sugai Kogy Ind., Ltd., DE-OS 2 230 369, 1972 (K. Sawazaki, H. Fujii, M. Dehura).
785. Y. Tamai, Y. Ohtsuka, *Bull. Chem. Soc. Japan* **46** (1973) no. 7, 1996–2000.
786. Monsanto, US 3 029 296, 1962 (W. A. White, R. A. Rührwein).
787. O. Silberrad, *J. Chem. Soc. (London)* **121** (1922) 1015–1022; US 259 329, 1925.
788. M. Ballester, C. Molinet, J. Castaner, *J. Am. Chem. Soc.* **82** (1960) 4254–4258.
789. M. Ballester, *Bull. Soc. Chim. Fr.* 1966, 7–15.
790. S. Kobayashi, JP-Kokai 74/43 901, 1972; *Chem. Abstr.* **81** (1974) 35683.
791. Du Pont, US 4 327 036, 1982 (F. D. Marsh).
792. R. J. Cremlyn, T. Cronje, *Phosphorus Sulfur* **6** (1979) no. 3, 495–504.
793. N. C. Weinberg, "Tech. Elektroorg. Synth.," *Tech. Chem.* (N.Y.) 5 (1975) Pt. 2, 1–82.
794. Texaco Inc., US 3 692 646, 1972 (W. B. Mather, H. Junction, E. R. Kerr).
795. Sumgait Petrochem., SU 612 923, 1978.
796. C. Galli, *Tetrahedron Lett.* **21** (1960) no. 47, 4515–4516.
797. S. Oae, K. Shinhama, Y. H. Kim, *Bull. Chem. Soc. Japan* **53** (1980) no. 7, 2023–2026.
798. *BIOS* 986, 151.
799. A. A. Ponomarenko, *Zh. Obshch. Khim.* **32** (1962) no. 12, 4029–4040.
800. N. N. Vorozhtsov et al., *Zh. Obshch. Khim.* **31** (1961) no. 4, 1222–1226.
801. Monsanto, GB 948 281, 1962; DE 1 468 241, 1963 (E. B. McCall, W. Cummings).
802. J. Blum, *Tetrahedron Lett.* **26** (1966) 3041–3045.
803. Du Pont, DE 1 804 695, 1968 (C. C. Cumbo).
804. Bayer, DE 2 326 414, 1973; US 3 897 321, 1975; US 4 012 442, 1977 (H. U. Blank, K. Wedemeyer, J. Ebersberger).
805. Mitsui Toatsu Chem. Inc., JP-Kokai 80/104 220, 1979; *Chem. Abstr.* **94** (1974) 30 327.
806. Monsanto, GB 957 957, 1962 (E. B. McCall, P. J. S. Bain).
807. Dow Chemical, US 1 946 040, 1934 (W. C. Stösser, F. B. Smith).
808. S. Rittner, S. Steiner, *Chem. Ing. Tech.* **57** (1985) no. 2, 91–102.
809. Politechnika Slaska, PL 78 610, 1976 (W. Zielinski, J. Suwinski).
810. Mitsui Toatsu Chem., Inc., JP-Kokai 78/44 528, 1976.
811. Dow Chemical, US 3 170 961, 1965 (E. C. Britton, F. C. Beman).
812. Dow Chemical, US 1 923 419, 1933 (E. C. Britton).
813. Rhône-Progil, DE 2 332 889, 1973 (R. Chanel et al.).
814. Mitsui Toatsu Chem., Inc., JP-Kokai 77/6229, 1975 (T. Kiyoura); *Chem. Abstr.* **87** (1972) 84 178.
815. S. Kikkawa et al., *Kogyo Kagaku Zasshi* **73** (1970) no. 5, 964–969; *Chem. Abstr.* **73** (1970) 87 542.
816. E. Pfeil, *Angew. Chem.* **65** (1953) no. 6, 155–158.
817. W. Brackmann, P. J. Smit, *Recl. Trav. Chim. Pays-Bas* **85** (1966) no. 8, 857–864; Shell Int. Res., NL 6 515 451, 1967.
818. Mobil Oil Corp., US 3 358 046, 1967 (R. D. Offenhauer, P. G. Rodewald).
819. Sumitomo, JP-Kokai 73/86 828, 1972 (H. Suda, S. Yamamoto, Y. Sato); *Chem. Abstr.* **80** (1974) 59 655.
820. Allied, US 2 866 828, 1958 (J. A. Crowder, E. E. Gilbert).
821. Ethyl Corp., US 2 943 114, 1960 (H. E. Redman, P. E. Weimer).
822. Columbia-Southern Chem. Corp., US 2 819 321, 1958 (B. O. Pray).
823. BASF, DE 1 020 323, 1956 (E. Merkel).
824. Union Carbide, US 2 666 085, 1954 (J. T. Fitzpatrick).
825. Y. G. Erykalov et al., *J. Org. Chem. USSR (Engl. Transl.)* **9** (1973) 348–351; SU 392 059, 1973; SU 351 818, 1972.

826. Kanegabuchi Chem. Ind. Co., JP-Kokai 57/510, 1957 (J. Nishida, J. Mikami, S. Kawato);
Chem. Abstr. **52** (1958) 2905 E.
827. H. Scholz et al., DD 9869, 1955.
828. D. Vrzgula, Z. Stoka, J. Auerhan, CS 92 749, 1959.
829. Mitsubishi, JP-Kokai 75/32 132, 1973;
Chem. Abstr. **83** (1975) 96 688.
830. BASF, DE 1 002 302, 1952 (G. Grassl).
831. Asahi Glass Co., JP-Kokai 55/7421, 1955 (M. Fujioka et al.);
Chem. Abstr. **51** (1957) 17 989 G.
832. Ethyl Corp., GB 738 011, 1952 (L. Merritt).
833. Diamond Alkali Co., US 2 914 574, 1959 (E. Zinn, J. J. Lukes).
834. Pennsalt Chem. Corp., US 2 914 573, 1959 (G. McCoy, C. Inman).
835. Österreichische Chem. Werke, AT 261 576, 1967 (H. Maier).
836. Uzima Chimica Turda, BG 54 958, 1972 (J. Todea).
837. Combinatul Chim. Riminicu-Vilcea, BG 61 469, 1976 (L. Raduly et al.).
838. *Ullmann*, 4th ed., **9**, 508.
839. Ishihara Sangyo Kaisha, Ltd., JP-Kokai 81/92 227, 1979;
Chem. Abstr. **95** (1981) 203 526.
840. Hoechst, DE-OS 2 920 173, 1979 (G. Volkwein, K. Bässler, H. Wolfram).
841. Ishihara Sangyo Kaisha, Ltd., JP-Kokai 79/112 827, 1978;
US 4 368 340, 1983;
DE-OS 3 047 376, 1980 (T. R. Nishiyama et al.).
842. BASF, DE 947 304, 1954 (F. Becke, H. Sperber).
843. Bayer, DE-OS 3 134 726, 1981 (G.-M. Petruck, R. Wambach).
844. Rhône Poulenc Ind., EP 43 303, 1982 (G. Soula).
845. Sonford Chem. Co., US 3 557 227, 1971 (M. M. Fooladi).
846. Dover Chem. Corp., US 3 274 269, 1966 (R. S. Cohen).
847. Jefferson Chem. Co., Inc., DE-OS 2 054 072, 1971 (S. Burns).
848. T. Magdolen, D. Vrzgula, CS 122 646, 1967.
849. Hoechst, DE 1 028 978, 1955 (H. Klug, J. Kaupp).
850. Nippon Soda Co., JP-Kokai 75/7572, 1975 (K. Iwabuchi);
Chem. Abstr. **51** (1957) 17 989 F.
851. Pechiney, US 2 852 571, 1958 (J. Frejacques, H. Guinot).
852. C. Popa, C. Raducanu, A. Stefanescu, *Rev. Chim. (Bucharest)* **15** (1964) no. 10, 601–607.
853. Phillips Petr. Co., US 4 096 132, 1978 (J. T. Edmonds Jr.
854. Bayer, EP 23 314, 1981;
EP 40 747, 1981;
EP 65 689, 1982 (K. Idel et al.).
855. Swann Research Inc., US 1 954 469, 1934 (C. F. Booth).
856. Ullmann, 4th ed., **8**, 344, 377–379.
857. Ihara Chem. Ind. Co., Ltd., EP 2 749, 1978 (M. Shigeyasu et al.);
Dynamit Nobel AG, EP 70 393, 1982;
EP 121 684, 1984 (M. Feld).
858. Nippon, DE-OS 2 711 332, 1977 (H. Hayami, H. Shimizu, G. Takasaki).
859. N. N. Lebedev, J. J. Baltadzhi, *Zh. Org. Khim.* **5** (1969) no. 9, 1604–1605.
860. Heyden Newport Chem. Co., US 3 000 975, 1961 (E. P. Di Bella).
861. Moscow Gubkin Institute, SU 1 002 297, 1983 (N. N. Belov et al.).
862. *Ullmann*, 4th ed., **9**, 512–513.
863. Hooker Chem. & Plas. Corp., US 4 013 730, 1977 (J. C. Graham).
864. Tokuyama Soda Co. Ltd., JP-Kokai 81/99 429, 1980;
Chem. Abstr. **95** (1981) 186 820.
865. Tenneco Chem., US 4 013 144, 1977 (E. P. Di Bella).
866. Hooker Chem. & Plas. Corp., US 4 031 142, 1977;
DE-OS 2 634 340, 1976 (J. C. Graham).
867. Tenneco Chem., US 3 317 617, 1967 (E. P. Di Bella).
868. Hodogaya Chem. Co., Ltd., JP-Kokai 81/110 630, 1980;
Chem. Abstr. **96** (1982) 34 810.
EP 63 384, 1982 (R. Hattori et al.).
869. Ihara Chem. Ind. Co., Ltd., JP-Kokai 81/105 752, 1980;
US 4 289 916, 1981;
DE-OS 3 023 437, 1980 (J. Nahayama, C. Yazawa, K. Jamanski).
870. Hooker Chem. & Plas. Corp., US 4 024 198, 1977 (H. E. Buckholtz);
US 4 069 263, 1978;
US 4 069 264, 1978 (H. C. Lin).
871. Hooker Chem. & Plas. Corp., US 4 250 122, 1981 (H. C. Lin, S. Robots).
872. P. B. De La Mare, P. W. Robertson, *J. Chem. Soc.* 1943, 279–281.
873. H. G. Haring, *Chem. Eng. Monogr.* **15** (1982) 193–208, 208 A.

874. Hoechst, Krebskosmos GmbH, DE-OS 2 536 261, 1975 (H. Horst et al.).
875. J. M. Ballke, J. Liaskar, G. B. Lorentzen, *J. Prakt. Chem.* **324** (1982) no. 3, 488–490.
876. A. B. Solomonov, P. P. Gertsen, A. N. Ketov, *Zh. Prikl. Khim. (Leningrad)* **42** (1969) no. 6, 1418–1420; *Zh. Prikl. Khim. (Leningrad)* **43** (1970) no. 2, 471–472.
877. A. M. Potapov, S. R. Rafikov, *Izv. Akad. Nauk SSSR, Ser. Khim.* **31** (1982) no. 6, 1334–1338.
878. Ajinomoto Co., Inc., JP Kokai 73/81 822, 1972 (R. Fuse, T. Inoue, T. Kato); *Chem. Abstr.* **80** (1974) 108 160.
879. Gulf Research and Development Co., US 3 644 543, 1972 (V. A. Notaro); US 3 717 684, 1973 (V. A. Notaro, C. M. Selwitz); US 3 717 685, 1973 (H. G. Russell, V. A. Notaro, C. M. Selwitz).
880. Int. Minerals & Chem. Corp., JP Kokai 74/76 828, 1972 (J. T. Traxler); *Chem. Abstr.* **85** (1976) 6167.
881. H. Inoue, M. Izumi, E. Imoto, *Bull. Chem. Soc. Japan* **47** (1974) no. 7, 1712–1716.
882. T. Tkaczynski, Z. Winiarski, W. Markowski, *Przem. Chem.* **58** (1979) no. 12, 669–670; *Chem. Abstr.* **92** (1980) 180 745.
883. Sugai Chem. Ind. Co., Ltd., JP Kokai 83/61 285, 1981; *Chem. Abstr.* **99** (1983) 79 068.
884. Shell Int. Res., FR 1 343 178, 1963.
885. R. Nishiyama et al., *J. Soc. Org. Synth. Chem. Tak.* **23** (1965) 515–525.
886. Hoechst, EP 72 008, 1982 (H.-J. Arpe, H. Litterer, N. Mayer).
887. BIOS Final Rep. No. 1145.
888. Toray, EP 62 261, 1982 (K. Tada et al.).
889. Tenneco Chem. Inc., FR 1 489 686, 1967; US 3 366 698, 1968 (E. P. Di Bella).
890. Hooker Chem. & Plas. Corp., DE-OS 2 258 747, 1972; US 4 006 195, 1977 (S. Gelfand).
891. Bayer, DE-OS 2 523 104, 1975 (H. Rathjen).
892. Tenneco Chem. Inc., US 4 031 146, 1977 (E. P. Di Bella).
893. Bayer, EP 46 555, 1982 (G.-M. Petruck, R. Wambach).
894. Shell Int. Res., GB 1 110 030, 1968 (C. F. Kohll, H. D. Scharf, R. Van Helden).
895. Shell Int. Res., NL 6 907 390, 1970 (D. A. Was).
896. Rhône-Poulenc Ind., FR 2 475 535, 1981 (J.-C. Lauct, J. Bourdon).
897. Toray, EP 125 077, 1984 (K. Miwa et al.).
898. Tenneco Chem. Inc., US 4 031 145, 1977 (E. P. Di Bella).
899. T. Asinger, G. Loch, *Monatsh. Chem.* **62** (1933) 345.
900. *Ullmann*, 4th ed., **9**, pp 513–514.
901. H. C. Brimelow, R. L. Jones, T. P. Metcalfe, *J. Chem. Soc.* 1951, 1208–1212.
902. Tenneco Chem. Inc., US 3 517 056, 1970 (J. F. De Gaetano).
903. Tenneco Chem. Inc., US 3 692 850, 1972 (E. P. Di Bella).
904. Kureha Chem. Ind. Co., Ltd., JP-Kokai 70/28 367, 1965 (M. Ishida); *Chem. Abstr.* **74** (1971) 3397.
905. A. L. Englin et al., SU 1 727 736, 1965.
906. Swann Research, Inc., US 1 836 180, 1929 (C. R. McCullough, R. C. Jenkins).
907. C. H. Penning, *Ind. Eng. Chem.* **22** (1930) no. 11, 1180–1182.
908. K. Soldner, G. Gollmer, *Elektrizitätswirtschaft* **81** (1982) no. 17/18, 573–583.
909. H. R. Buser, H.-P. Bosshardt, C. Rappe, *Chemosphere* **1** (1978) 109–119.
910. H. R. Buser, C. Rappe, *Chemosphere* **3** (1979) 157–174.
911. R. R. Bumb et al., *Science (Washington, DC)* **210** (1980) no. 4468, 385–390.
912. G. G. Choudhry, O. Hutzinger, *Toxicol. Environ. Chem. Rev.* **5** (1982) 1–65, 67–93, 97–151, 153–165 and 295–309.
913. A. Schecter, *Chemosphere* **12** (1983) no. 4/5, 669–680.
914. Environment Committee, Paris, March 1982, Report on the implementation by member countries of decision by the council on the protection of the environment by control of polychlorinated biphenyls, ENV (82) 5.
915. *Ullmann*, 4th ed., **9**, pp. 515–518.
916. O. Hutzinger, S. Safe, V. Zitko: *The Chemistry of PCB's*, CRC Press, Cleveland 1974.
917. Zehnte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Tenth Implementing Order Issued Pursuant to the Federal Immission Protection Act of the FRG); restrictions on PCB, PCT and VC. July 26, 1978.
918. Council Directive of July 27, 1976, harmonizing the legal and administrative provisions of the member states restricting the sale and use of dangerous substances and preparations (76/769/EEC).

919. Gesetz über die Beseitigung von Abfällen (Waste Disposal Act) of January 5, 1977 (FRG).
920. Verordnung zur Bestimmung von Abfällen nach § 2 Abs. 2 des Abfallbeseitigungsgesetzes (Order Regulating the Classification of Wastes According to § 2 Para. 2 of the Waste Disposal Act), May 24, 1977 (FRG).
921. Gesetz über Maßnahmen zur Sicherung von Altölbeseitigung (act embodying measures to ensure the disposal of scrap oil), December 11, 1979 (FRG).
922. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (First General Administrative Provision Issued Pursuant to the Federal Imission Protection Act)–Technische Anleitung Luft (Technical Guideline – Air). August 28, 1974 (FRG).
923. U.S. Public Law 94–469, October 11, 1976, Section 6 (e) of TSCA.
924. U.S. Environmental Protection Agency 560/6–75–004, 1976 (Proceedings of the National Conference on PCBs, Chicago, November 19 – 21, 1975).
925. *U.S. Environmental Protection Agency 230–03/ 79–001*, 1979, Washington, D.C., (Support Document/Voluntary Environmental Impact Statement for PCBs).
926. NIOSH, Washington, D.C., U.S. Government Printing Office, 1977, pp. 40–49 (summary of the literature on the toxicity of polychlorinated biphenyls).
927. C. Nels, *Elektrizitätswirtschaft* **83** (1984) no. 8, 375–379 (disposal of askarels).
928. M. G. McGraw, *Electrical World, (The PCB Problem: separating fact from fiction)* February 1983, pp. 49–72.
929. G. T. Hunt, P. Wolf, P. F. Fennelly, *Environ. Sci. Technol.* **18** (1984) 171–179.
930. *NATO-CC MS Report II, (Disposal of Hazardous Wastes – Thermal Treatment)* March 23, 1981, pp. 84–89.
931. General Electric Co., US 4 353 798, 1982 (S. D. Foss).
932. General Electric Co., US 4 353 793, 1982 (D. J. Brunelle).
933. General Electric Co., US 4 351 718, 1982 (D. J. Brunelle).
934. Goodyear Tire & Rubber Co., US 4 284 516, 1981 (D. K. Parker, R. J. Steichen). D. K. Parker, W. L. Cox, *Plant Eng.* (1980) Aug. 21, 133–134.
935. Sunohio, GB 2 063 908, 1981 (O. C. Norman, L. H. Handler); GB 2 081 298, 1982 (O. D. Jordan).
936. ITO T., JP-Kokai 81/111 078, 1980.
937. Fuji Elektric Co., Ltd., DE-OS 3 036 000, 1980 (Masuda et al.).
938. University of Waterloo, US 4 326 090, 1982 (J. G. Schmitz, G. L. Bubbar).
939. The Franklin Institute, DE-OS 3 033 170, 1980; EP 60 089, 1982 (L. L. Pytlewski, K. Krevitz, A. B. Smith).
940. ITO T., JP Kokai 82/29 373, 1980; *Chem. Abstr.* **96** (1982) 222 772.
941. Queens University, Kingston ONT, US 4 345 983, 1982 (J. K. S. Wan).
942. Tokyo Shibaura Elec., Ltd., JP-Kokai 82/166 175, 1981; *Chem. Abstr.* **98** (1983) 59 404.
943. Rockwell Int. Corp., WO 8 002 116; EP 26 201, 1980 (L. F. Grantham et al.).
944. L. Bjorklund, I. Fäldt, WO 8 200 509, 1980.
945. SOTOO Z., JP-Kokai 79/151 953, 1978; *Chem. Abstr.* **93** (1980) 31 358.
946. ICI Australia, Ltd., WO 8 202 001, 1980 (V. Vasak, K. K. Mok, J. Sencar).
947. K. Ballschmiter, M. Zell, *Fresenius Z. Anal. Chem.* **302** (1980) 20–31.
948. H. M. Klimisch, D. N. Ingebrigtson, *Anal. Chem.* **52** (1980) no. 11, 1675–1678.
949. P. W. Albro, J. T. Corbett, J. C. Schröder, *J. Chromatogr.* **205** (1981) 103–111.
950. D. A. Newton et al., *J. Chromatogr. Sci.* **21** (1983) 161.
951. E. Schulte, R. Malisch, *Fresenius Z. Anal. Chem.* **314** (1983) 545–551 and 319 (1984) 54–59.
952. W. J. Dunn et al., *Anal. Chem.* **56** (1984) no. 8, 1308–1313.
953. R. H. Lin et al., *Anal. Chem.* **56** (1984) no. 11, 1808–1812.
954. Versar Inc., National Technical Informal Service, *Industrial Use and Environmental Distribution* (NTIS 252–012/3 W P) 1976.
955. Fireproof Products Co., US 914 222, 1909; US 914 223, 1909; Halogen Products Co., US 1 111 289, 1913 (J. W. Aylsworth).
956. F. D. Kover, *Environmental Hazard Assessment Report*, Chlorinated Naphthalenes, Environmental Protection Agency 560/8–75–001, 1975.
957. *U.*S. Federal Register vol. **48**, no. 89 of May 6, 1983, pp. 20 668–20 679; vol. **49**, no. 166 of August 24, 1984, pp. 33 649–33 654.

958. U. A. Th. Brinkman, H. G. M. Reymer, *J. Chromatogr.* **127** (1976) 203–243.
959. AUER-Technikum, no. 10, 1982, Published by Auergesellschaft GmbH, Berlin 44 (FRG).
960. P. Ferrero, C. Caflisch, *Helv. Chim. Acta* **11** (1928) 795–812.
961. D. H. S. Horn, F. L. Warrer, *J. Chem. Soc.* 1946, 144.
962. Frontier Chemical Co. (Division of Vulcan Materials Co.), US 3 413 357, 1968 (K. F. Bursack, E. L. Johnston, H. J. Moltzau).
963. E. G. Turner, W. P. Wynne, *J. Chem. Soc.* 1941, 244–245.
964. W. Schwemberger, W. Gordon, *Zh. Obshch. Khim.* **2** (1932) 926;
5 (1934) 695.
965. Bayer, DE 844 143, 1951;
DE 946 055, 1953 (H. Vollmann).
966. A. Ledwith, P. J. Russel, *J. Chem. Soc. Perkin Trans.* **2** (1975) no. 14, 1503–1508.
967. Dow Chemical Co., US 1 917 822, 1933 (E. C. Britton, W. R. Reed).
968. *Ullmann*, 4th ed., **9**, pp. 512–520.
969. S. Safe, O. Hutzinger, *J. C. S. Chem. Commun.* 1972, 260–262.
970. M. P. Gulan, D. D. Bills, T. B. Putnam, *Bull. Environ. Contam. Toxicol.* **11** (1974) no. 5, 438–441.
971. U. A. T. Brinkman, A. De Kok, *J. Chromatogr.* **129** (1976) 451.
972. U. A. T. Brinkman et al., *J. Chromatogr.* **152** (1978) 97–104.
973. M. Cooke et al., *J. Chromatogr.* **156** (1978) 293–299.
974. P. W. Albro, C. E. Parker, *J. Chromatogr.* **197** (1980) 155–169.
975. P. A. Kennedy, D. J. Roberts, M. Cooke, *J. Chromatogr.* **249** (1982) 257–265.
976. J. J. Donkerbroek, *J. Chromatogr.* **255** (1983) 581–590.
977. Koppers Company, Inc., U.S., Product bulletin: *Halowaxes, Chlorinated Naphthalene Oils and Waxlike Solids*.
978. J. Todt, *Chem. Ing. Tech.* **54** (1982) no. 10, 932.
979. Hodogaya Chem. Co., Ltd., JP-Kokai 79/78 399, 1977 (K. Imaizumi, A. Arai, G. Tsunoda);
Chem. Abstr. **91** (1979) 142 751.
980. Sumitomo Chem. Co., Ltd., JP-Kokai 79/43 194, 1977 (Y. Matsushita, J. Kamimura);
Chem. Abstr. **91** (1979) 76 268.
981. M. F. Nathan, *Chem. Eng.* Jan. 30. 1978, 93–100.
982. Celanese Corp., US 4 276 179, 1981 (J. W. Soehngen).
983. Hooker Chem. Corp., US 2 949 491, 1960 (J. R. Rucker).
984. Osaka Prefecture, JP-Kokai 78/84 923, 1976 (Y. Hatano);
Chem. Abstr. **89** (1978) 163 306.
985. Teijin K. K., JP-Kokai 79/59 233, 1977 (T. Yamaji);
Chem. Abstr. **91** (1979) 192 958.

General References

986. *Beilstein*, Benzyl chloride, 5, 292, **5** (1), 151, **5** (2) 227, **5** (3), 685, **5** (4) 809.
987. *Beilstein*, Benzalchloride, 5, 297, **5** (1), 152, **5** (2), 232, **5** (3), 696, **5** (4), 817.
988. *Beilstein*, Benzotrichloride, 5, 300, **5** (1), 152, **5** (2), 233, **5** (3), 699, **5** (4), 820.
989. *Beilstein*, o o-Xylene: 5, 364, **5** (1), 180, **5** (2), 283, **5** (3), 816, **5** (4), 926.
990. *Beilstein*, m-Xylene: 5, 373, **5** (1), 183, **5** (2), 291, **5** (3), 835, **5** (4), 943.
991. *Beilstein*, p-Xylene: 5, 384, **5** (1), 186, **5** (2), 299, **5** (3), 854, **5** (4), 966.
992. *Houben-Weyl*, Chlorine compounds 5/3, 503;
993. *Houben-Weyl*, Photochemistry: 4/5 a, 1.
994. *Ullmann* 4th ed., **3**, 305–319.

Specific References

995. *Kirk-Othmer* **5**, 828–838.
996. L. H. Horsley, *Azeotropic Data III*, Advances in Chemistry Series No. 116; Amer. Chem. Soc., Washington DC, 1973.
997. A. V. Egunov, B. J. Konobeev, E. A. Ryabov, T. J. Gubanova, *Zh. Prikl. Khim. (Leningrad)* **46** (1973) no. 8, 1855–1856;
Chem. Abstr. **79** (1973) 140 122 x.
998. H. C. Haas, D. J. Livingston, M. Saunders, *J. Polym. Sci.* **15** (1955) 503.
999. R. Pieck, J. C. Jungers, *Bull. Soc. Chim. Belg.* **60** (1951) 357–384.
1000. M. L. Poutsma in E. S. Huyser (ed.): *Methods in Free-Radical Chemistry*, Free-Radical Chlorination, vol. **I**, Marcel Dekker, New York 1969, pp. 79–193.
1001. H. G. Haring, H. W. Knol, *Chem. Process Eng.* **1964**, no. 10, 560–567; **1964**, no. 11, 619–623; **1964**, no. 12, 690–693; **1965**, no. 1, 38–40.
1002. H. G. Haring: "Chlorination of Toluene," *Chem. Eng. Monog.* **15** (1982) 193–208, 208 A.
1003. J. Y. Yang, C. C. Thomas, Jr., H. T. Cullinan, *Ind. Eng. Chem. Process Des. Dev.* **9** (1970) no. 2, 214–222.
1004. W. Hirschkind, *Ind. Eng. Chem.* **41** (1949) no. 12, 2749–2752.

1005. BASF, DE-AS 1 245 917, 1967 (W. Beckmann).
BASF, DE-AS 1 249 831, 1967 (W. Beckmann, W. L. Kengelbach, H. Metzger, M. Pape).
Toyo Rayon Kabushiki Kaisha, DE-AS 1 191 342, 1965 (Y. Ito, R. Endoh).
1006. Bayer, DE-OS 2 152 608, 1973;
DE-OS 2 227 337, 1974;
US 3 816 287, 1974 (W. Böckmann, R. Hornung).
1007. Hoechst, DE-AS 2 530 094, 1976 (R. Lademann, F. Landauer, H. Lenzmann, K. Schmiedel, W. Schwiersch).
1008. T. Yokota, T. Iwano, A. Saito, T. Tadaki, *Int. Chem. Eng.* **23** (1983) no. 3, 494–502.
1009. S. K. Fong, J. S. Ratcliffe, *Mech. Chem. Eng. Trans.* **8** (1972) no. 1, 1–8, 9–14.
1010. Hoechst, DE-AS 2 139 779, 1972 (H. Schubert, K. Baessler).
1011. The Goodyear Tire and Rubber Co., US 2 844 635, 1958 (R. H. Mayor).
1012. The Goodyear Tire and Rubber Co., US 2 817 632, 1957 (R. H. Mayor).
1013. Nihon Nohyaku Co., Ltd., Japan, *Kokai Tokkyo Koho* **80** (1980) 113 729;
Chem. Abstr. **94** (1981) 174 606 h.
Hodogaya Chemical Co., Ltd., JP-Kokai 76 08 221, 1976 (R. Sasaki, T. Kanashiki, K. Fukazawa);
Chem. Abstr. **84** (1976) 164 373 j.
1014. Hooker Electrochemical Co., US 2 695 873, 1954 (A. Loverde).
Mitsubishi Gas Chemical Co., Inc., DE-OS 2 303 115, 1973 (T. Sano, M. Doya).
1015. Hodogaya Chemical Co., Ltd., JP-Kokai 76 08 223, 1976 (M. Fuseda, K. Ezaki);
Chem. Abstr. **85** (1976) 32 599 y.
Hodogaya Chemical Co., Ltd., JP-Kokai 76 08 222, 1976 (R. Sasaki, T. Kanashiki, K. Fukazawa);
Chem. Abstr. **84** (1976) 164 374 k.
1016. Dow Chemical Co., Ltd., GB 2 087 377 A, 1982 (D. P. Clifford, R. A. Sewell);
Diamond Alkali Co., US 3 350 467, 1967 (R. H. Lasco).
G. Messina, M. Moretti, P. Ficcadenti, G. Cancellieri, *J. Org. Chem.* **44** (1979) no. 13, 2270–2274.
1017. Velsicol Chemical Co., US 3 363 013, 1968 (G. D. Kyker);
The Goodyear Tire and Rubber Co., US 2 817 633, 1957;
US 2 817 632, 1957 (R. H. Mayor).
1018. Diamond Alkali Co., US 3 580 854, 1971;
US 3 703 473, 1972 (R. H. Lasco).
1019. Hodogaya Chemical Co., Ltd., JP-Kokai 82 98 224, 1982;
Chem. Abstr. **97** (1982) 181 930 z.
1020. Hodogaya Chemical Co., Ltd., JP-Kokai 82 98 225, 1982;
Chem. Abstr. **97** (1982) 162 552 m.
1021. S. Niki, JP-Kokai 77 111 520, 1977;
Chem. Abstr. **88** (1978) 50 451 k.
1022. Shin-Etsu Chemical Industry Co., Ltd., JP-Kokai 78 77 022, 1978 (J. Hasegawa, Y. Kobayashi, T. Shimizu);
Chem. Abstr. **89** (1978) 215 045 s.
J. Besta, M. Soulek, CS 179 600, 1979;
Chem. Abstr. **92** (1980) 58 414 d.
1023. W. E. Vaughan, F. F. Rust, *J. Org. Chem.* **5** (1940) 449–471, 472–503.
1024. M. S. Kharasch, M. G. Berkman, *J. Org. Chem.* **6** (1941) 810–817.
1025. Y. A. Serguchev, V. F. Moshkin, Y. V. Konoval, G. A. Stetsyuk, *Zh. Org. Khim.* **19** (1983) no. 5, 1020–1023;
Chem. Abstr. **99** (1983) 69 885 d.
1026. M. Ritchie, W. I. H. Winning, *J. Chem. Soc.* 1950, 3579–3583.
1027. G. Benoy, L. de Maeyer, *R. 27e Congr. Int. Chim Ind.* **2** (1954);
Ind. Chim. Belge, **20** (1955) 160–162.
1028. H. P. Dürkes, DE-OS 1 643 922, 1971;
Du Pont de Nemours Co., US 2 446 430, 1948 (J. A. Norton).
Mitsubishi Chemical Industries Co., Ltd., JP-Kokai 75 62 906, 1975 (S. Ueda, K. Zaiga);
Chem. Abstr. **83** (1975) 113 912 f.
1029. E. Clippinger (Chevron Res. Co.) *Am. Chem. Soc., Div. Pet. Chem., Prepr.* **15** (1970) no. 1, B37–B40.
1030. V. A. Averyanov, S. E. Kirichenko, V. F. Shvets, Y. A. Treger, *Zh. Fiz. Khim* **56** (1982) no. 5, 1136–1140;
Chem. Abstr. **97** (1982) 38 235 w.
1031. VEB Farbenfabrik Wolfen, DD 8523, 1954 (A. Weissenborn).
Ullmann, 4th ed., **9**, 530.
1032. V. R. Rozenberg et al., SU 687 061, 1979;
Chem. Abstr. **92** (1980) 76 078 b.
1033. J. C. André et al. a) *J. Photochem.* **18** (1982) 47–45;
Chem. Abstr. **96** (1982) 133 058 d;
18 (1982) 57–79;
Chem. Abstr. **96** (1982) 131 388 n;
22 (1983) 7–24;
Chem. Abstr. **99** (1983) 37 958 p;
22 (1983) 137–155;

Chem. Abstr. **99** (1983) 139 410 w;
22 (1983) 213–221;
Chem. Abstr. **99** (1983) 79 916 d;
22 (1983) 223–232;
Chem. Abstr. **99** (1983) 96 704 a;
22 (1983) 313–332;
Chem. Abstr. **99** (1983) 141 891 x.
b) *React. Kinet. Catal. Lett.* **16** (1981) no. 2–3, 171–176;
Chem. Abstr. **95** (1981) 123 904 s;
16 (1981) no. 2–3, 177–183;
Chem. Abstr. **95** (1981) 122 382 h;
17 (1981) no. 3–4, 433–437;
Chem. Abstr. **96** (1982) 43 767 e;
21 (1982) no. 1–2, 1–5;
Chem. Abstr. **98** (1983) 88 526 h.
c) *J. Chim. Phys. Phys.-Chim. Biol.* **79** (1982) no. 7–8, 613–616;
Chem. Abstr. **98** (1983) 178 410 v.
d) *AIChE J.* **28** (1982) no. 1, 156–166;
Chem. Abstr. **97** (1982) 82 599 s.
e) *Entropie* **18** (1982) no. 107–108, 62–81;
Chem. Abstr. **98** (1983) 129 139 e.
f) *Oxid. Commun.* **4** (1983) no. 1–4, 13–26;
Chem. Abstr. **101** (1984) 229 721 n.

1034. B. Bonath, B. Foertsch, R. Saemann (Geigy AG), *Chemie. Ing.-Tech.* **38** (1966) no. 7, 739–742.
1035. E. Stoeva, Z. Angelova, M. Gramatikova, Z. Zhelyazkov, *Khim. Ind. (Sofia)* 1982, no. 7, 294–296;
Chem. Abstr. **98** (1983) 74 188 n.
1036. E. Borello, D. Pepori, *Ann. Chim. (Rome)* **45** (1955) 449–466.
1037. J. S. Ratcliffe, *Br. Chem. Eng.* **11** (1966) no. 12, 1535–1537.
1038. Bayer, DE-OS 2 003 932, 1971;
US 3 715 283, 1973 (W. Böckmann).
1039. ICI, DE-OS 2 105 254, 1971 (E. Illingworth, A. Fleming).
M. Vrana, V. Janderova, M. Danek, CS 134 477, 1969;
Chem. Abstr. **75** (1971) 5452 g.
1040. J. Besta, M. Soulek, CS 159 100, 1975;
Chem. Abstr. **84** (1976) 164 371 g.
Albright and Wilson Ltd., GB 1 410 474, 1975 (C. H. G. Hands).
1041. A. Scipioni, *Ann. Chim. (Rome)* **41** (1951) 491–498.
1042. Occidental Chemical Co., DE 2 604 276, 1984 (S. Gelfand).
1043. Hercules Powder Co., US 2 108 937, 1938 (L. H. Fisher).
Hercules Powder Co., US 2 100 822, 1937 (H. M. Spurlin).
1044. I.G. Farbenindustrie AG, DE 659 927, 1936 (O. Leuchs).
G. L. Kamalov et al., SU 872 525, 1981;
Chem. Abstr. **96** (1982) 85 217 f.
1045. H. C. Brown, K. L. Nelson, *J. Am. Chem. Soc.* **75** (1953) 6292.
R. C. Fuson, C. H. Mc Keever: *Organic Reactions*, vol. **1**, J. Wiley & Sons, New York 1947, p. 63.
Heyden Newport Chem. Co., US 2 859 253, 1958 (J. E. Snow).
1046. Kureha Chem. Ind. Co., JP 2173/66, 1966.
1047. FMC, US 2 493 427, 1950 (R. M. Thomas).
1048. Mitsui Toatsu Chemicals Co., Ltd., JP-Kokai 73 05 725, 1973 (N. Kato, Y. Sato);
Chem. Abstr. **78** (1973) 97 293 u.
1049. Velsicol Chemical Co., US 3 535 391, 1970 (G. D. Kyker).
1050. Bayer, DE-AS 2 206 300, 1974 (W. Böckmann, K. A. Lipper).
1051. Mitsui Toatsu Chemicals, Inc., JP-Kokai 73 05 726, 1972 (N. Kato, Y. Sato);
Chem. Abstr. **84** (1976) 121 408 g;
cf. *Chem. Abstr.* **78** (1973) 97 294 v.
1052. Monsanto, US 2 542 225, 1951 (J. L. West).
1053. Dow Chemical, DE-OS 2 410 248, 1974 (R. H. Hall, D. H. Haigh, W. L. Archer, P. West).
1054. EKA AB, EP-A 64 486, 1981 (R. K. Rantala, G. L. F. Hag).
1055. Bayer, DE-OS 2 752 612, 1977 (F. Brühne, K. A. Lipper).
1056. Ciba Geigy AG, DE-AS 2 044 832, 1969 (P. Liechti, F. Blattner).
1057. GAF, US 3 087 967, 1960 (D. E. Graham, W. C. Craig).
1058. Tenneco Chemicals Inc., US 3 524 885, 1967 (A. J. Deinet).
1059. Kureha Chem. Ind. Co., Ltd., JP 69 12 132, 1969 (H. Funamoto);
Chem. Abstr. **71** (1969) 80 948 u.
1060. J. H. Simons, C. J. Lewis, *J. Am. Chem. Soc.* **60** (1938) 492.
1061. E. Pouterman, A. Girardet, *Helv. Chim. Acta* **30** (1947) 107.
1062. S. F. Khokhlov et al., SU 530 019, 1976;
Chem. Abstr. **86** (1977) 43 386 y;
cf *Chem. Abstr.* **85** (1976) 142 731 t;
Chem. Abstr. **87** (1977) 22 600 u.
1063. Rütgerswerke AG, DE-AS 1 200 275, 1965 (G. Bison, H. Binder).
1064. Mitsubishi Gas Chemical Co., Inc., DE-AS 2 443 179, 1977 (S. Yoshinaka, M. Doya, S. Uchiyama).

1065. Dow Chemical, US 4 046 656, 1977 (R. A. Davis, R. G. Pews);
J. Chem. Soc., Chem. Commun. 1978, no. 3, 105–106.
1066. Bayer, DE-OS 1 909 523, 1970 (W. Böckmann).
1067. Diamond Alkali Co., US 2 979 448, 1961 (G. A. Miller).
1068. Diamond Alkali Co., US 2 994 653, 1959 (G. A. Miller).
1069. Asahi Electro-Chemical Co., JP 28 788 ('65), 1964 (K. Harasawa);
Chem. Abstr. **64** (1966) 9633 e.
1070. Dynamit Nobel AG, DE-OS 2 461 479, 1976 (H. Richtzenhain, P. Riegger).
Dynamit Nobel AG, DE-OS 2 535 696, 1977 (P. Riegger, H. Richtzenhain, G. Zoche).
1071. P. G. Harvey, F. Smith, M. Stacey, J. C. Tatlow, *J. Appl. Chem. (London)* **4** (1954) 319–325.
1072. Mitsubishi Gas Chemical Co., DE-OS 2 614 139, 1976 (S. Yoshinaka, M. Doya, S. Uchiyama, S. Nozaki).
1073. P. Beltrame, S. Carrà, S. Mori, *J. Phys. Chem.* **70** (1966) no. 4, 1150–1158;
Tetrahedron Lett. **44** (1965) 3909–3915.
1074. H. Fürst, H. Thorand, J. Laukner, *Chem. Tech. (Berlin)* **20** (1968) no. 1, 38–40.
1075. P. Riegger, K. D. Steffen, *Chem. Ztg.* **103** (1979) no. 1, 1–7.
1076. V. I. Titov, G. S. Mironov, I. V. Budnii, *Osnovn. Org. Sint. Neftekhim* **13** (1980) 93–96;
Chem. Abstr. **97** (1982) 162 006 m.
1077. O. Cerny, J. Hajek, *Collect. Czecho. Chem. Commun.* **26** (1961) 478–484.
1078. L. M. Kosheleva, V. R. Rozenberg, G. V. Motsarev, *Zh. Org. Khim.* **16** (1980) no. 9, 1890–1893;
Chem. Abstr. **94** (1981) 14 815 x.
1079. Bayer, EP 9787, 1982;
DE-OS 2 844 270, 1978;
DE-OS 2 905 081, 1979 (R. Schubart, E. Klauke, K. Naumann, R. Fuchs);
ICI EP-A 28 881, 1981 (J. O. Morley).
1080. Mitsubishi Gas Chemical Co., Inc., JP 8 225 009, 1982;
Chem. Abstr. **97** (1982) 184 357 d.
1081. Bergwerksverband GmbH, GB 830 052, 1960.
1082. T. Nishi, J. Onodera, *Kogyo Kagaku Zasshi* **71** (1968) no. 6, 869–871;
Chem. Abstr. **70** (1969) 3381 f.
1083. H. Trautmann, W. Seidel, K. Seiffarth, DD 116 451, 1975.
1084. Mitsui Toatsu Chemicals, Inc., JP-Kokai 78 65 830, 1978 (M. Oba, M. Kawamata, T. Shimokawa, S. Koga);
Chem. Abstr. **90** (1979) 6063 c.
1085. Dynamit Nobel AG, DE-OS 2 161 006, 1973 (K. Redecker, H. Richtzenhain).
1086. BASF, DE 845 503, 1943 (W. Rohland).
1087. Hodogaya Chem. Co., GB 1 442 122, 1974.
1088. Dynamit Nobel AG, DE-OS 2 358 949, 1975 (G. Blumenfeld).
1089. T. R. Torkelson, V. K. Rowe in G. D. Clayton, F. L. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, 3rd ed., vol. **2 B**, Wiley-Interscience, New York 1981, pp. 3433–3601.
1090. American Conference of Governmental Industrial Hygienists Inc. Documentation of the Threshold Limit Values 1980 (With Annual Supplements) 6500 Glenway Bldg. D-5, Cincinnati, OH 45211.
1091. Commission for Investigation of the Health Hazards of Chemical Compounds in the Work Environment:
Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten, Verlag Chemie, Weinheim 1984.
1092. Monograph on the Evaluation of the Carcinogenic Risk of Chemicals to Humans. Chemicals, Industrial Processes and Industries Associated with Cancer in Humans.
IARC Monogr. Suppl. **1–29** (1982) no. 4.
1093. J. A. John, D. J. Wroblewski, B. A. Schwetz: Teratogenicity of Experimental and Occupational Exposure to Industrial Chemicals in H. Kalter, ed.: *Issues and Reviews in Teratology*, vol. **2**, Plenum, 1984, p. 267–324.
1094. American Conference of Governmental Industrial Hygienists. TLV's® Threshold Limit Values for Chemical Substances in the Work *Environment Adopted by the ACGIH for 1985–1986*. ACGIH 6500 Glenway Bldg. D-5, Cincinnati, OH 45211.
1095. Deutsche Forschungsgemeinschaft: MAK, Verlag Chemie, Weinheim 1984.
1096. The Chemical Industry Institute of Toxicology: *Unpublished data on methyl chloride*. Research Triangle Park, North Carolina 27709.
1097. B. N. Ames, P. F. Infante, R. H. Reitz, *Banbury Report 5. Ethylene Dichloride: A Potential Health Risk*, Cold Spring Harbor Laboratory, Maine, USA, 1980.

1098. R. J. Kociba, B. A. Schwetz, D. G. Keyes, G. C. Jersey, J. J. Ballard, D. A. Dittenber, J. F. Quast, C. E. Wade, C. G. Humiston: *Environ. Health* **21** (1977) 49.
1099. Ambient Water Quality Criteria for: Halomethanes, Chloroform, Carbon Tetrachloride, Chlorinated Ethanes, Trichloroethylene and Tetrachloroethylene. U.S. Environmental Protection Agency, Oct. 1980.
1100. M. A. Mayes, H. C. Alexander, D. C. Dill, *Bull. Environ. Contam. Toxicol.* **31** (1983) 139–147.
1101. Dow unpublished data.
1102. C. Bushon, *Ecological Effects Branch Chief*, Office of Pesticides and Toxic Substances. U.S. EPA, Washington, DC, Dec. 16, 1983.
1103. Y. Correia, G. J. Martens, F. H. van Mensch, B. P. Whim, *Atmos. Environ.* **11** (1977) 1113–1116.
1104. P. R. Edwards, I. Campbell, G. S. Milne, *Chem. Ind.* **1** (1982) 574–578.
1105. K. K. Beutel: *Health and Environmental Aspects of Chlorinated Solvents in Industrial Applications*, Juris Druck + Verlag, Zürich 1984.
1106. W. L. Dilling in R. A. Conway (ed.): *Environmental Risk Analysis for Chemicals*, Van Nostrand Reinhold Co., New York 1982, pp. 166–174.
1107. M. A. K. Khalil, R. A. Rasmussen, *Chemosphere* **7** (1984) 789–800.
1108. National Academy of Sciences: *Nonfluorinated Halomethanes in the Environment*, Washington 1978, p. 97.
1109. NIOSH: Registry of Toxic Effects of Chemical Substances, U.S. Dept. of Health, Education and Welfare, NIOSH (1983).
1110. E. Loeser, M. H. Litchfield, *Fd. Chem. Toxic.* **21** (1983) 825–832.
1111. J. Kohli et al., *Can. J. Biochem.* **54** (1976) 203.
1112. R. D. Lingg et al., *Drug Metab. Dispos.* **10** (1982) 134.
1113. J. H. Koeman et al., *Nature (London)* **221** (1969) 126.
1114. L. Acker, R. Schilte, *Naturwissenschaften* **57** (1970) 497.
1115. F. K. Zimmermann et al., *Mutat. Res.* **133** (1984) 199.
1116. L. Fishbein, *J. Chromatogr.* **68** (1972) 345.
1117. M. Ogawa, *Fukuoka Igaku Zasshi* **62** (1971) 74;
cited in: J. R. Allen, D. H. Norback, *Science (Washington, D.C.)* **179** (1973) 498.
1118. J. H. Miller, *Public Health Rep.* **59** (1944) 1085.
1119. Report of the Cooperative Research Program between the Gesellschaft für Strahlen- und Umweltforschung mbH, München (GSF) and the Albany Medical College of the Union University, Albany, N.Y., 1974 – 1978, GSF-Bericht Ö 486.
1120. B. Bush et al., *Arch. Environ. Contam. Toxicol.* **13** (1984) 517.
1121. A. Hofmann et al., *Arbeitsmed. Sozialmed. Präventivmed.* **18** (1983) 181.
1122. U. A. T. Brinkman, H. G. M. Reymer, *J. Chromatogr.* **127** (1976) 203.
1123. G. Löfroth et al., *Mutat. Res.* **155** (1985) 91.
1124. S. Haworth et al., *Environ. Mutagen.* **5** (1983) Suppl. 1, 3.
1125. H. Druckrey et al., *Z. Krebsforsch.* **74** (1970) 241.
1126. W. F. von Oettingen: U.S. Department of Health, Education and Welfare, Public Health Service Publication No. 414, U.S. Government Printing Office, Washington, D.C., 1955.
1127. K. Yasuo et al., *Mutat. Res.* **58** (1978) 143.
1128. J. McCann et al., *Proc Natl. Acad. Sci. U.S.A.* **72** (1975) 979.
1129. K. Fukuda et al., *Gann* **72** (1981) 655.
1130. H. G. Bray et al., *Biochem. J.* **70** (1958) 570.
1131. Bayer AG, Bericht Nr. 8418 (1979), unpublished
1132. Verordnung über gefährliche Arbeitsstoffe (ArbStoffV), Fassung vom 11. Februar 1982, (BGBl. I. S. 144).
1133. M. Windholz (ed.): *The Merck Index*, Merck & Co., Rathway 1976.
1134. Bayer AG, Bericht Nr. 7839 (1978), unpublished.
1135. H. F. Smyth et al., *Arch. Ind. Hyg. Occup. Med.* **4** (1951) 119.
1136. H. Sakabe et al.: The New York Academy of Sciences, TSCA Sec. 8(e) Submission 8EHQ-0777–0001.
1137. H. Sakabe, K. Fukuda, *Ind. Health* **15** (1977) 173.

Chlorine

Peter Schmittinger, CREANOVA GmbH, Werk Lülsdorf, Niederkassel, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 9, 10, 11, 12, 13, 14 and 15)

Thomas Florkiewicz, OxyTech Systems, Chardon, Ohio, United States (Chap. 6 and 9)

L. Calvert Curlin, OxyTech Systems, Chardon, Ohio, United States (Chap. 6 and 9)

Benno Lüke, Uhde, Dortmund, Federal Republic of Germany (Chap. 7 and 9)

Robert Scannell, DeNora Deutschland GmbH, Rodenbach, Federal Republic of Germany (Chap. 8.1)

Thomas Navin, OxyTech Systems Inc., Chardon, Ohio, United States (Chap. 8.2)

Erich Zelfel, Infraserv, Knapsack, Federal Republic of Germany (Chap. 12)

Rüdiger Bartsch, Technische Universität München, München, Federal Republic of Germany (Chap. 16)

1. Introduction 2
2. Physical Properties 3
3. Chemical Properties 6
4. Chlor-Alkali Process 8
4.1. Brine Supply 11
4.2. Electricity Supply 13
5. Mercury Cell Process 13
5.1. Principles 14
5.2. Mercury Cells 19
5.2.1. Uhde Cell 19
5.2.2. De Nora Cell 20
5.2.3. Olin – Mathieson Cell 21
5.2.4. Solvay Cell 21
5.3. Operation 21
5.3.1. Brine System 21
5.3.2. Cell Room 23
5.3.3. Treatment of the Products 24
5.3.4. Measurement 24
5.3.5. Mercury Emissions 25
6. Diaphragm Process 27
6.1. Principles 28
6.2. Diaphragm Cells 31
6.2.1. Dow Cell 33
6.2.2. Glanor Electrolyzer 34
6.2.3. OxyTech "Hooker" Cells 35
6.2.4. HU Monopolar Cells 35
6.2.5. OxyTech MDC Cells 37
6.3. Operation 37
6.3.1. Brine System 37
6.3.2. Cell Room 39
6.3.3. Diaphragm Aging 41
6.3.4. Treatment of the Products 41
6.3.5. Measurement 43
7. Membrane Process 43
7.1. Principles 44
7.2. Process Specific Aspects 46
7.2.1. Brine Purification 47
7.2.2. Commercial Membranes 50
7.2.3. Power Consumption 51
7.2.4. Product Quality 51
7.3. Membrane Cells 51
7.3.1. Monopolar and Bipolar Designs . . 51
7.3.2. Commercial Electrolyzers 53
7.3.3. Comparison of Electrolyzers 57
7.3.4. Cell Room 57
8. Electrodes 57
8.1. Anodes 57
8.1.1. General Properties of the Anodes . 59
8.1.2. Anodes for Mercury Cells 60
8.1.3. Anodes for Diaphragm Cells 61
8.1.4. Anodes for Membrane Cells 62
8.2. Activated Cathode Coatings 62
9. Comparison of the Processes . . . 63
9.1. Product Quality 64
9.2. Economics 64
9.2.1. Equipment 64
9.2.2. Operating Costs 66
9.2.3. Summary 66
10. Other Production Processes 67
10.1. Electrolysis of Hydrochloric Acid 67
10.2. Chemical Processes 69
10.2.1. Catalytic Oxidation of Hydrogen Chloride by Oxygen 69
10.2.2. Oxidation of Hydrogen Chloride by Nitric Acid 71
10.2.3. Production of Chlorine from Chlorides 71
11. Chlorine Purification and Liquefaction 71
11.1. Cooling 71
11.2. Chlorine Purification 72
11.3. Drying 72

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_399.pub2

11.4. Transfer and Compression 73
11.5. Liquefaction 74
11.6. Chlorine Recovery 75
12. Chlorine Handling 76
12.1. Storage Systems 76
12.2. Transport 77
12.3. Chlorine Discharge Systems 78
12.4. Chlorine Vaporization 79
12.5. Treatment of Gaseous Effluents . 80
12.6. Materials 80
12.7. Safety 81
13. Quality Specifications and Analytical Methods 82
13.1. Quality Specifications 82
13.2. Analytical Methods 82
14. Uses 83
15. Economic Aspects 86
16. Toxicology 87
17. Acknowledgement 89
18. References 89

1. Introduction

Although C. W. Scheele reported the formation of chlorine gas from the reaction of manganese dioxide with hydrochloric acid in 1774, he did not recognize the gas as an element [37]. H. Davy is usually accepted as the discoverer (1808), and he named the gas *chlorine* from the Greek ϰλῶροσ (chloros), meaning greenish yellow. Chlorine for bleaching textiles was first produced from manganese dioxide and hydrochloric acid by a process developed by Weldon, the yield of chlorine being 35 % of the theoretical value. In 1866, Deacon developed a process based on the oxidation of hydrogen chloride gas by atmospheric oxygen in the presence of a copper salt, $CuCl_2$, as the catalyst and obtained yields up to 65 % of the theoretical value.

In 1800, Cruickshank was the first to prepare chlorine electrochemically [38]; however, the process was of little significance until the development of a suitable generator by Siemens and of synthetic graphite for anodes by Acheson and Castner in 1892. These two developments made possible the electrolytic production of chlorine, the *chlor-alkali process,* on an industrial scale. About the same time, both the *diaphragm cell* process (1885) and the *mercury cell* process (1892) were introduced. The *membrane cell* process was developed much more recently (1970). Currently, more than 95 % of world chlorine production is obtained by the chlor-alkali process. Since 1970 graphite anodes have been superseded by *activated titanium anodes* in the diaphragm and mercury cell processes. The newer membrane cell process uses only activated titanium anodes.

Other electrochemical processes in which chlorine is produced include the electrolysis of hydrochloric acid and the electrolysis of molten alkali metal and alkaline earth metal chlorides, in which the chlorine is a byproduct. Purely chemical methods of chlorine production are currently insignificant.

Since 1975, the membrane cell process has been developed to a high degree of sophistication. It has ecological advantages over the mercury processes and has become the most economically advantageous process. The membrane cell process has become widely accepted, and all new plants are using this technology. By 2000 more than 30 % of the chlorine will be produced in membrane cells.

World capacity for chlorine exceeds 45×10^6 t/a. With an annual energy consumption of about 1.5×10^{11} kWh, the chlor-alkali process is one of the largest industrial consumers of electrical energy. The chlorine production of a country is an indicator of the state of development of its chemical industry.

Occurrence and Formation. Chlorine is the 11th most abundant element in the lithosphere. Because it is highly reactive, it is rarely found in the free state and then mainly in volcanic gases. It exists mainly in the form of chlorides, as in sea water, which contains an average of 2.9 wt % sodium chloride and 0.3 wt % magnesium chloride. In salt deposits formed by evaporation of seas, there are large quantities of rock salt (NaCl) and sylvite (KCl), together with bischofite ($MgCl_2 \cdot 6\,H_2O$), carnallite ($KCl \cdot MgCl_2 \cdot 6\,H_2O$), tachhydrite ($CaCl_2 \cdot 2\,MgCl_2 \cdot 12\,H_2O$), kainite ($KCl \cdot MgSO_4 \cdot 3\,H_2O$), and others. Occasionally there are also heavy metal chlorides, usually in the form of double salts, such as atacamite ($CuCl_2 \cdot 3\,Cu(OH)_2$), and compounds of lead, iron, manganese, mercury, or silver. Chlorates and perchlorates occur to a small extent in Chile saltpeter. Free hydrochloric

acid is occasionally found in gases and springs of volcanic origin. Plants and animals always contain chlorine in the form of chlorides or free hydrochloric acid.

Chlorine is formed by oxidation of hydrochloric acid or chlorides by such compounds as manganese dioxide, permanganates, dichromates, chlorates, bleaching powder, nitric acid, or nitrogen oxides. Oxygen, including atmospheric oxygen, acts as an oxidizing agent in the presence of catalysts. Some metal chlorides produce chlorine when heated, for example, gold(III) chloride or platinum chloride.

2. Physical Properties

Chlorine [*7782-50-5*], EINECS no. 231–959–5, exists in all three physical states. At STP it is a greenish-yellow pungent, poisonous gas, which liquefies to a mobile yellow liquid. Solid chlorine forms pale yellow rhombic crystals. The principal properties are given below; more details, including thermodynamic values are given in [40] and in "New Property Tables of Chlorine in SI Units" [41].

Atomic number Z	17
Relative atomic mass A_r	35.453
Stable isotopes (abundance)	35 (75.53 %) 37 (24.47 %)
Electronic configuration in the ground state	[Ne] $3s^2 3p^5$
Term symbol in the ground state	$^2P_{3/2}$
Melting point *mp*	172.17 K (− 100.98 °C)
Boiling point *bp*	239.10 K (− 34.05 °C)
Critical density ϱ_{crit}	565.00 kg/m^3
Critical temperature T_{crit} (t_{crit})	417.15 K (144.0 °C)
Critical pressure p_{crit}	7.71083 MPa
Density of gas ϱ (0 °C, 101.3 kPa)	3.213 kg/m^3
Density relative to air d	2.48
Enthalpy of fusion ΔH_f	90.33 kJ/kg
Enthalpy of vaporization ΔH_v	287.1 kJ/kg
Standard electrode potential E°	1.359 V
Enthalpy of dissociation ΔH_{diss}	239.44 kJ/mol (2.481 eV)
Electron affinity A	364.25 kJ/mol (3.77 eV)
Enthalpy of hydration ΔH_{hyd} of Cl^-	405.7 kJ/mol
Ionization energies ΔE_i	13.01, 23.80, 39.9, 53.3, 67.8, 96.6, 114.2 eV
EC No.	017–001–00–7

The density of chlorine gas at 101.3 kPa is a function of temperature:

t, °C	0	50	100	150
ϱ, kg/m^3	3.213	2.700	2.330	2.051

The density up to 300 °C is higher than that of an ideal gas because of the existence of more complex molecules, for example, Cl_4. In the range 400 – 1450 °C, the density approximates that of an ideal gas, and above 1450 °C thermal dissociation takes place, reaching 50 % at 2250 °C. The density of chlorine gas as a function of temperature and pressure is shown in Figure 1. The gas state can be described by the van der Waals equation with

$$a = 6.580\ \mathrm{L^2 bar\ mol^{-2}},\ b = 0.05622\ \mathrm{L/mol}$$

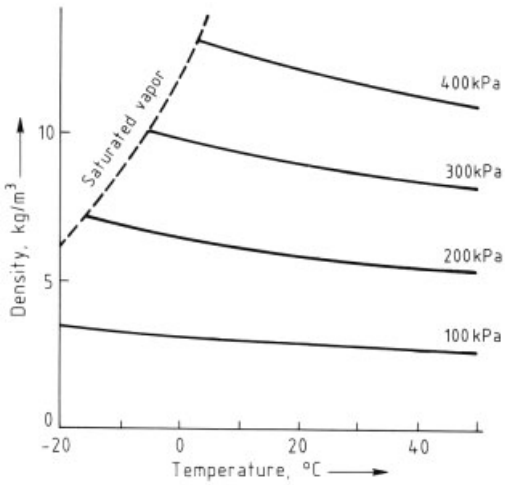

Figure 1. Density of chlorine gas as a function of temperature and pressure

The density of liquid chlorine is given in Figure 2. The compressibility of liquid chlorine is the greatest of all the elements. The volume coefficient per MPa at 20 °C over the range 0 – 10 MPa is 0.012 %. The coefficient increases rapidly with temperature: 0.023 % at 35 °C, 0.037 % at 64 °C, and 0.064 % at 91 °C. One liter of liquid chlorine at 0 °C produces 456.8 dm^3 of chlorine gas at STP; 1 kg of liquid produces 311 dm^3 of gas.

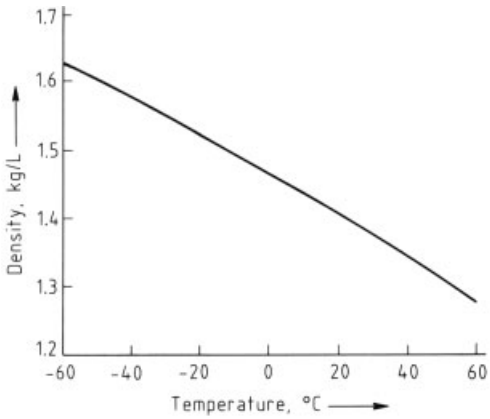

Figure 2. Density of liquid chlorine

The vapor-pressure curve for chlorine is shown in Figure 3.

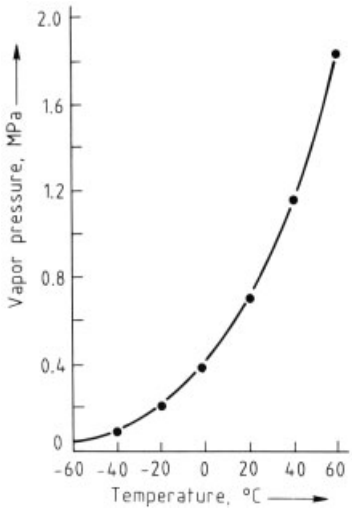

Figure 3. Vapor pressure of liquid chlorine

The vapor pressure can be calculated over the temperature range 172 – 417 K from the Martin – Shin – Kapoor equation [41]:

$$\ln P = A + \frac{B}{T} + C \ln T + DT + \frac{E\,(F-T)\ln(F-T)}{FT}$$

$A = 62.402508$
$B = -4343.5240$
$C = -7.8661534$
$D = 1.0666308 \times 10^{-2}$
$E = 95.248723$
$F = 424.90$

Thermodynamic information is given in Table 1, from which the data required for working with gaseous and liquid chlorine can be obtained [42]. The Joule – Thomson coefficient is 0.0308 K/kPa at STP.

At STP the specific heats of chlorine are

$c_p = 0.481\ \mathrm{kJ\,kg^{-1}\,K^{-1}}$
$c_v = 0.357\ \mathrm{kJ\,kg^{-1}\,K^{-1}}$
$\varkappa = c_p/c_v = 1.347$

The molar heat capacity at constant volume c_v increases with temperature [43]:

t, °C	0	100	200	500	1000
c_v, J/mol	24.9	26.4	28.1	28.9	29.7

The heat capacity of liquid chlorine decreases over the temperature range – 90 °C to 0 °C:

t, °C	– 90	– 70	– 50	– 30	0
c, $\mathrm{J\,kg^{-1}\,K^{-1}}$	0.9454	0.9404	0.9341	0.9270	0.9169
c, $\mathrm{J\,mol^{-1}\,K^{-1}}$	67.03	66.70	66.23	65.73	65.02

The thermal conductivities of chlorine gas and liquid are almost linear functions of temperature from – 50 °C to 150 °C:

t, °C	– 50	– 25	0	25	50	75	100
λ_g, $\mathrm{W\,m^{-1}\,K^{-1}} \times 10^2$	6.08	7.06	7.95	8.82	9.75	10.63	11.50
λ_l, $\mathrm{W\,m^{-1}\,K^{-1}}$	0.17	0.16	0.15	0.135	0.12	0.11	0.09

The viscosities of chlorine gas and liquid are shown in Figure 4 over the same temperature range. The surface tension at the liquid – gas interface falls rapidly with temperature:

t, °C	– 50	– 25	0	25	50
σ, $\mathrm{mJ/m^2}$	29.4	25.2	20.9	16.9	13.4

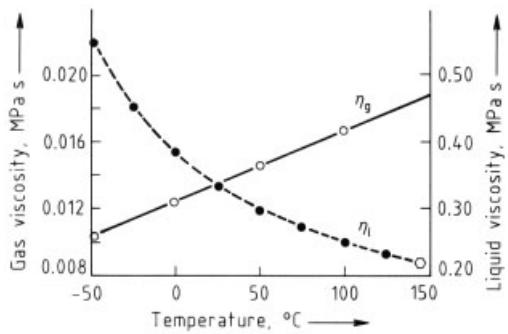

Figure 4. Viscosities of chlorine gas and liquid

Table 1. Properties of liquid and gaseous chlorine [41]. Lower values are quoted in more recent literature [38, 39], especially in the region of the critical points.

Temperature t, °C	Pressure, p, bar	Specific volumes, dm^3/kg		Specific enthalpies, kJ/kg*			Specific entropies, $kJ\,kg^{-1}\,K^{-1}$		
		liquid	vapor	liquid	vaporization	vapor	liquid	vaporization	vapor
−70	0.1513	0.6042	1563	351.11	306.89	658.00	3.9021	1.5106	5.4127
−60	0.2768	0.6135	894.4	360.69	301.58	662.27	3.9481	1.4147	5.3629
−50	0.4762	0.6233	541.8	370.15	296.29	666.41	3.9917	1.3276	5.3193
−40	0.7772	0.6336	344.9	379.70	290.73	670.43	4.0336	1.2468	5.2804
−30	1.212	0.6445	229.0	389.37	284.95	674.33	4.0737	1.1719	5.2456
−20	1.816	0.6560	157.7	399.21	278.84	678.05	4.1131	1.1015	5.2147
−10	2.628	0.6682	112.1	408.88	272.73	681.61	4.1508	1.0362	5.1870
0	3.689	0.6812	81.89	418.68 **	266.28	684.96	4.1868 **	0.9747	5.1615
10	5.043	0.6951	61.26	428.43	259.67	688.10	4.2215	0.9169	5.1385
20	6.731	0.7100	46.77	438.19	252.80	690.99	4.2546	0.8625	5.1171
30	8.800	0.7261	36.35	447.90	245.72	693.63	4.2873	0.8106	5.0978
40	11.30	0.7435	28.66	457.66	238.31	695.97	4.3183	0.7612	5.0790
50	14.27	0.7627	22.88	467.45	230.53	697.98	4.3480	0.7134	5.0614
60	17.76	0.7837	18.44	477.50	222.07	699.57	4.3781	0.6665	5.0447
70	21.84	0.8073	14.97	487.76	212.90	700.66	4.4074	0.6205	5.0279
80	26.55	0.8339	12.20	498.56	202.60	701.16	4.4376	0.5736	5.0112
90	31.95	0.8646	9.944	510.25	190.79	701.04	4.4665	0.5254	4.9919
100	38.14	0,9010	8.082	523.35	176.85	700.20	4.5004	0.4739	4.9743
110	45.18	0.9456	6.508	537.88	160.14	698.02	4.5372	0.4178	4.9551
120	53.18	1.0039	5.169	554.62	139.59	694.21	4.5787	0.3550	4.9337
130	62.24	1.0890	4.001	575.10	113.30	688.39	4.6277	0.2809	4.9086
140	72.50	1.2624	2.842	603.74	71.18	674.91	4.6934	0.1725	4.8659
144	77.01	1.7631	1.763	642.30	0	642.30	4.7825	0	4.7825

* These values have been calculated in S.I. units according to DIN 1345.
** The enthalpy of liquid chlorine at 0 °C was taken to be $H_0 = 418.66$ kJ/kg; the entropy of liquid chlorine at 0 °C was taken to be $\varrho_0 = 4.1868\,kJ\,kg^{-1}\,K^{-1}$.

The specific magnetic susceptibility at 20 °C is $-7.4 \times 10^{-9}\,m^3/kg$.

Liquid chlorine has a very low electrical conductivity, the value at −70 °C being $10^{-16}\,\Omega^{-1}\,cm^{-1}$. The dielectric constant of the liquid for wavelengths greater than 10 m is 2.15 at −60 °C, 2.03 at −20 °C, 1.97 at 0 °C, and 1.54 at 142 °C, near the critical temperature.

Chlorine gas can be absorbed in considerable quantities onto activated charcoal and silica gel, and this property can be used to concentrate chlorine from gas mixtures containing it.

Chlorine is soluble in cold water, usually less so in aqueous solutions. In salt solutions, the solubility decreases with salt concentration and temperature. In hydrochloric acid, chlorine is more soluble than in water, and the solubility increases with acid concentration (Fig. 5 and Fig. 6). In aqueous solutions, chlorine is partially hydrolyzed, and the solubility depends on the pH of the solution. Below 10 °C chlorine forms hydrates, which can be separated as greenish-yellow crystals. Chlorine hydrate is a clathrate, and there is no definite chlorine : water ratio. The chlorine – water system has a quadruple point at 28.7 °C; the phase diagram has been worked out by Ketelaar [44].

Chlorine is readily soluble in sulfur – chlorine compounds, which can be used as industrial solvents for chlorine. Disulfur dichloride [*10025-67-9*], S_2Cl_2, is converted to sulfur dichloride (SCl_2) and sulfur tetrachloride (SCl_4). Some metallic chlorides and oxide chlorides, such as vanadium oxide chloride, chromyl chloride, titanium tetrachloride, and tin(IV)chloride, are good solvents for chlorine. Many other chlorine-containing compounds dissolve chlorine readily. Examples are phosphoryl chloride, carbon tetrachloride (Fig. 7), tetrachloroethane, pentachloroethane, hexachlorobutadiene (Fig. 7), and chlorobenzene. Chlorine also dissolves in glacial acetic acid, dimethylformamide, and nitrobenzene. The solubility of chlorine in a number of these solvents is given in Table 2.

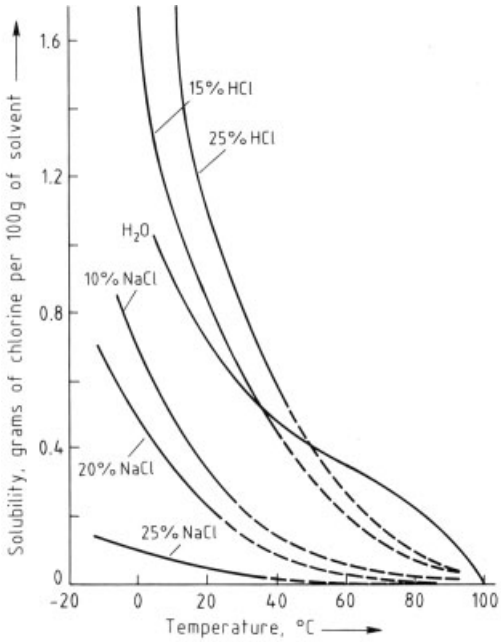

Figure 5. Solubility of chlorine in water, hydrochloric acid (two concentrations), and sodium chloride solutions (three concentrations)
All percentages are weight percents.

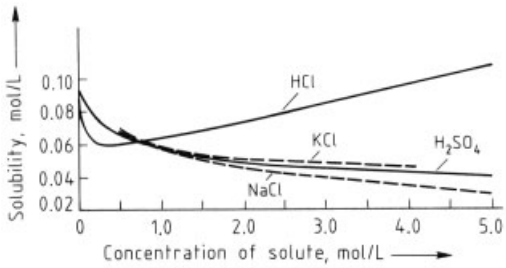

Figure 6. Solubility of chlorine in solutions of KCl, NaCl, H_2SO_4, and HCl at 25 °C

Table 2. Solubility of chlorine in various solvents

Solvent	Temperature, °C	Solubility, wt %
Sulfuryl chloride	0	12.0
Disulfur dichloride	0	58.5
Phosphoryl chloride	0	19.0
Silicon tetrachloride	0	15.6
Titanium tetrachloride	0	11.5
Benzene	10	24.7
Chloroform	10	20.0
Dimethylformamide	0	123 *
Acetic acid, 99.84 wt %	15	11.6 *

* g/100 cm^3

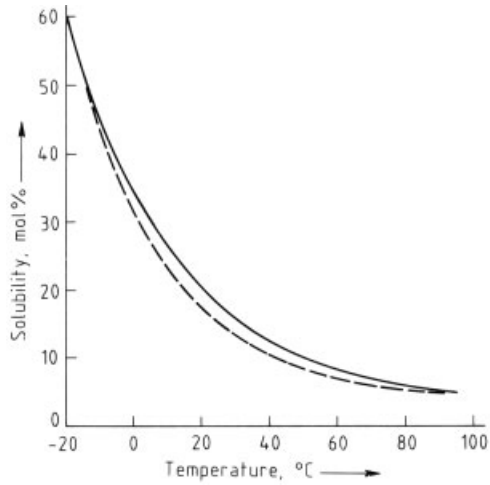

Figure 7. Solubility of chlorine in hexachlorobutadiene (——) and carbon tetrachloride (– – –) at 101 kPa as a function of temperature

3. Chemical Properties

Inorganic Compounds. Chlorine, fluorine, bromine, and iodine constitute the halogen group, which has marked nonmetallic properties. The valence of chlorine is determined by the seven electrons in the outer shell. By gaining one electron, the negatively charged chloride ion is formed; the chloride ion has a single negative charge and a complete shell of electrons (the argon structure). By sharing one to seven electrons from the outer shell with other elements, the various chlorine oxidation states can be formed, for example, in the oxides of chlorine, hypochlorites (+ 1), chlorates (+ 5), and perchlorates (+ 7).

The bonds between chlorine and the other halogens are mainly covalent. In the chlorine – fluorine compounds ClF and ClF_3, there is some ionic character to the bond, with chlorine the anion, and in the chlorine – iodine compounds ICl_3 and ICl, there is some ionic character to the bond, with chlorine the cation. Chlorine is very reactive, combining directly with most elements but only indirectly with nitrogen, oxygen, and carbon. Excess chlorine in the presence of ammonia salts forms the very explosive nitrogen trichloride, NCl_3. Hypochlorites react with ammonia to produce the chloramines NH_2Cl and $NHCl_2$.

Oxygen and chlorine form several chlorine oxides (→ Chlorine Oxides and Chlorine Oxygen Acids).

Chlorine gas does not react with hydrogen gas [*1333-74-0*] at normal temperatures in the absence of light. In sunlight or artificial light of wavelength ca. 470 nm or at temperatures over 250 °C, the two gases combine explosively to form hydrogen chloride. The explosive limits of mixtures of pure gases lie between ca. 8 vol % H_2 and ca. 14 vol % Cl_2 (the detonation limits). The limits depend on pressure, and the detonation range can be reduced by adding inert gases, such as nitrogen or carbon dioxide (Fig. 8) [45, 46].

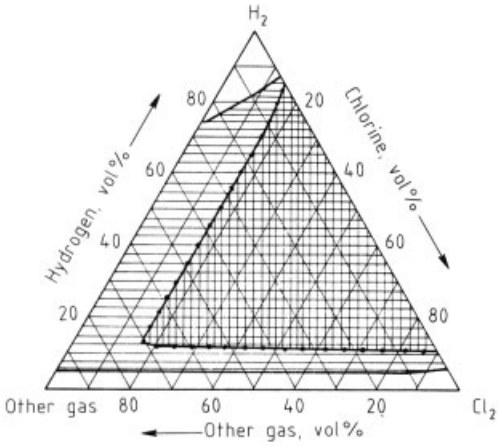

Figure 8. Explosive limits of chlorine – hydrogen – other gas mixture
Horizontally hatched area = Explosive region with residue gas from chlorine liquefaction (O_2, N_2, CO_2)
Checkered area = Explosive region with inert gas (N_2, CO_2)

Chlorine reacts vigorously with ammonia

$$3\,Cl_2 + 4\,NH_3 \rightarrow NCl_3 + 3\,NH_4Cl$$

In the presence of the catalyst bromine, chlorine reacts with nitric oxide to give nitrosyl chloride

$$NO + 0.5\,Cl_2 \rightarrow NOCl$$

Sulfur dioxide and chlorine in the presence of light or an activated carbon catalyst react to form sulfuryl chloride, SO_2Cl_2. Under these conditions carbon monoxide and chlorine react to produce the colorless, highly toxic carbonyl chloride (phosgene), $COCl_2$.

Chlorine reacts with sodium cyanide and sodium thiocyanate to produce cyanogen chloride and thiocyanogen chloride. The reaction of chlorine with sodium thiosulfate [*7772-98-7*] (Antichlor) is used to remove free chlorine from solutions.

$$Na_2S_2O_3 + 4\,Cl_2 + 5\,H_2O \rightarrow 2\,NaHSO_4 + 8\,HCl$$

Chlorine reacts with carbon disulfide to produce carbon tetrachloride and disulfur dichloride.

$$CS_2 + 3\,Cl_2 \rightarrow CCl_4 + S_2Cl_2$$

The reaction of chlorine with phosphorus produces phosphorus trichloride (PCl_3) and pentachloride (PCl_5). Wet chlorine attacks most metals to form chlorides. Although titanium [*7440-32-6*] is resistant to wet chlorine, it is rapidly attacked by dry chlorine. Tantalum is resistant to both wet and dry chlorine. Most metals are resistant to dry chlorine below 100 °C, but above a specific temperature for each metal, combustion takes place with a flame. This specific temperature, the ignition temperature, also depends on the particle size of the metal so that the following values are only approximate: iron at 140 °C, nickel at 500 °C, copper at 200 °C, and titanium at 20 °C.

Most metal chlorides are soluble in water [3, p. 668], notable exceptions being those of silver (AgCl) and mercury (Hg_2Cl_2). Chlorine liberates bromine and iodine from metallic bromides and iodides, but is itself liberated from metal chlorides by fluorine.

$$0.5\,Cl_2 + KBr \rightarrow KCl + 0.5\,Br_2$$

Selenium and tellurium react spontaneously with liquid chlorine, whereas sulfur begins to react only at the boiling point. Liquid chlorine reacts vigorously with iodine, red phosphorus, arsenic, antimony, tin, and bismuth. Potassium,

sodium, and magnesium are unaffected in liquid chlorine at temperatures below $-80\,^{\circ}C$. Aluminum is unattacked until the temperature rises to $-20\,^{\circ}C$, when it ignites. Gold is only slowly attacked by liquid chlorine to form the trichloride ($AuCl_3$). Cast iron, wrought iron, carbon steel, phosphor bronze, brass, copper, zinc, and lead are unaffected by dry liquid chlorine, even in the presence of concentrated sulfuric acid.

Organic Compounds. The chlorine–carbon bond is covalent in nature. Chlorine reacts with hydrocarbons either by substitution or by addition. In saturated hydrocarbons, chlorine replaces hydrogen, either completely or partially, to form chlorinated hydrocarbons and hydrogen chloride. Methane can be chlorinated in stages through to carbon tetrachloride.

In the reaction with unsaturated hydrocarbons, chlorine is added to the double or triple bond yielding dichloro- or tetrachlorohydrocarbons, respectively. In industry the reaction velocity is increased by light (photochlorination), heat (cracker furnace), or catalysts. In aromatic hydrocarbons, both addition and substitution are possible, depending on conditions (light, temperature, pressure, or catalysts).

Chlorine Compounds in Nature. More than 1800 chlorine-containing compounds have been identified in nature at present. These organic chlorine compounds, which range from chloromethane to the complex antibiotic vancomycin, are produced by marine and terrestrial plants, bacteria, fungi, lichen, insects, marine animals, some higher animals, and a few mammals. For example, the annual natural production of chloromethane is estimated to exceed 4×10^6 t/a [47].

4. Chlor-Alkali Process

In the chlor-alkali electrolysis process, an aqueous solution of sodium chloride is decomposed electrolytically by direct current, producing chlorine, hydrogen, and sodium hydroxide solution. The overall reaction of the process

$$2\,NaCl + 2\,H_2O \rightarrow Cl_2 + H_2 + 2\,NaOH$$

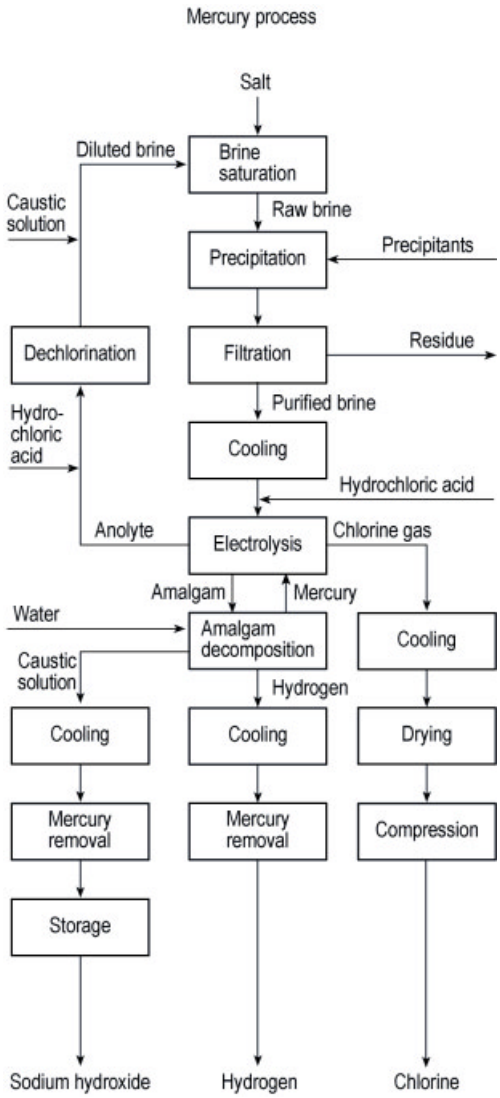

Figure 9. Flow diagram of the chlor-alkali mercury process

takes place in two parts, at the anode and at the cathode. The evolution of chlorine takes place at the anode:

$$2\,Cl^- \rightarrow 2\,Cl + 2\,e^- \rightarrow Cl_2 + 2\,e^-$$

There are three basic processes for the electrolytic production of chlorine, the nature of the cathode reaction depending on the specific

process. These three processes are (1) the diaphragm cell process (Griesheim cell, 1885), (2) the mercury cell process (Castner – Kellner cell, 1892), and (3) the membrane cell process (1970).

Each process represents a different method of keeping the chlorine produced at the anode separate from the caustic soda and hydrogen produced, directly or indirectly, at the cathode.

These three processes are described in detail in the following three chapters. The basic flow sheets of the three processes are shown in Figures 9, [10]], [11]]. In all three processes, nearly saturated, purified brine is introduced into the electrolysis cell.

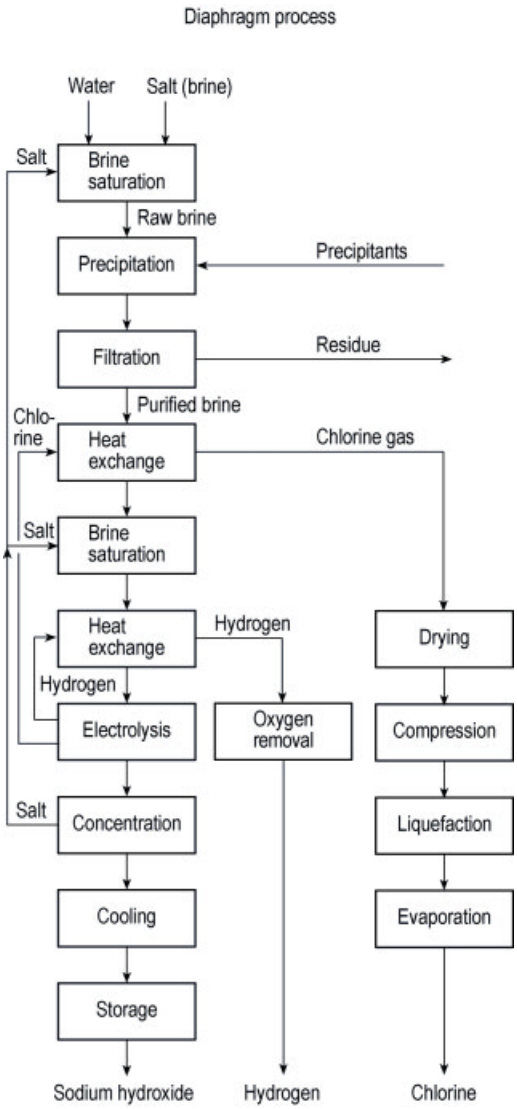

Figure 10. Flow diagram of the chlor-alkali diaphragm process

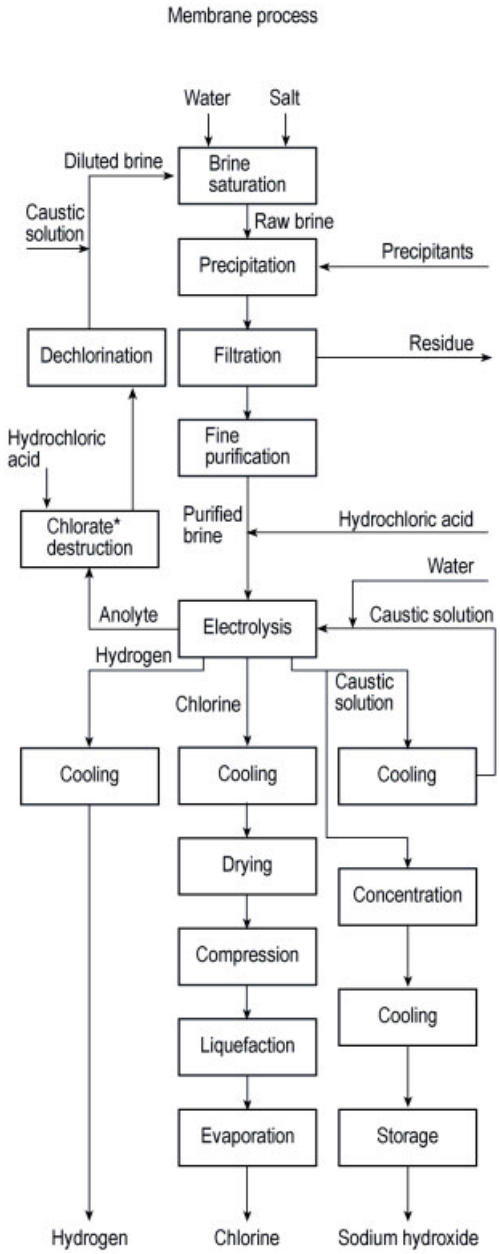

Figure 11. Flow diagram of the chlor-alkali membrane process
(* optional)

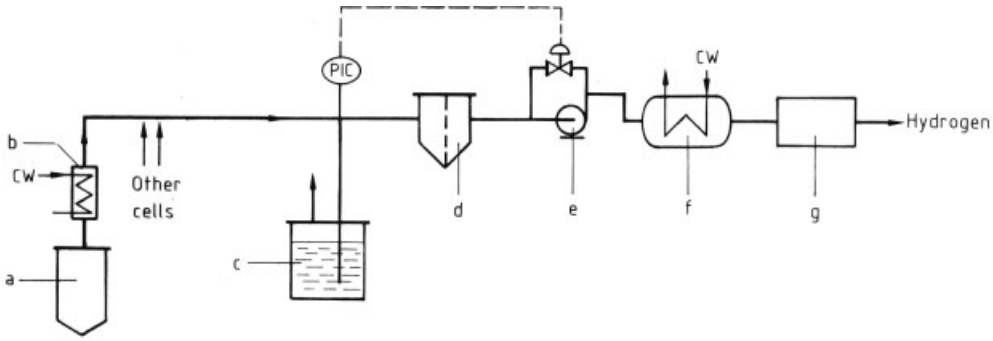

Figure 12. Processing of hydrogen gas from the amalgam decomposer
a) Vertical decomposer; b) Individual cell hydrogen cooler; c) Safety seal; d) Demister; e) Blower; f) Final hydrogen cooler; g) Mercury removal equipment

The *hydrogen* produced is cooled as it leaves the decomposer or the cathode compartment and is carried through electrically insulated pipework to a vessel fitted with a water seal (Fig. 12). If a hydrogen – air mixture forms because of a shutdown or breakdown, the seal allows the mixture to escape. A demister ensures that the gas is free of spray, whether water or sodium hydroxide solution. The hydrogen is compressed by Roots-type blowers or reciprocating compressors before it passes through coolers on its way to the consuming plants. At no stage is the pressure allowed to fall below ambient pressure.

Electrolytic hydrogen is very pure, >99.9 %; however, unwanted traces of oxygen can be removed by reaction with the hydrogen over a platinum catalyst. The hydrogen is used for organic hydrogenation, catalytic reductions, ammonia synthesis and to provide hot flames or protective atmospheres in welding technology, metallurgy, or glass manufacture. It is also used in the manufacture of high-purity hydrogen chloride by combustion with chlorine and as a fuel for heating and drying.

In the *mercury cell process,* sodium amalgam is produced at the cathode. The amalgam is reacted with water in a separate reactor, called the decomposer, to produce hydrogen gas and caustic soda solution.

Because the brine is recirculated, solid salt is required for resaturation. The brine, which must be quite pure, is first dechlorinated and then purified by a straightforward precipitation – filtration process.

The products are extremely pure. The chlorine, along with a little oxygen, generally can be used without further purification. The sodium hydroxide solution contains little chloride and leaves the decomposer with a 50 wt % concentration.

Of the three processes, the mercury process uses the most electric energy; however, no steam is required to concentrate the caustic solution. The use of large quantities of mercury demands measures to prevent environmental contamination. In addition, the hydrogen gas and sodium solution must be freed from mercury. Generally, the operation of the cells is not simple.

In the *diaphragm cell process,* the anode area is separated from the cathode area by a permeable, generally asbestos-based diaphragm. The brine is introduced into the anode compartment and flows through the diaphragm into the cathode compartment. Cheaper solution-mined brine can be used; the brine is purified by precipitation – filtration.

A caustic brine leaves the cell, and this brine must be freed from salt in an elaborate evaporative process. Even so, the resultant 50 wt % sodium hydroxide solution contains up to 1 wt % NaCl. The salt separated from the caustic brine can be used to saturate dilute brine. The chlorine contains oxygen and must be purified by liquefaction and evaporation.

The consumption of electric energy with the diaphragm cell process is ca. 15 % lower than for the mercury process, but the total energy consumption is higher because of the steam required to concentrate the caustic brine (see Fig. 62). Environmental contamination with asbestos must

be avoided. Under constant operating conditions, cell operation is relatively simple.

In the *membrane cell process,* the anode and cathode are separated by a cation-permeable ion-exchange membrane. Only sodium ions and a little water pass through the membrane.

As in the mercury process, the brine is dechlorinated and recirculated, which requires solid salt to resaturate the brine. The life of the expensive membrane depends on the purity of the brine. Therefore, after purification by precipitation – filtration, the brine is also purified with an ion exchanger.

The caustic solution leaves the cell with a concentration of 30 – 36 wt % and must be concentrated. The chloride content of the sodium hydroxide solution is almost as low as that from the mercury process. The chlorine gas contains some oxygen and must be purified by liquefaction and evaporation.

The consumption of electric energy with the membrane cell process is the lowest of the three processes, ca. 25 % less than for the mercury process, and the amount of steam needed for concentration of the caustic is relatively small (see Fig. 62). The energy consumption should be even lower when oxygen-consuming electrodes become common. There are no special environmental problems. The cells are easy to operate and are relatively insensitive to current changes, allowing greater use of the cheaper off-peak-time electric power.

4.1. Brine Supply

The brine used in the mercury cell and membrane cell processes is normally saturated with solid salt although there are some installations that use solution-mined brine on a once-through basis. The brine supply for diaphragm cells is always used on a once-through basis, although the salt recovered from caustic soda evaporators may be recycled into the brine supply.

Salt. The basic raw material for the mercury cell and membrane cell processes is usually solid salt. This may be obtained from three sources: rock salt, solar salt, or vacuum-evaporated salt from purifying and evaporating solution-mined brine.

In the United States and Europe, rock salt is most commonly used. The most important impurities are shown in Table 3. The concentrations of these impurities depend on the method of production and on the different grades: crude rock salt, prepared rock salt, and evaporated salt. Solar salt is used in Japan and many other parts of the world, the most important sources being Australia, Mexico, China, Chile, India, and Pakistan. The salt produced by solar evaporation is usually much less pure than rock salt. In a few cases the salt may be obtained from other processes, such as caustic soda evaporation in the diaphragm process.

A new upgrading process (Salex) has been developed by Krebs Swiss [48]. It removes the impurities by selective cracking of the salt crystals and a washing process. Salt losses are minimized, and the purity exceeds 99.95 % NaCl.

Table 3. Impurities in rock salt and sea salt, wt %

	Rock salt	Sea salt
Insolubles	≤ 2	0.1 – 0.3
Water	≤ 3	2.0 – 6.0
Calcium	0.2 – 0.3	0.1 – 0.3
Magnesium	0.03 – 0.1	0.08 – 0.3
Sulfate	≤ 0.8	0.3 – 1.2
Potassium	≤ 0.04	0.02 – 0.12

Brine Resaturation. In older plants, the open vessels or pits used for storing the salt are also used as resaturators. The depleted brine from the cells is sprayed onto the salt and is saturated, the NaCl concentration reaching 310 – 315 g/L. Modern resaturators are closed vessels, to reduce environmental pollution [49], which could otherwise occur by the emission of a salt spray or mist. The weak brine is fed in at the base of the resaturator, and the saturated brine is drawn off at the top. If the flow rates of the brine and the continously added salt are chosen carefully, the differing dissolution rates of NaCl and $CaSO_4$ result in little calcium sulfate dissolving within the saturator [50]. Organic additives also reduce the dissolution rate of calcium sulfate [51]. The solubility (g per 100 g of H_2O) of NaCl in water does not increase much with temperature (t, °C), whereas the solubility of KCl does:

t	0	20	40	60	80	100
c_{NaCl}	35.6	35.8	36.4	37.0	38.5	39.2
c_{KCl}	28.2	34.4	40.3	45.6	51.0	56.2

Brine Purification. In mercury cells, traces of heavy metals in the brine give rise to dangerous operating conditions (see page 15), as does the presence of magnesium and to a lesser extent calcium [52]. In membrane cells, divalent ions such as Ca^{2+} or Mg^{2+} are harmful to the membrane. The circulating brine must be rigorously purified to avoid any buildup of these substances to undesirable levels [7]. Calcium is usually precipitated as the carbonate with sodium carbonate; magnesium and iron, as hydroxides with sodium hydroxide; and sulfate, as barium sulfate.

The reagents are usually mixed with weak brine and added to the brine stream at a controlled rate. If solar salt is used, treatment costs may be reduced by prewashing the salt [53]. In order to precipitate calcium at low pH, sodium bicarbonate [54] or phosphoric acid [55] can be added.

The sulfate content can be reduced without the use of expensive barium salts by discharging a part of (purging) the brine [56], by crystallization of $Na_2SO_4 \cdot 10\,H_2O$ on cooling the brine stream [57], by precipitation of the double salt $Na_2SO_4 \cdot CaSO_4$ [58], by an ion-exchange process, or by membrane nanofiltration [59]. Hoechst [60] has a process for recovering barium sulfate of pure pigment quality by precipitation under acid conditions. Chlorate buildup can be avoided by addition of sodium metabisulfite $Na_2S_2O_5$ [61].

After stirring for 1 – 2 h, the precipitated impurities are removed by filtration alone or by sedimentation followed by filtration. Sedimentation is carried out in large circular settling tanks, from which the slurry is removed by mechanical raking equipment, e.g., Clariflocculator, Cyclator, or Dorr thickener. Filtration is carried out with a sand filter, a pressure-leaf filter with filter cloths of chlorine-resistant fabrics, or candle filters automatically cleaned by backflow of brine. The filter is cleaned by water jets, vibrating, or shaking. The separated filter cake is concentrated to 60 – 80 % solids content in rotary drum vacuum filters or centrifuges before disposal. Any soluble material present may be removed from the sludge by washing with water. Barium salts may be recovered by treating with sodium carbonate under pressure [62]. The purified brine should contain ideally $c_{Ca} < 2$ mg/L, $c_{Mg} < 1$ mg/L, and $c_{SO_4} < 5\ g/L$.

In the diaphragm process, the removal of sulfate is not always necessary because SO_4^{2-} can be removed from the cell liquor as pure Na_2SO_4 during the concentration process. In the membrane process, the brine must be purified to a much higher degree to avoid the deterioration of the membrane. The Ca^{2+} and Mg^{2+} concentration must be < 0.02 ppm (20 ppb), so a second, fine purification step is required (see Section 7.2.1).

Before the brine enters the electrolysis cells, it should be acidified with hydrochloric acid to pH < 6, which increases the life of the titanium anode coating, gives a purer chlorine product with higher yield, and reduces the formation of hypochlorite and chlorate in the brine.

Brine Dechlorination. In the mercury and membrane processes, the depleted brine leaving the cells must be dechlorinated before resaturation. Further acidification with hydrochloric acid to pH 2 – 2.5 reduces the solubility of chlorine by shifting the equilibrium point of hydrolysis and inhibits the formation of hypochlorite and chlorate. Chlorine discharged in the anolyte tank prior to dechlorination may be fed into the chlorine system. The dissolved chlorine of the brine then is still 400 – 1000 mg/L, depending on pH and temperature. The brine is passed down a packed column or sprayed into a vacuum of 50 – 60 kPa, which reduces the chlorine concentration in the brine to 10 – 30 mg/L. The vacuum is produced by steam jet or liquid-ring vacuum pump. The pure chlorine gas obtained is fed into the chlorine stream.

The water that evaporates from the dechlorinated brine is condensed in a cooler. The condensate, which may be chemically dechlorinated, is returned to the brine circulation system if necessary to maintain the volume of the brine circuit. If necessary, the remaining chlorine content can be further reduced by blowing with compressed air, by a second vacuum treatment, by treatment with activated carbon [63], or by chemical treatment with hydrogen sulfite, thiosulfate, sulfur dioxide, or sodium hydrogensulfide.

Brine Monitoring. The sodium chloride concentration in the brine is determined by density measured by equipment involving radioactive isotopes, vibration techniques, hydrometry, or weighing.

The pH following alkali or acid additions is determined with glass electrodes, and the redox potential following chlorine removal is determined with metal electrodes. Excess OH^- and CO_3^{2-} ions ensure adequate precipitation of dissolved calcium, iron, and magnesium. After filtration, a test sample of 100 mL should require 4 – 6 mL of 0.1 N acid to reach the phenolphthalein end point and a further 0.5 – 1.5 mL to reach the methyl orange end point. Inadequate filtration is detected by turbidimetry in transmitted light or by the Tyndall effect. Calcium and magnesium are determined hourly, and chlorate and sulfate about once per day, all by titration.

4.2. Electricity Supply

Since 1960 the direct current for electrolysis has been provided exclusively by silicon rectifiers. A set of rectifiers can supply up to 450 000 A. Voltages up to 4.0 kV per diode are feasible, but usually for safety, a peak a.c. voltage of 1500 V, corresponding to a d.c. output of 1200 V, is not exceeded. Liquid cooling of the diodes permits a compact design, and self-contained equipment reduces leakage losses. Modern membrane cell plants also use continuously variable thyristor converters in place of silicon diodes [64].

Rectification equipment is required to provide steady direct current at a voltage determined by the cell room. The current must remain steady even though the voltage is varied both by the operating condition of the cells and by the number of cells operating. The rectifier equipment usually consists of

transformer capable of variable output voltage with adequate compensation for changing input voltage
silicon rectifiers or thyristors
constant-current control gear
transducers for metering and control
control panels
isolators
cooling equipment
ancillary safety and monitoring equipment

Each set of rectifiers is connected through high-voltage switchgear to the three-phase supply [65]. Smaller units use a 10 – 30 kV supply, but large units can be connected into the high-voltage power system (> 100 kV) [66].

The unit cost of the d.c. supply decreases with increasing voltage and current. A plant is therefore most economical when as many high-current cells as possible are connected in series [67]. Total currents of 450 000 A are achieved. The switches for short-circuiting the cells are designed for 10 000 – 30 000 A and are operated by compressed air, hydraulically, or by spring action. Erosion of the main contacts is dealt with by using replaceable pre-contacts [68]. The contacts are protected from corrosion by installation in vacuum housings.

The current in the bus bars or in anode rods can be measured by means of iron-free transportable equipment with an accuracy of ca. 1 % [69].

5. Mercury Cell Process

The clean separation of chlorine from the cathode products is possible because of the high overvoltage of hydrogen at the mercury electrode. Hydrogen and sodium hydroxide are not produced at the cathode; instead, sodium is produced and dissolves in the mercury as an *amalgam.* The liquid amalgam is removed from the electrolytic cell to a separate reactor, called the decomposer or denuder, where it reacts with water in the presence of a catalyst to form the sodium hydroxide and hydrogen gas. The process may also be used to produce potassium hydroxide by feeding the cell with potassium chloride solution, although this is much less common. The sodium hydroxide is produced from the denuder at a concentration of ca. 50 wt %; the maximum value is 73 wt %. The hydroxide solution is very pure and almost free from chloride contamination.

The process was developed in 1892 almost simultaneously by H. Y. Castner and C. Kellner and used on an industrial scale, although the amount of chlorine produced remained relatively small until 1930, when the rapid growth of the rayon (artificial silk) industry, especially in Germany, increased the demand for pure

chloride-free sodium hydroxide solution. At this time, the horizontal high-current cell was developed and output increased rapidly. The development work in Germany was described in the FIAT final reports, published after World War II, and this led to widespread use of the process in Europe and Japan [70]. In the United States, the mercury cell process became more widespread, increasing its share of chlorine production from 3 – 4 % in 1945 to 20 % in 1960, reaching a maximum of 27 % in 1970.

The development of the mercury cell can be followed in the technical data: the cell current increased from 3.4 kA in 1895 to ca. 30 kA in 1945, 200 kA in 1960, and 450 kA in 1970. The current density rose from 2 kA/m^2 in 1950 to the current maximum of 15 kA/m^2. The cell area increased over the same period from ca. 7 m^2 to 37.5 m^2, while the *k*-factor (specific voltage coefficient, see page 16) was reduced by 50 %.

Since 1972 the importance of the mercury cell has decreased. Increasing concern about the effect of mercury on the environment has led to a considerable increase in the number and variation of statutory regulations that affect the mercury cell process.

In particular, widespread concern about cases of mercury poisoning in Japan, which were not related to the mercury cell process, caused the process to be legally banned since 1972. However, conversion to the alternative processes was delayed because of demand for low-chloride sodium hydroxide and because of the anticipated advantages of the rapidly developing membrane cell process. The last remaining mercury cell installations for NaCl were closed in 1986.

In the other countries, existing mercury cell plants are still in operation, but official regulations and uncertainty about possible further legal restrictions have hindered expansion.

In Europe and the United States great efforts are being made to develop methods of protecting the environment from mercury (see Section 5.3.5). These measures have greatly reduced emissions of mercury into the atmosphere and into wastewater to the extent that the present levels of emitted mercury are negligible in comparison to those arising from natural sources, such as volcanic action, geological erosion, or other nonnatural sources such as fuel combustion or metallurgical processes.

In 1984, the mercury cell process accounted for 45 % of world chlorine production [71]. Since then no new plants have been built. In the coming decades most of the existing mercury cell plants will be shut down or converted to membrane cell technology. Only plants with speciality products such as extremely pure sodium hydroxide, potassium hydroxide, alcoholates, and dithionites will use the mercury process in future. These plants will meet the highest emission control standards.

5.1. Principles

The cathode reaction

$$Na^+ + e^- + Hg_x \rightarrow NaHg_x$$

forming sodium amalgam, is followed by the decomposition reaction in a separate reactor

$$2\,NaHg_x + 2\,H_2O \rightarrow 2\,NaOH + H_2\,(g) + 2\,Hg_x$$

Process Description (Fig. 13). Mercury flows down the inclined base of the electrolytic cell (A). The base of the cell is electrically connected to the negative pole of the d.c. supply. On top of the mercury and flowing cocurrently with it is a concentrated brine with a sodium chloride content of ca. 310 g/L at the inlet. Anodes are placed in the brine so that there is a small gap between the anode and the mercury cathode. The concentration of the amalgam is maintained at 0.2 – 0.4 wt % Na, so that the amalgam flows freely (Fig. 14). The chlorine gas and depleted brine (270 g/L) flow out of the cell, either separately or as a two-phase mixture separated later in the process. The amalgam flows out of the cell through a weir and into the decomposer. The amalgam may be passed through a water wash between the cell and the decomposer to remove traces of sodium chloride. The amalgam flows through the decomposer countercurrent to a flow of softened or demineralized water in the presence of a catalyst to produce sodium hydroxide solution and hydrogen. Stripped of its sodium, the mercury flows out of the lower end of the decomposer and is recirculated through a pump back into the cell.

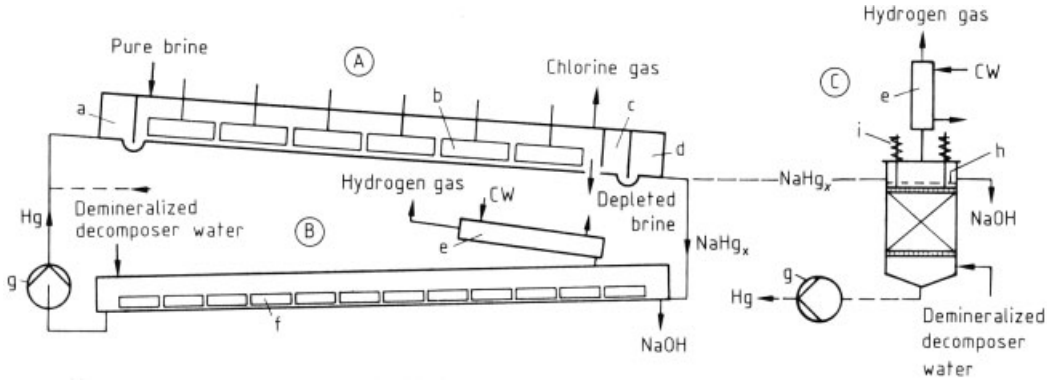

Figure 13. Schematic view of a mercury cell with decomposers
A) Mercury cell: a) Mercury inlet box; b) Anodes; c) End box; d) Wash box
B) Horizontal decomposer: e) Hydrogen gas cooler; f) Graphite blades; g) Mercury pump
C) Vertical decomposer: e) Hydrogen gas cooler; g) Mercury pump; h) Mercury distributor; i) Packing pressing springs

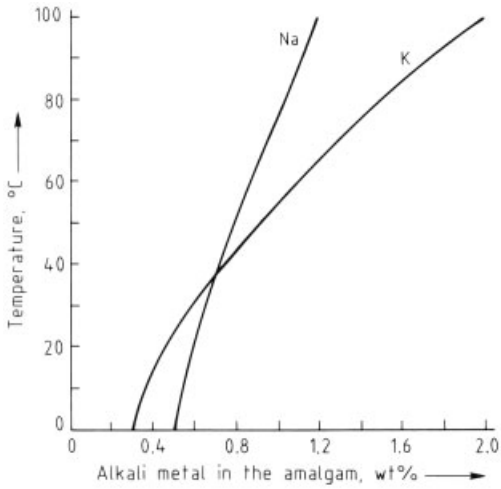

Figure 14. Freezing point curves of sodium amalgam and potassium amalgam

Anode Reactions. The oxidation of chloride ions to chlorine gas has a standard potential of 1.358 V. In a 300 g/L sodium chloride solution at 70 °C, the reversible reaction potential is reduced to 1.248 V [15, p. 339]. Some side reactions occur, such as the oxidation of OH^- and SO_4^{2-} ions and the electrochemical formation of chlorate ions. Nonelectrochemical reactions also take place in the region of the anode, such as hypochlorite formation (because of hydrolysis of chlorine) and chlorate formation. All of these side reactions represent a loss of efficiency.

Cathode Reactions. The standard potential of the hydrogen-liberating reaction is 0 V, which is considerably higher than the potential for the formation of 0.2 wt % sodium amalgam, − 1.868 V. However, hydrogen is not liberated at the mercury surface because the reaction is kinetically inhibited. Mainly sodium ions are discharged. At the sodium chloride concentrations used, the reversible potential is reduced by ca. 0.2 V. (Exact values of the discharge potential are given as a function of the sodium concentration in the amalgam, the sodium chloride concentration in the brine, and the temperature [72].) Electrochemical side reactions occur: the reduction of chlorine molecules or hypochlorous acid and the liberation of hydrogen gas. In addition, sodium in the amalgam can react directly with free chlorine, or chlorite and chlorate ions can be reduced to chloride by the action of nascent hydrogen at the cathode. All of these side reactions represent a loss of efficiency, normally ca. 2 – 4 % under good operating conditions.

Contamination of the system by heavy metals can lead to a reduction of the hydrogen discharge potential at the mercury cathode, thus increasing hydrogen liberation, and reducing amalgam formation [73]. The hydrogen concentration in the chlorine can increase to the point at which the cell and downstream chlorine handling equipment contains explosive mixtures. The probability of such problems is estimated by a hazard analysis of an existing plant [74 – 76].

The cell system is sensitive to trace quantities of catalysts in the brine, for example, vanadium, molybdenum, and chromium at the 0.01 – 0.1 ppm level or iron, cobalt, nickel, and tungsten at the parts per million level. Magnesium, calcium, aluminum, and barium are also active at the parts per million level.

In addition, relatively high concentrations of sodium in the amalgam (> 0.5 wt %) can cause increased hydrogen evolution in the cells. Potassium chloride electrolysis is considerably more sensitive to both catalysts and high concentration in the amalgam than the sodium chloride process.

Current Efficiency. The theoretical electrochemical equivalents representing the materials produced or consumed in the electrolysis of sodium chloride or potassium chloride brines are given in Table 4. In practice, the yield is ca. 95 – 97 % of the theoretical value, owing to side reactions at the electrodes and in the electrolyte. With activated titanium anodes, the yield is largely independent of the distance between the electrodes.

The decrease in salt concentration Δc is determined by the current I, the brine flow rate M, and the electrochemical equivalent f.

$$\Delta c = f\,I/M$$

The usual units are c in g/L, f in kg kA^{-1} h^{-1}, I in kA, and M in m^3/h.

Cell Voltage. The d.c. voltage across the cell circuit is determined by five factors:

1) The reversible decomposition voltage of the salt
2) The overpotentials of the chlorine and alkali metal at the electrodes
3) The voltage drop in the electrolyte
4) Voltage losses in the bus bars, switches, electrical conductors, anode materials, and cathode
5) The operating current density of the cells

Factor 1. The reversible decomposition potential of NaCl under standard conditions is $E^\circ = 3.226$ V (KCl $E^\circ = 3.234$ V). Under the operating conditions $c_{\mathrm{NaCl}} = 290$ g/L, $c_{\mathrm{amalgam}} = 0.15$ %, and 70 °C, the reversible decomposition voltage is $E = 3.095$ V [77].

Factor 2. The overpotential of chlorine depends on the material and shape of the anodes. At the high current densities (10 kA/m^2) present in modern cell rooms, the overpotential can reach several hundred millivolts, outweighing the effect of concentration changes in the electrolyte (concentration polarization) [78] and the retarding effect that formation of molecular chlorine has on the process of ion discharging [79]. Chlorine gas bubbles cover part of the anode surface, thereby increasing the current density at the free surface. The anode is designed so that the gas bubbles are liberated as quickly as possible. The rapid removal of these gas bubbles from the reaction zone is one of the advantages of titanium anodes over graphite anodes [80].

The overpotential of sodium on the amalgam cathode is caused by the limited diffusion rate of the liberated sodium atoms into the amalgam, but it is small compared to the chlorine overvoltage.

Factor 3. The specific conductivity of sodium chloride solutions increases with concentration and temperature (Fig. 15), but is independent of pH over the range 2 – 11. The brine normally enters the cells at 60 – 70 °C and leaves the cells at 75 – 85 °C. The conductivity of potassium chloride solutions at 70 °C is 30 % greater than that of sodium chloride solutions. Chlorine gas bubbles in the electrolyte increase the resistance between anodes and cathode. Better circulation of the electrolyte in the gap between electrodes allows more rapid removal of gas bubbles, thus reducing the voltage.

Factor 4. The voltage losses in the cell room are minimized by compactly arranging the cells, which shortens the current path. The relatively low conductivity of steel cell bases can be improved by copper or aluminum fittings. These measures also reduce problems caused by magnetic fields, which occur in wide cells [81].

Factor 5. In practice, cell current density and cell voltage have a linear relationship. The slope of the line is termed the specific voltage coefficient or k-factor, a useful measure of the specific energy requirement of cells produced by different manufacturers.

Table 4. Electrochemical equivalents f, kgkA^{-1}h^{-1}

Element	Element produced	Salt required	Alkali produced
Na	0.8580	2.1810 (NaCl)	1.4923 (NaOH)
K	1.4586	2.7816 (KCl)	2.0931 (KOH)
Cl_2	1.3228 (0.4115 m^3 STP)		
H_2	0.0376 (0.4185 m^3 STP)		

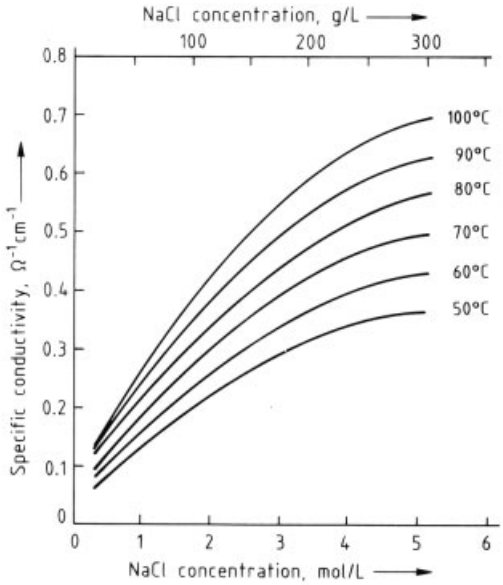

Figure 15. Specific conductivity of sodium chloride solutions

The cell voltage is given by $U_{\text{cell}} = 3.15 + kJ$, J = current density, kA/m^2, k = specific voltage coefficient, V m^2 kA^{-1}.

Computer-controlled cells with activated titanium anodes are run with k-factors from 0.085 to 0.11. The corresponding cell voltages at 10 kA/m^2 are 4.00 – 4.25 V (Fig. 16).

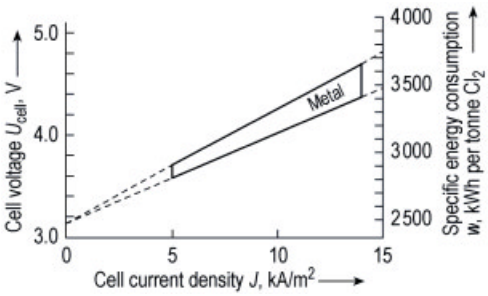

Figure 16. Cell voltage and specific energy consumption per tonne of Cl_2 versus cell current density

In addition to the d.c. voltages considered above, there are energy losses across the transformer and rectification equipment. All cell installations use a.c. power, which is rectified by silicon diodes in which the energy losses are minimized by operating at greater than 100 V. This voltage is achieved by operating at least 25 cells in series.

Energy Consumption. To operate a cell installation economically, the consumption of d.c. electrical energy per unit mass of product must be minimized. The specific energy consumption w is given by

$$w = 1000\, U_{\text{cell}}/a\, f$$

where w = kWh/t, a = yield factor or current efficiency, f = electrochemical equivalent, kg kA^{-1} h^{-1}

For example, if the cell voltage U_{cell} is 4.20 V and the current efficiency is 0.970, then ca. 3275 kWh is required to produce 1 t of chlorine + 1.13 t of NaOH. Since af is almost a constant, the specific energy consumption per tonne of chlorine w is effectively proportional to the cell voltage. In that case, w also depends on the cell current density (see Fig. 16). In the example, w = 3275 kWh corresponds to 10 kA/m^2.

The total energy requirement per tonne of Cl_2 must also include the transformer and rectifier losses (30 – 40 kWh/t) and the energy requirements of all of the ancillary equipment (120 – 160 kWh/t).

A mathematical model of the cell has been described [82].

Decomposition of the Amalgam. The amalgam is decomposed in horizontal decomposers, alongside or beneath the cell, or more often since ca. 1960 in vertical decomposers or denuders. The energy stored in the amalgam

has an emf of ca. 0.8 V. The hydrogen overpotential at the amalgam prevents spontaneous decomposition in contact with water, and a catalyst (depolarizer) must be used. The overall decomposition reaction is

$$2\,NaHg_x + 2\,H_2O \rightarrow 2\,NaOH + H_2 + 2\,Hg_x$$

and takes place in two stages, first as an anode reaction at the surface of the amalgam

$$2\,Na \rightarrow 2\,Na^+ + 2\,e^-$$

and then as a cathode reaction on the catalyst surface, where the water is decomposed

$$2\,H_2O + 2\,e^- \rightarrow 2\,OH^- + H_2$$

Industrial decomposers are essentially short-circuited electrochemical primary cells (Fig. 17). The most common *catalyst* is graphite [*7782-42-5*], usually activated by oxides of iron, nickel, or cobalt or by carbides of molybdenum or tungsten. The hydrogen overpotential on graphite (0.5 – 0.6 V at 2 kA/m^2 and 80 °C) increases with current density and decreases with temperature; therefore, the decomposer should be operated at as high a temperature as possible [83]. Good catalyst material must meet many requirements: resistance to alkali solutions, hydrogen, and mercury; low hydrogen overpotential; good electrical conductivity; long-lasting activity; wettability by amalgam; and incapability of amalgamation.

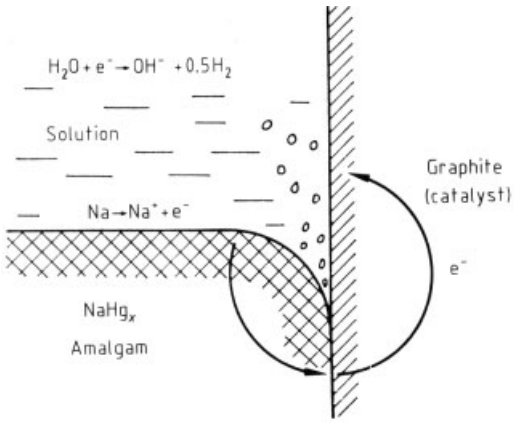

Figure 17. Principle of amalgam decomposition

Attempts to recover some of the energy stored in the amalgam by creating an electrical circuit by using the catalyst as the anode separated from the amalgam or by using the amalgam electrode with an oxygen gas diffusion electrode have so far had no practical outcome [84].

Horizontal decomposers are ducts with a rectangular cross section, which are installed with a 1 – 2.5 % slope near to or underneath the cells. The amalgam flows in a stream ca. 10 mm in depth, and the sodium content is thereby reduced to < 0.02 wt %. The catalyst consists of graphite blades 4 – 6 mm thick, which are immersed in the amalgam in a lengthwise direction (Fig. 18, also see Fig. 13 B). The water for the reaction, which is softened or demineralized by ion exchange, flows in the direction opposite the amalgam and is removed as 50 % caustic alkali solution. The hydrogen gas is cooled as it leaves the decomposer so that any condensed water and mercury run back into the decomposer. Advantages of the horizontal decomposer are serviceability, simple construction, and a pure product that is low in mercury. However, horizontal decomposers require a greater mercury inventory than vertical decomposers.

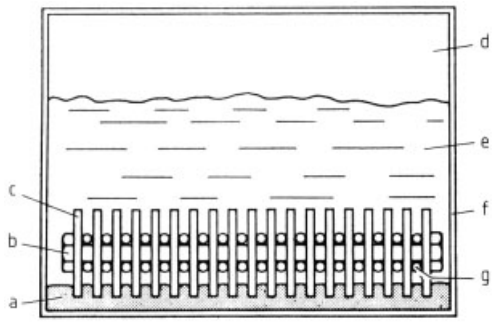

Figure 18. Cross section through a horizontal decomposer a) Amalgam; b) Bolt; c) Graphite blades; d) Hydrogen; e) Sodium hydroxide solution; f) Decomposer casing; g) Spacers

Vertical decomposers are designed as towers [85] containing packings of activated graphite spheres or other shapes 8 – 20 mm in diameter. The towers are packed 0.6 – 0.8 m high. The cross section of the tower is 0.35 m^2 per 100 kA of cell current. The amalgam is fed in via an overhead distributor, and the mercury is pumped

from the base of the tower back to the cell by a closed centrifugal pump (see Fig. 13C). The water for the reaction is fed into the base of the tower and flows upward counter to the amalgam. The 50 % caustic alkali solution flows out at the top. The smaller volume of the vertical decomposer leads to higher product temperature because of the greater energy intensity of the system. Cooling the hydrogen is essential. Compared with the horizontal decomposer, the amount of space required is small, and the mercury inventory is small, but the caustic alkali contains more mercury.

In alternative decomposition reactions, other products may be obtained from the amalgam in place of sodium or potassium hydroxide solutions [10, p. 518], [86]: sodium sulfide from sodium polysulfide solution, alcoholates from alcohols, sodium dithionite from sodium hydrogen sulfite, hydrazobenzene or aniline from nitrobenzene, adiponitrile from acrylonitrile, and alkali metals by distillation.

5.2. Mercury Cells

During the first decades after the rocking cells of Castner and Kellner were first commissioned, considerable efforts were made to develop suitable materials for the cells and the anodes. A large number of cell configurations were tested, resulting in the development of the continuous cell. Since 1950, the cell areas and the specific load were increased considerably.

In 1972, the changeover from graphite to metallic anodes began, with a parallel development of computer monitoring and control, leading to improved short-circuit protection and a reduction of the specific energy consumption by computer-controlled anode adjustment, of great significance in view of the drastic increase in electricity costs in the late 1970s. In the years following 1972, producers operating the electrolysis plants also concentrated on the development and installation of devices to reduce mercury emissions.

The cells currently available possess a number of common features. The mercury flows over a steel base that has a slope of 1.0 – 2.5 %. The flanged side walls are lined with rubber. The cell covers are mostly steel, lined with rubber or titanium on the underside, but they may also be rubberized fabric. The anodes, generally made of activated titanium, hang in groups from carrying devices that can be varied in height manually, hydraulically, or by motor-operated lifting devices. Each cell can be short-circuited externally by a switch. The cell bus bars are usually copper. The anodes are protected from internal short circuits by means of electronic monitoring systems. The size of the cells can be varied within a broad range to give the desired chlorine production rate. Computer programs optimize the cell size, number of cells, and optimum current density as a function of the electricity cost [87] and capital cost.

For comparison, a list of cells manufactured by leading engineering firms and cell characteristics is given in Table 5. Cathode surface areas are ca. 17 – 30 m^2, and nominal currents are ca. 170 – 300 kA [32, p. 204].

5.2.1. Uhde Cell

The Uhde cells (Fig. 19, also see Fig. 22) are available with a cathode surface area 4 – 30 m^2 for chlorine production rates 10 – 1000 t/d for the complete cell installation. The brine flows in via an inlet box fitted with two pipes for the removal of chlorine. The weak brine is removed at the end of the cell. The solid cover is fixed to the side walls by clamps. The anodes are suspended in groups in carrying frames supported near the cells on transverse girders with lifting gear. The anode rods are raised and lowered within a bellows seal. Short copper bus bars between the cells also serve for shunt measurement of the anode currents. The electric current is brought in above the cell covers via flexible copper straps that run immediately above the anode rods and are bolted to them. The compressed-air switches are situated under the cells. The cell bottom is usually a current conductor when cells are short-circuited, but in wide, heavily loaded cells the cathode current is carried by copper bus bars to prevent the occurrence of strong magnetic fields, which could interfere with the amalgam flow. The automatic equipment for protection and adjustment of the anodes depends on the shunt measurement of the currents and is controlled by a central computer. In this way, an optimum k-factor is selected for each cell. The vertical

Table 5. Characteristics of modern mercury cells

Characteristic	Manufacturer				
	Uhde	De Nora	Olin – Mathiesen	Solvay	Krebs Paris
Cell type	300 – 100	24M2	E 812	MAT 17	15 KFM
Cathode area, m^2	30.74	26.4	28.8	17	15.4
Cathode dimensions, $l \times b$, m^2	14.6 × 2.1	12.6 × 2.1	14.8 × 1.94	12.6 × 1.8	9.6 × 1.6
Slope of cell base, %	1.5	2.0	1.5	1.7	
Rated current, kA	350	270	288	170	160
Max. current density, kA/m^2	12.5	13	10	10	10.4
Cell voltage at 10 kA/m^2, V	4.25	3.95	4.24	4.10	4.30
Number of anodes	54	48	96	96	24
Stems per anode	4	4	2	1	4
Number of intercell bus bars	36	32	24	24	12
Quantity of mercury per cell, kg	5000	4550	3800		1650
Energy requirement per tonne of Cl_2, kWh d.c.	3300	3080	3300	3200	3400

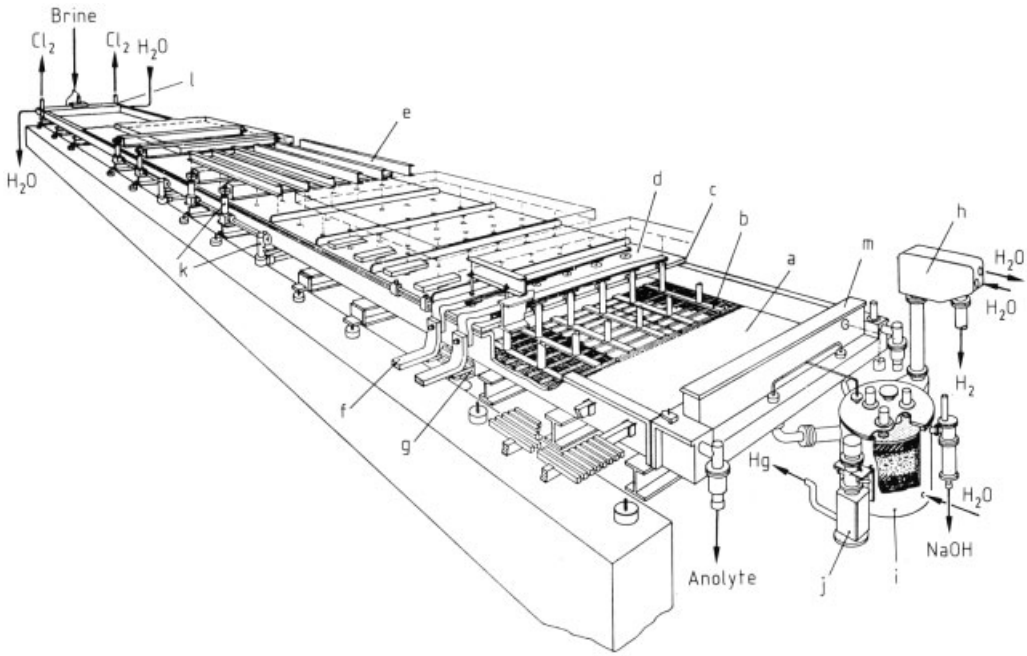

Figure 19. Uhde mercury cell
a) Cell base; b) Anode; c) Cover seal; d) Cell cover; e) Group adjusting gear; f) Intercell bus bar; g) Short-circuit switch; h) Hydrogen cooler; i) Vertical decomposer; j) Mercury pump; k) Anode adjusting gear; l) Inlet box; m) End box

decomposers are provided with hydrogen coolers and are situated at the end of each cell. The amalgam flows into the decomposer under the force of gravity [88].

5.2.2. De Nora Cell

The size of the De Nora cell (Fig. 20, also see Fig. 23) varies from 4.5 to 36 m^2, corresponding to electric currents from 45 to 400 kA. The cover is a flexible multilayer sheet of elastomer spread over the cell trough. This cover is supported by the anode rods and seals them. The DSA anodes (see Section 8.1) are held rigidly in strong carrying frames, which are automatically adjusted by electric motors. Individual anode adjustment is not provided. The anode rods are individually connected by flexible copper straps to the an-

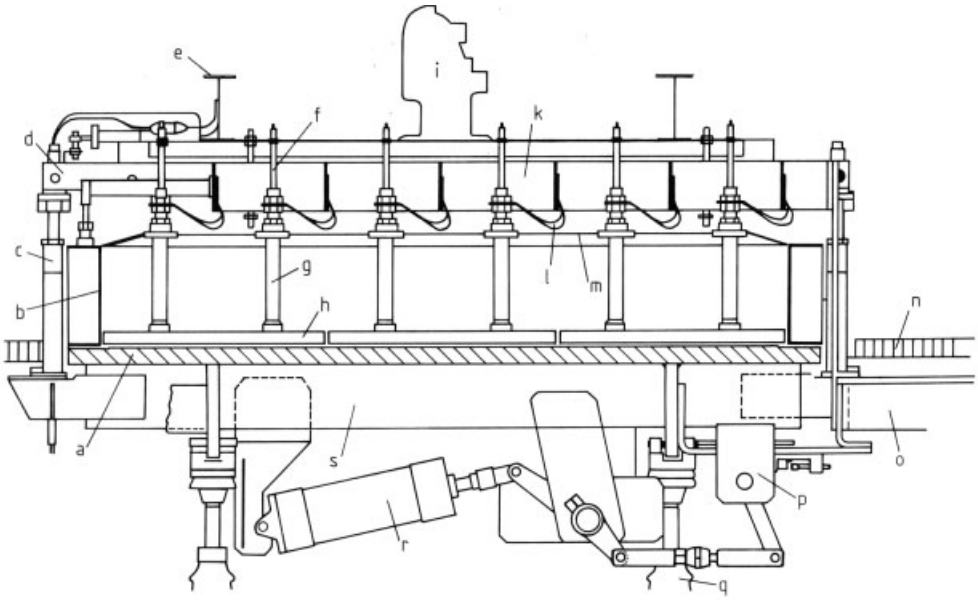

Figure 20. Cross section through the De Nora mercury cell
a) Cell base (steel); b) Side wall (rubber-lined steel); c) Lifting gear; d) Transverse support; e) Lengthwise support; f) Anode carrier; g) Anode rod; h) Anode surface; i) Adjusting motor; k) Bus bar; l) Flexible anode current strap; m) Multilayer cell cover; n) Service walkway; o) Intercell bus bar; p) Switch; q) Insulator; r) Switch drive; s) Support

ode bus bars. The cathode current is carried by copper bus bars. Devices for the improvement of brine circulation and gas removal within the cells reduce specific energy consumption. Consequently, the reduction in brine concentration can be increased from the usual 35 – 40 g/L to 60 – 70 g/L, and the brine circulating rate can be reduced by ca. 40 %. Separate outlets are present at the inlet box for the normal chlorine gas production and the weak chlorine gas produced during start-up. The graphite catalyst in the vertical decomposer is activated with molybdenum.

5.2.3. Olin – Mathieson Cell

The special feature of the Olin – Mathieson cell lies in the system of mounting and adjusting the anodes. Above each row of anode rods, a U-shaped copper or aluminum bus bar also serves to support the anode lifting gear. The anode rods are bolted to the U-shaped bus bar. The anodes are adjusted as a group, either manually or by a remote computer with the remote computerized anode adjuster (RCAA) system. The currents are measured independently of the cell potentials by means of reed contacts [89].

5.2.4. Solvay Cell

The bus bars in the Solvay cells are made primarily of aluminum. Above the cells is a cover that also serves as a convenient walkway, giving access to the anode rods. The titanium anodes are specially coated and are automatically adjusted by computer. The tall vertical decomposers are located under the cells.

5.3. Operation

The aspects of the operation of mercury cells that typically differ from those of the other processes are the brine circulation system, the cell room, treatment of the products, measurement and control, and reduction of mercury emissions.

5.3.1. Brine System

A typical brine circulation system for the mercury cell process is shown in Figure 21. In the cells the sodium chloride concentration of the brine is reduced by 35 – 60 g/L to 260 – 280 g/L

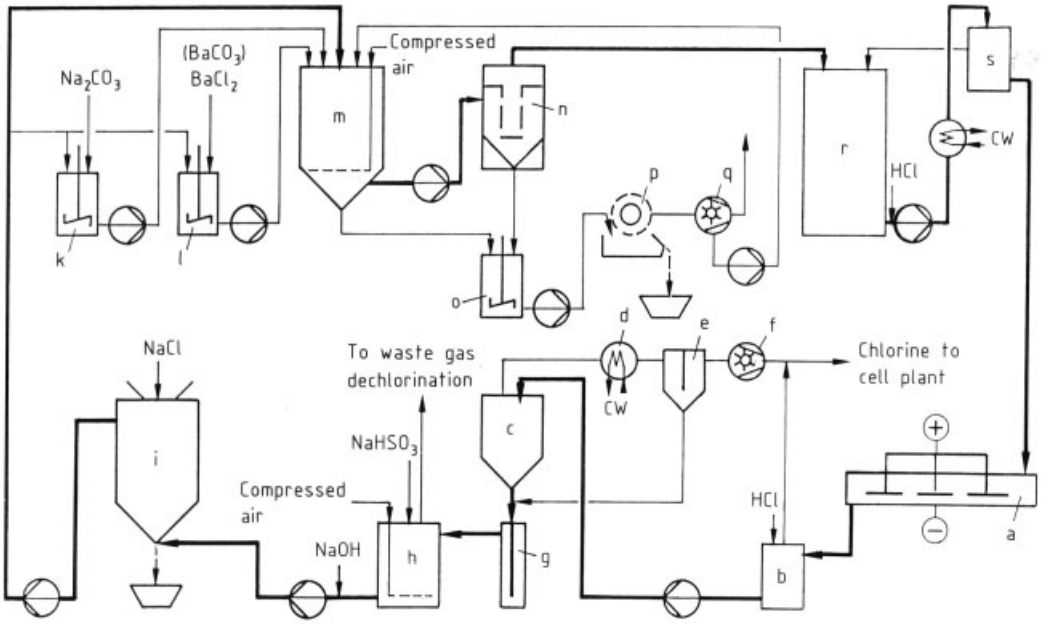

Figure 21. Schematic diagram of a brine circulation system in the mercury cell process
a) Electrolysis cell; b) Anolyte tank; c) Vacuum column dechlorinator; d) Cooler; e) Demister; f) Vacuum pump; g) Seal tank; h) Final dechlorination; i) Saturator; k) Sodium carbonate tank; l) Barium chloride tank; m) Brine reactor; n) Brine filter; o) Slurry agitation tank; p) Rotary vacuum filter; q) Vacuum pump; r) Brine storage tank; s) Brine supply tank

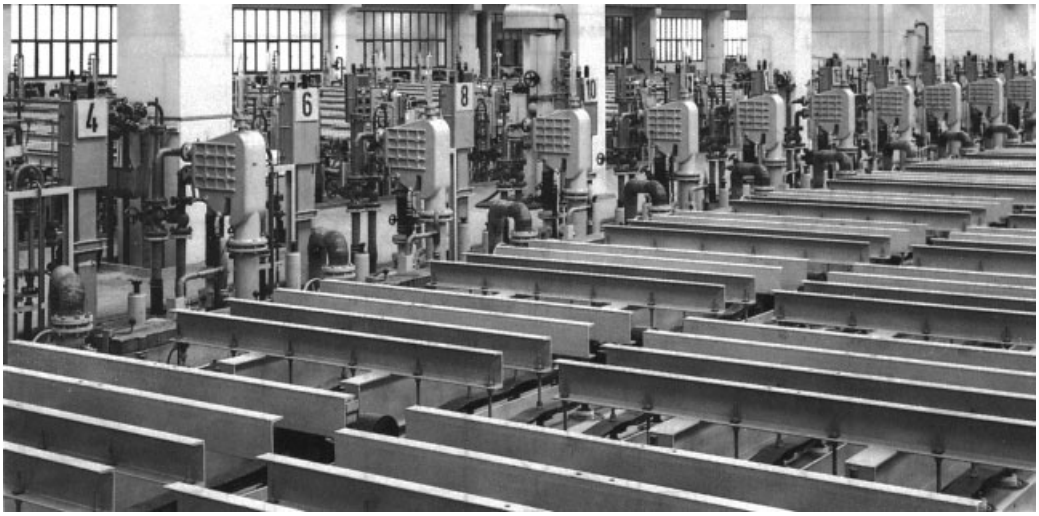

Figure 22. Cell room: Uhde mercury cells

at 70 – 85 °C. To avoid mercury emissions into the air, the resaturators are generally closed vessels. The mercury cathode is very sensitive to poisoning by heavy metals; therefore, a test [90] has been developed that allows rapid determination of the suitability of any particular salt or brine.

5.3.2. Cell Room

The cells are usually situated in a building (Fig. 22), although sometimes they are erected in open air (Fig. 23). Figure 24 shows a bird's eye view, and Figure 25 shows a cross section of cell room. The cells are arranged parallel to each other so that bus bars and supply lines are kept short. The cells stand on supporting structures and are insulated to prevent shorting to the earth. The transformer and rectifiers are situated at one end of the room, and the cell service and repair area is at the opposite end. Ancillary equipment is installed near the cell room in a spillage containment area.

Figure 23. Open-air cell room: De Nora cells

Cell floors, gangways, and spillage containment areas are constructed with smooth, sloping floors so that any mercury can be easily recovered or wash water can be conveniently collected for treatment. The supply pipes run under the cells and are connected to them by flexible, insulating connections. The heat given off by the cells and the decomposers is removed by a ventilation system.

The plant is operated with continuous 24 h/d supervision and control. An additional day-shift team carries out anode changes, repairs, and cleaning.

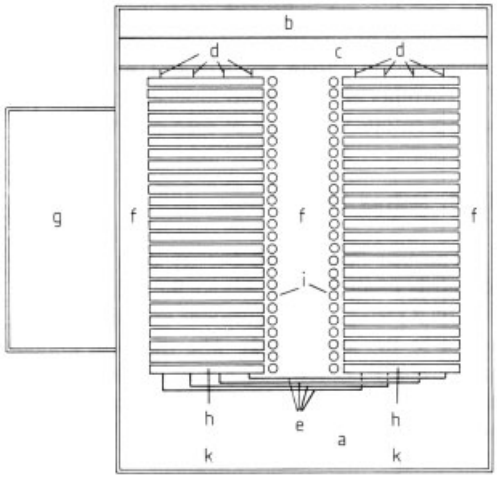

Figure 24. Mercury cell room (bird's-eye view, schematic) a) Cell room; b) Transformer room; c) Rectifier room; d) Bus bars; e) Turnaround bus bars; f) Service walkways; g) Ancillary equipment; h) Electrolysis cells; i) Vertical decomposers; k) Cell assembly and maintenance area

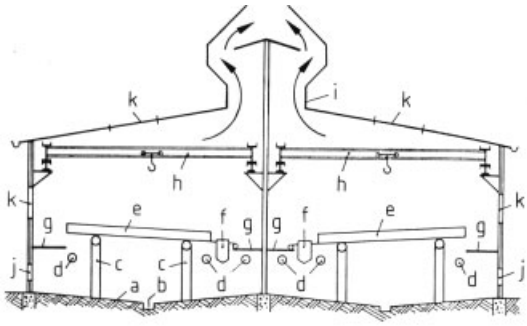

Figure 25. Mercury cell room (cross section, schematic) a) Basement floor; b) Floor drains; c) Cell supports with insulators; d) Supply pipes; e) Cells; f) Decomposers; g) Service walkways; h) Crane; i) Ridge ventilator; j) Ventilation air supply; k) Windows/lighting

Occupational Health. Anyone working in the cell area must undergo regular health checks. Euro Chlor has prepared a Code of Practice "Control of Worker Exposure to Mercury in the Chlor-Alkali Industry" [91]. The U.S. Chlorine Institute has released guidelines "Medical Surveillance and Hygiene Monitoring Practices of Worker Exposure to Mercury in the Chlor-Alkali Industry" [92]. The U.S. Environmental Protection Agency has established 18 rules re-

lating to cleanliness of the cell room [93]. Adherence to these rules eliminates any danger to the health of personnel caused by mercury. The maximum allowable concentration or threshold limit value (TLV) of mercury in the atmosphere in Western Europe and in the United States is between 0.025 and 0.100 mg/m^3.

5.3.3. Treatment of the Products

Chlorine. See Chapter 11.

Hydrogen. The treatment of the hydrogen gas leaving the decomposer is described in Chapter 4. It must pass special equipment for the removal of the traces of mercury before it is used (see page 25).

Sodium Hydroxide Solution. The great advantage of the mercury cell process is that very pure sodium hydroxide solution is produced (see Table 20) at a suitable concentration. The chloride content is only 5 – 50 mg/kg.

Sodium hydroxide from the decomposer usually has a concentration of 50 % and a temperature of 80 – 120 °C. It passes through rubber-lined steel pipe work to nickel or Incoloy coolers, where it is cooled to 40 – 60 °C. Any particles of graphite from the decomposer or traces of mercury are effectively removed by centrifuges, candle filters, or precoated leaf filters (Fig. 26).

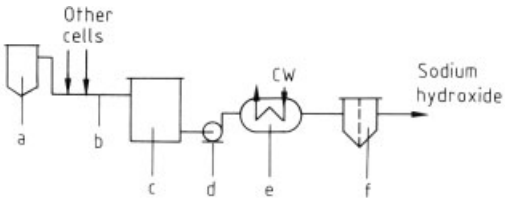

Figure 26. Processing of sodium hydroxide solution from the amalgam decomposer
a) Vertical decomposer; b) Collection main; c) Collecting tank; d) Pump; e) Cooler; f) Mercury removal filter

The freezing-point and boiling-point curves of sodium hydroxide solutions are shown in Figure 27. The phases separating from the solution, i.e., ice, hydrates, and NaOH, are indicated along the freezing-point curve. Sodium hydroxide is supplied to consumers as aqueous solution, solid block, flakes, prills, or powder. For the processes involved and for uses, see → Sodium Hydroxide.

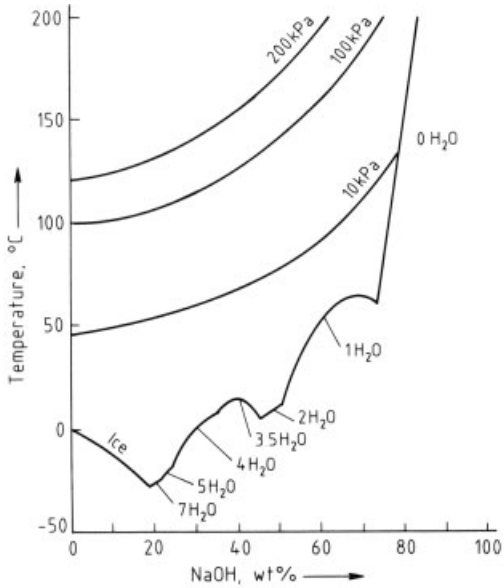

Figure 27. Freezing and boiling point curves of sodium hydroxide solutions

5.3.4. Measurement

The condition of the brine, the cells, and the products must be continuously and carefully monitored, since even small deviations from the correct conditions can increase the hydrogen concentration in the chlorine. The measuring operations are mostly automatic: critical limits are chosen and if these limits are exceeded, alarms are set off.

Cell Operating Conditions. The sodium concentration in the amalgam is determined at the cell inlet (max. 0.05 wt %) and outlet (max. 0.45 wt %).

A 20-g sample of the amalgam is reacted with 30 wt % aqueous sulfuric acid in an absorption pipette. The evolution of 1 cm^3 of gas corresponds to 0.01 wt % Na.

A portable analytical and recording apparatus is available that works electrochemically [94].

If the mercury pump stops, the steel cathode base of the cell is exposed to electrolyte, and hydrogen evolves to form an explosive mixture with the chlorine in the cell.

Failure of the mercury pump or mercury flow automatically short-circuits the cell. Mercury flow failure is detected by monitoring the mercury level at the lowest part of the mercury circulation system or in the inlet box, by direct flow measurement [95], or by loss of pressure at the pump delivery.

The motors for the mercury pumps, the chlorine absorption plant, and the most vital control equipment are all provided with an emergency power supply, ensuring safe shutdown of the plant if a power failure occurs. The installation is protected by a complex system of interlocks so that failure of important equipment, such as the chlorine compressor, shuts down the rectifiers.

A large number of systems have been developed for the protection of cells from short circuits. Titanium anodes are destroyed by short circuits and must be raised before any contact with the amalgam takes place.

The operation of the monitoring system depends on magnetic-field current measurement for individual anodes [89, 96, 97] or on shunt measurement of the supply bus bars [98]. Monitoring is achieved by comparison of the anode – cathode voltage of different cell sections [99] or by following the conductivity of the brine in the electrode gap [100].

The signals from the instruments are fed into central computers or local microprocessors at each cell [101] that control the anode lifting gear. The mercury inventory in each cell may be measured by a radioactive tracer technique once a year without affecting cell operation [102].

Products. The concentration of the sodium hydroxide solution is determined from its density, and the purity its checked by titration to determine hydroxide, carbonate, and chloride contents. The purity of the water for the decomposer is determined from its conductivity.

The oxygen content of the hydrogen gas is determined from the magnetic susceptibility, oxygen being paramagnetic.

5.3.5. Mercury Emissions (→ Mercury, Mercury Alloys, and Mercury Compounds, Chap. 5)

Any chlor-alkali plant up to modern technical standards is not a hazard to the environment. The residual emissions to water and air are ecologically acceptable. The mercury in the electrolytic cells circulates in a closed system. All materials that come into contact with the mercury – equipment, products, auxiliary chemicals, wash water, waste gases, other waste materials – may become contaminated with mercury, and must be treated before release to the environment or must be safely deposited. For the exact measurement of these trace amounts and for control of the effectiveness of the measures to reduce the emissions, analytical methods have been developed with sensitivities in the microgram region [103].

Many countries have set legal limits on emissions in waste air and water. The limits on the products of a chlor-alkali plant may depend on their end use, e.g., drinking water treatment or food processing. The sources of contamination are listed, and means of reducing them are described: Mercury cells are sealed vessels, and the products are conveyed in closed pipes. The cell rooms must have smooth joint-free floors with easily cleaned drainage surfaces and irrigated collection gutters (see Fig. 25). Spilled mercury is immediately washed away with water into collecting tanks or sucked up with a vacuum system.

Control of mercury loss is only possible if the mercury content of all the cells is known exactly. The gravimetric and volumetric methods formerly used were cumbersome and led to additional mercury emissions, disadvantages that are avoided by a radioactive tracer method.

Mercury in Products. Hot, moist *chlorine* leaving the cell contains small amounts of mercuric chloride. This is almost completely washed out in the subsequent cooling process and may be fed back into the brine with the condensate. In the cooled and dried chlorine gas, there are only minute traces of mercury: 0.001 – 0.01 mg/kg.

The equilibrium mercury concentration in *hydrogen gas* is a function of temperature and pressure. The mercury concentration (mg of Hg per m^3 of H_2 at 101.325 MPa) increases rapidly with temperature:

t, °C	0	20	40	60	80	100
c, mg/m^3	2.36	14.1	66.1	255	836	2404

Subjecting the hydrogen gas to pressure lowers the mercury content of the resulting prod-

uct gas at atmospheric pressure. For example, at 5 °C

$$c_{Hg}\, p = 0.37\ \text{Pa kg m}^{-3}$$

c_{Hg} = concentration of mercury in hydrogen gas at atmospheric pressure, mg of Hg per m^3 of H_2

p = pressure to which the hydrogen gas is subjected, MPa

When the mixture is cooled to 2 – 3 °C, the mercury concentration is reduced to ca. 3 mg/m^3 at standard pressure. This mercury content can be reduced by compressing and further cooling, adding chlorine to form mercurous chloride (calomel), which is collected on rock salt or similar material in a packed column, washing with solutions containing active chlorine, or by adsorption on activated carbon impregnated with sulfur or sulfuric acid, leaving a mercury concentration in hydrogen of 0.002 – 0.03 mg/m^3. The highest purity can be achieved by adsorption on copper/aluminum oxide or silver/zinc oxide, < 0.001 mg/m^3.

Centrifugation or filtration in candle filters or in disk filters precoated with charcoal gives *sodium hydroxide solutions* containing mercury concentrations of < 0.05 ppm (mg/kg of 50 % caustic soda).

The circulating brine contains mercury concentrations of 2 – 20 mg/L. Mercury emissions from the brine system can occur through losses of brine into the wastewater, by brine vaporization in the resaturators, or by disposal of the residues from the brine purification filter. These emissions are minimal at a chlorine concentration < 30 mg/L, giving a redox potential > 500 mV vs. NHE. Under these conditions mercury remains dissolved in the brine as a mercury chloride complex even if the brine is alkaline.

Mercury in Wastewater. Mercury-containing wastewater has several sources:

1) The process, e.g., condensate and wash liquor from treatment of chlorine, hydrogen, and brine; stuffing-box rinse water from pumps and blowers; brine leakages; ion-exchange eluate from process-water treatment
2) Cell cleaning operations
3) Cleaning of floors, tanks, pipes, and dismantled apparatus

The amount of wastewater can be reduced by separately disposing of the cooling water and process water and by feeding the condensate back into the brine, provided the water balance allows this. A wastewater rate of 0.3 – 1.0 m^3 per tonne of chlorine is achievable.

There are various methods of making wastewater suitable for discharge:

1) Chemical removal of mercury by reducing any compounds to the metal with hydrazine or sodium borohydride or by precipitating mercuric sulfide with thiourea or sodium sulfide. The mercury metal or sulfide is then filtered off.
2) Oxidation of the mercury by chlorine, hypochlorite, or hydrogen peroxide and adsorption on an ion-exchange medium. Elutriation is done with hydrochloric acid, which is then used to acidify the brine [104].

The Clean Water Act of 1972 (United States) demands the use of the "best available technology economically achievable." Since 1982 each plant has been limited to a maximum of 0.1 g of Hg per tonne of chlorine averaged over 30 d measured at the outlet of the wastewater treatment plant.

In Western Europe, an EC directive has been issued on the subject of the mercury content of wastewater from chlor-alkali plants, following various earlier agreements such as the Rhine protection agreement, the EC guidelines concerning the protection of natural waters, and the Paris Convention [105]. This directive requires plants with circulating brine systems to have a limit of 1.0 g of Hg per tonne of chlorine produced.

Mercury in Process Air. Air from the process, for example, the cell end box ventilation system, vents from liquid collection tanks (caustic, wastewater), from the vacuum cleaning system, or from the distillation unit for mercury-contaminated wastes can be treated to remove mercury by the methods used for hydrogen.

Ventilation of the Cell Room. The heat produced during electrolysis requires that the air must be changed 10 – 25 times per hour, depending on the type of building. Mercury spillage can

occur during essential operations involving cells or decomposers, for example, opening the cells for anode changing or cleaning, assembling or dismantling equipment, or replacing defective pipes. Spillage leads to small losses in the exhaust air owing to the vapor pressure of mercury. In addition, products that contain mercury, such as the sodium hydroxide solution, hydrogen, or process waste air, can escape via faulty seals in pipes and equipment, leading to emissions. Closed cell construction and special care in handling mercury, i.e., good housekeeping, by adhering to the EPA rules or the Code of Practice "Mercury Housekeeping" [106], keep the mercury concentrations and, hence, emissions below the allowable work place concentrations (MAK and TLV) [107].

Purification of large volumes of waste air containing mercury in very low concentrations is not effective. In the United States, the upper limit for the emission of mercury in process waste air and hydrogen is 1 kg per day per facility, and for ventilation air it is 1.3 kg per day per facility [18, p. 372]. In Western Europe the Parcom Decision 90/3 [108] requires a standard of 2 g Hg/t of chlorine capacity for emissions to the atmosphere from existing plants.

Mercury in Residues. Mercury-containing residues include brine filter slurry, spent decomposer catalyst, discarded cell components, residues from the purification of products, waste material from rinsing media, adsorption materials, ion-exchange media, etc. Mercury can be recovered from these materials by distillation in closed retorts. The residues after distillation must be disposed at special sites. Mercury in safely deposited wastes is not considered an emission to the environment.

Summary. The continued efforts of all producers have led to a steady decrease in mercury emissions over the years; for example, in Western Europe from 16 g in 1978 to 2 g Hg per tonne of chlorine capacity in 1996, as shown in Figure 28 [109]. With this low emission level, the contribution of the chlor-alkali industry to the total natural and anthropogenic mercury emissions is less than 0.1 % [111]. Euro Chlor is developing a BAT (best available techniques) for reducing mercury emissions from existing mercury-based plants, the application of which will ensure that in 2007 no plant emits more than 1.5 g Hg per tonne of chlorine capacity to air, water, and products [109].

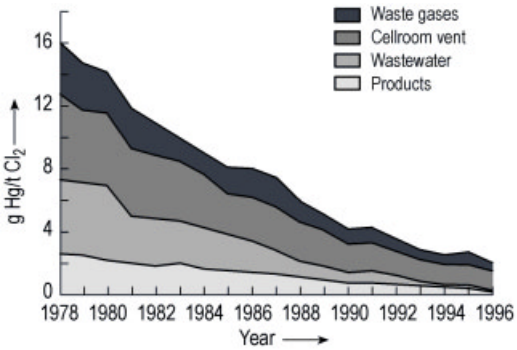

Figure 28. Mercury emissions from the European chlor-alkali industry

6. Diaphragm Process

The commercial production of chlorine by electrolytic processes began in Europe and the United States in the 1890s. Early cells of the bell-jar type had no diaphragm and relied on the flow of anolyte toward the cathode to prevent the hydroxide ion from back-migrating toward the anode. This method had limited capacity because gas evolution caused mixing and loss of efficiency. The Griesheim cell, another early design, used porous cement as the diaphragm.

E. A. Le Sueur is credited with the design of a cell incorporating a percolating asbestos diaphragm, which is the basis for all diaphragm chlor-alkali cells currently in use. When brine is caused to flow into the anolyte and subsequently through the diaphragm into the catholyte, continuous operation with much improved efficiency is obtained. This Le Sueur cell, and the similar Billiter cell developed in Germany, incorporated a horizontal asbestos sheet as the diaphragm. During the 1920s, the Billiter cell became the most widely used cell in the world; a few are still in operation.

Following the invention of synthetic graphite, numerous cells were developed. These fall into three basic types:

1) Rectangular vertical electrode cells

2) Cylindrical vertical electrode cells
3) The vertical electrode bipolar filter press cell developed by Dow Chemical

In 1913, C. W. Marsh developed a cell with finger cathodes and side-entering anodes and cathodes, which greatly increase the electrode area per unit of floor space. About 1928, Kenneth Stewart of Hooker Chemical (now Occidental Chemical Co.) developed a method of depositing asbestos fibers onto the cathode by immersing the cathode in a slurry of asbestos fibers and applying a vacuum. All significant installations of diaphragm cells currently in operation are derived from that development [2].

All diaphragm cells produce cell liquor that contains ca. 11 wt % caustic soda and 18 wt % sodium chloride. To market the caustic soda, its concentration must be increased to 50 %. During the evaporation and cooling processes, the salt becomes less soluble in the stronger caustic, and at 50 % NaOH the NaCl concentration is ca. 1 %.

6.1. Principles

The principles needed to understand the efficient operation of the diaphragm process involve the current efficiency, cell voltage, power consumption, and optimization of the operating conditions [112, 113].

The reaction at the positively charged anode is the same for all three chlor-alkali processes

$$2\,Cl^- \rightarrow Cl_2 + 2e^-$$

The reaction at the negatively charged cathode of the diaphragm cell is

$$2\,H_2O + 2\,e^- \rightarrow H_2 + 2\,OH^-$$

Figure 29 is a cutaway view of a diaphragm cell that shows the orientation of the various parts of the cell and the various reactions that take place. Figure 29 also shows the location of the diaphragm, which is deposited on the outside of the cathode screen and which separates the cell into two compartments, one containing the anodes and one containing the cathode. The sodium chloride solution (brine) enters the anode compartment and completely covers the anodes and the cathode tubes or fingers. The chlorine leaves the cell through an outlet in the cell head. The anolyte flows through the diaphragm into the cathode compartment because of the difference in liquid level between the two compartments. The catholyte is a solution of sodium chloride and sodium hydroxide because a portion of the water is converted to hydroxide at the cathode. The hydrogen produced at the same time leaves the cell through an outlet on the cathode. The solution of sodium chloride and sodium hydroxide overflows the cell through a level control pipe on the cathode, and is then commonly called cell liquor.

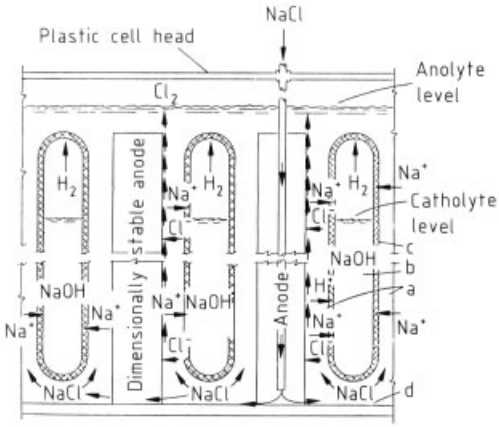

Figure 29. Basic chemical reactions within the cell
a) Anode compartment; b) Cathode compartment; c) Deposited diaphragm on cathode tubes, rims, and end screens

Current Efficiency. Current efficiency is defined as the amount of product actually produced divided by the amount of product that theoretically should have been produced on the basis of the amount of direct-current electrical energy input. The current efficiency is never 100 % because of side reactions. The efficiency of a diaphragm cell is usually based on the chlorine production.

The side reactions that lower the efficiency are a result of chlorine that enters the catholyte compartment or of hydroxide ions that enter the anolyte compartment. The amount of chlorine that enters the catholyte compartment is small.

The majority of the efficiency losses in a diaphragm cell are due to migration of hydroxide ions from the catholyte through the diaphragm and into the anolyte. This back migration takes place because the negatively charged hydroxide ions are attracted to the positively charged anodes and because of the hydroxide ion concentration gradient across the diaphragm. This migration of hydroxide ions through the diaphragm is in equilibrium with the opposing flow of brine through the diaphragm.

Three factors control the migration of the hydroxide ions into the anolyte:

1) Concentrations of hydroxide and chloride ions at the cathode side of the diaphragm
2) Flow rate of brine through the diaphragm
3) Condition of the diaphragm

These three factors in dynamic equilibrium determine the efficiency of the cell.

Factor 1. The higher the concentration of hydroxide ions in the catholyte, the larger the concentration gradient across the diaphragm, and the higher the probability of hydroxide ions crossing through the diaphragm. As a result, changing cell liquor strength strongly affects cell efficiency. The concentration of chloride ions in the catholyte also affects cell efficiency because some of the chloride ions migrate in place of hydroxide ions.

Factor 2. Decreasing brine flow rate to a cell increases the conversion of sodium chloride to sodium hydroxide and raises the hydroxide concentration in the catholyte because of reduced overflow from the cell. The decreased flow rate of brine through the diaphragm allows increased migration of hydroxide ions into the anolyte. These factors decrease cell efficiency.

Factor 3. The condition of the diaphragm is extremely important. Nonuniformity in the diaphragm results in high flow rates of brine through thin or loosely compacted areas and low flow rates through thick or compacted areas. In the areas where there is a low brine flow rate, back migration of hydroxide is increased.

The degree of inefficiency in a cell is indicated by the two products of the side reactions, oxygen in the chlorine and sodium chlorate in the cell liquor. Oxygen in the chlorine gas is the result of hydroxide ions that migrate through the diaphragm into the anolyte, where they are oxidized:

$$2\,OH^- \rightarrow 1/2\,O_2 + H_2O + 2\,e^-$$

Sodium chlorate in the cell liquor is a result of hydroxide ions that migrate through the diaphragm into the anolyte and react with chlorine before reaching the anode:

$$3\,Cl_2 + 6\,NaOH \rightarrow NaClO_3 + 5\,NaCl + 3\,H_2O$$

Equations. The simplest equations for calculating cell efficiency are based on the masses of products produced per unit of electrical input. Theoretically, 1.492 kg of sodium hydroxide and 1.323 kg of chlorine are produced per kiloampere-hour. It then follows that

$$\text{Cathode efficiency},\% = (\text{kg of NaOH} \times 100)\,/\,(Q \times 1.492 \times \text{the number of cells})$$

$$\text{Anode efficiency},\% = (\text{kg of } Cl_2 \times 100)\,/\,(Q \times 1.324 \times \text{the number of cells})$$

where Q is the quantity of electricity in kA h

Unfortunately, the production of a single cell cannot be measured with sufficient accuracy to give meaningful results. To get around this problem, the chlorine industry uses an equation based on the analysis of the chlorine gas, the cell liquor, and the anolyte:

$$\%CE = [\%Cl_2 \times 100]\,/\,[\%Cl_2 + 2\,(\%O_2) + (\%Cl_2 \times anox \times F)\,/\,c_{NaOH}]$$

where

% CE =	anode current efficiency, %
% Cl_2 =	percent chlorine in cell gas (air free)
% O_2 =	percent oxygen in cell gas (air free)
anox =	oxidizing power of anolyte expressed as grams of $NaClO_3$ per liter
c_{NaOH} =	NaOH concentration in the cell liquor, g/L
F =	conversion factor

Table 6. Typical voltage distribution [a]

Component voltages, V	Current density, kA/m^2				
	1.24	1.55	1.86	2.17	2.48
Anode potential [b]	1.30	1.30	1.30	1.30	1.31
Cathode potential [b]	1.12	1.13	1.15	1.16	1.17
Structure loss [c]	0.11	0.14	0.17	0.20	0.22
Brine loss	0.11	0.15	0.19	0.23	0.27
Diaphragm loss	0.24	0.31	0.36	0.41	0.47
Intercell bus	0.02	0.02	0.03	0.03	0.03
Total	*2.90*	*3.05*	*3.20*	*3.33*	*3.47*

[a] OxyTech MDC-55 cell with Modified Diaphragm and expandable anodes. Conditions: anolyte temperature 93 °C, anolyte NaCl concentration 250 g/L, catholyte NaOH concentration 130 g/L.
[b] Potential vs. NHE.
[c] Includes anode base, anodes, cathode, cathode screens, copper end connectors, and copper side plates.

The denominator is the amount of chlorine produced plus the amount of chlorine consumed in the side reactions. This is equivalent to the amount of chlorine that could have been produced theoretically from the input of current. The conversion factor F is the product of a volume factor, an electric field factor, and a stoichiometric factor. In practice, it is a function of cell liquor strength.

The *SIX equation* is a practical alternative to the previous equation and is often used with computers linked to a gas chromatograph and an automated cell liquor analyzer. The SIX equation is

$$\%\mathrm{CE} = [\%\mathrm{Cl_2} \times 100] \,/\, \left[\%\mathrm{Cl_2} + 2\,(\%\mathrm{O_2}) + \left(\%\mathrm{Cl_2} \times 6 \times c_{\mathrm{NaClO_3}}\right) / c_{\mathrm{NaOH}}\right]$$

This equation also accounts for chlorine lost to the anolyte. However, it approximates the oxidizing potential of the anolyte with the concentration of chlorate in the cell liquor and assumes a fixed conversion factor from anolyte concentration to catholyte, namely *SIX*. The SIX equation approximates the standard equation within 0.5 %.

Cell Voltage. The voltage of a cell is the sum of five component voltages: anode potential, cathode potential, cell structure voltage drop, diaphragm voltage drop, and anolyte – catholyte voltage drop. The anode and cathode potentials are sums of the reversible voltages, which are the thermodynamic minimum amounts of work to cause the reactions to take place, and the overpotentials, which are the additional voltages required for nonreversible kinetics. The cell structure voltage drop includes the voltage losses in the cathode, anodes, intercell bus, and all other connectors in the cell. The sum of the diaphragm voltage drop and the anolyte – catholyte (brine) voltage drop is the potential between the electrodes. All of these voltages are functions of current density. Table 6 shows how cell voltage is strongly affected by cell current. Cell temperature, feed brine NaCl concentration, and the cell liquor NaOH concentration also affect cell voltage (Table 7), because they affect the conductivity of the solutions between the electrodes. Table 7 clearly shows that current density is the most important factor.

Table 7. Factors affecting cell voltage: the change in cell voltage ΔU_{cell} divided by the change in four important cell factors

Factor	Modified cell *	Standard cell *
ΔU_{cell}/change in current density J, mV m^2 kA^{-1}	450	450
ΔU_{cell}/change in cell temperature t_{cell}, mV/°C	−7.7	−10.1
ΔU_{cell}/change in brine concentration c_{NaCl}, mV L g^{-1}	−0.7	−1.8
ΔU_{cell}/change in cell liquor concentration c_{NaOH}, mV L g^{-1}	0.26	0.6

* OxyTech MDC-55 cell. The modified cell is outfitted with the Modified Diaphragm and expandable anodes, whereas the standard cell is outfitted with the standard asbestos diaphragm and the standard DSA anode. Conditions: anolyte temperature of 93 °C, anolyte NaCl concentration of 250 g/L, and catholyte NaOH concentration of 130 g/L.

Excessive brine impurities or other severe operating problems can adversely affect the voltage of the cell.

Power Consumption. The power consumption of a cell, kWh per tonne of Cl_2, may be calculated from the cell voltage by the following equation:

$$\text{Power consumption} = U_{\text{cell}} \times 756/\varepsilon$$

where

U_{cell} = cell voltage, V

ε = cell efficiency expressed as a decimal

Optimization. The relationships described in the preceding paragraphs can be used to determine the optimum economical cell operating conditions. The optimizations that must be considered are the following:

1) Higher cell liquor caustic strength and lower steam usage in caustic evaporation versus lower cell efficiency and higher power consumption
2) Lower current density, lower voltage, and lower power consumption versus additional cells and higher capital costs
3) Lower feed brine temperature, thus decreased steam usage for brine heating, versus higher cell voltage, lower efficiency, and higher power consumption
4) High brine pH and reduced acidification costs versus lower chlorine efficiency, higher power consumption, and lower product purity

Each diaphragm cell chlorine plant must determine its own optimum conditions for the most economical operations.

6.2. Diaphragm Cells

Electrolyzers for the production of chlorine and sodium hydroxide, including both diaphragm and membrane cells, are classified as either monopolar or bipolar. The designation does not refer to the electrochemical reactions that take place, which of course require two poles or electrodes for all cells, but to the electrolyzer construction or assembly. There are many more chlor-alkali production facilities with monopolar cells than with bipolar cells.

Bipolar electrolyzers have unit assemblies of the anode of one cell unit directly connected to the cathode of the next cell unit, thus minimizing intercell voltage loss. These units are assembled in series like a filter press, and therefore, the voltage of an electrolyzer is the sum of the individual cell voltages created by the anode of one unit, a diaphragm, and the cathode of the next unit. Bipolar electrolyzers have high voltages and relatively low amperage; therefore, the cost of electrical rectification is lower per unit of production capacity. Bipolar electrolyzers either must be installed in a large number of electrical circuits or be designed with very large individual cell components. Developers have chosen the option for large components.

Dow Chemical was the only early developer to have chlorine production needs large enough to consider the bipolar option [2]. Later, following the development of the DSA anode, PPG Industries and Oronzio De Nora Impianti Elettrochimici designed, and PPG Industries installed Glanor bipolar electrolyzers in a large complex at Lake Charles, Louisiana [12].

The *monopolar* electrolyzer is assembled so that the anodes and cathodes are in parallel. Therefore, the potential difference of all cells in the electrolyzer is the same, and the amperage at any particular current density only depends on the electrode surface area. A monopolar electrolyzer has low voltage and high amperage. The highest amperage rating of the most common modern monopolar cells is ca. 150 kA. Because a monopolar electrolyzer has a voltage of only 3 – 4 V, circuits of up to 200 electrolyzers have been constructed, producing 900 t of chlorine per day.

Diaphragms. The earliest asbestos diaphragms were made of sheets of asbestos paper. Asbestos was chosen because of its good chemical stability and its ion-exchange properties. Asbestos has been relatively inexpensive, since it is a relatively abundant natural material that was already being mined and processed for other industrial purposes, such as insulation.

The deposited asbestos diaphragm developed by Hooker Chemical in 1928 was the most common diaphragm until 1971, when what is now OxyTech Systems developed the Modified Diaphragm. The Modified Diaphragm is a mixture of asbestos and a fibrous fluorocarbon polymer [114]. The polymer stabilizes the asbestos,

which in itself lowers cell voltage and also allows for the use of the expandable DSA anode [115]. In its various formulations, the Modified Diaphragm is the most common diaphragm. The Modified Diaphragm still contains a minimum of 75 % asbestos.

Environmental concern over the use of asbestos has increased. France, Saudi Arabia, and Norway have banned the use of asbestos as a separator in chlorine cells. These nations allowed local chlorine producers several years to install non-asbestos replacement separators in existing diaphragm cells or to replace the cells themselves with membrane cells. Other European nations issue permits for the continued use of asbestos until an agreed upon BAT (best available technology) for diaphragm cells is established.

There is also concern in the chlor-alkali-industry for the future supply of asbestos, as most of the North American mines producing the grades of asbestos previously used for chlorine cell diaphragms have closed. At present the most common source of asbestos for chlorine cell diaphragms is Zimbabwe. In addition asbestos disposal costs and regulation have continued to increase. All of theses factors have led the chlor-alkali industry to consider non-asbestos diaphragm technology. Three non-asbestos diaphragm systems are commercially available today.

Chloralp Asbestos-Free Technology. Chloralp, a joint venture of Rhône-Poulenc and La Roche, has developed an asbestos-free diaphragm based on the built-in activation concept. The Chloralp separator is made of two vacuum-deposited layers:

1) The first layer, known as the precathode, is a conductive mat of carbon fibers, containing an electrocatalytic powder to decrease the cathode overpotential
2) The second layer is the diaphragm itself, in which PTFE and inorganic materials have replaced asbestos

The cathodic activation provides energy savings from 50 to 150 mV, depending on the plant operating conditions, and the catalyst content in the mat. Other benefits from the precathode are electrocatalytic destruction of chlorates and lower hydrogen contents in the chlorine after plant shutdowns. The diagraphm is deposited on the precathode by standard vacuum techniques. To control the diaphragm porosity, a pore-forming agent is incorporated in the slurry. Optimized porosity leads to lower cell voltage and higher current efficiency.

The technology is currently evaluated on OxyTech HC-3B electrolyzers at the Chloralp Pont-De-Claix facility. Compared to polymer-modified diaphragms, Chloralp asbestos-free separators provide a 150 mV saving on voltage and a 2 % improvement in current efficiency. Combined benefits from the precathode and the diaphragm, as well as an extended lifetime, could lead in the near future to full conversion of the 240 000 t/a Pont-De-Claix plant.

OxyTech Polyramix Diaphragm. OxyTech Systems has developed and commercialized a synthetic non-asbestos diaphragm called the Polyramix diaphragm. The Polyramix fiber is composed of a PTFE [poly(tetraflouroethylene)] fibrid base with zirconium oxide ceramic particles embedded in and on the fiber. The Polyramix diaphragm is vacuum deposited onto the cathode and then baked in an oven to fuse the fibers together. The process is very similar to OxyTech's widely used Modified Diaphragms. The Vulcan Chemicals plant in Geismar Louisiana was fully converted to the use of Polyramix diaphragms in 1993. Most major diaphragm cell chlorine plants have operated from 2 to 40 Polyramix diaphragm cells. The longest life cell with a Polyramix diaphragms has been in operation for over ten years [116].

PPG Industries Tephram Diaphragm. PPG Industries has developed and commercialized the Tephram diaphragm as its entry into the non-asbestos diaphragm market. Major goals of PPG's non-asbestos program were to produce a diaphragm that deposited and operated similarly to asbestos, with longer diaphragm life and lower power consumption. This technology utilized vacuum deposition to produce a base diaphragm composed of PTFE and a topcoat from a slurry of inorganic particulate materials. Dopants are priodically added to the anolyte during cell operation to adjust the diaphragm permeability to maintain or improve cell operation. This diaphragm has been successfully tested of the major monopolar cell technologies, as well

as on PPG's Glanor bipolar cells. An active program is in place to operate diaphragms on a trial basis at various producer sites, allowing for site-specific evaluation of the overall economics of the diaphragm.

6.2.1. Dow Cell [2, 117, 118]

The Dow Chemical Company is the largest chlor-alkali producer, accounting for one-third of the U.S. production and one-fifth of the world capacity. Because Dow's production capacity is large and concentrated in a few sites, Dow's cell development followed a different path than other chlor-alkali technology developers. Dow uses its own cell design of the filter press *bipolar* type. Dow has operated filter press cells for over 90 years. Dow cell development occurred in several stages, characterized by simple rugged construction and relatively inexpensive materials.

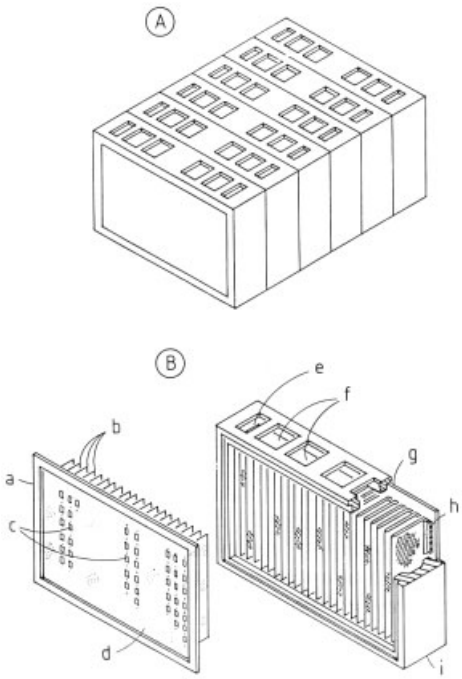

Figure 30. Dow diaphragm cell
A) Six-cell series
B) Internal cell parts: a) Cathode element; b) Cathode pocket elements; c) Copper spring clips; d) Perforated steel backplate; e) Brine inlet; f) Chlorine outlets; g) Copper backplate; h) Titanium backplate; i) Anode element

The cell employs vertical DSA coated titanium anodes, vertical cathodes of woven wire mesh bolted to a perforated steel backplate, and a vacuum-deposited modified asbestos diaphragm. A single bipolar element may have 100 m^2 of both anode and cathode active area. The anode of one element is connected to the cathode of the next by copper spring clips. This connection is immersed in the cell liquor during operation. Figures 30 B and 31 show these internal cell parts.

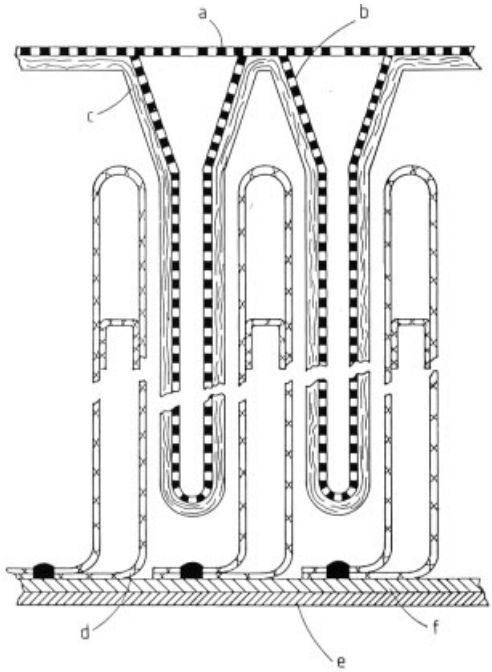

Figure 31. Dow diaphragm cell, section view
a) Perforated steel backplate; b) Cathode pocket; c) Asbestos diaphragm; d) DSA anode; e) Copper backplate; f) Titanium backplate

Dow operates at lower current densities than others in the chlor-alkali industry. The electrolyzers are normally operated with 50 or more cells in one unit or series. One electrical circuit may consist of only two of these electrolyzers. Figure 30 A shows a view of six electrolytic cells.

Treated saturated brine is fed to the anolyte compartment, where it percolates through the diaphragm into the catholyte chamber. The percolation rate is controlled by maintaining a level of anolyte to establish a positive, adjustable hy-

drostatic head. The optimum rate of brine flow usually results in the decomposition of ca. 50 % of the incoming NaCl, so that the cell liquor is a solution containing 8 – 12 wt % NaOH and 12 – 18 wt % NaCl.

The Dow diaphragm cell, optimized for low current density, consumes less electrical energy per unit of production than the rest of the industry. The cell voltage at these low current densities is only 300 – 400 mV above the decomposition potential of the cell. However, Dow has a larger investment in the electrolyzers, especially anodes.

The electrolyzers are operated at ca. 80 °C, lower than the 95 °C typical of other types of cells. This lower operating temperature allows cell construction with less expensive materials, such as vinyl ester resins and other plastics [117]. Operating data have not been published.

6.2.2. Glanor Electrolyzer [12, 119 – 121]

Glanor *bipolar* electrolyzers are a joint development of PPG Industries and Oronzio De Nora Impianti Elettrochimici S.p.A. The Glanor electrolyzer consists of several bipolar cells clamped between two end electrode assemblies by means of tie rods, thereby forming a filter press type electrolyzer (Fig. 32). The electrolyzer is equipped with DSA titanium anodes. Each electrolyzer normally consists of 11 or 12 cells. A lower number of cells can, however, be assembled in one electrolyzer. The Glanor electrolyzer was especially designed for large chlor-alkali plants.

The current is fed into the electrolyzer by means of anodic and cathodic end elements. The anodic compartment of each cell is connected to an independent brine feed tank by means of flanged connections.

Chlorine gas leaves each cell from the top through the brine feed tank and then passes to the cell room collection system. Hydrogen gas leaves from the top of the cathodic compartment of each cell, while the catholyte liquor leaves from the bottom through an adjustable level connection.

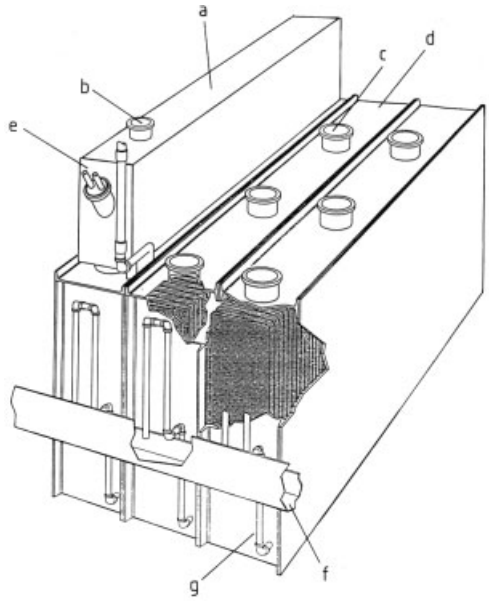

Figure 32. Glanor bipolar electrolyzer
a) Disengaging tank; b) Chlorine outlet; c) Hydrogen outlet; d) Bipolar element; e) Brine inlet; f) Cell liquor trough; g) Cell liquor outlet

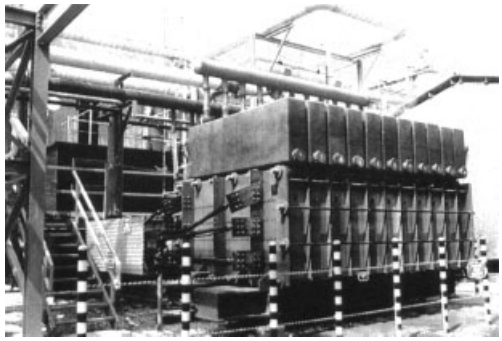

Figure 33. Glanor bipolar electrolyzer type V-1144

The V-1144 electrolyzer (Fig. 33) was the first commercial unit, and eight plants utilize this model. The second generation is the V-1161 electrolyzer, which employs Modified Diaphragms, narrower electrode gaps, lower current density, and DSA anodes to achieve lower power consumption than the V-1144 electrolyzer.

The operating characteristics of the Glanor electrolyzers are shown in Table 8.

Table 8. Glanor bipolar diaphragm electrolyzers: design and operating characteristics

Item	Model V-1144	Model V-1161
Cells per electrolyzer	11	11
Active anode area per cell, m^2	35	49
Electrode gap, mm	11	6
Current load, kA	72	72
Current density ϱ at 72 kA, kA/m^2	2.05	1.47
Cell voltage, V	3.50	3.08
Current efficiency, %	95–96	95–96
Power consumption (d.c.), kWh/t *	2500	2200
Anode gas composition (alkaline brine)		
Cl_2, %	97.3–98.0	97.0–98.0
O_2, %	1.5–2.2	1.5–2.2
H_2, %	<0.1	<0.1
CO_2, %	0.4	0.4
Cell liquor		
NaOH, g/L	135–145	135–145
$NaClO_3$, %	0.03–0.15	0.03–0.15
Production per electrolyzer		
Chlorine, t/d **	26.7	26.7
NaOH, t/d	29.8	29.8

* Per short ton of chlorine.
** Short tons.

6.2.3. OxyTech "Hooker" Cells [2, 12, 122, 123]

The first commercialized deposited asbestos diaphragm cell was the Hooker type S-1 monopolar cell, introduced in 1929. The basic design featured vertical graphite anode plates connected to a copper bus bar and a cathode with woven steel wire cloth or perforated steel fingers between the anodes. The cathode held vacuum-deposited asbestos fiber diaphragms that separated the anode and cathode compartments. The cathode fingers did not extend completely across the cell, but left a central circulation space. In the following 40 years, a family of S series cells with similar characteristics evolved, with over 12 000 having been installed in licensed plants.

In 1973, a new H series of *monopolar* cells was introduced. They incorporated the use of DSA anodes, which had been developed and commercialized in the late 1960s. These cells have significant voltage savings over the S series, thus allowing increases in cell capacity without corresponding increases in rectification capacity. The H series also incorporate cathode tubes with both ends open, extending across the cell, as the circulation space requirement was satisfied by the change from solid graphite anodes to the open DSA anodes (Fig. 34).

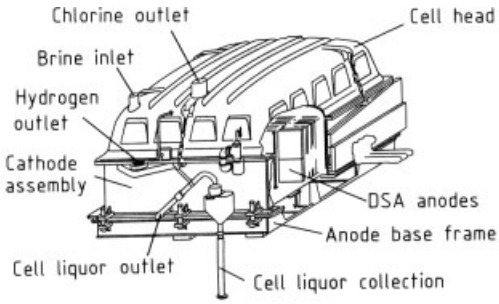

Figure 34. OxyTEch type H-4 cell

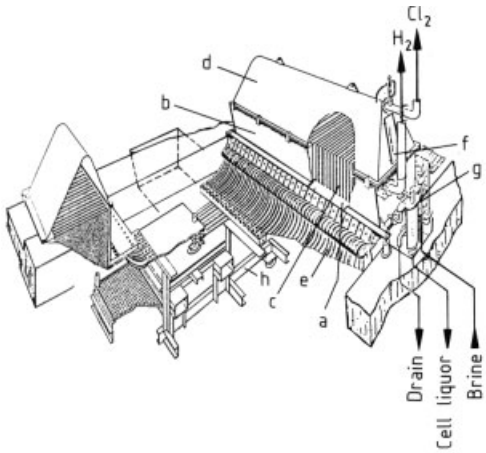

Figure 35. OxyTech/Uhde HU-type cells
a) Cell bottom; b) Cathode; c) Anode; d) Cell cover; e) Bus bars; f) Brine level gauge; g) Brine flow meter; h) Bypass switch

Table 9 is a summary of operating characteristics and current densities of the H-series cells currently available for license.

6.2.4. HU Monopolar Cells [123]

The HU type cells were a joint development of Hooker (now OxyTech Systems) and Uhde. The HU-type electrolyzer (Fig. 35) is rectangular, not cubic, and is narrow in the direction of current flow, since anodes are arranged in a sin-

Table 9. OxyTech Systems Hooker H-series diaphragm cells: design and operating characteristics

	H-2A		H-4	
Operating current, A	80 000		150 000	
Anode area, m^2	36.16		64.52	
in.^{2}	56 050		100 000	
Current density, A/m^2	2212		2325	
A/in.^{2}	1.43		1.50	
Cell voltage, V	3.44		3.44	
Approximate cell dimensions, m	1.87×2.66		2.58×3.11	
Diaphragm life, days	300–500		300–500	
Anode life, years	5–7	5–7		
Operating NaOH concentration, g/L	140	160	140	160
%	11.35	12.89	11.33	12.87
Current Efficiency, %	96.4	94.6	96.6	94.9
Chlorine output,				
metric ton/day	2.45	2.41	4.60	4.52
short ton/day	2.70	2.65	5.07	4.98
Caustic soda output,				
metric ton/day	2.76	2.71	5.19	5.10
short ton/day	3.05	2.99	5.72	5.62

Table 10. HU series diaphragm cells: design and operating characteristics

Item	Cell type						
	HU 24	HU 30	HU 36	HU 42	HU 48	HU 54	HU 60
Number of anodes	24	30	36	42	48	54	60
Anode surface area, m^2	20.6	25.8	31.0	36.1	41.3	46.4	51.6
Load, kA	30–45	40–60	50–70	55–85	60–95	70–105	80–120
Cl_2 production, t/d	0.90–1.36	1.19–1.82	1.49–2.12	1.64–2.58	1.79–2.88	2.09–3.18	2.39–3.64
NaOH (100 %) production, t/d	1.01–1.54	1.35–2.05	1.68–2.39	1.85–2.91	2.02–3.25	2.36–3.59	2.69–4.10
H_2 production, kg/d	25–39	34–52	42–60	47–73	51–82	59–91	68–103
Cell length, m	2.1	2.6	3.0	3.5	3.9	4.4	4.8
Distance, cell-to-cell, m	1.5	1.5	1.5	1.5	1.5	1.5	1.5

gle row. The cathode is long and narrow; consequently, the current density is lower through the cathode shell. The long, narrow cathode fabrication lends itself to closer anode–cathode tolerances and spacing. Copper on and around the cathode shell has been eliminated. Another advantage of the long, narrow design is shorter electrolysis current paths through the cell room, resulting in savings in piping and other materials. The HU-type cell incorporates a Modified Diaphragm.

A further novelty of the HU cell system is the design and arrangement of the bypass switch. The HU switch is installed underneath, not next to, the circuit of cells. This is accomplished by raising the cells from the floor, similar to mercury cells, creating a second operating floor. The interconnecting bus bars are flexible and are distributed over the entire length of the cell. The HU cell design incorporates a bus bar for each individual anode. This, as well as the elevation of the cell from the floor below, which allows access, enables connection of facilities for monitoring the current flowing through each anode. During operation of the bypass switch, connection is made for each individual anode, and no additional contact bus bars are required.

The HU-type cells are offered to cover 30–150 kA. All of the different cell types are equipped with cathodes and anodes of identical height and width. The only basic difference between the various cell models is the number of

elements and consequently the length of the cell (Table 10). Cell voltage and power consumption per tonne of chlorine, identical for all cell types, are shown in Table 11 for the specific current loads of 1.5 and 2.3 kA/m^2.

Table 11. HU series diaphragm cells: specific load, cell voltage, and power consumption

	Specific load, kA/m^2	
	1.5	2.3
Cell voltage, V	3.12	3.41
Power consumption (d.c., average), kWh/t*	2500	2700

* Per tonne of chlorine.

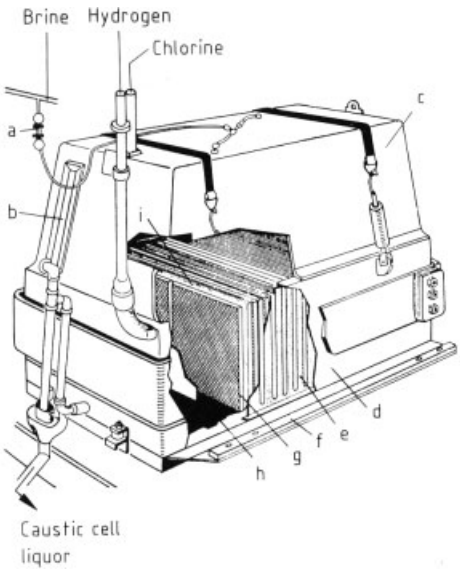

Figure 36. OxyTech Systems MDC cells
a) Brine feed rotometer; b) Head sight glass; c) Cell head; d) Cathode assembly; e) Tube sheet; f) Grid plate; g) Cathode tube; h) Grid protector; i) DSA expandable anode

6.2.5. OxyTech MDC Cells [12, 124]

OxyTech Systems manufactures and licenses the MDC series of *monopolar* diaphragm cells (Fig. 36). The MDC cells feature woven steel wire cathode screen tubes open at both ends, which are welded into thick steel tube sheets at each end. The tubes, tube sheets, and the outer steel cathode shell form the catholyte chamber of the cell (Fig. 37). Copper is bonded, rather than welded, to the rectangular cathode shell on the two long sides parallel to the tube sheets. Copper connectors attached at the ends of the bonded copper side plates complete the encompassing of the cathode with copper. Anodes are connected to a copper patented cell base, which is protected from the anolyte by a rubber cover or a titanium base cover (TIBAC) [125]. Orientation of the cathode tubes is parallel to the cell circuit, the opposite of a Hooker-type cell. This arrangement accommodates thermal expansion of the cell and circuit without changing the anode-to-cathode alignment.

The combination of the Modified Diaphragm and expandable DSA anodes reduces power consumption by 10 – 15 % from that of regular asbestos diaphragms and standard, fixed DSA anodes [126]. Table 12 presents performance data for the two most common MDC cell sizes [124]. The OxyTech MDC-29 is shown in Fig. 38. The licensed chlorine capacity of OxyTech cells now exceeds 20 000 t/d.

6.3. Operation

The process description in this section is intended to provide an overview of typical diaphragm cell process areas. A general block diagram for a diaphragm cell facility is shown in Figure 9. Included on the drawing are many process areas that may be optional, depending on the design of the plant and its end products. The operation of a cell room may be broken down into six areas: the two incoming systems, brine and electrical; the cells; and the three outgoing systems, chlorine, hydrogen, and cell liquor. Some of these are essentially the same for all three chlor-alkali processes and are described in Chapter 4 — the brine system (general), the electrical system, and the hydrogen system. The treatment of the chlorine is the subject of Chapter 11. Only aspects that are reasonably specific to the diaphragm cell process are described in this section.

6.3.1. Brine System

Most commonly, diaphragm cells are supplied with well brine on a once-through basis. The treated well brine flows to the treated brine storage tanks, which usually have 12-h capacity. From there the brine is fed to the cell room. The

Table 12. OxyTech Systems MDC cells: operating capacities and characteristics

Item	Model number and operating range, kA			
	MDC-29		MDC-55	
	35 to	80	75 to	150
Chlorine capacity,				
metric ton/day	1.05	2.41	2.33	4.53
short ton/day	1.16	2.66	2.48	5.00
Caustic capacity,				
metric ton/day	1.21	2.76	2.59	5.18
short ton/day	1.33	3.04	2.85	5.70
Hydrogen capacity,				
m^3/day	335	765	720	1435
cubic feet/day	11 830	27 010	25 420	50 670
Current density,				
kA/m^2	1.21	2.76	1.37	2.74
A/in.^{2}	0.78	1.78	0.88	1.76
Cell voltage, V [a]				
steel cathode	2.90	3.62	3.00	3.62
activated cathode	2.80	3.51	2.90	3.51
Power consumed (d.c., steel cathode) [b],				
kW h/t	2310	2876	2390	2870
kW h/short ton	2100	2610	2175	2610
Power consumed (d.c., activated cathode) [b],				
kW h/t	2230	2786	2310	2780
kW h/short ton	2025	2530	2100	2530
Diaphragm life, years	1 – 2	0.5 – 1.0	1 – 2	0.5 – 1.0
Anode life, years	8 – 10	5 – 8	8 – 10	5 – 8
Cathode life, years	10 – 15	10 – 15	10 – 15	10 – 15
Distance between cells [c],				
m	1.60		2.13	
inches	63		84	

[a] Cell voltage includes loss in intercell bus.
[b] Power consumed per ton (metric or short) of chlorine produced.
[c] Distance centerline-to-centerline and side-by-side with bus connecting.

flow to each individual electrolyzer is controlled by a rotameter. If the flow of brine to the cells is suddenly disrupted by failure of the brine feed pump, the rectifiers automatically shut down since an inadequate supply of brine to the cells is potentially dangerous. The specifications for brine for diaphragm cells are given in Table 13.

A brine recovery lagoon is usually available to handle any major upsets in the brine system. Brine sludges or out-of-spec brine can be sent to the lagoon. Supernatant clear brine can be recovered from the lagoon.

In most cases, operation with acidic brine is preferred because of the reduced amount of side-reaction products in the chlorine and the cell liquor.

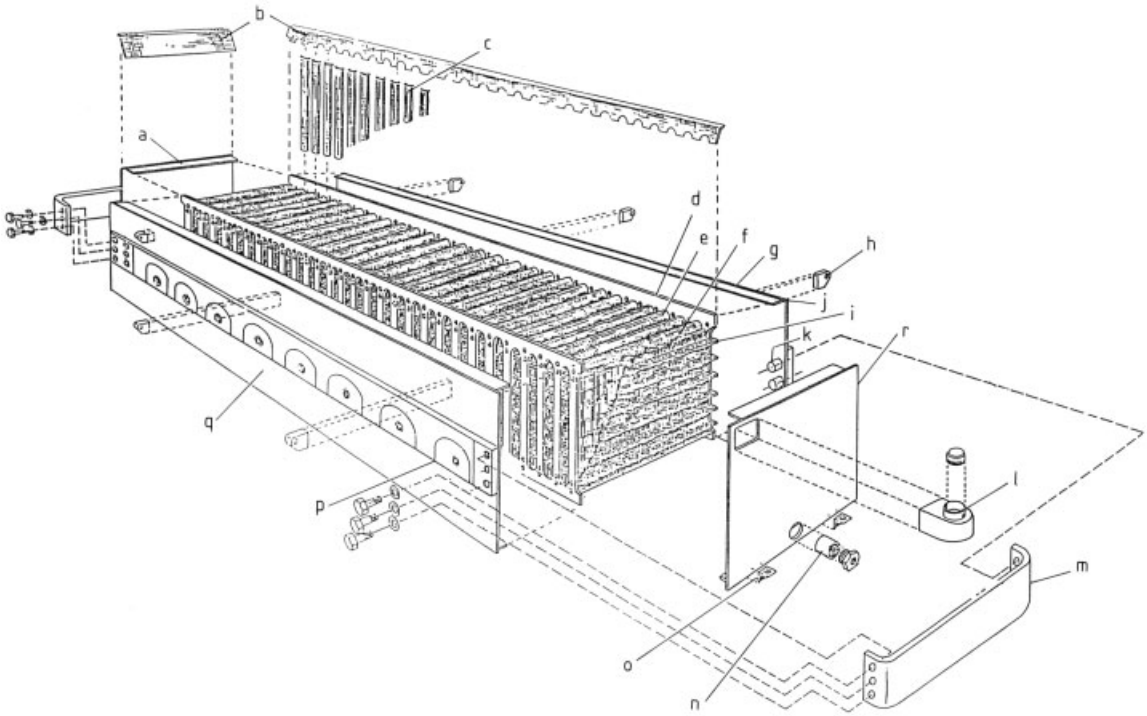

Figure 37. Exploded view of an OxyTech MDC-55 cathode
a) End plate; b) Rim screen; c) Side screens; d) Tube sheet; e) Full cathode tube; f) Half-cathode or end tube; g) Side plate; h) Lifting lug; i) Punched and coined stiffener strap; j) Bosses; k) End plate, operating aisle end; l) Hydrogen outlet; m) Connector bar; n) Caustic outlet; o) Clip angles; p) Grid bar, connector side; q) Side plate

6.3.2. Cell Room

Typical cell rooms are shown in Figures 39 (bipolar cells) and 40 (monopolar cells).

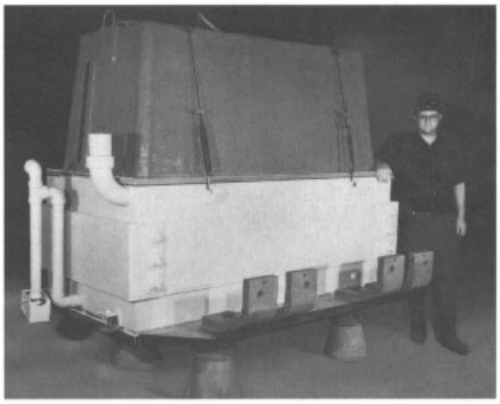

Figure 38. OxyTech MDC-29 with the author (1971)

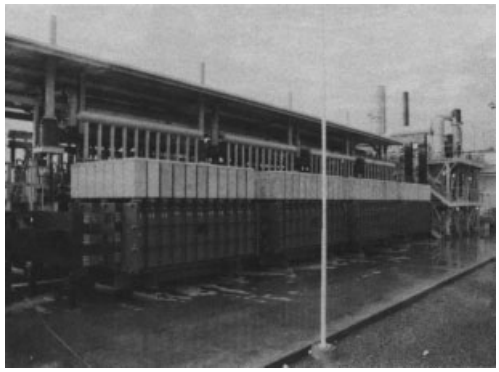

Figure 39. Cell room: bipolar PPG Industries/De Nora Glanor cells

A cell in normal operation requires little attention. The critical requirement is that the *brine flow rate* is sufficient to maintain an anolyte level above the cathode.

Table 13. Typical brine feed specifications for diaphragm cells

Parameter	Specification
NaCl	$\geqq$320 g/L
pH	2.5 – 3.5
Hardness ($Ca^{2+} + Mg^{2+}$)	<5 ppm
Magnesium	<0.4 ppm
Sodium sulfate (Na_2SO_4)	<5 g/L
Organics	<1 ppm
Manganese	<0.01 ppm
Barium	<0.01 ppm
Nickel	<0.1 ppm
Iron	<0.5 ppm
Silicon	<15 ppm
Cobalt	<0.02 ppm
Mercury	<1 ppm
Phosphate	<1 ppm

Figure 40. Cell room: monopolar OxyTech H-4 cells

Under no circumstances should a cell be operated with an inadequate or excessive anolyte level. Operation with the anolyte level not visible in a sight glass is unsafe. At least one operator should be in the cell room at all times. The cell room operator should inspect the anolyte level and brine flow to each cell at least once per hour. Any change in the anolyte level or brine flow rate should be investigated.

As the cell ages, the diaphragm will undergo changes in porosity because of the following:

1) Electrolysis effect
2) Brine impurities
3) Upsets in operation
4) Gradual wear of the diaphragm

A change in porosity may necessitate a change in brine flow rate. If the increase in porosity is severe, the cell may be replaced or doped with an asbestos slurry or inorganic salts.

Impurities in the brine often lead to decreased porosity. Decreased porosity can be offset to some extent by increasing the anolyte level and, if necessary, by lowering the catholyte level.

A cell operated with the anolyte level at the maximum value and the lowest catholyte level is called a sleeper. To gain additional diaphragm life after a cell has entered the sleeper position, the brine flow rate must be decreased below normal. This is not normally a recommended practice because current efficiencies of these cells are usually low.

For safe operation of diaphragm cells, the *header pressures* must be maintained at the proper values.

The chlorine header should be maintained at positive pressure to permit detection and correction of any chlorine piping leaks. The hydrogen header is also maintained at a positive pressure to avoid pulling air into the hydrogen, creating a potentially explosive mixture. The brine header pressure should be maintained to give the desired caustic concentration in the cell liquor. Normal practice is to adjust individual brine feed valves so that each cell receives the correct brine flow rate.

Load changes must be smooth to avoid fluctuations in the header pressures and detrimental effects on the diaphragms.

The brine feed rate to each cell should be increased to the new rate before circuit amperage is increased. The brine feed rate to each cell should be decreased immediately after amperage is decreased. During any period of operation when brine flow rates are being changed, extra attention should be given to the anolyte levels of the cells. Adjustment of the brine feed temperature may also be necessary when a load change occurs.

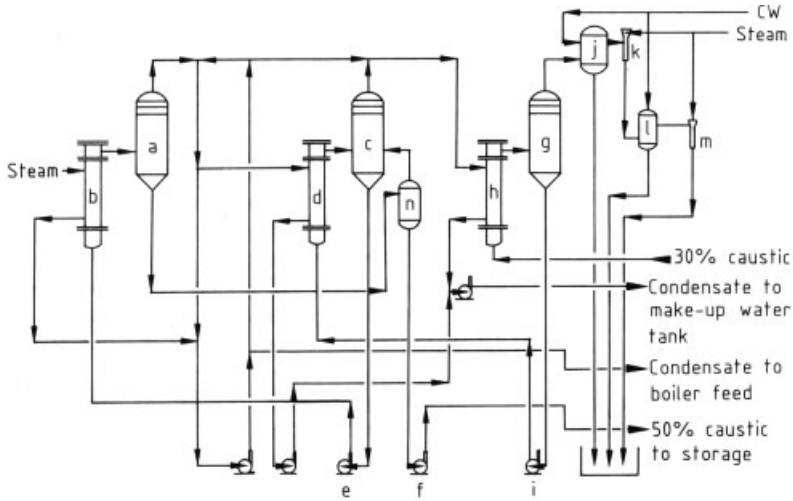

Figure 41. Process flow diagram: triple-effect caustic evaporator
a) First-effect vapor body; b) First-effect heat exchanger; c) Second-effect vapor body; d) Second-effect heat exchanger; e) Second-effect forwarding pumps; f) 50 % caustic transfer pumps; g) Third-effect vapor body; h) Third-effect heat exchanger; i) Third-effect forwarding pump; j) Barometric condenser; k) First-stage ejectors; l) Intercondenser; m) Second-stage ejector; n) Liquor flash tank

6.3.3. Diaphragm Aging

Of all the cell components, the diaphragm usually has the shortest life. The ability of a diaphragm to resist the back migration of hydroxide slowly becomes impaired with service life. The performance of the diaphragm deteriorates for the following reasons:

1) Chemical attack
2) Brine impurities
3) Unsteady operating conditions

The major reason for the deterioration is *chemical attack* on the asbestos by the alkaline catholyte and acidic anolyte. The rate of chemical attack can be minimized and diaphragm life maximized by careful operation of the cell. The most important situations to avoid are high concentrations of brine impurities and unsteady operating conditions. High *brine impurities* cause plugging of the diaphragm with insoluble hydroxides, which reduce the diaphragm's separation ability. The most common harmful impurities are calcium, magnesium, iron, nickel, silicates, aluminum, manganese, and barium.

Unsteady operation, such as electrical load changes, cell liquor strength changes, changes in brine concentration or pH, gas-pressure fluctuations, and shutdowns, change the pH of the various regions of the diaphragm, thus accelerating chemical attack on the asbestos. Diaphragm cell plant operators should strive to minimize these changes.

The real importance of the equations in Section 6.1 is as an aid in deciding when the diaphragms should be replaced.

6.3.4. Treatment of the Products

Chlorine. See Chapter 11.

Hydrogen. See Chapter 4.

Sodium Hydroxide Solution. The hydroxide produced at the cathode is associated with sodium ions and water to form a 10 – 12 wt % sodium hydroxide solution leaving the electrolytic cell. This cell liquor also contains 18 wt % unreacted sodium chloride.

Most large modern chlor-alkali plants have or will soon have an associated cogeneration power plant. In these facilities, the caustic evaporators are an important use for the byproduct steam.

Modern diaphragm cell plants use triple-effect evaporators and, in many cases, quadruple-effect evaporators.

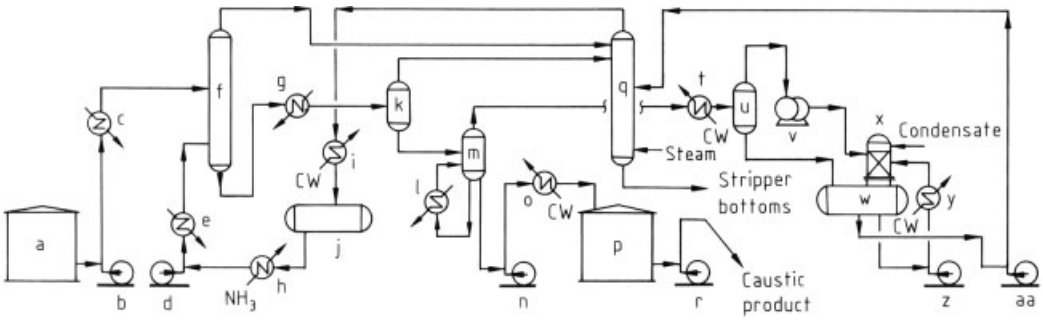

Figure 42. Caustic purification system
a) 50 % caustic feed tank; b) 50 % caustic feed pumps; c) Caustic feed preheater; d) Ammonia feed pumps; e) Ammonia feed preheater; f) Extractor; g) Trim heater; h) Ammonia subcooler; i) Stripper condenser; j) Anhydrous ammonia storage tank; k) Primary flash tank; l) Evaporator reboiler; m) Evaporator; n) Caustic product transfer pumps; o) Purified caustic product cooler; p) Purified caustic storage tank; q) Ammonia stripper; r) Purified caustic transfer pumps; t) Overheads condenser; u) Evaporator; v) Evaporator vacuum pump; w) Aqueous storage ammonia tank; x) Ammonia scrubber; y) Scrubber condenser; z) Ammonia recirculating pump; aa) Ammonia recycle pump

Caustic Soda Evaporation. A flow diagram for a typical triple-effect caustic soda evaporator is shown in Figure 41. The evaporator is of the backward-feed design and concentrates 10–11.3 wt % NaOH cell liquor to 50 wt % NaOH. Liquor flows from the third to the second to the first effect and from the first effect to the liquor flash tank. A cyclone is used for each effect to utilize the pressure drop across the circulating pump to clarify the transfer liquor.

The salt precipitated in the liquor flash tank is isolated from the rest of the salt precipitated in the evaporator and used as seed crystals in the cooling system to help diminish coil scaling and supersaturation of the product liquor with sodium chloride. The sodium chloride and triple salt ($NaOH-NaCl-Na_2SO_4$) precipitated in the liquor flash tank and cooling system is removed from the cooled product liquor with centrifuges. The salt precipitated in the three effects flows countercurrent to the liquor flow so that all of the salt is discharged from the last effect, the effect that has the coldest liquor and the lowest caustic soda concentration.

Two-stage steam-jet air ejectors with a common intercondenser are used to maintain vacuum in the evaporator. In the caustic cooling system, agitated tanks are used to cool the slurry discharged from the liquor flash tank. The slurry flows from the cooling system to the centrifuge feed tank from where it is pumped into centrifuges. Salt discharged from the centrifuges drops into the evaporator feed tank, where it is dissolved in cell liquor. The 50 wt % NaOH concentrate liquor, which flows by gravity to the pressure filter feed tank, contains ca. 1.0–1.5 wt % dissolved NaCl and ca. 0.1 wt % crystalline NaCl.

The liquor is pumped from the pressure filter feed tank into the pressure leaf filters, where the remaining traces of salt crystals are removed. The product caustic flows by gravity to the filtered product tank and then is pumped to storage. Salt removed in the pressure filters is reslurried with cell liquor and pumped to the evaporator feed tank via the filter backwash pump.

The salt discharged from the centrifuges drops into the leaching tank, where it is reslurried with condensate and recycled brine from the Glauber's salt ($Na_2SO_4 \cdot 10\,H_2O$) crystallizer. Concentrate from the pusher centrifuges flows by gravity into the evaporator feed tank.

The product salt is discharged from a cyclone into the salt reslurry tank. The overflow from the cyclone is returned to the leaching tank. The product salt is diluted with brine and pumped to the resaturator tank.

Brine containing the dissolved sodium sulfate is separated from the salt crystals in a cyclone. The underflow returns to the leaching tank. The overflow is collected in the feed tank for the Glauber's salt crystallizer. Sodium sulfate is crystallized from the liquor in a continuous vacuum cooled crystallizer.

Mother liquor removed from the crystallizer is pumped under level control to the brine tank. Slurry discharged from the crystallizer is thickened to ca. 50 wt %. The liquor from the thick-

ener is collected in the brine tank and pumped back to the leaching tank. The thickened slurry is redissolved in the Glauber's salt dissolving tank and pumped to a waste treatment system.

Caustic Purification [127]. Diaphragm-cell chlor-alkali producers requiring higher purity caustic than that produced by the diaphragm process can use caustic purification or DH process (Fig. 42). *Salt removal* in the purification unit is effected by contacting the 50 wt % caustic with anhydrous liquid ammonia under pressure sufficiently high to maintain all materials in the liquid state.

The liquid ammonia absorbs salt, chlorate, carbonate, water, and some caustic. It is then stripped, concentrated, and returned to the extraction process. The concentrated caustic leaving the extractor is stripped free of ammonia, which is recovered, concentrated, and recirculated. Typical purities before and after caustic purification are shown in Table 20. This process is offered for license by PPG Industries and OxyTech Systems.

In addition, producers and users of diaphragm-cell caustic may wish to *reduce metal impurities* by utilizing the porous cathode cell process (PPG Industries) [128]. The process consists of an electrolysis cell with porous nonmetallic cathodes. The caustic soda (50 wt %) is freed from iron, nickel, lead, and copper, which are deposited on the cathode. The cell must be regenerated periodically with water and hydrochloric acid. Typical feed and product analyses based on anhydrous NaOH are

Metal	Content, ppm	
	Feed	Product
Iron	10.0	2.0
Nickel	3.0	0.2
Lead	4.0	0.4
Copper	0.2	0.1

6.3.5. Measurement

Recorded data is an important tool for determining the operating condition of the plant and diagnosing problems.

The following should be recorded continuously or hourly:

ampere load on each circuit
voltage for each circuit
chlorine header pressure
hydrogen header pressure
brine header pressure or flow rate
brine temperature
brine pH
cell liquor temperature

Samples of brine should be taken every 4 h and combined into a daily composite. In addition, samples of cell liquor should be taken from each cell string, the sodium hydroxide content analyzed, and the temperature taken every 4 h. A daily composite should be made and samples should be analyzed by the laboratory for the following:

NaOH content
NaCl content
salt : caustic ratio
$NaClO_3$ content
NaOCl content
Fe content
average temperature
specific gravity at 25 °C

Chlorine gas from each cell circuit should be analyzed for chlorine and hydrogen content at least twice each 8-h shift. Each day a complete analysis of the chlorine header gas should be made. Additions or extensions to list may be dictated by plant operation.

Each plant must develop a procedure for taking *individual cell data* so that individual cells may be scheduled for renewal. The following is a minimum schedule:

weekly	voltage and cell liquor composition (NaOH, NaCl, $NaClO_3$)
monthly	chlorine composition (Cl_2, O_2, H_2, CO_2, N_2)

7. Membrane Process

In the membrane process, the anolyte and catholyte are separated by a cation-exchange membrane that selectively transmits sodium ions but supresses the migration of hydroxyl ions from the catholyte into the anolyte. A strong caustic soda solution with a very low sodium

chloride content can be obtained as the catholyte efflux. The advantages of the membrane process are its energy efficiency and its ability to produce caustic soda of high quality, with almost no impact on the environment. Depending on the particular design, membrane sizes from 0.2 to 5 m^2. The production capacity of an electrolyzer can be up to 90 t/d NaOH (100 %).

The process was started in the early 1970s with development of the perfluorosulfonate membrane Nafion by DuPont [129]. In 1975, a perfluorocarboxylate membrane capable of producing 35 wt % caustic soda became available, from Asahi Glass in Japan [130]. In 1978 the first two-layer membrane was developed, with low electrical resistance and high current efficiency [131].

The industrial success of the membrane process started in Japan, where the abolition of the mercury process on environmental grounds had been promoted by the government. Today the membrane process is the state of the art process for producing chlorine and caustic soda or potassium hydroxide.

The production capacity of chlor-alkali plants using the membrane process reached about 21 % of total world production capacity in 1995 and is predicted to increase to about 28 % in 2001 [132].

7.1. Principles

In a membrane cell a cation-exchange membrane separates the anolyte and catholyte, as shown in Figure 43. Saturated brine is fed into the anode compartment, where chlorine gas is evolved at the anode:

$$2\,Cl^- \rightarrow Cl_2 + 2\,e^-$$

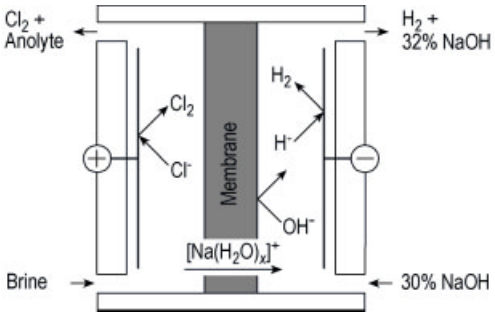

Figure 43. Principle of the membrane cell

The anolyte is discharged from the cell. The electric field causes hydrated sodium ions to migrate through the membrane into the catholyte. In the cathode compartment, hydrogen is evolved at the cathode, leaving hydroxyl ions, which together with permeating sodium ions constitute the caustic soda:

$$2\,H_2O + 2\,e^- \rightarrow H_2 + 2\,OH^-$$

$$Na^+ + OH^- \rightleftarrows NaOH$$

Liquid and gaseous phases anolyte/Cl_2 and catholyte/H_2 can be separated either in the cell compartment or downstream of the cell outlet. The chlorine-saturated anolyte is then treated in a dechlorination unit to recover the dissolved chlorine.

Membrane.

Structure. The membrane is exposed to chlorine and anolyte on one side and strong caustic solution on the other side at high temperature (90 °C). Only ion-exchange membranes made of perfluoropolymer can withstand such severe conditions. The ion-exchange groups of the original polymers are in the fluorosulfonate form, $-SO_3F$, or the carboxylate form, –COOR.

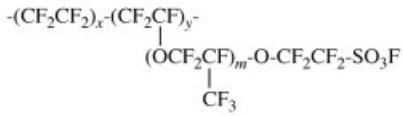

$m = 0 - 1$

Fluorosulfonate form

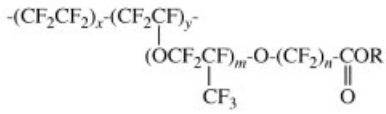

Carboxylate form $m = 0 - 1$ $n = 1 - 5$ $R =$ alkyl

The first membranes that showed significant potential for use in the chlor-alkali process were made of a perfluorosulfonate layer. These proved to be durable in chlor-alkali cells but were relatively inefficient. The perfluorocarboxylate polymers, with a lower water content,

showed higher selectivity but led to higher electrical resistance and high electrical power consumption [133]. Combining the advantages of high current efficiency and low electrical resistance a composite membrane (Fig. 44) was developed with a layer containing SO_3^- groups on the anode side and a layer containing COO^- groups on the cathode side [131, 134]. The perfluorosulfonate layer is thicker than the perfluorocarboxylate one and is the major constituent of the membrane [133].

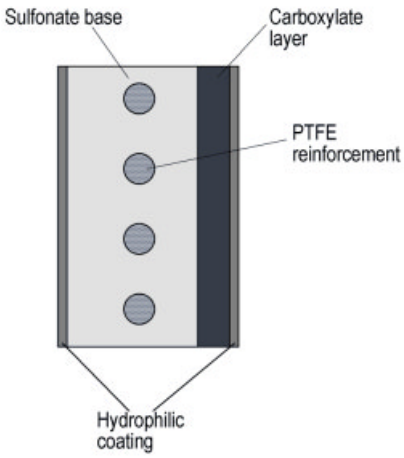

Figure 44. Membrane structure

Flux through the Membrane. The total flux through the membrane can be divided in to three parts [133]:

1) Migration due to electric field
2) Convection
3) Diffusion due to chemical gradients

Migration is the flux of ions through the membrane, driven by electric field. This includes the desired transfer of sodium ions to the cathode compartment and the undesired transfer of hydroxyl ions to the anode compartment. The capacity for selective separation of the cation exchange membrane is determined by its repulsive force for hydroxyl ions. This effect determines the current efficiency.

The backmigration of hydroxyl ions increases the formation of oxygen, hypochlorite, and chlorate in the anode compartment and causes a loss of current efficiency of 3 – 7 % in caustic soda production. The evolution of oxygen gas can be depressed by selecting an anode coating with suitable characteristics (see Section 8.1) or by decreasing the pH in the anode compartment by acidifying the inlet brine (Fig. 45).

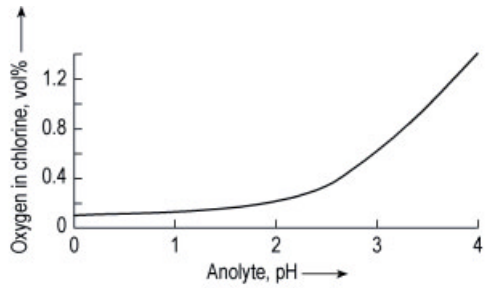

Figure 45. Oxygen content in chlorine [135]

Convection and diffusion determine the flow of uncharged compounds and ions through the membrane. Chloride anions in the catholyte are excluded by the cation-exchange membrane and repelled by the electric field, so that the transfer rate of chloride anions from the anolyte is extremly low. As a result, a caustic soda solution of about 32 – 35 wt % with a salt content of less than 20 ppm can be obtained. The water transport through the membrane is about 3.5 to 4.5 moles of water per mole of sodium ions, and can be regarded as the hydration sphere of the migrating sodium ions. Water flux increases with decreasing anolyte concentration [136].

Migration, convection and diffusion influence each other, and the resulting flux depends on membrane type, current density, temperature, and composition of anolyte and catholyte.

Cell Voltage. The cell voltage of a membrane cell is composed of the following terms:

1) Decomposition voltage
2) Membrane potential between anolyte and catholyte
3) Electrode overpotentials for chlorine and hydrogen
4) Ohmic drop in the membrane
5) Ohmic drop in the electrolytes
6) Ohmic drop in electrodes and conductors

Term 1. The decomposition voltage of the chlor-alkali process is about 2.20 V, depending on temperature, concentration, and pressure.

Term 2. This term describes the overpotentials at the surfaces of the membrane. Under standardized operating conditions (3 kA/m^2, 90 °C, 32 wt % caustic solution), the membrane potential is approximately 0.08 V.

Term 3. Titanium anodes coated with oxides of Ir, Ru or Pt are generally used in membrane cells and lead to a chlorine overvoltage of approximately 0.05 V at 3 kA/m^2. Hydrogen overpotentials of about 0.1 V at 3.0 kA/m^2 are attained with activated cathodes. Mainly nickel substrates are coated by painting and thermal treatment or galvanic deposition [137, 138]. Coating materials include Ni, Co, Ru and others (see Section 8.2).

Term 4. The ohmic drop of advanced commercial membranes under standardized operating conditions is about 0.25 – 0.30 V at 3 kA/m^2.

Term 5. To minimize the ohmic drop of an electrolyte, the gaps between the membrane and the electrodes are minimized in membrane cells. However, if the gap is very small, a rise in voltage is observed due to the entrapment of gas bubbles between the electrodes and the hydrophobic fluoropolymer membrane.

Term 6. The voltage losses in the electrolysis cell that occur due to unfavorable current paths along the metallic structure are reduced by an appropriate design. In modern chlor-alkali electrolysis cells the typical ohmic drop is 20 – 40 mV at 3 kA/m^2.

Current Efficiency. The current efficiency (CE) for caustic soda can be obtained either by directly measuring the quantity of caustic soda produced or by an anodic balance, i.e., the compositions of the anode gas and the anolyte with the following equation:

$$\mathrm{CE}\,(\%,\mathrm{NaOH}) = 100 - \eta\mathrm{O_2} - \eta\mathrm{ClO_3} - \eta\mathrm{ClO} + \eta\mathrm{NaOH} + \eta\mathrm{Na_2CO_3}$$

where ηO_2, ηClO, and ηClO_3 represent the loss of current efficiency due to the generation of oxygen hypochlorite, and chlorate, while $\eta NaOH$ and ηNa_2CO_3 take into account NaOH and Na_2CO_3 introduced in the feed brine.

The current efficiency is mainly dependent on membrane performance. On the cathode side the membrane is in contact with a concentrated NaOH solution, while the anode side has a pH of approximately 2 – 4. This leads to a pH gradient across the membrane cross section. The solubility of impurities, which are always present in the pure brine, depends on the pH. Therefore, depending on the type of impurities and on the pH, precipitation inside the membrane can take place. This leads to mechanical destruction of the membrane, which has a irreversible effect on current efficiency. In addition, the cell voltage rises due to the crystals formed inside the membrane.

7.2. Process Specific Aspects

The performance of a membrane cell depends on the following operating conditions:

1) Concentration of anolyte and catholyte
2) Current density
3) Temperature
4) Brine impurities

The optimum caustic strength depends on the composition of the membrane polymer. To achieve stable operation with high current efficiency, fluctuations in operating conditions or upsets must be avoided. Fluctuations in the caustic strength beyond the optimum range influence both the current efficiency and the cell voltage, as shown in Figure 46 [131]. Dilution of the anolyte caused by an upset in the brine feed also decreases the current efficiency. The sensitivity of membrane performance to operating conditions is attributed to changes in the water content of the membrane.

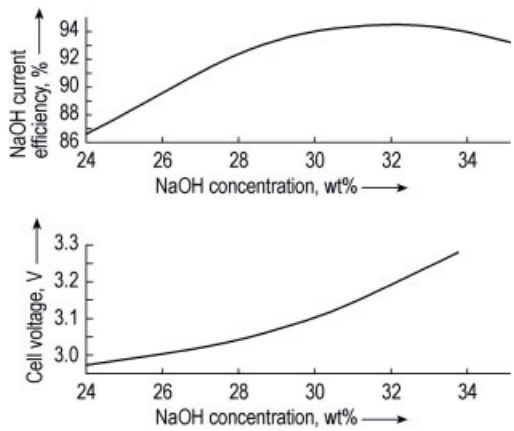

Figure 46. Dependence of cell voltage and current efficiency on NaOH concentration [139]

7.2.1. Brine Purification

The introduction of membrane technology into chlor-alkali electrolysis has dramatically increased the demands on brine purity [140]. The lifetime of chlor-alkali membrane cells is determined by the operating conditions and the quality and purity of the feed into the electrolyzers. Good long-term performance of the cells may be obtained if brine impurities are kept within the limits recommended in Table 14.

A major source of performance decline is the accumulation of solid material in the membrane [141]. Specific impurity levels are dependent on membrane design, cell design, operating conditions, the impurity itself and other impurities present. The prerequisite for long membrane life is to maintain low levels of, for example, Ca^{2+}, Mg^{2+}, Sr^{2+}, Ba^{2+}, Al^{3+}, SO_4^{2-} and SiO_2 in the brine. Traces of these impurities damage the membrane and/or electrodes and result in irrecoverable decreases in current efficiency and/or increased cell voltage. In the case of a closed brine loop with no purge, each impurity brought into or formed in the system must be removed to keep it below its specification level and to prevent accumulation.

The contaminants can be brought into the brine system by salt, by chemicals used in brine purification steps, by water for dissolving the salt, from materials of tanks, pipework, and cell components, or by the process itself [141]. The impurities in the salt depend upon the origin of the raw material. Rock salt, vacuum salt, sea salt, brine from well mining, or salt from waste incinerators serve as supplies of NaCl. The more varied the sources are, the more diverse the impurities.

Membrane and electrode damage effect cell performance, i.e., cause lower current efficiency, increased cell voltage, and, as a result, increased power consumption [142]. Some impurities affect the anode or cathode coating and cause an increase in overvoltage or simply deposit in the membrane, increasing its resistance and thus the cell voltage. The increase in voltage may in some cases be partially reversible when the impurity concentration drops to the recommended limits.

Current efficiency declines are strictly related to the membrane. Impurities lower the current efficiency by reducing the membrane's ability to reject anions, specifically the ability to prevent hydroxyl ions from migrating from the cathode compartment through the membrane to the anode compartment [143]. This is usually a result of physical damage caused by precipitation and crystallization of impurities inside the membrane. Impurities precipitate because the environment in the membrane changes from an acidic salt solution (pH 2 – 4) to a caustic solution (pH 14 – 15) over the 100 – 300 μm thickness of the membrane.

It is important to consider not only the impurities themselves but also their interaction. The presence of one impurity may not be harmful, but its synergistic combination with others may be [143]. For example, silica itself is not harmful for membranes. Only in the presence of calcium and aluminum do precipitates form and damage the membrane irreversibly. The concentration of silica and/or the concentration of aluminum plus calcium can be adjusted to give the optimum operating conditions. For example, with an effective secondary brine purification, higher levels of silica can be tolerated. Similarly, if aluminum concentration is high, calcium or silica concentration must be reduced to maintain acceptable membrane performance.

To meet the strict requirements on brine purity outlined in Table 14 brine treatment is generally carried out in the following main steps in the brine loop: saturation, precipitation, clarification, filtration, polishing filtration, ion exchange, electrolysis, chlorate decomposition and dechlorination.

Calcium and magnesium are precipitated and separated from the saturated brine with the insoluble materials. A ca. 10 wt % sodium carbonate and barium carbonate (barium chloride) solution and 32 wt % caustic soda are used as precipitants.

$$Ca^{2+} + Na_2CO_3 \rightarrow CaCO_3 + 2\,Na^+$$

$$SO_4^{2-} + BaCO_3 \rightarrow BaSO_4 + CO_3^{2-}$$

$$Mg^{2+} + 2\,NaOH \rightarrow Mg(OH)_2 + 2\,Na^+$$

Alternatively:

$$SO_4^{2-} + BaCl_2 \rightarrow BaSO_4 + 2\,Cl^-$$

Table 14. Impurities and their effects

Impurity	Source	Max. limit (w/w)	Reagents	Solubility in brine	Solubility in caustic	Mechanism	Damage	Negative effect on performance Voltage increase An.	Cath.	Mem.	CE	PQu	Methods of control
Ca^{2+}	salt	Ca^{2+}+Mg^{2+}	Ca^{2+}; OH^-	X	-	precipitation near the cathode	physical disruption of the membrane				++		precipitation with Na_2CO_3 plus ion exchange
		< 20 ppb	Ca^{2+}; SO_3^{2-}	X	-	side of the membrane, formation					++		
			Ca^{2+}; SO_3^{2-}; AL^{3+}	X	-	of large crystals	(blister formation)				++		
			Ca^{2+}; I^-	-	-	very fine precipitation in the membrane	-						
Mg^{2+}	salt	Ca^{2+}+Mg^{2+} < 20 ppb	Mg^{2+}; OH^-	X	-	fine precipitation near the anode side of the membrane	binding of ion-exchange sites in the membrane			++			precipitation with NaOH plus ion exchange
Sr^{2+}	salt	< 4 ppm	Sr^{2+}; OH^-	X	X/-	precipitation on the cathode side of the membrane, formation of crystals	physical disruption of the membrane			+	++		coprecipitation with Na_2CO_3 plus ion exchange
Ba^{2+}	salt	< 0.5 ppm	Ba^{2+}; OH^-	X	X								purge;
			Ba^{2+}; I^-	X	-	very fine precipitation in the membrane	minor damage on the membrane, minor interaction with ion-exchange sites			+	+		precipitation with $NaHSO_3$ plus ion exchange
			Ba^{2+}; SO_4^{2-}			coating of the anode		+					
Al^{3+}	salt	< 0.1 ppm	Al^{3+}	X	X		-						precipitation as hydroxide at pH 7 – 9, ion exchange under acid conditions
			Al^{3+}; SiO_3^{2-}	X	-	formation of crystals (zeolites, sodalites, faujacites) near the cathode side of the membrane	disruption of the membrane				++		
			SiO^{3-}; Ca^{2+}	X	-	precipitation near cathode	disruption of the membrane				++		Purge
			SiO^{3-}; Al^{3+}	X	-	side of the membrane and crystalization							
Hg^{2+}	parallel operation of an amalgam plant	< 0.2 ppm heavy metals	Hg^{2+}	X	X	deposition on the cathode	partially reversible, covering of active cathode coating		++				precipitation with Na_2S

Table 14. Continued

Impurity	Source	Max. limit (w/w)	Reagents	Solubility in brine	Solubility in caustic	Mechanism	Damage	Negative effect on performance Voltage increase An.	Cath.	Mem.	CE	PQu	Methods of control
Fe^{3+}	salt, pipework, tank material, anti-caking agent	< 0.1 ppm	Fe^{3+}	X	X	deposition on the cathode (in extreme cases: dendritic growth from cathode toward the anode)	covering of active coating (puncturing of the membrane)		++				precipitation with NaOH
Ni^{2+}	salta, pipework, tank material, cathode	< 0.2 ppm heavy metals	Ni^{2+}; OH^-	X	X	absorption of Ni in the membrane, deposition on the cathode			++	+			ion-exchange, purge
I^-	salt	< 0.2 ppm	I^-; Na^+	X	-	precipitation on the cathode side of the membrane, formation of crystals	physical disruption of the membrane				++		purge
			I^-; Ca^{2+}	-	-	very fine precipitation in the membrane	-						
			I^-; K^+	X	X		-						
F^-	salt	< 0.5 ppm		X			destruction of the anode coating	++					purge
SO_4^{2-}	salt, dechlorination with $NaHSO_3$	< 4 – 8 g/l Na_2SO_4	SO_4^{2-}; Na^+	X	-	precipitation near the cathode surface of the membrane	reduction of the OH^- ion rejection capability				++		purge, precipitation with $BaCO_3$ or $BaCl_2$ plus ion exchange
			SO_4^{2-}; Ba^+			coating of the anode		+					
CO_3^{2-}	salt, precipitation with Na_2CO_3 or $BaCO_3$	< 0.4 g/l Na_2CO_3	CO_3^{2-}; H_2O			formation of CO_2						++	
ClO_3^-	process, side reactions	< 10 g/l $NaClO_3$				chlorination of ion-exchange resin	destruction of ion-exchaange resin						chlorate decomposition by acidification
TOC	salt	< 1 ppm				increased foaming, overplating		+	+	+			filtration
$K_4(Fe(CN)_6)$	anti-caking-agent for salt	< 0.1 ppm Fe^{3+}	Fe^{3+}	X		as Fe^{3+}			++				oxidation with active chlorine plus precipitation with NaOH
						evolution of N_2						++	

∗ An. = Anode; Cath. = Cathode; Mem. = Membrane; CE = current efficiency; PQu = product quality.

From the precipitation tank, the brine is fed into the clarifier, where a defined quantity of flocculant is added to promote the settling of the precipitated solids and gels. The brine is then pumped to a filtration system followed by an ion exchange purification.

Additionally, if the brine circulating system of an existing mercury cell plant also serves membrane cells, all mercury must be removed in a chemical treatment facility. The brine from the primary filtration is acidified with hydrochloric acid to pH 2.0 – 2.5 and sodium sulfide is added to precipitate mercury sulfide. Subsequently, the brine is filtered, alkalized to pH 9.5 – 11 by adding caustic soda, and finally fed to the secondary purification section.

The content of calcium and magnesium must not exceed 20 ppb. Such low contents can be achieved by using ion-exchange columns. The polished brine is pumped at approximately 70 °C to an ion-exchange system with two resins beds operating in series according to the lead/lag principle. When the leading ion exchanger is exhausted it is put to the regeneration and conditioning mode, while the lagging one takes over the lead position. After treatment in this secondary purification step, the purity limits are met and the brine is fed to the membrane cells. In another secondary purification system two columns operate in series while the third is in regeneration mode. When the first column is exhausted, the regenerated column is put in second position.

Prior to the resaturation, the byproduct chlorate and dissolved chlorine must be eliminated from the anolyte. Chlorate concentration is controlled by acidification of a partial stream of anolyte with an excess of hydrochloric acid [143]. Chlorine is removed under vacuum followed by addition of sodium bisulfite and hydroxide.

$$NaClO_3 + 6\,HCl \rightarrow NaCl + 3\,Cl_2 + 3\,H_2O$$

$$2\,Cl_2 + 2\,NaHSO_3 + 6NaOH \rightarrow 4NaCl + 4\,H_2O + 2\,Na_2SO_4$$

All other impurities not precipitated, filtered out, or extracted by the ion exchangers can only be controlled by purging a partial stream of the anolyte to avoid accumulation. Resaturation then closes the loop.

7.2.2. Commercial Membranes

The ion-exchange membrane is the key component of the membrane cell. The energy consumption and the quality of the products depend on membrane performance. Requirements for the membrane are as follows:

1) Durability under the conditions of chlor-alkali electrolysis
2) High selectivity for sodium ion transport
3) Low electrical resistance
4) Sufficient mechanical strength for practical use
5) Low sensitivity to changing operating conditions

The importance of 1 – 3 is described in Section 7.1. High mechanical strength is necessary for installing the membrane and during service life, in which the membrane has to cope with deviations in temperature, concentration, and pressure.

As the performance of the membrane is the most important element in the economy of a membrane cell, many refinements have been made in membrane manufacturing. To reduce the current screening due to fabrics, membranes reinforced with dispersed microfibers and interwoven fabrics made of electrolyte-soluble fibers and PTFE have been developed [144, 145].

The improvement of hydrophilicity by covering the surface on the cathode side or on both sides with a nonconductive inorganic material brought about a significant reduction in the cell voltage. The surfaces of the membrane are covered with thin layers of a porous inorganic material. This material is an oxide, hydroxide, or carbide of the metals of groups 4, 5 and 6 or the iron triad (Fe, Co, Ni) [138, 146].

Figure 47 illustrates the effect of hydrophilic cathode surface modification. The surface-modified membrane (Type B) has a lower cell voltage than the conventional membrane (Type A). Tthe voltage of the surface-modified membrane decreases linearly with decreasing gap size. With these advanced membranes, so-called zero-gap cells have been made possible, and the ohmic loss in electrolytes has been reduced to a minimum.

The active life of a membrane is determined by the economic balance between membrane cost and energy cost in use [148].

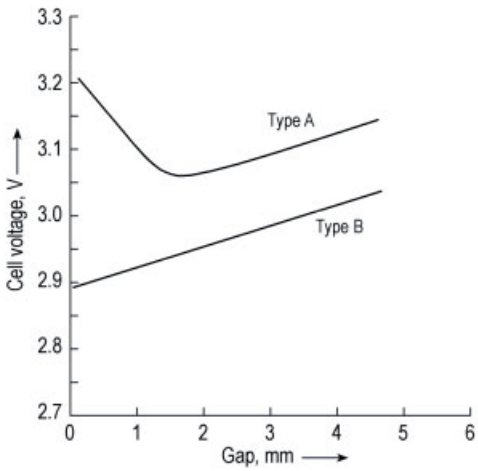

Figure 47. Effect of cathodic surface modification

The performance of membranes depends on the operating conditions, especially on the caustic strength of the solution (Figure 46). Commercially available membranes, delivered by Asahi Chemical, Asahi Glass and Du Pont, are designated for use in a specific strength of caustic. For economic production the selection of the appropriate membrane is essential. Table 15 gives an overview of the most widely used membranes. Most membranes are operated in the narrow- or zero-gap configuration to minimize power consumption.

7.2.3. Power Consumption

For monitoring cell performance and comparing different electrolyzer designs, the electric power required to produce one tonne of NaOH 100 % is considered. This figure is determined by the voltage drop over one cell and the NaOH current efficiency.

$$\frac{\text{DC Energy}}{100\%\ \text{NaOH produced}} = \frac{UIt}{M} = \frac{U}{F \cdot \text{CE}} = \left[\frac{\text{kWh}}{\text{t}}\right]$$

where U is the cell voltage (V), F the Faraday constant for NaOH (1.4923 kg/kAh), and CE the NaOH current efficiency (%).

The specific power consumption is the main indicator for economic plant operation, and continuous efforts are made to lower the voltage and increase the current efficiency. At a thermodynamic minimum the decomposition voltage of about 2.2 V limits the theoretical minimum energy requirement to about 1480 kWh/t 100 % NaOH, as shown in Figure 48. At practical current densities of 3.0 – 5.5 kA/m^2 for present-day commercial cells and membranes, power consumption measured at the electrolyzer terminals is in the range of 1950 to 2180 kWh/t 100 % NaOH dependent on the selected current density (anolyte/catholyte temperature 90 °C, NaOH concentration 32 wt %, NaCl concentration in the anolyte 220 g/L). The power consumption rises with increasing operating time due to aging effects, such as decreasing current efficiency and increasing voltage. Investment costs rise when operating at low current densities, as more cells are needed to meet production. Hence electrolyzers are operated at low current densities in countries with high energy prices, and at high current densities in countries with low energy prices.

7.2.4. Product Quality

The caustic soda solution has a concentration of up to 32 ± 1 wt % NaOH. If a NaOH concentration of 50 wt % is required, evaporation can be used. The typical NaCl content is 20 ppm in a 32 wt % caustic solution.

The hydrogen has almost synthesis quality with a concentration of about 99.9 vol % H_2 (dry basis). The chlorine has an oxygen content of about 1.5 vol % (dry basis). Chlorine with an oxygen content below 0.6 vol % (dry basis) can be obtained by acidifying the brine with hydrochloric acid.

7.3. Membrane Cells

7.3.1. Monopolar and Bipolar Designs

A commercial membrane plant has multiple cell elements combined into a single unit, called the

Table 15. Commercial membranes

		Available since	Tear strength, kg	Tensile strength, kg/cm	Ohmic drop (at 3 kA/m^2, $T = 90\,^\circ$C, $c = 32$ wt %), V	Caustic strength, wt %
Asahi Chem	F5201	1991	4	5.5	[a]	33 – 36
Aciplex	F4202	1993	4	5.5	[a]	30 – 34
	F4203	1997	4	5.5	[a]	30 – 34
	F890	1989	4.5	6	0.35	31.5 – 32.5
Asahi Glass	F892/old	1990	4.5	5	0.27	30 – 35
Flemion	F892/new	1994	4.5	5	0.28	31.5 – 33.5
	F893/new	1994	4.5	5	0.26	31.0 – 32.5
	N90209	1984	2.5	5.6	0.35	30 – 35
DuPont	N966	1988	5.5	7.3	[a]	30 – 35
Nafion	N981	1996		3.3	[a]	30 – 35

[a] Not published.

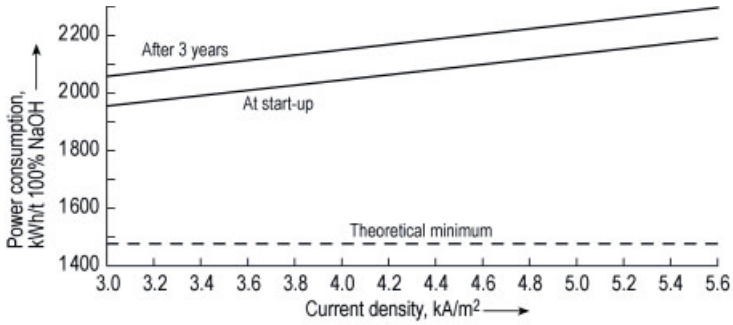

Figure 48. Specific power consumption

electrolyzer. The electrolyzers follow two basic designs: monopolar and bipolar [147].

In a bipolar arrangement the elements are connected in series with resultant low current and high voltage. The cathode of a cell is connected directly to the anode of the adjacent cell, as shown in Figure 49. The operation of a bipolar electrolyzer can be easily monitored by measurement of element voltages. If element upsets occur, a safety interlock system actuates the breakers (short-circuiting switches) and isolates the electrolyzer from the electric circuit. As the influx and efflux of electrolytes for the cells with different electric potential are gathered in common headers, problems of stray current may arise.

In the monopolar type all anodes and cathodes are connected in parallel, forming an electrolyzer with high current and low voltage (Figure 50). Due to the long current path, the voltage drop is high and can only be reduced by minimizing the size of cells or introducing internal copper conductors to lower the resistance. Because of this basic principle, ohmic losses in the monopolar cells are 80 – 100 kWh per tonne 100 % NaOH, which is much higher than in equivalent bipolar cells. Furthermore, the bipolar safety system is not applicable to the monopolar design, since the cell elements are arranged in parallel, which does not permit the monitoring of deviations in individual cell voltages.

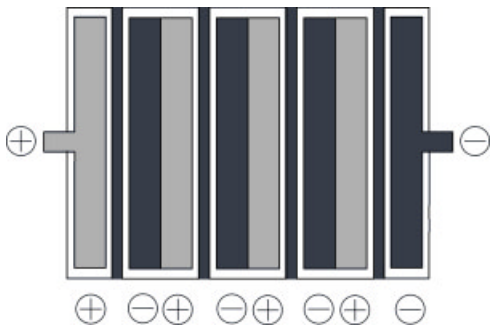

Figure 49. Bipolar electrolyzer

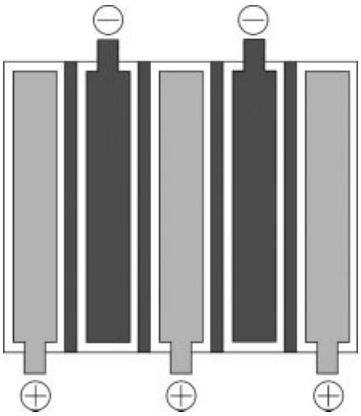

Figure 50. Monopolar electrolyzer

Multiple electrolyzers are employed in a single d.c. circuit (Fig. 51). Usually bipolar electrolyzers are connected in parallel with low current and high voltage. Monopolar electrolyzers are often connected in series, resulting in a high current circuit and low voltage. Though both principles still appear on the market, investment and operating cost considerations, such as for the rectifier system, the cell room space required, for piping, valves, instrumentation, busbars and switches, significantly favor the bipolar design.

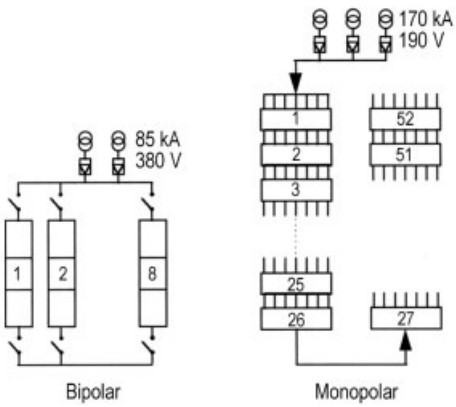

Figure 51. Electrolyzer architecture

7.3.2. Commercial Electrolyzers

Generally, membranes are clamped vertically between the meshlike metal anodes and cathodes. The effective membrane area of a cell ranges from 0.2 to 5.0 m^2. Current density varies between 1.5 and 7 kA/m^2.

The cells are filled with electrolytes, and gas-separating means are provided outside the cells. Many cells generally stacked like a filter press, constitute one electrolyzer with high production capacity.

The performance of a plant is determined by the electrolyzer, the cell voltages, and the current efficiency of the membrane. It is essential to design an electrolyzer with an homogeneous electrolyte concentration, temperature, and current density distribution across the whole area of the membrane.

The construction materials of the cell are selected to withstand the corrosive electrolytes. In most electrolyzers, titanium and nickel are used for the anode and cathode compartments of the cell. In older electrolyzers, stainless steel is used on the cathode side.

For economic and environmental reasons, mercury and diaphragm plants are increasingly being converted to membrane electrolyzers. The existing facilities, such as rectifiers, equipment for brine purification, and equipment for product treatment are utilized as much as possible.

Asahi Kasei ACILYZER-ML/NC Electrolyzer. The Asahi bipolar electrolyzer (Fig. 52) is of the filter-press type. The bipolar cell frames are suspended in a steel frame and compressed by a hydraulic device. Each cell frame consists of an anode and cathode compartment separated by a partition wall. The anode compartment is made of titanium, and the cathode compartment consists of special stainless steel and nickel. The anode and cathode structures are spot welded onto ribs in each compartment. Each compartment has an inlet nozzle for electrolytes at the bottom and an outlet nozzle for gas and electrolyte on top, connected to the gas/liquid separation chamber.

Two types of cell frames are available: frames with forced circulation of electrolytes by pumps, and frames with natural circulation in each compartment by means of a special arrangement of integrated ducts. The current is connected to the first and last element by flexible busbars.

CEC BITAC 800 Electrolyzer. The Chemical Engineers Corporation (CEC) bipolar BITAC electrolyzer was jointly developed with

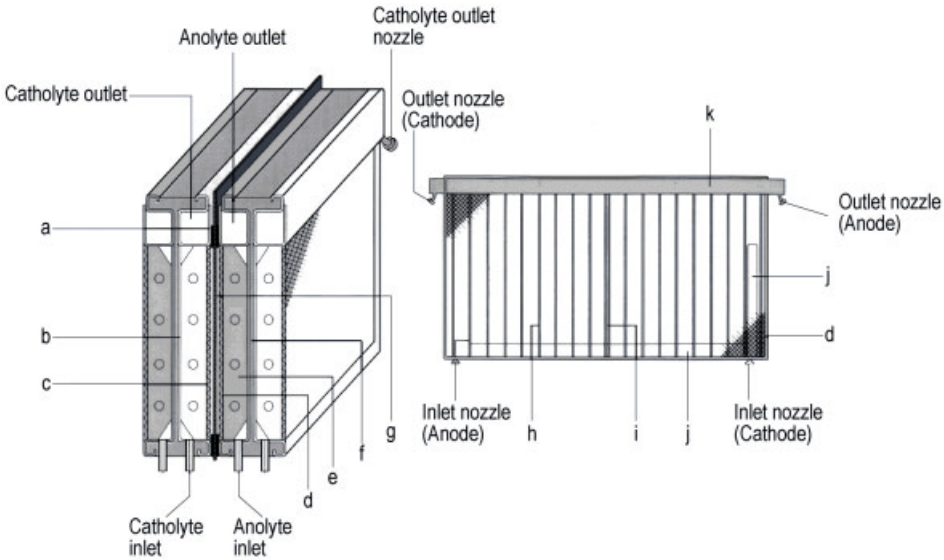

Figure 52. Cell structure of ACILYZER ML32NC
a) Gasket; b) Nickel; c) Cathode; d) Anode; e) Titanium; f) Partition wall; g) Membrane; h) Rib; i) Reinforcing rib; j) Duct; k) Gas – liquid separation chamber

Tosoh Corporation. The design follows the filter-press principle. Up to 80 bipolar electrode frames are clamped together by end plates and spring-loaded tie rods. The frames are made of special titanium alloy for the anode and nickel for the cathode. The electric current flows along the nickel pans, since the electrical conductivity of nickel is six times higher than that of titanium. Gas and electrolytes leave the cell compartment in overflow mode with little pressure fluctuation. Transparent PTFE tubes are attached at the electrolyte inlet and outlet nozzles of each element. Anolyte recirculation takes place through an external loop.

CEC CME DCM 400 Electrolyzer. The CME monopolar electrolyzer consists of large elements compressed in a filter-press arrangement. The electric current travels into each anode element through conductor rods and current distributors. This design achieves uniform current distribution over the large electrode area. The current distributors serve the additional role of a downcomer pipe, which creates a natural circulation within the cell, providing a uniformly distributed electrolyte concentration as well as good gas release. The anode frames are constructed from titanium and the cathode frames from a special stainless steel. The rods are cladded with titanium and stainless steel. Inlet and outlet tubes for liquids and gases and are made of transparent PTFE.

Uhde BM 2.7 Electrolyzer. The bipolar Uhde electrolyzer (Fig. 53) is a single-element concept. Each element comprises anode and cathode half-shells, electrodes, a membrane, flanges, and the sealing system. This enables long-term storage of pre-assembled and fully tested elements. The electrodes are attached with continuous laser weld to the current transfer and support blades and hence to the half-shells. The anode is made of titanium and the cathode of nickel. The individual cell elements of an electrolyzer are suspended in a steel frame in which they are lightly pressed together for electrical contact. Large sealing forces are not required in the single-element concept, as each element is a separate, stand-alone electrolysis cell. The feed and discharge lines of the cell are located underneath the cells and connected to the catholyte and anolyte headers. The area above the electrolyzer is free of piping or bus bars, simplifying access and eliminating the risk of leakage and associated corrosion problems. The current is conducted from cell to cell by continuously laser-welded, explosion-bonded titanium – nickel contact strips on the anode half-shell.

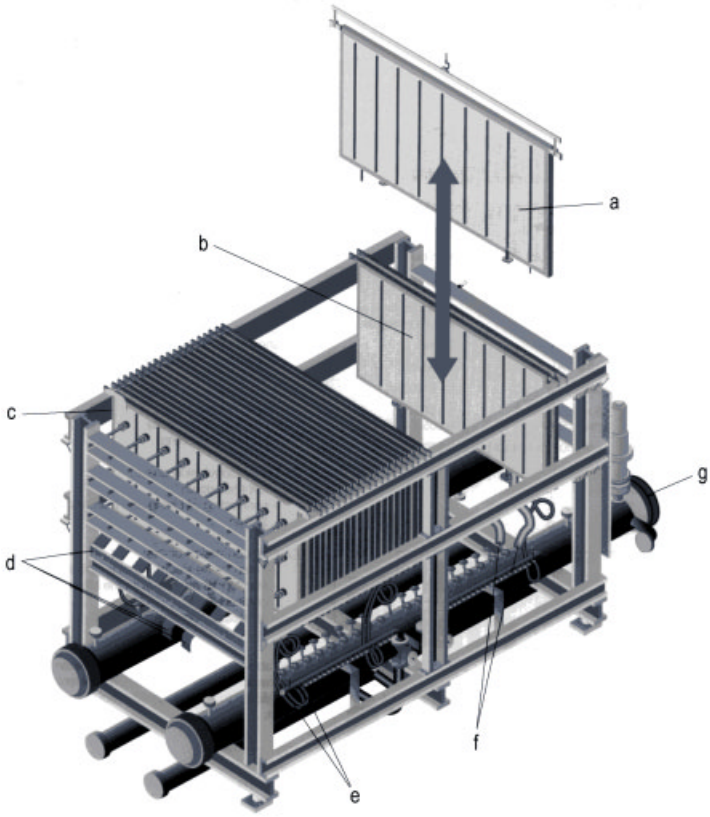

Figure 53. Uhde BM 2.7 electrolyzer
a) Single element; b) Contact strip; c) Cell rack; d) Busbars; e) Inlet hoses; f) Outlet hoses; g) Header

The brine and caustic soda feeds enter the cell at the bottom, and the product streams are discharged downwards through internal overflow pipes. The internal baffle plate at the top of the anode half-shell prevents gas-phase blistering of the membrane. The chlorine gas is effectively removed from the membrane, preventing contact and improving the inherent safety of the electrolyzer. Natural circulation around a downcomer plate and a distribution pipe for brine and caustic achieve homogeneous temperature and concentration profiles within the element and assist in achieving uniform current distribution.

EL-Tech ExL^{B} Bipolar Electrolyzer. The ExL^{B} bipolar electrolyzer (Fig. 54) is basically of the same design as the EL-Tech ExL^{M} electrolyzer. Instead of the copper distributors with interface material, which provide the parallel (monopolar) arrangement in the electric circuit, the elements are connected in series (bipolar), omitting the copper distributors and simply pressing the nickel cathode pan onto the nickel-plated back of the anode pan. The integral feed and discharge manifolds are designed to avoid current leakage.

EL-Tech ExL^{M} Monopolar Electrolyzer. The ExL^{M} monopolar electrolyzer is an improved version of the MGC electrolyzer, which has been in service for more than 15 years. The elements are sealed with O-rings in a staggered gasket design. The cathode O-ring is located closer to the liquid than the anode O-ring. This protects the anode O-ring from the chlorination degradation, making it a long-life back up seal. The elements are pressed together by tie rods with copper distributor plates and conductive interface material to provide good current distribution. Electrolytes and gases are fed and discharged to and from the elements through the manifold passage attached to the cell ele-

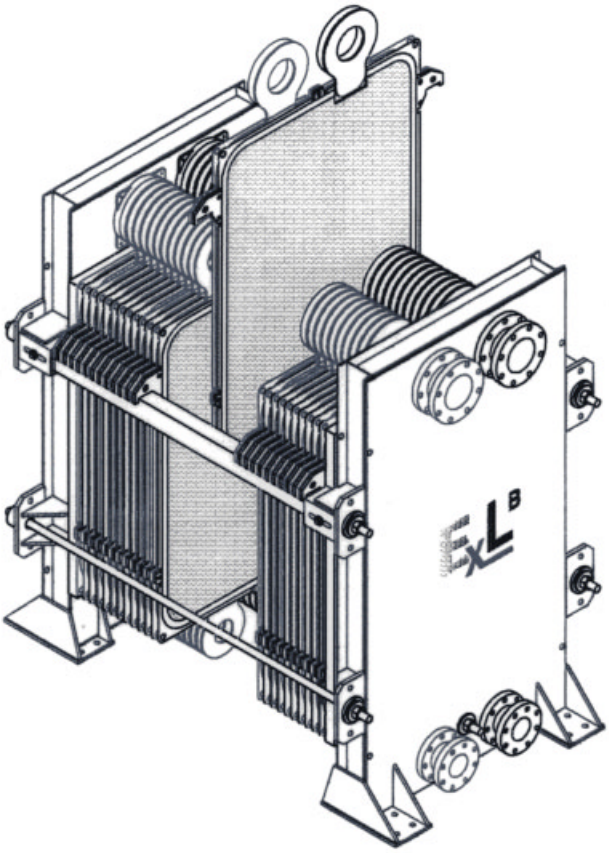

Figure 54. EL-Tech ExL^{B} bipolar electrolyzer

ments. Increased internal electrolyte circulation is achieved by an improved electrode design.

EL-Tech ExL^{DP} Dense Pak Unit Electrolyzer. The EL-Tech ExL^{DP} dense pak electrolyzer comprises monopolar sections in one electrolyzer filter press compression set using standard monopolar cell components. Each monopolar cell section is separated by an insulating Inter Pak Spacer. Mostly three monopolar electrolyzer sections are included, with 2 to 10 elements per section. The dense pak can be configurated to match special rectifier/transformer configurations of existing plants, making it suitable for mercury and diaphragm cell conversion projects. For new plants, the advantage of the ExL^{DP} electrolyzer is reduced current, increased voltage circuits compared to an equivalent monopolar cell unit.

INEUS FM21-SP Electrolyzer. The FM21-SP (Fig. 55) is a monopolar electrolyzer incorporating a simple pressed electrode structure. The anode assembly is composed of a 2 mm thick titanium panel between compression molded joints of a special cross-linked EPDM elastomer. The cathode assembly is composed of a 2 mm thick nickel panel between compression molded joints, also of EPDM.

The anodes and cathodes are assembled between 2 end plates until the number of electrodes required for the desired electrolyzer capacity is reached, up to 60 anodes in the FM21-SP and up to 90 anodes in the larger FM1500. A key feature of both designs is the elimination of any external piping to individual cell compartments by the use of a simple but effective internal header/manifold arrangement.

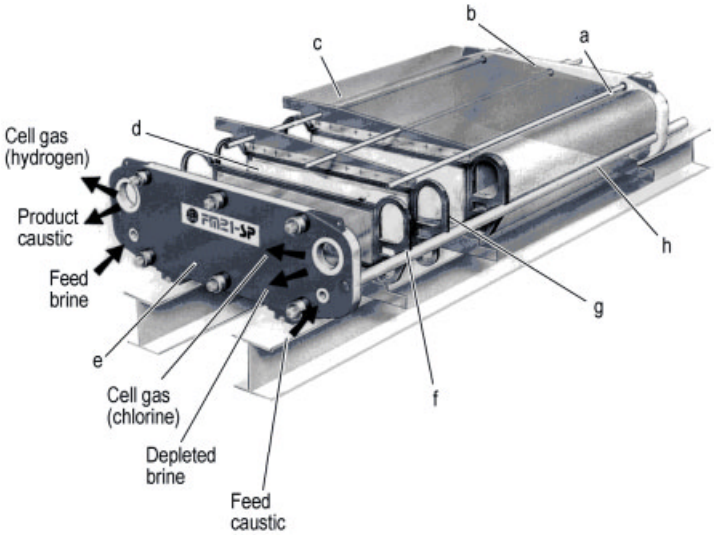

Figure 55. INEUS FM21-SP elelctrolyzer
a) Tie-rod; b) Floating end plate; c) Copper electrical connections; d) Ion exchange membrane; e) Fixed end plate; f) Anode electrode assembly (titanium panel between compression molded gaskets; g) Cathode electrode assembly (nickel panel between compression molded gaskets); h) Support rail

The electrolyzer has coated titanium anodes. The cathodes are pure nickel, also available with a coating to lower the hydrogen overpotential if necessary. Both electrodes are pressed from integral sheets of pure metal, and this makes recoating of the electrodes extremly simple and cost effective. Hence recoated structures can be sent to site prior to electrolyzer refurbishment from a pool of electrodes available to all customers.

Effective electrode area is $2 \times 0.21\,\mathrm{m}^2$ per electrode, which gives a very compact electrolyzer. The individual electrodes are readily handled without the need for lifting apparatus, which allows the electrolyzer to be rebuilt and refurbished in the minimum of time.

7.3.3. Comparison of Electrolyzers

Operating parameters of bipolar electrolyzers are compared in Table 16, and those of monopolar electrolyzers, in Table 17.

7.3.4. Cell Room

Typical bipolar membrane cell rooms are shown in the following Figures 56 and 57.

The media are fed to and discharged from the electrolyzers by a header system arranged along the walls of the cell room on one side.

From the other side power is supplied either from separate transformer/rectifier units for each electrolyzer or from one unit for two or more (up to six) electrolyzers in parallel. The switches are arranged close to the rectifiers. They are actuated automatically and connected to the common interlock system for safety reasons.

In the middle space remains available for the electrolyzers and their individual feed and discharge piping. Only a light crane is required to handle single electrode frames or elements. Thus only a light structure for the entire cell house is used.

8. Electrodes

8.1. Anodes

The initial anodes used for the electrolytic generation of chlorine were made of platinum or magnetite. However, as the plant grew in the size, the cost of platinum and limitations of the current density for magnetite led to the wide-scale introduction of graphite anodes, which were used exclusively up to 1970. The graphite of choice was

Table 16. Bipolar electrolyzers

Company	ASAHI KASEI		CEC	KRUPP UHDE	EL-TECH
Cell	ML 32	ML 60	BITAC 800	BM 2.7	ExL^B
Effective membrane area, m^2	2.72	5.05	3.276	2.72	1.5
Max. no. of elements	150	150	80	160	80
Current density, kA/m^2	up to 6.0	up to 6.0	1.5 – 6.0	1.5 – 6.0	1.5 – 7.0
Max. capacity of electrolyzer t/d NaOH 100 %	45	90	54	90	29
d.c. power consumption kWh/t NaOH	2100	2100	2150	2130	2100
(at current density)	(4.0)	(4.0)	(5.0)	(5.0)	(5.0)

Table 17. Monopolar electrolyzers

Company	CEC	INEUS	EL-TECH
Cell	CME DCM 400	FM 21-SP	ExL^M
Effective membrane area, m^2	3.03	0.21	1.5
Max. no. of elements	32	120	30
Current density, kA/m^2	1.5 – 4.0	1.5 – 4.0	1.5 – 6.0
Max. capacity of electrolyzer, t/d NaOH 100 %	13	7	9
d.c. power consumption kWh/t NaOH	2150	2140	2150
(at current density)	(3.5)	(4.0)	(5.0)

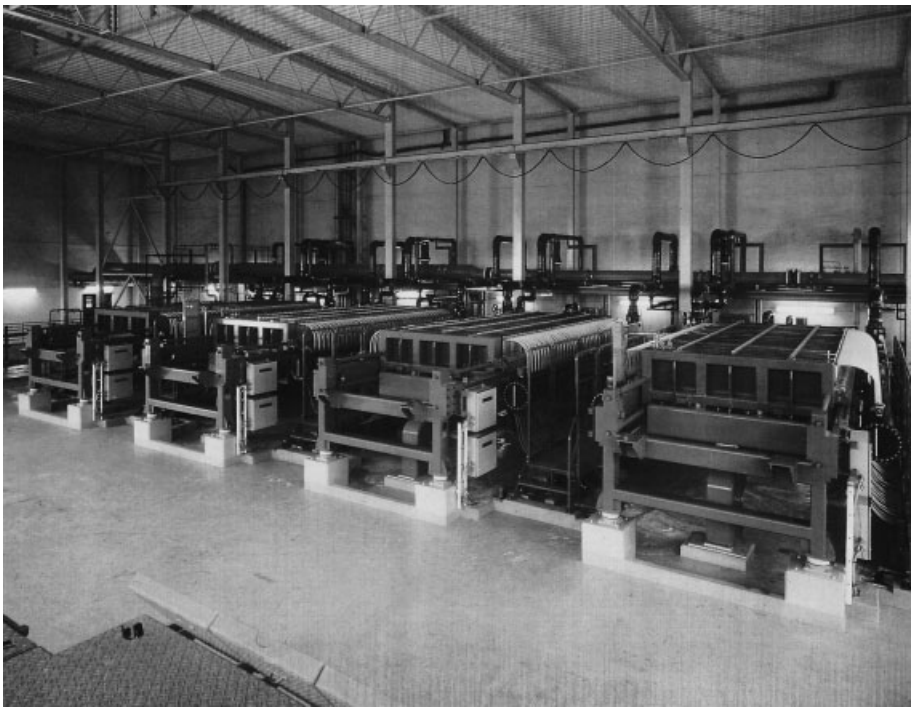

Figure 56. Bipolar cell room by Ashai Glass

Figure 57. Bipolar cell room by Krupp Uhde

low in ash and vanadium and composed of various types of particulate coke and pitch binder. Following extrusion, baking at ca. 1000 °C, and graphitization at 2600 – 2800 °C, the final shape of the electrode was achieved by machining. The shape of the horizontally suspended anodes with an initial thickness of 7 – 12 cm for the amalgam process was similar to that of modern titanium anodes due to the retrofitting of existing cells. The anodes had vertical slits and holes to allow the removal of the gaseous chlorine. Due to the cogeneration of oxygen and the resulting formation of CO and CO_2, electrode wear was high, in the range of 1.8 – 2.0 kg graphite per tonne of chlorine from NaCl and 3 – 4 kg per tonne from KCl. Even with a daily adjustment of the anodes to compensate for the changes in dimension a k value of only 0.12 to 0.14 $Vm^2\ kA^{-1}$ was achievable.

The initial attemps to replace the graphite anodes with activated titanium anodes began as early as 1957 with platinized titanium and Pt/lr-coated anodes. However because of the short lifetimes of the anodes, they were not economic. The use of mixed metal oxides was first patented by Beer in 1965 and 1967 [149]. The initial patent described a coated metal electrode in which the active material was a mixed metal oxide coating containing one or more of the platinum metal group oxides. The second patent described coatings in which mixed metal oxide crystals contained a non-platinum metal oxide in addition to the platinum metal oxide (including Ti, Ta, and Zr oxides).

Further improvements in the coating and the anode structures followed rapidly along with the commercialization of anodes by De Nora [150] under the trade name Dimensionally Stable Anode (DSA). Because of the dimensional stability and the lifetime of the coating and the ability to increase the current densities, rapid introduction of the activated titanium anodes was possible. At present only a few plants still use graphite anodes, largely due to the initial investment costs for titanium anodes.

8.1.1. General Properties of the Anodes

Coating Properties and Preparation. Comprehensive reviews on preparation and properties are given in [151, 152].

Chemical Composition. Because of its price and performance, Ru is the basic component in all commercial coatings at present,

along with an oxide of a non-platinum metal (e.g., Ti, Sn, or Zr). In most cases a second platinum metal oxide is added to increase the performance of the anode coating. There is an optimum ratio of platinum metal oxides to non-platinum metal oxides in terms of overpotential, wear rate, and costs. The optimum depends on the operating conditions and the method of preparation of the coating and normally lies in the range of 20 : 80 to 55 : 45 by weight. Some of these coatings may contain glassy fibers [153] and some contain pre-oxidized material such as $Li_{0.5}Pt_3O_4$ [154].

Preparation. The solvent used for the preparation of the precursors solutions are chosen on the basis of the desired electrochemical properties and the method of application, which is mainly determined by the anode structure. Most coating solutions are prepared by dissolving salts or organometallic complexes in aqueous, organic, or mixed solvents. The coating can be applied by spraying, brushing, dipping, or other techniques. Following evaporation of the bulk of the solvent, the anode is heated to 350 – 600 °C to form the oxidic coating then cooled prior to the next coating cycle. This is repeated until the desired coating thickness is applied. Optional post-thermal treatment can also be carried out.

The optimal performance of the coating depends on the above parameters and on the coating thickness per coating cycle, which must be optimized for each coating and surface pretreatment step.

Crystallographic Composition, Morphology, and Real Surface Area. A rutile phase is the electrochemically active phase of the coating, and although it is thermodynamically unstable, it remains even after many years of operation. The stable phase – anatase TiO_2 in the case of $TiO_2 - RuO_2$ coatings – is electrochemically inactive [155]. The degree of crystallinity and the composition are related to the processing parameters [156] and the various degrees of mixed crystals exhibit different stabilities. The real surface area of the coating is a function of both the titanium pretreatment and the coating composition. The surface of the chlorine-generating coating is often described as "cracked mud" due to its resemblance to a dry river bed. The BET surface area of the coatings or that determined electrochemically vary ca. 400 to 1000 times the geometric surface area [157].

Overpotential and Current – Voltage Relationship. The observed overpotential for chlorine evolution at 2 – 10 kA/m^2 is in the range of 80 – 110 mV [158 – 161], about 70 – 100 mV of which is due to diffusion overpotential effects [161]. The overpotential for the generation of oxygen under similar pH and temperature conditions lies is ca. 300 mV more anodic than that of chlorine generation. Other than oxygen evolution, the only other side reaction is formation of chlorate.

Coating Wear and Coating Lifetime. The coating lifetime is strongly dependent not only on the type of cell – membrane, diaphragm, or mercury – but also on a range of process parameters, including brine quality, current density, and membrane or diaphragm quality. The upper limit of the wear rate would seem to be in the region of 500 t Cl_2/m^2 anodic area for a standard commercial loading.

The wear rate mechanism is discussed in detail in [162, 163]. The effects of various impurities and materials in the brine can be divided into three types.

1) Compounds or ions which attack the substrate, e.g., fluoride or organic acids such as formic or oxalic acid.
2) Materials which built up blocking layers on the surface of the anode, e.g., hydraulic oil or polymer films resulting from delamination of membranes. The irreversible poisoning of coating is caused by ultrathin aluminum silicate layers.
3) Electrochemically active film-forming materials such as MnO_2, which may lead to an increased oxygen content in the chlorine.

Other examples, such as the insensitivity of the performance of diaphragm anodes to almost complete surface coverage by iron oxides illustrate the robustness of commercial coatings.

8.1.2. Anodes for Mercury Cells

Structure. The classical structure of anodes for this process still reflects the retrofitting concept used during the 1970s and the high current

operations at ca. 10 kA/m^2. A typical mercury cell anode consists of a number of copper shafts, protected by either a permanently welded or removable titanium outer sleave, from which the current is distributed to the active surface over distributor bars (Fig. 58).

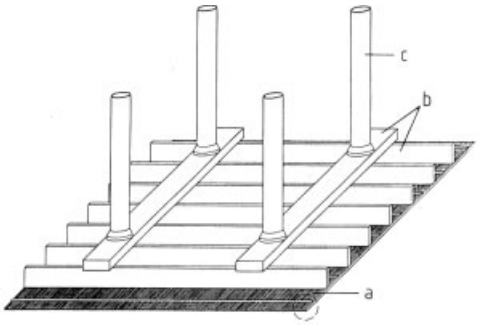

Figure 58. Four-stem anode for amalgam cells
a) Active surface; b) Current distributor; c) Riser tube to protect the copper bar inside

The quick release of gas and the supply of fresh brine to the active surface are the major requirements of an mercury cell anode, and a wide range of designs have been built. The most common types are shown in Figure 59. The differences are more evident at current densities > 7 kA/m^2. The use of baffles on the back of the active surface to enhance the gas lift and aid the supply of brine to the active surface is also common [164].

Flatness is critical for the optimal performance of the anodes in the cells. A typical value < 0.5 mm is achieved, mostly by manual straightening after manufacturing or recoating.

Coating Life. The coating life is determined by a wide range of practical aspects and normally not directly related to the electrochemical wear rate of the coating. These include:

- Mechanical damage to the anode caused by short circuiting [165, 166]
- The need to maintain a recoating schedule, due to production demands and the labor intensive refitting of a cell.
- Synchronization of recoating with the exchange of other consumable parts of the cells such as covers and gaskets.

The tendency has been to increase the lifetime from about 180 t Cl_2/m^2 to 300 – 400 t Cl_2/m^2. This has been achieved by the introduction of better control systems in the cells and the development of intermediate layers of plasma-sprayed conductive TiO_{2-x} between the active coating and the titanium substrate [159].

8.1.3. Anodes for Diaphragm Cells

The predominate determinants in the design of diaphragm anodes are:

- The relatively low current density of ca. 2.0 kA/m^2
- The minimization of the anode – diaphragm gap
- The need to remove and replace the anode array from the cathode, hence the use of retractible anodes

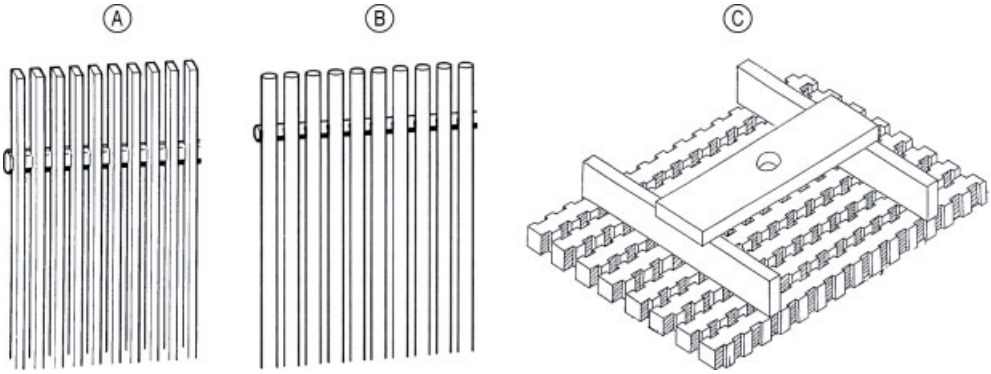

Figure 59. Anode designs for quick gas release
A) Flatt profile (channel blades); B) Rod type (3-, 4-, 5-mm diameter); C) 3D side profile anode

– The limiting height of the cathode, integrated with the diaphragm manufacturing technology

The conventional anodes shown in Figure 60 [167] have been further developed to optimize the energy consumption of the cells by replacing the simple flat expanded metal with complex structures [169]. At present very few plants are still operating without expandable anodes.

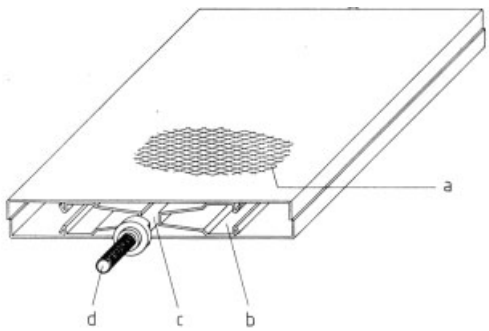

Figure 60. Anode for monopolar diaphragm cells
a) Activated (coated) expanded metal; b) Expanding spring; c) Titanium-clad copper bar; d) Copper thread to fix the anode to the cell base

Another type of diaphragm anode is used in the bipolar Ganor cells [168].

Coating Life and Mechanism of Deactivation. The coating lifetime of DSA coatings exceeds 12 years, and production of chlorine exceeds 240 t Cl_2/m^2. The wear is caused by the relatively high oxygen content in a diaphragm cell [170] of ca 1 – 2 %. The wear rate is nonlinear (Fig. 61). This nonlinearity is critical for determining the correct time to begin recoating so as to avoid unplanned stoppages.

8.1.4. Anodes for Membrane Cells

Structure. The variety of designs of membrane cells has led to a range of anodes active area structures; the common principles are the need to support the membrane and gas release to the back of the anode surface. Therefore, thin flattened expanded, perforated metals or louver type structures with and without perforations are used [171].

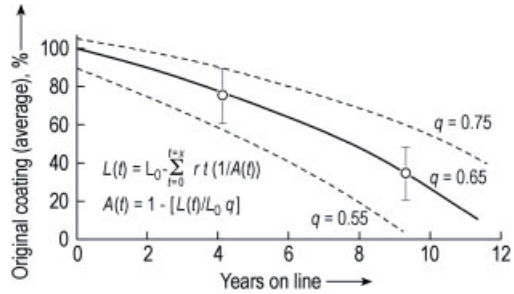

Figure 61. Empirical fit of observed nonlinear wear rate of coating thickness versus years on line.
$L(t)$ is the loading at time t, L the initial loading, r the wear rate, t time, $A(t)$ the active surface area at time t, q an empirical factor related to the current density sensitivity of the wear rate r

Coating Life. At present the second-generation coatings for membrane cells are showing lifetimes comparable to those of the diaphragm process. The actual lifetime of the anodes is dependent on the extent of damage by caustic flow through holes in the membrane or by contamination with the poisons.

Oxygen Content [151]. The oxygen content has also been improved by the optimization of internal circulation of the brine within the cells.

8.2. Activated Cathode Coatings

Since 1910 diaphragm brine electrolyzers have used carbon steel cathodes and continue to use carbon steel to this day. When the first ion-exchange membrane electrolyzers were introduced in the late 1970s, the cathodes were also carbon steel. By the early 1980s the design had evolved to stainless steel and nickel cathodes, and finally in the 1990s to exclusively nickel cathodes. Depending on current density, the hydrogen overpotential of carbon steel cathodes is about 300 mV. Active cathode coatings can lower the overpotential by 200 – 280 mV, thus providing significant energy savings. Active coatings have often been described in the literature and used in water electrolysis for over 40 years. With the development and evolution of the ion-exchange membrane technology, active cathode coatings are coming into general use.

The patent literature covers many different types of coatings, and new ones are being published regularly. The two basic approaches to activation are high-surface area coatings and catalytic coatings. Both bare nickel and carbon steel show lower hydrogen overpotential once in operation and their surfaces roughen. In fact by grit blasting bare nickel cathodes and roughening the surface, the long-term overpotential can be reduced by 30 – 40 mV. More common are porous nickel-type coatings that offer high surface area and good chemical resistance. These coatings consist of two or more components. At least one of the components is leached out in caustic to leave the porous high surface area nickel [172]. These coatings are typicall nickel-zinc [173], nickel – aluminium – Raney nickel [174], nickel – aluminium [175], or nickel – sulfur [176]. A variety of additives are recommended for strength, life, and resistance to poisoning by impurities. Rough coatings of nickel – nickel oxide mixtures [177] and nickel with embedded activating elements such as ruthenium [178] are also used. Sintered nickel coatings are described in patents [179] as well as being available from Huntington Alloys. Nickel coatings containing platinum group metals, primarily platinum and/or ruthenium, have been sold by Dow [180], Johnson Matthey, and ICI [181].

The coatings used for diaphragm and membrane electrolyzers differ because of the different substrates (carbon steel and nickel, respectively) and the different operating conditions. The weak 11 % caustic in diaphragm cell liquor is less corrosive than the strong 33 % caustic of a membrane electrolyzer. The less expensive and more fragile coatings like nickel – zinc can be used in diaphragm electrolyzers. Membrane electrolyzer suppliers favor the platinum group metal coatings.

The shape of the cathode structure is an important factor affecting the choice of cathode coating. The complex cathodes of diaphragm electrolyzers lend themselves to liquid systems (e.g., electroplating or electroless baths) that can coat the entire structure by immersion [182]. Membrane electrolyzers, which are primarily of a filter-press design, have flat cathodes that are easy to coat by spraying or painting. The lower operating current density of diaphragm electrolyzers means more cathode area per unit of production; this requires a less expensive coating. Most diaphragm electrolyzers use heat-cured polmer – asbestos separators (diaphragms) that are vacuum deposited after the cathode coating is applied. This curing operation can destroy the activity of certain coatings.

All cathode coatings are susceptible to poisoning by impurities that make their way into the catholyte with the deionized water or are components of the piping, electrolyzer etc. These impurities tend to blind the activity of the coatings over a period of time that depends on their concentration. Porous nickel coatings in diaphragm electrolyzers are less susceptible to blinding by impurities because spalling of the brittle coating makes the coatings self-cleaning. The platinum group metal coatings are subject to damage from reverse currents during electrolzer outages. Precautions are needed to protect the coatings with reducing agents [183] or by cathodic protection [184].

Active cathode coating have become the standard throughout the chlor-alkali industry for new construction with the ion-exchange membrane electrolyzer technology. In most of the older diaphragm electrolyzer plants, problems with application of the cathode coatings and generally lower power costs have obviated the use of active cathode coatings. While there are more recent developments applicable to the diaphragm technology in the way of active cathode coatings, many of these developments remain the proprietary information of the technology and coating suppliers.

9. Comparison of the Processes

The advantages and disadvantages of the three chlor-alkali processes are summarized in Table 18. The three chlor-alkali processes can be compared in respect to the quality of the chlorine and caustic produced, and the equipment and operating costs.

Today the membrane process is the state of the art for producing chlorine and sodium hydroxide or potassium hydroxide. All new plants are using this technology. The production capacity of chlor-alkali plants using the membrane process reached about 21 % of total world production capacity in 1995 and is predicted to increase to about 28 % by 2001 (Table 19) [132].

Table 18. Advantages and disadvantages of the three chlor-alkali processes

Process	Advantages	Disadvantages
Diaphragm process	use of well brine, low electrical energy consumption	use of asbestos, high steam consumption for caustic concentration in expensive multistage evaporators, low purity caustic, low chlorine quality, cell sensitivity to pressure variations
Mercury process	50 % caustic direct from cell, high purity chlorine and hydrogen, simple brine purification	use of mercury, use of solid salt, expensive cell operation, costly environmental protection, large floor space
Membrane process	low total energy consumption, low capital investment, inexpensive cell operation, high-purity caustic, insensitivity to cell load variations and shutdowns, further improvements expected	use of solid salt, high purity brine, high oxygen content in chlorine, high cost of membranes

The diaphragm cell capacity remains constant and there is a decline in mercury cell capacity.

Table 19. World chlorine market 1995 and 2001 (in %)

	1995	2001
Diaphragm	52	49
Mercury cell	22	18
Membrane	21	28
Others	5	5
Market, 10^6 t/a	41	43.3

The conditions for a conversion from the mercury and the diaphragm process to the membrane process are discussed below.

9.1. Product Quality

Table 20 shows typical composition values for the chlorine and caustic produced by the diaphragm, mercury, and membrane processes. Chlorine produced by the mercury process can be used directly for most uses. Chlorine produced by the diaphragm or membrane process contains up to 2 % O_2, depending on the pH of the anolyte. This oxygen can be removed by condensation and evaporation of the chlorine.

The sodium hydroxide solution from the mercury process is the purest of the three; the amounts of NaCl and $NaClO_3$ are especially low. However, the quality of caustic from the membrane process is almost as good. A main drawback of the diaphragm process is the high concentration of NaCl and $NaClO_3$ in the caustic solution. This sodium hydroxide solution cannot be used for some processes. A chloride-free grade, commonly referred to as rayon-grade caustic, is required for 20 – 30 % of the demand in industrialized countries. Even the use of purification processes (see page 24) does not reduce the NaCl content below 0.03 wt %. In addition to the NaCl and $NaClO_3$, the levels of Si, Ca, Mg, and sulfate impurities are higher than for the mercury and membrane processes.

9.2. Economics

The wide variation in the main cost factor, that for electrical energy, which varies from region to region by a factor of up to three, makes a direct comparison of production costs problematic. Further, the cost of electrical energy is increasing in different regions at drastically different rates, depending on the basic source of energy and customs. Rapidly changing foreign exchange rates also make international comparisons difficult.

A detailed discussion of the capital investment and operating costs for the three processes for a 200 000 t/a-plant in 1991 is given in [185]. A comparison of the investment costs does not make sense today for the mercury process, because no mercury cell plant and only a few diaphragm cell plants were built since then. All new plants are using the membrane process.

9.2.1. Equipment

The expenses for the rectifier, chlorine and hydrogen systems, HCl system, caustic storage, utilities, and engineering and construction overheads are approximately the same for the three processes.

Table 20. Product qualities: typical compositions of chlorine, caustic, and hydrogen

Product and contents	Process			
	Diaphragm		Mercury	Membrane
	Unpurified	Purified		
Chlorine gas (from cells), vol%				
Cl_2	96.5 – 98		98 – 99	97 – 99.5
O_2	0.5 – 2.0		0.1 – 0.3	0.5 – 2.0
CO_2	0.1 – 0.3		0.2 – 0.5	
H_2	0.1 – 0.5		0.1 – 0.5	0.03 – 0.3
N_2	1.0 – 3.0		0.2 – 0.5	
NaOH solution, wt %				
NaOH	50.0	50.0	50.0	50.0
NaCl	1.0	0.025	0.005	0.005
Na_2CO_3	0.1	0.1	0.05	0.04
Na_2SO_4	0.01	0.01	0.0005	0.0001
$NaClO_3$	0.1	0.001	0.0005	0.001
SiO_2	0.02	0.02	<0.001	0.002
CaO	0.001	0.001	0.001	0.0001
MgO	0.0015	0.0015	0.0002	0.0001
Al_2O_3	0.0005	0.0005	0.0005	0.0001
Fe	0.0007	0.0007	0.0005	0.0004
Ni				
Cu	0.0002	0.0002	0.00001	0.0001
Mn				
Hg	none*	none*	0.00001	none*
NH_3		0.001		
Hydrogen gas, vol%				
H_2	>99.9		>99.9**	>99.9

* $< 10^{-6}$ %.
** Hydrogen gas from the mercury process contains mercury: 1 μg/m^3 – 10 mg/m^3, depending on the purification process. The hydrogen gas from the other two processes is free of mercury.

Cells. The complex mercury cells are considerably more expensive than the simpler diaphragm and membrane cells. There is no development in mercury cell technology. Improvements are being made in diaphragm cells (higher current densities, longer service times), but the relative advantage of the membrane cells is rising fast with considerable increase in current density and improved membrane performance: Fewer cells are needed for a given production capacity.

Brine System. The brine system for the diaphragm process is the simplest of the three – there is neither sulfate precipitation nor dechlorination – and makes up only 3 – 4 % of the capital investment. The brine system is the most complex for the membrane process, for fine purification by ion exchange is necessary. However, the two- or three-fold greater depletion of the brine in the membrane process allows the brine system to be smaller than that for the mercury process. Therefore, the cost of the brine system for either process is approximately the same, 4 – 7 % of the total.

Caustic Concentration. The elaborate multistage evaporators required for the concentration of the diaphragm-cell caustic and the separation of NaCl and Na_2SO_4 must be nickel plated because of the corrosiveness of the cell liquor containing NaCl and $NaClO_3$. These evaporators cost 20 – 35 % of the total. The evaporators for the membrane process may be constructed of stainless steel and are much smaller because the essentially salt-free cell liquor is more concentrated, costing 3 – 4 % of the total. The mercury process produces 50 % caustic directly, evaporation is not required.

Facilities for Handling Salt. The mercury and membrane plants require storage and handling facilities for solid salt. If a diaphragm plant uses well brine, only small facility is needed for the recycling of the salt from the caustic evaporation.

Mercury. In addition to the capital cost of mercury itself, there is the expense of the equipment to prevent emission of mercury into the environment and to remove mercury from the products (see page 25). This equipment costs 10 – 15 % of the total capital investment.

The investment cost of a new (green-fields) chlor-alkali project in the USA is estimated to be between 250 000 and $ 300 000 per tonne per day chlorine capacity in 1998 [186].

9.2.2. Operating Costs

The fixed costs for operators and other personnel, taxes, insurance, repairs, and maintenance are about the same for all three processes. The 20 % lower depreciation of the membrane process is offset by the additional expense for purchase and replacement of the membranes and for the more elaborate brine purification.

Of the variable costs, the expense for salt, precipitants, and anode reactivation are roughly the same. The difference among the three processes shows up in the consumption of energy, as electricity and steam. If 1 t of steam is taken to be equivalent to 400 kWh of electrical energy, then the comparison in Table 21 can be made. The differing total energy consumptions are illustrated in Figure 62.

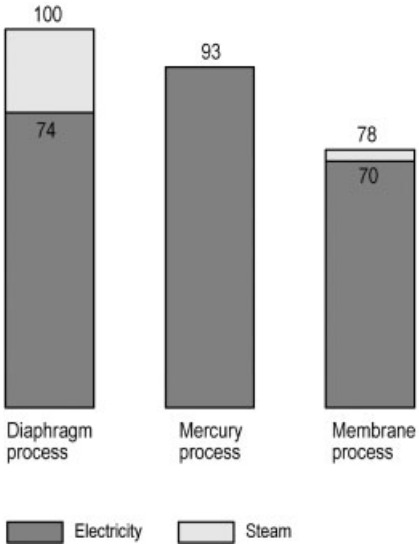

Figure 62. Relative consumption of energy (electricity and steam) in the three chlor-alkali processes in producing 50 wt % NaOH

The price of electrical energy varies widely from region to region. The relatively broad range of possible current densities combined with the steep increase in the cell voltage with current density for the diaphragm and membrane cells allows optimization of the current density with respect to the local energy price. That is, if electrical energy is relatively expensive, a greater number of cells, and thus a greater capital investment, can be tolerated to reduce the specific energy consumption and thus minimize total unit production cost [187].

9.2.3. Summary

In spite of the advantages of the membrane technology, about 75 % of all chlorine is produced in mercury and diaphragm cells, operating in ca. 500 plants around the world. Diaphragm technology prevails in the United States (70 %), Russia, and China, and mercury technology in Western Europe (64 %). Continued production from these plants is economical under special circumstances.

For mercury cell users, the question of today is whether the old, depreciated plant is competitive with new membrane cell plants. The alternatives are:

– Further production in the mercury cell plant
– Conversion to the membrane process
– Phasing out the old plant

Candidates for further production are plants of medium to large size, with low electricity costs, with very high quality products, with high emissions standards, with high maintenance standards (low repair costs), or which produce speciality products which cannot be obtained in membrane cells (e.g., alkoxides or dithionites).

All producers that do not fulfill one or more of these conditions are candidates for conversion. The more the existing infrastructure can be used, the greater the benefits resulting from conversion. The investment for the conversion of a middle-sized plant (100 000 t/a) is between $ 550 and $ 800 per tonne of chlorine capacity per year [188]. This investment includes the cost for the membrane cells, secondary brine purification and additional changes of the infrastructure. The costs are specific for each existing plant and depend on:

Table 21. Energy consumed to produce 1 t of chlorine plus 1.13 t of caustic soda (50 %) in the three chlor-alkali processes

Energy	Process		
	Diaphragm	Mercury	Membrane
Electricity for electrolysis, kWh	2300 – 2900	3100 – 3400	2200 – 2600
Steam equivalent, kWh	800 – 1000	0	200 – 400
Total, kWh	3100 – 3900	3100 – 3400	2400 – 3000
Relative energy costs	100 %	93 %	78 %

- Chlorine quality (e.g., the oxygen content)
- The use of existing buildings. The materials of the existing brine treatment area
- The possibility of using the dilute caustic (32 – 36 %) within the plant without concentration
- Use of the existing electrical equipment, rectifiers, busbars
- Possible capacity enlargement because of the lower specific energy per tonne of chlorine
- Dismantling and disposal of mercury-contaminated parts of the old plant.

Normally the decision for a conversion is initiated by plans for an expansion of the production capacity or by environmental legislation. Each change in the plant structure or in the cost structure may lead to reevaluation of the future of the electrolysis plant. Therefore, each plant has to be considered individually [189, 190]. For the European chlor-alkali industry a detailed analysis of the impact of a conversion of all mercury cells to the membrane technology on the competitiveness of the industry is given in [110].

The situation is different for diaphragm cell plants. These plants are still economic where inexpensive brine (e.g., from solution mining) is available, energy costs are comparably low (e.g., from cogeneration of electricity and steam on site), and when the market price for caustic is determined by the lower quality of diaphragm caustic. In countries like the United States it will be difficult to economically justify conversion. In contrast to the mercury process, improvements to the cells are still being made, resulting in lower operating costs and savings in solid waste disposal. The investment to convert a plant with a capacity of 1000 t/d to membrane technology is ca. 90 million dollars [186].

In the first few years after the introduction of membrane technology, diaphragm cells in several plants were equipped with membranes (retrofit) to reduce the cost of steam for cell liquor concentration, to give a small reduction in electricity consumption and better quality of caustic. This procedure is economic where steam is very expensive [187, 191].

10. Other Production Processes

10.1. Electrolysis of Hydrochloric Acid

Electrolytic decomposition of aqueous hydrochloric acid is used to produce chlorine and hydrogen. The first pilot plant was set up by G. Messner in 1942 in Bitterfeld, Germany, and since 1964 eight full-scale plants have been commissioned in Europe and the United States, a total capacity of 540 000 t/a [192]. Hydrogen chloride is a byproduct of many organic industrial processes. Electrolysis of hydrochloric acid competes with chemical processes in which either hydrogen chloride is used to produce chlorinated hydrocarbons directly, e.g., by oxychlorination, or where chlorine is produced by chemical reaction, e.g., in the KEL chlorine process (see page 69). The advantages of the electrolytic process are very pure products without further treatment, reliability (simple design), ease of operation, flexibility (5 : 1 turndown ratio), and low energy consumption even with small installations.

Principles. Hydrochloric acid (22 wt % HCl) is fed into the cells in two separate circuits, a catholyte circuit and an anolyte circuit. During electrolysis the concentration is reduced to ca. 17 %, and the temperature increases from 65 to 80 °C. A part of the depleted acid is separated from the catholyte stream, concentrated in the absorption plant to ca. 30 %, and fed back into the main stream. The electrolyzer is bipolar, with pairs of electrodes arranged like the leaves of a filter press. A diaphragm or membrane (e.g., Nafion 430) separates the anode compartment from the cathode compartment to prevent mixing of the gaseous products.

The reversible standard decomposition potential of hydrochloric acid is 1.358 V, made up of the anode potential, the discharge of chloride ions with formation of chlorine, and the cathode potential, the discharge of hydroxonium (H_3O^+) ions with formation of hydrogen. In practice (> 15 % HCl, 70 °C), the decomposition potential is $\leq$ 1.16 V.

The graphite electrode plates are not attacked by 22 % hydrochloric acid. A poly(vinyl chloride) (PVC) fabric constitutes the diaphragm. Chlorine dissolved in the anolyte diffuses through the diaphragm and is reduced at the cathode, causing a loss of 2 – 2.5 % of the theoretical current yield. The increase of cell voltage when current flows is mainly because of the hydrogen overpotential at the graphite cathode and the resistance of the electrolyte. Depolarizing agents (polyvalent metal ions) in the catholyte reduce the overpotential by $\leq$ 300 mV at 4 kA/m^2 [193].

The conductivity of hydrochloric acid is maximized at a concentration of 18.5 wt %. High temperatures improve the conductivity, but to avoid increased vapor pressure of HCl and material problems, the temperature is kept below 85 °C. Modern cells have a voltage of ca. 1.90 V at 4.8 kA/m^2, corresponding to an energy consumption of 1400 – 1500 kWh per tonne of chlorine.

Diaphragm Cells. Hydrochloric acid electrolysis cells are manufactured by Hoechst – Uhde [194]. Each Hoechst – Uhde electrolyzer consists of 30 – 36 individual cells that are formed from vertical graphite plates connected in series, between which there are diaphragms. To improve gas release, vertical slits are milled in the graphite plates, which are cemented in frames made of HCl-resistant plastics. At the bottom of the frames, channels feed in the electrolyte. The gases rise up the plates and pass through ducts into collection channels in the upper part of the cell. Chlorine leaves the cell with the anolyte, and hydrogen leaves with the catholyte. The end plates of the electrolyzer are made of steel lined internally with rubber and are held together by spring-loaded tension rods. The electric current is supplied via graphite terminals. The unit rests on insulated steel frames. The effective surface of the electrodes is 2.5 m^2, and the current loading can be up to 12 kA.

DeNora and General Electric are developing an electrolyzer with a solid polymer electrolyte (SPE) based on Nafion [195]. In addition to a voltage savings of 20 %, it is hoped that completely chloride-free hydrogen gas can be produced.

Operation. A simplified flow diagram of the process as operated by Bayer – Hoechst – Uhde is shown in Figure 63.

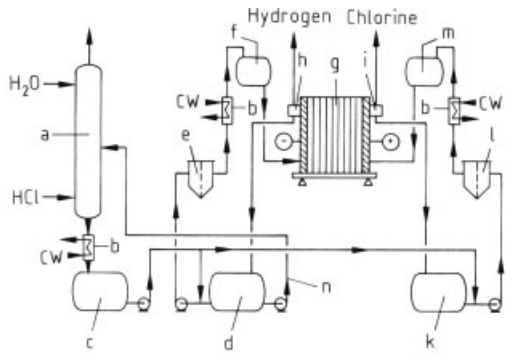

Figure 63. Simplified flow diagram of a hydrochloric acid electrolysis
a) Absorption column; b) Heat exchanger; c) Strong acid tank; d) Catholyte collecting tank; e) Catholyte filter; f) Catholyte supply tank; g) Electrolyzer; h) Hydrogen – catholyte separator; i) Chlorine – anolyte separator; k) Anolyte collecting tank; l) Anolyte filter; m) Anolyte supply tank; n) Weak acid line to absorber

In the absorption column, the hydrogen chloride gas is absorbed adiabatically by depleted hydrochloric acid from the catholyte. In the upper section of the column, an absorber removes the remaining hydrogen chloride and the water vapor by absorption in a water stream, which

makes up the water balance of the process. The 30 wt % acid that is produced is then cooled, purified if necessary by activated carbon, and supplied to the anolyte and catholyte circulation systems.

The electrolyte is pumped through a filter and heat exchanger to a gravity feed tank for the electrolyzer unit. The gases produced are freed from the electrolytes in separators, and the electrolytes flow back into their respective collecting tanks to be resaturated. The working life of the PVC diaphragms, 1 – 2 years, depends on the impurities in the acid. The concentrated acid is, therefore, purified carefully [196].

The product gases are saturated with water vapor and hydrogen chloride at the partial vapor pressures of 20 % hydrochloric acid. Both product streams are cooled. Sodium hydroxide solution is used to wash the hydrogen, removing chlorine and hydrogen chloride and producing a 99.9 % product. The chlorine, which is dried by sulfuric acid, contains ca. 0.5 % hydrogen and ca. 0.05 % carbon dioxide. The hydrogen overpotential can be reduced by activation of the cathodes.

Membrane Cells [197, 198]. Since 1992 Bayer has replaced the woven fabric cloth in the diaphragm cells by anion-exchange membranes of the sulfonate type. Only hydrated protons are able to pass from the anolyte to the catholyte, so that the whole cell and the electrolyte systems are simplified. Together with an optimized surface of the electrodes for better gas release, this leads to:

- Lower cell voltage of 300 mV, corresponding to a power consumption of 1300 kWh per tonne of chlorine at 4.8 kA/m^2
- Longer life of the cell components
- Higher product quality
- Improved safety of operation
- Simplified process

A similar electrolytic process for recovering chlorine from anhydrous HCl, also using membrane cell technology, has been developed by DuPont [199].

10.2. Chemical Processes

The chlor-alkali process produces chlorine and sodium hydroxide solution in fixed stoichiometric proportions. Experience has shown that there tends to be a surplus of either chlorine or sodium hydroxide. Chlorine may, however, be produced competitively without the byproduct sodium hydroxide by nonelectrolytic methods. The starting material is usually hydrogen chloride, which is catalytically oxidized to chlorine by oxygen, air, nitric acid, sulfur trioxide, or hydrogen peroxide. Other processes start from ammonium chloride or metal chlorides.

10.2.1. Catalytic Oxidation of Hydrogen Chloride by Oxygen

A catalyst is essential for the economic oxidation of hydrogen chloride to chlorine by air or oxygen (Deacon Process), and the catalyst must be active at low temperature and have adequate life. There are many patents claiming improved catalysts and equipment. Most of the catalysts are oxides and/or chlorides of metals on various substrates. Only three processes have been commercialized.

The KEL Chlorine Process. The process developed by Kellogg [197] uses concentrated sulfuric acid (ca. 80 %) with ca. 1 % nitrosylsulfuric acid as the catalyst. From 1975 to 1988 Du Pont operated a full-scale plant in Corpus Christi, Texas, recovering up to 600 t/d of chlorine. The plant was shut down due to a change in the structure of the plant and because of material problems after more than 10 years of operation. The raw material, from a fluorinated hydrocarbon plant, consisted of waste gases that contained hydrogen chloride [200]. Figure 64 shows a simplified flow diagram.

Sulfuric acid catalyst is fed into the top of the stripper column. The hydrogen chloride gas reacts with the catalyst to form nitrosyl chloride:

$$HCl + NOHSO_4 \rightarrow NOCl + H_2SO_4$$

The oxygen, the ultimate oxidizing agent, blows the remaining hydrogen chloride out of the sulfuric acid, which becomes more concentrated and also is cooled in a flash vaporizer. This acid is then fed back into the process. Nitrosyl chloride, hydrogen chloride, oxygen, and water

vapor flow as a gaseous stream into the oxidizer and react there, increasing the temperature:

$$2\,NOCl + O_2 \rightarrow 2\,NO_2 + Cl_2$$

$$NO_2 + 2\,HCl \rightarrow NO + Cl_2 + H_2O$$

In the absorber – oxidizer, the rest of the hydrogen chloride is oxidized. Concentrated sulfuric acid is fed in at the top, reacts with the oxides of nitrogen to form nitrosylsulfuric acid, absorbs the water that has formed, and is conducted back into the stripper:

$$NO + NO_2 + 2\,H_2SO_4 \rightarrow 2\,NOHSO_4 + H_2O$$

$$NOCl + H_2SO_4 \rightarrow NOHSO_4 + HCl$$

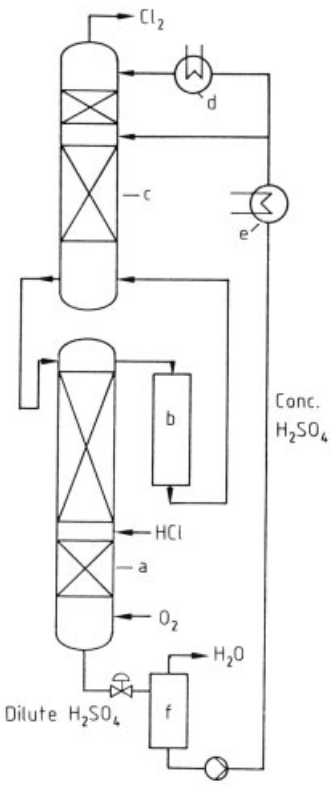

Figure 64. Flow diagram of the KEL chlorine process (simplified)
a) Stripper; b) Oxidizer; c) Absorber – oxidizer; d) Acid chiller; e) Acid cooler; f) Vacuum flash evaporator

The cooled, dried chlorine gas still contains ca. 2 % hydrogen chloride and up to 10 % oxygen. Both are removed by liquefaction.

The net reaction is

$$4\,HCl + O_2 \rightarrow 2\,Cl_2 + 2\,H_2O$$

The installation at Corpus Christi operated at 1.4 MPa and 120 – 180 °C. On account of the aggressive nature of the chemicals, expensive materials, such as tantalum-plated equipment and pipes, must be used. For outputs of 250 – 300 t of chlorine per day, this process can be more economical than the electrolysis of hydrochloric acid, depending on local conditions.

The Shell Chlorine Process. The catalyst developed by Shell consists of a mixture of copper(II) chloride and other metallic chlorides on a silicate carrier [201]. The reaction of the stoichiometric mixture of hydrogen chloride and air takes place in a fluidized-bed reactor at ca. 365 °C and 0.1 – 0.2 MPa. The yield is 75 %. The water condenses out from the gas stream, and the hydrogen chloride is removed by washing with dilute hydrochloric acid. After the residual gas has been dried with concentrated sulfuric acid, the chlorine is selectively absorbed, e.g., by disulfur dichloride. After desorption and liquefaction, the chlorine has a purity > 99.95 %.

A manufacturing unit was built by Shell in the Netherlands, 41 000 t/a, and another in India, 27 000 t/a, but both have been closed down owing to the prolonged surplus of chlorine on the market.

The Mitsui MT-Chlorine Process. The catalyst consists of chromium(III) oxide on a silicate carrier [202]. In a fluidized-bed reactor, hydrogen chloride is reacted with oxygen at a temperature of 415 °C to give chlorine gas with a conversion rate of 73 – 77 %. The reactor is made from nickel-lined low carbon steel. The concentration of the purified product is > 99.5 % Cl_2.

A commercial plant for 30 000 t Cl_2/a is successfully operating since 1988, with an expansion to 60 000 t Cl_2/a in 1990.

A two-stage cyclic fluidized bed process for converting HCl to chlorine is described in [203]. The catalytic oxidation process combines the exothermic oxidation of 60 – 70 % of the HCl at 380 – 400 °C in a fluidized bed of copper oxychlorides impregnated on zeolite with the transfer of the reaction products to a second reactor operating at 180 – 200 °C where the rest of HCl is converted.

10.2.2. Oxidation of Hydrogen Chloride by Nitric Acid

The nitrosyl chloride route to chlorine is based on the strongly oxidizing properties of nitric acid:

$$6\,HCl + 2\,HNO_3 \rightarrow 2\,Cl_2 + 2\,NOCl + 4\,H_2O$$

$$2\,NOCl + 2\,H_2O + O_2 \rightarrow 2\,HCl + 2\,HNO_3$$

The practical problems lie in the separation of the chlorine from the hydrogen chloride and nitrous gases. The dilute nitric acid must be reconcentrated. Corrosion problems are severe. Suggested improvements include (1) oxidation of concentrated solutions of chlorides, e.g., LiCl, by nitrates followed by separation of chlorine from nitrosyl chloride by distillation at 135 °C or (2) oxidation by a mixture of nitric and sulfuric acids with separation of the product chlorine and nitrogen dioxide by liquefaction and fractional distillation [204].

10.2.3. Production of Chlorine from Chlorides

Alkali-metal chlorides, ammonium chloride, and other metallic chlorides are reacted, usually with nitric acid, to produce nitrate fertilizers [205]. Chlorine is not produced directly, but it can be obtained from the intermediate products nitrosyl chloride or hydrochloric acid.

11. Chlorine Purification and Liquefaction

Chlorine produced by the various processes, especially by electrolysis, is saturated with water vapor at high temperature and may also contain brine mist and traces of chlorinated hydrocarbons, and is normally at atmospheric pressure. Before the chlorine can be used, it must be cooled, dried, purified, compressed, and where necessary, liquefied. A simplified flow sheet is shown in Figure 65.

11.1. Cooling

Table 22 shows the volume, water content, and heat content of 1 kg of chlorine gas at 101.3 kPa as a function of temperature. To avoid solid chlorine hydrate formation, the gas is not cooled below 10 °C [206]. Cooling is accomplished in either one stage with chilled water or in two stages with chilled water only in the second stage.

Table 22. Volume, moisture content, and enthalpy of 1 kg of chlorine gas at 101.3 kPa as a function of temperature *t*

t, °C	Volume, m^3		Water content, g/kg **	Heat content, kJ/kg *
	Dry	Saturated *		
0	0.312	0.314	1.54	3.81
20	0.335	0.342	5.95	24.45
40	0.357	0.385		69.50
			19.7	
60	0.380	0.473	61.5	188.41
70	0.392	0.565	112.0	325.73
80	0.404	0.756	222.0	623.83

* Chlorine gas saturated with water vapor at temperature *t*.
** Grams of H_2O per kg of Cl_2.

The chlorine gas can be *cooled indirectly* in a tubular titanium heat exchanger so that the cooling water is not contaminated and the pressure drop is small. The resultant condensate is either fed back into the brine system of the mercury process or dechlorinated by evaporation in the case of the diaphragm process.

The chlorine gas can be *cooled directly* in packed towers. Water is sprayed into the top and flows countercurrent to the chlorine. This treatment thoroughly washes the chlorine; however, dechlorination of the wastewater consumes a large amount of energy. The cooling water should be free of traces of ammonium salts to avoid the formation of nitrogen trichloride.

Closed-circuit direct cooling of chlorine combines the advantages of the two methods.

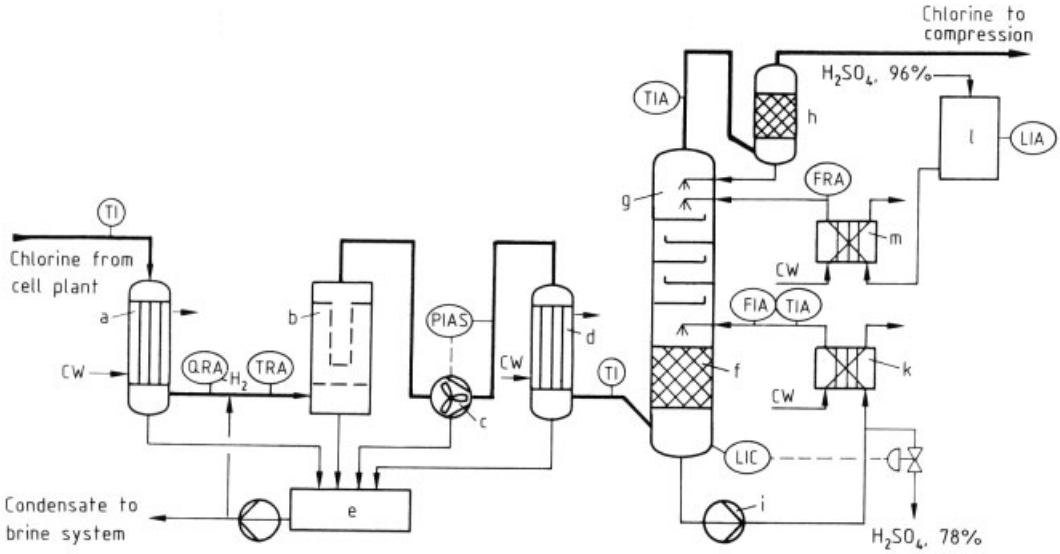

Figure 65. Simplified flow diagram of a chlorine processing plant
a) Chlorine gas cooler (primary); b) Chlorine demister; c) Blower or fan; d) Chlorine gas cooler/chiller (secondary); e) Condensate collection tank; f) Drier, first stage; g) Drier, second stage; h) Sulfuric acid mist separator; i) Sulfuric acid circulation pump; k) Cooler for circulating sulfuric acid; l) Sulfuric acid feed tank; m) Cooler for sulfuric acid feed

The chlorine-laden water from the cooling tower is cooled in titanium plate coolers and recycled. The surplus condensate is treated like the condensate from indirect cooling. Spray towers, as well as packed towers, are used. Water carryover is removed by demisters, which reduce the amount of sulfuric acid used for drying.

11.2. Chlorine Purification

Water droplets and impurities such as brine mist are mechanically removed by special filter elements with glass wool fillings. The efficiency varies with the gas throughput. A commonly used device is the Brink *demister* [207]. Instead of glass wool, porous quartz granules can be used.

In *electrostatic purification,* the wet chlorine gas is passed between wire electrodes in vertical tubes. The electrodes are maintained at a d.c. potential of 50 kV with a current density of 0.2 mA/m^2. The particles and droplets in the chlorine become charged and collect on the tube walls. The resultant liquid is fed back into the brine system or chemically treated before disposal.

Activated carbon filters can adsorb organic impurities and may be regenerated by heating to 200 °C.

Gaseous impurities can be removed by *absorption* of the chlorine in a suitable solvent, such as carbon tetrachloride, water, or disulfur dichloride, followed by *desorption.* This can be coupled with further processes, such as the recovery of chlorine from the waste gas remaining after liquefaction [15, pp. 418 – 422].

A wash with concentrated hydrochloric acid removes the dangerously explosive nitrogen trichloride [208]. *Scrubbing with liquid chlorine* (see Fig. 67) mainly reduces the content of organic impurities and carbon dioxide, but it can also lower the bromine content. When the chlorine is cooled down to near its dew point, liquid chlorine scrubbing is often combined with compression by turbo or reciprocating compressors.

11.3. Drying

Drying of chlorine is carried out almost exclusively with concentrated sulfuric acid (96 – 98 wt %) [209]. Depending on the desired final concentration of the waste acid, drying can be a two-, three-, or four-stage process. The acid and chlorine flow countercurrently. The fi-

nal moisture content depends on the concentration and temperature of the acid in the final stage (Fig. 66). An upper limit is 50 ppm H_2O. Low-temperature liquefaction (−70 °C) demands lower moisture content, which can be achieved with molecular sieves, whereby 2.5 ppm is possible [210].

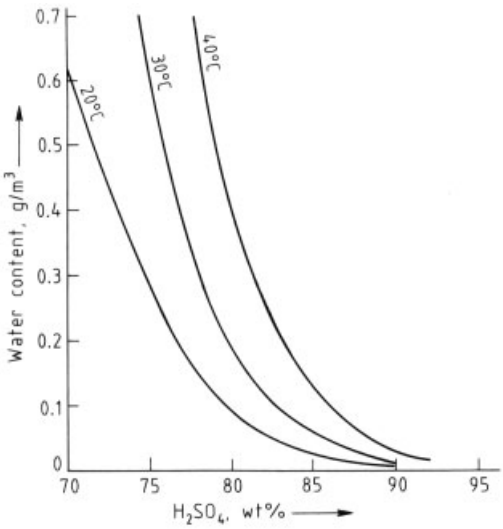

Figure 66. Drying chlorine with sulfuric acid
Attainable moisture content as a function of concentration and temperature of the acid

The *packed towers* usual in the first stages are constructed of rubber-lined steel or glass-fiber-reinforced poly(vinyl chloride). The heat liberated on dilution of the circulating acid is removed by titanium heat exchangers, and the weak acid is dechlorinated chemically or by blowing air. Often the acid is recirculated after reconcentration to 96 % by heating under vacuum. Generally, columns with *bubble cap plates* or *sieve trays* are used at the final stage. The drying is effective, but the pressure drop is great. Occasionally, spray towers are used to dry chlorine.

After drying, the chlorine gas is passed through a demister or a packed bed to remove sulfuric acid mist.

11.4. Transfer and Compression

In all operations involving compression, care must be exercised to prevent the heat of compression from increasing the temperature enough to ignite material in contact with the chlorine.

Wet chlorine gas can be compressed 20–50 kPa by a single-stage *blower* or *fan* with a rubber-lined steel casing and titanium impeller. It can also be compressed in *liquid-ring compressors,* so that further treatment of the chlorine can be accomplished in smaller equipment [211]. Sulfuric acid ring compressors are used for throughputs of 150 t of dry chlorine gas per day per compressor and for pressures of 0.4 MPa or, in two-stage compressors, 1.2 MPa. The heat of compression is removed by cooling the circulating liquid; cooling of the gas is not necessary. Advantages are simplicity of construction, strength, and reliability, but efficiency is low [212].

Reciprocating compressors were formerly lubricated with sulfuric acid, but are now available as *dry-ring compressors* (no lubrication). They can compress up to 200 t per day. Multistage compressors produce pressures up to 1.6 MPa. The heat of compression of each stage must be removed by heat exchangers or by injection of liquid chlorine (see Fig. 67). Well-purified chlorine gas is essential for trouble-free operation [213].

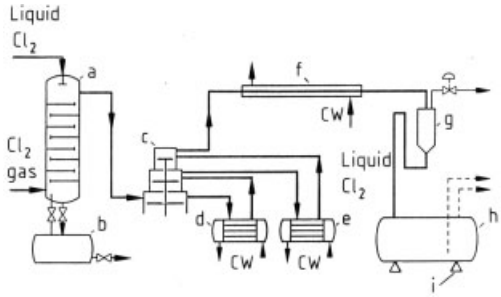

Figure 67. Multistage reciprocating compressor for chlorine liquefaction at 1 MPa with cooling water at 15 °C with liquid chlorine scrubbing
a) Low-temperature cooling and scrubbing column; b) Collection tank for impurities; c) Three-stage compressor; d) Intermediate cooler, stage 1; e) Intermediate cooler, stage 2; f) Liquefier; g) Chlorine collection vessel; h) Chlorine storage tank; i) Chlorine storage tank on load cells

Turbo compressors are most economical when they operate with large amounts of chlorine. Each unit compresses up to 1800 t/d. In multiple-stage operation, pressures up to

1.6 MPa are reached. Labyrinth seals are used on the high-speed shafts. Requirements for cooling and gas purity are like those of reciprocating compressors. *Screw compressors* handle low rates of chlorine and give pressures up to 0.6 MPa. *Sundyne blowers* are one-stage high-speed centrifugal compressors handling 80 – 250 t per day and giving pressures up to 0.3 MPa. Liquid chlorine injection is used for cooling [214]. *Membrane compressors* are used for pressurizing storage tanks with chlorine gas to transfer liquid chlorine to other vessels [215]. Liquid chlorine is pumped with *canned pumps* [216].

11.5. Liquefaction

The most suitable liquefaction conditions can be selected within wide limits. Important factors are the composition of the chlorine gas, the desired purity of the liquid chlorine, and the desired yield. There are nomograms that give the relationship between the chlorine concentrations of the incoming and residual gases, liquefaction yields, pressures, and temperatures [217]. Increasing the liquefaction pressure increases the energy cost of chlorine compression, although the necessary amount of cooling decreases, resulting in an overall reduction in energy requirement (Table 23) [218].

Table 23. Electrical energy requirement for compression and liquefaction of 1 t of chlorine gas

	Liquefaction pressure, MPa			
	0.1	0.3	0.8	1.6
Energy for compression, kWh/t	5	23	42	57
Energy for cooling, kWh/t	87	68	27	3
Combined energy, kWh/t	*92*	*91*	*69*	*60*
Starting temperature, °C	−36	−8	25	53
Final temperature, °C	−42	−17	14	40

Any hydrogen is concentrated in the residual gas. To keep the hydrogen concentration below the 6 % explosive limit, conversion of gas to liquid should be limited to 90 – 95 % in a single-stage installation. Higher yields may be obtained by condensing the chlorine from the residual gas in a second stage, which is constructed to reduce the risk from explosion [219]. This is achieved by the use of sufficiently strong equipment to withstand explosions or by the addition of enough inert gas to keep the mixture below the explosive limit. Multistage installations can liquefy over 99.8 % of the chlorine gas.

High-pressure (0.7 – 1.6 MPa) liquefaction with *water cooling* (Fig. 67) does not require a cooling plant. Therefore, it has the lowest energy cost of all methods; however, the high construction cost must be set against this.

Medium-pressure (0.2 – 0.6 MPa) liquefaction with *cooling* (− 10 to − 20 °C) is especially useful when only a part of the chlorine is to be liquefied and the remaining gas is to be reacted at the liquefaction pressure, e.g., with ethylene to form ethylene dichloride. The residual gas can be fed into the compressor suction systems, provided that the increased inert gas content does not interfere with the subsequent process. Otherwise, the residual gas must be scrubbed free of chlorine or liquefied in a second stage.

Figure 68 shows a two-stage liquefaction by the Uhde system, which operates at 0.3 – 0.4 MPa and − 20 °C in the first stage and − 60 °C in the second stage, with a yield of 99 % [220]. The refrigerant is difluoromonochloromethane. The gaseous refrigerant is compressed, liquefied by water cooling, and collected in a container. The liquid refrigerant is sprayed into the shell of the chlorine liquefier, where it evaporates, absorbing heat and cooling the chlorine, which flows from the liquefiers at − 15 °C (first liquefier) or − 55 °C (second liquefier) [221]. The residual gas from the first horizontal liquefier contains < 5 % hydrogen. It is fed into the second liquefier, which is at an angle of 60° and has a strong, low-volume construction. There the gas mixture passes through the explosive concentration limits. In case of an explosion, there is a comprehensive control system to ensure safety:

The explosion pressure is vented by means of a bursting disk to a residual gas absorber. Simultaneously the residual gas from the first stage is passed directly into this absorber. The chlorine gas to the second stage is shut off, and an inert gas purge is introduced. Finally, the liquid chlorine exit valve is closed to prevent back flow of the liquid chlorine into the second liquefier and from there into the absorber.

With normal-pressure (ca. 0.1 MPa) liquefaction and *low temperature* (< − 40 °C), cryogenic storage of the liquid chlorine is possible.

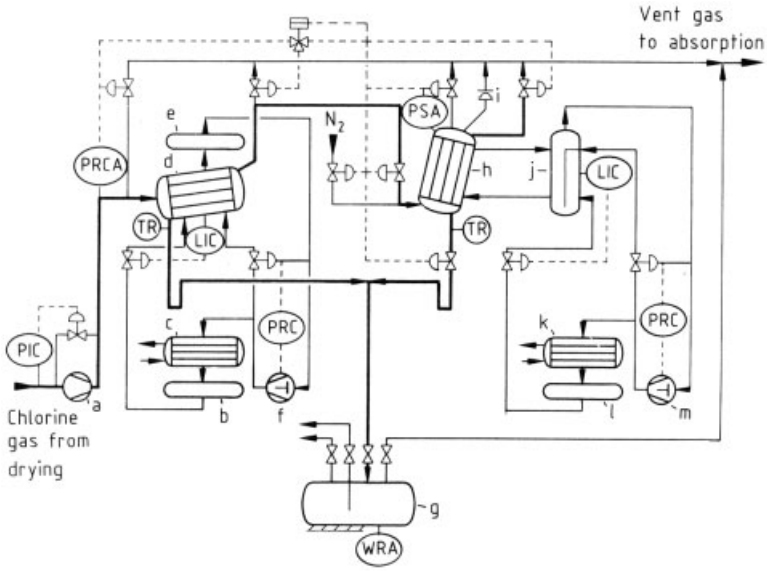

Figure 68. Flow diagram of a two-stage chlorine liquefaction plant at intermediate pressure – Uhde system
a) Chlorine gas compressor; b) Refrigerant collector, stage 1; c) Refrigerant condenser, stage 1; d) Chlorine liquefier, stage 1; e) Refrigerant separator, stage 1; f) Refrigerant compressor, stage 1; g) Liquid chlorine storage tank; h) Chlorine liquefier, stage 2; i) Bursting disk; j) Refrigerant separator, stage 2; k) Refrigerant condenser, stage 2; l) Refrigerant collector, stage 2; m) Refrigerant compressor, stage 2

This process is advantageous when large quantities of chlorine must be liquefied as completely as possible. Attention must be paid to the increased solubility of other gases at low temperatures, especially carbon dioxide [206]. This carbon dioxide can be removed from the liquid chlorine by passage of hot chlorine gas [219].

An *absorption – desorption process* by Akzo is based on carbon tetrachloride [222]. It requires little energy and yields over 99.8 % of a pure liquid chlorine that is almost free of carbon dioxide. A similar process by Diamond Shamrock has been described [224].

11.6. Chlorine Recovery

Chlorine can be recovered from the tail gas from liquefaction with a chlorine recovery system.

Tail gas from liquefaction and chlorine from the plant evacuation system together with the snift compressor and stripper recycle streams are supplied to a snift compressor suction knockout drum. The gas is compressed by the snift gas compressor to 7.0 kg/cm^2 with a discharge temperature of 85 °C.

The snift gas is then cooled by cooling water to 45 °C and then further cooled to − 12.2 °C by Freon. Gas is sent to the absorber, whereas liquid is either returned to chlorine storage or is used for reflux at the stripper.

The off-gas enters the bottom of the chlorine absorber and passes upward through the two packed sections of the tower while cold carbon tetrachloride flows downward. All of the chlorine and the nitrogen trichloride is absorbed in the carbon tetrachloride while the noncondensable gases remain in the gasphase and are removed from the system.

The chlorine-rich carbon tetrachloride leaves the bottom of the chlorine absorber at ca. 10 °C and is forced by pressure difference to the chlorine stripper. Chlorine stripper feed enters the middle of the column and flows downward through two packed sections, releasing chlorine as it is heated. A thermosiphon reboiler is provided at the base of the stripper. By heating the liquid above 65 °C, the absorbed nitrogen trichloride decomposes to nitrogen and chlorine gas.

Chlorine boiled off in the stripper passes upward through a packed top section of the column where it is scrubbed and purified by liq-

uid chlorine from the discharge knock-out drum. The stripper overhead stream, a mixture of chlorine and a small amount of inerts, is sent to the chlorine liquefaction system or recycled to the suction knock-out drum to maintain the stripper reflux [222, 223].

12. Chlorine Handling

Both the chlorine industry and governmental organizations are well aware of the risks of chlorine. In the United States and Canada, The Chlorine Institute [225] has established standards and recommendations for safe transport and handling of chlorine since 1924. In Europe, Euro Chlor, an association of major Western European chlorine manufacturers, publishes recommendations, codes, and memorandums for chlorine handling and transport concerning European conditions and regulations [226]. Both organizations distribute manuals and pamphlets worldwide. Surveys of existing national and international regulations for the handling and transport of hazardous chemicals are available [2, 34, 227, 229].

12.1. Storage Systems

Chlorine is liquified and stored at ambient or low temperature [230, 233]. In both cases the pressure in the storage system corresponds to the vapor pressure of liquefied chlorine at the temperature in the stock tank. *Pressure storage* is recommended for all usual customers [234, 235]. Euro Chlor recommends a maximum capacity of 300 – 400 t for individual tanks. For the large storage capacities required by producers, usually a *low-pressure storage* system, operating at a liquid chlorine temperature of ca. – 34 °C, is chosen. A low-pressure system needs a cooling or recompression system, and, for this reason, it is basically unsuitable for small chlorine consumers [236, 237].

A few major design aspects must be mentioned. Any risk of fire or explosion must be eliminated. All tanks having an external connection below the liquid level should be placed in a liquor-tight embankment (bund). In the event of leakage the liquid should be collected in a small area to reduce the rate of vaporization. The outer shell around a double-enveloped low-pressure storage tank can provide such a facility. To vent chlorine, there must be an absorption or liquefaction system. In the course of all operations, the design pressure should not be exceeded. The dimensions of branches and the amount of pipe work should be minimized. Bottom connections from storage tanks are not recommended for small chlorine users. Large branches should always be located in the gas space of a vessel. The pipework system should be provided with remotely operable valves to permit isolation in case of emergency.

Before being put into service, the whole storage system must be degreased, cleaned, and dried to achieve a dew point of – 40 °C in the purge gas at the outlet of the system. No substance that could react with the chlorine can be allowed to enter the storage system. The filling ratio in the tank should never exceed 95 % of the total volume of the vessel; for pressure storage tanks, this corresponds to 1.25 kg of liquid chlorine per liter of vessel capacity at 50 °C (Fig. 69).

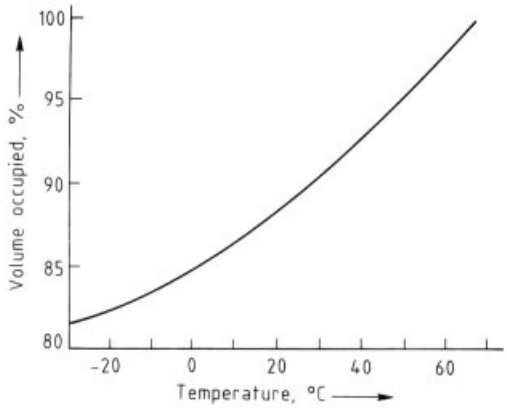

Figure 69. Proportion (%) of a one-liter vessel occupied by 1.25 kg of liquid chlorine as a function of temperature

Typical measuring and control equipment of a pressure storage tank is shown in Figure 70. The ISO codes for process measurement control functions and instrumentation are explained in Table 24 [238]. The measuring equipment of a low-pressure storage system needs supplementary devices, for example, a temperature indicator with an alarm and, in the case of a double-

shell vessel, a device to determine the quality of the purging gas inside the double shell. The vessel and an external envelope should be protected against overpressure or underpressure. In low-pressure systems, the chlorine is removed by vertical submerged pumps, canned pumps below the vessel, or ejector pumps operating with a flow of liquid chlorine produced by external pumps.

Table 24. ISO codes and miscellaneous symbols for process measurement control functions and instrumentation

Codes	Function or Instrumentation
AA	analysis alarm
CW	cooling water
dPI	difference pressure indicating
FA	flowrate alarm
FI	flowrate indicating
FIA	flowrate indicating alarm
FICA	flowrate indicating controlling alarm
FRA	flowrate recording alarm
HZ	hand operated emergency acting
H	high
L	low
LA	level alarm
LIA	level indicating alarm
LIC	level indicating controlling
M	moisture analysis
PA	pressure alarm
PCZA	pressure controlling emergency acting alarm
PI	pressure indicating
PIA	pressure indicating alarm
PIAS	pressure indicating alarm switching
PIC	pressure indicating controlling
PRC	pressure recording controlling
PSA	pressure switching alarm
QRA	quality recording alarm
TA	temperature alarm
TC	temperature controlling
TI	temperature indicating
TIA	temperature indicating alarm
TIC	temperature indicating controlling
TRA	temperature recording alarm
WI	weight indicating
WIA	weight indicating alarm
⬭	measuring device
⧖	remote control valve
– – –	control line

Periodic inspection and retesting of the whole system, including a visual examination, a thickness test of the wall of the vessel and pipes, and an examination of the welds and the surfaces under any thermal insulation, is recommended. Hydraulic retesting is accompanied by risk of corrosion and is, therefore, not favored.

12.2. Transport

Within a chemical plant and over distances of several kilometers, chlorine can be transported by pipelines, either as gas or liquid [24, 239]. Every precaution should be taken to avoid any vaporization of chlorine in a liquid-phase system or any condensation in a gas-phase system. Wherever liquid chlorine could be trapped between two closed valves or wherever the system could be overpressurized by thermal expansion, an expansion chamber, a relief valve, or a rupture disk should be provided [240, 241].

Commercial chlorine is transported as a liquid, either in small containers (cylinders and drums) or in bulk (road and rail tankers, barges, and ISO containers). The design, construction, system of labeling, inspection, and commissioning are covered by national and international regulations [227]. *Cylinders* have a chlorine content up to 70 kg. A protective hood is provided to cover the valve during transport. The ton containers (drums) have a capacity of 500 – 3000 kg of chlorine. Drums are equipped with two valves near the center of one end and connected with internal eductor pipes.

The capacity of *tank cars* (rail tankers) ranges from 15 to 90 t. Special angle valves are mounted on the manhole cover on top of the vessel. In Europe, pneumatic valves are normally used [230 – 232]. During loading and unloading, these valves can be closed rapidly and remotely in case of an accident. They have an internal safety plug, providing a tight seal against the passage of gas or liquid chlorine in the event of failure of the body of the valve.

In North America, the eductor pipe inside the vessel has an excess-flow valve at the top, immediately below the manhole cover. This valve closes the eductor pipe when the rate of liquid flow exceeds a set rate [2, 24]. North American tank cars have a spring-loaded safety relief valve, which protects the vessel against overpressure in case of external heat. The tanks have thermal insulation. In Europe thermal insulation and safety relief valves are not used or recommended.

Road tankers and ISO containers have a chlorine capacity of 15 – 20 t. The design of and the equipment on chlorine pressure road tankers is similar to these of rail tankers. In North America, large amounts of chlorine are transported by

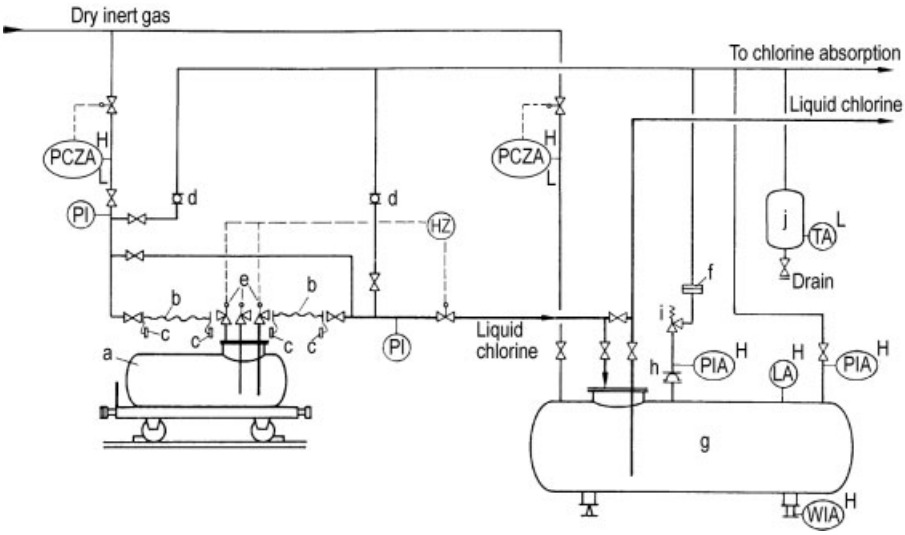

Figure 70. Discharge of liquid chlorine by padding to pressure storage
a) Liquid-chlorine rail tanker; b) Flexible connection; c) Plug; d) Viewing glass; e) Remote-control tank valves; f) Protective membrane; g) Storage vessel; h) Rupture disk; i) Relief safety valve; j) Buffer vessel for liquid chlorine

tank barges [24]. These barges usually are of the open-hopper type with several cylindrical uninsulated pressure vessels. The total capacity of barges ranges from 600 to 1200 t. The chlorine is transported at low temperature.

Classification and Labeling. According to Directive EEC 67/548, Annex I, chlorine (Index no. 017–001–00–7) is classified as toxic and dangerous to the environment. The R and S phrases are:

R 23	toxic by inhalation
R 36/37/38	irritating to eyes, respiratory system and skin
R 50	very toxic to aquatic organisms
S 7/9	keep container tightly closed and in a well-ventilated place

For transportation, chlorine is in class 2, no. 2TC (toxic, corrosive) in ADR, RID, and ADNR in Europe, and in class 2.3 in the IMDG Code (p. 2116) and ICAO Code. All vessels must be labeled with the denomination 268/1017 Chlorine and with labels for dangerous goods: cylinders and drums with labels 2.3 (toxic gas) and 8 (corrosive); railroad tankers with labels 2.3, 8, and 13 (shunt carefully); and for shipping, labels 2.3, 8, and MP (marine pollutant).

12.3. Chlorine Discharge Systems

All containers should be discharged in the same order as received. They must be placed where any external corrosion, risk of fire, explosion, or damage is avoided [24]. At normal room temperature, the discharge rate of chlorine gas from a single 70-kg cylinder is ca. 5 kg/h and the rate of a drum is ca. 50 kg/h. The flow of chlorine gas can be increased by a higher ambient temperature or by connecting two or more containers. A system of two or more containers must be carefully operated and controlled to avoid overfilling by transfer of chlorine from warm to cool containers. Direct heating of containers is not recommended [24]. The best way to determine the flow rate and container content is to observe the weight of the container [228]. A flexible tube is used to connect a mobile container with the fixed piping system. Any reverse suction from the consuming plant must be prevented by a barometric leg or other adequate precaution if the chlorination process runs at atmospheric pressure. Pressurized processes need a pressure controlling system with automatic isolation valves.

Uninsulated tanks have a maximum gaseous discharge rate of ca. 2 t/h. The chlorine gas can be used only for low-pressure chlorination processes and at low rates. This method increases

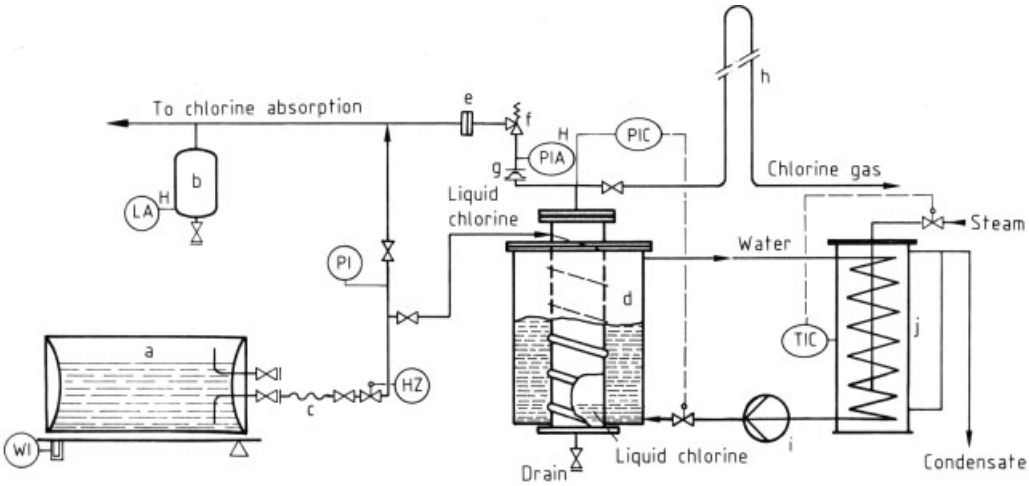

Figure 71. Liquid chlorine vaporizer
a) Liquid-chlorine drum; b) Buffer vessel; c) Flexible coil; d) Chlorine vaporizer; e) Protective membrane; f) Relief safety valve; g) Rupture disk; h) Barometric leg; i) Water pump; j) Water heater

the risk of concentrating nitrogen trichloride and other nonvolatile residues in the liquid phase within the tank. In all other circumstances, the liquid chlorine should be transferred into a fixed storage vessel and then vaporized in a special installation.

Liquid chlorine is discharged by putting the tank under pressure with dry inert gas or dry chlorine gas. The inert transfer gas must have a dew point below $-40\,^{\circ}C$ at atmospheric pressure and must be clean and free of impurities such as dust or oil. Before closing the valves, the tanks must be vented to avoid the risk of high pressure in the container on account of the additional partial pressure of the inert gas. The use of an inert gas requires the availability of a chlorine absorption or neutralization system. Discharge with pressurized chlorine gas requires a chlorine vaporizer or a special chlorine compressor. Articulated arms, flexible hoses, and steel coils are used for the flexible connections. Remote-control valves installed close to the ends of the flexible connections limit leakage in the event of a failure.

Recommendations on technical equipment, installation, taking into operation, checks, and handling are provided by The Chlorine Institute and Euro Chlor.

12.4. Chlorine Vaporization

When large amounts of chlorine gas are required or when the chlorination process needs pressurized gas, liquid chlorine must be vaporized and superheated to avoid liquefaction [242, 243]. It is advisable to operate the vaporizer at a sufficiently high temperature to accelerate the decomposition of nitrogen trichloride. As a source of heat, steam with a maximum allowed temperature of $120\,^{\circ}C$ is used when the vaporizing system is constructed of mild steel. Water above $60\,^{\circ}C$ is also suitable, as shown in Figure 71. Direct electrical heating is not appropriate because there is always a risk of overheating the steel.

Coil-in-bath vaporizers use a coiled tube or a spiral located in a vessel of hot water (Fig. 71). Generally, they are used for small throughputs; they are simple in design and construction. Double-envelope vaporizers have compact construction and are easy to operate and to maintain. Vertical tube vaporizers have a large surface area and allow a high flow rate. Kettle vaporizers are also constructed for large unit capacities [244]. Every effort must be taken to avoid the reverse suction of water or organic materials into the vaporizer. The recommended water and nitrogen trichloride content of introduced liquid chlorine must not be exceeded. Vaporizers operating at low temperature or with a constant liquid level

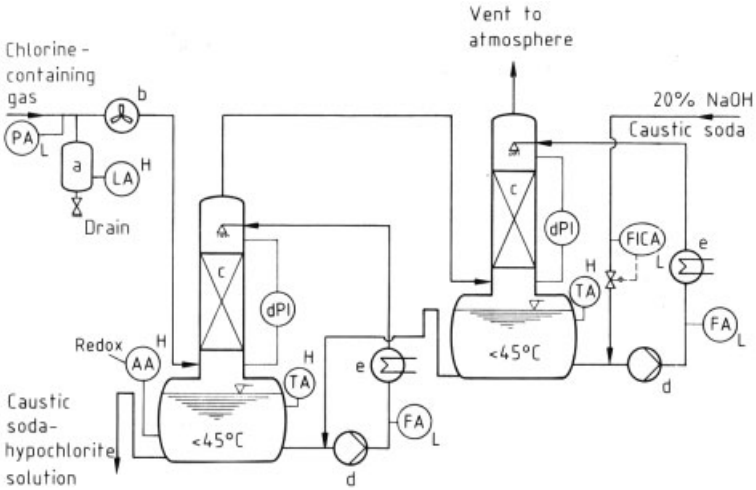

Figure 72. Absorption equipment for the treatment of gases containing chlorine
a) Buffer vessel; b) Vent fan; c) Packed tower; d) Circulating pump; e) Heat exchanger (cooler)

need to be purged to avoid dangerous concentration of nitrogen trichloride [245, 246].

12.5. Treatment of Gaseous Effluents

Gaseous effluents containing chlorine arise from various sources and must be treated in such a way as to obtain a tolerable concentration of chlorine when they are released into the air. The vent gas may contain other substances, such as hydrogen, organic compounds, CO_2, etc., which must be considered in design and operation of an effluent treatment installation [247, 248].

Operation of the collection system below atmospheric pressure facilitates the purging of chlorine vessels, pipes, etc. The risk of corrosion in dry chlorine installations by moisture from the treatment system must be excluded. The most commonly used and recommended reagent is caustic soda. The effluents are treated in an absorption system, such as packed absorption towers, venturi scrubbers, etc. An example of a flow sheet for a large plant is shown in Figure 72.

To avoid any formation of solid salts, the recommended concentration of caustic soda is < 22 wt %. The operating temperature should not exceed 55 °C; under normal conditions a temperature of ca. 45 °C is usual. A cooling system may be necessary. In large chlorine absorption units, the sodium hypochlorite solution that is produced can be used in other processes. Where this is not possible, several methods can be used to decompose the hypochlorite: controlled thermal decomposition, catalytic decomposition [249], acidification, for example, with sulfuric acid

$$NaCl + NaOCl + H_2SO_4 \rightarrow Na_2SO_4 + Cl_2 + H_2O$$

12.6. Materials

The choice of material [250] depends on the design and operating conditions and must take into account all circumstances. A chlorine manufacturer should be consulted to confirm the suitability of a material. Any use of silicone materials in chlorine equipment should be avoided.

Dry Chlorine Gas (water < 40 ppm by weight). Carbon steel is the material most used for dry chlorine gas. It is protected by a thin layer of ferric chloride. For practical purposes the recommended temperature of these materials is ≤ 120 °C. High-surface areas, such as steel wool, or the presence of rust and organic substances increase the risk of ignition of steel.

The resistance of stainless steels to chlorine at high temperature increases with the content of nickel. For stainless steels containing less than 10 wt % nickel, the upper temperature limit is

150 °C. High-nickel alloys, such as Monel, Inconel, or Hasteloy C, are suitable up to 350–500 °C. Poor mechanical strength limits the use of nickel. Copper is used for flexible connections and coils, but it becomes brittle when stressed frequently.

Because titanium ignites spontaneously in dry chlorine, it must be avoided. Graphite, glass, and glazed porcelain are used where there is a risk of moisture in the dry chlorine gas, and poly(vinyl chloride) (PVC) or chlorinated PVC and polyester resins are suitable if the temperature limits of these materials are regarded.

Liquid Chlorine. Unalloyed carbon steel and cast steel are used with liquid chlorine. Low-temperature chlorine systems apply fine-grain steels with a limited tensile strength to guarantee good conditions for welding. To avoid erosion of the protective layer, practice is to limit the velocity of the liquid to less than 2 m/s. Organic materials – rubber lining, ebonites, polyethylene, polypropylene, PVC, chlorinated PVC, polyester resins, and silicone – are dangerous [251]. Zinc, tin, aluminum, and titanium are not acceptable. For certain equipment, copper, silver, lead, and tantalum are appropriate.

Wet Chlorine Gas. Wet chlorine gas rapidly attacks most common metallic materials with the exception of tantalum and titanium. To assure a protective oxide layer on the surface of the titanium, sufficient water must be present in the chlorine gas. If the system does not remain sufficiently wet, titanium ignites spontaneously [252].

Most organic materials are slowly attacked by wet chlorine gas. Rubber-lined iron is successfully used up to 100 °C. At low pressure and temperature the use of plastic materials like PVC, chlorinated PVC, and reinforced polyester resins is advantageous. Polytetrafluoroethylene (PTFE), poly(vinylidene fluoride) (PVDF), and fluorinated copolymers like tetrafluoroethylene – hexafluoropropylene (FEP) are resistant even at higher temperature. Ceramics have been progressively replaced by plastics. Impregnated graphite is suitable up to 80 °C; the impregnation should be resistant to wet chlorine.

Materials for Special Parts. After the ban of asbestos as material for gaskets, substitutes are recommended [253]. In wet chlorine gas, rubber or synthetic elastomers are acceptable. Even at temperatures up to 200 °C, PTFE is resistant against wet and dry chlorine gas and liquid chlorine.

Materials resistant because of protection by a chloride surface layer are not recommended for protective membranes, rupture disks, and bellows. Suitable materials are tantalum, Hasteloy C, PTFE, PVDF, Monel, and nickel.

12.7. Safety

In hazard and risk assessment studies, the design of chlorine installations and equipment and the operating and maintenance concepts are examined in detail to minimize risks [254]. However, there remains a certain risk, and all efforts must be taken to protect people and the environment in the case of a chlorine emergency. The penetrating odor and the yellow-green color of a cloud indicate chlorine in the air. If around-the-clock surveillance by operators is not possible, automatic leak detectors are available. Safety in handling chlorine depends largely on the education and training of employees. An emergency plan should be brought to the attention of the personnel involved. Computer-assisted systems can be used in certain circumstances [255]. Periodic exercises and safety drills should be carried out.

All people on a chlorine plant are advised to carry escape-type respirators. The use of filter masks is prohibited where there is a risk of a high concentration of chlorine. Anyone who enters an area with high chlorine concentrations should be equipped with self-contained breathing apparatus and full protective clothing suitable for dealing with liquid chlorine. Protective equipment, safety showers, eye-wash facilities, and emergency kits [24] must be quickly accessible.

A means of indicating the actual wind direction should be located near the chlorine installation.

Fixed or mobile water curtains can be used to divert the dispersion of a chlorine gas cloud [256]. However, the direct discharge of water into liquid chlorine and on the area of a chlorine leak must be avoided.

In most countries, chlorine manufacturers have organized groups of experts who are well versed and drilled in handling chlorine and can be called at any time in case of chlorine emergency.

The Chlorine Institute has released pamphlets and recommendations covering all aspects of safety, e.g., first aid [257], emergency response plans [258], protective equipment [259], prevention of injuries to personnel [260], prevention of chlorine releases [261], and estimating the area affected by a chlorine release [262]. Emergency kits have been developed for sealing leaks in chlorine containers, drums, and tank cars [263].

Euro Chlor offers a Chlorine Safety Manual [264], recommendations for emergency intervention [265], and for safe design, construction, operation of equipment [266].

Emergency plans have been established for accidents during transportation and use, e.g., CHLOREP [267] in North America and TUIS (Transport Unfall Informations System) in Germany.

13. Quality Specifications and Analytical Methods

13.1. Quality Specifications

Liquid chlorine of commercial quality must have a purity of at least 99.5 wt % [268]. The water content is < 0.005 wt %, and solid residues are < 0.02 wt %. The impurities are mainly CO_2 (≤ 0.5 wt %), N_2, and O_2 (each 0.1 – 0.2 wt %). There are traces of chlorinated hydrocarbons (originating from rubberized or plastic piping) and inorganic salts such as ferric chloride. The chlorine may also contain small amounts of bromine or iodine, depending on the purity of the salt used in the electrolytic process.

13.2. Analytical Methods

Industrial liquid chlorine is mainly analyzed by the methods in ISO regulations. The liquid chlorine is evaporated at 20 °C, and this gas is then analyzed.

Sampling	ISO 1552 [269, 270]
Moisture	ISO 2121 [271], ASTM E410 [272] ISO 2202 [273]
Chlorine content	ISO 2120 [274], ASTM E412 [275]
Gaseous components	DIN 38 408, part 4 [276]
NCl_3	[277]
Mercury	[278], ASTM E506 [279]

The residue is weighed, and the organic constituents are taken up in acetone, hexane, or diethyl ether and determined by gas chromatography. The inorganic residue is analyzed. For quick analysis, liquid chlorine can be introduced directly onto a silica gel column of a gas chromatograph.

Chlorine Gas. The chlorine gas can be analyzed for chlorine content, gaseous impurities, hydrogen, organics, and moisture:

1) Chlorine content. One method for process monitoring and control of chlorine concentration is measurement of thermal conductivity.
2) Gaseous impurities. A known amount of chlorine gas is passed through a solution of potassium iodide or phenol to absorb the chlorine. The residual gases (O_2, N_2, H_2, CO, CO_2) are collected in a gas burette, measured, and analyzed by gas chromatography or with an Orsat apparatus.
3) Hydrogen. A known amount of air is added to the residual gas after removal of the chlorine to ensure excess oxygen, and the volume reduction is measured after the hydrogen is consumed on a heated platinum coil. The hydrogen content can be continuously monitored by thermal conductivity measurement.
4) Organics. Organic components can be determined most conveniently by gas chromatography.
5) Moisture. A known amount of chlorine gas is passed through a drying tube filled with a weighed amount of phosphorus pentoxide. The moisture content is determined from the weight gain of the drying tube (ISO 2121). Continuous determination can be carried out, e.g., by absorption with phosphorus pentoxide and measuring the current required to electrolyze the absorbed water or by the electrical conductivity after absorption in sulfuric acid.

Detection of Chlorine. Chlorine can be recognized by smell or color. Small amounts can be detected by the blue coloration of starch–iodide paper, although other oxidizing agents can produce the same effect. Another method for chlorine detection depends on its ability to combine with mercury. If the unknown gas mixture is shaken with water and mercury, all the chlorine disappears and the remaining water has a neutral reaction. However, if the chlorine contains some hydrogen chloride, the water becomes acidic and reacts with silver nitrate solution to give a white precipitate (AgCl) that is soluble in aqueous ammonia. Leaks in pipes or equipment are detected by testing with the vapor from aqueous ammonia: a thick white cloud of chloride forms.

Quantitative Determination of Free Chlorine. The gas mixture can be shaken with a *potassium iodide solution,* and the liberated iodine can be then determined by titration. Chlorine in alkaline solution can be reduced to chloride by potassium or sodium *arsenite,* and the arsenite can be then oxidized to arsenate. The end point is detected by spot tests with starch–iodide paper. Excess arsenite is back-titrated with acidified potassium bromate solution. Small amounts of chlorine, e.g., in drinking water, can be determined by *photometric measurement* of the yellow color produced by the reaction with *o*-tolidine in hydrochloric acid solution [280].

To determine both *chlorine and carbon dioxide,* the chlorine is absorbed by a solution that contains known amounts of acid and potassium arsenite, and the chlorine is determined by back-titration of the arsenite. The carbon dioxide, which is not absorbed by this solution, is then absorbed by potassium hydroxide solution.

Detection Tubes. Commercial detection tubes (Drägerwerk, Lübeck; Auergesellschaft, Berlin) are available for measuring chlorine in air. They have various ranges: 0.2–3, 2–30, and 50–500 ppm. The Chlorometer (Zeiss-Ikon, Berlin) can determine the free chlorine content of water in a few minutes.

For the protection of the environment and control of working conditions, traces of chlorine as small as 0.01–10 ppm must be determined. There are many types of apparatus on the market for measuring workplace concentrations or emissions. They depend on physicochemical methods, such as conductometry, galvanometry, potentiometry, colorimetry, and UV spectroscopy [281].

14. Uses

The first industrial use of chlorine was to produce bleaching agents for textiles and paper and for cleaning and disinfecting. These were liquid bleaches (solutions of sodium, potassium, or calcium hypochlorite) or bleaching powder (chlorinated lime). Chlorine was then regarded merely as a useful chemical agent.

Since 1900s, chlorine has achieved constantly increasing importance as a raw material for synthetic organic chemistry. Chlorine is an essential component of a multitude of end products, which are used as materials of construction, solvents, pesticides, etc. In addition, it is contained in intermediates that are used to make chlorine-free end products. It is these areas of use that allow chlorine production to increase. The direct use of elemental chlorine, e.g., for sterilizing water, has declined in some areas, but not in others. For example, in Germany, it is $< 0.1\,\%$, but in the United States it is $6\,\%$ (1982). The WHO recommends chlorine as the most useful agent for water disinfection [282].

The percentage of world chlorine consumption of various product groups in 1997 was as follows [283]:

Vinyl chloride	33
Misc. organic products	6
Solvents	19
Pulp and paper	5
Water treatment	6
Others (inorganic products, etc)	31

The number of possible reactions of chlorine, and therefore the number of intermediates and end products, is remarkably large. Some important reactions are shown in Figure 73 along with the areas of application of the end products [284]. Figure 74 also shows some of the applications of chlorine. About one-third of the finished products are chlorine-free, although made with the help of chlorine [285].

Inorganic Products. The use of chlorine to produce pure hydrochloric acid, bleaching agents, and other inorganic products is increasing, although slowly enough that this forms a

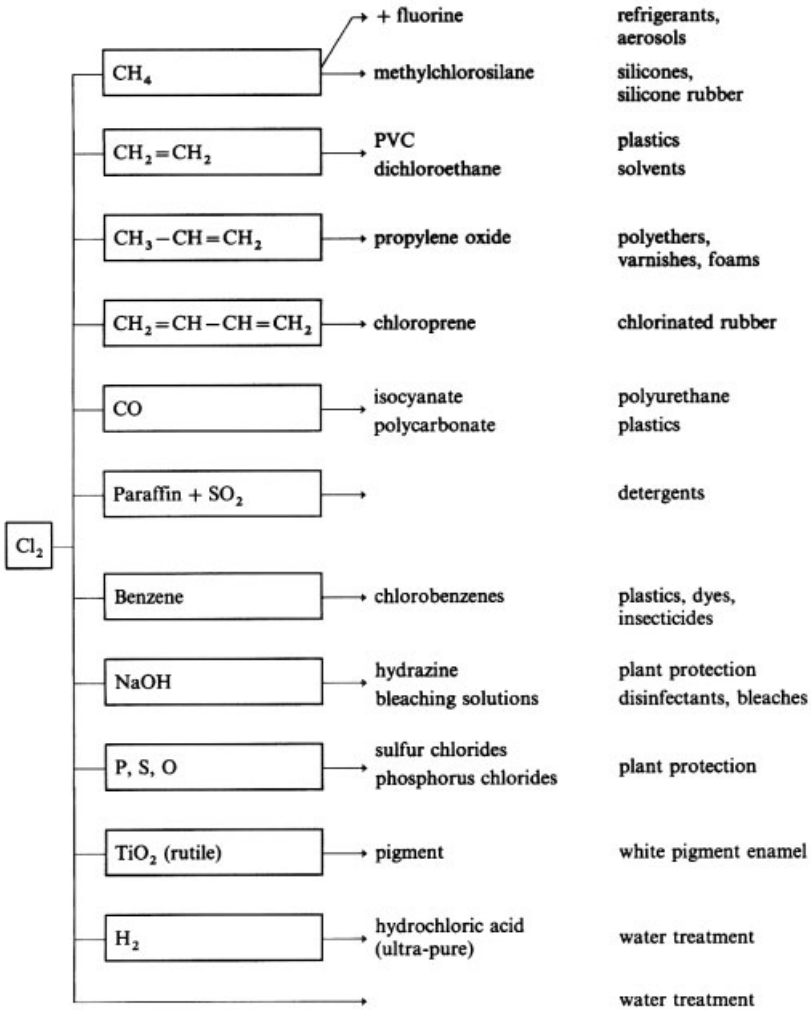

Figure 73. Important reactions of chlorine and the uses of the end products

steadily decreasing proportion of the total consumption. The production of some metal chlorides is increasing, e.g., titanium tetrachloride, aluminum chloride, and silicon tetrachloride.

Bleaching. Countries with large paper and cellulose industries had a large chlorine consumption for bleaching purposes. Ecological concerns about AOX (absorbable organic halogens) in the effluents of papermills have led to replacement of elemental chlorine by other bleaching agents, including ozone, hydrogen peroxide, and chlorine dioxide. As a consequence, the use of chlorine is dropping sharply; e.g., United States: 13 % in 1983, 9 % in 1994, 6 % in 2000, 0 % in 2010 (forecast); the decline in Canada and in Scandinavia is even sharper (1983: 60 – 70 %; 1998: near 0 %) [286].

Organic Products. Organic intermediates and end products accounted for up to 85 % of the chlorine use in developed countries. The product with the largest chlorine requirement is *vinyl chloride*, which accounted for 33 % of world chlorine consumption in 1997. This proportion will increase further and outweigh consumption losses in other areas and is therefore the reason for increasing chlorine capacities worldwide.

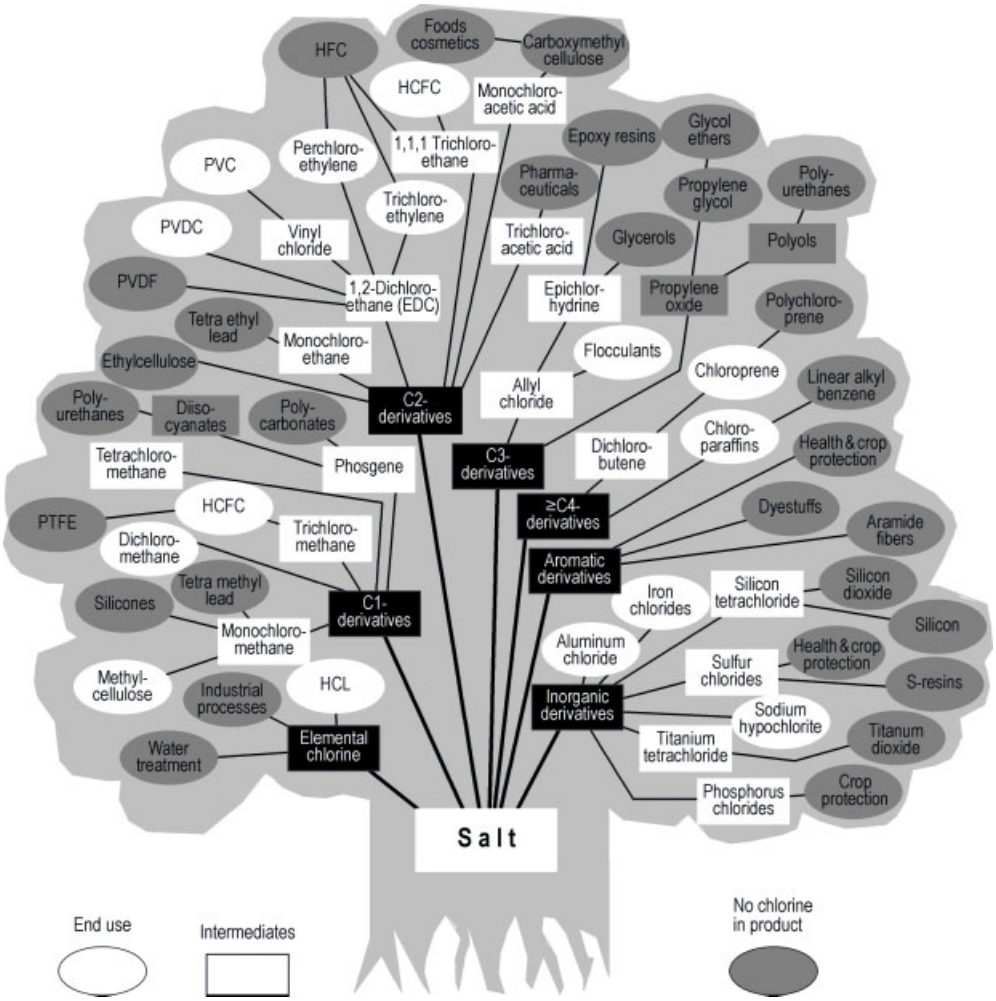

Figure 74. The chlorine tree (courtesy of Euro Chlor)

Chlorine reacts with methane to give the solvents methyl chloride, methylene chloride, chloroform, and carbon tetrachloride. Butane and pentane are the starting materials for hexachlorobutadiene and hexachloropentadiene, the basis of many pesticides. Chlorinated paraffins are starting materials for detergent raw materials, plasticizers, and lubrication additives.

The chlorination of ethylene first gives 1,2-dichloroethane, which is an important starting material for a variety of further products, including vinyl chloride, ethylene oxide, ethylene glycol, ethylenediamine, and chloral. Other products of the chlorination of ethylene are trichloroethylene, perchloroethylene, etc.

The following products are obtained from propylene: propylene oxide, propylene glycol, carbon tetrachloride (via propylene dichloride), glycerine (via allyl chloride), and epoxy resins and synthetic rubber (via epichlorhydrin).

Aromatic derivatives are also important. The reaction of chlorine with benzene produces the following compounds, among others: monochlorobenzene, dichlorobenzene, and hexachlorobenzene, which are used as solvents or pesticides. A more important use of monochlorobenzene, however, is as an intermediate in the manufacture of aniline, and dyes. A further important synthesis route is to start from toluene to produce toluenediamine and with the

help of phosgene to convert this to toluene diisocyanate, which is used in plastics.

Environmental Aspects. The production and use of chlorine, as well as certain chlorinated substances bear risks for human health and the environment. International discussions are underway on legally-binding conventions and protocols thereof. Matters of concern include the use of mercury, asbestos, chlorine containing materials which are persistant, toxic, and liable to bioaccumulate (e.g., polychlorinated biphenyls, dioxins and furans, DDT, endrine, chlordane), and chemicals having an ozone-depletion potential (e.g., chlorofluorocarbons). Regarding the principles of "Responsible Care" and of "Sustainable Development" [287] the chemical industry is managing this challenge: Once a hazard is realised, the substance in question is handled individually by investigating its properties and its impact on the environment on a scientific basis. Risk assessments and risk – benefit studies are prepared, the results are published as fact sheets, position papers, scientific publications and contributions to the discussion with governments, authorities, and the public. Practical measures for reducing the risks are continuous efforts to reduce or eliminate emissions (mercury, asbestos), offer recycling concepts for products (PVC, chlorinated solvents), improve production processes and transportation safety, cease the production (CFCs, pentachlorophenol in most countries) or marketing of problematic compoundes, and developing for alternative processes (bleaching in the pulp and paper industry). All these measures are influencing the use pattern of chlorine: the use for bleaching, for C_1 compounds, for solvents, and for tetraethyllead will further decrease; the importance of organic chemicals will rise and PVC will expand especially in the construction sector. A study conducted by ECOTEC [288] illustrates the chlorine flow within the industrial production system and to external consumers.

15. Economic Aspects

In Western Europe in 1995 [289]:

- Almost 2×10^6 jobs were related to chlorine
- 55 % of European chemical turnover depended on chlorine
- 85 % of pharmaceuticals are made using chlorine
- 98 % of Western Europe's drinking water is purified by chlorination

These figures demonstrate the economic importance of the chlorine industry.

Figure 75 shows how world chlorine production has developed from 1965 to 1995. The figures are only estimates, as many important nations do not publish this information.

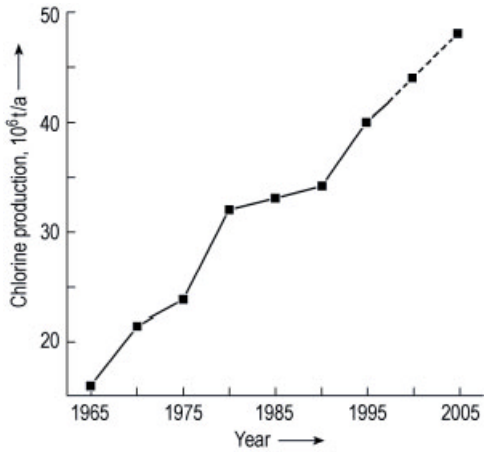

Figure 75. World chlorine production (1965 – 1995)

Chlorine capacities 1995 (in 10^3 t/a) in the main producer countries was as follows [290]:

United States	11 860
Japan	4250
Former Soviet Union	3800
China	3750
Germany	3690
Brazil	1660
Canada	1630
France	1580
United Kingdom	1270
India	1230
Italy	1130

World chlorine capacity was about 40×10^6 t/a in 1983, it rose to 48×10^6 t/a in 1998, and is forecast to reach 55×10^6 t/a by 2005 [283].

Figure 76 demonstrates the growth of chlorine capacity by economic regions [291]. Compared to 1983, the capacities in 1997 remained nearly constant in Western Europe, Eastern Europe, and Africa. In North America, reductions

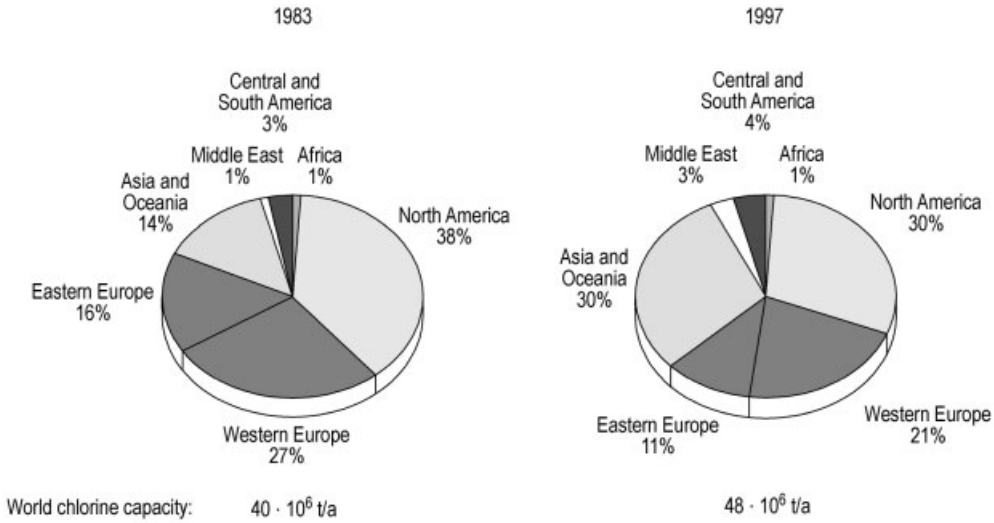

Figure 76. Chlorine production capacities by economic regions (1983 and 1997)

in Canada were outweighed by growth in the United States. Huge capacities were erected in the Middle East and above all in Asia and Oceania (China, India), where capacities were tripled. From the growth between 1997 and 2005, North America's share will be 35 %, Asia's 40 %, and the Middle East's 15 % [283]. After 2000 China will become the second biggest producer after the United States, with an annual chlorine capacity of about 5.5×10^6 t.

All chlorine producers are listed in [292]. A detailed review of the chlor-alkali market for 1994 is given in [293].

The production costs of 1 electrochemical unit ($\triangleq$ 1 t chlorine + 1.13 t sodium hydroxide) depends up to 60 % on the price of electricity. At 3.5 *d*/kWh, they are about 250 \$/ecu (Fig. 77), so the total value of production is ca 24×10^9 \$/a. The chlor-alkali industry is one of the biggest consumers of electrical energy, consuming ca. 0.15×10^{12} kWh/a.

A prediction of the market prices for chlorine and for caustic soda is very difficult. Firstly, most chlorine is used captively to avoid transportation, or it is sold on the basis of long-term contracts. In relation to production, the amounts on the spot markets are small. Secondly, the production of chlorine is strictly coupled to that of the caustic soda. A strong demand for chlorine creates a surplus of caustic, leading to high prices for the chlorine and a drop in prices for caustic; weak demand for chlorine reduces the available amount of caustic, leading to high caustic prices and low chlorine prices. This situation seems to change in cycles every 6 to 8 years. For example, from 1988 to 1991 the market value for chlorine fell from 150 \$/t down to 0, but from March 1992 to August 1993, the price rose to 230 \$/t. The price for caustic was 320 \$/t in 1991; it fell to 50 \$/t in March 1994 and climbed to 270 \$/t in December 1994 (Fig. 77) [294].

Because of the fixed ratio of chlorine to caustic in production and the different location of uses, worldwide trade flows have been created for both products. Chlorine is traded as chlorinated derivatives, particularly as EDC, VCM, and PVC, accounting for 85 %, and chlorinated solvents. The equivalent annual amount of chlorine is approximately 4.5×10^6 t (1995 to 2000). The main exporting countries are the United States (to South East Asia, South America), Russia (to West Europe), West Europe (to South East Asia, China), Japan (to China), and Middle East Asia (to Japan , China). [295].

16. Toxicology

Chlorine gas is dangerous to health because it is a powerful oxidant. In the physiological pH range, it is converted to hypochlorous acid, a cytotoxic substance. The extent to which the cells are damaged depends on the gas concentration, the exposure time, the water content of the tis-

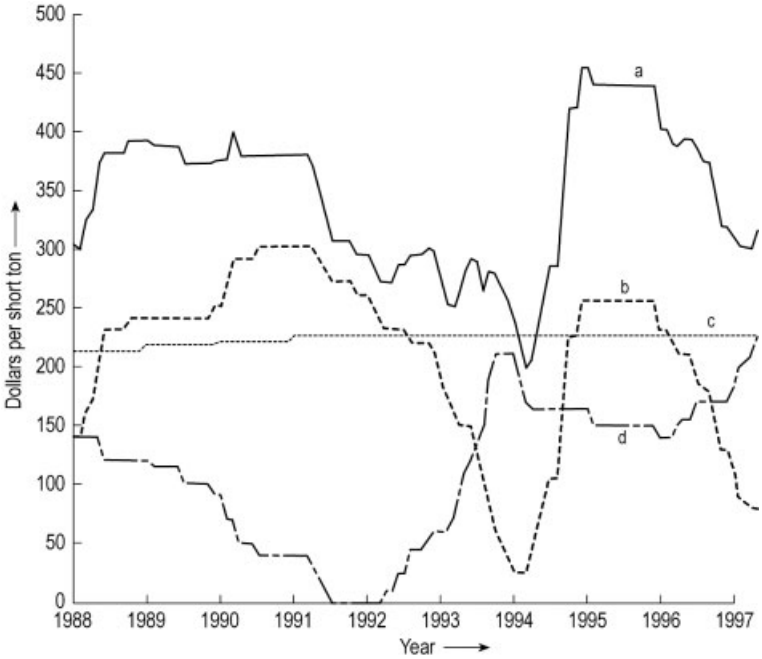

Figure 77. Production costs and product prices in the United states from 1987 to 1997 [294]
a) Product value for 1 ecu (1 t Cl_2 + 1.13 t NaOH; b) Price for caustic soda; c) Production costs for 1 ecu at 3.5 ¢/kWh electricity price; d) Price for chlorine into PVC

sue, and the health of the person exposed to the gas. Besides getting into the eyes, larynx, and trachea, chlorine also reaches the bronchi and the bronchioles. Because of its moderate solubility in water, chlorine affects the alveoli only at high concentrations. In this respect, chlorine differs from other gases with low water solubility and high lipid solubility, such as phosgene, nitrogen monoxide, and nitrogen dioxide. Initial moderate bronchial irritation is followed by the development of a toxic pulmonary edema because of increased alveolar injury.

The olfactory threshold of chlorine gas is 0.2 – 3.5 mL/m^3. Prolonged exposure seems to raise the olfactory threshold. Concentrations of 3 – 5 mL/m^3 are tolerated for up to 30 min without any subjective feeling of malaise. At concentrations between 5 and 8 mL/m^3, mild irritation of the upper respiratory tract and the conjunctiva is observed. In addition, running of the eyes and coughing are observed at concentrations of 15 mL/m^3 and higher. Above 30 mL/m^3, the following symptoms are observed: nausea, vomiting, oppressive feeling, shortness of breath, and fits of coughing, sometimes leading to bronchial spasms. Exposure to 40 – 60 mL/m^3 leads to the development of toxic tracheobronchitis. After a latent period of several hours with fewer symptoms, pulmonary edema may occur because of alveolar membrane destruction. This is indicated by increased shortness of breath, restlessness, and cyanosis. Subsequently, a further complication can occur after several days in the form of pneumonia caused by superinfection of the injured pulmonary tissue.

A clear dose – effect relationship of chlorine gas at different concentrations in humans has not yet been published. On the basis of results obtained from animal experiments, the LC_{50} for healthy humans is assumed to be 300 – 400 mL/m^3 at a 30-min exposure [296]. No deaths have occurred in animal experiments at 30-min exposures for concentrations below 50 mL/m^3. Death following acute intoxication is caused by a fulminant pulmonary edema.

Many investigations deal with the toxicity of *low chlorine concentrations.* Investigations in humans indicate the possibility of reversible damage to the lung function parameters at concentrations starting at 1.0 mL/m^3, whereas 0.5 mL/m^3 is a no-observable-effect-level [297, 298]. In monkeys no effects have been observed

at 0.5 mL/m^3 in a one-year study [299]. Long-term investigations of workers exposed to chlorine, e.g., in chlor-alkali electrolysis plants or in pulp manufacture, however, do not indicate increased rates of mortality or morbidity caused by pulmonary diseases [300 – 303].

No indications of *carcinogenicity* or *mutagenicity* of chlorine have been detected in animal experiments or encountered in industrial medicine [304]. In Germany, the MAK value is 0.5 mL/m^3 (1.5 mg/m^3) with a 15-min excursion factor of 1 [304]; in the USA, the TLV is 0.5 mL/m^3 (1.5 mg/m^3) with a STEL of 1.0 mL/m^3 (3 mg/m^3).

17. Acknowledgement

The entire topic was coordinated by Peter Schmittinger

18. References

General References

1. K. J. Vetter: *Elektrochemische Kinetik*, Springer Verlag, Berlin-Heidelberg 1961.
2. J. S. Sconce: *Chlorine, Its Manufacture, Properties and Uses*, Reinhold Publ. Co., New York 1962.
3. *Kirk-Othmer*, vol. **1**, 1978, pp. 799 – 883.
4. P. Gallone, G. Milazzo: *Electrochemistry*, Elsevier, Amsterdam-London-New York 1963.
5. F. Matthes, G. Wehner: *Anorganisch-Technische Verfahren für die Grundstoffindustrie*, VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1964.
6. *Gmelin*, system no. 6, Chlor (1927).
7. *Winnacker – Küchler*, 3rd ed., vol. **1**, pp. 228 – 239; *Winnacker – Küchler*, 4th ed., vol. **2**, pp. 379 – 480.
8. R. Powell: *Chlorine and Caustic Soda Manufacture*, Noyes Data Corp., Park Ridge, N.J. ,1971.
9. R. G. Smith: *Chlorine; An Annotated Bibliography*, The Chlorine Institute, New York 1971.
10. A. Kuhn: *Industrial Electrochemical Processes*, Elsevier, Amsterdam 1971.
11. P. Gallone: *Trattato di ingegneria elettrochimica*, Tamburine Editore, Milano 1973.
12. D. W. F. Hardie, W. W. Smith: *Electrolytic Manufacture of Chemicals from Salt*, The Chlorine Institute, New York 1975.
13. A. Schmidt: *Angewandte Elektrochemie*, Verlag Chemie, Weinheim 1976.
14. Society of Chemical Industry: *Diaphragm Cells for Chlorine Production*, London 1977.
15. J. J. McKetta, W. A. Cunningham: *Encyclopedia of Chemical Processing and Design*, vol. **7**, Marcel Dekker, New York-Basel 1977.
16. R. Dandres: *Le Chlore*, 3rd ed., Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionelles, Paris 1978.
17. M. O. Coulter: *Modern Chlor-Alkali Technology*, vol. **1**, Ellis Horwood, London 1980.
18. C. Jackson: *Modern Chlor Alkali Technology*, vol. **2**, Ellis Horwood, Chichester, Great Britain, 1983.
19. K. Wall, *Modern Chlor-Alkali Technology*, vol. **3**, Ellis Horwood, Chichester, Great Britain, 1986.
20. N. M. Prout, *Modern Chlor-Alkali Technology*, vol. **4**, Elsevier Appl. Science, London 1990.
21. T. C. Wellington, *Modern Chlor-Alkali Technology*, vol. **5**, Elsevier Appl. Science, Barking 1992.
22. M. G. Beal in R. W. Curry (ed.): *Modern Chlor-Alkali Technology*, vol. **6**, SCJ, London 1995.
23. S. Sealey: *Modern Chlor-Alkali Technology*, vol. **7**, The Graham House, Cambridge 1998.
24. The Chlorine Institute: Pamphlet 1 *Chlorine Manual*, Ed. 6, 1997.
25. Int. Soc. Electrochem.: *Extended Abstracts*, 23rd Meeting, Stockholm 1972.
26. The Electrochemical Society: *Extended Abstracts of Industrial Electrolytic Division*, The Chlorine Institute, Spring Meeting, San Francisco, May 12nd – 17th, 1974.
27. Int. Soc. of Electrochemistry: *Extended Abstracts*, 27th Meeting, Zürich 1976.
28. Technischer Arbeitskreis Elektrolyse: *Hinweise für den Umgang mit Chlor*, Verlag Chemie, Weinheim 1978.
29. The Electrochemical Society: *Extended Abstracts, Spring Meeting Seattle*, Princeton, N.J., May 1978.
30. AIChE: Sympos. Ser. **75**, no. 185, 1979.
31. Oronzio de Nora Symposium: *Chlorine Technology*, Venezia, May 1979.
32. AIChE: Sympos. Ser. **77**, no. 204, 1981.

33. J. S. Robinson: *Chlorine Production Processes*, Noyes Data Corp., Park Ridge, N.J., ,1981.
34. Joint Chlorine Institute: *BITC Meeting, London, June 1982*, The Chlorine Institute, New York.
35. Elektrochemische Verfahrenstechnik, *Dechema Monographie* vol. **97**, Verlag Chemie, Weinheim 1984.
36. Technische Elektrolysen, *Dechema Monographie* vol. **98**, Verlag Chemie, Weinheim 1985.

Specific References

37. C. W. Scheele, *Sv. Akad. Handl.* **35** (1774) 88.
38. J. Eidem, L. Lunevall in [26]
39. R. E. Means, T. R. Beck, *Chem. Eng. (N.Y.)* **91** (1984) no. 22, 46 – 51.
40. The Chlorine Institute: Properties of Chlorine in SI Units, *Pamphlet 72*, 2nd ed., July 1986.
41. J. J. Martin, D. M. Longpre, *J. Chem. Eng. Data* **29** (1984) 466 – 473.
42. L. Ziegler, *Chem. Ing. Techn.* **22** (1950) 229.
43. E. M. Aleta, P. V. Roberts, *J. Chem. Eng. Data* **31** (1986) 51 – 54.
44. I. A. A. Ketalaar, *Electrochem. Technol.* **5** (1967) 143.
45. R. F. Schwab, N. H. Doyle, *Electrochem. Technol.* **5** (1967) 228.
Y. E. Frolov, A. S. Maltsera, A. N. Baratov, A. I. Roslovskii, *Khim. Promst (Moscow)* **7** (1977) 530 –532.
46. The Chlorine Institute: Explosive Properties of Gaseous Mixtures Containing Hydrogen and Chlorine, *Pamphlet 121*, 2nd ed., June 1992.
47. IUPAC, *White Book on Chlorine* **68** (1996) 9.
48. V. M. Sedivy, *Industrial Minerals*, **343** (1996) April, 73.
49. Hoechst, EP 35 695-B1, 1983.
50. Solvay & Cie., EP 30 756-A1, 1981.
51. A. I. Postoronko et al., SU 537 027-A, 1976.
Jamestown Chem.: Calcium Sulphate Inhibitor.
52. S. L. Fuks, T. M. Orchinnikova, *Zh. Prikl. Khim (Leningrad)* **56** (1983) no. 8, 1757 – 1761.
53. K. Harada, Cleveland Meeting: The Electrochem. Soc., May 1966.
54. Olin, US 4 277 447-A, 1981 (R. F. Chambers, N. Rachima).
55. Toa Gosei Chem. Ind., JP 57–106 520-A2, 1982 (K. Suzuki, R. Nagai, H. Goto).
56. Hoechst, DE 3 037 818-A1, 1982 (S. Benninger).
57. Bayer, DE 3 216 418-A1, 1983 (R. Schäfer).
58. Bayer, DE 2 709 728-C2, 1981 (R. Schäfer).
59. K. Maycock et al., SCI Chlorine Symposium, London, June 1997.
60. Hoechst, DE 2 905 125-A1, 1980 (M. Schott, J. Russow).
61. Z. I. Lifatova et al., SU 865 800-A1, 1981.
62. Bayer, DE 1 177 122-B, 1964 (E. Zirngiebl).
63. Ebara Infilco, JP 54–81 174-A2, 1979 (M. Kanekawa, K. Akagi).
64. *Processing*, June 1978.
65. W. Glas, *Chem. Ing. Tech.* **43** (1973) 832.
66. F. Lüns, *BBC-Nachrichten*, 1/2 (1974) 36 – 42.
67. H. A. Horst, *Chem. Ing. Tech.* **43** (1971) 164.
68. Siemens, DE 3 123 877-C1, 1983 (G. Achilles).
Hundt & Weber Schaltgeräte, EP 11 820-A1, 1980 (E. Stratmann).
69. F. Lappe, *Chem. Ing. Tech.* **42** (1970) 1228.
70. Fiat Final Report 732, 797, 816, 834.
71. W. Kramer in [36]
72. T. Sugino, K. Aoki, *J. Electrochem. Soc. Jpn.* **27** (1959) no. 1 – 3, E 17.
73. G. Hauck, W. Dürr, *Chem. Ing. Tech.* **39** (1967) 720.
H. J. Antweiler, J. P. Schäfer, *Chem. Ing. Tech.* **43** (1971) 180.
74. D. S. Nielson, O. Platz, *Reliabil. Eng.* **4** (1983) 1 – 18.
75. PPG Ind., US 4 155 819-A, 1979 (W. W. Carlin).
76. Diamond Shamrock, US 4 073 706-A, 1978 (Z. Nagy).
77. H. Tsukuda et al., *J. Electrochem. Soc. Jpn.* **32** (1964) 1.
78. D. W. Wabner, P. Schmittinger, F. Hindelang, R. Huß, *Chem. Ing. Tech.* **49** (1977) 351.
79. L. J. J. Janssen in [18], p. 271.
80. Society of Chemical Industry, Electrochemical Technology Group: *Intern. Meeting on Electrolytic Bubbles*, 13 – 14 Sept. 1984, London.
81. BASF, US 3 502 561, 1970 (W. Rasche, E. Wygasch, G. Csizi).
82. A. Zimmer, L. Franke, *Chem. Tech. (Leipzig)* **36** (1984) no. 2, 64 – 69.
83. H. Hund, *Chem. Ing. Tech.* **39** (1967) 702.
84. W. Vielstich, *Chem. Ing. Tech.* **34** (1962) 346.
Bayer, DE 1 792 588-C3, 1977 (E. Zirngiebl).
Murgatroyd's Salt & Chem., DE 1 915 765, 1969 (M. J. Lockett).
85. M. M. Jaksic, I. M. Csonka, *Electrochem. Technol.* **4** (1966) 49.
86. G. Barthel, *Chem. Anlagen + Verfahren* **8** (1973) 59 – 60.

87. P. Theissing et al., *Chem. Ing. Tech.* **51** (1979) 237.
F. Hine: *Topics in Pure and Applied Electrochemistry*, SAEST, Karaikudi, India.
88. Uhde: *Alkalichloridelektrolyse nach dem Quecksilberverfahren*.
89. Olin, US 4 004 989-A, 1977 (R. W. Ralston).
90. G. Hauck, *Chem. Ing. Tech.* **34** (1962) 369.
91. Euro Chlor: *Control of Worker Exposure to Mercury in the Chlor-Alkali Industry*, 4th ed., October 1997.
92. The Chlorine Institute: Pamphlet 125, Guidelines: *Medical Surveillance and Hygiene Monitoring Practices for Control of Workers Exposure to Mercury in the Chlor-Alkali Industry*, 3rd ed., December 1994.
93. Environmental Protection Agency: *Control Techniques for Mercury Emissions from Extraction and Chlor-Alkali Plants*, Publ. No. AP-118, Washington, DC.
94. A. Gellera, L. Cavalli, G. Nucci, *Analyst (London)* **109** (1984) 1537 – 1539.
95. H. Klotz, *Chem. Ing. Tech.* **52** (1980) 444.
96. Akzo: *Anode Current Control by Computer ACCS-System*, 1972.
97. A. I. Vasilev et al., SU 1 024 528-A1, 1983.
98. Bayer, DE 1 767 840-B2, 1975 (R. Schlee, W. Büsing).
Hoechst, DE 3 244 033-A1, 1984 (K. Lehmann, H. Valentin, K. Gnann).
99. Montedison, EP 93 452-A2, 1983 (F. Lo Vullo, E. Malvezzi, P. Balboni).
100. Solvay & Cie., EP 85 999-A1, 1983 (J. P. Detournay, J. Defourny).
101. Heraeus Elektroden, EP 68 076-A2, 1983 (P. Fabian).
102. H. Becker, *Dechema Monographie*, vol. **61**, Verlag Chemie, Weinheim 1968, p. 59.
103. BITC Working Group Mercury Analysis, *Anal. Chim. Acta* **72** (1974) 37 – 48;
84 (1976) 231 – 257;
87 (1976) 273 – 281.
104. AKZO: *Das IMAC-TMR-Verfahren*.
105. EEC Directive 82/176, 22. 3. 1982.
106. Euro Chlor Publ. Env. Prot. 11, *Mercury Housekeeping*, 4th ed., September 1998.
107. Euro Chlor: *Good Mercury Housekeeping* Third Techn. Seminar, Proceedings, December 1993.
108. Parcom Decision 90/3, Reducing Atmospheric Emissions from Existing Chlor-Alkali Plants, June 14, 1990.
109. Euro Chlor: *Chlorine Industry Review 1996 – 1997*, p. 10.
110. A. A, Lindley, An Economic and Environmental Analysis of the Chlor-Alkali Production Process, prepared for the European Commision (DG III, Chemicals), Final Report, June 30, 1997.
111. Euro Chlor, GEST 97/228, The Eurochlor Program, Fourth Techn. Seminar, Proceedings, Feb. 1997.
112. ELTECH Systems Corporation Technical Service Department: *Diaphragm Electrolytic Chlorine Cell, Technical Data Manual*, ELTECH Systems Corporation, Chardon, Ohio 1982.
113. ELTECH Systems Corporation Technical Service Department: *Diaphragm Chlorine Cell, Operating Manual*, ELTECH Systems Corporation, Chardon, Ohio 1983.
114. ELTECH Systems Corporation, US 4 410 411, 1983 (R. W. Fenn, III, E. J. Pless, R. L. Harris, K. J. O'Leary).
115. ELTECH Systems Corporation, US 3 674 676, 1972 (Edward I. Fogelman).
116. L. C. Curlin, T. F. Florkiewicz, R. C. Matousek: "Polyramix, A Depositable Replacement for Asbestos Diaphragms", London International Chlorine Symposium, 1988.
117. R. N. Beaver: *The Dow Diaphragm Cell*, Dow Chemical, Freeport, Texas 1985.
118. Dow Chemical Company, US 4 497 112.
119. Oronzio De Nora Impianti Elettrochimici S.p.A.: Design Features of Glanor Diaphragm Electrolyzer, Oronzio DeNora Impianti Elettrochimici S.p.A., Milano.
120. PPG Industries: *PPG's Bipolar Diaphragm Electrolyzers*, PPG Industries, Inc., Pittsburgh, Pa.
121. PPG Industries: *Licensing From PPG Chemicals*, PPG Industries, Inc., Pittsburgh, Pa.
122. Occidental Chemical Corporation: *Hooker Chlor-Alkali Systems*, A Record of Achievement in Diaphragm Cell Technology, Occidental Chemical Corporation, Niagara Falls, N.Y.
123. Uhde GmbH: *Alkaline Chloride Elektrolysis by the Diaphragm Process; System Hooker*, Uhde GmbH, Dortmund.
124. ELTECH Systems Corporation: *ELTECH Modified Diaphragm Cells*, ELTECH Systems Corporation, Chardon, Ohio 1985.
125. ELTECH Systems Corporation, US 3 591 483, US 3 707 454, 1971 (Richard E. Loftfield, Henry W. Laub).

126. ELTECH Systems Corporation, US 3 928 166, 1975 (Kevin J. O'Leary, Charles P. Tomba, Robert W. Fenn, III).
127. ELTECH Systems Corporation: Caustic Purification System 1985.
128. G. A. Carlson, E. E. Estep, "Porous Cathode Cell for Metals Removal from Aqueous Solutions", paper presented at the Electrochemical Society Meeting, May, 1972.
129. W. G. Grot, *Chem. Ing. Tech.* **44** (1972) 167. Diamond Shamrock, US 4 025 405, 1977 (R. L. Dotson, K. J. O'Leary).
130. Asahi Glass, US 4 065 366, 1977 (Y. Oda, M. Suhara, E. Endoh).
131. H. Ukihashi, M. Yamabe, ACS Polymer Division Workshop on Perfluorinated Ionomer Membranes, Lake Buena Vista, Florida, Feb. 23 – 26, 1982.
132. *World Chlor-alkali Analysis*, Vol. **1**, CMAI, Texas, USA May 1997.
133. T. A. Davis, J. D. Genders, D. Pletcher, *Ion permeable membrane*, Alresford 1997.
134. Asahi Chemical Industry, US 4 178 218, 1979 (M. Seko).
135. K. Schneiders, J. Schulze: Improvement in the Hoechst-Uhde Membrane Cell, *Modern Chlor-Alkali-Technology*, Vol. **6**, 1995.
136. Akin, Optimierung einer Chloralkalielectrolyse-Membranzellen-Kaskade unter Berücksichtigung ihres Energieverbrauchs, Dissertation, Dortmund 1984.
137. Tokuyama Soda, US 4 190 514, 1980 (S. Matsuura, T. Oku, M. Kuramatani, N. Kramoto, Y. Ozaki).
138. PPG, US 4 323 595, 1982 (C. N. Welth, J. O. Snodgrass).
139. H. S. Burney: Chlor-Alkali Process, in Ralph. E. White et al. (eds.), *Modern Aspects of Electrochemistry*, no. **24**, Plenum Press, New York, 1993.
140. Du Pont: Recommended Brine Impurity Limits for Chloralkali Plants, Aug. 1991.
141. Bayer AG: The Importance of Electrolyte Quality to Membrane Chloralkali Cells, Uhde Chlorine Symposium 1995 (F. R. Minz).
142. Du Pont: Effect of Impurities on Membrane for Chloralkali Production, Wilmington, Delaware Aug. 1989 (James T. Keating).
143. Rolf Rodermund: Chlorat-Reduktion im Solekreislauf einer Chlor-Alkali-Electrolyse, Dissertation, Dortmund 1983.
144. Asahi Glass, US 4 255 523, 1981 (H. Ukihashi, T. Asawa, T. Gunjima).
145. Du Pont, US 4 021 327, 1977 (W. G. Grot).
146. Asahi Glass, JP 59/40 231, 1984 (Y. Oda, T. Morimoto, K. Suzuki).
147. The Dow Chemical Company: *Membrane Chlor-Alkali Process*, Texas Applied Science and Technology Laboratories, Freeport, Texas, 1993 (H. S. Burney).
148. D. Bergner, *Chem.-Ztg.* **107** (1983) 281.
149. H. B. Beer, GB 1 147 442, 1965. H. B. Beer, GB 1 195 871, 1967.
150. V. de Nora, J.-W. Kühn von Burgsdorff, *Chem. Ing. Tech.* **47** (1975) 125 – 128. Electronor Corporation, US 3 616 445, 1967 (G. Bianchi, V. de Nora, P. Gallone, A. Nidola).
151. S. Trasatti, G. Lodi in S. Trasatti (ed.): *Electrodes of Conductive Metallic Oxides*, part A, Elsevier Scientific Publishing Company, Amsterdam-Oxford-New York 1980, pp. 301 – 358.
152. S. Trasatti, W. E. O'Grady in H. Gerischer, C. W. Tobias (eds.): *Advances in Electrochemistry and Electrochemical Eng.*, vol. **12**, J. Wiley & Sons, New York-Chichester 1981, pp. 177 – 261.
153. GB 1 402 414, 1971 (N. W. J. Pumphrey, B. Hesketh). GB 1 484 015, 1973 (B. Hesketh, C. Pownall, N. W. J. Pumphrey).
154. C. Conradty, DE-OS 1 813 944, 1968 (G. Thiele, D. Zöllner, K. Koziol); C. Conradty, DE 2 255 690, 1972 (K. Koziol, K.-H. Sieberer, H.-C. Rathjen).
155. C. Modes, W. C. Heraeus: abstract and poster during the *13th Int. Congress and General Assembly,* Int. Union of Crystallography, Hamburg, August 1984.
156. Yu. E. Roginskaya, O. V. Morozova, *Electrochimica Acta* **40** (1995) 817 – 822.
157. D. V. Kokoulina, T. V. Ivanova, Y. I. Krasavitskaya, Z. I. Kudryavtseva et al., *Elektrokhimiya* **13** (1977) no. 10, 1511 – 1515.
158. J. E. Currey in *Encycl. of Chem. Proc. and Design*, vol. **7**, Marcel Dekker, New York 1978, pp. 305 –450.
159. D. Bergner, S. Kotowski, *J. Appl. Electrochem.* **13** (1983) 341 – 350.
160. Y. M. Kolotyrkin, *Denki Kagaku* **47** (1979) no. 7, 390 – 400.
161. D. W. Wabner, P. Schmittinger, F. Hindelang, R. Huß, *Chem. Ing. Tech.* **49** (1977) 351.
162. B. Busse, R. Scannell, Electrochemische Verfahrenstechnik-Energietechnik, Stoffgewinnung und Bioelektrochemie, Mono. GdCh. Tagung Monheim 1996.

163. F. Beck, *Electrochimica Acta* **34** (1989) 811 – 822.
164. Oronzio De Nora Impianti Elettrochimici S.p.A., GB 2 051 131, 1980 (A. Pellegri).
165. A. Nidola in S. Trasatti (ed.): *Electrode of Conductive Metallic Oxides*, part B, Scientific Publishing Company, Amsterdam-Oxford-New York 1981, pp. 627 – 659.
166. C. Traini, Proc. of O. De Nora Symposium, 15 – 18 May 1979.
167. Diamond Shamrock Corp., US 3 674 676, 1970 (E. I. Fogelman).
168. V. de Nora, *Chem. Ing. Tech.* **47** (1975) no. 4, 141.
169. Heraeus Elektrochemie, EP 0 687 312, 1994 (B. Busse, R. Scannell).
170. S. Kotowski, B. Busse, "Anodenbeschichtungen in Membranzellen zur Chlor- und Laugeproduktion", paper presented at the Annual DECHEMA Meeting 1984.
171. Uhde GmbH, DE 3 219 704, 1982 (H. Schmitt, H. Schurig, W. Strewe).
172. N. P. Fedot'ev, N. V. Berezina, E. G. Kruglova, *Zh. Prikl. Khim. (Leningrad)* **21** (1948) 317 – 328.
173. Toa Gosei Chemical Industry Co. Ltd., JP 31–6611, 1956 (K. Sasaki).
174. E. I. DuPont de Nemours & Co., US 4 116 804, 1978 (C. R. S. Needes).
175. Diamond Shamrock Corp., US 4 024 044, 1977 (J. R. Brannan, I. Malkin).
176. F. Hine, M. Yasuda, M. Watanabe: "Studies of the Nickel – Sulfur Electrodeposited Cathode", *Denki Kagaku* **47** (1979) no. 7, 401 – 408.
177. Asahi Chemical, EP-A 0 031 948, 1981 (M. Yoshida, H. Shiroki).
178. Chlorine Engineers Corp. Ltd., US 4 465 580, 1984 (K. Kasuya).
179. Tokuyama Soda, US 4 190 516, 1980 (Y. Kajimaya, T. Kojima, Y. Murakami, S. Matsura).
180. Dow Chemical, EP-A 0 129 734, 1985 (N. R. Beaver, L. E. Alexander, C. E. Byrd).
181. ICI, EP-A 0 129 374, 1984 (J. F. Cairns, D. A. Denton, P. A. Izard).
182. Diamond Shamrock Corp., US 4 104 133, 1978 (J. R. Brannan, I. Malkin, C. M. Brown).
183. Kanegafuchi Kagaku Kogyo Kabushiki Kaisha, EP-A 0 132 816, 1985 (Y. Samejima, M. Shiga, T. Kano, T. Kishi).
184. Olin Corp., US 4 169 775, 1979 (H. C. Kuo).
185. Harrimann Chemsult: Chlor-Alkali Prospects for Chlorine, Caustic Soda and Soda Ash in the 1990's, London 1991.
186. Th. F. Florkiewicz, The Future of Diaphragm Cell Technology, Chlorine Institute Spring Meeting, St. Louis, Missouri, March 1998.
187. M. Esayian, J. H. Austin: "Membrane Technology for Existing Chloralkali Plants", E. I. du Pont de Nemours Co., Wilmington, Del., 1984.
188. VCI Verband der Chemischen Industrie (ed.): *Zukunft der Alkalichlorid-Electrolyse-Anlagen nach dem Amalgamverfahren*, Frankfurt a. M. 1998.
189. M. R. Marcan: "Mercury/Membrane Cells – when to Change", *Modern Chlor-Alkali Technology*, vol. **5**, Elsevier Publ., London, 1992, p. 177.
190. R. W. Curry: "Issues involved in membrane conversion of a chlor-alkali business", *Modern Chlor-Alkali Technology*, vol. **5**, Elsevier Publ., London 1992, p. 185.
191. S. Higuchi, *Soda to Enso* **35** (1984) 383.
192. H. Isfort, W. Stockmans in [36]
193. K. Kerger, *Chem. Ing. Tech.* **43** (1971) 167.
194. Uhde: Chlor und Wasserstoff aus Salzsäure durch Elektrolyse.
195. General Electric, DE 2 856 882-A1, 1979 (R. M. Dempsey, A. B. Conti).
196. H. Klotz, P. Orth in [36]
197. The M. W. Kellogg Co.: Hydrogen Chloride to Chlorine, The Kel-Chlor Process, presentation at the 20th Chlorine Plant Managers Seminar, New Orleans, Feb. 9 1977.
198. Krupp Uhde: "Chlorine and Hydrogen from Acid by Electrolysis", company brochure, 1993.
199. *Chemical Week* 1996, 14. Aug., 26. M. Mortensen et al., *Chem. Eng. Sci.* **51** (1996) 10, 2031.
200. The M. W. Kellogg Co., *Hydrocarbon Process.* 60 (1981) no. 11, 143.
201. Shell, BE 599 241-A, 1961 (W. S. Engel, F. Wattimena).
202. Mitsui Chemicals Inc.: "The Mitsui MT-Chlor Process, Catalytic Oxidation of Hydrogen Chlordie", company brochure.
203. M. Mortensen et al., *Chem. Engng. Scie.* **51** (1996) 2031.
204. Institut Français du Pétrole, FR 1 294 706-A, 1962 (P. Bedagne et al.).
205. C. P. van Dijk, *Chem. Econ. Eng. Rev.* **4** (1972) no. 12, 42.
206. H. Schmidt, F. Holzinger, *Chem. Ing. Tech.* **35** (1963) 37.
207. Monsanto Enviro-Chem. Systems: *Fiber Bed Mist Eliminator*, 1979.

208. Dow Chemical, US 3 568 409-A, 1971 (B. Mac Ferguson, J. F. Gilbert, D. N. Glew).
209. *CJ 134 "The Drying and Liquefication of Chlorine and the Phase Diagram $Cl_2 - H_2O$"*, April 1967.
210. Koninkl. Nederlandsche Zoutindustrie, US 3 534 562-A, 1970 (D. Meyer, T. Thijssen).
211. ICI, DE 1 946 096-C3, 1973 (K. J. Howliston).
212. G. Alberti, *DECHEMA Monogr.* **66** (1971) 1193 – 1211.
R. Gabbioneta: *Garo*.
Nash International: HC-7 Chlorine Compressor.
213. Sulzer Burckhardt, Prospectus No. 1393 a.
214. Sundstrand: Sundyne Zentrifugalverdichter, 1980.
215. Euro Chlor Techn. Publ. 79/79 *Chlorine Transfer Compressors*, 3rd ed. February 1999.
216. Euro Chlor Techn. Publ. 83/119 *Canned Pumps for Use with liquid Chlorine*, 3 rd ed. September 1987.
217. H. Abel, F. Özvegyi, *Kälte-Klima-Rundsch.* **7** (1969) 100.
218. *Inf. Chim.* **98** (1971) 55.
219. H. Hagemann, *Chem. Ing. Tech.* **39** (1967) 744.
220. R. F. Schwab, W. H. Doyle, *Electrochem. Technol.* **5** (1967) 228.
221. B. Liggenstorfer: "Kälteanlage zum Verflüssigen von Chlor", *Tech. Rundsch. Sulzer* **63** (1981) no. 2, 53 – 56.
222. AKZO: *Chlorine Liquefaction by Absorption and Distillation*, 1974.
223. EC Position Paper PP VII: "Use of Carbontetrachloride in Chlorine Production", Feb. 1995.
224. T. A. Liederbach, *Chem. Eng. Prog.* **70** (1974) 64.
225. Euro Chlor, Ave E Van Nieuwenhuyse 4, Box 2, B-1160 Brussels, Fax: +32 2 6767241, e-mail: eurochlor@cefic.be: Publications.
226. *Technical and Safety Information, catalog 1997/98*, The Chlorine Institute, Inc., 2001 "L" Street, N.W. Suite 506, Washington D.C. 20036, Fax (202) 223–7225.
227. R. Dandres: *Le Chlore*, 3rd ed., L'Institut National de Recherche et de Sécurite pour la Prevention des Accidents du Travail et de Maladies Professionelles, Paris 1978.
H. Dorias: *Gefährliche Güter*, Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1984.
228. G. Payne: "Safe Handling and Storage of Liquefied Gases, in Particular Chlorine and Chlorofluorohydrocarbon Refrigerants", *Trans. Inst. Chem. Eng.* **42** (1964) no. 4, 92 – 99.
229. D. P. Meinhardt: "Chlorine", *Chem. Eng.* (1981) Oct. 5, 125 – 133.
230. N. C. Harris, J. P. Shaw, *European Chlorine Storage Practice*, paper presented at the Chlorine Institute's 23rd Plant Manager Seminar, 1980.
R. Papp, *Chlorine '74*, Chemistry and Industry, 15 March (1975) 243 – 246.
231. Euro Chlor Techn. Publ. 75/46, *Pneumatically Operated Valves for Use on Rail and Road Tankers, and ISO-Containers for Liquid Chlorine*, 8th ed., November 1996.
232. The Chlorine Institute, Pamphlet 93, *Pneumatically Operated Valves for Use on Tank Cars*, Ed. 1, September 1996.
233. *Winnacker-Küchler*: 4th ed., pp. 428 – 429.
234. Euro Chlor Techn. Publ. 72/10, *Pressure Storage of Liquid Chlorine*, 8th ed., June 1993.
235. The Chlorine Institute, Pamphlet 5, *Non-refrigerated Liquid Chlorine Storage*, Ed. 5,
October 1992.
236. Euro Chlor Techn. Publ. 73/17, *Low Pressure Storage of Liquid Chlorine*, 5th ed., February 1996.
237. The Chlorine Institute, Pamphlet 78, *Refrigerated Liquid Chlorine Storage*, Ed. 2, June 1994.
238. ISO 3511/1.
239. A. F. Timofeer et al.: "Steps to Prevent Failures of Chlorine Containers", *Khim. Promst (Moscow)* **10** (1978) no. 3, 202 – 205; *Sov. Chem. Ind. (Engl. Transl.)* **10** (1978) no. 3, 222 – 225.
240. Euro Chlor Techn. Publ. 79/81, *Dry Liquid Chlorine Piping Systems Located inside Producers' or Consumers' Plants*, 7th ed., September 1986.
241. The Chlorine Institute, Pamphlet 60, *Chlorine Pipelines*, Ed. 4, May 1997.
242. The Chlorine Institute, Pamphlet 9, *Chlorine Vaporizing Equipment*, Ed. 5, April 1997.
243. Euro Chlor, Techn. Publ. 75/47, *Chlorine Vaporizers*, 8th ed., January 1980.
244. Armstrong Chemtec, Form 54 – 2/97, *Vaporizers*.
245. Euro Chlor Techn. Publ. 76/55, *Acceptable Levels of Nitrogen Trichloride in Liquid Chlorine*, 9th ed., September 1990.
246. The Chlorine Institute, Pamphlet 21, *Nitrogen Trichloride – A Collection of Reports and Papers*, Ed. 4, 1997.
247. The Chlorine Institute, Pamphlet 89, *Chlorine Scrubbing Systems*, Ed. 1, December 1991.

248. Euro Chlor Techn. Publ. 76/52, *Equipment for the Treatment of Gaseous Effluents Containing Chorine*, 9th ed., October 1986.
249. M. E. Fakley et al., The ICI HYDECAT-Process in [21], p. 23.
250. Euro Chlor Techn. Publ. 79/82 *Choice of Materials of Construction for Use in Contact with Chlorine*, 7th ed., October 1995.
251. W. A. Statesir: "Explosive Reactivity of Organics and Chlorine", *Chem. Eng. Prog.* **69** (1973) no. 4, 52 – 54.
252. Euro Chlor Third Technical Seminar Proceedings, Techn. Publ. 93/192 *How to Use Steel and Titanium Safety*, December 1993.
253. The Chlorine Institute, Pamphlet 95, *Gaskets for Chlorine Service*, Ed. 1, October 1993.
254. T. B. Meslin: "Assessment and Management of Risk in the Transport of Dangerous Materials: The Case of Chlorine Transport in France", *Risk Anal.* **1** (1981) no. 2, 137 – 142.
G. B. Frame: "Determination of Risk from a Chlorine Spill or Major Release", *Chem. Can.* **32** (1980) no. 8, 27 – 29.
255. C. L. Melancon, *Dow Chemicals Emergency Response System*, paper presented at The Chlorine Institute's 22nd Plant Manager's Seminar 1979.
R. Sklarew et al., *Emergency System for Toxic Chemical Releases*, Pollution Engineering, July 1982.
256. J. M. Buchlin, *Aerodynamic Behaviour of Liquid Spray-Design Method of Water Spray Curtain*, von Karman Institute, Rhode Saint Genese, Belgium 1980, Test Report No. 171.
257. The Chlorine Institute, Pamphlet 63, *First Aid and Management of Chlorine Exposures*, Ed. 5, May 1993.
258. The Chlorine Institute, Pamphlet 64, *Emergency Response Plans for Chlorine Facilities*, Ed. 4, November 1995.
259. The Chlorine Institute, Pamphlet 65, *Personal Protective Equipment for Chlorine and Sodium Hydoxide*, Ed. 3, November 1995.
260. The Chlorine Institute, Pamphlet 85, *Recommendations for Prevention of Personnel Injuries for Chlorine Producer and User Facilities*, Ed. 3, November 1994.
261. The Chlorine Institute, Pamphlet 86, *Recommendations for Chlor-Alkali Manufacturing Facilities for the Prevention of Chlorine Releases*, Ed. 3, November 1994.
262. The Chlorine Institute, Pamphlet 74, *Estimating the Area Affected by a Chlorine Release*, Ed. 3, in revision.
263. The Chlorine Institute, Instruction Booklets, *Chlorine Institute Emergency Kit "A" for 100-lb. and 150-lb. Chlorine Cylinders, Kit "B" for Chlorine Ton Containers, Kit "C" for Chlorine Tank Cars and Tank Tucks*, Ed. 8, 1996.
264. Euro Chlor Chlorine Safety Manual 90/159, *Aspects of Transport Safety*, November 14 – 15, 1990.
265. Euro Chlor Techn. Publ. 92/176, *Chlorine Emergency Equipment*, 1st ed. September 1993.
266. Euro Chlor Techn. Publ. 93/191, 93/193, 93/194, "Safe Design", "Safe Constructions", "Safe Operations" in Eurochlor Third Technical Seminar Proceedings, December 2 – 3, 1993.
267. The Chlorine Institute, *Chlorep: How to Get Help in Chlorine Emergencies*.
268. British Standard Institution: *Specification for Liquid Chlorine*, BS 3947, 1965.
269. ISO 1552: Liquid Chlorine for Industrial Use — Method of Sampling.
270. The Chlorine Institute, Pamphlet 77, *Sampling Liquid Chlorine*, Ed. 3, April 1997.
271. ISO 2121: Liquid Chlorine for Industrial Use — Determination of the Water Content — Gravimetric Method.
272. ASTM – E410 *Standard Method of Test for Moisture and Residue in Liquid Chlorine*, 1986.
273. ISO 2202: Liquid Chlorine for Industrial Use — Determination of the Water Content — using an Electrolytic Analyzer.
274. ISO 2120: Liquid Chlorine for Industrial Use — Determination of the Content of Chlorine by Volume in the Vaporized Product.
275. ASTM – E412 *Standard Method of Assay of Liquid Chlorine*, 1986.
276. DIN 38 408 Teil 4: Gasförmige Bestandteile (Gruppe G) Bestimmung von freiem und Gesamtchlor (G 4).
277. *Anal. Chim. Acta*, **156** (1984) 221 – 233.
278. *Anal. Chim. Acta*, **87** (1976) 273 – 281.
279. ASTM – E506 *Standard Test Method for Mercury in Liquid Chlorine*, 1984.
280. *Deutsche Einheitsverfahren zur Wasseruntersuchung*, Verlag Chemie, Weinheim 1971.
281. M. L. Langhorst, *Am. Ind. Hyg. Assoc. J.* **43** (1982) 347 – 360.
The Chlorine Institute: *Atmospheric Monitoring Equipment for Chlorine*, 2nd ed.,

March 1982, CI Member Information Report 138.
W. König et al., *Chem. Ind.* (Feb. 1984) 69.
282. World Health Organisation, *"Recommendations for Drinking Water Quality"*, 2nd ed., Vol.**1**, December 1993.
W. König et al., *Chem. Ind.* **36** (Feb. 1984) 69.
283. Mary. C. Blackburn, Chemical Market Associates, Inc.: "Chlorine and Caustic – which Way is up", Krupp Uhde Symposium, Dortmund, May 6 – 8, 1998.
284. W. König, K. Krinke, H. Kron et al., *Chem. Ind.* **36** (Feb. 1984) 69.
285. Euro Chlor: Chlorine Industry Review 1994 – 1995.
286. *Chem. Engng. News*, 1994, Nov. 21, 17.
287. Euro Chlor, Proceedings of the Technical Seminar Leipzig "Care in Practise", Germany 1997.
288. Euro Chlor: Chlorine flow in Europe 1995, Jan. 1997.
289. Euro Chlor: Chlorine Industry review 1994 – 1995.
290. Harriman Chemsult, Chlor-Alkali Prospects for Chlorine, Caustic Soda and Soda Ash in the 1990's, Autumn 1991, London, England.
291. B. Hileman et al., *Chem. & Eng. News*, 21st November 1994, p. 12.
292. The Chlorine Institute, Pamphlet 10, *North American Chlor-Alkali Industry Plants and Production Data Book*, April 1997.
The Chlorine Institute, Pamphlet 16, *Chlor-Alkali Producers Outside North America* July 1997.
293. Chemical Economics Handbook – SRI International, CEH Marketing Report, "Chlorine/Sodium Hydroxide", 1996, USA, Zurich, Tokyo.
294. ANORGANICA Ltd, Monthly Newsletter "Soda – Chlorine" Issue Number 197, 29th May 1997.
295. Euro Chlor, GEST 97/230, D. Hutchinson: Derzeitige und zukünftige Tendenzen bei der Herstellung von Chlor, Natronlauge und Chlorderivaten, 4th Technical Seminar, Leipzig, 1997.
296. P. Davies, I. Hymes: "Chlorine toxicity criteria for hazard assessment", *Chem. Eng.*, June 1985, 30 – 33.
297. H. H. Rotman, M. J. Fliegelman, T. Moore, R. G. Smith, D. M. Anglen, C. J. Kowalski, J. G. Weg: "Effects of low concentrations of chlorine on pulmonary function in humans", *J. Appl. Physiol.: Respir. Environ. Excercise Physiol.* **54** (1983) no. 4, 1120 – 1124.
298. R. P. F. Schins, H. Emmen, L. Hoogendijk, P. J. A. Borm: "Nasal Inflammatory and Respiratory Parameters in Human Volunteers during and after Repeated Exposure to Chlorine", *Eur. Respir. J.* **16** (2000) 626 – 632.
299. "Chronic Inhalation Toxicity Study on Chlorine in Non-Human Primates", *CITT Activities Chemical Institute of Toxicology* **4** (1984) no. 8, 1 – 3.
300. B. Grenquist-Norden: "Respiratory effects of industrial chlorine and chlorine dioxide exposure", Institute of Occupational Health, University of Helsinki, Finland, 1983.
301. F. Schuckmann (Occupational Health Department of Hoechst AG): "Pulmonary Function Study on Workers with Long-Term Exposure to Chlorine", in: *Medichem proceedings, XI. International Congress, Calgary, Alberta, Canada,* 26. – 29. 9. 1983, pp. 475 –483.
302. B. G. Ferris, Jr., S. Puleo, H. Y. Chen: "Mortality and morbidity in a pulp and a paper mill in the United States: ten-year follow-up", *Br. J. Ind. Med.*, **36** (1979) no. 2, 127 – 134.
303. L. R. S. Patil, R. G. Smith, A. J. Vorwald, T. F. Mooney: "The health of diaphragm cell workers exposed to chlorine", *Am. Ind. Hyg. Assoc. J.* **31** (1970) 678 – 686.
304. H. Greim (ed.): "Chlor", *Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*, WILEY-VCH, Weinheim 2004.

Chlorine Oxides and Chlorine Oxygen Acids

HELMUT VOGT, Technische Fachhochschule Berlin, Berlin, Federal Republic of Germany (Chaps. 1, 4.2 – 4.6, 7, and 9)

JAN BALEJ, Ingenieurbüro für chemische Technik, Jülich, Federal Republic of Germany (Chap. 1)

JOHN E. BENNETT, Eltech Systems Corp., Fairport Harbor, Ohio 44077, United States (Chaps. 4.2 – 4.6)

PETER WINTZER, Cellulosefabrik Attisholz AG, Luterbach, Switzerland (Chaps. 7 and 9)

SAEED AKBAR SHEIKH, Davy McKee AG, Frankfurt, Federal Republic of Germany (Chaps. 2, 3, 4.1, 5, 6, and 9)

PATRIZIO GALLONE, Politecnico di Milano, Milano, Italy (Chaps. 8 and 9)

1. Introduction . . . 2
2. Hypochlorous Acid . . . 5
3. Solid Hypochlorites . . . 5
3.1. Properties . . . 5
3.2. Production . . . 6
3.3. Quality Specifications . . . 7
3.4. Uses . . . 7
4. Hypochlorite Solutions . . . 7
4.1. Chemical Production . . . 7
4.1.1. From Chlorine . . . 8
4.1.2. From Bleaching Powder . . . 8
4.2. Electrosynthesis . . . 9
4.2.1. Reaction Fundamentals . . . 9
4.2.2. Industrial Cells . . . 10
4.3. Storage . . . 14
4.4. Uses . . . 14
4.5. Economic Aspects . . . 15
4.6. Plant Safety . . . 15
5. Chlorine Dioxide . . . 16
5.1. Properties . . . 16
5.2. Production . . . 16
5.2.1. Day – Kesting Process . . . 17
5.2.2. R2 Process . . . 17
5.2.3. Mathieson Process . . . 19
5.2.4. Solvay Process . . . 19
5.2.5. Other Processes . . . 19
5.3. Economic Aspects . . . 19
5.4. Uses . . . 20
6. Sodium Chlorite . . . 20
6.1. Properties . . . 20
6.2. Production . . . 20
6.3. Uses . . . 21
7. Chloric Acid and Chlorates . . . 21
7.1. Properties . . . 22
7.2. Production Fundamentals . . . 24
7.2.1. Chlorate-Generating Reactions . . . 24
7.2.2. Loss Reactions . . . 26
7.3. Industrial Electrosynthesis Systems 27
7.3.1. Electrolysis Cell Types . . . 28
7.3.2. Electrodes . . . 28
7.3.3. Operational Parameters . . . 31
7.3.4. Brine Purification . . . 33
7.4. Crystallization . . . 33
7.5. Construction Materials . . . 33
7.6. Environmental Protection . . . 34
7.7. Quality Specifications . . . 34
7.8. Storage, Transportation, and Safety 34
7.9. Uses . . . 34
7.10. Economic Aspects . . . 35
8. Perchloric Acid and Perchlorates . 36
8.1. Physical and Chemical Properties . 36
8.1.1. Perchloric Acid . . . 36
8.1.2. Perchlorates . . . 36
8.2. Production . . . 37
8.2.1. Perchloric Acid . . . 37
8.2.2. Perchlorates . . . 38
8.3. Environmental Protection . . . 40
8.4. Chemical Analysis . . . 40
8.5. Storage, Transportation, and Safety 41
8.6. Uses . . . 41
8.7. Economic Aspects . . . 42
9. Toxicology and Occupational Health 42
10. References . . . 43

1. Introduction

Numerous chlorine oxides are known. However, only two anhydrides of chlorine oxygen acids, *dichlorine oxide,* Cl_2O, and *chlorine dioxide,* ClO_2, and two mixed anhydrides, *dichlorine hexoxide,* Cl_2O_6, and *dichlorine heptoxide,* Cl_2O_7, are fairly stable under certain condi-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_483

Table 1. Properties of chlorine oxides

	Dichlorine oxide [*7791-21-1*], Cl_2O	Chlorine dioxide [*10049-04-4*], ClO_2	Dichlorine hexoxide [*12442-63-6*], Cl_2O_6	Dichlorine heptoxide [*12015-53-1*], Cl_2O_7
Oxidation state	+1	+4	+6	+7
M_r	86.91	67.45	166.91	182.90
Melting point, °C	−116	−59	+3.5	−90
Boiling point, °C	+2	+11	–	82
Appearance	yellow-brown gas, red-brown liquid	orange-yellow gas, red liquid	red liquid	oily, colorless liquid
Stability	decomposes at 100 °C, explodes on heating or shock	unstable at ambient temperature	decomposes at melting point	decomposes slowly at ambient temperature

tions. Table 1 shows some important properties of these chlorine oxides. Other chlorine oxides, such as *dichlorine dioxide* [*12292-23-8*], Cl_2O_2, *dichlorine trioxide* [*17496-59-2*], Cl_2O_3, or *dichlorine tetroxide* [*27218-16-2*], Cl_2O_4, are unstable.

The chlorine oxygen acids are formed by reaction of the corresponding chlorine oxides with water.

Hypochlorous acid [*7790-92-3*], HClO:

$$Cl_2O + H_2O \longrightarrow 2\,HClO$$

Chlorous acid [*13898-47-0*], $HClO_2$:

$$2\,ClO_2 + H_2O \longrightarrow HClO_2 + HClO_3$$

Chloric acid [*7790-93-4*], $HClO_3$:

$$Cl_2O_6 + H_2O \longrightarrow HClO_3 + HClO_4$$

Perchloric acid [*7601-90-3*], $HClO_4$:

$$Cl_2O_7 + H_2O \longrightarrow 2\,HClO_4$$

Table 2 lists the thermodynamic properties of chlorine oxides, chlorine oxygen acids, and their sodium salts [1].

The *oxidation power* of individual chlorine oxygen compounds is characterized by the changes in the standard enthalpy ΔH° and Gibbs free energy ΔG° of the decomposition reactions forming molecular oxygen (Table 3). As shown in Table 3, the thermodynamic stability of chlorine oxygen acids and chlorine oxides increases with increasing oxidation state of the chlorine atom. Therefore, concentrated perchloric acid can be isolated, whereas all other oxygen acids are stable only in diluted form. The instability of the chlorine oxides and their acids determines their industrial significance. All chlorine oxygen compounds are strong oxidants; the strongest are those with the lowest oxidation state of the chlorine atom.

Dichlorine oxide, Cl_2O, chlorine dioxide, ClO_2, all oxygen acids, and their salts, particularly those of sodium and potassium, are used industrially.

History. Soon after the discovery of chlorine in 1774, scientific and commercial interest was directed to the chlorine oxides, the chlorine oxygen acids, and their salts.

Hypochlorite was first prepared in 1787 by C. L. Berthollet by feeding chlorine into potash lye. This bleach liquor (eau de Javel) was soon applied in bleaching textiles and in papermaking. Labarraque replaced potash lye by the cheaper soda lye (eau de Labarraque).

Hypochlorite was prepared by electrolysis of sodium chloride as early as 1801. However, commercial electrochemical production did not start for a long time.

At the beginning of the 20th century, the traditional routes to form bleach liquors [2] fell out of use by the rapidly expanding chlorine – caustic industry that made large quantities of cheap waste chlorine available. Within the last 20 years, however, electrosynthesis has found a widespread revival and is today an alternative to chemical hypochlorite production wherever safety risks of small plants are decisive.

In 1799 C. Tennant and C. McIntosh developed a process for the production of *bleaching powder* by absorbing chlorine onto dry calcium hydroxide. This bleaching powder was much more stable than previously obtained bleaching products. In 1906, G. Pistor succeeded in producing highly concentrated bleach-

Table 2. Thermodynamic properties of chlorine oxides, chlorine oxygen acids, and their salts at 25°C

Compound	Formula	CAS registry number	State∗	ΔH_f°, kJ/mol	ΔG_f°, kJ/mol
Dichlorine oxide	Cl_2O	[*7791-21-1*]	g	80.3	97.9
Chlorine dioxide	ClO_2	[*10049-04-4*]	g	102.5	120.5
Chlorine trioxide	ClO_3	[*13932-10-0*]	g	155	–
Dichlorine heptoxide	Cl_2O_7	[*12015-53-1*]	l	238.1	–
			g	272.0	–
Hydrochloric acid	HCl	[*7647-01-0*]	ao	– 167.2	– 131.2
Hypochlorous acid	HClO	[*7790-92-3*]	ao	– 120.9	– 79.9
Chlorous acid	$HClO_2$	[*13898-47-0*]	ao	– 51.9	+ 5.9
Chloric acid	$HClO_3$	[*7790-93-4*]	ai	– 104.0	– 7.95
Perchloric acid	$HClO_4$	[*7601-90-3*]	ai	– 129.3	– 8.52
Sodium chloride	NaCl	[*7647-14-5*]	ai	– 407.3	– 393.1
Sodium hypochlorite	NaClO	[*7681-52-9*]	ai	– 347.3	– 298.7
Sodium chlorite	$NaClO_2$	[*7758-19-2*]	ai	– 306.7	– 244.7
Sodium chlorate	$NaClO_3$	[*7775-09-9*]	ai	– 344.1	– 269.8
Sodium perchlorate	$NaClO_4$	[*7601-89-0*]	ai	– 369.5	– 270.4

∗ ao: undissociated solute in aqueous ideal solution at unit molality; ai: electrolyte in the hypothetical ideal solution at unit activity, dissociated into ions.

Table 3. Thermodynamic data of oxygen-forming decomposition reactions of chlorine oxygen compounds at 25°C

Reaction	ΔH°, kJ/mol	ΔG°, kJ/mol
$Cl_2O\,(g) + H_2O \longrightarrow 2\,HCl\,(aq) + O_2\,(g)$	– 128.8	– 123.2
$4/5\,ClO_2\,(g) + 2/5\,H_2O \longrightarrow 4/5\,HCl\,(aq) + O_2\,(g)$	– 101.4	– 106.5
$4/7\,ClO_3\,(g) + 2/7\,H_2O \longrightarrow 4/7\,HCl\,(aq) + O_2\,(g)$	– 102.4	
$1/4\,Cl_2O_7\,(l) + 1/4\,H_2O \longrightarrow 1/4\,HCl\,(aq) + O_2\,(g)$	– 71.6	
$1/4\,Cl_2O_7\,(g) + 1/4\,H_2O \longrightarrow 1/4\,HCl\,(aq) + O_2\,(g)$	– 80.1	
$2\,HClO\,(aq) \longrightarrow 2\,HCl\,(aq) + O_2\,(g)$	– 92.5	– 102.7
$HClO_2\,(aq) \longrightarrow HCl\,(aq) + O_2\,(g)$	– 115.3	– 137.1
$2/3\,HClO_3\,(aq) \longrightarrow 2/3\,HCl\,(aq) + O_2\,(g)$	– 42.1	– 82.2
$1/2\,HClO_4\,(aq) \longrightarrow 1/2\,HCl\,(aq) + O_2\,(g)$	– 18.9	– 61.4

ing powder with more than 70 % available chlorine.

R. Chevenix first made *chlorine dioxide* in 1802 by the reaction of concentrated sulfuric acid and potassium chlorate, but the product was first identified by H. Davy and F. von Stadion in 1815/1816. Calvert and Davies used oxalic acid instead of sulfuric acid in 1859 and obtained chlorine dioxide together with carbon dioxide, thus eliminating the explosion risk. This and the investigations of E. Schmidt in 1921 – 1923 [3] laid the ground to today's extensive use of chlorine dioxide as a bleaching agent in the textile, pulp, and paper industry.

J. R. Glauber probably prepared *chlorate* for the first time, but C. L. Berthollet first made chlorate in 1787 by the reaction of chlorine with potassium hydroxide and identified it as the salt of chloric acid. The first electrochemical chlorate preparation was performed by W. von Hisinger and J. J. Berzelius in 1802. In 1851 a cell patent was granted to C. Watt that had the essential features of later cells, but industrial production did not start before 1886. Originally, all cells were operated with an alkaline electrolyte until J. Landin added chromic acid to the electrolyte and solved the problem of cathodic hypochlorite reduction [4], [5]. Within the few years before and after 1900, scientific research focused on the mechanism of chlorate formation [6–12]; this was supplemented by numerous investigations within the last 30 years. However, the detailed reaction mechanism is still a subject of scientific debate.

The most important technological improvements were (1) the separation of the electrochemical reactor from the chemical reactor in 1933 (A. Schumann-Leclercq) [13], (2) the utilization of the hydrogen formed at the cathode for "stirring" the electrolyte [14], and (3) the introduction of dimensionally stable titanium anodes 20 years ago [15].

F. von Stadion first made *perchlorates* by oxidation of chlorate at platinum anodes in 1816; he also obtained perchloric acid by the reaction of concentrated sulfuric acid with potassium perchlorate and by anodic oxidation of hydrochloric acid. Preparation of perchloric acid by electrolysis of dilute chloric acid was first carried out by J. J. Berzelius in 1835. Like all other electrochemical production methods, electrosynthesis of perchlorate and perchloric acid did not gain industrial significance before the end of the 19th century. In 1890, O. Carlson was granted a patent for electrochemical perchlorate production, and in 1895 he operated the first commercial plant in Sweden. The first methodic studies of the reaction fundamentals were initiated by German researchers in 1898 [16–18].

General Significance. The importance of chlorine oxygen compounds is based predominantly on their oxidizing power. For 200 years, after the lawn bleaching of textiles became obsolete, the textile and paper industry has invariably been a main consumer of chlorine dioxide and hypochlorite. The importance of these products has steadily grown in accordance with the expansion of the pulp and paper industry. Traditional uses of chlorate as a herbicide or explosive became less important. On the other hand, the increasing demand for disinfection of process and drinking water and for sanitation in general has strongly favored the use of chlorine oxygen compounds. The importance of chlorinated lime as a disinfectant has declined. However, other chlorine oxygen compounds, above all hypochlorite, have grown. This trend was particularly favored by the danger of handling chlorine. Moreover, some of the chlorine oxygen compounds have found numerous novel uses that are reported in detail in the following chapters.

Today, chlorate production is one of the most important inorganic electrosyntheses. In addition, perchlorates are preferably made by electrochemical processes. On the basis of a deeper understanding of the chemical and electrochemical fundamentals and by introducing dimensionally stable anodes, design and operation of electrochemical plants have been revolutionized to such an extent that a modern electrochemical reactor has very little resemblance to a cell 20 years ago.

2. Hypochlorous Acid

Hypochlorous acid [*7790-92-3*], HOCl, M_r 52.5, is only moderately stable in aqueous solution. It is colorless when dilute and yellowish at higher concentrations. Hypochlorous acid is one of the most powerful oxidizing agents known.

Hypochlorous acid solution decomposes exothermically. The main decomposition products are hydrochloric acid and oxygen:

$$2\,HOCl \longrightarrow 2\,HCl + O_2$$

Minor amounts of chlorine and chloric acid are also formed.

Production. Hypochlorous acid is produced by the reversible reaction of chlorine and water:

$$Cl_2 + H_2O \rightleftharpoons HOCl + HCl$$

For efficient conversion, hydrochloric acid must be removed from the equilibrium mixture. This is achieved by limestone, $CaCO_3$, soda ash, Na_2CO_3, or calcium hypochlorite, $Ca(OCl)_2$.

One of the common ways to produce hypochlorous acid is to pass chlorinated water through towers packed with powdered limestone [19]. The overall reaction is as follows:

$$2\,Cl_2 + 2\,H_2O + 2\,CaCO_3 \longrightarrow 2\,HOCl + CaCl_2 + Ca(HCO_3)_2$$

Calcium hydrogen carbonate present in the resulting solution provides the buffer needed to stabilize the product. Any hydrochloric acid still produced is removed by the reaction

$$Ca(HCO_3)_2 + 2\,HCl \longrightarrow CaCl_2 + 2\,H_2O + 2\,CO_2$$

Hypochlorous acid can also be prepared by passing a mixture of dry chlorine and air through a column packed with yellow mercuric oxide, HgO. The resulting dichlorine oxide is dissolved in water to produce hypochlorous acid [20], [21].

Storage. Because of its limited stability, hypochlorous acid is best used soon after production. Its long-distance transport or intermediate storage should be avoided.

A solution containing less than 1 % hypochlorous acid can be stored in the dark over fairly long periods, provided that such metals as copper, nickel, or cobalt are absent. Solutions containing 30 % hypochlorous acid can also be stored, but at temperatures of −20 °C or lower.

Uses. Because of its instability, hypochlorous acid is not used extensively as an oxidizing or bleaching agent. It was used mainly in the water treatment industry for slime control, for treatment of drinking water, and for sterilization of swimming pools. These applications have now been taken over by the more stable hypochlorites.

3. Solid Hypochlorites

3.1. Properties

All solid hypochlorites are soft, white, dry powders. Some are almost odorless; others smell more or less strongly of chlorine or hydrochloric acid because of decomposition during storage (Eqs. 1 – 3). The stability of hypochlorites depends primarily on their water content, which is usually less than 1 %; tropical bleach contains even less than 0.3 %. They are stable up to 80 °C, tropical bleach even up to 100 °C. When heated to 180 °C, they decompose into chloride and oxygen. Such metals as iron, nickel, or cobalt decrease the stability of hypochlorites. Therefore, the raw materials used for the production of hypochlorites must be free of such metals. *Tropical bleach* (< 0.3 % water) free of heavy metals has a shelf life of more than 2 years, if properly stored.

If not properly stored in air-tight containers, hypochlorites suffer loss of available chlorine because of reaction with water:

$$Ca(OCl)_2 + CaCl_2 + 2\,H_2O \longrightarrow 2\,Ca(OH)_2 + 2\,Cl_2 \qquad (1)$$

or reaction with carbon dioxide:

$$Ca(OCl)_2 + CaCl_2 + 2\,CO_2 \longrightarrow 2\,CaCO_3 + 2\,Cl_2 \qquad (2)$$

or reaction with both:

$$Ca(OCl)_2 + CO_2 + H_2O \longrightarrow CaCO_3 + 2\,HOCl \qquad (3)$$

3.2. Production

Bleaching Powder. Standard bleaching powder is a mixture of calcium hypochlorite [*7778-54-3*], $Ca(OCl)_2$, calcium chloride, and calcium hydroxide containing varying amounts of water. It is made by passing chlorine over hydrated lime. The Rheinfelden bleaching powder process of Dynamit Nobel is a batch operation [22]. Dry, powdered lime hydrate is chlorinated at 45 °C and low pressure (5.3 kPa) in a horizontal reaction drum. Chlorine is injected as a liquid. The reaction mass is permanently mixed by a slowly rotating rake. The reaction of solid lime hydrate and chlorine leads to the formation of a mixture of dibasic calcium hypochlorite [*12394-14-8*], $Ca(OCl)_2 \cdot 2\,Ca(OH)_2$, and basic calcium chloride, corresponding to 40 % conversion of the available calcium hydroxide. The following equation characterizes the reaction [23], [24]:

$$5\,Ca(OH)_2 + 2\,Cl_2 \longrightarrow Ca(OCl)_2 \cdot 2\,Ca(OH)_2 + CaCl_2 \cdot Ca(OH)_2 \cdot H_2O + H_2O$$

On further chlorination, hemibasic calcium hypochlorite [*62974-42-9*], $Ca(OCl)_2 \cdot 1/2\,Ca(OH)_2$, and neutral calcium chloride hydrate are formed. After ca. 60 % of the available calcium hydroxide has been converted, the bleaching powder reaction stops; this can be represented by the following equation:

$$10\,Ca(OH)_2 + 6\,Cl_2 \longrightarrow Ca(OCl)_2 \cdot 2\,Ca(OH)_2 + 2\,Ca(OCl)_2 \cdot 1/2\,Ca(OH)_2 + CaCl_2 \cdot Ca(OH)_2 \cdot H_2O + 2\,CaCl_2 \cdot H_2O + 3\,H_2O$$

The reaction is strongly exothermic, generating 1100 kJ of heat per kg of chlorine converted. This heat and the low pressure cause the water formed during the reaction and the liquid chlorine to evaporate. Consequently, the reaction mass is dried completely under vacuum at a maximum temperature of 85 °C. The product is standard bleaching powder of 35 – 37 % available chlorine content (for definition, see Section 3.3).

Gaseous chlorine can also be used for this reaction, but then the reaction takes 2 – 3 times longer, chlorine losses are higher, and the available chlorine content of the product is smaller.

Tropical Bleach. To reduce the water content further, finest ground quicklime, CaO, is added to the bleaching powder. It absorbs any water still present and is converted into calcium hydroxide. Although this operation decreases the

available chlorine content by 1 – 2 %, the extra drying makes the resulting bleaching powder, known as tropical bleach, stable up to temperatures of 100 °C.

ICI has developed a continuous process for the production of bleaching powder, in which countercurrents of calcium hydroxide and chlorine react in a rotating drum [25]. The heat of reaction is removed by spraying the drum externally with water and by diluting the chlorine with cooled air; this gas stream also removes the water formed during the chemical reaction.

High-Percentage Hypochlorite. Solid hypochlorites with 70 % and higher available chlorine contents can be prepared by chlorinating slurries of such calcium compounds as calcium hydroxide, or bleaching powder. Initially, hemibasic calcium hypochlorite, $Ca(OCl)_2 \cdot 1/2\,Ca(OH)_2$, is formed. When further chlorinated, this gives neutral calcium hypochlorite dihydrate, $Ca(OCl)_2 \cdot 2\,H_2O$, which is then dried to the desired high-percentage hypochlorite. In all of these reactions, calcium chloride is formed as a byproduct [26–28]. Some processes recover the calcium values by adding sodium hypochlorite to the slurries:

$$2\,NaOCl + CaCl_2 \longrightarrow Ca(OCl)_2 + 2\,NaCl$$

In such cases, the product primarily consists of calcium hypochlorite, sodium chloride, and water, which is then removed [29].

Barium and Magnesium Hypochlorite. Barium hypochlorite [*13477-10-6*], $Ba(OCl)_2$, can be produced in large crystals with a maximum available chlorine content of 59 %. It is more stable than calcium hypochlorite but also more expensive because of the high cost of raw materials.

Magnesium hypochlorite [*10233-03-1*], $Mg(OCl)_2$, is extremely unstable and decomposes when dried.

3.3. Quality Specifications

The term *available chlorine content*, also called *active chlorine*, represents the mass fraction of liberated chlorine in bleaching powder when bleaching powder reacts with hydrochloric acid.

The following qualities of solid hypochlorites are available (content of available chlorine, wt %, in parentheses): tropical bleach (34 – 35), bleaching powder (35 – 37), and high-percentage hypochlorite (70).

3.4. Uses (→ Bleaching)

In the paper industry, calcium hypochlorite is used in single-stage bleaching. The more expensive sodium hypochlorite is used in the multistage process, which involves chlorination, caustic extraction, and hypochlorite oxidation. Kraft pulp is processed to higher brightness and greater strength when sodium hypochlorite is used instead of calcium hypochlorite.

For bleaching in laundry operations, bleaching powder is first suspended in water and then decanted. Only the solution is used because the insolubles could damage the fibers.

4. Hypochlorite Solutions

The reaction of gaseous chlorine with a slight excess of alkali produces highly concentrated hypochlorite solutions. Available chlorine concentrations of 170 – 220 g/L can be obtained, or even more if the residual chloride concentration is lowered extremely [30]. Hypochlorite solutions are relatively safe and are often chosen instead of chlorine for bleaching, disinfection, biofouling control, and odor control. Concern regarding the safety hazards associated with liquid chlorine has grown. Several major cities now restrict transportation of chlorine within their boundaries, and a great deal of attention has focused on accidents caused by the handling of liquid chlorine by unskilled labor [31]. This has increased the popularity of hypochlorite solutions in spite of their relatively high cost.

An attractive alternative to the chemical production of hypochlorite solutions described above is the on-site electrolysis of brine or seawater. Such processes are described in Section 4.2.

Stability. Hypochlorite solutions are more stable than solutions of hypochlorous acid, but they are active enough to be used as disinfectants or bleaching agents. The factors that affect the

stability of the parent acid also affect the stability of hypochlorite solutions: concentration, presence of such metals as copper, nickel, or cobalt, pH, temperature, and exposure to light.

Available Chlorine. When chlorine reacts with caustic soda, half of the chlorine is lost because inert sodium chloride is formed:

$$Cl_2 + 2\,NaOH \longrightarrow NaOCl + NaCl + H_2O$$

However, as an oxidant, sodium hypochlorite decomposes to sodium chloride and oxygen:

$$NaOCl \longrightarrow NaCl + [O]$$

The oxidizing power of one oxygen atom is equivalent to that of two chlorine atoms. Therefore, the complete oxidizing power of the original chlorine is available in the hypochlorite solution; it is expressed in grams of available chlorine per liter of finished solution. Thus, the "available chlorine" of hypochlorite solutions compares the oxidizing power of the agent to that of the equivalent amount of elemental chlorine used to make the solution.

4.1. Chemical Production

A hypochlorite unit is attached to each chlor-alkali plant to render harmless the dilute chlorine that cannot be recovered economically. These units make most of the industrially produced hypochlorite solutions [32–34].

Beyond that, hypochlorite solutions with available chlorine contents higher than 5 g/L are made by passing chlorine gas through dilute solutions of sodium hydroxide or potassium hydroxide. Calcium hydroxide suspensions may also be used, from which, after filtration or decantation, clear calcium hypochlorite solutions are obtained [35].

4.1.1. From Chlorine

In the commercial production of hypochlorite solutions from chlorine gas and alkali solutions of various concentrations, the following conditions must be maintained.

1) The temperature must be controlled at 30 – 35 °C.
2) The solution must be alkaline at any stage.
3) The equipment must provide for thorough mixing and escape of inert gas.
4) Such heavy metals as manganese, iron, cobalt, nickel, or copper must be avoided in the system.
5) Available chlorine contents of more than 150 g/L should be avoided. The high decomposition rate of such concentrated solutions more than offsets any savings in transportation [36].

Sodium or Potassium Hypochlorites. Figure 1 shows a schematic diagram of a typical process for the commercial production of hypochlorite solutions. The continuous process can produce hypochlorite solutions of any available chlorine content between 0 and 150 g/L and any amount between zero and the designated capacity. The only parameters to be adjusted are the redox potential of the analyzer (b) and the desired dilutionof the caustic soda solution. For a given amount of chlorine coming in, the system automatically adjusts the caustic soda, process water, and hypochlorite solution flows into and out of the system.

Chlorine gas, diluted with air, is introduced into the chlorination column (a), packed with Raschig rings. Caustic soda is diluted to the desired concentration with water. Tank (d) provides for buffer capacity and homogenization. The pumps (e) circulate the mixture of caustic soda and sodium hypochlorite through the titanium heat exchanger (f), the chlorination column (a), and the tank (d).

Any amount of chlorine between zero and the designated capacity is absorbed in the circulating caustic, producing sodium hypochlorite. When the available chlorine concentration reaches the desired value, the analyzer (b) signals to open control valve (c) for the withdrawal of sodium hypochlorite to the storage tank (g). In this case, the liquid in tank (d) must be replenished by fresh caustic.

The storage tank (g) is equipped with a circulation pump and a heat exchanger to keep the temperature of the stored hypochlorite solution below 35 °C. If necessary, the inert gas leaving column (a) can be scrubbed before entering the atmosphere.

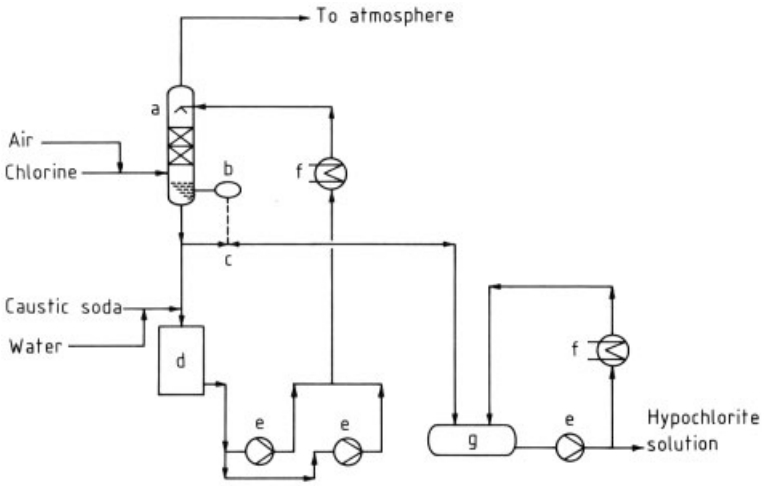

Figure 1. Production of hypochlorite solution from chlorine and caustic soda
a) Chlorination column; b) Analyzer; c) Control valve; d) Buffer tank; e) Pumps; f) Heat exchangers; g) Storage tank

Materials of Construction. All pumps have rubber-lined steel casings and titanium propellers. The heat exchangers are plated with titanium. The tanks (d) and (g) are made from fiberglass-reinforced plastic [37]. The same material is used for the chlorination column and the pipes; chlorinated poly(vinyl chloride) is an alternative material for the pipes.

Calcium Hypochlorite. If a solution of calcium hypochlorite is desired, milk of lime – a suspension of calcium hydroxide in water – is chlorinated. (The solubility of calcium hydroxide in water is 1.3 g/L at 20 °C.) In that case, the plant described in Figure 1 must be modified to handle the solid phase.

In addition to design changes in the equipment, no packing is needed in the chlorination column and the piping must be designed to ensure that pockets are avoided where solid particles can settle and block the pipes; a settler or filter is needed to remove all insoluble or suspended particles before a clear solution of calcium hypochlorite flows to the storage tank.

4.1.2. From Bleaching Powder

To avoid long-term storage of hypochlorite solutions, laundries prefer to store solid bleaching powder and then prepare the sodium hypochlorite solutions by using sodium carbonate, sodium sulfate, or caustic soda:

$$CaCl(OCl) + Na_2CO_3 \longrightarrow NaOCl + NaCl + CaCO_3$$

$$CaCl(OCl) + Na_2SO_4 \longrightarrow NaOCl + NaCl + CaSO_4$$

$$CaCl(OCl) + 2\,NaOH \longrightarrow NaOCl + NaCl + Ca(OH)_2$$

A disadvantage of this method is the precipitation of $CaCO_3$, $CaSO_4$, or $Ca(OH)_2$, which requires filtering or settling before the hypochlorite solution can be used.

4.2. Electrosynthesis

On-site, electrochemical production of dilute hypochlorite solution has long been recognized as an option wherever long-term storage of hypochlorite is unnecessary; it is now rapidly gaining popularity [38]. Hypochlorite solutions with an available chlorine content of up to 10 g/L are commonly produced on the site in electrochemical cells by using either prepared brine or natural seawater as feed.

4.2.1. Reaction Fundamentals

Electrolysis of sodium chloride yields chlorine at the anode:

$$2\,Cl^- \longrightarrow Cl_2 + 2\,e$$

The final product depends on the operational conditions of the cell. In the production of chlorine gas, special care is taken to prevent mixing of anode and cathode products (→ Chlorine). Chlorine hydrolyzes and hypochlorous acid dissociates, forming hypochlorite and chloride in solution [39], [40]:

$$Cl_2 + H_2O \rightleftharpoons HClO + Cl^- + H^+$$

$$HClO \rightleftharpoons ClO^- + H^+$$

The formation of hypochlorous acid and hypochlorite ceases when the electrolyte is saturated with chlorine and chlorine gas evolves at pH 2–3.

However, in the electrosynthesis of hypochlorite, as well as of chlorate, anolyte and catholyte are vigorously mixed. The hydroxyl ions formed at the cathode

$$2\,H_2O + 2\,e \longrightarrow 2\,OH^- + H_2$$

maintain the electrolyte near neutrality (pH 7–9). Under this condition, the concentration of dissolved chlorine near the anode surface remains too low to permit evolution of gaseous chlorine, and hypochlorite is the main product.

There are four main loss reactions that complicate industrial operation; they have been studied to optimize the operating conditions [41], [42] and are described in the following paragraphs in more detail.

Cathodic Reduction. Hypochlorite is reduced at the cathode to form chloride:

$$ClO^- + H_2O + 2\,e \longrightarrow Cl^- + 2\,OH^-$$

The rate of this reaction is controlled by mass transfer and is, thus, linearly proportional to the overall hypochlorite concentration; it increases with flow rate and temperature [42–45].

Loss by cathodic reduction also occurs in chlorate production; in that case it is minimized by adding dichromate to the electrolyte (see Section 7.2.2). This remedy is not possible in the simple flow-through hypochlorite cells, and loss in commercial cells is considerable, as Figure 2 shows.

Several means are used to lower reduction loss. Smooth cathode surfaces are superior to rough surfaces [45]. Loss is also lowered by decreasing the active area of the cathodes, thereby increasing the cathode current density. Cathodic reduction was greatly suppressed by decreasing the active area to ca. 1 % with a synthetic resin coating [42]; however, the attendant voltage increase makes this approach industrially unattractive. Bubbles always cover a portion of the cathode surface and contribute to lower cathodic reduction. Hydrogen evolution increases cathodic reduction because its stirring action enhances the mass transfer coefficient. Efforts to lower the mass transfer coefficient by covering the cathode with a porous plastic have also been successful at suppressing cathodic reduction [47]. Similarly, the inhibitory action of calcium chloride was related to the formation of insoluble compounds on the cathode surface [48].

Anodic Oxidation. Anodic oxidation of hypochlorite to chlorate is used in industry and is described in detail in Section 7.3:

$$3\,ClO^- + 1.5\,H_2O \longrightarrow ClO_3^- + 3\,H^+ + 2\,Cl^- + 0.75\,O_2 + 3\,e$$

This reaction is also controlled by mass transfer, and its rate increases with the hypochlorite concentration. Because hypochlorite is decomposed at the anode and cathode, commercial on-site production is restricted to a maximum concentration of available chlorine of ca. 10 g/L; this is usually sufficient for disinfecting and deodorizing water. The contribution of anodic oxidation to the overall loss can further be lowered by increasing the anodic current density; however, the anode lifetime may then decrease. The loss reaction may also depend on the anode material: dimensionally stable anodes based on RuO_2 were found to be more efficient than platinized titanium or graphite [42].

Anodic Water Electrolysis. This loss reaction competes with chlorine discharge, and its rate depends on the chloride concentration, various mass transfer coefficients, and the nature of the anode material.

$$H_2O \longrightarrow 1/2\,O_2 + 2\,H^+ + 2\,e$$

When NaCl concentrations are greater than 100 g/L, the loss is small. However, a typical NaCl concentration in a modern brine cell is ca.

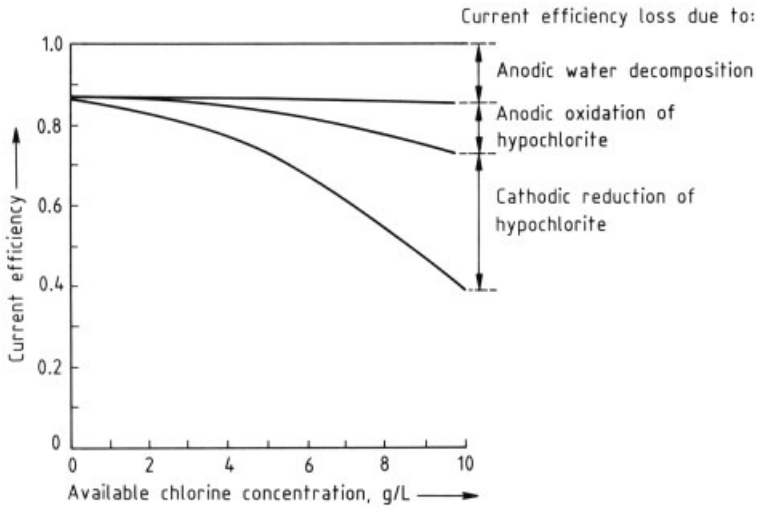

Figure 2. Current efficiency loss vs. available chlorine concentration (NaCl: 28 g/L, 25 °C, 1550 A/m^2) [46]

30 g/L, at which the rate of water decomposition is significant, as shown in Figure 2.

When salt must be purchased, the conversion of chloride into hypochlorite must be maximized. A typical on-site electrosynthesis from prepared brine consumes 3 – 5 kg of salt per kg of available chlorine produced. The question of chloride consumption is meaningless where natural seawater is used as a feed. However, in cases of low salinity or very cold seawater, oxygen evolution caused by anodic water electrolysis may decrease current efficiency by as much as 40 % [49].

As in the case of chlorate synthesis, the nature of the anode material strongly influences the amount of water decomposition. Platinized titanium and dimensionally stable anodes based on RuO_2 are more selective for chlorine evolution, whereas PbO_2 and graphite anodes have a greater tendency to evolve oxygen.

Chemical Chlorate Formation. Autoxidation of hypochlorite to chlorate is the preferred route of commercial chlorate production (Section 7.3). The rate of this reaction depends on pH and temperature; in industrial hypochlorite production, it is minimized by keeping the temperature below 40 °C and the pH above 7 [41].

4.2.2. Industrial Cells

The number of companies offering hypochlorite cells has grown quickly in the past few years. Over 20 suppliers offer a wide variety of cell designs. They can all be classified into three basic types: (1) tube cells, (2) parallel-plate cells, and (3) rotating or mechanical scraper-type cells. Of the modern hypochlorite cells, the tube cells were developed first. They generally consist of two concentric pipes, one being the anode and the other the cathode, with the annular space serving as the electrode gap. This type of cell may be operated under pressure and is well-suited for small applications. Parallel-plate cells achieve a much better packing of electrode area, and most large industrial installations use cells of this type. Only two of the manufacturers offer rotating or mechanical scraper-type cells. In theory, such cells should be able to operate indefinitely without deposit buildup, and would therefore result in low maintenance cost. In practice, however, their extra mechanical action is difficult to maintain, and such cells have not yet captured a significant market share.

Seawater Cells. Cells designed to operate by using natural seawater are different from those using prepared brince. Cells using seawater tend to operate at a higher electrolyte flow rate and a wider electrode gap than brine cells to minimize problems arising from cathode deposits. Seawater cells also produce a lower product con-

centration, typically solutions with an available chlorine content of 0.5 – 4.0 g/L. Hypochlorite concentration is kept low to maximize current efficiency whenever high salt usage is not a concern (see also Table 4).

Table 4. Typical operational data of electrolytic cells for hypochlorite generation

Parameter	Brine feed	Seawater feed
Current density, A/m^2	1500	1500
Current efficiency, %	65	90
Temperature, °C	25	5 – 25
Concentration		
NaCl (cell entrance), g/L	30	15 – 30
Available chlorine (cell exit), g/L	8 – 10	1.0 – 3.0
Energy consumption, kW/h/kg of available chlorine	4.5 – 5.0	3.3 – 4.1
Sodium chloride consumption, kg of NaCl per kg of available chlorine	3 – 3.5	

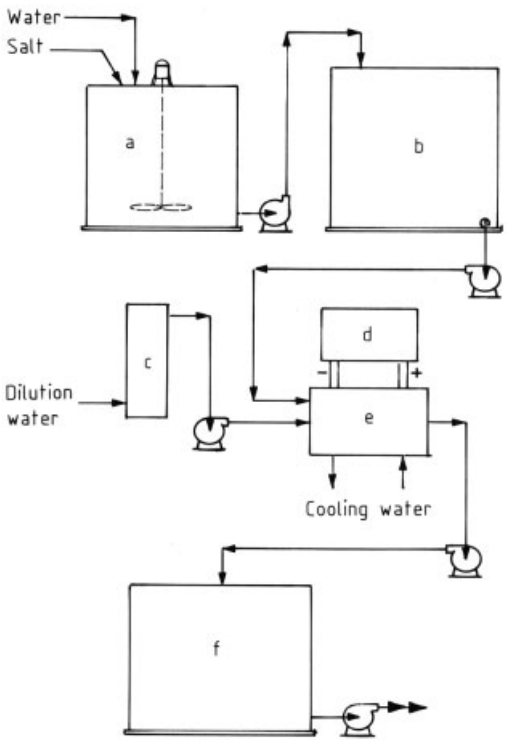

Figure 3. Typical layout for a brine hypochlorite cell system
a) Automatic brine makeup; b) Brine storage tank; c) Water softener; d) Rectifier; e) Hypochlorite cell; f) Hypochlorite storage tank

Brine Cells. Brine cells usually produce solutions with an available chlorine content of 7 – 10 g/L to keep salt cost low. Most manufacturers offer brine cells of the same design as their seawater cells, but electrolyte flow rate is lowered to maximize current efficiency at the higher concentration of available chlorine. Figure 3 shows a flow diagram for a typical system designed for electrolysis using prepared brine. Although brine cells were developed first, they now have only a small share of the total on-site hypochlorite cell market.

Table 4 shows typical operational data of seawater and brine feed cells.

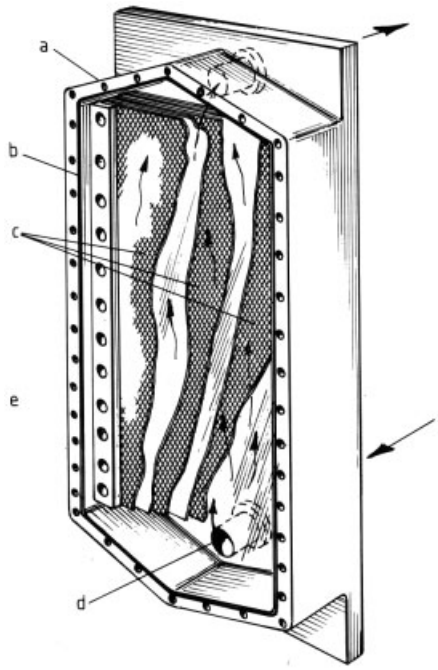

Figure 4. Sanilec seawater electrolysis cell
a) Molded polypropylene cell body; b) O-ring seal; c) Dimensionally stable anodes; d) Seawater inlet; e) Clear acrylic cover (not shown)

Sanilec System. Eltech System Corp. first offered the Sanilec system for seawater electrolysis in 1973. The cells are of a parallel-plate design (Fig. 4) and feature once-through operation without recycle [31]. Cells producing 30, 70, 140, or 155 kg/d can be banked in series to produce hypochlorite solutions of available chlorine concentrations up to 3.0 g/L. Full-load a.c. power consumption is as low as 4.1 kWh per kg of chlorine, partly because dimensionally stable anodes are used. These cells use cathodes made of nickel alloys and remove hydrogen pe-

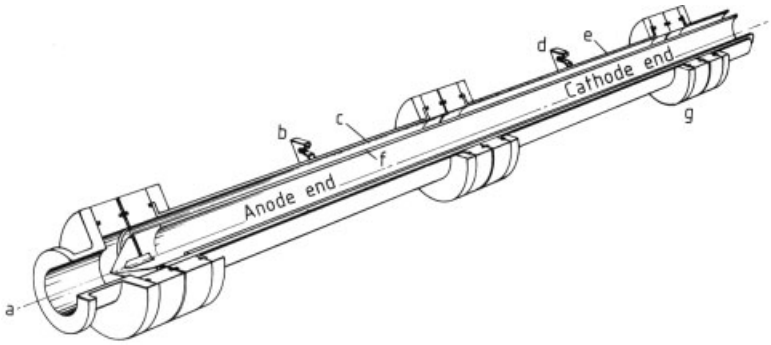

Figure 5. Chloropac seawater cell
a) Inlet; b) Cathode connector; c) Cathode; d) Anode connector; e) Anode; f) Bipolar electrode; g) Insulating flange

riodically to lower power consumption. Typical operational data are shown in Table 4.

Chloropac System. Englehard Minerals & Chemical Corp. produces tube cells with concentric titanium pipes as anode and cathode under the trade name of Chloropac System [50]. These cells are designed for an operating pressure of 1000 kPa. They utilize a high seawater flow rate to minimize deposit formation.

As shown in Figure 5, the electrodes are assembled in a bipolar arrangement, which is particularly suited for smaller capacities. The platinized titanium anodes are coated with 5 µm of platinum and have good current efficiency for chlorine evolution. The power consumption is claimed to be 3.5 – 5.0 kW h per kg of chlorine.

M. G. P. S. System. Mitsubishi Heavy Industries offers monopolar plate-type cells arranged in series under the name of the M. G. P. S. system. The cell body is made from mild steel lined with rubber, and the system is well-suited for large industrial applications. Cathodes are made from titanium and anodes from platinized titanium; precious metal oxide anodes are also used. The available chlorine concentration ranges from 0.2 to 1.0 g/L, and the power consumption is reported to be 5.8 kW h per kg of chlorine.

Seaclor System. The Oronzio de Nora Seaclor system features a bipolar parallel plate-type seawater cell [51] shown in Figure 6. Cathodes are made from titanium and anodes are based on RuO_2. Hypochlorite solutions with an available chlorine content of up to 2.5 g/L are produced; power consumption is cited to be 3.4 – 4.5 kW h per kg of chlorine. The most notable feature of the Seaclor system is the size of the individual cells, which may be large enough to produce nearly 1000 kg/d.

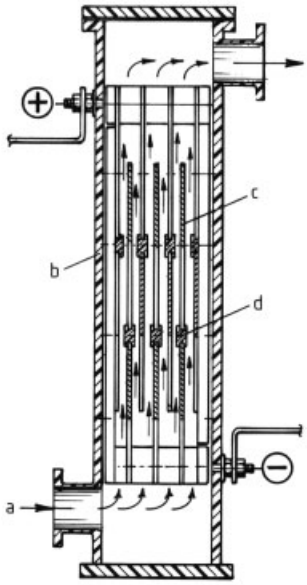

Figure 6. Seaclor hypochlorite cell
a) Inlet; b) PVC cell body with overlay; c) Bipolar electrode; d) Insulator

Pepcon System. The Pepcon system made by Pacific Engineering & Production Co. of Nevada is distinctly different from other commercial cells because it employs less expensive PbO_2 anodes. Pacific Engineering uses a tube-type design, with a steel or titanium pipe form-

ing the outside of the cell and a PbO_2-plated graphite rod at the center. Steel has a small overpotential for hydrogen evolution, but it requires careful protection to prevent corrosion during shutdown.

Maintenance. The major problem of on-site hypochlorite cells, especially those using natural seawater as feed, is that of deposit formation. Because the electrolyte adjacent to the cathode is strongly alkaline, magnesium hydroxide and calcium hydroxide deposit at the cathode surface. If allowed to build up, these deposits may bridge the electrode gap, reducing cell efficiency and ultimately causing anode failure [52].

Most manufacturers of industrial hypochlorite cells will tolerate deposit formation, but simultaneously attempt to minimize the problem by controlling the current density, turbulence, and cathode surface. In the case of seawater the deposit – chiefly $Mg(OH)_2$ – is soft and can generally be scoured from the cathode surface and flushed from the cell by using a high flow rate. Although at least two suppliers suggest that acid cleaning of deposits is never necessary, it seems likely that occasional removal of deposits is required on a cycle that varies from a few days to a few months. Usually, the deposit is removed by flushing the cell with hydrochloric acid. This dissolves the deposit quickly and consumes very little acid. Large industrial installations usually provide for convenient acid washing.

Brine cells may also form deposits as a result of hardness ions introduced from either feed water or from impurities in the salt. In this case, the deposit will usually consist of a hard calcium carbonate, which is also easily removed with acid. Unlike in seawater cells, however, it is often practical to soften the cell feed; brine cells can then be designed to operate with very little deposit formation.

A second major maintenance item is occasional replacement of anodes. The lifetime of both precious metal oxide and platinized titanium anodes is adversely affected by high current density, low salinity, low electrolyte temperature, and severe deposit formation. Manufacturers should be consulted on the expected anode lifetime and cost of replacement. Typically, anodes must be replaced or recoated after 2 – 5 years from startup.

4.3. Storage

Hypochlorite solutions slowly decompose if catalytic amounts of cobalt, nickel, or copper are present; iron and magnesium do not act catalytically [53]. Therefore, long-term storage and transport over long distances must be avoided. Hypochlorite solutions prepared from seawater are especially unstable and should be consumed directly after on-site preparation.

Hypochlorite solutions for household use have an available chlorine concentration of ca. 40 g/L. Special attention must be paid to minimize oxygen evolution. Improper storage of hypochlorite bottles may cause stoppers to be blown out and bottles to explode. Vented stoppers are used to avoid pressure buildup in the bottles.

4.4. Uses

Concentrated hypochlorite solutions are used primarily in the paper and textile industries for bleaching. Since chlorine dioxide (Chap. 5) produces a brighter product and is less harmful to fibers, the use of a hypochlorite solution as bleaching agent in these industries has declined.

On the other hand, the use of electrochemically produced dilute hypochlorite solutions has increased largely over the past several years. Electrolytic generators can be used whenever hypochlorite is needed for disinfection or for bleaching; their advantages are economy, safety, and convenience.

Currently, the broad area of biofouling control accounts for over half of the market for such equipment, especially at locations remote from chlor-alkali plants. The largest seawater electrolysis plants for the production of hypochlorite in the world, a 60 000-kg/d plant in Kuwait and a 48 000-kg/d plant in Saudi Arabia, were commissioned in 1980 as parts of large desalination projects. The hypochlorite is used to control slime and algae in piping and tubes and to eliminate odor in the desalted water.

Another important and quickly growing use for hypochlorite cells is disinfection of seawater for secondary oil recovery; otherwise, slime growth would clog the oil-bearing strata.

The largest plant for this purpose is a 5400-kg/d plant at Qurayyah, Saudi Arabia, where disinfected seawater is filtered, deaerated, and pumped 100 km inland for injection.

Shipboard applications account for many seawater electrolysis installations worldwide. In this case, smaller units are used to inhibit marine growth in seawater systems for sanitary services and for distilled water treatment. Fishing fleets also use hypochlorite as a disinfectant for storage.

Coastal utilities and industrial plants use hypochlorite from seawater electrolysis to control mollusks, algae, and slime, which may block seawater intakes, clog piping, and reduce heat transfer efficiency. Inland utility and industrial plants use brine hypochlorite cells when handling and safety are of primary concern. Nuclear power plants are particularly sensitive to the hazards of liquid chlorine.

Both coastal and inland wastewater treatment plants are major users of chlorine for disinfection before discharge. This was first started on the island of Guernsey in 1966 and was operated for 6 years [54].

On-site hypochlorite production from brine is also used for drinking water treatment, but it is only economical at remote locations or where safety is a major concern.

A number of companies have introduced on-site generators to supply chlorine demand for swimming pools. Many of these have been small chlor-alkali cells with separated anolyte and catholyte; but unseparated hypochlorite generators have gained acceptance and now represent an important portion of that market.

On-site hypochlorite generators are sometimes selected for other less common applications, such as textile manufacturing, industrial and laundry bleaching, cyanide destruction, odor control, ocean aquariums, and food processing [41], [55].

4.5. Economic Aspects

The Middle East now represents about half of the total market for electrolytic hypochlorite generation. The rest of the market is roughly balanced between East Asia, South America, and the United States. The total installed capacity has grown from ca. 200 t/d in 1979 to estimated 700 t/d in 1984. This probably represents total system sales approaching U.S. $\$200\times10^6$ through 1984.

On-site hypochlorite generation is difficult to compare with purchased chlorine or hypochlorite because chemical costs vary widely from nation to nation. Bulk liquid chlorine prices vary from ca. $ 150/t in the United States to well over $ 1000/t in remote locations, and hypochlorite prices vary from $1500/t of available chlorine content in the United States to over $ 2500/t where transportation cost is high. Therefore, a direct economic comparison with purchased chemicals can only be made for each location individually. There are many cases in which the selection of on-site generation is motivated primarily by economics, rather than by safety or convenience.

4.6. Plant Safety

Dilute sodium hypochlorite solution is much safer than liquid or gaseous chlorine. Sodium hypochlorite solution, as produced by on-site electrolytic generators (with an available chlorine content of 0.5 – 10.0 g/L), is regarded as corrosive and as an irritant when ingested or inhaled. It is also a mild skin irritant, and prolonged exposure may result in a burn or rash [56].

The primary safety concern associated with the electrolysis equipment is the explosion and fire hazard from byproduct hydrogen. This hazard is increased if oxygen is present in the cell gas. Oxygen content ranges from 3 % to 9 % in efficiently operating cells, but it varies widely in practice. Cells with anodes having poor selectivity for chlorine evolution or with poor design for operation at low temperature and low salinity may result in an oxygen concentration of 40 % or more in the cell gas.

The explosion limit for hydrogen – oxygen mixtures is 6.0 % O_2. Explosive impact will be soft at this point, but will increase quickly in severity as oxygen content increases. Two measures are taken in commercial cells against this: (1) dilute the byproduct hydrogen with sufficient air to less than 4.0 %, the explosion limit for hydrogen in air, or (2) allow the cell gas to vent with proper precaution, e.g., the installation of rupture disks.

A special hazard arises when cell deposits are washed with dilute hydrochloric acid. In addition to the hazards of handling hydrochloric acid, chlorine gas evolves when HCl is accidentally mixed with stored hypochlorite solution, or when the cells are started without first flushing out the acid.

5. Chlorine Dioxide

5.1. Properties

Chlorine dioxide [*10049-04-4*], ClO_2, M_r 67.45, is a yellowish-green to orange gas. It can be condensed to a reddish-brown liquid at 11 °C and solidified to orange red crystals at − 59 °C. The density of liquid ClO_2 is as follows [57]:

t, °C	− 33	− 21	− 17	5
ϱ, g/cm^3	1.907	1.788	1.735	1.635

Chlorine dioxide has an irritant, pungent odor that resembles that of a mixture of chlorine and ozone. Chlorine dioxide is an extremely unstable gas, readily decomposing into chlorine and oxygen even on mild heating. It is explosive as a gas or liquid at high concentration [58], [59]. However, it can be handled easily when it is diluted with air to less than 15 vol%.

Chlorine dioxide is easily soluble in water (heat of solution: − 26.8 kJ/mol); at 10 °C its solubility in water is 5 times that of chlorine (Table 5).

Chlorine dioxide can be easily driven out of aqueous solutions with a strong stream of air. When an aqueous solution is cooled, $ClO_2 \cdot 8\,H_2O$ crystals precipitate. Chlorine dioxide is also soluble in carbon tetrachloride, sulfuric acid, or acetic acid [61].

Other important properties of chlorine dioxide are its photochemical [62] and thermal [63] decomposition. Solutions of chlorine dioxide are relatively stable in the dark, but they decompose into chlorine and oxygen when exposed to light. Therefore, these solutions are very strong oxidants and extremely reactive and corrosive. They attack all metals except platinum, tantalum, and titanium.

In dilute aqueous solutions, chlorine dioxide oxidizes hydrocarbons to ketones and alcohols [64]. In the absence of water, chlorine dioxide loses its bleaching power.

5.2. Production

Because of the explosion risk, chlorine dioxide is manufactured on site. Its industrial production is based on the reduction of chlorate [65–67]. Undesirable byproducts are chlorine and chloride. For analysis of chlorine dioxide, see [68–71].

Reaction Mechanism [72], [73]. The basic reaction mechanism of chlorine dioxide formation is the same for all known processes. All processes use chlorate as the raw material; in all processes chlorine dioxide formation takes place in strong acidic solutions; and byproduct chloride is found in all generator solutions.

The presence of chloride ions is essential for the formation of chlorine dioxide [74], [75]; this is evident from the following facts:

1) No significant amount of chlorine dioxide is formed in acidified chlorate solutions with various reducing agents when chloride has previously been removed from the reaction system by adding silver sulfate.
2) If chlorine dioxide generators are allowed to stand overnight, minor ClO_2 formation consumes all of the chloride present in the reaction mass. When the supply of fresh raw materials is started the next morning, a specific chloride level must build up before chlorine dioxide can be produced at the desired rate. This chloride buildup phase can be eliminated by deliberately adding fresh chloride to the generator in the beginning.

Independent on the choice of reducing agent, the primary reaction for chlorine dioxide production is the reaction between chloric acid and hydrochloric acid to form chlorine dioxide and chlorine. Traces of Mn^{2+} and Ag^+ ions catalyze the reaction [76].

$$HClO_3 + HCl \longrightarrow HClO_2 + HClO$$
$$HClO_3 + HClO_2 \longrightarrow 2\,ClO_2 + H_2O$$
$$HClO + HCl \longrightarrow Cl_2 + H_2O$$
$$2\,HClO_3 + 2\,HCl \longrightarrow 2\,ClO_2 + Cl_2 + 2\,H_2O$$

In industrial chlorine dioxide production, sulfur dioxide, hydrochloric acid, or methanol are

Table 5. Solubility of chlorine dioxide in water [60]

25 °C		40 °C		60 °C	
p, mbar	*c*, g/L	*p*, mbar	*c*, g/L	*p*, mbar	*c*, g/L
46	3.01	74.9	2.63	141.2	2.65
29.5	1.82	45.7	1.6	71.6	1.18
17.9	1.13	25.2	0.83	28.4	0.58
11.2	0.69	13.2	0.47	16.0	0.26

used as reducing agents. Other reducing agents are not economical.

5.2.1. Day – Kesting Process [74], [77], [78]

This process combines the production of chlorine dioxide from sodium chlorate, with hydrochloric acid as reducing agent, and electrochemical sodium chlorate production from sodium chloride:

$$2\,NaCl + 6\,H_2O \xrightarrow{\text{Electrolysis}} 2\,NaClO_3 + 6\,H_2$$
$$2\,NaClO_3 + 4\,HCl \longrightarrow 2\,NaCl + 2\,ClO_2 + Cl_2 + 2\,H_2O$$

The efficiency of chlorine dioxide production depends on how far the rate of the competing chlorine production can be decreased:

$$NaClO_3 + 6\,HCl \longrightarrow 3\,Cl_2 + NaCl + 3\,H_2O$$

This can be achieved by maintaining a high chlorate concentration and a low hydrochloric acid concentration in the reaction system.

Chlorate solution is rapidly circulated from a large storage tank through the electrolytic cells, which oxidize chloride to chlorate. From the same tank the chlorate solution is also circulated slowly through the chlorine dioxide generator. The reaction takes place in a heated multicompartment column. The reactant solutions, chlorate and hydrochloric acid, are added from the top and air is introduced from the bottom of the column [79]. Chlorate is reduced with hydrochloric acid to chlorine dioxide and chlorine. Both gases are stripped from the system with air; at the same time, chlorine dioxide is diluted below the explosion limit [80]. The depleted chlorate solution is then returned to the storage tank and recycled to the chlorate electrolysis. The mixture of chlorine dioxide and chlorine is stripped with water to give a solution rich in chlorine dioxide [80], [81] and a gas phase rich in chlorine. Chlorine can then be recycled to the process as hydrochloric acid by reduction with hydrogen produced in the electrolytic cells or neutralized with NaOH solution to produce hypochlorite.

Although it has always been cheaper to produce chlorine dioxide by this process, it never gained momentum because of the high initial capital expenditure involved.

Lurgi [79], [82–86], Chemetics [87], [88], and others [89–91] have developed integrated systems that combine this process with the production of chlorine, caustic soda, and hydrochloric acid. Such combined systems use sodium chloride and electric power as the raw materials and produce chlorine dioxide solution and caustic soda (Fig. 7).

5.2.2. R2 Process [92–100]

Hooker Chemical Corporation, together with Electric Reduction Company, developed the R2 process, starting from the single-vessel process [101–103], and Electric Reduction Company further improved the details. Chlorine dioxide is produced according to the following overall reaction:

$$2\,NaClO_3 + 2\,NaCl + 2\,H_2SO_4$$
$$\longrightarrow 2\,ClO_2 + Cl_2 + 2\,Na_2SO_4 + 2\,H_2O$$

An undesirable side reaction produces additional chlorine:

$$NaClO_3 + 5\,NaCl + 3\,H_2SO_4$$
$$\longrightarrow 3\,Cl_2 + 3\,Na_2SO_4 + 3\,H_2O$$

In continuous industrial processes, the molar ratio of chlorine dioxide to chlorine is nearly 2 : 1, indicating that the ClO_2 production efficiency is almost 100 % (based on chlorate).

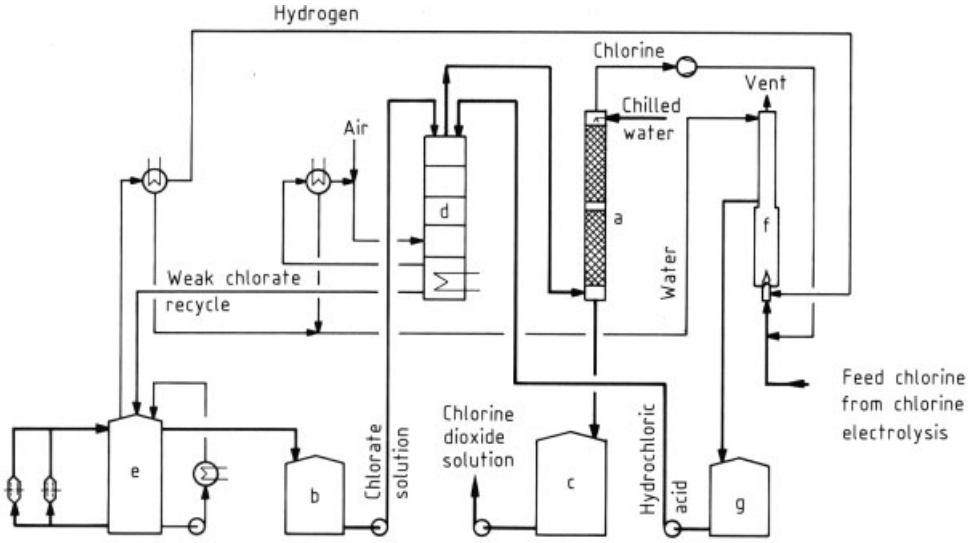

Figure 7. Schematic diagram of the integrated Day – Kesting process [82]
a) Chlorine stripper; b) Sodium chlorate solution; c) Chlorine dioxide solution storage; d) Chlorine dioxide generator; e) Chlorate electrolysis; f) Hydrochloric acid furnace; g) Hydrochloric acid storage

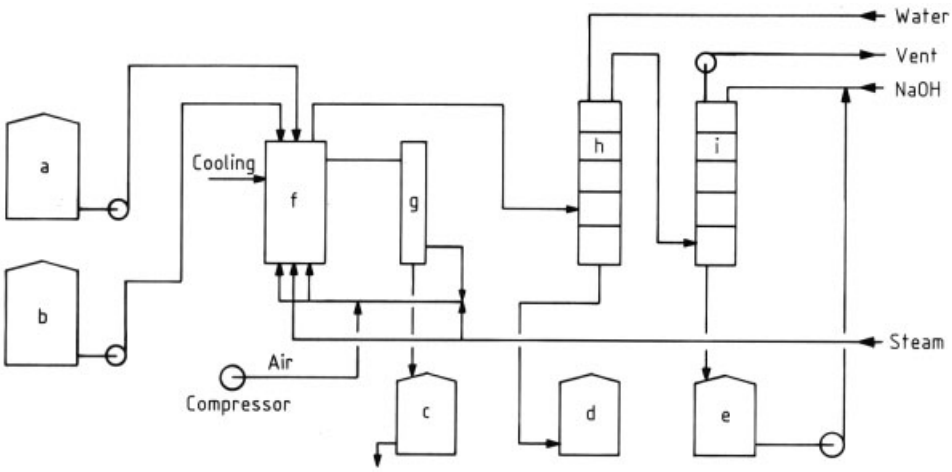

Figure 8. Schematic diagram of the R2 process [104]
a) $NaCl + NaClO_3$ storage; b) Acid storage; c) Spent acid; d) ClO_2 solution storage; e) NaOCl solution storage; f) ClO_2 generator; g) Stripper; h) ClO_2 absorption tower; i) Cl_2 absorption tower

Figure 8 shows a schematic diagram of the R2 process. Concentrated solutions of sodium chlorate, sodium chloride, and sulfuric acid (equimolar ratio) are metered to a vigorously agitated reaction vessel. Air is blown into the reactor through porous plates. Under optimum conditions, the reaction mass contains 0.1 – 0.2 mol/L of sodium chlorate, 4.5 – 5 mol/L of sulfuric adic and 0.02 – 0.08 mol/L of sodium chloride. Chlorine dioxide is absorbed from the gas phase in packed towers in cold water, and chlorine leaves the system as byproduct. The liquid effluent from the reactor is a mixture of sodium sulfate and sulfuric acid. The process can also be operated in such a way that sodium hydrogen sulfate crystallizes from the effluent solution [105], [106]. It provides sulfate for kraft pulping; the sulfuric acid is recovered for reuse.

The R2 process has been modified [106] and optimized [103], [107], [112] to suit the require-

ments of the industry. Currently, it is probably the most extensively used process for the production of chlorine dioxide.

5.2.3. Mathieson Process [67], [74], [96], [98], [104]

Sulfur dioxide is the reducing agent in the Mathieson process. The main overall reaction is as follows:

$$2\,NaClO_3 + H_2SO_4 + SO_2 \longrightarrow 2\,ClO_2 + 2\,NaHSO_4$$

Solutions of sodium chlorate and sulfuric acid are added continuously to a relatively large, cylindrical, lead-lined tank (primary reactor) from the top. Sulfur dioxide, diluted with air, is introduced through gas diffusion plates at four points in the bottom of the tank. The reaction mass overflows to a smaller secondary reactor of similar construction. The generated chlorine dioxide still contains some unreacted sulfur dioxide; it is stripped from the reaction mass by air and then washed in a scrubber packed with Raschig rings. The scrubber is installed on top of the primary generator, and the fresh sodium chlorate solution serves as a washing liquid on its way down to the primary generator. The mixture of chlorine dioxide, air, and chlorine coming out of the scrubber goes to the absorption tower.

5.2.4. Solvay Process [67], [74], [96], [98], [113]

The Solvay process uses methanol as reducing agent. The main overall reaction is as follows:

$$2\,NaClO_3 + CH_3OH + H_2SO_4 \longrightarrow 2\,ClO_2 + HCHO + Na_2SO_4 + 2\,H_2O$$

The reaction between sodium chlorate, methanol, and sulfuric acid takes place in two jacketed, lead-lined steel reactors. The chemicals are added to the bottom of the first vessel and flow by gravity from one reactor to the other. Each vessel has its own supply of methanol and air. The two reactors are operated at different temperatures and chlorate concentrations. Up to 70 % of the reaction is completed in the first reactor. Additional methanol and up to 10 % of the total acid requirement is fed to the second reactor; the conversion of chlorate is then more than 95 %.

5.2.5. Other Processes

The *Holst process* [114–116] is a batch version of the Mathieson process and has not been exploited because of its low chlorine dioxide yield. When solid sodium chlorate is successively added, the unreacted sulfuric acid from the previous operating period can be utilized and the conversion efficiency of chlorate increases to 83 %.

The *Persson process* [74], [117] has only historical value. The main raw materials were sodium chlorate and sulfur dioxide, but the latter was used to reduce chromic acid to chromic sulfate, and this in turn reduced chlorate to chlorine dioxide.

The *CIP process* [118], [119] was in operation long before it was published, maintaining secrecy for several years. Concentrated sodium chlorate solution is carefully metered and distributed to a packed reaction tower. Sulfur dioxide gas, diluted with air or nitrogen, enters from the bottom, reacts with the chlorate, and is oxidized to sulfuric acid. Chlorate is reduced to chlorine dioxide. The additional inert gas dilutes the ClO_2 formed to a safe concentration, and the mixture is led to the absorption tower.

Small-Scale Production from Sodium Chlorite [120–123]. Small-scale consumers produce chlorine dioxide by passing chlorine gas through a sodium chlorite solution:

$$2\,NaClO_2 + Cl_2 \longrightarrow 2\,NaCl + 2\,ClO_2$$

The reaction is almost quantitative.

The alternative acidification of sodium chlorite solution with hydrochloric acid is also commonly used:

$$5\,NaClO_2 + 4\,HCl \longrightarrow 4\,ClO_2 + 2\,H_2O + 5\,NaC$$

5.3. Economic Aspects

The economy of chlorine dioxide production depends on the efficient utilization of the byproducts of the process [97], [105]. When the chlorine dioxide production is combined with chloralkali membrane cells and hydrochloric acid

synthesis (Lurgi concept), chlorate cells work most efficiently [82–90]. Byproducts from one unit can then be used as raw materials or make-up chemicals for the others. In this way, high initial capital cost is paid back in the long run by a far lower operating cost. The R2 process offers similar possibilities if it is combined with a paper mill [95], [107], [111].

5.4. Uses

Chlorine dioxide is the most widely used bleaching agent, in particular for high-quality cellulose. It destroys lignin without attacking cellulose, yielding a characteristically white cellulose. The general trend is to eliminate chlorine and hypochlorite as bleaching agents altogether [124], [125] and replace them with chlorine dioxide [81], [126], [127].

Chlorine dioxide is used in the pulp and paper, textile, and food industries. In the pulp and paper industry, a unique whiteness can be achieved in kraft pulp, sulfite pulp, and soda pulp [96]. In the textile industry, chlorine dioxide produces high-quality fibers with additional special advantages. Shrinkproof wool owes its quality to the reaction of chlorine dioxide with the cross-linking sulfur atoms of the wool.

Chlorine dioxide is also used in sanitization, e.g., of industrial and municipal waters, sewage, algae, or decomposed vegetables. Waterworks use chlorine dioxide to handle taste and odor problems of household water [68], [122], [123], [128].

6. Sodium Chlorite

6.1. Properties

Sodium chlorite [*7758-19-2*], $NaClO_2$, M_r 90.45, is the sodium salt of the unstable chlorous acid; it exists as an anhydrous and a trihydrated form (transition point 38 °C). Very pure $NaClO_2$ crystals are white, but they usually have a greenish tint because traces of chlorine dioxide are present.

The stability of sodium chlorite lies between that of hypochlorite and chlorate. For further information, see [129]. Sodium chlorite is not sensitive to impact if organic matter is excluded. It can be struck with a clean metal surface without detonation. However, in the presence of organic matter – the film usually occurring on a hammer suffices – the impact may result in a spontaneous puffing. As a strong oxidizing agent, solid sodium chlorite forms explosive mixtures with such oxidizable materials as sulfur, powdered coal, metal powders, or organic compounds. Sodium chlorite solutions should never be allowed to dry on fabrics because this would result in a flammable combination.

Sodium chlorite is soluble in water: [129]

t, °C	5	17	20	30	40	45	50	60
Solubility, wt %	34	39	40.5	46	50.7	53	53.7	55

More important is its solubility in the presence of caustic soda (Table 6), sodium chlorate [130], sodium chloride, [131], and sodium carbonate [132]. For the system $NaClO_2 - NaCl - NaClO_3 - H_2O$, see [133], [134].

Aqueous solutions of sodium chlorite must be protected from light. At low pH (approximately 2), chlorite solutions contain chlorous acid that decomposes to form chlorine dioxide and chlorate:

$$4\,HClO_2 \longrightarrow 2\,ClO_2 + HClO_3 + HCl + H_2O$$

At pH 3 – 4, decomposition slows down. Alkaline solutions are stable, and dilute solutions can even be boiled without decomposition. Concentrated alkaline solutions of sodium chlorite slowly decompose when heated [135–137]:

$$3\,NaClO_2 \longrightarrow 2\,NaClO_3 + NaCl$$

The reaction of sodium chlorite with hypochlorite depends on pH [137]. At low pH, the reaction produces chlorine dioxide, whereas at high pH, chlorate is formed. The reaction with chlorine produces chlorine dioxide and sodium chloride:

$$2\,NaClO_2 + Cl_2 \longrightarrow 2\,ClO_2 + 2\,NaCl$$

Crystalline sodium chlorite is slightly hygroscopic without caking; it is stabilized with alkali for long-term storage. When heated to 180 – 200 °C, it decomposes partially to sodium chlorate and sodium chloride or completely to sodium chloride and oxygen [129]. For the analysis of sodium chlorite, see [69], [70], [138–140].

Table 6. Solubility in the $NaOH-NaClO_2-H_2O$ system at 30°C

d_4^{30}	c_{NaOH}		c_{NaClO_2}		Solid phase
	wt %	g/L	wt %	g/L	
–	55.3	–	0	–	
–	51.1	–	1.5	–	$NaOH-H_2O$
1.568	51.2	802.8	1.7	26.7	
1.575	50.6	797.0	2.45	38.6	$NaOH+NaClO_2$
1.546	48.9	756.0	2.5	38.7	
1.519	42.3	627.3	5.1	77.5	
1.496	35.9	537.1	7.65	114.4	
1.474	31.5	464.3	10.5	154.8	
1.453	26.8	389.4	14.9	216.5	$NaClO_2$
1.457	23.6	343.9	18.7	272.5	
1.447	20.3	293.7	22.4	322.7	
1.451	16.0	232.2	28.2	409.7	
1.457	15.4	224.4	29.2	425.4	
1.469	10.3	151.3	36.8	540.6	
1.469	9.5	139.6	38.0	558.2	
1.471	9.0	132.4	38.8	570.7	
1.473	7.86	115.7	40.5	596.6	
1.473	6.2	91.3	42.9	631.9	unstable
1.468	5.8	85.1	43.1	632.7	unstable
1.476	6.1	90.0	42.6	628.8	
1.441*	3.55	51.2	43.5	626.8	$NaClO_2 \cdot 3\,H_2O$
1.410**	0.0	–	45.7	644.7	

* The solid phase was analyzed.
** The solutions contain traces of NaCl (0.06 wt %).

6.2. Production

Sodium chlorite is produced by treating chlorine dioxide with caustic soda [141–143]:

$$2\,ClO_2 + 2\,NaOH \longrightarrow NaClO_2 + NaClO_3 + H_2O$$

The reaction products, sodium chlorite and sodium chlorate, have nearly the same solubility in water and are difficult to separate. Industrial sodium chlorite production uses the following procedure [144]. The absorption of chlorine dioxide in caustic soda solution and the simultaneous reduction of chlorate to chlorite are achieved by a suitable reducing agent, such as hydrogen peroxide [145]:

$$2\,ClO_2 + 2\,NaOH + H_2O_2 \longrightarrow 2\,NaClO_2 + 2\,H_2O + O_2$$

The product is a 33 wt % solution of sodium chlorite, which is then converted to a dry solid containing ca. 80 wt % of sodium chlorite, the rest being stabilizers.

Numerous complex chlorites have been prepared [146], but only sodium chlorite has proven to be of any commercial value.

Efforts to produce sodium chlorite by electrolysis, similar to hypochlorite or chlorate, have not yet been successful. Its synthesis from chlorine dioxide and sodium amalgam [147] could not be realized because redox potential [148] and pH [149] were difficult to control. Other processes for the direct reduction of chlorine dioxide were not efficient [150], [151].

6.3. Uses

Sodium chlorite is a very efficient bleaching agent. Its oxidation potential allows a controlled bleaching that is not attainable with other bleaching agents. Therefore, it is widely used as a bleaching agent in the textile industry [135–137], [152–155].

Another important use of sodium chlorite is the small-scale production of chlorine dioxide; see Section 5.2.5 [120–123].

7. Chloric Acid and Chlorates

Chloric acid is not being produced on an industrial scale. However, sodium chlorate and potassium chlorate have outstanding industrial significance. Sodium chlorate is produced on a very

large scale by one of the most important inorganic electrosyntheses. All other chlorates are produced in much smaller amounts for special purposes, usually from sodium chlorate.

Table 7. Physical properties of sodium chlorate and potassium chlorate

	$NaClO_3$ [*7775-09-9*]	$KClO_3$ [*3811-04-9*]
M_r	106.44	122.55
Crystal system	cubic	monoclinic
mp, °C	260	356
Enthalpy of fusion, kJ/mol	21.3	
Density, g/cm^3	2.487 (25 °C) 2.385 (252 °C)	2.338 (20 °C)
Molar heat capacity, J mol^{-1} K^{-1}	54.7 + 0.155 T ∗ (298 – 533 K)	99.8 (20 °C)
Standard enthalpy of formation, kJ/mol	− 365.8 (cryst.) − 344.1 (ai)∗∗	− 391 (cryst.)
Standard entropy, J mol^{-1} K^{-1}	123.4 (cryst.) 221.3 (ai)∗∗	143 (cryst.)
Enthalpy of dissolution (200 mol of H_2O/mol of chlorate, 25 °C), kJ/mol	+ 21.6	+ 40.9

∗ T = temperature, K.
∗∗ ai = ideal solution of unit activity.

7.1. Properties

Physical Properties. Anhydrous *chloric acid* [*7790-93-4*], $HClO_3$, M_r 84.46, is unstable and explosive. Dilute aqueous solutions are colorless and odorless; they are stable at low temperature if catalytically active contaminants are excluded. In the presence of such catalysts, the solutions may decompose vehemently, particularly at elevated temperature. Addition of polyphosphates or hydrogen peroxide lowers the decomposition tendency. Above 95 °C, pure chloric acid solutions decompose to form chlorine dioxide, chlorine, oxygen, and perchloric acid, but in the presence of hydrochloric acid, the products are chlorine and chlorine dioxide.

Concentrated chloric acid is a strong oxidant. In addition to the noble metals, only Hastelloy C exhibits satisfactory resistance against corrosion. At low temperature, dilute chloric acid may be kept in containers made of poly(vinyl chloride). At 18 °C, the density of an aqueous 13 % $HClO_3$ solution is 1.080 g/cm^3 while that of a 25 % solution is 1.166 g/cm^3.

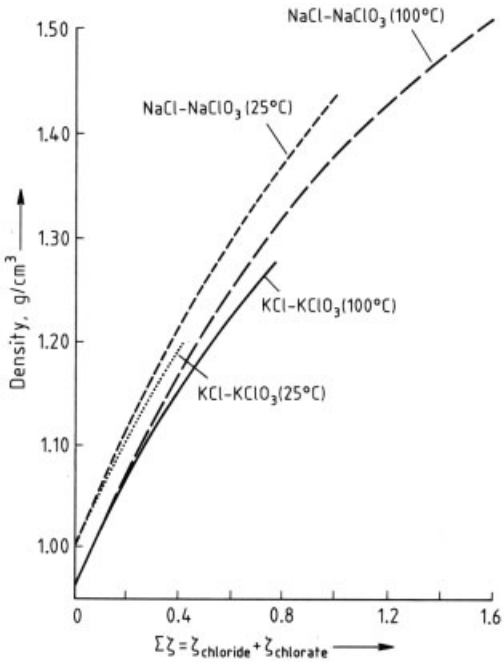

Figure 9. Densities of aqueous solutions of NaCl – $NaClO_3$ and KCl – $KClO_3$ [156]
$\Sigma\zeta$ is the sum of mass ratios $\zeta_{chloride}$ (kg of chloride/kg of H_2O) and $\zeta_{chlorate}$ (kg of chlorate/kg of H_2O)

Physical properties of *sodium chlorate* and *potassium chlorate* are shown in Table 7. Figure 9 shows the densities of aqueous chloride – chlorate solutions. Figure 10 shows the relative vapor pressure depression of aqueous solutions of chloride and chlorate. The electric conductivity of pure chlorate solutions is given in Table 8; for further data, see [158]. Solubility data of the aqueous chloride – chlorate system are shown in Figures 11 and 12 and in Table 9. The mass fraction of saturated sodium chlorate solutions in the range from 0 to 100 °C can be calculated from

$$w = 0.445 + 0.00226t$$

where

t = temperature, °C
w = mass fraction, kg of $NaClO_3$/kg of solution

The freezing point depression ΔT (in K) of aqueous solutions of sodium chlorate is given by

$$\Delta T = 33.64\zeta - 115.1\zeta^2$$

where ζ = mass ratio, kg of $NaClO_3$/kg of H_2O

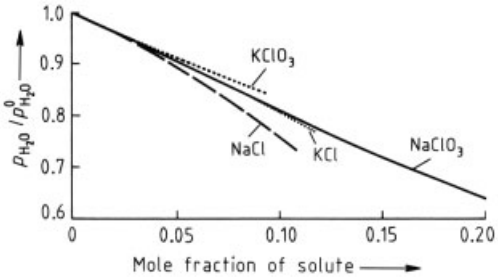

Figure 10. Vapor pressure depression of aqueous chloride and chlorate solutions

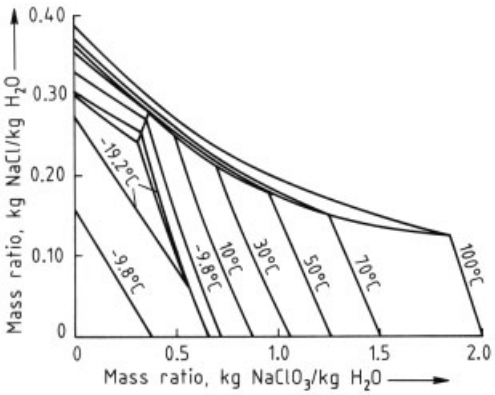

Figure 11. Solubility of aqueous solutions of NaCl and $NaClO_3$ [159]

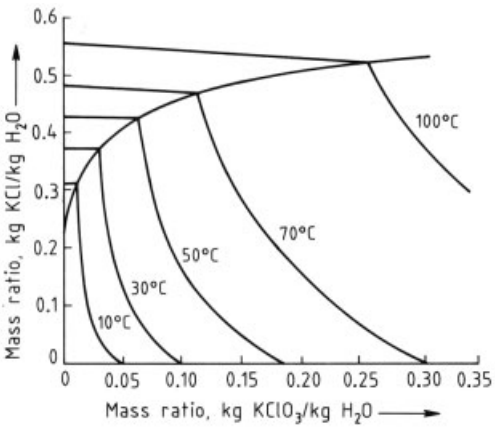

Figure 12. Solubility of aqueous solutions of KCl and $KClO_3$ [159]

Chemical Properties. Chlorates decompose to yield oxygen. They form flammable and explosive mixtures with organic substances, phosphorus, ammonium compounds, some sulfur compounds, and some metal salts, oxidizable solvents, or other oxidizable substances. Potassium chlorate decomposes below the melting temperature. Alkaline chlorate solutions do not exhibit strong oxidizing properties. With decreasing pH, however, the oxidizing activity of chlorate solutions increases. Concentrated acidic solutions are vigorous oxidants as a result of chloric acid formation.

Solutions containing more than 30 % $HClO_3$ decompose spontaneously. In the presence of organic matter or reducing agents the reaction may be violent, especially at elevated temperature.

7.2. Production Fundamentals

Sodium chlorate (and to a minor extent potassium chlorate) is produced by electrolysis of an aqueous sodium chloride (potassium chloride) solution. Hypochlorite forms as an intermediate that is further oxidized to chlorate along two competing reaction paths. The concentrated chlorate solution is either submitted to crystallization (see Section 7.4) or it is used directly, particularly in the production of chlorine dioxide.

An aqueous solution containing 450 – 550 g/L of sodium chlorate and 90 – 100 g/L of sodium chloride is generated directly by electrosynthesis; it can be used as feed in the Kesting process (Section 5.2.1) [160]. The spent solution from that process, containing 140 g/L of sodium chlorate and 220 g/L of sodium chloride, is then fed back to the electrolysis system.

The other, less common chlorates are chemically formed by conversion of sodium chlorate with the corresponding chloride. Other production methods, such as the chemical formation of chlorates by introducing gaseous chlorine into a warm hydroxide solution, are now obsolete.

7.2.1. Chlorate-Generating Reactions

An aqueous solution of sodium chloride is electrolyzed, usually in cells without a diaphragm.

Table 8. Electric conductivity of pure sodium chlorate solutions [157]

Concentration,	Conductivity, $\Omega^{-1}\,m^{-1}$		
g/L	20 °C	40 °C	60 °C
100	6.2	8.9	11.8
200	10.4	14.9	19.7
300	13.4	18.9	25.0
400	15.0	21.5	28.5
500	15.7	22.7	30.3
600	15.5	23.1	30.8
750		21.7	29.7

Table 9. Mass ratio (kg of salt/kg of H_2O) of the solution in equilibrium with crystalline chloride and chlorate [158]

t, °C	$NaCl-NaClO_3$ solution		$KCl-KClO_3$ solution	
	NaCl	$NaClO_3$	KCl	$KClO_3$
− 9.8	0.270	0.360	0.2466	0.0056
+ 10	0.249	0.499	0.3123	0.0144
30	0.2125	0.706	0.3703	0.0321
50	0.1785	0.958	0.4226	0.0635
70	0.1495	1.238	0.4651	0.1162
100	0.1245	1.85	0.518	0.2588

Hydrogen and sodium hydroxide are formed at the cathode, while chloride is discharged at the anode. Chlorine does not evolve as a gas, but undergoes hydrolysis (Fig. 13):

$$Cl_2 + H_2O \rightleftharpoons HClO + H^+ + Cl^- \quad (4)$$

Chlorate then forms simultaneously by two competing reactions: (1) predominantly by autoxidation of hypochlorite in the bulk electrolyte, and (2) to a small extent (ca. 20 %) by anodic chlorate formation. The detailed mechanism of the reaction was essentially clarified at the turn of the century, but continues to be the subject of intense studies [39, b].

Autoxidation. Autoxidation of hypochlorous acid, also called chemical chlorate formation, is a homogeneous reaction that proceeds according to the following overall equation, which gives no information on the individual steps:

$$3\,HClO \longrightarrow ClO_3^- + 2\,Cl^- + 3\,H^+ \quad (5)$$

Autoxidation is preceded by dissociation of a part of the total hypochlorous acid involved in autoxidation:

$$HClO \longrightarrow ClO^- + H^+ \quad (6)$$

This dissociation can occur to a significant degree only at some distance from the anode, where the electrolyte is sufficiently buffered by hydroxyl formed by the cathode reaction. The hypochlorite formed then reacts with the complementary amount of hypochlorous acid:

$$2\,HClO + ClO^- \longrightarrow ClO_3^- + 2\,Cl^- + 2\,H^+ \quad (7)$$

Equation (7) shows the classical form of Foerster [161], [162], which was attacked [163] but has since been confirmed [164], [165].

Anodic Chlorate Formation. The mechanism of anodic chlorate formation was established by Landolt and Ibl [166]. Hydrolysis of chlorine (Eq. 4) is considered to be fast. Because of the formation of H^+ ions, one might expect the electrolyte in the anodic boundary layer to be strongly acidic. However, this is the case only at low chloride concentration. Large chloride concentrations, as they occur in industrial processes, shift the hydrolysis equilibrium (Eq. 4) to the left. In the electrolyte layer adjacent to the anode, the H^+ concentration is too small to permit noticeable diffusion of H^+ into the inner electrolyte. Therefore, hydrogen is transported away from the anode as hypochlorous acid rather than H^+. In the bulk electrolyte where the pH is high, hypochlorous acid is largely dissociated. The hypochlorite ion diffuses back to the anode

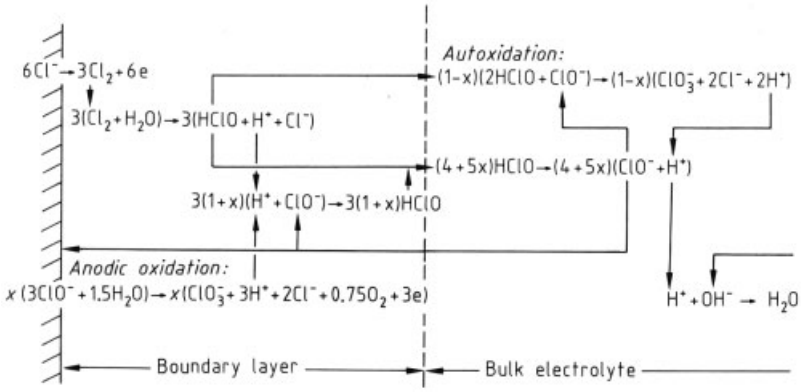

Figure 13. Chlorate formation by autoxidation in bulk electrolyte and by anodic oxidation

and more than two-thirds of it is consumed by buffering before reaching the anode. However, less than one-third of that hypochlorite is discharged at the anode to form chlorate and oxygen [166–168]:

$$3\,ClO^- + 1.5\,H_2O \longrightarrow ClO_3^- + 3\,H^+ + 2\,Cl^- + 0.75\,O_2 + 3\,e \qquad (8)$$

The stoichiometry of the anodic chlorate formation (Eq. 8) was recently reviewed [39, b].

As shown in Figure 13 the discharge of 6 mol of chloride yields 1 mol of chlorate; this is independent of the reaction route, either autoxidation or anodic discharge of hypochlorite. However, the anodic oxidation (Eq. 8) requires 50 % additional electric energy. Therefore, industrial processes endeavour to suppress the anodic oxidation in favor of the autoxidation. The effective means is a short residence time of the electrolyte solution inside the electrochemical reactor by applying large flow rates using rather short electrode lengths. Thereby the average bulk concentration of hypochlorite ($ClO^- + HClO$) is lowered and the undesired anodic chlorate formation controlled by mass transfer is minimized. However, increasing flow rates increase the mass transfer coefficient, thus counteracting the beneficial effect of short residence time. The optimum is at elevated values of the flow rate; this method to suppress anodic oxidation is used in all modern industrial chlorate systems. Under these conditions, chlorate is formed predominantly outside the interelectrode space, e.g., in a separate chemical reactor where the route is necessarily restricted to autoxidation (Fig. 14).

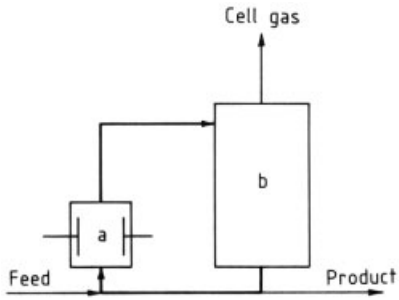

Figure 14. Concept of separate electrochemical reactor (a) and chemical reactor (b) for chlorate electrosynthesis

7.2.2. Loss Reactions

In addition to anodic chlorate formation, further loss reactions decrease current efficiency and must be suppressed in industrial systems.

Cathodic Reduction of Hypochlorite. The main loss occurs by cathodic reduction of hypochlorite, as described in Section 4.2.1:

$$ClO^- + H_2O + 2\,e \longrightarrow Cl^- + 2\,OH^-$$

The reaction competes with hydrogen generation. Without special precautions, its rate is controlled by mass transfer and, therefore, linearly proportional to the hypochlorite concentration in the bulk electrolyte. The reaction can widely be suppressed by adding a small amount of dichromate (1 – 5 g/L) to the electrolyte solution. By cathodic deposition a porous surface film of chromium hydroxide forms [169]. Inside

the pores of this film, the effective current density increases, and the resulting potential gradient across the film impedes the diffusion of any kind of anions to the cathode, whereas the access of cations to the cathode and their reduction are facilitated. The protective film ceases to grow after a certain thickness is reached; the increase in cell voltage caused by the ohmic resistance of the film is more than compensated by the gain in current efficiency [165], [170]. Increasing the electrolyte temperature favors mass transfer of hypochlorite to the cathode and, therefore, requires more dichromate. When the current is interrupted, the surface layer dissolves. It takes some time to reestablish a new film after the current is switched on, but cathodic reduction immediately takes place. Therefore, cells should be operated continuously, and a minimum cathodic current density is recommended to prevent dissolution of the film [171]; from industrial experience, a current density of 20 A/m^2 is considered satisfactory.

Cathodic Reduction of Chlorate. Chlorate is also reduced at the cathode, but this reaction is less significant than the reduction of hypochlorite described in the preceding paragraph:

$$ClO_3^- + 3\,H_2O + 6\,e \longrightarrow Cl^- + 6\,OH^-$$

The rate of reduction is strongly affected by the cathode material; reduction at iron is much faster than at nickel and platinum. Chlorate reduction is also restrained by the chromium hydroxide layer on the cathode [16], [171–174]. In closed-loop electrolyte systems (e.g., Fig. 14), the presence of residual dichromate in the solution has a beneficial effect because its additional buffering capacity stabilizes the various reaction equilibria [165], [175], [176].

Catalytic Decomposition of Hypochlorite. Hypochlorite solutions decompo e in the presence of catalytically active contaminants:

$$2\,ClO^- \longrightarrow 2\,Cl^- + O_2$$

Oxides of nickel, cobalt, and copper are very effective; their catalytic activity decreases in that order. The catalytic action of manganese, iron, lead, and tin has been demonstrated, but is at least one order of magnitude smaller [53], [177]. Autocatalytic decomposition is also possible, but its extent is negligible at ambient temperature [177], [178].

Chlorine Desorption. Hydrogen and oxygen gas bubbles evolved at the cathode and anode, respectively, tend to desorb some chlorine from the solution. This amount of chlorine is lost for chlorate production unless it is recovered from the cell gas by absorption outside the cell. Water vapor contained in the gas bubbles increases their volume and, thus, the desorption loss. Therefore, the operation temperature is limited to a maximum of 80 – 90 °C. Chlorine loss results not only in a corresponding loss in current efficiency, but also in a steady increase in pH. Therefore, chlorine gas or, more usually, hydrochloric acid must be permanently added to maintain the pH.

Further loss reactions are of minor importance: *perchlorate* may be formed by anodic oxidation of chlorate, particularly at low chloride concentration, provided the oxygen overvoltage is large enough (see Section 8.2.2). With the industrial anodes used in chlorate electrosynthesis, perchlorate formation is negligible. The *anodic decomposition of water,* which plays a major role in hypochlorite production, is negligible as long as the chloride concentration is larger than 100 g/L.

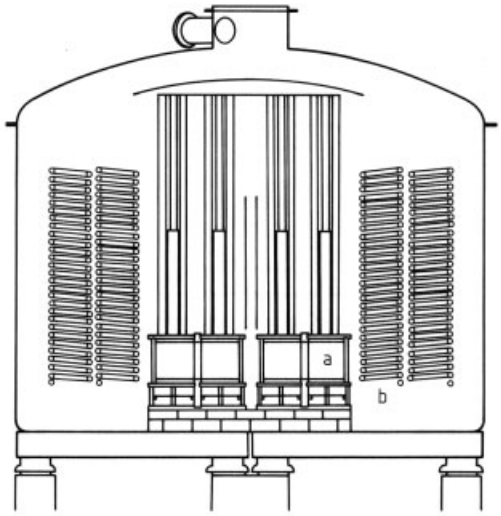

Figure 15. Huron electrosynthesis system with integrated electrolysis cells
a) Chemical reactor; b) Cooling coils

7.3. Industrial Electrosynthesis Systems

Chlorate cell design has changed much within the last 25 years by the introduction of coated titanium anodes [179] and by the systematic application of the results of chlorate formation theory. Hypochlorite autoxidation occurs at high temperature outside the interelectrode gap or even outside the electrochemical reactor to decrease the hypochlorite concentration before the electrolyte reenters the cell. The first industrial plant of this type was started in 1969 in Finland.

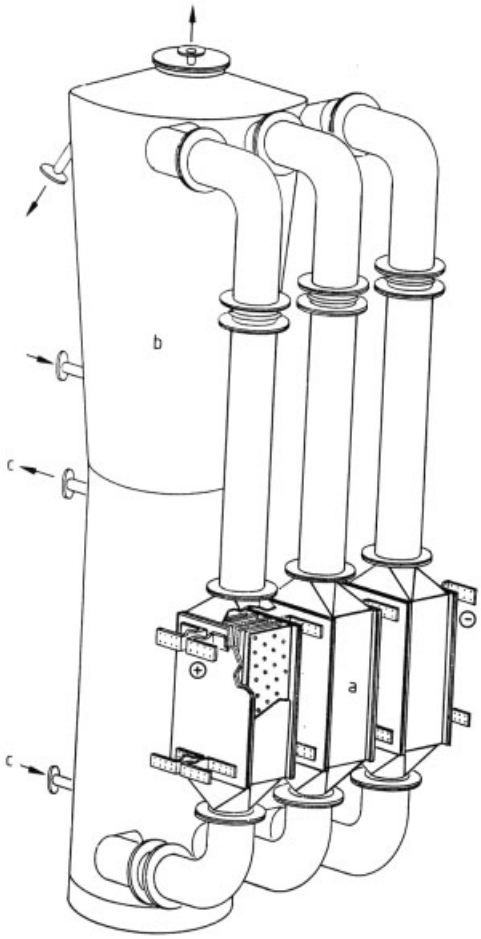

Figure 16. Electrosynthesis system with natural convection of electrolyte (Krebs, Paris)
a) Electrolysis cell; b) Chemical reactor; c) Cooler

The requirement for large electrolyte flow rate resulting from the considerations of Section 7.2 is taken into account in all modern cells. Hydrogen evolved at the cathode provides for a natural flow of the electrolyte and is used for its recirculation through the cell compartment and the chemical reactor without requiring additional pumping. Other systems use a mechanical pump to provide the required circulation rate. Hydrogen gas is released at the top of the unit. For safety reasons, the gas volume in the equipment must be minimized. To stabilize the pH value, the solution is continuously acidified (cf. Section 7.2.2.).

Some electrosynthesis systems combine the cell and the chemical reactor in a single unit (Huron, Fig. 15, and Atochem). Other industrial systems are composed of an electrochemical reactor and a separate chemical reactor [180] interconnected by electrolyte circulation pipes with integrated heat exchangers for the removal of excess heat (Krebs, Fig. 16, and Pennwalt, Fig. 17) or separate heat exchangers (Krebskosmo, Chemetics, and Fröhler – Lurgi – Uhde). None of these electrosynthesis systems is clearly superior from an engineering point of view.

7.3.1. Electrolysis Cell Types

Most cells are equipped with coated titanium anodes [181], and only some 5 – 10 % of all chlorate cells still use graphite anodes. Older cells have also been operated with anodes of magnetite (Fe_3O_4) [182], [183], platinum [184], graphite [185–188], or lead dioxide [189], [190]; reviews on these older anodes can be found in [39] and [191–193]. The following survey is restricted to modern cells with coated titanium anodes. Typical operational data are shown in Table 10.

Modern electrochemical reactors [194–196] cover the complete spectrum of unipolar and bipolar electrodes, including direct coupling of cells with unipolar electrodes, an arrangement sometimes called multipolar (Fig. 18).

Cells with *unipolar* electrodes (Krebs, Pennwalt, Fröhler – Lurgi – Uhde, and Atochem) are suitable for large currents up to 100 kA. Repair of one cell does not necessitate shutdown of the entire electrolyzer. Parasitic currents [197], corrosion, and malfunction of a cell unit are easy to locate. Pennwalt and Atochem use two cathodes per anode. This design allows circulation within the cell [194], [195]. Cells with *bipolar* elec-

Table 10. Typical operational data of sodium chlorate cells

Parameters	Krebs, Paris	Chemetics, Vancouver	Fröhler – Lurgi – Uhde	Pennwalt, Philadelphia	Atochem, Paris	Huron Chemicals, Kingston	Krebs, Zürich	Krebskosmo, Berlin
Current per cell, kA	75	15 – 85	6 – 50	30	30 – 75	480	30	17.5
Cell type	NC	DC	TDK	P8-3	TA2	cell-in-tank	ZMA	CZB 50 – 150
Cell connection	unipolar	multipolar	unipolar	unipolar	unipolar	multipolar	unipolar	bipolar
Anode surface, m^2	30	7 – 25	7.1 – 17.8	8.92	20/30	proprietary	10	30.6/91.8
Current density, kA/m^2	2.5	1.8 – 3.5	2.8	3.36	1.5 – 2.5	1.5 – 4	3.0	3.0
Operating voltage,								
min., V	2.85	2.70	2.9	3.4	2.92	3.1 at $3\,kA/m^2$	3.1	3.1
max., V	3.00	3.00	3.0	3.7	3.10	3.4 at $3\,kA/m^2$	3.3	3.4
Energy consumption (d.c.), kW h/t of $NaClO_3$	4500 – 4750	4250 – 4750	4610 – 4770	5650	4600 – 4900	5000 – 5500	5100	5000 – 5500
Current efficiency, %	95.5	96	95	96	96	> 93.5	95	94 – 96
Current concentration, A/L	30	25 – 30	22	20	13 – 22	6 – 20	20	25
Operating temperature, °C	80	80	75	90	70 – 80	65	80	80
pH of cell liquor	6.2	6 – 7	6.4	5.5 – 6.5	6.2 – 6.4	6.3 – 6.9	6.5	6.5
Interpolar distance, mm	4	–	3.5	4	6	proprietary	5	4
Kind of coating	$Pt/Ir/RuO_2$	IMI	platinate	Pt/Ir	RuO_2	Pt/Ir or RuO_2	$Pt/Ir/RuO_2$	noble metal oxide
Anode coating life, years	6 8 – 10	> 10	8 – 10	7 – 10	7 – 10	dep. on coating	> 6	4
Cathode material	spec. carbon steel	steel	mild steel	carbon steel	spec. steel	titanium	steel	steel
Cell liquor composition:								
Min. NaCl concentration, g/L	100	100	90	70 – 110	100	50	100	80
Max. $NaClO_3$ concentration, g/L	600	640	550	500 – 650	625	700	600	650
NaOCl concentration, g/L	1.3 – 1.8	–	3	1 – 2	1	–	2	2 – 3
$Na_2Cr_2O_7$ concentration, g/L	3 – 4	3 – 4	3	3 – 5	5	1.1	5	3
Cell gas composition:								
H_2, vol%	97.7 – 98.2	97.5	97.5	97.5	> 98	97 – 99	97.5	97.5
O_2, vol%	1.5 – 2	2	2	1.5	1 – 2	1.0 – 2.5	2	2
Cl_2, vol%	0.3	0.5	0.5	1.0	0.1 – 0.2	0.1 – 0.2	0.5	0.3
Requirements per t of $NaClO_3$ (100 %):								
NaCl, kg	560	560	–	560	550	–	550	570
HCl (33 %), kg	35	–	–	Cl_2	30	–	33	50
NaOH (30 %), kg	24	–	–	–	10	–	–	60
$Na_2Cr_2O_7 \cdot 2H_2O$, kg	0.05 – 0.1	–	–	5	0.3	–	neglig.	0.3
Production:								
$NaClO_3$ (100 %), kg/d	1137	–	–	446	–	7628	453	1370/4110
H_2, m^3 (STP)/d	753	–	–	291	–	4760	300	910/2730

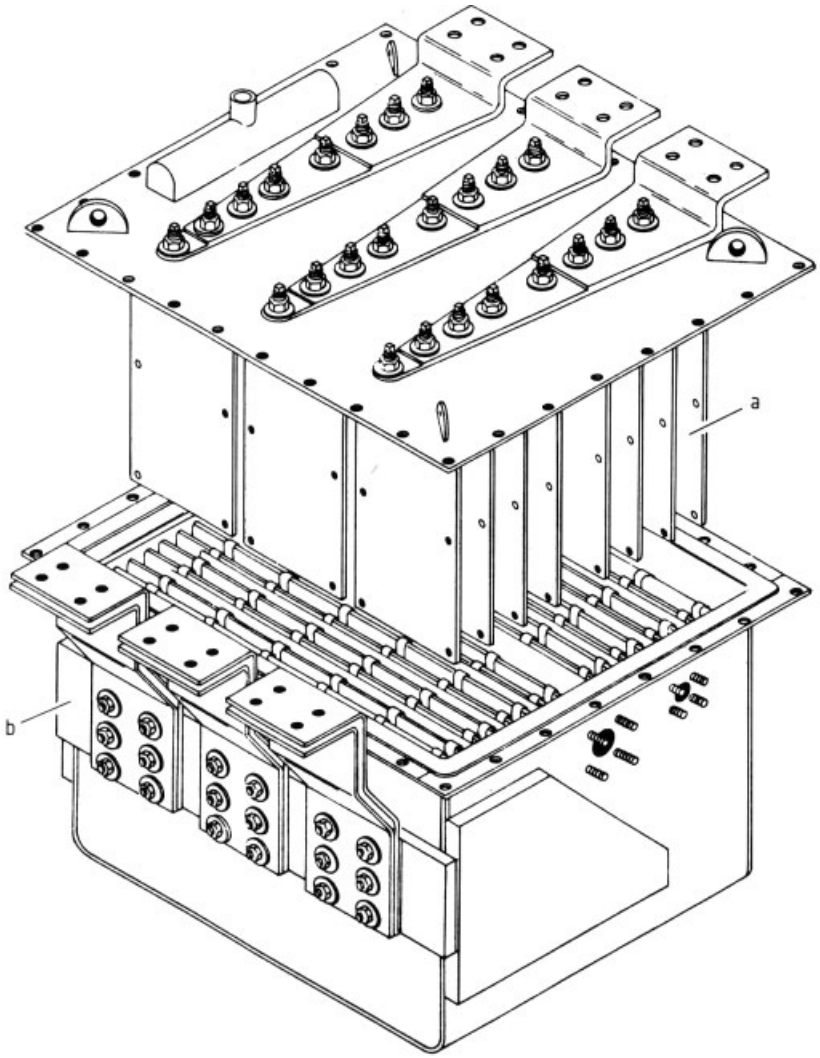

Figure 17. Pennwalt electrochemical reactor
a) Anodes; b) Cathodes

trodes (Krebskosmo) seldom exceed 30 kA (in charge of the voltage), thus lowering the rectifier cost. *Multipolar cells* (Chemetics and Huron) combine some of the advantages of cells with unipolar and bipolar electrodes and require a minimum of bus bars, resulting not only in savings in investment cost, but also in cell voltage. However, the entire electrolyzer system must be shut down for repair of a single unit.

7.3.2. Electrodes

Anodes. N. B. Beer made a decisive breakthrough in anode technology in the mid-1960s by introducing anode coatings based on ruthenium dioxide, RuO_2 [15], [198–201]. In fact, the anode surface is composed of a mixture of RuO_2 with other metal oxides, primarily titanium dioxide. These anodes maintain a low overvoltage for chlorine formation at high current densities for long periods without being consumed and changing their mechanical dimensions. Independent of the detailed nature of the mate-

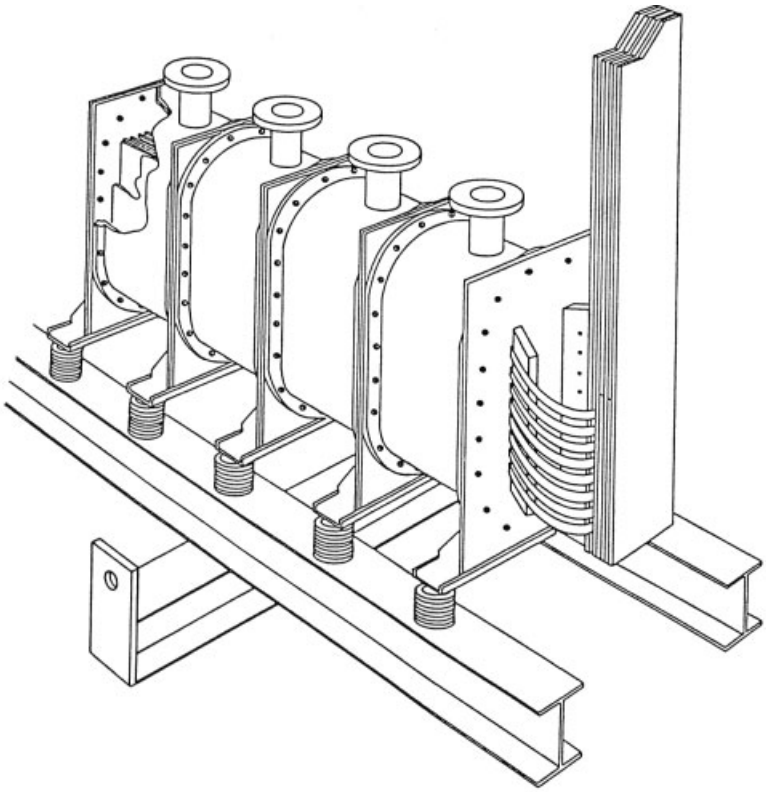

Figure 18. Directly connected electrochemical reactors (multipolar arrangement, Chemetics)

rial, they are called *dimensionally stable anodes* (DSA) (→ Chlorine) [202–204]. In the chlorate field, dimensionally stable anodes have been selected for 80 % of the total capacity installed since 1978.

Sintered titanium anodes (STA) have been developed by Sigri. In these electrodes, an intermediate layer of sintered titanium suboxide is applied between the titanium substrate and the coating, which gives improved mechanical stability, excellent adhesion to sintered titanium structures, and increased active surface area [205], [206]. Imperial Chemical Industries (ICI) has used these anodes in large-scale production since 1971. Other novel anodes have been developed by Imperial Metal Industries (IMI) and Marston Excelsior (Marex Electrode).

Titanium anodes normally consist of 1.5 – 4-mm sheets or of expanded mesh, activated by platinum group metal oxides. In troublefree operation, the durability of the coating based on ruthenium dioxide [207] is between 6 and 10 years. The literature on coated titanium anodes for chlorate electrosynthesis is reviewed in [193].

Cathodes. Mild steel has been the industrial standard for many years. It has a low hydrogen overvoltage and is cheap and stable under the operating conditions. However, steel corrodes when the cells are shut down, and it absorbs hydrogen, which tends to make the metal brittle and to form fissures in the welding areas and blisters [208]. The cathodes are usually 3 – 8 mm thick. The sheets are often perforated with holes of 3 – 6 mm diameter or they are slotted, which facilitates the release of hydrogen bubbles from the interelectrode gap. Mild steel cathodes tend to collect carbonate deposits, which increase the potential of the cathodes. This requires periodic acid washing unless the process water is deionized to contain less than 0.05 mg/kg of impurities.

The application of steel cathodes in bipolar cells is complicated because dissimilar metals must be joined. However, no connection problems arise by explosion bonding of titanium and steel [209].

Attempts to coat the steel cathodes or to replace them by other metals have not brought satisfactory results. Titanium has not been generally accepted because of its tendency to form titanium hydride, TiH_2, which gradually makes the cathode break off and limits the lifetime to 2 years. Attempts to lower the overvoltage by coating the cathodes with nickel, molybdenum, or their oxides [210] have not yet been successful because these materials catalyze the decomposition of hypochlorite during shutdown. Cathodes made from titanium alloys, e.g., Ti – 0.2 % Pd , remain stable for 2 years at a temperature of 95 °C, whereas the durability is much longer at 75 °C.

7.3.3. Operational Parameters

Temperature. The operation of electrochemical reactors benefits from temperature increase in two ways: (1) the cell potential decreases by lowering the resistivity of the electrolyte, and (2) autoxidation of hypochlorite to chlorate is more strongly favored than anodic oxidation. Consequently, the current efficiency increases with temperature. On the other hand, the temperature should not be near the boiling temperature of the electrolyte because then the rate of chlorine and water desorption increases. The introduction of dimensionally stable metal anodes has permitted raising the operating temperature from 40 – 45 °C (previously with graphite anodes) to 80 – 90 °C.

Pressure. Increasing the electrolyte pressure lowers not only the rate of chlorine desorption, but also the volume of the gas bubbles released from the electrolyte; thus, it acts favorably on the interelectrode resistivity. However, serious engineering problems arise when the pressure is increased. Industrial electrochemical reactors are usually operated at a gage pressure of less than 0.1 MPa (1 bar).

Interelectrode Distance. The electric interelectrode resistance strongly depends on the quantity of gas bubbles dispersed in the electrolyte [211]. When the electrolyte flow rate is increased and the electrode distance decreased, the resistance decreases. However, a very small electrode distance leads to a large frictional pressure drop in the reactor. Therefore, the minimum spacing of industrial reactors is ca. 3 mm, which roughly agrees with the results of optimizations [212]; distances of 3 – 5 mm are generally used between metal electrodes.

pH Value. According to Equation (7) (see Autoxidation, page 24), autoxidation of hypochlorite requires the simultaneous presence of hypochlorous acid and hypochlorite. The concentration of both species is very dependent on pH. A carefully controlled pH value, therefore, is a prerequisite of an industrially acceptable current efficiency. The optimum pH depends on the temperature and is maintained in the range of 6.1 – 6.4 in modern industrial electrochemical reactors operated at 80 – 90 °C.

Chloride Concentration. The brine feed to the cells is always close to being saturated to obtain a large final chlorate concentration of the product solution. The chloride concentration should not fall below 80 – 100 g/L because water decomposition is favored at low chloride concentration [213]. If necessary, the brine is resaturated with salt, but solubility limits must be observed (Figs. 11 and 12).

Current Density. High current density increases the acidity near the anode, thus depressing anodic hypochlorite oxidation. In addition, cathodic hypochlorite reduction is hampered. However, current density of industrial cells is limited by the concurrent increase in cell potential. Current densities of industrial cells are optimized on the basis of these considerations and thus permit operation at higher current efficiencies.

Current Efficiency. If x is the fraction of the hypochlorous acid (or chlorine) involved in the anodic oxidation and $(1-x)$ is the fraction that produces chlorate along the autoxidation reaction, the current efficiency ε is (cf. Fig. 13)

$$\varepsilon = \frac{1}{1 + 0.5x} \tag{9}$$

Exclusive anodic chlorate formation would result in a current efficiency of $\varepsilon = 66.7$ %, provided the contribution of loss reactions is negligible. The short survey of the operational parameters given in the previous sections highlights the complex system that governs current efficiency in chlorate electrosynthesis. The problem has been dealt with in detail in [39] and is the object of methodical investigations [215], [216]. Mathematical models have been established to provide means to theoretically predict current efficiency. A model assuming ideal stirred-tank-reactor conditions for both the electrochemical and the chemical reactor [214] was compared with industrial results [180]. A more recent model combines a plug flow electrolyzer with either a stirred tank or a plug flow chemical reactor [217]. The results show satisfactory agreement with data from two industrial systems [218].

The current efficiency of operating industrial cells with external recirculation of the electrolyte (Fig. 14) can be assessed by measuring the difference Δc of total hypochlorite concentration (mol/L) at the cell outlet and inlet. Because the difference of total hypochlorite concentration represents the amount of hypochlorite converted to chlorate by autoxidation outside the cell, the current efficiency is calculated to be

$$\varepsilon = \frac{2}{3}\left[1 + \frac{\dot{V}_L \cdot \Delta c \cdot F}{I}\right] \tag{10}$$

where

F = Faraday constant (96 487 A s/mol)
$\dot{V}_L$ = volume flow rate, m^3/s
I = total current, A

Equation (10) does not take account of the autoxidation in the interelectrode gap, and it neglects all loss reactions; both factors tend to balance.

In all electrosynthesis systems, including those without external electrolyte recirculation, the current efficiency may be estimated by analyzing the gas mixture leaving the electrochemical reactor. The amount of anodic chlorate formation (together with the catalytic decomposition) can be assessed from the volume fraction of oxygen (φ_{O_2}). The loss caused by chlorine desorption is represented by the volume fraction of chlorine (φ_{Cl_2}). Moreover, the ratio of the true hydrogen flow rate N_{H_2} to the flow rate N'_{H_2} as calculated from Faraday's law indicates the loss through cathodic reduction of hypochlorite and chlorate [219]:

$$\varepsilon = \frac{1 - 3\varphi_{O_2} - 2\varphi_{Cl_2}}{1 - \varphi_{O_2} - \varphi_{Cl_2}} \cdot \frac{N_{H_2}}{N'_{H_2}} \tag{11}$$

In industrial operation, the current efficiency is predominantly lowered by anodic chlorate formation; when neglecting any other loss reaction, one obtains a very simple approximation equation:

$$\varepsilon \approx 1 - 2\varphi_{O_2} \tag{12}$$

The error of Equation (12) as compared to Equation (11) is within the accuracy of gas analysis in the current efficiency range of modern cells. Automatic measurement of the oxygen concentration in the cell gas, therefore, serves as a simple and reliable means to control cell operation.

7.3.4. Brine Purification

Solar salt from seawater, rock salt, or very pure "vacuum salt" are used as raw material. The brine must not contain large amounts of magnesium and calcium, which could form deposits on the cathodes. Sodium carbonate and sodium hydroxide are added to the brine to raise its pH to at least 10. The sulfate concentration must be lower than 10 – 20 g/L. The brine must be free of heavy metals that favor the decomposition of hypochlorite (see Section 7.2.2). Maximum impurity concentrations are listed below:

SO_4^{2-}	10 g/L
SiO_2	10 mg/kg
Al, $P_2O_7^{4-}$, Ca	5 mg/kg
Fe, Mn, Sn	1 – 5 mg/kg
Cu, Mg, Pb, Ti	1 mg/kg
Cr, Mo	0.5 mg/kg
Ir, Co	0.2 mg/kg
Ni	0.05 mg/kg

7.4. Crystallization

For the crystallization of chlorate, a solution of 560 – 630 g/L of sodium chlorate and 90 – 120 g/L of sodium chloride is produced at 70 – 90 °C. The temperature at the cell outlets is

first raised to 85 – 95 °C to convert the remaining 1.5 g/L of hypochlorite into chlorate. Urea, ammonia, hydrogen peroxide, or sodium formate is added to complete the hypochlorite decomposition. The solution is then rendered slightly alkaline with NaOH; this reduces corrosion in the crystallizer. The $NaClO_3$ solution is subsequently fed to a vacuum crystallizer, based on flash cooling with subsequent crystallization.

Crystalline chlorate is separated from sodium chloride solution by cooling or evaporation [220]. Crystallization conditions depend on the concentration of NaCl and $NaClO_3$, on solution temperature, and on mixing intensity. The cooling rate has little influence on the crystal size; ca. 90 % of the crystals have 120 – 260 µm diameter. The solution temperature, initially raised to 90 – 95 °C, drops to 48 °C in the separator and to 35 °C in the vacuum crystallizer at 2.9 kPa. The concentrated slurry contains 15 – 20 wt % of crystals; it is passed through a hydrocyclone and then a pusher centrifuge. Energy consumption, efficiency and economic optimization of the overall process were studied [221]; an energy flow diagram is given in [222].

Potassium chlorate is generally made from potassium chloride and sodium chlorate:

$$NaClO_3 + KCl \longrightarrow KClO_3 + NaCl$$

Solid KCl is added to the sodium chlorate cell liquor in stoichiometric amounts. The mixture is then transferred to a crystallizer and the potassium chlorate slurry is removed as described in the preceding paragraph. The mother liquor is recycled to the cells, where the salt is converted to chlorate and the process is repeated. Quality requirements for solid KCl are high, because no purification is possible after it is added to the solution.

7.5. Construction Materials

Chlorate solutions are kept slightly alkaline except during electrolysis, where a specific acidity must be maintained.

Mild steel and cast iron are suitable for equipment if the liquor is alkaline and free of active chlorine. In acidic media, poly(vinyl chloride) (PVC) tubes and tanks made from fiber-reinforced PVC are used up to 50 °C. Poly(vinyl chloride) stabilized with Ba or Cd (PVC-C) is much better suited, but the consumption rate is nevertheless 1 mm per year at 60 °C [223].

Titanium, polytetrafluoroethylene (PTFE), glass, and poly(vinylidene fluoride) (PVDF) have excellent resistance at all temperatures and under all conditions. Evaporators are made from yellow brass, rubber-lined steel, monel, stainless steel (AISI: 316; DIN: 1.4401) with cathodic protection, or titanium. Stainless steel should be used for the dryer because it is imperative to avoid Fe_2O_3 contamination of the chlorate, which would act as an explosion catalyst. Pump seals must be of the noncombustible type. Titanium is increasingly used in all equipment as the material for anodes, heat exchangers, reaction vessels, pumps, and tubes.

7.6. Environmental Protection

The gas released from the electrochemical reactor always contains some chlorine in addition to hydrogen and oxygen. The mixture passes a cooler, where the major amount of chlorine condenses. Residual amounts of chlorine (0.3 %) are lowered to $(1-3)\times10^{-4}$ % by absorption in alkaline solution (15 % NaOH) and to $(1-10)\times10^{-7}$ % by subsequent adsorption on activated carbon. The resulting gas mixture of hydrogen and some oxygen can either be vented to the atmosphere or used as a fuel. The oxygen concentration (1.5 – 2.5 %) depends on the reaction conditions of the cell (see Section 7.2). In trouble-free operation, environmental pollution problems do not arise. Pure hydrogen can be produced by passing the residual gas mixture over a noble metal catalyzing combustion and subsequently purified by selective permeation in palladium at elevated temperature (350 °C).

Wastewater from cleaning tanks and building floors requires appropriate treatment to avoid damage of the water flora by chlorates acting as herbicides. A process to remove almost 100 % of dichromate from chlorate solutions by use of a fixed bed of standard ion-exchange resin has been operated for several years in a pilot plant and is about to be commercially exploited [224].

Table 11. Quality specifications for chlorates of sodium and potassium

	Sodium chlorate			
	"White crystals", wt %	Crystalline, wt %	Powder, wt %	Solution at 20 °C, g/L
$NaClO_3$	96.5 – 97.5	99.75 – 99.95	99.4 – 99.75	500
NaCl, max.	0.05	0.03	0.06	0.16
Na_2SO_4, max.	0.04	traces	0.04	0.11
CaO, max.	nil	nil	nil	0.11
$NaBrO_3$, max.	0.02	traces	0.02	0.004
Fe, max.	traces	nil	traces	traces
$Na_2Cr_2O_7$, max.	nil	nil	nil	0.004
Insolubles, max.	0.02	0.02	0.02	0.02
Humidity, max.	2.00	0.02	0.02	–
Bulk density, g/cm^3				
not vibrated	1.15	1.31	1.49	1.317
vibrated	1.55	1.58	1.65	

	Potassium chlorate	
	Typical specification, wt %	Aragonesas S.A., for matches, wt %
$KClO_3$, min.	99.7	99.8
KCl, max.	0.032	0.05
NaCl, max.	0.2	
$KClO_4$, max.	0.1	
K_2CrO_4	nil	
$KBrO_3$, max.	0.07	0.04
Fe, max.	0.01	nil
Insolubles, max.	0.01	0.02
Humidity, max.	0.05	0.02

7.7. Quality Specifications

The purity requirements of chlorate depend on the intended purpose and are usually negotiated between producer and user. Various products differ considerably in quality. Standard values for white and yellowish sodium chlorate and potassium chlorate are shown in Table 11. The given potassium chlorate quality is satifactory for many purposes, but may be recrystallized for special requirements (Table 11).

7.8. Storage, Transportation, and Safety

Chlorates should be stored in a cool, dry, fireproof building. Preferably, a separate storage building should be provided with cement floors and metal catwalks.

Wooden construction must be avoided because of the combustibility of chlorate-impregnated wood. Crystalline chlorate may be stored in concrete, lined steel bins, or glazed tile silos. Dried air should be supplied to prevent caking from atmospheric moisture. Drums and packages containing chlorates should not be stored where such noncompatible chemicals as acids, solvents, oils, organic substances, sulfur, or powdered metals could be spilled [225]. Ventilation must be provided for operations where fine chlorate dust arises. The ventilation system should discharge to a water scrubber and be designed for easy cleaning. Pumps handling chlorate solutions should be of the "packless" type. Electrical supply and distribution points are to be inspected periodically for dust. Motors and switches must not be housed in the plant or store and must conform to the National Electrical Code (United States). Metal or plastic pallets are recommended for handling containers.

Safety showers and water supply should be available to workers. The containers (barrels) must be kept closed and not be stored on top of each other. Chlorates are handled in polyethylene bags, in metal drums, or metal-lined fiberboard or plywood drums (25 – 170 kg). Bulk chlorates can be delivered by tank trucks or railroad tank cars either in the dry or wet state. They are usually unloaded as a slurry by recirculation of hot water (60 °C) between the car and the dissolver tank. The recommended material for storage tank construction is stainless steel (AISI

316). Tanks lined with rubber or plastics are not recommended because of possible fire or explosion hazards. For more details, see references [226–237].

7.9. Uses

Sodium Chlorate. The rapid growth of sodium chlorate production is mainly due to the widespread use of the chlorate-derived chlorine dioxide bleach by the pulp and paper industry. The once predominant use of sodium chlorate as a nonselective herbicide has strongly declined. Second in importance is its use as an intermediate in the production of other chlorates, mainly potassium chlorate, and of sodium perchlorate for conversion to ammonium perchlorate, which is used as an oxidizer in solid propellants. Sodium chlorate is further used as oxidizing agent in uranium refining and other metallurgical operations, as an additive to agricultural products and dyes, in textile and fur dyeing, metal etching, and in chemical laboratories and throughout the chemical industry as an oxidizing agent.

Potassium Chlorate. Potassium chlorate is used mainly in the manufacture of matches and in the pyrotechnics, explosives, cosmetics, and pharmaceutical industry; it is superior to sodium chlorate because of its smaller hygroscopicity.

Barium Chlorate. Barium chlorate is prepared by reaction of barium chloride with sodium chlorate solution. It is precipitated by cooling, purified by recrystallization, and used in pyrotechnics.

Calcium Chlorate. Solutions of calcium chlorate are used as herbicide.

7.10. Economic Aspects

Because of the steadily increasing demand for chlorine dioxide in the pulp industry, production of sodium chlorate has doubled in the past 10 years. Figure 19 shows annual production data. Canada is the world's leading producer with 0.5 Mt $NaClO_3$ in 1984, closely followed by the United States with 0.41 Mt $NaClO_3$.

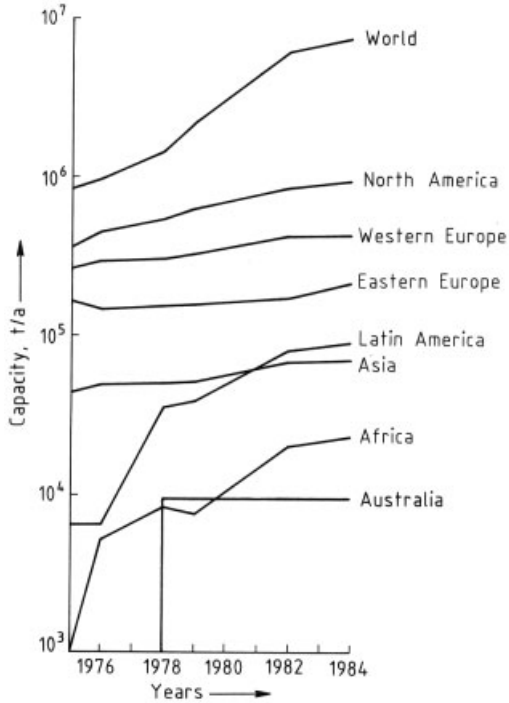

Figure 19. Chlorate production capacities

Since electric power cost amounts to ca. one-third of the total production cost of sodium chlorate, cheap energy is the key economic factor.

8. Perchloric Acid and Perchlorates

Perchloric acid [*7601-90-3*] and its salts, particularly ammonium perchlorate [*7790-98-9*], NH_4ClO_4, sodium perchlorate [*7601-89-0*], $NaClO_4$, and potassium perchlorate [*7778-74-7*], $KClO_4$, find many applications because of their strong oxidizing power; their chemical stability is sufficient to permit high-energy oxidation under controlled conditions. Perchloric acid is used on a limited scale mainly as a reagent for analytical purposes; its production has increased because of its use as a starting material for pure ammonium perchlorate, a basic ingredient of explosives and solid propellants for rockets and missiles. Minor amounts of potassium perchlorate occur in Chile, in natural deposits of sodium nitrate.

Table 12. Density, ϱ, and vapor pressure, p, of anhydrous perchloric acid

	t, °C			
	0	10	20	25
ϱ, g/cm^3	1.808	1.789	1.770	1.761
p, kPa	1.546	2.506	3.913	4.826

8.1. Physical and Chemical Properties

8.1.1. Perchloric Acid

Anhydrous perchloric acid, M_r 100.5, can be obtained by distilling at reduced pressure a mixture containing 1 part of 20 % $HClO_4$ and 4 parts of 20 % oleum. It is a colorless, strongly hygroscopic liquid, melting at − 102 °C. At atmospheric pressure, it decomposes at 75 °C. Its density and vapor pressure vs. temperature are given in Table 12.

Perchloric acid is miscible with water in all ratios and forms a series of hydrates as shown in Table 13. Boiling points and densities vs. concentration are given in Table 14. For other physical properties, see [239], [240].

Anhydrous perchloric acid is a very strong oxidizing agent. Some metals like nickel, copper, silver, and gold are only slightly oxidized at ambient temperature. Platinum is not attacked, but it decomposes the acid by catalytic action. In contact with combustible materials, the perchloric acid reacts violently, forming explosive mixtures with paper, charcoal, ethanol, acetic anhydride, and gelatin.

Perchloric acid is more stable in aqueous solution than in anhydrous form. The azeotropic mixture (72.5 wt % $HClO_4$) decomposes in the absence of oxidizable matter only above the boiling point. Concentrated perchloric acid is a strong oxidizing agent, especially at high temperature; it forms explosive mixtures with organic compounds, which can detonate on heating, percussion, or exposure to sparks or a flame.

Perchloric acid is a strong acid and reacts in aqueous solution with metals, metal oxides, and hydroxides, as well as with salts of volatile acids, forming the corresponding perchlorates. Some oxides, e.g., CuO, catalyze the decomposition of perchloric acid. This occurs through a chain of reactions that give chlorine, oxygen, and water as ultimate products.

8.1.2. Perchlorates

At least one element in each group of the periodic table, except the noble gases, forms perchlorates; this includes not only metals, but also nonmetallic elements, such as nitrogen in hydrazine perchlorate [*13762-80-6*] or fluorine in fluorine perchlorate [*37366-48-6*]. Organic ammonium, diazonium, and sulfonium perchlorates form another large class. Perchlorates are generally colorless and well (or fairly) soluble in water and such organic solvents as alcohols, ketones, or esters. Table 15 lists the solubilities of some metal perchlorates. Other physical properties can be found in [227], [241], and [242].

When heated, alkali metal and alkaline-earth metal perchlorates decompose before reaching the melting point, with the exception of $LiClO_4$ (*mp* 247 °C). The hydrates of the perchlorates generally have definite melting points, but they decompose on further heating, after liberation of water.

The outstanding chemical property of the perchlorates is their strong oxidizing power. This is exploited either in mixtures of perchlorates with combustible materials or by using organic perchlorates as explosives or propellants. On the other hand, all perchlorates, also in the pure state, are liable to undergo explosive decomposition if they are heated above a critical temperature, with liberation of chlorides and oxygen; ammonium perchlorate has been investigated extensively in this respect in view of its importance for jet propulsion [243–248].

8.2. Production

8.2.1. Perchloric Acid

Perchloric acid is made from sodium perchlorate and hydrogen chloride [239], [249]. A saturated $NaClO_4$ solution reacts with an excess of HCl,

Table 13. Perchloric acid and its hydrates, $HClO_4 \cdot n\,H_2O$

	n					
	0	1	2	2.5	3	3.5
CAS registry no.	[*7601-90-3*]	[*60477-26-1*]	[*13445-00-6*]	[*34099-94-0*]	[*35468-32-7*]	[*41371-23-1*]
M_r	100.46	118.47	136.49	145.5	154.51	163.5
c_{HClO_4}, wt %	100	84.8	73.6	69.1	65.0	61.5
mp, °C	− 112	+50	− 17.5	− 29.8	− 37 (α) − 43.2 (β)	− 41.4

Table 14. Boiling points and densities of aqueous perchloric acid solutions

	c_{HClO_4}, wt %							
	24.34	38.9	50.67	56.65	61.2	65.20	70.06	72.5
bp, °C	105.8	114.8	132.4	148.0	162.3	189.2	198.7	203
ϱ^{25}, g/cm^3	1.154	1.280	1.4058	1.4799	1.5413	1.5993	1.6748	1.7150

Table 15. Solubility of some perchlorates in water at 25 °C (grams in 100 g of water)

NH_4ClO_4	24.922	$Sr(ClO_4)_2$	309.67
$LiClO_4$	59.71	$Ba(ClO_4)_2$	198.33
$NaClO_4$	209.6	$Cu(ClO_4)_2 \cdot 2\,H_2O$	259
$KClO_4$	2.062	$AgClO_4$	540
$RbClO_4$	1.338	$Cd(ClO_4)_2 \cdot 6\,H_2O$	478
$CsClO_4$	2.000	$Mn(ClO_4)_2 \cdot 6\,H_2O$	268
$Mg(ClO_4)_2$	99.601	$Co(ClO_4)_2 \cdot 6\,H_2O$	292
$Ca(ClO_4)_2$	188.60	$Ni(ClO_4)_2 \cdot 6\,H_2O$	267

and NaCl precipitates. The dilute per-chloric acid produced after filtration (32 % $HClO_4$) is concentrated in three stages up to 70 – 71 % $HClO_4$; the last stage is vacuum distillation.

A continuous, electrochemical process for the production of perchloric acid has been developed in the Federal Republic of Germany. It involves the anodic oxidation of gaseous chlorine dissolved in chilled 40 % $HClO_4$ [250], [251]:

$$Cl_2 + 8\,H_2O - 14\,e \longrightarrow 2\,HClO_4 + 14\,H^+$$

The electrolyzer is of the filter press type; it is composed of poly(vinyl chloride) (PVC) frames and the electrodes are separated by a diaphragm of PVC fabric. The anodes are made from platinum foil, spot-welded to tantalum rods. The cathodes are horizontally slitted silver plates. A 40 % $HClO_4$ solution is circulated from the electrolyzer to an external cooler, which keeps the temperature within − 5 to + 3 °C. The operating conditions are as follows: current 5000 A; current density 2.5 – 5 kA/m^2; voltage 4.4 V; current efficiency 0.60; and chlorine concentration at inlet 3 g/L. The outflowing solution is distilled to remove residual chlorine and byproduct hydrogen chloride and to obtain high-purity perchloric acid. Platinum consumption is 0.025 g/t of 70 % $HClO_4$; it dissolves at the anode, but partly redeposits at the cathode and can thus be recovered. Per ton of 70 % $HClO_4$, 9600 kWh of electricity (d.c.) is used. The high purity of the product, directly obtained by a continuous process not requiring separation stages, is an advantageous tradeoff versus the relatively low current efficiency. The process further allows unusual perchlorates to be prepared by direct conversion with perchloric acid, thus avoiding the route via sodium chlorate.

8.2.2. Perchlorates

Various methods for the production of perchlorates are described in [239]. In practice, commercial production is based on the following steps:

1) Electrochemical production of sodium chlorate from sodium chloride (Chap. 7)
2) Electrochemical oxidation of sodium chlorate in aqueous solution to sodium perchlorate
3) Conversion of sodium perchlorate into another perchlorate, e.g.:

$$NaClO_4 + KCl \longrightarrow KClO_4 + NaCl$$

$$NaClO_4 + NH_4Cl \longrightarrow NH_4ClO_4 + NaCl$$

In the last step, the corresponding sulfate, which causes fewer corrosion problems, can be used instead of the chloride.

This indirect method for the production of perchlorates other than $NaClO_4$ is advantageous over their direct electrosynthesis from the corresponding chlorates, because of their relatively small solubility in water (Table 15). Attempts have been made to make sodium perchlorate directly from sodium chloride, by a single electrolysis [252–254]. The advantage, however, seems debatable, because the resulting current efficiency is only slightly better than 50 % under optimum conditions.

Anodic Oxidation. The anodic oxidation of chlorate occurs according to the following overall reaction:

$$ClO_3^- + H_2O - 2e \longrightarrow ClO_4^- + 2H^+$$

The standard potential of the anodic process (1.19 V) is very close to that of water oxidation (1.228 V), which gives rise to oxygen evolution. Both reactions compete strongly, independent of the acidity of the solution. To obtain a satisfactory current efficiency, the anode potential should be as large as possible, because a high polarization enhances the chlorate oxidation rate more than the oxygen evolution [255]. This is achieved by selecting a suitable anode material and a high current density. Several reaction mechanisms have been proposed for chlorate oxidation to perchlorate [39].

Cell design and operating conditions must be selected to optimize electricity consumption, which depends on cell voltage and current efficiency. Both are affected by a number of variables in conflicting ways. In particular, the current efficiency is improved by increasing the current density, but this also increases the cell voltage. A temperature rise operates in reverse in that not only the voltage, but also the current efficiency is diminished. A high sodium chlorate concentration in the effluent liquor improves current efficiency as well as voltage, but requires more expensive procedures and higher chlorate consumption to obtain a sufficiently pure perchlorate solution.

Industrial Cells. Contrary to electrochemical chlorate production, the residence time of the electrolyte in the perchlorate cell is not a critical parameter. The cell is undivided and hydrogen gas evolution provides sufficient agitation to decrease the sodium chlorate concentration gradient between the cell inlet and outlet. Because of the relatively high heat dissipation, the cell is normally provided with internal cooling.

Smooth platinum anodes yield the highest oxygen overpotential and, hence, the highest current efficiency (95 – 97 %). Platinum is either used in foils or cladded onto rods of some other metal, such as tantalum or copper. Lead dioxide anodes are also used [256]; they consist of a lead dioxide deposit plated on some conductive substrate, such as graphite rods or plates [257–260]. They are less expensive than platinum, which more than offsets their lower current efficiency (85 %). When a small amount of sodium fluoride (2 g/L) is added to the electrolyte, the current efficiency increases [257]. More recently, platinate-coated titanium has been claimed to substitute profitably for massive smooth platinum in perchlorate manufacture [261].

Cathodes are made from carbon steel, chromium – nickel steel, nickel, or bronze. As in chlorate electrosynthesis, sodium dichromate is added to the solution to minimize the current efficiency loss caused by cathodic reduction of perchlorate. At the same time, dichromate inhibits corrosion of the steel parts that are directly exposed to the electrolyte, in addition to their cathodic protection. However, dichromate cannot be used with lead dioxide anodes, because it catalyzes oxygen evolution [255]. In that case, the most suitable cathode material is nickel or chromium – nickel steel.

Figure 20 shows some typical cell models [262]. In the *Bitterfeld* cell, each anode, made from platinum foil, works between a pair of cathodes made from perforated steel plates. The cell is kept at 35 °C by cooling water running through a steel pipe bundle.

In the *Cardox* cell, the anodes consist of platinum cladded onto copper rods 1.3 cm in diameter. Each anode is surrounded by a cathodic tube 7.6 cm in diameter, which is made from steel and perforated at both ends. This allows hydrogen gas to escape into the interelectrode space and ensures convective circulation.

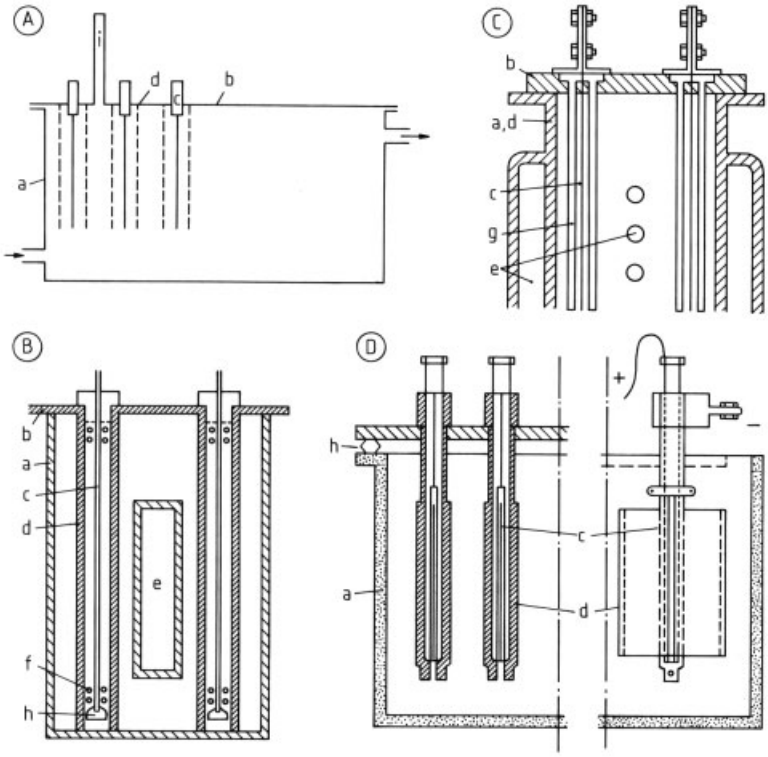

Figure 20. Some typical models of perchlorate cells
A) I. G. Farbenindustrie, Bitterfeld; B) Cardox Corp.; C) American Potash and Chemical Corp.; D) Pechiney, Chedde
a) Cell tank; b) Cover; c) Anode; d) Cathode; e) Cooling system; f) Circulation holes; g) Glass rods; h) Porcelain insulators; i) Gas outlet

Table 16. Typical operational data for perchlorate cells

Current	500 – 12 000 A
Current density	1500 – 5000 A/m^2
Cell potential	5 – 6.5 V
Current efficiency	
Platinum anodes	90 – 97 %
Lead dioxide anodes	85 %
Electric energy (d.c.)	
per ton of $NaClO_4$	2500 – 3000 kWh
Electrolyte temperature	35 – 50 °C
pH	6 – 10
Concentrations, g/L	
$Na_2Cr_2O_7$	0 – 5
Cell inlet	
$NaClO_3$	400 – 700
$NaClO_4$	0 – 100
Cell outlet	
$NaClO_3$	3 – 50
$NaClO_4$	800 – 1000
Platinum consumption per ton of $NaClO_4$	2 – 7 g

In the model of *American Potash and Chemicals* (formerly Western Electrochemical Co.), a set of anodic platinum foils is arranged around the inner wall of the cylindrical cell body performing as the cathode; the interelectrode gap is stabilized by a number of glass rod spacers. The temperature is kept at 40 – 45 °C by internal cooling and by a water jacket surrounding the cylindrical cell tank.

In the *Pechiney* model, the anodes are platinum foils and the cathodes are plates made from bronze.

Operation. Table 16 shows typical operational data [39], [239], [262]. Sodium perchlorate can be separated from the cell effluent either as hydrate or as anhydrous salt. Depending on concentration, separation is carried out by cooling alone or by evaporation, followed by cooling. The monohydrate precipitates from the solution below ca. 52 °C; above this temperature, the perchlorate crystallizes as the anhydrous salt. In either case, the mother liquor still contains a large amount of perchlorate. After enrichment

with sodium chlorate, this solution is recycled to the electrolysis process.

When the sodium perchlorate is used for conversion to other perchlorates, it need not be separated from the cell effluent. Sodium chlorate is destroyed by chemical treatment and dichromate precipitated as insoluble chromic hydroxide. The purified and concentrated solution is then ready for conversion to other salts [263].

8.3. Environmental Protection

On account of the limited volume produced and their slight toxicity, perchlorates do not constitute an environmental pollution problem. Aquatic life (fish, leeches, and tadpoles) survives indefinitely even when the perchlorate concentration in water exceeds 500 mg/L [264]. However, because of the antithyroid effect of perchlorate, some chronic symptoms may appear at low levels.

Some bacteria, such as *Vibrio dechloraticans,* can metabolize perchlorates; this is exploited for perchlorate destruction in the sanitation of contaminated sewage waters.

8.4. Chemical Analysis

Perchlorate is analyzed by its decomposition to chloride. This occurs by melting the perchlorate salt in a platinum dish with sodium carbonate, which accelerates the thermal decomposition. Sodium carbonate can be replaced with ammonium chloride if alkali-metal perchlorates are analyzed that decompose rapidly at 450 – 550 °C. The cooled melt is then dissolved in water or dilute nitric acid, and the total chloride is determined by normal volumetric or gravimetric methods. The chlorate or chloride content initially present in the sample must be subtracted to obtain the amount of perchlorate in the sample.

Quantitative precipitation of perchlorate from an aqueous solution by one of the following reagents is also possible: methylene blue [265]; potassium, rubidium, or cesium salts in cold ethanol – water solution [266]; or tetraphenylphosphonium chloride [267], [268]. The most sensitive quantitative determination of perchlorate ($10^{-6} - 10^{-7}$) is possible by means of ion-specific electrodes [269], [270].

8.5. Storage, Transportation, and Safety

The same storage and shipping regulations that apply to chlorates (Section 7.8) are also applicable to perchlorates.

According to the U.S. Department of Transportation (DOT) regulations for hazardous materials, perchloric acid and perchlorates are shipped in glass containers or metal drums as specified in [271]. They are classified as oxidizing substances and require oxidizer shipping labels. Transportation is forbidden on passenger-carrying aircraft or railcars. Shipment of perchloric acid in concentrations greater than 72 % is forbidden. Ammonium perchlorate may be shipped in steel drums with plastic liners; large quantities are transported in portable aluminum containers holding up to 2.27 t. Lower-side or hopper-type product discharge openings are not permitted.

Sodium or magnesium perchlorate may be shipped wet in tank cars with a minimum of 10 % water, which must be equally distributed. All perchlorates not specifically covered by DOT regulations should carry special precautionary labels indicating the specific fire or explosion hazards expected from the individual perchlorate.

Despite their limited toxicity, perchloric acid and perchlorates must be considered, like chlorates, as hazardous chemicals during fabrication, transportation, and handling. Because of its tendency toward spontaneous explosion, anhydrous perchloric acid should be prepared only in very small batches and in the absolute absence of impurities. It can be stored for a limited time at low temperature and must be protected from any kind of contamination. Distillation must be carried out under vacuum (2.4 kPa at 16 °C) and with protective shielding. One must never try to obtain the acid in anhydrous form by treatment with a drying agent. It should be stored in glass containers with glass stoppers, possibly embedded in kieselguhr or glass wool for protection.

Ammonium perchlorate and alkali-metal and alkaline-earth metal perchlorates are relatively insensitive to rubbing, heating, and shock. They also require special precautions in manufacturing and further handling [225], [226]. Clean work clothing must be worn each day and laundered afterward; it must not be taken home.

Clothing wet with perchlorate solution should be changed before drying. No smoking can be permitted in perchlorate working areas or while wearing work clothes. Deluge type safety showers or jump tanks should be provided. Only rubber shoes and rubber or rubberized gloves are permitted.

Perchlorates should not be stored close to flammable materials, reducing agents, or other hazardous substances. Buildings in working areas should be fireproof. Dust control or dust prevention in perchlorate solutions is particularly important. Since mixtures of such solutions with oil or grease are violently explosive, motor and pump bearings must be provided with special lubricating and washing devices [225]. Fires must be extinguished with water, but carbon dioxide may provide sufficient cooling to extinguish small fires. Dry powder is ineffective because it cannot smother a self-sustaining fire. For burns, cold water treatment should be used as quickly as possible. For other first aid measures, consult [226].

A second class of more hazardous compounds is formed by inorganic perchlorates containing nitrogen or heavy metals, organic perchlorates, or mixtures of inorganic perchlorates with organic substances, finely divided metals, or sulfur. They are all very sensitive to rubbing, shock, percussion, sparks, and heating and must, therefore, be handled with the same precautions as high explosives (→ Explosives).

8.6. Uses

Perchloric Acid. The commercial product is the azeotropic, aqueous solution of 72.5 % $HClO_4$ (*bp* 203 °C). In analytical chemistry, it serves to determine the metallic elements present in oxidizable substances, such as organic compounds [272–277]. Perchloric acid is an acetylation catalyst for cellulose and glucose and is used in the preparation of cellulose fibers.

Perchlorates. The most outstanding property of ammonium perchlorate is its high oxygen concentration (54.5 % O_2) and the fact that it decomposes without leaving a solid residue. Therefore, it is used as an oxidizing component in solid rocket propellants.

Lithium perchlorate is used in lithium–nickel sulfide dry batteries [278], and sodium perchlorate in electrochemical machining [279]; potassium perchlorate is a component of pyrotechnics and an ignition ingredient in flash bulbs [280]. Magnesium perchlorate is known as a very effective drying agent.

Ammonium perchlorate mixed with an epoxy resin forms a temporary adhesive between two metallic surfaces, such as two steel plates. These can be separated whenever desired by heating at ca. 300 °C because of the self-sustained combustion of the adhesive layer [281]. Ammonium perchlorate has been tried as a feed supplement to increase the weight of livestock, with an optimum dose of 2 – 5 mg/kg [282–286]. Perchlorates have also proven to be helpful in oxygen-regenerating systems to be used in enclosed environments, such as submarines and spacecraft, and in breathing equipment [287]. Potassium values from enriched bittern, obtained from the Dead Sea or the Great Salt Lake, can be recovered by precipitating potassium perchlorate with sodium perchlorate [288], [289].

8.7. Economic Aspects

Statistical data about the production and use of perchlorates are not easily accessible because of their strategic importance. The United States' production in 1974 was reported to be ca. 25 000 t, or ca. 12 % of the chlorate production [290]. On account of the increasing developments of artificial satellites and space shuttle programs, perchlorate production will increase, due to their major use in solid propellants.

9. Toxicology and Occupational Health

Hypochlorous acid and hypochlorites are toxic because they liberate chlorine on contact with acid. Sodium hypochlorite solution produced by on-site electrolyzers (available chlorine concentration 0.5 – 10 g/L) is regarded as corrosive and as an irritant when ingested or inhaled. It is also a mild skin irritant, and prolonged exposure may result in a burn or rash [56]. However, concentrated hypochlorous acid burns human skin in seconds.

Chlorine Dioxide. Chlorine dioxide gas is the most toxic and hazardous of all chlorine oxides. Even when small amounts of chlorine dioxide are inhaled, the respiratory system is severely damaged [291]. The symptoms of chlorine dioxide intoxication depend on its concentration and on the exposure time; they include lacrimation, headache, vomiting, severe cough, asthmatic bronchitis, dyspnea, and even death. Such defects as dyspnea or asthmatic bronchitis only heal slowly after the exposure to chlorine dioxide has ceased.

Exposure to 5 ppm of chlorine dioxide in air during several hours severely irritates the mucous membranes [292]; short-term exposure to 20 ppm may cause death. According to animal experiments and observations of employees in chlorine dioxide production plants, the human and animal organisms are extremely sensitive to chlorine dioxide, in particular when the concentrations are higher than 1 ppm [291–294].

Chlorine dioxide levels below 0.1 ppm are relatively harmless [295], [296]. Exposure of rats to 0.1 ppm of chlorine dioxide (5 h/d; 10 weeks) showed no toxic effects whereas repeated inhalation of 10 ppm of chlorine dioxide caused death.

The odor threshold of chlorine dioxide is less than 0.1 ppm [297], [298]. The MAK and the TLV of chlorine dioxide have been established at 0.1 ppm (0.3 mg/m^3).

Chlorine dioxide production plants must be operated at slightly subatmospheric pressure. The plant design must also account for the high explosiveness of chlorine dioxide; therefore, chlorine dioxide is diluted with air to below 15 % before it leaves the generator. No attempt should ever be made to transport chlorine dioxide over long distances in any form.

Chlorates. The major health hazard of the chlorates arises from their extreme danger of flammability on contact with oxidizable substances. Irritation of the skin, mucous membranes, and eyes may occur from prolonged exposure to dusty atmospheres. Contrary to chlorides, the chlorates are strong blood toxins. Chlorates are readily absorbed by the mucous membranes; doses of a few grams of chlorate are lethal for humans. Abdominal pain, nausea and vomiting, diarrhea, pallor, blueness, shortness of breath, and unconsciousness are the immediate symptoms when toxic amounts of chlorates are ingested. In workplaces with dust formation, dust masks must be worn. Cases of chronic toxicity have not been reported. The instructions for industrial safety must carefully be observed [226], [227], [229]. For detailed information on precautionary and first aid procedures, see [228], [233–237].

Perchlorates. Sodium perchlorate appears in the urine of humans within 10 min after ingestion and is largely eliminated within 5 h. During many years of large-scale production of alkali-metal perchlorates, no case of perchlorate intoxication has been reported [239]. Industrial experience also indicates that these salts are not particularly irritating to the skin or mucous membranes, although inhalation should be avoided. A sodium perchlorate dose of 2 – 4 g/kg is lethal to rabbits.

10. References

1. D. D. Wagman et al., *J. Phys. Chem. Ref. Data* **11** (1982) suppl. no. 2.
2. V. Engelhard: *Hypochlorite und elektrische Bleiche. Technisch-konstruktiver Teil,* Knapp, Halle (Saale) 1903.
3. E. Schmidt, E. Graumann, *Ber. Dtsch. Chem. Ges.* **54** (1921) 1861 – 1873.
4. H. Vogt, *J. Electrochem. Soc.* **128** (1981) 29C – 32C.
5. W. Wallace, *J. Electrochem. Soc.* **99** (1952) 309C –310C.
6. F. Oettel, *Z. Elektrotech. Elektrochem.* **1** (1894) 354 – 361; **1** (1895) 474 – 480.
7. H. Wohlwill, *Z. Elektrochem. Angew. Phys. Chem.* **5** (1898) 52 – 76.
8. F. Foerster, E. Müller, *Z. Elektrochem. Angew. Phys. Chem.* **8** (1902) 515 – 540.
9. F. Foerster, E. Müller, *Z. Elektrochem. Angew. Phys. Chem.* **8** (1902) 633 – 665, 665 – 672.
10. F. Foerster, E. Müller, F. Jorre, *Z. Elektrochem. Angew. Phys. Chem.* **6** (1899) 11 – 23.
11. F. Foerster, *Trans. Electrochem. Soc.* **46** (1924) 23 – 50.
12. J. B. C. Kershaw: *Die elektrolytische Chloratindustrie,* Knapp, Halle (Saale) 1905.
13. A. Schumann-Leclercq, FR 772 326, 1933.
14. F. Oettel, *Z. Elektrochem. Angew. Phys. Chem.* **7** (1901) 315 – 320.
15. H. B. Beer, *J. Electrochem. Soc.* **127** (1980) 303C –307C.

16. F. Foerster, *Z. Elektrochem. Angew. Phys. Chem.* **4** (1898) 386 – 388.
17. F. Haber, S. Grinberg, *Z. Anorg. Chem.* **16** (1898) 198 – 228, 329 – 361.
18. F. Winteler, *Chem.-Ztg.* **22** (1898) 89 – 90.
19. Shell-Development Co., US 2 347 151, 1944.
20. D. S. Davis, *Ind. Eng. Chem.* **34** (1942) 624.
21. *Gmelin,* System no. 6, Chlor, p. 228.
22. I. G. Rheinfelden, DE 656 413, 1931.
23. J. Ourisson, *Atti Congr. Int. Chim.* **4** (1938) 40 – 51.
24. C. W. Bunn, L. M. Clark, J. L. Clifford, *Proc. R. Soc., London, Ser. A* **151** (1935) 141 – 167.
25. ICI, DE 541 821, 1930.
26. Olin Mathieson, US 1 787 048, 1930.
27. Olin Mathieson, US 1 713 650, 1929.
28. Olin Mathieson, US 2 195 755, 1940; US 2 195 757, 1940.
29. Olin Corp., EP 0 086 914 A 1, 1982.
30. W. H. Sheltmire in J. S. Sconce (ed.): *Chlorine, Its Manufacture, Properties and Uses,* Reinhold, New York 1962.
31. G. L. Culp, R. L. Culp: *New Concepts in Water Purification,* Van Nostrand Reinhold, New York 1974, pp. 188 – 196.
32. Murgatroyd's Salt and Chemical Co., GB 984 378, 1962.
33. R. E. Stanton, US 3 222 269, 1962.
34. R. E. Stanton, *Manuf. Chem.* 1962, no. 10, 400 – 404.
35. B. M. Milwidsky, *Manuf. Chem.* 1962, no. 10, 400 – 404.
36. H. B. H. Cooper, *Chem. Process (Chicago)* **33** (1969) 28.
37. B. Alt, *Chem. Tech. (Heidelberg)* **4** (1975) no. 4.
38. M. I. Ismail, T. Z. Fahidy, H. Vogt, H. Wendt, *Fortschr. Verfahrenstech.,* **vol. 19,** VDI-Verlag, Düsseldorf 1981, pp. 381 – 414 (Engl.).
39. N. Ibl, H. Vogt in J. O'M. Bockris, B. E. Conway, E. Yeager, R. E. White (eds.): *Comprehensive Treatise of Electrochemistry,* **vol. 2,** Plenum Publishing, New York 1981, a) pp. 201 – 208; b) pp. 169 – 201; c) p. 210.
40. L. R. Czarnetzki, L. J. J. Janssen, *J. Appl. Electrochem.* **19** (1989) 630 – 636.
41. G. R. Heal, A. T. Kuhn, R. B. Lartey, *J. Electrochem. Soc.* **124** (1977) 1690 – 1697.
42. P. M. Robertson, W. Gnehm, L. Ponto, *J. Appl. Electrochem.* **13** (1983) 307 – 315.
43. L. Hammar, G. Wranglén, *Electrochim. Acta* **9** (1964) 1 – 16.
44. F. Hine, M. Yasuda, *J. Electrochem. Soc.* **118** (1971) 182 – 183.
45. A. T. Kuhn, H. B. H. Hamzah, *Chem. Ing. Tech.* **52** (1980) 762 – 763.
46. J. E. Bennett, *Chem. Eng. Prog.* **70** (1974) no. 12, 60 – 63.
47. P. M. Robertson, R. Oberlin, N. Ibl, *Electrochim. Acta* **26** (1981) 941 – 949.
48. V. I. Skripchenko, E. P. Drozdetskaya, K. G. Il'in, *Zh. Prikl. Khim. (Leningrad)* **49** (1976) 887 – 889.
49. J. E. Bennett, *Int. J. Hydrogen Energy* **5** (1980) 383 – 394.
50. Engelhard Minerals & Chemicals Corp., US 3 873 438, 1975 (E. P. Anderson, T. L. Lamb).
51. Oronzio de Nora Impianti Elettrochimici S.p.A., US 4 032 426, 1977 (O. de Nora, V. de Nora, P. M. Spaziante).
52. C.-Y. Chan, K. H. Khoo, T. K. Lim, *Surf. Technol.* **15** (1982) 383 – 394.
53. M. W. Lister, *Can. J. Chem.* **34** (1956) 479 – 488.
54. *Chem. Eng. (N.Y.)* **76** (1966), 9th May, 98.
55. A. T. Kuhn, R. B. Lartey, *Chem. Ing. Tech.* **47** (1975) 129 – 135.
56. G. D. Muir: *Hazards in the Chemical Laboratory,* The Chemical Society, London 1977, p. 388.
57. D. S. Davis, *Ind. Eng. Chem.* **34** (1942) 624.
58. F. Göbel, *Dissertation* TH Dresden (1961).
59. R. A. Crawford, B. I. Dewitt, *Tappi* **50** (1967) no. 7, 138.
60. T. F. Haller, W. W. Northgraves, *Tappi* **38** (1955) no. 4, 199.
61. J. Kepinski, J. Trzeszczynski, *Chemii,* **38**Ann. Soc. Chim. Pol. (1964) 201.
62. W. Finkelnburg, H. J. Schumacher, *Z. Phys. Chem. Bodenstein-Festband* (1931) 704.
63. H. J. Schumacher, G. Steiger, *Z. Phys. Chem. Abt. B* **7** (1930) 363.
64. A. S. C. Chen, R. A. Larson, V. L. Snoeyink, *Environ. Sci. Technol.* **16** (1982) no. 5, 268 – 273.
65. P. L. Gilmon, *Tappi* **51** (1968) no. 4, 62A – 66A.
66. R. Swindells, G. Cowley, R. A. F. Latham, *Pulp Pap. Can.* **81** (1980) no. 4, 49 – 52, 54, 56, 58.
67. W. H. Rapson, *Tappi* **37** (1954) no. 4, 129 – 137.
68. H. Karge, *Z. Anal. Chem.* **200** (1964) 57.
69. T. Ozawa, T. Kwan, *Chem. Pharm. Bull.* **31** (1983) no. 8, 2864 – 2867.
70. E. T. Eriksen, J. Lind, G. Merenyi, *J. Chem. Soc. Faraday Trans. 1* **77** (1981) no. 9, 2115 – 2123.
71. E. Asmus, H. Garschagen, *Z. Anal. Chem.* **138** (1953) 404.

72. W. H. Rapson, *Tappi* **39** (1956) no. 8, 554 – 556.
73. C. C. Hong, F. Lenzi, W. H. Rapson, *Can. J. Chem. Eng.* **45** (Dec. 1967) 349 – 355.
74. W. H. Rapson, *Can. J. Chem. Eng.* **36** (1958) 262.
75. Hooker Chem. Corp., US 2 936 219, 1961.
76. R. Hirschberg, *Papier (Darmstadt)* **17** (1963) 25.
77. E. Kesting, *Tappi* **36** (1953) no. 4, 166.
78. E. Kesting, DE 831 542, 1948.
79. Metallgesellschaft AG, EP 0 094 718 A1, 1983.
80. R. S. Saltzmann, *Adv. Instrum.* **30** (1975) Pt. 2, 609.
81. C. A. Lindholm, *Pap. Puu* **60** (1978) no. 5, 359 – 362, 365 – 366, 369 – 371.
82. "Production of Bleaching Chemicals – Chlorine Dioxide, Chlorine, Caustic Soda," *Chem. Econ. Eng. Rev.* **15** (1983) no. 10 (no. 172).
83. Lurgi, Express Information, Bleaching Chemicals, 1983.
84. Metallgesellschaft AG, EP 0 011 326 B1, 1979.
85. K. Lohrberg, G. Klamp, *Pulp Pap.* **56** (1982) no. 11, 114 – 116.
86. K. Lohrberg, *Pulp Pap.* **54** (1980) no. 10, 185 – 187.
87. T. D. Hughes, *Pulp Pap. Can.* **81** (1980) no. 4, 30 – 32, 34.
88. T. D. Hughes, *Chemetics, Integrated Chlorine Dioxide System,* International Pulp Bleaching Conference, June 13, 1979, Toronto, Canada.
89. P. Wintzer, *Chem.-Ing.-Techn.* **52** (1980) no. 5, 392 – 398.
90. "Clean Process for Making Chlorine Dioxide," *Environ. Sci. Technol.* **13** (1979) no. 7, 788 – 790.
91. Uhde: *Bleaching Chemicals, Chlorine Dioxide, Chlorine, Sodium Hypochlorite, Sodium Hydroxide,* Dortmund 1984.
92. W. H. Rapson, *Tappi* **41** (1958) no. 4, 181 – 185.
93. J. W. Robinson, *Tappi* **46** (1963) no. 2, 120 – 123.
94. P. Orlich, R. Simonette, *Pulp Pap.* **40** (1966) no. 7.
95. Hooker Chem. Corp., 1184-D.
96. F. M. Ernest, *Pap. Trade J.* **143** (1959) no. 34, 46 – 50.
97. E. S. Atkinson, R. Simonette, *Pulp Pap.* **42** (1968) no. 4.
98. E. S. Atkinson, R. Simonette, *Pulp Pap.* **42** (1968) no. 8.
99. H. D. Partridge, W. H. Rapson, *Tappi* **44** (1961) no. 10, 698.
100. ERCO, Chlorine Dioxide Processes, R2, R3, R3H, R5, R6, R7, R8, 1983.
101. E. S. Atkinson, *Chem. Eng. (N.Y.)* **81** (1974) no. 3, 36 – 37.
102. H. de V. Partridge, E. S. Atkinson, *Pulp. Pap. Mag. Can.* **73** (1972) no. 8, 203 – 206.
103. H. D. Partridge, E. S. Atkinson, A. C. Schulz, *Tappi* **54** (1971) no. 9, 1484 – 1487.
104. J. F. Serafin, H. C. Scribner, *Pulp Pap. Mag. Can.* **62** (1961) 473 – 476.
105. D. W. Reeve, W. H. Rapson, *Pulp Pap. Mag. Can.* **74** (1973) no. 1, 19 – 27.
106. R. K. Bradley, G. N. Werezak, *Tappi* **57** (1974) no. 4, 98 – 101.
107. E. S. Atkinson, *Tappi* **61** (1978) no. 8, 61 – 64.
108. R. A. F. Latham, *Pulp Pap.* **53** (1979) no. 11, 74 – 77.
109. D. Lobley, J. Howard, *Pulp Pap. Can.* **80** (1979) no. 2, 48 – 51.
110. M. C. Fredette, *Proc. Tech. Assoc., Pulp Pap. Ind., (Int. Pulp Bleaching Conf.)* 1982, 77 – 86.
111. M. C. Fredette, *Pulp Pap. Can.* **85** (1984) no. 1, 52 – 56.
112. D. D. Koch, L. L. Edwards, *Tappi* **55** (1972) no. 5, 752 – 756.
113. J. Schuber, W. A. Kraske, *Chem. Eng. N.Y.* **60** (1953) no. 9, 205.
114. G. L. Cunningham, US 2 089 913, 1937.
115. T. G. Holst, *Sven. Papperstidn* **47** (1944) 537 – 546.
116. T. G. Holst, US 2 373 830, 1945.
117. S. H. Persson, US 2 376 935, 1945.
118. US 2 481 241, 1949,(W. H. Rapson, M. Waymann).
119. US 2 598 087, 1952, (M. Waymann, W. H. Rapson,).
120. E. R. Woodward, G. P. Vincent, *Trans. Am. Inst. Chem. Eng.* **40** (1944) 271.
121. W. J. Masschelein, *J. Am. Water Works Assoc.* **76** (1984) no. 1, 70 – 76.
122. J. Valenta, W. Gahler, O. Mohr, *BBR, Brunnenbau, Bau Wasserwerken, Rohrleitungsbau* **32** (1981) no. 3, 106 – 109.
123. R. W. Jordan, J. A. Kosinski, R. J. Baker, *Water Sewage Works* **127** (1980) no. 10, 44 – 45, 48.
124. W. H. Rapson, K. A. Hakim, *Pulp Pap. Mag. Can.* **151** (July 1957) 151.
125. W. H. Rapson, C. B. Anderson, G. F. King, *Tappi* **41** (1958) 442.
126. M. Gummerus, C. A. Lindholm, N. E. Virkola, *Pap. Puu* **59** (1977) no. 5, 335 – 338, 341 – 344, 347 – 348, 353 – 355.
127. F. Kocevar, H. Talovic, *Melliand Textilber. Int.* **54** (1973) no. 10, 1064 – 1066.

128. C. Rav-Acha, R. Blits, E. Choshen, A. Serri, B. Limoni, *J. Environ. Sci. Health, Part A* **A18** (1983) no. 5, 651 – 671.
129. M. C. Taylor, J. E. White, G. P. Vincent, G. L. Cunningham, *Ind. Eng. Chem.* **32** (1940) 890.
130. G. Cunningham, T. S. Oey, *J. Amer. Chem. Soc.* **77** (1955) 4498.
131. G. Cunningham, T. S. Oey, *J. Amer. Chem. Soc.* **77** (1955) no. 3, 799.
132. V. Popjankow, N. Kolarow, *Z. Anorg. Allg. Chem.* **346** (1966) 322.
133. I. Nakamori, Y. Nagino, K. Hideshima, T. Hirai, *Kogyo Kagaku Zasshi* **61** (1958) 147 – 149; *Chem. Abstr.* **53,** 21322 a.
134. T. S. Oey, G. Cunningham, J. Koopman, *Chem. Eng. Data* **5** (1960) no. 3, 248.
135. H. Korte, W. Kaufmann, *Melliand Textilber.* **23** (1942) 234 – 239.
136. M. C. Tayler et al., *Ind. Eng. Chem.* **32** (1940) 899 – 903.
137. J. F. White, M. C. Tayler, G. P. Vincent, *Ind. Eng. Chem.* **34** (1942) 782 – 792.
138. L. F. Ynterna, T. Fleming, *Ind. Eng. Chem. Anal. Ed.* **11** (1939) 375 – 377.
139. I. F. Haller, S. Listek, *Analytic.* **20** (1948) 639 – 642.
140. I. F. White, *Am. Dyest. Rep.* **31** (1942) 484 – 487.
141. W. C. Gardiner, E. H. Karr, *Office Tech. Serv. Rept.,* **PB-33** (1946) 218.
142. I. G. Farbenind., DE 744 369, 1944 (F. Erbe).
143. Mathieson Alkali Works, US 2 092 944, Sept. 1937.
144. Potasse et Produits Chimiques, FR 1 019 027, 1953.
145. A. R. Reychler, *Bull. Soc. Chim. Fr.* **25** (1901) 663.
146. G. R. Levi, *Atti V Congr. Nazl. Chim. Pura Applicata, Rome 1935*, 1936, part I, 382 – 386; *Chem. Abstr.* **31,** 5290[5].
147. Solvay + Cie., CH 258 842, 1946.
148. Nederld. Zoutindustrie, DE 963 505, 1954.
149. Farbwerke Hoechst, DE 1 059 891, 1958.
150. Mathieson Alkali Works, DE 668 140, 1935.
151. Farbw. Hoechst, DE 906 447, 1942.
152. E. Elöd, A. Klein, *Melliand Textilber.* **27** (1947) 191, 263.
153. H. Baier, *Melliand Textilber.* **32** (1951) no. 2, 141 – 146.
154. W. Hundt, K. Vieweg, *Text.-Prax.* **6** (1951) 439.
155. S. K. Kapoor, Y. V. Sood, T. K. Roy, A. K. Kohli, R. Pant, *IPPTA,* **20** (1983) no. 1, 27 – 32.
156. H. Vogt, *Chem. Age India* **26** (1975) 540 – 544.
157. N. V. S. Knibbs, H. Palfreeman, *Trans. Faraday Soc.* **16** (1920) 402.
158. R. A. Crawford, W. B. Darlington, L. B. Kliever, *J. Electrochem. Soc.* **117** (1970) 279 – 282.
159. A. Nallet, R. A. Pâris, *Bull. Soc. Chim. Fr.* 1956, 488 – 494.
160. P. Wintzer, *Chem.-Ing.-Tech.* **52** (1980) 392 – 398.
161. F. Foerster, *J. Prakt. Chem.* **63** (1901) 141 – 166.
162. F. Foerster, E. Müller, *Z. Elektrochem.* **9** (1903) 171 – 185.
163. I. E. Flis, M. K. Bynyaeva, *Zh. Prikl. Khim. (Leningrad)* **30** (1957) 339 – 345; *J. Appl. Chem. USSR (Engl. Transl.)* **30** (1957) 359 – 365.
164. I. Taniguchi, T. Sekine, *Denki Kagaku* **43** (1975) 715 – 720.
165. B. V. Tilak, K. Viswanathan, C. G. Rader, *J. Electrochem. Soc.* **128** (1981) 1228 – 1232.
166. D. Landolt, N. Ibl, *Electrochim. Acta* **15** (1970) 1165 – 1183.
167. N. V. S. Knibbs, *Trans. Faraday Soc.* **16** (1920) 424 – 433.
168. V. A. Shlyapnikov, *Zh. Prikl. Khim. (Leningrad)* **42** (1969) 2182 – 2188, *J. Appl. Chem. USSR (Engl. Transl.)* **42** (1969) 2051 – 2056.
169. I. Taniguchi, T. Sekine, *Denki Kagaku* **43** (1975) 632 – 637.
170. C. Wagner, *J. Electrochem. Soc.* **101** (1954) 181 –185.
171. V. A. Shlyapnikov, T. S. Filippov, *Elektrokhim.* **5** (1969) 866 – 868; *Sov. Electrochem. (Engl. Transl.)* **5** (1969) 806 – 807.
172. F. Hine, M. Yasuda, *J. Electrochem. Soc.* **118** (1971) 182 – 183.
173. J. S. Booth, H. Hamzah, A. T. Kuhn, *Electrochim. Acta* **25** (1980) 1347 – 1350.
174. I. E. Veselovskaya, E. M. Kuchinski, L. V. Morochko, *Zh. Prikl. Khim.* **37** (1964) 76 – 83; *J. Appl. Chem. USSR (Engl. Transl.)* **37** (1964) 85 – 91.
175. I. Taniguchi, T. Sekine, *Denki Kagaku* **43** (1975) 532 – 534.
176. V. de Valera, *Trans. Faraday Soc.* **49** (1953) 1338 – 1351.
177. M. W. Lister, R. C. Petterson, *Can. J. Chem.* **40** (1962) 729 – 733.
178. J. M. González Barredo, *An. Fis Quim.* **37** (1941) 123 – 157.

179. J. R. Newberry, W. C. Gardiner, A. J. Holmes, R. F. Fogle, *J. Electrochem. Soc.* **116** (1969) 114 – 118.
180. T. R. Beck, R. Brännland, *J. Electrochem. Soc.* **119** (1972) 320 – 325.
181. J. Horacek, S. Puschaver, *Chem. Eng. Prog.* **67** (1971) no. 3, 71 – 74.
182. T. Matsumura, R. Itai, M. Shibuya, G. Ishi, *Electrochem. Technol.* **6** (1968) 402 – 404.
183. T. Nagai, T. Takei, *J. Electrochem. Soc. Japan* **25** (1957) no. 2, 11.
184. J. Billiter: *Die technische Elektrolyse der Nichtmetalle,* Springer-Verlag, Wien 1954.
185. M. M. Jakšić, *J. Appl. Electrochem.* **3** (1973) 219 – 225.
186. M. M. Jakšić, A. R. Despić, B. Z. Nikolić, *Elektrokhim.* **8** (1972) 1573 – 1584 (Russ.); *Sov. Electrochem. (Engl. Transl.)* **8** (1972) 1533 – 1542.
187. D. G. Elliott, *Tappi* **51** (1968) no. 4, 88A – 90A.
188. M. M. Jakšić, B. Z. Nikolić, D. M. Karanović, C. R. Milovanović, *J. Electrochem. Soc.* **116** (1969) 394 –398.
189. *Chem. Eng. (N.Y.)* **72** (19 July 1965) 82 – 83.
190. K. C. Narasimham, H. V. K. Udupa, *J. Electrochem. Soc.* **123** (1976) 1294 – 1298.
191. *Ullmann,* 3rd ed., **5,** 539.
192. M. Y. Fioshin, *Uspekhi v oblasti elektrosinteza neorganiceskikh soedinenii. (Progress of electrosyntheses of inorganic compounds)* Isdatelstwo "Khimia," Moscow 1974.
193. W. Schreiter, H. Wendt, H. Vogt, *Fortschr. Verfahrenstechnik,* **vol. 17,** VDI-Verlag, Düsseldorf 1979, pp. 379 – 412.
194. J. E. Colman, *AIChE Symp. Ser.* **77** (1981) no. 204, 244 – 263.
195. K. Viswanathan, B. V. Tilak, *J. Electrochem. Soc.* **131** (1984) 1551 – 1559.
196. C. Oloman: *Electrochemical Processing for the Pulp and Paper Industry,* The Electrochemical Consultancy, Romsey, Hants, 1996.
197. I. Roušar, V. Cezner, *J. Electrochem. Soc.* **121** (1974) 648 – 651.
198. H. B. Beer, DE 1 571 721, 1966.
199. H. B. Beer, *J. Electrochem. Soc.* **127** (1980) 303C – 305C.
200. H. B. Beer, DE 1 671 422, 1968.
201. H. B. Beer, CH 492 480, 1966.
202. S. Trasatti, W. E. O'Grady, *Adv. Electrochem. Electrochem. Eng.* **12** (1981) 177 – 261.
203. H. B. Beer, *Chem. Ind. (London)* 1978, 491 – 496.
204. O. De Nora, *Chem.-Ing.-Tech.* **42** (1970) 222 – 226.
205. G. Bewer, H. Debrodt, H. Herbst, *J. Met.* **34** (1982) 37 – 41.
206. F. Brandmeir, H. Böder, G. Bewer, H. Herbst, *Chem.-Ing.-Tech.* **52** (1980) 443 – 444.
207. L. M. Elina, V. M. Gitneva, V. I. Brystow, N. M. Shmygul', *Elektrokhim.* **10** (1974) 68 – 70 (Russ.); *Sov. Electrochem. (Engl. Transl.)* **10** (1974) 59 – 61.
208. *Ullmann,* 4th ed., **22,** 41, 69.
209. Heraeus, DE 3 239 535, 1982.
210. N. V. Krstajić, M. D. Spasojević, R. T. Atanasoski, *J. Appl. Electrochem.* **14** (1984) 131 – 134.
211. H. Vogt, in *Fortschritte der Verfahrenstechnik,* **vol. 20,** VDI-Verlag, Düsseldorf 1982, pp. 369 – 404.
212. W. Thiele, M. Schleiff, *Chem. Tech. (Leipzig)* **31** (1979) 623 – 624.
213. F. Hine, M. Yasuda, T. Noda, T. Yoshida, J. Okura, *J. Electrochem. Soc.* **126** (1979) 1439 – 1445.
214. T. R. Beck, *J. Electrochem. Soc.* **116** (1969) 1038 – 1041.
215. M. M. Jakšić, B. Z. Nikolić, M. D. Spasojević, *Chem. Tech. (Leipzig)* **27** (1975) 158 – 162, 534 – 538.
216. M. M. Jakšić, *Electrochim. Acta* **21** (1976) 1127 – 1136.
217. H. Vogt, *J. Appl. Electrochem.* **22** (1994) 1185 – 1191.
218. H. Vogt, *Electrochim. Acta* **39** (1994) 2173 – 2179.
219. M. M. Jakšić, A. R. Despić, I. M. Csonka, B. Z. Nikolić, *J. Electrochem. Soc.* **116** (1969) 1316 – 1322.
220. Pennwalt Corp., US 3 883 406, 1973.
221. H.-P. Neumann, J. Kardos, *Chem. Tech.* **39** (1987) 205 – 207.
222. T. R. Beck, R. T. Ruggeri, *Adv. Electrochem. Electrochem. Eng.* **12** (1981) 263 – 354.
223. J. Fleck, *Chem.-Ing.-Tech.* **43** (1971) 173 – 177.
224. J. R. Hodges, *J. Electrochem. Soc.* **131** (1984) 96C.
225. C. M. Olson, *J. Electrochem. Soc.* **116** (1969) 33C –35C.
226. *Safety Data Sheet SD-42,* Manufacturing Chemists Ass., Washington D.C., 1952.
227. *Data Sheet 371, Chlorates,* National Safety Council, Chicago, Ill., 1977.
228. *Merkblatt für Arbeiten mit Kalium-Natriumchlorat,* Verlag Chemie, Weinheim 1965.

229. I.C.C. Regulations for Transportation of Explosives and other Dangerous Articles by Rail Freight, Rail Express, Rail Baggage and Motor Carrier.
230. RID/ADR International Rules for the Transportation of Dangerous Goods by Rail and Road, Office Central des Transports par chemins de fer, Bern 1985.
231. *Accident Prevention Manual for Industrial Operations,* National Safety Council, 7th ed., Chicago, Ill., 1978.
232. *Unfallverhütungsvorschriften der Berufsgenossenschaft der chemischen Industrie,* Carl Heymanns Verlag, Köln, published annually.
233. *Merkblätter Gefährliche Arbeitsstoffe* (R. Kühn, K. Birett eds.), 4th ed., Verlag Moderne Industrie, München 1979, p. N04.
234. *Hygienic Guide Series, Sodium Chlorate,* American Industrial Hygiene Association, Detroit, Mich., United States.
235. L. Roth, M. Daunderer: *Toxikologische Enzyklopädie,* **l. 1 – 3,** voVerlag Moderne Industrie, München 1976.
236. G. Hommel: *Handbuch der gefährlichen Güter,* **vol. 1,** no. 142, Springer-Verlag, Berlin 1980.
237. L. Roth, U. Weller: *Gefährliche chemische Reaktionen,* ecomed, Landshut 1982, pp. III-N, 1 – 4.
238. W. C. Gardiner, *J. Electrochem. Soc.* **125** (1978) 22C – 29C.
239. J. C. Schumacher (ed.): *Perchlorates – their Properties, Manufacture and Uses,* Am. Chem. Soc. Mon. no. 146, Reinhold Publ. Corp., New York 1960.
240. *Gmelin,* System no. 6, Chlor, Verlag Chemie, Berlin 1927, pp. 362 – 391.
241. H. V. Casson, G. J. Crane, G. E. Styan, *Pulp Pap. Mag. Can.* **69** (1968) 39.
242. G. K. Loseva, *Tr. Novocherkassk, Politekh. Inst.* **266** (1972) 78 – 81.
243. F. Shadman-Yazdi, E. E. Petersen, US Nat. Tech. Inform. Serv., AD Rep. 1972, no. 746 728; *Gov. Rep. Announc.* **72** (1972) 172.
244. A. V. Nikolaev, G. G. Tsurinov, L. A. Khripin, *Izv. Sib. Otd. Akad. Nauk SSSR, Ser. Khim. Nauk.* 1972, 98 – 99.
245. P. Jacobs, W. McCarthy, H. M. Whitehead, *Chem. Rev.* **69** (1969) 551.
246. P. Barret, *Cah. Therm.* **4** (1924) 13.
247. C. N. R. Rao, B. Prakash, *Natl. Stand. Ref. Data Ser. (U.S. Natl. Bur. Stand.)* **56** (1975) 28.
248. F. Solymosi, *Acta Phys. Chim.* **19** (1973) no. 1 – 2, 67.
249. Oldbury Electrochem. Co., US 2 392 861, 1945 (J. C. Pernert).
250. E. Merck AG, DE 1 031 288, 1956.
251. W. Müller, P. Jönck, *Chem. Ing. Tech.* **35** (1963) 78 – 80.
252. Y. Kato, K. Sugino, K. Koizumi, S. Kitahara, *Electrotech. J.* **5** (1941) 45 – 48.
253. M. Nagalingam, R. P. Govinda, R. C. J. Narasimham, S. Sampath, H. V. K. Udupa, *Chem. Ing. Tech.* **41** (1969) 1301 – 1308.
254. H. V. K. Udupa, K. C. Narasimham, M. Nagalingam, N. Thiagarajan et al., *J. Appl. Electrochem.* **1** (1971) 207 – 212.
255. O. de Nora, P. Gallone, C. Traini, G. Meneghini, *J. Electrochem. Soc.* **116** (1969) 146 – 151.
256. J. C. Grigger in C. A. Hampel (ed.): *The Encyclopedia of Electrochemistry,* Reinhold Pub. Corp., New York 1964, pp. 762 – 764.
257. T. Osuga, S. Fujii, K. Sugino, T. Sekine, *J. Electrochem. Soc.* **116** (1969) 203 – 207.
258. Pacific Eng. and Prod. Co., US 2 945 791, 1960.
259. J. C. Schumacher, D. R. Stern, P. R. Graham, *J. Electrochem. Soc.* **105** (1958) 151 – 155.
260. J. C. Schumacher, *J. Electrochem. Soc.* **129** (1982) 397C – 403C.
261. K. R. Koziol, K. H. Sieberer, H.-C. Rathjen, J. B. Zenk, E. F. Wenk, *Chem. Ing. Tech.* **49** (1977) 288 – 293.
262. A. Legendre, *Chem. Ing. Tech.* **34** (1962) 379 – 386.
263. T. W. Clapper in C. A. Hampel (ed.), ref. [252], pp. 886 – 890.
264. D. Burrows, J. C. Dacre, U.S. NTIS, AD-A, Rep. no. 010 660, Army Med. Bioeng. Res. Dev. Lab. Fort Detrick, Md., 1975.
265. F. D. Snell, C. T. Snell: *Colorimetric Methods of Analysis,* **vol. 2,** Van Nostrand Co., Princeton, N.J., 1957, pp. 718 – 719.
266. H. H. Willard, H. Diehl: *Advanced Quantitative Analysis,* Van Nostrand Co., Princeton, N.J., 1944, pp. 254 – 257.
267. G. M. Smith, *Ind. Eng. Chem. Anal. Ed.* **11** (1939) 186.
268. G. M. Smith, *Ind. Eng. Chem. Anal. Ed.* **11** (1939) 269.
269. R. J. Bacquk, R. J. Dubois, *Anal. Chem.* **40** (1968) 685.
270. M. J. Smith, S. E. Manahan, *Anal. Chim. Acta* **48** (1969) 315.

271. R. M. Graziano: *Hazardous Materials Regulations,* Tarif No. 31, Dept. of Transportation, Washington DC, 1977.
272. G. F. Smith: *Mixed Perchloric, Sulfuric and Phosphoric Acids and their Applications in Analysis,* G. F. Smith Chem. Co., Columbus, Ohio, 1942.
273. G. F. Smith, *Anal. Chim. Acta* **8** (1953) 397.
274. G. F. Smith, *Anal. Chim. Acta* **17** (1957) 175.
275. E. Kahane: *L'Action de l'Acide Perchlorique sur les Matières Organiques,* Herman et Cie, Paris 1934.
276. H. Diehl, G. F. Smith, *Talanta* **2** (1959) 209.
277. T. T. Gorsuch: *The Destruction of Organic Matter,* Pergamon Press, Oxford 1970.
278. L. Gaines, R. Jasinski, US Natl. Tech. Inform. Serv., AD Rep. no. 749 861 (1972).
279. J. P. Hoare, K. W. Mao, A. J. Wallace, *Corrosion (Houston)* **27** (1971) 211 – 215.
280. General Electric Company, US 3 724 991, 1971.
281. Nissan Motor Co., US 3 993 524, 1976 (Y. Okada, S. Kensho).
282. P. N. Razumovskii, *Khim Sel'sk Khoz.* 1976, no. 14, 71.
283. P. N. Razumovskii, G. S. Semanin, G. I. Balk, *Kompleksn. Ispol'z Biol. Akt. Veshchestv. Korml. S-kh Zhivotn. Mater. Vses. Soveshch.* **1** (1973) 370.
284. A. S. Solan, *Zhivotnovodstvo* 1974, no. 11, 63.
285. V. I. Mikhailov, B. R. Gotsulenko, V. P. Kardivari, *Zhivotnovodstvo* 1976, no. 5, 83.
286. H. G. Pena, *Poult. Sci.* 1976, 188.
287. Thiokol Corp., US 3 993 514 (E. J. Pacanovsky, E. A. Martino).
288. J. A. Epstein, *Hydrometallurgy* **1** (1975) 39.
289. D. R. George, J. M. Riley, J. R. Ross, paper presented at 62nd Nat. Meet. Inst. Chem. Eng., Salt Lake City, May 21 – 24, 1967.
290. W. C. Gardiner, *J. Electrochem. Soc.* **125** (1978) 22C – 29C.
291. H. Petry, *Arch. Gewerbepath. Gewerbehyg.* **13** (1964) 363.
292. H. B. Elkins: *The Chemistry of Industrial Toxicology,* 2nd ed., J. Wiley & Sons, New York 1958.
293. J. Gloemme et al., *AMA Arch. Ind. Health* **16** (1957) 169.
294. A. Graefe, *Dtsch. Med. Wochenschr.* **8** (1902) 191.
295. L. T. Fairhall: *Industrial Toxicology.* Williams & Wilkins Co., Baltimore 1957.
296. T. Dalhamm, *AMA Arch. Ind. Health* **15** (1957) 101.
297. "Hygienic Guide, Chlorine Dioxide," *Ind. Hyg. Ass. J.* **21** (1958) 381.
298. M. N. Gleason, R. E. Gosselin, H. C. Hodge: *Clinical Toxicology of Commercial Products,* Williams & Wilkins Co., Baltimore 1963.

Chlorites → Chlorine Oxides and Chlorine Oxygen Acids

Chloroacetaldehydes

Reinhard Jira, Wacker-Chemie GmbH, Burghausen, Germany
Erwin Kopp, Wacker-Chemie GmbH, München, Germany
Blaine C. McKusick, Wilmington, Delaware, USA
Gerhard Röderer, Wacker-Chemie GmbH, Burghausen, Germany
Axel Bosch, Wacker Chemie AG, Burghausen, Germany
Gerald Fleischmann, Wacker Chemie AG, Burghausen, Germany

1. Introduction 1
2. Monochloroacetaldehyde 2
2.1. Physical Properties 2
2.2. Chemical Properties and Uses 3
2.2.1. Reactions of the Aldehyde Group . . . 3
2.2.2. Reactions at the Chlorine Atom 3
2.2.3. Reactions of the Chlorine and Aldehyde Groups 3
2.2.4. Uses of Chloroacetaldehyde 4
2.3. Production 4
2.3.1. Chlorination of Acetaldehyde and Paraldehyde 4
2.3.2. Chlorination of Vinyl Compounds in Aqueous Media and Alcohols . . . 4
2.3.3. Preparation of Anhydrous Chloroacetaldehyde 4
2.4. Analysis 5
2.5. Transportation, Storage, and Safety 5
2.5.1. Transportation 5
2.5.2. Storage 5
2.5.3. Safety 5
2.6. Polymeric Chloroacetaldehydes . . . 6
3. Dichloroacetaldehyde 6
3.1. Physical Properties 6
3.2. Chemical Properties and Uses 6
3.3. Production 7
3.3.1. Chlorination of Acetaldehyde or Paraldehyde 7
3.3.2. Hypochlorination of 1,2-Dichloroethylene 7
3.3.3. Chlorination of Ethanol 7
3.4. Availability, Storage, and Transportation 7
3.5. Polymeric Dichloroacetaldehydes . 7
4. Trichloroacetaldehyde 7
4.1. Physical Properties 8
4.2. Chemical Properties 8
4.3. Production 8
4.4. Analysis 8
4.5. Legal Aspects 9
4.6. Uses and Occurrence 9
4.7. Economic Aspects 9
5. Toxicology 9
5.1. Chloroacetaldehyde 9
5.2. Dichloroacetaldehyde 10
5.3. Trichloroacetaldehyde (Chloral) . . 10
6. References 10

1. Introduction

Chloroacetaldehydes are described by the general formula $CH_{3-x}Cl_xCH{=}O$ (x = 1, 2, 3). Chloroacetaldehydes are colorless liquids which form solid hydrates with water. The stability of the hydrates increases with the increasing number of chlorine atoms due to changed reactivity of the aldehyde group. Whereas monochloroacetaldehyde undergoes typical aldehyde reactions, the aldehyde group of trichloroacetaldehyde forms addition compounds with nucleophilic reagents.

Chloroacetaldehydes became important in organic synthetic chemistry. Today mainly monochloroacetaldehyde is of commercial and industrial importance as intermediate in the synthesis of pharmaceuticals, agrochemicals, and fragrances.

Chloroacetaldehydes are used as pure substances, hydrates, or aqueous solutions. Monochloroacetaldehyde is mainly used as an aqueous solution because isolation of pure

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_527.pub2

monochloroacetaldehyde is expensive and an aqueous solution is suitable for most reactions.

2. Monochloroacetaldehyde

Monochloroacetaldehyde [*107-20-0*], 2-chloroethanal, CH_2ClCHO, M_r 78.50, was first prepared in pure form by K. Natterer in 1882 by heating chloroacetaldehyde diethyl acetal, $CH_2ClCH(OC_2H_5)_2$ [*621-62-5*], with anhydrous oxalic acid at 100–150 °C. Monochloroacetaldehyde was obtained from the aqueous solution as a crystalline hemihydrate.

2.1. Physical Properties

Anhydrous monochloroacetaldehyde is a colorless, mobile liquid with extremely pungent odor; *bp* 85–85.5 °C (99.7 kPa), enthalpy of formation of the liquid from the elements −256.4 kJ/mol, enthalpy of combustion −981.4 kJ/mol (at 101.3 kPa, 20 °C), dipole moment 1.99 D (at 298 K in benzene). It is soluble in water (to form a hydrate) and in common organic solvents. The octanol–water partition coefficient of the substance is very low (0.16).

Monochloroacetaldehyde hemihydrate 1,1′-dihydroxy-2,2′-dichlorodiethyl ether, 1,1′-oxybis-(2-chloroethanol), $CH_2ClCH(OH)-O-CH(OH)CH_2Cl$, M_r175.01, colorless crystals, *mp* 43–50 °C (elimination of water), *bp* (with decomposition into the components) 84 °C, is soluble in water (see Tables 1 and 2), ethanol, ether, benzene, methylene chloride, and chloroform.

Two CAS numbers, [*34789-09-8*] and [*7737-02-2*], are given for chloroacetaldehyde hemihydrate. The first is valid for the term "acetaldehyde, chloro dimer, hydrate" and the second for "1,1′-oxybis(2-chloroethanol)." These compounds are identical.

The hemihydrate distills from the aqueous solution as an azeotropic mixture. Both components, chloroacetaldehyde and water, are present in the vapor phase, but they cannot be separated by a simple distillation. Figure 1 shows the composition of the vapor and the liquid of a monochloroacetaldehyde–water mixture, and Figure 2 shows the boiling and condensation points as functions of the composition.

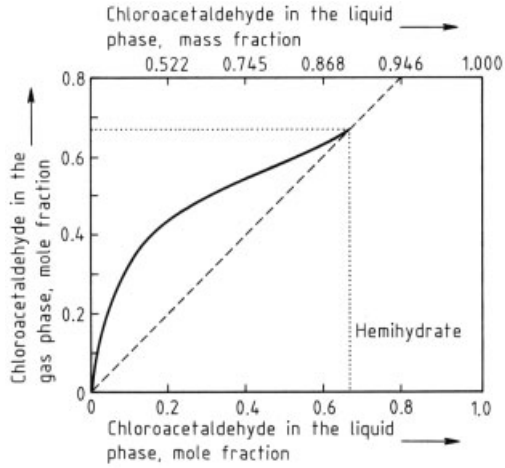

Figure 1. Vapor–liquid equilibrium curve for monochloroacetaldehyde–water [1]

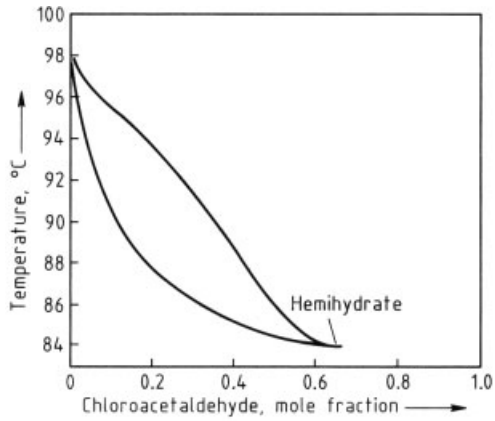

Figure 2. Boiling and condensation curve of aqueous chloroacetaldehyde solutions [1]

Safety Data (for the hemihydrate):

Flash point 60 °C
Ignition point 405 °C [2]
Explosive limits in air (at 20 °C, 101.3 kPa), lower 305 g/m^3
MAK (Germany) deferred; substantial suspicion of a carcinogenic potential, Cat. 3; H (skin absorptive) [3]
MAK (Germany): annulled [4]
TLV-STEL (USA): 1 ppm [5]

For chloroacetaldehyde NIOSH has defined an immediately dangerous to life or health concentration (IDLH) of 45 ppm based on acute inhalation toxicity data in humans [6].

Table 1. Solubility of monochloroacetaldehyde hemihydrate in water as a function of temperature [1]

Temperature, °C	1	10	20	30	40
Concentration, wt %	13.35	22.2	44.3	62.7	81.5

Table 2. Density of aqueous chloroacetaldehyde hemihydrate solutions at 15 °C as a function of concentration [1]

Concentration, wt %	10	20	30	40	50	60	72.5
Density, g/cm^3	1.041	1.085	1.137	1.188	1.238	1.290	1.355

2.2. Chemical Properties and Uses

Like other α-halocarbonyl compounds, chloroacetaldehyde is exceptionally reactive. The aldehyde group takes part in reactions typical of this group, e.g., addition and condensation. Other reactions involve the chloromethyl group. Nucleophilic substitution of the chlorine is facilitated by the inductive effect of the aldehyde group. The α-hydrogens atoms are likewise strongly activated by the inductive effects of the aldehyde group and the chlorine atom.

2.2.1. Reactions of the Aldehyde Group

Like acetaldehyde, anhydrous chloroacetaldehyde self-condenses easily at room temperature to form a cyclic trimer and tetramer as well as polymers with an acetal (polyoxymethylene) structure. Cotrimers and copolymers with acetaldehyde are also known. The hemihydrate has the structure 1,1′-dihydroxy-2,2′-dichlorodiethyl ether, $ClCH_2CH(OH)-O-CH(OH)CH_2Cl$, and reacts like an aldehyde.

Among industrially important derivatives are the crystalline sodium bisulfite addition compound, sodium 2-chloro-1-hydroxyethanesulfonate [*13064-50-1*], the dimethyl acetal [*97-97-2*] (*bp* 127–128 °C), and the diethyl acetal [*621-62-5*] (*bp* 157.4 °C), which are used in preparing pharmaceutical precursors and pesticides.

Oxidation with air in the presence of cobalt salts [7], hydrogen peroxide, nitric acid [8], or other oxidants yields monochloroacetic acid. Aldol addition in an alkaline medium gives 2,4-dichloroacetaldol, which can be dehydrated to 2,4-dichlorocrotonaldehyde in an acid medium. In coaldolization with acetaldehyde, chloroacetaldehyde reacts exclusively as the methylene component to produce 2-chlorocrotonaldehyde. This aldehyde is also obtained as a byproduct of acetaldehyde production (→ Acetaldehyde).

Other typical aldehyde reactions are the formation of the aldehyde ammonia, 1,3,5-tris-(chloromethyl)hexahydro-2,4,6-triazine [*7038-17-7*] [9], the cyanohydrin [*33965-80-9*], $ClCH_2CH(OH)CN$ [10, 11], and 2-chloroethyl chloroacetate, $ClCH_2COOCH_2CH_2Cl$ [*3848-12-2*], (Tishchenko reaction) [9].

2.2.2. Reactions at the Chlorine Atom

The chlorine atom can be removed by hydrogenation using a palladium–carbon catalyst in aqueous solution [12], and can be replaced by nucleophilic agents [13].

Mercaptoacetaldehyde is formed by the reaction of chloroacetaldehyde with sodium mercaptan. It is used as an intermediate in the synthesis of thienodiazepines, a new class of tranquilizers [14, 15], of thiophenic azo dyes, of *d,l*-cystein, and of certain 2-alkylthiazoline food flavors [16]. Mercaptoacetaldehyde exists as an equilibrium mixture of the dimer, 2,5-dihydroxy-1,4-dithian [*40018-26-6*] [17], and oligomers [18].

Substitution of the chlorine atom by the amino or methylamino groups is carried out on the acetal to protect the aldehyde group. The amino and methylamino derivatives are intermediates in the synthesis of pesticides and pharmaceuticals [19].

Reaction of chloroacetaldehyde dimethyl acetal with sodium methoxide yields 1,1,2-trimethoxyethane, which is also an intermediate in pharmaceutical syntheses, e.g., sulfamethoxydiazine.

2.2.3. Reactions of the Chlorine and Aldehyde Groups

Chloroacetaldehyde, usually in form of the hemihydrate or an acetal, is used as a starting compound for preparing numerous heterocyclic

compounds. The most industrially important reaction is the synthesis of 2-aminothiazole by reaction with thiourea.

2-Aminothiazole is an intermediate in the preparation of pharmaceuticals, e.g., sulfathiazole, and dyes. Cyclocondensation of chloroacetaldehyde with methyl acetoacetate gives 2-methyl-3-carbomethoxyfuran, an intermediate in making the seed disinfectant fenfuram [20]. 5-Chlorothiadiazole, an intermediate for the cotton defoliant and plant growth regulator thidiazuron, is obtained by condensation of chloroacetaldehyde with carboethoxyhydrazine and subsequent cyclization with thionyl chloride [21].

Stilbenes are produced by the reaction of chloroacetaldehyde with aromatic hydrocarbons, phenols, or phenol ethers [22]. Vinyl dialkyl phosphates are produced by reaction with trialkyl phosphites (Perkow reaction) [23]. Reaction with sodium acetate and acetic anhydride forms 1,1,2-triacetoxyethane, which can be used as a starting material for producing *d,l*-serine [24].

2.2.4. Uses of Chloroacetaldehyde

Chloroacetaldehyde finds application in the manufacture of pharmaceuticals, e.g., as starting material to synthesize statine intermediates. The mercapto derivative of chloroacetaldehyde, 2,5-dihydroxy-1,4-dithiane, is an important building block for drugs against AIDS [25]. Besides its application in the synthesis of pharmaceuticals, chloroacetaldehyde is used as a raw material to make dyestuffs. Dialkylacetal derivatives of chloroacetaldehyde such as chloroacetaldehyde dimethyl acetal are important starting materials for the manufacture of agrochemicals. The acetals are also used to make fragrances.

2.3. Production

2.3.1. Chlorination of Acetaldehyde and Paraldehyde

Chlorination of acetaldehyde or paraldehyde yields monochloroacetaldehyde mixed with small amounts of dichloroacetaldehyde and trichloroacetaldehyde. These three compounds are also obtained as byproducts of the production of acetaldehyde by the Wacker process (→ Acetaldehyde).

2.3.2. Chlorination of Vinyl Compounds in Aqueous Media and Alcohols

Simultaneous introduction of vinyl chloride and chlorine into water at ca. 20 °C forms chloroacetaldehyde in a nearly quantitative yield if its concentration in the reaction solution is not allowed to rise above ca. 5 wt %. At a higher concentration, increasing amounts of 1,1,2-trichloroethane are formed [26].

$$CH_2 = CHCl + aHOCl \rightarrow [ClCH_2CH(OH)Cl] \rightarrow ClCH_2CHO + HCl$$

This method yields only a dilute solution of chloroacetaldehyde. Concentrated aqueous solutions of chloroacetaldehyde are obtained in nearly quantitative yields by the reaction of chlorine with vinyl acetate in water at room temperature [27 – 29]. As an intermediate 1,2-dichloroethyl acetate is formed, which is quantitatively hydrolyzed at elevated temperatures.

$$CH_2 = CHOOCCH_3 + Cl_2 \longrightarrow ClCH_2CHClOOCCH_3 \xrightarrow{H_2O} ClCH_2CHO + HCl + CH_3COOH$$

The resulting solution is purified by distillation. Side products are only small amounts of higher chlorinated acetaldehydes, acetaldehyde, and condensation products.

Chloroacetaldehyde acetals are obtained in high yield and purity by carrying out the above reactions in alcohols [30, 31].

2.3.3. Preparation of Anhydrous Chloroacetaldehyde

Direct chlorination of dry acetaldehyde or paraldehyde gives only a low yield of monochloroacetaldehyde. Azeotropic dehydration of the hemihydrate with chloroform, toluene [32], or carbon tetrachloride [33] followed by distillation over anhydrous calcium chloride gives better yields. Anhydrous monochloroacetaldehyde is obtained in high yield by the depolymerization of trichloroparaldehyde or polychloroacetaldehyde at 145 °C in the presence of oxalic acid or

trichloroacetic acid [34]. Monochloroacetaldehyde is also obtained in high yield by pyrolysis of chloroethylene carbonate in the presence of quaternary ammonium salts [35, 36].

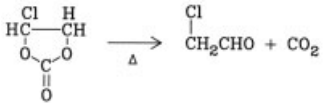

2.4. Analysis

Chloroacetaldehyde is commercially available as a 45 wt % aqueous solution. The monochloroacetaldehyde content can be determined by reaction with excess hydrazine sulfate to form chloride ion, which can be determined by the Volhard method. Due to side reactions and reaction with other aldehydes this method is not very specific to monochloroacetaldehyde. The monochloroacetaldehyde content of an aqueous solution can be determined by density and side products by GC [37].

The usual aldehyde reactions are suitable for identification purposes. Monochloroacetaldehyde forms derivatives with characteristic boiling or melting points: oxime, *bp* 61 °C at 2.7 kPa; semicarbazone, *mp* 146–148 °C; thiosemicarbazone, *mp* 141–142 °C; 2,4-dinitrophenylhydrazone, *mp* 154–156 °C; 2,4,6-trinitrophenylhydrazone, *mp* 159–160 °C (crystallized from glacial acetic acid).

Gas chromatography can separate and determine chloroacetaldehyde and its impurities [13]. Due to formation of hydrates this method is not specific to the content of chloroacetaldehyde.

2.5. Transportation, Storage, and Safety

2.5.1. Transportation

A 45 % aqueous chloroacetaldehyde solution can be shipped or stored in glass, enamel, enamelled steel, and phenolic-resin-coated steel containers. Stainless steel and polyethylene are other suitable materials for transporting the product and for storing it, but only for a short time.

Transportation of chloroacetaldehyde is regulated (UN 2232 class 6.1, I; IACO/IATA: forbidden).

2.5.2. Storage

Aqueous 45 % chloroacetaldehyde should be stored in an inert atmosphere, and it remains in good condition for at least 3 months if kept in the dark at room temperature. If it is exposed to sunlight or allowed to become warm, it discolors, and eventually brown, resinous products form with the liberation of hydrogen chloride. If the solution is kept at temperatures below 15 °C, the hemihydrate may crystallize out.

2.5.3. Safety

Chloroacetaldehyde solutions must not be heated. At elevated temperatures a strong exothermic condensation may occur. Tightly fitting goggles and gloves made of butyl or nitrile rubber or coated with neoprene (wear gloves for short periods only, < 10 min) should be worn when chloroacetaldehyde is handled because of its strongly irritating effect. The compound should not be allowed to come in contact with the skin, eyes, or clothes, and its vapors should not be inhaled. Adequate ventilation should be provided. Areas in which chloroacetaldehyde is being handled should be provided with a safety shower or comparable means of quickly washing the skin.

A gas mask fitted with filter ABEK should be worn if there is a risk of inhaling the vapor. In case of long or strong exposure a positive-pressure self-contained breathing apparatus is necessary. In the event of a spill, all potential sources of ignition must be kept far away. Use of personal protection equipment is necessary. The spill should be mopped up with an absorbent substance, such as sawdust, fine sand, or silica granules. Emission of the substance into surface waters, drains or sewers, and open soil must be prevented due to the toxic effects on aquatic organisms and inhibiting effects on bacteria. Disposal of chloroacetaldehyde can be effected by treating a dilute water solution with alkali at elevated temperature followed by biological treatment or incineration, always in compliance with local regulations.

Because of its high biocidal activity, chloroacetaldehyde interferes with the biological degradation of organic substances in wastewater treatment plants. Alkali treatment at elevated temperature converts it into nontoxic biodegradable compounds by eliminating chlorine.

2.6. Polymeric Chloroacetaldehydes

Trichloroparaldehyde [1129-52-8] 2,4,6-tris (chloromethyl)-1,3,5-trioxane, $C_6H_9Cl_3O_3$, M_r 235.50, *mp* 88–89 °C, *bp* 140–144 °C (1.3 kPa), colorless crystals, soluble in common organic solvents, can be recrystallized from ethanol and forms monochloroacetaldehyde on heating in the presence of acids, e.g., oxalic acid.

Trichloroparaldehyde is formed by the reaction of concentrated sulfuric acid with the hemihydrate of monochloroacetaldehyde [38] and is also part of the residue (with tetrachlorometaldehyde) left by azeotropic distillation of the hemihydrate [9]. Trichloroparaldehyde can also be obtained by decomposing 1,2-dichloroethyl nitrate in the presence of Lewis acids, such as tin tetrachloride [39].

Tetrachlorometaldehyde [7038-25-7] 2,4,6,8-tetrakis(chloromethyl)-1,3,5,7-tetroxocane, $C_8H_{12}Cl_4O_4$, M_r 313.99, is obtained as a residue following azeotropic dehydration of chloroacetaldehyde hemihydrate [9]. It forms colorless crystals, *mp* 65–67 °C, *bp* 127–130 °C at 1.3 Pa.

Polychloroacetaldehydes [27577-42-0], $(C_2H_3ClO)_x$, are formed by the polymerization of anhydrous monochloroacetaldehyde at −40 to −78 °C, particularly in the presence of such Lewis acids as boron trifluoride in ether or chloroform, to form amorphous, elastomeric products with a polyacetal (polyoxymethylene) structure [40, 41].

Crystalline polymers are obtained at −78 °C in the presence of catalytic amounts of organometallic compounds, such as triethylaluminum, diethylzinc, or butyllithium [42].

Polymeric chloroacetaldehydes are commercially unimportant.

3. Dichloroacetaldehyde

Dichloroacetaldehyde [*79-02-7*], 2,2-dichloroethanal, $CHCl_2CHO$, M_r112.94, was first obtained in 1868 by F. Paterno by distillation of dichlorodiethyl acetal [*619-33-0*], $CHCl_2CH(OC_2H_5)_2$, with sulfuric acid.

3.1. Physical Properties

Anhydrous dichloroacetaldehyde is a colorless liquid with a strong odor, *mp* −37.6 to −37.4 °C, *bp* 89.2 °C, d_{25} 1.4113, dipole moment 2.36 D at 30 °C, soluble in water with formation of a hydrate and in common organic solvents.

Dichloroacetaldehyde monohydrate [16086-14-9], $CHCl_2CH(OH)_2$, 2,2-dichloro-1,1-ethanediol, crystallizes from water, M_r 130.96, *mp* 35–50 °C, *bp* 85–95 °C at 101.3 kPa, d_4^{20} 1.53–1.54, vapor pressure ca. 6.5 kPa at 20 °C and ca. 25 kPa at 50 °C, soluble in polar organic solvents, insoluble in nonpolar organic solvents. The flash point is ca. 95 °C and the ignition temperature ca. 605 °C.

3.2. Chemical Properties and Uses

On standing, dichloroacetaldehyde forms a solid, colorless polymer which depolymerizes if heated to 120 °C [43]. Dichloroacetaldehyde shows all of the typical aldehyde reactions.

Oxidation of dichloroacetaldehyde with chromic acid yields dichloroacetic acid [44]. Reduction with ethylaluminum gives 2,2-dichloroethanol [45]. Aldol condensation yields 2,2,4,4-tetrachloroacetaldol [46].

Condensation with ethylbenzene yields *p,p′*-diethyl-1,1-diphenyl-2,2-dichloroethane [*72-56-0*], which is known as Perthane and is used as an insecticide. Condensation with chlorobenzene produces *p,p′*-dichloro-1,1-diphenyl-2,2-dichloroethane [*72-54-8*], which is also an insecticide and is known as TDE or DDD.

$$CHCl_2CHO + 2C_6H_5Cl \rightarrow CHCl_2CH(C_6H_4Cl)_2 + H_2O$$

The most important use for dichloroacetaldehyde and its acetals is in the synthesis of the diuretic trichloromethiazid and the cytostatic mitotane [*53-19-0*],1,1-dichloro-2-(*o*-chlorophenyl)-2-(*p*-chlorophenyl)ethane.

3.3. Production

The most important industrial production process is the chlorination of acetaldehyde or paraldehyde. Hypochlorination of 1,2-dichloroethylene yields pure dichloroacetaldehyde.

3.3.1. Chlorination of Acetaldehyde or Paraldehyde

Dichloroacetaldehyde is formed, along with mono- and trichloroacetaldehyde, by the chlorination of acetaldehyde or paraldehyde [47]. High yields are obtained in the presence of antimony trichloride (83 % yield) [48] or phosphoric acid (> 90 % yield) [49].

Dichloroacetaldehyde is also formed, together with mono- and trichloroacetaldehyde, by oxychlorination in the preparation of acetaldehyde by the Wacker process (→ Acetaldehyde).

3.3.2. Hypochlorination of 1,2-Dichloroethylene

Dichloroacetaldehyde is obtained without impurities such as chloral and monochloroacetaldehyde, by treating 1,2-dichloroethylene with 1 mol of chlorine at 0–20 °C in aqueous dioxane [50] or in water alone with vigorous mixing [51].

3.3.3. Chlorination of Ethanol

Chlorination of ethanol can be controlled in the presence of nickel(II) chloride to obtain principally dichloroacetaldehyde along with some chloral and monochloroacetaldehyde. Dichloroacetaldehyde forms a hemiacetal, which can be used in the same manner as the pure aldehyde or the hydrate [52].

3.4. Availability, Storage, and Transportation

Dichloroacetaldehyde is available in a polyethylene container either as the solid hydrate (minimum 82 % anhydrous product) or as an aqueous solution (minimum 65 % anhydrous product).

3.5. Polymeric Dichloroacetaldehydes

Hexachloroparaldehyde [17352-16-8,] 2,4,6-tris(dichloromethyl)-1,3,5-trioxane, $C_6H_6Cl_6O_3$, M_r 338.83, *mp* 131–132 °C, colorless crystals, *bp* 210–220 °C with the formation of small amounts of dichloroacetaldehyde.

Hexachloroparaldehyde is made by treating dichloroacetaldehyde with such Lewis acids as antimony trichloride, iron(III) chloride, aluminum trichloride, tin(IV) chloride, or boron trifluoride [53]. Hexachloroparaldehyde is also formed by the direct chlorination of paraldehyde with 6–7 mol of chlorine at 35 °C under anhydrous conditions. It forms dichloroacetaldehyde hydrate in the presence of aqueous acids [54].

Polydichloroacetaldehydes [28388-44-5] , $(C_2H_2Cl_2O)_x$, are formed by polymerization of dichloroacetaldehyde. Crystalline polymers are obtained by treating dichloroacetaldehyde at −78 °C with organometallic compounds such as triethylaluminum, diethylzinc, or butyllithium [42]. Amorphous polymers are obtained by using Lewis acids at a temperature below 0 °C [55]. These products have a polyoxymethylene (polyacetal) structure and are more or less soluble in common organic solvents, depending on the degree of polymerization. They dissolve in concentrated sulfuric acid at 100 °C with discoloration, are soluble in dimethylformamide, and form copolymers with other aldehydes.

4. Trichloroacetaldehyde

Trichloroacetaldehyde[*75-87-6*], chloral, 2,2,2-trichloroethanal, CCl_3CHO, M_r 147.39, was first synthesized by J. von Liebig in 1832 by chlorination of ethanol. It was the first hypnotic drug (1869) and is a precursor of the famous insecticide DDT, 1,1,1-trichloro-2,2-bis(4-chlorophenyl)ethane [*50-29-3*], discovered in 1941. Chloral was an important chemical in the 1960s, but its importance has steadily declined since then because the use of DDT and other chlorinated insecticides has been restricted.

4.1. Physical Properties

Chloral, *mp* −56.5 °C, *bp* 97.8 °C at 101.3 kPa, d_4^{20} 1.5121, n_D^{20} 1.4557, vapor pressure 4.7 kPa at 20 °C, is a colorless, mobile, oily liquid.

It has a penetrating odor and is readily soluble in water, alcohol, ether, and chloroform. The octanol–water partition coefficient is low (0.99). It forms azeotropes with 1,2-dichloroethane, heptane, and benzene.

4.2. Chemical Properties

The strong electron-withdrawing effect of its three chlorine atoms gives chloral chemical properties different from those of most aldehydes [56]. It has a strong tendency to form addition compounds with nucleophilic reagents. For example, with alcohols (ROH, R = alkyl) it forms isolable hemiacetals, $CCl_3CH(OH)OR$, and with water isolable chloral hydrate, $CCl_3CH(OH)_2$ [*302-17-0*], *mp* 53 °C; *bp* 97.5 °C at 101.3 kPa; d_4^{20} 1.908. Chloral cannot undergo many of the base-catalyzed condensations typical of aldehydes because strongly basic reagents tend to cleave its carbon–carbon bond.

$$CCl_3CHO + NaOH \rightarrow CHCl_3 + HCO_2Na$$
$$CCl_3CHO + R_2NH \rightarrow CHCl_3 + HCONR_2 \quad (R = \text{alkyl})$$

Under acidic conditions chloral usually reacts as a typical aldehyde, forming acetals with alcohols and condensing with aromatic compounds to form diaryltrichloroethanes, e.g., DDT from chlorobenzene.

$$CCl_3CHO + 2\,C_6H_5Cl \xrightarrow{H_2SO_4} CCl_3CH(C_6H_4Cl\text{-}p)_2 + H_2O$$

Anionic initiators such as lithium *tert*-butoxide convert chloral to an insoluble, nonflammable high polymer, $[CH(CCl_3)O]_n$, that has not found commercial use [57].

4.3. Production

Chloral is produced by the chlorination of acetaldehyde [58] (→ Acetaldehyde) or ethanol (→ Ethanol). Acetaldehyde is usually the more economical starting material.

$$CH_3CHO + 3Cl_2 \rightarrow CCl_3CHO + 3HCl$$

The chlorination is carried out in hydrochloric acid, which speeds up the reaction, represses condensation and aldehyde-oxidation reactions, and facilitates temperature control. Antimony trichloride is sometimes used as a catalyst. The process can be performed either in batches or continuously. The rate of chlorination decreases as the stepwise chlorination proceeds; thus, the temperature is gradually increased from 0 °C to 90 °C during the process. Chloral is distilled from the reaction mixture as the hydrate. The hydrate is then mixed with concentrated sulfuric acid, the heavier acid layer is drawn off, and chloral is distilled through a fractionating column of moderate height.

Chlorination of a mole of ethanol requires 4 mol of chlorine, with the first serving to oxidize the ethanol to acetaldehyde.

$$CH_3CH_2OH + Cl_2 \rightarrow CH_3CHO + 2HCl$$

To withstand hot hydrochloric acid, the reaction vessels and piping for these processes are lined with a ceramic or other acid-resistant material, and valves of nickel–molybdenum alloys such as Hastelloy B are used.

4.4. Analysis

Technical-grade chloral ranges in purity from 94 to 99 wt %, with water being the main impurity. Other impurities sometimes present are chloroform, hydrogen chloride, dichloroacetaldehyde, and phosgene.

A common analytical method is to treat the chloral for 2 min with 1*N* sodium hydroxide, which cleaves chloral to chloroform and sodium formate; the excess alkali is then titrated with acid [59]. Alternatively, chloral is treated with quinaldine ethyl iodide [*606-55-3*] to form a blue cyanine dye whose quantity is measured spectrophotometrically [60].

Gas chromatography can be used for quantitative analysis of chloral and its hydrate, which breaks down to chloral on vaporization.

4.5. Legal Aspects

Chloral hydrate and other forms of chloral are classified by the U.S. Bureau of Narcotics and Dangerous Drugs (BNDD) as Schedule IV drugs, i.e., drugs of relatively low potential for abuse that lead only to limited physical or psychological dependence. The substance is approved by FDA as a habit-forming CNS depressant for use in adult and pediatric medicine [61].

To possess chloral, one must have a license from the BNDD, keep stored chloral locked up, and maintain records of where it goes and for what purposes. Licensees are periodically audited by the BNDD. Similar regulations are found in other countries [62]. The WHO has established a provisional guideline of 10 μg/L for chloral and chloral hydrate in drinking water. The drinking water guideline for Australia and New Zealand is 20 μg/L. The minimization of the concentration of all chlorination byproducts is encouraged in different countries and in the EU [63].

4.6. Uses and Occurrence

The principal use of chloral is in the production of the insecticide DDT. Much smaller amounts are used to make other insecticides (methoxychlor, naled, trichlorfon, dichlorvos), a herbicide (trichloroacetic acid), and hypnotic drugs (chloral hydrate, chloral betaine, α-chloralose, and triclofos sodium). Chloral is not known to occur as natural product. For the general public, the major route of exposure to this substance is from drinking water. Chloral hydrate and its metabolites are formed as byproducts when water is disinfected with chlorine [61, 63].

4.7. Economic Aspects

Chloral (as the hydrate) has been produced for use as a hypnotic drug in relatively low and gradually declining volume for many years. As an indication of the scale, U.S. production for this purpose was about 135 t in 1978. A far greater demand for chloral arose during World War II to produce DDT, which was proving highly successful in controlling the transmission of typhus and malaria. Demand for chloral for insecticide production rose rapidly after the war and reached a peak around 1963 when 40 000 t was produced in the United States. It then declined as insects developed resistance to chlorinated insecticides, some animals at the top of food chains showed adverse effects from concentrating DDT in their bodies, and public concern arose about the possibility of harmful effects in humans. The use of DDT was banned in the United States in 1972 and subsequently in many other countries, but it is still produced for use in the tropics. Consumption of chloral in the United States for other pesticides was 1400 t in 1972. Montrose Chemical Corporation was the only United States producer of chloral in 1978 [64].

5. Toxicology

5.1. Chloroacetaldehyde

Chloroacetaldehyde is toxic if swallowed and in contact with the skin and it is highly toxic by inhalation, causing instant to delayed death in experimental animals. The acute toxicities in laboratory animals are: rat, oral: LD_{50} = 89 mg/kg; rabbit, dermal: LD_{50} = 267 mg/kg [65]; and rat, inhalation: LC_{50} (1-h) = 650 mg/m^3 [66]. For mice exposed to flowing air containing chloroacetaldehyde vapor, the LT_{50} (exposure time lethal to 50 % of the test animals) was 2.57 min. A skin sensitization test with guinea pigs (maximization test) was negative. Due to the corrosive nature of the compound it had to be diluted to 0.002 % [67]. Even ca. 0.03 % aqueous solutions cause definite eye irritation, and a 0.5 % solution causes skin irritation.

No signs of toxicity were observed in rats, mice, rabbits, and guinea pigs after inhalative exposure to a chloracetaldehyde vapor concentration of 5.2 mg/m^3 (7 h/d, 5 d/week) over six months [68].

Chloroacetaldehyde is mutagenic in a variety of test systems, such as bacteria, molds, fish and mammalian cells, and in human lymphocytes. The compound exhibits alkylating properties and reacts with single-strand DNA [69]. In two initiation/promotion experiments in which chloroacetaldehyde was subcutaneously and dermally administered to mice, no increase in tumor incidence was observed [70, 71]. No carcinogenic effects have been observed after oral application (one application per week) of 12.5 mg/kg b.w. to females and males over 630 (female) and 636 d (male). In a carcinogenicity study with male mice, drinking water with 0.1 g/L chloroacetaldehyde was administered over two years. In this test an increased incidence of liver tumors could be observed [68].

Human experience has shown that chloroacetaldehyde vapor is extremely irritating to the skin, mucous membranes, and respiratory tract

[67]. A 40 % aqueous solution is corrosive to skin and it represents a serious hazard of eye injury. Upon skin contact it causes skin effects resembling burn blisters and tissue damage after a latency period of 1–2 h [72]. Human exposure to 32.6 mg/m^3 of chloroacetaldehyde in air leads within some minutes to lacrimation and nasal irritation [68].

The recommended TLV-STEL value in the USA is 1 ppm [5], this also being the maximum airborne concentration to which workers should be exposed. In Germany the former MAK value of 3 mg/m^3 (1 ppm) was annulled due to the carcinogenic potential of the substance. Adequate personal protection (goggles, gloves, protection mask) must be applied to prevent contact with this highly dangerous compound.

The substance is officially classified in the EU (Annex 1, published list of substances with a harmonised classification and labelling) in accordance with Council Directive 67/548/EEC as "Very Toxic" (T+) and "Dangerous to the Environment" (N) with R-phrases 24/25-26-34-40-50 (Toxic in contact with skin and if swallowed—Very toxic by inhalation—Causes burns—Limited evidence of a carcinogenic effect—Very toxic to aquatic organisms).

5.2. Dichloroacetaldehyde

Dichloroacetaldehyde is known to be irritating to breathe, but no formal testing has been reported. It should be handled carefully, as with the other two chloroacetaldehydes [65, 66].

5.3. Trichloroacetaldehyde (Chloral)

Chloral and chloral hydrate have the same physiological properties and can be discussed together. Chloral hydrate was extensively used for many years to induce sleep and to calm or sedate, but in recent decades it has been relegated to a minor role by the barbiturates, benzodiazepines, etc. (→ Hypnotics and → Sedatives).

Because of chloral hydrate's long use as a drug, much is known about the effects of chloral on humans [62]. It is a central nervous system depressant. In therapeutic doses of 0.5–1.0 g, it has little effect on respiration and blood pressure, but in toxic doses it depresses both. Ethanol enhances chloral's effects. Chloral is quite irritating to skin and mucous membranes. In the body it is readily reduced to trichloroethanol, which is responsible for most of chloral's physiological action. The trichloroethanol is excreted in the urine, mainly as the glucuronide. Oral doses of 4–30 g have been lethal to adults [62].

Rat studies indicate that chloral is highly toxic by inhalation [64]. The LC_{50} for 4-h exposure is 440 mg/m^3. On exposure to 78 mg/m^3 of chloral 4 h/d for 10 d, rats had severe lung injury and 6 of 10 died; similar results were obtained with 75 mg/m^3 for 2 h/d for 13 d. It is mutagenic in the Ames test. TLV and MAK values have not been assigned for chloral. However, the inhalation studies in rats indicate that chloral should be handled at least as carefully as chloroacetaldehyde [66].

The carcinogenic potential of chloral and chloral hydrate has been evaluated by the IARC with the conclusion that there is inadequate evidence in humans for carcinogenicity of chloral and chloral hydrate and limited evidence for carcinogenicity of chloral hydrate in experimental animals. Therefore chloral hydrate is not classifiable as to its carcinogenicity to humans (Group 3) [63].

Chloral and Chloral hydrate are officially classified in the EU (Annex 1, published list of substances with a harmonised classification and labelling) in accordance with Council Directive 67/548/EEC as "Toxic" (T) with R-phrases 25-36/38 (Toxic if swallowed—Irritating to eyes and skin).

6. References

1. E. Kopp, Wacker-Chemie, unpublished result, 1967.
2. Physikalisch-Technische Bundesanstalt, Braunschweig, Federal Republic of Germany, 1980.
3. Deutsche Forschungsgemeinschaft, MAK-und BAT-Werte Liste, 2005.
4. Ausschuss für Gefahrstoffe bei der Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Technische Regel für Gefahrstoffe 900 Arbeitsplatzgrenzwerte, Ausgabe Januar 2006, Bundesarbeitsblatt 1-2006, 41–55.

5. American Conference of Governmental Industrial Hygienists (ACGIH), Threshold Limit Values for Chemical Substances and Physical Agents & Biological Exposure Indices, 2005.
6. National Institute for Occupational Safety and Health (NIOSH), Documentation of Immediately Dangerous to Life or Health Concentration (IDLH): NIOSH Chemical Listing and Documentation of Revised IDLH Values (as of 3/1/95); http://www.cdc.gov/niosh.
7. Soc. des Usines Chimiques Rhône-Poulenc, FR 1 423 671, 1964.
8. B. G. Yasnitskij, A. P. Zaitsev, *Zh. Org. Khim.* **2** (1966) 1022; *Chem. Abstr.* **65** (1966) 18 450; SU 119 875, 1959; *Chem. Abstr.* **54** (1960) 2178.
9. E. Kopp, J. Smidt, *Justus Liebigs Ann. Chem.* **693** (1966) 117.
10. Saint-Gobain Chauny & Cirey, FR 1 120 234, 1955.
11. R. M. Novak, *J. Org. Chem.* **28** (1963) 1182.
12. Consortium für elektrochemische Industrie, DE 1 141 631, 1961.
13. Wacker Chemie company publication, "Chloracetaldehyd" 1983.
14. Hoffmann La Roche Inc, US 3 872 089, 1975; *Chem. Abstr.* **83** (1975) 79 293.
15. Hoffmann La Roche Inc, DE-OS 2 221 623, 1972 (O. Hromatka and B. Binder); *Chem. Abstr.* **83** (1975) 486.
16. L. Givaudan & Ci, S.A., DE-OS 2 226 780, 1973 (P. Dubs M. Pesaro); *Chem. Abstr.* **78** (1973) 481.
17. O. Hromatka, R. Haberl, *Monatsh. Chem.* 1954, no. 5, 1088.
18. G. Künstle and Ch. Solbrig, Wacker-Chemie, unpublished results.
19. Velsicol Chem. Corp., CH 617 932, 1980 (J. Kreuzer); *Chem. Abstr.* **94** (1981) 15 741; A. Wohl et al., *Ber. Dtsch. Chem. Ges.* **22** (1889) 1354.
20. Shell Int. Res. Maatschapij Ger., DE-OS 1 914 954, 1969.
21. Schering AG, DE-OS 2 636 994, 1978; DE-OS 2 214 632, 1973.
22. R. H. Sieber Wacker-Chemie, *Justus Liebigs Ann. Chem.* **730** (1969) 31; DE 1 593 718, 1966.
23. W. Perkow, E. W. Krockow, K. Knoevenagel, *Chem. Ber.* **88** (1955) 662.
24. Degussa, DE-OS 2 239 278, 1972.
25. Wacker Chemie AG: market observation.
26. *Houben-Weyl*, **7/1**, p. 369.
27. T. Tsukamoto, Mitsubishi Chemical Ind. Co., JP 8027, 1954; *Chem. Abstr.* **50**, (1956) 13 998.
28. V. Valenta, CS 122 819, 1967; *Chem. Abstr.* **68** (1968) 49 067.
29. L. L. Shukovskaya, S. N. Ushakov, N. K. Galanina, *Izv. Akad. Nauk USSR, Otd. Khim. Nauk* 1962, 1692; *Chem. Abstr.* **58** (1963) 8891.
30. Dow Chemical Corp., US 2 803 668, 1956.
31. Dynamit-Nobel, FR 1 592 333, 1968; DE-OS 1 643 899, 1967.
32. Pechiney Compagnie de Produits Chimiques et Electrometallurgiques, FR 1 033 574, 1951.
33. B. G. Yasnitskij, A. P. Zaitsev, SU 162 136, 1964; *Chem. Abstr.* **61** (1964) 13 195.
34. Chemische Werke Hüls, DE 1 130 425, 1961.
35. Chemische Werke Hüls, DE 1 136 996, 1961.
36. H. Gross, *J. Prakt. Chem.* **21** (1963) 99.
37. F. Grüger, Wacker Chemie AG, unpublished results, 2006.
38. K. Natterer, *Monatsh. Chem.* **3** (1882) 447.
39. W. Fink, *Angew. Chem.* **73** (1961) 466; Monsanto Co., US 3 164 612, 1962.
40. T. Iwata, G. Wasai, T. Saegusa, J. Furukawa, *Makromol. Chem.* **77** (1964) 229.
41. Consortium für elektrochemische Industrie, DE 1 189 714, 1961.
42. G. Wasai, T. Iwata, K. Hirano, M. Suragano, T. Saegusa, J. Furukawa, *Kogyo Kagaku Zasshi* **67** (1964) 1920; *Chem. Abstr.* **62** (1965) 13 249.
43. M. N. Shchukina, *Zh. Obshch. Khim.* **18** (1948) 1653; *Chem. Abstr.* **43** (1949) 2575.
44. J. Rocek, *Tetrahedron Lett.* 1959, no. 5, 1.
45. J. Boeseken, *Rec. Trav. Chim. Pays-Bas.* **57** (1938) 75.
46. W. Deinhammer, *Consortium für elektrochemische Industrie*, unpublished results.
47. Shawinigan Chem. Ltd., GB 644 914, 1947.
48. FMC Corp., FR 1 337 054, 1962. FMC Corp., US 3 499 935, 1965.
49. FMC Corp., US 3 253 041, 1963; US 3 562 334, 1964.
50. Olin Mathieson, US 3 381 036, 1966.
51. Consortium für elektrochemische Industrie, DE-OS 2 306 335, 1973.
52. K. Akashi, *Bull. Inst. Phys. Chem. Res. Tokyo* **12** (1933) 329; *Chem. Abstr.* **27** (1933) 3447.

53. FMC Corp., US 3 322 792, 1963.
54. Michigan Chem. Corp., US 2 768 173, 1952.
55. Diamond Alkali Co., NL 6 408 665, 1963.
56. F. I. Luknitskii, *Chem. Rev.* **75** (1975) 259–289.
57. P. Kubisa, L. S. Corley, T. Kondo, O. Vogl et al., *Polym. Eng. Sci.* **21** (1981) 829–838.
58. W. T. Cave, *Ind. Eng. Chem.* **45** (1953) 1853–1857.
59. *United States Pharmacopeia*, Mack Easton, Pa., 1980, p. 129.
60. Association of Official Analytical Chemists: *Official Methods of Analysis*, Washington 1980, p. 623.
61. U.S. Environmental Protection Agency Toxicological Review of Chloral Hydrate (CAS No. 302-17-0); EPA/635/R-00/006; August 2000.
62. L. S. Goodman, A. G. Gilman, A. Gilman: *The Pharmacological Basis of Therapeutics*, MacMillan New York 1980, pp. 361–363, 1671.
63. World Health Organisation International Agency for Research on Cancer, IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 84, Some Drinking-Water Disinfectants and Contaminants, including Arsenic, Monograph Chloral and Chloral Hydrate, 2004, 317–358.
64. SRI International: *Chemical Economics Handbook*, Menlo Park Cal., 1979, p. 565, 52 000.
65. Registry of Toxic Effects of Chemical Substances (RTECS), NIOSH, Cincinnati, Ohio, 1997.
66. Wacker-Chemie, Report No. V 87.094/261236, 1987.
67. W. H. Lawrence et al., *J. Pharm. Sci.* **61** (1972) 19–25.
68. Deutsche Forschungsgemeinschaft, Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten, 1.-40. Lieferung 2006, Chloracetaldehyd, Wiley-VCH, Weinheim 2006.
69. J. S. Jacobson et al., *Genetics* **121** (1989) 213–222.
70. F. Zajdela et al., *Cancer Res.* **40** (1980) 352–356.
71. B. L. van Duuren et al., *J. Nat. Cancer Inst.* **63** (1979) 1433–1439.
72. Wacker-Chemie, personal communication, 1996.

Chloroacetic Acids

GÜNTER KOENIG, Hoechst Aktiengesellschaft, Augsburg, Federal Republic of Germany (Chaps. 2.6 – 4.6)

ELMAR LOHMAR, Hoechst Aktiengesellschaft, Köln, Federal Republic of Germany (Chaps. 2 – 2.5 and 5 – 8)

NORBERT RUPPRICH, Bundesanstalt für Arbeitsschutz, Dortmund, Federal Republic of Germany (Chap. 9)

1. Introduction 2
2. Chloroacetic Acid 2
2.1. Physical Properties 2
2.2. Chemical Properties 3
2.3. Production 3
2.3.1. Hydrolysis of Trichloroethylene 4
2.3.2. Chlorination of Acetic Acid 4
2.4. Quality Specifications 5
2.5. Uses 5
2.6. Derivatives 6
2.6.1. Sodium Chloroacetate 6
2.6.2. Chloroacetyl Chloride 7
2.6.3. Chloroacetic Acid Esters 7
2.6.4. Chloroacetamide 7
3. Dichloroacetic Acid 8
3.1. Physical Properties 8
3.2. Chemical Properties 8
3.3. Production 8
3.4. Quality Specifications 8
3.5. Uses 9
3.6. Derivatives 9
3.6.1. Dichloroacetyl Chloride 9
3.6.2. Dichloroacetic Acid Esters 9
4. Trichloroacetic Acid 10
4.1. Physical Properties 10
4.2. Chemical Properties 10
4.3. Production 10
4.4. Quality Specifications 10
4.5. Uses 10
4.6. Derivatives 10
4.6.1. Trichloroacetyl Chloride 10
4.6.2. Trichloroacetic Acid Esters 11
4.6.3. Trichloroacetic Acid Salts 11
5. Environmental Protection 12
6. Chemical Analysis 13
7. Containment Materials, Storage, and Transportation 13
8. Economic Aspects 14
9. Toxicology and Occupational Health 14
10. References 16

1. Introduction

Chloroacetic acid and its sodium salt are the most industrially and economically important of the three chlorination products of acetic acid discussed in this review. The sections on physical and chemical properties, production, quality specifications, uses, and derivatives are reported separately for each of these three acids, whereas those on environmental protection, chemical analysis, containment materials, storage, transportation, and economic aspects are considered together.

2. Chloroacetic Acid

Chlorinated acetic acids have become important intermediates in organic synthesis because of the ease of substitution of the Cl atoms. Chloroacetic acid [*79-11-8*], $ClCH_2COOH$, M_r 94.50, monochloroacetic acid, chloroethanoic acid, is the most industrially significant [1]. It does not occur in nature and was first discovered as a chlorination product of acetic acid by N. LEBLANC in 1841. It was synthesized in 1857 by R. HOFFMANN, who chlorinated acetic acid by using sunlight to initiate the reaction. Discovery of other reaction accelerators, such as phosphorus, iodine, sulfur, or acetic anhydride, followed rapidly. Development of commercial processes, based mainly on acetic acid chlorination and later on acid hydrolysis of trichloroethylene, followed.

2.1. Physical Properties

Pure chloroacetic acid is a colorless, hygroscopic, crystalline solid, which occurs in monoclinically prismatic structures existing in the α-, β-, γ-, and also possibly the δ-form. Of these the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_537

α-form is the most stable and the most important industrially.

Published physical data vary widely [1]. Some of the most common values are as follows:

Modification:	α	β	γ	δ
fp, °C:	62 – 63	55 – 56	50 – 51	43.8
Latent heat of fusion ΔH_f, kJ/mol:	19.38	18.63	15.87	–

Density:	d_4^{65} 1.3703 (liquid) d_4^{20} 1.58 (solid)
Refractive index:	n_D^{65} 1.4297
Surface tension:	σ 35.17 mN/m (100 °C)
Viscosity:	η 2.16 mPa · s (70 °C) 1.32 mPa · s (100 °C) 1.30 mPa · s (130 °C)
Degree of dissociation in water (potentiometric):	1.52×10^{-3} (25 °C)
Dielectric constant:	16.8 (100 °C)
Electrical conductivity, lowest value measured (70 °C):	3.1 µS/cm, (rises steeply if traces of water present)
Specific heat capacity:	c_p 144.02 J mol^{-1} K^{-1} (solid, 45 – 15 °C) 180.45 J mol^{-1} K^{-1} (liquid, 70 °C) 187.11 J mol^{-1} K^{-1} (liquid, 130 °C)
Heat of combustion:	ΔH_c 715.9 kJ/mol
Heat of evaporation:	ΔH_v 50.09 kJ/mol
Heat of formation (100 °C):	ΔH_f − 490.1 kJ/mol
Heat of sublimation (25 °C):	ΔH_{subl} 88.1 kJ/mol
Heat of solution in H_2O (16 °C):	
solid	$\Delta H_{solv,}$ − 14.0 kJ/mol
liquid	$\Delta H_{solv.}$ 1.12 kJ/mol
Flash point (DIN 51 758):	126 °C
Ignition temperature (DIN 51 794):	470 °C
Lower explosion limit in air (101.3 kPa):	8 vol%

Vapor pressure

t, °C	189	160	150	140	130	100	90	80
p, kPa	101.3	40.0	28.0	19.0	13.0	4.3	2.6	1.1

Solubility in water, α-form

t, °C	0	10	20	30	40	50	60
g/100 g soln.	71.0	76.0	80.8	85.8	90.8	95.0	99
g/100 g H_2O	245	317	421	604	987	1900	–

Chloroacetic acid has excellent solubility in water and good solubility in methanol, acetone, diethyl ether, and ethanol, but is only sparingly soluble in hydrocarbons and chlorinated hydrocarbons. Chloroacetic acid forms azeotropes with a number of organic compounds [2]. The freezing points of various binary mixtures of chloroacetic acids are shown in Figure 1.

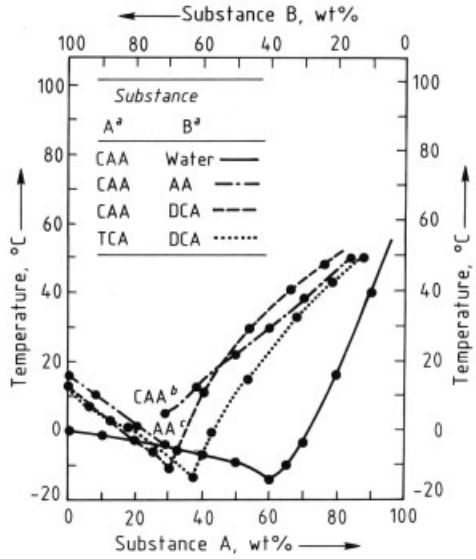

Figure 1. Freezing points of binary mixtures
[a] Acetic acid (AA), chloroacetic acid (CAA), dichloroacetic acid (DCA), trichloroacetic acid (TCA). [b] Crystalline phase CAA. [c] Crystalline phase AA.

2.2. Chemical Properties

The high reactivity of the carboxylic acid group and the ease of substitution of the α-Cl atom are directly related. As a result, chloroacetic acid is a common synthetic organic intermediate, either as the acid itself or as an acid derivative (e.g., salt, ester, amide, hydrazide, etc.). Some important reactions that are used for industrial applications are as follows.

Reaction with inorganic bases, oxides, and carbonates or with organic bases gives salts; some salts form adducts with chloroacetic acid. Sodium chloroacetate [*3926-62-3*] is an important commercial product.

Chloroacetic acid esters are obtained by reaction with alcohols or olefins; methyl chloroacetate [*96-34-4*] and ethyl chloroacetate [*105-39-5*] are also important industrially. Chloroacetyl chloride [*79-04-9*] is produced from the acid by reaction with $POCl_3$, PCl_3, PCl_5, thionyl chloride ($SOCl_2$), phosgene ($COCl_2$), etc. (see Section 2.6.2.).

Chloroacetic acid reacts with chloroacetyl chloride to form bis(chloroacetic)anhydride [*541-88-8*], which can also be obtained by dehydration of chloroacetic acid with P_2O_5 or by reaction of chloroacetic acid with acetic anhydride. Chloroacetyl chloride forms mixed anhydrides with other carboxylic acids, e.g., acetic chloroacetic anhydride [*4015-58-1*].

Nucleophilic substitution of the chlorine atom is an important reaction when the product is used as an intermediate in organic syntheses. For example, heating neutral or basic aqueous solutions hydrolyzes the chlorine atom. This is an industrial method of producing glycolic acid [*79-14-1*] (hydroxyacetic acid) and diglycolic acid [*110-99-6*] (2,2′-oxydiacetic acid).

Heating the salts gives glycolide, 1,4-dioxine-2,5-dione [*502-97-6*]. Reaction with sodium or potassium hydrogen sulfide forms thioglycolic acid [*68-11-1*] and thiodiglycolic acid [*123-93-3*].

Reaction with ammonia gives either aminoacetic acid [*56-40-6*] (glycine) as the main product or, depending on reaction conditions, nitrilotriacetic acid [*139-13-9*]. If methyl chloroacetate reacts with ammonia at low temperature, chloroacetamide [*79-07-2*] is obtained.

Aromatic compounds, such as naphthalene, undergo electrophilic substitution with chloroacetic acid over suitable catalysts to form arylacetic acids.

Reaction with potassium cyanide in a neutral solution gives the commercially important cyanoacetic acid [*372-09-8*]. Reaction with potassium iodide forms iodoacetic acid [*64-69-7*].

The corresponding phenoxyacetic acids, some of which are of industrial importance, are made by phenol etherification in the presence of sodium hydroxide. Another industrially significant ether formation process gives carboxymethyl derivatives with a relatively high degree of etherification by reacting polysaccharides, such as cellulose, starch, guar, etc., in a strongly alkaline sodium hydroxide medium.

2.3. Production

A multitude of methods have been proposed and patented for production of chloroacetic acid [1], [3–15]. However, of all the methods described, only two are used on an industrial scale: the hydrolysis of 1,1,2-trichloroethylene [*79-01-6*] catalyzed with sulfuric acid (Eq. 1), and the catalyzed chlorination of acetic acid with chlorine (Eq. 2).

$$ClCH{=}CCl_2 + 2\,H_2O \xrightarrow[\text{Heat}]{H_2SO_4} ClCH_2COOH + 2\,HCl \qquad (1)$$

$$CH_3COOH + Cl_2 \xrightarrow[\text{Heat}]{\text{Cat.}} ClCH_2COOH + HCl \qquad (2)$$

The first method (Eq. 1) has become of technological significance; the second (Eq. 2) is older, historically and commercially, and currently is the more important process.

2.3.1. Hydrolysis of Trichloroethylene [13–15]

Equal amounts of trichloroethylene and 75 % sulfuric acid are reacted at 130 – 140 °C in a continuous process so that with complete trichloroethylene conversion, the resultant reaction mixture contains about 50 % chloroacetic acid and 1 – 2 % water. This blend is vacuum distilled to give pure chloroacetic acid. During this process the vapors are washed with water, which is returned to the sulfuric acid as a diluent. The resultant hydrogen chloride gas is washed with the fresh trichloroethylene and then purified by freezing and absorbing in water. Trichloroethylene (1500 – 1850 kg) and H_2SO_4 (600 kg, 95 %) gives 1000 kg of finished product and 700 – 750 kg of HCl gas as byproduct.

The trichloroethylene method produces highly pure chloroacetic acid free of di- or trichloroacetic acid. The purification procedure consists of separation from trichloroethylene, sulfuric acid, and water. Despite the purity of the chloroacetic acid formed, this method has fallen into disuse because of the high cost of trichloroethylene and the large amount of HCl produced. It is, however, still used on an industrial scale in several countries (Spain, former States of USSR, Romania, Yugoslavia, Japan, German Democratic Republic, etc.).

2.3.2. Chlorination of Acetic Acid

Synthesis. In this method the object is to convert acetic acid to chloroacetic acid with high

selectivity [1]. This is achieved by using suitable catalysts [13], [16–33]. When acetic anhydride is the catalyst, the reaction mechanism is as follows [22]:

$$(CH_3CO)_2O + Cl_2 \longrightarrow ClCH_2\overset{O}{\overset{\|}{C}}-O-\overset{O}{\overset{\|}{C}}CH_3 + HCl$$

$$ClCH_2\overset{O}{\overset{\|}{C}}-O-\overset{O}{\overset{\|}{C}}CH_3 + CH_3COOH \longrightarrow ClCH_2COOH + (CH_3CO)_2O$$

$$(CH_3CO)_2O + HCl \longrightarrow CH_3COCl + CH_3COOH$$

$$CH_3COCl + Cl_2 \longrightarrow ClCH_2COCl + HCl$$

$$ClCH_2COCl + CH_3COOH \longrightarrow ClCH_2COOH + CH_3COCl$$

Various inhibitors have also been proposed to suppress formation of dichloroacetic acid, which results from chlorination of chloroacetic acid in the crude mixture. This eliminates or reduces the purification process for technical grades [34–38].

Purification. The high degree of purity required for many products can only be obtained by separating the di- and trichloroacetic acids. Fractional distillation is unsuitable because the boiling points of the three chlorinated acetic acids are so close. Azeotropic distillation [39] and extractive distillation [40], [41] have been suggested for separating dichloroacetic acid; it is doubtful, however, that these methods are used.

One of the most widely practiced purification processes is *crystallization* without use of a solvent. It is based on the higher melting point of the α-modification of chloroacetic acid. The di- and trichloroacetic acids are removed in the mother-liquor.

Crystallization is carried out either in stationary finger crystallizers [42] or in agitated stirrer crystallizers. With the latter, the mother-liquor is separated from the crystal slurry after crystallization by using a centrifuge [43], [44]. The product is washed with water or acetic acid and discharged. The chloroacetic acid is usually melted and converted to flakes. In stationary machines, crystallization is carried out by using cold fingers. When all the chloroacetic acid has crystallized, the mother-liquor is drained; the pure crystalline product is then melted and flaked.

Another purification method that has been described is a thin-layer crystallization process with the raw material [45]. Use of a water content of 5 – 25 % without organic solvent is also possible [46].

Solvents have also been used for crystallization. Solvents, such as carbon tetrachloride [44], [47], dichloromethane [48], or hydrocarbons with three chlorine atoms (e.g., trichloroethylene), give crystals that are easy to filter [49].

Common to all of these crystallization methods is obtaining a mother-liquor consisting of acetic acid, chloroacetic acid, and di- and trichloroacetic acids. In the most favorable cases, this mixture is further chlorinated to form the industrially useful trichloroacetic acid (see Section 4.3).

In many instances, di- and trichloroacetic acids in either the mother-liquors, resulting from crystallization, or in the total chlorinated crude acid can be dechlorinated by catalytic hydrogenation at elevated temperature to form chloroacetic acid or acetic acid. Palladium on a carrier, such as carbon or silica gel, is normally employed.

When the reaction is carried out in the vapor phase, dichloroacetic acid is dechlorinated primarily to acetic acid [50]. However, when Pd is used on a finely dispersed, inert carrier in the liquid phase at 130 – 150 °C, dichloroacetic acid is dechlorinated selectively to form chloroacetic acid [51]. Modifications of this procedure, such as spraying the crude acid with the hydrogen gas under vacuum [52] or trickling the acid over the catalyst in the fixed bed [53], [54] have also been described.

Selectivity is increased if HCl is mixed with the crude acid before it and the circulating hydrogen contact the catalyst in the fixed bed [55]. Acid chlorides and anhydrides are saponified before dechlorination [56]. A particularly active and selective catalyst is Pd on silica gel (particle size 40 – 200 μm) [57]. Especially good results are obtained by employing a cocatalytically effective additive, such as sodium acetate [58]. The catalyst can be made more effective by enrichment of the nobel metal on its surface [59], [60]. Spent catalyst can be reactivated by treatment with chlorine [61], [62].

An industrial-scale chlorination process is shown in Figure 2. The mixture of acetic acid, acetic anhydride, and recycled acetyl chloride is chlorinated at 90 – 140 °C in reactor (a) or in several cascade reactors. Only traces of chlorine are still present in the HCl gas formed. Chloroacetic

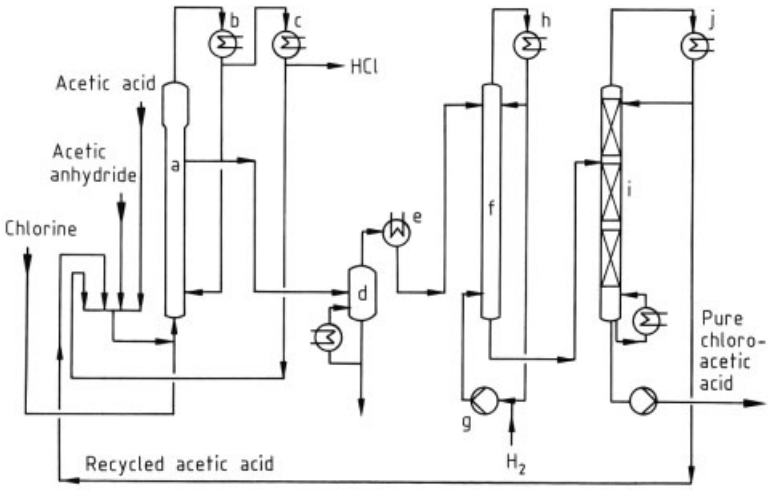

Figure 2. Chloroacetic acid obtained by the chlorination – hydrogenation process
a) Chlorinating reactor; b) Condenser for acetic acid, acetic anhydride, chloroacetic acid; c) Condenser for acetyl chloride; d) Evaporator; e) Condenser for chloroacetic acid; f) Hydrogenation reactor; g) Hydrogen compressor; h) Condenser; i) Distillation column; j) Condenser for acetic acid

acid, acetic acid, and acetic anhydride are condensed by using water-cooled condensers (b) and then returned to the reactor. Acetyl chloride entrained in the HCl gas is condensed (c) in a subsequent low-temperature process and recycled. The HCl gas is further purified and usually converted to concentrated aqueous hydrochloric acid.

The crude acetic acids are vacuum distilled (d). Di- and trichloroacetic acids in the crude distilled material are dechlorinated (f) to chloroacetic acid at 120 – 150 °C, using a palladium catalyst and a large excess of hydrogen. Acetic acid is taken overhead from the vacuum fractionation column (i); the bottom product is pure chloroacetic acid.

A total of 660 – 780 kg of acetic acid and 780 – 1020 kg of chlorine are required per 1000 kg of pure acid, depending on the method used (in crystallization processes, the mother-liquors are regarded as lost). The process also gives 400 – 420 kg of HCl.

2.4. Quality Specifications

Chloroacetic acid is usually marketed to the following specifications:

Chloroacetic acid	min. 99.0 wt %
Dichloroacetic acid	max. 0.5 wt %
Acetic acid	max. 0.2 wt %
Sulfate	max. 0.5 wt %
Water	max. 0.2 wt %
Iron	max. 5 mg/kg
Lead	max. 1 mg/kg

Acid manufactured by the trichloroethylene method is virtually free of dichloroacetic acid; specially purified acid from the chlorination of acetic acid is marketed with an acetic acid content of max. 0.1 % and max. 0.3 % dichloroacetic acid. Technical grades contain up to 5 % dichloroacetic acid.

2.5. Uses

Most of the chloroacetic acid produced is used to manufacture several hundred thousand tons annually of *carboxymethyl cellulose* [*9004-32-4*], CMC (→ Cellulose Ethers, Chap. 6.2.). Starch can be reacted with chloroacetic acid to give carboxymethyl starch, which is as widely used as CMC (→ Starch). Other polysaccharides modified with chloroacetic acid are less important.

Another major application is the production of *herbicides* based on arylhydroxyacetic acids (→ Weed Control, → Chlorophenoxyalkanoic

Acids). These herbicides are some of the most widely used. Chloroacetic acid and methyl chloroacetate are also employed for making the insecticide dimethoate and the herbicides benazoline and methyl β-naphthyloxyacetate.

A third important outlet for chloroacetic acid is the manufacture of *thioglycolic acid* (mercaptoacetic acid [*68-11-1*]), obtained fro sodium or potassium hydrogen sulfide or other sulfur compounds (→ Thiols and Organic Sulfides). It is used as its salt, ester, or another derivative. The largest amount is employed to produce stabilizers for poly(vinyl chloride). Moreover, thioglycolic acid and its esters are used in hair cosmetics (→ Hair Preparations).

A minor use of chloroacetic acid is in the production of glycolic acid by saponification with an alkali hydroxide. Glycolic acid is used as an auxiliary in textile printing, leather treating, furs finishing, and as a component for cleaners; the butyl ester is employed as a paint additive, and glycolic acid esters acylated with *o*-phthalic acid half-esters are used as plasticizers (→ Plasticizers).

Apart from the major fields of application mentioned above, chloroacetic acid and its derivatives are used in a multitude of other organic synthetic reactions. For example, caffeine and barbiturates, which are important as sleeping tablets, can be made from cyanoacetic acid or its esters. Chloroacetic acid condenses with aromatic hydrocarbons to form arylacetic acids. Reaction with naphthalene gives 1-naphthylacetic acid as the main product and 2-naphthylacetic acid as the byproduct. Both substances promote plant growth. Chloroacetic acid also is important in the syntheses of coumarin and vitamin B_6 (→ Vitamins).

2.6. Derivatives

2.6.1. Sodium Chloroacetate

Physical, Chemical Properties. The sodium salt $ClCH_2COONa$ [*3926-62-3*], M_r 116.5, is of particular importance. It is colorless and slightly hygroscopic, and has good storage stability. It dissolves readily in water (44 wt % at 20 °C), giving a neutral solution. It has limited solubility in other polar solvents and is insoluble in nonpolar solvents. It hydrolyzes in water, depending on temperature and time, forming glycolic acid and sodium chloride.

Production. This salt is manufactured by reacting sodium carbonate with chloroacetic acid in batches or in a continuous process [63], [64].

In exceptional cases, localized superheating (about 150 °C) may occur during the reaction. If it does, slow thermal decomposition can take place, producing sodium chloride and polyglycolide as the main products with pronounced evolution of gas.

Another method of manufacturing sodium chloroacetate is spraying molten chloroacetic acid together with 50 % caustic soda solution into a spray drier [65].

The production of salts of chloroacetic acids in fluidized beds has been described [66].

Uses. The uses for sodium chloroacetate are virtually the same as those for its free acid. The amount of salt required is less and depends on whether the free chloroacetic acid is used as an 80 % aqueous solution or as a melt.

2.6.2. Chloroacetyl Chloride

Physical Properties. Chloroacetyl chloride [*79-04-9*], $ClCH_2COCl$, M_r 112.95, is a colorless, highly corrosive liquid that has a pungent odor and fumes when exposed to moist air; *bp* 105 °C (101.3 kPa), d_4^{20} 1.42, n_D^{20} 1.454.

Production. Chloroacetyl chloride is usually manufactured from chloroacetic acid by reaction with phosphorus trichloride, thionyl chloride, sulfuryl chloride, or phosgene. It is also obtained by chlorination of acetyl chloride in the presence of stronger aliphatic acids, preferably chloroacetic acids, or from sodium chloroacetate and the usual chlorinating agents.

One patent describes the manufacture of chloroacetyl chloride by chlorination of a mixture of 5 – 50 wt % acetyl chloride in acetic anhydride at 70 – 100 °C [67]. Another claims reaction of chloroacetic acid and trichloroethylene in the presence of iron(III) chloride and hydrochloric acid at 150 °C and 2 MPa [68]. Chloroacetyl chloride also has been obtained in 97.1 % yield and with a purity of 99.8 % by reacting chloroacetic acid with phosgene in the

presence of palladium chloride at 110 °C [69]. Chlorination of ketene, which must be present in an excess of at least 50 %, at 100 – 200 °C gives chloroacetyl chloride with less than 7 % dichloroacetyl chloride [70]. Chloroacetyl chloride can also be made from 1,2-dichloroethylene and oxygen by using catalysts, such as bromine.

Uses. Chloroacetyl chloride is used for many syntheses, e.g., to make adrenaline, chloroacetic acid esters, and the anhydride.

2.6.3. Chloroacetic Acid Esters

Physical Properties. The *methyl ester* [*96-34-4*], $ClCH_2COOCH_3$, M_r 108.53, is of particular importance. It is a colorless liquid with a pungent odor, *bp* 130 – 131 °C (101.3 kPa), *fp* – 32.7 °C, d^{20} 1.236, soluble in alcohol and ether, and only sparingly soluble in water.

Also important is the *ethyl ester* [*105-39-5*], $ClCH_2COOCH_2CH_3$, M_r 122.55, *bp* 144.2 °C (101.3 kPa), *fp* – 26.0 °C, d^{20} 1.159, insoluble in water, and readily soluble in alcohol and ether.

Production. These esters can be manufactured from chloroacetic acid and either methyl or ethyl alcohol. In another method, trichloroethylene is converted to ethyl 1,2-dichlorovinyl ether, which can be readily hydrolyzed to form chloroacetic acid ethyl ester.

Uses. The reactivity of the ester, which is greater than that of the free acid, makes it suitable for many syntheses, e.g., sarcosine, chloroacetamide, thioglycolic acid ester [71] for pharmaceuticals (vitamin A), and crop protection agents (dimethoate). Condensation with aldehydes and ketones gives glycide esters [72]. Other uses are the synthesis of heterocyclic compounds, e.g., 2-phenylimino-4-oxooxazolidine from the ester and phenyl urea [73], and the well-known condensation of chloroacetic acid and its esters with thioureas to form pseudothiohydantoins.

2.6.4. Chloroacetamide

Physical Properties. Chloroacetamide [*79-07-2*], $ClCH_2CONH_2$, M_r 93.52, colorless crystalline needles, *bp* 224 °C (101.3 kPa), *fp* 121 °C, is soluble in water and alcohol, and sparingly soluble in all nonpolar solvents.

Production. Chloroacetamide is obtained on an industrial scale by reaction of methylchloroacetate with ammonia at low temperature [74]. Manufacture from chloroacetic acid and cyanamide at 150 – 200 °C also has been described [75].

Uses. Chloroacetamide is a versatile intermediate. In addition, it has biocidal properties and, therefore, is used as an industrial preservative. Because of its good solubility in water, chloroacetamide is a particularly suitable biocide for protection of the aqueous phase, e.g., in drilling fluids [76], in water-containing paints [77], and as a wood preservative [78]. Its insecticidal action on aphids [79] and its use as a hardener for urea and melamine resins [80] have also been described.

Various derivatives have the same biocidal effect as chloroacetamide. For instance, *N*-octadecylchloroacetamide is used as an antimicrobial plasticizer [81].

3. Dichloroacetic Acid

3.1. Physical Properties

Dichloroacetic acid [*79-43-6*], 2,2-dichloroethanoic acid, $Cl_2CHCOOH$, M_r 128.95, *bp* 192 °C (101.3 kPa), 102 °C (2.7 kPa), *fp* 13.5 °C, d_4^{20} 1.564, n_D^{20} 1.466, vapor pressure 0.19 kPa (at 20 °C), dissociation constant 5×10^{-2} mol/L (at 18 °C), is a colorless, highly corrosive liquid that gives off acidic vapors, which irritate the mucous membranes. It is miscible with water in any proportion. Dichloroacetic acid is readily soluble in the usual organic solvents, such as alcohols, ketones, hydrocarbons, and chlorinated hydrocarbons.

3.2. Chemical Properties

The two chlorine atoms of dichloroacetic acid are susceptible to displacement. For instance,

with aromatic compounds, diaryl acetic acids are formed, and with phenol, diphenoxy acetic acids are the products. However, dichloroacetic acid is less prone to hydrolysis than chloroacetic acid. In the manufacture of carboxymethyl celluloses and starches, the dichloroacetic acid impurity in the chloroacetic acid gives rise to cross-linking, which is either desirable or undesirable, depending on the use of the end product.

3.3. Production

The most cost-effective production method is the hydrolysis of dichloroacetyl chloride (see Section 3.6.1). Moreover, 98 % dichloroacetic acid can be obtained in 90 % yield by hydrolysis of pentachloroethane with 88 – 99 % sulfuric acid [82] or by oxidation of 1,1-dichloroacetone with nitric acid and air [83]. Extremely pure dichloroacetic acid can be produced by hydrolysis of the methyl ester [84], which is readily available by esterification of crude dichloroacetic acid. Furthermore, dichloroacetic acid and ethyl dichloroacetate can be obtained by catalytic dechlorination of trichloroacetic acid or ethyl trichloroacetate with hydrogen over a palladium catalyst [85].

Separation of pure dichloroacetic acid from the other chloroacetic acids cannot be carried out by physical methods, especially fractional distillation, because the differences in boiling points, particularly between di- and trichloroacetic acid, are too small. The ester mixtures, on the other hand, can be satisfactorily fractionated in efficient distillation columns. In addition, mixtures of the salts of the three chloroacetic acids can be washed with water, alcohol, or water – alcohol solutions. The dichloroacetate can be dissolved preferentially and acidified to give pure dichloroacetic acid.

Dichloroacetic acid can be produced in the laboratory by reacting chloral hydrate and potassium or sodium cyanide:

$$Cl_3CCHO + H_2O + KCN \longrightarrow HCN + KCl + Cl_2CHCOOH$$

3.4. Quality Specifications

Dichloroacetic acid is a colorless liquid that is marketed, for instance, by Hoechst, with the following technical data:

Content	ca. 99 %
bp at 101.3 kPa	192 °C
Water, max.	0.3 %
fp	> 12 °C

3.5. Uses

Dichloroacetic acid is used as a test reagent for analytical measurements during fiber manufacture [poly(ethylene terephthalate)] and as a medicinal disinfectant (substitute for formalin).

Dichloroacetic acid, particularly in the form of its esters, is an important intermediate in organic synthesis. It is a reactive starting material for the production of glyoxylic acid, dialkoxy and diaroxy acids, and sulfonamides.

3.6. Derivatives

3.6.1. Dichloroacetyl Chloride

Physical Properties. Dichloroacetyl chloride [*79-36-7*], $Cl_2CHCOCl$, M_r 147.40, is a colorless liquid, which has an unpleasant odor and fumes when exposed to moist air; *bp* 107 °C (101.3 kPa), d_4^{16} 1.5315, n_D^{16} 1.4638, vapor pressure 3.06 kPa (at 20 °C).

Chemical Properties. Dichloroacetyl chloride undergoes not only reactions typical of acid chlorides, but also displacement reactions of the chlorine atoms in the 2-position. The chemistry of dichloroacetyl chloride and its derivatives is analogous in certain respects to that of glyoxylic acid. Thus, dichloroacetyl chloride reacts with ammonia and amines to form amino acids, with alcohols to form ester acetals and acetals, with benzene to form diarylacetic acids, and with phenols to form diphenoxyacetic acids.

Production. Dichloroacetyl chloride is produced by *oxidation of trichloroethylene.* Oxidation with oxygen to form a mixture of dichloroacetyl chloride and chloral has been known since the early 1900s [86]. More recent patents describe methods to manufacture dichloroacetyl chloride of > 98 % purity, e.g., at 65 – 200 °C and 0.2 – 2 MPa in the presence of free-radical

initiators [87], using catalytic quantities of azo compounds and amines [88], and initiating oxidation with UV light and only adding organic nitrogen bases in quantities of 0.005 – 0.05 % once oxidation has begun [89]. Correspondingly, chloroacetyl chloride and trichloroacetyl chloride can be produced from 1,2-di- and tetrachloroethylene, but these methods are of little industrial significance because the products can be obtained more easily by other processes.

Furthermore, dichloroacetyl chloride is manufactured from pentachloroethane and fuming sulfuric acid or from chloroform and carbon dioxide in the presence of aluminum chloride at high pressure [90]. It has also been produced from ketene and chlorine in the presence of sulfur dioxide [91].

Uses. Dichloroacetyl chloride is used to manufacture dichloroacetic acid (see Section 3.3). In addition, it can be employed for the production of esters and anhydrides. It is used as the starting material to manufacture pharmaceuticals, such as sulfonamides and antibiotics (chloramphenicol), and to produce crop protection agents.

3.6.2. Dichloroacetic Acid Esters

Physical Properties. *Methyl dichloroacetate* [*116-54-1*], $Cl_2CHCOOCH_3$, M_r 142.98, *bp* 143.3 °C (101.3 kPa), d_4^{20} 1.3808, n_D^{20} 1.4428. *Ethyl dichloroacetate* [*535-15-9*], $Cl_2CHCOOCH_2CH_3$, M_r 157.00, *bp* 158.5 °C (101.3 kPa), d_4^{20} 1.2821, n_D^{20} 1.4386.

Chemical Properties. These esters are highly reactive and, for example, condense with aldehydes and ketones in the presence of dilute metal amalgams to form α-chloro-β-oxyacid esters, which can be converted to the corresponding aldehydes via glycide esters [92]. Dichloroacetic acid esters are readily saponified by boiling water.

Production. These esters are obtained by the usual esterification methods. The methyl and ethyl esters are best produced by alcoholysis of dichloroacetyl chloride.

Uses. These esters are employed as intermediates in the manufacture of chloramphenicol (→ Antibiotics), dichloroacetamide, and crop protection agents, and are employed as paint resin solvents. The glyceryl and ethylene glycol esters also serve as plasticizers for cellulose derivatives.

The vapors of dichloroacetic acid esters are heavier than air and can form explosive mixtures with air.

4. Trichloroacetic Acid

4.1. Physical Properties

Pure trichloroacetic acid [*76-03-9*] forms hygroscopic, rhombohedral crystals that are extremely soluble in water and soluble in many organic solvents; Cl_3CCOOH, M_r 163.40, *bp* 197.6 °C (101.3 kPa), *bp* ca. 107 °C (2.8 kPa), *fp* 59.2 °C, d_4^{60} 1.630, n_D^{65} 1.459.

4.2. Chemical Properties

The complete substitution of the hydrogen atoms on the methyl radical by chlorine makes trichloroacetic acid a strong acid ($K = 2\times10^{-1}$ mol/L); it also makes the chlorine atoms less easily substituted than those of chloroacetic acid and dichloroacetic acid. Trichloroacetic acid decomposes to chloroform and carbon dioxide when its aqueous solution is exposed to heat. The decomposition is particularly fast in the presence of organic or inorganic bases. When water is absent, aniline, resorcinol, and activated carbon can catalyze the decomposition. Purely thermal decomposition takes place only when the boiling point has been exceeded, resulting in the formation of chlorinated hydrocarbons, carbon monoxide and dioxide, and phosgene. Trichloroacetic acid salts with inorganic and many organic bases are known. They decompose when heated in the presence of water to form chloroform.

4.3. Production

Trichloroacetic acid is produced on an industrial scale by chlorination of acetic acid or

chloroacetic acid mother-liquors at 140 – 160 °C [93]. If necessary, calcium hypochlorite is added as a chlorination accelerator [94]. There are conflicting views concerning adding heavy metal salts as chlorination catalysts. Examples of catalysts that have been used are iron and copper compounds, which are precipitated with sulfuric acid or phosphoric acid if decomposition of the reaction mixture occurs [95]; 2 % phosphoric acid [96]; and catalysts and UV light [97]. Trichloroacetic acid has also been produced without catalysts [98], [99].

The crude product, containing about 95 % trichloroacetic acid, is best isolated by crystallizing the melt, removing the mother-liquor with most of its impurities, and increasing the purity by centrifugation or recrystallization.

4.4. Quality Specifications

Trichloroacetic acid is marketed in various degrees of purity. The specifications in Table 1 are typical (see below).

4.5. Uses

The main application for trichloroacetic acid is production of its sodium salt, which is used as a selective herbicide and in formulations with 2,4-D and 2,4,5-T preparations as a total herbicide.

Trichloroacetic acid also is employed as an etching or pickling agent in the surface treatment of metals, as a swelling agent and solvent in the plastics industry, as an albumin precipitating agent in medicine, as an auxiliary in textile finishing, and as an additive to improve high-pressure properties in mineral lubricating oils. Because it is strongly corrosive, trichloroacetic acid is used to remove warts and hard skin and to treat various skin afflictions [101]. Trichloroacetic acid, particularly its esters, are important starting materials in organic syntheses. The acid undergoes numerous color reactions that can serve to identify a wide range of organic compounds [102].

4.6. Derivatives

4.6.1. Trichloroacetyl Chloride

Physical Properties. Trichloroacetyl chloride [*76-02-8*], Cl_3CCOCl, M_r 197.9, *bp* 118 °C (101.3 kPa), d_4^{20} 1.620, n_D^{20} 1.4695, is similar to dichloroacetyl chloride.

Chemical Properties. The acid chloride can be hydrolyzed at 75 – 85 °C with water to form the free acid and hydrolyzed with ammonium hydroxide or concentrated sodium carbonate solution to form the salts.

Production. The acetyl chloride is prepared from trichloroacetic acid and various inorganic acid chlorides (e.g., $SOCl_2$, PCl_3) or with P_2O_5

Table 1. Typical quality specifications

Property	Grade			
	Crude [a]	Industrial [a]	Ph. Eur. [b]	ACS [c]
Freezing point, min., °C	55	56	–	–
Trichloroacetic acid, % min.	96.5	98	98	99
Dichloroacetic acid, % max.	2.5	1.2	–	–
Sulfuric acid, % max.	0.5	0.3	–	0.02
Sulfated ash, % max.	–	–	0.1	0.03
Water, % max.	0.5	0.2	–	–
Heavy metals in the form of:				
Lead, ppm max.	10	5	–	20
Iron, ppm max.	20	10	–	10
Phosphate, ppm max.	–	–	–	5
Chloride, ppm max.	–	–	100	10
Nitrate, ppm max.	–	–	–	20

[a] Specifications of Hoechst.
[b] European Pharmacopeia (formerly DAB 8) [100].
[c] Specifications of ACS.

and HCl. More useful methods are the oxidation of tetrachloroethylene with fuming sulfuric acid, oxygen, or fuming nitric acid and sulfuric acid at 18 – 20 °C [103], or from the reaction of pentachloroethane and dry oxygen under UV light [104]. It has been obtained in 37 % yield from carbon tetrachloride and carbon monoxide in the presence of aluminum chloride at 200 °C and high pressure [105].

The most common production method is the gas-phase, photochemical oxidation of tetrachloroethylene with oxygen. The reaction is initiated with UV light, with radioactive irradiation, or it is sensitized with chlorine or iodine [106].

Uses. Trichloroacetyl chloride can be used for the manufacture of the esters and the anhydrides of trichloroacetic acid.

4.6.2. Trichloroacetic Acid Esters

Only the methyl and ethyl esters are of industrial interest. They can be used as solvents, if necessary, and for the production of the amide and polyalcohol esters, which have been suggested for use as plasticizers. *Methyl trichloroacetate* [*598-99-2*], $Cl_3CCOOCH_3$, M_r 177.43, *bp* 153 °C (101.3 kPa), d_4^{20} 1.4864, n_D^{20} 1.4572. *Ethyl trichloroacetate* [*515-84-4*], $Cl_3CCOOCH_2CH_3$, M_r 191.45, *bp* 167.5 °C (101.3 kPa), d_4^{20} 1.3823, n_D^{20} 1.4505.

4.6.3. Trichloroacetic Acid Salts

The sodium salt is the only one of industrial importance.

Physical Properties. Sodium trichloroacetate [*650-51-1*], TCA, sodium TCA, STCA, NaTA, $Cl_3CCOONa$, M_r 185.40, is a colorless salt that decomposes at temperatures below the melting point. The salt is very soluble in water and methanol and soluble in polar solvents. The solubility in water at − 5 °C is 50 wt %, at 20 °C 60 wt %, and at 40 °C 70 wt %.

Chemical Properties. The salt is stable in the dry state and at normal storage temperature. Hydrolytic cleavage takes place in aqueous solution:

$$Cl_3CCOONa + H_2O \longrightarrow Cl_3CH + NaHCO_3$$

At the boiling point, this reaction takes place quantitatively within an hour and is used for quantitative analysis of sodium trichloroacetate [107]. At 20 – 25 °C the aqueous solution is relatively stable; with a 50 % solution only ca. 1 % is hydrolyzed in 4 – 6 weeks.

In aprotic solvents, such as 1,2-dimethoxyethane, the reactive intermediate dichlorocarbene is formed as the primary product at elevated temperature (ca. 80 °C) [108]:

$$Cl_3CCOONa \longrightarrow NaCl + CO_2 + :CCl_2 \quad (3)$$

When phase-transfer catalysis is used, the dichlorocarbene can add, for example, across olefinic double bonds to give dichlorocyclopropane derivatives [108]. If dichlorocarbene is formed in the absence of other acceptors, it reacts with additional sodium trichloroacetate:

$$Cl_3CCOONa + :CCl_2 \longrightarrow Cl_3CCOCl + CO + NaCl \quad (4)$$

Consequently, the main products formed during the thermal decomposition of sodium trichloroacetate (combination of Eqs. 3 and 4) were sodium chloride, trichloroacetyl chloride, carbon monoxide, and carbon dioxide; also observed were chloroform, carbon tetrachloride, and traces of phosgene [109].

Thermal decomposition is exothermic (42 kJ/mol) and starts between 125 and 170 °C. This decomposition reaction is quenched by adding water; hydrolysis then takes place.

Production. Sodium trichloroacetate is produced industrially by neutralizing trichloroacetic acid with sodium hydroxide solution or sodium carbonate [110], [111].

Uses. Sodium trichloroacetate, together with reducing agents and traces of heavy-metal salts, is recommended as a polymerization catalyst for vinyl compounds [112]. It is useful as a dyeing auxiliary because of the formation of $NaHCO_3$ during hydrolysis [113]. It also improves absorption of disperse dye systems on polyester and cellulosic fibers [114]. Moreover, in diazo papers developed by heat, sodium trichloroacetate is used to liberate base at 100 – 200 °C [115].

Most important, however, is its application as a selective herbicide to control monocotyledonous annual or perennial weeds [116–118]. As a soil-applied herbicide its half-life is 1 – 2 months [119]. For use as a crop protection agent, a minimum content of 95 % sodium trichloroacetate is required. Because the salt in powder form causes severe irritation, the product is usually marketed as granules [120].

5. Environmental Protection

The waste gases formed during chlorination of acetic acid depend on the degree of chlorination. When purification is done by fractional crystallization, less than 1 % of unreacted acetic acid is necessary in the crude acid. In this case chlorine occurs in these waste gases; recovery of the chlorine has been described [121].

Formation of chlorinated hydrocarbon impurities also depends on the production method used. They form as a result of decarboxylation and occur particularly in the manufacture of trichloroacetic acid.

In the trichloroethylene method and in the manufacture of dichloroacetic acid by catalytic oxidation of trichloroethylene, residues of the starting materials are always present in the waste gases. Such waste gases are purified by adsorption on activated charcoal or are incinerated.

Moreover, the high sulfate salt load can be a wastewater problem in the trichloroethylene method. Chlorinated acetic acids are biodegradable by activated sludge floc, but they can cause serious interference particularly in community wastewater treatment plants that have not been bioprepared. Preventing effluent contaminated with products of acetic acid chlorination from draining into public sewer systems is essential.

Spills of chlorinated acetic acids and their derivatives, resulting, for instance, from transport accidents, should be contained with inert materials (sand or earth) and collected for safe disposal, or neutralized with agents, such as Na_2CO_3 or $NaHCO_3$. Any contaminated soil must be removed (the chlorinated acetic acids are strong pollutants to groundwater). Waste material should be taken to an approved disposal facility (special landfill disposal, incinerator, or wastewater treatment plant).

Chloroacetic acid is flammable at higher temperature and forms explosive mixtures with air (lower explosion limit 8 vol%). The two other acids are nonflammable. If the acids are involved in a fire, the decomposition gases of trichloroacetic acid and dichloroacetic acid may contain phosgene, whereas chloroacetic acid gases are free of phosgene.

6. Chemical Analysis

Rapid analysis of the purity of mono-, di-, and trichloroacetic acid is adequately provided by determining the melting and freezing points, particularly when the melting depression is caused by known contamination.

Quantitative analysis of chloroacetic acid should include total acidity and a separate determination of dichloroacetic acid, sulfuric acid, hydrochloric acid, water, and heavy metals, as well as total chlorine content by approved methods. From these determinations acetic acid and chloroacetic acid content can be calculated.

Dichloroacetic acid can be determined with a high degree of accuracy by saponification with strong base and subsequent manganometric titration. Dichloroacetic acid values are exaggerated in the presence of trichloroacetic acid. If trichloroacetic acid is present, preliminary saponification with dilute base is advisable.

Acetic acid and dichloroacetic acid can be directly separated and quantitatively determined by gas chromatography. In addition to these two acids, trichloroacetic acid can also be detected in the form of its methyl ester by gas chromatography. Dichloroacetic acid and acetic acid in the presence of chloroacetic acid can also be determined accurately by HPLC [122]. This method is not recommended for chloroacetic acid itself because titration to determine total chlorine content is more accurate.

Trichloroacetic acid is determined by hydrolysis in acid solution to chloroform and quantitative determination of the hydrogen ions needed for saponification. Formation of a deep-red complex from pyridine and chloroform can be used for a colorimetric method [123]. Nitron [1,4-diphenyl-3-(phenylamino)-1*H*-1,2,4-triazolium hydroxide] produces a salt with trichloroacetic acid, which is suitable for detecting the acid [124].

Dichloroacetic acid can be determined by polarography in mixtures with trichloroacetic acid [125]. Large amounts of chloroacetic acid interfere with the determination.

Determination of heavy metals in all chlorinated acetic acids and derivatives is carried out by familiar methods, such as atomic absorption or colorimetry. For lead, inverse voltametry is also used.

7. Containment Materials, Storage, and Transportation

Because of the corrosiveness of these chemicals, most metals are unsuitable for use as *container materials.* Thus, chlorination is carried out in ceramic-lined, lead-coated steel containers or in glass-lined vessels. Pipelines are made of glass-lined steel or steel lined with polytetrafluoroethylene (PTFE) or perfluoroalkoxy (PFA) polymers (e.g., tetrafluoroethylene). Silver that has been rendered inert by a silver chloride layer can be used for parts of the equipment (e.g., valves, heat exchangers, distillation column linings, stirrers). Tantalum and titanium (with certain temperature limits) are other suitable metallic materials. Carbon (Diabon) and glass are effective materials for heat exchangers. In addition to PTFE and PFA/TFA, fluorinated ethylene – propylene copolymers (FEP), poly(vinyl chloride), polyethylene, and polypropylene can be used at relatively low temperature and for aqueous solutions.

The aqueous solution can be handled in special stainless steel or rubber-lined steel containers at temperatures below 40 °C. Iron with a baked phenolic resin coating can be used below 100 °C. In the trichloroethylene method, steel (cast iron) is an adequate material in many parts of the process.

During *storage and transportation,* high corrosivity and, with chloroacetic acid, toxicity must also be considered. Chlorinated acetic acids and their derivatives are classified as dangerous substances. They are subject to nearly all national and international handling regulations. Containers must be labeled in the stipulated manner. In the case of chloroacetic acid, the most important forms that are shipped are the melt (in glass-lined railroad cars or titanium-lined tank trucks) and the 80 % aqueous solution (in stainless steel or rubber-lined tank trucks and containers).

Crystallized, flaked chloroacetic acid and trichloroacetic acid are obtained by utilizing chilled rolling flakers and, more recently, with chilled conveyor-type flakers; the coarser flakes manufactured in this manner tend to agglomerate less and, therefore, are preferred. Flakes are packed in multilayer paper sacks or in polyethylene bags contained in fiberboard or iron drums. Small quantities are supplied in polyethylene containers or in glass bottles. Dichloroacetic acid and most of its liquid derivatives usually are shipped in iron drums with polyethylene inner containers.

If stored in a cool, dry place, the chlorinated acetic acids and their derivatives are stable almost indefinitely, although the crystallized products have a marked tendency to agglomerate.

8. Economic Aspects

The worldwide manufacturing capacity for chloroacetic acid is estimated at 420 000 t/a, divided as follows:

Western Europe Federal Republic of Germany (ca. 23 %), Finland, France, United Kingdom, Netherlands, Sweden, Spain	ca. 50 %
Eastern bloc	ca. 22 %
Asia	ca. 12 %
United States	ca. 11 %
South America, South Africa	ca. 5 %

Because French manufacturers have switched to the acetic acid method, 80 – 85 % of worldwide production is now manufactured by the chlorination method. One manufacturer uses 1,1-dichloroethene/chloroacetyl chloride as the starting material, whereas the remainder are probably still using the older trichloroethylene process.

Recognized manufacturers are AKZO, Atochem, Billerud/Uddeholm, Courtaulds Acetate/UK, Cros, Dow, Hercules, Hoechst, and Mätselliito.

Worldwide production cannot be determined precisely because no statistics have been published. Worldwide consumption was estimated

to be 290 000 t in 1983. The main consuming region is Western Europe (ca. 125 000 t in 1983), followed by the Eastern bloc, the United States, and Japan. An increase is expected in some areas in the next few years.

Accurate figures are not available on the amounts of chloroacetic acid and its derivatives used for different applications. However, by far the largest amount goes into the production of carboxymethyl cellulose, starch, and polysaccharides. This is followed by herbicides based on phenoxyacetic acids (2,4-D, MCPA, 2,4,5-T) and thioglycolic acid (ca. 70 % of which is used as esters in industry and about 30 % as salts in cosmetics). All the other applications (glycolic acid, glycine, vitamins, and derivatives of chloroacetic acid for organic syntheses) account for probably not more than 10 % of annual production.

Dichloroacetic acid is of little economic importance. More significant are its acid chloride and methyl ester, which are used as intermediates in the manufacture of agrochemicals and chloramphenicol.

In the case of trichloroacetic acid, only the sodium salt is important. About 21 000–23 000 t/a are used worldwide as a selective herbicide.

9. Toxicology and Occupational Health

Chloroacetic Acid and Chloroacetate. Chloroacetic acid is a corrosive solid material. The lowest published LD_{50} is 76 mg/kg for oral administration to rats [126]. Single oral administration of lethal doses resulted in apathy and loss of body mass. Death occurred several days after administration [127].

Acute inhalational exposure of rats led to eye and pulmonary irritation. The irritation threshold is 23.7 mg/m^3. The LC_{50} is 180 mg/m^3 [128].

Prolonged inhalation of chloroacetic acid resulted in reduced body mass, reduced oxygen consumption, hemoglobinemia, and inflammation of the respiratory tract. A threshold limit value for chloroacetic acid of 1 mg/m^3 is recommended if these experimental animal findings are taken as a basis [128]. However, the ACGIH has not established a TLV; neither has an MAK been set.

Repeated intake of 0.1 % chloroacetate in the diet led to depressed growth curves and an increased level of liver glycogen [129]. A major metabolic pathway for chloroacetic acid is reaction with glutathion, finally yielding thiodiacetic acid [130]. One report has shown that chloroacetic acid elevates mutant rates in the mouse lymphoma forward mutation assay [131].

Dichloroacetic Acid and Dichloroacetate. Dichloroacetic acid, a clear colorless liquid, is corrosive to the skin and mucous membranes. Its oral LD_{50} is 2820 mg/kg for rats, which is of the same order of magnitude as that for trichloroacetic acid. Dichloroacetic acid is able to penetrate rabbit skin with a dermal LD_{50} of 797 mg/kg [126]. The lowest oral LD_{50} reported for dichloroacetate is 4480 mg/kg for fasted rats [127]. Sodium dichloroacetate was examined in a 3-month oral toxicity study in rat and dogs. Both species manifested toxic effects, such as depressed appetites and increased body mass, degenerative changes of testicular germinal epithelium, hind limb muscular weakness or paralysis, and brain lesions. Cessation of administration brought about an amelioration or recovery from all effects except brain lesions. Dogs proved to be more sensitive than rats, showing toxic effects at 50 mg kg^{-1} d^{-1}. The CNS observations are extremely relevant in light of a report that described polyneuropathy of a 21-year-old man whose severe hypercholesterolemia had been treated orally with dichloroacetate (50 mg/kg) for 16 consecutive weeks. The polyneuropathy was characterized by weakness of facial, finger, and lower extremity muscles, by diminished deep tendon reflexes, and by slowing of nerve conduction velocity. The neuropathy improved when the therapy was stopped [132].

Dichloroacetate demonstrated low-grade mutagenicity in the Ames Salmonella/mammalian microsome mutagenicity test [133].

Trichloroacetic Acid and Trichloroacetate. Trichloroacetic acid is a strong protein precipitant and is corrosive to the skin and eye. The solid material is not readily absorbed through the skin [133]. The lowest reported oral LD_{50} in rats for trichloroacetic acid was 3320 mg/kg

and 5060 mg/kg for trichloroacetate. The lowest lethal dose of trichloroacetate for female dogs was 1590 mg/kg with toxic effects, such as salivation, emesis, sedation, ataxia, and coma [134], [135]. A 4-h exposure to 4800 mg/kg trichloroacetate in ambient air did not result in any clinical symptoms or macroscopically detectable pathologic lesions in rats, guinea pigs, rabbits, and cats [135].

Medical reports of acute exposure effects of trichloroacetic acid showed mild to moderate skin and eye burns [136]. In humans, trichloroacetate was an irritant to skin and mucous membranes. Opaqueness of the cornea may occur. Oral ingestion led to intestinal colics, salivation, emesis, diarrhea, muscular weakness, anorexia, and apathy [135].

Little relevant chronic toxicity data exists for trichloroacetic acid. The toxicological profile of airborne trichloroacetic acid is assumed to be dominated by lesions of mucous membranes of the upper respiratory tract. A TLV-TWA of 1 ppm, 7 mg/m^3 is suggested, based largely on analogy to 2,2-dichloropropionic acid (TLV-TWA 1 ppm, 6 mg/m^3) [136].

Trichloroacetic acid strongly binds to plasma proteins. High concentration in blood is capable of displacing drugs from protein-binding sites. Because of this systemic effect of potential drug displacement from plasma proteins, a TLV of 1 – 2 ppm may be accepted for trichloroacetic acid [134].

For trichloroacetate, a provisionally tolerated daily intake of 0.075 mg/kg body mass has been established. This limit value takes into account the results of a subchronic dog feeding study with concentrations in food of $\geq$ 2000 mg/kg, resulting in loss of body mass, malaise, necrosis of gingival and buccal mucosae, changes in white blood count, pathologic urine findings, lesions of the liver and myocardium, atrophy of skeletal muscle, and impairment of spermiogenesis. The nontoxic effect level was 500 mg/kg in the diet [135].

On the basis of available in vitro data, trichloroacetate does not exert mutagenic effects [135].

10. References

1. *Beilstein,* 4th ed., **1,** 194 – 197; **2(1),** 87 – 88; **2(2),** 187 – 191; **2(3),** 428 – 438; **2(4),** 474 – 480.
2. "Azeotropic Data 3" in R. F. Gould (ed.): *Advances in Chem., Series 116,* Amer. Chem. Soc., Washington, D.C., 1973, pp. 97 – 99.
3. Du Pont, US 2 455 405, 1943 (L. S. Burrows, M. F. Fuller).
4. SU 110 145, 1958 (B. G. Yasnitskii, A. P. Zaĭtzev); *Chem. Abstr.* **52** (1958) 17 110 g.
5. SU 119 875, 1959 (B. G. Yasnitskiĭ, A. P. Zaĭtzev); *Chem. Abstr.* **54** (1960) 2178 a.
6. Rhône-Poulenc, FR 1 423 671, 1964 (P. Mounier).
7. Rhône-Poulenc, FR 1 413 926, 1964 (G. Boullay).
8. SU 1 004 346, 1981 (J. N. Norikov, V. J. Kondratenko, G. S. Cibulskaja); *Chem. Abstr.* **99** (1983) 5218 f.
9. Gulf Research, US 3 627 826, 1970 (C. M. Selwitz).
10. Chevron Research, US 4 221 921, 1976 (S. Suzuki).
11. Mitsubishi Gas, JP-Kokai 7 844 520, 1976 (T. Isshiki, W. Yoshino, T. Kato).
12. Plantex, JL 41 782, 1973 (S. Roiter, Y. Fein).
13. British Intelligence Objectives Subcommittee (BIOS), Final Report no. 929, Item no. 22, London 1947, pp. 1 – 4, 7 – 9.
14. British Intelligence Objectives Subcommittee (BIOS), Final Report no. 1154, Item no. 22, London 1947, pp. 37, 38.
15. Field Information Agency Technical (FIAT), Report no. 1051, London 1947.
16. Hooker Electrochem. Co., US 2 595 899, 1949 (J. A. Sonia, C. E. Lisman).
17. BASF, DE 1 224 298, 1960 (H. Schlecht, H. Albers, R. Oster).
18. H. Brückner, *Angew. Chem.* **40** (1927) 973 – 974; **41** (1928) 226 – 229.
19. Akzo, DE 2 163 849, 1971 (H. D. Rupp, G. Meyer, H. Zengel).
20. Iron Enterprise Co., JP 528, 1959 (H. Fukuda).
21. Solvay Cie., GB 759 450, 1956 (J. Viriot).
22. W. Richarz, A. Mathéy, *Chimia* **21** (1967) 388 – 395.
23. CS 113 592, 1962 (J. Hrabovecky, S. Regula, J. Kozacek); *Chem. Abstr.* **63** (1965) 17 908 b.
24. Comp. Espanola de Petroleos, ES 464 673, 1977 (C. R. Huertas, A. Y. Chinchilla, J. M. Lopez Bahamonde).
25. Politechnika Lodzka, PL 63 908, 1968 (A. Kowalski); *Chem. Abstr.* **76** (1972) 139 941 v.

26. Celanese Corp., US 2 503 334, 1950 (A. R. Hammond, J. A. John, R. Page).
27. Monsanto, US 2 539 238, 1949 (C. M. Eaker).
28. Dow Chemical, US 2 826 610, 1955 (E. K. Morris).
29. Uddeholms, GB 928 178, 1960.
30. Hoechst, DE 1 919 476, 1969 (W. Opitz, A. Jacobowsky, W. Burkhard).
31. Fondazione de Nora, US 4 281 184, 1978 (P. Spaziante, C. Sioli, L. Giuffre).
32. Daicel, JP 81 127 329, 1980.
33. GB 1 176 109, 1968 (A. L. Englin, E. V. Sergeev, E. R. Berlin); *Chem. Abstr.* **72** (1970) 110 815 p.
34. Dow Chemical, US 3 152 174, 1960 (E. K. Morris, W. W. Bakke).
35. Frontier Chemical Co., US 2 917 542, 1958 (J. I. Jordan).
36. Shawinigan Chemicals, US 2 809 993, 1955 (G. M. Glavin, H. B. Stevens).
37. New York-Ohio-Chemical, US 2 688 634, 1950 (J. T. Pinkston).
38. Monsanto, US 3 576 860, 1967 (D. A. Zazaris).
39. Dow Chemical, US 3 772 157, 1971 (L. H. Horsley).
40. Mitsui Toatsu Chemical, JP 7 230 165, 1970 (R. Fujiwara, H. Ohira).
41. Mitsui Toatsu Chemical, JP 7 229 886, 1970 (R. Fujiwara, H. Ohira).
42. Knapsack, DE 1 919 751, 1969 (H. Baader, G. Berger, A. Jacobowsky).
43. Knapsack-Griesheim, DE 1 033 653, 1958 (R. Wesselmann).
44. New York-Ohio Chemical, US 2 809 214, 1957 (I. N. Haimsohn).
45. SU 374 278, 1968 (E. V. Sergeev, V. N. Egorova, R. P. Milyukova); *Chem. Abstr.* **79** (1973) 41 946 u.
46. Tsukishima Kikai Co., EP 32 816, 1981 (R. Sugamiya, K. Nakamaru, K. Takegami).
47. Comp. Espanola de Petroleos, ES 468 685, 1978 (C. R. Huertas).
48. Dow Chemical, US 3 365 493, 1964 (D. D. Deline, F. B. Wortman, A. A. Holzschuh).
49. Uddeholms, DE 1 268 130, 1960 (S. H. Persson, P. T. Akerström, P. E. Larberg).
50. Knapsack-Griesheim, DE 910 778, 1951 (K. Sennewald, A. Wolfram).
51. Hooker Chemical Co., US 2 863 917, 1954 (J. T. Rucker, J. S. Scone).
52. Knapsack-Griesheim, DE 1 201 326, 1963 (W. Opitz).
53. Hoechst, DE 1 668 023, 1967 (W. Freyer, M. Gscheidmeier, R. Höltermann).
54. Knapsack-Griesheim, DE 1 072 980, 1958 (W. Opitz, K. Sennewald).
55. N. V. Koninklijke Nederlandsche Zoutindustrie, NL 109 768, 1962 (G. van Messel).
56. N. V. Koninklijke Nederlandsche Zoutindustrie, NL 109 769, 1962 (A. Blaauw, G. van Messel).
57. Knapsack, DE 1 816 931, 1968 (K. Sennewald, A. Ohorodnik, W. Mittler).
58. Hoechst, DE 1 915 037, 1969 (K. Sennewald, J. Hundeck, A. Ohorodnik).
59. Hoechst, DE 2 240 466, 1972 (A. Ohorodnik, K. Gehrmann, J. Hundeck).
60. Hoechst, DE 2 323 777, 1973 (M. Gscheidmeier).
61. Knapsack, DE 1 920 805, 1969 (H. Baader, W. Opitz, A. Jacobowsky).
62. Knapsack, DE 1 920 806, 1969 (K. Sennewald, J. Hundeck, W. Opitz).
63. Hoechst, DE 871 890, 1942 (V. Hilcken, H. Petri). Hoechst, DE 860 354, 1942 (H. Petri, P. Landmann).
64. Hercules Powder, US 2 446 233, 1946 (B. T. Lamborn, N. J. Matawan).
65. Svenska Cellulosa, GB 782 479, 1955.
66. Hoechst, DE 2 432 567, 1974 (H. Scholz, G. Koenig).
67. Merck, DE 2 120 194, 1971 (H. Hornhardt).
68. Pechiney St. Gobain, DE 2 059 597, 1970 (J. C. Strini).
69. Knapsack, DE 1 804 436, 1968 (W. Opitz).
70. Montecatini, US 2 843 633, 1954 (G. Natta, E. Beati).
71. Monsanto, US 2 262 686, 1940 (L. P. Kyrides, W. Groves, F. B. Zienty).
72. *Bull. Soc. Chim. Fr. Mem.* **5** (1939) no. 6, 1616 – 1625. L. Givaudan & Cie., FR 854 038, 1938.
73. H. Aspelund, *Soum. Kemistiseuran Tied.* **49** (1940) 49 – 63.
74. Dow Chemical, US 2 321 278, 1941 (E. C. Britton, W. R. Shawver).
75. A. E. Kretov, A. P. Momsenko, SU 141 150, 1961.
76. Phillips Petroleum Co., US 2 815 323, 1956 (R. P. Schneider).
77. Dow Chemical, US 2 758 103, 1952 (W. A. Henson, W. M. Westveer).
78. M. Schlecker, DE 1 492 569, 1964 (M. Schlecker).
79. Leek Chemicals, GB 859 714, 1959 (F. N. Morris).
80. General Electric Co., US 2 328 424, 1940 (G. F. D'Alelio).

81. Ciba, CH 496 843, 1967 (A. Hiestand, A. Therwil).
82. Comp. des Prod. Chim. et Electrometallurgiques Alais, DE 610 317, 1933.
83. Shell, US 2 051 470, 1934 (M. de Simo, C. C. Allen).
84. Dow Chemical, US 2 495 440, 1946 (E. C. Britton, L. F. Berhenke).
85. Hoechst, DE 1 085 518, 1957 (W. Protzer).
86. Consortium f. Elektrochem. Ind., DE 340 872, 1919; DE 391 674, 1921; DE 531 579, 1929 (M. Mugdan, J. Wimmer).
87. Shell Int. Research, NL 6 606 933, 1965 (G. W. Gaertner, D. E. Ramey).
88. Shell Int. Research, DE 1 793 446, 1968 (R. K. June).
89. Hoechst, DE 1 568 547, 1966 (K. Petz).
90. C. A. Frank, A. T. Hallowell, C. W. Theobald, *Ind. Eng. Chem.* **41** (1949) 2061 – 2062.
91. F. B. Erickson, E. J. Prill, *J. Org. Chem.* **23** (1958) 141 – 143. Monsanto, US 2 889 365, 1957 (E. J. Prill). Distillers Co., US 2 862 964, 1956 (R. M. Lacey).
92. G. Darzens, *C. R. Hebd. Acad. Sci.* **203** (1936) 1374 – 1376.
93. Dow Chemical, US 1 757 100, 1924 (C. J. Strosacker).
94. Lech-Chemie Gersthofen, DE 860 211, 1944 (P. Heisel).
95. Hoechst, DE 1 020 619, 1957 (H. Kolb, T. Sievers, H. Hoyer).
96. Bozel-Maltera, FR 1 156 426, 1956 (F. Peto, L. Marcheguet, M. Girard).
97. Amer. Cyanamid, US 2 382 803, 1943 (A. A. Miller).
98. Monsanto, US 2 613 220, 1950 (C. M. Eaker).
99. Knapsack-Griesheim, DE 1 031 778, 1954 (K. Sennewald, A. Wolfram).
100. European Pharmacopeia, vol. 3.
101. E. Sidi, G. E. Goetschel, F. Casalis, *Presse Med.* **32** (1950) 570.
102. L. Rosenthaler, *Parfuem. Kosmet.* **36** (1955) 271.
103. V. P. Rudav'skii, *Khim. Promst. Kiev* 1965, 22 – 23.
104. E. Müller, K. Ehrmann, *Ber. Dtsch. Chem. Ges.* **69** (1936) 2207 – 2210.
105. Du Pont, US 2 378 048, 1944 (C. W. Theobald).
106. Uddeholms, DE 2 118 540, 1971 (C. Östlund, J. Dahlberg, A. Moden). SU 176 286, 1964 (V. A. Poluektov).
107. G. R. Raw (ed.): *CIPAC-Handbook: Analysis of Technical and Formulated Pesticides,* **vol. 1,** W. Heffer & Sons, Cambridge, England, 1970, p. 691.
108. E. V. Dehmlow, *Tetrahedron Lett.* 1976, no. 2, 91 – 94.
109. E. V. Dehmlow, K. H. Franke, *Z. Naturforsch.* **336** (1978) 686 – 687.
110. Nobel Bozel, FR 1 215 896, 1959 (M. L.-J. Girard).
111. Monsanto, US 2 643 220, 1950 (J. Loumiet-Lauigne).
112. Dynamit Nobel, DE 1 203 957, 1962 (W. Trautvetter).
113. Ciba, BE 646 398, 1964 (F. Raff).
114. Ciba, GB 1 035 340, 1964 (F. Raff).
115. General Anilin & Film, BE 672 553, 1966 (W. J. Welch, B. E. Tripp).
116. G. M. Tsukermann, *Best. Sel'sk. Nauk. Min. Sel'sk. Khoz. Kaz. SSR* **8** (1965) 117 – 121.
117. H. Ansorge, U. Koss, *Albrecht-Thaer-Arch.* **7** (1963) 889 – 899.
118. *Z. Pflanzenkr. Pflanzenschutz (Sonderheft)* **1** (1964) 49 – 53.
119. H. Neururer, *Pflanzenschutzberichte* **28** (1962) 145 – 180.
120. Hoechst, DE 1 166 761, 1962 (H. Kolb, R. Hartmann, H. Buckmiller).
121. Hoechst, DE 3 246 953, 1984 (G. Koenig, J. Maginot).
122. N. E. Skelly, *Anal. Chem.* **54** (1982) no. 4, 712 – 715.
123. *Chem. Ztg.* **77** (1953) 585.
124. W. L. Dulière, *Bull. Soc. Chim. Biol.* **34** (1952) 991.
125. P. J. Elving, Chang-Slang Tang, *Anal. Chem.* **23** (1950) 34 – 43.
126. R. J. Lewis, S., R. L. Tatken (eds.): *Registry of Toxic Effects of Chemical Substances,* NIOSH (RTECS Online Data Base), 1984.
127. G. Woodward et al., *J. Ind. Hyg. Toxicol.* **23** (1941) 78 – 81.
128. G. G. Maksimov, O. N. Dubinina, *Gig. Tr. Prof. Zabol.* 1974, no. 9, 32 – 35.
129. S. Dalgaard-Mikkelson et al., *Acta Pharmacol. Toxicol.* **11** (1955) 13.
130. S. Yllner, *Acta Phamacol. Toxicol.* **30** (1971) 69 – 80.
131. D. E. Amacher, G. N. Turner, *Mutat. Res.* **97** (1982) 49 – 65.
132. R. Katz et al., *Toxicol. Appl. Pharmacol.* **57** (1981) 273 – 287.
133. V. Herbert et al., *Am. J. Clin. Nutr.* **33** (1980) no. 6, 1179 – 1182.

134. D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-Arbeitsmedizinische Begründung von MAK-Werten,* Verlag Chemie, Weinheim 1981, Part 8.
135. DFG Pflanzenschutzkommission, *Toxikologie der Herbizide,* Verlag Chemie, Weinheim 1982, Part 4.
136. ACGIH (ed.): *Threshold Limit Values (TLV)* , Cincinnati, Ohio, 1985 – 1986.

Chloroamines

YASUKAZU URA, Central Research Institute, Nissan Chemical Ind., Ltd., Chiba, Japan

GOZYO SAKATA, Central Research Institute, Nissan Chemical Ind., Ltd., Chiba, Japan

1. **Introduction** 1
2. **Chemical Properties** 1
3. **Reactions** 2
4. **Inorganic *N*-Chloroamines** 2
5. ***N*-Chloroisocyanuric Acids** 3
5.1. **Production and Chemical Properties** . 3
5.2. **Uses** 3
6. **Other Organic *N*-Chloroamines** 4
7. **Toxicology** 5
8. **References** 6

1. Introduction

Chloroamines are inorganic or organic nitrogen compounds that contain one or more chlorine atoms attached to a nitrogen atom. A familiar example is monochloroamine (NH_2Cl), which has been known since the beginning of the 19th century.

Chloroamines are used in synthetic reactions and as bleaching agents, disinfectants, and bactericides because they function as chlorinating agents and oxidants. Since chloroamines are safer and easier to handle than chlorine gas or metal hypochlorites, they are widely used in the purification of drinking water and as sanitizing agents in swimming pools. *N*-Chloroisocyanuric acids and 1-bromo-3-chloro-5,5-dimethylhydantoin have become important in the sanitizing and disinfectant markets, replacing calcium hypochlorite, 1,3-dichloro-5,5-dimethylhydantoin, and chloramine-T.

2. Chemical Properties

N-Chloroamines can act as oxidizing agents by absorbing two electrons:

$$Cl^+ + 2e^- \longrightarrow Cl^-$$

Thus, an *N*-chloroamine oxidizes hydriodic acid to liberate iodine, and this reaction is used for the quantitative analysis of *N*-chloroamines:

$$RR'NCl + 2HI \longrightarrow HCl + RR'NH + I_2$$

The theoretical *available chlorine content* of an *N*-chloroamine is expressed as twice the mass fraction of chlorine in the molecule. However, for practical purposes, the value is expressed as the equivalent of elemental chlorine as determined by the oxidizing ability of the N-chloroamine.

The quantitative hydrolysis constant (*K* value) of an *N*-chloroamine is expressed by the equation below, and is generally in the range $10^{-4} - 10^{-10}$ (Table 1).

$$RR'NCl + H_2O \rightleftarrows RR'NH + HOCl$$

$$K = \frac{c_{RR'NH} \cdot c_{HOCl}}{c_{RR'NCl}}$$

Table 1. *K* values of important *N*-chloroamines

N-Chloroamine	CAS registry no.	*K* value
Trichloroisocyanuric acid	[*87-90-1*]	6.7×10^{-4}
1,3-Dichloro-5,5-dimethylhydantoin	[*118-52-5*]	2.5×10^{-4}
N-Chlorosuccinimide	[*128-09-6*]	6.6×10^{-5}
Dichloramine-T	[*473-34-7*]	8.0×10^{-7}
Chloramine-T	[*127-65-1*]	4.9×10^{-8}
Monochloroamine	[*10599-90-3*]	2.8×10^{-10}

The *K* value is used to express the bactericidal power of *N*-chloroamines, which depends on their generating hypochlorous acid in water.

The N – Cl bonds of *N*-chloroamines are covalent and are readily hydrolyzed by water with release of hypochlorous acid. They are thermally unstable; many do not melt congruently and can explode at elevated temperature. Trichloroamine (nitrogen trichloride, NCl_3) is particularly unstable and can cause explosions when present as an impurity in other *N*-chloroamines. *N*-Chloroamines should be stored cold and pro-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_553

tected from light, water, amines and ammonium compounds, strong acids and bases, and easily oxidized organic materials.

3. Reactions

Many *N*-chloroamines are useful as reagents or intermediates in organic synthesis. For example, *N*-chloroamides of organic carboxylic acids can be reduced by alkali to the corresponding amines in good yield. This reaction, called the Hofmann degradation, is widely used for the synthesis of aromatic, heterocyclic, and alicyclic amines [1–5]:

$$RCONH_2 \xrightarrow{Cl_2} RCONHCl \xrightarrow{OH^-} RNH_2$$

N-Chloroureas can be converted into hydrazines by treatment with base in an alcohol or water solution [6], [7]:

$$RNHCONHCl \xrightarrow[OH^-]{R'OH} RNHNHCO_2R' \text{ or } RNHNH_2$$

N,N-Dichlorocarbamates, Cl_2NCOOR, are useful chlorinating agents and can also be added to dienes [8], [9]:

$$RO_2CNCl_2 + CH_2{=}CH{-}CH{=}CH_2 \longrightarrow RO_2CNClCH_2CH{=}CHCH_2Cl$$

They are also the starting materials for the synthesis of *N*-halo-*N*-metallocarbamidates, which are intermediates in the synthesis of carbamate derivatives of physiologically active substances [10–14]:

$$C_2H_5O_2CNCl_2 \xrightarrow{C_2H_5O_2CNH_2} C_2H_5O_2CNHCl \xrightarrow[\text{metal}]{\text{Alkali}} C_2H_5O_2CNNaCl$$

8-Azaisatoic anhydride (**3**), an intermediate to several agrichemicals, is synthesized in high yield by chlorinating 2,3-pyridinedicarboximide (**1**) to *N*-chloro-2,3-pyridinedicarboximide (**2**), followed by treatment with alkali [15], [16]:

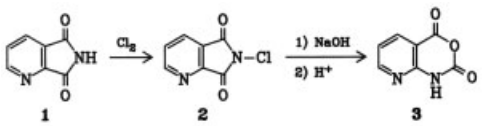

N-Chloroamidines and *N*-chloroguanidines, which have excellent bactericidal and fungicidal activity, are widely used as intermediates in the syntheses of heterocyclic ring systems such as imidazoles, thiazoles, and oxadiazoles [17–21].

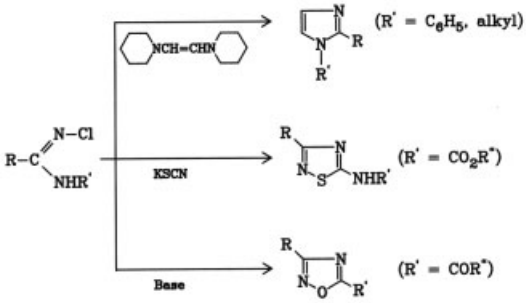

where R″ = alkyl

4. Inorganic *N*-Chloroamines

Monochloroamine [*10599-90-3*], NH_2Cl, M_r 51.47, *mp* −66 °C, is a colorless liquid with a strong pungent odor. It is more stable in aqueous solution than dichloroamine [*3400-09-7*], $NHCl_2$, M_r 85.91, or trichloroamine [*10025-85-1*], NCl_3, M_r 120.36, *mp* −92 °C, *bp* 70 °C, and is mainly used for primary purification of drinking water. It is also used in organic synthesis as an oxidizing agent for trisubstituted phosphines [22]. All of these *N*-chloroamines are prepared in pH-controlled reactions by the action of hypochlorous acid or chlorine on ammonium salts [23], [24].

$$NH_3 + x\,HOCl \longrightarrow NH_{3-x}Cl_x + x\,H_2O \qquad x = 1, 2, 3$$
$$NH_3 + 3\,Cl_2 \longrightarrow NCl_3 + 3\,HCl$$

Extreme care should be taken when any of these *N*-chloroamines in pure form are isolated, because they are unstable and explosive in the pure state, even at room temperature.

N-Chlorosulfamic acid (**4**) [*17172-27-9*], M_r 131.55, *N,N*-dichlorosulfamic acid (**5**) [*17085-87-9*], M_r 166.00, and sodium *N*-chloroimidodisulfonate (**6**) [*67700-32-7*], M_r 255.56, are prepared from sulfamic acid and hypochlorous acid. They are widely used as disinfectants and bleaching agents for paper [25].

$ClNHSO_3H$	Cl_2NSO_3H	$ClN(SO_3Na)_2$
4	**5**	**6**

The activities of fabric-bleaching compositions based on **6** are described in detail in [26], [27].

5. *N*-Chloroisocyanuric Acids

N-Chloroisocyanuric acids (chloroisocyanurates), such as 1,3,5-trichloro-1,3,5-triazine-2,4,6(1*H*,3*H*,5*H*)trione [*87-90-1*], trichloroisocyanuric acid, Symclosene, TCC (**7**), and 1,3-dichloro-1,3,5-triazine-2,4,6(1*H*,3*H*,5*H*)trione sodium salt [*2893-78-9*], sodium dichloroisocyanurate, DCC-Na (**8**), have become increasingly important as disinfectants because they are more stable and easier to handle than metal hypochlorites. Table 2 gives physical properties of TCC and DCC-Na.

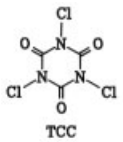

TCC

7

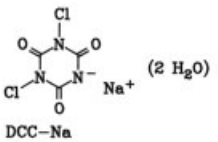

DCC–Na

(DCC–Na · 2 H_2O[*51580–86–0*])

8

Table 2. Physical constants of TCC and DCC-Na [28]

Property	TCC [*87-90-1*]	DCC-Na [*2893-78-9*]
Physical form	White powder	White powder
Formula	$(ClNCO)_3$	$Cl_2Na(NCO)_3$
M_r	232.44	219.98
mp	234 °C (decomp.)	225 °C (decomp.)
Available chlorine		
Theoretical (%)	91.5	64.5
Typical value	90.0	62.0
pH (1 % aqueous solution)	2.7 – 3.3	6.2 – 6.8
Solubility (g/100 g)		
Water (25 °C)	1.0	30.0
Acetone (30 °C)	35.0	0.5

In particular, TCC and DCC-Na have acquired a substantial share of the market traditionally served by calcium hypochlorite, pine oils, quaternary ammonium salts, and phenols.

5.1. Production and Chemical Properties

N-Chloroisocyanuric acids are prepared by continuous reaction of chlorine with isocyanuric acid in aqueous sodium hydroxide at 0 – 15 °C. Careful control of pH and reaction temperature are essential to prevent the formation of explosive NCl_3 [29], [30]. Trichloroisocyanuric acid (TCC) is gradually decomposed in an alkaline medium, yielding NCl_3, $NHCl_2$, and NH_2Cl. Both DCC-Na and its dihydrate have high water solubility compared with TCC; the dihydrate is thermally more stable than DCC-Na. The dihydrate of DCC-Na is obtained by cooling a saturated aqueous solution of DCC-Na (45 °C) to 10 °C.

5.2. Uses

The oxidizing and bioactive properties of the *N*-chloroisocyanuric acids derive from the hypochlorous acid that is slowly released from them in water. These compounds are used for many purposes such as nonshrinking treatment of wool, disinfectants for swimming pools, cleaning and sterilizing of bathrooms, and laundry bleach.

Cleaners and sanitizers based on *N*-chloroisocyanuric acids contain the following other components: phosphates, sodium metasilicates, surfactants, and such neutral salts as sodium sulfate and sodium carbonate. They are effective in removing oil and protein on stainless steel; they are also recommended for dishwashing in hotels, hospitals, restaurants, and food factories.

TCC and DCC-Na are sold in the form of granules or tablets. In general, TCC is used to achieve an active chlorine level of 1 – 2 mg/L in swimming pools, and DCC-Na a level of 100 – 300 mg/L in food processing.

Economic Aspects. In 1985, the estimated world production of *N*-chloroisocyanuric acids was 85 000 t. Suppliers include Olin, Monsanto, CDF Chimie, Chlorchem, Delsa, Nissan Chem. Ind., and Shikoku Chem. Ind. The demand for these compounds is increasing by 8 – 10 % per year for swimming pool use and 3 – 5 % per year for food processing.

6. Other Organic *N*-Chloroamines

1,3-Dichloro-5,5-dimethylhydantoin [*118-52-5*], Dactin, Halane (BASF Wyandotte), M_r 197.03, *mp* 132 °C, theoretical chlorine content 77.6 %, water solubility 2.1 g/L, is prepared by passing chlorine through an aqueous solution of 5,5-dimethylhydantoin [31], [32]. It has been used since the 1930s as a bactericide and disinfectant, but its use and market are decreased by the use of other dominant *N*-chloroamines, such as TCC and Di-Halo. 1-Bromo-3-chloro-5,5-dimethylhydantoin (**9**) [*16079-88-2*], M_r 241.48, *mp* 130 °C (decomp.), pH 2.88 (0.1 % solution), water solubility 2 g/L (20 °C), is prepared by successive bromination and chlorination of 5,5-dimethylhydantoin [33], [34]. The compound is a stable white powder, and the formation of chloroamines is not observed in weakly basic aqueous solution.

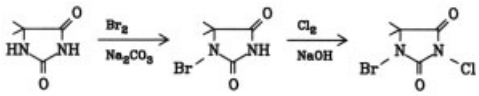

9

It has become widely used at an active halogen level of 0.5 – 3 mg/L for swimming pool disinfectants because of its stability in water and in storage, as well as its broad bactericidal spectrum [35], [36]. It is marketed under the name of Di-Halo (Aquabrom/TESCO; Division of Great Lakes Chemicals). Production in 1985 was estimated at ca. 2500 t, growing at a rate of 10 – 20 % per year. 1-Chloro-2,5-pyrrolidinedione (**10**) [*128-09-6*], *N*-chlorosuccinimide, M_r 133.54, *mp* 150 – 151 °C, is a white solid that is sparingly soluble in ether, chloroform, and carbon tetrachloride. It has 50 – 54 % available chlorine and is used principally as a chlorinating agent in organic synthesis [37].

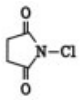

10

N-Chloroglycolurils have been studied as protective agents against poison gas and are prepared by condensation of 1,2-diketones with urea followed by chlorination. An important example is 2,4,6,8-tetrachloro-2,4,6,8-tetrazabicyclo[3.3.0]octane-3,7-dione (**11**) [*776-19-2*], M_r 279.82.

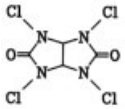

11

1,3,4,6-Tetrachloro-3α,6α-diphenylglycoluril [*51592-06-4*], Iodogen (**12**), M_r 431.94, is used as a bactericide and as an oxidizing agent in syntheses of peptides [38–40].

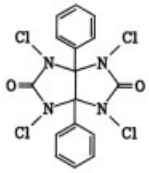

12

3-Chloro-4,4-dimethyl-2-oxazolidinon (**13**) [*58629-01-9*], M_r 149.50, is an example of a new class of *N*-chloroamines that are being investigated as disinfectants [41], [42].

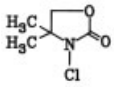

13

Sodium *N*-chloro-*p*-toluenesulfonamide (**14**) [*127-65-1*], chloramine-T, M_r 227.67 (hydrate: *mp* 175 °C; explosive), has an available chlorine content of 25 %. It is prepared in 75 – 95 % yield by passing chlorine into a sodium hydroxide solution of *p*-toluenesulfonamide [43], [44]. It is a strong electrolyte in acid solution and a good oxidizing agent in base. It is fairly soluble in water, and practically insoluble in benzene, chloroform, and ether. The compound reacts readily with mustard gas to yield a harmless crystalline sulfimide; chloramine-T derivatives are being studied as protective agents against poison gas [14].

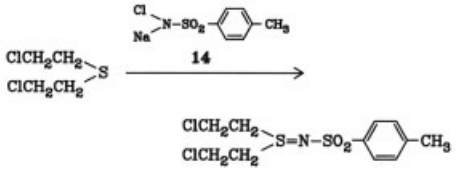

14

Other examples of this group of compounds are *N*-chlorobenzenesulfonamide (**15**) [*127-52-6*], chloramine-B, M_r 213.64 (hydrate: *mp* 180 °C; explosive), and *N*-dichlorobenzenesulfonamide (**16**) [*473-29-0*], dichloramine-B, M_r 226.08, *mp* 76 °C (decomp.).

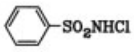

15

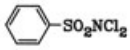

16

N,N′,N″-Trichloromelamine (**17**) [*12379-38-3*], M_r 229.46, *mp* 175 °C (decomp.), theoretical chlorine content 92.8 %, is used as a sterilizing agent to a smaller extent because of its low solubility in water.

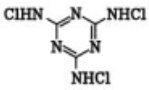

17

N-Chlorinated derivatives of various amino acids are being studied as bleaching agents and bactericides [45–47].

7. Toxicology

The toxicity of *N*-chloroamines is very important because of their widespread use in drinking water, swimming pools, and food processing. When they are used in water that contains organic compounds, special care must be taken to prevent the formation of halomethanes, which are weak carcinogens. The formation of halomethanes in water is measured by the head space method, purge trap method, or solvent extraction, followed by analysis with GC (electron capture detector) or GC – MS [48], [49].

Most *N*-chloroamines irritate the eyes, moist skin, and the upper respiratory tract. Solutions of N-chloroisocyanuric acids containing up to 100 ppm of available chlorine are not toxic, irritating, or sensitizing. However, ingestion of the pure solid or a fresh strong suspension attacks the stomach lining. The isocyanuric acid formed after hydrolysis has low toxicity. Table 3 gives a few acute toxicity values.

Table 3. Acute toxicity of *N*-chloroamines

Compound	LD_{50}, mg/kg (rat, oral)	Reference
TCC	1300	[28]
DCC-Na	1420	[28]
1,3-Dichloro-5,5-dimethylhydantoin	542	[50]
N-Chlorosuccinimide	2700∗	[51]

∗ Minimum lethal dose.

Chloramine-T is a skin irritant; it is highly toxic when absorbed in the blood and should, therefore, be used only in diluted solution. Monochloroamine is widely used as a primary disinfectant for drinking water because of the low rate at which it forms halomethanes from organic compounds in the water. It has a long duration of action, although its specific bactericidal activity is rather low. However, it has been implicated as a mutagen and as a toxic agent for aquatic life, and the U.S. Environmental Protection Agency proposes to prohibit its use in drinking water [52]. 1-Bromo-3-chloro-5,5-dimethylhydantoin is less toxic to test fish species than 1,3-dichloro-5,5-dimethylhydantoin [53].

8. References

1. Kyowa Gas Chem., JP-Kokai 77-53860, 1977 (K. Aida, H. Segawa, K. Nakao).
2. AKZO GmbH, DE 2 313 496, 1974 (H. Zengel, M. Bergfeld).
3. AKZO GmbH, DE 2 710 595, 1978 (H. Zengel, M. Bergfeld).
4. SNIA Viscosa, US 4 198 348, 1980 (C. A. Pauri).
5. Nippon Kayaku Co., JP-Kokai 78-130648, 1978 (R. Hasegawa, S. Kawamoto).
6. Pennwalt Corp., US 3 956 366, 1976 (C. S. Sheppard, L. E. Korczykowski).
7. Pennwalt Corp., US 3 657 324, 1972 (C. S. Sheppard, L. E. Korczykowski).
8. Exxon Research & Eng. Co., US 4 146 557, 1979 (F. A. Daniher, A. A. Oswald).
9. Exxon Research & Eng. Co., US 4 171 449, 1979 (F. A. Daniher, A. A. Oswald).
10. M. M. Campbell, G. Johnson, *J. Chem. Soc. Chem. Commun.* 1975, 479.
11. D. H. Bremner, M. M. Campbell, G. Johnson, *Tetrahedron Lett.* 1975, 2955 – 2958.
12. W. Traube, H. Gockel, *Ber. Dtsch. Chem. Ges.* **56 B** (1923) 384 – 391.

13. P. Chabrier, *C. R. Acad. Sci. Ser. B* **214** (1942) 362.
14. M. M. Campbell, G. Johnson, *Chem. Rev.* **78** (1978) 65 – 79.
15. Sherwin Williams Co., US 3 734 921, 1973 (L. C. Vacek).
16. Sherwin Williams Co., US 3 828 038, 1974 (L. C. Vacek).
17. Ciba Geigy, DE 2 242 772, 1973 (J. C. Petitpierre, C. Weis).
18. Ube Ind. Ltd., JP-Kokai 82-32205, 1982 (Y. Hirota, T. Yorie).
19. Medeau Researches S.r.l., DE 2 812 304, 1978 (R. Stradi).
20. Sankyo Co. Ltd., JP-Kokai 74-18899, 1974 (T. Konotsune, T. Yanai).
21. Sankyo Co. Ltd., JP-Kokai 73-80560, 1973 (T. Konotsune, T. Yanai, M. Suzuki).
22. H. H. Sisler, A. Sarkis, H. S. Ahuja, R. J. Drago et al., *J. Am. Chem. Soc.* **81** (1959) 2982 – 2985.
23. R. G. Laughlin, *Chem. Ztg.* **92** (1968) 383.
24. W. A. Noyes, *J. Am. Chem. Soc.* **42** (1920) 2173 –2179.
25. Du Pont, US 3 767 586, 1973 (F. Rutkiewic, F. Andrew).
26. Monsanto, US 4 201 687, 1980 (M. M. Crutchfield, R. P. Langguth, J. M. Mayer).
27. Monsanto, US 4 233 173, 1980 (J. M. Mayer, J. H. Payne).
28. Nissan Chem. Ind., Technical Information on Nissan Chlorinated Isocyanurates, 1985.
29. E. Nachbaur, W. Gottardi, *Monatsh. Chem.* **97** (1966) 115 – 120.
30. Shikoku Chem. Ind., FR 1 571 705, 1969; *Chem. Abstr.* **72** (1970) 121588.
31. H. Biltz, K. H. Slotta, *J. Prakt. Chem.* **113** (1926) 248; *Chem. Abstr.* **21** (1927) 1794.
32. O. O. Orazi, O. A. Orio, *Chem. Abstr.* **48** (1954) 13634.
33. Drug Research Inc., US 2 779 764, 1957 (L. O. Paterson).
34. Drug Research Inc., US 3 147 259, 1964 (L. O. Paterson).
35. L. O. Paterson, US 3 147 219, 1964.
36. L. M. G. Padilla, GB 1 139 188, 1969.
37. H. Zimmer, L. F. Audrieth, *J. Am. Chem. Soc.* **76** (1954) 3856 – 3857.
38. P. Palitzsch, DD 123 467, 1976.
39. P. R. P. Salacinski, C. McLean, L. E. Sykes, V. V. Clement-Jones, *Anal. Biochem.* **117** (1981) 136 – 146.
40. T. B. Schulz, R. Jorde, H. L. Waldum, P. G. Burhol, *Scand. J. Gastroenterol.* **17** (1982) 379 – 382.
41. Interx Research Corp., US 4 000 293, 1976 (J. J. Kaminski, N. S. Bordor).
42. *Chem. Week* 1985, May 8, 33.
43. M. Bugla, J. Hok, M. Veger, CS 159 409, 1975.
44. Toyo Chem., JP-Kokai 76-4141, 1976 (S. Masutani).
45. Interx Research Corp., US 4 036 843, 1977 (J. J. Kaminski, N. S. Bordor).
46. Interx Research Corp., US 4 045 578, 1977 (J. J. Kaminski, N. S. Bordor).
47. S. H. Gerson, S. D. Worley, N. S. Bordor, J. J. Kaminski, *J. Med. Chem.* **21** (1978) 686 – 688.
48. A. A. Nicholson, O. Meresz, B. Lemyk, *Anal. Chem.* **49** (1977) 814 – 819.
49. W. H. Glaze, G. R. Peyton, R. Rawley, *Environ. Sci. Technol.* **11** (1977) 685 – 690.
50. R. J. Lewis: *Registry of Toxic Effects of Chemical Substances 1981 – 2,* National Institute of Occupational Safety and Health, NIOSH.
51. *The Merck Index – An Encyclopedia of Chemicals, Drugs and Biologicals,* 9th ed.,1983.
52. G. S. Moore, E. J. Calabrese, *J. Environ. Pathol. Toxicol.* **4** (1980) 257 – 263.
53. E. W. Wilde, R. J. Soracco, L. A. Mayack et al., *Bull. Environ. Contam. Toxicol.* **31** (1983) 309 – 314.

Chloroanilines → Aniline
Chlorobenzenes → Chlorinated Hydrocarbons
Chlorobiphenyls → Chlorinated Hydrocarbons
Chlorocarbons → Chlorinated Hydrocarbons
Chloroethanes → Chlorinated Hydrocarbons
Chlorofluorocarbons → Fluorine Compounds, Organic
Chloroform → Chlorinated Hydrocarbons

Chloroformic Esters

Siegfried Böhm, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 1, 2, 3, 4, 5, 6, 7 and 8)
Maren Beth-Hübner, BG Chemie, Heidelberg, Federal Republic of Germany (Chap. 9)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Production 2
5. Safety 3
6. Quality Specifications and Analysis . . 3
7. Storage and Transportation 3
8. Uses and Economic Aspects 4
9. Toxicology 5
10. References 6

1. Introduction

Chloroformic esters are also known as chloroformates, chlorocarbonates, or chlorocarbonic esters (Chemical Abstracts Service calls them carbonochloridic acid esters). They are represented by the general formula ClCOOR, wherein R can be any alkyl or aryl group [1, 2]. They were first described by Dumas in 1833 [3]. They are highly reactive compounds and, since about 1950, have attained considerable importance as intermediates in organic syntheses [4].

2. Physical Properties

Commercially important chloroformic esters are predominantly colorless liquids having boiling points only slightly higher than the corresponding alcohols. The esters of low-boiling alcohols have pungent odors and are highly volatile.

In general, chloroformic esters have densities greater than 1.00 and are immiscible or only slightly miscible with water. They are, however, miscible with inert organic solvents. The order of thermal stability of the different types of esters is aryl > primary alkyl > secondary alkyl > tertiary alkyl. The phenyl ester decomposes only slightly at its boiling point of 189 °C, but the *tert*-butylester starts decomposing at 10 °C. Decomposition is catalyzed by metal salt impurities (iron, aluminum, copper, cobalt, magnesium, and nickel) and by tertiary amines. It can proceed in two ways and lead to two different sets of products both of which include gases, which cause increased pressure in closed containers.

$$ClCOOCH_2CH_2R \rightarrow CH_2 = CHR + HCl + CO_2$$
$$ClCOOCH_2CH_2R \rightarrow ClCH_2CH_2R + CO_2$$

where R may be any alkyl or aryl group

The physical properties of the commercially important chloroformic esters are given in Table 1.

3. Chemical Properties

Chloroformic esters exhibit the characteristic chemical properties of acid chlorides. They react with all nucleophiles including carbon acids. They react with water to produce hydrogen chloride and their corresponding alcohol. Moist chloroformic esters are, therefore, highly corrosive. They form carbonates (→ Carbonic Esters) and thiocarbonates in reactions with alcohols and mercaptans, respectively. Chloroformic esters react with carboxylic acids to form carboxylic acid anhydrides [6] and with primary and secondary amines to form carbamic acid esters [7].

Their reactions with alkali peroxides form dialkyl peroxydicarbonates [8]. Bis(chloroformic) esters obtained from glycols or dihydric phenols undergo reactions with dihydric alcohols to produce polycarbonates, with dicarboxylic acids to give polyesters, or with amino compounds to yield polyurethanes [1]. They react with nitriles to form amides [9].

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_559.pub2

Table 1. Physical properties of the commercially important chloroformic esters

Ester	Empirical formula	M_r	*bp*, °C/kPa	$d^{20}{}_4$	$n^{20}{}_D$	Flash point, °C	Autoignition temperature, °C
Methyl [*79-22-1*]	$C_2H_3ClO_2$	94.50	71/101.3	1.2298	1.3880	13	510
Ethyl [*541-41-3*]	$C_3H_5ClO_2$	108.53	94/101.3	1.1403	1.3947	15	500
Allyl [*2937-50-0*]	$C_4H_5ClO_2$	120.54	110/101.3 56.5/12.9	1.1360	1.4223	28	
n-Propyl [*109-61-5*]	$C_4H_7ClO_2$	122.55	115/101.3	1.09202	1.4034	34	
Isopropyl [*108-23-6*]	$C_4H_7ClO_2$	122.55	105/101.3 (decomp.) 66.3/26.7	1.0777	1.3996	14	535
n-Butyl [*592-34-7*]	$C_5H_9ClO_2$	136.58	142/101.3 35/1.7	1.0513	1.4121	40	435
Phenyl [*1885-14-9*]	$C_7H_5ClO_2$	156.57	189/101.3 88/2.67	1.2350	1.5160	69	500
Cyclohexyl [*13248-54-9*]	$C_7H_{11}ClO_2$	162.62	ca. 180/101.3 (decomp.) 72/2.67	1.1260	1.4588	65	320
2-Ethylhexyl [*24468-13-1*]	$C_9H_{17}ClO_2$	192.69	183/101.3 (decomp.)	0.9840	1.4330	74	235
4-*tert*-Butylcyclohexyl [*42125-46-2*] *cis*-[*15595-62-7*]* *trans*- [*15595-61-6*]*	$C_{11}H_{19}ClO_2$	218.73	137 (decomp.) 58 – 62/0.13	1.0450	1.4636	64	390
Diethylene glycol bis- [*106-75-2*]	$C_6H_8Cl_2O_5$	231.03	95 – 97/0.05	1.3880	1.4536	160	
Cetyl [*26272-90-2*]	$C_{17}H_{33}ClO_2$	304.90	ca. 140/101.3 (decomp.)	0.9250	1.448	140	ca. 300

* No physical data reported; identification by IR and NMR data [5].

4. Production

The most important method used to produce chloroformic esters is the reaction of phosgene with anhydrous alcohols.

$$ROH + COCl_2 \rightarrow ClCOOR + HCl$$

where R may be any alkyl or aryl group

The reaction with aliphatic alcohols proceeds with no need for hydrogen chloride acceptors or catalysts. The conversion of phenols with phosgene occurs less readily. The reaction depends on the presence of catalysts such as alkali salts [10], tertiary amines or quaternary ammonium salts [11], carboxylic acid amides (e.g., dimethylformamide or *N*-methylpyrrolidone) [12], or organic phosphorus compounds (e.g., triphenylphosphine, triphenylphosphine oxide) [13] and is carried out discontinuously at high temperature. The formation of carbonates by a competing reaction depends on the nature of the catalyst. Alkali salt catalysts lead to substantially more carbonate formation than phosphorus compounds. To suppress carbonate formation, the reaction heat is rapidly carried off and an excess of phosgene is maintained in the reaction zone.

Continuous and discontinuous processes are available for the preparation of chloroformic esters. In the laboratory the synthesis is carried out by passing alcohol into liquid phosgene at low temperature. The discontinuous method is generally not applicable to commercial syntheses for reasons of safety. In special cases where this method is employed, an inert diluent is required, e.g., the ester itself [14] or aliphatic, aromatic, or chlorinated hydrocarbons, and the alcohol and phosgene enter the reaction vessel simultaneously. The solvent, containing dissolved hydrogen chloride and excess phosgene, is removed by distillation [15]. The discontinuous method is used for the production of chloroformic esters from glycols and the conversion of sterically hindered or acid-sensitive alcohols. An auxiliary base, e.g., pyridine or dimethylaniline, increases the reaction rate.

The continuous procedure is the method of choice for the commercial production of aliphatic chloroformic esters. It can be carried

out in either cocurrent or countercurrent flow. The required amounts of alcohol and phosgene can be passed in cocurrent flow into a circulation reactor, cascade vessel [16], or nozzle atomizer [17], or the components can react in the gas phase [18, 19]. The countercurrent method is carried out in a packed column [20, 21] because hydrogen chloride and excess phosgene then escape from the top of the column and the liquid ester collects at the bottom. A product of 99 % purity can be obtained if the temperature in the reaction zone is maintained below 10 °C for the formation of the methyl ester or below 20 °C for the ethyl ester [21]. Esters of alcohols with an intermediate chain length ($C_3 - C_8$) are produced at 30 – 70 °C. Purity of 97 – 99 % and a yield greater than 90 % are achieved with esters of primary alcohols, whereas the more unstable esters of secondary and tertiary alcohols give 70 – 80 % yields.

Chloroformic esters can also be made by reacting alcohols with carbon monoxide and chlorine under pressure [22], by chlorinating formates or thioesters, by reacting phosgene with aldehydes, or by reacting phosphorus pentachloride with cyclic carbonates [2, 4]. Chloroformic esters with β-chlorine atoms are made by reacting phosgene with the corresponding epoxide [1], by reacting alcohols with trichloroacetyl chloride [1], and by cleaving α-keto sulfoxides with hydrogen chloride [2, 23]. The crude ester is purified by distillation unless, as is the case for the aliphatic esters, thermal instability precludes this method.

5. Safety

Safety precautions necessary for the handling of phosgene apply to the production of chloroformic esters. The exhaust gas is hydrolyzed and then burned. Undesirable substances, e.g., phenol, are extracted from wastewater.

Some chloroformic esters are classified as dangerous working substances [24] and may be handled only when protective clothing and equipment are used (gas masks with filter B, rubber gloves, etc.). They must not enter the atmosphere or the water supply. If spilled or leaked, they can be hydrolyzed by alkaline substances. Vapors must be destroyed in a scrubber with alkali or with water in the presence of activated carbon or by incineration. Clothing that comes into contact with chloroformic esters must be removed immediately. Any part of the body that is touched by a chloroformic ester must be washed well with water immediately. All important chloroformic esters are flammable; they form explosive mixtures with air. Electrostatic charges must, therefore, be avoided.

6. Quality Specifications and Analysis

Commercial products have concentrations of 97 – 99 wt %. Along with small amounts of carbon dioxide, hydrogen chloride, phosgene, and alkyl chlorides, the main impurities are the corresponding carbonates. The amounts of carbon dioxide, hydrogen chloride, and alkyl chlorides increase with the duration of storage, depending on the stability of the ester. A purity of 97 % is adequate for most applications. If especially pure product is required, 99 % pure, distilled methyl chloroformate is available. The methyl ester can be distilled under normal pressure, but the ethyl ester requires vacuum conditions and continuous distillation. Esters obtained from secondary or tertiary alcohols are not sufficiently stable to be purified by distillation. They must be washed with cold water and dried with calcium chloride.

The quantitative analysis of chloroformic esters can be carried out by titration, gas chromatography, or high-performance liquid chromatography (HPLC). High-boiling or thermally labile esters are determined by titration. They are treated with aniline in methanol, and the liberated hydrochloric acid is titrated with sodium hydroxide. They can also be treated with excess dibutylamine in chlorobenzene; the remaining amine is then back-titrated with hydrochloric acid.

Bivalent esters are assayed by reacting them with an aniline derivative, e.g., methylaniline, and using HPLC to measure the dicarbamate formed.

7. Storage and Transportation

The thermal stability of chloroformic esters is important for storage and transportation. Esters

of primary alcohols and phenols are stable if free of moisture and metal salts. Traces of moisture and heat can generate hydrogen chloride and increase the internal pressure. For this reason, containers should not be exposed to direct sunlight. The esters of secondary and tertiary alcohols (e.g., isopropyl ester or cyclohexyl ester) require rapid transportation and no more than brief storage because they are unstable.

The storage temperature should be maintained below 10 °C and the product used as soon as possible. Iron drums lined with polyethylene or lead-lined tank cars are used to transport chloroformic esters. The methyl and ethyl esters are mentioned in the regulations for the transportation of dangerous materials (GGVE), clause 6.1, class 4. The corrosiveness, toxicity, and flammability (primary label) of the esters must be declared on the container and EEC guidelines require that dangerous working substances be provided with an additional dangerous-materials label giving instructions for handling (R and S label).

8. Uses and Economic Aspects

Chloroformic esters are highly reactive and can be used as intermediates in the synthesis of numerous compounds. They are employed in the production of unsymmetrical carbonic acid esters (→ Carbonic Esters) such as diethylene glycol bis(allyl) carbonate [25], which is an intermediate in the synthesis of plastics of excellent optical quality. Other aliphatic or aromatic bis(chloroformic) esters are also used to make plastics such as polycarbonates (Makrolon) [26, 27] and polyurethanes having a high temperature resistance [1, 28]. Alkyl and aryl esters of chloroformic acid serve as catalysts in the presence of initiator radicals in the low-pressure polymerization of ethylene [29]. The methyl and ethyl esters are used as solvents in the photographic industry. They are also employed in the production of carbamates that are used to synthesize dyes (e.g., Resolin dyes), drugs, (e.g., the tranquilizer meprobamate [*57-53-4*]), veterinary medicines (e.g., furazolidone [*67-45-8*] or rintal [*58306-30-2*]), herbicides, (e.g., betanal [*13684-63-4*] and betanex [*13684-56-5*]), and insecticides, (e.g., Sevin [*63-25-2*]).

The herbicide IPC, isopropyl *N*-phenylcarbamate [*122-42-9*], is made from isopropyl chloroformate. Another important application is the preparation of the dialkyl peroxycarbonates used as catalysts in the sodium peroxide production of PVC, poly(vinyl chloride) [*78320-64-6*] [30]. Such commercially important peroxycarbonates as ethyl [*14666-78-5*], *n*-butyl [*16215-49-9*], isopropyl [*105-64-6*], cetyl [*26322-14-5*], cyclohexyl [*1561-49-5*], and 4-*tert*-butylcyclohexyl peroxycarbonate [*26523-73-9*] can be made from the corresponding chloroformates. Methyl chloroformate is the starting material for the synthesis of the preservative velcorin (dimethyl pyrocarbonic ester) [*4525-33-1*]. Phenyl chloroformate is used in the production of Sirius dyes [31]. Benzyl chloroformate is used to substitute the carbobenzyloxy protecting group on amino functions of amino acids in peptide synthesis [32], on hydroxyl groups [33], and on hydrazino groups [34]. The protecting group is easily removed by hydrogenolysis over a catalyst [35].

Ethyl chloroformate is the most important of these esters. It reacts with ethyl xanthate [*140-90-9*] to produce diethylxanthogen formate [*3278-35-1*] and diethylxanthic anhydride [*2905-52-4*], which are used in the flotation of ores [8]. Ethyl chloroformate is also used as a stabilizer for PVC and in the production of modified penicillins, e.g., binotal [*69-53-4*]. It is used to produce carboxylic acid anhydrides from carboxylic acids [6], secondary amines from tertiary amines [36], and nitriles from amides [9] and to make heterocyclic compounds [37].

The cyclohexyl, 4-*tert*-butylcyclohexyl, butyl, and cetyl esters of chloroformic acid are used almost exclusively to produce peroxydicarbonates. The majority of the isopropyl ester produced is also used for this purpose (total production is several thousand metric tons per year).

Chloroformic esters are not produced on a large scale. Their economic importance is due to their usefulness as intermediates in organic syntheses. The largest producers in Western Europe are BASF, Société Nationale des Poudres et Explosifs, Hoechst, and Bayer. In the United States, Chemetron Corp., Essex Chem. Corp., Minerec Corp., and PPG Industries are the major producers. Hodogaya Co. in Japan and Chemolimpex in Hungary are also major suppliers. The pro-

Table 2. LD_{50} values of chloroformic esters [39 – 42, 49, 51, 52, 54 – 62]

Esters increasing chain length	Oral LD_{50}, mg/kg b.w.*	Dermal LD_{50}, mg/kg b.w.*	Inhalation LC_{50}, mL/m^3 (ppm)
Methyl	40 – 313 (rat)	894 – 1230 (rat)	15 – 115 (rat, 4 h)
	40 – 67 (mouse)	1750 (mouse)	88 – 103 (rat, 1 h)
	140 (guinea pig)	7120 (rabbit)	47 (mouse, 2h)
Ethyl	205 – 470 (rat)	7120 (rabbit)	145 – 196 (rat, 1 h)
Allyl	244 (rat)		6.55 (rat)
	210 (mouse)		4.67 (mouse)
	178**	1 470**	
n-Propyl	872 – 1212 (rat)		
	650 (mouse)	10 (mouse)	319 (mouse, 1 h)
Isopropyl	1070 (rat)		
	178 (mouse)	12 (mouse)	299 (mouse, 1 h)
		11 300 (rabbit)	
n-Butyl	1325 – 2610 (rat)		
Phenyl	1581 (rat)	4800 (rabbit)	44 (rat, 4 h)
2-Ethylhexyl	5420 (rat)		34.21 (rat, 4 h)
4-*tert*-Butylcyclohexyl	>5000 (rat)	>2000 (rat)	80.35 (rat, 4 h)
Diethylene glycol bis-	813 (mouse)	3400 (mouse)	169 (mouse, 1 h)
Cetyl	5000 (rat)		no symptoms

* b.w. = body weight
** Test animals not specified.

duction capacity in Western Europe is ca. 20 – 25 kt/a.

9. Toxicology

Chloroformic esters cause acute reactions of the respiratory tract and mucous membranes. Contact with the gastrointestinal tract or absorption by the skin leads to less toxic reactions (Table 2). Chloroformic esters burn the skin, leaving its surface hard and stained brown [38 – 42], and they are corrosive to the eyes [39 – 42]. Inhalation of vapors of short-chain chloroformic esters results in coughing, choking, and respiratory distress, and with some chloroformates, such as methyl chloroformate, inhalation can cause death as a result of the onset of pulmonary edema, which may appear only several hours after exposure, as seen in phosgene poisoning [39, 43, 44]. The inhalative toxicity of chloroformic esters decreases with increasing chain length [39 – 42]. Cetyl chloroformate is not toxic even when saturated air is inhaled [45, 46]. The introduction of a halogen in the α-position of aliphatic esters increases their inhalative toxicity [47]. Inhalation of methyl chloroformate (shortest chain length) in a 90-d study caused marked damage throughout the respiratory tract in rats. Exposure to 8 mL/m^3 (31 mg/m^3) caused some lethality. No treatment-related toxicological changes were observed in organs other than the respiratory tract. Irrespective of the number of exposures, the no observed adverse effect concentration (NOAEC) was 0.4 mL/m^3 (1.6 mg/m^3) for 3, 10, 20, and 65 exposures in this study. The same NOAEC was found in a 28-d study in rats [39]. In an exploratory subacute inhalation study with ethyl chloroformate, 20 exposures to 1 mL/m^3 (4.5 mg/m^3) were tolerated by rats without remarkable findings, whereas 10 exposures to 20 mL/m^3 (90 mg/m^3) resulted in signs of irritation of the respiratory tract and poor general condition [40, 48]. Butyl chloroformate had an NOAEC of 1.8 mL/m^3 (10 mg/m^3) in a 28-d inhalation study [42].

On the basis of these studies, the German MAK Commission established MAK values of 0.2 mL/m^3 (0.78 mg/m^3) for methyl chloroformate and 0.2 mL/m^3 (1.1 mg/m^3) for butyl chloroformate. Furthermore, both substances have been assigned to pregnancy risk group C, i.e., substances for which "there is no reason to fear a risk of damage to the embryo when MAK and BAT values are observed" [39, 42]. Further limit values exist in the Netherlands and Australia/UK: ethyl chloroformate 1 mL/m^3 (4.4 mg/m^3) [49, 50], isopropyl chloroformate 1 mL/m^3 (5.1 mg/m^3) [51], and 2-ethylhexyl

chloroformate 0.13 mL/m^3 (7.9 mg/m^3) [49, 52].

In Salmonella/microsome tests, with or without metabolic activation, neither methyl nor ethyl, nor propyl nor butyl chloroformate revealed a mutagenic potential. Only with methyl and butyl chloroformate was a chromosomal aberration test conducted in V79 cell lines of the Chinese hamster. Both substances showed no clastogenic effect in the absence of metabolic activation, whereas in the presence of metabolic activation the aberration rate was increased, a result that was not reproducible in an independent study with butyl chloroformate [39 – 42].

In an exploratory carcinogenicity study, male Sprague – Dawley rats underwent whole-body exposure to ethyl chloroformate at levels of 1.5, 3.0, or 6.0 mL/m^3 (ca. 6.75, 13.5, and 27 mg/m^3 air) 6 h/d, 5 d per week for 30 d and were subsequently held for life-span observation. One out of 50 animals exposed to 6 mL/m^3 and examined at the age of ca. 2 years (700 d) were found to have squamous cell carcinoma of the nasal mucosa. None of the 98 controls exposed to test-substance-free air under the same conditions exhibited tumors of that type. Among animals treated with 6 mL/m^3 ethyl chloroformate, 15 out of 50 (30 %) exhibited squamous metaplasia of the nasal mucosa, and 45 out of 50 (90 %) had rhinitis; in the control group, 9 out of 98 (9.1 %) displayed squamous metaplasia, and 80 out of 98 (ca. 82 %) had rhinitis. The investigators did not analyze their data for statistical significance. Further exploratory carcinogenicity studies in female ICR/Ha Swiss mice (30 – 50 mice per group) receiving lifetime (18 – 22 months) dermal or subcutaneous administrations of ethyl chloroformate showed no increase in tumor incidence compared with controls. Ethyl chloroformate was administered to the 30 – 50 animals per group dermally three times per week at doses of 3, 4.3 or 5.5 mg per administration or subcutaneously once weekly at doses of 0.3 or 1.1 mg per administration. According to the investigators, the highest doses corresponded to the maximum tolerated dose (MTD) in each case. In parallel, the same investigators carried out an initiation – promotion test with single dermal application of 5.5 mg ethyl chloroformate followed by a 14-d treatment-free period and lifetime dermal applications of phorbol myristate acetate (PMA) as promoter three times per week. According to the investigators, animals treated with ethyl chloroformate and PMA exhibited marginally increased tumor incidences (6 out of 50 animals developed tumors, 4 papillomas, and 2 squamous carcinomas). It is unclear precisely what amount of PMA was used to treat animals on initiation with ethyl chloroformate and, consequently, which PMA control group was used for comparison. Therefore, the evaluation lacks clear plausibility [40].

In Germany the MAK Commission has assigned ethyl chloroformate to Category 3B of carcinogenic substances (i.e., "Substances that cause concern that they could be carcinogenic for man but cannot be assessed conclusively because of lack of data. The classification in Category 3 is provisional.") [53].

10. References

1. M. Matzner, P. Kurkjy, J. Cotter, *Chem. Rev.* **64** (1964) no. 6, 645–687; *Chem. Abstr.* **62** (1965) 1525a.
2. *Houben-Weyl*, **8**, 101–105. *Houben-Weyl*, **E4**, 15–28.
3. J. Dumas, *Ann. Chim. Phys., 2. Série*, **54** (1833) 225–247.
4. R. D. Concez, *Inf. Chem.* **106** (1972) 139.
5. P. Beak, J. T. Adams, J. A. Barron, *J. Am. Chem. Soc.* **96** (1974) 2494.
6. E. Schipper, J. Nichols, *J. Am. Chem. Soc.* **80** (1958) 5714.
7. *Org. Synth. Coll. Vol.* **IV** (1963) p. 780.
8. G. Crozier, US 4 454 051, 1984. Bosordi Vegyi Kombinat, HU 30 021, 1984 (B. Mariasi, L. Molnar, J. Scabo, P. Jubasz, A. Karcagi).
9. C. K. Sams, R. J. Cotter, *J. Org. Chem.* **26** (1961) 6.
10. Bayer, DE 1 117 598, 1960 (W. Altner, E. Meisert, G. Rockstroh).
11. Union Carbide, US 3 255 230, 1966 (R. P. Kurkjy, M. Matzner, R. J. Cotter). J. Mouralova, J. Hajicek, J. Trojanek, CS 202 454, 1983.
12. Hoechst, DE 2 131 555, 1971 (G. Semler, G. Schaeffer). Hoechst, DE 2 213 408, 1972 (G. Semler, G. Schaeffer).
13. Bayer, DE 3 019 526, 1980 (G. Rauchschwalbe, H.-U. Blank, K. Mannes, D. Mayer).

14. PPG, FR 2 484 406, 1981 (A. Sathe).
15. Synvar Associates, DE 2 144 963, 1972 (M. E. Packard, E. F. Ullmann, T. L. Burkoth).
16. BASF, DE 2 251 206, 1972 (G. Merkel, J. Datow, J. Paetsch, H. Toussaint, H. Hoffmann, S. Winderl).
Minerec Corp., DE 2 704 262, 1977 (F. S. Bell, R. D. Crozier, L. E. Strow).
F. František, L. Skalicky, J. Pscheidt, F. Janda, CS 190 921, 1981.
17. BASF, DE 2 453 284, 1974 (F. Neumayr, M. Decker, J. Paetsch).
Bayer, DE 2 847 484, 1980 (W. Schulte-Huermann, E. Schellmann, J. Lahrs).
18. Mitsubishi Chemical, JP-Kokai 51/43 719, 1974 (S. Ueda, Y. Mijagawa, H. Yoshioka, E. Hirayawa).
Mitsubishi Chemical, JP-Kokai 51/43 721, 1974 (S. Ueda, Y. Mijagawa, H. Yoshioka, E. Hirayawa).
19. Eszakmagyarorszagi Vegyimüvek, Sajobabony (Hungary), DE 2 159 967, 1971 (L. Rosza, L. Meszaros, F. Mogyorodi).
20. ICI, FR 1 336 606, 1961.
21. Bayer, DE 3 135 947, 1983 (D. Bauer, H. Dohm, W. Schulte-Huermann, H. Hemmerich).
22. Ube Industries, JP-Kokai 54/61 121, 1979 (S. Nakatomi, Y. Kamitoku).
23. T. Wakin, Y. Nakamina, S. Motoki, *Bull. Chem. Soc. Jpn.* **51** (1978) 3081.
24. EEC Guideline 79/831/EEC.
25. PPG, US 4 273 726, 1979 (A. Senol).
PPG, EP 80 339, 1981 (C. W. Eads, J. C. Crano).
26. Bayer, DE 971 790, 1953 (H. Schnell, L. Bottenbruch, H. Krimm).
27. K. Johnson, *Polycarbonates Recent Developments*, Chemical Process Review no. 47 Noyes Data Corp., Park Ridge, N.J., 1970.
28. Bayer, DE 1 720 761, 1967 (H. Krimm, G. Lenz, H. Schnell).
29. Union Rheinische Braunkohlen Kraftstoff, DE 1 126 139, 1958 (G. Nettesheim, R. Schulze-Bentrop).
30. Goodyear, DE 2 039 010, 1970 (F. A. Cox, D. R. Glenn).
Akzona Inc., US 4 269 726, 1981 (J. R. Kolczynski, G. A. Schultz).
31. *Ullmann*, 3rd ed., **4**, 105, 108, 109.
32. A. C. Farthing, *J. Chem. Soc.* 1950, 3 213.
33. L. Hough, J. E. Pridde, *J. Chem. Soc.* 1961, 3178.
34. H. Niedrich, W. Knobloch, *J. prakt. Chem.* **17** (1962) 263, 273.
35. R. A. Boisonnes, G. Breitner, *Helv. Chim. Acta* **36** (1953) 875.
B. F. Daubert, C. G. King, *J. Am. Chem. Soc.* **61** (1939) 3328.
36. Ciba Geigy, BE 615 026, 1962 (W. Schindler).
37. H. Krauch, W. Kunz: *Reaktionen der Org. Chemie*, Hüthig Verlag 4th ed., Heidelberg 1969, p. 596.
38. G. Schreiber, Fraunhofer Institut für Toxikologie und Aerosolforschung, Bericht vom 29. 10. 1980.
39. BG Chemie, Toxicological Evaluation No. 36, Chloroformic acid methyl ester, www.bgchemie.de/toxicologicalevaluations, 02/05 (2005).
40. BG Chemie, Toxicological Evaluation No. 77, Chloroformic acid ethyl ester, www.bgchemie.de/toxicologicalevaluations, 02/05 (2005).
41. BG Chemie, Toxicological Evaluation No. 159, Chloroformic acid propyl ester, www.bgchemie.de/toxicologicalevaluations, 02/05 (2005).
42. BG Chemie, Toxicological Evaluation No. 160, Chloroformic acid butyl ester, www.bgchemie.de/toxicologicalevaluations, 02/05 (2005).
43. W. Hey et al., *Arch. Toxikol.* **23** (1968) 186.
44. A. M. Thies et al., *Zentralbl. Arbeitsmed. Arbeitsschutz* **18** (1968) 141.
45. G. Schreiber, Fraunhofer Institut für Toxikologie und Aerosolforschung, Bericht 1 vom 20. 02. 1980.
46. G. Schreiber, Fraunhofer Institut für Toxikologie und Aerosolforschung, Bericht 2 vom 20. 02. 1980.
47. F. Flury, F. Zernik: *Schädliche Gase*, Springer Verlag Berlin 1931.
48. J. C. Gage, *Br. J. Ind. Med.* **27** (1970) 1–18.
49. BIA (Berufsgenossenschaftliches Institut für Arbeitssicherheit), Gefahrstoffliste 2004 – Gefahrstoffe am Arbeitsplatz, BIA-Report 1/2004, Hauptverband der gewerblichen Berufsgenossenschaften, 2004.
50. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Formic acid, chloro-, ethyl ester, last updated 11, 2004.
51. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Formic acid, chloro-, isopropyl ester, last updated 11, 2004.
52. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Carbonochloridic acid, 2-ethylhexyl ester, last updated 11, 2004.

53. DFG: List of MAK and BAT Values 2005, Wiley-VCH, Weinheim 2005.
54. R. J. Lewis, Sr. and R. L. Tatken (eds.): *Registry of Toxic Effects of Chemical Substances 1981 – 1982*, NIOSH Cincinnati, Ohio, 1983.
55. *Kirk-Othmer*, 3rd ed., **4**, 765.
56. Safety data sheet Bayer AG, Leverkusen.
57. Toxicological Institute, Bayer AG, report no. 9741, 1981.
58. *Am. Ind. Hyg. Assoc. J.*, **30** (1969) 470.
59. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Carbonochloridic acid, 4-(1,1-dimethylethyl)cyclohexyl ester, last updated 12, 2000.
60. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Formic acid, chloro-, allyl ester, last updated 07, 2000.
61. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Formic acid, chloro-, oxydiethylene ester, last updated 06, 1998.
62. MDL Information Systems, RTECS — Registry of Toxic Effects on Chemical Substances, Carbonochloridic acid, phenyl ester, last updated 01, 1997.

Chlorohydrins

Gordon Y. T. Liu, Dow Chemical, Midland, Michigan 48640, United States
W. Frank Richey, Dow Chemical, Midland, Michigan 48640, United States
Joanne E. Betso, Dow Chemical, Midland, Michigan 48640, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 5
5. **Environmental Protection** 8
6. **Chemical Analysis** 8
7. **Storage and Transportation** 8
8. **Uses** 9
9. **Toxicology and Occupational Health** . 9
10. **References** 11

1. Introduction

Chlorohydrins are compounds with one or more chlorine and hydroxyl groups in the nonaromatic portion of their structure. The compounds of greatest industrial importance are the propylene chlorohydrins and glyceryl chlorohydrins (see Table 1). Ethylene chlorohydrin is no longer an important intermediate for the production of ethylene oxide, one of the classic commodity chemicals.

Nomenclature. Many different names for the various chlorohydrins appear in the literature. Some of these *synonyms* for the major chlorohydrins are given in Table 1.

2. Physical Properties

Most dilute solutions of chlorohydrins in water have a somewhat sweet, pleasant odor, particularly 3-chloro-1,2-propanediol (**5**), 1,3-dichloro-2-propanol (**7**), and 3-chloro-1-propanol (**4**). The general physical properties of the more significant chlorohydrins are listed in Table 2.

3. Chemical Properties

Chlorohydrins undergo reactions which are characteristic of both alcohols and alkyl chlorides. This combination is responsible for the most common reaction of chlorohydrins, which is *dehydrochlorination to form epoxides:*

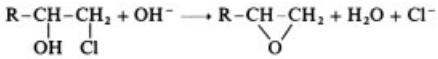

The mechanism of dehydrochlorination of ethylene chlorohydrin has been studied by comparing the relative rates of reaction in H_2O and D_2O [1]. The kinetics of the propylene chlorohydrin conversion to propylene oxide were determined with both NaOH and $Ca(OH)_2$ [2]. The relative rates of solvolysis of ethylene chlorohydrin, propylene chlorohydrin, and butylene chlorohydrin in water at 97 °C are 1.0 : 0.81 : 5.5 [3].

Hydroxyalkyl ethers may be prepared by reacting chlorohydrins with alcohols or phenols under basic conditions [4]. *Ethers of cellulose and starch* are manufactured by reacting these materials with various chlorohydrins. *Cationic starch* is made by reacting 3-chloro-2-hydroxypropyl trimethylammonium chloride with a starch slurry at pH 11 – 12 [5].

Chlorohydrins react with carboxylates in the presence of base to give *hydroxyalkyl esters.* On the other hand, *β-chloro esters* are formed by reaction with carboxylic acids under acid conditions [6] or from acid chlorides [7]. *Cyclic carbonates* may be produced by the reaction of a chlorohydrin with carbon dioxide in the presence of an amine [8].

A comparison of the rate of reaction of ethylene chlorohydrin with various *amines* gives the following order: *n*-amylamine > cyclohexylamine > aniline [9]. The reaction of ethylene chlorohydrin with ammonia gives

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a06_565

Table 1. Chlorohydrin nomenclature

Chlorohydrin	CAS registry number	Formula	M_r	Synonyms
Ethylene		C_2H_5ClO	80.52	
2-Chloro-1-ethanol (**1**)	[*107-07-3*]	$ClCH_2CH_2OH$		ethylene chlorohydrin, 2-chloroethyl alcohol, 1-chloro-2-hydroxyethane, 2-chloro-1-hydroxyethane, β-chloroethyl alcohol, β-chloroethanol, ethylene glycol chlorohydrin, glycol chlorohydrin, glycol monochlorohydrin, glycolmonochlorohydrin, 2-monochloroethanol
Propene		C_3H_7ClO	94.54	
2-Chloro-1-propanol (**2**)	[*78-89-7*]	$CH_3CHClCH_2OH$		2-propylene chlorohydrin, 2-chloropropyl alcohol, β-chloropropyl alcohol, propylene β-chlorohydrin, 2-chloro-1-hydroxypropane
1-Chloro-2-propanol (**3**)	[*127-00-4*]	$CH_3CHOHCH_2Cl$		β-chloroisopropyl alcohol, *sec*-propylene chlorohydrin, 1-chloroisopropyl alcohol, propylene α-chlorohydrin, 1-propylene chlorohydrin, 1-chloro-2-hydroxypropane
3-Chloro-1-propanol (**4**)	[*627-30-5*]	$Cl(CH_2)_3OH$		1-chloro-3-hydroxypropane, trimethylene chlorohydrin
Glycerol		$C_3H_7ClO_2$	110.54	
3-Chloro-1,2-propanediol(**5**)	[*96-24-2*]	$CH_2OHCHOHCH_2Cl$		α-chlorohydrin, 1-chloropropane-2,3-diol, glycerol α-chlorohydrin, 3-chloro-1,2-dihydroxypropane, 3-chloro-1,2-propanediol, 3-chloropropylene glycol, glycerol α-monochlorohydrin, α-monochlorohydrin, 1-chlorohydrin, chlorodeoxyglycerol, 1-chloro-2,3-dihydroxypropane, 3-chloropropane-1,2-diol, 2,3-dihydroxypropyl chloride, glycerin α-monochlorohydrin, α-glycerol chlorohydrin, monochlorhydrin, monochlorohydrin
2-Chloro-1,3-propanediol (**6**)	[*497-04-1*]	$HOCH_2CHClCH_2OH$		2-chlorotrimethylene glycol, glyceryl β-monochlorohydrin, β-chlorohydrin, 2-chlorohydrin, chloro-1,3-propanediol, chloropropanediol-1,3
Glycerol dichlorohydrins		$C_3H_6Cl_2O$	128.99	
1,3-Dichloro-2-propanol (**7**)	[*96-23-1*]	$CH_2ClCHOHCH_2Cl$		1,3-dichlorohydrin, a,γ-dichlorohydrin, β,β-dichloroisopropyl alcohol, glycerin α,α'-dichlorohydrin, α-dichlorohydrin, 1,3-dichloro-2-hydroxypropane
2,3-Dichloro-1-propanol (**8**)	[*616-23-9*]	$CH_2ClCHClCH_2OH$		1,2-dichlorohydrin, α,β-dichlorohydrin, β,γ-dichloropropyl alcohol, allyl alcohol dichloride, asymmetric glycerin dichlorohydrin, β-dichlorohydrin, 2,3-dichloro-1-hydroxypropane
Butene		C_4H_9ClO	108.57	
2-Chloro-1-butanol (**9**)	[*26106-95-6*]	$CH_3CH_2CHClCH_2OH$		2-chloro-*n*-butyl alcohol, 2-chloro-1-hydroxybutane
3-Chloro-1-butanol (**10**)	[*2203-35-2*]	$CH_3CHClCH_2CH_2OH$		3-chloro-*n*-butyl alcohol, 3-chloro-1-hydroxybutane
4-Chloro-1-butanol (**11**)	[*928-51-8*]	$CH_2ClCH_2CH_2CH_2OH$		4-chloro-*n*-butyl alcohol, 4-chloro-1-hydroxybutane, 4-chloro-1-butane-ol, 4-chlorobutanol, tetramethylene chlorohydrin
1-Chloro-2-butanol (**12**)	[*1873-25-2*]	$CH_3CH_2CHOHCH_2Cl$		1-chloro-*sec*-butyl alcohol, 1-chloro-2-hydroxybutane
3-Chloro-2-butanol (**13**)	[*563-84-8*]	$CH_3CHClCHOHCH_3$		3-chloro-*sec*-butyl alcohol, 3-chloro-2-hydroxybutane
4-Chloro-2-butanol (**14**)	[*2203-34-1*]	$CH_2ClCH_2CHOHCH_3$		4-chloro-*sec*-butyl alcohol, 4-chloro-2-hydroxybutane
Isobutene		C_4H_9ClO	108.57	
2-Chloro-2-methyl-1-propanol (**15**)	[*558-38-3*]	$(CH_3)_2CClCH_2OH$		β-isobutylene chlorohydrin
1-Chloro-2-methyl-2-propanol (**16**)	[*558-42-9*]	$(CH_3)_2COHCH_2Cl$		
Isobutene dichlorohydrins		$C_4H_8Cl_2O$	143.01	
1-Chloro-2-chloro-methyl-2-propanol (**17**)	[*597-32-0*]	$CH_3C(OH)(CH_2Cl)_2$		
1,2-Dichloro-2-methyl-3-propanol (**18**)	[*42151-64-4*]	$HOCH_2CCl(CH_3)CH_2Cl$		
Other				
2-Chloro-1-phenyl-1-ethanol (**19**)	[*1004-99-5*]	C_8H_9ClO $C_6H_5CH(OH)CH_2Cl$	151.61	styrene chlorohydrin

Table 2. Physical properties of chlorohydrins

Chlorohydrin	*mp*, °C (*p*, kPa)	*bp*, °C point, °C	Flash	d_4^{20}, g/mL	η, mPa · s	Azeotrope			Solubility (at 25 °C)			
						Component	*bp*, °C	wt %	Water	Ethanol	Acetone	Benzene
Ethylene												
2-Chloro-1-ethanol (**1**)	− 69 (− 63 to − 40 °C)	129	57	1.2133	3.43	water	97.8	42	miscible	miscible	miscible	miscible
Propylene												
2-Chloro-1-propanol (**2**)	–	133	–	1.1032	–	water	96.0	48	very soluble	soluble	soluble	–
1-Chloro-2-propanol (**3**)	–	127	52	1.1154	4.67	water	95.4	54	miscible	miscible	miscible	miscible
3-Chloro-1-propanol (**4**)	–	16	–	1.1318	–	–	–	–	very soluble	soluble	soluble	soluble
Glycerol												
3-Chloro-1,2-propanediol (**5**)	–	213	138	1.3204	159	–	–	–	soluble	miscible	miscible	miscible
2-Chloro-1,3-propanediol (**6**)	–	146 (2.4)	–	1.3219	300	–	–	–	high	soluble	soluble	–
Glycerol dichlorohydrins												
1,3-Dichloro-2-propanol (**7**)	− 4	174	58	1.3645	–	water	99	23	15.6*	very soluble	soluble	–
2,3-Dichloro-1-propanol (**8**)	–	182	91	1.3607	–	dichloro-benzene	170.8	30	12.7*	miscible	miscible	miscible
Butene												
2-Chloro-1-butanol (**9**)	–	75 (3.3)	–	1.0622	–	–	–	–	–	high	–	–
3-Chloro-1-butanol (**10**)	–	138	–	1.0671	–	–	–	–	–	–	–	–
4-Chloro-1-butanol (**11**)	–	84 (2.7)	–	1.0883	–	–	–	–	–	high	–	–

Table 2. (Continued)

Chlorohydrin	*mp*, °C (*p*, kPa)	*bp*, °C point, °C	Flash	d_4^{20}, g/mL	η, mPa · s	Azeotrope			Solubility (at 25 °C)			
						Component	*bp*, °C	wt %	Water	Ethanol	Acetone	Benzene
1-Chloro-2-butanol (**12**)	–	141	–	1.0683	–	–	–	–	–	soluble	–	–
3-Chloro-2-butanol (**13**)	–	139	–	1.0669	–	water	94.5	59	–	high	–	–
4-Chloro-2-butanol (**14**)	–	70 (1.7)	–	–	–	–	–	–	–	high	–	–
Isobutene												
2-Chloro-2-methyl-1-propanol (**15**)	–	132.5	–	1.0472	–	–	–	–	decomp.	–	–	–
1-Chloro-2-methyl-2-propanol (**16**)	–	128	–	–	–	–	93.5	66.0	–	–	–	–
Isobutene dichlorohydrins												
1-Chloro-2-chloromethyl-2-propanol (**17**)	–	174	–	–	–	–	–	–	–	–	–	–
1,2-Dichloro-2-methyl-3-propanol (**18**)	–	180	–	–	–	–	–	–	–	–	–	–
Other												
Styrene chlorohydrin (**19**)	–	110	–	–	–	–	–	–	soluble	soluble	soluble	soluble

∗ In g/100 g.

monoethanolamine [10]. *Quaternary ammonium compounds* result from chlorohydrins and tertiary amines [11], [12].

Other chlorohydrin reactions include formation of *nitriles* from cyanides [13], *acetals* from aldehydes [14], *oxazolidinones* from cyanates [15], and the oxidation of ethylene chlorohydrin to *monochloroacetic acid* [16].

4. Production

In 1863, Carius [17] reported that chlorohydrins could be synthesized by reacting olefins with hypochlorous acid. As early as 1904, BASF began production of ethylene chlorohydrin by introducing ethylene and CO_2 into an aqueous solution of bleaching powder [18].

Ethylene Chlorohydrin. The reaction of chlorine and water (1) produces very little hypochlorous acid because of an unfavorable equilibrium, $K = 4.2 \times 10^{-4}$ [19].

$$Cl_2 + H_2O \rightleftharpoons HOCl + HCl \qquad (1)$$

Gomberg [20] reasoned that if the reaction between ethylene and hypochlorous acid (2) is significantly more rapid than the reaction of ethylene and chlorine to form dichloroethane (3), then as the ethylene is added to the aqueous system, chlorohydrin **1** should be produced preferentially.

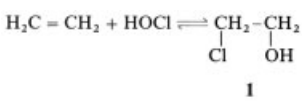

(2)

1

(3)

He showed that with good stirring to minimize reaction (3) in the gas phase, little dichloroethane is produced until the chlorohydrin concentration reaches 6 – 8 %.

In reaction (1), the concentration of hypochlorous acid declines steadily as the HCl level increases. Significant quantities of dichloroethane are observed when the HCl concentration goes above 3 %. Because the solubility of dichloroethane in water is low (0.869 g/100 mL at 20 °C), it quickly forms a separate phase into which both chlorine and ethylene are preferentially dissolved to produce even more dichloroethane.

In a laboratory process study of the hypochlorination of ethylene, the effects of various reaction parameters were determined with both single-column and recycle-reactor arrangements [21]. Reaction temperatures of 35 – 50 °C and ca. 50 % excess of ethylene are preferred because these conditions allow much of the dichloroethane to be stripped from the reaction medium as it is formed.

For a continuous run with recycle at a chlorohydrin concentration of 6.4 %, a reactor temperature of 35 °C, an ethylene : chlorine ratio of 1.42, and a chlorine feed rate of 71 g/h, the average yield of ethylene chlorohydrin was 88.0 % with a dichloroethane yield of 10.1 %. Lowering the chlorine feed rate to 42 g/h raised the chlorohydrin yield slightly to 88.9 % and lowered the dichloroethane yield to 9.1 %.

For many years, ethylene chlorohydrin was manufactured on a large industrial scale as a *precursor to ethylene oxide.* This process has been almost completely supplanted by the direct oxidation of ethylene to ethylene oxide using silver catalysts (→ Ethylene Oxide). However, since other commercially important epoxides, such as propylene oxide and epichlorohydrin, cannot be made by direct oxidation of the parent olefin, chlorohydrin intermediates are still important in manufacturing these products. Although it is scarcely practiced today, a review of the ethylene chlorohydrin technology is valuable because of the close similarity of all of the chlorohydrin processes.

In a *typical industrial plant,* chlorine, ethylene, and water are concurrently passed upward through packed towers [22], [23]. These reactors are designed to minimize the contact between ethylene and gaseous chlorine and to provide adequate contact between the hydrocarbon feed and the liquid phase [24]. The feed lines are positioned a sufficient distance apart so that the chlorine can be completely dissolved in the water before the ethylene is introduced into the system. In some plants, there is a separate mixing column for chlorine and water followed by an ethylene reaction tower.

The reaction product is removed at the top of the tower as a 4.5 – 5.0 % solution of ethylene chlorohydrin, the yield of which may be as high

as 85 – 89 % of the ethylene converted. In addition to dichloroethane, small amounts of bis(2-chloroethyl) ether are formed as byproducts.

Propylene Chlorohydrin. The *hypochlorination of propene* gives two isomers:

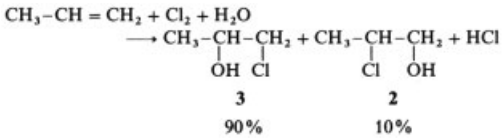

The major side reactions are chlorine addition to give 1,2-dichloropropane [*78-87-5*], and formation of bis(2-chloro-1-methylethyl) ether [*108-60-1*]:

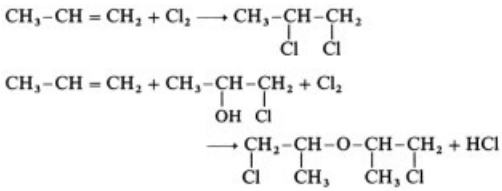

Propylene chlorohydrin reacts with a strong base, such as calcium hydroxide, to give *propylene oxide* [*75-56-9*] by dehydrochlorination (→ Propylene Oxide):

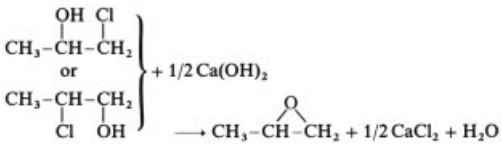

A typical flowchart for the production of propylene oxide through the chlorohydrin is shown in Figure 1 [25].

In the propylene chlorohydrin process, the separation of the feed points for chlorine and propene is even more critical than in the case with ethylene. The benefits of a recirculating-type, double-chamber reaction system have been described [26]. Chlorine was dissolved in a recycle stream of dilute aqueous propylene chlorohydrin together with makeup water. From this vessel, the effluent passed to a second chamber into which the propene was fed. This arrangement gave a propylene chlorohydrin yield of 87.5 % with 11.0 % dichloropropane and 1.5 % dichloroisopropyl ether. In contrast, a single-reactor system produced a 69.2 % chlorohydrin yield with 21.6 % dichloropropane and 9.2 % dichloroisopropyl ether.

Modern plants show propylene chlorohydrin yields of 87 – 90 % with 6 – 9 % dichloropropane. There are numerous patents that claim technology to boost chlorohydrin yields to 90 – 95 % [27–33].

A novel approach to hypochlorination involves the use of *tertiary alkyl hypochlorites* [34] and is based on the following chemistry:

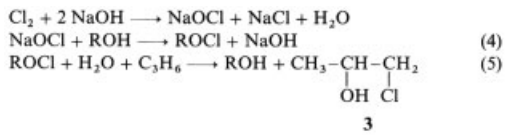

The preferred alcohol in Reaction (4) is *tert*-butanol. The resulting *tert*-butyl hypochlorite has very slight water solubility and may be separated by phase. Thus, the reaction of the hypochlorite with propene and water (Eq. 5) occurs in an environment essentially free of chloride ions to reduce formation of the dichloride. A 95.8 mol % selectivity to propylene chlorohydrin is claimed [34]. A diagram of this process is shown in Figure 2 [35].

Glycerol Monochlorohydrins. *Hypochlorination of allyl alcohol* gives glycerol monochlorohydrins:

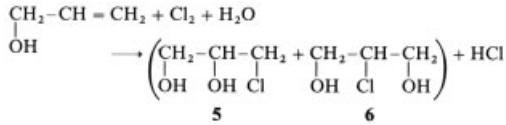

The reaction of allyl alcohol with chlorine and water at 50 – 60 °C is reported to give an 88 % yield of monochlorohydrins and 9 % dichlorohydrins [36].

Glycerol Dichlorohydrins. The *hypochlorination of allyl chloride* generates a mixture of the glycerol dichlorohydrins, 2,3-dichloro-1-propanol (**8**) and 1,3-dichloro-2-propanol (**7**), in about a 70 : 30 ratio.

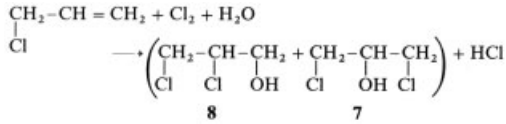

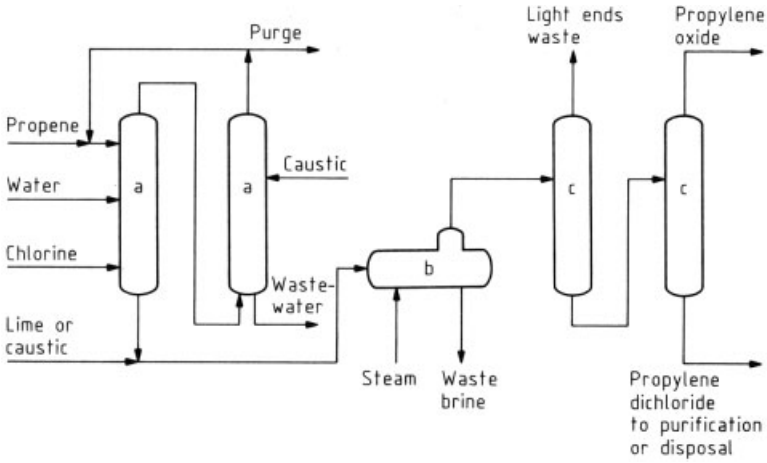

Figure 1. Typical arrangement for propylene oxide via the chlorohydrin route
a) Chlorohydrin absorber reactor; b) Chlorohydrin saponification and flashing; c) Propylene oxide purification train
Reproduced with permission from [25].

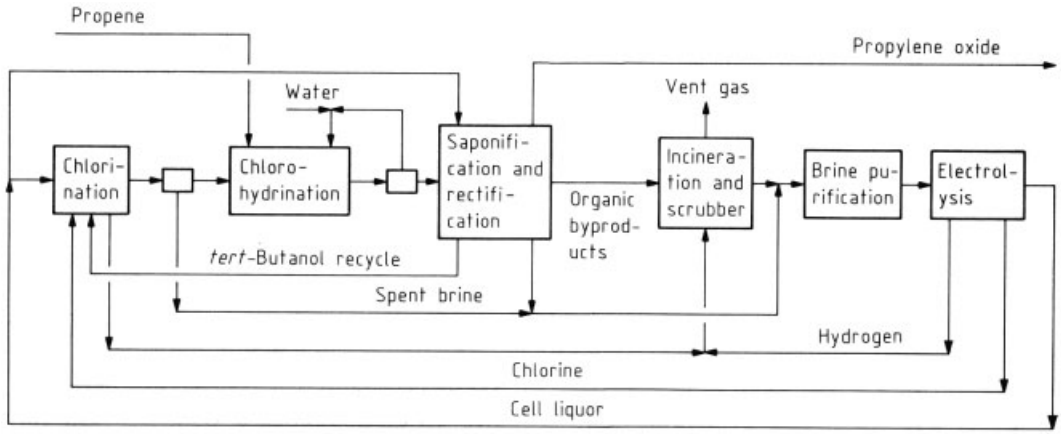

Figure 2. Hypochlorite route to propylene oxide
Reproduced with permission from [35].

These may be subsequently saponified to *epichlorohydrin* [*106-89-8*]:

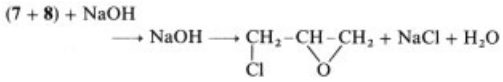

Because of the low water solubility of allyl chloride, it is essential to minimize formation of an organic phase, which leads to formation of the undesired byproduct 1,2,3-trichloropropane [24], [37]. This is commonly done by maintaining the allyl chloride as a fine dispersion by vigorous agitation in an aqueous system containing 10 – 50 volumes of water for each volume of allyl chloride [38–40]. Under such conditions, dichlorohydrin selectivities of 88 – 93 % are reported.

Another method of achieving a well-dispersed allyl chloride is by the use of a *surfactant* [41], [42]. Anionic or nonionic surfactants at a level of 0.2 – 0.5 wt % based on allyl chloride give reported dichlorohydrin yields of 93 – 95 % based on allyl chloride.

An alternative way of making dichlorohydrins is by *chlorination of allyl alcohol* [*107-18-6*], which may be produced by the hydrogenation of acrylaldehyde [*107-02-8*]. This route gives reported dichlorohydrin yields

of 94 – 98 % with a 2,3-dichloro-1-propanol yield of 88 – 97 % [43], [44].

Butene Chlorohydrins. Butylene chlorohydrin is prepared as a precursor to butene oxide. A laboratory study of the *hypochlorination of 2-butene* has been reported [45]. The dehydrochlorination of 3-chloro-2-butanol (**13**) with calcium hydroxide gave an 87 % yield of 2,3-butene oxide with 11 % glycol and 2 % methyl ethyl ketone [46].

Styrene Chlorohydrin. A laboratory preparation of styrene chlorohydrin (**19**) has been described [47]. Styrene was dispersed in water with a surfactant. Calcium hypochlorite solution was added to the vigorously stirred suspension through which carbon dioxide was bubbled. A 76 % yield of styrene chlorohydrin was obtained.

Long-Chain Chlorohydrins. Fatty acids containing unsaturation, such as oleic or elaidic acids, are converted to chlorohydrins as intermediates in making the epoxides or diols of these materials [48]. Long-chain olefins may be hypochlorinated with good selectivity (80 – 90 %) by using short reaction times, which result in low olefin conversion (5 – 25 %) [49].

Process Improvements. Some process improvements and new approaches for chlorohydrin production are the continuous production of glycerol dichlorohydrins [50], glycerol dichlorohydrins from allyl chloride [51], electrochemical chlorohydroxylation of olefins [52], 4-halo alcohols from ring cleavage of tetrahydrofuran [53], improved continuous production of aqueous propylene chlorohydrin solutions [54], high-yield continuous preparation of propylene chlorohydrin [55], ethylene chlorohydrin from ethanol [56], and chlorohydrins from immobilized halogenating enzymes [57], [58].

5. Environmental Protection

Unpublished data [59] for two of the chlorohydrin compounds are shown in Table 3 (there were no other data). Ethylene chlorohydrin is practically nontoxic to the fathead minnow according to a categorization system used by the U.S. Environmental Protection Agency. The two compounds in Table 3 are expected to be readily biodegradable.

Table 3. Environmental data on two chlorohydrins

	Biodegradation BOD/ThOD∗	Fathead minnow toxicity
Ethylene chlorohydrin	73 %	LC_{50}∗∗ > 100 mg/L
1,3-Butene chlorohydrin	67 %	no data

∗ BOD/ThOD = biochemical oxygen demand/theoretical oxygen demand.
∗∗ LC_{50} = lethal concentration to 50 % of the test organisms in a specified time.

6. Chemical Analysis

In addition to the conventional wet quantitative method, other modern spectroscopic analyses have been used, including Raman spectroscopy [60] and gas chromatographic analysis [61]. Potentiometric determination of propylene chlorohydrins [62] and enzymatic determination of ethylene chlorohydrin [63] were also reported.

7. Storage and Transportation

Chlorohydrins generally can be stored in acid-resistant tanks with a small or low-pressure rating. Titanium metal is resistant to corrosion by most chlorohydrins [64]. Potentiometric study of the corrosion of chlorohydrins up to 120 °C on titanium and steel has been reported [65]. Chlorohydrins in general are not classified as hazardous materials by the U.S. Department of Transportation. They can be shipped in acid tank railroad cars or trucks [64].

8. Uses

Chlorohydrins find their greatest utility as intermediates in the manufacture of epoxides, especially propylene oxide and epichlorohydrin. For reviews of these processes, see the following references: [25], [66–68], → Epoxides, → Ethylene Oxide, → Propylene Oxide.

Adducts of ethylene chlorohydrin with tungsten or molybdenum halides plus an organoaluminum compound have been patented as catalysts for olefin metathesis and cycloalkene ring-opening polymerization [69–72].

Chlorohydrins have also been used in the following ways:

– intermediate for plasticizers [73]
– wet strengthening agent for paper [74], [75]
– etherizing agent of phenolic resin to produce a high-temperature heavy-metal chelating compound [76]
– preservation of biological fluids and solutions [77]
– treatment of sulfonyl halides and dimethylamine aromatic compounds for fluorescent whitening agents and photosensitizers [78]
– raw material for lubricating oil intermediates and additives [79], [80]
– reaction with 1,3-bis(dimethylamino)-2-propanol to form a flocculant [81]
– production of low molecular mass epoxy resins [82]
– raw material for flame retardants for polymers [83]
– soil-resistant yarn-treating agent [84], [85]
– hardener for polyurethane elastomers [86]

Pharmaceutical and agricultural chemical applications have been explored. Chlorohydrins are used as base chemicals for various drugs and chemicals, such as the following:

– carbazole derivative for treatment of anxiety [87]
– pharmaceutically acceptable salts [88]
– quaternary ammonium polymeric antimicrobial agents [89]
– substituted nitrodiphenyl ethers [90]
– phenoxyalkanol antirhinovirus agents [91]
– quaternary salts of promethazine bronchoconstriction releasing agent [92]
– substituted oxobenzothiazolines plant growth regulators [93]
– phenyl urea derivative herbicides [94]

9. Toxicology and Occupational Health

Of the various chlorohydrins, ethylene chlorohydrin and α-chlorohydrin have been the most extensively studied, the former because it is a residue produced when poly(vinyl chloride) plastics are sterilized with ethylene oxide.

Ethylene Chlorohydrin. The American Conference of Governmental Industrial Hygienists TLV-TWA is 1 ppm (3 mg/m^3) as a ceiling limit with a skin notation, which indicates that the dermal route is significant. The MAK values are the same. The Occupational Safety and Health Act PEL is 5 ppm with a skin notation.

Ethylene chlorohydrin is moderate to high in acute oral toxicity with an LD_{50} of 89 mg/kg for rats [95]. It is also high in dermal toxicity as its dermal LD_{50} is 84 mg/kg in rabbits [95], a value that places ethylene chlorohydrin into the class B poison category for the U.S. Department of Transportation. A volume of ca. 5 cm^3 could be lethal to the average adult male (70 kg) if it contacts the skin and is not washed off immediately. Ethylene chlorohydrin is not particularly irritating to the skin or eyes; however, it causes slight to moderate irritation, depending on the circumstances. Excessive vapor concentration of ethylene chlorohydrin is attainable, and a single exposure at elevated concentration may cause death. One fatality due to a 2-h inhalation of ethylene chlorohydrin vapor at ca. 1 mg/L (ca. 306 ppm) has been reported [96]. Two fatal and several nonfatal cases of intoxication by ethylene chlorohydrin were investigated [97]. The average concentration in the nonfatal cases was 18 ppm. Nonfatal cases showed signs of circulatory shock, lack of coordination, repeated vomiting, epigastric pain, headache, heavy urine output, cough, and skin reddening. In another case report [98], autopsy findings showed damage to the lungs, liver, kidneys, brain, and other organs.

Ethylene chlorohydrin is considered to be *noncarcinogenic* on the basis of a test in which it was given to rats by subcutaneous injection for 1 year with a 6-month hold [99], and also on the basis of a skin painting test in rats and mice [100]. Ethylene chlorohydrin did not cause birth defects in rabbits when administered intravenously in doses up to 36 mg kg^{-1} d^{-1} during the period of organogenesis. When administered to mice at higher doses, 120 mg kg^{-1} d^{-1} caused a significant increase in the number of malformed fetuses and resorptions when accompanied by maternal toxicity [101]. There have

been no studies of ethylene chlorohydrin and its effect on male or female reproductive systems.

In the *Ames bacterial mutagenicity test,* ethylene chlorohydrin is sometimes positive and sometimes negative, with some of the positives only occurring after alteration by tissue homogenates (presumably creating metabolites that are known to be active) [102–104]. Ethylene chlorohydrin is also weakly and sporadically mutagenic when tested in animal cell systems. It is reported to have induced chromosome aberrations in rat bone marrow cells [105], but it did not produce a significant increase in the number of translocation heterozygotes (heritable translocation) when administered to male mice at 30 – 60 mg kg^{-1} d^{-1} for 5 weeks [106]. It was also negative in tests with fruit flies [107].

The toxicity of ethylene chlorohydrin (and presumably other activity, such as mutagenicity) is believed to be due to its *conversion to chloroacetaldehyde* in vivo. Chloroacetaldehyde and chloroacetic acid, which are metabolic intermediates, are to a great degree conjugated and inactivated by reaction with glutathione. Thus, when ethylene chlorohydrin levels are highly elevated, glutathione is depleted and chloroacetaldehyde cannot be detoxified. When chloroacetaldehyde levels at intracellular sites are elevated, other cell components are alkylated, damage to major organs occurs, and toxic symptoms are produced [108], [109].

α-Chlorohydrin. This member of the chlorohydrin series has also been extensively studied for its toxicological effects. Its chronic toxicological profile differs greatly from that of ethylene chlorohydrin. There is apparently no industrial hygiene exposure guide for α-chlorohydrin. α-Chlorohydrin is also high in acute oral toxicity, with an LD_{50} of 55 mg/kg in rats [110], and is apparently also high in dermal toxicity. The rabbits placed on skin irritation tests all died with one small application, and one was examined and found to have "widespread liver damage" [59]. The material was essentially nonirritating to skin. The vast majority of toxicological information on α-chlorohydrin is from investigation of its effect on male reproduction. This antifertility effect, which α-chlorohydrin has on the male, was demonstrated in 1969 and 1970 [111], [112], and has been shown to be predictable for rats by oral, injection, and inhalation routes.

Because of the potential utility of α-chlorohydrin as a *male contraceptive*, its mechanism of action has been thoroughly studied. Only the S isomer has been determined to possess antifertility activity; it has none of the detrimental effect on the kidneys associated with the R isomer [113]. α-Chlorohydrin and epichlorohydrin produce similar effects on the male reproductive system of the rat, i.e., epididymal sperm granulomas, spermatocoeles, and an increase in the number of morphologically abnormal spermatozoa [114]. There is evidence that the profile of action may be slightly different in primates [115].

Like ethylene chlorohydrin, α-chlorohydrin is *weakly mutagenic* in some tests [116], but α-chlorohydrin may be a direct-acting mutagen.

α-Chlorohydrin was tested for *carcinogenicity* and gave ambiguous results in an oral test using rats. Toxicity (testicular effects) was noted at the two dose levels used; and although not statistically significant, there were tumors of the parathyroid that were in question in this test [117]. In mice, α-chlorohydrin produced no carcinogenic effects via skin painting and subcutaneous injection [118].

Other Chlorohydrins. Data for other chlorohydrins are primarily of the acute variety. *1-Chloro-2-propanol* is moderate in acute oral toxicity with an LD_{50} of 100 – 300 mg/kg in rats and moderate in acute dermal toxicity with an LD_{50} of ca. 500 mg/kg in rabbits. It is only slightly irritating to the skin, and slightly to moderately irritating to the eyes [59]. The Dow Industrial Hygiene Guide is 3 ppm as a ceiling limit with a skin notation and is based on the following: (1) acute data which showed that rats were able to survive at most a 15-min exposure to a saturated atmosphere (nominal concentration 13 000 ppm) [59]; (2) a 22-week intubation study in rats, which showed only slight liver effects (organ weight increase) at 25 mg kg^{-1} d^{-1} [119]; and (3) an inhalation test in rats, in which 14 exposures to 30 ppm caused no adverse effects, but higher concentrations caused lung congestion and lethargy [120].

A *mixture of 2-chloro-1-propanol and 1-chloro-2-propanol* was mutagenic in the Ames strains TA1535 and TA100 with and without

metabolic activation but with enhanced activity after activation. The mixture was also mutagenic in a mouse lymphoma assay and in a rat cytogenetic test [121]. Such a mixture was also fed to rats and dogs for 90 days at doses ranging from 1 to 50 mg kg^{-1} d^{-1} [122]. No significant effects occurred in rats at any dose nor in dogs at 1 or 7 mg kg^{-1} d^{-1}. Only reduced food consumption, reduced body weight, and increased relative liver weight occurred at 50 mg kg^{-1} d^{-1} in dogs.

1,3-Dichloro-2-propanol is moderate in acute oral toxicity with an LD_{50} of between 126 and 252 mg/kg in rats, is moderate in acute dermal toxicity with an LD_{50} of between 500 and 1000 mg/kg in rabbits, and is moderately irritating to the eyes and moderately to severely irritating to the skin [59]. In inhalation tests, a 4-h exposure to 1000 ppm was lethal to five of five rats, while no rats died as a result of inhaling 1000 ppm for 2 h. Rats were able to survive a 4-h exposure at 300 ppm but not 7 h at 300 ppm [59]. This chlorohydrin was mutagenic toward Ames tester strains TA100 [116] and TA1535 [123] and in other such Ames tests.

2,3-Dichloro-1-propanol is moderate to high in acute oral toxicity with an LD_{50} of 90 mg/kg in rats, moderate in acute dermal toxicity with an LD_{50} of 0.2 mL/kg in rabbits, and also moderately irritating to the eyes and skin [124]. It is mutagenic toward Ames strain TA1535 [123] and in other Ames tests.

There are no toxicity data for either 3-chloro-2-butanol or 2-chloro-1-butanol (this should not be confused with chlorobutanol, which is the anesthetic having the full name trichloro-*tert*-butyl alcohol). There are also no data for the other glycerol monochlorohydrin, 2-chloro-1,3-propanediol.

Acute data indicate that *styrene chlorohydrin* is low in acute oral toxicity with an LD_{50} between 500 and 2000 mg/kg in rats [59], but there is no dermal LD_{50} available. Styrene chlorohydrin is quite irritating to the eyes and skin. Inhalation presents no hazard from exposure at room temperature; however, if the material is heated, the vapors thus generated may cause irritation and possibly organ effects (kidney and liver) [59].

10. References

1. G. C. Swain, A. D. Ketley, R. F. W. Bader, *J. Am. Chem. Soc.* **81** (1959) 2353.
2. S. Carra et al., *Chem. Eng. Sci.* **34** (1979) 1123.
3. I. K. Gregor, N. V. Riggs, V. R. Stimson, *J. Chem. Soc.* 1956, 76.
4. N. K. Bliznyuk et al., *Zh. Obshch. Khim.* **36** (1966) 480; *Chem. Abstr.* **65** (1966) 638 d.
5. Corn Products Co., US 2 876 217, 1959 (E. F. Paschall).
6. I. G. Rodier, *Bull. Soc. Chim. Fr.* 1948, 637.
7. H. V. R. Iengar, P. D. Ritchie, *J. Chem. Soc.* 1957, 2556.
8. Phillips Petroleum Co., US 3 923 842, 1975 (Y. Wu).
9. J. Plucinski, *Rocz. Chem.* **41** (1967) 1135; *Chem. Abstr.* **68** (1968) 104 329 j.
10. V. A. Krishnamurthy, M. R. A. Rao, *J. Indian Inst. Sci.* **40** (1958) 145; *Chem. Abstr.* **53** (1960) 19 860 g.
11. DD 105 207, 1974(E. Fibitz, B. Nussbuecker); *Chem. Abstr.* **81** (1974)151 522 v.
12. N. Parris, J. K. Weil, W. M. Linfield, *J. Am. Oil Chem. Soc.* **53** (1976) 97.
13. I. Shinji, *J. Soc. Chem. Jpn.* **45** (1942) 359.
14. M. F. Shostakovskii, T. T. Minakova, F. P. Sidel'kovskaya, *Izv. Akad. Nauk SSSR, Ser. Khim.* 1964, 2197; *Chem. Abstr.* **62** (1965) 8996 d.
15. B. L. Phillips, P. A. Argabright, *J. Heterocycl. Chem.* **3** (1966) 84.
16. Du Pont, US 2 455 405, 1948 (L. A. Burrows, M. F. Fuller).
17. L. Carius, *Ann. Chem. Pharm.* **126** (1863) 195.
18. J. F. Norris, *Ind. Eng. Chem.* **11** (1919) 817.
19. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* Wiley-Interscience, New York 1962, p. 446.
20. M. Gomberg, *J. Am. Chem. Soc.* **41** (1919) 1414.
21. W. G. Domask, K. A. Kobe, *Ind. Eng. Chem.* **46** (1954) 680.
22. F. Asinger in B. J. Hazzard (trans.): *Mono-Olefins, Chemistry and Technology,* Pergamon Press, Oxford 1968, pp. 556 – 568.
23. D. G. Weaver, J. L. Smart, *Ind. Eng. Chem.* **51** (1959) 894.
24. P. W. Sherwood, *Pet. Eng.* **27** (1955) C – 41.
25. R. B. Stobaugh, V. A. Calarco, R. A. Morris, L. W. Stroud, *Hydrocarbon Process.* **52** (1973) no. 1, 99 – 108.
26. P. Ferrero, L. R. Flamme, M. Fourez, *Ind. Chim. Belg.* **19** (1954) 113.

27. Shell, BE 630 446, 1963.
28. Bayer, FR 1 357 443, 1964.
29. Naphthachimie, DE-OS 2 101 119, 1971 (R. Bouchet).
30. BASF, DE-OS 2 022 819, 1971 (E. Bartholomé et al.).
31. G. Mikula, CS 152 702, 1974; *Chem. Abstr.* **81** (1974) 63135 r.
32. Petrocarbon Developments, FR 2 194 673, 1974.
33. Mitsui Toatsu Chemicals, JP-Kokai 77 48 606, 1977 (Y. Watanabe et al.).
34. Lummus, US 4 008 133, 1977 (A. P. Gelbein, J. T. Kwon).
35. *Hydrocarbon Process.* **59** (1979) 239.
36. J. Myszkowski, A. Z. Zielinski, *Przem. Chem.* **44** (1965) 249; *Chem. Abstr.* **63** (1965) 5516 e.
37. Shell, US 2 605 293, 1952 (T. T. Tymstra).
38. Shell, US 2 714 121, 1955 (J. Anderson, G. F. Johnson, W. C. Smith).
39. T. Reis, FR 1 557 589, 1969.
40. I. Ondrus, P. Klucovsky, CS 120 030, 1966; *Chem. Abstr.* **67** (1967) 63 741 w.
41. Tokuyama Soda Co., JP 73 21 924, 1973 (Y. Onoue, Y. Fujii, M. Ueda).
42. Euteco, GB 2 029 821, 1980 (C. Divo, M. Petri, M. Lazzari, A. Bigozzi).
43. Hoechst, DE-OS 2 007 867, 1971 (D. Freudenberger, H. Fernholz).
44. Nippon Soda Co., JP 73 18 207, 1973 (Y. Enoki et al.).
45. A. Z. Zielinski, J. Myszkowski, A. Czubowiez, *Chem. Tech. (Leipzig)* **14** (1962) 456; *Chem. Abstr.* **58** (1963) 3305 c.
46. J. Myszkowski, A. Z. Zielinski, *Przem. Chem.* **43** (1964) 324; *Chem. Abstr.* **61** (1964) 14 611 a.
47. W. S. Emerson, *J. Am. Chem. Soc.* **67** (1945) 516.
48. A. W. Ralston: *Fatty Acids and Their Derivatives,* J. Wiley & Sons, New York 1948, pp. 450 – 452.
49. Procter & Gamble Co., US 3 598 847, 1971 (R. K. Kloss, G. W. Claybaugh, D. D. Whyte).
50. G. S. Sharifov et al., DE-OS 2 951 770, 1981.
51. Euteco SpA, GB-A 2 029 821, 1980 (G. Divo et al.).
52. V. Kazarinov et al., *Collect. Czech. Chem. Commun.* **47** (1982) no. 11, 2849 – 2857.
53. Texaco Development Corp., DE-OS 2 952 277, 1980 (K. G. Hammond).
54. VEB Chemische Werke Buna, DD 144 907, 1980 (K. H. Kuessner et al.).
55. BASF, DE-OS 2 910 675, 1980 (W. Koehler et al.).
56. Financiadora de Estudos e Projetos (FINEP); Scientia-Engenharia de Sistemas, Desenvolvimento de Prototipos e Processos, Braz. Pedido PI BR 82 312 A, 1983 (A. Telles).
57. Cetus Corp., US 4 247 641, 1981 (S. L. Neidleman et al.).
58. J. Geigert et al., *Appl. Environ. Microbiol.* **45** (1983) no. 2, 366 – 374.
59. Unpublished results, Dow Chemical.
60. A. Gupta et al., *Spectrochim. Acta, Part A* **36 A** (1980) no. 6, 601 – 606.
61. S. A. Shakun et al., *Khim. Prom.-st., Ser.: Metody Anal. Kontrolya Kach. Prod. Khim. Prom-sti.* **1** (1980) 12 – 14; *Chem. Abstr.* **93** (1980) 88 211 y.
62. T. V. Shevchenko et al., *Khim. Prom.-St., Ser.: Metody Anal. Kontrolya Kach. Prod. Khim. Prom.-Sti.* **3** (1980) 43 – 46; *Chem. Abstr.* **93** (1980) 36 395 s.
63. American Sterilizer Co., US 4 162 942, 1979 (D. A. Gunther).
64. A. E. Romanushkina, *Zashch. Met.* **10** (1974) no. 1, 39 – 40.
65. L. I. Komarova, *Ispol'z. Sovrem. Fiz. Khim. Metodov Issled. Protsessov. Prod. Khim. Uglekhim. Proizvod. Tedzizy Dokl. Nauchno Tekh. Konf.* 1976, 96 – 97; *Chem. Abstr.* **89** (1978) no. 22, 187 881 r.
66. K. H. Simmrock, *Hydrocarbon Process.* **57** (1978) no. 11, 105 – 113.
67. Lummus, US 4 008 133, 1977 (A. P. Gelbein, J. T. Kwon).
68. "Glycerol (Epichlorohydrin)," *Pet. Refiner* **38** (1959) no. 11, 252 – 253.
69. Bayer, DE-OS 2 056 178, 1972 (D. Maertens, J. Witte, M. Beck).
70. Bayer, NL-A 7 408 469, 1975 (G. Lehnert, D. Maertens, G. Pampus, J. Witte).
71. Bayer, DE-OS 2 332 565, 1975 (G. Pampus, G. Lehnert, J. Witte).
72. Bayer, DE-OS 2 357 193, 1975 (J. Witte, G. Lehnert).
73. A. Courtier, FR 2 144 084, 1973.
74. Hercules, EP-A 416 511, 1981 (R. A. Bankert).
75. G. Troemel, *Farben Rev. USA* 1970, no. 18, 55 – 63.
76. Unitika, JP-Kokai-A2 145 719, 1983.
77. *Chem. Mark. Rep.* 1984, 28.
78. Mita Industrial Co., JP-Kokai 29 335, 1979 (S. Torii).
79. Chevron Research Co., GB 2 108 485, 1983 (V. R. Small).
80. Texaco Development Corp., FR 2 460 297, 1981 (K. G. Hammond).

81. Kewanee Industries, US 4 188 293, 1980 (H. A. Green).
82. Dow Chemical, US 4 408 062, 1983 (J. L. Bertram et al.).
83. Monsanto, US 4 179 483, 1979 (G. H. Birum et al.).
84. Allied, US 4 317 736, 1980 (R. M. Marshall).
85. Allied, DE-OS 3 304 351, 1983 (R. H. Thomas).
86. Ihara Chemical Industry Co., US 4 342 859, 1982 (T. Harada et al.).
87. Burroughs Wellcome Co., US-A 4 400 383, 1983 (J. R. T. Davison).
88. Wellcome Foundation, JP-Kokai 109 482, 1983.
89. Kewanee Industries, US-A 4 304 910, 1981 (H. A. Green).
90. Rohm & Haas Co., EP-A 21 692, 1981 (C. Swithenbank et al.).
91. Richardson-Merrell, BE 887 423, 1981 (R. A. Parker).
92. Am. Home Products Corp., US 4 183 912, 1980 (M. E. Rosenthale).
93. Monsanto, JP-Kokai 151 966, 1979.
94. Shell Oil Co., US 4 226 612, 1980 (K. H. Pilgram).
95. *Gig. Sanit.* **43** (1978) no. 8, 13 (in NIOSH Registry of Toxic Effects of Chemical Substances, 1983).
96. H. Dierker, P. G. Brown, *J. Ind. Hyg. Toxicol.* **26** (1944) 277 – 279.
97. M. W. Goldblatt, W. E. Chiesman, *Br. J. Ind. Med.* **1** (1944) 207.
98. V. Miller, R. J. Dobbs, S. I. Jacobs, *Arch. Dis. Child.* **45** (1970) 589 – 590.
99. T. Balazs, *FDA By-Lines* **3** (1976) 150.
100. National Toxicology Program (NTP) Technical Report on the Toxicology and Carcinogenesis Studies of 2-Chloroethanol (Ethylene Chlorohydrin) in F344/N Rats and Swiss CD – 1 Mice (Dermal Studies), NTP TR 275, Board Draft, Jun. 1, 1984.
101. J. B. LaBorde, C. A. Kimmel, C. Jones-Price, T. A. Marks et al., *Toxicologist* **2** (1982) no. 1, 71.
102. H. S. Rosenkranz, T. J. Wlodkowski, *J. Agric. Food Chem.* **22** (1974) 407.
103. J. McCann, V. Simmon, D. Streitwieser, B. N. Ames, *Proc. Natl. Acad. Sci.* U.S.A. 72 (1975) no. 8, 3190 – 3193.
104. J. D. Elmore, J. L. Wong, A. D. Laumbach, U. N. Streips, *Biochim. Biophys. Acta* **442** (1976) no. 3, 405 – 419.
105. V. N. Semenova, N. Y. Ramul, *Proc. Nat. Acad. Sci.* **72** (1977) no. 8, 179 – 181.
106. C. W. Sheu, K. T. Cain, R. M. Grydev, W. M. Generoso, *J. Am. Coll. Toxicol.* **2** (1983) no. 2, 221 – 223.
107. A. G. A. C. Knaap, C. E. Voogd, P. G. N. Kramers, *Mutat. Res.* **101** (1982) 199 – 208.
108. W. Grunow, H. J. Altmann, *Arch. Toxicol.* **49** (1982) 275 – 284.
109. W. H. Lawrence, K. Itoh, J. E. Turner, J. Autian, *J. Pharm. Sci.* **60** (1971) 1163 – 1168.
110. R. Paul, R. P. Williams, E. Cohen, *Contraception* **9** (1974) 451 – 457 (in NIOSH Registry of Toxic Effects of Chemical Substances, 1981 – 1982).
111. J. A. Coppola, *Life Sci.* **8** (1969) no. 1, 43 – 48.
112. R. J. Ericsson, V. F. Baker, *J. Reprod. Fertil.* **21** (1970) 267 – 273.
113. K. E. Porter, A. R. Jones, *Chem.-Biol. Interact.* **41** (1982) no. 1, 95 – 104.
114. W. M. Kluwe, B. N. Gupta, J. C. Lamb, *Toxicol. Appl. Pharmacol.* **70** (1983) no. 1, 67 – 86.
115. I. Braz, L. N. Shandilya, L. S. Ramaswami, *Andrologia* **8** (1976) no. 4, 290 – 296.
116. S. J. Stolzenberg, C. H. Hine, *J. Toxicol. Environ. Health* **5** (1979) no. 6, 1149 – 1158.
117. E. K. Weisburger, B. M. Ulland, J.-M. Nam, J. J. Gart et al., *J. Natl. Cancer Inst.* **67** (1981) no. 1, 75 – 88.
118. B. L. Van Duuren, B. M. Goldschmidt, C. Katz, I. Seidman et al., *J. Natl. Cancer Inst.* **53** (1974) no. 3, 695 – 700.
119. FDA Report, 1969.
120. J. C. Gage, *Br. J. Ind. Med.* **27** (1970) 1 – 18.
121. R. W. Biles, C. E. Piper, *Toxicologist* **1** (1981) no. 1, 41.
122. I. W. Daly, R. D. Phillips, *Toxicologist* **1** (1981) no. 1, 13.
123. L. A. Silhankova, F. Smid, M. Cerna, J. Davidek et al., *Mutat. Res.* **103** (1982) 77 – 81.
124. H. F. Smyth, C. P. Carpenter, *J. Ind. Hyg. Toxicol.* **30** (1948) no. 1, 63 – 68.

Chloronaphthalenes → Chlorinated Hydrocarbons

Chlorophenols

François Muller, Rhône-Poulenc, Le Pont-de-Claix, France (Chaps. 2 – 10)

Liliane Caillard, Rhône-Poulenc, Courbevoie, France (Chap. 11)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
3.1. Reactions of the Hydroxyl Group . . 2
3.2. Electrophilic Substitution of the Aromatic Ring 2
3.3. Nucleophilic Substitution of the Aromatic Ring 3
3.4. Oxidation Reactions 3
4. Production 3
4.1. Mono-, Di-, and Trichlorophenols with no Chlorine in a Meta Position; Tetrachlorophenols; and Pentachlorophenol 3
4.2. Mono-, Di-, and Trichlorophenols with Chlorine in a Meta Position . . 4
5. Analysis 4
6. Quality Specifications 5
7. Economic Aspects 6
8. Uses 6
9. Storage and Transportation 6
10. Environmental Considerations 6
10.1. Toxicity to Fauna and Flora 6
10.2. Persistence in the Environment . . . 7
11. Toxicology 7
12. References 8

1. Introduction

In chlorophenols the aromatic ring of phenol is substituted with one to five chlorine atoms. Laurent discovered this group of compounds when he chlorinated coal tar in 1836 [1], [2].

Chlorophenols are industrially important because of their broad spectrum of antimicrobial properties and their uses as fungicides, herbicides, insecticides, ovicides, and algicides. A mixture of tetrachlorophenols and pentachlorophenol, either as such or as salts or esters, finds widespread use as a preservative for wood, glue, paint, vegetable fibers, and leather. In addition, chlorophenols are widely used as intermediates in chemical syntheses.

2. Physical Properties

Important characteristics and physical properties of chlorophenols are shown in Tables 1 and 2. With the exception of 2-chlorophenol, these compounds are solid at ambient temperature; they all have strong characteristic odors. Several of them sublime readily. They can be steam-distilled. The chlorophenols are only slightly soluble in water, but highly soluble in alcohols. They impart a strongly disagreeable taste to drinking water.

3. Chemical Properties

Chlorophenols are versatile intermediates in chemical syntheses because both the hydroxyl group and the aromatic ring can react by both electrophilic and nucleophilic substitution. They are readily oxidized.

3.1. Reactions of the Hydroxyl Group

Chlorophenols are more acidic than phenol and readily form salts with alkali and alkaline-earth hydroxides or carbonates or with amines. These salts are stable and soluble in water.

Chlorophenols readily form esters of both organic and mineral acids [3–5]. Esters such as chlorophenyl phosphates, carbamates [6], [7], chloroformates [8], carbonates, and laurates are widely used. Pentachlorophenyl 2-ethylhexanoate is used as a textile preservative.

Chlorophenols are also readily O-alkylated, and ethers derived from aliphatic, aromatic, and heterocyclic reagents comprise a broad class of industrial products [9], [10]. One of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_001

Table 1. Characteristics of the chlorophenols

Chlorophenol	Formula	M_r	*mp*, °C	*bp*, °C	pK_a
2-Chlorophenol [*95-57-8*]	C_6H_5ClO	128.56	8.7	174.5	8.52
3-Chlorophenol [*108-43-0*]	C_6H_5ClO	128.56	32.8	216	8.97
4-Chlorophenol [*106-48-9*]	C_6H_5ClO	128.56	42.8	217	9.37
2,3-Dichlorophenol [*576-24-9*]	$C_6H_4Cl_2O$	163.00	58	206	7.71
2,4-Dichlorophenol [*120-83-2*]	$C_6H_4Cl_2O$	163.00	45	210	7.90
2,5-Dichlorophenol [*583-78-8*]	$C_6H_4Cl_2O$	163.00	59	210	7.51
2,6-Dichlorophenol [*87-65-0*]	$C_6H_4Cl_2O$	163.00	66	220	6.78
3,4-Dichlorophenol [*95-77-2*]	$C_6H_4Cl_2O$	163.00	65	253	8.62
3,5-Dichlorophenol [*591-35-5*]	$C_6H_4Cl_2O$	163.00	68	233	8.25
2,3,5-Trichlorophenol [*933-78-8*]	$C_6H_3Cl_3O$	197.45	62	248 – 249	
2,4,5-Trichlorophenol [*95-95-4*]	$C_6H_3Cl_3O$	197.45	68	245 – 246	6.72
2,4,6-Trichlorophenol [*88-06-2*]	$C_6H_3Cl_3O$	197.45	68	246	5.99
2,3,4-Trichlorophenol [*15950-66-0*]	$C_6H_3Cl_3O$	197.45	83.5	subl.	6.97
3,4,5-Trichlorophenol [*609-19-8*]	$C_6H_3Cl_3O$	197.45	101	275	7.55
2,3,6-Trichlorophenol [*933-75-5*]	$C_6H_3Cl_3O$	197.45	101	272	5.80
2,3,5,6-Tetrachlorophenol [*935-95-5*]	$C_6H_2Cl_4O$	231.89	115	subl.	5.03
2,3,4,6-Tetrachlorophenol [*58-90-2*]	$C_6H_2Cl_4O$	231.89	69 – 70	164 (3 kPa)	5.22
2,3,4,5-Tetrachlorophenol [*4901-51-3*]	$C_6H_2Cl_4O$	231.89	116 – 117	subl.	5.64
Pentachlorophenol [*87-86-5*]	C_6HCl_5O	266.34	190	309 – 310	4.74

Table 2. Physical properties of the most important chlorophenols

Property	2-Chlorophenol	4-Chlorophenol	2,4-Dichlorophenol	2,6-Dichlorophenol	2,4,6-Trichlorophenol	Pentachlorophenol
bp, °C						
at 0.1333 kPa	12	50	53	59.5	76.5	132
at 1.333 kPa	51	92	93	101	120.2	178
at 5.333 kPa	82	125	123	131.6	152.2	
at 13.33 kPa	106	150	146	154.6	177.8	
at 53.33 kPa	150	190	187	197.7	222.5	
Specific heat at 20 °C, $J \cdot mol^{-1} K^{-1}$	172	177	190			
Heat of formation at 50 °C, kJ/mol	185	198.9				295
Viscosity at 50 °C, mPa · s	2	5	2.65			
Solubility in water at 20 °C, g/L	28.5	27.1	4.5		0.8	0.014
Flash point (closed cup), °C	121	84	113			

the most widely known is 2,4-dichlorophenoxyacetic acid [*95-75-7*] (2,4-D), which is still used as a selective herbicide for broadleaf plants [11–13] (→ Chlorophenoxyalkanoic Acids).

3.2. Electrophilic Substitution of the Aromatic Ring

Electrophilic substitution is favored by the presence of chlorine atoms on the aromatic nucleus; thus, most chlorophenols partake in the standard reactions, such as alkylation, acylation, nitration, sulfonation, halogenation, [14–16], carboxylation, and condensation with aldehydes, e.g., formaldehyde, chloral, glyoxylic acid [17–19].

3.3. Nucleophilic Substitution of the Aromatic Ring

Nucleophilic substitution for one or more of the chlorine atoms, although disfavored by the presence of the other chlorine atoms, is nevertheless used widely.

The hydrolyses of 2-chlorophenol and 4-chlorophenol were used at one time to obtain pyrocatechol [*120-80-9*] and hydroquinone [*123-31-9*], respectively. The ammonolysis of 2-chlorophenol or 4-chlorophenol produces 2-aminophenol or 4-aminophenol, respectively. This type of reaction is used primarily to prepare various substituted diphenyl ethers, which serve as efficient herbicides [20–22].

Chlorophenols can be converted into poly-chlorinated alicyclic ketones such as "hexachlorophenol" [*599-52-0*] by reac-

tion of chlorine in aqueous acetic acid. "Hexachlorophenol" has pesticidal properties and is hydrolyzed to chloranil [*2435-53-2*] in the presence of an acidic catalyst such as sulfuric acid.

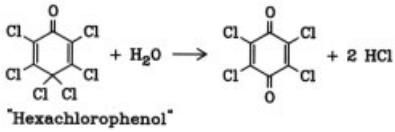

3.4. Oxidation Reactions

Like other phenols, chlorophenols can be oxidized; air oxidation forms chloroquinones, whereas oxidation with hydrogen peroxide leads to the corresponding chlorinated diphenol. Oxidation by a peroxide such as perpropionic acid in benzene at 80 °C yields chlorodiphenols (85 % yield) [23]:

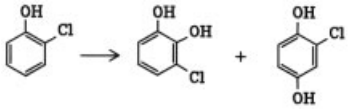

4. Production

There are two types of chlorophenols: (1) those with one to three chlorine atoms but none in a meta (3 or 5) position relative to OH and those which have four or five chlorine atoms, and (2) those with one to three chlorine atoms, at least one being in a meta position relative to OH.

4.1. Mono-, Di-, and Trichlorophenols with no Chlorine in a Meta Position; Tetrachlorophenols; and Pentachlorophenol

The mono-, di-, and trichlorophenols which have no chlorine atom in the meta (3 or 5) position relative to OH are 2-chlorophenol, 4-chlorophenol, 2,4-dichlorophenol, 2,6-dichlorophenol, and 2,4,6-trichlorophenol. The tetrachlorophenols and pentachlorophenol are also included with this type. These compounds are obtained by the direct chlorination of phenol.

The chlorination of melted phenol [*108-95-2*] using chlorine gas [*7782-50-5*] yields a mixture of ortho and para monochlorophenols with an *o*/*p* ratio of ca. 0.54 (35/65), which is scarcely dependent on temperature. The distribution between the two isomers depends especially on the nature of the chlorination reagent and the characteristics of the reaction medium. The formation of 2-chlorophenol is favored by an aprotic nonpolar solvent medium, such as hexane, carbon tetrachloride, or dichloroethane. The *o*/*p* ratio is also higher for reactions with a low concentration of phenol and a high temperature [24]. The chlorination of a 5 % solution of phenol in perchloroethylene at 110 °C in the presence of 0.01 wt % diisopropylamine [*108-18-9*] gives a very high *o*/*p* ratio, ca. 15 [25]. Chlorinations using aqueous sodium hypochlorite [*7681-52-9*] at pH 10, *tert*-butyl hypochlorite [*507-40-4*] and alkali phenolates dispersed in an anhydrous medium [26], or chlorine in acetic acid in the presence of acetic anhydride [*108-24-7*] also favor the formation of 2-chlorophenol [27].

The formation of 4-chlorophenol is favored by chlorination in a polar medium, such as acetonitrile, nitrobenzene, or ether [28–31]. Substitution in the para position is also strongly favored by the use of sulfuryl chloride [*7791-25-5*] as a chlorinating agent ($o/p = 0.35$) [28], [31]. Certain organic sulfides, e.g., diphenyl sulfide [*139-66-2*], activate chlorination and favor substitution at the para position [32]. Chlorination using copper(II) chloride in aqueous hydrochloric acid in the presence of oxygen leads to 4-chlorophenol (almost 100 %) [33].

Mono-, di-, and trichlorophenols are made industrially by chlorination of phenol with chlorine gas in a melt in cast-iron reactors. The isomer distribution in the product may be adjusted to market demands by the chlorination level and by recycling various intermediates. The yield, based on chlorine and phenol, is greater than 95 %.

Tetrachlorophenols and pentachlorophenol are produced batchwise in nickel reactors by the chlorination of less halogenated chlorophenols in the presence of aluminum trichloride [*7446-70-0*] or iron trichloride [*7705-08-0*].

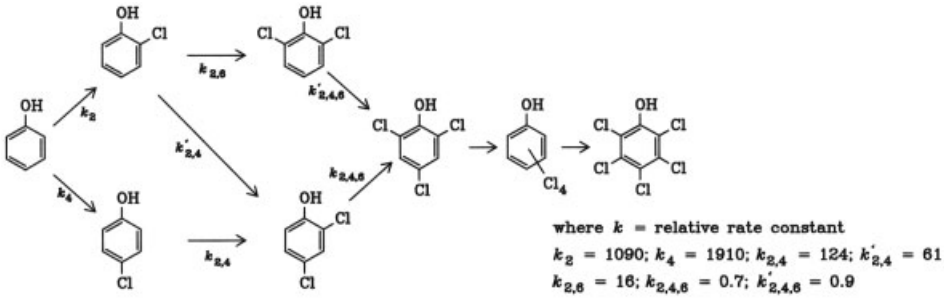

4.2. Mono-, Di-, and Trichlorophenols with Chlorine in a Meta Position

The mono-, di-, or trihalogenated phenols having at least one chlorine atom in a meta position are 3-chlorophenol, 2,3-dichlorophenol, 2,5-dichlorophenol, 3,4-dichlorophenol, 3,5-dichlorophenol, 2,3,4-trichlorophenol, 3,4,5-trichlorophenol, and 2,3,6-trichlorophenol.

These compounds cannot be obtained by the chlorination of phenol, but must be made by other types of reactions, such as hydrolysis, sulfonation, hydrodechlorination, hydroxylation, and alkylation.

Hydrolysis of chlorobenzenes can be carried out in alkaline aqueous media, with or without the presence of copper salts [34]. It can be done in the vapor phase at 250 – 400 °C on phosphates [35] or on silica at 500 – 550 °C [36]. Hydrolysis is most frequently carried out in the presence of methanolic sodium or potassium at 150 – 200 °C and 1.0 – 3.0 MPa [37], [38].

Sulfonation – Desulfonation. Sulfonation of chlorobenzenes followed by the desulfonation of chlorobenzenesulfonic acids can be used to produce numerous chlorophenols [39–44]. The chlorobenzenes are sulfonated by using 15 – 20 % oleum at 60 – 80 °C. The subsequent desulfonation is usually carried out by using 15 – 25 % aqueous sodium hydroxide solution at 180 –220 °C and 1.5 – 3.0 MPa.

Sandmeyer Reaction. *m*-Chlorophenols can also be made by diazotization of the corresponding chloroanilines that have been obtained either by nitration of chlorobenzenes followed by hydrogenation or by chlorination of nitrobenzenes [42].

Hydrodechlorination. Hydrodechlorination of polychlorophenols produces 3-chlorophenols. This reaction has long been carried out either in the vapor phase at 300 – 350 °C on an alumina catalyst [45] or in the liquid phase at 150 – 250 °C and 15.0 – 20.0 MPa, usually in the presence of sulfur [46], [47]. More recently, *catalytic hydrodechlorination* of polychlorophenols in organic or aqueous media with a palladium catalyst was used to obtain 3-chlorophenol and 3,5-dichlorophenol [48]. Hydrodechlorination of polychlorophenols with hydriodic acid in an aqueous or organic medium has also been reported [49].

Hydroxylation. Hydroxylation of chlorobenzenes gives a mixture of 55 wt % 4-chlorophenol, 30 – 35 wt % 2-chlorophenol, and 15 wt % 3-chlorophenol [50].

Hock Phenol Synthesis. The propylene alkylation of a chlorobenzene followed by oxidation and decomposition of the hydroperoxide in an acid medium produces the corresponding chlorophenol [46], [51].

The physical properties of the principal commercial chlorophenols are given in Table 2.

5. Analysis

The acidity of the phenol function in chlorophenols increases with increasing chlorination of the benzene ring. Potentiometric analysis with either soda in an aqueous medium or an amine in an organic medium is, therefore, suitable. Good sensitivity is obtained by using cyclohexylamine in ethylene glycol monoethyl ether as solvent [53], [54].

Chlorophenols are difficult to separate by gas chromatography because of their high acidity. However, the light homologues can be separated by using an acidic solid support in a glass column. The heavy homologues can be separated only after the phenol function is blocked by means of methylation, acetylation, silylation, or phthalation. Traces of chlorophenols may be determined as their trifluoroacetylated derivatives by gas chromatography with electron capture detection (limit of detection ca. 20 mg/L) [52], [54].

All phenols are correctly separated by reverse-phase HPLC on a silica column grafted with C_{18} species. Grafts of the propylamino type enable separations based on the degree of chlorination (the retention time depends on the pK_a). UV (280 nm) detection is normally used; amperometric detection is 100 times more sensitive [55–58].

IR spectrometry can distinguish between the different isomers, which absorb in the region 1600 – 400 cm^{-1} [59]. Proton NMR spectrometry can identify most of the different chlorophenols by the chemical shift and the coupling of their aromatic protons. Pentachlorophenol can be identified after the phenol group has been complexed with hexamethylphosphoric triamide [*680-31-9*]. Carbon-13 NMR is less sensitive than proton NMR, but it can provide even more useful information [60].

The individual chlorophenols can be identified and determined with great sensitivity by means of gas chromatography followed by chemical ionization mass spectrometry. The chlorophenols are first converted to their pentafluorobenzyl ether or pentafluorobenzoic ester derivatives. The analysis is then carried out on these species because of their higher electronegativity [61].

Thermal or chemical degradation of chlorophenols can produce substances that are deleterious to the environment. Certain polychlorodibenzoparadioxins, polychlorodibenzofurans, and polychlorophenoxyphenols are examples. These substances can be determined by HPLC with UV detector at a sensitivity of ca. 1 mg/kg [50], [52]. Determinations at the µg/kg level require that the sample be pretreated to remove the major portion of the chemical interferences. The sample is first extracted from an alkaline medium with hexane and then separated on an alumina liquid chromatography column.

The chosen compounds can then be collected into homologous groups [62]. The final determination is made by using GC with electron capture detection or by combined gas chromatography and mass spectrometry. The limit of detection for 2,3,7,8-tetrachlorodibenzo-*p*-dioxin [*1746-01-6*] (TCDD) is less than <1 µg/kg [52].

6. Quality Specifications

The minimum assay for commercial chlorophenols is 98.5 wt %. More often than not, it is >99 wt % and at times even 99.5 wt %. The APHA color number is always <100. The principal impurities are other isomers of the chlorophenol or chlorophenols with more or fewer chlorine atoms. In the majority of cases, the 2,4,5-trichlorophenol content is <20 mg/kg. The absence of chlorobenzoparadioxins, chlorobenzofuran, and most often chlorobenzenes should be noted.

Polychlorophenoxyphenols are the major impurity (up to several percent) in the heavy chlorophenols, which are marketed as mixtures of tetrachlorophenols and pentachlorophenol. Trace quantities of chlorobenzoparadioxins and chlorobenzofuran are also found. The highly toxic compound 2,3,7,8-tetrachlorodibenzo-*p*-dioxin has never been detected in products made by chlorination [52]. The sodium salts of the heavy chlorophenols contain the same impurities as the parent compounds.

7. Economic Aspects

The world market, excluding the Comecon countries and China, is fairly stable and is ca. 100 000 t/a.

The production of heavy chlorophenols is 25 – 30 kt/a. Of this 16 – 20 kt/a is produced in the United States (Vulcan, Reichhold), 8 – 10 kt/a in Western Europe (Rhône-Poulenc, Dynamit Nobel), and 2 kt/a elsewhere in the West.

Production of light chlorophenols is ca. 60 kt/a. Of this, 18 – 20 kt/a is made in the United States (Dow) and 30 kt/a in Western Europe (Rhône-Poulenc, BASF, Coalite, Bayer, Chemie Linz, Esbjerg, KVK).

8. Uses

Chlorophenols are used as agricultural chemicals, pharmaceuticals, biocides, and dyes.

Agricultural Chemicals. The preparation of agricultural chemicals consumes 80 – 90 % of chlorophenol production. The annual consumption of the herbicides 2,4-dichlorophenoxyacetic acid (2,4-D) and 2,4-dichlorophenoxy-4-butyric acid is 30 – 40 kt. Hoelon (Hoechst), Modown (Mobil), Ronstar (Rhône-Poulenc), and TOK E-25 (Rohm and Haas) are among the major *herbicides.* Curacron (Ciba Geigy) exemplifies the *insecticides* made from chlorophenols. Among the *fungicides* are Bayleton (Bayer), MO (Mitsui Toatsu Chemicals), and Procloraz or Sportak (Boots).

Pharmaceuticals. Pharmaceuticals derived from chlorophenols include clofibrate, ethyl 2-(4-chlorophenoxy)-2-methylpropionate (ICI), used in the treatment for high serum cholesterol, and Mervan or Aldofene, an antiinflammatory and analgesic drug (Continental Pharma).

Biocides. Biocides that are based on chlorophenols include the molluscicide Bayluscide (Bayer) and the bactericide Santophen or Chlorophen (Monsanto). Pentachlorophenol is also used to make wood and cellulose preservatives, esters for the textile and leather industries, paints, varnishes, and organic adhesives.

The sodium salt of pentachlorophenol is an active, water-soluble material. It is used as a preservative for such organic materials as freshly sawn wood (against blue rot), paints, adhesives, textiles, leather, and furs.

Dyes. Chlorophenols are used in the synthesis of anthraquinone dyes (→ Anthraquinone Dyes and Intermediates). Quinizarin, 1,4-dihydroxy-9,10-anthracenedione [*81-64-1*], is the most important example.

9. Storage and Transportation

The production, storage, transportation, and use of chlorophenols are subject to regulations and stipulations that depend on the products and on the countries concerned. Within the EEC, directive 67/548 governs classification, packing, and labeling; directive "131" 76/464 covers discharges from production plants into aquatic media; directives 77/728 and 78/631 regulate the relevant paints and pesticides, respectively.

Transportation of chlorophenols is governed by international regulations such as IMDG Code (sea), IATA-DGR (air), RID (rail), and ADR (road).

10. Environmental Considerations

Chlorophenols are hazardous to the environment and have been the subject of numerous ecological surveys of fauna and flora in aquatic, terrestrial, and aerial environments [63].

10.1. Toxicity to Fauna and Flora

Microorganisms. All chlorophenols possess bactericidal activities that increase with the degree of chlorination and attain a maximum for the trichlorophenols. Their metabolic effects include enzyme inhibitions of oxidative phosphorylation and the synthesis of adenosine triphosphate. Chlorophenols are highly toxic to algae. Trichlorophenols are at least 100 times more toxic than dichlorophenols, and pentachlorophenol is used as an algicide.

Plants. Chlorophenols possess a phytotoxicity that increases with the degree of chlorination. The absorption of chlorophenols by plants depends on the solubility of the product and pH of the environment. Terrestrial and aquatic plants can in certain cases absorb, transform, and eliminate chlorophenols without harm, but most often these plants are very sensitive to the phytotoxicity of chlorophenols.

Aquatic Organisms. Fish and other aquatic organisms absorb chlorophenols through their gills, gastrointestinal tract, or skin. Measurements over 2 – 4 d show that acute toxicity for invertebrates, crustacea, and freshwater and seawater fish increases with the level of chlorination. For mono-, di-, and trichlorophenols, the 2 – 4 d LC_{50} is 0.6 – 10.3 mg/L, depending on

the species. For pentachlorophenol, it is 0.03 – 0.6 mg/L, except for certain resistant invertebrates that can survive concentrations of 5 – 10 mg/L.

Chlorophenols also have a long-term toxic effect at low concentrations. 2,4-Dichlorophenol and 2,4,6-trichlorophenol at concentrations of 6.2 – 10.3 mg/L affect the fertilization of sea urchin eggs, and pentachlorophenol causes renal and hepatic lesions in fish. The EPA recommends that the maximum average pentachlorophenol concentration in surface waters not exceed 0.055 mg/L. For 2,4-dichlorophenol and 2-chlorophenol, the EPA set the maximum concentrations at 2.020 and 4.380 mg/L, respectively [64]. This limitation, based on toxicity, is strengthened by the fact that chlorophenols bioaccumulate to only a small extent [65], [66].

10.2. Persistence in the Environment

Aquatic Environment. Chlorophenols may be present in the aquatic environment in many forms. They may be dissolved in free or complexed form, adsorbed on suspended inert solid or biological materials or benthic sediments, or carried in biological tissues. They are removed by biodegradation, photodecomposition, volatilization to the atmosphere, and adsorption.

*Biodegradation*is the principal means by which chlorophenols are removed. It must be induced because the antimicrobial activities of these products require that the bacteria adapt. Biodegradation is rapid when adapted microorganisms are already present, but pentachlorophenol is more difficult to biodegrade than the other chlorophenols.

Chlorophenols *photodecompose* when exposed to UV light, such as sunlight on the water surface.

*Volatilization*transfers the chlorophenol from the water to the air but does not otherwise affect it. The *adsorption* of chlorophenols on suspended or benthic materials also plays a role in the amount of chlorophenols in water. Light chlorophenols are fixed to only a small extent, but pentachlorophenol adsorbs very strongly on certain types of particles when conditions are suitable.

Terrestrial Environment. The persistence of chlorophenols in the soil depends on their adsorption – desorption characteristics. Only the adsorption of pentachlorophenol has been studied in depth. It attaches very strongly to earth and is not easily washed away by rain [67].

Atmosphere. The presence of chlorophenols in the atmosphere has been inferred from the detection of minute quantities of pentachlorophenol in rainwater and snow.

11. Toxicology

The toxicity of chlorophenols to animals depends upon the degree of chlorination, the positions of the chlorine atoms, and the purity of the sample. The LD_{50} values for rats are given in Table 3 [30]. In rats and rabbits, 50 % of the administered dose of pentachlorophenol is eliminated within 24 h. In mice, 70 – 85 % of the dose is eliminated within 4 d. The elimination of tetrachlorophenols is comparable to that of pentachlorophenol [63].

Table 3. Lethal doses (LD_{50}) of chlorophenols for rats

Chlorophenol	LD_{50}, mg/kg	
	oral	percutaneous
2-Chlorophenol	670	950
3-Chlorophenol	570	1030
4-Chlorophenol	261	1390
2,4-Dichlorophenol	580	1730
2,4,5-Trichlorophenol	820	2260
2,4,5-Trichlorophenol, sodium salt	1620	
2,4,6-Trichlorophenol	820	
2,3,4,6-Tetrachlorophenol	140	210
Pentachlorophenol	50	100
Pentachlorophenol, sodium salt	210	72

Toxic doses of chlorophenols cause convulsions, as well as other symptoms [63]. Trembling and clonic convulsions can be induced by tactile or auditory stimulation. Agitation and polypnea are observed before muscular weakness; dyspnea and coma occur and finally death [37], [66]. The toxic effects of pentachlorophenol and the tetrachlorophenols are due to the inhibition of oxidative phosphorylation.

After repeated administration, toxic doses of chlorophenols cause mainly hepatic and renal lesions. The no-effect doses were ca. 100 mg/kg for 2,4-dichlorophenol and 2,4,6-trichlorophenol and 1 – 5 mg/kg for pentachlorophenol [63].

Investigations of the elimination and metabolism of chlorophenols have been inconclusive. Dogs excrete 2-chlorophenol in the urine as both the free and conjugate forms (87 %). However, with rabbits 2-chlorophenol is detected after the administration of chlorobenzene. 2,4-Dichlorophenol, 2,4,6-trichlorophenol, and traces of 2,4,5-trichlorophenol are found in their free and conjugate forms, following the administration of β- or γ-hexachlorocyclohexane to mice.

2,4,5-Trichlorophenol and its conjugate forms are found in the urine following the administration of 2,4,5-trichlorophenoxy-4′-butyric acid to mice; 2,3,4,5-, 2,3,4,6-, and 2,3,5,6-tetrachlorophenols, free and in their conjugate forms, are detected after the administration of pentachlorophenol [61].

Chronic effects of human exposure to low doses of chlorophenol have not been reported, but toxic doses of pentachlorophenol cause severe pathological modifications in the gastrointestinal tract, lungs, spleen, liver, kidneys, and cardiovascular system. Low levels of pentachlorophenol have been found in the urine, blood, and body fat of exposed and unexposed persons, but no explanation for their presence is available. Certain chlorophenols cause dermatoses such as photoallergic contact dermatitis, papulofollicular lesions, comedones, and sebaceous cysts [40], [41], [63], [68].

12. References

1. A. Laurent, *Ann. Chim. Phys.* **3** (1841) 195.
2. Grignard: *Traité de chimie organigue,* 1st ed., 1940, T6-915.
3. Velsicol Chemical, US 4 226 813, 1979 (J. A. Albright).
4. Dow Chemical, US 4 334 111, 1978 (R. A. Davis).
5. Synthelabo, FR 2 475 543, 1980 (J. P. Kaplan).
6. ANVAR, FR 2 487 343, 1980 (J. Martinez).
7. Hitachi Manell, JP 81/10 521, 1979 (N. Ogoshi).
8. Bayer, DE 3 019 526, 1980 (G. Rauchschwalbe).
9. Spofa United Pharma, US 4 221 919, 1975; CS 8681, 1974.
10. Rhône-Poulenc, EP 46 719, 1980 (S. Ratton).
11. UBE, JP 80/62 033, 1978 (Y. Umemura).
12. VEB Berlin Chemie, DE 2 903 343, 1979 (E. Gores).
13. PPG, US 4 293 707, 1979 (S. B. Richter).
14. Coalite, EP 49 126, 1980; GB 31 041, 1980 (K. Gladwin).
15. Akad. Wiss. DDR, DD 142 333, 1979 (D. Martin).
16. Ciba Geigy, US 4 223 166, 1977 (R. Jaeger).
17. Nippon Synthetic Chem., JP 81/55 337, 1979 (K. Nakashima).
18. Nippon Synthetic Chem., JP 81/55 338, 1979 (K. Gougi).
19. Nippon Synthetic Chem., JP 81/55 339, 1979 (K. Nakashima).
20. ICI, US 4 242 121, 1976 (A. F. Hawkins).
21. UBE, EP 23 725, 1980; JP 79/99 914, 1979 (H. Sekioka).
22. Dow Chemical, US 4 332 820, 1980 (D. M. Lowell).
23. Bayer, DE 2 658 866, 1976 (H. Seifert).
24. W. D. Watson, *J. Org. Chem.* **39** (1974) no. 8, 1161.
25. Ciba Geigy, DE 3 318 791, 1983; CH 3234, 1982 (I. Szekely).
26. Y. Ogata et al., *J. Chem. Soc. Perkin Trans. 2* 1984, no. 3, 451.
27. Y. Furuya, *Chem. Ind. London* 1976, no. 15, 649.
28. A. Campbell, D. J. Shields, *Tetrahedron* **21** (1965) 211.
29. S. U. Zubarev et al., *J. Org. Chem. USSR (Engl. Transl.)* **4** (1968) no. 10, 1769.
30. *Registry of Toxic Effects of Chemical Substances 1981 – 1982,* Cincinnati, Ohio, 45 226, June 1983.
31. W. D. Watson, *J. Org. Chem.* **50** (1985) 245.
32. W. D. Watson, *Tetrahedron Lett.* 1976, no. 30, 2591.
33. H. P. Crocker, R. Walser, *J. Chem. Soc.* **14** (1982) 1970.
34. Bayer, DE 2 259 433, 1974 (K. Wedemeyer).
35. A. H. Haymob, *Zh. Obshch. Khim.* **26** (1956) 1647 – 1650.
36. A. Galat, *J. Am. Chem. Soc.* **74** (1952) 3890 – 3891.
37. *Rep. Gov. Chem. Ind. Res. Inst. Tokyo* **47** (1952) 327 – 336.
38. Velsicol Chemical, US 4 094 913, 1977 (A. W. Carlson).
39. Dow Chemical, US 2 835 707, 1956 (W. C. Stoesser).
40. *Notices d'information médicale sur les affections figurant dans la liste européenne des maladies professionnelles* 23/7/1962 and 20/7/1966, Commission des Communautés Européennes, ed. 1969.

41. J. C. Limasset et al., *Chlorophénols et dioxines: risques pour les travailleurs de l'industrie chimique et pour les utilisateurs,* INRS Cahier de notes documentaires 1980, note ND 1249-99-80.
42. V. D. Simonov et al., *Zh. Prikl. Khim. (Leningrad)* **45** (1972) no. 12, 2765.
43. Rhóne-Progil, FR 2 212 827, 1972 (G. Soula).
44. Reichhold Chemicals, GB 845 287, 1956 (F. J. Shelton).
45. Progil, FR 2 161 861, 1971 (G. Rivier).
46. Bayer, DE 2 443 152, 1974 (K. Wedemeyer).
47. Bayer, DE 2 344 925, 1972 (W. Kiel).
48. Rhône-Poulenc, EP 55 196, 1980; EP 55 197, 1980; EP 55 198, 1980 (G. Cordier).
49. Rhône-Poulenc, EP 118 377, 1983 (G. Cordier).
50. L. Castle et al., *J. Chem. Soc. Chem. Commun.* 1978, no. 16, 704.
51. M. I. Faraberov et al., *Zh. Prikl. Khim.* **39** (1966) no. 9, 2094 – 3101.
52. Rhône-Poulenc, internal report.
53. W. Selig, *Mikrochim. Acta* **2** (1971) no. 3 – 4, 359.
54. K. Abrahamsson, T. Min Xie, *J. Chromatogr.* **279** (1983) 199.
55. K. Ugland, *J. Chromatogr.* **213** (1981) 83 – 90.
56. J. Nair et al., *J. Liq. Chromatogr.* **6** (1983) no. 14, 2829 – 2837.
57. N. G. Buckman et al., *J. Chromatogr.* **284** (1984) 441 – 446.
58. R. Beauchamp, P. Boinay, *J. Chromatogr.* **204** (1981) 123 – 130.
59. L. J. Bellamy: *The Infra Red Spectra of Complex Molecules,* Chapman & Hall, London 1975.
60. M. T. Tribble, J. G. Traynham, *J. Am. Chem. Soc.* **91** (1969) no. 21, 379 – 388.
61. S. S. Zhi, A. H. Duffield, *J. Chromatogr.* **284** (1984) 157 – 165.
62. H. R. Buser, *J. Chromatogr.* **107** (1975) 295 – 310.
63. V. P. Kozak et al., *Reviews of the Environmental Effects of Pollutants* – XI Chlorophenols – Report EPA – 600/1-179-012.
64. K. L. E. Kaiser, *Can. J. Chem.* **60** (1982) no. 16, 2104.
65. *Fed. Regist.* **45** (1980) no. 231, 79 318.
66. K. Kobayashi, *Nippon Suisan Gakkaishi* **45** (1979) no. 2, 173.
67. M. Alexander, M I. H. Aleem, *J. Agric. Food Chem.* **9** (1961) 44 – 47.
68. INRS, Cahier de notes documentaires 1984, fiche toxicologique no. 196.

Chlorophenoxyalkanoic Acids

Weed Control and Plant Growth Regulators are separate keywords.

MARGUERITE L. LENG, Dow Chemical, Midland, Michigan 48640 United States

1. **Introduction** 1
2. **Production** 2
3. **Quality Specifications** 4
4. **Uses** 4
5. **Toxicology and Occupational Health** . . 5
6. **Biochemical and Environmental Aspects** 7
7. **References** 8

1. Introduction

Research on chlorinated phenoxy compounds during World War II led to the discovery of the selective *herbicidal action* of 2,4-dichlorophenoxyacetic acid (2,4-D) for control of broadleaf weeds in grasses and related crops, such as cereal grains, rice, and sugarcane. Use patents issued before 1950 included several in the United States assigned to American Chemical Paint Co. (Amchem, now Union Carbide) and Dow Chemical [1].

Although all patents have expired and many new plant growth regulator type weedkillers have been developed, phenoxy compounds still comprise a large sector of the worldwide herbicide market. The estimated total *capacity* for manufacture of chlorophenoxyalkanoic acids and derivatives approached 2×10^5 t in 1982, with a value of about \$ 1.75/kg (\$ 0.80/lb.). Total production decreased about 10 % worldwide in 1983 because of weak prices and lower demand.

The *principal chlorophenoxyalkanoic acids used as herbicides* are listed in Table 1. They are formulated as water-soluble salts and oil-soluble or emulsifiable esters. Major products, in terms of volume, are 2,4-D and MCPA followed by mecoprop and dichlorprop. Most production of 2,4,5-T and fenoprop has been discontinued because of the high toxicity of the dioxin contaminant formed during manufacture of the 2,4,5-trichlorophenol intermediate used for making these herbicides.

Detailed information on individual chlorophenoxyalkanoic acids and their derivatives can be found in herbicide manuals and handbooks [1–4] and in extensive reviews published since 1970 [5–18].

2. Production

Pokorny Process. Most technical phenoxy acids are produced by using the simple reaction first described in 1941 by POKORNY [19] for synthesis of 2,4-D and 2,4,5-T. The appropriate chlorophenol is reacted with a chloroalkanoic acid under alkaline conditions at about 100 °C in an aqueous medium or in an organic solvent, such as toluene, xylene, or chlorobenzene:

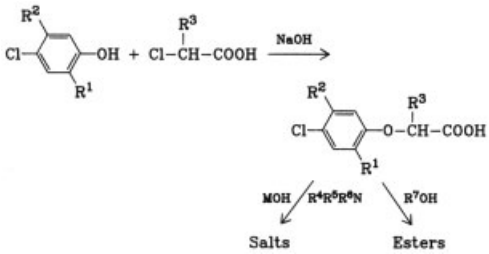

where R^1 is CH_3 or Cl, R^2 is H or Cl, R^3 is H or CH_3, $R^4R^5R^6N$ is an alkylamine, R^7OH is an alcohol, and MOH is an alkali metal base. The resultant salt is hydrolyzed to the free acid, which is purified by recrystallization from an organic solvent or by steam distillation. Chlorophenoxy acids prepared from 2-chloropropionic acid contain an asymmetric carbon atom and are isolated as racemic mixtures. A schematic diagram of the Pokorny process used to produce 2,4-D, its salts, and its esters is shown in Figure 1.

Various methods for synthesis of chlorophenoxyalkanoic acids, esters, and salts have been reviewed [10, pp. 107 – 123]. The nature and quantity of byproducts depend largely on the purity of the starting chlorophenols (→ Chlorophenols). For example, the 2,4-dichlorophenol used for making 2,4-D, dichlorprop, and 2,4-DB is pre-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_009

Table 1. Chlorophenoxyalkanoic acids used as herbicides ∗

Common name	Chemical name (IUPAC)	CAS registry number	Formula	M_r
2,4-D	(2,4-dichlorophenoxy)acetic acid	[*94-75-7*]	$C_8H_6Cl_2O_3$	221.04
2,4-DB	4-(2,4-dichlorophenoxy)butyric acid	[*94-82-6*]	$C_{10}H_{10}Cl_2O_3$	249.09
Dichlorprop	2-(2,4-dichlorophenoxy)propionic acid	[*120-36-5*]	$C_9H_8Cl_2O_3$	235.05
MCPA ∗∗	(4-chloro-o-tolyloxy)acetic acid	[*94-74-6*]	$C_9H_9ClO_3$	200.63
MCPB ∗∗	4-(4-chloro-o-tolyloxy)butyric acid	[*94-81-5*]	$C_{11}H_{13}ClO_3$	228.68
Mecoprop ∗∗	2-(4-chloro-o-tolyloxy)propionic acid	[*93-65-2*]	$C_{10}H_{11}ClO_3$	214.66
2,4,5-T	(2,4,5-trichlorophenoxy)acetic acid	[*93-76-5*]	$C_8H_5Cl_3O_3$	225.49
Fenoprop	2-(2,4,5-trichlorophenoxy)propionic acid	[*93-72-1*]	$C_9H_7Cl_3O_3$	269.53

∗ Melting point data in Table 2.
∗∗ The trivial names MCPA, MCPB, and mecoprop are derived from former chemical nomenclature using 2-methyl-4-chlorophenoxy as the basic unit.

pared by chlorination of phenol and contains 2,6-dichlorophenol and 2,4,6-trichlorophenol as major impurities. These byproducts are not formed if the chlorination is carried out in liquid sulfur dioxide below its boiling point [10].

Similarly, the chlorophenol used to manufacture MCPA, MCPB, and mecoprop is obtained by chlorination of *o*-cresol, a starting material in abundant supply in Europe particularly in Britain, as a product of coal tar distillation. Although chlorination occurs mostly at the 4-position, large amounts of 6-chloro- and 4,6-dichloro-2-methylphenol are also produced when chlorine or alkali hypochlorites are used. A much purer product is obtained when sulfuryl chloride is used for chlorination [10].

A completely different route is used for production of 2,4,5-trichlorophenol by *hydrolysis of 1,2,4,5-tetrachlorobenzene*:

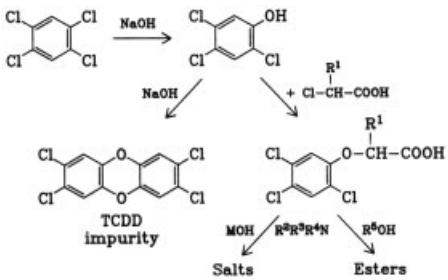

where R^1 is H or CH_3. $R^2R^3R^4N$ is an alkylamine, R^5OH is an alcohol, and MOH is an alkali metal base. One chlorine substituent is replaced by using alcoholic alkali under pressure, with careful control of temperature to avoid formation of highly toxic 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) by exothermic condensation of two molecules of 2,4,5-trichlorophenol. Trace amounts of TCDD are carried over into 2,4,5-T and, to a lesser extent, into fenoprop (also known as silvex or 2,4,5-TP). The impurities in technical-grade 2,4,5-T may vary among producers and among batches of the same product [9, p. 38].

Chlorination of phenoxyacetic acid is another important synthetic route for production of 2,4-D. However, the product may contain a high level of 2,4-dichlorophenol, which has a very strong odor [5]. Chlorination of tolyloxyacetic acid produces mainly the 4-chloro derivative, with small amounts of 6-chloro and other impurities [10].

Phenoxybutyric acid derivatives, such as 2,4-DB and MCPB, are generally prepared by using *γ-butyrolactone* to acylate the chlorophenols [5].

An alternate route to phenoxy acids is by formation of the appropriate *chlorophenolate* and reaction with an ester of the appropriate chloroalkanoic acid [10]. The resulting phenoxy ester can be saponified and acidified to release the free acid, or the phenoxy ester can be used directly for preparation of oil-soluble herbicide formulations. However, these reaction conditions could lead to condensation of chlorophenols to dioxins, and may have been responsible for the di-, tri-, and tetrachlorodibenzo-*p*-dioxins found in certain 2,4-D products in Canada [20]. The 2,3,7,8-tetrachloro isomer has not been detected in products other than 2,4,5-T and fenoprop.

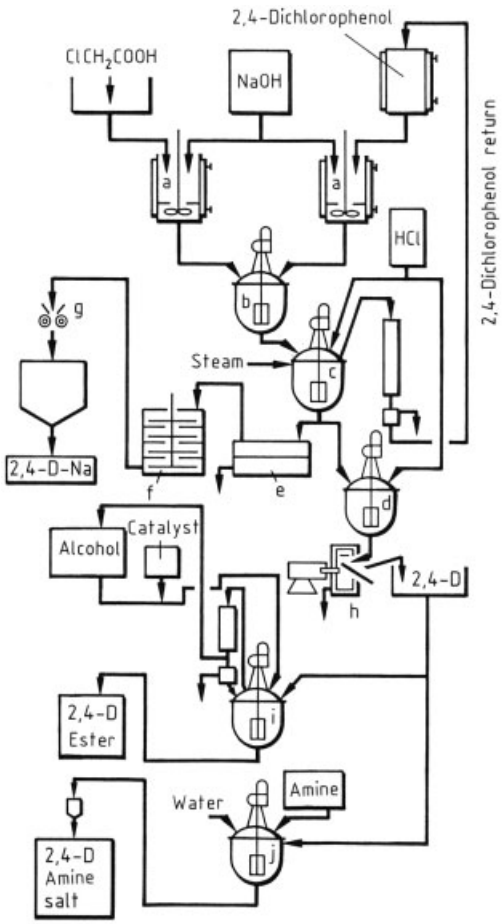

Figure 1. Production of 2,4-dichlorophenoxyacetic acid (2,4-D), as well as its salts and esters
a) Neutralization vessel; b) Reaction kettle; c) Acidification kettle; d) Precipitation vessel; e) Suction filter; f) Drier; g) Mill; h) Filter; i) Esterification kettle; j) Amination kettle

Platinum-catalyzed oxidation of 2-phenoxyethanols has also been used to synthesize phenoxyacetic acids [21]. However, this method may not be commercially feasible because of cost.

Volatile products containing isopropyl or butyl esters of 2,4-D and 2,4,5-T were produced in large quantities prior to 1970, particularly for use as defoliants in the Vietnam conflict. However, most countries now permit use of only *low-volatile products* containing salts or derivatives with a long side-chain, such as isooctyl, butoxyethyl, or PGBE (polypropylene glycol butyl ether) esters of chlorophenoxyalkanoic acids [7], [9], [10].

*Water-soluble salts*of phenoxy acids are formed with alkali metal cations, ammonia, and organic amines, such as dimethylamine, triethylamine, or triisopropanolamine. Magnesium and calcium salts of 2,4-D and MCPA are considerably less soluble in water than are sodium and potassium salts [10]. Oil-soluble salts are prepared from long-chain amines derived from fatty acids [1], [10].

Process *waste streams* are treated by a combination of methods to reduce levels of chlorophenolic and chlorophenoxy compounds to low ppm values prior to discharge to public waters. Methods include steam stripping, solvent extraction, carbon adsorption, resin adsorption, chlorine oxidation, biological oxidation, and incineration.

Principal *manufacturers* of technical-grade acids and derivatives are Akzo Zout in the Netherlands, BASF and Bayer in the Federal Republic of Germany, Chemie-Linz in Austria, Dow in the United States and Brazil, Koge in Denmark, A. H. Marks and May & Baker in the United Kingdom, Rhône-Poulenc in France, and Vertac in the United States. About 25 % of the world capacity in 1982 was located in East Europe, mainly in Czechoslovakia, German Democratic Republic, Hungary, Poland, Romania, and USSR. Phenoxy herbicides are also produced in Argentina, Australia, India, Japan, Mexico, New Zealand, the People's Republic of China, the Philippines, and Taiwan. Several major manufacturers have discontinued production or sold their facilities since 1982.

End-use products are formulated by most manufacturers of technical chlorophenoxyalkanoic acids and derivatives, as well as by many additional distributors worldwide. Numerous trade names have been listed for phenoxy herbicide products [1–4], [7], [10], [18], but many such names were discontinued or were transferred to other companies.

3. Quality Specifications

International specifications for phenoxy herbicides and other pesticides have been developed by the United Nations Food and Agriculture Organization (FAO). Methods of analysis have been tested and approved by the Collaborative International Pesticides Analytical Council

(CIPAC) and the Association of Official Analytical Chemists (AOAC).

Updated specifications for phenoxyalkanoic herbicides were consolidated in a booklet issued by FAO in 1984 [22]. The values for technical acids summarized in Table 2 have "full" status for MCPA and 2,4,5-T and "provisional" status for the others. Other parameters for identification include IR spectra, GLC retention times, and R_f values for TLC analyses. The FAO also lists specifications for salts and esters, aqueous salt solutions, emulsifiable ester concentrates, and mixtures containing phenoxy salts or esters, all given in terms of the acids released by hydrolysis of these derivatives. Complete names are required of the base components of salts and the alcohol components of esters in phenoxy herbicide products.

The U.S. Environmental Protection Agency (EPA) has issued regulations and guidelines for registration of pesticide products [23]. These include chemistry requirements for the identification of all impurities and byproducts down to 0.1 % in technical-grade products.

For many years, the *analytical methods* used for assessing the quality of phenoxy herbicides involved simple acid–base titration, or determination of total chlorine released by pyrolysis of technical-grade acids and derivatives [2–4], [18]. These nonspecific, and therefore inaccurate methods have been superseded by GLC procedures [1], [24], [25], and, by isomer-specific high-pressure liquid chromatography [26–28]. Similar methods for other phenoxy herbicides are undergoing collaborative testing and will likely be required in the future [10].

4. Uses

Distribution, marketing, and use of pesticides are closely regulated in most countries, either through national statutes or by indirect reliance on reviews by expert committees under the auspices of international agencies, such as FAO, the World Health Organization (WHO), and the Codex Food Standards Programme of the United Nations.

Labels on containers of phenoxy herbicides are registered individually in most exporting and importing countries (and even in states within countries). These labels specify the nature and content of active ingredients; give precautions to be taken to avoid hazard to humans, livestock, wildlife, or the environment; and describe precise conditions for each approved use.

Phenoxy herbicides are used primarily for *selective control of broadleaf weeds* in cereal grains, pastures, and turf; and for removing unwanted brushy species in rangeland, forests, and noncropland. Rates of application range from as low as 0.25 kg/ha in grain crops to as high as 16 kg/ha for spot treatment of individual trees in rights-of-way. Very dilute solutions of 2,4-D and fenoprop derivatives have also been used as plant growth regulators in fruit orchards.

The choice of products depends on the species of plants to be controlled [1–4], [6]. For example, 2,4,5-T and fenoprop are effective on woody plants and herbaceous weeds that are resistant to 2,4-D. Dichlorprop, fenoprop, and mecoprop are effective for controlling chickweed and other unwanted plants in lawns and turf. The butyric acid derivatives 2,4-DB and MCPB can be used on sensitive crops, such as peas, peanuts, soybeans, and seedling forage legumes (alfalfa and clover), because they are inactive as herbicides until converted to the acetic acid derivatives by β-oxidation of the side-chain [6].

Phenoxy herbicides are applied alone or as mixtures with other herbicides in solutions, dispersions, or emulsions in water or oil; equipment that produces large droplets is used to avoid spray drift. Although volatile derivatives, such as isopropyl and butyl esters of 2,4-D and 2,4,5-T, are effective for controlling weeds and brush in noncropland, treatments with low-volatile long-chain esters and amine salts are less likely to cause damage to susceptible off-target crops or desirable vegetation. Granular salt formulations of 2,4-D or fenoprop have been used for aquatic weed control, but longer-acting ester formulations are more effective.

Formulated phenoxy herbicide products generally contain inert ingredients to prevent precipitation of salts when these formulations are diluted with hard water, or to help disperse or emulsify esters in oil or oil/water mixtures. Oil concentrates are less corrosive than emulsifiable concentrates or salt solutions, but are more flammable and require more restrictive labeling for safe transport, storage, and handling.

Table 2. International FAO specifications for technical chlorophenoxyalkanoic acids [22]

Common name	*mp*, °C	Acid (anhydrous) [a], g/kg min.	Free phenol [b], g/kg max.
2,4-D	137 – 141	890 ± 25	5
2,4-DB	114 – 119	890 ± 30	20
Dichlorprop	113 – 118	890 ± 25	15
MCPA	118 – 119 [c]	840 ± 20	10
MCPB	100 [c]	840 ± 30	30
Mecoprop	94 – 95 [c]	840 ± 40	15
2,4,5-T [d]	150 – 156	910 ± 30	5
Fenoprop [d]	175 – 181	890 ± 40	15

[a] The water content of dry acids should not exceed 15 g/kg; the approximate water content should be stated for wet acids containing more than 15 g/kg.
[b] The free phenol content is expressed as the parent phenol in terms of the anhydrous phenoxy acid content.
[c] As listed in [1], [2]; may be 99 – 107 °C for technical-grade MCPA.
[d] Maximum content of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) is 1×10^{-8} expressed as 0.01 mg/kg of the 2,4,5-T or fenoprop content on an anhydrous basis.

Only inert ingredients approved by the EPA are permitted in products to be used on food crops in the United States [23].

5. Toxicology and Occupational Health

Numerous toxicological studies have been conducted with phenoxy herbicides in laboratory animals, livestock, birds, and fish. In addition, pharmacokinetic studies in animals and in humans have demonstrated that these compounds do not accumulate in the body. Detailed summaries of findings are provided in several extensive reviews [7], [9], [11–18], [29–31].

Chlorophenoxyalkanoic acids, derivatives, and formulated products are *classified as moderately toxic,* based on acute studies using single large doses administered by oral, dermal, or inhalation routes in various species of animals [1–4]. According to World Health Organization (WHO) guidelines, 2,4-D and 2,4,5-T are in class II (moderately hazardous) with acute oral LD_{50} values of 375 and 500 mg/kg of body weight in rats [32]. Other phenoxy herbicides are in class III (slightly hazardous), with acute oral LD_{50} values in mg/kg for rats as follows: fenoprop (650), MCPB (680), 2,4-DB (700), MCPA (700), dichlorprop (800), and mecoprop (930). Acute oral toxicity values for salts and esters are comparable to those for the acids, on an acid equivalent basis [7], [9].

Dogs appear to be more sensitive to chlorophenoxyalkanoic acids than rats, whereas birds are less sensitive [7] and livestock can tolerate higher doses without apparent effect [9]. Phenoxy herbicides are not toxic to bees. The LC_{50} in fish may range from less than 1 mg/L for esters to greater than 100 mg/L for acids and salts [7].

Dry acids are only slightly irritating to eyes and skin, but concentrated solutions or wetted powders may cause burns if contact is prolonged or repeated [31]. In rabbits, no significant adverse systemic effects resulted from exposure for 3 weeks to dilute solutions of formulated products containing dimethylamine salt, isooctyl ester, or butyl ester of 2,4-D [31].

Although phenoxy compounds are not acutely toxic in dermal and inhalation studies, potential occupational exposure is greater by these routes than by ingestion. Adequate precautions should be taken to avoid contamination of skin with technical-grade materials or concentrates, and to avoid prolonged breathing of the vapors or dusts. Occupational exposure limits have been set at 10 mg/m^3 for 2,4-D and 2,4,5-T in both the United States (TWA) and the Federal Republic of Germany (MAK).

Effects of 2,4-D in humans were reviewed in 1984 by an international committee convened by WHO [17]. Signs and symptoms of acute overexposure occurred only after ingestion of large amounts of 2,4-D products in suicide cases, or where poor occupational hygiene led to pronounced dermal exposure. The acute no-observed – adverse-effect level for biological effects was estimated to be as high as 36 mg/kg of body weight in humans for 2,4-D or its equiv-

alent in alkali salts, amine salts, or esters. On the basis of available studies on amounts of 2,4-D or 2,4,5-T absorbed through occupational exposure, the margin of safety is at least 360-fold for spraying crews and much greater than 1000-fold for the general population in areas where 2,4-D is used [17].

Little or no effects were seen in livestock or laboratory animals other than dogs given repeated oral doses up to 50 mg/kg of 2,4-D, 2,4,5-T, MCPA, or fenoprop, as acids, salts, or esters [7], [9]. Reversible embryotoxic or fetotoxic effects were seen in offspring of pregnant rats and mice given daily doses above 40 mg/kg during organogenesis [7], [9], [11–18]. However, true teratogenic effects, such as cleft palate, were seen only in mice given 2,4,5-T, possibly due to high levels of TCDD contaminant in some samples of technical-grade acid and esters [9, p. 115].

Dietary feeding studies have also been conducted with phenoxy acids and derivatives in rats, dogs, and mice for periods ranging from 3 months to 2.5 years to evaluate potential for causing chronic toxicity, cancer, or reproductive effects. Details are provided in extensive reviews by National Research Council of Canada [7], WHO [11–18], and others [30–37]. The major findings are as follows:

In chronic feeding studies with *2,4-D* conducted by U.S. Food and Drug Administration [33], dogs and rats were fed for 2 years on diets containing up to 500 and 1250 ppm technical-grade acid, respectively, and pregnant rats were fed up to 1500 ppm in the diet during three generations of reproduction with two litters per generation. No significant effects were seen in any of the dogs, nor in rats fed up to 625 ppm in the diet equivalent to a dosage level of 31 mg kg^{-1} d^{-1} in this species. There was no effect on fertility nor on average litter size at any dose, but at 1500 ppm the percentage of young born that survived to 21 days was sharply reduced and the weaning weight was depressed. On the basis of these studies, the acceptable daily intake (ADI) for 2,4-D in humans was estimated by the 1971 WHO Expert Committee on Pesticide Residues at 0.3 mg kg^{-1} d^{-1} with a 100-fold safety factor over the no-effect level in rats [12]. Additional studies are under way in the United States to fulfill EPA requirements for registration of pesticides used in food crops [23].

Four long-term feeding studies have been conducted with *2,4,5-T,* each study at dosage levels of 3, 10, and 30 mg kg^{-1} d^{-1} in rats. Effects were seen at the high dose, and to a lesser extent at the middle dose, in studies conducted by Dow in the United States using purified 2,4,5-T ($<5\times10^{-10}$ TCDD) in the diet of rats for 2 years [34] or during three generations of reproduction [35]. On the other hand, all dose groups were comparable to controls in studies conducted in the Federal Republic of Germany using technical-grade 2,4,5-T (5×10^{-8} TCDD) in the diet of rats from before conception through weaning to age 130 weeks or through three generations of reproduction [14]. On the basis of these studies, the 1981 WHO Expert Committee estimated the ADI in humans at 0.03 mg kg^{-1} d^{-1} for 2,4,5-T containing no more than 1×10^{-8} TCDD [16].

Unpublished studies with *fenoprop* have been summarized in two reviews [30], [31]. The dosage levels causing no significant toxicological effect were 7 mg kg^{-1} d^{-1} in 90-day feeding studies with the sodium salt and PGBE esters in rats and the potassium salt in dogs, and essentially 2.6 mg kg^{-1} d^{-1} in 2-year feeding studies with the potassium salt in both rats and dogs.

Subchronic feeding studies with *MCPA* and *mecoprop* have also been reviewed [7], [9]. When diethanolamine salts of these compounds were fed to rats at up to 2500 ppm in the diet for 7 months, the only effect noted was unspecified mortality, mainly due to infection at the maximum dose [36]. On the other hand, no mortalities were reported when MCPA as the sodium salt or mecoprop as the acid was fed at levels up to 3200 ppm in the diet of rats for 3 months. The only effects noted were reduced food consumption and decreased weight gain at the highest dose [37]. Additional studies with MCPA are under way to meet current requirements in the United States for reregistration of pesticides used in food crops [23].

The *potential human carcinogenicity* of chlorophenoxyalkanoic acids and their derivatives has been evaluated by several expert groups. Committees convened by the International Agency for Research on Cancer (IARC) reviewed all of the above studies in rats, numerous studies in mice conducted by U.S. National Cancer Institute, and several epidemiological studies in humans occupationally exposed to

formulated products [18]. Both *2,4-D* and *2,4,5-T* and their esters were assigned to group 3 (evidence inadequate for carcinogenicity to humans and animals, and for activity in short-term tests) [18, 1982, pp. 101 – 103, 235 – 238]. For *MCPA*, no adequate data were available to evaluate carcinogenicity to experimental animals or humans [18, 1983, pp. 255 – 269]. On the other hand, *TCDD* was classified in group 2B (sufficient evidence for carcinogenicity to animals and inadequate evidence in humans and in short-term tests [18, 1982, pp. 238 – 243], whereas *chlorophenols* and *phenoxyacetic acid herbicides* (occupational exposure) were generically classified in group 2B on the basis of only limited evidence for carcinogenicity to humans [18, 1982, pp. 88 – 89, 211 –212].

6. Biochemical and Environmental Aspects

The following observations summarize findings reported in extensive reviews on biochemistry, metabolism, degradation, and residues of phenoxy herbicides [6], [7], [9], [11–18], [31], [38], [39].

The *solubility* of chlorophenoxyalkanoic acids ranges from 45 to 1300 mg/L in water and from 0.1 to 153 g/100 mL in CCl_4 and ethanol, respectively [6]. These compounds are relatively strong acids, with p*K* values ranging from 2.64 for 2,4-D to 6.21 for MCPB [6], and exist mainly as organic anions at the pH of living tissues. Salts dissociate in aqueous media, and esters are rapidly hydrolyzed to acids by enzymatic action in animals, plants, and soil [6].

Phenoxy acids and derivatives are rapidly *absorbed* from the gastrointestinal tract in animals and are excreted in urine, largely as free acids, with a half-life of less than 1 day [38]. Doses greater than 50 mg/kg of body weight may saturate the excretory mechanism, particularly in dogs and pigs [9]. Fenoprop is excreted more slowly than 2,4-D and 2,4,5-T, and appears partly in conjugated form in urine. In humans, small oral doses of 2,4-D or 2,4,5-T (5 mg/kg) or fenoprop (1 mg/kg) were handled much as in rats, indicating little likelihood for accumulation from repeated exposure [31]. Absorption into the body is slower from dermal exposure than from oral ingestion [17], [39].

The phenoxy ether link is stable in plants and animals, but is cleaved by bacterial action in soil and in the rumen of cattle and sheep [38]. The rate of cleavage is inhibited by substitution of a third chlorine at the meta position on the ring and is sterically hindered by the angular methyl group in the 2-propionic acid side-chain. Thus, 2,4,5-T and fenoprop accumulate in the liver and kidney of cattle and sheep ingesting high levels in the diet [38].

The first step in *degradation* of chlorophenoxy acids and chlorophenols occurs by hydroxylation at the 4-position with shift of the 4-chloro group to the 3- or 5-position (NIH shift). This reaction also is inhibited by the presence of a *m*-chloro substituent; thus, residues of 2,4,5-trichlorophenol appeared in the milk of cows and in the liver and kidney of cattle maintained on feed containing 100 or 300 ppm 2,4,5-T, respectively [38].

The rate of *microbial degradation* of chlorophenoxyalkanoic acids in the environment depends on a number of factors, including sunlight, temperature, moisture and organic matter in soil, and whether the organisms were adapted by repeated treatment [6], [7], [9]. Under field conditions favorable for degradation, 2,4-D disappears within about 2 – 3 weeks, whereas other phenoxy herbicides disappear more slowly [6]. The average half-lifes in a laboratory study were 4, 10, 17, and 20 days for 2,4-D, dichlorprop, fenoprop, and 2,4,5-T, respectively, when added as dimethylamine salts at about 5 ppm in three soils containing partially decomposed litter obtained from two forest sites and one grassland site in Oklahoma [40].

Photodegradation is rapid when phenoxy herbicides are exposed to sunlight in water containing dissolved organic matter, and the resultant chlorophenols are photolyzed even more rapidly than the parent acids [7], [9]. The dioxin impurity TCDD is also degraded rapidly by sunlight, particularly in the presence of its herbicide carrier on the surface of leaves or in the presence of other hydrogen donors dissolved in water [7].

Plants treated with phenoxy herbicides rapidly hydrolyze esters and salts to the parent acids, which form ether-soluble *conjugates* with amino acids or water-soluble conjugates with sugars, depending on the relative susceptibility or resistance of the plants to the herbicides [6]. *Analyses for residues* in treated crops

and environmental samples involve extraction with organic solvents under acid conditions [10] or with polar solvents under alkaline conditions [41]. Quantification is best accomplished by conversion to volatile esters and measurement by electron capture gas chromatography (EC/GC). Other methods include colorimetry and thin-layer chromatography [4], [10].

Tolerances have been established in the United States for residues of 2,4-D and MCPA in a number of food and feed crops, and in milk and meat [42]. Permitted levels range from 0.05 ppm (the detection limit) in strawberries, poultry, and eggs to 1000 ppm on grass immediately after treatment with 2,4-D at high rates. Similarly, international maximum residue limits (MRLs) have been recommended for 2,4-D and 2,4,5-T by FAO [11–13], [15], and have been accepted by many member countries attending annual meetings of the United Nations Codex Committee on Pesticide Residues [43]. However, residues of chlorophenoxyalkanoic acids have seldom been detected in monitoring studies conducted on domestic and imported foods in a number of countries [44].

Guidelines for *drinking water* quality issued in 1982 by WHO recommend a maximum concentration level of 100 µg/L for 2,4-D in water [45].

7. References

1. *Herbicide Handbook of the Weed Science Society of America (WSSA),* 5th ed., WSSA, Champaign, Ill., 1983, pp. 128, 141, 167, 292, 295, 297, 429, 443.
2. D. Hartley, H. Kidd (eds.): *Agrochemicals Handbook,* Royal Society of Chemistry, Nottingham, England, 1983, updated 1984.
3. British Crop Protection Council, C. R. Worthing (ed.): *Pesticide Manual, a World Compendium,* 7th ed., Lavenham Press, Lavenham, Suffolk, England, 1983, 655 pp.
4. E. Y. Spencer (ed.): *Guide to the Chemicals Used in Crop Protection,* 7th ed., Publication 1093 of Research Branch of Agriculture Canada, London, Ontario, Canada, 1982, 595 pp.
5. N. N. Melnikov: "Aryloxalkylcarboxylic Acids," Chap. 14 in "Chemistry of Pesticides," *Residue Rev.* **36** (1971) 157 – 176.
6. M. A. Loos: "Phenoxyalkanoic Acids," in P. C. Kearney, D. D. Kaufman (eds.): *Herbicides: Chemistry, Degradation and Mode of Action,* 2nd ed., vol. 1, Marcel Dekker, New York 1975, pp. 1 – 128.
7. National Research Council of Canada: *Phenoxy Herbicides – Their Effects on Environmental Quality,* NRC no. 16075, NRCC/CNRC Publications, Ottawa, Canada, 1978, 440 pp.
8. National Research Council of Canada: *2,4-D: Some Current Issues,* NRC no. 20 647, NRCC/CNRC Publications, Ottawa, Canada, 1983, 99 pp.
9. R. W. Bovey, A. L. Young: *The Science of 2,4,5-T and Associated Phenoxy Herbicides,* Wiley-Interscience, New York 1980, 462 pp.
10. S. S. Que Hee, R. G. Sutherland: *The Phenoxyalkanoic Herbicides, Chemistry, Analysis, and Environmental Pollution,* **vol. 1,** CRC Series in Pesticide Chemistry, CRC Press, Boca Raton, Fla., 1981, 321 pp.
11. FAO/WHO: *1970 Evaluations of Some Pesticide Residues in Food,* WHO/Food Add./71.42, FAO, Rome 1971, pp. 59 – 86, 459 – 477 (2,4-D and 2,4,5-T, respectively).
12. FAO/WHO: *1971 Evaluations of Some Pesticide Residues in Food,* WHO Pesticide Residue Series no. 1, WHO, Geneva 1972, pp. 83 – 97 (2,4-D).
13. FAO/WHO: *1974 Evaluations of Some Pesticide Residues in Food,* WHO Pesticide Residue Series no. 4, WHO, Geneva 1975, pp. 159 – 183 (2,4-D).
14. FAO/WHO: *1975 Evaluations of Some Pesticide Residues in Food,* WHO Pesticide Residue Series no. 5, WHO, Geneva 1976, pp. 173 – 188 (2,4-D).
15. FAO/WHO: *Pesticide Residues in Food, 1979 Evaluations,* FAO Plant Production and Protection Paper 20 Suppl., FAO, Rome 1980, pp. 469 – 498 (2,4,5-T).
16. FAO/WHO: *Pesticide Residues in Food, 1981 Evaluations,* FAO Plant Production and Protection Paper 42, FAO, Rome 1982, pp. 494 – 501 (2,4,5-T).
17. International Programme on Chemical Safety (IPCS): *Environmental Health Criteria 29, 2,4-Dichlorophenoxyacetic Acid (2,4-D),* WHO, Geneva 1984, 151 pp.
18. IARC: *Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man,* IARC, Lyon, France, vol. 15, 1977, pp. 111 – 138, 273 – 299; vol. 1 – 29 suppl. 4,

1982, pp. 88 – 89, 101 – 103, 211 – 212, 235 – 243; vol. 30, 1983, pp. 255 – 269.
19. R. Pokorny: "New Compounds, Some Chlorophenoxyacetic Acids," *J. Am. Chem. Soc.* **63** (1941) 1768.
20. W. P. Cochrane, J. Singh, W. Miles, B. Wakeford, *J. Chromatogr.* **217** (1981) 289 – 299.
21. H. Fiege, K. Wedemeyer, *Angew. Chem. Int. Ed. Engl.* **20** (1981) no. 9, 783.
22. FAO Specifications for Plant Protection Products, "Phenoxyalkanoic Herbicides," AGP : CP/100, FAO, Rome 1984.
23. EPA, "Data Requirements for Pesticide Registrations, Final Rules," CFR 40, 1984, 158. *Fed. Regist.* **49** (1984) 42 856 – 42 905.
24. Collaborative International Pesticides Analytical Council (CIPAC): *CIPAC Handbook 1, 1 A, 1 B,* Heffers Printers, Cambridge, England, 1970, 1980, 1983, respectively.
25. Collaborative International Pesticides Analytical Council (CIPAC): *CIPAC Proceedings 1980, 1981,* Heffers Printers, Cambridge, England, 1980, 1981.
26. N. E. Skelly, T. S. Stevens, D. A. Mapes, *J. Assoc. Off. Anal. Chem.* **60** (1977) 868 – 872.
27. T. S. A. Stevens, N. E. Skelly, R. B. Grorud, *J. Assoc. Off. Anal. Chem.* **61** (1978) 1163 – 1165.
28. W. P. Cochrane, M. Lanouette, J. Singh, *J. Assoc. Offic. Anal. Chem.* **66** (1983) 804 – 809.
29. J. M. Way, *Residue Rev.* **26** (1969) 37 – 62.
30. W. R. Mullison: "Some Toxicological Aspects of Silvex," *Proc. South. Weed Control Conf.* **19** (1966) 426 – 435.
31. P. J. Gehring, J. E. Betso: "Phenoxy Acids: Effects and Fate in Mammals," in C. Ramel (ed.): "Chlorinated Phenoxy Acids and Their Dioxins," *Ecol. Bull. (Stockholm)* **27** (1978) 122 – 133.
32. "WHO Recommended Classification of Pesticides by Hazard – Guidelines for Classification 1984 – 1985," VBC/84.2, WHO, Geneva 1984.
33. W. H. Hansen, M. L. Quaife, R. T. Habermann, O. G. Fitzhugh, *Toxicol. Appl. Pharmacol.* **20** (1971) 122 – 129.
34. R. J. Kociba, D. G. Keyes, R. W. Lisowe, R. P. Kalnins et al., *Food Cosmet. Toxicol.* **17** (1979) 205 – 221.
35. F. A. Smith, F. J. Murray, J. A. John, K. D. Nitschke et al., *Food Cosmet. Toxicol.* **19** (1981) 41 – 45.
36. M. R. Gurd, G. L. M. Harmer, B. Lessel, *Food Cosmet. Toxicol.* **3** (1965) 883 – 885.
37. H. G. Verschuuren, R. Kroes, E. M. Den Tonkelaar, *Toxicology* **3** (1975) 349 – 359.
38. M. L. Leng: "Comparative Metabolism of Phenoxy Herbicides in Animals," in G. W. Ivie, H. W. Dorough (eds.): *Fate of Pesticides in Large Animals,* Academic Press, New York 1977, pp. 53 – 76.
39. M. L. Leng, J. C. Ramsey, W. H. Braun, T. L. Lavy: "Review of Studies with 2,4,5-T in Humans Including Applicators Under Field Conditions," in J. R. Plimmer (ed.): *Pesticide Residues and Exposure,* ACS Symposium Series 182, Am. Chem. Soc., Washington, D.C., 1982, pp. 133 – 156.
40. J. T. Altom, J. F. Stritzke, *Weed Sci.* **21** (1973) 556 – 560.
41. D. J. Jensen, R. D. Glass: "Analysis of Residues of Acidic Herbicides," in H. A. Moye (ed.): *Analysis of Pesticide Residues,* J. Wiley & Sons, New York 1981, pp. 223 – 261.
42. EPA, CFR 40, "Tolerances for 2,4-D," Section 180.142; "Tolerances for MCPA," Section 180.339, Washington, D.C., 1986.
43. Codex Alimentarius Commission (CAC), Joint FAO/WHO Food Standards Programme: "Codex Maximum Limits for Pesticide Residues," CAC, 1st ed., vol. 13, FAO, Rome 1983, p. 20-iv (2,4-D).
44. International Group of National Assoc. of Pesticide Manufacturers (GIFAP), "Pesticide Residues in Food," GIFAP, Brussels, Belgium, Mar., 1984, 26 pp.
45. WHO: "Recommendations," *Guidelines for Drinking Water Quality,* vol. 1, EFP/82.39, WHO, Geneva 1982, p. 82.

Chloroprene → Chlorinated Hydrocarbons
Chloroprene Rubber → Rubber, 3. Synthetic
Chlorosulfonic Acid → Chlorosulfuric Acid

Chlorosulfuric Acid

JOACHIM MAAS, Bayer AG, Leverkusen, Federal Republic of Germany

FRITZ BAUNACK, Wuppertal, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
5. Safety Precautions, Transportation . . . 3
6. Uses and Economic Aspects 4
7. Toxicology and Occupational Hygiene . 4
8. References 5

1. Introduction

Chlorosulfuric acid [*7790-94-5*], M_r 116.53, is also referred to in the literature as chlorosulfonic acid, sulfuryl hydroxychloride, sulfuric chlorohydrin, sulfonic acid monochloride, and chlorohydrated sulfuric acid.

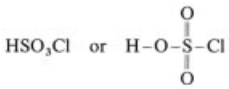

Chlorosulfuric acid is a colorless, mobile, extremely reactive liquid. It fumes very strongly in air and, consequently, has been used as a military smoke-generating agent, being mixed with oleum for this purpose (ca. 50 – 65 wt % HSO_3Cl, 50 – 35 wt % SO_3). A very thick, opaque mist is produced when 2 L of chlorosulfuric acid is mixed with 1000 m^3 of air.

Chlorosulfuric acid is highly corrosive and hygroscopic. It decomposes explosively in the presence of water, forming sulfuric acid and hydrogen chloride. A. WILLIAMSON first prepared chlorosulfuric acid in 1854 [5].

2. Physical Properties

The values of physical constants quoted in the literature should not be accepted uncritically. Chlorosulfuric acid is difficult to purify and has frequently been used in an impure state for experimental determinations. Table 1 shows the best values available. Very pure chlorosulfuric acid can be obtained by fractional crystallization [6].

Table 1. Physical properties of chlorosulfuric acid

Property	Value
Freezing point *fp*	− 80 to − 81 °C
Boiling point *bp* (decomp.)	151 – 152 °C
Vapor pressure	
(p in mbar at T in K)	$\log p = 9.495 - 2752/T$
(p in Pa at T in K)	$\log p = 11.495 - 2752/T$
Density∗	$d_4^{20} = 1.753$
	$d_4^{-70} = 1.90$
	$d_4^{-10} = 1.80$
	$d_4{}^t$ (0 – 100 °C)
	$= 1.7847 - 1.616\,t \times 10^{-3}$
	$+ 1.21\,t^2 \times 10^{-6}$
	$- 4.1\,t^3 \times 10^{-9}$
Specific heat capacity (15 – 80 °C)	1.18 kJ kg^{-1} K^{-1}
Heat of vaporization (at boiling point)	53.6 kJ/mol (460 J/g)
Heat of solution in water (at 18 °C)	168.7 kJ/mol
Heat of formation from elements (at 25 °C)	− 597.5 kJ/mol
Viscosity at − 31.6 °C	10 mPa s
at − 17.8 °C	6.4 mPa s
at 15.6 °C	3.0 mPa s
at 49 °C	1.7 mPa s
Refractive index n_D^{14}	1.437
Dielectric constant (at 15 °C)	60 ± 10
Electrical conductivity (at 25 °C)	$0.2 - 0.3 \times 10^{-3}\ \Omega^{-1}\ cm^{-1}$
Specific magnetic susceptibility	− 0.402

∗ Because chlorosulfuric acid decomposes at the boiling point, its vapor density cannot be determined precisely.

The large heat of vaporization is explained by the strong association of chlorosulfuric acid molecules, which results from hydrogen bonding [7]. Chlorosulfuric acid is soluble in methylene chloride, chloroform, 1,1,2,2-tetrachloroethane, 1,2-dichloroethane, acetic acid, acetic anhydride, trifluoroacetic acid, trifluoroacetic anhydride, sulfuryl chloride, liquid

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_017

sulfur dioxide, and sparingly soluble in carbon tetrachloride and carbon disulfide. It reacts with alcohols, ketones, ethers, and dimethyl sulfoxide.

Whereas sulfur trioxide is miscible in all proportions with chlorosulfuric acid, the limit of solubility of free hydrogen chloride is 0.5 wt % at 20 °C and atmospheric pressure [8].

Warning: Care must be taken when working with chlorosulfuric acid in the presence of solvents; an increase in temperature can lead to violent reactions, particularly when catalysts are also present.

3. Chemical Properties

Chlorosulfuric acid is a strong acid, stable under normal conditions. The tetrahedral molecule [9] contains a relatively weak S – Cl bond [10]. Its force constant is only about one-third that of the single bond in S – O(H) [11]. Depending on the co-reactant, the solvent (if any), the temperature, and the amount of excess HSO_3Cl present, the first step of the reaction will be the formation of free chlorine atoms [12], chlorine ions, or protons [10], [13]. Prolonged heating or distillation, especially in vacuo, leads to partial decomposition to hydrogen chloride, chlorine, sulfur dioxide, sulfuryl chloride, pyrosulfuryl chloride, and sulfuric acid [6].

If chlorosulfuric acid is stored in the presence of an excess of sulfur trioxide, the following reaction occurs [6]:

$$2\,HSO_3Cl + SO_3 \longrightarrow H_2SO_4 + S_2O_5Cl_2$$

In freshly prepared and cooled solutions chloropyrosulfuric acids are formed by the following equilibrium reaction:

$$H(SO_3)_nCl + SO_3 \rightleftharpoons H(SO_3)_{n+1}Cl$$

where $n = 1$, 2, or 3 [14]. Chloropyrosulfates are obtained from the reaction of alkali metal chlorides with sulfur trioxide in liquid sulfur dioxide at − 12 °C [15].

Many reactions with chlorosulfuric acid can be carried out at low temperatures, which is facilitated by its low melting point. Sulfonic acids are produced with equimolar quantities of aromatic hydrocarbons under mild conditions; the reaction usually does not affect other substituents present in the molecule. However, in some cases, amino groups must be protected by acylation from reacting with chlorosulfuric acid. Side reactions lead to the formation of small quantities of sulfonyl chlorides, which usually react further with aromatic hydrocarbons to form sulfones (→ Benzenesulfonic Acids and Their Derivatives). The sulfonyl chlorides are used as intermediates in the manufacture of pharmaceuticals or dyes.

Depending on the reaction conditions, sulfonyl chlorides, sulfones, or chlorinated aromatic hydrocarbons can be obtained by using an excess of chlorosulfuric acid [16], [17]. Sulfonyl chloride formation takes place in two steps: formation of a sulfonic acid followed by an equilibrium reaction [10], [18], [19]:

$$ArH + ClSO_3H \rightleftharpoons ArSO_3H + HCl$$

$$ArSO_3H + ClSO_3H \rightleftharpoons ArSO_2Cl + H_2SO_4$$

This equilibrium can be shifted to the right by reacting the sulfuric acid formed with pyrosulfuryl chloride or with thionyl chloride. The reaction with pyrosulfuryl chloride leads to the formation of chlorosulfuric acid and sulfur trioxide; thionyl chloride reacts with the nascent sulfuric acid to form chlorosulfuric acid, hydrogen chloride, and sulfur dioxide [20].

The reaction of chlorosulfuric acid with aliphatic hydrocarbons is slow, provided there are no double bonds or other reactive groups present. Therefore, long-chain saturated fatty alcohols react with chlorosulfuric acid to form the corresponding sulfates without chain cleavage or discoloration of the product. This type of reaction is used on a large scale in the manufacture of detergents:

$$ROH + HSO_3Cl \longrightarrow HCl + ROSO_3H$$

Both chlorosulfuric acid and fluorosulfuric acid have been proposed as polymerization catalysts [21]. Many of the reactions of chlorosulfuric acid have been reviewed by Jackson [22]. Special information on this subject can be found in refs. [23–31].

4. Production

All the industrial routes to chlorosulfuric acid are based on the reaction of hydrogen chloride with sulfur trioxide:

$$HCl + SO_3 \longrightarrow HSO_3Cl$$

Earlier processes made use of contact process gas which contains 6 – 7 % SO_3 [32], but today it is more usual to use pure sulfur trioxide. Waste-gas problems are thereby dealt with more easily, and a more compact plant design is possible. Production processes vary both in the manner of bringing the two raw materials into contact and in the methods of removing the heat of reaction.

The reactor can be a packed column with a chlorosulfuric acid spray at the top and hydrogen chloride and sulfur trioxide entering at the bottom [33]. In this case, the reaction and the heat removal both take place in one piece of equipment. Alternatively, it is possible to separate these steps by first mixing the components intensively in a separate mixer such as a mixing nozzle [34]; the hot reaction product is then quickly cooled with cold chlorosulfuric acid in a packed column or other suitable unit. Cooling has also been suggested by means of a water-cooled condenser [35]. In one modification of the process, the first product is chlorosulfuric acid which contains a small excess of sulfur trioxide; this is then cooled and saturated with hydrogen chloride in a bubble column [36]. The off-gas is normally scrubbed, first with 98 % sulfuric acid and then with water.

Several patents describe the recycling of recovered hydrogen chloride and sulfur trioxide [37]. The heat of formation of chlorosulfuric acid can be used to evaporate sulfur trioxide from low-percentage oleum. The sulfur trioxide then reacts with hydrogen chloride at subatmospheric pressure [38].

Construction Materials. Because commercial chlorosulfuric acid should contain very little or no iron, suitable construction materials for reactors, containers, etc., are enamel, glass, aluminum, or steel lined with polytetrafluoroethylene (PTFE) [39].

Analysis and Quality Requirements. The usual method for analyzing chlorosulfuric acid involves hydrolysis of the acid to form SO_3 and HCl, followed by determination of the total acidity and chloride content. When the content of HSO_3Cl, H_2SO_4, and free SO_3 or HCl is calculated, the amount of water that is absorbed by SO_3 to form H_2SO_4 must be taken into account. Variations of this method are described in ref. [31]. The presence of sulfuryl and pyrosulfuryl chloride (SO_2Cl_2 and $S_2O_5Cl_2$) must be determined by special methods [32]. Specific color reactions in combination with too low a melting point identify chlorosulfuric acid producing a cherry red color with tellurium powder or a moss green color with selenium [6].

A product low in iron is normally required, with the following analysis:

HSO_3Cl	98 – 99.5 %
H_2SO_4	0.2 – 2 %
Free SO_3	0 – 2 %
Free HCl	0 – 0.5 %
Fe	5 – 30 ppm

5. Safety Precautions, Transportation

Because chlorosulfuric acid is very aggressive, the regulations dealing with hazardous materials apply [40]. Chlorosulfuric acid reacts violently with water and, with moist air, decomposes to form hydrogen chloride and sulfuric acid mist. All vessels, pipes, metallic hoses, etc., should be dry. Even small leakages must be prevented. Containers should be handled with care: they must never be dropped because a slight internal pressure may build up. The stopper should be opened only with a wrench and not with hammer and chisel. Containers must be emptied by siphoning and not by the use of compressed air. The access of moist air can be avoided by tight connection to a gas bypass. After each removal of chlorosulfuric acid the container must be tightly closed.

Although chlorosulfuric acid itself is not combustible, it can induce ignition of combustible materials. Reaction with moist metals can produce hydrogen. Naked flames must therefore be avoided in the neighborhood of containers or pipes. If welding must take place, the safety precautions pertaining to hydrogen should be observed. Enclosed spaces in which

chlorosulfuric acid is being handled must always be well ventilated. Acid-proof clothing, goggles, and gloves must be worn. A full face gas mask with a filter element for absorbing acid fumes, and safety boots must be at hand. Polychloroprene coated with a copolymer of vinylidene fluoride and hexafluoropropylene would be a suitable material for gloves, protective clothing, and full-face gas masks. Boots could be made of poly (vinyl chloride). Contaminated clothing must be changed without delay.

Small spillages of chlorosulfuric acid must be washed away with water, and all necessary safety measures should be observed; the spilled material should be downwind and the area cordoned off. A high-velocity jet of water should not be used; the spillage should rather be worked from the outside toward the middle, preferably by means of a hose with a spray nozzle attachment. The area can then be neutralized with sodium carbonate, sodium hydrogen carbonate, or lime.

Chlorosulfuric acid must not be allowed to leak into the sewerage system or the soil. Spilled acid may be blanketed with paraffin oil or FEP film (perfluorinated ethylene – propylene copolymer). The contaminated area will reveal itself by fog formation; and this can be prevented from spreading by shielding with water. Personnel involved in such cleaning operations must wear full protective clothing and use compressed air breathing apparatus.

Transportation. Differing regulations apply to transportation by road, rail, or sea [41]. Road transportation may utilize barrels made from aluminum or iron.

Chlorosulfuric acid should, however, be carried by rail whenever possible. The construction material for tanks of rail or road trucks is usually stainless steel, but enameled steel may also be used.

6. Uses and Economic Aspects [42]

The breakdown of chlorosulfuric acid production according to end use is as follows:

Detergents	40 %
Pharmaceuticals	20 %
Dyes	15 %
Crop protection	10 %
Ion-exchange resins[43], plasticizers, and others	15 %

The following companies are published as manufacturers of chlorosulfuric acid:

Europe [42]:	BASF, Ludwigshafen	D
	Bayer, Leverkusen	D
	UCB, Ostende	B
	EniChem, Pieve-Vergonte	I
	Säurefabrik, Schweizerhalle	CH
	ICI, Runcorn	GB
	Tiszamenti, Szolnok	H
USA [43]:	DuPont, Wurtland, Kent.	US
	Gabrieil Chem., Geismar, Lous.	US
Asia [44]:	Mitsubishi, Kitakyushu	J
	Nippon Soda, Toaoka	J
	Nissan Chem., Neigun	J
	Takuyama, Tokuyama	J
	Choheung, Seoul	KOR
	Chen Yeh, Taipei	Taiwan
	Chung Hwa, Yingko	Taiwan
	Kuo Kai, Taipei	Taiwan

With regard to the annual production of chlorosulfuric acid in recent years, no reliable figures are available in the relevant literature.

7. Toxicology and Occupational Hygiene [45]

Chlorosulfuric acid is a strong acid and dehydrating agent that produces severe burn of the skin. The acid fog formed in moist air strongly irritates the respiratory tract and eyes. Contact with the skin or inhalation of the vapor must be avoided under any circumstances.

If there is accidental contact with the skin or eyes, the affected part must be washed immediately with large amounts of water; medical help should then be sought as soon as possible.

No MAK or TLV has been established for HSO_3Cl but the values for its principal decomposition products are:

H_2SO_4: MAK and TLV 1 mg/m^3 (1995)
HCl: MAK and TLV 7 mg/m^3 (1995)

8. References

General References

1. *Gmelin,* 8th ed.,(1963) Syst. Nr. 9: Schwefel.
2. *Houben-Weyl,* 4th ed., Georg Thieme, Stuttgart 1955, vols. IV/2 and IX.
3. C. M. Suter, A. W. Weston in R. Adams (ed.): *Organic Reactions,* vol. III, John Wiley & Sons, New York 1946.
4. H. Beyer, W. Walter: *Lehrbuch der Organischen Chemie,* 19th ed., S. Hirzel, Stuttgart 1981.

Specific References

5. A. Williamson, *Proc. Royal Soc. VII, London* **1854/55,** 11 – 15.
6. C. R. Sanger, E. R. Riegel, *Z. Anorg. Chem.* **76** (1912) 79 – 128.
7. A. Simon, H. Kriegsmann, *Z. Phys. Chem.* **204** (1955) 369 – 392.
8. E. Korinth, *Angew. Chem.* **72** (1960) 108 – 109.
9. R. Vogel-Högler, *Acta Physic. Austriaca* **1** (1948)323 – 338.
10. R. J. Gillespie, E. A. Robinson, *Can. J. Chem.* **40** (1962) 644 – 657.
11. Unpublished results, R. Seelemann, Bayer AG, Leverkusen.
12. G. E. Chivers, R. J. W. Cremlyn, T. N. Cronje, R. A. Martin, *Aust. J. Chem.* **29** (1976) 1573 – 1582.
13. M. Goehring, in *Scientia Chimica,* vol. 9, Akademie-Verlag, Berlin 1957, pp. 110 – 113.
14. H. Gerding, *J. Chim. Phys. Phys. Biol.* **46** (1949) 118 – 119. R. J. Gillespie, E. A. Robinson, *Can. J. Chem.* **40** (1962) 675 – 685.
15. G. H. Weinreich, *Bull. Soc. Chim. Fr.* 1960, 791. W. Traube, *Ber. Dtsch. Chem. Ges.* **46** (1913) 2513 – 2524.
16. R. J. W. Cremlyn, T. N. Cronje, *Phosphorus Sulfur* **6** (1979) 459 – 504.
17. E. Gebauer-Fuelnegg, H. Figdor, *Monatsh. Chem.* **48**(1927) 627 – 638.
18. *Ullmanns Encyklopädie,* 4th ed., **vol. 8, p. 414,**Verlag Chemie, Weinheim 1975. J. C. D. Brand, W. C. Horning, *J. Chem. Soc. (London)* 1952, 3922.
19. G. Schroeter, DE 634 687, 1933.
20. Chemische Fabrik von Heyden, DE 752 572, 1941 (E. Haack, R. Jacob).
21. H. Staudinger, H. A. Bruson, *Liebigs Ann. Chem.* **447** (1926) 110 – 122.
22. K. E. Jackson, *Chem. Rev.* **25** (1939) 67 – 119.
23. O. Ruff, *Ber. Dtsch. Chem. Ges.* **34** (1901) 3509 – 3515.
24. D. J. Salley, *J. Amer. Chem. Soc.* **61** (1939)834 – 838.
25. P. Claesson, *J. Prakt. Chem. 2,* **19** (1879) 231 – 265.
26. R. Levaillant, L.-J. Simon, *C. R. Acad. Sci.* **169** (1919) 140 – 143.
27. P. Baumgarten, *Ber. Dtsch. Chem. Ges.* **64** (1931) 1505.
28. M. Müller, *Ber. Dtsch. Chem. Ges.* **6** (1873)227 – 231.
29. K. Fuchs, E. Katscher, *Ber. Dtsch. Chem. Ges.* **60** (1927) 2288 – 2296; DE 505 687, 1928.
30. H. Beckurts, R. Otto, *Ber. Dtsch. Chem. Ges.* **11**(1878) 2058 – 2066.
31. I. G. Farbenind. AG, DE 582 265, 1930 (E. Konrad, H. Kleiner).
32. Soc. An. Industrie Chimiche, IT 395 404, 1942 (Baslini) DE 914 733, 1941 (H. Beyer)
33. BASF, DE 1 226 991,1964 (H. Wolf, W. Klingler).
34. BASF, DE 2 059 293,1970 (H. Wolf, W. Gösele, S. Schreiner).
35. The Sulfur Institute, Washington, D.C., DE 1 767 169, 1968 (M. Schmidt).
36. Hoechst, DE 2 730 011, 1977 (R. Borger, E. Malow, A. Renken, G. Riess).
37. Ciba-Geigy, CH 597 094, 1978 (M. Buerli, A. Henz).
38. Bayer, DE 3 129 976,1983 (K. H. Schultz, F. Baunack).
39. *DECHEMA-Werkstofftabelle,* DWT 475 – 476.
40. Kühn, Birret (eds.): Merkblätter Gefährliche Arbeitsstoffe; Registerband, p. 56; Bd. 2, p. 72 (GefStoffV); Bd. 5, VI – 4, Wassergefährdende Stoffe, p. 36; Bd. 7, C24; Ecomed Verlag, München – Landsberg 6/1994. Merkblätter der Berufsgenossenschaft der Chemischen Industrie, M004 (1987), M051 (1985), M053 (1989), Jedermann Verlag, Heidelberg. StörfallV (Bundesgesetzblatt Teil 1, Nr. 54 of 28.09.1991, p. 1891) and 1st, 2nd and 3rd VwV zur StörfallV. *European Community:* Guideline of 14. 7. 1976 (76/907 EEC), EG-Nr. 016-017-001. *United States:* Manufacturing Chemists' Assoc., Inc., Washington, D.C., 1949: Chemical Safety Data Sheet SD-33. *Japan:* Chem. Sub. Control Law, Existing Chem. Sub. 1 – 327.
41. IMDG Code; class 8, UN-No. 1754; RID/ADR/ADNR: class 8, Ziffer 12 A; GGVSee/GGVE/GGVS: Class 8, Ziffer 12 A, Warntafel: Gefahr-Nr. 088; ICAO/IATA-DGR: not allowed; Post and Express transportation in

Germany not allowed; Kühn, Birret (eds.): Merkblätter Gefährliche Arbeitsstoffe; Bd. 5, Merkblatt 175400 (6/1994); Merkblatt T015 (1990) der Berufsgenossenschaft der chemischen Industrie.
42. *Directory of Chemical Producers – Europe – Vol. 2.*
43. *Chemical Economics Handbook – 10/94.*
44. *Directory of Chemical Producers – East Asia*
45. L. Arzt, *Dermatol. Wochenschr.* **97** (1933) 995 – 997. F. Roulet, O. Straub, *Arch. Gewerbepathol. Gewerbehyg.* **10**(1941) 451 – 459. G. R. Cameron, *J. Pathol. Bacteriol. (Edinburgh)* **68** (1954) 197 – 204. Chlorschwefelsäure, Nr. 248, Toxikologische Bewertung, Ausgabe 12/1995, ISSN 0927 – 4248, Berufsgenossenschaft der chemischen Industrie, Jedermann Verlag, Heidelberg.

Chlorotoluenes → Chlorinated Hydrocarbons
Chlorotoluidines → Toluidines
Chloroxylidines → Xylidines

Chocolate

Jürg Kleinert-Zollinger, Lindt & Sprüngli AG (retired), Zürich, Switzerland
Stephen T. Beckett, Nestlé PTC York, York, United Kingdom

1. Definitions 1
2. Historical 2
3. Raw Materials 3
4. Fermentation and Drying 4
5. Composition of the Fermented Cocoa Bean 5
6. Chocolate Production 5
6.1. Cocoa Bean Processing 5
6.2. Preparation of Cocoa Powder and Cocoa Butter 9
6.3. Chocolate Production 10
6.3.1. Basic Chocolate Mass 10
6.3.2. Precrystallization (Tempering) 13
6.3.3. Processing Precrystallized Chocolate 13
7. Other Chocolate Processes 14
8. Packaging, Storage, Spoilage, Analysis 16
9. Economic Aspects 17
10. References 18

To be called chocolate a product must partially consist of ingredients derived from cocoa beans and strict regulations exist as to how the different chocolate types can be described on their packaging. This article looks at the history of drinking and solid eating chocolate, before describing the growing and processing of the cocoa beans themselves. The chocolate making procedure is then reviewed, followed by some of the methods used to process the liquid chocolate. As the main fat, cocoa butter, can solidify in six different forms, the crystallisation principles are particularly important. Finally the most important countries are compared with respect to their cocoa growing tonnage, chocolate manufacturing output and consumption per head of population.

1. Definitions

Cocoa beans are the fermented and dried seeds of cocoa trees. *Cocoa nibs* (the cotyledons from the beans) are the product obtained from cleaned raw or roasted cocoa beans by breaking them and removing the cocoa bean shell. The residual amount of shell and germ should not exceed 5 % (calculated on fat-free dry cocoa mass) although this is no longer a legal requirement within the EC.

Cocoa mass, also known as *cocoa liquor*, is the product obtained by grinding raw or roasted non-alkalized or alkalized cocoa nibs from which no cocoa butter has been removed.

Cocoa powder is made by partly defatting the cocoa mass by pressing it and then pulverizing the remaining cocoa press cake. Cocoa nibs for producing cocoa mass may be treated (alkalized) with alkali carbonate or other materials which are nutritionally and physiologically harmless, such as ammonium salts, magnesium oxide, or calcium oxide. This makes the powder less likely to sediment out when it is used as a drink. The alkalinity of the cocoa ash must not exceed the equivalent of 160 mL of 1 N alkali hydroxide solution per 100 g fat-free cocoa solids. Ordinary *cocoa powder* must contain a minimum of 20 % cocoa butter calculated on dry solids [1]. *Fat reduced cocoa powder* in the EC describes any powder with less than 20 % cocoa butter. *Sweetened cocoa powder* is a mixture of cocoa powder and sucrose. The proportion of cocoa powder must be a minimum of 25 %, which is the same as for *drinking chocolate*.

Although no longer described in the regulations several types of cocoa butter are found commercially. *Pressed cocoa butter* must have been obtained exclusively by mechanical extraction (pressing) from broken nibs or cocoa mass. The nibs or cocoa mass may have been treated with alkali prior to pressing.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_023.pub2

Expeller cocoa butter may be made by pressing the whole beans including the shell, whereas *refined cocoa butter* can be obtained by solvent extracting the residues from the cocoa press.
Chocolate must contain cocoa mass — for the dark or milk varieties — or cocoa butter, for white chocolate. The proportions of the other ingredients that can be used for a product which can be described as chocolate, varies from country to country. Most, however, follow one of three sets of regulations, i.e., European [2], USA and Codex Alimentarius [3]:

1) *Dark chocolates* are mixtures of cocoa mass or cocoa powder, cocoa butter and sucrose. The proportion of cocoa constituents (fat-free dry cocoa solids and cocoa butter) must be a minimum of 35 % and the proportion of fat-free dry cocoa mass must not be less than 14 % with a minimum of 18 % of cocoa butter. Some dark chocolates contain milk fat in order to extend the shelf-life by maintaining the gloss for a longer period.
2) *Milk chocolates* are mixtures of cocoa mass or cocoa powder with cocoa butter, sucrose, and dried milk solids. The proportion of cocoa constituents (fat-free dry cocoa solids and cocoa butter) must be a minimum of 25 % and the content of milk powder 14 %, of which milk fat must be 3.5 %. In the USA the legislation states a minimum of 12 % milk solids of which 3.39 % must be fat. The dried milk products employed must not contain any antiseptic substances or preservatives. In Europe as well as sucrose other sugars such as glucose and lactose can be used.
3) *Household milk chocolates* (called family milk chocolate in UK and Ireland) have a milk content with a minimum of 20 % of which as least 5 % must be milk fat. The proportion of cocoa may be lower, with a minimum of 20 % dry cocoa solids being permitted. Both types of milk chocolate must have a total fat content above 25 %. Limits are also set on the amounts of emulsifier (usually lecithin).
4) *White chocolates* (or cocoa butter confectionery) are mixtures of milk powder, sucrose, and cocoa butter. The proportion of cocoa butter must be a minimum of 20 % and that of milk powder 14 %, 3.5 % of which is milk fat. The total fat content must be a minimum of 25 %.

The chocolates mentioned above may contain additions of dried fruits, hazelnuts, almonds, walnuts, pistachios, pine kernels, etc., either whole or chopped. Filled chocolates are products which contain a solid, plastic, or liquid filling. In many countries special vegetable fats are added in order to modify texture or provide added heat resistance in hot climates. As long as less than 5 % is used, the product can still be called chocolate, although the packaging must clearly state that the product contains vegetable fat.

2. Historical

The cocoa tree originated in the northern part of South America. It was already being cultivated in Central America at the beginning of the 12th century, and was known in Mexico long before the arrival of Christopher Columbus. Cocoa was held in high regard by some of the peoples and was used both as a food and as a currency. Some of the taxes and tributes were paid in cocoa beans. Cocoa beans were a versatile resource during wars when they served as either money or food, depending on the immediate need. When Fernando Cortez conquered Mexico, he discovered large stores of cocoa beans in Montezuma's royal palace. The Spaniards soon learned to prepare a thick beverage, called chocolatl, from roasted cocoa beans, water, maize, and other ingredients along with such seasonings as Spanish pepper, vanilla, and cinnamon.

Columbus introduced this cocoa beverage to Spain, where it was initially received with little acclaim. The Spaniards began to cultivate sugar cane in the newly discovered American territories and soon recognized that the addition of sugar to the cocoa mass considerably improved the taste of the beverage. Chocolate made from cocoa products probably became known in France at the beginning of the 17th century, when it was introduced by Jews expelled from Spain. Between 1606 and 1640, cocoa beans became known in Italy, Holland, England, Germany, and Switzerland, even before coffee and tea. Initially chocolate was a luxury item. Chocolate making developed into a craft in the 18th century, but the processing of cocoa beans still closely resembled the old Mexican method. The cocoa kernels

were ground with a hand roller on a heated stone slab [4, 5].

Commercial production reduced prices and the sweet delicacies became widely accessible. In the 19th century, large factories mechanically producing chocolate rapidly increased the supply.

In Holland, Conrad van Houten improved drinking chocolate in 1828 by pressing out part of the fat from the cocoa mass and grinding the resultant cocoa press cake into a powder. Milk chocolate was invented by Daniel Peter in Switzerland in 1876. Rodolph Lindt invented a method for making chocolate taste smother. In 1879 he began to process chocolate in a machine, called a conche because of its form (Fig. 1). In the conche, chocolate was stirred in a reciprocating manner for several days. Heat, caused by the internal friction, drove off volatile substances with the moisture. External heat was only applied when conching was interrupted to avoid cooling of the mass. During the final phase, pieces of cocoa butter were added. These melted gradually and coated the nonfat constituents, producing chocolate with a pleasant taste, which melted rapidly in the mouth.

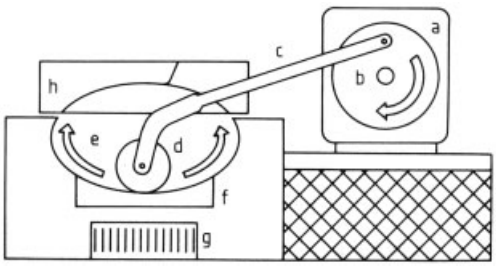

Figure 1. The Lindt longitudinal conche
a) Motor; b) Drive wheel; c) Connecting rod; d) Roller; e) Conche basin; f) Conche bottom (granite slab); g) Heater; h) Conche cover

3. Raw Materials

The cocoa tree *Theobroma cacao* is cultivated today in tropical rain forest areas around the world, i.e., within 20° north and south of the equator. In warm and humid areas, it grows commercially up to altitudes of 600 m [6].

Most areas of cultivation produce two crops a year: a main crop with high yield and good quality and a second crop of lower yield and quality. The time of harvest depends on climatic conditions. Each tree produces 20 – 30, and sometimes as many as 50, pods, 10 – 35 cm long and 7 – 10 cm wide. Each pod contains 20 – 50 flat, oval beans which are ca. 2 cm long and ca. 1.5 cm wide and weigh ca. 1 g each when dried (Fig. 2). The cocoa bean is surrounded by a thin shell and contains two cotyledons (nibs). Embedded between these is the germ, which is 5 mm long and 1 mm wide.

Cocoa pods must be harvested when fully mature but not overripe; the beans are then fermented and dried. Cocoa pods must be carefully cut from the trunk and branches to avoid damaging the flower cushions.

The yield of fermented and dried cocoa beans per tree is normally 0.5 – 2.0 kg, with a maximum of 4 kg. The useful life of a cocoa tree is ca. 40 years [6]. Cocoa trees are of three basic types, Criollo (native to Spanish America), Forastero (foreign), and Trinitario, usually regarded as a hybridization between Forastero and Criollo trees.

Criollo Cocoa. The pods are elongated with a diameter less than half of the length. The husk shows five deep furrows and five shallow furrows. Its surface is warty and the wall is relatively soft. Criollo trees are very sensitive to external influences such as wind, temperature, dryness, and pests. They produce beans of superior quality but in low yield.

Forastero Cocoa. The pods are rounder; i.e., the diameter is more than half of the length. The hard, smooth wall of the husk has either indistinct longitudinal furrows or none at all. Forastero cocoa trees are robust and produce high yields, but the beans are of lower quality.

Trinitario Cocoa. The pods have a range of shapes and sizes and are often reddish in color. The beans are sometimes colored as those of Criollo. Trinitario beans have the strong flavors of Forastero cocoa, with additional notes, which may be fruity, caramel, or spicy in nature. The trees are higher yielding than Criollo and are the major source of fine flavor beans.

Pests and Diseases of the Cocoa Tree. Cocoa trees in plantations are vulnerable to an attack by pests and diseases [6]. The output of an entire country may suffer as a consequence and the diminished supply may drive the price up.

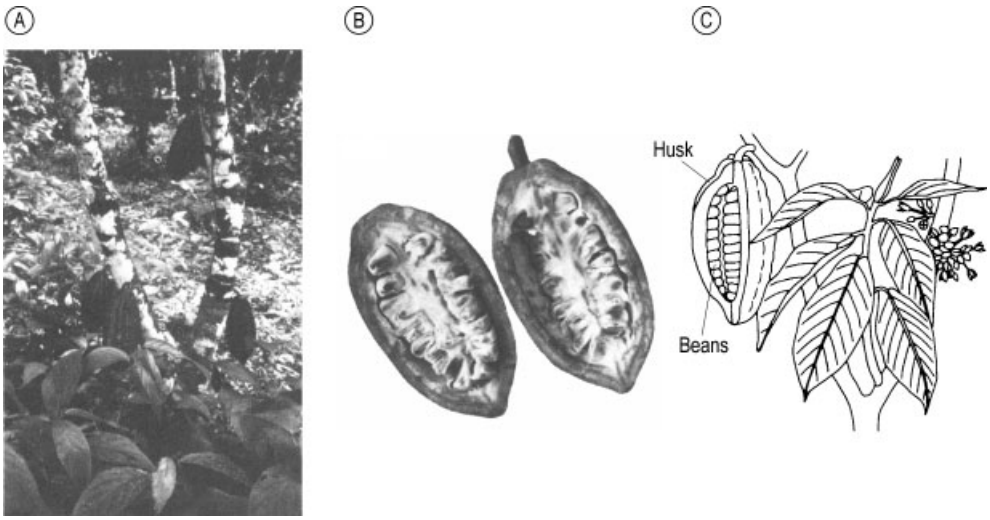

Figure 2. Cocoa plantation (A), cross-section of a cocoa pod (10 – 35 cm long) (B), and depiction of a cocoa tree with a pod cut open (C)

Black pod disease (Phytophthora) is more or less widespread in almost every cocoa-growing country. It can attack the fruit to cause the pods to rot or the trunk and branches to cause cocoa canker. The Phytophthora fungi produce spores which are spread by rain or insects. Infected pods cause far more damage than infected tree trunks and branches. The fungus develops particularly well in very humid areas of cultivation with high precipitation and relatively low temperatures such as Costa Rica, Nigeria, and Cameroon, and can cause losses of up to 50 %. Fungicides and prompt removal of infected pods help control this disease.

Witches' broom disease causes broom-like growths and deformities. These shoots are thicker than the healthy ones and carry many short laterals with underdeveloped leaves. This disease has caused considerable damage in Brazil, Surinam, Trinidad, Ecuador, Columbia, Grenada, and Venezuela. Cutting out and burning the affected shoots reduces loss.

Swollen-shoot disease has caused serious damage in West Africa, mainly Ghana. It is caused by a virus spread by insects. The virus infection can be lethal, and infected trees should be removed and burned.

Insect Pests. The most destructive insect pests are the capsids (*Sahlbergella singularis, Distantiella theobroma,* and *Helopeltis*). They deposit their eggs in the pods, fruit stems, and branches. Both the insects and their young feed on the cocoa plant tissue. Insecticides are effective against capsids, but they can also damage young cocoa plants.

4. Fermentation and Drying

The harvested cocoa pods are broken carefully using a machete or wooden club. The seed mass is removed by hand.

Fermentation facilitates the separation of the cocoa bean from the pulp, destroys its viability, and forms the flavor precursors in the cocoa cotyledons (nibs).

Heap Fermentation. The cocoa beans removed from the pods are heaped on banana leaves and covered with more leaves. The beans are turned and mixed after 48 – 96 h and then recovered, depending on the progress of the fermentation. Complete fermentation requires about 5 d and is followed by sun-drying.

Box Fermentation. The cocoa beans are placed in containers of wood or (less often) concrete, turned with a shovel every 1 – 2 d and subsequently dried. The complete fermentation process, including drying, normally takes 5 – 7 d.

Basket Fermentation. The cocoa beans are placed in leaf-lined baskets and mixed by transfer into other baskets.

Fermentation juice can drain from the basket, and the cocoa beans are thoroughly fermented.

Criollo beans require 2 – 3 d for fermentation, whereas Forastero varieties require 5 – 7 d. The fermentation yield (i.e., recovery of dried beans) is 31 – 46 % if the residual moisture content is 6 – 8 % (after drying). The fermentation time of Trinitario is very variable, depending on the country where it is grown.

Drying. Fermented cocoa beans contain ca. 60 % water. This must be reduced to less than 7 % to obtain produce of commercial quality. The fermented beans can be sun-dried directly on concrete terraces, on bamboo mats, or on mobile drying racks. In some countries beans are dried using hot air, which is heated artificially. Care must be taken to ensure that smoke does not come in contact with the beans, otherwise the beans take up a smoky flavor and become unusable. Drying must proceed slowly to avoid case hardening and to allow the volatile acids to escape. Drying requires 6 – 7 d, depending on weather conditions, or at least 14 h under artificial conditions. The flavor development continues throughout the drying process.

Chemical and Biological Processes [7 – 9]. Correct fermentation is essential to produce a good chocolate flavor in the final chocolate. Although not imparting the actual flavor itself fermentation does give the chemical precursors, which are later developed at the roasting and conching processing stages. During fermentation the pulp surrounding the beans undergoes an ethanolic, acetic, and lactic acid fermentation. The acid and heat kill the beans, change the cell membranes and facilitates enzyme reactions. The actual flavors and colors produced depend upon the type of bean and fermentation used.

In early fermentation the major saccharide present is sucrose, but this is soon hydrolyzed into glucose and fructose. These sugars are capable of reacting with amino compounds and are therefore very important flavor precursors.

In addition numerous nitrogen-containing precursors are formed including amino acids and peptides.

During the last stages of fermentation oxygen-mediated reactions take place, including the oxidation of protein – polyphenol complexes, which had been formed earlier. This results in a decrease in the bitterness and astringency of the cocoa [6].

A wide range of microbiological flora is present on the beans, ranging from lactic bacteria, to bacillus, yeasts, and molds [7].

The lactic bacteria are important in developing some of the acidic flavors, sometimes required for certain chocolate. Other flora such as molds are able to produce lipases, which will cause off-flavors following the subsequent cocoa powder or chocolate production.

5. Composition of the Fermented Cocoa Bean

There are normally 90 – 100 beans per 100 g of cocoa, depending upon the variety and origin [10]. The proportion of shell is normally 12 – 15 % [11]. The composition of the fermented and dried cotyledons (nibs and broken nibs) depends upon the variety, country of origin, climatic conditions, and time of the harvest. Table 1 shows the average compositions of broken nibs, husks, and germ.

6. Chocolate Production [4, 12]

Cocoa is processed in three stages: preparation of cocoa mass and cocoa powder, preparation of cocoa butter, and manufacture of chocolate. Figure 3 is a flow diagram of the production processes.

6.1. Cocoa Bean Processing

Storage. Cocoa beans are stored in 60 – 65-kg jute or burlap sacks or in silos, or in bin/bays where up to several thousand tons may be stored.

Cleaning. Fermented and dried cocoa beans frequently contain dust, sand, stones, iron, wood, glass, leather, rope, etc. These impurities must be removed before processing, as combustion gas from some of these impurities is particularly detrimental to cocoa flavor, whilst others are very hard and may damage subsequent grinding equipment.

Thermal Pretreatment of Cocoa Beans. To fully develop the flavor, the cocoa beans must be roasted. This can be carried out on the complete beans, on the cocoa nib, or on the liquid

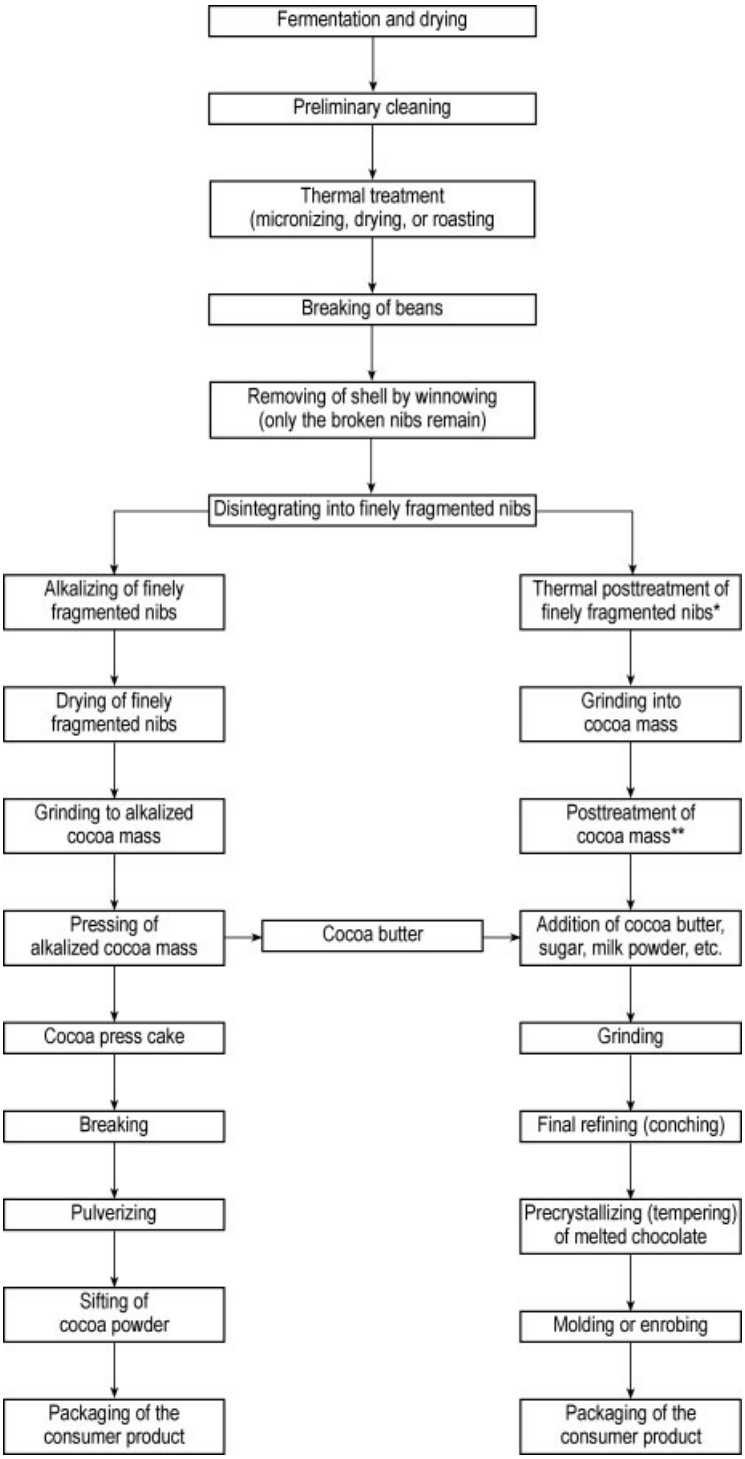

Figure 3. Flow chart for the preparation of cocoa powder and chocolate
* Optional.
** Posttreatment is necessary only if the finely fragmented nibs have received no preliminary treatment.

Table 1. Average compositions of cocoa broken nibs, shell, and germ [4]

Component	Proportion, %		
	Broken nibs	Shell	Germ
Moisture	5.0	4.5	8.5
Fat	54.0	1.5	3.5
Ash	2.5	8.0	6.3
Theobromine	1.2	1.4	
Caffeine	0.2		
Polyhydroxyphenols	6.0		
Total protein *	11.5	10.9	25.1
Sugars	1.0	0.1	2.3
Starch	6.0		
Pentosans	1.5	7.0	
Cellulose	9.0	26.5	4.3
Carboxylic acids	1.5		
Fibers (lignin, cellulose, hemicellulose, pectins, etc.)	0.5	40.1	50.0

* Calculated as (total nitrogen minus purine nitrogen) $\times 6.25$.

cocoa mass. For the latter two processes, the shell must be removed before roasting. As the raw beans have their shells tightly attached to the cotyledons, some form of thermal pretreatment is required to aid the separation. This can be done batchwise or by a continuous process. In the latter case the temperature should be $\leq$ 110 °C and the relative humidity $\geq$ 3.5 %. In the micronizing process [13, 14], infrared radiation heats only the surface of the bean to a high temperature. This incinerates pests and destroys insect fragments. It also detaches the cotyledons from the shell without significantly reducing their moisture content. This pre-roasting separation has the advantage that it minimizes the migration of cocoa butter into the shell. With whole bean roasting a significant amount of migration takes place and this cocoa butter is then thrown away with the shell. Immediately following thermal treatment with either steam or infrared radiation, cocoa beans can be broken while still warm. This improves the yield of broken nibs. Because of the increasing demand for products with a low microbiological count, some companies employ a debacterization process, irrespective of the type of roasting process used. This subjects the beans to a thermal pretreatment under high humidity.

Breaking of Beans and Separation of the Shell. Following thermal pretreatment of the cocoa beans or, in the case of the one-stage process, following the roasting of the beans, the latter are broken and the shell and some of the germ are separated from the broken nibs as completely as possible. Cocoa bean shell is extremely hard and often has sand adhering to it. Both of these materials will severely damage any subsequent milling equipment. A good separation of the shell is, therefore, of the utmost importance. The separation, known as winnowing, can be carried out because the shell is in a platelet form, which can rise in an air stream, whereas the heavier "round" nib particles will fall. A modern cocoa bean breaking and winnowing plant is shown in Figure 4.

In the first screening unit, any broken beans are separated in the drum sifter to reduce (a) the load on the impact smasher (b) and the formation of fines. The broken beans (nibs) are conveyed directly to the oscillating sieves (c). The unbroken beans pass through the impact smasher, where they are hurled singly at high speed against radially arranged exchangeable impact plates to produce a minimum of fines pieces. The mixture consisting of broken nibs and shell is then separated by the oscillating sieves (c); the shell is removed and the unbroken beans returned to (b) via (d). The exhaust air is cleaned in the cyclonic separators (e) and the dust filter (g). Coarse or finely fragmented nibs can either be subjected to alkalization for the production of cocoa butter and cocoa powder or can be ground into cocoa mass.

Alkalization of Cocoa Nibs. Coarse and finely fragmented cocoa beans contain butter, starch and other constituents that are for the

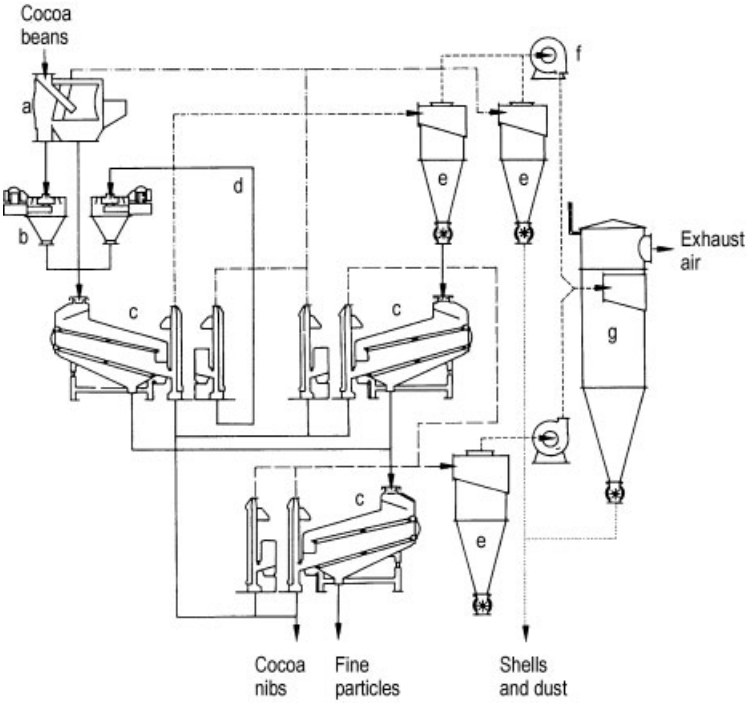

Figure 4. Equipment for breaking cocoa beans and winnowing (separating) out shells [15]
a) Drum sifter; b) Impact smasher; c) Oscillating sieves (separate nibs from the shells); d) Return of unbroken cocoa beans; e) Cyclone separator; f) Aspiration and blower fan; g) Dust filter

most part still enclosed within their cells. The fine nib fragments have a particle size of 1 – 4 mm and are more uniform than cocoa beans in this respect. Chemical treatment, therefore, begins with these particles. For cocoa powder production the nibs are normally treated with alkaline solutions.

As distinct from cocoa mass, in which fat-free cocoa particles are embedded in a continuous phase consisting of cocoa butter, the reaction solution that forms the continuous phase around the broken nibs has unimpeded access to the cocoa constituents with which it can freely react [16, 17]. Subsequent to the reaction, the steam-volatile substances can be removed without hindrance because there is no inhibiting fat layer as in the case of cocoa mass. When cocoa mass is treated, the aqueous reagent solution must diffuse across the hydrophobic fat cover. The adsorption of unpleasant smell- and taste-carrying substances by the cocoa butter and the cocoa starch during drying is then unavoidable [16, 17].

After the reaction, the finely broken nibs are dried and subjected to roasting. This process develops the color and flavor constituents and ensures the destruction of microorganisms, molds, and yeasts.

Grinding of Cocoa Nibs. Cocoa nibs are always ground into a fine cocoa mass. This process is usually carried out in two stages. Pregrinding is performed using cutter or beater blade mills. The nibs are liquefied by cocoa butter, melting during this grinding process. The liquid mass is further refined by roll refiners, ball mills, or disk mills. The cells within the nib must be completely opened up to release the cocoa butter. In general, the finer the cocoa mass is ground, the lower will be its viscosity. Eventually, there is no more fat to release and additional grinding only produces more fine particles which increase the viscosity once again, as is shown in Figure 5. Equipment for grinding nibs and refining cocoa liquor is shown in Fig. 6.

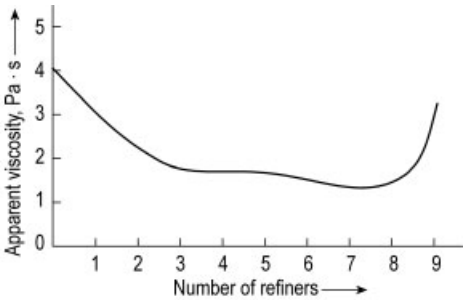

Figure 5. Viscosity of cocoa mass following milling by a three-roll refiner (apparent viscosity at 24 s^{-1} shear rate) after each refiner

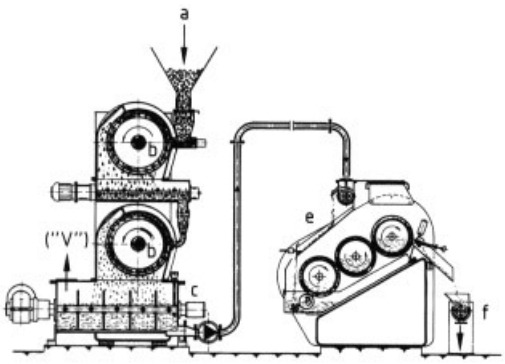

Figure 6. Equipment for grinding nibs and refining cocoa liquor [15]
a) Intake; b) Impact and shear mill; c) Homogenizer; d) Pump; e) Three-roller liquor refiner; f) Liquefier

Treatment of Cocoa Mass [12]**.** Because cocoa mass is a homogeneous liquid at temperatures above 40 °C, it can be roasted uniformly or treated by an alkalizing solution. As was noted earlier, some authors believe that this is not as satisfactory as alkalizing the nibs, because the water repellent fat layers inhibit the wetting of the solid cocoa particles [16, 17].

If cocoa mass is roasted, flavor can be developed by a suitable thin-layer heat exchanger treatment.

Quality Determination. The flavor index [17] (Fig. 7) can be used to determine damage caused by incorrect thermal treatment at any of the processing stages.

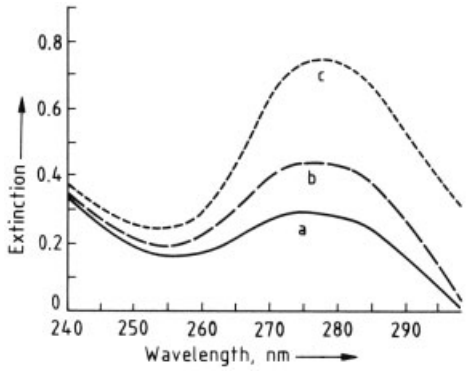

Figure 7. Extinction curves for the determination of the flavor index in cocoa masses [17]
a) Raw cocoa mass; b) Correctly treated cocoa mass; c) Thermally damaged cocoa mass

In order to determine the flavor index the sample is distilled with steam. The extinction of the distillate is measured between 230 and 300 nm with water as described in [17].

6.2. Preparation of Cocoa Powder and Cocoa Butter

Cocoa Powder. Cocoa powder can be made from cocoa mass prepared from cocoa beans that have only been roasted. However, to provide an attractive color and good suspension characteristics in milk or other aqueous solutions, cocoa powder is usually made from alkalized cocoa mass.

Powdered products are made from cocoa mass partially defatted by pressing out some cocoa butter. Modern presses consist of single pots placed side by side. The base of each pot is an exchangeable stainless steel mesh.

The finely ground hot cocoa mass is fed into the press chamber and partially defatted by driving in steel rams, producing a pressure of typically 40 – 50 MPa.

The cocoa press cakes are ejected automatically, fall onto a conveyor band system, and are transported to air-cooled bins. They contain 8 – 24 % fat. The cooled press cakes are broken up by counter-rotating spiked rollers. Figure 8 is a diagram of a pulverizing and sifting plant.

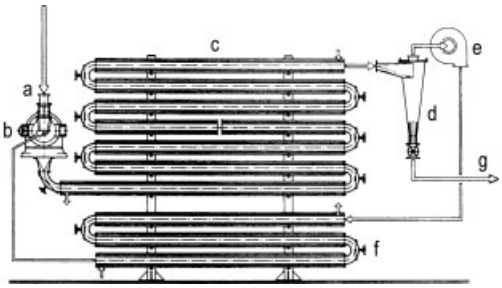

Figure 8. Preparation of cocoa powder [15]
a) Broken cocoa press cake inlet; b) Broken cocoa press cake milling unit; c) Cocoa powder refrigeration unit; d) Cocoa powder cyclone separator; e) Fan; f) Air and cocoa dust cooler; g) To cocoa powder bagging unit

Cocoa powder is widely used as an intermediate product in the food industry for producing fillings, coatings, desserts, ice creams, cocoa and chocolate beverages, etc. Chocolate flavored coatings may be produced by using low fat cocoa powder and a lauric fat (i.e., palm kernel oil or coconut oil that contain high proportions of short-chain fatty acids, e.g., lauric acid) to replace the cocoa mass.

Cocoa Butter. Cocoa butter pressed from the cocoa mass is collected as a pale yellow oil. To prevent the formation of hard layers, which slow the filtration, the ram pressure must be applied gradually. To obtain cocoa powder with about 21 – 22 % fat, about 50 % of the cocoa butter must be pressed out of the cocoa mass. Highly defatted powders require the removal of even more fat.

Continuously operated expeller presses are used to obtain cocoa butter from whole cocoa beans or from low-quality or impaired cocoa beans. Expeller press cakes can either be used as animal feed, or the residual fat can be extracted with a solvent. Expeller fat is of inferior quality compared with cocoa butter pressed from cocoa mass and must be filtered to remove cloudiness.

The demand for cocoa butter with a neutral smell and taste is increasing. Deodorized cocoa butter is produced by steam distillation under vacuum in a batchwise or continuous process.

6.3. Chocolate Production

Chocolate is a mixture of finely pulverized cocoa mass, sugar, cocoa butter, and flavoring substances such as vanillin (chocolate or milk flavors may not be used). Other ingredients, if so declared, may be added either finely ground, e.g., milk powder, and up to 5 % of nuts, coffee pastes, etc., or in coarse pieces, e.g., almonds, nuts, pine kernels, pistachios, raisins, etc.

6.3.1. Basic Chocolate Mass

Mixing and Kneading of Raw Materials. Finely ground cocoa mass, granulated sugar, and part of the cocoa butter are combined in a heavy-duty mixing and kneading machine, e.g., a double-trough mixing unit (Fig. 9). Intensive kneading action produces a homogeneous plasticized mass that can be ground on a roller-refiner. This requires a fat content of 20 – 28 %, depending on the other constituents. For milk chocolate, full-cream milk powder and/or skim milk are also added.

Milling. The most common method of grinding chocolate mass is based on the two-stage process using roll refiners for pre-refining and final refining. Refining breaks the components, wets them with fat, and helps develop the flavor of the basic chocolate mass. A single two-roll refiner (less commonly a three- or a five-roll refiner) is used to pre-refine the ingredients and this is followed by several five-roll refiners operating in parallel, which reduce the solid particles to the final required particle size. The rollers on the refiners are barrel shaped, but become parallel under the high pressures used to crush the chocolate.

Prerefining. Sugar is important because it is a flavor carrier [18]. The chocolate mass that results from grinding crystal sugar and cocoa together is said to be organoleptically superior to that obtained with sugar that has been prerefined on its own. The crushing/shearing action also coats some of the broken particles with fat. Crushing the particles in two stages produces a smaller proportion of very fine particles, than if all the milling was carried out at a single stage, or if the sugar was premilled alone and then added to the other ingredients for roll refining.

Final Refining. Having gone through the prerefiner, the basic chocolate mass is milled to the required final particle size in a five-roll refiner. The rollers are hollow and can be cooled or heated with water. The rolls increase in speed as the chocolate mass passes up the refiner. This

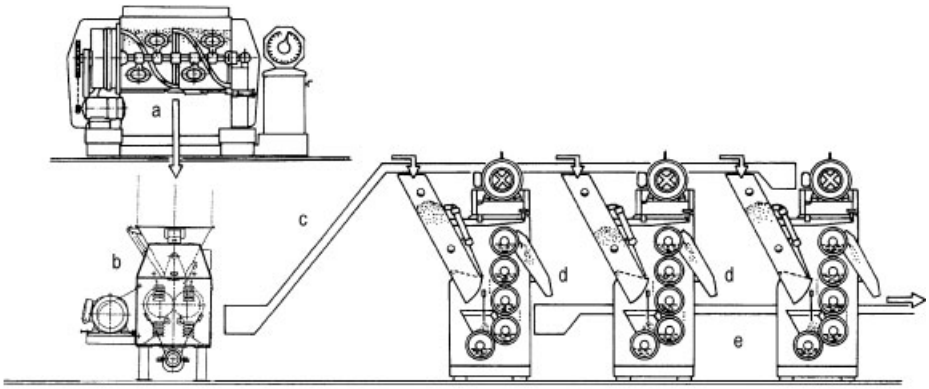

Figure 9. Mixing and two-stage refining system [15]
a) Chocolate mixer; b) Chocolate prerefiner; c) Conveyor (prerefined chocolate); d) Chocolate fine-refiners; e) Conveyor (refined chocolate)

causes the film of material between the rolls to become thinner and each subsequent gap between the rolls to decrease. The final fineness is largely controlled by adjusting the gap between the first two rolls, manipulating the roll temperature and by changing the relative speeds of the rolls. As with the two-roll refiner, the shearing action coats some of the solid particles with fat. A knife scrapes a thin film of chocolate powder or flakes from the fifth roll and conveys it to the final texture- and flavor-developing process (Fig. 10). The specific surface area of the nonfat constituents grows rapidly with increasing fineness. In addition, the organoleptic properties of the chocolate are significantly influenced by its fineness and texture. Top-quality products leave less than 1.5 % of the fat-free chocolate constituents on a 30-µm sieve, or have a 90th percentile of the mass distribution of < 25 µm.

Final Texture and Flavor Development. The basic chocolate mass is obtained after refining as a dry powder or thick paste; it must then be plasticized and liquefied to give the flow properties needed for further processing and to develop the required organoleptic properties. This can be achieved by a number of very different machines (conches) which operate on the kneader or rotation principle. The conching process normally takes place in three stages: (1) dry conching; (2) plasticizing; and (3) liquefying.

Dry Conching. The basic chocolate mass obtained after refining is directly conveyed to a conche, e.g., the Petzholdt High-Speed Texture

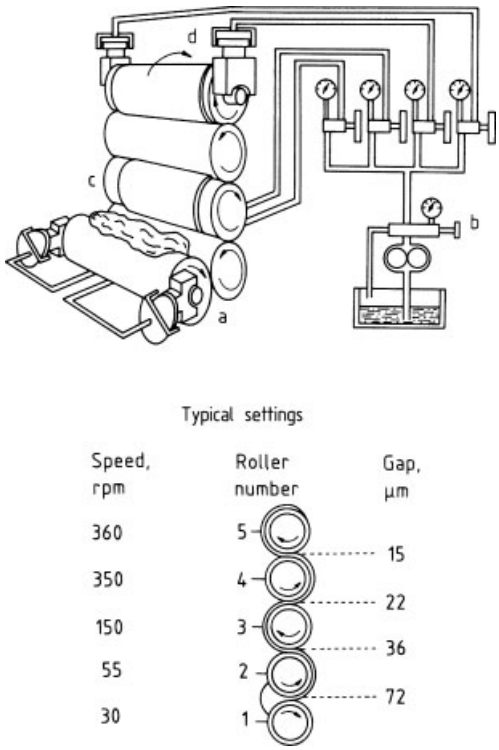

Figure 10. Hydraulically controlled five-roll chocolate refiner [15]
a) Hydraulic roller system; b) Hydraulic control; c) Feed; d) Removal of refined mass

and Flavor Developing Machine or the Thouet DRC Dry Conche (Fig. 11). Inner friction raises the temperature steadily, which causes the evaporation of moisture and some volatile flavor

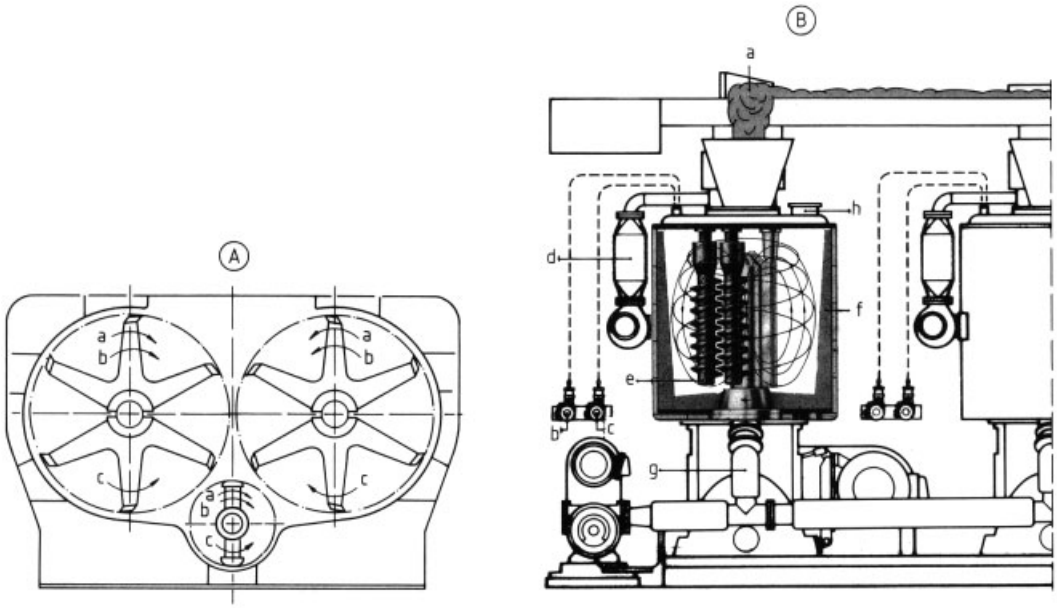

Figure 11. Diagram of two types of machines for the organoleptic modification of chocolate
A: Thouet DRC Dry Conche [19]. The arrows indicate the direction of turning during:
a) Dry conching, dehumidifying, and removal of volatile substances; b) Plastizing with cocoa butter; c) Liquefying with residual cocoa butter and lecithin
B: Petzholdt High-Speed Texture and Flavor Developing Machine
a) Refined chocolate; b) Cocoa butter; c) Lecithin; d) Air heating; e) Twin screws; f) Scraper; g) Chocolate discharge; h) Exhaust air

components. The chemical reactions are not completely understood. The volatile substances can be identified by gas chromatography. These are mainly acetaldehyde, acetone, isobutanol, ethanol, isopropanol, ethyl acetate, isopentanol, methanol, and diacetyl, which occur in cocoa in traces. The temperature should not rise too rapidly; otherwise the moisture can not escape from the conche and instead will re-enter the chocolate and cause the sugar to stick together in agglomerates. This in turn gives a gritty tasting chocolate. The dry conching stage may take up to a few hours, the actual time depending on the moisture in the original ingredients and the conche type and temperature. The latter also affects the taste of the chocolate. Very mild flavored chocolates are usually conched below 60 °C, with high temperatures of 80 °C to above 100 °C being used to introduce a cooked flavor.

Plasticizing. Once most of the moisture has been evaporated the plasticization begins. This consist of deagglomeration of the finely ground particles, optimum fat distribution and the surrounding of the nonfat constituents by a film of fat.

Liquefaction. After plasticizing, the chocolate mass is liquefied by the gradual addition of the residual cocoa butter and emulsifier while being subjected to continuous mechanical working. It is normally considered preferable to add the lecithin or other emulsifiers once the chocolate has been cooled below 50 °C. Chocolate mass with optimum flow properties is the end result. The liquefied chocolate is then pumped into storage tanks fitted with agitators.

Storage. In the past, the finished chocolate was poured into trays holding 40 – 50 kg each, turned out after solidifying, and kept in blocks in cooled storage rooms. Although these are still available, nowadays chocolate is often sold as smaller pieces or chips, up to about 2 cm in diameter. However, both cooling and remelting require energy, and melted chocolate masses are now held in large heated tanks, which are provided with agitators, until they are processed further. Because aging reactions accelerate with

rising temperatures, storage of chocolate in liquid form is advisable for short periods only. The temperature must be kept above 40 °C to prevent crystallization of the higher-melting glycerides, which are important for the subsequent precrystallizing (tempering) process, although white chocolate should be stored at lower temperatures. Vertical tanks are provided with agitators to prevent sedimentation of the bigger particles and separation of the fat.

6.3.2. Precrystallization (Tempering) [20, 21]

The precrystallization process (tempering) is of equal importance as the quality of the raw materials, the formulation, and the correct processing of chocolate masses with regard to the keeping qualities, heat resistance, gloss, snap, melting properties, texture, etc. of a chocolate. Cocoa butter can crystallize in six different forms (Fig. 12). Only form V (β_2) will give these desirable properties. If chocolate is made with the less stable forms, not only will it not contract properly (because the crystals are not so compact), but it will rapidly change its form, pushing fat to the surface. This gives a white mold-like appearance known as chocolate bloom.

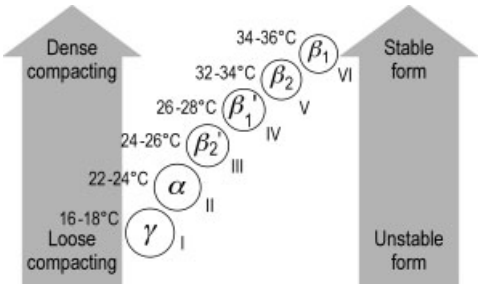

Figure 12. Different crystal forms of coca butter as a function of temperature

During the solidification of the melted chocolate mass, the fat phase must build a structure that consists of these type V (β_2) crystals throughout. Tempering is, therefore, the most important factor in the entire solidification process.

A large number of crystals of the correct type must be distributed throughout the chocolate. This must be done rapidly in order to meet with the needs of the chocolate usage plants. Cooling cocoa butter to 32 °C would take many hours to reach this state, and so alternative methods must be used. The most commonly used tempering procedure is to cool the chocolate much further, so as to produce a lot of both stable and unstable crystals (Fig. 13). The chocolate is then reheated to melt the unstable ones. The process takes place inside a rapidly rotating machine, which breaks up the larger crystals and distributes the resulting smaller ones throughout the chocolate.

The degree to which a chocolate mass is tempered can be evaluated by means of a tempering diagram. This is obtained by rapidly cooling a small quantity of chocolate and measuring its temperature as it sets. Figure 14 A shows a tempered sample, which cools slowly, because the latent heat of solidification comes out quickly and matches the cooling. The chocolate then sets rapidly. Figure 14 B shows a lightly tempered chocolate, which cools slowly and even reheats, but sets more slowly than that in Figure 14 A. The sample in Figure 14 C is untempered and will take a very long time to set and does not contract properly.

6.3.3. Processing Precrystallized Chocolate

The precrystallized chocolate mass can be used for molding and enrobing. Both cases require correct precrystallization, followed by careful cooling. Too fast a cooling rate may result in the wrong types of crystals being formed.

Molding. The molding process permits the preparation not only of blocks, bars, or single pieces with or without the addition of coarsely chopped nuts, raisins, croquants, etc., but also of products with a solid or liquid center completely enclosed in chocolate.

Solid products are made by dosing the precrystallized chocolate, with or without coarsely chopped additions, into molds where it is distributed in the mold by vibration and scraping. It is subsequently cooled and demolded. The preparation of filled chocolate bars and chocolates requires more complex equipment.

The hollow chocolate shells are usually made by first of all filling the molds as for the solids bars. The chocolate is allowed to partially set before the mold is inverted and the central liquid chocolate allowed to run out. The center filling,

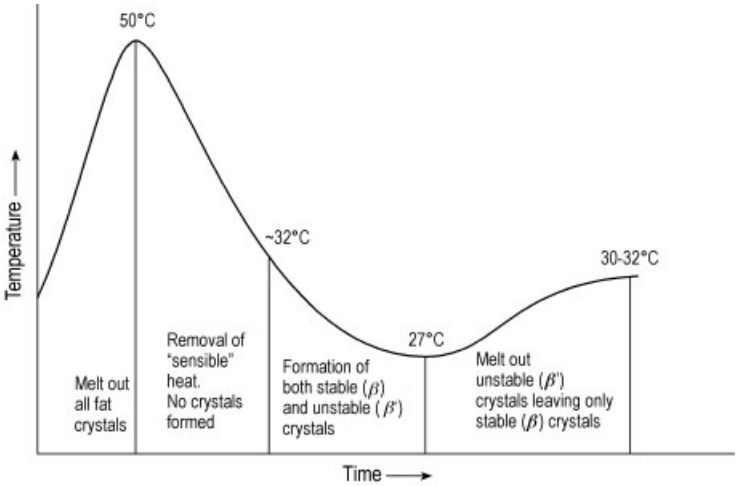

Figure 13. Milk chocolate tempering sequence (Reprinted with the permission of Loders Croklaan)

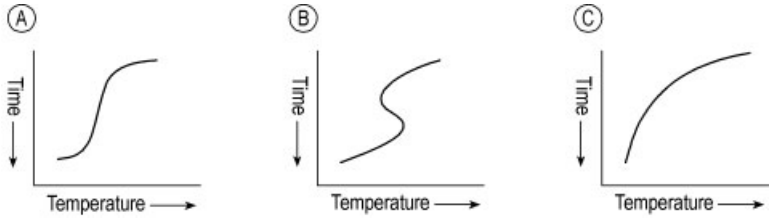

Figure 14. Cooling curves of tempered (A); lightly tempered (B); and untempered chocolate (C)

e.g., praline, fondant, or caramel is then poured in, before a coating of liquid chocolate is used to complete the shell.

Enrobing. The centers to be enrobed (e.g., nuts, marzipan, wafers, etc.) are heated to a slightly lower temperature than that of the tempered chocolate, before being placed on a wire grid conveyor belt. This transports them through one or two curtains of tempered chocolate (see Fig. 15). The product is then subjected to vibration and being blown in an air curtain, which removes the excess chocolate and gives the required weight. The excess chocolate falls through the wire grid to be recirculated onto another batch of centers. Once again careful cooling is required. Where the center is very rough or a thick layer of chocolate is required, two enrobers may be used, before the product is cooled.

7. Other Chocolate Processes

Crumb Process [22]. Crumb originated in the 1930s when milk powder had poor keeping properties. By drying the milk together with the sugar and cocoa mass (which contains natural antioxidants) a chocolate making ingredient was developed, which has a long shelf-life. The milk is partly evaporated, mixed with a 60 % sugar solution, and concentrated to a form that is 87 – 90 % dry solids. The cocoa mass or cocoa powder is then added and, after being dried to a residual moisture content of ca. 2 %, the crumb is pulverized. The proportion of cocoa mass is 6 – 12 %.

Figure 15. Chocolate enrober

To produce milk chocolate, the crumb powder requires only the addition of cocoa butter, flavors, and lecithin. The flavors of the cocoa are readily absorbed by the milk crumb because the carbohydrates, sucrose and lactose, in many crumb making processes are present predominantly in an amorphous form and only subsequently become crystalline. The proteins are also distributed in the crumb in a colloidal form. The crumb powder is worked in heavy-duty mixing and kneading machines while cocoa butter and flavorings are added. The resulting mass is then roll refined and then conched as for chocolates manufactured from milk powders. Sometimes the conching time is shortened, however, because some of the flavor changes, in particular cooked Maillard type notes, have already been developed during the crumb making process.

Low Calorie Chocolate. As with many other foods, there has been a demand in many of the developed countries for low calorie or low fat products.

One of the most widely used methods of reducing the calorific value of chocolate is to replace the sugar by sugar alcohols such a maltitol or lactitol together with some polydextrose. The sugar alcohols often form agglomerates and become gritty at higher conching temperatures, so some care has to be taken in processing them. In addition they must be eaten in limited amounts due to their laxative effects. Polydextrose has a lower calorific value than the sugar alcohols, but tends to give a drying effect in the mouth.

Another way to reduce the calorific value is to use a low calorie fat such as Benefat [23]. This is only permitted in a few countries, and because the fat is not very compatible with cocoa butter, the chocolate must normally be manufactured from cocoa powder.

The calorific value can also be reduced by lowering the fat content. However, it is necessary to maintain a viscosity similar to that of standard chocolate in order to enable processing, and to obtain satisfactory eating properties. Maintenance of viscosity is normally achieved by modifying the particle size distribution to have less surface of solid particles to coat with fat; by increasing levels and modifying types of emulsifiers; and by improved mixing/conching machinery [24].

Heat-Resistant Chocolate [25]**.** There are four processes which can be used to make chocolate more heat resistante for sale in hot climates;

1) Modifying the fat phase
2) Heat cycling
3) Adding moisture
4) Building a carbohydrate structure

Modifying the Fat Phase. Milk fat softens chocolate, so this type of fat should be reduced. In addition cocoa butter obtained from cocoa grown nearer the equator is harder to melt than that grown further away. Consequently chocolate made with Malaysian cocoa butter is much more heat resistant than with that obtained from West African, which in turn is better than most Brazilian product. Certain vegetable fats will also make chocolate much more heat resistant, but in many countries the product cannot be sold under the name chocolate. Some of these fats are also very effective at reducing chocolate bloom.

Heat Cycling. By carefully adjusting the temperature of the chocolate before manufacture, or of the finished product, it is possible to improve both heat and bloom resistance. Figure 16 [26] illustrates the use of cyclothermic precrystallization to mold better (dark) chocolate at 35 – 36 °C. Finished product can be improved by heating to 29.5 ± 0.5 °C [27] for 6 h.

Adding Moisture. Water rapidly increases the viscosity of chocolate by sticking together the sugar particles. This makes it very difficult to process the chocolate. A number of ways of getting around this problem have been devised; such as adding the water as an emulsion, or mixing it very rapidly with tempered chocolate just before molding. Another patent describes wrapping the product in a water permeable wrapper and then storing it in a moist atmosphere.

Building a Carbohydrate Structure. This process strengthens the texture of the chocolate. The nonfat substances of conventional chocolates are enveloped by a fat film that liquefies when heated and causes the system to flow. In heat-resistant chocolates, the nonfat substances form a continuous phase and produce a structure by partial adhesion. The fat is enclosed in isolated hollow spaces [22].

One possible method for building a carbohydrate structure is as follows: a sugar solution is made up of, for example, 80 % sucrose and 20 % invert sugar or glucose syrup. This solu-

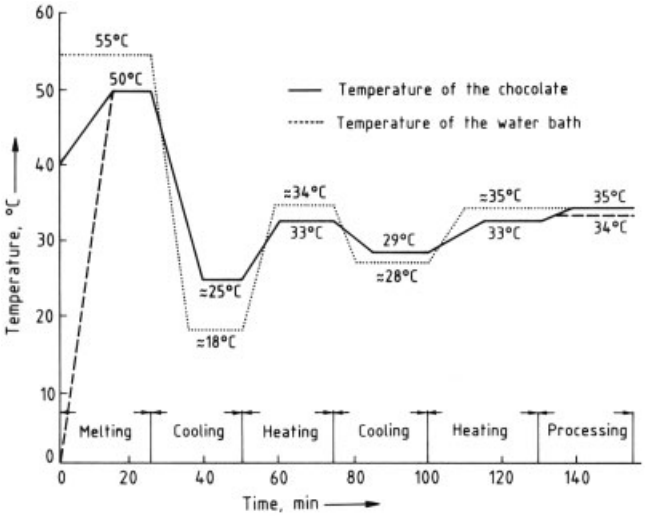

Figure 16. Temperature program for the cyclothermic preparation of heat-resistant chocolates [26]

tion is concentrated under vacuum to a residual moisture content of 1 – 2 % and cooled from a maximum of 150 °C to ca. 80 °C. This concentrated sugar is made into a homogeneous mixture of ca. 90 % sugar and 10 % butter or other fat (at ca. 150 °C for easy processing). The mixture solidifies when cooled. It is then pulverized to an extremely fine powder in a micronizer mill; the particle size should be <30 µm. Grinding must be carried out at 20 – 25 % relative humidity. The sugar – fat mixture is next homogeneously mixed with a finished bitter chocolate or milk chocolate mass at <50 °C, and the resultant mass is precrystallized, molded, and cooled to 15 ± 1 °C. The chocolate bars are wrapped in aluminum foil immediately after demolding and are hermetically sealed. The wrapped bars must be stored for 30 d at 25 ± 1 °C to allow the sugar structure to develop.

The heat-resistant texture can be tested by immersing pieces of the chocolate in ether for several hours; whereas ordinary chocolate breaks down because the fat phase is dissolved, chocolate with a sugar matrix remains practically unchanged [22].

8. Packaging, Storage, Spoilage, Analysis

Packaging. Chocolate bars and filled chocolates must be packaged in odor-free materials (no odor or taint may come from the material itself, inks, lacquers, or the cold seals applied). Light protection is also required in case of white chocolate. To guarantee the quality of the product over several years, packaging with an oxygen barrier is occasionally required. Aroma barriers are also sometimes necessary, depending on the environment and length of storage of the packaged products.

For chocolate tablets, hermetically sealed lacquered or laminated aluminum foil provides the best protection against the entrance of foreign odors, loss of flavor, and infestation of pests. Good quality foil should be used as thin material with numerous pinholes will not provide good protection. The foil, which is laminated to a heat-sealing film, for example, polypropylene, is heat-sealed and wrapped in a printed sleeve, envelope or carton. Double folds and envelope-type wraps (traditional chocolate tablet packaging, that is not heat-sealed) only offer limited protection. Printed opaque or metallized oriented polypropylene films offer very good protection and are frequently used in the flow wrapping of chocolate bars. These materials are more cost effective than aluminum foil laminates and can also be used to flow wrap chocolate tablets. In this case, however, the traditional appearance of the pack (paper outer envelope wrap with foil inside) is lost.

Table 2. World production of cocoa beans in 2001/02 and 2002/3 in 1000 t [30]

Country	2001/2002	2002/2003*
Côte d'Ivoire	1265	1300
Ghana	341	475
Indonesia	455	440
Nigeria	180	165
Brazil	124	150
Cameroon	126	135

[a] Other countries producing between 25 000 t and 90 000 t are Ecuador, Dominican Republic, Papua New Guinea, Malaysia, Colombia, and Mexico.
* Estimated.

Storage. Cocoa and chocolate products must always be stored in a cool, dry, dark, and odor-free atmosphere. The most suitable condition for extended storage is a temperature of 10 – 18 °C with 55 – 65 % relative humidity. Larger temperature fluctuations are particularly detrimental to product shelf life because fat crystallizes out during cooling and causes contraction. When the temperature is again increased, the liquid proportion of the fat phase increases, causing an increase in volume and a consequent increase in pressure. This "pumping effect" changes the texture and accelerates the aging of cocoa and chocolate products.

Spoilage. The visual damage called *fat bloom* is caused not only by improper precrystallization of the melted chocolate (see Section 6.3.2), but also by excessive heat or by soft fat such as hazelnut oil migrating into the chocolate from the center filling.

Sugar bloom only occurs when the surface of cocoa and chocolate products is wet, e.g., through microcondensation, and the sugar dissolves. On drying, the sugar crystallizes out on the surface in coarse crystals. This happens when the cooling tunnels, used to set the chocolate, are operated below their dew point.

Analysis. The main constituents of cocoa and chocolate products are sugar, cocoa mass, cocoa butter, dry milk solids, lecithin, and vanillin. Of analytical interest are moisture, sugars, fat, proteins, and alkaloids. Specific analytical methods have been published by the International Office of Cocoa, Chocolate and Sugar Confectionery (IOCCC) in Brussels. Appropriate measuring methods have also been developed on an international level for the evaluation of chocolate fineness and the determination of flow properties.

9. Economic Aspects

Cocoa beans are referred to commercially according to either their country of origin, main area of cultivation, or port of dispatch. Cocoa beans from different areas of cultivation differ quite considerably in flavor and organoleptic properties, and this is reflected in their prices. The great trading centers for cocoa beans are in New York and London. Chocolate and cocoa powders are usually produced from mixtures of cocoa beans of different origins and blended according to strictly secret recipes.

The production of cocoa beans in the main growing countries is given in Table 2. Due to changes in the economy in countries such as Nigeria and Malaysia, the production of cocoa has tended to decrease, whilst witches' broom disease has had a devastating effect upon Brazilian production. On the other hand, new plantings in the Côte d'Ivoire and Indonesia have in-

Table 3. Production and consumption figures for the top ten (1996) consumer countries for chocolate confectionery in 2003 (metric tons)

	Production	Exports	Imports	Consumption
Switzerland	118 185	61 640	19 090	75 635
Norway	30 300	6 645	16 000	39 655
Germany	985 940	359 570	236 930	863 300
Ireland	35 705	31 970	30 260	33 995
Belgium	189 110	146 795	69 155	111 470
Austria	67 105	47 475	56 345	75 975
UK	486 650	87 120	144 530	544 060
Denmark	31 200	19 520	35 535	47 215
Australia	101 990	29 240	14 873	87 623
USA	1 524 755	124 465	162 210	1 562 500

Source: CAOBISCO Statistical Review 2003 published 2004.

creased the tonnages produced by these countries.

The manufacture of chocolate confectionery has increased dramatically over the last 50 years. The production, export, import, and consumption figures for the top ten consumer countries is given in Table 3 [29] and the consumption figures (kg per capita and year) for 20 major countries are given in Table 4 [29].

Table 4. Chocolate confectionery consumption data (kg/head per annum) [29]

	1996*	2003
Australia	5.4	4.4
Austria	8.2	9.4
Belgium	8.3	7.7
Brazil	1.9	0.8.
Denmark	7.4	8.2
Finland	3.6	5.4
France	4.8	5.2
Germany	8.4	8.6
Greece	2.3	2.1
Ireland	8.4	8.8
Italy	2.6	3.4
Japan	2.0	1.8
Netherland	4.5	4.4
Norway	8.6	8.7
Portugal	1.7	1.8
Spain	1.6	1.5
Sweden	5.2	7.0
Switzerland	9.5	9.6
UK	7.9	8.6
USA	5.3	5.4

* Source: IOCCC/CAOBISCO Statistical Review 1996 published 1997 and 2003 published 2004.

10. References

1. Codex Standard for Cocoa Powders and Dry Mixtures of Cocoa and Sugars, CODEX STAN 105-1981, Rev.1-2001.
2. Directive 2000/36/EC relating to cocoa and chocolate products intended for human consumption, Official Journal of the European Communities, L 197, 3/8/2000, 19-25.
3. Codex Standard for Chocolate and Chocolate Products, CODEX STAN 87-1981, Rev.1-2003.
4. A. Fincke, H. Lange, J. Kleinert: *Handbuch der Kakaoerzeugnisse*, 2nd ed., Springer Verlag, Berlin-Heidelberg-New York 1965.
5. R. Whymper: *Cocoa and Chocolate*, Their Chemistry and Manufacture, 2nd ed., J. and A. Churchill, London 1921.
6. G. A. R. Wood, R. A. Lass: *Cocoa*, 4th ed., John Wiley & Sons, New York 1986.
7. P. S. Dimick, J. C. Hoskin: "Chemistry of Flavour Development in Chocolate," in S. T. Beckett(ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, p. 137 – 152.
8. D. M. Mazigh: "Microbiology and Risk Assessment of Cocoa Ingredients and Chocolate," in S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, p. 377 – 387.
9. P. S. Dimick, J. C. Hoskins, *Can. Inst. Food Sci. Technol. J.* **14** (1981) 269.
10. M. F. Fowler: CAOBISCO Proposal on Improvements to the Cocoa Contract, Manufacturing Confectioner, 1999, 57 – 59.
11. W. Wirth, *Zucker Süsswaren Wirtsch.* **8** (1955) 801.
12. S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999.
13. D. Newton, Micronizing Company Ltd., Framlingham, N. Woodbridge, Suffolk IP 13 9 PT, England.
14. K. Meyer-Potschak: *Kakao Infrarot-Vorbehandlung*, G. W. Barth, Information Kakao, no. 5, p. 4, Ludwigsburg 1984.
15. Bühler AG, Maschinenfabrik, CH 9240 Uzwil, Switzerland.
16. J. Kleinert: "Schokoladeherstellung nach dem LSCP-Verfahren", *Rev. Int. Choc.* **26** (1971) no. 1, 2 – 8.
17. J. Kleinert: *Verfahren zur Veredlung von Kakao-Kernbruch und Kakaomasse*, Chocotechnik, Zentralfachschule der Deutschen Süsswarenwirtschaft (ZDS), Solingen-Gräfrath 1982.
18. E. A. Niediek: "Amorphisierung von Zucker durch das Feinwalzen von Schokolademassen", *Gordian* **1980**, no. 11, 267 – 269.
19. Thouet GmbH, Maschinenfabrik, D-5100 Aachen, Germany.
20. G. Talbot: "Chocolate Temper," in S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, pp. 218 – 230.

21. R. Nelson: "Tempering," in S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, pp. 231 – 258.
22. CH 192 944, 1980 (L. A. Pirsch, G. F. Schubiger, W. Rostagno).
23. K. Jackson: "Recipes," in S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, pp. 323 – 346.
24. S. T. Beckett: Influencing the Rheological Behaviour of Chocolate Mass, *Penn. Manufacturing Confectioners' Ass'n Research Notes XVI*, Nov 1992, pp. 1 – 6.
25. G. Hogenbirk: "Turning up the Heat," *Kennedy's Confection* **5** (1994) no. 1, 19 – 23.
26. J. Kleinert: "Kakaobutterkennzahlen in Korrelation zur Vorkristallisation der Fettschmelze," *Rev. Int. Choc.* **19** (1964) no. 10, 446 – 463.
27. J. Kleinert: "Studien über die Fettreifbildung und deren wirksame Verzögerung," *Rev. Int. Choc.* **16** (1961) 249 – 276.
28. *ICCO Quarterly Bulletin of Cocoa Statistics* **29** (3) 2002–2003.
29. CAOBISCO Statistical Review 2003 published 2004.
30. C. Nuttal, W. A. Hart: "Chocolate Marketing and other Aspects of the Confectionery Industry," in S. T. Beckett (ed.): *Industrial Chocolate Manufacture and Use*, 3rd ed., Blackwells, Oxford 1999, pp. 439 – 459.

Choline

Charles Atwater, Technical Consultant, DuCoa L.P., a DCV company, Wilmington, Delaware 19898, United States

1. **Introduction** 1
2. **Properties** 2
3. **Production** 2
4. **Analysis** 2
5. **Economic Aspects** 3
6. **Salts** 3
7. **Uses** 3
8. **Toxicology** 4
9. **References** 4

1. Introduction

Choline [*62-49-7*], $C_5H_{15}O_2N$, M_r 121.18, trimethyl (2-hydroxyethyl) ammonium hydroxide was first isolated from pig bile by Stecker in 1849. Later, choline was isolated from brain tissue. Choline is a quaternary ammonium compound that is found in most biological materials:

$$[(CH_3)_3N^+CH_2CH_2OH]\ OH^-$$

Free choline makes up only a small proportion of the total choline found in biological materials. More common forms include phosphorylcholine, phosphatidylcholine (→ Lecithin), and acetylcholine. In addition, choline is associated with the metabolism and synthesis of glycine, betaine, cysteine, serine, methionine, and many other methyl-group-containing biological compounds.

Choline has four primary functions [1]:

1) Structural component of cell membranes in the form of phosphatidylcholine (lecithin) and sphingomyelin,
2) Lipotropic agent in the prevention of fatty liver,
3) Neurotransmitter in the form of acetylcholine, and
4) Source of labile methyl groups.

Choline has been referred to as a vitamin, although from a classical standpoint it is not, since it is not known to function as a cofactor in any enzymatic reaction. Like the vitamins, choline plays a significant role in nutrition but its daily requirement is significantly greater than that of vitamins. The recommended daily intake for humans is given in Table 1.

Table 1. Choline, recommended daily intake [2]

Population	Category	Adequate intake mg/d
Infants	0 – 5 months	125
	6 – 11 months	150
Childrens and adolescents	1 – 3 years	200
	4 – 8 years	250
	9 – 13 years	375
	14 – 18 years	550
males		
	14 – 18 years	400
females		
Adults		
	19 years and older	550
males		
	19 years and older	425
females		
Pregnancy	all ages	450
Lactation	all ages	550

The United States Pharmacopeia is drafting USP monographs for choline chloride and choline bitartrate for inclusion in the second or third supplement of USP 24 in the year 2000 [3].

In growing animals, fatty liver and reduced growth rate are the most common choline deficiency symptoms. Numerous animal species, including rats, mice, puppies, young pigs, calves, guinea pigs, chicks, ducklings, turkey poults, trout, and monkeys, have demonstrated a requirement for choline to produce maximum growth rate and prevent fatty liver [4]. Several salts of choline (chloride, dihydrogen citrate, and bitartrate) are often added to animal feeds as a dietary supplement.

While the nutritional needs of humans and animals dominate the choline industry in both on-going research and industry production of dietary supplements, there are other significant uses for choline. A review of the Chemical Abstracts Chemical Substance Index from

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_039

1995 – 1998 revealed six areas of significant patent activity (see Table 2).

Table 2. Industrial patent activity

Industry	Uses	Choline compound
Integrated circuits/semiconductors	cleaning, resist removal, etc.	choline and salts
Printed wire boards	resist removal, developer, etc.	choline and salts
Silica/silicic acid	catalyst in manufacturing	choline salts
Health	antifungal and wound healing	choline salts
Textile	laundering, finishing agent, dyeing, etc.	choline salts and esters
Oil drilling	drilling fluid, soil stabilization	choline salts

2. Properties

Choline is a strong base (pk = 5.06) which absorbs carbon dioxide and water from the air. Choline is usually a colorless, viscous liquid with a fishy odor. The products of commerce are solutions in methanol or water with a concentration of 45 vol % or less. Choline is also soluble in ethanol; hardly soluble in acetone and chloroform; and insoluble in ether, benzene, toluene, and carbon tetrachloride. Choline reacts with acids to form stable salts. Most of the choline salts except the chloride, bitartrate and dihydrogen citrate are difficult to crystallize. Choline also reacts with acids to form esters. Choline, a quaternary ammonium hydroxide, undergoes self-decomposition when it is heated. The reaction, known as the Hoffman Elimination is [5]:

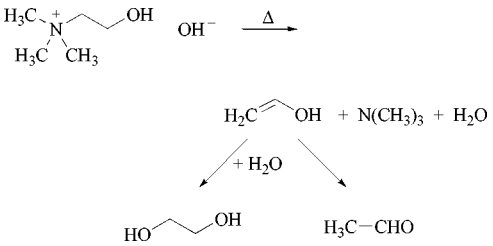

The unstable vinyl alcohol rearranges to acetaldehyde or can react with water to form ethylene glycol. Acetaldehyde can undergo aldol condensation leading to color forming products, while the trimethylamine gives the solution a strong "fishy" odor.

3. Production

Choline is normally produced by the reaction of trimethylamine, ethylene oxide, and water.

Its chloride can be produced by the reaction of trimethylamine with chlorohydrin.

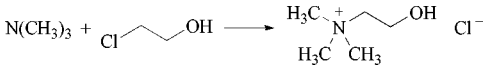

While other syntheses exist, the dominating industrial process is reacting ethylene oxide with trimethylamine and water, or an acid if one of its salts is desired, in a batch or continuous manufacturing process.

4. Analysis

Analytical procedures for choline, until recently, could be characterized as laborious, imprecise and relatively unspecific. Analytical methods for feed include perchloric acid titration [6] and the Reineckate procedure. The Reineckate procedure is used most frequently. Engel first described the Reineckate procedure for choline analysis in 1942. Basically it involves the extraction of choline-containing lipids, formation and separation of a water insoluble Reinecke complex and then spectrophotometric determination of the acetone soluble Reinecke complex at 520 nm. Sensitivity for choline analysis using the Reineckate procedure is as little as 100 ppm in plant and animal tissues.

More recent methods for determining choline rely on enzymatic procedures. Takayama et al. (1977) described a method that was specific for choline in serum. This method offers possible specific determination of choline in animal and plant tissues.

Numerous other enzyme-linked assays, gas, and liquid chromatographic, radiochemical and fluorometric procedures for choline have been described [7].

The perchloric acid method is satisfactory when applied to pure choline and its salts.

5. Economic Aspects

Market requirements for all forms of choline in the animal nutrition, human health, printed wire boards and integrated circuits are closely guarded by industry members. Worldwide uses in animal nutrition are probably $> 30\,000$ t/a while human health, printed wired boards and integrated circuits markets are probably less than 2000 t/a for each industry.

6. Salts

Choline chloride [*67-48-1*], $C_5H_{14}ONCl$, M_r 139.63, white crystalline solid with a slight amine-like odor and a strong brackish taste. It is hygroscopic and decomposes at ca. 180 °C to form dimethylaminoethanol and methyl chloride. Choline chloride is very soluble in water, freely soluble in alcohol, slightly soluble in acetone and chloroform, and almost insoluble in ether and benzene. Its solution is neutral to litmus paper.

Choline hydrogen tartrate [*87-67-2*], $C_9H_{19}O_7N$, M_r 253.26, *mp* 149 – 153 °C, is a white crystalline powder with an acidic taste and a slight amine odor. It is less hygroscopic than other choline salts. It is very soluble in water, slightly soluble in alcohol, and almost insoluble in ether, benzene, and chloroform. The pH of a 25 % aqueous solution is about 3.5.

Choline dihydrogen citrate [*77-91-8*], $C_{11}H_{21}O_8N$, M_r 295.30, *mp* 105 – 107.5 °C is a white granular to fine-crystalline powder with a faint amine-like odor and a acidic taste. It is hygroscopic and very soluble in water, soluble in alcohol and almost insoluble in ether, benzene and chloroform. The pH of a 25 % aqueous solution is about 4.3.

Choline bicarbonate [*78-73-9*], $C_6H_{15}NO_4$, M_r 165.2, its solution is a clear liquid, colorless to slightly yellow with a characteristic amine-like odor. An aqueous solution containing about 75 % choline bicarbonate has a pH between 9.0 and 11.5. Besides being very soluble in water, choline bicarbonate is freely soluble in alcohol and slightly soluble in benzene.

Choline gluconate [*608-59-3*], $C_{11}H_{25}O_8N$, M_r 299.33, is a hygroscopic yellow mass, soluble in water and slightly soluble in ethanol.

7. Uses

Choline deficiency in domestic animals is fairly well characterized and is preventable by the addition of choline salts to the diet. Research in human nutrition has focused on the potential for choline to alleviate various disease conditions. Many of these investigations are in their infancy and should provide further information concerning the function of choline in biological systems. Choline salts, such as the chloride, hydrogen tartrate, dihydrogen citrate and gluconate are used therapeutically for patients with liver cirrhosis that has not advanced to the irreversible fibrous stage. Oral choline does appear to be effective in reversing fatty infiltration and arresting the cirrhosis process. Common therapy includes the use of a low fat, high protein, high calorie diet with the usual vitamins and choline [8]. Orally, choline salts are used to provide the recommended dietary intake of choline for humans. Acetylcholine [*51-84-3*] is used as a vasodilator. Methacholine chloride [*62-51-1*], (*O*-acetyl-β-methylcholine chloride) is used as a parasympathetic stimulant and as an antiepinephrine substance. Carbamoylcholine chloride [*51-83-2*] is used to increase peripheral circulation in cases of vasospasms caused by peripheral vascular disorders.

The food and cosmetic industry has had an increasing interest in choline, brought about by the demand for harmless surface-active agents. Some acylcholine compounds, such as stearoylcholine chloride, have been suggested as ingredients in hair conditioners [9]. Chlorocholine chloride [*999-81-5*], (2-chloroethyltrimethylammonium chloride) is used as a plant growth regulator to maintain the compactness of the plant [10] as opposed to the growth-accelerating effects of gibberellic acid. It is used for ornamental plants, especially poinsettias and azaleas. It is also used to prevent the lodging and flattening of wheat by wind, as well as the etiolation of greenhouse tomatoes, thus leading to earlier and larger fruit yields [11].

8. Toxicology

Choline has a low toxicity. An acute toxicity test of choline chloride administered orally to rats shows an LD_{50} value of 3 – 6 g/kg [12]. Because of its high basicity choline is classified as corrosive.

9. References

1. DuCoa L.P.: Choline, Functions and Requirements, Technical Review, DuCoa L.P., Wilmington, DEL 1994.
2. Standing Committee on the Scientific Evaluation of Dietary Reference Intakes: *Dietary Reference Intakes, Choline,* Chap. 12, Institute of Medicine, National Academy Press, Washington D.C. 1998, p. 20.
3. Pharmacopeial Forum, vol. 26, no. 4, July-August, 2000, pp. 1002 – 1006.
4. DuCoa L.P.: Choline, Functions and Requirements, Technical Review, DuCoa L.P., Wilmington, DEL 1994, p. 11.
5. R. T. Morrison, R. N. Boyd: *Organic Chemistry,* 3rd ed., Allyn and Bacon, Boston 1973, p. 753.
6. DuCoa L.P.: Choline, Functions and Requirements, Technical Review, DuCoa L.P., Wilmington, DEL 1994, p. 36.
7. J. Hanin: "Measurement of lecithin and choline" in *Nutrition and the Brain. Choline and Lecithin in Brain Disorders.* **vol. 5,** Raven Press, New York 1979, pp. 113 – 127.
8. DuCoa L.P.: Choline, Functions and Requirements, Technical Review, DuCoa L.P., Wilmington, DEL 1994, p. 23.
9. Lion, JP-Kokai 109 709, 1982. *Chem. Abstr.* **97** (82) 168722p.
10. N. E. Tolbert, *J. Biol. Chem.* **235** (1960) 475.
11. N. P. Budykina et al., *Kim. Sel'sk Khoz.* (1984) 38 – 40; *Chem. Abstr.* **100** (84) 169852y.
12. M. W. Neuman, H. C. Hodge, *Proc. Soc. Exp. Biol. Med.* **58** (1945) 87 – 88.

Chromium and Chromium Alloys

James H. Downing, Elkem Metals Co., Niagara Falls, New York 14302, United States (Chap. 1)

Paul D. Deeley, Shieldalloy Corp., Newfield, New Jersey 08344, United States (Chap. 2)

Rudolf Fichte, Gesellschaft für Elektrometallurgie, Nürnberg, Federal Republic of Germany (Chap. 3)

1. Chromium 1
1.1. History 1
1.2. Chemical and Physical Properties . 2
1.3. Resources and Raw Materials 2
1.4. Production 2
1.5. Uses 5
1.6. Economic Aspects 6
2. Chromium Alloys 6
3. Ferrochromium and Other Chromium Master Alloys 8
3.1. Physical and Chemical Properties . 9
3.2. Raw Materials 9
3.3. Production 10
3.3.1. High-Carbon Ferrochromium 11
3.3.2. Medium-Carbon Ferrochromium . . . 14
3.3.3. Low-Carbon Ferrochromium 15
3.3.4. Nitrogen-Containing Low-Carbon Ferrochromium 18
3.3.5. Other Chromium Master Alloys 18
3.4. Quality Specifications, Storage and Transportation, and Trade Names . 19
3.5. Uses 20
3.6. Economic Aspects 21
3.7. Environmental Protection; Toxicology and Occupational Health 22
4. References 23

1. Chromium

1.1. History [1–5]

Chromium [*7440-47-3*], Cr, A_r 51.9961, was discovered by Vauquelin, in the mineral crocoite [*14654-05-8*], $PbCrO_4$, in 1797. In 1798 he isolated chromium metal by reducing the oxide with carbon. Soon after the discovery of chromium, the commercial process for manufacturing chromates by roasting chromite with soda ash was developed.

During the 19th century, ferrochromium and chromium were produced by a variety of techniques. However, a commercial process was not developed until 1893, when Moissan produced ferrochromium in an electric furnace by the reaction of chromium oxide (Cr_2O_3) and carbon. In 1898, Goldschmidt produced chromium by the aluminothermic reduction of Cr_2O_3. Other advances have included the application of silicothermics to the production of low-carbon ferrochromium and chromium, production of chromium by aqueous electrolysis, and production of low-carbon ferrochromium and refining of electrolytic chromium by high-temperature vacuum processing.

Although Cr is found in many minerals, chromite is the only commercial source of Cr. Most of the mineral came from the Ural Mountains up to 1827, when deposits were discovered in the United States. These supplied a limited market up to about 1860, when large deposits were found in Turkey. Since that time chromite has been mined primarily in the eastern hemisphere.

Chromium was electrolyzed from a solution of chromium chloride by Bunsen in 1854. However, large-scale commercial production of electrolytic Cr did not begin until 1954.

1.2. Chemical and Physical Properties [6], [7], [8]

At room temperature Cr is resistant to ordinary corrosive agents, which explains its use as an electroplated, protective coating. It dissolves in nonoxidizing mineral acids, such as hydrochloric and sulfuric acids, but not in cold aqua regia or nitric acid, which passivate the metal. At elevated temperature it reacts with halogens, silicon, boron, nitrogen, oxygen, and carbon.

Chromium and chromium-rich alloys are brittle at room temperature and this has limited their application [9].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_043

Selected physical properties of chromium are as follows:

A_r	51.9961
Atomic number	24
mp	1857 °C
bp	2672 °C
Density ϱ at 20 °C	7.19 g/cm^3
Crystal structure	cubic, body centered
Specific heat at 25 °C	23.25 J mol^{-1} K^{-1}
Molar entropy S°_{298}	23.64 J mol^{-1} K^{-1}
Heat of fusion	16.93 kJ/mol
Latent heat of vaporization at *bp*	344.3 kJ/mol
Linear coefficient of thermal expansion at 20 °C	6.2×10^{-6}
Resistivity at 20 °C	12.9×10^{-8} Ω m
Thermal conductivity at 20 °C	67 W m^{-1} K^{-1}

1.3. Resources and Raw Materials [4]

Chromite [*53293-42-8*] is a spinel $FeO \cdot Cr_2O_3$. In nature it is a mixture described by the formula $(Fe^{2+}, Mg)O \cdot (Cr, Al, Fe^{3+})_2O_3$. Chromite ore rarely contains more than 50 % Cr_2O_3; other minerals such as SiO_2 can also be present.

Chromite is found in stratiform and podiform deposits. Stratiform deposits occur in layers up to a meter thick. The Bushweld Igneous Complex (Transvaal, Republic of South Africa), the Great Dyke (Zimbabwe) and the Stillwater Complex (Montana, United States) are examples. Podiform deposits range in size but a typical commercial deposit will be over 100 000 t. Deposits occur in the Ural Mountains, Albania, Zimbabwe, and the Philippines.

The 1983 production and reserves of chromite are shown in Table 1 (see right). Generally, richer lumpy Cr bearing ores have been preferred for smelting, whereas those with Cr_2O_3 content less than 40 % have been used in refractories. For a detailed discussion of chromium ores, see → Chromium Compounds.

1.4. Production

Extraction. Chromium is extracted from its ores by alkaline or acidic dissolution. In alkali dissolution, finely ground chrome ore is roasted with Na_2CO_3 under oxidizing conditions at ca. 1100 °C. The sodium chromate is leached from the calcine; most of the gangue is insoluble. The solution containing hexavalent Cr can be reduced with SO_2 and used for electrowinning, or $Na_2Cr_2O_7$ can be crystallized from it. The $Na_2Cr_2O_7$ can be converted to CrO_3 for use in electrolysis or to Cr_2O_3 for use in metallothermics.

Chrome ore can be dissolved in acid if an oxidizing agent is present. However, Fe, Al, and Mg also dissolve and must be removed. The preferred acidic dissolution technique is to reduce the ore with carbon, forming ferrochromium, which is ground and dissolved in sulfuric acid. The only significant impurity carried over is Fe, which is removed by crystallization as iron(II) ammonium sulfate. The Cr in the solution is in the +3 valence state and with additional purification is used to produce electrolytic Cr.

Table 1. Chromite production and reserves, 1983 [4]

	Production [a], 10^3 t	Reserves [b], 10^6 t	Reserve base [b, c], 10^6 t
North America			
Canada	0	–	4
South America			
Brazil	82	8	9
Cuba	8	3	3
Europe			
Albania	277	6	20
Finland	56	17	29
Greece	15	1	1
Former States of USSR	855	129	129
Africa			
Madagascar	13	7	7
Republic of South Africa	688	828	5715
Sudan	8	2	2
Zimbabwe	147	17	75
Asia			
India	112	14	60
Iran	16	1	1
Pakistan	–	1	1
Philippines	75	14	29
Turkey	103	5	73
Vietnam	5	1	1
Oceania			
New Caledonia	30	2	2
Australia	–	–	2

[a] Estimated.
[b] Shipping grade ore is deposit quantity and grade normalized to 45 % Cr_2O_3 for high-Cr and high-Fe chromite, and 35 % Cr_2O_3 for high-alumina chromite.
[c] Reserve base includes deposits currently economic (reserves), marginally economic, and some currently subeconomic.

Production of Chromium Metal. Chromium metal is produced by the reduction of Cr_2O_3 or the electrolysis of Cr(III) solutions. The metal can also be obtained from Cr(VI) solutions by electrolysis, but the process is less efficient and is used primarily for plating.

Aluminothermic Reduction. Aluminum is the most important reducing agent for producing Cr from Cr_2O_3. Closely sized high-purity Al powder and Cr_2O_3 are blended and charged into a vessel lined with refractory, which is usually Al_2O_3. The charge is ignited either with a $KClO_4$– Al powder "wick" or electrically. The

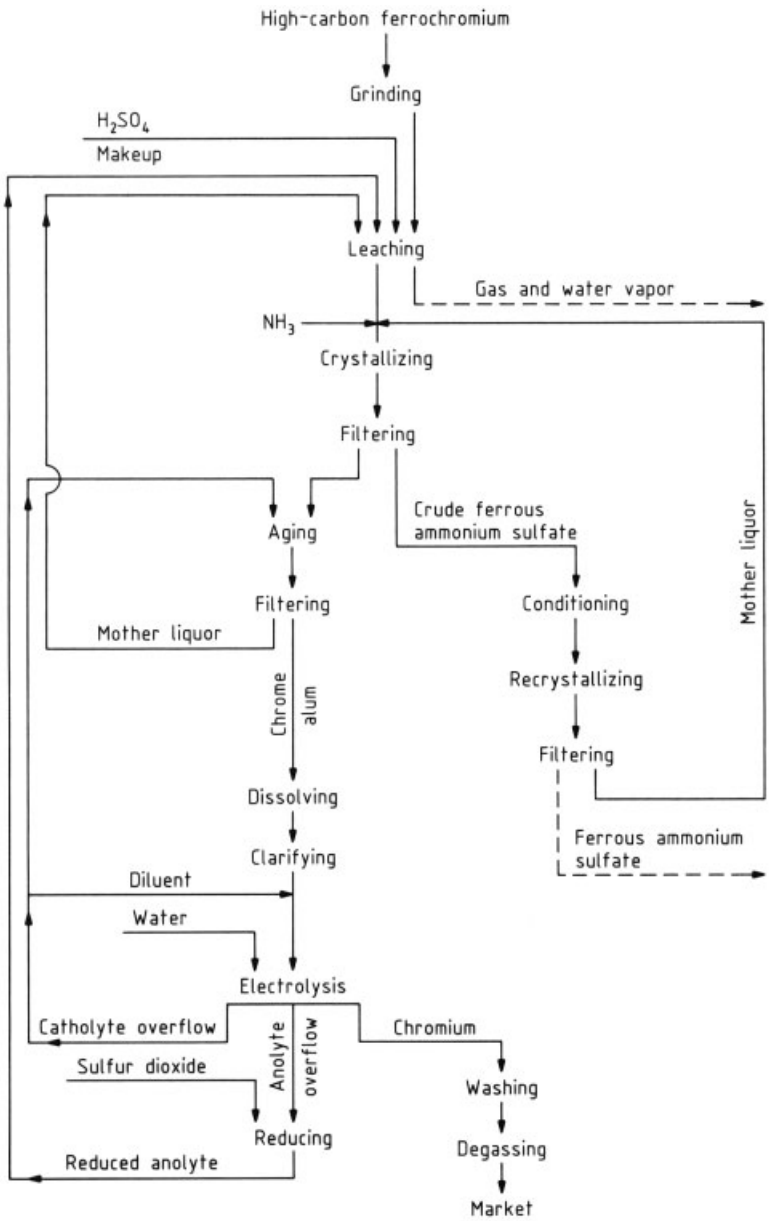

Figure 1. Flow sheet for producing electrolytic chromium from ferrochromium [10]

exothermic reaction results in a temperature greater than 2000 °C, which leads to a clean separation of slag from metal. The purity of the Cr metal depends on the purity of the reactants, particularly the Cr_2O_3. Originally this oxide was made by reduction of sodium chromate with sulfur, resulting in Cr_2O_3 of high sulfur content. Proprietary processes have been developed to produce Cr_2O_3 of higher purity. Analysis of aluminothermically reduced Cr is given in Table 2.

Chromium is also produced by *carbon reduction.* Chromium oxide and carbon are carefully weighed, mixed, briquetted, and heated in a furnace at 1400 °C at a minimum pressure of 40 Pa. The heating cycle is 100 h. The C + O content is $\leqq$ 1.5 % (see Table 2).

Solutions suitable for *electrolytic production* of Cr can be derived from ore by oxidative roasting in alkali, or by direct solution of chromite in sulfuric acid; however, a commercial process was not achieved until the electrolyte was made by dissolving ferrochromium in H_2SO_4 and reduced anolyte. Figure 1 (see previous page) [10] shows the essential aspects of the process for electrolytic production of chromium metal.

Milled FeCr is leached with hot reduced recycled anolyte, chrome alum mother liquor, and makeup H_2SO_4. The iron precipitates on cooling as iron(II) ammonium sulfate $[Fe(NH_4)_2(SO_4)_2]$, which is recrystallized to recover the coprecipitated Cr.

The Cr is further purified by precipitation of chrome alum. Transformation of the Cr(III) to an insoluble salt requires aging at 30 °C. After filtration the mother liquor is recirculated to the leach tank. The chrome alum crystals are redissolved to make the cathode feed. The overflow anolyte contains Cr(VI), which is reduced by SO_2, thereby generating additional H_2SO_4. The reduced anolyte is also cycled to the leach tank. The stripped catholyte is recycled. Dissolved chrome alum crystals are used to bring the Cr concentration in the catholyte feed up to the desired level.

The details of the *cell reactions* are shown in Figure 2 (see next page) [11]. A diaphragm is necessary to prevent migration of Cr(VI) into the cathode compartment where its reduction by Cr(II) would lead to loss of current efficiency. Flow is maintained into the anode compartment from the cathode compartment by a higher level of solution in the latter. The pH of the catholyte must be controlled. At too low a value, H_2 evolution increases; at too high a value, precipitation of $Cr(OH)_3$ occurs. This can be seen by an examination of the pH – potential diagram [12]. The pH is controlled by Cr concentration, current density, temperature, and differential height of solution between cathode and anode compartments. Water is fed to the anode to control the concentration of H_2SO_4.

Chromium is plated onto stainless steel cathodes until it attains a thickness of ca. 3 mm. The plate is stripped from the cathode and degassed by heating at 420 °C. The analysis of the flake is shown in Table 2.

For many applications the oxygen content of the electrolytic chromium is too high. Deoxidation is carried out on a commercial scale by two techniques [13]. In the first either flake or briquettes of powdered flake are contacted with H_2 at elevated temperature. The procedure is less effective on Cr produced from chrome alum solutions than on Cr from hexavalent solutions [14]. The second technique involves heating briquettes of ground electrolytic flake and carefully controlled trace amounts of C in a vacuum furnace to form CO. The heating cycle is 90 h, the maximum temperature is 1400 °C, and the pressure is 13 Pa. The product is cooled in helium to

Table 2. Analysis of various grades of chromium metal (in wt %)

	Cr	C	O	Si	S	P	N	Al
Aluminothermic [a]	99.4	0.05	0.10	0.10	0.010	0.010	0.02	0.10
Electrolytic [b]	99.1	0.02	0.5	0.01	0.03	0.01	0.05	0.01
Electrolytic [b, c]	99.5	0.05	0.02	0.05	0.04	0.01	0.002	0.015
Carbon reduced [b]	98	1.5 [d]		–	0.1	0.02	0.001	0.1

[a] Shieldalloy Corp.
[b] Elkem Metals Co.
[c] Vacuum treated.
[d] C + O.

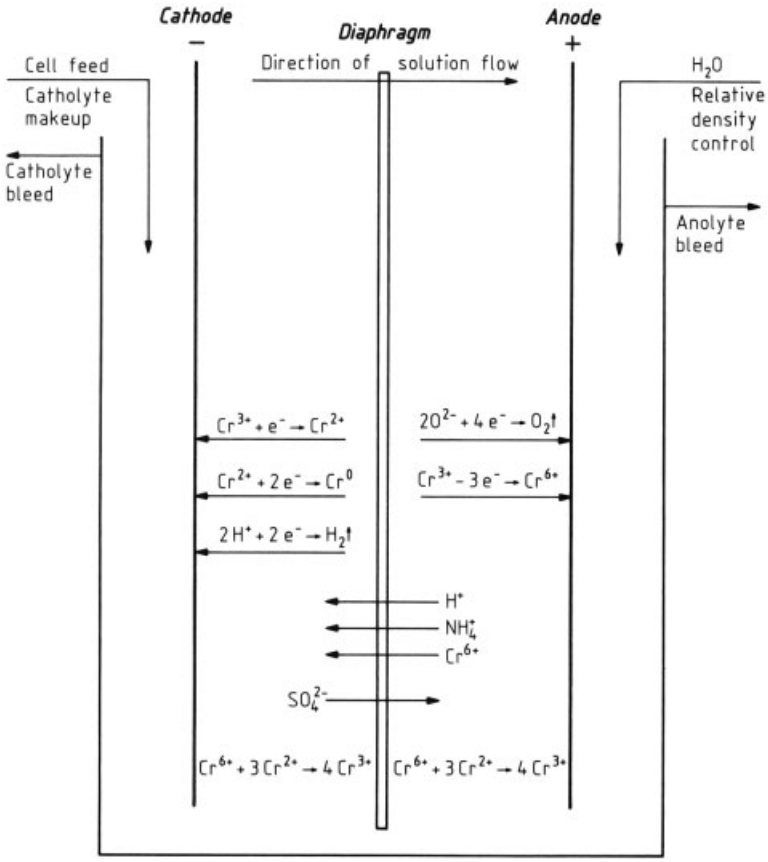

Figure 2. Idealized electrolytic cell reactions

prevent contamination. Analysis of the vacuum-treated product is shown in Table 2.

There are several other techniques for purifying Cr. These include iodide refining [9], [14], zone melting [9], [14], and treating Cr with a flux containing an alkaline earth metal [15]. Chromium prepared by these methods is purer but more expensive, and therefore is used only in electronic applications.

1.5. Uses [4]

Chromium is used in ferrous and nonferrous alloys, in refractories, and in chemicals. Chromium enhances an alloy's hardenability, creep and impact strength, and resistance to corrosion, oxidation, and wear. Ferrous alloys, mainly stainless steels, account for most of the consumption. These steels have a wide range of mechanical properties as well as being corrosion and oxidation resistant. Cast irons may contain from 0.5 % to 30 % Cr, which provides hardenability, toughness, hardness, and corrosion and wear resistance. Chromium is widely used in nonferrous alloys, including those based on nickel, iron – nickel, cobalt, aluminum, titanium, and copper. In Ni, FeNi, and Co, Cr is used for oxidation and corrosion resistance. In Al, Ti, and Cu it controls the microstructure.

Chromium chemicals are used in a variety of applications. The largest amount is consumed to manufacture pigments for use in paints and inks. Other applications include leather tanning, metal corrosion inhibition, drilling muds, textile dyes, catalysts, and wood and water treatment.

Chromite is used in the refractory industry to make bricks, mortar, and ramming and gunning mixes. Chromite enhances thermal shock and slag resistance, volume stability, and strength.

1.6. Economic Aspects

The world production of chromium fluctuated between 1973 and 1983; however, growth is expected to the end of the century. Data for worldwide production are as follows [4]:

Year	Production, 10^3 t	Year 10^3 t	Production,
1973	1999	1979	2936
1974	2207	1980	2973
1975	2540	1981	2786
1976	2669	1983	2490
1977	2883	1982	2489
1978	2820	2000	5625 *

* Probable projected demand.

2. Chromium Alloys

In the manufacture of steel, chromium is added usually in the form of ferrochrome (see Chap. 3). Pure chromium metal, produced by electrolytic or aluminothermic processes, is used for alloying nonferrous engineering materials. The most common materials are nickel-based and cobalt-based alloys, most of which are used at high temperature.

The only true chromium-based alloys that have been developed and used commercially are a few chromium – nickel alloys developed by the International Nickel Co. (INCO). These materials contain 50 – 60 % chromium; the remaining percentage is nickel, with either niobium (columbium) or titanium specified as carbon and nitrogen scavengers. These alloys, as well as materials cited in [16], [17], often referred to as *superalloys*, are used in constructing high-temperature exposed components in chemical and petrochemical industries.

Chromium-based alloys have found a unique application in power stations as supports for heat exchanger pipes. The balance of cobalt- and nickel-based high-chromium alloys are used mainly in components for gas turbine engines and for parts requiring resistance to elevated temperature, oxidation, and hot corrosion.

Chromium alloys of Ni, Fe – Ni, and Co are produced in a ratio of about 10 : 5 : 1; total annual production is ca. 7×10^4 t. By far the alloys produced in greatest quantity are the solution-hardened types, such as alloy 600 (75 % Ni – 15 % Cr – 10 % Fe) and alloy 800 (35 % Ni – 20 % Cr – balance Fe). Special purpose Ni – Cr – Mo alloys, which have generated a significant volume, include Hastelloy X and alloy 625. These alloys are used in sheet, plate, bar, wire, and tube forms for a wide variety of applications in industries, such as petrochemical, chemical processing, heat treatment, pollution control, nuclear power, and transportation.

The most widely used *wrought* nickel-based age-hardened superalloys are Waspaloy and alloys 700 and 718. Popular iron – nickel superalloys include 901 and A – 286. The only wrought cobalt-based alloy used in any volume is Haynes 188. *Investment cast superalloys* include 713 and 718 and the cobalt-based alloy X – 40. Wrought and cast superalloys are used primarily in static and rotating parts for gas turbine engines, although they are also used for other applications, including petroleum production. Because of the demanding service required of these very expensive alloys, many special purpose proprietary compositions have been developed.

Advances in *nickel-based alloys* include prealloyed powders, coating technology, rapid solidification, single-crystal technology, oxide dispersion, strengthening, and metal matrix composites. Although all of these developments are for products with relatively low commercial volume, they represent the leading edge of materials technology and have spurred a new series of unique alloys.

Producer and user specifications may allow for substitution of a special quality ferrochromium for chromium metal [18]. Chromium metal is also required in a few examples of iron-based engineering alloys in which the use of ferrochromium would introduce undesirable levels of certain elements deleterious to the physical properties of the end product, such as aircraft landing gears [19].

Chromium metal in powder form is used in the manufacture of *cermets* (→ Ceramics, Ceramic – Metal Systems), [20]. The cermets of commercial importance are as follows:

Cermet LT – 1	77 % Cr, 23 % Al_2O_3
Cermet LT – 1 B	59 % Cr, 19 % Al_2O_3, 20 % Mo, 2 % TiO_2
Cermet LT – 2	25 % Cr, 15 % Al_2O_3, 60 % W

These particular cermets have very good thermal stability and corrosion resistance. Both LT–1 and LT–1 B exhibit good creep resistance up to 1200 °C. Cermets are used for thermocouple tubes, crucibles for ovens, shuttles in the textile industry, gas tubes and rings in gas turbines, components of sulfuric acid pumps (LT–1), and for inserts in rocket nozzles (LT–2). Other applications are crucibles, pans, casting molds, casting forms, and nozzles.

Small alloying additions of chromium to *copper* between 0.3 % and 1.2 % are used to provide higher strength and, in some applications, abrasion resistance, while reducing conductivity only slightly. Copper – chromium alloys are used for electrode tips for spot welding in the automotive industry, rotating contact parts, rings, and other electrical applications [20]. Rapid solidification processing and solid-state microblending has resulted in the development of a more homogeneous microstructure in copper – chromium alloys [21].

Wrought aluminum alloys contain small amounts of chromium (0.10 – 0.3 %) in addition to other alloying elements, such as magnesium and manganese [22].

Cast aluminum alloys contain 0.05 – 0.40 % chromium [23]. In both the cast and wrought aluminum alloys, the chromium imparts increased corrosion resistance and improved physical properties.

Chromium metal from electrolytic or aluminothermic processes is used in a briquetted aluminum powder compact to provide the alloying addition to both cast and wrought aluminum products [24]. Two grades of smelted binary aluminothermic chromium – aluminum alloys (10 % chromium and 20 % chromium) are also used.

Other binary aluminothermic chromium alloys available for special alloying requirements are chromium – molybdenum (30 % Mo) and niobium – chromium (30 % Nb). Chromium – tungsten binary alloys were once used but are now redundant [25], [26].

Metallic chromium powders are used in both coated and cored welding electrodes. Small quantities of pure chromium are used for anodes in X-ray tubes and also for vacuum vaporization or sputtering.

Chromium powder, as well as ferrochromium powder, has been used in considerable quantities to produce *chromium coatings.*

The so-called pack chrome coating [26], [27] is applied to cast and wrought steel parts by immersing the article in a mixture of chromium powder; an inert material, e.g., kaolin, alumina, or magnesia; and mixtures of various salts, such as ammonium, iodide, or chloride. The packed part is then heated in an inert atmosphere at 900 – 1300 °C (heating time is dependent on part size).

Another procedure is the so-called gas chromium coating [27]. Steel parts and the chrome coating mixture are placed in a vessel at 1050 – 1250 °C. Volatile chrome compounds react with the steel components. The coating mixture consists mostly of chromium powder (both chromium metal and ferrochromium have been used [28]); inert materials (similar to those used in the pack chromium coating); and a relatively high concentration of a salt (fluoride, ammonia, or cryolite).

The gas chromium coating has been used to produce a so-called tin-free steel and to treat sheet steel on a continuous line.

Although chrome plating is used to produce a shiny as well as a hard satin finish, the source is from chromium salts rather than from aluminothermic or electrolytic chromium.

Chromium as the main alloying component has been reported in the literature [29], [30]. To date, none of these alloys is industrially significant. The cited reasons for lack of commercial application are the lack of ductility and the difficulty of processing these chromium-based alloys. Some of the alloys investigated that appear promising and the properties studied are listed in Table 3 (see next page) [29], [30].

3. Ferrochromium and Other Chromium Master Alloys

Ferrochromium is a master alloy of iron and chromium, containing 45 – 95 % Cr and various amounts of iron, carbon, and other elements. The ferrochromium alloys are classified by their carbon content and are known by their French names because basic work in this field was carried out mainly in France:

Table 3. Composition and properties of some chromium alloys under development [29], [30]

	Alloying elements, wt % (remainder is Cr)										Tensile strength, 100 h at 1150 °C, N/mm^2	Ductile – brittle transition temperature, °C	
	W	Mo	Ta	V	Ti	Zr	Si	MgO	C	Other		Tension	Impact
United States													
C-207	7.5	–	–	–	0.2	0.8	–	–	0.1	0.15 (Y)	83	120	370
C1-41	–	7.1	2	–	–	–	–	–	0.09	0.1 (Y + La)	110	230	–
IM-15	–	–	1.7	–	–	–	–	–	–	0.1 (Y), 0.1 (B)	137	380	330
Chrome 30	–	–	–	–	0.5	–	–	6	–	–	41∗	−12	160
Chrome 90	–	–	–	2.5	–	–	0.5	3	–	–	55∗	−18	290
Chrome 90 S	–	–	2	2.5	0.5	–	1	3	0.5	–	17	180	590
Australia													
Alloy E	–	–	2	–	0.5	–	0.5	–	–	–	52	25	315
Alloy H	–	–	2	–	–	–	0.5	–	–	0.5 (Re)∗∗	–	25	–
Alloy J	–	–	2	–	–	–	0.5	–	–	–	–	–	–

∗ 980 °C.
∗∗ Re = rare earths.

1) *High-carbon ferrochromium* (HC ferrochromium) with 4 – 10 % C, "ferrochrome carbure"
2) *Medium-carbon ferrochromium* (MC ferrochromium) with 0.5 – 4 % C, "ferrochrome affiné"
3) *Low-carbon ferrochromium* (LC ferrochromium) with 0.01 – 0.5 % C, "ferrochrome suraffiné"

The mechanical and chemical properties of steel can be improved by alloying it with ferrochromium. Chromium combined with nickel gives stainless steel excellent chemical resistance (see Chap. 1).

The first industrial use of ferrochromium in producing low-chromium alloy steels was in 1860 – 1870 in France. Previously P. Berthier (1821) and E. Fremy (1857) produced small quantities of high-carbon ferrochromium in crucibles by reducing chromite or combinations of chromium and iron oxides with charcoal. This direct reduction process was transferred to the blast furnace or cupola to produce low-chromium alloys with 7 – 8 % Cr, and later on alloys with 30 – 40 % Cr. In the United States (1869), Sweden (1886), and Russia (1875), similar processes were developed for producing high-carbon ferrochromium [26, pp. 1 – 3], [31].

The fundamental work by Moissan (1893) on using the electric arc furnace and its industrial application for the direct carbothermic production of high-carbon ferrochromium by Héroult (1899) were major improvements over the blast furnace that led to the modern large-scale production.

Table 4. Some physical properties of ferrochromium and ferrosilicochromium [32]

Alloy	Density (ϱ), g/cm^3	Melting range, °C	
		Liquidus	Solidus
Chromium metal (electrolytic)	7.2	1900	
Chromium metal (aluminothermic)	7.2	1850	
Low-carbon ferrochromium (72 % Cr, 0.01 % C)	7.35	1690	1660
Low-carbon ferrochromium (72 % Cr, 0.05 % C)	7.35	1670	1639
Low-carbon ferrochromium (69 % Cr, 0.1 % C)	7.35	1604	1343
High-carbon ferrochromium (69 % Cr, 4 – 6 % C, 1 % Si)	7.2	1500	1350
High-carbon ferrochromium (64 % Cr, 5 % C, 1 % Si)	7.1	1450	1340
Ferrochromium, charge grade (63 % Cr, 5.5 % C, 7 % Si) ∗	6.7	1500	1400
Ferrochromium, charge grade (56 % Cr, 6 % C, 5 % Si) ∗	6.8	1493	1660
Ferrochromium silicon (36 % Cr, 40 % Si)	5.3	1388	1360

∗ Charge chrome (see Table 10, and Section 3.4,)

The reduction of chromite by silicon for the production of low-carbon ferrochromium was developed by F. M. Beckett (1907) and improved by G. Jean (1909).

The aluminothermic production of low-carbon ferrochromium has proved to be too expensive and the method is seldom used now.

3.1. Physical and Chemical Properties

Low-carbon ferrochromium has a bright silvery appearance; as carbon content increases, the metal turns from gray to dark gray. The density and melting range for different grades of ferrochromium are summarized in Table 4 (see previous page) [32].

The Fe – Cr phase diagram (see Fig. 3, [33]) exhibits a continuous series of solid solutions at higher temperature and a σ-phase at lower temperature. This brittle σ-phase has a tetragonal structure ($a = 8.7995 \times 10^{-10}$ m, $c = 4.5442 \times 10^{-10}$ m).

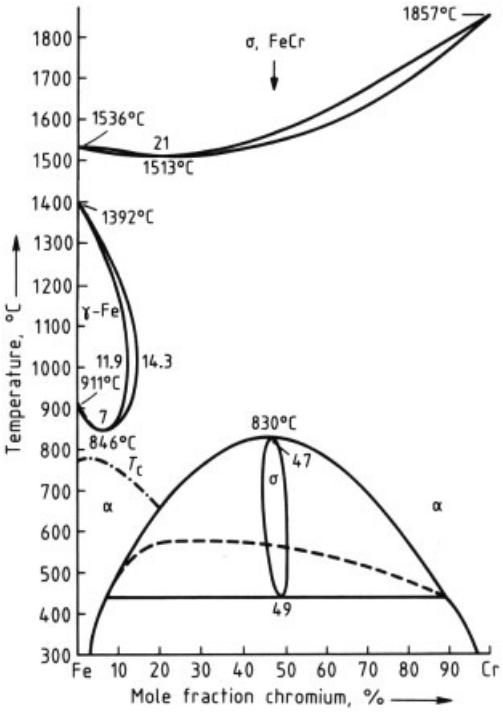

Figure 3. Fe – Cr phase diagram [33]

Commercial ferrochromium contains mainly carbon as a constituent element because of the high affinity of chromium for the carbon used in the reduction process. The constitution and structure of the C – Cr – Fe system has been reviewed [34]. A projection of the liquidus temperatures and solid-phase equilibrium by isothermal sections from 1150 to 600 °C have been published [34].

3.2. Raw Materials

The only raw materials for the production of ferrochromium are chromite ores. The mineral chromite has a spinel structure and its formula may be written as $(Fe^{2+}, Mg)O \cdot (Cr, Al, Fe^{3+})_2O_3$ (→ Chromium Compounds). A high Cr : Fe ratio is advantageous to produce an alloy with high chromium content. Chromite ores are classified as follows [35]:

1) Ores rich in chromium: > 46 % Cr_2O_3, Cr : Fe > 2 : 1; for the production of ferrochromium
2) Ores rich in iron: 40 – 46 % Cr_2O_3, Cr : Fe < 2 : 1; for the production of charge chrome and for the chemical industry
3) Ores rich in aluminum: > 60 % ($Cr_2O_3 + Al_2O_3$), > 20 % Al_2O_3; for refractories

Metallurgical-grade chromite ores are classified as hard lumpy or friable lump types, fines, and concentrates. Concentrates are produced by mechanical upgrading of lean ores or fines [36].

In the production of high-carbon ferrochromium, which is by far the alloy in greatest demand, generally a lumpy type of chromite ore is necessary. The submerged arc smelting of high-carbon ferrochromium by direct reduction of carbon in large low-shaft electric furnaces generally requires lumpy chromite ores to allow the reaction gases to pass from the lower reaction zone to the top of the furnace, where the burden (i.e., charge) is continuously charged.

About 80 % of chromite ores in the western world are fines (< 10 mm). Therefore, efforts have been made to agglomerate these fines, by either sintering, briquetting, or pelletizing [37]. Fines of chromite ores can be used to produce low-carbon ferrochromium.

Typical analyses of some important metallurgical chromite ores are summarized in Table 5 (see next page) [38].

Table 5. Analyses of some chromite ores, metallurgical grade

Compound	Chromite ore, wt %				
	Transvaal lump [a]	Zimbabwe friable [b]	Former USSR lump [c]	Turkey lump [d]	Albania lump
Cr_2O_3	42.55	49.53	53.73	47.58	40.5
FeO	21.85	11.6	8.5	9.45	11.4
Fe_2O_3	4.85	2.2	4.3	3.1	–
MgO	9.26	17.52	17.3	18.7	23.3
Al_2O_3	15.5	7.3	9.4	8.8	7.3
SiO_2	5.54	5.14	4.3	8.33	12.6
L.O.I. [e]	0.2	2.3	2.1	2.25	3.4
Cr : Fe	1.43	3.20	3.81	3.42	3.13

[a] Bushveldmassiv [38].
[b] Great Dyke (friable lump) [38].
[c] Kempirsajski [38].
[d] Anatolia (Fethiye) [38].
[e] L.O.I. = loss on ignition.

The reducing agent for chromite is usually carbon in the form of coke (gas coke, coal, or charcoal); its contents of S and P should be low. Silicon as a reducing agent is used in the form of ferrosilicochromium or ferrosilicon to produce low-carbon ferrochromium. Fluxing agents, e.g., quartzite or alumina (corundum or bauxite) and lime, are charged with the burden for slag formation. In the carbothermic production of ferrosilicochromium, chromite and quartzite are used as the raw materials.

3.3. Production

The oxides of iron and chromium present in the chromite can be readily reduced at high temperature with carbon. Because of the tendency of chromium to form carbides, a carbon-containing alloy is always obtained. The oxides can also be reduced with silicon, aluminum, or magnesium. However, only carbothermic and silicothermic reductions are used commercially. The reducibility of chromite depends on its composition. A chromite rich in iron ($FeO \cdot Cr_2O_3$) can be reduced by carbon at lower temperature than one rich in magnesium ($MgO \cdot Cr_2O_3$) [39], [40]. Iron oxide is reduced by carbon at a lower temperature than chromium oxide.

A thermodynamic analysis of carbothermic reactions in the field of ferroalloys has been performed [41]. For the carbothermic reduction of FeO, Cr_2O_3, and SiO_2, equilibrium temperatures at different CO pressures were calculated. The values in Table 6 were found at 101.3 kPa. Carbides with higher carbon content formed initially at lower temperature react at higher temperature with Cr_2O_3 and form carbides with lower carbon content; finally, reduction of SiO_2 starts at higher temperature. Therefore, production of ferrosilicochromium alloys requires high temperature.

Table 6. Equilibrium temperatures for various carbothermic reduction reactions [41]

Condensed reactants	Condensed products	Gases	Equilibrium temperature, °C
FeO, C	Fe	CO_2, CO	670
Cr_2O_3, C	Cr_3C_2	CO	1150
Cr_2O_3, Cr_3C_2	Cr_7C_3	CO	1190
Cr_2O_3, Cr_7C_3	$Cr_{23}C_6$	CO	1530
Cr_2O_3, $Cr_{23}C_6$	Cr	CO	1810
SiO_2, C	SiC	CO	1480
SiO_2, SiC	Si	CO, Si	1710

In practice the reactions are somewhat more complicated because iron-containing chromium carbides are formed. In high-carbon ferrochromium, the double carbide $(Cr, Fe)_7C_3$ is present. In this compound, two to four Cr atoms can be substituted by iron atoms. Equilibrium temperatures have been calculated for the FeO reduction in chromite [40]:

$$FeO \cdot Cr_2O_3 + C \rightleftharpoons Fe\,(\alpha, \beta, \gamma) + Cr_2O_3 + CO \qquad t = 950\,°C$$

For the Cr_2O_3 reduction:

$$MgO \cdot Cr_2O_3 + 13/3\,C \rightleftharpoons 2/3\,Cr_3C_2 + MgO + 3\,CO \qquad t = 1200\,°C$$

Because the difference in temperature between these two reactions is slight and because iron also facilitates reduction of chromium

oxide, selective reduction of iron is difficult. In solid-state reduction studies on Transvaal chromite, the amount of iron and chromium found in the metallic state at 1000 °C was 11.0 % of the iron and 1.3 % of the chromium, and at 1200 °C was 98.3 % of the iron and 38.5 % of the chromium [38].

In the carbothermic reduction process, unreduced oxides from the chromite (MgO, Al_2O_3) and from the gangue (SiO_2 in serpentine and olivine) are collected in a slag, which generally contains 30 % SiO_2, 30 % MgO, and 30 % Al_2O_3. The remaining 10 % is composed of Cr_2O_3, CaO, MnO, and FeO [39]. Control of slag composition is important with respect to melting temperature and fluidity. The approximate melting temperature may be derived from the ternary diagram for $MgO-Al_2O_3-SiO_2$ [42].

Low-carbon ferrochromium is produced by the silicothermic reduction of chromite ore. Silicon is used in the form of ferrosilicochromium, which is produced in submerged arc furnaces (as is high-carbon ferrochromium) by carbon reduction of chromite ore and quartzite. The solubility of carbon in the FeSiCr alloy depends on the silicon content; if the silicon content is higher, the carbon content is lower. The carbon solubility at different silicon contents and temperatures is shown in Table 7 [43]. The reduction of Cr_2O_3 by Si is enhanced by addition of lime (CaO), which reduces the activity of SiO_2 in the slag. The reduction may be written as follows:

$$FeO \cdot Cr_2O_3 + 2\,Si + 4\,CaO \longrightarrow (2\,Cr + Fe) + 2\,Ca_2SiO_4$$

Special chromium master alloys like CrMo, CrW, and CrNb are produced aluminothermically by coreduction of the relevant oxides together with chromium oxide (see Section 1.4).

3.3.1. High-Carbon Ferrochromium

High-carbon ferrochromium is produced by direct reduction of chromite ores with carbon (coke, coal, or charcoal) in large, three-phase submerged arc furnaces with 10 – 50 MV · A capacity (corresponding to 15 000 – 60 000 t/a of ferrochromium production). Elkem in Sweden operates the largest known electric arc furnace with 105 MV · A capacity. This furnace is used to produce high-carbon ferrochromium and ferrosilicochromium.

Submerged arc furnaces work continuously as a low-shaft electric furnace, where the burden is charged around the self-burning Søderberg electrodes. These electrodes are deeply immersed in the burden column and discharge to the reduced liquid products, i.e., the metal and slag, on the bottom of the furnace. Metal and slag are tapped at regular intervals through tapholes near the furnace bottom; they flow into a ladle with a slag overflow leading to another ladle or to a slag pan. After the slag has been skimmed, the metal is poured into heavy, flat cast-iron molds lined with sand or into sand molds. The furnace hearth and the walls in the reaction zone are normally lined with carbon blocks or a ramming mix; in special cases, magnesite is used. The walls of the shaft are lined with fireclay or magnesite.

In older plants with relatively small submerged arc furnaces, the CO reaction gas is burned at the top of the furnace and then scrubbed. Modern submerged arc furnaces with high capacities are totally closed, and the unburned reaction gas (ca. 90 % CO) is scrubbed. Thus, the volume of the off-gas can be minimized (by a factor of 10 – 20 compared to that of the open top furnace), and the investment costs for the gas-cleaning plant are correspondingly lower. Furthermore, the combustible gas can be

Table 7. Solubility of carbon in ferrosilicochromium as a function of temperature, calculated from equations given in [43]

Silicon content ∗,	Carbon content, wt %			
wt %	1550 °C	1450 °C	1350 °C	1300 °C
20	2.34	2.21	2.08	2.02
30	0.66	0.60	0.54	(solid)
40	0.07	0.05	0.03	0.02
50	0.03	0.02	(solid)	(solid)

∗ Cr : Fe = 2.67.

used as a fuel for processes, such as calcining the limestone and drying and preheating the ore or the whole burden, or for producing energy.

The coke rate is calculated on the basis of the stoichiometric requirement of the oxides and on the amount of dissolved carbon in the alloy; allowance is made for some combustion at the top of the furnace and for reaction with moisture. The slag composition is important to produce metal of desired quality and to maintain smooth furnace operation. Therefore, the slag is analyzed from tap to tap or once per shift, and the additives are altered accordingly. In modern plants, calculation of the burden and collection of data from the furnace are computerized [44].

The silicon content in high-carbon ferrochromium is dependent on the reduction temperature. High-melting slags lead to higher silicon content in the alloy (see below; ferrochromium with 4 – 6 % C). A typical slag composition is 30 – 33 % SiO_2, 26 – 28 % Al_2O_3, 20 – 25 % MgO, 3 – 7 % CaO, and 8 – 13 % Cr_2O_3 for charge chrome containing 52 – 54 % Cr, 6 – 7 % C, and 2 – 4 % Si produced from Transvaal ore or charge chrome containing 63 – 67 % Cr, 5 – 7 % C, and 3 – 6 % Si produced from Zimbabwe ore.

The lumpiness of the ore and the coke quality are important to maintain a proper submerged arc process. Because of the high coke rate (ca. 25 %), the coke properties (size, bulk density, volatile matter, and fixed carbon) are mainly responsible for the electrical resistance of the burden.

The current/potential ratio (*I/U*) is particularly important for large furnaces to maintain a reasonable power factor [45]; as a rule, *I/U* is approximately 200. Therefore, the electric power must be supplied at high current and low potential (e.g., 50 kA at 250 V). Specially constructed leads to the electrodes are necessary to carry the secondary current in order to avoid self-induction and energy losses [46].

Because more than 80 % of chromite ores mined in the western world are finer than 10 mm, agglomeration processes have been adopted to provide a good burden porosity in the electric arc furnaces, especially in the larger ones [37]. In some ferrochromium plants, fines and lumpy ores are blended in ratios between 1 : 1 and 4 : 1. Thus, Union Carbide in Tubatse, South Africa, produces 120 000 t/a of charge chrome in three open submerged arc furnaces, each with a capacity of 30 MV · A. A blend of 80 % fines and 20 % lumpy ore is used. A submerged arc process can be maintained by using raking machines to break the crusts that are formed [47]. A Cr recovery of 75 % was achieved for the production of a charge chrome containing 52.5 % Cr, 6.4 % C, 3.5 % Si, 0.04 % S, and 0.02 % P, and a slag containing 38 % SiO_2, 25.5 % MgO, 4.5 % CaO, 32 % Al_2O_3, and 13 % Cr_2O_3. The power consumption was 4000 kW · h/t of alloy.

Middelburg Steel & Alloys in South Africa installed a briquetting plant with a capacity of 250 000 t/a. The use of briquettes improved the furnace performance and resulted in higher productivity and lower specific energy consumption. In two semiclosed submerged arc furnaces, each with a capacity of 30 MV · A, 100 000 t/a of charge chrome (53.5 % Cr, 7.4 % C, 2 – 3 % Si, 0.015 % S, and 0.015 % P) was produced with a power consumption of 3800 kW · h/t of alloy [48].

Pelletizing of chromite fines and concentrates as an agglomeration process has been adopted by Gesellschaft für Elektrometallurgie (GfE) in the Federal Republic of Germany using the Lepol (grate kiln) process [49], and by Ferrolegeringar in Trollhättan, Sweden, using the Cobo (cold bond) process [50]. These pellets can be used for the production of ferrochromium and silicochromium.

Sintered chrome ore fines were used successfully in Japan. However, briquetting is the main agglomeration method used [37].

Figure 4 (see next page) shows the new process developed by Outokumpu Oy in Finland for the production of ferrochromium from their own chromite-containing deposit near the town of Kemi [51], [52].

The chromite concentrate is pelletized, using bentonite as a binder. After sintering in a shaft furnace, the pellets are blended with fluxes and coke. This burden is then preheated in a rotary kiln at 1000 – 1100 °C and charged to a fully closed 24 MV · A submerged arc furnace producing 60 000 t/a of charge chrome (53.5 % Cr, 7 % C, and 2.5 % Si) and slag (30.2 % SiO_2, 24.9 % MgO, 6.8 % CaO, 25.9 % Al_2O_3, 6.8 % Cr, and 1.8 % Fe); Cr recovery is 84 %. The off-gas from the furnace (ca. 90 % CO) has a heat value of 3.07 kW · h/m^3 (STP) and is used as a fuel for the shaft furnace and for heating the kiln.

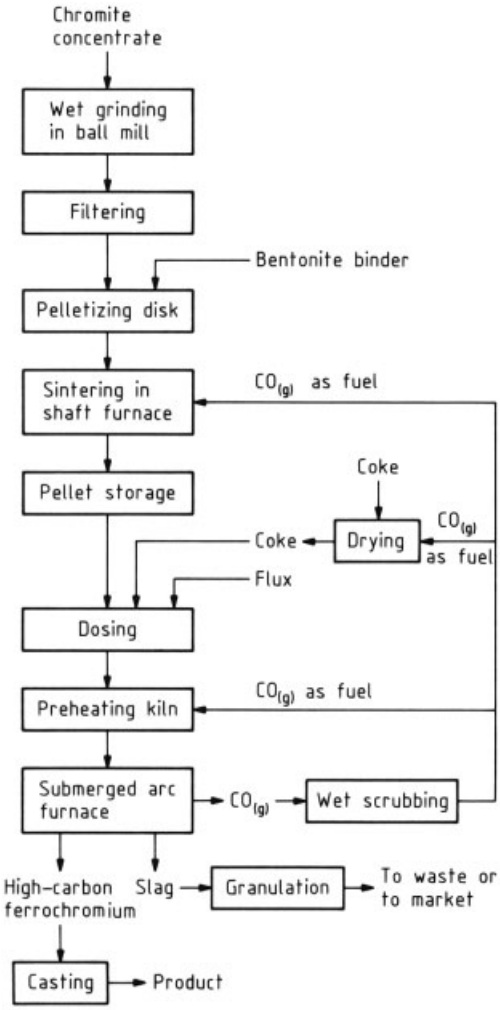

Figure 4. Outokumpu Oy high-carbon ferrochromium process [51], [52]

This technology results in a low specific energy consumption of 2600 – 2800 kW · h/t of charge chrome. The process has been adopted in other countries (Orissa Mining in India, Elazig in Turkey, Hellenic Ferroalloys in Greece, and Ferrochrome Philippines in the Philippines). Ferrochrome Philippines successfully began production in 1984 in a 20 MV · A furnace rated for 50 000 t/a of high-carbon ferrochromium [53].

A further improvement in specific energy consumption was achieved by the SRC process (solid-state reduction of chrome ores), developed by Showa Denko K. K. in Japan [54], [55]. This process is shown in Figure 5. The addition of carbon and flux during pelletizing resulted in a reduction of iron oxide and a partial reduction of chromium oxide during sintering in a rotary kiln at 1350 – 1450 °C. Hot charging a burden containing 60 % prereduced pellets in a closed 18 MV · A submerged arc furnace required an energy consumption of 2000 – 2100 kW · h/t of alloy for an annual production of 50 000 t high-carbon ferrochromium (57 – 60 % Cr, ca. 8 % C, and ca. 3 % Si). This process was also adopted by Johannesburg Consolidated Investment (South Africa) in a 180 000 t/a ferrochromium plant.

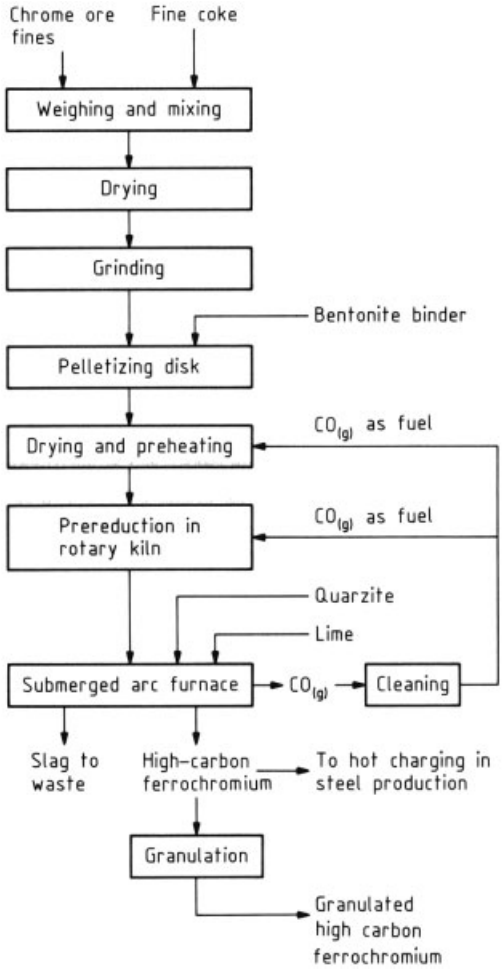

Figure 5. Showa Denko high-carbon ferrochromium process [54], [55]

A special ferrochromium 4 – 6 % C grade with 60 – 72 % Cr and max. 1.5 % Si was used for many years in the steel industry before inexpensive charge chrome from South Africa became more attractive. This ferrochromium is not sat-

urated with carbon. It is produced the same way as charge chrome in large submerged arc furnaces, but with a magnesite lining instead of a carbon lining. This process has been improved by using chromite ore mixes containing a definite fraction of hard and lumpy refractory ores with a high MgO content to produce a high-melting slag (e.g., 29 – 32 % SiO_2, 32 – 35 % MgO, 29 – 32 % Al_2O_3, and 3.5 – 6.0 % Cr_2O_3) [56], [57]. At a higher reduction temperature of 1500 – 1700 °C, the following refining reaction of the refractory ore takes place [40]:

$$3\,(Cr, Fe)_7C_3 + Cr_2O_3 \longrightarrow (Cr, Fe)_{23}C_6 + 3\,CO$$

Because of the low sulfur specification (< 0.05 %) in this grade, coke with low sulfur content is used as the reductant; 30 – 35 % of the sulfur in the coke is transferred to the metal. Alternatively, desulfurizing slags may be used. The ferrochromium can be poured into a highly basic fluid slag, e.g., slag from the LC-ferrochromium process. The sulfur forms CaS, which is removed in the slag.

A new process for producing ferrochromium with 5 % C and < 1 % Si from unagglomerated chromite fines in a transferred arc plasma furnace has been developed by Tetronics Research & Development [58]. Commercialization of this process has been accomplished in South Africa, where a 10.8 MV · A plasma furnace has been built [59].

3.3.2. Medium-Carbon Ferrochromium

Medium-carbon ferrochromium with 0.5 – 4 % C can be produced by refining high-carbon ferrochromium or by silicothermic reduction of chromite ores. Batch refining of high-carbon ferrochromium with refractory chromite ores in an electric arc furnace is no longer used because of the high power consumption of 8000 – 9000 kW · h/t of ferrochromium [46].

Decarburization of high-carbon ferrochromium in an oxygen-blown converter is more economical, especially on a large scale. In the United States and Japan, a top-blowing process with oxygen using water-cooled lances to the metal surface was used. In the Federal Republic of Germany, a bottom-blowing process (OBM = oxygen bottom Maxhütte or Q-BOP) was introduced in the 1970s [60]. This process is characterized by oxygen injection from the bottom into the metal bath with the oxygen jet, resulting in a high decarburization rate of ca 0.3 % C/min. Because of the high bath temperature (1800 – 1850 °C), mainly the carbon is oxidized. Chromium recovery is high; for example, recovery of product with 0.8 – 1 % C is 85 %, and it is improved to 90 – 93 % by adding silicochromium and lime. A typical sequence for a blow is shown in Figure 6 [61].

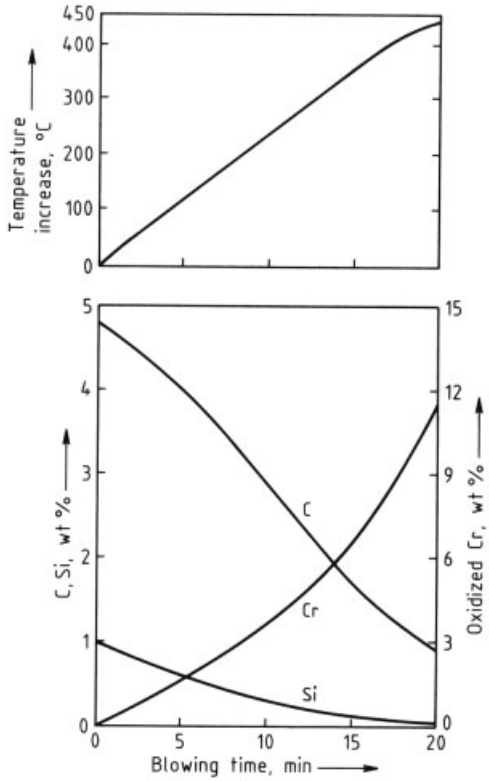

Figure 6. Sequence of an OBM-oxygen blow of high-carbon ferrochromium [61]

Because demand for medium-carbon ferrochromium is small compared to demand for the high-carbon material, the decarburization processes are rarely used. However, reduction of chromite ores with silicon in the form of silicochromium is used and is an economical production method for medium-carbon ferrochromium because the low-carbon grade can be produced as well. To make ferrochromium with 1 – 2 % C, a silicochromium alloy with a Si content of 25 – 30 % can be used (see Table 7, p. 52). The

power consumption, including the energy for silicochromium, amounts to 5000 – 6000 kW · h/t of ferrochromium (see Section 3.3.3).

3.3.3. Low-Carbon Ferrochromium

Low-carbon ferrochromium is produced mainly by silicothermic reduction of chromite ores. This process dates back to the Swedish *three-step process* introduced in 1920 by A. B. Ferrolegeringar in Trollhättan [46].

In step one, chromite ore was totally reduced with carbon in an electric arc furnace to an intermediate ferrochromium (Halbprodukt) with 60 – 65 % Cr, 5 – 6 % C, and 5 – 7 % Si, and a discard slag. In step two, this intermediate ferrochromium was smelted with quartzite and coke in a slagless process to produce silicochromium (40 – 45 % Si, ca. 40 % Cr, and < 0.5 % C). Finally, this silicochromium was used as a reductant in the third step; the crushed alloy was thrown onto the surface of a chrome ore – lime melt in an arc furnace. The tapped ferrochromium contained ca. 70 % Cr, ca. 1 % Si, and 0.03 – 0.05 % C. The slag, which contained 20 – 25 % Cr_2O_3, was returned to the first step to improve Cr recovery.

This Swedish process was modified, especially when the direct silicochromium production by carbothermic coreduction of chromite ore and quartzite in large submerged arc furnaces in one step proved more economical than the slagless process. Furthermore, the silicothermic conversion was improved by adding more lime to form a highly basic slag (CaO/SiO_2 ratio of ca. 2). Thus, ferrochromium with < 1.5 % Si was produced with a lean slag (ca. 5 % Cr_2O_3) in one step. Residual elements, i.e., S and N, were low (< 0.01 % and < 0.02 %, respectively). The power consumption was 3000 – 3500 kW · h/t of ferrochromium for the silicothermic process (arc furnaces with 2 – 10 MV · A capacity) and, including the power consumption for the silicochromium production, was 8000 – 9000 kW · h/t ferrochromium.

The most important process for the production of LC ferrochromium is the Perrin process, developed in 1937 by the Sociétié Ugine in France [62]. This process combines the direct reduction of silicochromium in a submerged arc furnace with a countercurrent silicothermic reduction of a chrome ore – lime melt (produced in an arc furnace) in a reaction ladle.

Production of Silicochromium. Silicochromium is not only used as a reductant in the production of LC ferrochromium, but also as an alloying and deoxidizing agent in the steel industry.

Silicochromium is made in large three-phase submerged arc furnaces of 10 – 36 MV · A capacity. These furnaces are of the same type as those used to produce HC ferrochromium. The lower reaction zone is lined with carbon and the upper side walls of the furnace shaft are lined with fireclay bricks. Maintaining the Søderberg electrodes deep in the burden column (30 – 50 cm above the hearth) is important to reduce the SiO_2 to Si. Therefore, coke with low electric conductivity and low bulk density (< 400 kg/m^3) is used (e.g., gas coke).

The electrical conditions in silicochromium furnaces are somewhat different from those in furnaces used to produce HC ferrochromium; e.g., the current/potential ratio is higher (I/U = ca. 500) [46]. The volume of $CO_{(g)}$ from silicochromium production is much higher than from HC ferrochromium production. Therefore, a porous burden is essential. To scrub the huge volumes of $CO_{(g)}$ produced, modern silicochromium furnaces are closed systems, and the unburned gas is used as a fuel.

To produce an alloy composed of 45 % Si, 40 % Cr, and 0.02 % C, a burden containing 100 parts of a refractory chromite (high Cr/Fe ratio of ca. 3 and a MgO/Al_2O_3 ratio of 1.2 – 1.4), 110 parts of quartzite, and 65 parts of coke (85 % fixed C) is continuously charged to the submerged arc furnace. Metal and slag are tapped every few hours. The slag contains 48 – 50 % SiO_2, 25 – 27 % MgO, 19 – 21 % Al_2O_3, and 1 – 3 % Cr_2O_3. The Cr recovery is 93 – 95 %, and the power consumption is 7500 – 8500 kW · h/t of alloy (45 % Si). The power consumption is primarily dependent on the Si content of the alloy. For a silicochromium with 40 % Si and 45 % Cr, the power consumption is 6500 – 7500 kW · h/t.

Production of LC Ferrochromium by the Perrin Process. Single-stage silicochromium production is one portion of the Perrin process. This process requires two furnaces: a submerged

arc furnace to produce silicochromium and an open-top electric arc furnace to melt a chromite ore – lime slag. A flow sheet of the entire process is shown in Figure 7.

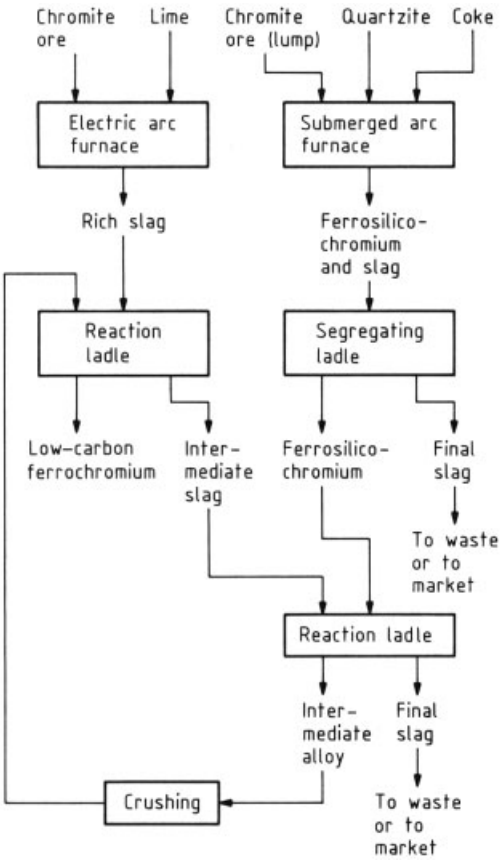

Figure 7. Low-carbon ferrochromium production by the Perrin process [62] (according to [39])

The silicochromium and slag from the submerged arc furnace are tapped to a lined segregating ladle and are allowed to settle for 1 – 2 h. Dissolved SiC floats to the top. The separated alloy (40 – 45 % Si, 45 – 40 % Cr, and 0.05 – 0.02 % C) is removed through a taphole at the bottom of the ladle. The alloy is reacted with an intermediate liquid slag (8 – 10 % Cr_2O_3) in a second basic-lined reaction ladle to produce an intermediate alloy and a discard slag. The intermediate alloy (20 – 25 % Si, 60 – 55 % Cr, and 0.05 – 0.03 % C) is allowed to solidify and is separated from the slag (< 1 % Cr_2O_3). Because of its high basicity, the slag disintegrates to powder (falling slag).

The rich slag (chrome ore – lime melt with 25 – 27 % Cr_2O_3, 7 – 8 % FeO, 2 – 3 % SiO_2, and 45 – 48 % CaO) is tapped from the open-top electric arc furnace in weighed amounts to another basic-lined reaction ladle. Crushed intermediate alloy from the first reaction ladle is added to form the desired LC ferrochromium (ca. 70 % Cr, < 1.5 % Si, and 0.02 – 0.05 % C) and an intermediate slag (8 – 10 % Cr_2O_3). This slag is reacted with silicochromium to lower the Cr_2O_3 content to < 1 %.

Production of 1 t of silicochromium by this process typically requires 1450 kg of chrome ore, 1500 kg of quartzite, 870 kg of coke, 32 – 35 kg of electrodes, and 7700 kW · h power [39]. Production of 1 t of ferrochromium requires 1440 kg of chrome ore, 1250 kg of lime, 660 kg of silicochromium (45 % Si and 40 % Cr), 22 kg of electrodes, and 3200 kW · h power.

Therefore, the total energy consumed to produce 1 t of LC ferrochromium, including silicochromium production, is 8200 kW · h. The overall recovery of Cr by the Perrin process is 90 – 92 %.

Low-carbon ferrochromium is tough and difficult to break. Therefore, it is poured into flat molds, sometimes filled with liquid slag to protect the mold. Quenching the hot alloy in water facilitates breaking it into small lumps, which is performed with heavy pneumatic hammers.

Many variations of the Perrin process have been developed. For example, chrome ore or lime have been added to the exothermic reactions, which occur in the two reaction ladles (Fig. 7), to "dampen" the reaction and to save energy. Another variation is the selective reduction of FeO, producing an iron-rich alloy; the remaining slag, high in Cr_2O_3, is reacted with silicochromium high in Si, to give LC ferrochromium (80 – 90 % Cr). By selectively reducing FeO, CoO is also reduced in the iron-rich alloy and a Co-free ferrochromium can be produced. This is used to produce alloyed steels for atomic energy plants.

Production of LC Ferrochromium by the Simplex Process. From 1943 to 1953 Union Carbide developed a new production process for LC ferrochromium, the so-called Simplex process, in which finely ground HC ferrochromium was decarburized in the solid state with oxidized ferrochromium by vacuum annealing [63–65].

The plant in Marietta, Ohio, went into production in 1953 and is now managed by Elkem Metals Co., Pittsburgh, Pennsylvania. The flow sheet of the process is shown in Figure 8.

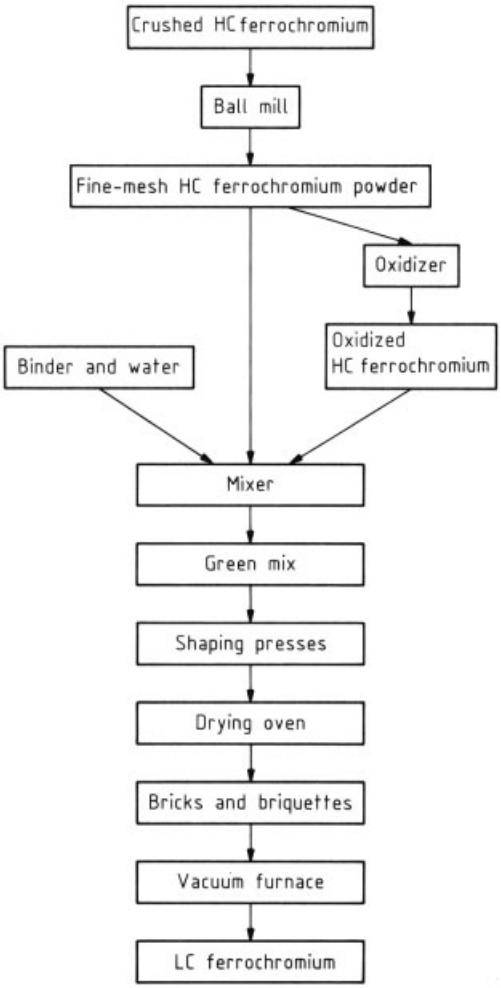

Figure 8. Production of low-carbon ferrochromium by the Simplex process [64]

High-carbon ferrochromium is crushed, pulverized in ball mills, and then oxidized in suspension in a vertical gas- or oil-fired shaft furnace [66]. The proper stoichiometric C/O ratio for decarburizing is attained by mixing the oxidized material with HC ferrochromium powder according to the following reactions:

$$3\,Cr_7C_3 + Cr_2O_3 \longrightarrow Cr_{23}C_6 + 3\,CO$$
$$1/2\,Cr_{23}C_6 + Cr_2O_3 \longrightarrow 27/2\,Cr + 3\,CO$$

The mixture is either formed into briquettes or loosely packed in corrugated pasteboard boxes, placed on a refractory topped bogie hearth (33.5 m × 3.35 m), and rolled into the vacuum chamber, which is a horizontal steel cylinder (43 m in length × 4.6 m in diameter). The graphite resistance heating elements are suspended above the charge from the refractory-lined roof. The vacuum (0.01 – 0.4 kPa) is generated by multistage steam ejectors. The process requires a special heating program up to 1370 °C, which follows the progress of decarburization. When the decarburization of the charge is complete, the power is shut off and the furnace is cooled by flooding with argon.

A complete reaction cycle requires 4 – 5 d; ca. 100 t of material is produced. A typical analysis of LC ferrochromium from the Simplex process is ca. 70 % Cr, ca. 1 % Si, and 0.008 – 0.010 % C. Oxygen is higher (ca. 1 %) than in LC ferrochromium from the Perrin process (ca. 0.2 %).

3.3.4. Nitrogen-Containing Low-Carbon Ferrochromium

Chromium has a high affinity for nitrogen and forms the nitrides Cr_2N and CrN (→ Nitrides). Low-carbon ferrochromium can form nitrides either in the solid or in the liquid state. These alloys with up to 10 wt % N are used in steel containing nitrogen as well as chromium [32]. The Simplex process is convenient for forming nitrides in LC ferrochromium in the solid state. During the cooling cycle, nitrogen is added to the vacuum vessel and the optimum temperature range of 1200 – 800 °C is held over a longer period to produce an alloy with 7 – 10 % N.

Silicothermically produced LC ferrochromium can form nitrides when pulverized and heated in boxes in a continuous annealing furnace at 1000 °C under nitrogen [67]. The sintered product (ϱ 5 – 6 g/cm^3) contains 5 – 10 % N as Cr_2N.

On the other hand, LC ferrochromium can form nitrides in the liquid state with nitrogen gas in a melting furnace, e.g., an induction furnace [68]. The tapped alloy contains 2 – 4 % N and is very homogeneous (ϱ 7.0 g/cm^3) [69].

3.3.5. Other Chromium Master Alloys

Chromium master alloys other than ferrochromium are produced in much smaller amounts. They are used mainly to produce heat-resistant alloys, based on nickel or cobalt (superalloys), and abrasion-resistant materials, and are used to alloy small amounts of chromium to copper or aluminum (see Chap. 2).

Nickel – chromium master alloys containing 50 – 80 % chromium may be made by aluminothermic processes or by melting chromium metal and nickel in induction-heated furnaces. The aluminothermic process must be performed in water-cooled molds in an inert atmosphere because of the high quality required of these master alloys [70]. For direct melting of chromium metal and nickel, a vacuum induction furnace or an inert gas atmosphere is also preferred; refining occurs simultaneously during melting [71].

Binary chromium master alloys, such as *chromium – molybdenum* (with 30 % Mo) or *chromium – niobium* (with 30 – 80 % Nb) [72], [73], are made by aluminothermic reduction of the corresponding oxides. Master alloys containing more than two metals, e.g., *cobalt – chromium – tungsten* (ca. 43 wt % Co, ca. 37 wt % Cr, ca. 16 wt % W, ca. 3 wt % C), or *cobalt – chromium – molybdenum* (26 – 30 wt % Cr, 4 – 5 wt % Mo, rest Co) [73], are made by direct melting of the metals in an inert atmosphere or in vacuum induction furnaces. *Copper – chromium* master alloys containing up to 10 wt % Cr, e.g., V – CuCr 10 [74], are made by directly alloying chromium with an oxygen-free copper melt; the resulting alloy is quickly cast into flat pigs. The same method is used for the production of *aluminum – chromium* master alloys with 5, 10, or 20 wt % Cr by alloying comminuted chromium metal with a superheated aluminum melt. The master alloy containing 5 – 6 % Al is standard in the Federal Republic of Germany, i.e., V – AlCr 5 [75]. Another aluminum – chromium master alloy containing ca. 50 wt % Cr and 50 wt % Al is made for special purposes [73].

3.4. Quality Specifications, Storage and Transportation, and Trade Names

International standards of specifications and conditions of delivery for ferrochromium and ferrosilicochromium are given in ISO 5448 [76] and ISO 5449 [77], respectively. The following specifications are contained in tables in reference [76]:

1) Chromium content of ferrochromium alloys ranges from 45 to 95 % [76, Table 1]; e.g., the designation is FeCr 50 for 45 – 55 % Cr content.
2) Composition of HC ferrochromiums with normal phosphorus [76, Table 2] and with low phosphorus content [76, Table 3]; e.g., FeCr 50 C 50 (45 – 55 % Cr, 4 – 6 % C, max. 1.5 % Si, max. 0.050 % P, and max. 0.10 % S); or FeCr 60 C 50 Si 2 LP (55 – 65 % Cr, 4 – 6 % C, 1.5 – 3 % Si, and max. 0.03 % P). HC ferrochromium is standardized to 10 % C. Ferrochromium with low sulfur content is designated LS (low sulfur, i.e., max. 0.05 %).
3) Analysis of MC ferrochromiums with normal phosphorus [76, Table 4] and with low phosphorus [76, Table 5]; carbon content ranges from 0.5 – 1.0 %, 1.0 – 2.0 %, up to 2.0 – 4.0 %, e.g., FeCr 70 C 20 (65 – 75 % Cr, 1.0 – 2.0 % C, max. 1.5 % Si, max. 0.050 % P, and max. 0.050 % S); or FeCr 60 C 10 LP (55 – 65 % Cr, 0.5 – 1.0 % C, max. 1.5 % Si, max. 0.03 % P, and max. 0.050 % S).
4) Standard chemical composition of LC ferrochromium for normal [76, Table 6] and for low [76, Table 7] phosphorus alloys; carbon content ranges from less than 0.015 %, 0.015 – 0.030 %, up to 0.25 – 0.50 %; e.g., FeCr 60 C 05 (55 – 65 % Cr, 0.030 – 0.050 % C, max. 1.5 % Si, max. 0.050 % P, max. 0.030 % S, and max. 0.15 % N) or FeCr 70 C 03 LP (65 – 75 % Cr, 0.015 – 0.030 % C, max. 1.5 % Si, max. 0.030 % P, max. 0.030 % S, and max. 0.15 % N).
5) LC ferrochromium with high chromium content (75 – 95 %) [76, Table 8]; e.g., FeCr 90 C 03 (85 – 95 % Cr, 0.015 – 0.030 % C, max. 1.5 % Si, max. 0.020 % P, max. 0.030 % S, max. 0.02 % Co, and max. 0.20 % N).
6) Nitrogenated LC ferrochromiums [76, Table 9]; the three given are smelted FeCr... C1 N3, sintered FeCr... C1 N7, and sintered

FeCr... C1 N7 Si. The Cr contents given in [76, Table 1] are valid for these alloys. The nitrogen contents are 2 – 4 % in the smelted qualities and 4 – 10 % in the sintered qualities.

7) Particle size ranges for ferrochromium deliveries [76, Table 9]; there are seven classes from 100 – 315 mm (Class 1) to $\leqq$ 3.15 mm (Class 7).

ISO Standard 5448 [76] also covers testing (sampling and chemical analysis) as well as storage and transportation according to international regulations. ISO Standard 5449 [77] for ferrosilicochromium specifications and conditions of delivery recognizes 12 different compositions with increasing Si content, ranging from FeCrSi 15 (min. 55.0 % Cr, 10.0 – 18.0 % Si, max. 6.0 % C, max. 0.050 % P, and max. 0.030 % S) up to FeCrSi 50 (min. 20.0 % Cr, 45 – 60 % Si, max. 0.1 % C, max. 0.030 % P, and max. 0.030 % S). Low-carbon and low-phosphorus grades are FeCrSi 50 LC (max. 0.05 % C) and FeCrSi 48 LP (min. 35.0 % Cr, 42 – 55 % Si, max. 0.05 % C, max. 0.020 % P, and max. 0.010 % S). The particle size of ferrosilicochromium deliveries is specified in seven classes as for ferrochromium. Ferrochromium and ferrosilicochromium are shipped in bulk or packaged in containers or steel drums.

In the United States, standards for ferrochromium and ferrosilicochromium are specified in ASTM A 101 – 73 [78] and ASTM A 482 – 76 [79]. Also recommended are methods for sampling (ASTM – E32) and chemical analyses (ASTM –E31). Comparisons between the standards in the Federal Republic of Germany for ferrochromium and ferrosilicochromium in DIN 17 565 [80] and the ISO standards are given in Tables 8 and 9 [81].

Internationally approved methods for sampling [82] and analytical determination of chromium in ferrochromium and ferrosilicochromium [83] and of silicon in ferrosilicochromium [84] are available as a draft.

Table 8. Comparison of the chemical composition of ferrochromium ISO 5448 and DIN 17565 [81]

Designation	% Cr		% C		% Si		% P max.		% S max.	
	DIN	ISO	DIN	ISO	DIN	ISO	DIN	ISO	DIN	ISO
FeCr 70 C 5	60 – 72	–	4.0 – 6.0	–	< 1.5	–	0.030	–	0.050	–
FeCr 50/60/70 C 50	–	45 – 75 [a]	–	4.0 – 6.0	–	< 1.5	–	0.050	–	0.10
FeCr 50/60/70 LSLP	–	45 – 75 [a]	–	4.0 – 6.0	–	< 1.5	–	0.030	–	0.05
FeCr 70 C	65 – 75	–	0.5 – 4.0 [b]	–	< 1.5	–	0.030	–	0.050	–
FeCr 50/60/70 C	–	45 – 75 [a]	–	0.5 – 4.0 [b]	–	< 1.5	–	0.050	–	0.050
FeCr 70 C [c]	65 – 75	–	0.01 – 0.50 [d]	–	< 1.5	–	0.030	–	0.010	–
FeCr 50/60/70 C [c]	–	45 – 75 [a]	–	0.015 – 0.50 [e]	–	< 1.5	–	0.050	–	0.030

[a] FeCr 50 means 45 – 55 % Cr, FeCr 60 means 55 – 65 % Cr, FeCr 70 means 65 – 75 % Cr.
[b] In four grades.
[c] 0.15 % N.
[d] In seven grades.
[e] In six grades.

Table 9. Comparison of the chemical composition of ferrosilicochromium ISO 5449 and DIN 17565 [81]

Designation	% Cr		% Si		% C max.		% P max.		% S max.	
	DIN	ISO	DIN	ISO	DIN	ISO	DIN	ISO	DIN	ISO
FeCrSi 15	–	> 55.0	–	10.0 – 18.0	–	6.0	–	0.050	–	0.030
FeCrSi 20/22	55 – 65	> 55.0	25 – 20	20.0 – 25.0	0.050	0.05	0.020	0.030	0.010	0.030
FeCrSi 23	–	> 45.0	–	18.0 – 28.0	–	3.5	–	0.050	–	0.030
FeCrSi 26	–	> 45.0	–	24.0 – 28.0	–	1.5	–	0.030	–	0.030
FeCrSi 33	–	> 43.0	–	28.0 – 38.0	–	1.0	–	0.050	–	0.030
FeCrSi 40	40 – 45	> 35.0	45 – 35	35.0 – 40.0	0.050	0.2	0.020	0.030	0.010	0.030
FeCrSi 45	–	> 28.0	–	40.0 – 45.0	–	0.1	–	0.030	–	0.030
FeCrSi 50 LC	–	> 20.0	–	45.0 – 60.0	–	0.05	–	0.030	–	0.030
FeCrSi 55	–	> 28.0	–	50.0 – 55.0	–	0.03	–	0.030	–	0.030
FeCrSi 48 LP	–	> 35.0	–	42.0 – 55.0	–	0.05	–	0.020	–	0.010

Table 10. Special and nonstandardized ferrochromium and ferrosilicochromium alloys

Trade name	Analysis, wt %							Producer
	Cr	C	Si	P	S	N	Co	
Charge chrome [a]	53 – 58	5 – 8	6 – 3	–	–	–	–	–
EHC ferro-	min. 66	min. 9.0	0.5	max. 0.03	max. 0.05	–	–	GfE/EWW [b]
chromium			1.0					
Simplex ferro- chromium	ca. 70	max. 0.01	ca. 1.0	–	–	–	–	Union Carbide/Elkem
LCN ferro-	ca. 70	max. 0.05	max. 1.5	–	–	max. 0.020	–	GfE/EWW [b]
chromium								
LCNSi ferro-	ca. 70	max. 0.05	max. 0.30	–	–	max. 0.020	–	GfE/EWW [b]
chromium	ca. 70	max. 0.03	max. 0.20	–	–	max. 0.020	–	GfE/EWW [b]
Ferrochromium	80 – 85	max. 0.06	max. 2.5	–	–	max. 0.20	max. 0.02	ABF/FTA [c]
reactor grade	80 – 85	max. 0.05	max. 1.50	–	–	max. 0.20	max. 0.02	ABF/FTA [c]
Cromax-	85 – 90	max. 0.10	max. 2.0	max. 0.020	max. 0.020	0.03 – 0.05	0.01 – 0.02	Showa
90 chrome		max. 0.06	max. 1.0	max. 0.015	max. 0.015	–	–	Denko
		max. 0.03	max. 0.5	min. 0.007	min. 0.007	–	–	
Silicochrom 40	35 – 40	max. 0.05	35 – 40	max. 0.03	–	–	–	GfE/EWW [b]
Silicochrom 60	55 – 65	max. 0.05	20 – 25	max. 0.03	–	–	–	GfE/EWW [b]

[a] Alloy has no definite specification.
[b] Gesellschaft für Elektrometallurgie/Elektrowerk Weisweiler.
[c] Aktiebolaget Ferrolegeringar/Trollhätteverken.

Table 10 contains trade names and the relevant analyses of special and nonstandardized ferrochromium and ferrosilicochromium alloys. Charge chrome is a relatively cheap high-carbon ferrochromium with no definite specification (see Section 3.3.1). It can be added to a basic oxygen-blown stainless steel charge [39]. Extra-high-carbon (EHC) ferrochromium is supplied as a powder for abrasive surface-welding purposes. Simplex ferrochromium is a low-carbon ferrochromium produced by the Simplex process (see Section 3.3.3). Low-carbon and -nitrogen (LCN) and low-carbon, -nitrogen, and -silicon (LCNSi) ferrochromium are trade names for special grades used to manufacture corrosion- and heat-resistant steel. A low-carbon ferrochromium reactor grade is used to make corrosion- and heat-resistant steel for nuclear equipment. Cromax is a low-carbon, high-chromium ferrochromium used to produce heat-, corrosion-, and wear-resistant special steel. The high-chromium alloys can partly replace chromium metal in superalloys. Silicochrom 40 is mainly used for deoxidizing steel and for Cr_2O_3 reduction in slags; silicochrom 60 for alloying corrosion- and heat-resistant steel.

3.5. Uses

Ferrochromium is used principally as a master alloy to produce Cr-containing steel and cast iron [32]. Chromium alloyed in steel imparts oxidation and corrosion resistance because of the formation of a thin, continuous, impervious chromium oxide film on the steel surface.

Most ferrochromium is used to manufacture *stainless steel.* Martensitic grades with ca. 13 % Cr are used for applications such as knives, whereas ferritic grades with 18 – 22 % Cr are used as deep-drawing sheets. Austenitic CrNi steels exhibit especially good corrosion (acid) resistance and are used for equipment in the chemical industry as well as for food-processing machinery. The most widely known type is AISI 302, which contains 18 % Cr and 8 % Ni. A small amount of molybdenum further improves the acid resistance. Heat-resistant steel (24 – 26 % Cr and 19 – 21 % Ni) is used in power stations for steam boilers and heat exchangers; superalloys based on nickel or cobalt (ca. 30 % Cr) are used for gas turbine parts.

Chromium addition to steel also improves its mechanical properties (ability to harden,

wear resistance, and retention of strength at elevated temperature). Steel with 0.5 – 2.0 % Cr is widely used in the automotive industry as case-hardening steel ($\leqq$ 0.2 % C and 1.4 – 2.1 % Ni) and heat-treatable steel (0.2 – 0.5 % C, 0.9 – 2.1 % Ni, and 0.15 – 0.35 % Mo). Abrasion-resistant steel for ball bearings is alloyed with 0.5 – 1.6 % Cr (ca. 1 % C and 0.2 % Mo). Permanent magnetic steel contains 1 – 3 % Cr (1.0 – 1.3 % C). In tool steel for cold and hot working and high-speed steel, up to 14 % Cr is alloyed as carbide (in addition to carbides of W, Mo, and V). Stellites (35 – 70 % Co, 25 – 33 % Cr, 10 – 25 % W, and 2 – 4 % C) are highly resistant to abrasion and are used for cutting tools.

Addition of up to 30 % chromium to cast iron improves hardness and heat resistance. Powdered ferrochromium (mainly the high-carbon grade) is used for chromizing the surface of steel or cast iron parts (pack-chromizing or case-chromizing, → Metals, Surface Treatment). A high-carbon ferrochromium with extra-high-carbon content is used in powdered form for abrasive surface welding (see Table 10, EHC Ferrochromium).

3.6. Economic Aspects

Ferrochromium is mainly consumed in highly industrialized countries producing a major amount of high-grade (stainless) steel, e.g., the United States, Europe, and Japan. However, production is shifting from these consuming countries to those with large ore resources; 93 % of the known potential chromite reserves of the world are in Africa, mainly in the Republic of South Africa, and ca. 3 % in the former States of USSR and Albania [85] (see Table 1). This shift is a consequence of modern oxygen-blowing processes in combination with argon and vacuum treatment, i.e., argon – oxygen decarburization (AOD) and vacuum – oxygen decarburization (VOD), for the production of stainless steel. These processes allow the use of cheap HC ferrochromium (charge chrome) instead of LC grades. Formerly only LC ferrochromium could be used. Because power consumption for production of HC ferrochromium is only about half that of the LC grade, the availability of cheap hydroelectric power in Scandinavia is no longer as advantageous. The change in ferrochromium consumption from LC to HC grades first took place in the United States and later, in the 1970s, in Europe [86].

In 1984, world production of marketable, i.e., lumpy, ore and concentrates was 9.2×10^6 t, 11 % of which was produced in Western Europe, Finland, Turkey, Greece, and Cyprus, 37 % in Eastern Europe (USSR and Albania), 4 % in Latin America (Brazil and Cuba), 9 % in Asia (Philippines, India, Iran, Vietnam, Japan, and Pakistan), and 38 % in Africa (South Africa, Zimbabwe, Madagascar and Sudan) [87]. In 1984, the world production of chromite ore and concentrates was 9.9×10^6 t, 72 % of which was consumed for metallurgical, 12 % for refractory, and 17 % for chemical purposes.

The metallurgical industry was the main user of chromium and consumed the following amounts in 1980 (based on United States statistics) [88]: 72 % for wrought stainless and heat-resistant steels, 13 % for wrought alloy steel, 3 % for tool steel, 7 % for cast iron and cast steel, and 5 % for other uses. In 1984, the total consumption of primary chromium units (from ferrochromium) in the United States was 195 380 t of which 79 % was for the production of stainless and heat-resisting steels, i.e., 96 kg of chromium units per ton of stainless and heat-resisting steels produced [89]. The ferrochromium consumption in the Federal Republic of Germany was 263 300 t in 1985. The change in consumption of LC to HC ferrochromium over the last 25 years is evident from Table 11 (see next page).

In 1981, consumption of ferrochromium in the western world was estimated to be $(1.32 - 1.36) \times 10^6$ t; ca. 0.5×10^6 t of this was consumed in the European Economic Community (EEC). In 1982, consumption fell drastically because of the recession in the steel industry. Consumption in the western world has since increased to ca. 2×10^6 t [90]. In 1981, the production capacity for ferrochromium plants in the western world was 3.1×10^6 t [86].

Prices of ferrochromium and ferrosilicochromium are reported weekly in *Metal Bulletin* (London). *Erzmetall* (published monthly) reports prices twice a year for raw materials and for ferrochromium. *American Metal Market* is a daily publication that also gives chromium ore and alloy prices.

Table 11. Ferrochromium consumption in the Federal Republic of Germany

	Consumption, t/a					
	1960	1965	1970	1973	1983	1985
LC ferrochromium	43 450	34 650	44 500	33 550	13 000	14 350
MC ferrochromium	20 300	24 400	16 700	18 400	7 500	11 600
HC ferrochromium	7 500	17 550	70 600	113 150	210 000	237 350
Total	71 250	76 600	131 800	165 100	220 500	263 300

Because the United States does not possess chromite deposits, most of the ferrochromium consumed is imported, mainly from Africa [91].

3.7. Environmental Protection; Toxicology and Occupational Health

Considerable attention must be paid to environmental protection during ferrochromium production. Dust evolution from ores and raw materials must be controlled during transportation and storage, and also during preparatory processes.

The off-gas from the electric arc furnaces is cleaned by using devices, such as bag filters, scrubbers, and electrostatic precipitators [46]. Bag filters are mainly used in preparatory processes, e.g., grinding, drying, briquetting [48], and pelletizing [49], and in open or semiclosed submerged arc furnaces [47], [48]. Scrubbers (Venturi and disintegrators) are preferred for cleaning unburned CO gas from closed submerged arc furnaces [51]. Effluent water from the scrubber is treated to meet wastewater regulations. Electrostatic precipitators (Cottrell filters) are also used for cleaning the off-gas from submerged arc furnaces for HC ferrochromium [92] and ferrosilicochromium [93] production. Gas from the chrome ore – lime melting furnace and from the ladle reactions during the Perrin process are cleaned with bag filters [94]. The dust content can be lowered from 150 to 20 mg/m^3 and the emission of Cr(VI) to < 1 mg/m^3 by optimizing the working conditions. Chromium(VI) is mainly responsible for the toxicity of chromium and its compounds (→ Chromium Compounds). Chromic acid, CrO_3, is carcinogenic. The Federal Republic of Germany has enacted an emission protection law and a technical guidance for clean air (TA-Luft) [95]. The TA-Luft restricts the mass of carcinogenic substances, i.e., Cr(VI) compounds (specified as Cr), in respirative form to 1 mg/m^3 (Class II). The total dust concentration in emissions is limited to 20 mg/m^3, and other inorganic materials in dust form, i.e., chromium and its compounds (Class III), to 5 mg/m^3. The maximum allowable concentration (MAK) of CrO_3 (measured as CrO_3) in the air at the workplace is 0.1 mg/m^3 [96].

In the United States the Clean Air Act Amendments, enacted in 1970, mandated the states set standards to limit air pollution. In pursuing this amendment, the U.S. Environmental Protection Agency (EPA) in a joint effort with the Ferroalloys Association studied dust emissions from alloy plants to establish ambient air quality standards [97], [98]. The states enacted process weight regulations and applied them to ferroalloy furnaces. The EPA published their studies "Engineering and Cost Study of the Ferroalloy Industry" and proposed New Source Performance Standards (NSPS) for new furnaces. A comparison of the state regulations and the NSPS (promulgated October, 1974) limitations for HC ferrochromium, charge chrome, and ferrosilicochromium for a 30-MW furnace are shown in Table 12 (see next page) [98].

Regulations on water pollution control (the federal act was passed in 1972) were also proposed by the Ferroalloys Association Environmental Committee and EPA [98]. Finally, the disposal of solid waste (slags, dusts, etc.) is also subject to regulations.

The TLV – TWA (1985 – 1986), which have been adopted, are as follows [99]: 0.5 mg/m^3 for chromium metal and chromium(II) and chromium(III) compounds, and 0.05 mg/m^3 for chromite ore processing (chromate) as Cr and chromium(VI) compounds as Cr. Chromite ore processing and certain water-insoluble Cr(VI)

Table 12. Comparison of NSPS limitations and state regulations for ferrochromium dust emissions [98]

Material	NSPS limitation,		State regulation, kg/h	
	kg/(mw · h)	kg/h	Most stringent	Least stringent
HC ferrochromium	0.23	6.94	13.00	17.55
Charge chrome	0.23	6.94	13.00	17.55
Ferrosilicochromium	0.23	6.94	8.29	10.80

compounds are listed in Appendix A1a as human carcinogens.

4. References

1. *Gmelin,* Chromium, Part A (1962) pp. 1 – 5.
2. J. W. Mellor: *Comprehensive Treatise on Inorganic and Theoretical Chemistry,* **vol. 11,** Longmans Green & Co., London 1931, p. 122.
3. M. J. Udy: *Chromium,* **vol. 1,** Reinhold Publ. Co., New York 1956, p. 1.
4. J. F. Papp: "Mineral Facts and Problems," U.S. Bureau of Mines Bulletin no. 675, 1985, Chapter on Chromium.
5. T. Matsumura, J. Hagiya, *Toya Soda Kenkyu Hokoku* **22** (1978) no. 2, 75 – 89.
6. C. L. Rollinson: *Comprehensive Inorganic Chemistry,* **vol. 3,** Pergamon Press, Oxford 1973, p. 624.
7. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* Interscience, New York 1966, p. 818.
8. L. B. Pankratz: "Thermodynamic Properties of Elements and Oxides," U.S. Bureau of Mines Bulletin no. 672, 1982, p. 117.
9. A. H. Sully: *Chromium,* Academic Press, New York 1954, pp. 21 – 29.
10. M. C. Carosella, J. D. Mettler, *Met. Prog.* **69** (1956) no. 6, 51.
11. J. B. Rosenbaum, R. R. Lloyd, C. C. Merrill, U.S. Bureau of Mines, R.I. no. 5322, March, 1957.
12. M. Pourbaix: *Atlas of Electrochemical Equilibria,* Pergamon Press, Oxford 1966, p. 262.
13. H. F. Reichard, H. R. Spendelow, "Chromium," London Metal Bulletin Survey, London 1977.
14. A. N. Rakitskii, V. I. Trefilov, *Poroshk. Metall.* **160** (1976) no. 4, 20 – 30.
15. Shin Nippon Iron K.K., JP 54 156 013, 1979 (K. Katayama).
16. SAE, ASTM (eds.): *Metals & Alloys in the Unified Numbering System,* 4th ed., SAE, Warrendale, Pa., 1986, pp. 185 – 195, 209 – 211, 217, 359 – 361.
17. R. C. Gibbons: *Woldman's Engineering Alloys,* 6th ed., American Society for Metals, Metals Park, Ohio, 1979.
18. Personal correspondence, J. J. de Barbadillo, INCO Alloys Int., Huntington, W.Va., Sept., 1985.
19. "Precipitation – Hardening Stainless Steel Bar, Wire, Plate and Forging Billets," Armco PH 13-8 MO, Product Data Sheet S-336, Armco Steel Corp., Baltimore 1970.
20. K. Wellinger, P. Gimmel, M. Bodenstein: *Werkstoff-tabellen der Metalle,* 7th ed., Kroner, Stuttgart 1972.
21. "American Metal Market," Battelle Memorial Laboratory, Columbus, Ohio, Nov. 19, 1984, p. 18.
22. Registration Record of Aluminum Association Designations and Chemical Composition Limits for Wrought Aluminum and Wrought Aluminum Alloys, The Aluminum Association, Washington, D.C., Jan., 1984.
23. Registration Record of Aluminum Association Alloy Designations and Chemical Composition Limits for Aluminum Alloys in the Form of Castings and Ingot, The Aluminum Association, Washington, D.C., Jan., 1984.
24. Union Carbide Corp., US 3 592 637, 1971.
25. P. D. Deeley, K. J. A. Kundig, H. R. Spendelow, Jr.: *Ferroalloys and Alloying Additives Handbook,* Shieldalloy Corp., Newfield, N.J., 1982, p. 29.
26. A. H. Sully, E. A. Brandes: *Metallurgy of the Rare-Earth Metals – Chromium,* 2nd ed., Butterworth, London 1967.
27. Alloy Surfaces Co., US 3 222 212, US 3 375 128, 1962.
28. G. M. Baughman, *Iron & Steel Eng.* 1972 (Sept.), 91 – 101.
29. C. T. Sims, W. C. Hagel: *The Superalloys,* J. Wiley & Sons, New York 1972.
30. J. D. Maykuth, E. S. Bartlett, V. D. Barth, "Refractory Metals Review," Defense Metals Information Center (DMIC), Battelle Memorial Institute, Columbus, Ohio, Jan., 1970.

31. A. Coutagne: *La Fabrication des Ferro-Alliages,* Baillière Fils, Paris 1924.
32. P. D. Deeley, K. J. A. Kundig, H. R. Spendelow, Jr.: *Ferroalloys and Alloying Additives Handbook,* 1st ed., Shieldalloy Corp., Newfield, N.J., 1982. Metallurgy Alloy Corp., New York, N.Y., 1981.
33. O. Kubaschewski: *Iron – Binary Phase Diagrams,* Springer-Verlag, Berlin and Verlag Stahleisen mbH, Düsseldorf 1982, pp. 31 – 34.
34. V. G. Rivlin, *Int. Metal. Rev.* **29** (1984) no. 4, 299 – 327.
35. P. Wille, *Stahl Eisen* **103** (1983) no. 3, 127 – 132.
36. F. Ergunalp, *Trans. Inst. Min. Metall.* **89** (1980, Oct.) A179 – A184.
37. D. P. O'Shaughnessy, *Mining Mag.* **147** (1982, Oct.) 291 – 299.
38. H. Fuchs: "Untersuchung des Petrographischen und Mineralogischen Aufbaus von Chromerzen für die Metallurgische Verwertung und ihrer Reduzierbarkeit mit dem Ziel der Aufstellung von Bewertungsrichtlinien," Dissertation Rheinisch-Westfälische-Technische-Hochschule (RWTH), Aachen 1962.
39. A. G. E. Robiette: *Electric Smelting Processes,* C. Griffin Co., London 1973, pp. 150 – 178.
40. K. Willand: "Untersuchungen zur Herstellung von Ferrochrom mit 4 – 6 % Kohlenstoff," Dissertation RWTH, Aachen 1971.
41. J. H. Downing: "Theoretical Basis for Smelting Ferroalloys," *34th Electric Furnace Conference, St. Louis,* Miss., Dec., 1976.
42. E. M. Levin, C. R. Robbins, H. F. McMurdie in M. K. Reser (ed.): *Phase Diagrams for Ceramists,* 2nd ed., compiled at National Bureau of Standards, American Ceramic Society, Columbus, Ohio, 1969, p. 246 (Fig. 712).
43. F. Breuer: "Untersuchungen über das System Chrom-Eisen-Silizium-Kohlenstoff und ihre Nutzanwendung im Bereich der Technischen Silikochrom-Legierungen," Dissertation RWTH, Aachen 1961.
44. M. S. Rennie, *Electr. Furn. Conf. Proc.* **37** (1979), 202 – 209.
45. J. H. Downing, L. Urban, *J. Met.* **18** (1966) 337 – 344.
46. G. Volkert, W. Dautzenberg, J. Willems, G. Zieger et al.: "Ferrochrom und Chrommetall," in Durrer, G. Volkert, K.-D. Frank: *Metallurgie der Ferrolegierungen,* 2nd ed., Springer, Berlin-Heidelberg-New York 1972, pp. 292 – 365.
47. H. D. Rowley, D. M. Legg: "Construction and Operation of Charge Chrome Facilities at Tubatse Ferrochrome (Pty.), Steelport, Transvaal, South Africa," *35th Electric Furnace Conference,* Chicago, Dec. 1977, Special Arc Prepr..
48. W. D. Winship: "Briquetting – An Economic Solution for the Production of Ferro-Chrome in South Africa," *Proc. of the 15th Biennial Conf.,* Inst. for Briquetting and Agglomeration, Montreal, Quebec, 1977, pp. 139 – 152.
49. E. Lankes, W. Boehm: "Experience Made and Operational Results Obtained with a Chrome Ore Pelletizing Plant Working on the LEPOL Process," *Int. Ferroalloy Conf.* (INFACON), Johannesburg, South Africa, Apr., 1974.
50. N. G. Lindberg, T. S. Falk, *CIM Bull.* **69** (1976, Sept.) 117 – 126.
51. J. Relander, M. Honkaniemi: "Production of Ferrochrome from Low-Grade Ores," *Symposium on Extraction of Steel Alloying Metals,* Technical Univ. of Luleå, Sweden, Mar. 15 – 18, 1983, Prepr. pp. 295 – 301.
52. H. Tuovinen, J. Relander, M. Honkaniemi, *Electr. Furn. Conf. Proc.* **36** (1978) 127 – 133.
53. W. Schiffner, *Stahl Eisen* **104** (1984) no. 21, 1099 – 1101.
54. Y. Kanoh, *Met. Bull. (London)*, Ferro Alloys Special Issue, 1971, 83 – 85.
55. K. Ichikawa, *Chem. Eng.* **81** (1974, Apr.) 36 – 37.
56. R. A. Leeper, T. J. Dyrdek, *Electr. Furn. Conf. Proc.* **23** (1965) 110 – 114.
57. H.-J. Retelsdorf, R. M. Fichte, F. Breuer, H. Zimmermann, *Metall (Berlin)* **36** (1982) 140 – 143.
58. N. A. Barcza, T. R. Curr, W. D. Winship, C. P. Heanley, *Electr. Furn. Conf. Proc.* **39** (1981) 243 – 260.
59. *Met. Bull. Mon.* **151** (1983, Jul.) 91 – 93.
60. Ges. f. Elektrometallurgie mbH, DE 2 201 388, 1972 (F. Breuer, K. Brotzmann, G. Duderstadt, R. Fichte et al.); DE 2 531 034, 1975 (F. Breuer, G. Duderstadt, G. Nassauer, W. M. Dresler et al.); DE 2 540 290, 1975 (F. Breuer, G. Duderstadt, W. Dresler, R. Fichte et al.).
61. H. Franke, G. Duderstadt: "Production of Medium Carbon Ferrochromium in a Bottom Blowing Converter and its Application in Stainless Steel Production," *Int. Ferroalloy Conf. (INFACON),* Johannesburg, South Africa, Apr., 1974.
62. Soc. d'Electrochimie d'Electrométallurgie et Aciéries Electriques d'Ugine, US 2 100 265, 1932 (P. Perrin).

63. Electric Furnace Products Co., DE 860 073, 1950 (H. de Wet Erasmus, H. R. Spendelow, Jr.).
64. C. G. Chadwick, *J. Met.* **13** (1961) 806 – 808.
65. J. Krüger in O. Winkler, R. Bakish (eds.): *Vacuum Metallurgy,* Elsevier, Amsterdam-London-New York 1971, pp. 265 – 268.
66. Union Carbide Corp., DE 1 302 764, 1961 (F. D. Hamilton, Jr., O. B. Chamberlain, Jr.).
67. Gesellschaft für Elektrometallurgie mbH, DE 1 558 500, 1967 (R. Fichte, H. Franke, H.-J. Retelsdorf).
68. V. P. Nemchenko et al., *Stal'* 1982, no. 7, 37 – 39 (Engl. Transl. from Russ. *BISI* 21 593).
69. GfE-Lieferprogramm, 1974.
70. Reading Alloys, US 4 331 475, 1982 (F. H. Perfect).
71. R. Hähn, H. Andörfer, H.-J. Retelsdorf, *Erzmetall* **36** (1983) 459 – 465.
72. Shieldalloy Corp., Product Specifications MA-104, MA-106, Newfield, N.J., 1979.
73. GfE Legierungen – Legierungsmetalle, Gesellschaft für Elektrometallurgie mbH, Düsseldorf, Mar., 1980.
74. DIN 17 657, 1973.
75. DIN 1725 Blatt 3, 1973.
76. International Standard ISO 5448, Ferrochromium-Specification and Conditions of Delivery, International Organization for Standardization, 1981.
77. International Standard ISO 5449, 1980.
78. ASTM Designation, A101-73, 1973.
79. ASTM Designation, A482.76, 1976.
80. DIN 17 565, 1968.
81. H.-J. Fischer, K.-D. Frank, *Erzmetall* **35** (1982) 403 – 409.
82. International Standard ISO 3713, Draft Proposal, Ferroalloys Sampling and Preparation of Samples, General Rules, ed. Dec., 1985.
83. Draft International Standard ISO/DIS 4140, Ferrochromium and Ferrosilicochromium – Determination of Chromium Content – Potentiometric Method, International Organization for Standardization, 1978.
84. Draft International Standard ISO/DIS 4158, Ferrosilicon, Ferrosilicomanganese and Ferrosilicochromium – Determination of Silicon Content – Gravimetric, International Organization for Standardization, 1977.
85. W.-H. Grebe, H. Kästner, C. Kippenberger, U. Krauss et al.: *Untersuchungen über Angebot und Nachfrage Mineralischer Rohstoffe, VII Chrom,* Bundesanstalt für Geowissenschaften &Rohstoffe, Hannover; Deutsches Institut für Wirtschaftsforschung, Berlin 1975, 147 pp.
86. P. Wille, *Stahl Eisen* **103** (1983) 127 – 132.
87. B. M. Coope, "Chromite," *Mining Annual Review* 1984, 65 – 68, published by *Min. J. (London)* 1984 (Jun.).
88. *Chromium Review* (1983, Apr.) no. 1, 13.
89. D. P. O'Shaughnessy, *J. Met.* **37** (1985, Jul.) 57 – 58.
90. *Erzmetall* **37** (1984) 478 – 479.
91. *Met. Bull. Mon.* **131** (1981, Nov.) 59 – 63.
92. J. W. Scott, *Electr. Furn. Conf. Proc.* **29** (1971) 80 – 82.
93. H.-J. Retelsdorf, E. Hodapp, N. Endell, *Electr. Furn. Conf. Proc.* **27** (1969) 109 – 114.
94. D. Liesegang, *Erzmetall* **37** (1984) 18 – 21.
95. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz, TA-Luft (Technische Anleitung zur Reinhaltung der Luft) Feb. 27, 1986.
96. D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe, Chrom und seine Verbindungen,* Verlag Chemie, Weinheim 1983, Part 9.
97. R. A. Person, *Electr. Furn. Conf. Proc.* **33** (1975) 39 – 46.
98. R. Wintersteen, *Electr. Furn. Conf. Proc.* **39** (1981) 221 – 225.
99. ACGIH (ed.): *Threshold Limit Values (TLV) for Chemical Substances in the Work Environment,* 1985 – 1986.

Chromium Compounds

Gerd Anger, Leverkusen, Federal Republic of Germany (Chap. 2)

Jost Halstenberg, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 3 – 4.1, 4.3 – 8, and 14)

Klaus Hochgeschwender, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 9 – 13)

Christoph Scherhag, Bayer AG, Uerdingen, Federal Republic of Germany (Chaps. 9 – 13)

Ulrich Korallus, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 13)

Herbert Knopf, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 3 – 4.1, 4.3 – 7, and 14)

Peter Schmidt, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 3 – 4.1, 4.3 – 7, and 14)

Manfred Ohlinger, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 4.2 and 13)

1. Introduction 2
2. Chromium Ores 2
2.1. Ore Deposits 4
2.2. Ore Beneficiation 7
3. Production of Sodium Dichromate . . 7
3.1. Alkaline Roasting 7
3.2. Leaching of the Roast 8
3.3. Acidification 10
3.4. Crystallization 11
4. Chromium Oxides 11
4.1. Chromium(III) Oxide and Chromium Hydroxide 11
4.2. Chromium(IV) Oxide (Chromium Dioxide) 13
4.3. Chromium(VI) Oxide 14
5. Chromium(III) Salts 16
5.1. General Properties 16
5.2. Chromium(III) Sulfates and Chrome Tanning Agents 17
5.3. Other Chromium(III) Salts 19
6. Chromic Acids and Chromates(VI) . 21
6.1. Chromic Acids 21
6.2. Alkali Chromates and Dichromates . 22
6.3. Other Chromates 24
7. Other Chromium Compounds 26
8. Analysis 26
9. Transportation, Storage, and Handling 27
10. Environmental Protection 27
11. Ecotoxicology 28
12. Nutrition 28
13. Toxicology and Occupational Health . 29
14. Economic Aspects 32
15. References 33

1. Introduction

Historical. Chrome iron ore (chromite) was discovered in 1798. A few decades later this ore was being subjected to oxidative roasting in the presence of soda and lime in manually operated furnaces to produce water-soluble sodium dichromate. This was processed further to yield yellow, red, and green chromium pigments which were used, among other things, for dyeing wallpaper; they replaced the toxic arsenic dyes that had been used until then. Chromium salts soon found their way into the textile industry as mordants for the dyeing of wool.

The importance of dichromates increased considerably in the period following 1870 when the rising coal tar dye industry needed large quantities for the oxidation of chemical intermediates. With the advent of the 20th century, chrome tanning was introduced in leather factories and in many areas replaced vegetable tanning.

The manufacture of chromium compounds received a further boost after 1930, when metallic chromium was successfully precipitated from chromic acid solutions by special additives. Since then this possibility has been used extensively in electroplating for bright and hard chromium plating.

Chromium compounds are used in numerous fields. In addition to the applications mentioned, chromates have long been used in printing as an aid in photomechanical reproduction. For some time, chromium dioxide has been a component

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_067

of magnetic tapes for information storage. Table 1 lists important applications of chromium chemicals.

Table 1. Uses of chromium compounds

Branch of industry	Product	Use
Building industry	chromium(III) oxide	pigment for coloring building materials
Chemical industry	dichromates, chromium(VI) oxide	oxidation of organic compounds, bleaching of montan waxes, manufacture of chromium complex dyes
	chromium(III) oxide	catalysts
Printing industry	dichromates	photomechanical reproduction processes
	chromium(VI) oxide	chromium plating of printing cylinders
Petroleum industry	chromates(VI)	corrosion protection
Paints and lacquers	chromates, chromium(III) oxide	pigments
Refractory industry	chromium(III) oxide	additive for increasing slag resistance
Electro-plating	chromium(VI) oxide	bright and hard chromium plating
Wood industry	chromates, chromium(VI) oxide	in mixtures of salts for protecting wood against fungi and insects
Leather industry	basic chromium(III) sulfates	tanning of smoothed skins
Metal industry	chromium boride, chromium carbide	flame sprays
	chromium(III) oxide	polishing agents
Metallurgy	chromium(III) oxide	aluminothermic extraction of pure chromium metal
Textile industry	dichromates	dyeing with chrome dyes
	basic chromium(III) acetates and chromium(III) fluorides	mordanting of textiles
Recording industry	chromium(VI) oxide	magnetic information storage
Pyrotechnics industry	dichromates	additive to igniting mixtures

2. Chromium Ores [4–16]

The distribution of chromium in terrestrial rocks is closely linked to magmatic intrusions and their crystallization. The average content in the ten-mile crust of the earth is 100 ppm of chromium [10]. Table 2 contains a worldwide estimate of chromium ore resources.

The most important applications of chromium ores are in the manufacture of stainless steel, grey cast iron, iron-free high-temperature alloys, and chromium plating for surface protection. In the nonmetallic mineral industry, chromite is processed in conjunction with magnesite (sintered magnesia, calcined magnesia) and binders (clay, lime, gypsum, bauxite, corundum). The products are intended to have good resistance to pressure, fire, and temperature change, as well as good insulating properties between basic and acidic masonry. The chemical industry uses chromium ores in the production of chromium compounds (Chaps. 34567). Table 3 shows quality requirements of chromium ores for different areas of application.

Minerals. Of the many minerals that contain chromium only the chromium spinels are of economic importance. The formula for the series of isomorphous mixtures of chromium spinels that form geological deposits is

$$(Fe^{II}, Mg)O \cdot (Cr, Al, Fe^{III})_2O_3$$

The proportion of Cr_2O_3 in the chromium spinels varies widely, causing the Cr : Fe ratio (also known as the Cr – Fe factor) to vary as well; this can have a profound effect on the evaluation of a deposit. In an ideal chromium spinel ($FeO \cdot Cr_2O_3$; 67.8 % Cr_2O_3, 32.2 % FeO) the chromium : iron ratio is 2. As a result of the isomorphous inclusion of MgO, the Cr : Fe ratio may rise to between 2.5 and 5. Figure 1 shows the region of isomorphism with varying composition of the spinels. Natural chromium spinels usually contain 33 – 55 % Cr_2O_3, 0 – 30 % Fe_2O_3, 0 –30 % Al_2O_3, 6 – 18 % FeO, and 10 – 32 % MgO. Table 4 lists some physical properties of chromium spinels.

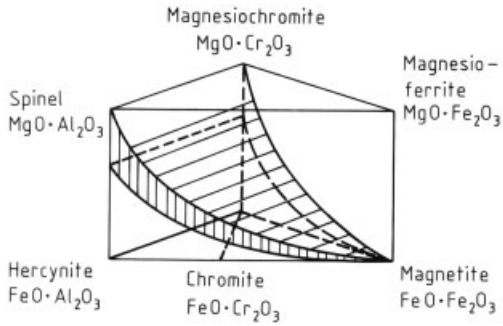

Figure 1. Ternary spinel system showing main isomorphous region

Chromium also occurs in all groups of silicates where chromium replaces Al^{3+}, Fe^{3+}, and Mg^{2+}. Sulfidic chromium ores do not occur on earth. Chromates and chromium iodates

Table 2. Estimated reserves of chromium ore [70]

	Reserves, 10^6 t			
	Total	Metallurgical [a], >45 % Cr_2O_3	Chemical [a], >40 % Cr_2O_3	Refractory [a], >20 % Al_2O_3
Republic of South Africa	2000 [b]	100 (5 %)	1900 (95 %)	–
Zimbabwe	600	300 (50 %)	300 (50 %)	–
Turkey	10	9 (90 %)	–	1 (10 %)
Philippines	7.5	1.5 (20 %)	–	6 (80 %)
United States	8	0.4 (5 %)	7.4 (92.5 %)	0.2 (2.5 %)
Canada	5	–	0.5 (100 %)	–
Finland	7.5	–	7.5 (100 %)	–
Others	11.35	8.175 (72 %)	0.2 (2 %)	2.975 (26 %)
Total	2649.35	419.075 (16 %)	2220.1 (84 %)	10.175 (0.4 %)
Former USSR and other Eastern bloc countries	51.5	26.5 (51 %)	15 (29 %)	10 (20 %)
Total worldwide (rounded off)	2701	446 (17 %)	2235 (83 %)	20 (1 %)

[a] Graded according to Cr_2O_3 or Al_2O_3 contents.
[b] Ores containing 30–50 % Cr_2O_3.

Table 3. Quality requirements (mass fractions in %) for chromium ores (according to U.S. Bureau of Mines)

	Metallurgical [a] (high-chromium chromite)	Refractory [b] (high-aluminum chromite)	Chemical [c] (high-iron chromite)
Cr : Fe ratio	3 : 1 or higher	–	–
Cr_2O_3	>48	>31	>44
$Cr_2O_3 + Al_2O_3$	–	>58	–
Fe	–	<12	–
SiO_2	<8	<6	<5
S	<0.08	–	–
P	<0.04	–	–
CaO	–	<1	–

[a] Zimbabwe, South Africa, Turkey, former USSR.
[b] Turkey, Philippines.
[c] South Africa, former USSR.

are described, which originate from the weathering zone of sulfidic lead deposits (e.g., crocoite, $PbCrO_4$).

Table 4. Physical properties of chromium spinels

Properties	Notes
Specific density: 3.8–4.8	increases as Fe and Cr contents increase
Hardness (Mohs): 4.5–8	increases with increase in the ferrochromite component, very high for Al spinels
Melting point: 1545–1730 °C	inclusion of Mg raises melting point, inclusion of Fe^{2+} reduces it
Color: dark brown to black	reddish with high Cr_2O_3 content
Streak on porcelain plate; hammer striking mark: brown	important feature for differentiating from serpentine

2.1. Ore Deposits

Chromium ore deposits can be divided into two genetically different types:

1) Seam-like deposits, also called stratiform or anorogenic deposits. Main representatives are Bushveld, Great Dyke, and Stillwater.
2) Deposits which are shaped like sacks or tubes; they are called podiform or orogenic deposits. Main representatives are Selukwe, Guleman, and Tiébaghi.

Various intermediate types such as adjacent "seam" pockets, striated chromite slabs, mottled ores, and vein-like deposits also occur. "Placer" deposits, i.e., enrichment due to chromite lumps

and grains on or near primary deposits, are now achieving economic importance.

The seam-like deposits reveal layers or strata of chromite enrichment, with thicknesses ranging from centimeters to decimeters; the layers are regularly interlaminated with banded series of olivine-rich or pyroxene-rich rocks. The Main Seam of the Western Bushveld is, for example, 1.10 – 1.30 m thick and can be traced for over 65 km without any significant change in the mineral composition or thickness.

The demarcation between the chromite enrichment and the underlying bed is usually razor-sharp; in the direction of the overlying layer, disintegration into layers or mottled ores as a result of increased silicate content is observed.

The chromite bodies that are sack-like to tube-like in appearance are usually aligned with the direction of the magmatic stratification, i.e., the lowest sections are massive chromite ores; in the direction of the overlying layer, these merge into striated slabs or mottled ores.

The internal texture of the chromite ore bodies varies widely. The closest chromite crystal packing results in the formation of massive ores containing 75 – 85 vol% of chromite. Sphere or leopard ores, which consist of round chromite crystal aggregates 0.5 – 2 cm in diameter in a silicate matrix (olivine, pyroxene, serpentine), are also characteristic. Banded ores are closely related to the massive ores, but they are frequently richer in silicate and then form a link with the mottled ores (chromite single crystals in silicate matrix).

During transformation (serpentinization), the silicate content within the chromite ore bodies has resulted in the formation of friable and pulverizable masses (friable ore) which are encountered not only near the surface but also at depths of several hundred meters below the present-day land surface.

Chromite transformation in the course of more recent tectonic superficial modification under pneumatolytic or hydrothermal conditions has resulted in the striking colors of recent uvarovite, smaragdite, and kammererite formations which act as pathfinders in prospecting and exploring for chromite deposits. Table 5 shows some analyses of selected chromium ores.

Former Soviet Union. The former Soviet Union is one of the most important producers of chromium ore in the world. All the deposits are distributed in ultrabasite massifs in the Central and Southern Urals. The deposits that are most important at present were found in the late 1930s in the Akhtiubinsk region (North Kazakhstan). The Donskoye deposit, which is associated with the mining settlement of Khrom Tau, contains high-grade chromium ores for ferrochromium production and low-grade ores for chemical purposes. Mining is carried on in numerous open-pit mines, which implies that the ore bodies are small. A new open-cast mine was put into production near Donskoye, as is a processing plant with a throughput of 10^6 t/a. Strong prospection effort for new occurrences is being made in the Northern Urals, but because of the rough climate no mine has been opened up to 1986.

Bushveld. In the Bushveld (Republic of South Africa) mining began in the 1920s in two districts: the Lydenburg district (Eastern Bushveld) and the Rustenburg district (Western Bushveld).

From a geological and petrological point of view this is a large intrusion of 500×250 km with a thickness of over 5 km. The chromite "seams" are located in the pyroxenite – norite zone of the basal section of the intrusion, always below the platinum-bearing Merensky Reef. In the case of Rustenburg there are up to 25 chromite seams on top of each other. The thickness of the individual seams varies from a few centimeters to 1.80 m. The seams are workable from 0.35 m upward, especially if they can be combined into mining units (Cr_2O_3 content in the crude ore 30 – 40 %; Cr : Fe ratio = 1.6 – 2.3). In the Lydenburg district, which is genetically very similar to the Rustenburg district, only two seams are being mined; the Cr_2O_3 content is 44 % and the chromium : iron ratio 1.6 – 1.7. The iron content is frequently high, which may cause difficulties in the case of metallurgical ores; however, these ores are highly valued as chemical ores.

Great Dyke. The Great Dyke (Zimbabwe) is an intrusion which is 610 km long and 6 – 9 km thick – a remarkable length : thickness ratio which is unique in the world. The internal structure is similar to that of the Bushveld. From north to south, the individual complexes are Musengezi, Hartley, Selukwe, and Wedza. Selukwe

Table 5. Chemical analyses (mass fractions in %) of some chromium ores (crude ores, concentrates)

Country	Cr_2O_3	FeO	SiO_2	MgO	Al_2O_3	CaO	V_2O_5	Cr : Fe ratio
South Africa								
Rustenburg (c) [a]	44.5	26.4	3.5	10.6	14.4	n.d. [b]		1.7 : 1
Lydenburg (c)	44.3	24.6	2.3	11.2	16.1	0.4		1.8 : 1
Zimbabwe								
Great Dyke (m) [a]	48.5	18.3	5.6	13.4	11.5	0.8		2.6 : 1
Great Dyke (r) [a]	50.7	16.4	4.3	13.2	13.0	0.8		3.1 : 1
Selukwe (m)	47.0	12.0	5.7	15.5	12.6	1.8		3.9 : 1
Selukwe (r)	42.0	15.7	8.6	15.8	13.8	0.3		2.7 : 1
Turkey								
(m)	48.3	14.1	5.1	16.8	13.0	0.9		3.4 : 1
(r)	37.0	15.2	4.3	17.7	24.3	0.2		2.4 : 1
Philippines								
(Masinloc) (r)	33.3	13.2	4.6	19.6	28.2	0.4		2.5 : 1
Finland (Kemi)								
Crude ore	26.5	15.0	18.5	19.5	9.5	–	0.04 – 0.1	1.8 : 1
Concentrate	45.7	33.8	0.4	2.9	13.6	n.d.	0.1	1.4 : 1
Albania								
(m, r)	43.0	16.2	9.8	22.2	7.9	0.1		2.6 : 1
Former USSR								
(m)	53.9	12.6	5.8	13.3	9.6	1.1		4.3 : 1
(r)	39.1	14.0	9.4	16.1	17.4	0.7		2.8 : 1

[a] (m) = metallurgical, (r) = refractory, (c) = chemical.
[b] n.d. = not determined.

Table 6. Analysis of chromium ores from Zimbabwe

	Cr_2O_3 content, wt %	Cr : Fe ratio	Proportion, wt %
Metallurgical	over 48	over 2.8	80
Chemical	45 – 48	2.2 – 2.5	17
Refractory	42 – 46	1.8 – 2.0	3

consists of sack-like deposits containing 48 % Cr_2O_3 and even more, with a chromium : iron ratio greater than 2.8 (a highly valued metallurgical ore). In the Hartley region, on the other hand, numerous bands and seams 2 – 75 cm thick are being mined; these are separated by serpentinized peridotite layers, some of which are very thick and make mining very difficult. However, the Cr_2O_3 content varies between 48 and 57 %, and the chromium : iron ratio is over 2.8 (Table 6).

According to conservative estimates, 1 km^2 of the Great Dyke contains around 1.4×10^6 t of crude ore, which corresponds to assured reserves of 600×10^6 t (geologically 4.6×10^9 t are possible).

Other Ore Deposits

Madagascar. On the island of Madagascar, chromium ores are being mined since 1967 with an annual production of around 60 000 t of metallurgical grade ore. Total output is calculated to be almost 2×10^6 t since the beginning of the operation (50 – 52 % Cr_2O_3). The reserves are said to be around 5.5×10^6 t.

Turkey. Turkey still is the traditional country for chromite deposits of metallurgical quality, but because of falling prices on the world market and exhaustion of reserves, many mines have been forced to close. The most important regions belong essentially to the alpidic era, e.g., Bursa, Mugla district, and Elazig, including the Guleman chrome ore field. Open-pit mining, and

in some places underground mining at shallow depths, are employed.

Iran. Chromite deposits are described in two regions of Iran: northwest of Sabzawar near Mashad and 200 km northeast of the Gulf port of Bandar Abbas. These deposits are pocket-like, sometimes containing only 500 t of ore. Extraction is by open-pit mining and by primitive underground mining. Only hard lump ore for metallurgical applications is sometimes exported.

Philippines. On the Philippine island of Luzon the most important chromite deposits are to be found in the Coto region (near Masinloc, province of Zambales). These are chromite seams and pockets within layered dunites and harzburgites. They are classical metallurgical and refractory ores. More recently a new type of chrome ore has been put into production: chromite from lateritic soils. The concentrates are suited for the chemical industry.

Finland. In 1959 a fairly large deposit of chromite was discovered near Kemi, which has been developed into a productive mine. Chromium ore occurs in a serpentinite – anorthosite massif 12 km long and 1 – 2 km wide; the ore zone, however, is only 15 – 100 m thick and dips at an angle of 60° toward the north. The Cr_2O_3 content of the crude ore of the various ore bodies varies between 17.5 and 21.9 % (locally even up to 30.5 % Cr_2O_3); the chromium : iron ratio is low (0.81 – 1.87).

Former Yugoslavia. The deposits of chromite in former Yugoslavia are restricted to the Raduša massif near Skopje, but the mining of metallurgical ores there has fallen considerably. Native ores are being processed in a new plant, whereas ores imported from Albania are being processed in the old one. The chromium ores are always associated with serpentinized ultrabasites. The striated slab type predominates, but massive chrome ores are encountered in some places.

Albania. Since 1960, Albania has become the third largest producer in the world. All actual data are based on estimates because the Albanian government withholds production and export figures. Albanian deposits belong to the podiform type and normal grades are reported to be 43 % Cr_2O_3 with a Cr : Fe ratio of 3 : 1. The largest chrome ore mines are Bulquize and Matanesh with concentration plants of 300 000 t/a each.

2.2. Ore Beneficiation

The simplest method of concentrating chromite is by hand picking; this is still employed today at many pits, including those in Turkey, Brazil, Iran, and the Philippines. Because the mining of richer ores continues to decline, concentrating procedures, chiefly using the gravity method, have been developed to separate the serpentine from the chromite. For example, in South Africa or Brazil the chromite ores are enriched by crushing, milling, screening, and sophisticated gravity procedures. In South Africa, spirals and diamond pans are standard equipment. A combination of Reichert cones and Reichert spirals has also been employed. Although the costs are higher, the use of hydrocyclones for separating the fine chromite grains from waste is of great importance in the recycling of tailing dumps. Some chromium ores contain magnetite which can be removed by means of magnetic separation. However, if the magnetite is present as an individual phase within the chromite grains or as a fringe around the grains, this method is only suitable if Cr_2O_3 is further enriched. Flotation and electrostatic processes have so far enjoyed little success in the concentration of chromium ores. If the Cr_2O_3 content or the chromium : iron ratio is sufficient, fine-grained chromite concentrates can be briquetted or pelleted with the aid of binders.

The yield (65 – 85 % of the chromite actually contained in the crude ore) depends on many factors including the nature of the chromite – serpentine intergrowth, grain size, and Cr_2O_3 content of the ore or individual grain.

At the chromite concentration plant at Kemi in Finland a fraction of the crude ore (70 %) is crushed to below 10 mm in a primary crusher plant at the open-pit mine. After further grinding (rod and ball mill) and removal of sludge, the intermediate product is dried in a rotary kiln. The magnetic separation (a combination of weak and strong fields) produces two concentrates: concentrate 1 containing 45.9 % Cr_2O_3, which is

sold or used as molding sand, and concentrate 2 containing 42.0 % Cr_2O_3 for the production of ferrochromium.

3. Production of Sodium Dichromate

Directly or via several intermediate stages, sodium dichromate [*7789-12-0*], $Na_2Cr_2O_7 \cdot 2\,H_2O$, is the starting material for the production of all chromium compounds and pure chromium metal.

Sodium dichromate is made in a three-step process: (1) alkaline roast of chromite under oxidizing conditions (Eq. 1), (2) leaching, and (3) conversion of sodium monochromate to sodium dichromate by means of an acid (Eq. 2).

$$4\,FeO \cdot Cr_2O_3 + 8\,Na_2CO_3 + 7\,O_2 \xrightarrow{\text{ca. } 1000\,^\circ C} 8\,Na_2CrO_4 + 2\,Fe_2O_3 + 8\,CO_2 \quad (1)$$

$$2\,Na_2CrO_4 + 2\,H^+ \longrightarrow Na_2Cr_2O_7 + H_2O + 2\,Na^+ \quad (2)$$

3.1. Alkaline Roasting

Soda ash (sodium carbonate) is generally used as the alkali component but sodium hydroxide may also be employed [17–19]. The degree of solubilization of chromites by the roasting process depends on their composition. For optimum results, the process is controlled by adding so-called carrier materials. These ensure sufficient porosity of the material so that oxygen can diffuse into the roast. Porosity is maintained by means of such materials as iron oxide, bauxite, or dried leach residue; CO_2-emitting additives include lime and/or dolomite. The inert additives dilute the sodium carbonate and sodium chromate, which both melt at the reaction temperature. In the *low-lime process* the carbonates evolve CO_2, decrease the reaction temperature to below 1000 °C, and raise the melting point of the reaction products; the amount of lime added must be controlled so that the compound $5\,Na_2CrO_4 \cdot CaCrO_4$ [20] is produced in the roast. Temperatures above 1150 °C must be avoided because they result in the subsidiary components of the ore being attacked. At still higher temperatures the degree of conversion is markedly decreased. The optimum temperature range is very narrow and depends strongly on the type of ore used and the composition of the mixture.

Process Description. A typical roast mixture contains 100 parts of ore, 60 – 75 parts of sodium carbonate, 0 – 100 parts of lime or dolomite, and 50 – 200 parts of inert materials. The components are first finely ground, then mixed, and fed into the furnace. Annular hearth furnaces or rotary kilns are commonly used in large plants today.

The *annular hearth furnace* (Fig. 2) is made from steel with a refractory lining (inner diameter ca. 20 m; outer diameter ca. 30 m); it is driven by a gear wheel underneath and has rails running on rollers. The rotating hearth is sealed from the stationary parts of the furnace by sheets of metal dipping into annular water troughs. The furnace is heated by several burners from the side or from the top with gas, coal dust, or oil. The exhaust gas can be utilized to preheat the burner air or to generate steam. The mixture is fed to the outer edge of the annular hearth by a feed screw. A water-cooled ribbon screw transports the mixture inward, each time the annular hearth revolves, and finally removes it in the middle.

In the annular hearth furnace, the mixture is uniformly heated to the reaction temperature and made to travel toward the center of the hearth with a well-defined layer thickness. The furnace process is fairly independent of the sintering of the roast; it allows the production of melts that contain 40 wt % of water-soluble sodium chromate. The yield is 80 – 95 %, based on the chromite feed. The roast takes 2 – 6 h, depending on the composition of the mix.

Most of the kiln tube of the *rotary kiln* (Fig. 3) between the feed point and the reaction zone is used to heat the mixture. Shortly before the mix reaches the actual reaction zone, the soda melts and calcines. At this point the mixture bakes, and pellets or wreath-shaped cakes may be formed. If the furnace is operated inexpertly (temperature too high) or the composition of the mixture is wrong (too little carrier material) the kiln tube may get substantially clogged. In such cases, the constriction can be cleared by using an industrial gun.

The roast from the rotary kiln contains up to 30 wt % of water-soluble sodium chromate. The yield is 75 – 90 %, based on the chromite feed.

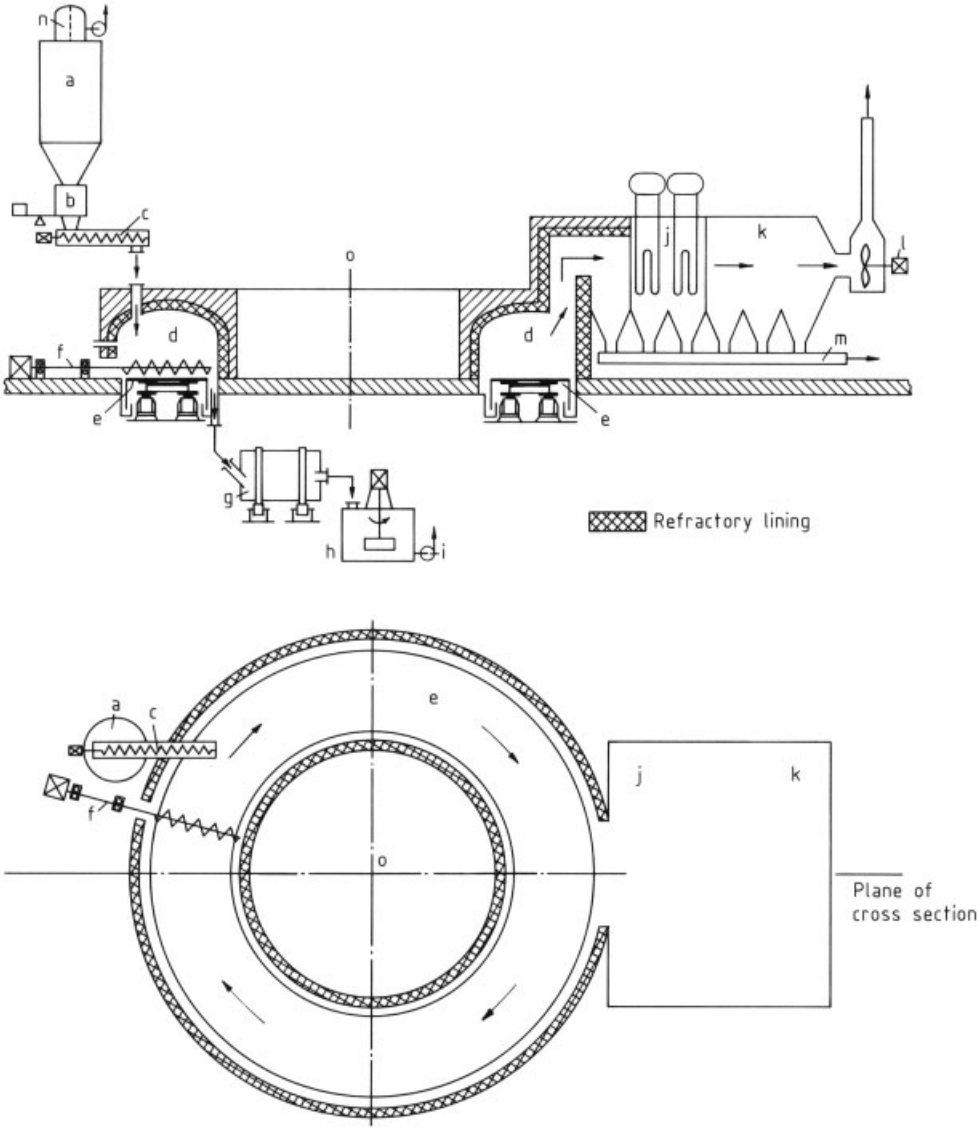

Figure 2. Annular hearth furnace
a) Mixture silo; b) Scales; c) Feed screw; d) Furnace; e) Annular hearth; f) Water-cooled ribbon screw; g) Wet tube mill; h) Stirred vessel; i) Pump for filtering system; j) Waste heat boiler; k) Electrostatic gas purification; l) Exhaust gas fan; m) Dust drag chain; n) Bin filter; o) Rotation axis

The roast takes 3 – 8 h, depending on the composition of the mixture. The hot exhaust gas from the rotary kiln can be used to preheat the burner air.

Gas Purification. Exhaust gas purification systems able to achieve a high degree of separation are required for dust collection. They essentially consist of two components: (1) an exhaust gas cooling system, with optional energy recovery, for example, steam generation; and (2) the exhaust gas purification system, usually an electrostatic separator. The size and the design of the purification system depend on the type of furnace. In the rotary kiln, ca. 10 % of the feed mixture is carried off by the exhaust gas, whereas in the annular hearth furnace, less than

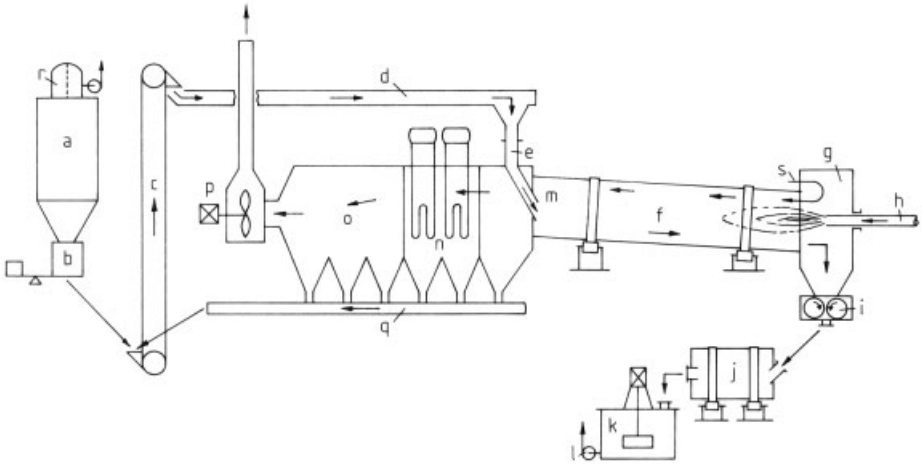

Figure 3. Rotary kiln for roasting chromium ores
a) Mixture silo; b) Scales; c) Elevator; d) Drag chain; e) Inlet tube; f) Rotary kiln; g) Combustion chamber; h) Burner; i) Crusher; j) Wet tube mill; k) Stirred vessel; l) Pump for filtering system; m) Kiln inlet; n) Waste heat boiler; o) Electrostatic gas purification; p) Exhaust gas fan; q) Dust drag chain; r) Bin filter; s) Air inlet

1 % is carried off. However, operation of the annular hearth furnace necessitates a considerable expenditure on gas cooling.

Other Processes. In the literature, other processes are proposed but so far these have not achieved any industrial significance. Thus, attempts have been made to roast chromium ore in a shaft furnace [21] or in a fluidized-bed reactor [22]. A fundamentally different process involves the reaction of the chromium ore and soda in a molten salt mixture with oxygen-containing gases being injected [23].

3.2. Leaching of the Roast

After the oxidative process, the roast is a mixture of soluble salts and insoluble components. It contains sodium chromate, sodium aluminate, magnesium oxide, sodium vanadate(V), iron(III) oxide, unused alkali, unchanged chromite, and small amounts of sodium chloride originating from the soda.

When the roast is extracted with hot water, a pH of 10.5 – 11.2 results. The pH is controlled by adding acids or carbonates so that all chromate dissolves, whereas the alkali-soluble impurities hydrolyze and form a readily filterable precipitate along with the iron hydroxide and the unchanged ore components.

The roast is first cooled on a Fuller grate or in a cooling drum. Then it is either ground in a wet tube mill after addition of water or wash solution (see below) with carbonates or acids added, or it is dissolved in a stirred vessel. The insoluble residue is separated from the sodium chromate solution and thoroughly washed with a countercurrent of water. Continuous multistage Dorr plants or rotary filters are used; after separation, the insoluble residue is extracted two to three more times in counterflow. Dorr plants only exhibit satisfactory separation of residue and solution if the sodium chromate concentration is not too high. Rotary filter plants can be employed without difficulty for nearly saturated hot sodium chromate solutions; such filters are frequently preferred because the higher water consumption of the Dorr plant results in unnecessary steam costs in the subsequent evaporation process.

After removal of residual aluminum hydroxide and other undissolved components in a final purification process (e.g., thickener), the concentrated sodium chromate solution is acidified (Section 3.3).

Some of the filter residue is dried and added to the roasting mix (Section 3.1). The remainder is subjected to reducing treatment to convert the residual chromate content into an ecologically harmless form. To do this, the residue

is suspended in water and treated with sulfuric acid and sodium hydrogen sulfite or iron(II) sulfate; chromate residues are converted into chromium(III) compounds in this way. Subsequent addition of alkali precipitates trivalent chromium (and iron(III), if present) as hydroxide. The suspension is then filtered and the cake (optionally after further removal of water) is dumped.

3.3. Acidification

The sodium chromate solution is converted into sodium dichromate solution by acidification with sulfuric acid (Eq. 3) or carbon dioxide (Eq. 4). The sequence of individual steps depends on the acid.

$$2\,Na_2CrO_4 + H_2SO_4 \longrightarrow Na_2Cr_2O_7 + Na_2SO_4 + H_2O \quad (3)$$

$$2\,Na_2CrO_4 + 2\,CO_2 + H_2O \rightleftharpoons Na_2Cr_2O_7 + 2\,NaHCO_3 \quad (4)$$

Sulfuric Acid Acidification. Sulfuric acid is added to the concentrated sodium monochromate solution in an agitated vessel until the pH is about 4. The sodium dichromate solution is then concentrated in a continuous evaporation plant. Each liter of sodium dichromate solution yields 400 – 500 g of anhydrous crystalline sodium sulfate. The sulfate is removed by centrifugation. The clear, dark-red sodium dichromate solution contains 900 – 1200 g of $Na_2Cr_2O_7 \cdot 2\,H_2O$ per liter and additional small amounts of sodium sulfate; it is dispatched in tanks, e.g., of steel. The solution is either used directly as an oxidizing agent or processed to yield dichromate crystals.

Carbon Dioxide Acidification. After filtration the sodium monochromate solution from the first filtration stage is concentrated to ca. 850 g of Na_2CrO_4 per liter. The saturated sodium chromate solution is then acidified with a countercurrent of carbon dioxide at 0.5 – 1.5 MPa (5 – 10 bar) to yield sodium dichromate and sodium hydrogen carbonate. A series of stirred autoclaves is preferred for this reaction. They must be cooled to remove the heat of neutralization; the slurry leaves the last reactor at room temperature. The degree of conversion is about 80 – 90 %.

The sodium hydrogen carbonate is removed by centrifugation or filtration, preferably under pressure to prevent reaction (4) from being reversed. After it has been washed, water can be removed in a pusher centrifuge. Still moist, the sodium hydrogen carbonate is then transferred to a calcining furnace. The carbon dioxide produced in the furnace may be fed back to the acidification autoclaves [2]; the sodium carbonate obtained is recycled for alkaline ore roasting [24].

To obtain commercial sodium dichromate solution, further evaporation to a concentration of 1000 g of sodium dichromate per liter is required. This is then followed by a second acidification with carbon dioxide or sulfuric acid.

Production of soda is the main advantage of carbon dioxide acidification. However, this is offset by a number of difficulties, particularly the formation of deposits on the evaporator during the concentration of the sodium monochromate solution. Other problems include mastering the pressure technology and cooling, separation of the sodium hydrogen carbonate, yield loss due to reverse reaction, and clogging of the calcination furnace.

3.4. Crystallization

For the purpose of crystallization the sodium dichromate solution (950 – 1200 g/L) is further concentrated and may, if necessary, be filtered while hot to remove additional sodium sulfate or sodium chromate. It is then slowly cooled to 30 – 35 °C with constant stirring to obtain orange-red crystals of $Na_2Cr_2O_7 \cdot 2\,H_2O$.

Today continuous vacuum crystallization is carried out to an increasing extent. The initial difficulties of this process, particularly in obtaining coarse, dust-free crystals, have largely been overcome. The crystalline slurry is continuously separated from the mother liquor and dried. Precise control of the drying temperature is important because hydrated sodium dichromate is converted into anhydrous sodium dichromate above 84.6 °C and, therefore, cakes if overheated.

For reasons of occupational health, care should be taken to ensure that workplaces and production plants are dust free when sodium dichromate, especially the dried product, is being handled. In such locations, extensive venti-

lation and dust removal systems (wet scrubbers, electrostatic separators) are necessary.

4. Chromium Oxides

4.1. Chromium(III) Oxide and Chromium Hydroxide

Chromium(III) oxide [*1308-38-9*], Cr_2O_3, M_r 151.99, ϱ 5.2 g/cm^3, is green in finely dispersed form, whereas fairly large crystals have a blackish green hue and a metallic luster. The crystals have a hexagonal rhombohedral structure of the corundum type. The compound melts at 2435 °C, but begins to evaporate at 2000 °C to form clouds of green smoke; the boiling point is estimated to be 3000 – 4000 °C. The enthalpy of formation is − 1141 kJ/mol. Macrocrystalline chromium(III) oxide has a hardness of 9 on the Mohs scale. An amorphous form of the oxide is also known; this crystallizes on heating.

Chromium(III) oxide does not dissolve in water, acid, alkali, or alcohols. It is converted by a molten bath of sodium peroxide into soluble sodium monochromate(VI). Chromium(III) oxide and chromates(III) are used in organic chemistry as catalysts, e.g., in the hydrogenation of esters or aldehydes to form alcohols and in the cyclization of hydrocarbons. They also catalyze the formation of ammonia from hydrogen and nitrogen.

Production. The industrial production of chromium(III) oxide involves the reduction of solid sodium dichromate, generally with sulfur. The finely divided components are thoroughly mixed, fed into a brick-lined furnace, and brought to dark-red heat. The reaction proceeds exothermically. After the reaction mass has cooled, it is broken up and the sodium sulfate produced is leached out with water. The remaining solid is separated, rinsed, dried, and ground. To obtain 100 kg of chromium(III) oxide, 200 kg of sodium dichromate must react with at least 22 kg of sulfur; usually an excess of sulfur is used. Additives such as ammonium chloride or starch in the crude mixture affect the pigment properties. When sodium dichromate is replaced by the corresponding potassium salt, the hue of the pigment becomes more bluish.

The compound can also be prepared by a wet route involving reduction of sodium chromate by sulfur [25], with sodium thiosulfate being produced as a coproduct. The hydrate initially obtained is washed by decanting, filtered, and calcined to form the oxide.

Chromium(III) oxide destined for aluminothermic production of pure chromium metal must be heated additionally at 1000 °C to increase its grain size. If products particularly low in sulfur are to be produced for this purpose, charcoal can be used for the reduction instead of sulfur. High-purity oxides can also be obtained by thermal decomposition of chromium(VI) oxide or ammonium dichromate(VI), the latter yielding a material of very low density.

Chromium (III) oxide pigments contain 99.1 – 99.5 % Cr_2O_3. The aluminum oxide and silicon dioxide impurities each amount to ca. 0.1 %; the annealing loss at 1000 °C is about 0.3 %. The individual particles are spherical, with a diameter of 0.3 µm predominating. Chromium(III) oxide finds widespread application as a green pigment resistant to atmospheric conditions and heat. In addition, it is used as a colorant in glass products and printing inks, as a vitrifiable pigment in the ceramics industry, and as a polishing agent because of its considerable hardness (→ Pigments, Inorganic).

Chromium(III) Aquoxides [*12292-46-5*], [*12182-82-0*]. Pure chromium(III) hydroxide [*1308-14-1*], $Cr(OH)_3$, M_r 103.02, can only be prepared with difficulty because the hydrates initially obtained by precipitation are subject to aging.

After drying in air, specimens prepared by precipitation with alkali in the cold from violet chromium(III) salt solutions have a composition corresponding to $Cr_2O_3 \cdot 9\,H_2O$ [*12182-81-9*], and are usually formulated as $Cr(H_2O)_3(OH)_3$ [*41993-26-4*]. They are bright bluish green powders with limited life. All three hydroxyl groups react immediately with acids. Upon careful heating, dehydration occurs in steps and compounds containing 8, 5, 3, and 1 mol of water are formed. The density increases as the water content falls. Above 50 °C, conversion to a gelatinous green aging product occurs, and the solubility and chemical reactivity decrease; oxygen bridges are formed through the elimination of water, and polynuclear complexes are produced. The composition approaches that of chromium(III) oxide

hydroxide, CrO(OH) [*20770-05-2*]. Aging is accelerated by the presence of hydroxide ions.

In freshly precipitated hydroxides a crystalline phase isomorphic with bayerite [$Al(OH)_3$] is observed (→ Aluminum Oxide), whereas aged compounds are X-ray amorphous. The chromium(III) hydroxide hydrates are amphoteric compounds. With acids they form Cr^{3+} salts, whereas they dissolve in strong hydroxide solution to form chromates(III), e.g., the deep green sodium chromate(III), $Na_2Cr_2O_4$ (previously known as sodium chromite). When ammonium hydroxide is added, red solutions are formed. Oxidizing agents in the presence of alkali produce chromates(VI). With halogens this takes place immediately on gentle heating, but with oxygen several hours are required at a pressure of 4 MPa (40 bar) at 175 °C.

Chromium(III) hydroxide forms a stable colloid solution, whose isoelectric point is at pH 5. At higher pH the sol becomes negatively charged and adsorbs cations, whereas below pH 5 the charge is positive and anions are adsorbed. The adsorption capacity of chromium(III) hydroxide sols is higher than that of aluminum or iron(III) hydroxide sols. Sols containing 127 g of Cr_2O_3 per liter have been obtained from concentrated chromium(III) chloride solution by addition of ammonium carbonate and dialysis while the solution is hot [26].

Production. In industry chromium(III) hydroxide hydrates are usually produced from solutions of chromium(III) sulfate or chromium alum by precipitation with soda, sodium hydroxide solution, or ammonium hydroxide. Production by reduction of sodium chromate with sodium sulfide [27] has also been proposed.

For production from potassium chromium alum 54 kg of soda is dissolved in 300 L of water and a solution of 180 kg of alum in 900 L of water is added slowly. After the evolution of carbon dioxide has subsided, about 220 kg of moist chromium hydroxide containing 12 % Cr_2O_3 is obtained by filtration. Chromium(III) hydroxide hydrates are used for the production of chromium(III) salts by reaction with the corresponding acids.

Hydrated chromium(III) oxide, $Cr_2O_3 \cdot x\,H_2O$, is a brilliant emerald-green pigment known as Guignet's green [*12001-99-9*] that consists of very finely divided chromium(III) oxide to which water is bonded by adsorption. It is produced by heating a ground mixture of one part by weight of potassium dichromate and three parts by weight of boric acid in a muffle furnace to a faint red heat, which results in the formation of chromium(III) and potassium tetraborates. The molten mass still contains 6 – 7 % water and, after cooling, already has a deep green color. When this mass is boiled with water, it decomposes into chromium(III) oxide hydrate and boric acid. The product is coarse-grained and difficult to grind. Use of sodium dichromate as raw material results in a more yellowish color, whereas addition of thiourea or polysulfide to the reaction mixture produces a pigment with a bluish hue. The composition of the commercial products varies; typical values are: Cr_2O_3 79.3 – 82.5 %, H_2O 16.0 – 18.0 %, B_2O_3 1.5 –2.7 %.

Hydrated chromium oxide green is a pigment with properties similar to those of Guignet's green but with a somewhat less intensive coloration. This pigment is prepared by reducing sodium chromate or sodium dichromate in aqueous solution with sulfur or sodium formate in a stirred autoclave or pressure tube [28]. The temperature required is 250 – 270 °C. The solid is separated by filtration, washed, dried, and ground. The finished pigment consists of fine needles, with a particle size of 0.02×0.1 µm predominating. The product contains 79 – 80 % Cr_2O_3, the annealing loss is about 19 %, and the density 3.7 g/cm^3. The coloration changes at elevated temperature. Because of its high reflecting power at infrared wavelengths the product had at times been of special importance in camouflage paints.

4.2. Chromium(IV) Oxide (Chromium Dioxide)

Wöhler discovered ferromagnetic chromium dioxide in 1859 when he decomposed chromyl chloride [29]. About 100 years later, Du Pont produced it in pure form by decomposition of chromic acid under hydrothermal conditions [30], [31]. Industrial exploitation began after the morphological and magnetic properties had been modified by doping chromium dioxide with heavy metals [32], [33] in order to meet the requirements for a magnetic pigment [34]. The marketing of chromium dioxide at the begin-

ning of the seventies initiated the development of cobalt-doped iron oxides which are nowadays used as an alternative to chromium dioxide for information storage.

Physical and Chemical Properties [35]. Chromium dioxide [*12018-01-8*], CrO_2, M_r 84.00, crystallizes in black tetragonal needles. The lattice is of the rutile type and belongs to the space group 4/mmm. The dimensions of the unit cell are $a = b = 44.21$ nm and $c = 29.16$ nm. The X-ray density is 4.89 g/cm^3, and the phase width is between $CrO_{1.89}$ and $CrO_{2.02}$ [36]. The enthalpy of formation is -590 kJ/mol [37]. The temperature coefficient for the *c*-axis is negative [38]. At 100 °C, agglomerated blocks have a linear coefficient of expansion of -6×10^{-6} K^{-1} [39].

At room temperature chromium dioxide is ferromagnetic, the magnetic moment being 2 Bohr magnetons. The Curie temperature is 120 °C and increases to 155 °C as a result of doping with iron [40]. Finely crystalline needle-shaped chromium dioxide has a specific magnetic saturation M_s/ϱ of 77 – 92 A m^2/kg, whereas in single crystals M_s/ϱ rises to 100 A m^2/kg [32]. The magnetocrystalline anisotropy constant is 22×10^3 J/m^3 [41]. The coercivity H^c depends on the size of the crystals and on their shape and magnetocrystalline anisotropies. The coercivity is affected to differing extents by various heavy metals [42]. Iron, antimony, and tellurium increase H^c from 35 kA/m to 60 kA/m, whereas iridium increases it to 220 kA/m [43].

Chromium dioxide behaves as a metallic conductor [35], [44], with a specific electrical resistivity between 2.5×10^{-4} and 4×10^{-2} Ω cm [38], [45].

At room temperature and normal pressure chromium dioxide is metastable; when heated to temperatures above 350 °C, it decomposes into chromium(III) oxide and oxygen. Chromium dioxide has an oxidizing action on reactive organic compounds [46]; the reactivity is considerably decreased by enveloping it with iron(III) and chromium(III) oxides [47–49]. Chromium dioxide is insoluble in water. Reaction with water occurs at the crystal surface, with disproportionation to chromate and Cr^{3+} ions. The aqueous suspension has a pH of 3. However, chromium dioxide is soluble in concentrated sulfuric acid or concentrated alkali solution.

Production [50]. Chromium dioxide is made by decomposition of chromyl chloride, chromic acid anhydride [50], and chromium(III) chromate [51], or by oxidation of chromium(III) compounds with oxygen, hydrogen peroxide, chromic acid anhydride [52], or ammonium perchlorate [43].

Industrial production employs a process originally carried out under licence from Du Pont [42], which involves hydrothermal oxidation of chromium(III) oxide with excess chromic acid:

$$Cr_2O_3 + 3\,CrO_3 \longrightarrow 5\,CrO_2 + O_2$$

Iron(III) oxide and antimony(III) oxide are used as a dopant. Finely divided chromium(III) oxide is obtained either by thermal decomposition of ammonium dichromate or by dehydration of chromium(III) hydroxide [52].

A highly viscous paste (50 – 100 Pa · s) is produced by intensively homogenizing the starting materials. This paste is heated at 300 °C and 35 MPa (350 bar) to form a hard agglomerate of fine chromium dioxide needles which must be drilled out of the reactor trays, broken, and carefully ground. If the residual moisture exceeds 5 %, the product is reheated in a rotary kiln. The chromium dioxide is deagglomerated in an aqueous sodium sulfite suspension, and the crystal surface is simultaneously reduced; this forms a chromium(III) oxide hydroxide layer about 1 nm thick. To do this, the suspension is circulated through a mill which generates intense shear fields and the fine component is removed by using a hydrocyclone. After filtration and washing, drying is carried out in a spray tower. The chromium dioxide obtained has a bulk density of 0.8 g/cm^3.

Production is carried out to a large extent in closed equipment. Less than 2 mg of dust is emitted per 1 m^3 of exhaust air (STP). The chromium-containing wastewater from the production is worked up by reduction.

Use and Economic Importance. Chromium dioxide is used as a magnetic pigment. So far no other uses have achieved any significance [33].

In the audio field, chromium dioxide is used in mono- and multilayer tapes, the layers in the

Table 7. Powder data for typical chromium dioxide pigments

Application	Particle geometry				Magnetic data [d]		
	SSA [a], m^2/g	$\bar{l}$ [b], µm	$\bar{l}/d$ [c]	V [c], 10^{-4} µm	H_c, kA/m	M_r/ϱ [e], $A\,m^2/kg$	M_s/ϱ [f], $A\,m^2/kg$
Audio	28	0.29	9	2.5	41	35	77
Video	35	0.29	11	1.5	49	34	74
Data storage	24	0.32	8	3.5	39	35.5	79

[a] Specific surface area (SSA) determined by N_2 adsorption using the BET method (1-point measurement).
[b] Mean length $\bar{l}$ determined by electron microscope photography with a magnification of 20 000 times.
[c] Diameter d and volume V calculated from SSA and $\bar{l}$.
[d] Measured with a vibration magnetometer, $H_m = 800$ kA/m.
[e] M_r/ϱ = specific remanent magnetization.
[f] M_s/ϱ = specific saturation magnetization.

latter containing chromium dioxide of different coercivity. In video tapes, it is employed either on its own or mixed with cobalt-doped iron oxides. Because of its low Curie point, chromium dioxide allows high-speed thermomagnetic copying of prerecorded audio and video tapes [53]. Chromium dioxide has been used in digital data storage since 1985. Since the magnetostriction of chromium dioxide is low, repeated playing results in virtually no level losses [54]. Table 7 contains data on chromium dioxide powder intended for various applications.

In 1984 the demand for magnetic pigments was about 50 000 t, about 10 % of this being chromium dioxide. The most important manufacturers are Du Pont (United States) and BASF (Federal Republic of Germany). In addition, there is a chromium dioxide plant in the former Soviet Union.

4.3. Chromium(VI) Oxide

Chromium trioxide [*1333-82-0*], chromic acid anhydride, chromic acid, CrO_3, M_r 99.99, ϱ 2.7 g/cm^3, forms dark red crystals which deliquesce in air. The enthalpy of formation is − 594.5 kJ/mol. The oxide melts at 198 °C and starts to decompose, giving off oxygen and brownish red vapors with a pungent smell. The rate of decomposition reaches a maximum at 290 °C, chromium(III) oxide, Cr_2O_3, being produced as the final product via various intermediate stages. Chromium(VI) oxide dissolves in water to form chromic acids (see 6.1); the solubility depends only slightly on temperature. A saturated solution contains 166 g of CrO_3 at 20 °C and 199 g of CrO_3 at 90 °C per 100 mL of water. The compound also dissolves in sulfuric acid and nitric acid. Chromium(VI) oxide is a powerful oxidizing agent, particularly in the presence of acids. Reactions with alkali metals and numerous organic compounds, e.g., low-boiling hydrocarbons, acetone, or benzene and its derivatives, proceed explosively with considerable heat being produced. Esters of chromic acid are also known, e.g., with such cyclic tertiary alcohols as methylfenchol and methylborneol.

Chromium(VI) oxide is made by the reaction of sodium dichromate with sulfuric acid. The reaction can be carried out with solid sodium dichromate or with solutions or suspensions. Both methods are in use industrially. The reaction proceeds rapidly and completely after the components have been mixed, with heat being evolved. Isolation of chromium(VI) oxide from the reaction mixture or purification of the crude product obtained from the aqueous solution is difficult. Quantitative separation of sodium hydrogen sulfate is possible only if the chromium(VI) oxide is melted, but at this temperature the product begins to decompose. The melting process must therefore be controlled very precisely.

Dry Process. Even today the old *discontinuous* process is in use to some extent. The reaction vessels are made from carbon steel or stainless steel. The conical containers are equipped with stirrer, exhaust facilities, and external heating. Sulfuric acid and sodium dichromate are added simultaneously with stirring. The paste heats up to 80 °C during mixing and is heated further

to evaporate water. At 170 °C sodium hydrogen sulfate melts, followed by chromium(VI) oxide at 198 °C. As soon as the reaction products are liquid, the heating and the stirrer are turned off. After a few minutes the heavier chromium(VI) oxide (ϱ 2.2 g/cm^3) settles at the bottom, and is covered by a layer of the lighter sodium hydrogen sulfate (ϱ 2.0 g/cm^3). Liquid chromium(VI) oxide is drawn off the bottom and conveyed to a cooling drum where it solidifies to form scales. Sodium hydrogen sulfate is subsequently drained. The yield of chromium(VI) oxide is about 85 %; 175 kg of sodium dichromate and 140 kg of 96 % sulfuric acid are required to obtain 100 kg of chromium(VI) oxide. About 150 kg of sodium hydrogen sulfate is obtained as byproduct.

Figure 4 shows a *continuous* dry process [55]. The raw materials are fed to a mixing screw for intimate mixing. A viscous paste of chromium(VI) oxide, sodium hydrogen sulfate, and water forms which is fed into a heated rotary kiln of stainless steel where it is melted. The heating must be controlled very carefully. The melt flows into a separator, where the heavier chromium(VI) oxide collects at the bottom of the trough, is removed by means of a rising pipe, and is converted into scales on cooling drums. The upper sodium hydrogen sulfate layer leaves the separating cell via an overflow and is also cooled on drums. Exhaust air from the various pieces of equipment is purified in a wash tower. The yield of this process is over 90 %.

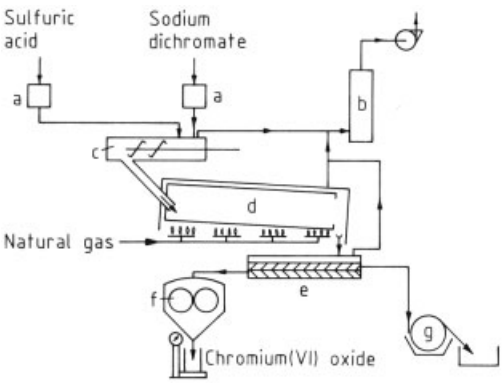

Figure 4. Continuous production of chromium(VI) oxide a) Metering device; b) Wash tower; c) Mixer; d) Rotary kiln; e) Separation cell; f) Cooling drum for chromium(VI) oxide; g) Cooling drum for sodium hydrogen sulfate

If a highly concentrated solution is used instead of dichromate crystals, the kiln can be heated directly [56].

Wet Process. A hot saturated solution of sodium dichromate, which may still contain dichromate crystals, reacts with sulfuric acid [57]. In the course of 30 – 60 min the chromium(VI) oxide precipitates from the hot solution. On filtration a crude product is obtained, with a yield of about 80 %. Sodium hydrogen sulfate must be removed from the crude product by fusion. The filtrate can be recycled for converting sodium monochromate into dichromate; it can also be used in a fresh reaction mixture [58] if sodium hydrogen sulfate is first crystallized and removed at 20 – 25 °C. The crude chromium(VI) oxide is purified by continuous successive fusion and decanting [59].

As an alternative to production by reaction with sulfuric acid, Diamond – Shamrock developed an electrochemical process [60–62] in which chromic acid is produced from sodium dichromate in a two- or three-compartment cell.

Chromium trioxide is usually sold in the form of flakes, but the coarsely or finely ground product is also marketed. Good commercial products contain 99.5 – 99.7 % CrO_3 and a maximum of 0.1 % of sulfate. In the form of flakes, the product has a bulk density of 1.1 kg/L whereas that of the ground product is 1.4 kg/L. Steel drums must be used as containers and they must be tightly sealed because the product absorbs moisture from the air.

Chromium trioxide is classified as a dangerous substance in the EEC list and must be marked as fire-promoting and corrosive. In the IMDG code chromic acid has been put in class 5.1., UN No. 1463. The MAK of CrO_3 is 0.1 mg/m^3.

Electroplating is the most important field of application of chromium(VI) oxide. Numerous mixtures containing chromium trioxide are on the market; these "compounds" often contain hexafluorosilicates which improve the properties of the chromium coatings. Chromic acid solutions are also used for passivating zinc, aluminum, cadmium, and brass. Proprietary mixtures predominantly contain additions of fluoride, nitrate, and phosphate ions. Other uses for chromic acid are in the production of chromium dioxide and in wood preservation.

5. Chromium(III) Salts

5.1. General Properties

Water Content. In contrast to many other inorganic salts, chromium(III) salts occur in a variety of forms that depend on water content and on the particular conditions under which they are formed. Anhydrous compounds do not dissolve in pure water.

However, some of them, e.g., chromium(III) chloride or chromium(III) sulfate, dissolve in the presence of chromium(II) ions. In this process, one dissolved divalent ion transfers an electron via an anion bridge to a trivalent chromium ion in the solid crystal. Having become divalent, this ion detaches itself, acts in a similar manner on another chromium(III) ion in the crystal array, and reverts again to the trivalent state.

Complex Formation. Dissolved chromium(III) ions are always coordinated by various ligands. In the simplest case of the hexaaquochromium(III) ion, $[Cr(H_2O)_6]^{3+}$, six water molecules surround the central chromium ion in an octahedral arrangement as ligands. In addition to the aquo complexes, numerous coordination compounds with other molecules are known, and research on them, particularly studies of the ammine complexes (NH_3 as ligand), has played an important part in the development of coordination compound chemistry [63].

When negatively charged ligands enter the chromium complex, the charge is decreased appropriately. If the sum of the negative charges is four or more, the complex becomes anionic, an example of this being the diaquodisulfatochromium(III) ion $[Cr(SO_4)_2(H_2O)_2]^-$. In this case, each sulfate radical with a double negative charge occupies the position of two ligands.

The tendency of negatively charged ligands to form complex compounds with chromium increases in the following order:

$Cl^- < SO_4^{2-} < CH_3CO_2^- < HCO_2^- < OH^-$

Nitrato complexes are unknown.

Hydrate Isomerism. Chromium(III) complexes exhibit hydrate isomerism due to the positioning of anions and water molecules.

Thus chromium(III) chloride hexahydrate [*10060-12-5*] is known in three different forms:

1) Hexaaquochromium(III) chloride, $[Cr(H_2O)_6]Cl_3$, bluish grey
2) Pentaaquochlorochromium(III) chloride hydrate $[Cr(H_2O)_5Cl]Cl_2 \cdot H_2O$, bright green
3) Tetraaquodichlorochromium(III) chloride dihydrate $[Cr(H_2O)_4Cl_2]Cl \cdot 2\,H_2O$, dark green

The anions that are directly bound to the central atom do not dissociate in water and consequently do not react with the common precipitating agents; therefore, in the two green chlorides, only one-half and one-third of the chloride ions, respectively, are precipitated by silver nitrate.

Basic Salts. Hydroxide ions form coordinate bonds, with central ion of the hexaaquochromium complex being hydrolyzed. In this process, the pentaaquohydroxochromium(III) ion with a double positive charge is first formed:

$$[Cr(H_2O)_6]^{3+} \xrightarrow{OH^-} [Cr(H_2O)_5OH]^{2+} + H^+$$

When more alkali is added, chromium hydroxides are precipitated immediately. Finally, with a strong hydroxide solution, a soluble deep green hydroxo salt is produced:

$$Cr(OH)_3 + 3\,NaOH \longrightarrow Na_3[Cr(OH)_6]$$

Pentaaquohydroxochromium(III) complexes are very weak bases. Their salts hydrolyze and the pH of aqueous solutions usually is 2. The basicity of these salts is defined as the ratio of hydroxyl groups (in percent) bound to chromium to the number of hydroxyl groups in chromium(III) hydroxide, that could theoretically be bound to chromium. Pentaaquohydroxochromium(III) complexes therefore have a basicity of 33 %. When a second hydroxyl group enters the complex the basicity increases to 67 %. However, from a basicity of 60 % onward, chromium(III) hydroxide precipitates and these compounds are not used in practice.

5.2. Chromium(III) Sulfates and Chrome Tanning Agents

Chromium(III) Sulfate. *Anhydrous chromium sulfate* [*10101-53-8*], $Cr_2(SO_4)_3$, M_r

392.18, ϱ 3.0 g/cm^3, is a violet powder which is insoluble in water but dissolves to form complexes when reducing agents are added. For its preparation, chromium metal or chromite is heated over 250 °C with sulfuric acid.

The *octadecahydrate* [*13520-66-6*], $[Cr(H_2O)_6]_2(SO_4)_3 \cdot 6\,H_2O$, M_r 716.45, ϱ 1.86 g/cm^3, forms cubic crystals. The violet compound gives off water on heating and above 70 °C it is converted, with further loss of water, into a dark-green crystalline *pentadecahydrate* [*10031-37-5*]. As the water content diminishes, the solubility decreases.

*Solutions*of chromium(III) sulfates can be made by treating chromite with sulfuric acid in the presence of chromium(VI) compounds [64]. Since other components of the ore are solubilized at the same time, the solutions are strongly contaminated; separation of magnesium, aluminum, and iron presents such great difficulties that this process has not yet gone beyond the experimental scale. Chromium(III) sulfate solutions are also obtained by dissolving ferrochromium in sulfuric acid, a process in which iron(II) sulfate is obtained as a coproduct. So far, the economical preparation of a pure product has been only partially successful.

Large quantities of chromium(III) sulfate solution are produced in the oxidation of organic substances with chromic acid or sodium dichromate in sulfuric acid solution. Examples of this are the preparation of anthraquinone from anthracene, the preparation of benzoquinone from aniline, or the bleaching of montan wax. These solutions are used to produce other chromium products; an electrolytic regeneration to dichromate is also possible.

Tanning Agents. *Basic chromium(III) sulfates* are used on a large scale as tanning agents for leather. Industrially, two processes are available for the reduction of sodium dichromate: (1) reaction with organic compounds (molasses, sugar) in the presence of sulfuric acid and (2) reduction with sulfur dioxide.

Reduction with molasses is carried out in aqueous solution; about 30 kg of molasses or 15 kg of cane sugar is required for 100 kg of sodium dichromate dihydrate. The amount of sulfuric acid required depends on the desired basicity. To adjust the basicity to 33 %, about 103 kg of 96 % acid is needed. The reaction is strongly exothermic; water evaporates in abundance and must be continuously replenished. Lead-lined vats have proved successful as reaction vessels. The exhaust gases have an unpleasant smell, but this can be eliminated by scrubbing with water in a wash tower or by heating them after condensing the water vapor.

The properties of the final product are, to a certain extent, dependent on how the reaction is performed. If sulfuric acid is added first to the dichromate solution and the reducing agent is then added slowly, relatively few organic acids are produced as result of side reactions. The proportion of these acids becomes considerably larger if the dichromate solution is mixed first with the reducing agent and the sulfuric acid is added last. The organic acids form chromium complexes and mask the tanning agent. This masking delays the tanning process.

In the *reduction with sulfur dioxide* sulfuric acid is generated in such proportions that the tanning agent produced has a basicity of 33 %. For the reduction of 100 kg of sodium dichromate dihydrate, 65 kg of SO_2 is theoretically required.

The reaction is carried out in lead-lined or brick-lined absorption towers containing ceramic packing material. Sulfur dioxide is produced by combustion of liquid sulfur which yields a gas containing 8 – 18 % SO_2; SO_2-containing gases from other manufacturing processes are also suitable for the reaction.

For the production of chrome tanning agents, chromium(III) sulfate solutions which are generated in the manufacture of organic intermediates may also be used. Impurities must be removed from such solutions, and the solutions are then converted to the correct basicity by acidification or neutralization. They must also be concentrated by evaporation.

To manufacture *solid tanning agents,* the concentrated solutions are dried in spray driers made of stainless steel. With a basicity of 33 %, the amorphous green powder obtained generally contains 24 – 26 % Cr_2O_3, 25 – 27 % SO_3, 22 – 25 % Na_2SO_4, and 22 – 25 % water. In air, the powder absorbs moisture and the particles stick together. Under the microscope, the individual particles, which are often hollow spheres or fragments of such spheres, have a glassy appearance.

The product obtained by spray drying is sold under numerous trade names, e.g., Chrometan (British Chrome and Chemicals, UK); Chromi-

tan (BASF, Federal Republic of Germany); Chromosal (Bayer, Federal Republic of Germany); Salcromo (Stoppani, Italy); Tanolin (Hamblett & Hayes Co., Mass., USA).

Paper or jute sacks with watertight polyethylene liners or wrappings, or plastic sacks are used for packing.

Solutions of basic chromium(III) sulfate containing 12 – 18 % Cr_2O_3 are also available commercially. Solutions of higher concentration must be kept warm because sodium sulfate precipitates at room temperature. The solutions are transported in rubber-lined rail or road tankers; lead-lined tankers are also suitable.

Besides the standard 33 % basic type, a large number of products of higher basicity are available. To improve their stability toward alkali, these contain various quantities of organic acids. The market importance of chrome tanning agents containing 30 % chromium oxide and having a basicity of 50 % has increased.

In addition to the standard products, mixtures of chrome tanning agents have established themselves on the market. These contain basifying agents which react slowly and eliminate the need for the tedious basifying process. Mixed products containing special organic masking agents and having a high total basicity have been developed. These are used in combination with conventional chrome tanning agents and afford a high degree of chromium exhaustion in the liquors.

Potassium chromium(III) sulfate [*10141-00-1*], [*10279-63-7*], [*7788-99-0*], potassium chrome alum, $KCr(SO_4)_2 \cdot 12\,H_2O$, M_r 499.11, ϱ 1.813 g/cm^3, crystallizes in the cubic system forming violet regular octahedra which decay in air and melt at 89 °C, the color changing to green; the enthalpy of formation is − 5788 kJ/mol. The solubility in water at 25 °C is 11.1 wt %. The solution is violet when cold, but becomes green above 50 °C, this change being accompanied by a decrease in the molar conductance. For 0.125 M solution at 50 °C, the molar conductance is 221 Ω^{-1} cm^2 mol^{-1} for the violet form and 202 Ω^{-1} cm^2 mol^{-1} for the green form. This change is reversible and its rate is increased by acids. The green form always occurs as an amorphous solid and crystals are unknown.

In addition to the alum containing 12 molecules of water of crystallization, potassium chromium(III) sulfates containing one [*35177-45-8*], two [*35177-44-7*], and six [*35177-43-6*] molecules of water are known.

For the preparation of potassium chromium alum, a saturated potassium dichromate solution is reduced with sulfur dioxide in the presence of sulfuric acid.

During the reaction, the temperature must be kept below 40 °C by cooling to prevent the green modification being produced. Apart from sulfur dioxide, such organic compounds as formaldehyde, methanol, or starch are suitable as reducing agents. Crystallization starts after sulfuric acid has been added and the temperature is kept further below 40 °C. If the solution is allowed to settle in vats, large crystals are produced, whereas fine ones result if the solution is stirred. The industrial product contains 15 % Cr_2O_3 and 0.01 – 0.03 % Fe. The mother liquor may be reused in the production of the alum solution, but after several cycles it is so enriched in magnesium and sodium sulfate that crystallization of the alum is retarded. The trivalent chromium is then precipitated from the solution with alkali and may be recycled. Lead-lined equipment is used for the production.

The alum can also be prepared from ferrochromium. The reaction with sulfuric acid first results in a solution of chromium(III) and iron(II) sulfates. The majority of the iron(II) sulfate can be removed by crystallization. With potassium sulfate added, the filtrate yields the alum which is contaminated with 0.1 – 0.2 % Fe.

Potassium chrome alum was formerly used on a large scale as a tanning agent in the leather industry, but its importance in this field has receded with the introduction of the basic chromium sulfates. Other fields of application are the textile industry and the film and photographic industry.

Ammonium chromium(III) sulfate [*10141-00-1*], ammonium chrome alum, $NH_4Cr(SO_4)_2 \cdot 12\,H_2O$, M_r 478.47, ϱ 1.72 g/cm^3, crystallizes in the cubic system forming bluish violet octahedra which appear ruby red when held against the light and slowly decay in air. When heated to 70 °C they turn green, and at 94 °C the compound melts in its water of crystallization. Ammonium chrome alum is obtained from chromium(III) sulfate solutions by addition of a stoichiometric amount of ammonium sulfate. Ammonium chrome alum

crystallizes much more easily than potassium chrome alum. The preparation of ammonium chrome alum from carbon-rich ferrochromium has assumed relatively great importance in the electrochemical production of pure chromium metal [65].

5.3. Other Chromium(III) Salts

Chromium(III) fluoride [*7788-97-8*], CrF_3, M_r 108.99, ϱ 3.8 g/cm^3. The anhydrous compound forms highly refractive rhombohedral crystals. It melts above 1000 °C and is distinctly volatile between 1100 and 1200 °C. Chromium(III) fluoride is insoluble in water if no divalent chromium is present. Double compounds are formed with other metal fluorides, e.g., green $CrF_3 \cdot 2\,KF \cdot H_2O$.

Hydrates are known which contain three [*16671-27-5*] to nine [*68374-27-6*] molecules of water. The violet hexaaquochromium(III) fluoride, $[Cr(H_2O)_6]F_3$, and its trihydrate, $[Cr(H_2O)_6]F_3 \cdot 3\,H_2O$, can be obtained from hexaaquochromium(III) salt solutions and alkali fluorides. Products containing less water are green. The composition of the industrial product corresponds approximately to $CrF_3 \cdot 3.5\,H_2O$ and the product contains about 30 % chromium.

For the production of chromium(III) fluoride hydrate, chromium(III) oxide hydrate is dissolved in hot aqueous hydrofluoric acid and the green salt crystallizes. Chromium(III) fluoride is used in the textile industry for mordanting wool, for chromating dyestuffs, and in vigoureux printing. Chromium(III) fluorides have also found application in rust-prevention paints as corrosion inhibitors.

Chromium(III) Chloride.

Anhydrous chromium(III) chloride [*10025-73-7*], $CrCl_3$, M_r 158.36, forms hexagonal reddish violet flakes which sublime at 950 °C yielding a vapor that dissociates above 1300 °C. The enthalpy of formation is −554.8 kJ/mol. Chromium(III) chloride is insoluble in water if no reducing agent is present. On roasting in air, chromium(III) oxide is produced.

Anhydrous chromium(III) chloride is obtained along with iron(II) chloride by chlorinating roasting of chromite in the presence of carbon at 900–1050 °C [66]. Oxygen is added to the chlorine to prevent nonvolatile residues, in particular minor constituents of the ore, from sintering together. Fractionating condensation between 400 and 640 °C has been suggested for separating the chloride vapors [67]. The compound can also be obtained by chlorinating chromium(III) oxide in the presence of reducing agents or by treating ferrochromium with chlorine [68].

Chromium(III) chloride can be prepared readily from chromyl chloride by reaction with carbon monoxide and chlorine. The reaction proceeds rapidly in the gas phase at 750–850 °C [69]. Since chromium(III) chloride evaporates only at a higher temperature, a considerable portion of the product, which varies as a function of the partial pressure, is produced in the form of fine crystals. As the smoke cools down, these act as crystallization nuclei for any gaseous chromium(III) chloride still present. This procedure prevents the deposition of solid on the cooling surfaces.

Anhydrous chromium(III) chloride has been suggested for chromizing steel parts by surface diffusion; it can be used for the production of high-purity ductile chromium metal by reduction with magnesium and for the synthesis of organic chromium compounds.

Chromium(III) chloride hexahydrate [*10060-12-5*] is obtained in pure form by introducing hydrogen chloride and methanol into an aqueous solution of chromic acid. The reaction proceeds exothermally, and adequate cooling must be provided. Of the three isomeric hydrates (see 5.1) the dihydrate of the tetraaquodichlorochromium(III) chloride crystallizes in the cold. This is used as an intermediate in the production of chromium complex dyes and other chromium salts, e.g., chromium stearates, which are of interest as impregnating agents for textiles or paper. The solution is also used as a mordant in the textile industry.

Chromium(III) Acetate [*1066-30-4*]. The bluish violet hexaaquo salt [*66851-10-3*], $[Cr(H_2O)_6](CH_3COO)_3$, M_r 337.22, forms needle-shaped crystals. Basic chromium(III) acetates are green. For their preparation chromium(III) hydroxide hydrate is dissolved in dilute acetic acid, and the solid is obtained by drying on drums or in a spray drier. Basic chromium acetates are used as mordants

in calico printing and worsted top printing, and also for fixing vigoureux dyes. Combinations of basic chromium(III) acetate and basic chromium(III) formate also find application in the textile industry as mordants. In addition, chromium(III) acetate is used as a starting compound in the production of organic chromium dyes.

Chromium(III) Nitrate [*13548-38-4*]. Normally, chromium(III) nitrate crystallizes with nine molecules of water, $[Cr(H_2O)_6](NO_3)_3 \cdot 3\,H_2O$ [*7789-02-8*], M_r 400.15, ϱ 1.8 g/cm^3. The dark violet rhombic prisms become green above 36 °C and melt at 66 °C. The nitrate group is not bound to the trivalent chromium in a coordination compound. The salt is readily soluble in water, acid, alkali, and alcohol. To prepare chromium(III) nitrate, chromium(III) oxide hydrate is dissolved in nitric acid and the nitrate is allowed to crystallize. The compound is also produced by reduction of chromic acid with methanol in the presence of nitric acid. During the exothermal reaction, the temperature is kept between 55 and 65 °C by cooling. Chromium(III) nitrate is used to a limited extent as a mordant in cotton printing, usually together with basic chromium acetates. In addition, the salt is suitable for producing alkali-free catalysts.

Chromium(III) Phosphate. *Anhydrous chromium(III) phosphate* [*7789-04-0*], $CrPO_4$, M_r 146.97, ϱ 2.99 g/cm^3, is a black powder which belongs to the orthorhombic crystal system. It is insoluble in water, hydrochloric acid, and aqua regia but is attacked by boiling sulfuric acid. It is obtained by calcining its hydrates. The violet *hexahydrate* [*13475-98-4*], $[Cr(H_2O)_6]PO_4$, M_r 255.06, ϱ 2.12 g/cm^3, precipitates from chromium(III) salt solutions upon addition of phosphoric acid and disodium hydrogen phosphate. The crystals are triclinic and only sparingly soluble in water. Green chromium(III) phosphates precipitated hot contain two to three molecules of water after drying. A salt containing four molecules of water is also known. The green products are virtually insoluble in water. They are used to a small extent as pigments and have corrosion-inhibiting properties.

Chromium lignosulfonates [*9066-50-6*] are prepared by reaction of sulfite waste liquor from the pulp industry with sodium dichromate solution, the hexavalent chromium being reduced by the organic material to trivalent chromium. After filtration, the pH is adjusted by means of alkali and the product is dried in spray driers. It is a free-flowing, water-soluble powder and normally contains cations in the following quantities: ammonium 1.5 – 2.5 %, chromium 2.5 – 4.2 %, iron 0 – 5 %, and sodium 1.5 – 2.5 %. These saltlike compounds are used to a large extent in the petroleum industry as additives for lowering the viscosity of drilling muds and decreasing liquid loss [70].

6. Chromic Acids and Chromates(VI)

6.1. Chromic Acids

Chromic acids are not known in the free state. Depending on the method of preparation, mono-, di-, tri-, or tetrachromic acids are formed in aqueous solution. In alkaline or dilute solution, formation of the yellow monochromate ion is favored, but in acid solution or at high concentrations, the orange-red dichromate ion is formed preferentially. Aqueous solutions of chromic acids are, therefore, yellow or red depending on their concentrations. Figure 5 shows the density as a function of concentration.

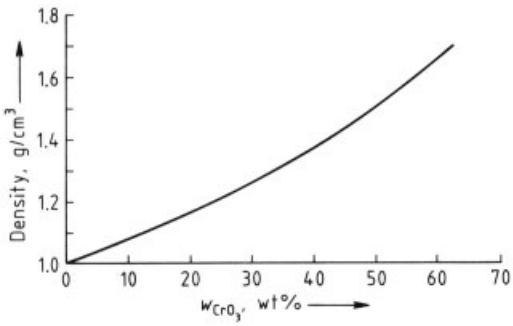

Figure 5. Density of aqueous chromic acid solutions at 15.6 °C (60 °F)

Dissociation constants (at 25 °C) are as follows:

$$H_2CrO_4 \rightleftharpoons H^+ + HCrO_4^- \quad K_1 = 1.21$$

$$HCrO_4^- \rightleftharpoons H^+ + CrO_4^{2-} \quad K_2 = 3.7 \times 10^{-7}$$

$$H_2Cr_2O_7 \rightleftharpoons H^+ + HCr_2O_7^- \quad K_3 = 1$$

$$HCr_2O_7^- \rightleftharpoons H^+ + Cr_2O_7^{2-} \quad K_4 = 0.85$$

$$Cr_2O_7^{2-} + H_2O \rightleftharpoons 2\,H^+ + 2\,CrO_4^{2-} \quad K_5 = 3 \times 10^{-15}$$

$$Cr_2O_7^{2-} + H_2O \rightleftharpoons 2\,HCrO_4^- \quad K_6 = 0.023$$

The standard redox potential for the reaction

$$Cr_2O_7^{2-} + 14\,H^+ + 6\,e \rightleftharpoons 2\,Cr^{3+} + 7\,H_2O$$

is 1.36 V.

Chromic acid solutions are strong oxidizing agents with a strongly acidic character; they form salts with metals and bases. The monochromates(VI), M_2CrO_4, which are derived from chromic acid, hydrolyze in aqueous solution:

$$CrO_4^{2-} + H_2O \rightleftharpoons HCrO_4^- + OH^-$$

The easiest method of preparing chromic acid solutions is to dissolve chromium(VI) oxide in water. In industry, chromic acid is often produced from sodium dichromate(VI) and sulfuric acid.

Chromic acid solutions can also be prepared by anodic oxidation of chromium(III) sulfate solutions [71]; lead-lined cells with a diaphragm and lead electrodes are employed. To keep the concentration of sulfuric acid constant, the chromium(III) sulfate solution is introduced first into the cathode space, where it becomes depleted of sulfuric acid, and then into the anode space. Here, oxidation to chromic acid takes place and the concentration of sulfuric acid is restored to its original value.

In practice, several electrolytic cells are combined to form a unit. At a current density of 3 A/dm^2 the voltage is 3.5 V. Current efficiency is 80 %. Lost chromium is periodically replenished by adding chromium(III) oxide. The electrolytic preparation of chromic acid can also start from chromium hydroxide hydrate with chromic acid as electrolyte [72].

Aqueous chromic acid solutions are used as pickling and chromium-plating baths in the metal processing and plastics processing industries.

6.2. Alkali Chromates and Dichromates

Sodium chromate [*7775-11-3*], Na_2CrO_4, M_r 161.97, *mp* 792 °C, ϱ 2.723 g/cm^3, ΔH^0_{298} −1329 kJ/mol, crystallizes in the orthorhombic system in small yellow needles or columns; transformation to the hexagonal form takes place at 413 °C. The bulk density of the powder is 0.7 g/cm^3; that of the crystals is 1.67 g/cm^3. For solubility, see Table 8. The compound is hygroscopic and forms several hydrates: below 19.5 °C the decahydrate [*13517-17-4*]; between 19.5 and 25.9 °C, the hexahydrate; and between 25.9 and 62.8 °C, the tetrahydrate [*10034-82-9*] which undergoes transformation into anhydrous sodium chromate above 62.8 °C.

To prepare the salt, sodium dichromate solution is usually mixed with a stoichiometric amount of sodium hydroxide, and the salt solution is then crystallized or spray dried. The 96.5 – 98.5 % product (0.4 % NaCl, 2 % Na_2SO_4) is stored and dispatched in watertight steel drums.

Sodium chromate is used as a corrosion inhibitor in the petroleum industry and as a dyeing auxiliary in the textile industry.

Sodium Dichromate. The *dihydrate* [*7789-12-0*], $Na_2Cr_2O_7 \cdot 2\,H_2O$, M_r298.0, ϱ 2.348 g/cm^3, ΔH°_{298} − 2194 kJ/mol, forms orange-red, monoclinic, translucent needles which are converted into the anhydrous salt above 84.6 °C. The bulk density is 1.2 g/cm^3. For solubility, see Table 8. The heat of solution is − 118 kJ/kg. The compound is very hygroscopic and deliquesces in air; in acid solution, it is a strong oxidizing agent.

The preparation of sodium dichromate is described in Chapter 3.

Sodium dichromate is the most important of the industrial chromium chemicals and is used as the starting compound for almost all chromium compounds. Large quantities are used in numerous industrial fields. In the textile industry (wool, cotton, silk, and synthetics), sodium dichromate is used in mordanting and in aftertreatment baths. The leather industry virtually no longer uses sodium dichromate. Sodium dichromate has a variety of uses in the surface treatment of metals, e.g., in the pickling of steel, aluminum, magnesium, and other metals and

Table 8. Solubility (in wt %) of various chromates in water

	Temperature, °C								
	0	20	25	40	50	60	75	80	100
Sodium chromate		44.3		48.8		53.5		55.8	56.1
Sodium dichromate dihydrate	70.6	73.18		77.09		82.04		88.39	91.43
Potassium chromate		39.96							45.0
Potassium dichromate	4.3	11.7		20.9		31.3		42.0	50.2
Ammonium chromate	19.78		27.02		34.4		41.2		
Ammonium dichromate	15.16	26.67		36.99		46.14		54.10	60.89
Silver chromate		0.0025			0.0053				0.0041

their alloys. The ability of the chromates to convert gelatin or protein into an insoluble form on exposure to light is exploited on a large scale in printing technology (lithography). A further field of application for the dichromates is in corrosion protection; they are added to crude oil in pipelines and to water in closed cooling systems as direct corrosion inhibitors. Sodium dichromate is also used for the manufacture of wood preservatives.

In the chemical industry, sodium dichromate is used as a strong oxidizing agent in numerous cases; the most important include oxidation of anthracene to anthraquinone (dyes), of aniline to quinone (hydroquinone for the photographic industry), of camphene to camphor, and of contaminants in oils, fats, tallow, and waxes (soap industry, wax bleaching).

To these classical applications, the wide field of catalysts and catalyst carriers containing chromium(III) oxide or chromates has been added over the past thirty years. These are important in a variety of oxidation and carbonizing processes.

Anhydrous sodium dichromate [*10588-01-9*], $Na_2Cr_2O_7$, M_r 261.96, *mp* 356.7 °C, ϱ 2.52 g/cm^3, bulk density 1 g/cm^3, heat of solution ca. − 33.5 kJ/kg, forms light-brown to orange-red plates which are strongly hygroscopic. They decompose above 400 °C with the formation of sodium monochromate(VI), chromium(III) oxide, and oxygen.

Anhydrous sodium dichromate can be prepared by melting down sodium dichromate dihydrate, by crystallizing aqueous dichromate solutions above 86 °C, or by drying sodium dichromate solutions in spray driers.

Anhydrous sodium dichromate is required for cases in which the water content of the dihydrate has an interfering action. Thus, for example, the energy liberated in the oxidation with anhydrous sodium dichromate is greater than that liberated in the case of sodium dichromate dihydrate. Anhydrous sodium dichromate is, therefore, used in the preparation of chromium(III) oxide by the dry process (4.1), in pyrotechnics, and in anhydrous oxidation processes where it replaces the more expensive potassium dichromate. Compared with sodium dichromate dihydrate, anhydrous sodium dichromate has the advantage that it can first absorb two molecules of water (13 wt %) instead of deliquescing immediately when moisture is admitted.

Potassium chromate [*7789-00-6*], K_2CrO_4, M_r 194.2, *mp* 968.3 °C, ϱ 2.73 g/cm^3, ΔH^0_{298} − 1383 kJ/mol, occurs as the stable β-modification. The lemon-yellow, nonhygroscopic prisms are isostructural with K_2SO_4. At 666 °C they are converted into hexagonal α-potassium chromate. For solubility, see Table 8. The heat of solution is −71.3 kJ/kg. The salt crystallizes from aqueous solution in anhydrous form and is thermally stable.

Potassium chromate is obtained by reacting potash with potassium dichromate. The potassium salt has been supplanted nearly completely by the cheaper sodium chromate and is used only for very specific purposes such as in the photographic industry.

Potassium dichromate [*7778-50-9*], $K_2Cr_2O_7$, M_r 294.19, *mp* 397.5 °C, occurs in two modifications. α-$K_2Cr_2O_7$, tabular or prismatic, bright orange-red triclinic crystals, ϱ 2.676 g/cm^3, has a bulk density of about 1.3 – 1.6 g/cm^3; at 241.6 °C α-$K_2Cr_2O_7$ transforms to β-$K_2Cr_2O_7$. For solubility, see Table 8. The heat of solution is − 258.3 kJ/kg. The thermodynamic data

are as follows: c_p 219.7 J mol^{-1} K^{-1}, ΔH°_{298} − 2033 kJ/mol, S° 291.2 J mol^{-1} K^{-1}, heat of fusion 36.7 kJ/mol. The substance is not hygroscopic and, above the melting point, decomposes into potassium chromate, chromium oxides, and oxygen.

Today potassium dichromate is obtained primarily by conversion of sodium dichromate with potassium chloride. Potassium dichromate has largely been supplanted by the cheaper sodium dichromate but is still used whenever its advantage of being nonhygroscopic is important, for example, in the match, firework, film, and photographic industries. Potassium dichromate is of interest in the preparation of yellow and green zinc pigments (→ Pigments, Inorganic).

Ammonium chromate [*7788-98-9*], $(NH_4)_2CrO_4$, M_r 152.07, smells of ammonia and forms golden yellow needles, ϱ 1.886 g/cm^3, ΔH^0_{298} − 1152 kJ/mol, S^0_{298} 656 J mol^{-1} K^{-1}. For solubility, see Table 8. In air, it decomposes into ammonia, water, and ammonium dichromate. On heating it ignites and decomposes into chromium(III) oxide, ammonia, and nitrogen. Ammonium chromate is prepared from ammonium dichromate with the addition of ammonia.

Ammonium dichromate [*7789-09-5*], $(NH_4)_2Cr_2O_7$, M_r252.06, forms large, bright orange-red crystals, ϱ 2.155 g/cm^3, bulk density ca. 1.0 – 1.3 g/cm^3. For solubility, see Table 8. The heat of solution is − 230.9 kJ/mol. Ammonium dichromate crystallizes in anhydrous form from aqueous solution and is not hygroscopic. Decomposition, which is not preceded by melting, sets in on heating to 180 °C; this becomes self-maintaining at 225 °C and above. Decomposition proceeds with displays of fire and heat, and large amounts of gas are developed. The products of decomposition are chromium(III) oxide, nitrogen, and water vapor. Ammonium dichromate reacts very violently with organic solvents.

Ammonium dichromate is prepared by reaction of sodium dichromate with ammonium chloride or, less frequently, ammonium sulfate. Ammonium dichromate is used as the starting material for preparing very finely divided chromium(III) oxide and, in addition, finds application primarily in pyrotechnics, wood preservation, and photography (lithography) for the preparation of light-sensitive solutions of gelatin or proteins. Ammonium dichromate is also used to prepare catalysts for organic syntheses. A further field of application is the production of magnetic chromium(IV) oxide. Because of its self-ignition properties and explosiveness, ammonium dichromate is subject to the German Explosives Law and the IMDG code, class 5.1, UN No. 1439. It is also distributed moist.

6.3. Other Chromates

Barium chromate [*10294-40-3*], $BaCrO_4$, M_r 253.33, *mp* 1400 °C (decomp.), ϱ 4.498 g/cm^3, ΔH^0_{298} − 1156 kJ/mol, crystallizes as light-yellow transparent rhombic crystals which are isomorphous with barium sulfate. Barium chromate is only sparingly soluble in water but dissolves readily in acids.

In the presence of excess alkali chromate or dichromate, barium chromate has a tendency to form double salts, among which special mention may be made of potassium barium chromate [*13819-19-7*], $K_2CrO_4 \cdot BaCrO_4$, and ammonium barium chromate [*13819-20-0*], $(NH_4)_2CrO_4 \cdot BaCrO_4$, both of which are light yellow.

These barium chromate double salts can be prepared by reaction of soluble barium salts or barium hydroxide with alkali chromate(VI) or dichromate(VI). In weakly acid solutions, as is the case, for example, if dichromates(VI) are used, the precipitation is incomplete. Quantitative precipitation is achieved by adding sodium acetate.

Yellow barium chromate and its double salts can be used for the production of chrome pigments for paints; the double salts, in particular, are excellent corrosion protection paints for all metals. They form sparingly soluble metal chromates, which prevent attack by moisture (condensation, seawater) even more readily than zinc chromate.

Calcium chromate [*13765-19-0*], $CaCrO_4$, M_r 156.07, *mp* 1020 °C (decomp.), ϱ 3.12 g/cm^3, is a yellow powder which is sparingly soluble in water (4.3 wt % at 0 °C, 0.42 wt % at 100 °C). The compound has acquired no industrial importance, but its prepara-

tion directly from chromium ores is described in several patents [73], [74].

Calcium dichromate [*14307-33-6*], $CaCr_2O_7$, M_r 256.06, ΔH^0_{298} −1821 kJ/mol, is thought not to constitute a uniform crystalline phase but to be a mixture of phases consisting of calcium monochromate and chromium(VI) oxide. The compound forms a series of readily soluble hydrates. Thus, below 10 °C, the hexahydrate $CaCr_2O_7 \cdot 6\,H_2O$ exists; between 20 and 40 °C, the pentahydrate [*61204-19-1*]; between 50 and 60 °C, the tetrahydrate; and above 70 °C, the monohydrate [*85752-77-8*]. A red deliquescent calcium dichromate trihydrate has also been prepared from the tetrahydrate.

Calcium dichromate can be made industrially by oxidative roasting of chromium-containing ores with calcium carbonate or calcium oxide and subsequent leaching of the cake with an acid, e.g., chromic or sulfuric acid [74].

Copper Chromates. Neutral copper chromate [*13548-42-0*], $CuCrO_4$, M_r 179.53, is produced as a yellowish brown, water-containing compound by precipitation from a copper sulfate solution with sodium or potassium dichromate. The compound is used as the starting material for "chrome black" which is prepared by calcining the neutral salt under oxidizing conditions and subsequently leaching with hydrochloric acid.

The double salt *copper ammonium chromate* (cupric ammonium chromate) is required in dyeworks along with logwood and fustic extracts to obtain olive green wool or cotton dyes. Copper chromate is used as such or in reduced form as a catalyst in a number of petrochemical reactions.

Iron Chromates. No anhydrous iron(II) chromate is known. On the other hand, the watersoluble double salts, $KFe^{III}(CrO_4)_2 \cdot 2\,H_2O$ [*20161-12-0*] and $NH_4Fe^{III}(CrO_4)_2 \cdot 2\,H_2O$ [*20161-14-2*] do exist, and potassium iron(III) chromate is a corrosion inhibitor. A similar corrosion-inhibiting compound is also thought to be formed on steel surfaces that are treated with chromate solutions. Industrially, the compound is produced by the reaction of iron(III) chloride solution with potassium dichromate in an autoclave at 130–160 °C. The precipitate is filtered and dried.

Lead chromate [*7758-97-6*], $PbCrO_4$, M_r 323.18, *mp* 844 °C (with evolution of oxygen), ϱ 6.123 g/cm^3, ΔH^0_{298} −910 kJ/mol, is a yellowish orange powder which occurs in three modifications:

$$\underset{\textit{yellowish orange}}{\text{monoclinic}} \xrightarrow{707\,^\circ\text{C}} \underset{\textit{yellow}}{\text{orthorhombic}} \xrightarrow{783\,^\circ\text{C}} \underset{\textit{red}}{\text{tetragonal}}$$

The solubility product in water is 1.5×10^{-14} at 18 °C. Lead chromate forms mixed crystals with lead(II) oxide, lead sulfate, and lead molybdate. All these salts are virtually insoluble in water; the molybdenum-containing compound is particularly well-known as molybdenum red. Lead chromate is prepared in a manner similar to that used for barium chromate. The composition, color, and quality of lead chromates depend on the conditions of precipitation; they are used widely as yellow to red pigments in the lacquer and paint industry (→ Pigments, Inorganic).

Silver Chromate and Silver Dichromate. Ag_2CrO_4 [*7784-01-2*], M_r 331.73, ϱ 5.625 g/cm^3, ΔH°_{298} −711.7 kJ/mol, S°_{298} 216 J mol^{-1} K^{-1}, c_p 142.3 J mol^{-1} K^{-1} (for solubility, see Table 8), and $Ag_2Cr_2O_7$ [*7784-02-3*], M_r 431.73, ϱ 4.770 g/cm^3, ΔH°_{298} −1218 kJ/mol, form dark red crystals and are soluble in acids, ammonia, and potassium cyanide solutions but not in water. The salts can be precipitated from a silver salt solution with chromate or dichromate solution and are used in the photographic industry.

Zinc chromate [*13530-65-9*], $ZnCrO_4$, M_r 181.36, is sparingly soluble in water but dissolves readily in acids. A series of zinc chromate hydrates exists having the composition $n\,ZnO \cdot m\,CrO_3 \cdot x\,H_2O$. In industry zinc chromate is known as zinc yellow. Its color may be controlled by the mode of preparation. It is made by the reaction of either a suspension of finely ground zinc white (ZnO) in concentrated sulfuric acid or water-soluble zinc salts ($ZnCl_2$, $ZnSO_4$) with potassium or ammonium dichromate. The zinc chromates prepared in this manner always incorporate potassium or ammonium ions into their lattice. Zinc chromate is used in lacquer primers as a corrosion inhibitor instead of minium (red lead) because of its ability to

form insoluble iron(III) chromates with iron or to passivate metal surfaces by oxidation.

7. Other Chromium Compounds

Chromyl chloride [*14977-61-8*], CrO_2Cl_2, M_r 154.90, *mp* − 96.5 °C, *bp* 116.7 °C, ϱ 1.912 g/cm^3, is a blood red oily liquid with a pungent smell, c_p 545 J kg^{-1} K^{-1}, ΔH°_{298} − 567.8 kJ/mol, S^0_{298} 510 J kg^{-1} K^{-1}, heat of fusion 268 kJ/kg. In the temperature range 178 – 390 K the vapor pressure obeys the equation

$$\log p = -3340T^{-1} - 9.08 \log T + 35.06$$

where p is in hPa (mbar) and T in K. The liquid is electrically conducting. Chromyl chloride is easily hydrolyzed; it is an extremely powerful oxidizing and chlorinating agent and, therefore, reacts with organic solvents, often violently. It decomposes in daylight within a week via CrO_2 and Cl_2 to form a series of chromium oxides and chlorides with a low degree of oxidation.

Industrially chromyl chloride is produced from chromium(VI) oxide and hydrogen chloride gas, with concentrated sulfuric acid (> 68 %) being used primarily to bind the water of reaction. The higher density chromyl chloride is drained, distilled, and collected in cooled receptacles.

Chromyl chloride can also be prepared by using mixtures or molten baths of sodium chloride and alkali chromates or dichromates with fuming sulfuric acid. An elegant route involves the reaction of chromium(VI) oxide with liquid thionyl chloride; these reactants are converted quantitatively into chromyl chloride by elimination of SO_2.

Hexacarbonylchromium [*13007-92-6*], chromium hexacarbonyl, $Cr(CO)_6$, M_r 220.06, ϱ 1.77 g/cm^3, forms colorless, highly refractive crystals which belong to the orthorhombic system. The compound sublimes slowly even at room temperature; when heated in a sealed tube it melts at 149 – 150 °C. The boiling point has been calculated to be 147 °C. At 210 °C, explosive decomposition occurs. The enthalpy of formation is − 1077 kJ/mol. Hexacarbonylchromium is somewhat soluble in chloroform and carbon tetrachloride but insoluble in benzene, ether, alcohol, and acetic acid. It is resistant to water and dilute acids at room temperature; there is no sign of attack even by concentrated hydrochloric acid or sulfuric acid in the cold, but concentrated nitric acid causes decomposition. No reaction with alkali occurs, but derivatives are produced with ammonia, pyridine, cyclopentadiene, and other organic ligands.

Attempts to synthesize hexacarbonylchromium directly from chromium and carbon monoxide have been unsuccessful. The action of carbon monoxide and a solution of phenylmagnesium bromide in ether on anhydrous chromium(III) chloride suspended in a mixture of benzene and ether produces intermediates that lead to hexacarbonylchromium upon addition of acid and distilling the ether in vacuo. Finely divided sodium can also be used for reduction instead of phenylmagnesium bromide. Yields of up to 80 % are reported when very finely divided sodium reacts with anhydrous chromium(III) chloride below 0 °C with carbon monoxide at a pressure of about 6 MPa (60 bar); the reaction proceeds in the presence of diethylene glycol dimethyl ether; the mixture is subsequently hydrolyzed while the carbon monoxide pressure is maintained [75].

Applications quoted for hexacarbonylchromium include the tempering and hardening of metal surfaces by chromizing, use as a fuel additive, an intermediate in the preparation of organic chromium compounds, and a catalyst for oxo syntheses.

Chromium(II) Compounds. In general, compounds of divalent chromium are extremely unstable in air and, therefore, are of only minor importance in industry. Gaseous chromium(II) halides (for preparation, see [76]) are used to produce chromium diffusion layers on iron and nickel parts. Air stable sodium fluorochromate(II) is said to be suitable for the precipitation of metals in electroplating operations and for corrosion-protection coatings [77].

8. Analysis

Compounds that are insoluble in water and acids, for example, chromium ore, are solubilized by roasting with sodium peroxide [78] in

the presence of soda. Chrome-tanned leathers are incinerated before being treated with sodium peroxide. Under these conditions, chromium is completely converted into water-soluble chromate(VI). The roasted mass is leached with hot water and after cooling sulfuric acid is added. Compounds of lower oxidation state that are soluble in water or acids are oxidized with ammonium persulfate to chromate(VI) in dilute boiling sulfuric acid in the presence of silver ions as catalysts [79]. Chloride ions must be removed beforehand. Traces of chloride ions are precipitated by adding a few drops of silver nitrate solution. The chromate(VI) solutions obtained in this way or solutions of chromium (VI) compounds that are soluble in water or sulfuric acid, are then titrated with iron(II) sulfate solution after phosphoric acid is added [80]. The end point of this titration can be determined potentiometrically or by means of a redox indicator, e.g., sodium diphenylamine-4-sulfonate [*6152-67-6*], [80]. Iodometric titration of chromate(VI) is also possible [81]. For very low concentrations, e.g., in the analysis of water, atomic absorption spectrometry (AAS) and inductively coupled plasma atomic emission spectrometry (ICP-AES) have proved successful [83]. Trace amounts of chromium are also detected photometrically at 540 nm after oxidation to chromate(VI) and addition of diphenylcarbazide [*140-22-7*] (formation of a reddish violet complex) [82].

9. Transportation, Storage, and Handling

Both hexavalent and trivalent chromium compounds are of commercial significance. Whereas the trivalent compounds and chromium(IV) oxide are not subject to regulations governing transport, the hexavalent compounds are classified internationally as dangerous goods. According to the IMDG code [85], the following classifications are applicable for hexavalent chromium compounds:

Compound	Hazard class	UN No.
$Na_2Cr_2O_7 \cdot 2\,H_2O$, crystals	6.1	3288
$Na_2Cr_2O_7$ solution	6.1	3287
$Na_2Cr_2O_7$ anhydr., crystals	5.1	3087
$K_2Cr_2O_7 \cdot 2\,H_2O$, crystals	6.1	3288
$(NH_4)_2Cr_2O_7$	5.1	1439
CrO_3, solid chromic acid	5.1	1463
Chromic acid solution	8	1755

During handling and storage, chromates, dichromates, and chromic acid must not be brought into contact with readily oxidizable substances.

A limit of 50 µg/L of chromium recommended by the WHO has been adopted in drinking water regulations. This figure was estimated by the U.S. Public Health Services from toxicity data. The value considers the No Observable Adverse Effect Level (NOAEL) from animal tests and the calculated Average Daily Intake (ADI) of drinking water consumers [86].

10. Environmental Protection

Wastewater. In Germany limits for the chromium content of wastewaters have been set up. The permitted concentrations depend on the type of industry. For the leather tanning industry, for instance, the concentration of Cr(IV) must not exceed 0.5 mg/L, and that of total chromium, 1 mg/L [87].

To avoid the pollution of water supply facilities, the risk of a release to surface and ground water has to be minimized during transportation, storage, and handling. In Germany, soluble hexavalent chromium compounds are classified in the highest water pollution class [88].

Exhaust Gas. In Germany, the emission of such carcinogenic chromates as calcium chromate, strontium chromate, chromium(III) chromate, and zinc chromate is limited to 1 mg/m^3 max. (as Cr) [89]; the emission of the other chromium compounds is limited to 5 mg/m^3 (as Cr).

Waste. Within the European Community, waste containing Cr(VI) compounds is considered hazardous [90]. It can only be deposited in dumps with particular safety measures and drainage-water treatment.

In Germany, sewage sludge containing not more than 900 mg per kilogram of chromium may be used as fertilizers in soils for agricultural purposes provided the chromium content of the soil does not exceed 100 mg chromium per kilogram soil before the application [91]. According to the U.S. Environmental Protection Agency, the chromium content of the soil should

not be regarded as a limiting factor for the application of sewage sludge [92]. Studies carried out in Germany on the use of sewage sludge containing a few percent of chromium(III) hydroxide as fertilizer did not reveal any adverse effects on soil and plants [93].

11. Ecotoxicology

Insoluble inert chromium(III) oxide, Cr_2O_3, is the stable mineral end product into which chromium compounds are converted in the environment as a result of natural processes.

Inland Waters. The natural concentration of dissolved trivalent chromium in surface waters is $< 1 - 10\,\mu g/L$ [94–96]. Hexavalent chromium does not occur in natural fresh waters. If dissolved trivalent chromium from anthropogenic sources enters an inland body of water, it precipitates under neutral conditions as chromium hydroxide. This ages and becomes increasingly insoluble, and only a small proportion remains in solution. Chromium(VI) compounds entering inland waters are reduced to chromium(III) compounds by the natural content of organic substances in water and sediments [97], [98].

Table 9 summarizes toxicity data for chromium(III) and chromium(VI) compounds in relation to fish, bacteria, algae, daphnia, and plants.

Limits for Chromium in Fresh Water. Various countries have set guidelines for the content of chromium in fresh water which can be considered tolerable [113–115]. Whereas the U.S. recommendations distinguish between trivalent and hexavalent chromium, both the Canadian and English guidelines consider total chromium.

Soil. The chromium content of the earth's crust varies from a few mg/kg to ca. 50 % in chromite deposits. The average chromium content of rock is 83 mg/kg. The chromium content of soil varies very widely, depending on the geological conditions. The range is from 5 to 1500 mg/kg, and the average concentration about 50 mg/kg [116]. The highest chromium contents are found in soils over basaltic rock. Chromium in soils is linked to the relicts of primary minerals such as chromite or olivine or bound by adsorption to clay minerals. Chromium is found everywhere in the soil of Germany [117], and levels of over 1100 mg/kg may be reached. Such levels are caused solely by geological conditions. All natural chromium is trivalent. Minerals of hexavalent chromium (eg. Crokoite) are extremely rare and can be found only in very few places worldwide.

Plants. As chromium occurs in soil only in the form of sparingly soluble chromium(III) compounds, which are available to plants only to a small extent, chromium is not enriched in the food chain [104], [105], [118–120].

Animals. Grazing animals are not subjected unduly to chromium since only small amounts enter the parts of plants above the ground [105]. Less than 1 % of the chromium contained in plants is available to animal and human organisms. To cover essential needs, the chromium must be in a special, biologically suitable form [121].

12. Nutrition

Chromium is an essential trace element both for the animal and human organism. A number of physiological functions in animal and humans can no longer proceed normally in a state of chromium deficiency [122], [123]. Examples include the role of chromium in glucose-tolerance factor, a cofactor for the action of insulin [124], in cytochrome C reductase [125], and in other metabolic processes [126].

The chromium content in food is in the range of 0.01 – 10 mg Cr/kg dry substance. Consumers in the western world ingest an average of 60 µg Cr per day in their food. The U.S. Food and Nutrition Board (1980) recommends [120] as the safe, but also adequate amount (average dietary intake): 50 to 200 µg per day.

The amount of chromium necessary to humans is not always guaranteed in the dietary intake and even with additional uptake from the environment. Therefore, nutrition science literature contains more references to deficiencies in chromium supply than any toxic effects due to chronic excess supply of chromium [127], [128].

Table 9. Ecotoxicology of chromium compounds

Species or medium	Chromium(III) compounds	Chromium(VI) compounds
Freshwater fish	**$CrCl_3$:** LC_{50} (48 h) minnows, static, 400 mg/L LC_{50} (48 h) trout, static, >1000 mg/L [99] LC_{50} (48 h) ides, static, 300 mg/L **$Cr_2(OH)_2(SO_4) \cdot Na_2SO_4$** LC_0 (30 days) *Brachydanio rerio*, static >1000 mg/L [100], [101]	**$K_2Cr_2O_7$:** 14 day no observable effect level (NOEL) for zebra fish (*Brachydanio rerio*) at 50 mg/L [107] or 80 mg/L [108]
Bacteria	**$KCr(SO_4)_2 \cdot 12\,H_2O$:** 100 mg/L proved not to be poisonous for bacteria of the genus *Escherichia* [100], [101]	**$Na_2Cr_2O_7 \cdot 2\,H_2O$:** toxic limiting concentration for *Pseudomonas putida* 0.78 mg/L [109]
Algae	**$KCr(SO_4)_2 \cdot 12\,H_2O$:** incipient injurious effect at 4 – 6 mg/L (*Scenedesmus*) [100]	**$K_2Cr_2O_7$** (*Scenedesmus subspicatus*): EC_{10} (96 h) 0.5 mg/L [107]; 0.3 – 1.3 mg/L [110]; 1.8 mg/L [108] EC_{50} (96 h) 1.4 mg/L [107]; 1.6 – 4.7 mg/L [110]
Daphnia	**$KCr(SO_4)_2 \cdot 12\,H_2O$** marked injurious effect at 42 mg/L [100]	**$K_2Cr_2O_7$** (21-day test): concentrations of > 0.1 mg/L markedly b decreased the production of offspring [108]; swimming ability of *Daphnia magna:* EC_0 = 0.3 mg/L; EC_{50} = 0.9 mg/L; EC_{100} = 2.2 mg/L [111]
Mammals (oral, rat)	**$Cr(NO_3)_3$:** LD_{50} 3250 mg/kg [102]	**$Na_2Cr_2O_7 \cdot 2\,H_2O$**: LD_{50} 160 mg/kg [112]
Soil mobility	**$Cr(NO_3)_3$:** low, only certain complex compounds are biologically available [103]	**$K_2Cr_2O_7$:** low since chromate is strongly adsorbed on the podsol type of soil and is also reduced by organic soil constituents [108]
Bioaccumulation	**$Cr(NO)_3$:** plants, particularly the parts above ground, do not absorb much chromium; accumulation of Cr is prevented at the very beginning of the food chain [104], [105]	**$K_2Cr_2O_7$:** carp (*Cyprinus carpio*): no bioaccumulation detected [108]
Higher plants **EC_{50} * (14 d)**	**$CrCl_3 \cdot 6\,H_2O$:** *Avena sativa* (oat) 560 mg of Cr/kg of soil [106] *Brassica rapa* (turnip) 230 mg of Cr/kg of soil [106]	**$K_2Cr_2O_7$**: oat: 27 mg/kg [107]; 32 mg/kg [108]; 96 mg/kg [110] turnip: 23 mg/kg [107]; 22 mg/kg [108]; 24 mg/kg [110]
Earthworm **LC_{50} (28 d)**	**$CrCl_3 \cdot 6\,H_2O$:** not known	**$K_2Cr_2O_7$:** > 2000 mg/kg of soil [107] no lethal effects at 1000 mg/kg [110]

* EC = effective concentration; at EC_{50} growth is retarded by 50 % compared with control.

For chromium the gap between essential and toxic concentrations is particularly large. Mammals can tolerate 100 to 200 times the normal chromium content of their bodies without adverse effects [129].

13. Toxicology and Occupational Health

The properties of chromium and its compounds which determine the toxicity and the impact on the environment differ greatly depending on the valency state. Only hexavalent chromium compounds are biologically active. Metallic chromium and the trivalent compounds, including those in chromium ores, are neither irritating, mutagenic, nor carcinogenic [129].

Reviews of the toxicity of chromium and its compounds are given in [130–132].

Chromium Dioxide. In a single administration, chromium dioxide is resorptively nontoxic; the lethal dose is greater than 17000 mg/kg (rats, oral). In rabbits, chromium dioxide causes slight

Table 10. Output of chromium ore and concentrates (in 1000 t) [133–136]

	1901 – 1982	1982	1983	1984
Former USSR∗	55 000	3 400	3 400	3 400
South Africa	50 000	2 200	2 250	2 450
Zimbabwe	22 000	430	450	450
Turkey	21 000	370	360	400
Philippines	18 000	350	360	400
Albania∗	15 000	1 200	1 350	1 350
India	7 900		360	410
		340		
Brazil	4 100	400	370	400
Finland	3 700	400	360	400
Cuba∗	3 700	30	40	30
Iran	3 400	40	40	30
Greece	2 000	40	50	50
Madagascar	2 000	60	40	40
Remainder	800	40	70	90
Total worldwide	208 600	9 300	9 500	9 900

∗ Estimated.

primary irritation of skin and mucous membranes.

Since any chromium dioxide in the normal atmosphere may contain traces of hexavalent chromium, care is to be taken during production and further processing to ensure that the concentration of chromium(VI) compounds at the workplace does not exceed the TLV for chromium(VI).

Employees of chromium dioxide plants are examined annually for chromium exposure. So far no indications have been found of any disease caused by chromium in chromium dioxide production plants.

Table 11. Estimated consumption of chromium ore in terms of products in 1984 [134], [137]

Product	World	United States
Ferrochromium, FeSi chromium, chromium metal, and foundry sands	72 %	60 %
Refractory bricks	12 %	20 %
Chemicals	17 %	20 %

Hexavalent Chromium Compounds. Particularly chromic acid and the alkali chromates corrode and irritate the skin and mucous membranes.

Acute Effects. Sodium monochromate and dichromate, potassium dichromate, and ammonium dichromate are soluble chromium(VI) compounds. These substances are toxic when swallowed, very toxic when inhaled, and dangerous to health when they come into contact with the skin. In humans, inhalation of these compounds, even at low doses, leads to coughing, chest pain, difficult breathing and fever.

Sodium monochromate and dichromate, potassium dichromate, and ammonium dichromate irritate the skin, eyes and respiratory tract. The risk of serious eye damage cannot be ruled out. In humans, repeated contact of hexavalent chromium compounds with the skin may lead to so-called chromium ulcers, but only on previously damaged parts of the skin.

Potassium dichromate and sodium monochromate show a skin-sensitizing effect in animal studies. There is evidence that soluble chromium(VI) compounds have a skin-sensitizing effect in humans. Skin allergy in humans is generally tested with potassium dichromate.

Hexavalent chromium compounds are capable of sensitizing the skin strongly, which may lead to chronic eczemas, particularly when the source is cement dust containing chromium(VI). In contrast, no sensitization of the respiratory tract occurs.

Uptake by the digestive system (mostly suicidal cases) causes serious intestinal inflammation, sometimes with loss of blood. Damage to the renal tubules occurs mainly after dermal absorption. This can lead to kidney failure if the spontaneous reduction capacity of plasma of about 2 ppm (20 min) is not sufficient to reduce chromium(VI) to nontoxic chromium(III). The

administration of high doses of ascorbic acid facilitates this reduction, and chromium(III) is then excreted in the urine without causing kidney damage [140].

Chronic Effects. Repeated inhalation of ammonium dichromate and sodium dichromate primarily affects the respiratory tract, while no corresponding studies are available with sodium monochromate and potassium dichromate. There are no significant data on repeated oral intake of ammonium dichromate or sodium dichromate, while liver and kidney damage have been described for sodium monochromate and potassium dichromate. In humans, prolonged inhalation of the dust damages the nasal mucosa or even the nasal septum (septal perforation) [139].

Chronic irritation of the nasal mucous membrane can lead to its atrophy. Chronic bronchitis has been reported following long-term exposure to hexavalent chromium compounds, but this does not occur normally under today's manufacturing conditions [129]. Impairment of breathing has not been observed.

Mutagenicity. Sodium monochromate and dichromate and potassium dichromate show mutagenic activity in tests with bacteria, microorganisms, cell cultures, and in animal studies. Ammonium dichromate shows mutagenic activity in tests with bacteria.

All hexavalent chromium compounds are mutagenic, but trivalent chromium compounds do not have any mutagenic potential. In some cases, the hexavalent compounds must first be solubilized to show a mutagenic effect. This effect is counteracted by the reduction of chromium(VI) to chromium(III), for example, by adding body fluids or organ homogenates [141], [142].

Carcinogenicity. Many animal experiments have been carried out to test the carcinogenicity of chromium compounds. These include subcutaneous injection, inhalation, intratracheal instillation, and introduction of pellets into the bronchial tree [143–147].

Depending on solubility, chromium(VI) compounds show carcinogenic activity after long-term inhalation. For this reason, the MAK Commission classed them under Group III A2 (substances with carcinogenic potential in animal studies).

Chromium(VI) compounds have been classed by IARC under Group I (carcinogenic in humans) based on early epidemiological studies, although clear classification of the findings to the category of carcinogenic substances was not possible.

The EU has classified chromic acid and zinc chromates as category 1 carcinogens (human carcinogens), other chromates/dichromates ($SrCrO_4$, $CaCrO_4$, $Na_2Cr_2O_7$, $K_2Cr_2O_7$, K_2CrO_4) as category 2 carcinogens (animal carcinogens), and PbCrO4 as category 3 carcinogen (suspected carcinogens) [148].

In the United States [149], water-soluble Cr(VI) compounds and insoluble Cr(VI) compound – both "not otherwise classified" – are considered A1, i.e., "confirmed humans carcinogens"

Reproductive Toxicity. In animal studies, sodium dichromate and potassium dichromate do not induce deformities at nonmaternotoxic doses. An embryotoxic effect has been described for potassium dichromate. There are no data on an embryotoxic effect of ammonium dichromate or sodium monochromate.

In animal studies, inhalation of sodium dichromate does not impair fertility. Animal studies with potassium dichromate and sodium monochromate at doses that are nontoxic to the animals yield no evidence of a fertility-impairing effect. There are no data on a fertility-impairing effect of ammonium dichromate.

Epidemiology. Epidemiological studies among workers in chrome-tanning plants and chrome(III) pigment manufacturing plants have proved trivalent chromium compounds to be noncarcinogenic [150].

Earliest indications of increased incidences of lung cancer among workers in the chromate manufacturing industry were observed in the 1930s. More evidence in support of this observation came from a large number of epidemiological studies carried out after 1948 [151], [152]. Studies carried out between 1948 and 1956 showed a 25- to 29-fold increase in the incidence of lung cancer. In contrast, investigations performed since 1979 reflect the effects of improved hygienic working conditions and production processes, for example, by avoiding the use of lime in the oxidative roasting of chromium ores (low-lime process). These measures have led to a convergence between the incidences of

Table 12. Imports of chromium ore∗ into the Federal Republic of Germany [138]

	1981	1982	1983
1000 t	268	244	247
Millions of €	25.0	22.4	21.6
Price, €/t	93.00	91.57	87.58

∗ Includes all types of chromium ore.

lung cancer observed in the workers and that expected in unexposed persons [139], [153–156].

An increased incidence of lung cancer was observed among persons employed in the chromate pigment industry only after exposure to zinc chromate. Lead chromate showed no carcinogenic effect [157–159]. Results of epidemiological studies carried out during the handling of chromic acid, particularly during chromium plating, are blurred by confounding factors; a statistically significant increase in the rate of lung cancer was not observed [143], [160]. An increased risk of lung cancer has also not been clearly established in the manufacture of ferrochromium [161], [162]. Occupational health care has been provided in the chromate-producing industry for several decades.

Classification. According to the EU List of Dangerous Substances [148] the following labelling has to be applied for

Occupational Exposure Limits.

Germany [163]. The TRK values for Cr(IV) compounds, except those which are practically insoluble in water, are (as CrO_3in total dust) for

Arc welding	0.1 mg/m^3 (corresponding to 0.05 mg/m^3 as Cr)
Manufacturing of soluble Cr(IV) compounds	0.1 mg/m^3 (0.05 as Cr)
Others	0.05 mg/m^3 (0.025 as Cr)

United States [149]. The TLV values (as Cr) are as follows;

Cr metal and Cr(III) compounds	0.5 mg/m^3
Water-soluble Cr(IV) compounds	0.05 mg/m^3
Insoluble Cr(IV) compounds	0.01 mg/m^3

Biological Monitoring. In addition to the classical assay for chromium in blood or urine, the degree of previous exposure to chromium can be estimated by determining the extent of bound chromium in erythrocytes [164].

14. Economic Aspects

The output of chromium ore is subject to considerable variations. Rich ores which are obtained easily by handpicking have declined. This shrinkage in output has to be counterbalanced by exploiting low-grade ores and upgrading them to saleable concentrates. The statistical documents are incomplete and contradictory. Table 10 shows the output of chromium ores according to various sources of information.

Even the most pessimistic estimates of the worldwide reserves and the extraction possibilities in individual countries do not predict a chromium shortage before the end of this century. In addition, the huge lean-ore deposits in South Africa, Southeast Asia, and the United States, which are still completely untouched, afford a replacement for the rich ores being mined at present.

In Table 11 the consumption of chromium ore is broken down according to products. The refractory brick share is declining because Siemens – Martin steel is being increasingly supplanted by oxygen-blown steel produced in converters with a basic lining.

Table 12 shows chromium ore imports into the Federal Republic of Germany and Table 13 the share of individual applications in the consumption of chromium chemicals.

Table 13. Estimated share (in %) of individual applications in the consumption of chromium chemicals, 1985∗

	United States	Worldwide
Pigments	22	19
Metal processing	15	24
Tanning	12	32
Wood preservation	26	11
Corrosion protection	5	3
Petroleum industry	5	2
Textile dyes	2	1
Catalysts	2	1
Video tapes	2	2
Remainder	9	5

∗ Excluding countries with state trading organizations.

$K_2Cr_2O_7$, $Na_2Cr_2O_7 \cdot 2H_2O$

T+	very toxic
N	dangerous for the environment
R49	may cause cancer by inhalation
R46	may cause heritable genetic damage
R21	harmful in contact with skin
R25	toxic if swallowed
R26	very toxic by inhalation
R37/38	irritating to respiratory sytem and skin
R41	risk of serious damage to eyes
R43	may cause sensitization by skin contact
R50/53	very toxic to aquatic organisms; may cause long-term adverse effects in the aquatic environment
S53	avoid exposure; obtain special instructions before use
S45	in case of accident or if you feel unwell, seek medical advice immediately (show label where possible)
S60	this material and its container must be disposed of as hazardous waste
S61	avoid release to the environment; refer to special instructions / safety data sheet

$(NH_4)_2Cr_2O_7$

E	explosive
T	very toxic
N	dangerous for the environment
R49 – 46	(see above)
R1	explosive when dry
R8	contact with combustible material may cause fire
R21 – 25 – 26 – 37/38 – 41 – 43 – 50/53	(see above)
S53 – 45 – 60 – 61	(see above)

Chromic acid (anhydride), CrO_3, carries the following labelling:

O	oxidizing
T	toxic
C	corrosive
N	dangerous for the environment
R49	may cause cancer by inhalation
R8	contact with combustible material may cause fire
R25	toxic if swallowed
R35	causes severe burns
R43	may cause sensitization by skin contact
R50/53	very toxic to aquatic organisms; may cause long-term adverse effects in the aquatic environment
S53 – 45 – 60 – 61	(see above)

The starting material for virtually all chromium chemicals is sodium dichromate prepared from chromium ore. Table 14 summarizes the production capacities for $Na_2Cr_2O_7 \cdot 2H_2O$.

In Table 15 the chromic acid production capacities are broken down according to countries.

Table 14. Estimated 1985 sodium dichromate capacities in 1000 t of $Na_2Cr_2O_7 \cdot 2H_2O$ per year

Western Europe (EEC)	260
State trading countries	250
United States	220
Asia, Africa	110
Worldwide	840

Table 15. Estimated 1985 chromic acid capacities in 1000 t/a

Western Europe (EEC)	42
United States	40
State trading countries	32
Japan	15
Others	8
Worldwide	137

15. References

General References

1. *Gmelin,* System no. 52, Chrom.
2. M. J. Udy: *Chromium,* vol. 1, Reinhold, New York 1956.
3. A. H. Sully, E. A. Brandes: *Chromium,* 2nd ed., Butterworth, London 1967.

Specific References

4. Minerals Yearbook US Bureau of Mines, Washington, D.C., 1983.
5. *Chromium minerals. World Survey of Production and Consumption.* ROSKILL Information Services Ltd., London 1972.
6. *Mining J. (London),* Mining annual review 1985.
7. Mineral Facts and Problems, US Bureau of Mines Bull. 650, 1970.
8. H. Borchert: "Principles of the genesis and enrichment of chromite ore deposits," in *Methods of prospection for chromite,* OECD, Paris 1964, pp. 175–202.
9. S. Janković: *Wirtschaftsgeologie der Erze,* Springer, Wien 1967.
10. H. Kern: "Zur Geochemie und Lagerstättenkunde des Chroms und zur Mikroskopie und Genese der Chromerze," *Clausthaler Hefte zur Lagerstättenkunde und Geochemie der Mineralischen Rohstoffe,* no. 6, Borntraeger, Berlin 1968.
11. T. Lukkarinen, L. Heikkilä: "Beneficiation of chromite ore, Kemi, Finland,". *10th International Processing Congress 1973,* Inst. Min. Met., London 1973.
12. H. Schneiderhöhn: *Die Erzlagerstätten der Erde,* vol. I, Gustav Fischer, Stuttgart 1958.
13. T. Shabad: *Basic industrial resources of USSR,* Columbia University Press, New York 1969.
14. H. Strunz: *Mineralogische Tabellen,* Akadem. Verlagsges., Leipzig 1966.
15. A. Sutulov: *Mineral resources and the economy of the USSR,* McGraw-Hill, New York 1973.
16. H. D. B. Wilson (ed.): *Magmatic ore deposits, a symposium,* monograph 4, Economic Geology Publ. Co., Lancaster (Pa., USA) 1969.
17. Bayer AG, DE-AS 1 533 076, 1966.
18. Bayer AG, FR 1 531 069, 1967.
19. Bayer AG, DE-OS 1 926 660, 1969.
20. S. K. Cirkov, *Zh. Prikl. Khim. (Leningrad)* **13** (1940) 521–527.
21. Associated Chemical Comp., DE-OS 1 467 298, 1963.
22. G. N. Bogachov et al., *Khim. Promst. (Moscow)* 1961, 655.
23. Produits Chimiques Ugine Kuhlmann, DE-OS 2 329 925, 1972.
24. Bayer AG, DE-AS 1 533 077, 1966.
25. C. K. Williams, US 2 695 215, 1950.
26. R. N. Mittra, N. R. Dhar, *J. Indian Chem. Soc.* **9** (1932) 315–327.
27. Pacific Bridge Co., US 2 431 075, 1945/47.
28. I. G. Farben, DE 492 684, 1927.
29. F. Wöhler, *Justus Liebigs Ann. Chem.* **111** (1859) 117–121.
30. DuPont, US 2 885 365, 1959 (A. L. Oppegard).
31. DuPont, US 2 956 955, 1960 (P. Arthur).
32. T. J. Swoboda, P. Arthur, N. L. Cox, J. N. Ingraham, A. L. Oppegard et al., *J. Appl. Phys.* **32** (1961) 374–375.
33. F. Hund, *Farbe + Lack* **78** (1972) 11–16.
34. E. Köster, "Neue Werkstoffe für magn. Speicherschichten," in H. H. Mende (ed.): *Neuere magn. Werkstoffe und Anwendungen magn. Methoden,* Verlag Stahleisen, Düsseldorf 1983, pp. 149–170.
35. B. L. Chamberland, *CRC Crit. Rev. Solid State Mater. Sci.* **7** (1977) 1–31.
36. D. S. Chapin, J. A. Kafalas, J. M. Honig, *J. Phys. Chem.* **69** (1965) 1402–1409.
37. L. Brewer, *Chem. Rev.* **52** (1953) 9.
38. F. J. Darnell, W. H. Cloud, *Bull. Soc. Chim. Fr.* 1965, 1164–1166.
39. BASF, unpublished, 1983.
40. DuPont, US 3 034 988, 1962 (J. N. Ingraham, T. J. Swoboda).
41. F. J. Darnell, *J. Appl. Phys.* **32** (1961) 1269.
42. H. Y. Chen, D. M. Hiller, J. E. Hudson, C. J. A. Westenbroek, *IEEE Trans. Magn.* **MAG-20** (1984) 24–26.
43. CNRS, DE-OS 3 209 739, 1982 (G. Demazeau, P. Maestro, M. Pouchard, P. Hagenmuller et al.).
44. J. B. Goodenough, *Bull. Soc. Chim. Fr.* 1965, 1200–1207.
45. D. S. Rodbell, J. M. Lommel, R. C. DeVries, *J. Phys. Soc. Jpn.* **21** (1966) 2430.
46. J. Boháĉek, *J. Signalaufzeichnungsmater.* **8** (1980) 55–60.
47. DuPont, US 3 512 930, 1970 (W. G. Bottjer, H. G. Ingersoll).
48. BASF, DE-OS 2 749 757, 1977 (G. Väth, M. Ohlinger, H. J. Hartmann, M. Velic et al.).
49. Montedison, EP 29 687, 1980 (G. Basile, G. Boero, E. Mello Ceresa, F. Montino).
50. W. Ostertag, W. Stumpfi, R. Falk, M. Ohlinger, *Elektronik-Anzeiger* **11** (1972) 225–227.
51. Montedison, DE 2 648 305, 1976 (G. Basile, A. Mazza, M. Spinetta).
52. DuPont, US 3 278 263, 1966 (N. L. Cox).
53. G. R. Cole, L. C. Bancroft, M. P. Chouinard, J. W. McCloud, *IEEE Trans. Magn.* **MAG-20** (1984) 19–23.
54. P. J. Flanders, *IEEE Trans. Magn.* **MAG-12** (1976) 348–355.
55. Bayer AG, DE 1 203 748, 1961.
56. Aktjubinskij Sawod Chromowych Isdelij, DE-OS 2 119 450, 1971.

57. Bayer AG, DE 1 065 394, 1958.
58. Diamond Alkali Co., US 2 993 756, 1959.
59. Diamond Alkali Co., DE-OS 2 018 602, 1970.
60. Diamond-Shamrock Corp., DE-OS 3 020 260, 1980.
61. Diamond-Shamrock Corp., DE-OS 3 020 261, 1980.
62. Diamond-Shamrock Corp., DE-OS 3 020 280, 1980.
63. A. Werner, P. Pfeiffer: *Neue Anschauungen auf dem Gebiet der anorganischen Chemie,* 5th ed., Vieweg, Braunschweig 1923.
64. Griesheim Elektron, DE 369 816, 1920.
65. R. B. Norden, *Chem. Eng.* **63** (1956) 308 – 311.
66. Degussa, FR 817 502, 1936.
67. Pittsburgh Plate Glass Co., US 2 185 218, 1938.
68. I. G. Farben, DE 514 571, 1925.
69. Bayer AG, DE 1 467 327, 1963.
70. W. F. Rogers: *Composition and Properties of Oil Well Drilling Fluids,* Gulf Publ. Co., Houston, Texas, 1963, pp. 459 – 464.
71. Farbenfabriken Meister Lucius Brüning, DE 103 860, 1898.
72. M. J. Udy, US 1 739 107, 1925.
73. M. J. Udy, GB 546 681, 1940.
74. Ass. Chem. Comp., DE-AS 1 467 298, 1964.
75. Ethyl Corp., DE-AS 1 159 913, 1959.
76. G. Brauer: *Handbuch der präparativen anorganischen Chemie,* 2nd ed., vol. 2, Enke Verlag, Stuttgart 1962.
77. DuPont, DE-AS 1 272 909, 1960.
78. I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 8, Interscience, New York-London 1963, pp. 301 ff.
79. I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 8, Interscience, New York-London 1963, p. 326.
80. I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 8, Interscience, New York-London 1963, pp. 327 - 328.
81. I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 8, Interscience, New York-London 1963, pp. 330
82. I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 8, Interscience, New York-London 1963, pp. 338 - 339.
83. C. Veillon, K. Y. Patterson, N. A. Bryden, *Analytica Chim. Acta* **164** (1984) 67 – 76. A. G. Cox, G. Cook, C. W. McLeod, *The Analyst* **110** (1985) 331 – 333.
84. *EEC Directive of* June 27, 1967 (67/548/EEC); *EEC Directive of* July 14, 1976, 1st amendment (69/907/EEC).
85. *International Maritime Dangerous Goods Code (IMDG Code).*
86. EPA: *Ambient Water Quality Criteria for Chromium,* Washington, D.C., NTIS No. PB 81-117647, 1980.
87. Anhang 25 der Abwasserverwaltungsvorschrift nach § 7A des Wasserhaushaltsgesetzes, veröffentlicht als Beilage 164a des Bundesanzeiger vom 31.7. 1996, p. 30.
88. Verwaltungsvorschrift wassergefährdende Stoffe (VwVwS), Verwaltungsvorschrift vom 18. April 1996 zum Wasserhaushaltgesetz über die Einstufung wassergefährdender Stoffe in Wassergefährdungsklassen.
89. *Technische Anleitung zur Reinhaltung der Luft (TA Luft)* , Feb. 28, 1986.
90. *EEC Directive on poisonous and hazardous wastes,* 1991 (91/689/EEC).
91. Klärschlammverordnung, April 15, 1992, *BGBl* 1992, p. 912 – 934.
92. EPA, Municipal Construction Div.: *"Application of Sewage Sludge to Cropland. Appraisal of Potential Hazards of the Heavy Metals to Plants and Animals,"* NTIS PB-264015, EPA-430/9-76-013,Nov. 15, 1976.
93. G. Schmid, W. Pauckner, *Leder* **35** (1984) 165 – 171.
94. C. N. Durfor, E. Becker, *U.S. Geol. Survey Water Supply, Paper No.1812,* 1964.
95. NAS-U.S. *Geochemistry and the Environment,* Washington, D.C., U.S. Government Printing Office, 1974.
96. RIWA annual report 1983, Part A: *Der Rhein,* Aug. 1984, p. 60/61.
97. Industrial Health Foundation (IHF): "Kinetic Reactions of the Fate of Chromium in the Environment," IHF, Pittsburgh 1988.
98. C. A. Harzdorf, *Int. J. Environ. Analyt. Chem.* **29** (1987) 249. C. A. Harzdorf: *Spurenanalytik des Chroms,* Thieme Verlag, Stuttgart 1990.
99. B. Hamburger, H. Häberling, H. R. Hirtz, *Arch. Fischereiwiss.* **28** (1977) 45 – 55.
100. G. Bringmann, R. Kühn, *Gesund. Ing.* **80** (1959) no. 4, 115 – 120.
101. D. M. M. Adema, A. de Ruiter: "The influence of "Basisches Chromsulfat" on the early life stages of *Brachydanio rerio* (Els-test)," Netherlands Organization for Applied Scientific Research, Report R 89/399, Delft 1990.
102. NIOSH, Registry of Toxic Effects of Chemical Substances 1985.

103. R. A. Griffin et al., *J. Environ. Sci. Health Part A* **A 12** (1977) 431 – 449.
104. H. Rasp, *Leder* **32** (1981) 188 – 203.
105. D. Sauerbeck, *Zeitschrift des Verbandes Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (VDLUFA)* , Landwirtschaftliche Forschung, special issue no. 39, Kongreßband 1982, pp. 108 – 129.
106. Unpublished results, Bayer AG, May 1985.
107. P. Friesel et al. (Bundesgesundheitsamt): *Überprüfung der Durchführbarkeit und der Aussagekraft der Stufe 1 und 2 des Chemikaliengesetzes,* Umweltbundesamt (UBA) Forschungsbericht 1984.
108. H. W. Marquart et al. (Bayer AG), same publication as [107] .
109. G. Bringmann, R. Kühn, *Haustech. Bauphys. Umwelttech. Gesund. Ing.* **100** (1979) 250.
110. W. Kördel et al. (Fraunhofer-Institut für Toxikologie und Aerosolforschung), same publication as [107] .
111. G. Bringmann, R. Kühn, *Z. Wasser Abwasser Forsch.* **15** (1982) 1 – 6.
112. Unpublished results, Bayer AG, 1984.
113. U.S. Water quality criteria documents, U.S. Federal Register 45, No. 231, Nov. 28 1980, p. 79331.
114. A. V. Pawliz, R. A. Kent, U. A. Schneider, C. Jefferson, "Canadian Water Quality Guidelines for Chromium," *Environmental Toxicology and Water Quality* **12** (1997) 123 – 183.
115. S. M. Hunt, S. Hedgecott: Revised Environmental Quality Standards for Chromium in Water, WRc Medmenham, Report DoE 2858(P), 1994.
116. H. J. M. Bowens: *Environmental Chemistry of the Elements,* Academic Press, London 1979.
117. *Chrom, Bachsedimente (creek sediments)* , Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover 1984.
118. Umweltbundesamt, Jahresbericht 1982, Berlin 1983, pp. 33, 35.
119. D. J. Swaine, R. L. Mitchell, *J. Soil Sci.* **11** (1960) no. 2, 347 – 368.
120. S. Langard: *Biological and Environmental Aspects of Chromium,* Elsevier, Amsterdam 1982, p. 58.
121. S. Langard, see [120] , p. 60, p. 62.
122. R. A. Anderson, *Sci. Total Environ.* **17** (1981) 13 – 28.
123. W. Mertz, *Proc. Nutr. Soc.* **33** (1974) 307 – 313.
124. W. Merz, K. Schwarz, *Am. J. Physiol.* **196** (1959) 614 – 618.
125. B. L. Horecker, E. Stotz, T. R. Hogness, *J. Biol. Chem.* **128** (1939) 251.
126. W. Mertz, *Physiol. Rev.* **49** (1969) 163.
127. F. J. Diehl in *Zeitschrift des Verbandes Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (VDLUFA), Landwirtschaftliche Forschung* (1982) special issue no. 39, 58.
128. Deutsche Forschungsgemeinschaft: *Forschungsbericht Schadstoffe im Wasser,* Deutsche Forschungsgemeinschaft, Bonn 1982 , p. 411.
129. D. Henschler: *Gesundheitsschädl. Arbeitsstoffe,* Verlag Chemie, Weinheim 1973.
130. S. Fairhurst, C. A. Minty: "Toxicity of chromium and Inorganic Chromium Compounds," *Toxicity Review,* **vol. 21** , HSE, London 1989.
131. Environmental Health Criteria 61, Chromium, WHO, Geneva 1988.
132. Toxicological Profile for Chromium, TP-92/08, U.S. Department of Health and Human Services, Public Health Service, Agency for Toxic Substances and Disease Registry, Atlanta, 1993.
133. U.S. Geological Survey Circular 930-B, 1984.
134. U.S. Bureau of Mines: Mineral Commodity Summaries 1984.
135. U.S. Bureau of Mines: Mineral Commodity Summaries 1985.
136. T. Power: "Chromite – the non metallurgical markets," *Industrial Minerals* 1985, April, 17 – 51.
137. Minerals Bureau, International Report No. 86, Braamfontein.
138. Statistisches Bundesamt Wiesbaden, Fachserie: Außenhandel Reihe 2, Außenhandel nach Waren und Ländern (Spezialhandel) 1983.
139. U. Korallus, H. Lange, A. Neiss, E. Wüstefeld, T. Zwingers, *Arb. Med., Soz. Med., Praev. Med.* **17** (1982) 159 – 167.
140. U. Korallus, C. Harzdorf, J. Lewalter, *Int. Arch. Occup. Environ. Health* **53** (1984) 247 – 250.
141. F. L. Petrilli, S. de Flora, *Appl. Environ. Microbiol.* **33** (1977) 805 – 809.
142. F. L. Petrilli, S. de Flora in: *Mutagens in our Environment,* A. R. Liss, New York 1982, pp. 453 – 464.
143. EPA Health Assessment Document for Chromium, August 1984 (EPA-600/8-83-D14F).
144. W. C. Hueper, W. W. Payne, *Arch. Environ. Health* **5** (1962) 445 – 462.

145. L. S. Levy, Fac. of Sc., University of London, 1975.
146. NIOSH (1975) Criteria for a recommended standard. Occ. exposure to chromium(VI), US DHEW, Washington, D.C.
147. D. Steinhoff, C. Gad, G. T. Hatfield, U. Mohr, *Exp. Pathol.* **30** (1986) no. 3, 129 – 141.
148. EC Commission Directive 96/64/EC of July 30, 1996 on the classification, packing and labelling of dangerous substances; 22nd adaptation of EEC Directive 67/548.
149. American Conference of Governmental Industrial Hygienists (ACGIH): Threshold Limit Values, Biological Exposure Indices, ACGIH, Cincinatti, OH 1996.
150. U. Korallus, H. Ehrlicher, E. Wüstefeld, *Arb. Med., Soz. Med., Praev. Med.* **3** (1974) 51 – 54;76 – 79; 248 – 252.
151. W. Machle, F. Gregorius, *U.S. Publ. Hlth. Rep.* **63** (1948) no. 35, 1114 – 1127.
152. T. Mancuso, W. C. Hueper, *Ind. Med. Surg.* **20** (1951) 358 – 363.
153. M. R. Alderson, N. S. Rattan, L. Bidstrup, *Br. J. Ind. Med.* **38** (1981) 117 – 124.
154. R. B. Hayes, A. M. Lilienfeld, L. M. Snell, *Int. J. Epidemiol.* **8** (1979) 365 – 374.
155. W. J. Hill, W. S. Ferguson, *JOM J. Occup. Med.* **21** (1979) 103 – 106.
156. H. Satoh, Y. Fukuda, K. Teric, N. Chike Katsumo, *JOM J. Occup. Med.* **23** (1981) 835 – 838.
157. J. M. Davies, *Br. J. Ind. Med.* **41** (1984) 148 – 169.
158. R. Frentzel-Beyme, *J. Cancer Res. Clin. Oncol.* **105** (1983) 183 – 188.
159. S. Langard, T. Norseth, *Br. J. Ind. Med.* **32** (1975) 62 – 65.
160. H. Royle, *Environ. Res.* **10** (1975) 141 – 163.
161. G. Axelsson, R. Rylander, A. Schmidt, *Br. J. Ind. Med.* **37** (1980) 121 – 127.
162. S. Langard, A. Andersen, B. Gylseth, *Br. J. Ind. Med.* **37** (1980) 114 – 180.
163. Deutsche Forschungsgemeinschaft: *MAK- und BAT-Werte-Liste 1996,* **Mitteilung 32** , VCH Verlagsgesellschaft, Weinheim 1996.
164. J. Lewalter, U. Korallus, C. Harzdorf, H. Weidemann, *Internat. Arch. Occ. Environ. Health* **55** (1985) 305 – 318.

Cigarettes and Cigars → Tobacco
Cinnamaldehyde → Flavors and Fragrances

Cinnamic Acid

DOROTHEA GARBE, Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Occurrence in Nature** 2
5. **Production** 2
5.1. **Perkin Reaction** 2
5.2. **Claisen Condensation** 2
5.3. **Benzal Chloride Reaction** 2
5.4. **Oxidation of Cinnamaldehyde** 2
5.5. **Preparation of Cinnamic Acid Esters** 3
6. **Uses** 3
7. **Quality Specifications** 3
8. **Economic Aspects** 3
9. **Toxicology** 3
10. **References** 3

1. Introduction

Cinnamic acid [*621-82-9*], 3-phenylpropenoic acid, $C_9H_8O_2$ M_r 148.16, exists in cis and trans forms.

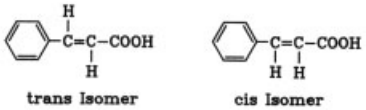

The more stable isomer is the trans isomer, which occurs naturally and is the usual commercial product.

Cinnamic acid was first isolated as crystals from cinnamon oil by TROMMSDORF in 1780. He thought it was benzoic acid. DUMAS and PÉLIGOT identified it in 1835, and in 1856 BERTAGNINI succeeded in synthesizing it from benzaldehyde and acetyl chloride.

2. Physical Properties

trans-Cinnamic acid [*140-10-3*] forms colorless crystals with a faint balsamic odor, *mp* 133 °C, *bp* 300 °C (101.3 kPa) and 173 °C (13.3 kPa), d_4^4 1.2475, d_4^{180} 1.0270. The heat of combustion is 30.5 kJ/g. *trans*-Cinnamic acid is only slightly soluble in water, but it is highly soluble in polar organic solvents.

cis-Cinnamic acid [*102-94-3*] exists in three crystalline forms: *cis*-allocinnamic acid, *mp* 68 °C, and two *cis*-isocinnamic acids, *mp* 58 °C and 42 °C. The lower melting *cis*-isocinnamic acid is extremely unstable and isomerizes spontaneously to *cis*-allocinnamic acid. The three *cis*-cinnamic acids can be interconverted, e.g., by inoculating their melts.

3. Chemical Properties

Cinnamic acid undergoes reactions typical of a carboxyl group and an olefinic double bond. The carboxyl group can be esterified to form cinnamates, some of which are important flavorings and fragrances. When reacted with inorganic acid chlorides, such as thionyl chloride or phosphorus chlorides, cinnamic acid gives cinnamoyl chloride [*102-92-1*]. When heated, cinnamic acid forms styrene [*100-42-5*] and carbon dioxide. With oxidizing agents or when heated with alkali, the olefinic double bond cleaves to give benzaldehyde [*98-87-3*].

Hydrogenation of cinnamic acid leads to 3-phenylpropionic acid [*501-52-0*] or 3-cyclohexylpropionic acid [*701-97-3*], depending on the reaction conditions. The addition of bromine to the olefinic double bond of cinnamic acid forms 2,3-dibromo-3-phenylpropionic acid [*6286-30-2*], an intermediate in the manufacture of the fragrant substance β-bromostyrene [*103-64-0*]. Hydrohalogenation gives 2-halo–3-phenylpropionic acids. Dimers, such as truxinic acid (3,4-diphenyl-1,2-cyclobutanedicarboxylic acid) [*4482-52-4*] and α-truxillic acid (2,4-diphenyl-1,3-cyclobutanedicarboxylic acid) [*490-20-0*], are formed from cinnamic acid under the influence of sunlight. Cinnamic acids with substituents on the benzene ring are seldom prepared by direct substitution reactions of cinnamic acid.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_099

4. Occurrence in Nature

Cinnamic acid, usually the trans isomer, occurs naturally in the free state and as an ester, e.g., in essential oils. Cinnamic acid is a constituent of cassia oil, of the extract oil from *Populus balsamifera,* and of the essential oils from leaves and peels of *Citrus bigaradia*. Cinnamates, especially methyl- [*1754-62-7*], benzyl- [*103-41-3*], cinnamyl- [*122-69-0*], and hydrocinnamyl cinnamates [*122-68-9*], are found in numerous essential and balsam oils (methyl cinnamate in oils of the *Alpinia* species and varieties of *Ocimum canum;* benzyl, cinnamyl, and hydrocinnamyl cinnamates in Peru, Tolu, and storax balsam oils). Siam and Sumatra benzoin [*9000-73-1*] resins also contain cinnamates.

5. Production

Commercial synthesis of cinnamic acid almost always results in the trans isomer.

5.1. Perkin Reaction

The Perkin reaction is the oldest known method of producing cinnamic acid commercially. In this reaction benzaldehyde [*100-52-7*] is condensed with acetic anhydride in the presence of sodium acetate as catalyst.

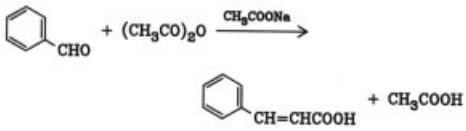

Other possible catalysts are potassium acetate, tertiary amines, potassium phosphate [2] and trimethyl borate [3].

5.2. Claisen Condensation

The Claisen condensation of benzaldehyde with acetic acid esters in the presence of alkali alcoholates yields cinnamic acid esters, which can be saponified to form cinnamic acid.

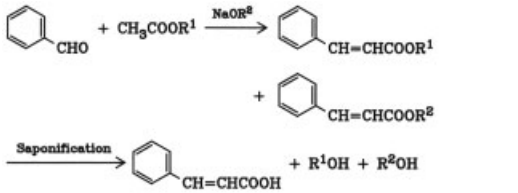

where R^1 and R^2 may be the same or different alkyl groups

5.3. Benzal Chloride Reaction

Benzal chloride [*98-87-3*] reacts with alkali acetate in an alkaline medium to give a high yield of cinnamic acid. Cinnamic acid can be obtained by this reaction in the presence of amines such as pyridine in more than 80 % yield [4].

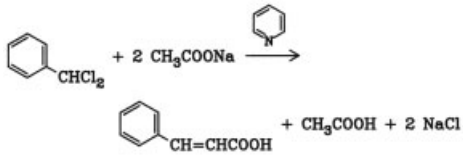

5.4. Oxidation of Cinnamaldehyde

Cinnamaldehyde [*104-55-2*] can be oxidized to cinnamic acid by using oxygen and a catalyst, e.g., silver [5] or palladium on charcoal [6].

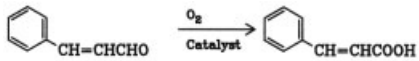

5.5. Preparation of Cinnamic Acid Esters

Cinnamic acid esters can be prepared from styrene, alcohols, carbon monoxide, and oxygen by using catalysts such as platinum [7] or palladium [8].

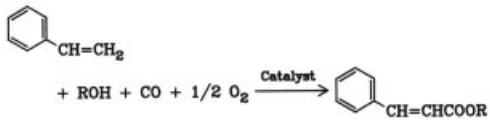

6. Uses

Cinnamic acid is an important intermediate in the preparation of its esters, which are used

as fragrances, for pharmaceuticals, and for the enzymatic production of L-phenylalanine, the starting material for peptide sweeteners. Sodium cinnamate is a known corrosion inhibitor. Cinnamic acid is also used as a brightener in cyanide-free zinc electroplating baths [9], a corrosion inhibitor during removal of scale from zinc [10] and in aerosol cans [11], a low-toxicity heat stabilizer for poly(vinyl chloride) [12], a cross-linking agent for dimethyl terephthalate – ethylene glycol copolymer [13] and polyurethanes [14], a fireproofing agent for polycaprolactam [15], in laundry-resistant polyurethane adhesives for polyester fibers [16], and for improvement of the storage stability of drying-oil-modified alkyd resin coatings [17].

7. Quality Specifications

Cinnamic acid used in fragrances and foods must be 99 % pure after drying. The melting point should be 133 °C, and the chlorine content not more than 0.005 wt % [18].

8. Economic Aspects

The annual worldwide production of cinnamic acid has increased to several thousand tons due to the steadily increasing demand for artificial dipeptide sweeteners (→ Sweeteners). Bayer (Federal Republic of Germany), DSM (the Netherlands), and Kay Fries (United States) are among the major producers.

9. Toxicology

The acute oral LD_{50} in rats is 2.5 g/kg, and the acute dermal LD_{50} in rabbits exceeds 5 g/kg. Cinnamic acid applied neat to intact or abraded rabbit skin for 24 h was slightly irritating; a 4 % solution in petrolatum produced no sensitization in man [19].

10. References

General References

1. E. Gildemeister, F. Hoffmann: *Die Ätherischen Öle,* Akademie Verlag, Berlin 1966.

Specific References

2. Agency of Industrial Sciences and Technology, JP 76 01 026, 1967 (H. Tanaka, M. Tsuda); see *Chem. Abstr.* **85** (1976) 142845h.
3. Y. Y. Makarov-Zemlyanskii, V. V. Malyavkin, SU 175 975, 1962; see *Chem. Abstr.* **64** (1966) 8095h.
4. Kureha Chemical Ind. Co., JP-Kokai 73 081 830, 1973 (K. Shinoda, K. Kobayashi); see *Chem. Abstr.* **80** (1974) 108194g.
5. Union Carbide and Carbon Corp., FR 1 113 852, 1954.
6. Rhône-Poulenc, FR 2 069 836, 1969.
7. Ube Ind., JP-Kokai 82 070 836, 1980; see *Chem. Abstr.* **97** (1982) 109703k.
8. Mitsubishi Chem. Ind., JP-Kokai 81 071 039, 1979; see *Chem. Abstr.* **95** (1981) 186892q.
9. M and T Chemicals Inc., GB 2 009 790, 1977 (D. A. Arcilesi).
10. F. Boehm, Erzeugung Chem. Produkte GmbH, DE 2 405 861, 1974 (W. Boehm).
11. Seitetsu Kagaku Co., JP-Kokai 81 082 872, 1979; see *Chem. Abstr.* **95** (1981) 208175e.
12. Adeka Argus Chemical Co., JP-Kokai 79 011 948, 1977 (M. Ito, T. Aoki); see *Chem. Abstr.* **90** (1979) 187987a.
13. Teijin Ltd., JP-Kokai 79 011 997, 1977 (H. Inata, T. Morinage, M. Ogasawara); see *Chem. Abstr.* **92** (1980) 95081p.
14. Gambro Dialysatoren GmbH & Co. KG, JP-Kokai 79 060 398, 1977; see *Chem. Abstr.* **91** (1979) 92517d.
15. Ube Industries, JP-Kokai 78 133 256, 1977 (Y. Sasaki, T. Fujimoto, M. Yamanaka); see *Chem. Abstr.* **90** (1979) 122557r.
16. Honny Chemicals Co., JP 79 038 131, 1971 (T. Ichinomiya, A. Imaizumi); see *Chem. Abstr.* **92** (1980) 148084a.
17. Meidensha Co., JP 71 001 709, 1966 (T. Horiuchi, K. Yonahara, K. Tokubata); see *Chem. Abstr.* **76** (1972) 47458q.
18. *Food Chemicals Codex 1981,* National Academy Press, Washington, D.C., pp. 364 – 365.
19. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **16** (1978) 687.

Cinnamyl Alcohol → Flavors and Fragrances
Citraconic Acid → Maleic Acid and Fumaric Acid
Citral → Flavors and Fragrances

Citric Acid

FRANK H. VERHOFF, Miles Laboratories, Elkhart, Indiana 46515, United States

1.	**Introduction**	1
2.	**Physical Properties**	2
3.	**Chemical Properties**	3
4.	**Production**	4
4.1.	**Laboratory Synthesis**	4
4.2.	**Early Production**	4
4.3.	**Fermentation**	4
4.3.1.	Surface Fermentation	4
4.3.2.	Submerged Fermentation	5
4.3.3.	Other Fermentation Methods	5
4.4.	**Recovery**	5
5.	**Quality Specifications and Uses**	6
6.	**Economic Aspects**	6
7.	**References**	6

1. Introduction

Citric acid [*77-92-9*], $C_6H_8O_7$, M_r 192.13, is a tricarboxylic acid. According to IUPAC-nomenclature, citric acid is 2-hydroxypropane-1,2,3-tricarboxylic acid; it is also known as β-hydroxytricarballylic acid.

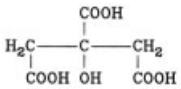

Citric acid was isolated and crystallized in 1784 by SCHEELE, who precipitated calcium citrate by adding lime to lemon juice. The presence of three carboxyl and one hydroxyl groups was recognized by LIEBIG in 1838. Citric acid was first prepared from calcium citrate in 1860 in the United Kingdom. By 1880 citric acid was being produced from calcium citrate in France, Germany, and the United States.

Citrus fruits contain citric acid in large quantities, from 5 % in the fruit [1] to 9 % in the juice [2]. Citric acid is also found in many other fruits. The citric acid content (in mg/100 mL) of various common fruits is as follows [3], [4]:

Lime	7000
Lemon	5630
Raspberry	2480
Black currant	1170
Tomato	1018
Pineapple	605
Strawberry	580
Cranberry	202
Apple	14

In addition to fruits citric acid is found in almost all plant and animal species. The *citric acid cycle,* also known as the Krebs cycle or tricarboxylic acid cycle (see Fig. 1), is pivotal in the oxidation of sugars and acetates to carbon dioxide and water, releasing energy for physiological functions [5], [6].

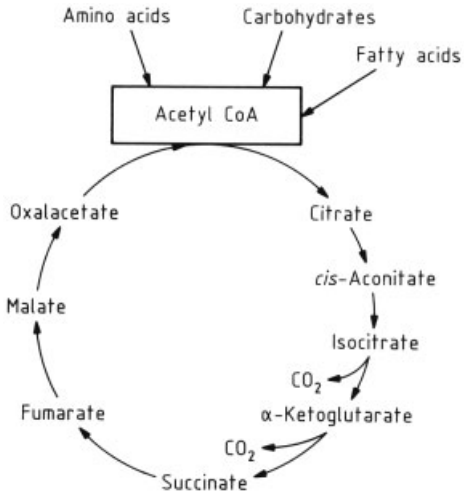

Figure 1. Citric acid cycle of plants and animals [5], [6]

About 100 years after citric acid had been isolated from citrus fruits, it was synthesized from glycerol. The finding, in 1893, that certain microorganisms secrete citric acid when grown on sugar media suggested the possibility of producing the acid commercially by fermentation. The first successful operation of this type began 1923 in New York; later large-scale fermentation processes were developed in the former Czechoslovakia, the United Kingdom, Belgium, and Germany. Submerged fermentation has been used since the middle of this century.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_103

Citric acid fermentation, whether surface or submerged, used carbohydrate substrates with certain strains of the mold *Aspergillus niger*. Since the early 1970s, other microorganisms have been introduced, including molds, yeasts, and bacteria, and other substrates, e.g., various carbohydrates and n-alkanes from crude oil.

2. Physical Properties

Citric acid is a colorless, odorless substance with an agreeable, though strongly acid taste. The monohydrate crystallizes from a saturated solution below 36.6 °C and the anhydrous form at higher temperature. The *monohydrate*, M_r 210.13, has a relative density of 1.542 and a heat of combustion at 30 °C of 1973 kJ/mol; it softens at 75 °C and melts at 100 °C. Its translucent, colorless crystals belong to the orthorhombic system. *Anhydrous citric acid, mp* 153 °C, has a relative density of 1.665 and a heat of combustion at 30 °C of 1990 kJ/mol. The crystals are translucent and colorless, belonging to the holohedral class of the monoclinic system. Both forms exist under conditions of normal humidity; dehydration of the monohydrate occurs in very dry air and is more rapid when it is carried out under vacuum in the presence of concentrated sulfuric acid. The anhydrous crystals gradually absorb water in moist air; both crystalline forms clump and harden in moist air.

The solubility of citric acid in water as a function of temperature is shown in Figure 2; the temperature of the solution drops as the acid dissolves. The pH of citric acid solutions as a function of concentration is shown in Figure 3. The freezing point depression and boiling point elevation of citric acid solutions are given in Table 1. The dissociation constants at 18 °C are K_1 8.2×10^{-4}, K_2 1.8×10^{-5}, and K_3 4.0×10^{-6}.

Table 1. Freezing point depression and boiling point elevation of aqueous citric acid solutions

Concentration, mol/kg H_2O	Freezing point depression, °C	Boiling point elevation, °C
0.01	0.023	–
0.05	0.042	–
0.10	0.203	–
0.50	0.965	0.284
1.00	1.940	0.577
2.00	1.000	1.214
5.00	–	3.512
10.00	–	8.390
20.00	–	16.600

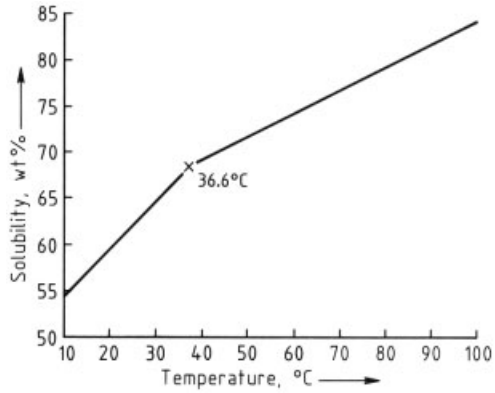

Figure 2. Solubility of citric acid in water as a function of temperature

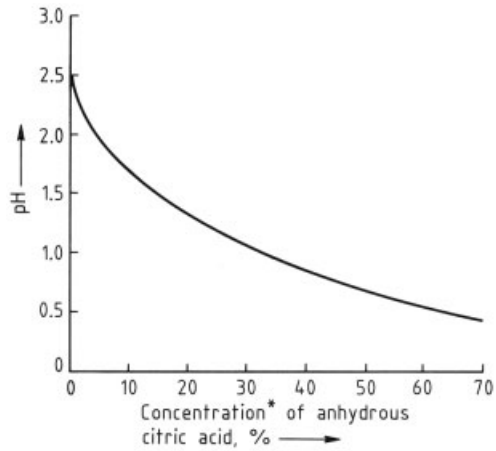

Figure 3. Concentration of citric acid in water vs. pH
∗ Concentration is in (grams per 100 mL solution) × 100 %

Table 2. Solubility of hydrated and anhydrous citric acid in organic solvents at 25° C and density of solutions

Form Solvent	Solubility,∗ g/100 g	d_{25}
Monohydrate		
Amyl acetate	5.980	0.8917
Amyl alcohol	15.430	0.8774
Ethyl acetate	5.276	0.9175
Ether	2.174	0.7228
Chloroform	0.007	1.4850
Anhydrous		
Amyl acetate	4.22	0.8861
Ether (absolute)	1.05	0.7160

∗ Saturated solution.

The solubility of citric acid in organic solvents is listed in Table 2, along with the relative

density of the solutions. The distribution between ether and water is shown in Table 3.

Table 3. Distribution of citric acid between water and ether

Concentration, mol/L		Distribution coefficient
In water	In ether	
At 15 °C		
0.902	0.0077	117
0.460	0.0036	128
0.220	0.0017	129
0.297	0.0023	129
At 25 °C		
0.9175	0.0063	114
0.487	0.0031	155
0.241	0.00155	155
0.315	0.0020	158

3. Chemical Properties

Citric acid loses water at 175 °C to form aconitic acid, HOOCCH=C(COOH) (CH_2COOH), which loses carbon dioxide to yield itaconic anhydride [*2170-03-8*]. Itaconic anhydride rearranges to citraconic anhydride [*616-02-4*] or adds water to form itaconic acid [*97-65-4*], (HOOCCH_2) (HOOC)C=CH_2.

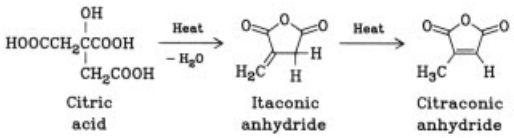

Addition of water to citraconic anhydride gives citraconic acid [*498-23-7*], *cis*-HOOCCH=C(CH_3) (COOH). Evaporation of a citraconicacid solution in the presence of nitric acid yields mesaconic acid [*498-24-8*], the trans isomer of citraconic acid.

Potassium permanganate oxidation of citric acid at 35 °C yields 1,3-acetonedicarboxylic acid [*542-05-2*], HOOCCH_2 $COCH_2$ COOH; at 85 °C the productis oxalic acid [*144-62-7*], HOOCCOOH. Fusion with potassiumhydroxide produces oxalic and acetic [*64-19-7*] acids.

Citric acid forms crystalline mono-, di-, and tribasic salts with manycations. The degree of hydration of these salts varies; thus, trisodium citrate can crystallize with 2 or 5.5 molecules of water. Mixtures of metal cations give complex salts, such as $ZnNa_3H(C_6H_5O_7)_2$, $ZnNa_4(C_6H_5O_7)_2$, and $Zn(Na_3C_6H_4O_7)_2$.

Citric acid forms stable complexes with many metals; some, such as ferroammonium citrates, can be crystallized. Copper gives a complex resembling the copper – tartaric acid complex of Fehling's solution. These complexes have been studied by various classical methods.

Citric acid can chelate many metal ions in solution by forming bonds between the metal and the carboxyl or hydroxyl groups of the citric acid molecule. Sometimes two or more citric acid molecules are involved in the interaction with the metal ion. This property is valuable for preventing precipitation (e.g., bathtub ring from soap), changing chemical potential, and altering other chemical properties.

Citric acid is easily esterified with many alcohols under the usual conditions in the presence of a catalyst, such as sulfuric acid, *p*-toluenesulfonic acid, or an acid ion-exchange resin; esterification with alcohols boiling above 150 °C requires no catalyst. Benzyl chloride and sodium citrate yield di- or tribenzyl esters. Trimethyl, triethyl, and tributyl citrate are used as plasticizers in food-packaging materials.

Dihydric alcohols, dihydric phenols, and polyhydric alcohols, such as mannitol, sorbitol, and glycerol, form polyesters with citric acid. In some cases the esterification reaction can be stopped before completion, leaving at least one of the carboxylic acid moieties free to form a salt; the resulting polyester may be soluble in water.

Acid chlorides and anhydrides react with the hydroxyl group of citric acid; usually the ester of citric acid is used and the allyl ester is formed. Dicarboxylic acid chlorides give bis(citric acid esters).

Epoxides, such as ethylene oxide, propylene oxide, and styrene oxide, form polymers by reacting with citric acid or its esters at the available hydroxyl and carboxyl groups.

Ammonia, amines, amides, and carbamides react with citric acid as they do with simple carboxylic acids. Aldehydes form adducts with citric acid; e.g., formaldehyde gives anhydromethylenecitric acid [*144-16-1*] 5-oxo-1,3-dioxolane-4,4-diaceticacid:

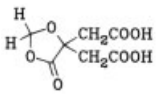

4. Production

4.1. Laboratory Synthesis

The classic laboratory synthesis of citric acid beginning with acetone has not been used on an industrial scale:

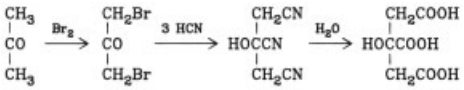

Reaction conditions at each stage are critical and use of bromine and hydrogen cyanide is hazardous. A simplified version of this synthesis has been patented [7].

Another synthesis is as follows:

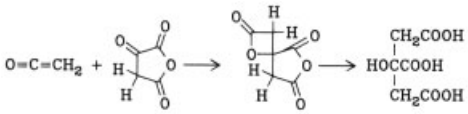

The key step in this route is the catalytic condensation of ketene with oxalacetic anhydride, which can be obtained by catalytic oxidation of malic or fumaric acid.

Other syntheses are found in the patent and scientific literature, but apparently none have been commercialized [8], [9].

4.2. Early Production

For about 50 years, citric acid was obtained exclusively from unripe citrus fruit. In southern Italy, an entire industry with orchards and processing facilities was devoted to this purpose. The juice was pressed from unripe fruit and mixed with lime (CaO) to precipitate calcium citrate. The solid was treated with sulfuric acid to precipitate calcium sulfate and form a solution from which citric acid was crystallized. The yield of purified product was about 2 – 3 wt % of the fruit; citrus oils were also obtained. This procedure has been displaced by fermentation.

4.3. Fermentation

The most economical method for producing citric acid since the 1930s has been fermentation, which employs a strain of *Aspergillus niger* to convert sugar to citric acid. Both surface fermentation and submerged fermentation have been used.

4.3.1. Surface Fermentation

In surface fermentation, *A. niger* is grown on a liquid substrate in pans stacked vertically in a chamber. The chamber and pans are sterilized either before or after introduction of the substrate. The pans are filled manually or automatically. The chamber is warmed by the introduction of moist, sterile air at a controlled temperature. The liquid and the surface microorganisms are removed manually or automatically from the pans. The pans are cleaned before the next batch is introduced.

The substrate for the fermentation is a carbohydrate, usually molasses or a sugar, such as raw beet, refined beet, or cane sugars, or a syrup. Glucose syrups can be prepared from wheat, corn, potato, or other starch. The sugar content of the syrup can vary from about 10 to 25 wt %. Certain inorganic nutrients, such as ammonium nitrate, potassium phosphate, magnesium sulfate, zinc sulfate, and potassium ferrocyanide, are added. The pH is adjusted to between 3 and 7, depending on the carbohydrate source. Sterilization may be batchwise or continuous; the latter uses less energy and is usually faster. After sterilization, the temperature is adjusted as required.

The surface of the sterile substrate in the pans is inoculated with *A. niger* spores, which germinate and cover the surface of the liquid with a mat of mold. After two to three days the surface is completely covered and citric acid production begins, continuing at almost a constant rate until 80 – 90 % of the sugar is consumed. Fermentation then continues more slowly for an additional six to ten days.

The theoretical yield from 100 kg of sucrose is 123 kg of citric acid monohydrate or 112 kg of anhydrous acid. However, the *A. niger* uses some sugar for growth and respiration, and the actual yield varies between 57 and 77 % of theoretical, depending on such factors as substrate purity, the particular strain of organism, and the control of fermentation.

4.3.2. Submerged Fermentation

Submerged fermentation is similar to surface fermentation, but takes place in large fermentation tanks. This method is used more frequently because labor costs are lower with large tanks than with small pans; equipment costs are also lower.

The fermentation vessel can be short and wide or tall and narrow, and equipped with mixing devices, such as top-entering or side-entering agitators of the turbine or propeller type. Agitation can be increased by use of a draft tube, a recirculation loop, or a nozzle through which air and recirculated substrate is pumped. Spargers located at the bottom of the vessel or under the stirrer supply air, which may be enriched with oxygen. Oxygen is usually recovered from the exhaust gas. The air is supplied by a compressor and passes through a sterile filter; if necessary, the air is cooled. Because the process is exothermic, the vessel must be equipped with heat-exchange surfaces, which can be the outside walls or internal coils. Ports are provided for introducing substrate, inoculum, and steam or other sterilizing agents; sampling and exhaust ports are also provided.

The substrate is prepared in a separate tank and its pH adjusted; the micronutrients may be added to this tank or directly to the fermentor. The substrate is sterilized by a batchwise or, more commonly, by a continuous operation. The fermentor is sterilized, charged with substrate, and inoculated. Fermentation requires 3 – 14 days. After it is completed, the air supply is stopped to prevent the microorganisms from consuming the citric acid.

4.3.3. Other Fermentation Methods

Citric acid can also be produced from hydrocarbons using yeast. The process resembles submerged fermentation, but requires more air. It is a four-phase process, i.e., air, organic liquid, aqueous liquid, and solid, and some operational details are different.

4.4. Recovery

The citric acid broth from the surface or submerged fermentation processes must be purified. First, biological solids usually are removed by filtration using a rotary vacuum filter or the more recent belt-press filter, or by centrifugation. The solids are washed to improve recovery of citric acid.

The dissolved citric acid must then be separated from residual sugars, proteins generated by the fermentation, and other soluble impurities. This has traditionally been accomplished by precipitation and crystallization. Addition of lime precipitates calcium citrate, which is filtered and stirred in dilute sulfuric acid to form a precipitate of calcium sulfate; filtration yields a purified citric acid solution. Control of pH and temperature in these operations helps to optimize the results. Citric acid is then crystallized from solution and recrystallized from water; the mother liquors are recycled to remove accumulated impurities.

Other purification techniques can be used, including ion exchange, carbon adsorption, membrane filtration, chromatography, and liquid extraction.

After filtration or centrifugation, the citric acid crystals are dried, classified, and sold in bags or in bulk. Crystals of the wrong size are recycled. Citric acid is also sold as an aqueous solution of varying degrees of purity.

Byproducts of these purification processes include biomass from the primary filtration, which can be sold for feed, and calcium sulfate, which can be used for landfill or for industrial purposes, such as manufacture of gypsum board. Soluble organic impurities are given normal wastewater treatment.

Few changes in production methods have occurred in recent years, but research has been done on the fermentation process [10] and on new separation techniques [11].

5. Quality Specifications and Uses

Specifications. The Food Chemical Code specifications for citric acid are as follows [12]:

> *Description:* colorless, translucent crystals or white, granular to fine crystalline powder; odorless, strongly acid taste; hydrate is efflorescent in dry air

Solubility: 1 g of anhydrous acid dissolves in 0.5 mL of water, in 2 mL of alcohol, and in 30 mL of ether
Identification: 1-in-10 solution gives positive tests for citrate
Assay: not less than 99.5 % $C_6H_8O_7$, calculation based on the anhydrous form
Water: not more than 0.5 % (anhydrous form) and not more than 8.8 % (hydrate)
Limits of impurities: arsenic (as As): not more than 3 ppm (0.0003 %)
heavy metals (as Pb): not more than 10 ppm (0.001 %)
oxalate: passes test; no turbidity produced
readily carbonizable substances: passes test
residue on ignition: not more than 0.05 %

For specific test methods, see [12].

Uses. Citric acid and its sodium or potassium salts are used in many food products, including carbonated beverages, dry-packaged drinks, fruit drinks, jams, jellies, and canned fruits. They are also important in producing vegetable oils and in preserving the color, flavor, and vitamin content of fresh and frozen vegetables and fruit.

Citric acid is widely used in cleaning, e.g., for boilers and heat exchangers. Its chelating ability assists in removal of scale. Citric acid is employed as a detergent builder, especially in liquid formulations. Its solutions can remove sulfur dioxide from gases and chelate micronutrients in fertilizer. Citric acid esters are plasticizers for food-grade plastic containers.

6. Economic Aspects

Citric acid is manufactured worldwide. The United States and Western Europe are the major producers; however, capacity has recently increased in the Far East and in Eastern Europe. The location of new facilities depends on access to raw materials and markets.

7. References

1. H. Moellering, W. Gruber, *Anal. Biochem.* **17** (1966) 369 – 376.
2. H. Fricke, S. B. Hensen, *Food Process. Ind.* 1975, 38 – 44.
3. J. J. Ryan, J. A. Du Pont, *J. Agri. Food Chem.* **21** (1973) 45 – 49.
4. K. N. Paulson, M. A. Stevens, *J. Food Sci.* **39** (1974) 354 – 357.
5. H. A. Krebs, *Angew. Chem.* **66** (1954) 313 – 319.
6. A. L. Lehninger: *A Short Course in Biochemistry,* Worth Publ., New York 1973, p. 217.
7. Montecatini Edison, DE-OS 2 245 892, 1972.
8. Chevron Research Corp., DE-OS 2 225 986, 1972.
9. Ethyl Corp., DE-OS 2 240 723, 1972.
10. M. Rohr, C. P. Kubicek, J. Kominek: "Citric Acid" in H. Dellweg (ed.): *Biotechnology Microbiology Products, Biomass, and Primary Products,* vol. 3, Verlag Chemie, Weinheim 1983, pp. 456 – 465.
11. D. T. Friesen et al.: "Separation of Citric Acid from Fermentation Beer Using Supported Liquid Membranes", Paper no. 194, Am. Chem. Soc. Meeting, New York, April 1984.
12. *Food Chemicals Codex*, 3rd. ed., National Academy Press, Washington, D.C., 1981.

Citronellal → Flavors and Fragrances
Citrus Juice → Beverages, Nonalcoholic
Cladding of Metals → Metals, Surface Treatment
Claus Process → Sulfur

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Clays

Standard Article

Haydn H. Murray[1]

[1]Indiana University, Department of Geological Sciences, Bloomington, Indiana, United States

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a07_109.pub2
Article Online Posting Date: December 15, 2006

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Structure and Composition of Clay Minerals
3. Geology and Occurrence of Major Clay Deposits
3.1. Kaolin
3.2. Smectite
3.3. Palygorskite and Sepiolite
3.4. Miscellaneous Clays
4. Mining and Processing
4.1. Kaolin
4.1.1. Dry Process
4.1.2. Wet Process
4.1.3. Combined Dry and Wet Processing
4.1.4. Special Processes
4.2. Smectite
4.3. Palygorskite and Sepiolite
5. Properties and Uses
5.1. Kaolin
5.1.1. Physical Properties
5.1.2. Paper
5.1.3. Paint
5.1.4. Ceramics
5.1.5. Rubber
5.1.6. Plastics
5.1.7. Ink
5.1.8. Catalysts
5.1.9. Fiberglass
5.2. Smectite
5.3. Palygorskite and Sepiolites
6. Environmental Aspects
7. Production and Consumption

Clays are important industrial minerals that are utilized around the world. The clay minerals and materials discussed here are kaolins, smectites, palygorskite and sepiolite, and miscellaneous or common clays. Kaolin is both a rock term and a group minerals name for kaolinite, dickite, nacrite, and halloysite. Bentonite is a rock term and the dominant minerals in bentonite are smectites, which is a group mineral name. Important minerals in the smectite group are sodium, calcium, magnesium, iron, and lithium montmorillonites. Miscellaneous or common clays include shales, mudstones, soil clays, lacustrine clays, glacial tills, and so on, which are used for making brick, tile, other structural clay products, and stoneware pottery. They are usually comprised of a mixture of clay and nonclay minerals, the most common of which are illite, chlorite, mixed-layer clays, and quartz. The geology and location of clay deposits that are marketed regionally and worldwide are described. Structure,

composition, and chemical and physical properties of the clay minerals are discussed. These control the many industrial applications of clay minerals and materials. Important applications are described. World production figures of the various clays are given as of 2004. A brief discussion on health-related and environmental issues is included.

[Top of Page]

1. Introduction

The term *clay* has a double meaning and, therefore, should be defined when it is used. Clay is used both as a rock term and as a particle size term. As a rock term, clay is used for a natural, earthy, fine-grained material composed largely of a limited group of crystalline minerals known as the clay minerals. As a particle size term, clay is used for the category that includes the smallest particles. Soil investigators and mineralogists generally use 2 μm as the maximum size, although the Wentworth scale [1] defines clay as material finer than 4 μm. GRIM [2] uses the term clay material for any fine-grained, natural, earthy, argillaceous material; in this way, the term can include clays, shales, or argillites, and some soils if they are argillaceous.

Clay is an abundant raw material and has an amazing variety of uses and properties that depend on the clay mineral composition and other factors enumerated by GRIM [2]. These factors are clay mineral composition, nonclay mineral composition, presence of organic material, the type and amount of exchangeable ions and soluble salts, and texture.

Clays are comprised of certain groups of hydrous aluminum, magnesium, and iron silicates that may contain sodium, calcium, potassium, and other ions. These silicates are called the clay minerals, and the major clay mineral groups are kaolins, smectites, illites, chlorites, and palygorskite and sepiolite. The specific clay minerals are identified by several techniques, including X-ray diffraction [3], differential thermal analysis [4], electron microscopy [5], and infrared spectrometry [6]. Identification and quantification of the clay minerals and nonclay minerals present in a clay material is important because the uses and engineering properties are controlled largely by these two factors.

Clays and clay minerals are important geologically, industrially, and agriculturally; they are important to the engineer in building foundations, tunnels, road cuts and fills, and dams. There are several scientific groups whose work is largely devoted to clays, including AIPEA (Association Internationale pour l'Étude des Argiles), Clay Minerals Group of the Mineralogical Society of Great Britain, CMS (Clay Minerals Society, United States), European Clay Minerals Society, Clay Minerals Society of Japan, and the Australian Clay Minerals Society. Major publications include the *Proceedings of International Clay Conferences*, published by AIPEA; *Clays and Clay Minerals*, published by CMS; *Clay Minerals*, published by the Mineralogical Society of Great Britain; *Clay Mineral Science*, published by the Clay Society of Japan; and *Applied Clay Science*, published by Elsevier.

[Top of Page]

2. Structure and Composition of Clay Minerals

GRIM [2] proposed a tentative classification based on the structural attributes of the various clay minerals. This classification is relatively simple:

- I. Amorphous
  - Allophane group [12172-71-3]
- II. Crystalline
  - A. Two-layer types (sheet structures composed of units of one layer of silica tetrahedrons and one layer of alumina octahedrons)
    - 1. Equidimensional
      - Kaolin group [1332-58-7]
        - kaolinite [1318-74-7],
        - dickite [1318-45-2],
        - nacrite [12279-65-1]
    - 2. Elongate
      - halloysite [12244-16-5]
- III. Three-layer types (sheet structures composed of two layers of silica tetrahedrons and one central dioctahedral or trioctahedral layer)
  - 1. Expanding lattice
    - Smectite group [12199-37-0]
    - a. Equidimensional
      - montmorillonite [1318-93-0],
      - sauconite [12424-32-7],
      - vermiculite [1318-00-9]
    - b. Elongate
      - nontronite [12174-06-0],
      - saponite [1319-41-1],
      - hectorite [12173-47-6]
  - 2. Nonexpanding lattice
    - Illite group [12173-60-3]
- IV. Regular mixed-layer types (ordered stacking of alternate layers of different types)
  - Chlorite group
- V. Chain-structure types (hornblende-like chains of silica tetrahedrons linked together by octahedral groups of oxygens and hydroxyls containing aluminum and magnesium ions)
  - palygorskite [12174-11-7]
  - (attapulgite [1337-76-4]),

sepolite [15501-74-3]

The materials described as *allophane* are amorphous [7] and contain variable proportions of silica, alumina, and water. Allophane is commonly associated with halloysite and generally has a glassy appearance.

The kaolin minerals [8] are hydrous aluminum silicates and have the approximate composition $2\ H_2O \cdot Al_2O_3 \cdot 2\ SiO_2$. *Kaolinite* is the most common of the kaolin minerals. The structure of kaolinite consists of a single silica tetrahedral sheet and a single alumina octahedral sheet combined to form the kaolin unit layer (Fig. 1). These unit layers are stacked on top of each other. Variations in orientation of the unit layers in stacking cause differences in the kaolin mineral itself and lead to the differentiation of *nacrite* and *dickite* in the kaolin group [2]. Platy and vermicular forms of kaolinite are shown in the electron micrograph in Figure 2.

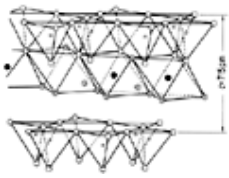

Figure 1. Structure of kaolinite

○ Oxygen; ⊙ Hydroxyl; ○ Silicon tetrahedrally coordinated; ● Aluminum octahedrally coordinated

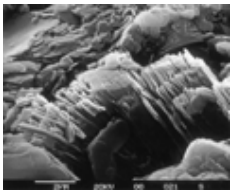

Figure 2. Scanning electron micrograph of kaolinite stacks and plates

Halloysite is an elongate mineral that occurs in a hydrated and dehydrated form [7]. BATES et al. [9] proposed that the hydrated form of halloysite consists of curved sheets of kaolin unit layers (Fig. 3), which is consistent with electron micrographs (Fig. 4).

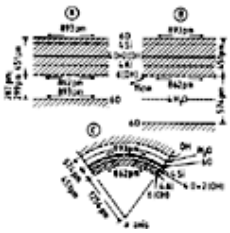

Figure 3. Structure of halloysite

A) Dehydrated halloysite; B) Hydrated halloysite; C) Proposed curved structure of halloysite

Figure 4. Scanning electron micrograph of elongate halloysite (bar = 2 μm)

Smectite is composed of units consisting of two silica tetrahedral sheets with a central alumina octahedral sheet. The lattice has an unbalanced charge because of substitution of alumina for silica in the tetrahedral sheet and iron and magnesium for alumina in the octahedral sheet. Because of this and because oxygen layers are adjacent when these unit layers are stacked, the attraction holding these layers together is weak, and cations and polar molecules can enter between the layers and cause expansion (Fig. 5). Sodium *montmorillonite* is a smectite in which sodium and water molecules are the interlayer material, and calcium montmorillonite is a smectite in which calcium and water molecules are the interlayer material. Figure 6 shows an electron micrograph of a sodium montmorillonite.

Figure 5. Structure of smectite

○ Oxygen; ⊙ Hydroxyl; ● Aluminum, iron, magnesium; ○• Silicon, occasionally aluminum

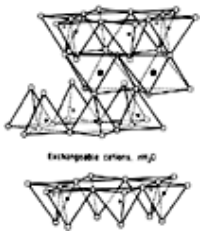

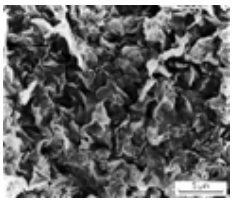

Figure 6. Scanning electron micrograph of a sodium montmorillonite (Wyoming)

Sauconite is a smectite in which zinc has replaced some of the aluminum. *Nontronite* results when iron replaces the aluminum, and *saponite* results when magnesium replaces aluminum. *Hectorite* results when magnesium replaces aluminum and lithium is the ion between the layers.

Vermiculite is an equidimensional expandable clay mineral. The structure [10] consists of a layer made up of two silica tetrahedral sheets with a central magnesium – iron octahedral sheet separated by two layers of water molecules and adsorbed magnesium and calcium ions. Vermiculite differs from montmorillonite in that the layers do not expand as much and show less random stacking [2].

Illite is a general term for the micalike clay mineral [11]. The basic structural unit is a layer composed of two silica tetrahedral sheets and a central octahedral sheet. The unit is similar to that of montmorillonite except that more aluminum ions replace silicon in the tetrahedral sheet, which causes a charge deficiency. This charge deficiency is balanced by potassium ions that act as a bridge between the unit layers and bind them together so that illites are nonexpandable (Fig. 7).

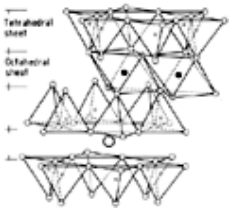

Figure 7. Structure of illite

○ Potassium; ○ Oxygen; ● Aluminum (only 2/3 of available positions filled); ⊙ Hydroxyl; ○ Silicon (1/4 replaced by aluminum)

Another common group of clay minerals is the *chlorite* minerals. The structure consists of alternate mica layers and brucite layers (Fig. 8). The brucite-like layer can be a magnesium, aluminum, or iron hydroxide, or a combination of all three. The various members of the chlorite group are differentiated on the kind and amount of substitution in the brucite and the mica layers [12].

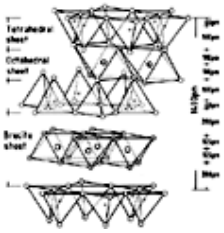

Figure 8. Structure of chlorite

○ Magnesium, iron (all available positions filled); ⊙ Hydroxyl; ○ Oxygen; ° Silicon (1/4 replaced by aluminum)

The clay minerals *palygorskite* (attapulgite) and *sepiolite* have a chainlike structure (Fig. 9). Palygorskite and attapulgite [13] are synonymous, palygorskite being the preferred name. The differentiation between sepiolite and palygorskite is the substitution of iron for magnesium, which causes some minor structural differences. These minerals consist of double silica tetrahedral chains linked together by octahedral oxygen and hydroxyl groups containing aluminum and magnesium ions. An electron micrograph of palygorskite is shown in Figure 10.

Figure 9. Structure of palygorskite (attapulgite), $(OH_2)_4(OH)_2Mg_5Si_8O_{20} \cdot 4\ H_2O$

○ H_2O; Dashed circle: Hydroxyl; ○ Magnesium or aluminum; ⊙ OH_2; ○ Oxygen; ● Silicon

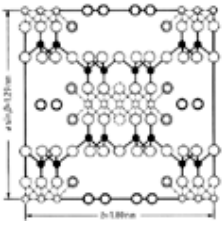

Figure 10. Scanning electron micrograph of palygorskite (attapulgite)

Mixed-layer clays are mixtures of the unit layers on a layer-by-layer basis and are relatively common in nature. Illite – smectite, illite – chlorite, smectite – chlorite, vermiculite – illite, and even smectite – kaolinite have been described. Mixed-layer clays with both random and regular layering were treated by REYNOLDS [14].

[Top of Page]

3. Geology and Occurrence of Major Clay Deposits

Certain clays, such as kaolins, smectites, and palygorskite and sepiolite, can be found in relatively pure occurrences or can be beneficiated to relatively high purity. These particular clays have special industrial applications. Illites and chlorites are usually in shales or clays that are used for structural clay ceramics; they are not utilized where the application depends on the particular physical or chemical properties of the individual clay mineral.

3.1. Kaolin

Kaolin is a rock term, a clay mineral group, and an industrial mineral commodity. It is used interchangeably with the term china clay. The term kaolin is derived from the Chinese word *kauling*, meaning high ridge, the name for a hill near Jauchau Fu in China where a kaolin clay was mined for many centuries. Kaolin is a clay that

1. consists of substantially pure kaolinite or related clay minerals;
2. is naturally or can be beneficiated to be white or nearly white;
3. fires white or nearly white; and
4. is amenable to beneficiation by known methods to make it suitable for use in whiteware, paper, rubber, paint, and similar uses.

The term is applied without direct relation to the purity of deposits. Many large kaolin deposits are essentially pure and require little concentration during preparation for market. Most, however, are slightly off-color and require beneficiation to improve the brightness and whiteness and to control the particle size. Other kaolins contain as little as 10 % kaolinite; these must be washed and concentrated to recover a marketable product.

Kaolin deposits are classed as primary or secondary. Primary or residual kaolins are those that have formed by the alteration of crystalline rocks such as granite and remain in the place where they formed. Secondary deposits of kaolin are sedimentary and have been transported from their place of origin and deposited in beds or lenses associated with other sedimentary rocks such as sands. Kaolins occur on every continent, but only a few (Georgia, England, Brazil, Czechoslovakia, Eastern Germany, Spain, the former Soviet Union, and Australia) can be mined and beneficiated to meet the rigid industrial specifications required for use as a filler, extender pigment, or ceramic raw material. Only those in Georgia, England, and Brazil can meet specifications for coating clays for paper. Over half of the world's kaolin is produced in the United States and England.

The largest kaolin producing district in the world is located in Georgia and South Carolina in the United States. These deposits are late Cretaceous and early Tertiary [15] (Table 1) and occur in lenses and beds in relatively coarse sands (Fig. 11). The kaolin was derived from granitic rocks on the Piedmont plateau that were altered by weathering over a long period of geologic time. During the Cretaceous period, the Piedmont was uplifted, and the weathered residue consisting of kaolinite, quartz, feldspar, mica, traces of iron and titanium minerals, and organic matter was transported by streams onto a deltaic platform where it was deposited as sinuous lenticular deposits in lakes, oxbows, and swamps along the Cretaceous shoreline. The environment in which the deposition took place was similar to the one now existing along the coast of Georgia and South Carolina, with deltas, offshore bars, swampy, muddy lagoons, and winding estuaries. Extensive diagenetic alteration by ground water movement and oxidation has taken place since deposition. The deposits are concentrated in a belt ca. 30 km wide extending from central Georgia northeastward into South Carolina south of the fall line (Fig. 12).

Table 1. Geologic eras, periods, and epochs

Eras	Periods	Epochs	Years ago
Cenozoic	Quaternary	Holocene	
		Pleistocene	
	Tertiary	*Pliocene*	
		Miocene	
		Oligocene	
		Eocene	
		Paleocene	70×10^6
Mesozoic	*Cretaceous*	*Upper*	
		Lower	
	Jurassic		160×10^6
	Triassic		
Paleozoic	Permian		230×10^6
	Pennsylvanian		
	Mississippian		
	Devonian		
	Silurian		
	Ordovician		
	Cambrian		

The terms in italics are mentioned in the text.

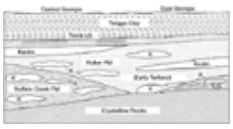

Figure 11. Diagrammatic sketch showing kaolin beds (K) in sands of Cretaceous (Buffalo Creek Fm.) and Tertiary periods (Huber Fm.)

Figure 12. Location of kaolin deposits in Georgia and South Carolina

These minable sedimentary deposits range in thickness from 3 to 15 m and extend laterally from a few hundred meters to kilometers in length and width. The reserves of the coarser Cretaceous kaolins are being rapidly depleted and will probably last only another twenty or thirty years. However, the finer Tertiary kaolins are more abundant and the reserves will last for over a hundred years. The relatively coarse Cretaceous kaolins which are known as soft kaolins have a particle size that ranges from about 60 to 70 % finer than 2 µm, and the finer Tertiary kaolins which are known as hard kaolins range from about 85 to 95 % finer than 2 µm. In 2005, about 6.2×10^6 t of kaolin were produced in Georgia and South Carolina. The Georgia and South Carolina kaolins generally contain about 85 – 95 % kaolinite. The other minerals present are quartz, muscovite, biotite, smectite, ilmenite, anatase, rutile, leucoxene, goethite, and traces of zircon, tourmaline, kyanite, and graphite.

The English kaolins that occur in southwest England in Cornwall (Fig. 13) are primary. These kaolins are the result of alteration of large granite bodies by warm acidic hydrothermal solutions migrating upward along structural channels such as faults, fractures, and joints [16]. These warm solutions attacked the granite (Fig. 14), altering most of the feldspar and leaving quartz, mica, and tourmaline relatively unaltered. Superimposed on the hydrothermally altered granites is additional alteration by surface weathering. About 2 000 000 t of kaolin is produced currently from the Cornwall area. The mines are located in the area around St. Austell. Drilling has shown that kaolinization exists at depths over 800 feet (> 250 m) [17].

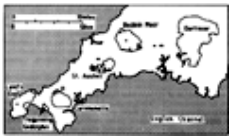

Figure 13. Location of kaolin deposits in southwestern England

° Working china clay pit

Figure 14. Cross section showing a hydrothermal kaolin deposit in granite in Cornwall

A large sedimentary deposit of kaolin of Pliocene age [18] is mined along the Jari River, a tributary of the Amazon, on the border of the states of Para and Amapa (Fig. 15). This deposit is 40 m thick and extends over an area of 60 km^2. This kaolin is fine in particle size, similar to the Tertiary deposits in Georgia and South Carolina. Currently 2005 about 900 000 t/a are being shipped from the Jari location in Brazil. The deposit is thought to be lacustrine in origin, and the kaolin was derived from weathering of the feldspathic granites on the Guyana shield. Kaolinite is the major mineral present, comprising 80 – 98 % of the deposit. The major nonclay mineral present is quartz, along with small amounts of gibbsite, zircon, rutile, anatase, ilmenite, tourmaline, and goethite.

Another large sedimentary kaolin area in Brazil is being mined and processed (Fig. 16). This kaolin is coarse, somewhat similar to the Cretaceous kaolins in Georgia. These Rio Capim crude kaolins are of exceptionally high quality [18]. As the Cretaceous kaolin deposits in Georgia become depleted, the Rio Capim kaolins will replace them in the world's paper-coating markets. Currently there are two kaolin operations along the Rio Capim producing about 1600 000 t/a. However, both plants are increasing their capacities, so that the annual tonnage will be about 2 000 000 in 2008.

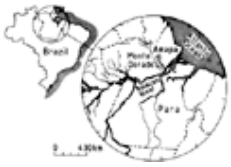

Figure 15. Location of Jari kaolin deposit at Monte Dorado

Figure 16. Location of two kaolin mines in the Capim basin: Rio Capim Caulim (RCC) and Para Pigmentos (PPSA).

Former Czechoslovakia is the principal producer of kaolin in eastern Europe. The major kaolin deposits are located in the vicinity of Karlovy-Vary, Pilzen, and Podborany (Fig. 17). The kaolin is primary in origin, and the deposits are extensive, occurring as a weathering crust on crystalline rocks of the Bohemian Massif [19]. The parent rocks are granites and gneisses. The feldspar in these rocks is almost completely kaolinized. The major nonclay minerals are quartz and mica.

Figure 17. Locations of the major kaolin deposits near Karlovy Vary

1) Bo Žicany; 2) Sadov; 3) Otivice; 4) Sedlec; 5) Všeborovice; 6) representation in Karlovy Vary

In Germany, both primary and secondary deposits are being mined. In the eastern part of Germany, near Dresden, are two large primary deposits at Caminau and Kemmlitz [20]. These kaolins are the result of weathering of the Bohemian Massif which extends into this area of Germany. The major secondary or sedimentary deposits are in Bavaria near the village of Hirschau. The kaolin is the alteration product of feldspar contained in a Triassic arkose [21]. The deposit extends for about 15 km and ranges from 3 to 8 m in thickness. Kaolinite constitutes only about 5 to 15 % of the total minerals present, with quartz being the major mineral constituent along with partially altered feldspar.

Kaolin is mined in Spain in two locations. A Primary kaolin are mined in the northwestern corner of Spain in Galicia. These deposit are residues resulting from the weathering of a large granitic body [22]. A secondary or sedimentary kaolin of Cretaceous age is mined in the state of Guadalajara, east of Madrid. This kaolin is a minor constituent (10 – 20 %) in a friable white sand. There are extensive deposits of these kaolinitic sands extending from Guadalajara to Valencia [23].

In the Ukraine, a large primary deposit is being mined and processed near Proyanovski [18]. The kaolin is washed to remove quartz and mica, which are the major nonclay minerals present.

In Australia, both primary and secondary kaolin deposits are mined [25]. The largest primary deposit being mined is north of Melbourne near Pittong. The kaolin is a residual weathering product of altered granite. The major nonkaolin minerals are quartz and muscovite. There are large sedimentary kaolin deposits of late Cretaceous or early Tertiary age beneath bauxite on the Cape York Peninsula in extreme northeastern Australia. A deposit near Weipa was in production from 1986 to 1997 shipping about 100 000 t/a of a fine particle size kaolin to the paper markets in Japan, Southeast Asia, and Australia. This operation closed in 1997 because of quality problems and high production costs. Another deposit north of Weipa near the Skardon River is currently reported to be producing calcined kaolin [24].

In China, several kaolin deposits are being mined. The largest is located near Suzhou, southwest of Shanghai. The Suzhou deposit is primary, resulting from both weathering and hydrothermal alteration of granites and other acid igneous rocks [26]. The annual production from Suzhou is about 300 000 t/a. Other processing plants are located near Maoming in Guangzhou Province in southern China. This deposit is a very large sedimentary kaolinite sand containing about 80 % quartz and 20 % kaolinite [27].

Other kaolin deposits that are smaller and have low production are listed in the literature [28].

Ball clays and refractory clays are special types of kaolinitic clays that are used primarily for ceramics. Ball clays are secondary clays characterized by the presence of organic matter, high plasticity, high dry strength, long vitrification ranges, and light color when fired. Kaolinite is the principal mineral constituent of ball clay and typically makes up more than 70 % of this type of clay. Ball clays are mined in the United States [29], England [17], Germany [30], and Ukraine [30].

Refractory clays include several varieties of kaolinitic clays that are used in the manufacture of products requiring resistance to high temperatures [29]. Fire clay and refractory clay are used interchangeably by many authors, but NORTON [31] excludes white burning clays in his definition of fire clay. Refractory clays are common and are found on every continent except Antarctica, where they are probably present but have not yet been discovered.

3.2. Smectite

Smectite is the name for a group of sodium, calcium, magnesium, iron, lithium aluminum silicates, which include the individual minerals sodium montmorillonite, calcium montmorillonite, nontronite, saponite, and hectorite. The rock in which these smectite minerals are usually dominant is *bentonite*. The name bentonite was first suggested in 1898 by KNIGHT [32] and is the term commonly used to describe the industrial mineral. The term bentonite was defined by ROSS and SHANNON [33], who restricted it to a clay material altered from a glassy igneous material, usually a tuff or volcanic ash. WRIGHT [34] suggested that bentonite was any clay composed dominantly of a smectite clay mineral and whose physical properties are dictated by this clay mineral. GRIM and GUVEN [35] used WRIGHT's definition in their book on bentonites because there are many clays designated as bentonite that did not originate by the alteration of volcanic ash or tuff. Therefore, the term bentonite usually does not include the mode of origin.

Bentonites in which sodium montmorillonites are the major mineral constituent commonly have a high swelling capacity [36]. The largest and highest quality sodium bentonite deposits in the world are located in South Dakota, Wyoming, and Montana [37]. These clays are commonly called Western or Wyoming bentonites. The bentonite beds are Cretaceous and are part of the Mowry formation [37]. There are several bentonite beds in the Mowry, but the thickest and most extensive is the Clay Spur bentonite bed (Fig. 18). This Clay Spur bentonite bed extends west from Belle Fourche, South Dakota, and then north into Alberta, Canada (Fig. 19).

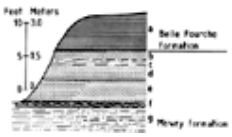

Figure 18. Diagrammatic sketch of the Clay Spur bentonite bed in the Mowry formation

a) Dark shale, nonsiliceous; b) Dark bentonite; c) Interlaminated siliceous shale; d) Dark bentonite; e) Lightcolored bentonite; f) Chertlike floor; g) Siliceous shale

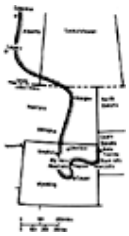

Figure 19. Outcrop area of the Mowry bentonites

The Clay Spur bentonite usually ranges from 0.5 to 2 m thick, although thicknesses up to 4 m have been reported [38]. The bentonite is generally light yellowish green on the outcrop, becoming bluish green away from the outcrop. The highest quality swelling clay is the yellowish green bentonite, which has been weathered and oxidized. This Western bentonite was formed by alteration of volcanic ash. The major nonclay minerals present in the Clay Spur bentonite bed are quartz, opal-CT, feldspar, mica, and some zeolite. Locally calcite and gypsum may also be present.

Bentonites in which calcium montmorillonites are the major mineral constituent commonly have a low swelling capacity [36]. These clays are generally referred to as Southern or subbentonites. Large deposits of these calcium bentonites occur in Texas and Mississippi [35]. In Mississippi, the calcium bentonite occurs in the Eutaw formation of the Upper Cretaceous [39]. It is commercially produced in Itawamba and Monroe Counties, where some bentonites reach a thickness of 4 m. The bentonite is waxy and varies from blue when fresh to yellow when weathered. Quartz, feldspar, and micas are the major nonclay minerals present. Another calcium bentonite has been produced from the Tertiary Vicksburg formation in Wayne and

Smith Counties of Mississippi [40]. Several beds have been identified and are generally 0.5 – 1 m thick. This Tertiary bentonite ranges from white to gray to yellow. Quartz and micas are the major nonclay minerals. This central Mississippi bentonite is often referred to as the Polkville, Mississippi, bentonite.

The Texas calcium bentonites are found in a belt paralleling the present gulf coast in Cretaceous and Tertiary sediments. They are best developed in the Tertiary Jackson and Gueydan formations in Gonzales and Fayette Counties [41]. The mined bentonites range from 0.5 to 3 m thick. They vary from white to yellow to green to dark brown. Some of the white bentonites in Gonzales County contain kaolinite [42]. The major nonclay minerals are quartz, opal-CT, feldspar, and mica.

Other calcium bentonites are produced in the United States in Arizona, Alabama, and Nevada. The best known, most extensively mined deposit is at Cheto, Arizona, where a thick (5 m) ash bed in the Pliocene Bidahochi formation has altered to calcium montmorillonite [43]. The bentonite is white to light gray. This clay also contains a small amount of kaolinite. Quartz, micas, and feldspars are the common nonclay impurities. A small amount of calcium bentonite is produced from the Amargosa Valley in Nevada [44]. This bentonite is Pliocene in age and is white to light gray. The major nonclay minerals include quartz, mica, calcite, and dolomite. In Alabama, a calcium bentonite is produced at Sandy Ridge from the Ripley formation [45].

Some special bentonites, including hectorite and saponite, are produced in the United States. Hectorite, a lithium-bearing smectite, is mined at Hector, California. The hectorite was formed where siliceous lithium – fluorine solutions reacted with carbonates [46]. The hectorite deposits are Tertiary. Saponite is mined in the Ash Meadows area of the Amargosa Valley in Nevada. It is called Pliocene [44], although DENNY and DREWES [47] mapped these rocks as Pleistocene.

Bentonites are produced in many other areas of the world and are commonly of the calcium montmorillonite variety. In England, there are two well-known occurrences of smectites, which are of the calcium montmorillonite variety. In the English literature, they are referred to as *fuller's earth*. The term fuller's earth refers to any natural material that has the capacity to decolorize oil to an extent that is of commercial value (also see Palygorskite and Sepiolite) [35]. At Redhill in Surrey, near London, is a calcium bentonite in the Upper Cretaceous [48]. The clay is waxy and blue. Another calcium bentonite, of the Jurassic period, is found in Somerset [49].

In the Federal Republic of Germany, there are extensive deposits of calcium bentonite in the vicinity of Moosberg and Landshut in Bavaria [50]. These Bavarian bentonites occur in a marine section of marls and tuffaceous sands in the Upper Miocene molasse. Some beds attain a thickness of 3 m. The bentonite is light yellow to gray and has a waxy appearance. These deposits are an important source of commercial bentonite in European and world markets.

In Argentina, a bentonite deposit near Neuquen is being mined primarily for use as a drilling mud and a bond for foundry sands [18]. A white sodium bentonite is being developed in northern San Juan Province near Rodeo. This deposit was formed by the hydrothermal alteration of volcanic ash and tuff [51].

In Greece, several of the volcanic islands contain bentonite. The most extensively developed deposit, a Pliocene calcium bentonite, is on the island of Milos [35].

In Hungary, bentonite has been produced for about 50 years and is now an important product. The most important deposits are located at Istenmezo, Nagyteteny, Band, Komloska, Mad, and Vegardo [52]. The last three are the result of hydrothermal alteration of rhyolites and pyroxene andesites associated with Tertiary volcanism; the first three are the result of alteration of tuff that fell into the sea and altered under marine conditions. These altered tuff deposits are Surmatian (Tertiary).

In Italy, important calcium bentonite deposits are located on the island of Sardinia. The Sardinian bentonites are of two types, hydrothermally altered trachytic tuffs and sedimentary bentonite of the Miocene, which is a marine-altered ash [53], [54].

In the territories of former Yugoslavia, there are several calcium bentonites produced commercially. These are located near Pula, Kutina, Laktasi, Kriva Palanka, and Petrovac. The Petrovac deposit is Triassic, and the others are Middle to Upper Tertiary [35]. These bentonites are altered ashes and tuffs [55].

In India, bentonite is produced in the Barmer district of Rajasthan near Akli and Hathi-ki-Dhani [56]. These are quite pure calcium montmorillonites up to 3 m thick. They are Lower Tertiary and are associated with calcareous sands and conglomerates.

In Japan, commercial bentonites are located in the Yamagata, Gumma Niigata, and Nagano Prefectures of northern Honshu and on Hokkaido. The deposits are Miocene and Pliocene, formed by alteration of ash, pumice, and tuff [57]. Some of the bentonites in the Gumma Prefecture are up to 5 m thick.

In the Republic of Georgia, near the capital city of Tbilisi, are two well-known bentonite deposits, Askana and Gumbri [35]. The Askana bentonite is part of a series of Eocene tuffs and breccias ranging in composition from trachytes to basalts. Both sodium and calcium montmorillonites are produced from the Askana deposit. RATEEV [58] believes the Askana bentonites are hydrothermal in origin. The Gumbri bentonite occurs in beds up to 8 m thick and is Cretaceous in age [59]. This bentonite was formed by the submarine alteration of volcanic glass.

Other bentonite deposits along with their geologic age and occurrence are described in [35].

3.3. Palygorskite and Sepiolite

Palygorskite and sepiolite are clay minerals with a chain structure. These include palygorskite and sepiolite. The name attapulgite is synonymous with palygorskite, but the more recent name attapulgite has been discredited as a mineral name. However, it is so prevalent in the literature that the term still persists.

These clays are sometimes referred to as sorptive clays because they have a large surface area and can absorb and adsorb many materials. The term *fuller's earth* is more or less a catchall term for clays or other fine-grained earthy materials suitable for bleaching and sorbent uses. It has no compositional or mineralogical meaning. The term was first applied to earthy material used in cleansing and fulling wool, thereby removing the lanolin and dirt [60]; thus, it acquired the name fuller's earth. Industrially the term fuller's earth is used quite often for any sorptive clay.

Palygorskite (attapulgite) is produced in the southeastern United States near the Georgia – Florida border in the vicinity of Quincy, Florida, and Attapulgus, Georgia, in Senegal near Dakar, in China at Mingguang in Anhui Province and at Xuyi in Jiangsui Province, and in the Ukraine [18]. Sepiolite is mined commercially in Spain, which has the largest deposits, in Turkey, and in Nevada in the United States. These clays are lathlike or elongate in shape (see Fig. 10). Sepiolite is structurally similar to palygorskite except the former has a slightly larger unit cell and the individual laths are ca. one-third wider [61]. This may be due to the presence of more iron substituting for magnesium in the lattice, because generally palygorskite has a higher iron content.

The palygorskite deposits in southeastern United States are of the Miocene age and were deposited in marine lagoons [62]. The clay ranges from 0.5 to 3 m thick and is gray to tan or brown. These deposits are mixtures of palygorskite and smectite with the palygorskite content higher in the southern part of the district (Fig. 20).

Figure 20. Fuller's earth deposits in Florida and Georgia

The palygorskite in Senegal [63] is a large deposit that extends from Theis, where it immediately overlies a commercial aluminum phosphate deposit, south-southwest about 100 km to the southern border of Senegal. This deposit has been in production for ca. 20 years. The age of the deposit is early Eocene, and the thickness ranges from 1.5 to 6 m. The overburden thickness is low, usually less than 2 m. The color of the palygorskite is light tan to gray green. There are calcite, dolomite, and chert impurities in some areas. Smectite is also a common constituent, but in some areas it is barely detectable on X-ray diffraction patterns.

In Russia, palygorskite was first described in the Urals in the Palygorsk Range [64]. Palygorskite is mined in the Ukraine in the central part of the Ukrainian crystalline massif along the borders of the Cherkassay and Kiev regions. The palygorskite occurs in the middle section of a productive bentonite and is Lower Miocene [64]. The thickness of the palygorskite layer is ca. 2 m. Smectite, hydromica, and quartz are the other major constituents.

Palygorskite is produced in two areas of China, at Mingguang in Anhui Province and at Xuyi, which is directly north of Nanjing in Jiangsui Province. At Mingguang, the palygorskite occurs directly under a basalt and ranges between 4 and 6 m in thickness [18]. Near Xuyi the palygorskite is very near the surface (under a smectite rich layer that is 1 – 2 m thick) and is about 6 m thick. There are lenses of a light tan dolomitic palygorskite interspersed in the palygorskite at Xuyi. The ages of these two palygorskites are Middle Miocene, and the reserves at both locations are very large.

Sepiolite is mined in Spain near Toledo from a large Tertiary lacustrine deposit [65]. It is light tan and ranges from 1 to 5 m thick. Palygorskite is mined at Trujillo in the province of Caceres in western central Spain.

Sepiolite is mined in small quantities in the Ash Meadows area in the Amargosa Valley in Nevada [66]. The sepiolite occurs as 0.3 – 1 m stringers in saponite clays of lacustrine origin. The sepiolite is lightweight, very porous, and brown to brownish gray. The age of these sepiolite beds is Pliocene or Pleistocene [44], [47].

Sepiolite is produced in Turkey in the vicinity of Eskisehir in Anatolia [61]. The sepiolite is white and is interbedded with palygorskite shales. The sepiolite contains up to 3 wt % organic carbon in the form of humic acid, which enhances the dispersability of the clay.

Palygorskite is produced in the Mudh District in India [67]. The deposits are Tertiary and occur between fossiliferous limestones, thus apparently marine in origin. The thickness averages ca. 1.5 m.

Sepiolite is mined in Somalia where it is used to produce carved figurines and other art objects.

3.4. Miscellaneous Clays

Illite or chlorite deposits are not individually mined because they are normally mixed together and with other clay and nonclay minerals. Many clays and shales contain these clay minerals and are used for structural clay products, lightweight aggregate, and cement. Generally, clays and shales that contain illite and chlorite in major proportions have properties that are important in the manufacture of structural clay products. Structural clay products are a broad category that includes brick, drain tile, sewer piper, conduit tile, floor tile, wall tile, terra cotta, and other items (⟶ Construction Ceramics). The properties that are important in the manufacture of structural clay products are plasticity, green strength, dry strength, drying and firing shrinkage, vitrification range, and fired color. These physical properties depend on the constituents of the clay and their grain size [68].

Illitic and chloritic shales and clays are not restricted geologically and can be found in many areas of the world. In the United

States, shales and clays with high illite and chlorite content are found in all geologic periods with a preponderance in Pennsylvanian, Cretaceous, and Tertiary formations. Usually these clays and shales are relatively thick with low overburden.

Lightweight aggregate is a material produced from clays and shales. The material is fired rapidly in rotary kilns or sintering machines up to the temperature between incipient and complete fusion. At this temperature, substances that release gas cause bloating and vesiculation after fusion has developed a molten jacket around the particle, which prevents the escape of the gas. The mineral content is important, i.e., the clay minerals and the minerals such as calcite or dolomite that produce a gas on firing [69]. Clays and shales used for lightweight aggregate production are found in many areas of the United States and again are not restricted geologically or geographically.

Large quantities of clays and shale are used in the manufacture of cement. These clays and shales provide silica and alumina for the cement charge. Practically all clay or shale with an average alumina and silica content can be used in the cement industry. Usually these clays and shales can be found in the area in which the cement is being manufactured (→ Cement).

[Top of Page]

4. Mining and Processing

Almost all clay is mined by open pit. There are a few underground mines; however, in most instances the cost of production and poor roof conditions prohibit underground mining of clay deposits unless the clay has exceptional properties.

An understanding of the geology and the origin of a particular type of clay plays an important part in clay exploration. After a clay deposit is discovered, it must be evaluated to determine the quantity and quality of the clay. In most instances, this is done by core drilling and testing the recovered cores [70]. The spacing of the drill holes depends on the geologic and surface conditions associated with the deposit. For example, the drilling pattern used to evaluate a sedimentary kaolin deposit is much different than that used to evaluate a residual or hydrothermal kaolin deposit.

A small section that is contaminated with a deleterious material can affect the results of the total length of the core if it is composited. For this reason, the core is normally tested in increments of a meter or less. If the core is 5 m long, the core is composited for testing in five increments. In this way, a section that does not meet a particular specification can be isolated. Further drilling can corroborate the thickness and extent of the material that is out of specification, and a mining plan can be designed to handle the particular problem. Drilling is normally done on a pattern grid so that maps indicating thickness of overburden and thickness and quality of clay can be prepared.

4.1. Kaolin

All kaolins that are used as paper coating clays and most that are used as filler clays are processed because practically no deposits are naturally pure. Usually beneficiation is accomplished by one of two basic processes, a dry process or a wet process [70]. The higher quality grades of kaolin, except for some kaolins with special ceramic properties, are prepared by wet processing because the product is more uniform, relatively free from impurities, and of better color.

Major kaolin processing developments and their significance to the growth of the kaolin industry have been summarized [71]. There has been a steady growth of the kaolin industry since 1930, and the only physical property that has not been significantly improved by process developments is viscosity.

4.1.1. *Dry Process*

The dry process [72] is relatively simple and yields a lower cost, lower quality product than the wet process. In the dry process, the properties of the finished kaolin reflect the quality found in the crude kaolin. Therefore, deposits containing the desired qualities of brightness, low grit content, and satisfactory particle-size distribution are located by drilling and testing.

The overburden on the kaolin is stripped by motorized scrapers and bulldozers or draglines, or shovels. Once the overburden is removed, the kaolin is mined by using draglines, shovels, or front-end loaders. The kaolin is transported to the processing plant in trucks. The kaolin is usually shredded or crushed and stored in large storage sheds. Usually the kaolin is segregated into two or three bins according to specific properties, such as brightness and grit percentage. (Grit is material, usually quartz, that is coarser than 44 μm, i.e., the material that is retained on a 325-mesh screen.) The general flow sheet utilized in dry processing is shown in Figure 21.

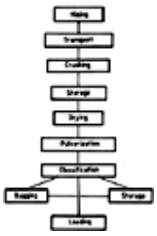

Figure 21. Dry-process flow sheet for kaolin production

The material, as it enters the storage shed or enters the process, is shredded or crushed to particles equal to or less than the size of an egg. Normally, the crude kaolin contains ca. 20 – 25 % moisture and, therefore, must be dried. This is commonly done in a rotary dryer. After drying, the kaolin is disintegrated in a roller mill, hammer mill, disk grinder, or some other disintegrating device. Commonly, heat is applied to the kaolin during the disintegration to further reduce the moisture content. The disintegrated kaolin may then be air-classified to separate the fine particles from the coarse. The top size can be controlled to a degree so that various grades, based upon particle size, can be produced. The material is shipped in bulk or

in bags by truck or railcar.

4.1.2. *Wet Process*

The general flow sheet for wet processing Georgia kaolin [73] to produce both filler and coating grades is shown in Figure 22. The kaolin is mined with shovels, draglines, motorized scrapers, or front-end loaders and can be either transported to the processing plant or fed into a stationary or mobile blunger (Fig. 23). The blunger separates the kaolin into small particles, which are mixed with water and a dispersing chemical to form a clay – water slip or slurry. The dispersing chemical can be sodium polyphosphate, sodium silicate, or sodium polyacrylate, each of which can be blended with soda ash to reduce cost.

Figure 22. Wet-process flow sheet for kaolin coating and filler production

Figure 23. Dragline feeding kaolin into a mobile blunger

The percent solids of the slurry is normally between 30 and 40 %, but in some special circumstances it can be in excess of 60 % and as high as 70 % solids (see Section Combined Dry and Wet Processing). This clay – water slurry is pumped from the blunger to rake classifiers or hydrocyclones and screens to remove the grit. The grit is removed from the kaolin slurry and discarded into waste ponds or into mined-out areas. The degritted slurry is collected in large storage tanks with agitators and is then pumped to the processing plant, which may be several miles away.

The kaolin slurry is collected in large storage tanks at the plant before it is processed. Although each kaolin producer may process the kaolin slightly differently, the steps in the processing are generally as shown in Figure 22. The first step is to separate the kaolin particles into a coarse and fine fraction through continuous centrifuges. Some operators pass the degritted slurry through a high-gradient magnetic separator prior to centrifugation to upgrade the crude clay by removing iron and titanium minerals (see Section Special Processes).

The coarse fraction can be used as a filler in paper, plastics, paint, and adhesives, as a casting clay for ceramics, or as a feed to delaminators, which can be a subsequent processing step to produce a special paper coating grade (see Section Special Processes). The fine particle size fraction is used in paper coatings, high-gloss paints, inks, special ceramics, and rubber. As shown in Figure 22, there are many different steps that can be taken after the clay is fractionated. Normally, the coarse kaolin takes one of two routes:

1. It can go directly to the leaching department, where it is chemically treated to solubilize some of the iron if the brightness needs to be upgraded, or it can be flocculated so that the dewatering step is facilitated.
2. It can go through magnetic separation and into the delaminators or directly to the delaminators and then to the magnetic separator or to the delaminators without any magnetic separation step.

In the leaching operation, the kaolin slurry is acidified with sulfuric acid to a pH of ca. 3.0 to solubilize some of the colloidal iron. Sodium dithionite, a strong reducing agent, is then added to reduce the iron to the iron(II) state, which is more soluble and forms a clear iron sulfate that is removed with the filtrate in the dewatering step. In some cases, the iron in the kaolin is in a reduced state rather than the normal oxidized state. Usually such kaolins are slightly gray. These gray kaolins are treated with a strong oxidizing agent, such as ozone or sodium hypochlorite, prior to acidification and the normal reduction leach procedure. To facilitate flocculation prior to dewatering, in many cases alum is added along with the sulfuric acid.

At this point in the process, the kaolin slurry is between 20 and 30 % solids and is flocced. This flocced slip is then dewatered, which raises the solids to ≥ 60 %. The dewatering is accomplished by large rotary vacuum filters or plate and frame filter presses. The dewatered filter cake is then prepared for drying by using rotary, apron, drum, or spray dryers. Most of the clay that is shipped in dry bulk form is spray dried in a dispersed state. The usual process is to take the filter cake, which is flocced, and disperse it by adding a long-chain sodium polyphosphate or sodium polyacrylate. This dispersed high-solids slurry (≥ 60 %) is fed into the spray dryer, where it is sprayed or atomized and dried to meet the final moisture specifications, which normally range from less than 1 % to as high as 6 %. This dispersed clay slurry can also be dried by rotary drum dryers for some special uses that require the flake form.

Slurry clay, which is normally shipped in tank cars or trucks at 70 % solids, is produced by mixing the dispersed clay from the filters with enough spray-dried clay to bring the final solids to 70 %. Another process to bring the solids to 70 % is to use large evaporators. About 70 % of all shipments of coating and fine filler clays to the paper industry is now being shipped in slurry form.

If the filter cake is to be dried with rotary or apron dryers, then the filter cake is extruded into noodles ca. 3/8 in. (1 cm) in diameter and fed into the rotary dryer or onto an apron dryer. The final moisture from these dryers is generally ca. 6 %. This type of clay can be shipped in lump form, or if necessary, the clay can be further dried and disintegrated, using hammer mills swept with hot air. The final moisture of this disintegrated kaolin is ≤ 0.5 %. It is commonly referred to in the kaolin industry as acid clay because the pH is ca. 4.5.

The wet processes are used on the sedimentary kaolins in Georgia and Brazil. Primary kaolins such as the English china clays produced in the Cornwall district of southwest England are wet processed in much the same manner as the Georgia kaolins except for the mining methods. The kaolin content of the weathered granites ranges from 10 to 20 %. Mining uses high-pressure hydraulic monitors that play a stream of water on the mine face, washing out the fine particle kaolin and leaving the coarse quartz and mica residue. The fine kaolin is suspended and is transported in small rivulets to a collecting basin in the bottom of the open pit. This slurry is pumped to large thickeners, where the clay slurry is concentrated, and then processed generally following Figure 22.

4.1.3. *Combined Dry and Wet Processing*

A sort of hybrid development, a blend of dry processing and wet processing, is the whole clay slurry produced from fine particle size East Georgia Tertiary crude kaolins. Two processes are used to produce this high-solids whole clay slurry. One is to dry process the kaolin and then slurry the product at 70 % solids and degrit the clay slurry with a scalping cut on a centrifuge. A second method is to blunge the crude clay at high solids (65 – 68 %), degrit with a centrifuge, and add dry clay to bring the solids to 70 %.

4.1.4. *Special Processes*

Several special processes are used to produce unique and special quality grades of kaolin. One of these special processes is *delamination*, which is a process that takes a large coarse kaolin stack and separates it into several thin, large-diameter plates (Fig. 24). These thin, large-diameter plates have excellent covering power on rough base sheets of paper and are also used to produce high-quality lightweight coatings.

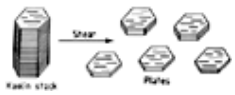

Figure 24. Diagrammatic representation of delamination of a kaolin stack into large thin kaolin plates

The process of delamination involves attrition mills into which fine media, such as glass beads or nylon pellets, are placed along with the coarse kaolin stacks and intensely agitated. The fine media impact the kaolin stacks, separating them into thin plates [74]. The brightness and whiteness of the delaminated kaolin is very good: the clean newly separated basal plane surfaces are white because they have been protected from groundwater and iron staining. The delamination process has enabled the kaolin industry to convert low-priced coarse clay into a higher priced special coating and filler clay.

Another special process is *ultraflotation*, which is a special flotation process to remove iron and titanium contaminants to make a 90 % brightness kaolin product [75]. *Selective flocculation* [76-78] is another special process that produces a 90 % brightness kaolin product by selectively removing iron and titanium impurities.

High-gradient magnetic separation (HGMS) is another special process, although it has become a standard processing technique in the kaolin industry. The high-gradient magnetic separation process [79] uses a canister filled with a fine stainless steel wool that removes iron, titanium, and some mica minerals from the kaolin slurry as it passes through the canister (Fig. 25). The process can be used to upgrade marginal low-brightness crude kaolins and to produce high-brightness coating clays. It also reduces the amount of chemicals that are needed to leach out the iron in a subsequent processing step. The development of HGMS has dramatically increased the usable reserves of kaolin in Georgia. Superconducting cryogenic magnets with field strengths up to 5 T are now being used in addition to the 2 T electromagnets. The superconducting magnets use very little power and are effective in removing ultrafine paramagnetic minerals [80].

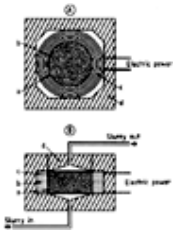

Figure 25. Diagrammatic sketch of a high-gradient magnetic separator

A) Top view: a) Canister wall; b) Steel-wool matrix; c) Magnet coil; d) Magnet steel

B) Side view section through magnet: a) Canister; b) Magnet coil; c) Matrix; d) Magnet steel

Calcining is another special treatment that is used to produce special grades. One grade is thermally heated to a temperature just above the point where the structural hydroxyl groups are driven out as water vapor, which is between 650 and 700 °C. This produces a bulky product that is used as a paper coating additive to enhance resiliency and opacity in low basis weight sheets. A second grade is thermally heated to 1000 – 1050 °C. By proper selection of the feed kaolin and careful control of the calcination and the final processing [81], the abrasiveness of the calcined product can be reduced to acceptable levels. The brightness of this fully calcined, fine-particle kaolin is 92 – 95 %, depending on the feed material. It is used as an extender for titanium dioxide in paper coating and filling and in paint and plastic formulations.

Another special process is *surface treatment*. Kaolin is hydrophilic and can be easily dispersed in water. Because of the nature of the kaolinite surface, it can be chemically modified to produce hydrophobic and organophilic characteristics. Generally, an ionic or a polar nonionic surfactant is used as a surface treating agent. These surface-modified kaolins are

used in paper [82], paint [83], plastics [84], and rubber [85].

4.2. Smectite

Virtually all bentonite, which is comprised mainly of a smectite mineral, is strip-mined. Bulldozers and motorized scrapers are most commonly used for removing overburden. In a typical open pit, the overburden is removed in panels. The exposed bentonite is then mined by loading the material into trucks with draglines, shovels, or front-end loaders. Once the bentonite is removed, the overburden from the next panel is shifted to the mined-out panel.

Thicknesses of removed overburdens vary considerably. Most Wyoming-type bentonite has 10 m or less overburden. The thickness of overburden removed from some Mississippi bentonite deposits is as much as 30 m [86], about the same as that on some of the Arizona bentonites near Chambers. In Bavaria, some overburden on the bentonites is up to 50 m thick.

Bentonite beneficiation and processing involves relatively simple milling techniques that involve crushing or shredding, drying, and grinding and screening to suitable sizes (Fig. 26). The high-swelling Wyoming or sodium bentonites contain ca. 30 % moisture when delivered to the plant. The Southern or calcium bentonites have ca. 25 % moisture. The processed bentonite generally contains only 7 – 8 % moisture, although because bentonite is hygroscopic, it may contain considerably more moisture when used.

Figure 26. Flow sheet for bentonite production

The raw bentonite is passed through some sort of crushing or shredding device to break up the large chunks before drying. Drying in most plants is accomplished by gas- or oil-fired rotary dryers, but in one bentonite plant in Wyoming, a fluidized-bed dryer is used. The bentonite properties can be seriously affected by overdrying, so the drying temperature must be carefully controlled.

The dried bentonite is ground and sized in several ways. Granular bentonite is cracked by using roll crushers and screened to select the proper size range of granules. Most powdered bentonite is ground with roll or hammer mills to ca. 90 % finer than 200 mesh. The bentonite is shipped mainly in bags, but some is shipped in bulk.

Special processing is used in some plants to produce special products. Extruders can be used to pug the clay and mix it with additives that may improve the viscosity and dispersion properties. Some hectorite and Southern bentonites are beneficiated wet. In this process, the bentonite is dispersed in water and degritted by centrifuges, hydrocyclones, settling devices, or screens. The slurry is either filtered and then flash-dried or sent directly to rotary drum or spray dryers.

Organic-clad bentonite is bentonite that is processed to replace the inorganic exchangeable ions with alkylamine cations, which produces a hydrophobic clay [87]. These organic-clad bentonites are used in paints, greases, oil-base drilling muds, and to gel organic liquids.

Acid-activated bentonites are special sorptive clays used for bleaching edible oils and decolorizing special lubricating oils. The process flow sheet for acid activation is shown in Figure 27. The process involves slurrying the bentonite in warm water, removing the grit by hydrocyclones, reacting the clay with either sulfuric or hydrochloric acid at elevated temperature, dewatering in plate and frame filter presses, and flash drying. A dry activation process is also used and involves adding concentrated sulfuric acid to the clay in a blender and then pulverizing and drying the acid – clay mixture to ensure uniform distribution of the acid on the clay. Acid treatment removes some of the aluminum, magnesium, and iron octahedral ions, resulting in a highly charged particle saturated with hydrogen ions, which makes for a good bleaching and decolorizing clay.

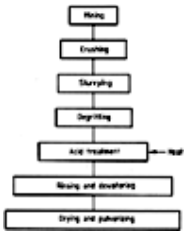

Figure 27. Flow sheet for producing acid-activated bentonite

4.3. Palygorskite and Sepiolite

The mining of palygorskite and sepiolite clays is no different than mining bentonites. The processing is also similar (Fig. 28). The hormite clay is shredded or crushed, dried, screened, air-classified, and packaged. Some of this hormite clay is thermally treated to harden the particles to prevent disintegration of the granules during bagging, transport, and handling prior to arrival at the customer. The temperature of this thermal treatment is less than 400 °C so that no structural modification results, but all of the adsorbed water is driven from the surface and interior of the particles.

Figure 28. Flow sheet for producing palygorskite

When palygorskite (attapulgite) is to be used for salt-water drilling mud, the clay is extruded at high pressure through small orifices. This shears apart the bundles of palygorskite into individual elongate particles, which produces a higher viscosity drilling fluid.

[Top of Page]

5. Properties and Uses

Kaolins, smectites, palygorskite and sepiolite, and miscellaneous clays form an important industrial minerals group. Because of their fine particle size, particle shape, and surface chemistry, they have unique colloidal and physical properties.

5.1. Kaolin

Kaolin is one of the most versatile of the industrial minerals [88] and is used extensively for many applications [89]. It is a unique industrial mineral because it

1. is chemically inert over a relatively wide pH range (except for catalytic activity in some organic systems);
2. is white or near white;
3. has good covering or hiding power when used as a pigment or extender in coating and filling;
4. is soft and nonabrasive;
5. has low conductivity of both heat and electricity; and
6. is lower in cost than most materials with which it competes.

Some uses of kaolin require rigid specifications, including particle size, brightness, color, and viscosity. On the other hand, some uses have no critical specifications, e.g., cement, where the only concern is light color and chemical composition. Ceramic specifications are variable in that individual users may have different requirements as to strength, plasticity, fired color, shrinkage, and pyrometric cone equivalent (PCE). The major process industries in which kaolin finds a substantial use are paper, paint, ceramics, rubber, plastics, ink, catalyst, and fiberglass, as well as some special uses.

5.1.1. *Physical Properties*

Diagnostic tests are used to evaluate kaolin physical properties [90]. Some properties that can be used as screening tests for the many potential uses for kaolin are as follows:

mineralogy
screen residue (grit)
particle-size distribution
brightness
viscosity
pH

Mineralogy. The mineral content of a kaolin is important in assessing the uses, the results of many tests determining the physical properties, and the beneficiation processes that may be needed to produce a saleable product. X-ray diffraction of the whole sample gives a quick assessment of the gross mineralogy, but in many instances minor quantities of illite and smectite are not detected. Separation of the 2-μm and 0.5-μm fractions and preparation of oriented slides for X-ray enhances the basal reflections of the clay minerals and enables the detection of very small quantities of smectite and illite.

The presence of quartz, opal-CT, alunite, smectite, illite, muscovite, biotite, chlorite, gibbsite, feldspar, anatase, pyrite, iron oxides and hydroxides, and halloysite affects the beneficiation processes and the possible uses of the kaolin. In many instances, the presence of halloysite in a kaolin cannot be detected by X-ray diffraction; therefore, electron micrographs and differential thermal analysis must be used.

Screen Residue. The grit test gives the percentage of particles that are retained on a 325-mesh screen (44-μm openings). Quartz and mica, along with hard agglomerates of kaolin, are the most common minerals retained on the screen. The amount of screen residue is important in estimating the recovery of the minus 325 mesh material since this is the fraction that is the usable portion of the kaolin. Generally, if the kaolin contains more than 5 – 8 % grit, it cannot be dry processed.

Particle-Size Distribution. Particle-size distribution (Fig. 29) is one of the more important properties of a kaolin because it affects viscosity, brightness, opacity, gloss, ceramic strength and shrinkage, and many other properties. Sedimentation methods based on Stoke's law of settling are used for measuring particle size. The 2-μm content of a kaolin is particularly important because it is this fraction that constitutes the major portion of paper coating clays and high-glossing paint clays. The particle-size distribution also determines the beneficiation procedures that are used in processing a kaolin for industrial uses.

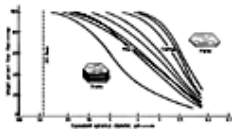

Figure 29. Particle-size distribution of some typical kaolin coating and filler grades

Brightness. The brightness of a kaolin is an important property because it determines the potential uses and prices. In general, the higher the brightness of the kaolin the more value the kaolin has. Standard brightness values are measured at a wavelength of 457 nm, and a common standard against which the samples are compared is smoked magnesium oxide. As a general rule, filler clays have a brightness of 80 – 85 %, and paper coating clays and high-glossing paint clays have a brightness of 85 – 93 %.

Viscosity. The flow properties of kaolin used for paper coating are very important. Both low-shear and high-shear viscosity measurements are made at 70 – 71 % solids. Low-viscosity kaolins are required for paper coating because the kaolin coating color must flow easily as it is applied to the paper surface, not leaving streaks or blotches, which can be caused by high-viscosity coating colors.

pH. The pH of untreated kaolin slurries normally ranges from 4.5 to 6.5. A high pH generally indicates the presence of soluble salts that can cause problems in many applications.

5.1.2. *Paper*

The mineral kaolinite has many physical properties that are advantageous to the paper industry. It is white, fine, soft and nonabrasive, and relatively inert. In the United States, ca. 70 % of all of the wet processed kaolin is used by the paper industry as a coating clay and as a filler. Some physical constants that are representative of kaolins from Georgia are the following:

Specific gravity d_4^{20}	2.62
Index of refraction	1.57
Hardness (Mohs' scale)	1.5 – 2
Fusion temperature, °C	1850
Valley abrasion number	4 – 10
Dry brightness (at 457 nm), %	75 – 91
Crystal system	triclinic

The Georgia kaolins and English kaolins are the most widely used kaolins in the paper industry. Table 2 lists typical chemical analyses of two Georgia kaolins and an English kaolin, two Brazilian kaolins, and theoretical kaolin.

Table 2. Typical chemical analyses, wt %

Component	Cretaceous middle Georgia kaolin	Capim soft kaolin	Tertiary east Georgia kaolin	Jari hard kaolin	English kaolin	Theoretical kaolin
SiO_2	45.30	46.56	44.00	44.45	46.77	46.3
Al_2O_3	38.38	38.03	39.5	37.37	37.79	39.8
Fe_2O_3	0.30	0.59	1.13	1.93	0.56	
TiO_2	1.44	0.78	2.43	1.39	0.02	
MgO	0.25	0.01	0.03	0.02	0.24	
CaO	0.05	0.01	0.03	0.01	0.13	
Na_2O	0.27	0.03	0.08	0.01	0.05	
K_2O	0.04	0.02	0.06	0.12	1.49	
Ignition loss	13.97	13.8	13.9	14.45	12.79	13.9

Important Properties. *Particle-size* and *shape distribution* are probably the most important factors in controlling the many physical properties of kaolin important to the papermaker. Figure 29 shows the range of particle-size distributions available commercially. The 2-µm size is used as the commercial control point. The coarse particle size kaolins are normally used as filler clays, and the fine particle size kaolins are used as coating clays.

Some of the properties of importance to the papermaker are dispersion, rheology, brightness and whiteness, opacity, gloss and smoothness, orientation, adhesive demand, film strength, and ink receptivity.

Complete *dispersion* of the kaolin particles is necessary to obtain maximum efficiency in a coating or filler clay. Kaolin is

easily dispersed because of its hydrophilic surface. In the natural state, kaolin is acid and is flocculated. This flocced state is the result of the presence of both positive and negative charges on the particles, which attract each other and form aggregates. Therefore, dispersing agents must be used to deflocculate the kaolin by making the particles all the same charge order, thus obtaining maximum fluidity. The most efficient dispersing agents for kaolin in aqueous systems are alkali-metal polyphosphates, polyacrylates, and silicates [91].

The *rheology* or flow properties of coating clays and coating colors are complex and exert a controlling influence on the coating operation. Such factors as coat weight, smoothness, and freedom from pattern on the coated sheet are related to viscosity. Low-viscosity kaolins can be dispersed (made down) readily at high solids concentration (71 %). In addition, these low-viscosity suspensions can be pumped and screened with relative ease. The coating process (i.e., trailing blade, roll coater, air knife, etc.), the coating color formula, the type and amount of adhesive, and the percent solids of the coating color determine the high- and low-shear viscosity values that can be tolerated for the coating clay.

Brightness and whiteness of kaolin coatings are controlled primarily by the presence of iron and titanium minerals in the kaolin. However, the coating structure is also important. Significant differences in refractive indices between coating components are necessary to obtain maximum light scattering and, thus, optimum brightness and whiteness. The coating structure must have a maximum of uniformly distributed air-filled pores. Differences in brightness of one percentage point of the dry kaolin powder may result in a difference of less than 0.2 % brightness of the coated sheet. Coating structure and weight, particle-size distribution, particle shape, and brightness and whiteness of the dry kaolin powder all affect the coated sheet brightness and whiteness.

Opacity of kaolin coatings is strongly influenced by particle packing, which depends largely on particle size and shape and particle-size distribution. Particle size and particle-size distribution must be optimal to develop efficient scattering of light and consequent opacity. The presence of ultrafine particles of ca. 0.1 μm seriously lowers the opacity [92]. The opacity of the coating layer depends on the difference in the refractive indices and the relative proportions of the various components, including air, pigment, and adhesive.

Gloss and smoothness are related to particle size and shape. Superior gloss is attained with kaolins of fine particle size. Smoothness is developed with fine, thin, small-diameter particles. The object of coating paper is to smooth over the fine structure of the base paper to make high-quality ink transfer possible.

Parallel *orientation* of the kaolin particles results in conformation to substrate irregularities, thus reflecting imperfections of the substrate surface. The method of application of the coating affects the orientation and, hence, the smoothness. For optimum effects on irregular substrates, the particle orientation should be random adjacent to the substrate and parallel at the top of the coating [93].

Coating clays are characterized by a low *adhesive demand*, ranging between 10 and 18 %. Finer particle size clays require slightly higher amounts of adhesive. The introduction of adhesive can disrupt orientation and reduce smoothness, gloss, and other optical qualities. The selective drainage of adhesive into the substrate or the concentration at the coating surface strongly affects the coating properties [93]. The adhesive requirement is directly related to the surface area of the clay. The amount of adhesive in the coating affects the opacity, brightness, color, and smoothness, as well as film strength. If the amount of adhesive is excessive, the printability is poor.

Film strength is related to preferential adhesive migration into the substrate and to the preferred orientation of the kaolin particles [94]. Maximum film strength is attained with well-oriented, dense films, which imply a highly bonded surface area. If pick occurs within the coating layer, it is generally caused by insufficient adhesive, excessive adhesive migration into the base sheet, or poorly dispersed pigment.

Kaolins are considered to have low ink absorption even though they readily accept printing inks. Substantial differences exist between the *ink receptivity* of coatings made from different grades of kaolin. This is related to film permeability, which is influenced by void volume. Thus, small-diameter particles that are randomly oriented give excellent ink receptivity. The interrelationship of the kaolin and adhesive is a major control over ink receptivity and holdout.

There are many grades of coating and filler clays available to the paper industry. The properties that define these coating and filler clays are based on brightness, particle size and particle-size distribution, and particle shape. Table 3 shows the major coating grades and their particle size and brightness.

Table 3. Particle size and brightness of Georgia coating kaolins

Grades	Mass fraction with particle size < 2 μm, %	Brightness, %
Regular coating clays		
No. 3	73	85.0 – 86.5
No. 2	80 – 82	85.5 – 87.0
No. 1	90 – 92	87.0 – 88.0
Fine no. 1	95	86.0 – 87.5
Delaminated coating clays		

Regular	≈70 – 80	88.0 – 90.0
Fine	≈90 – 95	87.0 – 88.0
High-brightness coating clays		
No. 2	80	89.0 – 91.0
No. 1	92	89.0 – 91.0
Fine no. 1	95	89.0 – 91.0
Special coating clays		
Low-glossing clays	≈50	87.0 – 88.0
Calcined extender	85	92.5 – 94.0

Coating Clays. The no. 3 and no. 2 grades are generally used in high-solids formulations for publication-grade paper and medium-finish lower cost enamels. Slick sheet magazines with large circulation are generally coated with one of these two regular grades often mixed with the regular delaminated grade. The no. 1 grade is used to coat high-quality boxboard and enamel papers. This grade gives excellent gloss and brightness and relatively uniform ink receptivity and good smoothness. The fine no. 1 grade is used where maximum gloss and very good ink holdout is desired.

The regular delaminated grade is used on lightweight paper and paper that has a rough surface where good printability must be maintained. This clay gives coatings that are exceptionally smooth with excellent ink and varnish holdout. The fine delaminated grade is used to achieve maximum gloss, good runnability, and high opacity.

The high-brightness grades, having a blue-white tint, are very bright. The coatings that use these grades are very smooth and give excellent gloss ink holdout and good printability. These high-brightness grades are used to coat premium-grade paper.

The special coating grades were developed to meet specific coating needs. The low-glossing coating clays are used when a dull or matte finish is required. The calcined kaolin extender is used to replace titanium dioxide in the coating to reduce the cost of the coating formulation and still maintain the desired optical properties.

Filler Clays. Fillers are highly desirable in printing papers because they increase the opacity, raise the brightness, increase the smoothness or finish, and improve the ink receptivity and printability. Filler clays also improve the appearance and absorbency of paper, as well as increase the density. A perfect filler, if available, would have the following characteristics [95]:

1. Reflectance of 100 % at all wavelengths of light
2. High index of refraction
3. Grit-free and a particle size close to 0.3 μm, approximately half the wavelength of light
4. Low specific gravity, soft, and nonabrasive
5. Ability to impart to paper a surface capable of taking any finish, from the lowest matte to the highest gloss
6. Complete retention in the paper web
7. Completely inert and insoluble
8. Reasonable in price

Obviously, such a perfect filler is not available, and filling paper is a matter of compromise. Kaolin is the usual filler in white or essentially white papers such as newsprint, printing grades, uncoated book, to develop certain desirable properties, and in paper may be the only filler used. In some grades, kaolin may be partially replaced by other mineral fillers to develop certain desirable properties, or may be used alone without clay.

In general, the lighter the weight of the paper, the greater the need for a filler to improve its opacity and brightness. Cost reduction is an important criterion for using fillers. In most cases, the filler is much less expensive than the pulp it replaces. Table 4 gives the filler grades of kaolin that are available.

Table 4. Filler grades of kaolin

Type	Brightness, %	Screen residue, %
Air-floated filler	80 – 81	0.30
Whole clay filler	81 – 83	0.10
Whole clay filler	83 – 85	0.10
Water-washed filler	81 – 83	0.09
Water-washed filler	83 – 84	0.09
Water-washed filler	84 – 86	0.05
Delaminated filler	86 – 87	0.005

High-brightness delaminated filler	87 – 89	0.005
Calcined kaolin extender	91 – 94	0.01

In general, filler grades are produced to meet brightness and grit specifications. The lowest price filler grade is air-floated, whereas the highest price filler grade is the calcined kaolin extender. A good review of the properties of paper affected by fillers is presented by CASEY [96] (⟶ Paper and Pulp).

5.1.3. *Paint*

Kaolins are used extensively in water-based paints. Kaolin is a functional extender pigment that has good covering or hiding power, imparts desirable flow properties to the paint, and is inexpensive. *Brightness* and *particle size* are important physical properties, as is a *resistivity specification*. A resistivity test gives an index of the amount of residual soluble salts that remain in the clay. Higher resistivity values reflect a low soluble salt content, which is important because high soluble salt content adversely affects the dispersion of the paint ingredients and alters paint properties.

Another important physical property is *oil absorption*, which gives an indication of surface area and which is related to a parameter called vehicle demand. A test commonly used by the paint manufacturer to measure dispersion is a fineness of grind measurement using a Hegman gage. Actual paint formulations are made, and many performance tests are run, including gloss, color, smoothness, flow characteristics, dispersion, stability, weathering characteristics, aging, washability, and hiding power.

The kaolins used in paint vary from very coarse, which produces a matte finish, to very fine, which gives a high-gloss finish. Calcined kaolins are used in paint as a titanium dioxide replacement and extender. Delaminated clays are used to give a high-gloss finish with good covering power, good color, and excellent washability (⟶ Paints and Coatings).

5.1.4. *Ceramics*

The ceramic industry is a large user of kaolin clays in whiteware, sanitary ware, insulators, and refractories. Several tests are used to evaluate kaolins for use in ceramics, including plasticity, shrinkage, modulus of rupture, absorption, fired color, casting rate, PCE (see Section Kaolin), and chemical analysis.

Coarseparticle-size clays are used for casting clays in sanitary ware, and fine-particle kaolins are used when high strength is important, such as in whiteware and as a bond in certain refractory applications. Ball clays, which are a fine-particlesize, very plastic kaolinitic clay, are used in whiteware, stoneware, enamels, etc. Calcined kaolin called grog is marketed to the refractory industry for use in high-alumina refractories. Both dry-process and wet-process kaolins are used by the ceramic industry (⟶ Ceramics, General Survey).

5.1.5. *Rubber*

Kaolin is used by rubber manufacturers because of its reinforcing and stiffening properties and because of its low cost. Fine-particle-size kaolins give good resistance to abrasion and are used extensively in nonblack rubber goods. The kaolins that are used in rubber contain a maximum of 1 % free moisture. Both dry-process and wet-process kaolins are used, but by far the largest tonnage are dry-process grades. Brightness, particle size, and grit percentage are important physical properties. Other tests that are important in evaluating kaolin for use in rubber are water settling characteristics and oil absorption. The water settling test gives a good idea of the reinforcing characteristics of the clay, and oil absorption also correlates with reinforcement. Tests on the clay-filled rubber itself are stress – strain, tear resistance, abrasion resistance, heat generation, energy rebound, extrusion and plasticity, hardness, aging characteristics, and water absorption [97] (⟶ Rubber, 4. Chemicals and Additives).

5.1.6. *Plastics*

Kaolin is used in plastics because it aids in producing a smooth surface finish, reduces cracking and shrinkage during curing, obscures the fiber pattern when fiberglass is used as reinforcement, contributes to a high dielectric strength, improves resistance to chemical action and weathering, and helps control the flow properties. Kaolin is a functional filler in that it contributes some beneficial effects. Loadings in various plastic compositions vary from ca. 15 % to as high as 60 % (⟶ Plastics, General Survey).

5.1.7. *Ink*

Kaolin clays are used extensively in printing inks. In black inks the pigment is normally carbon black, but in white and other colored inks kaolin is a common extender pigment. The kaolins used in ink must be fine particle size and relatively white. A major disadvantage to the use of kaolin in ink is its hydrophilic nature. However, coating the kaolin particles with hydrophobic material increases its usefulness in inks. Surface-modified kaolins have found a significant market in printing inks (⟶ Drawing and Writing Materials).

5.1.8. *Catalysts*

Kaolin clays are used extensively in the production of cracking catalysts and molecular sieves. The kaolin clay should have relatively low iron and sulfate content. Although the surface of the kaolin does contribute some catalytic activity, the kaolin serves largely as a diluent for the silica – alumina gels [98]. Partially calcined kaolins, which are heated to a temperature just above the dehydroxylation temperature of ca. 600 °C, are used in the manufacture of molecular sieves [99] (⟶ Heterogeneous Catalysis and Solid Catalysts).

5.1.9. *Fiberglass*

Kaolin clays with a low iron content are used as a source of alumina and silica in the manufacture of fiberglass. A good grade of air-float kaolin is generally used (⟶ Fibers, 5. Synthetic Inorganic).

5.2. Smectite

Smectite is a group name, and many minerals comprise the smectite group, including sodium montmorillonite, calcium montmorillonite, saponite, nontronite, hectorite, and sauconite. Each of these minerals has unique properties, and therefore, the individual smectite minerals and the other minerals that are present in the bentonite must be determined. This is done by X-ray diffraction, electron microscopy, differential thermal analysis, infrared spectrometry, and selective chemical analysis. Once the smectite mineral has been identified, then some of the properties can be established, for example, viscosity, thixotropy, plasticity, shrinkage, bonding strength, shear strength, and water impedance. These properties are not the same for all smectites and, in fact, are quite different.

The exchangeable cations present between the silica and alumina sheets (see Fig. 5) have a strong influence on the use and properties of the smectite. The major uses of smectites (bentonites) are in drilling muds, foundry sand bond, and iron-ore pelletizing. In addition to these three major uses, bentonite is used for many miscellaneous products, including filtering agents, water impedance, cosmetics, animal feed, pharmaceuticals, paint, ceramics, slurry trenching, catalysts, and decolorizing.

Drilling Fluids. Sodium montmorillonite, which is the major constituent of the Wyoming or Western bentonites, is a high-swelling clay with some unique properties. It is used throughout the world as an ingredient in drilling fluids because commonly this Western bentonite yields 100 barrels per ton (20 m^3/t) [35]. Only 5 % of this bentonite is necessary to produce the high viscosity, thixotropy, and low filter-cake permeability required in the drilling fluid. Hectorite, the lithium smectite, also gives high yields and has excellent properties as an ingredient in drilling fluids. Some calcium montmorillonites can be treated with a sodium salt, usually soda ash, to produce a bentonite suitable for use in drilling fluids, but in general such treatment does not substantially improve the water loss property of the bentonite.

Foundry Molding Sands. The molding sands used in foundries are composed of sand and clay. The clay provides bonding strength and plasticity. A small amount of tempering water is added to the mixture to make it plastic. It can then be molded around a pattern and be cohesive enough to maintain the shape after the pattern is removed and while molten metal is poured into it. The important foundry properties are green compression strength, dry compression strength, hot strength, flowability, permeability, and durability. These properties vary greatly with the amount of tempering water [35]. Both sodium and calcium bentonites (montmorillonites) are used as bonding clays. Calcium bentonite has a higher green strength, lower dry strength, lower hot strength, and better flowability than sodium bentonite. Final evaluations of the bonding clay can only be made after the bentonite is actually used in foundry practice so that flowability, durability, ease of shakeout of the casting from the mold, and cleanness of the surface of the cast metal can be determined. Commonly, blends of sodium and calcium bentonites are used to gain the optimum properties from each.

Iron-Ore Pelletizing. Bentonites are used extensively in pelletizing iron ores [100]. Finely pulverized ore concentrate must be pelletized into units of ca. 1 in. (2.5 cm) or more in diameter before it is used as blast furnace feed. Pelletized ore is a superior blast furnace feed. Bentonite is ca. 0.5 wt % of the ore. Because of its superior dry strength sodium bentonite is the preferred bentonite for pelletizing ores.

Other Uses. Acid-activated bentonites are widely used to *decolorize* mineral, vegetable, and animal oils. The clay also serves to deodorize, dehydrate, and neutralize the oils. In addition, the clay must have low oil retention and good filtration characteristics. It must not change the character of the oil nor give an objectionable taste to the oil. Some bentonitic clays naturally possess adequate decolorizing ability. Such materials are called fuller's earth. The bentonites that work best as both natural and acid-activation decolorizers are the calcium variety.

Some clay *catalysts* for cracking petroleum are produced from bentonites with very low iron content. The bentonite is reacted with sulfuric or hydrochloric acid at elevated temperatures followed by washing, drying, and calcining at ca. 500 °C. This process removes the adsorbed alkali metals and alkaline-earth metals along with the removal of iron and partial removal of magnesium and aluminum from the lattice. This increases the surface area and alters the pore-size distribution of the product. Calcium bentonite is the preferred material for making a smectite clay catalyst.

Sodium bentonites are used in *water impedance*. The clay is used to impede the movement of water through earthen structures such as dams, to stop seepage of water from ponds and irrigation ditches, and to stop water from entering basements of homes. One use of high-swelling sodium bentonite is in the slurry-trench or diaphragm-wall method of excavation in construction in areas of unconsolidated rock or soil [101]. In this method, the trench being excavated is filled with bentonite slurry, and the earth being excavated is removed through it. A thin filter cake on the walls of the excavation prevents loss of fluid, and the hydrostatic head of the slurry prevents caving and running of loose soil, which makes costly shoring unnecessary. Small amounts of sodium bentonites are added to cement to improve workability, lessen aggregate segregation, and enhance impermeability. Another growing use is for liners in landfills for waste disposal and for toxic wastes [104].

Sodium bentonites are widely used as *emulsifying* and *suspending agents* in many oil and water systems and in medicinal, pharmaceutical, and cosmetic formulations [102]. Sodium bentonites are used in both oil- and water-based paints. In water-based paints, the bentonite acts as a suspending and thickening agent. In oil-based paints, it is used as an emulsifying agent.

In *paper*, sodium bentonite is sometimes used in the deinking process to recover the cellulose fibers. These deinking processes generally involve heating the paper to be deinked in a caustic soda solution to break down the ink, thus freeing the ink pigment. Next a detergent is added to force the pigment from the paper fiber. Smectite is then added to disperse the pigment particles and adsorb them. Washing removes the smectite, which carries the ink pigment with it. Sodium bentonite is also used in papermaking to prevent the agglomeration of pitch, tar, wax, and resinous material. The addition of 0.5 %

smectite, based on the dry weight of the paper stock, prevents agglomerates so that the globules will not stick to screens, machine wires, and press rolls and thus cause defects in the paper.

Sodium montmorillonite is the base for the organoclays described earlier. The organoclad clays are used in lubricating greases, inks, paints, and oil-based drilling fluids. These clays improve pigment suspension, viscosity, and thixotropy control. They also improve brushability and spraying characteristics of paints [105]. Another area of research and utilization is to process sodium montmorillonite to produce nanocomposites for use in improving the physical properties of certain plastic compositions [106]

Calcium montmorillonite is used in adhesives; as granular pet litter; to clarify wines, cider, and beer; as floor absorbents; carriers and diluents in pesticide preparations; and in water clarification. White calcium smectites are used in detergents, paints, enamel suspensions, pharmaceuticals, cosmetics, food, pelletized animal feed, and in no-carbon-required copy paper [28].

Both calcium and sodium smectites are used as additives in ceramics, particularly because of their plasticity and relatively high green and dry strengths. For example, many brick clays are silty materials. The addition of a small amount of smectite improves the plasticity and green and dry strengths. In addition, both types are used as a liner for toxic and low-level radioactive waste disposal sites and as additives to latex for thickening and stabilizing.

5.3. Palygorskite and Sepiolite

The use of the palygorskite and sepiolite is almost as varied as that of the kaolins and smectites. Because of the particle shape of the these clays, i.e., palygorskite (attapulgite) and sepiolite are elongate (see Fig. 10), they have some special applications. The specific type can be identified with X-ray diffraction and is generally characterized by using scanning electron microscopy to determine the crystal size and length. In addition to the crystal shape, another set of important properties is the absorption and adsorption characteristics.

One of the more important, larger uses is *drilling fluids*. Considerable quantities of palygorskite and sepiolite are used in drilling fluids because the viscosity and gel strength of the mud are not affected by variations in electrolytic content as are those of muds comprised of bentonite. Thus, palygorskite and sepiolite drilling muds can be used with salt water or when formation brines become a serious problem. In recent years, the offshore drilling industry has used salt-water-based muds, so that the quantities of palygorskite and sepiolite used have increased dramatically.

Because palygorskite and sepiolite are not easily flocculated on account of their particle shape, these clay minerals are used as *suspending agents* in paint, medicines, pharmaceuticals, and cosmetics. In paint, palygorskite is used because it improves the thickening and thixotropic properties. Another major use is as *floor sweep compounds*, for absorbing oil and grease spills on factory floors, in service stations, and other areas where oil and grease spills are a problem. Granular particles of palygorskite are the most effective floor absorbent.

Palygorskite and sepiolite are also used extensively in *agriculture* as absorbents and adsorbents for chemicals and pesticides. The chemical is mixed onto the granular hormite clay particle, and the treated particle is placed in the ground with the seed particle. The pesticide or fertilizer chemical is slowly released to provide the necessary fertilizer or protection for the growing plant. Finely pulverized hormite clays are also used to adsorb chemicals, which can then be dusted or sprayed on the plant or on the surface of the ground. A growing application is to use palygorskite (attapulgite) to make suspension fertilizers.

Another large use for palygorskite and sepiolite is *pet litter*. Granular particles of are an effective litter for absorbing animal waste, particularly for domestic cats. The clay effectively absorbs the waste and odors.

HADEN [103] described the properties and uses of palygorskite and pointed out both the colloidal and noncolloidal characteristics that make this mineral important. Other uses of hormites include foundry sand binder, polishes, wax emulsions, metal drawing lubricants, laundry washing powders, bonding agents for granulation of powders, decolorizing oils, anti-caking agents, flattening agent in paints, catalyst in no-carbon-required paper, and catalyst carrier.

[Top of Page]

6. Environmental Aspects

Almost all clays are surface mined, so the industry is required to reclaim the disturbed land in most countries. Common practice is to open a cut and then spoil the overburden from the following panels or cuts into the mined out areas. The land is leveled or sloped to meet the governmental requirements and then planted with grasses or trees.

In the processing plants the waste materials, which include chemicals and separated clay particles or other minerals, are collected in impounds. The clay and other particles are flocculated with alum or other chemical flocculants, and the clear water is released into streams after adjusting the pH to 6 – 8.

Air quality is maintained in the processing plants by using dust collectors on the dryers and by enclosing transfer points where dry clay may cause dust. In areas such as the bagging departments, the workers may be required to wear dust masks. None of the clays, kaolins, smectites, or hormites are health hazards. Any dust, if inhaled in large quantities, can be a problem, but the industries use dust collectors and require workers to wear masks. Therefore, no serious lung ailments result. The clays themselves contain no deleterious trace elements or chemicals that are hazardous to human health.

[Top of Page]

7. Production and Consumption

Clays are important industrial minerals in most parts of the world. As indicated in Chapter Properties and Uses they are utilized in many process industries and are necessary ingredients in a great many products: kaolins are used in paper, ceramics, paint, etc.; bentonites, in drilling fluids and foundries; and palygorskite and sepiolite, in special drilling fluids and as suspending agents in paints and pharmaceuticals.

The total kaolin production in the world is about 40 × 10^6 t/a. The largest producer is the United States with about 6.2 × 10^6 t/a. The paper and ceramic industries are the largest consumers worldwide. Paper-coating kaolins are produced in Brazil, England, and the United States. The largest markets for these coating clays are in Europe, Japan, and North America.

Bentonite production is about 11.5 × 10^6 t/a in the world, and the United States has the largest production of ca. 6.2 × 10^6 t/a. The high-viscosity, high-swelling sodium bentonites from Wyoming, Montana, and South Dakota are exported around the world for use in drilling muds for oil and gas exploration, as high-strength foundry bond clays, and for pelletizing iron ore for shipment.

Palygorskite and sepiolite production in the world is ca. 2.8 × 10^6 t/a, and the United States is the largest producing country with ca. 1 × 10^6 t/a. Spain is the second largest producer with ca. 800 000 t/a. The largest consumers of these clays are pet litter, drilling muds, carriers for agricultural chemicals, and oil and grease absorbents.

No figures are available for the world production of common clays and shales but it is probably over 100 × 10^6 t/a. These common clays and shales are used extensively in most countries to make brick and other structural clay products.

In summary, clays are one of the most important and versatile industrial minerals. They are used in construction, agriculture, the many process industries, and environmental applications [107]. Clay minerals and materials are about the least expensive ingredients that are used in many products even though they are very functional. Some important value added products such as calcined kaolins, organoclays, and nanocomposites have been developed, and future research efforts will continue to develop new value-added products. The higher quality clays — kaolins for paper coating, sodium bentonite for drilling muds, and palygorskite and sepiolite for sorbent uses — are exported to markets all around the world, which continue to grow. Fortunately, the reserves of these clays are large so that the markets will continue to be served for the foreseeable future.

[Top of Page]

References

1. C. K. Wentworth: “A Scale of Grade and Class Terms for Clastic Sediments,” *J. Geol.* **30** (1922) 377 – 392.
2. R. E. Grim: *Clay Mineralogy*, 1st ed., McGraw-Hill, New York 1953, p. 384.
3. G. W. Brindley, G. Brown (eds.): *Crystal Structures of Clay Minerals and Their X-ray Identification*, Monograph 5, Mineralogical Soc., London 1980, p. 495.
4. W. J. Smothers, Y. Chiang: *Handbook of Differential Thermal Analysis*, Chem. Pub. Co., New York 1966, p. 633.
5. H. Beutelspacher, H. W. Van Der Marel: *Atlas of Electron Microscopy of Clay Minerals and Their Admixtures*, Elsevier, Amsterdam 1968, p. 333.
6. H. W. Van Der Marel, H. Beutelspacher: *Atlas of Infrared Spectroscopy of Clay Minerals and Their Admixtures*, Elsevier, Amsterdam 1976, p. 396.
7. C. S. Ross, P. F. Kerr: “Halloysite and Allophane,” Geol. Surv. Prof. Paper (U.S.) no. 185-G, (1934) 135 – 158.
8. C. S. Ross, P. F. Kerr: “The Kaolin Minerals,” Geol. Surv. Prof. Paper (U.S.) no. 165-E, (1931) 151 – 180.
9. T. F. Bates *et al.*: “Morphology and Structure of Endellite and Halloysite,” *Am. Mineral.* **35** (1950) 463 – 484. Links
10. J. W. Gruner: “Vermiculite and Hydrobiotite Structures,” *Am. Mineral.* **19** (1934) 557 – 575. Links
11. R. E. Grim *et al.*: “The Mica in Argillaceous Sediments,” *Am. Mineral.* **22** (1937) 813 – 829. Links
12. S. W. Bailey: “Chlorites” in J. E. Gieseking (ed.): *Soil Components*, vol. II., Springer-Verlag, New York 1975, pp. 191 – 263.
13. W. F. Bradley: “The Structural Scheme of Attapulgite,” *Am. Mineral.* **25** (1940) 405 – 410. Links
14. R. C. Reynolds in [3], pp. 249 – 303.
15. H. H. Murray in S. Shimod (ed.): *7th Int. Kaolin Sym. Proc.*, Univ. of Tokyo, 1976, pp. 114 – 125.
16. E. M. Durrance *et al.*: “Hydrothermal Circulation and Post-Magnetic Changes in Granites of Southwest England,” *Proc. Ann. Conf., Usher Soc.*, 1982, pp. 304 – 320.
17. C. M. Bristow: “Kaolin Deposits of the United Kingdom of Great Britain and Northern Ireland,” *Int. Geol. Congr. Rep. Sess. 23rd*, (1969) 275 – 288.
18. H. H. Murray: *New World Class Industrial Clay Deposits*, Preprint 99-135, SME Annual Meeting, Denver, Colorado, 1999.
19. M. Kuzvart: “Kaolin Deposits of Czechoslovakia,” *Int. Geol. Congr. Rep. Sess. 23rd*, (1969) 25 – 32.
20. M. Storr: *Kaolin Deposits of The GDR in The Northern Region of the Bohemian Massif, Guidebook, 5th Int. Kaolin Sym.*, Ernst-Moritz-Arndt Univ., Greifswald 1975, p. 243.
21. H. M. Koster in S. Shimoda (ed.): *7th Int. Kaolin Sym. Proc.*, Univ. of Tokyo, 1976, p. 83.
22. L. Martin-Vivaldi: “Kaolin Deposits of Spain,” *Int. Geol. Congr. Rep. Sess. 23rd*, (1969) 225 – 261.
23. E. Galan Huertos, “Caolines Espanoles: Geologia, Mineralogia, y Genesis”. Tesis Doctoral, Sociedad Espanola de Ceramica y Vidrio (1972) VIII Partes.
24. P. Moore: “Kaolin — White Gold or White Dirt”, *Ind. Minerals Mag.* (2003) August, 24–35.

25. H. H. Murray, C. C. Harvey: *Australian and New Zealand Clays*, Preprint 82 – 83, 1st Int. SME-AIME Fall Mtg., Honolulu, Hawaii, 1982, p. 5.
26. Z. Zheng *et al.*: "Clay Deposits of China," in *Mineral Deposits of China*, vol. **4**, 1996, pp. 235 – 287.
27. J. Yuan, H. H. Murray: *Mineralogical and Physical Properties of the Maoming Kaolin from Guandong Province, South China*, Special Publication Number 1, Clay Minerals Society, 1993, pp. 249 – 260.
28. H. H. Murray in R. W. Hagemeyer (ed.): *Paper Coating Pigments*, Tappi Press, Atlanta, Ga., 1984, pp. 95 – 141.
29. S. H. Patterson, H. H. Murray: "Kaolin, Refractory Clay, Ball Clay, and Halloysite in N. America, Hawaii, and the Caribbean Region," *Geol. Surv. Prof. Pap. U.S.* no. 1306, (1984) 56.
30. M. Willis: "An Eye on the Ball", *Ind. Minerals Mag.* (2002) October, 26–33.
31. F. H. Norton: *Refractories*, 4th ed., McGraw-Hill, New York 1968, p. 450.
32. W. C. Knight: "Bentonite," *Eng. Min. J.* **66** (1898) 491.
33. C. S. Ross, E. V. Shannon: "Minerals of Bentonite and Related Clays and Their Physical Properties," *J. Am. Ceram. Soc.* **9** (1926) 77 – 96. Links
34. P. C. Wright: "Meandu Creek Bentonite — A Reply," *J. Geol. Soc. Aust.* **15** (1968) 347 – 350.
35. R. E. Grim, N. Guven: "Bentonites, Geology, Mineralogy, Properties and Uses," *Devel. Sedimentol.* **24** (1978) 256.
36. J. Elzea, H. H. Murray, "Bentonite," in D. D. Carr (ed.): *Industrial Minerals and Rocks*, 6th ed., SME, 1994, pp. 233 – 246.
37. M. M. Knechtel, S. H. Patterson: "Bentonite Deposits of the Northern Black Hills District, Wyoming, Montana, and South Dakota," *U.S. Geol. Surv. Bull.* **1082** (1962) 893 – 1029.
38. M. Slaughter, J. M. Early: "Mineralogy and Geological Significance of the Mowry Bentonites," *Spec. Pap. Geol. Soc. Am.* **83** (1965) 116.
39. R. E. Grim: "Preliminary Report on Bentonite in Mississippi," *Miss. State Geol. Surv. Bull.* **22** (1928) 14.
40. R. E. Grim, G. Kulbicki: "Montmorillonite: High Temperature Reaction and Classification," *Am. Mineral.* **46** (1961) 1329 – 1369. Links
41. A. F. Hagner: "Adsorptive Clays of the Texas Gulf Coast," *Am. Mineral.* **24** (1939) 67 – 108. Links
42. P. Y. Chen: *Geology and Mineralogy of the White Bentonite Beds of Gonzales County, Texas*, Ph. D. Thesis, University of Texas 1968.
43. R. L. Sloan, J. M. Gilbert: "Electron-Optical Study of Alterations in the Cheto Clay Deposit," *Clays Clay Miner.* **15** (1966) 35 – 44.
44. K. G. Papke: "Montmorillonite Deposits in Nevada," *Clays Clay Miner.* **17** (1969) 211 – 222.
45. W. H. Monroe: "Notes on Deposits of Selma-Ripley Age in Alabama," *Ala. Geol. Surv. Bull.* **48** (1941) 150.
46. L. L. Ames, Jr., *et al.*: "A Contribution on the Hector, California, Bentonite Deposit," *Econ. Geol.* **53** (1958) 22 – 37.
47. C. S. Denny, H. Drewes: "Geology of the Ash Meadows Quadrangle, Nevada – California", *U.S. Geol. Surv. Bull.* **1181-L** (1965) 56.
48. I. A. Cowperthwaite et al.: "Sedimentation, Petrogenesis, and Radioisotopic Age of the Cretaceous Fuller's Earth of Southern England," *Clay Miner. Bull.* **9** (1972) 309 – 327.
49. A. Hallam, B. W. Selwood: "Origin of the Fuller's Earth in the Mesozoic of Southern England," *Nature (London)* **220** (1968) 1193 – 1195.
50. R. Fahn: "The Mining and Preparation of Bentonite," *InterCeram* **12** (1965) 119 – 122.
51. W. A. Allo, H. H. Murray: "Mineralogy, Chemistry, and Potential Applications of a White Bentonite in San Juan Province, Argentina", *Appl. Clay Sci.* **25** (2004), 237–243. Links
52. V. Szeky-Fux: "The Hydrothermal Genesis of Bentonites on the Basis of Studies in Komloska," *Acta Geol. Acad. Sci. Hung.* **4** (1957) 361 – 382.
53. V. Annedda: "Deposits of Bentonite on Sadali and Vallanova Tulo Territories, Sardinia", *Resoconti Assoc. Min. Sarda* **60** (1956) 5 – 9.
54. T. E. Wayland: "Geologic Occurrence and Evaluation of Bentonite Deposits," *Trans. AIME* **250** (1971) 120 – 132.
55. E. Lukas: "State of Exploration of Zaluska Gorica Bentonites in Comparison with other Yugoslavian Bentonites," *Rud. Metal. Zb.* 1968, no. 3, 277 – 286.
56. N. N. Siddiguie, D. P. Bahl: "Geology of the Bentonite Deposits of the Barmer District, Rajasthan," *Mem. Geol. Surv. India* **96** (1965) 36.
57. S. Iwao: "The Clays of Japan," *Int. Clay Conf. Proc. 1969*, 209 pp.,
58. M. A. Rateev: "Sequence of Hydrothermal Alteration of Volcanic Rocks into Bentonite Clays of the Askana Deposit in the Georgian S.S.R.," *Dokl. Akad. Nauk SSSR* **175** (1967) 675 – 678.
59. G. S. Dsotsenidze, E. A. Matchabeley: "The Genesis of Bentonites of Georgian S.S.R.," *Int. Clay Conf. Proc. 1963*, pp. 203 – 210.
60. R. H. S. Robertson: *Fuller's Earth — A History of Calcium Montmorillonite*, Min. Soc. Occasional Publication, Volturna Press, Hyth, Kent, U.K., 1986.
61. "Bentonite, Sepiolite, Attapulgite," *Industrial Minerals Magazine* 1978, no. 126, 50.
62. S. H. Patterson: "Fuller's Earth and Other Industrial Mineral Resources of the Meigs-Attapulgus-Quincy District, Georgia and Florida," *Geol. Surv. Prof. Pap. (U.S.)* no. 828 (1974) 45.
63. L. Wirth: *Attapulgites du Sénégal Occidental*, Rapport No. 26, Laboratoire de Géologie, Faculté des Sciences, Université de Dakar 1968, p. 55.
64. F. D. Ovcharenko *et al.*: *The Colloid Chemistry of Palygorskite*, Israel Program for Scientific Translations, Jerusalem 1964, p. 101.

65. E. Galan, J. J. Brell, A. La Iglesia, R. H. S. Robertson: “The Caceres Palygorskite Deposit,” *Proc. Int. Clay Conf. 1975*, (1976) 81 – 94.
66. H. N. Khoury, D. D. Eberl, B. F. Jones: “Origin of Magnesium Clays from the Amargosa Desert, Nevada,” *Clays Clay Miner.* **30** (1982) 327 – 336.
67. M. K. H. Siddiqui: *Bleaching Earths*, Pergamon Press, New York 1968, p. 86.
68. H. H. Murray: “Clay,” in J. L. Gilson (ed.): *Industrial Minerals and Rocks*, AIME, New York 1960, pp. 259 – 284.
69. H. H. Murray, J. M. Smith: “Lightweight Aggregate Potentialities of Some Indiana Shales”, *Indiana Geol. Sur. Rep. Progr.* **12** (1958) 42.
70. S. M. Pickering, H. H. Murray, “Kaolin” in D. D. Carr (ed.): *Industrial Minerals and Rocks*, 6th ed., SME, 1994, pp. 233 – 246.
71. H. H. Murray: “Major Kaolin Processing Developments,” *Int. J. Miner. Process.* **7** (1980) 263 – 274. Links
72. H. H. Murray: “Dry Processing of Clay and Kaolin,” *InterCeram.* **31** (1982) no. 2, 108 – 110.
73. H. H. Murray: “Equipment for Wet Processing of Clay Material,” *InterCeram* **31** (1982) no. 3, 196 – 197.
74. F. A. Gunn, H. H. Morris, US 3 171 718, 1965.
75. E. W. Greene, J. B. Duke: “Selective Froth Flotation of Ultrafine Minerals or Slimes,” *Trans. Soc. Min. Eng. AIME* **223** (1962) 389 – 395.
76. W. M. Bundy, J. P. Berberich, US 3 477 809, 1969.
77. R. N. Maynard, B. R. Skipper, N. Millman, US 3 371 988, 1968.
78. V. V. Mercade, US 3 701 417, 1972.
79. J. Iannicelli: “New Developments in Magnetic Separation,” *IEEE Trans. Magn., MAG-12* (1976) 436 – 443.
80. J. A. Wernham, T. H. Falconer: *The Eriez Commercial Superconducting High Gradient Separators (Abstract)*, Conf. on Superconducting Magnetic Separation, Imperial College, London 1985, p. 2.
81. H. R. Fanselow, D. A. Jacobs, US 3 586 523, 1971.
82. W. M. Bundy, J. L. Harrison, J. G. Ishley: “Chemically Induced Kaolin Floc Structures for Improved Paper Coating,” *Proc. Tech. Assoc. Pulp Pap. Ind.* 1983, 175 – 187.
83. R. F. Grim: *Applied Clay Mineralogy*, McGraw-Hill, New York 1962, p. 375.
84. P. G. Nahin, US 3 248 314, 1966.
85. P. W. Libby, J. Iannicelli, C. R. McGill: “Elastomer Reinforcement with Amino Silane Grafted Kaolin,” *Abstr. Pap. Am. Chem. Soc. 154th 1967* (Div. of Rubber Chem.).
86. K. H. Teague, “Southern Bentonite,” SME Preprint No. 72 H 328 (1972), 7 pp.
87. J. W. Jordan, US 2 531 440, 1950.
88. H. H. Murray: “Kaolin — A Versatile Pigment, Extender, and Filler,” *Prepr. Pap. Annu. Meet. Tech. Sect. C.P.P.A. 63rd*, 1977, A29 – 34.
89. H. H. Murray: *Industrial Applications of Kaolin, Clays and Clay Minerals*, vol. **10**, Pergamon Press, London 1963, pp. 291 – 298.
90. H. H. Murray: “Diagnostic Tests for Evaluation of Kaolin Physical Properties,” *Acta Mineral.-Petrogr.*, **23** (1980) 67 – 76.
91. A. S. Michaels: “Deflocculation of Kaolinite by Alkali Polyphosphates,” *Ind. Eng. Chem.* **59** (1958) 951 – 958.
92. W. M. Bundy, W. D. Johns, H. H. Murray: “Physico-Chemical Properties of Kaolinite and Relationship to Paper Coating Quality,” *Tappi* **48** (1965) 688 – 696.
93. G. A. Hemstock, J. W. Swanson: “The IGT Printability Tester as an Instrument for Measuring the Adhesive Strength of Pigment-Coated Paper,” *Tappi* **40** (1957) 794 – 801.
94. A. C. Eames: “The Transverse Tensile Strength of Clay-Starch Coatings,” *Tappi* **43** (1960) 2 – 10.
95. W. R. Willets: *Paper Loading Materials*, Introduction-Tappi Monograph 19, New York 1958, p. 5.
96. J. P. Casey: *Pulp and Paper. Papermaking* (vol. **2**), Interscience, New York 1960, pp. 985 – 1020.
97. *Kaolin Clays and Their Industrial Uses*, J. M. Huber Corp., New York 1955, p. 214.
98. J. S. Magee, J. J. Blazek in “Preparation and Performance of Zeolite Cracking Catalysts”, *ACS Monogr.* **171** (1976) 615 – 679.
99. D. W. Breck: *Zeolite Molecular Sieves*, J. Wiley & Sons, New York 1974, pp. 314 – 320.
100. F. D. DeVaney, US 2 743 172, 1956.
101. W. J. Lang: *Bentonite: The Demand and Markets for the Future*, SME Preprint No. 71 H 29, AIME, 1971, p. 9.
102. G. Novelli, “Properties and Possible Use of Montmorillonite in Pharmaceutical and Cosmetic Preparation,” *Riv. Ital. Essenze, Profumi, Piante Off., Olii Veg., Saponi* **54** (1972) 267 – 270.
103. W. L. Haden, Jr., “Attapulgite: Properties and Uses,” *Clays Clay Miner.* **10** (1963) 284 – 290.
104. K. S. Keith, H. H. Murray: “Clay Liners and Barriers,” in D. D. Carr (ed.): *Industrial Minerals and Rocks*, 6th ed., SME, 1994, pp. 435 – 452.
105. W. F. Moll, Jr.: *Organobentonites — Constitution and Applications*, Abstracts and Program, SME Annual Meeting, Orlando, Florida, 1998, p. 150.
106. G. W. Beall, S. J. Tsipursky: *Computational Modeling of Intercalated Clays Utilized in Nanocomposites*, Meeting Program and Abstracts, Clay Minerals Society Annual Meeting, Cleveland, Ohio, 1998, p. 87.
107. H. H. Murray: “Clays in Industry and the Environment,” in G. J. Churchman, R. W. Fitzpatrick, R. A. Eggleton (eds.): *Clays Controlling the Environment*, Proceedings of the 10th International Clay Conference, Adelaide, Australia, 1993, pp. 49 – 55.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Cleansing Agents

Christian Nitsch, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 1.3)

Hans-Joachim Heitland, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 1)

Horst Marsen, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 1)

Hans-Joachim Schlüssler, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 2)

1. Household and Institutional Cleansing Products 1
1.1. All-Purpose Cleansers 2
1.2. Scouring Cleansers 3
1.3. Dishwashing Agents 4
1.3.1. Manual Dishwashing Detergents . . 4
1.3.1.1. Classical Dishwashing Detergents . 5
1.3.1.2. Dishwashing Concentrates 6
1.3.1.3. Skin-Friendly Products 6
1.3.1.4. Liquids with Antibacterial Action . 7
1.3.1.5. Dishwashing Habits 7
1.3.2. Automatic Dishwashing Detergents 7
1.3.2.1. Dishwashers 8
1.3.2.2. Cleaners 9
1.3.2.3. Rinse Aids 12
1.3.2.4. Regeneration Salt 12
1.3.2.5. Other Products 13
1.3.2.6. Performance Assessment 13
1.4. Floor Cleaners 14
1.5. Carpet Cleaners 15
1.6. Toilet Bowl Cleaners 17
1.7. Toilet Hygiene and Deodorant Products 17
1.8. Drain Cleaners 17
1.9. Metal Cleaners 18
1.10. Window Cleaners 18
1.11. Automotive Cleansers 18
1.12. Building Facade Cleaners 19
2. Cleansers and Disinfectants for Use in the Food Industry 20
2.1. Cleansers 20
2.2. Disinfectants 22
3. References 23

1. Household and Institutional Cleansing Products

For centuries, people cleaned pots, pans, kettles, and other household objects with fine sand, occasionally mixing it with wood ashes or soda. However, with the advent of goods fabricated from increasingly sensitive materials, including polished metal, painted wood, porcelain, and plastics, a need for more specialized cleansing agents arose.

The varied mechanical and chemical resistances of the surfaces to be cleaned and the chemical heterogeneity of the dirt to be removed require a multiplicity of chemical agents, not all of which are compatible with one another. Moreover, cleansing agents that leave no visible residues, produce adequate sudsing, and have pleasant aromas are desired. Finally, the components should be gentle to the skin, they must be regarded as safe from a toxicological standpoint, and the products containing them need to have a long shelf life. In some countries, both the individual components and the cleansing mixtures themselves must be biodegradable. Many types of cleansing agents have emerged to satisfy all of these demands.

Mode of Action. Household cleansing operations are normally carried out at room temperature, primarily because the small amounts of cleansing solution typically employed rapidly acquire the temperature of the surface being cleaned. Dishwashing represents a special case, in which utensils are submerged in a hot solution of the dishwashing agent, whereby they assume approximately the temperature of the bath. These higher temperatures simplify the cleansing operation both by melting grease residues and by accelerating physicochemical and purely chemical processes.

Modern cleansing agents consist largely of surfactants and sequestering agents; hence, their primary action is physicochemical in nature. Strictly chemical phenomena (such as the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_137

Table 1. Partial compositions for three types of aqueous all-purpose cleansers

Component	Type of cleanser		
	Cleanser containing no organic solvent, wt %	Cleanser containing an organic solvent, wt %	Cleanser containing a disinfectant, wt %
Alkylbenzenesulfonate [*75534-59-7*]	10	6	10
Fatty alcohol polyglycol ether [*68213-23-0*]	2	4	4
Pentasodium triphosphate [*7758-29-4*]	3	5	4
Urea [*57-13-6*]	5	5	4
Glycol ether [*20324-32-7*]		5	
Aldehyde – amine condensation product [*4719-04-4*]			6
Colorant/perfume	present	present	present

saponification of fats or the degradation of starches and proteins) are of lesser significance. The chelating properties of the polymeric phosphates, citrates, or other sequestering agents that are present is of considerable importance, however, in that their ability to remove calcium and magnesium ions reduces the extent to which dirty residues adhere, making them easier to remove.

The primary steps in cleansing are adsorption phenomena occurring at liquid – vapor, liquid – liquid, and liquid – solid phase boundaries. Surfactant adsorption occurs in a nonspecific fashion at all interfaces, whereas that of sequestering agents is specific to polar solids. Surfactant adsorption leads to a reduction in interfacial tension, which allows for improved wetting of solids and, if oily residues are present on the surface, displacement of the oils by roll-up or by micelle formation. The removal of pigmented soils is eased by the fact that when surfactants or sequestering agents are adsorbed at a liquid – solid boundary, they increase the surface potential of both the substrate and the soil, which increases their mutual electrostatic repulsion. As a result, the dirt particles are readily removed from the surface and enter the cleansing solution, where they remain well dispersed [1–3].

1.1. All-Purpose Cleansers

All-purpose cleansers have a wide range of applicability. Their composition is such that they are compatible with any household or commercial surface amenable to wet treatment. Originally, most all-purpose cleansers were in powdered form and contained soda. More recently, however, liquid agents have gained wide popularity. These contain mixtures of anionic surfactants, nonionic surfactants, polymeric phosphates or other sequestering agents, solvents, hydrotropic substances, polymeric compounds, corrosion inhibitors, skin protective agents, perfumes, and colorants. Typical formulations are given in Table 1.

All-purpose cleansers are normally applied as dilute aqueous solutions containing ca. 10 – 15 g of liquid or 2 – 4 g of powdered cleansing agent per liter. The soiled object is wiped with a cloth that has been wrung out after being dipped in the cleansing solution. All-purpose cleansers can also be used in undiluted form to attack heavier dirt layers. The latter procedure requires a subsequent water rinse to prevent visible residues of the cleaner.

All-purpose cleansers are effective with most common kinds of dirt. Their dilute solutions are neutral or weakly alkaline, and are safe for use on most surfaces.

Trade names: Mr. Clean (Procter & Gamble), Ajax (Colgate Palmolive), Dor (Henkel), Magic Clean (KAO).

Surface compatibility is, however, somewhat restricted for the more strongly alkaline powdered cleansers, which are based on formulations of soda and surfactants. Many paints, for example, are sensitive to alkali. Alkaline powders are best suited to the removal of the large quantities of grease commonly encountered in meat markets or commercial kitchens. A typical composition for such a material is:

5 – 25 wt %	anionic surfactant
0 – 5 wt %	nonionic surfactant
10 – 25 wt %	trisodium orthophosphate [*7601-54-9*]
0 – 50 wt %	sodium carbonate [*497-19-8*]
0 – 0.5 wt %	colorant and perfume
	Balance sodium sulfate

Liquid all-purpose cleansers that contain water-soluble organic solvents more effectively remove large amounts of oily or fatty soil, fresh paint, or other colored material than cleaners without solvents. Appropriate additives include glycol ethers and higher molecular mass alcohols. The proper choice of amount and type of solvent can improve cleansing without causing a loss of surface protection.

Trade names: Der Generalreiniger (Henkel).

All-purpose cleansers can be supplemented with antimicrobial agents to confer microbistatic or microbicidal properties. Disinfecting all-purpose cleansers have long been in household use in the United States and in South America, although they have only recently begun to have an impact in Europe. Typical active disinfectant ingredients include quaternary ammonium compounds, phenol derivatives, terpene alcohols (pine oil), aldehydes, and aldehyde – amine condensation products [4].

Trade names: Lysol (Lehn & Fink), Pine Sol (American Cyanamid), Pinho Sol (American Cyanamid, Brazil).

Spray cleansers and foam cleansers are alternative forms of application for liquid all-purpose cleansing agents. The solutions are supplied in containers either equipped with pump-type sprayers or pressurized with a propellant gas. In the latter case, the product is dispensed on the surface as a foam. Some degree of retention on vertical surfaces results, thereby prolonging the time available for action. The major advantage of spray and foam cleansers is their convenience.

Trade name: Formula 409 (Clorox).

1.2. Scouring Cleansers

Scouring cleansers are used to remove particularly stubborn or burned-on dirt from surfaces resistant to the effects of mechanical action. Cleansing results largely from mechanical abrasion by finely divided mineral powder, although active ingredients such as surfactants, sequestering agents, or alkaline salts play a supportive role. Quartz and marble powder, less often diatomaceous earth (kieselgur), pumice, or corundum, serve as the basis of these products. To prevent scratching of surfaces, the mineral powders must be ground to a particle size of less than 0.15 mm. Particle size must be as uniform as possible and larger particles must be absent.

Trade names: Comet (Procter & Gamble), Ajax (Colgate Palmolive), ATA (Henkel), Supa Homing (KAO).

A distinction is made between bleaching and nonbleaching scouring powders. The bleaching agents most often used are organic compounds that decompose to liberate hypochlorite, thereby also conferring disinfectant properties. Examples include sodium dichloroisocyanurate [*2893-78-9*] and trichloroisocyanuric acid [*87-90-1*], as well as the hydrated double salt of sodium hypochlorite and trisodium orthophosphate ("chlorinated orthophosphate") [*56802-99-4*]. Scouring powders containing oxygen bleaches, such as magnesium perphthalate [*93094-22-5*], may also be prepared. These have more pleasant odors than the chlorine compounds.

Liquid scouring agents can be prepared with the same characteristics as powders and are more convenient to use. Since consumers tend to employ liquid scouring agents on more sensitive surfaces, most of the products currently marketed contain the softer abrasive marble dust. Products of this kind will not scratch enamel, tile, or melamine laminate (Formica) counter tops. Examples of typical scouring cleanser formulations are given in Table 2 (see next page).

Trade names: Soft Scrub (Clorox), Viss, Cif (Unilever), ATA (Henkel), Liquid Homing Tough (KAO).

Safety and Handling. Personnel working with powdered quartz must be protected from inhaling the dust by appropriate air filtration equipment because particles smaller than 0.5 μm can cause silicosis. Maximal exposure values (MAK or TLV values) to dust containing quartz have been established in a number of countries. These values are approximately 0.15 mg/m^3 [5]. Scouring agents containing quartz powder are filled by using completely isolated filling machines operating under a slight negative pressure.

Storage. Liquid scouring agents contain 25 – 50 % mineral powder with a relative density more than twice that of the carrier liquid. Proper choice of thickeners and stabilizers permits the preparation of metastable suspensions with an

Table 2. Partial compositions for three types of scouring cleansers

Component	Type of cleanser		
	Powdered scouring cleanser containing bleach *	Powdered scouring cleanser not containing bleach *	Liquid scouring cleanser **
Anionic surfactants, wt %	3.5	3	
Nonionic surfactants, wt %		1	3
Sodium dichloroisocyanurate, wt %	1		
Pentasodium triphosphate, wt %	5	5	
Acrylic acid – methacrylic acid copolymers [*9003-01-4*], wt %			5
Diethanolamine [*111-42-2*], wt %			1.5
Colorant/perfume	present	present	present
Marble powder, wt %			50

* The powdered cleansers use a quartz powder matrix.
** The liquid cleanser has a water matrix.

adequate shelf life for most purposes (about one year). The required high viscosity of the liquid phase is achieved either by a combination of anionic and nonionic surfactants [6] or with polymeric thickeners such as sodium polymethacrylate or acrylic acid – methacrylic acid copolymers [7].

Quality Control. Quality control of scouring agents primarily entails measuring particle size and particle-size distribution by sieving and measurement of settling times (elutriation).

1.3. Dishwashing Agents

1.3.1. Manual Dishwashing Detergents

Manual dishwashing detergents (MDDs) are everyday products used worldwide for cleaning used dishes, cups, glasses, and pots. The superior performance and acceptance of this product group is due to the liquid formulation of MDDs. Liquid dishwashing detergents are easy to use and offer very good fat and stain removal.

In Europe manual dishwashing detergents are available in classical and in concentrated formulations, for example, under the trade names Palmolive (Colgate Palmolive), Pril (Henkel KGaA), Fairy (Procter & Gamble), and various no-name products of supplier chains.

1.3.1.1. Classical Dishwashing Detergents

The largest product group on the market are the classical manual dishwashing detergents. These products are manufactured with special surfactant combinations and contain roughly 15 % active material. The main ingredients are surfactants, water, perfume, preservatives, and colorants. Formulation ranges for classical MDDs follow:

Anionic surfactant	8 – 12 %
Cosurfactant	1 – 4 %
Dyes, fragrance, preservatives, hydrotropic agents, salt	
Water to 100 %	

A mixture of two to four surfactants is used. The workhorse surfactant is linear alkylbenzenesulfonate (LAS), which offers very good fat removal and a good foam profile. This petrochemical surfactant has a natural counterpart: fatty alcohol ether sulfate (FES). For high-performance MDDs, FES is the preferred surfactant [38], [39]. The type mostly used in MDDs is based on coconut alcohol ($C_{12} - C_{14}$) with two moles of ethylene oxide (EO) (Table 3; see next page). Another widely used basic surfactant is SAS (secondary alkyl sulfonate), which offers good cleaning performance with reduced amounts of hydrotropes [40]. Hydrotropic additives such as urea, cumenesulfonate, or xylenesulfonate help to eliminate turbidity in solution. The performance of anionic surfactants is boosted by adding cosurfactants.

One of the most widely used tests of the performance of MDDs is the plate test. Plates are soiled with different kinds of fats/protein, aged for a defined time, and washed in an apparatus. Evaluators count the number of washed plates until the foam dies down [41].

Table 3. Anionic surfactants for MDDs

Surfactant (abbreviation)	Formula	Trade name	Producer
Linear alkyl sulfonate (LAS)	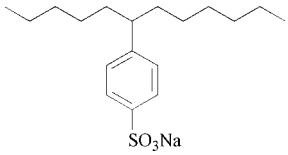	Marlon	Condea
Fatty alcohol ether sulfate (FES)	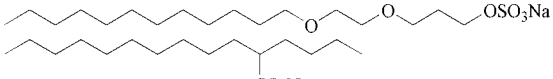	Texapon	Cognis
Secondary alkyl sulfonate (SAS)		Hostapur	Clariant

1.3.1.2. Dishwashing Concentrates

During the few last years consumer habits have changed. Today's consumer prefers concentrated products with a high content of surfactants and reduced amounts of water. Concentrates are packed in small bottles, typically 500 mL. These highly concentrated products have excellent cleaning performance. In some formulas up to five different surfactants are combined. Formulation ranges for MDD concentrates follow:

Anionic surfactant	20 – 35 %
Cosurfactant	2 – 10 %
Betaine	
APG/NMG	
Amine oxide	
Nonionic surfactant	0 – 4 %
Ethanol	2 – 7 %
Cumenesulfonate/hydrotropes (optional)	
Dye, fragrance, salts	
Water to 100 %	

The optimum cleaning performance is related to the combination of basic surfactant (anionic) and cosurfactants. Different cosurfactants with special performance profiles are used (Table 4).

Cosurfactants increase the foam profile. Consumers perceive a high-foaming product as having better fat removal. Since concentrated MDDs contain large amounts of surfactants, it is necessary to stabilize the formulation to avoid

Table 4. Cosurfactants for MDDs

Cosurfactant (abbreviation)	Formula	Trade name	Manufacturer
Coco amido propyl betaine (CAPB)	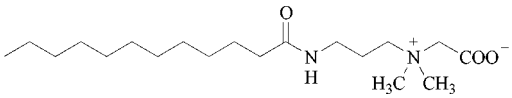	Tego Betain	Goldschmidt
Alkylpolyglucoside (APG)	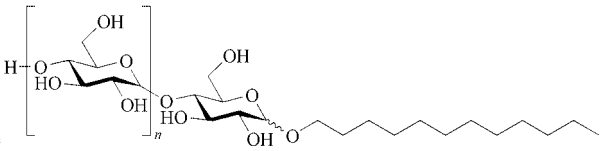	Glucopon	Henkel/Cognis
N-methylglucamide (NMG)	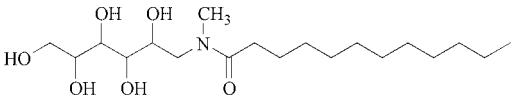	*N*-methylglucyamide	Clariant
Amine oxide (AO)	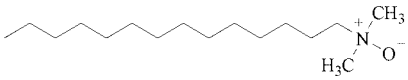	Rewominox	Rewo

phase separation. Ethanol or other solvents like glycerin and polyglycols are helpful but reduce the viscosity of the product. Consumers associate the viscosity with performance: high viscosity indicates high performance [42]. Cosurfactants do not reduce the viscosity as much as solvents and are surface-active. Another benefit of cosurfactants is mildness to the skin [43–45]. One of the preferred cosurfactants is coco amido propyl betaine (CAPB) [46–48], often in combination with APG.

1.3.1.3. Skin-Friendly Products

In general MDDs are neutrally formulated with a pH range of 5 – 8 [46]. There is only a minor tendency to irritate the skin, because MDDs are used in very low concentrations (e.g., 2 mL of product in 5000 mL of water). Consumers with sensitive skin or skin problems prefer to use mild products. To minimize skin attack softening ingredients like Lamesoft (protein hydrolysate), refatting agents like fatty acid amides, APG (alkyl polyglucosides), aloe vera, and/or panthenol are added to the basic formulation. Dyes and perfume are eliminated or reduced to a very small amount. Formulation ranges for skin-friendly MDDs follow:

Anionic surfactant	10 – 20 %
Betaine	5 – 10 %
Nonionic surfactant	0 – 20 %
APG	0 – 10 %
Ethanol	2 – 5 %
Aloe vera, panthenol, protein hydrolysate	
Small amounts of dye, fragrance	
Water to 100 %	

Some manufactures use large amounts of nonionic surfactants (FAEO with a particular alkyl chain length and a special degree of ethoxylation). Another way to provide a skin-friendly formulation is the use of APG. The alkyl polyglycoside containing product has the best skin compatibility of tested products (Fig. 1). A standard test design is the patch test. A dermatologist evaluates the degree of changes in the skin according to various criteria. A low irritation score (sum of erythema, edema, squamation, and fissures) indicates excellent skin compatibility.

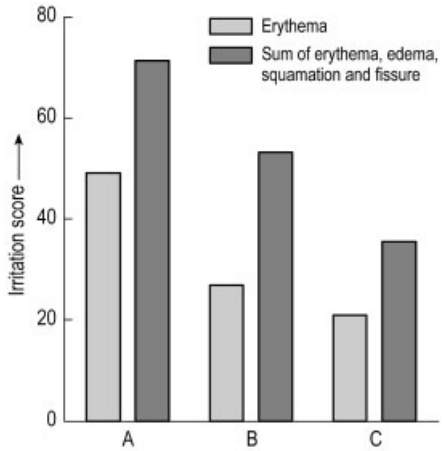

Figure 1. Results of patch tests on different skin-friendly MDDs
Test conditions: 20 healthy volunteers (male and female); concentraion: 1 % active substance; mode of application: once, occlusive, 24-h duration
A) Brand claiming good skin compatability; B) Market product for sensitive skin; C) Market product for sensitive skin with APG

1.3.1.4. Liquids with Antibacterial Action

This new group of 2-in-1 products is based on consumer habits in Europe. Today consumers want products with an excellent performance profile and additional benefits like easy dosing. Therefore, it is important to optimize the packaging and the formulation. Henkel introduced Pril 2-in-1 in late 1999. This product is a combination of an MDD and a handsoap with an antibacterial action. It is very convenient to use this product for both doing the dishes and washing the hands. The package design is optimized for this application by using a pump dispenser. Consumers are able to dose the product easy and exactly. The formulation includes an antibacterial component which is tested according to EN 1276, the European standard determing the biocidal activity of ingredients in a formulation against microorganisms. Common microorganism in household areas include *Staphylococcus aureus* (ATCC 6538), *Escherichia coli* (ATCC 10536), *Pseudomonas aeruginosa* (ATCC 15442), and *Salmonella typhimurium* (ATCC 13311). Optimized formulations offer a reduction in live microorganism numbers by five orders of magnitude after 5 min

at room temperature. Formulation ranges for 2-in-1 MDD & handsoap follow:

Anionic surfactant	20 – 40 %
Betaine	2 – 10 %
APG	2 – 10 %
Ethanol	4 – 10 %
Na benzoate, benzoic acid Dye, fragrance	1 – 4 %
Water to 100 %	

1.3.1.5. Dishwashing Habits

Different countries have different cultures, and consumers use manual dishwashing detergents in different ways. In northern European countries such as the Netherlands and Germany, consumers prefer to fill the sink in the kitchen with (hot) water and add the MDD. They use about 5 – 8 L of water and add ca. 2 mL of concentrated or 4 mL of classical products. In southern countries such as Italy and Spain consumers use MDDs on a sponge under running tap water, mainly cold water. During cleaning they apply concentrated product to the sponge or other cleaning devices.

1.3.2. Automatic Dishwashing Detergents

Automatic dishwashing detergents (ADDs) are one of the fastest growing segments of the market for home-care products. The performance of the dishwashers and detergents has reached a high technical standard. Especially cleaners have become a field of high dynamics and innovation. The success of ADDs reflects the desire of consumers for maximum convenience. The product system for automatic dishwashing consists of cleaners, rinse aids, and regeneration salt, reflecting the construction of the dishwasher and the different tasks of the dishwashing process.

1.3.2.1. Dishwashers

Since 1990 automatic dishwashers have become one of the most important appliances in modern households in Western Europe and the United States, with penetration rates up to 60 %. With an average distribution in the EU of 36 % in 1999 the growth potential is still high. On the international market, the most important manufacturers of automatic dishwashing machines are: Brandt Group, BSH, Candy, Electrolux, General Electric, Maytag, Merloni, Miele, and Whirlpool.

Three sizes of dishwashers are available in Europe: normal 60-cm-wide machines (50 cm in Switzerland) with capacities of 12 – 14 international place settings, narrow 45-cm machines for 7 or 8 settings, and table-top ma-

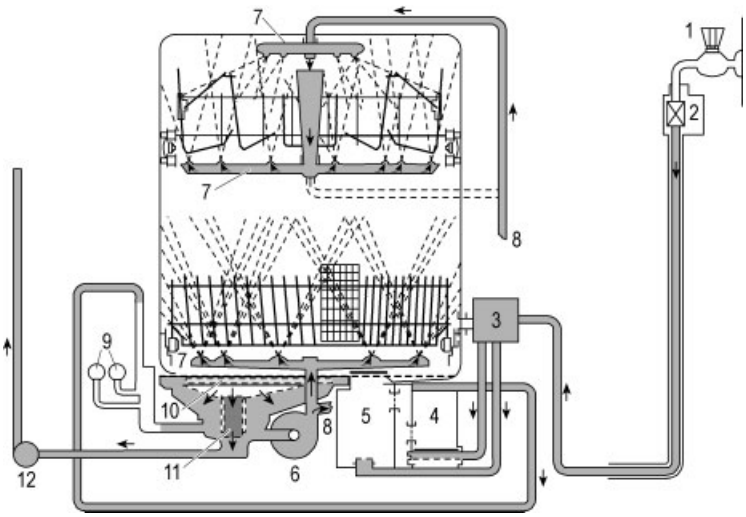

Figure 2. Functionality of a normal household automatic dishwasher
1) Water supply; 2) Magnetic inlet valve; 3) Inlet chamber; 4) Ion exchanger; 5) Salt storage; 6) Circulation pump; 7) Spraying arms; 8) Supply tube for upper spray level; 9) Water level control; 10) Heating element; 11) Filter combination; 12) Effluent pump

Table 5. Wash temperature and water and energy consumption in modern dishwasher programs

Program	Wash temp., °C	Rinse temp., °C	Water consumption, L	Energy consumption, kWh
Normal/universal	50 – 65	50 – 70	15 – 22	1.0 – 1.5
Strong	65 – 75	65 – 70	18 – 25	1.4 – 1.8
Special/energy-saving	35 – 55	50 – 60	10 – 18	0.6 – 1.3

chines for 4 or 5 place settings. Dishwasher suppliers normally offer different models with up to 10 washing programs. These programs can roughly be classified as normal/universal, intensive/heavy/pots & pans, and special energy-saving low-temperature programs such as eco/delicate/rapid/economy/glass, etc. Wash and rinse temperatures as well as water and energy consumption of the three main program types in modern type dishwashers are presented in Table 5 (see above).

The functionality of an automatic dishwasher is shown in Figure 2 (see previous page) [49].

A typical normal wash program consists of four water fills, as shown in Figure 3.

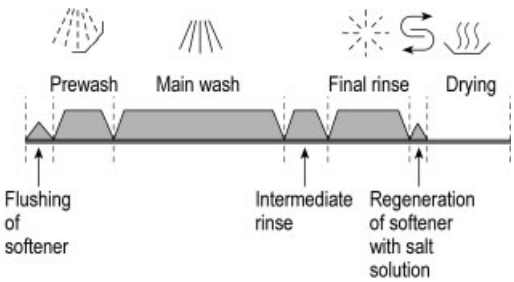

Figure 3. Normal wash program

After a cold prewash step, the detergent is automatically added at the beginning of the main wash by a special dispensing device, located in the front door of the machine. The dispenser also includes a rinse-aid unit with storage tank and injection system, which automatically adds the appropriate amount of rinse aid when a temperature of about 50 °C is reached in the rinsing step. According to the water quality used in the dishwasher (e.g., residual hardness), the volume of rinse aid can be adjusted by a mechanical switch and amounts to 1 – 6 mL.

In contrast to the United States, most European dishwashers (except in Scandinavia) are equipped with a softening device, in which the incoming water is softened by ion exchange. Depending on the type of dishwasher the ion exchanger is regenerated by addition of NaCl solution for about 20 min, either during the cleaning or final rinse step. After flushing the ion exchange resin with fresh water, the system is ready for a new wash program. Modern dishwashers are commonly equipped with signal lights on the front panel to indicate the the need for regeneration-salt and rinse-aid refill.

In principle, dishwashers can be operated with cold or hot water supply. In the USA, dishwashers are recommended for use with hot water (ca. 50 °C). Therefore, the electric heating capacity of about 1.0 – 1.5 kW is lower than that of European dishwashers (2.0 – 3.0 kW).

In modern dishwashers the traditional electric heating rods at the bottom of the interior are replaced by through-flow heating pipes. Consequently, the final drying process can no longer be supported by additional dry heating. For better drying efficiency, some dishwashers use instead an electric fan or a heat exchanger at the rear of the machine.

The developments in dishwasher technology in the late 1990s were driven by the need to save energy and water to meet European legislation. As with other household appliances automatic dishwashers have to carry an Energy Label [50], [51] which lists energy and water consumption together with cleaning and drying performance. To reach the goal of decreased energy and water consumption at sufficiently high cleaning performance [52], wash temperatures in the normal programs have been decreased from 65 to 55 °C, accompanied by the development of special technologies such as soil-sensor-controlled automatic programs, fuzzy logic control, half-load programs, variation between upper and lower spray level, pulsation spraying, and hardness-controlled regeneration systems and the introduction of new electronically controlled dishwashers.

1.3.2.2. Cleaners

Parallel to the increase in dishwasher penetration in households the European ADD cleaner market has doubled in value since ca.1990.

The main step in the development of cleaner formulations was the change from classical high-alkali cleaners to the new generation of the lower alkali cleaners in the early 1990s. The low-alkali system served as the chemical basis for the develoment of compact ADD cleaners [53].

The *classical system* was based on phosphate (preferably pentasodium triphosphate, $Na_5P_3O_{10}$) as complexing agent for the calcium and magnesium ions in the wash liquor to avoid scale formation. Metasilicate (disodium dihydrogensilicate, $Na_2H_2SiO_4$) was used as a highly alkaline component to soak and remove the dirt from the dishes [54].

Classical cleaners containing metasilicate (pH of wash > 12) must be labeled as corrosive according to the EC formulation guideline. In addition, usually an active chlorine carrier (e.g., trichloroisocyanuric acid) was used as bleach. Today, only 1 – 2 % of ADD cleaners in Europe still contain chlorine bleach.

The development of the *new cleaner generation* can be divided into two main steps. In the first step chlorine bleach was substituted by active oxygen bleach. Oxygen bleach in current ADD cleaners contains peroxo salts such as sodium peroxoborate or sodium peroxocarbonate. Usually bleach activators such as tetraacetylethylenediamine (TAED) are added to enhance the bleaching action. Metasilicate as alkaline carrier was replaced by sodium carbonate, in combination with sodium hydrogencarbonate or less alkaline silicates such as disilicate (tetrasodium silicate). In some products phosphate was substituted by citrate (preferably trisodium citrate dihydrate) and partially complemented by polycarboxylates (e.g., acrylic acid – maleic acid copolymer, sodium salt) and/or phosphonates (e.g., 1-hydroxyethylidene 1,1-diphosphonic acid) to more complex builder systems. In addition, enzymes (e.g., α-amylases and savinases) were incorporated in ADD formulations to increase the cleaning performance against starch- and protein-containing soils under at lower pH (wash liquor pH 9 – 11) [55].

In the second step, since the mid-1990s, phosphate was re-introduced as the main builder component, replacing citrate. Due to its high complexing power phosphate is not only very efficient in scale inhibition but also helps in removing stains that contain calcium and magnesium (e.g., tea stains). Furthermore, phosphate helps to disperse soil particles in the liquor. From the ecological point of view the use of phosphate for ADD seems acceptable at present, since the relative contribution to the overall environmental phosphate input is quite low [56]. However, there are some regulations for maximum phosphorus content of ADD formulas or with respect to the dosage, e.g., in Italy, Switzerland, and the USA.

In the field of active oxygen bleach specific metal complexes (e.g., pentaammineacetatocobalt(III) nitrate and 1,4,7-trimethyl-1,4,7-triazacyclononanmanganese(IV) [Mn_3(TACN)] were identified that can equal or even exceed the action of conventional bleach activators (TAED) especially at lower temperatures of 40 – 55 °C. Today metal catalysts are used in some compact ADD cleaners. In the field of enzymes, changing amino acid sequences by genetic engineering led to amylases and proteases having better compatibility with bleach components and improved performance profiles.

In addition, the compact cleaners usually contain low-foaming nonionic surfactants such as addition products of ethylene oxide with fatty alcohols and silver-protection agents such as 1*H*-1,2,3-benzotriazole [57]. Due to their lower alkalinity and depending on the amount of irritant components (e.g., disilicates) the compact cleaners are either not labeled at all or labeled as irritant.

Table 6 (see next page) lists examples of ADD formulations.

ADD cleaner products can be divided into three basic types: liquids, powder, and tablets.

Liquids were introduced in Germany as early as 1987 but no longer play any role. In some European countries, e.g., France and Great Britain, they still have some importance. Overall they have lost market relevance year by year. Liquids suffer mainly from their low content of active ingredients and chemical and/or rheological instability. These negative aspects outweigh positive aspects like good solubility. Liq-

Table 6. Percentage ADD cleaner formulations (examples)

Ingredient	Classic	Liquid	Compact Step 1		Step 2
	High-alkali	High-alkali	Low-alkali	No label	Low-alkali
Phosphate	30	20			50
Citrate			30	45	
Polycarboxylate			8		
Metasilicate	64				
Soda	5	10	35	10	20
Hydrogencarbonate				23	
Silicates		10	10		5
Cl compounds	1	1			
Perborate			7	10	10
TAED			2	4	2
Amylase, protease			2	3 – 4	3 – 6
Surfactant		0.5	1	3	2
Silver care	–	–	+	+	+
Fragrance	–	+	–	+	+
Water (balance)	–	55	+	+	+

uid cleaners are mostly based on classical formulations. Up to now all efforts to establish powerful and successful liquid or so-called gel products of the new-generation formula with enzymes and active oxygen bleach on the market failed due to compatibility problems.

Powders were the first ADD products in Europe (1960) and clearly dominated the market until the mid-1990s. Then, in parallel to the rise of the tablets, they constantly lost market shares. However, in contrast to the liquids powders still are of some importance in the market (< 30 % share of market value in 2000). Today the existing powder products in the European market reflect all stages of development. Although the relative importance of the classical products has been decreasing over the years, compact powders gained only limited importance in Europe, except in Germany, Austria, and the Netherlands. As chemical basis for the noncompact powder segment, metasilicate formulations were mostly replaced by "diluted" low-alkali formulations ("traditional powders"). They usually have low contents of oxygen bleach and enzymes and contain certain amounts of, e.g., sodium sulfate as filler. Often part of the components of modern powder products are colored with bleachable dyes to obtain speckles.

Recently, a so-called 2-in-1 powder variety appeared on the market. Compared to the standard powder type it contains a high amount of rinse-aid surfactant, thereby combining cleaning and rinsing properties in a single product.

Powders are always dosed in the dosage chamber of the dishwasher to act in the main-wash cycle. The recommended amount of powder detergent in the wash varies from 20 to 30 g for compact powders and from 40 g to complete filling of the dosage chamber for noncompact powders, depending on soil load and water hardness.

ADD powders are usually based on powdered and granular components and are made by mixing or by special agglomeration processes [58]. Today powders are mainly packed in flexible plastic bags with gussets in composite packaging material. Refill bags are able to stand upright for transport and use. Plastic bottles and folding cardboard boxes are also used.

Tablets were first introduced with a classical formulation (1990, Germany) and since then constantly gained importance. In Europe they are now the dominant ADD type, approaching a total market share of ca. 70 % in 2001. Modern ADD tablets are mainly based on low-alkali compact formulas with phosphate or, to a lesser extent, citrate.

In contrast to powders tablets can in principle be dosed in the prewash or in the main-wash cycle. In the cold prewash the tablet is partially dissolved to start soaking and partially removing soil from the dishes. For example, classical tablets of 30 g weight releasing highly alkaline metasilicate used to work like this.

Tablets with compact formulas are nowadays loaded into the dosage chamber where they re-

lease all ingredients in the main-wash. The currently available enzymes and the active oxygen bleach require certain temperatures for optimal action. Due to the dosage chamber dimensions compact tablets have maximum weights of 25 g.

Tablets are still are the most innovative field in ADD development. The increasing consumer acceptance due to convenience was further promoted by introduction of novel two-phase tablets in the mid-1990s. Today there are various two- and, since 1999, even three-phase tablets with different structures on the EU market. The structures range from parallel layers to corelike.

Phase separation allows possible negative interactions of potentially less compatible components (e.g., bleach and enzymes) to be avoided, thus leading to improved storage stability. Furthermore, phase separation in combination with different dissolution profiles strongly increases the possibility of time-delayed release of specific ingredients relative to one-phase tablets. For example, two-phase tablets are on the market that provide for controlled and separate release of enzymes and bleach, respectively, during the main-wash cycle.

The current state of the art are multiphase 2-in-1 and even 3-in-1 tablets that combine several ADD system components in a single product. The 2-in-1 tablets were introduced in 1999 and combine cleaning and rinse-aid properties. Compared to standard tablets they contain a significantly increased amount of rinse-aid surfactant (up to ca. 10 %), that is distributed in the tablet phases in different ways. Depending on their specific adsorption characteristics the surfactants are partially transferred from the main-wash to the intermediate and the rinse cycles, respectively. In some cases the 2-in-1 tablets contain special rinse-aid phases (e.g., cores) in which the surfactants are processed in a solid, but water-soluble or dispersible matrix. In addition the matrix and/or further core coatings may exhibit controlled-release effects that exploit the change in various physical or chemical parameters on the transition from the main-wash to the subsequent cycles.

The 3-in-1 tablets were introduced in 2001 and combine cleaning and rinse-aid properties and, in addition, contain so-called regeneration-salt substitutes (see Section 1.3.2.4). They exhibit all features of 2-in-1 tablets, as described above, but contain special builder/cobuilder ingredients to cope with tap water hardness. With 3-in-1 tablets the detergent manufacturers claim that there is no need for water softening by the ion exchanger up to certain limiting water hardnesses, (e.g., 21 °d = 3.8 mmol CaO/L). The binding and/or dispersal of the water hardness occurs by means of phosphate and/or polymers, thereby avoiding residues on glass, china, and the dishwasher interior [59].

ADW tabs are manufacured by using a tableting process. A main prerequisite for this technology is the preparation of a free-flowing powder premix. A suitable process unit is, e.g., a batch or continuous granulation step in which powdered or liquid ingredients are mixed to form granules. Afterwards the premix is conditioned by drying, cooling, and sieving and is then finished by postaddition of further specific raw materials or compounds (e.g., enzymes, TAED). The powder premix is fed into a standard tablet press to form the required tablets. Special multistage presses are equipped with two or more dosing and pressing devices to produce multilayer tablets [60–62].

Usually the tablets are single-wrapped in a plastic film of orientated polypropylene (OPP) as a protection against moisture and mechanical stress. They are loose packed in reclosable folded cardboard boxes.

Trade names: Somat (Henkel), Calgonit (Benckiser), Fairy (Procter & Gamble), Sun (Unilever).

1.3.2.3. Rinse Aids

Rinse aids have been used from the beginning of automatic dishwashing in private households. Rinse aids are designed to ensure rapid and film-like draining of the wash solution by optimized wetting in the rinse cycle. This supports spotless drying of the relevant surfaces during water evaporation in the heating and drying cycle at the end of the dishwashing process. The use of rinse aid is especially important if the water contains considerable amounts of dissolved salts even after softening by the ion exchanger.

Foaming must be minimized, as for the cleaner, since the rinsing solution must be vigorously circulated. To an extent that strongly depends on the overall conditions, the rinse

aid must cope with hardness, soil, and alkaline cleaner residues from main-wash liquor tranfer.

Rinse-aid formulations have not changed significantly with time. To fulfill the above-mentioned requirements, they contain typically 10 – 20 % of nonionic surfactants, e.g., addition products of ethylene oxide and propylene oxide with fatty alcohols, 3 – 10 % organic acid (in one case up to 35 % citric acid), 3 – 10 % solubilizers (e.g., cumenesulfonate), perfumes, and water.

The recommended dosage is 2 – 6 mL, depending on rinsing water volume and hardness. Conventionally, the rinse aid is dosed via an automatical dispenser in the dishwasher during the rinse cycle. Since 1999 also so-called 2-in-1 tablets (see Section 1.3.2.2) are available. Detergent manufacturers claim that with these tablets a separate (e.g., liquid) rinse aid is no longer neccessary.

1.3.2.4. Regeneration Salt

In contrast to laundering, automatic dishwashing only delivers acceptable cleaning results when soft water is used. Therefore, dishwashers are equipped with softeners based on ion exchange. The ion exchangers, which are based on conventional ion-exchange resins, are regenerated either in every wash program or, in new electronically controlled models that monitor the hardness of the tap water, when the exchange capacity is almost expired. For the regeneration process pure sodium chloride solution is used. Regeneration salts are special grades with low contents of insoluble residues, free from additives such as anticaking agents. The granulometry must also be controlled with little or no dust particles smaller than 0.3 mm to avoid sedimentation or clogging of the regeneration system.

Since 2001 so-called 3-in-1 tablets containing regeneration salt substitutes are also available.

1.3.2.5. Other Products

Due to the attractivity of the ADD market various additives with additional consumer benefit have become established. Whereas pretreatment sprays based on alkaline or amylase solutions did not reach the market, machine-care and machine-deodorizing products gained importance in the few last years.

Machine-care formulations contain mainly 3 – 10 % nonionic surfactants to wet and loosen the soil residues, 10 – 30 % organic acids to dissolve scale, and 5 – 15 % alkohols such as glycerine as care component. Machine-care products can be applied by hand to rubber seals to keep them pliable and dosed automatically in the cleaning cycle by special devices.

Deodorants are designed to overcome unpleasant odors derived from the soil load after when the dishwasher is opened. They consist of hanger devices for the rack and contain perfumes, which are released through appropriate membranes during the washing program.

1.3.2.6. Performance Assessment

Cleaning Performance. The usual methods for cleaning performance tests are EN 50242 [63] and IEC 436 [64]. However, both methods are designed to measure the performance of electric dishwashers. They were developed at a time when the new compact detergents containing enzymes had not yet become established on the market. Consequently, these methods are often incapable of adequately reflecting the complete wash profile of such detergents. It therefore became necessary to develop a new system for a more comprehensive assessment of the wash performance of dishwasher detergents.

In the mid-1990s, under the auspices of the German Industrial Association of the Manufacturers of Toiletries and Detergents [*Industrieverband Körperpflege- und Waschmittel* (IKW)], a new method was created to select forms of soils from existing methods or develop new soils if necessary for determining the wash profile of the detergent [65].

The types of soil can be roughly classified as:

- Colored, bleachable (tea, ketchup, lipstick, etc.)
- Persistant, burnt-in (burnt meat, burnt milk, burnt casseroles, etc.)
- Dried starch-containing (porridge, starchy sauces, pasta, potatoes, etc.)
- Dried proteinaceous (egg yolk, meat, etc.)

The laboratory soil is specially designed to be difficult to remove in order to distinguish between similar formulations.

The cleaning performance is assessed for the different types of soil on a scale ranging from 0 to 100 %. The evaluation of the cleaning performance is either based on visual inspection of washed and unwashed articles or gravimetrically. For the removal of tea stains a hypochlorite-containing classical detergent and a high-phosphate compact detergent display performance advantages compared with a phosphate-free system with conventional bleach activator. Owing to its higher alkalinity the classical formulation would also remove stubborn and burnt-in soil better than the phosphate-free compact product. However, due to the application of enzymes, the newer generations of phosphate-containing and phosphate-free compact detergents exhibit clear performance advantages on starch and proteinacious soils. This indicates that there are differences between the cleaning performance aspects of the classical cleaner and the two types of low-alkali compact detergents.

Material Care. The European draft standard EN 12875 [66], [67] specifies a dynamic, multicycle test method for the determination of the mechanical dishwashing resistance of domestic articles. This method is also recommended to differentiating between different dishwashing detergents.

With regard to their corrosion-protection effect compact detergents with a high disilicate content are advantageous for porcelain on-glaze decoration. However, thanks to an improved burning technique and the use of in-glaze and/or sub-glaze decorations, sensitive on-glaze decorations are today hardly found in the market. On the other hand, after a few dishwashing programmes cleaners with a high disilicate content cause corrosion phenomena such as iridescence and clouding on glassware (e.g. soda lime glass, potash crystal, and lead crystal glass). Iridescence is caused by the formation of a thin silica layer with different refractive index causing interference patterns which result in a colored appearance.

Permanent clouding, mostly symmetric and ring-shaped, is a further phenomenon that arises from corrosion of glass in automatic dishwashing. The reason for this is the leaching of alkali metal ions due to an uneven temperature distribution in the glass surface during glass production [68]. This occurs, for example, during the smelting process of the top edges or the attachment of stems. Finally, this leads to the conclusion that corrosion protection of delicate on-glaze decorated porcelain can only achieved at the expense of increased glass corrosion.

Additional, glass corrosion is mainly effected by high temperatures ($> 65\,^{\circ}C$) and very low water hardness (< 18 ppm $CaCO_3$) in the dishwashing machine and the composition and manufacturing process of the glass itself.

The phenomenon of silver tarnishing has increased with the introduction of low-alkali compact detergent. The formulation parameters (lower pH, activated oxygen bleach) results in a more sensitive reaction of the silver surface. Additional parameters like higher water hardness, chloride ion concentration, soil, and temperature influence the corrosion behavior as well [69]. Silver tarnishing is prevented by adding agents such as benzotriazole that form an adsorptive layer.

Scale Inhibition. Sufficient complexation of residual water hardness during dishwashing will prevent the formation of insoluble precipitates (e.g., carbonates) on washed dishes and machine parts. Depending on the type of detergent, this is achieved by complexing agents like citrate [55], citrate – phosphate or phosphate, sometimes in combination with phosphonates and cobuilders like polycarboxylates [70].

Evaluation methods [71], [72] are mainly based on visual inspection of the tableware, especially glasses and cutlery, with different rating systems, sometimes also with addition of soil in multicycle tests.

Rinse and Shine Effect. Specific test methods for the evalution of rinse or shine effects depending on various parameters during the dishwash program are described in [73], [74]. As in the scale-inhibition evaluation they are based on visual inspection of tableware, especially glasses and cutlery, on effects like spotting using rating systems, also with addition of soil.

1.4. Floor Cleaners

The all-purpose cleansers described in Section 1.1 are suitable for daily floor cleaning use.

However, strippers are better for cleaning badly soiled floors and removing built-up layers of self-polishing floor wax. These are ammoniacal surfactant preparations containing glycol ethers (e.g., 2-butoxyethanol [*111-76-2*]).

A typical stripper contains 2 – 10 wt % surfactant, 2 – 8 wt % phosphate, 1 – 3 wt % ammonia, 5 – 20 wt % glycol ether, and a perfume in aqueous solution.

Solvent cleaners (mineral spirits) are needed to clean and remove old wax from floors made of moisture-sensitive substrates, such as unsealed wood or parquet. Combination, single-step floor finishes are convenient for routine maintenance of floors. These are products that simultaneously clean and wax. Floor cleaners stress cleaning power, whereas floor polishes are primarily floor finishers (waxes). The cleaners are designed to be added in small quantity to ordinary wash water. They consist of a combination of soap and other detergents, including alkanesulfonates or fatty acid alkanolamides, and they may also contain dispersions of wax or waxlike substances (e.g., montanic acid and montanic acid esters).

A typical floor cleaner contains 10 – 30 wt % potassium soap, 2 – 10 wt % other surfactants, 0 – 6 wt % waxes, and a perfume in aqueous solution.

Floor polishes are self-polishing wax emulsions with a high content of polymeric substances and a smaller amount of surfactants. Products of this type are spread on the floor in undiluted form to produce a thin layer. When dried, they leave a protective film with a high polish. Treated floors can be maintained by regular washing with water that contains a small amount of the same product.

A typical floor polish contains 20 – 40 wt % polymer dispersion (40 %) (e.g., polyacrylate or polystyrene), 2 – 5 wt % surfactant (e.g., fatty alcohol or alkylphenol polyglycol ether), 2 – 6 wt % waxes (e.g., montanic acid esters, polyethylene wax), 1 – 3 wt % plasticizers (e.g., tributoxyethyl phosphate or dimethyl phthalate), 2 – 4 wt % solubilizers (e.g., glycol ethers), a perfume, and some preservatives in aqueous solution.

Quality control tests and standards for floor finishes, floor cleaners, and strippers are, e.g., solids content, solvent content [8–11].

1.5. Carpet Cleaners

Only loose carpets and rugs can be washed or dry cleaned by the techniques employed with other textiles. Rugs can be washed by using gentle, low-sudsing detergents or dry cleaned with organic solvent mixtures. These processes are carried out in very large, dedicated machines. Various techniques are used to clean permanently installed carpets [12], [13]. Carpet sweepers and vacuum cleaners are commonly used to remove loose dust and dirt, but loose rugs are better cleaned by brushing devices. The removal of stubborn, oily dirt requires special cleansing agents. Adhering soil must be loosened from the fibers and transformed into a state that allows its removal by brushing or vacuuming.

Wet processes and dry processes can be distinguished, depending on whether the dirt that has become associated with the cleaning agent is removed immediately along with the cleansing solution (wet process) or after the carpet has been allowed to dry (dry process) [14]. The dry processes are further subdivided into *shampooing* and *powder cleaning,* whereas the usual wet process is known as *spray extraction cleaning* or *steam cleaning.*

Shampooing is accomplished by applying a dilute aqueous solution of carpet cleaner directly to the surface to be cleaned [15]. Typically, a special device containing rotating brushes, which ensure production of a heavy foam and simultaneously work the foam into the carpet, is used. Excess foam is then removed, and after the carpet dries (which may require several days), the dirt-laden cleanser residue is removed by vacuuming. Given its mode of action, a carpet shampoo must be capable of producing a stable foam and leaving behind a dry solid residue that is easily collected. Only a very few surfactants generate sufficient dry residue. Appropriate formulations contain anionic surfactants, such as sodium (C_{15-18})olefinsulfonate [*91082-14-3*], lithium (C_{12-14})alkylsulfate [*90583-20-3*], or sodium (C_{12-16})alkylsulfate [*73296-89-6*], along with such polymers as styrene maleinate resin, polystyrene [*9003-53-6*], polyacrylic resin [*9003-01-4*], or poly(methyl methacrylate) [*9011-14-7*]. The partial composition of a typical carpet shampoo is given, along with the

Table 7. Partial compositions of a carpet shampoo, a spray extraction cleaner, and a cleaning powder

Component	Carpet shampoo ∗	Spray extraction cleaner ∗	Cleaning powder
Surfactant, wt %	5 – 20	2 – 30	1 – 10
Polymers, wt %	2 – 8		
Sequestering agent (e.g., sodium triphosphate, EDTA, NTA), wt %	0 – 2	2 – 30	
Solvent (isopropanol, hydrocarbon), wt %	0 – 20		0 – 25
Carrier material, wt %			15 – 60
Perfume	present	absent	present
Preservative	present	present	present

∗ The two liquid cleaners are aqueous solutions

composition of a spray extraction cleaner and a cleaning powder, in Table 7 (see above).

In the *spray extraction method,* a highly dilute cleansing solution is applied on and into the surface layers of the carpet with special equipment that then removes the wash solution [16]. The carpet fibers are effectively flooded with cleansing agent, and dirt is immediately flushed out of the depths of the pile as a part of the cleaning process. The cleansers contain mixtures of low-sudsing surfactants and sequestering agents in aqueous solution (see Table 7).

Powdered cleansers are especially suitable for home carpet cleaning because their use does not involve any special equipment. *Cleansing powder* is spread over the carpet, brushed in, allowed to collect the dirt, and then removed with an ordinary vacuum cleaner. The powders themselves are easily dispensed preparations of a carrier material, which has been presaturated with an aqueous surfactant solution, occasionally supplemented with alcohol or hydrocarbons. Suitable carriers include sawdust and finely granular preparations of either polyurethane foam or urea – formaldehyde resins (see Table 7).

1.6. Toilet Bowl Cleaners

Toilet bowl cleaners must remove calcium and rust deposits from porcelain surfaces; thus, acidic preparations are effective agents [17]. The most commonly used products are powders that contain sodium hydrogen sulfate [*7664-38-2*] as the major component. They may also contain sodium carbonate [*497-19-8*] or sodium hydrogen carbonate [*144-55-2*], along with small amounts of surfactants such as sodium olefinsulfonate [*91082-14-3*]. The carbon dioxide that is liberated when these products are sprinkled into water serves to agitate the solution and enhance the cleansing action. The partial composition of a typical powdered toilet bowl cleanser, along with the compositions of a liquid cleanser and an alkaline cleanser, is shown in Table 8 (see below).

Liquid bowl cleaners usually contain hydrochloric [*7647-01-0*], phosphoric [*7664-38-2*], or formic acid [*64-18-6*], along with surfactants (see Table 8). The use of a 1 : 4 mixture of sodium alkylbenzenesulfonate and an alkylphenol polyglycol ether containing

Table 8. Partial compositions of several types of toilet bowl cleansers

Component	Powdered cleanser	Liquid cleanser∗	Alkaline cleanser∗
Sodium hydrogen sulfate, wt %	58 – 100		
Phosphoric, hydrochloric, formic acid, wt %		5 – 50	
Sodium carbonate, sodium hydrogen carbonate, wt %	0 – 40		
Chlorine bleach (13 % active Cl), wt %			25 – 75
Surfactant, wt %	0 – 2	1 – 10	0 – 5
Sodium hydroxide, wt %			0 – 5
Colorant	present	absent	absent
Perfume	present	present	present

∗ The two liquid cleaners are aqueous solutions

ten moles of ethylene oxide has been found to be particularly advantageous.

Alkaline products containing bleach (hypochlorite) are often used for cleaning toilet bowls even though they are not able to remove calcium deposits (see Table 8). Such materials must not be used in combination with acidic products because the evolution of deadly chlorine gas would result.

1.7. Toilet Hygiene and Deodorant Products

Toilet hygiene and deodorant products include disinfectant cakes, tank disinfectant dispensers, and toilet bowl deodorants.

The *deodorant and bowl cakes* are solid surfactant preparations with a high perfume content. These are normally supplied in a plastic holder that can be suspended from the edge of a toilet bowl. Each time the toilet is flushed, the incoming water dissolves a small amount of the cake material, which both cleanses and deodorizes. Deodorant properties are also sometimes conferred by cakes containing *p*-dichlorobenzene [*106-46-7*] (with or without surfactant), although the use of this compound has raised serious ecological questions. The effectiveness of *p*-dichlorobenzene stems from its inherently strong odor and its high vapor pressure. Table 9 (see below) gives the compositions of typical toilet hygiene and deodorant products.

Automatic tank dispensers are highly colored surfactant blocks that are packaged in plastic containers or in a water-soluble film. They are placed directly in the toilet tank to lower the surface tension and color the water used for flushing. The surfactants employed are highly foaming anionic agents, preferably sodium alkylbenzenesulfonates, and structure-building nonionic compounds, such as highly ethoxylated (> 25 ethylene oxide units) fatty alcohol [*68439-49-6*] or alkylphenol polyglycol ethers [*9016-45-9*]. The solubility of the product is adjusted by adding waxy substances such as stearic acid [*57-11-4*], cetyl alcohol [*36653-82-4*], montanic acid [*68476-03-9*], or montanic acid esters [*73138-46-2*] (see Table 9).

1.8. Drain Cleaners

Drain cleaners are strongly alkaline products designed to unclog the drains and traps of sinks, bathtubs, etc., when these become plugged with hair, grease, food residues, soap deposits, or the gelatinous masses that can result from the action of microorganisms [18].

Powdered drain cleaners are effective partly because their dissolution in water generates heat. A temperature of 90 °C can be achieved by using ca. 25 g/100 mL. Many drain cleaners also contain finely granulated aluminum, which reacts with the hot aqueous alkali to evolve hydrogen gas. Sodium nitrate converts this hydrogen to ammonia. This gas evolution provides mechanical assistance for the removal of residues and deposits.

Liquid drain cleaners are used largely to prevent clogging. The inclusion of substances that release chlorine ensures the decomposition of hair. Compositions of drain cleaners are shown in Table 10.

Drain cleaners in aerosol cans – so-called *pressure cleaners* – are an example of propellant packaging. In the course of their use, a burst of

Table 9. Partial compositions of toilet hygiene and deodorant products

Component	Toilet-bowl deodorant cake	Toilet-bowl *p*-dichlorobenzene cake	Toilet-tank agent
Surfactant, wt %	10 – 94	0 – 30	15 – 94
Polyglycol ether, wt %	5 – 30	0 – 20	5 – 30
Waxy substances, wt %	0 – 5		0 – 5
Structural material (e.g., sodium sulfate, phosphates), wt %	0 – 40	0 – 30	0 – 40
Perfume, wt %	1 – 15	1 – 10	0 – 5
Preservative	present	absent	present
Colorant, wt % (if applicable)	present	present	1 – 5
p-Dichlorobenzene, wt %	absent	10 – 99	absent

Table 11. Partial compositions of some metal cleaners

Component	Stainless steel, nickel, and chromium cleaner	Nonferrous metal cleaner	Silver polish
Abrasive (e.g., Neuburger siliceous chalk, diatomaceous earth), wt %	10 – 30	10 – 30	
Acid, wt %	1 – 5		1 – 5
Ammonium soaps, wt %		5 – 10	
Nonionic surfactants, wt %	2 – 10	0 – 5	1 – 5
Chelating agents, wt %		0 – 5	
Solvent (e.g., mineral spirits), wt %	20 – 40		
Thiourea, wt %			5 – 10
Perfume	present	present or absent	absent

∗ All three types of cleaners are prepared in aqueous solution

pressure is developed that dislodges the blocking agent.

Table 10. Partial compositions of drain cleaners

Component	Powdered cleaner	Liquid cleaner∗
Sodium hydroxide, wt %	25 – 100	5 – 25
Sodium nitrate, wt %	0 – 40	
Sodium chloride, wt %	0 – 30	
Aluminum, wt %	0 – 5	
Surfactant, wt %		0 – 5
Chlorine bleach, wt %		0 – 60

∗ The liquid cleaner is an aqueous solution

1.9. Metal Cleaners

Special metal polishes are used to clean such metal surfaces as stainless steel sinks, faucets, metal trim, silverware, etc. These products contain fine abrasives with a particle size $< 20\,\mu m$. Weakly acidic materials containing lactic [*50-21-5*], citric [*77-92-9*], or phosphoric acid [*7664-38-2*] are preferred for cleaning *stainless steel, nickel,* and *chromium.* Fatty alcohol [*68439-49-6*] or alkylphenol polyglycol ethers [*9016-45-9*] with 7 – 12 ethylene oxide units are added as emulsifiers and dispersing agents.

Polishes for the *nonferrous metals* are weakly ammoniacal. The usual dispersing agents are the ammonium salts of oleic [*544-60-5*] or stearic acid [*1002-89-7*]. Chelating agents are often added to enhance cleaning. These include ammonium citrate [*3458-72-8*] and ammonium oxalate [*1113-38-8*]. *Silver* objects can be freed of sulfide tarnish by dipping them in products containing thiourea [*62-56-6*] along with either hydrochloric [*7647-01-0*] or sulfuric acid [*7664-93-9*]. Table 11 shows the compositions of several types of metal cleaners.

1.10. Window Cleaners

Window cleaners must leave no residue on glass surfaces. Such products are composed almost exclusively of liquid with small amounts of necessary surfactants. Two types of window cleaners are available. One comes packaged in a plastic bottle equipped with a sprayer or spray pump and is ready to use. The other, known as window washing concentrate, is a mixture that must be diluted with water before use [19]. The compositions of the two types of window cleaners are shown in Table 12.

Table 12. Partial compositions of window cleaners∗

Component	Ready-to-use window cleaner	Window washing concentrate
Surfactants (e.g., olefinsulfonates, fatty alcohol ether sulfates), wt %	0.1 – 0.2	2 – 6
Ethanol, isopropyl alcohol, wt %	2 – 20	0 – 10
Glycol ether, wt %	0 – 15	
Ammonium hydroxide, wt %	0 – 1	0 – 1
Colorant	present or absent	present
Perfume	present or absent	present

∗ These mixtures are prepared in aqueous solution

1.11. Automotive Cleansers

Automotive cleansers must not damage auto paint. They must be freely soluble in and compatible with hard water, be roughly neutral (i.e., pH), and possess effective wetting and cleaning powers. The active detergent in most liquid

car cleaners is comprised of alkylbenzenesulfonates, alkylsulfonates, or fatty alcohol sulfates together with fatty alcohol or alkylphenol polyglycol ethers [20].

Commercial car washes require low-sudsing products with a high content of polyphosphates. These normally leave behind trace residues of hydrophilic cleaning agents. Consequently, washing should always be followed by treatment with a protective agent to return the paint to its normal hydrophobic state. This ensures that water on the surface will gather into droplets and run off. Actual drying is necessary only on horizontal surfaces, and this is accomplished in automatic car washes with the aid of an air stream. The protective agents commonly employed are cationic preparations containing dimethyldistearylammonium chloride [*61789-80-8*] [21].

The compositions of typical automobile-finish cleaners and automobile-finish protective agents are given in Table 13.

Table 13. Partial compositions of automobile-finish cleaners and protective agents

Component	Automobile-finish cleaner ∗	Automobile-finish protective agent ∗
Anionic surfactant, wt %	10 – 20	
Nonionic surfactant, wt %	5 – 10	4 – 8
Cationic surfactant, wt %		5 – 10
Solvating agent (e.g., isopropyl alcohol, urea), wt %	2 – 5	
Polyphosphate, wt %	0 – 10	
White oils, wt %		10 – 20
Preservative	present	absent

∗ These preparations are in aqueous solution

In general, *car windshields* are washed with common window cleaners. Especially stubborn dirt such as silicone films is removed by using a suspension of a mild abrasive such as siliceous chalk or finely dispersed silicic acid in an alcoholic solution containing a wetting agent. To clean the windshield effectively while a vehicle is in motion, wetting agents compatible with hard water are added to the windshield washer reservoir. Appropriate substances include alkyl polyglycol ether sulfates or fatty alcohol polyglycol ethers. For winter use, addition of an antifreeze such as alcohol or some other freezing point depressant is necessary [22–24]. The compositions of windshield washer additives for summer and winter use are shown in Table 14.

Table 14. Partial compositions of windshield washer additives for summer and winter use

Component	Additive for summer use∗	Additive for winter use
Surfactant, wt %	15 – 20	1 – 2
Sequestering agent, wt %	1 – 5	
Ethanol, wt %		78 – 94
Ethylene glycol, wt %		5 – 20

∗ The preparation is used in aqueous solution

Tar removers, surfactant-containing solvent cleaners, are used to remove spots of tar or asphalt. Suitable surfactants include alkylphenol polyglycol ethers having 8 or 9 ethylene oxide units. The product is sprayed or painted on the spot, and then removed with water [25].

A typical tar remover contains 60 – 80 wt % mineral spirits, 15 – 25 wt % perchloroethylene or citrus terpenes, and 5 – 15 wt % surfactants.

1.12. Building Facade Cleaners

Because of the number of materials employed in the facades of buildings and the variability of their resistance to chemicals, any cleansers to be used on them must be thoroughly tested for suitability [26]. Facades made of *acid-resistant masonry* (including brick) are usually treated with strongly acidic cleansers, preferably those containing acids such as hydrofluoric [*7664-39-3*] or phosphoric acid [*7664-38-2*] that form sparingly soluble calcium salts. A typical such cleanser contains 10 – 20 wt % hydrofluoric acid, 10 – 20 wt % phosphoric acid, and 5 – 15 wt % surfactants in aqueous solution.

Acid-sensitive materials, such as concrete or plaster, require the use of strongly alkaline media. Alkaline treatment must be followed by neutralization with dilute acid to prevent possible subsequent efflorescence. A typical such alkaline cleanser contains 10 – 30 wt % sodium hydroxide and 1 – 4 wt % thickener (e.g., polyacrylic acid) in an aqueous solution.

Stainless steel facades are cleaned with mineral spirit – mineral oil emulsions. Heavily soiled or weathered objects require the use of either emulsions containing abrasives or weakly acidic cleansers containing 20 – 40 wt % phosphoric acid and 5 – 15 wt % surfactants in an aqueous medium. Such emulsions with abrasives typically contain 15 – 25 wt % mineral oil,

15 – 25 wt % mineral spirits, 5 – 10 wt % emulsifiers, and 10 –20 wt % abrasive (e.g., Neuburger siliceous chalk or diatomaceous earth) in an aqueous solution. *Anodized aluminum*[27] facades are cleaned with mineral spirit – mineral oil emulsions, although special acidic cleaners containing nitric acid [*7697-36-2*] [28], [29], or fluorosilicic acid [*16961-83-4*] can also be used [30], [31]. Mineral spirit – mineral oil emulsions typically contain 20 – 30 wt % mineral oil, 20 – 30 wt % mineral spirits, and 5 – 10 wt % emulsifiers in an aqueous solution.

2. Cleansers and Disinfectants for Use in the Food Industry

All surfaces that come in contact with foodstuffs during their manufacture, processing, or transport must undergo cleansing at prescribed intervals. The various branches of the food industry (e.g., beverage and canned goods manufacturers and the sugar, dairy, meat, and fats industries) are susceptible to different kinds of contamination; thus, cleansing in this context necessitates the availability of a wide variety of different chemical agents. Specific bacteriological concerns must also be addressed. Most can be satisfied only by a special disinfectant treatment, either subsequent to or as a part of the cleaning process. Numerous special disinfectants have been developed, and combination agents that contain mixtures of detergents and disinfectants are available.

2.1. Cleansers

Cleansing agents must be capable of removing contaminants containing proteins, fats, carbohydrates, and certain inorganic contaminants. The corresponding products are invariably mixtures of several components and are supplied as concentrated powders or liquids, less commonly as pastes; these are to be diluted with water to final concentrations of 0.5 – 2 wt % (0.05 –0.5 wt % for disinfecting agents). Such products must not corrode the various treated surfaces. They must dissolve readily and show a limited foaming tendency in largely automatic processes; they must also prevent the deposition that might otherwise result from hard water.

Decontaminators. The primary decontaminators usually are alkaline. These include sodium or potassium hydroxide, sodium or potassium carbonate, alkali-metal salts of orthophosphoric acid and organic acids, and sodium and potassium water glasses having silicon dioxide to alkali-metal oxide ratios of 0.7 – 3.3 [32].

Inorganic acids, such as sulfuric, nitric, and phosphoric acids, are used to remove largely inorganic contaminants such as milk or beer scale [33]. Hydrochloric and hydrofluoric acids are employed only in exceptional cases, such as in cleansers for masonry floors.

Sequestering Agents ("Builders"). A number of suitable organic and inorganic sequestering agents have been employed to complex the calcium and magnesium ions present in water and, thereby, prevent the formation of incrustations or deposits on boundary surfaces [34]. These substances also aid in the cleansing operation by breaking up soil lattices. The alkali-metal salts of the oligophosphoric acids, diphosphates, triphosphates, tetraphosphates, and more highly condensed phosphates (e.g., hexametaphosphate) remain the most important inorganic sequestering agents, but their hypertrophic effects on streams are a serious environmental concern.

Unlike phosphate, organic sequestering agents are not decomposed by hard industrial waters, and they are useful for alkaline and high-temperature applications. Nitrilotriacetic acid (NTA), ethylenediaminetetraacetic acid (EDTA), gluconic and polyacrylic acids and their alkali-metal salts, and phosphonic acid or phosphonates are examples of such organic sequestering agents. Both organic and inorganic sequestering agents usually perform several functions. They serve simultaneously as alkaline soil removers, as dispersants, and as water softeners and are among the most important cleanser ingredients.

Surface Active Agents. The removal of greasy dirt requires additional surface active agents [34]. Such anionic surfactants as fatty alcohol sulfates, alkylbenzenesulfonates, and soaps are used. The nonionic polyglycol ethers of fatty alcohols and alkylphenols, the latter syn-

thesized by reaction of the corresponding alcohols with ethylene oxide, are also suitable. The solubility and foaming characteristics of such compounds depend on the degree of ethoxylation. Propylene oxide adducts also find some application, as do alkylamine oxides. Heavily foaming mixtures of these latter surfactant, often in combination with alkaline or acidic cleansing agents, are used in cleansing operations that entail production of stable foams. Such foaming processes are preferred in the meat-cutting and meat-packing industries for external cleansing operations on processing and transport equipment. The stable foam permits the cleansing or disinfecting agent to act over a prolonged period of time.

Corrosion Inhibitors. Light metals such as aluminum and its alloys are attacked more or less vigorously by alkaline and most acidic cleansers [35]. Water glass (sodium silicate) is virtually the only agent used to minimize destructive corrosion under alkaline conditions. A number of inhibitors can be employed in acidic situations. The proper choice depends on the specific material and acidic cleanser. The increasing use of stainless steel is reducing the need for inhibitors. Tin- and zinc-plated utensils and equipment are rarely encountered.

Foam Inhibitors. Foam is a consequence of the normal behavior of components of the cleansing agents, and it can also be produced by certain contaminants, such as those encountered in beverage bottles. Special ethoxylated and propoxylated compounds can be introduced to reduce foaming, as can such water-insoluble materials as paraffin oil and silicones [34], [36]. Ethoxylated – propoxylated block polymers are particularly ideal for cleansing operations because these materials can show decreasing solubility with increasing temperature and thereby increasingly inhibit foaming as the temperature is raised during cleansing. At the low temperatures found in subsequent rinse steps, they regain their solubility and can be rinsed cleanly from the newly washed surfaces.

Dust Inhibitors, Anticaking Agents, and Solubilizers. Products marketed in powdered form often contain compounds that inhibit dust formation, thereby minimizing dust exposure for personnel working with them. Other additives may be needed to prevent caking of the powders. Aqueous products often require the presence of solubilizers, commonly derived from mono- and polyfunctional alcohols.

Enzymes. Enzymes such as proteases and amylases are useful in attacking protein- and starch-containing contaminants. They are particularly valuable for cleaning corrodible objects. An example is the cleaning of membranes employed in apparatus for ultrafiltration (reverse osmosis).

2.2. Disinfectants (→ Disinfectants)

The food industry employs as disinfectants mainly active chlorine compounds, quaternary ammonium salts, ampholyte soaps, iodophors, peroxy derivatives, and certain aldehydes [36]. For disinfection of the hands, various mono- and polyfunctional alcohols are used. Phenolic disinfectants are seldom used in the food processing industry. In addition to their microbiological activity [37], disinfectants must not be corrosive or cause foam, odor, or skin irritation. They may not leave residues on food unless these can be considered absolutely harmless.

Chlorine-Producing Compounds. The disinfectant use of salts of hypochlorous acid, both in powdered and liquid form, continues because they are inexpensive and have a broad spectrum of microbiological activity. Liquid products usually contain alkali-metal hypochlorites. Among powdered materials, the adducts of alkali-metal hypochlorites with trisodium phosphate and also the salts of mono-, di-, and, to some extent, trichlorocyanuric acid are used. Chloramines, such as chloramine-T (→ Chloramines), which form hypochlorites in aqueous solution are also used.

Hypochlorite is also used in combination disinfectant cleansers, which contain ingredients other than the cited alkaline substances to enhance cleansing. Dispersants, wetting agents, and corrosion inhibitors such as silicates are among these additives. Hypochlorite disinfectants typically contain 60 – 70 wt % sodium or potassium hypochlorite, 10 – 20 wt % sodium

or potassium hydroxide, and 3 – 5 wt % pentasodium or pentapotassium triphosphate in an aqueous solution.

Peroxy Derivatives. The common peroxy derivative disinfectants are hydrogen peroxide and peroxyborate, although salts of peroxodisulfuric acid are also occasionally used. Hydrogen peroxide and peroxyborate are noted for their long shelf life, but their disinfecting action is too slow for most purposes. Peroxy derivatives of the lower fatty acids, especially peroxyacetic acid, are more suitable because they display more rapid microbiological effects. Such products are used almost exclusively in acidic to neutral environments and usually contain stabilizers of various types. Both organic peracids and hydrogen peroxide are used in one-step acidic cleansing products.

Iodophors. Disinfectants based on iodine contain high molecular mass substances with surfactant properties, preferably nonionic wetting agents. These are particularly effective dispersants for elemental iodine and thereby promote its dissolution in water. Such disinfectants also contain varying amounts of phosphoric acid, antifoam agents, and corrosion inhibitors. They are usually offered as liquids, although materials in powdered form also exist. A disinfectant of this kind typically contains 10 –40 wt % phosphoric acid, 1 – 3 wt % iodine, 1 – 2 wt % corrosion inhibitor, 0 – 20 wt % isopropyl alcohol, and 1 – 6 wt % ethoxylated – propoxylated block polymer in an aqueous solution.

Quaternary Ammonium Compounds. Salts under this heading are neutral disinfectants that possess surfactant properties. The (C_{12-18})alkylbenzyldimethylammonium chlorides are examples. Various quaternary ammonium compounds are often combined to increase the activity of the product against specific microorganisms. Some quaternary ammonium compounds can be combined with phosphoric or sulfuric acid or with alkaline cleansing agents to yield disinfectants with cleansing properties. A typical quaternary ammonium salt disinfectant contains 5 – 20 wt % (C_{12-18})alkylbenzyldimethylammonium chloride, 5 – 10 wt % (C_{12-18})alkyldichlorobenzyldimethylammonium chloride, 1 – 3 wt % alkyl polyglycol ether, and 0 – 3 wt % ethoxylated – propoxylated block polymer in aqueous solution.

Ampholytes. Ampholytes, ampholytic surfactants, or ampholytic soaps have a betaine structure. They find use as disinfectants, often in combination with related materials or, occasionally, with quaternary ammonium compounds.

Halocarboxylic Acids. Halocarboxylic acids, particularly monochloro-, monobromo-, and monoiodoacetic acids, are used as disinfectants, especially in the beverage industry. The haloacids are supplied in combination with sulfuric or phosphoric acid, or both. Combinations including surfactants, foam inhibitors, sequestering agents, and solubilizers provide one-step disinfectant cleansers. Products based on a halocarboxylic acid are more stable than peroxide- or iodophor-based acidic disinfectant cleansers and, consequently, display a longer active lifetime. Their use in a carbon dioxide atmosphere for the cleansing and disinfecting of fermentation tanks in the brewing industry is particularly noteworthy (→ Beer).

Aldehydes and Alcohols. Only formaldehyde and glutaraldehyde are used as disinfectants to any extent. They are generally supplied in the form of mixtures containing other surfactants. Aqueous alcoholic products, often containing perfumes, are commonly employed for disinfecting the hands.

3. References

1. H. Lange in W. G. Cutler, R. C. Davies (eds.): *Surfactants Science Series,* **vol. 5,** Marcel Dekker, New York 1972, pp. 65 – 100.
2. M. J. Schwuger in Lukassen-Reynders (ed.): *Surfactants Science Series,* **vol. 11,** Marcel Dekker, New York 1980, pp. 267 – 314.
3. H. Lange in Henkel KGaA (ed.): *Waschmittelchemie,* Hüthig Verlag, Heidelberg 1976.
4. K. H. Disch, H. J. Heitland, E. Kiewert, *Hauswirtsch. Wissensch.* **28** (1980) 263.
5. D. Hentschler (ed.): *Gesundheitsschädliche Arbeitsstoffe,* vol. 3, Verlag Chemie, Weinheim 1972.

6. Unilever, DE-AS 1 293 372, 1961 (T. T. Jones, D. W. Stephens).
7. Henkel, DE 2 618 977, 1980 (K. Disch, E. Kiewert).
8. *Seifen Öle Fette Wachse* **108** (1982) 526 – 528.
9. *Seifen Öle Fette Wachse* **108** (1982) 558.
10. *Seifen Öle Fette Wachse* **108** (1982) 590.
11. *Seifen Öle Fette Wachse* **108** (1982) 648.
12. E. Nebe: *Teppichboden-Pflege,* Allround-Verlag, Solingen 1972.
13. H. Ellwanger, *Gebäudereiniger-Handwerk* **24** (1973) no. 5, 8 – 34;no. 8, 10 – 17; no. 9, 6 – 22.
14. *Gebäudereiniger-Handwerk* **24** (1973) no. 12, 6 (RAL 991a/2).
15. *Seifen Öle Fette Wachse* **109** (1983) 207 – 210.
16. *Seifen Öle Fette Wachse* **109** (1983) 241 – 242.
17. *Seifen Öle Fette Wachse* **109** (1983) 566.
18. *Seifen Öle Fette Wachse* **109** (1983) 582.
19. *Seifen Öle Fette Wachse* **109** (1983) 460.
20. *Seifen Öle Fette Wachse* **104** (1978) 49 – 50.
21. *Seifen Öle Fette Wachse* **104** (1978) 26.
22. *Seifen Öle Fette Wachse* **104** (1978) 170.
23. *Seifen Öle Fette Wachse* **104** (1978) 198.
24. *Seifen Öle Fette Wachse* **104** (1978) 264.
25. *Seifen Öle Fette Wachse* **104** (1978) 112.
26. H. Ellwanger, *Gebäudereiniger-Handwerk* **20** (1969) no. 6, 16, 18, 20 – 25; no. 7, 6, 8, 10 – 11; no. 8, 3, 4, 6, 8.
27. K. Broockmann, H. Ellwanger, A. Osberghaus, *Aluminium (Düsseldorf)* **58** (1982) 215 – 217.
28. Henkel, EP 0 066 117, 1984 (R. Osberghanl, H. Ellwanger, H. Marsen, H. Bossek).
29. H. Ellwanger, *Rationell Reinigen* **33** (1982) no. 3, 12 – 13.
30. Metallgesellschaft, DE 2 721 573, 1979 (J. Rones, S. Möller, H. Hoch).
31. J. Rones, *Aluminium (Düsseldorf)* **53** (1977) 493 –496.
32. G. Wildbrett: *Symposium über Reinigen und Desinfizieren lebensmittelverarbeitender Anlagen,* VDI-Ges. Verfahrenstechnik, Chemiewesen, Karlsruhe 1975, pp. 1 – 1/1 ff.
33. W. Mohr: *Die Reinigung und Desinfektion in der Milchwirtschaft,* T. Mann Verlag, Hildesheim 1954.
34. Henkel KGaA: *Waschmittelchemie,* Hüthig Verlag, Heidelberg 1976.
35. Aluminiumzentrale e. V.: *Das chemische Verhalten von Aluminium,* Aluminium-Verlag, Düsseldorf 1975.
36. H. Horn, M. Privora, W. Weuffen: *Handbuch der Desinfektion und Sterilisation,* vol. 1, VEB-Verlag Volk und Gesundheit, Berlin 1972.
37. H.-J. Schlüssler, *Z. Lebensm. Technol. Verfahrenstech.* 1980, no. 2, 46 – 49.
38. P. Hövelmann, B. Brackmann, *Seifen Öle Fette Wachse* **123** (1997) 697 – 701, 707.
39. B. Brackmann, *Seifen Öle Fette Wachse* **125** (1999) 40 – 42, 44 – 46.
40. H. G. Hauthal, *Tenside Surf. Det.* **36** (1999) 230 – 237.
41. IKW Arbeitskreis Handgeschirrspülmittel, Empfehlung zur Qualitätsbewertung von handgeschirrspülmitteln *Seifen, Öle, Fette* in press
42. A. Matero, A. Mattsson, M. Svensson, *J. Surfactants Deterg.* **1** (1998) 485.
43. *Jahrbuch für den Praktiker,* Verlag für Chemische Industrie, 1999, p. 147.
44. A Böhme, S. Brouwers, Henkel informiert, www.henkel.de
45. C. Fox, *Cosmet. Toiletries* **114** (1999) no. 10, 22.
46. H. I. Leidreiter, *Seifen Öle Fette Wachse* **122** (1996) 674.
47. H. I. Leidreiter, *Int. J. Cosmetic Sci.* **19** (1997) 239.
48. F. Müller, *Seifen Öle Fette Wachse* **125** (1999) 125.
49. HEA-Bilderdienst, “Geschirrspülmaschinen”, *Informationen über Elektrizität und ihre Anwendung, Teil 6.6,* 5. Auflage, Frankfurt/Main 1994.
50. R. Stamminger, *Hauswirtsch. Wissensch.* **46** (1998) 149 – 161.
51. Amtsblatt der Europäischen Gemeinschaften, Richtlinie 97/17/EG, 16. April 1997.
52. DIN EN 50242, Elektrische Geschirrspüler für den Hausgebrauch, Meßverfahren für Gebrauchseigenschaften.
53. C. Gutzschebauch, J. Härer, P. Jeschke, R. Schröder, D. Zaika, *Hauswirtsch. Wissensch.* **2** (1996) 51 – 63.
54. T. Altenschöpfer, *Fetten, Seifen, Anstrichmittel* **74** (1972) 3 – 20.
55. P. Jeschke, H. Andree, W. Buchmeier, B. Burg, J. Härer, *Seifen, Öle, Fette, Wachse* **119** (1993) 208 – 215.
56. T. Müller-Kirschbaum, J. Härer, C. Nitsch, WfK-Tagung, Krefeld, 1998.
57. B. Burg, J. Härer, P. Jeschke, D. Speckmann, W. von Rybinski, *Seifen, Öle, Fette, Wachse* **120** (1994) 400 – 404.
58. Henkel, DE 2 913 145, 1980 (F. J. Carduck et al.).
59. B. Middelhauve, J. Penninger, *Int. J. Appl. Sci. – Personal Care – Detergents – Specialities (SÖFW J.)* **127** (2001) no. 6, 3 – 7.

60. H. Burlinson, *Tablets and Tabletting,* Heinemann, London, 1968.
61. J. Zimmer, *Einfluß der Press- und Lageratmosphäre auf die physikalischen Eigenschaften von Tabletten,* VDI, Duesseldorf, 1996.
62. G. Heinze, *Handbuch der Agglomerationstechnik,* Wiley-VCH, Weinheim, 1999
63. DIN EN 50242, Elektrische Geschirrspüler für den Hausgebrauch, Meßverfahren für Gebrauchseigenschaften.
64. IEC 436, Methods for Measuring the Performance of Electric Dishwashers, Bureau Central de la Commission Electrotechnique Internationale, Genève 1981.
65. IKW Arbeitskreis Maschinenspülmittel, "Methoden zur Bestimmung der Reinigungsleistung von maschinellen Geschirrspülmitteln", *Seifen, Öle, Fette, Wachse,* **124** (1998) nos. 11 (Teil A) and 14 (Teil B).
66. DIN-EN 12875-1, Mechanische Geschirrspülmaschinenbeständigkeit von Haushaltswaren, Referenz-Prüfverfahren.
67. DIN-EN 12875-2, Mechanische Geschirrspülmaschinenbeständigkeit von Haushaltswaren, Inspektion von nichtmetallischen Gegenständen.
68. W. Buchmeier, P. Jeschke, R. Sorg, *Glastech. Ber. Glass Sci. Technol.* **69** (1996) no. 6.
69. B. Burg, J. Härer, P. Jeschke, D. Speckmann, W. von Rybinski, "Korrosionsphänomene an Silberoberflächen beim maschinellen Geschirrspülen", *Seifen, Öle, Fette, Wachse,* **120** (1994) no. 7.
70. J. Shulman, M. S. Robertson, Resin Review, A Rohm and Haas Company Publication, Vol. XLII, 1992, No. 2.
71. Stiftung Warentest, Test, Heft 1 (2000) S. 67.
72. ASTM D 3556-85, Standard Test Method for Deposition on Glassware During Mechanical Dishwashing.
73. W. Schirmer, T. Altenschöpfer, W. Wickelham, H. Andree *Tenside Surf. Det.* **28** (1991) 313 – 319.
74. H. Magg, 1. International Fresenius Fachtagung Wasch- und Reinigungsmittel, Feb. 2000.

Coagulants → Blood

Coal

John C. Crelling, Southern Illinois University at Carbondale, Carbondale, Illinois, United States (Chap. 1, 2, 3, 4, 10, 11 and 12)

Hans W. Hagemann, Technische Universität Aachen, Aachen, Federal Republic of Germany (Chap. 1, 2, 3 and 4)

Dieter H. Sauter, Lurgi GmbH, Frankfurt, Federal Republic of Germany (Chap. 5 and 8)

Raja V. Ramani, The Pennsylvania State University, University Park, Pennsylvania, United States (Chap. 6)

Werner Vogt, Technische Universität Clausthal, Clausthal-Zellerfeld, Federal Republic of Germany (Chap. 6.3)

Dieter Leininger, Bergbauforschung GmbH, Essen, Federal Republic of Germany (Chap. 7)

Steffen Krzack, Technische Universität Bergakademie Freiberg, Freiberg, Federal Republic of Germany (Chap. 7.2, 9, 9.1, 9.3 and 9.6)

Bernd Meyer, Technische Universität Bergakademie Freiberg, Freiberg, Federal Republic of Germany (Chap. 7.2, 9, 9.1, 9.3 and 9.6)

Fred Orywal, Heiligenhaus, Federal Republic of Germany (Chap. 9.2)

Rainer Reimert, Universität Karlsruhe, Karlsruhe, Federal Republic of Germany (Chap. 8 and 9.4)

Bernhard Bonn, Essen, Federal Republic of Germany (Chap. 9.5)

Udo Bertmann, Bergbauforschung GmbH, Essen, Federal Republic of Germany (Chap. 9)

Wolfgang Klose, Universität Gesamthochschule Kassel, Kassel, Federal Republic of Germany (Chap. 12)

Günter Dach, German Hard Coal Association, Essen, Federal Republic of Germany (Chap. 13)

1. Coal Petrology 2
1.1. Coal Characterization 2
1.2. The Maceral Concept 2
1.3. Maceral Types and Properties . . 3
2. Coalification 5
3. Occurrence 6
3.1. Coal Seam Occurrence 6
3.2. Coal Seam Structures 6
3.3. Coal Seam Distribution 7
4. Classification 8
5. Chemical Structure of Coal 10
5.1. Characterization of Coals 12
5.1.1. Standard Analytical Methods 12
5.1.2. Laboratory and Bench-Scale Simulation Tests 13
5.1.3. Petrographic Studies 13
5.2. Structural Deductions from Analytical and Bench-Scale Data . . . 13
5.3. Bonding of Elements in Coal . . . 14
5.4. Structural Evidence of Coals . . . 14
6. Coal Mining 14
6.1. General Considerations 14
6.2. Surface or Underground Mining 15
6.3. Coal Methods 16
6.4. Underground Coal Mining Methods 20
7. Coal Preparation 24
7.1. Hard Coal Preparation 25
7.1.1. Preliminary Treatment 27
7.1.2. Wet Treatment 27
7.1.2.1. Cleaning in Jigs 27
7.1.2.2. Cleaning in Dense-Media Separators 28
7.1.2.3. Cleaning by Flotation 28
7.1.2.4. Other Cleaning Methods 28
7.1.3. Dewatering 28
7.1.4. Decantation and Thickening of the Process Water 29
7.1.5. Dosing and Blending 29
7.1.6. Removal of Pyritic Sulfur 30
7.1.7. Thermal Drying 30
7.2. Brown Coal Preparation 30
8. Agglomeration 30
9. Coal Conversion (Uses) 31
9.1. Pyrolysis 32
9.2. Carbonization and Coking 33
9.3. Coal Liquefaction 33
9.3.1. Direct Liquefaction (Hydrogenation) 34
9.3.2. Indirect Liquefaction 34

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_153.pub2

9.4. Coal Gasification 35
9.5. Coal Combustion 36
9.6. Nonenergetic Use of Coal 38
9.6.1. Extraction of Montan Wax 38
9.6.2. Coal Tar Chemical Industry 38
9.6.3. Production of Carbonaceous Adsorbents 39
10. Transportation 39
11. Coal Storage 40
12. Quality and Quality Testing 40
12.1. Chemical Analyses 40
12.2. Mineral Matter 41
12.3. Thermal Properties 42
13. Economic Aspects 43
13.1. World Outlook 43
13.2. Some Major Coal Producing Countries 44
13.2.1. China 44
13.2.2. United States 44
13.2.3. Former Soviet Union 45
13.2.4. India 45
13.2.5. Australia 45
13.2.6. Republic of South Africa 45
13.2.7. Indonesia 45
13.2.8. Poland 46
13.2.9. Germany 46
13.2.10. Columbia 46
14. References 46

1. Coal Petrology

1.1. Coal Characterization

Coal is an extremely heterogeneous, complex material that is difficult to characterize. Coal is a rock formed by geological processes and is composed of a number of distinct organic entities (macerals, i.e., the microscopically recognizable individual constituents of organic matter) and lesser amounts of inorganic substances (minerals). Each of the coal macerals and minerals has a unique set of physical and chemical properties; these in turn control the overall behavior of coal. Although much is known about the properties of minerals in coal, for example, the crystal chemistry, crystallography, and magnetic and electrical properties, surprisingly little is known about the chemical and physical properties of individual coal macerals [1]. Even though coal is composed of macerals and minerals, it is not a uniform mixture of these substances. The macerals and minerals occur in distinct associations. These macroscopic association are called lithotypes, and each lithotype has a set of physical and chemical properties that also affect coal behavior.

Coal seams, the basic units in which coal occurs, are in turn composed of layers of coal lithotypes, and individual coal seams may also have their own sets of physical and chemical properties. For example, even if two coal seams have the same average maceral and mineral composition, the seams may have significantly different properties if the maceral associations in lithotypes in the two seams are different. The enclosing rocks immediately adjacent to a coal seam can also affect the properties of the coal. This aspect is particularly important in mine design, production, and strata control. The compositional characterization of a coal seam must cover the nature of the macerals, the lithotypes, the entire seam, and the association of the seam with the neighboring strata.

In addition to compositional factors, coal properties also change with the rank or the degree of coalification of a given sample of coal. Coal is part of a metamorphic series ranging from peat, through lignite and subbituminous and bituminous coal, to anthracite. Temperature, pressure, and time mainly alter the botanical precursors of coal (plant fragments) through this metamorphic series. As the rank of the coal changes, the properties of the coal macerals change progressively and, therefore, so also do the properties of lithotypes and the entire seam.

Because of these factors, coal characterization requires a detailed knowledge of the type (maceral composition), the rank (degree of coalification) and the grade (content of impurities) of the coal. All coal properties are ultimately a function of these three factors.

1.2. The Maceral Concept

The term maceral was introduced by Stopes to distinguish the organic components of coal (macerals) from the inorganic components (minerals) [2]. The term is now interpreted in two conflicting ways. The interpretation of the International Committee for Coal Petrology (ICCP)

is that a maceral is the smallest microscopically recognizable component of coal [3]. This conception, generally held by most coal petrographers all over the world, is based on morphology and other criteria such as size, shape, botanical affinity, and occurrence. The ICCP concept implies that the properties of the macerals change with rank. Thus, the same maceral, vitrinite, can exist in a bituminous coal as well as in an anthracite.

The basis of the petrographic study of coal composition is the idea that coal is composed of a number of distinct macerals. The entire body of coal petrographic literature supports this idea and is in direct contrast to the earlier chemical concept of coal being composed of a single unique molecular substance. Although the concept and the term "coal molecule" may once have been useful, modern chemical studies of coal reveal different molecular frameworks depending on coal's heterogeneous maceral composition and rank.

1.3. Maceral Types and Properties

A large number of macerals have been described in terms of their appearance in reflected light when observed through an oil immersion objective. The Stopes – Heerlen system used by the ICCP is given in part in Table 1. For routine maceral analysis, the conventional subdivision into three maceral groups, i.e., vitrinite, liptinite (formerly called exinite) and inertinite is useful.

As shown in Table 1, coal macerals fall into three distinct groups: vitrinite, liptinite, and inertinite. *Vitrinite macerals* are generally the most abundant, commonly making up 50 to more than 90 % of the Eurameric coals (European and North American coals of Carboniferous age). This group is not as abundant in Permian Gondwana coals that originate in the southern hemisphere.

Most vitrinite macerals are derived from the cell wall material (woody tissue) of plants. Although the details of the complexity of the vitrinization process are not well understood, it is generally believed that during the coalification process the plant cell wall material is chemically altered and broken down into colloidal particles that are later deposited and desiccated. This process commonly homogenizes the components so that the resulting macerals are structureless. The variation in vitrinite macerals is usually thought to be due to differences in the original plant material or to different conditions of alteration at the peat stage or during later coalification.

Table 1. Macerals of hard coal according to the Stopes – Heerlen system (after ICCP 1957, 1963, 1971, 1975, 1985 [3])

Maceral group	Maceral	Submaceral*
Vitrinite	telinite	telinite 1
		telinite 2
	collinite	telocollinite
		gelocollinite
		desmocollinite
		corpocollinite
	vitrodetrinite	
Liptinite	sporinite	
	cutinite	
	resinite	
	alginite	
	liptodetrinite	
Inertinite	micrinite	
	macrinite	
	semifusinite	
	fusinite	pyrofusinite
		degradofusinite
	sclerotinite	fungosclerotinite
	inertodetrinite	

* This category can be arbitrarily extended, if required

Under the microscope, vitrinite macerals have a reflectance (brightness) between that of the liptinite and inertinite macerals. Because the reflectance of the vitrinite macerals shows a more or less uniform increase with coal rank, reflectance measurements for the determination of rank are always taken exclusively on vitrinite submacerals (collinite, telocollinite, desmocollinite). The reflectance of vitrinite macerals is also anisotropic, so that in most orientations a particle of vitrinite will display two maxima and two minima with complete rotation. Vitrinite is almost always the most abundant maceral present and makes up the groundmass in which the various liptinite and inertinite macerals are dispersed. It has a uniform gray color and is always anisotropic. With UV excitation, some vitrinites will fluoresce.

A modernized vitrinite classification and nomenclature system was introduced by the ICCP in 1994 [4]. As shown in Table 2, six vitrinite macerals have been grouped into three subgroups according to their degree of degradation: telovitrinite, detrovitrinite, and gelovitrinite.

Table 2. Subdivision of the group of vitrinites according to the ICCP System 1994

Subgroup	Maceral
Telovitrinite	telinite
	collotelinite[a]
Detrovitrinite	vitrodetrinite
	collodetrinite[b]
Gelovitrinite	corpogelinite[c]
	gelinite[d]

According to ICCP (1963, 1971, 1975, 1985 [3]):
[a] elocollinite;
[b] desmocollinite;
[c] corpocollinite;
[d] gelocollinite

Liptinite macerals are derived from the hydrogen-rich parts of plants, i.e., the spores and pollen, cuticles, resins, and algae. This group generally makes up 5 – 15 vol % of most Carboniferous coals, although it dominates some unusual types of coal such as cannel and boghead. In any given coal, liptinite macerals have the lowest reflectance. Liptinite macerals are the most resistant to alteration or metamorphism in the early stages of coalification; thus, the reflectance changes are slight up to the rank of medium-volatile coal. In this range the reflectance of liptinite macerals increases rapidly until it matches or exceeds the reflectance of vitrinite macerals in the same coal and, thus, essentially disappears. Sporinite is the most common of the liptinite macerals and is derived from the outer waxy cell walls of fossil spores and pollen. It generally has the form of a flattened spheroid with upper and lower hemispheres compressed until they fuse. The outer surfaces of the sporinite macerals often show various kinds of ornamentation. In Paleozoic coals, two sizes of spores are common. The smaller ones, usually $< 150\,\mu m$ in diameter, are called microspores, and the larger ones, ranging up to several millimeters in diameter, are called megaspores. Cutinite, found as a minor component in most coals, is derived from cuticles, which are formed within the outer cell wall of the epidermis of leaves, roots, stems, etc. of plants. It occurs as long stringers, which often have one fairly flat surface and another that is crenulated. Cutinite usually has a reflectance equal to that of sporinite. Occasionally the stringers of cutinite are distorted. Resinite is also common in most coals and usually occurs as ovoid bodies with a reflectance slightly greater than that of sporinite and cutinite but still less than that of vitrinite. Most resinites originate from resins and waxes. Some resinite may appear to be transparent and strongly yellow-fluorescing, at the subbituminous stage of coalification. In some cretaceous and tertiary coals, a number of different forms of resinite may be distinguished by using fluorescent microscopy. Alginite is derived from fossil algae colonies. It is rare in most coals and is often difficult to distinguish from mineral matter. However, under excitation with blue light it fluoresces with a brilliant yellow color and can display a distinctive cauliflower-like appearance.

The *inertinite* maceral groups are mainly derived from the same plant material as the vitrinites, usually woody tissue, that has been strongly altered by charring in a forest fire or by biochemical processes such as composting either before or shortly after deposition. These processes are described as fusinitization. These macerals can make up 5 – 40 % of most Eurameric coals. In southern hemisphere coals of Permian/Triasic age, this group commonly is more abundant than vitrinite. The inertinite macerals have the highest magnitude and greatest range of reflectance of all the macerals. They are distinguished by their relative reflectances and by the presence or absence of cell texture. Fusinite is seen in most coals and has a charcoal-like structure (pyrofusinite). It is always the highest reflecting maceral present and is distinguished by a cell texture that is commonly broken into small shards and fragments. Semifusinite can have the cell texture and general features of fusinite except that it is of lower reflectance. But normally the cell textures of these macerals are poorly preserved. Semifusinite is also the most abundant of the inertinite macerals in most Permian Gondwana coals. Macrinite is a very minor component of most coals and usually occurs as structureless ovoid bodies or as structureless groundmass with the same reflectance as fusinite. Micrinite occurs as very fine granular particles of high reflectance. It is commonly associated with the liptinite macerals and sometimes gives the appearance of actually replacing the liptinite. The genesis of macrinite and micrinite is uncertain or insufficiently understood. Sclerotinite, found as a minor component in most hard coals, has been referred to as fungal remains, but the greater portion of sclerotinite, which is

found in carboniferous and permian bituminous coals originates from cell secretions. The maceral inertodetrinite contains mainly small fragments of fusinite, semifusinite, sclerotinite, and macrinite. It is concentrated in many subaquatic coal. Some southern hemisphere coals contain characteristic concentrations of inertodetrinite.

2. Coalification

A unique and often troublesome feature of coal that distinguishes it from other fuels and bulk commodities is its property of rank. All coal starts out as peat, which is then changed into progressively higher ranks of coal. This transformation is generally divided into two phases. The first, which occurs in the peat stage, is called diagenesis or biochemical coalification. In this phase most of the plant material making up the peat is biochemically broken down. Specifically, most of the cellulose in the plant material is digested away by bacteria and fungi, and the lignin in the plant material is transformed into humic acids and humic compounds, humins. Some plant material is also thermally altered by partial combustion or biochemical charring. Still other plant material, such as spores and pollen, survive the diagenesis stage without much change. After diagenesis is complete and the altered peat is buried, geological forces begin to act in the geological or metamorphic phase of coalification.

Because of the common occurrence of high-rank coals in geologically deformed areas, for example, anthracite in the eastern Appalachian Mountains and low-volatile bituminous coal in the folded and faulted Canadian Rocky Mountains, it was assumed that tectonic pressure was responsible for the high rank of most coals. However, this position has now lost most of its support because studies in the folded coal seams of the Ruhr Basin show that isorank lines (isovols) follow the folds in the seams [5]. This shows that the coal achieved its rank before the structural deformation and was not altered during or after such deformation. Similar studies of stratigraphic relationships in the Canadian Rocky Mountains showed that the higher rank coals correlated with increasing depth of burial and not tectonism [6]. Laboratory experiments using increased confining pressure have not led to increased coalification [7].

The majority of the geological evidence suggests that temperature is the major factor in coalification and that the temperature range in which most coalification takes place is 50–150 °C. The depth at which these temperatures occur is a function of the natural geothermal gradient (dT/dZ), which ranges from 0.8 °C/ 100 m to 4 °C/100 m. Therefore, at these two geothermal gradient extremes, the depth at which 150 °C would be reached, if a surface temperature of 20 °C is assumed, is 16.25 km (53 300 feet, 10.1 miles) for the lowest gradient and 3.25 km (10 700 feet, 2.02 miles) for the highest. The major lines of evidence for the importance of temperature in coalification are as follows:

1) An increase in coal rank and temperature with depth is confirmed by thousands of bore holes.
2) Lines of equal rank contour igneous intrusions and show an increase in rank in the direction of the intrusion. In some cases, natural coke is found at the contact of a coal seam with an intrusion.
3) Laboratory experiments show that wood can be coalified by heating in an inert atmosphere. Increasing the pressure alone has no effect on the coalification of wood.
4) When thermodynamic factors are considered, it is clear that temperature should have more of an effect than pressure. While an increase in temperature will encourage most coalification reactions by increasing the available energy, an increase in pressure may often inhibit such reactions by raising the energy requirements for such reactions to take place. However, pressure changes the physical structure of the coal.

It was long thought that time was not a factor in coalification, and the example of the Russian paper coals of lignite to lower subbituminous rank and Lower Carboniferous age was often cited as evidence. However, it is now believed that if the peat has been buried deeply enough to effect coalification, then time does have a "soaking effect". For example, in Venezuela there is a subbituminous coal of Eocene age that has been buried at a depth of 1120 – 1220 m at 125 °C for $(10-20)\times10^6$ years. In Germany there is a coal that has been at the same depth and temperature for 300×10^6 years and is of anthracite rank [5]. Numerous other examples have been found and a number of correlation charts relating depths of

burial, temperature, rank, and time are based on this type of data.

Changes in the coal rank correspond to changes in most of the properties of the coal. For example, as rank increases, moisture, volatile matter, and ultimate oxygen and hydrogen decrease, whereas fixed and ultimate carbon, calorific value, and reflectance increase. All of these measurements and even some combinations of these have been used as measures of coal rank. However, with the exception of reflectance, they all suffer from two major drawbacks. First, none actually change uniformly across the rank range of coal. Second, they are all bulk properties of coal and, thus, can be significantly affected by changes in maceral composition having nothing to do with rank. For example, a coal with a higher-than-normal content of liptinite macerals can have a higher-than-normal hydrogen content and, therefore, appear to have a lower rank than it actually does on the basis of a rank parameter independent of composition such as reflectance. The reflectance parameter of coal is based on the amount of light reflected from the vitrinite macerals in a coal compared to a calibration standard of known refractive index and reflectance. The vitrinite reflectance of coal changes uniformly across most of the coal rank range. However, it is not very sensitive in the lowest rank range (lignite to lower subbituminous).

3. Occurrence

3.1. Coal Seam Occurrence

Although coal particles are scattered throughout many rock units, most coal occurs in seams, which can range in thickness from a few centimeters to over 100 m. In most of the world, anthracite and bituminous coals occur in thinner seams of 1 – 2 m while the lower rank coals, lignite and subbituminous, commonly occur in thicker seams of up to 20 – 30 m, and sometimes more than 100 m (Germany, Australia). While many seams, even minable seams, are of a limited size area, some seams continuously underlie large areas. The Pittsburgh coal seam, for example, extends over 77 700 km^2 in Pennsylvania, Ohio, West Virginia, and Maryland, and it is of minable thickness for over 15 540 km^2.

In some coal basins, there are hundreds of named coal seams, but normally only a few are of the quality, thickness, and size of area to be extensively exploited commercially. Other coal basins exist, where only one to three coal seams are developed in minable tickness.

3.2. Coal Seam Structures

All coal seams have a number of structural features, including partings, splits, rolls, cutouts, cleats, faults, folds, and alterations caused by igneous intrusions. These features strongly affect the mining and economical recovery of the coal. *Partings* are layers of rock, usually shale or sandstone, that occur within a coal seam and were caused by an influx of sediment into the original coal swamp. In commercial seams, the partings must, of necessity, be few and thin (a few centimeters). However, even in such seams the partings can be very extensive. When the thickness of a layer of rock within a seam increases to the point where it is no longer practical to mine the parting with the coal, the seam can be considered to have *split.* Although some seams split into three or more minable seams, the splitting of one seam into multiple seams can present serious problems in mining, including correlation (i.e., tracing a given seam across a zone, e.g., a valley, where it is missing) and loss of minable thickness.

In addition to partings and splits within the coal seam, the upper and lower surfaces of a seam often pinch and swell to change the coal seam thickness. Such features have been called *pinches, rolls, horsebacks*, and *swells*, and their occurrence at the base of a seam is usually attributed to differential compaction. Roof rolls are common and the protrusion of rock into the coal causes some serious problems with mine roof stability. *Cutouts* or stream washouts are the extreme cases where the protrusion actually eliminates the coal seam. These features are clearly the result of nondeposition or erosion by ancient streams associated with the coal swamps.

Another important feature of coal seams, especially in bituminous coals, is the presence of closely spaced fractures within the coal. These fractures, called *cleats*, are usually perpendicular to the bedding plane of the coal and com-

monly occur in two sets perpendicular to each other. This gives the coal a tendency to break into blocks. The cleat controls the ease with which the coal breaks up, and it has long been used in coal mining. The most prominent cleat is called the face cleat because the working face of a mine is often parallel to it. The other cleat is called the butt cleat. Cleats give coal high permeability to gas and groundwater and also act as sites of mineral deposition. Calcite and pyrite are the most common cleat-filling minerals, although other minerals including gypsum can occur.

Faulting and folding of coal seams by geological forces can also be important features. Seams that are faulted or folded usually cannot be mined by surface methods and are more expensive to mine than flat seams. When faulting is extensive, even thick seams of high quality may be too discontinuous to mine. On the other hand, folding can double the minable thickness of some seams.

The *alteration* of coal seams by dikes, sills, and other types of igneous intrusions is widespread and is a serious problem in some coal fields, such as those in the Rocky Mountains of the United States or in the Natal coal fields in South Africa, which have been affected by dolerite intrusions. The alteration is thermal and causes an increase in carbon content and a decrease in hydrogen, moisture, and volatile matter. In the contact zone, the coal may be transformed into natural coke. On the other hand, mining is rendered difficult because of the hard igneous rocks.

3.3. Coal Seam Distribution

Although coal seams are found in rocks of all geologic ages since the Devonian, the age distribution is not even. The first sequences of real coal seams formed from higher terrestrial plants occur in Carbonaceous shales and sandstones of Upper Devonian age. The classical example is the coal deposit of the Bear Island (Norway). Major coal deposits of Carboniferous age occur in eastern and central North America (Pennsylvania, Illinois, Newfoundland, Nova Scotia, Cape Breton Island), in north western Europe (British Isles, Ruhr and Saar-Lorraine coal fields, Upper Silesian coal fields), and the Ukrainian coal fields. Major deposits of Permian age occur in South Africa (Witbank and Natal coal fields), South America (Morungava coal basin) India (e.g., Raniganj and Darjeeling coal fields), Australia (Greta and Tomago coal measures, Collie and Coorabin coal fields), China, and Antarctica. In late Triassic and Jurassic times the major coal accumulation was in Australia (Calide and Milmerran coal, Gunnedah and Eromanga coal basin, Wallon coal field), and parts of the former Soviet Union and China. Coal deposits of the Cretaceous period are found in the Rocky Mountains of North America (Baltimore canyon trough, Cordillera, Foothill belt, Arctic archipelago, Kootenay coal), in New Zealand, and Spitzbergen. The last great period of coal deposition was in the Tertiary period. Coals originating from this time are found in central and south eastern Europe (Germany, Poland, Czechia, former Yugoslavia, Romania, Bulgaria, Greece), Russia, Australia (Victoria, Gippsland basin), New Zealand, Japan, China, Indonesia, United States (Western coal, Texas) and Canada (Arctic Archipelago, British Columbia). Because the coal deposits of Cretaceous and Tertiary age are younger than the Carboniferous and Permian coals, these coals tend to be of lower rank, usually subbituminous or lignitic, in comparison with the older bituminous and anthracitic coals. However, there are many exceptions to the rule.

The distribution of coal seams throughout the world is also not uniform. As shown in Figure 1, more than 90 % of the world's coal resources are located in the northern hemisphere. Although the figures vary from source, China, the former Soviet Union, and the United States together have about 75 % of the total coal resources and reserves. Each of these countries has a coal production of $> 500\times10^6$ to $> 1000\times10^6$ t/a.

Figure 1. Geographic location of the world's coal

Table 3. Classification of coal by rank [a]

Class	Group	Fixed carbon limits, % (dry, mineral-matter-free basis)		Volatile matter limits, % (dry, mineral-matter-free basis)		Calorific value limits, Btu/lb (moist, mineral-matter-free basis) [b]		Agglomerating character
Anthracite	metaanthracite	≧98	–	–	≦2	–	–	nonagglomerating
	anthracite	≧92	< 98	> 2	≦8	–	–	nonagglomerating
	semianthracite [c]	≧86	< 92	> 8	≦14	–	–	nonagglomerating
Bituminous	low-volatile bituminous coal	≧78	< 86	> 14	≦22	–	–	commonly agglomerating [e]
	medium-volatile bituminous coal	≧69	< 78	> 22	≦31	–	–	commonly agglomerating [e]
	high-volatile A bituminous coal	–	< 69	> 31	–	≧14 000 [d]	–	commonly agglomerating [e]
	high-volatile B bituminous coal	–	–	–	–	≧13 000 [d]	< 14 000 [d]	commonly agglomerating [e]
	high-volatile C bituminous coal	–	–	–	–	≧11 500	< 13 000 [d]	
						≧10 500	< 11 500	agglomerating
Subbituminous	subbituminous A coal	–	–	–	–	≧10 500	< 11 500	nonagglomerating
	subbituminous B coal	–	–	–	–	≧ 9 500	< 10 500	nonagglomerating
	subbituminous C coal	–	–	–	–	≧ 8 300	< 9 500	nonagglomerating
Lignitic	lignite A	–	–	–	–	≧ 6 300	< 8 300	nonagglomerating
	lignite B	–	–	–	–	–	< 6 300	nonagglomerating

[a] This classification does not include a few coals, principally nonbanded varieties, which have unusual physical and chemical properties and which come within the limits of fixed carbon or caloric value of the high-volatile bituminous and subbituminous ranks. All of these coals either contain < 48 % dry, mineral-matter-free fixed carbon or have > 15 500 Btu/lb moist, mineral-matter-free.
[b] Moist refers to coal containing its natural inherent moisture but not including visible water on the surface of the coal.
[c] If agglomerating, classify in low-volatile group of the bituminous class.
[d] Coals having 69 % or more fixed carbon on the dry, mineral-matter-free basis shall be classified according to fixed carbon, regardless of caloric value.
[e] It is recognized that there may be nonagglomerating varieties in these groups of the bituminous class, and there are notable exceptions in high-volatile C bituminous group.

4. Classification

Coal is combustible and should be composed of more than 50 wt % carbonaceous material [8, 9]. Commercially, coal is classified in a number of ways on the basis of (1) the original plant or maceral composition, sometimes called coal type, (2) the degree of maturity or metamorphism, called coal rank, (3) the amount of impurities such as ash or sulfur, called coal grade, and (4) the industrial properties such as coking or agglomeration.

The various bands or layers in coal evident to the unaided eye have also been classified into four types [10]. *Vitrain* layers appear bright and vitreous; *clarain* appears as relatively less bright, striated layers; *durain* is dull and featureless; *fusain* layers are dull gray and like charcoal. Although these terms (all ending in ain) are megascopic terms meant to be applied to hand-specimen samples, they do have some compositional implications at the microscopic level. For example, vitrain layers contain mainly vitrinite macerals, fusain layers contain mainly inertinite macerals, and clarain and durain are mixtures of all three maceral types.

The most important classification for commercial purposes in the United States and many other countries is the ASTM classification by rank. It is the basis on which most of the coal in the United States is bought and sold. This classification, ASTM Standard D 388 shown in Table 3, divides coals into 4 classes, anthracite, bituminous, subbituminous, and lignitic, which are further subdivided into 13 groups on the basis of fixed carbon and volatile matter content, calorific value, and agglomerating character. The fixed carbon and volatile matter values

are on a dry, mineral-matter-free basis and the calorific values are on a moist, mineral-matter-free basis. In this system, coals with $\geq 69\,\%$ fixed carbon are classified by fixed carbon content and those with $< 69\,\%$ fixed carbon content are classified by calorific value. Thus, all lignitic and subbituminous coals and the lower rank bituminous coals are classified by their calorific value. It is also important to note that not all coals can be fitted into this system. This is especially true of coals with a high liptinitic maceral content, such as cannel and boghead types.

In the last two decades coal classification has been discussed extensively. These discussions involve exchange of ideas, e.g., with the Economic Commission for Europe of the United Nations (ECE-UN). As a result the "International Codification System for Medium and High Rank Coals" (exclusively for commercial and utilization purposes, involving the international coal trade) has been formalized and approved. With this new codification system (ECE-Coal-115, 1988) producers, sellers, and purchasers can communicate unambiguously about the quality coal must possess to be suitable for certain industrial applications (e.g., combustion, carbonization, gasification, or hydrogenation) as well as general and environmental aspects of coal quality. This system is applicable to all bituminous coals of different origin and geological age. It is also relevant for all seam coals, multiseam blends of run-of-mine coals, raw, and washed coals. The threshold between low-rank and higher-rank (medium- and high-rank) coals is defined and set at a gross calorific value (moist, ash-free basis, i. e., maf) of ≥ 24 MJ/kg, provided the mean random reflectance of the vitrinite is $\geq 0.6\,\%$.

The coals are characterized by eight coal property-related parameters which use a 14-digit code number to describe the rank of the composition of the coal types and the quality of the coal (see Table 4). These include the mean random reflectance (Rr %), reflectogram (distribution of reflectance readings), maceral group composition index, swelling crucible number, volatile matter content (dry, ash-free basis, i..e., daf) ash content (dry basis, db, in %), and gross calorific value (daf, MJ/kg).

A summary of the codes for medium- and high-rank coals is given in Tables 5 and 6. Table 7 presents an example of coal coding, a carboniferous coal blend from the northern hemisphere with good coking properties.

Table 4. International codification system for medium- and high-rank coals

Basic parameters	Digit
(a) Mean random reflectance of vitrinite Rr, %	1; 2
(b) Characteristics of the reflectrogram (standard deviation, gap(s))	3
(c) Maceral group composition index, % by volume (mmf) [a] digit 4 = Inertinite, digit 5 = Liptinite	4; 5
(d) Crucible swelling number	6
(e) Volatile matter content, mass % (daf) [b]	7; 8
(f) Ash content, mass % (db) [c]	9; 10
(g) Total sulfur content, mass % (db)	11; 12
(h) Gross calorific value, MJ/kg (daf)	13; 14

[a] mmf = mineral-matter-free;
[b] daf = dry, ash-free;
[c] db = dry basis

A codification system for low-rank coals, i.e., coals with a gross calorific value (maf) of < 24 MJ/kg and a mean random reflectance of the vitrinite (Rr) $< 0.6\,\%$ has not been drawn up until now and will be dealt with by the ECE-UN at a later stage. Additional parameter for subdivision of these low-rank coals may be total moisture content (as-received basis, mass %), ash content (db, mass %), and total sulfur content (db, mass %).

An "International Classification of In-seam Coals" for geological assessment of coal deposits, coal mining, and industrial purposes has been developed by the ECE-UN Working Party on Coal (Fig. 2). The work was supported by the ICCP, the International Energy Agency (IEA/OECD) and the ISO. The classification is not to be used for commercial or trade purposes. The systematization of coal is based on three fundamental characteristics which, according to the present knowledge, provide clear discrimination (a schematic version is given in Figure 3). The characteristics are (1) rank (degree of coalification; (2) type (petrographic composition); and (3) grade (amount of impurities). The main rank divisions lignite, subbituminous, bituminous, or, alternatively, low-, medium-, and high-rank coals are based on the vitrinite mean random reflectance (Rr, %) and gross calorific value (GCV, MJ/kg). Petrographic composition is expressed in terms of maceral group analysis, i. e., the vitrinite (V), liptinite (L), and inertinite (I) content in vol %, all mineral matter free (mmf).

Table 5. International codification system (digits 1 – 6) for medium- and high-rank coals

	Random reflectance of vitrinite, %		Characteristics of reflectogram			Maceral group composition index, vol % (mmf) [a] 4 = Inertinite; 5 = Liptinite				Swelling crucible number	
Digit	1 ; 2		3			4		5		6	
Code no.	02	0.20 – 0.29	0	≤ 0.1;	no gap	0	0 – < 10	0	exempted	0	0 – 1/2
	03	0.30 – 0.39	1	> 0.1;	≤ 0.2; no gap	1	10 – < 20	1	0 – < 5	1	1 – 11/2
	04	0.40 – 0.49	2	> 0.2	no gap	2	20 – < 30	2	5 – < 10	2	2 – 21/2
	05	0.50 – 0.59	3		1 gap	3	30 – < 40	3	10 – < 15	3	3 – 31/2
	06	0.60 – 0.69	4		2 gaps	4	40 – < 50	4	15 – < 20	4	4 – 41/2
	07	0.70 – 0.79	5	more than 2 gaps		5	50 – < 60	5	20 – < 25	5	5 – 51/2
	08	0.80 – 0.89				6	60 – < 70	6	25 – < 30	6	6 – 61/2
	09	0.90 – 0.99				7	70 – < 80	7	30 – < 35	7	7 – 71/2
	10	1.00 – 1.09				8	80 – < 90	8	35 – < 40	8	8 – 81/2
	11	1.10 – 1.19				9	≥ 90	9	≥ 40	9	9 – 91/2
	12	1.20 – 1.29									
	13	1.30 – 1.39									
	14	1.40 – 1.49									
	15	1.50 – 1.59									
	16	1.60 – 1.69									
	17	1.70 – 1.79									
	18	1.80 – 1.89									
	19	1.90 – 1.99									
	20	2.00 – 2.09									
	21	2.10 – 2.19									
	22	2.20 – 2.29									
	23	2.30 – 2.39									
	24	2.40 – 2.49									
	25	2.50 – 2.59									
	26	2.60 – 2.69									
	27	2.70 – 2.79									
	28	2.80 – 2.89									
	29	2.90 – 2.99									
	30	3.00 – 3.09									

[a] mmf = mineral-matter-free

A differentiation between banded (mainly humic) and nonbanded coals (sapropleic coals) has also been made. The boundary limits for the grades are defined by ash content (db) in mass %.

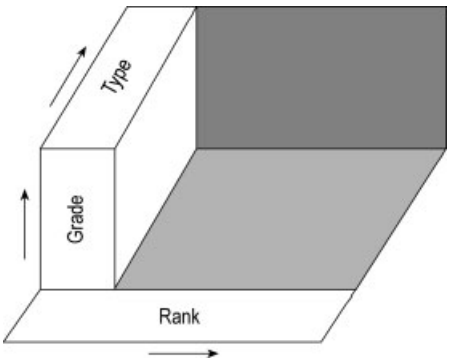

Figure 2. Schematic version of the International Classification of In-seam coal (1995)

5. Chemical Structure of Coal

The *organic components* of coal consist of a complex mixture of macromolecular carbon compounds of varied chemical constitution. The heterogeneity of coal is well-known from optical studies of thin coal sections. The microscopically distinct and distinguishable maceral groups of vitrinite, liptinite, and inertinite, their proportions, and the association of macerals of several groups with each other are essentially responsible for the physical, chemical, and technological properties of a particular coal. Preformed macerals derived from transformation of the original lipidic, woody, and waxy plant matter under varying oxidizing and reducing conditions and processes are present in *peat* and *soft lignites.* Temperature-induced chemical changes of these macerals and increased pressures influencing the physical structure of the material undergoing coalification have led to the formation

Table 6. International codification system (digits 7 – 14) for medium- and high-rank coals

	Volatile matter	mass % (daf[a])	Ash content,	mass % (db[b])	Total sulfur,	mass % (db)	Gross calorific value,	MJ/kg (daf[a])
Digit	7 ; 8		9 ; 10		11 ; 12		13 ; 14	
Code no.	48	≥ 48	00	$0 - < 1$	00	$0.0 - < 0.1$	21	< 22
	46	$46 - < 48$	01	$0 - < 2$	01	$0.1 - < 0.2$	22	$22 - < 23$
	44	$44 - < 46$	02	$2 - < 3$	02	$0.2 - < 0.3$	23	$23 - < 24$
	42	$42 - < 44$	03	$3 - < 4$	03	$0.3 - < 0.4$	24	$24 - < 25$
	40	$40 - < 42$	04	$4 - < 5$	04	$0.4 - < 0.5$	25	$25 - < 26$
	38	$38 - < 40$	05	$5 - < 6$	05	$0.5 - < 0.6$	26	$26 - < 27$
	36	$36 - < 38$	06	$6 - < 7$	06	$0.6 - < 0.7$	27	$27 - < 28$
	34	$34 - < 36$	07	$7 - < 8$	07	$0.7 - < 0.8$	28	$28 - < 29$
	32	$32 - < 34$	08	$8 - < 9$	08	$0.8 - < 0.9$	29	$29 - < 30$
	30	$30 - < 32$	09	$9 - < 10$	09	$0.9 - < 1.0$	30	$30 - < 31$
	28	$28 - < 30$	10	$10 - < 11$	10	$1.0 - < 1.1$	31	$31 - < 32$
	26	$26 - < 28$	11	$11 - < 12$	11	$1.1 - < 1.2$	32	$32 - < 33$
	24	$24 - < 26$	12	$12 - < 13$	12	$1.2 - < 1.3$	33	$33 - < 34$
	22	$22 - < 24$	13	$13 - < 14$	13	$1.3 - < 1.4$	34	$34 - < 35$
	20	$20 - < 22$	14	$14 - < 15$	14	$1.4 - < 1.5$	35	$35 - < 36$
	18	$18 - < 20$	15	$15 - < 16$	15	$1.5 - < 1.6$	36	$36 - < 37$
	16	$16 - < 18$	16	$16 - < 17$	16	$1.6 - < 1.7$	37	$37 - < 38$
	14	$14 - < 16$	17	$17 - < 18$	17	$1.7 - < 1.8$	38	$38 - < 39$
	12	$12 - < 14$	18	$18 - < 19$	18	$1.8 - < 1.9$	39	≥ 39
	10	$10 - < 12$	19	$19 - < 20$	19	$1.9 - < 2.0$		
	09	$9 - < 10$	20	$20 - < 21$	20	$2.0 - < 2.1$		
	08	$8 - < 9$			21	$2.1 - < 2.2$		
	07	$7 - < 8$			22	$2.2 - < 2.3$		
	06	$6 - < 7$			23	$2.3 - < 2.4$		
	05	$5 - < 6$			24	$2.4 - < 2.5$		
	04	$4 - < 5$			25	$2.5 - < 2.6$		
	03	$3 - < 4$			26	$2.6 - < 2.7$		
	02	$2 - < 3$			27	$2.7 - < 2.8$		
	01	$1 - < 2$			28	$2.8 - < 2.9$		

[a] daf = Dry, ash-free; [b] db = dry basis

of *lignitic, subbituminous, bituminous*, and *anthracitic* coals, forming the basis of coal classification by *rank* [11, 12].

Table 7. Example of coal coding

Basic parameters	Code no.
(a) : 1.31	13
(b) : s = 0.22; no gap; type = complex blend	2
(c) : Inertinite : 14; Liptinite : 8;	12
(d) : 7	7
(e) : 25.3	24
(f) : 6.8	06
(g) : 0.73	07
(h) : 36.4	36

Code number: 13 2 12 7 24 06 07 36

The *minerals* of coal play a considerable part in the industrial use of coals of all ranks. The mineral content, its distribution within the coal, and its composition may affect carbonization, gasification, combustion, and liquefaction processes by modifying the process of coal depolymerization and influencing hydrogenation and the thermal behavior of the resulting ash under both oxidizing and reducing conditions.

Besides microscopically unidentifiable plant ash, coals contain varying amounts of *syngenetic* and *epigenetic* minerals, which entered the organic coal substance during the first or second phase of coalification, respectively. They have either grown together with the organic substance or are physically separable. Generally, medium to high ash coals contain epigenetic minerals. Structure and hardness bear considerable influence on the distribution in the various fractions of the crushed coal from coal preparation plants. The mineral structure of coal ash is extremely difficult to determine. Low-temperature combustion in oxygen plasma produces an ash that can be analyzed by X-ray diffraction and IR absorption spectrometry. Deductions must also be made from the chemical analysis of the oxides present in the ash. The major constituents of coal ash as determined by this chemical analysis are

SiO_2, Al_2O_3, Fe_2O_3, and CaO, in addition to MgO, Na_2O, K_2O, BaO, TiO_2, P_2O_5, and SO_3.

These oxides may be derived from minerals of the following types originally present in the coal: kaolinite, illite, and bentonite as major clay minerals; calcite or siderite as major carbonates; pyrite or marcasite as the sulfides; and many other minor minerals such as apatite, hematite, rutile, quartz, etc.

5.1. Characterization of Coals

5.1.1. Standard Analytical Methods

Coals are characterized by various chemical and physical methods that have been established as national and international standards. These include proximate analysis, carbonization assay, ultimate analysis, heating value (calorific value) determination, sulfur distribution, caking and coking behavior, ash composition, and fusibility.

The above tests are carried out on representative prepared coals, using either as-received (by the laboratory) or predried samples. The results are commonly expressed on an as-received, dry (d), dry ash-free (daf), or dry mineral-matter-free (dmmf) basis.

Figure 3 shows the comparison and relationship between the proximate and ultimate analyses as well as the carbonization assay.

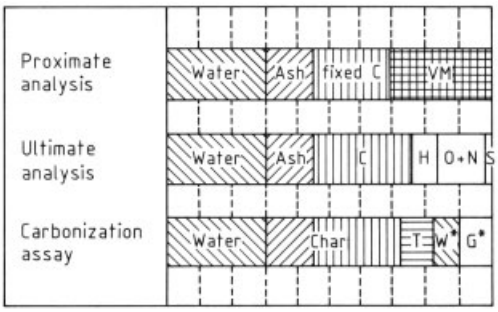

Figure 3. Comparison of proximate and ultimate analysis with carbonization assay C, H, O, N, S = elements; T = tar; W* = decomposition water; G* = carbonization gas; VM = volatile matter

In Figure 4 the sulfur forms occurring in coals are illustrated. Figure 5 shows selected typical distributions of mineral components.

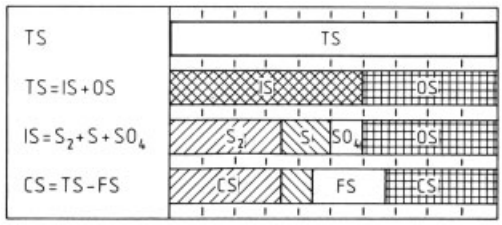

Figure 4. Scheme of sulfur distribution in coal
TS = total sulfur; IS = inorganic sulfur; OS = organic sulfur; S_2 = pyrite sulfur; S = sulfide sulfur; SO_4 = sulfate sulfur; CS = combustible sulfur; FS = fixed sulfur

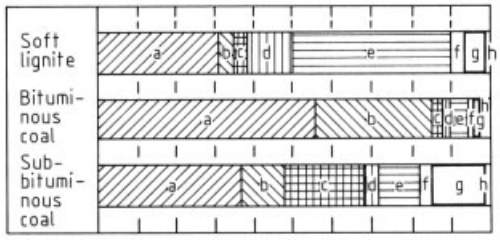

Figure 5. Distribution of mineral components of ash
a) SiO_2; b) Al_2O_3; c) Fe_2O_3; d) MgO; e) CaO; f) $Na_2O + K_2O$; g) SO_3; h) Balance

The coalification series, expressed by plotting atomic ratios of H/C vs. O/C, is graphically represented in Figure 6. As one moves from right to left, i.e., from peat to bituminous coals, the diagram shows only a slight shift to somewhat lower H/C ratios while the O/C ratios steadily decrease. At an O/C level of ca. 0.05, a steep decline in the H/C ratio occurs when the anthracitic coals are reached. Examples of actual compositions of various coals of different degrees of coalification are given in Table 8.

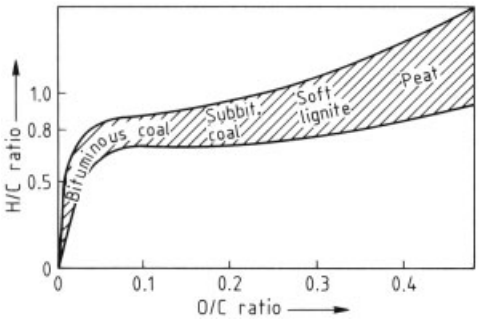

Figure 6. Coalification series: atomic ratios H/C vs. O/C

Table 8. Analytical data for coals of different degree of coalification

Analytical parameter	Peat	Soft lignite	Lignite	Subbituminous coal	Bituminous coal	Anthracitic coal
Moisture (as received), wt %	> 75	56.7	38.7	31.2	3.7	1.0
Ultimate analysis (daf)						
Carbon, wt %	58.20	70.30	71.40	73.40	82.60	92.20
Hydrogen, wt %	5.63	4.85	4.79	4.86	4.97	3.30
Nitrogen, wt %	1.94	0.74	1.34	1.16	1.55	0.15
Sulfur, wt %	0.21	0.27	0.60	0.31	1.50	0.98
Oxygen (difference), wt %	34.02	23.84	21.87	20.27	9.38	3.37
Elemental ratio						
H/C	1.15	0.82	0.80	0.79	0.72	0.43
O/C	0.44	0.25	0.23	0.21	0.09	0.03
Heating value (dry, ash-free), kJ/kg	23 500	27 500	28 500	29 400	30 600	35 700

5.1.2. Laboratory and Bench-Scale Simulation Tests

Because of the complex structure of coal and its vastly varying composition, its depolymerization, liquefaction, and combustion characteristics in thermal and extractive applications are somewhat unpredictable. In commercially operated processes, they depend very much on the prevailing conditions, such as heating and flow rates, pressure, and gas or solvent composition.

Therefore, it is often necessary to evaluate coal properties further by subjecting samples to thermal, extractive, and mechanical tests, assessing on a small scale the expected behavior and at the same time providing for study of the types and compositions of specific reaction products.

Such experiments that yield fragmental decomposition compounds can also be employed for the structural identification of coals.

The determinations include, e.g., pyrolytic or progressive devolatilization tests, extraction or solubility studies, performance of selected coking tests, reactivity measurements on specially prepared cokes, grain stability evaluation, and ash slagging behavior.

5.1.3. Petrographic Studies

The determination of the petrographic composition of coals by maceral group analysis is a valuable tool for assessing structural aspects and predicting coal behavior. Usually it also incorporates the determination of the degree of coalification by measuring the reflectance of the vitrinite as the most important and abundant group of macerals. This method allows a ranking or classification of coals by their degree of coalification and, hence, identifies the type of coal. It also reveals specific phenomena that may have induced structural coal changes within the geochemical stage of coal formation, such as aging or contact coalification due to inflowing magma.

5.2. Structural Deductions from Analytical and Bench-Scale Data

The results of coal analyses predict thermal decomposition products (proximate analysis, carbonization assay, caking and coking tests) and the element composition (ultimate analysis). Furthermore, the moisture level of run-of-mine coals is indicative of the capillary structure, allowing a first tentative classification of a particular coal. The lack of moisture holding capacity of highly coalified coals is proof of an increasingly densified structure. The element ratio, particularly that of H/C and its change to lower values as coalification progresses, points toward the aromatization of structural groups. Likewise, the dramatic decrease in oxygen within the series

peat to anthracite strongly suggests the loss of functional end groups.

The softening properties of bituminous coals (caking and coking tests) reveal the mobility of fragmented decomposition products, observed particularly in coals with high hydrogen and, hence, high vitrinite and liptinite contents. On the other hand, oxygen-rich coals, i.e., low-rank coals with an associated appreciable inertinite content or highly aromatic coals with a high degree of cross-linkage, yield recondensed solid depolymerization products. The tar and gas composition (carbonization assay, devolatilization, caking, and coking tests) shows that hydrogen preferentially enters liquid and gaseous components due to rearrangement [13].

Stable fragments, obtained from specific pyrolysis or hydropyrolysis studies, indicate the major types and the volatility of aromatic molecules present in a particular coal. Liquid-phase treatment produces larger unit fragments, which can be determined analytically. A petrographic survey permits selective composition analyses and structural identification of individual major maceral groups such as vitrinite.

5.3. Bonding of Elements in Coal

Thermal depolymerization by controlled slow or flash pyrolysis of coals, oxidation, hydrogenation, liquefaction of coals, supercritical and selective solvent extraction, as well as the application of modern physical techniques such as X-ray diffraction, IR spectroscopy, nuclear magnetic resonance, or pyrolysis mass spectrometry have furnished extensive data on the bonding of elements in coals. Major findings from these investigations are [14]:

Carbon: The aromaticity increases from 40–50 % C in subbituminous coals to 70–80 % C in bituminous coals and is over 90 % C in anthracites.

Hydrogen: Aromatic bonding to carbon and as aliphatic hydrogen in, e.g., methylene ($-CH_2-$) and methyl ($-CH_3$) groups occurs.

Oxygen: Major functional groups are hydroxyl ($-OH$), carboxyl ($-COOH$), carbonyl ($=C=O$), etheric ($-O-$) and heterocyclic oxygen.

Sulfur: Bonding is predominantly as thiophenes, also in the thiolic form ($R-SH$), decreasing with higher rank coals; up to ca. 25 % S is present as aliphatic sulfides ($R-S-R$). Heterocyclic sulfur compounds are also known to exist.

Nitrogen: Very little information on organic nitrogen bonding is available. It appears to occur in heterocycles.

5.4. Structural Evidence of Coals

Much data on the structural aspects of coal has been accumulated and published [15, 16]. The overall picture, however, is still very incomplete. The difficulty is that coal is a highly heterogeneous mixture and, as such, is not a clearly defined macromolecular aromatic compound of uniform molecular mass and structure. Thus, evaluated structural parameters refer to isolated fragments, components (vitrinite), or constituents. Nevertheless, coal may be described structurally as a three-dimensional skeleton of generally four to five highly stable condensed aromatic and hydroaromatic units with cross-linkage by weaker short-chain aliphatic groups. Reactive functional end groups and aliphatic structures are attached to the aromatic skeleton. The average molecular mass distribution is ca. 500–800 for low-rank coals, increasing to ca. 3000–6000 for bituminous coals, and possibly values in excess of 100 000 for anthracites.

Various overall models for the complex organic structure of coals, combining available experimental data on structural fragment analysis, have been tentatively proposed. However, none as yet fully accounts for all the phenomenological characteristics [14].

6. Coal Mining

6.1. General Considerations

A coal deposit is the final result of the cumulative effects of decaying vegetation, deposition of sediments, movements of the earth crust, and forces of erosion. The transformation of the vegetation in swamps into peat, lignite, anthracite, and bituminous deposits is associated with high pressure or high temperature. Movements of the earth crust (folding and faulting), igneous intrusions, and wind and water erosion have played

Table 9. Salient statistics on leading coal-producing countries and world coal production by method of mining and coal rank [a] (1983 production; all figures are in 10^6 short tons)

Country	Bituminous			Anthracite [e]	Lignite [e]	Total	Proportion
	Underground [b]	Surface [b]	Total				of total, %
Australia	54.2 [c]	88.5 [c]	133.4		38.6	172.0	3.9
China	555.5	23.2	578.7	209.4		788.1	17.8
West Germany (FRG)	98.8		98.8	6.6	136.9	242.3	5.5
India	75.1	75.1	150.2		8.1	158.3	3.6
Poland	180.7	29.9	210.6		46.8	257.5	5.8
South Africa	100.1	42.9	143.0	2.4		145.4	3.3
United Kingdom	112.3	16.2	127.1	2.9		130.1	3.0
United States	294.2	420.0	719.6	3.8	58.3	781.0	17.7
Former States of USSR	380.9	142.7	523.6	87.1	178.5	789.4	17.8
Others			204.0	77.9	669.7	951.7 [d]	21.6
Total			2889.0	390.2	1136.9	4416.1	100
Proportion of the total coal production, %			65.4	8.8	25.8	100	

[a] Compiled from *Statistics in International Coal*, 1985 Edition, National Coal Association, Washington, D.C., USA. Numbers may not add up due to independent rounding off
[b] Estimated, except for the United States
[c] Fiscal 1983 – 1984 data
[d] Estimated on the basis of world production
[e] Anthracite production is mostly from underground mining and lignite production is mostly from surface mining

important roles in determining the inclination, continuity, and quality of the seams (see Section 3.2). These have also resulted in (1) changing the position of the coal seams relative to the surface of the earth, (2) altering the condition of the roof and floor of the coal seams, (3) eroding away of entire seams or portions of seams, (4) outcropping of the surface of otherwise concealed seams, and (5) determining the economic and technical feasibility of mining these coal seams [17, 18].

The various methods of mining a coal seam can be classified as surface or underground, depending on whether the coal seam is exposed and recovered from operations on the surface of the earth or from operations in underground openings, approached from suitably located surface openings. The results of an analysis of the world coal production by mining method and rank of coal are summarized in Table 9. Bituminous and lignite seams account for over 65 and 25 % of the world production, respectively. Ten countries account for nearly 75 % of the total production. Contributions from surface and underground mining to worldwide production are approximately equal; however, by country the production distributions are quite variable. In general, most anthracite production is from underground mines, and almost all lignite production is from surface mines.

Decisions on whether a coal seam will be mined and the mining method to be practiced on a given coal seam depend on many technological, economic, and social factors. Very often, the technological factors, the most important of which are associated with geology, constrain the mining engineer to a single mining method. However, at any given time, economic and social factors dictate whether a coal seam can be mined at all.

6.2. Surface or Underground Mining

The key to determining whether to mine a coal seam by a surface or underground method is the *stripping ratio.* Distinction is often made between several stripping ratio calculations. The first is the calculation of the stripping ratio of the areas over the coal seam. It is generally defined as the number of cubic meters of the strata overlying the coal seam for every ton of coal in the seam. The basic inputs are the thicknesses of the coal seam and the strata overlying the coal seam.

$$\text{stripping ratio} = \frac{\text{cubic meters of strata material over a specified area of coal seam}}{\text{tons of coal in the same specified area of coal seam}} \quad (1)$$

The second is the calculation of a stripping ratio at which recovery by surface mining is equivalent in cost to recovery by the underground method. The basic inputs are the costs of mining that reflect the mining technology to be used.

$$\text{break-even stripping ratio} = \frac{\left\{\begin{matrix}\text{underground}\\ \text{mining cost}\\ \text{per ton of coal}\end{matrix}\right\} - \left\{\begin{matrix}\text{surface mining and}\\ \text{reclamation cost}\\ \text{per ton of coal}\end{matrix}\right\}}{\text{surface mining cost per cubic meters of strata above the coal seam}} \quad (2)$$

Only those parts of the coal seam for which the stripping ratio calculated by Equation (1) does not exceed the stripping ratio shown in Equation (2) should be mined by surface mining methods.

A third stripping ratio, also called break-even stripping ratio, incorporates the technology of mining and the market value of coal produced. It is calculated as follows:

$$\text{break-even stripping ratio} = \frac{\left\{\begin{matrix}\text{sales realiza-}\\ \text{tion per ton}\\ \text{of coal}\end{matrix}\right\} - \left\{\begin{matrix}\text{total cost of produc-}\\ \text{ing and marketing}\\ \text{per ton of coal}\end{matrix}\right\}}{\text{surface mining cost per cubic meters of strata above the coal seam}} \quad (3)$$

6.3. Coal Methods

The mining methods in coal surface mines are dictated by geological and topographical conditions, i. e., the thickness and inclination of the coal seams, the thickness of the overlying strata, and the topographical relief. The hardness – or better: cutting resistance – of the overburden and coal is of special importance for the selection of the excavation method. Hard rock formations having an uniaxial compressive strength of about more than 20 MPa must be loosened before digging, whereas soft material (loose rock and soil) can be excavated directly. The loosening of the hard rock is as a rule effected by blasting, in special cases, e.g., when blasting is not possible for environmental reasons, by ripping.

Figure 7. Rope controlled shovel excavator working in the Gilette coal mining district (Wyoming/USA)

Figure 8. Dragline working in a typical coal strip mine

Typical mining machines in hard rock mines are shovel excavators and wheel loaders. Shovel excavators are either rope controlled (Fig. 7) or operate with a hydraulic system. Shovels and wheel loaders are *discontinuously working mining machines* equipped with one bucket. Also draglines (Fig. 8) belong to this group of machines.

Soft material is usually mined by bucket wheel excavators (Fig. 9) or bucket chain excavators (Figs. 10 and 11), which belong to the *continuous mining systems* because here the bucket wheel or the bucket chain create a continuous flow of material.

Also bulldozers and scrapers find extensive application in surface mines, e.g., in scalping the surface, removing the topsoil and preparing the mining area for subsequent operations.

Figure 9. Bucket wheel excavator with a daily capacity of 240 000 m^3 working in the Renish lignite district in Germany, the machine weighs 13 000 t (Courtesy: Rheinbraun, Cologne)

Overburden removal is the most important unit operation in many mines. The equipment used varies from wheel loaders to shovel excavators, draglines, bucket wheel excavators, and bucket chain excavators. In multi-seam coal mines, where inert partings have to be removed between the split seams, usually on the same bench overburden and coal have to be excavated by the same machine. Two basic methods exist to transport the waste material to the inside dump: Cross-pit hauling and around-the-pit hauling.

Cross-Pit Hauling. Because of the short hauling distance the cross-pit method is the most economic one, but it cannot be applied everywhere. The following cross-pit equipment systems are in use:

- Draglines or shovels for direct casting (Strip mining systems)
- Bucket wheel excavators with long discharging boom
- Direct connection of excavator and spreader
- Conveying bridges

Bucket wheel excavators with a long discharging boom are applied in several mines. In Texas two systems of this type are in operation. Figure 12 shows a model of the type of machines working in the "Big Brown" lignite mines. The discharging boom has a length of 205 m and crosses an underlying dragline system and the exposed coal seam. Typical strip mining systems are usually equipped with draglines (Fig. 13) [19]. In the German lignite district south-east of Berlin giant conveyor bridges are in operation. They are equipped with three big bucket chain excavators and can remove 100×10^6 m^3 of overburden per year (Fig. 11).

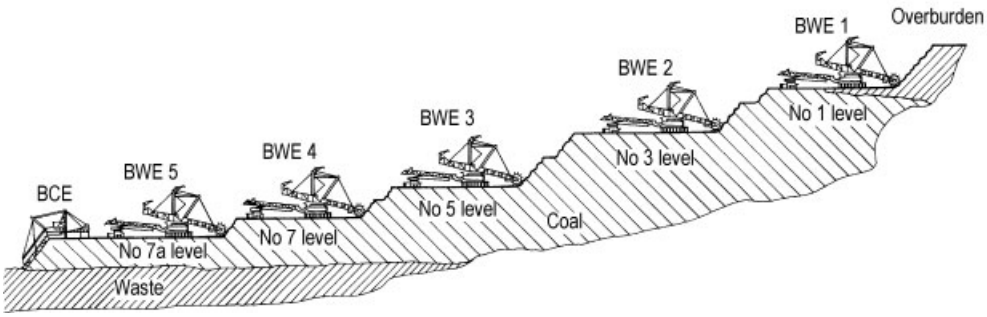

Figure 10. Cross-section of an Australian lignite mine equipped with bucket wheel excavators (BWE) and a bucket chain excavator (BCE) on the lowest bench [19]

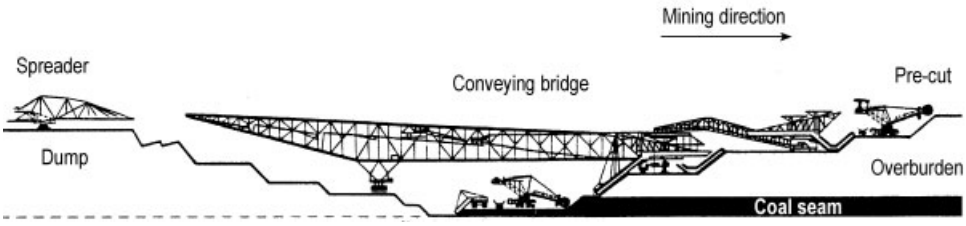

Figure 11. Cross-section of a conveying bridge of the F 60 type, equipped with three bucket chain excavators. The unit has a daily capacity of up to 500 000 m^3 and weighs 26 000 t. Several machines of this type are installed in the LAUBAG lignite mines south-east of Berlin/Germany

Figure 12. Cross-pit system consisting of two compact bucket wheel excavators and one directly connected spreader with a boom length of 205 m [20]

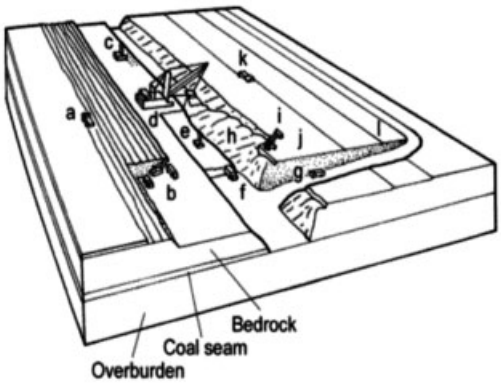

Figure 13. Typical strip mining system [19]
a) Topsoil stripping; b) Pre-stripping with a face shovel; c) Drilling and blasting of deeper overburden; d) Dragline; e) Drilling and blasting; f) Coal loading; g) Specialized tankers; h) Parallel saw-tooth spoil piles; i) In-filling of dragline spoil piles; j) Reshaping of spoil surface; k) Replacement of topsoil; l) Revegetation

Around-the-Pit Hauling. For the around-the-pit-hauling process either off-road trucks or belt conveyors are used. Railway systems, which were applied in surface mining earlier, have meanwhile been replaced nearly everywhere. Particularly in hard rock mines the blasted material is transported by *trucks* with a loading capacity of up to 300 t. For heavy trucks two driving systems are available. The mechanical system, derived from road trucks and equipped with a special multi-disk, oil-cooled brake system, and the diesel-electric system. Here the drive assembly consists of a diesel engine, a generator and an electrical wheel motor in the rear axle. These trucks have an electrical brake system.

Continuously operating *belt conveying systems* are used in many branches of industry involving bulk handling. Prerequisites for their application on the operating benches of surface mines were – besides a high performance – an easy switch of material flow and shiftability without dismantling, because the conveyors must follow the mining advance.

In many coal mines, where the seam is split, coal and overburden have to be excavated from the same bench. Therefore a flexible conveying connection is required. The common solution for this problem is a conveyor distribution yard at the flank of the mine (Fig. 14). Here, by means of telescopic drive heads the material flow can be switched, e.g., from the coal bunker to the waste dump and vice versa.

The second aspect is the track shifting of the bench conveyors. The belt has to follow the operating advance without dismantling. This operation is carried out by heavy bulldozers equipped with a side-crane. After release of the belt ten-

Figure 14. Conveyor distribution yard in a Renish lignite mine in Germany. The shiftable conveyors on the different benches (in the background) feed via link conveyors onto conveyors with telescopic driveheads. Their feeding points can be put on the required conveyor, depending on the material (Courtesy: Rheinbraun, Cologne)

sion the conveyor becomes flexible and can be pulled sideways using a rail with a fixed roller set located at the crane head. For loading the blasted material belt conveyors are applied – instead of or besides trucks – also in hard rock mines, where shovel excavators or wheel loaders are used. But contrary to trucks, belt conveyors cannot take the blasted rock as it is. Before it can be loaded onto the conveyor, it has to be crushed to a grain size of about 300 mm. The crushers are either directly connected with the mining machine or they are – as semi-mobile crushers – located at a central point of the mine. In this case trucks are used for the first section of hauling distance. This hauling concept is advantageous in deep open-pit mines with a high lift, because belt conveyors have, contrary to trucks, no lifting deadweight.

In multi-seam coal mines selective operation is necessary in order to separate coal from inert intercalations. Suitable machines for selective operation are bucket wheel excavators or backhoe shovels. But if a very accurate cut at the interface of the different layers is required, also continuous surface miners are used. This type of machines is derived from surface cutting machines applied for road repair works and operates successfully in several coal and limestone mines all over the world. Because of their high cutting force ground preparation by blasting can be dispensed with.

Reclamation (→ Soil). The reclamation of land used for surface mining operations has become extremely important and is usually carried out very carefully. In the large German lignite mines where the overburden dumps are operated with conveyor belts and spreaders (Fig. 15) these machines cover the dump surface with fertile soil (loess). For *agricultural reclamation* the surface is leveled, a road network combined with a sur-

face drainage system is constructed, and subsequently, for five to seven years, farmers employed by the mine operator cultivate the farmland until the full crop is achieved and the land can be given back to farmers. For *forest reclamation* either loess or – particularly in slope areas for stability reasons – a mixture of loess and quaternary gravel is used as topsoil.

Figure 15. Spreader in a Rheinish lignite mine. The top material, which is laid-on by the machine, consists of fertile soil (loess) for land reclamation [21]

6.4. Underground Coal Mining Methods

Underground coal mining is a complex operation where the working environment of the miners and equipment is completely enclosed by the geological medium overlying and underlying the coal seam. Therefore, the correct assessment of the geological factors is more critical for an underground mine than for a surface mine. The depth to the coal seam, its thickness and quality, the conditions of roof and floor strata, and the problems associated with methane and water inflows must be analyzed in detail to select the appropriate applicable methods and equipment.

An underground mine can be classified as a shaft mine, a drift mine, or a slope mine, depending on the type of access from the surface to the coal seam, as shown in Figure 16. Where applicable, drift is the obvious choice for access to a coal seam. In other cases, the choice between slope and shaft for deep mines can be difficult. A mine with large daily production and a long life would benefit from a slope opening for hauling coal.

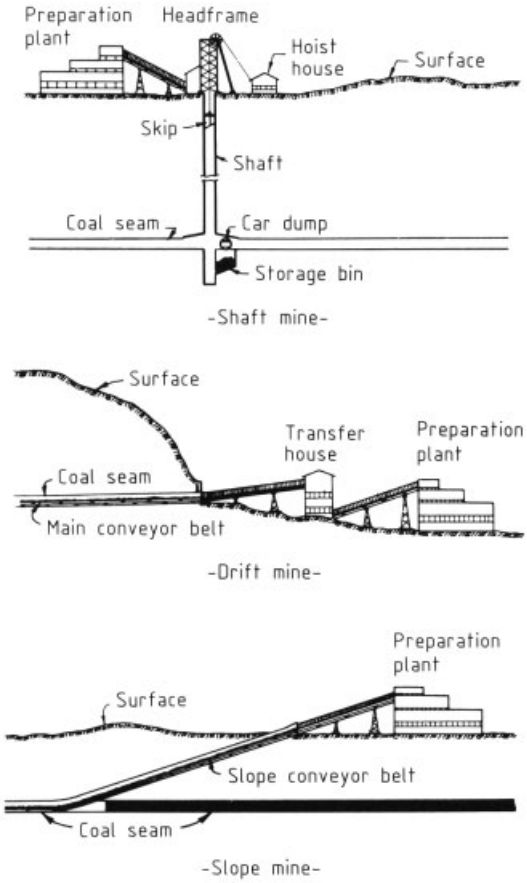

Figure 16. Types of access to a coal seam
Deep coal seams can be accessed only by sinking vertical shafts

Two distinct activities in underground coal mining are the unit operations in the production and hauling of coal and the auxiliary operations that are essential to maintaining a safe productive operation. Specifically, the unit operations include cutting, drilling, blasting, loading, and hauling. Auxiliary operations include ventilation, drainage, power, communications, and lighting. Roof support operations in the face area (active face where coal is being mined) can be considered as a unit operation. Support of other areas in the mine can be included in the auxiliary operations. The roof support used in most coal mines is a 1.5 – 2 m long, 15 – 25 mm diameter

steel bolt, usually set in rows such that the bolts are 1.2 – 1.8 m apart in a row, with the rows 1.2 – 1.8 m apart. Additional supports may be used as required.

There are several theories which explain how rock bolts hold the roof. These include the beam theory (roof bolts tie together several weak roof strata into one strong beam), the suspension theory (the weak members are suspended from a strong anchor horizon), and the keying effect theory (the roof bolts act much like the keystone in an arch).

There are two major systems of underground coal mining: the room and pillar and the longwall. There are many variations in each of the systems. Several types of equipment – cutting machines, drills, loaders, roof bolters, shuttle cars, belts, etc. – are used in the mining process, depending on the method chosen to extract a coal seam; the choice is determined by local geological conditions and operating practices.

Two main types of room and pillar systems are the conventional and continuous mining systems, shown in Figure 17. In the former, the operations of cutting, drilling, blasting, and loading are performed by separate machine units. In the latter, a continuous miner rips the coal from a coal seam and loads it directly into a hauling unit, all in one step. In either system, the objective is to create blocks of coal, usually rectangular or square, during the first stage of the mine,

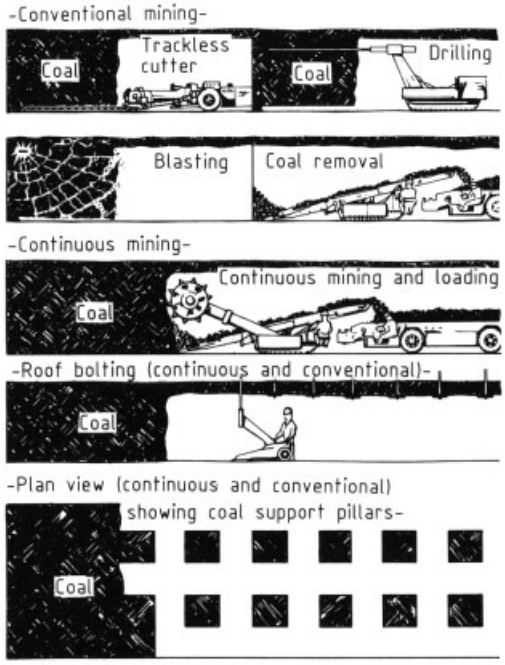

Figure 17. Conventional and continuous room and pillar underground mining systems

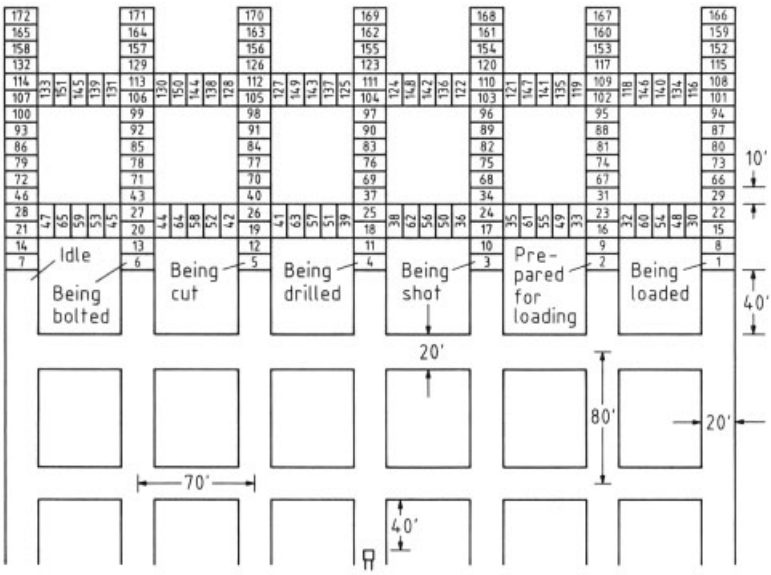

Figure 18. Development plan for a conventional room and pillar mining section (lengths in feet; 1′ = 0.305 m)

Figure 19. The milling-head continuous miner
The coal is cut by the picks mounted on a number of oscillating and rotating disks. The cut coal is loaded onto a chain conveyor in the machine by a gathering arm in the loading head

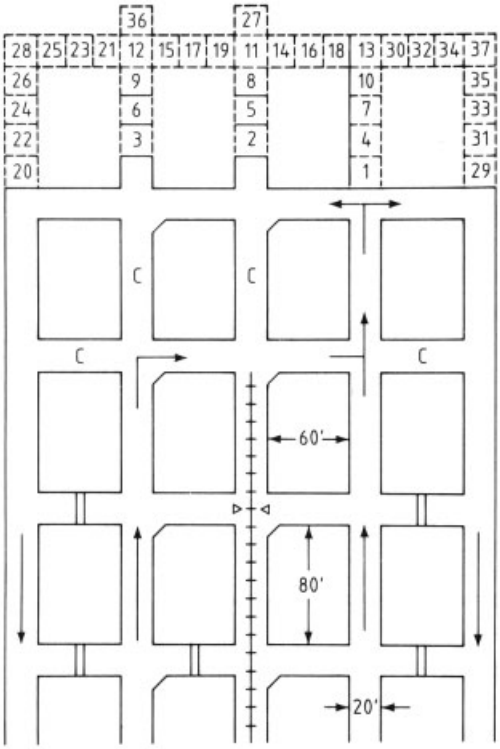

Figure 20. Development plan for a five-entry continuous room and pillar mining section (lengths in feet; 1′ = 0.305 m)

called "development." These blocks of coal may be left in place. However, they may also be recovered during the second stage of mining, usually called "pillaring." Where the surface features (such as houses, roads, streams, etc.) have to be protected, pillars underneath these features are not recovered. In the United States, the average recovery from a coal seam, when room and pillar methods are used, is ca. 50 %. The sequence of operations for the various machines in a conventional mining development plan is shown in Figure 18. After finishing its operation in a working area (or face), a machine moves to the next higher numbered working area. Because of the large number of machines, there is a need to have a large number of working faces to avoid excessive delays.

Continuous mining incorporates the unit operations of cutting, drilling, blasting, and loading into one machine: the continuous miner. There is no need for a large number of working faces. While there are many types of continuous miners, the milling-head miner, shown in Figure 19, is by far the most popular machine in operation. The sequence of operations for the continuous miner in a five-entry development is shown in Figure 20. The only other machine unit that cycles with the miner is the roof bolter. The ventilation plan has been laid out such that the miner and the roof bolter operations can have their own split of ventilating air current. To a large extent, continuous mining has replaced conventional mining. However, in thick and hard coal seams, conventional mining has great potential for increased production.

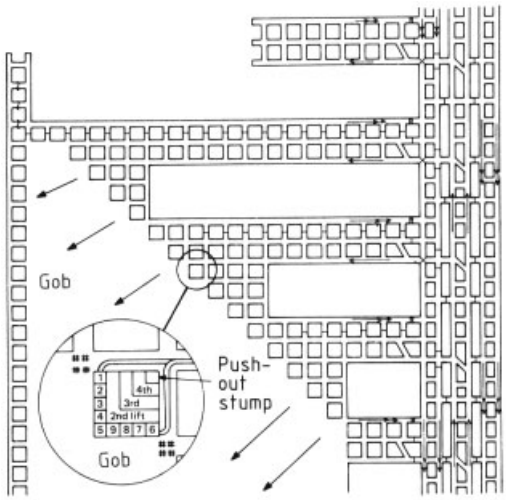

Figure 21. Open-end method of pillar extraction [22]
The pillar extraction must be sequenced to maintain the 45° lineThe methods that are practiced to extract the coal left in the pillars depend greatly on the na-

ture of roof and floor and the manner in which the increased pressure generated due to pillaring is transferred to other pillars and the gob. The necessary evaluations are specific to each site and must ensure that sudden bumps and coal eruptions do not result. In addition, the unrecovered pillars should not be crushed to a point where loss of life, equipment, or coal is possible. Two methods are shown in Figures 21 and 22.

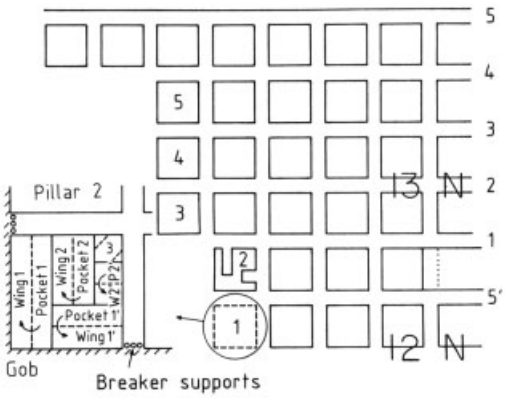

Figure 22. Pocket and wing method of Pittsburgh block room and pillar slicing system [22]

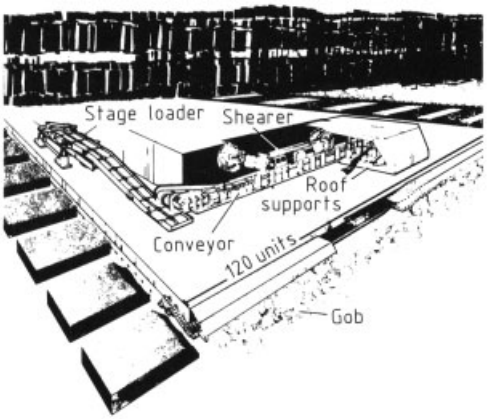

Figure 23. Perspective and close-up view of longwall system showing the equipment and its arrangement
Note that this is a retreating longwall system

In longwall mining systems, a block of coal 200 – 300 m wide and 2000 – 3000 m long is extracted by a machine that runs up and down the width of the face and takes a 0.75 – 1 m slice of the block as shown in Figure 23. A modern mechanized longwall system is a true continuous coal production system. It offers great productivity and safety. The system is most common in Europe; however, it is gaining popularity in the United States and elsewhere. There are essentially four machines: the cutting machine (usually a shearer or a plough), the armored face conveyor, the line of roof supports, and the stage loader. The cutting machine rides on the conveyor, cutting and transferring coal to it (Figs. 24 and 25). Both of these machines and all of the men in the face are under a protective canopy of steel, formed by a continuous line of roof supports. In modern longwall systems, shields are most common for supporting the roof over the immediate face area.

Figure 24. Plough longwall face in a German coal mine

Figure 25. Shearer longwall face in an American coal mine

There are three main types of longwall system layout. The retreating longwall, shown in Figure 26, is most common in the United States. The advancing longwall, shown in Figure 27, is most common in Europe. The third method, alternate retreat and advance, is not in common practice.

A method combining the longwall layout with continuous miners is known as the shortwall method and is shown in Figure 28. It is also not in common practice, but where conditions are favorable, this method may have great potential.

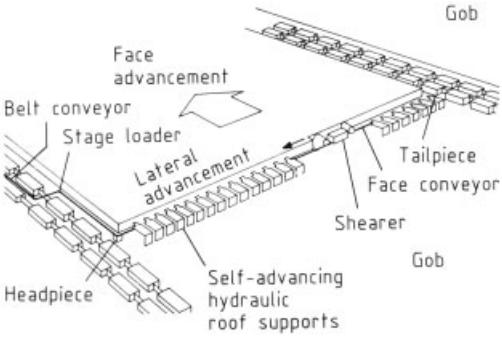

Figure 26. Equipment arrangement in a retreating longwall face [22]

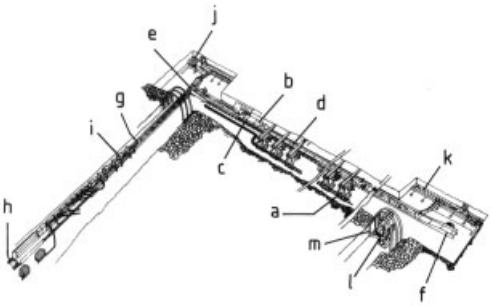

Figure 27. Equipment arrangement in an advancing longwall face [22].
a) Powered roof supports; b) Shearer mining machine; c) Cable-handling equipment; d) Face chain conveyor; e) Chain conveyor head drive; f) Face chain conveyor tail drive; g) Stage loader; h) Belt conveyor; i) Hydraulic power pack for face supports; j) Head gate stable driving equipment; k) Tail gate stable driving equipment; l) Ripping machine; m) Yieldable arches

Mention has already been made about thick and inclined coal seams. Thick coal seams are a special gift of nature. However, thick and steeply pitching seams present challenges to the ingenuity of miners to devise safe and efficient methods for their extraction. A generalized classification of thick-seam mining methods is shown in Table 10.

The cutting action of a water jet has a long history (at least 40 years) of successful application in the Former States of theSoviet Union and other European countries for the extraction of coal. Some important requirements for the application of such hydraulic mining underground are that the coal should be soft or have friable bands to aid in the breaking and that the roof should be competent and part easily from the coal. This allows the maximum extraction of coal before the roof collapses and also minimizes dilution. The floor must resist degradation by water. Where the coal is hard, blasting can aid subsequent water jetting. Under proper conditions, hydraulic mining is very efficient, safe, and productive.

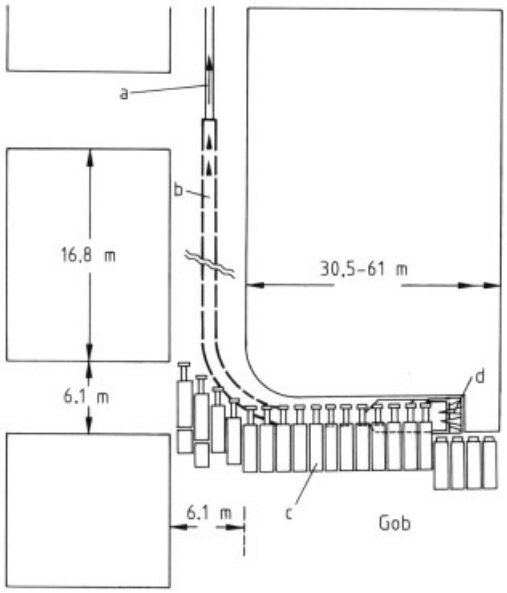

Figure 28. Shortwall mining method
Shortwall is generally 40 – 60 m wide. The cut width is usually 3 m. The roof support equipment (chocks) must be specially designed for the longer cantilever support and is more costly. a) Panel belt conveyor; b) Flexible conveyor train; c) Chock; d) Continuous miner

7. Coal Preparation

The various coal conversion processes require the use of different coal types and qualities. Coal preparation can particularly affect water, ash and sulfur contents, and grain size. Preparation is intended to turn the raw coal into more manageable products with defined properties. Coal preparation includes homogenization, breaking, crushing and screening; cleaning (reduction of sulfur and ash contents); dewatering; and blending. The choice of preparation processes and the

Table 10. A generalized classification of thick-seam mining methods

Method	Variations	Operational dip constraint	Thickness constraint	Ground control	Haulage constraint	Practicing countries
Full face	longwall	0 – 15°	5 m	caving stowing	conveyors	Germany (FRG) Poland
	conventional	depends on haulage	7 m	caving stowing	shuttle cars < 12° conveyors < 15°	United States India
Slicing	longwall	0 – 15°	3 m/pass (max. 14 m)	caving stowing	chain conveyors	France Poland
	continuous	depends on haulage	3 m/pass (max. 7 m)	caving	shuttle cars < 12° conveyors < 15°	United States
	conventional	depends on haulage	3 m/pass (max. 7 m)	caving stowing	chain conveyors	India
Caving	longwall	0 – 15°	< 14 m	caving	chain conveyors	France
	conventional	dip is used	< 7 m	caving	none	Former States of USSR
	hydraulic	> 2.5°	penetration up to 20 m	caving	fluming > 2.5°	Canada Former States of USSR

degree of preparation depend (1) on raw coal properties and the demands of the technological processes that follow, (2) the quality requirements on the end products, and (3) on economic and environmental considerations. Coal value but also its selling price increase with the degree of preparation.

About 80 % of hard coal and nearly all of the brown coal presently consumed are used for heat and power generation. Pulverized-coal furnaces are almost exclusively used in large power stations. In this case, the coal is usually mill dried to a water content of 1 – 2 wt % for hard coal and of 15 – 20 wt % for brown coal directly before entering the boilers.

7.1. Hard Coal Preparation

Hard coal preparation is part of the colliery operation. The ROM (run of mine) coal brought to the surface contains various types of accompanying minerals and interstratifications, which must be removed to obtain a coal that complies with market demands for the different types and grades. Grain size, ash content, and moisture content can be controlled within a narrow range by applying mechanical and physical cleaning methods. In general, run-of-mine (ROM) coal is only prepared by size reduction and/or screening. Depending on the use, grain size will vary from micrometer to centimeter size. Pulverized coal can be slurried with water for transportation in pipelines (Chap. 10) [23 – 25]. With water or oil it is slurried in order to replace fuel oil in power stations and industrial boilers and for some gasification processes (e.g., Texaco process, Section 9.4; → Gas Production).

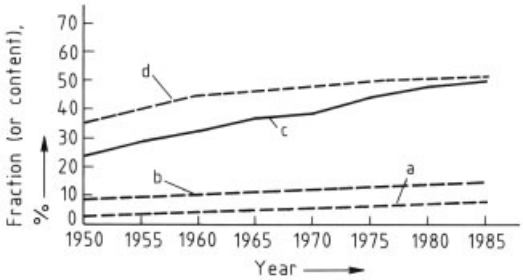

Figure 29. Change in the characteristics of ROM coal
a) Moisture content in particles < 6 mm; b) Fraction of particles < 0.5 mm; c) Refuse content of the ROM coal; d) Fraction of particles < 10 mm

Only about one-third of the hard coal undergoes preparation including cleaning and dewatering. Particularly coal intended for coking process requires extensive preparation in view of the high quality demands on metallurgical coke. Since ash and water are roughage, their contents are reduced for economic reasons. Sulfur has to be removed because it negatively influences the utilization of coke in metallurgical processes. Simultaneously, the emission of sulfur dioxide can be reduced in the primary stages, if coal/coke is used as fuel.

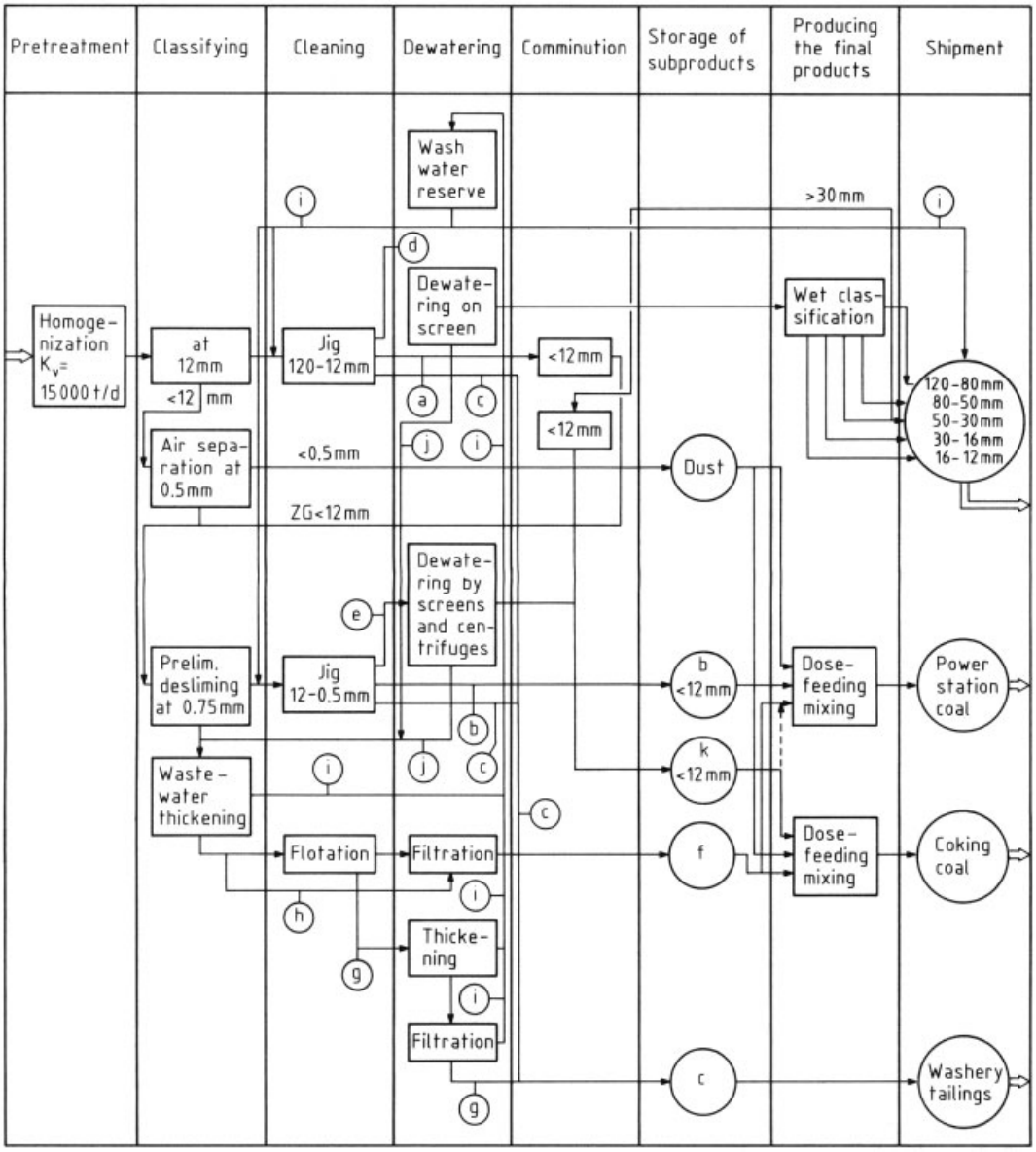

Figure 30. Basic flow sheet of a modern coal preparation plant
K_v Capacity of homogenization; a) Intermediate products; b) Middlings; c) Refuse; d) Cleaned coarse; e) Cleaned smalls; f) Froth fines; g) Tailings; h) Filtrate; i) Clarified water; j) Dirty water; k) Cleaned coal

Figure 29 shows that the characteristics of raw coals have increasingly deteriorated due to the mechanization of coal retrieval and to the measures adopted to suppress dust. Large mining complexes concentrating on few face areas produce coal with characteristics that fluctuate within a wide range. A raw coal homogenization step can minimize investment and operational costs for new or extended washery buildings needed to handle such diverse coals. It can also improve the precision of separation and, thus, the yield of saleable product and provide a more consistent product.

The sequence of hard coal cleaning procedures is depicted in the flow sheet of a modern washery shown in Figure 30.

7.1.1. Preliminary Treatment

The upper grain size of raw coal is defined by the dimensions of the coal cleaning equipment. Before being fed to that equipment, the coal is subjected to preliminary screening and crushing to a diameter of 120 – 150 mm; foreign matter is simultaneously removed. Coarse dirt is normally reduced in particle size along with the coal, and only in exceptional cases is it removed by preliminary treatment. Such preliminary removal of coarse dirt is done exclusively in drum or inclined separators. Dual eccentric and resonance screens are used for lump coal; each of these machines has a throughput of up to 2×10^3 t/h.

After preliminary screening and removal of foreign matter, the coal in the 0 – 150 (or 0 – 120) mm diameter range is defined as raw feed coal. To an increasing extent, this coal is fed to a storage and homogenization plant without further treatment. Homogenization then produces a consistent grain size and uniform contents of moisture, dirt, and volatile materials. The uniform quality of the coal feed makes possible a uniform product and optimal utilization of the washery capacity [26].

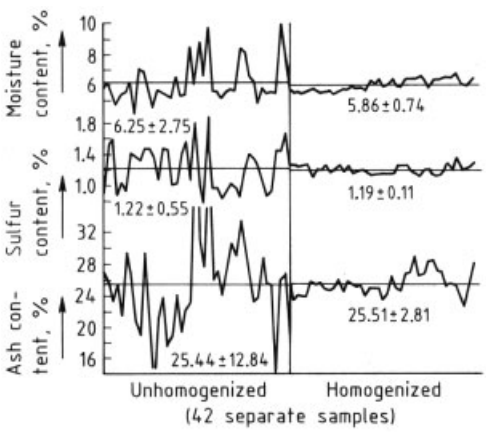

Figure 31. Homogenization of raw coal in a blending silo

Homogenization of the raw feed coal can be carried out in blending silos or on blending yards. Figure 31 shows the homogenization results of a silo. A critical factor in homogenization is the size of the storage location for containing one day's output.

In the washery the coal is first classified to ensure optimal utilization of the subsequent cleaning equipment for coarse coal and smalls. Normally the raw feed coal is screened off between 10 and 12 mm and subdivided into coarse raw coal and raw smalls. Such classification is carried out either on vibrating screens or on tensioned screens equipped with varying types of bottoms of perforated plate or tissue. The throughput for a unit may be as high as 1×10^3 t/h [27]. The raw coarse fraction is then subjected to wet mechanical cleaning, which completely removes adhering smalls and ultrafines by spraying.

Prior to cleaning of the raw smalls 0 – 12 (or 0 – 10) mm, the ultrafines are removed by a complex method. Normally an air separation method, e.g., dry classifying, or wet desliming on screens is used. Among the air separators, cyclone separators with mechanical dust discharge have proven to be most successful. These are designed for throughputs of up to 700 t/h. Vibrating separators are used in some exceptional cases where dust must be removed from all of the raw feed 0 – 150 (or 0 – 120) mm [28]. With some rare exceptions, the separator dust is added unscreened to the saleable products; in the future, however, an increasing proportion of the dust will need to be cleaned.

7.1.2. Wet Treatment

Cleaning of the raw feed is done by wet treatment in jigs and dense-medium separators, which separate according to differences in density or flotation (surface) properties.

7.1.2.1. Cleaning in Jigs

Almost all of the jigs used in Germany are of the air-pulse type or, in exceptional cases, of the wash-box type [29, 30]. They usually yield three products. Separation is accomplished by density, under the action of a pulsating flow of water. Pulsation may be applied either by a simple stroke or a superimposed double stroke.

The Batac jig with compressed air chamber underneath the bed proved to be the most successful machine. With such a system, all of the machine width, including the chamber width, is available for jigging. An electronic valve control

or a mechanical rotational valve is used to adjust the compressed-air load. Air pulsation coupled to the hydraulic inductive discharge controls enables the operator to modify the cut point by simply actuating a selector. This allows the cleaning of two different coal types successively in the same jig. The jigs usually accept coarse coal 10 – 150 (or 10 – 120) mm in diameter and smalls < 10 mm.

7.1.2.2. Cleaning in Dense-Media Separators

The coarse fraction is also cleaned by dense media, particularly in star-wheel extractors or inclined separators [31].

Whereas previously baryte, clay, ultrafine dirt, loess, and pyrite were used as dense media, the more recent equipment runs almost exclusively on magnetite and ilmenite. Selection of the dense medium depends on the method of regeneration (magnetic separation or gravity). The regeneration method also depends on the grain size of the dense medium.

A jig can be controlled by using either floaters or a layer of a given weight. It can also be fully automatic, controlled by gamma-ray counters. Dense-media equipment is controlled by maintaining the densities constant via automatic addition of the dense liquid. This adjustment can be set so precisely that it approximates separation by a true solution. While the cut point with dense-media cleaning is, in general, more precise than that with jigs, the latest achievements in jig cleaning have resulted in the widely expanded use of jigs within the Federal Republic of Germany because the coal normally is highly amenable to jig cleaning and because machine expenditure and costs are lower.

7.1.2.3. Cleaning by Flotation

Some of the raw coal feed in Germany is subjected to flotation. The diverging degrees of wettability of different mineral components allow cleaning of very fine particles by means of air bubbles. The bubbles are created by dispersion of the air feed in the dense medium; they remove the coal particles that have been rendered hydrophobic by flotation agents [32]. Flotation is applied to slurries if they contain an excessive percentage of mineral components so that their inclusion in the high-grade product is economically impossible.

Flotation plants consist of five to seven separate cells or flotation troughs with the corresponding number of stirrers. Each cell is between 5 and 14 m^3 and has a throughput up to 100 t/h particulate matter corresponding to 1000 m^3/h of pulp. Flotation is done either in one or two steps, with a flotation agent is of the collecting and foaming type. A ton of particulate matter requires 0.3 – 0.8 kg of flotation agent for the process.

7.1.2.4. Other Cleaning Methods

Cleaning on tables and in dense-media cyclones has been used in certain cases to reduce the sulfur content of the product and to obtain a coal extremely low in ash.

7.1.3. Dewatering

Moisture must be removed after wet cleaning [32 – 35]. There are static and dynamic methods. Among the dynamic methods are screen dewatering of nut-size coal and preliminary dewatering of small products on stationary screening tables or in refuse elevators. The static methods are being replaced by dynamic ones (centrifuge, vacuum, and pressurized dewatering) because of the demand for low final moisture content. The concurrent higher ultrafine and moisture contents in the raw coal have necessitated additional process steps.

Smalls are normally dewatered on vibrating screen centrifuges. Several improvements in their design – particularly the conversion from vertical to horizontal arrangements – have allowed high throughputs and easy repair and maintenance. Centrifuges with drum diameters of 1.3 m and rated capacities of 250 t/h (water-free) are the best. The smalls are dewatered down to 5 or 7 % at centrifugal forces of 80 – 100 g, depending on grain size distribution.

Coarse slurries are dewatered in pusher screens where the centrifuge forces (400 – 500 g) are higher than in vibrating screen centrifuges and where residences are considerably longer. Figure 32 shows a pusher screen. A first

(comparatively short) step provides preliminary dewatering and an even distribution of particles on the drum periphery. This is followed by the secondary dewatering step. Recent pusher screen designs have drum diameters up to 1.2 m with a rated throughput of coarse slurrying of more than 40 t/h. The coarse slurry is dewatered to 9 – 15 %, depending on grain size distribution.

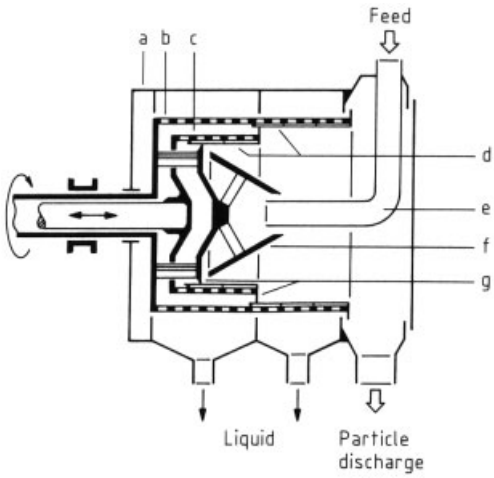

Figure 32. Two-step pusher screen
a) Enclosure; b) External screen drum; c) Internal screen drum; d) Slot screens; e) Product feed; f) Vortex; g) Pusher rings

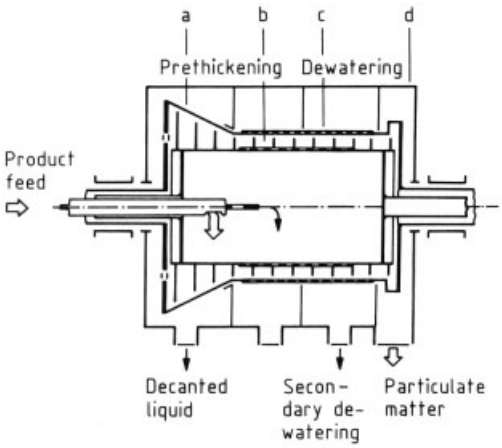

Figure 33. Solid-bowl centrifuge
a) Solid bowl; b) Conveyor screw; c) Screen bowl; d) Enclosure

Drum filters with surface areas up to 120 m^2 and disk filters of up to 500 m^2 are available for ultrafine dewatering. With this equipment, specific feed rates of 100 – 800 kg m^{-2} h^{-1} (wf) and moisture contents of 18 – 28 % are attainable.

The flotation concentrate is also dewatered by means of a solid-bowl screen centrifuge where a solid-bowl section is followed by a cylindrical screen section, which dewaters the prethickened feed under favorable conditions independent of the particle concentration in the feed medium. Such a solid-bowl centrifuge is depicted in Figure 33. Flotation tailings are dewatered either by chamber filter presses or by solid-bowl centrifuges. More recently, screen belt presses have also been used. The selection of these alternatives depends on the desired moisture content of the waste.

7.1.4. Decantation and Thickening of the Process Water

Wet-type coal preparation plants involve a number of particle – water circuits where up to 6000 m^3/h medium are circulated. If the water is to be reused for further cleaning steps, the particulate matter must be removed from it as far as possible. Decantation units, which usually consist of circular thickeners with diameters up to 40 m, are used. Rectangular thickeners are used less often. Water circulation varies, depending on the mechanical equipment, and can be as high as 6 m^3 per ton of raw feed coal. Water consumption oscillates near 0.2 m^3/t feed for a closed circuit. It can rise to as high as 1 m^3/t feed if, e.g., salt content is excessive or if the flotation waste must be discharged to an external disposal site.

The decantation units are meant to bring about simultaneously a desirable degree of thickening for subsequent dewatering or cleaning procedures. Thickener outputs fluctuate between final concentrations of 120 g/L for flotation feed and 650 g/L for pusher screen dewatering. About 0.5 g of organic sedimentation accelerators are added per cubic meter to promote sedimentation for wash-water decantation.

7.1.5. Dosing and Blending

The products from coal preparation (nuts, dewatered smalls, air separation dust, flotation concentrate, and middlings, i.e., product that has a

high content of intergrown material) are stored separately in bunkers. The demand for nut-size particles has been declining, and they are reduced to < 10 mm diameter by impact or hammer mills. They then serve as a constituent of high-grade coals. The cleaned products are withdrawn from the bunkers and blended according to a set program. During this operation, the ash contents can be monitored or adjusted by automatic rapid-measuring instruments. The addition of those components that are highest in ash (air separation dust) is adjusted, depending on the measurements, so that the final product is brought to the desired ash content. Instruments for rapid and continuous measurement of moisture content will soon be introduced. Other mechanical equipment used for dose-feeding and blending includes filling-level indicators, dose-feeding devices, and belt weighers.

7.1.6. Removal of Pyritic Sulfur

Sulfur may be present in hard coal as elementary sulfur, sulfate, organic sulfur, or sulfide (pyrite).

Pyritic sulfur is removed by modern cleaning procedures [36, 37]. The raw coal is screened off at ca. 40 mm. The dirt of the > 40 mm product is removed in a jig. The jig floatings are reduced to < 40 mm, fed along with the < 40 mm product from sizing to a blending yard, and homogenized. The product is then cleaned in another jig; the coal and intergrowth of this step are reduced to a diameter of < 10 mm and fed to dense-media cyclones to yield, at low separation density, a raw coal low in sulfur. The intergrowth from the cyclone is reduced to a diameter < 3 mm and then classified in cyclones at 0.063 mm and subjected to secondary cleaning on tables. The < 0.063 mm product is then subjected to flotation. Dewatering of the intermediate products is done in vibrating screen centrifuges, pusher screens, and solid-bowl screen centrifuges.

7.1.7. Thermal Drying

Thermal drying is restricted to a few special cases in the treatment of coking and power station coals because the procedure involves high costs. The continuing increase of the ultrafine proportion, which is difficult to dewater, and the increasingly stringent quality demands necessitate more advanced dissociation and cleaning of ultrafines, as well as novel techniques of coal extraction, haulage, and treatment. These factors have led to increasing reliance on thermal drying in the aforementioned sectors [37].

Drum-type dryers are usually superior to the gas suspension dryers in terms of energy balance. The quality requirements for coals for coking and power stations require that only part of the coal volumes be dried to stay within the moisture tolerances required for the final smalls product. Less drying also reduces problems of dust.

7.2. Brown Coal Preparation

Brown coal is generally delivered without any preparation from the coal mines. Due to transport costs (high water content up to 60 wt %), the coal users have to be relatively close to the mines. The preparation of brown coal includes breaking, crushing, and screening as well as thermal drying. For decentralized supply of energy, fluidized-bed coal is produced for low emission fluidized-bed combustion (coal with a water content of about 16 wt %, grain size < 4 mm). Pulverized coal (water content about 11 wt %, grain size < 0.3 mm) is also applied in the cement and lime industries. Predried and fine-grained coal is used for briquetting, fine coke production, and montan wax extraction.

8. Agglomeration

Numerous industrial processes require coal in its coarse state for handling and transportation and as a feed for such thermal applications as combustion, gasification, iron ore reduction, and carbonization. However, highly mechanized mining operations, coal preparation, and beneficiation yield a high proportion of fines. Upgrading of such coal fines by agglomeration is frequently applied and considered [38 – 43].

The choice of agglomeration process and of a possible binder depends very much on the degree of coalification of the coal and the intended application.

The agglomeration ability of a coal is mainly influenced by its moisture-holding capacity, ash

content and composition, hydrophobicity, size distribution, and plastic or elastic behavior.

Fundamental coal agglomeration methods are grouped into

1) briquetting and extrusion [39, 43 – 47];
2) balling and tumbling [39, 41, 42], [48 – 51].

Coal briquetting is intended to turn fine-grained coal (< 6 mm) into a lump form. This conversion leads to a better manageable solid fuel especially for domestic heating. Briquetting is also necessary if coal is reacted in a fixed bed subjected to a gas flow (e.g., in fixed-bed gasification) and for lump or formed coke production from noncaking coal.

Regarding the temperature range, briquetting processes are subdivided into *cold briquetting* (< 100 °C) and *hot briquetting* processes (400 – 500 °C). The production of high quality briquettes by cold briquetting without binding agent is only possible with soft brown coal. Binding agents such as pitch, tar, and bitumen were formerly used for low volatile coals. Because of their carcinogenic effect (which is particularly pronounced if such binding agents are based on hard coal), they are being replaced by other binding agents, e.g., biomass materials (for example molasses). Hot briquetting is a less common method of compacting. Here, a caking coal in its softening range is used as the binding agent. Low sulfur emission briquettes can be produced by adding sulfur binding components such as milk of lime to reduce the sulfur dioxide emission during low- and medium-temperature combustion.

Balling of finely ground prepared coal generally smaller than 0.315 mm in particle diameter with 2 – 6 wt % of organic or inorganic binders has only recently evolved as a commercially viable coal agglomeration technique by using pelletizing disks designed for light bulk materials [41, 51, 52]. Both moist and dry pellets may be used industrially. Although the disk produces pellets of almost uniform size, tumbling in, e.g., rotary drums or mixers yields pellets of a wide size distribution. This mode of agglomeration is preferentially applied to subbituminous, bituminous, and anthracitic coals of both low and high ash contents, but predried lignites can also be pelletized [39, 48, 49]. Drying and thermal treatment of these pellets, however, cause severe cracking and significant losses in strength and stability.

Coal agglomerates must withstand mechanical and thermal treatment when used in industry. The type and severity of stress depend on the particular use and mode of treatment which mainly include transportation, drying, thermal shock, pyrolysis, gasification, and combustion [43, 46, 48]. Suitable procedures for the agglomeration step may be developed and selected from bench-scale simulation studies. Subsequent handling and thermal experiments with these agglomerates lead to a final, optimized product, which may be tested on a larger scale in pilot or commercial plants.

9. Coal Conversion (Uses)

The aim of coal conversion processes is the improvement of the fuel's utility value (e.g., production of briquettes and coke) and/or the transformation into other energy carriers or energy forms (e.g., electric power generation, gasoline production), or the production of products for nonenergetic use (e.g., synthesis gas, activated carbon).

Coal refining may be divided into three stages:

1) Mechanical conversion processes (e.g., crushing, screening, separating) and mechanical-thermal processes (e.g., drying, briquetting)
2) Thermochemical conversion processes (e.g., combustion, coking, gasification, direct liquefaction)
3) Carbochemical processes (e.g., processing of liquid coal reaction products) and gas treatment (e.g., cleaning, conversion)

The most important refining routes, the individual conversion processes, and some of the intermediates and end products are shown in Figure 34. The bold lines depict the most important routes worldwide. The thinner lines represent common routes with more local importance, and the broken lines show those that have no commercial applications at present.

By *thermochemical coal refining* coal can be converted into gaseous, liquid, and solid products and its energy content possibly into sensible heat. Product distribution and composition vary

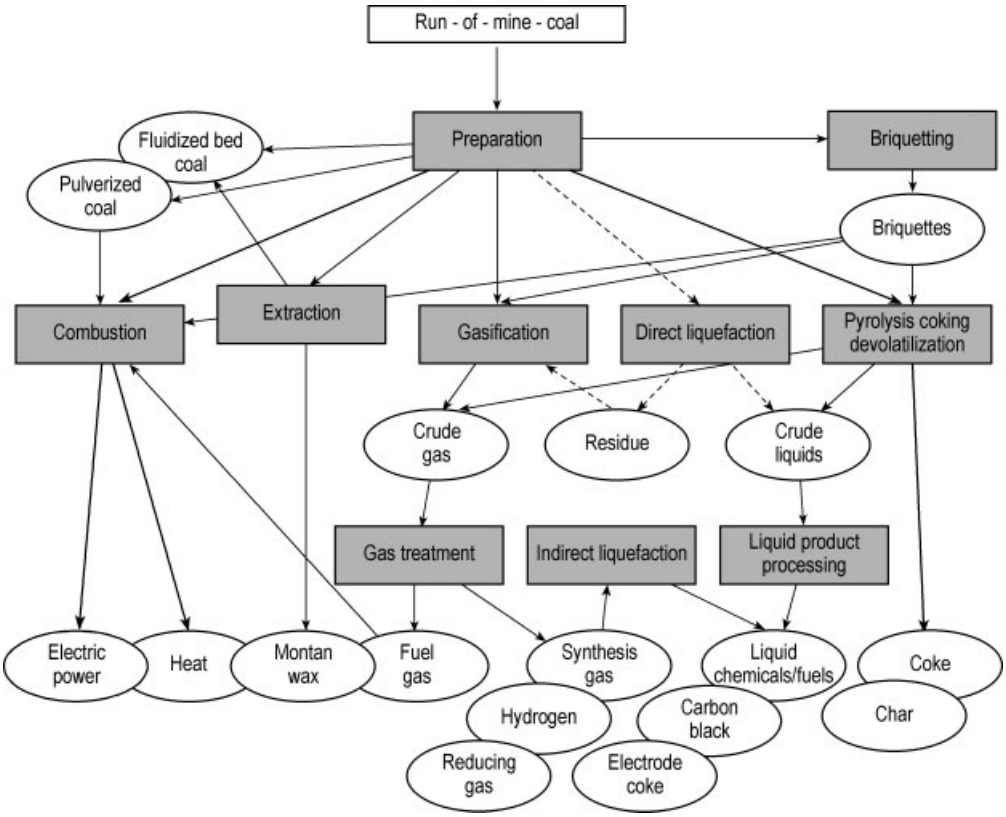

Figure 34. Coal utilization and conversion processes

over a wide range as a function of the process conditions, especially the reaction atmosphere; but also temperature, pressure and use of catalysts have to be taken into account. These are also the parameters to subdivide the individual thermochemical coal conversion processes into combustion, gasification, liquefaction, and pyrolysis. But there are transitional ranges which make the classification of some processes difficult.

The large experience in industrial-scale coal conversion processes is increasingly exploited for energetic and nonenergetic uses of biomass, industrial residues, and waste material.

9.1. Pyrolysis

Pyrolysis (also called thermolysis) is the thermal decomposition of organic substances such as coal during heating to more than 300 °C in a oxygen-free atmosphere. It is the generic term for carbonization, coking, and devolatilization. It is also the starting reaction in gasification, combustion, and direct liquefaction.

The decomposition products of pyrolysis are the pyrolysis gas (mainly hydrogen, carbon monoxide, carbon dioxide, methane, and C_2–C_5 hydrocarbons), the liquid products (tar, oil, crude benzene, and water), and coke as a solid residue and main product. Depending on coal properties, different sulfur and nitrogen compounds are formed in the pyrolysis process.

The pyrolysis product distribution and composition are mainly determined by the coal type but can be influenced by process parameters such as heating rate, temperature, atmosphere, and pressure.

The *hydropyrolysis* process operates under high hydrogen pressure and leads to a high yield of gaseous and liquid products, especially BTX aromatics (benzene, toluene, and xylene). In

flash pyrolysis a very high heating rate is used (500 – 20 000 K/min), in order to increase the yield of gaseous and liquid products. *Plasma pyrolysis* is carried out with a high heating rate in a high-temperature plasma atmosphere (2000 – 5000 °C), particularly to produce acetylene. These special pyrolysis processes with extreme process conditions currently have no or only few commercial applications.

9.2. Carbonization and Coking

Low-temperature carbonization and coking involve heating of coal with the exclusion of air. This process removes the condensable hydrocarbons (pitch, tar, and oil), gas, and gas liquor, leaving a solid residue of coke. Low-temperature carbonization (up to 800 °C) and coking (> 900 °C) are differentiated by the final temperature. The two processes also differ sharply in the rate of heating of the coal and in the residence time in the reactor. These parameters have a direct effect on the product yields. Low-temperature carbonization produces fine coke and fairly large quantities of liquid and gaseous products, whereas high-temperature coking is used primarily for the production of a high-temperature lump coke for blast furnaces and cupola ovens. After the oil crisis of 1973, attempts were made to submit the coal used in power stations to preliminary low-temperature carbonization. The coal byproducts were then to be refined while the coke was used for firing. However, the process could not be made profitable.

high-temperature coking of coal is carried out entirely in batch-operated coke ovens, of which the majority are of the horizontal chamber type. The feedstock is a coking coal of given size composition. The coking properties depend chiefly on softening and resolidification temperatures and on swelling behavior. Coking takes place at 1000 – 1300 °C. The coking time of 15 – 30 h depends on the operating conditions and width of oven. The main product is metallurgical coke required for the production of pig iron. Metalurgical coke is characterized by its suitable size and high resistance to abrasion even under blast furnace conditions. Coke oven gas and liquid byproducts are also produced. In Western Europe, these byproducts influence the economy of coking and, therefore, are reprocessed.

Considerable technical improvements in coke production have led to greater cost effectiveness. They include mechanization and automation of oven operation, reduction of coking time and increase of specific throughput by the use of thinner bricks of higher thermal conductivity, and increased oven sizes.

Coking in horizontal chamber ovens is becoming increasingly difficult for the long term due to the deteriorating quality of coal feedstocks. Therefore, coking technology must be assisted with additives and auxiliary technology. Potential additives include petroleum coke, bitumen, and oil. Among the auxiliary technologies to be considered are the ramming operation, such as is used in Germany and some East European countries, and the various modifications of coal preheating, which are used in the United States and Japan.

Dry cooling of coke has found increased use in Western Europe to recover stored heat. For climatic reasons this technique has been used for a considerable time in the (former) Soviet Union. Dry cooling of coke leads to considerable savings in energy and pollution, especially when used in combination with preheating.

At the start of the 1960s, many countries considered developing a continuous production process for formed coke in expectation of a further worldwide increase in demand for coke for steel production. Numerous process developments, such as briquette coking or hot briquetting of a mixture of char and caking coal, have been demonstrated on an industrial scale. Even so, these processes have not been able to gain acceptance. This is partly because unexpected improvements in chamber coking have been realized and partly because the steel industry has been having difficulty for more than a decade, and thus, the anticipated shortage of coking coal has not yet occurred.

9.3. Coal Liquefaction

Coal liquefaction is a conversion process in which liquid fuels and liquid chemicals are obtained from solid coal. To this end, the H : C ratio has to be increased by raising the hydrogen content to about 10 – 15 wt %.

Coal liquefaction can be accomplished in two ways. In the first, called *direct liquefaction* or *coal hydrogenation*, the coal is suspended in suitable oils and treated either with hydrogen in the presence of a catalyst or with hydrogenating solvents to yield oil products and some unreactive residue (→ Coal Liquefaction, Chap. 3). In *indirect liquefaction*, coal is gasified (Section 9.4) to yield a mixture of hydrogen and carbon monoxide (synthesis gas) from which liquid products can be synthesized in one or more steps (→ Coal Liquefaction, Chap. 2). Both routes developed into industrial-scale processes during the 1930s and were extensively used during World War II in Germany. Currently, (indirect) coal liquefaction is employed on an industrial scale only in South Africa (Sasol plants I, II, and III). Further developments were initiated to the point of large pilot plant operation, mainly in Germany, the United States, and Japan. These activities reached their peak between 1975 and 1985, and have partly been continued up to now (→ Coal Liquefaction).

9.3.1. Direct Liquefaction (Hydrogenation)

See also, → Coal Liquefaction, Chap. 3.

Coal hydrogenation is a hydrogenating digestion of the coal molecule. Hydrogenation using a solvent brings about depolymerization at ca. 15 MPa and 430 – 460 °C to yield an asphalt-like product. This product can be deashed, e.g., by hot filtration, and it can serve as a boiler fuel or as a feedstock, e.g., for the production of high-quality carbon products. Many development schemes went on to further hydrogenation of the filtration extract in a separate stage and thereby produce a synthetic oil that can be distilled and refined into marketable products [53, 54].

Catalytic hydrogenation of coal by the prewar Bergius – Pier process usually requires a pressure of tens of megapascals and a temperature of 450 – 480 °C. These strong reaction conditions yield a light oil fraction that can be further refined.

New generation processes are closely related to the earlier German developments but surpass these by (1) increased selectivity and yield of useful liquid products, (2) by a considerable reduction in operating pressure to 20 – 30 MPa from ca. 70 MPa, (3) by increased volumetric reactor throughput, (4) by improved availability of the plant, and (5) last but not least, by conditioning the unreacted residue in such a way as to make it accessible to steam reforming for hydrogen production. Depending on coal type and technology, the yield of liquid hydrocarbons reaches usually 50 – 7 wt %.

Although direct coal liquefaction is suited for production of low-sulfur and low-nitrogen boiler fuels, it is more useful for producing synthetic crude oil, which can be refined into light fuel oil, gasoline, and diesel as well as chemical feedstocks by petrochemical processing. However, direct coal liquefaction has presently no industrial use due to the availability of comparatively inexpensive liquid fuels and chemicals obtained from petroleum.

9.3.2. Indirect Liquefaction

See also, → Coal Liquefaction, Chap. 2.

The synthesis of hydrocarbons from synthesis gas is called indirect liquefaction. This technology is also named *Fischer – Tropsch synthesis* after its discoverers. The synthesis of liquid fuels by this process requires a synthesis gas with a H_2 : CO ratio of about 1 : 2.5. This ratio can be influenced by gasification conditions and subsequent gas conversion (Section 9.4). A great variety of catalysts were developed for the production of several liquid hydrocarbons.

Another way for indirect production of liquid fuels is the two-stage *Mobil methanol-to-gasoline* (*MTG*) *process* (→ Coal Liquefaction, Chap. 2.3). The first stage is catalytic methanol synthesis from synthesis gas. The second stage includes the catalytic reaction of methanol to gasoline.

The coal-based Fischer – Tropsch process used in South Africa is set up downstream of Lurgi fixed bed gasifiers. In the Sasol I plant, two different Fischer – Tropsch process technologies are employed, the fixed bed ARGE process (Lurgi – Ruhrchemie) and the entrained-flow Synthol process (Kellogg – SASOL). The ARGE process produces mainly gasoline fuel oil and waxes, whereas the Synthol process yields a larger fraction of gasoline and low molecular mass products (methane and propene). The SASOL II and III plants, built in the late 1970s,

use only the Synthol process. On an energy basis, the total liquid product yield obtained by indirect liquefaction of coal in the Sasol II plant is ca. 32 %.

9.4. Coal Gasification

Systematic development of coal gasification began in the first half of the 19th century. A mixture of carbon monoxide and hydrogen was produced. This was generally used for chemical purposes, which remained the main application of coal gasification for nearly 100 years. Town gas for lighting, cooking and – very much later – also for heating, in many cases was obtained as a byproduct from coke oven operation (→ Coal Pyrolysis). However, coal or coke gasification was also employed mainly for peak shaving in large gas schemes and also for the supply of gas to industrial furnaces. These applications provided strong development incentives, e.g., the first pressurized coal gasification for the manufacture of town gas to be distributed over larger distances [55, 56]. Natural gas and mineral oil in large amounts and at low cost replaced gas manufactured from coal not only in the heating market, but also in the chemical industry in most countries. The oil price increase of 1973 – 1974, and the fear for supply shortages, triggered new interest in coal gasification as well as the increased use of biomasses and waste material for energetic purposes in recent years (→ Gas Production).

Gasification can be accomplished by reacting coal with oxidizing or reducing agents. Using oxygen (or air) and steam as gasifying agents yields water gas.

$$C + H_2O \rightarrow CO + H_2$$

To supply the energy for this endothermic reaction either the carbon is preheated (water gas) or exothermic combustion and gasification reactions run in parallel (producer gas):

$$C + \frac{1}{2}\,O_2 \rightarrow CO$$

$$C + O_2 \rightarrow CO_2$$

The composition chosen for the blast depends on individual process conditions and governs the carbon monoxide : hydrogen ratio of the product. Varying the steam : oxygen ratio in the blast is a very sensitive and convenient means of controlling the reaction temperature.

Gasification with hydrogen (hydrogasification) leads primarily to methane in an exothermic reaction.

$$C + 2\,H_2 \rightarrow CH_4$$

However, due to thermodynamic and kinetic restrictions hydrogasification is usually incomplete and produces a char residue. Thus, reductive gasification is considered as only a preliminary step in the complete gasification of coal. However, it is a very favorable step if a gas with a high heating value, e.g., substitute natural gas (SNG), is to be produced. Hydrogasification is also involved in most fixed-bed countercurrent-flow gasification processes, contributing to the methane content of the product. Several other reactions take place simultaneously, including the Boudouard reaction, complete combustion of carbon, and methane formation from carbon monoxide and hydrogen.

In coal gasification processes, it is necessary to optimize heat transfer, mass transfer, and chemical reaction conditions, and to handle large streams of solid and gaseous agents. Many varieties of coal gasification processes have been developed, reflecting compromises of the many constraints and conditions.

First consideration must be given to the intended use of the product gas:

1) a fuel gas, e.g., a clean boiler or turbine fuel
2) synthesis gas, for the production of commodity chemicals
3) hydrogen, for ammonia production or for energy purposes
4) reducing gas, for direct reduction of iron ore
5) a gas with a high heating value, e.g., SNG

A second, equally important, consideration applies to the available coal. Of particular importance are coal rank, grain size, caking properties, ash content, and ash melting behavior. Third, environmental regulations, site specific infrastructure, etc. have to be taken into account. Considering all these factors, a decision in favor of

fixed-bed, fluidized-bed, or entrained-flow gasification can be made.

In a *fixed-bed* gasifier, coal and blast are usually contacted in countercurrent flow. This leads to the splitting of the overall gasification reaction into several zones with favorable conditions for the various reaction steps involved. Fixed-bed gasification usually results in excellent carbon utilization and high efficiency. It is applicable to coals of all ranks. However, it requires lump, noncaking or weakly caking coal with a high ash softening temperature. The product contains tars and other liquid byproducts that complicate the gas treatment. Atmospheric fixed-bed gasifiers of various design are still occasionally found in small-scale industrial use. On a large scale, some Lurgi fixed-bed pressurized gasification plants are operating commercially, e.g., in the Republic of South Africa (Sasol I, II, and III) and in the USA (Great Plaines). A more advanced process is the British Gas/Lurgi slagging gasifier requiring specifically less steam and being not sensitive regarding ash softening temperature.

Fluidized-bed gasification, invented in 1922 by Winkler at BASF, has the advantage of fairly simple reactor design. However, overall reaction temperature is limited by the need to avoid ash agglomeration. Therefore, Winkler gasifiers operating at atmospheric pressure to produce synthesis gas were restricted to reactive fuels such as subbituminous coals or lignite. These gasifiers are in operation, e.g., in Germany and in India. Further developments successfully led to the applications of higher pressure (HTW-Gasification) [57, 58]; of higher temperature allowing for local ash agglomeration (U-Gas) [59, 60]; and of higher superficial velocity (Circulating Fluidized Bed) [61].

Entrained-flow gasification takes place in a flamelike reaction zone, usually at a very high temperature to produce a liquid slag. For economical operation, a high-standard heat recovery system is mandatory, but the product gas is free of methane, tars, etc., thereby considerably simplifying gas and water treatment.

Entrained-flow gasifiers of the Koppers–Totzek design operated at atmospheric pressure are used industrially in many countries to produce hydrogen or synthesis gas. Further developments of this process type are aimed at the application of elevated pressure. The available processes differ in the modes of heat recovery and of coal feeding. Dry coal feeding is applied in the PRENFLO [62], the Shell [63] and the GSP processes, whereas in the DOW/Destec [64] and in the Texaco [65] process a coal water mixture is fed to the gasifier. Preferably these gasifiers are applied in conjunction with coal based combined cycle power systems (→ Gas Production).

Apart from these well-established or recently introduced processes, further intense development is under way on existing processes, as well as on new concepts. Such new concepts as pressurized gasification in various modes, hydrogasification, multistage processes, slag bath and molten iron processes, and allothermal processes, e.g., using process heat from gas-cooled nuclear reactors, have been proposed and tested [66 – 71].

Some of the new development projects have been terminated or shelved because of the present availability of cheap mineral oil and natural gas. However, the long-term incentive of securing and diversifying the world energy supply is still valid. Environmental considerations may favor converting coal into a clean fuel gas that can be used either conventionally or in highly efficient technologies like fuel cells or combined cycle power plants. Furthermore, low-quality coals, which are located in remote places and are therefore not marketable as such, can be gasified and converted on site into transportable products such as commodity chemicals or hydrocarbons and thereby contribute to worldwide supplies. Moreover, principles and know how of coal gasification are going to be applied in gasifying biomass and waste, either solely or in combination with coal.

9.5. Coal Combustion

In combustion only the energy content of coal rather than its carbon or hydrogen content is exploited. Since combustion products usually represent useless waste, combustion is not classically considered a coal conversion or refining process but rather an energy conversion process. The main reaction in coal combustion is complete oxidation of the carbon and hydrogen content with atmospheric oxygen. The heat released is utilized either directly, e.g., for room heating or indirectly, e.g., via steam generation. Although overall chemistry of coal combustion is

principally quite simple, the actual mechanisms involved are fairly complicated. They depend on reaction conditions and on many fuel parameters including chemical composition, intrinsic reactivity, volatiles and ash contents, particle size, porosity and pore structure, mineral matter properties, etc.

Entrained-Phase Combustion. The standard process applied in large-scale units is entrained-phase combustion of pulverized coal. Finely ground coal is injected through a burner into the furnace where it is mixed with the combustion air. Combustion occurs in a flame as main reaction zone. In the largest units several burners fire into one furnace either in separate flames or they form a common fire ball in the center of the furnace. At the high temperatures prevailing, reactions are fast and mixing of oxygen and fuel becomes rate determining. Hence, a large number of sophisticated burner designs have been developed to optimize the balance between the processes involved, i.e., heat release, heat transfer, gas and particle flow, ignition of the incoming fuel, and completion of the burn-out in the after-flame zones.

A recent development (1997) aims at combustion of pulverized coal under pressure [72] to make possible combined cycles with increased power generation efficiency.

Fluidized-Bed Combustion. In fluidized-bed combustion crushed coal is burned in a bed of hot inert material (usually ash or sand), fluidized by the combustion air. In this way combustion takes place in the very closely controlled temperature regime of ca. 750 – 950 °C. The high amount of hot inert bed material allows to fire also low-grade high-ash fuels with good efficiency. Moreover, because of the low combustion temperature sulfur dioxide (released by combustion of the fuel's sulfur content) can be effectively captured simply by addition of limestone to the fuel.

Fluidized combustion has been realized also at elevated pressure on an industrial scale. Application of elevated pressure offers the advantages of significantly smaller sized combustors and the possibility of gas turbine operation [73, 74].

Combustion on a moving grate. The oldest and still widely used combustion technology is combustion of lump coal on a fixed or moving grate. These stokers are particularly well suited for small- to medium-size units and feature simplicity of operation and a very wide range of part-load operation at still high efficiency.

Since these combustion systems operate most efficiently only with fuels meeting specific qualities, one of the main development issues in coal combustion is, on the one hand, the design of burner/furnace systems that can process coals with a wider margin of specification limits and fuel blends. On the other hand, characterization of coals/coal blends for use in combustion systems is another field of intense research [75 – 77].

Besides combustion of the main constituents of coal also unwanted side reactions occur, leading to pollution problems such as formation of oxides of nitrogen and sulfur and the emission of particulates. For large combustion facilities add-on units like catalytic converters, scrubbers, and filters are widely installed to control emissions. Modern combustion systems employ intrinsic combustion modifications such as low-NO_x burners or two-stage combustion as more cost effective means of pollution control [78, 79].

The advent of high-performance computing equipment as well as powerful codes for the simulation of combustion processes rendered possible the development of combustor models that – with often fair or good accuracy – can predict the response of the respective system to changes in fuel or operating condition with respect to temperature distributions, gas flows and concentrations, burn-out and pollutant levels. In the 1990s rapid development in this field also led to an improved understanding of the detailed mechanisms involved [80, 81].

Concerns over excessive emission of carbon dioxide into the atmosphere that might lead to inadverted climatic changes (greenhouse effect) have raised the interest in co-combustion of coal and biomass and other nonfossil solid fuels. Coping with the problems involved in handling and reliably and cleanly burning these fuels (or fuel blends) forms a new challenge for combustion experts [82 – 84].

9.6. Nonenergetic Use of Coal

Like petroleum, coal can be used as a basic feedstock in the chemical industry. Corresponding coal conversion processes were developed on an industrial scale. However, because of the more costly technology, coal-derived liquid fuels and chemicals are more expensive than those obtained from petroleum refining. Therefore the commercial application depends on availability and price of the different fuels and the economic or political situation. As a result of present (1998) relatively low global prices for petroleum and natural gas, the nonenergetic use of coal is of more local importance or is applied for the extraction of special products. That situation will change in the medium- or long-term future, when the prices for petroleum and natural gas will increase due to supply shortage (petroleum and natural gas constitute only about one-third of the global fossil fuel reserves, about two-third consist of coal; the current consumption of fossil fuels is in inverse proportion).

Several coal conversion processes can be used for the nonenergetic use of coal. Crude gas from coal gasification (Section 9.4; → Gas Production) is aftertreated corresponding to the quality requirements of subsequent processes and can then be used, e.g., for ammonia and methanol synthesis, for the Fischer – Tropsch synthesis and oxo synthesis. Liquid hydrocarbons obtained in coal pyrolysis (e.g., in coking or carbonization processes) (→ Coal Pyrolysis) and coal liquefaction (at present without commercial application) (→ Coal Liquefaction) can be refined to various basic materials for chemical industries. Other coal-based products are montan wax, electrode carbon, carbon black, and carbonaceous adsorbents.

9.6.1. Extraction of Montan Wax

Brown coal extraction is a thermal material separation process. The montan wax obtained in this way is a multicomponent mixture, consisting in particular of long-chain aliphatic acids, their esters, and long-chain alcohols (→ Waxes, Chap. 3). Montan wax is solid under ambient conditions and melts at about 85 °C without decomposition.

Only special types of bituminous brown coal are suitable for montan wax production. Therefore, this process is only of local importance. The biggest coal extraction plant is in Amsdorf in eastern Germany with a present annual output of about 18 000 t montan wax (these are about 85 % of the world's capacity). Since the beginning of the 20th century montan wax from bituminous brown coal has been produced on an industrial scale in this plant. The bituminous brown coal is selectively mined and prepared by crushing, screening and thermal drying (grain size 0.2 – 3 mm, water content 14 – 16 wt %). Coal extraction is carried out with toluene as extracting agent (other organic solvents are also possible) at 80 – 90 °C in a continuous belt extractor. The solvent is thermally recovered; the extracted and almost solvent-free coal can be used as fuel.

About three quarters of the crude montan wax are bleached and used for a variety of purposes, e.g., for carbon paper, as release agent for plastics, ceramics and insulating materials, as paint additive, as impregnating agent for leather, and for the building and construction industry.

9.6.2. Coal Tar Chemical Industry

Coal tar and crude benzene formed as byproducts in the coking process are used as aromatic chemical feedstocks in considerable quantities. About 15×10^6 t of hard coal tar are annually obtained in the production of metallurgical coke. In the caking process ca. 3 wt % crude tar, ca. 1 wt % crude benzene and also ca. 0.3 wt % ammonia are produced from dry coking coal.

Benzene is extracted from crude benzene and can be processed, e.g., to plastics intermediate products such as styrene and phenol.

Coal tar is first distilled into different fractions. These are further processed by crystallization, extraction, and carbonization. Important products are pure aromatics and aromatic oils, phenols and heterocyclic compounds – which are important basic materials for pigments and dyes and for pharmaceuticals – and carbon black, used, e.g., for colored printing ink and as filling material for rubber.

Pitch is the tar distillation residue with a boiling range > 450 °C (about 50 wt % of coal tar). It is particularly used for electrodes and carbon fibers (→ Tar and Pitch).

9.6.3. Production of Carbonaceous Adsorbents

Carbonaceous adsorbents are divided into three groups: activated coke, activated carbon, and carbon molecular sieves. Carbonaceous adsorbents can be produced from both hard coal and brown coal (besides other raw materials). If the adsorption-relevant pore system is primarily formed by pyrolysis, the solid residue can be used as *activated coke* for flue gas cleaning or for waste-water purification. The next stage is the activation of coke by partial removal of carbon by gasification. Thereby the pore system is modified, and the specific surface can be increased considerably. The *activated carbon* produced in this way is the most important carbonaceous adsorbent. It is used for recovery of solvents and other recycling materials, drinking water treatment, etc. *Carbonaceous molecular sieves* are obtained if carbon is added by pyrolysis of hydrocarbons in the pore system after the activation process. In this way the size of micropores can be fitted to the size of certain molecules, and the adsorbent is applied for separation of air, methane, hydrogen from coke oven gas, etc.

Besides the adsorption properties of these carbonaceous materials, also their ignition behavior and their resistance to abrasion, which have to be taken into consideration for many applications, can be influenced by the production process (→ Carbon).

10. Transportation

The large amount of coal that is mined every year is a major commodity for the transportation industry and one that requires a significant capital investment. In the USA, ca. 50 % of the $700-800 \times 10^6$ t of coal produced per year is shipped by rail; 18 % is shipped by river barges; and the rest is handled by trucks, conveyor belts, and pipelines. The actual cost of long-distance shipping is usually quite modest, but the loading and unloading of the coal adds greatly to the total cost of shipping. In a typical surface mining operation, for example, coal is moved from the mine by truck or conveyor belt to a nearby preparation plant, where it is loaded onto a unit train for rail shipment to a utility or port facility. At the port it may then be loaded onto river barges or ocean-going bulk carriers.

Although some of the coal transported by *rail* is shipped at single-car or bulk rates, most is carried at lower rates in unit trains of up to 100 cars that travel continuously to a single destination and return. The unit train carries only coal and each car can carry up to 100 t. The cars are of either the bottom-dumping or roll-over type, and each car can be unloaded in 1.5 – 5 min in modern processing facilities, which have a capacity of 2000 – 6000 t/h.

The least expensive means of transporting coal is by *river barges*, whereby the cost is as low as $ 0.003 per kilometer. The barges have open tops and commonly hold up to 1000 – 1500 t each. They are moved in groups or tows of 20 – 30 barges powered by a single towboat. Although there can be some problems with delivery schedules due to congestion, the ease and economy of river transport are major factors in the location of many coal-burning utilities along inland waterways in the USA and in Europe.

In the USA, ca. 12 % of coal production is carried by *trucks*. Much of this is done at surface mines, where the coal is hauled from the working face to preparation plants. This is usually done with dedicated high-capacity vehicles; however, significant amounts of coal are hauled by trucks on public highways. *Conveyor belts* are also used for transporting coal from mines to preparation plants, as well as for short cross-country hauls of over 16 km to loading points. Because they are dependable and economical, conveyor belts are also widely used for in-mine coal transport.

Although > 80 % of the coal mined in the USA is used domestically as steam coal, some of the metallurgical coal production is shipped overseas. This coal is shipped in dry bulk carriers that can be loaded and unloaded in modern port facilities at rates of up to 100 000 t/h. These *ocean-going ships* are usually in the 60 000 t range, although some are larger. In addition, some smaller vessels in the 30 000 t range that are self-unloading are also in service. These ships have the advantage of being able to use a much greater number of ports.

The transportation of coal by *slurry pipeline* has been demonstrated to be dependable and economical, even though there are only a few such pipelines in use around the world. In this

technique the coal is reduced to a diameter of < 1 mm, treated with chemicals to prevent corrosion and improve flow characteristics, and mixed with water. It is then pumped as a slurry at a velocity of 1.5 – 2 m/s. The fact that pipelines can be built above ground or buried eliminates some environmental problems, but their high water requirements can be a serious drawback. Sustained delivery rates of > 600 t/h have been demonstrated.

11. Coal Storage

The use of any bulk commodity requires the storage of sufficient material to ensure efficient operations. In the case of coal, this requirement varies from a few days supply at the mine to a few months supply at a power plant. Although some covered storage in bins or silos is practical at mines and small operations, most of the larger volumes of coal are stored in open piles, where it is exposed to the air and precipitation. When coal is exposed to air, it loses moisture and begins to oxidize. This oxidation can degrade coal quality, and more importantly, it can cause spontaneous combustion that destroys the coal. The main coal properties that influence oxidation are particle size, pyrite content, and rank. Coal oxidation leading to spontaneous combustion is enhanced as the particle size decreases and as the pyrite content increases. Low-rank coals such as lignite and subbituminous coal are very difficult to store because of their strong tendency toward spontaneous combustion, and care must be taken even with the lower rank bituminous coals. Repeated wetting and drying also exacerbate oxidation. The continued oxidation generates more heat than can be dissipated, and this leads to hot spots, which eventually ignite.

In a stockpile even mild oxidation can lead to degradation of coal quality. For steam coals there can be some loss in calorific value, but in coking coals there can be a total loss of quality. The fluid properties can be destroyed, and the heat-transfer properties and coke production yields can also be reduced. The quality of coke made from oxidized coal is also seriously altered. The prime coke property, coke strength or stability, is reduced, and this can result in the loss of iron production in the blast furnace. Oxidized coal also increases the reactivity of coke and leads to higher coke consumption in the blast furnace. Because all of the effects of oxidation are undesirable, every effort should be made to prevent it.

The main steps to prevent oxidation involve limiting the access of air and protection from wind and precipitation. Stacking the coal to eliminate size segregation and compaction of the pile with earth-moving equipment to reduce pore space have proven helpful. Coating the pile with a sealant also helps protect it from air and moisture, although this may be expensive. Another costly but useful step is to provide the storage piles with some kind of wind protection. The wind not only enhances oxidation, but it also removes the fine particles from the coal pile and can thereby create environmental problems.

Another aspect of coal storage is the need to mix and blend the stockpile coal to homogenize the product coal reclaimed from it, so that a consistent, uniform product can be delivered to the plant. This can be a serious concern in cases where coal enters an operation from a number of sources. The usual way to solve this problem is to develop a bedding and blending system. A typical system consists of building a stockpile with long, thin layers of the various incoming coals and reclaiming the coal from the pile with vertical cuts across these layers. The key design factors are the variability of a target blending parameter such as ash or sulfur content and the desired degree of homogeneity. These are evaluated to determine the necessary pile parameters such as individual layer thickness.

12. Quality and Quality Testing

Because coal is so variable in its maceral and mineral composition, its suitability for a given use must be determined by a variety of tests. Although some tests such as chemical analysis are quite general in nature, others such as ash fusion are specific for particular uses.

12.1. Chemical Analyses

Although coal is composed of a large number of organic chemical components, no true standard organic chemical test exists for coal. The two

major kinds of chemical analyses used are the proximate analysis and ultimate analysis, which are standard tests defined by the International Organization for Standardization (ISO).

The proximate analysis consists of a determination of the moisture, ash, and volatile matter, and a calculation of the fixed carbon value. The *moisture* (ISO 331–83) is determined by heating the sample at 105 – 110 °C to a constant weight. The percent weight loss is reported as the moisture value. The *ash content* (ISO 1171–97) in this analysis is the incombustible residue after the coal is burned in air to a constant weight. The ash content is not a measure of the kinds or relative amounts of the minerals in the coal. The *volatile matter* content (ISO 562–98) is a measure of the amount of gas and tar in a coal sample. It is reported as the weight loss minus the moisture after the coal is heated at 900 °C for 7 min. The *fixed carbon* is not a distinct chemical entity. It is reported as the difference between 100 % and the sum of the moisture, ash, and volatile matter values. The standard tests defined by the American Society for Testing and Materials (ASTM) differ usually only in the amount of the sample and/or the heating rate and temperature conditions (ASTM D 3172–89). While the proximate analysis as described above is a rather simple assay of the chemical nature of coal, its value as a quality parameter is well established and is widely used in commerce.

The ultimate analysis consists of direct determinations of ash, carbon, hydrogen, nitrogen, sulfur, and an indirect determination of the oxygen. The ash is determined as in the proximate analysis, and because all of the values are reported on a moisture-free basis, moisture must also be determined (Table 11).

Table 11. Ultimate analysis, test standards

Test	ISO	ASTM D 3176–93
Carbon/hydrogen	609–96, 625–96	D 3178–93
Nitrogen	332–81, 333–96	D 3179–93
Total sulfur	334–92, 351–06	D 3177–93
Oxygen	1994–76	
Ash	1171–97	D 3174–93
Moisture	331–83	D 3173–92

Both proximate and ultimate analyses are reported in a number of different ways, and care must be taken to be certain of the method of reporting. On the "as-received" basis, the results are based on the moisture state of the coal sample as it was received for testing. With the "dry" basis, the results are calculated back to a condition of no moisture, and with the "dry, ash-free" basis, the results are calculated to a condition of no moisture and no ash. These calculations are done so that different coals can be compared on their inherent organic nature. The reporting basis (ISO 1170–77, ASTM D 3180) can cause significant changes in the values reported, and it is essential that a target value and the value of a sample in question be on the same basis.

Because some components of the minerals such as water from the clays and carbon dioxide from calcite are lost in the high-temperature ashing process, the ash value determined is less than the actual mineral matter in the raw coal. A number of corrections for this loss are in use, but the one most used in the United States is the Parr formula, where the corrected mineral matter is equal to 1.08 times the ash percentage plus 0.55 times the sulfur percentage. The average mineral matter factor for German Ruhr coals is 1.09. Results reported with this correction are considered to be on a dry, mineral-matter-free basis.

12.2. Mineral Matter

The mineral matter content of a coal is the actual weight percent of the minerals present. It is the best measure of the inorganic content of a coal, but it is difficult to determine. Determination of the mineral matter content or mineral matter factor follows a procedure which is described in the standard test ISO 602–83:

- Partial demineralization of a sample of the coal by treatment with hydrochloric and hydrofluoric acids under such conditions that the coal substance remains unaffected.
- Recording of the loss in mass of the coal due to the acid treatment and determination of the undissolved part of the mineral matter by ashing the partially demineralized coal.
- Determination of the iron content of the ash so that the pyrites present in the extracted coal can be calculated.

– Determination of the amount of hydrochloric acid absorbed by the coal substance.

No corresponding ASTM standard exists, however a procedure widely used in the United States is low temperature ashing ($< 150\,^{\circ}$C) with an oxygen plasma device. The X-ray diffraction analysis of the low-temperature ash is used to identify the actual minerals in the ash. Other techniques used to analyze the minerals are infrared spectroscopy, differential thermal analysis, electron microscopy, Mössbauer spectroscopy, and light microscopy for petrographic investigations. However, with present techniques it is impossible to accurately determine the amounts of the various minerals present in a given coal.

Although a large number of different minerals have been identified in coal, the four most common substance classes are clays, sulfides, carbonates, and quartz. The minerals get into the coal in a variety of ways. Some can be part of the original plant material itself. Silica (quartz) is present in some of the saw grasses of modern swamps, and similar types of plants are known from ancient swamps. Certainly much of the mineral matter is transported into the coal-forming swamp from the environment. Much of the clays and quartz minerals were brought into the coal swamp as clastic material in streams or in airborne dust. Most of the common minerals could be chemically precipitated from solution under conditions that can exist in swamps. Some minerals such as sulfides and carbonates can also be precipitated into cleats and fractures in coal after it has formed and coalified beyond the peat stage.

The presence of sulfur in coal is of great interest because of the problems it causes in utilization, especially air pollution. There are three commonly recognized forms of sulfur in coal. Sulfide sulfur is the sulfur tied up in, e.g., pyrite, FeS_2; it is determined by leaching with nitric acid. Sulfate sulfur is of minor importance and thought to be formed by the weathering of pyrite; it is determined by leaching with hydrochloric acid. Organic sulfur is that portion of the total sulfur that is organically bound with the various coal macerals; it is determined indirectly by difference. There also may be some minor elemental sulfur present in coal. The forms of sulfur are determined in accordance with the ASTM standard D 2492 or ISO 157–96.

The occurrence of pyrite in coal is of special interest because it is the only form of sulfur that can usually be removed by mechanical cleaning methods and because it is a major source of air and water pollution. Although some pyrite is formed by the chemical combination of the sulfur and iron that occur naturally in peat, most of the pyrite found in coal is thought to form from Fe(III) ions absorbed on clay minerals that are transported into coal swamps, and from sulfate ions that are introduced into the swamp in seawater. Marine water usually has about two orders of magnitude more sulfate ions than fresh water. In the swamp the ferric Fe(III) is reduced to Fe(II), which combines with the sulfur in the sulfate ions to form pyrite. Although this process is clearly not the only way of pyrite formation in coal seams, it is a good model for many U. S. coal seams, especially those in the Illinois Basin. In these seams it is reported [85 – 87] that both the total sulfur and pyritic sulfur contents are controlled by the nature of overlying rocks; they are both high under marine rocks and low under nonmarine rocks.

Pyrite can occur in coal in a number of forms such as single crystals, void fillings, irregular and dendritic masses, and framboids (raspberry-like clusters). The most significant forms are the more massive ones, which are easier to remove from the coal by washing, and the framboidal forms, which are more chemically active because of their high surface area. When pyrite is exposed to air and water, it oxidizes in a series of chemical reactions that generate sulfuric acid and cause acid mine drainage. In fact, the reaction of 1 mol of pyrite with air and water results in the generation of 2 mol of sulfuric acid.

12.3. Thermal Properties

The most widely used thermal property is the *gross calorific value* (synonymous superior calorific value, higher heating value), which is the specific energy of combustion for unit mass of a solid burned in oxygen under given conditions. The products of combustion are assumed to consist of specified gaseous products and liquid water (ASTM D 3286 & D 2015, ISO 1928–95). The *net calorific value* (synonymous infe-

rior calorific value, lower heating value) is defined as above, however all the water formed as reaction product remains in the vapor phase. For the classification of coals by rank (ASTM D 388–95) the gross calorific value is expressed on a moist, mineral-matter-free basis. In the "International Classification of Hard coals by Type" and for the use in steam generation the gross calorific value is expressed on an air-dried, ash-free basis (DIN 23003, Publication des Nations Unités Sales No. 1956 11E4). In many utility company contracts, coal is bought on the basis of total calorific value per unit mass.

13. Economic Aspects

13.1. World Outlook

Reserves. World proven recoverable reserves of coal are an estimated 519×10^9 t of hard coal and 465×10^9 t of brown coal (subbituminous and lignite) (Table 12). Proven reserves are those which geological and engineering information indicate with reasonable certainty can be recovered in the future from known deposits under existing economic conditions.

Table 12. Proven recoverable world reserves of coal as of end 2001 [88] (Quantity, 10^6 t)

Region or country	Anthracite and bituminous	Subbituminous and lignite
Africa and Middle East	56 881	196
North America	120 222	137 744
South and Central America	7 738	14 014
Asia	147 042	66 456
Former Soviet Union	97 362	132 613
Europe	47 234	74 289
Australia and New Zealand	42 583	40 079
Total	519 062	465 391

Production. Total production of hard coal (anthracite and bituminous) worldwide has increased clearly only since 2002. However, the individual regions of production have developed to different extents. For instance, production in Europe has been reduced, while the coal output of countries such as Australia, China, India, Indonesia, Columbia, and South Africa has increased, in some areas considerably. In 2003 production of hard coal reached 3.9×10^9 t (Table 13). World production of brown coal (subbituminous and lignite) has remained at the same level of 0.9×10^9 t since 1998 (Table 14). This amount yields approximately the same heat as 0.3×10^9 t of hard coal.

Table 13. World hard coal production in 2003 [89]

Region or country	Quantity, 10^6 t	Proportion, %
Africa		
South Africa	238.2	6.1
Others	9.7	0.3
Africa total	247.9	6.4
Americas		
United States	895.6	23.1
Canada	26.6	0.7
Brazil	5.0	0.1
Columbia	48.0	1.2
Mexico	6.5	0.2
Others	7.7	0.2
Americas total	989.4	25.5
Asia		
China*	1315.2	33.9
India	348.4	9.0
Indonesia	106.8	2.7
North Korea	53.0	1.4
South Korea	3.3	0.1
Pakistan	3.3	0.1
Turkey	3.0	0.0
Vietnam	6.5	0.2
Others	2.0	0.0
Asia total	1841.5	47.4
Europe		
Czech Republic	13.7	0.4
Poland	102.2	2.6
Former Soviet Union	353.0	9.1
Germany	28.8	0.7
Spain**	12.6	0.3
United Kingdom	27.8	0.7
Others	7.1	0.2
Europe total	545.2	14.0
Australia & New Zealand	259.6	6.7
World total	3883.6	100

* Including brown coal
** Including lignito negro

Consumption. Approximately 20 % of the 3.9×10^9 t/a of hard coal used in the world is allocated to coking coal, and 80 % to steam coal. Virtually all brown coal is used in power stations. The largest growth in both production and consumption in the last years has been in the former Soviet Union, Australia and Asia.

Trade. Most coal is consumed in the regions of production. World trade in hard coal amounts to only 17 % of world production. Nevertheless, since some important consumer countries are without adequate resources of their own,

Table 14. World brown coal production in 2003 [89]

Region or country	Quantity, 10^6 t	Proportion, %
Americas		
United States	74.7	8.6
Canada	35.7	4.1
Americas total	110.4	12.7
Asia		
Mongolia	5.0	0.6
India	26.0	3.0
Noth Korea	23.0	2.6
Turkey	43.5	5.0
Thailand	19.8	2.3
Asia total	117.3	13.5
Europe		
Bulgaria	27.2	3.1
Czech Republic	50.2	5.8
Slovakia	3.1	0.4
Hungary	12.0	1.4
Poland	60.9	7.0
Romania	30.2	3.5
Former Soviet Union	99.4	11.4
Former Yugoslavia	34.0	3.9
Germany	179.1	20.6
Greece	68.2	7.8
Ireland	2.0	0.2
Spain	8.0	0.9
Others	2.2	0.3
Europe total	576.5	66.3
Australia & New Zealand	65.2	7.5
World total	869.4	100.0

Table 15. World trade in hard coal in 2003 [89]

Exports	10^6 t
United States	39
Canada	26
Columbia	46
Venezuela	7
China	95
Indonesia	90
Poland	20
Former Soviet Union	34
South Africa	71
Australia	215
Others	21
	664
Imports	
European Union	200
Other Europe	4
Japan	162
South Korea	72
Taiwan	51
Hong Kong	11
India	24
OtherAsia	31
Latin America	24
Others	85
	664

trade in hard coal has assumed worldwide proportions (Table 15). Australia, China, Indonesia, and South Africa are the main net exporting regions; Japan, Western Europe, South Korea, and Taiwan are the main importing regions. Brown coal is not traded over long distances.

13.2. Some Major Coal Producing Countries

13.2.1. China

Reserves. China accounts for ca. 12 % of the world's recoverable hard-coal reserves. Hard-coal reserves are concentrated in the north and northwest of the country. Substantial lignite deposits (about 11 % of the world's reserves) are distributed throughout the country.

Production. China is the largest producer in the world (2003: 1315×10^6 t hard coal including about 100×10^6 t lignite). Between 1996 and 2001 production decreased clearly, but increased again to the previous level recently.

Consumption. The vast majority of China's coal production is consumed in the country itself. Only a small portion is exported. About two-thirds of primary energy consumption is based on coal, of which more than half is used for electricity generation. The Chinese economy influences the global energy and raw-materials markets and is geared to further expansion.

13.2.2. United States

Reserves. The United States possesses economically recoverable hard coal reserves of ca. 116×10^9 t–one-fifth of total world reserves. For brown coal the respective figures are 134×10^9 t and nearly 30 %.

Production. The United States has maintained a high level of production for decades, producing one-fourth of the world's coal. Longwall mining accounts for an increasing proportion of underground coal production.

Consumption. Domestic coal consumption, which absorbs nearly 95 % of production, has varied little in the past few years, with power stations being the main consumers. Disregarding the traditional deliveries to Canada, coal imports, mainly from South America, are meanwhile higher than coal exports.

13.2.3. Former Soviet Union

Reserves. With 230×10^9 t of recoverable coal reserves, 97×10^9 t of which is hard coal, the former Soviet Union accounts for 25 % of total world reserves.

Production. The former Soviet Union still ranks third among world producers. The political changes were followed by major economic disruptions, which affected coal production in particular. Since 1998 hard coal production has increased continuously whilst lignite production has stagnated since 1995.

Consumption. The majority of production is consumed domestically. Exports to the EU and Asia show an increasing trend.

13.2.4. India

Reserves. India possesses ca. 15 % of world reserves of hard coal.

Production. Coal is the main source of energy and vital for India's economic development. Efforts to develop the industry have resulted in a steady increase in production to nearly 350×10^6 t of hard coal.

Consumption. Consumption doubled since 1984. Despite increased production, coal is imported, mainly from Australia. For the coming years a sustained upturn in growth with an impact also on the coal sector is expected for India.

13.2.5. Australia

Reserves. Australia possesses economically recoverable reserves of 43×10^9 t of hard coal and 40×10^9 t of brown coal, each representing 8 % of world reserves.

Production. Coal production has risen steadily in the last few decades (2003: 257×10^6 t) and is expected to do so in future.

Consumption. The high-quality coals found in Queensland and New South Wales are of value as steam and metallurgical coals, and most of production is for export. With 215×10^6 t Australia accounts for one-third of the entire global coal trade. In the world coking coal market it even holds a share of 60 %. Production from Western Australia, Victoria, and South Australia (bituminous and sub bituminous coal, lignite) is geared to local thermal electricity generation.

13.2.6. Republic of South Africa

Reserves. South Africa has recoverable reserves of ca. 50×10^9 t of hard coal, which at current production levels would last nearly 200 years.

Production. Following a phase of stagnation until 2002 production increased clearly only in 2004 (238×10^6 t).

Consumption. The slight increase in production in the last years was initiated by higher internal demand. Coal is the main primary energy for the generation of electricity (95 %). Exports are over 40×10^6 t/a.

13.2.7. Indonesia

Reserves. Indonesia has vast resources of high-quality accessible coal. In terms of tonnage, southern Sumatra contains the highest proportion, although much of it is lignite. The coal from Kalimatan includes some deposits of very high quality (low ash and sulfur, high volatile matter).

Production. Recently, annual production has risen significantly to meanwhile over 100×10^6 t.

Consumption. The national policy is to expand internal use; the target is for 80 % of electricity to be generated from coal-fired plants. But the main impulse to increase the production is the expanding export market. Indonesia has become the world's third largest exporter of coal (2003: 90×10^6 t).

13.2.8. Poland

Reserves. Polish recoverable coal reserves amount to 20×10^9 t of hard coal and 2×10^9 t of brown coal.

Production. Poland boosted its production for export reasons for a long period. In the mid-1980s government officials expected to increase hard-coal output to 205×10^6 t and its brown-coal output to 100×10^6 t by the year 2000. Instead, production in 2003 fell to 102×10^6 t of hard coal and 60×10^6 t of brown coal.

Consumption. Inland consumption decreased over the last decade and may decrease even further. In the meantime, Poland tried to maintain its exports. In 1984, when it was able to export 45×10^6 t Poland set its intermediate export goal at 42×10^6 t/a of hard coal. In 2003, exports reached only 20×10^6 t.

13.2.9. Germany

Reserves. Germany possesses nearly 7 % of the world's economically recoverable coal reserves.

Production. Output of hard coal has fallen from a peak of 150×10^6 t in 1957 to 29×10^6 t in 2003 and further reductions are planned. Brown-coal production dropped with the closure of mines in the eastern Länder. Nevertheless, with ca. 180×10^6 t, Germany is by far the largest producer of brown coal in the world.

Consumption. The majority of the hard coal (65 %) and nearly all of the brown coal is channeled into the generation of electricity. The second largest consumer is the steel industry. Following the decline of indigenous production, hard-coal imports have steadily increased (2003: 35×10^6 t).

13.2.10. Columbia

Reserves. Figures on hard coal reserves vary widely.

Production. As a result of growing demand for exports production has been expanded significantly over recent years (2003: 48×10^6 t).

Consumption. The major share (2003: 46×10^6 t) of production is exported. Columbia has become an important supplier of steam coal for the European Union and the USA.

14. References

1. D. W. van Krevelen: *Coal: Typology, Chemistry, Physics, Constitution*, 3rd ed., Elsevier, Amsterdam 1993, p. 979 ff.
2. M. C. Stopes: "On the petrology of banded bituminous coals," *Fuel* **14** (1935) 4 – 13.
3. International Committee of Coal Petrology (ICCP): *International Handbook of Coal Petrography*, 1st ed., CRNS, Paris 1957; 2nd ed., 1st Suppl., CRNS, Paris 1963; 2nd ed., 2nd Suppl., CRNS, Paris 1975; Suppl. to 2nd ed., CRNS, University of New Castle upon Tyne 1985.
4. International Committee for Coal Organic Petrology (ICCP): "The new vitrinite classification (ICCP System 1994)," *Fuel* **77** (1998) no. 5, 349 – 358.
5. M. Teichmuller, R. Teichmuller: "Geological cause of coalification," in R. F. Gould (ed.): "Coal Science," *Adv. Chem. Ser.* **55** (1966) 133 – 163.
6. P. A. Hacquebard, J. D. Donaldson: "Rank studies of coal in the Rocky Mountains and Inner Foothills Belt Canada," in R. R. Dutcher, P. A. Hacquebard, J. M. Schopf, J. A. Simon (eds.): "Carbonaceous materials as indicators of metamorphism," *Spec. Pap.-Geol. Soc. Am.* **153** (1974) 75 – 93.
7. E. Hryckowian, R. R. Dutcher, F. Dachille: "Experimental studies of anthracite coals at high pressures and temperatures," *Econ. Geol.* **62** (1967) no. 4, 517 – 539.
8. J. M. Schopf: "A definition of coal," *Econ. Geol.* **51** (1956) 521 – 527.

9. R. Thiessen: "What is coal?" *Int. Circ. U.S. Bur. Mines* **7397** (1947) 48 p.
10. M. C. Stopes: "On the four visible ingredients in banded bituminous coals," *Proc. R. Soc. London, Ser. B* **90** (1919) 470.
11. M. L. Gorbaty, K. Ouchi (eds.): "Coal Structure," *Adv. Chem. Ser.* **192**, American Chemical Society, Washington, D.C., 1981.
12. R. C. Neavel: "Origin, Petrography and Classification of Coal," in M. A. Elliot (ed.): *Chemistry of Coal Utilization*, 2nd suppl. vol., Chap. 3, J. Wiley & Sons, New York, Chichester, Brisbane, Toronto 1981, pp. 91 – 158.
13. D. W. van Krevelen: *Coal and Its Properties Related to Conversion*, Int. Conference on Coal Conversion, Pretoria, RSA, 16 – 20 August 1982.
14. R. M. Davidson: *Molecular Structure of Coal*, Rep. No. ICTIS/TR 08, Jan. 1980, IEA Coal Research, London.
15. G. R. Gavalas, P. H.-K. Cheong, R. Jain, *Ind. Eng. Chem. Fundam.* **20** (1981) 113 – 122.
16. G. J. Pitt, G. R. Millward: *Coal and Modern Coal Processing, An Introduction*, Academic Press, London, New York, San Francisco 1979.
17. R. V. Ramani et al.: "User's Manual for Premining Planning of Eastern Surface Coal Mining, vol. 2: Surface Mine Engineering," *National Technical Information Service Report No. PB81–109–415*, U.S. Dept. of Commerce, Springfield, VA, 1980.
18. L. J. Thomas: *An Introduction to Mining*, Hicks, Smith & Sons, Sydney 1973.
19. M. K. Hessling: "Mine Productivity," *IEACR/35* (1991) March.
20. W. Durst, W. Vogt: *Bucketwheel Excavator*, Trans Tech Publications, Clausthal-Zellerfeld 1988.
21. *Braunkohle Surface Mining* (1996) no. 3.
22. R. Stefanko: *Coal Mining Technology: Theory and Practice*, Society of Mining Engineers of AIME, New York 1983.
23. D. Rebb: "Latest Design in Coal Pipelining," *Can. Min. J.* **104** (1983) no. 3, 20 – 22.
24. T. Wheeler: "Hydraulic Transport of Coal: Slurry Pipelines," *Mine Quarry*, **14** (1985) 35 – 38.
25. J. R. Siemon: "Economic Potential of Coal-Water Mixtures," *IEA Coal Res.*, London, Sept. 1985.
26. R. von der Gathen, K.-H. Kubitza, B. Bogenschneider: "Die Auswirkungen der Vergleichmäßigung von Rohförderkohle auf Kosten und Produkte der Aufbereitung," *Glückauf* **119** (1983) 22 – 27.
27. M. Hampel: "Großsiebmaschinen für die Vorklassierung in Steinkohlenaufbereitungsanlagen," *Glückauf* **113** (1977) 80 – 85.
28. K.-H. Kubitza, P. Wilczynski: "Trockene Feinstkornabscheidung in Sichtern: neuere Entwicklungen bei der Ruhrkohle AG," *Glückauf* **110** (1974) 480 – 484.
29. E. Fellensiek: "Das Setzverhalten unterbettgepulster Feinkorn-Großsetzmaschinen," *Glückauf-Forschungsh.* **39** (1978) 207 – 212.
30. E. Fellensiek: "Feinkornsortierung auf einer neuartigen, in Doppelfrequenz gepulsten Durchsetzmaschine," *Glückauf-Forschungsh.* **42** (1981) 130 – 136.
31. S. Heintges: "Die Entwicklung der Schwertrübesortierung von Steinkohlen in der Bundesrepublik Deutschland," *Glückauf* **109** (1973) 955 – 960.
32. M. Becker: "Der Weg zu den Großraumflotationsanlagen bei der Ruhrkohle AG," *Glückauf* **113** (1977) 952 – 955.
33. W. Blankmeister, B. Bogenschneider, K.-H. Kubitza, D. Leininger, L. Angerstein, R. Köhling: "Optimierung der Fein- und Feinstkohlenentwässerung im Bereich unter 10 mm," *Glückauf* **112** (1976) 758 – 762.
34. W. Erdmann: "Neuere Entwicklungen bei der Entwässerung von fein- und feinstkörnigen Steinkohlenerzeugnissen," *Aufbereit. Tech.* **26** (1985) 249 – 258.
35. D. Leininger, P. Wilczynski, R. Köhling, W. Erdmann, T. Schieder: "Behandlung und Verwertung von Flotationsbergen in der Bundesrepublik Deutschland," *Glückauf* **115** (1979) 467 – 472.
36. R. von der Gathen: "Möglichkeiten und Grenzen der Entschwefelung von Steinkohle," *Glückauf* **115** (1979) 112 – 118.
37. W. P. Bethe, G. Koch: "Neue Bauformen für Aufbereitungsanlagen," *Glückauf* **119** (1983) 368 – 373.
38. M. Teper, D. F. Hemming, W. C. Ulrich, EAS-Report E2/80, London, January 1983.
39. S. A. Elmquist et al., DOE/FE/05147–1488, May 1983.
40. M. K. Schad, C. F. Hafke, CEP May 1983, 45 – 51.
41. K. V. S. Sastry, D. W. Fuerstenau, EPRI CS-2198, Project 1030–1, January 1982.
42. F. P. Calhoun, *Min. Congr. J.* **49** (1962) 38 – 39.

43. N. Galbenis, O. Abel, J. Lehmann, W. Peters, *Erdöl Kohle, Erdgas, Petrochem.* **34** (1981) 59 – 65.
44. F. H. Beckmann, *Stahl Eisen* **100** (1980) 803 – 813.
45. E. Dunger, P. Dittmann, H. Reißmann, *Neue Bergbautech.* **10** (1980) 427 – 430.
46. H. Krug, W. Naundorf, *Energietechnik* **31** (1981) 62 – 68, 228 – 232.
47. R. Kurtz, *Braunkohle (Düsseldorf)* **32** (1980) 368 – 372, 443 – 448.
48. A. F. Baker, R. E. McKeever, A. W. Deurbrouck, DOE, RI-PMTC-12 (82), March 1982.
49. M. A. Colaluca, TENRAC/EDF-043, July 1981.
50. H. P. Hudson, J. E. Landon, J. H. Walsh: *The Canadian Mining and Metallurgical Bull. for January 1964*, Montreal, pp. 52 – 58.
51. K. V. S. Sastry, V. P. Mehrotra, 3rd Int. Symp. Agglomeration, Nürnberg, May 1981, H 36 – 38.
52. M. Schad, D. Sauter et al., BMFT-FB-T-85, Oct. 1985, pp. 108 – 109.
53. E. Ahland, F. Friedrich, I. Romey, B. Strobel, H. Weber, *Erdöl, Erdgas* **102** (1986) no. 3, 148 – 154.
54. A. G. Comalli, J. B. MacArthur, H. H. Stotler: "H-Coal Process Demonstrations, Development and Research Activities," *Prepr. Pap. Am. Chem. Soc. Div. Fuel Chem.* **27** (1982) no. 3/4, 104 – 113.
55. R. Drawe, *GWF Gas Wasserfach* **80** (1937) 806 – 810.
56. E. Kapp, *Braunkohle, Wärme Energ.* **11/12** (1955) 7 248 – 254.
57. G. Franken, W. Adlhoch, W. Koch, *Chem. Ing. Tech.* **52** (1980) 324 – 327.
58. K. A. Theis, E. Nitschke: "Make Syngas from Lignite," *Hydrocarbon Process.* **66** (1982) no. 9, 233 – 237.
59. P. Palat, *Rev. Gén. Therm. Fr.* **275** (1984) 738 – 742.
60. R. J. Dihn, J. G. Patel, *Int. J. Hydrogen Energy* **8** (1983) 175 – 182.
61. G. Schaub, R. Reimert, K. Janssen, P. R. Rodriguez, 3rd Int. Conf. on Heavy Crude and Tar Sands, Long Beach, USA, July 1985.
62. W. Schellberg: "Gasification. The Gateway to a Cleaner Future", Conf. Dresden, Germany, 23. – 24.9.1998.
63. M. J. van der Bergt, J. E. Naber, 3rd BOC Priestly Conf., London, September 1983.
64. P. Ameck: "Gasification. The Gateway to a Cleaner Future", Conf. Dresden, Germany, 23. – 24.9.1998.
65. P. F. Curran, R. F. Tyree: "Gasification. The Gateway to a Cleaner Future", Conf. Dresden, Germany, 23. – 24.9.1998.
66. H.-D. Schilling, B. Bonn, U. Krauß: *Kohlenvergasung*, 3rd ed., vol. 22: "Bergbau, Rohstoffe, Energie," Verlag Glückauf, Essen 1982.
67. H.-D. Schilling, B. Bonn, U. Krauß: *Coal Gasification*, Graham & Trotmann, London 1981.
68. R. Specks, *Glückauf* **119** (1983) no. 23, 1147 – 1159.
69. P. Nowacki (ed.): "Coal Gasification Processes," *Energy Technology Review No. 70*, Noyes Data Corp., Park Ridge, N.J., 1981.
70. H. Teggers, H. Jüntgen, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) no. 4, 163 – 173.
71. J. M. Caffin: "Industrial Coal Gasification; Technology, Applications and Economics," *Energy Prog.* **4** (1984) no. 3, 131 – 137.
72. K. Hannes, F. Neumann, W. Thielen, M. Pracht: "Kohlenstaub-Druckverbrennung," *VGB Kraftwerkstechn.* **77** (1997) 393 – 400.
73. A. Dahl: *"Operation experience of the PFBC plant in Värtan, Stockholm,"* Proc. 12th Int'l Conf. on Fluidized Bed Combustion, May 9 – 13, 1993; San Diego, Ca. (USA), Vol. 2, pp. 931 – 935, ASME New York (1993); ISBN 0–7918–0679–0.
74. H. Goto: *"Operation experience from the 71 MW Wakamatsu PFBC demonstration plant,"* Proc. 13th Int'l Conf. on Fluidized Bed Combustion, May 7 – 10, 1995, Orlando, Fl. (USA), Vol. 2, pp. 911 – 918, ASME New York (1995); ISBN 0–7918–1305–3.
75. D. H. Scott, A. M. Carpenter: *"Advanced power systems and coal quality,"* IEA Coal Research, London (1996), IEACR/87, ISBN 92–9029–269–5.
76. A. Carpenter: *"Switching to cheaper coals for power generation,"* IEA Coal Research, London (1998), CCC/01, ISBN 92–9029–301–2.
77. N. M. Skorupska: *"Coal specifications – impact on power station performance,"* IEA Coal Research, London (1993), IEACR/52, ISBN 92–9029–2105–5.
78. A.-K. Hjalmarsson, H. N. Soud: *"NOx control installations on coalfired plants,"* IEA Coal Research, London (1991), IEACR/34, ISBN 92–9029–186–9.

79. H. Soud: *"FGD installations on coal fired plants,"* IEA Coal Research, London (1994), IEACR/71, ISBN 92–9029–239–3.
80. S. Niksa: *"Coal Combustion Modelling,"* IEA Coal Research, London (1996), IEAPER/31, ISBN 92–9029–279–2.
81. C. Di Blasi: "Modelling and simulation of combustion processes of charring and non-charring solid fuels," *Prog. Energy Combust. Sci.* **19** (1993) 71 – 104.
82. J. M. Ekman et al.: *"Cofiring of coal and waste,"* IEA Coal Research, London (1996), IEACR/90, ISBN 92–9029–274–1.
83. J. M. Bemtgen, K. R. G. Hein, A. J. Minchener (eds.): *Combined Combustion of Biomass/Sewage Sludge and Coals*, **vol II** European Commission, DG XII: Science, Research and Development, APAS Clean Coal Technology Programme 1992 – 1994.
84. M. Noer, C. Nielsen, I. Rasmussen, J. C. Clausen: "Clean Coal – Technologische Initiativen in Dänemark," *VGB Kraftwerkstechn.* **74** (1994) 317 – 321.
85. G. H. Cady, *Illinois Geological Survey Report of Investigations*, no. 35, 1935, pp. 25 – 39.
86. H. J. Gluskoter, J. A. Simon, *Illinois Geological Survey Circular*, no. 432, 1968, p. 28.
87. C. N. Kravits, J. C. Crelling, *Int. J. Coal Geol.* **1** (1981) 195 – 212.
88. Jahrbuch der europäischen Energie- und Rohstoffwirtschaft, Essen 2004.
89. Der Kohlenbergbau in der Energiewirtschaft der Bundesrepublik Deutschland im Jahre 2003, Essen und Köln, Nov. 2004.

Coal Liquefaction

Coal, Coal Pyrolysis, and Gas Production are separate keywords.

Takao Kaneko, Mitsubishi Chemical Corp., Takasago Coal Liquefaction Center of Kobe Steel Ltd., Hyogo 676-8670, Japan (Chaps. 1, 3)

Frank Derbyshire, The Pennsylvania State University, University Park, Pennsylvania 16802, United States (Chaps. 1, 3)

Eiichiro Makino, NISSHO IWAI Corp., Coal Liquefaction Project Office, Tokyo 107-8655, Japan (Chap. 2)

David Gray, Mitre Corp., McLean, Virginia 22102, United States (Chap. 2)

Masaaki Tamura, Kobe Steel Ltd., Takasago Coal Liquefaction Center, Hyogo 676-8670, Japan (Chap. 3)

1. Introduction 2
1.1. Commercial Development 3
1.2. Historical Background 5
2. Indirect Coal Liquefaction 13
2.1. Preparation of Synthesis Gas 15
2.1.1. Coal Gasification 15
2.1.2. Water – Gas Shift 17
2.1.3. Synthesis Gas Purification 17
2.2. Fischer – Tropsch Synthesis 17
2.2.1. Mechanisms 20
2.2.2. Product Selectivity and Characteristics 22
2.2.3. Catalytic Selectivity 24
2.2.4. Refining of Crude Products 26
2.2.5. Integrated Plant Configurations 26
2.2.6. Commercial Practice at SASOL 28
2.2.7. New and Developing Technologies . . 29
2.2.8. Advanced Systems 31
2.3. Methanol and Methanol to Gasoline (MTG) 31
2.4. Dimethyl Ether (DME) 33
2.5. Environmental Protection 34
2.6. Economic Aspects 36
3. Direct Coal Liquefaction 37
3.1. Mechanisms 37
3.2. Liquefaction Products 39
3.3. Solvents 40
3.3.1. Hydrogen Donors 40
3.3.2. Coal Conversion and Hydrogen Transfer 42
3.3.3. Interactive Solvent Effects 43
3.3.4. Side Reactions of Solvents 44
3.4. Coal Composition 44
3.4.1. Coal Rank 45
3.4.2. Bituminous Coal 45
3.4.3. Low-Rank Coal 46
3.4.4. Coal Drying and Oxidation 46
3.5. Gaseous Hydrogen 47
3.6. Liquefaction in $CO-H_2O$ and $CO-H_2-H_2O$ 47
3.7. Catalysis 49
3.7.1. Mechanisms 49
3.7.2. Disposable Catalysts 50
3.7.3. Supported Catalysts 54
3.7.4. Molten Halide Catalysts 57
3.8. Process Description 58
3.9. Commercial Processes 59
3.9.1. Kohleoel (Ruhrkohle-Veba Oel) Process 59
3.9.2. NEDOL Process 60
3.9.3. Solvent-Refined Coal (SRC-I and SRC-II) Processes 61
3.9.4. Exxon Donor Solvent Process 62
3.9.5. H-Coal Process 63
3.9.6. Catalytic Two-Stage Process 65
3.9.7. HTI Coal Process 66
3.9.8. BCL Process 66
3.9.9. Liquefaction Solvent Extraction Process 68
3.10. Environmental Considerations . . . 70
3.11. Economic Aspects 70
4. References 71

Abbreviations and Symbols used in this Article

ARGE Arbeitsgemeinschaft
ATTM ammonium tetrathiomolybdate
BCL Brown Coal Liquefaction
BSU bench-scale unit

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_197

CCUJ	Center for Coal Utilization of Japan
CFB	circulating fluidized-bed
CLB	coal liquid bottoms
daf	dry, ash-free
DAO	deashed oil
dmmf	dry, mineral-matter free
EDS	Exxon Donor Solvent
ESP	electrostatic precipitator
FBP	final boiling point
FGD	flue gas desulfurization
FT	Fischer – Tropsch
GAC	granular activated carbon
GFR	gas flow rate
GTL	gas-to-liquid
HDN	hydrodenitrogenation
HDS	hydrodesulfurization
HTFT	high-temperature Fischer – Tropsch
HTI	Hydrocarbon Technologies Incorporated
hvc	high-volatile bituminous coal
IBP	initial boiling point
IGOR	Integrated Gross Oil Refining
LTFT	low-temperature Fischer – Tropsch
maf	moisture and ash-free
MTG	methanol-to-gasoline
NEDO	New Energy and Industrial Development Organization
PDU	process development unit
PP	pilot plant
RHT	radical hydrogen transfer
SAS	SASOL Advanced Synthol
SASOL	South African Coal, Oil, & Gas Corp.
SMDS	Shell Middle Distillate Synthesis
SPD	Slurry Phase Distillate
SRC	Solvent Refined Coal
SSF	SASOL Synthetic Fuels
SSL	single-staged liquefaction
TFB	tubular fixed-bed
TPL	temperature-programmed liquefaction
TSL	temperature-staged liquefaction

1. Introduction

Coal liquefaction is a general term used to describe a process in which a portion of the organic coal substance is converted to liquid products. In the widest sense this includes liquids extracted by solvents. Waxes and resins have been produced commercially by extracting certain peats and low-rank coals with solvents, such as benzene or benzene – methanol mixtures [1]. Solvent extraction has also been used extensively in research on coal structure [2], [3].

More specifically, coal liquefaction involves producing liquids by chemically altering the coal structure and increasing the content of hydrogen relative to carbon. At the same time, nitrogen, sulfur, oxygen, and mineral matter are removed. Compared with other fuels (see Table 1), coal is particularly low in hydrogen. The hydrogen consumed in producing liquids of increased hydrogen content and in removing heteroatoms represents a large part of the cost of coal liquefaction.

Liquefaction methods can be classified as pyrolysis, indirect liquefaction, and direct liquefaction. In *pyrolysis* (→ Coal Pyrolysis), coal is converted at a temperature above 400 °C in a nonoxdizing atmosphere to gases, liquids, and char. The solid char is usually the main product. Liquid yields can be increased by employing a hydrogen atmosphere and by operating under conditions where the volatile products are rapidly diluted to minimize secondary reactions.

In *indirect liquefaction* (see Chap. 2), coal reacts with oxygen and steam at high temperature to produce a mixture of carbon monoxide and hydrogen (synthesis gas), which can be catalytically converted to liquid products. The best-known process, the Fischer – Tropsch (FT) synthesis, produces a mixture of gaseous, liquid, and solid (waxy) hydrocarbons. Other processes of growing importance are methanol syntheses and dimethylether (DME) production, which can offer a more selective route to fuels and chemicals [4–10].

In *direct liquefaction* (see Chap. 3), coal reacts at elevated temperature and pressure with gaseous hydrogen and a solvent that is a hydrogen-donor. Conversion to liquids can be realized with or without added catalysts. According to the conditions chosen, the principal products can be high molecular mass fuels, distillate fuel oils, gasoline, or chemical feedstocks.

The products of both indirect and direct liquefaction have a higher energy content than coal. Coal typically contains 60 wt % of the heating value of the liquefied products. The liquids produced by the two methods are very different in chemical composition. Synthesis gas conversion processes give saturated hydrocarbon liquids, whereas the products of direct liquefaction are highly aromatic. These chemical differences affect the end use. Indirect liquefaction

Table 1. Composition of selected fuels

Fuel	Element, wt %					H : C
	C	H	O	S	N	
Anthracite	93.0	3.1	2.4	0.7	0.8	0.4
Bituminous coal	78.1	5.5	10.0	5.3	1.2	0.8
Crude oil	86.0	11.2	0.6	1.6	0.5	1.6
Fuel oil	86.0	13.3	0.1	0.2	0.1	1.8
Gasoline	85.0	14.9	0.0	0.1	0.0	2.1
Methane	75.0	25.0	0.0	0.0	0.0	4.0

provides high-quality diesel fuel and intermediates for olefin production; direct liquefaction yields high-octane gasoline and excellent feedstocks for aromatic chemicals [11–14]. A hybrid plant concept where both direct and indirect technologies are sited at the same location may have some considerable merits [9], [10].

1.1. Commercial Development

Direct and indirect liquefaction have both been used commercially to produce liquids from coal. The first developments took place in Germany during the 1920s and 1930s. The original Bergius process was developed and modified by I. G. Farbenindustrie. The first module of a commercial plant was started up at Leuna in 1927. Indirect liquefaction was commercialized by Ruhrchemie, which built a plant in 1936 to produce 1000 t/a of liquids by the Fischer – Tropsch synthesis. During this same period, Imperial Chemical Industries (ICI) began work on the hydrogenation of bituminous coal at Billingham-on-Tees (United Kingdom), which led to the commissioning, in 1935, of the only commercial coal hydrogenation plant to be built in that country. The plant was originally designed for an output of 100 000 t/a of liquid fuel, which was later increased by 50 %. The plant operated until 1958.

Intensive efforts in Germany, before and during World War II, led to the construction of large-scale plants to supplement limited indigenous petroleum resources. In 1943, 12 plants were producing liquids by the hydrogenation of coal or coal tar and 9 plants by the Fischer – Tropsch synthesis. Liquids produced from coal by various processes reached a maximum output of over 5×10^6 t/a in 1944. Hydrogenation processes were the biggest producers, attaining a production rate of 3.25×10^6 t/a. Liquefaction by Fischer – Tropsch synthesis was comparatively small at 0.58×10^6 t/a. Coal-derived liquid fuels supplied all of Germany's aviation gasoline and were a major source of motor gasoline, diesel fuel, fuel oil, and some lubricating oils. After the war, three plants continued to operate in the German Democratic Republic until 1965 [5], [15–17]. Kieselguhr supported mixtures of cobalt and thorium oxide was the "standard" catalyst for the German Fischer – Tropsch process using tubular fixed-bed reactors [18].

German efforts laid the foundation for all later research and development. A commercial indirect liquefaction process was established in South Africa by South African Coal, Oil, & Gas Corp. (SASOL) using Fischer – Tropsch technology. The first plant, SASOL I, was brought on stream in 1955 with a rated capacity of ca. 800 m^3/d. Since then two more plants have been put into operation, SASOL II in 1980 and SASOL III in 1982. Each plant has a capacity of ca. 8000 m^3/d. South Africa, which has large coal reserves and no oil, adopted indirect coal liquefaction in order to become independent of petroleum imports [5], [8], [19].

All Fischer – Tropsch reactors at Secunda (SASOL II and SASOL III) were replaced by SASOL Advanced Synthol reactors during 1998 and 1999. As a result of this and other technology improvements the capacity of these plants has increased by about 20 % while simultaneously decreasing operating costs.

The technologies used to convert synthesis gas derived from natural gas to liquid products can in principle also be applied to coal-derived synthesis gas. For example, the SASOL Synthol technology first used with a coal-based plant is also used with synthesis gas derived from natural gas. Two other examples of natural gas-based commercial operations are the

Shell Middle Distillate Synthesis (SMDS) process at Bintulu, Malaysia with a capacity of about 2000 m^3/d [20] and the Mobil technology used for a plant in New Zealand with a capacity of about 2700 m^3/d [21]. The SMDS process uses fixed-bed Fischer – Tropsch reactors similar to those used previously in Germany and by SASOL, but with a more advanced supported cobalt catalyst. The New Zealand plant converts the synthesis gas to methanol and the methanol can then be converted to gasoline using the Mobil MTG (methanol-to-gasoline) process.

The commercial viability of coal liquefaction rests with the overall economy of the process. The conceptual design and economic evaluations performed on the direct liquefaction process in the early 1980s indicated that the price of equivalent crude oil from coal was above $ 50 per barrel. Economic evaluations performed by Bechtel Corporation and Amoco Oil Company (1992 – 1993) indicate that, under the operation conditions in the Catalytic Two-Stage Liquefaction (CTSL) process demonstrated at Willsonville, Alabama, the price of equivalent crude oil produced by direct liquefaction would be about $ 33 per barrel. Advanced liquefaction technologies will result in further reduction in the cost of coal liquids to the level of $ 20 per barrel so that may compete with the cost of petroleum crude by the early 2000s [10].

In 1997, China signed agreements with three countries, the state government, of Nordrhein – Westfalen, Germany (IGOR$^+$ process), Hydrocarbon Technologies Incorporation (HTI) of the United States, and New Energy and Industrial Technology Development Organization (NEDO) of Japan (NEDOL process), to execute feasibility studies for the construction of a direct coal liquefaction plant in China [22–25]. China has changed a net oil importer from an exporter since 1993. Coal liquefaction is an important way for ensuring energy supply according to China's current conditions, it is of practical significance particularly for the provinces with abundant coal and deficient oil resources.

In 1999, NEDO of Japan exchanged a memorandum of understanding on a feasibility study for the construction of a direct coal liquefaction plant, based on the Improved Brown Coal Liquefaction (BCL) process, at the request of the Agency of the Assessment and Application Technology (BPPT) of Indonesia [26]. Indonesia will become a net oil importer in the early 21st century, because of the increasing domestic oil consumption by the recent rapid economic growth. Liquefied oil from brown coal will effectively replace crude oil as a transportation fuel, and thus reduce Indonesia's dependency on imported oil.

These current feasibility studies for Asian countries, collaborated with Germany, the United States, and Japan (see Table 2), may lead to the commercialization of the direct coal liquefaction technology in the early 21st century.

These cases all represent government decisions to supply or underwrite most of the enormous capital requirements. Under normal peacetime conditions, coal liquefaction cannot complete with petroleum and is unattractive to private industry [5]. A notable exception is a privately financed Tennessee Eastman (USA) process for converting coal to acetic anhydride via the production of methanol from synthesis gas. The plant came on stream late in 1983 and gasifies 900 t/d of coal [6].

1.2. Historical Background

The events of major importance in the direct and indirect liquefaction of coal are as follows:

Direct Liquefaction

1869: Berthelot discovered that coal can be converted to oil by chemical reduction [15], [16].

1911: Bergius demonstrated the noncatalytic conversion of coal to heavy crude oil by using a solvent and hydrogen pressure [15], [16].

1925: I. G. Farbenindustrie demonstrated the first large-scale catalytic hydrogenation of coal [15].

1926: I. G. Farbenindustrie modified the Bergius process from one to two reactor stages for gasoline production [15], [16].

1927: I. G. Farbenindustrie started the first module of a commercial liquefaction plant at Leuna, Germany [15], [16], [27]. Pott and Broche initiated research to produce high molecular mass coal extract by using an H-donor solvent under mild

Table 2. Feasibility studies in Asian countries

Feasibility study	China (1997 – 2000)				Indonesia (1999 – 2002)
Organization	Germany NWF state government RUR, RAG	Japan MITI, JICA NEDO, CCUJ	United States DOE HTI	Japan MITI NEDO, PSU	Japan MITI NEDO, Kobe Steel
Coal	Yunnan Xianfeng Coal	Helong Jin Yilan Coal	Shaanxi Shenhua Coal	Shaanxi Shenhua Coal	Sumatra, Banco Coal
Rank	lignite	subbituminous	bituminous	bituminous	lignite
Carbon content	69 %	78 – 80 %	81 %	81 %	71 %
Plant scale	5000 t/d (2500 t/d × 2 trains)	5000 t/d (2500 t/d × 2 trains)	12 000 t/d (4000 t/d × 3 trains)	5000 t/d (2500 t/d × 2 trains)	30 000 t/d (6000 t/d × 5 trains)
Process	IGOR$^+$	NEDOL	HTI-Coal	NEDOL	Improved BCL
First stage	slurry phase	slurry phase	slurry phase	slurry phase	slurry phase
Catalyst	Fe (red mud)	Fe (pyrite)	Fe (Gelcat)	Fe (pyrite)	Fe (limonite)
Conditions	30 MPa, 475 – 490 °C	17 – 19 MPa, 460 – 470 °C	17 MPa, 400 – 460 °C	17 – 19 MPa, 400 – 470 °C	15 MPa, 450 – 460 °C
Second stage	in-line fixed-bed	separate fixed-bed	in-line fixed-bed	separate fixed-bed	in-line fixed-bed
Catalyst	Ni-Mo/Al_2O_3	Ni-Mo/Al_2O_3	Ni-Mo/Al_2O_3	Ni-Mo/Al_2O_3	Ni-Mo/Al_2O_3
References	[20], [22], [23]	[20], [23]	[18], [20], [23]	[20], [23]	[24], [44]

conditions; the plant was built at Welheim, Germany (capacity 26 000 t/a) and operated from 1938 to 1944; some of the extract was used as hydrogenation feedstock [2], [15], [27].

1935: ICI commissioned a liquefaction plant in Billingham, England (original capacity 10^5 t/a); the plant was closed in 1958 [15], [27].
Semicommercial plant was built in Lievin, France (capacity 15 000 t/a). Two coal tar hydrogenation plants were built in Italy (combined gasoline production 24 000 t) [15], [16].

1939: Semicommercial liquefaction plants were built in Manchuria and Korea; both were subsidized by the Japanese government; success was limited [15].

1943: Total of 12 German hydrogenation plants were in operation; 9 in Germany, 2 in Poland, and 1 in Czechoslovakia (max. output 5×10^6 t in 1944) [15], [16], [27].

1949: The United States Bureau of Mines operated a semicommercial plant (50 t/d), based on German technology, in Louisiana, Missouri; the plant was closed in 1954 [15], [16], [27].

1952: Union Carbide developed a coal hydrogenation plant (300 t/d) at Institute, West Virginia for chemicals production; the plant was closed in 1956 [15], [16], [27].

1963: Consolidation Coal Co. developed Consol Synthetic Fuels (CSF) process and investigations continued to 1972; the process involved extraction followed by catalytic hydrogenation; 20 t/d plant at Cresap, West Virginia operated from 1966 to 1970; success was limited and the project was terminated [28], [29].

1963: Consolidation Coal Co. studied molten salt catalyzed liquefaction until 1968; the process involved use of large amounts of molten zinc chloride catalyst to selectively produce distillates; development was resumed in 1975; technical problems have limited progress [28].

United States (developments following formation of OPEC in 1973)

1973: Dow Chemical developed a process using disposable catalyst in a single-stage configuration; high dispersion of molybdenum catalyst was achieved by introduction in the form of an emulsion, allowing operation at low concentration; process was investigated to scale of 90 kg/d; it is not believed to be currently pursued [30].

1974: Chevron Research Co. began research that led to development of a two-stage process, first publicly reported in 1982; thermal liquefaction in the first stage was followed by catalytic hydroprocessing in the second; direct feed of effluent from the

first to the second reactor was claimed to minimize condensation of thermal products; it was scaled to 6 t/d; it is not believed to be currently pursued [31].

1974: Pittsburgh Energy Research Center has developed a catalytic process since 1967 that employed turbulent flow of coal-oil slurry through a fixed-catalyst bed; the operation involved 22 and 180 kg/d units; a proposed 8 t/d pilot plant has been terminated [28], [29].

1974: Gulf Oil operated a solvent-refined coal (SRC) 50 t/d pilot plant at Fort Lewis, Washington until 1981; a single-stage process to produce coal extract by non-catalytic reaction with solvent under hydrogen pressure was used; development was started in the 1960s; a later process variant SRC II was explored from 1977 to maximize distillates [19], [28], [29].

Southern Company Services operated and managed a 6 t/d SRC plant at Wilsonville, Alabama; development was in SRC-I mode to 1981; in 1981 the plant was modified to a two-stage process by addition of catalytic hydrotreater; plant is still in operation [19], [28], [29], [32].

1980: C-E Lummus Co. began operation of a process development unit to study two-stage liquefaction; research dates to 1976; the process combines a short residence time thermal dissolution stage with catalytic hydrotreatment of the higher boiling products in an ebullated-bed reactor; before the second stage, solids are removed by using a light solvent to promote separation and gravity settling; the unit operated until 1984 [33].

1980: A 200 – 600 t/d H-Coal plant at Catlettsburg, Kentucky was operated from May until August, 1982; the unit employed a single-stage ebullated-bed catalytic reactor. Since 1963 Hydrocarbon Research has developed pilot units (up to 3 t/d) using basic concepts derived from H-Oil process [28], [29], [34].

Exxon donor solvent plant (250 t/d) was in operation at Baytown, Texas until August, 1982; the single-stage process used donor solvent that was hydrogenated in a separate reactor; since 1966 Exxon has developed pilot units up to 1 t/d [28], [29], [35].

1982: DOE and the Electric Power Research Institute began developing and testing the two-stage liquefaction process at the 5.5 t/d facility in Wilsonville, Alabama. Two-stage liquefaction went through several process configurations before focusing on close-coupled catalytic two-stage liquefaction. Simultaneously, HRI performed bench scale development (25 kg/d) of close-coupled catalytic two-stage liquefaction [10], [36].

1983: The Wilsonville process development unit was the only remaining US test facility to allow proof-of-concept scale development of direct liquefaction. Operations at this unit ceased in 1992. HRI is now operation a 2.7 t/d unit in Lawrenceville, New Jersey, to continue the development of the direct liquefaction process at a proof of content-scale [36].

1992: Economic evaluations were performed by Bechtel Corporation and Amoco Oil Company (1992 – 1993). Under the high space velocity operating conditions demonstrated at Wilsonville, the price of equivalent crude oil produced by direct liquefaction would be about $ 33 per barrel. If steam reforming of natural gas were used instead of coal gasification for hydrogen production, this price would be reduced to about $ 31 per barrel [10].

1997: HTI signed an agreement with Shenhua Clean Coal Technology Development (SCCT) and China Coal Research Institute (CCRI) to execute a feasibility study for the construction of a Direct Coal Liquefaction plant in China, one of the major projects in China's tenth five-year plan. The project is to built a mine-mouth liquefaction plant (12 000 t/d) using coals (bituminous coals) produced in the Shenhua coal-field in Northern Shaanxi Province and Inner Mongolia near Baotou City. The HTI Coal Process consists primary of two backmixing reactor stages utilizing a proprietary dispersed superfine iron catalyst (GelCat) plus a fixed-bed in-line hydrotreater. The process operates at a pressure of 17 MPa and reactor temperatures in the range of 400 – 460 °C. Phase-l

is a preliminary feasibility study that involves a bench-scale continuous flow unit (CFU) test (30 – 50 kg/d) at HRI and a preliminary economic evaluation and local economic data for the plant. Phase-2 consists of a process development unit (PDU) scale test and a more in-depth economic analysis. The results of the feasibility study will be issued by 1999 – 2000 [22].

Federal Republic of Germany

1976: Bergbau-Forschung operated a 0.5 t/d pilot unit in Essen; the single-stage process used disposable iron catalyst, and was based on the original German process modified to operate at lower pressure by excluding liquids that could not be distilled from the recycle stream; the process is currently in operation to provide support for a 200 t/d plant at Bottrop [5], [37], [38].
Saarbergwerke operated a 0.5 t/d plant at Reden (and, since 1980, a 6 t/d plant at Völklingen); the process configuration was similar to that at Bergbau-Forschung [5], [38].

1978: Rheinische Braunkohlenwerke operated a 0.25 t/d unit at Wesseling for development of brown coal liquefaction; the process was similar to Bergbau-Forschung [38], [39].

1981: Ruhrkohle – Veba Oel commissioned a 200 t/d plant at Bottrop; scaleup was based on pilot unit at Essen, and is currently in operation [37].

1984: The integrated refining concept was initially developed in the PDU at Deutsche Montan-Technologie (DMT), and subsequently up-scaled in the 200 t/d Bottrop coal oil plant in 1985 – 86. Here it was successfully demonstrated through 5000 operating hours until the end of the test series in May 1987 [24].

1987: Ruhrkohle – Veba Oel successfully completed their test runs in the 200 t/d Bottrop Coal Oil Plant, which was subsequently modified to permit hydrogenation of vacuum residues from crude refining. Since 1989, the spectrum of feedstocks had been steadily extended to other chemical residues, e.g., used solvents, waste oil, spent charcoal, and polychlorinated biphenyls.

1992: Coprocessing test runs with mixed waste plastics were started. In 1995, the rather deficit-ridden hydrogenation of vacuum residues was abandoned. In future, only waste plastics will be converted at a rate of 80 000 t/a [24].

1995: Coal-based tests, which were continued in a PDU jointly operated by RAG Umwelt Rohstoffe GmbH (RUR) and DMT to provide technical support and process optimization during the test runs performed in the 200 t/d Bottrop Coal Oil Plant, were accomplished.

1997: The PDU was reactivated to support the work of a feasibility study on a hydrogenation plant to be set up in Yunnann, People's Republic of China. Process basis shall be the test runs on Xianfeng coal (lignite) performed in the proven PDU at the end of 1997 and beginning of 1998. Coal feed stock is a lignite with a carbon content near 70 %. The feasibility study is scheduled for presentation in Beijing in August, 1999 [24].

Japan

1974: The government established the national energy strategy, "Project Sunshine," with imported oil to be supplemented by products from coal [40].

1980: New Energy Development Organization (NEDO) centralized research and development [40].

1981: NEDO decided to construct a 50 t/d plant for brown coal liquefaction in the Latrobe Valley, Australia; the plant, currently under construction, will utilize a two-stage process, using disposable iron catalyst in the first stage and a fixed-bed catalytic reactor in the second stage [41].

1983: Three separate pilot scale processes for bituminous coal liquefaction were integrated into a single development project, using a two-stage configuration, similar to that planned for brown coal liquefaction; the 250 t/d plant is in the design stage [40–43].

1984: Design of a 250 t/d pilot plant was started in 1984, but the plant capacity was reduced to 150 t/d in 1987 [44].

1985: The PSU project started and its operation has been conducted by Nippon Steel, Mitsui SRC Development, and Nippon Coal Oil (NCOL) at Kimitu, Chiba since 1989. The adaptability of various kinds of coal (subbituminous and bituminous coals) and operating conditions were studied and the mechanism of coal liquefaction was being investigated. Improvement of the quality of coal liquid and economic optimization of the NEDOL process were performed until March 1999. The cumulative operation time is ca. 26 000 h using nine types of coal. The operation of a 1 t/d PSU has been carried out for supporting the operation of 150 t/d pilot plant using the same coals and same operation conditions as for the pilot plant since 1994. A series of the examination for expanding the range of applicable coal rank in the NEDOL process was implemented with Adaro, Ikeshima and Yilan coal; the applicable range of carbon content is 73 – 83 wt % [23], [45].

1985: A 50 t/d pilot plant for brown coal liquefaction at Morwel in the Latrobe Valley, Australia was successfully operated from June, 1985 until October, 1990. The Brown Coal Liquefaction (BCL) process has been developed by Nippon Brown Coal Liquefaction Co., Ltd (NBCL) which consists of five companies: Kobe Steel, Mitsubishi Chemical, Nissho Iwai, Idemitsu Kosan, and Cosumo Oil. This project was implemented as the national energy project between the Australia and Japanese government. A 0.1 t/d PDU has been operated at Kobe since 1981 to obtain basic data and to support the operation of a 50 t/d pilot plant. The conceptual design of a demonstration plant and follow-up study were continued until 1993 [46].

1991: NCOL has carried out the design, construction, and operation of the 150 t/d pilot plant of the NEDOL process under the support by NEDO since 1984. Design of a 150 t/d pilot plant was completed, followed by the construction work at Kashima, Ibaragi until June in 1996 [44].

1993: NBCL completed all programs related to the pilot plant experiments, and Takasago Research Laboratory (TRL) was established in Takasago works of Kobe Steel Ltd. to improve the reliability, the economics, and the energy efficiency of the BCL process and to promote the international collaboration on coal liquefaction [46].

1994: New R&D on the improvement of coal liquefaction technologies was started as one of the "New Sunshine Projects" supported by NEDO. Development of improved BCL process was carried out at TRL by using a 0.1 t/d bench scale unit (BSU) until March 1998. Use of a highly dispersed limonite catalyst improves oil yield and process reliability. The quality of recycle solvent and product oils are improved by the use of an in-line hydrotreater. An applicability study on direct coal liquefaction of Banko coal in Indonesia was simultaneously implemented under the Memorandum of Understanding for the research collaboration on coal liquefaction between NEDO and the Agency of the Assessment and Application Technology (BPPT) of Indonesia [26], [46].

1997: The operation of the 150 t/d pilot plant was started in March and continued for two years until September 1998. During the five operation runs, three types of coal with different ranks (Tanito Harum and Adaro coals in Indonesia and Ikeshima coal in Japan) were liquefied without serious problems [44], [47]. Japan International Cooperation Agency (JICA) signed an agreement with China Coal Research Institute (CCRI) to execute a feasibility study on direct coal liquefaction in China. Liquefaction of Yilan coal was performed by using a 1 t/d PSU to support the work of the feasibility study on the direct coal hydrogenation plant to be set up in Helong Jin, People's Republic of China. Feed stock coal is a subbituminous coal with a carbon content near 79 % which is ranked between Tanito Harum and Ikeshima coal. This feasibility study has been completed by the end of 1999 [23], [45].

1998: The R&D activities of NBCL since 1981 were successfully completed in March

and succeeded by the Takasago Coal Liquefaction Center (TCLC) in Kobe Steel Ltd. [46].

1999: NEDO signed an agreement with Shenhua Clean Coal Technology Development (SCCT) to execute a feasibility study on the liquefaction of Shenhua coal Shaanxi in China. Liquefaction tests were performed by using a 1 t/d PSU to support the work of the feasibility study on the construction of the coal liquefaction plant. This feasibility study has been completed by March 2000 [23], [45].

1999: NEDO also exchanged a Memorandum of Understanding on the feasibility study for a direct coal liquefaction plant at the request of the Agency of the Assessment and Application Technology (BPPT) of Indonesia. Along with the collaborative activities, the feasibility study was conducted by Kobe Steel, Ltd. supported by NEDO. The major objective of this study is to ascertain the suitability of Indonesian low-rank coal (lignite) for the use with the improved BCL process through a selection of plant location, a conceptual design of a commercial plant followed by a financial analysis. This feasibility study is scheduled for three years until March 2002 [45], [48].

United Kingdom

1976: The National Coal Board (NCB) investigated two liquefaction processes at Stoke Orchard, Gloucester; in one, coal was extracted in a heavy solvent without hydrogen or catalyst, and after filtration, the extract was catalytically hydrocracked to distillates; in the other, coal was extracted in an aromatic solvent under supercritical conditions, with extract considered as a feedstock for hydrocracking; both processes were developed to a scale of 50 kg/d and plans were made to build two 24 t/d pilot plants; the construction of a 2.5 t/d unit using a liquid solvent extraction process was planed [5], [16], [48], [49].

1985: The next stage in the development of the NBC's Liquid Solvent Extraction (LSE) process was to design, build and operate a 2.5 t/d process development pilot plant. This project was approved in October by BCC's Board and agreements reached with the Department of Energy and the EEC. Ruhrkohle Oel und Gas GmbH and Amoco corporation joined the project in 1987 and 1991, respectively [50].

1992: A 2.5 t/d pilot plant (LSE process) was operated until 1995 at Point of Ayr, North Wales. Point of Ayr coal (bituminous coal) was chosen as the standard coal for use in experiments from 1981 on [50].

Indirect Liquefaction

1902: Sabatier and Senderens hydrogenated carbon monoxide over nickel to produce methanol [51].

1913: BASF was awarded a patent covering production of hydrocarbons and oxygenates by catalytic hydrogenation of carbon monoxide at high pressure over cobalt catalysts [52].

1923: Franz Fischer and Hans Tropsch reported their work on the hydrogenation of carbon monoxide over alkaline iron catalysts at 10 – 15 MPa and 400 °C [53], calling their mixture of oxygenates and hydrocarbons "synthol".

1923: BASF and I. G. Farben were awarded patents covering zinc-based, iron-free catalysts for the synthesis of methanol.

1926: Fischer and Tropsch were awarded a patent describing the synthesis of hydrocarbons over iron catalysts at 0.1 MPa and 240 – 300 °C [54]; however, activity of these catalysts under these conditions and conversions of synthesis gas to oil was low [16].

1928 – 1930: P. K. Frohlich and W. K. Lewis worked on the synthesis of higher alcohols using alkalized $ZnO - Cr_2O_3$ catalysts [55].

1936: The first four German Fischer – Tropsch (FT) plants were commissioned, having a total capacity of 200 000 t/a of hydrocarbons [56]; the catalyst used was based on cobalt, having the approximate composition by mass Co : ThO_2 : MgO : kieselguhr of 100 : 5 : 8 : 200 [57].

1936 – 1937: Fischer and Helmut Pichler discovered that hydrocarbon synthesis over iron catalysts was greatly improved at a higher operating pressure

(0.5 – 2 MPa), making alkalyzed iron catalysts a possible replacement for the cobalt-based systems; the lifetime of iron catalysts was greatly improved under these conditions [57].

1939 – 1944: The Germans embarked on a research and development program to perfect an iron-based catalyst to replace the cobalt catalysts being used in existing Fischer – Tropsch plants; reactor development was intensified during this period, including fixed-bed reactors with recycle, as well as slurry reactors [55].

1943 – 1944: Herbert Kölbel coordinated a series of tests held at Schwarzheide to screen suitable iron catalysts to replace the cobalt catalysts being used in existing reactors [58]; tests were successful but iron catalysts were not used in the German production plants during World War II [55], [57].

1944: Nine Fischer – Tropsch plants were operated in Germany, producing ca. 700 000 t/a [57]; all plants used the Ruhrchemie synthesis process and the cobalt – thorium – magnesium catalyst [16].

1947 – 1950: Ruhrchemie and Lurgi formed the so-called Arbeitsgemeinschaft (ARGE) to develop a multitubular, fixed-bed Fischer – Tropsch reactor using a precipitated alkaline iron catalyst, which produced high wax yields [57].

1950: Hydrocarbon Research of Trenton, New Jersey developed a fixed, fluidized-bed reactor to produce FT hydrocarbons from reformed natural gas [55]; the plant, known as Carthage Hydrocol, was located in Brownsville, Texas; in 1953 the plant was shut down and dismantled because of the high cost of natural gas [55].

1953: H. Kölbel and P. Ackerman directed the successful operation of a 10 t/d, pilot-scale, slurry-phase Fisher – Tropsch reactor system at Rheinpreussen–Koppers in Germany; this reactor used a precipitated iron FT catalyst suspended in a heavy liquid heat transfer medium through which synthesis gas was bubbled.

1955: The first SASOL plant (SASOL I) was commissioned in South Africa (construction began in 1952); this indirect liquefaction plant used Lurgi dry-ash gasifiers to produce the synthesis gas from coal and two types of F – T synthesis processes: (1) ARGE fixed-bed process, developed by Ruhrchemie and Lurgi, (2) circulating entrained-flow reactor developed in the United States by the M. W. Kellogg Co.; both synthesis reactors used iron catalysts; the fixed-bed FT unit operated as expected, but the entrained system was inoperable commercially until substantial modifications were made by SASOL; extensive entrained-flow catalyst research was necessary before a viable catalyst with acceptable life and selectivity was formulated [57]; not until 1960 did the circulating entrained-flow FT system first achieve steady-state, reliable, continuous operation (system is now known as the SASOL Synthol system).

1966: ICI of the United Kingdom developed a low-pressure process for production of methanol from synthesis gas using a copper – zinc catalyst in a gas-cooled reactor.

1975: SASOL decided to build another much larger plant known as SASOL II, at Secunda, South Africa [59]; the SASOL II plant came on stream in 1980 and produced ca. 8000 m^3/d of liquid fuels; SASOL II used Lurgi dry-ash gasifiers to gasify the coal and Synthol circulating entrained-flow Fischer – Tropsch synthesis units to produce liquid hydrocarbons.

1976: Mobil developed a process for converting methanol to gasoline by using a shape-selective zeolite catalyst known as ZSM – 5; this process was known as the Mobil methanol-to-gasoline (MTG) process.

1979: SASOL decided to build SASOL III, a mirror image of SASOL II, in South Africa because of the continuing deterioration in the crude oil situation [59]; this plant came on stream in 1982.

1981: A plant was initiated in New Zealand to produce methanol from abundant cheap natural gas and to convert this methanol to gasoline by using the MTG process; this plant, designed to produce 2700 m^3/d of gasoline, came on stream in 1985 [21].

1987: The South African Government announced that the Mossgas project would be established for the production of synthetic fuels from two gas fields on the south of Mossel Bay. The offshore platform supplied the first gas in 1992 and the onshore plant including a Synthol reactor started production with full capacity in 1993.

1989: The first commercial SASOL Advanced Synthol (SAS) reactor, a conventional fluidized-bed high-temperature Fischer – Tropsch reactor, was brought into operation at Sasolburg, South Africa. The reactor capacity was about 530 m^3/d. This development was preceded by the operation of a demonstration-scale reactor. Both the demonstration-scale reactor and the commercial reactor were subsequently modified to operate as slurry-phase reactors [60].

1990: SASOL operates a demonstration-scale slurry-phase bubble-column reactor using an iron-based catalyst to convert coal-derived synthesis gas to liquid products. The plant is located in Sasolburg, South Africa.

1992: Slurry-phase Fischer – Tropsch technology was demonstrated at the LaPorte Alternative Fuels Development Unit, Texas, sponsored by the DOE, Air Products, Exxon, Shell Oil, Statoil, and UOP. DOE's advanced catalyst development program includes iron- and cobalt-based catalysts with specific goals for improved catalytic activity, selectivity, attrition resistance, and stability. An efficient catalyst – wax separation system is being devoted to prevent catalyst loss and to guarantee product quality. For oxygenates development, several different technologies are being developed concurrently for the conversion of coal-derived synthesis gas into marketable oxygenate fuels and chemicals, such as alcohols, ethers, and olefins. Air Products successfully developed and demonstrated a one-step process for synthesizing dimethyl ether from coal-derived synthesis gas [10].

1993: A Commercial slurry-phase reactor was commissioned at SASOL I in Sasolburg and the Synthol reactors ceased operate at this site. The slurry-phase reactor uses an iron based catalyst and produces 400 m^3/d of liquid products with the same product spectrum as the ARGE reactors. At the same time an 800 t/d ammonia plant was added to the site as well as expanding the town gas production [61].

1993: The Mossgas plant was commissioned. The plant converts natural gas to liquid fuels using the Synthol technology licensed from SASOL. Total conversion capacity installed amounts to about 4000 m^3/d with three reactors.

1993: Shell starts up a commercial plant using four reactors to convert synthesis gas derived from natural gas to mainly middle distillate products at Bintulu in the State of Sarawak, Malaysia. The plant capacity is about 2000 m^3/d. The plant has operating flexibility to also produce low molecular mass paraffins and waxes in addition to middle distillate fuels and naphtha. The production rate of the former products can be varied to match the smaller market demand [20].

1999: All the circulating fluidized-bed (CFB) Synthol reactors at SASOL II and SASOL III at Secunda, South Africa are replaced with new SASOL Advanced Synthol (SAS) reactors. The 16 CFB reactors each with a capacity of 1200 m^3/d are replaced by 4 SAS reactors each with a capacity of 1700 m^3/d and a further 4 SAS reactors each with a capacity of 3200 m^3/d, giving a total installed capacity of 19 600 m^3/d [62].

1999: SASOL proceeds to basic engineering for a 5000 m^3/d plant to convert natural gas to liquids at Escravos, Nigeria in a joint venture with Chevron. SASOL's Slurry Phase Distillate (SPD) Process licensed to the venture is based on the operation of a supported cobalt catalyst using coal-derived synthesis gas in the demonstration slurry-phase reactor at the Sasolburg site in South Africa. SASOL and Chevron announce a global alliance to use and license the SASOL SPD Process for future gas to liquids (GTL) projects.

1999: A DME pilot plant with a capacity of 5 t/d was commissioned at Taiheiyo Coal

Mining's site in Hokkaido. NKK Corporation started to establish the direct synthesis technology for the low-cost mass production of DME [63].

Although the commercial-scale operation of direct coal liquefaction plants, had been demonstrated in Germany, coal-derived liquids could only be produced at high costs. Incentives for further research were limited until the 1970s, primarily due to the availability of vast reserves of inexpensive oils discovered in the Middle East in the late 1940s. The two oil crises in 1973 and 1978 sky-rocketed the price of crude oil, which jumped to $ 40/bbl at maximum. Since then, developed countries have paid great efforts to reduce the ratio of petroleum in the energy consumption.

Coal is a major oil substitute (see Table 3), its clean use technologies have been pursued and developed. Coal liquefaction is expected to supply liquid fuels for transportation. The price of petroleum is now stable after the success of developed countries to reduce the petroleum ratio, and this leads to sufficient supply of petroleum at present. No necessity of coal liquefaction may be recognized right now. In the longer term, the differences between the rates of consumption and the reserves of the different energy forms will inevitably provide the incentive for more widespread and efficient use of coal. The studies in coal liquefaction (direct and indirect) are being planned and implemented with the potential of being key technology processes developed as one of the modules that will be combined into highly flexible energy complexes "Vision 21" in the United States [64], [65].

A rapid economic growth in East Asia with a large population requiring a huge amount of energy, especially transportation fuels, and limited growth of petroleum supply are anticipated to bring about a significant gap between fuel supply and demand. Coal liquefaction is expected to be developed to supply commercially the liquid fuels for transportation at acceptable cost.

In China or Indonesia, the dominant position of coal in the primary energy production and consumption will not change for a considerably long time period. Direct coal liquefaction is an important way to make up for the oil reserves in those Asian countries which have cheap and abundant coal resources suitable for liquefaction. The current feasibility studies, collaborated with Germany, the United States, and Japan, will provide a reliable basis for the construction of direct coal liquefaction plants in commercial scale in the near future.

2. Indirect Coal Liquefaction

Indirect liquefaction of coal is a stepwise process (see Fig. 1). The coal first reacts with steam and oxygen (gasification) to produce a raw mixture of carbon monoxide and hydrogen known as synthesis gas or *syngas.* The water – gas shift reactor adjusts the ratio of hydrogen to carbon monoxide to the requirements of a particular synthesis by means of a reversible reaction between steam and carbon monoxide:

$$H_2O_{(g)} + CO \rightleftharpoons H_2 + CO_2 \tag{1}$$

After purification, the clean syngas is catalytically converted to a wide range of products, such as hydrocarbons, alcohols, aldehydes, ketones, and acids. Production of predominantly liquid hydrocarbons from syngas is known as the Fischer – Tropsch synthesis, which is a commercial process for converting coal into substitute natural gas (SNG), gasoline, diesel oil, wax, and alcohols. Production of hydrocarbons and alcohols by the Fischer – Tropsch synthesis can be represented as follows:

$$2n\,H_2 + n\,CO \longrightarrow C_nH_{2n} + n\,H_2O \tag{2}$$

$$(2n+1)\,H_2 + n\,CO \longrightarrow C_nH_{2n+2} + n\,H_2O \tag{3}$$

$$2n\,H_2 + n\,CO \longrightarrow C_nH_{2n+1}OH + (n-1)\,H_2O \tag{4}$$

where n is an integer.

Reaction (2) represents formation of olefins, reaction (3) of alkanes, and reaction (4) of alcohols. Hydrocarbon products are in the gasoline boiling range (ca. 50 – 180 °C), but not directly suitable for gasoline use because straight-chain structures predominate, giving an intermediate octane number as low as 55 – 65. The intermediate octane number is defined as the average of the research and motor octane numbers. Extensive further treatment is needed to meet octane requirements for gasoline (ca. 87 – 95). However, the fraction that boils between ca. 180 and

Table 3. Worldwide fuel demands, 10^6 t

Area	Crude oil				Natural gas				Coal			
	1995	2010	2020	Annual growth rate 1995 – 2020	1995	2010	2020	Annual growth rate 1995 – 2020	1995	2010	2020	Annual growth rate 1995 – 2020
OECD	1832	2159	2262	0.8 %	840	1199	1301	2.3 %	904	1097	1219	1.2 %
North America	873	1025	1050	0.7 %	489	608	579	0.6 %	501	649	835	2.1 %
Europe	650	779	850	1.1 %	280	475	593	2.1 %	282	314	245	−0.6 %
Pacific	309	354	361	0.6 %	71	117	130	1.8 %	121	133	138	0.5 %
Non-OECD	1363	2135	3749	2.9 %	706	1130	1665	2.3 %	1301	2013	2556	2.7 %
Transition economics	275	329	391	1.4 %	447	578	750	1.6 %	300	357	360	0.7 %
Africa	97	145	180	2.5 %	26	51	76	2.6 %	82	112	137	2.1 %
China	164	356	506	4.6 %	14	49	70	3.2 %	664	1087	1416	3.1 %
East Asia	264	472	639	3.6 %	46	110	180	3.0 %	85	146	219	3.9 %
South Asia	99	191	278	4.2 %	31	81	146	3.2 %	140	256	348	3.7 %
Latin America	282	424	520	2.5 %	70	141	244	2.9 %	25	44	59	3.5 %
Middle East	183	219	281	1.7 %	73	120	200	2.5 %	5	13	18	5.0 %
Total	3195	4294	5056	1.9 %	1546	2329	2966	1.9 %	2205	3110	3775	2.2 %

* Ref.: World Energy Outlook, International Energy Agency, 1998 Edition.

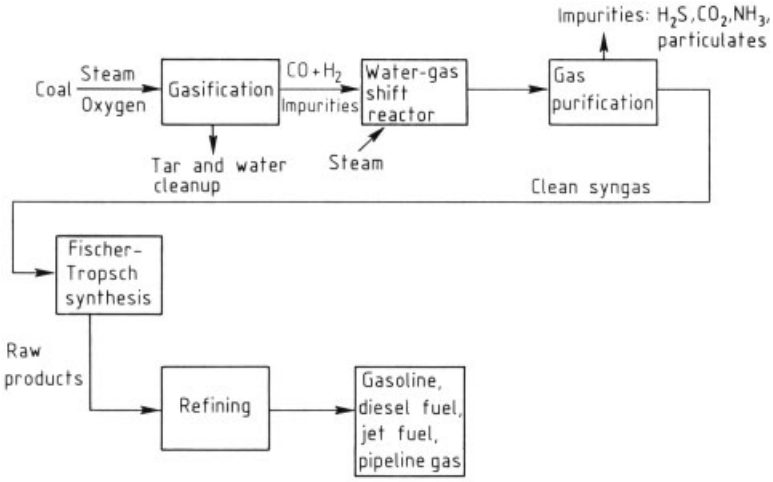

Figure 1. Flow sheet for indirect liquefaction process

320 °C has a high cetane number (65 – 75) and provides valuable diesel fuel with little or no additional refining.

A variant of the Fischer – Tropsch synthesis gives methanol in yields that, unlike the hydrocarbon synthesis, approach 100 %. Methanol can be used as turbine and transportation fuel, although its energy content on a volume basis is only half that of gasoline. It can also be converted almost quantitatively to high-octane gasoline by the Mobil methanol-to-gasoline (MTG) process, which uses the highly selective, synthetic zeolite catalyst ZSM-5.

The primary step in the development of indirect liquefaction is the production of clean synthesis gas; thus, development of coal gasification parallels the success of indirect liquefaction. Synthesis gas for the German Fischer – Tropsch plants during World War II was produced from low-temperature cokes gasified in conventional

water – gas generators [16]; development of coal gasification was also proceeding at that time. In the early 1930s, the fixed-bed coal gasifier was being developed jointly by Lurgi Mineralöltechnik, and Sächsische Werke. In 1936, the first commercial Lurgi plant was built at Hirschfelde in Germany for the production of town gas by lignite gasification [66]. The first commercial Winkler fluidized-bed gasification plant was built in 1926 at Leuna, Germany [67]. By 1949, the first commercial Koppers – Totzek (K – T) entrained gasification plant was in operation at Marzingarbe, France [66].

The most reliable commercial coal gasifier available in 1950 was the fixed-bed, dry-ash Lurgi system, which was chosen for the SASOL I plant in South Africa. With extensive improvements, this system has been used in all subsequent SASOL plants.

Continuing advances in gasification in the United States and elsewhere since the 1950s promise to increase the efficiency of indirect liquefaction. The most noteworthy improvements in coal gasification technology have occurred in entrained-flow systems (Shell and Texaco), the agglomerating-ash fluidized system (Kellogg – Rust – Westinghouse, KRW), and the slagging-ash, fixed-bed system (British Gas Corp., BGC).

Currently, the three SASOL plants operating in South Africa represent the only commercial examples of indirect coal liquefaction by the Fischer – Tropsch synthesis in the world. These plants produce a total of more than 20 000 m^3/d of liquid products (supplying over 30 % of South Africa's liquid transportation fuel needs) by using more than 46×10^6 t/a of low-grade coal. Part of this coal is gasified and another part is used to generate the steam and power for the process. SASOL Synthetic Fuels (SSF) produced about 6.8×10^6 t of products in 1998 with a market value about fivefold that of the coal used. SSF's products comprise liquid fuels (sulfur-free gasoline, diesel, and liquefied petroleum gas), pipeline gas, and various chemical and polymer feedstocks, including alcohols, ketones, organic acids, and other solvents, ammonia, sulfur, ethylene, propene, and α-olefins.

The indirect liquefaction of coal provides high-octane gasoline and high-cetane diesel fuel equal or superior to the conventional products of petroleum refining so that no modifications are required to existing distribution systems or existing engines. Indirect liquefaction also allows coal to be used in an environmentally acceptable manner.

The following steps are required for the indirect liquefaction of coal to transportation fuels:

1) Gasifying coal by using oxygen and steam to produce raw synthesis gas
2) Adjusting hydrogen : carbon monoxide ratio (water – gas shift) for compatibility with the FT catalyst system
3) Purifying the raw synthesis gas and removing catalyst poison and excess carbon dioxide
4) Converting the carbon monoxide and hydrogen to hydrocarbons by FT synthesis
5) Refining the raw FT products to produce marketable liquefied petroleum gases, high-octane gasoline, diesel fuel, solvents, and olefins for petrochemical use.

2.1. Preparation of Synthesis Gas

2.1.1. Coal Gasification

Coal gasification represents a thermal balance between the endothermic reactions of carbon with steam and carbon dioxide and the exothermic reaction of carbon with oxygen. Additional processes occurring in gasifiers include coal devolatilization, the water – gas shift, and methanation [68].

	ΔH, kJ/mol
Gasification	
$C + H_2O \longrightarrow CO + H_2$	135.1
$C + 2\,H_2O \longrightarrow CO_2 + 2\,H_2$	96.2
$C + CO_2 \longrightarrow 2\,CO$	173.2
Combustion	
$C + 1/2\,O_2 \longrightarrow CO$	−110.5
$C + O_2 \longrightarrow CO_2$	−393.7
Devolatilization	
Coal $\longrightarrow$ Char + Volatiles	
Water – Gas Shift	
$CO + H_2O \longrightarrow CO_2 + H_2$	−38.5
Methanation	
$CO + 3\,H_2 \longrightarrow CH_4 + H_2O$	−217.6

Gasification technology has advanced greatly since the World War II German indirect liquefaction plants. These plants produced gas with blue water – gas sets that used alternating cycles of "blows," during which the coke in the generator was heated with air; and "runs" or gas-producing periods, when steam was passed through the bed

to produce carbon monoxide and hydrogen [69]. The net lower heating value (LHV) efficiency of this process was ca. 50 %. The net efficiency (LHV) is defined as the energy content of the produced gas divided by the sum of the energy content of the coke consumed plus the energy used to produce steam.

The commercial SASOL processes in South Africa employ the reliable dry-ash Lurgi fixed-bed gasifiers [70], which use low-grade or high-ash coals and operate at high pressure with low oxygen consumption and high carbon conversion. They do, however, suffer from serious disadvantages, including high steam consumption and the production of methane, tars, oils, and phenols. Furthermore, they cannot handle agglomerating coals or fines, but require a coal feed of uniform size.

The performance of different systems is compared in Table 4 [69–74]. The BGC demonstration gasifier, with a coal throughput of ca. 550 – 730 t/d, is located at Westfield, Scotland [70]. The Westinghouse (KRW) gasifier has been operating on a pilot-plant scale in Waltz Mill, Pennsylvania since 1976 with a capacity of ca. 20 t/d [72], [73]. The Shell (formerly Shell – Koppers) gasifier has been operating since 1979 at Hamburg, Federal Republic of Germany, with a capacity of 140 t/d [74].

The net efficiency of synthesis gas production is important, but it is not the only criterion in gasifier selection. Figure 2 compares the output of four gasifiers in terms of product energy; input coal is the total amount of coal used in the indirect liquefaction plant. Fixed-bed gasifiers burn some of the coal in boilers to provide high-pressure process steam. Dry-ash and slagging Lurgi gasifiers use 18 – 20 wt % of the total coal supply for steam production, i.e., only 80 % can be gasified. Fluidized-bed Westinghouse and entrained-bed Shell gasifiers gasify all the coal. Waste heat boilers use the exit gas for steam production, which is essential for improved process efficiency, because 15 – 20 % of the energy is contained in the sensible heat of the exit gas.

Fixed-bed gasification produces considerable quantities of methane and ethane, which are inert in the Fischer – Tropsch synthesis. If exclusively liquid fuels are required, these gases must be reformed to carbon monoxide and hydrogen and recycled, with considerable loss in efficiency. Only 61 % of the energy content of the methane and ethane appears in the liquid products of reforming and resynthesis [71].

Coal pyrolysis in the upper portion of the fixed bed forms tars, oils, and phenols representing more than 10 % of the energy content of the coal feed. These impurities must be removed before further gas treatment.

The ratio of hydrogen to carbon monoxide in the exit gas is an important consideration in the Fischer – Tropsch synthesis. The conventional FT synthesis used at SASOL requires ratios of 1.7 and 2.5, depending on conditions. More advanced gasification systems have low H_2 : CO ratios (see Table 4) and thus, require adjustment by the water – gas shift.

2.1.2. Water – Gas Shift

Modern catalysts for the water – gas shift are resistant to poisoning by sulfur compounds; this ability to process sulfur containing gases is referred to as sour or raw-gas shift. With the older nickel or iron catalysts poisoning was prevented by preliminary gas purification. Commercially important shift catalysts include [68]:

1) High-temperature shift catalysts (440 – 700 °C), e.g., chromium – iron oxide or zinc – chromium oxide
2) Low-temperature shift catalysts (230 – 350 °C), e.g., copper – chromium oxide
3) Raw gas shift catalysts (at 180 – 550 °C), e.g., cobalt – molybdenum sulfide on aluminum oxide or magnesium oxide

2.1.3. Synthesis Gas Purification

Gasifier exit gas contains impurities that interfere with synthesis. Some impurities can be removed mechanically by water washing or by cyclones. Coal dust, ash, tars, and oils can be removed by cooling and scrubbing. Water washing also removes ammonia and phenols. The remaining impurities include hydrogen sulfide, carbonyl sulfide, carbon dioxide, and small quantities of hydrogen cyanide. The upper limit of sulfur allowable in the Fischer – Tropsch and methanol syntheses is 1 ppm (by vol) [68];

Table 4. Gasification performances

Parameter	Blue water-gas	Dry-ash Lurgi	BGC Lurgi	KRW	Shell
Gasification pressure, MPa	atmospheric	2.4	2.4	2.8	4.5
Exit gas temperature, °C	700	540	450	980	1420
Gasification type	water – gas set	fixed-bed	fixed-bed	fluidized-bed	entrained
Input ratios of reactants					
Steam: coal, daf [a]	1.5 [b]	1.34	0.33	0.37	0.03
Oxygen: coal, daf [a]	–	0.36	0.39	0.80	0.83
Raw gas composition, mol %					
H_2	47.3	23.0	21.9	24.9	30.9
CO	37.0	11.1	43.8	32.7	64.4
CH_4	1.3	6.7	5.2	4.3	0
C_2	0.0	0.3	0.3	0	0
CO_2	5.4	17.6	2.1	16.6	2.4
H_2O	–	41.0	26.3	20.8	1.4
Others	9.0	0.3	0.3	0.8	0.9
H_2: CO ratio	1.28	2.07	0.50	0.76	0.48
Net efficiency (LHV)	50.0 [c]	65.0	76.0	74.0	83.0

[a] Dry, ash-free.
[b] Steam to dry coke ratio.
[c] Based on dry coke.

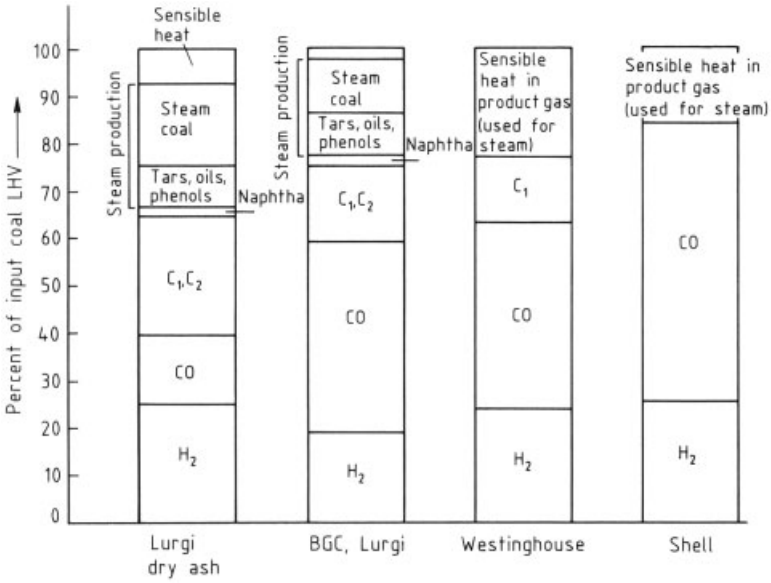

Figure 2. Output of different gasifiers

larger amounts rapidly poison the catalysts. Carbon dioxide is less critical and can, in fact, actually improve catalyst performance in the Fischer – Tropsch synthesis.

Hydrogen sulfide and carbon dioxide are acidic and can be removed by physical or chemical absorption [75], [76]. Physical adsorption, following Henry's law, is favored by low temperature and high pressure.

The removal of acid gas may be nonselective or selective. In the former hydrogen sulfide and carbon dioxide are simultaneously removed in a single stage. If acid gas is removed after shift, the gas stream consists predominantly of carbon dioxide, with a few percent of hydrogen sulfide. Selective gas removal produces one stream containing hydrogen sulfide and the other, carbon dioxide [75]. Sulfur can be recovered from the

selective system by the Claus sulfur-recovery process [76].

The Rectisol process is licensed by Lurgi and employed in the SASOL complexes in South Africa [77], [78]; methanol is the adsorbent at low temperature. The Selexol process, licensed by Allied Chemical, uses poly(ethylene glycol) dimethyl ether [77], [78]. The Lurgi Purisol process uses *N*-methyl-2-pyrrolidine (NMP) [77], [78], and the Fluor solvent process uses propylene carbonate [78].

Chemical absorption systems primarily use amines or carbonates to neutralize acidic gases:

$$2\,RNH_2 + H_2O + CO_2 \longrightarrow (RNH_3)_2 \cdot CO_3$$

$$RNH_2 + H_2S \longrightarrow RNH_3 \cdot HS$$

$$RNH_2 + HCN \longrightarrow RNH_3 \cdot CN$$

Heat reverses these reactions, liberating the gases [67].

Carbonates are neutralized as follows:

$$K_2CO_3 + H_2S \longrightarrow KHCO_3 + KHS$$

Carbonyl sulfide (COS) is hydrolyzed to CO_2 and H_2S [77]; the latter is absorbed by the carbonate solution.

The principal chemical absorption processes are Benfield and Catacarb, both potassium carbonate systems [67], and diisopropanolamine (ADIP) [67], *N*-methyldiethanolamine (MDEA) [67], Sulfiban [67], Fluor Econamine [78], and monoethanolamine (MEA) [78], all amine systems.

After purification and shift of the hydrogen: carbon monoxide ratio, the synthesis gas can be converted to hydrocarbons.

2.2. Fischer – Tropsch Synthesis

The Fischer – Tropsch synthesis is a nonselective process that produces a wide range of compounds containing from 1 to > 100 carbon atoms. Only C_1 compounds can be produced with 100 % selectivity. Selectivity can also be high for high molecular mass wax. However, for intermediate products, only certain maximum values are possible; in general selectivities are interrelated [57], [79]. All postulated mechanisms assume that the synthesis occurs by a stepwise growth of hydrocarbon chains, one carbon being added at a time. A probability mechanism of this chain growth, which limits production of gasoline and diesel fuel hydrocarbons (see Section 2.2.3), can be calculated.

At SASOL I in South Africa, iron-based catalysts promoted with potassium and copper are used. One type of iron catalyst is used for the low-temperature Fischer – Tropsch (LTFT) processes used at SASOL I (i.e., ARGE and slurry phase processes) and another one in the high-temperature Fischer–Tropsch (HTFT) process, i.e., the Synthol Process used at Secunda and Mossel Bay in South Africa. A precipitated iron catalyst is used with the LTFT processes while a fused iron catalyst is used with the HTFT processes. Supported cobalt catalysts can also be considered for LTFT processes.

Fischer – Tropsch processes can be used to produce either a light synthetic crude oil (syncrude) and light olefins or heavy waxy hydrocarbons. The syncrude can be refined to environmentally friendly gasoline and diesel and the heavy hydrocarbons to specialty waxes or, if hydrocracked and/or isomerized, to produce excellent diesel fuel, lube oils, and a naphtha, which is an ideal feedstock for cracking to olefins. To give some idea of the flexibility of the potential product spectrum, the gasoline:diesel ratio may vary from 80 : 20 to 20 : 80. In addition, for a total of 50 % potential olefinic petrochemical feedstock, the light olefins (C_2 to C_4) can amount to 25 % of the products, a further 25 % of the product being in the form of olefins in the carbon number range C_5 to C_{10}. On the other hand, waxes can comprise more than 50 % of the total products.

For production of gasoline and light olefins, the FT process is operated at high temperature (330 – 350 °C), for production of waxes and/or diesel fuel, at low temperatures (220 – 250 °C). Traditionally the high-temperature FT process was performed in Synthol circulating fluidized bed (CFB) reactors and the low-temperature FT process in ARGE tubular fixed-bed (TFB) reactors. In the 1990s, much better reactor systems have been developed for both HTFT and LTFT processes. The conditions of both processes are compared in Table 5.

For HTFT the SASOL Advanced Synthol (SAS) reactor with solid – gas fluidization was developed. The SAS reactor gives very much

the same product spectrum as the CFB reactors, but much more effectively and cheaply.

Table 5. Process conditions for Fischer – Tropsch reactors

Parameter	LTFT	HTFT (SAS)
Temperature, °C	220 – 250	330 – 350
Pressure, MPa	2.5 – 4.5	2.5
$CO + H_2$ conversion, %	60 – 90	85

For LTFT, the SASOL Slurry Phase Distillate (SPD) reactor, of the bubble column type, was developed which is a significant improvement over the TFB reactor used in the ARGE process [61], [80]. In the Slurry Phase reactor products can be made with the same carbon chain length distribution as in the TFB reactor with Schulz – Flory distribution α-values of 0.95 and higher (for a definition of α-values see Section 2.2.1). The SPD reactor has a greater flexibility with respect to product distribution than the TFB reactor.

The heavy hydrocarbons from the SASOL SPD process are hydrogenated and converted to specialty waxes. Alternatively these waxy hydrocarbons can be very easily hydrocracked to excellent diesel fuel with a cetane number > 70 and negligible aromatics and zero sulfur content. At 780 kg/m^3 its density is somewhat lower than that of conventional diesel. If hydroisomerized, the waxy hydrocarbons can be used to produce lube oils.

The primary product spectra of the HTFT process are discussed in more detail in Section 2.2.2. The olefins from SASOL's HTFT plants are recovered commercially and ethylene, propene, 1-hexene, and 1-octene are produced more cheaply than from crude oil. The production of butenes, heptenes, and higher olefins in a HTFT plant is also being considered.

Low-Temperature Fischer – Tropsch Reactors. The two types of LTFT reactors are described below [81].

Tubular Fixed-Bed Reactor. The tubular fixed-bed reactor which has been operated by SASOL (using an iron catalyst) since 1953 and by Shell in Bintulu, Malaysia (using a supported cobalt catalyst) since 1993, makes use of tubes into which the catalyst is paced. Syngas is passed downward through the catalyst bed and is catalytically converted to hydrocarbons. The heat evolved is removed through the tube walls to produce steam on the shell side of the TFB reactor.

Slurry Bed Reactor. The slurry bed reactor which was successfully developed and operated by SASOL commercially since 1993, consists of a vessel containing a slurry of process-derived wax with catalyst dispersed in it. Syngas is bubbled through this slurry bed and is converted to hydrocarbons. The heat generated is passed from the slurry to the cooling coils inside the reactor to generate steam.

The light hydrocarbons, which are in the vapor phase, are removed from the freeboard at the top of the SSPD reactor with the unconverted reactants and are condensed in the downstream condensing train. The heavier liquid hydrocarbons are mixed into the slurry from which they are removed in a proprietary solid separation process developed by SASOL. Excellent results are obtained by this separation step.

The conventional TFB ARGE reactor and the state of the art slurry phase reactor are shown in Figure 3. The maximum capacity for a slurry phase reactor is an order of magnitude higher than that of the TFB reactor. The slurry phase reactor is much simpler in construction than the TFB reactor. The suspended cooling coils and a gas distributor give a much cheaper arrangement than the tube and tube sheet arrangement in the TFB reactor. Because of this, the slurry phase reactor lends itself much better to scale-up. At high gas throughputs, the TFB reactor becomes very complex and mechanically difficult to manufacture and the pressure drops across the tubes become excessive. The pressure drop across the TFB reactor varies from 0.3 – 0.7 MPa, depending on operating pressure, whereas it is typically less than 0.2 MPa across the slurry phase reactor. These differences make it possible to design (at present operating pressures of 2.5 MPa) single slurry phase reactors with capacities of 3000 m^3/d.

The most significant advantage of the slurry phase reactor is that it is well mixed and can be operated isothermally. The absence of axial and radial temperature gradients as in a TFB reactor, allows much higher average operating temperatures in the slurry phase reactor and hence higher reaction rates. Together with the more effective contact between the syngas and the catalyst particles, this leads to higher production rates for the same overall reactor dimensions even though

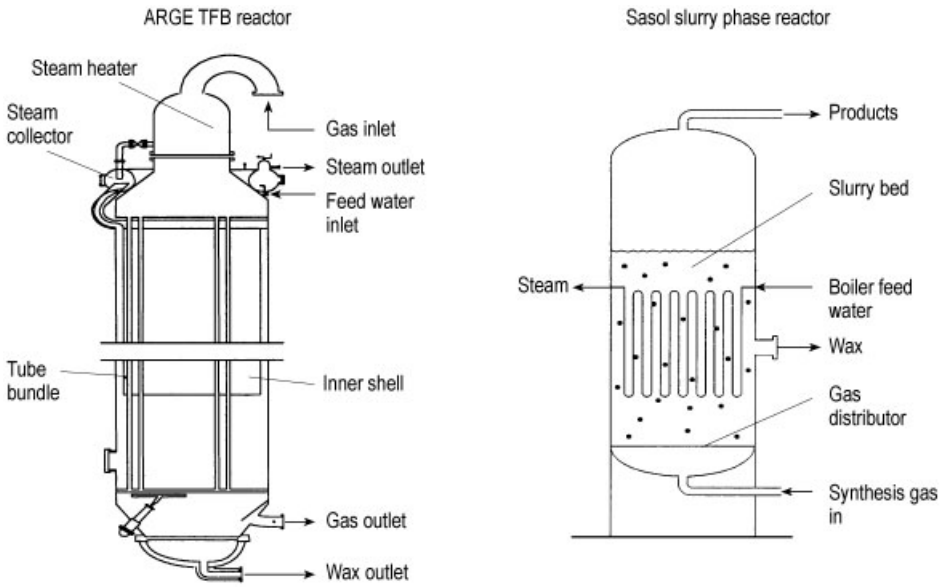

Figure 3. SASOL reactor system for LTFT

the catalyst inventory in the slurry phase reactor is lower than in the TFB reactor. Because all the catalyst particles are flexibly mixed with the slurry at more optimal process and conditions, the yields per reactor volume are higher and the catalyst consumption per ton of product is only 20 to 30 % of that of the TFB reactor. Because of the isothermal nature of the reactor, the easier control of average catalyst life, and the much smaller pressure drops across the reactor, the control of the reactor is much simpler and operating costs are significantly reduced.

High-Temperature Fluidized Bed Reactors. Also two types of HTFT reactors exist, which are described below.

Circulating Fluidized-Bed Reactor. Three Synthol CFB reactors were used commercially by SASOL from 1955 and another sixteen from the early 1980s on. For this type of reactor, fused iron catalyst is circulated with synthesis gas through a complex reactor – hopper – standpipe system and heat is removed as steam through coils suspended in the reactor section. The reactor system is complex and needs a complex support system to cope with the circulating catalyst loads and temperature differences which makes it expensive. The large tonnages of catalyst circulated cause relatively high pressure drops across the reactor system. These and other disadvantages are eliminated when using a SAS reactor which makes use of conventional solid – gas fluidization.

SAS Reactor. The Synthol CFB and SAS reactors are shown in Figure 4. The SAS reactor is a vessel containing a fluidized bed consisting of fused and reduced iron catalyst. Syngas is bubbled by means of a gas distributor through the bed where it is catalytically converted to hydrocarbons which at the process conditions of about 340 °C and 2.5 MPa, are in the vapor phase. The products and unconverted gases leave the reactor through internal cyclones. The SAS reactor is operated such that the cyclones can retain the catalyst so effectively that, in contrast to the CFB reactors, scrubber towers are not needed to remove the last traces of catalyst before the product stream is passed to the condensing train.

The major advantages of the SAS reactor over the CFB reactor are its simplicity, ease of operation, lower operating cost due to elimination of the catalyst recycle and lower maintenance, and in general higher conversions at higher gas loads. The latter together with the fact that more cooling coils can be installed in the SAS reactor and more heat can be removed, allows for larger capacity equipment which translates into advantages of economy of scale. Maximum capacities

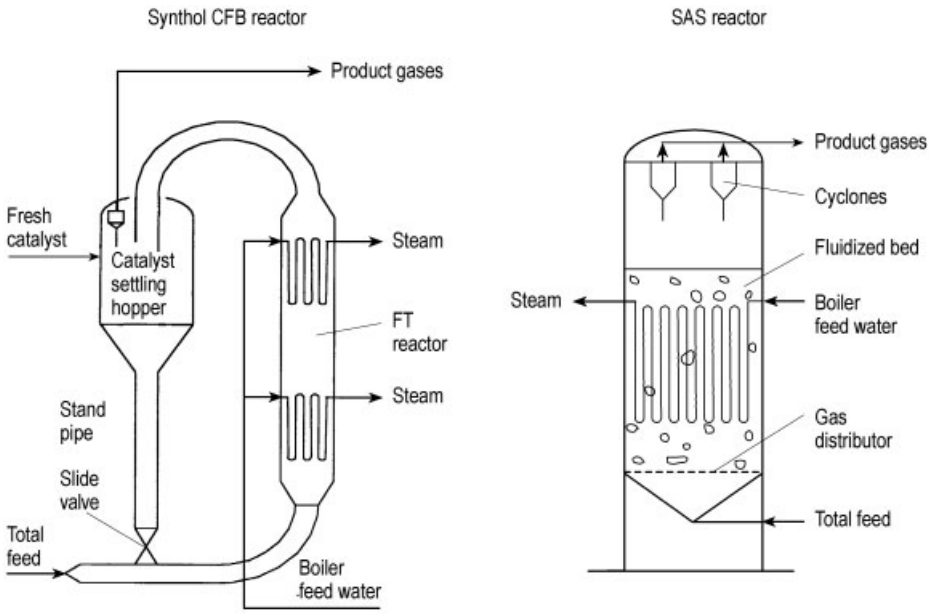

Figure 4. SASOL reactor system for HTFT

of 3200 m^3/d per reactor have been installed at the Secunda site.

2.2.1. Mechanisms

The mechanism of the Fischer–Tropsch synthesis is uncertain, but three theories have been developed [57], [82], [83].

In the first, the carbon monoxide dissociates at the catalyst surface (**M**) to carbon and oxygen atoms. Chemisorbed hydrogen reacts with the carbon to form CH_x entities, which combine into hydrocarbon chains. Chain termination produces α-olefins or alkanes (see reaction scheme 1). To account for the formation of oxygen compounds, a parallel mechanism postulates their production on oxide components of the catalyst surface [83].

In the second theory, carbon monoxide does not dissociate, but is hydrogenated to form oxymethylene species, which condense on the catalyst surface to propagate the hydrocarbon chain (reaction scheme 2).

In the third theory, the CO molecule does not dissociate, but inserts in an **M**–H or an **M**–C bond (reaction scheme 3) [84].

Reaction scheme 1

Initiation

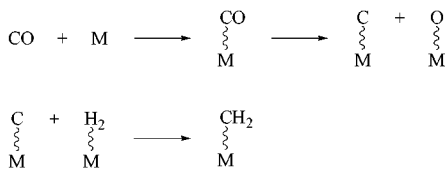

Propagation

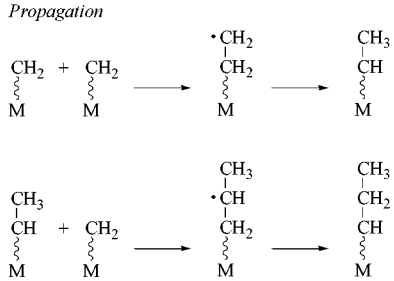

Termination

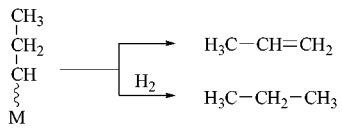

denotes a surface bond of indefinite order

Table 6. Fischer – Tropsch product selectivities, wt %

Product	SASOL I		SASOL II and III	
	ARGE [79]	ARGE [87]	Synthol [57]	Synthol [68]
CH_4	2.0	5.0	10	11.0
C_2H_4	0.1	0.2	4	7.5
C_2H_6	1.8	2.4	4	7.5
C_3H_6	2.7	2.0	12	13.0
C_3H_8	1.7	2.8	2	13.0
C_4H_8	2.8	3.0	9	11.0
C_4H_{10}	1.7	2.2	2	11.0
C_5–C_{11}, gasoline	18.0	22.5	40	37.0
C_{12}–C_{18}, diesel	14	15.0	7	11.0
C_{19}–C_{23}	7	6.0	4	3.5
C_{24}–C_{35}, medium wax	20	17.0	4	3.5
C_{35}, hard wax	25	18.0	–	–
Nonacid chemicals	3	3.5	5	6.0
Acids	0.2	0.4	1	NA ∗
Total	100.0	100.0	100.0	100.0

∗ NA = not available.

This mechanism has been widely accepted because CO insertion is known to occur with homogeneous catalysts [57]; however, the evidence is inconclusive for heterogeneous systems. A generalized mechanism that incorporates all three of these schemes has been proposed [59] (see nextpage).

Table 7. Product selectivities (carbon basis) for SASOL reactors, wt %[85]

Product	LTFT	SAS
Methane	4	7
C_2 – C_4 olefins	4	24
C_2 – C_4 paraffins	4	6
Gasoline	18	36
Middle distillates	19	12
Heavy cut and waxes	48	9
Water soluble oxygenates	3	6
Breakdown of the main oxygenate components for the SAS reactor (mass %)		
Main nonacid chemicals		
Acetaldehyde		3
Acetone		10
Ethanol		55
Methyl ethyl ketone		3
Isopropanol		3
n-Propanol		13
Isobutanol		3
n-Butanol		4
Acids		
Acetic acid		70
Propionic acid		16
Butyric acid		9
Valeric acid and higher		5

2.2.2. Product Selectivity and Characteristics

The ARGE fixed-bed reactor operates at relatively low temperature and is intended to produce a large quantity of wax and a minimum of methane. The objective of the Synthol reactor, on the other hand, is to maximize production of materials in the boiling range of gasoline. As mentioned previously, formation of waxy material must be avoided or catalyst particles stick together, resulting in bed defluidization. Product selectivities obtained with the SASOL fixed-bed ARGE system and Synthol reactors are shown in Tables 6 and 7. The two sets of data for the ARGE system probably represent different catalyst ages.

Table 8 shows the typical product composition for LTFT and HTFT reactors with iron catalyst [85].

Because of the higher operating temperature, Synthol gasoline contains less paraffinic and more olefinic, aromatic, and oxygenated components than ARGE gasoline. The latter has a high percentage of *n*-alkanes and a low research octane number (RON) of 35, but the diesel has a high cetane number of 75. Synthol gasoline, with its high olefinic and some aromatic content, has a higher RON of 65, but an octane number that is low for gasoline. The Synthol diesel fuel has a cetane number of 55.

Reaction scheme 2

Initiation

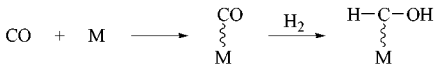

Propagation

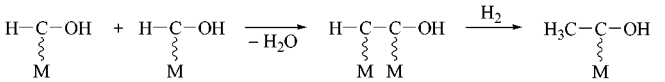

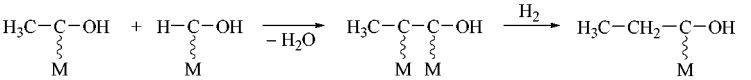

Termination

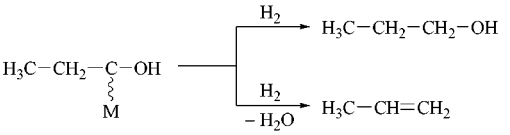

denotes a surface bond of indefinite order

Reaction scheme 3

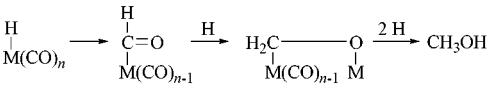

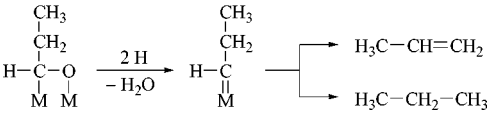

Table 8. Typical product compositions for LTFT and HTFT reactors (with iron catalyst) [85]

	TFB reactor		SSPD reactor		HTFT	
	C_5-C_{12}	$C_{13}-C_{18}$	C_5-C_{12}	$C_{13}-C_{18}$	C_5-C_{10}	$C_{11}-C_{14}$
Paraffins, %	53	65	29	44	13	15
Olefins, %	40	28	64	50	70	60
Aromatics, %	0	0	0	0	5	15
Oxygenates, %	7	7	7	6	12	10
n-Paraffins, %	95	99	96	95	55	60

The C_3 Synthol hydrocarbons are 86 % olefinic, compared to 60 % for those from the ARGE process. For the C_4 hydrocarbons, the corresponding figures are 82 % and 56 %. The high percentage of C_3 and C_4 olefins is a great advantage, because these olefins produce more gasoline by catalytic polymerization and alkylation.

The selectivities obtained in the SAS reactors are very similar to those in the Synthol CFB reactors because both processes use a catalyst with the same composition. A slightly lower methane selectivity can be attained in the SAS reactor because it is operated more isothermally.

The Slurry Phase Distillate process is designed to produce predominantly middle distillate fuels such as gasoline or diesel. At the same time a small amount of naphtha is manufactured as coproduct. These fuels are compatible with existing petroleum derived fuels, and thus no engine modifications or special distribution infrastructure are necessary [86]. At the same time, this synthetic diesel has several environmental advantages over conventional fuels. It has superior combustion characteristics due to its high cetane number (> 70), and is free of sulfur and aromatic compounds (see Table 9).

Environmental concerns have already led to specifications on fuels being tightened in several regions. This is expected to continue, resulting in more stringent specifications in the future. Some refiners may face difficulties in meeting these specifications due to either equipment or feedstock limitations. The environmental advantages of the synthetic product will in this case result in synthetic diesel selling at a premium rate compared with conventional diesel (expected at $ 2.4/bbl premium) for use as a blending component.

Table 9. Comparison of the quality for diesel fractions [86]

	SPD process	Conventional
Sulfur	< 10 ppm	500 ppm
Aromatics	< 1 %	10 – 25 %
Cetane number	> 70	45

The emissions from diesel engines operating on standard, "reformulated", and Fischer – Tropsch diesel fuels have been measured by the South West Research Institute, a large independent fuels and lubricants testing organization in the USA (see Table 10). These tests showed that Fischer – Tropsch diesel results in reduced exhaust emissions (particulates, nitrogen oxides, carbon monoxide, and unburned hydrocarbons) compared to both the standard and the "reformulated"diesels [86].

Table 10. Comparison of Fischer – Tropsch diesel in SWRI diesel engine emission tests [86]

Reduction in	Average diesel	Reformulated diesel
Hydrocarbons	33 %	20 %
CO	46 %	36 %
NO_x	8 %	4 %
Particulates	30 %	26 %

The coproduction of some chemicals is possible with the Slurry Phase Distillate process. The products are primarily a mixture of linear paraffins and olefins, with a wide range of carbon chain lengths. It is feasible to recover linear paraffins that can be used for the manufacture of detergent alkylates and other chemical uses. The separation of linear olefins for the production of detergent feedstock, and the production of lubricating oil base stocks, are also possible. The recovery of these fractions is economically justified for a large scale facility consisting of several modules, as these cuts form only a small fraction of the overall production [86].

2.2.3. Catalytic Selectivity

Fischer – Tropsch processes include catalytic carbon – carbon and carbon – hydrogen bond formation and carbon – oxygen bond rupture [82]. Whatever the mechanisms, the resulting hydrocarbon chain distribution can be explained by a polymerization process, adding one carbon atom at a time to the growing alkyl chain. At any point in the chain, growth can be terminated by hydrogen atom addition or by α-hydrogen abstraction, followed by desorption from the catalyst surface. This process results in a distribution of chain lengths, as in all polymerization processes. Product distribution can be predicted accordingly. The Schulz – Flory distribution can be used to predict selectivity, assuming that all hydrocarbon species on the catalyst surface, regardless of chain length, add a carbon atom with equal probability to produce a chain that is one unit longer [88], [89]. For a hydrocarbon chain having n carbon atoms, the probability of poly-

merization at the first carbon and at each successive carbon is α; the probability that polymerization does not occur at the final carbon atom is $(1-\alpha)$. Thus, the probability that the hydrocarbon chain is composed of n carbon atoms is:

$$P_n = \alpha^{n-1}\,(1-\alpha)$$

Therefore, P_n is equivalent to the mole fraction of the nth oligomer. If the assumption is made that the molecular mass of the monomer is the same, independent of position, the mass fraction (W_n) is simply the mole fraction multiplied by the chain length n and divided by 1 $(1-\alpha)$ (the average degree of polymerization). The equation becomes:

$$W_n = n\alpha^{n-1}\,(1-\alpha)^2$$

Rearranging and taking logarithms gives the Schulz – Flory equation:

$$\log\frac{W_n}{n} = n\log\alpha + \log\frac{(1-\alpha)^2}{\alpha}$$

When $\log\frac{W_n}{n}$ is plotted against n, a straight line of slope $= \log\alpha$ should be obtained if the Schulz – Flory polymerization mechanism is obeyed. With increasing α, the product contains higher molecular mass hydrocarbons. The chain growth probability distribution results in a broad Fischer – Tropsch product spectrum. In addition, only C_1 hydrocarbons can be produced with 100 % selectivity (i.e., when $\alpha = 0$). The Schulz – Flory distribution places a significant limitation on the maximum yield of the transportation fuel fraction. Selectivity as a function of α is shown in Figure 5 [57]. Chain growth is influenced by temperature, pressure, feed composition, and catalyst [55], [90], [91].

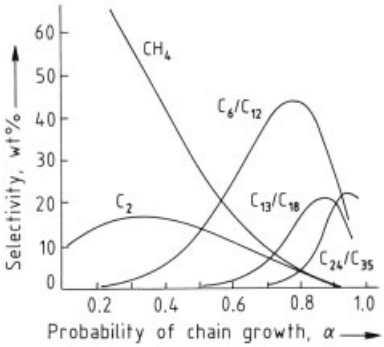

Figure 5. Calculated product selectivities from Schulz – Flory analysis

The SASOL operation uses a precipitated iron catalyst in the LTFT reactors (ARGE and Slurry Phase) and a fused iron catalyst in the HTFT Synthol reactors. The fixed-bed catalyst is activated with hydrogen under mild conditions to produce the large metallic surface area necessary for high activity [57]. Selectivity is apparently influenced by the activation or conditioning process. Copper increases the rate of iron oxide reduction and the activity of the catalyst, but not its selectivity for wax [57]. Synthesis gas converts the metallic iron in the partly reduced catalyst to a carbide form (Fe_5C_2) known as Hagg carbide [57]. Promoters, such as K_2O, increase α, i.e., selectivity for longer hydrocarbons, and production of acid. Catalytic activity passes through a maximum with increasing K_2O content. Among support materials for the precipitated iron catalyst, silica gives both the highest activity and the highest wax selectivity [57].

The Synthol catalyst is prepared from mill scale (waste iron) containing small amounts of impurities, such as SiO_2, TiO_2, MgO, and Al_2O_3. The catalyst and promoters are electrically fused into ingots, which are crushed to the necessary particle size and then reduced with hydrogen in fixed fluidized-bed reactors. The surface area increases according to the degree of reduction to a final value between 2 and 30 mg^2/g; pore volume also increases [57], [58]. Potassium is the most important promoter [59]. Structural promoters (Al_2O_3, MgO, and Cr_2O_3) increase and stabilize the surface [57].

Increased temperature in the Fischer – Tropsch synthesis reduces α and favors lower molecular mass products, the chemical characteristics of which also vary (see Tables 11 and 12) [57], [59].

Table 11. Influence of temperature on selectivity of Synthol catalysts [57], [59]

Temperature, °C	Methane selectivity,	Ratio $C_3H_6 : C_3H_8$	Aromatics in gasoline, wt %
310	10	11	4
330	14	10	8
350	17	9	10
360	20	8	13
370	23	6	18
380	28	4	26

Table 12. Influence of temperature on selectivity of ARGE catalysts [57], [59]

Temperature, °C	Wax selectivity, %	Olefins in diesel, wt %
213	47	45
227	34	39
237	24	33
247	17	40

Gas composition and pressure also influence catalyst selectivity [57], [59], probably because, at the catalyst surface, hydrogen favors chain termination and carbon monoxide favors chain propagation. This suggests a correlation between product selectivity and the ratio of hydrogen to carbon monoxide in the feed gas. Wax formation in the fixed-bed reactor is in fact correlated with the H_2 : CO ratio; lower values favor polymerization and higher values favor wax production (higher α). The total reactor pressure, however, has little or no effect on selectivity [59].

This correlation is not observed with the fluidized-bed Synthol catalyst system. At SASOL methane selectivity has been found to correlate with the function:

$$\frac{P_{H_2}^{0.5}}{(P_{CO}+P_{CO_2}+P_{H_2O})}$$

Sulfur poisoning decreases the activity of the Synthol catalyst but increases the formation of olefins and higher hydrocarbons [92]. The structure of the Synthol catalyst, under synthesis conditions, changes continuously from the freshly reduced catalyst to the fully mature catalyst [62]. The main reason for this change is carbon formation during the synthesis process. The change in morphology of the catalyst can affect the catalyst's product selectivity after low-level sulfur poisoning.

2.2.4. Refining of Crude Products

Crude Fischer – Tropsch products from Synthol and ARGE units require extensive refining for use as gasoline and diesel fuel (see Table 13).

Table 13. Refining of Fischer – Tropsch products

Refining process	Purpose
FT product fractionation	separation into light gases and feeds for polymerization and hydrotreating
FT product hydrotreating	saturation of olefins and destruction of alcohols and acids
Hydrotreated product fractionator	separation of hydrotreated products into pentane–hexane stream for isomerization, C_7-C_{11} for reforming, and diesel and fuel oil products
Catalytic reforming	aromatization to increase octane number of gasoline stream
C_5-C_6 Isomerization	isomerization of straight chains to increase octane number of hexanes and pentanes
Catalytic polymerization	polymerization of propene and butene to gasoline
Alkylation	catalytic alkylation of isobutane and unpolymerized C_3-C_4 olefins to increase gasoline yield
Polygas hydrogenation	saturation of olefins in heavy catalytic polygasoline
Alcohol recovery	preparation of marketable alcohol mixture
Wax hydrocracking	cracking of wax components into gasoline and diesel hydrocarbons

Figure 6 illustrates an indirect liquefaction plant producing transportation fuels. If only Synthol reactors are used, as at SASOL II and III, the wax hydrocracker is eliminated. Wax produced in ARGE units can be selectively cracked to $C_1 - C_4$ hydrocarbon gases, gasoline, and diesel in the proportions 5 : 15 : 80 (by mass) [93]. SASOL II and III operations resemble those described in Table 13 [94] (see Section 2.2.6).

2.2.5. Integrated Plant Configurations

The flow sheet of an integrated indirect coal conversion plant, combining the steps discussed in previous sections, is shown in Figure 7. In the upper configuration, designated as the *mixed-output mode,* liquid transportation fuels and substitute natural gas (SNG) are produced. In the lower configuration, the *all-liquid-output mode,* the light hydrocarbon gases are autothermally reformed back to synthesis gas; the net output contains only C_3 and higher liquid hydrocarbons. The SASOL II and III operations are of this

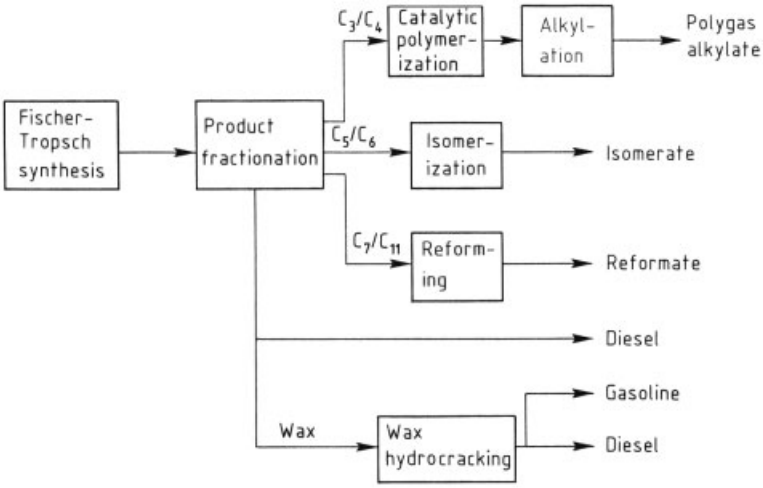

Figure 6. Flow sheet for Fischer – Tropsch products

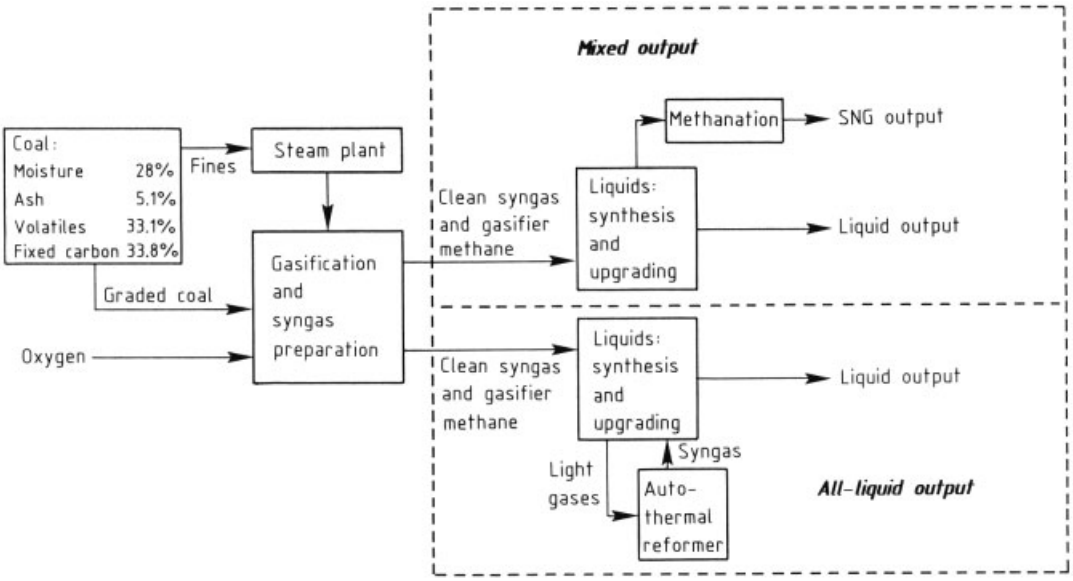

Figure 7. Flow sheet for integrated liquefaction

type. The flow sheet was adapted from studies conducted by Mobil Research and Development Corp. [71] and Mitre Corp. [70].

The plant represented in Figure 7 is designed to process 25 200 t/d of Wyoming subbituminous, noncaking reactive coal. To meet plant requirements, ca. 18 % of the coal is needed to produce high-quality steam, most of which is used for oxygen production for the gasification section. Fixed-bed, dry-bottom Lurgi gasifiers produce the synthesis gas, which is very wet, with a molar hydrogen : carbon monoxide ratio of ca. 2 : 1. After being quenched with water to remove tars and particulates, the synthesis gas is split into two streams. About 20 % is passed to shift reactors, where the hydrogen : carbon monoxide ratio is raised to 10 : 1. Recombination of the two streams gives an overall H_2 : CO ratio of 2.5 : 1, as required for Synthol reactors. The combined gas is purified by the Lurgi Rectisol process, passed to the Synthol units, and refined (see Section 2.2.4).

Table 14. Calculated yields from model liquefaction plants

Products	Mixed output	All-liquid output
SNG, m^3/d (STP)	4.91×10^6	–
Gasoline, m^3/d	2160	4465
Propane, m^3/d	176	387
Butane, m^3/d	23	51
Diesel, m^3/d	324	807
Fuel oil, m^3/d	99	217
Alcohols, m^3/d	291	640
Annual thermal output (LHV ∗), GJ/a	85 627 000	65 418 000
Liquid yield, m^3 C_4^+/t dry coal	0.159	0.334
Overall efficiency (LHV ∗), %	54	41
Total plant coal (LHV ∗), GJ/a	158 424 000	158 424 000

∗ LHV = lower heating value.

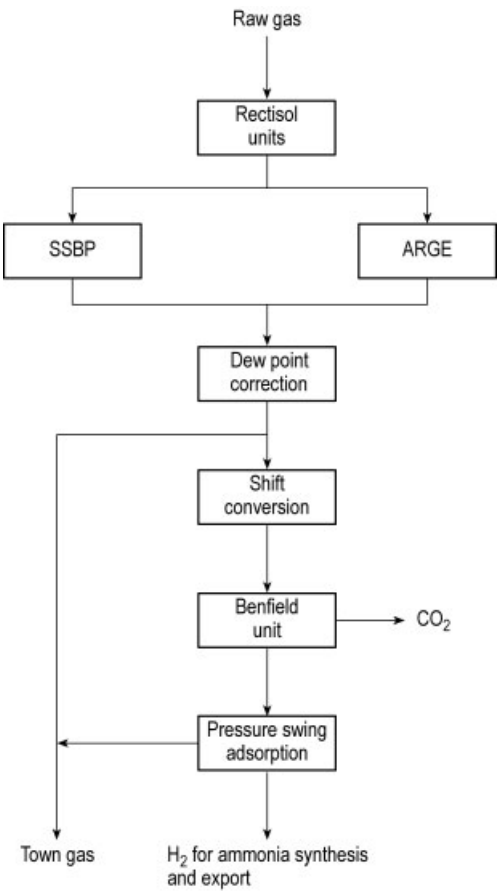

Figure 8. Flow sheet for SASOL I synthesis

The gasification process includes auxiliary units for gas – liquid separation and recovery of phenol, sulfur, and ammonia. These units contribute to pollution control (see Section 2.5).

Off-site units are required for oxygen production, steam boilers and superheaters, stack gas precipitation and flue gas desulfurization, coal handling and preparation, ash removal, water supply for boilers and cooling towers, power generation, wastewater treatment, storage, fuel blending, and catalyst preparation.

The calculated yields based on a model commercial plant for mixed-output and all-liquid operation are given in Table 14. In the mixed-output mode, 65 % of the thermal output is in the form of SNG, which originates from the Lurgi gasifier and also from the Synthol reactor. In the all-liquid mode, the methane and ethane are reformed and resynthesized to liquids, reducing overall plant efficiency from 54 to 41 %. Overall plant efficiency is defined as the LHV energy content of all the products divided by the LHV energy content of the total feed coal.

Energy relations in the all-liquid mode as a percent of total plant coal input energy (LHV) are as follows:

Input coal	100 %
Gasifier output	74 %
Synthesis gas	63 %
Clean gas	62 %
Raw products	45 %
Refined products	41 %

The raw gasifier output represents ca. 74 % of the energy of the total feed coal and includes the tars, oils, and phenols recovered from the wash and wastewaters. A calculation based only on gases produced would reduce the efficiency to ca. 65 %. After the shift and cleaning steps, the energy remaining in the gas represents 62 % of the total input coal. In the all-liquid mode, the methane and ethane are reformed to synthesis gas with a loss of energy; thus, the crude products represent only ca. 45 % of the input coal energy. Refining to final products lowers this value to 41 %.

2.2.6. Commercial Practice at SASOL

The SASOL plants, in operation in South Africa since 1956, are the only indirect coal liquefaction plants in the world that currently produce liquid transportation fuels by the Fischer – Tropsch synthesis. The original Sasol-

burg site has been converted entirely to produce products other than liquid transportation fuels, mainly waxes, paraffins, ammonia, and town gas (see Fig. 8).

The Secunda site consists of two identical plants previously called SASOL II and SASOL III but now referred to as SASOL West and SASOL East. These plants are virtually identical and they now also produce a wide range of petrochemical products and a methane-rich pipeline gas.

There are 36 gasifiers at each plant (72 in total at Secunda). The Mark-IV Lurgi dry-ash gasifiers used have 3.85 m internal diameter. They were originally designed to produce ca. 55 000 m^3/h (STP) of dry raw gas [59], but are now capable of producing about 85 000 m^3/h. The raw gas from these gasifiers is cooled in waste-heat boilers and washed with water. Particulates, ammonia, tars, oils, and phenols are removed. The cooled, stripped gas, still containing hydrogen sulfide and carbon dioxide, is purified by a chilled methanol wash in Rectisol units, leaving less than 40 ppb (by volume) of H_2S and less than 1 % CO_2.

The product work-up system used for the Secunda plants is shown in Figure 9. The tail gas from the separator is scrubbed to remove carbon dioxide and then fractionated in a cryogenic unit. The recovered ethylene from this cryogenic unit is a valuable chemical feedstock [59]. Some pipeline or town gas is withdrawn from the methane-rich product stream from the cryogenic unit and the remainder is autothermally reformed to synthesis gas and recycled. The propene was originally polymerized to either gasoline or diesel in a catalytic unit that can also accept butene as a feedstock if higher diesel yields are desired. Nowadays the propene is mostly recovered as a valuable petrochemical feedstock and the butene is processed to gasoline in an alkylation unit.

As shown in Figure 9, pentane and hexane are isomerized and the $C_7 - C_{11}$ hydrocarbons are reformed over a platinum catalyst to increase the octane rating of the gasoline. Higher carbon number fractions are catalytically hydrodewaxed and marketed as a zero sulfur diesel fuel [94]. Increasingly α-olefins such as 1-hexene and 1-octene are being recovered as petrochemical feedstocks.

2.2.7. New and Developing Technologies

The conventional fluidized-bed reactor concept has been developed and scaled-up by SASOL to the limits of this technology at the Secunda site in South Africa. This type of reactor was first operated at the Hydrocol plant in Brownsville, Texas in the early 1950s. The tubular fixed-bed reactor concept used in Germany in the early applications (maximum production in 1944) has now been developed further and scaled up by Shell to the limits of this technology for the commercial plant in Bintulu, Malaysia.

The slurry phase or slurry bubble column reactor (the successor to the TFB reactor for LTFT) is potentially the best reactor choice for the Fischer – Tropsch synthesis and it has yet to be developed to its full potential in a commercial application. This type of reactor was first applied in a demonstration unit at Rheinpreussen, Germany, in the 1950s using an iron-based catalyst for the production of gasoline. SASOL, Statoil, Exxon, Shell, and, more recently, IFP/AGIP have been working on the concept of using a supported cobalt catalyst with this type of reactor. SASOL and Exxon have operated demonstration-scale reactors, and Shell and Air Products have participated in demonstration scale tests using the US DOE facilities at La Porte, Texas. The concept is to produce wax as the main primary product from the reactor which also produces straight-run diesel and naphtha and optionally kerosene and liquefied petroleum gas (LPG). Only SASOL has operated a commercial-scale slurry phase reactor. The characteristics of the Rheinpreussen plant were as follows [95]:

Reactor dimensions	155 cm diameter 8.6 m height
Slurry volume	10 000 L
Catalyst, Fe	800 kg
Production output	11.5 t/d
Average temperature	268 °C
Pressure	1.2 MPa
Feed gas rate at STP	2700 m^3/h
Hourly space velocity of gas	270 L of gas per liter of expanded slurry
Carbon monoxide conversion	91 mol %
Overall syngas conversion	89 mol %
Syngas feed ratio H_2 : CO	0.67
Linear velocity of syngas	9.5 cm/s
Catalyst composition:	
Fe	1000 g/kg
Cu	1 g/kg
K_2O	0.5 – 5 g/kg

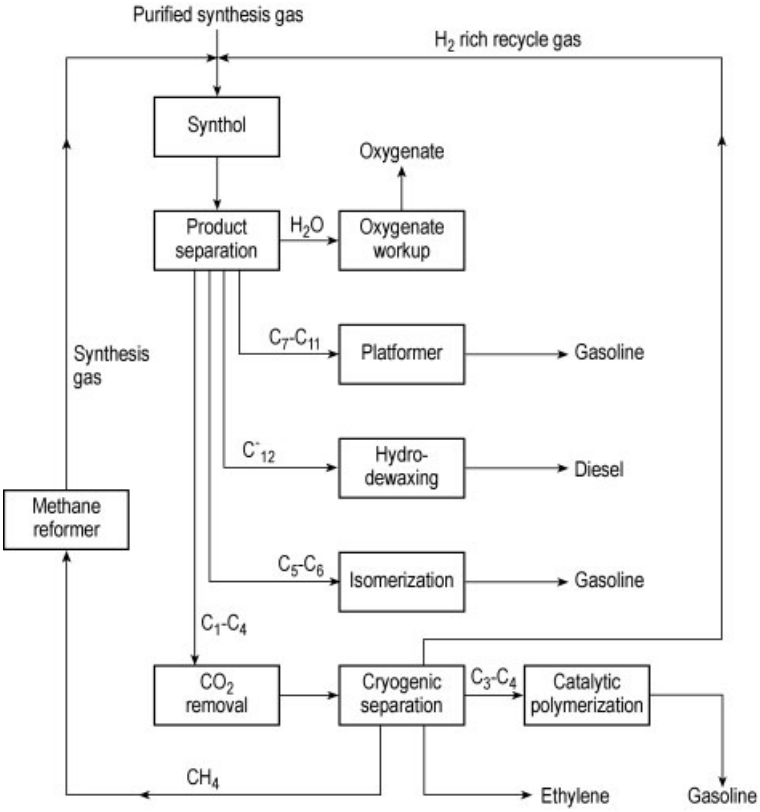

Figure 9. Flow sheet for SASOL II products workup

The height of the expanded liquid-catalyst bed was kept constant by adding melted wax or by withdrawing the products with a high boiling point through a pressure filter to remove the catalyst. The product distribution (in wt %) obtained from the operating plant at maximum gasoline production was as follows:

Methane and ethane	3.2
Ethylene	3.6
Propene	18.0
Propane	4.5
Butene	3.8
Butane	1.3
Boiling range	
40 – 180 °C	53.6
180 – 220 °C	4.0
220 – 320 °C	6.0
>320 °C	1.9

High selectivity for gasoline was achieved, almost 54 % compared to the 48 % predicted from probability considerations. If propene and butene polymerization was included, gasoline production reached 73 % compared to 62 % for Synthol gasoline. Attempts to repeat these results have been unsuccessful [97–100].

In a later investigation at Mobil, slurry-phase Fischer – Tropsch synthesis was combined with fixed-bed zeolite processing of the crude products [101]. Processing of such products over zeolite ZSM-5 was known to form high-octane gasoline by a combination of isomerization, aromatization, alkylation, and elimination of oxygen compounds [102].

In the first stage of the *Mobil two-stage process,* the Fischer – Tropsch synthesis was carried out in a slurry reactor containing alkaline-promoted iron catalysts in a liquid heat transfer medium. Product selectivities were as follows [101–103]:

Reactor conditions	
Pressure	1.48 MPa
Temperature	260 °C
H_2 : CO Molar ratio	0.68
Hourly space velocity at STP, L/g Fe catalyst	2.6
Conversion data, mol %	
Hydrogen	79
Carbon monoxide	92
Syngas	87
Selectivities, wt %	
CH_4	6.8
C_2H_4	1.6
C_2H_6	2.8
C_3H_6	7.5
C_3H_8	1.8
C_4H_8	6.2
C_4H_{10}	1.8
C_5–C_{11}	18.6
Light hydrocarbons	14.3
Heavy hydrocarbons	27.6
Wax	10.0

When the reactor was operated at low C_1 and C_2 yields (ca. 5 wt %), large quantities of wax (C_{23}+) were produced; this wax was removed to prevent excess liquid buildup, the catalyst was returned to the reactor without deactivation, which may be difficult in a commercial operation. The effect of the zeolite second stage is shown in Table 15.

Table 15. Selectivities before and after ZSM-5 reactor

Products	First stage, wt %	Second stage, wt %
C_1	6.5	6.8
C_2*/C_2**	1.7/2.7	1.0/2.9
C_3*/C_3**	7.1/1.9	2.0/5.6
C_4	5.9	4.7
Iso-C_4 /*n*-C_4	0/1.8	6.6/6.1
C_5–C_{11}	27.1	51.5
C_{12+} (liquid)	33.3	0.8
Reactor wax	12	12
C_5–C_{11}, wt %		
Alkanes	–	49.4
Olefins	–	22.8
Naphthenes	–	4.5
Aromatics	–	23.3

* Saturated.
** Unsaturated.

Research on the Fischer–Tropsch process is continuing at SASOL and elsewhere. Much of this work is aimed at improving selectivity by developing novel catalysts and reactors. SASOL has developed high-activity, long-life, supported cobalt catalysts, that have a longer service life and higher activity than the conventional iron-based catalysts [85]. This more than compensates for the higher raw material costs. However, further work on improving the life of the iron-based catalyst in slurry phase reactors may still result in the use of iron catalyst with low H_2 : CO ratio syngas derived from coal. The iron catalyst exhibits water–gas shift activity which is absent from cobalt-based catalysts. Water–gas shift activity is desirable for low H_2 : CO ratio coal-derived syngas but not for the higher H_2 : CO ratio natural gas-derived syngas.

The selective hydrocracking of Fischer–Tropsch waxes yields 80 % diesel fuel (cetane number ca 70), 17 % naphtha, and 3 % C_1–C_4 hydrocarbon gases. The wax is an ideal hydrocracker feedstock because it contains no aromatic, nitrogen, or sulfur compounds. When combined with the straight-run products (which are typically hydrotreated to remove olefins and oxygenates) the final diesel : naphtha ratio is typically 75 : 25. There is considerable flexibility for the processing of the wax. It may be fed to a Fluidized Catalytic Cracking (FCC) unit to produce gasoline; it may be fractionated and hydroisomerized to produce lubricant base oils, or intermediate severity hydrocracking/hydroisomerization may produce a combination of FCC feedstock, lubricants, diesel fuel, and naphtha. A kerosene or jet fuel product intermediate between diesel and naphtha may also be produced. The naphtha is an ideal feedstock for cracking to olefins since it is completely paraffinic [96].

2.2.8. Advanced Systems

Integrated indirect liquefaction by SASOL technology with two modes of operation was discussed in Section 2.2.5. In the mixed-output mode, gas and liquid fuels are produced together. In the all-liquid-output mode, the C_1 and C_2 hydrocarbons are reformed to synthesis gas and only liquid fuels are produced. The expected effect of using advanced entrained gasification, such as the Shell system, in place of the dry-ash Lurgi is significant, with an increase in net efficiency of ca. 23 % [104]. This would result in a 23 % reduction in gasoline cost if commercial operation matches pilot-plant results.

Predicting the effect of the slurry-phase system and the Mobil two-stage slurry ZSM-5 system is more difficult. However, if the Rheinpreussen-Koppers pilot plant results could

be scaled up, the plant efficiency of a Shell gasification system would increase by an estimated additional 10 % [104].

2.3. Methanol and Methanol to Gasoline (MTG)

Methanol formation can be seen as a variant of the Fischer – Tropsch synthesis:

$$CO + 2\,H_2 \longrightarrow CH_3OH \quad \Delta H_{298} = -90.84\,\text{kJ/mol}$$

$$CO_2 + 3\,H_2 \longrightarrow CH_3OH + H_2O \quad \Delta H_{298} = -49.57\,\text{kJ/mol}$$

Both reactions are exothermic and proceed with volume contraction; they are consequently favored by a low temperature and high pressure.

Both high- and low-pressure processes are available to hydrogenate carbon monoxide. The *high-pressure method* was commercialized by BASF in 1923 [105] and dominated the market in various versions until 1966 [106]. These processes used a zinc – chromium oxide catalyst at or above 340 °C and required a pressure of 30 – 50 MPa to obtain methanol concentrations of ca. 5 – 6 vol % in the effluent gases from the reactor, together with significant amounts of methane, dimethyl ether, ethanol, and higher alcohols.

In 1966 ICI introduced a *low-pressure methanol synthesis* that used a copper – zinc – chromium catalyst at 5 MPa [105]. Other low-pressure processes were developed later by Lurgi Kohle & Mineralöltechnik [107], Mitsubishi Gas & Chemical Co., and Haldor Topsøe [108]. These processes differ primarily in operating pressure, catalyst, and reactor design, especially with respect to heat recovery and temperature control. The ICI and Lurgi processes are dominant.

In the *ICI methanol process,* sulfur-free synthesis gas is mixed with part of the recycle gas, heated, and fed to the converter at 5 – 10 MPa. The remainder of the recycle is used as quench gas and is admitted to the catalyst bed through a series of lozenge-shaped distributors. In the reactor, the CO, CO_2, and H_2 are partially converted to methanol. The hot exit gases are cooled to condense the methanol. Part of the unreacted gases is purged to prevent inert impurities from accumulating; the remainder is compressed and recycled to the reactor. The typical recycle ratio ranges from 5 at 10 MPa to 8 at 5 MPa [108]. Conversion of CO to methanol per pass is normally 16 – 20 %. Feed-gas composition is typically 59 vol % H_2 and 27 vol % CO, to give an H_2 : CO ratio of 2.17.

The ICI process uses a hot-wall, catalytic, fixed-bed, vapor phase, adiabatic reactor designed for rapid catalyst loading and unloading and for efficient mixing of cold and hot gases. The reactor holds a bed of catalyst with quench gas distributors at several levels to control temperature. The bottom catalyst layer is usually kept at ca. 230 °C. Catalyst lifetime is typically 2 – 3 years.

The *Lurgi process* uses a tubular reactor with copper – zinc catalyst packed in vertical tubes that are contained within a reactor shell filled with boiling water. This closely resembles the ARGE Fischer – Tropsch reactor; it differs from the ICI methanol system in having only one feed location. Hot synthesis gas is fed to the top of the reactor and distributed uniformly through the tubes. The heat of reaction is removed by the circulating boiling water [107].

The U.S. Department of Energy (DOE) and the Electrical Power Research Institute (EPRI) have supported research on methanol synthesis by a novel system known as *liquid-phase methanol* (LPM), in which synthesis gas reacts with a catalyst suspended in a liquid heat transfer medium. This reactor resembles a slurry reactor (Section 2.2.7) except for the circulation through an external heat-transfer loop attached to the reactor. The experimental LPM reactor at LaPorte, Texas, is 57 cm in diameter and 5.5 m high [109].

Mobil methanol-to-gasoline process. Methanol can be used as a fuel alone or blended with gasoline. In addition, it can be converted to gasoline by the Mobil MTG process [110]. Methanol as the first stage in the indirect liquefaction affords 100 % selectivity for gasoline, much higher than that obtainable in the Fischer–Tropsch synthesis with its wide range of hydrocarbon products.

In the Mobil MTG process, methanol is dehydrated over the zeolite ZSM-5 catalyst [111], first to dimethyl ether and then to hydrocarbons that are predominantly in the gasoline boiling range, with some LPG and fuel gas [68]. The reaction exotherm ranges from 1.5 to 1.75 kJ/g

methanol, depending on the product distribution [68]. The complex reaction sequence can be represented as follows [68]:

$$2\,CH_3OH \longrightarrow (CH_3)_2O + H_2O$$

$$(CH_3)_2O \longrightarrow (CH_2)_2 + H_2O$$

Light olefins $\longrightarrow$ Heavier olefins

Heavy olefins $\longrightarrow$ Aromatics, alkanes, cycloalkanes

For 100 g of methanol consumed, the stoichiometric yields are 43.75 g of hydrocarbons and 56.25 g of water.

Mobil has developed two reactor configurations: the adiabatic fixed-bed and the fluidized-bed processes. In the *adiabatic fixed-bed process*, two reactors are used in series, the first for dehydrating methanol to an equilibrium mixture of dimethyl ether and water. This mixture is combined with recycle gas from the second reactor and converted to hydrocarbons and water over the zeolite catalyst. The recycle gas, which consists of light hydrocarbons, aids in heat removal from the second stage [68]. The catalyst in the MTG reactor requires regeneration about every 20 d.

Much of the development work has been performed in a 0.64 m^3/d pilot plant at Mobil's New Jersey facility. The dehydration reactor is 5 cm in diameter by 4 m and the conversion reactor is 10 cm in diameter by 4 m. In 1980, the government of New Zealand decided to use this technology in a 2700 m^3/d plant for converting natural gas to gasoline [112].

Development of the *fluidized-bed MTG process* was partially supported by the U.S. DOE. A 0.64 m^3/d pilot plant was operated successfully, demonstrating the following advantages over the fixed-bed process [110]: heat removal at high temperature and mixing of catalyst and reactants are highly efficient, high catalyst activity is maintained with continuous regeneration, and the heat of reaction is utilized to vaporize feed methanol.

Table 16. MTG operations [68], [110]

	Fixed-bed	Fluidized-bed	
	0.64 m^3/d	0.64 m^3/d	16 m^3/d
MTG inlet temp., °C	360	413	407
outlet temp., °C	412	413	407
Gas recycle ratio (molar)	9	0	0
Weight hourly space velocity, (pure methanol)	1.6	1.0	1.1
Reactor pressure, MPa	2.17	0.27	0.41
Product, wt % [a]			
Hydrocarbons	43.7	43.5	40.0
Water	56.2	56.0	NA [b]
CO, CO_2, H_2, other	0.1	0.5	NA
Hydrocarbon product, wt %			
$C_1 + C_2$	2.1	5.6	NA
Propane	8.6	5.9	NA
Propene	0.2	5.0	NA
n-Butane	4.2	1.7	NA
Isobutane	8.4	14.5	NA
Butenes	0.7	7.3	NA
C_5+ gasoline	75.8	60.0	NA
Gasoline yield, wt % hydrocarbons [c]	80.2	88.0	89.0
Research octane number [d]	95	96	95

[a] Based on methanol feed.
[b] NA = not available.
[c] Includes alkylate.
[d] Unleaded product.

The 0.64 m^3/d pilot plant's reactor is 10 cm in diameter and 7.6 m high, with provision for external catalyst circulation and regeneration [113]. In 1980, an international project was initiated to construct and operate the fluidized-bed MTG process on a 16 m^3/d demonstration scale at Wesseling, Federal Republic of Germany. Participants included Mobil, Uhde, and Union Rheinische Braunkohlen Kraftstoff (URBK). Financial assistance was also provided by the U.S. DOE and the Ministry for Research and Technology (BMFT) of the Federal Republic of Germany [110]. The conversion rate has averaged 850 kg/h of methanol, and the plant has met design specifications and target quantities [110]. Conditions and yields are shown in Table 16 for fixed-bed and fluidized-bed operations. The product has the properties of premium unleaded gasoline [68]. The content of durene (1,2,4,5-tetramethylbenzene), which can crystallize in the engine, is kept below 5 wt % by regulating reactor conditions.

2.4. Dimethyl Ether (DME) (see also → Dimethyl Ether)

Dimethyl ether is a new promising substitute fuel for gas turbines, for LPG, and perhaps more importantly, as an alternative diesel fuel. DME has found use as a substitute propellant for the ozone depleting chlorofluorocarbons presently being phased out. The world production (1999) is around 150 000 t/a. The manufacturing process today is based on simple dehydration of methanol. The relatively small quantities of DME produced do not justify investment in new more expensive and sophisticated production facilities yielding higher volumes and lower unit prices. However, recent studies have shown that DME has the potential of being a new significant energy carrier in the future global energy market similar to liquefied natural gas.

DME can be manufactured from nearly all carbon-bearing substances at a price not much above that of standard diesel [114], [115]. DME production from natural gas is outlined in [116–118]. A brief discussion of the production process from coal derived synthesis gas follows. Coal gasification generates a fairly impure synthesis gas, which has to be cleaned thoroughly. The purification of such gas may be done by hot gas cleaning at around 400 °C in order to conserve energy. The conversion of synthesis gas directly through methanol into DME takes place in a synthesis loop consisting of three catalytic fixed-bed adiabatic reactors with indirect cooling. The catalyst applied is a proprietary dual-function catalyst, catalyzing both steps (i.e., methanol and DME synthesis) in the sequential reaction. Significant advantages arise by permitting the methanol synthesis, the water–gas shift, and the DME synthesis reaction to take place simultaneously. The synthesis of DME from synthesis gas involves three reactions:

$$CO_2 + 3\,H_2 \rightarrow CH_3OH + H_2O \quad \Delta H = -50.1\ \text{k/mol} \qquad (5)$$

$$H_2O + CO \rightarrow H_2 + CO_2 \quad \Delta H = -40.9\ \text{k/mol} \qquad (6)$$

$$2\,CH_3OH \rightarrow CH_3OCH_3 + H_2O \quad \Delta H = -23.4\ \text{k/mol} \qquad (7)$$

Reactions (5) and (6) represent a conventional methanol synthesis. This methanol synthesis is restricted by equilibrium, which requires high pressure in order to reach an acceptable conversion. The introduction of Reaction (7), the DME synthesis, serves to relieve the equilibrium constraints inherent to the methanol synthesis by transforming the methanol into DME. Moreover, the water formed in Reaction (7) is to some extent driving Reaction (6) to produce more hydrogen, which in turn will drive Reaction (5) to produce more methanol. Thus, the combination of these reactions results in a strong synergetic effect, which dramatically increases the synthesis gas conversion potential.

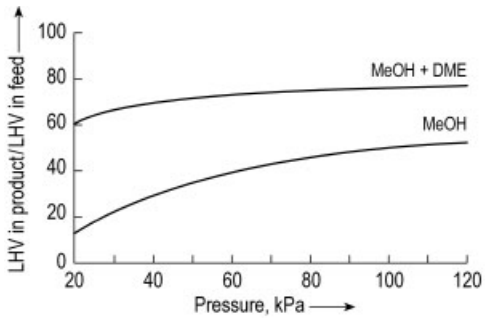

Figure 10. Conversion of $C + CO_2$ at equilibrium as a function of synthesis pressure for a typical synthesis gas composition

Figure 10 shows the conversion of $CO + CO_2$ at equilibrium as a function of synthesis pressure for a typical synthesis gas composition. The efficiency of the process and the dual-function catalyst have been demonstrated in a 50 kg/d process demonstration pilot in Haldor Topsøe's facilities in Copenhagen, Denmark. The original catalyst charge has been in operation for more than 15 000 h and demonstrated an excellent stability with an aging profile slower that the of commercial methanol synthesis catalysts.

Subsequently to the start up of the Topsøe pilot plant, Air Products of USA and NKK of Japan have piloted the synthesis of DME from synthesis gas based on slurry bed reactors [119].

NKK began working to develop the direct synthesis technology for DME in 1989 and has established a 50 kg/d bench plant. In 1997, NKK started an R&D project on direct synthesis technology for the low cost mass production of DME using a 5 t/d pilot plant to evaluate the whole system. The basic features of the NKK process are its slurry bed and its newly developed catalyst system which realize the direct reaction (Eq. 8)

of synthesis gas with a H_2 : CO ratio of 1 in the slurry phase. NKK uses a slurry bubble column reactor with an inner diameter of 90 mm and a height of 2 m. The hybrid catalyst containing $CuO-ZnO/Al_2O_3$ and copper-modified γ-alumina showed a high activity for the synthesis of DME in the slurry phase. Optimum synthesis gas conversion (56 %) and high selectivity to DME (98 %) were achieved at the initial H_2 : CO ratio of 1 [120–123]:

$$3\,CO + 3\,H_2 \rightarrow CH_3OCH_3 + CO_2 \qquad \Delta H = -246.1\ \text{kJ/mol} \tag{8}$$

Coal-derived synthesis gas obtained by high-temperature gasification has a H_2 : CO ratio of 0.5 to 1, which has to be adjusted to $H_2 : CO = 2$ by an expensive shift converter before the DME synthesis in order to synthesize DME according to Equation (7). The reaction according to Equation (8) has a substantial advantage regarding synthesis gas preparation and DME conversion [124].

With funds provided by the Agency of Natural Resources and Energy of the Ministry of International Trade and Industry, the Center for Coal Utilization of Japan, Japan is conducting a five year project to evaluate the whole system of the direct synthesis of DME. The project started in the fiscal year 1997 with the cooperation of NKK, Taiheiyo Coal Mining Co., Ltd. and Sumitomo Metal Industries, LTD. In 1999, a 5 t/d pilot plant was constructed at Taiheiyo Coal Mining's site in Kushiro, Hokkaido and the operation was started. The dimensions of the slurry phase reactor are 55 cm inner diameter and 15 m height. The aimed single-pass synthetic gas conversion is around 50 % [124].

2.5. Environmental Protection

Liquefaction plants, like SASOL in South Africa, produce gaseous, liquid, and solid effluents. Emissions into the environment are controlled by integral plant design technologies, some of which are used in related industries, such as utilities or petroleum refining.

Atmospheric pollutants arise from (1) burning coal and fuel gases (2) gasification, (3) coal handling and preparation, and (4) fuel refining. *Combustion-generated pollutants* from the steam and utilities plant are primarily particulates, sulfur dioxide, nitrogen oxides, and volatile organic compounds. Other inorganic pollutants are found in coal, including fluoride, beryllium, lead, and mercury. The known technology for controlling pollution created by coal combustion is, of course, applicable here. Flue gas desulfurization (FGD) units [125], electrostatic precipitators (ESP), and bag houses [126] can be used to control sulfur dioxide and particulates. Nitrogen oxides (NO_x) emissions can be minimized by combustion techniques or chemically [127], for example with Thermal DeNox. The ash and sulfur contents depend on the characteristics of the coal. In the United States, regulations currently exist for coal-fired utilities but not as yet for coal conversion plants; these regulations will most likely be extended as needed [128]. With appropriate design, indirect liquefaction plants can eliminate coal combustion for steam and utilities production. If gasifiers process feed coal, including fines, all the steam and power requirements could be met by combustion in gas turbines of a portion of the cleaned synthesis gas.

The crude gasification product contains almost 100 % of the coal sulfur, mostly as hydrogen sulfide. The sulfur content must be reduced to ca. 1 mg/kg so as not to poison the catalyst. Various processes convert the hydrogen sulfide to elemental sulfur, which is marketable. The most common of these are the Stretford, used at SASOL, and the Claus process, used at several petroleum refineries. The Stretford process was developed by BGC and is used worldwide in over 50 plants [77], [78].

The Claus process has been well established in the refinery and natural gas industries for many years. Several hundred plants operate worldwide, with an average capacity of 100 t/d of sulfur. The correct stoichiometric portion of feed H_2S is oxidized by combustion in a front-end furnace to yield SO_2, which oxidizes the remaining H_2S in downstream catalytic reactors:

$$2\,H_2S + 3\,O_2 \longrightarrow 2\,H_2O + 2\,SO_2$$

$$2\,H_2S + SO_2 \longrightarrow 2\,H_2O + 3\,S$$

Residual sulfur compounds can be removed by the Shell Claus Offgas Treating (SCOT) process

developed in 1972 [129]. The combined gas purification and recovery processes convert 99 % of the sulfur in gaseous compounds to elemental form.

Particulates from *coal handling* and preparation can be controlled by enclosures, sprays, and bag houses. Gaseous pollutants from refining operations and vapor leakages can be controlled by recovery systems and proper maintenance. In general, indirect coal conversion plants do not create major problems of atmospheric pollution; any gaseous effluents can be controlled by known methods [128].

Aqueous effluents or wastewaters arise principally from steam plant blowdowns, gasification wash waters, and Fischer – Tropsch wastewaters. *Steam plant blowdown* creates wastewater identical to that produced by conventional coal-fired utility plants and are not discussed here.

Gasification wash water from Lurgi dry-ash gasification, also called process gas liquor, is treated as follows in the SASOL II and III plants [128]. All the liquor produced by washing and cooling of the exit gases is combined in large separation vessels. This wastewater contains dissolved acid gas, organic acids, tar components, phenols, and ammonia [128]. The last two are removed before attempting to remove the others because they are dangerous to activated sludge microbes employed in water biotreatment. The Lurgi Phenosolvan process, which is also used at the Great Plains gasification plant in North Dakota, removes phenols by solvent extraction [130], [131]. Ammonia is removed by the Chemie Linz Lurgi ammonia process [71], and the wastewater is then treated biologically by activated sludge [128], [132], which, however, does not remove all chemical oxygen demand (COD). Residual COD is attributed to humic acid components, which are apparently harmless [128], but the remaining 1 mg/kg of phenol must be removed by adsorption on granular activated carbon (GAC). The effluent water, which is then colorless and odorless, can be discharged or used as cooling tower makeup water. At SASOL II and III, it is reused for cooling without fouling or corrosion problems [128].

Process wastewater from the Fischer – Tropsch synthesis, after recovery of alcohols and other oxygen compounds, is also given activated-sludge biotreatment.

Solid wastes arise from gasification residue (i.e., ash, slag, and char); wastes generated by grinding and grading the coal; steam plant wastes, such as fly and bottom ash and flue-gas desulfurization residue; water treatment wastes, such as biosludge and salts from blowdowns; and spent catalysts.

The quantity and character of these solid wastes depend, largely, on the gasification process used and the characteristics of the feed coal. The residue from the dry-ash Lurgi gasification process contains some ungasified carbon. The coal used at SASOL II and III produces ca. 12 000 t/d of ash, with a density of ca. 1440 kg/m^3. Fly ash, bottom ash, slag wastes, and FGD wastes from coal combustion have been classified as nonhazardous by the Resource Conservation and Recovery Act (RCRA) of 1976, the major regulation in the United States governing the management of solid wastes [133].

A later regulation temporarily excludes coal gasification wastes from being classified as hazardous [134]. These regulations allow solid wastes from gasification and boiler plants to be disposed of in nonhazardous landfills or settling lagoons. None of the ashes or slags from coal gasification that have been tested by the EPA-approved extraction procedure have been found hazardous [128]. Wastewater treatment sludge is usually incinerated and activated carbon regenerated. Concentrated inorganic salts from cooling tower blowdowns are hazardous and require disposal in lined landfills.

The prior description applies to Lurgi dry-ash gasification combined with the Synthol Fischer – Tropsch synthesis. The *fixed-bed Lurgi gasification system* produces tars, oils, phenols, and such quantities of wastewater as to represent a worst-case environmental scenario. Pollution control is greatly simplified by the use of advanced entrained gasification. The Shell system, for example, produces only carbon monoxide and hydrogen without tars, oils, or phenols, transforms the mineral matter in the coal to a glassy, nonleachable slag, and consumes minimal steam.

Pollution control for indirect coal liquefaction represents 10 – 15 % of the total plant cost

[128]. The degree of protection to be attained depends primarily on cost considerations rather than technical limitations [128].

2.6. Economic Aspects

SASOL, in their 1999 interim report, set an operating cost target for their high temperature Fischer – Tropsch operations at Secunda, excluding capital charges, of $ 11 per barrel. This corresponds to a product cost of just over $ 16 per barrel when allowing for the 90 % yield of a typical modern refinery and a $ 4 per barrel mark-up for the refinery processing. When compared to the figure of $ 18 per barrel for the product cost published in SASOL's 1999 annual report it can be seen that continued improvements are required to decrease the operating costs of coal liquefaction towards the above-mentioned target. The capital charges for future coal-based operations would be site specific. Perhaps the benchmark should be those locations that use coal as a competitive feedstock for other synthesis gas-based processes, such as methanol or ammonia production. These locations may also be considered as a potential source of syngas for a Fischer – Tropsch based operation. In addition, there is synergy between various syngas users, e.g., ammonia and methanol synthesis.

SASOL has demonstrated that coal liquefaction using HTFT technology can be commercially viable, given a number of prerequisites. Firstly, such plants are capital intensive and require very large economies of scale to be viable. Secondly, higher value products other than liquid fuels must also be produced. Potential petrochemical products are light olefins (ethylene and propene) and various linear α-olefins, such as 1-butene, 1-pentene, 1-hexene, 1-heptene, and 1-octene. Alcohols are another important coproduct. Depending on the source of gasification technology, there are a number of gasifier byproducts that may also play a role. SASOL, for example, uses Lurgi Coal Gasifier, which produce valuable byproducts such as ammonia, phenol, cresol, etc. Advances in HTFT technology have, however, decreased the capital cost considerably since the CFB reactors, originally used in Secunda plants, were replaced by the SASOL Advanced Synthol reactors. In addition the energy efficiency of the Fischer – Tropsch process has been dramatically improved by recent technology developments. The Fischer – Tropsch unit is now also a large steam exporter substantially decreasing the need to generate high-pressure steam from coal. Cogeneration of power may also be an option at some locations.

Another aspect to consider are the recent developments in low temperature Fischer – Tropsch (LTFT) technologies. This results in considerable simplification of the primary product recovery and refining processes. The simplification leads to further decreases in the capital costs allowing viable applications at smaller plant capacities when compared to the HTFT route. At SASOL's Sasolburg site coproduction of methanol and hydrogen for ammonia production and refinery uses has been shown to produce synergistic benefits. The choice between HTFT and LTFT technologies is decisively determined by the desired products. The potential products from the various Fischer – Tropsch technologies are discussed in Section 2.2.6.

3. Direct Coal Liquefaction

3.1. Mechanisms

Because the exact structure of coal is unknown, understanding and close control of the processes by which coal is converted to liquids are difficult. The available structural information pertains predominantly to the vitrinite maceral, the most abundant organic coal constituent. Attempts to describe the vitrinite structure have focused either on the concept of an "average" structure, which embodies the measurable chemical parameters of vitrinite [135], [136], or on a less rigidly defined concept in which the maceral is considered to consist of a three-dimensional macromolecular network that can accommodate lower molecular mass species in open and closed pores [137–139].

Bituminous coal consists of highly substituted, cyclical carbon structures of one or more condensed rings linked by alkyl or ether (oxygen or sulfur) bridges. Low-rank coals contain a higher proportion of oxygen groups. With increasing rank, progressively fewer functional groups are present and the ring system becomes more polycondensed [140].

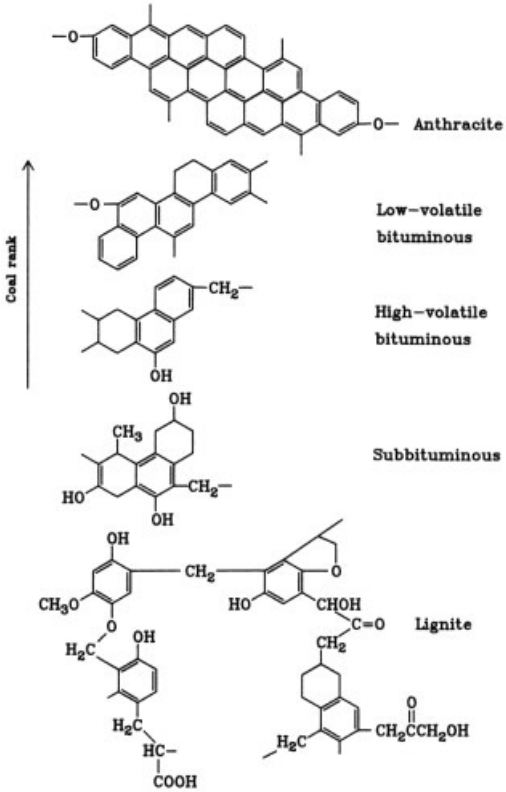

Uncatalyzed liquefaction is considered to be a pyrolytic process; the bridges linking the cyclical units are ruptured when heated to 400 – 450 °C. Ether linkages are cleaved first [141–143] and then methylene bridges [144–146]. In the early stages of liquefaction, hydroxyl groups are apparently lost more slowly than ether oxygen [147], [148]. Organic sulfur and oxygen are lost at similar rates [148], [149].

Bond rupture liberates free radicals, which can be stabilized by the addition of a hydrogen atom; the low molecular mass fragments are then dispersed in the solvent. If hydrogen capture does not occur, the radicals can combine to form stable, high molecular mass products, leading ultimately to char or coke formation (Fig. 11).

It is well recognized that such retrogressive reactions occur to some extent during coal liquefaction and that they have a negative impact on the liquid yield [150], [151]. It has become increasingly clear that retrogressive reactions substantially hinder the liquefaction of low-rank coals and that oxygen-containing functional groups in the coal structure are major actors in these retrogressive reactions. Carboxyl functions have been implicated with the cross-linking of coals during heating at relatively low temperatures [152–156]. Benzyl-type radicals can lead to retrogressive reactions and these reactions can be induced by decomposition of the oxygen-containing functional groups and condensation of the polyhydroxyl structure, such as aromatic polyphenols and aromatic carboxylic acids [157], [158]. A two-stage liquefaction, using a lower preheating temperature, is advantageous compared with one-stage liquefaction for the same reaction time [159], [160].

In addition, the oil yield was successfully increased by demineralization [159], solvent swelling [161], [162], or alkylation of coal [163], [164].

Liquefaction, therefore, depends critically on a supply of hydrogen atoms commensurate with the rate of free-radical formation, which is a function of coal structure and reaction temperature. Hydrogen sources are the liquefaction solvent, the coal, or H_2 gas. A new mechanism was proposed for hydrogen donation and cleavage of relatively strong bonds, involving a solvent-assisted process [165], [166]. This mechanism, called radical hydrogen transfer (RHT), is illustrated at the bottom of Figure 12. According to this reaction pathway, H-atoms transfer occurs through solvent-derived radicals to ipso-positions of the coal matrix which leads to the scission of the strong bonds in the coal structure. This pathway also provides the possibility that the selectivity in coal liquefaction might be controlled by tailoring the solvent medium appropriately [167]. Other exciting developments include discovery of new bond-breaking chemistry, such as the use of water to cleave refractory diaryl ether covalent bonds [168–170].

A major limitation to understanding the process of coal liquefaction is the lack of a realistic working model of the coal structure. Recent studies emphasize the importance of noncovalent bonds connecting the constituent macromolecules through their polar groups. Examples of these bonds are hydrogen bonds, charge-transfer interactions, cationic bridges, and layered stacking aromatic planes, which are responsible for low solubility, facile coking of coal

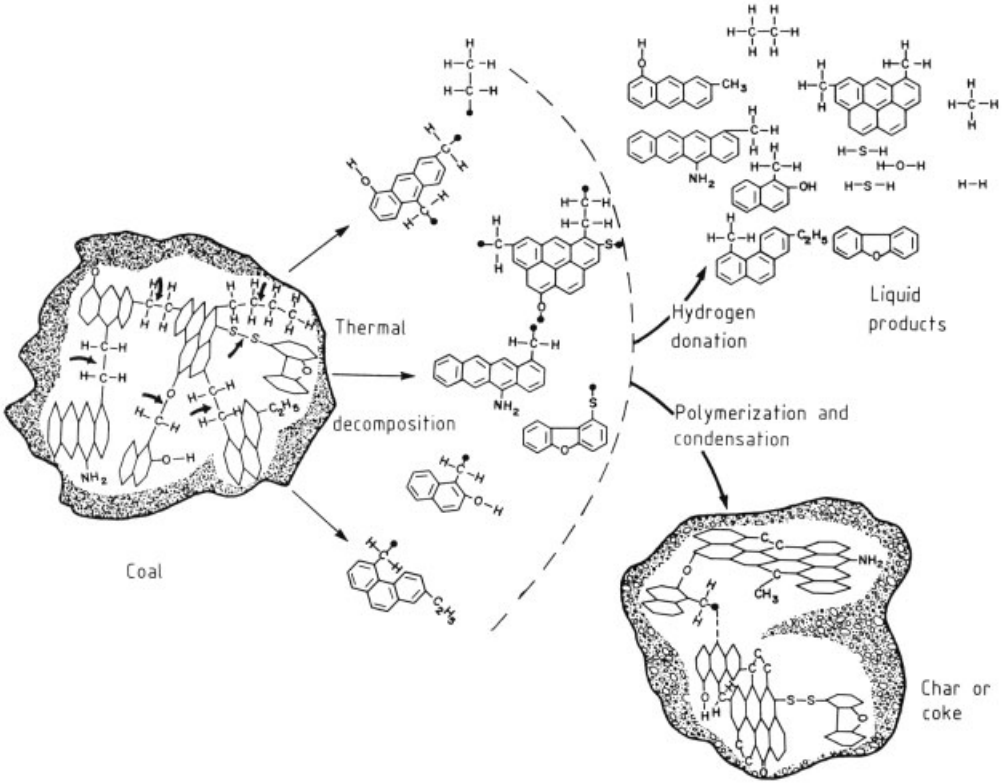

Figure 11. Representation of processes occurring during direct coal liquefaction; structures are representations only and are hexagonal combinations of aromatic and nonaromatic rings

components, and limited access to solid catalysts [171–175].

Although progress in improving the direct coal liquefaction process in the 1990s has been particularly significant, the fundamentals of the physical and chemical processes during the coal liquefaction are still incompletely understood [176]. A comprehensively knowledge of basic coal structure must be acquired for the effective development of coal liquefaction [177–182].

3.2. Liquefaction Products

The initially solubilized coal is high in molecular mass and cannot be distilled. Extended thermal or catalytic reaction is necessary for the production of distillates. In all cases a distribution of products is obtained that includes light gases, distillate liquids, and nondistillable liquids, which may be solid at room temperature.

The nondistillable products are usually characterized by solvent fractionation, using techniques developed for the petroleum industry. The liquids are separated on the basis of solubility into oils (soluble in pentane or another light alkane) and asphaltenes (benzene-soluble and pentane-insoluble). A third category, not used in petroleum processing, is preasphaltenes (pyridine-soluble and benzene-insoluble); this category is applied to coal liquids [183]. The distinction between the three fractions is based solely on solubility and has no chemical significance. The quantities of the fractions isolated from coal liquids are sensitive to the solvent separation procedure [184–187].

The chemical character of asphaltenes depends on the structure of the parent coal and on the processing history. Long hydrotreatment increases aromaticity and reduces the concentration of polar functional groups and molecular mass [188–190].

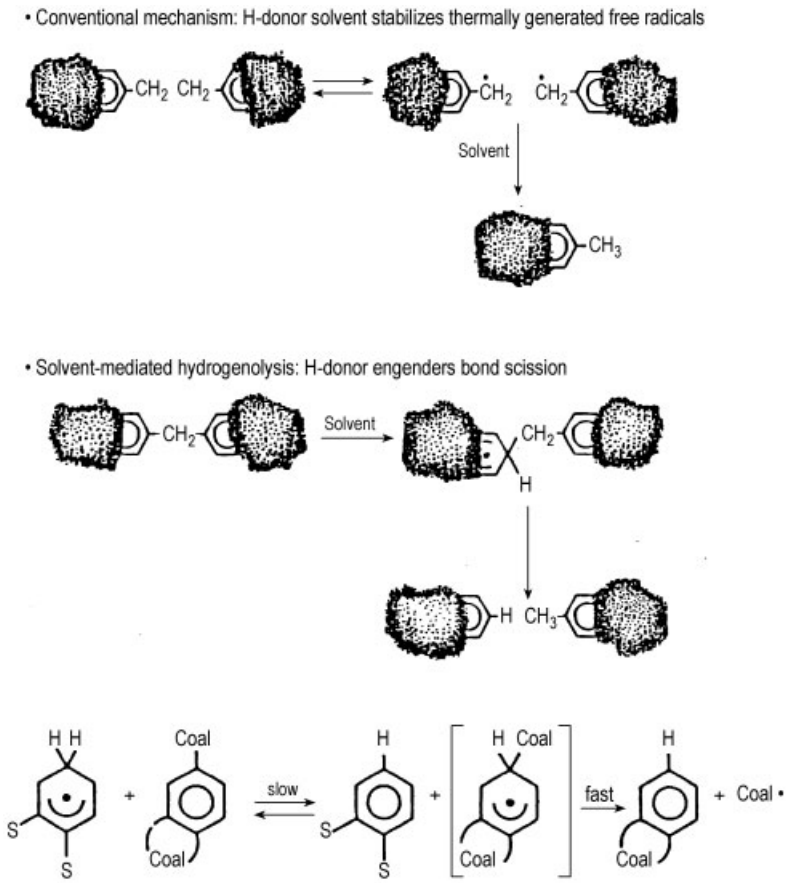

Figure 12. Conventional coal liquefaction mechanism and solvent-mediated hydrogenolysis [165]

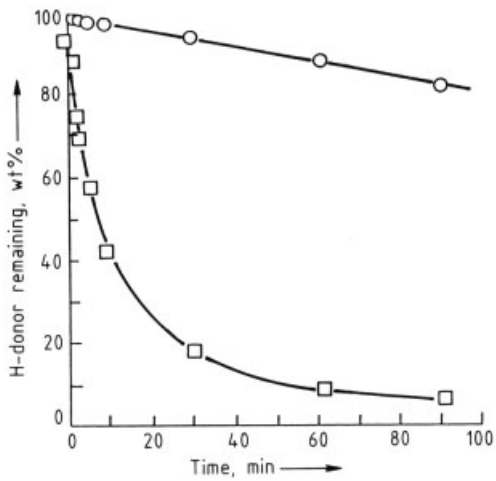

Figure 13. Reaction of H-donors with coal [149]
○ Tetralin; □ Dihydrophenanthrene

These considerations are often ignored in the literature, where, for example, kinetic models are based on the erroneous assumption that a solubility fraction behaves as a single compound.

3.3. Solvents

The solvent used in coal liquefaction serves a number of important purposes. It provides a transport medium for the coal and facilitates heat transfer. During reaction it serves as an important source of transferable hydrogen, and it disperses the liquefaction products.

Most processes are designed to be independent of external sources of solvent and hydrogen gas. Therefore, the solvent must be derived entirely from the liquefaction products for recycle and produced in sufficient quantity and of suitable quality to sustain continuous operation. A

sufficient degree of freedom within the process is necessary to control recycle solvent composition. Much research has been devoted to the reactions of solvents with coal to identify those products that improve liquefaction so that they can be selectively separated for recycle or generated in an additional process stage.

The stipulation for liquefaction processes to generate their own solvent has been subject to reevaluation. The co-processing of coals with heavy hydrocarbon feeds, such as tar sands bitumen, shale oil kerogen, heavy petroleum crudes, or petroleum residuals, can upgrade the hydrocarbon feedstock to distillate products as well as produce liquids from coal [191–195], [196–204]. Other benefits include simultaneous reductions in the concentrations of metals, sulfur, and asphaltenes (see Table 17). Co-processing eliminates the need for recycling, and is consequently more economical than integrated liquefaction.

3.3.1. Hydrogen Donors

The solvent components that are the most effective hydrogen-donor compounds (H-donors) are those with a *partially hydrogenated aromatic structure* [141], [183], [205–209]. The hydroaromatic hydrogens are readily abstracted, giving rise to stable aromatic compounds and stabilizing the coal free radicals [210].

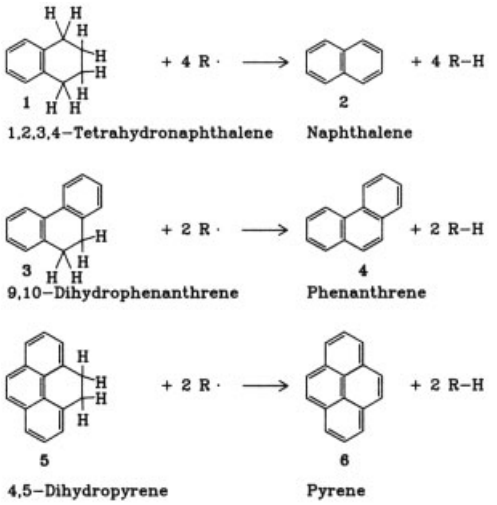

where R· represents any group fragmented from the coal structure

Electron spin resonance (ESR) spectrometry permits direct observation of free radicals in coal and in its liquefaction products [209], [211], [212]. Although the mechanistic significance of ESR data is debatable [209], the free-radical population has been shown to be substantially lowered by active H-donors.

Table 17. Co-processing bituminous coal with heavy residual crude ∗ [191]

Property	Feed oil	Product oil
Ni, mg/kg	89	9
V, mg/kg	94	19
S, wt %	5.3	0.6
N, wt %	0.6	0.4
Gravity, API ∗∗	10.0	22.4

∗ Conversion: coal 89 %, residuum 76 %.
∗∗ American Petroleum Institute.

The value of hydrogen-donor compounds has long been recognized, and their activity is related to the degree of saturation and to the number of condensed rings [2], [15], [27]. Complete saturation produces stable naphthenic compounds (e.g., decahydronaphthalene) that are poor H-donors compared to partially saturated compounds (e.g., tetrahydronaphthalene, **1**) [141], [208]. In turn, 9,10-dihydrophenanthrene (**3**) is a more active H-donor than tetrahydronaphthalene (see Fig. 13).

Polycyclic compounds are easily hydrogenated or dehydrogenated. Pyrene (**6**) is readily converted to dihydropyrene (**5**) during coal liquefaction (see Fig. 14, [213]). Although the reaction can proceed thermally, it is enhanced by catalysts [214–216]. Thus, polycyclic aromatics can be converted to hydrogen donors in situ and provide hydrogen via a hydrogenation – dehydrogenation cycle. However, naphthalene (**2**) is not converted in significant quantity to tetrahydronaphthalene (**1**) under liquefaction conditions [183], [217]. Dihydropyrene (**5**) can also be formed by the transfer of labile hydrogen from coal [213]. The redistribution of hydrogen by repeated hydrogenation and dehydrogenation is called *hydrogen shuttling* [208].

The presence of solvent can aid in the transmission of H-atoms, generated at catalyst sites, to the coal. The effectiveness of hydrogen distribution by this means will be determined by the solvent composition and it will be accentuated when a significant concentration of H-shuttlers (e.g., polycondensed aromatics) is present. The

H-atoms produced by catalytic dissociation will increase the pool of available hydrogen within the coal or coal – solvent system and will be available for hydrogenation of aromatics, promotion of bond cleavage reactions, and stabilization of radicals [214–216], [218–224].

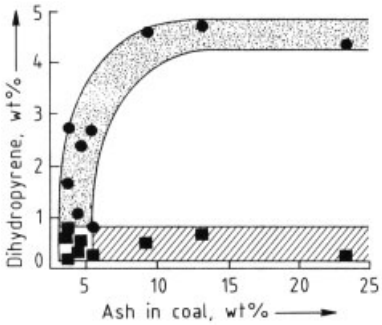

Figure 14. Formation of dihydropyrene during liquefaction [213]
● Hydrogen atmosphere; ■ Argon atmosphere

Certain *heterocyclic nitrogen compounds,* such as piperidine and 1,2,3,4-tetrahydroquinoline, are effective liquefaction solvents [225–228]. They give high coal conversions under mild conditions. However, these compounds are easily and irreversibly incorporated into the products. Such processes evidently involve more than simple hydrogen transfer. Liquefaction solvents typically contain 10 – 15 % nitrogen compounds, and their presence may be important to coal conversion [229].

3.3.2. Coal Conversion and Hydrogen Transfer

Coal liquefaction can be considered to take place in two kinetically distinct stages. First, part of the coal is rapidly converted with low hydrogen consumption. This typically occurs in 5 min or less above 400 °C, with a hydrogen consumption of 0.2 – 0.5 wt % of the coal. Subsequent conversion is gradual and consumes much more hydrogen [141], [208], [230]. Attempts have been made to separate these processes by using two reactors under different conditions (see Section 3.9.6).

The first-stage conversion depends on the structure of the coal; the bulk of the hydrogen demand is probably satisfied by the coal itself. Despite some opposing evidence [208], conversion and product distribution are apparently strongly affected by the composition of the solvent, even at very short reaction times (Fig. 15); conversion increases with increasing H-donor concentration [230–232]. The presence of selected nondistillable liquids can improve both conversion and distillate yield, while allowing a reduction in reaction severity [233]. Reactions that occur during the first stages of coal dissolution may strongly influence the extent of subsequent conversion and the product selectivity [217].

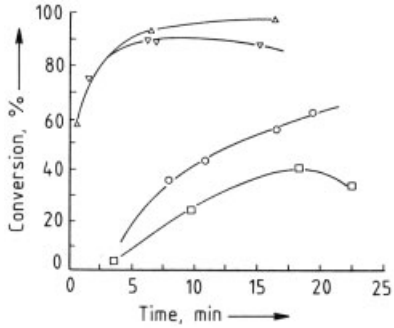

Figure 15. Coal conversion at short times at 427 °C [217]
△ Bituminous coal, H-donor solvent; ▽ Bituminous coal, poor donor solvent; ○ Subbituminous coal, H-donor solvent; □ Subbituminous coal, poor donor solvent

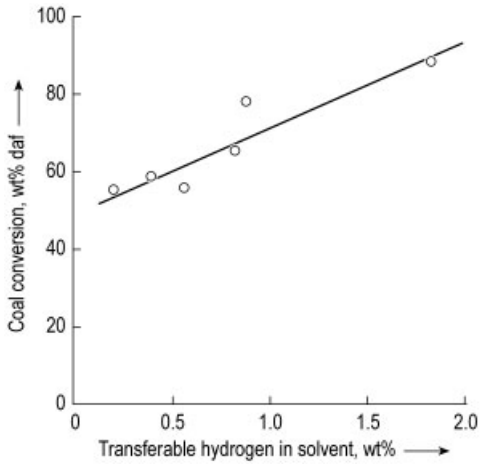

Figure 16. Conversion of Morwell brown coal to benzene-solubles: dependence on transferable hydrogen (TH) content in coal-derived solvent [234]
Process conditions: 425 °C, 60 min, 12 MPa (cold) H_2 pressure, catalyst: Fe_2O_3 + S (1 wt % daf)
Copyright (1984) with permission from Elsevier Science

Second-stage conversion is more clearly dependent on H-donor concentration (Fig. 16). During prolonged reaction in a donor-limited system, conversion may be reduced because of the prevalence of condensation reactions [234]. A high solvent : coal ratio can also retard condensation reactions by increasing H-donor capacity and by diluting the dissolved coal [230] (see Fig. 17). In actual operation, practical considerations limit the ratio to values between 1.5 and 2.0.

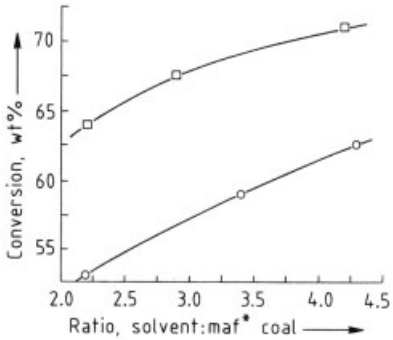

Figure 17. Effect of solvent : coal ratio on subbituminous coal conversion at short times [217]
□ Hydrogenated solvent; ○ Nonhydrogenated solvent
∗maf = moisture and ash-free.

The thermally initiated reactions of coal, especially of low-rank coals, can take place very rapidly at lower temperature producing free radicals. Whether these free radicals are stabilized by hydrogenation or by retrogressive reactions is the critical point in the liquefaction of low-rank coals. The important factors which have significant impact on the hydrogenation or retrogressive reactions are reaction conditions (temperature, H_2 pressure, and reaction time), solvents, and catalysts. The reaction conditions have been studied by many researchers and the procedures for single-staged liquefaction (SSL), temperature-staged liquefaction (TSL), and temperature-programmed liquefaction (TPL) have been proposed [235–237]. For catalytic liquefaction of a subbituminous coal using a dispersed molybdenum catalyst the TPL was found to be superior to SSL in terms of higher conversion and oil yield (see Table 18). Low-temperature catalytic pretreatment or preconversion may be a promising approach to enhance coal conversion due to the suppression of retrogressive reactions.

3.3.3. Interactive Solvent Effects

Coal-derived solvents are highly complex mixtures containing compounds that can beneficially or adversely influence liquefaction. The effectiveness of H-donors can be augmented by the presence of compounds that are strong solvents, such as phenols [209], [238–242]. Combinations of certain solvents act synergistically, e.g., mixtures of tetrahydronaphthalene and pyrene [233]. Process solvent blends, e.g., distillate with nondistillate, can offer advantages to coal conversion because of improved hydrogen transfer or physical solvation [243–248].

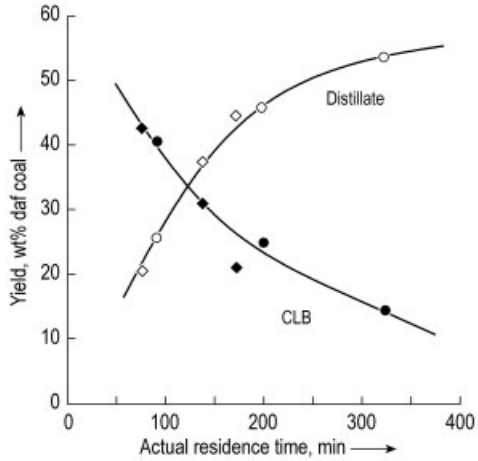

Figure 18. Relation between actual residence time in the reactors and yields of distillate and CLB (coal liquid bottoms: $bp > 420\,°C$): liquefaction of Yallourn coal at 450 °C, 14.7 MPa, and solvent : coal weight ratio 2.5, catalyst $Fe_2O_3 + S$ [249]
○, ● = Solvent (180 – 420 °C); ◇, ◆ = Solvent (220 – 420 °C)
Copyright (1996) with permission from Elsevier Science

The partition of solvent components between the vapor and liquid phase could strongly affect the degree to which they can contact other liquid and solid reactants (coal and catalyst) in the slurry phase. In flow reactors, the stripping of lower boiling fractions by the gas stream will lead to a reduced residence time of these fractions relative to the high-boiling fractions. By reducing the boiling range of the recycle solvent in a continuous flow reactor, the residence time of the high-boiling reactants was selectively increased, leading to an increased oil yield [249–251] (see Fig. 18). An analysis of the composi-

Table 18. Temperature-programmed (TPL) and single-staged liquefaction (SSL) with H-donor tetralin and nondonor 1-methylnaphthalene (1-MN) solvents at 400 °C for 30 min [237]

Coal Feature Solvent	DECS-9 Mont Sub TPL Tetralin	DECS-9 Mont Sub SSL Tetralin	DECS-1 Texas Lig TPL Tetralin	DECS-1 Texas Lig SSL Tetralin	DECS-9 Mont Sub TPL 1-MN	DECS-9 Mont Sub SSL 1-MN
Product dmmf wt %						
THF conversion *	79.2	71.4	78.0	69.8	34.1	32.2
Toluene conversion **	55.5	50.1	66.1	58.0	27.1	25.4
Oil + gas	34.4	29.4	48.5	45.1	18.9	16.0
Asphaltenes	21.1	20.6	17.5	12.9	8.2	9.4
Preasphaltenes	23.7	21.4	11.9	11.8	7.0	6.8

* Total conversion to THF-solubles. ** Total conversion to toluene-solubles plus gas.

tion of the liquid in the reactors under liquefaction conditions confirmed that the increase in gas flow rate (GFR) enhanced remarkably the stripping of the solvent fraction in the feed slurry and the lighter fraction derived from the coal. This results in an increased actual residence time of the liquids consisting of the concentrated catalyst and heavy fractions [251], [252]. Among the high-boiling point components, the hydrocarbon fractions of the recycle solvent can enhance liquefaction, whereas the heterocyclic fractions can be detrimental.

Table 19. Solvent incorporation in coal liquefaction ∗ [258]

Solvent	Solvent molecules incorporated per 1000 C atoms of liquid product		
	As-phaltene	Preas-phaltene	THF ∗∗ insolubles
Tetrahydronaphthalene	8.1	1.6	2.1
Dimethyltetralin	4.0	0.8	1.6
Naphthalene	8.0	1.6	1.8
Octahydrophenanthrene	2.9	0.65	2.0
Phenanthrene	6.3	1.9	2.0

∗ Ratio of solvent to coal, 2.3 : 1; reaction temperature, 450 °C; reaction time 30 min; hydrogen pressure 3.45 MPa.
∗∗ THF = tetrahydrofuran.

Among high-boiling components, the hydrocarbon fractions of recycle solvents can enhance liquefaction, whereas the heterocyclic fractions can be detrimental [253]. The tendency to char is strongly affected by chemical composition [254]. If a solvent contains too high a concentration of constituents that tend to carbonize, such as polyfunctional components, condensation reactions predominate, which result in low conversion [246], [255].

3.3.4. Side Reactions of Solvents

In addition to performing the functions described in Section 3.3.1, solvent components can undergo isomerization, cracking, and polymerization. The resulting losses can lead to solvent imbalance [256–258].

The incorporation of nitrogen-containing solvents into the liquefaction products has already been mentioned (see Section 3.3.1) [209]. Tetrahydronaphthalene dimerizes and is converted to methylindan and butylbenzene [141], [230], [259–264]. The isomerization of hydroaromatic compounds is attributed to the presence of free radicals [265].

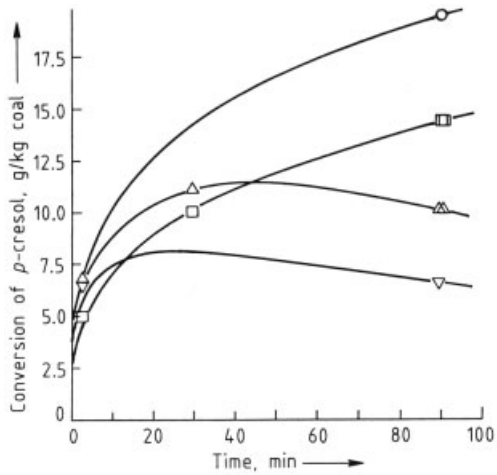

Figure 19. Effect of H-donor concentration on solvent side reactions [149]
Tetrahydronaphthalene concentration: ○ = 0 %; □ = 4 %; △ = 8 %; ▽ = 40 %

Even active H-donors, such as tetrahydronaphthalene and octahydrophenanthrene,

are readily incorporated into high molecular mass liquefaction products (see Table 19). The effect of tetrahydronaphthalene concentration on the extent of side reactions during liquefaction has been studied by mixing it with 2-methylnaphthalene, *p*-cresol, and 4-picoline [149], [266]. At low tetrahydronaphthalene concentration, *p*-cresol forms ethyl and dimethyl phenols (see Fig. 19); 2-methylnaphthalene dimerizes and condenses with *p*-cresol. Reaction with coal can result in solvent alkylation. In pyrene, as many as 1 in 60 of the carbon atoms in coal can participate in the formation of methyl and dimethyl pyrenes [213], [266].

3.4. Coal Composition

The behavior of coal during liquefaction is usually governed by rank, petrographic composition, and inorganic constituents [267]. The macerals of the vitrinite group, which constitutes >80 vol % of the organic content of most coals and of the liptinite group, are generally accepted to be the most reactive in liquefaction. Inertinite macerals, with certain exceptions, are much less reactive [268], [269].

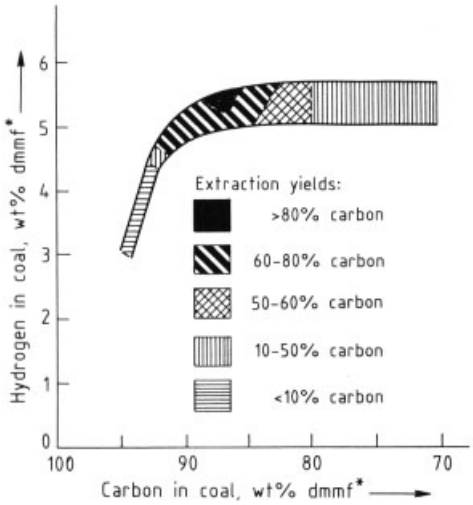

Figure 20. Dependence of conversion on coal rank [270]
∗ dmmf = dry, mineral matter-free.

3.4.1. Coal Rank

The influence of rank on the conversion of a series of British coals in a poor H-donor solvent is shown in Figure 20 [270], [271]. Mid-rank coals (ca. 83 – 88 % maf carbon) liquefy most readily, whereas coals of higher rank are almost inert [272]. Lower rank coals require an additional source of hydrogen (H_2 or H-donor) to attain high conversion levels, as shown in Figure 21.

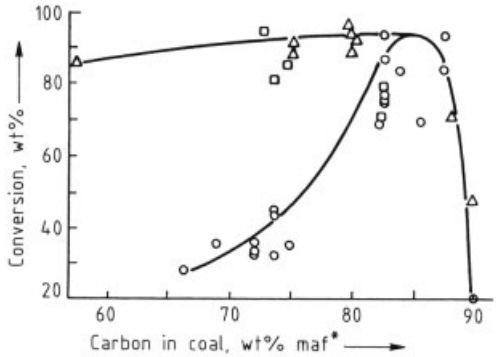

Figure 21. Dependence of conversion on coal rank and hydrogen source [213]
○ Pyrene solvent, inert atmosphere; □ Pyrene solvent, hydrogen atmosphere; △ Tetralin-containing solvent, hydrogen atmosphere
∗ maf = moisture- and ash-free.

Although this account of rank dependence is generally accepted, differences of opinion exist concerning the liquefaction behavior of low-rank coals [15], [268], [269], [273].

3.4.2. Bituminous Coal

A strong correlation appears to exist between the liquefaction and coking properties of bituminous coals [274–277]. Low-molecular mass, hydrogen-rich bitumens present in coal may serve as H-donors and as solvents for molecular fragments as they are liberated [276], [278].

The yield of liquids extractable in chloroform, which reaches a maximum at 85 – 90 % moisture- and ash-free carbon [279], correlates with the fluid properties of the coal [278], [280]. The yields of pyridine and quinoline extracts show a similar correlation with coal fluidity

[281]. Such extraction can destroy the fluid behavior of the coal [3]. Liquefaction yields in nondonor systems are related to the fluid characteristics of the coal, decreasing as the softening point of the coal increases. In an H-donor solvent, high conversions are obtained for most bituminous coals [274]. Coals with relatively low softening points (commensurate with a high content of chloroform-extractable liquids) do not require H-donor solvents for high conversions; the necessary hydrogen can be provided by internal redistribution. The removal of low molecular mass liquids by extraction reduces coal conversion, especially in nondonor systems [282].

3.4.3. Low-Rank Coal

The liquefaction behavior of subbituminous coals and lignites differs from that of bituminous coals. Lignites tend to be more readily converted in moist mixtures of carbon monoxide and hydrogen than in hydrogen alone (see Section 3.6) [283]. Liquefaction of low-rank coals has caused pumping difficulties because of the high viscosity of the residual products. Such problems can sometimes be surmounted by increasing the liquefaction residence time [210] or by adding disposable catalysts (see Section 3.7.2).

Another source of difficulty is the formation of calcium carbonate deposits. Lignites and subbituminous coals can contain up to 3 wt % of calcium, present as humic acid salts. The liquefaction process liberates calcium oxide, which reacts with carbon dioxide to form calcium carbonate. The carbonate can deposit on reactor walls or grow and accumulate in the reactor and other units. Accumulation can be prevented by periodic withdrawal of solids from the reactor or by pretreating the coal with sulfur dioxide gas. Deposits do not form from the resulting stable calcium salts, such as calcium sulfate [210], [284], [285].

Compared with bituminous coals, low-rank coals contain less sulfur and more moisture and organic oxygen. High oxygen content has been associated with the absence of plastic properties, which precludes the liquefaction mechanisms available to bituminous coals [276]. Fluid properties are related to oxygen cross-linking that may be inherent in low-rank, high-oxygen coals [286] or may arise from mild energy input [287], [288].

Low-rank coals convert more slowly and incompletely than bituminous coals. They are more sensitive to solvent H-donor concentration and require more severe reaction conditions to maximize liquid yield [210], [255], [276], [289]. Formation of water (from organic oxygen) and gas is high, accounting for greater hydrogen consumption [289]. Most of the oxygen in low-rank coals forms carbon oxides on liquefaction (above 400 °C) [290].

Although total liquid yield tend to decrease with decreasing rank, the proportion of low-boiling distillates increases [291]. Early German work suggested that, compared to bituminous coals, the low-rank coals contained smaller polynuclear units [15]. The liquids produced by bituminous and subbituminous coal liquefaction have been analyzed in detail [291]. The subbituminous coal products were more volatile, contained more saturated hydrocarbons, were less aromatic, and had a higher hydrogen content and fewer heteroatoms.

3.4.4. Coal Drying and Oxidation

Drying or storage in an oxidizing atmosphere (aging or weathering) can deleteriously affect coal liquefaction behavior [276]. Even exposure to mild oxidizing conditions can reduce yields, partly because of polymerization of coal-derived liquids [213], [292–295]. Oxidation similarly reduces the yield of pyrolysis liquids while increasing their aromaticity and molecular mass [296], [297].

Oxidation of bituminous coals adversely affects plastic properties (see Fig. 22) and the yield of extractable liquids (as shown by the following data) [2], [298–302]:

Weathering time, d	Mass fraction extracted in tetrahydrofuran (dmmf ∗), %
0	21.3
5	19.6
35	18.3
56	14.5

∗ dmmf = dry, mineral matter-free.

Plastic properties can be restored by reduction [298].

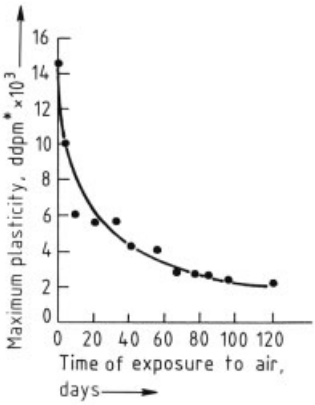

Figure 22. Reduction in coal fluidity with progressive oxidation [299]
∗ ddpm = dial divisions per minute.

The mechanism of coal oxidation has been studied [299], [301–303]. Oxidation introduces ether, carbonyl, carboxyl, and phenolic groups, which results in the loss of aliphatic structures. The conditions of oxidation and the structure of the coal appear to have a strong influence on the extent of oxygen uptake and the functional group distribution. The reduction in plastic properties has been attributed to formation of stable ether cross-links and to loss of aliphatic groups [301], [302]. Cross-linking may also result from functional group condensations or from the formation of unstable ethers, which proceed to cleave at still higher temperature, leading to carbon – carbon bond formation [276]. The oxidation tendency generally increases with decreasing rank, as can be seen from Table 20 [298], [303].

A drying process, with minimal effect on the reactivity of coal, is important in coal liquefaction studies [304]. Drying of coal can cause an irreversible change in the pore structure such as collapse of pores, which could limit the accessibility of the reacting components during liquefaction [305–311]. A negative impact of severe oxidation on coal liquefaction is considered to arise from the significantly increased oxygen functionality which enhances the cross-link formation in the early stage of coal liquefaction.

Drying lignites before liquefaction lowers the yield of light oil products, possibly because of oxidation [304–310], [313–315]. These decreases in yield can be reduced by driving off moisture after the coal is slurried in the liquefaction solvent [312], [316].

3.5. Gaseous Hydrogen

Reaction under hydrogen pressure provides a source of stabilizing hydrogen in addition to that available from the solvent and the coal. The consumption of hydrogen gas increases with decreasing H-donor solvent concentration, even at short reaction time, and the presence of hydrogen gas can alleviate constraints on solvent quality [317].

Gaseous hydrogen may react directly with coal or through intermediate solvent hydrogenation. Partial hydrogenation of compounds, such as quinoline and pyrene, can produce active H-donors by thermal reaction [217], [318], but catalytic processes are regarded as being of primary importance (see Section 3.7). Molecular hydrogen may also participate in free-radical reactions, competing with donor hydrogen and promoting the cracking of C – C bonds [143].

3.6. Liquefaction in $CO - H_2O$ and $CO - H_2 - H_2O$

Since the economy of the direct coal liquefaction process using high-pressure pure molecular hydrogen is aggravated by the cost of hydrogen production, development of a process using an alternate hydrogen source is of great importance [319–329]. Excellent accounts of coal liquefaction in mixtures of water with carbon monoxide or with synthesis gas are available [183], [315], [330]. Low-rank coals give higher yields in a mixture of water and carbon monoxide than in hydrogen. This advantage decreases with increasing coal rank; for bituminous coals hydrogen is superior [283], [331]. Reactions of some Victorian brown coals with $CO - H_2O$ in the presence of sodium aluminate have been shown to give very high conversions to products (THF solubles) under mild conditions (see Table 21). The influence of water loading, of reaction time and temperature, of sodium aluminate concentration, of gas composition, and of CO partial pressure have been investigated [326]. Oil yield was significantly higher for $CO - H_2O$ than for H_2 – tetralin in the liquefaction of brown coal.

Table 20. Effect of oxidation on oxygen distribution [303]

Carbon	Oxygen gain, g/kg dry coal						
in coal, wt %	In gas			In solid			Total
	CO_2	CO	H_2O	COOH	OH	C=O	
89	1	0	3	1	11	1	17
87	2	0	3	0	3	1	9
85	3	0	5	−1	5	2	14
84	4	1	8	1	9	1	24
67	36	3	34	15	4	9	101

Table 21. Effect of $NaAlO_2$ concentration on conversion and products yields of Loy Yang coal * in CO – H_2O system [326] **

			Product yields (wt % daf)		
Catalyst loading, mmol/kg dry coal	Reaction time, min	Conversion (THF solubles)	Asphaltols	Asphaltenes	Oil + gas + water (by diff.)
1000	30	89	10	18	61
500	30	77	7	15	55
1000	60	88	9	15	64
500	60	89	11	16	62
0	30	51	1	5	45

* Reaction conditions: 15 g coal, 60 g water, $NaAlO_2$ at 3 MPa (cold) CO, 375 °C.
** Copyright (1994) with permission from Elsevier Science.

The CO – H_2O system requires a basic catalyst, such as an alkaline component or a basic nitrogen-containing compound in the liquefaction solvent. Because most lignites contain high proportions of Na^+ and Ca^{2+} ions, they possess an indigenous catalyst. The addition of sodium formate, alkali carbonates, iron oxide, or metal chlorides promotes activity [332].

Carbon monoxide may function by reducing the carbonyl groups in the coal, thus minimizing condensation reactions. The mechanism may involve formate intermediates produced by the reaction of CO and H_2O with alkali compounds or nitrogen bases [322–325]. Above ca. 400 °C, the formate begins to decompose to produce carbon monoxide and molecular hydrogen. The fact that the latter is less reactive than the formate hydrogen contributes to the reduced effectiveness of CO –water reactions at higher temperature [315].

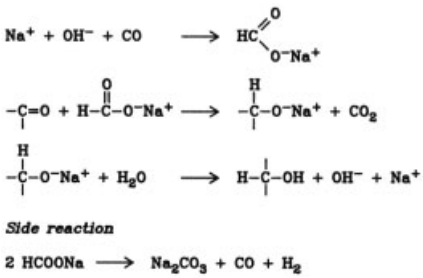

Hydrogen is also supplied by the water – gas shift reaction (see Section 2.1.2). Hydrogenolysis by nascent hydrogen from the shift reaction may be one of the reasons for effective coal liquefaction [333].

For economic reasons, there is an incentive to replace carbon monoxide by less expensive synthesis gas. Conversion of lignites and subbituminous and bituminous coals in CO – H_2 – H_2O mixtures is different than in CO – H_2O [315]. The highest reactivities were obtained with high-volatile bituminous coals (hvc) having high mineral and sulfur contents. The reactivity of subbituminous coals correlated well with the sum of their iron, sulfur, and alkali-metal contents, and the liquefaction of lignite was greatly improved by the addition of pyrite. Increased consumption of CO indicated that pyrite also catalyzes the water – gas shift reaction. A synthetic pyrite of large surface area was found to be one of the most excellent catalyst precursors for the liquefaction of Australian brown coals in the CO – H_2 – H_2O system. Similar results were obtained with a $Fe(CO)_5$ + S catalyst [326–328]. The presence of CO or H_2O may have a positive side effect in enhancing the dispersion of Mo catalysts [329].

The behavior of both H_2 and CO is affected by mineral matter and added catalysts. However,

these gases function differently. Carbon monoxide is most effective below 400 °C, supposedly by inhibiting cross-linking reactions. Hydrogen is more reactive at higher temperature, participating in hydrocracking reactions that can increase oil yield and reduce product viscosity. Although the application of CO-containing gases has been the subject of numerous liquefaction studies, this approach has never emerged as a serious candidate for large-scale development.

3.7. Catalysis

The use of extraneously added catalyst is essential to developing a viable liquefaction process. Catalysts have been used either in the dispersed form in the coal – solvent slurry, or as supported metals in fixed or ebullated bed reactors [334–338]. The former are used to promote the primary hydrogenation of coal, while the latter are most effective for upgrading of coal-derived liquids. Improved performance can be realized both during primary coal liquefaction and coal liquids upgrading by using suitable catalysts. The advantages of application of catalysts include accommodation of a wider range of coal feeds, higher selectivity to liquid products with reduced light gas yield, reduced reaction severity, and improved product quality. Effective catalysis can also help to reduce investment and operating costs through lowering the reaction severity. More efficient use of hydrogen can greatly improve process economics [339]. The following cost analysis (1998 – 2000) gives the range for the HTI-Coal process, NEDOL process, and Improved BCL process (see Section 3.9).

Capital cost (of coal)	\$ 170 200 – 199 000 t^{-1} d^{-1}
Liquid products	0.53 – 0.71 m^3/t of coal
Capital cost (of syncrude)	\$ 248 000 – 372 000 m^{-3} d^{-1}
Capital cost distribution	
Liquefaction	39 – 49 %
Hydrogen manufacture and cleanup	19 – 32 %
Others (power, utilities, tankage, and offsite facilities)	19 – 42 %
Total costs	
Coal	20 – 27 %
Capital	45 – 57 %
Operations	17 – 35 %

Pilot plant tests at Wilsonville in the United States have demonstrated that the use of dispersed catalysts could be superior to supported catalysts for primary liquefaction of coal, particularly low-rank coals such as subbituminous coals or lignite [340], [341]. Liquefaction processing and economics may be improved more efficiently through the development and appropriate application of highly dispersed catalysts [342].

3.7.1. Mechanisms

Coal liquefaction catalysts function in a number of ways as shown in Figure 23. In the initial stage of coal liquefaction thermal decomposition of coal occurs to give free radicals. Free radicals thus formed from coal can be stabilized by hydrogen abstraction from either hydrogen donor substances or molecular hydrogen activated by a catalyst. Hydrogenation of aromatic component and hydrocracking of C – C bonds in the coal-derived initial products could result in a decrease in their molecular mass to give lighter oil fractions. In the early stages of coal liquefaction, promoting hydrogenation reactions appears to be the most critical requirement.

Liquefaction with a highly dispersed iron catalyst in tetralin and tetralin – naphthalene mixtures suggests that the most significant role of the catalyst is to activate molecular hydrogen and transfer it to coal fragment radicals [343]. Catalysts may indirectly increase hydrogen transfer to the coal by hydrogenating solvent species to produce H-donor compounds. Consequently, retrogressive reactions of thermally decomposed coal fragment radicals are suppressed and increased coal conversion and yield of oil fractions are obtained.

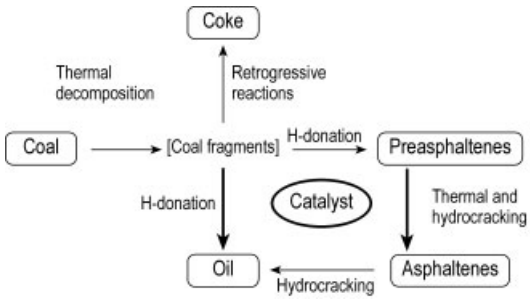

Figure 23. Catalyst functions in direct coal liquefaction

Iron-based catalysts exhibit a much lower activity for the heteroatom removing reactions,

such as hydrodesulfurization (HDS) and hydrodenitrogenation (HDN) than molybdenum-based catalyst. Heteroatoms, such as O, S, and N, in aromatic rings require sulfided Ni – Mo or Co – Mo catalysts for their expensive removal to the levels experienced in petroleum refining [338].

Supported catalysts, consisting of metals deposited on high surface area supports, have the advantage of high dispersion at relatively low metal concentrations. However, a large portion of the metal may be contained in pores whose size restricts the access of large molecules, thereby rendering those site ineffective for facilitating the dissolution of solid coal particles. Consequently, supported catalysts are more suitable for hydroprocessing of distillate liquids.

3.7.2. Disposable Catalysts

Disposable (i.e., once-through) catalysts are materials whose recovery is not economically essential. Catalyst cost is low because the material is cheap or because it is highly active at low concentration.

Some coal minerals are catalytically active [344], but the content and composition of these components are too variable for stable process operation. However, research has been conducted on mineral matter catalysis to identify prospective disposable catalysts. Dispersed catalysts generally have very little internal porosity and the active sites are on the exterior surfaces of the particles. These catalysts are more effective for promoting the initial conversion of coal. A variety of metals have been tested for possible use as dispersed catalysts over the years, including molybdenum, iron, tungsten, cobalt, nickel, tin, zinc, and mixtures of these metals. The high dispersion of the catalyst makes recovery difficult if not impossible, and therefore the cost of the catalyst is a critical factor.

Pyrite and other iron compounds have attracted widespread interest as disposable catalysts [345], [334]. Molybdenum and tin, used in Germany during the 1930s and 1940s, were replaced by less active, cheaper iron ores [15], [346]. In the liquefaction of low-rank coals, molybdenum (0.2 wt % daf) was as effective as a high concentration of pyrite (2.5 wt % daf); however, the latter was preferred because of cost [347, part 2]. More recently, there has been renewed interest in the use of iron-based catalysts as disposable catalysts because of economic and environmental benefits. Highly active iron-based catalysts have been extensively investigated in Japan and the United States [335], [342].

Clays. The mineral content of most coals consists largely of clays, which have complex polymeric aluminosilicate structures (→ Clays). The hydroxyl groups attached to Si and Al are strongly acidic and provide active sites for catalysis. Kaolinite and montmorillonite catalyze hydrogen transfer between hydroaromatic and aromatic compounds, and kaolinite increases the conversion of coal to gas and benzene-soluble products [348], [349]. Natural clay minerals may possess cracking activity.

Iron Sulfides. The activity of iron catalysts is generally low, necessitating the use of relatively high concentrations (several wt % on daf coal) of iron. A variety of methods have been employed to increase the activity of iron-based catalysts [350]. Catalyst precursors have been introduced by physically mixing very small particles into the coal – solvent slurry [351–361], by addition of iron sulfides precipitated from aqueous solutions [362], by addition of the catalyst in oil-soluble forms [363–369], and by impregnation of the feed coal with water-soluble compounds [370–374].

The activity of iron based catalyst depends on their composition and the extent of their dispersion. The extent of dispersion is directly related to the catalyst effectiveness. Until the coal is dissolved, there is no available mechanism by which reactants can be transported to the catalyst, and hence the basic requirement is to transport the catalyst to the reacting species. Some progresses are being made in the development of techniques to determine particle size and size distribution [375].

The synthesis of nanometer size catalyst particles could provide a means to enhance the activity of dispersed iron-based catalysts. Several techniques have been used to produce this type of precursor, including flame pyrolysis [352], laser pyrolysis [357], and a hydrothermal process [358]. The flame pyrolysis of metal chloride vapors has been used to produce aerosol oxides

with particle diameter of < 50 nm and surface areas of $20-50\ m^2/g$. Iron oxide particles generated by this technique have demonstrated a high catalytic activity for the liquefaction of bituminous coal, resulting in conversions only slightly lower than obtained with a similarly produced molybdenum oxide [376].

Table 22. Effect of FeOOH mode of addition on coal conversion at 425 °C, 0.5 h, 1000 psig (6.9 MPa) (cold) of H_2, 5000 ppm Fe, 0.1 g of S added to 9.9 g of a 2 : 1 mixture of PANASOL* to coal [374]

Precursor	Coal conversion to, %	
	THF solubles	Heptane solubles
Blind Canyon		
No catalyst	58	30
Physically mixed FeOOH	66	34
Impregnated FeOOH	85	41
Illinois no.6		
No catalyst	53	25
Physically mixed FeOOH	67	40
Impregnated FeOOH	83	50
Black Thunder		
No catalyst	54	30
Physically mixed FeOOH	64	35
Impregnated FeOOH	79	39

* PANASOL = a mixture of alkylated naphtalenes (Crowley Chemical Co.)

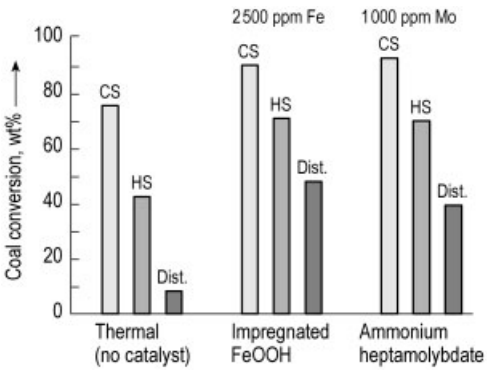

Figure 24. Coal conversion in semibatch tests of Illinois, No. 6 coal and V-1074∗ solvent in thermal mode and with added iron and molybdenum. Semibatch tests of Illinois No. 6 coal and V-1074, 30 wt % coal, 425 °C, 1 h, 2500 psig (17200 kPa), 4 SCFH ∗∗ H_2/3 % H_2S [337]; CS = CH_2Cl^2 solubles; HS = Heptane solubles; Dist. = Distillates < 950 °F (< 510 °C)
∗from the Wilsonville Test Facility
∗∗SCFH = Standard Cubic Feet per Hour (ft^3/h), 1 SCFH = 0.0283 Nm^3/h
Copyright (1994) with permission from Elsevier Science

A laser pyrolysis technique has been used to produce a variety of iron-based particles including two phases of iron carbide (Fe_3C and Fe_7C_3), two phases of iron nitride (Fe_3N and Fe_4N), an iron sulfide ($Fe_{1-X}S$), and α-iron. The particles were in the range of 4 – 20 nm in diameter. Experimental studies have shown that the carbides transform to pyrrhotite ($Fe_{1-X}S$) under liquefaction conditions and that the addition of sulfur appears to inhibit sintering. These particles have demonstrated moderate catalytic activity for the conversion of bituminous coal [377].

High activities have been reported for catalysts introduced as oil-soluble organometallic precursors, such as carbonyls and naphthenates [378], [363–370], [379]. The results of some studies indicate that, even with the soluble precursors, quite large crystallites or agglomerates can be formed during liquefaction and hence the potentially high dispersion is not maintained [335]. The use of an incipient wetness technique to impregnate the coal with Fe^{3+} from an aqueous solution has been shown to give high dispersion of the catalyst precursor on a coal substrate (see Table 22) [337], [374]. The coal matrix may also aid in maintaining the high dispersion of catalyst particles subsequent to the activation. As a catalyst loading of 2500 ppm the precipitate FeOOH yielded the same catalytic activity as 1500 ppm molybdenum, added as ammonium heptamolybdate (see Fig. 24) [337].

Another route to improving the activity of iron catalysts is to modify their composition. One approach has been to change the surface properties of hematite particles either by treating the precipitates with H_2SO_4 or by precipitating in the presence of sulfate anions, resulting in the formation of Fe_2O_3/SO_4^{2-} particles after calcination [232], [380–385]. Sulfated Fe_2O_3, SnO_2, and Mo/Fe_2O_3 have focused attention to the catalytic activities in coal liquefaction [380], [383]. The sulfate group is believed to be responsible for the superacidity of catalyst, and it can also help to prevent sintering of the catalyst particles, which may relate to their enhanced catalytic activities [380].

The effectiveness of iron-based catalyst can be greatly improved by the addition of small concentrations of molybdenum [383], [386–388]. The activity of Mo-added iron oxide aerosol catalysts was higher than that of supported Ni – Mo catalysts [388]. These fundamental studies have provided evidence for an increase in catalytic activity with decreasing iron particle size, decreas-

ing availability of sulfur donating species, and suppression of the crystal growth of iron sulfide. Elemental sulfur was often added with the iron oxide catalyst when low-sulfur coals were processed. Iron oxide can be converted to sulfide in situ and the active form of iron catalyst is generally believed to be a pyrrhotite [15], [17], [204], [345],[346], [389–399]. For the iron catalyst, the activity reached a maximum level at an S : Fe atomic ratio > 1. In many cases, the attainment of this phase relies upon in situ sulfiding, and there is scant information on the relative kinetics of dissolution and sulfiding reactions: the rate of conversion of the precursor to a sufficiently active form may, in some cases, be too slow to influence the process of coal dissolution. The mechanism of catalysis by iron sulfide is controversially discussed [204], [345], [389–405]. The roles of pyrite and pyrrhotite are unclear. Pyrrhotite is a nonstoichiometric mineral, which may be catalytically active: it has a variable number of lattice vacancies. The partial pressure of H_2S is important because it influences the concentration of metal vacancies [389], [400–403], [406]. Another mechanism suggests that sulfide minerals generate H_2S, which acts as a hydrogen donor and initiates the free-radical chain reactions [344], [403], [404]. In the working states, the active sites may be sulfate species formed on the surface of iron-based catalyst during the coal liquefaction, because the catalyst may be always exposed to oxidants in the reaction system [405].

Iron-based catalysts are used in Germany [24], [38], [39], [273], [408], Japan [41], [44], [342], [353–355], [407], [409][411] and the United States [22], [32], [347].

In the German IGOR process, a disposable iron catalyst known as "red mud" or Bayermasse is used at a concentration of ca. 4 – 6 wt % daf; it contains ca. 30 % iron(II) oxide and some abrasive material such as alumina [15], [39], [273], [407], [408]. Removal of inorganic matter before the liquefaction can afford processing benefits and extend the life of plant equipment [412]. The use of high concentrations of low-activity catalyst, often associated with abrasive material, may lead to operating problems.

In the NEDOL process of Japan, finely pulverized natural pyrite catalyst is used at a concentration of 2 – 3 wt % daf in the liquefaction of subbituminous or bituminous coal [409–411]. Natural pyrite was dried to 1.5 wt % moisture, and then finely pulverized to 0.7 μm particle size in the process solvent by a two-stage pulverizing method; to 2 μm by a ball mill in the first stage and then to 0.7 μm by an agitated beads mill in the second stage [409]. Finely pulverized pyrite catalyst increases coal conversion and the selective production of oils [413], [414]; milder operating conditions become possible and gas formation is reduced. Pyrite catalysis facilitated the continuous processing of low-rank coals [347], [347, part 2]. Highly active iron catalysts (synthetic γ-FeOOH and finely pulverized limonite ore) have been developed in the Improved BCL process of Japan for the liquefaction of brown coals [342], [353–355], [415]. Both catalysts are well dispersed in the coal – solvent slurry by pulverization in the process solvent to submicron particle size before the coal liquefaction. Synthetic γ-FeOOH (30 nm crystallite size) catalyst exhibits a high oil yield over 60 wt % in the liquefaction of Yallourn coal even at the catalyst loading of 0.3 wt % as Fe with partial recycling [415].

This γ-FeOOH catalyst can be transformed into the smaller crystallite size of pyrrhotites at lower temperatures around 250 °C by the addition of elemental sulfur (see Fig. 25). The amount of elemental sulfur added to the iron catalyst was controlled at 1.0 vol % of H_2S in the reactor in order to suppress both formation of less active troilite (FeS) and the crystal growth of pyrrotite particles [355].

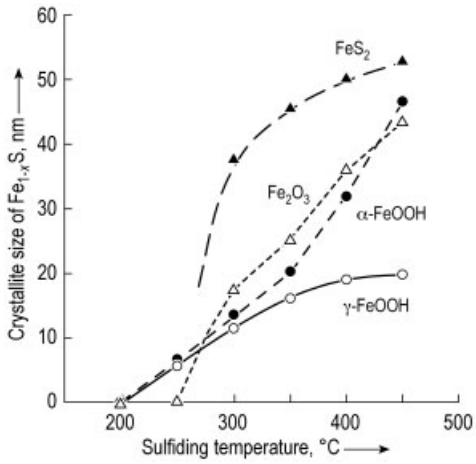

Figure 25. Increase in crystallite size of pyrrhotites ($Fe_{1-X}S$) with sulfiding temperature at 0 min of holding time, 100 °C/min of heating rate, S : Fe ratio 2.0 with no coal in coal-derived solvent [355]

Table 23. Oil yield in the liquefaction of Yallourn coal with fresh or used * catalyst. Liquefaction conditions: 450 °C, 1 h, 200 g daf of Yallourn coal, 500 g of coal-derived solvent, 6.0 g of catalyst as Fe, 4.13 g of elemental sulfur (S : Fe = 1.2), 7.5 MPa initial H_2 pressure [416]

Catalyst		Pyrrhotite crystallite size, nm	Oil yield ** (wt % daf)
γ-FeOOH	fresh	34	49.3
(Synthesized powder)	used (CLB-THFI)	42	44.6
Limonite (YE)	fresh	34	49.1
(Yandi Yellow, Australia)	used (CLB-THFI)	41	46.9
Limonite (RO)	fresh	41	45.0
(Robe River, Australia)	used (CLB-THFI)	50	40.7
Pyrite	fresh	63	42.4
(Mt. Lyell, Australia)	used (CLB-THFI)	66	35.7

* CLB – THFI (THF insolubles in coal liquid bottoms) was employed as used catalyst.
** Oil yield: distillate oil (C_5 – 420 °C) yield.

Limonite ore consists of natural goethite (α-FeOOH), containing a considerable amount of hematite (α-Fe_2O_3). The hematite content can affect the catalytic activity: the lower the hematite content the higher the coal conversion. The activity of a finely pulverized limonite catalyst based on a specific limonite ore (named "Yandi Yellow") deposited in Australia, was comparable to that of γ-FeOOH catalyst (see Table 23). This limonite catalyst looks promising as a disposable catalyst in the commercial plant of BCL process due to its high catalytic activity, low pulverization cost, and abundant resources [416].

In HTI' s direct coal liquefaction process a proprietary dispersed superfine iron catalyst, "Gel Cat", is used [22]. Unfortunately, little information is available on its physical properties and chemical composition.

Nonferrous Metals. Molybdenum and tin are active liquefaction catalysts that were used in Germany before being replaced by cheaper, less active iron catalysts [346]. As little as 0.05 % molybdenum gave satisfactory results in brown coal liquefaction, provided that the coal alkalinity was first neutralized with sulfuric acid [417]. For bituminous coals, higher concentrations were required [15], [346].

In a plant at Billingham, United Kingdom, tin compounds and hydrochloric or hydroiodic acid proved to be active for bituminous coal liquefaction. The usual combination was 0.04 –0.06 % tin oxalate and 0.4 – 1.0 % ammonium chloride on the basis of feed. A similar catalyst was employed in some German plants, but because of scarcity and corrosion problems, was replaced by iron catalysts [15].

Many disposable liquefaction catalysts have been investigated, some in batch-autoclave experiments [418–426]. Tin promoted with a halogen was highly active; germanium oxide promoted with ammonium chloride and molybdenum, when applied by impregnation, were comparable [417]. Metal naphthenates were the most effective catalysts studied.

The influence of the mode of catalyst addition can be seen from Table 24 [421]. For molybdenum, the disulfide appears to be the active form, its effectiveness reaching a maximum when reaction with sulfur gives an S : Mo ratio of 2.0 [378].

Although iron remains the catalyst metal of choice because of cost and availability, the effectiveness of well dispersed molybdenum at low concentrations has led to much research on precursors based on that metal, and on possible recovery and recycle of the catalyst. Interest also remains in determining the chemical fate of precursors, and on using catalysis as a tool in further understanding the complex mechanism of coal liquefaction.

Extensive prior studies in the United States have demonstrated the potential of a dispersed molybdenum catalyst for coal liquefaction. In the Dow process, highly dispersed molybdenum was used as a disposable catalyst with partial recycling in the overflow of a hydrocyclone used for solids separation [30], [427]. Ammonium molybdate was added to the coal – solvent slurry as a water-in-oil emulsion at 100 – 200 mg per kilogram of slurry. Some molybdenum was recycled with the process solvent, increasing the concentration of active catalyst by a factor of 2.0 – 2.5. The microemulsion of molybdate is converted to very small particles of MoS_2 under

Table 24. Effect of catalyst distribution on conversion of a bituminous coal in the absence of solvent ∗ [421]

Catalyst	Catalyst distribution	Coal (maf ∗∗), %			
		Total conversion	Asphaltene	Oil	Gaseous hydrocarbons
$SnCl_2$ (1 % Sn)	powder added	82.3	26.5	29.3	14.5
$SnCl_2$ (1 % Sn)	impregnated	88.3	19.9	41.4	15.5
$NiCl_2$ (1 % Ni)	powder added	44.2	6.8	13.2	13.4
$NiCl_2$ (1 % Ni)	impregnated	88.3	15.5	45.3	18.0
$FeSO_4$ (1 % Fe)	powder added	38.9	6.9	8.1	13.1
$FeSO_4$ (1 % Fe)	impregnated	84.9	38.9	21.7	15.0
$(NH_4)_2MoO_4$ (1 % Mo)	powder added	33.7	1.0	13.8	8.1
$(NH_4)_2MoO_4$ (1 % Mo)	impregnated	92.7	27.2	41.1	13.6
None	–	33.4	2.8	10.4	9.0

∗ Reaction 1 h at 450 °C, 7 MPa initial H_2 pressure.
∗∗ maf = moisture and ash-free.

the reaction conditions, and a substantial portion of these species survives for recycling to the reactor.

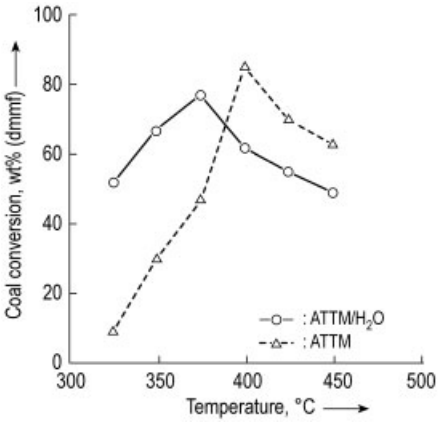

Figure 26. Effect of reaction temperature on conversion of DECS-8 Wyodak coal with ATTM in the presence and absence of added water without using an organic solvent [429] Copyright (2000) with permission from Elsevier Science

The nonstoichiometry of molybdenum disulfide originating from ammonium tetrathiomolybdate (ATTM) is also of interest, since the exact composition may play a role in ATTM-catalyzed liquefaction [428]. The conditions used for impregnation can strongly influence catalyst dispersion and activity in dry catalytic hydrogenation with coal impregnated with ATTM. In particular, water removal after impregnation gave greatly improved dispersion of Mo, as evidenced by scanning electron microscoppy/energy dispersive X-ray spectroscopy. A strong promoting effect of water addition on catalytic conversion using ATTM was reported for the liquefaction of subbituminous coals at low temperature of 350 °C (see Fig. 26) [429]. Temperature-programmed liquefaction using ATTM with added water at 350 °C, followed by water removal and subsequent reaction at 400 °C gave good coal conversion and oil yield [430]. A plausible extension of the ATTM concept is the use of its tungsten analogue, ammonium tetrathiotungstate (ATTW), as a precursor [431], [432].

Another type of catalyst materials are the chlorides of transition metals, such as $ZnCl_2$ and $SnCl_4$ [433–438]. Also the combination of a mild superacid (CF_3-SO_3H) and iodine was studied. Iit was reported that the acid enhances the depolymerization of coal to asphaltenes and that the major effect of iodine is to hydrogenate and hydrocrack asphaltenes to oil [439]. These catalysts are primary Friedel – Crafts type catalysts, which because of their acidic function catalyze C – C bond cleavage. The corrosive nature and instability of these catalysts may exclude their practical application.

3.7.3. Supported Catalysts

Supported catalysts are used in the form of 1.5 mm or larger particles in a fixed-bed [15], [440] or ebullated-bed reactor [32–34], [334–336]. Because the catalyst pores provide virtually all of the active surface area, the catalyzed reactions probably do not directly involve the coal, but rather the solvent and the liquefaction products. In some processes, primary coal dissolution is carried out in a separate reactor; in

others, most of the coal may dissolve during passage through the preheater [183].

In the US H-coal process, Ni – Mo or Co – Mo/Al_2O_3 catalysts have been used in the ebullated-bed reactor. Ebullated-bed reactors provide for catalyst addition and withdrawal, thus maintaining activity during operation [5]. The performance and deactivation characteristics of catalyst used in the Wilsonville Advanced Coal Liquefaction Test Facility were examined [441–443]. Catalyst deactivation is due to the buildup of both carbonaceous deposit and contaminant metals; steam sintering is somewhat less important [444]. A great deal of research has been devoted to examining the deactivation of supported catalysts during coal dissolution and upgrading [445–450].

Carbon deposition is rapid during the first 30 – 50 h on stream and is accompanied by a loss of surface area (see Fig. 27) and pore volume [452], [453]. After this decline, catalyst activity changes more gradually. Deactivation by metal deposition is slower (see Fig. 28) [454], the chief contaminants being iron, titanium, and boron [455]. The metal precursors are organometallics or other soluble compounds. Particulate minerals apparently do not contribute directly to poisoning, which is caused primarily by carbonaceous deposits distributed throughout the catalyst particles [183], [454]. Deactivation by metals takes place through a shell-progressive mechanism [454]. Contaminant metals cause loss of intrinsic activity by covering active sites and also cause some minor permanent loss of pores due to the pore-mouth blockage. The metal deposits are concentrated in an outer layer of the aged catalysts and may cause deactivation of the catalyst due to slow diffusion through this barrier. Carbonaceous deposits cause deactivation presumably due to active-site coverage and yield a decrease in catalyst average pore size due to the filling of larger pores. Nitrogen compounds, aromatics, and phenols may act as precursors for carbon deposits [456], [448]. Sintering and metal losses are among the other causes of catalyst deactivation [455], [450].

The earliest coal liquefaction catalysts were designed for hydrofining petroleum distillates and were ineffective for high molecular mass coal liquids because of their small and easily clogged pores [417], [457]. Increasing the proportion of larger pores has improved activity [417], [458], [459]. Catalyst requirements are similar for upgrading coal liquids and high-boiling-point petroleum residues [460].

Calcium-doped Ni – Mo/Al_2O_3 with bimodal pore distribution has been developed as a secondary hydrogenation catalyst in the BCL process, which has been shown to have a stable activity for about 8000 h, included the 3400 h operations of a 50 t/d pilot plant in Australia [461–465]. Doping of alkaline-earth metals, such as Ca or Mg, modifies the acidic properties of the catalyst and decreases the carbonaceous deposition on the catalyst in the hydrotreating of coal-derived liquids [465–468]. The modification of the acidic properties of the catalyst could be an effective way to suppress its deactivation (see Fig. 29).

Continuous reactor tests performed by Amoco have shown that the activities of Ni – Mo/hydrous titanium oxide (HTO) catalysts were significantly greater than that of commercial Ni – Mo/Al_2O_3 catalysts for the second-stage upgrading of coal-derived residues [469–472]. The catalysts prepared from HTO-coated supports, containing one-tenth the active metals of commercial alumina-supported catalysts, have been found to exhibit high hydrogenation and desulfurization activities based on weight of Mo. HTO-coated support materials may offer a promising method to produce cost-effective catalysts with mechanical, physical, and catalytic properties tailored to specific direct liquefaction process applications [472].

The high activity of Ni – Mo catalyst supported on carbon nanoparticles with extremely high surface area and low density has been demonstrated (see Fig. 30) [473–481]. This high activity is attributed to the high dispersion of active metals on the KB carbon nanoparticles. These catalysts, being hollow carbon particles, have a very low density and can potentially be recovered from the coal liquefaction products by gravimetric separation. Various combinations of sulfides such as Fe – Ni and Fe – Mo can also be candidates for catalysts with higher activity and lower cost [480].

Removal of heteroatoms is another important consideration in catalytic upgrading of crude liquefied oils. The storage stability of crude liquefied oil should be improved by a hydrotreatment step in the coal liquefaction process, as the heteroatoms in liquefied oils can cause the forma-

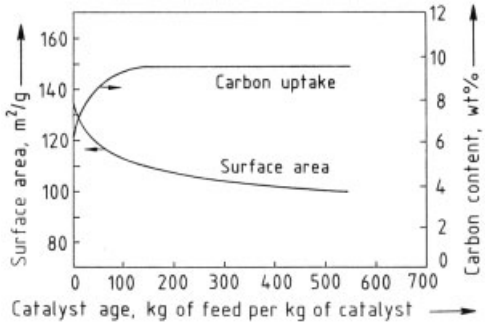

Figure 27. Changes in surface area and carbon deposition with catalyst age [451]

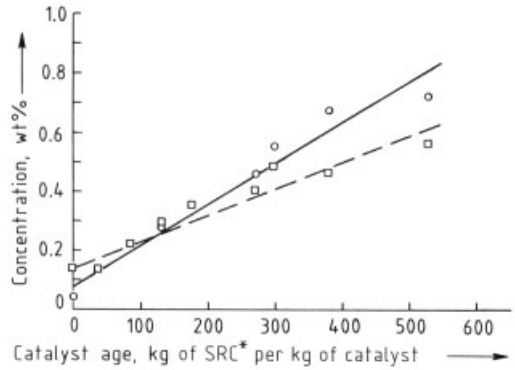

Figure 28. Accumulation of metals as a function of catalyst age [454]
–○ Fe; – – –□ Ti
∗ SRC = solvent-refined coal.

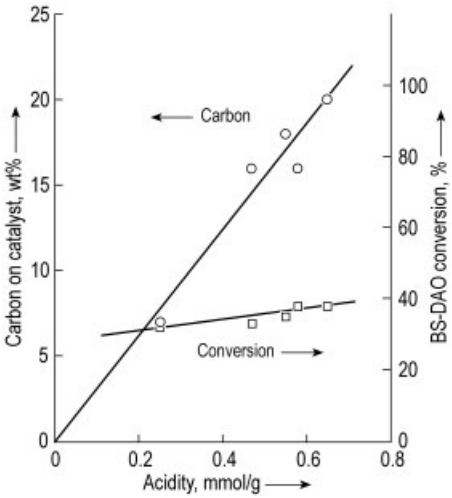

Figure 29. Effect of acidic properties of catalyst on formation of coke-like substances and BS-DAO (Benzene Solubles in De-Ashed Oil, b.p. > 420 °C) conversion at 400 – 450 °C. Acidity was defined as the amount of pyridine adsorbed at equilibrium on the catalyst in benzene solution at room temperature [461]. Copyright (1990) with permission from Elsevier Science

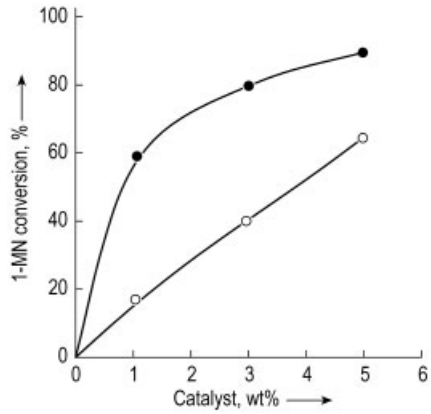

Figure 30. Effect of catalyst amount on the hydrogenation conversion of 1-methylnaphthalene(1-MN) with Ni-Mo/carbon black (KB JD, ●) or Ni – Mo/Al_2O_3 (○) [477]; Reaction conditions: 9.5 MPa, 1-MN : Decaline = 1 : 9, 380 °C for 40 min

tion of gum or sludge during storage or transportation to the oil refinery. In the current liquefaction processes, such as the IGOR+ process [24], the HTI process [482], and the Improved BCL process [46], [483], the crude liquefied oil in the gas phase of the high temperature – high pressure separator is directly introduced into a fixed-bed reactor and hydrotreated in the presence of an Ni – Mo/Al_2O_3 catalyst, so-called in-line hydrotreatment (see Section 3.8). Table 25 shows the improved oil yield and quality achieved by integrated refining in the IGOR process [408].

The catalyst deactivation in the hydrotreating of crude liquefied oil derived from the BCL process was examined with a commercial Ni – Mo/Al_2O_3 catalyst during the operation for 2000 h under model conditions with coexisting

Table 25. Improved oil yield and quality through integrated refining [408]

Refining mode	None	IGOR*	None	IGOR	IGOR+
Coal feed rate, $kg^{-1}\,L^{-1}\,h^{-1}$	0.5	0.5	0.6	0.6	0.6
C_5^+ distillate oil yield, wt % daf	50	56	45	54	53.5
Oil inspection					
Density, g/mL	0.930	0.844	0.935	0.853	0.830
50 % boiling point ASTM D86, °C	236	218	233	225	225
FBP, ASTM D86, °C	330	331	330	333	328
H, %	10.0	13.0	10.0	12.6	13.7
O, ppm	27.000	160	33.000	400	12
N, ppm	9.000	85	10.000	110	6

* IGOR = Integrated Gross Oil Refining. ** FBP = Final boiling point.

Table 26. Properties of liquefied and hydrotreated oils derived from brown coal (Banko coal) [483]

		Elemental analysis					
Oil		C, wt %	H, wt %	N, ppm	S, ppm	Odiff.*, wt %	H/C
Crude liquefied oil	LO	85.2	13.7	5000	1000	0.5	1.93
	MO	87.6	9.8	8000	1000	17	1.34
1st Hydrotreated oil	LO	86.2	13.6	308	5	0.2	1.89
(350 °C, Ni-Mo/Al_2O_3)	MO	88.0	11.1	1061	5	0.8	1.51
2nd Hydrotreated oil	LO	86.1	13.9	1		0.0	1.94
(380 °C, Ni-Mo/Al_2O_3)	MO	87.0	13.0	1		0.0	1.79

LO: Light oil (*bp* < 180 °C), MO: middle oil (*bp* 180 – 300 °C).
* Odiff. = 100 – (C + H + N + S)

gases, such as NH_3, H_2S, H_2O, CO, and H_2, at 15 MPa. It was reported that the service life of the catalyst was about a year, although the catalytic activity slightly decreased with process time [483]. No serious problems have been reported on the in-line hydrotreatment of crude liquefied oils with boiling points below 420 °C. The nitrogen and sulfur contents in liquefied oil could be reduced to a negligible level (< 10 ppm, see Table 26).

3.7.4. Molten Halide Catalysts

Acid-catalyzed coal depolymerization has been used as an analytical tool and a technique for liquefaction [183], [484], [485].

The process is different from those described in Sections 3.7.2 and 3.7.3. Bond cleavage is heterolytic rather than homolytic, probably involving carbocation formation. The reaction is faster and hydrogenation and hydrocracking reactions can be catalyzed at much lower temperature and hydrogen pressure. The most important catalysts are metal halides, especially zinc chloride, which is comparatively inexpensive and selective for liquids in the gasoline boiling range. Nitrogen and sulfur contents in the distillates are low.

Shell Oil Co. patented zinc halide catalysts for the hydroconversion of heavy petroleum fractions [486], [487]. Consolidation Coal Co. developed a coal liquefaction process that uses large quantities of molten zinc chloride (1 : 1 wt ratio of $ZnCl_2$: coal). Large proportions of catalyst are required to control product composition and to maintain activity. The catalyst is rapidly poisoned by NH_3 and H_2S liberated during hydrogenation and hydrocracking.

Despite its potential, the process has not been scaled up because of corrosion and zinc chloride recovery problems. Coal mineral matter constituents, particularly those containing potassium or sodium, can form stable eutectic salts with the melt. This can complicate recovery, especially for low-rank coals with high alkali content. Research was concentrated on the use of mild conditions and lower catalyst concentration [433–435].

3.8. Process Description

Direct liquefaction is the most efficient coal liquefaction route currently available. Its liquid products can be used for most stationary fuel applications. However, these products require further upgrading before being used directly as transportation fuels. This upgrading utilizes conventional petroleum industry techniques, allowing the products from a liquefaction plant to be blended into the feedstock streams of a petroleum refinery.

Direct liquefaction processes can conveniently be divided into two main groups, depending on whether the initial dissolution of the coal is separated from the conversion of the dissolved coal into distillable products. Single-stage direct liquefaction processes give distillates via the primary reaction system. Such processes may include integrated in-line hydrotreating, which is intended to upgrade the primary distillates without directly increasing the overall conversion. In the mid- to late 1960s, all available processes were single-stage [488]. Most development therefore continued to adopt a single-stage approach. Some developers added a second stage in the 1970s to increase the production of light oils.

The *single-stage processes* developed furthest are:

Kohleoel (Ruhrkohle, Germany)
NEDOL (NEDO and NCOL, Japan)
H-Coal (HRI, USA)
Exxon Donor Solvent (Exxon, USA)
SRC-I and II (Gulf Oil, USA)
Imhausen high-pressure (Germany)
Conoco zinc chloride (Conoco, USA)

Most of these have since been superseded and abandoned. Two exceptions are the Kohleoel and NEDOL processes, both of which are considered ready for commercialization by their developers. Several other processes were developed to a modest scale in USA. Other countries, notably Russia and Poland, also carried out R&D on single-stage liquefaction at a significant scale; the approaches adopted are believed to be similar in most respect to the Kohleoel process.

A *two-stage direct liquefaction process* is designed to give distillate products via two reaction systems. The primary function of the first stage is coal dissolution. This stage is operated either without a catalyst or with a disposable catalyst. After the solids, consisting of mineral matter and undissolved coal, are removed by deashing, the heavy coal liquids or coal liquid bottoms are hydrotreated in the secondary stage in the presence of a high-activity catalyst to produce additional distillate. Development of most two-stage direct coal liquefaction processes started after the first oil crises in the early 1970s. The developments were performed in many countries, but few processes were proceeded to pilot-plant scale.

The processes include:

Catalytic Two-stage Liquefaction (CTSL) (DOE and HRI, now HTI, USA)
HTI Coal (HTI, USA)
Liquid Solvent Extraction (LSE) (British Coal Corporation, UK)
Brown Coal Liquefaction (BCL) (NEDO and NBCL, Japan)
Consol Synthetic Fuel (CSF) (Consolidation Coal Co. , USA)
Lummus ITSL (Lummus Crest, USA)
Chevron Coal Liquefaction (CCLP) (Chevron, USA)
Kerr – McGee ITSL (Kerr – McGee, USA)
Solvolysis (Mitsubishi Heavy Industries, Japan)
Pyrosol (Saarbergwerke, Germany)
Amoco CC – ITSL (Amoco, USA)
Supercritical Gas Extraction (SGE) (British Coal Corporation, UK)

Only the CTSL, LSE, and BCL processes continued in development beyond the late 1980s. But somehow these surviving processes have evolved or revolved to single-stage processes because of the development of highly active, disposable catalysts and economical reasons.

A simplified block flow diagram of most direct coal liquefaction processes is shown in Figure 31. A slurry of pulverized dry coal in recycle solvent, containing sometimes coal liquid bottoms, is mixed and transferred through a fired preheater into the liquefaction reactors together with hydrogen and disposable catalyst. The reactors are maintained at a temperature of 400 – 470 °C and a pressure of 12 – 30 MPa. The temperature is normally controlled by addition of recycle gas. The effluent from the reactor is cooled, depressurized, and finally separated into gaseous, liquid, and bottom products. The major liquefaction products are liquids, which amount

to more than 50 wt % based on dry, ash-free coal (dafc).

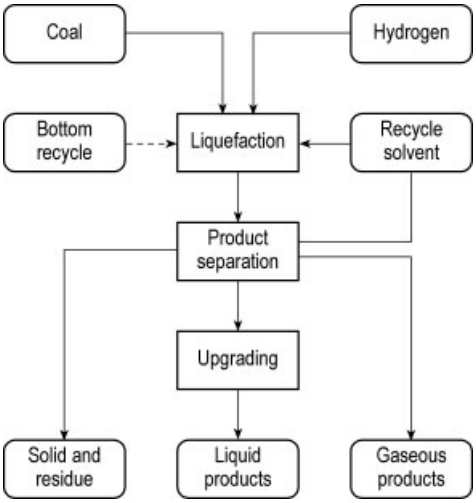

Figure 31. Block diagram of direct coal liquefaction

The bottom product consists of mineral matter, undissolved coal, and bottoms which cannot be distilled by vacuum distillation. The bottoms may be recycled as a portion of the recycle solvent or sent to a solids separation unit. Use of filters, centrifuges, hydrocyclones, or solvent precipitation is required if the mineral matter and undissolved coal must be separated from the liquids. Pilot plant experience has demonstrated that use of filters, centrifuges, and hydrocyclones makes scale-up difficult and decreases efficiency. Consequently solvent deashing processes such as the Lummus Antisolvent Process [489], the Kerr – McGee Super-critical Solvent Process [490], and the BCL Solvent Deashing Process [491], [492] have been recognized as deashing processes for a pioneer plant because of their ease of scale-up with relatively low capital costs.

The bottoms and solids are fed to partial oxidation and/or a boiler for hydrogen generation and heat recovery, respectively. Hydrogen is a major cost component in coal liquefaction. Usually, hydrogen is produced by the gasification of residual coal products and raw coal or by steam reforming of natural gas. Usually, light hydrocarbon products are used as fuel for fired furnaces in the liquefaction plant. In order to produce distillable liquids, hydrogen is required in the range of 6 – 8 wt % based on dry and ash-free coal.

3.9. Commercial Processes

Significant advances have been made in liquefaction processes since the 1940s. Milder operating conditions have been achieved, distillate yields have been increased, and efficiency of hydrogen utilization has been improved. Notable developments are the recycle of heavy liquids, the adoption of multistage processing, and the recognition that selective catalysts are required in all reaction stages.

3.9.1. Kohleoel (Ruhrkohle-Veba Oel) Process

In Germany, the first oil crisis initiated a revival of coal to oil technologies which had been in industrial-scale application in the early 1940s. National energy planner recognized the need for intensified conversion efforts, so that four companies got involved in three development projects all based on the old German IG Farben Technology:

Ruhrkohle AG (RAG, RUR)/Veba Oel AG
Saarbergwerke AG
Rheinische Braunkohlenwerke AG (Rheinbraun)

Proof of concept tests was carried out using process development units (PDU) at a scale of 0.1 – 0.2 t/d, and two pilot plants (PP) were built to demonstrate the technology, one designed for 200 t/d (RAG/Veba Oel) at Bottrop and another for 6 t/d (Saarberg). In 1986, the 6 t/d pilot plant and PDU of Rheinbraun were closed down. The 200 t/d PP was operated from 1981 to 1987, producing over 85 000 t of distillate products from 170 000 t of coal over approximately 22 000 operating hours [24]. In May 1987, RAG/Veba Oel successfully completed their 200 t/d PP test runs to modify the hydrogenation of vacuum residue from crude refining. Coal-based tests were continued in a PDU by RUR and Deutsche Montan-Technologie (DMT) to support the Bottrop PP during its test runs and also for process optimization until 1995. In 1997, the PDU was restarted to support the work for a feasibility study of

5000 t/d demonstration plant targeting the liquefaction of Xianfeng lignite in Yunnan province, China.

Therefore, the Kohleoel process with Integrated Gross Oil Refining (IGOR$^+$) (see Fig. 32) is a relatively recent development. This process aims to obtain "straight-run" fuels finished directly. Coal is mixed with a recycle solvent and relatively low-active iron catalyst (red mud) before the slurry is hydrogenated at elevated pressure and temperature, typically at 30 MPa and 470 °C. The specific coal feed rate to an upflow tubular reactor (c) is in the range 0.5 – 0.6 t m^{-3}h^{-1}. The reactor effluent is passed to a hot separator (d) and products remaining in the gas phase are hydrotreated in fixed-bed reactors (f) at 350 – 420 °C under the same pressure as in the main reactor. The hydrotreated products are depressurized and cooled. The liquid products from the first of these stages are recycled to the slurry-making step as part of the solvent. The liquid product from the second stage is routed to an atmospheric distillation column, yielding a light oil (C_5 – 200 °C) and a middle oil (200 – 350 °C) product. The bottoms from the hot separator pass to a vacuum column (e) to recover distillable products. These products are added to the hydrotreating reactor feed, and are subsequently largely recycled as solvent. The vacuum column bottoms, consisting of pitch, mineral matter, unreacted coal, and waste catalyst, would be used as a gasifier feedstock for hydrogen generation if the process were in a commercial stage. Process yields and quality [24], using German bituminous coal (Prosper coal), are summarized in Tables 27 and 28.

Table 27. Product yields of Kohleoel process

Product	Yield
Gaseous hydrocarbons (C_1 – C_4)	19.0
Light oil (C_5 – 200 °C)	25.3
Middle oil (200 – 325 °C)	32.6
Unreacted coal and pitch	22.1

Table 28. Kohleoel process – product quality obtained with Prosper coal

Product quality	Light oil	Middle oil
Hydrogen, %	13.6	11.9
Nitrogen, ppm	39	174
Oxygen, ppm	153	84
Sulfur, ppm	12	< 5
Density, kg/m^3	772	912

3.9.2. NEDOL Process

In 1983, New Energy and Industrial Technology Development Organization (NEDO) of Japan has developed a new process concept (NEDOL process) integrating the advantages of three different bituminous coal liquefaction processes (see Table 29) to produce liquid fuels for the transportation sector.

Table 29. Bituminous coal liquefaction processes developed between 1978 and 1983 in Japan

Process	Developer	Unit scale
Direct Coal Liquefaction	Nippon Kokan	2.4 t/d
Hydrogen Donor Process	Sumitomo Metals Industries	1 t/d
Solvolysis Process	Mitsubishi Heavy Industries	0.1 t/d

The NEDOL process was developed through the 1 t/d PSU operation at Kimitsu in 1989 – 1998, and a 150 t/d pilot-plant operation at Kashima in 1996 – 1998. In the pilot plant tests, seven runs were carried out using three types of coal of different rank (Tanito Harum and Adaro coals from Indonesia and Ikeshima coal from Japan) which were liquefied without serious problems. A 80 days continuous coal liquefaction operation was achieved successfully during these test runs [44],[47],[493], obtaining 58 wt % daf oil yield, with high coal concentration (50 wt %) in the slurry.

CCRI of China, NEDO, and Centre for Coal Utilization of Japan (CCUJ) signed an agreement to carry out a feasibility study for a 5000 t/d demonstration plant in Heilongjiang province using Yilan coal. Process evaluation, environmental impact assessment and economic analysis was carried out from 1997 to 1998.

The NEDOL process is a single-stage dispersed catalytic liquefaction process as shown in Figure 33. The distillate recycle solvent is hydrogenated catalytically in external fixed bed reactors to increase the concentration of H-donor compounds. Bituminous and subbituminous coals are slurried with the H-donor solvent and iron-based catalyst (a) to be hydrogenated at 450 – 465 °C and 17 – 19 MPa. The nominal slurry residence time in simple upflow tubular reactors (c) is 60 min while the actual slurry phase residence times varies between 90

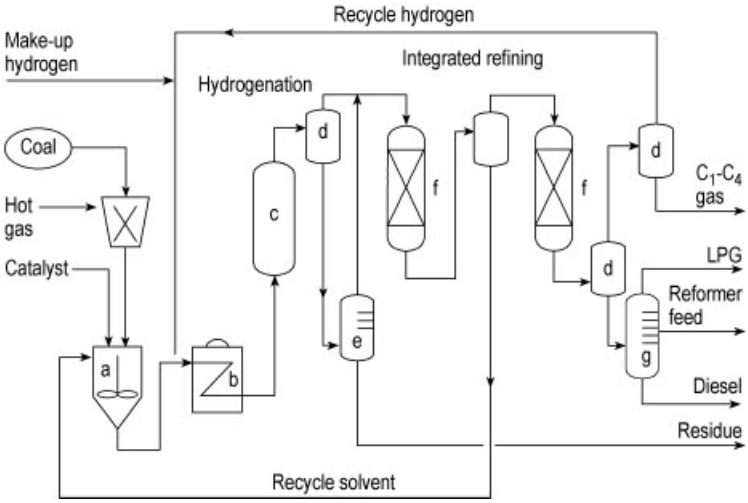

Figure 32. Schematic diagram of the IGOR$^+$ coal liquefaction process
a) Coal preparation; b) Slurry preheater; c) Hydrogenation reactor; d) Hot separator; e) Vacuum distillation column; f) Fixed-bed reactor; g) Atmospheric distillation column

and 180 min depending on the gas : slurry feed ratio in the reactors. The products are cooled, depressurized and separated by atmospheric (d) and vacuum distillation columns (e). Part of the middle distillate and all of the heavy distillate are recycled as solvent via the hydrogenation step. The vacuum column bottoms, which contain unreacted coal, mineral matter, and catalyst are withdrawn from the process for production of hydrogen. In this process, no bottoms are recycled (one-through operation) to simplify the process. Since the organic matter in the bottom withdrawn with the solids represents a substantial loss of potential products, the process is limited to coal of relatively low ash content.

The middle and heavy distillates from the vacuum column are fed to solvent hydrogenation reactors (g) which are of the trickle-flow fixed-bed type, operating at 320 – 400 °C and 10 – 15 MPa. The catalysts used are variants of those used in the conventional petroleum industry for hydrodesulfurization of crude oils. The reactor products are depressurized and fed to a flash stripper (f) where the hydrotreated naphtha is removed. The main stream from the flash stripper is recycled to the coal slurry making step as the solvent. Typical operation results of a 150 t/d pilot plant using Tanito Harum Indonesian subbituminous coal are summarized in Tables 30, 31, 32 [23], [44].

Table 30. NEDOL process conditions with Tanito Harum coal

Reaction temperature	460 °C
Reaction pressure	17 MPa
Coal in slurry	50 wt %
Gas : slurry ratio	850 m^3/t (STP)
Catalyst loading (pyrite)	2.9 wt % (daf)

Table 31. NEDOL process yields (wt % daf) with Tanito Harum coal

Hydrocarbon gases ($C_1 - C_4$)	16.9
Light oil (C_5 – 220 °C)	28.8
Middle oil (220 – 350 °C)	19.3
Heavy oil (350 – 538 °C)	3.3
Unreacted coal and pitch	18.8

Table 32. NEDOL process quality with Tanito Harum coal

Product quality	Light oil	Middle oil
Carbon, wt %	83.3	87.3
Hydrogen, wt %	11.7	9.8
Nitrogen, wt %	0.5	0.8
Oxygen, wt %	4.3	2.0
Sulfur, wt %	0.1	0.0
Density, kg/m^3	857	978

3.9.3. Solvent-Refined Coal (SRC-I and SRC-II) Processes

Interest in producing a substitute fuel for electric utility boilers led the Southern Company and Electric Power Research Institute to explore the Solvent Refined Coal (SRC-I) process. This is a

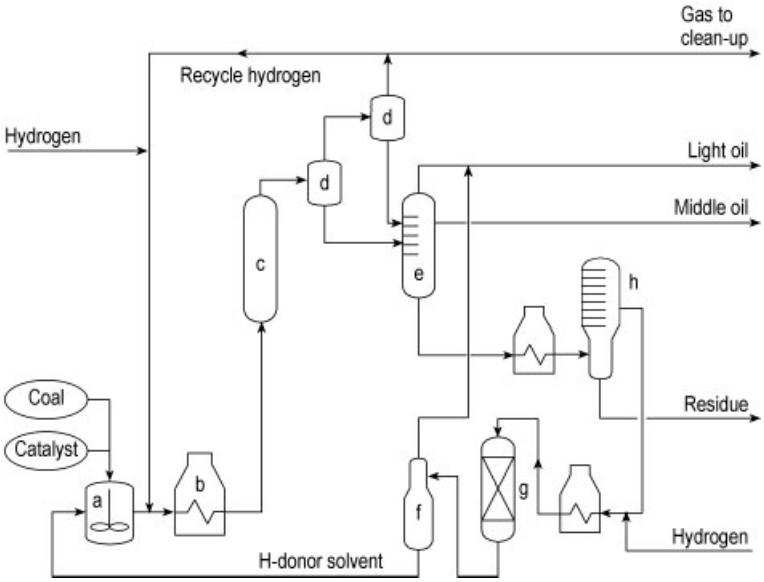

Figure 33. Schematic diagram of the NEDOL process
a) Slurry mix tank; b) Slurry heater; c) Liquefaction reactor; d) Separator; e) Atmospheric distillation column; f) Stripper; g) Catalytic solvent hydrogenation; h) Vacuum distillation column

one-stage thermal reaction process in which coal is solubilized (at ca. 450 °C and 10 – 14 MPa for 30 – 60 min) in a recycle solvent, and deashed to remove mineral matter, producing an ash-free reduced-sulfur solid fuel. The 6 t/d pilot plant operation of SRC-I, which started in 1974 at Wilsonville, Alabama, USA, was initially unsuccessful, because of problems of solvent imbalance and degeneration of solvent quality with successive recycling [243], [244]. An ebullated-bed catalytic hydrotreater at the SRC-I plant provided additional flexibility in solvent balance and in maintaining the quality of the recycle solvent.

At Fort Lewis, Washington, USA, the process was modified to SRC-II by Gulf Oil Company and others in a 50 t/d pilot plant in 1977 to produce liquid fuels (see Fig. 34). In this process, coal is hydrogenated in a single-stage reactor (c) without catalyst, but the distillation bottoms are recycled to take advantage of the catalytic activity of the coal mineral matter. Bottoms recycle also provides another important function: it allows residue to be selectively cracked to distillate without excessive production of gaseous hydrocarbons. For this reason, this recycle is an integral part of essentially all processes designed to produce a distillate product. In the normal SRC-II mode, most low-rank coals gave low yields and caused operability problems, because of the high product viscosity. The use of dispersed catalysts could overcome these difficulties [494]. The further development of SRC-I and SRC-II has moved away from the original aim of refining coal in favor of distillate production. The Wilsonville plant continued to be funded by the DOE until 1992 as a pilot-scale test facility for the whole US direct liquefaction development program. The SRC processes have now been abandoned in their original form, but elements have been incorporated in the more recent US processes.

3.9.4. Exxon Donor Solvent Process

Exxon started the development of the Exxon Donor Solvent (EDS) process in the 1970s, the 250 t/d pilot plant started operation at Baytown, Texas, in 1980. The pilot plant was operated until 1982, further research continued until at least 1985. At this point, Exxon considered that the process was ready for commercialization.

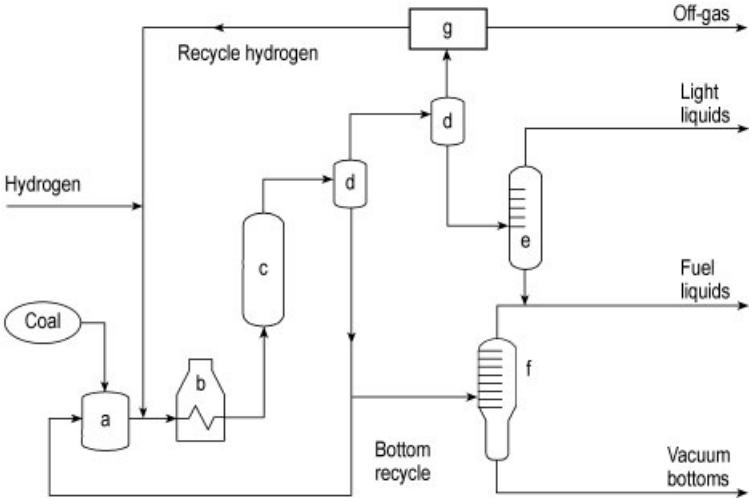

Figure 34. Schematic diagram of the SRC-II process
a) Slurry mix tank; b) Slurry heater; c) Reactor; d) Separator; e) Atmospheric distillation; f) Vacuum distillation; g) Gas purification

The EDS process is a single-stage noncatalytic liquefaction process (see Fig. 35). The distillate recycle solvent is hydrogenated catalytically in an external fixed bed reactor to increase the concentration of H-donor compounds. In its original version, coal is slurried with the H-donor solvent (200 – 425 °C) and hydrogenated at 425 – 450 °C and 10.3 – 17.2 MPa in a plug-flow tubular reactor (c). The reactor effluent passes to gas – liquid separators (d), from which the liquid product is fed to a vacuum distillation column (e). All of the naphtha and part of the middle distillates are recovered as products. The rest of the middle distillate is recombined with the heavy distillate to form the recycle solvent. The vacuum distillation column bottoms, containing the solid, are fed to a flexicoker unit for producing additional liquids and fuel gases. The coke is gasified to produce hydrogen, supplemented, if necessary, by gasification of raw coal or steam reforming of the light hydrocarbon gases.

Hydrogenation of the recycle solvent is carried out in a fixed-bed catalytic reactor (f), using conventional refinery catalysts for hydrodesulfurization. The reactor is operated in the region of 370 °C/11 MPa, although conditions are varied to control the degree of hydrogenation of the solvent and thus maintain its quality.

In a later development, a bottoms recycle operation was added to the EDS feed to provide additional residence time for conversion, and perhaps to utilize the catalytic activity of the mineral matter. This feature has been taken over from processes such as the CONSOL CSF process and the British Coal LSE process. These processes have in common that the primary coal liquefaction takes place in the absence of any added catalyst, ostensibly to avoid the rapid deactivation of the catalyst that is presumed to occur in the presence of the coal mineral matter. Distillate yields in the 250 t/d PP were lower than in the smaller test facilities because the high gas holdup in the reactor – probably a result of the high slenderness ratio of the reactor – led to a reduced residence time [19], [495], [496].

3.9.5. H-Coal Process

The H-Coal process is a catalytic single-stage liquefaction process (see Fig. 36) developed by Hydrocarbon Research Incorporated (HRI, now Hydrocarbon Technologies Incorporated, HTI) as an outgrowth of their H-Oil process for upgrading heavy oils. The process was successfully scaled up from 3 t/d to a 200 t/d pilot plant (600 t/d in the fuel oil mode) at Catletts-

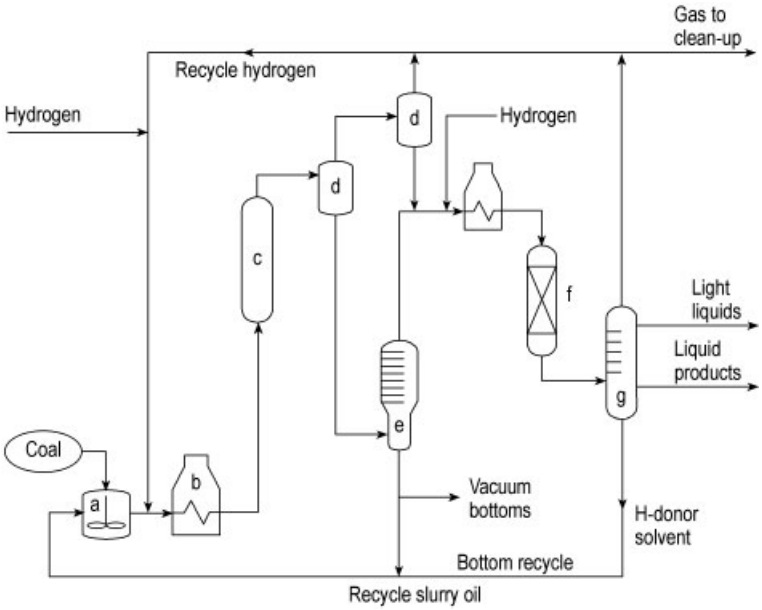

Figure 35. Schematic diagram of the EDS process
a) Slurry mix tank; b) Slurry heater; c) Plug-flow reactor; d) Separator; e) Vacuum distillation; f) Catalytic solvent hydrogenation; g) Atmospheric distillation

burg, Kentucky, from 1980 to 1983. At high coal throughput (fuel mode), a low-sulfur fuel oil was produced. Separating solids from heavy liquid products in the fuel oil mode was difficult. At reduced throughput (syncrude mode), the product was an all-distillate "syncrude". In this mode, the solids were separated by vacuum distillation. The distillation bottoms could be used for hydrogen production via gasification. Over the two years period, the plant proved capable of prolonged continuous operation, however, the aim of its commercialization has been abandoned. The H-Coal process remained the basis of most subsequent process development sponsored by the US DOE and was effectively incorporated into the Catalytic Two-Stage Liquefaction (CTSL) process [19], [28], [497], [498].

A key feature is the novel reactor design (see Fig. 37) with an ebullated bed of conventional supported hydrocracking catalysts. Coal, recycle solvents, that contain unreacted coal solids, and hydrogen gas are heated and fed to the reactor to contact the catalyst at temperatures of 425 – 455 °C and a pressure of 20 MPa. The catalyst particles, which are in form of 1.5 mm extrudates, are fluidized by the upward flow of internally recycled reactants. In the expanded bed, reactants flow over the weir to the inlet of the ebullating pump. The catalyst particles are too dense to enter the pump or to leave with the effluent, although fines undoubtedly permeate the whole system. Spent catalyst is periodically withdrawn and replaced with fresh catalyst to maintain a constant level of activity. The ebullated bed design promotes gas – liquid – solid contact, facilitates isothermal operation, and allows solids, which could plug a fixed-bed system, to pass through the reactor. The reactor products pass to a flash separator. Liquids in the gas – vapor stream are condensed and routed to an atmospheric distillation column, separating naphtha and distillate products. The flash bottoms are fed to a bank of hydrocyclones. The overhead stream, which contains 1 – 2 % solid, is recycled as a part of solvent. The underflow is routed to a vacuum distillation column. Solids are removed with the vacuum bottoms, while the vacuum distillate forms a part of the products.

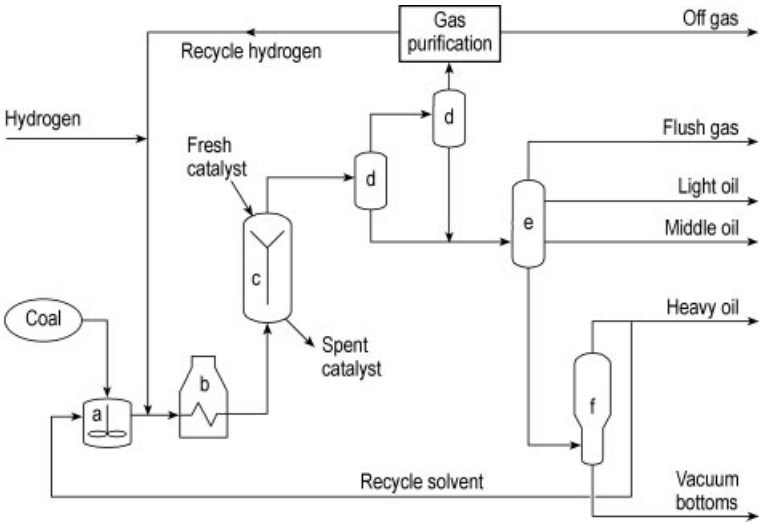

Figure 36. Schematic diagram of the H-Coal process
a) Slurry mix tank; b) Slurry heater; c) Ebullated-bed reactor; d) Separator; e) Atmospheric distillation; f) Vacuum distillation

3.9.6. Catalytic Two-Stage Process

The CTSL process is a development of the H-Coal single-stage process. The process was further developed over a period of almost 15 years in the pilot plant at Wilsonville. This plant closed in 1992, but development work has been continued by bench-scale and PDU testing at HTI's Lawrenceville facilities under support of the US DOE. The CTSL process has now evolved into a generic composite of much of the liquefaction development work funded by the DOE in the 1980s and 1990s. A closed-coupled configuration has been adopted in which both stages use an active supported catalyst. In more recent work, HTI has included an in-line hydrotreater after the second-stage reactor to improve the product qualities.

One problem with the single catalytic stage is that hydrogenation reactions are best carried out at relatively low temperatures where the equilibrium is more favorable, but conversion reactions are faster at higher temperature. To overcome this, research was conducted on the use of multiple reaction stages. In the Lummus/Cities Service bench unit, the coal was converted to a soluble form in the first stage at relatively high temperature but short contact time. The first stage product was deashed, and converted to distillate liquids at around 400 °C in an ebullated bed LC-Finer unit. A similar configuration was installed at the Wilsonville plant, which employed a Kerr – McGee Critical Solvent Deasher between a thermal first stage and an ebullated-bed second stage supplied by HRI. It was recognized that the location of the deashing unit between the two stages had little effect on catalyst life. The most important factor in initial catalyst deactivation was deposition of carbonaceous material, not poisoning by metals in the mineral matter [441–443]. This led to "close coupling" of the two reactor stages, with deashing applied to the second-stage product. By concentrating solids in the recycle stream, the amount of material processed by the deasher was significantly reduced, and the unconverted coal was given additional residence time.

The Wilsonville unit was subsequently used to explore a variety of staged reactor configurations, including two ebullated-bed catalyst stages, the use of a dispersed iron oxide as a first-stage catalyst, and alternative temperature staging. This resulted in substantial reductions in production of gaseous hydrocarbons and residues, with a corresponding increase in distillate liquids [340], [341]. Subsequent works supported by the DOE investigated the use of dispersed catalysts based on iron, molybdenum, and other metals [335–337].

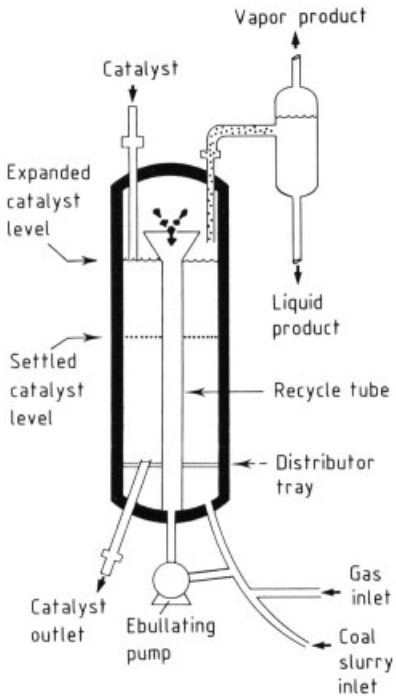

Figure 37. H-coal ebullated bed reactor [497]

In a collaborative project, CONSOL, the University of Kentucky, LDP Associates, HTI, and Sandia National laboratories have been exploring several advanced concepts for improving direct liquefaction process economics. A schematic diagram of the process [36] is shown in Figure 38.

3.9.7. HTI Coal Process

HTI chose a proprietary dispersed superfine, iron catalyst instead of the supported catalysts in the closed-coupled two-stage liquefaction reactor configuration [22], [482]. The latest process is called HTI Coal Process. In 1997, HTI signed a two-year agreement with Shenhau Clean Coal Technology Development (SCCT) and CCRI to execute a feasibility study for a direct coal liquefaction plant using Shenhau coal. The project is to build a mine-mouth liquefaction plant of 12 000 t/d capacity using coals produced in the Shenhua coalfield in Northern Shaanxi Province and Inner Mongolia near Baotou city [22].

The HTI Coal Process (see Fig. 39) consists primarily of two backmixing reactor stages utilizing a iron catalyst (GelCat) plus a fixed-bed in-line hydrotreater. The process operates at a pressure of 17 MPa and reactor temperatures in the range of 400 – 460 °C. A slurry consisting of pulverized coal, recycle solvent, and recycled residue is mixed and pressurized through a fired preheater (b) and passed along with hydrogen and catalyst into the first-stage reactor (d). The effluent from the first reactor is separated in an interstage separator (e) to remove gases and hydrocarbon vapors, the slurry is sent to the higher temperature second-stage reactor (f). Effluent from the second-stage reactor is flashed and the overhead gas – vapor stream is fed to the in-line hydrotreater (h) together with the overheads from the interstage separator for enhanced upgrading to clean fuels. A part of the atmospheric bottoms stream, containing solids, is used as recycle with a portion going to a vacuum distillation column (j). A typical operation result of the continuous flow unit (30 – 50 kg/d) using Shenhau coal [22] is summarized in Tables 33 and 34.

Table 33. HTI Coal process yield (wt % daf) with Shenhau coal

Coal conversion	90 – 93
Distillate (C_5 – 524 °C)	52 – 68
Residue (524 °C)	7 – 22

Table 34. HTI Coal product quality with Shenhau coal

Product quality	Whole distillate
Carbon, wt %	87.6
Hydrogen, wt %	12.7
Nitrogen, wt %	0.14
Sulfur, wt %	0.10
API gravity	857

3.9.8. BCL Process

The Brown Coal Liquefaction (BCL) process is a two-stage liquefaction process developed by Nippon Brown Coal Liquefaction (NBCL) through the 50 t/d pilot plant operation at Morwell, Victoria, Australia. The pilot plant was operated over the period 1985 – 1990 under full support of NEDO, processing in total 70 000 t of coal. The process is designed specifically to handle very low-rank coals such as those found

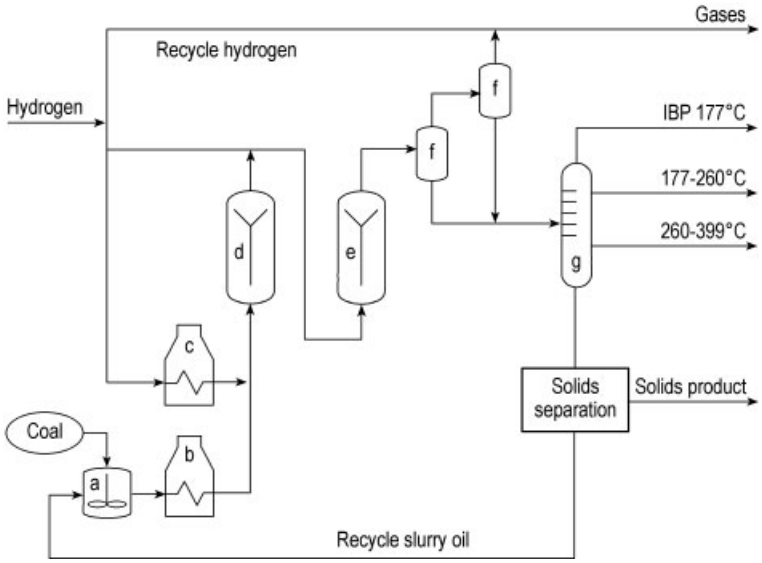

Figure 38. Schematic diagram of the CTSL process
IBP = Initial boiling point

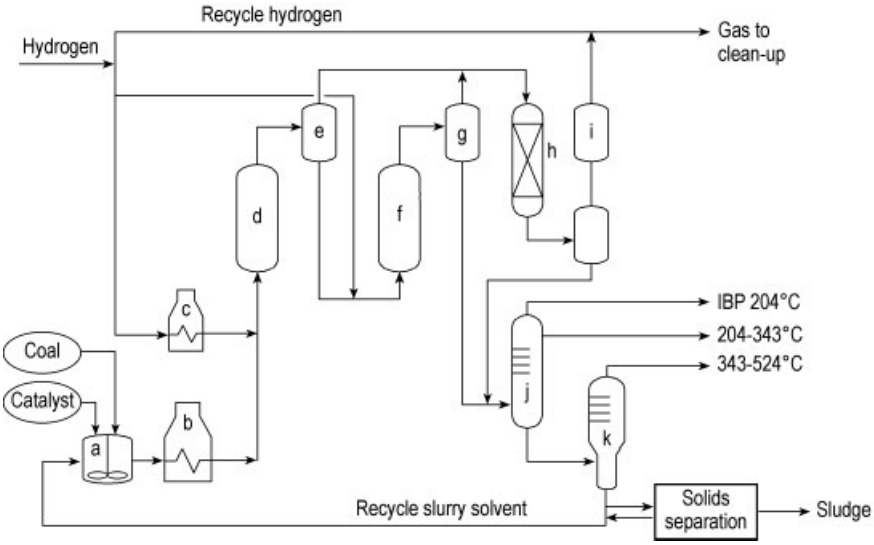

Figure 39. Schematic diagram of the HTI Coal process
a) Slurry mix tank; b) Slurry heater; c) Hydrogen treater; d) First-stage reactor; e) Separator; f) Second-stage reactor; g) Flash separator; h) Fixed-bed hydrotreater; i) Gas purification; j) Atmospheric distillation; k) Vacuum distillation

in the Latrobe Valley of Victoria, which contain more than 60 % of moisture. This process consists of slurry dewatering, primary hydrogenation, solvent deashing, and secondary hydrogenation process units. Raw brown coal is slurried by adding a fine iron catalyst and recycle solvent which includes some bottoms from primary and secondary hydrogenation. The moisture content in the raw brown coal is removed by heating the slurry. The slurry is transferred through a fired preheater along with hydrogen into liquefaction reactors which are maintained at 430 – 460 °C and 15 – 20 MPa. Effluent from the reactor is cooled, depressurized and separated. The bottom product ($bp \geq 420$ °C) is deashed and fed into the secondary hydrogen-

ation unit to be hydrogenated further to produce light oil and hydrogenated heavy solvent. In 1993, NBCL completed the pilot plant program, and attained all the developmental objectives: 52 wt % oils yield, 1700 h continuous operation, 700 ppm ash concentration in the bottom at the outlet of the deashing unit, development of a secondary hydrogenation catalyst, and establishment of a more economical slurry dewatering method [46].

Following extensive pilot plant operation, R&D using a 0.1 t/d bench scale unit and related equipment was continued from 1994 until 1997 at Takasago, Japan to improve the reliability, economics, and environmental compatibility of the process. Based on the R&D results, an Improved BCL Process was proposed [46]. The Improved BCL Process is a modification of the BCL Process and a single-stage liquefaction method as shown in Figure 40. This process consists of the four operations slurry dewatering, liquefaction, in-line hydrotreating, and solvent deashing. The Improved BCL Process is a simplified liquefaction process with the following advantages: high efficiency and low investment, use of a highly active and inexpensive iron catalyst such as limonite ore pulverized with process solvent, improved quality of recycle solvent and product oils by the in-line hydrotreater. The deashed bottoms recycling operation achieves a high oil yield with mitigation of erosion problems.

Compared with the results of the pilot plant, in the bench-scale unit oil yield could be increased, product quality improved, and scale formation in the reactors suppressed using 1 wt % catalyst or less. It was estimated that the improved BCL Process could decrease the crude oil equivalent price by 24 % compared with the BCL Process at the Australia pilot plant. Typical test run results using Victorian brown coal and Indonesian low-rank coal are shown in Tables 35, 36, 37.

Table 35. BCL process conditions

	BCL	Improved BCL
Coal used	Yalloum	Banko
Tested by	50 t/d	0.1 t/d
Reaction pressure, MPa	15	15
Reaction temperature, °C	450	450
Solvent : coal, wt/wt	2.5	2.0
Catalyst	hematite	limonite
Catalyst concentration, wt % daf	3.0 – 5.0	1.0 – 1.5

Table 36. BCL process yields

Fraction	Yields, wt % daf	
	BCL	Improved BCL
$C_1 - C_4$	11.1	13.6
C_5 – 220 °C	16.4	36.7
220 – 300 °C	16.8	15.1
300 – 420 °C	14.9	11.6
Bottom ($\geq$ 420 °C)	18.4	8.1

Table 37. BCL process product quality (whole distillate)

	BCL	Improved BCL
Coal used	Yalloum	Banko
Tested by	50 t/d PP	0.1 t/d BSU*
Carbon, wt %	84.5	87.5
Hydrogen, wt %	10.6	12.1
Nitrogen, wt %	0.47	0.01
Oxygen, wt %	4.6	0.4
Sulfur, wt %	0.1	0.00
H : C ratio	1.50	1.67
Density, kg/m^3	925	910

* BSU = Bench Scale Unit

In 1999, NEDO and BPPT (Agency for the Assessment and Application of Technology) of Indonesia signed an agreement to carry out a feasibility study for a 30 000 t/d direct coal liquefaction plant in Tanjung Enim, South Sumatra using Banko coal (lignite). Conceptual design, environmental impact assessment and economic analysis will be carried out until 2002 [26].

3.9.9. Liquefaction Solvent Extraction Process

The Liquefaction Solvent Extraction (LSE) process was developed by British Coal Corporation between 1973 and 1995 using a 2.5 t/d pilot plant at Point of Ayr, North Wales. After the operation of the pilot plant, an outline design for a 65 t/d demonstration plant has been developed, however, currently there are no plans to construct the plant. Figure 41 shows a schematic flow diagram of the process. Coal is slurried with a process-derived recycle solvent, preheated, and fed to an noncatalytic digestion step, which consists of two or more continuous-stirred tank reactors in series. These reactors are operated at 410 – 440 °C and 1 – 2 MPa, the pressure required solely to reduce solvent vaporization. No hydrogen is used in this step, but the solvent acts as a hydrogen donor, transferring up to 2 wt % of

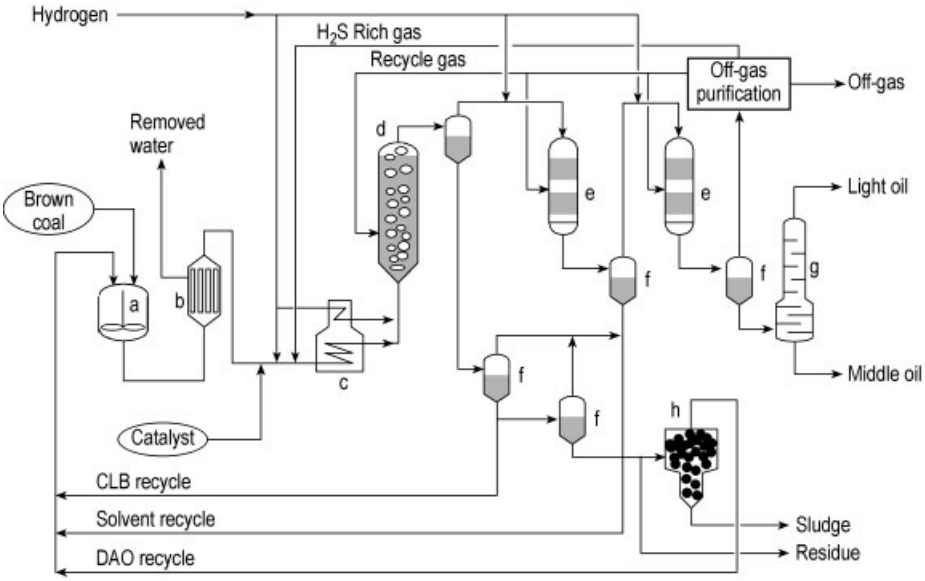

Figure 40. Schematic diagram of the Improved BCL process
a) Slurry mix tank; b) Evaporator; c) Slurry preheater; d) Reactor; e) Fixed-bed hydrotreater; f) Separator; g) Distillation; h) Settler
CLB = Coal liquid bottom; DAO = Deasphalted oil

hydrogen to the coal [499]. The digester product is partially cooled and filtered in a vertical-leaf pressure filter to remove solids. The filter cake is washed with a light recycle oil fraction to recover product and dried under vacuum. In commercial operation the filter cake would be gasified to produce hydrogen. The filtered coal extract passes to a distillation column to recover the light oil wash solvent and is then preheated, mixed with hydrogen, and routed to two ebullated-bed reactors in series in which the coal extract is hydrocracked at 400 – 440 °C and up to 20 MPa. The reactor effluent is cooled, depressurized and fed to a atmospheric distillation column to recover distillate products. The cut-point of this column is adjusted to maintain solvent balance, resulting in a product that typically boils below 300 °C. The column bottoms are routed partially to a vacuum distillation column, used to control the level of pitch in the recycle solvent. The overheads from the vacuum distillation column are recombined with the main atmospheric column bottoms stream and recycled as solvent to the slurrying step.

The pilot plant operating program concluded with a steady-state demonstration run. The results from this run indicate that it would be possible to operate the process in an all-distillate product mode, removing the requirement for a vacuum column. Solvent quality was maintained without the need for a separate, thermal cracking step. Primary product quality data are shown in Tables 38, 39, 40. The total distillate product yield is in the range of 60 – 65 wt % on dafc [50].

Table 38. LSE process conditions

Solvent : coal ratio, wt/wt	2.2
Digestion pressure, MPa	1.5
Digestion temperature, °C	431
Nominal residence time, min	50
Hydrocracking pressure, MPa	20
Hydrocracking temperature, °C	434
Space velocity, kg feed kg cat^{-1} h^{-1}	0.75

Table 39. LSE process yield (wt % daf) using Point of Ayr coal

$C_1 - C_4$ hydrocarbon gases	15.4
C_5 – 300 °C distillate product	49.9
300 – 450 °C distillate product	12.4
Pitch (450 °C)	0.8
Filter cake organics	23.9

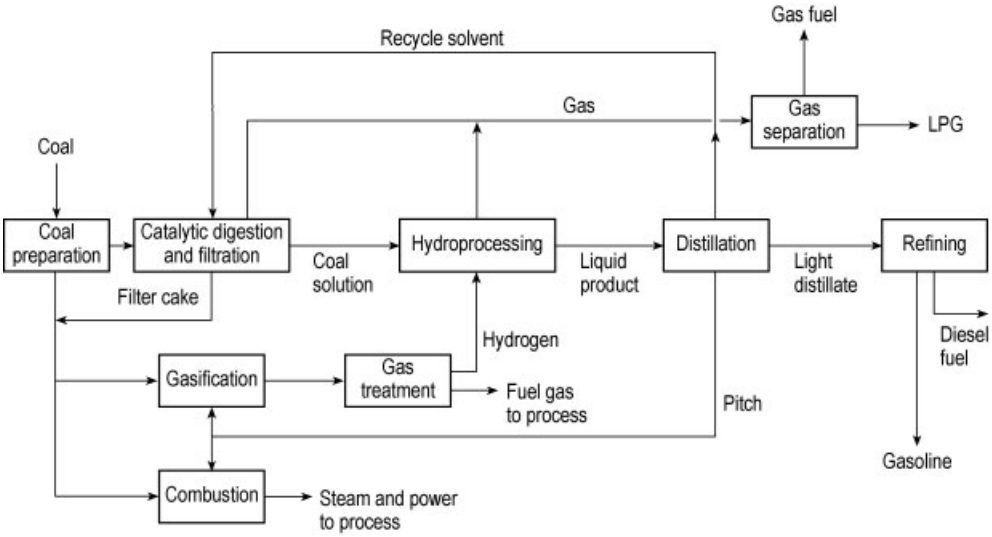

Figure 41. Block flow sheet of the LSE process

Table 40. LSE product quality using Point of Ayr coal

Hydrogen, wt %	12.14
Nitrogen, wt %	0.14
Sulfur, wt %	0.04

3.10. Environmental Considerations

Direct and indirect coal liquefaction create similar environmental concerns, although the former produces more aromatic liquids [500]. The toxicity and carcinogenicity of coal products are largely due to polycyclic aromatic hydrocarbons (PAH) and their nitrogen-containing analogues (N-PAC), which are found in fractions boiling at ca. 370 °C or higher [501].

Carcinogens and mutagens are removed by denitrogenation, hydrogenation, and hydrocracking. In processes where the higher boiling products are catalytically refined or recycled to extinction, mutagenic activity may be substantially eliminated [501], [502]. Refined coal and shale products are generally equivalent, in terms of potential health hazard, to their petroleum derived equivalents [503].

Preferably, high-boiling, coal-derived liquids should not be stored, but fed directly to a gasifier, if this is their intended disposition.

Health hazards from intermittent contact with refined coal liquids appear to be negligible. The main danger does not confront the general public, but the plant operators. Strong precautionary practices should be instituted in the direct area of process operation.

The liquefaction technologies have advanced significantly since the mid-1980s, which has produced great benefit not only for improving economics but also for environmental adaptation. The liquefaction plants to produce coal-derived fuels will be designed to meet the highest standards for environmental compliance. The transportation fuels produced by advanced coal liquefaction technologies will be environmentally superior to their petroleum-derived counterparts and will be capable of meeting all requirements of clean air standards and regulations in major countries.

3.11. Economic Aspects

In 1985, oil price fell suddenly and had remained low until spring 1999. Therefore, interest in coal liquefaction for the production of transportation fuels has declined except in a few countries. China and Indonesia are now executing collaborative feasibility studies on direct coal liquefaction plants. It remains to be seen whether a future increase in oil prices will result in renewed interest and possibly in full-scale commercialization of these plants. Because of resource limitation of

conventional petroleum caused by rapid rate of economic growth particularly in the Pacific rim, there will be a worldwide shortfall in petroleum liquids in the 2025 – 2030 timeframe [504]. This shortfall will impact all countries by constraining their ability to maintain their share of imports, and as a result, the world is likely to experience a liquid fuels shortfall anytime between 2005 and 2030. The liquefaction technologies must be in the state of readiness for commercial deployment about 15 years earlier, because lead times for the introduction of new energy technologies are on the order of 10 to 15 years, even after technologies are technically ready for commercial deployment.

Although direct liquefaction technology has undergone very significant improvements in the 1990s and achieved a high level of technical readiness, liquefaction products are still not cost competitive against crude oils, with prices being in the range of $ 10 – 20/bbl over the last decade. Continuing R&D is, therefore, needed to reduce costs to meet the target of 2015 for the start of commercial deployment. The US DOE contracted with Bechtel to develop a conceptual commercial design to produce hydrotreated liquid products from either bituminous or subbituminous coal [10]. The Bechtel design yielded a cost of about $34/bbl of crude oil equivalent (COE). This cost is significantly lower than earlier estimates of $ 40 – 50/bbl because of process improvements from R&D undertaken over the last decade. Because direct liquefaction technology is still evolving and additional process improvements are expected, costs will decrease further as improvements are incorporated. Areas of most importance in reducing costs include decreasing capital investment, improving product yields, and lowering catalyst cost.

The key to commercialization will be integration with the existing petroleum refining/distribution infrastructure. In achieving commercialization, two intermediate technologies are important. The first is co-processing of petroleum-derived wastes (plastics, tires, waste oil) with coal. Developing this technology is being driven by dwindling landfill availability and increase in tipping fees. The second technology is coprocessing of coal with heavy petroleum residues or oils. This option is likely to account for the first production of coal-base transportation fuels in existing petroleum refineries.

4. References

1. H. Steinbrecher, *Braunkohle* **12** (1926) 40.
2. I. G. C. Dryden in H. H. Lowry (ed): *Chemistry of Coal Utilization,* **suppl. vol.,** J. Wiley & Sons, New York 1963, p. 237.
3. D. W. van Krevelen: *Coal,* Elsevier, Amsterdam 1981.
4. R. F. Probstein, R. E. Hicks: *Synthetic Fuels,* McGraw-Hill, New York 1982.
5. L. Grainger, J. Gibson: *Coal Utilization: Technology, Economics and Policy,* Graham & Trotman, London 1981.
6. H. W. Coover, Jr., R. C. Hart, *Chem. Eng. Prog.* **78** (1982, Apr.) 72 – 75.
7. J. L. Ehrlen, B. Juran, *Chemtech.* **13** (1983, Mar.) 188 – 192.
8. J. Haggin, *Chem. Eng. News,* 1981, (Feb. 23) 39.
9. G. V. McGurl, S. R. Lee, R. D. Srivastava, "Coal Liquefaction: From the Moral Equivalent of War to an Option for the Nation's Future", *Proceedings of the 10th International Pittsburgh Coal Conference,* 1993, p. 270.
10. D. Gray, G. Tomlinson, E. Elsawy, "The Economics of Liquid Transportation Fuels From Coal: Past, Present and Future", *Proceedings of the 10th International Pittsburgh Coal Conference,* 1993, p. 277.
11. G. Collin, *Erdoel Kohle* **35** (1982) no. 6, 294.
12. A. S. Nislick, *Chem. Eng. Prog.* **78** (1982, May) 41 – 46.
13. W. H. Calkins, *Catal. Rev. Sci. Engl.* **26** (1984) nos. 3, 4, 347.
14. D. Gray, M. Lytton, M. B. Neuworth, G. C. Tomlinson: *Adv. Coal Util. Technol. IV,* Institute of Gas Technology (IGT), Apr. 20 – 24, 1981, p. 93.
15. W. R. K. Wu, H. H. Storch: "Hydrogenation of Coal Tar," U.S. Bureau of Mines Bulletin no. 633, 1968.
16. "Liquid Fuels From Coal," National Coal Board Report, U.K., Aug., 1978.
17. "Report on the Petroleum and Synthetic Oil Industry of Germany," Ministry of Fuel & Power, His Majesty's Stationary Office, London 1947.
18. B. H. Weil, J. C. Lane (eds.): *The Technology of the Fischer – Tropsch Process,* Constable and Company Ltd., London 1949.
19. B. S. Lee: "Synthetic Fuels from Coal," *AIChE Monograph Series* **78** (1982) no. 14, 1 – 42.

20. J. A. Tijm, J. M. Marriott, H. Hasenack, M. M. G. Senden, Th. van Herwijnen: "The Markets for Shell Middle Distillate Synthesis Products," *Alternate Energy '95,* Vancouver, May 2 – 4, 1995.
21. W. Lee, *Synergy* **2** (1982) no. 1, 5.
22. A. G. Comolli, L. K. Lee, G. A. Popper, P. Zhou, *Fuel Process. Technol.* **59** (1999) 207.
23. H. Yoshida: Results of NEDOL Process: Coal Liquefaction Technology Development, *Proceedings of Advanced Clean Coal Technology International Symposium '98,* Tokyo, October 1998, p. 198.
24. K. D. Dohms, "Revival and Status of Coal Liquefaction Technology in Germany," *ibid* 1998, p. 235.
25. Jin Jialu: "Status of Energy resources and Commercialized Prospect of Direct Coal Liquefaction Technology in China," *ibid* 1998, p. 255.
26. H. Soedioto, M. Tamura: "Applicability Study on Coal Liquefaction Technology in Indonesia," *ibid* 1998, p. 265.
27. E. E. Donath in H. H. Lowry (ed): *Chemistry of Coal Utilization,* **suppl. vol.,** J. Wiley & Sons, New York 1963, pp. 1041 – 1080.
28. S. B. Alpert, R. H. Wolk in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* **2nd suppl. vol.,** J. Wiley & Sons, New York 1981, pp. 1919 – 1990.
29. P. Nowacki (ed.): *Coal Liquefaction Processes,* Noyes Data Corp., N.J., 1979, pp. 76 – 159.
30. N. G. Moll, G. J. Quarderer, *Chem. Eng. Prog.* **75** (1979, Nov.) 46 – 50.
31. J. W. Rosenthal, A. J. Dahlberg, C. W. Kuehler, D. R. Cash et al., *Fuel* **61** (1982) 1045.
32. C. W. Lamb, J. M. Lee, R. V. Nalitham, M. J. Moniz et al.: *Proc. Dept. of Energy Direct Coal Liquefaction Contractors' Review Meeting,* Albuquerque Oct. 17 – 18, 1984, pp. 1 – 33.
33. A. N. Schiffer, M. Peluso, J. Chen, H. D. Schindler, *Energy Prog.* **2** (1982) no. 4, 220.
34. A. Talib, D. Gray, G. Tomlinson, M. Neuworth et al.: "A Technical Assessment of Environmental Control Technologies for the Proposed H-Coal Commercial Plant," MTR-83W26, Mitre Corp., McLean, April 1983.
35. D. Gray, M. Neuworth, A. Talib, G. Tomlinson: "A Technical Assessment of Environmental Control Technologies for the Exxon Donor Solvent Commercial Plant," MTR-83 W230, Mitre Corp., McLean, December 1983.
36. J. B. MacArthur, "Evolution of HRI's Coal Liquefaction Technologies," *Proceedings of the 10th International Pittsburgh Coal Conference,* 1993, p. 283.
37. E. Wolowski, H. Hosang: paper presented at the *27th Deutsche Gesellschaft für Mineralölwissenschaft und Kohlechemie (DGMK) Annual Meeting,* Aachen, Oct.6 – 8, 1982.
38. G. Koelling, I. Romey, E. Wolowski: "Current German Developments in Coal Liquefaction Technology," *Energy Technology VI, Energy Conference,* Washington, D.C., Feb. 26 – 28, 1979.
39. F. Friedrich, I. Romey, B. Strobel: *88th Meeting of the American Institute of Chemical Engineers,* Philadelphia 1980.
40. Y. Kitaoka, M. Ueda, M. Yoshitake: *Pan-Pacific Synfuels Conference,* **vol. 1,** The Japan Petroleum Institute, Tokyo, November 17 – 19, 1982, pp. 356 – 364.
41. S. Ueda, Y. Miki, M. Naka, M. Wada: "Status of Synfuels Development in Japan," *19th Intersociety Energy Conversion Engineering Conference (IECEC),* San Francisco, Aug. 19 – 24, 1984.
42. T. Hosoi, T. Yamada, K. Hayakawa, *Pan-Pacific Synfuels Conference,* **vol. 1,** The Japan Petroleum Institute, Tokyo, November 17 – 19, 1982, pp. 349 – 355.
43. H. Kajimoto, T. Shimada, M. Kaneko, *Pan-Pacific Synfuels Conference,* **vol. 1,** The Japan Petroleum Institute, Tokyo, November 17 – 19, 1982, pp. 365 – 371.
44. T. Kano: "Results of 150 t/d Pilot Plant Operation," *Proceedings of the Advanced Clean Coal Technology International Symposium '98,* Tokyo, October 1998, p. 181.
45. K. Imada: "Summarization of the Research and Development on the 1 t/d PSU," *ibid* 1998, p. 208.
46. K. Shimasaki: "Brown Coal Liquefaction Development," *ibid* 1998, p. 217.
47. K. Hirano, *Fuel Process. Technol.* **62,** (2000) p. 109.
48. *Synfuels Week* **6** (1985, Feb. 4) 5.
49. *International Energy Annual 1983,* DOE/EIA-0219 (83), Energy Information Administration, Washington, D.C., Nov., 1984.
50. G. M. Kimber: "A History of U.K. Coal Liquefaction," *Coal* R078, ETSU, Harwell, December 1996.
51. P. Sabatier, J. D. Senderens, *C. R. Hebd. Seances Acad. Sci* **134** (1902) 514 – 680.
52. BASF, DE 293 787, 1913.

53. F. Fischer, H. Tropsch, *Brennst.-Chem.* **4** (1923) 276.
54. GB 255 818, 1926 (F. Fischer, H. Tropsch).
55. R. B. Anderson: *The Fischer–Tropsch Synthesis,* Academic Press, Orlando 1984.
56. C. D. Frohning in I. Falbe (ed.): *Fischer–Tropsch Synthesis: Chemierohstoffe aus Kohle,* Thieme, Stuttgart 1977.
57. M. E. Dry in J. R. Anderson (ed.): *The Fischer–Tropsch Synthesis, Catalysis Science and Technology,* **vol. 1,** Springer-Verlag, Berlin 1981, Chap. 4.
58. H. Pichler, *Adv. Catal.* **4** (1952) 271.
59. M. E. Dry in B. Leach (ed.): *The Sasol Fischer–Tropsch Processes, Applied Industrial Catalysis,* **vol. 2,** Academic Press, Orlando 1983, pp. 167 – 213.
60. B. Jager: "Sasol's Advanced Commercial Fischer – Tropsch Processes," *AIChE Spring Meeting,* Houston March 10 – 13, 1997.
61. B. Jager, R. C. Kelfkens, A. P. Steynberg, H. E. Curry-Hyde, R. F. Hove (eds.): *Natural Gas Conversion II,* Elsevier Science B.V., Amsterdam 1994, p. 419.
62. A. P. Steynberg, R. L. Espinoza, B. Jager, A. C. Vosloo: "High Temperature Fischer – Tropsch Synthesis in Commercial Practice," *Appl. Catal., A: General* (1999) p. 1 – 14.
63. Y. Ohno: "Dimethyl Ether Synthesis form Coal Derived Gas", *Proceedings of Advanced Clean Coal Technology International Symposium '99,* Tokyo 1999, p. 251.
64. Sun. W. Chun: "Coal Gasification-Coal Liquefaction: A New Philosophy for A Changing Time", *Proceedings of the Advanced Clean Coal Liquefaction International Symposium '98,* Tokyo, October 1998, p. 172.
65. E. Schmetz: "Coal Liquefaction Program – Overview", *Proceedings of the 14th AIST-NEDO/DOE-FECT Joint Technical Meeting,* Tokyo, October 1997.
66. "Coal Gasification Systems: A Guide to Status, Applications and Economics," Electric Power Research Inst. (EPRI), report number AP 3109, June 1983.
67. *Handbook of Gasifiers and Gas Treatment Systems,* report no. FE–1772–11, prepared for Energy Research and Development Agency (ERDA) by Dravo Corp., Pittsburgh, February 1976.
68. J. C. W. Kuo: "Gasification and Indirect Liquefaction," in B. R. Cooper, W. A. Ellingson (eds.): *The Science and Technology of Coal and Coal Utilization,* Plenum Publ., New York 1984, Chap. 5.
69. H. H. Lowry (ed.): *Chemistry of Coal Utilization,* vol. 2, J. Wiley & Sons, New York 1947.
70. "Impact of Developing Technology on Indirect Liquefaction," report no. MTR 80W326, prepared for DOE by Mitre Corp., McLean, November 1980.
71. Max Schriener: "Research Guidance Studies to Assess Gasoline from Coal by Methanol-to-Gasoline and Sasol Type Fischer–Tropsch Technologies," final report no. FE 2447–13, prepared for DOE by Mobil Research & Development Corp., Paulsboro, August 1978.
72. "Further Studies of Developing Technology for Indirect Liquefaction," report no. MTR 82W32, prepared for DOE by Mitre Corp., McLean, March 1982.
73. J. D. Holmgren et al., *Mod. Power Systems* **44** (1981, March).
74. E. V. Vogt, M. J. Van der Burgt, *Chem. Eng. Prog.* **76** (1980, March) 65.
75. "Comparison of Acid Gas Removal Processes," report no. FE 2240 – 49, prepared for DOE by C. F. Braun & Co., Alhambra 1978.
76. D. K. Beavon, B. Kouzel, J. W. Ward, in M. Raymont (ed.): *Claus Processing of Novel Acid Gas Streams Sulfur: New Sources and Uses,* ACS Symposium Series no. 183, Am. Chem. Soc., Washington, D.C., 1982, Chap. 4.
77. *Hydrocarbon Process.* **61** (1982, April) 102.
78. R. D. Parekh: *Handbook of Gasifiers and Gas-Treatment Systems,* report no. DOE/ET/10159–T24, prepared for DOE by UOP/SDC, McLean, 1982.
79. M. E. Dry, *Ind. Eng. Chem. Prod. Res. Dev.* **15** (1976) no. 4, 282.
80. R. L. Espinoza, A. P. Steynberg, B. Jager, A. C. Vosloo: "Low Temperature Fischer – Tropsch Synthesis from a Sasol Perspective," *Appl. Catal., A: General* 1999, p. 4669.
81. B. Jager in M. de Pontes, R. L. Espinoza, C. P. Nicolaides, H. H. Scholtz, M. S. Scurell (eds.): *Natural Gas Conversion IV,* Elsevier Science B.V., Amsterdam 1997, p. 219.
82. P. Biloen, W. M. H. Sachtler: *Mechanism of Hydrocarbon Synthesis Over F–T Catalysis, Advances in Catalysis,* **vol. 30,** Academic Press, New York 1981, pp. 165 – 216.
83. V. Ponec, *Catal. Rev. Sci. Eng.* **18** (1978) no. 1, 151.
84. H. Pichler, H. Schulz, *Chem.-Ing.-Tech.* **42** (1970) 1162.

85. B. Jager, R. L. Espinoza: "Advanced in Low Temperature Fischer – Tropsch Synthesis", *Catal. Today* **23** (1995) 17.
86. J. H. Fourie: "Gas Conversion to Middle Distillates and Chemicals," *Gastech 96,* Vienna, December 1996.
87. J. C. Hoogendoorn: "New Applications of the Fischer – Tropsch Process," *Clean Fuels from Coal Symposium, Ill,* Institute of Gas Technology (IGT), Chicago, Ill., Jun. 23 – 27, 1975, pp. 343 – 358.
88. G. V. Schulz, *Z. Phys. Chem. Abt. B.* **29** (1935) 299.
89. P. J. Flory, *J. Am. Chem. Soc.* **58** (1936) 1877.
90. M. A. Vannice, *J. Catal.* **37** (1975) 449, 462.
91. H. Pichler, *Adv. Catal.* **4** (1952) 289.
92. R. J. Madon, H. Shaw, *Catal. Rev.* **15** (1977) no. 1, 69.
93. M. E. Dry: "High Yield, High Quality Diesel from Fischer–Tropsch," *Coal Liquefaction Conference,* Pretoria, August 1982.
94. A. Brink: "Fischer–Tropsch Product Workup," *Coal Liquefaction Conference,* Pretoria, August 1982.
95. H. Koelbel, M. Ralek, *Catal. Rev.* **21** (1980) no. 2, 225.
96. M. E. Dry, *Hydrocarbon Process.* **61** (1982, Aug.) 121.
97. R. Farley, D. J. Ray, *J. Inst. Pet. London* **50** (1964) no. 482, 27.
98. M. D. Schlesinger, H. Crowell, M. Leva, H. Storch, *Ind. Eng. Chem.* **43** (1951) no. 6, 1474.
99. A. K. Mitra, A. N. Roy, *Indian Chem. Eng.* **27** (1963, July).
100. T. Kunugi, W. Negishi, *Sekiyu Gakkai Shi* **11** (1968) 44.
101. J. C. W. Kuo: "Slurry F–T/Mobil Two-Stage Process of Converting Syngas to High Octane Gasoline," no. DOE/PC/30022-10, final report prepared for DOE by Mobil Research & Development Corp., Paulsboro, 1983.
102. Mobil Oil Corp., US 4 126 644, 1978 (P. D. Ceasar, W. E. Garwood, A. W. Peters).
103. A. Techno: "Economic Assessment of the Mobil Two-Stage Slurry Fischer–Tropsch/ZSM-5 Process," report no. MTR-84W173, prepared for Sandia National Laboratories by Mitre Corp., McLean, November 1984.
104. D. Gray, M. Lytton, M. Neuworth,: "Potential Improvements in SASOL Technology," *Coal Liquefaction Conference,* Pretoria, August 1982.
105. P. L. Rogerson, *Chem. Eng. Prog. Symp. Ser.* **98** (1970) 66.
106. J. K. Paul (ed.): "Methanol Technology and Application in Motor Fuels, Chemical Technology," review no. 114, Noyes Data Corp., Park Ridge 1978.
107. E. Supp, *Hydrocarbon Process.* **60** (1981, Mar.) 71.
108. R. Salmon, M. S. Edwards, R. M. Wham: "Production of Methanol and Methanol Related Fuels from Coal," Oak Ridge National Laboratory (ORNL), report ORNL-5564, Oak Ridge, May, 1980.
109. Liquid Phase Methanol, Technical Progress Report no. 5, prepared for DOE by Air Products &Chemicals and Chem. Systems, January 1983.
110. J. Maziuk, R. Gould, K. H. Keim, H. Gierlich: "The MTG Process," *Petro Pacific '84 Conference,* Melbourne, September 1984.
111. S. L. Meisel, J. P. McCullough, C. H. Lechthaler, P. B. Weisz, *Chemtech.* **6** (1976) 86.
112. *Chem. Eng.* 1980, (Apr. 7) 43.
113. A. Y. Kam, W. Lee: "Fluid Bed Process Studies on Selective Conversion of Methanol to High Octane Gasoline," report no. DOE/FE 2490-15, prepared for DOE by Mobil Research & Development Corp., Paulsboro, 1978.
114. T. H. Fleisch, P. C. Meurer: "DME: The Diesel Fuel for the 21st Century?", *Engine and Environment,* Graz, 1995.
115. J. B. Hansen, S.-E. Mikkelsen: "DME as Neat Diesel Fuel", Gas-to-Liquid Processing – Bringing Clean Fuels to Market – , San Antonio, March 1998.
116. J. B. Hansen, B. Voss, F. Joensen, I. D. Sigurŏardóttir: "Large Scale Manufacture of Dimethyl Ether: A New Alternative Diesel Fuel from Natural Gas", SAE Paper 950063, 1995.
117. I. Dybkjær, J. B. Hansen: "Large Scale Production of Alternative Synthetic Fuels from Natural Gas", *Proceedings of the 4th international gas conversion symposium,* Krüger National Park, South Africa, November 1995.
118. DME Newsletter 1 – 3: "Dimethyl Ether as Automotive Fuel", TNO Road Vehicles Research institute IEA, Implementing Agreement on Alternative Motor Fuels (IEA/AMF), June 1999.

119. M. Mizuguchi et al., "Slurry Phase Synthesis of Dimethylether from Syngas – A Reactor Model Simulation", *Proceedings of the 15th International Pittsburgh Coal Conference,* September 1998.
120. Y. Ohno, T. Shikada, T. Ogawa, M. Ono, M. Mizuguchi, K. Fujimoto: "New Clean Fuel from Coal – Dimethyl Ether – ", *213th ACS National Meeting,* San Francisco, April 1997.
121. T. Ogawa, Y. Ohno, T. Shikasa, M. Ono, M. Mizuguchi, K. Fujimoto: "New Clean Fuel from Coal – Direct Dimethyl Ether Synthesis from Hydrogen and Carbon Monoxide – ", *14th Annual International Pittsburgh Coal Conference & Workshop,* Taiyuan, September 1997.
122. T. Shikada, Y. Ohno, T. Ogawa, M. Ono, M. Mizuguchi, K. Tmura, K. Fujimoto: "Direct Synthesis of Dimethyl Ether from Synthesis Gas", *Stud. Surf. Sci. and Catal.* **119** (1998) 515.
123. Y. Adachi, M. Komoto, I. Watanabe, Y. Ohno, K. Fujimoto, *Fuel* **79** (2000) 229.
124. Y. Ohno, T. Ogawa, N. Inoue, H. Miyano, T. Tao: "Dimethyl Ether (DME): An Ultra Clean and Economical Fuel from Coal", *The 9th Japan/Australia Joint Technical Meeting on Coal,* Melbourne, June 1999.
125. R. C. Weber, D. R. Herman, M. E. Smock: "Sulfur Emission Control Technology for Coal Conversion Plants," DOE/ET/14746-3, topical report prepared for DOE by Environmental Research & Technology and Carnegie-Mellon Univ., October 1981.
126. M. N. Krauss: "Baghouses," *Chem. Eng.* 1979, (Apr. 9) 94.
127. R. K. Lyon, *Hydrocarbon Process.* 1979 (Oct.) 109.
128. "Workshop on Environmental Control Technology for Coal Gasification," Electric Power Research Inst. (EPRI), report AP-3006, Palo Alto, April 1983.
129. C. D. Swaim, Jr.: "The Scot Process," *Sulfur Removal and Recovery from Industrial Processes,* Advances in Chemistry Series no. 139, Amer. Chem. Soc., Washington D.C., 1975, Chap. 9.
130. R. E. Hicks: "Wastewater Treatment in Coal Conversion," EPA report 600/7–79–133, Water Purification Associates, Cambridge (MA), June 1979.
131. E. C. Cavanaugh: *Environmental Assessment Data Base for Low/Medium Btu, Gasification Technology,* vol. 2, EPA 600/7–77–125 b, November 1977.
132. M. Beychok: "State of the Art Wastewater Treatment," in J. D. Wall (compiler): *Environmental Management Handbook for the Hydrocarbon Processing Industries*, Gulf Publ. Co., Houston 1980.
133. *Fed. Regist.* **45** (1980, May 19) no. 98, Part 261.4.
134. *Fed. Regist.* **45** (1980, Nov. 19) no. 225.
135. R. M. Davidson in M. L. Gorbaty, J. W. Larsen, J. Wender (eds.): *Coal Science,* **vol. 1,** Academic Press, New York 1982, pp. 84 – 160.
136. P. H. Given in M. L. Gorbaty, J. W. Larsen, J. Wender (eds.): *Coal Science,* **vol. 3,** Academic Press, Orlando, 1984, pp. 63 – 252.
137. J. R. Nelson, O. P. Mahajan, P. L. Walker, Jr., *Fuel* **59** (1980) 831.
138. L. M. Lucht, N. A. Peppas in B. D. Blaustein, B. C. Bockrath, S. Friedman (eds.): *New Approaches in Coal Chemistry,* ACS Symposium Series no. 169, Amer. Chem. Soc., Washington D.C., 1981, pp.43 – 59.
139. T. Green, J. Kovac, D. Brenner, J. W. Larsen in R. A. Meyers (ed.): *Coal Structure,* Academic Press, New York 1982, pp. 199 – 282.
140. I. Wender, *Catal. Rev., Sci. Eng.* **14** (1976) no. 1, 97 – 129.
141. G. P. Curran, R. T. Struck, E. Gorin, *Ind. Eng. Chem. Process Des. Dev.* **6** (1967) 166.
142. F. R. Mayo, J. G. Huntington, N. A. Kirshen in J. W. Larsen (ed.): *Organic Chemistry of Coal, ACS Symposium Series no. 71,* Amer. Chem. Soc., Washington, D.C., 1978, pp. 126 – 130.
143. L. W. Vernon, *Fuel* **59** (1980) 102.
144. N. C. Deno, A. Greigger, A. D. Jones, W. G. Rakitsky et al., *Fuel* **59** (1980) 701.
145. N. C. Deno, K. Curry, A. D. Jones, R. Minard et al. in B. R. Cooper, L. Petrakis (eds.): *Chemistry and Physics of Coal Utilization,* Amer. Inst. Phys., New York 1980, pp. 154 – 166.
146. E. Kuhlmann, E. Boerwinkle, M. Orchin, *Fuel* **60** (1981) 1002.
147. A. J. Szladow, P. H. Given: "Some Aspects of the Mechanism and Kinetics of Coal Liquefaction," report no. FE–2494–TR–3, special technical report prepared for DOE by The Pennsylvania State University under contract no. AC22 – 76–ET–10587, 1979, pp. 171.
148. J. S. Youtcheff, P. H. Given, *Fuel* **61** (1982) 980.
149. D. D. Whitehurst, T. O. Mitchell, M. Farcasiu, *Coal Liquefaction,* Academic Press, New York 1980.

150. H. Huang, K. Wang, S. Wang, M. T. Klein, W. H. Clakins, *Energy Fuels* **10** (1996) 641.
151. K. Mae, C. Song, M. Shimada, K. Miura, *Energy Fuels* **12** (1998) 975.
152. E. M. Suuberg, D. Lee, J. W. Larsen, *Fuel* **64** (1985) 1668.
153. P. R. Salomon, M. A. Serio, G. V. Desphande, E. Kroo, *Energy Fuels* **4** (1990) 42.
154. M. A. Serio, D. A. Hamblen, J. R. Markham, P. R. Salomon, *Energy Fuels* **1** (1987) 138.
155. J. A. Manion, D. F. McMillen, R. Malhotra, *Energy Fuels* **10** (1996) 776.
156. T. P. Eskay, P. F. Britt, A. C. Buchanan, *Energy Fuels* **10** (1996) 1257.
157. A. K. Saini, M. M. Coleman, C. Song, H. H. Schobert, *Energy Fuels* **7** (1993) 328.
158. C. Song, H. H. Schobert, P. G. Hatcher, *Energy Fuels* **6** (1992) 326.
159. I. Mochida, *Fuel* **62** (1983) 659.
160. Y. Inukai, S. Arita, H. Hirosue, *Energy Fuels* **9** (1995) 67.
161. J. T. Josepf, T. R. Farrai, *Fuel* **71** (1992) 75.
162. K. Miura, K. Mae, S. Asaoka, T. Yoshimura, K. Hashimoto, *Energy Fuels* **5** (1991) 340.
163. C. J. Chu, S. A. Cannon, R. H. Hauge, J. L. Margrave, *Fuel* **65** (1986) 1740.
164. K. Ofosu-Asante, L. M. Stock, R. F. Zabransky, *Fuel* **68** (1989) 567.
165. D. F. McMillen, R. Malhotra, G. P. Hum, S. J. Chang, *Energy Fuels* **1** (1987) 193.
166. D. F. McMillen, R. Malhotra, D. S. Tse, *Energy Fuels* **5** (1991) 179.
167. K. Kidena, N. Bandoh, M. Kouchi, S. Murata, M. Nomura, *Fuel* **79** (2000) 317.
168. M. Siskin, A. R. Katrizky, M. Balasubramanian, *Energy Fuels* **5** (1991) 770.
169. C. E. Snape, C. J. Lafferty, H. P. Stephens, R. G. Dosch, E. Klavetter, *Fuel* **70** (1991) 393.
170. C. E. Snape, C. Bolton, R. G. Dosch, H. P. Stephens, *Energy Fuels* **3** (1989) 421.
171. J. W. Larsen, T. K. Green, J. Kovac, *J. Org. Chem.* **50** (1985) 4729.
172. M. Miyake, L. M. Stock, *Energy Fuels* **2** (1988) 815.
173. M. Iino, T. Takanoashi, H. Ohsuga, K. Toda, *Fuel* **67** (1988) 1639.
174. T. Takanoashi, M. Iino, *Energy Fuels* **8** (1994) 395.
175. T. Aida, *Journal of Japan Institute of Energy* **70** (1991) 820.
176. M. L. Gorbaty, *Fuel* **73** (1994) 1819.
177. M. Nishioka, W. Laird, P. G. Bendale, R. A. Zeli, *Energy Fuels* **8** (1994) 643.
178. J. W. Larsen, P. Lee, S. E. Shawver, *Fuel Process. Technol.* **12** (1986) 51.
179. F. J. Derbyshire, A. Marzec, H. R. Schuten, M. A. Wilson, A. Davis, P. Tekely, J. J. Delpuech, A. Jurkiewicz, C. E. Bronnimann, R. A. Wind, G. E. Maciel, R. Narayan, K. Bartle, C. Snape, *Fuel* **68** (1989) 1091.
180. P. Painter, *Energy Fuels* **6** (1992) 863.
181. M. Nishioka, *Fuel* **71** (1992) 941.
182. K. Miura, K. Mae, S. Takebe, H. Wakiyasu, K. Hashimoto, *Energy Fuels* **8** (1994) 874.
183. E. Gorin in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* **suppl. vol.,** J. Wiley & Sons, New York 1981,, pp. 1845 – 1918.
184. R. G. Ruberto, D. C. Cronauer, L. G. Galya, *Fuel Process. Technol.* **5** (1981) 103.
185. R. G. Ruberto, D. C. Cronauer, *Fuel Proc. Technol.* **2** (1979) 215.
186. F. W. Steffgen, K. T. Schroeder, B. C. Bockrath, *Anal. Chem.* **51** (1979) 1164.
187. B. C. Bockrath, K. T. Schroeder, F. W. Steffgen, *Anal. Chem.* **51** (1979) 1168.
188. F. K. Schweighardt, B. M. Thames, *Anal. Chem.* **50** (1978) 1381.
189. B. C. Bockrath, R. P. Noceti, *Fuel Process. Technol.* **2** (1979) 143.
190. L. Shadle, P. H. Given, *Fuel* **61** (1982) 972.
191. J. H. Shinn, A. J. Dahlberg, C. W. Kuehler, J. W. Rosenthal: "The Chevron Co-Refining Process," *Proc. 9th Annual Electric Power Research Inst. (EPRI) Contractors' Conference on Coal Liquefaction,* Palo Alto, May 8 – 10, 1984.
192. S. E. Moschopedis, B. Ozum, *Liq. Fuels Technol.* **2** (1984) no. 2, 193.
193. S. E. Moschopedis, *Liq. Fuels Technol.* **2** (1984) no. 2, 177
194. S. E. Moschopedis, *Fuel* **59** (1980) 67.
195. S. E. Moschopedis, R. W. Hawkins, J. F. Fryer, J. G. Speight, *Fuel* **59** (1980) 647.
196. Z. Liu, J. W. Zondlo, D. B. Dadyburjor, *Energy Fuels* **9** (1995) 673.
197. C. Bengoa, J. Font, A. Moros, A. Fabregat, F. Giralt, *Fuel* **74** (1995) 1704.
198. T. Yoshida, H. Nagaishi, M. Sasaki, M. Yamamoto, T. Kotanigawa, A. Sasaki, K. Idogawa, T. Fukuda, R. Yoshida, Y. Maekawa, *Energy Fuels* **9** (1995) 685.
199. K. S. Rothenberger, A. V. Cugini, R. L. Thompson, M. V. Ciocco, *Energy Fuels* **11** (1997) 849.
200. H. K. Joo, C. W. Curtis, *Energy Fuels* **11** (1997) 801.
201. R. K. Sharma, J. W. Zondlo, D. B. Dadyburjor, *Energy Fuels* **12** (1988) 589.

202. S-L. Wang, C. W. Curtis, *Energy Fuels* **8** (1994) 446.
203. J. Higashine, A. Kido, K. Takahashi, M. Miyake, M. Nomura, *Proceedings of the 10th International Pittsburgh Coal Conference,* 1993, p. 241.
204. D. E. Herrick, J. W. Tierny, I. Wender, *Energy Fuels* **4** (1990) 231.
205. D. S. Ross, J. E. Blessing, *Fuel* **57** (1978) 379.
206. F. Mondragon, R. Kamoshita, T. Katoh, H. Itoh et al., *Fuel* **63** (1984) 579.
207. C. S. Carlson, A. W. Langer, J. Stewart, R. M. Hill, *Ind. Eng. Chem.* **50** (1958) 1067.
208. R. C. Neavel, *Fuel* **55** (1976) 237.
209. B. C. Bockrath in M. L. Gorbaty, J. W. Larsen, I. Wender (eds.): *Coal Science,* **vol. 2,** Academic Press, New York 1983, pp. 65 – 124.
210. R. C. Neavel, *Philos. Trans. R. Soc. London, Ser. A* **300** (1981) 141.
211. I. Petrakis, D. W. Grandy, G. L. Jones, *Fuel* **62** (1983) 665.
212. R. F. Sprecher, H. L. Retcofsky, *Fuel* **62** (1983) 473.
213. F. J. Derbyshire, D. D. Whitehurst, *Fuel* **60** (1981) 655.
214. I. Mochida, A. Yufu, K. Sakanashi, Y. Korai, *Fuel* **67** (1988) 114.
215. B. Chawla, H. A. Dabbagh, B. H. Davis, *Energy Fuels* **8** (1994) 355.
216. M. Godo, A. Ishihara, K. Tabe, *Energy Fuels* **11** (1997) 724.
217. F. J. Derbyshire, G. A. Odoerfer, L. R. Rudnick, P. Varghese et al., Electric Power Research Institute (EPRI), report AP–2117, vol. 1, Research Project 1655–1, 1981.
218. N. Ikenaga, T. Sakoda, T. Matsui, K. Ohno, T. Suzuki, *Energy Fuels* **11** (1997) 183.
219. D. Mayer, A. P. Oviawe, D. Nicole, J. C. Lauer, *Fuel* **69** (1990) 1317.
220. A. P. Oviawe, D. Nicole, J. L. Fringant, *Fuel* **72** (1993) 65.
221. A. P. Oviawe, D. Nicole, R. Gerardin, *Fuel* **74** (1995) 259.
222. P. M. Rahimi, W. H. Dawson, J. F. Kelly, *Fuel* **70** (1991) 95.
223. M. W. Bedell, C. W. Curtis, J. H. Hool, *Energy Fuels* **7** (1993) 200.
224. K. Aimoto, I. Nakamura, K. Fujimoto, *Energy Fuels* **5** (1991) 739.
225. R. Bruecker, G. Koelling, *Brennst. Chem.* **46** (1965) 41.
226. D. Von Hausigk, G. Koelling, F. Ziegler, *Brennst. Chem.* **50** (1969) no. 1, 8.
227. L. F. Atherton, C. Kulik: "Advanced Coal Liquefaction," *Annual AIChE Meeting,* Los Angeles, 1982.
228. F. J. Derbyshire, G. A. Odoerfer, D. D. Whitehurst, *Fuel* **63** (1984) 56.
229. F. Heinz-G., J. W. Stadelhofer, D. Biermann, *Fuel* **62** (1983) 78.
230. D. D. Whitehurst in D. D. Whitehurst (ed.): *Coal Liquefaction Fundamentals,* ACS Symposium Series no. 139, Amer. Chem. Soc., Washington, D.C., 1980, pp. 133 – 165.
231. H. Huang, K. Wang, S. Wang, M. T. Klein, W. H. Calkins, *Energy Fuels* **10** (1996) 641.
232. V. R. Pradhan, G. D. Holder, I. Wender, J. W. Tierny, *Ind. Eng. Chem. Res.* **31** (1992) 2051.
233. F. J. Derbyshire, P. Varghese, D. D. Whitehurst, *Fuel* **61** (1982) 859.
234. M. Aiura, T. Masunaga, K. Moriya, Y. Kageyama, *Fuel* **63** (1984) 1138.
235. L. Huang, C. Song, H. H. Schobert, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **37** (1992) 223.
236. C. Song, H. H. Schobert, *Energy Fuels* **6** (1992) 326.
237. C. Song, H. H. Schobert, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **37** (1992) 976.
238. Y. Kamiya, H. Sato, T. Yao, *Fuel* **57** (1978) 681.
239. D. D. Whitehurst, M. Farcasiu, T. O. Mitchell, Electric Power Research Inst. (EPRI), report AF-252, project RP-410, 1976.
240. A. Pott, H. Broche, *Glueckauf* **66** (1933) 903.
241. M. Orchin, H. H. Storch, *Ind. Eng. Chem.* **40** (1948) 1385.
242. T. Yao, Y. Kamiya, *Nippon Kagaku Kaishi,* 1980, 893.
243. F. P. Burke, J. A. Kleinpeter, P. J. Dudt, D. C. Jones: "Short Residence Time and Low Severity SRC Coal Liquefaction with Selective Product Recycle," *National AIChE Meeting,* Orlando, Feb. 28 – Mar. 4, 1982.
244. R. H. Wolk: "Evolutionary Changes in the Perception of the Role of the Recycle Solvent in the Direct Hydroliquefaction of Coal," *AIChE Spring 1981 National Meeting,* Houston, Apr. 5 – 9, 1981.
245. D. Kang, E. N. Givens, *Chem. Eng. Prog.* **80** (1984, November) 38.
246. F. J. Derbyshire, G. A. Odoerfer, P. Varghese, D. D. Whitehurst, *Fuel* **61** (1982) 899.
247. O. Okuma, K. Saito, A. Kawashima, K. Okazaki, Y. Nakako, *Fuel Process. Technol.* **14** (1986) 23.
248. M. Aiura, T. Masunaga, Y. Kageyama, S. Kawai, *Fuel Process. Technol.* **14** (1986) 13.
249. O. Okuma, M. Yasumuro, T. Matsumara, *Fuel* **75** (1996) 313.

250. O. Okuma, M. Yasumuro, S. Yanai, *Fuel* **77** (1998) 797.
251. O. Okuma, M. Yasumuro, Y. Kageyama, T. Matsumura, *Proceedings of Pittsburgh Coal Conference,* 1993, p. 235.
252. O. Okuma, M. Yasumuro, *Fuel* **77** (1998) 1755.
253. J. R. Longanbach, *Chem. Eng. Prog.* **80** (1984, Nov.) 29.
254. P. L. Walker, Jr., W. Spackman, P. H. Given, A. Davis et al., Electric Power Research Inst. (EPRI), report AF-832, project 366-1, 1978.
255. D. D. Whitehurst, M. Farcasiu, T. O. Mitchell, J. J. Dickert, Jr., Electric Power Research Inst. (EPRI), report AF-1298, project RP-410, vol. 2, 1979.
256. R. G. Ruberto, *Fuel Process. Technol.* **3** (1980) 7.
257. F. J. Derbyshire, D. D. Whitehurst, *High Temp. High Pressures* **13** (1981) 177.
258. R. I. McNeil, D. C. Young, D. C. Cronauer, *Fuel* **62** (1983) 806.
259. D. C. Cronauer, D. M. Jewell, Y. T. Shah, K. A. Kueser, *Ind. Eng. Chem. Fundam.* **17** (1978) 291.
260. J. M. L. Penninger, H. W. Slotboom, *Recl. Trav. Chim. Pays-Bas* **92** (1973) 513.
261. R. J. Hooper, H. A. J. Battaerd, D. G. Evans, *Fuel* **58** (1979) 132.
262. J. A. Franz, D. M. Camaioni, *Fuel* **59** (1980) 803.
263. J. A. Franz, D. M. Camaioni, *J. Org. Chem.* **45** (1980) 5247.
264. J. J. de Vlieger, A. P. G. Kieboom, H. van Bekkum, *Fuel* **63** (1984) 334.
265. R. I. McNeil, D. C. Cronauer, D. C. Young, *Fuel* **62** (1983) 401.
266. C. J. Collins, V. F. Raaen, B. M. Benjamin, P. H. Maupin et al., *J. Am. Chem. Soc.* **97** (1979) 5009.
267. P. H. Given, D. C. Cronauer, W. Spackman, H. L. Lovell et al., *Fuel* **54** (1975) 40.
268. J. R. Pullen in M. L. Gorbaty, J. W. Larsen, I. Wender (eds.): *Coal Science,* **vol. 2,** Academic Press, New York 1983, pp. 173 – 288.
269. P. H. Given, W. Spackman, A. Davis, R. G. Jenkins in D. D. Whitehurst (ed.): *Coal Liquefaction Fundamentals, ACS Symposium Series no. 139,* Amer. Chem. Soc., Washington, D.C., 1980.
270. J. W. Clarke, G. M. Kimber, T. D. Rantell, D. E. Shipley, in D. D. Whitehurst (ed.): *Coal Liquefaction Fundamentals, ACS Symposium Series no. 139,* Amer. Chem. Soc., Washington, D.C., 1980. pp. 111 – 129.
271. J. W. Stadelhofer, R. Gerhards, *Fuel* **60** (1981) 367.
272. B. O. Strobel, F. Friedrich, *Proc.-Int. Conf. Coal Sci.,* Sydney 1985, p. 7.
273. O. Okuma, *Fuel* **79** (2000) 355.
274. I. Mochida, A. Takarabe, K. Takeshita, *Fuel* **58** (1979) 17.
275. I. Mochida, K. Iwamoto, T. Tahara, Y. Korai et al., *Fuel* **61** (1982) 603.
276. R. C. Neavel in M. L. Gorbaty, J. W. Larsen, J. Wender (eds.): *Coal Science,* **vol. 1,** Academic Press, New York 1982, pp. 1 – 19.
277. H. Marsh, R. C. Neavel, *Fuel* **59** (1980) 511.
278. H. R. Brown, P. L. Waters, *Fuel* **45** (1966) 17.
279. H. N. M. Dormans, D. W. van Krevelen, *Fuel* **39** (1960) 273.
280. J. G. Walters, J. L. Shultz, J. A. Glaenzer, U.S. Bureau of Mines, Report of Investigations 6973, 1967.
281. K. Ouchi, K. Tanimoto, M. Makabe, H. Itoh, *Fuel* **62** (1983) 1227.
282. J. W. Larsen, T. L. Sams, B. R. Rodgers, *Fuel* **59** (1980) 666.
283. E. A. Sondreal, W. G. Wilson, V. I. Stenberg, *Fuel* **61** (1982) 925.
284. J. B. Stone, K. I. Trachte, S. K. Poddar, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **24** (1979) no. 2, 255.
285. M. L. Gorgaty, *Proc.-Int. Conf. Coal Sci.,* Tokyo 1989, p. 691.
286. P. H. Given, V. Lupton, M. E. Peover: "Studies of the Aromatic Systems in Coals by Chemical and Electrochemical Reduction Processes," *Proc. Resid. Conf. Sci. Use Coal,* Inst. of Fuel, London 1959, pp. 38 – 42.
287. S. Parkash, A. J. Szladow, N. Berkowitz, *Can. Inst. Min. Metall. Bull.* **64** (1971, Nov.) 92 – 94.
288. F. Goodarzi, D. G. Murchison, *Fuel* **51** (1972) 322.
289. K. Robinson, K. D. Kim, "Catalyst Development for Liquefying Eastern and Western Coal," *Proceedings of EPRI Contractors' Conference on Coal Liquefaction,* Palo Alto, May 8 - 10, 1979, p. 9–1.
290. T. Yoshida, Y. Nakata, R. Yoshida, S. Ueda et al., *Fuel* **61** (1982) 824.
291. M. Becker, J. G. Bendoraitis, M. G. Bloch, A. V. Cabal et al.: "Analytical Studies for the H-Coal Process," prepared for DOE by Mobil Research and Development Corp. under contract no. EF-77–C-01–2676, Paulsboro, 1979.

292. G. O. Davies, F. J. Derbyshire, G. M. Kimber: *Economic Commission for Europe (ECE) Symposium on Gasification and Liquefaction of Coal,* Düsseldorf, January 1976.
293. D. C. Cronauer, R. G. Ruberto, R. S. Silver, R. G. Jenkins et al., *Fuel* **62** (1983) 1116.
294. D. C. Cronauer, R. G. Ruberto, R. G. Jenkins, A. Davis et al., *Fuel* **62** (1983) 1124.
295. D. C. Cronauer, R. G. Ruberto, R. S. Silver, R. G. Jenkins et al., *Fuel* **63** (1984) 71.
296. H. Wachowska, W. Pawlak, *Fuel* **56** (1977) 342.
297. E. Furimsky, J. A. MacPhee, L. Vancea, I. Ciavaglia et al., *Fuel* **62** (1983) 395.
298. F. Boyer, A. Peytavy, R. P. Loison, H. H. Lowry (ed): *Chemistry of Coal Utilization,* **suppl. vol.,** J. Wiley & Sons, New York 1963, pp.150 – 201.
299. J. T. Senftle, A. Davis, *Int. J. Coal Geol.* **3** (1984) 375.
300. J. K. Brown, I. G. C. Dryden, D. H. Dunevein, W. K. Joy et al., *J. Inst. Fuel* **31** (1958) 259.
301. R. Liotta, G. Brons, J. Isaacs, *Fuel* **62** (1983) 781.
302. C. Rhoads, J. T. Senftle, M. M. Coleman, A. Davis et al., *Fuel* **62** (1983) 1387.
303. P. D. Swann, D. G. Evans, *Fuel* **58** (1979) 276.
304. F. P. Miknis, D. A. Netzel, T. F. Turner, J. C. Wallace, C. H. Butcher, *Energy Fuels* **10** (1996) 631.
305. K. S. Vorres, D. L. Wertz, V. Malhotra, Y. Dang, J. T. Joseph, R. Fischer, *Fuel* **71** (1992) 1047.
306. K. S. Vorres, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **38** (1993) 587.
307. A. E. Wroblewwski, J. G. Verkade, *Energy Fuels* **6** (1992) 331.
308. F. P. Miknis, D. A. Netzel, T. F. Turner, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **38** (1993) 609.
309. M. A. Serio, P. R. Soloman, E. Kroo, R. Bassilakis, D. McMillan, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **35** (1990) 61.
310. E. M. Suuberg, Y. Otake, Y. Yun, S. C. Deevi, *Energy Fuels* **7** (1993) 384.
311. C. Song, A. K. Saini, H. H. Schobert, *Energy Fuels* **8** (1994) 301.
312. O. Okuma, K. Mae, T. Hirano, Y. Nakako, *Fuel Process. Technol.* **23** (1989) 117.
313. P. Nowacki (ed.): *Lignite Technology,* Noyes Data Corp. Park Ridge 1980.
314. B. Stanmore, D. N. Baria, L. E. Paulson: ”Steam Drying of Lignite: A Review of Processes and Performance,” DOE/GFETC/RI-82/1 (DE 82 007 849), Technical Information Center, DOE, Washington, D.C., 1982.
315. H. R. Appell, R. D. Miller, E. G. Illig, E. C. Moroni et al.: “Coal Liquefaction with Synthesis Gas,” Pittsburgh Energy Technology Center, report PETC/ TR-79/1, DOE, Washington, D.C., September 1979.
316. F. J. Garnish, P. C. Madden, *Chem. Eng. Prog.* **80** (1984, Nov.) 42.
317. L. E. Furlong, E. Effron, L. W. Vernon, E. L. Wilson, *Chem. Eng. Prog.* **72** (1975) 71.
318. D. D. Whitehurst, T. O. Mitchell, M. Farcasiu, J. J. Dickert, Jr., Electric Power Research Inst. (EPRI), report AF-1184, project 779–18, September 1979.
319. S. C. Lim, R. F. Rathbone, A. M. Rubel, E. N. Givens, F. J. Derbyshire, *Energy Fuels* **8** (1994) 294.
320. A. C. Sofianos, A. C. Butler, H. B. Louwrens, *Fuel Process. Technol.* **22** (1989) 175.
321. L. A. Amestica, E. E. Wolf, *Fuel* **65** (1986) 1226.
322. P. J. Cassidy, W. R. Jackson, F. P. Larkins, M. B. Louey, R. J. Sakurovs, *Fuel Process. Technol.* **14** (1986) 231.
323. P. J. Cassidy, W. R. Jackson, F. P. Larkins, R. J. Sakurovs, *Fuel* **65** (1986) 1057.
324. P. J. Cassidy, W. R. Jackson, F. P. Larkins, R. J. Sakurovs, *Fuel* **65** (1986) 374.
325. D. S. Ross, T. K. Green, R. Mansani, G. P. Hum, *Energy Fuels* **1** (1987) 287.
326. S. T. Chan, W. R. Jackson, M. Marshall, *Fuel* **73** (1994) 1628.
327. Y. Watanabe, K. Hata, N. Kawasaki, K. Wada, T. Mitsudo, *Energy Fuels* **8** (1994) 806.
328. K. Hata, K. Wada, T. Mitsudo, *Energy Fuels* **12** (1998) 1181.
329. M. Shibaoka, Y. Takemura, N. J. Russel, K. Ouchi, *Fuel* **65** (1986) 362.
330. D. S. Ross, in M. L. Gorbaty, J. W. Larsen, J. Wender (eds.): *Coal Science,* **vol. 3,** Academic Press, Orlando, Fla., 1984, pp. 301 – 338.
331. H. P. Appell, R. D. Miller, I. Wender: “On the Mechanism of Lignite Liquefaction with Carbon Monoxide and Water”, *Amer. Chem. Soc., Div. Fuel Chem., 163rd National Meeting,* Boston, April 1972.
332. V. I. Stenberg, R. J. Baltisberger, K. J. Klabunde, N. F. Woolsey et al.: “Chemistry of Lignite Liquefaction,” FE-2211–7, U.S. Energy Research & Development Agency Quarterly Report, July/September 1977.
333. Y. Takemura, K. Ouchi, *Fuel* **62** (1983) 1133.

334. F. J. Derbyshire, "Catalysis in Coal Liquefaction: New Direction for Research", IEACR/08, ICBN 92-9029-158-3, IEA Coal Research, London, June 1988.
335. F. Derbyshire, T. Hager, *Fuel* **73** (1994) 1087.
336. S. W. Weller, *Energy Fuels* **8** (1994) 415.
337. A. V. Cugini, D. Krastman, R. G. Lett, V. D. Balsone, *Catal. Today* **19** (1994) 395.
338. I. Mochida, K. Sakanishi, *Adv. Catal.* **40** (1994) 39.
339. S. S. Penner, S. B. Alpert, V. Bendanillo, J. Clardy et al.: *New Sources of Oil and Gas,* Pergamon Press, Oxford, 1982, pp. 19 – 26.
340. L. K. Lee, V. R. Pradhan, R. H. Stalzer, E. S. Johanson, A. G. Comolli, *Proceedings of the Coal Liquefaction and Gas Conversion Contractor's Review Conference,* U.S. DOE, Pittsburgh, September 1994, p. 33.
341. A. G. Comolli, L. K. Lee, V. R. Pradhan, R. H. Stalzer, *Proceedings of the Coal Liquefaction and Gas Conversion Contractor's Review Conference,* U.S. DOE, Pittsburgh, September 1994, p. 19.
342. I. Mochida, K. Sakanishi, N. Suzuki, M. Sakurai, T. Kaneko, *Catal. Surv. Jpn.* **2** (1998) 17.
343. T. Suzuki, *Energy Fuels* **8** (1994) 341.
344. R. M. Davidson: "Mineral Effects in Coal Conversion," report number ICTIS/TR22, International Energy Agency (IEA) Coal Research, London, January 1983.
345. D. Gray, M. B. Neuworth: "Disposable Catalysts in Coal Liquefaction: The Effect of Iron Sulfides," DOE contract no. DEAC01–77ET10280, Mitre Corp., McLean, September 1981.
346. E. E. Donath, M. Hoering, *Fuel Process. Technol.* **1** (1977) 3.
347. "Liquefaction of Low Rank Coals with Slurry Catalyst," prepared for DOE by contract DE-AC22–82PC50001, Gulf Research and Development Co., Jul., 1984, Part 1.
348. P. H. Given: "The Dependence of Coal Liquefaction Behavior on Coal Characteristics," final technical report, part I, Mar., 1981–12 Feb. 1984, prepared for the U.S. DOE, contract no. DE-AC22–81–PC-40784, 1983, pp. 26 – 27.
349. D. K. Mukherjee, P. B. Chowdhury, *Fuel* **55** (1976) 4.
350. G. T. Hager, F. J. Derbyshire, *Proceedings of the 10th International Pittsburgh Coal Conference,* 1993, p. 289.
351. J. G. Darab, J. C. Linehan, D. W. Matson, *Energy Fuels* **8** (1994) 1004.
352. R. Bacaud, M. Besson, G. Djega-Mariadassou, *Energy Fuels* **8** (1994) 3.
353. T. Kaneko, T. Koyama, K. Tazawa, T. Shindo, K. Shimasaki, Y. Kageyama, *J. Jpn. Inst. Energy* **77** (1998) 313.
354. T. Kaneko, K. Tazawa, K. Shimasaki, *J. Jpn. Inst. Energy* **77** (1998) 805.
355. T. Kaneko, K. Tazawa, T. Koyama, K. Satou, K. Shimasaki, Y. Kageyama, *Energy Fuels* **12** (1998) 897.
356. J. Zhao, Z. Feng, F. E. Huggins, G. P. Huffmann, *Energy Fuels* **8** (1994) 1152.
357. G. T. Hager, X. X. Bi, P. C. Eklund, E. N. Givens, F. J. Derbyshire, *Energy Fuels* **8** (1994) 88.
358. D. W. Matson, J. C. Linehan, J. G. Darab, M. F. Buehler, *Energy Fuels* **8** (1994) 88.
359. M. Farcasiu, P. A. Eldredge, S. C. Petrosius, *Energy Fuels* **8** (1994) 48.
360. V. L. Cebolla, M. Diack, D. Cagniant, M. Oberson, R. Bacaud, B. N-P-Donat, *Energy Fuels* **8** (1994) 77.
361. J. Zhao, Z. Feng, F. E. Huggins, G. P. Huffman, *Energy Fuels* **8** (1994) 38.
362. Y. Nakao, S. Yokoyama, Y. Maekawa, K. Kaeriyama, *Fuel* **63** (1984) 721.
363. R. R. Anderson, B. C. Bockrath, *Fuel* **63** (1984) 329.
364. Y. Watanabe, O. Yamada, K. Fujita, Y. Takegami, *Fuel* **63** (1984) 752.
365. N. Ikenaga, Y. Kobayashi, S. Saeki, T. Sakota, Y. Watanabe, H. Yamada, T. Suzuki, *Energy Fuels* **8** (1994) 947.
366. T. Suzuki, O. Yamada, K. Fujita, Y. Takegami, Y. Watanabe, *Ind. Eng. Chem. Prod. Res. Dev.* **24** (1985) 832.
367. T. Suzuki, O. Yamadam, Y. Takahashi, Y. Watanabe, *Fuel Process. Technol.* **10** (1985) 33.
368. T. Suzuki, T. Ando, Y. Watanabe, *Energy Fuels* **1** (1987) 294.
369. Y. Tang, C. W. Curtis, *Energy Fuels* **8** (1994) 63.
370. K. Miki, Y. Yamamoto, A. Inaba, *Fuel* **71** (1992) 825.
371. J. T. Joseph, *Fuel* **70** (1991) 459.
372. S. Okamoto, A. Kitajjima, H. Taniguchi, N. Ikenaga, T. Suzuki, *Energy Fuels* **8** (1994) 1077.
373. M. M. Taghiei, F. E. Huggins, V. Mahajan, G. P. Huffman, *Energy Fuels* **8** (1994) 31.
374. A. V. Cugini, D. Krastman, D. V. Martello, E. F. Frommell, A. W. Wells, G. D. Holder, *Energy Fuels* **8** (1994) 83.

375. V. U. S. Rao, *Energy Fuels* **8** (1994) 44.
376. V. L. Cebolla, M. Diack, M. Oberson, R. Becaud, D. Cagniant, B. N. P. Donat, *Fuel Process. Technol.* **28** (1991) 183.
377. G. T. Hager, X. X. Bi, F. J. Derbyshire, P. C. Eklund, J. M. Stencel, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **36** (1991) 1900.
378. R. R. Anderson, B. C. Bockrath, *Fuel* **63** (1984) 329.
379. C. E. Burgess, H. H. Schobert, *Fuel* **70** (1991) 372.
380. V. R. Pradhan, J. W. Tierny, I. Wender, G. P. Huffman, *Energy Fuels* **5** (1991) 497.
381. T. Yamaguchi, T. Jin, K. Tanabe, *J. Phys. Chem.* **90** (1986) 3148.
382. M. Farcasiu, C. Smith, *Fuel Process. Technol.* **29** (1991) 199.
383. V. R. Pradhan, D. E. Herrick, J. W. Tierney, I. Wender, *Energy Fuels* **5** (1991) 712.
384. H. Hattori, T. Yamaguchi, K. Tanabe, S. Yokoyama, J. Uematsu, Y. Sanada, *Fuel Process. Technol.* **8** (1994) 117.
385. M. M. Ibrahim, M. S. Seehra, *Energy Fuels* **8** (1994) 48.
386. F. J. Derbyshire, A. Davis, H. H. Schobert, P. G. Stansbeny, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **35** (1990) 51.
387. D. Grag, E. N. Givens, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **28** (1983) 200.
388. R. Bacaud, H. Charcosset, M. Jamond, *Fuel Process. Technol.* **24** (1990) 163.
389. P. A. Montano, A. Bommannavar, V. Shah, *Fuel* **61** (1982) 1288.
390. D. B. Dadyburjor, W. R. Stewart, A. H. Stiller, C. D. Stinespring, J. P. Wann, J. W. Zondlo, *Energy Fuels* **8** (1994) 19.
391. J. Y. Kim, P. J. Reucroft, M. Taghiei, V. R. Pradhan, I. Wender, *Energy Fuels* **8** (1994) 886.
392. R. Bacaud, M. Besson, G. Djega-Mariadassou, *Energy Fuels* **8** (1994) 3.
393. J. G. Darab, J. C. Linehan, D. W. Matson, *Energy Fuels* **8** (1994) 1004.
394. N. Shah, J. Zhao, F. E. Huggins, G. P. Huffman, *Energy Fuels* **10** (1996) 417.
395. G. P. Huffman, B. Ganguly, J. Zhao, K. R. P. M. Rao, N. Shah, Z. Feng, F. E. Huggins, M. M. Taghiei, F. Lu, I. Wender, V. R. Pradhan, J. W. Tierney, M. S. Seehra, M. M. Ibrahim, J. Shabatai, E. M. Eyring, *Energy Fuels* **7** (1993) 285.
396. J. C. Linehan, D. W. Matson, G. Darab, *Energy Fuels* **8** (1994) 56.
397. J. M. Lambert, G. Simkovich, P. L. Walker, *Fuel* **59** (1980) 687.
398. G. Djega-Mariadassou, M. Besson, D. Brodzki, H. Charcosset, T. V. Huu, J. Varloud, *Fuel Process. Technol.* **12** (1986) 143.
399. N. Ikenaga, H. Taniguchi, A. Watamabe, T. Suzuki, *Fuel* **79** (2000) 273.
400. P. A. Montano, A. S. Bommannavar, V. Shah, *Fuel* **60** (1981) 703.
401. P. A. Montano, P. P. Vaishnava, J. A. King, E. E. Eisentrout, *Fuel* **60** (1981) 712.
402. A. S. Bommannovar, P. A. Montano, *Fuel* **61** (1982) 523.
403. M. G. Thomas, T. D. Padrick, F. V. Stohl, H. P. Stephens, *Fuel* **61** (1982) 761.
404. J. M. Lambert, Jr., *Fuel* **61** (1982) 777.
405. T. Kotanigawa, M. Yamamoto, M. Sasaki, N. Wang, H. Nagaishi, T. Yoshida, *Energy Fuels* **11** (1997) 190.
406. P. A. Montano, B. Granoff, *Fuel* **59** (1980) 214.
407. G. Langhoff, E. Wolowski, U. Graese: *Proc. 8th Annual Electric Power Research Institute (EPRI) Contractors' Conference on Coal Liquefaction,* Palo Alto, May 1983, pp. 19–1 to 19–21.
408. B. O. Strobel, F. Friedrich, *Proc.-Int. Conf. Coal Sci.,* Amsterdam 1987, p. 395.
409. K. Hirano, N. Suzuki, Y. Miyake, M. Kouzu, S. Ueda, M. Kobayashi, *J. Jpn. Inst. Energy* **78** (1999) 42.
410. K. Hirano, M. Kouzu, T. Okada, S. Ueda, M. Kobayashi, O. Tokari, *J. Jpn. Inst. Energy* **75** (1996) 909.
411. K. Ikeda, K. Sakawaki, Y. Nogami, K. Inokuchi, K. Imada, *Fuel* **79** (2000) 373.
412. D. Gray, M. G. Neuworth, G. C. Tomlinson:"The Potential Impact of Mineral Matter Reduction and Disposable Catalyst Addition on a Single-Stage, Slurry Phase Coal Liquefaction Process,"working paper 81W00427, Mitre Corp., McLean, July 1981.
413. P. G. Stansberry, J. P. Wann, W. R. Stewart, J. Yang, J. W. Zondlo, A. H. Stiller, D. B. Dadyburjor, *Fuel* **72** (1993) 793.
414. T. Kotanigawa, S. Yokoyama, M. Yamamoto, A. Maekawa, *Fuel* **68** (1989) 618.
415. T. Kaneko, K. Tazawa, N. Okuyama, M. Tamura, K. Shimasaki, *Fuel* **79** (2000) 263.
416. T. Kaneko, K. Tazawa, S. Yanai, M. Tamura, K. Shimasaki, *Proc.-Int. Conf. Coal Sci., Taiyuan* 1999, p. 625.
417. S. W. Weller: "Coal Liquefaction with Molybdenum Catalysts,"*4th International Conference on Chemistry and Uses of Molybdenum,* Golden, Colo., 1982.

418. C. O. Hawk, R. W. Hiteshue:"Hydrogenation of Coal in the Batch Autoclave," U.S. Dept. of the Interior, Bureau of Mines, Bulletin 622, 1965.
419. S. Weller, M. G. Pelipetz, S. Friedman, H. H. Storch, *Ind. Eng. Chem.* **42** (1950) 330.
420. S. Weller, M. G. Pelipetz: "Catalysis in Liquid-Phase Coal Hydrogenation," *Proc. 3rd World Petroleum Congress,* Leiden 1951.
421. S. Weller, M. G. Pelipetz, *Ind. Eng. Chem.* **43** (1951) 1243.
422. S. Weller in P. H. Emmett (ed.): *Catalysis,* **vol. 4,** Rheinhold Publ., New York 1956, pp. 513 – 528.
423. P. H. Given, D. C. Cronauer, W. Spackman, H. L. Lovell et al., *Fuel* **54** (1975) 34.
424. H. H. Ginsberg, S. Friedman, P. S. Lewis, M. D. Schlesinger et al.: "Hydrogenating Coal in a Pilot Plant with a Molybdenum Catalyst," U.S. Department of the Interior, Bureau of Mines Report of Investigations 5673, 1960.
425. M. D. Schlesinger, L. V. Frank, R. W. Hiteshue: "Relative Activity of Impregnated and Mixed Molybdenum Catalysts for Coal Hydrogenation," U.S. Department of the Interior, Bureau of Mines, Report of Investigations 6021, 1962.
426. W. Kawa, S. Friedman, W. R. K. Wu, L. V. Frank et al.: "Evaluation of Catalysts for Hydrodesulfurization and Liquefaction of Coal," *167th National Meeting of Amer. Chem. Soc.,* Los Angeles 1974.
427. N. G. Moll, V. T. May, G. J. Quarderer, G. A. Stuart: "Dow Coal Liquefaction Process Update," *Conference on Synthetic Fuels Status and Directions,* San Francisco, 1980.
428. F. J. Derbyshire, A. Davies, R. Lin, P. G. Stansberry, M. T. Terrer, *Fuel Process. Technol.* **12** (1986) 127.
429. C. Song, A. K. Saini, Y. Yoneyama, *Fuel* **79** (2000) 249.
430. C. Song, A. K. Saini, *Energy Fuels* **9** (1995) 188.
431. A. B. Garcia, H. H. Schobert, *Fuel* **68** (1989) 1613.
432. A. B. Garcia, H. H. Schobert, *Fuel Process. Technol.* **26** (1990) 99.
433. M. Nomura, H. Sakashita, M. Miyake, S. Kikkawa, *Fuel* **62** (1983) 73.
434. S. Yokoyama, D. M. Bodily, W. H. Wiser, *Fuel* **62** (1983) 4.
435. S. S. Salim, A. T. Bell, *Fuel* **63** (1984) 469.
436. D. P. Mobley, A. T. Bell, *J. Catal.* **64** (1980) 494.
437. M. Mizumoto, H. Yamashita, S. Matsuda, *Ind. Eng. Chem. Prod. Res. Dev.* **24** (1985) 394.
438. C. R. Green, F. E. Biasca, *Coal Process. Technol.* **6** (1980) 102.
439. D. Fraenkel, V. R. Pradhan, J. W. Tierney, I. Wender, *Fuel* **70** (1991) 64.
440. Chevron Research Co., DE 2 756 976A1, Jun. 29, 1978.
441. G. J. Stiegel, R. E. Tischer, D. L. Cillo, N. K. Narain, *Ind. Eng. Chem. Prod. Res. Dev.* **24** (1985) 206.
442. F. V. Stohl, H. P. Stephens, *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 2466.
443. F. V. Stohl, Q. A. Quader, F. E. Maasoth, D. S. Thakur, *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 840.
444. C. C. Kang, E. S. Johanson, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **21** (1976) no. 5, 32.
445. D. S. Thakur, M. G. Thomas, *Ind. Eng. Chem. Prod. Res. Dev.* **23** (1984) 349.
446. T. Zhang, P. D. Jacobs, H. W. Haynes, A. J. Swanson, *Fuel* **74** (1995) 431.
447. D. S. Thakur, M. G. Thomas, *Appl. Catal.* **15** (1985) 197.
448. I. Mochida, X. Z. Zho, K. Sakanishi, *Fuel* **67** (1988) 1101.
449. Y. Yoshimura, S. Endo, S. Yoshitomi, T. Sato, H. Shimada, N. Matsubayashi, A. Nishijima, *Fuel* **70** (1991) 733.
450. H. Shimada, T. Sato, Y. Yoshimura, M. Kubota, A. Nishijima, *Sekiyu Gakkaishi* **30** (1987) 258.
451. F. V. Stohl, H. P. Stephens, G. J. Stiegel, R. E. Tischer, *9th Electric Power Research Inst. Contractors' Meeting on Coal Liquefaction,* Palo Alto, May 1984.
452. G. J. Stiegel, L. M. Polinski, R. E. Tischer, *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 477.
453. G. J. Stiegel, R. E. Tischer, L. M. Polinski, *Ind. Eng. Chem. Prod. Res. Dev.* **22** (1983) 411.
454. F. V. Stohl, H. P. Stephens: "Deactivation of Direct Coal Liquefaction Catalysts," *Proc. DOE Direct Coal Liquefaction Contractors' Review Meeting,* Albuquerque October 17 – 18, 1984.
455. M. G. Thomas, D. G. Sample:"Coal Liquefaction Catalyst Characterization," Sandia Laboratories Energy, report no. SAND-79–0085, May, 1979.
456. E. Furimsky, *Fuel Process. Technol.* **6** (1982) 1.
457. S. Parkash: "Studies on Coal Liquefaction Catalysts," Information Series 89, Alberta Research Council, Alberta 1980.

458. R. J. Bertolacini, L. C. Gutberlet, D. K. Kim, K. K. Robinson:"Catalyst Development for Coal Liquefaction,"Electric Power Research Inst. (EPRI), report AF-574, 1977.
459. R. J. Bertolacini, L. C. Gutberlet, D. K. Kim, K. K. Robinson, Electric Power Research Inst. (EPRI), report AF–1084, 1979.
460. D. C. Green, D. H. Broderick, *Chem. Eng. Prog.* **1981,** (Dec.) 33.
461. T. Masuyama, Y. Kageyama, S. Kawai, *Fuel* **69** (1990) 245.
462. T. Kaneko, Y. Kageyama, T. Masunaga, *Kagaku Kogaku Ronbunshu* **21** (1995) 1110.
463. T. Kaneko, Y. Kageyama, Y. Kitamura, O. Okuma, T. Matsumura, *J. Jpn. Inst. Energy* **72** (1993) 970.
464. T. Kimura, T. Kaneko, Y. Kageyama, S. Kawai, *Sekiyu Gakkaishi* **33** (1990) 311.
465. Y. Kageyama, T. Masuyama, *Proc.-Int. Conf. Coal Sci.,* Sydney 1985, p. 157.
466. H. Shimada, T. Sato, Y. Yoshimura, A. Nishijima, M. Matsuda, T. Konakahara, K. Sato, *Sekiyu Gakkaishi* **31** (1988) 227.
467. A. Papaioannou, H. W. Haynes, *Energy Fuels* **4** (1990) 38.
468. S. M. Richardson, M. R. Gray, *Energy Fuels* **11** (1997) 1119.
469. H. P. Stephens, C. Bolton, C. E. Snape, *Proc.-Int. Conf. Coal Sci.,* Maastricht 1987, p. 703.
470. H. P. Stephens, R. G. Dosch, F. V. Stohl, *Ind. Eng. Chem. Prod. Res. Dev.* **24** (1985) 15.
471. R. G. Dosch, F. V. Stohl, H. P. Stephens, "Development of Hydrous Metal Oxide-Supported Catalysis for Direct Coal Liquefaction", *Proceedings of the 1990 Direct Liquefaction Contractors' Review Meeting,* U.S. Department of Energy, PETC, Pittsburgh, September 1990, p. 56.
472. R. G. Dosch, F. V. Stohl, J. T. Richardson, "Novel Materials in Heterogeneous Catalysis", *ACS Symp. Ser. 437* 1990, p. 279.
473. A. W. Scaroni, R. G. Jenkins, P. L. Walker, *Appl. Catal.* **14** (1985) 173.
474. V. H. Beer, F. J. Derbyshire, C. K. Groot, R. Prins, A. W. Scaroni, J. M. Solar, *Fuel* **63** (1984) 1095.
475. M. Farcasiu, C. Smith, *Energy Fuels* **5** (1991) 83.
476. I. Mochida, K. Sakanishi, R. Sakata, K. Honda, T. Umezawa, *Energy Fuels* **8** (1994) 25.
477. K. Sakanishi, H. Hasuo, I. Mochida, O. Okuma, *Energy Fuels* **9** (1995) 995.
478. K. Sakanishi, H. Hasuo, M. Kishino, I. Mochida, O. Okuma, *Energy Fuels* **10** (1996) 21.
479. K. Sakanishi, H. Hasuo, M. Kishino, I. Mochida, *Energy Fuels* **12** (1998) 284.
480. K. Sakanishi, H. Taniguchi, H. Hasuo, I. Mochida, *Ind. Eng. Chem. Res.* **36** (1997) 306.
481. I. Mochida, K. Sakanishi, *Fuel* **79** (2000) 221.
482. A. G. Comolli, T. K. Lee, G. A. Popper, R. H. Stalzer, P. Zhou, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **43** (1998) 441.
483. N. Okuyama, N. Komatsu, S. Yanai, T. Kaneko, M. Tamura, K. Shimasaki, *Proc.-Int. Conf. Coal Sci., Taiyuan* 1999, p. 629.
484. G. A. Olah, M. R. Bruce, E. H. Edelson, A. Husain, *Fuel* **63** (1984) 1130.
485. N. D. Taylor, A. T. Bell, *Fuel* **59** (1980) 49.
486. Shell Oil Co., US 3 844 928, 1974 (D. O. Geymer).
487. Shell Oil Co., US 3 677 932, 1974 (D. E. Hardesty, T. A. Rodgers).
488. IEA Technology Status Report 010: "Coal Liquefaction", DTI/Pub URN 99/1120, October 1999.
489. A. E. Sawy, M. B. Neuworth: "Critical Evaluation of Solvent Deashing of Coal Liquids," *C. E. Lummus Antisolvent Deashing Process,* **vol. 1, MTR–83W42–01,** Mitre Corp., McLean 1983.
490. A. Sawy, D. Gray, M. G. Neuworth, G. Tomlinson: "Critical Evaluation of Solvent Deashing of Coal Liquids," *Kerr–McGee Critical Solvent Deashing Process,* **vol. 2, MTR–83W42,** Mitre Corp., McLean, November 1983.
491. O. Okuma, K. Masuda, N. Okuyama, T. Hirano, *Fuel Process. Technol.* **56** (1998) 229.
492. O. Okuma, N. Okuyama, T. Hirano, K. Masuda, *Fuel Process. Technol.* **60** (1999) 119.
493. M. Kouzu, K. Koyama, M. Oneyama, T. Aramaki, T. Hayashi, H. Itoh, H. Hattori, *Fuel* **79** (2000) 365.
494. G. Tomlinson, D. Gray, M. Neuworth: "Effect of Coal Rank on Direct Coal Liquefaction Processes: SRC II Process Experience," WP–84W00308, Mitre Corp., McLean, July 31, 1984.
495. G. C. Tomlinson, D. Gray, M. B. Neuworth: "The Impact of Pilot Plant Operations and Concurrent Development Efforts on the Projected Performance and Economics of a Commercial EDS Facility, MTR–83W235, Mitre Corp., February 1984.

496. L. L. Ansell, J. W. Taunton, K. L. Trachte: "Bottom Recycle Studies of EDS Process Development," *Proceedings of the 5th Electric Power Research Institute Annual (EPRI) Contractors' Conference on Coal Liquefaction,* Palo Alto, May 7 – 8, 1980, pp. 1–1 to 1–20.
497. A. Talib, D. Gray, M. Neuworth: "Assessment of H-Coal Process Developments," MTR–83W199, Mitre Corp., McLean, January 1984.
498. A. G. Comolli, J. B. MacArthur, J. B. McLean: *Proc. DOE Direct Coal Liquefaction Contractors' Review Meeting,* Albuquerque, Oct. 17 – 18, 1984, pp. 60 – 71.
499. G. O. Davies, *Prepr. - Am. Chem. Soc., Div. Fuel Chem.* **26** (1981) no. 1, 67.
500. K. E. Cowser, C. R. Richmond: *Synthetic Fossil Fuel Technology,* Ann Arbor Science, Mich. 1980.
501. B. W. Wilson, D. W. Later: "The Chevron Co-Refining Process," *Proc. 9th Annual Electric Power Research Inst. (EPRI) Contractors' Conference on Coal Liquefaction,* Palo Alto, May 8 – 10, 1984, pp. 22–1 to 22–18.
502. B. W. Wilson, D. L. Stewart, D. D. Mahlum, C. W. Wright et al.: "Chemical Basis for Toxicologic Activity in Coal Liquids: A Comparison of Direct Liquefaction Materials," *Proc. DOE Direct Coal Liquefaction Contractors' Review Meeting,* Albuquerque, Oct. 17 – 18, 1984, pp. 328 – 350.
503. M. R. Guerin, W. H. Griest, L. H. Smith et al.: "Comparative Toxicological and Chemical Properties of Fuels Developed from Coal, Shale, or Petroleum: Coal Liquids Update," *Proc. DOE Direct Coal Liquefaction Contractors' Review Meeting,* Albuquerque, October 17 – 18, 1984, pp. 306 – 327.
504. S. W. Gouse, D. Gray, G. C. Tomlinson, D. L. Morrison, "Potential World Development Through 2100: The Impacts on Energy Demand, Resources, and the Environment", World Energy Council Journal, December 1992.

Coal Pyrolysis

Coal, Coal Liquefaction, and Gas Production are separate keywords

Alan W. Scaroni, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

M. Rashid Khan USDOE, Morgantown Energy Technology Center, Morgantown, West Virginia 26505, United States

Semih Eser, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

Ljubisa R. Radovic, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

1. Introduction 2
2. Thermoplastic Properties of Coal . 2
2.1. Experimental Techniques 4
2.2. Experimental Variables 4
2.2.1. Coal Rank and Petrographic Composition 4
2.2.2. Heating Rate 5
2.2.3. Particle Size 6
2.2.4. Pressure 6
2.2.5. Gaseous Atmosphere 7
2.2.6. Preoxidation 7
2.2.7. Additives 8
2.3. Development of Thermoplasticity . 9
2.3.1. Softening 9
2.3.2. Swelling 10
2.3.3. Resolidification 11
2.4. Coke Microstructure and Mesophase Formation 11
2.4.1. Microscopic Characterization 11
2.4.2. Carbonaceous Mesophase 13
2.4.3. Modification of Optical Texture and Properties 17
3. Yield and Distribution of Pyrolysis Products 17
3.1. Experimental Methods 17
3.2. Experimental Variables 18
3.2.1. Time and Temperature 18
3.2.2. Heating Rate 20
3.2.3. Pressure 21
3.2.4. Particle Size 21
3.2.5. Coal Rank 23
3.3. Mechanism 23
4. Kinetics 24
5. Hydropyrolysis 26
6. Pyrolysis Processes 28
6.1. Low-Temperature Pyrolysis Processes 29
6.1.1. Early Processes 29
6.1.2. Contemporary Processes 30
6.2. High-Temperature Processes (Coke-Making) 31
6.2.1. Coke Ovens 31
6.2.2. Coke Properties 32
6.2.3. Byproducts 32
6.2.4. Formed-Coke Processes 33
6.2.5. Environmental Aspects 33
6.3. Contemporary Hydropyrolysis and Pyrolysis Processes 34
7. References 37

1. Introduction

Coal pyrolysis is the heating of coal to produce gases, liquids, and a solid residue (char or coke). Pyrolysis occurs in all coal utilization processes, i.e., combustion, gasification, liquefaction, and carbonization. The nature of pyrolysis and of the products is intimately related to the operating conditions and to the composition and properties of the coal. Consequently, control of pyrolysis is important in coal utilization processes.

The thermal decomposition (or devolatilization) of coal is illustrated using Wiser's model of coal structure (see Fig. 1) [1]. This model, one of many in the literature, represents the types of structures found in coals. Although it does not necessarily depict the structure of a particular coal, it is consistent with knowledge of coal physics and chemistry.

Pyrolytic rupture of functional groups attached to aromatic and hydroaromatic units of the coal structure leads to the formation of gases

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_245

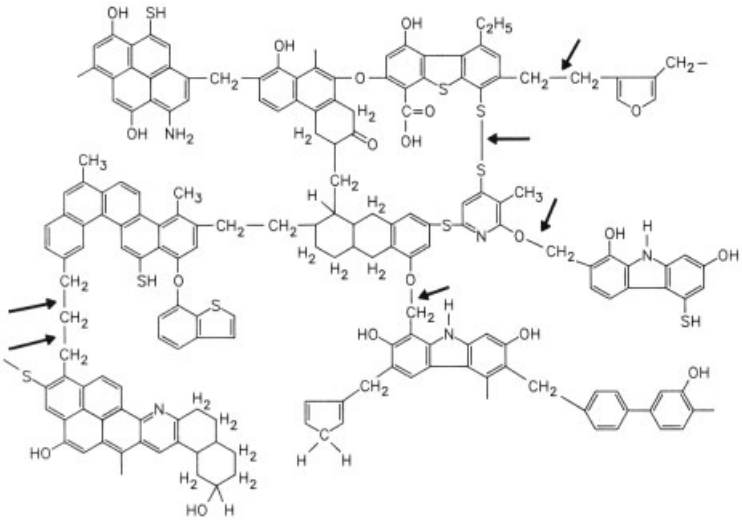

Figure 1. WISER's model of coal structure [1]

(CO, CO_2, H_2O, CH_4, C_2H_4, etc.). In addition the cross-links (indicated by arrows in Fig. 1) break and release reactive free radicals (fragments). The fate of these radicals controls the overall yield and distribution of products. Stabilization of the fragments with hydrogen gives primary volatile products and favors a high yield of gases and liquids (see Fig. 2). Some primary products are also obtained by release of low molecular mass species (the so-called mobile phase), which are believed to be trapped within the coal network [2]. However, if the fragments undergo secondary cracking, polymerization, or condensation reactions, part of the primary product is converted to char.

The four fundamental aspects of this complex process are the following:

1) The thermoplastic behavior of some coals (Chap. 2),
2) The yield and distribution of pyrolysis products (Chap. 3),
3) The temperature and time dependency (kinetics) of the pyrolysis process (Chap. 4), and
4) The source of hydrogen for the stabilization of reactive fragments (Chap. 5).

2. Thermoplastic Properties of Coal

When bituminous coal is heated, it undergoes marked chemical and physical changes, which are influenced by rank, petrographic composition, particle size, heating rate, temperature, and the atmosphere in which pyrolysis occurs. Bituminous coal softens and melts at 350 – 400 °C. In contrast, lignite, subbituminous coal, and anthracite do not soften when heated. Fusion and agglomeration may produce a fluid mass. Concurrently, thermal decomposition may liberate gases and vapors, producing a foamlike material with significant volume change. At some point, the foam solidifies.

These macroscopic changes are accompanied by microstructural rearrangements within the fluid phase, which govern the properties of the resulting coke. Development of the coke microstructure often involves formation of an intermediate phase called carbonaceous mesophase (see Section 2.4).

Volume changes (expansion and contraction) occur in some bituminous coals on heating. Initial softening at temperature T_s produces a fluid. As softening increases, pores and voids are eliminated and the material undergoes volume contraction, reaching a minimum V^c at temperature T^c. At this point, rapid gas evolution leads to expansion, producing a maximum V_s at temperature T_e. With a further increase in temperature, the fluid coal resolidifies at temperature T_r; the corresponding volume is V_r. These phenomena, i.e., softening, swelling, and resolidification are collectively referred to as the *thermoplastic*

properties of a coal, otherwise called the softening, plastic, fluid, agglomerating, caking, and coking properties of coal.

The terms "caking" and "coking" are sometimes used incorrectly as synonyms. In a strict sense, *caking* describes softening and agglomeration of coal on heating, without any implications about the nature of the resulting solid. *Coking coals,* on the other hand, are those that form strong, coherent solids (metallurgical cokes) on softening and resolidifying in conventional coke ovens. In other words, all coking coals exhibit caking behavior, but not all caking coals can be coked. Coal that does not become fluid when heated is generally termed nonswelling or noncaking.

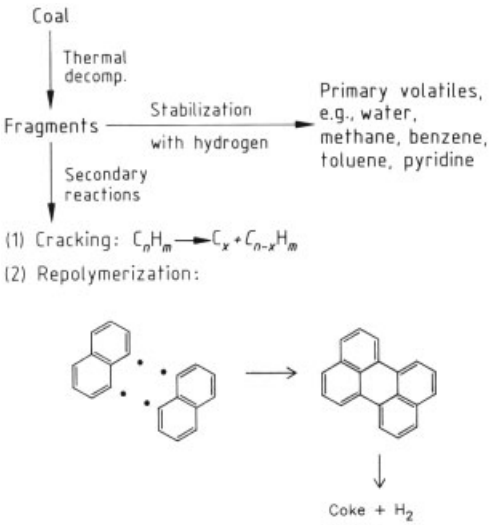

Figure 2. Primary and secondary reactions in coal pyrolysis

Continued heating of a *semicoke,* i.e., a solid residue still containing appreciable volatile matter, to ca. 1000 °C leads to further decomposition with formation of a solid residue termed *coke.* Coke contains little or no volatile matter and undergoes little further volume change when heated. The volume changes occurring during the transformation of a coal into a semicoke depend on the fluidity of the plastic mass and its resistance to the escape of gases and vapors.

An understanding of the thermoplastic behavior of coal is important for nearly all coal conversion and coke-making processes. For example, plasticity development and agglomeration may create severe operating difficulties in gasifiers.

In a *moving-bed gasifier*, as the coal proceeds down the (sometimes pressurized) reactor, it devolatilizes at relatively low heating rates (25 – 65 °C/min) in the presence of a reducing atmosphere. Under these conditions, bituminous coal undergoes, in varying degrees, thermoplastic transformations that can lead to the formation of coke masses. These are detrimental to the flow of gases and solids, resulting in poor operation of the gasifier. In the extreme, coke agglomerates can completely plug the reactor.

In a *fluidized-bed gasifier*, an agglomerated mass may disrupt the fluidization. As a complicating factor, highly swollen coke may greatly reduce the bed density and lower the throughput. As a further complication, a highly swollen coke mass tends to disintegrate when it is fluidized, producing excessive fines that are carried from the bed and must be separated from the exhaust gas.

In contrast to the general objectives in gasification and combustion, extensive plasticity development is essential during carbonization to produce metallurgical coke. The main objective of the coking industry is to produce coke for blast furnace operation. In this context, research focused on the identification of coke microconstituents and the mechanism of their formation. The improvement of coke microstructure depends largely on the ability to modify the thermoplastic properties of the coal feedstocks and the associated mesophase. This is of particular significance because of the interest in the plastic behavior of weakly caking coal as a supplement to the reserves of prime coking coal.

2.1. Experimental Techniques

The industrial importance of the thermoplastic properties of coal is reflected in the many methods available for its characterization [3]. A common objective of these techniques is to predict the behavior of coal in various processes.

The *free-swelling index* (Gray – King) provides a visual comparison of cokes carbonized under well-defined conditions. The method is inexpensive, rapid, and reproducible, but lacks flexibility. *Hot-stage optical microscopy* permits the visual observation of morphological changes caused by heat. The *Foxwell gas flow method*

measures the resistance to the flow of a highly purified gas through a bed of coal during continuous heating; the pressure drop indirectly measures the fluidity of the plastic mass. The *constant torque plastometer* records the velocity, as a function of temperature, of a rotating shaft passing through a bed of coal. In the variable torque plastometer, the velocity is kept constant and the required torque measured as a function of temperature. The resulting measurements of fluidity provide a sensitive index of oxidation or weathering, but not of swelling. The heating rate is limited to ca. 3 °C/min. In the *dilatometer,* the volume change on solidification is measured as a function of temperature. This is applicable over a wide range of heating rates and provides data on dilation parameters and transition temperatures, but not on fluidity changes.

These empirical tests require that experimental conditions be specified when results are reported. Furthermore, these tests measure different aspects of thermoplastic behavior; therefore, there is no single best method or technique that can predict the changes which occur in all industrial processes. Selection of a test method depends on the utilization process.

2.2. Experimental Variables

The thermoplasticity of a coal is a complicated phenomenon. It is affected by softening and rheological properties and by formation of gas bubbles [7], which are strongly influenced by pyrolysis conditions and feedstocks.

2.2.1. Coal Rank and Petrographic Composition

The thermoplastic properties of coal vary widely as a function of rank. Moreover, the plastic properties of some bituminous coals serve to differentiate them from others of similar rank. A number of attempts have been made to *correlate plastic properties with rank.* Swelling, measured by dilatometry, exhibits a maximum for a volatile-matter content of ca. 25 – 28 wt %; the Gieseler fluidity reaches a maximum at ca. 28 – 32 wt % of volatile matter [5]. Correlation of the free-swelling index (FSI) with the maximum Gieseler fluidity is poor, which is not surprising; rather, it illustrates the heterogeneity of coal [4]. Some variations can be expected because of the obvious differences in test conditions. In the FSI test, a 1-g sample of coal (250 µm particle size or 60 mesh) is heated in a silica crucible at 820 ± 2 °C for 2.5 min; the swelling is compared to that of a standard. The Gieseler fluidity test measures the rotary speed (in dial division per minute, ddpm) of a stirrer under a constant torque in a compacted 5-g charge, heated at 3 ± 0.1 °C/min from 300 °C until it resolidifies. However, the FSI does not increase monotonically with increasing Gieseler fluidity. In any case, the data support the concept of an optimum fluidity that is essential for maximum swelling.

The principal maceral responsible for coal plasticity is *vitrinite,* the predominant petrographic constituent of most United States coals. *Exinite* differs from vitrinite of the same rank by containing more hydrogen and volatile matter. Exinite exhibits a lower softening temperature, wider plastic range, greater Gieseler fluidity, and higher dilatometric swelling than vitrinites (see Fig. 3) [3].

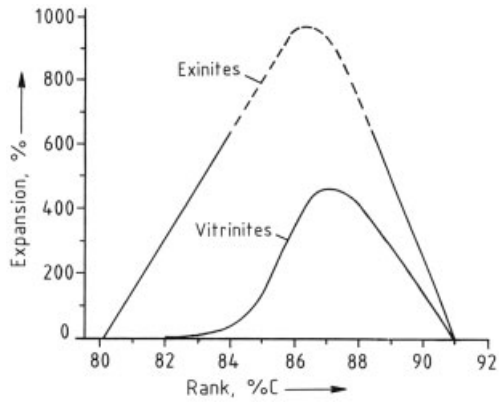

Figure 3. Expansion of vitrinites and exinites [3]

Inertinite is infusible and inert, regardless of coal rank. It is distinguished from vitrinite of the same rank by its higher density and lower hydrogen and volatile matter contents [5].

Plasticity usually reflects the properties of the constituent macerals, but it also depends on the dispersion of exinite and inertinite within the vitrinite. This dispersion determines the degree of interaction between the different macerals. Consequently, predicting the thermoplastic behavior of a coal from its petrographic (maceral) analysis alone is impossible.

2.2.2. Heating Rate

The thermoplasticity of coals increases as the heating rate is increased [6]. Shock heating rates (above 100 °C/min) significantly increase the maximum swelling [6]. However, beyond a limiting heating rate (100 – 300 °C/min) a further increase does not increase swelling. On the other hand, at a low heating rate (below 3 °C/min), even a highly plastic coal may not soften or swell.

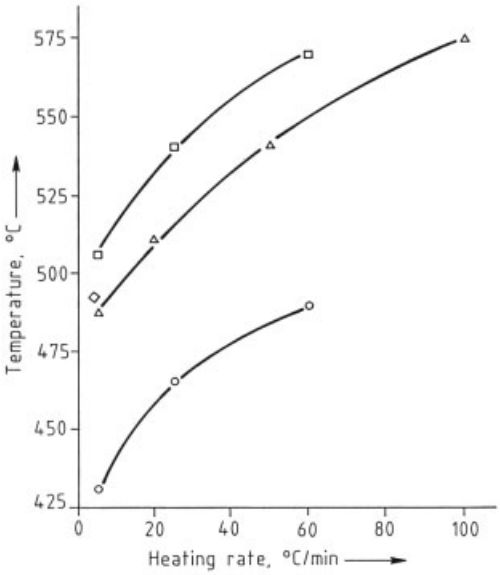

Figure 4. Effect of heating rate on the characteristic temperatures of a low-volatile bituminous coal (particle size: – 74 µm; pressure: 101.3 kPa of He) [4]
○ Softening temperature; △ Temperature of maximum devolatilization; ◇ Temperature of maximum fluidity [7]; □ Temperature of maximum swelling (dilatometry)

The effect of heating rate can be explained in terms of a competition between evolution of volatile matter and thermal decomposition. At a higher heating rate, reactive components of the volatile matter, which would otherwise be lost before plasticity develops, can participate in secondary reactions that promote fluidity.

With an increase in heating rate, the characteristic temperatures (for softening, maximum swelling, and maximum fluidity) increase, as measured by plastometry or dilatometry (see Fig. 4). Because these temperatures do not shift to the same extent, the temperature range over which the coal is plastic is generally widened.

An important effect of heating rate on the thermoplastic properties of coals is related to the maximum rate of devolatilization [3]. As the heating rate increases from ca. 1 to ca. 7 °C/min, the maximum swelling, maximum Gieseler fluidity, and maximum devolatilization rate all increase [4]. An increase in the heating rate produces a linear increase in the maximum rate of devolatilization. In addition, the maximum swelling increases with increasing heating rate, although this parameter is dependent on particle size. The maximum rate of devolatilization is slightly lower for the larger fraction.

The sequence of events occurring during coal pyrolysis is a function of heating rate. Softening is preceded by a rapid increase in evolution of volatile matter. Within the same temperature range, the viscosity of the sample reaches a minimum [7]. In response to this rapid rise in formation of volatile substances the coal particles and coal bed expand, producing the maximum swelling, determined by dilatometry, and the maximum rate of devolatilization, determined by thermogravimetric analysis [4].

2.2.3. Particle Size

The effect of particle-size variations on the thermoplastic properties of coal has not been widely reported. The data published before 1980 conflict. A few investigations have suggested that extremely fine grinding of the coal particles reduces thermoplastic properties. However, these results are possibly due to increased ease of weathering and oxidation of the more finely ground coal, or to maceral and mineral segregation that occurs during grinding.

Experimental data on particle size must be interpreted with caution. Studies conducted in a Hoffmann dilatometer, which accommodates coarse samples, have shown that the effect of particle size varies according to coal type and petrographic constitutents [5].

The effect of particle size is dependent on heating rate [4]. At a high rate (60 °C/min), a considerable decrease in the particle size resulted in a significant increase in swelling. This confirmed earlier findings that, at a higher heating rate, the maximum swelling was greatest for the smallest size fraction [8], [9]. The behavior at 60 °C/min has been explained as follows:

the smaller coal particles, with their greater surface : volume ratio, fuse more easily with the neighboring particles and produce a homogeneous mixed phase [4]. In contrast, the larger coal particles tend to maintain their individuality [4], [6].

With slower heating (5 °C/min), the swelling parameter for the smaller fraction (– 74 µm) was less than that at 60 °C/min. This was attributed to slower devolatilization at 5 °C/min, resulting in a lower swelling pressure, i.e., the pressure generated inside the particles. In addition, the diffusion distance (the distance traveled by the volatile substances before escaping) is shorter for smaller particles. Therefore, an increase in particle size with slower heating results in greater swelling.

A particle-size effect is also noted in the contraction parameter. In general, the initial contraction of the coal particles is significantly greater for the smallest fraction [4], [8], [9]. This is explained by the suggestion that larger coal particles fuse less than smaller ones because of their lower surface : volume ratios.

2.2.4. Pressure

Until recently, little work had been published on the effects of inert gas pressure on the thermoplastic properties of coal. In the absence of data at elevated pressure, the behavior of coal under typical gasification conditions (elevated pressure and reducing atmosphere) has generally been extrapolated from data obtained at atmospheric pressure. However, this treatment has serious deficiencies.

Several investigations have shown that pressure increases the coke and gas yields at the expense of tar production. This has been attributed to an increased residence time of volatile and plasticizing substances in the coal melt and to a corresponding effect on rheological behavior. Pyrolysis of bituminous coal at elevated pressure has also been shown to increase the Gieseler fluidity (see Fig. 5) [10].

Swelling of coal in dilatometers results from a balance between opposing forces and occurs when the internal (intraparticle or intrabed) gas pressure (swelling pressure) generated in the plastic state is higher than the applied forces that tend to contain it. This internal pressure depends on the coal type, particle size, heating rate, and temperature. Consequently, at elevated pressure, coal swelling depends largely on experimental conditions. For a range of bituminous coals, the rank dependency of the effect of pressure on the maximum swelling parameters was highly variable [4]. In addition, the extent of maximum swelling varied markedly.

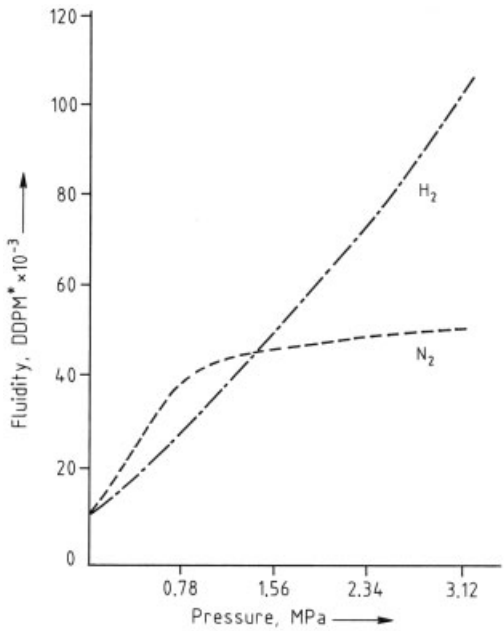

Figure 5. Effect of pressure on the Gieseler fluidity of a bituminous coal [10]
∗ DDPM = dial divisions per minute

Other observed effects of pressure, irrespective of coal rank, were an increased contraction parameter and an expanded plastic range. In contrast, with vacuum pyrolysis, maximum swelling, contraction, and plastic range were significantly reduced [4].

The effects of pressure can be explained qualitatively by invoking the concept of optimum fluidity, which is essential for high expansion. If the melted coal is very fluid, swelling is not extensive because the volatile substances escape easily. At the other extreme, minimum fluidity is essential for the material to fuse and entrap volatile material. This implies that an optimum fluidity exists, which is necessary to permit a high degree of swelling. Under vacuum, for example, volatile and plasticizing components are readily removed before sufficient fluidity or swelling

pressure can develop. Hence, thermoplastic behavior is inhibited [4].

Production of volatile matter during pyrolysis is a necessary but insufficient condition for a coal to become plastic and swell. The combined effects of the generated volatile matter and the rheological properties of coal determine thermoplastic behavior [4], [9], [11].

2.2.5. Gaseous Atmosphere

The effects of a gaseous atmosphere on the thermoplastic properties of coal are of interest because in many processes devolatilization occurs in the presence of reactive gases, e.g., H_2, CO/H_2, or CO_2.

The effects of H_2 and He depend on the conditions. However, an increase in the maximum swelling parameter can be correlated with the degree of hydrogenation. At heating rates above 60 °C/min in a high-pressure microdilatometer, the thermoplastic properties of a low-volatile bituminous coal were not significantly different in atmospheres of H_2 or He [4]. The total loss of mass during pyrolysis at ca. 630 °C in these two gases was similar, indicating little hydrogenation at high heating rates.

In contrast, at a low heating rate of 5 °C/min, the maximum swelling is significantly greater at elevated H_2 pressure compared to He. The loss of mass during pyrolysis is also significantly greater in H_2 than in He, indicating some hydrogenation at this heating rate [4].

The maximum Gieseler fluidity (heating rate, 5 °C/min) of a weakly swelling coal (FSI = 0.5) at atmospheric pressure was markedly greater in H_2 than in He or N_2 [10], [12].

2.2.6. Preoxidation

In many processes, thermoplastic behavior of bituminous coals can have undesirable consequences. Caking can be reduced or eliminated by low-temperature (below the softening point) preoxidation, which can destroy thermoplastic properties [5], [6], [13], [14]. In addition, carbonization of preoxidized bituminous coal produces a char with a higher gasification reactivity and surface area than coke produced from untreated coal [15].

Although many studies have been conducted on oxidized or weathered coals, oxidation mechanisms are still poorly understood. Small amounts of oxygen (0.01 – 0.02 wt %) can produce substantial changes in coal plasticity [16]. The first stage of oxidation was postulated as peroxide formation; however, this has not been confirmed experimentally. It was further suggested that the peroxide is unstable above 70 °C and decomposes to yield more stable coal – oxygen complexes with evolution of CO_2, CO, and H_2O.

Another proposed mechanism was that oxidation increases hydroxyl, carboxyl, and carbonyl content [3]. As the temperature rises the hydroxyl groups can condense with the liberation of water:

$$\mathrm{Ar-OH + Ar-OH \longrightarrow Ar-O-Ar + H_2O}$$

where Ar is an aromatic radical.

Fourier transform infrared (FTIR) studies suggest that in the early stages of oxidation, the main functional groups formed are carbonyl and carboxyl [17]. At higher degrees of oxidation, evidence indicates a significant increase in ether, ester, and phenolic groups. Loss of plastic properties on oxidation is attributed to the formation of ether and ester cross-links. The decline in fluid behavior on oxidation can be correlated to a reduction in the aliphatic hydrogen content [18].

The swelling of a highly fluid coal increased on mild oxidation [5], [6], [8]. More severe oxidation sharply reduced swelling.

Little work has been done on carbonization of preoxidized coal under the pressure of a reducing gas. Successful operation of fixed- and fluidized-bed processes depends on the behavior of preoxidized coal when it is devolatilized under these conditions.

A preoxidized coal shows a higher loss of mass during pyrolysis in H_2 than in N_2 [4]. The thermoplastic properties of preoxidized coal are highly dependent on pressure, and maximum swelling increases significantly at an elevated pressure of H_2 or He [4], [19], [20]. This suggests that if preoxidized coal is used as a feed material for a pressurized process, where devolatilization occurs in a reducing atmosphere, the effects of preoxidation are reduced.

Although oxidation is an inexpensive technique for reducing the plastic behavior of coal,

deleterious consequences have been reported [4]. Preoxidation results in an overall loss of carbon (as CO_2) and hydrogen (as H_2O); hence, the heating value is significantly reduced. Consequently, other means of reducing thermoplastic properties are being investigated.

2.2.7. Additives

Inorganic additives are an alternative to preoxidation. They reduce the caking of coal carbonized at atmospheric pressure (see Table 1); however, they have economic and operational drawbacks.

The mechanism of decaking is not known. In general, an additive may function as an inert diluent or in a more active role by chemically interacting with the softened coal. Some inert additives may provide sites for complex reactions of the coal. Some inert additives reduce fluidity by adsorbing primary decomposition products. Removal of 2 – 3 wt % of the extractable bitumen prevents caking completely [5].

Additives may also increase the permeability to gas flow and, therefore, reduce swelling in the coal melt. Compounds like SiO_2, however, serve simply as diluents.

Additives may also react chemically with the pyrolysis products. Substances, such as K_2CO_3, KOH, CaO, Fe_2O_3, Fe_3O_4, HCl, and BF_3, reduce the swelling of coal at atmospheric pressure.

The effectiveness of additives in decaking under gasification conditions has been examined [4], [22], [23]. Compounds of K, Ca, or Fe significantly reduced thermoplasticity under most conditions. Although K_2CO_3 and KOH are effective decaking agents, KCl is not. Similarly, CaO and $Ca(OCOCH_3)_2$ were more effective than $CaCO_3$. Both Fe_2O_3 and Fe_3O_4 destroyed thermoplasticity at 0.1 MPa (H_2 or He), even at ca. 5 wt %.

Effective additives may promote char formation during pyrolysis. Pyrolysis at elevated hydrogen pressure usually increases V_s (see 2). This was attributed to a reduction of char-forming reactions, thereby increasing fluidity of the system [4].

2.3. Development of Thermoplasticity

2.3.1. Softening

Although thermoplastic properties of coal have been studied for many years, softening still lacks a satisfactory explanation. However, a general pattern has been established.

Direct microscopic observations are revealing. Between 325 and 450 °C, depending on coal rank and experimental conditions, visible softening is manifested as follows [6]:

1) Coal particles become mobile and fill the bed space. Larger particles, which soften less than smaller ones, fuse with the smaller particles.
2) So-called fusion pores begin to form as softening continues, depending on particle size and nature of the coal. However, with further softening, the material becomes more homogeneous, making distinction between pores formed by fusion and those formed by devolatilization difficult. As softening continues, viscosity decreases, particle boundaries disappear, and gas evolution increases.
3) With a further increase in temperature, viscosity increases until resolidification occurs, pore volume and vitrinite reflectance increase, and in some cases, anisotropy increases.

Softening can be measured in a dilatometric test by the onset of contraction when a bed of coal is heated under a weighted plunger. A more precise measurement is given by the attainment of a specific degree of fluidity within the plastometer. In all theories of plasticity, it is acknowledged that a fusible material is formed during softening, but differences of opinion exist as to the origin, nature, and function of this material [9].

The homogeneous-melt theory, for example, proposes that coal undergoes pastelike melting throughout the entire mass, but decomposes at the same time, giving rise to gas and an infusible solid [24], [25]; the composition of this mixture depends on time and temperature. Resolidification is observed when the transformation to infusible material is complete. If decomposition proceeds too rapidly before the inception of softening, fusion does not occur and char is formed. On the other hand, if softening occurs before the

Table 1. Effects of additives on thermoplastic behavior of coal [21]

Additive	Concentration	Pretreatment	Caking index [a]	Changes in caking index	
				Before	After
Dry sand	20 wt %	dry mixed	FSI	4.5	4.1
NaOH	20 wt %	dry mixed	FSI	4.5	1.0
Na_2CO_3	20 wt %	dry mixed	FSI	4.5	1.8
Na_2CO_3	10 wt %	dry mixed	FSI	4.5	3.0
Na_2CO_3	5 wt %	dry mixed	dpm	275	15
K_2CO_3	20 wt %	dry mixed	FSI	7.0	1.5
Boric acid	2 wt %	dry mixed	BS swelling no.	8.0	0.5
NaCl	1 *N* solution	20 g of coal	FSI	4.5	2.0
Na_2CO_3	1 *N* solution	in 200 mL of solution	FSI	4.5	1.3
$NaHCO_3$	1 *N* solution	dried at	FSI	4.5	1.5
NaOH	1 *N* solution	105 °C for 40 min	FSI	4.5	1.0
BF_3	0.25 mmol	gravimetric	FSI	8.0	2.0
	BF_3/100 g	gas sorption	FSI	8.0	2.0
BF_3	0.27	at saturated	FSI	9.0	2.0
BF_3	0.21	pressure	FSI	7.0	2.5
$AlCl_3$, $(CH_3)_2CHCl$ [b]	–	–	coke inspection	agglomerated	powder
HCl	not stated	1 h at 150 °C	coke inspection	agglomerated	powder
Excess tri-*n*-butyl borate [c]	–	–	BS swelling no.	8.0	0
Phosphorylation [c]	–	–	BS swelling no.	8.0	0.5
Acetylation [c]	–	–	BS swelling no.	8.0	2.5
NH_3 solution with Ni catalyst	3 – 3.5 wt % Ni	autoclaved with liquid NH_3 at 120 °C for 4 h	swelling no.	2.2	1.1

[a] FSI = Free Swelling Index; dpm = Divisions per minute as determined by a Gieseler plastomer; BS = British Standard.
[b] Friedel – Crafts alkylation.
[c] Esterification of phenolic hydroxyl groups.

evolution of volatiles, a coal fuses and swells to form a coke.

According to the *partial-melt theory,* only a fraction of the coal melts and the remainder is plasticized. The fusible fraction, the coking principle or bitumen, has been identified in benzene and chloroform extracts.

In the *thermobitumen theory,* softening is defined as a consequence of pyrolysis that transforms coal into fluid products. The appearance of these fluid products causes softening and plasticity. Their disappearance by evaporation or thermal decomposition results in resolidification. The fluid products are termed metaplast or thermobitumen, implying that they are the products of pyrolysis.

The *physical and physicochemical theory* proposes that thermal softening is initially a physical process of partial melting of the coal substance to a fluid mass. However, the rheological properties of this mass are subseqently modified by chemical reactions during decomposition. Therefore, development of fluidity is believed to be the result of both physical and physicochemical processes.

In noncaking coal, the decomposition point is lower than the softening point. The rank dependency of these temperatures is illustrated in Figure 6, which shows that plastic coals occupy only a narrow range of rank. Other theories suggest that the fraction of a coal that fuses plasticizes the remainder [26–35].

Thermal softening of coal is apparently not related to physical melting but instead to kinetic phenomena [3], [36]. The development of fluidity may be the result of physical and physicochemical processes without appreciable rupture of covalent bonds during pyrolysis. After melting of the mass, the properties are modified by pyrolytic reactions [31]. This mechanism is consistent with three conditions that have been proposed as being necessary and sufficient for plasticity development [37]:

1) lamellar bridging structures that can be thermally broken
2) a supply of hydroaromatic hydrogen

3) intrinsic ability of micelles and lamellae to become mobile, i.e., ability of the coal to melt independent of thermal bond rupture.

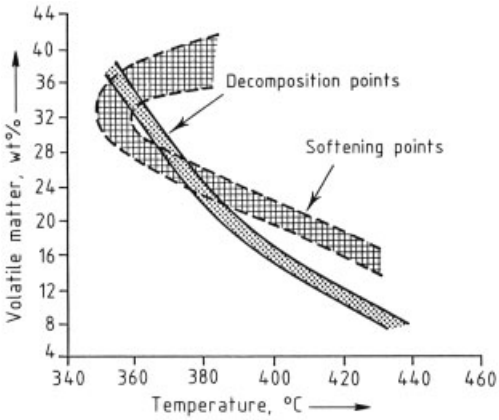

Figure 6. Relationship of volatile matter content to softening points and decomposition points of coals [26]

2.3.2. Swelling

Investigators agree that the softening of coal particles is caused by the melting or solvation of low molecular mass substances or by primary pyrolytic decomposition [5]. Softening causes swelling and the formation and liberation of gas. The gas cannot escape rapidly because of the low permeability of the plastic mass. Hence, gas bubbles are formed with high internal pressures. In response to this internal pressure, the viscous plastic mass swells like baker's dough. The formation of an impermeable plastic mass is a precondition for swelling.

Swelling can be observed in coal particles individually or collectively. *Intraparticle* (single-particle) *swelling* results from the formation of a plastic mass that eliminates the fissures and macropores through which gas could escape. The resulting internal pressure swells the plastic mass. Swelling of particle beds, i.e., *interparticle swelling*, is caused similarly when the plastic mass fills the interparticle spaces, preventing gas from escaping.

The importance of each type of swelling depends on the particle size and the nature of the coal. Intraparticle swelling is only important for weakly coking coal. For coal to form a coherent and strong coke, interparticle swelling is necessary. In highly plastic coal distinction between the two types of swelling is often impossible [22].

The extent of swelling depends on the rate of devolatilization [3], [4]. A minimum fluidity is essential to reduce the permeability of the plastic mass to gas flow and generate swelling pressure. However, if the plastic mass is too fluid, gas bubbles escape at low pressure.

Swelling has been viewed as a process independent of softening [33], [34], [38]. In this light, swelling is a consequence of the pore structure of the starting coal. It is caused by the pressure of the evolved gas trapped within the micropores.

Porosity exhibits a minimum for the higher rank bituminous coals, for which swelling is at a maximum. This mechanism explains intraparticle swelling, but not interparticle swelling. Because microporosity is a prominent feature of agglomerating coal, an empirical relationship between swelling and porosity may exist without necessarily implying that swelling is a consequence of porosity. In fact, swelling and porosity may be consequences of a third factor, such as cross-linking density.

2.3.3. Resolidification

Resolidification is usually attributed to a series of cracking reactions followed by condensation and polymerization. A large number of free radicals and small molecules are generated by cracking. This is consistent with the observation that the original solvating agent, the bitumen, decomposes at ca. 480 °C, before resolidification of the coal to form a semicoke [31]. Because of the loss of hydrogen-donating and solvating agents, which would otherwise stabilize the free radicals, the solid material polymerizes forming semi-coke. This viewpoint is supported by the fact that the semicoke is deficient in hydrogen compared to the precursor.

2.4. Coke Microstructure and Mesophase Formation

The properties of cokes are determined by their microstructural characteristics. The structure – property relationship is, however, difficult to establish for coke because of its extreme heterogeneity in chemical constitution

and physical structure. Nevertheless, considerable progress has been made since the 1920s, primarily with the aid of optical microscopy.

The most significant industrial use of coke is in the production of iron and steel. Only prime coking coal produces coke appropriate for blast furnaces, where reactivity and stability under thermal stress are required [39], [40]. These requirements are met by metallurgical coke.

In general, the principal feature of the coke microstructure is the anisotropy of the organic coke substance as observed by optical microscopy [41–45]. The critical role of an intermediate plastic phase (carbonaceous mesophase) in the development of the characteristic anisotropic structures was first recognized in 1965 [46], [47]. Subsequent research has been devoted to the significance of mesophase formation during the carbonization of coking coal [48], the effect of coal rank and chemical constitution on mesophase formation [49–53], the use of optical and scanning electron microscopy [54], and the chemistry of mesophase formation [55–59]. These investigations may help to improve coke manufacture by modifying the microstructure through the development of a carbonaceous mesophase.

2.4.1. Microscopic Characterization

Optical microscopy is the principal means of analyzing coke morphology and anisotropic structures [54], [57]. The resolution limit of the optical microscope is ca. 0.2 µm. Optical anisotropy is an expression of structural anisotropy on the same scale [60].

For microscopic examinations, polished coke sections are viewed in plane-polarized light [61]. The appearance of the polished surface is termed the optical texture. Isotropic and anisotropic structures can be distinguished on the basis of different behavior [62], [63]. A common technique employs *reflection interference colors* [63]. These are yellow, blue, and dark purple for anisotropic materials and light purple for isotropic materials. Thus, cokes can be characterized by color and the shape and size of the isochromatic areas [54].

Differences in optical activity between isotropic and anisotropic cokes are due to differences in the range of crystallographic order [55]. The crystallographic order of isotropic coke is considered to be only small in scale; i.e., the parallelism of associated constituent lamellar molecules extends over short distances (ca. 1 – 5 nm). Although the structure can be claimed to be anisotropic for 1 – 5 nm, it is essentially isotropic to the light beam because this range is far below the resolution of the optical microscope (ca. 200 nm). However, for anisotropic coke, the range of crystallographic order extends from 0.3 to ca. 200 µm; in the light beam, the material has bulk anisotropy [55].

An optical microscopic classification of the organic solid residues derived from the hydrogenation of bituminous coal can be applied to coke obtained by carbonization. Table 2 includes the type of macerals and the transient phases responsible for coke microstructures. The formation of the transient, fluid phases (vitroplast and mesophase) is indicative of fusion (or plasticity) during carbonization and is limited to certain bituminous coals.

The term *vitroplast* designates a plastic or formerly plastic carbonization product of vitrinite that, unlike the starting material, is optically isotropic (e.g., the particle marked VP in Fig. 7). In coke it is characterized by flow structures and spherical morphologies.

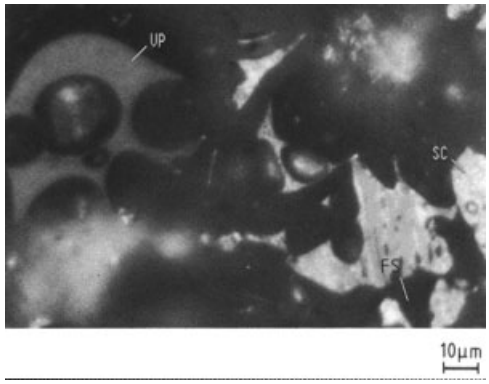

Figure 7. Optical micrograph of the surface of solids from the carbonization of a low-volatile bituminous coal in an entrained-flow reactor (temperature: 900 °C; residence time: 0.15 s; heating rate: 10^4 °C/s)
VP: Vitroplast; SC: Semicoke; FS: Fusinite

The semifusinite maceral can also make a small contribution to the formation of spherical vitroplast. Further alteration of vitrinite and the spherical vitroplast derived from it is seen in the development of *cenospheres.* This term was

Table 2. Optical microscopic classification of solid carbonization products [64]

Classification	Optical nature	Intermediate fluid phase	Maceral precursor
Vitroplast (VP∗)	isotropic	vitroplast	vitrinite, semifusinite
Cenosphere (CS∗)	isotropic and anisotropic	vitroplast	vitrinite
Semicoke, (SC∗)	anisotropic	mesophase	vitrinite, semifusinite
Unreacted or slightly altered			
vitrinite, (UV∗)	anisotropic	–	vitrinite
fusinite, (FS∗)	isotropic	–	fusinite

∗ Abbreviation.

first used to designate structures formed by rapid heating of pulverized coal [65], [66]. A cenosphere is defined as a reticulated hollow sphere composed of ribs or frames and windows (e.g., the particle marked CS in Fig. 8). Cenospheres are formed by the gas liberated within vitroplast spheres during the plastic stage [64].

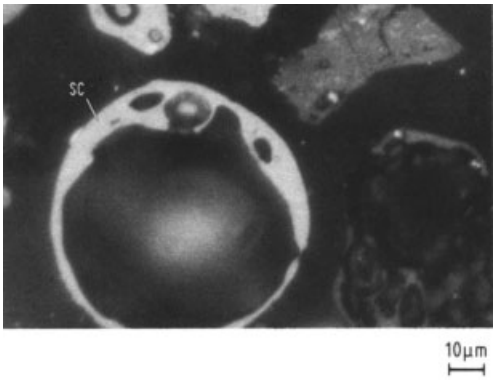

Figure 8. Optical micrograph of the surface of solids from the carbonization of a high-volatile bituminous coal in an entrained-flow reactor (temperature: 900 °C; residence time: 0.15 s; heating rate: 10^4 °C/s)
∗ CS: Cenosphere.

In contrast to vitroplasts, the *mesophase* is optically anisotropic (e.g., the particles marked SC in Figs. 7 and 9) and represents the main route for the formation of anisotropic structures in coke. It is considered to be a unique, ordered, fluid system consisting of high-molecular-mass, planar molecules. The solid product formed via a mesophase is called a semicoke, a term used in a more general sense for the solid products of low-temperature (500 – 600 °C) carbonization of coking coals or heavy hydrocarbon feedstocks, such as coal – tar and petroleum pitches.

Both vitroplasts and cenospheres may develop intrinsic anisotropic structures (depending on the conditions of carbonization) while retaining their characteristic morphology. Another type of anisotropy in coke is associated with the unreacted vitrinite fragments. The distinction among these anisotropic microconstituents has a significant bearing on coke properties [64].

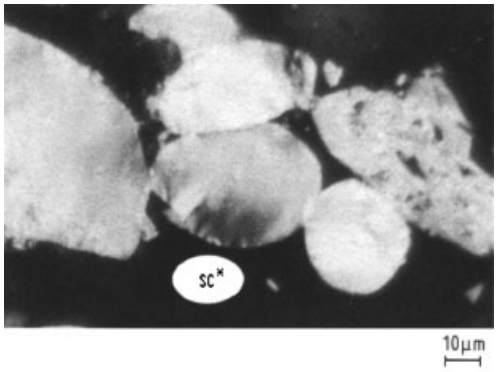

Figure 9. Optical micrograph of the surface of solids from the carbonization of a low-volatile bituminous coal in a packed-bed reactor showing extensive mesophase development (temperature: 900 °C; residence time: 1 h; heating rate: 10 °C/min)
∗ SC: Semicoke.

Scanning electron microscopy (SEM) is also useful for examining the microstructural characteristics of solids [67]. Its high resolution (ca. 3 nm) and depth of focus permit characterization of surface topography at very low and high magnifications (up to ×200 000). Specimens are examined by sweeping a finely focused electron beam across the surface. The resulting signals include secondary electrons, backscattered electrons, and characteristic X-rays. In the examination of coke, the most useful signal is the variation in secondary electron emission with changing surface topography. Optical microscopy and SEM are complementary techniques for coke characterization, providing structural and topographical information, respectively [55].

2.4.2. Carbonaceous Mesophase

The important properties of cokes, including mechanical strength and reactivity, are governed by the arrangement of the constituent carbon atoms. The principal features of the atomic arrangement are the alignment and size of *carbon crystallites.* The highest degree of alignment is seen in natural graphite crystals, which consist of extended layers of fused hexagonal rings of carbon atoms. This structure, with an interlayer spacing of 0.3354 nm, is approached by certain carbons manufactured from coal – tar and petroleum pitches and by polynuclear aromatic hydrocarbons on heating at 2500 – 3000 °C. Depending on their behavior when heat-treated, carbonaceous solids can be classified as graphitizing or nongraphitizing, e.g., coke and char, respectively, in the case of coal-derived solids.

Although the crystallite alignment in carbonaceous solids is only observable after heating at 1000 °C, crystallite alignment starts during carbonization over a narrow temperature range near 450 °C [57]. This stage, coincident with plasticity development, is characterized by the formation of carbonaceous mesophase, which leads to the anisotropic structures found in cokes. The shape and size of these anisotropic structures usually determine the behavior of the coke [68]. Substances that form graphitizing coke via a mesophase include the vitrinites of medium-volatile coking coal, coal – tar and petroleum pitches, polymers such as poly(vinyl chloride), and polynuclear aromatic compounds such as anthracene.

Lignites, thermosetting resins, polymers such as poly(vinylidene chloride), and hydrocarbons such as biphenyl do not form mesophase; instead, they produce isotropic nongraphitizing coke commonly referred to as *char* [57].

An outstanding characteristic of carbonaceous mesophase is the prevailing *molecular order,* which represents an intermediate state between totally isotropic liquids and anisotropic solid crystals. A crystal is characterized by long-range order. Isotropic liquids, on the other hand, possess no long-range order [69]. In liquid crystals, some but not all of the long-range order is maintained. The state is neither crystalline nor strictly liquid, and in 1922 the term *mesophase* was proposed [70–72]. The plastic, anisotropic phase formed during carbonization is designated interchangeably as liquid crystal or (carbonaceous) mesophase.

Of the several types of *liquid crystals,* the nematic (threadlike) type is the most relevant to carbonaceous mesophase [73]. Nematic liquid crystals separate from the isotropic liquid as anisotropic spherical droplets that eventually coalesce to give nematic domains. The molecular arrangement is envisaged as rod-shaped, similarly sized molecules lying parallel to each other but with no stacking order.

The parallel alignment of the molecules in carbonaceous mesophase and the separation from the isotropic liquid in the form of spheres suggest the behavior of nematic liquid crystals. The molecular units of carbonaceous mesophase are, however, disk-shaped and widely varied in size [74]. Furthermore, the formation of carbonaceous mesophase usually involves irreversible chemical processes, in contrast to the physical and reversible formation of nematic liquid crystals. Carbonaceous mesophase consists essentially of large polynuclear aromatic hydrocarbons [75]. Mesophase spheres formed during the carbonization of pitch may contain smaller molecules incorporated in the matrix of planar condensed-ring compounds (see Fig. 10) [74]. This gives a molecular mass distribution in the range of 400 – 3000 or even more.

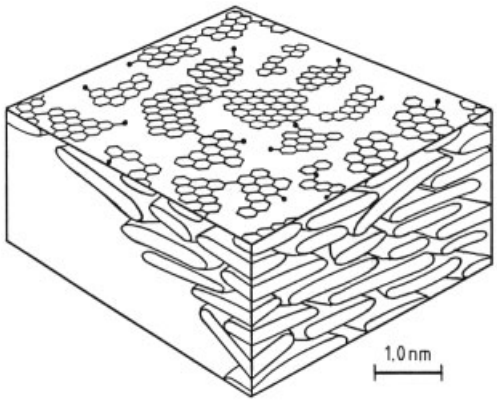

Figure 10. Nematic molecular order of carbonaceous mesophase comprised of methyl-substituted polynuclear aromatic compounds connected by methylene bridges and biaryl linkages [74], [76]

The most commonly observed arrangement of molecules in mesophase spheres is the *Brooks – Taylor configuration*, in which layers of planar molecules lie parallel to the equato-

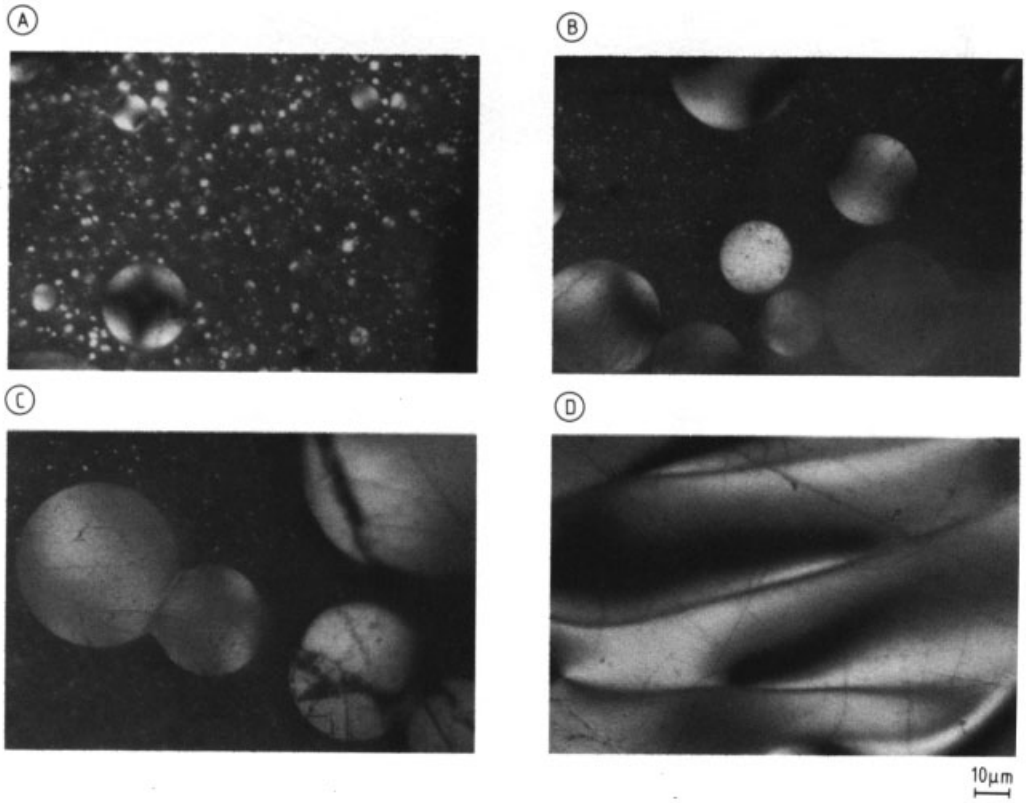

Figure 11. Optical micrographs showing the stages of mesophase formation during the carbonization of a decant oil in closed-tubing reactors (temperature: 500 °C; reaction time: 15 min, 30 min, 1 h, and 2 h, respectively)
A) Nucleation. B) Growth. C) Coalescence. D) Semicoke (flow domains)

rial plane around the polar axis, but curve to reach the surface at an angle [57], [77]. Other configurations have been observed, especially in the presence of induced magnetic fields and solid particles, e.g., carbon black [78], [79]. These configurations apparently show no different properties.

The uninhibited *formation and development of mesophase* occurs in three sequential stages: nucleation, growth, and coalescence. This progression, although clearly seen with the optical microscope in pitch and aromatic hydrocarbons, is not always observed during coal carbonization because of the inhibition of growth and coalescence or the very rapid formation of the final texture of the semicoke.

Nucleation is observed as the separation of anisotropic spheres from the isotropic melt, when they become sufficiently large (ca. 0.3 µm in diameter) to be visible (Fig. 11 A). This is a homogeneous process that does not require a specific site. Under favorable conditions, the spheres can grow at the expense of the isotropic melt (Fig. 11 B) and coalesce (Fig. 11 C) to form large domains of ordered regions (Fig. 11 D), which eventually solidify to semicoke. This type of extensive mesophase development (flow domains) leads to highly graphitizable coke.

If growth and coalescence are inhibited, irregularly shaped anisotropic spheres of small-scale order (0.3 – 10 µm grain size) are formed (see Fig. 12). These structures, frequently encountered in coal carbonization, are termed *mosaics.* A widely used classification, based on the shape and size of the anisotropic structures found in coke, is as follows [51]:

Isotropic no optical activity, uniform, light purple color on the surface
Anisotropic
Via mesophase dark purple, yellow, and blue areas of varying shape and size

Fine-grained mosaics irregularly shaped, isochromatic areas (ISIA), 0.5 – 1.5 µm long
Medium-grained mosaics ISIA, 1.5 – 5.0 µm long
Coarse-grained mosaics ISIA, 5.0 – 10 µm long
Coarse flow rectangular, flow-type structures 30 – 60 µm long; > 10 µm wide
Flow domains flow-type structures > 60 µm long; > 10 µm wide
Domains isometric structures > 60 µm in diameter
Basic anisotropy flat, featureless anisotropy associated with the parent unreacted (or slightly altered) vitrains from high-rank coals

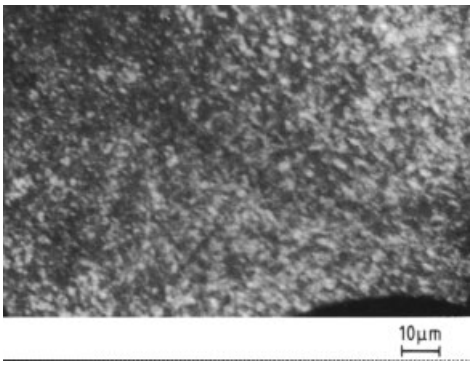

Figure 12. The fine mosaic texture of a semicoke from the carbonization of a petroleum distillation residue in a closed-tubing reactor (temperature: 500 °C; reaction time: 1 h)

The sequential formation of mesophase is apparently not the sole mechanism leading to anisotropic structures in coal-derived coke. *Vitrains* from high-rank coals behave somewhat differently from coking and caking coals [49]. The inherent (basic) anisotropy of vitrain, which is rather featureless, is unchanged or converted directly to flow-type anisotropy. In this respect, it does not follow the progressive development of flow-type anisotropy.

Mesophase formation results from extremely complex and interdependent *physical and chemical processes* occurring during carbonization. The first chemical process is the rupture of weak bonds, breaking down the molecular network and forming free radicals, followed by such reactions as cyclization, aromatization, and polymerization.

The *fluid carbonization medium* at this stage can be envisaged as a dynamic chemical system, with increasing molecular mass and size [55]. When the planar molecules reach a critical concentration and size, van der Waals forces become sufficiently strong to establish parallel orientation. The resultant anisotropic miscellae separate from the isotropic fluid phase in the form of spheres. If the molecules are not sufficiently planar, parallel stacking is sterically hindered. The spherical shape of the separated anisotropic miscella results from the usual requirement of minimum surface area and surface energy at the interface of a two-phase system [55].

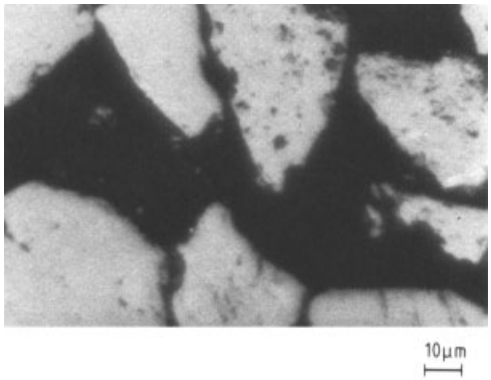

Figure 13. Isotropic texture of a char from the carbonization of a lignite in a packed-bed reactor (temperature: 900 °C; residence time: 1 h; heating rate: 10 °C/min)

The principles governing mesophase formation can be summarized as follows:

1) Heteroatoms and alkyl groups inhibit mesophase formation by cross-linking reactions, which impair planarity.
2) Naphthenic hydrogen promotes mesophase formation via hydrogen-transfer reactions, which retard cross-linking and increase fluidity.
3) Aromatic systems promote mesophase formation by virtue of their planarity and low reactivity.

These findings are based on studies of model compounds thought to resemble the building blocks of bituminous coal and petroleum feedstocks [56], [59], [80], [81]. In addition, mesophase formation depends on carbonization conditions.

The carbonization of coal of different rank produces coke with corresponding optical prop-

erties [53]. Low-rank coals (lignite and subbituminous) do not soften during carbonization, thus producing isotropic chars (see Fig. 13) [54].

The size of the anisotropic structures in coke generally increases with increasing coal rank (see Fig. 14) [51]. Whole coal behaves like vitrain. Caking coal with high volatile content forms isotropic coke containing small amounts of finegrained mosaics. With increasing rank up to the prime coking coals, these anisotropic structures become larger (5 – 10 µm) and more numerous.

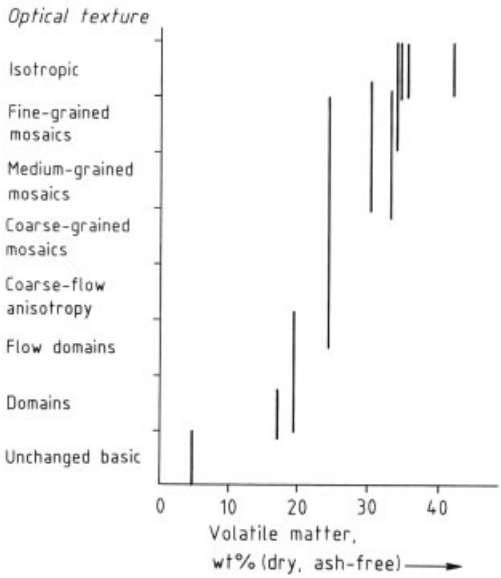

Figure 14. Variation of coke optical texture with volatile matter content of the parent vitrains [51]

In coke from prime coking coal, the optical texture (apart from inorganic and infusible organic petrographic components) is essentially and continuously anisotropic, with structures varying in size from fine-grained mosaics to coarse particles up to 20 µm. The anisotropic structures in coke from fusing coal (high-to-low-volatile-content bituminous) develop mostly via mesophase formation.

Semianthracites, anthracites, and some noncaking low-volatile bituminous coals have no fluid stage during carbonization. The large anisotropic structures found in these cokes originate from the inherent anisotropy of the parent coal [49].

The variations in the optical texture of coke from coals of different rank reflect the differences in mesophase formation. These differences can be explained in terms of the variations in chemical constitution.

Coals become more aromatic with increasing rank, which is also associated with decreasing oxygen content and increasing incorporation of oxygen into cyclic structures [82], [83], [84].

In low-rank coal, most oxygen is in the form of carboxyl (–COOH) and hydroxyl (–OH) groups; both are reactive cross-linking agents. Oxygen in cyclic structures does not participate in cross-linking. In light of the principles discussed in this section, increasing aromaticity and decreasing oxygen content can explain the improvement in mesophase formation with increasing coal rank.

Carbonization conditions also affect mesophase formation. These effects are due to changes in the rate of chemical and physical processes [85], [86].

2.4.3. Modification of Optical Texture and Properties

The most significant feature of the microstructure of metallurgical coke is the unique anisotropy that is associated with low porosity and takes the form of mosaics bridged by small flow-type structures in the pore wall [53]. Most macropores are hundreds of micrometers in size; the pore wall is of significant thickness [53]. The limited porosity provides a coke density suitable for blast furnace operations, whereas the bonding of the small growth units of the anisotropic structures strengthens the coke. In addition, the anisotropic and nonporous nature of the coke reduces reactivity toward carbon dioxide, as is desired in blast furnace operations.

Consequently, coke properties can be modified by altering the optical texture and morphology. This makes possible manufacture of metallurgical coke from the large reserves of noncoking coal of low rank.

Modification of mesophase formation has therefore, great value in improving coke. The cocarbonization of noncoking coal with coking coal or pitch is also useful [87], [88].

3. Yield and Distribution of Pyrolysis Products

3.1. Experimental Methods

The yield of pyrolysis products (gases, liquids, and solids) is determined by monitoring the loss of coal mass or the quantity of gas evolved. In the traditional determination of proximate volatile matter (e.g., ASTM or DIN), the yield is measured as the mass lost from a fixed bed at standard conditions. Deviations from these conditions can change mass losses significantly (see Fig. 15). Smaller samples in a standard crucible (shorter bed depths) and higher heating rates and temperatures increase volatile yield.

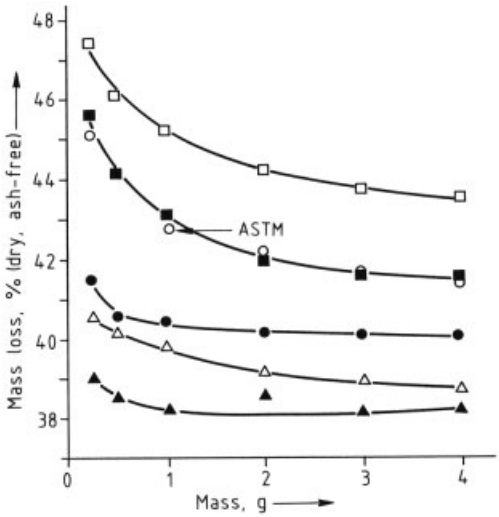

Figure 15. Effect of sample mass (bed depth), heating rate, and pyrolysis temperature on mass loss [89]

Symbol	Heating rate, °C/s	Temp., °C
△	6	750
○∗	14	950
□	22	1150
▲	0.5	750
●	0.5	950
■	0.5	1150

∗ Standard ASTM volatile matter test.

Yields are also increased by reducing gas pressure and particle size. Thus, the volatile-matter content is highly dependent on pyrolysis conditions and can exceed 150 % of the ASTM value (e.g., last line, Table 3). When volatile products are desired, increased yield is of great interest. Furthermore, factors that influence the product distribution are important because in many processes the liquids (not the gases) are of primary interest.

In contrast to coal gasification, for example, with its few and easily identified products, the prediction of yield and product distribution in coal pyrolysis is difficult. The extreme complexity of coal pyrolysis precludes accurate thermodynamic analysis. Moreover, the temperature and pressure are often ill-defined. In rapid pyrolysis, internal pressures as high as 100 MPa have been estimated [89]. The design of reactors and processes is usually based on experimental determination of the yield and product distribution as functions of temperature, heating rate, pressure, particle size, and coal rank [90].

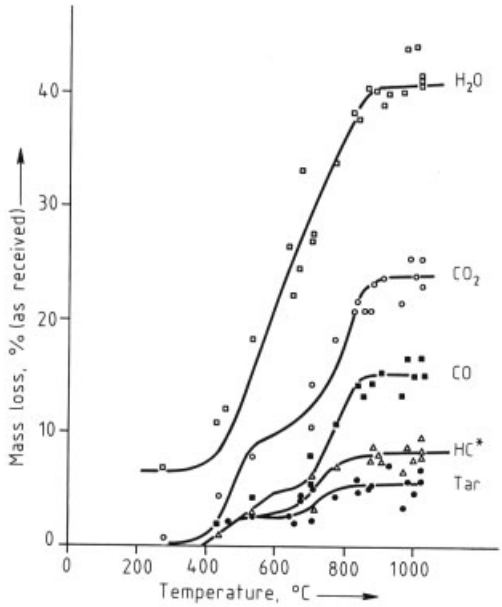

Figure 16. Effect of temperature on the distribution of volatile products from a lignite (pressure: 101.3 kPa of He; mean particle diameter: 74 µm; heating rate: 270–10 000 °C/s)
□ Total; ○ Tar + HC∗ gases and $H_2 + CO + CO_2$; ■ Tar + HC∗ gases and $H_2 + CO$; △ Tar + HC∗ gases and H_2; ● Tar [90]
∗ HC: Hydrocarbon.

3.2. Experimental Variables

3.2.1. Time and Temperature

The determination of temperature and residence time in entrained-flow reactors presents special

Table 3. Comparison of experimental pyrolysis yields with proximate volatile matter [90]

Conditions	Coal rank	Proximate analysis∗, *VM*, %	Experimental yield, V', %	V'/VM∗∗
Dry basis,	medium-volatile bituminous, (mvb)	25.3	19	0.75
heated to 950 °C	high-volatile A bituminous, (hvAb)	31.6	36	1.14
at 600 °C/s	high-volatile A bituminous, (hvAb)	37.7	49	1.30
	subbituminous	40.7	48	1.17
	high-volatile bituminous, (hvb)	44.0	41	0.93
Dry, ash-free basis,	bituminous	46.2	53.7	1.16
heated to 1000 °C at 10^3 °C/s	lignite	44.3	44.7	1.01
	bituminous	44.6	52.2	1.17
heated to 2000 °C at ca. 10^5 °C/s	bituminous	42.5	65	1.53

∗ ASTM method.
∗∗
V' = total volatile matter; VM = proximate volatile matter.

problems. Particle temperature is usually estimated from measurements of gas temperature (by suction pyrometry) with the aid of a heat-transfer model. Particle residence time (≤ 1 s) is calculated from known or assumed trajectories by using an appropriate model for the fluid and solid dynamics. The total yield of volatile substances, when not obtained from material balances, is frequently determined from the mineral matter or ash in the coal or char. The distribution of volatile products is usually determined by gas chromatography or mass spectrometry.

The yield of volatile products in the pyrolysis of coals of different rank under different heating rates increases continuously with increasing temperature. This results from the existence of a hierarchy of bond energies in the structure of coals (see Fig. 1). In most cases, even at heating rates as high as 10 000 °C/s in entrained-flow reactors, pyrolysis is complete at ca. 1000 °C. For a longer residence time, a lower temperature is needed to achieve a particular yield, in accordance with Arrhenius-type behavior (see Chap. 4). In particular, the yield is a sensitive function of temperature for short residence times in entrained-flow reactors.

The effect of temperature, the most important influence on product distribution, is illustrated in Tables 4 and 5, and Figures 16 and 17. The influence of temperature on the distribution of volatiles obtained from rapid pyrolysis of a lignite (270 – 10 000 °C/s) is clearly seen in Figure 16. The yield of tars reaches a low maximum at relatively low temperature. The hydrocarbon gases consist mainly of methane, ethylene, and hydrogen. The oxygen-containing products, water, carbon dioxide, and carbon monoxide, dominate the gas composition. A result of this temperature dependence is that the pyrolysis products have a range of heating values. This is illustrated in Figure 18 for lignite and is typical of the problems associated with obtaining clean-burning fuel from low-rank coals.

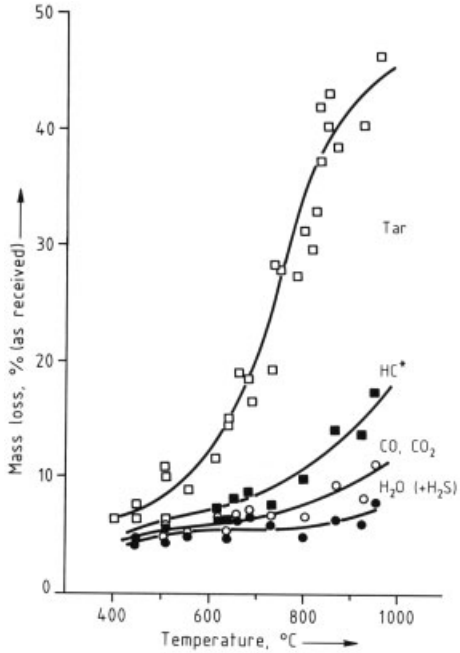

Figure 17. Effect of temperature on the distribution of volatile products from a bituminous coal (pressure: 101.3 kPa of He; mean particle diameter: 74 µm; heating rate: 1000 °C/s)
□ Total; ■ HC∗ gases and $H_2 + CO$ and $CO_2 + H_2O$ and H_2S; ○ CO and $CO_2 + H_2O$ and H_2S; ● H_2O and H_2S [90]
∗ HC: Hydrocarbon.

In contrast, a high yield of tar and other combustible matter is obtained by the rapid pyrolysis

Table 4. Distribution of volatile products in coal pyrolysis [90]

Method	Coal type	Volatile products, % (dry)			
		Tar [a]	Gas	Water	V [b]
Fischer assay, 600 °C	Fürst Leopold	11.2	10.0	8.1	29.3
Stirred bed, 600 °C		18.7	7.0	7.5	33.2
Slow heating, 0.08 °C/s	Moscow brown coal	4.6	11.9	6.0	22.5
Rapid heating to 500 °C		4.0	8.1	9.6	21.7
Thin layer of particles	Maigre Oignies no. 2	1.8	5.1	1.5	8.4
heated to 1050 – 1100 °C	Bergmannsglück	41.7	5.9	3.0	23.6
at 1500 °C/s	Emma fines	15.9	6.2	2.4	24.5
	Lens-Lievin no. 10	17.1	7.0	2.6	26.7
	Flenus de Bruay	25.2	9.6	4.4	39.2
	Wendell III	22.5	7.9	4.1	34.5
	Faulquemont	18.9	12.6	5.2	36.7
Entrained flow, 600 °C	West Virginia, mvb [c]	18.7	1.8	6.5 [e]	27.0
	Illinois, mvb [c]	19.6	19.7	4.6 [e]	43.9
	Pittsburgh, hvAb [d]	31.5	22.0	8.0 [e]	61.5
	Wyoming, subbituminous B	11.0	26.4	5.6 [e]	43.0
Fischer assay, 500 °C	Western Kentucky	16.3	5.0	9.4	30.7
Flow tube, 530 – 650 °C	hvBb [c]	33.0	6.6	1.7	41.3
Thin layer of particles heated	Pittsburgh, hvAb [d]	25.8	11.3	6.5	43.6
to 1000 °C at 1000 °C/s	Montana lignite	6.8	22.7	10.4	39.9

[a] Volatiles condensable at 20 °C.
[b] Combined volatiles from tar, gas, and water.
[c] See Table 3 for abbreviation.
[d] High-volatile A bituminous.
[e] Estimated as mass difference between original dry coal and combined tar, gas, and char yields.

Table 5. Composition of gaseous products in coal pyrolysis [90]

Conditions	Temperature,	Volatiles,	Gas composition, vol% (dry)			
	°C	%	CO_2	CO	H_2	CH_4
Thin layer of particles	105 – 1100 (at 1500 °C/s)	8.4	2.2	10.1	66.2	8.6
		18.1	1.6	5.0	67.2	21.4
		20.4	2.0	6.0	62.9	20.3
		24.4	3.8	9.6	64.2	21.0
		31.0	4.1	15.8	54.3	16.9
		33.9	2.4	10.0	52.8	21.8
		36.4	6.1	20.6	60.3	13.1
Fluidized bed	600	31	12.0	12.3	14.7	50.5
Fischer assay	600	31	10.3	8.7	22.4	54.2
Entrained flow	600	31.6	9.2	3.6	19.4	24.7
		34.6	11.7	0.0	23.0	40.7
		37.7	9.5	9.5	17.1	42.9
		40.7	16.0	10.5	25.6	33.9
Flow tube	530 – 650	41	10.6	4.9	25	42.3
Thin layer of particles	1000 (at 1000 °C/s)	39.5	3.2	9.9	57.8	17.8
		39.6	24.0	37.2	27.8	8.8

of bituminous coal (Fig. 17). In both cases, as in all pyrolysis processes, the yield of char or coke is substantial, even at 1000 °C.

3.2.2. Heating Rate

The effect of heating rate on coal pyrolysis has often been misunderstood. If slow heating (e.g., 10 °C/min in fixed-bed reactors) is compared with rapid heating (e.g., 10 000 °C/s in entrained-flow reactors), the latter usually gives a better yield. In most such comparisons, however, the heating rate is not the only variable. A change from a fixed-bed reactor to an entrained-flow reactor involves a sharp decrease in the density of the pyrolyzing medium, with important implications for secondary char-forming reactions (see Section 3.3). However, when the only change is an increased heating rate (see Table 6 and Fig. 19), yield and product distribution are

Table 6. Effect of heating rate on product distribution from bituminous coals [90]

Product	Yield, % (as received)		
	350 – 450 °C/s	1000 °C/s [b]	13 000 – 15 000 °C/s
CO	2.4	2.4	2.3
CO_2	1.6	1.2	1.7
H_2O [c]	7.6	7.8	7.7
CH_4	2.2	2.5	2.4
C_2H_4	0.40	0.83	0.66
C_2H_6	0.59	0.51	0.59
C_3S	1.1	1.3	1.2
Other HC gas	1.1	1.3	1.4
Light HC liquid	2.3	2.4	2.7
Tar	22.4	23.0	23.0
H_2	–	1.0	–
Char	54.0	53.0	53.0
Total	95.69	97.24	96.65
Error (loss)	4.3	2.8	3.3
Number of runs	3	20	2

[a] Pressure: 101.3 kPa of He; temperature: 850 – 1000 °C; mean particle diameter: 74 µm; residence time: 2 – 10 s.
[b] Base case.
[c] Includes coal moisture (1.4 %); may include some H_2S.

unaffected [90], but the time required to reach a given yield is reduced.

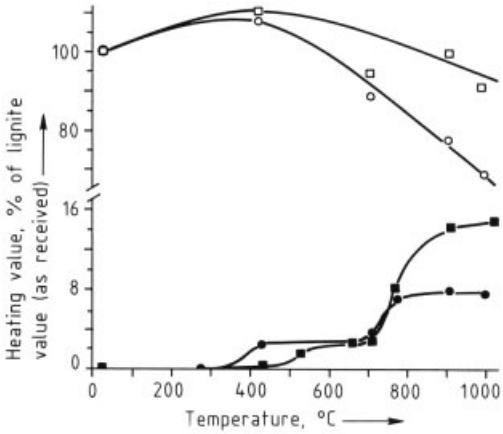

Figure 18. Effect of temperature on the distribution of heating value among the products of pyrolysis of a lignite (pressure: 101.3 kPa of He; heating rate: 1000 °C/s)
□ Total; ○ Char; ■ Gas; ● Tar [90]

3.2.3. Pressure

In the rapid pyrolysis of bituminous coal, yield usually decreases with increasing pressure (see Fig. 20). A drop in the yield of tar is not compensated by an increase in the yield of hydrocarbon gases, such as methane (see Fig. 21 and Table 7). Increased pressure evidently favors char-forming reactions, particularly cracking and polymerization.

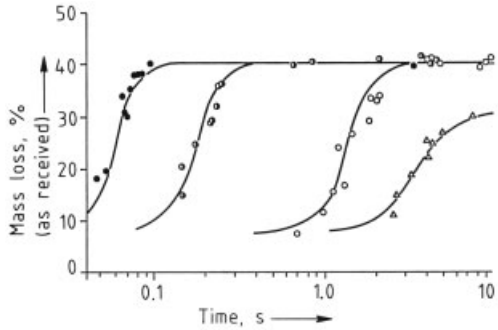

Figure 19. Effect of time on pyrolysis mass loss from a lignite (pressure: 101.3 kPa of He; final temperature (if attained): 1000 °C; mean particle diameter: 70 µm)
● 10^4 °C/s; ◗ 3000 °C/s; ○ 650 °C/s; △ 180 °C/s [90]

3.2.4. Particle Size

As with heating rate, the effect of coal particle size has been difficult to determine. Changes in particle size are usually accompanied by changes in temperature and residence time. In entrained-flow systems, the calculated residence time is dependent on particle size. Obtaining identical temperature – time histories for a range of sizes is difficult; therefore, reports in the literature are conflicting.

As might be intuitively expected (see Section 3.3), larger particles give a lower yield of volatiles. Increase in char formation (due to

Table 7. Effect of pressure on product distribution from lignite and bituminous coal [90]

Product	Yield, % (as received)					
	Lignite			Bituminous [b]		
	6.6 Pa of He [a]	0.1 MPa of He [a]	6.9 MPa of He [b]	6.6 Pa of He	0.1 MPa of He	6.9 MPa of He
CO	6.1	7.1	9.0	2.0	2.4	2.5
CO_2	7.6	8.4	10.6	1.4	1.2	1.7
H_2O [c]	17.7	16.5	13.4	6.8	7.8	9.5
H_2	–	0.50	–	0.75	1.0	–
CH_4	0.94	1.3	2.5	1.6	2.5	3.2
C_2H_4	0.43	0.56	0.55	0.45	0.83	0.46
C_2H_6	0.21	0.18	0.17	0.44	0.51	0.89
$C_3H_6 + C_3H_8$	0.46	0.37	0.38	0.71	1.3	0.71
Other HC [d] gases	0.60		0.21	0.98	1.3	1.6
		0.47				
Light HC [d] liquids	0.81		1.1	1.6	2.4	2.0
Tar	6.9	5.4	2.8	31.9	23.0	12.0
Char	55.2	58.7	59.0	48.5	53.0	62.4
Total	97.0	99.5	99.7	97.1	97.2	97.0
Error (loss)	3.0	0.5	0.3	2.9	2.8	3.0

[a] Final temperature: 900 – 1035 °C; no residence time.
[b] Final temperature: 850 – 1070 °C; mean particle diameter: 74 µm; heating rate: 1000 °C/s; residence time: 2 – 10 s.
[c] Includes coal moisture (lignite: 6.8 %; bituminous: 1.4 %); may include H_2S.
[d] HC = Hydrocarbon.

Table 8. Effect of particle size on product distribution from a bituminous coal [a] [90]

Product	Yield, % (as received)			
	53 – 88 µm (average, 74 µm)	< 300 µm [b]	300 – 830 µm	830 – 990 µm
CO	2.4	2.7	3.2	3.0
CO_2	1.2	1.1	1.2	1.3
H_2O [c]	7.8	5.4	5.3	7.2
H_2	1.0	–	–	0.99
CH_4	2.5	2.9	3.0	3.2
C_2H_4	0.83	1.0	1.1	1.3
C_2H_6	0.51	0.50	0.55	0.63
C_3S	1.3	0.92	0.84	1.1
Other HC [d] gases	1.3	1.4	1.1	1.2
Light HC [d] liquids	2.4	2.5	2.6	2.7
Tar	23.0	24.2	21.3	18.4
Char	53.0	57.1	56.5	55.8
Total	97.2	99.7	96.7	96.8
Error (loss)	2.8	0.3	3.3	3.2
Number of runs	20	1	2	3

[a] Pressure: 101.3 kPa of He; final temperature: 850 – 1070 °C; heating rate: 1000 °C/s; residence time: 3 – 10 s.
[b] 830 – 990-µm sample ground to pass 297-µm sieve (50 mesh).
[c] Includes coal moisture (1.4 %); may include some H_2S.
[d] HC = hydrocarbon.

tar cracking or polymerization) is not compensated by increased yield of hydrocarbon gas (see Fig. 22 and Table 8).

3.2.5. Coal Rank

Coals over a wide range of rank can give yields of pyrolysis products that are higher than their standard content of volatile matter (as determined by ASTM proximate analysis) (see Figs. 16 and 17, Tables 4 and 9). The most important result is the high tar yield given by entrained-flow pyrolysis of bituminous coal.

3.3. Mechanism

The objective of most coal pyrolysis processes, whether fuels or chemicals are being produced, is to maximize the yield of gaseous and liquid products (coke production is a notable exception). This is achieved by reducing secondary char-forming reactions (cracking and polymerization) of high-molecular-mass products.

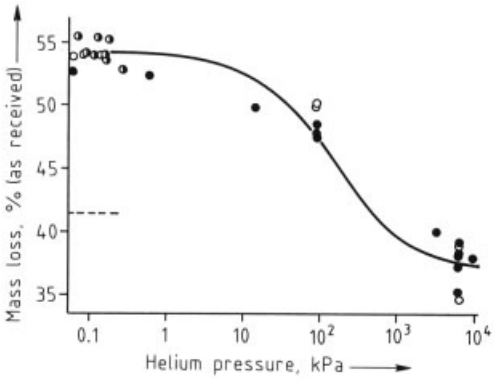

Figure 20. Effect of pressure on pyrolysis mass loss from a bituminous coal (final temperature: 1000 °C; mean particle diameter: 70 µm; residence time: 5 – 20 s). Nominal heating rate, ° C/s:
• 650 – 750; ◗ 3 000; ○ 10 000; - - - - - Proximate volatile matter (ASTM).

Cracking and polymerization can occur within and between the coal particles. These reactions are usually subject to surface catalytic effects and to the nature of the medium to which the primary products are exposed [91].

Volatile material can be driven from the interior of a coal particle by increasing the internal pressure or by reducing the external (ambient) pressure. High pressure gradients accelerate volatilization, thus reducing residence time within the particle, which decreases the probability of secondary reactions and char formation. This explains the effect of external pressure (shown in Fig. 20) and of high heating rates, which result in rapid heat transfer and a high rate of thermal decomposition. This higher rate of formation of volatiles increases the internal pressure and reduces char formation.

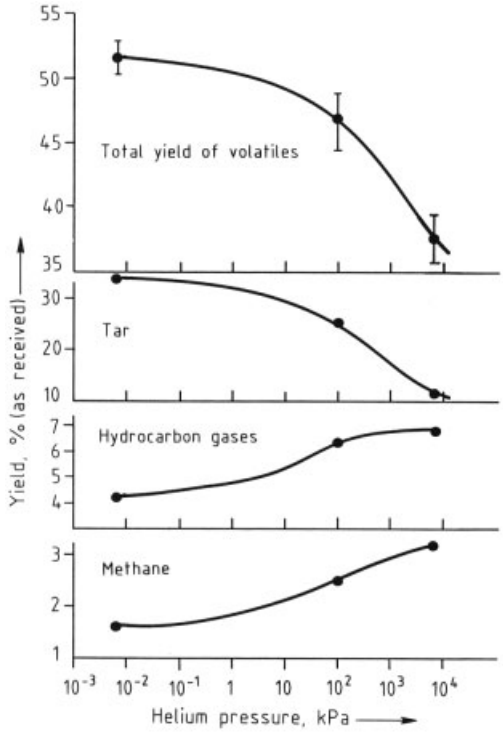

Figure 21. Effect of pressure on the yield and distribution of volatile products from a bituminous coal (final temperature: 1000 °C; mean particle diameter: 74 µm; heating rate: 1000 °C/s; residence time: 2 – 10 s) [90]

Minimizing interparticle secondary reactions also is important for a maximum yield of volatiles. This is achieved by minimizing particle loading. In entrained-flow reactors, particle loading is far lower than in fixed-bed reactors and the heating rate is far higher. Fluidized-bed reactors are intermediate between these two types.

When pyrolyzed, low-rank coal yields much less tar than bituminous coal (Section 3.2.5). In part this is due to the presence of high molecular mass aromatic units in high-rank coal, which

form liquid rather than gaseous products. Another factor is the presence of cations in low-rank coal. Acid washing removes these cations and substantially increases the yield of volatiles (see Fig. 23). This is attributed to a higher tar yield and the elimination of char-forming reactions catalyzed by the highly dispersed cations (principally Ca^{2+}).

Table 9. Effect of coal type on yields of gaseous pyrolysis products [90]

Element/ compound	Composition, wt %		
	Lignite	Subbituminous	Bituminous
Coal analysis (as received)			
C∗	71.2	73.5	77.7
H∗	4.6	5.8	5.5
N∗	1.1	1.2	1.5
S∗	1.3	0.81	6.1
O∗, by difference	21.8	18.7	9.2
H_2O	6.8	34.7	1.4
Ash, dry	10.6	9.1	11.5
VM∗	44.3	–	11.5
Product analysis ∗			
CO	8.5	5.8	2.8
CO_2	10.1	11.3	1.4
H_2O	11.6	–	7.4
H_2	0.60	1.1	1.2
N_2	–	–	–
CH_4	1.6	4.8	2.9
C_2H_4	0.67	0.24	0.95
C_2H_6	0.24	0.89	0.58
C_3H_8	0.17	0.47	1.5∗∗
Total	33.5	24.6	18.7

∗ Dry, ash-free.
∗∗ Includes C_3H_6.

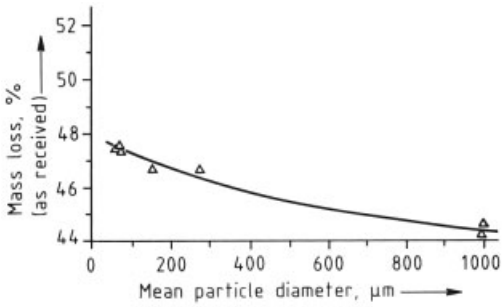

Figure 22. Effect of particle size on pyrolysis mass loss from a bituminous coal (pressure: 101.3 kPa of He; final temperature: 1000 °C; heating rate: 650 – 750 °C/s; residence time: 5 – 20 s) [90]

Various options are available when selecting the operating conditions of a pyrolysis reactor (see Chap. 6) [93]. Residence times of 0.1 s or less are attainable only in entrained-flow reactors; at the other extreme, residence times of 100 000 s or more are practical only in fixed-bed reactors. At low temperature (ca. 525 °C), the total yield of volatiles is low because thermal decomposition is incomplete. Entrained-flow conditions minimize secondary reactions and give the highest overall yield. Furthermore, high molecular mass products are stable at low temperature, and the product distribution favors liquids. At ca. 1000 °C, volatilization is complete, with a correspondingly high yield. Again, entrained-flow conditions reduce secondary reactions and increase overall yield. However, even under dilute conditions in entrained-flow reactors, high temperature causes substantial cracking and gives gas as the principal volatile product.

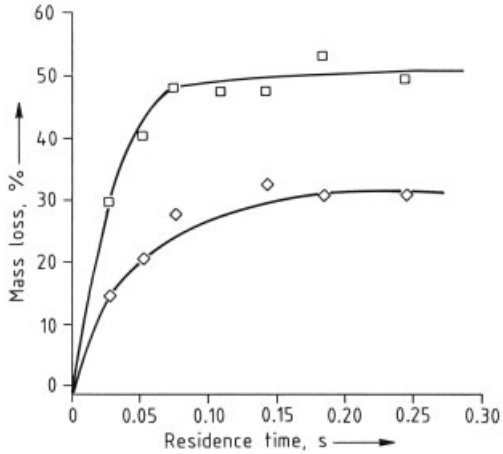

Figure 23. Effect of cations on pyrolysis mass loss from a lignite (pressure: 101.3 kPa of N_2; final temperature: 1173 K; mean particle diameter: 70 µm; heating rate: 10^4 °C/s)
□ Acid-washed; ◇ As received [92]

4. Kinetics

The practical importance of the kinetics of the processes described in Chapter 3 depends on the reactor type [90], [94], [95]. In *fixed-bed reactors*, such as coke ovens where the heating rate is ca. 100 °C/min or less, kinetics are not important. The time required for complete pyrolysis is much less than the time required for heating

to the final temperature, which determines yield and product distribution.

On the other hand, residence time at pyrolysis temperatures is important in *entrained-flow reactors,* such as pulverized-coal combustors and some modern pyrolysis reactors, which are increasingly important.

A simple kinetic treatment for determining the rate constants that define the size of a reactor or reaction zone is to assume first-order decomposition of the volatile matter in the coal:

$$\frac{\mathrm{d}V}{\mathrm{d}t} = k\left(V' - V\right) \tag{1}$$

where k = rate constant (s^{-1}), V = fraction of volatiles evolved at time t, and V' = total amount of volatiles in the coal.

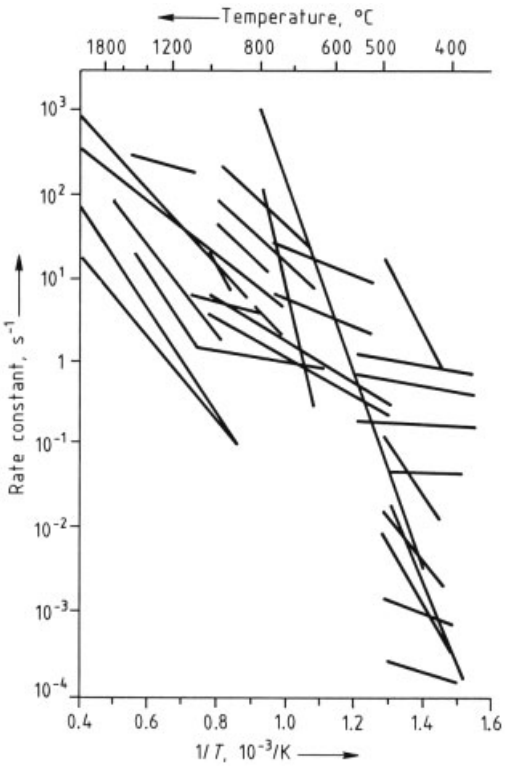

Figure 24. Arrhenius diagram for first-order kinetics of coal pyrolysis [94]

The temperature dependency of the rate constant (Arrhenius diagram) is given in Figure 24 for a number of experimental studies using different coals and reaction conditions [94]. According to the Arrhenius expression $k = A\exp\left(-E/RT\right)$, where A is the preexponential or frequency factor, E the activation energy, R the ideal gas law constant, and T the temperature; the rate constant is only a function of T. The large discrepancies observed result mostly from differences in measured reaction conditions and variations in coal rank.

The magnitude of the effect of coal rank is debatable. In one report the kinetics of pyrolysis were found to be insensitive to changes in rank from lignite to bituminous coal [94]. A possible explanation for this observation is that rapid pyrolysis rates are not determined by the chemical process of bond rupture, but by the physical processes of heat or mass transfer.

The modeling of the complex pyrolysis process by a single first-order reaction is an oversimplification. Many alternative kinetic schemes, proposed to represent more closely the actual mechanism of coal pyrolysis, range from two independent parallel reactions corresponding to the postulated two components in coals [96], to 42 reactions of the postulated 14 different functional groups in coal [97].

A reasonable compromise, which incorporates process complexity and is a mathematically adequate yet physically reasonable kinetic model [98], is based on a hierarchy of bond energies in the coal structure (see Fig. 1). The pyrolysis is assumed to consist of a large number, i, of independent chemical reactions representing the rupture of various bonds within the coal structure:

$$\frac{\mathrm{d}V_i}{\mathrm{d}t} = k_i\left(V_i' - V_i\right) \tag{2}$$

where the symbols are analogous to those of Equation (1), with the subscript i denoting a particular reaction. The rate constants k_i are assumed to be of Arrhenius form and to differ only in the values of their activation energies. The number of reactions is assumed to be large enough to permit the activation energies E_i to be expressed as a continuous Gaussian distribution function

$$\mathrm{f}\left(E\right) = \frac{1}{\sigma\left(2\pi\right)^{1/2}}\exp\left[-\frac{\left(E - E_0\right)^2}{2\sigma^2}\right]$$

with

$$\int_0^\infty \mathrm{f}\left(E\right)\mathrm{d}E = 1$$

where E_0 = mean activation energy and σ = standard deviation.

The term $f(E)\,dE$ represents the fraction of potential volatiles that have an activation energy between E and $E + dE$,

$$\frac{dV'}{V'} = f(E)\,dE \tag{3}$$

The total amount of volatiles remaining in the coal is obtained by combining the contribution from each reaction, i.e., by integrating Equation (2) over all values of E by using Equation (3):

$$\frac{V'-V}{V'} = \int_0^\infty \left(\exp - \int_0^t k\,dt \right) f(E)\,dE \tag{4a}$$

or

$$\frac{V'-V}{V'} = \frac{1}{\sigma(2\pi)^{1/2}} \cdot \int_0^\infty \exp\left[-A \int_0^t \exp\left(-\frac{E}{RT}\right) dt - \frac{(E-E_0)^2}{2\sigma^2} \right] dE \tag{4b}$$

Equation (4) shows that in addition to the temperature – time history, four parameters are needed to correlate coal pyrolysis data. The total yield of volatiles, V', should be determined experimentally and is not necessarily equal to the ASTM volatile matter content.

This model represents a significant improvement over the single first-order reaction scheme. Curves drawn through the experimental points in Figures 16 and 17 were obtained by using this model.

If, for simplicity and convenience, the single first-order reaction model of coal pyrolysis is assumed, the treatment that in the past has been applied to complex continuous reaction mixtures [99] and more recently to model coal liquefaction [100] should be used. It results in a temporal distribution of activation energies rather than the spatial distribution just described. It considers coal to be composed of constituents C_i; pyrolysis results in products P_i,

$$\Sigma C_i \overset{[ar]}{\underset{i}{}}$$

where the pyrolysis of each constituent is a first-order reaction:

$$-\frac{dC_i}{dt} = k_i(T)\,f(C_i)$$

Because C_i and P_i are not identifiable chemical species, a grouping procedure is used:

$$\Sigma C_i = V$$

The first-order rate expression, analogous to Equation (1), becomes [101]

$$-\frac{dV}{dt} = k(T,V)\,f(V)$$

Thus, the first-order rate constant k is not only a function of temperature, but also of the concentration of the volatiles, i.e., of conversion or time of pyrolysis. Limited experimental evidence [102] suggests that k decreases with conversion because of an increase in the observed activation energy. This can be interpreted as the intuitively reasonable initial predominance of the rupture of relatively weak bonds in the coal structure.

The main conclusion from this brief analysis is that some of the variation in rate constants observed in Figure 24 can be attributed to the fact that the constants may not have been compared at the same level of conversion.

For optimization of product distribution, having rate parameters for the total yield of volatiles and for individual compounds is desirable. The principal issue is the number of reactions necessary to describe the pyrolysis. In particular, determination of tar evolution rates is difficult because the rates are dependent on the transport properties of the experimental system, as discussed in Chapter 3. In one report these rate constants seem to correlate equally well the pyrolysis of lignite, subbituminous coal, and bituminous coal [94]. This means that pyrolysis rates may be dependent only on the quantity of the various functional groups in the coal and not on its origin or type. This conclusion, if established as generally valid for chemically rather than physically controlled pyrolysis processes, would simplify the design of pyrolysis reactors.

5. Hydropyrolysis

Most coal conversion processes involve hydrogen transfer reactions; pyrolysis involves *hydrogen redistribution.* Free radicals formed by bond rupture usually capture hydrogen atoms

(see Fig. 2). Under entrained-flow pyrolysis conditions, the total yield of volatiles is typically 20 – 40 % higher than the ASTM volatile content of the coal, i.e., at least 30 – 40 wt % of the product is a solid. The available hydrogen is used primarily to stabilize gaseous radicals, such as methyl or ethyl, leaving the higher molecular mass radicals to polymerize and form char.

To reduce char formation and increase volatile yield, the coal is pyrolyzed in the presence of hydrogen (see Fig. 25) [103]. The effect is more pronounced at higher hydrogen pressure, but is not observed if pyrolysis in an inert atmosphere is followed by exposure to hydrogen. Coal can be converted completely to methane in a few seconds at 950 °C and 50 MPa of hydrogen [104]. These effects are not fully understood, but seem to resemble the effect of hydrogen in coal liquefaction, where external hydrogen stabilizes the free radicals formed by bond scission and inhibits char formation. Another possible role of hydrogen in hydropyrolysis is to attack the highly reactive, but transient nascent carbon sites. Aromatic carbon radicals are formed by cleavage of functional groups. Hydrogen reacts with these active sites eventually forming methane, and, thus, the overall yield of volatiles is increased.

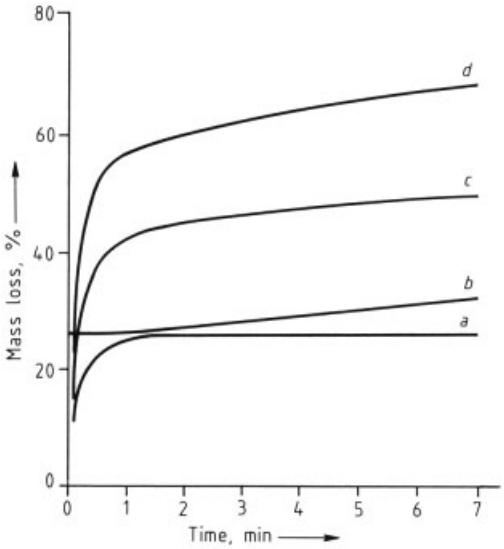

Figure 25. Effect of hydrogen pressure during pyrolysis on the yield of volatiles from a bituminous coal
[a] Coal: 3.5 MPa of N_2. [b] Char: 3.5 MPa of H_2. [c] Coal: 3.5 MPa of H_2. [d] Coal: 6.9 MPa of H_2 [103]

The effects of hydrogen pressure on the yield and distribution of pyrolysis products is illustrated in Figures 26 and 27 (see also Table 10). As discussed previously (see Section 3.2.3), an increase in the pressure of an inert gas reduces the yield of volatiles. With hydrogen, however, the yield first drops (see Fig. 26), and then increases with an increase in pressure.

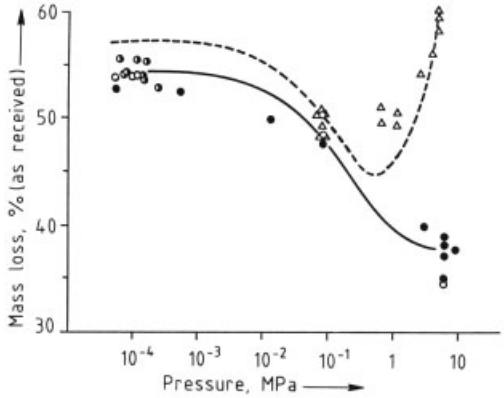

Figure 26. Effect of helium and hydrogen pressure on pyrolysis mass loss from a bituminous coal (final temperature: 1000 °C; mean particle diameter: 70 µm; residence time: 5 – 20 s)
○ He (10^4 °C/s); ◗ He (3000 °C/s); ● He (650 – 750 °C/s); △ H_2 (750 °C/s) [90]

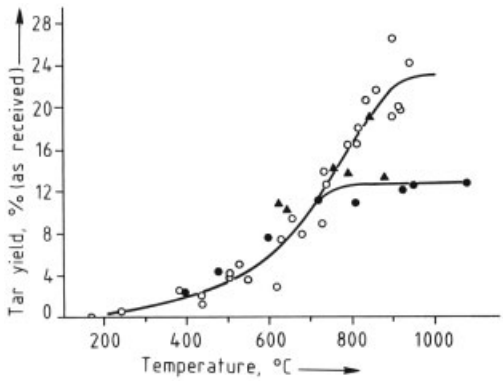

Figure 27. Comparison of tar yields from pyrolysis and hydropyrolysis of a bituminous coal (mean particle diameter: 74 µm; heating rate: 1000 °C/s)
○ 0.1 MPa of He; ● 6.9 MPa of He; ▲ 6.9 MPa of H_2 [90]

Experimental evidence suggests that the principal effect of H_2 is not in preventing secondary char-forming reactions (see Fig. 27). The yield of tar under the experimental conditions employed decreased under high hydrogen pressure.

Table 10. Yields from pyrolysis and hydropyrolysis of bituminous coal [a] [90]

Product	Yield, % (as received)					
	0.1 MPa of He [b]		6.9 MPa of He [b]		6.9 MPa of H_2 [c]	
	Bituminous [d]	Lignite [e]	Bituminous [d]	Lignite [e]	Bituminous [d]	Lignite [e]
CO	2.4	9.4	2.5	9.0	–	7.1
CO_2	1.2	9.5	1.7	10.6	1.3	8.5
H_2O	6.8	16.5	9.5	12.9	–	16.0
CH_4	2.5	1.3	3.2	2.5	23.2	9.5
C_2H_4	0.8	0.6	0.5	0.6	0.4	0.2
C_2H_6	0.5	0.2	0.9	0.2	2.3	1.4
$C_3H_6 + C_3H_8$	1.3	–	0.7	–	0.7	–
Other $HC_{(g)}$ [f]	1.3	–	1.6	–	2.0	–
Light $HC_{(l)}$ [f]	2.4	–	2.0	–	5.3	–
Tar	23	5.4	12	3	12	8
Char	53.0	56.0	62.4	59.8	40.2	48.5

[a] Final temperature: 850 – 1070 °C; mean particle diameter: 74 µm; heating rate: 1000 °C/s.
[b] Pyrolysis.
[c] Hydropyrolysis.
[d] Residence time: 2 – 20 s.
[e] Residence time: 3 – 10 s.
[f] HC = Hydrocarbon.

Table 11. Early pyrolysis processes [105]

Pyrolysis process	Final temp., °C	Products	Processes
Low temp.	500 – 700	more coke, high tar yield	Rexco (700 °C), cylindrical vertical retorts; coalite (650 °C), vertical tubes
Medium temp.	700 – 900	more coke, high gas yield, domestic briquettes	town gas (obsolete) and coke, phurnacite, using low-volatile steaming coal, pitch-bound briquettes carbonized at 800 °C
High temp.	900 – 1050	hard, unreactive coke for metallurgical use	foundry coke (900 °C); blast furnace coke (950 – 1050 °C)

In this instance, the detrimental effect of a high external pressure overcame the beneficial stabilizing effect of hydrogen. In the case of a bituminous coal and a lignite, the overall yield enhancement was due to an increase in methane production (see Table 10).

Pyrolysis and hydropyrolysis processes are depicted in Figure 28 from a kinetic viewpoint (see also Fig. 2). On thermal decomposition, the pores of coal particles are filled with reactive (e.g., radicals) and less reactive volatiles. The latter escape unchanged, whereas the former are either quenched and escape, are cracked, or polymerize. Active char (or coke) contains active sites that react with H_2 to give methane. The deactivated sites contribute little or nothing to the overall volatile yield. The corresponding kinetic expressions are given below [90]:

Pyrolysis:

$$V' = V''_{\mathrm{nr}} + V''_{\mathrm{r}} \frac{K_{\mathrm{c}}}{K_{\mathrm{c}} + k_1} = V''_{\mathrm{nr}} + V''_{\mathrm{r}} \frac{1}{1 + \frac{k_1 P}{k_{\mathrm{c}}}}$$

Hydropyrolysis:

$$V' = V''_{\mathrm{nr}} + V''_{\mathrm{r}} \frac{K_{\mathrm{c}} + k_2 P_{\mathrm{H_2}}}{K_{\mathrm{c}} + k_2 P_{\mathrm{H_2}} + k_1} + k_3 P_{\mathrm{H_2}}$$

The curves in Figure 26 were obtained by using these expressions.

6. Pyrolysis Processes

Pyrolysis and hydropyrolysis give gaseous, liquid, and solid products; a further hydrogenation step may be required. Fixed-bed processes (low heating rate) give the highest quality coke. They predominate commercially and

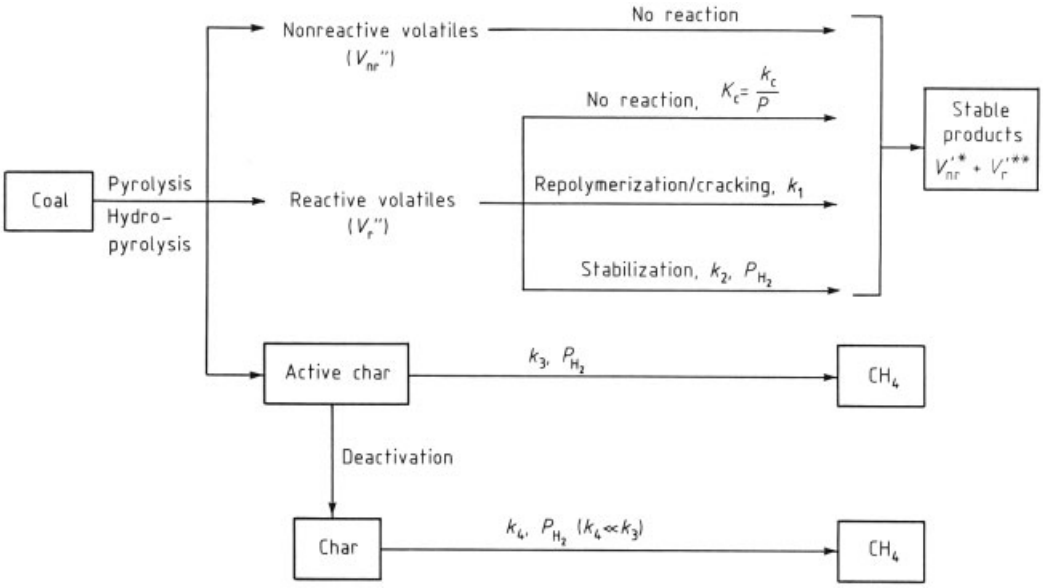

Figure 28. Kinetic model for coal pyrolysis and hydropyrolysis [90]
$* V_{nr} = V'_{nr}$. $** V'_r \leq V''_r$.

are often referred to as carbonization processes. Dilute-phase fluidized-bed and entrained-flow processes (high heating rate), suitable for high yield of gaseous and liquid products (see Chap. 3), are at various stages of precommercial development.

Pyrolysis processes can be classified as early (Table 11) [105] or contemporary (advanced). The *early processes* generally involved low-cost systems and were intended to produce solids (coke or char) rather than liquids.

Contemporary processes, on the other hand, have been designed to maximize production of liquid by using conditions (e.g., dilute phase, pressurized H_2, and rapid heating and quenching) that minimize secondary pyrolysis reactions. This requires more expensive equipment and severer operating conditions [93], [106].

6.1. Low-Temperature Pyrolysis Processes

6.1.1. Early Processes

Low-temperature pyrolysis (below 700 °C) was developed mainly to supply gas for public lighting and smokeless (devolatilized) solid fuel for home use. Byproduct tars were valuable as chemical feedstocks and could be converted to gasolines, heating oils, and lubricants. Commercial low-temperature coal pyrolysis was utilized extensively in Europe, but was almost abandoned after 1945 as oil and natural gas became widely available. Rising oil prices have reactivated interest in this process.

The many low-temperature processes developed in Europe can be classified as batch or continuous, vertical or horizontal, and directly or indirectly heated (see Tables 12 [107] and Table 13).

Retorts or byproduct coke ovens were employed. Horizontal retorts (D retorts), 2.5 – 6 m long, were usually grouped together to form a bank heated by flue gas, with manual loading and unloading. Vertical retorts, both batch and continuous, were preferred in the United States and Europe because they permit gravity loading and unloading [108].

Rates of heat-transfer and carbonization were increased by using steel refractories as materials of construction. Internal heating was sometimes necessary to accelerate heat transfer by circulating hot gas. Only noncaking or weakly caking coals were used in continuous vertical retorts because caking coal softens and flows, forming large masses of coke that stick to the retort walls.

Table 12. Low-temperature batch processes [107]

	Process						
	Krupp–Lurgi	Brennstoff–Technik	Otto	Weber	Phurnacite	Parker retort	Rexco
Country (year)	Germany (1930s)	Germany (1940s)	Germany (1940s)	Germany (1940s)	United Kingdom (1940s)	United Kingdom (1920s)	United Kingdom (1930s)
Objective	char production	lumpy char from weakly coking coals	char and gas production	char production	smokeless domestic fuel	smokeless domestic fuel	smokeless domestic fuel
Plants (year)	Wanne-Eickel (1943)	Berlin-Neukölln (1943)	–	–	South Wales (1942)	Barnsley (1927), Bolsover (1936)	Nottingham, England (1936)
Yields							
Char	2.04 $\times 10^8$ kg/a	0.75 kg/kg dry coal	–	–	0.7 kg/kg coal	–	–
Tar	1.2×10^6 kg fuel oil/a 1.9×10^6 kg motor fuel/a	0.1 L/kg dry coal	–	–	0.06 L/kg coal –	0.08 L/kg coal –	0.08 L/kg coal –
Gas	105 L/kg coal	142 L/kg dry coal	–	–	–	125 L/kjg coal	unknown
Gas HV∗	–	(2.5–3) $\times 10^4$ kJ/m^3	–	–	–	26 000 kJ/m^3	5200 kJ/m^3
Reactor	vertical fixed-bed retort	vertical fixed-bed retort	vertical fixed-bed retort	vertical fixed-bed retort	vertical fixed- bed retort	vertical fixed- bed retort	vertical fixed-bed retort
Capacity	270 kg/d	10^5 kg briquettes/d	3600 kg/d	unknown	11 340 kg/d	0.3 kg/kg charge	50 000 kg/d
Heating	indirect hot gas recycling	indirect hot gas recycling	indirect heating by gas burned in flues	unknown	indirect heating by recycling hot combustion gas	indirect radiant heat by combustion gas	direct heat by combustion gas
Temperature range	560–620 °C	650 °C	–	–	–	650 °C	650 °C
Residence time	2–3 h	2–4 h	–	–	4 h	4.5 h	13.5 h

∗ HV = heating value.

Two noteworthy low-temperature pyrolysis processes are the Disco and Coalite processes. These were successful when most low-temperature processes failed [108], [109].

6.1.2. Contemporary Processes

Lurgi – Ruhrgas Process. The Lurgi–Ruhrgas process (see Fig. 29) for the low-temperature production of liquids from low-rank coals is in commercial use in Europe.

Crushed coal is rapidly heated in a mixer (f) to 450–600 °C by contact with recirculating char particles that have been heated by partial oxidation in an entrained-flow reactor. The gas from the mixer is freed of particulate matter in a cyclone (h) and is passed through a series of condensers (i) to collect the liquid products; the latter are hydrogenated to yield stable products.

The char constitutes ca. 50 % of the products, whereas liquid yield is ca. 18 %; the remaining 32 % is gas with a heating value of $(26-31)\times 10^6$ J/m^3.

Table 13. Low-temperature continuous processes [107]

	Process			
	Coalite & Chemical Products	Lurgi – Spuelgas	Disco	Koppers continuous vertical ovens
Country (year)	United Kingdom (1950s)	Germany (1930s)	United States (1930s)	Germany (1940s)
Objective	char, liquid fuels	char, automotive fuels	lump char	char
Plants (year)	Bolsover, England (1952)	Offleben, Germany, Japan (1941), New Zealand (1931), Lehigh, North Dakota (1940)	Pittsburgh, PA.	Kattowitz (1945)
Yields				
Char	0.74 kg/kg coal	0.45 kg/kg briquette	6.8×10^5 kg/d	1.5×10^4 kg/d-retort
Tar	0.06 L/kg coal	0.125 kg/kg briquette	–	2.1×10^3 kg/d-retort
Gas	111 L/kg coal	144 L/kg briquette	–	0.32 L/kg coal
Gas HV∗	26 000 kJ/m^3	8380 kJ/m^3	–	1.8×10^4 kJ/m^3
Reactor	Parker vertical retort	continuous vertical retort	continuous rotating horizontal retort	continuous vertical retort
Capacity	0.3 kg coal charge	2.7×10^5 kg briquettes/d	1.4×10^5 kg/d	25 000 kg/d
Heating	radiant heat supplied by combustion gas	direct contact with combustion gas	indirect heating by circulating hot combustion gas through flues built in hearths	indirect and direct heating by passing hot gases alternately upward and downward through flues and regenerators
Temperature range	combustion chamber held between 600 and 700 °C	600 – 700 °C	550 °C	700 °C
Residence time	coal carbonized for 4 h	20 h	1.5 h	21 h/retort

∗ HV = heating value.

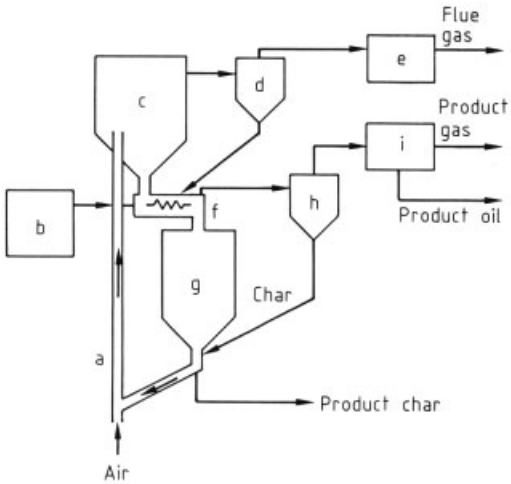

Figure 29. Lurgi – Ruhrgas process
a) Transport reactor and lift line; b) Coal preparation; c) Collecting bin; d) Cyclone; e) Heat recovery; f) Mixer – carbonizer; g) Surge hopper; h) Cyclone; i) Condenser

Toscoal Process. The Toscoal process, an adaptation of the Toscoal II oil – shale retorting process, was developed at a pilot plant near Golden, Colorado, with a capacity of 22.5×10^3 kg/d.

Crushed dry coal is preheated in lift pipes with flue gas from a ball heater and is fed to a rotating pyrolysis drum containing hot ceramic balls. Pyrolysis at ca. 500 °C produces char and hydrocarbon vapors. The char is separated by a trommel screen and withdrawn. The balls are reheated and returned to the drum. Pyrolysis products are cooled and condensed to recover light gases and coal liquids.

6.2. High-Temperature Processes (Coke-Making)

6.2.1. Coke Ovens

High-temperature carbonization is employed primarily to produce blast-furnace and foundry cokes for use in iron and steel manufacture. The process is carried out in coke ovens at ca. 900 °C.

In the 16th century, coke was found to burn hotter and cleaner than coal [110]. The earliest type of coke oven, the so-called *beehive oven*,

was developed in the 1850s; its main characteristic is that the heat necessary for coking is produced by burning the volatile constituents within the oven [111], [112]. All the gaseous and liquid byproducts are lost, together with large amounts of heat.

Beehive ovens have persisted well into the 20th century with progressive improvements in design and operation, such as heat recovery systems and vertical flues to supply process heat to the coking chamber.

A parallel development during the 19th century led to *slot ovens.* At first, these were simple rectangular firebrick structures holding a layer of coal ca. 1 m deep; later they were equipped with external heating flues [87]. In a further development, volatile byproducts were collected and used as fuel.

By the 1940s, the basic design of *modern coke ovens* had been developed. The ovens were ca. 12 m long, 4 m high, and 0.5 m wide, equipped with doors on both sides. The air supply was preheated by the hot exit gas. The recovery of the waste heat created higher temperatures and increased coking rates [113]. Since the 1940s, the process has been mechanized and the construction materials have been improved without significant design modifications. Assemblies may contain up to 60 ovens as large as 14 m long and 6 m high. Because of heat transfer considerations, widths have remained 0.3 – 0.6 m. Each oven in the battery holds up to 30 t of coal and operates on a 15-h cycle [114].

The coking process illustrated in Figure 30. The coal is conveyed from storage bunkers (a) to blending (b) and crushing (c) units, and then to the coke oven bunker (d). The ovens are loaded from a mobile larry car (e) located on top of the battery. Coking takes place in completely sealed ovens (f). The byproduct vapors and gases are collected and sent to recovery and processing units.

After carbonization, the oven doors are opened and the ram (g) pushes the red-hot coke into a quenching car. The coke is transported to the quenching unit (h) and then dumped on a sloping wharf (i) for cooling and drying. The dry coke is transferred to the screening and loading unit (j).

6.2.2. Coke Properties

The properties of coke depend on the coal and the processing conditions employed. The particle size, mineral matter, and moisture content, which can be controlled, affect the bulk density of the coal. Rank and maceral composition determine the fluidity of the plastic phase formed at intermediate temperatures. Significant process conditions include the wall temperature, coking time, and oven width; the heating rate and final temperature also affect coke structure (see Section 2.4).

The coal is prepared for coking by washing, drying, sizing, preheating, briquetting, and blending, with or without addition of pitch or breeze (fine coke particles) [111]. Foreign matter is removed mechanically and particle size is adjusted. Preheating (up to ca. 600 °C) increases production by ca. 50 % and reduces coking time without loss of quality [111].

Briquetting, with or without binder, increases the strength of the coke produced and permits noncoking or poorly coking coals to be used to make metallurgical coke. Cold briquetting is performed below the softening point of the coal (ca. 400 °C) and hot briquetting above the softening point (450 – 520 °C) [113].

Coal blending improves coke properties and yield and permits the use of noncoking or weakly coking coals [115]. Blending may also prevent damage to ovens when high coking pressure develops.

6.2.3. Byproducts

Carbonization byproducts were important in the development of the organic chemicals industry, but their use has greatly diminished because of competition from petroleum-based products. The principal byproducts are tar and coke-oven gas. Gas produced in the early slot ovens was used for domestic heating. Coal tar can be processed by petroleum-refining techniques. High-temperature tar can be fractionally distilled; three oil cuts are taken and a solid residue remains [33], [116]:

1) *Light oil* ($bp < 200$ °C), composed primarily of benzene, toluene, xylenes, and styrene.

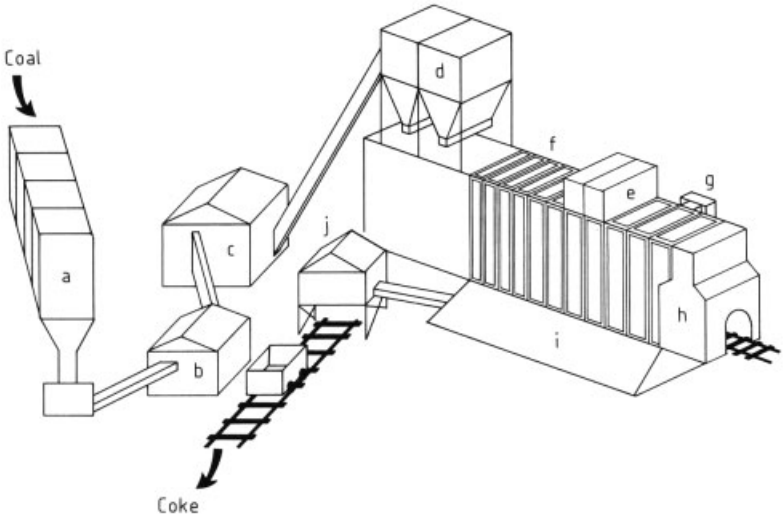

Figure 30. Coke-making process [87]
a) Storage bunkers; b) Blending plant; c) Crusher; d) Service bunker; e) Charging car; f) Coke oven; g) Ram machine; h) Cooling tower; i) Cooling wharf; j) oke screen

2) *Middle oil* (*bp* $< 370\,^{\circ}$C), containing tar acids (phenols), tar bases (e.g., pyridine, anilines, and quinolines), and neutral oils (naphthalene).
3) *Heavy oil* (*bp* $< 550\,^{\circ}$C), containing aromatic compounds, such as anthracene, phenanthrene, carbazole, and chrysene.
4) *Coal tar pitch* (solid residue), consisting of a wide range of polycondensed aromatic compounds; it can be used as a binder for coal briquetting.

Byproducts that are obtained more cheaply from petroleum are no longer recovered exhaustively from coal tar.

6.2.4. Formed-Coke Processes

The preparation techniques and reactor configurations employed by some of the coking processes developed since the 1960s are different. Among these, the formed-coke processes have been the most successful (see Table 14). The term "formed coke" is used to describe cokes obtained by the carbonization of briquettes of weakly coking or noncoking coals. These processes provide the possibility of using weakly coking or noncoking coals, of controlling coke size to improve blast furnace operation, and of continuous operation in a contained system.

Blast furnace tests of these cokes have been successful. They are usually higher in moisture, volatile matter, sulfur, and reactivity to carbon dioxide, and lower in abrasion resistance than conventional metallurgical cokes [87].

6.2.5. Environmental Aspects

Coking plants create environmental problems like other large installations producing dusty materials and gaseous and liquid effluents [86]. Gas and volatile emissions may arise from leaks. Coal and coke particulates may be discharged during loading and unloading, and water used for quenching may be contaminated.

Particulate, volatile, and gas emissions can be reduced by using sheds or traveling hoods with dust and fume collectors [87]. Polluted water must be treated in wastewater plants. In some plants, dry quenching is used instead of water quenching, and the coke is cooled by gas or steam in a closed system. A further advantage of this process is that energy can be recovered via sensible heat and combustible gases.

In addition to air pollution, machinery and working conditions are hazardous. Plant location and accident prevention require serious consideration.

Table 14. Formed-coke processes [87], [113], [117]

Process (Country)	Raw Materials	Process steps
FMC (USA)	noncoking or coking coals and tar	1) oxidation and carbonization of coal in fluidized beds 2) cold briquetting of char with tar 3) preheating and carbonization of briquettes
DKS (Japan)	noncoking, coking coals and pitch	1) cold briquetting of coal with pitch 2) carbonization of briquettes
HPNC (France)	weakly coking, coking coals and pitch	1) cold briquetting of coal with pitch 2) carbonization of briquettes
BFL (FRG)	high-volatile noncoking coals and coking coals or pitch	1) carbonization of noncoking coal 2) hot briquetting of char with coking coal or pitch 3) carbonization of briquettes
ANCIT (FRG)	low-volatile noncoking coals and coking coals	1) flash heating of low-volatile noncoking coal 2) hot briquetting of char with coking coal 3) reheating of briquettes
Sapozhnikov (former states of the USSR)	high-volatile weakly coking coals and low-volatile weakly coking coals	1) flash heating of coal 2) hot briquetting 3) carbonization of briquettes
CCC-BNR (USA)	high-volatile noncoking coals and coking coals on pitch as binder	1) preheating of high-volatile coal 2) hot briquetting of char with binder 3) carbonization of briquettes

Table 15. Characteristics of selected pyrolysis processes [93]

Process	Reactors	Temperature, °C	Pressure, MPa	Yield, % (dry)			
				Coke	Tar and oil	Gas	H_2O
COED	fluidized-bed	290 – 565	0.12 – 0.19	62	21	14	3
Occidental	entrained-flow	610	0.3	56	35	7	2
Rockwell	entrained-flow	845	3.5	46	38	16	–

6.3. Contemporary Hydropyrolysis and Pyrolysis Processes

The three processes shown in Table 15 illustrate the principles described in Chapter 3. Their main objective is production of liquid hydrocarbons, for which entrained-flow conditions are the most favorable. However, char or coke is still the principal product in all three processes; its efficient use is one of the main issues in their economic analysis. The use of hydrogen in the Rockwell process reduces char production to some extent but as discussed in Chapter 5, this is achieved primarily by increasing the yield of gas, not liquids.

The *Coal Oil Energy Development (COED) process* (see Fig. 31), is a part of the COGAS process for complete conversion of coal to gases and liquids. The residual char from pyrolysis reacts with H_2O and O_2 in a fluidized-bed gasifier (g) to produce the hot synthesis gas, a mixture of CO and H_2. The synthesis gas serves as fluidizing and heat-transfer medium in the pyrolysis reactors (p_1–p_4).

The conditions in the gasifier do not permit complete conversion to gases. Part of the char that leaves is recycled to the pyrolysis section, where it provides additional heat for the endothermic devolatilization. The entrained fine particles and the remaining char are burned in a combustor (c). The hot flue gas is also recycled to the pyrolysis section to provide more heat.

This brief description illustrates the principal technical issues involved in the evaluation of pyrolysis processes: the type of contact between solids and fluids; the method of supplying heat; and the amount and fate of the residual char. In this context, several features of the

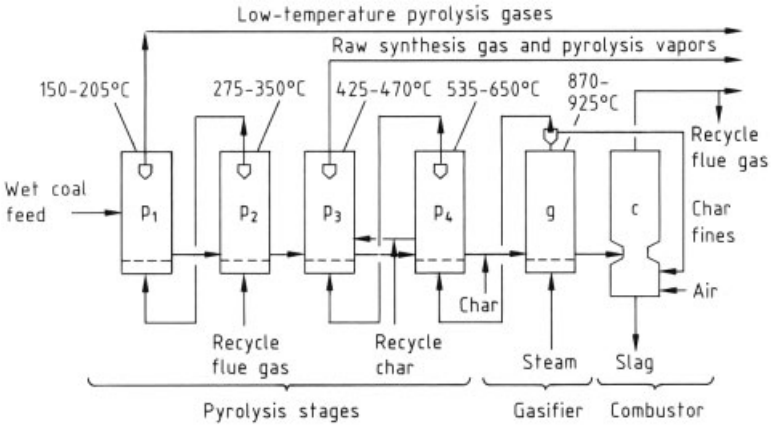

Figure 31. COED pyrolysis process [93]
$p_1 - p_4$) Pyrolysis reactors; g) Fluidized-bed gasifier; c) Combustor

COED process increase the production of liquids. The countercurrent flow of solids and gases (or vapors) and the increasing operating temperatures in the various pyrolysis stages help to reduce cracking by avoiding subsequent exposure to higher temperatures. The relatively dilute medium in the fluidized beds also helps to reduce the probability of secondary interparticle reactions.

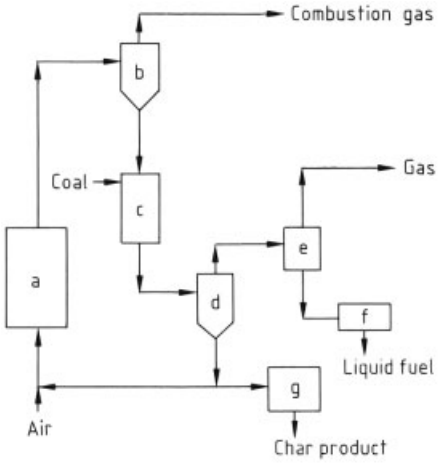

Figure 32. Occidental pyrolysis process [118]
a) Char burner; b) Cyclone; c) Reactor; d) Cyclone; e) Oil collection system; f) Liquid upgrading; g) Char desulfurization plant

The low final temperature favors production of liquids. However, the heating rates are relatively low and intraparticle secondary reactions may still occur. The use of internally generated and recirculated hot gas and char to provide energy for devolatilization has the advantage of efficiency but it introduces a degree of complexity.

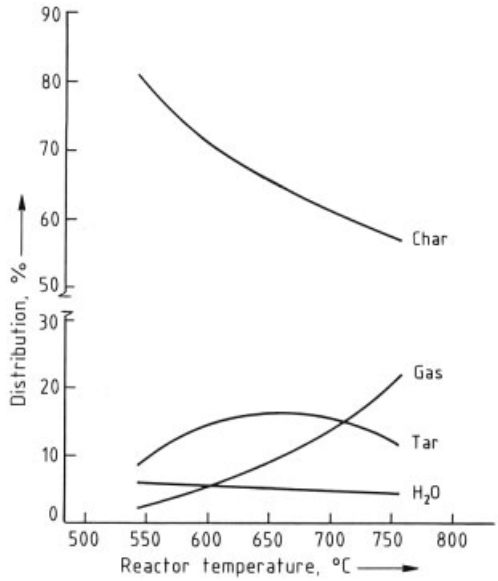

Figure 33. Effect of temperature on the distribution of products from a subbituminous coal in the Occidental process (pressure: 21 kPa of N_2; particle size: 80 % passing a 250 µm screen, 60 mesh; residence time: 1.5 s) [118]

The complete internal utilization of the coal is a virtue of the COED/COGAS process. The relatively low final pyrolysis temperature forms a

char that is highly reactive and, thus, suitable for gasification with steam. The resulting synthesis gas is a highly valuable and versatile product that can be converted, for example, to a natural gas substitute by methanation.

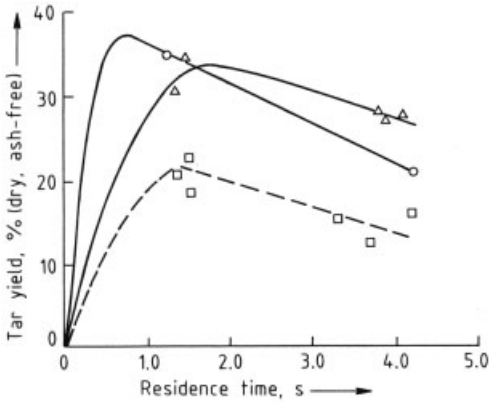

Figure 34. Effect of residence time on tar yield in the Occidental process (pressure: 21 kPa of N_2; particle size: 80 % passing a 250 µm screen, 60 mesh)
□ Subbituminous coal (649 °C); △ Bituminous coal (607 °C); ○ Bituminous coal (660 °C) [118]

Entrained-flow reactors are favored by a new generation of coal-conversion processes for several reasons. Both fixed-bed and fluidized-bed reactors need auxiliary equipment to avoid difficulties in processing of agglomerating bituminous coal. The fusion of coal particles during pyrolysis can block these reactors. In the COED process (Fig. 31), for example, where bituminous coal is the preferred feedstock because of its potentially high liquid yield, one of the stages (p_2) introduces oxygen cross-links into the coal structure and subsequently reduces or eliminates thermoplasticity (see Chap. 2), thus preventing agglomeration.

High throughputs are possible in entrained-flow reactors. The very dilute medium reduces the probability of particle agglomeration. Furthermore, rapid heating and a short residence time sharply reduce the importance of the plastic phase of pyrolysis of bituminous coals.

Finally, as discussed in Chapter 3, these conditions maximize the yield of volatiles and favor production of chars that may be sufficiently reactive to give valuable products, as in the COGAS process. Less reactive chars serve as fuel.

In the *Occidental process* (see Fig. 32) [118], the reactor operates at ca. 625 °C, thus maximizing production of liquids. Heat is supplied by mixing the coal with recirculated char that is made incandescent by partial combustion. The conditions for entrained flow require particles smaller than 200 µm (typically 80 % passing through a 60 mesh screen) and, thus, pulverization is necessary in contrast to fixed-bed

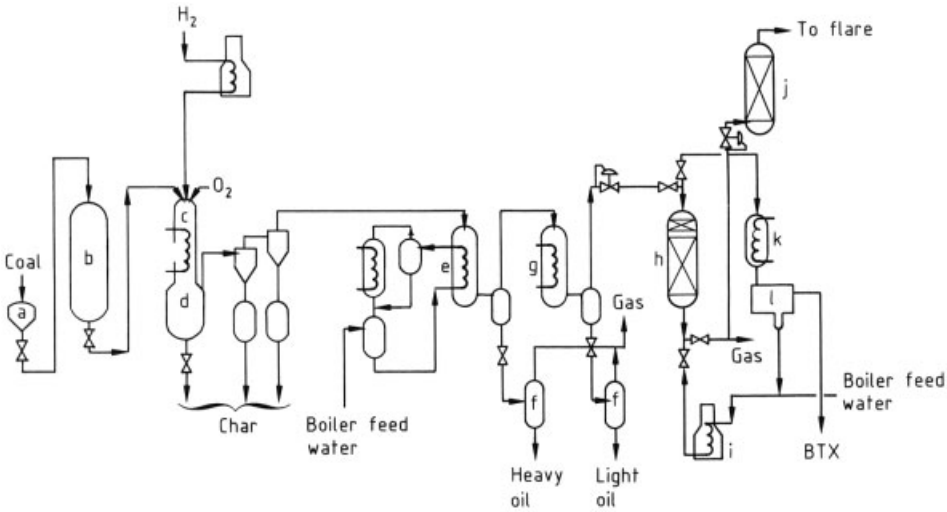

Figure 35. Rockwell hydropyrolysis process [119]
a) Loading feeder; b) Coal feeder; c) Reactor; d) Char – vapor separator; e) Heavy oil condenser; f) Flash drum; g) Light oil condenser; h) BTX adsorber; i) Boiler; j) Scrubber; k) BTX – steam condenser; l) Decanter

and fluidized-bed processes. The residual char is usually desulfurized for sale as "smokeless" fuel.

The effect of temperature and residence time on product distribution is shown in Figures 33 and 34.

The *Rockwell hydropyrolysis process* (see Fig. 35) incorporates the entrained-flow reactor as a minor component; other operations are similar to those encountered in petroleum refining. The hydrogen feed is preheated by partial combustion. After hydropyrolysis lasting 0.02 – 0.2 s, the products are rapidly quenched to prevent secondary char formation. A maximum liquid yield is obtained at an intermediate reactor temperature [119]. Changing hydrogen pressure affects product distribution, but not the yield of liquids; very high pressure is apparently needed to stabilize the reactive fragments in the absence of a solvent [119].

7. References

1. D. D. Whitehurst, T. O. Mitchell, M. Farcasiu: *Coal Liquefaction,* Academic Press, New York 1980.
2. T. Green, J. Kovac, D. Brenner, J. Larsen in R. A. Meyers (ed.): *Coal Structure,* Academic Press, New York 1982, pp. 199 – 282.
3. D. W. Van Krevelen: *Coal: Typology-Chemistry-Physics-Constitution,* Elsevier, Amsterdam 1961.
4. M. R. Khan, Ph. D. Thesis, The Pennsylvania State Univ., 1985.
5. R. Loison, A. Peytavy, A. Boyer, R. Grillot in H. H. Lowry (ed.): *Chemistry of Coal Utilization,* J. Wiley & Sons, New York 1963,Suppl. vol., pp. 150 – 201.
6. D. Hebermehl, F. Orywal, H. Beyer in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* 2nd Suppl. vol., J. Wiley & Sons, New York 1981, pp. 317 – 351.
7. J. T. Senftle, Ph. D. Thesis, The Pennsylvania State Univ., 1982.
8. D. J. Maloney, M. S. Thesis, The Pennsylvania State Univ., 1980.
9. M. R. Khan, M. S. Thesis, The Pennsylvania State Univ., 1985.
10. M. S. Lancet, F. A. Sim, *Am. Chem. Soc., Div. Fuel Chem., Prepr.* **26** (1981) no. 3, 167.
11. M. R. Khan, R. G. Jenkins, *Fuel* **63** (1984) 109 – 115.
12. M. Kaiho, Y. Toda, *Fuel* **58** (1979) 397.
13. D. J. Maloney, Ph. D. Thesis, The Pennsylvania State Univ., 1983.
14. A. F. Boyer: *Proc. International Conference on Chemical Engineering in Coal Industry,* Pergamon Press, Paris 1957, p. 141.
15. O. P. Mahajan, M. Komatsu, P. L. Walker, Jr., *Fuel* **59** (1980) 3.
16. B. S. Ignasiak, A. J. Szladow, D. S. Montgomery, *Fuel* **53** (1974) 12.
17. P. C. Painter, M. M. Coleman, R. W. Snyder, O. P. Mahajan et al., *Appl. Spectrosc.* **35** (1981) 106.
18. C. Rhoads, J. T. Senftle, M. M. Coleman, A. Davis et al., *Fuel* **63** (1984) 245 – 250.
19. M. R. Khan, R. G. Jenkins, *Fuel* **65** (1986) 1291.
20. M. R. Khan, R. G. Jenkins, *Fuel* **65** (1986) 1203.
21. S. K. Chakrabartty, S. Parkash, N. Berkowitz, *Fuel* **55** (1976) 270. G. F. Crewe, U. Gat, V. Dhir, *Fuel* **54** (1975) 20. A. C. Cunningham, W. F. Wyss, *Fuel* **46** (1967) 137. W. R. Epperly, H. M. Siegel: *Proc. of 11th Symposium Intersociety Energy Conversion Engineering Conference,* State Line, Nevada, Sept. 12 – 17, 1976, pp. 249 – 267. Y. Nishiyamo, Y. Tamai, *Fuel* **57** (1978) 559. J. W. Patrick, F. H. Shaw, *Fuel* **51** (1972) 69. R. H. Schlosberg, M. L. Gorbaty, R. L. Lang, *Fuel* **57** (1978) 424.
22. M. R. Khan, R. G. Jenkins: *Proc. International Conference on Coal Science,* Int. Energy Agency (IEA), Pittsburgh, Pa., 1983, pp. 495 – 498.
23. M. R. Khan, R. G. Jenkins, *Fuel Process. Technol.* **8** (1984) 307 – 311.
24. E. Audibert, *Fuel* **5** (1926) 229.
25. E. Audibert, L. Delmas, *Fuel* 6 (1927) 131 – 140, 182 – 189.
26. J. G. Bennett, *J. Inst. Fuel* **14** (1941) 175.
27. S. R. Illingworth, *Fuel* **1** (1922) 3 – 6, 17 – 19, 33 – 35, 65 – 67.
28. R. Thiessen, *Ind. Eng. Chem.* **24** (1932) 1032.
29. R. G. Atkinson, R. E. Brewer, J. D. Davis, *Ind. Eng. Chem.* **29** (1937) 840 – 844.
30. H. L. Riley, H. E. Blayden, J. Gibson: *Proc. Conference Ultrafine Structure of Coals and Cokes,* British Coal Utilization Research Assn. (BCURA), London 1943, pp. 176 – 231.
31. H. R. Brown, P. L. Waters, *Fuel* **45** (1966) 17 – 41.
32. W. Hirst, *Proc. Conference Ultrafine Structure of Coals and Cokes,* British Coal Utilization Research Assn. (BCURA), London 1943, pp. 80 – 94.

33. N. Berkowitz: *An Introduction to Coal Technology,* Academic Press, New York 1979.
34. D. H. Bangham, F. A. P. Maggs: *Proc. Conference Ultrafine Structure of Coals and Cokes,* BCURA, 1944, p. 118.
35. I. G. C. Dryden, M. Griffith, *BCURA Rev.* **134** (1954) 62.
36. D. Fitzgerald, *Trans. Faraday Soc.* **52** (1956) 362 – 369.
37. R. C. Neavel in M. L. Gorbaty, J. W. Larsen, I. Wender (eds.): *Coal Science,* **vol. 1,** Academic Press, New York 1982, pp. 1 – 19.
38. P. H. Given: *Coal Agglomeration and Conversion Symposium,* Morgantown, W.V., 1975.
39. L. Grainger: *Coke Oven Managers Association Yearbook,* Mexborough, U.K., 1975, p. 282.
40. J. Dartnell, *Ironmaking Steelmaking* **1** (1978) 18 – 24.
41. P. Ramdohr, *Eisenhüttenwesen* **1** (1928) 669.
42. C. E. Marshall, *Fuel* **24** (1945) 120.
43.
44. C. Abramski, M. T. Mackowsky in H. Freund (ed.): *Handbuch der Mikroskopie in der Technik,* **vol. 2,** Verlag Umschau, Frankfurt 1952, Part 1, p. 311.
45. B. Alpern, *Brennst. Chem.* **37** (1956) 194.
46. G. H. Taylor, *Fuel* **40** (1961) 465.
47. J. D. Brooks, G. H. Taylor, *Carbon* **3** (1965) 185.
48. H. Marsh, *Fuel* **52** (1973) 205.
49. J. W. Patrick, M. J. Reynolds, F. H. Shaw, *Fuel* **52** (1973) 198.
50. H. Marsh, F. Dachille, M. Iley, P. L. Walker, Jr. et al., *Fuel* **52** (1973) 253.
51. A. Grint, U. Sweitlik, H. Marsh, *Fuel* **58** (1979) 642.
52. A. Grint, H. Marsh, *Fuel* **60** (1981) 1115.
53. H. Marsh, P. L. Walker, Jr., *Fuel Process. Technol.* **2** (1979) 61 – 75.
54. H. Marsh, J. Smith in C. Karr (ed.): *Analytical Methods for Coal and Coal Products,* **vol. 2,** Academic Press, New York 1978, pp. 371 – 414.
55. H. Marsh, P. L. Walker, Jr. in P. L. Walker, Jr., P. A. Thrower (eds.): *Chemistry and Physics of Carbon,* **vol. 15,** Marcel Dekker, New York 1979, pp. 229 – 286.
56. E. Fitzer, K. Mueller, W. Schaefer, in P. L. Walker, Jr. (ed.): *Chemistry and Physics of Carbon,* **vol. 7,** Marcel Dekker, New York 1971, pp. 237 – 383.
57. J. D. Brooks, G. H. Taylor, *Chemistry and Physics of Carbon,* **vol. 4,** Marcel Dekker, New York 1968, pp. 243 – 286.
58. H. Marsh, R. C. Neavel, *Fuel* **59** (1980) 511.
59. I. C. Lewis, *Carbon* **20** (1982) 519.
60. W. J. Schmidt in H. Freund (ed.): *Handbuch der Mikroskopie in der Technik,* **vol. 1,** Verlag Umschau, Frankfurt 1957, Part 1, p. 147.
61. E. A. Rosauer: *Instruments for Materials Analysis,* Iowa State Univ. Press, Ames 1981.
62. J. L. White in J. O. McCaldin, G. Somorjai (eds.): *Progress in Solid-State Chemistry,* **vol. 9,** Pergamon Press, Oxford 1974, pp. 59 – 104.
63. R. A. Forrest, H. Marsh, *Carbon* **15** (1977) 348.
64. G. D. Mitchell, A. Davis, W. Spackman in R. T. Ellington (ed.): *Liquid Fuels from Coal,* Academic Press, New York 1977, pp. 255 – 270.
65. H. E. Newall, F. S. Sinnatt, *Fuel* **3** (1924) 424.
66. P. J. Street, R. P. Weight, P. Lightman, *Fuel* **48** (1969) 342.
67. M. Von Heimendahl: *Electron Microscopy of Materials – An Introduction,* Academic Press, New York 1980.
68. H. Marsh, C. Cornford in M. L. Deviney, T. M. O'Grady (eds.): *Petroleum Derived Carbons, Amer. Chem. Soc. Symposium Series 21,* Amer. Chem. Soc., Washington, D.C., 1976, pp. 266 – 281.
69. J. N. Murrell, E. A. Boucher: *Properties of Liquids and Solutions,* J. Wiley & Sons, New York 1982, pp. 92 – 95.
70. F. Reinitzer, *Monatsh. Chem.* **9** (1888) 421.
71. O. Lehmann, *Z. Phys. Chem. (Frankfurt am Main)* **4** (1889) 462.
72. G. Friedel, *Ann. Phys. (Paris)* **18** (1922) 273.
73. F. D. Saeva (ed.): *Liquid Crystals,* Marcel Dekker, New York 1979.
74. I. Mochida, K. Maeda, K. Takeshita, *Carbon* **16** (1978) 459.
75. I. C. Lewis, *Carbon* **16** (1978) 503.
76. J. L. White in M. L. Deviney, T. M. O'Grady (eds.): *Petroleum Derived Carbons, Amer. Chem. Soc. Symposium Series 21,* Amer. Chem. Soc., Washington, D.C., 1976, pp. 282 – 314.
77. J. Dubois, C. Agache, J. L. White, *Metallography* **3** (1970) 337 – 369.
78. T. Imamura, Y. Yamada, S. Oi, H. Honda, *Carbon* **16** (1978) 481.
79. K. J. Huettinger, *Carbon* **10** (1972) 5.
80. I. C. Lewis, L. S. Singer in P. L. Walker, Jr. (ed.): *Chemistry and Physics of Carbon,* **vol. 17,** Marcel Dekker, New York 1981, pp. 1 – 88.

81. T. Yokono, H. Marsh in H. D. Shultz (ed.): *Coal Liquefaction Products,* **vol. 1,** J. Wiley & Sons, New York 1983, pp. 125 – 138.
82. [3] pp. 445 – 452.
83. I. G. C. Dryden in H. H. Lowry (ed.): *Chemistry of Coal Utilization,* J. Wiley & Sons, New York 1963,Suppl. vol., pp. 262 – 268.
84. P. H. Given, *Fuel* **39** (1960) 147.
85. J. W. Patrick, M. J. Reynolds, F. H. Shaw, *Carbon* **13** (1975) 509.
86. F. Goodarzi, D. G. Murchison, *Fuel* **57** (1978) 273.
87. L. Grainger, J. Gibson: *Coal Utilization Technology, Economy and Policy,* Holsted Press, London 1981, pp. 137 – 150.
88. M. Forrest, H. Marsh in E. L. Fuller, Jr. (ed.): *Coal and Coal Products: Analytical Characterization Techniques, Amer. Chem. Soc. Symposium Series 205,* Amer. Chem. Soc., Washington, D.C., 1982, pp. 1 – 25.
89. R. H. Essenhigh in C. Y. Wen, E. S. Lee (eds.): *Coal Conversion Technology,* Addison-Wesley, Reading, Mass., 1979, Chapter 3.
90. J. B. Howard in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* 2nd Suppl. vol., J. Wiley & Sons, New York 1981, Chapter 12.
91. A. W. Scaroni, P. L. Walker, Jr., R. G. Jenkins, *Fuel* **60** (1981) 558.
92. M. E. Morgan, R. G. Jenkins in H. H. Shobert (ed.): *The Chemistry of Low-Rank Coals, Amer. Chem. Soc. Symposium Series 264,* Amer. Chem. Soc., Washington, D.C., 1984, p. 213.
93. R. F. Probstein, R. E. Hicks: *Synthetic Fuels,* McGraw-Hill, New York 1982.
94. P. R. Solomon, D. G. Hamblen, *Prog. Energy Combust. Sci.* **9** (1983) 323.
95. H. Juentgen, K. H. van Heek, *Fuel Process. Technol.* **2** (1979) 261.
96. N. Y. Nsakala, R. H. Essenhigh, P. L. Walker, Jr., *Combust. Sci. Technol.* **16** (1977) 153.
97. G. R. Gavalas, P. H. K. Cheong, R. Jain, *Ind. Eng. Chem., Fundam.* **20** (1981) 113.
98. D. B. Anthony, J. B. Howard, *AIChE J.* **22** (1976) 625.
99. R. Aris, G. R. Gavalas, *Philos. Trans. R. Soc. London, Ser. A* **260** (1966) 43.
100. A. Szladow, P. Given, *Ind. Eng. Chem. Process Des. Dev.* **20** (1981) 27.
101. S. V. Golikeri, D. Luss, *AIChE J.* **18** (1972) 277.
102. P. L. Walker, Jr., R. G. Jenkins, L. R. Radovic: "Importance of Active Sites for Char Gasification in Oxygen (Air) and CO2," Final Report, Gas Research Institute, Chicago, Oct., 1982.
103. J. L. Johnson in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* 2nd Suppl. vol., J. Wiley & Sons, New York 1981, Chapter 23.
104. F. Moseley, D. Paterson, *J. Inst. Fuel* **40** (1967) 523.
105. G. J. Pitt, G. R. Millward (eds.): *Coal and Modern Coal Processing: An Introduction,* Academic Press, New York 1979, p. 152.
106. L. Seglin, S. A. Bresler in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* 2nd Suppl. vol., J. Wiley & Sons, New York 1981, pp. 785 – 846.
107. M. R. Khan, T. Kurata: "The Feasibility of Mild Gasification of Coal: Research Needs," U.S. Dept. of Energy, Morgantown Energy Technology Center, Technical Note TM-85/4019, 1985, NTIS DE 85 013 625.
108. P. J. Wilson, Jr., D. D. Clendenin in H. H. Lowry (ed.): *Chemistry of Coal Utilization,* J. Wiley & Sons, New York 1963, pp. 150 – 201.
109. G. S. Pound, *J. Inst. Fuel* **25** (1952) 355.
110. M. O. Holwatey in R. A. Meyers (ed.): *Coal Handbook,* Marcel Dekker, New York 1981, Chapter 9.
111. F. Denig in H. H. Lowry (ed.): *Chemistry of Coal Utilization,* **vol. 1,** J. Wiley & Sons, New York 1945, Chapter 24.
112. J. Speight in J. Falbe (ed.): *The Chemistry and Technology of Coal,* Marcel Dekker, New York 1983, Chapter 12.
113. E. Ahland, G. Nashan, W. Peters, W. Weskamp in J. Falbe (ed.): *Chemical Feedstocks from Coal,* J. Wiley & Sons, New York 1982, pp. 12 – 77.
114. A. E. Williams, L. O. Smith, K. A. Koenig, K. A. Basciani, *Iron Steel Eng.* **60** (1983, May) 45.
115. N. Nakamura, Y. Togino, T. Adachi, *Ironmaking Steelmaking,* **5** (1978) no. 2, 49 – 60.
116. T. F. Edgar: *Coal Processing and Pollution Control,* Gulf Publ. Co., Houston 1983, Chapter 6.
117. W. Eisenhut in M. A. Elliott (ed.): *Chemistry of Coal Utilization,* 2nd Suppl. vol., J. Wiley & Sons, New York 1981,Chapter 14.
118. P. W. Chang, K. Durai-Swamy, E. W. Knell, *Coal Process. Technol.* **6** (1980) 20.
119. C. L. Oberg, A. Y. Falk, *Coal Process. Technol.* **6** (1980) 159.

Coal Tar → Tar and Pitch
Coated Fabrics → Textile Technology
Coatings → Paints and Coatings
Cobalamines → Vitamins

Cobalt and Cobalt Compounds

John Dallas Donaldson, Brunel University, Uxbridge, United Kingdom (Chap. 1, 2, 3, 4, 5, 6, 7, 8 and 9)
Detmar Beyersmann, University of Bremen, Bremen, Germany (Chap. 10)

1. Introduction 1
2. Physical Properties 3
3. Occurrence 5
4. Production 7
4.1. Concentration of the Ores 7
4.2. Extraction 8
4.2.1. Leaching of Cobalt Ores and Concentrates 8
4.2.2. Separation of Cobalt from Other Metal Ions in Leach Solutions 11
4.2.3. Electrowinning 13
5. Cobalt Powders 13
5.1. Production 13
5.2. Uses 13
5.2.1. Cemented Carbides 13
5.2.2. Cobalt Powders in Powder Metallurgy and Cobalt-containing Metal 14
6. Cobalt Alloys 14
6.1. General 14
6.2. High Temperature Alloys (Superalloys) 15
6.3. Magnetic Alloys 16
6.4. Hard Metal Alloys 19
6.5. Cobalt-containing High-Strength Steels 20
6.6. Electrodeposited Alloys 20
6.7. Alloys with Special Properties 20
7. Cobalt Compounds 21
7.1. Chemical Properties of Cobalt . . . 21
7.2. Commercially Important Cobalt Compounds 24
7.3. Industrial Applications of Cobalt Compounds 28
7.3.1. Glasses, Ceramics, and Refractories 28
7.3.2. Driers, Paints, Varnishes, and Dressings 29
7.3.3. Catalysts 29
7.3.4. Electroplating 31
7.3.5. Electronics, Solid-State Devices, and Batteries 32
7.3.6. Agriculture, Nutrition, and Medicine 33
8. Analysis 33
9. Economic Aspects 33
10. Toxicology 34
11. References 37

1. Introduction [1 – 4]

Cobalt, Co [*7440-48-4*], is a metallic element. Along with iron and nickel, it is one of the first-row transition-metal members of group VIII of the periodic table (old IUPAC recommendation; according to the new IUPAC recommendation, cobalt belongs to group 9). Its electronic configuration is [Ar] 3 d^7 4 s^2, giving it an atomic number of 27. The relative atomic mass of cobalt is 58.9332. Only one of its isotopes, ^{59}Co, is stable and occurs naturally; the other 12 known isotopes are radioactive and have the following mass numbers (half-lifes in parentheses): 54 (0.2 s), 55 (18.2 h), 56 (80 d), 57 (270 d), 58 (9 h), 58 (72 d), 60 (10.1 months), 60 (5.3 years), 61 (99 months), 62 (1.6 months), 62 (13.9 months), and 64 ($<$ 28 s).

The γ-rays emitted in the decay of ^{60}Co have energies of 1.17 and 1.33 MeV, and these taken with the 5.3-year half-life of the isotope provide a widely used source of radioactivity for use in food sterilization, radiography, and radiotherapy as an external source. The isotope is also used in chemical and metallurgical analysis and in biological studies as a radioactive tracer. The ^{57}Co isotope decays by electron capture to give ^{57}Fe, the most widely used isotope in γ-resonance (Mössbauer) spectroscopy [5].

Pure metallic cobalt has few applications, but its use as an alloying element and as a source of chemicals makes it a strategically important metal. End uses of cobalt-containing alloys include superalloys for aircraft engines, magnetic alloys for powerful permanent magnets, hard metal alloys for cutting-tool mate-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_281.pub2

rials, cemented carbides, wear-resistant alloys, corrosion-resistant alloys, and electrodeposited alloys to provide wear and corrosion-resistant metal coatings. Cobalt chemicals, among their many applications, are used as pigments in the glass, ceramics, and paint industries; as catalysts in the petroleum industry; as paint driers; and as trace metal additives for agricultural and medical use. About 36 % of the worldwide annual production of cobalt is converted to chemicals, whereas high-temperature and magnetic alloys account for 41 % and 14 % of the consumption, respectively. Detail on all aspects of the science and technology of cobalt in general reviews [1 – 4] can be supplemented by reference to published abstracts [6, 7].

History. Although very little cobalt metal was used until the 20th century, its ores have been used for thousands of years as blue coloring agents for glass and pottery. Blue glazed pottery found in Egyptian tombs and dated at ca. 2600 B.C. has been found to contain cobalt, as have Persian glass beads dating from 2250 B.C. The Portland vase in the British Museum provides evidence that cobalt pigments were used by Greek glassworkers around the beginning of the Christian era. Cobalt-containing materials were also used to impart blue coloration to Chinese pottery during the Tang (600 – 900 A.D.) and Ming (1350 – 1650 A.D.) dynasties and to Venetian glass produced in the first part of the 15th century. The brilliant blue pigment used for these purposes can be produced by fusing an ore containing cobalt oxide with potash and silica to produce a vitreous material called smalt, which is powdered to produce the pigment. The secret of making the pigment was apparently lost during the Middle Ages and rediscovered during the 15th century. LEONARDO DA VINCI was one of the first artists to use the rediscovered blue pigment when painting "The Madonna of the Rocks." In the 16th century, P. WEIDENHAMMER produced a blue pigment, which he called zaffre, from the silver – cobalt – bismuth – nickel – arsenate ores found in Saxony. The next development in the use of cobalt-containing materials was the discovery, early in the 18th century, that solutions containing bismuth and cobalt could be used as sympathetic inks. The invisible writing obtained by using the inks became green when heated, with the color change apparently due to the presence of cobalt. It was not until 1735 that G. BRANDT, a Swedish scientist, first isolated cobalt metal as an impure sample by reducing an ore; in 1780 T. O. BERGMAN showed that the metal was in fact an element. The use of cobalt as a metal dates from 1907, when E. HAYNES patented a series of cobalt – chromium alloys named Stellites that were the forerunners of modern superalloys. In 1930 it was shown that addition of cobalt to certain alloys of iron, nickel, and aluminum enhanced their properties as permanent magnets.

From the 16th to the 19th century, the world's supply of cobalt — mainly as smalt and zaffre — was produced in Norway, Sweden, Hungary, and Saxony. Outside of Europe cobalt-containing ores were worked in Burma from 1651 and in New Caledonia from 1864. Major developments in cobalt recovery in the 20th century have stemmed from the discovery of new ore bodies and improvements in the methods of winning the metal from ores. Cobalt-containing copper ores were discovered in Zaire in 1914, and extraction of cobalt from the ores began in 1924, whereas recovery of cobalt from pyrite roasting residues was started in Germany in 1926. Mining of cobalt – nickel – gold – silver ores began in Morocco in 1932, and production of cobalt from copper – cobalt ores began in Zambia in 1933. The Outokumpu copper – cobalt – zinc deposit was discovered in 1913 and has been exploited on a large scale since 1928. In 1940 the International Nickel Co. (INCO) introduced a method of recovering cobalt from nickel ores, and from 1952 to 1955 the production of cobalt by the Canadian companies, Falconbridge Nickel Mines and the Sherritt Gordon Mining Co., began. Developments include the commissioning in 1975 of a cobalt refining plant based on a combination of pressure leaching, solvent extraction, and electrowinning by the Sumitomo Metal Mining Co. in Japan, and centering of interest on Laterite ores, particularly in Australia.

Worldwide production of cobalt in 1997 was just over 28 000 t, but this is expected to increase to 34 000 t/a.

The production of cobalt is usually subsidiary to that of other metals, such as copper and nickel, and its output cannot be significantly increased without a corresponding increase in the markets of the primary metals. Another important factor in the economics of the supply of cobalt is the

fact that a large proportion of the world's ore comes from one country, Zaire, and is therefore dependent on the political stability of that country. Cobalt is regarded as a strategic metal and is stockpiled in some countries.

The name cobalt is derived from the Greek word κόβαλος or the German word kobold for goblin or evil spirit. The term kobold was applied by miners in the Harz mountains of Germany to an ore that gave no metal when smelted and that also produced highly toxic fumes of arsenic oxide.

2. Physical Properties [1 – 4]

The physical and mechanical properties of cobalt metal are susceptible to variations because of variations in the structures of metal samples arising from a slow cubic to hexagonal phase transformation. The data quoted in this section are for the highest purity metal available; details of the effects of impurities on the properties of cobalt are to be found in a review by W. Betteridge [8].

Cobalt exists in two allotropic modifications, a close-packed hexagonal ϵ-form stable below ca. 400 °C and a face-centered cubic α-form stable at high temperature.

The transformation temperature is 421.5 °C. The free energy change associated with the transformation is low, being ca. 500 J/mol for the $\epsilon \rightarrow \alpha$ change and ca. 360 J/mol for the $\alpha \rightarrow \epsilon$ change; this accounts for the slowness of the transformation. The mechanism of the phase change is martensitic and involves dislocation movements on the octahedral planes of the cubic lattice. Grain size affects the stability of the two allotropes, with finer grain size favoring the cubic high-temperature form. For this reason the fine grain sizes found in cobalt powders, sponge, thin films, and fibers are responsible for the retention of essentially cubic structures down to ambient temperature. The cubic form is also favored by the presence of a few percent of iron in the lattice. Table 1 shows the cell parameters and other properties of the allotropes. The density of the metal decreases with increasing temperature and shows an anomalous fall of ca. 0.15 % at the transformation temperature. The liquid density just above the melting point is 7.73 g/cm^3, and at 2000 °C it is 7.26 g/cm^3.

Table 1. Properties of the allotropic forms of cobalt

Hexagonal ϵ-form	
Cell dimensions	
(ambient temperature)	a = 0.250 71 nm
	c/a = 1.6233
(417 °C)	a = 0.2541 nm
	c/a = 1.631
Density (20 °C)	8.832 g/cm^3
Co – Co interatomic distances	0.294 – 0.251 nm
Stacking fault energy	
(20 °C)	31×10^{-7} J/cm^2
(370 °C)	20.5×10^{-7} J/cm^2
Cubic α-form	
Cell dimension	
(ambient temperature)	a = 0.35441 nm
(520 °C)	a = 0.35688 nm
(1398 °C)	a = 0.36214 nm
Density	
(ambient temperature)	8.80 g/cm^3
(1495 °C)	8.18 g/cm^3
Co – Co interatomic distance	0.251 nm
Stacking fault energy	
(500 °C)	13.5×10^{-7} J/cm^2
(710 °C)	18.5×10^{-7} J/cm^2

Thermal Properties. The *melting point* of cobalt is 1495 ± 2 °C, and its boiling point at normal pressure is 2800 ± 50 °C. The *vapor pressure* of cobalt varies from less than 10^{-5} Pa (at 1250 °C) to greater than 10^5 Pa (at 3200 °C). The *heats of fusion and vaporization* of cobalt are 17.2 and 425 kJ/mol, respectively. The *thermal conductivity* (λ) of the metal falls steadily with increasing temperature; there is no marked change in conductivity at the transformation temperature, but the data show a minimum in the region of the Curie temperature at 1121 °C. The values of λ at 0, 100, and 800 °C are 102, 85, and 50 W m^{-1}, respectively. The *heat capacity* of cobalt rises steadily with increasing temperature. There is only a slight inflection in the c_p vs. T curve at the transformation temperature, but a large change does occur at the magnetic transformation (see page 4). The *thermal expansion coefficient* of hexagonal cobalt depends on the orientation of the crystals and, for example, is 14.62×10^{-6} K^{-1} along the $\langle 0001 \rangle$ and 10.96×10^{-6} K^{-1} along the $\langle 1120 \rangle$ direction (at 20 °C). The average value for polycrystalline cobalt is 12.14×10^{-6} K^{-1}. The curve of the coefficient versus temperature shows a sharp change at the transformation temperature because of the

volume expansion of ca. 0.36 % associated with the phase change from hexagonal to cubic crystals. Cubic cobalt has a higher mean thermal expansion coefficient with values of (14.2, 15.7, 16.0, and 16.8) $\times 10^{-6}$ K^{-1} at 200, 400, 600, and 750 °C, respectively.

Mechanical Properties. The mechanical properties of cobalt are critically dependent on the purity of the metal and its thermal history. For well-annealed samples of high-purity cobalt, the *Vickers hardness* at normal temperature has values between 140 and 160 N/mm^2. The Vickers hardness of annealed cobalt decreases with increasing temperature to values of under 50 N/mm^2 at 750 °C with no marked change at the phase transformation. Microhardness studies on zone-refined single crystals of the metal show variations with crystal orientation and give values between 81 and 250 N/mm^2. Electrodeposited cobalt has high hardness (270 – 310 N/mm^2), presumably because of its fine grain size.

Measurements of elastic properties on polycrystalline bars of high-purity sintered metal annealed at 1000 °C give a *Young's modulus* of 211×10^3 MN/m and a *shear modulus* of 82 $\times 10^3$ MN/m at room temperature. The moduli decrease with increasing temperature, with the shear modulus showing an inflection and the Young's modulus a hysteresis at the phase change. For single crystals the moduli do vary with crystal direction, as shown in Table 2. Values of *Poisson's ratio* for cobalt are found to be in the range of 0.29 – 0.32.

Table 2. Elastic moduli at 20°C for cobalt along different crystal directions

Crystal direction	Young's modulus, 10^3 MN/m^2	Sheer modulus, 10^3 MN/m^2
1120	174	62.2
1010	175	62.2
1012	169	74.1
0001	213	62.4

Vacuum-melted, hot-worked cobalt annealed at 800 – 1000 °C has *tensile strengths* of 800 – 875 MN/m^2 and an elongation of 15 – 30 %. Air-melted samples are much less ductile. Cobalt has a maximum *ductility* at 500 °C, and this is the best working temperature for the metal, although the vacuum-degassed metal free from lead, sulfur, and zinc, can be hot-rolled in the temperature range of 600 – 1000 °C. Both elongation and tensile strength values show changes at the phase transformation.

The activation energy for *creep* is higher for hexagonal cobalt than for the cubic form; the values near the phase change temperature are 540 and 190 kJ/mol, respectively. The hexagonal crystals have lower values of friction and wear than the cubic crystals. The *friction coefficient* of cobalt rises steeply at temperatures above 300 °C to values of ca. 1.6 at 400 °C. The low coefficient associated with the hexagonal form is due to easy slip on the basal plane accompanied by the formation of a surface layer with the basal planes oriented parallel to the surface. In actual applications, however, the frictional and wear behavior of the metal will be determined by surface oxide films.

Magnetic Properties. Pure cobalt has the highest known Curie temperature. Hexagonal cobalt is ferromagnetic at all temperatures, but the cubic form becomes paramagnetic at 1121 $\pm$ 3 °C. The magnetic properties of single crystals show marked anisotropy. For hexagonal crystals at ambient temperature, the *c*-axis is the direction of easiest magnetization, but with increasing temperature magnetization becomes easier in a direction perpendicular to the *c*-axis; however, at ca. 250 °C the metal is magnetically isotropic. At even higher temperature, the *c*-axis becomes the most difficult to magnetize. Near the phase change temperature, the direction of easiest magnetization of the cubic form is along the $\langle 111 \rangle$ cube diagonal. Differences between this and other directions diminish with increasing temperatures until, at 1000 °C, the cubic form is magnetically isotropic.

The magnetic properties of polycrystalline samples of cobalt depend on the purity of the metal and its thermal history; this applies particularly to the development of preferred orientation. Magnetic anisotropy can, for example, be developed by slowly cooling the metal over the phase transformation temperature range in a strong magnetic field. Typical values for magnetic property parameters are listed in Table 3. The magnetorestrictive behavior of cobalt is also strongly anisotropic and reaches a maximum value for hexagonal crystals of $\Delta l/l = -27 \times 10^{-6}$ in a magnetic field of 560 kA/m along

the ⟨1120⟩ direction. The volume magnetorestriction shows a maximum contraction of 10^{-6} in a field of 480 kA/m and a magnetorestrictive expansion when the fields are above 760 kA/m.

Table 3. Magnetic properties of cobalt

Curie temperature	1121 °C
Maximum permeability	3.1×10^{-4} H/m
Coercive force	707 A/m
Remanence	0.49 T
Saturation magnetization	1.79 T

Electrical Properties. The electrical resistivity and the temperature coefficient of the resistivity of high-purity (99.999 %) polycrystalline cobalt are given in Table 4. The cubic form has a lower resistivity than the hexagonal modification whenever the forms coexist. The resistivity of single crystals of cobalt shows a marked dependence on orientation.

Table 4. Electrical resistivity and temperature coefficient of resistivity for cobalt

Temperature, °C	Resistivity, 10^{-8} Ω m	Temperature coefficient, 10^{-11} Ω m/K
− 100	2.65	–
− 50	3.84	–
0	5.25	30
50	6.81	34
100	8.52	39
200	12.80	50
300	18.38	60
400	24.72	71
500	30.65	75
600	39.1	87
800	57.9	105
1000	78.5	100
1200	92.7	41
1400	100.0	34

Other Properties. The self-diffusion of cobalt in the temperature range between 1057 and 1306 °C follows an Arrhenius-type law, the *diffusion coefficient* being given by the formula 2.2 (exp − 295/*RT*). The rate of diffusion is very dependent on grain size, being about 10 times faster for powder-metal samples than for bulk samples prepared by melting.

The *optical reflectance* of polished cobalt at ambient temperature and near normal incidence increases with wavelength from 60 % (at 1 µm) to 97 % (at 12 µm). The reflectance is much less at high angles of incidence. The optical emission spectrum of cobalt is complex and has over 330 lines in the range of 974 – 201 nm. The most important lines for analytical purposes are listed in Table 5. The characteristic K α_1, K α_2, and K β_1 X-radiation from a cobalt target occur at 0.1789, 0.1793, and 0.1621 nm, respectively; the K absorption edge is at 0.1608 nm.

Table 5. Optical emission lines of cobalt used in atomic emission analysis

Wavelength, nm	Relative intensity	
	Arc source	Spark source
352.9813	1000	30
346.5800	2000	25
345.3505	3000	200
340.5120	2000	150
251.9822	40	200
238.8918	10	35
237.8622	25	50
236.3787	25	50
230.7857	25	50
228.6156	40	300

3. Occurrence [1, 2, 4, 9]

Cobalt occurs in nature in a widespread but dispersed form in trace quantities in many rocks, soils, and plants. It is also found in sea water and in manganese-rich marine nodules. The cobalt content of the earth's crust is about 20 mg/kg, whereas its concentration in sea water has been reported as being 0.1 – 1 parts in 10^9. The concentration of the element in marine nodules is usually 0.1 – 1 %. The largest concentrations of cobalt are found in mafic and ultramafic igneous rocks; the concentration of the element and the nickel – cobalt ratio decreases from ultramafic to acidic rocks, as shown in Table 6. The nickel – cobalt ratio changes because cobalt enters the lattice of early crystallizing magnesium silicates less readily than nickel. Sedimentary rocks contain varying amounts of cobalt, with average values of 4, 6, and 40 mg/kg being reported for sandstone, carbonate rocks, and clays or shales, respectively. During the formation of metamorphic rocks, very little movement or concentration of cobalt took place; hence, the levels of cobalt found in metamorphic rocks depend essentially on the amount of the element in the

original igneous or sedimentary source. Those formed from ultramafic or mafic sources contain an average cobalt content of ca. 100 mg/kg, whereas gneissic granites and metasedimentary rocks contain an average of 16 and 8 mg/kg, respectively.

Table 6. Average cobalt content of igneous rocks

Rock type	Cobalt content, mg/kg	Ni – Co ratio
Ultramafic	270	7
Gabbro	51	2.6
Basalt	41	2.5
Diabase	31	2.3
Intermediate igneous	14	1.9
Felsic	5	1.1

Under oxidizing conditions, cobalt shows a strong tendency to concentrate with manganese oxides. In the weathering process of mafic and ultramafic rocks to form laterites, nickel tended to be leached downward with magnesia and silica, whereas cobalt with manganese oxides was residually enriched near the surface of the deposit. The concentration of cobalt and other metals in marine nodules has been attributed to the strong ion-exchange properties of the submicroscopic particles of colloidal manganese dioxide from which the nodules were formed.

Cobalt is a major constituent of about 70 minerals [10] and is a minor or trace constituent of several hundred more, particularly those containing nickel, iron, and manganese. The minerals that have been mined or concentrated for their cobalt content and those that are relatively high in cobalt are listed in Table 7. The sulfide minerals include the copper-containing carrollite, which is one of the main sources of cobalt in Zaire, linnaeite found in Zaire, Zambia, and the United States, and cattierite, which is also present in ores from Zaire. The arsenide ores include smaltite, which is found in the silver – copper ores from Cobalt, Ontario and in Morocco, and skutterudite, the main cobalt mineral in Canadian and Moroccan deposits. The sulfoarsenide, cobaltite, is found in ore bodies in Zaire, Canada, and the United States. The oxide mineral heterogenite is a hydrated metal oxide containing varying amounts of cobalt and copper and is one of the main cobalt-bearing components of the Zaire deposits. The hydrated manganese – cobalt mineral, asbolite, is the source of most of the cobalt in ores from New Caledonia.

Table 7. Cobalt minerals

Mineral	CAS reg. no.	Cobalt content, %
Cattierite, CoS_2 (pure)	[*12017-06-0*]	47.8
Linnaeite, Co_3S_4 (pure)	[*1308-08-3*]	58.0
Siegenite, $(Co,Ni)_3S_4$	[*12174-56-0*]	20.4 – 26.0
Carrollite, $(Co_2Cu)S_4$	[*12285-42-6*]	35.2 – 36.0
Cobaltite, (Co,Fe)AsS	[*1303-15-7*]	26.0 – 32.4
Safflorite, $(Co,Fe)As_2$	[*12044-43-8*]	13.0 – 18.6
Smaltite, $(Ca, Ni)As_2$	[*12006-41-6*]	ca. 21
Glaucodot, (Co,Fe)AsS	[*12198-14-0*]	12.0 – 31.6
Skutterudite, $(Co,Fe)As_3$	[*12196-91-7*]	10.9 – 20.9
Heterogenite, CoO(OH) (pure)	[*12323-83-0*]	64.1
Asbolite	[*12413-71-7*]	0.5 – 5.0
Erythrite $(CoNi)_3(AsO_4)_2 \cdot 8 H_2O$	[*149-32-6*]	18.7 – 26.3
Gersdorffite, (Ni,Co)AsS	[*12255-11-7*]	(low)
Pyrrhotite, $(Fe,Ni,Co)_{x-1}S_x$	[*1310-50-5*]	up to 1.00
Pentlandite, $(Fe,Ni,Co)_9S_8$	[*53809-86-2*]	up to 1.50
Pyrite, $(Fe,Ni,Co)S_2$	[*1309-36-0*]	up to 13.00
Sphalerite, Zn(Co)S	[*12169-28-7*]	up to 0.30
Arsenopyrite, Fe(Co)AsS	[*1303-18-0*]	up to 0.38
Manganese oxide minerals		0.10 – 1.0

Ore Deposits. Cobalt is produced mainly as a byproduct of the mining and processing of the ores of other metals, particularly those of copper, nickel, and silver, but also those of gold, lead, and zinc.

The deposits of cobalt can be classified under the following headings:

1) *Hypogene deposits* associated with mafic intrusive igneous rocks. The massive and disseminated iron – nickel – copper sulfides containing cobalt are important examples of this type of deposit. They include the ore from the Sudbury district of Ontario, which has an average cobalt content of 0.07 %. The suite of ore minerals consists of pyrrhotite, pentlandite, pyrite, marcasite, cobaltite, and gersdorffite in veins stringers and disseminated grains within an igneous host rock of Precambrian age.
2) *Contact metamorphic deposits* associated with mafic rocks. Deposits of magnetite, chalcopyrite, and cobalt-containing pyrite, formed by contact metamorphism of carbonate rock by sills and dikes of diabase, gave the ore deposits at Cornwall and Morgantown in the United States.

3) *Lateritic Deposits.* The weathering of peridotite and serpentine generally gives laterite that is rich in iron, nickel, cobalt, and chromium. Commercially valuable deposits contain 40 –50 % iron, 1 – 2 % nickel, and 0.01 – 0.1 % cobalt. Major lateritic deposits occur in Cuba, New Caledonia, Australia, the United States, and Russia.
4) *Massive sulfide deposits* in metamorphic rocks, largely of volcanic sedimentary origin. These deposits consist mainly of pyrite and pyrrhotite and are mined in the United States.
5) *Hydrothermal deposits* are subdivided into two classes: vein deposits and replacement deposits. Some of these deposits are the only ones that have been mined specifically as sources of cobalt. In Canada, veins that contain as much as 10 % cobalt occur in the Cobalt-Gowganda region of Ontario, whereas veins in the Bou-Azzer area of Morocco contain an average of 1.2 % cobalt. Hydrothermal replacement deposits containing 0.5 % cobalt as gersdorffite are found in Burma and at Outokumpu in Finland; a copper-rich sulfide deposit contains 0.2 % cobalt mainly in linnaeite.
6) *Strata-bound deposits:* the copper – cobalt deposits of Zaire and Zambia are of this type. They occur in folded shale and dolomite and contain a number of minerals, including chalcopyrite, bornite, chalcosite, linnaeite, and carroleite. The ore bed is 6 – 24 m thick and currently provides the world's major source of cobalt.
7) *Deposits formed as chemical precipitates* usually contain chemically precipitated cobalt in associated marine manganese nodules.

Reserves. As generally accepted, there are $(3 – 5) \times 10^6$ t of minable cobalt reserves and a further $(4 – 5) \times 10^6$ t of potential reserves. Table 8 lists the reported reserves in a number of countries. However, the minable reserves do not have the same degree of profitability. Cuba has minable resources of cobalt, but provides only 5 % of the world production because its ore bodies are laterites with only small quantities of nickel and cobalt that are costly to refine. Zaire produces more than half of the world's cobalt because its ores can be treated more profitably.

In addition to the land-based reserves listed in Table 8, almost 6×10^9 t of copper is available in marine nodules if it can be recovered and worked economically.

4. Production

4.1. Concentration of the Ores [1, 2, 9]

The first stages of the production of cobalt from its ores involve the separation of cobalt-bearing minerals from the gangue and from minerals containing other desirable metals but not cobalt. The concentrates obtained by applying physical separation methods such as gravity separation or froth flotation to ores can increase the cobalt content to 10 – 15 % from cobalt-rich ores. However, in general, these processes only increase the level of cobalt from 0.1 – 0.6 % in ores to a few percent. The concentration procedures for the main types of ores are described in this section.

Arsenide Ores. Cobalt arsenide minerals have specific gravities of 6.5 – 7.2 g/cm^3 and can be separated at relatively large grain size by gravity methods. Concentrates containing 13 – 14 % cobalt have been produced by these methods from Moroccan ores at Bou-Azzer containing 2 % cobalt.

Sulfoarsenide Ores. The main mineral in this type of ore, cobaltite, can be effectively separated by froth flotation with xanthate at pH 4 – 5, although iron sulfide minerals are also floated under similar conditions and must be separated from the cobalt concentrate.

Sulfide Ores. The chalococite and carrollite fractions of the ores of the Shaba province in Zaire are floated together at pH 9.2 by using lime, xanthates as collectors, and triethoxybutane as the frothing agent to give a concentrate containing 43 – 53 % copper and 0.5 – 3 % cobalt. The copper – cobalt sulfide ores of Zambia are first treated to float the copper minerals selectively at pH 10.5 – 11; at this stage the presence of sodium cyanide depresses the cobalt minerals. These are subsequently activated by sulfuric acid and floated at pH 8.5 – 9 with sodium isopropyl xanthate; the concentrates contain up to 3.5 % cobalt. The flotation

Table 8. Cobalt reserves

Country	Reserves, t	Country	Reserves, t
Australia	295 000	New Caledonia	385 000
Brazil	>9 000	New Guinea	18 000
Burma	16 000	Philippines	159 000
Canada	250 000	Puerto Rico	68 000
Columbia	22 500	Solomon Islands	22 500
Cuba	1 048 500	former Soviet Union	181 500
Dominican Republic	89 000	Uganda	8 000
Finland	22 500	United States	>764 000
Guatemala	45 500	Venezuela	60 000
Japan	2 500	Zaire	1 920 000
		Zambia	370 000

process at the Outokumpu Keretti Mill first separates the copper minerals with xanthates at pH 11 – 11.5 in the presence of lime and sodium cyanide. The cobalt minerals are then floated at pH 9 to recover 65 – 75 % of the cobalt from an ore containing 0.23 – 0.31 % cobalt; the concentrate contains 0.7 – 0.82 % cobalt. The nickel – copper – cobalt sulfide ores of Ontario are first treated to separate the sulfide minerals from gangue, and the cobalt minerals are then contained in the nickel sulfide concentrate.

Oxide Ores. Two methods have been used on a commercial scale to float oxide ores. The *palm oil process* collects the concentrate in a 3 : 1 mixture of hydrolyzed palm oil and gas oil emulsified in carbonated hot water; this process can give concentrates containing 23 – 25 % copper and 2 – 3 % cobalt, representing a cobalt recovery of 50 – 70 % from the ore. The *sulfidization process* is used for mixed oxide – sulfide ores to float sulfide ores first and to concentrate oxide minerals subsequently in a xanthate flotation.

4.2. Extraction [1, 2, 9]

Cobalt can be extracted from concentrates and occasionally directly from the ore itself by hydrometallurgical, pyrometallurgical, and electrometallurgical processes. Although most methods of extraction are based on hydrometallurgy, cobalt concentrates, mattes, and alloys have been reduced to metal by pyrometallurgical methods. Most arsenic-free cobalt concentrates can, for example, be mixed with lime and coal and melted in a reducing atmosphere to give cobalt – copper – iron alloys. Two types of alloys are obtained, a *white alloy* containing ca. 40 % cobalt that is processed for cobalt and a *red alloy* containing ca. 4 % cobalt that is processed for copper. Recovery of copper from the red alloy produces a cobalt-rich slag that can be recycled with the ores. The hydrometallurgical processes involve (1) leaching of concentrates to give a cobalt-containing solution, (2) separation of cobalt from the other metal ions in solution, and (3) reduction of cobalt ions to metal. Electrometallurgical processes are used in the electrowinning of the metal from leach solutions and in refining the cobalt that has been extracted by hydrometallurgical or pyrometallurgical methods.

4.2.1. Leaching of Cobalt Ores and Concentrates

Cobalt can be leached from its ores and concentrates in both acidic and alkaline media. Processes involving acid-sulfate leaching include (1) treatment of oxide ore concentrates with sulfuric acid containing a reducing agent, (2) extraction of cobalt sulfate from sulfide ores following an oxidizing roast, (3) extraction following a sulfatizing roast, and (4) extraction by pressure leaching with sulfuric acid, often in the presence of oxygen; the *Chemico process* for the treatment of arsenic-containing ores is, for example, based on an acid pressure leach. Acid leaching processes based on hydrochloric acid extraction and chloridation roasting have also been used.

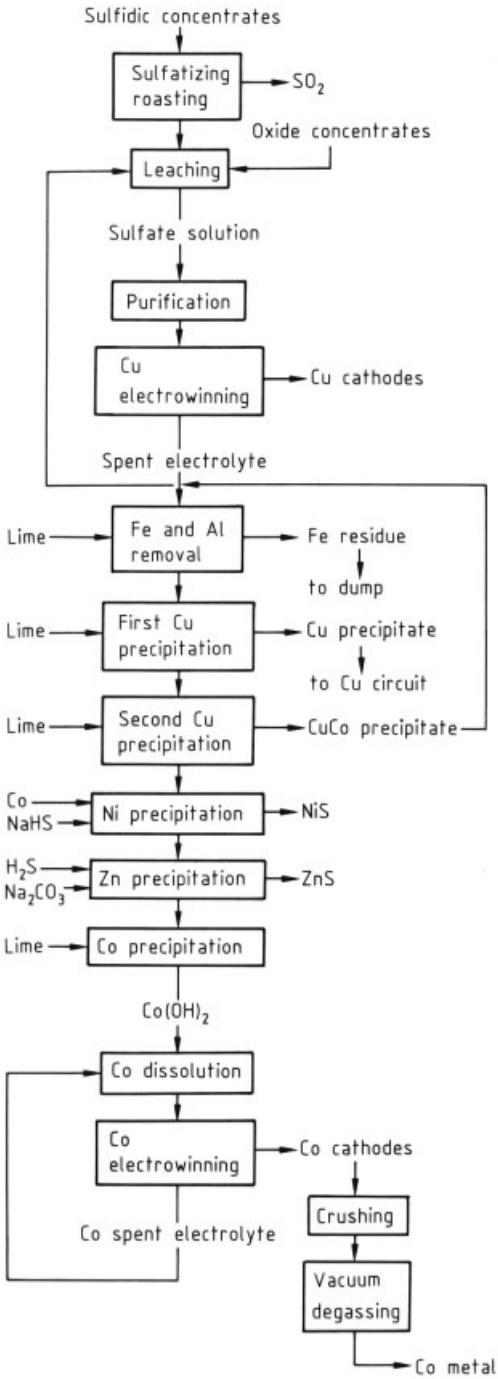

Figure 1. Flow diagram of the Gécamines process

Processes that involve the extraction of cobalt into alkaline solutions depend on the formation of the hexamminocobalt(II) species $[Co(NH_3)_6]^{2+}$, which remains in solution for long periods because of its kinetic stability, even though it has low thermodynamic stability. Ammoniacal leaching has been successfully applied to ores after a reducing roast, and the Sherritt Gordon process depends on ammonia solution pressure leaching.

Cobalt from Copper – Cobalt Concentrates. The copper – cobalt ores of Zaire and Zambia are treated by a sulfatizing roast in a fluidized-bed furnace to convert copper and cobalt sulfides into soluble oxides and iron into insoluble hematite. The calcine is subsequently leached with sulfuric acid from the spent copper recovery electrolyte. Oxide concentrates are introduced at this leaching step to maintain the acid balance in the circuit. Iron and aluminum are removed from the leach solution, and copper is electrowon on copper cathodes. A part of the spent electrolyte enters the cobalt recovery circuit and is purified by removal of iron, copper, nickel, and zinc prior to precipitation of cobalt as its hydroxide. In the final stages, this cobalt hydroxide is redissolved and the metal is refined by electrolysis and by degassing to remove any traces of hydrogen. Figure 1 shows the flow diagram of the process used at the Générale des Carrières et des Mines (Gécamines) plant in Luilu.

Cobalt from Nickel Sulfide Concentrates. Nickel sulfide concentrates can be treated by either roasting or flash smelting to give a matte from which nickel and cobalt can be recovered hydrometallurgically, or they may be treated directly by an ammonia solution pressure leach. In the *Sherritt Gordon process* used at Fort Saskatchewan in Canada, which is illustrated as a flow diagram in Figure 2, a feed of matte and sulfide concentrate, containing about 10 % nickel, 2 % copper, 0.4 % cobalt, 33 % iron, and 30 % sulfur, is pressure leached in autoclaves at 83 °C and 0.7 MPa (7 bar) with an ammoniacal medium. This converts most of the sulfide via thiosulfate and polythionate to sulfate and solubilizes nickel, copper, and cobalt as ammine complexes. Most of the copper in the leach solution is precipitated as copper sulfides when the solution is distilled to recycle ammonia. The solution is treated with hydrogen sulfide to re-

move any residual copper, and any sulfide or sulfamate remaining at this stage is converted to sulfate by a pressure oxidation hydrolysis reaction at 6.5 MPa (65 bar). The solution then enters reduction autoclaves, where nickel powder is precipitated by reduction with hydrogen at 3.6 MPa (36 bar). The remaining solution, which contains 1 – 1.5 g/L of both cobalt and nickel, is treated in an autoclave with hydrogen sulfide to precipitate cobalt and nickel sulfides, which are subsequently leached with sulfuric acid at 140 °C and 6.4 MPa (64 bar). Iron is removed from this leach liquid at pH 2.5 – 3.0 by addition of ammonia solution. The cobalt(II) in solution is oxidized in air to give a very soluble cobalt(III) pentammine complex. Nickel is subsequently precipitated from the solution in two steps as a nickel ammonium sulfate. Cobalt powder is then added along with sulfuric acid to reduce the cobalt(III) pentammine complex to cobalt(II), which is finally reduced to cobalt powder by treatment with hydrogen at 120 °C and 4.6 MPa (46 bar).

Cobalt-containing mattes and mixed sulfide concentrates have also been treated by pressure leaching with sulfuric acid, for example, at Amax at Port Nickel in the United States, MRR in South Africa, Nippon Mining and Sumitomo Metals Mining Co. in Japan, and Outokumpu Oy in Finland. The flow diagram for the Outokumpu Kukkola plant (Fig. 3) is used to illustrate this type of cobalt recovery process. Two companies, Falconbridge in Canada and Norway and Société le Nickel in France, leach nickel mattes and sulfide concentrates in chloride media at ambient pressure. Falconbridge use a hydrochloric acid leach at 70 °C for about 15 % of their feed and a direct chlorine leach for the remainder. The leach medium used by Société le Nickel is chlorine in iron(III) chloride. At the INCO Copper Cliff plant nickel – copper mattes are cooled slowly and crushed prior to magnetic separation, which isolates a nickel – copper – cobalt alloy. The alloy is smelted in a top-blown rotary converter, desulfurized by oxygen lancing, and granulated. Nickel is volatilized from the granules as $[Ni(CO)_4]$ by the *INCO carbonyl process* to give a residue from which cobalt can be leached; the solution obtained allows cobalt to be electrowon after some purification steps.

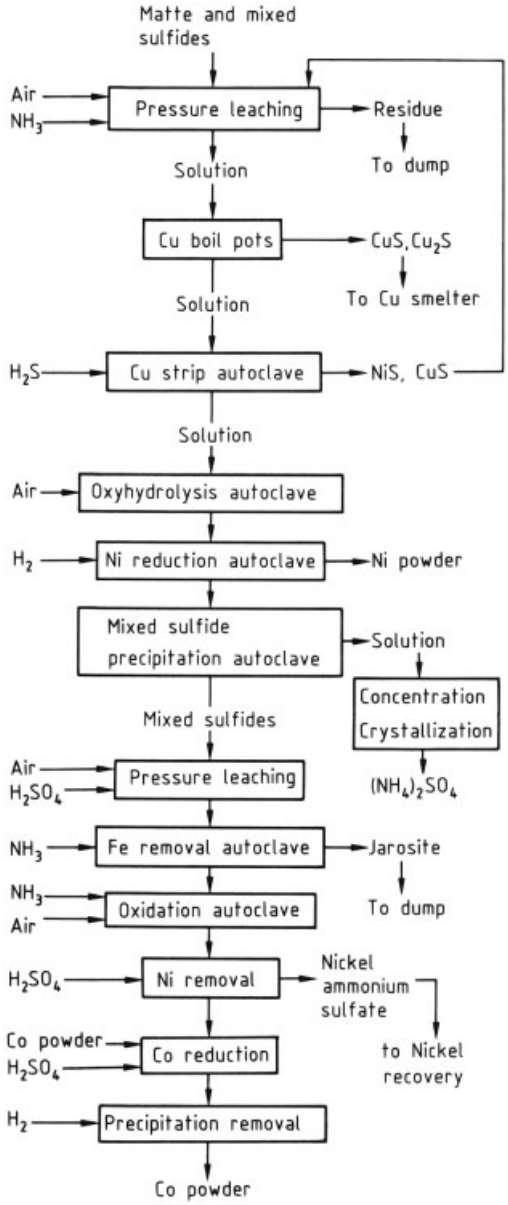

Figure 2. Flow diagram of the Sherritt Gordon process

Cobalt from Laterite Ores. There has been an increasing interest in the recovery of cobalt from laterite ores. At least 150×10^6 t of laterite reserves containing up to 15 % of cobalt have been identified in Australia alone. It is predicted that this will lead to the production of about 10 000 t/a of cobalt in Australia and about 3000 t/a from both Papua New Guinea and New Caledonia.

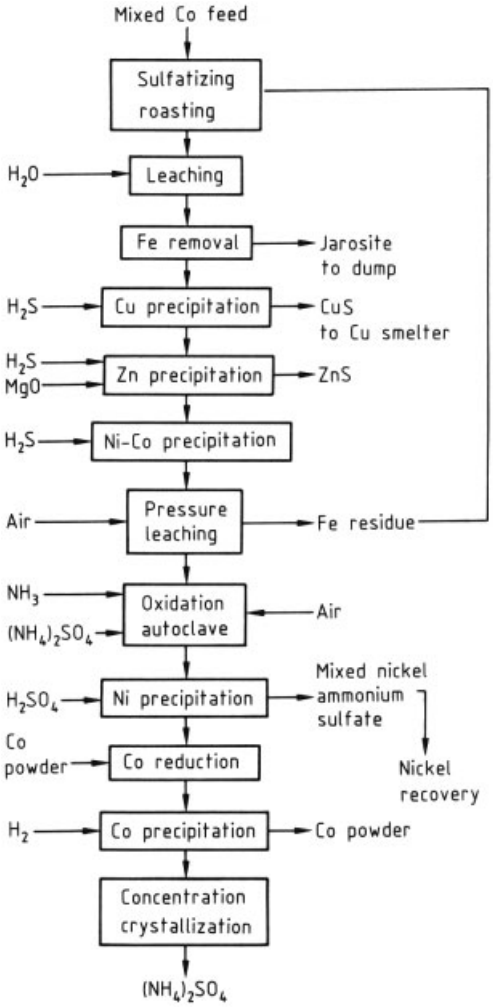

Figure 3. Flow diagram of the Outokumpu process

The typical laterite ore process involves a pressure sulfuric acid leach of slurried ore at about 250 °C followed by neutralization and clarification of the leach solution to give a pregnant liquor containing dissolved cobalt and nickel at a pH of about 4.4. Extraction efficiencies for cobalt and nickel are claimed to be grater than 90 %. Cobalt is separated from the pregnant liquor by solvent extraction prior to precipitation as sulfide. The cobalt sulfide concentrate is then refined, redissolved, and the metal extracted by electrowinning.

Cobalt from Arsenide Ores. Arsenic-containing concentrates are roasted in a fluidized bed at 600 – 700 °C to remove 60 – 70 % of the arsenic present as arsenic(III) oxide. The roasted ores can be treated with hydrochloric acid and chlorine or with sulfuric acid to give a leach solution that can be purified by hydrometallurgical methods and from which cobalt can be recovered by electrolysis or by carbonate precipitation.

4.2.2. Separation of Cobalt from Other Metal Ions in Leach Solutions [9]

A major part of cobalt recovery from leach solutions is concerned with its separation from other elements. Because of the similarities in chemical behavior of cobalt and nickel, the separation of these two elements has been studied extensively and is well known. This section deals with chemical methods of separation of cobalt from nickel and with physicochemical methods of separation of cobalt from other elements.

Chemical Methods of Separation. The chemical methods of separating cobalt from other metals use the different solubilities and kinetic or thermodynamic stabilities of their compounds. The most important compounds that have been used in separation are hydroxides, carbonates, ammine complexes, double ammonium sulfates, sulfides, chlorides, and carbonyls; the cementation to metal is also utilized. For examples, see Section 4.2.1. The relative solubilities of cobalt(III) and nickel(II) hydroxides in the presence and absence of complexing agents and of cobalt(II) and nickel(II) carbonates in solutions containing NH_3 and CO_2 have also been used to separate these elements. Fractional crystallization of chloride-containing media can be used to concentrate cobalt in solution and nickel in solid nickel chloride dihydrate.

Physical Methods of Separation. Many physical methods have been used to separate cobalt from other elements, but the three most important are electrolysis (described in Section 4.2.3.), solvent extraction, and ion exchange.

In most solvent extraction processes, cobalt is extracted with a tertiary amine from chloride solutions, although cobalt and nickel can be separated from each other with phosphate or carboxylic acid reagents from ammoniacal sulfate

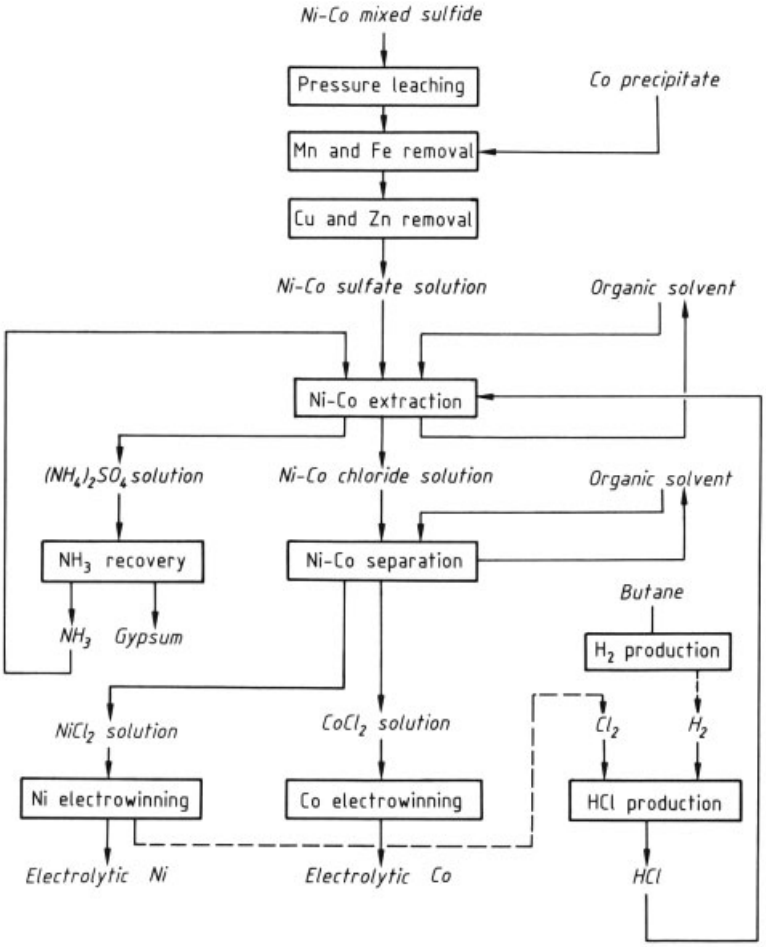

Figure 4. Flow diagram of the Sumitomo process

solutions. In the sulfuric acid leach process of Nippon Mining at Hitachi, solvent extraction is used to extract zinc from the leach solution with diethylhexyl phosphonic acid (DEHPA) at pH 2 – 3 prior to solvent extraction of cobalt with an alkylphosphonic acid. The Falconbridge and Société le Nickel processes both involve solvent extraction of cobalt from chloride solution with triisooctylamine (TIOA). The Sumitomo Metals Mining Co. uses a solvent extraction process to coextract cobalt and nickel with versatic acid. The extract is then stripped with hydrochloric acid, and cobalt is reextracted from the chloride solution with tri-*n*-octylamine (TNOA) to give an extract from which cobalt is stripped and electrowon. The flow diagram for this process is shown in Figure 4, and the solvent extraction process can be expressed in terms of the following chemical equations:

Extraction:

$$2\,\mathrm{RH\,(org)} + \mathrm{Ni\,(Co)\,SO_4\,(aq)} + 2\,\mathrm{NH_4OH} \rightarrow \mathrm{R_2Ni\,(Co)\,(org)} + (\mathrm{NH_4})_2\mathrm{SO_4\,(aq)} + \mathrm{H_2O}$$

Stripping:

$$\mathrm{R_2Ni\,(Co)\,(org)} + 2\,\mathrm{HCl} \rightarrow 2\,\mathrm{RH\,(org)} + \mathrm{Ni\,(Co)\,Cl_2\,(aq)}$$

Extraction:

$$\mathrm{Ni\,(Co)\,Cl_2\,(aq)} + 2\,\mathrm{R_3^1NHCl\,(org)} \rightarrow (\mathrm{R_3^1NH})_2\mathrm{CoCl_4\,(org)} + \mathrm{NiCl_2\,(aq)}$$

Stripping:

$$(R_3^1NH)_2CoCl_4(org) \xrightarrow{dil.\ HCl} 2\ R_3^1N \cdot HCl(aq) + CoCl_2(aq)$$

where RH is versatic acid in kerosene and $R_3{}^1N$ is tri-*n*-octylamine in xylene; (org) and (aq) are organic and aqueous phases, respectively.

4.2.3. Electrowinning [9]

Cobalt can be extracted and obtained as high-purity metal by electrolysis of sulfate or chloride solutions. Solutions of pH 1 – 4 are necessary for the electrowinning of cobalt, and the pH is often maintained by suspending cobalt hydroxide or cobalt carbonate in the electrolyte to consume the acid produced at the anode. Current densities of 5 A/dm^2 are used with current consumptions of ca. 6.5 kW h per kilogram of cobalt. Electrolytically refined cobalt is usually purer than 99.5 %.

5. Cobalt Powders [4, 11, 12]

5.1. Production

Cobalt powder can be produced by a number of methods, but those of industrial importance involve the reduction of oxides, the pyrolysis of carboxylates, and the reduction of cobalt ions in aqueous solution with hydrogen under pressure. Very pure cobalt powder is prepared by the decomposition of cobalt carbonyls.

Reduction of Oxides. Gray cobalt(II) oxide (CoO) or black cobalt(II)/cobalt(III) oxide (Co_3O_4) is reduced to metal powder with carbon monoxide or hydrogen. The reactions occur under conditions well below the melting point of the oxides or the metal. The purity of the powder obtained is 99.5 % with a particle size of ca. 4 µm, although the density and the particle size of the final product depend on the reduction conditions and on the particle size of the parent oxide. The finely powdered metal is stored at very low temperature.

Pyrolysis of Carboxylates. The thermal decomposition of such cobalt carboxylates as formate or oxalate in a controlled reducing or neutral atmosphere produces a high-purity (about 99.9 %), light, malleable cobalt powder with a particle size of ca. 1 µm that is particularly suitable for the manufacture of cemented carbides. The particle size, form, and porosity of the powder grains can be changed by altering the pyrolysis conditions.

Reduction of Cobalt Ions in Solution. Purified leach solutions containing cobalt pentammine complex ions can be treated in autoclaves with hydrogen under pressure and at high temperature to give an irregular chainlike form of the powder that is suitable for consolidation by direct strip rolling.

5.2. Uses

5.2.1. Cemented Carbides [4, 13]

One of the most important uses of metallic cobalt is as a bonding agent in cemented carbides that are used extensively as cutting tools for metals, rocks, and other high-strength materials (→ Hard Materials). No suitable substitute has been found for cobalt as a cementing agent for carbides. The important properties of cobalt for this use are (1) high melting point, (2) high-temperature strength, (3) the ability to form a liquid phase with WC at 1275 °C, (4) the ability to dissolve WC by the formation of an eutectic containing 10 % of WC at 1275 – 1350 °C, and (5) the ability to reprecipitate WC from solution as the mixture cools. The most commonly used cemented carbide, tungsten carbide containing 2 – 30 % cobalt, is manufactured as follows: pure cobalt powder is added to tungsten carbide (particle size ca. 1 µm), and the mixture is ball-milled for a long period by using hard metal balls. The mixed powders are then consolidated by cold pressing, by hot pressing at ca. 1300 – 1400 °C, by cold extension, or by slip casting to produce small artifacts, large artifacts, constant-section rods, and complex-shaped articles, respectively. The consolidated parts are sintered in a reducing atmosphere at 1300 – 1600 °C. The sintering process involves the formation of a liquid phase; the cobalt liquefies at ca. 1320 °C and dissolves tungsten and carbon from the carbide. When the mixture is cooled,

most of the tungsten and carbon reprecipitates, but the binder phase is much stronger than pure cobalt because enough tungsten and carbon are present to stabilize the metal in its cubic form. During the sintering process, the consolidated powders shrink by 20 %. The binding mechanism depends on (1) the liquid metal phase to wet the carbide particles so that surface tension forces help densify the sintered phases and (2) on the ability of the liquid to dissolve and retain traces of carbide impurities.

The properties of cemented carbides are controlled by the amount of cobalt present and the particle size of the carbide used. The hardness of the cemented carbides increases with decreasing particle size and with increasing cobalt content. The impact strength of cemented carbides is directly proportional to the cobalt content; it rises from less than 1 kg m for a sample with 6 % cobalt to ca. 2.5 kg m for a sample with 25 % cobalt in tungsten carbide with a particle size of 1.4 – 3.1 µm. The properties of cemented tungsten carbides can be improved by the addition of other carbides such as those of niobium, tantalum, or titanium.

Cobalt powders are also used as binders for the diamond tool industry [14].

5.2.2. Cobalt Powders in Powder Metallurgy and Cobalt-containing Metal [13]

Cobalt powders have been used in the formation of a number of alloy phases including maraging steel by hot extension of prealloyed powders, cobalt-based superalloys by consolidation of powders involving liquid phase sintering, fine-particle magnetic alloys, and bearing materials impregnated with low-friction substances such as graphite, lead, nylon, and molybdenum disulfide.

6. Cobalt Alloys [1 – 4, 9]

6.1. General

About 64 % of the cobalt is consumed in alloy compositions. The most important cobalt alloys are discussed in this article under the following headings:

- high-temperature alloys
- magnetic alloys
- hard metal alloys
- cobalt-containing high-strength steels
- electrodeposited alloys
- alloys with special properties

Solute metals in cobalt can be classified in terms of their effect on the phase transition (cubic to hexagonal) of the host metal. Those elements that raise the transition temperature are classified as *restricted-field components,* whereas those that lower the temperature are *enlarged-field components.* Some alloying elements must be classified as combined components because they both raise the $\epsilon \rightarrow \alpha$ transition temperature and lower the $\alpha \rightarrow \epsilon$ transition temperature. Table 9 lists some alloying elements according to their classification. The phase diagrams of a number of binary and ternary cobalt alloy systems have been reported [1, 3]. According to the phase diagrams, a number of binary alloys have stable hexagonal structures at high temperature and these may be of practical importance where the hexagonal structure is required, for example, to maintain low friction and wear characteristics. Alloys with high-temperature stable hexagonal lattices (temperature in parentheses) are cobalt – chromium, 90 : 10 (800 °C); cobalt – molybdenum, 90 : 10 (950 °C); cobalt – osmium, 60 : 40 (1250 °C); cobalt – platinum, 80 : 20 (650 °C); cobalt – rhenium, 80 : 20 (750 – 950 °C); cobalt –ruthenium, 70 : 30 (1100 – 1400 °C); cobalt – tungsten, 80 : 20 (950 °C).

Table 9. Classification of alloying elements in cobalt-rich binary alloys

Enlarged-field	Restricted-field	Combined
Al Nb	As Re	Au
B Ni	Cr Rh	Be
C Sn	Ge Ru	Ga
Cu Ti	Ir Sb	Pb
Fe Zr	Mo Si	Pd
Mn	Os Ta	V
	Pt W	

6.2. High Temperature Alloys (Superalloys) [13, 15]

Cobalt-based superalloys and cobalt-containing nickel superalloys are used in aircraft engines as turbine vanes and blades and in sheet form for exhaust case assemblies and burner liners. Cobalt-based superalloys are complex combinations of elements intended to accomplish one or more specific purposes. Normally the matrix is face-centered cubic. Commercial alloys are commonly hardened by carbide formation and by solid-solution strengthening. The properties responsible for the industrial importance of carbide-strengthened alloys are high creep strength at high temperature, high thermal fatigue resistance, good oxidation resistance, excellent hot corrosion resistance, an air melting capability, an air or argon remelting capability, good welding properties, and ease of casting. The effects of added elements on the properties of superalloys are listed in Table 10. A number of elements are detrimental to the alloy properties. Sulfur, for example, affects the ductility and must be kept at a concentration of less than 0.006 %. Other elements that have been reported to show deleterious effects are bismuth, lead, antimony, thallium, selenium, cadmium, mercury, tin, and silver. The cobalt-based alloys consist essentially of a cobalt – chromium matrix with ca. 20 % chromium, which provides major resistance to high-temperature corrosion. Up to 30 % of the cobalt can be replaced with nickel, and up to 15 % tungsten may be added as well as smaller levels of other beneficial metals. The carbon content lies in the range of 0.25 – 1 % for casting alloys and 0.05 – 0.4 % for hot worked alloys. Table 11 lists the compositions of several cobalt-based high-temperature alloys, indicating the alloys that are normally cast; Table 12 shows the stress rupture strengths for some cobalt-based superalloys.

The presence of carbides in the grain boundaries of superalloys is important in determining their mechanical properties. The grain boundaries are strengthened by carbides in one of two ways: (1) their precipitation as discrete particles in the grain boundaries hinders sliding by a pinning action and (2) the formation of a grain boundary skeletal network that can support load as well as hinder sliding. Both mechanisms also strengthen by limiting grain-boundary migration. Solid-solution strengthening is principally due to contents of chromium, tungsten, niobium, and tantalum, not to the precipitation of carbides. Table 13 lists the strengthening mechanisms in cobalt-based superalloys.

Table 10. Effects of several elements in cobalt superalloys

Element	Effect*
Chromium	improves oxidation and hot-corrosion resistance; produces strengthening by formation of M_7C_3 and $M_{23}C_6$ carbides
Molbydenum and Tungsten	solid-solution strengtheners; produce strengthening by formation of intermetallic compound Co_3M; formation of M_6C carbide
Tantalum and Niobium	solid-solution strengtheners; produce strengthening by formation of intermetallic compound Co_3M and MC carbide; formation of M_6C carbide
Aluminum	improves oxidation resistance; formation of intermetallic compound CoAl
Titanium	produces strengthening by formation of MC carbide and intermetallic compound Co_3Ti; with sufficient nickel produces strengthening by formation of intermetallic compound Ni_3Ti
Nickel	stabilizes the face-centered cubic form of matrix; produces strengthening by formation of intermetallic compound Ni_3Ti; improves forgeability
Boron and Zirconium	produce strengthening by effect on grain boundaries and by precipitate formation; zirconium produces strengthening by formation of MC carbide
Carbon	produces strengthening by formation of MC, M_7C_3, $M_{23}C_6$, and possibly M_6C carbides
Yttrium and Lanthanum	increase oxidation resistance

* M = metal.

The important properties of the alloys for high-temperature use are time-dependent creep, stress rupture, or fatigue strength at the operating temperature. Properties such as tensile strength, hardness, and modulus are not important for high-temperature operation, but do have an effect on manufacturing operations involved in the production of parts. The oxidation of uncoated cobalt-based superalloys can be quite severe, and protective coatings are required when the alloys are used in hot sections of turbines in jet engines.

Cobalt-containing Nickel-based Alloys. Most highly stressed components in high-temperature engineering use are nickel-based superalloys strengthened by the precipitation

Table 11. Composition of superalloys

Trade name		Element											
		C	Cr	Ni	Mo	W	Nb	Ta	Ti	B	Zr	Fe	Other
(C)*	HS-31 (X-40)	0.5	25.5	10.5	–	7.5	–	–	–	0.01	–	2	–
(C)	HS-21 (Mod. Vitallium)	0.25	27	3	5	–	–	–	–	–	–	1	–
	S-816	0.38	20	20	4	4	4	–	–	–	–	4	–
	HA-25 (L-605)	0.1	20	10	–	15	–	–	–	–	–	–	–
(C)	HE-1049	0.4	26	10	–	15	–	–	–	0.4	–	≤ 3	–
	UMCo-50	0.1	28	–	–	–	–	–	–	–	–	21	–
(C)	ML-1700	0.2	25	–	–	15	–	–	–	0.4	–	–	–
	J-1570	0.2	20	28	–	7	–	–	4	–	–	2	–
	J-1650	0.2	19	27	–	12	–	2	3.8	0.02	–	–	–
(C)	WI-52	0.45	21	≤ 1	–	11	2	–	–	–	–	2	–
(C)	HA-151	0.5	20	–	–	12.7	–	–	–	0.05	–	–	–
(C)	MAR-M 302	0.85	21.5	–	–	10	–	9	–	0.005	0.2	–	–
(C)	MAR-M 322	1.0	21.5	–	–	9	–	4.5	0.75	–	2.25	–	–
(C)	X-45	0.25	25.5	10.5	–	7.5	–	–	–	0.01	–	2	–
	NASA-CoWRe	0.4	3	–	–	25	–	–	1	–	1	–	2 RE**
(C)	AiResist 13	0.45	21	≤ 1	–	11	2	–	–	–	–	≤ 2.5	3.5 Al; 0.1 Y
(C)	MAR-M 509	0.6	24	10	–	7	–	3.5	0.2	–	0.5	–	–
	UMCo-51	0.3	28	–	–	–	2.1	–	–	–	–	19	–
	AiResist 213	0.18	19	–	–	4.7	–	6.5	–	–	0.15	–	3.5 Al; 0.1 Y
	MAR-M 918	0.05	20	20	–	–	–	7.5	–	–	0.1	–	–
	HA-188	0.1	22	22	–	14	–	–	–	–	–	1.5	0.08 La
(C)	AiResist 215	0.35	19	–	–	4.5	–	7.5	–	–	0.13	–	4.3 Al; 0.17 Y
	CM-7	0.1	20	15	–	15	–	–	1.3	–	–	–	0.5 Al
(C)	FSX-414	0.35	29.5	10.5	–	7	–	–	–	0.01	–	2	–

* C = cast alloy.
** RE = rare earth.

of the intermetallic compound $Ni_3(TiAl)$. Addition of ca. 20 % cobalt improves the high-temperature properties of the nickel-based alloys. The influence of cobalt arises from several effects, namely:

1) Its relatively minor contribution to solid-solution hardening.
2) Its influence on the solubility of the matrix for titanium and aluminum. At temperatures below 1080 °C, the cobalt-containing alloys have a lower solubility for these metals than the cobalt-free alloys and, therefore, increase the amount of $Ni_3(TiAl)$ precipitated at a given level of alloying; this is a major factor in giving increased high-temperature strength.
3) Its effect on the formation of carbides in the grain boundary. The presence of cobalt increases the solubility of carbon in the matrix and modifies the carbides formed in the grain boundary.
4) It improves the hot workability.

The presence of cobalt in nickel-based alloys has a pronounced effect in decreasing the stacking fault energy, which results in a decrease in the creep rate of nickel or nickel – chromium phases.

6.3. Magnetic Alloys [1, 2, 4, 16, 17]

Cobalt is used in the manufacture of three main types of permanent magnet materials: quench-hardened carbon steels, nickel – iron – aluminum – cobalt (Alnico) [*12605-54-8*] alloys, and rare-earth cobalt alloys (→ Magnetic Materials).

Quench-hardened Carbon Steels. The addition of cobalt to quench-hardened steels increases the coercive force and the BH_{max}. The improved properties are mainly due to magnetorestrictive strain effects that increase the coercive force together with a small increase in saturation magnetization. Currently, magnet steels are used less, but they still have advantages in situations where the magnetic alloy must be machined and worked.

Table 12. Characteristic rupture strengths of selected cobalt- and nickel-based superalloys

Trade name		Characteristic rupture strength, MN/m^2							
		815 °C		870 °C		980 °C		1095 °C	
		100 h	1000 h	100 h	1000 h	100 h	1000 h	100 h	1000 h
(C)	HS-21 (Mod. Vitallium)	152	98	115	91	65	48	–	–
(C)	HS-31 (X-40)	179	138	134	103	76	55	–	–
(C)	X-45	131	103	96	69	48	31	–	–
(C)	FSX-414	152	117	110	83	55	34	21	–
(C)	WI-52 (HA-152)	–	–	172	152	96	76	34	–
(C)	HA-151	255	228	186	165	96	79	–	–
	S-816	172	124	107	69	–	–	–	–
(C)	S-816 + 1 % B	301	223	206	145	100	54	–	–
	J-1570	228	165	158	110	–	–	–	–
	J-1650	317	228	220	145	90	–	–	–
(C)	MAR-M 302	276	207	207	158	110	76	41	28
(C)	MAR-M 322	276	193	228	158	138	103	69	55
(C)	MAR-M 509	269	228	200	138	117	90	55	38
	HA-25 (L-605)	165	117	107	72	48	26	–	–
	HA-188	154	110	105	70	41	25	15	–
	MAR-M 918	207	138	110	76	41	22	17	–
	Nimonic 80 A	193	116	–	–	–	–	–	–
	Nimonic 90	239	154	139	77	–	–	–	–
	Nimonic 105	324	234	139	134	68	32	–	–
	Nimonic 115	402	309	301	201	121	70	–	–
	Udimet 700	400	296	290	200	110	52	–	–
(C)	B-1900	503	379	386	255	179	106	–	–
(C)	MAR-M 200	524	413	400	290	186	131	76	45
	Hastelloy X	107	69	69	48	36	21	–	–
	Rene 41	310	200	193	117	69	–	–	–

C = cast alloy.

Nickel – Iron – Aluminum – Cobalt Alloys (Alnico Alloys). Nickel – iron – aluminum alloys (see also → Magnetic Materials) have a high coercive force, but are very brittle and must be cast into their final shape. Addition of cobalt to the alloys greatly improves their magnetic properties. The gains are progressive with increasing cobalt content up to 20 – 25 % cobalt. Typical compositions of cobalt-containing Alnico alloys are given in Table 14 along with their magnetic properties. The magnetic properties of these alloys, particularly remanence, can be improved by applying a strong external field to them as they cool from the solution treatment temperature. The resulting magnetic properties are anisotropic, being higher in the direction parallel to the field than perpendicular to it. The properties of anisotropic magnets can be further improved by careful control of thermal gradients during cooling, which results in the growth of columnar grains with the ⟨100⟩ crystal axes lying along the direction of growth. When these columnar alloys are subjected to a magnetic field during cooling, the magnetic properties are again markedly improved.

The permanent magnet properties of Alnico alloys depend on the separation of the original body-centered cubic lattice into two isomorphous phases: (1) a nickel – aluminum-rich nonmagnetic matrix and (2) magnetic iron – cobalt-rich precipitated particles that are epitaxially orientated in the matrix and elongated along one of the ⟨100⟩ directions.

The Alnico alloys can be prepared by casting in shell molds or by powder metallurgy. The use of powder metallurgy depends on a knowledge of the domain theory of ferromagnetism. In the process the particle sizes of the alloy powders used are reduced so that they approximate the size of single magnetic domains. The fine particles (0.02 – 0.05 µm) of the iron – cobalt alloys used can be prepared by thermal reduction of salts such as formates or oxalates or by controlled electrolysis of mixed iron – cobalt salt solutions.

Rare-Earth – Cobalt Alloys. The rare-earth – cobalt alloys often have the composition MCo_5, where M is a rare-earth metal, and contain 60 – 65 % cobalt (see also → Magnetic

Table 13. Strengthening mechanisms in cobalt-based superalloys

Mechanism	Comments
Carbide formation (MC, M_6C, $M_{23}C_6$, M_7C_3)	effective as strengthening agents but overaging a problem; optimum distribution not always achieved
Solid-solution hardening (Ta, W, Mo, Nb)	effective at temperatures above ca. 1095 °C, but not potent below this temperature
Intermetallics (Co_3W, Co_3Mo, Co_3Ti, Co_3Ta, σ-, μ-, or Laves phases)	overaging above ∼ 870 °C; possible deterioration of ductility and oxidation resistance due to σ, μ and Laves; appreciable hardening can be achieved below 815 °C
Grain structure (directional, single, fine)	directional solidification yields better thermal fatigue and ductility, but higher cost because of special processing; fine-grained more ductile, possible better fatigue but poorer creep
Dispersoids (ThO_2, other oxides, carbides)	expensive, technique very critical, no outstanding alloy yet developed
Minor elements (B, Zr)	fairly well explored, effective; excessive boron must be avoided due to low ductility and incipient melting
Eutectics (intermetallics)	requires further investigation, offers promise of high-temperature stability and good strength
Face-centered cubic to hexagonal close packing transformation and dislocation structure	field requires much more investigation before assessment can be made, question of stability
Fiber reinforcement	not enough information developed as yet
Combinations of the above	adds to expense so that properties developed must be improved enough for economic justification

Table 14. Composition and magnetic properties of Alnico alloys

Composition, %						Method of manufacture	Remanence, T	Coercive force, kA/m	BH_{max}, kJ/m^3
Co	Ni	Al	Cu	Ti	Nb				
3 – 5	21 – 28	11 – 13	2 – 4	0 – 1	–	Cast	0.55 – 0.75	36 – 56	11 – 12
12 – 14	16 – 20	9 – 11	3 – 6	0 – 1	–	Cast	0.65 – 0.80	40 – 50	13 – 14
17 – 20	18 – 21	8 – 10	2 – 4	4 – 8	–	Cast	0.6 – 0.7	60 – 72	14 – 16
23 – 25	12 – 15	7.8 – 8.5	2 – 4	0 – 0.5	–	field treated	1.2 – 1.3	46 – 52	40 – 44
32 – 36	14 – 16	7 – 8	4	4 – 6	–	field treated	0.0 – 0.9	110 – 140	40 – 45
24 – 25	13 – 15	7.8 – 8.5	2 – 4	–	0 – 1	columnar	1.3 – 1.4	56 – 62	56 – 64
32 – 36	14 – 16	7 – 8	4	4 – 6	0 – 1	columnar	1.0 – 1.3	110 – 140	60 – 74

Materials). The anisotropy that gives high coercivity in these alloys arises from a hexagonal lattice that has the direction of easy magnetization along the hexagonal axis. Single domain particles compacted in an external magnetic field retain the high coercivity of the individual particles. Exceptionally high values of coercive force have been obtained in these alloys — the values for alloys based on samarium or praseodymium are the highest and can be as high as 650 kA/m.

The rare-earth – cobalt magnets are normally manufactured by melting the components in a protective atmosphere. The cooled melt is then crushed and milled to a fine powder, again in a protective atmosphere, before it is compacted to shape in an applied field. Two dominant groups of rare-earth – cobalt magnets are in commercial production. One group is based on single-phase $SmCo_5$ [*12017-68-4*] and the other on precipitation-hardened alloys near the compo-

sition $Sm(Co, Cu, Fe)_{7.5}$. A large amount of the samarium in $SmCo_5$ can be substituted by neodymium, praseodymium, or the natural mixture of rare-earth metals (mischmetals); this either increases the saturation polarization or decreases the cost of the alloy components.

Cobalt is also the major component of zero magnetorestrictive amorphous alloys. These are metallic glasses with the lowest technically realized coercivities. They have no directionality in terms of their magnetic properties and are relatively hard and strong materials giving rise to low hysteresis eddy current losses. Their properties favor their use in applications such as magnetic shielding, magnetic heads, amplifiers, pulse transformers, and HV/HW transformers.

The uses of the magnets include focusing magnets in traveling wave tubes, magnetic bearings in ultrahigh-speed centrifugal separators and inertia wheels, and actuators, motors, and generators of various sizes from watches to 100-horsepower motors, including brushless motors. The magnetic properties of these alloys depend critically on their manufacturing history, but are characterized by an almost linear demagnetization curve between the remanence and coercive force. Typical magnetic properties of some of the MCo_5 magnets are listed in Table 15. The $Sm(Co, Cu, Fe)_x$ phase owes its coercive force to pinning of domain walls on coherent precipitates embedded inside the grains either of a M_2Co_{17} phase embedded in an MCo_5 matrix if $x < 7$ or discrete MCo_5 particles embedded in M_2Co_{17} for $x > 7.8$. The presence of copper is essential for this type of hardening, but it decreases the saturation polarization.

Table 15. Properties of MCo_5 magnets

Magnet	Remanence, T	Coercive force, kA/m	BH_{max}, kJ/m^3
$SmCo_5$	0.85	600	140
$Sm_{0.5}Pr_{0.5}Co_5$	0.95	600	176
(Mischmetal)Co_5	0.65	300	64
$PrCo_5$	0.94	660	150

6.4. Hard Metal Alloys [4]

Because of their inherent high strength, corrosion resistance, and ability to retain hardness at elevated temperature, cobalt-based alloys have been used extensively for over 50 years in wear-related engineering applications.

Cobalt – Chromium Alloys (Stellite Alloys). The Stellite alloys were developed around 1900 and over 20 compositions are now commercially available. They can be divided into two classes: (1) alloys of cobalt, chromium, tungsten, and carbon and (2) alloys of cobalt, chromium, tungsten, molybdenum, nickel, iron, and carbon. For special hard-facing applications, these alloys can also be modified by the addition of small percentages of silicon and boron.

The traditional *alloys of cobalt, chromium, tungsten, and carbon* contain ca. 30 % chromium with varying amounts of tungsten (4 – 17 %) and carbon (1 – 3.2 %). They are used primarily in lubricated or high-temperature wear applications. The composition of a typical tough-grade hard alloy is 30 % chromium, 4 % tungsten, and 1 % carbon, which has a Vickers hardness of 375 N/mm^2 at room temperature and of 100 N/mm^2 at 900 °C. A typical medium-grade hard alloy contains 30 % chromium, 9 % tungsten, and 1.5 % carbon and has hardness values of 470 N/mm^2 at room temperature and of 150 N/mm^2 at 900 °C. A typical hard-grade hard alloy contains 32 % chromium, 18 % tungsten, 2.5 % carbon and has hardness values of 600 N/mm^2 at room temperature and of 200 N/mm^2 at 900 °C.

The *Co – Cr – W – Mo – Ni – Fe – C alloys* contain 25 – 32 % chromium, 0 – 19 % tungsten, 3.22 % nickel, 3 – 10 % molybdenum, 5 – 20 % iron, and 0.1 – 2 % carbon in a cobalt base. The microstructure of the Stellite alloys consists of hard particles of Cr_7C_3 dispersed in a cobalt-rich solid-solution matrix. The adhesion and cavitation-erosion wear characteristics that have been attributed to the matrix alloy are influenced by a phase change from a strain-induced cubic form to a hexagonal form in the matrix alloy. The toughness and abrasion resistance of the alloys are, however, due mainly to the volume fraction of carbides and their morphology.

Cemented Carbides. The precipitation and applications of cemented carbides as hard alloys are discussed in Section 5.2.1.

6.5. Cobalt-containing High-Strength Steels [18]

Cobalt is not one of the elements commonly added to steels because its direct effects on the properties of steel are usually less than those achieved by other alloying elements. Cobalt is, however, an important component of steel when high strength is a requirement, and it is a unique alloying element in favoring the austenitic structure at the ferrite – austenite boundary and in raising the martensitic transformation temperature.

The addition of a few percent of cobalt to low-alloy steels that contain strong carbide-forming elements, such as molybdenum or tungsten, increases the tempering resistance of the steel. Solid-solution hardening of martensitic steels by the addition of cobalt is particularly noticeable in chromium stainless steels. High chromium ferritic and martensitic steels containing 6 – 7 % cobalt are used for high-temperature applications up to 550 °C. Cobalt is a particularly useful alloying element in controlled-transformation steels in which initial cooling of the steel gives austenite, which can be cold worked and subsequently transformed into martensite by tempering or refrigeration. A typical steel of this type contains 6 % cobalt.

Maraging steels are ultrahigh-strength alloys and contain cobalt, nickel, and molybdenum with 5 – 20 % cobalt. Cobalt plays an indirect role in the formation of intermetallic compounds containing minor alloy components such as titanium, which accelerates and enhances the hardening process. Cobalt also plays a prominent role in the short-range ordering of the solid solution that leads to additional hardening. The advantages of maraging steels are (1) good tensile strength and toughness that is maintained up to 350 °C and (2) their weldability either in the annealed or aged conditions without preheat or postheat stress-relieving treatments.

Cobalt (5 – 12 %) can be added to improve the hot hardness of high-speed tool steels.

6.6. Electrodeposited Alloys [9]

Protective coatings of electroplated nickel – cobalt alloys have been used since the mid-1930s. The 25 % cobalt alloy deposited from sulfate electrolytes is about three times harder than electroplated nickel and has about the same hardness as pure cobalt. The Vickers hardnesses of the electrodeposits vary with cobalt content, and for deposits from sulfamate – chloride solutions, they rise from ca. 200 N/mm^2 for nickel plate to ca. 540 N/mm^2 for a deposit containing 40 % of cobalt. The cobalt – nickel alloy deposits also have improved corrosion resistance in many environments and have been used in the production of mirrors, decorative coatings, and for electroforming. Electrodeposits of cobalt – tungsten alloys are of interest because of their hardness at high temperature. Hardness values in the range of 400 – 600 N/mm^2 are obtained for deposits with 30 % tungsten; this increases to 500 –700 N/mm^2 after heat treatment at 300 – 350 °C and remains as high as 650 N/mm^2 at 600 °C. These electrodeposited alloys have been used to improve the wear resistance of hot forging dies.

The magnetic properties of a number of electroplated alloys have attracted attention for use in recording systems and computer applications. The alloys with suitable magnetic properties include cobalt – iron, cobalt – nickel, cobalt – platinum, and cobalt – phosphorus.

6.7. Alloys with Special Properties

A number of cobalt-containing alloys have special applications; they are described in this section.

Dental Alloys. Cobalt – chromium alloys have been used in dentistry since the mid-1930s (→ Dental Materials). The most widely used alloy is Vitallium [*12629-02-6*], which is a cobalt-based alloy containing 30 % chromium, 5 % molybdenum, and 0.5 % each of carbon and silicon. The alloys are used for cast denture bases, complex partial dentures, and some types of bridgework. The properties of the alloys that make them suitable for dental applications are their good castability, resistance to tarnish, compatibility with mouth tissues, high strength and stiffness, and low density. A modified alloy is used for porcelain coatings for crown restorations; it enables the crown to be fused to metal bridges. Alloys containing 12 % chromium, 9 % nickel, 8 % titanium, and 1 % molybdenum

have expansion characteristics that match those of dental porcelain.

Surgical Implants. Cobalt – chromium alloys similar to Vitallium are used for surgical implants and for bone replacement and repair because they are very resistant to corrosion in body fluids. The best known prosthesis is hip joint replacement, but the alloys are also used for plates, screws, and nails for fracture fixations. Because the dynamic loadings on bone replacement implants amount to several million cycles artificial joints must also have high fatigue strength.

Low-Expansion Alloys. Cobalt is used as an alloying element in low-expansion iron – nickel alloys of the Invar type. The cobalt content can be up to 25 % and the alloys are used for sealing metals to borosilicate glasses. These alloys have a face-centered cubic solid solution structure that is ferromagnetic within the low-expansion temperature range. The low-expansion behavior is probably caused by a contraction that is due to magnetorestriction canceling out the normal thermal expansion of the alloy. Low-expansion behavior can also be found in the cobalt –chromium – iron alloy system and the composition with 54 % cobalt, 9 % chromium, and 37 % iron has an expansion coefficient close to zero over the temperature range from − 60 to + 60 °C.

Spring Alloys. The alloy Elgiloy is typical of the alloys used in springs. It contains 40 % cobalt, 20 % chromium, 15 % nickel, 7 % molybdenum, 2 % manganese, 0.15 % carbon, 0.04 % beryllium, and the balance being iron. After special treatment, this alloy has a Vickers hardness of 650 N/mm^2, an elastic limit of ca. 160 MN/m^2, and a tensile strength of 2300 MN/m^2.

Alloys with temperature-independent elastic moduli can be used as hair springs for watches and chronometers and for spring balances. Although many of the alloys used for this purpose are based on nickel, a series of cobalt-based alloys containing 50 – 60 % cobalt, 25 – 35 % iron, and 15 – 18 % chromium have suitable properties and have been used for this purpose.

Magnetic Recording Alloys. Much work has been carried out on the development of thin films of cobalt-containing alloys for use as magnetic recording materials [16]. The alloy films can be manufactured by vapor deposition, sputtering, ion-plating, and calendering. Several alloy compositions have been described for use in recording applications; they include cobalt – nickel, copper – nickel – zinc – phosphorus, cobalt – iron – magnesium – carbon, and cobalt – molybdenum – zirconium – niobium. Microcrystalline or amorphous films of the alloys are deposited on polymers and alloys such as Co – Ni (80 : 20), and Co – Mo – Zr – Nb (81 : 7 : 10 : 2) have been used. Video recording tapes with smooth surfaces have also been produced by depositing a ferromagnetic layer of γ-Fe_2O_3– Co powder. The magnetic properties of the films are important in recording applications. The films should have a nearly square *B/H* hysteresis loop and no magnetorestriction.

7. Cobalt Compounds [9, 20]

7.1. Chemical Properties of Cobalt

Cobalt is much less reactive than iron. It is stable to atmospheric oxygen unless heated. When heated, it is first oxidized to Co_3O_4 and then, above 900 °C, to CoO, which is also the product of the reaction between the red-hot metal and steam. The activation energy of the oxidation above 900 °C has been calculated as 155 – 160 kJ/mol. The metal does not combine directly with hydrogen or nitrogen, but it combines with carbon, phosphorus, and sulfur on heating. The reaction with sulfur is influenced by the formation of a low melting eutectic (877 °C) between the metal and the Co_4S_3 phase; the reaction between cobalt and sulfur is rapid above this temperature. Below 877 °C, a protective layer of sulfide scale is formed. In an atmosphere of hydrogen sulfide, cobalt also forms a scale of sulfide, but in air containing sulfur dioxide, a mixed oxide – sulfide scale is formed.

In bulk form, cobalt is resistant to many mild corrosive agents, but it is more readily attacked when it is finely divided. Table 16 gives the corrosion rates of cobalt in some aqueous media. Cobalt is strongly attacked by concentrated nitric acid at ambient temperature. The

metal dissolves slowly in dilute mineral acids, the Co^{2+}/Co potential being − 0.277 V.

Table 16. Corrosion of cobalt in aqueous media at 25 °C

Corrosive medium	Rate of cobalt corrosion, mg dm^{-2} d^{-1}
Distilled water	1.1
5 vol% Ammonia	5.3
10 vol% Sodium hydroxide	5.6
Conc. phosphoric acid	7.4
50 % Phosphoric acid (aqueous)	65.1
5 vol% Acetic acid	12.5
5 vol% Sulfuric acid	56.8
50 % Hydrofluoric acid (aqueous)	178.6

The solubility of oxygen in cobalt is 0.006, 0.013, 0.125, and 0.4 wt % (at 600, 1200, 1510, and 1700 °C, respectively). Nitrogen is only slightly soluble (40 mg/kg) in cobalt at its melting point; the solubility rises to 60 – 70 mg/kg at 1750 °C. The solubility of hydrogen in cobalt increases with temperature from 1 mg/kg at 700 °C to ca. 8 mg/kg at the melting point. The solubility of hydrogen in the liquid metal increases to ca. 20 mg/kg at 1500 °C and to ca. 27 mg/kg at 1750 °C.

The main oxidation states of cobalt are Co^{2+} [*22541-53-3*] and Co^{3+} [*22541-63-5*]. In acid solution and in the absence of complexing agents, Co^{2+} is the stable oxidation state, with oxidation to Co^{3+} being difficult.

$$\left[Co\,(H_2O)_6\right]^{3+} + e^- \rightarrow \left[Co(H_2O)_6\right]^{2+} \Delta E^\circ = 1.84V$$

The oxidation can be achieved by electrolysis or ozone, but the Co^{3+} ion is very unstable and is rapidly reduced to Co^{2+}, even at room temperature, with evolution of oxygen from the water. The solution chemistry of cobalt in acid solution in the absence of complexing agents is dominated by Co^{2+}. The most common species present is $[Co(H_2O)_6]^{2+}$, although the ions $[CoX_3]^-$, $[CoX_4]^{2-}$, and $[CoX_6]^{4-}$ (X = halide) are also found in solutions of hydrogen halides. Cobalt is often removed from solution as its sulfide, and it is interesting to note that although the precipitates of cobalt sulfide obtained by using H_2S are not readily soluble in dilute acids, those obtained by using Na_2S or $(NH_4)_2S$ are soluble. All sulfide precipitates become less soluble as they age in the atmosphere to form Co(OH)S. The cobalt dihalides (except fluoride) are also readily soluble in some organic solvents such as alcohol, acetone, and methyl acetate.

In alkali, Co^{2+} is more readily oxidized to Co^{3+}

$$CoO\,(OH)\,(s) + H_2O + e^- \rightarrow Co(OH)_2\ (s) + OH^- \qquad \Delta E^\circ = 0.17\ V$$

In the presence of complexing agents, oxidation is very easy in any solution because Co^{3+} has a particularly high affinity for complex formation. With N-donor ligands the redox reaction in the presence of ammonia is

$$\left[Co(NH_3)_6\right]^{3+} + e^- \rightarrow \left[Co(NH_3)_6\right]^{2+} \quad \Delta E^\circ = 0.1\ V$$

Oxidation of Co^{2+} solutions containing complexing agents can be achieved with air or hydrogen peroxide. Thus, the solution chemistry of cobalt in the presence of complexing agents is dominated by the complex chemistry of Co^{3+}. The sensitivity of the reduction potential of the Co^{3+}/Co^{2+} couple to different ligands whose presence renders Co^{2+} unstable to air oxidation is shown by the data in Table 17.

Table 17. ΔE° for some Co^{3+}/Co^{2+} couples in acid solution

Redox reaction	ΔE°
$[Co(C_2O_4)_3]^{3-} + e^- \rightleftharpoons [Co(C_2O_4)_3]^{4-}$	0.57
$[Co(EDTA)]^- + e^- \rightleftharpoons [Co(EDTA)]^{2-}$	0.37
$[Co(bipy)_3]^{3+} + e^- \rightleftharpoons [Co(bipy)_3]^{2+}$	0.31
$[Co(en)_3]^{3+} + e^- \rightleftharpoons [Co(en)_3]^{2+}$	0.18
$[Co(CN)_6]^{3-} + H_2O + e^- \rightleftharpoons [Co(CN)_5(H_2O)]^{3-} + CN^-$	− 0.8

EDTA = ethylenediaminetetraacetic acid; bipy = bipyridyl; en = ethylene diamine.

Cobalt(II) Compounds. Cobalt combines with oxygen to form cobalt(II) oxide [*1307-96-6*], CoO, which is stable above 900 °C. This oxide has the sodium chloride structure and is antiferromagnetic at ordinary temperature. When cobalt(II) oxide is heated at 400 – 500 °C in an atmosphere of oxygen, the cobalt(II,III) oxide, Co_2O_3 [*1308-06-1*], is formed. This mixed-valence oxide has the spinel structure

with cobalt(II) in tetrahedral sites and cobalt(III) in octahedral sites. Cobalt hydroxide [*1307-86-4*], $Co(OH)_2$, is a product of the hydrolysis of solutions containing Co^{2+} ions. The hydroxide is amphoteric, dissolving both in alkali, to give blue solutions containing the $[Co(OH)_4]^{2-}$ ion, and in acids. Cobalt(II) forms an extensive range of simple and hydrated salts with all of the common anions including acetate, bromide, carbonate, chloride, fluoride, nitrate, perchlorate, and sulfate. Many of the hydrated salts and their solutions contain the pink octahedral $[Co(H_2O)_6]^{2+}$ ion. Although complexes of cobalt(II) are generally unstable to oxidation, a number of octahedral species are formed, including (1) a series of $[Co(N-N)_3]^{2+}$ complexes with neutral bidentate donor ligands, such as ethylenediamine and bipyridyl, and (2) the acetylacetonates $[Co(acac)_2] \cdot 2\,H_2O$ and $[Co(acac)_2]$. Tetrahedral complexes $[CoX_4]^{2-}$ are usually formed with monodentate anionic ligands, such as chloride, bromide, iodide, thiocyanate, azide, and hydroxyl. Tetrahedral $[CoX_2L_2]$ complexes are formed by a combination of two such ligands with two neutral ligands (L). Addition of cyanide to a solution of Co^{2+} produces a dark green color attributed to the $[Co(CN)_5]^{3-}$ ion.

The colors produced by cobalt(II) complexes in aqueous media have been used to distinguish between octahedral and tetrahedral coordination in the complexes. In general, octahedral species are pink to violet, whereas tetrahedral species are blue. This is not an infallible distinction, but it does provide a useful guide that can be improved by analyzing the electronic spectra of the complexes.

Cobalt(III) Compounds. In addition to the mixed-valence oxide Co_2O_3, impure forms of the unstable cobalt(III) oxide, Co_2O_3, have been prepared. Very few simple cobalt(III) salts are known. The blue sulfate, $Co_2(SO_4)_3 \cdot 18\,H_2O$ [*13478-09-6*], which contains the $[Co(H_2O)_6]^{3+}$ species, can be obtained by the electrolytic oxidation of cobalt(II) in 4M sulfuric acid solution. It is stable when dry, but decomposes in the presence of moisture. The alums $MCo^{III}(SO_4)_2 \cdot 2\,H_2O$ (M = K, Rb, Cs, NH_4) are also known, and a hydrated fluoride $2\,CoF_3 \cdot 7\,H_2O$ has been reported.

The chemistry of cobalt(III) is dominated by complex formation. Cobalt(III) complexes are kinetically inert, and for this reason, indirect methods of syntheses are used to obtain them. Usually the ligand is added to a solution of Co^{2+}, which is then oxidized with some convenient oxidant, often in the presence of a catalyst such as activated charcoal. A wide range of cobaltammines have been prepared and studied; the species identified in solution or as solid derivatives include $[Co(NH_3)_6]^{3+}$, $[Co(NH_3)_5H_2O]^{3+}$, $[Co(NH_3)_5X]^{2+}$ (X = Cl, Br, NO_2, NO_3), and *cis*- and *trans*-$[Co(NH_3)_4X_2]^{+}$ (X = Cl, NO_2).

In addition to these simple cobaltammines, a number of polynuclear species containing bridging groups such as NH_2^-, NH^{2-}, NO_2, OH^-, and O_2 have been prepared. Polynuclear cobaltammines that have been identified include:

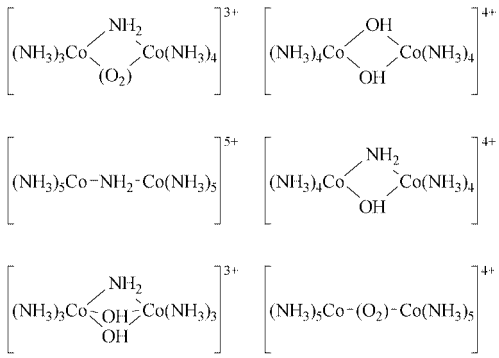

Complexes of cobalt(III) with O-donor ligands are generally less stable than those with N-donor species, although $[Co(acac)_3]$ [*14024-48-7*] and $[Co(C_2O_4)_3]^{3-}$ are known. Apart from the O_2-bridged compounds mentioned above, the octahedral fluoro complexes $[CoF_6]^{3-}$ and $[CoF_3(H_2O)_3]$ are the only known high-spin cobalt(III) complexes, the paramagnetic moment of $[CoF_6]^{3-}$ at room temperature being about 5.8 μ_B. All other cobalt(III) complexes are low-spin and diamagnetic, including the hexacyanocobaltates (III), $[Co(CN)_6]^{3-}$, and the hexanitritocobaltates(III), $[Co(NO_2)_6]^{3-}$.

In addition to the two most stable oxidation states, 2+ and 3+, cobalt also forms compounds in the 1−, 0, 1+, and 4+ oxidation states. There are only a few reported examples of Co^{IV} [*20499-79-0*] compounds, mainly fluoro complexes, $[CoF_6]^{2-}$, and mixed-metal oxides (the purity of compounds in this oxidation state is questionable).

The formation of cobalt compounds in oxidation states lower than 2+ requires the presence of stabilizing π-acceptor ligands. The 1+ state [*16610-75-6*] is not common for cobalt, and most of the known examples are pentacoordinated complexes of the type $[Co(NCR)_5]^+$ (NCR = organic nitrile). Cobalt forms a wide range of complexes in which its formal oxidation state is 0 or 1 – [*16727-18-7*]. Many of these contain ligands such as CO, CN^-, NO^+, and RNC, but other ligands such as tertiary phosphines also stabilize lower oxidation states to give compounds like $[Co(P(CH_3)_3)_4]^+$, which is prepared by reducing an ethereal solution of cobalt(II) chloride with sodium or magnesium amalgam in the presence of trimethylphosphane.

Cobalt Carbonyls [19, 20]. The carbonyl complexes of cobalt are important because of their uses as hydroformylation catalysts (see Section 7.3.3). Because cobalt has an odd number of valence electrons, it can only satisfy the eighteen-electron rule in its carbonyls if Co – Co bonds are formed. For this reason, the principal binary carbonyls of the element are octacarbonyldicobalt [*15226-74-1*], $[Co_2(CO)_8]$, dodecacarbonyl-tetracobalt [*19212-11-4*], $[Co_4(CO)_{12}]$, and hexadecacarbonylhexacobalt [*12182-17-1*], $[Co_6(CO)_{16}]$.

Octacarbonyldicobalt is prepared by heating the metal to 250 – 300 °C at 200 – 300 bar of carbon monoxide or by heating cobalt carbonate under similar conditions in the presence of hydrogen. It is an air-sensitive orange-red solid with a melting point of 51 °C. The compound can be reduced with sodium amalgam in benzene to give the tetrahedral monomeric ion $[Co(CO)_4]^-$, acidification of which leads to tetracarbonylhydridocobalt [*16842-03-8*], $[HCo(CO)_4]$. This hydride is a yellow liquid that forms a colorless vapor. It melts at – 26 °C and decomposes above this temperature to H_2 and $[Co_2(CO)_8]$. It is partly soluble in water to give a strong acid solution containing $[Co(CO)_4]^-$ ions. The hydrogen atom in $[HCo(CO)_4]$ is bound directly to the cobalt, giving a Co – H infrared stretching frequency of ca. 1934 cm^{-1}. The $[Co(CO)_4]^-$ ion is reoxidized to $[Co_2(CO)_8]$ by carbon tetrachloride, and further reaction with this reagent leads to a triply bridged chloromethynyl derivative, from which a carbidocarbonyl compound, disodium carbidopentadecacarbonylhexacobaltate, $Na_2[Co_6C(CO)_{15}]$, can be obtained:

$$10\,[Co_2(CO)_8] + 6\,CCl_4 \xrightarrow[\text{90 \% yield}]{40\,^\circ C} 4\,[Co_3(CO)_9(CCl)] + 8\,CoCl_2 + 44\,CO + C_2Cl_4$$

$$[Co_3(CO)_9(CCl)] + 3\,Na[Co(CO)_4] \xrightarrow[\text{Propylene oxide}]{25\,^\circ C} Na_2[Co_6C(CO)_{15}] + NaCl + 6\,CO$$

Other carbidocarbonyls are obtained by similar routes, including $[Co_6C(CO)_4]^-$, $[Co_8C(CO)_{18}]^{2-}$, and $[Co_{13}(C)_2(CO)_{24}H]^{4-}$. Dodecacarbonyltetracobalt is obtained as a green black solid by heating $[Co_2(CO)_8]$ in an inert atmosphere at 50 °C.

The structure of $[Co_2(CO)_8]$ in the solid state is different from that in solution. The solid-state structure shown in Figure 5 A involves two bridging carbonyl groups and can be best rationalized in terms of the formation of a bent Co – Co bond. In solution, however, this structure is in equilibrium with the form shown in Figure 5 B, in which the dimer is held together by a Co – Co bond. The structure of $[Co_4(CO)_{12}]$ is shown in Figure 5 C. The structure of the carbidocarbonyl (Fig. 5 D) shows the presence of a carbon atom at the center of a distorted square antiprismatic cobalt cluster.

In addition to the carbonyls, cobalt also forms complexes with N_2 and NO^+, which are isoelectronic with CO. Examples of this type of complex are $[CoH(PR_3)_3(N_2)]$ and $[Co(CO)_3NO]$, where R is an alkyl group.

7.2. Commercially Important Cobalt Compounds [9, 16, 21, 22]

Table 18 lists the known applications of cobalt compounds. The most important commercially available compounds are the oxides, as well as hydroxide, chloride, sulfate, nitrate, phosphate, carbonate, acetate, oxalate, and other carboxylic acid derivatives.

Cobalt Oxides. Two main types of cobalt oxide distinguishable by their colors are available: *gray* cobalt(II) oxide, containing 75 – 78 %

Table 18. Industrial uses of cobalt compounds

Compound	CAS reg. no.	Formula	Uses
Acetate(III)	[*917-69-1*]	$Co(CH_3COO)_3$	catalyst
Acetate(II)	[*71-48-7*]	$Co(CH_3COO)_2 \cdot 4\ H_2O$	driers for lacquers and varnishes, sympathetic inks, catalysts, pigment for oil cloth, mineral supplement, anodizing, stabilizer for malt beverages
Acetylacetonate	[*14024-48-7*]	$Co(C_5H_7O_2)_3$	vapor plating of cobalt
Aluminate	[*13820-62-7*]	$CoAl_2O_4$	pigment, catalysts, grain refining
Ammonium sulfate	[*13586-38-4*]	$CoSO_4(NH_4)_2SO_4 \cdot 6\ H_2O$	catalysts, plating solutions
Arsenate	[*24719-19-5*]	$Co_3(AsO_4)_2 \cdot 8\ H_2O$	pigment for paint, glass, and porcelain
Bromide (II)	[*7789-43-7*]	$CoBr_2$	catalyst, hydrometers
Carbonate	[*513-79-1*]	$CoCO_3$	pigment, ceramics, feed supplements, catalyst
Carbonate (basic)	[*7542-09-8*]	$2\ CoCO_3 \cdot Co(OH)_2 \cdot H_2O$	chemicals
Carbonyl	[*15226-74-1*]	$Co_2(CO)_8$	catalyst
Chloride	[*7791-13-1*]	$CoCl_2 \cdot 6\ H_2O$	chemicals, sympathetic inks, hydrometers, plating baths, metal refining, pigment, catalyst
Chromate	[*24613-38-5*]	$CoCrO_4$	pigment
Citrate	[*18727-04-3*]	$Co_3(C_6H_5O_7)_2 \cdot 2\ H_2O$	therapeutic agents, vitamin preparations
Ferrate	[*12052-28-7*]	$CoFe_2O_4$	catalyst, pigment
Fluoride	[*10026-17-2*]	CoF_2	fluorinating agents
	[*10026-18-3*]	CoF_3	
Fluoride	[*13817-37-3*]	$CoF_2 \cdot 4\ H_2O$	catalyst
Fluorosilicate	[*15415-49-3*]	$CoSiF_6 \cdot 6\ H_2O$	ceramics
Formate	[*6424-20-0*]	$Co(CHO_2)_2 \cdot 2\ H_2O$	catalyst
Hydroxide	[*1307-86-4*]	$Co(OH)_2$	paints, chemicals, catalysts, printing inks
Iodide	[*15238-00-3*]	CoI_2	moisture indicator
Linoleate	[*14666-96-7*]	$Co(C_{18}H_{31}O_2)_2$	paint and varnish drier
Manganate	[*12139-69-4*]	$CoMn_2O_4$	catalyst, electrocatalyst
Naphthenate		$Co(C_{11}H_{10}O_2)_2$	catalyst, paint and varnish drier
Nitrate	[*10026-22-9*]	$Co(NO_3)_2 \cdot 6\ H_2O$	pigments, chemicals, ceramics, feed supplements, catalyst
2-Ethylhexanoate	[*136-52-7*]	$Co(C_8H_{15}O_2)_2$	paint and varnish drier
Oleate	[*14666-94-5*]	$Co(C_{18}H_{33}O_2)_2$	paint and varnish drier
Oxalate	[*814-89-1*]	CoC_2O_4	catalysts, cobalt powders
Oxide	[*1307-96-6*]	CoO	chemicals, catalysts, pigments
Oxide	[*1308-06-1*]	Co_3O_4	enamels, semiconductors
Oxides		(mixed metal)	pigments
Dilanthanum tetroxide	[*39449-41-7*]	La_2CoO_4	catalyst, anode
Tricobalt tetralanthanum decaoxide	[*60241-06-7*]	$La_4Co_3O_{10}$	catalyst
Lithium oxide	[*12190-79-3*]	$LiCoO_2$	battery electrode
Sodium oxide	[*37216-69-6*]	$NaCoO_2$	battery electrode
Dicobalt manganese tetroxide	[*12139-92-3*]	$MnCo_2O_4$	catalyst
Dicobalt nickel tetroxide	[*12017-35-5*]	$NiCo_2O_4$	catalyst, anode
Lanthanum trioxide	[*12016-86-3*]	$LaCoO_3$	oxygen, electrode
Phosphate	[*10294-50-5*]	$Co_2(PO_4)_2 \cdot 8\ H_2O$	glazes, enamels, pigments, steel pretreatment
Potassium nitrite	[*17120-39-7*]	$K_3Co(NO_2)_6 \cdot 1.5\ H_2O$	pigment
Resinate		$Co(C_{44}H_{62}O_4)_2$	paint and varnish drier, catalyst
Succinate	[*3267-76-3*]	$Co(C_4H_4O_4) \cdot 4\ H_2O$	therapeutic agents, vitamin preparations
Sulfamate		$Co(NH_2SO_3) \cdot 3\ H_2O$	plating baths
Sulfate	[*10026-24-1*]	$CoSO_4 \cdot x\ H_2O$	chemicals, ceramics, pigments
Sulfide	[*1317-42-6*]	CoS	catalysts
Tungstate	[*12640-47-0*]	$CoWO_4$	drier for paints and varnishes

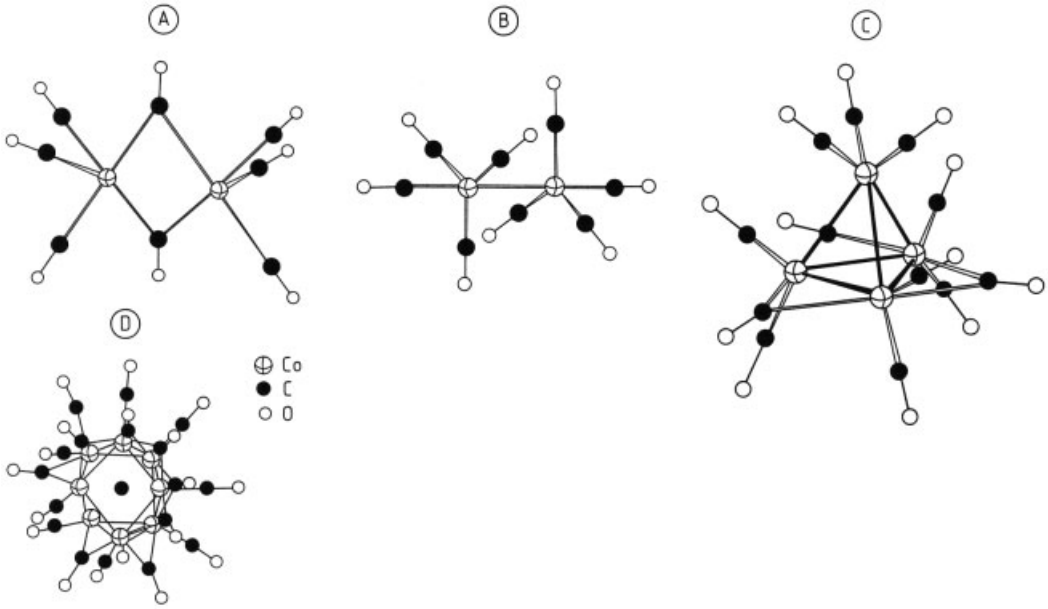

Figure 5. Structures of carbonyl complexes of cobalt
A) $[Co_2(CO)_8]$ (solid state); B) $[Co_2(CO)_8]$ (solution); C) $[Co_4(CO)_{12}]$; D) $[Co_8C(CO)_{18}]$

cobalt, and *black* cobalt(II) dicobalt(III) oxide, containing 70 – 74 % cobalt.

Cobalt(II) oxide, CoO (78.66 % Co), is usually prepared by the controlled oxidation of the metal above 900 °C followed by cooling in a protective atmosphere to prevent partial oxidation to Co_3O_4. Cobalt(II) oxide has a cubic unit cell with $a = 0.425$ nm. It is insoluble in water, ammonium hydroxide, and alcohol, but dissolves in strong acids in the cold and in weak acids on heating. Cobalt oxide (CoO) absorbs a large amount of oxygen at room temperature. Cobalt(II) dicobalt(III) tetroxide, Co_3O_4 (73.44 % Co), can be prepared by the controlled oxidation of cobalt metal or CoO or by the thermal decomposition of cobalt(II) salts at temperatures below 900 °C. The cubic spinel lattice has $a = 0.807$ nm; the solid material readily absorbs oxygen at room temperature, but never transforms into cobalt(III) trioxide, Co_2O_3. The mixed-valence oxide is insoluble in water and only slightly soluble in acids.

Cobalt(II) Hydroxide. $Co(OH)_2$ (63.43 % Co) is prepared commercially as a pink solid by precipitation from a cobalt(II) salt solution with sodium hydroxide. It has a hexagonal crystal structure with $a = 0.317$ nm and $c/a = 1.46$. It is insoluble in water and alkaline solutions, but dissolves readily in most inorganic and organic acids; for this reason, it is commonly used as a starting material in the synthesis of cobalt chemicals. $Co(OH)_2$ decomposes thermally by loss of water, starting at 150 °C, to give the anhydrous oxide at 300 °C. Care must be taken to store the hydroxide in the absence of air because slow air oxidation leads to a product that is poorly soluble in weak acids.

Cobalt(II) Chloride. Cobalt(II) chloride hexahydrate, $CoCl_2 \cdot 6\,H_2O$ (24.79 % Co), is a dark-red deliquescent crystalline compound with a monoclinic unit cell ($a = 0.886$ nm, $b = 0.707$ nm, $c = 1.312$ nm, $\beta = 97°17'$). It is prepared by concentrating a hydrochloric acid solution of cobalt oxide or carbonate. The chloride is very soluble in water, alcohols, and a number of other organic solvents. The solubility of the chloride in aqueous media does, however, decrease with increasing hydrochloric acid content. The hexahydrate dehydrates thermally in three stages, giving the dihydrate at 50 °C, the monohydrate at 90 °C, and the anhydrous chloride at 130 – 140 °C.

The anhydrous chloride and the lower hydrates are very hygroscopic and transform to the

hexahydrate in a moist atmosphere. The fact that the anhydrous chloride is blue and the hexahydrate red is used as a humidity indicator in silica gel desiccants. The chloride is also used in the electroplating, ceramics, glass, chemical, agricultural, and pharmaceutical industries.

Cobalt(II) Sulfate. Cobalt(II) sulfate hexahydrate, $CoSO_4 \cdot 6\,H_2O$ (20.98 % Co), is a brownish-red crystalline compound with a monoclinic unit cell (a = 1.545 nm, b = 1.308 nm, c = 2.004 nm, β = 104.7°). It is prepared by concentrating a sulfuric acid solution of cobalt oxide or carbonate. It is an efflorescent substance and loses one molecule of water when exposed to dry air or when heated gently. The hexahydrate obtained loses water in two stages to give the monohydrate at 100 °C and the anhydrous sulfate above 250 °C. The monohydrate, $CoSO_4 \cdot H_2O$ (34.08 % Co), has been manufactured and sold commercially. Cobalt sulfates are very soluble in water and are generally more stable than cobalt(II) chlorides or nitrates; for this reason they have been widely used as sources of cobalt(II) in solution for the manufacture of chemicals and for the electroplating industries. The sulfates are also used in the ceramics, linoleum, and agricultural industries.

Cobalt(II) Nitrate. Cobalt(II) nitrate hexahydrate, $Co(NO_3)_2 \cdot 6\,H_2O$ (20.26 % Co), is a red-brown crystalline compound with a monoclinic unit cell (a = 1.509 nm, b = 0.612 nm, c = 1.269 nm, β = 119°). It is prepared by concentrating a nitric acid solution of cobalt oxide or carbonate. The hexahydrate loses water rapidly at 55 °C to give the trihydrate; the monohydrate can also be prepared. The nitrates are very soluble in water, alcohols, and acetone. Cobalt nitrate is an important source of high-purity cobalt for use in the electronics and related industries, and the compound has also found uses in the chemical and ceramics industries.

Cobalt(II) Phosphate. Cobalt(II) phosphate octahydrate [*10294-50-5*], $Co_3(PO_4)_2 \cdot 8\,H_2O$ (34.63 % Co), is obtained as a purple flocculent precipitate when an alkali-metal phosphate is added to the solution of a cobalt(II) salt. The phosphate is soluble in inorganic acids and particularly in phosphoric acid, but is insoluble in water or alkaline solutions. It is used in the paints and ceramics industries and is a component of some steel phosphating formulations.

Cobalt(II) Carbonate. Although pure $CoCO_3$ has a cobalt content of 49.57 % and a pseudohexagonal unit cell, the material available commercially is a mauve basic carbonate, $[CoCO_3]_x\,[Co(OH)_2]_y \cdot z\,H_2O$, of indeterminate composition and a cobalt content of 45 – 47 %. The basic carbonate is insoluble in water, but dissolves easily in most inorganic and organic acids. For this reason it is often used as a starting material for the manufacture of other chemicals. The basic carbonate is also used in the ceramics and agricultural industries.

Cobalt(II) Acetate. Cobalt(II) acetate tetrahydrate, $Co(CH_3CO_2)_2 \cdot 4\,H_2O$ (23.68 % Co), is a deliquescent mauve-pink crystalline compound with a monoclinic unit cell (a = 0.847 nm, b = 1.190 nm, c = 0.482 nm, β = 94.3°). It is prepared by concentrating solutions of cobalt hydroxide or carbonate in acetic acid. The tetrahydrate is soluble in water, alcohols, inorganic, and organic acids including acetic acid. The compound loses its water of crystallization at ca. 140 °C. It is used in the manufacture of drying agents for inks and varnishes, catalysts, and pigments, as well as in the anodizing and agricultural industries.

Cobalt(II) Oxalate. Cobalt(II) oxalate dihydrate, $Co(C_2O_4) \cdot 2\,H_2O$ (32.23 % Co), is obtained as a pink precipitate when oxalic acid or an alkali-metal oxalate is added to the solution of a cobalt(II) salt. The dihydrate loses water when it is heated gently in air, and at 200 °C it decomposes to cobalt(II) oxide. The oxalate is insoluble in water, slightly soluble in acids, but soluble in solutions containing ammonia or an ammonium salt. The main use of cobalt oxalate is as a starting material for the preparation of cobalt metal powders.

Cobalt(II) Carboxylates. The cobalt salts of carboxylic acids can be made by the direct reaction of cobalt powder, oxide, or hydroxide with the organic acid or by precipitation reactions involving the addition of the sodium salt of the acid to an aqueous solution of a cobalt salt, such as the sulfate. Cobalt(II) resinate, oleate, linoleate, soyate, naphthenate, ethylhexanoate,

formate, acetylacetonate, and citrate have been produced for use in a variety of applications including catalysts, drying agents, metal film production, and medical uses.

7.3. Industrial Applications of Cobalt Compounds [4, 9, 16, 21 – 24]

The more important applications of cobalt chemicals listed in Table 10 can be considered under the following headings : glasses, ceramics, and refractories; driers, paints, varnishes, and dressings; catalysts; electroplating; electronics and solid-state devices; and agricultural, nutritional, and medical uses.

7.3.1. Glasses, Ceramics, and Refractories [9, 23]

The addition of cobalt oxides to provide a blue pigment for glass, ceramics, and enamels has been used for many centuries (→ Ceramic Colorants). The level of cobalt in the final product depends on the color intensity required, but is typically 0.4 – 0.5 % for colored ceramic bodies, 0.5 % for the blue glass used in welder's goggles, and a few mg/kg in camera lens glass. A range of cobalt-containing mixed-metal oxides and cobalt tripotassium hexanitrite (a yellow pigment) is also available for use as pigments. An indication of the colors obtained from mixed-metal oxide pigments is given in Table 19. The shades of the colors also depend on the exact composition of the oxide and on the method of preparing the pigment. The compositions of some ceramic pigments are given in Table 20.

Table 19. Colors of cobalt-containing pigments

Metals in mixed-metal oxide	Color
Co Al	blue
Co P	violet
Co Zn	green
Co Sn Si	light blue
Co Cr Al	turquoise
Co Mg	pink
Co Fe	brown

Table 20. Compositions of cobalt-containing pigments

Color	Content of Co_3O_4, %	Content of other components (% in parentheses)
Mazarine blue	68.0	SiO_2 (12), cornish stone (16), $CaCO_3$ (4)
Willow blue	33.3	$CaCO_3$ (50), SiO_2 (16.7)
Dark-blue	44.6	Al_2O_3 (55.4)
Matt blue	20.0	Al_2O_3 (60), ZnO (20)
Blue-green	41.8	Al_2O_3 (39), Cr_2O_3 (19.2)
Black	20.6	Fe_2O_3 (41.1), Cr_2O_3 (32.4), MnO_2 (5.9)

The pigments are normally prepared by mixing the ingredients as oxides or as readily decomposable salts and then calcining the mixture at 1100 – 1300 °C prior to milling the product to obtain the pigment as a fine powder. In ceramic applications, the pigments can be added to the ceramic base materials to give a body color or, after being mixed with suitable fluxes, applied as an underglaze (as on Delft china), or overglaze decoration. The final shade after the ceramic has been fired may be modified by reaction between the pigment and the clay base.

Cobalt pigments are used for decolorizing both glass and pottery articles that contain iron oxide, which would give a yellow coloration to the products in the absence of a decolorizer. The yellow color is masked by the complementary blue color of the cobalt pigment added. The levels of cobalt required for the compensation of iron oxide colorations in typical pottery bodies, enamels, and glasses are 0.003 – 0.02 %, 0.002 – 0.01 %, and 1 – 2 mg/kg, respectively. Cobalt compounds have also been used to remove color that can arise in the polymerization of poly(ethylene terephthalate) used in clear plastic bottle manufacture.

The color of the pigments and of the compounds used for decolorizing arise from the optical absorption spectra of cobalt atoms modified by lattice effects in the pigments and in the final products.

In the vitreous enameling of steel sheet, a small amount of cobalt oxide (0.15 – 1.0 %) is included in the ground coat mixture of feldspar, sand, borax, and soda ash to improve the adhesion of the enamel to the steel. As generally accepted, the cobalt does not contribute directly to the adhesion process, but it is involved in cre-

ating suitable conditions for the development of good adhesion.

Ceramic coatings for low-carbon and low-alloy steels for high-temperature use can consist of a mixture of aluminum oxide and two cobalt coats, one of which is high-firing and the other low-firing. The purpose of the latter is to seal off oxygen at low temperatures to protect the base metal from oxidation from the start of the firing. These coatings are claimed to withstand combustion products and corrosive vapors up to 750 °C and have been used to protect flue gas and gas turbine exhaust pipes.

The metal – glass coatings used to protect plain carbon steels typically contain 80 % powdered chromium and 20 % of a borosilicate glass containing 5 % cobalt oxide. Superalloys have also been coated with a ceramic consisting of high-firing cobalt ground coats with high aluminum oxide content or with boron nitride containing a lithium cobalt oxide binding agent.

7.3.2. Driers, Paints, Varnishes, and Dressings [9]

A number of cobalt pigments have been used in oil paintings, including Cobalt blue smalt [*1345-16-0*] (a silicate), Thenard's blue [*13820-62-7*] (an aluminate), Cerulean blue [*68187-05-3*] (a stannate), Cobalt violet [*82196-89-2*] (a phosphate), and Aureolin yellow [*13782-01-9*] (potassium cobalt(III) nitrite). Some of the pigments have also been used for inks, for printing on fabrics and paper, and for coloring plastics.

Cobalt acetate is used for coloring the oxide layers in anodized aluminum to give shades varying from bronze to black. The anodized metal is immersed in a 20 % solution of cobalt acetate and then either in a solution of ammonium sulfide to produce cobalt sulfide (black) or in a solution of potassium permanganate (bronze). The different depths of bronze shades are achieved by different numbers of alternate immersions in the cobalt acetate and permanganate solutions. In addition to its use as a coloring agent, cobalt acetate is also used to improve the lightfastness of anodized aluminum dyed with organic pigments.

The cobalt(II) salts of a number of organic acids have been used as drying agents for unsaturated oils in paints, varnishes, and inks. The salts used for this purpose include the oleate, ethylhexanoate, naphthenate, soyate, linoleate, resinate, and tallate. These cobalt salts are either soluble in the unsaturated oils or they react with them to form soluble compounds that then act both as oxidation accelerators of the oil and as polymerization catalysts, forming a film with increased stability, resistance, and flexibility. The drying action is due entirely to the redox behavior of the cobalt species present; the carboxylate moieties simply act as oleophilic groups to solubilize the cobalt compounds in the oils. The amounts of cobalt added to the oils as drying agents are in the range 0.01 – 0.6 %; the salts can be added as solutions in organic solvents or as dispersions of ultrafine powders (→ Drying Oils and Related Products). Cobalt-containing drying agents are also used in low-temperature curing processes for silicone resins. The cobalt species promote drying at the surface of the film in contact with atmospheric oxygen. Its action is so fast that a hard polymeric film can be formed that prevents deep drying; for this reason cobalt is often used in conjunction with other metals such as zinc and calcium that do not have a drying action but slow down and control the effects of cobalt.

7.3.3. Catalysts [24]

Cobalt compounds are versatile catalysts. Cobalt-containing materials have been used to catalyze the following types of reactions: hydrogenation; dehydrogenation; hydrogenolysis including hydrodenitrification and hydrodesulfurization; hydrotreatment of petroleum products; selective oxidation; ammonoxidation; oxidation; hydroformylation; polymerization; selective decomposition; and ammonia synthesis. In addition to these more general reactions, there are a number of reactions for which cobalt chemicals are among the best catalysts. These include their uses as driers, described in Section 7.3.2., their use in the oxidation of ammonia to nitric acid, and the syntheses of fluorocarbons.

The effectiveness of cobalt as a catalyst is related to the ease with which the element forms complexes and particularly to the large variety of ligands in these complexes.

The main industrial processes using cobalt catalysts are the removal of sulfur from vari-

ous petroleum-based feedstocks (hydrodesulfurization), selective liquid-phase oxidation, and hydroformylation. The catalysts used in the hydropurification reactions including *hydrodesulfurization* consist of cobalt and molybdenum oxides supported on an inert material such as alumina and contain 2.5 – 3.7 % cobalt. The homogeneous catalysts used in the *liquid-phase oxidation* processes are soluble cobalt(II) carboxylates such as acetate, naphthenoate, and oleate. Relatively large concentrations of cobalt (0.1 – 1 %) are used as cobalt salts in the catalysis of *hydroformylation* reactions in which $[HCo(CO)_4]$ is the active material. The amounts of cobalt catalysts used in hydrodesulfurization, oxidation, and hydroformylation have been estimated at 800 – 1800, 200, and 800 t/a, respectively.

In *hydrotreating,* the cobalt – molybdenum catalysts function in the sulfide form and catalyze two reactions, hydrogenation and hydrogenolysis of carbon – heteroatom bonds. Two theories have been used to explain the enhanced catalytic activity obtained by having both cobalt and molybdenum oxides in the catalysts: the *pseudointercalation model* and the *remote control model.* The first model assumes an association between molybdenum sulfide (MoS_2) and cobalt to produce active sites for the catalysis. The remote control model, on the other hand, assumes that active centers are created in the MoS_2 lattice by a mobile species of hydrogen produced at a cobalt sulfide component. The cobalt in the catalyst thus controls the nature and the number of catalytically active sites formed.

The types of *oxidations* catalyzed by cobalt salt solutions include *p*-xylene to terephthalic acid, cyclohexane to adipic acid, and hydrocarbons or acetaldehyde to acetic acid. All of these liquid-phase oxidations follow classical radical chain mechanisms. The following reaction mechanism for cyclohexane oxidation shows the role of cobalt in decomposing the hydroperoxides formed in the reaction and thus avoiding unwanted side reactions:

Initiation:

$$C_6H_{12}+R\cdot \rightarrow C_6H_{11}\cdot + RH$$

Chain reaction:

$$C_6H_{11}\cdot + O_2 \rightarrow C_6H_{11}OO\cdot$$

$$C_6H_{11}OO\cdot + C_6H_{12} \rightarrow C_6H_{11}OOH + C_6H_{11}\cdot$$

Hydroperoxide decomposition:

a) $C_6H_{11}OOH \longrightarrow$ cyclohexanone + H_2O

b) $Co^{2+} + C_6H_{11}OOH \longrightarrow Co^{3+} + C_6H_{11}O\cdot + OH^-$
$Co^{3+} + C_6H_{11}OOH \longrightarrow Co^{2+} + C_6H_{11}OO\cdot + H^+$
$\longrightarrow$ chain

c) $C_6H_{11}OOH$ + all radicals $\longrightarrow$ degradation products, cyclohexanol, and cyclohexanone

In *hydroformylation* and *hydroesterification,* oxo reactions such as those listed below occur (→ Carbonylation; → Oxo Synthesis):

$$CH_3OH + CO \rightarrow CH_3COOH$$

$$RCH=CH_2 + CO + H_2 \rightarrow RCH_2CH_2CHO$$

$$RCH=CH_2 + CO + 2H_2 \rightarrow RCH_2CH_2CH_2OH$$

$$RCH=CH_2 + CO + R'OH \rightarrow RCH_2CH_2COOR'$$

The cobalt salts added to the reaction mixture are converted into $[HCo(CO)_4]$, which catalyzes the oxo reaction by the following types of mechanism:

$$[HCo(CO)_4] + RCH=CH_2 \rightarrow [RCH_2CH_2Co(CO)_4]$$

$$[RCH_2CH_2Co(CO)_4] + CO \rightarrow [RCH_2CH_2CO-Co(CO)_4]$$

$$[RCH_2CH_2CO-Co(CO)_4] + H_2 \rightarrow RCH_2CH_2COH + [HCo(CO)_4]$$

In addition to the main types of reactions catalyzed by cobalt species, cobalt compounds have also been used as catalysts for automobile exhaust gas purification, for the manufacture of nitric acid, for heterogeneous oxidation reactions, and for the oxidation of toluene to benzoic acid. Cobalt(II) dicobalt(III) tetroxide is one of the

most effective catalysts for the oxidation of carbon monoxide and also has a high activity in oxidizing nitrogen monoxide; these properties have led to the use of Co_3O_4 and the $LaCoO_3$ (perovskite phase) as components of exhaust gas purification catalysts. The oxide Co_3O_4 is also a highly selective catalyst in the oxidation of ammonia to nitric acid, although it loses activity due to sintering after repeated use and due to poisoning by sulfur dioxide.

Cobalt is often cited in the patent literature as an additive in nearly all heterogeneous oxidation catalyses, including the oxidations of propene, butane, butenes, and methanol. New systems containing cobalt can also be used to catalyze the conversion of thiols to disulfides in gasoline sweetening.

Cobalt compounds have considerable potential in *homogeneous catalytic systems* and are used with copper in the oxidation of toluene to benzoic acid; good selectivities for the conversion of methanol to ethanol have been achieved by using cobalt cluster catalysts.

Cobalt fluoride can be used as a catalyst in the *fluorination* of hydrocarbons. These reactions utilize the cobalt fluoride as a fluorine carrier and take place in two steps – oxidation of CoF_2 to CoF_3 at 150 – 200 °C followed by fluorination of the hydrocarbon at 250 – 350 °C. Both stages of the reaction are exothermic, and careful control of the reaction is needed to obtain optimum yields.

Cobalt has been a component of the catalysts used in the *Fischer – Tropsch synthesis* of hydrocarbons from synthesis gas. The mixed cobalt – thorium oxide – magnesium oxide – kieselguhr catalyst (100 : 5 : 8 : 200), for example, is generally prepared by boiling solutions of sodium carbonate containing cobalt and thorium nitrates to precipitate the metals as carbonates prior to adding magnesium oxide and kieselguhr. The mixture of metal carbonates and support materials is then filtered off, washed, dried, milled, and reduced in a hydrogen atmosphere to produce the Fischer – Tropsch catalyst.

Raney cobalt can be prepared by leaching a finely powdered aluminum – cobalt alloy containing 40 – 50 % cobalt with sodium hydroxide at 15 – 20 °C, which dissolves the aluminum to leave a very active porous cobalt residue. The properties of Raney cobalt are similar to those of Raney nickel, and the materials can be used to catalyze the same types of reaction.

7.3.4. Electroplating [9]

Cobalt is readily electrodeposited from a number of electrolyte solutions. Pure plated layers of cobalt are of little commercial value, but cobalt is an important component of a number of alloy electrodeposits. The electrochemical equivalent of cobalt is 1.099 g A^{-1} h^{-1} and, with cathode efficiencies of 90 – 100 % for most electrolytes, the deposition rates are ca. 1 g A^{-1} h^{-1}. Solutions of cobalt sulfate, chloride, sulfamate, or fluoroborate have been used as electrolytes. The compositions of typical plating baths are listed in Table 21. To obtain a smooth metal deposit at a cathode efficiency of 90 – 100 % and a current density of 5 A/dm^2, the bath should have a high cobalt ion concentration (65 – 100 g/L) and should be operated at a temperature of 25 – 50 °C and a pH of 4 – 5.

Additions of cobalt sulfate to nickel plating baths have been used to produce bright nickel plates consisting of nickel – cobalt alloys. The proportion of cobalt in the electrolyte controls the composition of the alloy deposited. Since the deposition potential of cobalt is less than that of nickel, the cobalt – nickel ratio is higher in the deposit than in the electrolyte. A typical bath composition is given in Table 21. If this bath is operated at 55 – 60 °C, pH 4 – 4.3, and with a current density of 2.5 – 4 A/dm^2, an alloy deposit containing 18 % cobalt is produced. The cobalt content of the bath must be maintained by the addition of the cobalt salt when nickel anodes are used. Cobalt sulfamate is used as an alternative to cobalt sulfate in some chloride – sulfate nickel plating baths.

Cobalt – tungsten alloys with tungsten contents of up to 10 % can be deposited from baths containing cobalt sulfate and sodium tungstate (typical electrolyte composition given in Table 21). A temperature of 70 – 90 °C and a pH of 7 – 9 are used to obtain smooth crack-free cobalt –tungsten deposits. Cobalt – molybdenum alloys can be obtained under similar conditions when sodium molybdate replaces the sodium tungstate in the electrolyte. A number of other cobalt-containing alloys can be obtained as electrodeposits, including cobalt – iron, cobalt –

Table 21. Compositions of cobalt-containing plating baths

Electrolyte composition			Content,	Metal
Metal salts	Content, g/L	Additives	g/L	deposited
$CoSO_4 \cdot 7\,H_2O$	332	H_3BO_3	30	Co
$CoCl_2 \cdot 6\,H_2O$	300	H_3BO_3	30	Co
$CoSO_4(NH_4)_2SO_4 \cdot 6\,H_2O$	200	H_3BO_3	25	Co
$Co(SO_3NH)_2 \cdot 4\,H_2O$	450	$HCoNH_2$	30*	Co
$Co(BF_4)_2$	116 – 154	H_3BO_3	15	Co
$NiSO_4 \cdot 7\,H_2O$	240	H_3BO_3	30	Co – Ni
$NiCl_2 \cdot 6\,H_2O$	45	$NaHCO_2$	35	Co – Ni
$CoSO_4 \cdot 7\,H_2O$	3 – 15			
$CoSO_4 \cdot 7\,H_2O$	150	NaCl	28	Co – W
$NaWO_3$	15 – 50	H_3BO_3	40	Co – W
		Na heptanoate	100	Co – W

* mL/L.

platinum, cobalt – gold, and cobalt – phosphorus.

Electrodeposited cobalt and cobalt alloys have been used as matrices for composite wear-resistant coatings in systems that codeposit suspended particles of materials such as alumina and silicon carbide with the plating metal.

7.3.5. Electronics, Solid-State Devices, and Batteries [16, 22]

Since 1990 there has been increasing interest in the use of cobalt-containing materials in the electronics industry. The most important applications are magnetic recording and batteries.

Cobalt-containing magnetic recording media are of two types, (1) oxide or metal particles and (2) metal films. Cobalt-modified iron oxide particles are presently the predominant material in video tapes, while cobalt and particularly cobalt-chromium alloys are used in perpendicular recording applications. In digital storage of optical data and bubble memories, high recording densities are required, and cobalt alloys such as Gd – Co or Gd – Fe – Co are used in these magneto-optic devices.

The three major types of high energy density battery use substantial amounts of cobalt, and the expected consumption of the metal for this application is about 3000 t/a. All lithium ion batteries use *p*-type semiconductor $LiCoO_2$ as the positive electrode active material to operate in the range around 4 V that is unique to this type of battery. The purpose of the lithium cobalt oxide phase is to achieve charging and discharging by the reversible exchange of Li between an LiC_6 electrode and the Li^+ ions in the oxide lattice. The electrochemical reactions in Ni – Cd storage batteries involve the reduction of Ni(III) to Ni(II) and the oxidation of another metal species during discharge with the reverse reactions occurring during charging. When Co is used as the other metal in the cell the discharge reaction is:

$$Co + 2\,NiO\,(OH) + 2\,H_2O \rightarrow Co(OH)_2 + 2\,Ni(OH)_2$$

The cobalt in these cells is part of a porous positive electrode formed by impregnating Ni powder on a perforated metal foil with co-precipitated nickel and cobalt hydroxides. Cobalt is an important component of both electrodes of nickel/metal hydride rechargeable batteries. In the hydride electrode, cobalt reduces the volume change during the hydrogenation – dehydrogenation cycles and is an important component of the protective surface layer formed on the electrode. Both functions improve the electrode stability by minimizing corrosion and decrepitation and are essential factors in determining the cycle life of the cells. In the positive electrode the presence of cobalt helps to reduce the effects of swelling on the sintered nickel electrodes. The advantages of using cobalt along with nickel in the Ni – Cd and Ni hydride batteries stem from three factors: The similarities between the crystal structures of cobalt and nickel compounds, the greater ease of oxidation of Co(II) compared with Ni(II), and the higher electrical conductivities of cobalt oxide phases.

Black cobalt oxide, Co_3O_4, is a selective coating material for high-temperature solar collectors. The cobalt oxide coatings are superior to the black chrome coatings often used. The function of cobalt oxide as a coating intended to operate up to 1000 °C is to concentrate the solar radiation on the collector surface by a factor of up to 2000. For this purpose, a high solar absorbance with low infrared emittance is required. Cobalt oxide coatings can be prepared by plating, pyrolysis of cobalt salts on heated substrates, vacuum deposition, and by use of Co_3O_4-based paints. Table 22 gives absorbance and emittance data for cobalt oxide coated solar panels prepared in a variety of ways on different substrates.

Cobalt(II) oxide has been used in thermistors to improve both the resistivity and the temperature coefficient of the resistivity of the device. A number of mixed-metal cobalt oxides have been used in devices, including $NiCoO_4$ and La_2CoO_4 as anodic materials in water electrolysis, $NaCoO_2$ as cathodes in sodium batteries, and $CoMoO_4$ as a fuel cell electrode.

7.3.6. Agriculture, Nutrition, and Medicine [21]

Cobalt chemicals are used to correct cobalt deficiencies in soils and in animals. Soil treatments usually involve top dressings containing cobalt sulfates, whereas treatment of ruminant animals involves the use of either salt licks containing ca. 0.1 % of cobalt as sulfate, or concentrated feeds, or pellets of cobalt oxide bound in an inert material such as china clay.

The medicinal uses of cobalt are dominated by the use of vitamin B_{12}. The vitamin is obtained from the mother liquors of the microbial formation of antibiotics such as streptomycin, aureomycin, and terramycin after the removal of the antibiotic. Vitamin B_{12} and cobalt treatments are used as remedies in certain types of anemia (→ Vitamins). Cobalt salts, administered intravenously, have been used as an antidote in cyanide poisoning.

The mechanisms of the agricultural and medicinal uses of cobalt chemicals are discussed in Chapter 10 of this article.

8. Analysis [25, 26]

Cobalt can be determined gravimetrically as anthranilate, as Co_2O_3 or Co after precipitation of the 1-nitroso-2-naphthol complex and subsequent decomposition, and as a pyridine thiocyanate. Volumetric methods include the use of 8-hydroxyquinoline, ethylenediamine tetraacetic acid (EDTA), and iron(III) cyanide solutions as titrants. In the case of complexometric methods, cobalt is often back-titrated with magnesium or zinc solutions after an excess of EDTA has been added. The metal can also be determined by electrodeposition on perforated platinum electrodes.

Trace levels of cobalt in most materials can be determined by a number of instrumental techniques, including X-ray fluorescence, activation analysis, atomic absorption, atomic emission, and polarography. The reagents that can be used to measure cobalt by colorimetry include dimethylglyoxime, dithizone, 8-hydroxyquinoline, nitrosonaphthols, diethyldithiocarbamate, and tripyridyl.

9. Economic Aspects [4, 27]

Although in the third quarter of the 20th century, cobalt production was very dependent upon the situation in African countries, this dependence has been eliminated with the increase in production in the rest of the world. In the period 1970 – 1979, over 70 % of available production came from Zambia and Zaire, but this had been reduced to about 33 % by 1998. The statistics for cobalt production by the major producer companies and by other countries are given in Table 23 for the period 1992 – 1997. These supply figures can be compared with the current demand for the metal (excluding demand in the CIS) of about 29 500 t. Estimated demand by end use for 1998 was: superalloys (24.3 %), hardfacing and other alloys (6.9 %), magnets (8.5 %), hardmetals (15.2 %), catalysts (8 %), pigments (11.5 %), batteries (9.5 %)), tires and driers (11 %), other chemicals (5.1 %). Some forecasts suggest that the cobalt consumption could double in the period up to 2012 and that the price sensitivity to any such increase will depend on the speed with which new sources of cobalt from laterite ores are developed. The marginal cost of production

Table 22. Properties of cobalt oxide coated solar panels

Coating method	Substrate	Absorbance, %	Emittance, %
Plated	Cu, Ni	93	28
Plated	Ni-plated steel	90	7
Plated/thermal decomposition	Ag-coated steel	> 90	20
Plated	Ni	95	–
Plated	Cu-coated steel	96	13
Oxidized paint	Al	91	30
Paint	Ni alloy	85	10
Spray pyrolysis	Al, galvanized Fe	91, 92	13, 12
Spray pyrolysis	Stainless steel	93	14

has been estimated at about $ 6 per pound. The cobalt price (for 99.8 % metal) has fluctuated from a low of about $ 7.8 per pound in 1990 to peaks of between $ 30 and $ 35 in 1992 and 1996 to $ 15 – 16 per pound in 1999.

10. Toxicology

Distribution in the Environment. Data on the distribution of cobalt in the environment have been reviewed [29, 30]. Cobalt concentrations in ambient air are exceedingly low; in remote areas atmospheric concentrations are below 1 ng/m^3, in urban environments in the range of 1 – 10 ng/m^3, and in the vicinity of cobalt processing industries above 10 ng/m^3. The cobalt content of soils is in the range of 5 – 70 mg/kg; the average is estimated to be 8 mg/kg. In fresh water, cobalt concentrations are generally low, between 0.1 and 10 µg/L [29], but considerably higher in the vicinity of metal mining and processing industries. Cobalt is mobilized into surface waters by weathering of rocks and transported into the oceans, where it is mainly coprecipitated with iron as hydroxide. Hence, the cobalt concentration in ocean waters is very low (ca. 2 ng/L) and cobalt is enriched in deep-sea sediments.

Radioactive cobalt isotopes, especially ^{60}Co, have been released into the environment from nuclear power plants, radioactive waste dumping, and nuclear accidents (Chernobyl). Because health effects of ^{60}Co are predominantly attributable to radiation and not chemical properties of cobalt, radioactive cobalt is not treated here and readers are referred to other sources, e.g., the U.S. Toxicological Profile for Cobalt [30].

Nutrition. Levels of cobalt in drinking water usually are very low (0.1 – 5 µg/L) and do not significantly contribute to the human supply of cobalt [30, 31]. The human dietary intake of cobalt varies depending on the type of food [30]. Relatively high concentrations of about 1 mg/kg dry matter (d.m.) are found in cocoa, tea, and coffee. Green vegetables and unprocessed cereals have cobalt contents ranging between 10 and 600 µg/kg d.m., and meat and fish contain only 1 – 50 µg/kg d.m. The total human intake of cobalt is estimated to be about 10 – 30 µg/d [29]. The cobalt content of nutrients has gained much attention because of the role of cobalt as the central constituent of vitamin B_{12} (cobalamin). This vitamin is the only utilizable form of cobalt, whereas inorganic cobalt is of no use for mammalian health. Vitamin B_{12} occurs almost entirely in food of animal origin, and it constitutes only a small fraction of the total cobalt in the human diet. The recommended human intake for vitamin B_{12} is 2.5 – 5 µg/d, corresponding to only 0.1 – 0.2 µg Co/d [29, 30].

Essentiality and Toxicity.

Essentiality. Inorganic cobalt is essential for the nutrition of ruminants, which require cobalt for the bacterial biosynthesis of vitamin B_{12} in their rumen (first stomach). The normative cobalt requirement of ruminants is 0.13 mg/kg feed d.m. [32]. Additions of 2 – 4 mg/d to the feed of sheep and 20 mg/d to the feed of cattle are tolerated, whereas larger doses are toxic [29]. Vitamin B_{12} deficiency in humans causes pernicious anemia. This is a very rare disease and has been observed only with strict vegetarians after decades because of the long half-life of the vitamin in the body. The symptoms of the

Table 23. Annual production statistics for refined cobalt for major countries and producers (excluding production in CIS for internal consumption)

Country	Producer	1992	1993	1994	1995	1996	1997
USA	OMG	2100	2200	3000	3610	4160	5000
Zambia	ZMCC	4610	4211	2639	2934	4799	3949
Canada	Falconbridge	2300	2414	2923	2804	3099	3417
Zaire/Congo	Cecamines	5023	829	2439	3422	3540	2808
Bahamas	ICCI	688	1218	1820	1730	2070	2250
Canada	INCO	1465	1410	1130	1362	1544	1500
Australia	QNI						657
Japan	Sumitomo	105	190	161	222	228	263
Belgium		1000	1000	1000	1200	1200	1200
China			933	800	1076	1200	1200
RSA		234	243	246	190	292	294
Brazil		88	158	165	166	193	266
Morocco						80	220
France		121	144	141	161	174	159
India							110
CIS exports		1199	1960	2070	2787	1654	3200
Total		**18933**	**16910**	**18534**	**21664**	**24233**	**26493**

disease are neurological and mental defects and low erythrocyte counts.

Resorption, Distribution, and Elimination. After oral uptake, the absorption of cobalt occurs in the small intestine by the iron uptake system. In the blood, cobalt ions are preferentially bound to serum albumin and distributed to the organs [29]. The average concentrations in human organs are 0.09 µg/L in whole blood, 0.25 µg/L in urine, and 42.3 µg/kg in liver. The liver is the main location of vitamin B_{12}; it contains about 50 – 90 % of the total human vitamin B_{12} content [29].

After inhalation by exposed workers of cobalt metal dust or sparingly soluble cobalt compounds, these are taken up by lung macrophages, solubilized, and resorbed from the lung into the blood and excreted via the urine. Elevated internal concentrations of cobalt have been found in workers of the cobalt and hard metal industries [30, 31]. Inorganic cobalt is excreted slowly via the kidney with half-lives up to 90 d. Vitamin B_{12} is retained even longer with a half-life of about 400 d [29].

Toxic Effects.

Oral Toxicity. Oral toxicity of cobalt compounds has been observed rarely. After administration of high doses of cobalt chloride (100 mg/d) as a remedy for certain forms of anemia, adverse side effects on the alimentary tract were observed [30, 31]. Severe outbreaks of cardiomyopathy were recorded among heavy consumers of beer fortified with cobalt as an foam-controlling agent [30, 31]. Also, in experimental animals, the acute oral toxicity of cobalt is moderate. The oral LD50 values for various inorganic cobalt salts ranged from 150 to 500 mg/kg body weight (b.w.) for the anhydrous compounds, corresponding to 1.5 – 3 mmol Co/kg b.w. [30, 31].

Skin Sensitization. Repeated dermal exposure of metal workers to cobalt and/or cobalt compounds has resulted in a contact dermatitis manifesting as eczema and erythema [30].

Pulmonary Effects. The major concern in cobalt toxicology are pulmonary diseases caused by inhalation of cobalt and cobalt compounds, as occurs in the production and processing of cobalt and cobalt compounds and in the production of hard metals composed of cobalt and tungsten carbide. The symptoms include decreased pulmonary function, asthma, inflammation, and ultimately lung carcinoma [30, 31, 33]. For the nonmalignant pulmonary effects, the U.S. Agency for Toxic Substances and Disease Registry has estimated a LOAEL (lowest observed adverse effect level) of 0.0151 mg Co/m^3 and a NOAEL (no observed adverse effect level) of 0.0053 mg Co/m^3 [30].

Carcinogenicity. Few epidemiological studies on cohorts of cobalt-exposed workers have been performed. They showed some evidence for lung carcinogenicity of cobalt and cobalt compounds [31, 34]. However, no firm conclu-

sions about the human carcinogenicity of cobalt and cobalt compounds can be drawn because of small numbers, limited statistical confidence, and co-exposure to other metals. Studies on workers employed in the production of hard metals revealed a significant correlation between exposure and incidence of lung cancer [33]. In experimental inhalation studies with both rats and mice, cobaltous sulfate evoked the formation of lung tumors [35].

Table 24. Classification of cobalt and cobalt compounds as carcinogens [30], [31], [33], [34]

Agency	Substances	Classification
IARC 1990 [31]	cobalt and cobalt compounds	2B (possibly carcinogenic to humans)
EU 2004	cobalt dichloride	2 (considered to be carcinogenic for humans)
	cobalt sulfate	2 (considered to be carcinogenic for humans)
DFG 2001 [34]	cobalt and cobalt compounds	2 (considered to be carcinogenic for humans)
DFG 2004 [33]	hard metal, composed of tungsten carbide and cobalt	1 (causes cancer in humans)
Germany 2004	cobalt (elemental)	3 (possibly carcinogenic to humans)
	soluble cobaltous salts	2 (considered to be carcinogenic for humans)
	cobalt oxide and cobalt sulfide	3 (possibly carcinogenic to humans)
ACGIH 2003	cobalt, elemental and inorganic compounds	A3 (confirmed animal carcinogen with unknown relevance to humans)

The epidemiological and animal data are supported by mechanistic studies in vitro which reveal that cobalt compounds are weakly genotoxic but that cobalt enhances the genotoxicity of other agents because it inhibits the repair of damaged DNA [34]. Furthermore, the genotoxicity of cobalt-containing hard metals is severalfold higher than the genotoxicity of the cobalt component alone. This enhancing effect is attributed to a catalytic role of tungsten carbide in the formation of reactive oxygen species [30, 33].

Reproductive Toxicity. There are no data about the reproductive toxicity of cobalt in humans. In experiments with male mice, high oral doses of $CoCl_2$ caused reduced sperm count, testes weight, and fertility [30, 34]. After mating of treated males with untreated females, embryonic losses were registered. Also with male rats, oral administration of $CoCl_2$ caused degeneration and necrosis of the testes and reduced spermatogenesis [30, 34].

Environmental Toxicity. There are little data about the environmental toxicity of cobalt and cobalt compounds in microorganisms, plants, and animals except farm animals. Accidental feeding of several grams of cobalt to sheep have caused fatalities [29]. Extrapolating from experimental data, it is probable that the levels of cobalt encountered in the environment do not exert significant toxicity on environmental species. Of course, this conclusion does not apply to biological effects of effluents and waste dumping sites of metal industries.

Regulations and Guidelines. The qualitative classification of cobalt and cobalt compounds with respect to carcinogenicity is summarized in Table 24. Whereas previously some regulatory agencies have classified cobalt and its compounds as a whole group, recent classifications tend to classify cobalt compounds individually according to the degree of evidence for carcinogenicity to humans. Table 25 summarizes quantitative toxicity and limit values issued by various agencies. This table reveals considerable differences in quantitative risk estimation between agencies of different countries. Obviously, there is a lack of data with respect to quantitative effect – exposure relations and a need for harmonization of regulations in the future.

Table 25. Synopsis of limit values for cobalt and cobalt compounds (as elemental cobalt).

Description	Limit values	Source
Occupational limit values for Cobalt and inorganic compounds	0.5 mg/m^3 TELa	Germany 2004
	0.1 mg/m^3 TELb	Germany 2004
	0.1 mg/m^3 PELc	OSHA 2003
	0.05 mg/m^3 RELd	NIOSH 2001
	0.02 mg/m^3 TLVe	ACGIH 2000
Occupational biological exposure equivalent for cobalt and its compounds	25 µg/L in blooda300 µg/L in urinea	Germany 2004
	5 µg/L in bloodb60 µg/L in urineb	Germany 2004
Occupational biological exposure index for cobalt and its compounds	15 µg/L in urine	ACGIH 2000
	1 µg/L in blood, both at end of shift at end of work week	ACGIH 2000
Food. Cobaltous salts and their derivatives	Prohibited from use in human food	FDA 2000

a Technical exposure limit, production of cobalt powder, catalysts, hard metals, magnets.
b Technical exposure limit, all other processes.
c Permissible exposure limit.
d Recommended exposure limit.
e Threshold limit value.

11. References

1. *Cobalt Monograph*, Centre d'Information du Cobalt, Brussels 1960.
2. R. S. Young: "Cobalt," *ACS Monogr. Ser.* **149**, New York 1960.
3. W. Betteridge: *Cobalt and its Alloys*, Ellis Horwood, Chichester 1982.
4. "Cobalt Facts", Cobalt Development Institute, Guildford UK — a regularly updated document.
5. N. N. Greenwood, T. C. Gibb: *Mössbauer Spectroscopy*, Chapman Hall, London 1971.
6. *Cobalt and Cobalt Abstracts*, Centre d'Information du Cobalt, Brussels 1958 – 1975.
7. S. M. Grimes (ed.): "Cobalt Abstracts," *ATB Metall.* **24** (1984)
   Cobalt Abstracts in *ATB Metall.*, Cobalt Development Institute, UK, 1985 – 1994 and *Cobalt News*, Cobalt Development Institute, UK 1995 – 1999.
8. W. Betteridge, *Prog. Mater. Sci.* **24** (1979) 2.
9. J. D. Donaldson, S. M. Grimes, S. J. Clark: *Cobalt in Chemicals*, The Cobalt Development Institute, Guildford 1986.
10. R. W. Andrews: *Cobalt: Overseas Geol. Surveys*, Mineral. Resources Div., H. M. Stationary Office, London 1962.
11. H. Hagon, *Metall (Berlin)* **29** (1975) no. 11, 1157 – 1158.
12. H. Hagon, *Met. Powder Rep.* **31** (1976) no. 1.
13. J. D. Donaldson, S. J. Clark: "*Cobalt in Superalloys*," The Monograph Series Cobalt Development Institute, London 1985.
14. S. Rimlinger: *Proc. The Cobalt Conference*, The Cobalt Development Institute, Guildford 1999.
15. C. T. Sims et al.: *Superalloys*, vol. **II**, Wiley, New York 1987.
16. J. D. Donaldson, S. J. Clark, S. M. Grimes: *Cobalt in Electronic Technology*, The Cobalt Development Institute, Guildford 1998.
17. M. S. Pierre, *Cobalt News* **1999**, no. 2, 3.
18. A. Magnee, J. M. Drapier, J. Dumont, D. Coutsouradis, L. Habraken: *Cobalt-Containing High-Strength Steels*, Centre D'Information du Cobalt, Brussels 1974.
19. *Gmelin*, vol. 58.
20. B. F. G. Johnson, R. E. Benfield in G. L. Geoffroy (ed.): *Topics in Inorganic Stereochemistry*, Wiley, New York 1981, pp. 253 – 335.
21. J. D. Donaldson, S. J. Clark, S. M. Grimes: *Cobalt in Medicine*, Agriculture and the Environment, The Cobalt Development Institute, Guildford 1986.
22. J. D. Donaldson, S. J. Clark, S. M. Grimes: *Cobalt in Batteries*, The Cobalt Development Institute, Guildford 2000.
23. J. D. Donaldson, S. J. Clark, S. M. Grimes, Z. Hussain, *Cobalt News* **1999**, no. 4, 11.
24. J. D. Donaldson, S. J. Clark, S. M. Grimes: *Cobalt in Catalysts*, The Cobalt Development Institute, Guildford 1990.
25. C. Tombu, *Cobalt* **1963,** no. 20, 103 – 110; no. 21, 185 – 189.
26. M. Pinta: *Modern Methods for Trace Element Analysis*, Publ. Ann. Arbor Science, Ann Arbor, Michigan, 1978.
27. "Cobalt Statistics," in *Cobalt News* **1999**, nos. 1, 2.
28. *Cobalt and the Environment*, The Cobalt Development Institute, Guildford 1994.
29. G. Schrauzer: "Cobalt", in E. Merian, M. Anke, M. Ihnat, M. Stoeppler (eds.): *Elements and Their Compounds in the Environment*, Wiley-VCH, Weinheim 2004, pp. 825 – 839.

30. Agency for Toxic Substances and Disease Registry: *Toxicological Profile for Cobalt*, U.S. Department of Health and Human Services, Washington 2004. (www.atsdr.cdc.gov.).
31. IARC: "Cobalt and Cobalt Compounds", in IARC (ed.) *Monographs on the Carcinogenic Risk to Humans*, vol. 52 (*Chlorinated Drinking Water; Chlorination By-products; Some other Halogenated Compounds; Cobalt and Cobalt Compounds*), International Agency for Research on Cancer, Geneva 1991, pp. 363 – 472.
32. M. Anke: "Essential and Toxic Effects of Macro-, Trace and Ultratrace Elements in the Nutrition of Animals" in E. Merian, M. Anke, M. Ihnat, M. Stoeppler (eds.): *Elements and Their Compounds in the Environment*, Wiley-VCH, Weinheim 2004, pp. 305 – 341.
33. Deutsche Forschungsgemeinschaft: "Hartmetall" in Deutsche Forschungsgemeinschaft (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*, 39. Lieferung, Wiley-VCH, Weinheim 2004.
34. Deutsche Forschungsgemeinschaft: "Cobalt und Cobaltverbindungen" in Deutsche Forschungsgemeinschaft(ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*, 33. Lieferung, Wiley-VCH, Weinheim 2001.
35. NTP (National Toxicology Program, USA): "Toxicology and Carcinogenesis Studies of Cobalt Sulfate Heptahydrate in F344/ Rats and B6C3F1 Mice (Inhalation Studies)", *NTP Technical Report* 471, Research Triangle Park 1998.

Coffee

Rinantonio Viani, Association Scientifique Internationale du Café (ASIC), 1802 Corseaux, Switzerland
Marino Petracco, illycaffé S.p.A, Trieste, Italy

1. Introduction 1
2. History 2
3. Green (Raw) Coffee 3
3.1. Botany and Culture 3
3.2. Harvesting and Processing 4
3.2.1. Arabica 4
3.2.2. Robusta 6
3.2.3. Liberica and Excelsa 6
3.2.4. Processing at the Mill 6
3.3. Marketing 7
4. Decaffeination 10
4.1. Low-Moisture Processes 10
4.2. High-Moisture Processes 10
4.3. Processes Using Supercritical Media 11
4.4. Coffee Treatments Other than Decaffeination 12
5. Roasting 12
5.1. Roasters 13
5.2. Grinding 14
5.3. Packaging 14
6. Brewed and Instant (Soluble) Coffee 14
6.1. Brewing 14
6.2. Instant Coffee Manufacture 15
6.2.1. Extraction 15
6.2.2. Concentration 15
6.2.3. Spray-Drying 16
6.2.4. Freeze-Drying 16
6.2.5. Aromatization 17
7. Composition 17
7.1. Carbohydrates 17
7.2. Acids 19
7.3. Lipids 21
7.4. Nitrogen Compounds 22
7.4.1. Amino Acids, Peptides, and Proteins 22
7.4.2. Purines and Nucleic Acids 22
7.4.3. Other Bases 23
7.5. Minerals 23
7.6. Products Formed during Roasting . 23
7.6.1. Volatile Aroma 24
7.6.2. Brown Polymers 24
8. Consumption and Physiology 25
8.1. Consumption 25
8.2. Dietary Value 26
8.3. Physiological Value 28
9. References 29

1. Introduction [1 – 3]

Commercial coffee, the dried seed from two species of the genus *Coffea, C. arabica* and *C. robusta,* is grown as a short tree in the tropics at an altitude ranging from sea level to more than 2000 m. The ripe cherries (sometimes incorrectly called "berries"), which usually contain two seeds – the coffee beans – are harvested and processed in one of two ways.

In the *dry process*, the cherries are dried in the sun for a period of up to three weeks, and the outer and inner shells are mechanically removed to liberate the coffee beans ("green coffee"), which can be cleaned and bagged. This method is usually applied to Arabicas from Brazil, the main coffee-producing country, and Ethiopia, as well as to all Robustas, of which Indonesia and Vietnam are the main producers.

In the *wet process*, the outer skin of the cherry is squeezed through slots and the pulp surrounding the beans is fermented and washed away. The beans, still covered by the inner shell, the parchment, are dried either for a period of up to 1 week in the sun or for a shorter period in a dryer. After mechanical removal of the parchment, the beans are cleaned and bagged. The main producer of wet-processed Arabicas, sold as "mild coffees" or "milds", is Colombia, followed by the Central American and Eastern African countries.

Europe, North America, and Japan are the main consumers. Of the producing countries, Brazil is the main consumer. For consumption, coffee is roasted (and ground) for sale and consumed as a brew, or further treated by extraction, concentration, and spray- or freeze-drying and sold as soluble (instant) coffee.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_315.pub2

Caffeine, the active principle of coffee, can be removed from the green bean after soaking in water, by dissolving it in a solvent, such as dichloromethane, ethyl acetate, or supercritical carbon dioxide. Adsorption on activated carbon is also employed. In addition to caffeine, ca. 1 % in Arabicas and 2 % in Robustas, the dry bean contains ca. 50 % carbohydrates, 8–17 % lipids, up to 11 % proteins and amino acids, and ca. 5 % minerals. During roasting, these components form the typical coffee flavor, of which more than 1000 components, many of them fragrant or sapid, have been identified.

2. History [4]

Unlike most cultivated plants whose primitive strains have long ago disappeared by genetic erosion, wild coffee plants, related to the actual cultivars of *C. arabica*, still exist in East Africa in the region where the main species originated. The plant spread from the rainy forests around Lake Victoria to the Kaffa province in the Ethiopian highlands, north to the arid Sudan, where its sweet red cherries, *bunna*, were picked and chewed. Dried cherries are easily stored and transported; thus, their use reached the coastal area, where they are known as *m'buni* or *buni*, and are still consumed sweetened and unroasted on special occasions as a ceremonial food.

Intense commercial and political exchanges across the Red Sea introduced coffee use to Arabia, and its culture may have started as early as the sixth century in the fertile valleys of the country now known as Yemen.

The advent of Mohammedanism, forbidding the use of alcoholic beverages and encouraging exchanges through pilgrimages to Mecca, together with the flourishing of the Arabian and Persian civilizations, asserted the strength of coffee around the harbor of Mocha and its extension eastward. The Persian scholars Rhazes (850–922) and Avicenna (980–1037) mention the digestive properties of coffee, probably identified with *bunchum* by the former and called *bon* or *ban* by the latter.

According to Abdelkadir, who in 1587 wrote the first historical account dealing specifically with coffee, the expansion northward appears to have been slower. It was only by the second half of the 15th century that coffee, which was then being roasted and consumed as a brew, *kahwa*, had conquered the whole Arabian peninsula and reached the Mediterranean Sea. During the 16th century, coffee became very popular throughout the Ottoman Empire.

The stimulating properties of coffee are described by the legends reported in Abdelkadir's manuscript, where strength to survive in the desert or to stay awake and pray is restored to holy people by chewing the sweet berries or by drinking a beverage prepared by roasting the berries to make them softer and to reduce their bitterness. Leonhard Rauwolf (ca. 1540–1596), physician and botanist from Augsburg, was one of the first Europeans to describe coffee as a drink "almost as black as ink" and to give a vivid description of Aleppo's coffeehouses. Prospero Alpini (1553–1617), physician and professor of botany at the University of Padua, drew the first picture of a coffee branch from a tree seen in the garden of the Bey in Cairo. By way of Venice, coffee from Mocha conquered Europe in the 17th century. Coffee culture was introduced to the Malabar coast of India and to Ceylon toward the turn of the 16th century. By the end of the 17th century the Dutch had succeeded in cultivating coffee on a commercial basis in Java; they broke the Mocha monopoly in 1711, when the first coffee produced in the East Indies was auctioned in Amsterdam. Seedlings from a plant brought to Amsterdam in 1706 were distributed to the European botanical gardens. Due to a state of war, the French waited until 1713, when a plant 5 ft (ca. 2 m) tall was sent to Louis XIV, the Sun King. It was planted in the Jardin des Plantes in Paris and described by Antoine de Jussieu. By royal bequest it became the ancestor, with related plants from Amsterdam, of most of the coffee trees planted during the 18th century in the Western hemisphere, which eventually led to the decay of Dutch supremacy in the trading of coffee.

The Swedish botanist Linnaeus, a friend of the Jussieu family, classified the plant as *Coffea arabica* in 1753. Caffeine [*58-08-2*], the active ingredient of coffee, was isolated by Runge in 1820 from a few beans received from Goethe. The formula of caffeine, which had been determined from combustion data by Pfaff and Liebig in 1832, was established by Medicus in 1875 as 1,3,7-trimethylxanthine, $C_8H_{10}N_4O_2$,

M_r 194.19, and confirmed by E. Fischer in 1897 by total synthesis.

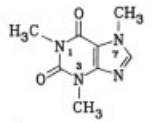

Caffeine (1,3,7-trimethylxanthine)

The year 1869 saw the destruction of coffee culture in Ceylon (Sri Lanka) by the fungus *Hemileia vastatrix* (coffee rust) and the spread of infection to the low-altitude plantations in India, Java, and Sumatra. The Dutch botanist Pierre studied the sources of coffee and in 1895 found a rust-resistant species in Congo, *Coffea canephora,* var. *robusta*, now the second most economically important type of coffee. Roselius, a Bremen coffee merchant, devised the technique of steaming coffee beans to enlarge their pores, then washing out the caffeine with a solvent. A patent was issued in 1906 for the decaffeination of green coffee beans by a technique very similar to those applied currently.

From the beginning of the 20th century on, coffee cooking ceased to be considered a plain kitchen task and stimulated the ingenuity of inventors that produced a range of increasingly sophisticated and complex brewing devices. From the first steam kettle patented in 1901 to the lever-operated and the electrical pump-driven machines, this effort culminated with the present-day computer-controlled "espresso" machines.

For the sake of convenience, instantly soluble coffee has gained nowadays a wide diffusion. A patent for making instant coffee had already been granted to Gale in 1865, and commercial soluble coffee made its faint appearance at the beginning of the 20th century. It was not until 1938 that Morgenthaler succeeded in obtaining the first 19 t of a free-flowing, stable, and readily soluble coffee extract, which the United States army found to be very convenient during World War II. Pure coffee extract was introduced in 1950, and in 1964 freeze-dried granules appeared on the market, having a flavor that approached that of brew. The introduction of vacuum packing and the developments of home brewers have, in the 1980s and 1990s, brought roast and ground coffee to a level of convenience approaching that of soluble coffee.

3. Green (Raw) Coffee

3.1. Botany and Culture [5 – 7]

The genus *Coffea* L. (Rubiaceae) consists of about 100 species classed as *eucoffea*. The subgenera *erythrocoffea*, which contains all of the cultivated species, and *argocoffea* are both native to continental Africa. *Mascarocoffea* from Madagascar and *paracoffea* from the Indian subcontinent are low in caffeine. Two species only are economically important, both cultivated around the world in the tropical belt, *Coffea arabica* L., Arabica coffee, with two main varieties, *bourbon* and *typica*, accounting for two-thirds of world production, and *Coffea canephora* Pierre, one main variety *robusta*, one-third of world production (Table 1). Two other species, *C. liberica* Bull ex Hiern, Liberica coffee, and *C. dewevrey* De Wild., Excelsa coffee, have lost all commercial importance. Arabica cultivars have been developed for visual characteristics, such as "maragogype", a mutant from *typica* producing giant beans; productivity, such as "Mundo Novo" in Brazil; or resistance to diseases, like "Ruiru eleven" in Kenya, resistant to the coffee berry disease, but which has lost some of the typical flavor of Kenyan coffee.

The genetic variability of Arabica coffee has remained until quite recently extremely limited; as a consequence, susceptibility to two fungal diseases, coffee leaf rust (CLR), caused by *Hemileia vastatrix*, now encountered worldwide at low altitudes, and coffee berry disease (CBD), produced by *Colletotrichum kahawae*, previously known as *C. caffeanum*, which attacks the green and ripening fruits in cool areas at high altitudes in Africa, has increased. Breeding programs aiming at resistance to diseases and pests, increased yield, special shapes and size of plant, and improvement of cup quality have produced new commercial varieties. Natural and bred botanical interspecific hybrids, such as "Hibrido de Timor", a spontaneous hybrid of *C. arabica* and *C. canephora*, are used to create progenies that are resistant to rust like Robusta but have Arabica flavor characteristics. This is accomplished starting from traditional varieties with a small genetic base (Caturra, Catuai), and introducing genetic variability from

spontaneous trees from Ethiopia or Sudan. Commercially important ones are Catimor: a hybrid created by backcross selections of a hybrid between "Hibrido de Timor" and "Caturra vermelho" (a natural mutant of *bourbon* found in Brazil), and Icatu: an interspecific hybrid between a colchicine-induced tetraploid Robusta plant and the Arabica cultivar "Mundo Novo". A similar progeny has also been developed in Colombia as "Colombia". The selection cycle can now be reduced from 30 years down to less than 10 years by multiplying in-vitro selected F1 hybrids [8]. An interesting low-caffeine somaclonal variation, obtained from *Arabica* x *racemosa* hybrids, "Aramosa", has been found highly resistant to drought and early ripening and has good cup characteristics [9]. *Coffea canephora* has a very high genetic variability and is usually propagated by cuttings in nurseries.

The best soil for planting is friable, porous, well drained, rich in organic matter, and slightly acidic. The soil must be deep for Arabica. The plants are cultivated on sheltered slopes arranged in squares or following the contour of the land protected from wind and frost also by intercropping at early stage with various suitable crops, and by rows of trees. For Arabica, the spacing between each plant can even be less than 1×1 m, with a density of 10 000 or more trees per hectare, depending on variety, climate, and the choice of either single-stem or multiple-stem cut. The best spacing for Robusta is 3×2 m.

Various pests attack coffee. Beside CLR and CBD, other harmful mycetes that deserve mentioning are *Fusarium* (*Gibberella*) *stilboides*, fusarium bark disease, and *Fusarium xylarioides*, tracheomycosis or fusarium vascular wilt, the most serious disease affecting Robusta, also affecting Arabica. Nematodes can invade the roots (e.g. *Meloidogyne coffeicola*). Among the insects that attack coffee the most damaging are *Hypothenemus hampei*, "broca", coffee berry borer (CBB) and *Perileucoptera coffeella*, "bicho mineiro". Now that rust-resistant varieties have been introduced, CBB has become the most important pest, present almost in all producing countries, causing a premature fall of berries and damaging the seeds within the berries.

Table 1. Commercial coffees

	***Coffea arabica* L. – Arabica coffee**	***Coffea canephora* Pierre – Robusta coffee**
First described by, date	LINNAEUS, 1753	PIERRE, 1895
Botanical varieties	*typica*, *bourbon*	*robusta*
Indigenous to	Eastern highland	Eastern central rainforest
Optimal growth	temperate equable climate	warm humid climate
Altitude	700–2000 m	0–800 m
Average yearly temperature	15–23 °C	18–27 °C
Maximum temperature	25 °C	30 °C
Annual rainfall	1500–2200 mm/m^2	2200–3000 mm/m^2
Plant	self-fertile shrub or small tree bearing after 3–4 years	self-sterile shrub or small tree bearing after 2–3 years
Chromosomes (2n)	44	22
Root system	deep	shallow
Leaf	small, glossy, oval, (12–15) × 6 cm	large, broad, corrugated, (15–30) × (5–15) cm
Flower	after rain, small white to cream	irregularly, large, white, sometimes pink
Fruit, a drupe (cherry or berry; it contains two seeds surrounded by parchment with the flat grooved faces against each other)	oblong ellipsoid, 15 mm long, several on each leaf axil in loose clusters, crimson or yellow, 8–9 months to ripen	ellipsoid, striated when dry, 12 mm long, in tight axillary clusters, crimson then black, 9–10 months to ripen, dries on branch
Seed (bean)	oval, flat, deeply grooved, 5–13 mm long	oval to round, grooved, 4–8 mm long
Trade name	Arabica	Robusta
Relative importance (2006)	66 %	34 %
Cup flavor of roasted and brewed beans	mild, aromatic, acid	woody, earthy
Caffeine content of bean (dry basis)	0.6–1.4 %, average 1.2 %	1.7–4.0 %, average 2.2 %

3.2. Harvesting and Processing [10, 11]

A transverse section of a coffee cherry is shown in Figure 1.

3.2.1. Arabica

Arabica coffee is harvested by selectively hand picking the red ripe cherries every 1–2 weeks. In some countries near the tropics, where seasons are recognizable, ripening occurs in a rather uniform way. In such cases, when not enough la-

bor is available at harvest time, immature and ripe cherries are stripped off the branches and collected from the ground along with the overripe and dry ones. The leaves and twigs are winnowed away in the field. Often a water siphon, in which the heavier cherries sink, while leaves, twigs, and dry cherries float and are diverted, is used to carry out an initial grading of the cherries. The middle fraction containing the ripe cherries is further treated by either the dry process, in which the whole cherry is dried, or the wet process, in which the pulp surrounding the bean is fermented and washed away before the beans are dried in their parchment [12]. Raw beans recently processed in either of these ways are recognized by their color: dry-processed (or natural) beans are yellow-green to pale brown, whereas wet-processed (or washed) beans are green to blue-green.

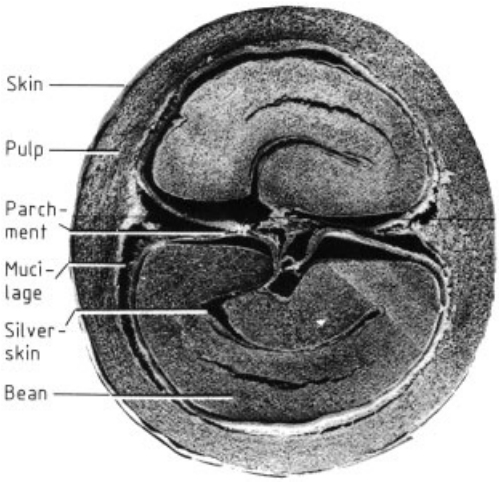

Figure 1. Transverse section of a coffee cherry containing two seeds (courtesy Dr. E. Dentan)
Pericarp: "Skin" or exocarp, red (or yellow); "Pulp" or mesocarp, yellow-white; "Parchment" or endocarp, pale yellow
"Silverskin" or spermoderm, silvery
"Bean" or endosperm, yellow-green when dry
The site occupied by the "mucilage," yellow-white, appears black in the figure, having been dissolved away during the preparation of the sample.
A coffee cherry may contain one, two, or rarely, three seeds.

The *dry process* is applied where water is scarce (as in Ethiopia and Brazil) and to the off-grade cherries discarded during wet processing. Freshly harvested cherries are spread out to dry in the sun. They are placed in layers up to 6 cm thick, turned regularly, and sheltered from rain showers and dew. The coffee dries in 2–3 weeks, and the husks can then be removed mechanically from the beans.

The *wet process*, which uses water flowing down a slight gradient, was introduced by the Dutch in the West Indies during the first half of the 18th century and is used where enough water is available and the climate is humid. Harvested cherries are handled in bulk if they can be processed in less than one day; otherwise, they are stored under water to avoid overheating. Where the water supply is abundant, cherries are pregraded in a washing channel or a siphon tank. Where water is scarce, the cherries are supplied dry to the pulper. This reduces water consumption by half (60 vs. 120 L/kg). Pulping is done by squeezing the cherries through slots in drum or disk machines to separate the outer skin and some of the fleshy pulp (the pericarp) from the parchment coffee (the bean surrounded by the endocarp). Parchment coffee is then sorted by gravity, and floaters, light parchment, peaberries (oval beans from cherries containing only one seed), and heavy parchment are sent to separate tanks. The sticky pulp still adhering to the parchment is loosened by fermentation in a tank, either dry in the open for 24–48 h or under water for up to 72 h, depending on the altitude. The fine acid taste of some of the coffees grown at high altitudes is attributed to the long fermentation time. Adding pectic enzymes or alkali may accelerate fermentation. Fermentation is stopped by soaking parchment coffee in water.

Pulping and fermentation are sometimes combined in the "aquapulper," a pulper and demucilager, where the pulp is rubbed off the parchment mechanically. All pulp residues are washed away in a channel where the remaining skins and whole cherries, which are lighter, are flushed away and collected on wire trays. The cherries are dried in the sun for 5–10 d on wire strapping, which ensures thorough air circulation, or on barbecues. Where the climate is not sufficiently dry, sun drying can be partially or completely replaced by artificial drying in tray or tower air dryers for 24–36 h. Particular care must be taken to avoid overheating the coffee, which results in the subsequent development of a sour taste.

3.2.2. Robusta

Robusta cherries remain on the tree at maturity, undergoing a slight fermentation, which is interrupted when the fruit dries. The ripe cherries are selectively picked from the large shrubs by starting at the lower branches. The practice of stripping cherries off the tree at various degrees of maturity from green to dry, and drying, husking, and bagging them without sorting is becoming common.

Figure 2. Coffee beans of various types
a) roasted Arabica; b) roasted Robusta; c) raw Arabica; d) raw Robusta

3.2.3. Liberica and Excelsa

Liberica cherries are rather pulpy and difficult to dry. Excelsa cherries ripen a few at a time and then drop to the soil when dry. In both cases, uncontrolled fermentation occurs, which contributes to the unclean taste of their brew. Their importance is waning and they will not be discussed further.

Coffee beans of various types are shown in Figure 2.

3.2.4. Processing at the Mill

Dry cherry or parchment coffee is normally cured in centralized mills. Impurities are removed from coffee by sizing on oscillating screens and by density differences in a pneumatic "catador" or on densimetric vibrating tables. Processing rate is up to 1 t/h for cherry and 2 t/h for parchment coffee. Dehulling of the dry parchment or of the shell, which becomes brittle when it is dried, is done by impact in a centrifugal mill, usually a ribbed steel drum or a crossbar revolving in a perforated case, which also polishes the silverskin off the surface of the beans and winnows away the shells. Capacities are 0.1 t/h for hand-operated mills and up to 3 t/h for motor-operated mills. The raw coffee beans are graded by size, through revolving or flat screens with round holes, at the rate of 0.1–1.5 t/h. Peaberries, which are oval, can be separated from flat, regular beans through slot holes or by rolling them down a conveyor belt that runs upward at a slight angle. They are sold separately under the name caracoli. Grading by density is done by winnowing the beans on densimetric tables (capacity 1–4 t/h) or in catadors (capacity 0.8–1.5 t/h), where sand and stones are again removed and the product is separated into healthy beans of full weight, medium beans, lighter inferior beans (empty, insect damaged, etc.), "triage" (material rejected from superior grades), and "brisures" (bits).

Grading by color, recommended to get rid of visible defects [13], is accomplished in several ways: by hand picking (capacity < 1 kg/h per person); electronically with monochromatic equipment, used with Robustas to sort out black beans; bichromatically, by which both brown

and bleached beans are eliminated in Arabicas; or fluorimetrically at 360 nm, to eliminate "stinker" beans, which have an impure taste, in freshly cured washed Arabicas. The beans are fed into chutes on counterrotating rollers or on belts and viewed from different angles. The rejected beans are driven out by compressed air (capacity up to 150 kg/h per channel).

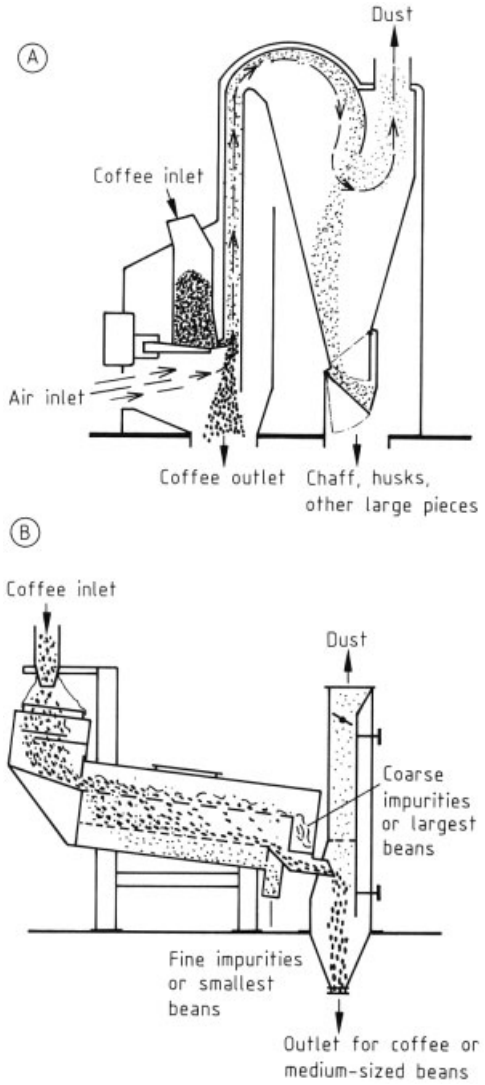

Figure 3. Equipment used in coffee processing (courtesy of Bühler, Uzwil, Switzerland)
A) Pneumatic aspirator, the catador, which classifies according to density; B) Vibrating separator, which classifies by size

Grading by cupping is sometimes performed in the producing countries, but more often it is done in the consuming countries. Representative samples of coffee are examined for defective beans and extraneous matter, roasted and noted for uniformity of roasting, ground and brewed in porcelain equipment, tested by smelling, and tasted by swiftly drawing a spoonful against the palate [14]. Washed Arabicas are graded according to acidity, body, and flavor. Dry-processed Brazilian coffees are subdivided in order of decreasing quality as strictly soft, soft, softish, hardish, hard, and Rioy (a typical medicinal off-taste). The best Robustas are neutral and clean, while defective ones are harsh, rubbery, or earthy.

Placing 60 or 69 kg of coffee (70 kg in Colombia) into jute bags marked with the name of the producer and the year of harvest readies the coffee for shipping. The shipping of Robusta coffee in bulk containers is now practiced.

Equipment used for separating impurities and for grading green coffee beans is shown in Figure 3.

3.3. Marketing [15]

Today one of the most important food commodities, the world green coffee market has been characterized in the past by long periods of oversupply alternating with brief periods of shortage, linked to the size of the crop. A series of short-term agreements, concluded by producing countries in an effort to avoid wild fluctuations in prices, finally led to the first *International Coffee Agreement* (ICA), signed in 1962 by both producing and consuming countries at the United Nations. The agreement, which is administered by the *International Coffee Organization* (ICO) with a head office in London, has been reviewed several times. The present (in 2006) agreement entered into force in 2001 with the signature of 44 exporting (> 99 % of all coffee produced) and 30 importing members of the ICO. All major consumer countries are represented, including the United States as a new entry. The objectives of the new ICA, which have switched in the 1990s from regulation of supply and demand through export quotas to assistance to international trade, are:

- Encouraging members to develop a sustainable coffee economy
- Promoting coffee consumption
- Promoting quality of coffee
- Providing a forum for the private sector for representatives of the private sector (traders, roasters, etc.)
- Promoting training and information programs designed to assist the transfer of technology relevant to member countries
- Analyzing and advising on the preparation of projects to the benefit of the world coffee economy

Table 2 [16] lists the total coffee production for the crop years 1993 to 1998 in thousands of 60-kg bags – the traditional trade unit – of all ICO exporting member countries.

Marketing systems have involved different types of semigovernmental agencies in the producing countries. Liberalization of trade, particularly noticeable in the 1990s, has led to price transparency and eliminated the monopoly of Marketing Boards in the sale of coffee.

In the consuming countries, coffee is traded in the following ways:

1) The "forwards market" applies to coffee before it leaves the producing country. This is usually only available to large traders and roasters who need to ensure large quantities (thousands of bags) of well-defined grades and origins throughout the year.
2) The "spot market" covers coffee once it has actually landed in an importing country. Practically all origins are available at one of the main importing ports. This is often the only way of purchasing available to small roasters who buy a few bags at a time and wish to avoid price risks.
3) The exchange or terminal market covers coffee scheduled for delivery at a set date and price in the future. Futures contracts are taken either to avoid (hedging) or to exploit (speculating) price fluctuations. The main markets are the New York Coffee, Sugar, and Cocoa Exchange (NYCSCE), where lots of 37 500 lb (ca. 250 bags) of "milds" are traded under the "C" contract, with prices in US cents per pound ex-dock, and the London International Financial Futures and Optional Exchange (LIFFE), which deals in Robustas, with 5-t lots and prices in US$ per ton, CIF (Cost Insurance Freight) selected European ports.

Three consumer issues have acquired some importance after the 1990s in the coffee trade.

Fair Trade coffee is a commodity produced in many less developed countries, sometimes in conditions and of a quality below current food standards, further exacerbated after liberalization of trade in the late 1980s and early 1990s, when world coffee prices reached a very low level, rendering the living conditions of small coffee farmers precarious. As a reaction, a new way of trading, namely "Fair Trade", was introduced in 1988 in the Netherlands to improve the working and living conditions of small farmers by direct trade, a guaranteed minimum price, prefinancing, and long-term contracts, and to reduce their dependence on the local wholesaler. Direct trade has been extended in the 1990s to other commodities and many importing countries. Although reaching only a small fraction of the coffee traded, small-scale direct procurement systems, now benefiting from the financial backing of governmental and private funds, have helped in drawing the attention of consumers in the developed countries to the question of redistribution of profits in the production areas. Major industrial users now revise their trading practices, introducing some form of direct procurement, which allows them to assist farmers to produce coffee according to their quality standards.

Organically grown coffee has picked up a few percent of trade since it appeared in the 1980s among the specialty coffees in the United States with an image of wholesomeness [17]. In the 1990s its image has moved to one of environmental protection, according to the aims of organic agriculture of working with renewable resources, maintaining genetic diversity, minimizing all sources of agricultural pollution, and considering the ecological and social impact of agricultural systems. Regulations for organic products, including coffee, which must be certified by recognized inspection agencies, have been issued [18].

Sustainability is a general approach in coffee economy that, according to a preliminary draft wording, implies conditions of production, processing, and trade for all parties involved in the supply chain that:

Table 2. Green coffee world production (thousands of 60-kg bags; source: International Coffee Organization)

Country	Species*	2000	2001	2002	2003	2004	2005	Average
TOTAL		115 064	106 645	121 913	103 849	113 631	106 386	111 248
Bolivia	(A)	173	118	149	125	161	115	140
Colombia	(A)	10532	11999	11889	11197	12043	11900	11593
Costa Rica	(A)	2253	2166	1938	1802	1726	1758	1941
Cuba	(A)	313	285	239	254	257	275	271
Dominican Republic	(A)	439	435	433	354	467	791	487
El Salvador	(A)	1707	1667	1438	1457	1447	1372	1515
Ethiopia	(A)	2768	3756	3693	3874	5000	4500	3932
Haiti	(A)	420	402	374	373	355	375	383
Honduras	(A)	2667	3036	2497	2968	2575	2990	2789
Jamaica	(A)	37	30	38	36	26	35	34
Kenya	(A)	1001	991	945	673	709	1002	887
Malawi	(A)	63	60	42	48	21	25	43
Mexico	(A)	4815	4200	4000	4550	3407	4000	4162
Nicaragua	(A)	1595	1116	1199	1405	1127	1400	1307
Panama	(A)	170	160	140	172	90	170	150
Paraguay	(A)	22	20	23	20	20	23	21
Peru	(A)	2596	2749	2900	2616	3355	2420	2773
Rwanda	(A)	273	296	319	265	450	300	317
Venezuela	(A)	956	721	869	786	701	820	809
Zambia	(A)	94	100	119	101	111	117	107
Zimbabwe	(A)	109	121	106	84	96	58	96
Brazil	(A/R) 9:1	34 100	30 837	48 617	28 820	39 272	32 944	35 765
Burundi	(A/R) 9:1	337	259	475	333	430	350	364
Guatemala	(A/R) 10:0	4940	3669	4070	3610	3703	3675	3945
Papua New Guinea	(A/R) 10:0	1041	1062	1085	1155	997	1267	1101
Tanzania	(A/R) 6:4	809	624	824	611	763	750	730
Angola	(R)	50	21	57	38	15	20	34
Benin	(R)	0	0	0	0	0	1	0
Central African Rep.	(R)	122	75	92	43	61	100	82
Congo, Rep. of	(R)	3	3	3	3	3	3	3
Côte d'Ivoire	(R)	4846	3595	3145	2689	2328	2171	3129
Gabon	(R)	0	1	1	0	0	2	1
Ghana	(R)	76	13	32	16	11	25	29
Guinea	(R)	368	254	272	407	245	310	309
Nigeria	(R)	48	43	51	45	42	45	46
Sierra Leone	(R)	53	53	32	24	5	25	32
Thailand	(R)	1692	548	758	846	797	764	901
Togo	(R)	197	113	68	144	166	168	143
Trinidad and Tobago	(R)	14	15	15	15	14	10	14
Vietnam	(R)	14 939	13 133	11 555	15 230	13 844	11 000	13 284
Cameroon	(R/A) 9:1	1113	686	801	900	727	1000	871
Congo, Dem. Rep. of	(R/A) 5:5	362	421	319	427	360	575	411
Ecuador	(R/A) 6:4	872	893	732	766	938	1125	888
India	(R/A) 6:4	4516	4970	4683	4495	3844	4630	4523
Indonesia	(R/A) 9:1	6978	6833	6785	6571	7536	7654	7060
Madagascar	(R/A) 10:0	366	147	445	434	388	425	368
Philippines	(R/A) 8:2	775	759	721	433	373	500	594
Sri Lanka	(R/A) 10:0	43	32	35	36	32	35	36
Uganda	(R/A) 7:3	3401	3158	2890	2598	2593	2366	2834
Total Arabica		**74 783**	**71 927**	**87 278**	**68 460**	**78 805**	**74 067**	**75 887**
Total Robusta		**40 281**	**34 718**	**34 636**	**35 389**	**34 826**	**32 319**	**35 361**
Ratio A/R		1.9	2.1	2.5	1.9	2.3	2.3	2.1

* A = Arabica, R = Robusta, A/R = predominantly Arabica, R/A = predominantly Robusta.

1) Provide an economic return that covers production and living costs plus a further margin for development,
2) Treat the environment responsibly so that natural resources remain available to future generations,
3) Secure social and working conditions in accordance with international standards, conducive to the maintenance of stable communities.

Table 3. Decaffeination by low-moisture processes

Operation	Medium	Conditions	
		t, °C	Time, h
Soaking	steam and water	22–100	0.25–5
Decaffeination	water-saturated solvent		
	dichloromethane [22]	60–100	2–12
	ethyl acetate [23]	80–105	3–12
	fats and oils* [24, 25]	95–105	6–9
Deodorization	steam	100–110	1–4
Drying	heat	40–80	0.5–10

* No longer used.

4. Decaffeination [19]

The stimulating effect of coffee is due to the presence of the alkaloid caffeine (structure, see 2); decaffeinated coffee gives a brew devoid of this effect. The presence of water (caffeic acid as a dope has been proposed [20]) is essential in decaffeination to soften and open the cellular structure of the beans and to ensure the diffusion of caffeine across the cell walls by solubilizing the caffeine – potassium chlorogenate complex. The ease of decaffeination increases with the moisture level in the beans, while the selectivity for caffeine extraction with respect to other water-soluble constituents decreases. Industrial processes, which are usually carried out with green beans, can be broadly divided into two main types: those performed at moisture levels below 50 % (usually between 30 and 45 %), and those performed above 50 % (usually between 55 and 65 %).

4.1. Low-Moisture Processes

Low-moisture decaffeination processes are applied to green coffee to avoid removing or altering fragile aromatic constituents formed during roasting. The beans are soaked in a mixer, column, or rotating drum and extracted with a nonpolar solvent in one or more columns or rotating drums. The residual caffeine level in the beans is lowered to less than 0.1 % in the dried decaffeinated beans by most national regulations, or to less than 3 % of the amount initially present (the commercial practice in some countries). The beans are then deodorized (solvent traces are removed by steam stripping) and dried to their initial moisture content in a tower or a fluidized-bed drier under atmospheric or reduced pressure. The solvent is recovered by batch or continuous evaporation, or by steam stripping of the caffeine under vacuum. Losses due to leaching of hydrosoluble material are low, and caffeine, containing a few percent of waxes from the surface of the beans, is recovered and purified by successive crystallizations to pharmaceutical grade for use in soft drinks and medicines [21]. Typical operating conditions are given in Table 3.

4.2. High-Moisture Processes

In high-moisture processes, green or roasted coffee beans are soaked in water until they have reached the required moisture level. Decaffeination is carried out by dissolution with an aqueous solution of 10–30 % noncaffeinic, soluble coffee solids. The resulting caffeine-enriched solution is then decaffeinated either by liquid – liquid extraction with any of the solvents used in the low-moisture processes or by adsorption of the caffeine on activated carbon or on ion-exchange resin. This solution is then used to decaffeinate new beans.

Alternatively, the swollen green beans may be decaffeinated with fresh water. After removal of the caffeine the solution is concentrated and reabsorbed on the predried decaffeinated beans from which it has been obtained, thus replacing noncaffeine solids otherwise lost. Losses of coffee solids are usually higher with adsorption processes than with solvent processes. Losses on the adsorbent may be reduced by precharging it with sucrose or by regenerating it with acetic acid or ethanol solutions. In addition, pretreatment of active carbon with formic or hydrochloric acid reduces discoloration of the beans. Table 4 gives typical operating conditions.

Table 4. Decaffeination by high-moisture processes

Operation	Medium	Conditions	
		t, °C	Time, h
Soaking	water [26]	70–100	0.2–2
Decaffeination of extract			
Liquid – liquid extraction	dichloromethane [27]	80–85	10
Deodorization	steam	100–110	2
Adsorption	Amberlite [28] [a]	ambient	
	Duolite [29] [b]	95	
	acid-activated carbon [30 – 32]	60–100	5–8
Decaffeination of beans			
Same beans in a loop		60–100	5–8
New beans counter-current		70–90	8
Drying	heat	40–80	0.5–10

[a] Amberlite XAD-2, 4, 7 or 8 (Rohm & Haas); not used.
[b] Duolite S561 (Diamond Shamrock); not used.

Table 5. Decaffeination using supercritical media

Medium	Conditions			
	t, °C	*p*, MPa	Time, h	Reference
Carbon dioxide	40–80	20–30	5–30	[33]
Nitrous oxide *	75	20–38	4	[34]
Propane *	97–120	10–15	6	[35]

* Not used.

4.3. Processes Using Supercritical Media

Supercritical carbon dioxide, i.e., at temperatures and pressures above the critical point of 31.06 °C and 7.38 MPa (73.8 bar), can be used to decaffeinate coffee by a process called destraction by its inventor K. Zosel. The supercritical medium in a pressure vessel is circulated through moist coffee, where it becomes charged with caffeine. It is then passed through a second pressurized vessel containing an adsorbing medium such as activated carbon, resin, or water, which retains the caffeine. The solubility of caffeine increases with the density of supercritical carbon dioxide, thus increasing with pressure at constant temperature and decreasing with temperature increases at constant pressure. This property can be used in the separation step. Besides carbon dioxide, which is employed in at least one industrial decaffeination plant, other gases have also been proposed for decaffeination (Table 5).

Both carbon dioxide and propane can also dissolve caffeine in the liquid state below the critical temperature. The pressure – temperature diagram of carbon dioxide is given in Figure 4, and a scheme of decaffeination using supercritical carbon dioxide is shown in Figure 5.

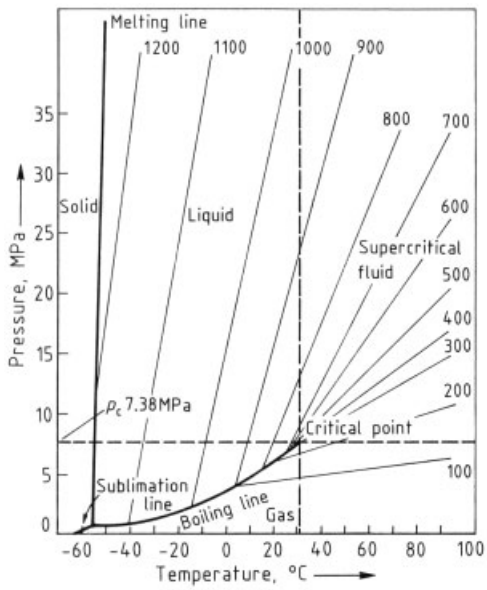

Figure 4. Pressure – temperature diagram of carbon dioxide
Densities from 100 to 1200 g/L are given as the third dimension [36].

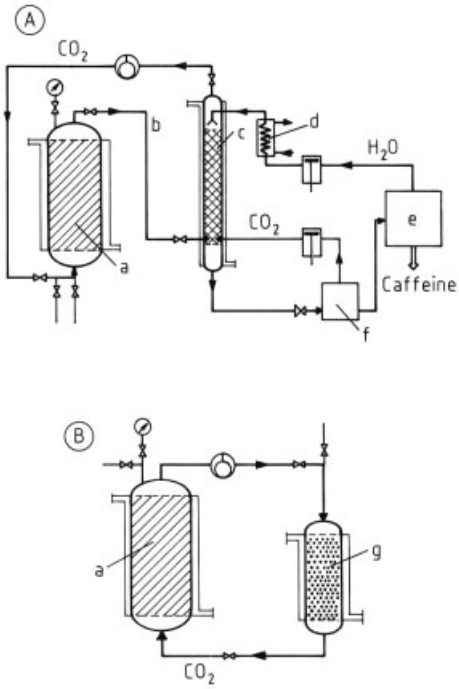

Figure 5. Decaffeination with supercritical media [36]
A) Caffeine is washed off the fluid with water in a washing tower and recuperated; B) Caffeine is adsorbed on a column containing activated carbon or resin, and the fluid is continuously recycled
a) Moist green beans; b) CO_2 + caffeine; c) washing tower; d) heat exchanger; e) distillation chamber; f) degassing chamber; g) active carbon or adsorbing resin column

4.4. Coffee Treatments Other than Decaffeination

Coffees processed by methods other than decaffeination are marketed in Germany. Green coffee beans are either treated for several hours with supersaturated steam (the Lendrich process) or washed with dichloromethane. Roast coffee beans can be washed with liquid carbon dioxide (the Bach process). All of these processes are claimed to make coffee more easily digested by removing irritating substances formed at roasting or their precursors in the green bean. Clinical and epidemiological research does not show a direct relationship between the removal of specific constituents other than caffeine (and cafestol/kahweol) and the physiological effects of drinking coffee.

5. Roasting [37, 38]

The stone-hard blue- to yellow-green raw coffee bean of density 550–700 g/L has a sweet astringent taste and a weak "green" aroma. During the roasting process, which occurs at ca. 200 °C, the coffee bean develops a brittle structure, a dark brown color, and a characteristic flavor and aroma [39]. At the beginning of roasting, for temperature increases up to 120–150 °C, moisture is released while the beans contract slightly and become paler. Browning reactions start, and the beans begin to swell and release some gas, mostly carbon dioxide, with moisture and some carbon monoxide. The reaction becomes exothermic, and above 180 °C, the main aromatic constituents form. Pressure buildup in the beans is said to reach 1.5–2.5 MPa, and, when the bean has almost doubled its volume [40], the cellular structure ruptures [41]. The final temperature reached by the bean is 190–240 °C, with roasting losses ranging from 13 to 15 % for a light-roasted coffee, consumed in northern Europe, and up to 20–22 % for a dark roast, typical of southern Europe.

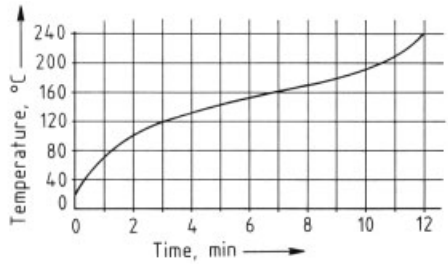

Parameter*													
Roasting losses, %	0	0.7	1.3	2.3	3.1	3.8	5.5	6.8	7.8	9.2	10.8	12.1	16.2
Moisture, %	10	8.8	8.7	8.4	8.0	7.6	6.1	4.9	4.3	3.6	2.6	1.5	1.1
T.O.C.**, $\frac{mg}{100g}$							0.05	2.00	2.35	2.90	3.80	4.40	4.95
Apparent bulk density, g/cm^3	0.71	0.70	0.69	0.68	0.65	0.63	0.61	0.60	0.57	0.56	0.47	0.40	0.34

*Average of figures obtained with four different washed Arabica samples. **Total organic carbon is quantitatively linked to the volatile aroma formed during roasting.

Figure 6. Changes occurring in coffee beans during roasting
(courtesy Dr. G. Pictet)

Figure 6 shows the temperature profile, roasting losses, moisture content, apparent bulk density, and volatile aroma released by Arabica beans with an initial moisture content of 10 %, heated up to 240 °C in 12 min. Similar graphs are obtained with roasting cycles of 6–18 min by adjusting the ratio of beans to hot air. With faster roasting cycles, down to less than 2 min, uniformity of roasting throughout the bean is reduced. For a given roasting loss, the apparent bulk density, quantity of aroma released, and extractable solids at brewing increase. Robustas need longer roasting times to attain the same roasting losses as Arabicas. On subsequent cooling, which may be accelerated by quenching the roasted beans with water, residual carbon dioxide is still present in the beans and is slowly released over a period of days.

5.1. Roasters [42]

Green coffee is delivered either in palletized bags or in bulk. After the coffee is cleaned, it is stored in silos and then automatically conveyed for volumetric or gravimetric dosage and roasting. Roasters can be subdivided into conduction roasters, with direct flame heating, and convection roasters, employing preheated gas either in a laminar or a turbulent flow (fluidization). There are many roaster models, operating both batchwise or continuously, with capacities ranging from a few kilograms to more than 5 t/h. These range from manual to fully automatic, some having the capability of programming the roasting cycle. They are operated with either gas or fuel oil, consuming energy at the rate of 1–2 GJ per ton of green coffee. Models with gas recirculation and catalyzers can reduce gas emission below 50 mg/m^3. Odors and emissions are suppressed by a water spray lasting 10–15 s and by passing steam through the catalyzer [43]. Regular automatic cleaning cycles in some models avoid the buildup of tars. In the newer convection models, roasted coffee with a density of 300–450 g/L can be prepared by modulation of the temperature and the ratio between the hot gas and beans. That is, a roasting temperature ca. 300 °C, short exposure time (2–3 min), and high air : bean ratios produce low-density roasted coffees; this so-called high-extraction or fast-roasted coffee can yield up to 20 % more extractable matter when brewed and a more pronounced taste. Less spectacular results can also be attained by using traditional conduction – convection roasters at high final temperature conditions and low gas volumes [44 – 46]. Figures 7, 8, and 9 show examples of batch and continuous industrial roasters (courtesy of Probat, Emmerich, Germany).

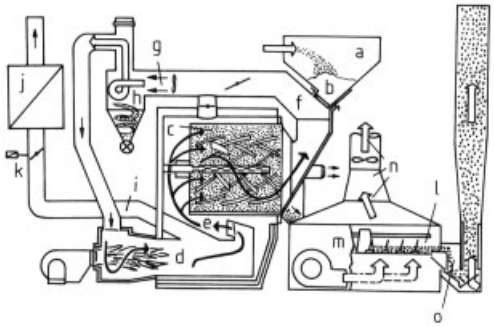

Figure 7. Conduction – convection drum batch roaster; roasting cycle 8–12 min and hourly capacity up to 3 t, a versatile model
a, b) Raw coffee supply; c) Roasting drum with mixing vanes and central quenching system; d, e) Burner with roasting gas at 550 °C; f – k) Recirculation with cyclone for chaff removal; exhaust fumes are vented through a catalyzer; l – n) Air flow to cool the roasted coffee, which is fluidized through a destoner; o) output

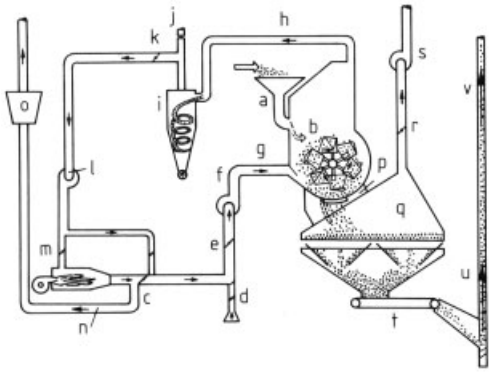

Figure 8. Convection tangential batch roaster; roasting cycle 1–5.5 min and hourly capacity of 750 kg, a high-yield model
a) Raw coffee supply; b) Tangential roaster; c – g) Burner with high-velocity roasting air mixed for the desired temperature; h – o) Recirculation with cyclone for chaff removal; exhaust fumes are vented through a catalyzer; p – u) Air flow to cool the roasted coffee, which is fluidized through a destoner; v) output

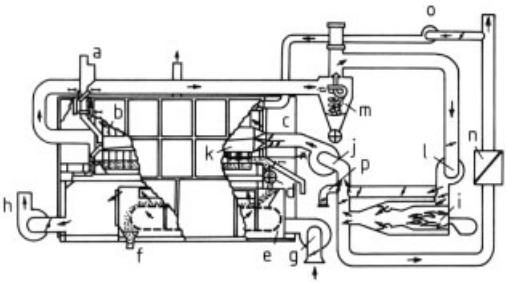

Figure 9. Convection continuous roaster; roasting time up to 6 min and hourly capacity up to 4 t, suitable for regular production in instant coffee plants
a) Raw coffee inlet; b) Roasting drum with small batch compartments; c – h) Air cooling of roasted coffee; i – p) Burner with high-velocity roasting air, which can be adjusted in one or more heating stages between 240 and 350 °C; recirculation with cyclone for chaff removal; exhaust fumes vented through a catalyzer

5.2. Grinding [47]

Grinding of the brittle beans is easily done in mills that vary in capacity from a few grams for household equipment to 4 t/h for industrial roll mills. The average particle size of the ground coffee depends on the extraction equipment to be used. Characteristic particle sizes in millimeters for various brewing methods follow:

Industrial extracts	1.0 to bean halves
U.S. drip, percolator	0.7–1.0
Filter	0.4–0.6
Espresso	0.3–0.4
Middle East	< 0.1

5.3. Packaging [48]

Unlike raw beans, roasted coffee spoils relatively quickly if unprotected from oxygen and moisture. At ambient temperature, some detrimental changes in brew flavor occur.

	Beans	Ground
Fresh roast	6 d	48 h
Weaker aroma	10–12 d	4–6 d
Stale	4–6 weeks	2 weeks

The presence of residual carbon dioxide of ca. 4 cm^3/g, which corresponds to an overpressure of 0.5 MPa [49] in the beans, precludes packing of freshly roasted coffee in airtight packages. Coffee is either prepared for market as beans in non-airtight packs or allowed to degas for up to 48 h and then filled under vacuum in metal cans or in plastic containers impermeable to air and to carbon dioxide. Coffee is more often prepared for market preground and packed in brick packs or cans after degassing or under an initial slight vacuum in flexible bags with a one-way degassing valve. In this case, 2–4 h of degassing is sufficient [50]. In pressure-tight cans, optimal aroma is reached after a few weeks if packeged in a protective atmosphere of inert gas.

6. Brewed and Instant (Soluble) Coffee

Roasted coffee is either sold as whole beans or already ground for brewing and immediate consumption, or further processed at the coffee plant where, after extraction, the soluble solids are concentrated and dried to produce instant coffee.

6.1. Brewing [51]

In spite of the apparent simplicity of coffee brewing–just a pinch of grounds dispersed in hot water–a bewildering array of domestic and professional brewing tools have made their appearance, giving origin to different consumption styles and cultures. The traditional brewing technique, still applied in several parts of the world, consists of boiling coarsely ground coffee (in the Nordic countries) or very finely ground coffee (in the Middle East) in water, infusing it for several minutes (up to 30), and serving it without separation of the grounds.

The main problem being the separation of the unpleasant spent grounds from the beverage, various ingenious filtering devices–sometimes lined with filter paper or cloth–evolved, fitting to plain or reversible pots. For example, in a typical household percolator, hot water is continuously recirculated through a bed of coarsely ground coffee (average particle diameter 0.7–1.0 mm). The extract is then filtered when it has reached the desired concentration of soluble solids. In more recent automatic drip coffee machines, soluble solids can be leached from a bed of ground coffee (average diameter 0.4–0.6 mm) kept on a

filter at slightly above atmospheric pressure for up to 15 min with or without recirculation.

The request for faster preparation of a cup arose around the beginning of the 20th century, when the first so-called "espresso" machine was built by Bezzera. Speed of percolation was increased at that time by steam-generated water pressure. By use of overpressure, as in household and food service espresso machines, the water passes through a bed of coffee (average diameter 0.3–0.4 mm) at a temperature depending on the desired pressure. A nimbler control of brewing conditions was made possible using compressed air boosters (as in a patent by Illy [52]), or lever-operated pistons or electrically-driven pumps. In all of these instances, production of water pressure is disconnected from the temperature, which remains slightly below its boiling point. The yield of soluble solids can attain 20–30 %, depending on the grind and the technique employed [53].

6.2. Instant Coffee Manufacture [54, 56]

6.2.1. Extraction

In an instant coffee plant, coarsely ground coffee (average diameter 1.0 mm to bean halves) is extracted with softened water in a battery of five to eight percolation columns, with cell capacity from a few kilograms up to 1 t. The process operates semicontinuously, with water at 160–180 °C passing as a countercurrent to the coffee from the most extracted cell to the one just filled with freshly ground, roasted coffee. The cells are divided into two groups: the *hot* cells at a temperature of 140–180 °C under a pressure of 1.4–1.6 MPa (14–16 bar), and the *cold* cells at a temperature of ca. 100 °C. In the hot cells, the extraction of difficult material, carbohydrates in particular, takes place; in the cold cells, the most flavorful material is extracted. Extract is withdrawn from the fresh cell, cooled to 4–5 °C in a heat exchanger, and then sent into a scale in amounts that depend on the desired yield (35–55 %) from the roast and ground coffee and on the concentration of the extract. At the end of the *withdrawal* period, a fresh cell enters the circuit, while the spent grounds are evacuated from the bottom of the most extracted cell. Figure 10 shows an extraction battery with six cells [56].

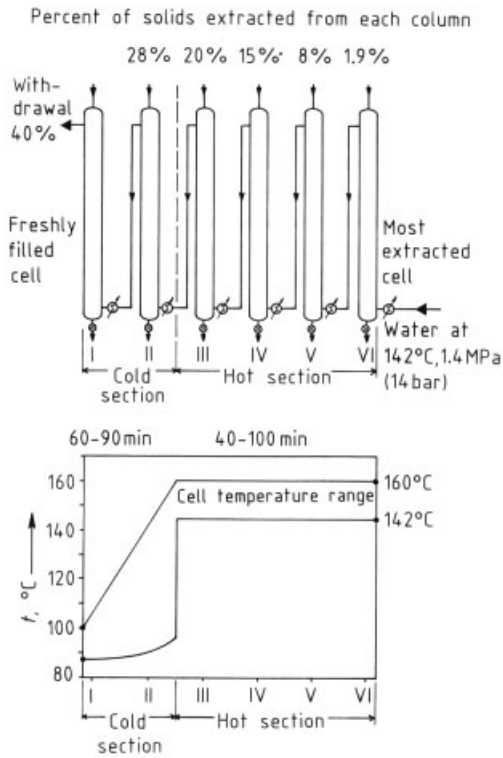

Figure 10. Extraction battery of six percolators with temperature profile [56]

The major manufacturers now use shorter percolators with more finely ground roast coffee. The shorter extraction times thus obtained, together with new techniques to protect and rectify the volatile aroma, improve quality, which now approaches that of a brew, even at extraction yields around 50 % [57].

6.2.2. Concentration

Extract at a concentration of 30 % or more can be dried directly, but with a high loss of aromatic constituents. The extract coming out of the hot section contains fewer volatile aromatic constituents than cold extract. The hot extract can be evaporated to a concentration of 50–60 % after flashing or steam stripping the volatile aroma or by separating hot from cold extracts; it then can be mixed with diluted aroma or *cold* extract to reach a *standardized* extract at ca. 40 % concentration. A more gentle concentration process

involves freezing the extract and then mechanically separating the almost pure ice crystals. By this technique, coffee extract can be efficiently concentrated up to 35–40 % with relatively low flavor or product losses.

6.2.3. Spray-Drying

Cooled and clarified extract is nebulized through a rotary or static pressure nozzle into the top of a spray tower, dried by a cocurrent flow of hot air at ca. 250 °C, and collected in the conical bottom of the tower at a temperature ca. 50 °C. Coarse, free-flowing nondusty powders with a mean size of 200–300 µm and a residual moisture of 2–4 % are obtained. Bulk density can be adjusted between 230 and 280 g/L by foaming the extract with an inert gas, usually carbon dioxide or nitrogen. The powder can be agglomerated to a coarse structure by rewetting, which makes it easy to spoon and mix with hot water. A spray-drying tower is shown in Figure 11.

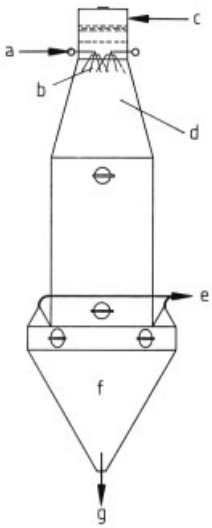

Figure 11. Spray-drying tower
(courtesy of Niro Atomizer, Copenhagen) [58]
a) Cooled and clarified coffee extract (30–50 % solids); b) Nozzle assembly; c) Hot air (ca. 250 °C); d) Transition cone for uniform spray – air mixing flow in the upper region of the dryer; e) Exhaust air to dry collector for fines recovery by cyclone or bag filter (reslurried in the feed or agglomerated); f) Chamber cone; g) Powder discharge (over 90 % recovery)

6.2.4. Freeze-Drying

During this operation, extract is frozen with separation of ice crystals, which are then sublimated under vacuum. The process entails several steps:

Primary Freezing. Cold extract at 4–5 °C is prechilled to soft ice slush at − 6 to − 8 °C. The bulk density and, to some extent, the color of the final product depend on the degree and nature of foaming, with nitrogen giving a lighter product than carbon dioxide.

Secondary Freezing. The soft ice is cooled down stepwise to − 40 to − 50 °C on a steel belt, on trays, or on drums. The size and the rate of formation of ice crystals influence the color and aroma retention of the product. Larger crystals give a darker product, with freezing times of 10–180 min, while light-colored products can be obtained by fast freezing for 30–120 s on freezing drums.

Grinding. The ice slabs at a temperature of − 40 to − 45 °C are ground and sieved to the desired size. Undersized or fine particles are reworked. The frozen particulate is put into trays, which are loaded on carriers and introduced into the drying section.

Drying. The drying chamber may be a batch type, where the trays remain fixed on heating plates through the drying operation, or a semicontinuous type, in which the tray carriers move through a vacuum lock into a drying tunnel, where they pass through successive temperature zones and exit through another vacuum lock. In a fully continuous freeze dryer, the frozen coffee particulate is fed directly into the vacuum chamber from a feed tank, progressing through the dryer to a discharge tank. The drying chamber is fitted with heating plates, which provide the sublimation energy to convert ice into water vapor. The vapor is collected on cooled condensers, and a pumping device keeps the chamber pressure lower than the triple point of water (613 Pa). A product with 1–4 % moisture at a temperature of 45–50 °C is obtained.

Table 6. Composition of dried raw coffee beans *

Constituent	Proportion, wt %*		Components
	Arabica	Robusta	
Carbohydrates	48.3–54.0	48.6–55.0	
Monosaccharides	0.2–0.4	0.5–0.7	fructose, glucose, galactose, arabinose
Oligosaccharides	5.1–8.6	2.2–6.6	sucrose
Polysaccharides	43.0–45.0	46.9–48.3	soluble and insoluble polymers containing mannose, galactose, glucose, arabinose
Acids and phenols	8.2–11.0	10.1–16.3	
volatile	traces to 0.06		acetic, formic acids
nonvolatile aliphatic	2.0–2.9	1.3–2.2	citric, malic, oxalic, quinic acids
phenolic	6.2–8.1	8.8–14.1	depsides of caffeic, ferulic, *p*-coumaroyl acids, and quinic acid; in Robustas also with amino acids
Lipids	8–17		
	15 average	10 average	fatty acids as triglycerides, diterpene esters, sterol esters, free acids
Amino acids	10.3		mainly as proteins
free	0.5	0.8	
Methylxanthines	1.2	2.2	caffeine, traces of other purines
Trigonelline	1.0	0.7	
Minerals	4.2	4.4	
potassium	1.7	1.8	

* The contribution of the analytical error to inter- and intraspecies variations may be larger than natural variation.

6.2.5. Aromatization

Volatile aroma from grinding gas, or from steaming of ground, roasted coffee prior to extraction is emulsified with oil from pressed, roasted coffee or spent grounds and sprayed on the powder, at a level of 0.3–1 % during the filling operation.

7. Composition [59]

The main differences in the composition of Arabica and Robusta raw beans lie in the relative amounts of lipids and trigonelline, which are higher in Arabicas, and of chlorogenic acids and caffeine, which are higher in Robustas. Table 6 shows the composition of Arabica and Robusta coffees.

7.1. Carbohydrates [60]

Carbohydrates constitute ca. 50 % of the raw bean. Although their content appears to be marginally higher in Robustas than in Arabicas, there are no major qualitative differences between the two botanical species. Sucrose is the main oligosaccharide, and polymers of arabinose, galactose, glucose, and mannose constitute both the soluble polysaccharides in the cells and the insoluble ones, which form, with proteins and the chlorogenic acids, the structure of the cell walls [61]. Table 7 lists the makeup of the carbohydrate portions of raw Arabica and Robusta coffee beans. Some structures that have been partially identified are listed in Table 8

Table 7. Carbohydrates in dried raw beans

Fraction	Proportion, wt %	
	Arabica	Robusta
Monosaccharides *	0.2–0.4	0.5–0.7
Oligosaccharides *	5.1–8.6	2.2.–6.6
Polysaccharides, composed of:	43.0–45.0	46.9–48.3
Mannose	21.3–22.5	21.7–22.4
Galactose	10.4–11.9	12.4–14.0
Glucose	6.7–7.8	7.8–8.7
Arabinose	3.4–4.0	2.8–4.1
Total	48.3–54.0	48.6–55.0

Fructose, glucose, galactose, and arabinose (traces). ** Mainly sucrose.

Roasting favors polysaccharide depolymerization, improving mannan extractability at high temperature (180 °C), liberation of oligo- and monosaccharides, reaching a maximum at medium roasting, followed by pyrolysis at dark roasting. Table 9 lists the carbohydrate makeup of roasted coffee beans.

During high-temperature industrial extraction, polysaccharides are partially hydrolyzed to mono- and oligosaccharides [63] independently of species. Tables 10 and 11 show the extraction behavior as a function of roasting degree of the

Table 8. Carbohydrate structures that have been partially identified in dried raw coffee beans

Fraction	Proportion, wt %		Structure
	Arabica	Robusta	
[Galacto]mannan	22		Mainly mannose as straight chains of β-(1-4)-mannan, with little galactose. Storage carbohydrates of molecular mass range 800–800 000; extremely hard and poorly soluble.
Arabinogalactan	14–15	16–17	Polymers composed of β-(1-3)-arabinogalactan, probably covalently linked to proteins. In Arabica there is a higher branching frequency and a higher level of terminal mannose than in Robusta. Structural cell-wall components of molecular mass range 200–200 000
Cellulose	8		β-(1-4)-Homoglucan. Structural cell-wall components.
Hemicelluloses	traces		Contain rhamnose (0.3 %), xylose (0.2 %) from residual parchment.

Table 9. Carbohydrates in dry roasted beans

Fraction	Proportion, wt %	Components	Behavior during roasting
Monosaccharides	0–2.0	glucose, fructose, arabinose (traces), and galactose	disappear at dark roasting
Oligosaccharides	0.4–2.8	sucrose	decreases, disappearing at dark roasting
Polysaccharides [62]		mannan	is depolymerized to a molecular mass range of 200–80000 and its high-temperature extraction behavior improves sharply
		arabinogalactan	is depolymerized to a molecular mass range of 200–50000 with a decrease in the degree of substitution of galactan chains from 1:7 to 1:12 arabinose:galactose
		cellulose	rather stable

Table 10. Carbohydrates in soluble extract from coffee roasted at three different degrees, wt % by monosaccharide constituents and by their polymers (glycans) (determined as single sugars)

Components	Light roast	Medium roast	Dark roast
Arabinose	2.4	1.7	1.5
Galactose	20.0	17.8	13.9
Glucose	0.9	10.4	0.7
Mannose	8.7	10.4	16.0
Total	32.6	31.2	32.7

Table 11. Carbohydrates in soluble extract from coffee roasted at three different degrees, wt % by degree of polymerization observed (determined as total monosaccharide)

Components	Light roast	Medium roast	Dark roast
Monosaccharides	3.34	2.89	2.01
Oligosaccharides	8.03	9.57	7.38
Polysaccharides	20.85	18.46	23.22

carbohydrates of an Arabica coffee at extraction yields of around 39 %.

While the amount of mannan in a coffee brewed at ca. 95 °C is rather low, it increases to up to two-thirds of the amount present in green beans when coffee is extracted at 180 °C. This material is, however, poorly soluble and precipitates on standing [64]. The molecular mass profile of the carbohydrates in a commercial soluble coffee allows inferences to be made about the way it was manufactured, and, in particular, the degree of high-temperature hydrolysis that was applied during industrial extraction to be estimated. The free and total carbohydrate content of commercial soluble coffees is given in Table 12.

The amount of specific free and total carbohydrates in a soluble coffee indicates if extraneous material, such as cheap maltodextrins or worthless husk and hulls, discarded during dry processed green coffee production, were used, either dry or roasted [66, 67] (Table 13).

Table 12. Free and total carbohydrate content of soluble coffee, wt % [65]

Component	Free		Total	
	Range	Average	Range	Average
Xylose			0.11–0.28	0.17
Arabinose	0.75–2.01	1.17	2.35–5.80	3.908
Fructose	0.05–0.42	0.12		
Mannose	0.13–2.62	0.53	10.2–19.7	14.2
Glucose	0.00–0.43	0.12	0.57–1.51	0.97
Galactose	0.27–0.72	0.47	13.5–24.7	19.5

Table 13. Carbohydrates used to determine the purity of commercial soluble coffees, wt %

	Pure soluble coffee	Maximum tolerable level
Free		
Fructose	0.05–0.42	0.60
Glucose	0–0.43	0.50
Sucrose	0.01–0.64	0.80
Maltose	0	absent
Total		
Glucose	0.57–1.51	1.80
Xylose *	0.11–0.28	0.40
Sugar alcohol	0.11–0.28	0.40
Mannitol **	0.02–0.22	0.30

* Xylose makes up around 50 % of the woody parchment surrounding the beans.
** From husks [68] somewhat unstable on roasting.

7.2. Acids [69]

A series of aliphatic acids, common in cellular metabolism, has been identified in the raw bean. The concentration of some increases during uncontrolled fermentation, and the C_2– C_6 aliphatic acids with their ethyl esters may contribute to the fruity and impure off-flavor of the "stinker" beans, which survives roasting. Table 14 lists the aliphatic acids in dried raw coffee beans.

Volatile acids are formed during roasting by thermal degradation of carbohydrates and nonvolatile acids. Table 15 shows the aliphatic acids in roasted coffee beans. Phosphoric acid (Arabica 1.55 g/kg, Robusta 2.8 g/kg) increases with roasting degree, particularly in Robusta [70].

More characteristic are the phenolic acids, which exist as depsides (i.e. intermolecular esters of aromatic hydroxycarboxylic acids, where the carboxyl group of one acid is esterified with the phenolic OH group of the other) with quinic acid and are known as chlorogenic acids (see Table 16). Quinic acid itself is also present in unusually large quantities in the free state (0.55 % in Arabica and 0.35 % in Robusta beans), with caffeic acid (0.2–0.3 % free in the raw bean) and ferulic acid (also present free in trace amounts in the raw bean).

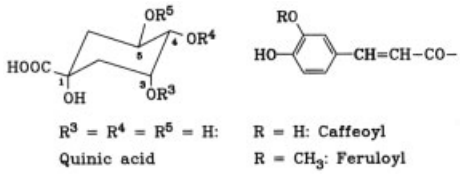

The ratio of monochlorogenic to dichlorogenic acids is related to the quality of a cup of coffee. It is lower in Robusta than in Arabica beans, and in immature or discolored beans, in which an excess of dichlorogenic acids may be responsible for the lingering metallic taste of the roasted infusion. During roasting, many reactions occur, such as decarboxylation, dehydration of the quinic acid moiety, lactonization, isomerization, polymerization, and reactions with sugars like sucrose. The feruloylquinic acids are thermally stable and survive roasting, while free quinic acid increases. At short roasting times (< 120 s), dCQAs are hydrolyzed into CQA and CA. 5-CQA rapidly undergoes esterification with carbohydrates and proteins, producing bound chlorogenic acid. On longer roasting times, phenolic and quinic moieties start degrading, generating phenols (4-ethylcatechol, catechol, phenol hydroquinone, pyrogallol [74]. The proportion of chlorogenic acids in beans of various roasting degrees is shown in Table 17.

Between 50 and 90 % of the chlorogenic acids pass into domestically brewed coffee and up to 100 % into industrial extracts, where neochlorogenic acid becomes the most important phenolic acid present (30 % of the total). Further isomerization or hydrolysis of the lactones formed at roasting may be responsible for this increase. Three stable lignan-like acidic dimers and trimers, named caffeicins, are formed on

Table 14. Aliphatic acids in dried raw beans

Acid fraction	Proportion, wt %			
	Arabica	Robusta	Component	
Volatile acids	traces – 0.1		mostly acetic acid	traces – 0.06
Nonvolatile acids	2–2.9	1.3–2.2	citric acid	0.7–1.4
			malic acid	0.3–0.7
			oxalic acid	0–0.2
			quinic acid	0.3–0.5
Total	2–3	1.3–2.2		

Table 15. Aliphatic acids in roasted bean

Acid fraction	Proportion, wt %	Components		Behavior at roasting
Volatile acids	0.4–1.0	formic acid	0.1–0.3	amount is maximum at medium roast
		acetic acid	0.3–0.7	amount is maximum at high roast
		C_2– C_{10} acids	traces	
Nonvolatile acids	1.4–3.6 [71]	lactic acid	0–0.3	amount increases slowly
		glycolic acid	0.1–0.3	amount increases
		pyruvic acid	0–0.2	amount decreases at high roast
		malic acid	0.1–0.4	amount decreases
		citric acid	0.3–1.1	amount decreases
		pyroglutamic acid	0.1–0.2	amount increases
		quinic acid	0.6–1.2	amount increases
Total	1.8–4.6			

Table 16. Depsides of quinic acid in raw beans [72, 73]

Component	Proportion, wt %		R^3	R^4	R^5
	Arabica	Robusta			
n-Chlorogenic acid [*372-97-7*]	5.4–6.2	4.9–6.4	H	H	caffeoyl
Cryptochlorogenic acid [*17608-52-5*]	0.4–0.7	0.9	H	caffeoyl	H
Neochlorogenic acid [*906-33-2*]	0.1–1.0	0.4–0.8	caffeoyl	H	H
5-Feruloylquinic acid [*40242-06-6*]	0.2–0.3	0.6–1.0	H	H	feruloyl
4-Feruloylquinic acid	traces	0.1	H	feruloyl	H
3-Feruloylquinic acid [*1899-29-2*]	traces	0.1	feruloyl	H	H
4,5-Dicaffeoylquinic acid [*14534-61-3*]	0.3–0.4	0.2–1.1	H	caffeoyl	caffeoyl
3,5-Dicaffeoylquinic acid [*2450-53-5*]	0.2–0.4	0.3–0.9	caffeoyl	H	caffeoyl
3,4-Dicaffeoylquinic acid [*57378-72-0*]	0.1–0.2	0.1–0.8	caffeoyl	caffeoyl	H
p-Coumaroylquinic acid	0.03–0.07	0.05–0.06			
p-Caffeoyltryptophan	0	0.2–0.3			
p-Coumaroytryptophan	0	0.14–0.17			
Caffeoyltyrosin	0	0.13–0.48			
"Angola II"	0	0.19–0.37			

* Only in Angola Robusta.

Table 17. Chlorogenic acids in roasted Arabica and Robusta beans, wt % [75]

Species	State of the bean			
	Raw	Light roast	Medium roast	Dark roast
Arabica	6.9	2.7	2.2	0.2
Robusta	8.8	3.5	2.1	0.2

roasting, and are present in roasted and instant coffees [76]. Table 18 shows the proportion of chlorogenic acids in industrial extracts.

A good correlation exists between titratable acidity of the roasted coffee infusion, to which phosphoric acid (0.2–0.3 %) also contributes, and the acid taste of the cup of coffee. The correlation with pH, which may vary between 4.7 and 5.5, is less clear; for example, pH is lower for washed coffees than for natural coffees and for extracts than for household infusions. The sour taste of infusions that have been kept warm for a period of time is due to an increase in free acids, in particular quinic acid formed by hydrolysis of esters, depsides, and lactones, as well as to loss of pyridine.

Table 18. Chlorogenic acids in industrial extracts, wt % [77]

Component	Quantity
n-Chlorogenic acid	0.7–1.9
Cryptochlorogenic acid	0.8–2.3
Neochlorogenic acid	1.0–3.5
Sum	2.5–7.7
3-Feruloylquinic acid	0.3–0.8
4-Feruloylquinic acid	0.3–0.8
5-Feruloylquinic acid	0.2–0.4
Sum	0.8–2.0
3,4-Dicaffeoylquinic acid	traces – 0.5
3,5-Dicaffeoylquinic acid	0.1–0.3
4,5-Dicaffeoylquinic acid	0.1–0.4
Sum	0.2–1.2

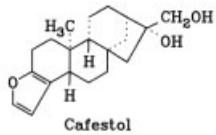

Cafestol

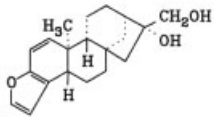

Kahweol (15–50% of cafestol in Arabicas)

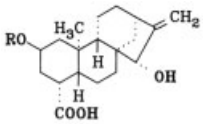

Atractyligenin (R = H)
and glycosides
(R = 2-O-β-D-glucopyranosyl
or 3'-O-β-D-glucopyranosyl-
2'-isovaleryl-β-glucopyranosyl)

7.3. Lipids [78]

Lipids constitute 15–18 %, average 15 %, of Arabica and 8–12 %, average 10 %, of Robusta raw beans. The actual range may be smaller.

The content of unsaponifiable matter is relatively high. The lipids can be subdivided into two chemically different fractions: the wax, which constitutes the outermost layer of the bean (0.2 – 0.3 %), and the oil, located inside the endosperm. The qualitative lipid composition differs only slightly between Arabicas and Robustas.

The main fatty acids are linoleic (40–45 %) and palmitic (25–35 %) acids. The fatty acid composition of the outside wax layer, rich in the stable saturated arachidic, behenic and lignoceric acids, differs from that of the oil, which contains mostly the $C_{16} - C_{18}$ acids. Triglycerides remain virtually unchanged during roasting. The free fatty acid content, higher in unwashed beans than in washed beans, increases during storage and roasting.

Fatty acid esters of diterpenes from the kaurane family are cafestol, present both in Arabica and in Robusta coffees, 16-*O*-methylcafestol, present only in Robusta beans, and kahweol, present mainly in Arabica beans. These three compounds are specific to coffee and constitute about 20 % of the lipid fraction[79]. The presence of 16-*O*-methylcafestol in a commercial roast and ground coffee sample indicates that the blend contains Robusta coffee [80]. Another diterpene of the kaurane family present in small amounts in coffee is atractyligenin, bound in various glycosides [81].

Table 19 shows the composition of lipids in dried raw coffee beans.

Table 19. Composition of the lipid fraction of dried raw beans

Lipid fraction	Proportion, wt %	Components
Triglycerides	70–80	mainly esters of linoleic and palmitic acids
Free fatty acids	0.7–2.7	higher in Robusta [83]
Diterpene esters	15–18.5	mainly esters of palmitic and linoleic acids
Triterpene, sterol, and methylsterol esters	1.4–3.2	
Free diterpenes	0.1–1.2	
Free triterpenes and sterols	1.3–2.2	mainly sitosterol, stigmasterol, and campesterol
Phospholipids	0.1	
Hydrocarbons	traces	mainly squalene and nonacosane
5-Hydroxytryptamides	0.3–1.0	amides of arachidic, behenic, and lignoceric acids
Tocopherols	0.3–0.7	α, β-, and γ-isomers

The free alcohols are liberated from the esters during roasting, and cafestol and kahweol are further decomposed to volatile dehydro derivatives, while 16-*O*-methylcafestol is stable even at high roasting temperatures. Free atractyligenin is formed by decomposition of its glycosides.

The composition of the sterol fraction is virtually unaffected during roasting and it was also

proposed as a means to evaluate the relative proportions of Arabica and Robusta beans in commercial roasted coffee blends.

Small quantities of the antioxidants, α-, β- and γ-tocopherols, and 5-hydroxytryptamides have been identified. The content of tryptamides in the wax increases with ripening of the cherry at the expense of tryptophan and decreases on storage and also after various treatments of the bean. For example, the reduced level in treated coffee (by steaming or by solvent washing) is claimed to be an index of the reduction of irritating substances in coffee. Tryptamides are decomposed by roasting, and their relationship, if any, to health effects of the brew is at least doubtful [84].

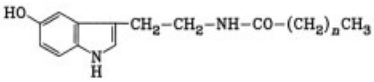

5-Hydroxytryptamides

n = 18: Arachidonamide 33%
n = 20: Behenamide 57%
n = 22: Lignoceramide 10%

0.5–2.5 g/kg of raw coffee; they correspond to 160–175 mg of serotonin; 25–50% is lost on roasting

The content of lipids in a cup of coffee depends on the brewing method and can reach a few percent in brews made from finely ground Turkish coffee or when ground coffee is boiled with water; it is insignificant in filtered coffee infusions (for physiological significance, see Section 8.3). Small amounts of oil (between 0.2 and 1 %), obtained from roast and ground coffee or from spent grounds, are often used as support for the reincorporation of volatile aroma into an instant coffee powder, but the amount of oil present in the cup is usually negligible.

7.4. Nitrogen Compounds

7.4.1. Amino Acids, Peptides, and Proteins [85]

The estimated sum of free and bound amino acids present in the raw bean is based on the total Kjeldahl nitrogen:

$$\text{Protein } \% = \text{totalnitrogen} - (\text{purinicnitrogen} + \text{betainicnitrogen}) \times 6.25 \ \%$$

Values range from 8.7–12.2 % of the dry weight. No major differences exist between Arabica and Robusta, although the higher concentration of sulfur-containing amino acids in Robustas may be responsible for the differences perceived in the volatile flavor between the two species. Table 20 shows amino acids and proteins in dried raw bean.

Table 20. Amino acids and proteins in dried raw beans

Nitrogenous fraction	Proportion, wt %	Components
Free amino acids	0.2–0.8	glutamic acid, aspartic acid, γ-aminobutyric acid, alanine, proline, and serine are the most important; total concentration is possibly lower in Arabicas than in Robustas
Proteins	7.9–12	composition near that of the free amino acids, with 20 % of glutamic acid and 10 % each of both aspartic acid and leucine; histidine, cysteine, and methionine are also present in small quantities
Soluble proteins	25–40 %	isoelectric point at pH 5.7 for Arabicas and pH 6.3 for Robustas
Structural proteins (15–25 % soluble in alkali)	60–75 %	linked with polysaccharides and chlorogenic acids in the cell wall

Only traces of free amino acids remain after roasting since they are thermally unstable, undergoing Maillard reactions with reducing sugars, to release volatile components and form melanoidins, a polymeric material. Up to 40 % of the proteins are denaturated and decompose during roasting, although one-half of the amino acids present in the raw bean remain as apparent polypeptides with the exception of arginine, which is completely degraded. All available polypeptides are extracted at 30 % extraction yield and constitute 10–12 % of an industrial extract.

7.4.2. Purines and Nucleic Acids

Several purines have been identified in the raw bean. Among them, only caffeine, a physiologically active ingredient of coffee, is present in appreciable amounts. Nucleic acids of unknown structure have also been isolated. These are shown in Table 21.

Table 21. Nucleic acids and purine bases of raw bean

Fraction	Proportion, wt %		Components
	Arabica	Robusta	
Nucleic acids	0.7	0.8	
Purine bases	0.9–1.4	1.7–4.0	caffeine, theobromine (0.2 mg/100 g), theophylline (0.06 mg/100 g), adenine, guanine, and hypoxanthine

Part of the caffeine is present as the free crystalline base in the cytoplasm surrounding the cell vacuole. Part of it is in the cell wall, where it may be bound as a 1 : 1 π-complex with potassium chlorogenates, although the complex may also be an artifact formed during extraction [86].

Caffeine is thermally stable, and only a small part of the quantity present in the raw bean is lost by sublimation during roasting. About 85 % passes into the domestic brew and almost 100 % into industrial extracts. The caffeine content of industrial extracts varies between 3.0 and 5.1 % (dry basis) for spray-dried powders, which usually contain mainly Robustas, and between 2.5 and 3.7 % for milder freeze-dried products, which contain more Arabicas.

7.4.3. Other Bases

Trigonelline [*535-83-1*] (Arabica 0.6–1.2 %, Robusta 0.3–0.9 %) and traces of betaine and choline are present in the raw bean. Some trigonelline decomposes on roasting to nicotinic acid (niacin or vitamin PP), pyridine, and other volatile aroma constituents. Residual trigonelline is easily extracted and constitutes 0.9–1.7 % of industrial extracts.

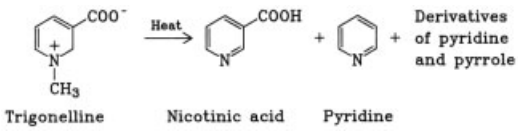

Traces of diamines, putrescine, spermine, and spermidine have also been identified in the raw bean. They are easily decomposed during roasting.

7.5. Minerals

The ash content of green coffee averages 4.0 %, varying between 3.0 and 5.4 % (dry basis) due to the botanical or geographical origin, method of preparation, and crop year. The mineral composition of dry coffee beans is shown in Table 22.

Table 22. Minerals in raw beans

Mineral	Proportion, wt %	
Potassium	1.68–2.00	
	Milds	1.68–1.70
	Brazils	1.77–1.88
	Robustas	1.84–2.00
Calcium	0.07–0.35	
Magnesium	0.16–0.31	
Phosphate	0.13–0.22	
Sulfate	0.13	

Sodium and chloride content are low, unless the coffee has been contaminated by seawater during transport. About 90 % of the ash is water-soluble and passes into the brew with only insignificant losses at roasting. The extraction yield of instant coffee can be estimated by the measure of the potassium content, when the blend and manufacturing losses are known [87].

Trace metal concentrations in the bean, particularly of manganese, may be correlated with cup quality [88]

7.6. Products Formed during Roasting

The principal thermally reactive constituents of the raw bean are the monosaccharides and sucrose, the chlorogenic acids, the free amino acids, and trigonelline (Table 23). The terminal arabinose and galactose units of polysaccharides are split off, and the basic, sulfur-containing, and hydroxyamino acids are also easily decomposed. The lipids are relatively stable, apart from an increase in free fatty acids due to hydrolysis of the esters and some degradation of linolenic acid and unsaponifiable matter. During roasting, volatile aromatic compounds and polymeric brown pigments are formed, while water and carbon dioxide are liberated.

Caramel, melanoidins, caffeine, and several aromatic constituents are responsible for the characteristic bitter taste of roasted coffee.

Table 23. Main reactive constituents lost during roasting

Constituent	Proportion in raw bean, wt %	Fraction reacted, wt %
Carbohydrates		
Monosaccharides	0.2–0.5	≈100
Sucrose	2–7	70–80
Polysaccharides		20–25
Mannose	4–6	20–30
Galactose	3	25–30
Glucose	1	5–10
Arabinose	1	50–60
Chlorogenic acids	4.7–8.6	40–98
Nitrogen compounds		
Free amino acids	0.2–0.8	100
Trigonelline	0.1–0.8	30–80
Lipids	1	3–5

7.6.1. Volatile Aroma [89]

Two hundred and thirty volatile compounds have been identified in the aroma of green coffee [90]. The majority of these compounds survives roasting and may contribute to the roasted coffee flavor [91]. Albeit made up of almost one thousand volatile compounds so far identified (most of them common to all roasted foods [92]), aroma constitutes just 0.1% of the weight of roasted ground coffee [93].

Table 24. Important off-flavors of (green) coffee

Flavor characteristics	Odor threshold in water	Responsible chemical
Medicinal, rioy	1.0 ng/L	2,4,6-trichloroanisole (TCA) [93] *
Earthy, robusta	2.5 ng/L	2-methylisoborneol (MIB) [94] *
Musty	5.0 ng/L	geosmin *
Fruity, rotten, stinker	5–10 mg/L	ethyl esters of 2- and 3-methylbutanoic and cyclohexanoic acids [95]

* Musty/earthy off-flavors in green coffee are usually due to different combinations of these substances and of methoxypyrazines; all these substances are probably of microbiological origin and apparently occur more readily in dry-processed beans.

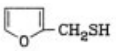

Furylmethanethiol
[98-02-2]

In the 1990s the focus of coffee aroma research has switched from the mere identification of more and more volatile substances to the selection of those that constitute the flavor of coffee. Two types of aroma impact substances are recognized: those that contribute to specific defects (Table 24), usually already present among the green coffee volatile compounds, and those that together constitute the aroma of roasted coffee.

A process [96] for improving the quality of Robusta coffee proposes the removal of MIB by steam heating. Since roasting destroys MIB, the Robusta flavor character of a brew may be due to a combination of several factors [97].

The coffee aroma obtained by extracting either the roasted beans or the brew with selective solvents, or by concentrating the headspace over the coffee or the brew, is separated in a gas chromatograph (GC). The smell of each peak is evaluated by sniffing the exit port of the GC, and described by notes such as burnt, cocoalike, buttery, floral, etc. (Figure 12). Once the key aroma components have been identified by mass spectrometry, their relative impact can be measured by the aroma dilution technique, which consists of injecting an increasingly diluted aroma into the GC until only a few peaks are still perceptible to the nose [98]. The most potent odorants identified in Arabica and Robusta brews are listed in Table 25 [99].

Various sensors, the so-called electronic noses, have become available for the detection of odors: chemical families such as α-dicarbonyls, pyrazines, and sulfur compounds can be recognized in the headspace when present at relatively high concentrations (at least 20 μg/L).

The cup profile can be presented as a star diagram (Figure 13) [100], in which each flavor is semi-quantitatively described by a panel of expert tasters. Star diagrams are very useful for comparing samples.

An important key aroma component is 3-mercapto-3-methylbutyl formate, which as a pure substance has a catty/blackcurrant smell, but when added at very high dilution to a (soluble) coffee infusion it strongly enhances the roasted coffee note [101].

7.6.2. Brown Polymers [102]

The nature of this heterogeneous class of yellow-brown to almost black pigments is still subject to speculation. They are formed by two main routes. In the first, small molecules in the cells, e.g., monosaccharides and sucrose, polymerize to yellow- to dark-brown caramel, which reacts with chlorogenic acids to form red- to

Table 25. Concentrations of the most potent odorants in Arabica and Robusta brews

Odorant	Arabica, μg/L	Robusta, μg/L	Odor threshold in water, μg/L
2-Furfurylthiol	19.1	39.0	0.01
(*E*)-*β*-Damascenone	1.3	1.5	0.00075
3-Mercapto-3-methylbutyl formate	5.5	4.3	0.0035
3-Methylbutanal	550	925	0.35
Methylpropanal	800	1380	0.7
Methanethiol	210	600	0.2
5-Ethyl-4-hydroxy-2-methyl-3(2*H*)-furanone	840	670	1.15
2-Methylbutanal	650	1300	1.3

black-brown humic acids [103]. These sugars also react with amino acids via the Maillard reaction [104] to form yellow- to black-brown melanoidins [105].

In the second, similar reactions may also occur in the cell wall following thermal degradation of the constituents. Partially soluble material, composed of intact mannose units linked together by galactose, arabinose, free phenolic acids, free amino acids, and peptide side chains, can be identified. These form the matrix on which the brown pigments are fixed. These processes are depicted in the following diagram:

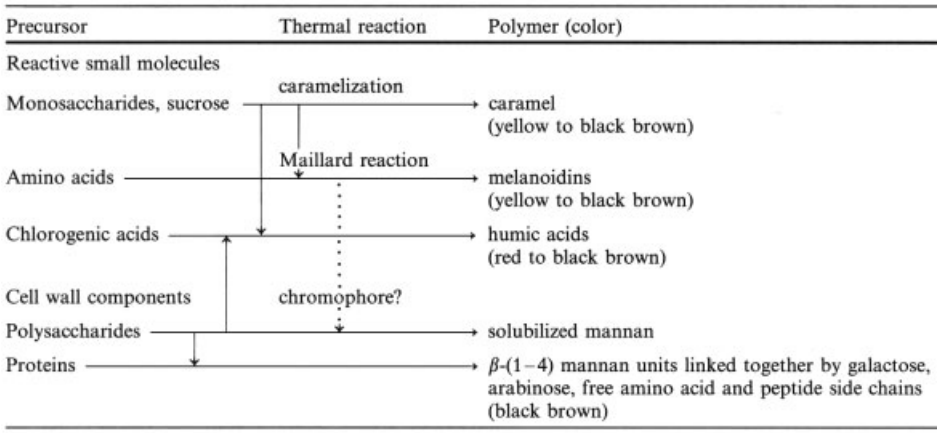

8. Consumption and Physiology

[106 – 108]

8.1. Consumption

The leading coffee-consuming countries with annual per capita consumption and green coffee disappearance are listed in Table 26 [109]. The Nordic countries have the largest per capita consumption, while the European Union is by far the main importer, followed by the United States, where the decline in consumption has now been stopped by the improved quality introduced with specialty coffees, particularly Italian-style espresso.

Table 27. Coffee consumption in some producing countries

Country	National consumption, thousands of 60-kg bags	Average individual intake, kg per capita per annum
Brazil	15 540	5.10
Colombia	1400	1.76
Indonesia	2000	0.55
India	1134	0.06
Ethiopia	1833	1.40
Mexico	1500	0.84
Guatemala	300	1.39
Thailand	500	0.44

Among the producing countries, consumption is high in Brazil and most of Latin America, intermediate in Asia, and very low in Africa with the exception of Ethiopia (Table 27), where domestic consumption nearly approaches the exported share of national production.

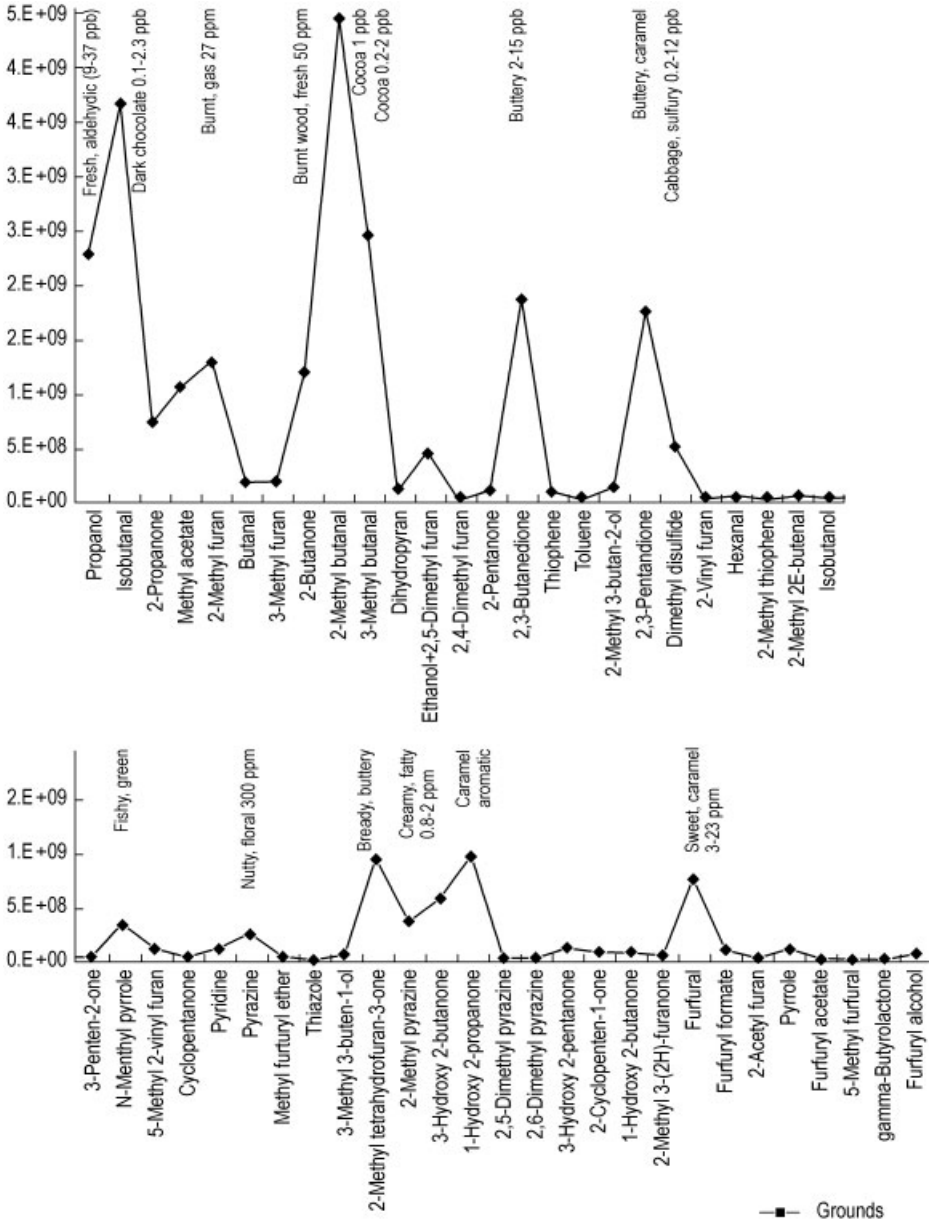

Figure 12. Odor profile of a dry ground roasted coffee. The vertical axis gives the mass spectrometric abundance. Known odor thresholds are indicated on the peaks (courtesy Dr. H. Brévard)

8.2. Dietary Value

Coffee is consumed at breakfast and between meals as a stimulant, or after a meal as an aid to digestion, particularly in Latin cultures. The nutrient profile shows that a cup of coffee contributes, via nicotinic acid, a fraction of the daily nutritional requirement for niacin (vitamin B_3 or PP), in amounts increasing with the degree of roast, and might play a role in the diet of undernourished people. The composition of a cup of coffee is given in Table 28.

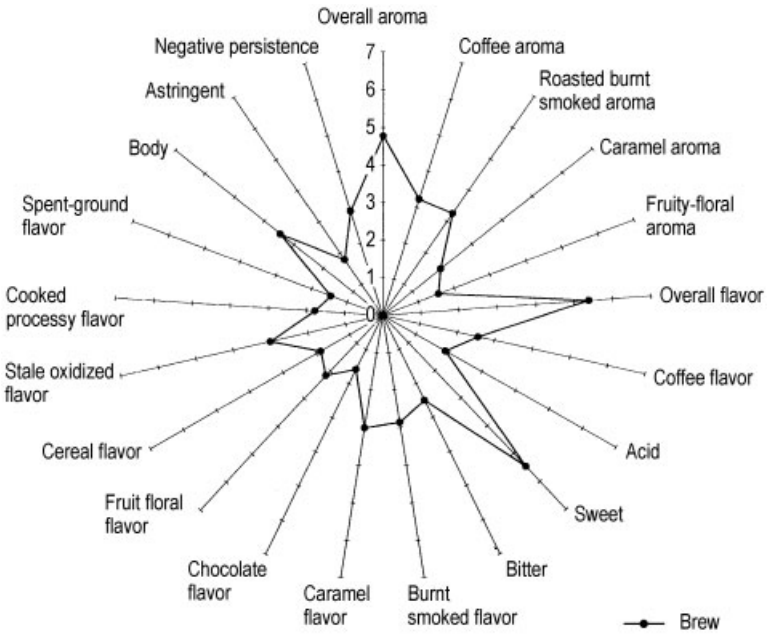

Figure 13. Sensory profile of a coffee brew. This sample has been characterized by a trained panel as possessing an intense overall aroma and flavor, a roasted/burnt/smoked note, and an off-flavor described as stale/oxidized (courtesy Dr. S. Bobillot)

Table 26. Green coffee disappearance in the principal importing countries [thousands of 60-kg bags (source: International Coffee Organization)] and consumption (kilograms per capita per year)

Country	2000	2001	2002	2003	2004	2005	Average	kg per capita
Austria	862	1029	945	753	1032	809	905	6.64
Belgium	1304	987	1635	1719	1458	1307	1402	8.26
Cyprus	68	55	53	54	57	71	60	4.48
Czech Republic	550	655	646	556	546	534	581	3.27
Denmark	784	863	806	725	849	814	807	9.17
Estonia	99	110	112	118	133	150	120	5.42
Finland	896	941	962	988	1030	1048	978	11.21
France	5446	5474	5627	5360	4972	4952	5305	5.28
Germany	9283	9475	9064	9076	10177	8052	9188	6.90
Greece	713	578	864	1003	928	911	833	4.65
Hungary	733	740	709	611	561	394	625	3.83
Ireland	83	147	143	152	220	224	161	2.50
Italy	5164	5252	5180	5503	5405	5460	5327	5.57
Latvia	112	156	151	169	167	166	154	3.60
Lithuania	172	203	207	189	234	251	209	3.67
Malta	19	15	26	34	26	23	24	3.66
Netherlands	1911	1732	1641	1827	1599	1454	1694	6.26
Norway	658	711	692	682	710	743	699	9.17
Poland	1765	1958	1952	2096	2202	2133	2018	3.06
Portugal	723	750	743	745	771	749	747	4.54
Slovakia	262	308	298	318	319	368	312	3.28
Slovenia	177	193	201	182	195	200	191	5.66
Spain	3058	2869	2907	2826	2791	2959	2902	4.45
Sweden	1182	1259	1235	1181	1242	1162	1210	8.03
United Kingdom	2381	2261	2134	2196	2423	2174	2261	2.29
European Union	38 405	38 719	38 935	39 064	40 045	37 107	38 713	5.10
Switzerland	859	876	778	891	714	1056	862	7.06
USA	21 524	19 148	19 193	20 275	20 368	20 363	20 145	4.21
Japan	6876	6919	7209	6786	7167	7441	7066	3.32
Total (including others)	67 664	65 662	66 114	67 015	68 294	65 968	66 786	0.63

Table 28. Composition of a cup of coffee, mg

Component	Brew [a]	Instant [b]
Monosaccharides	0–150	15–150
Sucrose	0–200	0–200
Aliphatic acids	130–350	60–175
Chlorogenic acids	100–250	150–200
Lipids	traces	0–10
Volatile aroma	traces	traces
Caffeine	50–150	50–100
Trigonelline	30–60	15–35
Minerals	100	200
Potassium	50	100
Niacin	traces	traces
Polymeric substances [c] (by difference)	390–1230	920–1500

[a] 7.5 g of ground roasted coffee at 22 % extraction yield.
[b] 2.0 g of soluble coffee.
[c] Polymeric substances contribute to the bitter taste in the cup [110] and to the retention of aroma constituents in the soluble powder [111]; although measured as proteins and carbohydrates, they are unavailable as such after roasting.

8.3. Physiological Value

Coffee has been the subject of extensive scientific research addressing its potential impact on human health [112], caffeine attracting over 90 % of it.

Caffeine. Caffeine is the principal active ingredient in coffee, and its stimulating action is the basis for the popularity of coffee and the other caffeine-containing beverages [113, 114].

Indicative caffeine contents of coffee products are given in Table 29.

Table 29. Caffeine content of different coffee products

Product	Portion, mL	Caffeine, mg
Filter coffee	100–200	35–175
Filter decaffeinated	100–200	3–6
Espresso coffee	20–60	30–75
Instant coffee	100–200	35–115
Instant decaffeinated	100–200	2–3
Ready-to-drink	225–285	60–200

Within minutes from ingestion, caffeine is readily adsorbed and distributed through all of the tissues and organs, then it is degraded to several methylxanthines and excreted via the kidneys within a few hours [120], with only a minor proportion appearing in the urine as unchanged caffeine. Metabolization time is longer in children and in pregnant women. The physiological effects of caffeine at normal dose levels are well known [115], even if important individual variations exist [116, 117]. It is a diuretic that stimulates the central nervous system, shortens reaction time, improves alertness, and reduces mental fatigue [118].

It is a cardiac muscle stimulant, a smooth muscle relaxant, and improves physical performance [119], but not up to the point to be considered as a doping agent by the International Olympic Committee (2004). It does not induce addiction [121] and is neither mutagenic nor teratogenic for humans at normal dosages and consumption levels. At excessive levels, other symptoms occur, such as stimulation of gastric acid secretion, increase of heart rate, sweating, nervousness, and sleeplessness. To date, the evidence gathered from many studies "does not suggest that caffeine at present levels of consumption poses a hazard to public health" according to the FDA, which has kept it on the GRAS (Generally Recognized As Safe) list [122].

Other Physiologically Active Substances. An association between coffee consumption and serum cholesterol levels, reported in 1983 [123], is linked with the presence of cafestol (and kahweol) esters in the beverage [124]. The amount present depends on the brewing technique (particularly on fineness of grinding, means of separation of grounds) and is independent of technological parameters (decaffeination, industrial high-temperature extraction). The ester content is high for boiled, French press, and Middle Eastern coffees, where there is no separation of grounds, intermediate in espresso coffee, and negligible in instant and filtered coffees.

In recent years, more and more epidemiologically evident positive effects on human health deriving from a regular coffee consumption have been highlighted [125]. Preventative effects against diabetes [126], hepatic cirrhosis and cancer [127], and Parkinson's disease [128] have been demonstrated by a large number of studies. Regrettably, explanations for the underlying mechanisms are still missing. On the other hand, a chemoprotective effect against colon cancer has been associated with the presence in coffee, beside cafestol and kahweol esters [129], of antioxidants [130] and enhancers of endogenous detoxifying mechanisms [131].

Contaminants. Several types of contaminants are sometimes encountered in coffee. Pesticides [132], present in trace amounts in green

coffee, disappear during storage, roasting and brewing. Paraffins used in the manufacture of jute/sisal bags may contaminate the green beans [133]. Polycyclic hydrocarbons are formed during roasting [134]; if allowed to accumulate in the cold spots of the roaster and of the cooler, where they tend to co-sublime with caffeine, forming whitish deposits, they may constitute a work health hazard. Ochratoxin A, a mycotoxin formed during cherry processing, storage or transport of the beans if moisture is uncontrolled, may be present in trace amounts in commercial coffee products [135], although the substance is partially destroyed during processing [136]. Relatively high levels of ochratoxin A may be present in instant coffees adulterated with husks [137].

Acrylamide, a Maillard product derived by heating sugars with asparagine and present in baked potato products and in bread crust [138], is formed in coffee during the early phases of roasting but largely destroyed during the last ones [139]. Its carcinogenic effects, while observable in rats, have not been evidenced in humans.

9. References

1. ASIC (Association Scientifique Internationale du Café), Reports of the Scientific Symposia on Coffee 1–17, Paris 1963–2004.
2. R. J. Clarke, R. Macrae (eds.), *Coffee 1–6*, Elsevier, London, 1985–1988.
3. A. Illy, R. Viani (eds.), *Espresso Coffee – The Science of Quality*, Elsevier, Oxford 2005.
4. W. H. Ukers: "All about Coffee," in *The Tea and Coffee Trade Journal*, 2nd ed., New York 1935, reprinted by Gale Research Co., Book Tower, 1976.
5. M. N. Clifford, K. Willson (eds.), *Coffee – Botany, Biochemistry and Production of Beans and Beverage*, Croom Helm, London, 1985.
6. G. Wrigley, *Coffee*, Longmans Scientific & Technical, Harlow, Essex, 1988.
7. R. Coste, *Caféiers et Cafés*, G. P. Maisonneuve & Larose, Paris, 1989.
8. D. J. Walyaro, *17th ASIC* 1997, 391.
9. M. R. Sondahl, M. Petracco, L. Zambolim, *17th ASIC* 1997, 447.
10. B. Rothfos: *Coffee Production*, Gordian-Max Rieck, Hamburg 1980.
11. S. Bee et al.: "The raw bean" in A. Illy, R. Viani (eds.): *Espresso Coffee – The Science of Quality*, Elsevier Academic Press, Oxford 2005, pp. 87–178.
12. ISO 3509:2005 *Coffee and coffee products – Vocabulary*, International Organization for Standardization, Geneve 2005.
13. ISO 10470:2004 *Green coffee – Defect reference chart*, International Organization for Standardization, Geneve 2004.
14. ISO 6668:1991 *Green coffee – Preparation of samples for use in sensory analysis*, International Organization for Standardization, Geneve 1991.
15. *Marketing and Processing of Coffee: Areas of International Cooperation*, UNCTAD, Geneva TD/B/C.1/PSC/31 1983.
16. ICO: Trade Statistics, International Coffee Organization, London, 2006, www.ico.org/trade_statistics.asp (accessed July 2006).
17. "Organic coffees", *World Coffee & Tea*, 1994, Oct. 19.
18. Council Regulation (EEC) No 2091/91 of 24. June 1991, *Official J. Europ. Communities* No L 198.1, 22.7.91; Council Regulation (EC) No 1935/95 of 22. June 1995, *Official J. Europ. Communities* No L 186/1, 5.8.95.
19. W. Heilmann: "Decaffeination of Coffee" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford, 2001, pp. 108–123.
20. B. L. Zeller, F. Z. Saleeb, 18th ASIC, 1999, 168.
21. General Foods, US 2 508 545, 1950 (A. C. Shuman).
22. Procter & Gamble, US 3 671 263, 1972 (J. M. Patel, A. B. Wolfson).
23. Procter & Gamble, EP-A 0 114 426, 1983 (R. L. Morrison, J. P. Phillips).
24. Nestlé, US 4 465 699, 1984 (F. A. Pagliaro, J. G. Franklin, R. J. Gasser).
25. General Foods, US 4 446 162, 1984 (P. D. Malizia, J. F. Trumbetas).
26. General Foods, US 2 309 092, 1943 (N. E. Berry, R. H. Walters).
27. S. N. Katz, *9th ASIC*, 1980, 295.
28. Nestlé, US 4 031 251, 1977 (G. Margolis, D. F. Rushmore, R. T. S. Liu).
29. Douwe Egberts, EP-A 0 078 088, 1982 (G. D. Mooiweer).
30. Coffex, EP-A 0 008 398, 1979 (A. Fischer, P. Kummer).
31. Nestlé, EP-A 0 040 712, 1981 (D. Green, M. Blanc).
32. General Foods, US 4 298 736, 1981 (S. N. Katz, G. E. Proscia).

33. Studiengesellschaft Kohle, US 4 260 639, 1981 (K. Zosel).
34. Kaffee-Veredelungs-Werke, US 4 334 974, 1982 (W. Sirtl).
35. Fried, Krupp, FR 2 439 553, 1979 (H. Coenen, E. Kriegel).
36. H. Martin, *10th ASIC*, 1982, 21.
37. B. Rothfos: *Kaffee – der Verbrauch*, Gordian-Max Rieck, Hamburg, 1984.
38. R. Eggers, A. Pietsch: "Preparation: Brewing Trends for the New Millennium" in R. J. Clarke, O. G. Vitzthum(eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 90–107.
39. S. Schenker et al., *J. Food Sci.* **67** (2002) 60.
40. R. Perren, R. Geiger, F. Escher, 19th ASIC, 2001.
41. S. Schenker et al., 18th ASIC, 1999, 127.
42. R. Eggers: "Roasting Techniques" in A. Illy, R. Viani (eds.): *Espresso Coffee – The Science of Quality*, Elsevier Academic Press, Oxford 2005, pp. 184–191.
43. G. Lührs, *16th ASIC* 1995, 406.
44. Hills Bros., US 4 169 164, 1979 (M. H. Hubbard, R. R. Phipps, R. L. Thompson).
45. Hills Bros., US 4 322 447, 1982 (M. H. Hubbard).
46. General Foods, US 4 349 533, 1982 (A. Stefanucci, M. G. Protomastro).
47. M. Petracco: "Grinding" in A. Illy, R. Viani (eds.): *Espresso Coffee – The Science of Quality*, Elsevier Academic Press, Oxford 2005, pp. 215–229.
48. M. C. Nicoli, O. Savonitti: "Storage and packaging" in A. Illy, R. Viani (eds.): *Espresso Coffee – The Science of Quality*, Elsevier Academic Press, Oxford 2005, pp. 230–258.
49. R. Heiss, R. Radke, L. Robinson, *8th ASIC*, 1977, 163.
50. Kenco Coffee, US 3 560 446, 1970. US 3 613 549, 1971 (R. C. Champion).
51. M. Petracco: "Preparation: Brewing Trends for the New Millennium" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 140–164.
52. F. Illy: "Apparecchio per la rapida ed automatica preparazione dell' infusione di cafè", Italian Patent, IP 333293/26, 1935
53. G. Pictet, L. Vuataz, *8th ASIC*, 1977, 261.
54. M. Sivetz, N. W. Desrosier: *Coffee Technology*, AVI, Westport, Conn., 1979.
55. R. J. Clarke: "Technology III: Instant Coffee" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 125–139.
56. General Foods, US 2 915 399, 1959 (H. Guggenheim).
57. Kraft Jacobs Suchard, EP-A 0 489 401, 1990 (O. G. Vitzthum, K. Koch).
58. A. Stoltze, K. Masters, *Food Chem.* **4** (1979) 31.
59. H. G. Maier: *Kaffee*, Paul Parey, Berlin 1981.
60. A. G. W. Bradbury: "Carbohydrates" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 1–17.
61. A. G. W. Bradbury, D. J. Halliday, *J. Agric. Food. Chem.* **38** (1990) 389.
62. V. Leloup, R. Liardon, *15th ASIC* 1993, 863.
63. V. Leloup, J. H. De Michieli, R. Liardon, *17th ASIC* 1997, 120.
64. A. G. W. Bradbury, E. D. T. Atkins, *17th ASIC* 1997, 128.
65. M.-B. Blanc, G. E. Davis, J.-M. Parchet, R. Viani, *J. Agric. Food. Chem.* **37** (1989) 926.
66. ISO 11292, "Instant Coffee – Determination of Free and Total Carbohydrate Contents", 1997.
67. R. Viani, *ISO Bulletin* 1996, Dec. 2.
68. G. E. Davis et al., *J. Agric. Food Chem.* **38** (1990) 1347.
69. H. H. Balzer: "Acids in Coffee" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 18–32.
70. M. Weers, H. Balzer, A. Bradbury, O. G. Vitzthum, *16th ASIC* 1995, 218.
71. H. G. Maier, U. H. Engelhardt, A. Scholze, *Dtsch. Lebensm. Rundschau* **80** (1984) 265 and previous articles in the series.
72. IUPAC, *Biochem. J.* **153** (1976) 23.
73. M. N. Clifford, *17th ASIC* 1997, 79.
74. V. Leloup, A. Louvrier, R. Liardon, *16th ASIC* 1995, 192.
75. L. C. Trugo, R. Macrae, *Food Chem.* **15** (1984) 219.
76. R. H. Stadler, D. H. Welti, A. A. Stämpfli, L. B. Fay, *J. Agric. Food Chem.* **44** (1996) 905.
77. L. C. Trugo, R. Macrae, *Analyst (London)* **109** (1984) 263.
78. K. Speer, I. Kölling-Speer: "Lipids" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 33–49.
79. T. Kurzroch, K. Speer, *17th ASIC* 1997, 133.
80. K. Speer, R. Tewis, A. Montag, *14th ASIC* 1991, 237.
81. R. Aeschbach, A. Kusy, H. G. Maier, *Z. Lebensm. Unters. Forsch.* **175** (1982) 337.
82. U. Mätzel, H. G. Maier, *Z. Lebensm. Unters. Forsch.* **176** (1983) 281.

83. K. Speer, N. Sehat, H. Montag, *15th ASIC* 1993, 583.
84. J. Wurziger, *Ann. Falsif. Expert. Chim.* **66** (1973) 1.
85. S. Homma: "Amino acids and Proteins" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford, 2001 pp. 50–53.
86. I. Horman, R. Viani, *J. Food Sci.* **37** (1972) 925.
87. R. J. Clarke, L. J. Walker, *7th ASIC*, 1975, 159.
88. R. Macrae, M. Petracco, E. Illy, *15th ASIC* 1993, 650.
89. W. Grosch: "Volatile compounds" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 68–89.
90. W. Holscher, H. Steinhart: "Aroma compounds in green coffee," in G. Charalambous (ed.): *Food Flavors: Generation*, *Analysis and Process Influence*, Elsevier Science, Amsterdam 1995, p. 785.
91. I. Flament: *Coffee Flavor Chemistry*, Wiley-VCH, Weinheim, 2001.
92. T. Yamanishi: "Coffee," in M. R. Teranishi, R. A. Flath, H. Sugisawa (eds.): *Flavor Research*, Marcel Dekker, New York 1983, p. 278.
93. J. C. Spadone, G. Takeoka, R. Liardon, *J. Agric. Food. Chem.* **38** (1990) 226.
94. H. Bade-Wegner, W. Holscher, O. G. Vitzthum, *15th ASIC* 1993, 537.
95. H. Bade-Wegner, I. Bendig, W. Holscher, R. Wollmann, *17th ASIC* 1997, 176.
96. Jacobs Suchard, US 5 019 4130, 1991 (R. Becker, B. Schlabs, C. Weisemann).
97. F. Rouge, C. Gretsch, K. Christiansen, R. Liardon, L. Fay, *15th ASIC* 1993, 866.
98. W. Schmidt, W. Grosch, *Z. Lebensm. Unters. Forsch.* **182** (1986) 407.
99. W. Grosch, *Chem. unserer Zeit* **30** (1996) 126.
100. S. Bobillot, H. Brévard, A. Feria-Morales, Y. Krebs, A.-F. Mermoud, *17th ASIC* 1997, 214.
101. W. Holscher, O. G. Vitzthum, *Lebensmittelchemie* **45** (1990) 12.
102. S. Homma: "Colored macromolecular compounds" in R. J. Clarke, O. G. Vitzthum (eds.): *Coffee Recent Advances*, Blackwell Science, Oxford 2001, pp. 58–65.
103. T. Nakabayashi, *Nippon Shokuhin Kogyo Gakkaishi* **31** (1984) 421, and previous articles in the series.
104. U. H. Engelhardt: "Melanoidins" in A. Illy, R. Viani (eds.): *Espresso Coffee – The Science of Quality*, Elsevier Academic Press, Oxford 2005, pp. 204–208.
105. F. M. Nunes, M. A. J. Coimbra, *Agric. Food Chem.* **49** (2001) 1773.
106. *Coffee, Tea, Maté, Methylxanthines, and Methylglyoxal*, IARC Monographs on the Evaluation of Carcinogenic Risk to Humans, Volume 51, IARC, Lyon, 1991.
107. J. E. James, *Caffeine and Health*, Academic Press, London, 1991.
108. S. Garattini (ed.), *Caffeine*, Coffee and Health, Raven Press, New York, 1992.
109. ICO: Trade Statistics, International Coffee Organization, London, 2006, www.ico.org/historical.asp (accessed July 2006).
110. H. D. Belitz, *7th ASIC*, 1975, 243.
111. H. Thaler, *Food Chem.* **4** (1979) 13.
112. J. Gray, *Nutr. Food Sci.* **6** (1998) 314.
113. B. A. Weinberg, B. K. Bealer: *The World of Caffeine*, Routledge, New York 2001.
114. P. Nawrot et al., *Food Addit. Contam.* **20** (2003) no. 1, 1.
115. R. W. Von Borstel, *Food Technol.* **37** (1983) 40.
116. K. Alsene et al., *Neuropsychopharmacology*, **28** (2003) no. 9, 1694.
117. M. C. Cornelis et al., *JAMA* **295** (2006) no. 10, 1135.
118. A. P. Smith, *Food Chem. Toxicol.* **40** (2002) no. 9, 1243.
119. T. E. Graham, *Can. J. Appl. Physiol.* **26** (2001) 103.
120. J. M. Ritchie: "The Xanthines," in L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics*, 5th ed., Macmillan, London 1975, p. 367.
121. A. Nehlig, *Neurosci. Biobehav. R.* **23** (1999) 563.
122. *FDA Consumer* **18** (1984) 14.
123. D. S. Thelle, E. Arnesen, O. H. Førde, *New England J. Med.* **308** (1983) 1154.
124. M. P. Weusten-van der Wouw et al., *J. Lipid Res.* **35** (1994) 721.
125. B. Schilter, D. Holzhäuser, C. Cavin, 19th ASIC, 2001.
126. ICO: Positively Coffee, International Coffee Organization, London, 2006, www.positivelycoffee.org/topic_dm_ref.aspx (accessed July 2006).
127. ICO: Positively Coffee, International Coffee Organization, London, 2006, www.positivelycoffee.org/topic_liver_references.aspx (accessed July 2006).
128. ICO: Positively Coffee, International Coffee Organization, London, 2006, www.positivelycoffee.org/topic_parkinson_refs.aspx (accessed Jul 2006).

129. C. Cavin et al., *Food Chem Toxicol.* **40** (2002) no. 8, 1155.
130. T. Tanaka, H. Mori, *16th ASIC* 1995, 79.
131. V. Somoza et al., *J. Agric. Food Chem.* **51** (2003) 6861.
132. J. P. McCarthy et al., *14th ASIC* 1991, 175.
133. K. Grob, *Z. Lebensm. Unters. Forsch.* **193** (1991) 213.
134. H. G. Maier, *Cáfe Cacao Thé* **35** (1991) 133.
135. G. van der Stegen et al., *Food Add. Contam.* **14** (1997) 211.
136. M. Blanc, A. Pittet, R. Muñoz-Box, R. Viani, *J. Agric. Food Chem.* **46** (1998) 673.
137. A. Pittet, D. Tornare, A. Huggett, R. Viani, *J. Agric. Food Chem.* **44** (1996) 3564.
138. CAST: Acrylamide in Food, Issue Paper 52, Council for Agricultural Science and Technology, 2006, http://www.cast-science.org/cast/src/cast_publications.php (accessed July 2006).
139. I. Lantz et al., *Mol. Nat. Food Res.* **50** (2006) 1039.

Colloids

GERHARD LAGALY, Institut für Anorganische Chemie der Universität Kiel, Kiel, Germany

1. **Definitions** 1
2. **Colloids in Industrial Uses** 2
3. **Preparation of Colloidal Particles** 4
4. **Properties of Colloidal Systems** 5
5. **Colloid Stability** 11
6. **Factors Controlling Colloid Stability** . . 20
7. **Heterocoagulation** 29
8. **References** 31

1. Definitions

Colloid chemistry involves a number of disciplines, including chemistry, physics, material sciences, and biology.

The particle dimension is the primary parameter that governs the properties of colloidal systems. In a practical sense, any particle that has a linear dimension in the range 10^{-3} to 1 µm can be defined as a colloidal particle. Very small colloidal particles (nanoparticles) have received increasing attention [1].

Colloidal particles can consist of finely subdivided bulk matter or of more or less flexible macromolecules. Particulate colloids are broadly classed as lyophobic (hydrophobic when dispersed in water) or lyophilic (hydrophilic). In addition, micellar solutions can be considered as associative colloids because the micelles formed by association of surface active agents often have colloidal dimensions [2]. This type of colloidal system is not considered here (see → Surfactants, → Laundry Detergents [3]).

Lyophobic and lyophilic colloids differ primarily in the relationship that exists between the colloidal particle and the medium in which it is dispersed. Lyophobic colloidal particles form a two- (or multi-) phase system with the medium, and a phase boundary exists between the particles (*dispersed phase*) and the medium (*continuous phase*) (Table 1).

The interactions between the particles and between the particles and the medium are governed by the structure of the particle surface, in particular the surface charge density and the extent to which it responds to changes in system parameters such as salt concentrations, organic materials, and pH.

Table 1. Classification of colloidal systems

Dispersed phase	Continuous phase	Descriptive names
liquid	gas	aerosol, fog, mist
solid	gas	smoke, aerosol
gas	liquid	foam
liquid	liquid	emulsion, microemulsion
solid	liquid	dispersion, suspension, sol, colloidal solution
gas	solid	solid foam
liquid	solid	solid emulsion, solid foam
solid	solid	solid sol, solid – solid dispersion, alloy

Lyophilic dispersions are true solutions in a thermodynamic sense and are thus one-phase systems. The dispersed and continuous phases are considered as solute and solvent, respectively. The colloidal particles are formed by coils of macromolecules. The macromolecules may be a long thread, more or less stretched, rolled up, or folded. In any case, the coil is open to the dispersion medium, and large parts of the macromolecules are in contact with it. Thus, the shape and size of the coils can change considerably with changes in solvent properties. The system is governed by the laws of macromolecular solutions [4–7].

Lyophobic and lyophilic colloids are also called *electrocratic* and *solvatocratic colloids,* reflecting the important role of electrostatic interactions in the former and of solvent – particle interactions in the latter.

Dispersions of lyophilic colloids form spontaneously when solute and solvent are brought together, as is characteristic of true solutions. In the case of lyophobic colloids, dispersions may or may not be produced spontaneously when finely subdivided matter is added to the dispersion medium. For these reasons, the terms

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_341

"reversible" for lyophilic and "irreversible" for lyophobic colloids should be avoided, because they do not describe correctly the behavior of colloidal dispersions.

The dispersed and continuous phases of colloidal dispersions can be solid, liquid, or gaseous. "Aerosol" is used for both liquid/gas and solid/gas dispersions. "Gel" designates solid/liquid dispersions which show a certain stability against mechanical forces [2], [8]. For instance, a test tube containing such a dispersion can be turned over without the dispersion flowing out. Coacervation is the separation of a colloidal dispersion (solid/liquid or liquid/liquid) into a phase with a high contents of the dispersed material (coacervate) and a phase with low content. It is an important step in microencapsulation (→ Microencapsulation, Chap. 4.1.).

2. Colloids in Industrial Uses

Although colloidal dispersions are used in numerous areas of science and technology, very few colloidal systems are produced solely because of their colloidal nature. The scientists in industry use the laws of colloid stability to improve the applicability of materials, to find new fields of applications, or to develop new products. Methods for stabilizing or destabilizing colloidal systems are indispensable for practical applications. For example, it may be necessary to produce a dispersion that is stable for a long period of storage but that destabilizes during application. Thus, it is important to understand the principles that determine stability and instability of colloidal systems.

Areas in which particles of colloidal size are regularly encountered either in stable dispersions or in coagulated or flocculated states are the following [2], [3], [8]:

Chemical Products. Paints, pigments, paper coatings, adhesives, thickening agents, lubricants, catalysts, decolorizing and absorbent materials, and light-sensitive coatings (photography).

Paper Industry and Printing. Fillers, coating, printing inks and pastes, carbonless copying papers, deinking.

Pharmaceutical Industry. Emulsions, microemulsions, cremes, ointments, and absorbing materials.

Materials Science. Powder metallurgy, pelletization, flotation, ore enrichment, alloys, dispersion hardening, superalloys, superplasticity, ceramics, cement, fibers, plastics, additives for reinforcing plastics, tar, bitumen, and native and artificial asphalts.

Oil Recovery. Enhanced oil recovery, emulsification and de-emulsification, and drilling fluids.

Soil Science. Soil engineering, soil stabilization, permeability, adsorption, and ion-exchange processes.

Environmental Science. Fog and smog, aerosols, foams, cloud seeding, water purification, sludges, and supporting materials for biocides and pesticides.

Consumer Products. Dairy products, milk, margarine, beer, waterproofing, cosmetics, cleansing agents, encapsulated products, and additives for foodstuffs.

This list can be supplemented by describing some specific systems of practical importance:

Colloidal Silica. Colloidal aqueous dispersions of silica are easy to prepare by adding sodium silicate (water glass) solutions to dilute hydrochloric acids (silica sols; see → Silica, Chap. 4.) [9], [10]. They are produced on a large scale, and different types of products are available. They are also obtained by dispersing fine-grained silicas such as Aerosil in water or in organic solvents. Silica dispersions are widely used in paints, in pharmaceutical products, and as thickening agents for other aqueous and non-aqueous dispersions.

Oxides and Hydroxides. Titania (TiO_2) is the best white pigment for paints. Aluminum oxides (alumina) and hydroxides (hydrated alumina) are used in many products such as abrasives, toothpastes, papers, plastic materials and rubber (as fillers), and in paints (as pigments, sometimes replacing titania in lower quality products).

Iron oxide pigments are used on a large scale (→ Pigments, Inorganic). A well-known application is the use of γ-Fe_2O_3 (maghemite) as magnetic pigment in magnetic tapes, which requires colloidal particles of distinct shape and sizes (→ Information Storage Materials). They are also used for preparing ferrites (→ Magnetic Materials).

Sulfides and Sulfides. Sulfides of various metals can be obtained as colloidal dispersions. Colloidal arsenic sulfide (As_2S_3) has been known for a long time. The stability of colloidal nickel and cobalt sulfides is familiar to students of analytical chemistry because these sulfides form readily instead of precipitates in the reaction of the metal ions with sulfide ions. Colloidal sulfur with particle sizes of 0.5 – 2 µm, produced by ball milling of sulfur, is used for vulcanization and as fungicide in viniculture and fruit farming.

Silver Iodide and Other Silver Halides. Besides the well-known application of silver halides in light-sensitive coatings (→ Photography), the use of AgI crystals as nuclei in cloud seeding has been studied extensively.

Clays. Clay minerals, major components of bentonites and kaolins, form platelike or sometimes lath-shaped colloidal particles (→ Clays). The annual world use of these natural colloids (excluding China and the former Soviet Union) has grown to more than 10×10^6 t of bentonite, 20×10^6 t of kaolins, and 50×10^6 t or more (exact values are unknown) of common clays. The widespread industrial uses of clays in colloidal dispersions as thixotropic gels or as dry materials have been described extensively [11–15].

Zeolites. Zeolites of almost colloidal particle sizes (1 – 12 µm) are now used as phosphate substitutes in detergents [16].

Colloidal metals are usually obtained by reduction of their salts in solution [8]. Colloidal gold dispersions illustrate the effect of particle size on color. Dispersions of 16-nm particles are orange, those with 41-nm particles are red, and those with 98-nm particles are violet with a yellow Tyndall effect of the scattered light [10], [17]. Some unsupported metals in colloidal dispersion are used as catalysts.

Alkali and alkaline earth metal colloids are available in organic solvents (Na in benzene; Ca, Sr, and Ba in propyl and isopropyl alcohol). Blue rock salt (sodium chloride) contains sodium in colloidal dispersion.

Semiconductor Colloids. Research in semiconductor colloids is developing in two directions: (1) preparation of particles with narrow size distribution and regular shape and investigation of their properties, and (2) their self-assembly into new materials and devices [18], [19].

The quantum size effect of colloidal semiconductor particles (ca. 2 – 4 nm in diameter) was first recognized by Brus. A nanoparticle represents a three-dimensional potential well for the charge carriers that are generated by absorption of a photon. The consequence is a blue shift of the threshold of absorption and fluorescence band with decreasing particle size. For semiconductors, this blue shift can be very pronounced [20]. The color of Cd_3P_2 particles changes from black (coarse particles) to brown (3 nm diameter), red, yellow, and white (1.5 nm)

Latices. Polymer particles of different structure and surface morphology (Fig. 1) are formed by emulsion, dispersion, and suspension polymerization [8, Chap. 6].

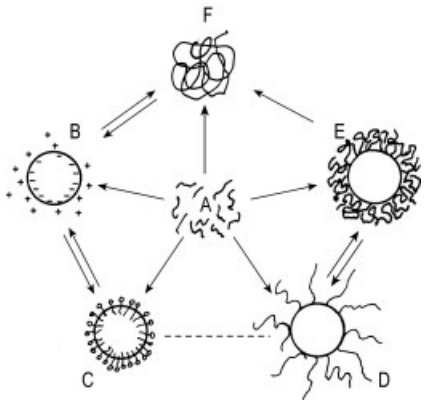

Figure 1. Types of latex particles [8]
A) Monomers; B) "Classical" particles; C) Bristly latices (surface with adsorbed emulsifier molecules); D) Hairy latices; E) Core – shell particles; F) Polymer coils

The particles usually carry negative surface charges, although positively charged latices and amphoteric forms (with both positive and negative surface charges) are also available.

Latices are used in many applications, for instance as binders in paints, for paper filling, coating, impregnation, diving equipment, latex, and latex foam [21].

Hydrophilic Colloids. Typical examples are gelatin, pectins, alginates, alginic acids, and proteins. Synthetic products are, for instance, polyacrylic acid and its salts. Alginates, gelatin, and pectins are used as thickening and thixotropic agents, emulsifiers for pharmaceutical products, cosmetics, and in the food-processing industry.

3. Preparation of Colloidal Particles

Colloidal particles are prepared by dispersion of larger particles or by condensation of atoms, ions, or molecules [2], [8], [22].

Dispersion Methods. Dispersion of larger particles into particles of colloidal size often requires high mechanical forces and can only be attained in high-energy mills such as colloid mills. Cavitation processes in the solvent [23] during milling promote particle size reduction [8, Chap. 5]. However, agglomeration processes become effective at high inputs of mechanical energy, so that the particle size reaches a lower limit. On the other hand, the forces in high-energy mills are so strong that solid materials can be alloyed like iron and silver which usually do not form alloys. Mechanical alloying is an important technical process, for instance for hardening of metals by dispersed oxides (dispersion hardening).

During particle size reduction new surfaces are created which must be covered by sufficient stabilizing agent to prevent reaggregation. Therefore, more powerful equipment which yields smaller particle sizes often requires higher amounts of stabilizers.

In the early 1900s Bredig and Svedberg studied the scattering of metals by arcs formed between two metal electrodes within a solvent. A modern version is ion beam sputtering, which has become important for preparing nanocrystalline materials.

Dispersion of a liquid phase into an other liquid phase (emulsification) is much more easier, as everyone knows from the preparation of mayonnaise. Nevertheless, technical preparation of emulsions often needs considerable amounts of energy. In all cases a careful selection of effective emulsifiers or emulsifier mixtures is imperative.

Condensation Reactions. A well-known condensation method for preparing colloidal dispersions is the reduction of nobel metal salts in aqueous or nonaqueous solvents [8], [24]. Another simple method is precipitation of low-solubilty compounds from solution. In many cases particles with broad size distributions and irregular shape are obtained. Precipitation from homogeneous systems is of great advantage. The starting system is a homogeneous solution, and precipitation is slowly initiated by hydrolytic reactions or the slow decomposition of complexes. Often, a precursor compound is dissolved in organic solvents, then slowly hydrolyzed by addition of water [8, Chap. 5]. A variety of monodisperse colloidal particles of different sizes and shapes were prepared by Matijević et al. [8, Chap. 5], [22], [25–28].

Polymer colloids are prepared by emulsion, dispersion, and suspension polymerization. Latices are obtained by emulsion polymerization (with or without emulsifier) of styrene, butadiene/styrol, acrylamide, *N*-isopropylacrylamide, etc. They are usually obtained as spherical particles, but many other forms and surface morphologies can also be produced [8, Chap. 6], [21], [29].

An example for gas phase reactions is the flame hydrolysis of $SiCl_4$ and $TiCl_4$ forming finely dispersed silica (→ Silica, Chap. 6.) and titania. Chemical vapor deposition can be used to coat preformed particles.

A problem in the preparation of colloidal particles is the need to stop particle growth when the desired size is attained. Preparation in confined systems provides advantages. One method consists of nucleation and growth of the particles within the dispersed water of microemulsions [8, Chap. 5], [30–32] or within the water cores of reverse micelles [33], [34]. Different types of nanosized colloidal particles were prepared in block copolymer micelles [35–37]. The use of templating polymer gels is an other way to produce porous materials, particles of unusal shape, and to stabilize less stable modifications [38], [39].

Dékány et al. [40], [41] described the production of nanoparticles in the confined spaces between clay mineral particles. Gold clusters and colloidal gold particles were formed in alumina nanotubes [42].

Peptization. In several cases the particles are of colloidal size but are aggregated. Colloidal dispersions are then prepared by peptization. The methods required for successful peptization depend on the chemical nature of the aggregated particles.

4. Properties of Colloidal Systems

Shape and Size of Colloidal Particles. Many colloids form spherical particles (e.g., silica, alumina, titania, metals, and latices), but particles with other shapes can also be obtained. Particle anisometry affects the colloid stability, particularly, when both positive and negative surface charges are present, as in clay colloids.

The shape of the particles can be determined by transmission or raster electron microscopy. Difficulty arises with smectitic clay minerals, which often form broad undulating mosaic sheets, making individual particles barely discernible. Modern transmission electron microscopy, however, allows the observation of very thin particles [43]. The jet-freezing technique allows the imaging of particles dispersed in the solvent phase [44] or of the microstructure of microemulsions [45]. More recently, colloidal particles and aggregation processes in aqueous media were made visible in situ by X-ray microscopy [46–49].

Colloidally dispersed particles can be monodisperse or polydisperse. The size distribution can be narrow or wide, uniform or bimodal, or even more complex [10]. Modern preparative techniques can produce particles with very narrow size distribution (monodisperse systems).

In the case of particle size distributions or anisometric particles [50], [51], the different experimental methods yield different types of average diameters. The light-scattering mean value is larger than the volume-average mean diameter, which in turn is larger than the number average diameter [8, Chap. 8].

The particle size can be obtained from electron micrographs, by light scattering (turbidity measurement, photon correlation spectroscopy, dynamic light scattering), sedimentation methods, cuvette and disk centrifuge photosedimentation, and streaming methods (Coulter counter, capillary fractionation, field flow fractionation) [8, Chap. 8], [52–54].

Shape, particle size, and size distribution are very important for highly concentrated dispersions like paper coatings and ceramic masses. The desired fluidity of such dispersions not only requires the appropriate chemical treatment but also careful selection of particle sizes and size distribution. A broader size distribution is advantageous for obtaining a lower viscosity [8, Chap. 12], [55], [56]

Light scattering [2], [10], [53], [57] is theoretically complicated, but rigorous solutions of scattering patterns are available. The use of dynamic light scattering for particle size determination is very common. However, the data are often uncritically accepted. Polydispersity cannot be recognized in every case. As the scattered light intensity is proportional to d^6, the presence of even small amounts of larger particles strongly influences the average particle diameter [8, Chap. 8], [52–54], [58].

The Tyndall effect is a well known light scattering phenomenon. When the dispersion is viewed against the light, it appears to be perfectly transparent, but when viewed from the side, the dispersion seems to be turbid. This effect is demonstrated by colloidal sulfur. Addition of dilute hydrochloric acid to sodium thiosulfate solution precipitates colloidal particles of sulfur, which grow in size with time [10]. When the dispersion is viewed directly in white light, the initial white color becomes yellow, then orange, then red, and finally dark brown. However, if the illuminated dispersion is viewed from the side, it looks turbid, milky, and the optical density increases with time.

Stability. Dispersions of lyophobic colloids may be stable for weeks, months or years, despite their inherent thermodynamic instability. Such dispersions are said to be *peptized.* These dispersions are destabilized by the addition of salts, and when specific threshold concentrations are exceeded, the dispersion coagulates. The critical coagulation concentrations c_K are essentially dependent on the valency of the ion that is charged oppositely to the particle surface (counterion). This fact was established in the early days of colloid science by the Schulze – Hardy rule. The chemical nature of the counterion is far less important, as are the valency and nature of the ion with the same sign of charge as the particle (co-ion), if there is no chemical interaction. Variations of c_K with the nature of the counterions are observable, but are small compared with the changes due to different valencies.

According to the Hardy – Schulze rule the ratio of c_K values is

$$c_K(1) : c_K(2) : c_K(3) = 1 : x : x^2$$

where x varies from 1/40 to 1/30. $c_K(i)$ is the critical coagulation concentration of counterions with valency i. One can also write

$$c_K(1) = (30-40)\, c_K(2) = (900-1600)\, c_K(3)$$

The experimental c_K values of monovalent cations range from 25 to 500 mmol/L. For di- and trivalent counterions, c_K values of 0.1 – 10 mmol/L and 0.01 – 1 mmol/L, respectively, are typical [8, Chap. 3].

Hence, di- or trivalent counterions are preferred for coagulation of peptized dispersions. In practice, aluminum or iron(III) salts are used.

Within the series of alkali cations as counterions, the c_K decreases in the order (Hofmeister series)

$$Li^+ > Na^+ > K^+ > NH_4^+ > Cs^+ > Rb^+$$

and for anions as monovalent counterions

$$SCN^- > I^- > NO_3^- > Br^- > Cl^- > F^-$$

Hence, a negative AgI dispersion is coagulated by 165 mmol/L of $LiNO_3$ and 126 mmol/L of $RbNO_3$, a positive Al_2O_3 dispersion by 60 mmol/L of KNO_3 and 46 mmol/L of KCl (see also Fig. 21).

Dispersions of lyophobic colloids are also affected by addition of polymers (see page 27).

Peptization and Repeptization. Colloidal dispersions are formed spontaneously when certain finely subdivided materials are added to a liquid under suitable conditions. For instance, fine grained quartz, amorphous silica, or sodium clays peptize by simply shaking the powders with water. Other materials require small quantities of specific electrolytes for peptization [Ag^+ or I^- ions for AgI, SH^- ions for sulfides, H^+ or OH^- ions for oxides and hydroxides (see Chapter 6)].

Dispersions that have been coagulated by salts may or may not repeptize by removing the salt, for instance by simply washing with the solvent (usually water) or by dialysis. Generally, colloids coagulated with di- or trivalent counterions are more difficult to redisperse than colloids coagulated with monovalent ions. An exchange of the multivalent ions by sodium ions is often imperative for repeptization. This is particularly true for clay minerals that do not peptize in the presence of calcium ions and in the absence of sodium ions. The Ca/Na ratio is thus a very important factor in practical uses of clays. On an industrial scale clays and clay minerals are peptized by addition of sodium ions (in amounts similar to the cation-exchange capacity), mostly in form of soda (soda-activated or alkali-activated clays) [8], [14], [59].

In other cases the pH must be changed (see page 9, page 23 and Chap. 3). Surfactants can also be used as dispersants. One of the most important methods is the use of protective colloids. These are macromolecular compounds that are adsorbed on the particles and give them enhanced surface charge density or stabilize them sterically (see page 17). Steric stabilization is often needed in ceramics [60], [61], and drilling fluids [14], [62], [63].

Particles flocculated from sterically stabilized dispersions are repeptized by changing the solvency of the solvent or, in some cases, by changing the temperature (see page 17).

In many cases, repeptization is prevented by secondary reactions between the coagulated particles (aging processes). The particles can grow together or may undergo structural changes that render repeptization difficult or impossible.

Aggregation. The theory of colloid stability (Chap. 5) pertains to the initial states of aggregation. The subsequent steps and the way in which the particles aggregate and settle to form a sediment are very important in practical applications [8, Chap. 12]. Limiting cases are illustrated in Figure 2. Sediments that are produced from highly peptized dispersions are compact and difficult to stir. Such sediments form when particles settle slowly from well-peptized dispersions at low salt concentrations (below c_K), on standing for longer periods, by slow evaporation of the solvent, or by filtration. At higher salt concentrations, if the dispersion coagulates more rapidly, voluminous sediments are formed that are easy to redisperse by stirring. Types of sediments intermediate between the examples shown in Figure 2 can be obtained by selecting appropriate experimental conditions, for instance, by carefully adjusting the salt (or polymer) concentration or by changing pH (see page 23) or the type of counterion. Coagulation with divalent counterions often favors formation of voluminous sediments.

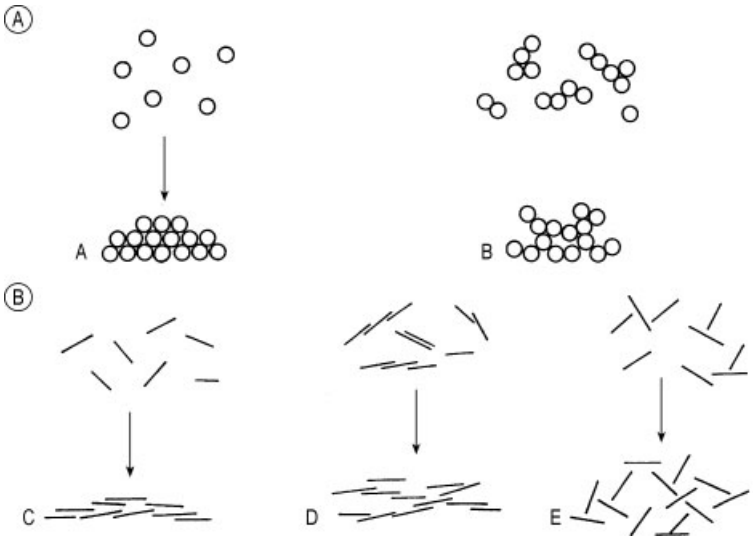

Figure 2. Sediments from peptized and coagulated or flocculated dispersions
a) Spherical particles
A) Dense sediment with repulsive interaction between the particles; B) Voluminous sediment with attractive interaction between the particles.
b) Platelike particles
C) Dense sediment with repulsive interactions between the particles; D) Band-type structure with attractive face/face interactions between the particles; E) House-of-cards structure with attractive edge/face interactions between the particles

Compact sediments form if the interaction between the particles is repulsive. Such particles are able to move, sliding and rolling past each other as they settle to form a closely packed sediment, but without coming into actual contact because of their mutual repulsion.

With sufficiently monodisperse spherical particles of latices [65], [66] and various other materials, for instance, silica [9], [67], the particles can settle in regular arrays which give Bragg reflections with light, similar to those from crystals of closely packed spheres with X-rays [64], [68]. Due to the presence of ordered domains, the dispersion shows a pronounced play of color when viewed in white light at different angles (iridescence) [67], [69]. The name *colloidal crystals* is sometimes used for these highly ordered aggregates [70].

If conditions in solution are such that the particle – particle interaction becomes attractive, the particles aggregate to some extent in the dispersion. The haphazardly formed flocs settle and pile up at the bottom of the vessel to form a voluminous sediment. The large voids between the particles are filled with solvent. Figure 3 shows different types of voluminous aggregates. Spherical particles form a three-dimensional network consisting of strings of particles tied together like strings of beads. Depending on the experimental conditions and the charge distribution, platelike particles aggregate to band-type structures (face/face contacts) or houses of cards (edge/face contacts). The different types of aggregation are very important in practical uses of kaolins, bentonites, and clays [8, Chap. 12], [71–73].

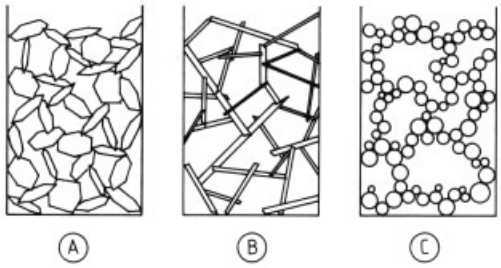

Figure 3. Network structure of A) platelike particles (house-of-cards), B) lath-shaped, and C) spherical particles

The formation of a stable voluminous network requires attractive forces at the interparticle contact sites. These are the same electrostatic

forces that cause colloidal particles to coagulate and that are attractive only in the presence of salt (see page 13). The volume of the aggregate can increase with an increase in salt concentration above the critical coagulation concentration because of the enhanced attractive forces. Divalent counterions strengthen the forces at the contact points which may increase the stability and volume of the network. The salt concentrations required are distinctly lower than those for monovalent counterions. If the attractive forces are too strong, the volume-filling network of particles contracts and fragments, and distinct aggregates form.

Amorphous silica like Aerosil form networks (Fig. 3) in many water-immiscible or organic solvents. The particles are held together by hydrogen bonds and by bridges of water. This type of network formation is used practically for thickening organic liquids [8, Chap. 12].

Finally, the term *gel* warrants discussion. It is used as a counterpart to *sol* (sol: peptized colloid; gel: coagulated colloid). However, coagulated dispersions do not usually have those properties that one associates with "gel". Thus, *gel* may be used for voluminous network structures that possess a certain degree of stiffness (*gel strength*). Many hydrophobic colloids (SiO_2 and Al_2O_3) form gels under certain conditions. The voluminous band-type and house-of-cards structures of the clay minerals show this property very clearly. Gelation is a general phenomenon of lyophilic colloids [5].

Many practical applications are very sensitive to the stability of the dispersion and the type of aggregates formed. The delicate adjustment of the stability conditions in paints is an example. The dispersed pigments in paints settle slowly during standing. To avoid compact sediments, the suspension must be held in some kind of flocculated condition, but the stability parameters must be adjusted so that the paint can be applied properly. On the other hand, the formation of thin surface coatings by sedimentation or electrodeposition (insulating films and fluorescent coatings) demands compact sediments. Again, the compactness is adjusted by varying the stability of the suspension from which the coating is obtained.

The widespread uses of clays are based on their ability to form all types of aqueous and nonaqueous colloidal systems ranging from highly fluid dispersions to stiff gels with pronounced thixotropy.

Filterability. The finely subdivided solid matter of colloidal dispersions cannot be separated from the dispersion medium by filtration through ususal filters because of the small particle size. Filter materials are now available with pore diameters as small as a few nanometers, so that larger colloidal particles can be retained (colloidal filtration [74], [75]).

The contrasting behavior of dispersions of sodium smectite (→ Clays) is noteworthy. These dispersions are difficult to filter, even through filters with large pores. A thin filter cake forms at the very beginning of filtration and makes the filter rather impermeable. The clay platelets become aggregated parallel to each other and form a dense sediment. For filtration, the dispersion must be held in a coagulated or flocculated state, for instance by adding Ca^{2+} ions, organic solvents such as ethanol, or small amounts of polymeric materials. Flocculation by polyanions is widely used in the processing of clays [14], [62]. On the other hand, engineering work takes advantage of the formation of almost impermeable clay films to reduce the permeability of rocks and soils. This "plastering" is also important in drilling operations [11], [14], [62].

Sedimentation. When spherical colloidal particles in dilute dispersions move independently, they settle under the influence of gravity [2], [10], [57], [74–77]. The particle falls through a distance h in time t (Fig. 4)

$$t = 18\eta_0 / (\varrho - \varrho_0)\, g \cdot h / d^2 \qquad (1)$$

(ϱ, d: density and diameter of the particle; ϱ_0, η_0: density and viscosity of the dispersion medium, g: gravitational acceleration: 9.81 m/s^2).

In water, 10-µm quartz particles ($\varrho - \varrho_0 = 1.65$ g/cm^3, $\eta_0 = 1.001 \times 10^{-2}$ g cm^{-1}s^{-1}, 20 °C) need $t = 18$ min 34 s to fall a distance of 10 cm, 1-µm particles need $t = 30$ h and 57 min. With too small particles (< 1 µm for quartz or clay minerals) the sedimentation time becomes so long that sedimentation is disturbed by Brownian motion. A centrifugal field must be used for sedimentation of such particles. The gravitational acceleration in the Stokes equation (Eq. 1)

is replaced by the radial acceleration, which can be increased to ca. $10^3\,g$. The particles then fall the distance $h = r - r_0$ (Fig. 4) in the time t [78]:

$$t = \frac{K_c}{d^2 U^2} \ln r/r_0 \quad \text{(in seconds)} \tag{2}$$

where $K_c = 18\,\eta_0 \cdot 60^2/(\varrho - \varrho_0)\,4\,\pi^2$ with $U = \text{rpm}$ (rotations $\cdot$ min^{-1}). For $U = 5000$, $r = 12$ cm, $r_0 = 10$ cm, the sedimentation times of 0.1-µm and 0.01-µm quartz particles in water at 20 °C are 12 min 7 s and 20 h 12 min.

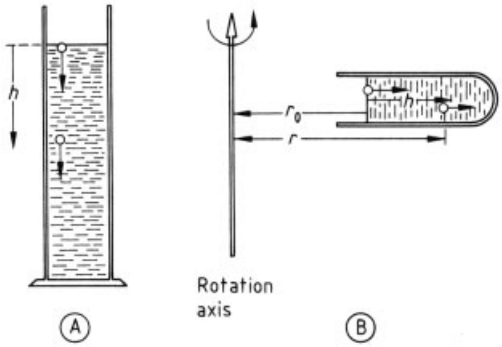

Figure 4. Sedimentation under gravity (A) and under centrifugal forces (B)

The high centrifugal acceleration required for sedimentation of particles < 0.01 µm in diameter in reasonable times can only be realized in ultracentrifuges, in which the radial acceleration can easily be increased to $10^5\,g$. This instrument is used extensively for characterization and determination of the molecular mass of macromolecular colloids, particularly those of biological origin such as proteins, nucleic acids, and viruses.

When the particles are not spherical, the application of the above formulas leads to *Stokes equivalent diameters* [50], [51]. In the sedimentation of nonspherical particles, the friction factor f is larger than that for spherical particles ($f_0 = 3\,\pi\,\eta_0\,d$), and the viscous resistance is increased. Thus the particles settle more slowly than spherical particles of the same mass. The same effect is produced by particle solvation. A particle containing a given volume of dry material has its smallest possible friction factor if it is an unsolvated sphere [10], [57], [76].

Sedimentation under gravity and in the centrifugal field finds practical application in separating particles into fractions of different sizes and for determination of particle size distributions. For commercial uses, the fractionation is often confined to the removal of particles of too large or too small dimensions. Fractionation by sedimentation is applied particularly in processing clay minerals and pigments. Experimental details for such laboratory techniques as fractionation and determination of particle size distribution have been described [8, Chap. 6], [78]. Several "particle size analyzers" are available, but manual control experiments are recommended.

Rheological Phenomena. In dilute colloidal systems the particles move independently of each other and the system behaves as a Newtonian fluid [2]. With increasing particle concentration, a gradual transition from Newtonian to plastic or pseudoplastic behavior is generally observed [10], [57], [76], [79]. The transition occurs when mutual interactions between the particles become significant. This can be illustrated by dispersions of silica (Aerosil) [79]. At pH 3 the dispersion exhibits Newtonian flow up to silica concentrations of ca. 10 wt %, above which shear thickening at high shear rates and increasing hysteresis effects are observed. At pH 9, there is a gradual transition from Newtonian to pseudoplastic flow as the silica content is increased above 5 wt %. Figure 5 shows a plot of the plastic viscosity (obtained from the slope of the linear portion of the shear rate – shear stress curves) at different pH values as a function of particle concentration. Near pH 7 the viscosity increases strongly above a particle concentration of 8 wt %. At pH 9 and pH 3, higher particle concentrations are required to obtain dispersions of high viscosity.

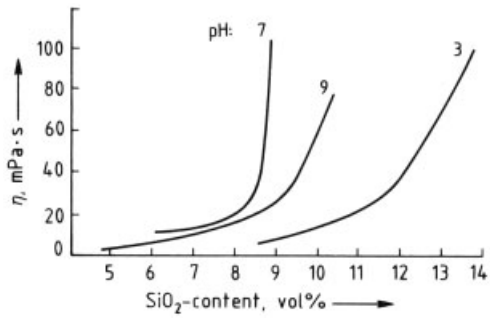

Figure 5. Plastic viscosity of silica dispersions at different pH as a function of solid content [79]

Figure 5 demonstrates the effect of particle – particle interactions on rheological

phenomena. At low pH the surface charge density of the silica particles is low (see Chapter 6), and the particles aggregate to dense flocs [80] with Newtonian flow behavior. Because of the low surface charge density, the concentration of the particles can be increased considerably without significant viscosity changes. At high pH, the increased negative surface charge density and electrostatic repulsion prevent the particles from forming large dense flocs so that the number of moving particles and small aggregates is increased. This effect and the stronger electrostatic interaction lead to higher viscosities than those at pH 3. In the neutral range, the particles interact strongly through hydrogen bonds and form network structures with considerable increase in viscosity. The electrostatic repulsion between the particles is too weak to disrupt the network but strong enough to prevent the formation of dense flocs as at pH 3.

It should be noted that the viscosity of silica dispersions depends on the nature of the silica (fumed silica such as Aerosil and precipitated silica such as Kieselgel) and the dispersing procedure (ultrasonication or not [80]).

The dependence of the viscosity of colloidal silica dispersions on concentration, pH, and electrolyte or polymer addition is widely used in practical applications. Fumed silica serves as a thickening and gelling agent for aqueous and nonaqueous liquids. In polar solvents the thickening is optimal in the neutral pH range. The thickening efficiency of silica can be improved by additives that stabilize the network structure. Suitable additives for dispersions in nonpolar solvents are water, ethylene glycol, or ethylene diamine [8, Chap. 12].

Clays find widespread applications because of their special rheological properties and pronounced thixotropic behavior. Thixotropy is the property of dispersed systems to become more fluid (lower viscosity) under the influence of mechanical forces (stirring, shaking, shearing) and to regain the higher viscosity at rest. It is caused by fragmentation of the network structure by the mechanical forces. At rest, these fragments are pushed into new contact positions by the Brownian motion of the solvent particles and reaggregate to the volume-filling network. Practical applications of thixotropy are found in tomato ketchup, paints, paper coating, drilling fluids, and civil engineering (slurry walling, underground sealing) [2], [8, Chap. 12], [11], [12], [14] [62], [81].

Electrokinetic phenomena are observed when two phases move relative to each other under the influence of an electric field. Electrophoresis describes the motion of charged particles or droplets in an applied electric field (for other electrokinetic phenomena see [8], [10], [53], [57], [76], [82]). A thin layer of liquid adheres to the particle or droplet surface so that the shearing plane is located anywhere in the liquid. The potential at the shearing plane is called the electrokinetic or zeta potential ζ. Because of the unknown distance of the shearing plane from the particle surface, electrokinetic phenomena are often difficult to interpret.

The velocity u of a particle in an electric field E depends on the dimensions of the particle and the atmosphere of counterions around it, i.e. on $\varkappa a$ (see page 13; for details see [8], [53], [57], [82]): $\varkappa a \ll 1$: small particles with large diffuse ionic layers:

$$u = 2\varepsilon\varepsilon_0\zeta E/(3\eta_0)\ \text{(Hückel – Onsager equation)} \quad (3)$$

(ε: relative dielectric constant, $\varepsilon_0 = 8.854 \cdot 10^{-12}\,\mathrm{AsV^{-1}m^{-1}}$, E: applied electric field strength, η_0: viscosity of the dispersion medium). $\varkappa a \gg 1$: large particles with diffuse ionic layers much thinner than the particle radius

$$u = \varepsilon\varepsilon_0\zeta E/\eta_0 \quad (4)$$

(Helmholtz – Smoluchowski equation)

$\varkappa a \approx 1$

$$u = 2\varepsilon\varepsilon_0\zeta E/3\eta_0 \cdot f(\varkappa a)\ \text{(Henry equation)} \quad (5)$$

[$f(\varkappa a)$ [76], [82] $f(\varkappa a) = 1$ for $\log(\varkappa a) \leq -1$, $f(\varkappa a) = 1.5$ for $\log(\varkappa a) \geq 3$; but note a different variation for metal colloids]. Table 2 illustrates the applicability of the Equations (3) – (5) for different particle diameters and electrolyte concentrations. Usually particles between 0.1 µm and 10 µm are observable in electrophoretic cells so that the Smoluchowsky and Henry equations are more applicable than the Hückel equation.

Corrections to these equations can be made to account for the particle – particle interactions in dispersions that contain more than 5 wt % solid particles [82, Chap. 4]. In such dispersions the

Table 2. Applicability of the Helmholtz – Smoluchowski, Henry, and Hückel – Onsager equations (Eqs. 3 – 5) for determining zeta potentials

Normality of solution	Approximate particle diameter (nm) for which the equations are valid		
	Helmholtz – Smoluchowski	Henry	Hückel – Onsager
1	≥ 80	0 – 80	–
0.1	≥ 200	0 – 200	–
0.01	≥ 800	1 – 800	0 – 1
0.001	≥ 1800	2 – 1800	0 – 2
0.0001	≥ 8000	10 – 8000	0 – 10
0.00001	$\geq 16\,000$	20 – 16 000	0 – 20

mobility of the particles is reduced, and too small zeta potentials would be obtained without the corrections. Further corrections are needed for electrophoretic retardation and relaxation [83–85]. Since the counterions with their hydration shells tend to move in an opposite direction from the particles, the movement of the particles is retarded. Furthermore the double layer must be constantly restored in front of the particle and dispersed behind the particle. This restoration takes a finite time (relaxation). The double layer is therefore shifted somewhat with respect to the particle and a field is created that opposes the applied electric field and reduces the mobility. For practical purposes, the corrections can be neglected for low particle concentrations and small zeta potentials. Recently, the influence of surface conduction on the mobility has become evident [8, Chap. 10], [82, Chap. 4], [86].

The electokinetic mobility is an important quantity in colloid chemistry and is a fundamental property of colloidal particles. It gives important information such as the sign of the surface charge, the way in which the charge changes when experimental parameters are changed, and the conditions that establish the isoelectric point. It responds to the addition of inorganic compounds and monomeric and polymeric organic materials, thus providing information about the interaction of the colloid particles with these materials. Zeta potential measurements can also be used for some kinds of quality control. For example, commercial TiO_2 pigments, both anatase and rutile, can exhibit isoelectric points between pH 1 and 7, depending on the way they have been prepared. The variation is caused by adsorption of anions, mainly sulfate and phosphate anions, which are introduced during industrial production. The isoelectric point of pure TiO_2 (both modifications) is near pH 6.9. Measurements of zeta potentials can also be used to obtain information on the flotability of minerals in flotation processes [10], [87].

The electrokinetic mobility usually is measured by microelectrophoresis and laser-doppler electrophoresis (commercial instruments are available). In the 1990s the instrumentation and theory of the electroacoustic characterization of colloidal dispersions and emulsions made great strides. An important advantage is that the particle dynamic mobility can be determined in dispersions of essentially any concentration above about 1 wt %. Apparatus is available to measure particle sizes in the range 0.1 – 10 µm and the zeta potential of particles down to nanometer dimensions [88]. The moving boundary method [76], [89–91] is simple and does not require large and expensive instrumentation. Horn and Ditter [92] describe a self-made laser interferometer for mobility measurements of pigment particles in organic solvents.

In most commercial instruments the zeta potential is calculated from the mobility by means of the Helmholtz – Smoluchowski equation (Eq. 4). However, conversion of mobility data to zeta potentials is not without arbitrariness [8]. It is recommended to report the electrokinetic mobility instead of possibly false zeta potentials.

5. Colloid Stability

Origin of Colloid Stability. No finely subdivided material dispersed in a continuous phase is ever stable in a thermodynamic sense. The dispersed particles or droplets tend to reduce the large interfacial area by aggregation to a continuous phase. The coarsening process may be slow, but it is inexorable. If special stabilization mechanisms are absent, the particles or droplets aggregate spontaneously (Fig. 6). Two mecha-

nisms are known for stabilizing the colloidal state, electrostatic stabilization and steric stabilization. Electrostatic stabilization occurs when the surface of the particle or droplet carries an electric charge. The particles are surrounded by atmospheres of counterions and an electrostatic repulsion is created between these atmospheres. As long as the repulsion V_R exceeds the van der Waals attraction V_A between the particles, the colloidal dispersion is stable. When the van der Waals attraction overcomes the electrostatic repulsion, the system becomes unstable. This occurs when the concentration of salts added to the dispersion exceeds a critical value.

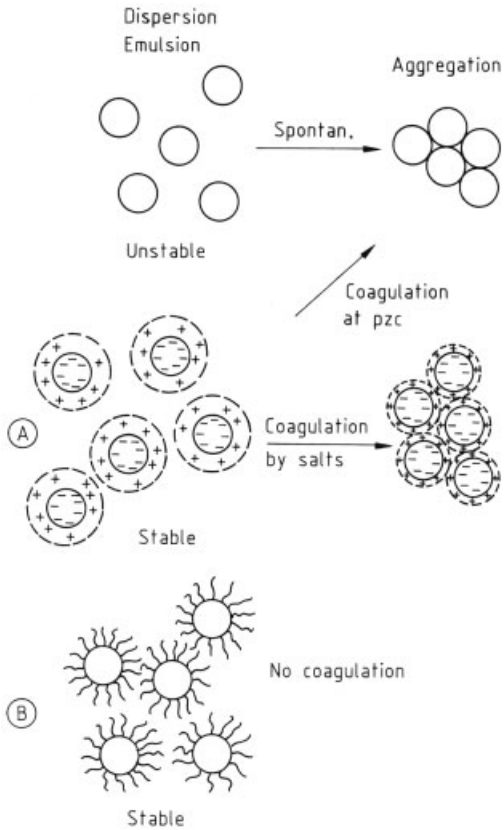

Figure 6. Electrostatic (A) and steric (B) stabilization of colloidal dispersions and emulsions

The dispersion also becomes unstable when the surface charge disappears. This occurs when the concentration of protons or other potential-determining ions (see Chapter 6) is changed or when counterions are introduced that are strongly absorbed (see Chapter 6).

Long-chain molecules or macromolecules adhering to the particle surface cause steric stabilization. Envelopes of organic matter around the particles prevent the particles from aggregating. Such dispersions are often difficult to destabilize because salts exert little influence. Destabilization can be achieved by changing the temperature or the quality of the solvent, thus changing the polymer – solvent interaction (see page 17).

Liquid particles in emulsions have the same kinds of interactions as solid colloidal particles. They can be stabilized electrically by surface charges or sterically by macromolecules (→ Emulsions). Important emulsifiers are nonionic surfactants like alkyl poly(ethylene oxides). Ionic surfactants are used in combination with nonionic surfactants [8, Chap. 7].

Stabilization by solid particles is one of the pathways that lead to new types of emulsions [93–95].

DLVO Theory: the Ionic Layer. The DLVO theory (DLVO: Derjaguin, Landau, Verwey, and Overbeek [8], [10], [53], [57], [76], [82], [96–99]) describes the interplay between electrostatic repulsion and van der Waals attraction. The theory may be illustrated for particles with a negative surface charge density (σ_0), dispersed in a continuous liquid phase of dielectric constant ε, for instance water with $\varepsilon = 78.5$ (25 °C). An electric potential $\psi = f(x)$ then exists that decays from $\psi = \psi_0$ at $x = 0$ (ψ_0 = surface potential) to $\psi = 0$ far away from the surface ($x \to \infty$) (Fig. 7 and Fig. 19). A cation (counterion) at position x is then attracted by the surface, but the electrostatic force is counteracted by the thermal motion of the ions, and the distribution of the counterions is determined by the Boltzmann equation. The concentration of counterions, therefore, decreases from a value at $x = 0$ to the bulk concentration at $x \to \infty$ (Fig. 8). Anions (co-ions) are repelled from the surface.

The counterions and the co-ions are not arranged in a distinct pattern on the particle surface but are distributed within an atmosphere surrounding the particle. The dimension of this diffuse (or Gouy – Chapman) layer can be substantial (Fig. 8). The thickness of the diffuse layer is given by $1/\varkappa$, the Debye – Hückel length. The quantity $\varkappa$ contains a cluster of

constants (concentration c_i (mol/m^3) and valencies v_i of the ions, $F = 96\,485.29$ A s mol^{-1}, $\varepsilon_0 = 8.854 \times 10^{-12}$ A s V^{-1} m^{-1}):

$$\varkappa^2 = F^2 \sum v_i^2 c_i / (\varepsilon\varepsilon_0 RT) \tag{6a}$$

When only one type of salt is present with $v^+ = v^- = v$ (for instance: NaCl, KNO_3, $MgSO_4$) $\sum v_i^2 c_i = 2v^2 c$ and

$$\varkappa^2 = 2F^2 v^2 c / (\varepsilon\varepsilon_0 RT) \tag{6b}$$

The reciprocal length $\varkappa$ is of general importance and is widely used in colloid science instead of the concentration. For water ($\varepsilon = 78.5$ at 25 °C) one obtains with c expressed in mol/L:

$$\varkappa^2 = 32.87 \times 10^6 v\sqrt{c}\ \mathrm{cm}^{-1} \tag{6c}$$

For 10^{-5}, 10^{-3}, and 10^{-1} mol/L solutions, $\varkappa$ is ca. 10^5, 10^6, and 10^7 cm^{-1}, respectively ($v = 1$). The reciprocal value $1/\varkappa$ is generally considered a measure of the thickness of the diffuse ionic layer. For large or platelike particles the potential is reduced to $\psi_0/e = 0.37\,\psi_0$ at the distance $x = 1/\varkappa$ from the surface. The ionic layer extends further into the continuous phase, and ψ becomes $0.01\,\psi_0$ at $x = 4.6 \cdot 1/\varkappa$.

The extension of the ionic layer is very sensitive to salt addition. This is seen from the variation of $1/\varkappa$ with the salt concentration (Eq. 6 c). As the concentration of salt ($v = 1$) is increased by a factor of 10 or 100, the thickness of the ionic layer $1/\varkappa$ is reduced by the factors 0.32 and 0.1, respectively. At a salt concentration $c = 10^{-5}$ mol/L, the thickness of the ionic layer is $1/\varkappa = 100$ nm. Increasing the concentration to 10^{-3} or 10^{-1} mol/L compresses the double layer to $1/\varkappa = 10$ or $1/\varkappa = 1$ nm.

Polyvalent ions exert a similar effect (Eq. 6 c). Di- and trivalent ions compress the diffuse layer thickness by a factor of 2 and 3, respectively.

DLVO Theory: Electrostatic Repulsion and van der Waals Attraction. The compression of the double layer has dramatic effects on the stability of colloidal dispersions. At low electrolyte concentration, the atmosphere of counterions is extended, and particles are widely separated. Higher electrolyte concentrations compress the diffuse layer, and the particles approach more closely (Fig. 9). At $\varkappa = 10^7$ ($c = 10^{-1}$ mol/L, $v = 1$) the particles can approach to ca. 5 nm until they "feel" the mutual repulsion.

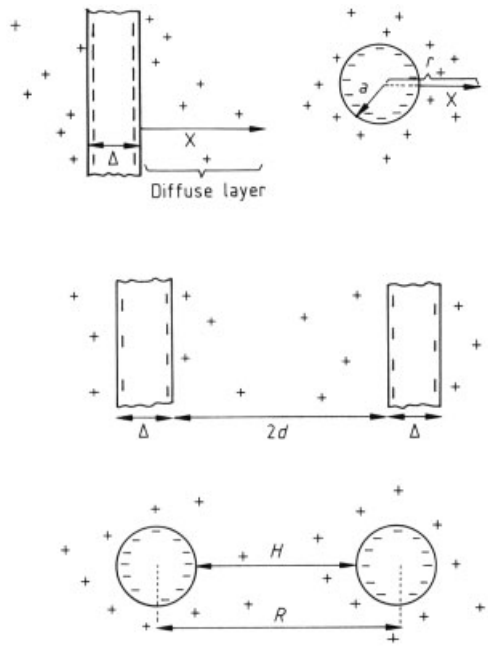

Figure 7. Diffuse ionic layers around and between spherical and platelike colloidal particles

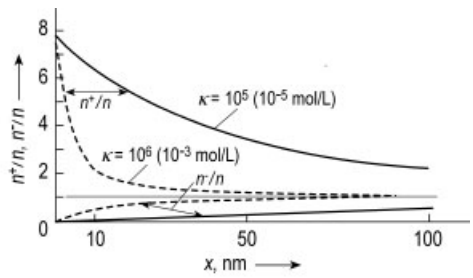

Figure 8. Concentration of monovalent counterions (n^+) and co-ions (n^-) around platelike particles or large spheres at different electrolyte concentrations (surface potential ≈ 50 mV). $\varkappa$ in cm^{-1}

The electrostatic repulsion between two particles can be calculated by the approximation:

$$V_\mathrm{R} = 32\pi a\varepsilon\varepsilon_0 \left(RT/(vF)\right)^2 \gamma^2 e^{-\varkappa H}\ \mathrm{J} \tag{7}$$

where H is the nearest distance between two particles (Fig. 7), and the expression for γ is

$$\gamma = \left(e^{z/2} - 1\right) / \left(e^{z/2} + 1\right)$$
$$\text{with } z = vF\psi_0/(RT) \tag{8}$$

It should be noted that γ approaches 1 at higher surface potentials ($\psi_0 \geq 150$ mV). The electrostatic repulsion then becomes almost independent of the surface potential.

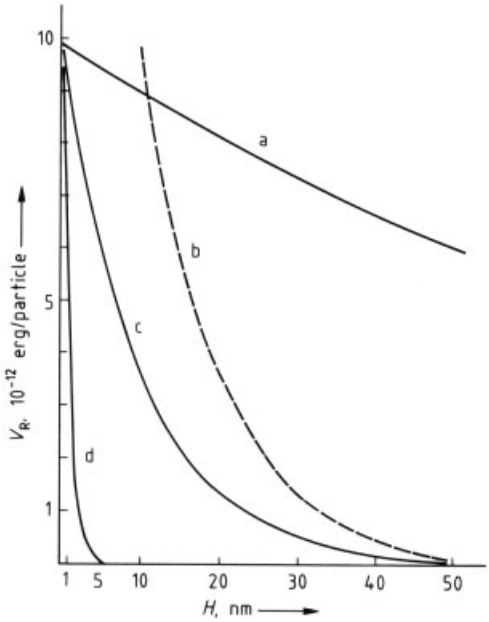

Figure 9. Electrostatic repulsion V_R between spherical particles (radius 100 nm) at different electrolyte concentrations, $v = v^- = 1$, surface potential: ≈ 50 and 100 mV, H: see Fig.7
a) $z = 2$ ($\psi_0 \approx 50$ mV), $\varkappa = 10^5$ cm^{-1} (10^{-5} mol/L);
b) $z = 4$ ($\psi_0 \approx 100$ mV), $\varkappa = 10^6$ cm^{-1} (10^{-3} mol/L);
c) $z = 2$ ($\psi_0 \approx 50$ mV), $\varkappa = 10^6$ cm^{-1} (10^{-3} mol/L);
d) $z = 2$ ($\psi_0 \approx 50$ mV), $\varkappa = 10^7$ cm^{-1} (10^{-1} mol/L)

The van der Waals attraction is nearly independent of the salt concentration. A good approximation for spherical particles is

$$V_A = -\frac{A}{12}\left(\frac{L}{H} + 2\ln\frac{H}{L}\right) \tag{9}$$

with $L = a + 0.75\,H$ [100]. The Hamaker constant A is dependent on the chemical composition of the particles. Typical values for aqueous dispersions follow [8], [101], [102]:

Metals:
$A \approx (15-30) \times 10^{-20}$ J
Oxides, silicates:
$A \approx (1-5) \times 10^{-20}$ J
Inorganic salts:
$A \approx (0.3-5) \times 10^{-20}$ J
Polymers (latices):
$A \approx (0.3-1.4) \times 10^{-20}$ J

A more precise value for latices has been reported [103].

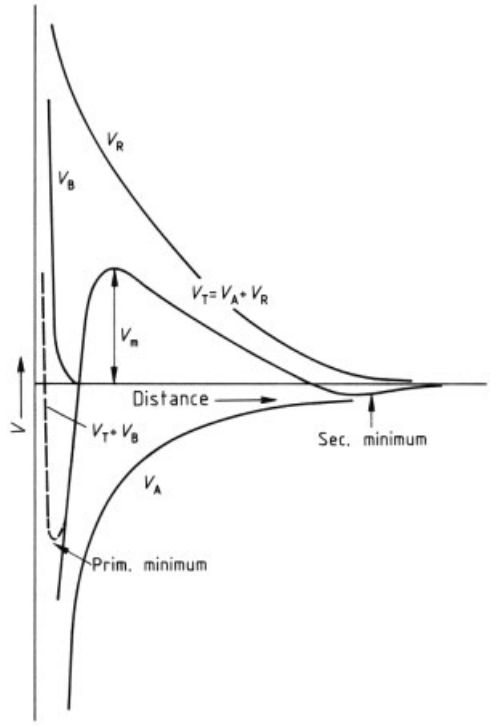

Figure 10. Schematic representation of the interactions between colloidal particles
V_R electrostatic repulsion; V_A van der Waals attraction; V_B Born repulsion

The van der Waals attraction decays more rapidly with particle distance than does the double layer repulsion. It becomes operative only when the double layer is sufficiently compressed that the particle can approach closely. Under these conditions the attraction overcomes the double layer repulsion, and the particles attract each other; thus coagulation is initiated. Figure 10 illustrates schematically how the attraction curve V_A (negative sign) and the repulsion curve V_R (positive sign) are superimposed to give the total interaction curve $V_T = V_A + V_R$. With decreasing particle distance, V_T increases from a shallow minimum (secondary minimum) at a great distance to a maximum and then falls off sharply. When the particles come into direct contact, Born repulsion V_B prevents further approach. Superimposition of V_B and V_T creates the primary minimum.

Because the repulsion V_R varies with the salt concentration, V_T responds sensitively to salt addition. Calculated curves are shown in Figure 11 for latex particles with surface potentials

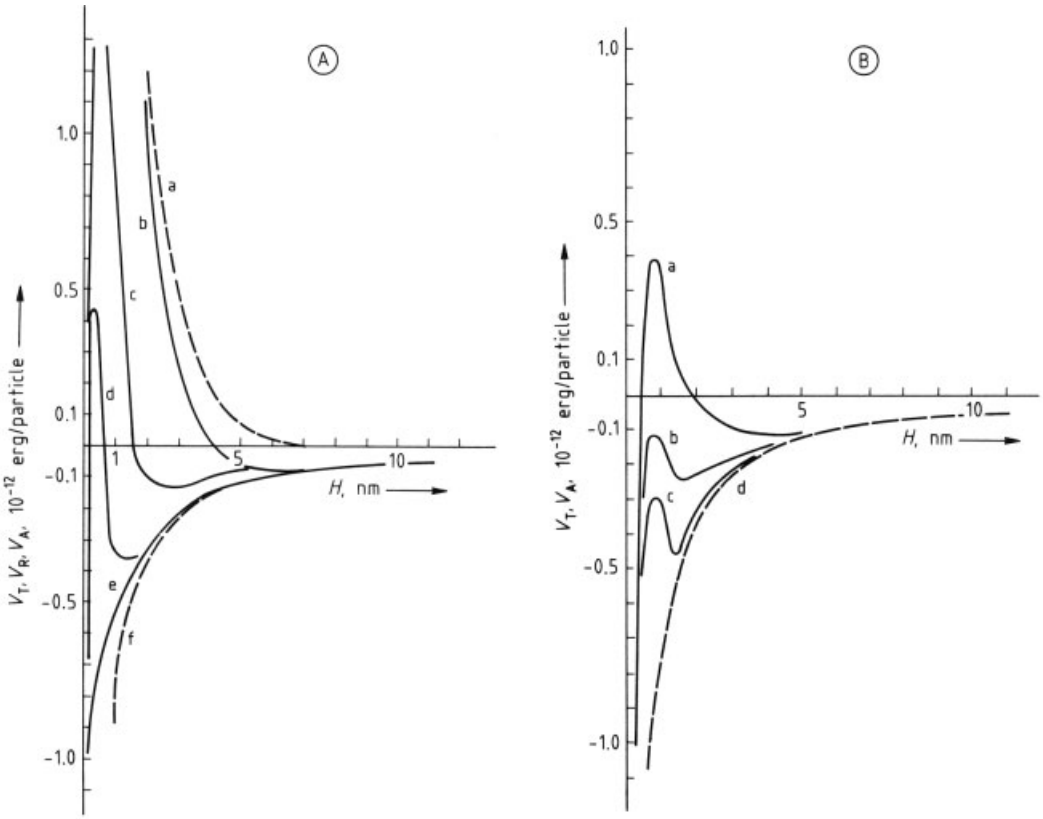

Figure 11. V_{T}, V_{R}, and V_{A} for latex dispersions at different concentrations of electrolyte ($\nu^+ = \nu^- = 1$) and surface potentials z
A) $z = 2$ ($\psi_0 = 51.4$ mV), Hamaker constant $A = 10^{-20}$ J; $\varkappa$ in cm^{-1}
a) V_{R}, $\varkappa = 10^7$ (0.1 mol/L);
b) V_{T}, $\varkappa = 10^7$ (0.1 mol/L);
c) V_{T}, $\varkappa = 1.80 \times 10^7$ (0.3 mol/L);
d) V_{T}, $\varkappa = 3.16 \times 10^7$ (0.9 mol/L);
e) V_{T}, $\varkappa = 4.66 \times 10^7$ (2 mol/L);
f) V_{A}
B) $z = 1$ ($\psi_0 = 25.7$ mV), Hamaker constant $A = 10^{-20}$ J;
a) V_{T}, $\varkappa = 10^7$ (0.1 mol/L);
b) V_{T}, $\varkappa = 1.47 \times 10^7$ (0.2 mol/L);
c) V_{T}, $\varkappa = 1.80 \times 10^7$ (0.3 mol/L);
d) V_{A}

of $z = 1$ and $z = 2$. As long as a maximum V_{m} separates the secondary and primary minima, the colloidal dispersion is stable. When V_{m} is reduced below ca. 10 kT, the particles aggregate and the dispersion becomes unstable. The salt concentration that initiates this process is the critical coagulation concentration c_{K}.

If coagulation really occurred at the primary minimum, it would be impossible to repeptize the colloid by removing the salt. However, the particles rarely come into direct contact as required for coagulation at the primary minimum. Instead, two or more water layers remain between the particles, so that the "distance of closest approach" is larger and the coagulated particles do not assume the high negative interaction energies of the primary minimum. Figure 12 shows that repeptization is then possible [97], [104].

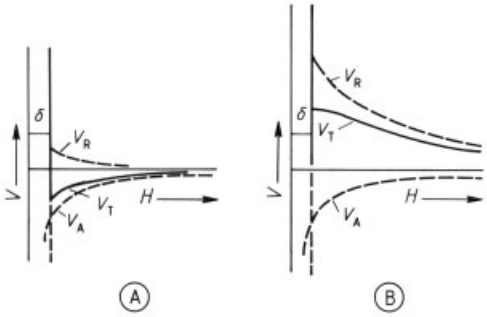

Figure 12. Energy diagram leading to coagulation (A) and to repeptization (B) [97]

The above formulas refer to spherical particles. For platelike particles the energy of repulsion can be calculated from the following expression

$$V_{\mathrm{R}} = 32\varepsilon\varepsilon_0 \left(RT/(vF)\right)^2 \varkappa\gamma^2 e^{-2\varkappa d} \ \mathrm{J/m^2} \tag{10}$$

The energy of van der Waals attraction is obtained from

$$V_{\mathrm{A}} = -\frac{A}{48\pi}\left[\frac{1}{d^2} + \frac{1}{(d+\Delta)^2} - \frac{2}{(d+\Delta/2)^2}\right] \tag{11}$$

(d, Δ see Fig. 7) [57], [96–99]. The calculation procedure for clay minerals is described in detail by VAN OLPHEN [62]. See also [105–108].

Stability Ratio. Whether coarsening of colloidal dispersions by coagulation occurs rapidly, slowly, or hardly at all depends on the forces between the colloidal particles. In the absence of repulsion forces, the frequency of Brownian encounters determines the rate of coagulation, which is the maximum rate possible (rapid coagulation). Repulsion forces reduce the rate. The maximum V_{m} of the interaction curve (Fig. 10) acts as an activation energy that decreases the ratio of coagulation [97]. The stability ratio (stability factor) W is defined as the ratio between the rate constant (k_0) in the absence of repulsion forces and in their presence (k):

$$W = k_0/k \tag{12}$$

When repulsion forces are absent, $k = k_0$ and $W = 1$; then rapid coagulation occurs. Increasing repulsion decreases the frequency of particle collisions (decreasing k), and W increases. For the influence of the hydrodynamic effect, see [8, Chap. 2], [109], [110]. Colloidal dispersions are considered stable when $W \geq 10^5$. This corresponds to a maximum of $V_{\mathrm{m}} \gtrsim 15\,\varkappa\mathrm{T}$ [96, p. 167]. The dependence of W on the energy of interaction of two spherical particles (in water, 25 °C) is given by [111]:

$$\mathrm{d}\log W/\mathrm{d}\log c = -2.15 \times 10^7 a\gamma^2/v^2 \tag{13}$$

(a: particle radius in cm). Equation 13 is very useful because the effective surface potential ψ_0 or ψ_δ (see Section 6) can be derived easily without knowledge of the van der Waals interaction parameter (Hamaker constant). The dependence of W on the salt concentration can be measured by a simple photometric procedure [111]. When salt is added to a colloidal dispersion, the optical density ($D = \log I_0/I$) increases, as shown schematically in Figure 13. The increase of D with time at the very early states of coagulation (initial linear section of the curves in Figure 13 A) is related to W:

$$\log\left(\mathrm{d}t/\mathrm{d}D\right)_{\mathrm{c}} = \log W - K \tag{14}$$

(K = constant). The values of $\log W$ as a function of $\log c$ are obtained when $\log(\mathrm{d}t/\mathrm{d}D)$ is plotted against $\log c$ (Fig. 13 B, C). This method is widely used for measuring the stability ratio of colloidal dispersions [8, Chap. 11].

Critical Coagulation Concentration. The critical coagulation concentration c_{K} is of fundamental importance in colloid chemistry. It can be obtained directly from a series of V_{T} curves constructed for different salt concentrations. A useful approximation for spherical particles is [96], [100]:

$$c_{\mathrm{K}} = 49.63\,(4\pi\varepsilon\varepsilon_0)^3\,(RT)^5\,\gamma^4/\left(F^6 A^2 v^6\right) \ \mathrm{mol/L} \tag{15a}$$

A = Hamaker constant (in J); at 25 °C, $\varepsilon = 78.5$:

$$c_{\mathrm{K}} = 3.837 \times 10^{-39}\gamma^4/\left(A^2 v^6\right) \ \mathrm{mol/L} \tag{15b}$$

Corrections that may be introduced are discussed in [100]. For platelike particles, the expression is

$$c_{\mathrm{K}} = 8.308 \times 10^{-39}\gamma^4/\left(A^2 v^6\right) \ \mathrm{mol/L} \tag{15c}$$

As discussed before (see page 5) , c_{K} is highly dependent on the valency v (Schulze – Hardy

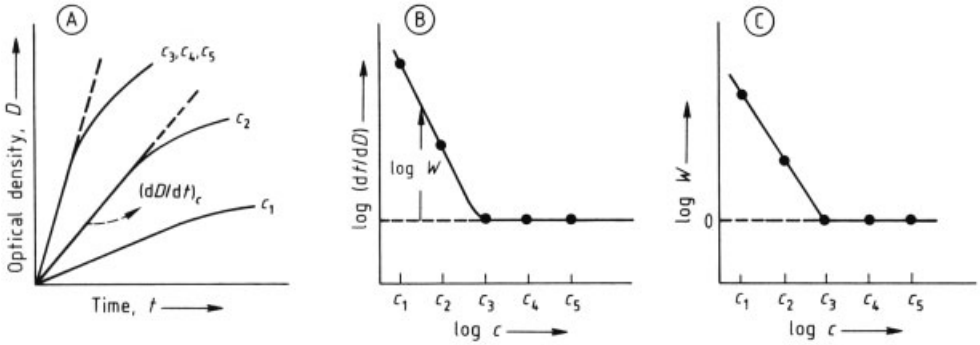

Figure 13. Experimental determination of the stability ratio W (D optical density)
A) Optical density as a function of time at different salt concentrations; B) and C) log (dt/dD) (from A) and log W (from B) as a function of salt concentration

rule) [8, Chap. 3]. Equation 15 a – c may give the impression that c_K varies with v^{-6}. This is true for high surface potentials ($\gamma \approx 1$), but does not hold at lower surface potentials because v is also included in γ. The quantity c_K then changes with v^{-5}, v^{-3}, or v^{-2} at potentials of 100, 50, and 25 mV, respectively [8, Chap. 3]. Critical coagulation concentrations are generally measured by two methods [8, Chap. 11]. The test-tube method is simple but sensitive. Increasing amounts of salt are added to a certain amount of the colloidal dispersion in a series of test tubes. The tubes are shaken and then allowed to stand for a certain time, usually 24 h. By visual inspection it is determined which dispersions are stable and which are coagulated. The particle concentration must be kept low so that formation of voluminous networks is avoided, concentrations below 1 wt % are generally required.

When the stability factor is determined as described in page 16 and Figure 13, the critical concentration is obtained immediately as the lowest salt concentration at which log W becomes zero or has a minimum. The stability of colloidal dispersions should be studied by both methods because the experiments are based on different time scales. The test-tube investigations reveal the changes that occur over several hours, whereas the photometric turbidity measurements record the initial stages of coagulation. Thus, different critical concentrations are determined in complex systems by the two methods. The differences can be helpful in understanding different coagulation processes. The test-tube method should not be considered obsolete because of the lack of electronic equipment. It is particularly recommended for testing industrial products because it responds sensitively to many experimental parameters and stability changes can be detected over long time periods by visual inspection. When deviations from "ideal" behavior occur, the test-tube tests provide information that may not be noticed by other methods.

Steric Stabilization. The stabilizing action of polymers is significant in many aqueous and nonaqueous systems. The principles of steric stabilization are well understood [8, Chap. 4], [98], [112]. If large molecules or polymers are absorbed on or bound to the particle surface (Fig. 14), a repulsion between the particles is created. When two particles with their envelopes of organic chains approach, the chains in the gap between the particles lose conformational entropy and resist a further approach (volume restriction effect or entropic stabilization, Fig. 14 A). In any case, there is also a local increase in concentration of organic material between the particles (region of the dotted line, Fig. 14 B) and an osmotic effect is produced that exerts a repulsive force between the particles (for terminology see [112]).

The interaction energy due to steric stabilization can be easily calculated by pragmatic formulae [113]. However, it is difficult to find the correct input data, for instance, the segment density distribution of the macromolecules or the contour length of the macromolecular envelope.

In several systems, particularly biological systems, a combination of electrostatic

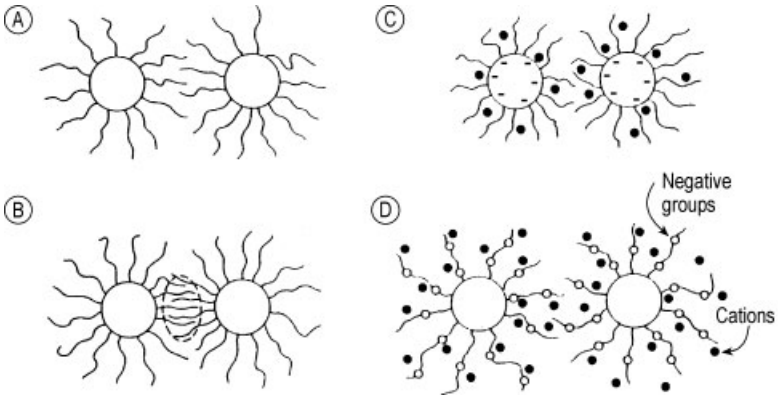

Figure 14. Steric stabilization
A) Entropic (volume restriction); B) Osmotic; C), D) Electrosteric

and steric stabilization occurs when particles carry surface charges or when the adherent macromolecules are polycations or polyanions (Fig. 14 C, D).

Steric stabilization has ben used by people since ancient times (egyptian "instant" ink, paints, and peptizing agents for clays). Its practical application has been exploited continuously. Industrial applications now comprise polymer dispersions, including latices, emulsions of all types, paints, inks, oils, lubricants, laundry products, and many others.

Typical requirements for steric stabilization are as follows:

1) Steric stabilization requires threshold concentrations of polymers. At low concentrations, sensibilization occurs (see page 27).
2) The amounts of absorbed polymer required for stabilization decrease with the molecular mass of the polymer.
3) The macromolecules should be firmly bound on the surface.
4) The interaction between the macromolecules and the solvent molecules is very important, and the polymer – solvent interaction parameters play a decisive role. Stabilization increases with the quality of the solvent. In a good solvent, interpenetration of the chains is disfavored, promoting stabilization. Conversely, interpenetration is thermodynamically favored in a worse-than-θ solvent, and the dispersion is destabilized. Figure 15 illustrates how sensitively stability responds to the polymer – solvent interaction parameter α.
5) Under θ conditions the system becomes unstable within a narrow temperature range.
6) Change of temperature can favor or disfavor steric stabilization (cf. Table 3).
7) In most cases, destabilization can only be obtained by changing the temperature or the quality of the solvent (changing χ or α) so that θ or worse-than-θ states are established.
8) Destabilization is generally reversible; changing the temperature or the quality of the solvent leads to redispersion and stabilization.

Steric stabilization is widely exploited in industrial processes [115] because of the following advantages:

1) Sterically stabilized dispersions are relatively insensitive to electrolytes.
2) Flocculation (destabilization) is generally reversible; simple dilution with an appropriate solvent is often sufficient to induce the particles to redisperse spontaneously. The particles also redisperse after drying.
3) Steric stabilization at high concentration of solid materials does not produce a comparable increase in viscosity, as occurs in electrostatic stabilization. In the latter case, gels of considerable strength are often formed.
4) Whereas electrostatic stabilization is less effective in nonaqueous dispersion media, steric stabilization can be as effective in aqueous as in non-aqueous media. Steric sta-

bilization is therefore the preferred method of stabilization in organic solvents.

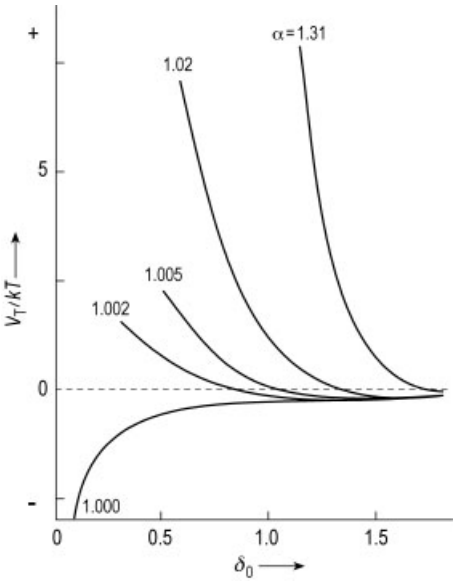

Figure 15. Steric interactions as a function of the solvent parameter α [114]. δ_0 is the normalized distance $H/(r^2)^{1/2}$ where $(r^2)^{1/2}$ is the root-mean-square end-to-end length of the polymer coil in solution.

Steric stabilizers for industrial products are found empirically. Amphipatic block or graft copolymers are widely used as steric stabilizers. They are made up of monomer units that anchor the polymer to the particle and other units that provide stabilization. The anchor unit is one that would not be soluble in the dispersion medium and that has affinity for the particle surface. The stabilizing unit is one that would be soluble in the medium and that imparts mutual repulsion to the dispersed particles. For example, aqueous dispersions can be stabilized by styrene/ethylene oxide copolymers in which the styrene units act as the anchor and the ethylene oxide units provide stabilization. Conversely, a non-aqueous dispersion is stabilized by styrene/acrylonitrile copolymers in which the acrylonitrile units provide the anchor and the styrene units provide stabilization.

An important technique is the grafting of macromolecules to the particle surface by covalent bonds [116–118].

Depletion Effects. Dissolved, nonadsorbable macromolecules can destabilize colloidal dispersions by depletion effects [8, Chap. 4], [119], [120]. The surface of the dispersed particles is surrounded by a depletion zone (thickness Δ) in which the polymer concentration is smaller than in the bulk. The reason is that the reduced conformational entropy of nonadsorbable macromolecules at the surface is not compensated by the adsorption enthalpy. When the particle distance H becomes smaller than $2\,\Delta$ (Fig. 16), the macromolecules are displaced from between the particles. Solvent molecules also leave the interparticle space, so that H decreases, ΔG becomes more negative, and the colloidal dispersion is destabilized by phase separation or flocculation, although the solvent is a better-than-θ solvent for the dissolved macromolecules [121], [122]. The theory claims that in certain cases restabilization can occur at very high polymer concentrations [113], [119], [122]. The typical form of the potential curve is shown in Figure 16. The depletion interaction energy can be calculated by pragmatic formulae [113], [123], [124].

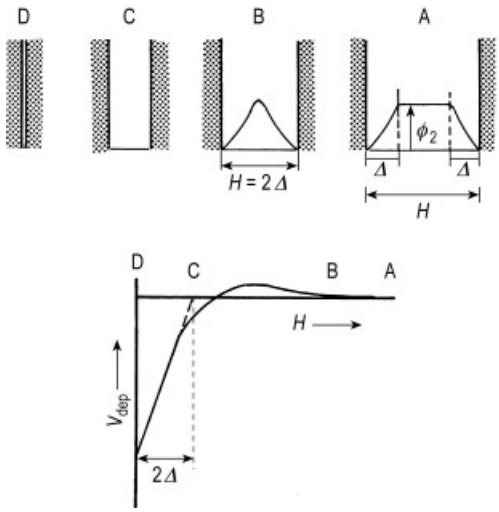

Figure 16. Depletion interaction. The macromolecules leave the space between two approaching particles when the distance H is smaller than $2\,\Delta$; the particle – particle interaction becomes attractive. ϕ_2 is the polymer concentration between the particles, Δ is the thickness of the depletion zone [113]; from [8]
The curve shows the depletion potential V_{dep} as a function of H.

Many systems, particularly practical dispersions like latex dispersions contain particles surrounded by macromolecules and dissolved poly-

mers, and both steric stabilization and depletion effects have to be considered. Also, electrostatic stabilization can be relevant. Stability diagrams are calculated from the corresponding equations by assuming the different interactions (electrostatic, van der Waals, steric, depletion) to be additive, which is not always the case. For examples of such calculations, see [125–128].

6. Factors Controlling Colloid Stability

Surface Charge. The imperative requirement for electrostatic stabilization is the presence of electric charges on the particle surface. There are several mechanisms for producing these charges:

1) In many silicates, including clay minerals and zeolites, isomorphous substitutions (most common is Al^{3+} for Si^{4+}) and cation vacancies produce negative surface charges. The surface charge density σ_0 is invariantly determined by the substitutions and vacancies and does not change with experimental conditions (concentration, pH, etc.).
2) Oxides like SiO_2, Al_2O_3, Fe_2O_3, and TiO_2 carry OH groups on the surface that can dissociate or adsorb protons. Silanol groups, for instance, are protonated at very low pH (< 2) and the surface is positively charged. With increasing pH, the surface charge density decreases and becomes increasingly negative above the point of zero charge (pzc) because of the dissociation of silanol groups (Fig. 17). Analogously, SH^- groups impart surface charges to colloidal sulfides. In both cases the surface charge density is not constant, but changes with pH.
3) Sparingly soluble salts such as silver iodide may have an excess of Ag^+ or I^- ions on the surface depending on the Ag^+ or I^- concentrations in the solution. (The concentrations of Ag^+ and I^- ions in solution are related to each other by the solubility product L: $p\,Ag + p\,I = p\,L = 16$). The silver and iodide ions act as potential-determining ions; thus AgI particles are positively charged at Ag^+ concentrations above the pzc (at $p\,Ag^0 = 6$) and are negatively charged at Ag^+ concentrations below $p\,Ag^0 = 6$.
4) Metal colloids are usually charged by absorption of ions, for instance Cl^- and citrate anions on Au and Pt particles ([8], [17]).
5) Commonly, latex particles carry $-SO_4H$, $-SO_3H$ or $-COOH$ groups on the surface. Sulfate and sulfonate groups dissociate the proton at low pH so that in pH ranges of practical applications the surface charge density is invariant with pH. The weakly acidic $-COOH$ groups dissociate at $pH \leq 4$, so that the surface charge increases considerably at $pH \leq 4$.

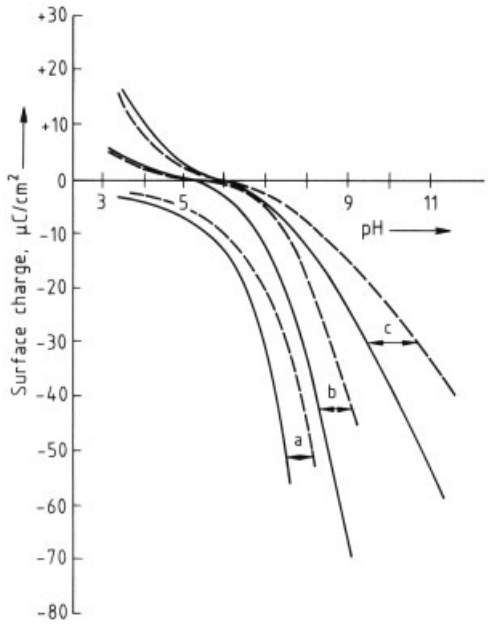

Figure 17. Variation of surface charge of oxides with pH in 0.1 mol/L (- - - -) and 1 mol/L KCl (–) solutions at 20 °C
a) Precipitated silica [129]; b) TiO_2 [130]; c) Hematite [130]

Consequences of Surface Charge Changes. Above several systems are described in which the surface charge density changes with experimental conditions (pH and concentration of potential-determining ions). When parameters are chosen so that the particle charge disappears (point of zero charge), the dispersion becomes highly unstable and coagulates. The stability factor has a sharp minimum (cf. Fig. 32 A) because the surface is recharged by crossing the pzc, and the system restabilizes. Thus, small pH changes near the pzc can have dramatic effects on the stability of colloidal dispersions.

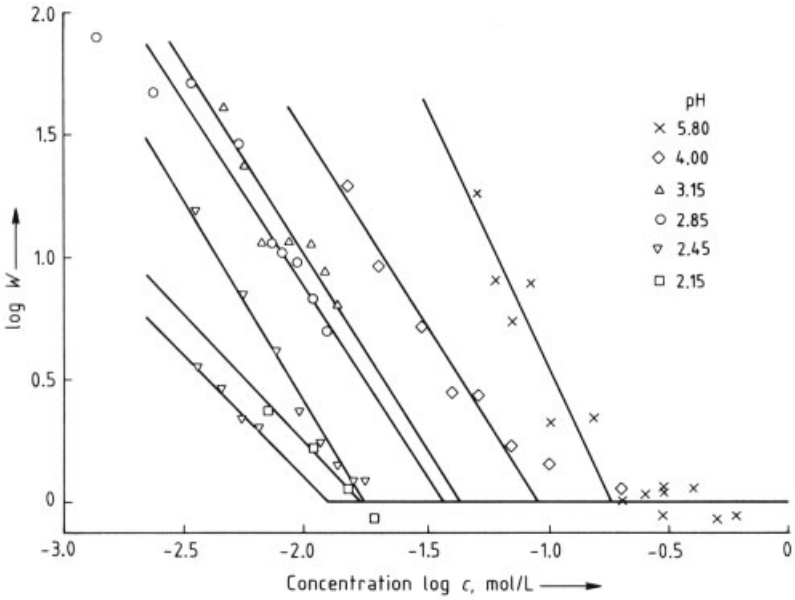

Figure 18. Stability ratio of PTFE latex as a function of pH in NaCl solutions [131]

The effect of pH changes on the stability of PTFE latex dispersions is shown in Figure 18. Because of the increased surface charge density the dispersion tolerates a tenfold higher salt concentration at pH = 5.8 than at pH < 2.5 before becoming destabilized.

Specific Adsorption. The simplest model, developed by Stern (1924), assumes that the number of ions adsorbed on a particle surface is restricted by their size (including their hydration or solvation shell) and that specific interactions between the counterions and the surface may occur [8, Chap. 9], [82], [132]. The aqueous part of the ionic layer is divided into a Stern layer and a diffuse or Gouy – Chapman layer. The "Stern surface" is drawn through the centers of the ions that are assumed to be adsorbed on the particle surface; the potential decreases linearly from ψ_0 at the particle surface to ψ_δ at the Stern surface and then exponentially in the diffuse layer (Fig. 19).

The thickness δ of the Stern layer is usually assumed to be about 0.4 nm (two water layers). Some generalizations can be made:

1) The existence of the Stern layer does not invalidate the expressions for the diffuse part of the double layer. The double layer interactions are determined by the Stern potential ψ_δ. Thus ψ_δ has to be used instead of ψ_0 in the Equations 7, 8, 10.
2) As adsorption in the Stern layer increases, ψ_δ is reduced and the interparticle repulsion is diminished. In the limiting case where the counterions are strongly absorbed, $\psi \approx 0$, and the system becomes unstable.
3) As the electrolyte concentration increases, adsorption in the Stern layer increases and reduces ψ_δ, even if specific interactions are absent.
4) ψ_δ varies only slightly with the energy of specific interactions until this energy is relatively large.
5) A study shows that, even in simple systems like AgI with potassium and barium counterions, Stern layer adsorption must be taken into consideration [133]. The specific adsorption energies are of the order of $4-6\,kT$ for potassium and $5-8\,kT$ for barium ions.

Several other models were developed to describe the behavior of the ions at the surface [8, Chap. 9], [82]. The triple-layer model is often used. On the basis of such models, equilibrium constants of the reactions of the surface groups were evaluated from potentiometric titrations. However, these values should be considered cautiously [132], [134].

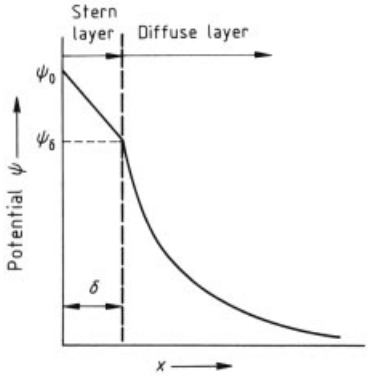

Figure 19. Stern layer and diffuse ionic layer around colloidal particles

Solvent Effects. Solvent parameters are of decisive importance for sterically stabilized systems (see page 17). This section describes the changes that occur when, under conditions of electrostatic stabilization, water is replaced by or mixed with polar organic solvents. A dependence of the critical coagulation concentration on ε is generally observed but does not follow the relation $c_K \sim \varepsilon^3$ as required by Equation (15) [135–137].

The principal effect of organic solvents (with $\varepsilon < 80$) is to press the counterions into the Stern layer, thus reducing ψ_δ and the stability of the dispersion toward salts [133]. The effect can be quite strong for multivalent counterions. In an extreme case, "equivalent coagulation" occurs [136]. When the bi- or trivalent counterions are added to the dispersion, the surface charges are neutralized, the particles become uncharged and the dispersion coagulates. The critical counterion concentration is then distinctly lower than that for coagulation by compression of the double layer and decreases with the concentration of the particles.

In some cases, high content of organic solvent suppresses the dissociation of the surface groups and the surface remains almost uncharged. Very low concentrations of monovalent counterions then coagulate the dispersion (≈ 0.001 mol/L for Ag and α-FeO(OH) particles in acetone [136]).

The effect of organic solvents is therefore not sufficiently described by changes in the dielectric constant. For instance, isopropanol and acetone affect the colloid stability more strongly than methanol and ethanol although their dielectric constants are not much lower. The effect of the organic solvent on the structure of the continuous phase, on the dissociation of surface groups, and on the solvation of the ions and their distribution between the Stern and the diffuse layer are important to the stability of dispersions in organic solvents or their mixtures with water.

Stabilization by charges can also occur in nonpolar solvents [8, Chap. 2]. As the concentration of charges is extremely low in such solvents, $\varkappa$ is very small (Eqs. 6 a – c) and the ionic layer very large, that is, the potential decreases very slowly with increasing distance from the surface. Therefore, high zeta potentials are often observed in such dispersions.

Effect of pH. Changes in pH can considerably influence the stability of colloidal dispersions when the surface charge density is pH-dependent. This is particularly true for oxides, sulfides, and latices with surface COOH groups (Fig. 18). The behavior of TiO_2 dispersions is shown as a further example (Fig. 21). The counterions are anions at pH below the zero point of charge, and cations at pH > 3. The above-mentioned influence of the ions of homologous series (see page 5) is clearly evident.

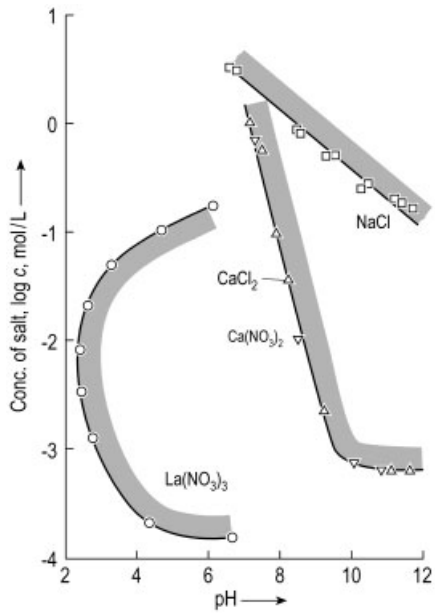

Figure 20. Dependence of the critical coagulation concentration c_K on pH for silica sol (Ludox sol, 0.2 g SiO_2 /100 mL) [138]

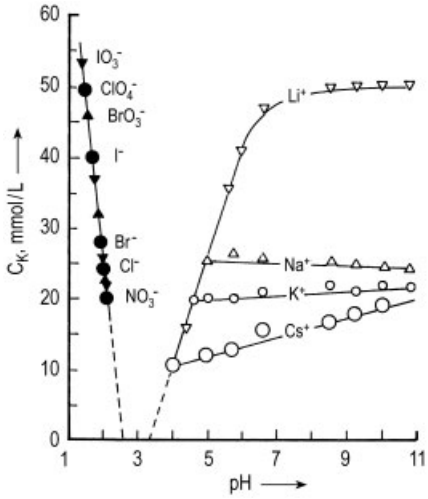

Figure 21. Specific ion effects for the coagulation of rutile (TiO_2) dispersions: the critical coagulation concentration c_K as a function of pH. At pH < 3 anions are the counterions; at pH > 3 cations are the counterions [139]; from [8]

The influence of pH on the stability of silica sols is striking (Fig. 20). The decreasing c_K for NaCl and $CaCl_2$ in the alkaline range, however, is opposite the behavior expected from Figure 17. This behavior is important in practical systems but is not understood completely. Deformation and decomposition of a hydration layer around the SiO_2 particles by the increased negative charge density is the most plausible explanation. Repulsive and attractive hydration forces are discussed in [8], [140]. The increase of the c_K values of the TiO_2 dispersions in the presence of alkali metal ions (Fig. 21) is not as strong as would be expected from Fig. 17, and this may be for similar reasons.

The flow behavior (viscosity, yield value) of dispersed oxides is also pH-dependent. Both properties often show maximum values at the point of zero charge (Fig. 22) [141] because the disappearance of the electrostatic repulsion promotes aggregation of the particles. This behavior is very important for practical applications. Sometimes, undesirable thickening or thinning of dispersions can be reduced by changing the pH. The viscosity and yield value at a given pH can also be decreased by adsorption of appropriate anions. Adsorption of polyacrylate, for instance, shifts the maximum yield value of ZrO_2

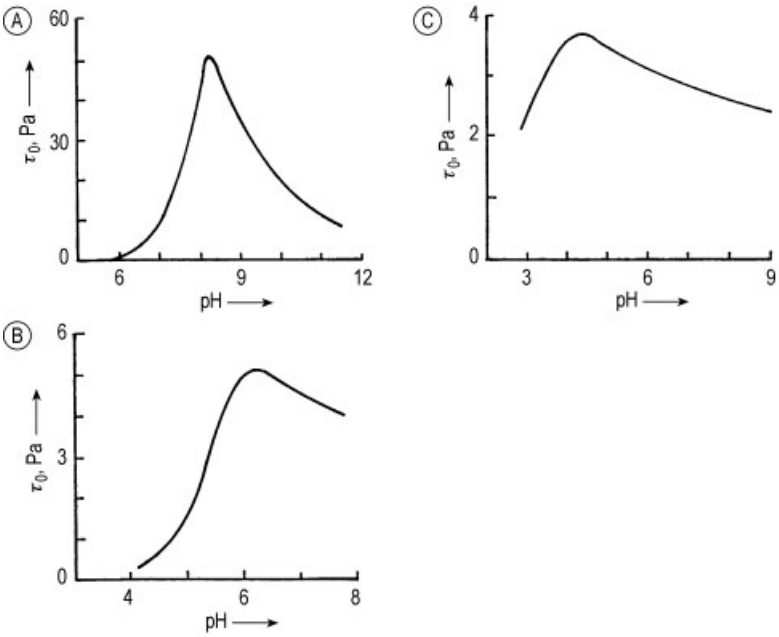

Figure 22. Yield value τ_0 of some dispersions as a function of pH at 25 °C
a) Alumina (Type C, Degussa, solid content 10 wt %). b) Titania (P25, Degussa, 7.5 wt %). c) Silica (Aerosil 200, Degussa, 8 wt %) [141]; from [8] pH $\triangleq$ pH in Fig.

dispersions to lower pH values (Fig. 23). Similar changes were observed with many other anions (phosphate, citrate, benzenecarboxylic acids, etc.) [142].

Multivalent Ions. Colloidal systems respond very sensitively to multivalent cations (Schulze – Hardy rule). Aluminum and iron ions are thus the preferable counterions for coagulating colloidal systems. Multivalent cations, in particular tri- and quadrivalent ions, can influence the colloid stability in a way different from the behavior expected from the Schulze – Hardy rule for $\nu = 3$ or 4:

1) First, aqueous solutions of these cations are acidic and change the pH of the system, so that pH effects as described can occur, in particular when the pH is near the pzc.
2) Multivalent hydrated cations condense to form larger complexes with a high charge, which decreases the stability. The stepwise decreasing critical concentration of aluminum sulfate and nitrate at pH 4 – 5 in AgI dispersions is explained on this basis [138].
3) The specific adsorption of counterions in the Stern layer is more pronounced for multivalent ions than for monovalent ones. In extreme cases, the surface charges are neutralized by the multivalent counterions so that the system reacts very sensitively to such ions ("equivalent coagulation" see page 23).
4) An excess of multivalent counterions on the surface recharges the surface:

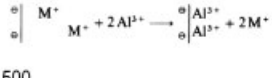

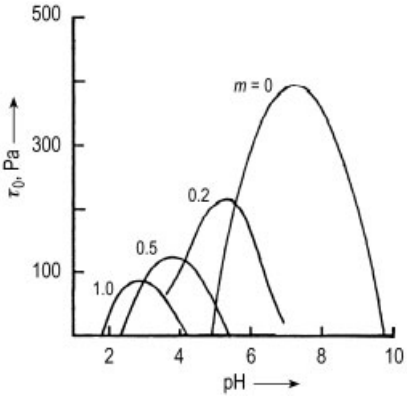

Figure 23. Yield value τ_0 of ZrO_2 dispersions in the presence of polyacrylic acid ($M = 2000$) as a function of pH. m = amount of polymer (g/100 g ZrO_2) [142]; from [8]

Thus when increasing amounts of multivalent ions are added to a dispersion, the system can pass from stabilization to destabilization (charge neutralization) and again to stabilization (recharging). Recharging and restabilization of SiO_2 sols with increasing concentrations of La^{3+} are perceptible in Figure 20. One might expect that recharging by aluminum salts is optimal at low pH when Al^{3+} ions are present. Figure 24, however, shows that recharging occurs at higher pH (pH 4 – 5, depending on the aluminum salt concentration). It is assumed that recharging is brought about by the adsorption of hydrolyzed metal ions [138]. At higher pH precipitation of the metal hydroxide is initiated. Surface precipitation usually occurs under conditions different from precipitation in homogeneous solution. As a consequence of surface reactions, the zeta potential can change in a complex manner (Fig. 25) [143].

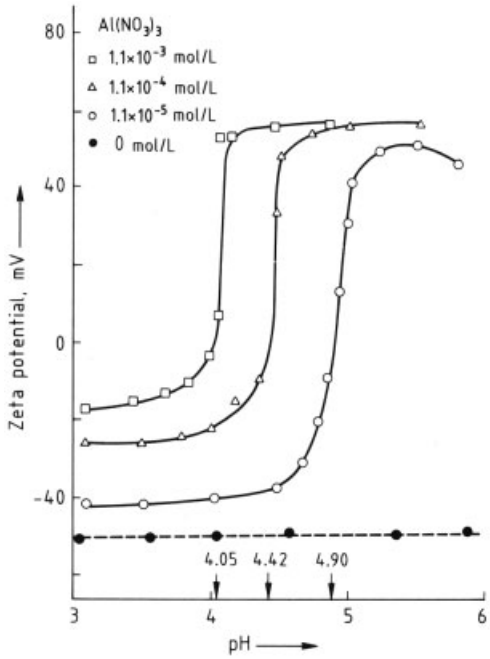

Figure 24. Zeta potential of PVC latex (0.003 wt %) as a function of pH in the absence and presence of $Al(NO_3)_3$ (Arrows indicate the isoelectric points. All systems contained 1×10^{-3} mol/L $NaNO_3$.) [138]

Multivalent anions (e.g., SO_4^{2-}, PO_4^{3-}, oxalate) are often strongly adsorbed by oxides or incorporated into the surface instead of oxygen or hydroxyl ions with considerable changes of the pzc.

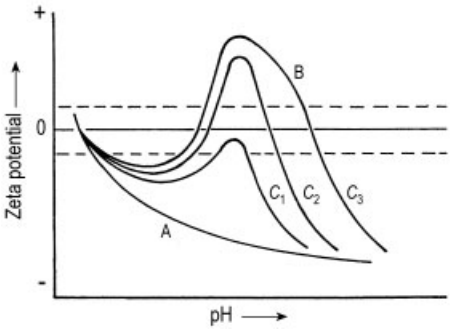

Figure 25. Dependence of zeta potential on pH in the presence of hydrolyzable multivalent metal cations of concentration $c_1 < c_2 < c_3$
A) Oxide in absence of metal cations; B) Colloidal metal hydroxide
The dispersion becomes unstable at zeta potentials lying between the dotted lines [143]; from [8]

Complexing Agents. Complexing of countercations by anions often causes antagonistic effects. Thus, fluoride and sulfate ions (at pH < 4) can decrease the coagulation power of aluminum ions (increasing c_K) [138]. An opposite effect is observed when the counterions are adsorbed as complexes. Addition of small amounts of complexing agents can then decrease the c_K by orders of magnitude [144].

Surface Active Agents. Surfactants with counterionic charge are strongly adsorbed at the particle surface. Charge neutralization ($\psi_\delta = 0$) with rapid destabilization of the colloidal system requires specific concentrations because recharching and restabilization can take place by adsorption of an excess of surfactant ions together with their counterions (Fig. 26). This process is promoted by the tendency of the long-chain ions to aggregate to monomolecular or bimolecular films. In the case of low surface charge densities, the surfactants form small surface aggregates [145]. Adsorption of surfactant counterions on oxidic surfaces is often accompanied by charge regulation. For example, after compensation of positive surface charges by surfactant anions, adsorption of further anionic surfactants causes the adsorption of protons by surface groups, so that the density of the positive surface charges increases and the zero point of charge increases with increasing pH [167].

The charge reversal by adsorption of alkylammonium ions on quartz is illustrated in Figure 27. Longer alkyl chains shift the pzc to lower surfactant concentrations by orders of magnitude. The effect of a particular surfactant depends on pH, which determines whether the surface of the oxide is positively or negatively charged. Thus, dodecyl sulfate causes charge reversal and restabilization of Al_2O_3 sols only at a pH below the pzc of pure Al_2O_3 sols (pH < 8) when the surface is positively charged. The importance of the way in which oxides are produced industrially can be exemplified by TiO_2. TiO_2 prepared in the absence of sulfate or phosphate ions adsorbs tetradecylpyridinium ions only at pH > 5. In the presence of sulfate ions, which are strongly held on the surface, the surfactant is adsorbed throughout the whole pH range, with a minimum at pH ≈ 6, and a complex surfactant layer is formed [146], [147].

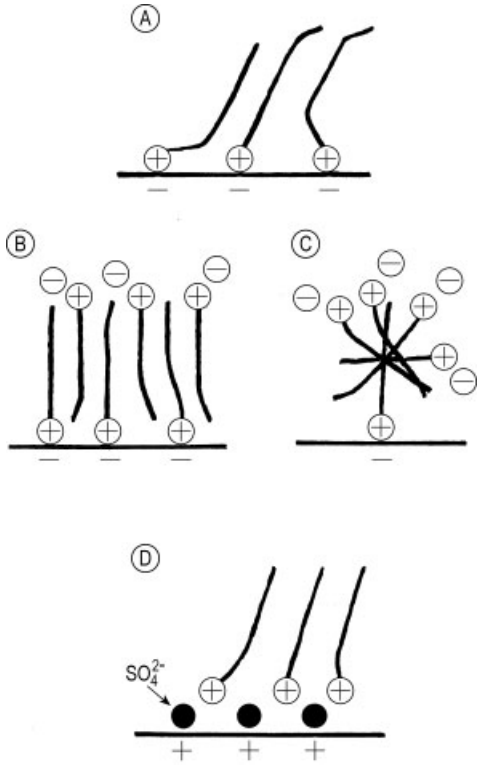

Figure 26. Adsorption of surfactant counterions at hydrophilic surfaces
A) charge neutralization; B), C) Recharging by an excess of surfactant counterions at high (B) and low (C) surface charge density; D) Adsorption of surfactant cations on the positively charged TiO_2 surface in the presence of sulfate ions [146], [147], [145].

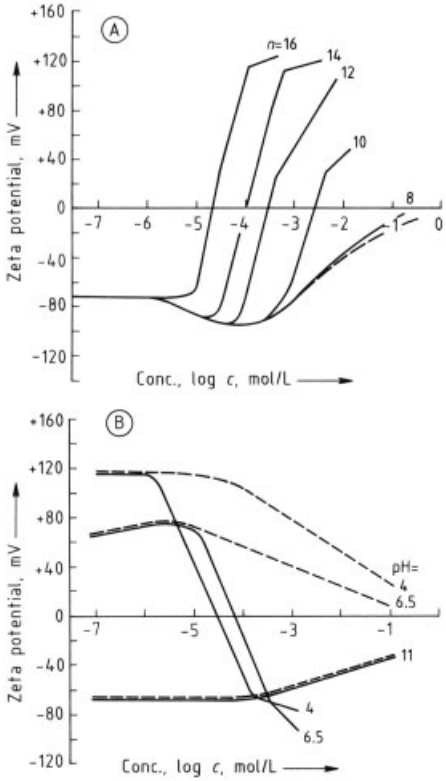

Figure 27. Influence of salts and surfactants on the zeta potential
A) Ammonium acetate (- - -) and alkylammonium acetates (–) of various chain lengths on quartz (n of alkylammonium ions $C_nH_{2n+1}NH_3^+$ as indicated) [148];
B) Sodium chloride (- - -) and dodecylsulfate (–) on corundum at different pH [148]

Polymers. Polymers influence the stability of colloid dispersions in different ways (Fig. 28) [8, Chap. 4]. Stabilization or destabilization of colloidally dispersed systems are used in many technical applications. Usually, low concentrations of macromolecules decrease the stability of colloidal dispersions (sensibilization) whereas higher concentrations stabilize the colloidal state, in many cases in extreme ways. The stability factor W as a function of polymer concentration proceeds through a minimum that indicates the highest degree of sensibilization (Fig. 29). Besides the stabilization effect, the adsorption of polymers can change the behavior of the systems to salts. PVA-stabilized silica responds sensitively to Na_2SO_4 and KCl but is not coagulated by $CaCl_2$ [149].

Sensibilization at low polymer concentration is brought about by macromolecules bridging between the particles or by charge neutralization when charged polymer ions displace the counterions from the surface (Fig. 30 A, B). Charge neutralization can also occur patchwise (Fig. 30 C). The charges are not compensated site by site, but the total charge of the particle becomes zero as a consequence of negative and positive domains. To bridge the particles, the macromolecular charge must not be opposite to the surface charge; it is only required that the macromolecule can attach itself at some sites at the surface so that larger parts of it remain free to form bridges between different particles. Practical applications prefer the sensibilization by particle bridging because charge neutralization requires a careful adjustment of parameters to prevent charge reversal and restabilization. Polyanions are therefore widely used as flocculating agents for clay materials [14], [150].

Formation of flocs by polymer addition can retard the adsorption equilibrium because solvent remains trapped within the flocs. The amounts adsorbed then depend on the manner of mixing, stirring, shaking, etc., and reproducibility of the measurement can be poor [151].

Stabilization at higher polymer concentrations is caused by steric stabilization. This process is often the only process that allows stabilization in nonaqueous solvents.

With highly anisometric particles the sensibilization region can be preceded by a stabilization region at very low polymer concentrations so that a sequence stabilization $\rightarrow$ sensibilization $\rightarrow$ stabilization is observed with increasing polymer concentration (polyacrylamide onto kaolinite [152]).

Mechanical Effects. In laboratory experiments colloid stability usually is measured under perikinetic conditions, that is, the particles collide owing to the Brownian motion of the solvent molecules. In practical applications, the dispersions are shaken, stirred, or transferred by pumping. The mechanical forces intensify the particle flux, which depends on the third power of the collision diameter of the particles [76]. Orthokinetic aggregation is therefore more important for larger than for smaller particles. Thus, orthokinetic destabilization can be very effective for emulsions.

Under orthokinetic conditions, hydrodynamic forces are in interplay with the forces bet-

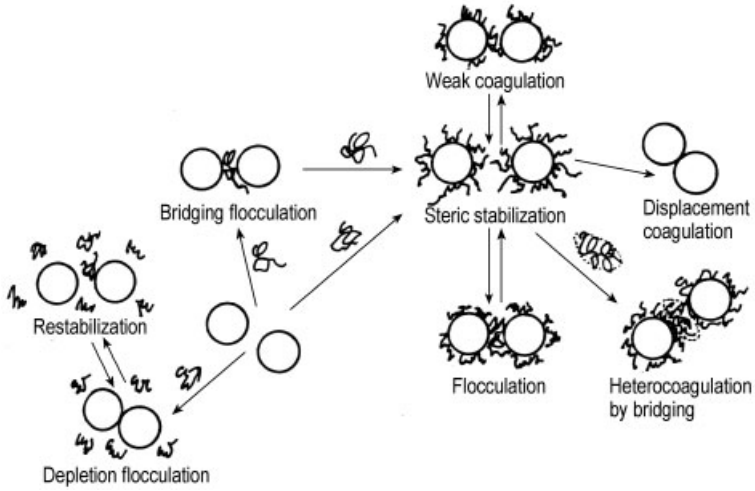

Figure 28. Stabilization and destabilization of colloidal dispersions by macromolecules [8]

Table 3. Effect of changing the stabilizing species on emulsion stability [158]

Type of emulsion	Stabilizing moiety ∗	Anchor polymer ∗	Dispersed phase	Dispersion medium	Flocculation on	Type of stabilization	Critical flocc. temp., K	θ-Temp., K
O/W	PEO	PVAc	toluene	0.39 M $MgSO_4$	heating	enthalpic	318	318
O/W	PVA	PVAc	toluene	2 M NaCl	heating	enthalpic	297	299
O/W	PAM	PVAc	toluene	2.1 M $(NH_4)_2SO_4$	cooling	entropic	297	298
O/W	PAA	PVAc	toluene	0.2 M HCl	cooling	entropic	298	287
W/O	PLM	PAA	water	*n*-pentanol	cooling	entropic	301	303
W/O	PS	PEO	water	cyclohexane	cooling	entropic	308	307

∗ PEO = Polyethylenoxide; PVA = Polyvinylalcohol; PAM = Polyacrylamide; PAA = Polyacrylic acid; PLM = Poly-(lauryl methacrylate); PS = Polystyrene; PVAc = Polyvinyl acetete

ween the colloidal particles [8, Chap. 3], [153–155]. It is difficult to predict the behavior of colloidal dispersions under shear. The stability and destability regions depend on the rate of shear, the particle size and size distribution, and the interparticle forces in a complicated way. Even small changes in parameters can strongly influence stability at low and high shear rates. A calculated diagram is shown in Figure 31. The dispersion is unstable ($\log W < 5$) at low but also at high shear rates. At low shear rates the dispersion coagulates in the secondary minimum (see Fig. 10), at high shear rates in the primary minimum [156]. In many technical processes coagulation occurs under turbulent flow but not much is known about this process [154].

Stability against shearing plays an important role in industrial processes such as coating and production of thin-layer ceramic films. Further, dispersions should not destabilize during shaking (transport), stirring, or pumping. Sometimes a dispersion is stable for years but coagulates in a short time on shearing. Because of the complicated dependence on the particle size, two latex dispersions with different particle sizes can be stable against shearing but the mixture can be unstable. To avoid undesirable shear coagulation the stabilization mechanisms, particle shape, size and size distribution and the chemical parameters (pH, salt content, presence of organic compounds) must be critically evaluated. Dispersions of oxides often coagulate under shear when pH is near the zero point of charge (see page 23).

Adsorbed stabilizer ions or macromolecules can be desorbed by shearing. It is known that emulgator molecules can be removed from the surface of latex particles by centrifugation [8, Chap. 6] which reduces the stability of the dispersions.

Shearing can change the structure of adsorbed macromolecules, and this can also reduce stability. Otsubo and Watanabe [157] observed, at low shear rates, a stepwise in-

crease of the viscosity of silica particles stabilized by adsorbed polyacrylamide. Depending on the molecular mass, the solvency of the solvent, the solid mass content, and the particle size, macromolecular bridges formed between colliding particles during shearing increased the viscosity.

Macromolecules which are desorbed by shearing can cause flocculation by depletion effects.

Temperature Effects. Temperature effects are most pronounced for sterically stabilized systems. It depends on the type of stabilization – mainly enthalpic or mainly entropic – whether a dispersion flocculates on cooling or heating. When enthalpy effects dominate, the system flocculates on heating. If enthropic effects are important, the sol flocculates on cooling. Examples of water-in-oil and oil-in-water emulsions are given in Table 3 [158]. Every sterically stabilized system flocculates under θ conditions and at the θ temperature (see page 17) [8, Chap. 4], [112].

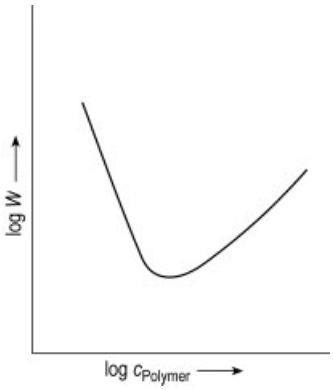

Figure 29. Variation of the stability ratio with the concentration of polymers

Electrically stabilized dispersions are not very sensitive to temperature provided structural changes do not occur. When such changes take place, the dispersion becomes sensitive to temperature and can destabilize within a narrow temperature range (thermosensitive latex dispersions). Temperature sensitivity can also be attained by choosing as counterions complex ions that decompose at higher temperature, for instance $[SiF_6]^{2-}$ counterions. Thermosensitive poly(*N*-isopropyl acrylamide) latices (NIPAM latices) were studied in detail [159], [160]. The van der Waals attraction increases when the particles shrink above 30 °C and reduce the stability of the dispersion [161].

7. Heterocoagulation

Colloidally dispersed oxides can be separated by choosing appropriate pH values. For example, in quartz-rutile dispersions, both types of particles are negatively charged at pH $\approx$ 9.5 but can be separated by making the dispersion 0.015 – 0.050 mol/L in NaCl [162].

If the different particles carry opposite surface charges, heterocoagulation occurs in specific pH ranges. This is illustrated for $TiO_2 - Al_2O_3$ dispersions in Figure 32 A. At any pH between the pzc of TiO_2 and Al_2O_3, the TiO_2 particles are negatively charged whereas the Al_2O_3 particles are positive, and heterocoagulation takes place. Thus, TiO_2 can be selectively separated from mixed dispersions only at a pH below the pzc of TiO_2 and, correspondingly, Al_2O_3 above the pzc of Al_2O_3. Experimental observations (Fig. 32 B, [163]) confirm the expected behavior, but only if coagulation is initiated immediately after the TiO_2 and Al_2O_3 dispersions are mixed. If the mixed dispersion is allowed to stand before coagulation, it behaves like a pure Al_2O_3 dispersion. Soluble aluminum species are adsorbed by the TiO_2 particles, so that the particles become coated by Al_2O_3 and behave like Al_2O_3 particles. This example demonstrates that the stability of a colloidal dispersion of two or more different particles can be considerably affected by the adsorption of soluble species.

When aluminum nitrate is added to a PVC latex, distinct regions of stabilization and destabilization are observed (Fig. 33, [164], [165]). Since the anionic PVC latex has an isoelectric point (iep) $\approx$ 3, domains a and g represent stable negative sols of PVC (a) and PVC + Al_2O_3 (g). In domain b the latex particles are coagulated by Al^{3+} ions. In region f, the latex particles absorb hydrolyzed aluminum ions, become positively charged (recharging, cf. Fig. 24), and form a stable dispersion. In region c, PVC and Al_2O_3 coprecipitate, probably because recharging of the PVC latex ceases (cf. [143]) so that heterocoagulation occurs between negative PVC and positive Al_2O_3 particles.

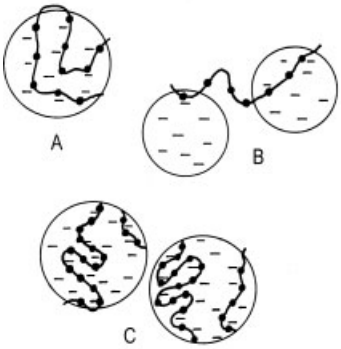

Figure 30. Destabilization of colloidal dispersions by macromolecules
A) Flocculation by charge compensation; B) Bridging flocculation; C) Flocculation by charge pattern interactions (mosaic model). The domains with adsorbed polycations receive a positive excess charge, which is compensated by the negative charge of the uncovered surface patches

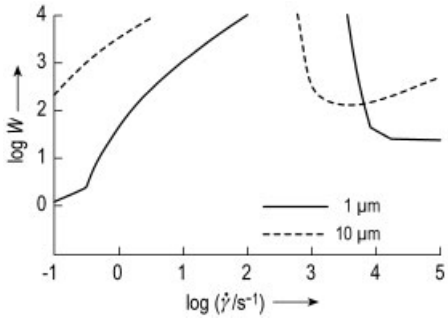

Figure 31. Influence of shearing on colloid stability. Stability factor W versus rate of shear for particles of 1 and 10 mm diameter. Stern potential 25 mV, $\varkappa = 10^8\ \mathrm{m}^{-1}$, Hamaker constant $A = 10^{-20}$ J [156]; from [8]

Heterocoagulation has also been studied with amphoteric latices [166]. By variation of the ratio of $-COOH$ and $-NH_2$ groups, latices with a different pzc can be prepared. The stability ratio of mixtures of latices with a different pzc follows the theoretical prediction. The behavior of negative latices in the presence of negative amphoteric latices toward salts exhibits some interesting features. The negative latex appears to be more easily coagulated in the presence of a negative amphoteric sol.

In practical applications heterocoagulation can occur during mixing of two different types of dispersions. To obtain a stable mixture, one kind of particle must be recharged by addition of suitable compounds like trivalent metal ions, complexed metal ions, or ionic surfactants.

A novel application of heterocoagulates is the stabilization of oil-in-water emulsions by clay minerals and layered double hydroxides without any surfactant emulsifier [47], [48], [94].

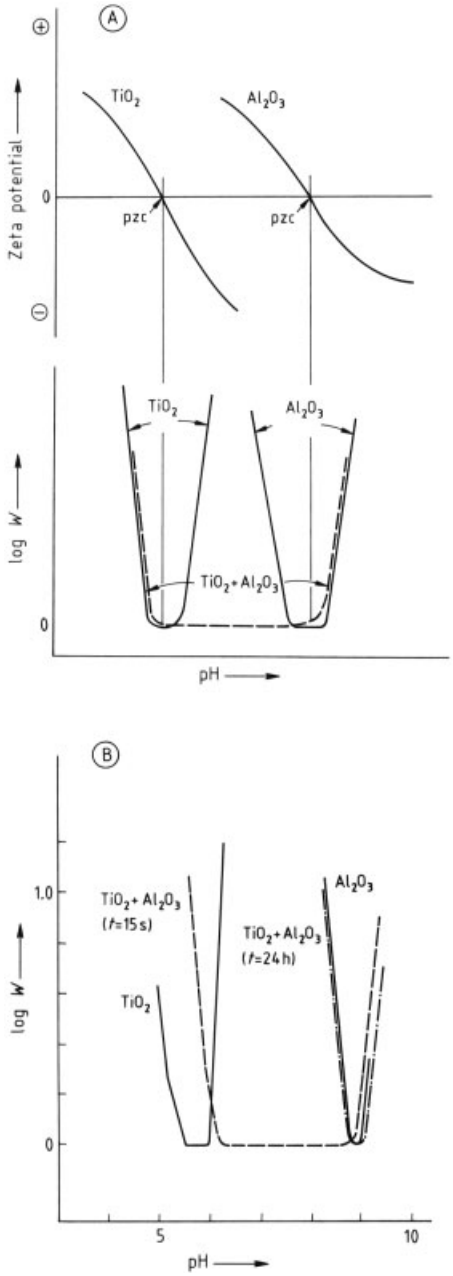

Figure 32. Coagulation of $Al_2O_3-TiO_2$ dispersions
A) Variation of surface potential and stability ratio with pH, schematically (–pure sols, – – – $Al_2O_3-TiO_2$ sol); B) Experimental results in 10^{-4} mol/L KNO_3 at 15 s and 24 h after mixing [163]

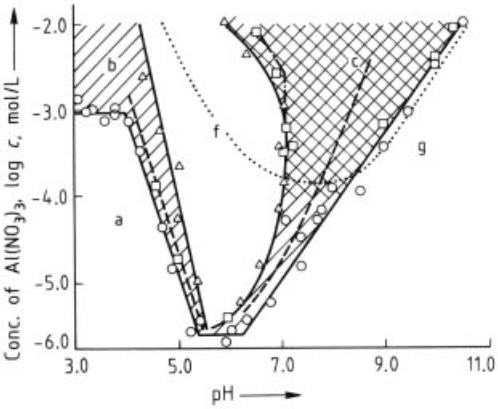

Figure 33. Stability diagram of a PVC latex (0.005 g/100 ml) in presence of $Al(NO)_3$; (- - -) variation of point of zero charge; (· · · ·) precipitation boundary of aluminum hydroxide; a) – g) see text [165]

8. References

Text books see:
[4], [5], [112], [120] (macromolecules); [2], [3], [8], [9], [14], [53], [76], [82], [96], [102], [112], [140] (colloids); [14], [62] (clay minerals).

1. R. W. Siegel, *Spektr. Wiss.* 1997 62 – 67.
2. H.-D. Dörfler: *Grenzflächen- und Kolloidchemie,* VCH Verlagsgesellschaft, Weinheim 1994.
3. M. J. Schwuger: *Lehrbuch der Grenzflächenchemie,* G. Thieme, Stuttgart 1996.
4. H. G. Elias: *Makromoleküle,* Hüthig Verlag, Basel, Heidelberg 1990 (vol.1), 1992 (vol. 2).
5. B. Vollmert: *Polymer chemistry,* Springer Verlag Berlin 1973.
6. J. Th. G. Overbeek, *Pure Appl. Chem.* **46** (1976) 91 – 101.
7. G. Ebert, *Top. in Curr. Chem.* **128** (1985) 1 – 36.
8. G. Lagaly, O. Schulz, R. Zimehl: *Dispersionen und Emulsionen,* Steinkopff Verlag, Darmstadt 1997.
9. R. K. Iler: *The chemistry of silica,* Wiley-Interscience, New York 1979.
10. R. D. Vold, M. J. Vold: *Colloid and Interface Chemistry,* Addison-Wesley Publ. Co., London 1983.
11. I. E. Odom in L. Fowden, R. M. Barrer, P. B. Thinker (eds.): *Clay minerals, their structure, behaviour and use,* The Royal Society, London 1984, pp. 171 – 190. *(Philos. Trans. R. Soc. London Ser. A* **311** (1984) 391 – 409)
12. W. B. Jepson in L. Fowden, R. M. Barrer, P. B. Thinker (eds.): *Clay minerals, their structure, behaviour and use,* The Royal Society, London 1984, pp. 191 – 212; (Philos. Trans. R. Soc. London Ser. A 311 (1984) 411 – 432).
13. T. R. Jones, *Clay Miner.* **18** (1983) 399 – 410.
14. K. Jasmund, G. Lagaly: *Tonminerale und Tone,* Steinkopff Verlag, Darmstadt 1993.
15. T. Kendall (ed.): *Industrial Minerals,* Industrial Minerals Information Ltd. 1996.
16. M. J. Schwuger, *Ber. Bunsenges. Phys. Chem.* **88** (1984) 1123 – 1137.
17. G. Frens, *Nature (London) Phys. Sci.* **241** (1973) 20 – 22.
18. H. Weller, *Angew. Chem.* **105** (1993) 43 – 55.
19. L. Brus, *Curr. Opinion Colloid Interface Sci.* **1** (1996) 197 – 201.
20. H. Weller et al., *Chem. Phys. Let.* **124** (1986) 557 – 560.
21. D. Distler (ed.): *Wäßrige Polymerdispersionen,* Wiley-VCH, Weinheim 1998.
22. V. Privman, D. V. Goia, J. Park, E. Matijević, *J. Colloid Interface Sci.* **213** (1996) 36 – 45.
23. D. Petters, *J. Mater. Chem.* **6** (1996) 1605 – 1608.
24. J. Belloni, *Curr. Opinion Colloid Interface Sci.* **1** (1996) 184 – 196.
25. E. Matijević, *J. Colloid Interface Sci.* **58** (1977) 374 – 389.
26. E. Matijević, *Pure Appl. Chem.* **60** (1988) 1479 – 1491.
27. E. Matijević, *Materials* **5** (1993) 412 – 426.
28. E. Matijević, *Curr. Opinion Colloid Interface Sci.* **1** (1996) 176 – 183.
29. E. S. Daniels, E. D. Sudol, M. S. El-Asser: "Polymer latexes. Preparation, Characterization and Applications," *ACS Symp. Ser.* **492** (1992).
30. S. Y. Chang, L. Liu, S. A. Asher, *J. Amer. Chem. Soc.* **116** (1994) 6739 – 6747.
31. F. A. Arriagada, K. Osseo-Asare, *J. Colloid Interface Sci.* **211** (1999) 210 – 220.
32. G. D. Rees, R. Evans-Gowing, S. J. Hammond, B. H. Robinson, *Langmuir* **15** (1999) 1993 – 2002.
33. M. P. Pileni, *Ber. Bunsenges. Phys. Chem.* **101** (1997) 1578 – 1587.
34. H. N. Stein, *Curr. Opinion Colloid Interface Sci.* **2** (1997) 165 – 170.
35. M. Antonietti, F. Gröhn, J. Hartmann, L. Bronstein, *Angew. Chem.* **109** (1997) 2170 – 2173.
36. S. Klingelhöfer, W. Heitz, A. Greiner, S. Oestreich, S. Förster, M. Antonietti, *J. Amer. Chem. Soc.* **119** (1997) 10116 – 10120.

37. G. Carrot et al., *Colloid Polymer Sci.* **276** (1998) 853 – 859.
38. R. A. Caruso, M. Giersig, F. Willig, M. Antonietti, *Langmuir* **14** (1998) 6333 – 6336.
39. H. Cölfen, M. Antonietti, *Langmuir* **14** (1998) 582 – 582.
40. I. Dékány et al., *Langmuir* **12** (1996) 3709 – 3715.
41. I. Dékány, L. Turi, Z. Király, *Appl. Clay Sci.* **15** (1999) 221 – 239.
42. G. Hornyak et al., *Chem. Eur. J.* **3** (1997) 1951 – 1956.
43. P. H. Nadeau, *Clay Miner.* **20** (1985) 499 – 514.
44. H. Vali, L. Bachmann, *J. Colloid Interface Sci.* **126** (1988) 278 – 291.
45. W. Jahn, R. Strey, *J. Phys. Chem.* **92** (1988) 2294 – 2301.
46. J. Thieme et al., *Progr. Colliod Polymer Sci.* **95** (1994) 135 – 138.
47. J. Thieme, S. Abend, G. Lagaly, *Colloid Polymer Sci.* **277** (1999) 257 – 260.
48. U. Neuhäusler, S. Abend, C. Jacobsen, G. Lagaly, *Colloid Polymer Sci.* **277** (1999) 719 – 726.
49. J. Thieme, J. Niemeyer, *Geol. Rundschau* **85** (1996) 852 – 856.
50. B. R. Jennings, K. Parslow, *Proc. Royal Soc. London* **A419** (1988) 137 – 149.
51. B. R. Jennings, *Clay Min.* **28** (1993) 485 – 494.
52. T. Allen: *Particle Size Measurements,* **vols. 1 + 2,** Chapman and Hall, London 1997.
53. R. J. Hunter: *Introduction to Modern Colloid Science,* Oxford Sci. Publ, 1993; R. J. Hunter: *Foundations of Colloid Science,* **vols. I + II,** Clarendon Press, Oxford, 1991.
54. T. P. Provder (ed): "Particle Size Distribution II," *ACS Symp. Ser.* **472** (1991).
55. P. A. Smith, R. A. Haber, *J. Am. Ceram. Soc.* **78** (1995) 1737 – 1744.
56. R. Greenwood, P. L. Luckham, T. Gregory, *J. Colloid Interface Sci.* **191** (1997) 11 – 21.
57. P. C. Hiemenz: *Principles of colloid and surface chemistry.* Marcel Dekker, New York 1977.
58. M. Filella, J. Zhang, M.-E. Newman, J. Buffle, *Colloids Surfaces A* **120** (1997) 27 – 46.
59. G. Lagaly in W.-M. Kulicke (ed.): "Anorganische Systeme," *Fließverhalten von Stoffen und Stoffgemischen,* Hüthig Verlag, Basel 1986, pp. 147 – 186.
60. F. Aldinger, H.-J. Kalz, *Angew. Chem.* **9** (1987) 381 – 391.
61. J. Böhnlein-Mauß et al., *Adv. Mater.* **4** (1992) 73 – 81.
62. H. van Olphen: *An introduction to clay colloid chemistry,* J. Wiley and Sons, New York 1977.
63. H. C. H. Darley, G. R. Gray: *Composition and Properties of Drilling and Completion Fluids,* Gulf Pub. Comp., Houston 1991.
64. R. H. Ottewill, *J. Colloid Interface Sci.* **58** (1977) 357 – 373. *Progr. Colloid Polym. Sci.* **67** (1980) 71 – 83.
65. J. T. Okubo, H. Fujita, *Colloid Polymer Sci.* **274** (1996) 368 – 374.
66. R. H. Ottewill, *Progr. Colloid Polymer Sci.* **67** (1980) 71 – 83.
67. C. Smits, W. J. Briels, J. K. G. Dhont, H. N. W. Lekkerkerker, *Progr. Colloid Polymer Sci.* **79** (1989) 287 – 289.
68. W. Luck, M. Klier, H. Wesslau, *Ber. Bunsenges. Phys. Chem.* **67** (1963) 75 – 85.
69. T. Okubo, H. Fujita, K. Kiriyama, H. Yamaoka, *Colloid Polymer Sci.* **274** (1996) 73 – 80.
70. P. Pieranski, *Contemp. Phys.* **24** (1983) 25 – 73.
71. N. Güven, R. M. Pollastro (eds): "Clay-Water Interface and Its Rheological Implications," CMS workshop lectures, vol. 4, The Clay Min. Soc., Boulder, CO 1992.
72. G. Lagaly, *Appl. Clay Sci.* **4** (1989) 105 – 123.
73. S. Abend, G. Lagaly, *Appl. Clay Sci.* **16** (2000) in press.
74. A. P. Philipse, B. C. Bonekamp, H. J. Veringa, *J. Am. Ceram. Soc.* **73** (1990) 2720 – 2727.
75. A. P. Philipse, *Curr. Opinion Colloid Interface Sci.* **2** (1997) 200 – 206.
76. D. J. Shaw: *Introduction to colloid and surface chemistry,* Butterworths, London 1980.
77. D. M. E. Thies-Weesie, A. P. Philipse, H. N. W. Lekkerkerker, *J. Colloid Interface Sci.* **177** (1996) 427 – 438.
78. H. Tributh, G. Lagaly, GIT, Fachz. Lab., **30** (1986) 524 – 529, 771 – 776.
79. D. Heath, Th. F. Tadros, *J. Colloid Interface Sci.* **93** (1983) 320 – 328.
80. J. Eisenlauer, E. Killmann, *J. Colloid Interface Sci.* **74** (1980) 108 – 119.
81. G. Lagaly in B. Dobias (ed.): "From Clay Minerals to Colloidal Clay Mineral Dispersions," *Coagulation and Flocculation – Theory and Applications,* Marcel Dekker, New York 1993, pp. 427 – 494.
82. J. Lyklema: *Fundamentals of Interface and Colloid Science, vol. II Solid-liquid interfaces,* Academic Press, New York 1995.
83. P. H. Wiersema, A. L. Loeb, J. Th. G. Overbeek, *J. Colloid Interface Sci.* **22** (1966) 78 – 99.

84. R. H. Ottewill, N. J. Shaw, *Electroanalyt. Chem. Interf. Electrochem.* **37** (1972) 133 – 142.
85. J. Th. G. Overbeek, B. H. Bijsterbosch in P. G. Righetti, C. J. van Oss, J. W. Vanderhoff (eds.): *Electrokinetic Separation Methods,* Elsevier, Amsterdam 1979, 1 – 32.
86. C. S. Mangelsdorf, L. R. White, *J. Chem. Soc. Faraday Trans.* **94** (1998) 2441 – 2452, 2583 – 2593.
87. P. Ney: *Zeta-Potentiale und Flotierbarkeit von Mineralen,* Springer Verlag, Wien–New York 1973.
88. R. J. Hunter, *Colloids Surf* **A 141** (1998) 37 – 65.
89. A. Tiselius, *Trans Faraday Soc.* **33** (1937) 524 – 531.
90. L. H. Allen, E. Matijević, *J. Colloid Interface Sci.* **31** (1969) 287 – 296.
91. R. H. Ottewill, J. N. Shaw, *Kolloid Z. Z. Polymere* **218** (1967) 34 – 40.
92. D. Horn, W. Ditter, *Progr. Colloid Polym. Sci.* **60** (1976) 203 – 212.
93. A. Tsugita, S. Takemoto, K. Mori, R. Yoneya, Y. Otani, *J. Colloid Interface Sci.* **95** (1983) 551 – 560.
94. S. Abend, N. Bonnke, U. Gutschner, G. Lagaly, *Colloid Polymer Sci.* **276** (1998) 730 – 737.
95. G. Lagaly, M. Reese, S. Abend, *Appl. Clay Sci.* **14** (1999) 83 – 103.
96. E. J. W. Verwey, J. Th. G. Overbeek: *Theory of the stability of lyophobic colloids,* Elsevier, Amsterdam-New York 1948.
97. J. Th. G. Overbeek, *J. Colloid Interface Sci.* **58** (1977) 408 – 421.
98. J. Th. G. Overbeek, *Adv. Colloid Interface Sci.* **16** (1982) 17 – 30.
99. G. Lagaly in H. Asche, D. Essig, P. C. Schmidt (eds.): *Technologie von Salben, Suspensionen und Emulsionen,* **APV vol. 10,** Wissenschaftl. Verlags GmbH, Stuttgart 1984, pp. 32 – 61.
100. J. Th. G. Overbeek, *Pure Appl. Chem.* **52** (1980) 1151 – 1161.
101. J. Visser, *Adv. Colloid Interface Sci.* **3** (1972) 331 – 363.
102. J. Lyklema: *Fundamentals of Interface and Colloid Science,* **vol. I Fundamentals,** Academic Press, New York 1991.
103. R. H. Ottewill, R. A. Richardson, *Colloid Polym. Sci.* **260** (1982) 708 – 719.
104. G. Frens, *Faraday Discuss Chem. Soc.* **65** (1978) 146 – 155.
105. E. Frey, G. Lagaly, *J. Colloid Interface Sci.* **70** (1979) 46 – 55.
106. S. D. Lubetkin, S. R. Middleton, R. H. Ottewill, *Philos. Trans. R. Soc. London Ser. A* **311** (1984) 353 – 368.
107. R. B. Secor, C. J. Radke, *J. Colloid Interface Sci.* **103** (1985) 237 – 244.
108. T. Permien, G. Lagaly, *Clay Min.* **29** (1994) 751 – 760.
109. E. P. Honig, G. J. Roebersen, P. H. Wiersema, *J. Colloid Interface Sci.* **36** (1971) 97 – 109.
110. T. G. M. van der Ven, *J. Colloid Interface Sci.* **124** (1988) 138 – 145.
111. H. Reerink, J. Th. G. Overbeek, *Discuss. Farad. Soc.* **18** (1954) 74 – 84.
112. D. H. Napper: *Polymeric stabilization of colloidal dispersions.* Academic Press, London 1983.
113. B. Vincent, J. Edwards, S. Emmett, A. Jones, *Colloids Surf.* **18** (1986) 261 – 281.
114. R. Evans, D. H. Napper, *Kolloid Z. Z. Polym.* **251** (1973) 329 – 336.
115. K. E. J. Barett: *Dispersion Polymerization in organic media,* J. Wiley and Sons, London 1975.
116. J. Clarke, B. Vincent, *J. Colloid Interface Sci.* **82** (1981) 208 – 216.
117. H. de Hek, A. Wrij, *J. Colloid Interface Sci.* **79** (1981) 289 – 294.
118. J. Edwards, D. H. Everett, R. O'Sullivan, I. Pangalou, B. Vincent, *J. Chem. Soc. Faraday Trans. I* **80** (1984) 2599 – 2607.
119. G. J. Fleer, J. H. M. H. Scheutjens, B. Vincent, *ACS Symp. Ser.* **240** (1984) 245 – 263.
120. G. J. Fleer, M. A. Cohen Stuart, J. M. H. M. Scheutjens, T. Cosgrove, B. Vincent: *Polymers at Interfaces,* Chapman & Hall, London 1993.
121. B. Vincent, J. Edwards, S. Emmett, R. Croot, *Colloids Surf.* **31** (1988) 267 – 298.
122. B. van Lent, R. Israel, J. M. H. M. Scheutjens, G. J. Fleer, *J. Colloid Interface Sci.* **137** (1990) 380 – 394.
123. B. Vincent, *Colloids Surf.* **50** (1990) 241 – 249.
124. A. Jones, B. Vincent, *Colloids Surf.* **42** (1989) 113 – 138.
125. J. E. Seebergh, J. C. Berg, *Langmuir* **10** (1994) 454 – 463.
126. J. N. Smith, J. Meadows, P. A. Williams, *Langmuir* **12** (1996) 3773 – 3778.
127. E. S. Pagac, R. D. Tilton, D. C. Prieve, *Langmuir* **14** (1998) 5106 – 5112.
128. M. B. Einarson, J. C. Berg, *Langmuir* **8** (1992) 2611 – 2615.
129. J. Lyklema in Ed. D. Goldberg (ed.): *The Nature of Seawater,* Physical and Chemical Sciences Res. Rep. 1, Dahlem Workshop Report, pp. 579 – 597.

130. S. M. Ahmed in J. W. Diggle (ed.): *Oxides and Oxide films,* vol. 1, Dekker, New York 1972.
131. R. M. Duckworth, A. Lips, *J. Colloid Interface Sci.* **64** (1982) 311 – 319.
132. L. K. Koopal, W. H. van Riemsdijk, M. G. Roffey, *J. Colloid Interface Sci.* **118** (1987) 117 – 136.
133. J. Lyklema, J. N. de Wit, *J. Electroanal. Chem.* **65** (1975) 443 – 452. *Colloid Polym. Sci.* **256** (1978) 1110 – 1120.
134. L. Righetto, G. Azimonti, T. Massana, G. Bidoglio, *Colloids Surf. A* **95** (1995) 141 – 157.
135. E. Matijević, M. E. Ronayne, J. P. Kratohvil, *J. Phys. Chem.* **70** (1966) 3830 – 3884.
136. N. de Roody, P. L. de Bruyn, J. Th. G. Overbeek, *J. Colloid Interface Sci.* **75** (1980) 542 – 554.
137. T. Permien, G. Lagaly, *Colloid Polymer Sci.* **272** (1994) 1306 – 1312.
138. E. Matijević, *J. Colloid Interface Sci.* **43** (1973) 217 – 245, **58** (1977) 374 – 389.
139. F. Dumont, J. Warlus, A. Watillon, *J. Colloid Interface Sci.* **138** (1990) 543 – 554.
140. J. Israelachvili: *Intermolecular and Surface Forces,* 2nd ed., Academic Press, London 1994.
141. E. Tombacz, M. Szekeres, I. Kertész, L. Turi, *Progr. Colloid. Polymer Sci.* **98** (1995) 160 – 168.
142. Y. Kwong Leong, P. J. Scales, T. W. Healy, D. V. Boger, *Colloids Surf. A* **95** (1995) 43 – 52.
143. R. O. James, G. R. Wiese, T. W. Healy, *J. Colloid Interface Sci.* **59** (1977) 381 – 385.
144. E. Matijević, *J. Colloid Interface Sci.* **58** (1977) 374 – 389.
145. H. Rupprecht, T. Gu, *Colloid Polymer Sci.* **269** (1991) 506 – 522.
146. H. Rupprecht, M. Biedermann, *Colloid Polym. Sci.* **252** (1974) 558 – 565.
147. H. Rupprecht, *Acta Pharm. Technol.* **22** (1976) 37 – 54.
148. F. F. Aplan, D. W. Fuerstenau in D. W. Fuerstenau (ed.): Froth Flotation, 50th Anniversary Volume *AIME 1962,* p. 170.
149. Th. F. Tadros, B. Vincent, *J. Colloid Interface Sci.* **72** (1979) 505 – 514.
150. G. Lagaly, in A. D. Wilson and H. J. Prosser (ed): *Development of Ionic Polymers II,* Elsevier, London, 1986, pp. 77 – 140; in L. G. Schultz, H. van Olphen, and F. Mumpton (ed): *Proc. Internat. Clay Conference, Denver 1985,* Elsevier 1986.
151. V. Chaplain, M. L. Janex, F. Lafuma, C. Graillat, R. Audebert, *Colloid Polymer Sci.* **273** (1995) 984 – 993.
152. L. Nabzar, E. Pefferkorn, R. Varoqui, *J. Colloid Interface Sci.* **102** (1984) 380 – 388.
153. H. C. Husband, J. M. Adams, *Colloid Polymer Sci.* **270** (1992) 1194 – 1200.
154. Y. Adachi, *Adv. Colloid Interface Sci.* **56** (1995) 1 – 31.
155. F. Le Berre, G. Chauveteau, E. Pefferkorn, *J. Colloid Interface Sci.* **199** (1998) 1 – 12.
156. Q. Wang, *J. Colloid Interface Sci.* **150** (1992) 418 – 427.
157. Y. Otsubo, K. Watanabe, *J. Colloid Interface Sci.* **127** (1989) 214 – 221; **133** (1989) 491 – 495.
158. G. C. March, D. H. Napper, *J. Colloid Interface Sci.* **61** (1977) 383 – 387.
159. K. Makino, S. Yamamoto, K. Fujimoto, H. Kawaguchi, H. Ohshima, *J. Colloid Interface Sci.* **166** (1994) 251 – 258.
160. H. Ohshima, *Electrophoresis* **16** (1995) 1360 – 1363; H. Ohshima, *Adv. Colloid Interface Sci.* **62** (1995) 189 – 235.
161. M. J. Snowden, N. F. Marston, B. Vincent, *Colloid Polymer Sci.* **272** (1994) 1273 – 1280.
162. R. J. Pugh, J. A. Kitchener, *J. Colloid Interface Sci.* **35** (1971) 656 – 664; **38** (1972) 656 – 657.
163. T. W. Healy, G. R. Wiese, D. E. Yates, B. V. Kavanagh, *J. Colloid Interface Sci.* **42** (1973) 647 – 649.
164. H. Sasaki, E. Matijević, E. Barouch, *J. Colloid Interface Sci.* **76** (1980) 319 – 329.
165. E. Matijević, *J. Pure Appl., Chem.* **53** (1981) 2167 – 2179.
166. R. O. James, A. Homola, W. Healy, *J. Faraday Trans. I,* (1977) 1436 – 1445.
167. M. R. Böhmer, L. K. Koopal, *Langmuir* **8** (1992) 2649 – 2659, 2660 – 2665.

Color → Dyes, General Survey
Color and Constitution → Dyes, General Survey
Colorants (Food) → Foods, 3.Additives
Colorants (Plastics) → Plastics, Additives
Colorants for Ceramics → Ceramic Colorants
Color Measurement → Dyes, General Survey
Color Photography → Photography
Columbium → Niobium and Niobium Compounds

Composite Materials

Separate keywords are: Reinforced Plastics, Metal-Matrix Composites; Cements see Hard Materials

Bernhard Ilschner, Ecole Polytechnique Fédération de Lausanne, Lausanne, Switzerland (Chaps. 1 – 3)

Joseph K. Lees, E.I. Du Pont de Nemours & Co., Wilmington, DE 19898, United States (Chaps. 4 – 7)

Ashok K. Dhingra, E.I. Du Pont de Nemours & Co., Wilmington, DE 19898, United States (Chaps. 4 – 7)

R. L. McCullough, University of Delaware, Newark, DE 19716, United States (Chaps. 4 – 7)

1. Definition 2
2. Classification 4
2.1. Geometric Classification 4
2.2. Material Classification 5
3. Principles of Fiber Reinforcement . 5
3.1. High-Strength Fiber Composites . . 5
3.2. Behavior in Compliant Matrices . . 6
3.3. Behavior in Brittle Matrices 8
4. Materials Used in Advanced Composites 9
4.1. Reinforcing Fibers 9
4.1.1. Carbon Fibers 10
4.1.2. Organic Fibers 12
4.1.3. Inorganic Fibers 14
4.1.4. Reinforcement Geometry 17
4.2. Matrix Materials 18
5. Processing and Fabrication Technology 22
5.1. Thermoset Resin Matrix Composites 24
5.1.1. Classic Low-Volume Production Methods 24
5.1.2. Automation of Traditional Production Methods 26
5.1.3. Alternate Production Methods 27
5.2. Thermoplastic Resin Matrix Composites 28
5.3. Design Integration 28
6. Properties of Advanced Composite Materials 29
6.1. Properties of Composite Structures 29
6.2. Laminate Analysis 32
6.3. Micromechanical Models 33
7. Applications 37
8. References 42

1. Definition

Structure as a Basis for Definition. The word *composite* is used in the technical sense to describe a product that arises from the incorporation of some basic structural material into a second substance, the *matrix*. That which is incorporated can be in the form of particles, whiskers, fibers, lamellae, or a mesh. Its task is to impart its own advantageous mechanical characteristics to the matrix material. Composites can be produced either by thermomechanical mixing of individual components or by means of chemical reactions that lead to phase separations within homogeneous starting mixtures (cf. page 2).

To obtain a composite with desirable mechanical properties, it is necessary that the separation between the incorporated elements be in the range of 0.01 – 1 mm and that the incorporation show directional characteristics. The extent of mutual adhesion and interpenetration of the components must also be favorable.

The key characteristics sought in an incorporated substance, which are in turn conferred on the composite, are above all elastic rigidity, tensile and fatigue strength, and hardness, as well as appropriate electrical and magnetic properties. Glass fiber reinforced resins are typical examples of composites, and these are probably the best known (→ Reinforced Plastics).

Composite materials and glass fiber reinforced plastics, respectively, are known in German as "Verbundwerkstoffe" and "glasfaserverstärkte Kunststoffe," and in French as "matériaux composites" and "matière plastique armée aux fibres de verre."

For further information in the form of introductory textbooks, monographs, and review articles, see [1–6].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_369

Modes of Preparation as a Basis for Definition. The most common way to produce composites entails thermomechanical introduction of one or more additive components R_1, R_2, $R_3, \ldots$, into a preexisting matrix. Particularly suitable techniques include winding and infiltration (see the literature sources cited above). Extrusion [7] and thermal injection [8] methods have also been described. The variety of possible forms that composite materials can take and the number of possible component combinations are large. Finishing of the materials presents no major problems; the appropriate technology is both proven and cost effective. Rather, it is the preparation of composites in suitable form that tends to be costly.

An interesting alternative to thermomechanical mixing is the possibility of using a phase-separation process within a homogeneous starting material to create the desired heterogeneity. The following processes lend themselves to this approach: eutectic crystallization from a melt, eutectic decomposition of solid solutions, or phase separation within a supersaturated solid solution [9], [10]. Composites that form to some extent in this way within partially prefabricated construction elements are known as *in situ composites.*

The method is a particularly elegant one, and it affords an opportunity to produce exceedingly fine geometric distribution of the phase components. Moreover, constituents that are difficult to prepare or handle can sometimes be introduced by this method. Nonetheless, the following are disadvantages:

1) The number of multicomponent systems suited to the approach and at the same time capable of yielding desirable properties is extremely limited.
2) The resulting ratio of matrix : included phase is often a fixed property of the system.
3) Reactions of this kind proceed under diffusion control and with a finite number of transport paths to the growing front. Moreover, both of these parameters change during the course of the reaction. To introduce a measure of selectivity on the directional characteristics of the composite, the reaction must be correspondingly controlled, for example, by provision of a steep temperature gradient in the direction of desired growth. Thus, production is slow, highly subject to interference, and limited to geometrically simple forms.

A material prepared in the latter way through controlled solidification of a eutectic mixture is referred to as a *directionally solidified eutectic,* commonly abbreviated DSE.

The method of phase separation is best suited to the manufacture of small items that are subject to rigid demands, where cost plays a minor role (e.g., for turbine blades or microcomponents) [11].

Both approaches are represented schematically in Figure 1.

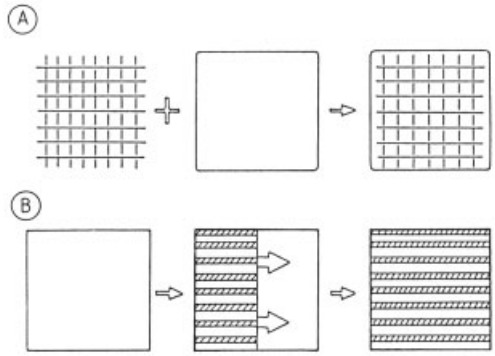

Figure 1. Symbolic comparison of the two principal approaches to the manufacture of composite materials
A) Thermomechanical mixing; B) In situ preparation by phase separation

Distinction from Other Classes of Materials. The definition provided above is occasionally criticized as being too narrow, where the objections arise from the point of view of the material sciences. Thus, it is argued that it is the nature of the interaction between two or more finely divided materials that determines measurable macroscopic properties of a composite, not the way the mixture is prepared or the purpose it is intended to serve. The very fact that oriented materials resulting from eutectic crystallization are called in situ composites demonstrates that the eutectic mixture itself determines the properties of the composite. This line of reasoning suggests that all eutectic mixtures (not only those that are oriented) should be regarded as composites – including cast iron, Al – 12 % Si alloys,

and soft solder. It is further asserted that no fundamental difference exists between an inclusion comprised of Ti/W carbides in a Co matrix and the product resulting from distribution of Cr and Nb carbides in a ferrite matrix, whereas only the former system (cemented carbides) is normally referred to as a composite. Similar argumentation can lead to the conclusion that paper and even the natural product wood are composite materials.

The term "composite" is not normally used in as broad a sense as this interpretation would suggest. Therefore, it is preferable to employ a more general term such as "multiphase" material (or solid) when one is discussing the general relationship between microstructure and macroscopic properties, and to reserve the word "composite" for cases in which some specific characteristic of an incorporated material is intentionally being exploited. The latter criterion does not apply to most eutectic or precipitation-hardened alloys. It is true that the hardness of a carbide or intermetallic compound that is introduced into a metal matrix does contribute strengthening characteristics, but this is not a property that is exploited as such. In glass fiber reinforced resins, on the other hand, the elastic rigidity of the glass fibers is the desired property. Similarly, the hardness of carbides is the reason these are added in the preparation of hardened alloys, and asbestos fibers are what supply the tensile strength of asbestos cement.

Cast iron represents an interesting borderline case. Gray cast iron with eutectically oriented graphite lamellae is often employed in applications in which the damping properties of the graphite lamellae are advantageous. White cast iron, on the other hand, is preferred for purposes that require hardness, where eutectic carbide is performing the same function as in hardened alloys. Nevertheless, the historical development of both forms of cast iron was unrelated to such considerations, so that it is reasonable to conform to tradition and refrain from designating cast iron as a "composite."

It is also inappropriate to describe as a composite a macroscopic material into which whole structural elements (wires, meshes), themselves homogeneous, have been introduced for the purpose of improving mechanical or other properties (e.g., steel-reinforced concrete and steel-belted tires). The same is true even if the union of materials is conducted in a layered fashion (e.g., plywood) or by coating. Similarly, sheet steel or pipe stock that has been electrolytically nickel plated, galvanized, plated in a rolling mill, or plastic coated is not a composite material.

2. Classification

2.1. Geometric Classification

A reinforcing agent is introduced into a given matrix in the form of particles. These are understood to exist as dispersed entities with planar (usually crystalline) or rounded boundary surfaces. For a particular particle shape, the relationship of particle radius r to the distance between centers of adjacent particles d determines the volume fraction f of the composite material, relative to the volume V:

$$f = V_R / (V_M + V_R) \tag{1}$$

Throughout this section the subscript R is used to refer to the incorporated reinforcing material, wheras the index M signifies the matrix. The maximum volume fraction that can be obtained by closest spherical packing of spherical particles of uniform radius is $f_{max} = 0.680$. Higher values of f can be attained by using particles of varying diameter if these are arranged (e.g., by applying vibration) so that smaller spheres occupy the spaces between larger ones. According to [12], space filling that approaches 95 %, for example, results from the layering of spherical particles of radius ratio of 1 : 7 : 40 : 320 in a volume relationship of 6 : 10 : 25 : 59. Even higher degrees of admission are possible if the spherical shape is abandoned, as shown by the polycrystalline structure of metals and ceramics. Thus, cubes and hexagonal dodecahedra would, in theory, lead to values of 100 %. Other particle shapes (ellipsoids) reduce the potential packing density, but at the same time they open the way to establishing preferential orientation.

Whiskers are monocrystalline fibers that are largely free of defects and are thus subject to elastic deformation even under extreme tension (→ Fibers, 5. Synthetic Inorganic). Such whiskers can be obtained by the chemical vapor deposition (CVD) process, which entails reactive deposition from the gas phase. Whisker

diameter ranges from 0.1 to 1 μm, and is thus an order of magnitude below that of ordinary fibers. Their length reaches 0.1 – 1 mm. Whiskers exhibit very high mechanical strength as a result of their lack of defects. Tensile strengths ca. 10 % as large as the elastic modulus have been measured, a value 5 times that associated with a high-quality polycrystalline fiber and 100 times as large as that for massive polycrystalline material. Unfortunately, the preparation and handling of whiskers is both difficult and costly. Despite 20 years of developmental effort, it has still not proven possible to mass produce them, although well over 100 substances have been prepared as whiskers on a laboratory scale. Expectations were high – probably unrealistically so – in the 1960s for future applications of whiskers, but these have been largely unfulfilled [13]. In principle, whiskers should be ideal components for use in composites, since this application is one that would take full advantage of their remarkable strength while at the same time making available in the form of the matrix a suitable container and protective coating.

Fibers are incorporated into composites in both short and continuous forms, although even the so-called short fibers rarely have length: radius ratios under 20. In most cases the fiber's high rigidity or tensile strength is to be conferred upon the composite material. Thus, a key role is played by the transfer of forces out of the matrix and, therefore, by the microgeometry and wettability of the boundary surfaces. For this reason, the fibrous materials are often either provided with a coating or in some other fashion "activated."

Most fiber-reinforced composites are distinguished by the rigid orientation of their fibers ("unidirectional composites," UDC). This leads to a distinctive directionality with respect to macroscopic (e.g., mechanical) properties of the materials as well – they show high strength in the direction of the fibers, but show poor characteristics in the transverse direction. Weakness of this kind intolerable is in many applications (e.g., in automotive components). One way to compensate for this weakness is through introducing fibers in the form of a mesh (laminates) rather than unidirectional single fibers. Resulting improvements in strength transverse to the principal axis of stress are obtained at a price, however: the effective volume fraction in the direction of primary stress is substantially reduced, as is the corresponding carrying capacity. The same holds true for irregular "fleecy" inclusions.

2.2. Material Classification

Since the volume fraction of fibers in a composite rarely exceeds 0.5, the general appearance of the material is conferred by the matrix. This fact is reflected in the terminology that is employed (e.g., fiber-reinforced resin). Almost all of the known structural material types can play the part of the *matrix.*

High polymers are used as building materials [14], [15]. Most often encountered are unsaturated polyester (UP) and epoxy (EP) resins. Other examples include polyacrylate dental materials containing silicate particles and polypropylene to which glass fibers have been added for use in shatterproof safety glass.

Metals are less commonly used for ordinary temperature composite materials because of their undesirable weight. Exceptions include aluminum and other light-metal alloys strengthened by the incorporation of boron, SiC, or carbon fibers [16], [17]. Al_2O_3 whiskers are, of course, also utilized, although the economics involved are far from ideal. Strengthening through the use of steel threads has also been studied [18].

One important domain of fiber-reinforced metal composites is the area of high-temperature applications. The use of organic matrices is here excluded, although ceramic reinforcing materials represent an alternative. A special case of growing importance is the category of in situ composites (oriented eutectic mixtures, already mentioned above). Examples of the latter include Ni – Co-based materials containing fibrous TaC, or ones based on Fe and reinforced with $(Cr,Mn)_7C_3$.

Cobalt-based alloys containing particulate WC/TiC (so-called hardened alloys) are also employed on a large scale for cutting tools.

A third domain of metal-based composites includes materials for electrical engineering applications, such as contacts [19], [20].

Ceramics containing fibrous materials (Al_2O_3, SiC, and B_4C) have come into use as lightweight bulletproof shields (e.g., helmets,

vests, and helicopter seats), though civilian applications exist as well.

Cement containing asbestos, glass, plastic, or steel threads has numerous uses in the construction industry, especially for poured concrete structures. Also noteworthy is the high impact resistance of concrete reinforced with polypropylene fibers [23–25].

Rubber can also serve as the basis for composites, as in preparations containing magnetic particles ("magnetic rubber").

Materials with proven value as *fibrous reinforcing agents* for composites include the following:

Glass fibers, especially those derived from E-type glass (aluminum – boron – silicate glass) and, for better elastic properties, S-type glass (SiO_2– Al_2O_3– MgO) – these materials are also introduced in fabric or fleece form.

Carbon fibers, prepared by the pyrolysis of cellulose or polyacrylonitrile fibers at high temperature and under inert atmosphere – if the final stage of the thermal treatment is carried out at ca. 2800 °C, the material is converted largely into a graphite possessing exceptional tensile strength [26], [27].

Tungsten-core fibers, including boron, B_4C, and SiC – these are prepared by thermal decomposition of halides on a tungsten core.

Metal fibers can be prepared from the extremely light but toxic beryllium, as well as from steel, niobium, molybdenum, and tungsten. Their advantage is their plastic ductility (albeit restricted within specific limits), but they have the disadvantage of great weight.

Ceramic fibers – these include important natural fibers (asbestos, especially in its chrysotile modification), but synthetic ceramic fibers are also known (though largely on the laboratory scale) [28].

Hybrid composite materials containing two or more additives (e.g., carbon and glass fibers) are increasingly important because they represent a means of creating a balance between the advantages and the disadvantages shown by individual components.

3. Principles of Fiber Reinforcement [1–6]

3.1. High-Strength Fiber Composites

Despite the complications attending their preparation, composite materials represent an interesting alternative to metals whenever there is a demand for great strength with minimal mass. To characterize this property, it is common to employ either the term *specific strength* $R_m/(\varrho \cdot G)$ (ϱ = density in g/m^3, $G = 9.80665$ m/s^2 = gravitational acceleration) or else the *specific modulus of elasticity* (Young's modulus) $E/(\varrho \cdot G)$. Since $\varrho \cdot G$ has the unit N/m^3, whereas R_m and E are in N/m^2, the resulting specific quantity has the unit m. In practice, data are normally reported in km; e.g. $E/(\varrho \cdot G)$ for graphitized carbon fibers is 2000 km. This measure corresponds to the breaking length data commonly reported for textile fibers.

High values of this parameter are only attainable with materials having high tensile strength and low density. These criteria apply mainly to the fibrous component, since the strength of the composite is determined almost exclusively by the fibers. The matrix serves to protect fiber surfaces from damage, distribute applied forces, and compensate for microscopic fiber fracturing. Thus, it is more important that the matrix material have a low density than that it be particularly strong.

Strong binding and high rigidity are achieved primarily by maximizing the number of covalent bonds per unit cross-section. This high concentration of bonds ("valences") can only be realized without undue increase in fiber weight if the constituent atoms are of low atomic mass.

For these reasons only a few elements can be considered candidates for the construction of high-strength fibers: those with low atomic number and covalent bonding character [3]. Beryllium can be largely excluded on the basis of its toxicity. Thus, there remain (apart from Al_2O_3) only boron, carbon, and silicon, as well as a few of their compounds with oxygen and nitrogen. Hydrocarbon and silane fibers generally possess low E moduli; their constituent hydrogen atoms occupy much space, but contribute nothing to material strength. The best results within this category of substances have been achieved with

aramide fibers, which combine benzene rings and – CONH – groups in rather strict alignment.

When one considers applications entailing the strengthening of metals, especially for use as high-temperature materials, the thermal stability of the fibrous component becomes another criterion, which limits the choice further. The best solution in a qualitative sense is graphitized carbon fibers [26], although B_4C and BN fibers of very high quality and with melting points above 2500 °C can also be prepared (by high-temperature boration of carbon fibers and high-temperature nitration of B_2O_3, respectively).

The choice is further influenced by price and production costs, factors that are highly susceptible to change in response to technological developments. Materials exist whose costs are more favorable than those of the components described above, but their properties are also less ideal. An example is mullite fiber (Al_2O_3–SiO_2–B_2O_3) produced by the centrifugal melt spinning process. Glass fibers fall at the low end of both the price scale and the thermal stability spectrum. It should be noted again that optimization of the balance of technical and economic factors is often possible with hybrid composites, such as those containing combinations of carbon and glass fibers.

Strength factors as they are actually observed, particularly in the case of graphite fibers, come very close to the theoretical values predicted from atomic bond strength data. Thus, it is unrealistic to anticipate that further significant increases will be obtained.

3.2. Behavior in Compliant Matrices [29–32]

The most easily described situation consists of a fiber composite material containing infinitely long fibers oriented in a single direction. If this material contains n_F fibers of cross-section a_F within an overall cross-sectional unit area, then the corresponding fiber volume fraction can be calculated as $f = n_F\, a_F$. When such a composite is placed under elastic stress, the extent of elongation is identical for both fiber and matrix (apart from small regions near the ends of the sample). Application of Hooke's law under these considerations illustrates the role of the modulus of elasticity E_F of the fibers. Thus, the average tension that the composite is capable of sustaining for a given elongation is

$$\sigma_c = f\sigma_F + (1-f)\,\sigma_M = [f E_F + (1-f)\, E_M]\,\varepsilon \qquad (2)$$

This can be rewritten as

$$\sigma_c = E_c \cdot \varepsilon$$

or, if the effective modulus of elasticity for the composite is expressed as E^c, it can be written as

$$E_c = f E_F + (1-f)\, E_M \qquad (3)$$

Equation (3) contains the so-called "rule of mixtures," which provides a first approximation to the behavior of a composite material. It is known, however, that this approximation can, for a number of reasons, lead to errors of ± 30 % or more in individual cases. Notice that if one assumes a fiber volume fraction of ca. 50 %, then $E_c \approx 1/2 E_F$, provided that (as is usually the case) $E_F \gg E_M$. Thus, the elastic properties of the matrix are irrelevant to the overall behavior of the composite. The fibers alone determine the strength of the whole.

Nevertheless, glass fibers and most of the other fibers that lend themselves to practical application are brittle; that is to say, they break rather than undergoing plastic deformation upon reaching some maximum load σ_F^*. This behavior is of considerable importance in the development of a composite material, because the total elongation ε (see above) at maximum load (the breaking point) remains very low due to the high E modulus of the fibers:

$$\varepsilon^* = \sigma_F^* / E_F \qquad (4)$$

At this elongation the amount of stress assumed by the matrix, determined by its elastic resilience, is correspondingly small, namely $\tilde{\sigma}_M = E_M \varepsilon^*$. The load supported by the matrix at the breaking point is thus

$$\sigma_c^* = f\sigma_F + {}^*(1-f)\,\tilde{\sigma}_M \qquad (5)$$

This composite strength increases in a linear manner from $\tilde{\sigma}_M$ as the degree of admission increases. Nonetheless, it is also necessary to take into account the fact that a nonreinforced matrix with the same cross-section $(1-f)$ has, in the

region $f = 0$, a greater strength than the composite (if it is assumed that the breakage of a high proportion of the fibers also leads to destruction of the matrix). Thus:

$$\sigma^*_{\mathrm{MN}} = (1-f)\, E_{\mathrm{M}} \varepsilon^*_{\mathrm{M}} \tag{6}$$

where $\varepsilon^*_{\mathrm{M}}$ is the breaking elongation of a sample of matrix from which a certain amount of material has been removed to create the channels that would otherwise be occupied by fibers. Figure 2 is a graph of Equations (5) and (6), clearly showing that at low volume fractions there is a decrease in overall tensile strength caused by reduction in the matrix cross-section.

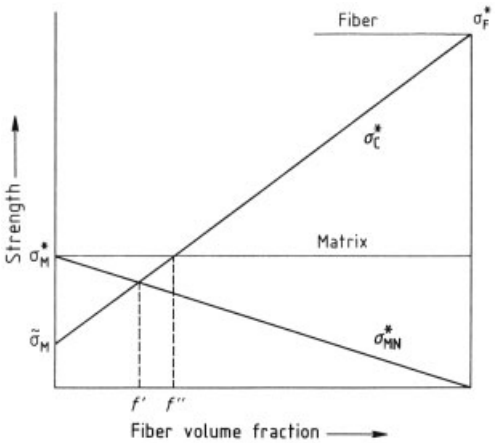

Figure 2. Influence of the volume fraction on strength σ^*_{c} for a composite whose fibers are oriented in a single direction as compared to the strength σ^*_{MN} of a nonreinforced matrix of the same cross-sectionOnly if $f > f''$ is there a true strengthening of the composite relative to matrix material.

Only after a *critical volume fraction* f' has been reached does the strength begin to rise. True strengthening by fibers can only be said to occur after a volume fraction f'' has been exceeded, where f'' is defined such that above this value, the strength of the composite is greater than that of a pure matrix of comparable cross-section. Thus:

$$f'' = (\sigma^*_{\mathrm{M}} - -\tilde{\sigma}_{\mathrm{M}}) \,/\, (\sigma^*_{\mathrm{F}} - -\tilde{\sigma}_{\mathrm{M}}) \tag{7}$$

In the event that the E modulus of the fibers is very large, $\tilde{\sigma}_{\mathrm{M}} \ll \sigma^*_{\mathrm{F}}$ and, in fact, $\ll \sigma^*_{\mathrm{MN}}$. Equation (7) can then be approximated by:

$$f'' \approx \sigma^*_{\mathrm{M}} / \sigma^*_{\mathrm{F}} \tag{7a}$$

Thus, if the fibers have a tensile strength 10 times that of the matrix, the degree of admission must be at least 10 % for a strengthening effect to be achieved. This would appear readily attainable, but certain limitations must also be recognized. For example, the tensile strength of a composite is reduced when, in the interest of improving transverse strength, uniaxially incorporated fibers are replaced by a fiber network. Moreover, defects are always present in the fibers of a real material. It is true that premature breakage of individual fibers will be somewhat confined as a result of localized crack branching and plastic microdeformation of the matrix. At the same time, however, the risk of crack spreading is increased if breakage of one fiber causes transmission of critical loads through the matrix to adjacent fibers. This risk makes it mandatory that an ample volume of matrix material be present even at high fiber volume fractions in order that it may serve to maintain the integrity of the composite and prevent the spreading of breaks from one fiber to another. Furthermore, the Weibull statistical probability of fiber breakage and the rate of subcritical expansion of cracks within the fibers must be known if one is to make a realistic estimate of the value of σ^*_{F}. In other words, account must be taken of the distribution of microdefects both in the fibers and at their surfaces.

All of these aspects of the subject have been thoroughly discussed [29–32]. The same is true for metal fibers, which undergo plastic deformation when subjected to stress above the flow limit, consequently become stiffened, and then break only when subjected to relatively high degrees of elongation. The matrix, too, is considerably stretched in the process, so that its fracture properties must also be taken into account. It is fundamentally true that the inherent strength of the matrix is here better utilized. The transmission of forces through a matrix by short fibers is thoroughly discussed in the literature [33], [34]. In this case, transverse stress along the fiber encasement surfaces plays a decisive role.

3.3. Behavior in Brittle Matrices

A brittle matrix represents a different situation, one in which the task of fibers is not to increase elastic strength or to raise the tensile strength relative to the matrix material. Instead, fibers are

expected to counter what is the principal disadvantage of a brittle matrix: the tendency for cracks to be rapidly propagated and – unlike metals – for breakage to occur due to the absence of damage-limiting plastic deformation in the vicinity of the crack front.

The "ductilization" that results from incorporating fibers into ceramic materials is a consequence of the fact that any cracks that form in the matrix transverse to the direction of stress fail to penetrate into the fibers: the lower *E* modulus (greater ductility) of the fibers allows them initially to undergo elongation (Fig. 3). In other words, those fibers that happen to lie in the direction of stress effectively "bridge" the crack and exert a force counterbalancing the opposing stress. Elastic elongation does lead automatically to cross-sectional contraction of the fibers, however, just as with the plastic stretching of metal fibers. Consequently, fiber separation from the matrix can occur in strain-free regions at the edge of a crack. This results in additional fiber length becoming available to accommodate further energy-consuming length changes, thus permitting the crack to open wider and increasing the potential for its further growth.

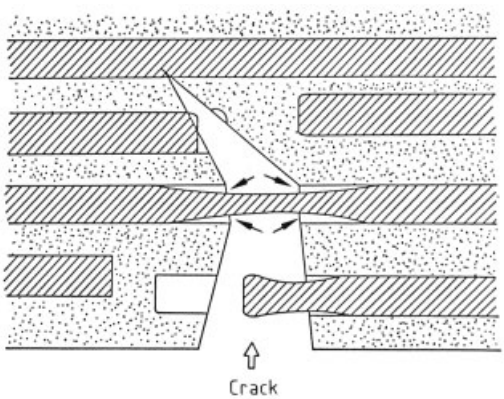

Figure 3. Elastic – plastic bridging of a crack and "pull-out" of fibers within a matrix at the onset of crack formation

Whenever short fibers are utilized, the probability (which can be easily quantified) exists that ends of fibers will be present in the vicinity of the walls of a crack. Thus, there arises the possibility that fibers will be pulled out of the matrix as the crack grows and cross-sectional contraction occurs. This process, known as "pullout," also consumes energy in the form of friction with the walls. The extent of the extra work expended in broadening of main cracks is clearly demonstrable with or without pullout. It appears as an increase in the effort required to break a sample (equal to the integral under the corresponding stress/elongation curve), and leads to an increase in ductility.

The following Chapters deal only with Advanced Composites; the others are separate keywords listened at the beginning of this article.

4. Materials Used in Advanced Composites

Approximately two decades ago, to meet the requirements of advanced military systems, a number of low-density fibers were produced that exhibited both high specific stiffness and high specific strength. In the ensuing years, the category of high-performance fibers has grown to include such diverse materials as carbon (graphite), p-aramid, boron, silicon carbide, alumina, silica, and glass. Incorporating one or more classes of these high-performance filaments in a suitable matrix to achieve properties that approach those of steel or aluminum produces a specialized class of composite materials, herein called advanced composite materials, that exhibit physical and structural properties not attainable with conventional engineering materials. The matrix or bonding agent can be a thermosetting resin, a thermoplastic resin, a metal, or a ceramic material.

On a functional basis, advanced composite materials may be described as a specialized class of composite materials that contain a high volume concentration of continuous high-performance fibers and, therefore, exhibit very high specific strength (strength/weight) and very high specific modulus (stiffness/weight).

The choice of production methods and most of the environmental properties of advanced composite materials are a function of the matrix material rather than the high-performance fiber. In this respect, resin matrix composites have the general characteristics of reinforced plastics, metal matrix composites behave like metals, and so forth. While there is much technical interest in the development of metal, ceramic, and carbon matrix advanced composite materials, resin matrix materials dominate the technology and commercial use of advanced composite materi-

als. The predominance of resin matrix systems is due to a number of reasons, which include:

1) The relatively low temperatures involved in the fabrication of resin matrix materials as compared to carbon, ceramic, or metal matrix materials, thus reducing the problems of internal stresses induced by differences in the coefficients of thermal expansion of the components and/or reaction products between components
2) The direct applicability or adaptability of fabrication technology developed for glass fiber reinforced plastics to the fabrication of advanced composite materials systems.

Carbon, ceramic, or metal matrix composites are being developed principally for use in high-temperature environments hostile to resin matrix materials, or in applications that require characteristics not attainable with organic resins, such as electrical conductivity or toughness of the level attainable with metal matrix materials.

4.1. Reinforcing Fibers

In general, reinforcing fibers used in advanced composite materials can be characterized as being light, stiff, and strong. These properties are usually exhibited by materials on the basis of the elements Be, Mg, B, Al, C, Si, Ti. Most of these relatively light elements form thermally stable covalent bonds, which are inherently stiffer and stronger than compounds formed through metallic or ionic bonds. However, many of these materials are quite brittle in bulk form because of the presence of internal or surface flaws. The probability of internal flaws, which detract from the potential strength of these materials, is minimized when the material is used in a fibrous form.

Of the elements listed above, carbon and silicon are the most important in terms of the composition of commercially available high-performance fibers. These fibers can be considered to belong to one of the three following broad families of materials: carbon, organic resins, and inorganic compounds. Within each family, several classes of high-performance fiber materials have been developed that satisfy the basic criteria of low density, high strength, and high stiffness to varying degrees. Each of these classes of fiber materials is unique in that it offers a different combination of properties and characteristics. The various fiber classes exhibit different combinations of mechanical properties, physiochemical properties, electromagnetic properties, and economic characteristics (cost and availability).

4.1.1. Carbon Fibers (→ Fibers, 5. Synthetic, Inorganic)

Carbon fibers are the predominant high-strength, high-modulus reinforcing fiber used in the manufacture of advanced composite materials. Carbon fibers are made by the pyrolytic degradation of a fibrous organic precursor. In the process, an organic polymeric fiber is heated under tension and in an inert atmosphere to a very high temperature to drive off the volatile constituents. The residual carbon atoms tend to orient themselves along the fiber axis into graphitic crystallites to form a high-strength, high-modulus fiber. The composition of the precursor and the time – temperature history of the fiber establish the properties of the carbon fiber product. In general, as the processing temperature to which the fiber is exposed increases, the extent of crystallite orientation parallel to the fiber axis increases, and thus the fiber modulus increases. Because of greater sensitivity to flaws, this increase in modulus is usually accompanied by a decrease in strength. Because of the high degree of internal structure orientation, graphite fibers are strongly anisotropic. Their transverse tensile and shear moduli are usually an order of magnitude lower than the axial modulus.

A good carbon fiber precursor has the following characteristics:

1) Its chemical structure favors the formation of an aligned graphitic carbon structure during pyrolysis
2) A high carbon fiber yield is obtained
3) Fiber integrity is maintained during the pyrolysis process
4) The fiber is strong enough to be handled during all phases of the pyrolysis process
5) It should be inexpensive and readily available from various commercial sources.

The first characteristic is essential to obtain the required mechanical properties. The others,

particularly achievement of a high carbon yield, are important for reasons of practicality and cost. Obtaining a high carbon yield is necessary not only to minimize raw material consumption, but also to reduce the volume of off-gas that is generated and must be treated.

The organic materials that have best satisfied these criteria to date are rayon (cellulose acetate), polyacrilonitrile (PAN), and petroleum pitch.

Rayon-Based Carbon Fibers. Carbon fibers based on rayon are now mainly of historical interest. The commercial use of carbon filaments produced by the pyrolysis of cellulose dates back to 1880 when THOMAS EDISON patented the incandescent electric lamp [35]. The development of methods of producing rayon-based carbon fibers that warranted their use as reinforcements in structural composites is a much more recent event, dating back to the mid-1950s and early 1960s. Key discoveries included recognition of the need to apply stress to the fiber prior to and during carbonization, which resulted in the initial development of high-modulus carbon fibers. These efforts led to the commercial production of Thornel 25 by Union Carbide. This material exhibited an average modulus of 25×10^6 psi (170 GPa) and an average tensile strength of 184×10^3 psi (1.27 GPa). As the process became more refined with time, Union Carbide was able to introduce rayon-based carbon fibers with improved performance characteristics, such as Thornel 50, and Thornel 75, which respectively exhibited an average modulus of 57×10^6 psi (380 GPa) and 75×10^6 psi (520 GPa) and average tensile strengths of 315×10^3 psi (2.17 GPa) and 385×10^3 psi (2.65 GPa) [36]. High-modulus rayon-based carbon fibers are, however, expensive to produce because of the low carbon yields (10 – 30 %) and because of the extremely high temperatures (>2800 °C) required for the "stretch graphitization" operation needed to attain moduli in excess of 100 GPa. Because of their high price (>$ 600/kg in the early 1970s), they have not been able to compete with lower cost carbon fibers that are based on other precursor materials.

PAN-Based Carbon Fibers. Most of the carbon fibers that are currently commercially available are obtained by the pyrolysis of polyacrylonitrile (PAN), an atactic, linear polymer that contains nitrile pendant groups attached to

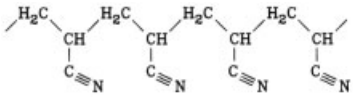

a carbon – hydrogen backbone. This technology is principally based on the work of SHINDO et al. in Japan [37] and of WATT and JOHNSON of the Royal Aircraft Establishment in Great Britain [38]. The essential elements of the production of PAN-based carbon fibers are outlined in Fig. 4. In the process, bundles (tows) that may contain from ca. 1000 to 320 000 individual fibers, each ca. 10 µm in diameter, are passed, while under tension, through a series of ovens where the temperature and atmosphere are closely controlled. The time – temperature history of the fiber establishes the characteristics of the carbon fiber produced: the higher the temperature, the higher the degree of graphitization and the higher the modulus of the resulting carbon fiber.

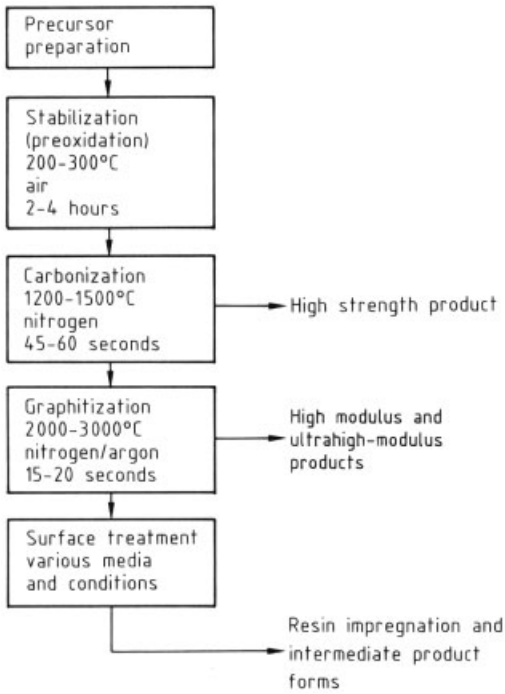

Figure 4. Graphite fiber process outline

Table 1. Representative properties of standard PAN-based carbon fibers

Fiber type Nominal designation	Type I High strength	Type II High modulus	Type III Ultrahigh modulus
Carbon content, wt %	92 – 94	>99	>99.9
Specific gravity	1.7 – 1.8	1.8 – 1.9	1.9 – 2.1
Filament diameter, µm	7 – 8	7 – 8	8 – 9
Tensile modulus, GPa	220 – 250	340 – 380	520 – 550
Tensile strength, GPa	2.5 – 3.5	2.2 – 2.4	1.8 – 1.9
Tensile elongation, %	1.2 – 1.4	0.6 – 0.7	0.3 – 0.4
Toughness, MPa	20	7.5	3
Electrical resistivity, µΩ/m	15 – 18	9 – 10	6 – 7
Longitudinal CTE *, $10\,mm^{-1}\,K^{-1}$	−0.5	−0.7	−0.9 (est.)
Range of available tow counts, 1000 fibers/bundle	1 – 320	1 – 12	<1
Price range, 1985 $/kg	40 – 300	90 – 1000	>300

* CTE = coefficient of thermal expansion.

Worldwide, there are currently over a dozen producers of PAN-based carbon fibers, each of whom offers a number of specific products. Most manufacturers offer a high-strength product (type I carbon fiber) and a high-modulus product (type II carbon fiber), with some manufacturers also offering an ultrahigh-modulus product (type III carbon fiber). Representative properties for these materials are given in Table 1. As can be noted from this table, there is a close correlation between carbon content and physical properties. Strain to failure is low for carbon and all types of carbon fibers, but especially for the higher modulus products. This makes the handling of carbon fibers difficult and is one of the major limitations of the products listed in Table 1. Carbon fibers differ from other high-perfomance reinforcing fibers because they are good conductors of electricity. They also exhibit unusual thermal expansion characteristics, since the coefficient of thermal expansion is negative along the fiber axis while being positive in the transverse directions.

The price of commercial PAN-based carbon fibers ranges from ca. $ 40/kg to $ 1000/kg (1986). These prices are mainly a function of fiber modulus and the number of filaments in a yarn bundle (end count). Type I (high-strength) fibers are less expensive than type II (high-modulus) fibers, in part because processing costs are lower and in part because they are manufactured in higher volumes. Type III (ultrahigh-modulus) fibers are expensive speciality materials. For any fiber type, price is also a strong function of end count. High end count material (40 K or more) is significantly less expensive than low end count material (12 K or less). This price differential is a reflection of the fact that plant production capacity is directly proportional to end count. Thus, production costs increase as the end count decreases. However, pyrolyzing low end count bundles results in a more uniform product. In most critical applications, high end count material is rarely used.

The low strain to failure characteristics have been the major limitation of the standard PAN carbon fibers listed in Table 1. High-strain carbon fibers are now available that exhibit tensile elongations of ca. 1.7 – 1.8 %, as shown in Table 2. These fibers are not only unusually strong, but also much easier to handle, which facilitates the fabrication of composite parts.

Table 2. Representative properties of high-strain type I PAN-based carbon fibers

Carbon content, wt %	92 – 94
Specific gravity	1.7 – 1.8
Filament diameter, µm	5 – 8
Tensile modulus, GPa	240 – 270
Tensile strength, GPa	4.0 – 4.7
Tensile elongation, %	1.7 – 1.8
Toughness, MPa	34 – 42
Longitudinal CTE, $10\,m\,m^{-1}\,K^{-1}$	−0.5
Range of available tow counts 1000 fibers/bundle	6 – 12
Price range, 1985 $/kg	90 – 100

Pitch-Based Carbon Fibers. Petroleum pitch is that fraction of the petroleum barrel that is uneconomical to further process into light hydrocarbons because of its tendency to carbonize and interfere with processing operations such as catalytic cracking. Because petroleum pitch is a high carbon content, readily available, and

low cost material, interest has been growing in its use as a precursor material in an attempt to reduce the manufacturing cost of carbon fibers.

Petroleum pitch must be purified and preconditioned to produce high-performance carbon fibers. An important step is the development of a liquid crystalline phase (mesophase) prior to spinning to form fibers with a high degree of preferred orientation. These preliminary operations add substantially to the fabrication costs.

4.1.2. Organic Fibers (→ Fibers, 4. Synthetic, Organic)

The measured tensile strength and modulus of most natural and synthetic organic fibers are typically less than 1 and 10 GPa, respectively. These values are significantly lower than the theoretical values of the maximum tensile strength (>20 GPa) and modulus (>100 GPa) calculated on the basis of force constants of bonds in the polymer chain and crystal extensions observed by X-rays on stressed fibers [39]. A number of polymeric fibers are now commercially available, whose strength and modulus more closely approach these theoretical values and which are high enough for these materials to be considered as reinforcements for advanced composites. As compared to other textile fibers, these high-strength, high-modulus materials all exhibit a high degree of molecular orientation and chain alignment. High-performance organic fibers of this class that have been commercialized to date include *p*-aramids, ultrahigh molecular mass polyolefins, and certain classes of polyesters.

***p*-Aramid Fibers.** Aromatic polyamides or aramids were introduced initially in the early 1960s to meet the need for fibers with improved heat and flammability resistance. The first example of this class of compounds was poly(*m*-phenyleneisophthalamide), which is sold commercially by Du Pont under the trademark Nomex. In the early 1970s, Du Pont introduced aramid fibers under the trademark Kevlar, based on para-substituted aromatic polyamides, most notably poly(*p*-phenyleneterephthalamide) (PPD-T), which is a condensation product of *p*-phenylenediamine and terephthalic acid.

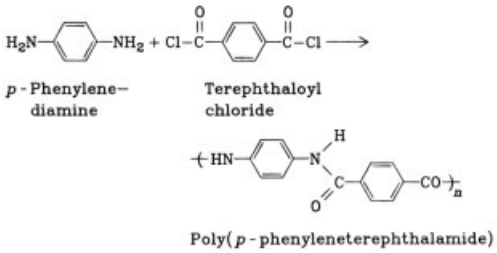

Other aramid fibers have been more recently introduced to the market by Enka (Netherlands) and Teijin (Japan).

Because of its rigid chain structure, PPD-T can exhibit liquid crystalline behavior in solution. The rodlike molecules of these materials can aggregate in nematic, ordered domains. When liquid crystalline solutions of these materials are subjected to shear, these ordered domains tend to be oriented in the direction of flow. When passed through a spinneret, these liquid crystalline solutions retain the high degree of orientation imparted by the spinning operation, leading to as-spun fibers with extraordinary levels of crystallinity and orientation.

As-spun fibers obtained by the spinning of a 20 % solution of PPD-T in 100 % sulfuric acid exhibit a crystalline orientation angle of ca. 12° (as determined from wide-angle X-ray diffraction) and a modulus of ca. 72 GPa. Heat treatment increases the degree of crystalline alignment. Heat-treated fibers have an orientation angle of ca. 9°, and a tensile modulus of ca. 120 GPa.

Du Pont markets several types of fibers under the Kevlar aramid trademark that offer different combinations of properties to satisfy the requirements of a variety of end use requirements. The major products are Kevlar, Kevlar 29, and Kevlar 49 aramid fibers. The major end uses of these materials are outlined in Table 3. Selected physical properties of Kevlar 49 and of Kevlar 29, the two products of interest in terms of plastics reinforcements, are presented in Table 4. The stress – strain graph for Kevlar 29 is nonlinear. At failure, its tensile modulus is about 20 % higher than the initial value reported in the table.

Table 3. Commercial applications of aramid fibers

Material	Kevlar	Kevlar 29	Kevlar 49
Tires	×		
Mechanical rubber goods	×		
Ropes and cables		×	(×)
Ballistics protection		×	(×)
Coated fabrics		×	
Tapes and webbing		×	
Friction products and gaskets		×	
Reinforced plastics			×

Supplier: E. I. du Pont de Nemours & Co., Inc.

Table 4. Representative properties of Kevlar *p*-aramid fibers

Fiber type Nominal designation	Kevlar 29 Standard	Kevlar 49 Standard	Kevlar 49 Improved
Specific gravity	1.44	1.44	1.44
Filament diameter,			
denier	1.5	1.5 – 2.25	2.25
µm	13	13 – 16	16
Tensile modulus, GPa	62 *	131	124
Tensile strength, GPa			
Unimpregnated yarn **	2.8	2.7	
Impregnated strand ***	3.8	3.8	4.1
Tensile elongation, %	4.0	2.8	3.3
Toughness, MPa	87	55	68
Longitudinal CTE, $10\,m\,m^{-1}\,K^{-1}$		−2	
Range of available tow counts 1000 fibers/strand	0.13 – 10	0.13 – 2	2

* Initial modulus,
** As per ASTM test D885,
*** As per ASTM test D2343.

Kevlar 49, because of its higher modulus, is the member of the Kevlar aramid product family normally used in advanced composites. Although Kevlar 49 behaves elastically in tension, it exhibits nonlinear characteristics under compressive stress. The application of compressive stress results in formation of structural defects called kink bands and eventual ductile failure [40]. The onset of this behavior occurs at compressive yield strains of ca. 0.3 – 0.5 %. This phenomenon has been attributed to a molecular rotation of the amide carbon – nitrogen bond from the normal extended trans configuration to a kinked cis configuration. This results in about a 45° bend in the chain, which propagates to form a kink band across the fiber.

As a result of this unusual behavior under compression, the use of *p*-aramid fibers in applications subject to high-strain compressive or flexural loads is limited. At the same time, however, because of their ductile response under compression, *p*-aramid fibers are inherently tough materials that exhibit good damage tolerance and energy absorption characteristics.

Polyolefin Fibers. The high degree of molecular orientation that results in a well-defined crystalline structure, and the corresponding high-modulus, high-strength properties, can also be obtained by the imposition of a very high degree of plastic extensional deformation on a polymer in the solid state. When this method is applied to ultrahigh molecular mass polyethylene, which exhibits a very high degree of crystallinity, very strong and stiff fibers can be prepared [41]. Allied has developed a high-strength, high-modulus polyethylene fiber, trademarked as Spectra 900, for tensile applications such as ropes and cordage. It may also have potential use in polymer matrix composites. Room-temperature properties of this fiber are presented in Table 5. While this fiber exhibits superior mechanical properties at ambient temperature, its utility decreases rapidly with increasing temperature above 50 °C due to the relatively low melting point of the polymer. A noticeable decrease in tensile properties occurs above 50 °C. At 120 °C, the tensile strength is approximately half the value at ambient temperature, with elongation increasing proportionally.

Table 5. Representative properties of Spectra 900 polyolefin fibers

Specific gravity	0.97
Filament diameter, µm	38
Tensile modulus, GPa	117
Tensile strength, GPa	2.6
Tensile elongation, %	2.2
Toughness, MPa	29
Range of available tow counts 1000 filaments/strand	0.06 – 0.12 %

Supplier: Allied Corp.

4.1.3. Inorganic Fibers (→ Fibers, 5. Synthetic, Inorganic)

Even though the densities of inorganic fibers are significantly higher than those of carbon and polymeric fibers, these materials are of interest

in the preparation of advanced composite materials because of their stability at high temperature and their compatibility with metal and ceramic matrices, which are also inorganic materials. In addition, any listing of reinforcing fibers for advanced composite applications must include glass fibers because of their low cost and economic importance as the primary fibrous reinforcing material used by the plastics industry.

Glass Fibers. Glass fibers, also called fiberglass, have become a major product since their commercial introduction by Owens-Corning Fiberglass Corp. in 1939. Worldwide, the reinforced plastics and composites industry is currently estimated to use over 500 000 metric tons annually of glass fibers.

Glass fibers are prepared by melting a suitable mixture of raw materials (silica sand, limestone, boric acid, alumina, etc.), extruding the molten glass through a platinum spinneret and cooling the streams to form a multitude of individual glass fibers, which are then lubricated and mechanically gathered into strands. Glass fibers produced in this manner are amorphous and isotropic.

The principal grades of glass fibers used in composite applications are E-glass (E: electrical) and S-glass (S: strength). A third type, C-glass (C: corrosion), is used where resistance to acid corrosion is particularly important. The nominal compositions of E-glass and S-glass are presented in Table 6. Representative properties of these grade of glass fibers are presented in Table 7. Glass fibers generally have a high strength: weight ratio, but their moduli, which are in the range of aluminum alloys, are significantly lower than those of most of the other high-performance fibers considered in this chapter. Glass fibers are extremely sensitive to abrasion damage. They are coated with protective sizings to minimize this problem. Sizing materials also lubricate the fiber and promote its resin adhesion.

E-glass is the most common grade of fibrous glass used by the plastics industry. While originally introduced for electrical applications because of its dielectric and loss tangent properties, E-glass is used in many nonelectrical applications because of its combination of mechanical properties and low cost. E-glass is the only reinforcing fibrous material available at a price of $ 1/kg.

S-glass is a high-strength glass initially developed for military applications. Its modulus is about 20 % greater than that of E-glass, and it is about one-third stronger. The creep rupture resistance of this material is significantly better than that of E-glass, and it is about one-third stronger. It is significantly less susceptible to acid corrosion than E-glass. The failure energy of S-glass is high, making it a very tough fiber.

S-2 glass is a fiber of the same chemical composition as S-glass that was developed for commercial applications. The primary difference in the two products is the sizing on the fibers. The cost of S-2 glass fibers is, however, significantly lower than that of S-glass fibers because they are not subject to the stringent and expensive quality control procedures imposed by military procurement requirements.

High-Silica and High-Quartz Fibers. A number of high-silica (>95 % silica) and high-quartz (>99.95 % silica) fibers, which melt at a higher temperature than glass fibers, have been developed for extreme-temperature textile applications. At lower temperatures, the mechanical properties of this family of fibers are not substantially different from those of glass fibers. They exhibit superior electric properties, in terms of exhibiting an even lower dielectric constant and a lower power dissipation factor than E-glass. The high-temperature properties of silica fibers are of little value in resin matrix composites, where temperature limitations are imposed by the thermal resistance of the matrix material. Because they are significantly more expensive than glass fibers, high-silica and high-quartz fibers are rarely used as composite reinforcing fibers, except in specialized electronic applications, where an extremely low dissipation factor is required.

Alumina-Based Fibers. A number of fibers with a high alumina content (>60 wt % alumina) have been developed both for high-temperature textile applications, where they compete with high-silica and high-quartz fibers, and as reinforcements for metal matrix and ceramic matrix composites. The nominal composition and representative properties of five such materials that are currently commercially available are listed

Table 6. Chemical composition (in wt %) of inorganic oxide fibers

Material	Tradename	Suppliers	Components						
			SiO_2	Al_2O_3	FeO_3	CaO	MgO	Na_2O+K_2O	B_2O_3
E-Glass		various	53.2	14.8		21.1	0.3	1.3	9.1
S-Glass		Owens-Corning	64.3	24.8	0.2	<0.01	10.3	0,27	<0.01
Quartz			>99.95						
Ceramic	Nextel 312	3M	24	62					14
Ceramic	Nextel 440	3M	28	70					2
Alumina		Sumitomo/Avco	15	85					
Alumina	Saffil	ICI	3	97					
Alumina	Fiber FP	Du Pont		>99					

Table 7. Representative properties of inorganic oxide fibers

Material designation	Unit	E-Glass	S-Glass S2-Glass	Quartz	Ceramic	Ceramic	Alumina	Alumina	Alumina
Trade name					Nextel 312	Nextel 440		Saffil	Fiber FP
Supplier		Various	Owens-Corning	Various	3M	3M	Sumito-mo/Avco	ICI	Du Pont
Specific gravity		2.60	2.49	2.2	2.7	3.1	3.25	3.3	3.95
Filament diameter	µm	9	9	10	8 – 12	8 – 12	17	3	20
Range of available tow counts	1000 filaments/ strand	0.2 – 2	0.2	0.3	0.39	0.39	1		0.21
Price range	1985 $/kg	1	6 – 30	100 – 400	180	440	760		440
Tensile modulus	GPa	72	87	69	152	220	200	300	377
Tensile strength	GPa	3.45	4.6	0.9	1.5	1.7	1.5	2	1.4
Tensile elongation	%	4.8	5.4	1.3	1	0.8	0.8	0.67	0.36
Toughness	MPa	83	124	5.9	7.7	6.6	5.6	13.3	2.5
Dielectric constant at 10^{10} Hz		6.1	5.2	4.5	4.7				9.5
Longitudinal CTE	$10^{-6}\,\mathrm{m\,m^{-1}\,K^{-1}}$	5	5.6						6.8
Melt temperature	°C	1260		1650	1800			>2000	2045
Max. use temperature	°C	600	760	900	1200			1600	
90 % Tensile strength retention temperature	°C	220	275		1100	1400	1250		1100

in Tables 6 and 7. These fibers are generally prepared by extruding an aqueous or organic precursor gel through spinnerets, drying the resulting fibers to remove most of the liquid phase, and then subjecting the resulting filaments to high-temperature (>1200 °C) heat treatment to form a continuous refractory yarn. ICI's Saffil fibers have a δ-alumina crystal phase, whereas Du Pont's FP fibers are polycrystalline α-alumina. Saffil alumina is not available as continuous filaments, but only as fibers 1 – 5 cm long (in mat form) or less (in bulk and milled form).

The Sumitomo and 3M Nextel fibers are alumina-rich mixtures of alumina and silica and exhibit properties intermediate to those of pure silica and pure alumina fibers.

The high-alumina fibers all exhibit superior high-temperature properties, as exemplified by the fact that for all these fibers, increasing temperature from ambient to 1100 °C results in a reduction in tensile strength of less than 10 %. These fibers are claimed to be able to sustain exposure temperatures that range from 1200 °C to 1600 °C. At 1100 °C, these fibers are wetted by and stable in molten light metal alloys, such as those of aluminum and magnesium. Thus, these fibers are of interest as reinforcing materials for metal matrix composites.

Silicon Carbide Fibers. Since the 1960s, there has been significant interest in development of continuous carbide, nitride, and boride fibers and filaments as reinforcing fibers for high-temperature structural materials. Of these materials, silicon carbide fibers offer the most promise and are currently the most fully developed. Silicon carbide filaments are of interest because of their thermal stability, oxidation re-

sistance, superior mechanical properties at elevated temperatures, and compatibility with most molten metals, as well as with many ceramic materials. Silicon carbide is currently available in two different types of continous filaments, as well as in whisker form. These fibers, as well as isotropic SiC powder, are all being actively examined as reinforcements for a variety of metal matrix and ceramic matrix applications.

Because of the refractory nature of silicon carbide, silicon carbide fibers cannot be produced by consolidation from the melt or by sintering. The silicon carbide fibers that are currently available are produced by very different process routes. Avco Corp. produces large-diameter silicon carbide filaments by a vapor chemical deposition process, Nippon Carbon Corp. produces finer diameter continuous filaments by the pyrolysis of polycarbosilane polymers, and Arco Chemical Co. produces silicon carbide whiskers by the pyrolysis of rice hulls.

Avco's CVD process is very similar to the one it developed for the production of boron filaments. Boron filaments, which were major reinforcing fibers about a decade ago, have been displaced by carbon and aramid fibers in resin matrix composites, principally because these fibers are expensive ($ 700/kg) and difficult to handle. In the Avco process, silicon carbide filaments are produced by reducing a blend of methyltrichlorosilane and methylhydrogendichlorosilane with hydrogen in the presence of an electrically heated, continuously moving carbon substrate filament. Representative properties of these filaments are given in Table 8.

The original impetus for developing SiC filaments was to obtain a fiber better suited than boron for incorporation in metal matrix composites. Boron rapidly loses its strength at temperatures above 600 °C and is unsuitable for high-temperature applications. Furthermore boron reacts with molten metals, such as aluminum, which makes the fabrication of boron-reinforced metal matrix composites by liquid infiltration or standard casting techniques not feasible. Silicon fibers made by this process are potentially less expensive than boron fibers because the rate of deposition is faster and lower cost raw materials are required. The price of $ 1100/kg (1986) reflects the developmental status of these fibers. Avco projects that in production quantities, the price of these fibers would be less than $ 300/kg.

Nippon Carbon Co.'s silicon carbide filaments are produced by the pyrolysis of a carbosilane polymer precursor, a process originally developed by Yajima of Tohoku University [42]. Yajima discovered that polydimethylsilane polymers, obtained by condensing dimethyldichlorosilane in an autoclave at ca. 450 °C, could be melt spun into fibers, which when subsequently pyrolyzed at ca. 1300 °C in an inert atmosphere were transformed into high-strength, high-modulus fibers that were predominantly β-silicon carbide ($>$ 80 wt %) and carbon. In spite of their carbon content, these fibers were observed to be oxidation resistant and wetted by molten metals. It was subsequently observed that the physical properties of the silicon carbide fibers were improved by slight modifications in the composition of the polycarbosilane precursor. In particular, the addition of diphenyldichlorosilane and of boron-modified silanes to the condensing reaction mixture was observed to have a beneficial effect on fiber properties. The properties of these silicon carbide fibers, which are marketed under the trademark Nicalon, are also given in Table 8. According to the Dow Corning Co., the United States distributor of these fibers the price of these fibers in production quantities could drop from the present level of $ 500/kg to ca. $ 220/kg.

4.1.4. Reinforcement Geometry

The fiber diameter of the various fibers discussed in the previous sections of this chapter is in the 5 – 20 µm range. Exceptions to this are the silicon carbide and boron filaments produced by chemical vapor phase deposition, which have a diameter of ca. 140 µm (Avco). As a result, these filaments are neither combined in strands nor treatable by standard textile fiber processes. Fibers of small diameter range are flexible enough to be amenable to processing by many standard textile operations in spite of their high modulus. At the same time, fibers of this diameter range are too large to be aspirated into the human respiratory tract, and thus, do not present the environmental health hazards associated with fibrous materials, such as asbestos. Most of these fibers also have a circular cross-section. Notable exceptions are rayon-based carbon fibers, which are crenulated,

Table 8. Representative properties of SiC fibers

Manufacturing process		CVD	Pyrolysis
Trade name			Fiber FP
Supplier		Avco	Nippon Carbon Co./Dow-Corning
Specific gravity		3	2.55
Filament diameter	µm	140	10 – 20
Range of available tow counts	1000 fibers/bundle	not applicable	0.5
Price range	1985 $/kg	1100	500
Tensile modulus	GPa	430	190
Tensile strength	GPa	2.4	2
Tensile elongation	%	0.56	1.1
Toughness	MPa	13	11
Electrical resistivity	µΩ/m		0.1
Longitudinal CTE	$10^{-6}\,\mathrm{m\,m^{-1}\,K^{-1}}$		3.1
Melt temperature	°C	2700 *	2700 *
Max. use temperature	°C	1150	
90 % Tensile strength retention temperature	°C	800	300

* Sublimes

and certain PAN-based ultrahighmodulus carbon fibers (e.g., Celion 70), which are bilobar.

Fibers in the above diameter range are too fragile to be handled individually and are only available as multifilament strands. The reinforcing materials used in most advanced composite applications typically have a tow count of 200 – 12 000 filaments per strand. It should be noted that carbon fibers are also available in much higher tow counts, as noted in Table 1. Carbon fiber tows with a high tow count are not commonly used in the manufacture of advanced composite parts, even though they are less expensive than low end count carbon tows. In general, low end count strands are preferred as reinforcements for critical applications of advanced composite materials. The higher the tow count, the greater the variation in filament properties, and the more difficult it becomes to assure uniform contact between the matrix and the individual filaments.

Strands of reinforcing fibers can also be woven into flat fabrics, or three-dimensional woven or knitted fabrics, or even braided shapes, by slight modifications of standard textile machines. So-called structural fabrics, which contain one or more reinforcing fibers, are available in a wide variety of weaves and fabric weights.

4.2. Matrix Materials

The matrix of an advanced composite is the continuous phase that binds the fibers together so that they can act in concert. The major characteristics required of a matrix material are that it be:

1) Chemically compatible with the reinforcing fibers
2) Compatible with the manufacturing methods used to fabricate the desired advanced composite component
3) Environmentally stable under conditions of use of the advanced composite component.

The matrix material is the primary influence on mechanical properties of interlaminar shear strength and compression and flexural strength, especially at high temperature. The matrix also dictates the processibility of the composite, and its environmental resistance. Since the processibility and the environmental properties of advanced composite components depend to a significant extent on the characteristics of the matrix material, no single ideal matrix material exists that satisfies the multitude of requirements imposed by different applications. Thus, advanced composites have been prepared with a

diversity of matrix materials, which include organic polymers, carbon, metals, and ceramics.

Polymeric materials historically have been the predominant matrix materials used in advanced composites. Such as high-performance fiber-reinforced epoxy resins. Many other high-performance fibers are now combined with many other thermosetting and thermoplastic resins.

Thermosetting Resins. Thermosets are those resins which in the presence of a catalyst, heat, radiation, and/or pressure undergo an irreversible chemical reaction, or cure. Prior to cure, thermosets may be liquid or made to flow under pressure to any desired form. Once cured, they cannot be returned to the uncured state, and can no longer flow. This cure reaction can be either an addition reaction, such as the free radical polymerization of unsaturated groups, or a condensation reaction, which occurs as a result of the stepwise combination of different functional groups. If the volatile byproducts can be removed, superior properties can be attained with condensation polymers. This generally requires long, high-temperature cure cycles.

Epoxy Resins (→ Epoxy Resins). Epoxies are, by far, the matrix materials most commonly used in advanced composite materials. The main reason is the good balance of properties obtained with relative ease of handling and processing.

Epoxies are generally obtained by reacting epichlorohydrins (ECH) with molecules containing active hydrogen, such as phenols, amines or acids, followed by dehydrochlorination:

Most of the commercial epoxies are based on bisphenol A; higher functionality resins for higher temperature are based on aminophenols, diaminodiphenyl methane, triphenol methane, or novolaks.

Epoxies can be cured either by ionic homopolymerization of the epoxy groups or by reaction with suitable polyfunctional curing agents, such as primary or secondary amines, anhydrides, phenols, acids, or dicyandiamide. Some typical curing agents are:

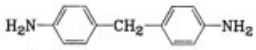

4,4′ - Diaminodiphenylmethane

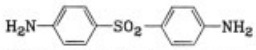

4,4′ - Diaminodiphenylsulfone

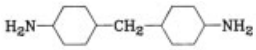

4,4′ - Diaminodicyclohexylmethane

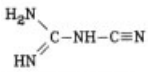

Dicyandiamide

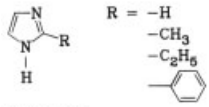

Imidazole

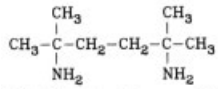

2,5 - Dimethyl hexan - 2,5 - diamine

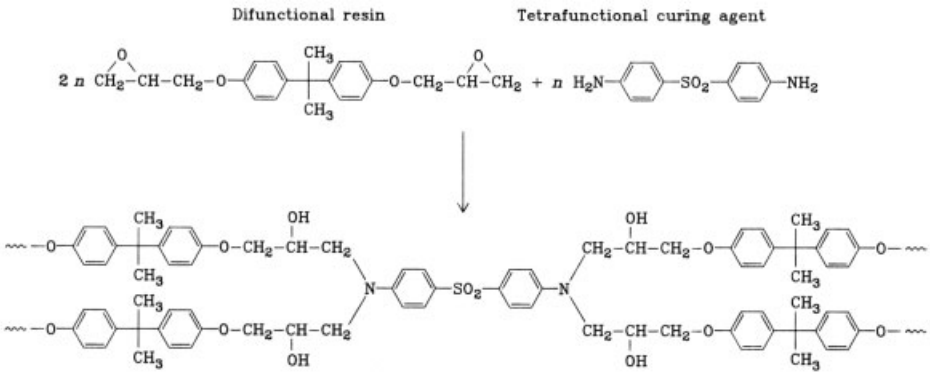

Cure and processing behavior are governed by structure and choice of components. Aliphatic amines allow ambient temperature curing, whereas slow-reacting, aromatic amines, such as diaminodiphenyl-sulfone (DDS), require a high temperature to cure.

The range of properties attainable for neat (unfilled), cured epoxy resins of different structure are summarized in Table 9. In general, increasing the concentration of aromatic rings and cross-links tends to improve resin properties such as thermal stability, chemical resistance, and rigidity, but lowers the strain to failure (increased brittleness).

Table 9. Property values of unfilled cast epoxy resin systems [43]

Property	Typical value at 23 °C (73 °F)
Specific gravity	1.2 – 1.3
Rockwell hardness	M100 – M110
Impact strength (notched bar, Izod test) J/m of notch	0.1 – 1.0
Coefficient of thermal conductivity, $Wm^{-1}\,K^{-1}$	0.17 – 0.21
Coefficient of thermal linear expansion, $m\,m^{-1}\,K^{-1}$	$(5-8)\times10^{-5}$
Specific heat, $Jkg^{-1}\,K^{-1}$	$(1.25-1.80)\times10^{3}$
Volume resistivity, $\Omega\cdot m$	$(1-50)\times10^{17}$
Dielectric constant (at 60 Hz)	2.5 – 4.5
Tensile strength, MPa	55 – 130
Tensile modulus of elasticity, MPa	2800 – 4200
Poisson's ratio	0.20 – 0.33
Flexural strength, MPa	125

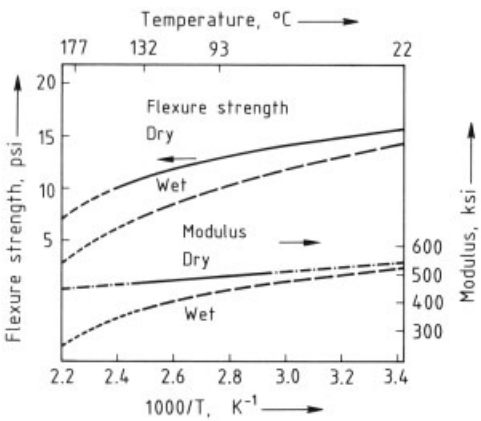

Figure 5. Environmental effects on mechanical properties of cured high-temperature epoxy resin (NEAT) [44]
System = Narmco 5208, wet: after 50 h water boil (3.1 % moisture pickup)
Note: 1 psi = 6.895 × 10^3 Pa; 1 ksi = 1000 psi

Most of the epoxy resins used by the aerospace industry in advanced composite structures, which must withstand exposure to high temperature (350 °F = 177 °C), are highly cross-linked aromatic resins. The properties of a commercial epoxy resin, which has been accepted for high-temperature composite applications, are given in Table 10. The major performance limitations of materials of this class are their low toughness and the significant loss in mechanical properties at high temperatures, particularly in the presence of moisture, as shown in Figure 5.

Bismaleimides. Bismaleimides are synthesized by the reaction of maleic anhydride with primary aromatic diamines and subsequent cyclodehydration of the intermediate thus formed. The double bonds can be reacted further by copolymerization with suitable comonomers or by a Michael-type addition of amines, for example, [48].

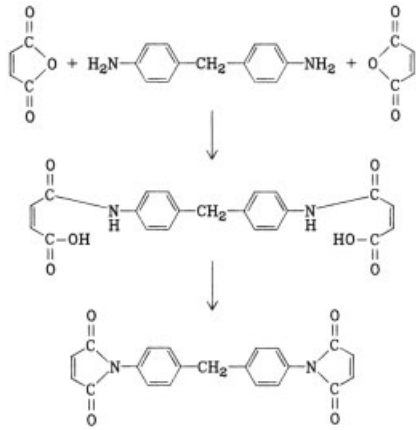

Bismaleimides offer high glass transition temperatures and good aging stability and flammability characteristics. These resins are still fairly brittle and exhibit limited high-temperature (>350 °C) stability. Significant interest in these resins exists because they may be modified or formulated to achieve better environmental resistance (e.g., better properties under hot/wet conditions) than accepted epoxy resins while retaining most of the processibility characteristics of epoxies. Representative properties of a commercial bismaleimide resin are given in Table 10.

Polyimides (→ Polyimides). Polyimides are a class of resins that can withstand contin-

Table 10. Representative properties of selected thermosetting resin matrix materials [44–47]

Polymer	Unit	Epoxy Narmco 5208	BMI * Ciba-Geigy XU-922	Polyimid Nat. Starch thermid 600	Rigid P'ester generic
Specific gravity		1.265	1.35	1.37	1.10 – 1.46
Glass transition temperature	°C	238	310	350	70 – 140
Processing temperature	°C	177	250	370	20 – 150
Mechanical properties at ambient					
Tensile strength	MPa	50	94	83	40 – 90
Tensile modulus	GPa	39	39	39	20 – 44
Tensile elongation	%	1.4	3.0	2.0	<2
Fracture toughness	kJ/m^2	0.08	0.21		
Solvent resistance		+	+	+	+
Price range	$/lb	10			1

uous exposure in air at temperatures above 300 °C. These materials have become of practical interest because they are more processible than most other resin systems that are stable at high temperature. Polyimide resins have evolved from early amic-acid type condensation products in high-boiling solvents (such as *N*-methylpyrrolidone). These were difficult to process and tended to form porous laminates because of the subsequent condensation of volatile reaction byproducts which could not be totally removed from the resin. Subsequent developments have led to resins in which final cure is obtained by addition-type reactions, which do not emit volatile byproducts [49]. Examples include PMR-type resins, initially developed by NASA Lewis Research Center. These entail the in situ condensation of an imidized prepolymer on the reinforcing fibers at a temperature of 150 – 220 °C. The composite is then cured by cross-linking via a retro-Diels – Alder reaction at 300 – 350 °C. More recently, a family of condensation polyimides with a flexible diamine has been developed by Du Pont. These materials have the potential to be prepolymerized and handled like thermoplastics; thus, they should provide a good processing balance [54].

Ethynyl-Terminated Resins. Ethynyl-terminated resins are another promising class of high-temperature thermosetting resins. A very stable thermosetting resin is obtained by the high-temperature polymerization and cyclotrimerization of terminal ethynyl groups on resin oligomers. This intramolecular cyclization (IMC) [46] results in the formation of benzene rings as the cross-link sites. Ethynyl-terminated resins have been prepared with a wide range of backbones, including polyimides and epoxies. These materials produce thermosets with high toughness and thermal and oxidative stability, low water absorption, and versatile handling and processing characteristics. Representative properties of a commercial resin of this type (National Starch and Chemical Corps., Thermid 600) are given in Table 10.

Other Thermosets. Some of the oldest and most widely used thermosets are generally not considered as suitable matrix materials for advanced composite materials. Unsaturated polyesters are obtained by condensation of unsaturated dicarboxylic acids with polyethylene or polypropylene diols. These resins are then cross-linked by radical copolymerization of the double bonds along the chain with suitable comonomers such as styrene, divinyl benzene, etc. Because of their low cost, unsaturated polyesters are extensively used in the manufacture of glass fiber reinforced plastic parts (→ Reinforced Plastics). However, this class of polymers finds little application in advanced composites because of its relatively poor thermal resistance (i.e., low glass transition temperature) and low mechanical properties, as indicated by the data presented in Table 10. Vinyl ester resins are obtained by reacting epoxies with methacrylic acid to form a bismethacrylate. Although the double bonds of this oligomer can be cured without comonomers, it is generally diluted with comonomers similar to those used in conjunction with unsaturated polyester resins. Properties and the cost of the vinyl esters generally lie between those of unsaturated polyesters and epoxy resins, and can be varied to meet specific requirements in this range. Phenolic resins are made by condensation of phenols with aldehydes, such as formaldehyde, to form mixtures of

hydroxymethylphenols. Resol or novolak type resins are obtained, depending on the stoichiometric ratio of monomers. These resins can be cross-linked by heat, by acidic catalysts, or by the addition of cross-linking agents, such as hexamethylenetetramine. During curing, the material condenses further to form a methylene-bridged three-dimensional network. Water, which is given off as a byproduct, tends to form voids, which result in relatively poor mechanical properties. However, phenolic resins are important in the manufacture of carbon fiber reinforced carbon matrix composites, which are obtained by the pyrolysis of carbon fiber reinforced phenolic resins.

Thermoplastic Resins. Thermoplastics are materials which are normally solids at room temperature, which melt or soften when heated to a sufficiently high temperature, and which become solid again when cooled. This cycle can be repeated. As a result, it is usually easier and less expensive to fabricate a complex part out of a thermoplastic resin than with a thermosetting resin. Thermoplastic resins lend themselves to rapid processing and to the formation of complex parts by processes such as injection molding, extrusion, and thermoforming. The mechanical properties at ambient temperature of many commercial thermoplastics are comparable to those of thermosets. Thermoplastics are generally more ductile and tougher than thermosets. Indeed, for many common applications, thermoplastics are the material of choice, as evidenced by the high-volume use of polyolefins, vinyls, polyamides, polyacrylics, polyesters, etc.

The ability of thermoplastics to be deformed by heat also has deleterious effects. The creep resistance of most thermoplastics at an elevated temperature is lower than that of thermosets. This has been a serious impediment to their wider use in structural applications. Furthermore, as a class, thermoplastics also tend to be more susceptible than thermosets to solvent attack. Representative properties of selected standard thermoplastics are presented in Table 11.

Short-fiber-reinforced thermoplastics have been in use for a number of years, but there have been relatively few applications of continuous fiber reinforcement. As a consequence, there has been relatively little use made of thermoplastics in the formulation of advanced composites. This situation is now changing for two reasons. First, thermoplastic materials are now available which exhibit good high-temperature properties and are reasonably resistant to solvent attack. Second, there has been a mutual adaptation of these materials and advanced composite fabrication methods. This combination has resulted in advanced composite products whose performance and cost are competitive with those made with thermoset matrices.

High-temperature thermoplastics that are being actively considered for use in advanced composites include aromatic polyesters, poly(phenylene ether), polyphenylene sulfide, polysulfone, aromatic polyetherether ketone (PEEK), and polyimides, such as ether --, ester --, or amide – imides. These materials all contain an aromatic backbone, but differ in the way the aryl groups are linked, as indicated in Table 12. Representative properties of these materials are given in Table 11. These materials can be either amorphous or crystalline. In general, crystalline polymers are more resistant to creep and solvents than the amorphous polymers, but are more difficult to process because they are more subject to shrinkage and warpage.

5. Processing and Fabrication Technology

The manufacture of advanced composite components entails the following basic operations:

1) Surrounding the fibers with the matrix material while the latter is in a nonrigid state
2) Arranging the reinforcing fibers in selected patterns and orientations
3) Solidifying the matrix material under constrained conditions that prevent fibers from moving during the solidification of the matrix
4) Integrating the advanced composite material into the finished component or system.

Most advanced composite components are made by laying up precoated fibers. However, the order of the first three steps can be changed, depending on the type of resin system and process technology selected.

Table 11. Representative properties of selected thermoplastic resin matrix material [50–53]

Polymer Supplier	Unit	Polyimide Du Pont	Polyimide Du Pont	Polyamide Du Pont	Polyamide–imide Amoco	PPS * Phillips	PEEK ** ICI	Nylon 66 Du Pont	PET *** Du Pont
Designation		K-I	K-II		Torlon C	Ryton	AFC-2		
Specific gravity		1.37	1.31	1.14	1.38	1.36	1.30	1.14	1.38
Glass transition temperature	°C	195	255	145	135	88	145	50	75
Melting temperature	°C	n.a	n.a		n.a	290	334	260	255
Processing temperature	°C	371	371	315	350	325	390	290	290
Mechanical properties at ambient									
Tensile strength	MPa	103	110	103	138	76	83	83	69
Tensile modulus	GPa	2.38	2.86	3.17	3.30	3.31	3.65	3.10	2.83
Tensile elongation	%	7	11	27	25	2–20	46	60	50
Fracture toughness	kJ/m^2	6.3	14		3.3				
Solvent resistance		+	+	−	+	+	+	+	−
Price range	$/lb	20	20	2	20	3	28	2	2

* PPS: Polyphenylene sulfide;
** PEEK: Polyetherether ketone;
*** PET: Polyethelene terephthalate

Table 12. High-temperature thermoplastics [48]

Polymer	Structure	Glass temp., °C	Melting temp., °C
Poly(phenylene ether)	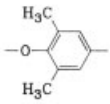	210	
Polyarylate	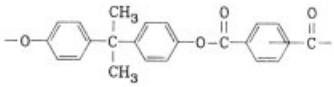	192	
Polyphenylene sulfide	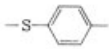	88	290
Polysulfone	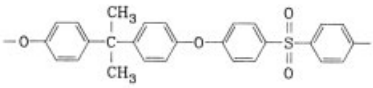	187	
Polyether sulfone	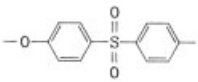	225	
Polyetherimide	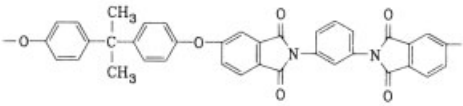	200	
Polyetherketone	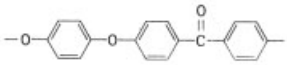	145	334

Advanced composite components can be manufactured by a variety of fabrication processes. The particular process choice made by a manufacturer depends on both the composition and the structure of the composite materials used to fabricate the part, as well as on the complexity of the design and part. Each step in the forming process is affected by a combination of fiber characteristics, the resin type, and the position in the fabrication process. The orientation process (step 2), if carried out with bare fiber, is dominated by the fiber handling properties such as yarn quality and toughness. If this step is carried out with matrix resin impregnated yarn, resin viscosity can significantly alter the process rate. In the choice of consolidation technology (step 3), resin characteristics such as cure chemistry, resin viscosity, heat transfer, and reaction kinetics will control the process. In the final step of joining the composite section to other components (step 4), the process, while affected by resin type, is primarily application specific.

It is self-evident that different processes are required to fabricate a polymeric matrix composite part and one made with a metallic matrix. In fact, even small differences in the properties of the matrix material significantly influence the choice of processing methods used to fabricate composite components. For this reason, the rest of this chapter is organized by the type of matrix material.

Processing technology choices are also influenced by non-material-product characteristics, such as part geometry, the number of parts required and their rate of production, reliability requirements, the usual economic considerations of the relative costs of capital, labor, and raw materials and the desire to minimize the total costs of producing a part to meet a given set of specifications.

5.1. Thermoset Resin Matrix Composites

5.1.1. Classic Low-Volume Production Methods

Traditional methods of fabrication of advanced composite structures in the aerospace industry have entailed cutting and manual lay-up of preimpregnated tapes of collimated fibers in a prearranged design, which are then cured by the application of heat and pressure to form the desired composite structure.

Combining the Matrix and the Reinforcement. Thermoset "prepreg" (preimpregnated) tape is the basic starting material now used in fabricating most advanced composite structures. These are most commonly unidirectional tapes that contain parallel strands of fibers that are bonded to each other by a thin film of partially cured resin. As indicated in Figure 6 nonwoven unidirectional prepreg tape is made by wetting parallel strands of filaments in a resin tank, passing the impregnated filaments through a set of doctor blades to control the resin content, collimating the filaments on a removable paper substrate, and then driving off the volatiles that are present by passing the tape through controlled-temperature ovens to attain a fiber concentration of 65 vol %. The resulting B stage product is a dry, relatively tack-free material that is relatively easy to handle. There are two categories of prepreg tapes: (1) continuous tapes, generally less than 3 in. (7.5 cm) in width, and (2) "broad goods," which are sheets whose dimensions are measured in meters.

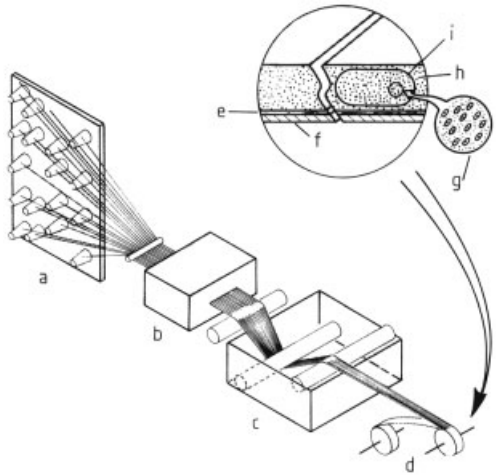

Figure 6. Typical steps in the fabrication of composite tapes a) Creel; b) Surface treatment; c) Resin; d) Backing paper; e) Carrier cloth; f) Removable backing; g) Glass or carbon filaments; h) Resin; i) Glass or carbon yarns

If the fiber can be woven, an alternate form of prepreg is obtained by drawing a woven fabric through the impregnating resin solution. The fiber content of woven prepregs is somewhat

less than that of unidirectional tape, from ca. 55 vol % to 65 vol %, depending on the fabric weave.

The various reinforcing fibers and thermoset resins discussed above have been combined by a number of suppliers to provide composite molders with a wide range of commercially available prepreg materials. While the exact compositions of prepreg materials are proprietary, prepreg suppliers certify their physical and processing properties, and their compliance to military and industry specifications and quality control standards. Specifications include the following:

1) Composition: fiber type and matrix type
2) Content: volatiles content and volume fraction of resin matrix or reinforcement in cured composite
3) Processing characteristics: shelf life as a function of storage temperature, tack (a measure of adhesion of the prepreg), flow (resin flow during curing), and gel time as a function of temperature.

Prepregs are attractive to molders of advanced composite components because they are relatively easy to handle and because their use generally results in cured laminate structures that exhibit reasonably reproducible properties. While these are significant advantages, there are also certain inherent disadvantages associated with thermoset prepregs. These include a very short shelf life (less than 24 h) at ambient conditions, necessitating refrigerated shipment and storage. A second limitation is that each prepreg material on the market is unique and generally not compatible with similar prepregs (i.e., prepregs that result in a composite with similar properties) offered by other suppliers. In general, the unit price of a prepreg is at least equal to the unit price of the contained fiber.

Fiber Orientation. The advantage of using prepreg tapes is the ability to cut them into specific patterns and then arrange the pieces individually to define the internal structure and contours of the composite part. This lay-up method, which is illustrated in Figure 7, is best suited to flat or slightly curved surfaces. More convoluted parts can be layed up by stacking woven prepregs over a suitably shaped preform. This forming method offers the advantages of versatility in selecting fiber orientation and in allowing the incorporation of different types of fiber (e.g., carbon and aramid, or aramid and glass) to achieve an optimized structure. The disadvantages of manual lay-up as a fabrication method are that it is slow, extremely labor intensive, subject to human error, and consequently an expensive operation.

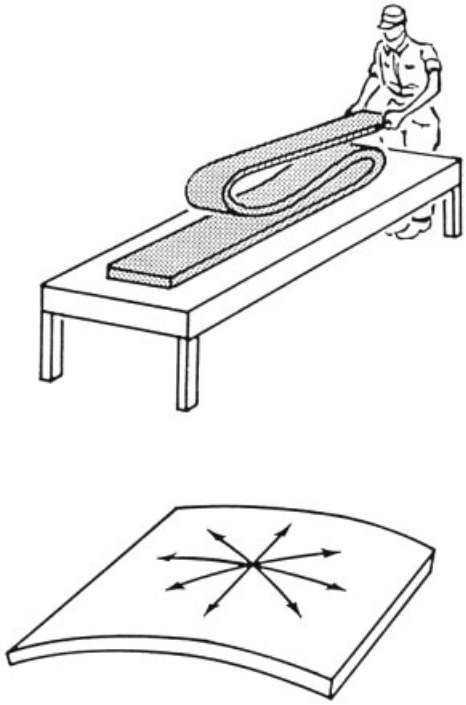

Figure 7. Elements of lay-up forming

Consolidation of the Composite. Consolidation, or curing, of the layed-up prepreg structure is most commonly achieved by the application of heat and pressure, as outlined in Figure 8. As previously noted, during the cure of thermosetting polymers, the molecules react with each other to form a more rigid network of cross-linked molecules. Epoxy matrix composites are cured at temperatures of 250 °F (121 °C) to 350 °F (177 °C), depending on their formulation. High-temperature resins, such as polyimides, require cure temperatures in excess of 500 °F (260 °C). Alternately, curing can be achieved by the application of other energy

sources, such as X-ray, microwave, or dielectric radiation.

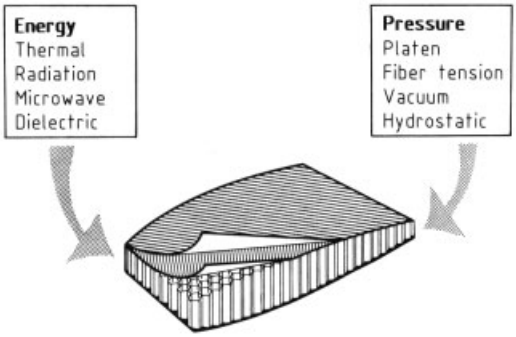

Figure 8. General methods of curing

Pressure is also applied to ensure compaction of the fibers and resin and to establish the dimensions of the part. This is most easily achieved by bag molding. In this method, the prepreg lay-up is encased in a flexible airtight sheet or blanket and a vacuum is applied to the material in the bag so that the atmosphere exerts pressure. The pressure differential may be increased by placing the bag in an autoclave. Evacuation is a useful method for reducing or eliminating volatiles prior to or during cure, thereby reducing the formation of voids in the composite structure. Bag molding is used for curing large and complex parts. However, it is a relatively slow process, which may entail the application of pressure to the part for a period of 2 – 4 h, followed by up to a day in a postcure oven.

Matched die compression molding and elastomeric tooling are also commonly used to cure prepreg lay-ups. In compression molding, prepreg is layed up in the cavity of a two-part metal mold. The mold assembly is then placed between the platens of a hydraulic press, and pressure and heat are applied until the resin cures. Elastomeric tooling is a variation of matched metal tooling, which does not require the use of a hydraulic press. In this instance, the prepreg layup is placed in a metal mold with a silicone rubber mandrel. The mold is then bolted shut and placed in an oven. Thermal expansion of the rubber provides the pressure necessary to consolidate the composite. The advantages of these methods over bag molding techniques are that they provide a rigid, predefined envelop for enclosing the composite part and that they result in shorter curing times because of higher heat transfer rates.

System Integration – Joining Methods. Not all functional characteristics required of an advanced composite material can be attained in a molded part. In spite of the fact that it is possible to mold advanced composites into very complex geometrical shapes, it is sometimes necessary to machine a cured composite material or to attach it to another material. Because of the directional characteristics of the mechanical properties of advanced composites and the brittle nature of thermoset resins, proper joint design is critical to prevent stress concentration around joints and premature failure of the assembly. In general, factors of safety used in the design of attachments of advanced composite parts are usually more conservative than those used in metallic designs.

Since thermoset resins cannot be joined by the equivalent of welding or brazing, adhesive bonding is the most commonly used method of forming permanent joints. This method is of limited value if coefficients of thermal expansion of the materials to be joined differ greatly or are dissimilar to that of the adhesive. Advanced composites can be fastened mechanically by drilling holes and joining with rivets, bolts, or pins. The fiber discontinuities caused by machining can be eliminated by molding in suitable inserts.

5.1.2. Automation of Traditional Production Methods

The major disadvantages of the traditional approach to the production of advanced composite structures described above have been its low productivity and high production costs. To overcome these economic disadvantages, while maintaining versatility in selecting fiber orientations and flexibility in terms of being able to produce a variety of composite structures in a given manufacturing facility, the aerospace industry has invested heavily in advanced manufacturing methods, such as robotics and other computer-aided manufacturing tools. Machine tool builders and material handling suppliers have developed a diversity of process automation modules to support the fabrication of composite structures, from the removal of the

prepreg from refrigerated storage to the placement of the composite part on an airframe. These modules include automated, guided vehicles to transfer material from one production site to another, integrated composite cutting centers, automated tape-laying machines, "intelligent" autoclaves which incorporate computer control of the major process parameters, automated on-line inspection methods based on ultrasonic scanning, and more recently on CAT (computer-aided tomography) scanning, and automation of secondary processes such as trimming, drilling, and fastening [55–65].

5.1.3. Alternate Production Methods

Advanced composite structures can also be produced directly from roving and resin. Filament winding and pultrusion are the most important alternate processes. These are continuous or nearly continuous automated processes that are better suited to high-volume production than the manufacturing methods based on hand lay-up of prepregs.

Filament winding is a method of fabricating hollow composite structures that have one or more axes of rotation from fiber roving or tape. As outlined in Figure 9, continuous fibers or rovings are fed from a creel, drawn through a resin bath, and wound under tension over a rotating mandrel in the shape of the finished part. By controlling the rate of rotation of the mandrel and the position of the carriage that deposits the filaments onto the mandrel, layers of reinforcing fibers can be placed in a variety of predetermined patterns over the length of the mandrel. Originally, filament winders were mechanical machines in which the relative motion between the carriage and the mandrel was determined by gearing. With these machines, wind patterns were those that had a constant wind angle and an adjustable dwell (no carriage motion while the mandrel turns). Other limitations of the mechanical machines were the difficult gearing calculations and the often lengthy gearing changes needed to change patterns. A major technological advance was the application of dedicated computer control to filament winding. Motion of the carriage relative to the rotating mandrel is controlled by software, which allows wind patterns to be modified without changing gears. There are many types and forms of filament winders available, including horizontal winders, vertical winders, two-axis winders, four-axis winders, and advanced machines with seven or more axes of controlled direction. Filament winders can be used to wind very large parts. For example, the mandrel of a large filament winding machine may be 4 m (160 in.) in diameter and 11.5 m (460 in.) long, and weigh in excess of 45 000 kg (100 000 lbs) [66].

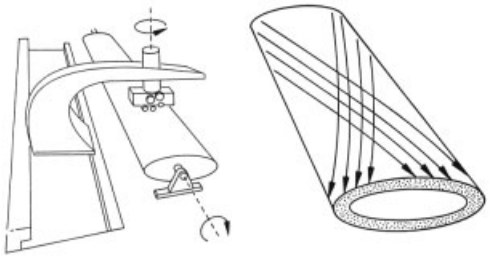

Figure 9. Elements of winding methods

Because of the need to deposit the filaments under tension and because of the rapid motion of the carriage, the filaments being wound into a composite part need to be somewhat ductile. The process has been used successfully with more ductile fibers such as aramid, high-strength carbon, and glass, which exhibit a strain to failure of 1 % or more. Very brittle high-performance fibers, such as silicon carbide and very high modulus carbon, are not compatible with filament winding.

After the filament is deposited on the mandrel, the resin is then cured either at room temperature or in an oven, depending on the size of the part and the characteristics of the resin matrix. The composite part is then stripped from the mandrel.

Pultrusion is a method of producing continuous reinforced composites of constant cross-section, as outlined in Figure 10. In pultrusion, continuous filaments are drawn through a resin bath for impregnation and then pulled through a hardened steel die. The die orients the reinforcement, sets the final shape of the laminate, and controls its resin content. Cure may be completed within the die or may require additional heating after the formed composite leaves the die. Many structural shapes such as I-beams, channels, solid rods and bars, hollow rectangular beams, and flat plates can be pultruded.

Some manufacturers have used postforming of partially cured pultruded stock to produce pultrusions of variable cross-section.

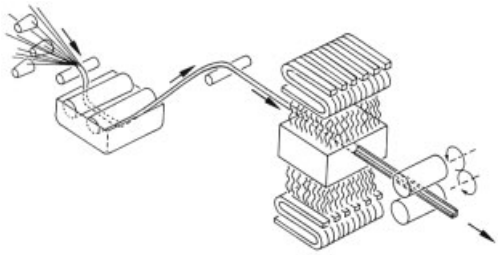

Figure 10. Pultrusion – an illustration of continuous forming and curing

5.2. Thermoplastic Resin Matrix Composites

The various production methods used to manufacture thermoset resin matrix composites are also applicable, with certain modifications, to the production of thermoplastic resin matrix composites. The key difference in the processing of these two classes of materials is in the applied time – temperature profile. Although long periods of time at elevated temperatures are required to cure properly thermosetting resin matrix composites, consolidation of thermopastic resin composites is achieved by heating to the fusion temperature of the resin and then cooling rapidly. The use of thermoplastic resin matrices eliminates the cure time delays that can stretch into days for large thermoset parts. Because of the resultant short cycle time, mass production applications of advanced composites are being enabled by the use of thermoplastic matrix materials. Other inherent manufacturing advantages of thermoplastic resin composites are that the raw materials have an inherently long storage life at room temperature, that molded parts can be postformed, and that scrap parts can be reclaimed.

Tape lay-up, filament winding and pultrusion are all adaptable to the production of thermoplastic resin matrix composites. A variety of uniform thermoplastic resin – fiber prepregs is available from commercial sources. Developmental products include tapes of glass, aramid, carbon, or other reinforcing fibers, which are combined with engineering thermoplastics such as polyetheretherketones, polyphenylene sulfide, polyamides, or thermoplastic polyester, as well as lower cost thermoplastic materials such as polypropylene, polyethylene, and ABS. These prepregs are made by a variety of processes, which include pulling the reinforcing fiber through a molten resin bath and a forming die, interweaving the reinforcing fiber with a thermoplastic monofilament of equal diameter, impregnating the reinforcement with minute particles of resin in a fluidized bed, or coextruding resin-sheathed fibers. It is also claimed that thermoplastic resin prepregs are better adapted to automatic tape laying systems than thermoset prepregs because of their lack of tack at ambient. Once the tapes are layed up, as with thermoset resin matrix composites, a thermoplastic resin matrix composite part can then be consolidated, if needed, by bag molding or compression molding, with the appropriate modifications of the time – temperature profile as mentioned above.

Thermoplastic filament winding systems are not radically different from thermosetting resin filament winding systems. The major difference is in providing means of preheating resin-coated rovings to the resin melt temperature prior to their being deposited on the mandrel. By heat loss to the surroundings the resin is resolidified and the composite structure is consolidated once the filaments are laid down. Heat sources that can be used range from ultrasonics and infrared for low-temperature materials to focused plasma guns and laser beams for high-melting engineering thermoplastic matrices.

The major difference between pultrusion systems for thermoplastic resin matrix composite materials and thermoset matrix composites is in the temperature of the resin impregnation bath and pultrusion die. Thermoplastic pultrusions are made by drawing filaments through a hot bath of molten resin and then through a chilled die, which shapes and cools the material to its final form.

5.3. Design Integration [67], [68]

The complexities of these materials force a different approach to design. The traditional design approach to any complex system was to treat it as an assemblage of relatively isolated systems to minimize interdependencies. The lim-

itations of the "open loop" process have been recognized, and a significant change is taking place. The emphasis is shifting to "configuration integration," by which technological advances, new products, or system requirements are permitted to impact the total configuration. Capabilities and advances in diverse disciplines can be coupled to achieve the valuable benefits of a multidisciplinary interaction and the full potential of these anisotropic materials. Thus, groups of independent specialists are being replaced by multidisciplinary teams.

The part performance depends not only on the geometric envelope, but also on the fiber orientation. In this way the material properties can be treated as a design parameter that can be varied to best meet the overall system requirements. The reduction to practice is best realized through the development of computer-aided design (CAD) systems that allow these very complex calculations to be made. This can then be coupled with a computer-aided manufacturing program that allows a process such as filament winding and tape lay down to be directly controlled by the design. The result is both an improved efficiency in the use of the materials and an efficient composite manufacturing process limited only by the limits of the available hardware.

6. Properties of Advanced Composite Materials

The design of load-bearing structures requires knowledge of a large number of material properties. Resistance to deformation is characterized by the stiffness of the material as described by Young's moduli, shear moduli, and Poisson's ratio. The ability to support loads is described by the tensile or compressive strength. Coefficients of thermal expansion characterize changes in dimensions under thermal loads.

The properties of the polymeric components of composite materials are dependent on the duration, rate, and frequency of applied loads or deformations. This viscoelastic behavior is translated into the response characteristics of the composite and is manifested as creep, stress relaxation, and energy dissipation, which may result in internal heating. Metals and ceramics may also exhibit such behavior near the melting point; however, at normal operating temperature, these effects are negligible and are not usually included in traditional design considerations. If the viscoelastic behavior of polymer composites is ignored, the consequences may be disastrous. Consequently, mechanical properties should be characterized as functions of time, frequency, and temperature. The dependence of polymer composites on time, frequency, and rate of loading is reflected in quantities such as fatigue life, fracture toughness, and impact strength.

In addition to mechanical requirements, composite materials may be employed in applications involving the shielding of electric or magnetic fields, electrical and heat conduction, and permeation by fluids or gases. These characteristics are described by the appropriate transport properties: dielectric constant, magnetic permeability, thermal conductivity, electrical resistivity, and diffusion coefficient.

All of the above properties may be modified by environmental effects such as infrared and ultraviolet radiation, moisture, etc.

6.1. Properties of Composite Structures

A unique feature of composite materials is the directional dependence (or anisotropy) of the various properties. The feature is intuitively clear for a unidirectional lamina; i.e., properties along the direction of the fiber reflect the higher performance characteristics of the fiber while properties perpendicular to the fiber tend toward the behavior of the matrix material. This directional dependence of the properties introduces the opportunity to further tailor the material to specific applications. The alignment of stiff and strong fibers along anticipated load paths can yield highly efficient structures.

The most common method of controlling fiber directions is shown in Figure 11. In this manual lay-up procedure, sheets of continuous fiber prepregs can be stacked in a variety of prescribed orientations. The effects of fiber orientation, developed through different stacking sequences, are illustrated in Table 13. Properties of aluminum and cold rolled steel are included for comparison.

Table 13. Effect of fiber orientation on mechanical properties of various fiber reinforced composites [69]

Material	Fiber lay-up geometry	Density	Elastic moduli, GPa			Ultimate strength, MPa			Yield strength	Fatigue strength
		g/cm^3	Longi-tudinal	Trans-verse	Shear	Longi-tudinal	Trans-verse	Shear	as % of ulti-mate strength	as % of ultimate strength
A type graphite/epoxy	unidirectional (0°)	1.57	138	6.9	4.5	1517	41	97		70
	crossply (0°, 90°)	1.57	74	74	4.5	838	838	97		59
	crossply (±45°)	1.57	17.2	17.2	31.0	138	138	345		
	isotropic (0°, 90°, ±45°)	1.57	48	48	17.8	604	604	221		78
	harness satin weave cloth (warp)	1.57	62	62		462	476			
IM type graphite/epoxy	unidirectional (0°)	1.60	221	6.9	4.8	1206	34	69		
	crossply (0°, 90°)	1.60			4.8			69		
	crossply (±45°)	1.60	17.2	17.2	44.8	124	124	290		
	isotropic (0°, 90°, ±45°)	1.60	73	73	24.8	345	345	179		
UHM type graphite/epoxy	unidirectional (0°)	1.68	303	6.9	6.6	758	28	48		
	crossply (0°, 90°)	1.68	159	159	6.6	402	402	48		
	crossply (±45°)	1.68	20.7	20.7	79.3	96.5	96.5	207		
	isotropic (0°, 90°, ±45°)	1.68	103	103	42.8	242	242	128		
Kevlar 49/epoxy	unidirectional (0°)	1.38	86 (tens) 41 (comp)	5.5	2.1	1517 (tens) 276 (comp)	28	41		70
	crossply (±90°)	1.38								
	crossply (±45°)		7.6	7.6	20.7	207	207	221		
	isotropic (0°, 90°, ±45°)									
	181 fabric (warp)(50v/e)	1.33	31	31	2.0	517 (tens) 172 (comp)	517 (tens) 172 (comp)	110		
S-Glass/epoxy	unidirectional (0°)	1.88	48	6.9	3.4	1730	40	10		
	crossply (0°, 90°)	1.88	31	31		980	900			
	crossply (±45°)	1.88	16	16		170	170			
	isotropic (0°, 90°, ±45°)	1.88	25	25		730	730			
E-Glass/epoxy	unidirectional (0°)	1.80	39	9.6	2.1	1104	20			23
	crossply (0°, 90°)	1.80	25	25		518	518		32	22
	crossply (±45°)	1.80	11	11		152	152			
	isotropic (0°, 90°, ±45°)	1.80	18	18		330	330		42	25
Aluminum 6061-T-6		2.70	72	72	26.2	310	310	207	88	31
Steel, cold rolled (0.25 carbon)		7.85	207	207	83	552	552	414	75 (tension) 60 (shear)	40–45

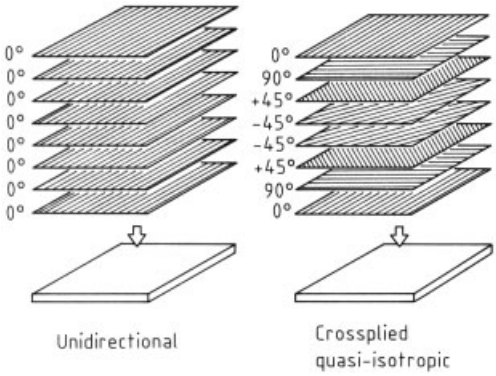

Figure 11. Representative composite lay-up configurations

To compensate for the low transverse properties of unidirectional material (0 °), laminates of unidirectional fiber may be cross-plied at right angles to each other (0, 90°). The resulting structure has improved transverse properties compared to the unidirectional structure, but poorer longitudinal properties. Furthermore, the in-plane shear strength is not significantly improved over that of the unidirectional structure. Shear strength can be improved, but at the expense of longitudinal and transverse properties, by a stacking sequence of ± 45 °. A stacking sequence of 0 °, 45 °, and 90 ° plies (also illustrated in Fig. 11) yields a quasi-isotropic structure in which the in-plane properties are independent of direction.

As shown in Table 13, woven fabric structures are similar to crossplied laminate structures; however, they are usually less stiff and strong due to the various weaving patterns. Fabric geometries offer the advantage of drapability and handling characteristics that facilitate the fabrication of geometrically complex shapes.

The structures described in Table 13 consist of a single-fiber component. However, the material properties may be further altered through hybrid composites consisting of two (or more) types of fibers in a common matrix. By judicious positioning of different fibers, it is possible to greatly expand the range of properties that can be achieved with composite materials systems [70].

The benefits of tailoring materials through control of composition and reinforcing geometry greatly complicate the direct characterization of the properties of composite structures. For homogeneous materials, such as aluminum and steel, it is feasible to develop testing programs to generate design data concerning stiffness, strength, fracture toughness, etc. In contrast, the properties of a composite structure are strongly dependent on the composition and the reinforcing geometries associated with the various methods of fabrication. The labor and expense required to acquire direct characterization data for the directionally dependent properties for each variation in composition and geometry prohibit the use of this traditional approach. Consequently, analytical techniques that predict behavior in terms of composition and reinforcing geometry have emerged as an impor-

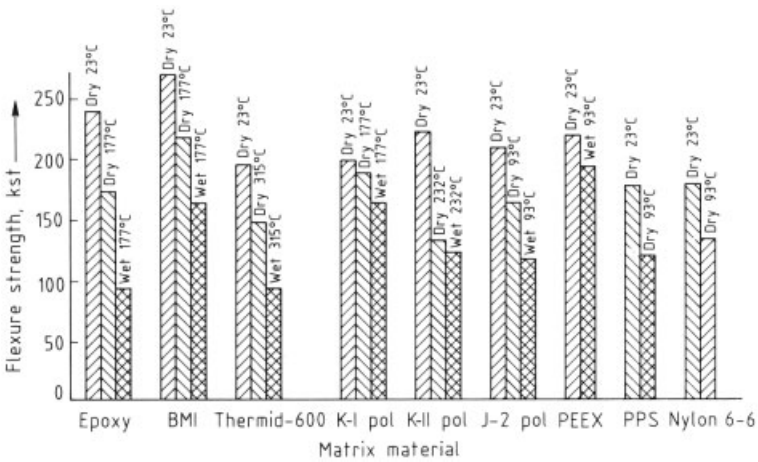

Figure 12. Effect of environment on flexure strength of unidirectional graphite reinforced composites

tant component of the technology of composite materials. These analyses proceed at two levels, laminate analysis and micromechanics, as further discussed below.

6.2. Laminate Analysis

In this treatment, as in manufacturing, the unidirectional lamina is taken as the basic structural element. The microstructure of the individual plies is ignored and each ply is treated as a homogeneous material with different properties along and perpendicular to the fiber direction. Building on techniques inherited from plywood, tires, and cloth reinforced plastics, analysts have constructed sophisticated techniques for defining the optimum orientation of plies for specific applications.

Laminate Theory. The starting point for laminate analysis is a knowledge of the properties of a unidirectional ply. The ply is treated as a homogeneous orthotropic material with different properties along and perpendicular to the fiber direction. The direction-dependent properties are formulated as matrix arrays according to the methods of anisotropic elasticity theory (an introduction to anisotropic elasticity is given in [71]). The structural member is considered to consist of variously oriented orthotropic plies, the boundary conditions on the structural member are formulated, and laminated plate (or laminated shell) theory is used to determine stresses on individual plies. Relative orientations of the plies, which best match the properties of the plies to the loading conditions imposed on the structural member, are determined.

The execution of laminate analysis requires mathematical treatments beyond the scope of this review. An introduction to laminate analysis is given in [72]. An important result of laminate analysis shows that a structure made from a sufficient number of plies stacked in the sequence illustrated in Figure 11 approaches the behavior of a quasi-isotropic material. For this structure, the in-plane properties are independent of direction. For example, the in-plane Young's modulus (E_0) is uniform and is related to the longitudinal modulus (E_L) and transverse modulus (E_r) of the ply through the following relationship: $E_0 = (3/8)\,E_L + (5/8)\,E_r$. Tsai [73] recommends that comparisons between different compositions of composite materials should be made in terms of these properties rather than the common practice of comparing longitudinal properties and ignoring the possibility that the transverse properties might not be acceptable.

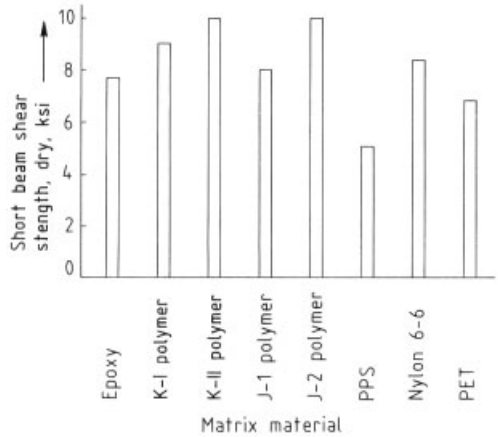

Figure 13. Effect of matrix material on interlaminar shear strength of *p*-aramid reinforced composites
1 ksi = 6.895 × 10^6 Pa

Typical Properties of Unidirectional Laminae. The application of laminate analysis requires a complete set of properties for the individual lamina. Some properties of representative, unidirectional materials are presented in Tables 14, 15, 16, 17, 18, 19. The influence of various reinforcing fibers on the room-temperature properties of otherwise comparable epoxy matrix composites is illustrated in Table 14. Conversely, the effects of various matrix materials for comparable groups of carbon fiber and *p*-aramid fiber reinforced composites are shown in Tables 15 and 17, respectively, as well as in Figures 12 and 13. The impact of various matrix materials on selected properties of carbon fiber reinforced composites is illustrated in Table 18. Tables 16 and 19 provide data on the effects of temperature and humidity on *p*-aramid and glass composites in epoxy matrix resins.

Although the utilization of laminate analyses reduces the characterization effort by focusing attention on the characterization of unidirectional specimens, the wide range of possible compositions and effects illustrated in Tables 14, 15, 17, 18 tends to preclude a total reliance on direct laboratory characterizations.

Table 14. Properties of unidirectional epoxy matrix composites with varying reinforcing fibers (normalized to 60 vol % fiber) [78]

Property	Unit	Reinforcing fibers										
		Standard carbon	1.5 % Strain carbon	1.8 % Strain carbon	Int mod carbon	HI mod carbon	UL HI mod carbon	Pitch 100 carbon	Kevlar 49 aramid	EGlass	S-2 Glass	Nicalon SiC
Specific gravity		1.58	1.60	1.61	1.60	1.80	1.83	1.83	1.45	1.90	2.02	1.90
Ultimate tensile strength	MPa	1516	1895	2584	2756	779	758	1034	1364	1034	1688	1378
Tensile modulus	GPa	131	134	138	165	239	314	420	65	41	52	110
Tensile strain	%	1.16	1.42	1.88	1.67	0.33	0.24	0.25	2.11	2.50	3.22	1.25
Ultimate compressive strength	MPa	1309	1585	1585	1378	345	338	255	207	827	827	1654
Compressive modulus	GPa	131	131	134	145	227	316	310	41	41	60	–
Transverse shear strength	MPa	65	65	69	65	34	37	34	59	–	–	–
Interlaminar shear	MPa	110	110	110	110	35	65	31	52	76	76	96

Table 15. Mechanical properties of 60 vol % Kevlar unidirectional *p*-aramid-reinforced composites of varying matrix composition [79]

Property	Unit	Thermoset resin matrix	Thermoplastic resin matrix						
			Polymer K-I	Polymer K-II	Polymer J-1	Polymer J-2	PPS	Nylon 6-6	PET
Tensile strength at 25 °C	MPa	1380	1070	945	1380	1366	1056		
at 121 °C	MPa	1172							
Compressive strength, dry									
at 25 °C	MPa	276			277				
at 121 °C	MPa	221							
Flexure strength, dry	MPa	621	642	607	676	710	566	642	634
Short beam shear, dry	MPa	53	62	69	55	69	35	58	47
Fracture toughness	kJ/m^2				1.40		0.9		

Table 16. Hot, humid aging of *p*-aramid fabric epoxy composites *

Property	Condition-ing	Tempera-ture, °C	*p*-Aramid resin		E-Glass resin	
			Narmco 5208	Hexcel F-161	Narmco 5208	Hexcel F-161
Flexural strength MPa	dry	23	434	426	500	639
		100	391	405	485	563
		160	343	345	465	530
Flexural strength MPa	wet **	23	374	341	470	466
		100	344	234	420	396
		160	174	136	315	275
Short beam shear, dry MPa	dry	23	32	37	56	53
		100	29	32	54	45
		160	25	28	46	38
Short beam shear MPa	wet **	23	31	36	54	51
		100	27	23	40	36
		160	19	13	28	26

* S-281 Fabric – Kevlar 49;
** 21-day exposure at 82.2 °C and 95 % relative humidity

6.3. Micromechanical Models

Starting ca. 1960, emphasis was given to predicting the properties of the individual unidirectional plies in terms of the properties and concentration of the fiber and matrix components. These models became a necessary supplement to laminate analysis as the choices of components became sufficiently numerous to make a strictly empirical approach economically impractical. In this approach, the inhomogeneous nature of the ply at a microscopic level is recognized. Various models have been developed by introducing approximations concerning the packing geome-

try and/or the response fields within the ply. By these models the mathematical analyses can be performed to relate the properties and concentrations of the fibers and matrix to the effective properties of the ply.

A detailed treatment of the various micromechanical models is beyond the scope and intent of this review. Introductory treatments are given in [74–76]; more advanced treatments are given in [77] and [78].

For the purposes of model development properties may be grouped under the classifications of averaged or localized properties. The class of properties within the first group can be treated in terms of average properties and average response characteristics. Thermoelastic properties, certain viscoelastic properties, and transport properties fall within this group. The class of the properties belonging to the localized group is strongly dependent on the statistical features of the material properties as well as fluctuations in microstructure; thus, properties belonging to this group are much more sensitive to the nature of the local environment. Properties such as tensile strength, fatigue life, and fracture toughness fall in this group. For such properties, the identification of modes of local failure (i.e., the weak links) is a prerequisite to the application of model relationships.

Averaged (Effective) Properties. Numerous important properties of advanced composites can be predicted by micromechanical models. They include thermoelastic properties such as Young's modulus, Poisson's ratio, shear modulus, and the thermal coefficient of expansion; transport properties such as magnetic permeability, electrical conductivity, thermal conductivity, and diffusion coefficients; as well as expansional strains induced by environmental factors such as moisture adsorption.

All of the above properties share the common feature that they can be treated in terms of average properties and the overall (average) response for the composite system. In principle, the average thermoelastic and transport properties of a ply can be obtained by specifying (1) the details of the geometry (fiber shape, packing

Table 17. Mechanical properties of unidirectional graphite-reinforced composites of varying matrix composition [79] – 82]

Property	Unit	Thermosetting resins			Thermoplastic resins							
Fiber, vol %		H.T. Epoxy	BMI	Thermid 600	Polymer K-I	Polymer K-II	Polymer J-1 *	Polymer J-2 **	PEEK (APC-2)	PPS	Nylon 6-6	PET
		AS 4 60	AS 4 65	HTS 60	AS 4 60	AS 4 60	AS 4 60	AS 4 60	AS 4 60	AS 4 60	AS 4 60	AS 4 60
Tensile strength	MPa	1860					2432		2439	2274	2322	2150
Compressive strength, dry	MPa	1378			1034		1068	1040	1040			
Flexure strength, dry	MPa	1654	1860	1344	1364	1530	1240	1447	1495	1226	1350	1220
Short beam shear, dry	MPa	110	123	83	110	96	83	103	117	83	83	83
Fracture toughness	kJ/m^2	0.26			1.05		2.10		1.75	1.40		

Table 18. Selected high-temperature mechanical properties of unidirectional graphite-reinforced composites of varying matrix composition [79–82]

Property	Temperature	Unit	Thermosetting resins			Thermoplastic resins							
			H.T. Epoxy AS 4	BMI AS 4	Thermid 600 HTS	Polymer K-I AS 4	Polymer K-II AS 4	Polymer J-1 AS 4	Polymer J-2 AS 4	PEEK (APC-2) AS 4	PPS AS 4	Nylon 6-6 AS 4	PET AS 4
	°C		60	65	60	60	60	60	60	60	60	60	60
Flexure strength, dry	23	MPa	1654	1860	1344	1364	1530	1240	1447	1495	1226	1350	1220
	93	MPa	1557			1261	1412	971	1116		827	923	951
	117	MPa	1192	1509		1123	1096						
	232	MPa					909						
	315	MPa			1020								
Flexure strength, wet ∗	23	MPa	1764		641	1426	1412						
	93	MPa	1392			1151			717	1323			
	177	MPa	648	1123		806	923						
	232	MPa					847						
	315	MPa			350								
Short beam strength, dry	23	MPa	110	123	83	110	96	83	103	117	69	83	83
	93	MPa	90			90	83	62	83		42	69	48
	177	MPa	65	82		54	63						
	232	MPa		79			56						
	315	MPa			55								
Short beam strength, wet ∗	23	MPa	103			95	86						
	93	MPa	75			73							
	177	MPa	40		53	52	65						
	232	MPa					47						

∗ 2-Week exposure at 71 °C and 95 % relative humidity

geometry, and spacing), (2) the distribution of the thermal, mechanical, or electrical loads on the surfaces of the ply, and (3) the connectivity between the fiber and the matrix phase. The average bulk response can be determined by taking the appropriate volume averages, thereby relating the volume fraction of the components and their properties to the average properties of the ply.

The simplest models for predicting thermoelastic and transport properties may be developed intuitively by treating response to mechanical, thermal, and electrical loads by analogy to springs (for mechanical loads and thermal expansion) or resistors (for thermal or electrical transport), which can be arranged either in parallel or in series. The analogies are illustrated in Figure 14.

The longitudinal response characteristics are treated as if the components of the ply react to the loads by parallel response. For mechanical loads, the central feature of this analogy is the assumption that each component is subject to the same strain. This analogy yields the following simple linear relationship:

$$P = \sum_{i=1}^{n} v_i P_i \tag{8}$$

where P_i is any one of the pertinent longitudinal mechanical, thermal, or transport properties of the ith component. The concentration terms, v_i, are the volume fractions of the ith component given by

$$v_i = V_i / V \tag{9}$$

where V_i is the overall volume of the ith component and V is the total volume of the ply. The volume fractions must satisfy the following relationship:

$$\sum_{i=1}^{n} v_i = 1 \tag{10}$$

The transverse response of the ply to mechanical and other loads is modeled by considering that the components react to these loads as though they were connected in series. The central feature of this analogy (for mechanical loads) is the assumption that each component is subject to

the same stress. This analogy yields the following simple reciprocal relationship:

$$\frac{1}{P}=\sum_{i=1}^{n}\frac{v_i}{P_i} \tag{11}$$

Table 19. Air aging of *p*-aramid unidirectional epoxy composites * at 150 °C

Hours exposure at 150 °C	Tensile strength (in MPa) ** at composite thickness		
	14 mils	42 mils	73 mils
0	1234	1207	1289
1000	1165	1165	1254
2000	1110	1138	1241
8760 (1 year)			1082
19600	676	883	

* SP-306 Epoxy, 50 vol % fiber;
** Measured at room temperature after exposure

For a two-component ($n=2$) fiber–resin composite, Eq. (10) becomes

$$v_f+v_r=1$$
or
$$v_r=1-v_f$$

and Eqs. (8) and (11) can be expressed as follows:

$$P=v_f P_f+v_r P_r$$
and
$$1/P=v_f/P_f+v_r/P_r$$

If the property of the fiber (P_f) is considerably greater than the corresponding property of the matrix (P_r), these equations are reduced to the following:

$$P\approx v_f P_f$$

for longitudinal properties and

$$P\approx P_r/v_r$$

for transverse properties. These simple relationships, which have become popularly known as the "Rule of Mixtures," indicate that for advanced composite materials in which fiber properties are considerably greater than matrix properties, the longitudinal properties of a ply are dominated by the fiber contributions and the transverse properties are dominated by the contributions of the matrix.

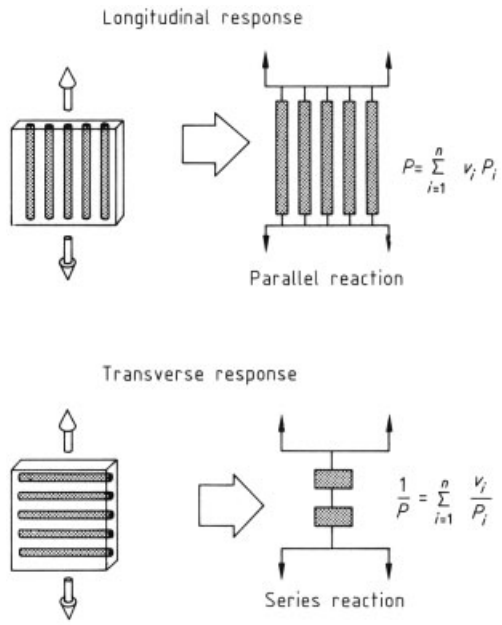

Figure 14. Simple composite models

While it has been demonstrated that the Rule of Mixtures relationship for longitudinal properties gives excellent agreement between the predicted and observed properties of continuous fiber composites, prediction of transverse properties is usually underestimated. Considerable effort has been devoted to developing more accurate and comprehensive models and methods of analysis. Improved models for predicting shear moduli and transverse properties are discussed in [72–77].

Strength of Composites (Localized Properties). In contrast to the thermoelastic properties, an understanding of the average response fields is insufficient for a description of strength. The strength of a system is dependent on the behavior at local levels, with the weaker sections of the material exerting more influence than the strong sections. Because of this weighting of the weak elements, the strength of the assembly is less than the average strength of the individual components. The appropriate weighting of the weak elements is dependent on the mechanism or mode of failure.

The weak elements or flaws can be due to characteristics either of the fibers or of the matrix. Even in the case of longitudinal ten-

sile strength, while the direct contribution of the strength of the matrix phase to the tensile strength of the composite is negligible, an improper choice for the matrix phase could seriously detract from the realization of the full strength potential of the reinforcing fibers.

The tensile, compressive, and shear strengths of a composite are determined by a complex interplay between several factors:

1) The mechanical properties of the fiber system, taking into account the statistical distribution in the tensile and/or compressive strengths of individual fibers
2) Statistical variations in the fiber – resin geometry in a ply
3) Coupling between the fiber and the resin phase, taking into account not only the bond strength at the fiber/resin interface, but also stresses induced by the difference in the responses of the fiber and resin to environmental changes; for example, a mismatch in the coefficients of thermal expansion of the fiber and resin phases can cause strains to develop near the fiber – resin interface when the composite is subjected to significant changes in temperature, as often occurs in the composite manufacturing process.
4) The viscoelastic properties of the resin phase, taking into account possible spatial variations in these properties due to presence of the fibers; one of the unique features of composite materials is the extremely large surface area over which interfacial contact is made. Because of the large surface : volume ratio and the heterogeneous nature of the matrix resins, particularly the multicomponent thermosets and semicrystalline thermoplastics, the probability is high that certain components of the resin phase are preferentially adsorbed on the fiber surface, thus resulting in spatial variations in both the composition and the properties of the matrix phase.
5) Flaws and discontinuities in the matrix phase.

In summary, the strength of a composite is determined by a complex interplay between all of the above factors; the achievement of the full capabilities potentially available from high-performance reinforcing fibers requires careful control over the extent to which each of the above factors exerts its influence. Theoretical models can offer only qualitative guidelines as to their relative significance [70–77].

7. Applications

Historically, advanced composite materials have been used for reasons of performance in applications where their value in use is the driving economic consideration. The major impetus has been their potential for reducing the mass of a structure due to their high specific stiffness and specific strength. More recently, however, the focus has shifted to the need for cost reduction through parts consolidation and total part count reduction.

Advanced composite materials have found greatest use in applications where mass savings result in significantly improved performance, thereby allowing the product to be value priced. This criterion applies to high-performance military aircraft, where lower mass translates into higher speed, greater acceleration, and/or greater payload, which are the essential elements of survival, much less victory in combat. It also applies to wide classes of equipment that use human muscle power as the motive force. Better control and ease of handling, and thus improved performance, have been the driving factors behind the use of advanced composites in sports equipment or medical prostheses.

Advanced composite materials are also being used in applications where mass reduction results in significantly lower life cycle costs. This criterion applies broadly to all forms of transportation equipment, where reduced system mass results in directly increased payload or reduced fuel consumption.

Missiles and rockets are extreme examples of transportation vehicles in which any small reduction in mass is extremely valuable. For example, the Space Shuttle System weighed 1.9×10^6 kg at liftoff. This was many times larger than the weight of the Orbiter, which was 68 000 kg, or of its low earth orbit altitude payload of ca. 30 000 kg [83]. On the basis of the cost of a Space Shuttle launch of ca. \$ 80×10^6, it cost ca. \$ 2700/kg to send an object into low earth orbit. Sending an object into geosynchronous orbit is even more constrained in terms of maximum payload, ca. 2000 kg, and costs, which range from ca. \$ 20 000 to \$ 60 000/kg.

Table 20. Reduction in lifetime fuel consumption per pound of mass savings for various modes of transportation

Mode of transport	Nominal gross mass, kg	Nominal life of equipment, km of use	Liters per km and kg of mass	Lifetime fuel reduction per kg of mass saved, L
Aircraft				
Light pleasure craft	1 400	3×10^5	1.3×10^{-4}	39
Business jet	5 500	3×10^6	1.4×10^{-4}	420
Commercial transport	95 000	3×10^7	8×10^{-5}	2 400
Turbine helicopter	1 400	1×10^6	3.3×10^{-4}	330
Automotive vehicles				
Passenger auto, gasoline	1 400	2×10^5	7.0×10^{-5}	14
Passenger auto, diesel	1 400	2×10^5	5.0×10^{-5}	10
Intercity bus	16 000	2×10^6	1.0×10^{-5}	20
Commercial truck	27 000	2×10^6	7.0×10^{-6}	14

The effects of mass reduction on the fuel consumption of various common modes of transportation are summarized in Table 20. The data presented in this table are based on published fuel consumption and vehicle utilization data. An economic value of mass savings for vehicles used to serve these various modes of transportation is obtained by assigning a net present value to the fuel that otherwise would have been consumed. Not surprisingly, mass reduction has greater value in aircraft, particularly commercial transport aircraft, than it does in ground transportation equipment.

The value of providing tailored anisotropic material properties for reasons other than mass savings is now also being recognized and is being applied in developmental aircraft, as discussed further below. Other general technical characteristics of advanced composite materials that have led to their use include corrosion resistance and their relatively high transparency to electromagnetic radiation.

In some instances, use of an advanced composite has been based on the unique characteristics of a specific fiber – matrix combination in addition to the general characteristics described above. In some applications, advanced composites reinforced with carbon fibers are used specifically because of (1) the high electrical and thermal conductivity of these fibers or (2) their unique thermal expansion characteristics, which makes these composites the materials of choice for constructing low-mass structures that do not distort under thermal gradients, such as telescope reflectors and antennae. Conversely, glass or *p*-aramid fiber reinforced composites are used in applications, such as radomes, that must be transparent to electromagnetic radiation and, thus, require a material with a low dielectric constant.

Advanced composite materials are now also used for reasons of cost as well as performance, as in the case of helicopter fuselages. In these applications, complex structures, normally made from a large number of metal parts, are now being fabricated from advanced composite materials as a single, integrated component. Cost savings derive from the smaller number of parts required and the lower assembly costs.

In the concluding sections of this chapter, a select number of representative applications of various composite materials are presented. These examples are chosen to illustrate various materials, manufacturing processes, or reasons for using advanced composites. They are not meant to be representative of all possible applications of advanced composite materials.

Satellite Systems. Satellite systems are now commonly utilized to perform a wide variety of military and civilian purposes. Commercial satellites are used for telecommunications and TV broadcasting. Government satellites are used to gather data for weather forecasting and geophysical studies, to conduct a variety of scientific experiments, and to perform a variety of military missions.

The generic technology issues and specific materials and structures criteria that address these issues are as follows [84]:

- Good dimension stability by low thermal expansion, high thermal conductivity, and no outgassing
- High structural rigidity (specific stiffness) by high modulus and low density
- Noncontaminating by no outgassing

- Rapid track and point capability by good vibrational damping characteristics
- Deployable/erectable in space by low mass (shuttle transportable) and joining and fastening capabilities
- Survivable in laser/nuclear natural environments by high strength, no induced outgassing, high-temperature integrity, controllable reflectivity, and good temperature dissipation control
- Minimal space charging by good electrical conductivity.

The specific criteria include minimum mass, low thermal distortions, well-separated natural frequencies, high damping, environmental stability, as well as good electrical conductivity and a low thermal coefficient of expansion. Carbon fiber reinforced composites with their high specific strength and specific stiffness, their conductivity, and their controlled coefficient of thermal expansion that approaches zero (or even a negative value, depending on their structure), have become the materials of choice for satellite structures.

Military Aircraft. The United States Department of Defense has supported the development and use of advanced composite materials since the early 1960s. Use of advanced composite materials has increased with each new type of aircraft. Use of advanced composite materials has progressed from replacement of metal in a few selected secondary structures, whose failure in flight would not endanger the aircraft, to designs in which advanced composites are used in primary structures critical to airworthiness and safety, and account for over 20 % of the mass of the airframe.

The first production part made of an advanced composite to reach flight status, in 1970, was a boron – epoxy horizontal stabilizer on the Grumman/U.S. Navy F-14. This 85-kg part, which is still in service, represents only 0.8 % of the structural mass of this airplane. In contrast, the McDonnell-Douglas/U.S. Navy AV-8B Advanced Harrier VSTOL aircraft contains over 450 kg of advanced composites, which is ca. 20 % of its structural mass. Major components made from carbon – epoxy composites include the forward fuselage, as well as wing substructures, such as spars, ribs, and skins, which account for over 70 % of the wing mass.

By the mid-1990s, it is anticipated that advanced performance military aircraft must be capable of maintaining supersonic flight for long missions without refueling and at the same time demonstrate superior agility, VSTOL capability, and superior firepower (e.g., high payload). To meet these requirements, aircraft size and mass will be severely constrained, and advanced composites will be extensively used. For example, not only will proposed United State Air Force Advanced Tactical Fighter (ATF) have a structure made largely (40 – 60 %) of advanced composites and, thus, weigh 25 % less than a comparable one made of aluminum or titanium, but also the total configuration of the aircraft will be influenced by advanced composite materials.

The ATF could have a forward-swept wing, which would be aerodynamically more efficient than the traditional rear-swept wing. Advantages of forward-swept wings are reduced drag, improved control, and greater resistance to spin leading to smaller engines and lighter, less costly aircraft. Despite its advantages, the forward-swept wing has been sparingly employed because the phenomenon of divergence makes it a structurally unstable design, as discussed in more detail in [85]. Conventional metallic structures of sufficient strength and stiffness to counter these divergence effects at desirable sweep angles are also very heavy, and result in unacceptable mass penalties.

The problems of divergence can be overcome by using advanced composite materials without incurring a major structural mass penalty. Since laminates can be fabricated with fiber directions selected to produce a wide variety of twisting characteristics, wing skins can be tailored to twist the leading edge down while bending, so that the wing is then structurally stable.

Radomes. Glass – epoxy and, more recently, *p*-aramid – epoxy composites are extensively used in radomes for both military and civilian aircraft. The combination of low mass, damage tolerance, and low dielectric constant (needed for signal transparency) uniquely qualifies these composites as the materials of choice.

Helicopters. Since the introduction of the first production helicopters, damage caused by fatigue of critical flight structures has had an

adverse effect on all aspects of rotorcraft operations. Thus, helicopter designers have been interested in using advanced composite materials since their commercialization, and the current designs of all major manufacturers of rotorcraft make extensive use of advanced composites for reasons of cost as well as performance.

The CH 53-E Blackhawk developed by Sikorsky for the U.S. Army was the first military helicopter to make extensive use of advanced composites. This rotorcraft contains ca. 350 kg of advanced composites, mainly *p*-aramid – epoxy composite. *p*-Aramid composites are also extensively used on the Boeing CH-47 Chinook helicopter and on the Hughes AAH-64A close ground support helicopter. The high impact resistance and toughness of *p*-aramid fibers led to these composites being preferred over carbon fiber reinforced composites, as well as over aluminum, for military helicopter structures. These properties result in a rotorcraft that is much less sensitive to damage from exploding shells and shrapnel.

Helicopter manufacturers are using the technology developed for military programs on the latest generation of commercial helicopters. For example, both the Sikorsky S-76 and the Boeing Commercial Chinook make extensive use of *p*-aramid – epoxy composites. By using *p*-aramid – epoxy composites, the frame of the S-76 weighs 30 % less than that of a structurally equivalent aluminum frame, resulting in a 20 % improvement in efficiency.

The U.S. Army has further pursued the advancement of using composites by supporting the development of the all-composite helicopter (ACAP). The Sikorsky ACAP candidate, designated the S-75, has met or exceeded all the objective and goals that the Army had set for the ACAP program. These included a mass reduction of 22 %, a cost savings of 17 %, and improved crashworthiness, survivability, reliability, and maintainability. The primary materials of construction on the S-75 are carbon and aramid reinforced epoxy.

In addition to a significant mass reduction, one of the major benefits of using advanced composites instead of metals was a sizable reduction in manufacturing costs. Previously complex metal assemblies can now be fabricated as a single, integrated composite part.

Commercial Transport Aircraft. In the past decade, the average price of aviation fuel has increased by an order of magnitude, from 10 ¢/gal to >$ 1/gal. As a result of this escalation, fuel costs have become a dominant cost factor and can now account for 60 % of the direct costs of operating a commercial transport jet. Since a marginal kilogram represents an additional consumption of 600 gal (2400 L) of fuel over the life of the aircraft (see Table 20), a strong economic incentive exists for developing more fuel-efficient transport aircraft. Technical advances that have led to more fuel-efficient transport aircraft have included improved propulsion systems, aerodynamically more efficient designs, and improved structural materials, particularly advanced composites.

Lockheed set the pace in the early 1970s by adapting 2500 lb (1140 kg) of *p*-aramid – epoxy composites on the secondary exterior structures of their L-1011-500 transport. This resulted in a structural mass reduction of 805 lb (366 kg), which contributed to the extended range of this model.

The current generation of commercial transport aircraft, such as the new Boeing 757 and 767 and the Airbus A-310, long-range, wide-bodied aircraft, makes somewhat more extensive use of advanced composite materials. In these aircraft, advanced composites are used in a variety of interior and secondary exterior structures and account for < 5 % of the structural mass. Typical applications include fairing, flaps, and parts of the empennage structure.

It is anticipated that the next generation of commercial transport aircraft will make more extensive use of advanced composites, and that these applications will include important primary structures, such as the horizontal tail plane and elevator, the fin box, and the flaps, in addition to the components already in use. Such increased use is predicated on statisfactory performance of the secondary components that are currently flying and on the development of lower cost, more damage-tolerant composite structures than the epoxy matrix materials now in use.

General Aviation. Some of the most innovative concepts for the integrated use of advanced composite materials have been developed by manufacturers of fixed-wing aircraft for

general aviation. Their incentive is to build a relatively inexpensive (ca. $\$ 1 \times 10^6$), fuel-efficient (>10 mpg), 6 – 10-passenger aircraft capable of transcontinental flight.

The Lear Fan 2100 is the first fixed-wing aircraft to make intensive use of advanced composites. This airplane has an all-composite primary structure that makes extensive use of graphite – epoxy composite materials. This aircraft is currently undergoing testing and evaluation by the FAA and is likely to be the first all-composite airplane to be certified.

The Avtec 400 represents a more radical design approach that is enabled by the use of advanced composites. This aircraft, which is currently undergoing development, uses a canard concept to improve lift and reduce drag. The skins are made of aramid – epoxy for damage tolerance, while carbon – epoxy is used in the spars and other parts that require maximum stiffness and compressive strength.

Another business aircraft of unusual design that makes extensive use of carbon – epoxy composites is the Beech Starship. This airplane has a forward wing that adjusts forward for takeoff and landing and back for flight. Beech is considering a filament wound structure.

Finally, a very special part of general aviation, human-powered flight, owes its realization to advanced composite materials. Without the availability of advanced composite materials, an aircraft light enough to be effectively powered by a human being simply could not have been built.

Ballistic Protection. *p*-Aramid-reinforced composites are being increasingly used in military helmets, for fragment protection on ships, and for military shelters. They are replacing fiberglass and metals in mass-critical applications [89]. The ability of *p*-aramid fabrics to stop high-speed objects is controlled by their dynamic energy-absorbing properties, primarily high tensile strength and specific modulus, complemented by good thermal resistance and high fracture toughness. Special woven fabrics need to be designed to balance positive and negative aspects of fabric cross-overs. Cross-overs help spread the impact by involving more yarns, but in excess, can reduce fabric deflection causing fiber shearing and loss in performance.

Automotive Transportation. Advanced composite materials have been accepted for primary body structures and chassis parts in speciality and racing vehicles, in which light mass and performance take precedence over cost. Their use in production automobiles, however, has been extremely limited for reasons of both high materials costs and lack of a suitable technology for mass production of composite parts and components. The most notable automotive application of advanced composites to date has been the continuous glass – epoxy composite transverse rear leaf spring on the Chevrolet Corvette. This single leaf spring, which weighs only 8 lb (3.6 kg), has a fatigue life that is an order of magnitude greater than that of the much heavier (45 lb, 22 kg) multileaf steel spring that it replaced [90].

This situation may change radically within the next decade because of recent developments in process and materials technology. Computer-controlled filament winding of thermoplastic matrix composites has generated major interest in the automobile industry as a means of producing large intricate parts, such as the frame, the suspension system, and the steering column. In particular, it has been reported that filament wound structural frames for cars with plastic bodies are currently under development by General Motors [91].

There is also active interest in the application of advanced composite materials to automotive engine parts. Prime candidates are those highly stressed parts that oscillate or rotate. High-speed reciprocating parts such as wrist pins, connecting rods, and valve lifters made from advanced composite materials would weigh significantly less than the metallic components now used, and their use would significantly improve engine performance, in terms of vibration, noise, and fuel efficiency.

General Industries/Electronics. While general use of advanced composite materials in the industrial sector on the basis of their mechanical properties has been limited to selected high-speed rotating or reciprocating parts, such as textile machinery components as picking sticks, use of advanced composite materials is increasing in applications that require tailored electromagnetic properties as well as superior mechanical properties. For example, X-ray ta-

bles are currently made of advanced composites because of the combination of stiffness and X-ray transparency. As the electronic industry moves to direct surface mounting of chip carriers, the ability to match the coefficients of expansion of the board and of the chip carrier, coupled with the desired electrical properties, make *p*-aramid reinforced composites attractive materials of construction for circuit boards.

Sporting Goods. The sporting goods and recreational equipment markets were the first significant commercial uses of advanced composite materials. Advanced composites have been successfully used in applications such as golf club shafts, fishing rods, tennis (and other) racquets, archery equipment, hand gliders, windsurfer boards and masts, etc. The successful penetration of advanced composites in these applications is based on psychological as well as technical reasons. While use of advanced composite materials does result in lighter mass and more responsive equipment that can marginally improve performance during play, the ego gratification that comes from using the most expensive, as well as the most technologically innovative, equipment available cannot be discounted in these markets.

Prospective Developments. In less than 30 years, advanced composite materials have evolved from their experimental beginnings to become well-established speciality products used in a wide variety of high-performance applications, where mass savings has a substantial value in use. Since a higher level of complexity is inherently needed to understand and properly use advanced composites than is needed with traditional materials, the successful development of this technology was greatly enabled by changes in the design environment brought about by the dramatic increases in computing power and computerized methods of analysis, design, and control that also occurred during this period.

The value of using advanced composite materials for reasons other than mass savings are now being recognized. These reasons include:

1) The ability to provide tailored anisotropic properties, which allows otherwise unfeasible structural design concepts to be realized by providing an additional degree of freedom in material properties
2) The ability to manufacture complex structures previously manufactured as assemblies of numerous metal parts as single, integrated components of advanced composite materials, which may result in a simpler and less expensive overall manufacturing process

Areas of future advanced composites research include the development of materials that can result in higher performance composites, as well as those that will allow advanced composites to be used more economically.

In terms of improved performance, the emphasis will be on developing composites that exhibit:

1) Superior mechanical properties at elevated temperatures
2) Improved three-dimensional properties
3) Improved toughness and ductility.

Penetration of advanced composite materials into mass production applications, in which their value in use would be modest, will depend on the development of suitable manufacturing processes, as well as lower cost materials. The most significant development in this regard may well be the filament winding of composites of high-strain fiber in a thermoplastic matrix.

8. References

1. D. Hull: *An Introduction to Composite Materials,* Cambridge University Press, Cambridge 1981.
2. S. W. Tsai, H. T. Hahn: *Introduction to Composite Materials,* Cambridge University Press, Cambridge 1981.
3. A. Kelly: *Strong Solids,* Clarendon Press, Oxford 1973.
4. A. Kelly, G. J. Davies: "The Principles of Fibre Reinforcement of Metals," *Met. Rev.* **38** (1965) 1 – 77.
5. T. W. Chou, A. Kelly: "Fibre Composites," *Mater. Sci. Eng.* **25** (1976) 35 – 40.
6. Deutsche Gesellschaft für Metallkunde (eds.): *Verbundwerkstoffe (Symposium 1980),* DGM Informationsges. mbH, Oberursel 1980.
7. D. Ruppin, K. Müller, *Z. Werkstofftech.* **12** (1981) 263 – 271.

8. H. D. Steffens, H. Kayser, K. N. Müller, *Z. Werkstofftech.* **2** (1971) 44.
9. W. Kurz, D. Sahm: *Gerichtet erstarrte eutektische Werkstoffe,* Springer Verlag, Berlin 1975.
10. H. Mangers, E. Blank, *Z. Werkstofftech.* **11** (1980) 367 – 373.
11. P. R. Sahm, U. W. Hildebrandt, *Z. Werkstofftech.* **10** (1979) 257 – 262.
12. A. Kelly, unpublished presentation.
13. C. A. Calohr, A. Moore: "No Hope for Ceramic Whiskers or Fibres as Reinforcement of Metal Matrices at High Temperatures," *J. Mater. Sci.* **7** (1972) 543.
14. P. H. Seldan: *Glasfaserverstärkte Kunststoffe,* Springer Verlag, Berlin-Heidelberg-New York 1967.
15. G. Mengers, H. Brintrup, R. Jonas, *Z. Werkstofftech.* **9** (1978) 164 – 171.
16. R. T. Pepper, G. W. Upp, R. S. Rossi, E. G. Kendall: "The Tensile Properties of a Graphite-Fiber Reinforced Al-Si-Alloy," *Metall. Trans.* **2** (1971) 117.
17. J. P. Giltrow, J. K. Lancaster: "Friction and Wear Properties of Carbon-Fiber Reinforced Metals," *Wear* **12** (1968) 91 – 105.
18. K. E. Saeger, *Aluminium (Düsseldorf)* **46** (1970) 681 – 686; see also W. Dawihl, W. Eicke, *Z. Werkstofftech.* **9** (1978) 27 – 30.
19. G. Rau: *Metallische Verbundwerkstoffe,* Werkstofftechnische Verlagsgesellschaft, Karlsruhe 1977.
20. G. Rosenkranz, *Z. Werkstofftech.* **12** (1981) 289 – 296.
21. K. H. Grünthaler, D. Harkort, G. Nixdorf, *Glas Email Keramo Tech.* **21** (1970) 313 – 324.
22. J. G. Zwissler, M. E. Fine, G. W. Groves: "Strength and Toughness of a Ceramic Reinforced with Metal Wires," *J. Am. Ceram. Soc.* **60** (1977) 390 – 396.
23. A. Neville (ed.): *Fibre Reinforced Cement and Concrete,* Construction Press, Lancaster, United Kingdom, 1975.
24. G. Batson: "Steel Fibre Reinforced Concrete," *Mater. Sci. Eng.* **25** (1976) 53 – 58.
25. J. A. Manson: "Modifications of Concretes with Polymers," *Mater. Sci. Eng.* **25** (1976) 41 – 52.
26. H. Böder, D. Gölden, P. Rose, H. Würmseher, *Z. Werkstofftech.* **11** (1980) 275 – 281.
27. R. M. Gill: *Carbon in Composite Materials,* Iliffe Books, London 1972.
28. E. Fitzer, J. Schlichting, *Z. Werkstofftech.* **11** (1980) 330 – 341.
29. R. M. Chaddell: *Deformation and Fracture in Solids,* Chap. 9: "Composites," Prentice-Hall, Englewood Cliffs 1980, pp. 252 – 270.
30. B. W. Rosen in ASM (eds.): Fibre Composite Materials, Metals Park, Ohio, USA, 1965, p. 37.
31. C. Sweben, B. W. Rosen, *J. Mech. Phys. Solids* **18** (1970) 189 – 196.
32. G. Meder, *Z. Werkstofftech.* **12** (1981) 366 – 374.
33. C. C. Chamis: "Mechanics of Load Transfer in the Interface," in E. P. Plueddemann (ed.): *Interface in Composite Materials,* Academic Press, New York-London 1974, pp. 32 – 78.
34. M. J. Folkes: *Short Fibre Reinforced Thermoplastics,* J. Wiley/Research Studies Press, London 1982.
35. T. Edison, US 223, 898, 1880.
36. W. E. Chambers, "Low Cost High Performance Fibers," *Mech. Eng.* December (1975).
37. A. Shindo, *J. Ceram. Assoc. Japan,* **69 C** (1961) 195.
38. W. Watt, W. Johnson, *Nature (London)* **220** (1960) 835.
39. E. E. Magat, *Philos. Trans. R. Soc. London A* **294** (1980) 463 – 472.
40. D. Tanner, *Future of Aramid Fiber Composites as a General Engineering Material* E. I. Du Pont de Nemours & Co., Textile Fibers Department, Wilmington, Del 19898, 1985.
41. I. M. Ward, *Philos. Trans. R. Soc. London, A* **294** (1980) 473 – 481.
42. S. Yajima, *Philos. Trans. R. Soc. London A* **294** (1980) 419 – 426.
43. L. S. Penn, T. T. Chiao in G. Lubin (ed.): "Epoxy Resins," *Handbook of Composites,* Chap. 5, Van Nostrand Reinhold Company, New York, 1982, p. 73.
44. "Rigidite 5254C, *Carbon Fiber Prepreg System*," Narmco Materials Inc., Anaheim, Calif., April 1984.
45. M. Chaudhari et al., Nat. SAMPE Symp. Exhib. Proc. **30** March (1985) 735 – 746.
46. *Technical Bulletins,* National Starch and Chemical Corp., 1985.
47. *Mod. Plast. Encycl.* **62** (1985) no. 10a, 449 – 480.
48. D. Nissen, H. Stutz, *Advanced Composite Matrices* CCM Annual Workshop, Center for Composite Materials, University of Delaware, Newark, Del. 19716, May 1985.
49. J. A. Harvey, *Proc. ICCM V,* 1985, pp. 1623 – 1632.
50. I. Y. Chang, *Composites Science & Technology,* **24,** 198561 – 79 .

51. H. H. Gibbs, *Proc. ICCM V,* 1985, pp. 971 – 993.
52. J. E. O'Connor, et al, *Proc. ICCM V,* 1985, pp. 963 – 969.
53. B. Cole, Proc. 30th National SAMPE Symposium, 799 – 808, 1985.
54. H. H. Gibbs, *Proc. ICCM V,* 1985, pp. 971 – 993.
55. G. W. Ewald, "Two Stage Tape Placement System," *Proc. Nat. SAMPE Symp.* 30 March 1985, 565 – 578.
56. E. L. Bonahan Jr., "F/A 18 Composite Wing Automated Drilling System," *Proc. Nat. SAMPE Symp.* 30 March 1985, 579 – 585.
57. R. A. Postier, "Factory Automation for Composite Structures Manufacturing," *Proc. Nat. SAMPE Symp.* 30 March 1985, 586 – 594.
58. O. Weingart, "Flexible Assembly Subsystems," *Proc. Nat. SAMPE Symp.* 30 March 1985, 820 – 831.
59. R. H. Sievers Jr., "Robotic Opportunities in Advanced Material Processing," *Proc. Nat. SAMPE Symp.* 30 March 1985, 832 – 842.
60. J. F. Wagner, "Intelligent Methodology for Automation and Robotics," *Proc. Nat. SAMPE Symp.* 30 March 1985, 843 – 854.
61. J. F. Kober, "Automated Fiber Placement – Process Creativity," *Proc. Nat. SAMPE Symp.* 30 March 1985, 1238 – 1245.
62. T. G. Gutowski, et al., "Advanced Composites Automation – Process Modelling," *Proc. Nat. SAMPE Symp.* 30 March 1985, 1275 – 1283.
63. D. J. McNally, J. R. Mitchell: "Inspection of Composite Rocket Motor Case Using AE," *Soc. Plast. Ind. Proc. Reinforced Plastics/Composites Institute* **40 Paper 5A** (1985).
64. J. W. Saveriano: "Automated Contour Tape Laying of Composite Materials," *Nat. SAMPE Tech. Conf.* 16 October 1984, 176 – 182.
65. G. Gugliotta: "Computer Aided Laser Gauging and Inspection," *Nat. SAMPE Tech. Conf.* 16 October 1984, 183 – 190.
66. T. M. Harper, J. S. Roberts: "Advanced Filament Winding Machines for Large Structures," *Society of the Plastics Industry, Proc. Reinforced Plastics/Composites Institute* **39 Paper 9E** (1984).
67. R. H. Tolson, J. Sobieszczanski-Sobieski: "Multi-disciplinary Analysis and Synthesis: Needs and Opportunities," *AIAA Pap.* **85-0584** (1985) Apr. 15, 1 – 12 (Proc. 26th Structures, Structural Dynamics and Materials Conference).
68. R. R. Emery, F. W. Crossman, *Proc.* 29th Nat. SAMPE Symp. April 1985, pp. 1465 – 1476.
69. R. Kaiser: *Automative Applications of Composite Materials,* Prepared for U.S. Dept. of Transportation, NHTSA, Report No. DOT HS-804745, July 1978.
70. H. S. Kliger: *Design of Cost Effective Hybrid Composites for Automotive Structures,* Paper 18-D, Proc. 33rd Annual Technical Conference, Reinforced Plastics/Composites Institute, The Society of the Plastics Industry, Washington, D.C., January 1978.
71. R. F. S. Hearmon: *An Introduction to Applied Anisotropic Elasticity,* Oxford University Press, London 1961.
72. J. E. Ashton, J. C. Halpin, P. W. Petit: *Primer on Composite Materials: Analysis,* Technomic Publ., 1969.
73. S. W. Tsai, N. J. Pagano: *Invariant Properties of Composite Materials,* Composite Materials Workshop, Technomic Publ., 1968, p. 233.
74. R. L. McCullough: *Concepts of Fiber-Resin Composites,* Marcel Dekker, 1971.
75. R. M. Jones: *Mechanics of Composites Materials,* Scripta Book Company, 1975.
76. R. M. Christensen: *Mechanics of Composite Materials,* J. Wiley and Sons, New York 1979.
77. C. T. D. Wu, R. L. McCullough in G. S. Holister, (ed.): *Constitutive Relationships for Heterogeneous Materials,* Chap. 7 in "Developments in Composite Materials-1," Applied Science Publishers, 1977.
78. *Composite Materials Brochure,* Fiberite Corporation, Winona, Minn, 1985.
79. I. Y. Chang, *Compos. Sci. Technol.* **24** 1985, 61 – 79.
80. M. Chaudhari et al., Proc. 30th National SAMPE Symposium, March 1985, 735 – 746.
81. *Technical Bulletins,* National Starch and Chemical Corp., 1985.
82. H. H. Gibbs, *Proc. ICCM V,* 1985, pp. 971 – 993.
83. C. Teixeira: *Space Shuttle/High Energy Upper Stage Capabilities for the 1990's,* Proc. 1982 National Telesystems Conference, IEEE Cat No 82CH1824-2, November 1982.
84. Raj N. Gounder: *Structures and Materials Technologies for Spacecraft Systems – An Overview,* Proc.29th National SAMPE Symposium, April 1984, pp. 1 – 7.
85. S. Dastin, H. L. Eidinoff, H. Armen: *Some Engineering Aspects of the X-29 Airplane,* Proc. 29th National SAMPE Symposium, April 1984, pp. 1438 – 1449.

86. R. C. Forney: *Advanced Composites,* Keynote Address, 5th International Conference on Composite Materials, San Diego, Calif., July 30, 1985.
87. Technical Symposium V, "Design and Use of Kevlar Aramid Fiber in Composite Structures," Du Pont, Wilmington, Del., March 1984.
88. Technical Symposium III, *Design and Use of Kevlar Aramid in Aircraft,* Du Pont, Wilmington, Del., October 1981.
89. D. Tanner et al., *Future of Aramid Fiber Composites as a General Engineering Material,* Du Pont, Wilmington, Del., 1985.
90. *A Leaf with Lots of Spring,* Exhibited Products Guide, 34th Annual Conference, Reinforced Plastics/Composites Institute, Society of the Plastics Industry, February 1979, p. 8.
91. Joseph A. Sneller, "Thermoplastics Open the Way to Mass-Produced RP Composites," *Mod. Plast.* **62** (1985), no. 2, 44 – 47.

Compounding of Rubber → Rubber, 5. Technology
Concrete → Cement and Concrete

Confectionery

Chocolate is a separate keyword.

Terry Richardson, Richardson Researches, Hayward, California 94545, United States

Geert Andersen, A. Storck E & S GmbH, Halle/Westfalen, Federal Republic of Germany

1. History 1
2. Raw Materials 2
2.1. Physical and Chemical Properties 2
2.2. Sweetening Ingredients 2
2.3. Other Raw Materials 3
3. Properties and Microbiology of Confectionery 4
4. Types of Confectionery 6
5. Processes and Machinery for Confectionery Production 9
5.1. Modern Developments 9
5.2. Production of Typical Confectionery Products 9
5.2.1. Fondant 9
5.2.2. Production of Fondant 9
5.2.3. Casting Fondant Articles 10
5.2.4. Soft-Boiled Sweets, Fudge, and Toffee 10
5.2.5. Forming of Bars and Small Units from Soft-Boiled Masses 11
5.2.6. Boiled Sweets, Hard Candy, Brittle, Croquante, and Butterscotch 12
5.2.7. Light Foams, Marshmallows, and White Nougat 13
5.2.8. Jellies and Gums 14
5.2.9. Panning 15
5.2.10. Other Products 16
6. Packaging, Storage, and Economic Aspects 16
7. References 16

1. History

Egyptian records describe the art of confectionery over 3000 years ago. Excavations of Herculaneum, buried under volcanic dust in A.D. 79, revealed a confectioner's workshop. The sweets of those ancient times were often based on honey, but sugar cane juices, boiled to concentrate, were also used in China and India.

The cultivation of sugar cane increased during the Middle Ages. Sugar refining was developed in Persia in the seventh century, and the growing of cane spread through southern Europe to the West Indies and, eventually, to most tropical and subtropical parts of the world. Sugar is also obtained from beet, and production from this source was accelerated in Europe by the Napoleonic Wars and by World Wars I and II.

At the same time, a search for alternative sweeteners led to the development of products from corn and other starches. These now constitute an integral part of confectionery formulations, and their production is a major industry [1–4].

2. Raw Materials

2.1. Physical and Chemical Properties

The properties of the raw materials determine the nature of the confection, its quality, and its shelf life. These include the solubility of sugars and the viscosity of their solutions, the melting points and stability of fats, and gelatinization by colloidal materials such as starch and pectin. The deterioration of sugar confectionery is in most cases due to uptake or loss of water. The reaction of milk products with sugars during cooking processes is an important factor (Maillard reaction). Microbiological activity arising from ingredients used or from external sources must be controlled to avoid spoilage.

2.2. Sweetening Ingredients

Sugar. The main sweetening substance of confectionery is sucrose. This sugar may be of cane or beet origin, and the modern commercial product has such a high degree of purity

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_411

(> 99.9 %) that its origin cannot be recognized. In addition, partially refined sugars play a useful part in the flavoring of confectionery. These are variously described as brown sugar, Demerara, molasses, and golden syrup. All are derived from cane sugar because the equivalent products derived from beet have unsatisfactory flavors (→ Sugar).

Invert Sugar. Invert sugar, a product of cane or beet sugar, is made by hydrolysis with acid or enzyme. Sucrose is a disaccharide which, when hydrolyzed, is split into the isomeric monosaccharides dextrose and fructose:

$$\underset{\text{Sucrose}}{C_{12}H_{22}O_{11}} + H_2O \longrightarrow \underset{\text{Dextrose}}{C_6H_{12}O_6} + \underset{\text{Fructose}}{C_6H_{12}O_6}$$

The reaction occurs to some extent upon prolonged boiling of a solution of sugar, but it is greatly accelerated by the presence of an acid or an acid salt such as cream of tartar (potassium hydrogentartrate). Under cold conditions, the enzyme invertase is also used to bring about inversion and is employed to soften the texture of fondant cream centers. Commercial invert sugar, a solution of dextrose and fructose of about 77 % concentration, is sometimes called artificial honey because its composition is similar to that of the natural product.

Corn Syrup, Glucose Syrup, and Liquid Glucose. Corn syrup is the name given in the United States to the product obtained by the hydrolysis of starch. Because different starches, e.g., potato, are used in other countries, glucose syrup seems to be a more appropriate and general name.

In 1811 the Russian chemist G. S. C. Kirchhoff discovered that a sweet substance could be made from starch by treatment with acid. Since then, many syrups of different viscosity, sweetness, and saccharide content, have been produced. Originally, glucose syrup was made entirely by acid hydrolysis under pressure, but now the process has been extended to a combination of acid and enzyme treatment. This has resulted in considerable variation in the proportions of dextrose, maltose, and the higher saccharides, known as oligosaccharides. A more recent development has been the isomerization of the dextrose to fructose by further enzyme treatment (→ Glucose and Glucose-Containing Syrups) [6].

Other Sugars. Dextrose is the ultimate conversion product of starch. The commercial forms are the monohydrate $C_6H_{12}O_6 \cdot H_2O$ and the anhydrous form. Dextrose has unusual crystallization properties and is less sweet than sucrose. These properties are useful in some confections. Fructose is used to a major extent for diabetic candies, because it is not obligatory to insulin.

Lactose is the natural sugar of mammalian milk and is an ingredient of confectionery in which milk products are used, e.g., caramels. Lactose is available commercially and has very low sweetness and solubility.

Sugar Alcohols. Catalytic hydrogenated glucose (sorbitol) has long been offered for diabetics as a substitute for sugar. Because of its pronounced hygroscopicity this polyalcohol is added in small amounts as a humectant for certain items, especially as hydrogenated glucose syrup.

Dental research showed that sugar alcohols, such as maltitol (hydrogenated maltose), lactitol (hydrogenated lactose) and xylitol (hydrogenated xylose) are practically indigestible and therefore noncariogenic. Consequently, confectionery manufactured with polyols was advertised as "good for the teeth" and as "light products", due to their very low calorific value. Most chewing and bubble gums contain exclusively sugar alcohols as sweetening component.

2.3. Other Raw Materials

Although sugars are the primary raw materials, other ingredients play an important part in the formulation of the variety of confections available today.

Milk. Milk in its various forms is used in the manufacture of caramel, fudge, and toffee. Liquid milk is rarely used because of its high water content (87.5 %), and concentrated or dried milk is the accepted ingredient (→ Milk and Dairy Products).

Sweetened condensed milk, skimmed milk, and increasingly whey (with lactose partly hydrolyzed to the sweeter mixture of glucose and

galactose) are considered best by most confectioners. The reason is, that mostly the sometimes poorly soluble lactose is by means of a sophisticated process better soluble (i.e. the seeding technique with ultrafine "seeding lactose" to stabilize the unavoidable supersaturation). Milk powders are sometimes not easily reconstituted, and incomplete reconstitution will result in a rough texture. Condensed sweetened milk has about 40 % added sugar; both this and milk powder may be produced in the nonfat form, in which the original milk fat has been removed. In caramel formulation, milk fat may be replaced by another vegetable fat for reasons of economy. Nonfat milk ingredients are sometimes used in other high-moisture confections where there is some danger of partial decomposition of the milk fat and the risk of rancidity.

Fat. Vegetable fats are used in many recipes and are usually refined and hardened natural oils (→ Fats and Fatty Oils). The one exception is cocoa butter, the natural fat of the cacao bean. The processing of fats is a specialized industry, and unrefined oils are unsatisfactory because of their poor stability and low melting point. Vegetable fats are produced mainly from palm kernel, coconut palm, soybean, peanut, and cottonseed oils, and their melting points are increased either by physical separation (stearines) or by hydrogenation. Palm kernel oil and coconut oil contain high proportions of short-chain fatty acids, e.g., lauric acid, and are called lauric fats. The remaining oils are then known as non-lauric and contain a larger proportion of long-chain fatty acids.

The type of fat influences the texture of confectionery. Catalytically hydrogenated hardened fats give a firm, and unhardened fats a softer quality. Consumers usually prefer the latter, but higher temperatures in summer set limits. This is especially so with lauric fats. Furthermore, enzymatic hydrolysis liberates sensorially unpleasant lower fatty acids. It is therefore better to use special palm oil fractions and (expensive) nonlauric oils from tropical palms (Shea and others). Due to their very narrow melting ranges they give similar cooling effects to unhardened cocoa nut oil.

In confectionery formulations, particularly those with high moisture, great care must be taken in the selection of other raw materials. Cocoa powder, milk powder, egg albumen, flour, and some spices, for example, must have a low microbiological count and be free from fat-splitting enzymes, e.g., lipase.

Minor Ingredients. Many other ingredients are used in compounding the very large number of confections. *Gelatinizing and aerating agents* are materials such as gelatin, agar, pectin, and egg albumen, used for making jellies and whipped or aerated products. *Starches and modified starches* are used as stabilizers and also in the preparation of a class of jellies. *Soy flour* may be regarded as a nutritional filler. *Preserved fruit, jams, and dried fruit* are often included in confectionery for their flavor, texture, and appearance. *Roasted nuts* may be added in a comminuted form, as in marzipan, or in chopped pieces. In the latter case, the confection must be one of low moisture, such as hard caramel, toffee, or nougat. In higher moisture material, like soft toffee, nuts lose their crispness because of the differences in water activity between the nuts and the matrix. *Emulsifiers* are needed to stabilize the fat dispersion when fats are included in syrup-based formulations. The main emulsifiers are vegetable lecithin and mono- and diglycerides.

The appearance and flavor of a food are of major importance. Many *flavors* are used: some are natural, such as the essential citrus oils; others, such as vanillin, are synthetic. Selection and blending of flavors constitute an art, and no food, however well prepared, will be liked if the flavor is wrong. The *color* of food plays a significant part in its acceptability, but the colorant used must conform to legal standards approved internationally or locally. Certain natural colors, such as those extracted from vegetables, have universal approval. In most countries nowadays, the list of synthetic colors allowed is relatively short. Also of importance are sweets containing health products like vitamins, mineral salts, and herbal extracts.

3. Properties and Microbiology of Confectionery

Solubility and Syrup Concentration. The solubility of substances in water varies considerably and generally increases with temperature.

Some materials are capable of forming supersaturated solutions; i.e., more substance is present in solution than would normally be present in a fully saturated solution at the same temperature. A supersaturated solution is formed by first heating the solid and solvent (sugar and water) until a concentrated solution is obtained. If this solution is then cooled without agitation and no solid is precipitated, a supersaturated solution is formed. Such solutions are unstable, and agitation will lead to recrystallization (graining).

The structure or texture of many confections is related to the solubility of the sugars and their ability to form supersaturated solutions. Crystallization may occur from some of these solutions upon agitation. This crystallization can be controlled by addition of other sugars or sugar alcohols and regulation of viscosity, to give many different types of products. Fondants and fudge, which will be described later, are typical examples.

Sugar itself, or in combination with glucose syrup in water, can be concentrated to such a high viscosity by boiling that, on cooling, a stable glasslike product is obtained. This is, in effect, a highly supersaturated solution, containing about 98 % solids and 2 % water, which is the basis of boiled sweet (hard candy) production. No change will occur in this product over long periods of time if it is stored under dry conditions at ambient temperature. The presence of glucose syrup in the formulation improves stability.

The solubility of various sugars and mixtures of sugars plays another important role. The maximum solubility of sucrose, for example, is 67 % at 20 °C, and this solution is still susceptible to microbiological decomposition by yeasts and molds. If the total sugar concentration is increased by combination with other sugars such as glucose syrup or invert sugar, much higher solubilities are ensured. This is illustrated in Table 1 [5].

High solubilities retard or prevent fermentation due to relatively higher osmotic pressure. Modern technology now recognizes that a minimum concentration of soluble solids and, consequently, a lower water activity in the syrup phase of a confection is necessary to curtail any microbial activity. The syrup phase is the liquid portion which contains some of the sucrose, all of the residual moisture, and all of the noncrystallizable carbohydrates. It may also contain a hydrated colloid, such as the milk protein used in fudge.

Water Activity. The equilibrium relative humidity (ERH), also called water activity (A/W), of a substance is the relative humidity of ambient air at which that substance will neither lose nor gain moisture. Temperature must be defined because the ERH of a substance will vary with temperature. Water activity is related to the vapor pressure of a substance, usually a solution. In confectionery, this is normally a solution of sugars, but solids such as starch also have water activity. The value depends on the molar concentration of the aqueous solution, and the values are expressed, respectively, as units or percentages: i.e., A/W 1.0 = ERH 100 % or A/W 0.6 = ERH 60 %.

Knowledge of the water activity of a confection or other foodstuff is important because shelf life will depend largely on this factor. Products with low water activity are likely to absorb moisture and become sticky or otherwise change texture. Those with high water activity will dry out under average climatic conditions.

Table 2. Water activities of some confections

Product	Water activity at 20 °C	Moisture content, %
Wafer biscuits	0.15 – 0.25	<2
Toasted cereals	0.15 – 0.25	<2
Roasted nuts	0.15 – 0.25	<2
Hard candy	0.15 – 0.25	<2
Brittles	0.15 – 0.25	2 – 3
Caramels	0.4 – 0.5	5 – 7
Ungrained nougat	0.4 – 0.5	5 – 7
Gums (jujubes)	0.5 – 0.6	10 – 15
Chewy jellies	0.5 – 0.6	10 – 15
Marshmallow – nongrained	0.6 – 0.7	15 – 18
Soft fruit jellies	0.6 – 0.75	18 – 22
Grained nougat	0.6 – 0.7	8 – 10
Fondants	0.65 – 0.75	10 – 12
Fudge	0.65 – 0.75	7 – 10
Cream pastes	0.65 – 0.75	7 – 10

The water activity of a product influences the conditions required in storage and workrooms, the type of packaging, and the combination of different confections in candy bars. For example, wafer biscuits cannot be combined with a soft fondant because the wafers become soft and the fondant dry. The water activities of some confections are given in Table 2. Each type of product varies somewhat according to moisture

Table 1. Solubilities of mixed sugars in solutions saturated with sucrose at 20°C

Proportion of solids, wt %		Solids content,
Sucrose	Glucose syrup	wt %
100	0	67.1
78.6	21.4	70.0
67.6	32.4	72.0
57.6	42.4	74.0
53.0	47.0	75.0
48.8	51.2	76.0
47.5	52.5	76.1
40.9	59.1	78.0
34.1	65.9	80.0
28.4	71.6	82.0
23.7	76.3	84.0

Solubilities of sucrose and invert sugar mixtures in saturated solutions at 20 °C

Proportion of solids, wt %		Solids content, wt %	
Sucrose	Invert sugar	Saturated sucrose solution	Saturated dextrose solution
100	0	67.1	–
78.6	21.4	70.0	–
67.6	32.4	72.0	–
57.6	42.4	74.0	–
48.8	51.2	76.0	–
47.5	52.5	–	76.1
40.0	60.0	–	73.6
30.0	70.0	–	70.5
20.0	80.0	–	67.7

content and structure. For example, two products can have the same residual moisture content, but if one is a two-phase system such as a grained confection, it would have a higher water activity because the residual moisture would be concentrated in only one part of the confection.

If the composition of a confection is relatively simple, the water activity can be calculated. In former times, desiccators with solutions of special salts, whose vapor pressure values were known, were used to measure the water activity. Nowadays, however, automatic electric devices are used to measure the water vapor contents in a hermetically sealed cell above the sample.

Microbiology. Microbiology plays an important part in confectionery technology. Sugar solutions are susceptible to deterioration by fermenting organisms or mold. Ingredients, such as egg albumen and gelatin in solution, can provide ideal media for the development of molds and bacteria. Powders, such as cocoa, spices, and cereal flours, may contain lipolytic enzymes capable of splitting fat ingredients and causing rancidity. Nuts and fruits often contain ferments and enzymes.

Some enzymes can be usefully employed. Invertase, for example, is used to prepare invert sugar and to bring about the partial liquefaction of fondant centers and to improve the moisture retention, e.g., of marzipan raw mass and formed items of ready made marzipan. Other enzymes are used to reclaim defective products in syrup form by the destruction of pectin and egg albumen residues, which hinder filtration.

Fermentation in chocolate centers was once a great nuisance. In boxed chocolates, fermentation of the fondant centers caused the chocolate covering to burst because of the evolution of carbon dioxide. Syrup exuding from the centers of only one or two pieces could contaminate the whole box. Alcoholic odors also prevailed. Isolation of the organisms responsible revealed that they were mostly osmophilic yeasts (*Zygosaccharomyces torulopsis*) which were able to ferment syrups of fairly high concentration.

While the maintenance of correct syrup phase concentration, previously mentioned, is essen-

tial and generally sufficient to prevent trouble, elimination of the virulent yeasts is also necessary. Good sanitation of plant and vessels is important. Syrup residues in pipelines and closed machinery will develop very active yeasts if the plant is used intermittently. Such machinery should be sterilized regularly with live steam and bactericidal detergent. Certain ingredients, such as preserved fruits, often contain yeasts that remain dormant until incorporated in suitable media – fondant centers, for example. Suspect ingredients must be sterilized by heat, preferably in boiling syrup.

Enzymes, particularly lipase, present a different problem. Many are active at low water activity, but their action is usually accelerated when more moisture is present. This can be a disturbing factor because action initially may be so slow that it is unnoticed, but ultimately, after several months, rancid flavors appear. This phenomenon has occurred with soft nougats containing cocoa powder and egg albumen when lauric fats were included in the formulation. Many ingredients such as cocoa powder, milk powder, and egg albumen are now guaranteed by their suppliers to be lipase free.

Methods for the detection of lipolytic activity are described in publications that also indicate the significance of microbiology in the confectionery industry [10].

4. Types of Confectionery

Some indication of the different types of confectionery has been given in the previous section. The solubility and crystallization properties of sucrose and its combination with other sugars are largely responsible for the physical differences in confectionery products. The major ingredients of all confectionery are sugars. In combination with fats, milk, gelatinizing agents, and minor substances such as emulsifiers and flavors, an extensive variety of confections is possible. The main products are hard- and soft-boiled candies, jellies and gums, foamed products, and nougat and marzipan fillings.

Boiled Sweets and Hard Candy. Hard candies and boiled sweets consist of a solution of sucrose and glucose syrup (the typical ratio is 3 : 2) boiled down to a concentration of about 98 % solids. The hot syrup is cooled to a plastic state at which stage it is flavored, colored, and formed into shapes. The final confections are in a glassy, single-phase, amorphous state, and will remain so under conditions of dry and cool storage. The high viscosity of the supersaturated liquid prevents the sucrose from recrystallizing.

Pulled Confections. The mass of hard- and soft-boiled sweets can be pulled and folded in the plastic state between 90 or 50 °C. Due to the many layers, air is incorporated in tiny bubbles, and the density is reduced drastically. The formerly glassy mass becomes opaque white. In supersaturated soft-boiled masses, this can lead to recrystallization of sucrose, which is often desirable. Pulled soft caramels show better or longer chewability. Addition of 3 – 6 % sorbitol reduces the risk of drying out on the shelf. Pulling in a humid atmosphere or storage of ready-made articles results in a sort of dry fondant, e.g., lozenges, after-dinner mints, or scotch rock.

Fondant. Fondant is a partly crystallized sugar – glucose syrup prepared by mechanical beating and cooling of syrup under controlled conditions.

The syrup is prepared by using a ratio of sugar to glucose syrup of 3 : 1 – 10 : 1, which is dissolved in water and then, to avoid any premature crystallization, carefully cooked up to a concentration of about 88 % solids. This is followed by precooling: a thin stream is continuously drawn into a cold fondant beater. Instantaneously, a creamy mass is formed due to the crystallization process. Final cooling gives a white, stiff paste, which is composed of two phases: a dry phase consisting of minute sugar crystals suspended in the wet phase of sugar, noncrystallizable carbohydrates, and water. The ratio of dry to liquid phase depends on the initial ratio of sugar to glucose syrup and moisture content. Because the moisture is in the wet phase only, the water activity of these confections is relatively high. Some air is also occluded during the beating stage.

The recrystallization of sucrose is the basis of many types of confections, especially fondant. The main factors affecting the growth and formation of crystals of pure substances are the degree of supersaturation, the amount of agitation, the temperature of the solution, the nature

and concentration of impurities (glucose syrup, colloids, etc.), and the shape and presence of existing crystal surfaces (seeds). Crystal growth is a two-step process. First, there is a mass transfer of sucrose molecules to the surface of the crystal; then these molecules aggregate at the surface of the crystal. The warmer the fondant mass becomes on reheating, the more sugar crystals disappear.

Caramels and Toffee. Caramels are relatively low-cooked (113 – 117 °C) confections (typical sugar to glucose syrup ratio of 1thinsp;: 1) [12]. They retain their shape at room temperature because of the presence of swollen milk proteins, especially caseins, and/or gelatin. Their sticky nature is rendered palatable by the presence of an emulsified fat which is also stabilized by the milk protein. Not only does the presence of hydrated milk protein prevent recrystallization of the sucrose because of its extreme viscosity, but the colloidal solution also prevents realignment of the sucrose molecules.

The syrup – milk mixture is cooked to varying degrees to give hard caramels (ca. 5 % moisture) or soft caramels (ca. 7 – 11 % moisture). Additional fat may be added, depending on the amount of fat contributed by the milk ingredient. The typical flavor and color of milk-containing caramels is a result of the Maillard reaction between the milk protein and the reducing sugars during the cooking process. Milk caramels and fruit chews are emulsions and have four phases: sugar solution, sugar or lactose crystals, fat globules, and air bubbles. Their water activity is relatively low.

Because caramels and fudges are emulsions, factors affecting the formation and stabilization of emulsions should be considered. Among these are the amount and type of agitation, the viscosity of the mass, the fat content, and the type and amount of emulsifying agents used (lecithin, milk proteins, etc.). Emulsions can be broken down by mechanical action, chemicals, and heat.

Fudge. If crystallization in caramels is initiated either by agitation or by seeding with the addition of powdered sugar or fondant, a short textured (i.e., one that breaks apart easily) paste is obtained. Fudge is like a fondant and has a dry phase of crystallized sugar and a wet phase of saturated sugar solution with some fat and milk constituents. Like fondants, fudge has a relatively high water activity and is often unstable due to insufficient emulsification. Fudge textures may vary greatly according to the proportion and size of sugar crystals formed, the dispersion of milk solids, and the fat content [7].

Foam Confectionery. This group consists of several different types of confections, ranging from very light with densities down to 0.15, up to almost regular product with 0.9 – 1.2 g/mL. Cones of foamy white mass of the light quality, covered with a thin layer of chocolate are very popular in Europe and are known as Foam Kiss (German: *Schaumkuß*). Other countries prefer heavier foams under the name of marshmallows. Other foamed items are fruit chews and white nougat (Nougat Montelimar). All these confections contain finely divided air cells dispersed throughout the syrup matrix. These air cells are stabilized by the viscosity of the cooled carbohydrate matrix and the aerating agent. In turn, the cold flow of the syrup matrix is slowed by the air cells. During the incorporation of air, the volume of the syrup increases about 250 %. Because of its increased surface area, the normally sticky syrup dissolves quickly in the mouth.

Factors affecting the formation and stabilization of foams are the type and concentration of aerating agent, the amount and type of agitation, the temperature of the syrup, the chemical and physical factors that affect aerating agents, the amount of residual moisture, the type of carbohydrates used, and the amount of air incorporated.

Nougats and marshmallows can be made in grained (controlled crystallization) or ungrained forms and, consequently, may have their carbohydrate constituents in a single phase (ungrained) or in two phases (grained). Nongrained nougats have a typical sugar to glucose syrup ratio of 2 : 3 and a residual moisture content of about 5 %. Foam confectionery of higher density can contain a small amount of fat to aid lubrication. Grained products have a higher sugar content and so a higher density. To bring about partial crystallization, normally fondant or finely ground icing sugar is used. These foams are mostly covered with fatty masses or icing to prevent drying out.

Egg albumin is used as foaming agent for light foams. For heavier ones like fruit chews and white nougat gelatin is sufficient.

Jellies and Gums. Jellies and gums are single-phase, low-boiled (105 – 108 °C), high-moisture (15 – 22 %) confections whose syrupy mass would be quite flowable, but in the presence of some colloidal setting agent, they remain dimensionally stable. The ratio of sugar to corn syrup is about 1 : 1. The recrystallization of sucrose is inhibited by the presence of hydrated colloid and the concomitant increase in viscosity.

Five main types of gelling agents are used in confectionery gels: modified starches, gelatin, agar, gum arabic, and pectins.

Each type forms its own characteristic gel. Starch forms cloudy gels with a chewy, stringy texture. High-methoxyl pectin produces clear gels with a tender short texture. Low-methoxyl pectin acts like a cross between starch and high-methoxyl pectin. Agar gels like pectin, but the gel is slightly more insoluble. Gelatin forms semiclear gels that are rubbery and elastic.

Although the mechanism of gelling is vastly different for each type, the formation of all gels comes about by loss of heat or by chemical bonding. The hydrated colloidal molecules organize and bond together at various sites to form long-chain polymers in which the carbohydrate syrup matrix is entrained and held mainly by hydrogen bonding and capillary attraction between the chains [11].

The gel strength depends on the concentration of colloid, the total solids of the solution, the competition from other sources for water of hydration, chemical bonding, and changes in pH. Because of their high moisture content, jellies have relatively high water activity and tend to dry out. Gums are dried (less than 15 % moisture) versions of starch jellies or gum arabic – gelatin jellies.

Nut Pastes. The prime examples of nut pastes are marzipan (with almonds) [13] and nut nougat (with hazelnuts/filberts). In some countries marzipan may contain other nuts or substitutes. Partly ground almonds are incorporated in a sugar – glucose or invert – sugar syrup by simple mixing and beating. Other pastes contain coconut or apricot kernels (persipan). The nut ingredients must be sterilized in the processing to destroy natural yeasts and enzymes.

Marzipan or "raw marzipan" usually consists of about two-thirds almonds and one-third sugar (invert, corn syrup, and sucrose). Defect-free blanched almonds are kibbled to produce nibs. The nibs, both sweet and bitter (approximately 1 % of total nut weight should be bitter), are blended and then mixed thoroughly with granulated sugar. At the mixing stage, it may be necessary to add water if previously dried almonds are being used (due to moisture loss in the drying process). The mixture is passed through a three-granite roll refiner and the resultant paste is cooked in steam jacketed kettles at 82 – 85 °C for about 20 min to reduce the water activity of the syrup phase to a safer level. The mass is then transferred to shallow water-jacketed pans for rapid cooling and packed after cooling.

Panned Products. Panning is a method of coating a center with layers of chocolate or sugar in a revolving pan. The centers may be nuts, preserved fruits, or other hard or soft items of confectionery. The process often lasts many hours and even days, depending on the desired amount of shell material. A precoating is necessary so create good adhesion between the nuts and the subsequent layers. It must prevent fat reaching the surface and guard against the edges of the pellets breaking or being eroded during the process. Due to the microcrystalline state of the outer sugar coating, such products are especially stable to humidity and have a long shelf life. The coating is usually glazed, which gives a protective quality as well as enhanced appearance.

Pralines, Truffles, and Noisettes. Numerous varieties of pralines, truffles, and noisettes are possible. Many rely on the roasting of nuts in high-boiled (149 – 157 °C) sugar syrup, followed by cooling and grinding to a paste. The natural oil of the nuts plus the very low moisture content of the syrup produce an attractively flavored paste. Incorporation of chocolate, other fats, and alcoholic liqueurs is possible.

Lozenge Pastes and Cream Pastes. Lozenge pastes and cream pastes are simple formulations consisting of a mixture of fine sugar with a solution of a gum (usually gum arabic or tragacanth) to produce a stiff paste

with the consistency of bakers' dough. This is stamped out into a variety of shapes and dried to less than 1 % moisture. Cream pastes are wetter versions of this type, usually made with gelatin and some reducing sugars. These pastes can be layered with other confections such as licorice or low-moisture jellies.

Licorice. Licorice is a special product with a very popular, unique flavor. All formulations are based on an extract of the root of the licorice plant, *Glycyrrhiza glabra*, which grows in the subtropical regions of Asia and Europe. Licorice confections generally take the form of stiff gelatinous pastes which can be partially dried and formed into layers or strips. In some countries, regulations now require health warnings, limited consumption, or reduction of the content of raw liquorice. This is because of the content of glycyrrhizine which can cause high blood pressure.

5. Processes and Machinery for Confectionery Production

5.1. Modern Developments

Confectionery at one time was an art and, within the realm of the experimentalist, remains so to this day. Not so long ago, in some factories skilled employees customarily made their own modifications of recipes as they went along. To some extent, this was necessary because the raw materials they worked with were by no means consistent.

As the size of factories increased along with the costs of raw materials and labor, more scientific and mechanical methods became necessary. The creative art of the experimentalist then had to be transferred to economic production without loss of quality. Initially, this meant increased batch size with mechanical handling and, sometimes, automatic weighing. Batch production is now giving way to continuous production, in which machinery from other industries is often adapted to confectionery processes.

5.2. Production of Typical Confectionery Products

5.2.1. Fondant

Fondant is a relatively simple primary product prepared from sugar and glucose syrup, with sorbitol or invert sugar included occasionally. By means of a secondary process of partly remelting and the inclusion of other ingredients such as preserves, flavors, and egg frappé, fondant can be transformed into a large variety of confectionery pieces suitable for coating with chocolate.

5.2.2. Production of Fondant

The basic solution has a concentration of about 80 % solids and is then further concentrated to 88 % by boiling to 117 – 119 °C. The syrup is cooled to 38 – 42 °C; it is now a supersaturated solution and, in this condition, is delivered to a beater – cooler, known as tablier machine. Beating is performed in a water-cooled tube with rotating internal spindles fitted with intermeshing pegs. Extreme agitation deposits excess sugar from solution in the form of minute crystals. This causes the evolution of much latent heat, which must be removed to avoid redissolution of the sugar. The sugar crystals in the cooled fondant from the beater should have a maximum diameter of 10 – 15 µm. At one time, this process was carried out manually by vigorous mixing of the syrup on a large marble table with a spadelike implement – a laborious task. Large machines are now employed, with outputs in excess of 1000 kg/h.

The sugar – glucose syrup is usually concentrated by means of a scraped-surface, thin-film, continuous cooker. Cooling is accomplished either by a large, rotating drum that is internally water cooled (Fig. 1) or by a water-jacketed tube with an internal screw.

5.2.3. Casting Fondant Articles

To prepare small fondant pieces for subsequent coating, the fondant must be partly remelted so that it can be cast into molds or chocolate shells. The conditions for remelting must ensure that only a small percentage of the crystals are dissolved. This process consists of heating

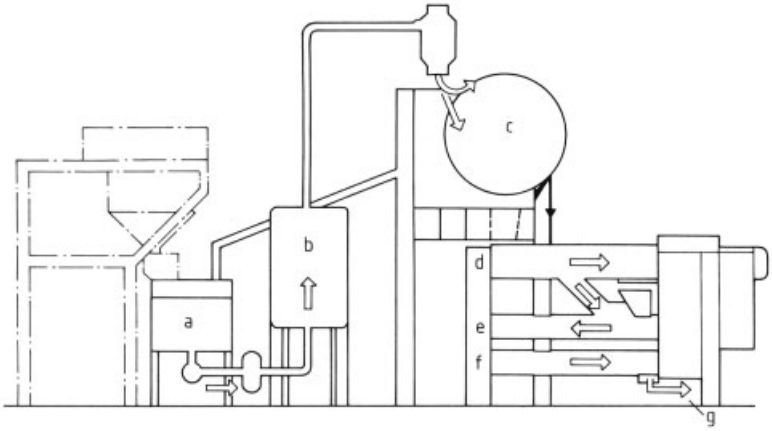

Figure 1. Baker – Perkins fondant machine
a) Premixer; b) Cooker; c) Cooling drum; d) Fondant beater; e) Creme mixer; f) Second stage mixer (color, flavor); g) Discharge

the fondant with a small amount of syrup of the same composition as the fondant or preferably sorbitol. The maximum temperature should be 65 °C. This will give the required fluidity without destroying the crystal phase.

This process temporarily lowers the viscosity and permits depositing into molds. Upon cooling, the excess sucrose in the liquid phase recrystallizes onto the existing crystal sites and cements them into a rigid structure. This occurs especially at the surface, where more evaporation has taken place, and permits the cast fondants to be handled. The remelting can be carried out in a steam-jacketed kettle or continuously in a tube provided with a heating and mixing system. At the same time, flavors, colors, fruit acids, and preserves are included. The established method of casting is to use starch molds. The large automatic machine developed for this purpose, called a mogul, is shown in Figure 2.

In this machine, shallow trays are filled with dried starch and printed by pressing a multiple former into the starch, which retains the impression of the shape of each protrusion when the printer is removed. By means of a hopper and multiple-pump system, the correct amount of melted fondant is deposited into each impression. Trays containing the starch and deposited pieces are removed automatically, stacked, and stored until the fondant has set. The starch trays containing the set fondant are fed to the front of the mogul, which empties the trays through a sieving and brushing system that completely removes the starch. The pieces of fondant are then delivered to trays and, subsequently, to a chocolate enrober for covering.

During the sieving process, the separated starch is returned to a drier and then used to refill the trays. The mogul has the disadvantage of being a very dusty operation, which leads to difficulty in maintaining a high degree of sanitation in the production area.

Modern developments utilize other methods of forming such as extrusion or molding. The use of metal molds was developed by Baker Perkins in conjunction with Cadbury. Each mold impression is coated with a release agent and provided with minute holes. After depositing, the molds are cooled and the pieces are ejected onto a belt by compressed air. They are then fed directly to an enrober. The rubber mold system uses flexible silicone rubber molds. Silicone rubber has exceptional nonstick properties and is very heat resistant. Continuous depositing, cooling, and demolding plants are now used for toffees, caramels, and some hard candies.

5.2.4. Soft-Boiled Sweets, Fudge, and Toffee

Sugar, glucose syrup, invert sugar, sorbitol, milk products, fat, and emulsifiers are the essential ingredients of this group.

For the production of milk caramel the ingredients are mixed with sufficient water and boiled over a gas burner or in a steam-jacketed kettle. Soft caramels are boiled to 118 °C, medium to

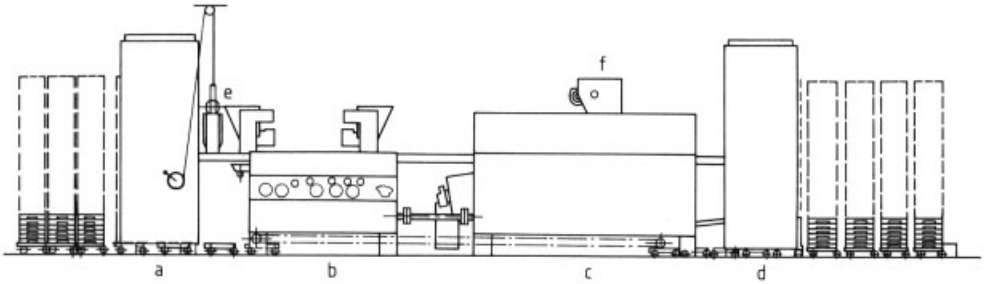

Figure 2. NID high speed mogul molder
a) Stacker; b) Depositor; c) Starch molder; d) Feeder; e) Suspended starch sprinkler; f) Starch sifter

122 °C, and hard to 130 °C; however, these figures are subject to some variation. In pan boiling, flavor develops over a period of 15 – 20 min and arises from the Maillard reaction between the milk protein and reducing sugars. The milk ingredient is important, and sweetened condensed milk, skimmed milk and whey are favored by most manufacturers.

Milk powder is also used, but this must have high solubility and be adequately hydrated in water before use. Modern cooking methods are continuous, but when cooking time is much reduced, a caramelizer must be included in the plant. This consists of a vessel in which the cooked caramel is held at a temperature of about 120 °C for 15 min. Cooking time in a continuous cooker can be as little as 1 min, which is insufficient for Maillard flavor development.

Many continuous cooking plants exist; most are based on the heating of a thin film to concentrate and cook the original mixture quickly. Immediately after cooking the product must be cooled by discharge onto a metal table, a water-cooled belt, or a cooling drum. Fruit chews do not normally contain milk constituents but gelatin. They are often aerated by pulling or adding frappe down to densities of 0.8 – 1.2 g/mL.

To make fudge, caramel is produced in the manner described above and then cooled to 80 °C. Fondant is added in the proportion of about 10 % and mixed in well. This mixture is discharged onto a table and allowed to cool until it sets to a short texture because of the gradual formation of a sugar crystal lattice.

In continuous fudge manufacture, the fondant may be extruded into a trough in which the caramel is being mixed and discharged. In the United States toffee is a high-cooked version of caramel but with a lower milk content. In Europe toffee is a chewable soft candy, rich in fat and milk constituents. Toffee is a high-cooked version of caramel but with a lower milk content.

5.2.5. Forming of Bars and Small Units from Soft-Boiled Masses

Soft milk boilings, fruit chews, and fudge can be formed into pieces by cutting, extrusion, or a combination of both. They can also be formed by the batch roller and rope-forming system described for boiled sweets, or it may be deposited into the silicone rubber molds mentioned under fondant. The principle of extrusion has been applied a great deal to confectionery products in recent years, particularly to plastic materials such as caramel, fudge, nougat, and fondant.

Extrusion machines operate by means of fluted rollers which are fed by a hopper and force the material through multiple orifices. The extruded confection may be sliced off into small pieces by wire or knife just beneath the orifices, or it can be allowed to form ropes. The ropes are subsequently guillotined into bars for feeding to a chocolate enrober.

5.2.6. Boiled Sweets, Hard Candy, Brittle, Croquante, and Butterscotch

Boiled sweets mostly have a relatively simple formulation consisting of sugar and glucose syrup. Some, however, contain milk product, cream, licorice, etc. High-quality sugar must be used. The glucose syrup varies in type and proportion according to the method of cooking and

the kind of hard candy required. Cooking may be by open pan, thin-film continuous cooker, or vacuum cooker (Fig. 3).

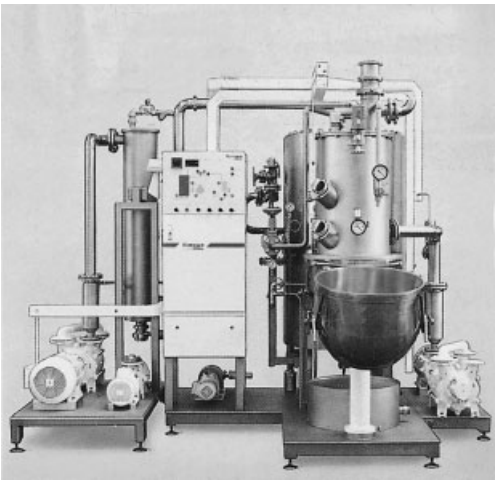

Figure 3. Vacuum cooker for hard-boiled sugar masses

In small-scale production, a gas-heated or steam-jacketed pan may be used. The batch is boiled to ca. 150 °C under atmospheric pressure or cooler under vacuum. Care must be taken to avoid any source of graining. The batch is poured onto a cooling table where it reaches a plastic condition; the outer edges of the batch are turned inward to give even cooling. Flavor, color, and acid are added to the center of the batch and incorporated by folding and turning; the mass must retain an even temperature and plastic condition. The whole of the material is then transferred to a group of machines consisting of a batch roller, a rope sizer, a former, and a cooler (Fig. 4).

The batch roller reduces the large piece to a rope which is fed through the sizing rollers to yield a uniform diameter. This rope is then delivered to a forming press which shapes and discharges individual pieces according to the design or the dies. These pieces pass along a conveyor supplied with cooling air. In large-scale production, syrup is boiled either by thin-film cookers or in vacuum pans and is conditioned prior to shaping by mechanical kneaders or water-cooled conveyors provided with plows. Flavors and colors can be fed into the syrup in continuous fashion. Some boiled sweets may be deposited into metal molds which are cooled; the pieces are removed by ejection pins in the mold. An interesting process is the technique of producing hard-boiled sweets with soft centers. A filling tube is injected into the hot plastic lump, which rotates in the batch roller. A pump presses the liquid filling into the center, so that, after passing the cutting station, each sweet has 20 – 25 % filling. It is important that during cutting the open ends are closed by squeezing the still plastic material together.

Because it is impossible to produce boiled sweets only with sugar, it is an old practice to add weakly acidic hydrolyzing salts to the batch. This leads to a certain degree of inversion of sucrose and formation of invert sugar, which helps to prevent the otherwise unavoidable recrystallization. Substances like potassium hydrogen tartrate are often called "doctoring agents", and sweets made by this technology, "all-sucrose candies". More common, however, is the addition of glucose syrup, which prevents recrystallization in the same way and without the risk of over-inversion.

Hard candy may be pulled by submitting the plastic mass to rotating and intermeshing cranks. This process occludes air and produces an opaque and shiny appearance.

Plastic masses of different flavors and colors can be combined and laminated to produce striped sweets. Both these laminated masses and the pulled candy can be shaped by the forming system previously described.

Brittle, butterscotch, and croquante are all high-boiled (149 – 157 °C) products. They are always cooked on a fire which gives the necessary high temperature (about 155 °C) and caramelization. Nuts may be included in the cooking process.

5.2.7. Light Foams, Marshmallows, and White Nougat

All these products are aerated. They can be made batchwise or in continuous processes (Fig. 5). The base is made from sugar, glucose syrup, and invert sugar or sorbitol. Aeration is brought about by including a foam or frappé produced from egg albumen, gelatin, or from modified milk or soy protein.

Products with high aeration (0.15 – 0.2 g/mL) are very fragile, and their manufacture requires great skill. Their shelf life is short, although they are covered with a film of chocolate ("foam

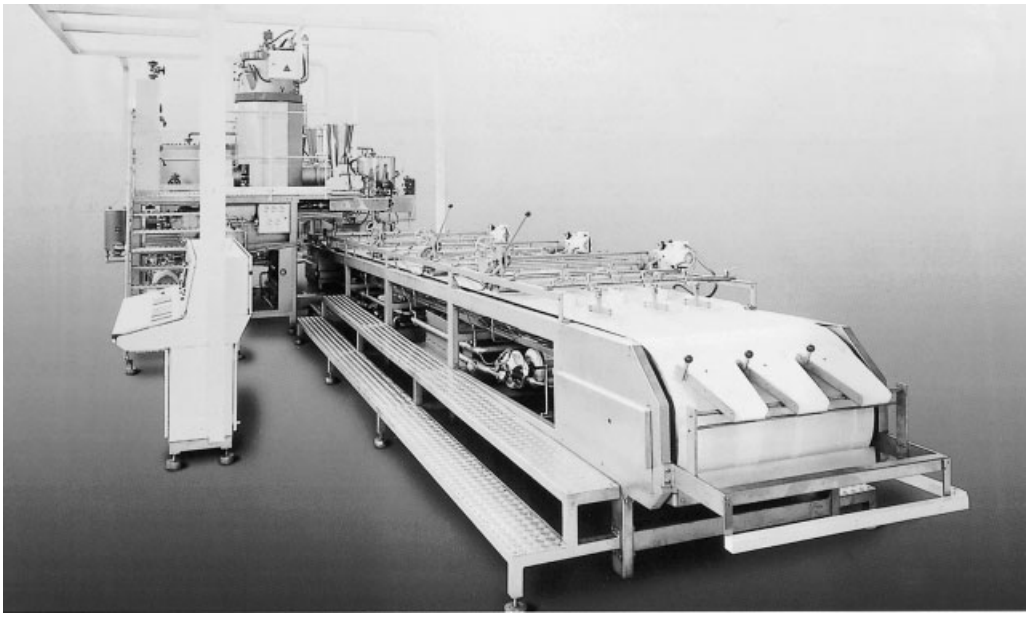

Figure 4. Forming line for hard-boiled candies for continuous large-scale production

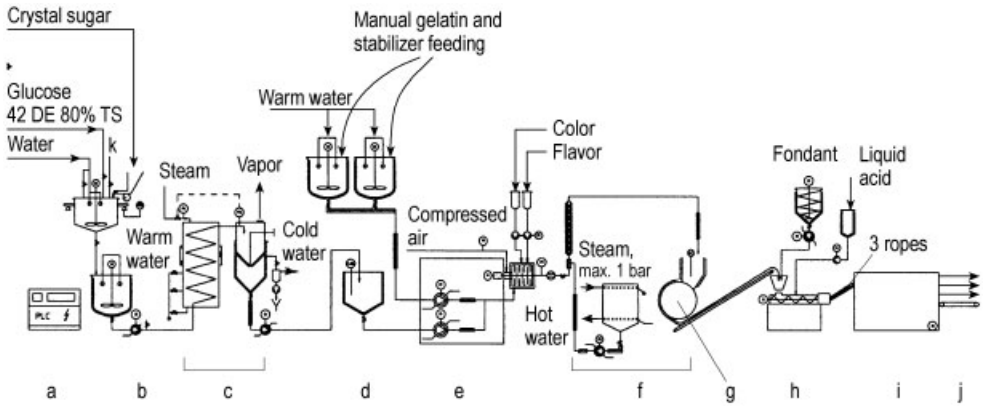

Figure 5. Chewy fruit toffee lines
a) Central control panel; b) Raw material weighing unit (Contigrav 1030); c) Predissolving/cooking/cooling (Coil Cooker 3606); d) Intermediate holding vessel; e) Turbomat 1115; f) Fat melting unit; g) Cooling drum; h) Mixing and forming extruder; i) Cooling tunnel; j) Belt system to three cut and wrap machines

kisses"). They are foamed with special fractions of egg albumin in powdered form.

Marshmallows – confectioneries with densities of about 0.4 – 0.5 g/mL – can be made with gelatin or egg albumin as foaming agent. White nougats like Torrone and Nougat Montelimar are even heavier and very tough. They are prepared with nuts, dried fruits, etc. The egg albumin powder and the gelatin have to be dissolved with some sugar, preferably by soaking with cold water for a period. Working batchwise, the foaming agents are beaten separately in a mixer with horizontal or planetary arms. Finally the hot sugar/syrup is added (frappé).

Modern methods use pressure beating, in either a batch or a continuous machine. The syrup and the aerating agent are beaten under pressure in a closed container that is controlled in time and pressure. Release of pressure is used as a means of discharge. With the continuous

method, the syrup, air, and egg – gelatin solution are each metered to a rotor – stator mixer against a back-pressure valve through which the foam is continuously discharged.

Marshmallows can be made with a variety of textures. Some may be deposited into starch as fondant; others are extruded into ropes onto a moving bed of starch and subsequently cut into pieces. Because of their lower moisture content, nougats need more powerful mixers, but the first process is similar to the atmospheric or pressure beating used for marshmallows. In the final process, fat is included together with other ingredients such as cocoa powder, milk powder, flavor, preserved fruits, and icing sugar. This mixing must be as short as necessary to completely disperse the fat because the fat acts as an antifoam and tends to reduce aeration.

Both nougat and marshmallow can be made to a short texture by the inclusion of icing sugar. This will aid crystallization of the sugar ingredient in the base syrup as well as provide a shortening effect. Air inclusion may have deleterious effects on flavor because of oxidation. Some essential oils will develop an objectionable taste and the fat may become rancid. The risk of microbiological deterioration also exists.

5.2.8. Jellies and Gums

Products of this group are either easy to bite, short and soft in texture, or gumlike and elastic. They are manufactured in a variety of textures, partly depending on the residual moisture content and partly on the gelatinizing agent used (Fig. 6) [14].

Pectin Jellies. Pectin jellies are considered by many to have the best texture, but they require careful technical control for the best results. Pectin is the natural gelling substance in fruit and is responsible for the setting of jam. To form a jelly, pectin must first be dissolved fully in water, then sugar and glucose must be added and the syrup boiled to 75 – 78 % solids. This is followed by the addition of citric acid to give a pH of 3.3 – 3.6. The hot mix is deposited at not less than 90 °C within a short space of time.

Normally, pectin jellies are cast in starch like fondant, but modern developments have made metal mold depositing possible. Two types of pectin are used commercially – high methoxyl and low methoxyl; the names denote a difference in the pectin molecule. High-methoxyl pectin requires an acid pH to give a correct jelly; low-methoxyl pectin needs a calcium salt to set it. Pectin jellies usually require the addition of a buffer salt such as sodium citrate to control the rate of setting. A good set is obtained with 1 – 1.5 % pectin in the jelly.

Gelatin Jellies. Gelatin jellies are relatively simple to prepare. The basis is a sugar – glucose syrup, and the quantity of gelatin used varies from 4 to 10 %, depending on the texture required. The essential factor is to soak the gelatin first and add it to the syrup after the syrup has been boiled to the correct concentration. Usually gelatin solution must never be boiled although modern cooking techniques have eliminated the problem of burning. The texture of jellies with gelatin is not very satisfactory. The higher the amount of gelatin, the tougher the jelly.

Agar Jellies. Agar is extracted from seaweed (red algae like Gelidium and Gracilaria). Even small percentages in a syrup form softly structured jellies. Agar is dissolved by boiling, but flavor acids must be added at the end of the boil and after cooling or the strength of the gel will be lost. Agar gels have a texture that is somewhat short, and some of them do not disperse readily in the mouth. Starch casting is usual, but agar jellies can also be set in slabs and cut into pieces.

Starch Jellies. Products made with modified starches are not really jellies, but intermediate between jellies and gums. Starch in its various forms has achieved much importance in the modern confectionery industry, in part because it can be used to produce low-cost jellies. The texture of some of these, however, leaves much to be desired. Modified starches such as boiling starch and high-amylose starch are normally gelatinized in jet cookers. A slurry of the starch and sugars is subjected continuously to high-pressure steam in a small chamber at about 170 °C. Mixes with low water content can be cooked successfully by this method. Starch casting is generally used for these jellies.

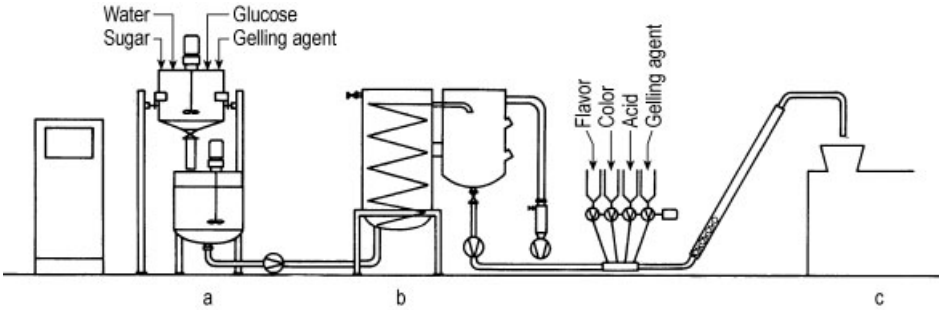

Figure 6. Process-controlled jelly plant; output up to 2500 kg/h
Semicontinuous automatic weighing of all raw materials in Contigrav (a) and continuous feeding to the mixer Coil Cooker 3606 (b); continuous cooking to the required concentration (with or without vacuum). Continuous feeding of the product to the depositing plant [Mogul hopper (c)]

Gums. Gums, wine gums, fruit gums and the like were prepared originally with gum arabic as the base, but gelatin, agar, and starch are also used today. High-quality items with the real firm bite, however, are still prepared with gum arabic in high percentages. Gelatin is used up to 25 – 40 % to achieve the desired texture. A gel is prepared in a manner similar to the jellies mentioned previously, but after casting in starch, the trays are stored for up to six days at 50 °C. After removal from starch, the gums are glazed by a process of steaming, drying, and lubricating.

5.2.9. Panning

Panning is an important part of confectionery production, and very large amounts of panned goods are now manufactured [15]. The two types are chocolate panning and sugar panning, but sugar panning can be further subdivided into hard and soft panning.

Panned confectioneries are always made in rotating coating pans. They have a special shape and are inclined so that when it is rotated, the confectionery pieces will roll round the inside of the pan and not slide. The shape and inclination of the pan can be varied to suit the product.

Whether the coating is chocolate or sugar, the operation is similar. The product to be coated (for example, almonds or hazelnuts) is placed in the pan, which is then set in motion. While the nuts are rotating, they are first coated with a sealing solution which may be gum arabic. Warm air is injected into the pan until the gum solution is dry. For chocolate panning, liquid chocolate at about 35 – 36 °C is applied intermittently by spray or drip feed and set by cool air. Each layer must be set before application of the next. Thus a thick coating is built up.

For hard sugar panning, supersaturated sugar syrups are applied in small amounts and then dried with air. By this means, a hard sugar coating is built up over a long period of time. For soft panning, the center is different and may be a preserved fruit, jelly, or paste. The outer coating consists of glucose syrup, finely granulated, and icing sugar. With all panned articles, a final glaze is applied. This may consist of an underglaze of syrup plus a gum, followed by a wax or edible shellac coating.

5.2.10. Other Products

Some products, such as chewing gum, chewing sweets, comprimated products and licorice, are almost special industries in themselves. Chocolate technology is closely integrated with confectionery [16]. The principles of chocolate tempering and enrobing should be fully understood by the confectionery technologist. Many confectionery manufacturers do not make their own chocolate but purchase it from companies that have large-scale plants and can produce it with maximum economy (→ Chocolate).

6. Packaging, Storage, and Economic Aspects

Package. The wrapping of a candy bar or, in fact, any confection, serves three main purposes:

Table 3. Sugar confectionery[a], consumption per capita (kg/head per annum)

Countries	1992	1993	1994	1995	1996	1997	97/96	97/92
EU								
Austria[b]	3.15	3.20	3.01	2.69	2.69	0.35	− 86.94 %	− 88.86 %
Belgium	5.02	4.64	3.44	4.11	4.11	3.91	− 4.82 %	− 22.13 %
Germany[c]	5.80	5.90	6.05	4.38	4.36	4.57	4.85 %	− 21.16 %
Denmark	6.65	7.52	7.76	7.55	8.01	8.25	2.93 %	24.03 %
Spain	2.88	2.96	3.11	3.04	3.03	3.01	− 0.38 %	4.62 %
Finland	4.58	4.67	4.62	4.91	5.01	5.53	10.24 %	20.57 %
France	2.93	3.20	3.29	3.52	3.48	3.57	2.36 %	21.86 %
Greece[c]	2.13	2.16	2.12	2.15	1.99	2.00	0.47 %	− 6.00 %
Italy	2.24	2.19	2.13	2.08	2.14	2.25	5.07 %	0.08 %
Ireland	5.93	5.90	5.85	5.56	5.45	5.42	− 0.61 %	− 8.70 %
Netherlands	5.80	6.20	6.27	6.18	6.55	6.46	− 1.37 %	11.30 %
Portugal	0.57	0.62	0.66	0.83	0.93	1.00	7.10 %	73.57 %
Sweden	5.06	5.01	5.05	5.11	5.17	5.30	2.66 %	4.87 %
UK	5.07	4.94	4.99	4.85	5.10	5.22	2.54 %	3.13 %
Total EU	4.03	4.10	4.13	3.77	3.83	3.88	1.41 %	− 3.62 %
EFTA								
Switzerland	2.90	2.94	2.91	2.92	3.00	2.99	− 0.14 %	3.03 %
Norway	4.37	4.59	4.30	4.22	4.38	4.59	4.73 %	4.88 %
Total EFTA	3.47	3.57	3.44	3.42	3.53	3.60	2.18 %	3.91 %
Grand Total	4.01	4.08	4.11	3.76	3.82	3.87	1.43 %	− 3.43 %

[a] Chewing-gum, Boiled sweets, Toffees, caramels and similar sweets, Gums & Jelly Confectionery including fruit pastes, Other types of confectionery
[b] Net trade for the year 1997; no data available from the Austrian Statistics Office.
[c] 1992 Estimate.
[d] Introduction of PRODCOM in 1995 entailed radical changes in German classification practices: comparison with earlier years is not possible; 1995 – 1996: including candied fruits.

Table 4. Sugar confectionery, all chocolate products, biscuits and other baked goods, key figures, 1997

Countries	Turnover in ECUS ('000)	Number of companies[a]
EU		
Austria (1996 data)	528 667	49
Belgium	2 063 054	141
Germany	9 835 539	270
Denmark	646 401	32
Spain (1996 data)	2 547 745	265
Finland[b] (1996 data)	279 662	42
France	6 139 794	261
Greece	387 904	30
Italy	6 177 266	220
Ireland	481 595	15
Netherlands	2 125 918	250
Portugal[c] (1996 data)	50 747	8
Sweden	681 989	70
UK[d]	7 077 816	170
Total EU	39 024 098	1 823
EFTA		
Switzerland	1 135 609	51
Norway[a]	370 997	5
Total EFTA	1 506 606	56
Hungary	206 021	
Czech Republic	293 474	
Grand total	41 030 200	1 879

[a] Total number of companies members + non-members in CAOBISCO's sector (except for, Norway).
[b] FIN: sugar confectionery and chocolate sector only.
[c] P: Chocolate sector only.
[d] Estimates.

(1) to provide a container; (2) to describe and advertise the product; (3) to protect the product.

Display in a supermarket is usually on multiple shelves or near the check-out counter (cashier). At both places, the contents of the package must be easily identified and, when necessary, the age of the product ascertained. The wrapping itself should have protective qualities. The material of the wrapper must prevent access of moisture or extraneous odors. The wrapper should be strong enough to resist damage by normal handling and preferably be heat-sealed. Sealing also discourages insect entry and tampering by humans.

Materials to meet these demands and machinery to seal the packages at high speed have been developed. Special laminated plastic films such as polyvinylidene chloride and polyethylene and many other plastics have been used to satisfy protective requirements.

High-speed wrapping and printing sometimes result in a serious problem with residual solvent odors which contaminate the product. Sufficient airing of the freshly printed wrappers, effective quality management in combination with certification according to ISO standards and application of high-tech laboratory equipment has reduced this defect.

Storage. With protective packaging of confectionery products, many problems have disappeared, with the exception of heat damage. No normal protective wrap will guard against excessive heat. Damage in storage often occurs because of failure to recognize the need for a strong enough outer carton or to avoid stacking loads too high. Many large companies now have computerized stockrooms where packages and loads are handled correctly and stock rotation is controlled accurately.

Economic Aspects. The chocolate (→ Chocolate, Chap. 9.) and confectionery industry is extremely large worldwide. Table 3 gives the sugar confectionery consumption from 1992 to 1997 in Europe. Table 4 is representative of the number of companies and sales volume in the EU.

7. References

1. M. T. Gillies: *Candies and other Confections,* Noyes Data Corporation, New Jersey 1979.
2. H. Hoffmann, W. Mauch: *Zucker und Zuckerwaren,* P. Parey, Berlin 1985.
3. E. B. Jackson: *Confectionery Manufacture,* 2nd ed., Blackie Academic and Professional, London 1995.
4. Ron Lees: *A Basic Course in Confectionery,* Specialized Publications, Surbiton, Surrey 1980.
5. A. Meiners, K. Kreiten: *Silesia Confectionery Manual No. 3,* **Vol. 1/1, 1/2, & 2,** Silesia Essenzenfabrik, Neuß 1984.
6. S. Z. Dziedzik, M. W. Kearsly: *Glucose Syrups Science and Technology,* Elsevier, London/New York 1984.
7. G. Andersen: "Zur Morphologie von Toffees und Fudge", *Süßwaren* **21** (1977) no. 7, 184 – 190.
8. G. Andersen: "The Microstructure of Hard Boilings and Caramels", *The Manufacturing Confectioner* **1968,** 39 – 45.
9. J. S. Kitt: "Caramels", *The Manufacturing Confectioner* **1990,** 37 – 40.
10. G. Andersen: "Gefüllte Hartkaramellen", *Süßwaren* **32** (1988) no. 4, 140 – 145.
11. P. W. Jung: "Aeration of Hard Candy", *Manufacturing Confectioner* **1987,** 57 – 60.
12. W. R. Captain: "Caramel Atmospheric and Pressure Cooking", *The Manufacture Confectioner* **1986,** 110 – 116.
13. A. Hoffmann: "Marzipan und Persipan", *Lebensmittel-Technik* **16** (1984) no. 9, 234 – 240.
14. M. Glicksman: *Gum Technology in the Food Industry,* Academic Press, New York 1969.
15. J. S. Kitt: "Panning Problems", *The Manufacturing Confectioner* **1986,** Special Issue.
16. L. Cook, E. H. Meursing: *Chocolate Production and Use,* Harcourt Brace Jovanovich Inc., New York 1982.

Construction Ceramics

Ceramic Colorants; Ceramics, General Survey; Clays; Whitewares are separate keywords.

Dieter Hauck, Institut für Ziegelforschung Essen e. V., Essen, Federal Republic of Germany (Chap. 1)

Ernst Hilker, Institut für Ziegelforschung Essen e. V., Essen, Federal Republic of Germany (Chap. 1)

Enno Hesse, Ibbenbueren, Federal Republic of Germany (Chap. 1)

Jochen Hartmann, Cremer + Breuer Keramische Betriebe GmbH, Frechen, Federal Republic of Germany (Chap. 2)

P. Schuster, Cremer + Breuer Keramische Betriebe GmbH, Frechen, Federal Republic of Germany (Chap. 2)

Hansheinz Vogel, Friedrichsfeld GmbH, Steinzeug- und Kunststoffwerke, Mannheim, Federal Republic of Germany (Chap. 2)

1. Bricks and Structural Tiles 2
1.1. Classification 2
1.2. History 3
1.3. Raw Materials 3
1.4. Preparation 7
1.5. Molding 8
1.5.1. Processing 8
1.5.2. Machinery 9
1.6. Drying 11
1.7. Firing 13
1.8. Properties of the Fired Body 14
1.9. Coloration and Surface Effects . . . 15
1.10. Quality Control and Standardization 15
1.11. Economic Aspects 16
2. Stoneware 17
2.1. History 17
2.2. Raw Materials 18
2.3. Composition 19
2.4. Preparation 20
2.5. Molding 22
2.6. Drying 23
2.7. Firing 24
2.8. Glazing 25
2.9. Stoneware Products 26
2.9.1. Clinker Ware 26
2.9.2. Sewer Pipe 26
2.9.3. Stable Ware 33
2.9.4. Products for the Chemical Industry . . 33
2.9.5. White Chemical Stoneware 36
2.9.6. Tiles and Slabs 36
3. References 38

1. Bricks and Structural Tiles

Bricks and structural tiles are mass-produced ceramic products requiring large amounts of raw materials for manufacture. Brick and tile making is mostly a low-value industry, but is nevertheless often problematic because of the dependence on suitable clay close to the manufacturing facility. These raw material problems, the increasingly higher quality standards, and competition from other building materials necessitate costly but economical manufacturing techniques.

1.1. Classification

In order to give an idea of the essential features of the various products, some porosity and strength values specified in the German DIN standards are given in the following.

Bricks. Ordinary bricks constitute the largest fraction of total brick production. They have a porous body. If there are no special requirements in regard to appearance or resistance to weathering, they are called *inner-wall bricks*, or *backing bricks*, and can be solid or hollow. For hollow bricks there are requirements for strength (4 – 28 MPa) and a brick bulk density of 1.0 – 2.2 kg/dm^3. (The brick bulk density is calculated by dividing the mass of the dried brick by its

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_425

volume *including* all channels, finger holes, and mortar pockets, or frogs.)

One special variety of brick is the *insulating brick* used to reduce heat flow [10]. For these the strength specification is 2 – 28 MPa. The brick density must be only 0.6 – 1.0 kg/dm^3. The porosity is increased by adding materials that burn during firing. Insulating brick can be as large as $49 \times 30 \times 23.8$ cm.

Progress is being made in producing building components with height equal to the height of the room. These are known as *brick planks* [11] and are produced as monolithic pieces with an extrusion press.

If the brick has surface and color suitable for facing and the brick is able to withstand frost, the brick is classified as *face brick*, or *facing brick.* If the average compression strength is ≥ 45 MPa, the brick is ranked as *extra strong.*

Clinker Bricks. To achieve clinker classification, the bricks must have a specified degree of sintering (water absorption below ca. 7 wt %), average compression strength > 35 MPa, and a *body density* > 1.9 kg/dm^3. (The body density is calculated by dividing the mass of the dried brick by its volume *excluding* all channels, finger holes, and mortar pockets.) Extra strong clinker bricks must have an average compression strength > 45 MPa.

Ceramic clinker bricks, which are manufactured from high-quality nonporous sintering clays, are face bricks having a water absorption capacity ≤ 6 %. The net density must average ≤ 2.0 kg/dm^3, and the average compression strength must be > 75 MPa. Furthermore, the surface must have a specified abrasive hardness and colorfastness in light. The clinker materials also include bricks for sewer construction, paving stones, and sintered flagstones, for which the water absorption must be below 3 – 6 %, depending on the intended use.

Roofing Tiles. Roofing tiles are manufactured to produce their particular shape, sometimes with their natural color and sometimes with a slip coating or a glazed finish (see Section 1.9.). They are manufactured by extrusion or stamping. They usually have a porous, unsintered structure, and to be weatherproof they must be accurately shaped, water impermeable, and able to bear sufficient load.

Special Products. There are a number of porous products other than roofing tiles: wall tiles, floor tiles, drain pipe, and cable conduits.

Expanded clay is a special product [1], [12], [13], usually consisting of spherical particles with a diameter of 1 – 16 mm, having a densely sintered surface and an expanded finely porous structure. The quality of this product depends critically on the packing density of the particles, the density of the outer skin and the uniformity of the pore structure. The bulk density varies from 0.3 to 0.8 kg/dm^3, permitting use as an aggregate in light concentrate or as an insulator for walls and floors.

Typical brick and tile products are shown in Figure 1.

1.2. History

Brick manufacture is one of the oldest mass-production trades. Excavations along the Nile have revealed the earliest bricks, estimated to be 15 000 years old, which were made by hand from Nile mud, sand, and powdered straw. The dimensions were approximately the same as those of the standard bricks used in Germany. They were air dried and used to construct small dwellings. The oldest description of brickmaking (ca. 1500 B.C.) was also discovered in Egypt, all stages of the process being described, from clay extraction to air drying of the product [1].

Bricks and brick fragments 6000 years old have been excavated in Mesopotamia. In Uruk (Sumeria) both air-dried and fired bricks have been found, as well as colored mosaic pieces of fired clay, these dating from 3100 to 2900 B.C. The magnitude of the industry may be appreciated from the 9.5-km brick city wall with 900 towers.

In the Punjab region of India there are remains of brick towers 8 m in height and dating from 2500 B.C. Other well-known examples are the glazed and embossed brick walls of Babylonia and Persia, and the 5000-year-old houses of Habuba Kabira, on the upper Euphrates, constructed of air-dried, die-formed bricks and provided with a sewer system of fired clay (Fig. 2), including sockets, pipes, and U-sections [14].

The 800-year-old pueblos of the American Southwest are recent by comparison. They were

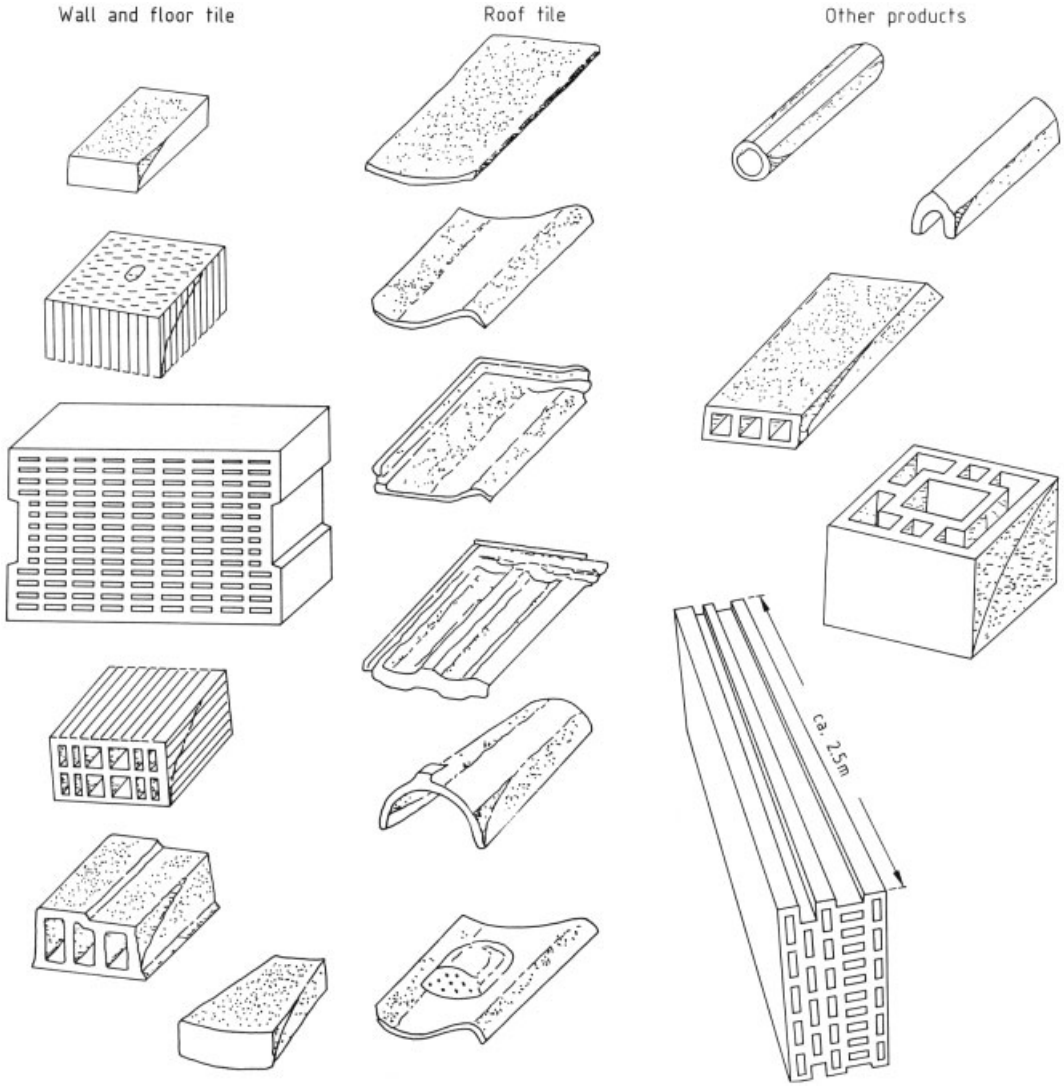

Figure 1. Typical brick and tile products

built four or five stories high from dried adobe bricks. A complex of buildings could contain up to 800 rooms [15].

As reported by Pindar (5th century B.C.), roofing tiles were invented in Corinth: later, these were adopted and used over the entire western European part of the Roman Empire. The words tile (English), tuile (French), Ziegel (German), and tegel (Dutch) are all derived from the Latin word for roofing tile, *tegula.*

Eaves trough tiles, antefix tiles, and ridge tiles up to 50 cm × 100 cm were produced. The tile shapes and methods of attaching them have changed little since Roman times [1].

1.3. Raw Materials

A great variety of clay sediments are used, ranging from loose to compact in consistency. The principal raw materials are loams, clays, marls, mudstone, shale, and slate together with smaller quantities of boulder clay, loess, and sand [16]. They can contain all types of suitable clay minerals (layered silicates → Clays), such as illite,

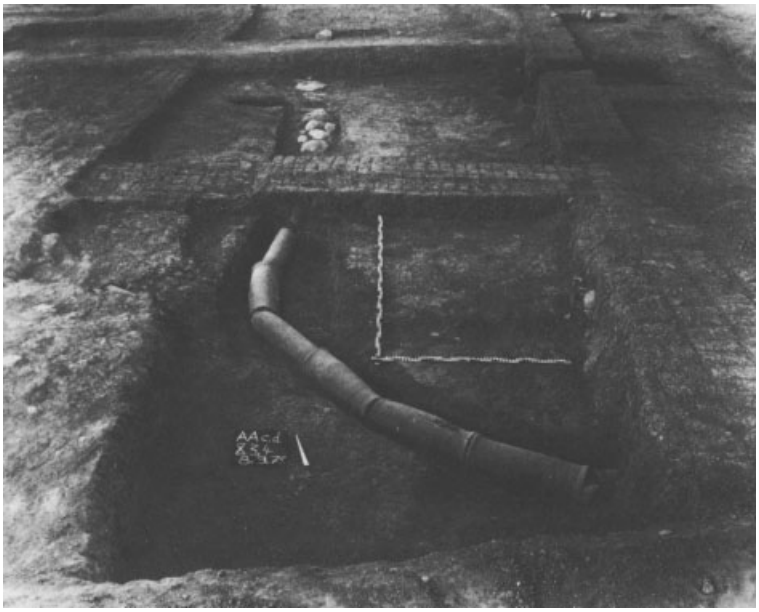

Figure 2. Ceramic sewer pipe and stone-covered sewers in a 5000-year-old house in Habuba Kabira (Courtesy of the Deutsche Orient-Gesellschaft)

sericite, kaolinite, chlorite, and montmorillonite. Quartz is an important component and is always present, and there can also be considerable amounts of calcite, dolomite, and iron compounds. The composition and particle size are the two principal factors that determine the nature of the products.

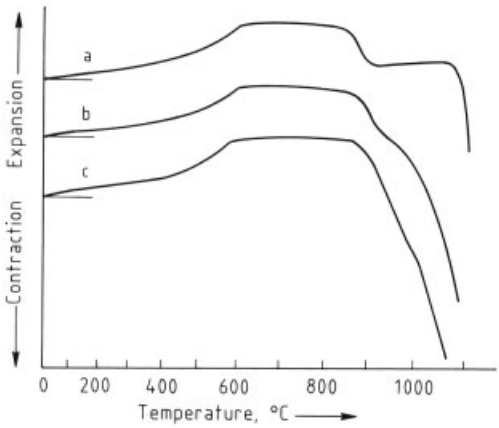

Figure 3. Dilatometer curves
a) Marl clay (calcareous); b) Clay (noncalcareous); c) Slate clay

Some constituents are harmful, mainly coarse pyrites and limestone inclusions, which can cause spalling; sulfur compounds, which can cause efflorescence [17]; and large carbonaceous particles and plant remains.

Whereas common bricks may be manufactured from fairly low-quality brick clays, large hollow bricks require adequate plasticity of the clay body. For hollow bricks, clays and marls are used primarily, sometimes mixed with loam and shale. Other products with a porous structure are produced from these raw materials, e.g., face bricks, paving bricks, roofing tiles, drainage pipes, and cable conduits. Clinker, with its impervious or near impervious structure is produced mainly from shale and stoneware clays. Shale clay tends to be used for the manufacture of clinker face bricks, high-compressive-strength solid or hollow bricks, paving bricks, and sewer channel bricks. Stoneware clays are the first choice for high-quality face bricks and for split clinker flags and flooring tiles (see Section 2.7).

Additives. Additives improve both the manufacturing process and the products. A variety of combustible organic materials are used, such as foamed polystyrene, sawdust, powdered straw, powdered brown coal, and materials with organic content including oily filtration aids

(fuller's earths), oily metallic hydroxide slimes and coal washery refuse. These additives may increase the porosity and thus the insulating quality of the brick. Furthermore, the amount of fuel required for firing may be reduced [18–20].

Powdered limestone or chalk may improve firing properties, increase porosity, and lighten the color [21], [22]. Powdered slags from blast furnaces or power stations are opening agents that add strength, while making the plastic mass leaner; glass powder is also a suitable additive [19]. Iron oxide powder and various sludges provide color [23].

As a consequence of ever more stringent quality requirements, some high-quality clay is commonly used for brick manufacture.

Raw-Materials Testing. Owing to the wide variety of raw materials, effective tests for suitability and effective methods for optimizing the proportions are needed.

Raw materials are characterized by their geological origin and condition as well as by mineralogical analysis, particle size analysis, and to a limited extent chemical analysis. Harmful constituents, e.g., coarse inclusions that could cause defects, salts that could cause efflorescence, and compounds that could produce noxious substances on firing, must be revealed by testing [17], [24], [25].

To assess the workability of the unfired mass, the water requirement, the relationship between water content and stiffness (Pfefferkorn method), the tensile and shear strength (Linseis method), the shrinkage on drying, and the transverse strength after drying are needed. Also needed is knowledge about how sensitive the material is to the drying process (Bigot curves showing shrinkage versus water loss, see Sections 1.6 and 2.6) and the water permeability.

The expected firing characteristics and properties of the fired body can be obtained by dilatometry, differential thermal analysis (DTA), thermogravimetric analysis (TGA), and test firings. Firing at various peak temperatures allows determination of the tendency to form reduction cores, the ability to retain shape (extent of softening at high temperature), the shrinkage on firing, water absorbancy (vapor and liquid) of the fired body, density, and color [1], [26].

Examples are given in Figure 3 of dilatometer curves for three different raw materials, showing the expansion and contraction during firing.

These testing methods are especially useful for working out optimum proportions of various clays and additives. In addition, there are some tests used for production control that are designed individually for particular manufacturing processes [27].

Clay Winning. The raw materials for brick manufacture are nearly always obtained by opencut mining. To reduce costs, all the various strata are usually mined, and these strata are mixed to the extent possible during the extraction process. The higher the quality of the products, the more worthwhile it is to separate the valuable clay strata from the lower value inclusions and strata.

For the extraction of soft or crumbly materials, especially when the strata are horizontal, bucket excavators are suitable because each bucket scrapes away thin layers of material from top to bottom of the strata, thoroughly mixing them in the process (Fig. 4). These excavators are, however, not especially mobile and are unsuitable for selective extraction of materials. For mainly vertical strata, bulldozers or scrapers achieve the best mixing effect (Fig. 5). These can also be used to transport the material over short distances.

Dragline excavators mix somewhat less effectively, but can be used where the geological formations are complex. A dredging shovel suspended by a cable from a boom is thrown outwards and then dragged back upwards through the material being extracted.

Power shovel excavators (Fig. 6) are widely used. These were formerly cable operated, but are now usually hydraulic and normally fitted with a grab. An advantage of shovel machines is their great mobility. Unwanted strata (gravel, chalk, etc.) are removed relatively easily, and lenticular deposits of clay can be extracted. Heavy power shovels have been used to extract diagenetically solidified raw materials such as shale. Hard clay stone must first be broken by explosives. Transportation of the materials within the mining area and to the processing area is by rail, belt conveying, or truck, depending on the distance [1].

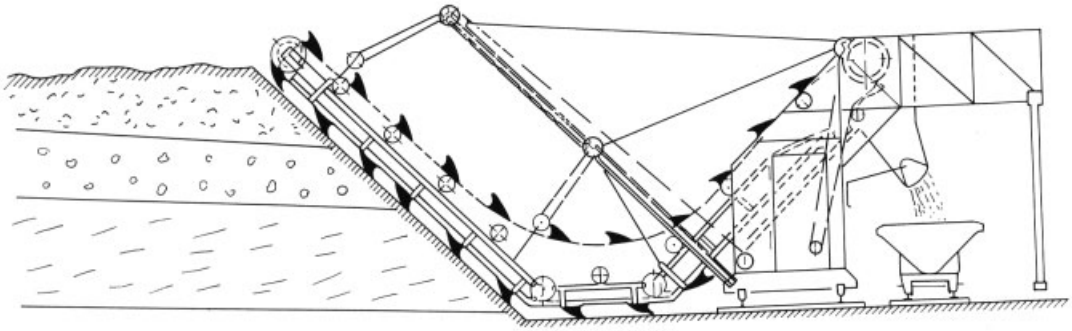

Figure 4. Bucket excavator

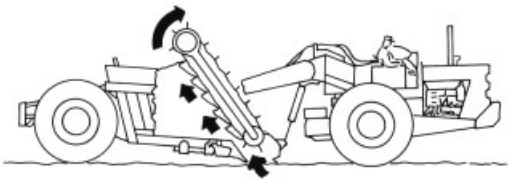

Figure 5. Scraper

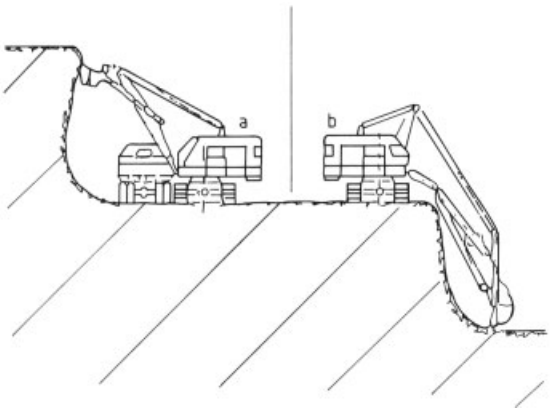

Figure 6. Power shovels
a) Excavating a face; b) Digging below ground level

1.4. Preparation

Raw materials and mixtures that are variable in consistency and composition are converted into homogeneous, easily moldable masses of uniform quality [28]. If the available equipment is unable to achieve adequate mixing, and therefore the quality required, a premix should be carried out with a suitable extraction process in the mine or with a large-scale mixer, e.g., one for mixing stocks of materials. Coarse, hard materials, such as limestone, pyrites, and quartz, should be removed as completely as possible; otherwise, the cost of producing a fine product is too high. Diagenetically hardened clays, such as shale, can be stored in the open air; the weathering that takes place favors deflocculation of the clay.

The choice of the preparation process is determined by the type and condition of the raw material and the end product. There are three processes: wet, semidry, and dry. The wet process is used most often.

Wet Processing. Most raw materials for brick manufacture have a high natural moisture content as mined. This moisture may have to be carefully supplemented during the preparation process with gauging water if the product is intended for wet molding. The constituents are precisely measured out with charging boxes (Fig. 7 A and 7 B) and wet-milled. Edge runner mills are still commonly used for this purpose, their heavy rollers (Fig. 7 C, b) effectively grinding quite hard clay stone. Only the hardest raw materials need an initial crushing operation with a roll crushing mill. The edge runner mill also acts as a mixer to some extent, incorporating most of the water needed to give good molding properties. The intensity of the size reduction and mixing processes is determined by the number and size of the slits in the sieve plate (c).

If necessary, the product may be further treated with coarse and fine crushing rollers,

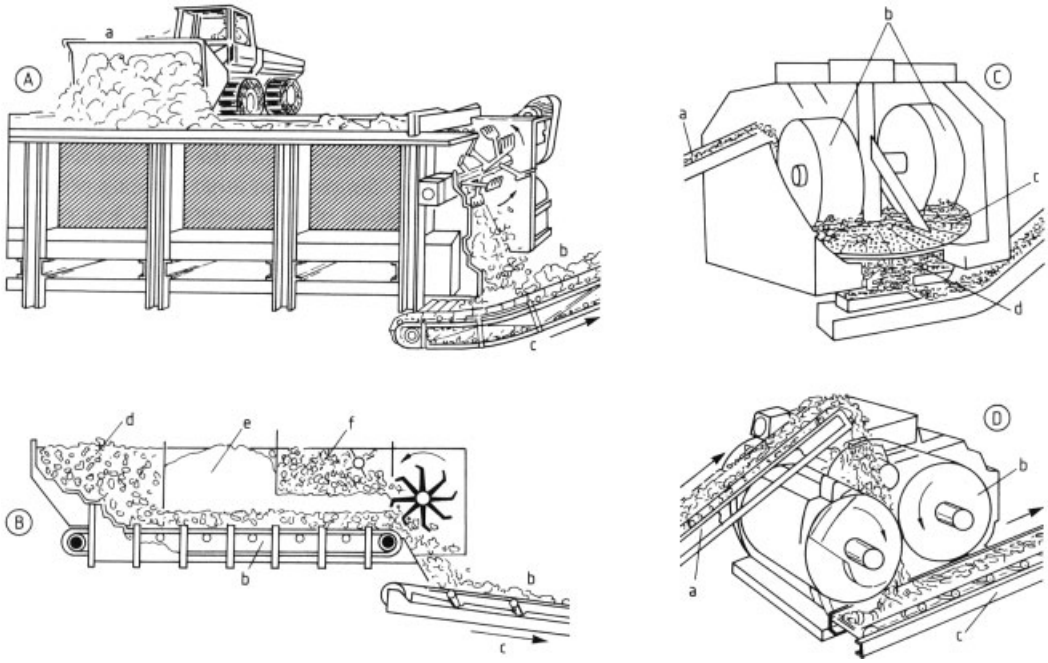

Figure 7. Mixing and grinding machines
A) Box feeder: a) Feed; b) Conveyer belt; c) Direction of the next processing machine
B) Box feeder (cutaway side view): b) Conveyer belt; c) Direction of the next processing machine; d) Clay; e) Sand; f) Sawdust
C) Kollergang (edge runner): a) Feed; b) Rollers (e.g., 1.8 m in diameter); c) Perforated plate; d) Collecting plate
D) Rolling mill: a) Conveyer belt from the kollergang; b) Rollers; c) Conveyer belt to the next processing step

which reduce the particle size to < 1 mm (Fig. 7 D). A few installations include a series of about three roll crushers, which progressively reduce the particle size of the raw material mixture. Another technique is to treat coarse material, while water is being added simultaneously with cone-shaped rim gears, which produces a fine, soft plastic mass. Lean mixtures, which break down more easily, may usually be molded directly after mechanical crushing. Very plastic materials are better processed intensively with mixing and homogenizing machines to break up the compressed clay mass into small pieces.

The material is often put into temporary storage in large containers, bunkers, or aging towers in a moist condition. This procedure encourages deflocculation and homogenization of the clay, and tends to prevent operating problems. It also ensures a stable supply and good cross mixing of material. However, prolonged storage of moist material is not advisable if pyrite is present, because this slowly oxidizes to sulfate, which produces severe efflorescence on drying.

There are mixing and homogenizing machines specially designed for the molding process, e.g., circular feeders and single and double trough mixers. They facilitate the addition of measured amounts of powders (e.g., barium carbonate), materials to promote porosity (e.g., polystyrene), and water or steam to heat up the mass (hot preparation). These machines, like the aging tower, fulfill two separate functions [1], [10].

Semidry and Dry Processing. These procedures are mainly used in the production of clinker flagstones, more rarely for face bricks, pipes, and clinker paving slabs. The best raw material for these products is shale, which has a low natural moisture content, < 10 wt %, and therefore does not need to be dried. If shale has a high quartz content it is very abrasive and perhaps should be wet ground in a heavy edge

runner mill. Dry processing is used from time to time in the manufacture of high-quality face bricks and floor tiles when highly plastic, moist clays, e.g., stoneware clays, are used. However, a large proportion of the water must be removed before processing.

Depending on the type and hardness of the raw material, various types of size reduction machinery are used: jaw crushers, kollergangs, impact mills, hammer mills, pendulum roller mills, and centrifugal grinders.

Moist raw materials can be dried during milling with the waste hot air from driers or kilns. When harmful inclusions interfere with the manufacture of high-quality products, these combined grinding/drying machines can also pay for themselves by finely milling these particles in the presence of high clay moisture contents [29].

An advantage of dry processing is that the material being milled can be mixed with the precise proportions of powdered additives required, e.g., powdered clay, brick dust, sand, metallic oxide pigments, etc. Then the product can be transported satisfactorily by screw conveyor and is easy to remove from the storage bunkers in a uniform manner.

Dry processing is especially suitable when dry pressing is to be used. If necessary, the material is carefully moistened and granulated on trays. In order to obtain the correct granule size and to cause densification, pan granulators may also be used.

1.5. Molding

1.5.1. Processing

The molding process most commonly used for brick is *wet pressing.* Other processes are stiff mud pressing for low water additions and the semidry and dry pressing processes. The choice of pressing process is determined by the raw materials and products [1], [30], [31]. The most suitable moisture content depends mainly on the type of raw material to be treated.

Wet Pressing. Materials with sufficient plasticity may be molded by wet pressing using extrusion presses or stamping presses to form almost all brick products, including wall slabs and flagstones. The correct water content (tempering water) is between 20 and 30 wt % and depends on the ratio of clay to nonclay material. Whereas ordinary bricks tended to be soft pressed, more recently there has been a tendency toward stiff pressing. The reasons are improved green strength and therefore dimensional stability, and a saving in energy for drying.

Stiff Pressing. Good plastic masses and fairly soft shale are sometimes molded by stiff pressing. The gauging water is then ca. 15 wt %. Mixtures that contain lean and silty minerals and low-quality clays are unsuitable because their molding properties are extremely sensitive to slight variations in moisture content. Stiff pressing, particularly for hollow bricks, requires high pressures and powerful presses. The main advantage is that the molded pieces are stacked on the kiln car immediately after forming and then can be passed through the dryer and kiln without repeated handling. This method of molding is not so suitable for making products with many holes, e.g., with thin webs or thin walls, nor can it be considered for pressing insulating bricks whose porous materials would be destroyed by high pressure.

Soft Mud Brick Making. By means of strickling machines, lean or silty clay mixtures that are unsuitable for the extrusion process may be molded to give face bricks with a very rough (rustic) surface to fulfill special architectural requirements or for restoration work (sometimes even hand molding is employed). These processes require soft plastic masses with gauging water contents of 25 – 35 wt %.

Semidry Pressing. Semidry pressing, which is, for example, commonly used in Great Britain to make bricks [1], requires a gauging water content of 10 – 15 wt %. As with stiff pressing, the molded pieces may be handled without difficulty and stacked directly onto the kiln cars without predrying. Compared to other molding processes, preparation of raw materials is less costly, normally only involving particle-size reduction and incorporation of the required tempering water.

Dry Pressing. Materials for dry pressing, which consist of shale, require as a rule 5–10 wt % gauging water. Stoneware clays and other rich clay materials, or those with complicated shapes, can have a water content exceeding 10 wt %. Diagenetically solidified clay stones, e.g., shale, are especially suitable for dry pressing with low moisture content because they can be fairly easily milled to the correct particle size. If plastic clays are used, they must, after drying, be brought to the appropriate grain size by granulation. Molding is then carried out by various stamping presses.

1.5.2. Machinery

Extrusion Molding. The most popular machine for extrusion molding is the screw or helical extruder (Fig. 8).

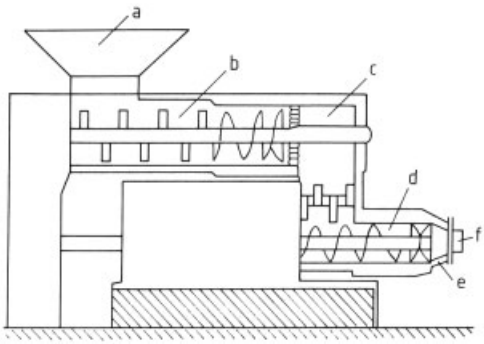

Figure 8. Screw extruder
a) Feed; b) Double shaft mixer; c) Vacuum chamber; d) Extrusion tube with screw; e) Die

The prepared mass is usually mixed first in a double trough mixer (b) with a sieve or annular discharge into the screw feed of the extruder (d). The molding head (e) forms an integral part of the die (f). The final pressed shape is formed in the die. Between the double trough mixer and the screw there is usually a vacuum chamber (c) in which the air is extracted from the shredded mass. This improves the binding and flow properties, and allows raw material of naturally poor molding properties to be molded. In addition, the formation of the extrudate by the die is improved by water or steam treatment.

Although the extrusion press has the advantages of continuous operation, high throughput, and varied forms, it has the disadvantage of texture formation, which can be aggravated by drying and firing. Lean mixtures also have a tendency to surface tearing and formation of "dragon's teeth" on the edges. The ring-shaped and S-shaped textures [32] caused by the action of the screw feeder cannot be wholly avoided. However, altering the granule characteristics (e.g., increasing as much as possible the proportion of easily fractured coarse granules) and modifying the screw, the die, and the molding head reduces the number of these textures so that the final product quality is not excessively affected.

Apart from the screw-fed extruder, the disk press (Europress, Fig. 9) is also used. In this press the screw is replaced by a roller feeder (c) resembling a V-belt pulley.

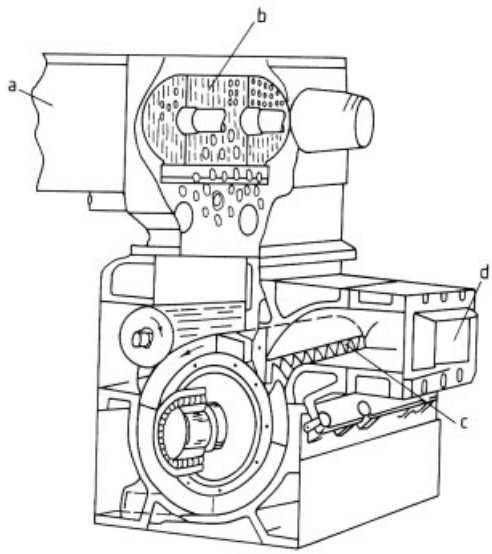

Figure 9. The Europress
a) Feed; b) Vacuum chamber; c) Roller; d) Die

The extrusion method makes possible the production of a wide range of surfaces on the molded shapes by the use of devices such as shaped rollers and brushes located near the die outlet. The surface of the extrudate may be coated with an aqueous sealant to avoid dry efflorescence.

Forming the extrudate with a suitably designed press and die (double die) allows the use of lower quality raw material for the molded shape with a surface layer of high-quality material a few millimeters thick.

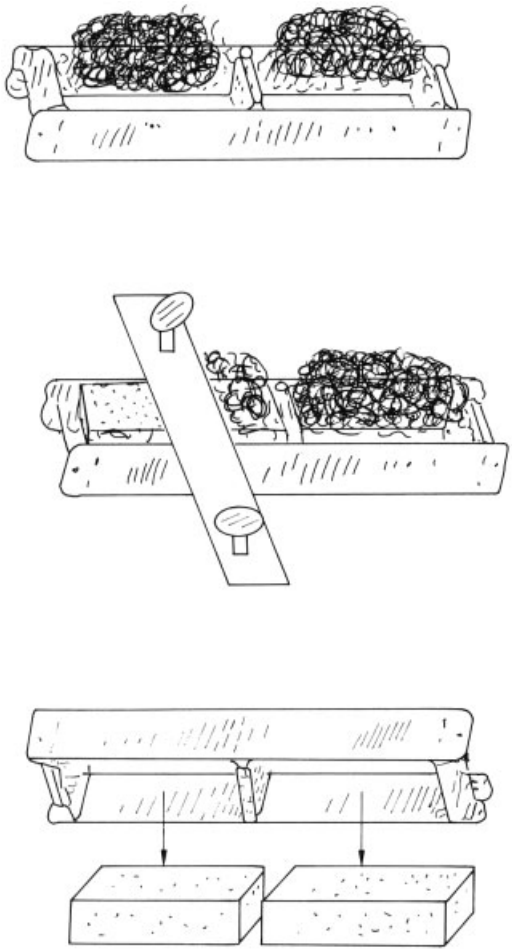

Figure 10. Soft mud molding by hand

After leaving the die, the extrudate is carried by rollers or band conveyors to the cutting machinery, where it is cut into separate pieces. Each type of product requires its own *cutting system,* the most common being stretched steel wires. For roofing tiles, split clinker flags, and large blocks formed from soft material, knives are also used. With a ribbon harp (i.e., an array of parallel ribbon cutters) a long length of extrudate can be cut into a number of pieces in one cutting action. Even though this is quite slow, a high output may be obtained. For extrudates that are difficult to cut, e.g., stiff plastic masses or molded pieces with low green strength, a vibrating or saw action cutter is an improvement [18].

Soft Mud Molding. Before the brick industry became mechanized, the usual process was hand molding (Fig. 10), and this is still used to a small extent. A lump of soft plastic clay was beaten into a sanded mold, the excess material was removed by strickling, and the mold was inverted to release the product. These techniques are usually carried out with mass-production strickling machines.

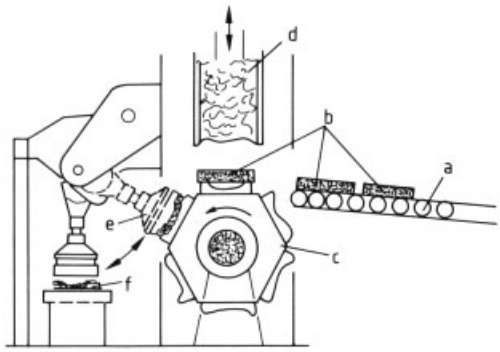

Figure 11. Revolver press (schematic)
a) Conveyer belt from the clot press; b) Clots; c) Drum with lower molds; d) Plunger with the upper mold; e) Green body removal; f) Conveyer belt to the dryer

Press Molding. Products with complex shapes cannot be produced by extrusion; complex shapes are produced by stamping.

Articles such as special types of roofing tiles, wall and floor tiles with embossed surfaces, and face bricks with a "handmade" surface are made on a machine known as a *revolver press* (Fig. 11). The clots of clay are produced on a screw-fed extrusion press, either with or without a deairing stage. All stages of the process are automatic: placing the clots in the pressure mold, the various steps of the compression process, the trimming of the shaped pieces, their removal from the mold, and placement on a carrying frame or band conveyor.

For molding ridge tiles, ventilation bricks, and other special shapes, semi and fully auto-

matic turntable presses, as well as high lift revolver presses, are used.

The molds for stamping presses are still mainly made from high-strength plaster, sometimes bonded with plastics. Other mold materials used are lubricated cast iron, steel, molds with rubber sheeting, and metal molds with vulcanized rubber coating [1].

Dry press material is usually molded with *rotating table presses*, which are driven mechanically, hydraulically, or by a combination of the two. Mechanical or hydraulic toggle lever presses or friction presses are also used. These last two give the highest compression force.

With dry presses, the prepared, loose feed is loaded into the mold, where it occupies about twice the volume of the final molded piece. Exactly reproducible amounts of material must be charged so that the pieces are of constant size, and they must be compressed to the same extent, a task that can be difficult.

While the production of thin wall and floor tiles is relatively easy, thick pieces and pieces containing holes can present more problems. There are large frictional forces between the material and the die walls, and between the grains of the material itself, and these effects reduce the densification produced by the applied pressure. Steps can be taken to reduce these density differences, e.g., lowering the viscosity of the added water with flow-promoting agents such as lubricating gums (sulfite liquor, etc.). Another approach is to even out the differences by applying pressure from both sides.

An important method of preventing delamination is to deair the mixture. It can be improved by stepwise compaction of the grains by vibration processes and also by carrying out the compression in several stages [31].

The machines for the semidry pressing are fairly simple and require little maintenance. The pieces are preformed in molding boxes between two mechanically operated rams and then compacted again in a second shallower molding box (double pressing process).

1.6. Drying

The gauging water, which was needed to make a plastic mass, is removed during drying, the process that immediately follows forming. This drying is associated with marked shrinkage, which varies considerably in extent, generally ranging between 10 and 25 % by volume. When the green bricks are formed under conditions of flow, as in an extrusion press, the amount of shrinkage in the direction of the flow can be strikingly different from that in the perpendicular direction as a result of the orientation of the clay particles.

Drying occurs in two stages. In the first stage, during which ca. 1/3 – 2/3 of the water evaporates, depending on the material, nearly all the shrinkage takes place. A large proportion of the pore water and the water film around each solid particle evaporates, and the particles move closer together. In the second stage, the remaining water evaporates without appreciable shrinkage. The shrinkage can be measured as a function of the water removed with a Barelettograph, which requires only a small sample, and can be displayed as Bigot curves.

Incorrect drying of green bricks can produce distortion and cracking. The drying conditions are determined by both the type of brick and the material, which can vary with respect to mineral composition and grain-size distribution. Fine-grained, unmixed clay minerals, especially montmorillonite, can easily cause defects on drying, even in the case of small amounts. Products made from mudstone or shale, however, give few problems.

The most important factor affecting the drying properties of ceramic masses is the moisture diffusion through the mass. Reducing the sensitivity to drying, and hence drying times, therefore depends largely on improving moisture-diffusion properties. Suitable additives are nonclay inorganic materials, such as sand, slag, and brick dust, or organic fibrous materials, such as paper pulp or sawdust. In some cases, flocculating agents, such as calcium hydroxide, may be used, provided that they do not cause excessive deterioration in the molding properties [33]. The effects of nonclay minerals, and especially their particle size, on drying have been fully discussed [34].

On an industrial scale, brick products are generally dried with an air current, usually preheated. In times past bricks were almost exclusively dried in the open air or in large rooms above an oven. This is still true in developing countries, where land and labor are less scarce than capital. Modern drying equipment is almost

always capable of year-round operation, and the air temperature, moisture content, and throughput are controlled to give optimum product quality [35]. Open-air drying is no longer common.

The systems most frequently used are chamber dryers and tunnel dryers. The *chamber dryers,* which have single or double chambers, are periodically charged and discharged by means of fork lift trucks, with the green bricks stacked in several layers on grids or frames. The whole of the chamber contents are in contact with the drying air.

In the *tunnel dryer,* the feed material, which is carried by frame or pallet cars, is transported through a tunnel so that the drying air contacts the green bricks, either countercurrently or cocurrently. The principal advantage of a tunnel dryer is that it is a continuous process, thus providing a continuous feed to the next stage.

In both types of dryer, the air can be circulated horizontally and vertically, improving the drying effect [36]. Uniformity of drying, and therefore reduction of the time required, can be achieved with batch driers by judicious direction of the air stream [37]. One special characteristic is the rhythmic drying of the Rotomixair system, a system provided with wide currents of air or with longitudinal slit nozzles in the walls. The warm air flows intermittently for brief periods at a high velocity onto the material [1].

The evenness of drying of a charge or a chamber is affected considerably by the arrangement of the green bricks; this determines the permeability of the stack. As a general rule, drying time exceeds 24 h, because even green bricks with many holes are mainly dried by the stream of air flowing around the outside of the bricks [34], [38], [39].

So-called *rapid dryers* achieve crack-free drying of bricks containing holes in less than 6 h. A large number of air jets are used, whereby every single brick is subjected to an air stream over and through it. Whether or not a rapid dryer can be used in an individual case depends both on the shape and on the raw materials of the brick. Raw materials that are particularly sensitive to drying conditions are unsuitable for rapid dryers. The type and significance of the drying stress when air jets are used has been investigated [40].

The tunnel cars that contain the dried bricks are usually kept in heat-retaining tunnels before going into the kilns. Even if the raw materials contain a high proportion of swelling clay, this avoids most reabsorption of moisture from the atmosphere, and the consequent reduction in quality.

1.7. Firing

Brick products are fired predominantly in tunnel kilns because these require fewer personnel, are easily automated, and may be operated under uniform conditions. The stacking of the green brick on the cars is often done by stacking machines, not by hand. The earlier ring kilns (Fig. 12), zigzag kilns, and chamber-ring kilns are now used only rarely. For firing special products in small quantities and for influencing the firing colors of roofing tiles and clinker bricks by means of a reducing atmosphere, batch chamber kilns are occasionally used because they permit individual firing regimes.

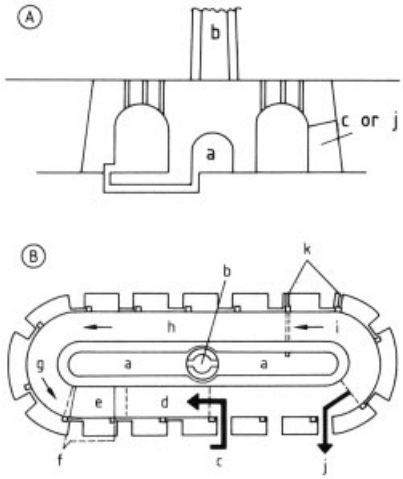

Figure 12. Ring kiln
a) Flue-gas collector; b) Chimney; c) Entrance; d) Stacks of the green bricks; e) Prefiring zone; f) Damper; g) Heating-up zone; h) Burning zone; i) Cooling zone; j) Exit; k) Flue-gas skimmers
The small arrows show the direction of motion

As a rule, bricks are direct-fired at peak temperatures of 900 – 1200 °C. The fuel can be gas, oil, or coal. Clinker bricks and slabs that should not come into contact with combustion products (which might affect the fired colors) are heated indirectly in tunnel kilns. High-quality wall and

floor tiles are occasionally fired in electrically heated kilns.

Bricks are usually heated by overhead firing, but sometimes also from the sides. The temperature distribution can be improved by high-velocity jet burners [1].

The energy requirement varies with the type of raw material, firing temperature, and kiln operating conditions. Ordinary bricks fired in a tunnel kiln require ca. 1 MJ/kg of fired material [39]. For the firing of insulating bricks, a considerable proportion of the heat energy is supplied by the combustible substances, e.g., sawdust, that are added to increase the porosity of the material [19], [20].

In order to fire at higher temperatures without causing distortion, e.g., to increase frost resistance or density of the fired body, roofing and wall tiles are fired inside boxes (saggars) to avoid deformation.

Firing Conditions. Because of the variety of raw material mixtures, as well as the sizes and shapes of the products, there are no universal firing conditions: each product needs its own firing regime [41].

The firing rate allowed at each stage of firing depends on the physical and chemical reactions of the raw materials: swelling, shrinkage, evaporation, combustion and decomposition processes, solid-state reactions, and formation of liquid phases, all depending on the thickness, size, and shape of the molded piece [42].

In the prefiring zone, with temperatures up to 300 °C, the rapid evaporation of the water can give rise to significant shrinkage, and consequent high distortion and crack formation, particularly with highly plastic masses containing large proportions of swelling clays.

Another critical temperature range lies between 450 and 650 °C. If the material contains quartz, this quartz can undergo large volume changes producing considerable internal stresses and therefore cracking.

For many materials, the rate at which the temperature may be increased above 600 °C is limited owing to a tendency to give *black cores,* the mottling of the outer surfaces or intumescence. The basic cause is the presence in the raw material of bituminous substances, plant remains, wood, coal, or combustible materials that have been added to increase porosity of the fired body or to economize on expensive fuels. The oxidation of these substances must be complete before the diffusion of gases is obstructed by vitrification of the material. This may be achieved by correct dwell times at set temperatures, which must be determined for each case [43].

Materials rich in limestone, such as marl clays, are useful for the manufacture of lightweight insulating bricks. However, significant shrinkage, causing cracks, can take place in the range of 800 – 900 °C because of silicate formation, especially in the case of large-sized bricks.

Cooling of brick products through 573 °C is a problem because at this temperature any free quartz changes from the high-temperature (β) to the low-temperature (α) form, with an accompanying reduction in volume (quartz transition). In products that require a high firing temperature, such as clinker bricks and floor tiles, some quartz can change to cristobalite, and this too gives rise to a contraction on cooling, in this case at 230 °C [44].

Figure 13 shows the typical firing curves for four different raw materials or product types.

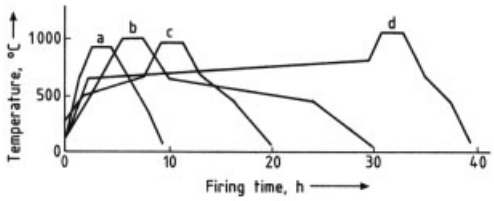

Figure 13. Typical firing curves

Possibilities of controlling the emission of harmful substances, e.g., fluorine, during firing has been described [25], [45].

Expanded clay products occupy a special position. The wet-formed granules, which consist of pellets from a pan granulator, or fragments produced from shale by size reduction [46], are usually (ca. 90 %) fired in rotary kilns [1]. The feed material is heated to ca. 1150 °C in 30 – 45 min, with the firing controlled so that the outer skin is densely sintered before the gas begins to form in the interior, which causes the clay spheres to swell [12], [47].

1.8. Properties of the Fired Body

The fired body acquires the properties that render it suitable for its particular use mainly in the finishing burn, i.e., the time that the piece spends near the peak temperature. At the same time the density and strength of the body increase. In general, a longer residence time around the peak temperature improves product quality.

In the course of this sintering process, new mineral phases, mainly silicates, are formed as well as the glassy phases that bind the whole body together. The final properties that are desired are produced not only by the nature of the raw materials, the pretreatment, and composition but also by the firing process [48]. Porosity data, such as water absorption and strength, are used to assess the degree of sintering. The designation *sintered* denotes in practice that a definite, prescribed degree of densification and hardening has taken place.

Most brick products, such as inner-wall bricks, face bricks, roofing tiles, floor tiles, and drainage pipes, can be and even should be porous. For these products then, the required fired body strength, weather resistance, etc., can be attained before the sintering process is so advanced that softening and deformation occur as the result of the excessive formation of liquid phase.

Lightweight insulating bricks are made especially porous (body density of ca. 1.2 kg/dm^3) by incorporating substances that burn out during firing. Convenient raw materials are those that on firing produce a porous structure without additives, either because of their naturally high organic content (e.g., the overburden clays in the brown coal regions) or because of a high content of finely divided chalk. The relatively high porosity of fired products from chalk-containing raw materials (marl clays) is produced not only by the formation of CO_2 from calcium carbonate but also by the formation of calcium silicates, which prevent shrinkage at 900 – 1000 °C, sometimes even leading to expansion. The fired strengths of lightweight bricks made from clays rich in chalk are comparatively high; therefore, calcium carbonate as limestone or powdered chalk, is sometimes added to clays poor in calcium carbonate [21]. A further feature of chalk-containing bodies is that they must not be heated above 1060 °C, or they soften when calcium iron silicates suddenly melt.

Clinker products are manufactured from materials low in chalk. These materials, on account of the minerals and potential fluxes present, can withstand higher temperatures, reducing atmospheres, and prolonged dwell times at the peak firing temperature, becoming dense without deformation.

A special firing method, known as the Hydrite process, was developed in the former German Democratic Republic. This process involves firing the green bricks in indirectly heated kilns with maximum temperatures of ca. 780 °C. The dehydratation of the clay minerals produces a moist atmosphere, which irreversibly strengthens the fired brick, and, it is reported, giving higher compression strengths than by normal firing methods [49].

1.9. Coloration and Surface Effects

The color of fired brick products is determined mainly by the ratio of iron (III) oxide to aluminum oxide if the lime content is low. If the lime content is high, the color is determined mainly by the iron oxide/calcium oxide ratio [22]. Other raw material properties and firing methods play a role. Iron oxide usually gives red brick; lime tends to produce whitish-yellow brick. Incorporating a reducing stage in the high temperature and cooling zones produces varied shades, from yellow to green, blue, and gray black, in place of the usual red. This is mainly caused by the reduction of iron (III) to iron (II). If the reducing stage is prolonged and is continued until the product has completely cooled, small particles of carbon deposit to give a silver-gray appearance. This process is mostly used for the production of silver-gray roofing tiles. The use of raw materials containing adequate lime combined with a firing in a reducing atmosphere produces light colors.

Glazing and Engobing. A broader product range may be achieved by the use of a glaze coating on face bricks, wall and floor tiles, and similar items. These glazes are fused onto the already fired clay body; they consist of a mixture of silica sand, clay, alkali-metal and alkaline-earth metal oxides, lead and boron oxides, and

colored metallic oxides, such as CoO, Cr_2O_3, MnO_2, and Fe_2O_3.

Sometimes roofing tiles are also given a glaze. Usually, however, they are engobed with a porous layer, i.e., they are coated with a fine clay that gives a color on firing [50]. A shiny surface may be obtained if these coatings also contain lead silicates, alkali-metal compounds, etc., which melt at low temperatures (sintered engobing). Because glazing and sintered engobing both have the effect of densifying the surface it may be necessary to allow for the loss of moisture at the back of the brick to avoid frost damage [51], [52].

Solid Coloring. Occasionally, face bricks, tiles, and roofing tiles are colored throughout. Some examples of substances used are as follows:

Iron (III) oxide red
Manganese clays or manganese/iron ores brown
Mixtures of manganese/iron ores with chromium oxide black

To reduce the cost, the oxides used are ores or byproducts from other branches of industry, if possible [23]. The use of chromium oxide to produce a green color is economical only for high-value floor tiles.

Other methods of modifying the surface appearance of face bricks are treatment with rollers or brushes, removal of the surface skin, application of sand, surface treatment with combustible materials, and rock facing, i.e., mechanical removal of the outer surface of the fired brick.

1.10. Quality Control and Standardization

Continuous quality control of construction ceramics by recognized institutes is gaining in importance because over the years the quality requirements have increased considerably. In addition, most producers have their own quality control. The trend is to control the quality in-house and to rely on external testing only to monitor in-house quality control [13], [52].

In the *Federal Republic of Germany,* for example, the product requirements vary significantly with the end use but are standardized:

DIN 105		Mauerziegel (bricks and tiles)
	Part 1	Vollziegel, Hochlochziegel (facing and inner-wall brick)
	Part 2	Leichthochlochziegel (insulating brick)
	Part 3	Hochfeste Ziegel und Klinker (high-strength bricks and clinker)
	Part 4	Keramikklinker (ceramic clinker)
	Part 5	Leichtlanglochziegel und Leichtlanglochplatten (brick planks)
DIN 456		Dachziegel (roof tiles)
DIN 4 159		Ziegel für Decken- und Wandtafeln (wall and floor tiles)
DIN 4 160		Ziegel für Decken- und Wandtafeln (wall and floor tiles)
DIN 278		Tonhohlplatten und Hohlziegel (hollowbricks and tiles)
DIN 4 051		Kanalklinker (bricks for sewer construction)
DIN 1 057		Mauersteine für freistehende Schornsteine (bricks for freestanding stacks)
DIN 18 503		Pflasterklinker (paving clinker)
DIN 18 158		Bodenklinkerplatten (flooring clinker)
DIN 1 180		Dränrohre aus Ton (ceramic drain pipes)

European Standards. For several years now European standards (EN) have tended to replace the various national standards. For example, the quality control for split flags has been established as an European standard.

International Standards. The International Organization for Standardization (ISO) consists of representatives of national standardization organizations. Some non-European countries, e.g., Australia, Brazil, Canada, China, India, the former Soviet Union, and the United States, are also represented. The technical commission ISO/TC 179 has published "Materials Testing Methods and Requirements," which is important for ceramic products. In addition, the design

and construction of masony structures, both reinforced and non-reinforced, has been standardized. For developing countries, the *simple rules* have been made available.

Table 1 shows the standard lengths of bricks in several countries. The testing of ceramic raw materials and products is treated in more detail under → Ceramics, General Survey.

Table 1. Dimensions of standard bricks [10]

Country	Nomenclature	Dimensions, mm
France	most usual size	220 × 110 × 60
Germany, Federal Republic of	NF	240 × 115 × 71
United Kingdom	BS	210 × 102.5 × 65
United States	standard	203 × 95 × 57

1.11. Economic Aspects

In 1984, brick production in the Federal Republic of Germany was ca. 5×10^9 units (one unit equaling the standard format 240×115×71 mm). This production is equivalent to ca. 82 000 homes. In 1981, the fraction of brick wall construction in homes in the Federal Republic of Germany was 43 %. Sand-lime blocks and expanded-concrete blocks accounted for 34 % and 9 %, respectively, of wall construction. Ceramic tile is used for 22 % of the nonflat roof surface in the Federal Republic of Germany.

The number of brick producers in the Federal Republic of Germany decreased from ca. 1000 to ca. 320 over the 20-year period 1965 – 1985, the number of employees decreasing from 58 000 to 12 500. In the same period, the total production of bricks decreased by 22 %; that of roof tiles, by 32 %. In countries such as Belgium, France, Finland, Great Britain, Holland, and Sweden, the number of producers has decreased 75 – 80 % [13].

In 1982, a modern brick plant in the Federal Republic of Germany could produce ca. 120 000 units (ca. two homes) per day, i.e., 44×10^6 units per year at full capacity. The capital cost of such a plant was ca. 12.8×10^6 €in 1982. The sales were ca. 9.2×10^6 €, wages accounting for ca. 20 % of the sales.

Table 2 compares brick and roof tile production in several European countries [53].

Brick production is reported by number of bricks or by cubic meters of brick. Tile production is reported by number of tile or square meters of tile.

2. Stoneware

Stoneware is a traditional, clay-based ceramic with a densely sintered body, which is not translucent and has a conchoidal, stonelike fracture. Its external characteristics distinguish stoneware from other ceramic products: The densely sintered body is not absorbent and does not adhere to the moist tongue, which clearly differentiates stoneware from brick, grog (chamotte), and earthenware. Stoneware differs from porcelain mainly in the color of the body and often also in its coarser structure. Both unglazed stoneware and unglazed porcelain are impermeable to liquids. Light-colored stoneware is a type of whiteware (→ Whitewares).

The highly plastic nature of the clay mass for stoneware, the many possible methods of forming the shape, and the useful physical properties of the sintered body have led to variegated stoneware products during the course of the historical development. These include tableware, art objects, construction ceramics, and industrial ceramics. The division into coarse and fine ceramic products is based on the visible macrostructure of the body. Among the coarse ceramics are clinker, sewer pipes, and stable equipment; fine stoneware includes all densely fired ceramic products that are not porcelain.

2.1. History

The earliest stoneware is Chinese and dates back to the 1st and 3rd centuries A.D. All earlier ceramic products were porous, underfired pottery. The Chinese were the first to make a completely nonporous, brown stoneware. The Koreans and Japanese followed. In the 9th century A.D. the Chinese succeeded in producing white stoneware, representing the transition to porcelain. The famous China red glazes were the model for many European developments.

In Europe, where only porous earthenware had been produced in ancient times, there was a course of development similar to that in the

Table 2. Production of European countries in 1984

Country	Plants	Employees	Bricks, 10^6	Roof tiles, 10^6
Austria	62 *	2 150 **	969	1.7
Belgium	61 **	2 275 **	1.4 m^3	
Denmark	47	1 170 **	409	20.6
Finland	29 **	930 **	136	
France	242	8 400	4.6 m^3	42.2 m^2
Germany, Federal Republic of	325	12 500	11.7 m^3	23.0 m^2
Great Britain	200	14 500 **	5.2 m^3	2.0 m^2
Italy	513	27 000 **	11.6 m^3	33.0 m^2
Netherlands	86	2 150	1 600	41.5
Switzerland	31	2 255	667	131.8

* 1982
** 1983

Far East. Roman water pipes show that the ancient Romans were capable of manufacturing a fairly nonporous fired body. These high-quality pipes can be regarded as the first forerunners of industrial stoneware. With the fall of the Roman Empire, however, the art of stoneware manufacture was lost.

Stoneware reappeared in Middle Europe in the 11th century, favored by the availability of suitable clay, mainly in the Rhineland in the Eifel region (northwest of the Moselle) and the Westerwald (between the Lahn and Sieg rivers). Higher firing temperatures than was usual produced densely sintered bodies of great hardness and strength. At first these ceramics were unglazed. Stoneware manufacture was established in many other localities during the Middle Ages, e.g., Thuringia, Saxony, Bavaria, Silesia, and the Brandenburg March. The salt glazing process was discovered in the Rhineland in the 11th century.

In the main centers of stoneware manufacture in the 14th – 16th centuries, especially in the Raeren/Aachen region, the product was brown. Colored oxides, such as cobalt blue enamels and manganese dioxide allowed brown to violet colors to be obtained. The earliest manufacture of stoneware with a whitish to light-yellow body was in Siegburg. Typical Siegburg ware of ca. 1400 was decorated with detailed raised reliefs, which required a very fine, plastic clay body. Westerwald ware of the 16th century, especially from the Höhr-Grenzhausen region, may be recognized by its bluish-green body, incised patterns, and blue, violet, and brown decorations. Salt-glazed ware with red-brown colors was also common by this time. Very dark brown types of stoneware made their appearance in the early 17th century in the Creussen region (near Bayreuth). These were the precursors of Böttger ware (1708 – 1709), which led to the development of porcelain in Germany.

The Bunzlauer brown tableware of Silesia was not salt glazed, but had a clay glaze that was made from a mixture of several locally available earths. The establishment of technical colleges specializing in ceramics in Höhr-Grenzhausen (1879) and Bunzlau (1897) gave new impetus to the pottery industry and was a turning point in the industrial development of these regions.

Dense stoneware was first manufactured in England in the 17th century. In 1688, the Ehler brothers from Nuremberg founded a large pottery in Staffordshire; they also founded the stoneware industry in Lambeth, where their company became known as Doulton. From the 18th century onwards, English stoneware attained a worldwide reputation through the Wedgwood Company.

In the 19th century, industrial products became important. The manufacture of sewage pipes led the way. In Germany, materials were developed that were watertight, acid proof, abrasion resistant, and mechanically strong, and these replaced the porous pipes of English origin. In England, the chemical industry had attained importance during the first half of the 19th century, and thus it was here that Wedgwood and Doulton produced the first vessels, apparatus, and cooling coils from stoneware for chemical use.

In 1904 several German stoneware factories amalgamated, forming the Deutsche Ton- und Steinzeug-Werke AG in Charlottenburg.

In 1922 they combined with the company Deutsche Steinzeugwarenfabrik für Kanalisation und chemische Industrie in Friedrichsfeld (Baden). The two companies were world leaders in the field of stoneware for industrial chemical application.

Friedrichsfeld became famous for its Hoffmann stoneware body, a gas tight material that was relatively insensitive to temperature changes, and for another product, *Korundsteinzeug* (corundum stoneware), which had extremely high mechanical strength. The Deutsche Ton- und Steinzeugwerke AG provided a completely densified stoneware, *DTS-Sillimanite,* which was suitable for high-voltage insulators.

2.2. Raw Materials

The most important raw materials for all types of stoneware are the stoneware clays that contain kaolinite and fireclay (25 – 40 % total), together with other main components, illite (25 – 75 %), and sericite (→ Clays). Their relatively high alkali-metal content gives them good sintering properties. Montmorillonite can be present, although only in small amounts. The feldspar content is also small. The quartz content is 20 – 30 %, and the quartz is usually very finely divided. Quartz is beneficial in the formation of a good salt glaze. Stoneware clays in general have a very small particle size, and this is closely related to their most important properties. Between 60 and 95 % of the particles are less than 2 µm in size; consequently, stoneware clays are very plastic.

Typical stoneware clays with no added flux must be completely sintered when fired at Seger cones 4 a – 10 (1160 – 1300 °C). At the same time they must have as wide a sinter interval as possible: there must be at least five Seger cones between sintering point and melting point. Good stoneware clays sinter between 1200 and 1300 °C (Seger cones 6 a – 10) and have a cone "squat" temperature of 1580 – 1750 °C (Seger cones 26 –34). Therefore, there is excellent rigidity on firing up to the sintering point, in contrast to porcelain, and thus large articles can be manufactured. The sinter interval can be influenced by the type and amount of clay minerals or fluxes present in the clay. Illite contains alkali metals, and consequently promotes the widening of the sinter interval, but lime, magnesia, and iron oxide, especially iron oxide in a reducing atmosphere, act as fluxes and can significantly reduce it.

Depending on the iron oxide and titanium oxide content, and also on whether the kiln atmosphere is oxidizing or reducing, stoneware color is gray white, stone gray, blue gray, yellowish, or red to brown.

Typical stoneware clays can also be supplemented by plastic clays with a high quartz content, such as kaolin clays and kaolins, depending on the type of product and desired properties. However, there is no need to make use of kaolins for the purpose of achieving fine particle size with consequent high plasticity, these properties being more easily attained through the stoneware clays. Stoneware clays of various qualities (Table 3) are extracted in Germany, former Czechoslovakia (Wildstein), and England (Devon).

The amount of emphasis to be placed on clay purity depends on the type of ware. In any case, the clays must be free from coarse particles of sand and quartz, which could have a harmful effect on molding and drying. More important is the effect of these particles in producing stresses and structural damage caused by volume changes associated with quartz transitions during firing and cooling.

Good quality stoneware clays can also sometimes contain pyrites or other iron compounds such as marcasite, hematite, and siderite. These can give rise to melting defects at high firing temperatures unless they are finely ground and thoroughly mixed into the clay body. Calcium carbonate must not be present in stoneware clays because the very small particles of quicklime produced during firing slake subsequently and give rise to spalling.

Nonplastic, nonclay minerals include sand and grog, and some substances that also have a fluxing action, e.g., feldspar, feldspar sands, porphyry, and basalt. Fluxes differ from grog in that they must be finely divided or finely ground. Too much coarse material can cause porosity. The flux content must not be too high, or it can lead to distortion and formation of melting defects, blistering, and pitting.

Table 3. Chemical composition of stoneware clays, weight percent

Clay	Region	SiO_2	Al_2O_3	Fe_2O_3	TiO_2	CaO	MgO	Alkalis		Ignition loss	Reference
								K_2O	Na_2O		
Ball clay	England	59.80	26.40	1.00	1.40	0.20	0.50	2.40	0.40	7.90	[54]
Duinger stoneware clay	Lippe ∗	56.96	26.58	2.57	0.91	0.91	1.38	1.65		9.06	[55]
Lämmersbach clay 101/W	Westerwald ∗	62.95	27.95	0.55		trace		0.48	0.67	7.42	[56]
Klardorf bonding clay	Oberpfalz ∗	50.50	32.80	1.67	1.10	0.14	0.40	1.52	0.08	11.30	[56]
Niesky stoneware clay	Oberlausitz	51.80	32.20	1.60	1.70	0.05	trace	1.22		11.80	[55]
Pfalz glass pot clay	Rheinpfalz ∗	56.39	37.18	1.54	0.96	0.28	0.77	1.46	1.42	11.52	[56]
Pfalz yellow clay	Rheinpfalz ∗	60.31	31.28	3.76	1.00	0.11	0.13	1.58	1.83	8.18	[56]
Westerwald standard clay FT-W	Westerwald ∗	64.60	23.00	1.00	1.30	0.20	0.50	2.70	0.20	6.50	[57]
Westerwald stoneware clay 204	Westerwald ∗	53.60	30.00	2.50	1.60	0.40	0.50	2.10	0.10	9.20	[57]
Wildstein clay AGB	Former Czechoslovakia	53.30	29.90	3.90	1.20	1.40	0.20	3.20		6.90	[56]

∗ Federal Republic of Germany.
∗∗ Former German Democratic Republic.

2.3. Composition

Today a single stoneware clay is generally not used to make pottery, the exceptions being art and decorative tableware. Usually the desired clay properties are obtained by mixing several clays, sometimes with nonplastic fluxes. The proportions of the various clays and nonclay additions are chosen so that the plasticity and sintering properties of the clay body are suitable for the molding and firing process.

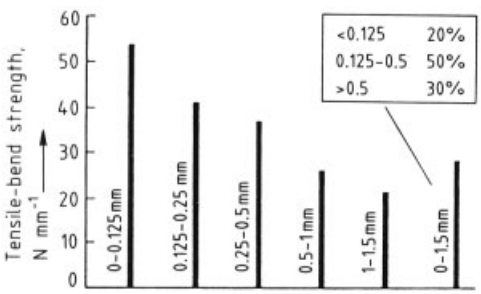

Figure 14. Effect of particle size of added grog (chamotte, chamotte : clay = 30 : 70) on strength of stoneware [59]

Clay bodies, normally quite plastic, are rendered less so by the addition of fired bodies ground to a definite particle size between 0.2 and 1.8 mm. This addition reduces drying and firing shrinkage and, therefore, sensitivity to the drying process. It also improves the rigidity of the mass during firing. Suitable additives are fired stoneware, fired porcelain, or grog produced by firing stoneware clay in a shaft furnace [58]. The amount, particle size, and particle shape of the added grog affect the fired density and strength of the body of the ware (Fig. 14).

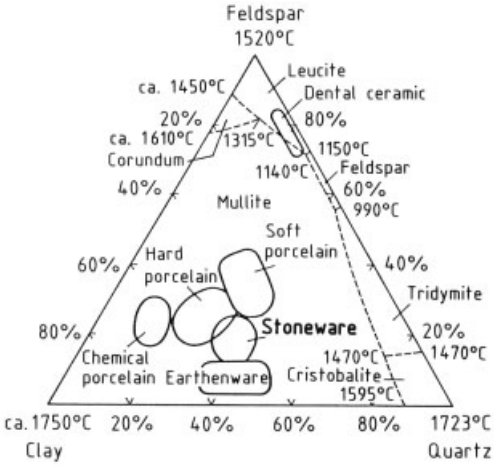

Figure 15. Some ceramic products in the three-phase system clay – feldspar – quartz (% = wt %) [5]

The chemical composition of the clay body (see Ceramics, General Survey, Chap. 5.1., for

the Seger convention) lies within the limits given by
$RO \cdot 0.33 - 7.0\ Al_2O_3 \cdot 4.0 - 44\ SiO_2$
with RO varying between 0.7 (CaO + MgO + FeO) + 0.3 K_2O/Na_2O and 0.3 (CaO+MgO+FeO) +0.7 K_2O/Na_2O

Table 4 shows the chemical compositions of stoneware clay bodies. The stoneware bodies lie, along with porcelain bodies, in the mullite precipitation zone (Fig. 15).

2.4. Preparation

The preparation process brings the raw materials into a workable form, giving a clay body mixed to the correct consistency for molding [61]. Sometimes these processes begin in the clay pit, where harmful impurities, such as pyrite or limestone, are separated as far as possible. The actual preparation process comprises the following steps:

1) Breakdown of the plastic raw material
2) Breakdown, milling, and screening of the nonclay minerals
3) Proportioning
4) Mixing of the constituents

The breakdown of the clay must disperse the clay into its finest particles to produce good plastic properties.

The type and quality of the products determine the choice of the preparation process. As the structure of the fired body becomes finer and the specifications more demanding, the preparation process generally becomes more costly. Basically, there are four preparation processes, dry, semidry, hot, and wet, with the last the one used most often.

Dry Preparation [62]. All raw materials are broken down or milled in a dry state to the required fineness. The clay and nonclay materials may be milled separately or together. The clay minerals must sometimes be predried to a moisture content of 5 % to facilitate the size reduction process. The nonclay minerals are ground to a predetermined particle size before mixing.

The coarse size reduction of the clay (down to 10 – 50 mm diameter) is done by gyratory crushers; medium size reduction is by edge runner mills, roll crushers, or cutters; and for the finest product, double hammer mills, pin disk mills, ball mills, centrifugal roll mills, single roller ring mills, or clay disintegration mills are used. A particle size of 0.5 mm can be obtained. If still greater fineness is required, wet processing is necessary.

The grog (chamotte) often needed for stoneware manufacture must first be broken down roughly with jaw crushers prior to medium and fine grinding. Cross-beater mills and pin mills reduce the particle size to 0.5 – 3 mm. Swing hammer mills, centrifugal roll mills, single roller ring mills and Maxecon ring roll mills reduce the particle size below 1 mm. The sieved milled material can be used directly. The milled material may be classified so that the predetermined particle-size distribution gives the greatest packed density. Classification of the grog is carried out with vibration, resonance, and sonic sieves or with Mogensen sizers. The sieving equipment is generally fitted with magnetic separators (drum magnets), which remove the iron particles produced by abrasion in the milling equipment.

The various plastic and nonplastic constituents are measured by weight or volume [63]. Mixing is the final stage of the preparation process. The breaking down of the clay mix is further improved, and at the same time the finely powdered nonclay components are mixed in homogeneously. Continuous mixers may be used, e.g., single and double trough mixers or sieving mixers. The countercurrent intensive mixer, sometimes with a high-speed agitator, is a batch mixer.

The pure mixing stages can be followed by a souring process, in which the plasticity of the moist clay body is improved by prolonged storage. The clay – water reaction is promoted by the action of microorganisms such as algae and bacteria. This process can be considerably accelerated by raising the temperature, e.g., by steam heating. Improved plasticity can be achieved in a shorter time with souring towers.

Semidry and Moist Preparation. Another approach involves size reduction in which the clays are treated in the dry state or moist as mined. By later addition of the finely ground nonclay minerals and the water, a plastic mass is obtained with a water content of 15 – 25 %.

Table 4. Chemical composition of important stoneware bodies, weight percent [60]

Source						Alkalis	
	SiO_2	Al_2O_3	Fe_2O_3	CaO	MgO	K_2O	Na_2O
Ancient Roman	65.62	27.94	1.60	1.25	1.33	0.39	1.42
Vauxhall	74.00	22.04	2.00	0.60	0.17	1.06	
Helsingborg	74.60	19.00	4.25	0.62	trace	1.30	
Voisonlieu	74.30	19.50	3.90	0.50	0.80	0.50	
Baltimore	67.40	29.00	2.00	0.60	trace	0.60	
Wedgwood	66.49	26.00	6.12	1.04	0.15	0.20	
China	62.00	22.00	14.00	0.50	trace	1.00	
China	62.04	20.30	15.58	1.08	trace	trace	
Japan, gray	71.29	21.07	1.25	2.82	1.98	1.03	0.44
Japan, brown	73.68	19.20	4.37	0.70	0.32	1.41	0.32
Bitterfeld	71.24	25.25	2.11	0.11	0.21	0.64	
Krauschwitz	53.77	41.34	3.34	0.03	0.01	1.40	0.10
Moscow	68.05	29.22	1.31	0.13	0.08	0.91	0.24
Rhineland	62.60	34.20	1.70	0.30	0.10	0.90	0.40

Hot Preparation. A temperature increase is produced by direct condensation of steam on the clay-body mixture, completely breaking down the clay and making molding easier to carry out. Drying time is also reduced. However, this method of preparation is used less frequently than the others.

Wet Preparation. The usual process for fine ceramics, wet preparation gives the most complete clay breakdown and finest milling, thus producing especially homogeneous mixtures. Milling and mixing take place in the presence of water. Nonclay minerals, such as feldspar and silica sand together with any hard or impure clays, are milled in rotating cylindrical mills with flint pebbles or balls made of porcelain, steatite, or aluminum oxide. Clays and kaolins sufficiently pure and finely divided not to require milling are slurried with water in blungers and then mixed with the wet-milled nonplastic materials. Both raw material streams are treated before mixing with magnets and vibration sieves to remove iron or remaining coarse impurities such as wood, coal, or coarse sand [64]. A filter press may be used to convert the slurry to a plastic mass with a water content of 20 – 25 %. Afterwards, grog or other material may be incorporated in a mixer. Alternatively, the filter cake may be dried, milled, and used in a dry press process or converted into a slip. The slurry can also be converted to granules in a spray dryer or simply to a plastic mass [65].

Table 5 sets out the preparation and molding processes used in the manufacture of various products [66].

2.5. Molding

Various molding methods are used to produce the shapes from prepared body [67], [68]. The usual processes are plastic molding, dry molding, semidry molding, and slip casting. The most important factor affecting the choice of process is the shape of the molded piece:

Small hollow bodies with uniform wall thickness by slip casting in a mold; larger ones by jollying

Bodies of any complicated shape with variable wall thickness by slip casting with a core; larger ones by molding

Small, symmetrical, thin-walled bodies by turning in plaster molds; larger ones by molding or extruding

Highly plastic, typical stoneware clay bodies are well suited for throwing, jollying, jiggering, or extruding, but unsuitable for slip casting, for which a leaner mass must be prepared. The method is often the most economical shaping method.

The most important forming process for continuous production of stoneware articles is *extrusion pressing* with a continuously operating vacuum screw press. The shaping process in the press, and especially the extrusion of the mass

Table 5. Preparation processes for various products [66]

	⟶ Fine stoneware ⟶					⟵ Coarse stoneware ⟵
	Floor tiles	Kiln linings	Split flags	Stoneware pipes	Clinker	Roofing tiles and bricks
Consistency and granule size	0 – 2 mm	plastic liquid	plastic liquid	plastic	plastic 0 – 3 mm	plastic
% H_2O	4 – 7	18 – 25 or 28 – 40	17 – 22 or 28 – 40	16 – 20	3 – 10 or 16 – 22	16 – 24
Molding process	dry pressing	plastic pressing, modeling, slip casting	extrusion, slip casting	pipe pressing (vertical or horizontal)	extrusion, dry pressing	revolver press, extrusion
Preparation *						
Dry	2	2	2	1	1	1
Semidry	0	2	2	2	2	2
Wet	2	1	0	0	0	0

* 2 = commonly used; 1 = restricted use; 0 = not used.

through the die, give rise to flow and the particles align, causing layering and "texturing" [69]. This effect can be reduced by adding grog and thus increasing particle size or by using clays with as large a proportion as possible of particles between 11 and 38 µm [70]. In the intermittent *piston press* the texturing is avoided, but the problem can occur at the die exit.

Moist or wet pressing of plastic masses is possible by using stamping presses, friction screw presses, and hydraulic presses. Heavy metal molds or porous plaster molds are filled with excess plastic mass. Surplus material becomes flowable during the pressing operation and comes away.

An important shaping method is *trimming,* in which preformed, extruded or slip-cast pieces are affixed to molded shapes or complicated apparatus.

Dry or semidry molding requires dry or nearly dry, crumbly or flowable granules with a moisture content of 5 – 8 %. Molding is done under high pressure by partly or wholly automatic hydraulic presses, e.g., friction screw presses, and turntable or toggle presses. Large blocks can be formed by stamping a crumbly mass containing ca. 6 % moisture through use of compressed-air stamping machines.

A slip with the lowest attainable water content is required for the slip-casting process. For this reason, 0.05 – 0.2 % electrolyte, based on dry clay [71], is added to produce a good, flowable slip. Sodium carbonate and/or silicate can be used. The viscosity can be adjusted to suit the wall thickness and corresponding standing time for a casting operation in a mold or around a core. The water content should, if possible, be no more then 3 – 6 % greater than that of the corresponding plastic mass, so that the drying process is not too prolonged. Stoneware clays are generally difficult to make into a liquid slip. They are slip cast in porous plaster molds with a water content of 20 – 26 %.

2.6. Drying

The water added to produce a moldable mass must be removed by drying before firing. The drying process must not distort the shape of the molded clay or crack it, which demands great care because the highly plastic stoneware clays require much water to make them moldable. The amount of shrinkage on drying is larger for the more plastic, finely divided clays containing larger amounts of water than for the less plastic clays.

In Bourry's diagram (Fig. 16), the volume change of the mass is shown during the course of drying. Although the rate of water loss remains fairly constant during the whole drying period, the volume reduction takes place mainly in the early stages, i.e., while the water film surrounding the particles is being lost. Pores are formed when the particles touch each other and cannot move any closer. Further water loss then merely increases the pore volume.

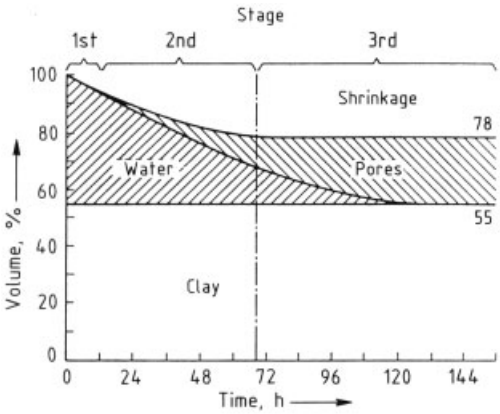

Figure 16. Bourry drying diagram for clay bodies [5]

The correlation between shrinkage on drying and water loss is shown in the Bigot curve, which has a characteristic form for each material (Fig. 17). The clay body, which is initially moldable, is brittle when the drying is complete. The drying operation must be carried out so that the changes in the state of the body take place evenly, and cause neither distortion nor cracking.

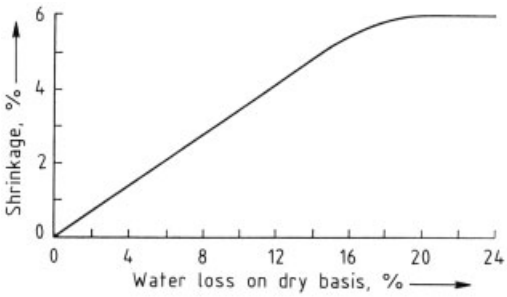

Figure 17. Bigot curve for drying a stoneware clay body

Inhomogeneities or texturing in the molded piece, sometimes produced by extrusion, causes drying stresses and raises the possibility of cracking.

The moist air drying process (→ Ceramics, General Survey) can be used for drying stoneware safely and economically [72].

After shrinkage is complete, usually when the residual water content has reached 6 – 8 %, drying can be safely speeded up by further increasing the temperature and reducing the relative humidity (Fig. 18). Total drying time and the course of each individual drying period must be established for each clay body and each type of product. The value of the lowest permissible residual moisture content should be set. The thicker the walls of the pieces and the finer the clay body, the lower this value must be, e.g., for stoneware pipes it must be < 1 %.

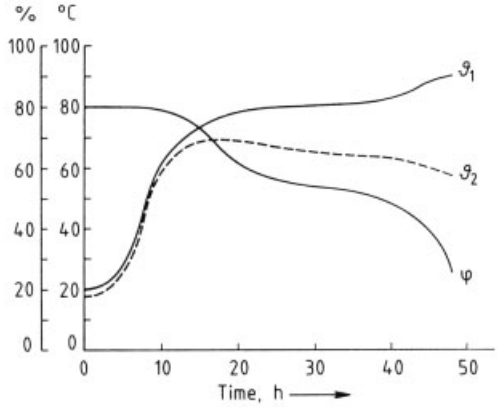

Figure 18. Moist air drying process for stoneware
φ = relative humidity, %ϑ_1 = dry-bulb chamber temperature, °Cϑ_2 = wet-bulb chamber temperature, °C

Depending on the type of product, batch dryers (floor dryers, chamber dryers, channel dryers) or continuous dryers (tunnel dryers) may be used [73].

2.7. Firing

The clay body or molded pieces after drying possess sufficient strength to undergo glazing, transportation, or setting in the kiln. During firing, processes take place within the body that are known collectively as sintering. Stoneware undergoes liquid sintering: a glassy phase is formed from the components of the body, which dissolves and binds together the other nonmelting components. The firing temperature must be chosen so that it is high enough to cause strengthening. However, it must not be so high that an excessive deformation occurs in addition to the normal uniform size reduction. The firing temperature of stoneware is 1100 – 1300 °C. The distinctive feature of the sintering of stoneware is that strengthening takes place while shape is retained.

The sintering properties of the stoneware body are dependent on a large number of factors: the chemical and mineralogical composition of the body; the condition of the raw materials,

e.g., particle size, standard of preparation, and packed density; firing temperature; time; and atmosphere. The porosity of the fired body is increased by loss of free water at the start of firing, combustion of organic components of the raw materials, subsequent splitting off of bonded water, breakdown of kaolinite, and decomposition of carbonates. It reaches its maximum at 900 °C. An oxidizing kiln atmosphere in this temperature region is advantageous because the remaining combustible components within the body are burned away. Above 900 °C the firing shrinkage begins, and porosity decreases. Increased flux content gives more of the glassy phase and thus a denser body (Fig. 19). This process takes place more readily when the clay body is finely milled, especially the fluxing component. Higher temperatures reduce the viscosity of the molten phase and cause the quartz to begin to dissolve, more readily if it is finely divided. Another factor promoting the sintering process is the packed density of the raw material. Every addition of grog reduces or delays sintering. All these relationships require that for each stoneware product the composition and fineness of the materials must be specified to suit the product.

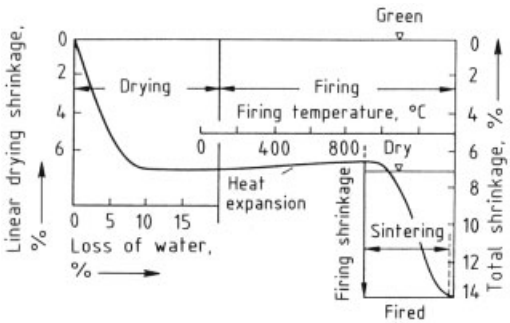

Figure 19. Drying and firing shrinkage

The firing time is also important. There are two possible ways to achieve densely fired stoneware: either by terminating the firing process after reaching a high temperature, or by holding at a low temperature for a longer time, until the optimum degree of sintering is reached. The second method produces less distortion and eliminates blistering caused by overfiring.

Rigidity and strength development during firing are partially a consequence of reactions between the decomposition product of kaolinite, metakaolinite, and molten feldspar to produce mullite crystals, which strengthen the structure.

After the first (oxidative) firing stage, which produces the greatest porosity of the body, the final sintering can be carried out in either an oxidizing or a reducing atmosphere. The choice depends on the desired product. In a reducing atmosphere, i.e., firing with an oxygen deficiency, the reduction must take place between 900 and 1000 °C, so that it can still be effective before the pores in the interior of the body become closed by sintering. If the reduction starts at a lower temperature, below 800 – 850 °C, the Boudouard reaction can take place:

$$2\,CO \rightarrow CO_2 + C$$

and this can give carbonaceous deposits and blistering at higher temperatures. Reduction causes the conversion of Fe_2O_3 in the raw material to FeO, which reduces the viscosity of the molten phase and promotes sintering by acting as a flux.

The depth of the color of the stoneware body depends on the iron content. Oxidatively fired bodies are light yellow to brown; reductively fired bodies are light to dark-gray or blue-gray.

Firing is followed by cooling. The rate of cooling is determined by the type and size of products. The stoneware body, because of its composition after firing, still contains free quartz, and must be cooled slowly in the regions of the quartz transition (575 °C) and cristobalite change (200 – 240 °C).

The furnaces may be intermittent, e.g., chamber kilns, bogie hearth kilns, and top hat kilns [74], or continuous, e.g., tunnel and fast firing kilns [75]. The fuel may be coal, heavy oil, light oil, or gas.

Setting the ware in the kiln is an important operation. Shrinkage must be trouble-free, with no quality problems resulting from deformation. If necessary, the ware may be placed on some supporting material that shrinks at the same or a similar rate to the ware. Because the setting arrangement is often open, i.e. without saggars, fired refractory materials are often used, such as grog, cordierite, or silicon carbide, so that the ware may be equally distributed over the whole of the firing space.

To prevent ware softened through sintering from sticking to the supporting surface, the surface is covered with a suitable engobing, or it

may be covered with sand, quartz, or alumina-fiber paper.

2.8. Glazing

The traditional stoneware glaze is *salt glazing,* which is formed at the end of the firing. When the ware is already partly sintered, common salt is thrown onto the burning bed of coal and blown into the kiln by a blast of air, or sprayed into the kiln as an aqueous solution. The process is repeated several times during the firing. The salt is decomposed by the steam in the kiln atmosphere:

$$2\,NaCl + H_2O \rightarrow Na_2O + 2\,HCl$$

The Na_2O reacts with the SiO_2 and Al_2O_3 on the surface of the body and forms a fused coating, which on cooling solidifies to a glass.

This glaze formation requires temperatures $\geq 1100\,°C$ and a well-sintered body to avoid penetration of the glaze into a porous surface. It also requires that the body have a sufficiently high silica content. The final color is determined by the Fe_2O_3 content of the body and by the kiln atmosphere during firing and cooling. It can be brown to red-brown, light brown, or gray. Salt glazing is rarely used today because of changed kiln technology and the need to reduce environmental pollution.

When stoneware is to be glazed, it is coated with the glaze before firing by spraying, dipping, pouring, or brushing. The glaze may simply be a clay, or a feldspar, which may be colored with metallic oxides or colored bodies. The color range is limited owing to the high firing temperature, but an attractive range of effects is obtainable with the help of glazing technology: running glazes, craquelure glazes, matt glazes, ash glazes, crystal glazes, and shrinkage glazes.

2.9. Stoneware Products

2.9.1. Clinker Ware

The term clinker includes relatively thick slabs for floors and façades, wall and street surfacing, and clinker bricks, all having a sintered body and made from usually strongly colored clays (brown to reddish-black when fired). They are extremely hard, strong, and abrasion resistant. The method of manufacture is similar to that for bricks. For shaping, the extrusion press is usual for bricks, and the dry press is usual for slabs. Firing temperatures are 1150 – 1250 °C (see 446).

Standards for unglazed floor slabs are DIN 105 and Preliminary Standard DIN 18 158 (Dec. 1978).

2.9.2. Sewer Pipe

Stoneware is the classical material for pipes carrying domestic and industrial sewage. In the Federal Republic of Germany between 1980 and 1985, the consumption of stoneware pipes was ca. 300 000 – 350 000 t/a. This extensive use of stoneware is a tribute to outstanding properties: impermeability, strength, and corrosion and abrasion resistance.

The first sewer systems, which were based on clay ceramics, were constructed by 6000 B.C. in Turkey, although, of course, they were not completely impermeable (cf. Fig. 2). The first sewer system in Frankfurt am Main was built in 1200 A.D.; the first in England, in 1840. Hamburg's sewer system, begun in 1842, first used clinker pipes, then in 1875 stoneware, and in 1900 some concrete. The first factory for stoneware pipes was founded in Germany in 1852.

Preparation. The manufacture of stoneware pipe requires a plastic clay body made up of as many as five stoneware clays, so that any negative effects caused by variability of the properties of a single clay are averaged out. Clays low in chalk may be added, and other additives include basalt, sand, porphyry, and feldspar. In addition, 20 – 35 % powdered grog of grain size 0 – 2 mm may be added to reduce shrinkage, thus improving dimensional accuracy, reducing sensitivity to drying, and increasing rigidity of the product during firing [76]. The grog can be made from waste kiln furniture (support rings) of the same material, pottery fragments, and specially fired clay grog, e.g., bought-in material made in a shaft furnace. Fine-grained porphyry (0.8 mm) may also be used, either alone or with other additives.

A stoneware grog for sewer pipes has approximately the following size distribution:

<0.063 mm	3 %	0.2 – 0.5 mm	30 %
	5 %	0.5 – 1.0 mm	35 %
0.063 – 0.1 mm			
0.1 – 0.2 mm	12 %	>1.0 mm	15 %

Coarse grog should first be broken down in a jaw or hammer crusher to a size of ca. 150 mm. Final size reduction to ca. 10 mm is carried out in a Symons cone crusher or an impact and gyratory crusher. For uniform fine grinding to ≤ 3.5 mm, the Hazemag impact mill may be used or, more often, the Maxecon ring roll mill. The Mogensen vibratory sieve, which cleans itself by electrostatic repulsion, is the most suitable for separation into definite size ranges. With this equipment the mesh size of the top sieve should correspond to the size of the coarsest feed material.

The ingredients for the clay body could be 30 % extremely plastic stoneware clay, 22 % plastic stoneware clay, 10 % lean clay, 8 % loam, and 30 % grog.

The *clays* for stoneware pipes are not always usable in the raw state, but must undergo preparation. The quality requirements for stoneware, particularly in regard to strength, require the semidry or moist processes. The dry preparation process is hardly used.

In the *semidry* process [77], the various clays first arrive at a circular feeder, which measures out the individual clays in definite proportions. The moist clay mixture produced is then squeezed through the small sieve holes of a purifier, which removes gross contaminants such as coal, stones, and iron minerals. After passing through a roll crusher, the clay mixture is carried by band conveyor to a double trough mixer, where the grog is added. Finally it goes to a sieve mixer.

Molding. Stoneware pipes with sockets are formed by vacuum extrusion mainly by vertical screw presses. The water content of the stiff plastic mass is < 15 – 17 %. Although stiff clay bodies have a tendency to texturing, they may be used for manufacturing straight pipes up to 2 m in length and placed immediately on palettes or trucks where they keep their shape unsupported. They also dry more quickly. Of course, such pipes must be formed with heavy, powerful presses, e.g., 125 – 150 kW with a compression cylinder diameter of 0.45 m.

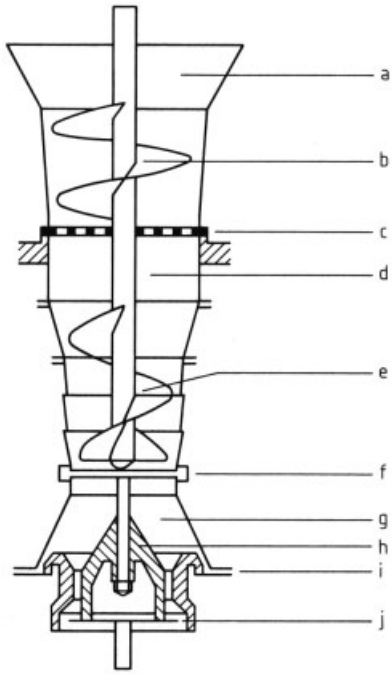

Figure 20. Screw press
a) Feed hopper; b) Prepress; c) Grid; d) Vacuum chamber; e) Press cylinder; f) Holder; g) Bell; h) Former; i) Die; j) Table

Both pipe and socket are formed by the extrusion press. In order to make the socket, a former, fixed to a vertically movable table, is brought up to the die exit before the start of the extrusion process and locked in position, where it exerts the necessary back-pressure against the thrust created by the screw press (Fig. 20). When the socket has been formed, the table is lowered to correspond to the rate of extrusion of the body of the pipe. Grooves are made in the plain end of the pipe, the socket end having been grooved during its formation. The pipe is then cut off and picked up by a gripping or suction device and placed on a palette or drying-chamber bogie. The density of the extruded article is influenced greatly by the form of the screw, i.e., the helix angle and the shape of the part nearest the mouth of the extruder.

The entire operation of the extrusion press, including groove making and cutting off and removing the finished pipe, is usually fully automatic [78]. So-called *top clay* pipes (Cremer +

Breuer, Frechen) constitute an interesting development in the automation of stoneware pipe manufacture. These pipes receive an inner glaze coating during the extrusion process and are immediately placed vertically on the tunnel kiln car. The setting ring, which is necessary to prevent deformation, is extruded at the same time in the form of an extra length of pipe. Firing in the tunnel kiln is carried out immediately after the drying process, and the usual transfer from drying cars to firing cars is thus omitted. *Top clay* pipes are unglazed on their outer surface.

The German Standard DIN 1230, Part 1, specifies that stoneware pipes be manufactured with a nominal diameter of 0.1 – 1.0 m and in lengths of 1.0, 1.25, 1.5, and 2.0 m. Larger diameters are available by special order only. Pipes with extra carrying capacity are specially strengthened (1.5 times normal wall thickness). The grooves formed in the pipes and sockets enable them to be joined tightly together.

In the United States, England, Belgium, and Holland, smaller diameter pipes are both extruded and dried horizontally. This technique is very popular in the United States, where the shale clays may be extruded in a rigid form and do not deform to oval.

Curved pipes are formed by using horizontal vacuum presses. After the socket has been formed, the emerging pipe is bent around a suitable guide. The process is automatic. The manufacture of branched pipes, at one time a matter of hand joining separate pipes, is now mechanized.

Drying. Chamber or, more often, tunnel dryers are generally used. The pipes, usually standing vertically, are dried in 24 – 48 h to < 1 % residual moisture by the moist air process. In tunnel dryers with so-called wandering cell drying, the pipes, standing on the grid floors of tunnel cars, pass through a drying tunnel with no blown or circulated air but with finned tubes at floor level carrying hot water. This hot water can be produced in the cooling zone of a tunnel kiln [79]. The heat rises from below through the vertical pipes and the moisture can escape through the roof of the tunnel, the permeability of which increases along the length of the tunnel.

Glazes and Glazing. Although the pipe body after firing is fully sintered and impervious, the surface is usually glazed, at least on the inside, to give a smooth, abrasion-resistant finish. The roughness factor K is thus decisively reduced, values between 0.02 and 0.15 mm being obtained. The earlier process of salt glazing of pipes in tunnel kilns is no longer carried out. Instead, the pipes are glazed by dipping in loam or feldspar glaze immediately after forming or drying [80]. With small- and medium-sized pipes the entire contents of a drying car can be glazed in one operation. The loam glaze consists of iron oxide-containing loam, marl, and flux-rich clay. Loam glazing is done at Seger cone 3 a – 12 (1140 – 1350 °C). Because the melting interval is only 3 – 4 Seger cones, the melting point is reduced by adding other fluxes such as pumice, basalt, calcite, dolomite, wollastonite, zinc oxide, feldspar, or nepheline syenite. The brown color produced by the loam or clay is itself insufficient, and is intensified with iron oxide and/or manganese dioxide.

The composition of a glaze for Seger cone 6a – 10 (1200 – 1300 °C) corresponds to:

0.4 CaO
0.25 MgO
0.25 K_2O
0.1 Na_2O
0.35 – 0.5 Al_2O_3
0.2 Fe_2O_3
3.5 – 4.5 SiO_2

A glaze composition for Seger cone 7 – 9 (1230 – 1280 °C) is 20 % glazing loam, 25 % Niederahr clay, 30 % feldspar K 40 (Mandt), 20 % dolomite, and 5 % calcite.

The glaze is ground in a well ball mill. This glaze must not have more than 5 – 10 % oversize on a standard 0.063 DIN 4188 sieve.

Firing. Chamber kilns and circular chamber kilns have now been replaced by bogie hearth kilns and tunnel kilns. Because of their greater economy of operation, tunnel kilns have generally taken over. Where chamber kilns are still in operation, they are fired with grid gas or oil. The heat consumption is 8 – 15 MJ/kg of ware. In contrast, large tunnel kilns specially constructed for stoneware pipe manufacture and having a width of 3 – 4 m, even 6 m (equal to the car width), require only 3 – 4 MJ per kilogram of ware. Tunnel kiln cars carrying 0.5 t/m^2 can give a daily output of 100 – 150 t. The kilns are 100 – 140 m in length and have a height that will ac-

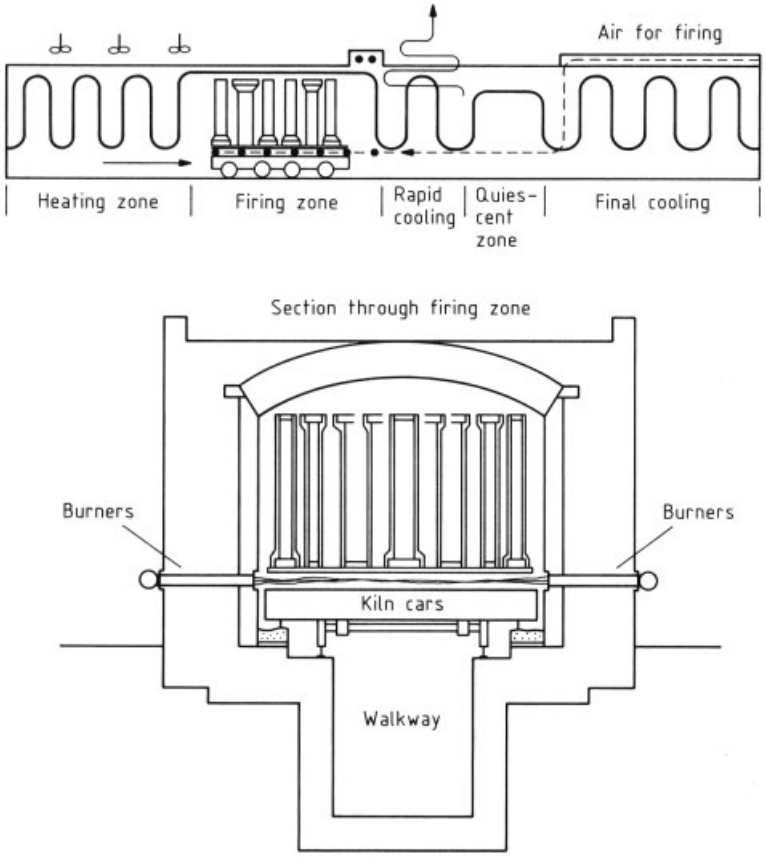

Figure 21. Cremer's tunnel kiln system

commodate the pipes in a vertical position. Pipes 1 – 1.5 m long are fired with additional curved and branched pipes placed on top, but 2-m pipes are fired alone.

To prevent deformation and loss of the circular cross section, the pipes either are fired with setting rings which contract along with the pipes, or are made extra long but scored to enable easy separation after firing.

During firing there must be equalization of temperature over the whole cross section, from top to bottom, to avoid stresses in the pipes, and the same is true during cooling. However, cooling can be rapid – 400 °C/h – until the 800 – 700 °C region is reached, when cooling must be slowed, becoming as slow and steady as possible in the region between 600 and 500 °C, where the quartz transition takes place.

The cooling system of a reliable tunnel kiln consists of water-cooled tubes in the roof of the tunnel. The heated water can be cooled in a heat exchanger. The cooling of the kiln may be made to follow a predetermined cooling curve by automatic control of dampers under the cold-water tubes. Other cooling systems remove the heat by injecting air into the cooling zones and extracting it at a higher temperature from other locations.

A particularly well-established system for firing stoneware pipes is Cremer's tunnel kiln system, which uses side firing, a hot water cooling system, and continuous operation (Fig. 21).

The tunnel kiln often has a preceding predrying space, where the pipes are heated to ca. 150 °C. During the following heating period, all the remaining water, including absorbed water, is completely driven off. However, the rate of heating must be low enough to avoid spalling. Also, in this zone all organic components should

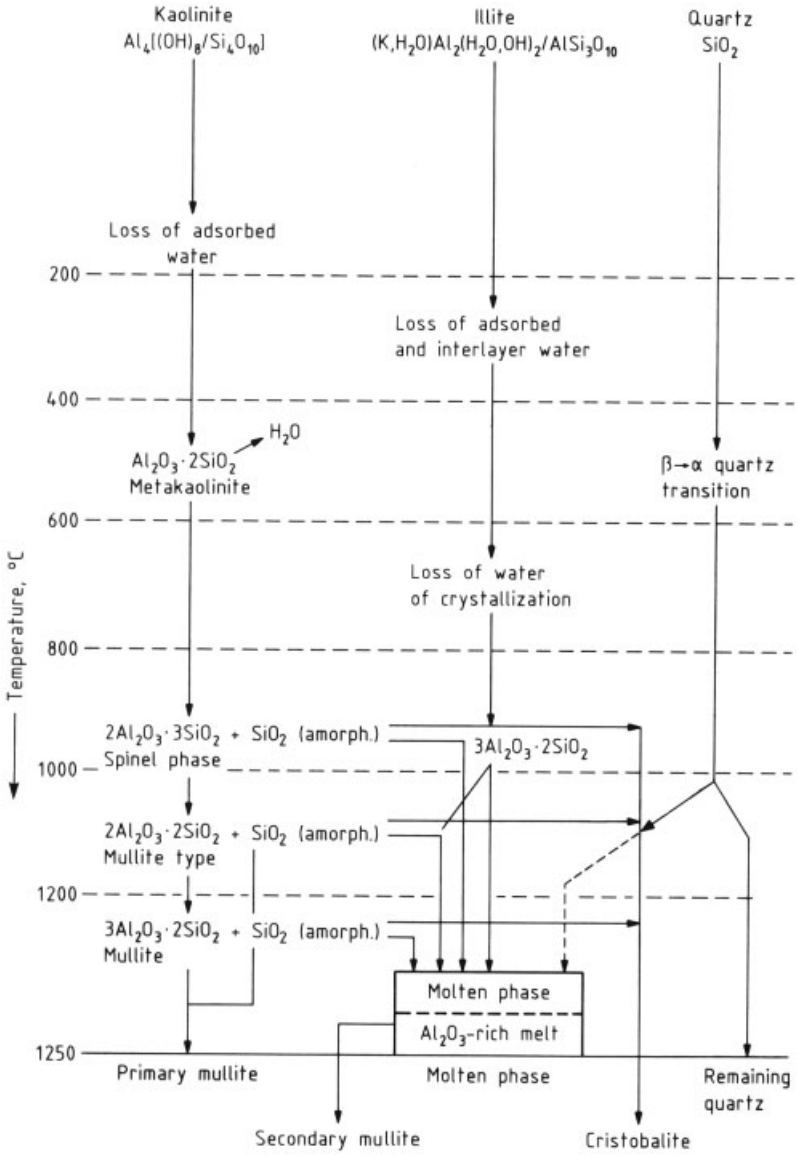

Figure 22. Reactions during firing

be burned off. The maximum firing rate is between 800 and 950 °C.

The firing zone must densify the body thoroughly. The vitrification, dissolution of quartz, and mullite formation must take place evenly over the entire cross section of the body.

Skillful arrangement of the burners, optimal air flow pattern, correct setting of the pipes on the kiln cars, and a suitable throughput time can ensure an even distribution of temperature across the whole tunnel cross section.

The processes taking place within the cooling zone determine the properties of the vitreous and quartz phases. As long as the vitreous phase still contains viscous components (the case down to ca. 750 °C), rapid cooling may be used, because the resulting stresses are relieved by plastic deformation. Immediately below this temperature, cooling must be very slow, especially in the re-

gion of the quartz transition at 573 °C, where a quiescent zone is necessary.

During firing, the reactions shown in Figure 22 take place.

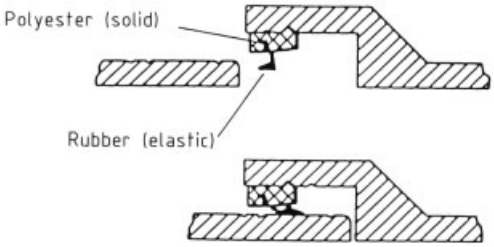

Figure 23. Steckmuffe L

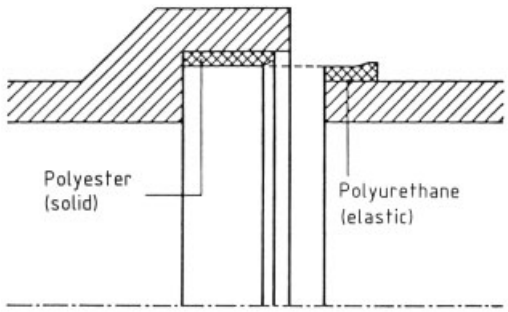

Figure 24. Steckmuffe K

Kaolinite, which has a layer structure, loses adsorbed water and water of crystallization to form metakaolinite, which then is converted to a spinel phase, a type of mullite, and finally, to primary mullite. During these reactions amorphous silica is liberated in several steps, and largely dissolves into the molten phase, but can also be changed to cristobalite. Illite gives up its adsorbed interlayer and crystallization water, and forms a molten phase at 920 °C and illite – mullite at 1000 – 1050 °C.

The chemical composition of fired stoneware materials can vary between the following limits:

60 – 70 %	SiO_2	0 – 1 %	CaO
20 – 30 %	Al_2O_3	0 – 1 %	MgO
1 – 4 %	Fe_2O_3	0.1 – 0.5 %	Na_2O
1 – 2 %	TiO_2	1.3 – 2.5 %	K_2O

However, the properties of the product depend less on the chemical analysis than on the mineralogical composition:

35 – 50 % vitreous phase 10 – 25 % quartz
15 – 30 % mullite 0 – 15 % cristobalite

If the proportion of vitreous phase falls below 25 %, the strength is low, while too much vitreous phase means a brittle product. The mechanical strength is also affected by the mullite, which is formed on firing, especially its particle size and the way in which it is embedded in the vitreous phase. Small crystals of mullite improve the properties of stoneware.

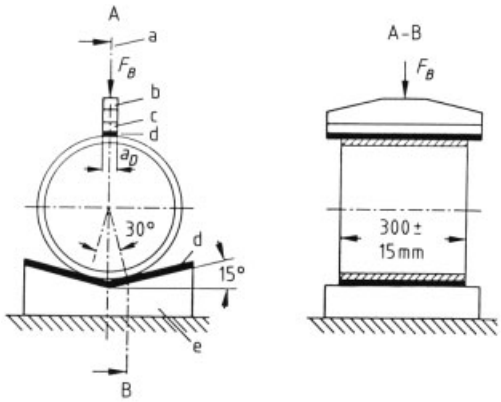

Figure 25. Compression strength testing of pipe sections

The density of stoneware sewer pipe is ca. 2.5 g/cm^3, the unfired density being 2.1 – 2.3 g/cm^3 and the total porosity (from texturing) between 2 and 15 %.

Jointing. The component parts of a stoneware sewer system (pipes, fittings, and special shapes) must be connected together by long-lasting, flexible, corrosion- and temperature-resistant, watertight joints. The seal must also resist tree roots and pressure from inside or outside the piping system. At one time common bituminous materials, which were melted and poured, and tarred rope, both of which were brought to the construction site, were used. This was a time-consuming task and did not always guarantee a leak-proof seal. These have been replaced by elastic seals that are tightly bonded to the pipe. The development of these prefabricated seals (by the Friedrichsfeld GmbH, Mannheim-Friedrichsfeld) was an important contribution to the continued existence of stoneware pipes in their competition with other piping materials. The socket seals were introduced into Germany under the names *Steckmuffe K* and *Steckmuffe L.*

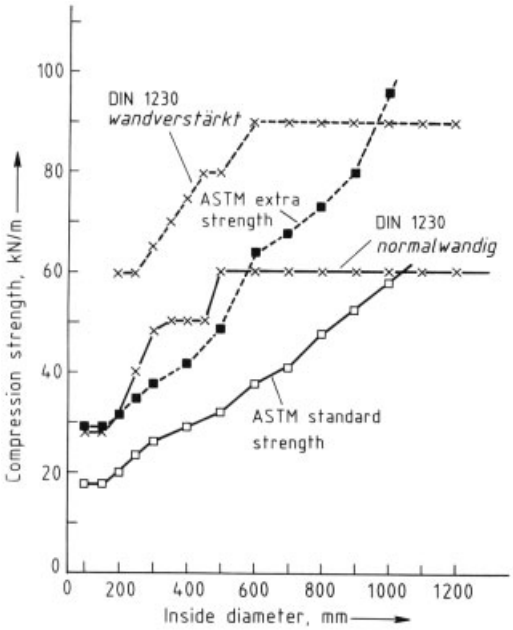

Figure 26. Minimum compression strength – DIN and ASTM

Compression provides the seal. Pipes ≤0.2 m in diameter often need to be shortened and, therefore, the seal only comprises a solid ring of synthetic polyester resin cast within the socket (*Steckmuffe L*, Fig. 23). There is an additional seal consisting of a rubber ring.

Pipes over 0.2 m in diameter are sealed with *Steckmuffe K* (Fig. 24), consisting of a silica-filled solid polyester ring inside the socket and an elastic polyurethane ring cast onto the outside of the plain end.

The compression gives increased chemical resistance.

Standard Dimensions. The standard dimensions in Germany are given for pipes with sockets and for seals in DIN 1230, Part 1 (1979) and Part 2 (1979). Testing conditions and methods of control are also given.

Testing. The most important properties to be tested, apart from dimensions, are watertightness and corrosion resistance.

Strength. Tests are carried out on pipe sections 0.3 m in length (method A; Fig. 25) or on complete pipes. In order to allow for the large effect of the socket on the compression strength, the results must be calculated by method A:

$$F_{\mathrm{N}} = \frac{F_{\mathrm{B}} \cdot 1000}{l} \cdot 1.09 \left(1 - \frac{295}{l + 1440}\right)$$

F_{N} = compression strength, kN/m
F_{B} = breaking force of pipe section, kN
l = length of pipe section, mm

The minimum values of compression strength that pipes must withstand when tested are given in DIN 1230 (Fig. 26). The ASTM standards are also given in Figure 26. The BSI standards are shown in Figure 27.

Watertightness. To prevent pollution of the groundwater, the pipe seals of a sewer system must be watertight. This requirement is tested with an excess pressure (0.5 bar) inside the pipe, corresponding to the maximum pressure of water at 5 m depth. The test takes place over a 15 min period, after 1 h during which the pipe stands full of water at the test pressure. The water loss is measured and the loss factor calculated:

$$W_{15} = \frac{V_{15}}{\pi d_1 l_1}$$

W_{15} = water loss factor, L m^{-2} (15 min)$^{-1}$
V_{15} = water loss, L/(15 min)
d_1 = inside diameter, m
l_1 = length of piping, m

The water-loss factor must not exceed 0.07 L/m^2 of inner piping surface, and there must be no formation of droplets or wet areas.

Corrosion Resistance. According to DIN 1230, Part 1, stoneware pipes must not be corroded by wastewater, groundwater, or earth material, with the exception of hydrofluoric acid; DIN 51 102, Part 1, describes the determination of corrosion resistance using test pieces of prescribed dimensions that are subjected to the action of 70 % sulfuric acid for 6 h in a boiling water bath. The loss must not exceed 0.5 wt %.

Other required properties of pipes relate to roughness and abrasion resistance of inner surfaces. The testing of seals relates to their sealing properties, mechanical properties, chemical resistance, and temperature properties. The observance of all required properties is controlled either internally or by an outside body such as the Güteschutzgemeinschaft Steinzeugindustrie, Köln (Cologne). The internal supervision comprises daily or weekly testing of the quality

of the pipes and seals, dimensions, load-bearing capacity, and watertightness. The outside supervision involves carrying out the same tests twice a year without previous notice.

2.9.3. Stable Ware

Stable ware is made from the same material as sewer pipes. The molding processes are extrusion pressing and slip casting in plaster molds. The glazed stoneware articles such as mangers, troughs, drinking basins, and gutters fulfill all hygienic requirements. They may be cleaned either roughly or thoroughly, they are not attacked by acids or fermenting materials produced by decaying fodder, and they do not allow penetration by harmful fungi or yeasts.

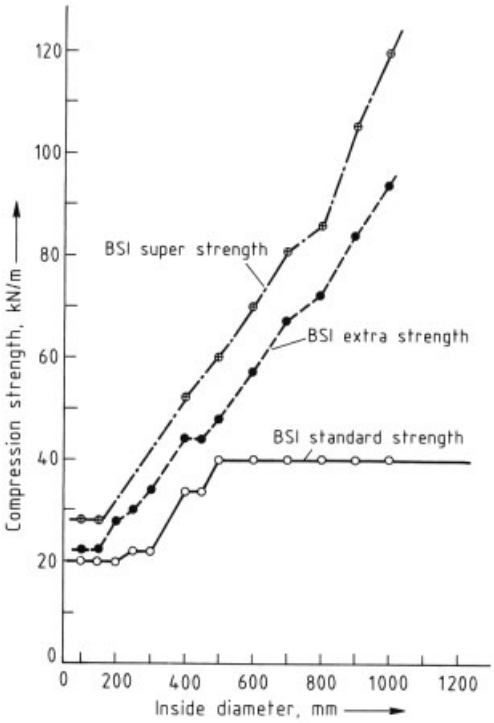

Figure 27. Minimum compression strength – BSI

2.9.4. Products for the Chemical Industry

The stoneware in the chemical industry has a variety of uses. Its properties closely resemble those of the material used for sewer pipes, split flags, and floor tiles, and resemble industrial porcelain. Chemical stoneware may have a coarse or a fine structure, although the latter is more common. Products include containers, heating vessels, filters, columns, troughs, basins, working surfaces, pipes, valves, rollers, cyclones, extractor fans, pumps, and bleaching equipment. They must resist corrosion and, very often, mechanical and thermal stresses and abrasion.

Raw Materials. High-quality raw materials such as pure, uniform stoneware clays or standardized clays are offerd by suppliers, as well as kaolins, nonclay minerals, and fluxes. If quartz is necessary, it is added as ground silica sand or quartz powder. Feldspar acts as a flux and also makes the mix leaner. Potassium feldspar and potassium sodium feldspar are used.

Composition. The composition of the clay body for chemical stoneware depends on the molding method and the desired properties. The proportion of clays with more or less plasticity is occasionally supplemented by kaolin, which can improve firing properties but can at the same time lead to greater shrinkage. The addition of fired broken stoneware ground to a particle size of 0.2 – 1.2 mm, sintered clay grog, or porcelain fragments can have several useful effects, including a leaner clay body that reduces shrinkage on drying and firing, reduced drying sensitivity, and increased rigidity during firing.

Chemical stoneware consists mainly of SiO_2 (40 – 70 %) and Al_2O_3 (25 – 50 %). The 40 – 60 % vitreous phase contains crystalline components such as mullite, quartz, and/or corundum. The approximate makeup of the clay body is:

Component	Proportion, %
Stoneware clay, very plastic	20 – 35
Stoneware clay, plastic	10 – 15
Stoneware clay, lean	10 – 15
Kaolin	5 – 15
Feldspar	10 – 20
Stoneware grog (0.2 – 1.2 mm)	20 – 30

Special Bodies. The use of selected ingredients and especially crystalline materials that may be added or that are formed during firing, such as corundum, indialith (cordierite), and

β-eucryptite, which are embedded in the vitreous phase, encourages the development of special properties: mechanical strength, abrasion resistance, thermal conductivity, and resistance to thermal shock. The enrichment or presence of the following oxides tend to produce the physical properties stated:

- MgO Better resistance to alkalis and molten metals; with other suitable additives, formation of cordierite, which has low conductivity and hence good resistance to thermal shock
- SrO Properties intermediate between those of calcium and barium stoneware
- BaO Better resistance to alkalis and molten metals
- ZnO Better resistance to molten metals without affecting resistance to acids
- Al_2O_3 Better mechanical strength
- Cr_2O_3 Better alkali resistance without loss of acid resistance
- ZrO_2 Better acid resistance

Additions of corundum and argillaceous earth increase mechanical strength and hardness. The development of these bodies by the Deutsche Steinzeugwarenfabrik für Kanalisation und Chemische Industrie, Friedrichsfeld (Baden), dates back only to the early 1920s. In an improved form as ("corundum stoneware") they are used for mechanically strong acid-resistant pumps and apparatus.

A low coefficient of expansion is necessary to achieve good thermal shock resistance. This usually incurs penalties such as difficulties in manufacture or in attaining a good standard for other properties. Quartz or vitreous silica may be added as a grog grain owing to its low thermal expansion coefficient (0.55×10^{-6}) [81]; however, below a grain size of 0.4 mm, vitreous silica can be converted on firing to cristobalite, which has poor thermal shock properties. Also, quartz with a grain size over 0.8 mm can lower impermeability. Another problem with quartz additions to the body can be the formation of hairline cracks in the glaze, giving poor densification. By using steatite, talc, or magnesium carbonate in clay bodies and firing over 1150 °C, indialith (cordierite, $2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$), can be formed. Such cordierite bodies have the low thermal expansion coefficient of $(1-2)\times10^{-6}$ between 20 and 100 °C. Furthermore, as a consequence of the small firing interval, production of densely sintered products with precise dimensions is difficult. Their use is still further restricted in the chemical industry by their poor acid resistance [82].

Products made from bodies containing barium oxide have lower expansion coefficients than mullitic stoneware, although not as low as those of cordierite stoneware. However, the products designed for good alkali resistance are better than mullitic stoneware in this respect [83].

The use of lithium-containing minerals and salts leads to the formation of β-eucryptite crystals, which have a very low and sometimes even negative thermal coefficient of expansion. The body has the extremely low expansion coefficient of $(0.6-2.0)\times10^{-6}$. The sinter interval is, however, small, and dimensionally accurate products are difficult to make. The thermal shock resistance of these products depends on size, shape, and wall thickness because of the anisotropy of the thermal expansion in different axial directions of the lithium aluminum silicate crystals.

Bodies with increased heat conductivity are obtained by the addition of silicon carbide, ferrosilicon, or silicon.

Preparation. The clay bodies are prepared by the semidry or the wet process (see page 21). The latter is preferred if superior mechanical or thermal properties are required or if the surface properties are especially important, e.g., for polished rollers, dense surfaces, or products whose glazed surface has aesthetic qualities.

Molding. Plastic molding and slip casting are the usual methods (see 2.5). The choice depends on type, shape, size, wall thickness, service conditions, and number of pieces to be produced. Rectangular vessels up to 1000 L in volume are extruded as box sections, and the side pieces are attached with slip. Larger containers are made entirely from vacuum-pressed plates (up to 2 m long) which are also attached with slip. Circular molded pieces up to a diameter of 1 m can likewise be extruded. Larger circular containers and vessels over 1000 L in volume are made by joining pre-extruded pieces in plaster molds, or they can be made on a rotating wheel in a plaster mold. For slip casting with plaster

shapes, the clay body is prepared with a water content of 22 – 26 % with the aid of electrolytes such as sodium carbonate, sodium silicate, or humic acid. The method of casting around a core is used most often.

Drying. See Chap. 2.6.

Glazing. See Chap. 2.8. A glazing process is frequently unnecessary, e.g., with polished ware, especially since the acid resistance and liquid impermeability are not produced by the glaze, but by the combination of densely sintered components.

Firing. See Chap. 2.7. The rate of heating is determined by the size and shape of the ware, the firing properties, the weight of the charge, and the kiln and burner construction.

The preconditions for an economical method of firing with rapid kiln turnaround, as well as consistent and improved quality, were provided by new circulatory and jet burners and corresponding kiln construction. The ware is now fired in gas-heated preprogrammed, rapid-firing kilns such as chamber kilns, bogie hearth kilns, and top hat kilns at Seger cone 8 – 10 (1250 – 1300 °C). One firing including cooling takes 4 – 5 days. Tunnel kilns are preferred for large batches of similar articles where the batch weight does not change continuously.

Finishing. Although shrinkage is taken into account at the molding stage, dimensional variations of up to ± 3 % are possible. For long production runs, tolerances of ± 0.5 % may be achieved. Demands for dimensions accurate within 0.01 mm may be met by machining, and this finishing process can also achieve a polished surface. The grinding medium can be a silicon carbide or often a diamond wheel. Diamond tools are suitable for drilling and cutting.

Properties. Chemical stoneware is resistant to corrosion by reactive media such as acids, solvents, and solutions of salts at all concentrations and temperatures, with the exception of hydrofluoric acid, which will damage ceramic materials even in trace amounts. Resistance to alkalis is strongly dependent on concentration and temperature (Fig. 28). If unglazed or surface ground bodies are to be corrosion resistant, they must be impermeable to liquids. Like all ceramics, chemical stoneware has high compression strength but low tensile strength. It is possible by structural means and by variation of the composition to produce special properties to suit the conditions in which pumps and apparatus operate in the chemical industry.

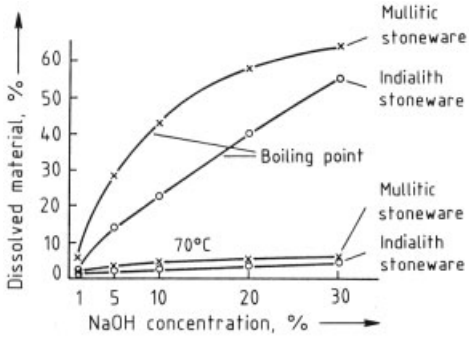

Figure 28. Solubility after 6 h of mullitic and indialith stoneware (1 g, grain size 0.25 – 0.60 mm) in sodium hydroxide solutions (DIN 51 103) [83]

Chemical stoneware is hard wearing, the hardness being 7 – 8 on the Mohs scale, equal to that of quartz or topaz. Corundum stoneware in which the corundum is very fine (< 50 µm), evenly distributed, and well bonded within the vitreous phase is especially hard wearing. Stoneware products are especially suitable for simultaneously corrosive and abrasive conditions [84]. In the "DKG-Werkstoffkennblätter für technische keramische Werkstoffe" ("German Ceramic Association Information Sheets on Industrial Ceramic Materials") [68] the areas of application, properties, and manufacturers of all industrial ceramic materials are described. Chemical stoneware materials are included under the headings DKG 150, DKG 152 and DKG 154 (see Table 6).

Use. Chemical stoneware can be used for pipes and shutoff devices, ventilation equipment [85], pumps, hydrocyclones, rollers, laboratory equipment (acid resistant basins and large working surfaces), and kitchen ceramics (sink units, working surfaces and cooking appliances).

Standard Specifications. Standard dimensions are given in DIN 7 000 – 7032 for pipes, valves, faucets, vessels etc. and in DIN 12 915

Table 6. Properties of chemical stoneware types [68]

Property	Symbol	Unit	DKG 150 chemical stoneware	DKG 152 chemical stoneware	DKG 154 corundum stoneware
Al_2O_3 content		wt %	30 – 35	25 – 35	40 – 50
Water absorbency (DIN 51 056, Section 5.2)	W_g	wt %	0 – 3	0 – 0.5	0 – 0.1
Green density	ϱ_R	kg/dm	2.2	2.3	2.5
Tensile strength	σ_{zB}	MPa	10 – 20	15 – 30	25 – 35
Compression strength	σ_{dB}	MPa	100 – 250	200 – 300	250 – 500
Transverse strength	σ_{bB}	MPa	30 – 40	45 – 65	50 – 90
Modulus of elasticity	E	kPa	50	55	60 – 70
Coefficient of linear expansion	α_t	$10^{-6}\ K^{-1}$			
20 – 100 °C			4	4 – 5	5
20 – 600 °C			4 – 5	4 – 5	5 – 5.5
Acid resistance (relative weight loss, DIN 51 102, Sheet 2)		wt %	0.5 – 0.8	0.3 – 0.6	0.2 – 0.6

and 12 916 for laboratory basins and large working surfaces in chemical stoneware, this information being also available in Werkstoffblatt (Materials Sheet) 71 of the DECHEMA-Werkstofftabelle (Materials Register). Special standard specifications govern the testing of resistance to acids and alkalis (DIN 51 102, Part 2, and 51 103).

2.9.5. White Chemical Stoneware

The special group of white chemical stoneware arose from a requirement for a material with the same properties as the brown salt-glazed stoneware but with a light-colored porcelain-like body. The *raw materials* are clay bodies of similar composition to the usual stoneware, but with clays as pure as possible, which fire to a light color. To achieve this aim, iron-free raw materials are used and prepared by the wet method. Molding processes must take into account the low plasticity of the clay body. Feldspar glazes opacified with tin or zircon are used for glazing. Flow properties and adherence (for dipping, brushing, or spraying) may be improved with organic thickeners, e.g., carboxymethylcellulose (Tylose from Kalle of Wiesbaden or Relatin from Henkel of Düsseldorf) [86]. The following Seger formula (soft porcelain) is suitable:

0.5 CaO
0.1 MgO
0.15 ZnO
0.25 K_2O
0.3 Al_2O_3
2.8 SiO_2

The ingredients are:

44 % Norwegian feldspar	12 % calcite
6 % dolomite	4 % kaolin
4 % zinc oxide	30 % silica flour

White stoneware is fired only in pure, oxidizing atmospheres at Seger cone 8 – 10 (1250 – 1300 °C) in chamber or bogie hearth kilns heated with gas or light oil.

It is only a small step from white stoneware to so-called industrial porcelain, which is produced from fine clay bodies without added grog, is fired at Seger cone 8 – 12, and resembles soft porcelain bodies. It is made into industrial articles such as kettles, columns, pipes, rollers, valves, and faucets whose outstanding property is mechanical strength. The clay bodies may be fired in oxidizing or reducing conditions.

2.9.6. Tiles and Slabs

Ceramic tiles of all kinds will soon be subject to a unified European standard laid down by the Comité Européen de Normalisation (CEN), and stoneware wall and floor coverings occupy an important place in this. According to the proposed CEN definition, tiles are regarded as building materials when they are used to cover floors and walls both indoors and outdoors, irrespective of size or shape. They can be unglazed, glazed, or engobed [87]. In the future these products will be classified according to molding method and water absorption. The following

will thus count as stoneware: split clinker flags, extruded slabs, and dry pressed or slip-cast tiles, all with a water absorption of not more than 3 %.

Split Clinker Flags. The name comes from the method of manufacture. Two slabs that are weakly bonded together back to back are split apart after firing. The parting line is formed by knives set in the die of an extrusion press. This makes a scratch on the core. The single slabs have grooves with a dovetail shape in the direction of extrusion on the back, and these give a good bond when mortar is applied. Split clinker flags are used as frost-resistant building components for indoor and outdoor wall coverings and for swimming pools. They are manufactured in various sizes, shapes, and colors, both glazed and unglazed.

The body corresponds to typical stoneware compositions with mixtures of stoneware clays of various plasticities:

30 – 50 % stoneware clays of varying plasticity
10 – 25 % grog or powdered fired body
10 – 20 % silica flour
5 – 20 % feldspar
0 – 5 % talc

The required grain size of the grog or the powdered fired body depends on the desired surface structure. Fine, smooth surfaces require a grog milled to a particle size $< 60\,\mu m$. For a rustic effect, which at the present time accounts for 60 – 65 % of production, the grog can be ≤ 1 mm.

For open-air swimming pools a grade made with fine grog is usually used. The raw materials are often prepared in high-output installations by the dry and semidry processes. The vacuum extrusion molding process is automated, including the cutting and setting of drying racks. The slabs are stacked on their longitudinal edges in chamber dryers with good air circulation or dried in continuous dryers with adjustable lateral air injection. Then they are carried on conveyor belts through one or more automatic glazing stages, where they are coated with feldspar glazes by spraying, pouring, or hosing. After this, they are automatically set onto kiln cars and fired at Seger cone 6 a – 10 (1200 – 1300 °C) in an oxidizing atmosphere in a tunnel kiln, where the energy consumption is 2.7 – 3.0 MJ/kg. The final processes of splitting, sorting, and packing are all automated.

Quality requirements and testing procedures are summarized in DIN 18 166. The following important individual test methods are planned: water absorption (DIN 51 056), transverse strength (DIN 51 090), acid and alkali resistance of the glaze (DIN 51 092) and acid and alkali resistance of the unglazed slabs (DIN 51 091).

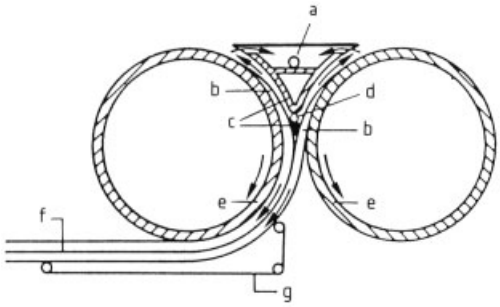

Figure 29. Schematic diagram of electrophoretic formation of clay body sheets
a) Outlet for reusable excess slip; b) Clay body particles deposited by electrophoresis; c) Counterelectrode (cathode); d) Inlet for slip; e) Zinc anode; f) Two-layer sheet of clay body; g) Belt conveyor

Single Extruded Slabs. These are also formed by a vacuum extrusion press and may be given their exact rectangular or square shape by a later punch press operation. Sometimes this can also include a dry glazing step to give a rustic effect. Slabs are also produced in a punch press by a process still being developed. Continuous sheets of clay body are produced from a slip by an electrophoretic process using the "Elephant" machine of the firm Karl Händle & Söhne, Mühlacker (Fig. 29). The desired shapes and sizes (from small mosaic tiles up to 60 × 60 cm slabs) can be punched out from this by means of the Lingl punch press [88]. The output of clay body sheet is 60 – 100 m^2/h, with a thickness of only 4 – 6 mm.

Large Slabs. Since the early 1970s, large Keraion slabs have been manufactured in stoneware by the company Buchtal, Keramische Betriebe, Schwarzenfeld. The largest size produced is 1.25 × 1.60 m and normal sizes are 0.6 × 0.6 m and 0.3 × 0.6 m. All the slabs are 8 mm thick, weighing only 18 kg/m^2. Again,

molding is done with an extrusion press, but the extruded slabs must be passed between rollers until the desired size and thickness are obtained. Keraion slabs may be used for covering walls, floors, or ceilings, and can be used for interiors or façades.

Test methods suitable for Keraion slabs are given in DIN 18 166, and the properties meet or exceed the specifications: compression strength ca. 200 N/mm^2, transverse strength 30 – 35 N/mm^2 if the glazed surface is under compression and 20 – 22.5 N/mm^2 if it is under tension, modulus of elasticity 65 000 N/mm^2.

Dry-Pressed Tiles and Slabs. The European tile industry underwent a change of direction in the 1950s after far-reaching developmental work in Italy and Spain. Completely automated factories came into being, with a completely new concept. Mass production of this kind requires that the greatest importance be attached to consistency and standardization of raw materials. These are stored in large silos with sophisticated weighing equipment. The preparation method often depends on the particular types of raw material available. Thus, the wet process with spray drying may be used or the semidry or dry process. The product is a flowable granular clay body with 5 – 8 % moisture, suitable for dry pressing in high-power presses. Grain-size distribution and an even moisture distribution are vital for defect-free molded pieces. The usual automatic presses used are hydraulic presses, friction screw presses, or combinations thereof. They operate at a maximum pressure of 500 t and 15 – 18 strokes/min, up to four 15×15 cm tiles being produced per stroke. All subsequent operations of removal, transport, drying, glazing, and setting in the kiln are completely automatic. Drying occurs in a rocking dryer.

Glazing is by various means, e.g., squirting, dripping, spraying, sheepskin-covered rollers, brushing equipment, etc. Dry glazes may be applied either evenly or unevenly to produce special effects. Flowable spray-dried glaze can be applied during the pressing operation by feeding it onto the clay body in the mold from a separate push feeder and, after the bottom molding plate has been lowered slightly, pressing the glaze onto the molded piece. Another possibility is to apply the glaze preprepared in the form of sheets or foil.

For firing in tunnel kilns there is automatic equipment to load and unload saggars stacked one on top of another. Flat flame kilns for so-called fast firing are energy saving and highly automated. The throughput times are up to 60 min. The output of a 60-m-long flatflame kiln, for shapes 200×300×10 mm, with throughput times of 65 – 100 min and firing temperatures of 1180 – 1200 °C, is said to be 1200 –1500 m^2/d with a heat consumption of 2.5 –2.9 MJ/kg. Flatflame kilns are designed as single and multilayer roller kilns and as bogie kilns. In the latter, the charge is placed on solid supports of aluminum oxide, Pythagoras body, or special steel fixed to the flat upper surface of the kiln car.

The Standard Specification DIN 18 155 (March 1976) applies to stoneware tiles; in Part 1, the concept and constitution of fine ceramic tiles are defined and their applications are explained. Part 2 gives shapes and measurements; Part 4, quality requirements and testing. Stoneware tiles, whether glazed or unglazed, must have a transverse strength of at least 25 N/mm^2 (for test method see DIN 51 090). Test methods for thermal shock resistance are given in DIN 51 090, and for household chemicals, acids, and alkalis they are given in DIN 51 092.

3. References

General References

1. W. Bender, F. Händle (eds.): *Handbuch der Ziegelindustrie, Verfahren und Betriebspraxis in der Grobkeramik,* Bauverlag, Wiesbaden-Berlin 1982.
2. T Haase: *Keramik,* Fernstudium der Bergakademie Freiberg, VEB-Verlag der Wissenschaften, Berlin 1961.
3. H. J. Oel, G. Tomandl: *Das Sintern in der Keramik,* Institut für Werkstoffwissenschaften III, Erlangen 1970.
4. T. Plaul: *Technologie der Grobkeramik, Herstellungs- und Prüfverfahren,* vol. 6, VEB-Verlag für Bauwesen, Berlin 1966.
5. H. Salmang, H. Scholze: *Die Keramik. Physikalische und chemische Grundlagen,* 5th ed., Springer Verlag, Berlin-Heidelberg-New York 1968.

6. F. Singer, S. Singer: *Industrielle Keramik,* vol. 1: Die Rohstoffe, Springer Verlag, Berlin-Heidelberg-New York 1964.
7. F. Singer, S. Singer: *Industrielle Keramik,* vol. 2: Massen, Glasuren, Farbkörper, Herstellungsverfahren, Springer Verlag, Berlin-Heidelberg-New York 1969.
8. F. Singer, S. Singer: *Industrielle Keramik,* vol. 3: Die keramischen Erzeugnisse, Springer Verlag, BerlinHeidelberg-New York 1966.
9. W. Bilke: *Handbuch der Keramik,* Verlag Schmid, Freiburg, since 1966.
10. C. O. Pels Leusden: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1977, pp. 277 – 359.
11. C. O. Pels Leusden, H. B. Weber, *Ziegelindustrie* 1975, no. 7, 254.
12. G. Piltz: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1972, pp. 218 – 261.
13. H. Kromer, W. Potschigmann: "Blähton" in *Handbuch der Keramik,* group II M, Verlag Schmid, Freiburg 1977.
14. E. Strommenger: *Habuba Kabira – Eine Stadt vor 5000 Jahren,* Verlag Philipp von Zabern, Mainz 1980, pp. 39 – 46.
15. K. Benesch: *Auf den Spuren großer Kulturen.* Lexikothek-Verlag, Gütersloh 1979, pp. 176 – 183.
16. G. Piltz, *Ziegelindustrie* 1964, no. 13, 493 – 498.
17. E. Schmidt: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1973, pp. 373 – 411.
18. W. Köther, E. Hilker, E. Hesse: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1981, pp. 177 – 207. G. Piltz" E. Hilker, *Ziegelindustrie* **1973,** no. 12, 453;**1974,** no. 1, 15; **1974,** no. 2, 60.
19. D. Hauck, E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1984, pp. 15 – 60.
20. G. Piltz, E. Hilker, *Ziegelindustrie* 1974, no. 9, 374.
21. W. Köther, E. Hilker, E. Hesse: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1981, pp. 369 – 377.
22. G. Piltz: "Untersuchung der Möglichkeiten der Aufhellung der Brennfarben von Ziegelrohstoffen," Forschungsberichte NRW, no. 1323, Westdeutscher Verlag, Köln-Opladen 1964.
23. E. Hilker, D. Hauck: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1986, pp. 17 – 48.
24. W. Köther, E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1982, pp. 198 – 211. H. Schmidt, H. Scholze, G. Tünker, *Sci. Ceram.* **11** (1981) 333 – 339. E. Schmidt: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1964, pp. 349 – 368.
25. W. Köther, N. Pauls: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1982, pp. 212 – 244.
26. G. Piltz: "Vergleiche in der Grobkeramik angewandter Untersuchungsmethoden," Forschungsberichte NRW, no. 1351, Westdeutscher Verlag, KölnOpladen 1964.
27. U. Troje, H.-G. Hopp, *ZI Ziegelind. Int.* 1980, no. 5, 288 – 291.
28. C. O. Pels Leusden: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1974, pp. 172 – 226.
29. W. Unger, *Keram. Z.* **31** (1979) 26 – 27.
30. C. O. Pels Leusden, *Keram. Z.* **29** (1977) 665 – 668; **30** (1978) 93 – 97. R. Grätz, *Sprechsaal Keram. Glas Email Silik.* **102** (1969) no. 18, 764 – 787; **102** (1969) no. 22, 990 – 998.
31. R. Grätz: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1972, p. 262 – 287.
32. R. Grätz, *Ziegelindustrie* 1969, no. 9/10, 197 – 203.
33. E. Hilker, *Ziegelindustrie* 1974, no. 8, 333.
34. C. O. Pels Leusden, E. Hilker: "Erhöhung der Maßhaltigkeit von keramischen Produkten insbesondere durch Zuschlagstoffe zur Verbesserung der Formgebung und des Trocknungsablaufs," Forschungsberichte NRW, no. 2960, Westdeutscher Verlag, Köln-Opladen 1980.
35. F. R. Stupperich, *Ziegelindustrie* 1975, no. 11, 400.
36. K. Junge, *ZI Ziegelind. Int.* **38** (1985) no. 1, 10 – 22; **38** (1985) no. 4, 227 – 235.
37. C. O. Pels Leusden, *ZI Ziegelind. Int.* 1979, no. 7, 384 – 397.
38. H.-B. Weber, *Ziegelindustrie* 1973, no. 2, 46 – 54.
39. C. O. Pels Leusden: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1979, pp. 202 – 352.
40. W. Köther, *ZI Ziegelind. Int.* 1979, no. 7, 409 – 416.
41. C. O. Pels Leusden, *Ber. Dtsch. Keram. Ges.* **46** (1969) 529 – 533.

42. E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1973, pp. 213 – 253.
43. D. Hauck, E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1985, pp. 46 – 95.
44. G. Piltz, *Ziegelindustrie* 1965, no. 20, 751 – 778.
45. D. Hauck, E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1986, pp. 58 – 116.
46. W. Bender, *Ziegelindustrie* 1968, no. 8, 196 – 199.
47. W. Schellmann, H. Fastabend: *Ziegelindustrie* 1963, no. 24, 899 – 905.
48. E. Hesse, E. Hilker, *ZI Ziegelind. Int.* 1978, no. 5, 256 – 265.
49. H. Hohmann, H.-G. Krüger, J. Geilich, *Silikattechnik* **22** (1971) no. 4, 115 – 120. H. Hohmann, S. Plüschke, *Baustoffindustrie* **24** (1981) no. 2, 47 – 50.
50. E. Hesse, *Ziegelindustrie* 1969, no. 11/12, 249 – 253.
51. E. Hesse, *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1982, pp. 118 – 145. G. Piltz, E. Hesse, *Ziegelindustrie* 1972, no. 9, 432 – 436. G. Schellbach, G. Piltz, E. Hilker: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden – Berlin 1977, pp. 360 – 426.
52. G. Schellbach: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1982, pp. 14 – 51.
53. G. Schellbach: *Ziegeleitechnisches Jahrbuch,* Bauverlag, Wiesbaden-Berlin 1983, pp. 101 – 144.
54. Firmenschrift Watts, Blake, Bearne & Co. Ltd., Devon, England.
55. O. Reumann: "Eigenschaften der keramischen Rohstoffe," in: *Handbuch der Keramik,* group I A 3, Verlag Schmid, Freiburg 1968.
56. *Databook 1975,* Sprechsaal-Verlag, Coburg 1975.
57. Firmenschrift Fuchssche Tongruben KG, Ransbach-Baumbach.
58. E. Gugel, K. Schröder, E. Frank, *Ber. Dtsch. Keram. Ges.* **49** (1972) 179 – 184.
59. P. Fischer, H. A. Müller, *Marsdorfer Tech. Mitt.* **1** (1965) 5 – 22.
60. F. Singer, S. Singer: *Industrielle Keramik,* **vol. 2,** Springer Verlag, Berlin-Heidelberg-New York 1968, p. 46.
61. H. B. Ries: "Aufbereitung keramischer Massen," in: *Handbuch der Keramik,* group I C, Verlag Schmid, Freiburg 1968.
62. H. Zimmermann, *Keram. Z.* **23** (1971) 381 – 384.
63. K. Suchowski, *Keram. Z.* **27** (1975) 401 – 402.
64. G. Lengersdorf, H. Röhr, *Keram. Z.* **21** (1969) 428 – 431.
65. H. B. Ries, *Keram. Z.* **25** (1973) 454 – 460.
66. H. B. Ries, *Euro-Ceram.* **13** (1963) 249.
67. S. Lenk, *Keram. Z.* **25** (1973) 134 – 136. K. Krahl, C. Richter, H. Hässlich, *Silikattechnik* **29** (1978) 151 – 153.
68. DKG-Werkstoffkennblätter für technische keramische Werkstoffe, Dtsch. Keram. Ges. Fachausschußber. no 23, 1978.
69. G. Teubner, *Sprechsaal* **101** (1968) 752 – 758.
70. H. G. F. Winkler, F. Freund, *Ber. Dtsch. Keram. Ges.* **35** (1958) 375.
71. W. Weiand, *Sprechsaal* **109** (1976) 332 – 335.
72. A. Bergholz, K. Herdt, Silikattechnik **19** (1968) 150 – 154.
73. F. Rüb, *Keram. Z.* **21** (1969) 98 – 106.
74. R. Lenz, *Keram. Z.* **21** (1969) 438 – 443.
75. I. Gatzke, *Keram. Z.* **21** (1969) 219 – 224.
76. P. Fischer: "Kanalisations-Steinzeug" in: Handbuch der Keramik, group II D 1, Verlag Schmid, Freiburg 1972. F. Gorn, *Sprechsaal* **105** (1972) 533 – 535.
77. H. B. Ries, *Euro-Ceram.* **10** (1960) no. 2.
78. W. Richter, *Keram. Z.* **26** (1974) 638 – 642.
79. G. Cremer, *Ber. Dtsch. Keram. Ges.* **32** (1955) 365 – 368. G. Cremer, *Ber. Dtsch. Keram. Ges.* **39** (1962) 175 – 180.
80. W. Richter, *Keram. Z.* **28** (1976) 581 – 583.
81. R. Masson, *Chimia* **8** (1954) 7.
82. E. Gugel, H. Vogel, *Ber. Dtsch. Keram. Ges.* **41** (1964) 197 – 205.
83. E. Gugel, H. Vogel, O. Osterried, *Ber. Dtsch. Keram. Ges.* **43** (1966) 587 – 594.
84. E. Dörre, A. Lipp, D. Rauschert, K.-H. Schüller, H. Vogel, *Ber. Dtsch. Keram. Ges.* **50** (1973) 4.
85. K. Pfeifer, R. Roth, *VFDB Z.* **16** (1967) no. 2.
86. W. Weiand, *Keram. Z.* **31** (1979) 148 – 151.
87. M. Drews: "Fliesen und Platten" in: *Handbuch der Keramik,* group II H 2, Verlag Schmid, Freiburg 1979.
88. E. W. Schmid, *ZI Ziegelind. Int.* 1978, 217 – 220. F. Händle, *Keram. Z.* **32** (1980) 185 – 188. H. Lingl, *Keram. Z.* **33** (1981) 41 – 42.

Contraception

Do Won Hahn, R. W. Johnson Pharmaceutical Research Institute, Raritan, New Jersey 08869, United States
John L. McGuire, Johnson & Johnson, New Brunswick, New Jersey 08933, United States

1. Introduction . . . 1
2. Natural Methods . . . 1
2.1. Periodic Methods . . . 1
2.2. Breast Feeding . . . 2
3. Barrier Methods . . . 2
3.1. Chemical Vaginal Contraceptives . . . 2
3.2. Physical Barriers . . . 3
4. Hormonal Methods . . . 4
4.1. Oral Contraception . . . 4
4.2. Progestogen-Only Oral Contraceptives . . . 9
4.3. Long-Acting Contraceptives . . . 9
5. Intrauterine Devices (IUDs) . . . 10
6. Sterilization . . . 11
7. Contragestational Drugs . . . 12
8. New Approaches . . . 14
9. References . . . 16

1. Introduction

Millions of people throughout the world today use a variety of contraceptive methods to regulate human fertility. These include chemical methods, such as oral contraceptives, vaginal contraceptives, injectable/implantable contraceptives, and contragestational agents. They also include other methods such as intrauterine devices, intravaginal barrier methods; male and female condoms, surgical sterilization, induced abortion, and natural family planning. It is generally accepted that none of these methods represents an ideal method of fertility regulation [1–3] for everyone. Each has its own advantages and disadvantages. A comprehensive listing of contraceptive products can be found in [4].

2. Natural Methods

2.1. Periodic Methods

Abstinence from sexual intercourse during the fertile period of the menstrual cycle with the intent of avoiding conception is reviewed in [5]. Knaus first found, in 1929, that the fertile period in women occurs before menstruation [6]. A year later Ogino in Japan independently confirmed this finding [7]. From their studies, Knaus and Ogino developed formulas to determine the fertile and infertile days of a woman's menstrual cycle. At about the same time, van de Velde found a link between ovulation and a cyclic shift in basal body temperature [8], which can be used to determine the time of ovulation. These studies gave women the possibility of calculating retrospectively the time at which they ovulate and of abstaining from intercourse during their fertile period.

The primary difficulty with utilizing periodic abstinence (rhythm method) to prevent conception lies in the month-to-month variation in the time of ovulation. Whereas the ovum can only be fertilized during the first 12 – 24 h after its release from the ovary, sperm remain viable in the female reproductive tract for 5 – 7 d and perhaps longer. Thus, intercourse several days prior to ovulation can result in pregnancy.

More recent findings may better identify the fertile period. Changes in the quality and quantity of cervical mucus occurring 5 – 6 d prior to the mid-cycle surge of luteinizing hormone (LH), which initiates ovulation, can be used to predict the fertile period. However, it is sometimes difficult to recognize the changes in mucus. In addition, colorimetric enzyme immunoassays have been developed for the measurement of LH in urine. Since LH is rapidly excreted, the increase in urine LH levels can be used to predict impending ovulation. Recent studies suggest that urine LH testing every evening is a more reliable method of predicting ovulation within the ensuing 48 h [9]. However, despite these advances and others which will be made in the future, one must always remember the life span of sperm in the female reproductive

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_461

tract and be cautious in predicting how useful the method will be in identifying the onset of the fertile period. Research on other methods of consistently and accurately predicting ovulation are in progress.

2.2. Breast Feeding

In many societies, it is commonly believed that women who are breast-feeding are incapable of becoming pregnant. Suckling leads to a release of the hormone prolactin, which in turn suppresses the release of estrogen from the ovary and luteinizing hormone (LH) from the pituitary gland. The suppression of these hormones results in the inhibition of ovulation.

However, although breast-feeding does affect ovulation, it is generally accepted that the duration of lactational amenorrhea and infertility is variable [10] and that maximum birth spacing is achieved only when a mother is "fully" or nearly fully breast-feeding. The consensus of researchers in the area of lactational infertility appears to be that, if used in this way, full breast-feeding can provide as high as 98 % protection from pregnancy but not beyond the first six months [11]. Thus, breast-feeding is not reliable generally as the sole method of fertility regulation.

3. Barrier Methods

3.1. Chemical Vaginal Contraceptives

Vaginal contraception dates back as far as 1850 B.C., when written instructions for vaginal contraceptives appeared in Egyptian papyri. Vaginal contraceptives are among the simplest and most widely utilized forms of birth control available today. Major advancements in vaginal contraceptives have been made by conducting research into sperm motility and by the development of spermicidal assays [12] which could be used to search for spermicidal agents for inclusion in vaginal contraceptives.

In 1855, KOLLIKER reported on the effects of chemical agents on sperm motility [13], prompting the first studies in vaginal contraceptive research. In 1880, RENDELL produced the first commercial vaginal contraceptive: a suppository of cocoa butter containing the spermicide quinine sulfate. Vaginal preparations containing other agents toxic to sperm, such as boric acid, tannic acid, or mercuric dichloride rapidly proliferated. Due to their lower spermicidal potency or potential toxicity, only a few of these are still in use.

In 1932, BAKER [14] provided the basis for in vitro spermicidal assays for the evaluation of new spermicidal agents. In the late 1930s Ortho Research Laboratories developed the Sander Cramer test [15], which has been widely used since then to find new spermicidal agents and to compare spermicidal formulations. In the 1940s, the first surface active spermicidal agents, octoxynol [*9002-93-1*], 4-diisobutylphenoxypolyethoxy-ethanol, trade name Triton X-100, and nonoxynol-9 [*26027-38-3*], 7-nonylphenoxypolyethoxyethanol, trade name Triton N, were discovered and developed at Ortho Research Laboratories [16]. These two agents rapidly became the principal active ingredients in vaginal contraceptives worldwide. A third surfactant spermicidal agent, menfegol [*57821-32-6*], 4-menthanylphenylpolyoxyethylene(8,8)ether, was discovered and developed by Eisai in the late 1960s [17].

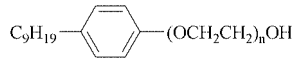

Nonoxynol-9

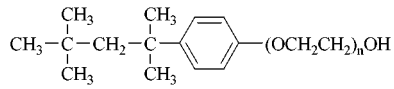

Triton X-100

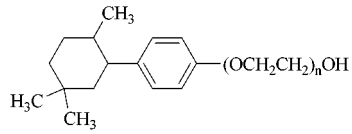

Menfegol

The prevalence of HIV infection and other sexually transmitted diseases (STD) have led to focused research efforts to develop new spermicidal agents which also have antimicrobial activity [18].

The formulations used in vaginal contraceptives consist of a relatively inert base material, which to some extent blocks the passage of sperm and also serves as a carrier for the chemically active spermicide. In the last several decades, it has been recognized that the inert formulation that contains and delivers the spermicide can affect the efficacy of vaginal contraceptives [19], [20]. As a result, physical properties of vaginal contraceptive formulations have been improved to deliver the spermicide more effectively and/or to enhance consumer compliance. Formulations include jellies, creams, suppositories, films, aerosol foams, and foaming tablets.

The major advantages of vaginal contraceptives are simplicity and safety. In addition, they do not require medical supervision. Previously considered to be relatively ineffective, it has since been shown that vaginal contraceptives can, but only if consistently used properly, be effective [20].

3.2. Physical Barriers

Various physical barrier devices are manufactured for use by men and women to regulate fertility. These include condoms, diaphragms, and cervical caps, and they may be used in conjunction with vaginal contraceptives.

Diaphragms are shallow rubber cups with a flexible metal rim that are placed over the cervix. They act both as a mechanical barrier to sperm and as a receptacle for a spermicidal agent. The major mechanism of action is believed to be spermicidal rather than barrier. To be effective, the diaphragm must fit, be inserted properly, and remain in place sufficiently long for the spermicide to act [21].

Cervical caps are available in Europe, although they have not yet been approved for use in the United States or Japan. Cervical caps are deeper and smaller in diameter than diaphragms, and they are often made of rigid plastics, such as lucite; others are made of rubber. They cover only the cervix and are held in place by suction. Cervical caps are believed to have about the same degree of effectiveness as the diaphragm.

In May, 1993, the FDA approved the first female condom, Reality, for nonprescription sale in the United States. The same device had previously been available in Europe and Canada [22]. The Reality female condom is a thin, soft, loose-fitting polyurethane sheath with two flexible rings. The female condom provides women with an opportunity to help prevent unintended pregnancy and STDs without participation of the male partner; the use of a female condom is entirely under female control [22].

Another type of barrier contraceptive device is the vaginal contraceptive sponge. Clinicians do not appear to be enthusiastic about this method, however. Studies show that the failure rate for the contraceptive sponge in parous women is higher than the failure rate for the diaphragm. A U.S. study of the contraceptive sponge found a first-year failure rate of 13.9 % in nulliparous users and a 28.3 % first-year failure rate in parous users. Other disadvantages of the method include allergic reactions in a small percentage of women and a slightly increased risk of nonmenstrual toxic shock syndrome (TSS). Compared with other over-the-counter methods, the sponge is expensive [23].

The male barrier contraceptive device is known as the condom and is widely available in most countries. The condom is a rubber sheath, sometimes packaged with a lubricant and spermicide, which serves as a cover for the penis and a receptacle for semen. The method is very effective if the condom is properly used. Usage appears to be increasing in conjunction with other methods of contraception for prevention of HIV or other STDs. By rough estimate, condoms may have been used in more than 13 billion acts of sexual intercourse that risked unwanted pregnancy, HIV, and/or other STDs [24].

4. Hormonal Methods

4.1. Oral Contraception

Combination oral contraceptives, first introduced in the early 1960s, contain a synthetic progestogen and a synthetic estrogen. There are over 60 million users of this method in the world today, and almost 150 million women have used oral contraceptives sometime during their reproductive lives [25]. The American College of Obstetrics and Gynecologists (ACOG) has published a Technical Bulletin on hormonal contraception [26]. According to ACOG, in addition to being extremely effective as a contraceptive

agent, the use of oral contraceptives is associated with a variety of health and noncontraceptive benefits, including quality of life benefits and protection from certain gynecological malignancies.

The idea that hormonal control of ovulation might be practical first appeared in 1921 in the work of HABERLANDT, who showed that extracts of the corpus luteum could make mice and rabbits infertile [27]. During the next 20 years, the female steroidal sex hormones progesterone and β-estradiol were isolated, identified, and synthesized. Unfortunately, evaluation of their clinical efficacy was severely restricted by insufficient quantities of compounds for testing. By 1943, MARKER had developed a method for synthesizing progesterone and other steroid derivatives from diosgenin, a naturally occurring steroid present in a species of Dioscorea. MARKER's studies led to the synthesis of larger quantities of these and other steroids which could be evaluated clinically for their effects on fertility regulation. For further information on hormones, see → Hormones.

Progesterone [*57-83-0*] is a natural female hormone which affects ovulation; it is largely deactivated when given orally. However, anelimination at C-19 of progesterone results in a compound that is orally active. Similarly, structural modification of steroids from other hormone classes, such as the elimination at C-19 from testosterone, results in compounds (progestogens) which have progesterone-like effects (they induce secretory changes in estrogen-primed endometrium; induce the formation of thick, viscous cervical mucus; and suppress the release of luteinizing hormone and ultimately ovulation) [28]. In 1951, DJERASSI at Syntex and COULTIN at Searle in the United States produced the substituted synthetic progestogens, norethisterone [*68-22-4*] (called norethindrone in the United States) and norethynordrel [*68-23-5*], which are still utilized in oral contraceptives today.

Other progestogens which were subsequently synthesized include lynestrenol [*52-76-7*], chlormadinoneacetate [*302-22-7*], medroxyprogesterone acetate [*71-58-9*], ethynodiol diacetate [*287-76-7*], levonorgestrel [*6533-00-2*], desogestrel [*54024-22-5*], norgestimate [*35189-28-7*], and gestodene [*60282-87-3*].

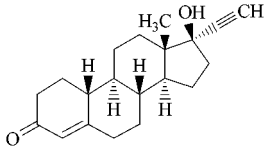

Norethindrone

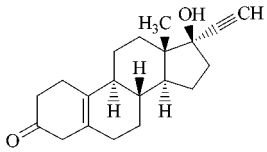

Norethynodrel

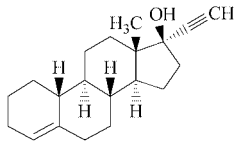

Lynetrenol

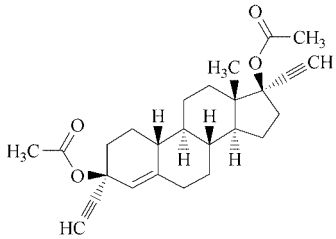

Ethynodiol diacetate

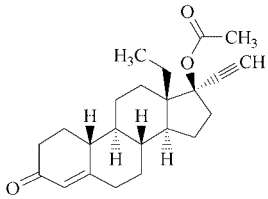

Norgestrel (levonogrestrel)

The first clinical study of an oral contraceptive began with a pill which investigators believed contained only the progestogen norethynodrel. Early clinical data demonstrated that this preparation inhibited ovulation. In subsequent studies with newly synthesized batches of the progestogen, ovulation was inhibited, but significant intermenstrual vaginal bleeding also occurred. It was then discovered that the original batch of norethynodrel, which gave better menstrual cycle control, was contaminated with an estrogen, mestranol. This led to the subse-

Table 1. Representative low-dose monophasic oral contraceptives

Trade name	Progestogen/dose, mg/d	Estrogen∗ dose, µg/d	Manufacturer
Ortho-Novum 1/35	norethindrone/1.0	35	Ortho McNeil/Janssen Cilag
Ortho Cyclen/Cilest	norgestimate/0.25	35	Ortho McNeil/Janssen Cilag
Desogen	desogestrel/0.15	30	Organon
Brevicon	norethindrone/0.5	35	Syntex
Lo-Ovral	norgestrel/0.3	30	Wyeth
Nordette	levonorgestrel/0.15	30	Wyeth
Demulen 1/35	ethynodiol diacetate/1.0	35	Searl
Loestrin 1/20	norethindrone acetate Norethindrone acetate/1.0	20	Parke-Davis
Loestrin 1.5/30	norethindrone acetate/1.5	30	Parke-Davis
Microgynon 30	levonorgestrel/0.15	30	Schering
Marvelon	desogestrel/0.15	30	Organon

∗ Ethinyl estradiol in all cases.

quent discovery that addition of estrogen to pure progestogen not only controlled intermenstrual bleeding but also increased antiovulatory efficacy. This led to the controlled inclusion of estrogen in future combination oral contraceptives [29].

Most oral contraceptives today are combinations of a synthetic estrogen and a progestational agent (progestogen). Estrogens are produced in the ovary and are important in preventing pregnancy; they work in conjunction with the progestogen to suppress ovulation. This combination also provides several ancillary contraceptive mechanisms by interfering with fertilization and implantation processes should ovulation occur [25].

The second- or third-generation oral contraceptives available today are much safer than the original products. The original combination products, introduced into clinical use in the 1960s, contained relatively high doses of estrogen and progestogen to ensure maximum contraceptive effectiveness in the absence of dose-ranging studies. The development of oral contraceptives preceded the introduction of liberal abortion laws in many countries. The knowledge that a pregnancy could not be terminated if women became pregnant in a clinical study prevented clinicians in those early studies from titrating the dose of hormone needed to prevent pregnancy down to the lowest effective dose. Epidemiological studies during the 1960s and 1970s indicated that oral contraceptive usage at that time was associated with an increased risk of thromboembolic disease. These cardiovascular complications were associated with the relatively high doses of estrogen used in original oral contraceptives, especially in women who smoke. Those findings led to the development of low-estrogen dose combination oral contraceptives (see Table 1, above).

For example, early oral contraception formulations contained up to 100 – 150 µg of estrogen. By the early 1990s, the majority of oral contraceptives contained only 30 or 35 µg of estrogen. These low-dose products are commonly referred to as second-generation oral contraceptives, and have almost totally replaced the original high-estrogen oral contraceptives. Numerous review articles and reports on cohort studies describing the use of oral contraceptives and the incidence of side effects are available [30–44].

As pharmaceutical companies continued to conduct studies to find the lowest doses of estrogen and progestogen effective as contraceptives and to market those products, women began to experience increased levels of intermenstrual bleeding. This observation led to the development of multiphasic oral contraceptives, a new approach to low-dose contraception.

Multiphasic oral contraceptives vary the dose of active ingredients or the ratio of progestogen to estrogen throughout the cycle, instead of maintaining a constant dose as in conventional fixed combination oral contraceptives. They utilize the lowest effective dose of active ingredients, but by varying the ratio, they attempt to better control intermenstrual bleeding. Some, but not all, of the low-dose multiphasic oral regimens studied significantly reduce intermenstrual bleeding and spotting [45]. The success of this approach was seen in the marketplace when the

first triphasic oral contraceptive was introduced in the United States (Ortho-Novum 7/7/7) in 1984. It rapidly became one of the most commonly prescribed oral contraceptives.

The next advance in the development of oral contraceptives was the discovery of new, pharmacologically more selective progestogens. The goal of synthesizing these more selective progestogens was to eliminate undesirable pharmacological activity but retain the required progestational/antiovulatory activity. Studies published in the 1970s and 1980s demonstrated that androgenicity, associated with the progestational component of some of the original progestogens, is associated with increased weight gain, an increased incidence of acne, and changes in lipid metabolism [46], [47], [48]. These latter changes may impact cardiovascular morbidity. Hence, oral contraceptives that do not disturb the various blood lipid fractions and do not lower HDL may be preferable to those that shift the lipid profile in an undesirable direction [49], [50]. Three companies, Johnson & Johnson's Ortho-McNeil division, Organon, and Schering AG, worked in this area, attempting to dissociate the unwanted androgenicity and desirable progestational activities of steroidal progestogens by using medicinal chemistry approaches. Norgestimate, desogestrel, and gestodene emerged from this research [51–56]. These new progestins had unique characteristics [48]. These more selective progestogens, in combination with ethinyl estradiol, compose the third-generation oral contraceptives (Table 2, see below). Norgestimate containing Ortho Tricyclen was ranked number one among U.S. oral contraceptives in 1998.

Synthetic progestogens used in today's oral contraceptives belong to one of three classes. Estranes are 19-nortestosterone derivatives; gonanes are 19-nortestosterone derivatives with a C-13 ethyl group; and pregnanesare 17-α-OH

Table 2. Representative multiphase combination oral contraceptives

Trade name	Active ingredients		Manufacturer
	Progestogen/dose∗, mg/d	Estrogen∗∗, dose, µg/d	
Ortho-Novum 7/7/7	norethindrone/0.5 (7)	35	Ortho McNeil/Janssen Cilag
	norethindrone/0.75 (7)	35	
	norethindrone/1.0 (7)	35	
Ortho Tri-Cyclen/ TriCiilest	norgestimate/0.18	35	Ortho McNeil/Janssen Cilag
	norgestimate/0.215	35	
	norgestimate/0.250	35	
Ortho Novum 10/11	norethindrone/0.5 (10)	35	Ortho McNeil/Janssen Cilag
	norethindrone/1.0/(11)	35	
Jenest	norethindrone/0.50 (7)	35	Organon
	norethindrone/1.00 (14)	35	
Tri-Norinyl	norethindrone/0.50 (7)	35	Syntex
	norethindrone/1.0 (9)	35	
	norethindrone/0.5 (5)	35	
Triphasil	levonorgestrel/0.05 (6)	30	Wyeth
	levonorgestrel/0.075	40	
	levonorgestrel/0.125 (10)	30	
Triquilar	levonorgestrel/0.05 (6)	30	Schering
	levonorgestrel/0.075 (5)	40	
	levonorgetrel 0.125 (10)	30	
Tristep	levonorgestrel/0.05 (6)	30	Asche
	levonorgestrel/0.05 (5)	50	
	levonorgestrel/0.125 (10)	40	

∗ Number of medication days in each phase in parenthesis. ∗∗ Ethinyl estradiol in all cases.

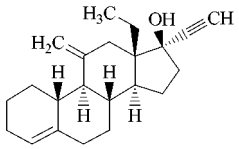

Desogestrel

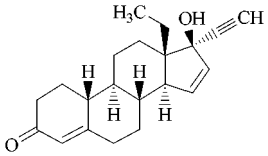

Gestodene

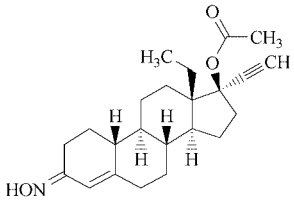

Norgestimate

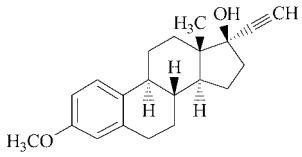

Mestranol

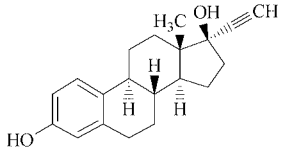

Ethinyl estradiol

progesterone derivatives similar in structure to progesterone itself.

Only two synthetic estrogens have been used in oral contraceptives to date, ethinyl estradiol [*57-53-6*], 19-nor-17-pregna-1,3,5(10)-trien-20-yne-3,17-diol, $C_{20}H_{24}O_2$, M_r269.39, *mp* (hemihydrate) 141 – 146 °C, and mestranol [*72-33-3*], 3-methoxy-19-norpregna-1,3,5(10)-trien-20-yn-17-ol, $C_{21}H_{26}O_2$, M_r310.42, *mp* 150 – 151 °C. Ethinyl estradiol differs from naturally occurring estradiol-17β by having an ethinyl group attached to C-17, whereas mestranol differs from ethinyl estradiol by methylation of the hydroxyl group at C-3. Both of these steroids resemble the natural estrogens qualitatively in their actions on the reproductive tract, pituitary gland and hypothalamus. Mestranol probably exerts most of its pharmacological activity as ethinyl estradiol following demethylation at C-3.

It is now known that estrogen receptors in various target tissues differ and that a synthetic nonsteroidal estrogen analogue can have tissue-specific agonist activity, antagonist activity, or mixed pharmacological activity. Analogously to research on progestogens, this opens the possibility of finding a new generation of more selective estrogens for use in contraceptives. As a result, there is an increasing effort within the pharmaceutical industry and academia directed toward the search for tissue-selective estrogens which retain the desired pharmacological activity but minimize potential side effects. These efforts are enhanced by deeper knowledge of the pharmacological mechanisms that underlie the selective action of estrogens [57], the recently cloned new estrogen receptor ERβ [58], and tissue distribution of estrogen receptors α and β [59].

Since oral contraceptives were introduced in the 1960s, knowledge on the benefits and risks of oral contraceptives has increased. First, numerous studies have shown that oral contraception is the most effective and reversible means of preventing pregnancy [60]. Secondly, however, it is generally agreed that the use of oral contraceptives can slightly increase the risk of venous thromboembolism, ischemic heart disease, cerebrovascular disease and hypertension [61]. The thrombolic risk appears to be concentrated among older women who smoke and who use the older higher dose products. However, several recently published articles have reported that users of third-generation oral contraceptives containing gestodene and desogestrel may be at increased risk of venous thromboembolism (VTE) [39–44]. This increased risk was not observed for third-generation pills containing the progestin norgestimate [62], [63]. Regulatory reaction has varied from advising patients to restrict use of desogestrel or gestodene-containing products to making labeling changes [63]. Thirdly, many studies have shown that women appear to obtain other, beneficial effects from oral contraceptives in addition to protection against preg-

nancy. They include protection against pelvic inflammatory disease, ectopic pregnancy, endometrial cancer, cancer of the ovaries, benign breast disease, and benign ovarian cysts [61]. Relief from a wide range of menstrual disorders is also a known beneficial effect of oral contraceptives [61].

4.2. Progestogen-Only Oral Contraceptives

Sometimes termed the mini-pill, progestogen-only oral contraceptives are widely available throughout the world but are not used as extensively as combination oral contraceptives. These preparations contain progestational 19-norsteroids such as norethindrone, lynoestrenol, ethynodiol diacetate, or norgestrel; or 17-β-acetoxy progestogens such as chlormadinone acetate and megestrol acetate. In general, the contraceptive effectiveness of these products is lower than that of combination oral contraceptives, and intermenstrual or breakthrough bleeding and spotting occur more frequently. However, these products are valuable for women who are intolerant of estrogens, and there are fewer effects on breast feeding.

4.3. Long-Acting Contraceptives

The establishment of fertility regulation with hormonally active oral contraceptives led to the development of long-acting agents which could be delivered parenterally. Long-acting contraceptives avoid compliance issues, and are useful in countries with fewer health professionals. Their popularity is based on simplicity of administration and a relatively high degree of effectiveness. Efforts to develop long-acting contraceptives have included both chemical modifications of known progestogens for injectable administration and the use of new drug-delivery systems [64].

Two well-known injectable long-acting products are medroxyprogesterone acetate (Depo Provera) and norethindrone enanthate [*3836-23-5*] (NET EN). Both are progestational products. Depo Provera (Upjohn PharmaciaCompany) is a once-every-three-months injectable administered as a microcrystalline aqueous suspension of the steroid. NET EN (Schering AG) is a two-month injectable administered as a solution of steroid in castor oil. Depo Provera is approved for use as a contraceptive in more than ninety countries. NET EN is used in more than forty countries. Other products developed by chemically modifying known progestogens include 17α-hydroxyprogesterone caproate [*630-56-8*], $C_{27}H_{40}O_4$, and dihydroxyprogesterone acetophenide [*24356-94-3*], $C_{29}H_{36}O_4$. These monthly injectables are widely usedin Latin America.

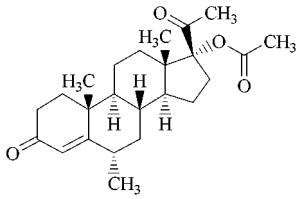

Medroxyprogesterone acetate

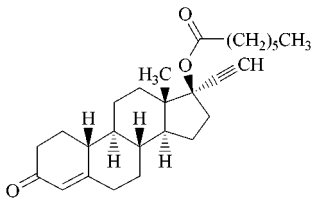

Norethindrone enanthate

Like progestogen-only oral contraceptives, parenteral long-acting progestins act primarily as ovulation inhibitors. An important second mechanism of action is their effect on the cervical mucus and endometrium, achieved at circulating blood drug levels below those required for ovulation inhibition [65].

The major disadvantage of these long-acting injectable contraceptives is the side effect of irregular menses or amenorrhea. In an attempt to reduce this, long-acting injectable contraceptives containing both an estrogen and a progestogen are being studied, including dihydroxprogesterone acetophenide, norethisterone enanthate, medroxyprogesterone acetate and norethisterone in combination with estradiol esters such as their valerate, enanthate, or cypionate esters.

Drug-Delivery Systems. Research efforts have been directed at the use of new drug-delivery systems that release drug slowly over a prolonged period of time, such as silastic capsules, rods, or vaginal rings, and biodegradable implantable systems [66].

Nonbiodegradable Implants. The best known contraceptive implant is Norplant, developed by the Population Council in the United States. Norplant is composed of six matchlike silastic cylinders containing the progestogen levonorgestrel. After implantation under the skin, the product can provide effective contraception for a period of five years [67], [68]. Since silastic is not biodegradable, the implant can be removed at any time. The main reason patients request the removal of Norplant is unpredictable vaginal bleeding episodes followed by amenorrhea. The bleeding problem is an unavoidable sequel of progestogen-only contraception.

Biogradable Implants. Utilization of implantable biodegradable polymers obviates the need for implant removal. Technically, however, it is not as easy to develop such a product as it may appear. For example, biodegradation should not take place before drug release is finished; before that, sufficient structural integrity to permit surgical removal of the implant must be maintained. In addition, release rates for the drug must be sufficient to maintain blood levels of drug which will insure effective contraception.

An example of a biodegradable contraceptive implant is Capronor. This device has walls composed of σ-caprolactone; it releases levonorgestrel and is a single implant with projected life span of 12 – 18 months [66]. The timecourse for biodegradation and the disappearance of the implant are still being defined.

Fused pellets may be composed of the drug alone or the drug fused with cholesterol [53], [69], and are formed into small cylinders by melting the drug and then solidifying it under pressure. Clinical studies with norethindrone pellets have been in progress for a number of years. Polymeric implants and pellets require minor surgery for insertion and removal.

Microencapsulated Steroids. Microencapsulation of steroids to date has utilized polylactic acid (PLA), polyglycolic acid (PGA), and their copolymers (PLGA). The ratio of PLA to PGA determines the rate of steroid release. These thermoplastic biodegradable forms have been constructed as films, micropellets, microspheres, rods, and fibers [70].

Vaginal Rings. Vaginal epithelium is readily permeable to contraceptive steroids. Since the vascular drainage of the vagina bypasses the liver, this route of administration potentially permits utilization of drugs that have low oral activity due to degradation in the liver.

Contraceptive vaginal rings consist of silastic shells or core rings of various sizes and membrane thicknesses. They have been developed for delivery of progestins alone or progestins combined with estrogens. The biology behind progestin-alone rings is conceptually similar to that of implants, but the use of vaginal rings is under direct control of the patient since they can be inserted or removed at any time without a trained professional being present. After the ring is removed, a predictable withdrawal bleeding takes place. Vaginal rings have been under development for nearly 30 years by the WHO and the Population Council [71], [72]. Progress is being made. Materials have been selected which allow release of the active ingredient. Large clinical studies must still be done, but it is already known that higher doses and/or release rates are required for heavier women than for lighter women. Acceptability of this route of contraceptive drug administration must still be determined.

5. Intrauterine Devices (IUDs)

Intrauterine devices are medical products that prevent conception when placed in the uterus. In spite of their ancient origins, modern intrauterine devices (IUDs) have been widely used only in the last 30 years. The two generic subclasses of IUDs are nonmedicated (inert) devices and medicated IUDs, i.e., progestin-releasing or copper IUDs.

IUDs are used throughout the world, with an estimated 79 million users in 1989 [73]. They are a highly effective contraceptive method with protection rates for some devices reported to be 94 – 99 per 100 women for one year. Reviews are available on the efficacy and safety of this contraceptive method [74].

Complications associated with IUDs include uterine perforation and pelvic inflammatory dis-

ease [74]. Uterine bleeding and cramping are the most common causes for discontinuation of this method. The IUD's relationship to pelvic infection, fueled by the high rate of septic abortion and pelvic inflammatory disease (PID) among users of the Dalkon Shield IUD, led to a decline in the popularity of IUDs. In fact, all IUDs, except the progesterone-releasing Progestasert, were withdrawn from the U.S. market in the mid-1980s. With time, however, the realization that this effect appears to be limited to the Dalkon Shield led to the renewed interest in IUDs in the USA, and in 1988 the copper T-380A IUD (ParaGard) was introduced into the USA.

Inert IUDs act by creating an environment hostile to sperm or fertilized ova and by blocking implantation [75], [76]. The exact mechanism of action of IUDs is not totally clear, but convincing evidence is mounting to support the idea that IUDs act primarily as contraceptives and not as abortifacients [77]. Compared with noncontraceptors, IUD users have fewer recoverable sperm in the uteri and tubes after intercourse. In addition, there are fewer recoverable ova in the uteri and tubes at mid-cycle. Those ova found are rarely fertilized. In nonhuman primates, the rate of recovery of degenerating embryos is not significantly different from that seen among controls, in contrast to what might be expected if the IUD works by preventing embryo implantation. Transient elevations of human chorionic gonadotropin (hCG), which may indicate early pregnancy, were not more frequent among IUD users than among noncontraceptors [77].

After insertion of an IUD, polymorphonuclear leukocytes and macrophages accumulate in the uterine cavity. These cells appear to phagocytize sperm and liberate a blastotoxic toxin [78], [79]. Intrauterine devices also may create a hostile environment, perhaps because antibodies are produced that interfere with implantation of the fertilized ovum [79].

The range of IUDs available worldwide is considerable. Outside of China, the Lippes Loop, made of polyethylene, is the most widely used unmedicated IUD. The other main type of unmedicated IUD, used mostly in China, is a flexible stainless steel ring, i.e., the Chinese IUD.

Medicated IUDs consist of an inert base reservoir for a uterus-affecting or spermicidal agent such as progesterone or copper. Copper-bearing IUDs have spermicidal activity and interfere with implantation [80]. Two of these, the TCu-200 and the Multiload-250 (MLCu-250), are widely available except in China and the USA. The inclusion of progestogens in IUDs does not appear to improve their efficacy but may reduce the menstrual cramping and bleeding associated with IUD usage [81].

The second generation of copper IUDs have more copper wire, copper sleeves, and/or silver core in the copper wire, denoted by Ag in the IUD name. Significant second-generation IUDs include the TCu-380A, TCu-220C, Nova T, and Multiload-375 (MLCu-375); these are available worldwide, except for China.

Progestasert is a hormone-releasing IUD containing 38 mg progesterone, released at a rate of 65 μg/d for one year. It is available only in France and the USA. LNG-20, a long-lasting levonorgestrel-releasing IUD, is still under development in Europe.

6. Sterilization

Voluntary surgical sterilization of both men and women has increased in popularity. Sterilization has been running neck and neck with the oral contraceptive in terms of popularity, each chosen by about one in four couples.

Worldwide, an estimated 42 million couples rely on vasectomy; nearly 140 million rely on female sterilization [73]. Vasectomy is a principal family planning method in the USA, New Zealand, Australia, the UK, Canada, the Netherlands, China, India and South Korea. The method is hardly used in other countries [82].

In the 1970s, questions were raised about certain immunological complications as a consequence of male sterilization or vasectomy. Clinical epidemiological data do not appear to indicate that this actually occurs in clinical practice. No significant long-term side effects of male sterilization have been demonstrated.

Voluntary female sterilization is the world's most widely used family planning method. An estimated 138 million women of reproductive age used the method in 1990, 43 million more than in 1984. Millions more were expected to ask for the method during the 1990s [83].

Because of the increasing worldwide interest and demand for simple, effective and inexpensive female sterilization, a variety of procedures and methods have been developed. These approaches differ as to whether they are performed postpartum, postabortum, or in interval situations. The choice of method also largely depends upon the physician's prior training, knowledge, and experience. Reviews on sterilization are available [84].

7. Contragestational Drugs

Pharmacological substances that either inhibit implantation or interrupt pregnancy after implantation were investigated during the 1980s and 1990s. A number of different terms have been used to describe these compounds [47–63], including anti-implantive agents, post-coital contraceptives, morning-after pills, emergency contraceptives, once-a-week pills, interceptives, abortifacients, and contragestational agents. This medical approach to fertility regulation presents several principal advantages, including potentially fewer long-term side effects as a result of short-term periodic administration and greater convenience.

However, in the decision to develop and market contragestational drugs, social, political, legal, ethical, and religious factors have been of critical importance.

Post-Coital Contraception. Post-coital contraception historically has been viewed as an emergency measure where regular contraception was not used or where the primary contraceptive methods may have failed. A number of approaches have been utilized in post-coital regimens. Early regimens utilized high doses of the estrogens [84] diethylstilbestrol, ethinyl estradiol, and the conjugated equine estrogens (Premarin). Characteristically, the drugs are taken for a period of three days. When given later than 72 h after coitus, the effectiveness is reduced; the drugs are ineffective if implantation has been established. Although the estrogens are highly effective, users suffer from a high incidence of nausea and vomiting; cycle regularity also is disturbed. A similar approach has been reported for levonorgestrel-containing contraceptives [86], [87].

The U.S. FDA recently approved specific regimen of combination oral contraceptives containing levonorgestrel and ethinyl estradiol as safe and effective for post-coital pregnancy prevention in women who have had unprotected sexual intercourse. The regimen consists of two oral contraceptive tablets taken within 72 h of intercourse and two pills taken 12 h later. When used in this manner, the treatment is about 75 % effective in preventing pregnancy. The treatment does not work if a woman is already pregnant.

Danazol is also highly effective in post-coital regimens and has a very low incidence of side effects. Danazol is marketed in many countries for the treatment of endometriosis and its availability is unrestricted. There is apparently no significant difference in effectiveness when doses of 800 to 1200 mg have been utilized for three days [86], [88]. When utilized in the post-coital mode, the precise mechanism of action of Danazol is not well-understood.

A nonsteroidal weekly pill, centchroman [*31477-60-8*] [16], was developed by the Central Drug Research Institute, Lucknow, India, and is being marketed as Choice-7 and Sahali (Hindustan Latex). Centchroman, (3,4-*trans*-2,2-dimethyl-3-phenyl-4-[*p*-β-pyrrolidino-ethoxy)-phenyl]-7-methoxychromane) inhibits implantation of the fertilized egg, thus avoiding pregnancy [89]. It exerts its antifertility effect by weak estrogenic and potent anti-estrogenic activity [90]. Although the synthesis and pharmacological actions of centchroman have been well documented [91], overall efficacy and side effects are not known or described except in a brief description of the product. A Pearl index (i.e., average of pregnancies out of 100 women who use the specific contraceptive for one year) of 3.05 was calculated for the administration of weekly doses of centchroman (30 mg) to

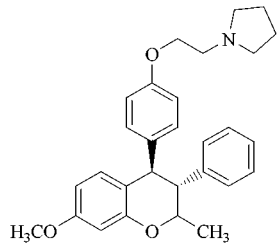

Centchroman

approximately 1600 women for a total of 20 000 months [89].

Abortifacients. It is estimated that between 30 and 40 million legal abortions and an equal number of illegal abortions are carried out each year worldwide. Most abortions are induced because the pregnancy is unintended and unwanted. In the USA, about 60 % of all pregnancies are unwanted or mistimed, half of which are aborted [92].

In mammals, removal of the corpus luteum during early pregnancy results in termination of pregnancy. Pharmacologically, this can be accomplished by various methods, including inhibition of gonadotropin support of the corpus luteum, inhibition of progesterone biosynthesis in the ovary, inhibition of progesterone binding to the uterine progestin receptor, and increase in the metabolism of progesterone. Many compounds that affect these processes have been reported [93]; some of these have been introduced for clinical study.

Chemically induced abortion originally involved the administration of the natural prostaglandin $F_{2\alpha}$ [*551-11-1*] (Dinoprost, $C_{20}H_{34}O_5$) and synthetic analogues such as sulprostone [*60325-46-4*], $C_{23}H_{31}NO_7S$, and prostaglandin ONO 802, [*64318-79-2*] (Gemeprost, $C_{23}H_{38}O_5$) have also been reported [94]. After administration of relatively small amounts of these prostaglandins, the muscle tone of the uterus increases, followed by contractions, cervical dilation and expulsion of the uterine contents. The usefulness of prostaglandins for termination of pregnancy is limited because of a high frequency of gastrointestinal side effects.

In 1988, the French government approved the marketing of an abortion pill. RU486 [*84371-65-3*] (Mifepristone), an antiprogestin developedby the French pharmaceutical company Roussel-UCLAF. It is the first clinicallyuseful progesterone antagonist. A review of the chemistry, pharmacology, and clinical applications of this compound [95] is available, as is a review on the use of RU486 alone or in combination with a prostaglandin analogue for termination of early pregnancy [96]. Other studies suggest the use of antiprogestins for contraception and for treatment of gynecological disorders related to hormone production [97]. The discovery of RU486 was followed by laboratory and clinical studies with other antiprogestins, including ZK98 299 and ZK98 734 [98]. Many issues, not only scientific, but also political and religious,

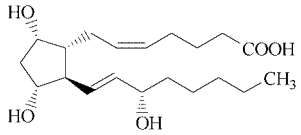

Prostaglandin $F_{2\alpha}$

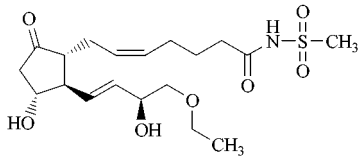

Sulprostone

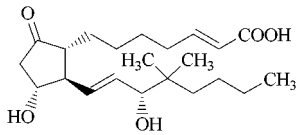

Prostaglandin ONO 802

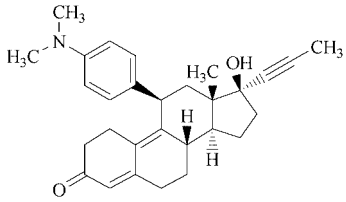

RU468 (Mifepristone)

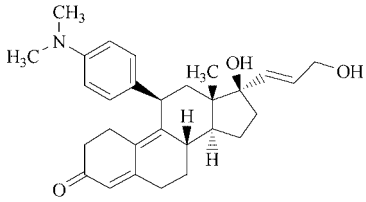

ZK98 734 (Lifopristone)

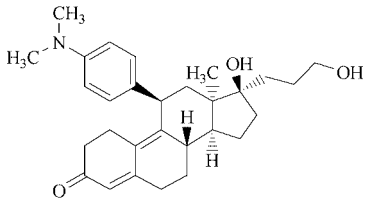

ZK98 299 (Onapristone)

surround the clinical application of progesterone antagonists, and it is difficult to project worldwide availability of RU486 and related products.

8. New Approaches

There are no methods of fertility regulation available today that are totally safe and effective or that appeal to all groups and societies in the world. Therefore, there continues to be great interest in developing new and improved contraceptives. In addition to utilizing new technology, new contraceptives should be superior to existing products. For example, oral contraceptives, used by millions of women over the last 30 years, are not only safe and effective but even protect women against some cancers. Because oral contraceptives are so effective, they have become a very high standard that other products must meet.

Contraceptive Vaccines. Major research efforts involve immunological approaches to fertility control; for reviews, see [99, 100, 102, 102]. The development of contraceptive vaccines is directed towards the immunoneutralization of the reproductive process or interfering with fertilization by inducing antibodies against oocytes and spermatozoa. Attempts have been made to develop vaccines against luteinizing hormone releasing hormone (LHRH) (also known as gonodotropin releasing hormone, or GnRH), LH, follicle stimulating hormone (FSH), human chorionic gonadotropin (hCG), placenta antigen, the zona pellucida of the ovum, and different sperm antigens.

Research on an hCG vaccine has been conducted over the past 20 years. The WHO has conducted a Phase I clinical study in Australia, using a vaccine based on a synthetic C-terminal peptide (109-141) of β-hCG conjugated to Diptheria Toxoid (CTP-DT). This study demonstrated that potentially effective contraceptive levels of antibodies were produced in vaccinated women without any adverse side effects. Phase II clinical studies are under consideration to determine if the immune response to the prototype anti-hCH vaccine is capable of preventing pregnancy in fertile women volunteers [103]. While research on the C-terminal peptide from the β-subunit of hCG has been carried out under the auspices of WHO, research supported by the Population Council and the NIH has involved two alternative vaccine candidates [101], [104].

By using advanced technologies, unique sperm antigens have been identified and partially characterized. Sperm antigens shown to have high immunocontraceptive potential are human sperm membrane antigen (SP-10) and guinea pig sperm membrane protein (PH-20). SP-10 is a sperm membrane-specific antigen of 24 – 34 kD, isolated by using a monoclonal antibody (MHS-10) that cross-reacts with the entire acrosomal region. It is associated with the outer aspect of the inner acrosomal membrane and the inner aspect of the outer acrosomal membrane of mature human sperm [104]. It has been produced recombinantly in an *E. coli* expression system. The recombinant SP-10 fusion protein is under study in the baboon.

PH-20, a guinea pig sperm protein of 64 kD, is present on both the plasma membrane and inner acrosomal membrane of sperm. It is essential for adhesion of sperm to the zona pellucida, the initial step in the fertilization process. Active immunization with PH-20 causes infertility in both male and female guinea pigs for a period ranging from six to fifteen months [105].

Another interesting sperm specific antigen is lactic dehydrogenase-x (LDH-x or $LDHC_4$), an isoenzyme of LDH confined to male germ cells. LDH-x is one of the best characterized antigens and its amino acid sequence is known. A synthetic peptide based on a portion of the molecule has been shown to reduce fertility in laboratory animals. The nucleotide sequence coding for human LDH-x has been defined and engineered into an expression vector system [100].

The zono pellucide (ZP) is the complex extracellular glycoprotein matrix that surrounds the oocyte. It plays an important role in sperm penetration and fertilization. It is a composite of several antigenic glycoproteins designated ZPI, ZPII, ZPIII, and, in some species, ZPIV. Immunologically, it does not cross-react with any other body tissues, but interspecies cross-reactivity has been observed among several species, including primates. Numerous studies indicate the immunocontraceptive potential of zona pellucida [100]. Several other antigens with good immunocontraceptive potential have been identified and investigated in laboratory animals.

In most immunocontraceptive studies to date, however, the rate and duration of immunocontraceptive effects have been less than acceptable. Two problems which are faced in immunological approaches to antifertility research is the need for a safe, effective adjuvant and suitable animal models for evaluating the efficacy and safety of the adjuvants [99]. Newer and more effective adjuvants are required for contraceptive vaccines and vaccines in general.

Luteinizing Hormone Releasing Hormone. The isolation and synthesis of luteinizing hormone releasing hormone (LHRH) was an important advance in reproductive research [122]. LHRH is a peptide hormone produced and secreted by the hypothalamus that stimulates the secretion of FSH and LH. A decapeptide with an amino acid sequence of pyroGlu-His-Trp-Ser-Tyr-Gly-Leu-Arg-Pro-Gly-NH_2, it is chemically and functionally similar in both males and females of all mammalian species studied thus far. More than 1000 analogues of LHRH have been synthesized by deletion or substitution of amino acids to introduce agonist or antagonist properties. A large number of reviews have appeared discussing their therapeutic importance [106–108].

These agonists and antagonists may provide useful therapy for several clinical conditions such as prostatic carcinoma, precocious puberty, and endometriosis. Scientists are also studying LHRH analogues for contraception [109]. Treatment with LHRH analogues blocks ovulation in women and spermatogenesis in men. However, treatment also results in a loss of estrogen and testosterone and causes other related side effects. Scientists are attempting to assess the use of these agents in combination with replacement estrogen, progesterone, and testosterone to determine if the side effects can be avoided. Several small nonhuman primate and clinical studies have suggested possible utility in this area, but large-scale clinical efficacy studies have not been conducted.

Progesterone Antagonists as Contraceptives. Progesterone is required to maintain pregnancy in women, and infertility results from failure of the corpus luteum to produce adequate amounts of progesterone. Inhibitors of progesterone synthesis, such as epostane [110], and inhibitors of progesterone-receptor binding, such as RU486, have been investigated for termination of pregnancy. Studies in the nonhuman primate and nonpregnant women indicate that progesterone antagonists may also have antifertility potential other than as an abortifacient [88], [111].

Inhibin and Activin. Inhibin, a water-soluble, gonadal factor known for over 50 years to inhibit pituitary function, has been isolated and identified [112–115]. Inhibin is a glycoprotein hormone that preferentially inhibits the secretion of FSH. It consists of an α-chain subunit, M_r 14000, linked by disulfide bonds to a β-chain subunit, M_r 18000. There are two forms of the β-chain subunit, β-A and β-B. The smaller subunit combines with either the β-A or β-B subunit to form inhibin-A or inhibin-B, respectively.

During the isolation of inhibin from follicular fluid, it was found that some chromatographic fractions stimulated FSH release from cultured anterior pituitary cells, suggesting the existence of FSH releasing proteins (FRPs). Two FRPs, given the generic term activins, were subsequently isolated [116], [117]. One is composed of two disulfide-linked β-A subunits (activin A); the other consists of similarly linked β-A and β-B subunits (activin AB).

Studies confirm that inhibin plays a role in regulating FSH secretion. However, the importance of this role in humans has not yet been determined. If inhibin-regulated FSH secretion is pivotal in follicular recruitment and growth, then it may be possible to block ovulation by means of inhibin antagonists.

Activin has the potential to serve the same therapeutic uses as GnRh analogues. This is because of its ability to regulate steroidogenesis directly at the gonadal level. Although the spectrum of functions of inhibin and activin are not completely understood at present, this peptide family has already demonstrated, by the nature of its differential subunit association, a powerful mechanism for the generation of dimers with opposing biologic actions. These characteristics of the inhibin peptide family warrant further study and evaluation as alternative approaches to fertility control.

Male Fertility Control. There is interest in male fertility control, both from a scientific as

well as from a sociological viewpoint. Many compounds have been identified as having male antifertility activity in various species, e.g., gossypol [*303-45-7*], ORF 5513, 5-thio-D-glucose [*20408-97-3*], and 6-chlorodeoxy-glucose [118]. A program based on the use of androgens has also been conducted [119].

Organic molecules thus far identified, such as those listed above, appear either to have irreversible antifertility effects, to be inherently toxic, or to affect libido. With respect to androgens, it has been demonstrated that sperm count can be depressed but not eliminated in men injected with large doses of androgens. However, questions about the potential utility of androgens as male antifertility agents are still debated, not least of which is whether lowering, but not eliminating sperm, will provide effective male contraception.

The ideal male contraceptive would produce azoospermia without compromising libido and sexual potency. While not totally fulfilling the criteria for a perfect male contraceptive, GnRH antagonists hold a greater potential than GnRH agonists. Unlike the agonists, GnRH antagonists inhibit gonodotropin secretion, decrease androgen levels, and induce azoospermia in male primates [120], [121]. Similar effects on hormone secretion have been reported in men [115].

In monkeys, testosterone replacement delays, but does not prevent, GnRH antagonist-induced azoospermia [122]. In men, the combination of testosterone and a GnRH antagonist results in a more complete gonadotropin and gonadal suppression than either agent alone [123]. These results suggest that GnRH antagonists, given in conjunction with androgens to maintain libido and sexual potency, have potential as male contraceptives. More recently the WHO and other research groups are actively pursuing the use of androgen plus progestogens as male contraceptives. Reviews of male contraceptive research are available [124], [125].

9. References

1. R. O. Greep, M. A. Koblinsky, F. S. Jaffe: *Reproduction and Human Welfare,* MIT Press Cambridge, Mass. 1976 p. 53.
2. C.- R. Garcia, D. L. Rosenfeld: *Fertility: The Regulation of Reproduction,* FA Davis Company, Philadelphia 1997, p. 59.
3. L. Mastroiani, Jr., P. J. Donaldson, T. T. Kane: *Developing New Contraceptives,* National Academic Press, Washington, D.C. 1990, p. 11.
4. R. A. Hatcher, J. Trussel, F. Steward (eds.): *Contraceptive Technology,* 17th ed., Ardent Media Inc., New York 1998.
5. R. L. Kleinman: *Periodic Abstinence for Family Planning,* IPPF, London, 1983.
6. H. Knaus, *Zentralbl. Gynäkol.* **53** (1929) 2193.
7. K. Ogino, *Zentralbl. Gynäkol.* **54** (1930) 464.
8. T. H. van de Velde: *Die Vollkommene Ehe, eine Studie über ihre Physiologie und Technik,* 21st ed., Benno Konegen, Medizinischer Verlag, Leipzig-Stuttgart 1928.
9. P. Miller, M. Soules, *Obstetrics and Gynecology,* **87** (1996) 1 p. 13.
10. J. Bonnar, in D. R. Mishell, Jr. (ed.): *Advances in Fertility Research,* **vol. 1,** Raven Press, New York 1982, p. 1.
11. K. L. Kennedy, R. Rivera, A. S. McNeilly, *Contraception,* **39** (1989) 477.
12. J. L. McGuire, F. C. Greenslade, in G. I. Zatuchni, A. J. Sobrero, J. J. Speidel, J. J. Sciarra (eds.): *Vaginal Contraception,* Harper and Row, New York 1980.
13. A. Kolliker: "Physiologische Studien über die Samenflüssigkeit," *Z. Wiss. Zool.* **7** (1855) 221.
14. J. R. Baker, *J. Hyg.* **32** (1932) 550.
15. F. V. Sander, S. D. Cramer, *Human Fertil.* **6** (1941) 134.
16. U.S. Patent No. 2 752 284, 1956 (V. R. Berliner, W. C. Mende, H. O. Singher).
17. S. Iwahara, K. Furuse, *Jpn. Public Health* **12** (1965) 123.
18. "Recommendations for the Development of Vaginal Microbicides, International Working Group on Vaginal Microbicides," *AIDS* **10** (1966) 8, 1 – 6.
19. D. W. Hahn, R. E. Homm, B. E. McKenzie, in G. I. Zatuchni, A. J. Sobrero, J. J. Speidel, J. J. Sciarra (eds.): *Vaginal Contraception,* Harper and Row, New York 1980, p. 232.
20. E. B. Connell: "Vaginal Contraception" in D. R. Mishell, Jr. (ed.): *Advances in Fertility Research,* Raven Press, New York 1982, p. 19.
21. The Contraceptive Diaphragm, Emory University, School of Medicine, Healthcare Communications Network, New York, 1989.
22. A. Melanie, Do Gold, *Pediatric Ann.* **24** (1995) 4, 211.

23. *Contraceptive Technology Update,* **11** (1990) 10, 145.
24. Condoms, Population Reports, Series H, no. 8, Population Information Program, The Johns Hopkins University, Baltimore, 1990.
25. R. A. Hatcher, et al.: *Contraceptive Technology,* Irvington Publishers, Inc., New York, 1990 – 1992, p. 227.
26. The American College of Obstetricians and Gynecologists (ACOG) Technical Bulletin No. 198, 1994.
27. L. Haberlandt, *Münch. Med. Wochenschr.* **68** (1921) 1577.
28. C. Djerassi, L. Miramontes, G. Rosenkranz, F. Sondheimer, *J. Am. Chem. Soc.* **76** (1954) 4092.
29. C. R. Kay, *J. Royal Col. Gen. Pract.* **30** (1980) 8.
30. W. H. Inman, M. P. Vessey, B. Westerhom, A. England, *Br. Med. J.* **2** (1970) 203.
31. Royal College of General Practitioners, Oral Contraceptives and Health: *An Interim Report from the Oral Contraception Study,* Pitman Medical, New York 1974.
32. M. P. Vessey et al., *J. Biosoc. Sci.* **8** (1976) 373.
33. A. Rosenfield, *Am. J. Obstet. Gynecol.* **132** (1978) 92.
34. H. W. Ory, A. Rosenfield, L. C. Landman, *Fam. Plann. Perspect.* **12** (1978) 278.
35. S. Ramcharan, et al.: *The Walnut Creek Contraceptive Drug Study: A Prospective Study of the Side Effects of Oral Contraceptives,* **Vol. 3** U.S. Dept. of Health and Human Services, Government Printing Office, Washington D.C. 1981.
36. M. Notelovitz, C. S. Kitchens, L. Coone, L. McKenzie, R. Carter, *Am. J. Obstet. Gynecol.,* **141** (1981) 71 – 75.
37. D. R. Mishell, Jr., *Am. J. Obstet. Gynecol.* **142** (1982) 809.
38. "Oral Contraceptives", in *Popul. Rep. [A]* **6** (May-June 1982).
39. "World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception," *Lancet* **346** (1995) 1575 – 1582.
40. "World Health Organization Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception," *Lancet* **346** (1995) 1582 – 1588.
41. K. W. M. Bloemenkamp et al., *Lancet* **346** (1995) 1593 – 1596.
42. H. Jick et al., *Lancet* **346** (1995) 1589 – 1593.
43. W. O. Spitzer et al., *BMJ* **312** (1996) 83 – 87.
44. M. A. Lewis et al., *BMJ* **312** (1996) 88 – 90.
45. S. A. Pasquale, *J. Reprod. Med.* **29 (Suppl.)** (1984) 560.
46. K. Fotherby, *Br. Family Planning* **11** (1985) 86.
47. G. Silfverstolpe, A. Gustafson, G. Samsoie, A. Svanborg, *Acta Obstet. Gynecol. Scand. (Suppl)* **88** (1979) 89.
48. B. R. Carr, *Int. J. Fertil.* **42** (1997) suppl. 1, 133.
49. T. Gordon et al., *Am. J. Med.* **62** (1977) 707.
50. I. F. Godsland, D. Crook, *Am. J. Obste. Gynec.* **170** (1994) 15 – 28.
51. A. Phillips, D. W. Hahn, S. Klimek, J. L. McGuire, *Contraception,* **36** (1987) 181.
52. A. Phillips, K. Demarest, D. W. Hahn, *Contraception,* **41** (1990) 399.
53. H. J. Kloosterboer, C. A. Von Knoordegraaf, E. W. Turpijn, *Contraception,* **38** (1988) 325.
54. D. W. Hahn, A. Phillips, J. L. McGuire, *Aktuelle Aspekte der Hormonalen Kontrazeption,* Karger, New York 1991, p. 46.
55. S. L. Corson, *Acta Obstet. Gynecol. Scand.* **69** (1990) (suppl. 152) 25 – 31.
56. R. S. London, et al., *Acta Obstet. Gynecol. Scand.* **71** (1992) (suppl. 156) 9 – 14.
57. J. A. Katzenellenbogen, B. W. O'Malley, B. S. Katzenellenbogen, *Molecular Endocrinology* **10** (1996) no. 2, 119.
58. G. J. M. Kuiper, E. Enmark, M. Pelto-Huikko, S. Nilsson, J. Gustafsson, *Proc. Nat'l. Acad. Sci. USA* **93** (1996) p. 5925.
59. G. J. M. Kuiper, et al., *Endocrinology,* **138** (1997) 863.
60. *"Oral Contraceptives-Update on Usage, Safety and Side Effect," Population Reports,* Series A, no. 5, Population Information Program, The Johns Hopkins University, Baltimore 1979.
61. *Oral Contraceptives in the 1980s. Population Reports,* Series A, no. 6, Population Information Program, The Johns Hopkins University, Baltimore 1982.
62. D. A. Grimes (ed.): *The Contraception Report,* **Vol. 7,** No. 1, Emron, Totowa, NJ 1996, p. 1.
63. J. S. Lippman, G. A. Shangold, *Int. J. Fertil.* **42** (1997) 4, 230.
64. D. R. Mishell, Jr. (ed.): "Long-Acting Steroid Contraception," *Advances in Human Fertility and Reproductive Endocrinology,* Raven Press New York, 1983.
65. "Hormonal Contraception: New Long-Acting Methods," Population Reports, Series K no. 3, Population Information Program, The Johns Hopkins University, Baltimore, 1987.

66. G. I. Zatuchni, A. Goldsmith, J. D. Shelton, J. J. Sciarra (eds.): *Long-Acting Contraceptive Delivery Systems,* Harper and Row, Philadelphia 1984.
67. Norplant: *A Summary of Scientific Data,* The Population Council, New York 1990.
68. *Contraceptive Technology Update* **12 (1)** (1991) 1.
69. F. Michael (ed.): *Safety Requirements for Contraceptive Steriods,* Cambridge University Press, Cambridge, 1980.
70. L. R. Beck, V. Z. Pope, *Drugs* **27** (1984) 528.
71. E. Diczfalusy, *Contraception* **33** (1986) 7.
72. *Contraceptive Technology Update* **12** (1991) 5, 79.
73. E. Diczfalusy, *Contraception* **43** (1991) 201.
74. *"Intrauterine Devices," Population Reports,* **Series B, no. 3** Population Information Program, The Johns Hopkins University, Baltimore 1982.
75. R. L. Kleinman: *Intrauterine Contraception,* 4th ed., International Planned Parenthood Federation, London 1977.
76. F. Alvares et al., *Fertil. Steril.* **49** (1988) 768.
77. I. Sivin, *Studies in Family Planning* **20** (1989) 355.
78. D. R. Mishell, Jr., in J. R. Newton (ed.): *Clinics in Obstetrics and Gynaecology,* **vol. 11,** W. B. Saunders, London 1984, p. 679.
79. H. J. Tatum, in D. R. Mishell, Jr. (ed.): *Advances in Fertility Research,* **vol. 1,** Raven Press, New York 1982, p. 47.
80. S. Iwahara, K. Furuse, *Jpn. Public Health* **12** (1965) 123.
81. D. A. Edelman, L. P. Cole, R. Apelo, P. Lavin, in G. I. Zatuchuni, A. Goldsmith, J. D. Shelton, J. J. Sciarra (eds.): *Long-Acting Contraceptive Delivery Systems,* Harper and Row, Philadelphia 1984, p. 621.
82. "Vasectomy: New Opportunities," Population Reports, Series D, no. 5, Population Information Program, The Johns Hopkins University, Baltimore, 1992.
83. "Voluntary Female Sterilization," Population Reports, Series C, no. 10, Population Information Program, The Johns Hopkins University, Baltimore, 1990.
84. J. R. Newton, in J. R. Newton (ed.): *Clinics in Obstetrics and Gynaecology,* **vol. 11,** W. B. Saunders, London 1984, p. 603.
85. *Outlook* **8** (1990) 3.
86. G. Zuliani, U. F. Colombo, R. Molla, *European J. OB/GYN and Reprod. Biol.* **37** (1990) 153.
87. Amer. College Obstet. Gynecol. (ACOG) Practice Patterns, No. 3, Dec. 1996.
88. M. Fasoli, F. Parazzini, G. Cecchetti, C. LaVacchia, *Contraception* **39** (1990) 459.
89. S. S. Lyer, *Drug News and Perspectives* **4** (1991) 40, 228.
90. V. P. Kamboj et al., *Ind. J. Exp. Biol.* **9** (1971) 103.
91. *Drugs of the Future* **2** (1977) 441.
92. C. Tietze and S. K. Henshaw: *Induced Abortion: A World Review,* The Alan Guttmacher Institute, New York – Washington, D.C. 1986, p. 29.
93. G. I. Zatuchni, M. H. Labbok, J. J. Sciarra (eds.): *Research Frontiers in Fertility Regulation,* Harper and Row, Hagerstown, MD 1980, p. 330.
94. *"Prostaglandins," Population Reports,* **Series E, no. 8,** Population Information Program, The Johns Hopkins University, Baltimore 1980.
95. E. E. Baulieu, S. J. Ssegal (eds.): *The Antiprogestin Steroid RU486 and Human Fertility Control,* Plenum Press, New York 1985.
96. O. M. Aurech et al., *Fert. Ster.* **56** (1991) 385.
97. G. D. Hodgen, *Fert. Ster.* **56** (1991) 394.
98. W. Elger et al., *J. Steroid. Biochem.* **25** (1986) 835.
99. V. C. Stevens, in G. Benagiano, E. Diczfalusy: *Endocrine Mechanisms in Fertility Regulation,* Raven Press, New York 1983, p. 141.
100. N. J. Alexander, D. Griffin, J. M. Spieler, G. M. H. Waites (eds.): *Gamete Interaction,* Wiley-Liss, New York 1990.
101. K. N. Sacco, O. Singh, R. Pal, G. P. Talwar: *Human Reproduction Update,* **Vol. 1,** no. 1, Oxford Press, 1985, p. 1.
102. W. R. Jones, *Bailliere's Clin. Obstet. Gynecol.* **10** (1996) 1, 69.
103. P. D. Griffin, *Human Reprod.* **6** (1991) 166.
104. J. C. Herr, *Amer. J. Reprod. Immunol.* **35** (1996) 184.
105. P. Primakoff et al., *Nature* **335** (1988) 543.
106. A. V. Schally, in F. Labrie, A. Belanger, A. Dupont (eds.): *LHRH and Its Analogues – Basic and Clinical Aspects,* Excerpta Medica, Amsterdam 1984, p. 3.
107. G. I. Zatuchni, J. D. Shelton, J. J. Sciarra (eds.): *LHRH Peptides as Female and Male Contraceptives,* Harper and Row, Philadelphia 1981.
108. B. H. Vickery, B. Luneneld: *GnRH Analogs in Cancer and Human Reproduction,* Kluwer Academic, Lancaster 1989.
109. M. C. Pike, J. R. Daniels, D. V. Spicer, *Endocr. Related Cancer* **4** (1997) 125.

110. H. P. Schane, J. E. Creange, in G. I. Zatuchni, M. Labbook, J. J. Sciarra (eds.): *Research Frontiers in Fertility Regulation,* Harper and Row, Hagerstown, MD 1980.
111. A. L. Goodman, G. D. Hodgen, in E. Y. Adashi, J. A. Rock, Z. Rosanulaks (eds.): *Reproductive Endocrinol, Surgery and Technology,* Lippincott Raven Publishers, Philadelphia 1996.
112. K. Miyamato et al., *Biochem. Biophys. Res. Comm.* **129** (1985) 396.
113. N. Ling et al., *Proc. Natl. Acad. Sci.* **82** (1985) 7217.
114. J. Rivier et al., *Biochem. Biophys. res. Comm.* **133** (1985) 120.
115. D. M. Robertson et al., *Biochem. Biophys. Res. Comm.* **126** (1985) 220.
116. N. Ling et al., *Nature* **321** (1986) 779.
117. W. Vale et al., *Nature* **321** (1986) 776.
118. A. Bartke, D. W. Hahn, R. G. Foldesy, J. L. McGuire in G. I. Zatuchni, A. Goldsmith, J. M. Spieler, J. J. Sciarra (eds.): *Male Contraception,* Harper and Row, Philadelphia 1985, p. 158.
119. C. A. Paulsen, W. J. Bremner, J. M. Leonard, in D. R. Mishell, Jr. (ed.): *Advances in Fertility Research,* **vol. 1,** Raven Press, New York 1982, p. 157.
120. F. B. Akhtar, G. F. Weinbauer, E. Nieschlag, *J. Endocrinol.* **104** (1984) 345.
121. G. F. Weinbauer et al., *Fertil. Steril.* **42** (1984) 906.
122. G. F. Weinbauer, F. J. Surmann, E. Nieschlag, *Acta Endocrinol.* **114** (1987) 138.
123. C. J. Bagatell et al., *J. Clin. Endocrinol. Metab.* **69** (1989) 43.
124. F. C. W. Wu, *Bailliere's Clin. Obstet. Gynecol.* **10** (1996) 1, 1.
125. M. J. Cosentino, S. A. Matlin, *Exp. Opin. Invest. Drugs* **6** (1997) x6, 635.

Controlled Release → Fertilizers
Controlled Release → Pharmaceuticals, General Survey and Development

Copper

ADALBERT LOSSIN, Norddeutsche Affinerie Aktiengesellschaft, Hamburg, Federal Republic of Germany

1. Introduction 2
2. Physical Properties 4
3. Chemical Properties 6
4. Occurrence 8
4.1. Copper Minerals 8
4.2. Origin of Copper Ores 9
4.3. Copper Ore Deposits 10
4.4. Copper Resources 10
4.5. Mining 10
5. Production 11
5.1. Beneficiation 12
5.2. Roasting 15
5.3. Pyrometallurgical Principles 17
5.3.1. Behavior of the Components 17
5.3.2. Matte 17
5.3.3. Slags 17
5.3.4. Oxidizing Smelting Processes 19
5.3.5. Proposals 19
5.4. Traditional Bath Smelting 21
5.4.1. Blast Furnace Smelting 21
5.4.2. Reverberatory Furnace Smelting . . . 22
5.4.3. Electric Furnace Smelting 22
5.4.4. Isasmelt Furnace 24
5.4.5. Noranda Process 24
5.4.6. CMT/Teniente Process 25
5.4.7. Vanyukov Process 25
5.4.8. Baiyin Process 26
5.5. Autogenous Smelting 26
5.5.1. Outokumpu Flash Smelting 27
5.5.2. Inco Flash Smelting 29
5.5.3. KIVCET Cyclone Smelting 30
5.5.4. Contop Matte Smelting 30
5.5.5. Flame Cyclone Smelting 31
5.6. Discontinuous Matte Conversion . . 32
5.7. Continuous Matte Conversion 35
5.7.1. Noranda Process 35
5.7.2. Mitsubishi Process 36
5.7.3. Kennecott/Outokumpu Flash Converting Process 38
5.8. Direct Blister Smelting 38
5.8.1. Blister Flash Smelting 38
5.8.2. QS Process 39
5.9. Copper Recycling 39
5.10. Hydrometallurgical Extraction . . . 40
6. Refining 45
6.1. Pyrometallurgical Refining 45
6.1.1. Discontinuous Fire Refining 45
6.1.2. Continuous Fire Refining 46
6.1.3. Casting of Anodes 46
6.2. Electrolytic Refining 47
6.2.1. Principles 47
6.2.2. Practice of Electrorefining 49
6.3. Melting and Casting 50
6.3.1. Remelting of Cathodes 51
6.3.2. Discontinuous Casting 51
6.3.3. Continuous Casting 51
6.3.4. Continuous Rod Casting and Rolling 51
6.4. Copper Powder 52
6.5. Copper Grades and Standardization 53
6.6. Quality Control and Analysis 54
7. Processing and Uses 55
7.1. Working Processes 55
7.2. Other Fabricating Methods 56
7.3. Uses 57
8. Economic Aspects 58
9. Environmental Protection 60
10. Toxicology 61
11. References 62

1. Introduction

Copper [*7440-50-8*], the red metal, apart from gold the only metallic element with a color different from a gray tone, has been known since the early days of the human race. It has always been one of the significant materials, and today it is the most frequently used heavy nonferrous metal. The utility of pure copper is based on its physical and chemical properties, above all, its electrical and thermal conductivity (exceeded only by silver), its outstanding ductility and thus excellent workability, and its corrosion resistance (a chemical behavior making it a half-noble metal).

Its common alloys, particularly brass and bronze, are of great practical importance (→ Copper Alloys). Copper compounds and ores are distinguished by bright coloration, especially reds, greens, and blues (→ Copper Compounds). Copper in soil is an essential trace element for most creatures, including humans.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_471

Etymology. According to mythology, the goddess Venus (or Aphrodite) was born on the Mediterranean island of Cyprus, formerly *Kypros* (Greek), where copper was exploited millennia before Christ. Therefore, in early times the Romans named it cyprium, later called cuprum. This name is the origin of *copper* and of the corresponding words in most Romance and Germanic languages, e.g., cobre (Spanish and Portuguese), cuivre (French), Kupfer (German), koper (Dutch), and koppar (Swedish).

History [21–24]. The first metals found by Neolithic man were gold and copper, later silver and meteoric iron. The earliest findings of copper are presumed to be nearly nine millennia old and came from the region near Konya in southern Anatolia (Turkey). Until recently the six-millennia-old copper implements from Iran (Tepe Sialk) were presumed to be the oldest. In the Old World, copper has been worked and used since approximately

7000 B.C.	Anatolia
4000 B.C.	Egypt, Mesopotamia, Palestine, Iran, and Turkestan
3000 B.C.	Aegean, India
2600 B.C.	Cyprus
2500 B.C.	Iberia, Transcaucasia, and China
2200 B.C.	Central Europe
2000 B.C.	British Isles
1500 B.C.	Scandinavia

Empirical experience over millennia has led to an astonishing knowledge of copper metallurgical operations:

1) Native copper was hardened by hammering (cold working) and softened by moderate heating (annealing).
2) Heating to higher temperatures (charcoal and bellows) produced molten copper and made possible the founding into forms of stone, clay, and later metal.
3) Similar treatment of the conspicuously colored oxidized copper ores formed copper metal.
4) The same treatment of sulfide copper ores (chalcopyrite), however, did not result in copper metal, but in copper matte (a sulfidic intermediate). Not before 2000 B.C. did people succeed in converting the matte into copper by repeated roasting and smelting.
5) In early times, bronze (copper – tin alloy) was won from complex ores, the Bronze Age beginning ca. 2800 B.C. At first, copper ores were smelted with tin ores; later, bronze was produced from metallic copper and tin. Brass (copper – zinc alloy) was known ca. 1000 B.C. and became widely used in the era of the Roman Empire.

In Roman times, most copper ore was mined in Spain (Rio Tinto) and Cyprus. With the fall of the Roman Empire, mining in Europe came to a virtual halt. In Germany (Saxony), mining activities were not resumed until 920 A.D. During the Middle Ages, mining and winning of metals expanded from Germany over the rest of Europe. In the middle of the 16th century, the current knowledge of metals was compiled in a detailed publication [23] by GEORGIUS AGRICOLA, *De Re Metallica* (1556).

Independent of the Old World, the Indians of North America had formed utensils by working native copper long before the time of Christ, although the skills of smelting and casting were unknown to them. On the other hand, the skill of copper casting was known in Peru ca. 500 A.D., and in the 15th century the Incas knew how to win the metal from sulfide ores.

Around 1500, Germany was the world leader in copper production, and the FUGGER family dominated world copper trade. By 1800, England had gained first place, processing ores from her own sources and foreign pits into metal. Near 1850, Chile became the most important producer of copper ores, and toward the end of the last century, the United States had taken the world lead in mining copper ores and in production of refined copper.

Technical development in the copper industry has made enormous progress in the last 120 years. The blast furnace, based on the oldest principle of copper production, was continually developed into more efficient units. Nevertheless, after World War I, it was increasingly replaced by the reverberatory furnace, first constructed in the United States. Since the end of World War II, this furnace has been superseded slowly by the flash smelting furnace invented in Finland. Recently, several even more modern methods, especially from Canada and Japan, have begun to compete with the older processes.

An important development in producing crude metal was the application of the Bessemer converter concept to copper metallurgy by

MANHÈS and DAVID (France, 1880): this principle is still the most widely used method for copper converting in the world.

Over time the requirements for copper purity have become increasingly stringent. The invention and development of electrolysis by J. B. ELKINGTON (England, 1865) and E. WOHLWILL (Germany, 1876) made refining of high-purity copper possible.

In addition, the quantity of copper produced has increased immensely (Table 1). Since 1800, ca. 375×10^6 t of primary copper has been mined in the world, but of this only ca. 10×10^6 t was mined between 1800 and 1900.

Table 1. World mine production of copper (approximate, from several sources)

Year	Production, 10^3 t	Year	Production, 10^3 t
1700	9	1970	6400
1800	17	1975	7300
1850	57	1980	7900
1900	450	1985	8300
1950	2500	1990	9225
1955	3100	1995	10 050
1960	4200	1997	11 525
1965	5000		

2. Physical Properties

Most properties of copper metal depend on the degree of purity and on the source of the metal. Variations in properties are caused by

1) Grade of copper, i.e., the oxygen content: tough-pitch copper, deoxidized copper, oxygen-free copper
2) Content of native impurities (e.g., arsenic, bismuth) or remnants of additives (e.g., phosphorus), which form solid solutions or separate phases at the grain boundaries
3) Thermal and mechanical pretreatment of the metal, which lead to states such as cast copper, hot-rolled copper, cold-worked (hard) copper, annealed (soft) copper, and sintered copper

These property differences are caused by the defects in the crystal lattice. Two groups of properties are to be distinguished:

1) Low dependence on crystal lattice defects, e.g., caloric and thermodynamic properties, magnetic behavior, and nuclear characteristics
2) High dependence on defects, e.g., electrical and thermal conductivity, plastic behavior, kinetic phenomena, and resistance to corrosion

The variations in properties are caused either by physical lattice imperfections (dislocations, lattice voids, and interstitial atoms) or by chemical imperfections (substitutional solid solutions).

Atomic and Nuclear Properties. The atomic number of copper is 29, and the atomic mass A_r is 63.546 ± 0.003 (IUPAC, 1983). Copper consists of two natural isotopes, ^{63}Cu (68.94 %) and ^{65}Cu (31.06 %). There are also nine synthetic radioactive isotopes with atomic masses between 58 and 68, of which ^{67}Cu has the longest half-life, ca. 58.5 h.

Crystal Structure. At moderate pressures, copper crystallizes from low temperatures up to its melting point in a cubic closest-packed (ccp) lattice, type A 1 (also F[1] or Cu) with the coordination number 12. X-ray structure analysis yields the following dimensions (at 20 °C):

Lattice constant	0.36152 nm
Minimum interatomic distance	0.2551 nm
Atomic radius	0.1276 nm
Atomic volume	7.114 cm^3/mol

There is also a high-pressure modification, which forms at ca. 400 MPa and 100 °C.

Density. The theoretical density at 20 °C, computed from lattice constant and atomic mass is 8.93 g/cm^3. The international standard was fixed at 8.89 g/cm^3 in 1913 by the IEC (International Electrotechnical Commission). The maximum value for 99.999 % copper reaches nearly 8.96 g/cm^3.

The density of commercial copper depends on its composition, especially the oxygen content, its mechanical and thermal pretreatment, and the temperature. At 20 °C, a wide range of values are found:

Cold-worked and annealed copper	8.89 – 8.93 g/cm^3
Cast tough-pitch electrolytic copper	8.30 – 8.70 g/cm^3
Cast oxygen-free electrolytic copper	8.85 – 8.93 g/cm^3

The values for cold-worked copper are higher than those of castings because the castings have pores and gas cavities.

The density of copper is nearly a linear function of temperature, with a discontinuity at the melting point:

Temperature, °C	Density, g/cm^3
solid copper	
20	8.93
600	8.68
900	8.47
1 083	8.32
liquid copper	
1 083	7.99
1 200	7.81

The solidification shrinkage is 4 %; the specific volume at 20 °C is 0.112 cm^3/g.

Mechanical Properties. Important mechanical values are given in Table 2. High-purity copper is an extremely ductile metal. Cold working increases the hardness and tensile strength (hard or hard-worked copper); subsequent annealing eliminates the hardening and strengthening so that the original soft state can be reproduced (soft copper). The working processes are based on this behavior (Section 7.1). Impurities that form solid solutions of the substitutional type likewise increase hardness and tensile strength.

Table 2. Mechanical properties of copper at room temperature

Property	Unit	Anealed (soft) copper	Cold-worked (hard) copper
Elastic modulus	GPa	100 – 120	120 – 130
Shearing modulus	GPa	40 – 45	45 – 50
Poisson's ratio		0.35	
Tensile strength	MPa	200 – 250	300 – 360
Yield strength	MPa	40 – 120	250 – 320
Elongation	%	30 – 40	3 – 5
Brinell hardness	(HB)	40 – 50	80 – 110
Vickers hardness	(HV)	45 – 55	90 – 120
Scratch hardness		≈3	

Pure copper has outstanding hot workability without hot brittleness, but the high-temperature strength is low. Detrimental impurities, those that decrease the strength at high temperatures, are principally lead, bismuth, antimony, selenium, tellurium, and sulfur. The concentration of oxides of such elements at the grain boundaries during heating causes the embrittlement. However, such an effect can be desirable when free cutting is required. At subzero temperatures, copper is a high-strength material without cold brittleness.

The changes in typical mechanical properties such as tensile strength, elongation, and hardness by heat treatment result from recrystallization [25]. The dependence of recrystallization temperature and grain size on the duration of heating, the amount of previous cold deformation, and the degree of purity of copper can be determined from diagrams. The recrystallization temperature is ca. 140 °C for high-purity copper and is 200 – 300 °C for common types of copper. A low recrystallization temperature is usually advantageous, but higher values are required to maintain strength and hardness if the metal is heated during use.

Thermal Properties. Important thermal values are compiled in Table 3. The thermal conductivity of copper is the highest of all metals except silver.

Table 3. Thermal properties of copper

Property	Unit	Value
Melting point	K	1356 (1083 °C)
Boiling point	K	2868 (2595 °C)
Heat of fusion	J/g	210
Heat of vaporization	J/g	4810
Vapor pressure (at *mp*)	Pa	0.073
Specific heat capacity		
at 293 K (20 °C)		
and 100 kPa (1 bar)	J g^{-1} K^{-1}	0.385
at 1230 K (957 °C)		
and 100 kPa		0.494
Average specific heat		
273 – 573 K (0 – 300 °C)		
at 100 kPa (1 bar)	J g^{-1} K^{-1}	0.411
273 – 1273 K (0 – 1000 °C)		
at 100 kPa		0.437
Coefficient of linear thermal expansion		
273 – 373 K (0 – 100 °C)	K^{-1}	16.9×10^{-6}
273 – 673 K (0 – 400 °C)		17.9×10^{-6}
between 273 and 1173 K (0 – 900 °C)		19.8×10^{-6}
Thermal conductivity		
at 293 K (20 °C)	W m^{-1} K^{-1}	394

Electrical Properties. In practice, the most important property of copper is its high electrical conductivity; among all metals only silver is a better conductor. Both electrical conductivity and thermal conductivity are connected with the Wiedemann – Franz relation and show strong dependence on temperature (Table 4). The old American standard, 100 %

IACS (International Annealed Copper Standard), corresponds to 58.0 MS/m at 20 °C, and it is still widely used in the United States. The corresponding electrical resistivity (ϱ) is $1.7241 \times 10^{-8}\,\Omega \cdot$ cm, and the less usual resistivity based on weight (density of 8.89 g/cm^3, IEC) is $0.1533\,\Omega\,\mathrm{g\,m^{-1}}$. The corresponding temperature coefficients are $0.0068 \times 10^{-8}\,\Omega\,\mathrm{m\,K^{-1}}$ ($\mathrm{d}\varrho/\mathrm{d}T$) and $0.00393\,\mathrm{K^{-1}}$ ($\varrho^{-1}\mathrm{d}\varrho/\mathrm{d}T$). The theoretical conductivity at 20 °C is nearly 60.0 MS/m or 103.4 % IACS, and today commercial oxygen-free copper (e.g., Cu-OF) has a conductivity of 101 % IACS.

Table 4. Temperature dependence of thermal and electrical conductivity of copper

Temperature		Thermal conductivity, $\mathrm{W\,m^{-1}\,K}$	Electrical conductivity, MS/m
K	°C		
17	−256	5 000	
73	−200	574	460
113	−160	450	
173	−100	435	110
273	0	398	60
293	20	394	58
373	100	385	44
473	200	381	34
573	300	377	27
973	700	338	15

The factors that increase the strength decrease electrical conductivity: cold working and elements that form solid solutions. Elements that form oxidic compounds that separate at grain boundaries affect electrical properties only slightly. Copper may lose up to ca. 3 % of its conductivity by cold working; however, subsequent annealing restores the original value. There is a simple rule: the harder the copper, the lower is its conductivity.

Other Properties. High-purity copper is diamagnetic with a mass susceptibility of -0.085×10^{-6} cm^3/g at room temperature. The dependence on temperature is small. However, a very low content of iron can strongly affect the magnetic properties of copper.

The lower the frequency of light, the higher the reflectivity of copper. The color of a clean, solid surface of high-purity copper is typically salmon red.

The surface tension of molten copper is 11.25×10^{-3} N/cm at 1150 °C, and the dynamic viscosity is 3.5×10^{-3} Pa · s at 1100 °C.

Detailed physical-property information and data are to be found in the literature, particularly as tabular compilations [25–30].

3. Chemical Properties

In the Periodic Table copper is placed in the first transition series (period 4). It belongs to Group 11 and, together with silver and gold, forms the coinage metals. Its electron configuration is [Ar] $3d^{10}\,4s^1$. Copper compounds are known in oxidation states ranging from +1 to +4, although the +2 (cupric) and the +1 (cuprous) are by far the most common. In aqueous solutions or below 800 °C, the +2 oxidation state is the most stable.

Copper(I) compounds such as CuCl and CuI are diamagnetic colorless materials, except for those whose color results from charge-transfer bands, for example, Cu_2O. Cu^+ ions, [Ar] $3d^{10}$, are coordinated in a linear (two ligands) or tetrahedral fashion (four ligands).

Copper(II) compounds such as $CuSO_4 \cdot 5\,H_2O$ are paramagnetic blue or green substances, the color of which results from strong absorption bands in the region between 600 and 900 nm caused by $d-d$ electron transfer processes. The Cu^{2+} ion is a d^9 system and generally sixfold coordinated in a distorted octahedral manner.

Copper(III) compounds are mostly diamagnetic. Cuprates like $NaCuO_2$ can be obtained by heating the oxides in pure oxygen. In chemistry only a few Cu^{3+} complexes are known, but it appears that Cu^{3+} plays an important role in biochemistry, especially with deprotonated peptides.

Copper(IV) compounds are not well known except for $Cs_2[CuF_6]$.

Behavior in Air. Copper in dry air at room temperature slowly develops a thin protective film of copper(I) oxide [*1317-39-1*]. On heating to a high temperature in the presence of oxygen, copper forms first copper(I) oxide, then copper(II) oxide [*1317-38-0*], both of which cover the metal as a loose scale.

In the atmosphere, the surface of copper oxidizes in the course of years to a mixture of green basic salts, the patina, which consists chiefly of the basic sulfate, with some basic carbonate. (In a marine atmosphere, there is also some basic chloride.) Such covering layers protect the metal.

Behavior versus Diverse Substances. While many substances scarcely react with copper under dry conditions, the rate of attack increases considerably in the presence of moisture. Copper has a high affinity for free halogens, molten sulfur or hydrogen sulfide.

Standard electron potentials of copper are as follows [31], [32]:

Potentials in standard (acid) solution:

$$Cu^{+} + e^{-} \longrightarrow Cu \quad E_0 = 0.521\ V$$

$$Cu^{2+} + 2e^{-} \longrightarrow Cu \quad E_0 = 0.153\ V$$

Potentials with complexing ligands:

$$[Cu(NH_3)_4]^{2+} + 2e^{-} \longrightarrow Cu + 4NH_3 \quad E_0 = -0.11\ V$$

$$[Cu(CN)_2]^{-} + e^{-} \longrightarrow Cu + 2CN^{-} \quad E_0 = -0.43\ V$$

As the standard electron potentials show, copper metal is stable to nonoxidizing acids like dilute sulfuric or hydrochloric acid, similar to the precious metals. Dissolution of copper is possible in oxidizing acids such as nitric acid or hot concentrated sulfuric acid. Also other redox systems such as iron(III) or copper(II) chloride solutions are suitable reagents for leaching copper in practice.

Copper dissolves not only in oxidizing acids but also, for example, in ammonia or cyanide solutions in the presence of oxygen because stable complexes are formed. Also acetic acid together with oxygen or hydrogen peroxide attacks copper forming a green pigment called verdigris.

Free Cu^{+} ions are not stable in aqueous solution although Cu^{+} ($3d^{10}$) has a filled d shell. Spontaneous disproportion into Cu^{2+} and Cu takes place.

$$2Cu^{+} \longrightarrow Cu^{2+} + Cu \quad E_0 = 0.37\ V$$

$$K = [Cu^{2+}]/[Cu^{+}] = 10^6$$

The distorted octahedral coordination of six water molecules around the Cu^{2+} ion (d^9) gives an additional stabilization energy (ligand-field effect). In aqueous solutions, Cu^{+} is only existent in form of very stable complexes like $[Cu(CN)_2]^{-}$ or in the presence of an excess of copper metal. Also, insoluble Cu^{+} compounds such as cuprous oxide do not disproportionate in water.

By virtue of its large ionic radius and low electrical charge, the Cu^{+} ion is a soft acid. Therefore, the chemistry of copper in the oxidation state + 1 is predominated by reactions with soft bases like iodine (CuI), sulfur (CuSCN), or unsaturated nitrogen ligands. In contrast, the chemistry of Cu^{2+}, which is smaller and more highly charged, is dominated by hard ligands like oxygen ($[Cu(H_2O)_6]^{2+}$) or nitrogen ($[Cu(NH_3)_4]^{2+}$).

Copper is very stable in fresh water and also in sea water or alkali metal hydroxide solutions. Wastewater containing organic sulfur compounds can be corrosive to copper.

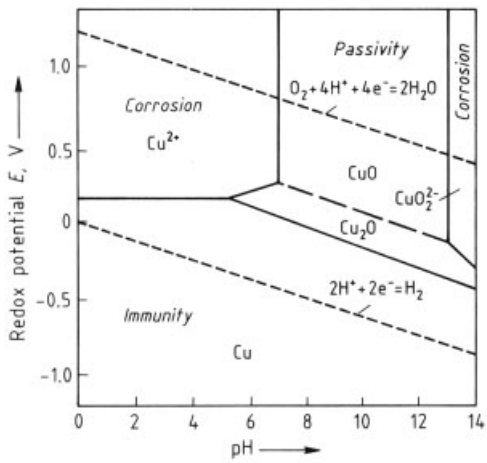

Figure 1. Pourbaix diagram for copper in highly dilute aqueous solution at normal temperature [35]

Corrosion [33], [34]. M. J. N. Pourbaix has developed potential – pH equilibrium diagrams for metals in dilute aqueous solutions [35]. Such graphs give a rough indication of the feasibility of electrochemical reactions. Figure 1 shows the behavior of copper at room temperature and atmospheric pressure. The $Cu - H_2O$ system contains three fields of different character:

1) Corrosion, in which the metal is attacked
2) Immunity, in which reaction is thermodynamically impossible
3) Passivity, in which there is no reaction because of kinetic phenomena

Gases and Copper [36], [37]. An exact knowledge of the behavior of solid and liquid copper toward gases is important for production and use of the metal. With the exception of hydrogen, [*1333-74-0*], the solubility of gases in molten copper follows Henry's law: the solubility is proportional to the partial pressure.

Oxygen [*7782-44-7*] dissolves in molten copper as copper(I) oxide up to a concentration of 12.6 wt % Cu_2O (corresponding to 1.4 wt % O) (also see Fig. 32). Copper(I) oxide in solid copper forms a separate solid phase.

Sulfur dioxide [*7446-09-5*] dissolves in molten copper and reacts:

$$6\,Cu + SO_2 \rightleftharpoons Cu_2S + 2\,Cu_2O$$

Hydrogen is considerably soluble in liquid copper, and after solidification some remains dissolved in the solid metal, although copper does not form a hydride. The solubility follows Sievert's law, being proportional to the square root of the partial pressure because the H_2 molecules dissociate into H atoms on dissolution. Hydrogen has high diffusibility because of its extremely small atomic volume.

Hydrogen dissolved in oxygen-bearing copper reacts with copper(I) oxide at high temperatures to form steam:

$$Cu_2O + 2\,H \rightleftarrows 2\,Cu + H_2O(g)$$

Steam is not soluble in copper; therefore, it either escapes or forms micropores.

Nitrogen, carbon monoxide, and carbon dioxide are practically insoluble in liquid or solid copper. Hydrocarbons generally do not react with copper. An exception is acetylene, which reacts at room temperature to form the highly explosive copper acetylides Cu_2C_2 and CuC_2; therefore, acetylene gas cylinders must not be equipped with copper fittings.

4. Occurrence

In the upper part of the earth's crust (16 km deep), the average copper content is ca. 50 ppm. It is thus about half as abundant as chromium, about twice as abundant as cobalt, and 26th in order of abundance of the elements in the accessible sphere of the earth. Table 5 shows average copper contents in natural materials.

Table 5. Typical copper contents of natural materials

Mineral	Content, ppm
Basalt	85
Diorite	30
Granite	10
Sandstone	1
Copper ores (poor)	5 000
Copper ores (rich)	50 000
Native copper	950 000
Seawater	0.003
Deep-sea clays	200
Manganese nodules	10 000
Marine ore sludges	10 000
Earth's crust (average)	50
Meteorites (average)	180

4.1. Copper Minerals

More than 200 minerals contain copper in definable amounts, but only about 20 are of importance as copper ores (Table 6) or as semiprecious stones (turquoise and malachite). Copper is a typical chalcophilic element; therefore, its principal minerals are sulfides, mostly chalcopyrite, bornite, and chalcocite, often accompanied by pyrite, galena, or sphalerite.

Secondary minerals are formed in sulfide ore bodies near the earth's surface in two stages. In the oxidation zone, oxygen-containing water forms copper oxides, basic salts (basic carbonates and basic sulfates), and silicates. In the deeper cementation zone, copper-bearing solutions from these salts are transformed into secondary copper sulfides (chalcocite and covellite) and even native copper of often high purity, e.g., in the Michigan copper district (Keweenaw Peninsula).

Other metallic elements frequently found in copper ores are iron, lead, zinc, antimony, and arsenic; less common are selenium, tellurium, bismuth, silver, and gold. Substantial enrichments sometimes occur in complex ores. For example, ores from Sudbury, Ontario, in Canada contain nickel and copper in nearly the same concentrations, as well as considerable amounts of platinum metals. The copper ores from Zaire and

Table 6. The most important copper minerals

Mineral	Formula	Copper, wt %	Crystal system	Density, g/cm^3
Native copper	Cu	≦99.92	cubic	8.9
Chalcocite	Cu_2S	79.9	orthorhombic	5.5 – 5.8
Digenite	Cu_9S_5	78.0	cubic	5.6
Covellite	CuS	66.5	hexagonal	4.7
Chalcopyrite	$CuFeS_2$	34.6	tetragonal	4.1 – 4.3
Bornite	Cu_5FeS_4/Cu_3FeS_3	55.5 – 69.7	tetragonal	4.9 – 5.3
Tennantite	$Cu_{12}As_4S_{13}$	42 – 52	cubic	4.4 – 4.8
Tetraedrite	$Cu_{12}Sb_4S_{13}$	30 – 45	cubic	4.6 – 5.1
Enargite	Cu_3AsS_4	48.4	orthorhombic	4.4 – 4.5
Bournonite	$CuPbSbS_3$	13.0	orthorhombic	5.7 – 5.9
Cuprite	Cu_2O	88.8	cubic	6.15
Tenorite	CuO	79.9	monoclinic	6.4
Malachite	$CuCO_3 \cdot Cu(OH)_2$	57.5	monoclinic	4.0
Azurite	$2\,CuCO_3 \cdot Cu(OH)_2$	55.3	monoclinic	3.8
Chrysocolla	$CuSiO_3 \cdot n\,H_2O$	30 – 36	(amorphous)	1.9 – 2.3
Dioptase	$Cu_6\,[Si_6O_{18}] \cdot 6\,H_2O$	40.3	rhombohedral	3.3
Brochantite	$CuSO_4 \cdot 3\,Cu(OH)_2$	56.2	monoclinic	4.0
Antlerite	$CuSO_4 \cdot 2\,Cu(OH)_2$	53.8	orthorhombic	3.9
Chalcanthite	$CuSO_4 \cdot 5\,H_2O$	25.5	triclinic	2.2 – 2.3
Atacamite	$CuCl_2 \cdot 3\,Cu(OH)_2$	59.5	orthorhombic	3.75

Zambia are useful sources of cobalt. Many porphyry copper ores in America contain significant amounts of molybdenum and are the most important single source of rhenium. The extraction of precious metals and other rare elements can be decisive for the profitability of copper mines, smelters, and refineries.

4.2. Origin of Copper Ores

Ore deposits are classified according to their mode of formation, but the origin of copper ores is geologically difficult to unravel, and some of the proposed origins are controversial. The classification distinguishes two main groups, the magmatic series and the sedimentary series.

Magmatic ore formation involves magma crystallization and comprises the following groups:

1) Liquid magmatic ore deposits originate by segregation of the molten mass so that the heavier sulfides (corresponding to matte) separate from the silicates (corresponding to slag) and form intrusive ore bodies. Examples: Sudbury, Ontario; Norilsk, western Siberia.
2) Pegmatitic – pneumatolytic ore deposits develop during the cooling of magma to ca. 374 °C, the critical temperature of water. Examples: Bisbee, Arizona; Cananea, Mexico.
3) Hydrothermal ore deposits result by further cooling of the hot, dilute metal-bearing solutions from ca. 350 °C downward, i.e., below the critical temperature of water. Such deposits contain copper primarily as chalcopyrite and satisfy ca. 50 % of the demand in the Western world. There are many examples of different types of hydrothermal deposits. Examples: Butte, Montana (gangue deposit); Tsumeb, Namibia (metasomatic deposit); Bingham Canyon, Utah; Chuquicamata, Chile; Toquepala, Peru; Bougainville, Solomon Islands (impregnation deposits). Impregnation deposits are also called disseminated copper ores or porphyry copper ores (or simply porphyries) because of their fine particle size.
4) Exhalative sedimentary ore deposits originate from submarine volcanic exhalations and thermal springs that enter into seawater, and constitute a transitional type to sedimentary deposits. These ores are third in economic importance in the Western world. The actual formation of such sulfidic precipitations can be observed, for example, the marine ore slimes in the Red Sea. Examples: Mount Isa, Queensland; Rio Tinto, Spain; Rammelsberg (Harz), Federal Republic of Germany.

The origin of sedimentary ore occurs in the exogenous cycle of rocks and may be subdivided into the following groups:

1) Arid sediments in sandstones and conglomerates occur widely in the former Soviet Union as widespread continental zones of weathering with uneven mineralization. Examples: Dsheskasgan, Kazakhstan; Exótica, Chile.
2) Partly metamorphized sedimentary ores in shales, marls, and dolomites form large strata-bound ore deposits, especially in the African copper belt, and represent the second most important source of copper to the Western world, as well as supplying nearly 75 % of its cobalt. Examples: Zaire (oxidation zone, oxidized ores $\leqq$6 % Cu); Zambia (cementation zone, secondary sulfide ores $\leqq$4 % Cu).
3) Marine precipitates have formed sedimentary ore deposits similar to the present phenomenon of sulfide precipitation by sulfur bacteria in the depths of the Black Sea. Examples: Silesia (copper marl), Poland.
4) Deep-sea concretions lie in abundance on the bottom of the oceans, especially the Pacific Ocean. These so-called *manganese nodules* could also be a source of copper in the future.

4.3. Copper Ore Deposits

Geologically, the main regions of copper ore deposits are found in two formations: the Precambrian shields and the Tertiary fold mountains and archipelagos. There are major producing countries on every continent [38], [39].

North America: United States (Arizona, Utah, New Mexico, Montana, Nevada, and Michigan), Canada (Ontario, Quebec, British Columbia, and Manitoba), and Mexico (Sonora)

South America: Chile, Peru, Argentina, and Brazil

Africa: Zaire, Zambia, Zimbabwe, South Africa, and Namibia

Australia and Oceania: Queensland, Papua New Guinea

Asia: Former Soviet Union (Siberia, Kazakhstan, and Uzbekistan), Japan, Philippines, Indonesia, India, Iran, and Turkey

Europe: Poland (Silesia), Yugoslavia, Portugal, Bulgaria, Sweden, and Finland

Antarctica may be an important source of copper ores in the foreseeable future.

4.4. Copper Resources

World primary copper reserves were estimated in 1991 at 552×10^6 t (Table 7) [41]. Reserves are identified resources and do not include undiscovered resources. With time the available reserves have increased because of technical progress in processing ores with low copper content and the discovery of new ore deposits [38], [39]. About 321×10^6 t were classified as minable copper ores under the technical and economical conditions at that time.

It is believed that a large potential of as-yet untouched deposits exist. Therefore, the potentially usable copper resources are estimated to be about three times as large as the reserve base. In addition to ores on land, there is an estimated amount of copper in deep-sea nodules of about 0.7×10^9 t.

Table 7. Copper ore reserves in 1991 [41]

Country	Ore reserves, 10^6 t	Percentage of world reserves
United States	90	16.3
Chile	120	21.7
Peru	31	5.6
Zambia	30	5.4
Zaire	30	6.3
Canada	23	4.2
Australia	21	3.8
Philippines	16	2.9
Indonesia	8	1.4
China	8	1.4
Poland	15	2.7
CIS	54	9.8
Other countries	106	19.2

If one assumes that total production of primary copper will remain stable, the identified reserves would last until 2040. With an increase in copper production of 2 – 3 %, which is more realistic, the duration of the known reserves will be reduced. However, these forecasts are quite unreliable because the growing use of secondary copper (recycling materials) and the discovery of new copper ore deposits are not considered.

4.5. Mining

Exploration, which is the search for ore deposits and their detailed investigation, is required to

ascertain the commercial feasibility of a potential mine. The used geological, geochemical, or geophysical methods are complicated and expensive. But often legal and political factors are more decisive for a new mine project than technological or financial aspects. The methods used for exhausting the copper ore are depending to a great extent to the type of deposit. Important parameters are metal content of the ore and geometry and depth of the ore body. These parameters determine the working method. Underground and open-cast mining are the two basic techniques. As a generalization mining can be divided into the stages drilling, blasting, loading, and haulage of the ore.

Technological developments like the LHD technique (load, haul, dump) and the use of concrete for stowing have made mining more cost efficient. But underground mining is still more labor-intensive and expensive than open cast mining. Therefore, copper ores with an average Cu content of 1 % or more or that contain other valuable metals in addition to copper (e.g., precious metals, nickel, cobalt) are mined underground, while ores with 0.5 % Cu, which represents a 100-fold enrichment of the average copper content in the earth's crust, are mined by open-cast methods. The porphyry copper deposits which are located near the surface can only be exploited by open-cast technique. They have become more important during the last decades. The first open pit was started at Bingham Canyon early this century. Today the terraced copper open pits are the largest ore mines in the world. They often cover more than one square kilometer and have working depths of several hundred meters. 100 000 t of crude ore are extracted per day. Today more than 50 % of the primary copper comes from open pits. Other less common methods for copper extraction are in situ leaching or ocean mining.

In situ leaching is a hydrometallurgical process in which copper is extracted by chemical dissolution in sulfuric acid. This method is suitable for low-grade copper ore bodies for which customary mining operations would be uneconomical, as well as for the leaching of remenant ores from abandoned mines. In some cases, the ore body must be broken before leaching by blasting with explosives to increase the surface area for chemical reaction.

Ocean mining involves obtaining metalliferous raw materials from the deep oceanic zones. Two groups of substances are of interest: deep-sea nodules [43] and marine ore slimes [44]. The nodules (manganese nodules; see → Manganese and Manganese Alloys, Chap. 9.) contain, in addition to iron oxides, ca. 25 % Mn, 1 % Ni, 0.35 % Co and 0.5 % (max. 1.4 %) Cu. Specially equipped ships have collected and lifted these nodules from depths of 3000 – 5000 m; specific metallurgical and chemical methods for processing the nodules have been developed in pilot plants. Because of the extremely high expenses, large-scale operations of this type have not yet been undertaken. Marine ore slimes from the Red Sea (2200-m depth) average ca. 4 % Zn, 1 % Cu, and a little silver. Although methods for processing these slimes have been investigated, this resource is not now economically important.

5. Production

Over the years copper production methods have been subjected to a continual selection and improvement process because of the need for (1) increased productivity through rationalization, (2) lower energy consumption, (3) increased environmental protection, (4) increased reliability of operation, and (5) improved safety in operation. During this development a number of tendencies have become apparent:

1) Decrease in the number of process steps
2) Preference for continuous processes over batch processes
3) Autogenous operation
4) Use of oxygen or oxygen-enriched air
5) Tendency toward electrometallurgical methods
6) Increased energy concentration per unit of volume and time
7) Electronic automation, measurement, and control
8) Recovery of sulfur for sale or disposal
9) Recovery of valuable byproducts

The selection of a particular production method depends essentially on the type of available raw materials, which is usually ore or concentrate and on the conditions at the plant location.

About 80 % of primary copper production comes from low-grade or poor sulfide ores. After enrichment steps, the copper concentrates are usually treated by pyrometallurgical methods. Generally, copper extraction follows the sequence:

1) Beneficiation by froth flotation of ore to give copper concentrate
2) Optional partial roasting to obtain oxidized material or calcines
3) Two-stage pyrometallurgical extraction
   a) smelting concentrates to matte
   b) converting matte by oxidation to crude (converter or blister) copper
4) Refining the crude copper, usually in two steps
   a) pyrometallurgically to fire-refined copper
   b) electrolytically to high-purity electrolytic copper

Figure 2 [13] illustrates the principal processes for extracting copper from sulfide ore.

About 15 %, with an increasing trend, of the primary copper originates from low-grade oxidized (oxide) or mixed (oxidized and sulfidic) ores. Such materials are generally treated by hydrometallurgical methods.

The very few high-grade or rich copper ores still available can be processed by traditional smelting in a shaft furnace. This process is also used for recovering copper from secondary materials, such as intermediate products, scrap, and wastes.

Figure 3 [13] illustrates the most important operations in copper extraction from various oxidic ores.

5.1. Beneficiation

Sulfidic copper ores are too dilute for direct smelting. Smelting these materials would require too much energy and very large furnace capacities. The copper ore coming from the mine (0.5 – 1 % Cu) must be concentrated by beneficiation. The valuable minerals like chalcopyrite are intergrown with gangue. Therefore, in the first step the lumpy ore is crushed and milled into fine particles ($< 100\,\mu m$) to liberate the individual mineral phases.

Typical equipment for crushing to about 20 cm are gyratory and cone crushers. Then wet grinding in semi-autogenous rod or autogenous ball mills takes place. Size classification takes is performed in cyclones. In the next step of beneficiation, valuable minerals and gangue are separated by froth flotation of the ore pulp, which exploits the different surface properties of the sulfidic copper ore and the gangue [46]. The hydrophobic sulfide particles become attached to the air bubbles, which are stirred into the pulp, rise with them to the surface of the pulp, and are skimmed off as a froth of fine concentrate. The hydrophilic gangue minerals remain in the pulp. Organic reagents with sulfur-containing groups at their polar end, such as xanthates, are used as collectors in the flotation process. Additionally, modifiers like hydroxyl ions (pH adjustment) are used to select different sulfide minerals, for example, chalcopyrite and pyrite. Alcohols are used to stabilize the froth.

To obtain concentrates with highest possible purity and recovery rate, the flotation process usually consists of several stages which are controlled by expert systems. Various sensors for particle size, pH, density, and other properties are installed. Figure 4 gives an overview of a typical beneficiation process at a concentrator. In the first flotation stage, as much copper as possible is recovered in a rougher concentrate so that as little as possible goes to the tailings. To increase the copper recovery rate, often these tailings are leached with sulfuric acid. After regrinding, the rough concentrate is cleaned in several flotation steps. After sedimentation in thickeners and filtration in automatic filter presses or vacuum filters (ceramic disk) the typical copper concentrate contains 25 – 35 % Cu and about 8 % moisture. The moisture content of the concentrate is a compromise between transporting water (cost) and avoiding dust generation during transport. Dewatered concentrates may heat spontaneously or even catch fire; therefore, appropriate precautions must be taken [47].

Copper concentrators typically treat up to 100 000 t of ore per day. They are located directly at the mines to achieve low transport costs. The copper recovery efficiency is over 90 %. About 95 % of the ore input goes into the tailings, which are stored in large dams near the mine and are used for water recycling to the flotation stages.

Separation of special copper ores such as those containing molybdenite or with high zinc

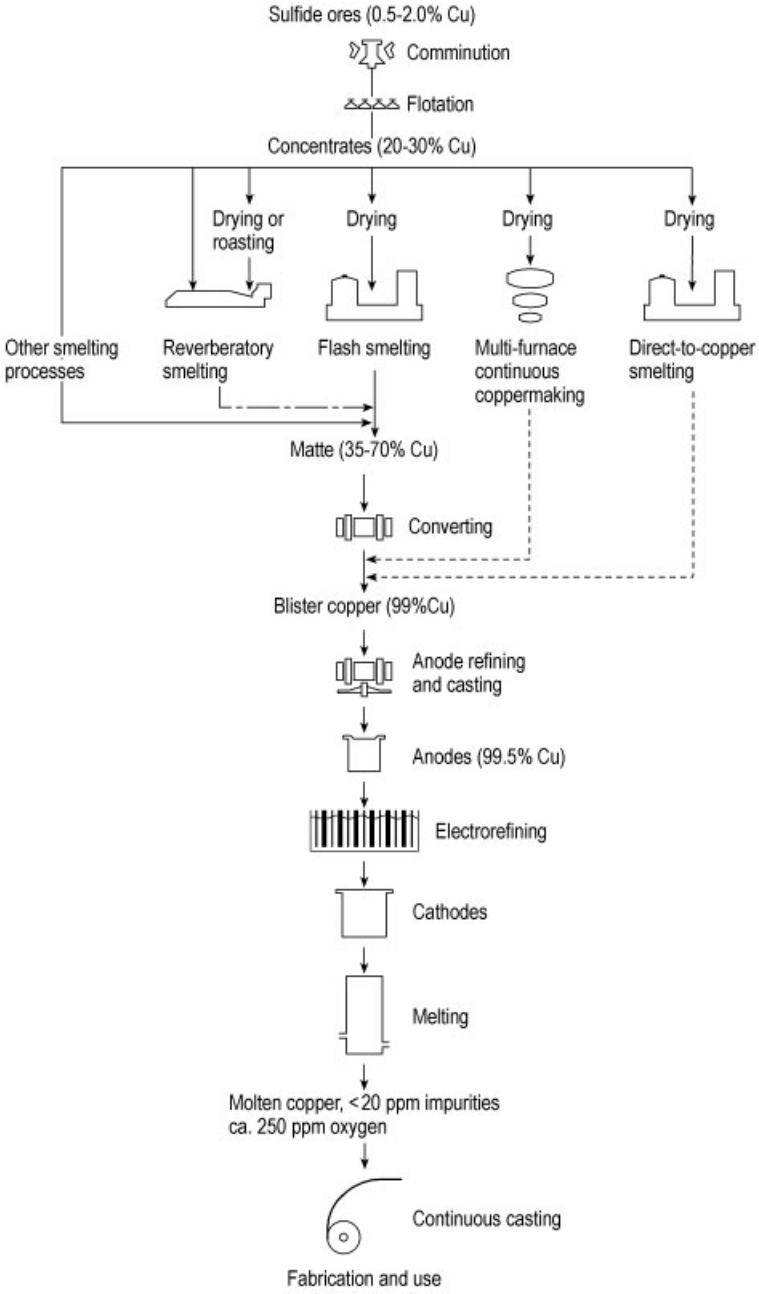

Figure 2. Principal processes for extracting copper from sulfide ore

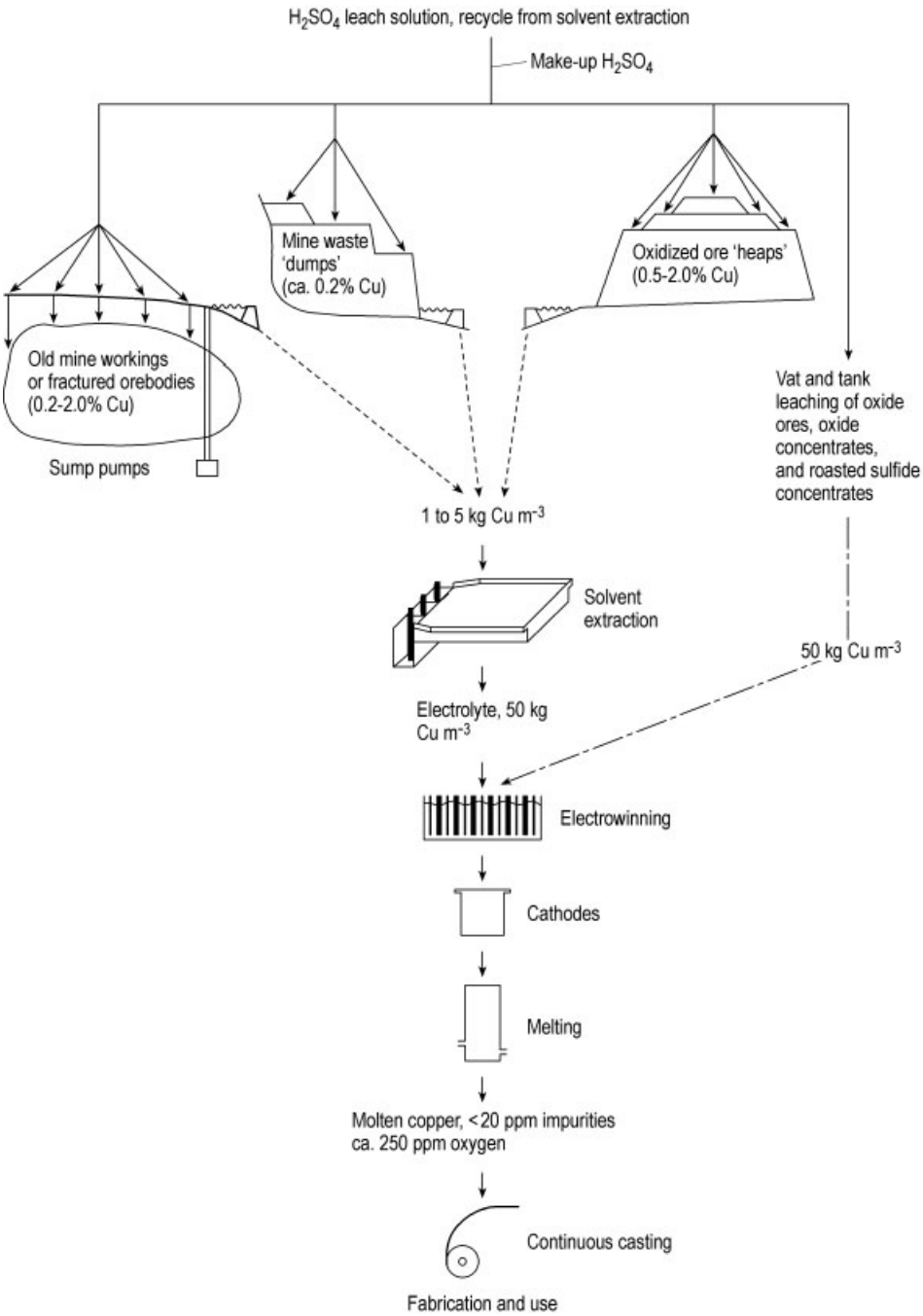

Figure 3. Principal processes for extracting copper from oxidic ore [45]

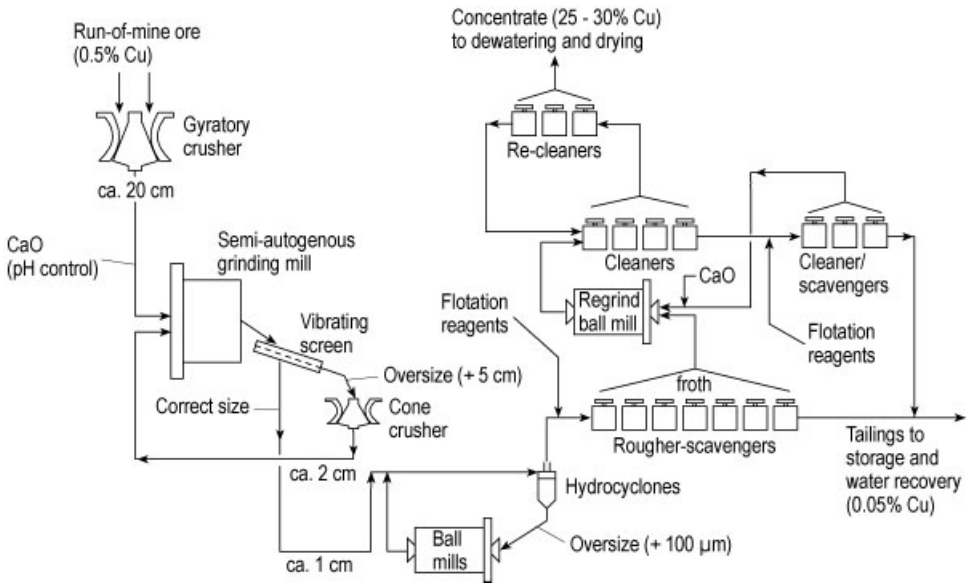

Figure 4. Overview of a typical beneficiation process at a concentrator

or lead content (Canada) is also possible by flotation methods. Flotation of non-sulfide copper minerals is rare because these ores are mostly subjected to hydrometallurgical copper recovery, for example, heap leaching. In Zambia and Zaire, however, siliceous copper oxide ores are floated with fatty acid collectors, and dolomitic copper oxide ores are sulfidized with sodium hydrogensulfide and then floated [48].

5.2. Roasting

Roasting can be used to prepare sulfide concentrates for subsequent pyrometallurgical or hydrometallurgical process. Partial roasting under oxidizing conditions may be carried out prior to smelting in reverberatory or electric furnaces. Complete oxidizing or sulfatizing roasting may be performed before leaching operations, especially if other valuable metals such as cobalt are present in the concentrate. Reducing roasting may be carried out if copper concentrates with very high contents of impurities such as As are to be smelted.

However, roasting processes are today not very important for the copper extraction process. Only a few plants are still operating, for example Boliden's Rönnskär Smelter [211] and Bor Smelter [212]. Since ca. 1975 combined roasting and matte smelting processes such as flash smelting have been favored because of their lower energy consumption and process gas handling advantages.

The roasting process has several effects:

1) Drying the concentrates
2) Oxidizing a part of the iron present
3) Controlling the sulfur content
4) Partially removing volatile impurities, especially arsenic
5) Preheating the calcined feed with added fluxes, chiefly silica and limestone

Chemical Reactions. When the moist concentrates, which contain many impurity elements, are heated, a multitude of chemical reactions occur. Because analysis of the many thermodynamic equilibria is not practical, a few fundamental systems are usually chosen. The most important is the ternary copper – oxygen – sulfur system (Fig. 5). The next most important system is the ternary iron – oxygen – sulfur system because most sulfidic copper ores contain significant amounts of iron.

Initially, sulfides such as pyrite and chalcopyrite decompose and generate sulfur vapor, which reacts with oxygen to form sulfur dioxide:

$$FeS_2 \longrightarrow FeS + S(g)$$

$$2\,CuFeS_2 \longrightarrow Cu_2S + 2\,FeS + S(g)$$

$$S(g) + O_2(g) \longrightarrow SO_2(g)$$

The principal reactions, i.e., the formation of metal oxides, sulfur trioxide, and metal sulfates, are exothermic.

$$MS + 1.5\,O_2 \rightleftharpoons MO + SO_2$$

$$SO_2 + 0.5\,O_2 \rightleftharpoons SO_3$$

$$MO + SO_3 \rightleftharpoons MSO_4$$

In addition, there are secondary reactions, such as the formation of basic sulfates, ferrites (especially magnetite), and silicates, the last providing most of the slag in the subsequent smelting:

$$MO + MSO_4 \longrightarrow MO \cdot MSO_4$$

$$MO + Fe_2O_3 \longrightarrow MFe_2O_4$$

$$FeO + Fe_2O_3 \longrightarrow Fe_3O_4$$

$$MO + SiO_2 \longrightarrow MSiO_3$$

Representative reductive roasting reactions are:

$$FeS_2 \longrightarrow FeS + S(g)$$

$$8\,FeAsS \longrightarrow 4\,FeAs + 4\,FeS + As_4S_4(g)$$

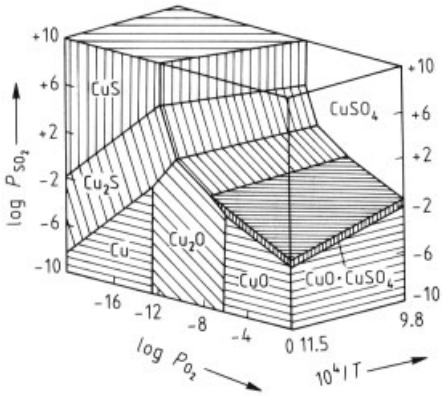

Figure 5. Partial phase diagram of the ternary Cu – O – S system [51]

Methods. There are several important roasting methods; all involve oxidation at an elevated temperature, generally between 500 and 750 °C:

1) *Partial (oxidizing) roasting* is the conventional way of extracting copper from sulfide concentrates. At 700 – 750 °C, the degree of roasting is determined by controlling the access of air. A predetermined amount of sulfur (30 – 50 % is removed, and only part of the iron sulfide is oxidized. The copper sulfide is relatively unchanged. These conditions are important for the formation of a suitable matte.
2) *Total, or dead, roasting* is occasionally used for complete oxidation of all sulfides for a subsequent reduction process or for special hydrometallurgical operations.
3) *Sulfatizing roasting* is carried out at 550 – 650 °C to form sulfates. This method yields calcines well-suited for hydrometallurgical treatment.

Roasters. Industrial roasting is done in two types of roasters: fluidized-bed and multiple-hearth roasters. Both are continuously operated processes.

Oxidizing roasting is usually carried out in fluidized-bed roasters with short residence times in the range of seconds and high production rates up to 50 t of moist concentrate per hour. The oxidation reactions supply most of the required heat. About 30 – 50 % of the incoming sulfide is oxidized to SO_2 by using slightly oxygen enriched air (up to 30 % O_2). The off-gas is rich in SO_2(6 – 12 %) and suitable for conversion to sulfuric acid. The hot calcine is usually sent to reverberatory or electric furnace. The advantage of a roaster in front of a smelting furnace is the lower energy requirement of the smelting furnace and the higher matte grade. Examples of fluidized-bed roasters are Boliden Rönnskär (Sweden), Bor Smelter (Yugoslavia) in front of pyrometallurgical copper extraction, and Chambishi (Zambia) ahead of a leaching plant.

Reductive roasting is usually carried out in multiple-hearth furnaces because of the long residence time (several hours) and the precise control of temperature and gas composition on each hearth. These roasters have lower production rates and are fired by natural gas burners. An example for the reducing process is the treatment

of El Indio (Chile) concentrate, which contains 8 % As and about 22 % Cu. About 97 % of the arsenic is removed during the roasting process at about 650 – 720 °C [213]. In former times the multiple-hearth roasters (Herreshoff furnaces) were widely used for the extraction of copper from concentrates.

5.3. Pyrometallurgical Principles

Smelting of unroasted or partially roasted sulfide ore concentrates produces two immiscible molten phases: a heavier sulfide phase containing most of the copper, the *matte*, and an oxide phase, the *slag*. In most copper extraction processes, matte is an intermediate.

5.3.1. Behavior of the Components

The most important equilibrium in copper matte smelting is that between the oxides and sulfides of copper and iron:

$$Cu_2O + FeS \rightleftharpoons Cu_2S + FeO$$

Iron(II) oxide [*1345-25-1*] reacts with added silica flux to form fayalite [*13918-37-1*], a ferrous silicate that is the main component of slag:

$$2\,FeO + SiO_2 \longrightarrow Fe_2SiO_4$$

Liquid iron sulfide [*1317-37-9*] reduces higher iron oxides to iron(II) oxide:

$$3\,Fe_2O_3 + FeS \longrightarrow 7\,FeO + SO_2(g)$$

$$3\,Fe_3O_4 + FeS \longrightarrow 10\,FeO + SO_2(g)$$

The second reaction serves to remove magnetite [*1309-38-2*], which complicates furnace operations because of its high melting point (1590 °C) [54].

The pyrometallurgical production of copper from sulfide ore concentrates may be considered as a rough separation of the three main elements as crude copper, iron(II) silicate slag, and sulfur dioxide. About 20 accompanying elements must be removed from the copper by subsequent refining. Table 8 shows the distribution of important impurities among matte, slag, and flue dust. Precious metals, such as silver, gold, platinum, and palladium, collect almost entirely in the matte, whereas calcium, magnesium, and aluminum go into the slag.

Table 8. Average percentage distribution of the accompanying elements in copper smelting, p. 591[20]

Element	Matte	Slag	Flue dust
Arsenic	35	55	10
Antimony	30	55	15
Bismuth	10	10	80
Selenium	40	–	60
Tellurium	40	–	60
Nickel	98	2	–
Cobalt	95	5	–
Lead	30	10	60
Zinc	40	50	10
Tin	10	50	40
Silver and gold	99	1	–

5.3.2. Matte

The ternary Cu – Fe – S system is discussed in detail in the literature [55–57]. Figure 6 shows the composition of the pyrometallurgically important copper mattes and the liquid-phase immiscibility gap between matte and the metallic phase. In the liquid state, copper matte is essentially a homogeneous mixture of copper(I) and iron(II) sulfides: the pseudobinary Cu_2S – FeS system.

Arsenides and antimonides are soluble in molten matte, but their solubility decreases with an increasing percentage of copper in the matte. Accordingly, when the arsenic concentration is high, a special phase, the so-called *speiss*, can separate. It is produced under reducing conditions in the blast or electric furnace, and its decomposition is complicated (→ Arsenic and Arsenic Compounds, → Antimony and Antimony Compounds).

Compositions of several copper mattes are shown in the partial diagram Cu_2S – FeS – (Fe_3O_4 + FeO) (Fig. 7), which is a section of the quaternary Cu – Fe – O – S system. The density of solid copper mattes ranges between 4.8 g/cm^3 (FeS) and 5.8 g/cm^3 (Cu_2S); liquid mattes have the following densities: 4.1 g/cm^3 (30 wt % Cu, 40 wt % Fe, 30 wt % S), 4.6 g/cm^3 (50 wt % Cu, 24 wt % Fe, 26 wt % S), and 5.2 g/cm^3 (80 wt % Cu, 20 wt % S).

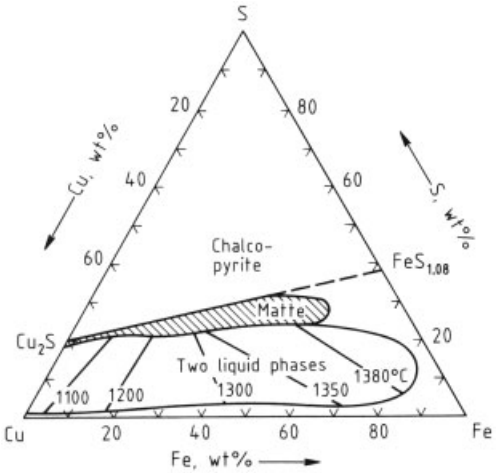

Figure 6. Ternary Cu – Fe – S diagram showing copper mattes and the miscibility gap [55]

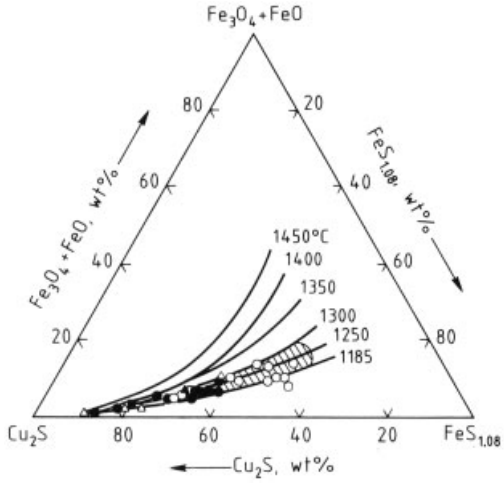

Figure 7. Partial ternary $Cu_2S - FeS_{1.08} - (Fe_3O_4 + FeO)$ diagram [58] showing mattes from various processes
○ Reverberatory furnace; △ Flash smelting furnace; ▲ Electric furnace; ● Blast furnace; ■ Converter

Table 9. Composition (wt %) of typical copper smelter slags [64]

Component	Reverberatory furnace	Flash furnace	Noranda reactor	Peirce – Smith converter
Copper	0.4 – 0.6	1 – 2	8 – 10	2 – 8
Iron (total)	35	40	35	50
Silica	38	30	21	25
Magnetite	7 – 12	13	25 – 29	20 – 25
Ratio of Fe to SiO_2	0.92	1.33	1.67	2.0

5.3.3. Slags

Slags from copper matte smelting contain 30 – 40 % iron in the form of oxides and about the same percentage of silica (SiO_2), mostly as iron(II) silicate. Such slags can be considered as complex oxides in the $CaO - FeO - SiO_2$ system [59] or, because of the relatively low CaO content of most slags, in the partial diagram $FeO - Fe_2O_3 - SiO_2$ [60] (Fig. 8). Ternary systems of these and other pertinent oxide systems are found in the literature [61], [62]. Table 9 shows the general composition of some slag types. Important properties of copper slag systems are compiled in [63].

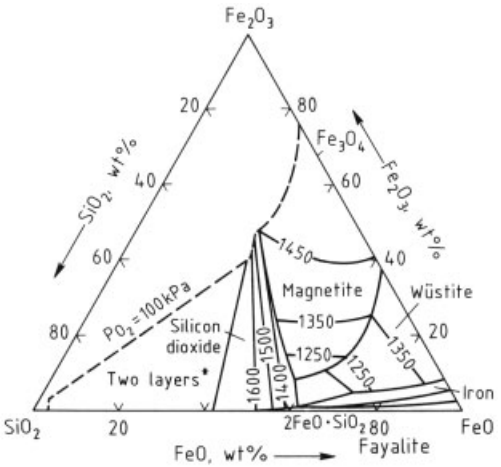

Figure 8. Ternary $FeO - Fe_2O_3 - SiO_2$ diagram [60]

The objectives of matte smelting are to achieve a rapid, complete *separation of matte*

and slag, the two immiscible phases, and a minimal copper content in the slag. The differing properties of slag and matte affect this separation:

1) the low, narrow melting interval of slag
2) the low density of liquid slag (ca. 3.1 – 3.6 g/cm^3) and the difference in density between molten matte and slag of ca. 1 g/cm^3
3) the low viscosity and high surface tension of the slag

The ratio of the weight percent of copper in matte to that of copper in slag should be between 50 and 100. High matte grades generally cause high copper losses in slag. Such losses depend on the mass ratio of slag to copper produced, which is usually between 2 and 3. Copper in slags occurs in various forms, including suspended matte, dissolved copper(I) sulfide, and slagged copper(I) oxide, partially as a silicate, which is typical of nonequilibrium processing.

Slags containing < 0.8 % copper are sold as products with properties similar to those of natural basalt (crystalline) or obsidian (amorphous) or discarded as waste. When liquid slag is cooled slowly, it forms a dense, hard, crystalline product that can be used as a large-size fill for riverbank protection or dike construction and as a medium-size crushed fill for roadbeds or railway ballast. Quick solidification, by pouring molten slag into water, gives amorphous granulated slag, an excellent abrasive that has partially supplanted quartz sand. Ground granulated slag is used as a trace element fertilizer because of its copper and other nonferrous metals.

Most of the newer copper smelting processes produce high-grade mattes, and the short residence time of the materials in the reaction chamber results in an incomplete approach to chemical equilibrium. Both factors lead to high amounts of copper in the slag, generally >1 wt %. Such slags must be treated by special methods for copper recovery (Section 5.5.1).

5.3.4. Oxidizing Smelting Processes

Nearly all pyrometallurgical copper processes are based on the principle of partial oxidation of the sulfide ore concentrates. Methods based on the total oxidation of sulfide ores with subsequent reduction to metal, avoiding the formation of copper matte, are used only rarely because of high fuel consumption, formation of copper-rich slags, and production of crude copper with a high level of impurities.

Prior to the 1960s, the most important way of producing copper was roasting sulfide concentrates, smelting the calcines in reverberatory furnaces, and converting the matte in Peirce – Smith converters. Since that time, the modern flash smelting process with subsequent converting has become predominant. Figure 9 shows the flow sheet of a modern copper smelter, from concentrate to pure cathode copper, including the use of oxygen, recovery of waste heat, and environmental protection. Table 10 compares the important stages and processes of copper production, showing the range of the matte composition for each process.

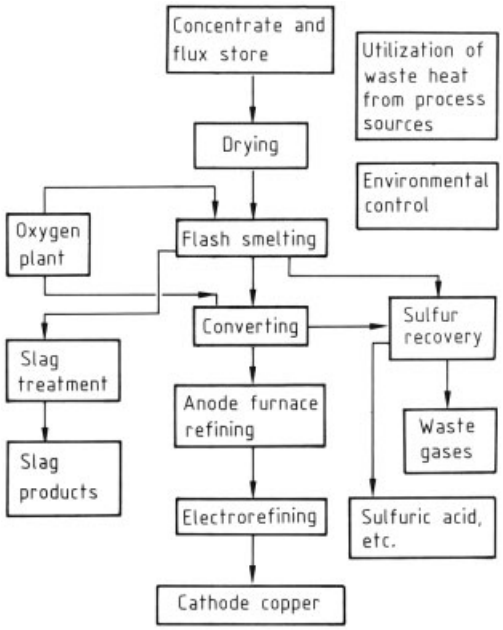

Figure 9. Typical flow sheet for pyrometallurgical copper production from ore concentrates [65]

5.3.5. Proposals

Numerous laboratory experiments and pilot-plant runs have been carried out to develop smelting methods based on elements other than

Table 10. Survey of pyrometallurgical processes for copper production [66]

Process	Matte composition											Description, final product
	wt%S		32	30	28	26	24	22	20		wt%S	
	wt%Fe		48	40	32	24	16	8	0		wt%Fe	
	wt%Cu	10	20	30	40	50	60	70	80	90	100 wt%Cu	
Roasting												solid-state desulfurization
Smelting												
Reverb smelting (green feed)												melting, natural-grade matte
(calcine feed)												
Electric smelting												
INCO oxygen flash smelting												oxygen flash smelting, medium-grade matte
KIVCET process												
Reverb (oxygen sprinkle smelting)												
Outokumpu flash smelting												flash smelting, high-grade matte
Noranda process (matte mode)												
El Teniente reactor												
Converting												
Peirce-Smith												converting
Hoboken												
TBRC												
Flash converter												
Continuous smelting and converting												
KHD Contop process												continuous smelting, blister copper
Mitsubishi process												
Q-S process												single-step smelting, blister copper
Noranda process (copper mode)												
TBRC												
Refining												
Fire refining												refining, anode copper
		concentrate						white metal		blister copper		

oxygen. Two lines of development have dominated, reduction with hydrogen and chlorination, but without leading to commercialization.

Reduction. A potential process involves the reduction of chalcopyrite [70]:

$$2\,CuFeS_2 + 3\,H_2 + 3\,CaO \longrightarrow Cu_2S + 2\,Fe + 3\,CaS + 3\,H_2O$$

$$CuFeS_2 + 2\,H_2 + 2\,CaO \longrightarrow Cu + Fe + 2\,CaS + 2\,H_2O$$

The reduction by hydrogen is endothermic, but the overall reaction with calcium oxide is exothermic. A similar proposal [71] is based on the reaction of a metal sulfide with steam in the presence of calcium oxide:

$$\begin{array}{l} MS + H_2O \longrightarrow MO + H_2S \\ H_2S + CaO \longrightarrow CaS + H_2O \\ \hline MS + CaO \longrightarrow CaS + MO \end{array}$$

Chlorination. The reactions of chalcopyrite with chlorine are also of interest [72]:

$>500\,°C$:

$$CuFeS_2 + 2\,Cl_2 \longrightarrow CuCl_2 + FeCl_2 + 2\,S$$

$<500\,°C$:

$$CuFeS_2 + 3.5\,Cl_2 \longrightarrow CuCl_2 + FeCl_3 + S_2Cl_2$$

followed by electrolysis of the molten copper (I) chloride.

The recently proposed thermoelectron chlorination process is another variation [73]:

$$CuFeS_2 + 2\,Cl_2 \longrightarrow CuCl + FeCl_3 + 2\,S$$

Electrolysis. Another approach, to avoid converting, proposed the electrolysis of molten copper matte [74].

5.4. Traditional Bath Smelting

At the end of the Middle Ages, copper was produced by the German or Swedish smelting process that involved roasting reduction with up to seven process steps in small shaft furnaces. Around 1700, reverberatory furnaces were constructed in which the ore was processed by roasting, the so-called English or Welsh copper smelting process, originally with ten process steps.

The large blast and reverberatory furnaces of the 1900s were derived from these principles. Later, the electric furnace for matte smelting was developed. Newer processes are the Isasmelt/Ausmelt/Csiromelt (furnace with vertical blowing lance), the Noranda and CMT/Teniente reactors (developed from converters), the Russian Vanyukov, and the Chinese Bayia process.

5.4.1. Blast Furnace Smelting

The blast or shaft furnace is well-suited for smelting high-grade, lumpy copper ore. If only fine concentrates are available, they must first be agglomerated by briquetting, pelletizing, or sintering. Because of this additional step and its overall low efficiency, the blast furnace lost its importance for primary copper production and is currently used in only a few places, for example, Glogow in Poland.

Smaller types of blast furnace, however, are used to process such copper-containing materials as intermediate products (e.g., cement copper or copper(I) oxide precipitates), reverts (e.g., converter slag, refining slag, or flue dusts), and copper-alloy scrap.

The construction of the furnace is basically related to that of the iron blast furnace, but there are considerable differences in design, especially in size and shape: the copper blast furnace is lower and smaller, and its cross section is rectangular. Developments adopted from the steel industry include use of preheated air (hot blast), oxygen-enriched air, and injection of liquid fuels.

The furnace is charged with alternate additions of *mixture* (copper-containing materials and accessory fluxes such as silica, limestone, and dolomite) and coke (which serves as both fuel and reducing agent). There are three zones in the furnace:

1) In the heating zone (the uppermost), water evaporates and less stable substances dissociate.
2) In the reduction zone, heterogeneous reactions between gases and the solid charge take place.

3) In the smelting zone, liquid phases react.

The usual mode of operation is *reducing smelting*, which yields two main products. Sulfide ores are used to produce a matte (40 – 50 wt % Cu) and a disposal slag (ca. 0.5 wt % Cu). In contrast, oxide ores are processed directly to impure black copper (≦95 wt % Cu) and to a copper-rich slag. The two ore types can be smelted together to produce matte and a slag with low copper content. Another product is top gas, which contains flue dust. Ores that contain high concentrations of arsenic and antimony also form speiss, which is difficult to decompose.

In Poland (KGHM Polska Meidz S.A. Smelters in Glogow I and Legnica) the blast-furnace technology is well adapted to Polish copper concentrates, which contain about 20 – 30 % Cu like normal chalcopyrites but also 5 – 10 % of organic carbon and only 9 – 12 % S. Also these concentrates have relatively high lead (up to 2.5 %) and arsenic (up to 0.3 %) content. The organic carbon compounds provide about 40 – 60 % of the process energy; the rest is added by coke. The matte has about 58 – 63 % Cu and 3 – 6 % Pb. The slag contains less than 0.5 % Cu. The off-gas from the blast furnaces (three in each plant) is mixed with the converter gases (Hoboken Converter). It contains 7 – 10 % SO_2 and is sent to sulfuric acid production. The production figures are 80 000 t/a converter copper in Legnica, and 200 000 t/a in Glogow I Smelter [214].

5.4.2. Reverberatory Furnace Smelting

The reverberatory furnace dominated copper matte smelting for much of the 1900s, because it was an excellent process for smelting fine concentrate from flotation. It is a fossil fuel fired hearth furnace for smelting concentrate and producing copper matte. The reverbs began to decline in the 1970s with the adoption of environmentally and energetically superior processes like flash smelting. Probably the last one was erected in 1976 in Sar Chemesh, Iran. In 1980 about 100 reverbs were in operation, but in 1994 the number had decreased to about 25. As shown in Figure 10 it is a rectangular furnace up to 10 m wide and 35 m long with internal brick lining. Throughputs of up to 1100 t/d concentrate or a mixture of concentrate and calcine could be processed. The charge is passed into the furnace near the burners through the roof or lateral openings. As fossil fuel, pulverized coal, heavy oil, or natural gas is used. Normally the burner is located in the front wall of the furnace. The atmosphere is slightly oxidizing, and the maximum flame temperature is up to 1500 °C. During the 1980s oxygen – fuel burners have been set in the roof to fire downwards directly on the top of the bath. This increases the smelting rate by up to 40 % and the energy efficiency to about 50 % [76], [215]. Another invention was the sprinkler burner for feeding concentrate, coal, and flux from the top of the reverb [77].

The back half of the furnace is the settling zone. A matte grade of between 35 and 60 % Cu is produced, depending on whether concentrate or calcine is fed. The slag has low copper content (< 1 %). Normally, the converter slag is also fed back to the reverb. Sometimes problems with solid magnetite accretions in the furnace occur.

The off-gas has a temperature of about 1250 °C. It is diluted by the combustion air and contains only about 1 – 2 % SO_2. This is too low for efficient SO_2 capture such as sulfuric acid production by the contact process. Improvements have been achieved by using higher oxygen enrichment of the burners, but the off-gas is still too dilute, and therefore this furnace is unsuitable for many regions in the world because of environmental problems. Another disadvantage is the very high energy consumption. The reverb has the highest energy consumption of all copper matte smelting processes. A good present example for operating reverbs is the Ilo Smelter in Peru (Table 11). There two reverbs smelt about 2000 t/d of concentrate, producing a matte containing 35 % copper for the subsequent Peirce – Smith converters. In the latest smelter enlargement one reverb was replaced by an Teniente converter (CMT), and a sulfuric acid plant for partial SO_2 recovery was built.

5.4.3. Electric Furnace Smelting

In regions where relatively cheap electrical power was available, electric furnaces were built. The Scandinavian countries were the first to perform electric matte smelting: 1929 at Sulitjelma, Norway; 1938 at Imatra, Finland, and

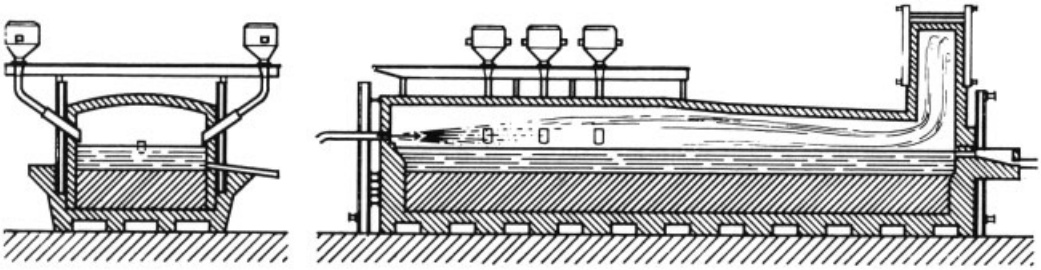

Figure 10. Schematic longitudinal and cross-section views through a reverberatory furnace [45]

1949 at Rönnskär, Sweden, which is still operating.

Electric furnaces have the rectangular ground plan and the dimensions of larger reverberatory furnaces. Along the centerline, up to six Söderberg continuous self-baking electrodes are used with alternating current (Table 11).

Table 11. Examples of traditional reverberatory and electric furnaces

	Boliden Rönnskär Smelter	Southern Peru Ilo Smelter
Type	electric furnace	reverbatory
Inside dimensions $l \times w \times h$, m	$24 \times 6 \times 3$	$36 \times 10 \times 3.4$
Electrodes/burners	6	8
Smelting capacity, t/d	950 (hot calcine)	1000 (wet concentrate)
Matte grade, % Cu	51	35
Off-gas volume, m^3/h (STP)	25 000	100 000
SO_2 content, %	4.5	1.2
Energy consumption	300 kW/t concentrate	194 kg oil/t concentrate

The smelting process in such units is similar to the operation in reverbs, but the concentrate is usually dried and roasted before charging to increase the smelting capacity. The converter slag is returned to the furnace. The composition of matte and slag resembles that of the reverb products, but the content of magnetite is lower. No fuel is burned, and the volume of waste gases and the quantity of flue dust are small. The SO_2 content of the waste gas can be 10 %. The off-gas also contains carbon oxides from the consumption of the graphite electrodes.

A significant difference between the electric furnace and the reverb is the inversion of the temperature gradient in the furnace cavity. In the reverb, the combustion gases have the highest temperature, whereas in electric matte smelting, the waste gases are ca. 500 °C cooler than the slag phase, which is heated by the electric energy. Accordingly, in the electric furnace, cheap refractories are sufficient for lining the walls above the slag zone and the roof; only a common arch is required.

The temperatures of both molten phases depend on the submergence of the electrodes, and the required heat is controlled by the electrical power supplied. Heat transfer takes place chiefly by convection, which causes intense circulation in the molten bath.

The current is divided into two partial circuits, through slag and through matte. The difference in conductivity is great, slag : matte ratio of $1 : 10^2$ to $1 : 10^3$; therefore, the depth of immersion of the electrodes into the liquid slag must be precisely controlled. As the electrodes immerse deeper into the slag, more current flows through the matte. If they touch the matte layer, a short circuit occurs. These considerations lead to an approximate relation between the electrical power input and the depth of the slag layer: 6000 kVA and 0.5 m, 30 000 kW and 1.0 m, and 50 000 kW and 1.5 m.

The smelting capacity of electrical furnaces is higher than that of reverbs.

Brixlegg Process. Lurgi developed and practiced at Brixlegg, Tirol, Austria, a modification of the old roasting reaction process. Nearly dead-roasted concentrates are reduced in batches by coal in a small special circular electric furnace (2500 kVA, 5-m diameter). The crude copper averaged only 95 wt % Cu, and the operation has been discontinued.

5.4.4. Isasmelt Furnace

The Isasmelt furnace and the closely related Ausmelt furnace were both developed in the 1980s on the basis of work conducted by CSIRO, Australia [216]. The process consists of a tall cylindrical furnace (small diameter but large height, internally brick lined) equipped with a vertical blowing lance. The lance is submerged into the bath (slag) and blows oxygen-enriched air and/or fossil fuel (natural gas). The moist concentrate and flux material is pelletized and fed through the roof of the furnace onto the bath. Most of the energy requirement comes from the reaction of the concentrate. The lance is made from stainless steel. It is cooled by swirling the air in the annulus between the pipes so that a protective layer freezes on that part of the lance which is immersed in the slag bath (0.3 m) [217]. The lance has to be replaced from time to time, for example, once a week. In the continuous Isas process, an emulsion of matte and slag is produced, which is tapped periodically to a electric or fuel-fired settling furnace for matte/slag separation. The off-gas is high in SO_2. For cooling it passes through a waste-heat boiler.

Two Isasmelter for copper matte are running (Mount Isa, Australia; Cypus Miami, Arizona). Figure 11 shows the Cyprus furnace schematically. About 90 t/h of copper concentrate is processed, and about 45 t/h of matte is produced (58 % Cu). The off-gas contains about 35 % SO_2 directly behind the furnace and is then diluted by air to 8 – 9 %. The Isasmelt process has replaced the electric (reverberatory) furnace, which is now used for matte/slag settling. Smaller Isasmelt or Ausmelt furnaces have also been proposed or built for secondary copper smelters and primary smelters like in China. It is possible to run the process batchwise, smelting and converting the material in the furnace for producing blister copper. Also Isasmelt furnaces have been built for lead refining.

5.4.5. Noranda Process

The Noranda process was initially constructed as a continuous smelting and converting process which produced blister copper from copper concentrate. The first reactor was built in 1973 at Horne smelter [105]. This direct-to-blister process operated from 1973 to 1975 [218]. It was switched to high-grade copper matte smelting (as it now still operates) because of excessive impurity levels in the copper anodes and to increase the smelting rate.

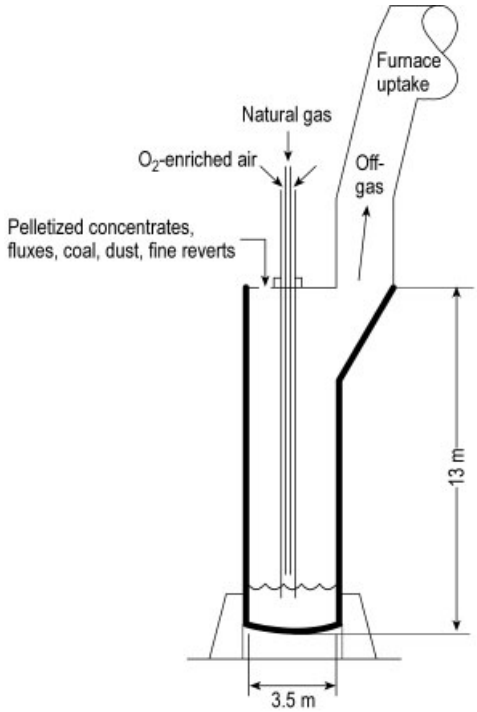

Figure 11. Isasmelt furnace

The Noranda furnace [219] is a horizontal steel barrel with an inner brick lining (Figure 12. The diameter is about 5 m and it is about 20 m long. The process runs continuously. At one end of the reactor, pelletized wet concentrate, coal, flux, revert materials, and scrap are thrown into the furnace and on to the top of the molten bath by a high-speed slinger belt. Feeding fine concentrate through the tuyeres is also possible. The feed material is absorbed and melted in the liquid matte/slag bath. On the side there are 20 – 40 tuyeres in the cylindrical part of the vessel. Oxygen-enriched air (ca. 40 % O_2) is blown through these tuyeres into the matte phase. The tuyeres have to be punched periodically, as in a Peirce – Smith converter. A layer of matte and slag must always be present in the furnace.

The oxidation reaction of the sulfides (mostly iron sulfide) and the added coal provide the process energy. Matte is tapped at the bottom of the vessel. Slag is periodically tapped at the

other end of the cylindrical furnace where there are no tuyeres and the bath is quiet enough for matte and slag settling. The matte usually contains 70 – 75 % Cu (white metal) and is sent to a subsequent converter. This operation gives a large heat evolution from Fe and S oxidation in the Noranda furnace. This avoids excessive retention of impurities like Sb in the matte. The slag contains about 3 – 7 % copper and goes to further treatment like electric furnace or solidification/comminution/flotation. The off-gas contains about 15 – 20 % SO_2. It passes through the large mouth at the top of the reactor into a hood and is sent to cooling, cleaning, and sulfuric acid production. In case of emergencies the furnace can be rolled out so that the tuyeres have a secure position above the molten bath.

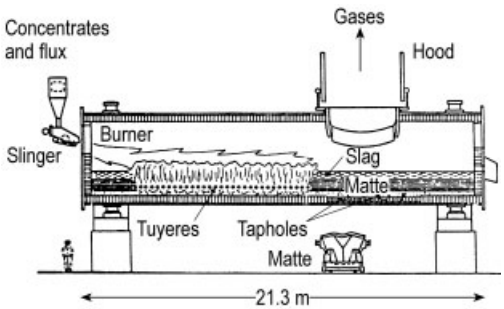

Figure 12. Noranda furnace

The Noranda process is well suited for smelting scrap and residues, ashes, precious metals scrap, telephone scrap, and computer/electronic scrap. Due to the high temperature and smelting intensity, potentially harmful organics are oxidized to CO_2 and H_2O. The productivity can be increased by using higher oxygen enrichment. Then less fossil fuel is needed and the amount of off-gas decreases.

5.4.6. CMT/Teniente Process

The Caletones Matte Treatment (CMT) or so called Teniente process [220], [221] is similar to the Noranda process. The important differences are that no coal is included in the charge and molten low-grade copper matte is periodically added as fuel. Therefore, the Teniente furnace is mostly regarded as a converter.

The process was developed at Teniente Smelter by adding dry concentrate to a Peirce – Smith converter for increasing existing smelter capacity. The Teniente process has been widely adopted in Latin America. Like the Noranda reactor, it is a robust technology but has still not demonstrated that it can meet the highest environmental standards. The furnaces are horizontal cylinders, 4 – 5 m in diameter and 14 – 22 m long (Figure 13). Inside, the converter is brick-lined, and it has 30 – 50 tuyeres for blowing oxygen-enriched air (30 – 32 % O_2). Solid feed such as wet concentrate, flux, or recycled materials are blown continuously through a high-pressure gun (Garr gun) on the top of the molten bath. Dry concentrate is injected additionally through four or five tuyeres in the bath. Molten matte (35 – 40 % Cu) coming from a reverb is added periodically through the mouth of the converter. The temperature in the converter is controlled by the ratio of wet (lowers temperature) and dry (increases temperature) feed material.

High-grade matte (72 – 77 % Cu) is produced and tapped periodically to the subsequent converting step. Slag with 4 – 7 % Cu is tapped for further processing. The off-gas is collected by a water-cooled hood above the mouth of the converter. The off-gas is mixed with air because the hood is not tight to the converters mouth. Therefore, the SO_2 content is only about 10 %. The gas is cooled with water sprays, the flue dust is stripped and then the gas is sent to sulfuric acid production (or in some cases to the stack).

With its molten-matte requirement, the CMT is dependent on the operation of a reverberatory furnace with all the associated environmental problems (SO_2 emission). To overcome this reliance, some possibilities have been tested, all of which increase the energy in the vessel: increased oxygen enrichment of the blast, increased injection of dry concentrate through the tuyeres, and burning small amounts of fossil fuel. Today there are CMTs which do not need low-grade molten matte.

5.4.7. Vanyukov Process

The Vanyukov process was developed in the 1970s in Norilsk, Russia. It is similar to the Noranda process, but the furnace is rectangular with tuyeres on both sides, blowing highly enriched air (70 – 90 % O_2) in the slag zone (Figure 14).

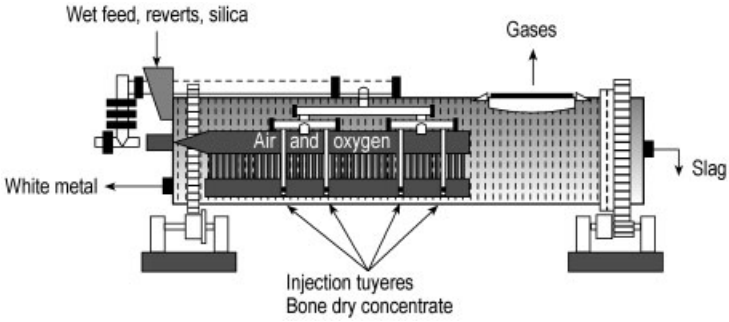

Figure 13. Teniente converter (CMT)

Moist copper concentrate is charged through ports in the roof of the furnace. The process produces matte (ca. 50 % Cu), slag (1.5 – 2 % Cu) and off-gas (ca. 30 % SO_2). The furnace operates continuously, and the matte and slag leave the furnace continuously through two siphons. The matte is sent to converting process, and the slag is cleaned in an electric furnace [222], [223].

In contrast to the Noranda furnace the Vanyukov reactor is stationary. The tuyeres can not be tilted out of the bath in emergency situations. The advantage of the stationary system is the simple and very efficient SO_2 capture. Six Vanyukov furnaces are running in the CIS. With a relatively small furnace (10 m long, 2 m wide, and 6 m high) about 1500 to 2000 t/d concentrate can be processed.

All the submerged-tuyere processes like Noranda, CMT, and Vanyukov have relatively short campaign lifetimes (one year) because of high erosion in the tuyere zone.

5.4.8. Baiyin Process

At the Baiyin Copper Smelter in China, a process similar to the Noranda furnace was developed in the 1980s [78]. Concentrate is added on the top of the molten bath, and air/oxygen is blown through submerged tuyeres into the melt. The Baiyin furnace has a wall inside to divide the reaction and the settling zone. In China there are at least two furnaces in operation, but few data are available.

5.5. Autogenous Smelting

Autogenous smelting involves the use of combustion heat generated by reactions of the feed in an oxidizing atmosphere in which the sulfide concentrate acts partly as a fuel. The formerly separate steps of roasting and smelting are combined into a roast – smelting process. The spatial and temporal coupling of exothermic and endothermic reactions leads to an economical process, but the sensible heat of nitrogen in the air causes a deficit in the heat balance. In practice, various measures must be taken:

1) Increasing the oxygen content of the combustion air and even using pure oxygen
2) Preheating combustion air with waste heat or in a preheater
3) Combustion of natural gas, fuel oil, or pulverized coal in supplementary burners

To achieve autogenous operation and prevent agglomeration of the feed, the moisture in the concentrate must be removed by drying before charging. The quantity of added fluxes is minimized as far as practical to save energy.

Because the residence time of the sulfide particles in the reaction chamber is only a few seconds, kinetic conditions predominate over the thermodynamic equilibrium. The reactants form a heterogeneous system, with the feed suspended in the gas flow, thus the term *smelting in suspension.*

These processes have several advantages:

1) High rate of reaction, increasing the production rate
2) Energy savings

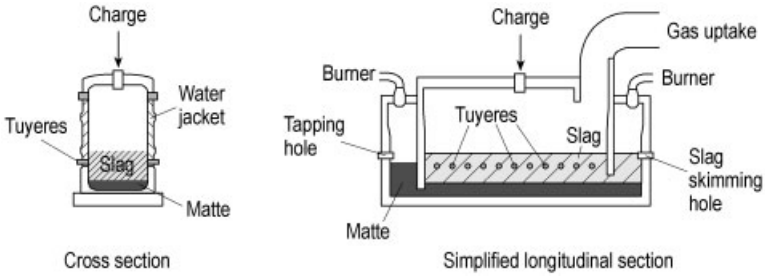

Figure 14. Vanyukov furnace

3) Low volume of off-gas and correspondingly high concentration of SO_2 and low quantity of flue dust, if oxygen is used

However, a typical disadvantage is the high copper content of the slags, and the relatively high flue dust content in the off-gas, which can cause problem in the waste-heat boiler.

5.5.1. Outokumpu Flash Smelting

After a preliminary test in 1946, the first full-scale flash smelting furnace started operation in 1949 at Outokumpu Oy, Harjavalta (Finland). Flash smelting has been the most widely adopted copper matte smelting process since 1970. More than 40 furnaces have been installed to replace reverberatory furnaces or at new smelting operations. Nowadays the Outokumpu-type smelters account for more than 50 % of world primary smelter capacity.

Furnace Design. The flash furnace can be envisioned as a combination of a reverb with a roaster in which smelting takes place simultaneously. Figure 15 shows a Outokumpu flash furnace designed in 1988 for smelting more than 3000 t/d of copper concentrate – a production rate several times higher than those of earlier furnaces. The current trend is towards even bigger units. The furnace is 24 m long, 7.5 m wide, and 2 m high in the settling area. The hearth or settler is equipped with oil or natural gas burners.

The reaction shaft is 6 m in diameter and 7 m in height. It is equipped with one concentrate burner. The gas off-take shaft is 9 m high and has a diameter of 5 m. (All dimensions inside the furnace, which is lined with chrome magnesia bricks.) The furnace has eight matte tapholes and four slag tapholes. The reaction shaft and large areas of the settler and the gas off take are water-cooled to prevent overheating and weakening of the furnace structure. The reaction shaft of many furnaces is cooled by cascading water outside the steel shell. The other areas are usually cooled by special copper cooling jackets, which are built into the furnace walls [224].

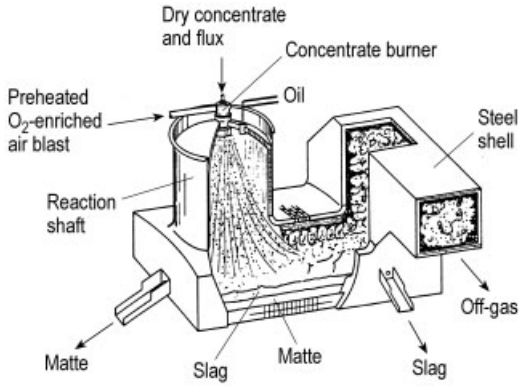

Figure 15. Outokumpu flash smelting furnace

An important feature of the flash smelter is the concentrate burner, which suspends the feed particles in the oxygen-enriched air in the hot reaction shaft. The aim is to ensure a rapid, uniform (over the shaft, to prevent hot spots), and efficient concentrate oxidation and generate minimum flue dust. Older flash smelters have four burners, but nowadays a single burner is used because it is easier to control the burner parameters. The burner (Fig. 16) consists of two concentric pipes: a central pipe for the solid feed and an annulus for the oxygen-enriched blast. In the inner pipe a distributor cone is installed to optimize the particle distribution [225]. The gas velocity out of the burner is up to 200 m/s.

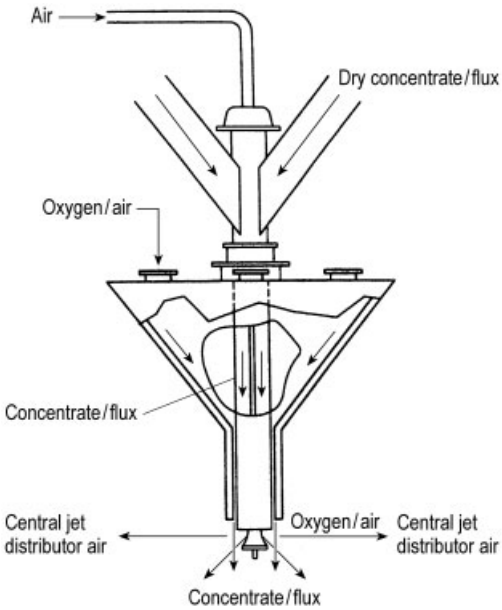

Figure 16. Schematic of a flash furnace concentrate burner

Flash furnaces need auxiliary equipment like solids dryers (rotary, flash, or steam dryers), delivery systems, oxygen plant, blower and preheater systems, waste-heat boiler, dust recovery and recycling system, slag-treatment furnace, and sulfuric acid plant for treating the highly SO_2 enriched off-gas.

Operation. Various sorts of concentrate and a few flux materials (mainly silica) are blended and then dried before feeding into the flash smelter. Recycle materials such as flue dust from the waste-heat boiler and the converters or others are also fed to the flash furnace. The process operates continuously 24 h a day, and campaigns are run for as long as possible with out interruption.

The feed mixture is suspended at the top of the reaction shaft, and sulfur and iron are oxidized partly. The reactions are strongly exothermic, and the particles melt to form copper matte and slag. At oxygen enrichments of between 45 and 80 % (depending on concentrate composition and preheated blast), the process runs autothermally. Less oxygen enrichment requires fossil fuel combustion in the reaction shaft. The temperature in the reaction shaft is about 1250 – 1300 °C. The optimal matte concentration is between 60 and 65 % Cu. Higher Cu concentrations the in the matte are also possible, for example, before a flash converting process.

The matte is tapped as needed by the subsequent converting step. Slag is tapped nearly continuously. The slag contains 1 – 2 % of copper, mainly in matte droplets. The slag temperature is about 1230 – 1250 °C; the matte is about 30 – 50 °C cooler. Typical operating data are shown in Table 12.

Table 12. Operating data for Outokumpu flash furnaces

	Norddeutsche Affinerie, Hamburg, Germany	BHP Magma Metals, San Manuel, Arizona
Furnace commissioning date	1972	1988
Size, inside brick, m hearth $w \times l \times h$	$6 \times 20 \times 3$	$7.5 \times 24 \times 2$
reaction shaft $d \times h$	6×9	6×6.9
Concentrate throughput, t/d	2000 (30 – 34 % Cu)	3000 (30 % Cu)
Blast temperature, °C	400 – 420	25
Oxygen enrichment, %	45 – 55	60 – 80
Blast flow-rate, m^3/h (STP)	25 000	51 000
Matte grade, % Cu	63	63
Off-gas volume, m^3/h (STP)	55 000	55 000
SO_2 Content	25 – 30	35
Recycling of converter slag	in flash furnace	flotation
Slag treatment	electric furnace	flotation
Fossil fuel input in shaft	0	0
Fossil fuel for settler burner	5 m^3/h (STP) natural gas/t conc.	15 – 20 L oil/t conc.

Slag Cleaning. In practice two methods are used: froth flotation and treatment in small electric slag furnaces. The decision for one or the other is determined by economics. In froth flotation the slag concentrate is returned to the feed, and the tailings are discarded. The second method involves the reduction of flash furnace slags in an electric furnace, which yields a matte and slag that is suitable for sale.

Off-gas and Flue Dust. The waste gases are separated from part of the flue dust at ca. 1300 °C in the off-take shaft and pass through a waste-heat boiler for generating steam, and subsequently to an electrostatic (Cottrell) precipitator for separating the mass of flue dust, which is recycled to the feed. The precleaned off-gas with SO_2 content of >20 % is usually processed to sulfuric acid.

The flue dust consists chiefly of sulfates and basic sulfates of copper, lead, and zinc, as well as some volatile compounds of arsenic, antimony, bismuth, and selenium. Repeated recycling makes it possible to enrich selected elements for later extraction. The quantity of flue dust is between 4 and 10 % of the input.

Process Control. Five main parameters have to be controlled and adjusted carefully to obtain the highest smelting rate of concentrate, constant matte composition, specified slag composition and furnace temperatures, and minimum energy consumption:

1) Concentrate feed rate
2) Flux and recycled materials feed rate
3) Blast input and temperature
4) Oxygen enrichment of the blast
5) Fossil fuel combustion

In the flash smelters constructed between 1980 and 1999 most of these parameters are measured and adjusted automatically.

Special Operations. There are several variations on the Outokumpu flash smelting process.

Two Outokumpu copper smelters (Tamano, Japan and Pasar, Philippines) have inserted carbon electrodes in the settler. The electrodes are submerged in the slag layer to heat the slag with electric power. The system was installed to recover Cu from the slag and reduce magnetite without using an additional slag-cleaning furnace [80], [226]. The system at Tamano Smelter ceased operation in 1988 because the operators found that combustion of coke – oil mixtures also gives low copper content in the slag (0.6 %) [227].

Occasionally molten converter slag is recycled through the flash furnace like in a reverb. It is added from ladles through a spoon and launder system high on the flash furnace [228].

At two smelters in the world Outokumpu flash furnaces are used to smelt blister directly from concentrate (see Section 5.8).

Trends. The diversity of modified furnace constructions and operating methods shows the adaptability of the Outokumpu process to different raw materials and smelter locations. As increasing oxygen content has been used in the blast air, the process has approached the principle of INCO flash smelting. On the other hand, newer developments aim at the production of blister copper or white metal (75 – 80 wt % Cu) from concentrates in only one vessel, thereby approaching continuous smelting and converting as in the Noranda process.

5.5.2. Inco Flash Smelting

Inco Metal Co. was the first company in the nonferrous metal industry to use commercially pure oxygen for autogenous flash smelting. After tests in the late 1940s, two smaller furnaces went into operation in 1953 [82]. Today a total of five Inco furnaces are operating worldwide (including Ni/Cu and one older plant in Uzbekistan [83]). Since 1990 no Inco furnace has been built.

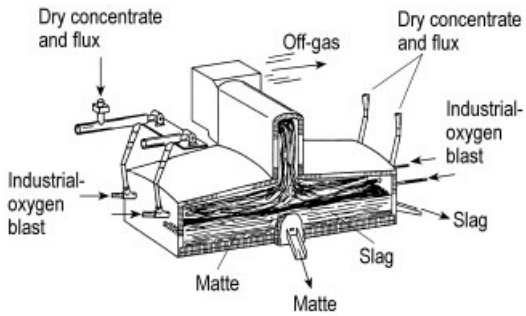

Figure 17. Inco flash smelting furnace

Furnace Design. In contrast to Outokumpu-type furnaces the Inco process uses pure commercial oxygen (95 – 98 %) instead of enriched air. The oxygen blast and concentrate are fed through the burners (generally two) in the side walls of the furnace. The furnace is smaller than the Outokumpu furnace and has only one shaft, which is located at the center of the roof and takes the off-gas to the cooling system. Using pure oxygen makes the off-gas volume much smaller than in the Outokumpu furnace. Figure 17 shows a sectional view of an Inco flash furnace. The burner is completely different from the Outokumpu type. It has low velocity of about 30 m/s into the furnace. This gives a short flame and also low flue dust generation. The Inco furnace is not equipped with other fossil-fuel burners. Because of the small size of the furnace and

the surrounding equipment, this process is suitable for replacing old furnaces within existing smelters.

Operation. Dry concentrate and flux are fed continuously to the Inco furnace. The products of the process are matte (50 – 60 % Cu), slag, and off-gas with 70 – 80 % SO_2. The slag is tapped at the sidewalls of the furnace, and the matte in the middle part under the offtake. The temperatures are similar to those of the Outokumpu furnace. In contrast to an Outokumpu furnace the amount of energy which is removed with the off-gas is much lower because of the small off-gas quantity. Therefore the Inco furnace is not equipped with a waste-heat boiler for energy recovery. The off-gas is cooled by different techniques such as splash towers, cyclones, or scrubbers [229]. The absence of a waste-heat boiler avoids out-of-service times of the furnace. Furthermore, the copper content in the slag of the Inco furnace is lower (0.5 – 1.5 %), and a slag-cleaning furnace is not needed. Only Hayden Smelter operates an additional electric furnace. Some operating data of the Chino, New Mexico furnace are listed in the following:

Furnace commissioning date	1984
Size, inside brick, m hearth $w \times l \times h$	$5.6 \times 23 \times 5$
gas-off take $w \times h$	3.9×7.7
Concentrate throughput, t/d	1520
Blast temperature, °C	40
Oxygen enrichment, %	98
Blast flow-rate, m^3/h (STP)	12 500
Matte grade, % Cu	58
Off-gas volume, m^3/h (STP)	20 000
SO_2 Content	75
Recycling of converter slag	in flash furnace
Slag treatment	discarded
Fossil fuel input	0

Process Control. The aim of running the Inco furnace is the same than the Outokumpu furnace: highest throughput rates with constant furnace properties such as matte concentration and furnace temperatures. In comparison with the Outokumpu process, it is more complicated to keep the matte grade constant because the oxygen enrichment is not adjustable. This could be done by adjusting the mix of the feed (high or low in copper) or adding fossil fuel for oxygen consumption. But this is in conflict with the principle of the process and does not seem economic.

5.5.3. KIVCET Cyclone Smelting

Developments in power-plant technology have led to adoption of the cyclone principle by the metallurgical industry. The acronym KIVCET uses the initial letters of the following Russian terms: oxygen, vortex, cyclone, electrothermic. The development began in 1963, and the first plant was operated by Irtysh Polymetal Combine in Glubokoe, Kazakhstan. The process is not widely used for copper smelting.

The method is aimed at processing copper sulfide concentrates that contain considerable amounts of other metals. The essential part of the continuously operated plant is the smelting cyclone, in which the concentrates are fed vertically, and technical-grade oxygen ($\leqq$95 %) is blown in horizontally, so that reaction takes place rapidly above 1500 °C.

The furnace is divided to allow separating and settling of the reaction products, in this respect similar to a reverb. In contrast to the separation chamber, the atmosphere in the electrically heated settler is maintained in a reducing state so that the slag does not need special posttreatment.

The off-gas volume is small, and the content of SO_2 can be up to ca. 80 %. Metals are recovered from the flue dust of both the separating chamber and the settler. Table 13 and Figure 18 explain the process.

Table 13. KIVCET process: analysis and yield [84]

Analysis and yield	Metal		
	Cu, wt %	Zn, wt %	Pb, wt %
Analysis			
Concentrate *	25.6	10.0	1.7
Copper matte	50.0	2.5	2.0
Slag	0.35	3.5	0.2
Yield			
In matte	99.1	12.7	60.0
In oxidic condensate	–	71.0	34.1

* Also 24.0 % Fe and 33.0 % S.

5.5.4. Contop Matte Smelting

Using a KIVCET license, KHD Humboldt Wedag AG, Cologne developed the combined Contop process [85]. Contop stands for continuous smelting and top blowing. It is a combination

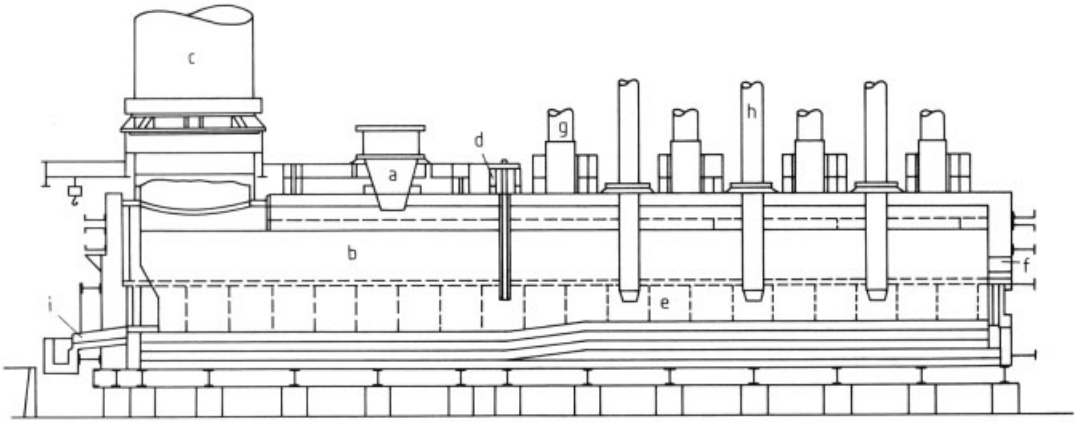

Figure 18. KIVCET furnace [84]
a) Smelting cyclone; b) Separating chamber; c) Cyclone waste-gas offtake; d) Partition wall; e) Settling reduction hearth; f) Slag tap hole; g) Feed of reductant and offtake for waste gas from the hearth; h) Electrical resistance heating; i) Matte taphole

of a high-intensity smelting process in a cyclone and secondary treatment of the molten phase with a top-blown jet. Contop-type cyclone burners have been used on reverberatory furnaces, for example, at Chuquicamata and Palabora. In 1993 a complete Contop smelter started operation at El Paso (Asarco) [230], replacing the old reverberatory furnace. The furnace is shown in Figure 19. It is a kind of a hearth furnace. Two cyclone burners are placed on the top of one end of the furnace. Each is fed with about 25 t/h of dry concentrate and flux and about 95 % oxygen blast. Natural gas is fed to the cyclone for energy adjustment. The cyclones are made of stainless steel and are water-cooled and about 2 m in diameter and 4 m in height. Inside they are protected by a layer of solidified matte – slag. Matte with 55 – 60 % Cu is produced. At the other side of the furnace the off-gas is passed through a waste-heat boiler. The amount of off-gas is about 65 000 m^3/h (STP) with about 24 % SO_2. Its temperature is about 1150 °C. After cooling, the gas is washed and sent to sulfuric acid production.

To keep the slag hot, oxy-fuel burners are installed on the roof of the furnace between the cyclones and the off-gas shaft [231]. The slag contains 0.8 % Cu and can be discarded. The advantage of the Contop cyclone is high flexibility in concentrate composition, low dust production (ca. 2 % of the feed material), high volatilization of zinc and bismuth from dirty concentrates, and potential steam generation from the cyclones. Compared with Outokumpu flash smelting, the Contop process is more energy intensive (higher specific cost per tonne of concentrate). Another disadvantage is the relatively short lifetime of the cyclone (about one year). Therefore only one plant was built until now.

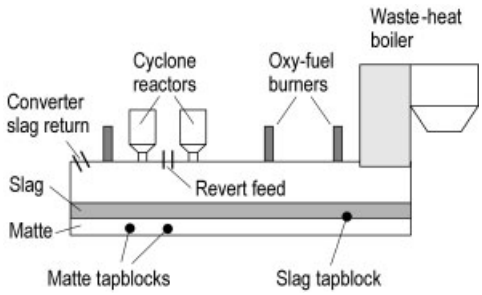

Figure 19. Contop furnace

5.5.5. Flame Cyclone Smelting

The flame cyclone reactor (FCR) process was suggested by LURGI and Deutsche Babcock AG ca. 1975 and has been demonstrated at a pilot plant of Norddeutsche Affinerie in Hamburg with a capacity of ca. 10 t/h. It is a high-temperature (> 1500 °C) reaction for autogenous smelting of copper sulfide concentrates in a high-oxygen atmosphere (up to ca. 75 %). A second characteristic is the simultaneous removal

of volatile compounds of other elements, such as zinc, cadmium, tin, lead, arsenic, antimony, and bismuth, as oxides and basic salts, in the flue dust. The SO_2 content of the off-gas is greater than 50 %. The products are a high-grade matte, containing up to 80 % Cu, and slag, which separates in a settler [86].

The principle of this method differs from that of the KIVCET process in that the reaction in the FCR process takes place in a special chamber situated before the cyclone, where the molten droplets are separated by centrifugal force. The process is well suited for processing complex or dirty concentrates, but until now never installed in practice.

5.6. Discontinuous Matte Conversion

Matte produced by smelting processes is treated in the molten state by blowing with air; this stage of concentration is known as converting. Copper and iron sulfides, the main constituents of matte, are oxidized to a crude copper, ferrous silicate slag, and sulfur dioxide.

The batch converting process has been employed for many decades in two operating steps at ca. 1200 °C in the same vessel. Investigations and development of continuous methods are being made [87], [88].

The conventional converting of matte is a batch process that yields in the first step an impure copper(I) sulfide containing ca. 75 – 80 wt % Cu, the so-called *white metal*, and in the second step the converter, or blister, copper averaging 98 – 99 wt % Cu. The name *blister copper* derives from the SO_2-containing blisters that are enclosed in the solidified metal.

First Step. The main reactions are oxidation of iron(II) sulfide and slagging of iron(II) oxide by added silica [*7631-86-9*] flux:

$$2\,FeS + 3\,O_2 \longrightarrow 2\,FeO + 2\,SO_2$$

$$2\,FeO + SiO_2 \longrightarrow Fe_2SiO_4$$

Formation of magnetite occurs near the tuyeres:

$$3\,FeS + 5\,O_2 \longrightarrow Fe_3O_4 + 3\,SO_2$$

Copper(I) sulfide is partially oxidized, but it is also reformed

$$Cu_2S + 1.5\,O_2 \longrightarrow Cu_2O + SO_2$$

$$Cu_2O + FeS \longrightarrow Cu_2S + FeO$$

In Figure 6, the first step corresponds to moving along the pseudobinary Cu_2S – FeS line to form an impure copper(I) sulfide.

Second Step. Continuing oxidation occurs as in a typical roasting reaction process:

$$\begin{array}{l} 2\,Cu_2S + 3\,O_2 \longrightarrow 2\,Cu_2O + 2\,SO_2 \\ 2\,Cu_2O + Cu_2S \longrightarrow 6\,Cu + SO_2 \\ \hline Cu_2S + O_2 \longrightarrow 2\,Cu + SO_2 \end{array}$$

In Figure 6 and Figure 20, the composition moves along the Cu_2S – Cu line from the copper(I) sulfide to crude metallic copper, the two phases being immiscible.

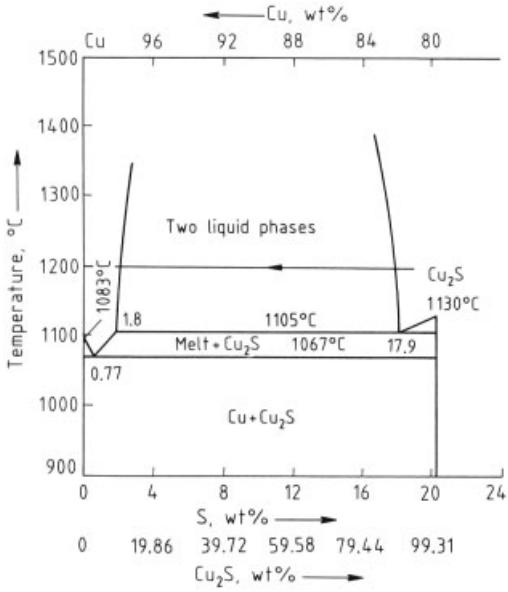

Figure 20. The Cu – Cu_2S system [89]

The blister copper contains < 0.1 wt % S, ca. 0.5 wt % O, and traces of other impurities.

Converter Slags. The slags from the first step are iron(II) silicates (40 – 50 wt % Fe) with high magnetite content (15 – 30 wt % Fe_3O_4). The initial copper content of 3 – 8 wt % can increase up to 15 wt % at the end of the reaction by formation of copper(I) oxide. This slag can

be decopperized by returning it to the smelting unit or by froth flotation (cf.page 28).

The high-viscosity small-volume converter slags from the second step have a high copper content (20 – 40 wt %) in the form of copper(I) oxide or silicate. When enough slag has accumulated, it is returned to the first converting stage.

Temperature. Converting is a strongly exothermic process that can overheat during oxidation of iron-rich mattes. The temperature must be held ca. 1200 °C by adding fluxes, copper scrap, precipitates from hydrometallurgical treatment (e.g., cement copper), or concentrates. The off-gas (5 – 10 vol % SO_2) is transferred to a sulfuric acid plant.

The blowing time per batch is a few hours; however, as the copper content of the matte increases, the converting time decreases. Occasionally, oxygen-enriched air is used to increase the throughput.

Impurities. The distribution of other elements among the phases during converting is as follows:

1) Noble metals and most of the nickel, cobalt, selenium, and tellurium collect in white metal and then in blister copper.
2) The bulk of zinc and some nickel and cobalt collect in converter slag.
3) The oxides and sulfates of arsenic, antimony, bismuth, tin, and the basic sulfates of lead are found in flue dust.

Converter Types. The copper converter was invented in 1880 by Manhès and David, based on the Bessemer converter, which had been used in the steel industry since 1855. This development led to the incorrect name "copper bessemerizing," although the true Bessemer process is a refining step. Originally, the copper converter was upright, and such obsolete units were in operation until the early 1980s, e.g., the Great Falls converter developed by Anaconda Mining Co., United States.

The following types are in use currently (Fig. 21):

1) The Peirce – Smith converter has been the most important apparatus for converting for many decades, and the number in operation may be in the range of a thousand. More than 80 % of the worldwide produced copper comes from this type of converter. It is a horizontal cylinder lined with basic bricks (magnesite, chrome – magnesite) that can be rotated about its long axis (Fig. 22); blast air is blown through a horizontal row of tuyeres. In practice, the punching of tuyeres with special devices is necessary to maintain the flow of air. The largest vessels are 12.5 m long with a diameter of 4.6 m.
2) Hoboken or syphon converter [91]. This variation of the P – S type was developed years ago by Metallurgie Hoboken N.V., Belgium, but is now used by only a few smelters in Europe and in North and South America; larger units are operated at Glogow smelter in Silesia, Poland; Cyprus Miami Smelter in Arizona; and Paipote smelter (ENAMI), Chile [92]. Its advantage is its freedom from sucking in air, so the off-gas can attain SO_2 levels up to 12 %. Special features of the design are the small converter mouth and the syphon or goose neck that guides the off-gas and flue dust flow.
3) Inspiration converter [93]. The vessel has two mouths, the smaller for charging, the larger for the off-gas. The latter is well hooded in all operating positions. It is only in operation at Cyprus Miami Smelter.
4) Top-blown rotary converter [95]. The TBRC, which is known in the steel industry as the Kaldo converter, was adopted by the nonferrous industry (first by INCO, Canada) because of its great flexibility. Air, oxygen-enriched air, or on occasion, commercial oxygen is blown through a suspended water-cooled lance onto the surface of a charge of copper-containing materials. In practice, the TBRC is used batchwise for special operations on a small scale, but generally not for converting copper matte. Tests of direct smelting of concentrates (clean, complex, or dirty) to white metal or blister copper were performed at Rönnskär smelter (Boliden Metall AB), Sweden, but the process was not realized technically. Copper extraction from copper scrap and other secondary materials is also carried out. The TBRC is also well suited for lead/precious metals metallurgy.

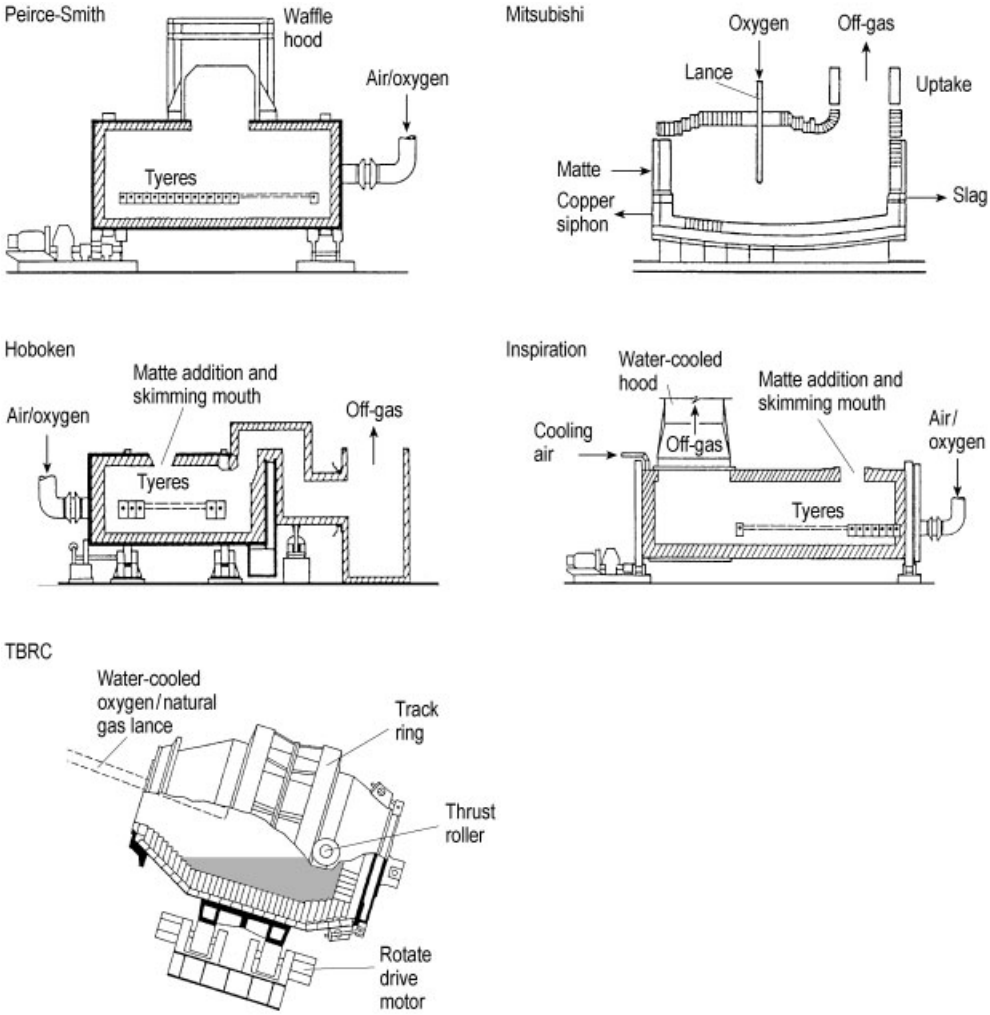

Figure 21. The evolution of the copper converter [90]

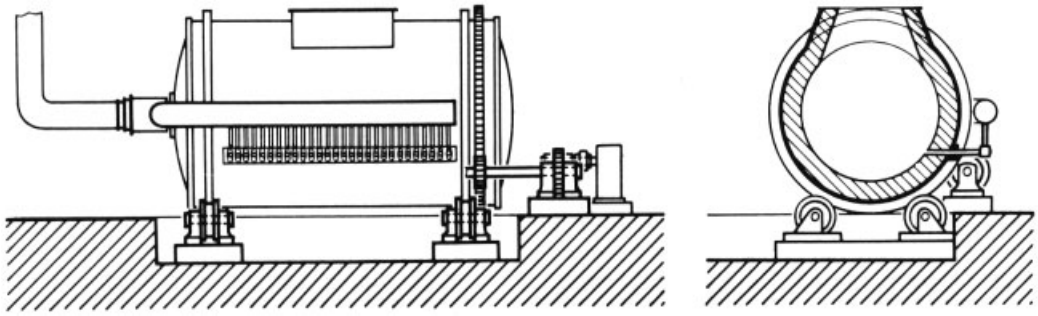

Figure 22. Schematic cross section and back view of a Peirce – Smith converter [96]

Developments for Increased Productivity and Environmental Protection To increase the productivity of a Peirce – Smith converter, usually the oxygen enrichment in the blast is increased. Depending on process step (slag or copper blowing) up to 30 % O_2 is used. The resulting increases in the temperature in the vessel must be controlled. Copper smelters mostly in Europe or Japan add copper scrap during the copper blow phase to adjust the temperature in the converter bath. Special systems (lift and conveyor) have been developed to add up to 70 % of the cooling material without stopping blowing [232], [233] (Figure 23).

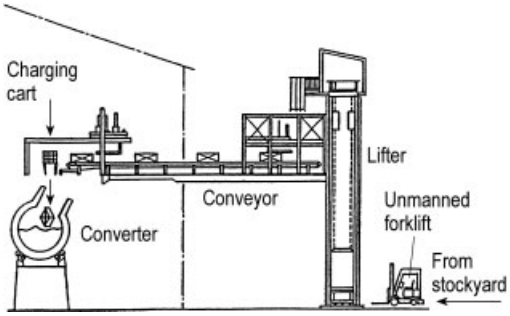

Figure 23. Machine for automatic charging of cooling material

Automatic temperature measurement (optical spectrometer or thermocouple through the tuyere) is used for process control and to avoid high erosion of the brick lining of the converter. Also the chemistry of the converting process is monitored by optical spectrometry [234]. This allows the blowing time to be easily controlled without overblowing the charge and thus minimizes the copper loss to the converter slag and magnetite formation in the slag. High-velocity tuyeres have been developed [235], [236] which can be operated with high oxygen enrichment without extensive refractory erosion. At the front of the tuyere a tubular accretion is formed which protects the brick lining. Another advantage is that these tuyeres need not be punched (saves labor and maintenance cost). A disadvantage is the relatively high pressure of about 3 bar of compressed air (high investment). These high-velocity tuyeres have been tested in Peirce – Smith and Hoboken converters. Today (ca. 2000) the first smelters (Hidalgo) are planing to install the system.

One of the problems of Peirce – Smith converters is leakage SO_2-containing off-gas during charging and pouring operations in the working environment. To avoid these fugitive emissions, special secondary hooding systems have been developed and installed. For example, Figure 24 shows the hooding system used at Norddeutsche Affinerie. The collected fugitive gases are dilute in SO_2 (up to 0.2 %) and are treated by various techniques (scrubbing with basic solutions or dry absorption of the SO_2 producing gypsum) [237], [238].

5.7. Continuous Matte Conversion

While continuous copper matte smelting processes have been in operation for many years, continuous converting of matte has come into use slowly. The potential benefits are minimization of materials handling (especially overhead crane transport of liquids), more efficient off-gas capture, and continuous SO_2 production for the sulfuric acid plant. Two multiple-furnace processes are used today for continuous smelting and converting: the Mitsubishi process and the Kennecott/Outokumpu Flash converting process. Noranda has developed a continuously running converter which has been in operation since 1997 at Horne Smelter.

5.7.1. Noranda Process

In the 1970s Noranda started with the Noranda reactor for directly smelting blister. This was not useful and was therefore switched to smelting high-grade matte. A second reactor similar to the smelting one is proposed for continuous conversion of matte (patented 1985 [239]). Since 1997 it has been operated at Horne smelter. It is a horizontal cylindrical vessel with two mouths (one for adding liquid matte by ladle, the other for the off-gas) and a row of submerged tuyeres. There is also the possibility to feed solid matte or coal by slinger belt through one end wall. The converter is usually fed with liquid high-grade matte (70 % Cu) from the Noranda smelting reactor. It produces semi-blister copper with high sulfur content (1 – 1.5 %). The semi-blister is poured into a ladle which is transported by special ladle car to the anode furnace. The Noranda converter can be operated in four

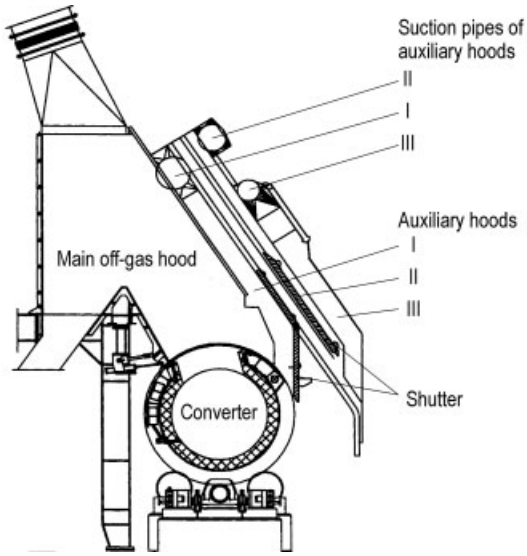

Figure 24. Hooding system for the prevention of fugitive emissions at Norddeutsche Affinerie

modes: with molten matte feed, with any combination of solid and molten matte feed, up to 100 % solid feed, in smelting mode with concentrate like the Noranda reactor and a conventional Peirce – Smith converter. The converter at Horne Smelter (Figure 25) [240] is 4.5 m in diameter and 19.8 m long (inside the brick lining). It has 42 tuyeres. The process off-gas is collected in a water-cooled hood and sent to the sulfuric acid plant. The converter has a secondary ventilation system, the gases of which are sent directly to a stack. This process operates continuously and has very high flexibility but not all of the potential benefits have been achieved. There is still crane transport and handling with ladles which causes fugitive emissions.

5.7.2. Mitsubishi Process

Mitsubishi Metals Corp., Japan, tested the new concept during the 1960s and started the first commercial plant at Naoshima smelter in 1974. Today (ca. 2000) four companies are operating this process (Naoshima, Japan; Kidd Creek, Canada; LG Metals, Korea; Gresik, Indonesia). The principle of this process is the interconnection of three furnaces, as shown in Figure 26 [241], [242]:

1) Smelting or S-furnace: Dried concentrates, flux material, pulverized coal, return converter slag, copper scrap, and flue dust are smelted in this furnace. The fine material (concentrate, coal) is fed through nine or ten vertical steel lances on the top of the molten bath. The blowing lances consist of two concentric pipes. Through the inner pipe, the dried concentrate is air blown, and through the outer one, oxygen-enriched blast (40 – 50 % O_2). The lances are rotated to prevent them sticking in the roof. They extend to 0.5 – 1 m above the molten bath. About 0.3 – 0.5 m of the lance is consumed each day. The off-gas contains 30 – 45 % SO_2 and is sent to sulfuric acid production.
2) Slag-cleaning furnace: Slag and matte from the S-furnace flow continuously by gravity into the elliptical slag-cleaning furnace, which is an electric furnace with three or six submerged graphite electrodes. The temperature of the slag is kept at 1250 °C. The slag is decopperized to 0.6 – 0.9 % Cu and discarded. The matte flows continuously through a siphon into the converting furnace.
3) Converting or C-furnace: The matte (68 % Cu) is converted continuously by blowing enriched air (30 – 35 % O_2) and $CaCO_3$ flux through six lances on the top of the bath. Also copper scrap is added through the sidewall of the furnace. Conversion only takes place where the oxygen comes into contact with the

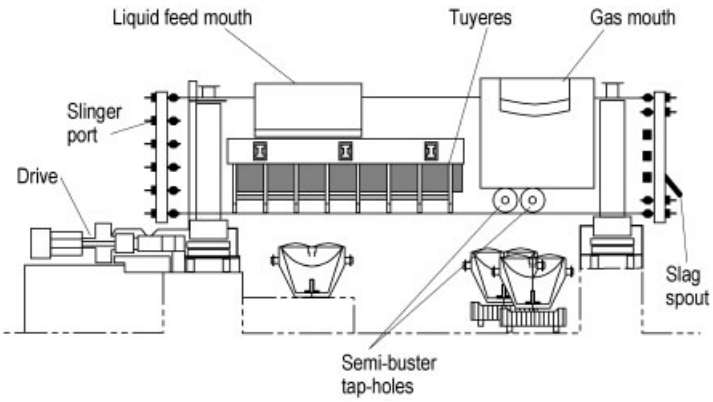

Figure 25. Noranda converter at Horne Smelter

sulfides. This is achieved by running the converter with a special basic calcium ferrite slag (40 – 50 % Fe, 15 – 20 % CaO). The usual silica slag is not possible because when oxygen is blown onto the top, solid magnetite is formed and blocks the surface. In the calcium ferrite slag dissolves the magnetite but also a large amount of copper (15 – 18 %). The converter slag is returned to the S-furnace. The blister copper contains slightly more sulfur (0.7 %) than from a Peirce – Smith converter (0.02 %). It is sent to an anode furnace.

Some operating data for the Mitsubishi process (Naoshima Smelter, Japan) are summarized in the following:

Commisioning date	1991
S-furnace	
Diameter inside brick, m	10
Number of lances	9
Concentrate throughput, t/d	2050
Copper scrap, t/d	20
Blast, m^3/h (STP)	37 500
O_2 enrichment, %	49
Off-gas volume, m^3/h (STP)	39 000
SO_2 in off-gas, %	29
Matte grade, %	68
Slag-cleaning furnace	
Size, inside brick $w \times l \times h$, m	$6 \times 12.5 \times 2$
Number of electrodes	6
Residence time	ca. 2 h
Slag	0.6 % Cu
Off-gas volume, m^3/h (STP)	50
C-furnace	
Diameter inside brick, m	8
Number of lances	10
oxygen enrichment, %	33
Blast flow rate, m^3/h (STP)	25 500
Copper Scrap, t/d	190
Off-gas volume, m^3/h (STP)	25 000
SO_2 in off-gas, %	26

The major advantage of the Mitsubishi process are the good SO_2 capture and the lower handling expense of materials. A disadvantage during the initial operating time was that no scrap material could be added. This is has now been solved [243]. With a smelting capacity of about 2000 t/d of copper concentrate, Naoshima Smelter processes also about 40 000 t/a of copper scrap from the market and additional anode scrap from the refinery. Another problem is the impurity behavior. Especially lead is a problem because of the calcium ferrite slag. If too much lead is in the feed, the copper anodes are too rich in lead for electrorefining. A comparison of the impurity behavior is given in Table 14.

Table 14. Impurity contributions to anodes (wt % of input)

Process	Reverbatory/Peirce – Smith	Mitsubishi
Pb	8 – 13	15 – 19
As	9 – 11	4 – 6
Sb	28 – 34	15 – 20
Bi	14 – 26	15 – 25
Ni	45 – 50	67 – 75

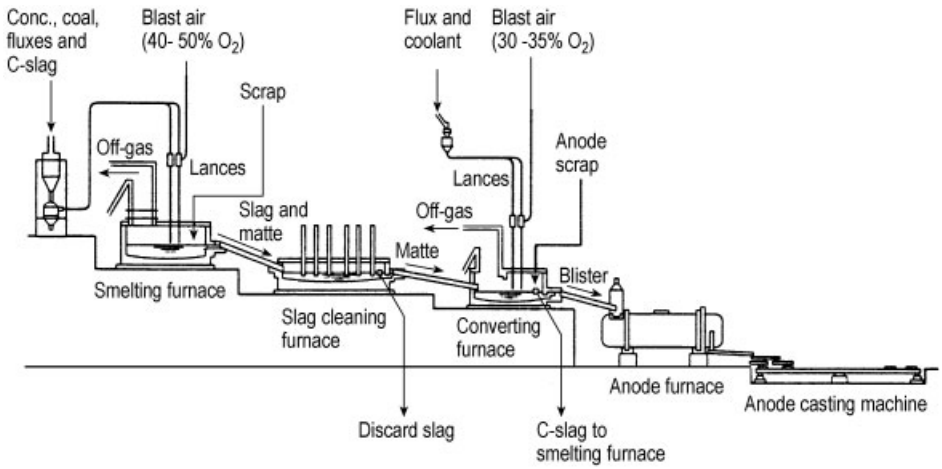

Figure 26. Mitsubishi process

5.7.3. Kennecott/Outokumpu Flash Converting Process

In the 1980s Kennecott announced the development of their Solid Matte Converting (SMOC) process [244]. This process is based on solidifying the molten matte from the smelter, grinding, and feeding the solid to an high-oxygen blast converter. This process decouples the smelting and converting steps and gives great flexibility. Later on, it was developed with Outokumpu, and for the converting step also a well-proven flash furnace was chosen. In 1995 the process went into operation at Garfield Smelter.

For matte smelting a flash furnace with a concentrate feed rate of about 140 t/h was built (oxygen enrichment 65 – 75 %). The liquid matte contains 68 – 70 % copper and is granulated in water. In the second step the matte is milled and then fed together with lime as flux material to the flash converter, which operates with an oxygen enrichment of 65 – 75 % O_2. The feed rate of matte is about 60 t/h (up to max. 80 t/h) [245]. This furnace is constructed like the smelting furnace. A blister copper and a calcium ferrite slag with about 18 % copper (like in the Mitsubishi process) is obtained. The slag is granulated and fed back to the flash smelter.

The off-gases contain >30 % SO_2 and are converted to sulfuric acid (input concentration of 14 % SO_2 in the sulfuric acid plant). This process has very low SO_2 emissions of only 3.5 kg SO_2 per tonne of copper, which is until now the lowest in the world. This is achieved because it has almost no fugitive emissions. The disadvantage is that no scrap can be added to the converter.

5.8. Direct Blister Smelting

Normally copper extraction takes place in two steps – concentrate smelting and matte converting – in which the chemical reactions are principally the same: oxidation of Fe and S. It has long been the aim to combine these processes in a single step/reactor and minimize the energy consumption and operating cost. Today (ca. 2000) direct blister smelting is carried out at two smelters (Glogow, Poland and Olympic Dam, Australia) in a Outokumpu flash furnace. Between 1973 and 1975, Noranda also used their reactor for blister smelting from concentrate, but it was switched to matte smelting because of impurity problems. Another proposal is the QS process.

5.8.1. Blister Flash Smelting [246]

This process takes place in a conventional Outokumpu flash smelting furnace. Concentrate and silica flux are fed with 60 – 85 % O_2 blast to the burner. The amount of O_2 is just enough to form blister copper. In practise, the particles in the reaction shaft are somewhat over-oxidized on the outside and incompletely oxidized on the inside. In the molten bath layers the reaction is

completed to form blister copper and iron silica slag. Formation of a layer of matte between blister and slag must be avoided, because otherwise a pronounced foaming reaction could take place.

In both operations copper concentrates with low iron content (e.g., chalcocite, bornite) are processed to minimize slag production and copper loss in the slag. The slag has a high copper content (15 – 21 % Cu), mainly in oxidized form, and has to be reduced in an electric furnace (produces copper – iron – lead alloy [247]) or concentrated by grinding and flotation. Reduction in an electric furnace with addition of coke takes about 12 h residence time. The furnace is operated at higher temperatures than the normal flash smelter (slag temperature 1300 °C) to obtain higher solubility of magnetite in the slag and decrease the viscosity of the slag. The blister copper contains 0.3 – 1 % S. At the Glogow Smelter, the process has been in operation since 1977 with a daily throughput of 1500 t of concentrate. At Olympic Dam it has been operated since 1988 and was enlarged in 1998.

5.8.2. QS Process

The QS process, based on the QSL leadmaking process, was invented by Queneau and Schuhmann in 1974 [109]. Blister copper is produced from concentrate in a single vessel. The reactor is similar to the Noranda reactor but has countercurrent flow of slag and matte. Gases are injected from the bottom [248] of the vessel to oxidize Fe and S and to reduce the copper content in the slag (Figure 27). Although the process has fundamental and technical merits, it has not been realized in a commercial plant.

5.9. Copper Recycling

Most copper (> 95 %) is used in metallic form, as copper metal or copper alloys. Recycling of copper and its alloys has been carried out since ancient times. In 1997, 37 % of copper consumption came from recycled copper. This low figure could be explained by the long lifetime of copper products (e.g., more than 30 years for copper wire) and the much increased copper production in the last 50 years. Virtually all products made from copper can be recycled, and copper can be recycled over and over again without losing its chemical or physical properties.

The processes used for copper recycling depend on the copper content of the secondary raw material, its size distribution, and other constituents (Table 15) [249], [250]. Three general types can be defined:

Type 1: copper scrap (new and old scrap), used for smelting and refining or direct smelting for products. This type accounts for about 95 % of all recycled copper. The value of the metal is much higher than its treatment costs.

Type 2: copper-containing special scrap such as cables and printed circuit boards. Pretreatment (e.g., cable comminution) is necessary before smelting the copper. Another example is copper in cars. Each car contains 10 – 30 kg of Cu which must be separated from the iron. The value of the material is in the range of the overall treatment cost.

Type 3: copper-containing residues, for example sludges from metal-plating industry. The copper content of these materials is low. The value of the copper is much lower than the treatment cost of the material.

High-purity secondary copper material is used for smelting and casting new products, and no fire and/or electrolytic refining is necessary. This is also true of pure alloy scrap, which is used for fresh alloy. With increasing impurity content, the material is fed to smelting and refining processes. The process steps and aggregates of secondary smelters are generally similar to those of primary production, but the secondary raw material is usually metallic or oxidic [253]. Therefore, mainly reducing conditions are used for secondary smelting. Often carbon in the form of coke or natural gas is used.

For smelting metallic scrap, often shaft furnaces (Asarco type or Contimelt) are used. For smelting oxidic materials, blast furnaces, electric furnaces, TBRCs, and Isasmelt furnaces are used. In the TBRC both the smelting and the converting step could be done. This saves energy in comparison with the blast furnace and Peirce – Smith converter. In some cases electric arc furnaces are used instead of the blast furnace. The advantage is that this furnace produces about 80 % less off-gas than the blast furnace. This makes the capture easier and more efficient.

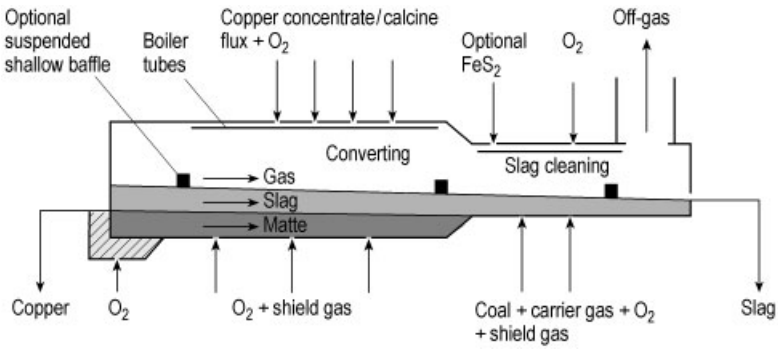

Figure 27. QS process

Table 15. Overview of different sorts of secondary copper sources

Type	Material	Cu content, %	Source	Recycling process
1	mixed copper scrap	90 – 95	sheets, gutters, water boilers, heaters, wires, pipes	converting, fire refining
1	pure copper scrap	99	semi-finished products, wire, strip, cuttings	melting and casting of semifinished products
1	red brass scrap	75 – 85	valves, taps, machine components, fittings, bearing boxes, car radiators	converting or alloy production
2	shredder material	60 – 65	cars	blast furnace, electric furnace, TBRC
2	cable	40 – 50	buildings, cars	shredder, firerefining
2	electronic scrap	10 – 20	electronics	converting, TBRC
3	sludges	5 – 10	electroplating	blast furnace, Electric furnace
3	copper – iron material	10 – 20	armatures, stators, rotors	blast furnace, electric furnace, TBRC
3	drosses, ashes, slags	20 – 25	foundries	blast furnace, electric furnace, TBRC

A typical flow sheet of a secondary smelter is shown in Figure 28.

Many primary smelters treat scrap besides their internal recycling materials like anode scrap from the tankhouse or converter flue dusts. Additional copper scrap bundles are fed to the Peirce – Smith matte converter. The scrap cools the converters and uses the energy from matte oxidation for heating and melting.

Also electronic scrap such as printed circuit boards is fed to copper matte converters (e.g., Norddeutsche Affinerie [251], [252]) or Noranda reactors (e.g., Noranda). The copper and the precious metals are captured and sent to subsequent refining. The organic polymers are burnt at the high temperatures, and water and CO_2 are formed; ceramics are slagged. Because of the high temperature and the high SO_2 content in the off-gas, no dioxins are formed.

5.10. Hydrometallurgical Extraction

About 15 % of the primary copper production in the western world, that is, about 1.8×10^6 t/a, results from sulfuric acid leaching of copper ore, which is combined with solvent extraction and electrowinning. Other proposed techniques like ammonia leaching or hydrochloric acid leaching have minor importance. They are mainly proposed for treating concentrates. A typical flow sheet of a sulfuric acid leaching operation is shown in Figure 29.

Chemistry Of Sulfuric Acid Leaching Processes. There are two chemically different classes of copper minerals:

1) Oxidic or secondary minerals like malachite and chrysocolla
2) Sulfidic or primary minerals like chalcopyrite and chalcocite

The sulfidic minerals are the most common in an ore body. Oxidic minerals are often located on the top of the ore body and are present in smaller amounts.

Chemical reactions during extraction of copper from primary and secondary minerals are as follows:

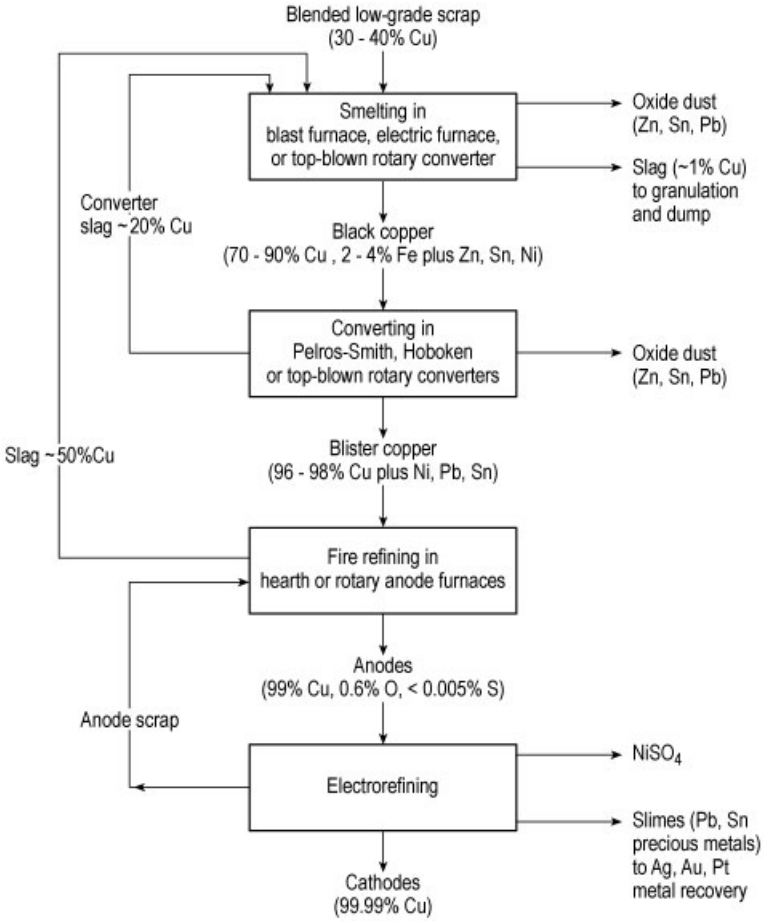

Figure 28. Flow sheet of a secondary smelter

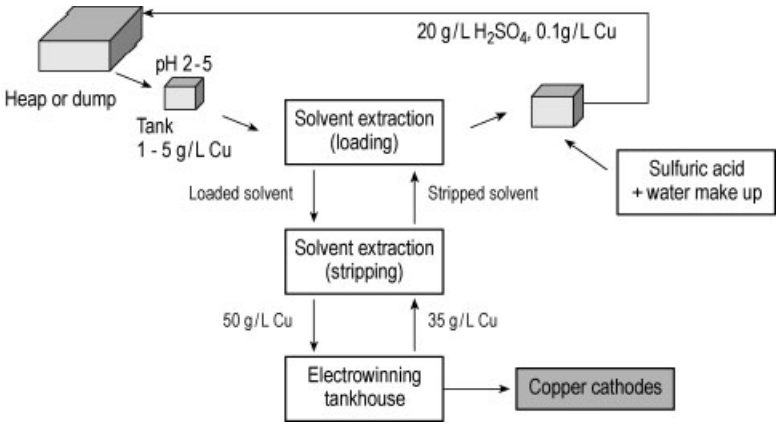

Figure 29. Sulfuric acid leaching of copper ores

Oxidic secondary copper minerals:

$$CuCO_3 \cdot Cu(OH)_2 + 2\,H_2SO_4 \longrightarrow 2\,CuSO_4 + 3\,H_2O + CO_2$$

Sulfidic primary copper minerals:

First step: generation of Fe^{3+}, which is the oxidant for copper leaching, by bacteria.

Thiobacillus thiooxidans:

$$2\,FeS_2 + 7\,O_2 + 2\,H_2O \longrightarrow 2\,FeSO_4 + 2\,H_2SO_4$$

Thiobacillus ferrooxidans:

$$2\,FeSO_4 + 0.5\,O_2 + H_2SO_4 \longrightarrow 2\,Fe^{3+} + 3\,SO_4^{2-} + H_2O$$

Second step: attacking the chalcopyrite:

$$2\,Fe_2(SO_4)_3 + CuFeS_2 + 3\,O_2 + H_2O \longrightarrow 5\,FeSO_4 + CuSO_4 + 2\,H_2SO_4$$

Secondary minerals are readily soluble in sulfuric acid, and reaction times are typically short. After several hours most of the copper is extracted, but sometimes the ore requires a few weeks reaction time. Recovery of copper from primary minerals is more difficult because the sulfides are very stable, and apart from sulfuric acid an oxidant is required as well. The most important oxidant is Fe^{3+}, which is generated from the mineral chalcopyrite by bacteria-assisted reactions with atmospheric oxygen. Although these bacteria, which normally exist in the mine water, speed up the process, very long reaction times like months or years are still necessary to extract most of the copper. The consumption of sulfuric acid depends on the type of copper mineral. The bacterial leaching process produces sulfuric acid itself.

Industrial Leaching Processes. The copper leaching techniques in practical use are listed in Table 16.

The most important technique is heap leaching. It is very common for oxidic copper ore. The typical copper content is between 0.25 and 1 %. The ore is crushed to 1 – 10 cm and stacked to heaps 3 – 10 m high with an area of about 10 000 to 100 000 m^2. Dilute sulfuric acid is sprayed on the top of the heap. The liquor trickles through the material, copper is dissolved, and the copper-bearing solution (1 – 5 g/L Cu) is collected at the bottom of the heap. This copper solution is purified and enriched by solvent extraction. Copper metal is generated by electrolysis of the enriched solution. After several months (1 – 6), leaching is stopped and a new heap is placed on the top of the previous and leaching is begun again.

Data from a typical heap leaching plant are shown in Table 17. To produce 54 000 t/a of copper, 27 000 t ore per day must be treated.

Table 17. Typical data of heap and dump leaching operation

	Heap leaching	Dump leaching
Cu cathode production, t/a	54 000	6600
Leach ore composition	chrysocolla	95 % chalcopyrite, 3 % chalcocite
Soluble Cu in ore, %	0.54	0.15
Total area under leach, m^2	200 000	300 000
Tonnes of ore per day	27 000	55 000
Annual Cu recovery rate, %	90	20
Sulfuric acid consumption, t/t Cu	1.7	0

The copper production of a typical dump-leaching operation is much lower, for example. 6600 t/a. The feed material is very low grade sulfidic ore, and several years are required to extract all the copper. The sulfuric acid consumption depends on the mineralization of the copper ore and the composition of the host rock. Typical figures are in the range of 1 – 5 t of sulfuric acid per tonne of copper product.

Minor methods are in situ leaching, tailings leaching, and vat leaching. Vat leaching was carried out in the past to produce a pregnant leach solution with about 30 g/L Cu which was directly suitable for electrowinning. This technique is replaced by heap leaching and SX. Agitation leaching is used mainly to leach oxidic Cu concentrates in Zaire and Zambia. Fine particles are treated with strong sulfuric acid (ca. 60 g/L) solution, and the reaction is complete after several hours.

Other Leaching Processes. Mainly for the leaching of sulfidic copper concentrates, many processes have been developed.

The *Arbiter/Escondida Process* was developed for leaching mainly chalcocite concentrates. In ammonia solution, half of the copper is extracted very quickly and CuS remains in the residue.

$$Cu_2S + 2\,NH_3 + 2\,NH_4^+ + 0.5\,O_2 \longrightarrow CuS + [Cu(NH_3)_4]^{2+} + H_2O$$

Table 16. Common techniques for sulfuric acid leaching of copper mine

Leaching technique	Mineralization	% Cu in ore	Leaching time	Estimated Cu production (1998), T/A
Heap	oxides, chalcocite	0.25 – 1	several months to years	700 000
Dump	chalcopyrite	< 0.25	1 – 5 decades	50 000
In Situ	all	> 0.5	decades	50 000
Vat	oxides	1 – 2	5 – 10 d	50 000
Agitation	oxides (carbonates)	20 – 40	2 – 5 h	200 000
Tailings	oxides (carbonates)	0.25 – 1	1 d	100 000

The reaction takes place in agitated vessels. The copper complex solution is sent to a solvent extraction and electrowinning plant. The residue is sent to a flotation plant, where copper is enriched, and then fed to a smelter. This process was operated at Escondida mine in Chile but at present is shut down.

The *Cuprex process* uses iron(III) chloride for the oxidation of copper from chalcopyrite:

$$CuFeS_2 + 4\,FeCl_3 \longrightarrow CuCl_2 + 5\,FeCl_2 + 2\,S$$

A solvent extraction step gives pure $CuCl_2$ solution (ca. 100 g/L Cu), which is fed to a membrane-divided electrowinning cell, where copper powder is produced and in at the anode the iron(II) chloride is oxidized to iron(III) chloride and recycled to the leaching process. The process has been tested in a 1 t/d pilot plant.

The *Intec process* is carried out in a NaCl/NaBr solution at 80 – 85 °C with $CuCl_2$ as oxidant. Leaching is carried out in four steps; gold is also leached in the last step. The residue consists of Fe(O)OH and elemental sulfur.

$$4\,CuFeS_2 + 4\,CuCl_2 + 3\,O_2 + 2\,H_2O \longrightarrow 8\,CuCl + 4\,FeOOH + 8\,S$$

The solution is cleaned by adding CaO and then sent to a membrane electrolytic cell. At the cathode copper is reduced to copper powder. At the anode two reactions take place: oxidation of Cu^+ to Cu^{2+} and oxidation of bromide ions to $BrCl^-$. This solution, which contains about 30 g/L Cu, is sent back to the leaching reactor (step 4, Au leaching).

Oxygen pressure leaching of chalcopyrite, bornite, or chalcocite concentrates in sulfuric acid under 8 bar oxygen pressure at 200 °C gives about 99 % copper recovery after a short reaction time of only one hour. The dissolved copper can be recovered by electrowinning. The leach residue can be treated with cyanide for gold recovery. This process was tested in laboratory scale. It might find use for impure concentrates or concentrates high in precious metals.

Bacterial leaching of chalcopyrite concentrates (Bio COP) is developed since 1997 at Codelco's Chuquicamata mine. The sulfide concentrate is leached in stirred tanks at 65 – 85 °C assisted by special heat-resistant bacteria (bacteria mesofila). The process needs large amounts of air/oxygen. The copper sulfate solution is fed to a SX-plant and then to copper electrowinning. A pilot plant will be built within the next 2 years.

Solvent Extraction. The pregnant leach solutions from heap, dump, or in situ leaching are too dilute in Cu (1 – 5 g/L) and too impure for direct production of high-grade copper cathodes. Industrial copper winning requires electrolytes with about 40 – 50 g/L copper. Copper enrichment and purification, mainly to remove iron, is done by solvent extraction. The dilute, impure solution from the leaching operation is mixed with an organic solution, which is based on kerosene or petroleum and contains about 5 – 10 % of an extractant which is selective for copper. The organic extractants used for the process are salicylaldoximes and ketoximes (RH) and commercially available as LIX (Henkel) or MOC (Allied Signal). The copper ions are complexed by these reagents and enter the organic phase:

$$2\,RH + CuSO_4 \longrightarrow R_2Cu + H_2SO_4$$

After mixing, the aqueous and organic phases are separated by gravity (settler). In a second step, copper is stripped from the organic phase with more highly concentrated sulfuric acid solution.

$$H_2SO_4 + R_2Cu \longrightarrow 2\,RH + CuSO_4$$

Normally, the depleted solution from electrowinning, which contains about 170 g/L H_2SO_4 and 35 g/l Cu, is used. It is enriched with copper to about 50 g/L and sent to the electrowinning cells. Depending on the composition of the dilute solution from the leaching plant several mixer/settler steps are needed for purification of the copper.

Copper Electrowinning. The copper-enriched sulfuric acid solution (ca. 40 – 50 g/L Cu, 140 – 170 g/L H_2SO_4) is sent to the electrowinning cells, which are similar in construction to refining copper cells. Copper is depleted at the cathode on stainless steel sheet (ISA Process) or copper starters. The cathode reaction is identical to the refining reaction:

$$Cu^{2+} + 2e^- \longrightarrow Cu \quad E_0 = 0.34\,V$$

The anode reaction is completely different. Inert anodes, mainly PbSnCa alloy, are used, and oxygen is formed at the anode.

$$H_2O \longrightarrow 0.5\,O_2 + 2\,H^+ + 2e^- \quad E_0 = 1.23\,V$$

The electrical potential needed for copper electrowinning is about 2 V, as opposed to about 0.3 V for copper electrorefining. About 2000 kW h is required for depleting 1 t of copper. The usual plating period is about 7 d, after which around one-third of the cathodes are taken out of the cell by crane and immediately replaced by fresh ones, in contrast to the refining tankhouse, where all cathodes are removed. The cathode replacement in the winning tankhouse is performed without cutting off the electrical power. This maintains a passivation layer on the lead anode and minimizes contamination of the copper with lead. In addition, the electrolyte contains about 100 ppm cobalt to prevent corrosion of the anodes. The lead anodes have lifetimes of up to 20 years. The current density in electrowinning plants is about 200 – 300 A/m^2. Modern electrowinning tankhouses have capacities up to 200 000 t/a of cathode copper.

Economics of Copper Production by L/SX/EW Processes Compared with copper production in smelters, L/SX/EW processes have lower production figures. But the arguments for the L/SX/EW techniques are low investment costs and increased copper output efficiency of mines. Furthermore, copper can be recovered from low-grade materials which are unsuitable for smelters. With improvements in solvent extraction and electrowinning technology, the quality of the recovered copper has become as good as that from classical refineries. Leaching facilities are mostly located at the copper mines. The disadvantage of leaching is that it does not recover the precious metals.

Production figures (in 10^3 t/a) for L/SX/EW copper by region in 1997 follow:

North America	567
Latin America	943
Oceania	49
Africa	65
Asia	4
Western Europe	2

The expected growth of western world copper production and the contributions from smelters and L/SX/EW operations is shown in Figure 30. Copper production is expected to increase from about 12×10^6 t/a in 1996 to about 14×10^6 t/a in the year 2000. The biggest part of this growth will come from L/SX/EW operations. Smelter production will also increase, but there the next step in capacity enlargement is expected after 2002.

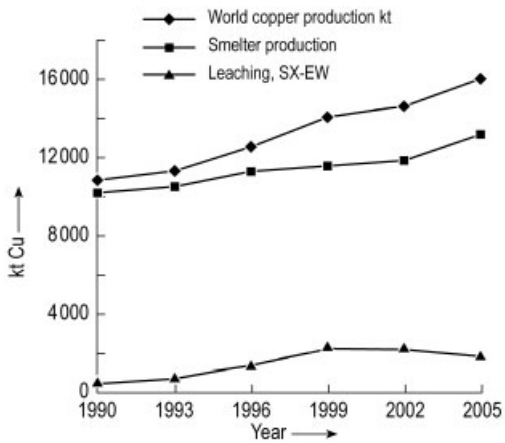

Figure 30. World copper production statistics

Most of the increase in copper production by leaching was accounted for by Latin America (Figure 31). By 2002 or 2003, the production of L/SX/EW copper is expected to have increased by nearly a factor of three. Several new operations are scheduled for the near future.

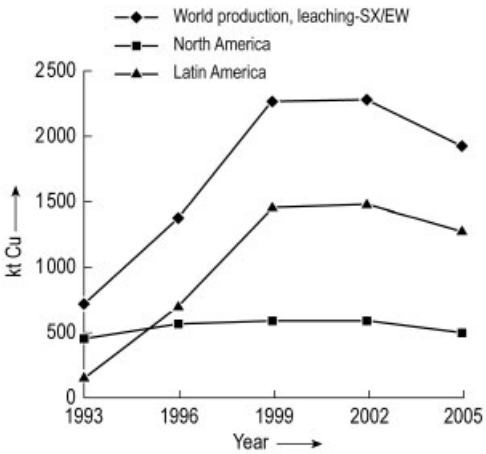

Figure 31. Copper production by the L/SX/EW technique

6. Refining

Conventional refining comprises three stages: (1) pyrometallurgical or fire refining, (2) electrolytic refining, and (3) remelting of cathodes and casting of shapes. Refining without electrolysis is adequate if the fire-refined copper has the necessary purity and if the content of precious metals can be neglected. If extremely high-purity copper is needed, zone melting or repeated electrolysis of cathodes is used.

6.1. Pyrometallurgical Refining

Fire refining is applied to crude copper such as blister copper from converters (ca. 97–99 wt % Cu), black copper from blast furnaces (ca. 90–95 wt % Cu), cement copper from hydrometallurgical operations (ca. 85–90 wt % Cu), anode scrap from electrolytic refining, and high-grade copper scrap, chiefly unalloyed wire scrap.

The refining of molten copper to anode copper for electrolysis or commercial fire-refined copper has the following functions:

1) Removing impurities by slagging and volatilization, with the precious metals remaining entirely in the metallic copper
2) Reducing the sulfur content to ca. 0.0005–0.005 wt % by oxidation
3) Decreasing the oxygen content to $< 0.1-0.25$ % by reduction (poling) to give a flat surface as a result of the water-gas equilibrium in molten copper

6.1.1. Discontinuous Fire Refining

Two furnace types are available for batch copper refining, the older *reverberatory furnace* and the more modern *rotary furnace.* The former, which resembles smaller reverbs for matte smelting from concentrates, has a capacity of 200–400 t of copper per charge, can be fed with molten or solid copper, and is used in secondary smelters. Rotary furnaces hold up to 350 t of molten metal per charge and are generally fed only with liquid copper. This type of furnace is preferred by primary smelters because the reduction is more efficient. An extra melting furnace, e.g., anode shaft type (Section 6.1.2), can be required for remelting of solid materials (scrap and anode rests).

Low-sulfur pulverized coal, fuel oil, natural gas, or reformed natural gas serve as fuel. The refractory lining consists of basic bricks, such as magnesite or the more spall-resistant chrome–magnesite bricks.

After charging and possibly melting, oxidation and reduction stages are carried out in sequence. At the beginning of the *oxidation period,* air is blown into the melt, partly slagging and partly volatilizing the impurities. During this blowing step, a part of the copper is oxidized to copper (I) oxide, which dissolves in the liquid metal (Fig. 32). If the content of Cu_2O in copper increases to ca. 10 wt % (corresponding to $\geqq 1$ wt % O), it acts as a selective oxidizing agent:

$$Cu_2S + 2\,Cu_2O \longrightarrow 6\,Cu + SO_2$$

In practice, large amounts of SO_2 are generated so that this final stage of the oxidation period is termed boiling. Reduction of the sulfur content limits SO_2 blisters in the solid copper.

The subsequent *poling* with wood is a centuries-old method that is still employed in older reverb plants. Large tree trunks (poles) of beech, birch, eucalyptus, etc. are plunged under the surface of the melt to generate reducing gases and steam by dry distillation of the wood. The escaping gas mixture reacts with copper (I)

oxide and mixes the molten bath [1, p. 441], [2, p. 392].

This awkward operation has been largely displaced by *gas poling* [143] in rotary furnaces: natural gas (CH_4), reformed gas (CO, H_2, and N_2), propane, or ammonia [144] is blown into the copper melt through tuyeres. This process has been introduced at most smelters in the world.

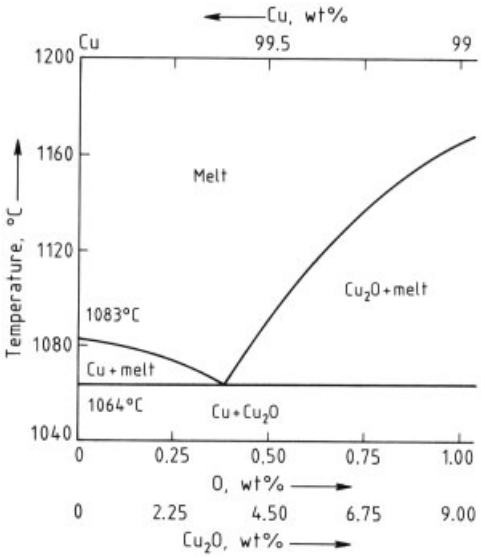

Figure 32. Partial phase diagram of the $Cu-Cu_2O$ system

The poling operation proceeds in two steps. During *tight poling,* the remaining sulfur dioxide from copper (I) sulfide is almost entirely flushed out by the escaping gas, a sample of liquid copper at the end of this stage solidifying without blisters or cavities. Next comes the *poling tough pitch,* which is necessary to reduce the copper (I) oxide and achieve the required low oxygen content. High Cu_2O content in the solidified metal causes brittleness and decreased strength; moreover, Cu_2O disproportionates to copper metal and copper(II) ions in sulfuric acid electrolytes, which disturbs the electrolytic refining operation.

The final oxygen content in fire-refined *tough pitch copper* must be 0.02 – 0.05 wt %; for anodes it can be 0.05 – 0.3 wt %. The fire-refining process can be understood from the Cu – O system [145] (see Fig. 32); the system Cu – O – S is important in the oxidation period (see Fig. 5) and the Cu – H – O system for the reduction or poling period.

When a copper melt solidifies, a shrinkage of ca. 5 vol % occurs, but a flat surface can be achieved by careful control of the equilibrium

$$Cu_2O + H_2 \rightleftharpoons 2\,Cu + H_2O$$

The steam of micropores can compensate the volume difference, and a *flat set* cast is obtained.

Surface and fracture of small samples of solidified copper from the molten bath are observed during the refining process to ascertain the current state. It is also possible to measure the oxygen content of the copper melt potentiometrically.

A special problem is the extremely high copper content (up to ca. 40 wt % Cu, chiefly as Cu_2O) in refining slags. Such products are treated as high-grade oxidized copper ores.

6.1.2. Continuous Fire Refining

The two-stage Contimelt process for copper melting and refining was developed in 1968 by Norddeutsche Affinerie at Hamburg, in cooperation with Metallurgie Hoboken-Overpelt in Olen, Belgium [147]. The first stage began operation in 1979, and the complete process has been operated since 1982 on a commercial scale. The continuous operations are performed successively in two units connected by launders. First is the anode shaft furnace, where charging, melting, and oxidation of crude copper take place. Second is the small drum-type furnace, where poling and casting of anodes are carried out. The oxygen content of copper from the anode shaft furnace averages 0.6 %, with 0.15 % after poling. A feature of the shaft furnace is the additional equipment with oxygen burners for regulating the composition of the furnace atmosphere and the overheating of copper. In comparison with conventional fire refining, the Contimelt process provides savings in energy and labor.

6.1.3. Casting of Anodes

The conventional method of producing anodes is the discontinuous casting on casting wheel machines. The pure copper molds must be sprayed

with a slurry that prevents the sticking of solidified anodes; baryte, alumina, or silica flour are suitable. (Calcium-containing material is not suitable because it forms gypsum, which is partially soluble in the electrolyte.) The casting rate can reach 100 t/h. The anode weights vary between 250 and 450 kg, depending on the refinery. Anodes from modern plants usually have the following dimensions: 0.9 – 1.1 m long; 0.9 – 1.0 m wide; and 3.5 – 6.0 cm thick. They weigh 300 – 450 kg. Economic considerations call for anodes of nearly the same weight within a plant; therefore, discontinuous casting is best controlled by electronic systems.

Contilanod Process. The Contilanod process, developed by Metallurgie Hoboken-Overpelt in Olen, Belgium, produces uniform anodes by using the continuous Hazelett twin-belt casting system [148]. The continuous cast strip of anode copper formed between two belts and damblock chains is 1 m wide and 1.5 – 6 cm thick; special cutting equipment separates the strip into individual anodes 1 m long. Some refining plants use this method, e.g., White Pine, Michigan, United States; Kidd Creek, Timmins, Canada. The advantage of this system is the uniformity of the anodes and the high degree of automation. But in comparison to the mold-on-wheel technique, the Hazelett caster has somewhat higher maintenance costs.

6.2. Electrolytic Refining

About 80 % of the world copper production is refined by electrolysis. This includes all primary copper and much secondary copper. This treatment yields copper with high electrical conductivity and provides for separation of valuable impurities, especially the precious metals.

The basic patent, GB 2838, for galvanic deposition of metals was awarded to J. B. Elkington in 1865. The most important technical problems were solved by E. Wohlwill at the Norddeutsche Affinerie in Hamburg, Germany, in 1876, and this method has been used ever since. The first electrolytic copper refinery in the United States was operated from 1883 to 1918 by the Balbach Smelting and Refining Co., Newark, New Jersey.

6.2.1. Principles [149], [150]

Several possible half-reactions can occur at the electrodes.

Anode reactions	Cathode reactions	Standard electrode potential E° (25 °C), V
$Cu \longrightarrow Cu^{2+} + 2e^-$	$Cu^{2+} + 2e^- \longrightarrow Cu$	0.337
$Cu \longrightarrow Cu^{+} + e^-$	$Cu^{+} + e^- \longrightarrow Cu$	0.521
$Cu^{+} \longrightarrow Cu^{2+} + e^-$	$Cu^{2+} + e^- \longrightarrow Cu^{+}$	0.153

Secondary reactions occur in the electrolyte:

$$2\,Cu^{+} \longrightarrow Cu^{2+} + Cu \text{ (disproportionation)}$$

$$2\,Cu^{+} + 2\,H^{+} + 0.5\,O_2 \longrightarrow 2\,Cu^{2+} + H_2O \text{ (air oxidation)}$$

$$Cu_2O + 2\,H^{+} \longrightarrow 2\,Cu^{+} + H_2O \text{ (dissolution of } Cu_2O)$$

Oxidation by air and disproportionation of copper (I) ions yield a surplus of copper (II) ions in the electrolyte. The copper metal powder formed by the disproportionation of Cu^{+} contributes to the accumulation of the anode slime.

The electrochemical equivalent of copper depends on the oxidation state of the copper:

Species	$g\,A^{-1}\,h^{-1}$	mg/L
Cu^{2+}	1.185	0.3294
Cu^{+}	2.371	0.6588

The greater electrochemical equivalent of copper (I) suggests the use of solutions of copper (I) ions instead of copper (II) ions. However, this concept has not been put into practice because of enormous industrial difficulties [151].

The two most important electrical parameters in electrolytic copper refining operations are the cell voltage and the current density. The cell voltage, which usually ranges between 0.25 and 0.3 V, is determined by several factors:

1) Ohmic resistance of the electrolyte, depending on composition, temperature, electrode distance, and cell construction
2) Polarization, especially concentration polarization of electrodes, which depends on the rate of electrolyte circulation

3) Overpotential because of organic additives (e.g., an *inhibitor* for achieving uniform electrocrystallization of copper)
4) Voltage loss in the circuit
5) Anode passivity, which may occur at high current densities

The interaction of these effects is difficult to predict. At any particular electrolytic facility, a continuing effort is made to optimize parameters that affect the cell voltage. The voltage loss in the conductors and contacts is minimized by good plant design and use of special contacts (Baltimore grooves, Whitehead contacts, wet contacts, etc.).

The second important parameter is the cathodic current density. With increasing current density the production of copper increases and the current efficiency decreases because the cathode potential depends on the current density.

Impurities. The behavior of impurities depends on their position in the electrochemical series: elements more noble than copper are insoluble, while less noble ones dissolve in or react with the electrolyte. For that reason, the anode material is distributed by electrolysis among three phases: cathode copper, electrolyte, and anode slime (Table 18) [152]. Table 19 lists the fractions of anode elements that enter residues and electrolyte.

Copper Cathodes. Cathode copper is produced currently in a purity between 99.97 and 99.99 %. Silver can be deposited in traces; however, this can be avoided by precipitating the silver from the electrolyte with chloride ion. Other impurities, such as suspended slime or droplets of the electrolyte, may be mechanically occluded. The following measures are taken to produce copper of high purity:

1) Maintenance of the optimum current density, to prevent cathodic deposition of other elements (e.g., arsenic)
2) Addition of organic inhibitors to avoid the formation of nodules on the cathode surface
3) Removal of impurities such as arsenic, antimony, and bismuth from the electrolyte by adsorption or chemisorption
4) Prevention or elimination of suspended slimes by regulating the electrolyte flow and occasionally filtering it.

Electrolyte. The composition of copper electrolytes from various plants is generally similar: ca. 35 – 45 g of copper and 150 – 220 g of sulfuric acid per liter at an operating temperature of 55 – 65 °C (see Table 14). As a result of secondary reactions during electrolysis, the concentration of copper (II) ions increases slowly; therefore, this copper surplus must be recovered by cathodic deposition in a few (ca. 2 %) *liberator cells* equipped with insoluble anodes, usually of antimonial lead. Soluble impurities, such as iron, cobalt, zinc, manganese, most of the nickel, and some arsenic and antimony, are also enriched in the electrolyte.

The upper limits of impurity content are ca. 10 g/L for arsenic and 20 – 25 g/L for nickel. Part of the electrolyte is withdrawn continuously from the circuit for purification, and the volume is compensated by adding sulfuric acid and cathode wash water. There are two methods of purification. In the first, the solution can be completely decopperized in a system of special liberator cells with insoluble anodes; arsenic and antimony are almost completely deposited and returned to pyrometallurgical operations. The electrolyte is then concentrated by vacuum evaporation to yield concentrated sulfuric acid and crude nickel sulfate, from which pure nickel sulfate or nickel metal can be produced.

The second method of purification is by producing copper sulfate. For this purpose, the sulfuric acid is usually neutralized by addition of copper shot in the presence of air. The copper sulfate is obtained by crystallization, and the mother liquor is cemented with iron scrap.

Anode Slimes [153], [154]. The content of insoluble substances is < 1 % of the anode weight, and they collect on the bottom of cells as anode slime. They contain precious metals (silver, gold, and platinum); selenides and tellurides of copper and silver; lead sulfate; stannic [tin (IV)] oxide hydrate; and complex compounds of arsenic, antimony, and bismuth (the undesired floating slimes). The main component is copper. In addition, gypsum and silica, alumina, or baryte from anode casting are present.

Table 18. Industrial range of copper anode and cathode compositions

Element	Anodes (%)	Cathodes (%)
Cu	99 – 99.8	99.99 + (excluding oxygen)
O	0.1 – 0.2	not determined
Ag	0.01 – 0.5	0.0005 – 0.002
S	0.0005 – 0.005	0.0004 – 0.001
Ni	0.002 – 0.7	trace – 0.0003
Pb	trace – 0.5	trace – 0.0004
Fe	0.001 – 0.006	0.0001 – 0.0002
Sb	trace – 0.1	trace – 0.0002
As	trace – 0.2	trace – 0.0002
Se	trace – 0.2	trace – 0.0001
Te	trace – 0.03	trace – 0.0001
Bi	trace – 0.03	trace – 0.0001
Au	trace – 0.004	trace

Table 19. Fractions of anode elements entering residues and electrolyte

Metal	Percentage into anode residues	Percentage into electrolyte
Cu	< 0.2	> 99.8
Au	99	< 1
Ag	98	2
Se and Te	98	2
Pb	98	2
Sb	50	50
As	30	70
Co	5	95
Ni	5	95
Fe	0	ca. 100
Zn	0	ca. 100
Bi *		

* Dissolves up to about 0.15 – 0.2 kg m^3 in electrolyte, then forms slimes.

The distribution of the elements (in %) varies over wide ranges:

copper	20 – 50
nickel	0.5 – 2
lead	5 – 10
arsenic	0.5 – 5
antimony	0.5 – 5
bismuth	0.5 – 2
selenium	5 – 20
tellurium	1 – 4
silver	≤ 25
gold	≤ 4

Although the separation techniques differ greatly from plant to plant, anode slimes are generally treated as follows:

1) Oxidizing leaching of copper with dilute sulfuric acid
2) Recovery of selenium and tellurium by pyrometallurgical or hydrometallurgical methods
3) Removal of detrimental elements and production of silver alloy (Doré bullion)
4) Separation of precious metals by electrolysis (silver and gold) and fractional precipitation (platinum metals)

The greatest part of the world selenium production comes from processing copper anode slimes.

6.2.2. Practice of Electrorefining [155–157]

Electrorefining is carried out in large tankhouses with capacities up to 500 000 t/a of pure cathode copper. Today two techniques with principally different starting cathode sheets are used. The conventional system uses thin copper sheets, while the modern systems (ISA developed by Copper Refineries [167], Australia and Kidd developed by Kidd Creek Divison, Falconbridge, Canada) utilize permanent stainless steel cathode sheets. The cathode copper is stripped mechanically from the steel sheets, which are used again, while copper starting sheets have to be prepared for each cathode. The advantages of the permanent cathode systems are high degree of automation, very low personnel costs, higher specific production rate, and higher cathode quality (lower content of impurities, e.g., lead).

Cells and Operation. The cells or electrolytic tanks are constructed of reinforced concrete lined with antimonial lead sheet or with flexible poly(vinyl chloride) or of polymer concrete without a lining. The cells are rectangular, 3.5 – 6 m long, 0.8 – 1.2 m wide and 1.2 – 1.5 m deep. Copper anodes and starting sheet cathodes are suspended alternately in the cells with precise spacing by overhead crane. In most refineries the anodes are prepared (pressed to make them uniformly flat and weighed) before they are inserted into the cells. Commercial cells contain between 40 and 60 electrodes of each type. The distance between the surfaces of the two adjacent electrodes should be small (about 2 cm) to avoid voltage loss, but not so small that the deposited copper is contaminated with anode slime. Electrolyte is circulated through each cell at ca. 0.02 m^3/min. At this flow rate the elec-

trolyte in one cell is exchanged every 5 – 6 h, which is important for removing impurities from the cell.

The anodes remain in the cell for 20 – 28 d. During this time the cathodes are usually changed two or three times, so that their weight is only 60 – 160 kg. The cathodes are removed by crane and the electrolyte on their surface is washed off with hot water. In tankhouses which use stainless steel sheets, the plated copper is stripped mechanically (automatic stripping machine) from both sides of the blank. In conventional tankhouses copper starter sheets are prepared by electrodeposition of copper on titanium sheets (< 1 mm, about 1 d) in special stripper cells. Then the copper sheets are attached to rolled copper support bars for hanging in the cells. This process is partly mechanized [164].

About 15 % of the anode is not dissolved during the refining process. This anode scrap is recycled to the smelter (converter or shaft furnace). The anode slime of the complete cycle is drained, filtered, washed, and sent to the precious metal plant. Part of the electrolyte is removed continuously from the circulation system for purifying. The sulfuric acid is sent back to the tankhouse.

Electrical System. The anodes and cathodes in a cell are connected in parallel. The cells are connected in series to form sections of 20 to 40 cells. Each section can be isolated electrically for inserting or removing anodes or cathodes. For example, 12 sections are connected with one silicon controlled rectifier for current input. Total electrical current through the cell is between 10 000 and 40 000 A. Electrical densities of up to 40 kA/m^2 are used in ISA tankhouses today. To obtain a high current efficiency of 92 – 98 %, the bath and section voltages are monitored [254], [255], and the temperature of the cells is measured by infrared techniques (from the crane). In this way short circuits, caused, for example, by the growth of copper nodules can be quickly detected and corrected. The energy consumption of electrorefining per tonne of plated copper is about 260 – 280 kW h.

Electrocrystallization [163]. The structure of an electrodeposited cathode surface can be influenced by adsorption of molecules on the crystallites. The addition of colloids or special organic compounds improves cathode quality by yielding an evenly grained copper deposition. Important surface-active additives, or inhibitors, are bone glue, gelatin, avitone A (sulfonated aliphatic hydrocarbons), goulac, or bindarene (sulfonated wood fibers). Effective substances with definite composition are thiourea and its derivatives and other sulfur – nitrogen compounds. These inhibitors, mostly in combination, are added to the electrolyte in extremely small amounts, although more is added at higher current density. Inhibitors increase the voltage and, therefore, the energy consumption. Several refineries automatically control their reagent addition rates by measuring glue and thiourea, for example, with the CollaMat [254–256] system. This stabilizes refinery operation and optimizes the process (lower energy consumption).

Special Developments. Jumbo tanks instead of conventional cells, first installed in 1972 at Onahama Tankhouse, Japan [165]. The jumbo tanks have about 20 times the volume of conventional cells. The electrolyte flows parallel to the electrode surface, in contrast to conventional cells. A newer installation of jumbo tanks is the Kidd Creek tankhouse in Canada.

To overcome anodic passivation caused by high current density and impure anodes a periodic current reversal system (PCR) [166] was developed. Optimum reversal conditions seem to be 20 units forward time and 1 unit reverse time [257]. This system has been adopted in some refineries, but due to the extra energy costs it is not widely used.

6.3. Melting and Casting

Copper cathodes must be remelted and cast to shapes because the structure of copper formed by electrocrystallization is not suitable for working to semifinished products. Cathodes are remelted in several types of special furnaces that perform the tasks of melting, postrefining (if necessary), holding, and casting. Casting can be carried out by the older discontinuous methods or by modern continuous casting.

6.3.1. Remelting of Cathodes

There are various kinds of furnaces that use either fossil fuels (coal, coke, fuel oil, natural gas, or reformed natural gas) or electric energy:

Small coke-fired crucible furnaces
Gas- or oil-fired rotary furnaces
Large hearth furnaces (reverbs)
Electric-arc furnaces
Low-frequency induction furnaces
Cathode shaft furnaces (e.g., ASARCO type)

Copper ready for pouring must be nearly free of sulfur, at most 10^{-3} % (10 ppm) S, because a higher content affects detrimentally the mechanical properties of the metal.

In practice, copper is treated in two ways:

1) After melting cathodes with sulfur-containing fuels, the copper melt must be fire-refined like blister copper, by oxidation and poling. This is the case when hearth furnaces are employed for casting.
2) Use of electric power or sulfur-free fuels allows the use of continuous units, such as induction or ASARCO furnaces.

The *ASARCO shaft furnace,* constructed by American Smelting and Refining Co. [168], [169], has a cylindrical shaft consisting of a steel jacket with a brick lining. Cathodes are charged near the top, and the combustion gases ascend in countercurrent flow from groups of burners; the liquid metal is collected in a holding furnace. Apart from its effectiveness and high productivity, a distinct advantage is the maintenance of a constant, slightly reducing atmosphere by automatic control. The largest types (1.8 m diameter and 8 m high) can have a throughput up to 80 t/h. Worldwide ca. 200 units are in operation.

6.3.2. Discontinuous Casting

The discontinuous casting of various shapes on horizontal casting wheels with open ingot molds, analogous to the casting of anodes, was formerly the most important casting method. It is being replaced by continuous casting processes.

6.3.3. Continuous Casting

Since the end of World War II, several continuous casting methods have been developed. A comprehensive synopsis, including patents, has been published [171].

Continuous casting has several technological and economic advantages over the older casting processes, such as excellent surface quality, uniform grain structure without blisters and shrinkage cavities, and energy savings.

Molten copper flows continuously into a relatively short water-cooled chill, usually lined with graphite and open at both top and bottom. The solidifying metal itself forms the lower closure. The shape is steadily withdrawn by clamping rolls and cooled by spraying with water.

There are three main continuously cast shapes: circular billets with a diameter up to 500 mm (for extrusion presses or tube rolling mills), square bars, and rectangular cakes with cross sections up to 1300 × 200 mm (for rolling to sheets and strips). Seamless tubes and other hollow shapes are occasionally produced. Continuous casting shapes are automatically sawn off by flying saws when they reach the required lengths (up to 7.5 m). In semicontinuous casting, the process must be interrupted, but the shapes can be up to ca. 12 m long.

The kind of chill is generally independent of the furnace: the liquid metal stream free-falls into the chill (e.g., Junghans system), or the less frequent fixed connection of chill with the furnace (e.g., ASARCO casting system).

Other methods employ a joint moving in unison with the solidifying metal and a mold with traveling parts forming a gap of the required cross section. (An example is the Hazelett process employed for continuous casting of anodes. Section 6.1.3.) Such casting methods are particularly suitable for continuous production of shapes with a small cross section. Section 6.3.4.

6.3.4. Continuous Rod Casting and Rolling

The continuous production of cast and rolled wire rod [172], [173] involves considerable energy savings because the solidified hot metal can be rolled immediately. Several plants use such a direct process starting from molten electrolytic tough-pitch copper. Worldwide >100 plants are in operation, using the following systems:

1) Properzi process [174]. The first continuous copper rod caster was constructed in 1960 following developments for other nonferrous metals. It operates on the wheel – belt casting principle, i.e., casting into the gap between the periphery of the casting wheel and the closing steel belt. About ten plants are in operation with a capacity up to 30 t/h.
2) Southwire process [175]. Started in 1963 as a further development of the wheel – belt casting principle with capacity up to ca. 50 t/h, the Southwire process directly introduces the continuous cast bar into the rolling mill (Fig. 33). After rolling, the rod, which is oxidized on the surface, is continuously treated by pickling with dilute sulfuric acid or alcohols, water or steam rinsing, and wax coating. The saleable product (8 – 20 mm diameter) is formed into "coils" of up to 10 t, which are packaged. About 30 plants exist at present.
3) Secor process [177]. Only two factories (Australia and Spain) use this modified wheel – belt casting concept, dating from 1975, with a capacity up to ca. 10 t/h.
4) Contirod or Krupp – Hazelett process [178]. As a variant of the Hazelett twin-belt process similar to the Contilanod process (Section 6.1.3.), the continuous cast bar solidifies between two belts and damblock chains and is directly moved to the rolling mill (Fig. 33). Metallurgie Hoboken-Overpelt, Belgium, developed this system in the 1960s; the capacity of the largest units is ca. 50 t/h. At present about 20 plants are operated.
5) General Electric dip forming process [179]. A process based on the "candlestick" principle has been operated since about 1970. A copper core rod is pulled upward through liquid copper so that its diameter increases; the thickened rod moves immediately to rolling. Oxygen-free copper can be produced by using a reducing atmosphere. The capacity is ca. 10 t/h. Nearly 20 plants exist at present.
6) Outokumpu up-cast process [180]. A new upward casting system developed in 1969 draws copper upward through a vertical die cooler with a cooled graphite mouthpiece dipping into the melt. The caster, comprising 8 or 12 strands, yields oxygen-free copper at the rate of ca. 2 t/h per line. Because of the small cross section (8 – 25 mm diameter), hot rolling is not required. Approximately 40 plants are in operation.

6.4. Copper Powder

Copper and copper-alloy powders are required for products prepared by powder metallurgical techniques, including friction materials, carbon brushes, self-lubricating bronze bearings, special filters, and catalysts and other sintered components.

The principal methods for producing copper powders are electrolytic deposition at high current densities and the atomization of molten metal, the latter more for copper-alloy powders. Copper powders are also formed by cementation or by pressurized precipitation from aqueous solutions, but such precipitates are of little commercial interest.

Atomizing is done by spraying a melt into a pressurized air or water flow. Various grain shapes are formed, depending on the cooling rate and on additives that change the surface tension. Additives that decrease surface tension, e.g., magnesium, form irregular powders; those that increase it, e.g., lead or phosphorus, yield globular particles. Spongy powders can be obtained by reduction of oxidized copper powders with hydrogen.

Electrolytically deposited powder particles have dendritic shape; a typical flow sheet is shown in Figure 34. For this purpose electrolysis is used as a shaping process rather than for refining because high-purity copper cathodes are the anodes. The main parameters of powder electrolysis are, as Figure 35 shows, the following [181]:

Cathodic current density
Electrolyte temperature
Concentration of copper (II) ions
Concentration of chloride ions

Varying these factors markedly change particle shapes and apparent densities.

Anodic and cathodic current densities differ. The former is normally 300 – 600 A/m^2, and the latter is 2000 – 4000 A/m^2, which is 10 – 20 times higher than the cathodic values in conventional electrolytic copper refining. This effect is obtained by using copper rod cathodes and plate-like anodes. The energy consumption in powder

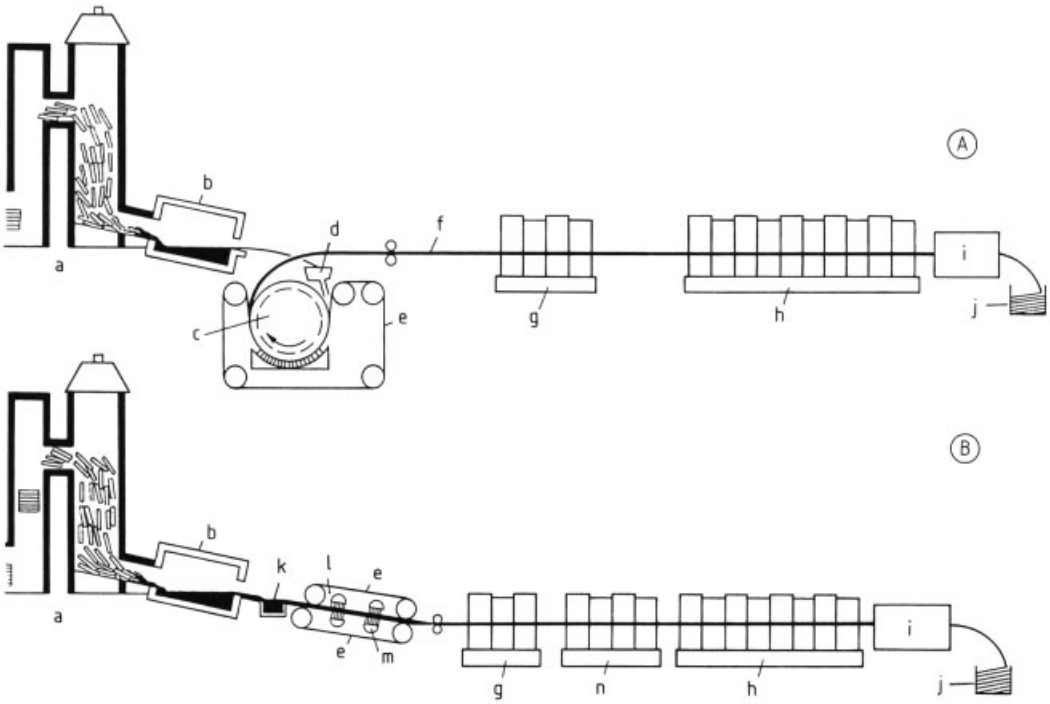

Figure 33. Scheme of continuous rod casting and rolling [176]
A) Southwire process; B) Contirod process.
a) Melting furnace (ASARCO); b) Holding furnace; c) Wheel; d) Tundish; e) Steel band; f) Continuously cast copper bar; g) Preliminary rolling mill train; h) Finishing rolling mill train; i) Pickler; j) Coiler; k) Casting receptacle; l) Casting canal; m) Stationary edge dams; n) Middle rolling mill train

electrolysis averages nearly $2\,\mathrm{kWh\,kg^{-1}}$. The powders are generally posttreated for various purposes. Electrolytic copper powders are characterized by dendritic particle shape, high purity, low oxygen content, favorable resistance to oxidation, and good green strength.

During the last 5 years the electrowinning technology to produce copper powder was developed to commercial scale. By using pregnant leach solution from ore leaching operation and subsequent solvent extraction and electrowinning in specially designed cells, it is possible to produce dendritic shape copper powder with well defined parameters like particle size [258].

6.5. Copper Grades and Standardization

The three main grades of refined copper are (1) tough-pitch copper, (2) deoxidized copper, and (3) oxygen-free copper.

Without regard to the method of refining, *tough-pitch copper* normally contains 0.02–0.04 wt % O (as Cu_2O) and ca. 0.01 wt % total of other impurities. This grade is easily worked and has an electrical conductivity of $\geqq 100$ % IACS, but it is unsuitable for welding and brazing because of the danger of hydrogen embrittlement. About 80 % of the world production of refined copper is tough pitch, mostly electrolytic tough-pitch copper (ETP).

Deoxidized copper, with no oxygen content is produced by addition of nonmetallic or metallic reducing agents, usually copper phosphide. As a result of the absence of oxygen, hydrogen embrittlement cannot occur, and such copper is well suited for welding and brazing. The content of residual phosphorus, however, increases the electrical resistivity. Deoxidized copper with low residual phosphorus (DLP, ca. 0.005 % P) is required if the metal is to be used as a conductor. Copper with high residual phosphorus (DHP,

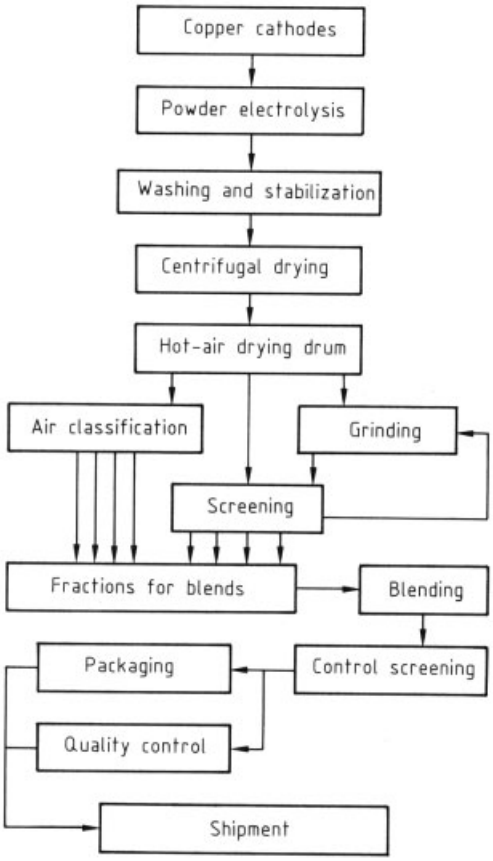

Figure 34. Flow sheet for electrolytic copper powder production [181]

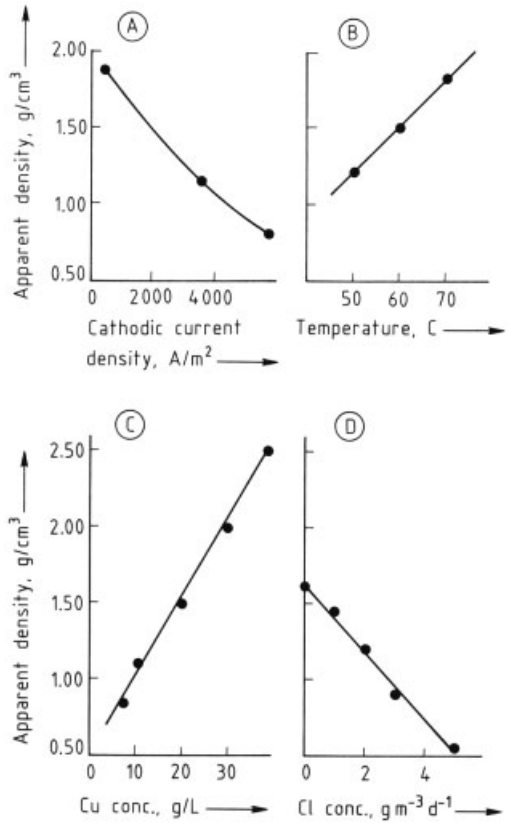

Figure 35. Apparent density of electrolytic copper powder as a function of four electrolysis parameters [181]
Other conditions: A) Cu 6 g/L, no Cl, 60 °C; B) Cu 17 g/L, no Cl, 3600 A/m^2; C) no Cl, 50 °C, 3600 A/m^2; D) Cu 13 g/L, 50 °C, 3600 A/m^2

ca. 0.04 % P) can be employed for nonelectrical purposes.

Oxygen-free (OF) copper is a special quality produced by keeping oxygen away from the copper melt in a controlled atmosphere. Copper of this type with an oxygen content less than 0.001 wt %, is suited for purposes requiring weldability and high electrical conductivity, especially electronics. OFHC, oxygen-free high-conductivity copper, is internationally known.

Standardization [14], [182], [183]. Most industrial countries have established standards for copper; these national specifications include detailed specifications for chemical composition, physical properties, and geometrical dimensions, but differences exist. Table 20 indicates the rough equivalent between international and European specifications for the most important copper properties.

6.6. Quality Control and Analysis

Tests for quality control of copper are carried out on samples of both refinery shapes and semifinished products. There is need to standardize the testing methods, but currently only some of them are fixed in national specifications. The most important tests are the measurement of electrical conductivity, mechanical properties, and quality of the metal surface.

Table 20. Comparison of international and selected national standards

International ISO R 1337	European EN prEN 1977 prEN 1978	Materials Number prEN 1412	Composition, %				Electric conductivity, % IACS	Remarks
			Cu min	O_2 max	P min	P max		
Cu-CATH1	Cu-CATH1	CR001A	99.99					higher grade cathodes
Cu-CATH2	Cu-CATH2	CR002A	99.9					standard cathodes
Cu-ETP1	Cu-ETP1	CR/CW003A	99.9	0.04			101	electrolytically refined, tough pitch copper
Cu-OF1	Cu-OF1	CW007A	99.99				101	oxygen-free copper
Copper not produced from Cu-CATH-1 Cathodes:								
Cu-ETP	Cu-ETP	CR/CW004A	99.99	0.04			100	electrolytically refined tough pitch copper
Cu-FRHC	Cu-FRHC	CR/CW005A	99.99	0.04			100	fire-refined tough pitch copper
Cu-OF	Cu-OF	CR/CW008A	99.95				100	oxygen-free copper
Cu-FRTP	Cu-FRTP	CR/CW006A	99.9	0.1				fire-refined copper
Phosphor containing Copper:								
Cu-PHC	Cu-PHC	CR/CW020A	99.95		0.001	0.006	100	
Cu-HCP	Cu-HCP	CR/CW021A	99.95		0.002	0.007	98.3	
Cu-DLP	Cu-DLP	CR/CW023A	99.9		0.005	0.013		deoxidized copper with low P content
Cu-DHP	Cu-DHP	CR/CW024A	99.9		0.015	0.04		deoxidized copper with high P content

The *electrical conductivity* is very sensitive to impurities and crystal lattice imperfections. *Mechanical tests* measure hardness, tensile strength, elongation at failure, and torsional fatigue strength. The spiral elongation test [184] is a complicated test method that assesses the purity and the mechanical behavior of the sample. Defects on the *surface* and subsurface can occur in various forms, e.g., folds, splashes, cracks, inclusions, and voids. The voids are caused by gas porosity, shrinkage porosity, and shrinkage cavities.

Nondestructive testing procedures, such as radiography, ultrasonic examination, or the eddy-current technique can be used. Metallographic methods (polishing and etching) give information by microscopic examination.

Analytical methods are of interest for determining the impurity level of copper products. Both wet chemical procedures and physicochemical procedures, such as atomic absorption spectrometry, optical emission spectroscopy, and X-ray fluorescence spectroscopy, are employed, the latter essentially for quick analysis of solid and liquid co- and byproducts [185], [186]. The classical analytical methods are gradually being superseded by the modern automatic instrumental techniques. It is especially important to analyze the oxygen content, and one effective modernized method is available, hot extraction, i.e., melting a copper sample in a small graphite crucible and determining the CO formed by IR absorption spectroscopy.

7. Processing and Uses

The pure metal produced in refineries or remelting plants is manufactured into semifabricated products.

7.1. Working Processes

Usually copper is treated initially by noncutting, shaping processes to obtain semifinished products or "semis". These processes are subdivided into hot working, cold working and, if necessary, process annealing.

Hot working means plastic forming above the recrystallization temperature. Generally copper is preheated to 800 – 900 °C, and the subsequent hot forming is finished at ca. 400 °C. Cast bars from modern combined continuous casting/rod-rolling systems already have sufficient temperature, thus saving thermal energy. After cooling, the hot-worked copper is *soft copper*. Its

mechanical and electrical properties are scarcely changed, but its density has increased to nearly 8.9 g/cm^3.

The next step is *cold working*, which involves plastic forming below the recrystallization temperature. In practice, the operation is done at room temperature. Unlike hot working, this procedure entails an essential strain hardening of the metal by increasing the number of lattice defects; however, simultaneously formed lattice voids cause a considerable decrease of electrical and thermal conductivity. After cold working, the metal is *hard copper*.

Process annealing is a heat treatment that is necessary if the hardened copper must be softened again, either for continued working or for producing (soft) copper with high electrical conductivity. Special furnaces are used for the purpose of steady heating and cooling of the metal – often in a nonoxidizing atmosphere. To achieve the intended microstructural change, the recrystallization temperature of 200 – 300 °C must be exceeded; in practice, the metal is heated to 400 – 500 °C for accelerated recrystallization. Copper products with exactly defined properties can be obtained if all annealing conditions are carefully controlled.

Table 21. Important fabricating processes for copper products

Refinery shapes	Hot-working process		Cold-working process	Semi-finished products
Cakes	hot rolling	→	cold rolling	sheets/strip
Rod		⟶	wire drawing	wires
Billets	extruding	→	cold drawing	rods/sections
Billets	extruding	→	cold drawing	tubes/pipes
		↗↘		
	or hot rotary piercing	→	or cold pilger rolling	

The engineering techniques are versatile. The following working methods are of special importance:

hot working	cold working
hot rolling	cold rolling
extrusion	cold drawing
drop forging	cupping

Foils only ca. 0.002 mm thick are manufactured by rolling, and wires to 0.004 mm diameter by drawing. Many products of varying size are fabricated by modern variants of the extrusion process [187]. The fabrication of tubes is also quite diverse [188]. The most widely used working processes are compiled in Table 21.

7.2. Other Fabricating Methods

In many cases, machining operations are required, e.g., cutting, turning, planing, drilling, and sawing. However, these are more important for copper alloys than for pure copper because of copper's tendency to gum. Noncontinuous casting processes are likewise more suitable for copper alloys because copper has a disadvantageous coolability. These include sand mold casting, permanent mold casting, gravity die casting, pressure die casting, and centrifugal casting. Continuous or semicontinuous casting processes, however, are well-suited for pure copper.

Galvanoplasty is an electrolytic operation for manufacturing complicated objects that require high precision and flawless surfaces such as hollow bodies, disk matrices, and electrotypes. A special galvanic method is copper plating, which involves electrolytic deposition of a thin layer of copper on another metal either for surface protection or as a base layer for electroplating with another metal (→ Electrochemical and Chemical Deposition).

Powder-metallurgical techniques are used primarily for the mass production of small pieces, especially intricate forms such as electrotechnical and mechanical structural parts. The metal powders are first compacted by pressure and then sintered in a controlled atmosphere. The copper powder is often mixed with other powdered metals, including those that do not form common copper alloys (→ Powder Metallurgy and Sintered Materials).

There are other important fabricating methods [189]. Joining is usually carried out above room temperature by soldering, brazing, or welding. Soldering may be used for all sorts of copper, owing to the low temperature. However, welding and brazing are feasible only with deoxidized or oxygen-free copper.

When tough-pitch copper is heated in an atmosphere containing hydrogen, the steam generated (see page 7) collects within the grain boundaries at high pressure and can destroy the grain

structure by forming cracks. This phenomenon is known as "hydrogen embrittlement."

Mechanical joining and metal bonding are also possible ways of joining copper with other materials.

Surface treatment of copper is a group of operations for surface protection or surface refinement. These include mechanical, electrical, or electrochemical handling, e.g., polishing, matte finishing, pickling by dilute sulfuric or nitric acid, metal coating or electroplating (with nickel, nickel and chromium, tin, silver, gold, or platinum metals), lacquering or coating with synthetic plastics (mainly for electrical insulation), enameling of objects (applied art), and chemical or electrochemical coloring (decoration). Coloration is effected by chemicals, mostly specially formulated metal salt solutions which form thin layers of insoluble green, red, brown, or black compounds.

7.3. Uses

Copper is a useful material with a wide range of applications because of its combination of properties. Because of its excellent electrical conductivity, it is the dominant conductor material. Copper is used primarily as round wire or rods, bare or insulated, for current generation, transmission, and conduction; various sorts of cables are produced for special applications. Substantial quantities of copper are made into generators, motors, transformers, and other electrical appliances. About 40 % of the world consumption of copper is for electrical purposes. As a result of its high thermal conductivity, copper is well-suited for vessels and pipes, especially for heating, cooling, and heat exchange.

While high-conductivity copper is required for electrotechnical and electronic uses, special copper qualities are chosen for other uses. About 30 % of world copper production is used for alloying. Copper alloys are usually cold-worked; only ca. 10 % of them are cast.

Copper is frequently used in the chemical and food industries because of its high resistance to corrosion. There is substantial use of copper in mechanical engineering, by fabricators of precision implements, in vehicle construction, and in ship building. There is increasing interest in copper building construction as a material for installation, wall lining, and roofing. Hydraulic engineers use copper sheets for tightening on dams, sluices (floodgates), and bridges.

Other areas of application are in the fabrication of household articles, art objects, coins and medals, and in military hardware as ammunition. There is a smaller demand for other purposes, such as electrodeposition; powder-metallurgical copper, special materials for brakes and self-lubricating bearings, small precision parts, filters, graphite brushes; and alloying additives for aluminum, iron, and steel. Use in copper compounds, chiefly copper sulfate and copper oxides, consumes only 1 – 2 % of the primary world production.

Table 22 lists the distribution of copper consumption among various industries.

Table 22. Industrial use of copper (including alloys) in the Western world in 1995, percentage by country [190]

Branch	United States	Europe	Asia
Electrical and electronic industry	25	37.5	50
Industrial machinery and equipment	11	9	9
Building construction	43	39.5	15
Transportation	9	6.5	11
Consumer and general products	12	7.5	15

Substitution and Miniaturization. Several materials compete with copper and may substitute for it, depending on the relative costs. Copper is partly replaced by aluminum in automotive radiators and in transmission cables, high-voltage long-distance lines, and household wiring. Copper wires and cables for telecommunications are being displaced by microwave technology and fiber optics. Copper is being replaced by plastics for water pipes in both residential and commercial construction. In the area of corrosion-resistant materials, in addition to plastics there are also stainless steel and titanium.

The movement toward making smaller and smaller parts has been one of the most pervasive and continuing pressures on the copper market. A dramatic drop in the use of copper has occurred in the widespread acceptance of printed circuits. The use of wire has plummeted. The number and size of the connectors have dropped. On the other hand, miniaturization has steadily

decreased the cost of the final products, thus increasing the number of units sold.

At the same time, however, this drive towards miniaturization, whether in the thickness of an automotive radiator or in the size of an electronic component, is a challenge to the copper industry to produce purer copper and more useful alloys and to the copper fabricating industry to produce the new miniaturized products.

In several applications copper is resisting substitution by using new technologies. For example in telecommunications, copper continues to be the preferred signal carrier for the last mile. The new xDSL (Digital Subscriber Line) technology allows the existing copper infrastructure of ordinary telephone wires to also carry high-speed data. The installation of optical fiber in communication trunk lines has led to a revolution in the telecom industry. Copper application was partly displaced, but overall this increased the demand for copper. Another development is the use of copper circuitry in silicon chip technology, which makes the microprocessors faster and lowers energy consumption. Another example is the automobile radiator, which was formerly made of copper, which was then displaced by aluminum. New technology was developed for producing smaller and lighter copper brass radiators with higher thermal conductivity than aluminum radiators. A final example for innovation in copper is the development of superconducting power cables made from high-temperature superconductor wire. This technology will improve energy efficiency, and now projects in Chicago and Tokyo have been started.

8. Economic Aspects

There are numerous tabular compilations of statistics on copper resources, production, and consumption [192–196]. Many books deal with economic relations and commercial problems, e.g., [197–200]. Compilations of companies in the nonferrous metal industry [201] and the mining industry [202] are published at irregular intervals.

Production and Consumption. The copper production of mines, smelters, and refineries and the consumption by country and region are given in Table 23. Over the decades, the annual per capita consumption of primary copper in the United States has grown [195]:

Late twenties	7.5 kg
Early thirties	2.5 kg
World War II	9.5 kg
Postwar period	7.5 kg
1970	13.3 kg
1979	14.6 kg

During the 1980s, there has been no increase in consumption in America and Europe and only a small increase worldwide. In the 1990s copper production increased due to industrial growth.

Figure 36 shows the development of world per capita consumption for 1950 to 1997.

Cost of Copper Production and Copper Price. The cost of copper production is characterized by high capital investment in mining projects and in smelters and refineries. Mining projects are financed by large consortiums and banks. The capital investment for a green-fields smelter is in the region of 2500 – 3000 $/t of design copper production. Smelter enlargement investment is approximately half that. Therefore, increased smelter production is preferably achieved by enlargement.

In the last 15 years copper leaching projects have been established. This is due to lower capital investment than smelters (about half), and also leaching operations built to increases the copper yield of ores.

Operational costs are high due to energy consumption, which is the most important factor. For primary copper production the overall energy consumption per tonne of copper is about 45 GJ, about half of which is consumed in mining and beneficiation and the rest in smelting and electrorefining. For secondary copper, coming for example from copper scrap smelting and refining, the overall energy consumption is only 20 GJ/t.

Due to several smelter enlargements, leaching operations, and also new energy-efficient milling and smelting/refining processes, the overall production cost of copper are falling.

The copper price is set primarily at the two metal exchanges: the London Metal Exchange (LME) and the New York Commodity Exchange (COMEX). Like the copper production cost also the copper prices have been in an overall declining trend since the World War II. The copper

Table 23. Production and consumption of copper in 1997 (in 10^3 t)

	Mine Production	Metal ∗∗ Production	Metal ∗∗ Consumption
Argentina	30.2	15.6	51.7
Armenia	1.9		
Australia	560.0	271.1	181.7
Austria		77.0	31.0
Belgium-Luxembourg		386.0	363.3
Botswana	20.7		
Brazil	40.0	177.1	254.6
Bulgaria	75.5	34.9	42.0
Canada	658.0	560.3	224.6
Chile	3 392.0	2 116.6	79.9
China	495.5	1 179.4	1 258.0
Colombia	1.8	4.0	
Congo	40.1	40.1	
Cyprus	3.9	3.9	
Czech Rep.			9.7
Egypt		4.4	4.3
France		35.1	558.0
Georgia	4.0		
Germany		673.6	1 040.0
Greece			96.8
Hungary			13.3
India	38.0	65.9	180.0
Indonesia	548.4		54.2
Iran	117.3	106.3	56.4
Italy		85.7	520.6
Japan	0.9	1 278.7	1 440.7
Kazakhstan	316.2	301.1	13.7
Korea (North)	10.0	30.0	15.0
Korea (Rep.)		262.6	624.3
Macedonia	13.0		
Malaysia	18.9		159.6
Mexico	393.1	297.0	239.7
Mongolia	130.0	3.0	
Marocco	15.4		
Myanmar	6.0		
Namibia	20.3		
Netherlands			40.4
Oman		23.6	
Pakistan			11.0
Papua New Guinea	111.5		
Peru	503.3	384.1	48.0
Philippines	48.6	146.6	45.5
Poland	414.7	440.6	233.0
Portugal	106.5		1.4
Romania	23.2	22.9	14.5
Russian Fed.	526.0	610.0	165.0
Saudi Arabia	1.0		145.0
Scandinavia ∗	101.8	277.2	264.5
Singapore			10.3
Slovakia		23.0	28.0
South Africa	185.6	130.2	81.8
Spain	38.3	292.0	203.0
Switzerland			7.5
Taipei, China			587.8
Thailand			151.9
Turkey	65.0	113.7	194.4
United Kingdom		60.4	408.5
United States	1 940.0	2 452.4	2 790.0
Uzbekistan	74.4	115.0	10.0
Venezuela			18.0
Yugoslavia Fed.	73.6	106.6	36.0
Zambia	352.9	338.4	17.7
Zimbabwe	6.8	17.9	15.0
Others	1.8		16.0
TOTAL	11 526	13 564	13 084

∗ Scandinavia includes Finland, Norway, and Sweden. ∗∗ Refined copper.

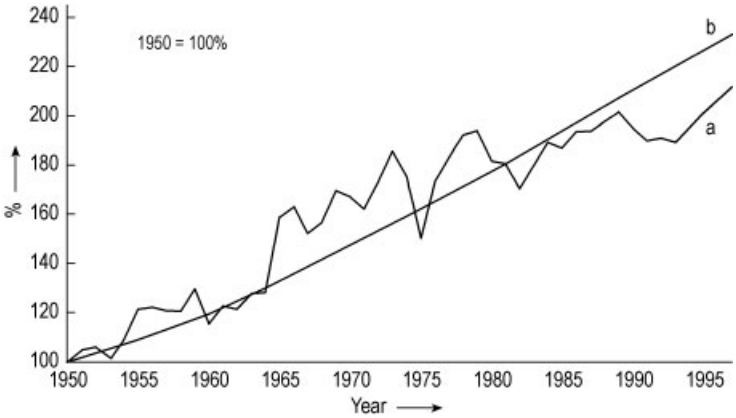

Figure 36. Per capita consumption of refined copper, 1950 – 1997
a) Relative per capita copper consumption; b) World population
Base index figures for 1950 are a copper consumption of 1.07 kg per capita and a world population of 2.5×10^9

price also increases and decreases in the same economic cycles as industrial growth and recession. The development of copper prices since 1960 is shown in Figure 37.

Product Information. Information on recent developments in copper are available at http://www.copper.org and http://www.kupfer.org. Statistical data are available at http://www.icsg.org.

9. Environmental Protection

The worldwide growth of industry and population has caused a series of environmental problems, particularly the following: (1) emission control; (2) water protection; (3) solid-waste disposal.

Emission Control. There are two important tasks in the treatment of off-gas from pyrometallurgical processes in copper metallurgy: the removal of sulfur dioxide and the containment of flue dust.

Because most copper comes from sulfide ores, *sulfur* is the main problem in copper extraction. In pyrometallurgical operations, it appears as sulfur dioxide (Table 24) [204], [205]. The mass ratio of sulfur to copper in sulfidic concentrates is usually between 0.8 and 1.6. Consequently, a large quantity of sulfur dioxide must be captured because of its harmful effects on health, vegetation, and property. In most cases, sulfuric acid is produced from the SO_2-containing off-gas (→ Sulfuric Acid and Sulfur Trioxide) [206].

Table 24. Comparison of SO_2 concentrations in copper smelting off-gas [203]

Process	SO_2, vol %
Multiple-hearth roasting	5 – 8
Fluidized-bed roasting	8 – 15
Sinter roasting	1 – 2
Blast furnace smelting	2 – 5
Reverberatory furnace smelting	0.5 – 2.5
Electric furnace smelting	0.5 – 5
Outokumpu flash smelting	10 – 30
INCO flash smelting	75 – 80
KIVCET process	80 – 85
Peirce – Smith converter	5 – 12
Hoboken converter	7 – 17
TBRC process	1 – 15
Mitsubishi process	15 – 20
Noranda process	10 – 20

Flue dust can be separated from off-gas to a high degree in modern gas-cleaning systems such as electrostatic precipitators, baghouses, cyclones, and wet scrubbers. This metal-containing dust is recycled.

Water Protection. Harmful wastewater does not usually result from pyrometallurgical copper production but water for direct or indirect cooling of furnaces, casting machines, and cast copper products is required on a large scale. This cooling water is moderately warmed, but does not acquire chemical impurities. Closed

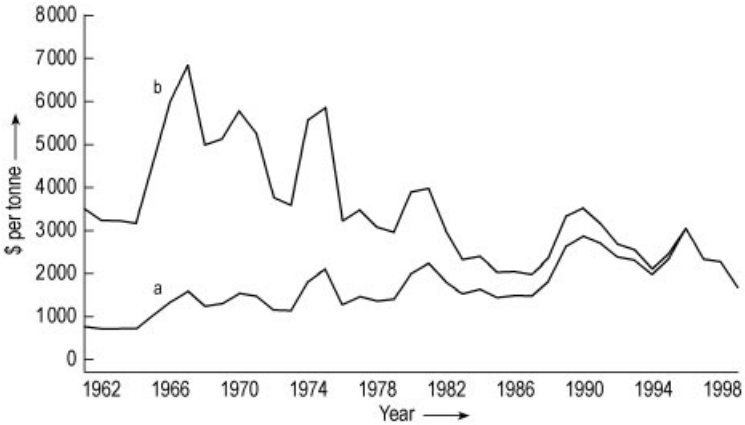

Figure 37. Development of copper prices since 1960
a) Cu price (current); b) Cu price (constant 1996)

circulation is used as much as possible in modern plants.

Hydrometallurgical operations for the extraction of copper from ores or concentrates present the risk of water pollution. These solutions, of various compositions, must be posttreated if they cannot be recycled. Such posttreatment consists of neutralization or precipitation of specific ions, chiefly anions bearing sulfur and the cations of heavy metals. Lime is an excellent precipitant. Thus, the sulfate ion in sulfuric acid solutions formed during hydrometallurgical extraction is precipitated as gypsum [208].

Solid-Waste Disposal. The following means are used for handling solid residues:

1) Recycling
2) Exploitation as raw material for the preparation of useful products
3) Dumping in deposits

One example of each method follows:

1) Flue dust from pyrometallurgical operations, e.g., from the Outokumpu flash smelting process, are added to the feed and recycled into suitable furnaces (Section 5.5.1) and occasionally into blast furnaces after agglomeration (Section 5.4.1). Zinc-containing flue dusts can be processed into zinc and zinc compounds.
2) Discarded slags with low copper content from some copper smelting processes can be sold after suitable treatment (Section 5.3.3).
3) Hydrometallurgical techniques yield various precipitates such as elemental sulfur or impure gypsum, which can easily be deposited.

10. Toxicology

Copper is a vital trace element for humans, most animals, and plants. The copper content of an adult human ranges from 50 to 120 mg. The average dietary intake of copper by adults ranges from 0.9 to 2.2 mg/d. For higher organisms, the compact metal is completely harmless. However Protista, especially bacteria, die in contact with metallic surfaces of copper and many of its alloys (oligodynamic effect) [209], [210].

Continued exposure to the metal dust or fumes can irritate mucous membranes. The following exposure limits have been established:

Form	Federal Republic of Germany	United States
Metal dust	MAK 1 mg/m^3	TLV/STEL 2 mg/m^3
Metal fumes	MAK 0.1 mg/m^3	TLV/TWA 0.2 mg/m^3

11. References

General References

1. V. Tafel: *Lehrbuch der Metallhüttenkunde,* 2nd ed., vol. 1, Hirzel, Leipzig 1951.
2. A. Butts: *Copper – The Science and Technology of the Metal, Its Alloys and Compounds,* Reinhold Publ. Co., New York 1954.
3. N. N. Muratsch: *Handbuch des Metallhüttenmannes,* vol. 1, VEB-Verlag Technik, Berlin 1954.
4. *Gmelin,* system no. 60, "Kupfer," Part A (1955).
5. H. Grothe (ed.): *Lueger-Lexikon der Technik,* 4th ed., vol. 5 (*Lexikon der Hüttentechnik*), Deutsche Verlagsanstalt, Stuttgart 1963.
6. A. Sutulov: *Copper Production in Russia,* University of Concepción, Chile, 1967.
7. R. P. Ehrlich: "Copper Metallurgy," *Symposium of the Metallurgical Society,* Denver/Colorado, Feb. 15 – 19, 1970.
8. M. J. Jones: "Advances in Extractive Metallurgy and Refining," *Symposium of the Institution of Mining and Metallurgy,* London, Oct. 4 – 6, 1971.
9. *Winnacker-Küchler,* 4th ed., **vol. 4,** p. 350.
10. A. Sutulov: *Copper Porphyries,* University of Utah, Salt Lake City 1974.
11. M. J. Jones: "Copper Metallurgy – Practice and Theory," *Symposium of the Institution of Mining and Metallurgy,* Brussels, Feb. 11, 1975.
12. J. C. Yannopoulos, J. C. Agarwal (eds.): "Pyrometallurgy and Electrolytic Refining,"*Extractive Metallurgy of Copper,* **vol. 1,** The Metallurgical Society of AIME, New York, N.Y., 1976; (AIME Annual Meeting, Las Vegas 1976).
13. A. K. Biswas, W. G. Davenport: *Extractive Metallurgy of Copper,* 3rd ed., Pergamon Press, Oxford-New York 1994.
14. Deutsches Kupfer-Institut: *Kupfer,* 2nd ed., Berlin 1980.
15. C. B. Gill: *Nonferrous Extractive Metallurgy,* J. Wiley & Sons, New York 1980.
16. D. B. George, J. C. Taylor (eds.): *Copper Smelting – An Update,* The Metallurgical Society of AIME, Warrendale, Pennsylvania, 1981.
17. E. G. West: *Copper and Its Alloys,* Ellis Horwood Ltd., Chichester, England, 1982.
18. T. K. Corwin, T. W. Devitt, M. A. Taft, A. C. Worrell: *International Technology for the Nonferrous Smelting Industry,* Noyes Data Corp., Park Ridge, New Jersey, 1982.
19. H. Y. Sohn, D. E. George, A. D. Zunkel (eds.): *Advances in Sulfide Smelting,* vol. 1 (*Basic Principles*), vol. 2 (*Technology and Practice*), The Metallurgical Society of AIME, Warrendale, Pennsylvania, 1983.
20. F. Pawlek: *Metallhüttenkunde,* De Gruyter, Berlin-New York 1983.

Specific References

21. L. Aitchison: *A History of Metals,* **vols. 1 and 2,** Mac Donald & Evans Ltd., London 1960.
22. R. F. Tylecote: *History of Metallurgy,* The Institute of Metals, London 1976.
23. Georgius Agricola: *Zwölf Bücher vom Berg- und Hüttenwesen* (1556), 5th ed., VDI-Verlag, Düsseldorf 1978.
24. L. Suhling: *Der Seigerhüttenprozeß,* Riederer-Verlag, Stuttgart 1976.
25. *Landolt-Börnstein,* **IV 2 b,** 639 – 648, 668 – 719.
26. Deutsche Ges. f. Metallkunde u. Verein Deutscher Ingenieure (eds): *Werkstoff-Handbuch Nichteisenmetalle,* 2nd ed., part 3, VDI-Verlag, Düsseldorf 1960.
27. American Society of Metals (eds.): *Metals Handbook,* 8th ed., vol. 1, Metals Park, Ohio, 1961.
28. K. Dies: *Kupfer und Kupferlegierungen in der Technik,* Springer Verlag, Berlin 1961.
29. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* 3rd ed., vol. 1, Springer Verlag, Berlin 1967.
30. Copper Development Assoc.: *CDA Technical Note TN 20: Copper Data,* London 1975.
31. J. O'M. Bockris, A. K. N. Reddy: *Modern Electrochemistry,* 6th ed., vols. 1 and 2, Plenum Press, New York 1977.
32. E. G. King, A. D. Mah, L. B. Pankratz: *Thermodynamic Data of Copper and Its Inorganic Compounds,* Int. Copper Res. Assoc., New York 1973.
33. H. Kaesche: *Die Korrosion der Metalle,* 2nd ed., Springer Verlag, Berlin 1979.
34. S. K. Coburn (ed.): *Corrosion Source Book,* American Society for Metals, Metals Park, Ohio, 1984.
35. M. Pourbaix: *Atlas of Electrochemical Equilibria in Aqueous Solutions,* Pergamon Press, Oxford 1966.
36. P. Klare: *Kupfer – Sauerstoff – Wasserstoff, Auszüge aus dem Schrifttum der letzten hundert Jahre,* Ges. Deutscher Metallhütten- und Bergleute, Clausthal-Zellerfeld 1962.

37. E. Fromm, E. Gebhardt (eds.): *Gase und Kohlenstoff in Metallen,* Springer Verlag, Berlin 1976.
38. Bundesanstalt für Bodenforschung, Hannover, Deutsches Institut für Wirtschaftsforschung, Berlin: *Untersuchungen über Angebot und Nachfrage mineralischer Rohstoffe,* vol. 2: Kupfer (1972).
39. The World of Metals: Copper, Metallgesellschaft, Frankfurt/Main 1993.
40. W. Gocht: *Handbuch für Metallmärkte,* 2nd ed., Springer Verlag, Berlin 1985.
41. US Bureau of Mines: Mineral Commodity Summaries, Washington D.C. 1992.
42. Z. S. Vukmanovic, *Metall (Berlin)* **38** (1984) 238.
43. Metallgesellschaft: *Mitteilungen aus dem Arbeitsbereich No. 18: Manganknollen – Metalle aus dem Meer,* Frankfurt/Main 1975.
44. D. Neuschütz, U. Scheffler, *Erzmetall* **30** (1977) 152 – 157.
45. Norddeutsche Affinerie: *Kupfer in Natur, Technik, Kunst und Wissenschaft,* Hamburg 1966.
46. K. S. E. Forssberg (ed.): *Flotation of Sulfide Minerals,* Elsevier Sci. Publ., New York 1985.
47. N. W. Kirshenbaum: *Transport and Handling of Sulfide Concentrates,* Stanford University, Department of Mineral Engineering, Standford, California, 1967.
48. M. S. Prasad: "Concentration Capacity in Treating Copper – Cobalt and Copper – Zinc Ores at Gecamines, Congo;" *Min. Eng.* **43** (1991) 129 – 133.
49. F. Pawlek, *Erzmetall* **22** (1969) 413 – 414.
50. M. Kaneko, *Eng. Min. J.* **175** (1974) no. 12, 61 – 64.
51. R. O. Thomas, D. W. Hopkins in [11], pp. 1 – 5.
52. H. Schackmann, *Erzmetall* **20** (1967) 499 – 511.
53. D. MacAskill, *Eng. Min. J.* **174** (1973) no. 7, 82 – 86.
54. S. D. Michaelson et al., *J. Met.* **18** (1966) 172 – 180.
55. H. Schlegel, A. Schüller, *Freiberg. Forschungsh. B* 1952, no. 2, 2 – 31.
56. W. A. Krivsky, R. Schuhmann, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **209** (1957) 981 – 988.
57. Y. A. Chang, Y. E. Lee, J. P. Neumann in [12]pp. 21 – 48.
58. F. Johannsen, H. Knahl, *Erzmetall* **16** (1963) 611 – 621.
59. E. M. Levin et al.: *Phase Diagrams for Ceramists,* vols. 1 and 2, The American Ceramic Soc., Columbus, Ohio, 1964; Suppl. 1969.
60. P. Röntgen, R. Winterhager, R. Kammel, *Erzmetall* **9** (1956) 207 – 214.
61. W. Wiese, *Erzmetall* **17** (1964) 298 – 305.
62. J. M. Floyd, P. J. Mackey: *Extraction Metallurgy '81 (Symposium)* , The Institution of Mining and Metallurgy, London 1981.
63. C. Diaz: "The Thermodynamic Properties of Copper-Slag Systems," *INCRA Monogr.* III, USA, 1974.
64. K. N. Subramanian, N. J. Themelis, *J. Met.* **24** (1972) no. 4, 33 – 38.
65. B. T. Andersson, P. Paarni in [19], p. 1005.
66. J. G. Eacott in [19], p. 583.
67. H. H. Kellogg, *Eng. Min. J.* **178** (1977) no. 4, 61 – 64.
68. H. Hilbrans, P. Paschen, *Erzmetall* **34** (1981) 639 – 644.
69. Ch. H. Pitt, M. E. Wadsworth, *J. Met.* **33** (1981) 25 – 34.
70. F. Habashi, R. Dugdale: *Research Reports 1969/70,* The Anaconda Co., Tucson, Arizona, 1970.
71. F. Habashi, B. I. Yostos, *J. Met.* **29** (1977) no. 7, 11 – 16.
72. R. S. Olsen et al., *Trans. Soc. Min. Eng. AIME* **254** (1973) no. 4, 301 – 305.
73. K. Parameswaran et al. in [16], pp. 323 – 339.
74. G. Dobbert, W. Wiese: *Periodische Wechsel-Reduktionselektrolyse von Spurstein unter Gewinnung von umformfähigem Elektrolytkupfer und Elementarschwefel,* Forschungsberichte Nordrhein-Westfalen, Westdeutscher Verlag, Opladen 1977.
75. A. H. Kinneberg in [7], pp. 173 – 176.
76. Onahama Smelting & Refining Co., US 4 001 013, 1977 (M. Goto et al.).
77. K. Cormack, *Eng. Min. J.* **185** (1985) no. 6, 44 – 48.
78. Ch. Baiqi, T. Xiangtin, J. Xigen, M. Yuebo: *Bai-Yin Copper Smelting Process,* Paper of the Bai-Yin Non-Ferrous Metals Co., China (1984).
79. M. Yasuda, T. Yuki, M. Kato, Y. Kawasaki in [16], pp. 251 – 263.
80. S. Okada, M. Miyake, A. Hara, M. Uekawa in [19], pp. 855 – 874.
81. Outokumpu Oy, DE 2 536 392, 1975 (O. A. Aaltonen, B. T. Andersson et al.).
82. T. N. Antonioni, C. M. Diaz, H. C. Garven, C. A. Landolt in [16], pp. 17 – 31.
83. K. I. Ushakov et al., *Tsvetn. Met. (N.Y.)* **16** (1975) no. 2, 5 – 9.
84. G. Melcher, E. Müller, H. Weigel, *J. Met.* **28** (1976) 4 – 8.

85. Humboldt Wedag's Cyclone, *Eng. Min. J.* **178** (1977) no. 10, 45, 49.
86. G. Berndt, K. Emicke: *Extraction Metallurgy '85,* The Institution of Mining and Metallurgy, London 1985.
87. F. E. Lathe, L. Hodnett: "Data on Copper Converter Practice in Various Countries," *Trans. TMS-AIME* **212** (1958) 603 – 617.
88. R. E. Johnson (ed.): *Copper and Nickel Converters,* The Metallurgical Soc. of AIME, Warrendale, Pennsylvania, 1979.
89. F. Johannsen, H. Vollmer, *Erzmetall* **13** (1960) 313.
90. P. J. Mackey, P. Tarassoff in [19], p. 408.
91. J. Leroy, P. J. Lenoir: "Hoboken Type of Copper Converter and Its Operation," *Institution of Mining and Metallurgy, Symposium,* London, April 1967 (Paper 15).
92. J. D. Gomez in [87], pp. 291 – 311.
93. Inspiration Consolidated Copper Co., GB 2 089 011 A, 1981 (A. F. Tittes et al.).
94. R. Campos, C. Queirolo in [87], pp. 257 – 273.
95. G. Lindquist, P.-L. Nystedt, S. Petersson in [16], pp. 41 – 49.
96. U. Kuxmann, *Erzmetall* **27** (1974) 55 – 64.
97. J. M. Floyd, N. C. Grave, B. W. Lightfoot: *Extractive Metallurgy Symposium 1980,* The Australian Institute of Mining and Metallurgy, Melbourne, Australia, 1980, pp. 63 – 74.
98. D. A. Diomidovskii et al., *Tsvetn. Met. (Moscow)* **32** (1959) no. 2, 27 – 34.
99. F. Sehnalek, J. Holeczy, J. Schmiedl, *J. Met.* **16** (1964) 416 – 420.
100. K. J. Richards, D. G. George, L. K. Bailey in [19], pp. 489 – 498.
101. M. Goto, N. Kikumoto: "Process Analysis of Mitsubishi Continuous Smelting and Converting Process," *110th AIME Annual Meeting,* Chicago, Febr. 1981.
102. T. Suzuki, T. Nagano: "Development of New Continuous Copper Smelting Process," *Jt. MMIJ–AIME Meet.* Tokyo, May 1972.
103. T. Suzuki et al. in [16], p. 60.
104. M. P. Amsden, R. M. Sweetin, D. G. Treilhard, *J. Met.* **30** (1974) no. 7, 16 – 26.
105. J. B. W. Bailey, A. G. Storey: "The Noranda Process after Six Years' Operation," *18th Annual CIM Conference of Metallurgists,* Sudbury, Ontario, Aug. 1979.
106. U. Kuxmann, *Erzmetall* **27** (1974) 55 – 64.
107. P. J. Mackey, G. C. McKerrow, P. Tarassoff: "Minor Elements in the Noranda Process," *104th AIME Annual Meeting,* New York, Febr. 1975.
108. H. K. Worner, *Eng. Min. J.* **172** (1971) no. 8, 64 – 68.
109. R. Schuhmann, P. E. Queneau: "Thermodynamics of the Q – S Oxygen Process for Coppermaking," *AIME Annual Meeting,* Las Vegas 1976.
110. N. Torres, G. Melcher, *CIM Bull.* **77** (1984) no. 871, 86 – 91.
111. G. J. Brittingham, DE 1 280 479, 1964.
112. R. B. Worthington, *Rep. Invest. U.S. Bur. Mines* 1973, 7705.
113. G. Fleischer, R. Kammel, U. Lembke, *Metall (Berlin)* **32** (1978) 29 – 34.
114. W. S. Nelmes, *Trans. Inst. Min. Metall. Sect. C* **93** (1984) 180 – 192.
115. W. S. Nelmes, J. A. Charles, A. G. Cowen, *J. Met.* **13** (1961) 216 – 220.
116. J. S. Jacobi, *J. Met.* **32** (1980) no. 2, 10 – 14.
117. W. Schwartz, *Metall (Berlin)* **34** (1980) 121 – 124.
118. H. P. Rajcevic, W. R. Opie, *J. Met.* **34** (1982) no. 3, 54 – 56.
119. R. Hubrich, K. Hein, *Neue Hütte* **28** (1983) 452 – 456.
120. A. Bahr, *Erzmetall* **33** (1980) 324 – 330.
121. J. Julius, *Metall (Berlin)* **38** (1984) 758 – 762.
122. J. C. Yannopoulos, J. C. Agarwal (eds.): "Hydrometallurgy and Electrowinning,*Extractive Metallurgy of Copper,* **vol. 2,** " The Metallurgical Society of AIME, New York, N.Y., 1976 (AIME Annual Meeting, Las Vegas 1976).
123. L. A. Haas, R. Weir (eds.): "Hydrometallurgy of Copper, its Byproducts and Rarer Metals," The Metallurgical Society of AIME, New York 1983.
124. A Pinches et al., *Hydrometallurgy* **2** (1976) 87 – 103.
125. W. Grote et al., *Proc. Australas. Inst. Min. Metall.* **278** (1981) 37 – 40.
126. D. S. Flett, *Trans. Inst. Min. Metall. Sect. C* **83** (1974) 30 – 38.
127. M. G. Atmore, K. J. Severs, R. B. G. Voyzey: "Past, Present and Future of Solvent Extraction of Copper," *Inst. Min. Metall. and Chinese Soc. of Metals, Internat., Conference,* Kunming, Yunnan, China, Oct. 1984.
128. R. R. Greenstead, *J. Met.* **31** (1979) no. 3, 13 – 16.
129. R. Kammel, H.-W. Lieber, *Galvanotechnik* **68** (1977) 413 – 418.
130. A. E. Back, *J. Met.* **19** (1967) 27 – 29.
131. W. R. Hopkins, G. Eggett, J. B. Scuffham: "Electrowinning of Copper from Solvent Extraction Electrolytes – Problems and

Possibilities," in: *International Symposium on Hydrometallurgy* AIME, New York 1973, pp. 127 – 144.
132. D. J. Robinson, St. E. James (eds.): *Anodes for Electrowinning,* TMS – AIME, Warrendale, Pennsylvania, 1984.
133. W. C. Cooper: "Recent Advances and Future Prospects in Copper Electrowinning," *22nd Annual CIM Conference of Metallurgists,* Edmonton, Alberta, Aug. 1983.
134. *Eng. Min. J.* **168** (1967) no. 1, 97 – 100.
135. W. A. Griffith, H. E. Day, T. S. Jordon, V. C. Nyman, *J. Met.* **27** (1975) no. 2, 17 – 25.
136. R. G. Bautista (ed.): *Hydrometallurgical Process Fundamentals,* Plenum Press, New York 1984.
137. K. Osseo-Asare, J. D. Miller (eds.): "Hydrometallurgy – Research, Development and Plant Practice," *12th AIME Annual Meeting,* Atlanta, Georgia, March 1983.
138. W. R. Hopkins, A. J. Lynch, *Eng. Min. J.* **178** (1977) no. 2, 56 – 64.
139. W. Schwartz, R. Michels, *Erzmetall* **17** (1964) 117 – 124.
140. M. C. Kuhn, N. Arbiter, H. Kling, *CIM Bull.* **67** (1974) no. 742, 62 – 73.
141. P. R. Kruesi, E. S. Allen, J. L. Lake, *CIM Bull.* **66** (1973) no. 734, 81 – 87.
142. G. E. Atwood, R. W. Livingston, *Erzmetall* **33** (1980) 251 – 255.
143. J. J. Oudiz, *J. Met.* **25** (1973) no. 12, 35 – 38.
144. R. Henych, F. Kadlec, V. Sedlacek, *J. Met.* **17** (1965) 386 – 388.
145. J. Osterwald, H. Sadat-Darbandi, *Metall (Berlin)* **30** (1976) 1057 – 1058.
146. G. Melcher, F. Sauert, *Erzmetall* **33** (1980) 451 – 455.
147. G. Kapell, W. Leutloff, *J. Met.* **32** (1980) no. 9, 36 – 40.
148. J. M. Dompas, *Min. Mag. (London)* (1983) no. 9, 169 – 175.
149. V. T. Isakov: *The Electrolytic Refining of Copper,* Technocopy, Stonehouse, Glos., England 1973.
150. GDMB Gesellschaft Deutscher Metallhütten- und Bergleute (ed.): *Elektrolyse der Nichteisenmetalle,* Verlag Chemie, Weinheim-Deerfield Beach-Basel 1982.
151. A. P. Brown et al., *J. Met.* **33** (1981) no. 7, 49 – 57.
152. H. Ykeda, Y. Matsubara, in [12], p. 601 (1970).
153. I. Fujimura, A. Katan. The Met. Soc. AIME, TMS Paper Selection A 82 – 12, 1982.
154. J. Sato, T. Imamura, M. Hojo, T. Suzuku, *Nippon Kogyo Kaishi* **1046** (1975) no. 97, 258.
155. J. C. Jenkins: "Copper Tank House Technology Reviewed and Assessed," *Symposium May 1985, Australas. Inst. Min. Metall.,* Victoria, Australia.
156. M. J. Jaskula, *Electrochim. Acta* **28** (1983) 1395 – 1406.
157. T. B. Braun, *J. Met.* **33** (1981) no. 2, 59 – 67.
158. B. Rühl in: Metallgesellschaft AG, Frankfurt a. M., Review of the Activities no. 11 (1968) p. 48.
159. Amarillo Copper Refining, *Eng. Min. J.* **182** (1981) no. 9, 67 – 78.
160. Wennberg, DE 2 618 679, 1976 (R. Bengtsson).
161. *Eng. Min. J.* **176** (1975) no. 4, 101.
162. W. W. Harvey, M. R. Randlett, K. I. Bangerskis, *J. Met.* **27** (1975) no. 7, 19 – 25.
163. R. Winand, *Trans. Inst. Min. Metall. Sect. C* **84** (1975) 67 – 75.
164. K. Rinne, in [150] pp. 72 – 76 (1982).
165. M. P. Amsten et al., *J. Met.* **30** (1978) no. 7, 16 – 26.
166. J. Bertha, J. Schwimann, H. Wöbking, H. Wörz, *Erzmetall* **32** (1979) 335 – 337.
167. I. J. Perry, J. O'Kane: A Review of Five Years of Commercial Operations of the ISA Electrorefining Process at Townsville, Australia," *Jt. GDMB – IMM Symposium,* Hamburg, Oct. 1983.
168. Asarco Inc., US 3 199 977, 1965 (A. J. Phillips, R. Baier).
169. C. L. Thomas, *Metal Bull. Monthly* (August 1983), 17 – 22.
170. K. Sczimarowsky, *Draht-Welt* **46** (1960) 329 – 335.
171. E. Herrmann: *Handbook on Continuous Casting,* Aluminium-Verlag, Düsseldorf 1980.
172. L. W. Collins, Jr. (ed.): *Nonferrous Wire Handbook,* vol. 1: "Nonferrous Wire Rod," The Wire Assoc. Int. Inc., Guilford, Connecticut 1977.
173. H.-D. Hirschfelder, *Draht* **29** (1978) 164 – 170.
174. L. Properzi, A. Ossani, *Draht-Welt* **66** (1980) 456 – 458.
175. E. H. Chia, R. D. Adams, *J. Met.* **33** (1981) no. 2, 68 – 74.
176. P. Wincierz, *Z. Metallkd.* **66** (1975) 235 – 248.
177. I. A. Dundurs, D. W. Hoey, C. W. Walter, *Wire J.* **13** (1980) no. 9, 120 – 125.
178. J.-P. Dosdat, J. M. Dompas, *Wire J.* **13** (1980) no. 8, 90–95.
179. B. Hansson, K.-G. Soderlund, E. Martinsson, *J. Inst. Met.* **98** (1970) 234 – 237.
180. M. Rantanen, *Wire J.* **13** (1980) no. 3, 102 – 104.

181. E. Peissker, *Int. J. Powder Metall. Powder Technol.* **20** (1984) no. 2, 87 – 101.
182. *1984 Annual Book of ASTM Standards,* **vol. 02.**01: "Copper and Copper Alloys," ASTM, Philadelphia (published annually).
183. *DIN Taschenbuch 26: Nichteisenmetalle, Normen über Kupfer und Kupferknetlegierungen,* 3rd ed., Beuth-Verlag, Berlin 1984.
184. ISO Technical Report 4745 (E),High Conductivity, Copper – Spiral Elongation Test (1978).
185. GDMB and VDEh (eds.): "Schiedsanalysen," *Analyse der Metalle,* **vol. 1,**" Springer Verlag, Berlin 1966; Supplement volume (Erg.-Bd.) 1980.
186. C. Engelmann, G. Kraft, J. Pauwels, C. Vandecasteele: *Modern Methods for the Determination of Non-Metals in Non-Ferrous Metals,* De Gruyter, Berlin-New York 1985.
187. M. Bauser, *Metall (Berlin)* **38** (1984) 513 – 516.
188. E. Tuschy, *Z. Metallkd.* **61** (1970) 488 – 492.
189. *Source Book on Copper and Copper Alloys,* American Soc. of Metals, Metals Park, Ohio, 1979.
190. Source: Commodities Research Unit, London.
191. Source: Copper Development Assoc. Inc., New York.
192. Metallgesellschaft AG (ed.): *Metallstatistik 1974 –1984,* vol. 72, Frankfurt/M. 1985.
193. World Bureau of Metal Statistics (ed.): *World Metal Statistics 1984,* London.
194. Service Etudes et Statistiques (ed.): *Annuaire Minemet 1984, Statistiques Annual, Groupe Metal,* Paris.
195. American Metal Market (ed.): *Metal Statistics 1984, The Purchaising Guide of the Metal Industries,* New York.
196. C. J. Schmitz: *World Nonferrous Metal Production and Prices 1700 – 1976,* Frank Cass & Co., London 1979.
197. H. P. Münster, G. Kirchner: *Taschenbuch des Metallhandels,* 7th ed., Metall Verlag, Berlin-Heidelberg 1982.
198. R. F. Mikesell: *The World Copper Industry – Structure and Economic Analysis,* J. Hopkins University Press, Baltimore-London 1979.
199. Metal Bulletin Conferences Ltd.: *Metal Bulletin's 2nd International Copper Conference, London, April 1984,* Surrey, England.
200. *Metall Bulletin Handbook '85,* Metal Bulletin Books Ltd., Surrey, England.
201. R. M. Serjeantson (ed.): *Non-Ferrous Metal Works of the World,* Met. Bull. Books Ltd., London 1985.
202. *Mining Companies of the World,* Mining Journal Books Ltd., London 1975.
203. J. G. Eacott in [16], p. 101.
204. T. D. Chatwin, N. Kikumoto (eds.): *Sulfur Dioxide Control in Pyrometallurgy,* Metallurg. Soc. AIME, Warrendale, Pennsylvania, 1982.
205. CS Survey: Sulfur Dioxide Emission Control in the Copper Industry, *Copper Studies* **12** (1985) no. 12 (June), 1 – 3, Commodities Research Unit Ltd., London.
206. L. J. Friedman in [19], pp. 1023 – 1040.
207. I. J. Weisenberg, P. S. Bakshi, A. E. Vervaert, *J. Met.* **31** (1979) no. 10, 38 – 44.
208. J. G. Eacott in [16], pp. 127 – 128.
209. *Commitee on Medical and Biological Effects of Environmental Pollutants: Copper,* Nat. Academy of Sci., Washington 1977.
210. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A,** Wiley-Interscience, New York 1981, p. 1620.
211. Th. Lehner, P. O. Lindgren, *Erzmetall* **46** (1993) 105 – 112.
212. M. Cirkovic, J. Marinkovic, D. Vucurovic, S. Stopic, *EPD Congress 1999,* The Minerals, Metals & Materials Society, Warrendale 1999, pp. 933 – 944.
213. E. H. Smith, J. W. Foster, Ph. Minet, Ph. Cauwe: Complex Sulfides – Processing of Ores Concentrates and By-Products," *TMS – AIME Annual Meet.* (1985) 421 – 440.
214. J. Czernecki, Z. Smieszek, St. Gizicki, J. Dobrzanski, M. Warmuz, *Sulfide Smelting '98* Current and Future Practices, TMS Warrendale (1998) pp. 315 – 343.
215. P. A. Gajardo, Copper 87, **vol. 4 Pyrometallurgy of Copper,** Alfabeta Impresores Santiago, Chile 1987, pp. 111 – 121.
216. E. N. Mounsey, K. R. Robilliard, *JOM* **46** (1994) 58 – 60.
217. R. L. Player, C. R. Fountain, T. V. Nguyen, F. R. Jorgensen, "International Symposium on Bath Smelting," TMS, Warrendale (1992) pp. 215 – 229.
218. L. A. Mills, D. D. Hallet, C. J. Newman, "Extractive Metallurgy of Copper," **vol. 1,** TMS, Warrendale 1976 , pp. 458 – 487.
219. Y. Prevost, "International Symposium on Bath Smelting," TMS, Warrendale (1992) pp. 189 – 202.

220. R. J. Campos, J. O. Achurra, O. C. Rojas, "Copper 91," *Pyrometallurgy of Copper,* **vol. 4,** Pergamon Press, New York 1991 , pp. 229 – 246.
221. R. Campos, L. Torres, "Extractive Metallurgy of Copper, Nickel and Cobalt," TMS, Warrendale 1993 , pp. 1441 – 1460.
222. H. H. Kellog, C. Diaz, "International Symposium on Bath Smelting," TMS, Warrendale (1992) pp. 39 – 63.
223. A. K. Biswas, W. G. Davenport, *Extractive Metallurgy of Copper,* Pergamon Press, New York 1994 , pp. 184 – 187.
224. T. P. T. Hanniala, J. S. Sulanto, "The Development Trends of the Outokumpu Flash Smelting Process for the Year 2000, TMS, Warrendale (1989).
225. T. Inami, K. Baba, H. Kurokawa, K. Nagai, Y. Kondo, *Copper 91,* **vol. 4,** Pergamon Press, New York 1991 , pp. 49 – 63.
226. A. K. Espelata, J. Hino, A. Yazawa, *Copper 91,* **vol. 4,** Pergamon Press, New York 1991 , pp. 109 – 123.
227. K. Sasaki, "Extractive Metallurgy of Copper, Nickel and Cobalt," **vol. 2,** TMS, Warrendale 1993 , pp. 1377 – 1385.
228. P. Willbrandt, "Extractive Metallurgy of Copper, Nickel and Cobalt," **vol. 2,** TMS, Warrendale 1993 , pp. 1361 – 1376.
229. B. G. Belew, E. H. Partelpoeg, Paper presented at the TMS annual meeting 1993.
230. M. Brueggemann, E. Caba, "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 159 – 166.
231. F. Sauert, L. W. Castor, S. S. Jones, Recent Developments in Non Ferrous Pyrometallurgy, paper 16.4, CIM Montreal (1994).
232. K. Fukushima, K. Baba, H. Kurokawa, M. Yamagiwa, "Process Control and Automation in Extractive Metallurgy," TMS, Warrendale (1998) pp. 113 – 130.
233. T. Maruyama, T. Saito, M. Kato, "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 219 – 226.
234. W. Persson, W. Wendt, S. Demetrio: *Copper 99,* **vol. 5 Smelting Operations and Advances,** TMS, Warrendale 1999, pp. 491 – 503.
235. A. A. Bustos, J. K. Brimacombe, G. G. Richards, A. Vahed, A. Pelletier, "Copper 87," *Pyrometallurgy of Copper,* **vol. 4,** Alfabeta Impresores, Santiago de Chile 1987 , pp. 347 – 373.
236. A. A. Bustos, J. K. Brimacombe, G. G. Richards, N. B. Gray, *Metallurgical Tranactions B,* **15 B** (1984) 77 – 89.
237. T. Shibata, Y. Oda, Paper presented at the Sixth International Flash Smelter Congress, Brazil (1990).
238. T. Inami, K. Baba, Y. Ojima, Paper presented at the Sixth International Flash Smelter Congress, Brazil (1990).
239. Noranda, US 4 544 141, 1985 and US 4 504 309, 1985 (P. J. Mackey, B. W. Baily).
240. M. Boisvert et al., "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 569 – 583.
241. T. Shibasaki, M. Hayashi, Y. Nishiyama, "Extractive Metallurgy of Copper, Nickel and Cobalt," *Copper and Nickel Smelting Operations,* **vol. 2,** TMS, Warrendale 1993 , pp. 1413 – 1428.
242. T. Shibasaki, M. Hayashi, "International Symposium on Injection in Process Metallurgy," TMS, Warrendale 1991 , pp. 199 – 213.
243. E. Oshima, T. Igarashi, N. Hasegawa, H. Kumada, "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 597 – 606.
244. K. J. Richards, D. B. George, L. K. Bailey, "Advances in Sulfide Smelting," *Technology and Practice,* **vol. 2,** TMS, Warrendale 1993 , pp. 489 – 498.
245. C. J. Newman, T. I. Probert, A. J. Weddick, "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 205 – 215.
246. Z. Smieszek, S. Sedzik, W. Grabowski, S. Musial, S. Sobierajski, *Extractive Metallurgy 85* (1985) 1049 – 1056.
247. J. Czernecki, Z. Smieszek, St. Gizicki, J. Dobrzanski, M. Warmuz, "Sulfide Smelting '98," *Current and Future Practices,* TMS, Warrendale 1998 , pp. 315 – 343.
248. P. E. Queneau, "Extractive Metallurgy of Copper, Nickel and Cobalt, *Fundamental Aspects,* **vol. 1,** TMS, Warrendale 1993), pp. 447 – 471.
249. E. Arpaci, T. Vendura, *Metall* **47** (1993) 340 – 345.
250. B. E. Langner, in W. Nickel (ed.): *Recycling Handbuch,* VDI Verlag, 1996, pp. 286 – 301.
251. K. Göckmann, *CIM Bulletin* **85** (1992) 150 – 165.
252. B. E. Langner, *Metall* **48** (1994) 880 – 885.
253. K. Hanush, H. Bussmann, *TMS 1995* (1995) 171 – 188.

254. G. Leuprecht, "Messungen in einer Kupferelektrolyse," *Schriftenreihe der GDMB* **74,** 102 – 119.
255. M. Landau, H. Traulsen, *Proceedings of Copper 91,* Pergamon Press, New York 1991.
256. Norddeutsche Affinerie AG, EP 0294001B1, 1984 (B. Langner, P. Stantke).
257. J. H. Schloen, *Copper 91,* **vol. 3,** Pergamon Press, New York 1991 , pp. 491 – 507.
258. S. J. Kohut, J. J. Pio, M. D. Precup: *Copper 99,* **vol. 3 Electrorefining and Electrowinning of Copper,** TMS, Warrendale 1999, pp. 547 – 559.

Copper Alloys

Clarence A. Isbell, Jr., (Olin Brass Corp.), Florissant, Missouri 63034, United States

1. Introduction 1
2. Wrought Coppers 3
2.1. Coppers, ≧ 99.3 % Cu 3
2.2. High-Copper Alloys, ≧96.0 % Cu . . 5
3. Wrought Brasses 11
4. Wrought Leaded Brasses 13
5. Wrought Tin Brasses 15
6. Wrought Tin Bronzes and Brazing Alloys 16
6.1. Phosphor Bronzes 17
6.2. Leaded Phosphor Bronzes 19
6.3. Brazing Alloys 19
7. Wrought Specialty Alloys 20
7.1. Copper–Aluminum Alloys (Aluminum Bronzes) 21
7.2. Copper – Silicon Alloys (Silicon Bronzes) 22
7.3. Miscellaneous Copper – Zinc Alloys 22
8. Wrought Copper – Nickel Alloys . . 24
8.1. Cupronickels 26
8.2. Copper – Nickel – Zinc Alloys 27
9. Casting Alloys 29
9.1. Cast Coppers 31
9.2. High-Copper Alloys 31
9.3. Copper – Tin – Zinc and Copper – Tin – Zinc – Lead Alloys 31
9.4. Copper – Manganese Alloys 32
9.5. Copper – Zinc – Silicon Alloys 34
9.6. Copper – Tin Alloys 35
9.7. Copper – Tin – Lead Alloys 36
9.8. Copper – Tin – Nickel Alloys 38
9.9. Copper – Aluminum – Iron and Copper – Aluminum – Iron – Nickel Alloys 38
9.10. Copper – Nickel – Iron Alloys 41
9.11. Copper – Nickel – Zinc – Tin – Lead Alloys 41
9.12. Special Cast Alloys 41
10. Master Alloys 41
11. Safety and EnvironmentalAspects . 43
12. Economic Aspects 44
13. References 44

1. Introduction

Copper alloys are of ancient origin, being the source of tools that led humans from the Stone Age into the Bronze Age. Over the centuries, people have learned to refine metals to a more pure state and alloy these metals together to form a more useful product.

As techniques improved, the number of copper alloys proliferated and became more complicated. In parallel, the alloy systems of other metals were also evolving and increasing in number. Each manufacturer had devised a nomenclature that was specifically applicable to his alloys, making it difficult for users to purchase from more than one source and still obtain equivalent products.

In the mid-1950s, the Aluminum Association established a four-digit numbering system to define the various aluminum alloys. The steel industry already had a four-digit system to describe many of its alloys. In the mid-1960s, the copper industry decided to adopt a three-digit system. There was a spurt of research activity in the copper industry, especially on aluminum bronzes, which became difficult to administer with the three-digit alloy identification system. Efforts were also being made internationally to develop a unified method of systematically identifying alloys. In the United States, a *Unified Numbering System* (UNS) was devised, while in Europe the International Standards Organization (ISO) began assembling an identification system.

The UNS designation system for metals and alloys in North America has been developed under the joint management of the American Society for Testing and Materials (ASTM) and the Society of Automotive Engineers (SAE). This designation system incorporates an alphanumeric identification that encompasses the commercially produced metals and alloys. The International Standards Organization (ISO) uses a designation that is based on composition (e.g., cartridge brass, CuZn3O), but as the number of specific alloying elements increases, the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_525

ISO designation becomes unwieldy, e.g., UNS C86300 vs. ISO CuZn25Al6Fe3Mn3.

The initial letter of the UNS system designates the metal and alloy system covered, e.g., A for aluminum and aluminum alloys, P for precious metals and alloys, N for nickel and nickel alloys, or C for copper and copper alloys. The copper and copper alloys system is administered by the Copper Development Association (CDA) and encompasses all coppers and copper alloys of commerical use that have been presented for inclusion in this system. The numeric portion of the alphanumeric UNS system classifies the various alloys into specific subgroupings on the basis of the quantity of copper present and the alloying elements present. These subgroups are identified as follows:

C10000	Wrought coppers
C20000	Wrought copper – zinc alloys (wrought brasses)
C30000	Wrought copper – zinc – lead alloys (wrought leaded brasses)
C40000	Wrought copper – zinc – tin alloys (wrought tin brasses)
C50000	Wrought copper – tin alloys (standard phosphor bronzes), wrought copper – tin – lead alloys (leaded phosphor bronzes), and copper – phosphorus alloys (brazing alloys)
C60000	Wrought speciality alloys: wrought copper – aluminum alloys (aluminum bronzes), wrought copper – silicon alloys (silicon bronzes), and miscellaneous copper – zinc alloys (manganese brasses and aluminum brasses)
C70000	Wrought copper – nickel alloys (cupronickels), wrought copper – nickel – zinc alloys (nickel silvers), wrought leaded copper – nickel – zinc alloys (leaded nickel silvers)
C80000	Cast coppers, cast high-copper alloys, cast copper – tin – zinc alloys, cast copper – tin – zinc – lead alloys (unleaded and leaded red brasses, semi-red brasses, yellow brasses, and manganese brasses), cast copper – zinc – silicon alloys (silicon bronzes and silicon brasses)
C90000	Cast copper – tin alloys (tin bronzes), cast copper – tin – lead alloys (leaded tin bronzes), cast copper – tin – nickel alloys (nickel tin bronzes), cast copper – aluminum – iron and copper – aluminum – iron – nickel alloys (cast aluminum bronzes), cast copper – nickel – iron alloys (copper nickels), cast copper – nickel – zinc alloys (nickel silvers), and cast special alloys

This classification system is the basis for discussing commercial coppers and copper alloys in an orderly manner.

The *relative electrical conductivity* is given in %IACS: In 1913, in France, an electrical standard was adopted on an international basis. A copper wire 1 m long and weighing 1 g, with an electrical resistance of 0.15328 Ω, was selected as the standard. A value of 100 % electrical conductivity was assigned to this wire, which has a volume resistivity of 0.017241 Ω mm^2 m at 20 °C. This is the International Annealed Copper Standard (IACS).

Some coppers may have a greater electrical conductivity than 100 %. Alloy C11000 has 101 % IACS due to its higher purity. Figure 1 illustrates how many of the common elements associated with copper, even at minute levels, degrade the electrical conductivity. Most high-conductivity copper alloys start with electrolytic copper as the basic building block for developing the alloy.

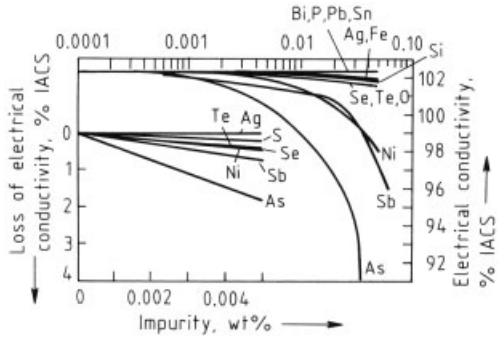

Figure 1. Effect of impurity levels on the electrical conductivity of copper [28, p. 66]

2. Wrought Coppers

The wrought coppers are further subdivided into two classes: (1) the coppers, i.e., copper alloys that are designated as C10000 through C15600 and have a minimum of 99.3 wt % Cu, and (2) high-copper alloys, which are designated as C16000 through C19600 and have a minimum of 96.0 wt % Cu.

2.1. Coppers, ≧ 99.3 % Cu

In coppers, there are several elements that are added for specific purposes and one element that is specifically excluded – oxygen [*7782-44-7*]. Unless special precautions are taken to prevent contact with air when the copper is molten, oxygen is absorbed into the liquid copper. It remains there unless the copper is deoxidized. The presence of oxygen in copper has several deleterious effects:

1) Hydrogen embrittlement is a probability during brazing or welding, or any time copper is heated above 450 °C and hydrogen is present in the atmosphere or environment.
2) The electrical conductivity is somewhat lowered and elements used to deoxidize the copper lower the conductivity even more.
3) Oxides present act as inclusions that reduce formability. Rolling further accentuates the effect of the oxides.
   When the oxides are randomly spaced throughout the matrix, the effect on the formability is relatively low. When processing causes these oxides to agglomerate within the matrix in the form of a string, however, formability is greatly reduced. This condition most often has its origin in the original casting of the alloy, where agglomeration is permitted to begin.
4) Copper oxides within the structure interfere with recrystallization and grain growth during annealing.

The oxygen problem is handled by one of three methods:

1) Processing the liquid copper in such a manner as to exclude oxygen through the use of vacuum chambers or reducing or inert atmospheres during melting, pouring, and casting. However, this does not preclude the absorption of oxygen at some later time (annealing, welding, etc.) unless these same precautions are followed.
2) Using normal melting and casting techniques, but adding deoxidizers to the melt to combine with the oxygen. With excess deoxidizer present, the chemical reaction is driven to completion and offers protection during subsequent processing. The common deoxidizers are phosphorus, boron, zirconium, and magnesium.
3) Recognizing that oxygen is present but working within the limitations imposed by the oxygen during subsequent processing.

Other elements that may be present as impurities are removed normally through fire and electrolytic refining processes, leaving very pure (less than 0.10 % other elements) copper.

In the coppers (≧ 99.3 wt % Cu), several alloying elements are used to provide specific attributes.

Silver [*7440-22-4*] is alloyed with copper to increase the softening temperature of rolled temper products. When silver is present in sufficient quantity, a rolled temper copper can be subjected to higher processing temperatures for longer times without softening the metal excessively. Figures 2 and 3 illustrate the effect of silver content on the softening characteristics of silver-bearing copper strip that has been reduced 40 % in thickness by cold rolling. Figure 2 illustrates that the quantity of silver alloyed with copper affects the amount of softening with time at 290 °C. Figure 3 illustrates the effect at 340 °C for the same group of silver-bearing coppers.

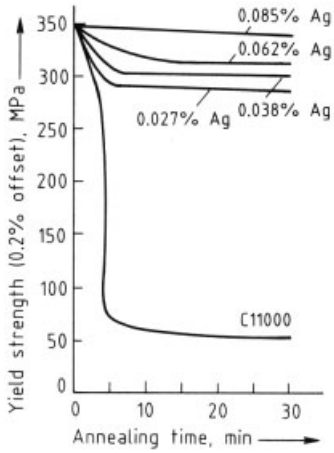

Figure 2. Effect of silver additions (wt %) on the annealing of copper following 40 % cold work [26], [27]
Annealing temperature, 290 °C

Phosphorus [*7723-14-0*] is added to molten copper as a deoxidizer. When normal control of oxygen content is exercised during the melting of copper, a minimum of 0.015 % P should be sufficient to deoxidize the melt completely. The copper should then be capable of passing the standard hydrogen embrittlement tests. Deoxidized high-phosphorus (DHP) copper (C12200) contains 0.015 –0.040 wt % P. Its electrical conductivity is approximately 85 % IACS. The high phosphorus leaves sufficient deoxidizing power to maintain deoxidizing during welding or brazing of the copper. With careful control of oxygen content during melting, a phosphorus residual in the range of 0.004 – 0.012 wt % can be achieved to produce deoxidized low-phosphorus (DLP) copper (C12000). Since part of the phosphorus analyzed is present as oxide, C12000

probably would fail the hydrogen embrittlement tests when the phosphorus is kept low to maximize the electrical conductivity (ca. 95 % ACS). If the phosphorus is kept toward the high side of the phosphorus range to meet the hydrogen embrittlement test standards, the electrical conductivity is reduced to ca. 90 % IACS.

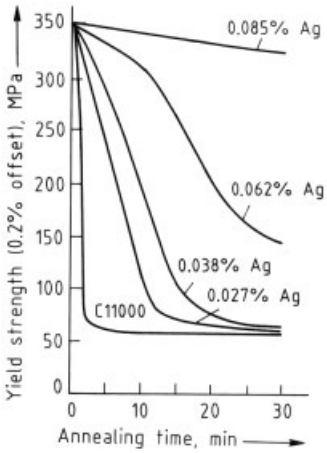

Figure 3. Effect of silver additions (wt %) on the annealing of copper following 40 % cold work [26], [27]
Annealing temperature, 340 °C

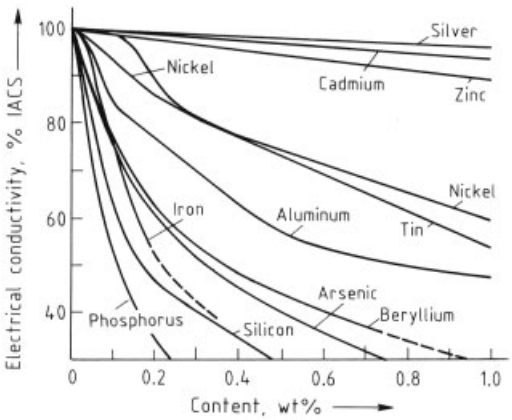

Figure 4. The effect of alloying elements on the conductivity of copper [28, p. 91]
Of the elements shown, silver has relatively little effect, whereas phosphorus greatly affects the conductivity.

Figure 4 illustrates the effect on the electrical conductivity of alloying various elements with copper. Phosphorus, which strongly lowers the conductivity, is the most widely used deoxidizer, but other elements can be used, e.g., lithium, sodium, boron, carbon, magnesium, or silicon. Beryllium, aluminum, calcium, manganese, and zinc are also deoxidizers, but they are usually added to copper for other reasons.

Boron [*7440-42-8*] is used as a deoxidizer. *Arsenic* [*7440-38-2*] is alloyed in copper to increase the corrosion resistance and to increase the temperature of recrystallization. Arsenic decreases the electrical conductivity (see Fig. 4). *Cadmium* [*7440-43-9*] is used as an alloying element in copper to increase the strength of the copper and, like silver, to raise the softening temperature. Cadmium lowers the electrical conductivity more than silver (see Fig. 4). Small amounts of tin [*7440-31-5*] may be added to copper as a strengthener. Tin also reduces the electrical conductivity.

Tellurium [*22541-49-7*] is alloyed with copper to increase machinability. The machinability rating of pure copper is 20 FMB (free machining brass (FMB), 100), but addition of tellurium increases the rating to 85 FMB. Tellurium has only a slight effect on electrical conductivity, because it is insoluble in copper in the solid state. Sulfur [*7704-34-9*] is used at times to increase the machinability of copper. Sulfur decreases casting quality by creating gas holes in the casting, making sulfur less desirable than tellurium.

Zirconium [*7440-67-7*] functions as a deoxidizer when added to copper. In fact, it reacts so readily with oxygen that a patented casting technique was developed to make zirconium addition feasible. The residual zirconium alloys with copper. Only a small amount of zirconium is added to the copper, 0.05 – 0.15 wt %. It provides strength through a precipitation hardening effect. The solid solubility of zirconium in copper decreases with decreasing temperature. The zirconium forms an intermetallic compound with copper that is a fine precipitate in the copper matrix. Zirconium copper has 95 % IACS conductivity, a high softening temperature, and good strength and formability. One difficulty in using zirconium is the effect on surface quality. The copper – zirconium precipitate can be large at times. In continued rolling, the precipitate produces a gray copper color at the surface of the strip, leaving a mottled *copper – gray copper* appearance that is undesirable in some applications.

Cobalt [*7440-48-4*] acts as a strengthener in copper, forming a precipitate of fine intermetallic compounds. Where phosphorus is present,

cobalt exhibits an even greater effect on alloy strength. Magnesium [*7439-95-4*] is a strong deoxidizer. It can create porosity problems in casting unless its addition is carefully controlled. As an alloying element, it forms a precipitate of $CuMg_2$ particles and strengthens the copper.

The coppers shown in Table 1 with their compositions and properties are representative of the types frequently used.

In addition to the coppers there is a series of *dispersion-strengthened coppers*, containing various quantities of aluminum oxide [*1344-28-1*], Al_2O_3, as the dispersant. The alloys are manufactured by using powder metallurgical technology, since Al_2O_3 would tend to agglomerate in the liquid phase during conventional casting.

The analytical oxygen content is high for copper standards, reaching 0.59 wt % O. However, the effective content is $\leqq$ 0.04 wt %, the rest of the oxygen being in the Al_2O_3 dispersant. The most popular alloys in this group are described in Table 2.

2.2. High-Copper Alloys, $\geqq$96.0 % Cu

The high-copper alloys ($\geqq$ 96.0 wt % Cu) contain several other alloying elements, as well as some of those in the coppers. Those alloying elements common to both groups are usually present in a greater percentage in the high-copper alloys. The high-copper alloys are grouped with respect to the major alloying element.

Cadmium Coppers. The cadmium coppers (C16200 – C16500) contain cadmium, from 0.6 to 1.2 wt %. The high-cadmium coppers have greater hardness and higher softening temperatures than the coppers with lower cadmium additions. The reduction in electrical conductivity from 96 to 90 % IACS is modest. Tin is an additional strengthener, giving a 15 % strength increase and a reduction to 60 % IACS in electrical conductivity, with no decrease in softening temperature. Table 3 describes the most popular cadmium – copper alloys.

Beryllium Coppers. The beryllium coppers (C17000 – C17700) constitute an alloy family in the high coppers. The addition of up to 2 wt % Be and a small amount (0.6 wt % max.) of cobalt yields one of the highest strength copper alloys (C17200) available, with upward of 1380 MPa ultimate tensile strength. Figure 5 is a partial copper – beryllium phase diagram.

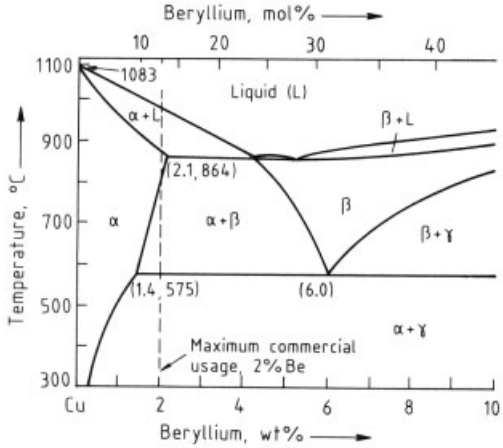

Figure 5. Copper–beryllium phase diagram [28, p. 63]

The very high strength, as in all of the copper – beryllium alloys, is obtained through heat treatment. The metal is subjected to a solution anneal and quenched. A saturated alpha Cu – Be – Co phase is produced. Work hardening the metal in the solutionized condition adds to the total strength obtainable by the subsequent precipitation of intermetallic particles. When the metal has been work-hardened to the extent desired, the metal is given a precipitation heat treatment (aged) and copper cobalt beryllides, (Cu, Co) Be, are precipitated, providing an extremely strong copper alloy. The precipitation process not only strengthens the product, but also increases the electrical conductivity. The electrical conductivity of C17200 can increase from 16 to 22 % IACS during the precipitation heat treatment.

Most of the beryllium – copper alloys are supplied as either *age hardenable* or *mill hardened.* The age-hardenable temper permits the manufacturer to fabricate parts from a relatively ductile material. The precipitation heat treatment is then performed on the fabricated parts to maximize their strength and electrical conductivity. The parts may require fixturing prior to and during the heat treatment to minimize warpage during the treatment cycle. Fabricating parts from mill-hardened metal does not require any further

Table 1. Wrought coppers

UNS (ISO) alloy	Alloying elements[a], wt%	Permissible impurities (max.[b]), wt%	Density, g/cm^3	Description, properties	Uses
C10100	Cu 99.90	P 0.0003 Te 0.0010[c] Hg 0.0001 Cd 0.0001 S 0.0018 Pb 0.0010 Se 0.0010 Bi 0.0010 O_2 0.0010	8.94	oxygen-free electronic; extremely pure copper with very low oxygen formulated for electronic applications where purity is essential to maintain electrical conductivity > 101% IACS as annealed	bus bars, waveguides, anodes, glass-to-metal seals
C10200 (CuOF)	Cu 99.95		8.94	oxygen-free, very pure copper with no residual deoxidants, electrical conductivity of 101% IACS	where the presence of oxygen would decrease deformability in fabricating the required products in transistor components, coaxial cables, klystrons
C10400	Cu + Ag 99.95 Ag 0.027		8.94	oxygen-free with silver added to raise the softening point	similar to C10200 where good creep strength at elevated temperature is essential
C10700	Cu + Ag 99.95 Ag 0.085		8.94	oxygen-free with silver added to raise the softening point	similar to C10400
C10920	Cu + Ag 99.90	O_2 0.02	8.94	low residual oxygen maintained by processing	replaces the more expensive oxygen-free copper and still retains maximum formability and electrical conductivity of 100% IACS
C10940	Cu + Ag 99.90 Ag 0.085	O_2 0.02	8.94	low residual oxygen, alloy C10920 with silver	silver-bearing copper, retains formability and electrical conductivity of 100% IACS with silver to raise the softening point, replaces the more expensive oxygen-free alloys
C11000 (CuETP)	Cu + Ag 99.90	O_2 0.04 nominal	8.94	electrolytic tough pitch, commercially pure copper: quantity of retained oxygen depends on processing, embrittles if heated above 370 °C in hydrogen	electrical wires, nails, switches, terminals, rivets
C11600 (CuAg0.1)	Cu + Ag 99.90 Ag 0.085	O_2 0.04 nominal	8.94	tough-pitch copper with silver, commercially pure copper with silver to raise the softening point	automotive radiator, fin stock, printed circuit foil
C12000 (CuDLP)	Cu + Ag 99.90 P 0.004–0.012		8.94	phosphorus deoxidized, low residual phosphorus, fully deoxidized to meet hydrogen embrittlement tests and maximize electrical conductivity of 95% IACS	electronic connectors, waveguides
C12200 (CuDHP)	Cu + Ag 99.9 P 0.015–0.040		8.94	phosphorus deoxidized, high residual phosphorus, phosphorus retained to protect against hydrogen embrittlement during welding or brazing, electrical conductivity reduced by phosphorus to 85% IACS	water and heat-exchanger tubing and plates

Table 1 Continued

C12800	Cu + Ag 99.88 Ag 0.034	As 0.12 Sb 0.003 Te + Se 0.025 Ni 0.050 Bi 0.003 Pb 0.004	8.9	fire-refined tough pitch with silver, lower grade of copper that has not been electrolytically refined, silver added to raise the softening point	roofing, screening, kettles, and radiator fin stock
C14200 (Cu,As(P))	Cu + Ag 99.4 As 0.15–0.50 P 0.015–0.040		8.94	phosphorus-deoxidized arsenical, electrical conductivity of 45% IACS, an aid in resistance welding, arsenic reduces corrosion	locomotive fireboxes, heat-exchanger and condenser tubes
C14310	Cu + Cd 99.90 Cd 0.10–0.30		8.94	cadmium-deoxidized copper, electrical conductivity of 96% IACS and ability to resist softening and hydrogen embrittlement	anneal-resistant electrical components such as lead frames, contacts, and terminals, furnace-brazed assemblies and components
C14500 (Cu,Te)	Cu + Te 99.90 P 0.004–0.012 Te 0.40–0.7		8.94	phosphorus-deoxidized, tellurium-bearing copper, free-machining copper, electrical conductivity of 93% IACS combined with good machinability and good hydrogen embrittlement, excellent choice when extensive machining is required	electrical connectors, soldering tips, welding tips, parts assembled by furnace brazing
C15000	Cu + Ag 99.80 Zr 0.10–0.20		8.89	zirconium copper: heat-treatable deoxidized copper, electrical conductivity of 93% IACS, good corrosion resistance, good creep strength	electrical switches, high-temperature circuit breakers, solderless wrapped terminals

[a] Minimum, except for ranges. [b] Maximum, except where stated otherwise. [c] Total Se, Te, Bi, As, Sb, Sn, and Mn not to exceed 0.0040 wt%

Table 2. Wrought dispersion-strengthened coppers

UNS (ISO) alloy	Alloying elements*, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C15710	Cu + Ag 99.71 Al 0.08–0.12 O_2 0.07–0.15	Fe 0.01 P 0.01	8.82	"glid-cop" alloys, produced through powder metallurgical techniques to obtain rolled strip, wire, rod, and shaped wire with electrical conductivity of 90% IACS	electrical connectors, electronic spring contacts, heat sinks
C15720	Cu + Ag 99.52 Al 0.18–0.22 O_2 0.16–0.24	Fe 0.01 P 0.01	8.81	similar to C15710, where high strength is required and electrical conductivity of 89% IACS is tolerable	lead frames, contact supports, circuit-breaker components
C15735	Cu + Ag 99.24 Al 0.33–0.37 O_2 0.29–0.37	Fe 0.01 P 0.01	8.80	similar to C15710 and C15720, where even higher strength and an electrical conductivity of 85% IACS are retained after exposure to high temperature	resistance-welding electrodes, circuit-breaker components, heat sinks

* Minimum, except for ranges. Aluminum is reported as Al_2O_3 and oxygen is reported as $Al_2O_3 + Cu_2O$.

Table 3. Wrought cadmium coppers

UNS (ISO) alloy	Alloying elements*, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Description, properties	Uses
C16200 (CuCd1)	Cd 0.7–1.2 Cu + Ag + Cd + Fe 99.8 Cu + Ag remainder	Fe 0.02	8.89	cadmium copper, high strength with electrical conductivity of 90% IACS	trolley wire, heating pads, electric-blanket elements, high-strength transmission lines
C16500	Cd 0.6–1.0 Sn 0.5–0.7 Cu + Ag + Cd + Sn + Fe 99.8 Cu + Ag remainder	Fe 0.02	8.89	tin-bearing cadmium copper, higher strength, electrical conductivity of 60% IACS	electrical springs, lead frames, electronic contacts, clips, flat cables

* Minimum, except for ranges.

heat treatment, but strength usually is sacrificed to have sufficient formability for part fabrication.

Beryllium is expensive. Use of the beryllium coppers is prohibited by the cost of the alloys, except where their very high strength, good relaxation, and corrosion resistance are necessary. Although some degrading of the high strength occurs, substitution of more cobalt and nickel for some of the beryllium produces less expensive grades of beryllium copper (C17400, C17410, C17500). One advantage of these lower priced products is that the electrical conductivity is 40 or 60% IACS, rather than 22% IACS. Since these alloys have better formability, they are supplied mill-hardened, eliminating the need for heat treatment after forming.

Where wire and rod are used to form the beryllium–copper components, machinability is often needed. Lead is added, up to 0.6 wt% (C17300), to aid in the machinability. Free machining brass has a *machinability* of 100 FMB, unleaded beryllium copper of 20 FMB, and leaded beryllium copper of 50 FMB. Some producers also add small quantities of tellurium for machinability, but lead gives superior results.

The high temperatures and long time cycles normally employed in the solution heat treatment of beryllium–copper alloys result in relatively large grain sizes, which decrease formability in the thinner products now in use. Some producers add cobalt or increase the amount of cobalt used as a grain refiner with good results. The cobalt also replaces part of the beryllium. New processing techniques are being incorporated to effect a grain size reduction. Table 4 describes the more popular beryllium–copper alloys.

Chromium Coppers. Chromium is added to copper up to 1.2 wt% to provide another family of alloys that are heat treatable (C18000–C18500). In heat treating, intermetallic particles are precipitated to yield a moderately strengthened copper with an ultimate tensile strength of 380 MPa vs. 220 MPa and a moderately high electrical conductivity of 80% IACS vs. 101% IACS for C11000. The chromium that is present not only is a strengthening agent, but also acts as a grain refiner. Some chromium coppers may contain up to 0.8 wt% Si as a substitute for chromium and as an aid in forming a fine precipitate. The more popular chromium–copper alloys are described in Table 5.

Leaded Copper and Specialty Copper Alloys. Alloys C18700–C19100 (UNS) are principally designed for wire and for welding rod products. Within this group, tin and nickel are added to strengthen the copper. For welding, silicon is added to increase the fluidity of the weld metal. Manganese and phosphorus are added to act as deoxidizers. Alloy C18700 has up to 1.5 wt% Pb added for machinability, giving a machinability of 85 FMB. The presence of the lead makes this alloy useful for machinable wire, but not for welding. Alloy C18900 contains tin, silicon, and manganese for strength and fluidity in welding. Alloy C19100 contains nickel, tellurium, and phosphorus, is heat treatable, and

Table 4. Wrought beryllium coppers

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C17000 (CuBe1.7-CoNi)	Be 1.60–1.79 Cu + Ag remainder	Ni + Co 0.20 Ni + Co + Fe 0.6 Si 0.20 Al 0.20	8.41	beryllium copper, 165 alloy, heat-treatable, supplied either age hardenable or mill hardened, exhibits hard-temper tensile strength of 1310 MPa with electrical conductivity of 22% IACS	bellows, springs, nonsparking safety tools, lock washers
C17200 (CuBe2-CoNi)	Be 1.80–2.00 Cu + Ag remainder	Ni + Co 0.20 Ni + Co + Fe 0.6 Si 0.20 Al 0.20	8.30	beryllium copper, 25 alloy, heat treatable, normally supplied mill hardened or age hardenable, exhibits hard-temper tensile strength of 1380 MPa with electrical conductivity of 22% IACS	switchblades, diaphragms, contacts, connectors, terminals
C17300 (CuBe2-Pb)	Be 1.80–2.00 Pb 0.20–0.6 Cu + Ag remainder	Ni + Co 0.20 Ni + Co + Fe 0.6 Si 0.20 Al 0.20	8.30	beryllium copper, M 25 alloy (machinable), same properties as C17200 with lead added for machinability (50% vs. 20% for other beryllium coppers), must be heat treated	parts manufactured by screw machine techniques
C17400	Co 0.15–0.35 Be 0.15–0.50 Cu + Ag remainder	Fe 0.20 Si 0.20 Al 0.20	8.80	relatively new beryllium copper, cobalt added as a grain refiner and strengthener, available mill hardened, tensile strength of 800 MPa, electrical conductivity of 40–55% IACS, very good resistance to stress relaxation	electrical and electronic components
C17410	Co 0.35–0.6 Be 0.15–0.50 Cu + Ag remainder	Fe 0.20 Si 0.20 Al 0.20	8.80	relatively new beryllium copper, similar to C17400, tensile strength of 900 MPa, electrical conductivity of 45–60% IACS	same as C17400
C17500 (CuCo2Be)	Co 2.4–2.7 Be 0.40–0.7 Cu + Ag remainder	Fe 0.10 Si 0.20 Al 0.20	8.75	low-beryllium copper, alloy 10, supplied either age hardened or mill hardened, similar to C17400, tensile strength of 900 MPa, electrical conductivity of 45–60% IACS	high-strength electrical contacts

Table 5. Wrought chromium coppers

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C18000	Ni(Co) 2.0–3.0 Cr 0.10–0.6 Si 0.40–0.8 Cu + Ag remainder	Fe 0.15	8.89	chromium copper, heat treatable, usually supplied aged	resistance welding electrodes, cable connectors, molds, spot-welding tips
C18200 (CuCr1)	Cr 0.6–1.2 Cu + Ag remainder	Fe 0.10 Si 0.10 Pb 0.05	8.89	chromium copper, heat treatable, usually supplied aged, excellent cold workability, electrical conductivity of 80% IACS	circuit-breaker components, spot-welding tips, electrical and thermal conductors

Table 6. Wrought leaded copper and specialty coppers

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C18700 (CuPb1)	Pb 0.8–1.5 Cu + Ag remainder		8.94	leaded copper, lead added to oxygen-free or deoxidized pure copper, electrical conductivity of 96% IACS, machinability of 85%	electrical connectors and screw machine parts requiring high electrical conductivity
C18900	Sn 0.6–0.9 Si 0.15–0.40 Mn 0.10–0.30 Cu + Ag remainder	Pb 0.02 P 0.05 Al 0.01 Zn 0.10	8.91	copper welding wire and rod, alloy strongly deoxidized, good strength but electrical conductivity of only 30% IACS	primarily welding
C19100	Ni 0.9–1.3 Te 0.35–0.6 P 0.15–0.35 Cu + Ag remainder	Fe 0.20 Pb 0.10 Zn 0.50	8.94	heat-treatable machinable copper, fully heat treated, electrical conductivity of 55% IACS, machinability of 75%	bolts, bushings, electrical components, gears, nuts, tie rods

has a machinability of 75 FMB. The leaded copper and specialty copper alloys are in Table 6.

High-Strength Coppers. The last alloy family in the high-copper alloys has iron as the major alloying element. Iron adds strength to the alloys, but dissolving iron in molten copper is a problem. A master alloy of copper and iron aids greatly in putting the iron into solution. These alloys contain phosphorus to deoxidize the melt, and there is a synergistic relationship among the copper, iron, and phosphorus that adds additional strength to the alloy.

The UNS numbers in this alloy series start with the moderate strength C19200, which contains 1 % Fe and 0.025 wt % P. A major alloy in this series is C19400, which contains 2.1–2.6 wt % Fe, with an ultimate tensile strength of 525 MPa in spring temper. During processing, the iron and phosphorus are first solutionized and then precipitated as fine spheroids of copper–iron phosphides. As in other alloys, precipitation of the alloy additions from solution increases the electrical conductivity. When fully solutionized, the alloy has an electrical conductivity as low as 35 % IACS. However, through special processing in which the maximum precipitation has been achieved, the alloy can exhibit an electrical conductivity as high as 75 % IACS. Alloy C19400 also has a small percentage (ca. 0.12 wt %) of zinc for improved control over process annealing.

Alloy C19500 is the next commercially important alloy in the high-copper series. It contains iron, phosphorus, tin, and cobalt as strengtheners and deoxidizers. The precipitation of intermetallic particles of iron, phosphorus, and cobalt is used to gain strength, and the tin is also a solid-solution strengthener. In the annealed, fully precipitated state, the alloy develops an ultimate tensile strength of 410 MPa with an electrical conductivity of 55 % IACS. Through cold work after the precipitation anneal, an ultimate tensile strength in excess of 670 MPa can be attained.

Another alloy in the system of high coppers is C19600, consisting of copper–iron–phosphorus. This alloy is very similar to C19200, except it has slightly higher phosphorus content. Its mechanical properties are similar to those of C19200. The higher phosphorus seems to increase precipitation, so the electrical conductivity is higher.

The high-strength coppers are further described in Table 7.

3. Wrought Brasses

The series of alloys from which a majority of the industry's products are made are the copper–zinc alloys – the brasses (C20000–C29999). The most popular alloy of the group is C26000 (cartridge brass, 70/30 brass).

Table 7. Wrought high-strength coppers

UNS (ISO) alloy	Alloying elements*, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C19200	Cu 98.7 Fe 0.8–1.2 P 0.01–0.04		8.87	HC copper iron, a high-strength moderate-conductivity alloy: tubing with electrical conductivity of 50% IACS and hard-drawn tensile strength of 390 MPa, hard-temper strip with electrical conductivity of 60% IACS and tensile strength of 450 MPa	tubing: heat exchangers, automotive hydraulic brake linings strip: electrical connectors, contact springs, lead frames, gaskets, hollowware
C19400	Cu 97.0 Fe 2.1–2.6 Zn 0.05–0.20 P 0.015–0.15	Pb 0.03	8.77	HSM copper, similar to C19200, strip or welded tube, hard-temper tensile strength of 450 MPa, electrical conductivity varies: 35% IACS (welding), 65% IACS (normal), 70% IACS (high condition)	electronic connectors and contacts, lead frames, cable shielding, gasketing
C19500	Cu 96.0 Fe 1.0–2.0 Sn 0.10–1.0 P 0.01–0.35 Co 0.30–1.3	Zn 0.20 Al 0.02 Pb 0.02	8.91	strescon, similar to C19200, electrical conductivity of 50% IACS with a hard-temper tensile strength of 590 MPa, excellent hot and cold workability	electrical springs, terminals, connectors

* Minimum, except for ranges.

Zinc is by far the most important alloying element in the copper alloys. Zinc is a deoxidizer, but when it is added in quantities of 5 % or less, *phosphorus* may also be added to ensure full deoxidation. As zinc content in copper is increased, the density, melting point, electrical and thermal conductivity, and the modulus of elasticity (Young's modulus) decrease, but the strength, hardness, and coefficient of expansion increase.

Figure 6 is a copper – zinc phase diagram that illustrates some of these effects. The effect of zinc content on the color of the alloy is also of importance commercially, because many brasses are used for decorative purposes. The red copper color predominates to alloy C22600 (12.5 wt % Zn), where a red-gold color, very close to that of 18-carat gold, is evident. The color then progresses toward a yellow gold on to brass yellow. In the range of 38 wt % Zn, a buff red color appears, especially on the fracture surface. The buff red is a result of the presence of the beta phase along with the alpha phase and is an indication that the beta phase is present.

The *alpha phase* has good hot-working characteristics, but at the higher zinc content, the hot-working temperature range narrows perceptibly. With the *beta phase* present in the very high-zinc brasses, hot working improves. The alpha phase brasses have good cold-working properties, and C26000 exhibits the best cold-working properties, especially in deep drawing, of any of the copper alloys. When the beta phase is present in the very high-zinc brasses, cold working is hindered because the beta phase has poor cold-working characteristics. Richer zinc alloys are hot worked to a thickness at which additional hot working is no longer practical. The hot-worked surface is removed by milling, and the metal is subjected to a 2-h homogenizing anneal. This causes diffusion of the zinc, transforming the beta phase into alpha phase: good cold-working properties are then available.

There are few other elements alloyed with the brasses in the C20000 system. Phosphorus is one. When the metal is subjected to a hot water environment, 0.02 – 0.08 wt % As may be added to inhibit dezincification (C26310). One alloy (C23030) has 0.20 – 0.40 wt % silicon added for additional strengthening and resistance welding.

Since the production of the brass alloys relies predominantly on recycling scrap, impurities must be controlled carefully. Impurities encountered in the brasses affect processing of the metal more than they affect uses.

When lead approaches 0.025 wt %, the hot-working characteristics become impaired. Even

though industry standards permit 2 – 3 times that quantity, the lead must be controlled by the brass mills to the lower level.

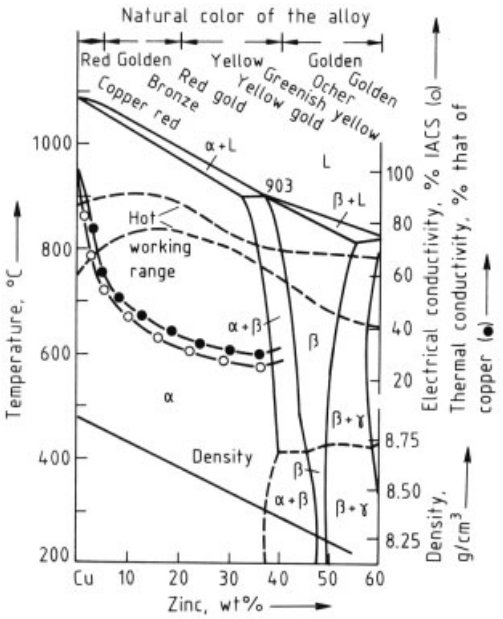

Figure 6. The copper–zinc phase diagram [26]
The thermal conductivity of the alloys is given as a percentage of that of copper. Density, electrical conductivity, and thermal conductivity are for 20 °C.

Iron is a grain refiner. An annealing cycle may be set with the expectation of a specific quantity of iron being present. When the alloy being processed is low in iron (0.01 wt % Fe), the resulting annealed grain sizes may be too large, exceeding the specification. When the alloy being processed is high in iron (0.025 wt % Fe), the resulting grain sizes may be too small. Reannealing the high-iron metal at a higher temperature frequently results in *critical grain growth,* and the resulting grain size greatly exceeds the specification. Commercial standards permit 0.05 % Fe generally, but internal controls must be much tighter.

Phosphorus impurity levels must be restricted. If no iron were present, phosphorus would be a small problem, but iron is almost always present. Phosphorus contents > 0.001 % can lead to hot rolling problems and, even more definitely, loss of grain growth in annealing. With phosphorus present in the alloy along with iron, the grain refinement is extreme, making anneals unpredictable. Normally silver is counted as copper, with little effect on the properties of brass. When the brass is used for decorative work or coinage, however, silver can be a problem, and silver-bearing copper scrap cannot be used in brass where a true brass color is required.

Normally nickel is not a problem in brass, except in conjunction with iron or iron and phosphorus. Then nickel in quantities < 0.05 wt % can act as an additional grain refiner, giving annealing problems in further processing. Nickel can also act like silver in changing the color of the brass.

Table 8 describes the more popular brass alloys.

4. Wrought Leaded Brasses

The wrought leaded brasses (C30000 series) are brasses to which lead is added. The main purpose of adding lead is for machinability. As the quantity of lead increases up to ca. 3 wt %, the machinability index increases to 100 FMB for free-machining brass (C36000), as shown in Figure 7.

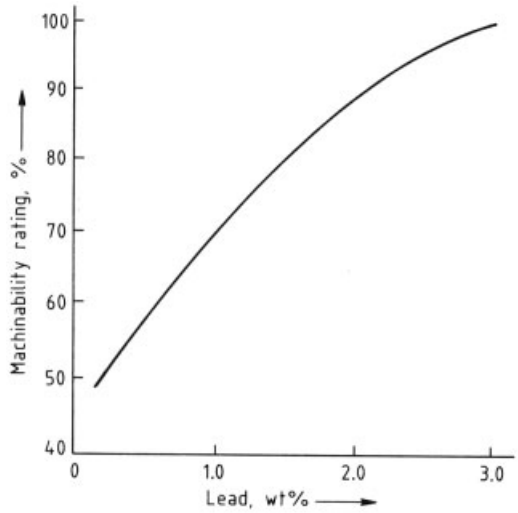

Figure 7. The effect of lead content on the machinability rating for leaded brass alloys

Because of the lead content, most leaded brasses are not easily worked either hot or cold. The copper-rich leaded brasses are best cold-rolled. The high-zinc-content leaded brasses have a good-to-excellent hot workability. At hot-working temperatures, the alpha-phase crys-

Table 8. Wrought copper–zinc alloys (wrought brasses)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C21000 (CuZn5)	Cu 94.0–96.0 Zn remainder	Pb 0.03 Fe 0.05	8.86	gilding metal (95% Cu), copper-red color, electrical conductivity of 56% IACS, hard-temper tensile strength of 390 MPa	coins, medals, tokens, bullet jackets, emblems, jewelry, craft work
C22000 (CuZn10)	Cu 89.0–91.0 Zn remainder	Pb 0.05 Fe 0.05	8.80	commercial bronze (90% Cu), bronze color, electrical conductivity of 44% IACS, hard-temper tensile strength of 420 MPa	weather stripping, hardware, primer caps, cosmetic cases, craft work, waveguides
C22600	Cu 86.0–89.0 Zn remainder	Pb 0.05 Fe 0.05	8.78	jewelry bronze (87% Cu), red-gold color, electrical conductivity of 40% IACS, hard-temper tensile strength of 460 MPa	costume jewelry, architectural, chain, slide fasteners, cosmetic containers, coinage
C23000 (CuZn15)	Cu 84.0–86.0 Zn remainder	Pb 0.05 Fe 0.05	8.75	red brass (85% Cu), gold color, electrical conductivity of 37% IACS, hard-temper tensile strength of 480 MPa	weather stripping, electrical and electronic connectors and contacts, condenser and heat-exchanger tubes, flexible hose, coins, badges, cosmetic containers, jewelry, dials
C24000 (CuZn20)	Cu 78.5–81.5 Zn remainder	Pb 0.05 Fe 0.05	8.67	low brass (80% Cu), yellow-gold color, electrical conductivity of 32% IACS, hard-temper tensile strength of 510 MPa	coins, tokens, medallions, clock dials, bellows, flexible hose, battery cups
C26000 (CuZn30)	Cu 68.5–71.5 Zn remainder	Pb 0.07 Fe 0.05	8.53	cartridge brass (70% Cu), yellow color, electrical conductivity of 28% IACS, hard-temper tensile strength of 525 MPa, highly susceptible to stress corrosion cracking	automotive radiator cores, bead chain, lamp fixtures, socket shells, screw shells, eyelets, hinges, locksets, cartridge cases, pumps
C26130	Cu 68.5–71.5 As 0.02–0.08 Zn remainder	Pb 0.05 Fe 0.05	8.53	arsenical cartridge brass (70% Cu), properties essentially same as C26000; arsenic helps prevent dezincification in automotive radiator tubes	automotive radiator tubes
C26800 (CuZn33)	Cu 64.0–68.5 Zn remainder	Pb 0.15 Fe 0.05	8.48	yellow brass (66% Cu), greenish-yellow color, electrical conductivity of 27% IACS, hard-temper tensile strength of 510 MPa	automotive radiator cores, tanks, reflectors, lamp fixtures, bead chain springs, screws
C28000 (CuZn40)	Cu 59.0–63.0 Zn remainder	Pb 0.30 Fe 0.07	8.39	Muntz metal (60% Cu), golden color, electrical conductivity of 28% IACS, hard-drawn tensile strength of 590 MPa, poor cold-working properties but excellent hot-working properties, subject to stress corrosion cracking and dezincification	decorative products, heat exchangers, brazing rods, sometimes used to match C23000 in color

talline structure is transformed to the beta-phase crystalline structure. The beta phase absorbs the lead and permits hot work to be done. Those alloys whose compositions skirt the alpha – beta phase boundary show good hot workability when in the beta phase. The beta phase can then be transformed to the alpha phase by diffusion annealing for good cold workability. These are the methods most brass mills use to cast, hot work, and cold work leaded brass alloys economically.

The lead in leaded brasses improves machinability because it is present as small inclusions, causing the machined chip to break away freely. This reduces the friction against the cutting tool, so heat does not build up or cause rapid dulling of the tool. Heat is more readily dissipated, and lubrication can be more beneficial. There may also be some slight lubrication from the lead.

There is another action operating that enhances the machinability of the leaded brasses. When some beta phase is permitted to remain, machinability increases. These beta particles retard cold workability, increasing fracturing tendencies. The beta fractures at the tool – metal interface, producing essentially the same effect as lead.

Lead and zinc are not the only two elements alloyed in the C30000 series. As in the C20000 series, several elements are added for special purposes:

Phosphorus is alloyed in the copper-rich leaded brasses to obtain additional deoxidizing for added strength and to inhibit dezincification. Nickel is added at times as a strengthener. Arsenic (up to 0.25 wt %) is added to help inhibit dezincification of the alloy, especially in water service. Antimony may be added to the leaded brasses for dezincification resistance. Quantities up to 0.10 wt % are used.

Aluminum is added to architectural bronze (C38000) to increase the resistance to atmospheric corrosion of the metal. Small quantities (< 0.6 wt %) are added to help develop aluminas (aluminum oxides) at the metal surface to aid in protection from the corrosive effect of the environment.

The effect of impurities in the leaded brasses is about the same as in the wrought brasses (C20000 series). The one major element that can give trouble is iron. As in the unleaded brasses, iron must be controlled carefully because of its effect on the recrystallization temperature of the alloys. Iron is also an effective grain refiner in the C30000 series. Table 9 contains information on the leaded brass alloys.

5. Wrought Tin Brasses

The copper – zinc – tin alloys (C40000 series), more commonly called the tin brasses, are probably the most underutilized alloys in the copper alloy families. In these alloys, both tin and zinc have been added to copper to produce alloys that in some compositions are stronger, have higher fatigue strength, better corrosion resistance, and better stress corrosion resistance than the high-zinc brasses. They are more economical than the phosphor bronzes (C50000 series) that are frequently used when these properties are needed.

Figure 8 illustrates this relationship on the basis of tensile strength. The strength of some C40000 series alloys is plotted against the sum of tin plus zinc alloyed. The strength of the C50000 series is plotted against the quantity of tin contained, and the strength of the C20000 series is plotted against the zinc contained.

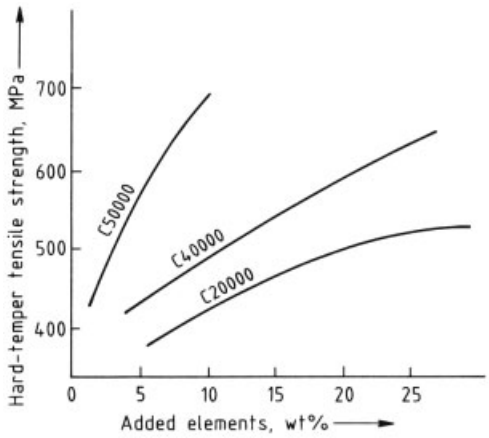

Figure 8. Effect of alloying elements on the hard-temper tensile strength of the alloy series C50000, C40000, and C20000

The C40000 series found ready use during and after World War II as a substitute for the phosphor bronze alloys (C50000 series). Tin was difficult to obtain and costly, so the tin – zinc combination was acceptable as a substitute for the high-tin alloys, although some strength was sacrificed to obtain equal formability.

Table 9. Wrought copper – zinc – lead alloys (wrought leaded brasses)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C31400	Cu 87.5–90.5 Pb 1.3–2.5 Zn remainder	Fe 0.10 Ni 0.7	8.83	leaded commercial bronze, rich bronze color, electrical conductivity of 42% IACS, normally supplied as rod and wire, machinability of 80%	screw machine parts, electrical plug-type connectors, builders hardware, screws, nuts, bolts
C33000	Cu 65.0–68.0 Pb 0.25–0.7 Zn remainder	Fe 0.07	8.50	low-leaded brass tubing, yellow color, electrical conductivity of 26% IACS, machinability of 60%	where degree of machinability and cold-working properties required: pump lines, T-bends, trap lines
C33200	Cu 65.0–68.0 Pb 1.5–2.5 Zn remainder	Fe 0.07		high-leaded brass tubing, yellow color, electrical conductivity of 26% IACS, machinability of 80%	same as C33000 except greater machinability
C33500 (CuZn35-Pb)	Cu 62.0–65.0 Pb 0.25–0.7 Zn remainder	Fe 0.15	8.47	low-leaded brass, yellow color, electrical conductivity of 26% IACS, machinability of 60%	clock parts, hinges
C34000 (CuZn35-Pb2)	Cu 62.0–65.0 Pb 0.8–1.5 Zn remainder	Fe 0.15	8.47	medium-leaded brass (64.5% Cu), yellow color, electrical conductivity of 26% IACS, supplied as bar, strip, rod, and wire, machinability of 60%	engravings, gears, nuts, bolts, knurled and threaded parts
C34200 (CuZn34-Pb2)	Cu 62.0–65.0 Pb 1.5–2.5 Zn remainder	Fe 0.15	8.5	high-leaded brass (64.5 Cu), grayish-yellow color, electrical conductivity of 26% IACS, supplied as bar, strip, rod, and wire, machinability of 90%	keys, clock parts, valve stems, general screw machine parts that also require cold working such as knurling and staking
C35000 (CuZn37-Pb1)	Cu 60.0–63.0 Pb 0.8–2.0 Zn remainder	Fe 0.15	8.47	medium-leaded brass (62% Cu), yellow color, electrical conductivity of 26% IACS, supplied as bar, strip, rod, and wire, machinability of 70%	clock parts, engravings, hinges, hose couplings, keys, zinc strainers, locks, washers
C36000 (CuZn36-Pb3)	Cu 60.0–63.0 Pb 2.5–3.7 Zn remainder	Fe 0.35		free-cutting brass, grayish-yellow color, electrical conductivity of 26% IACS, supplied as bar, rod, and wire, poor cold workability but fair hot workability, machinability of 100%	hardware parts, gears, pinions, high-speed screw-machining components

Resistance to stress relaxation has become an important property for many electrical and electronic connectors, which are frequently required to function while stressed at elevated temperatures. Examples are those found in computers and in the engine compartments of automobiles. Under such conditions, the spring member of a connector pair tends to relax, and a poor electrical connection results. The brasses, adequate for many room-temperature uses, may lose contact force at moderately high temperatures. The tin brasses have somewhat better resistance to stress relaxation than the brasses at temperatures up to 75 °C. (The phosphor bronzes have better resistance than the tin brasses.)

In choosing among these alloys for a particular use, strength, formability, stress relaxation resistance, and cost effectiveness all are factors that need to be considered. One of the stronger alloys, such as C42500, might be just as suitable as the higher cost phosphor bronzes. Full testing

of prototype parts under working conditions is always recommended.

The wrought tin brasses (C40000 series) can be subgrouped by the products or the forms in which these alloys are supplied and, consequently, how products are applied in their ultimate use:

1) Alloys C40000 – C43600, supplied most frequently in sheet and strip form, with minor quantities in rod, wire, and tube
2) Alloys C44300 – C49080, supplied most frequently in bar, rod, wire, or tube, with minor quantities in sheet and plate

In the alloys C40000 – C43600, tin contents range from 0.25 – 3.0 wt %. Zinc contents range from 4 – 11.5 wt %. No other elements are added to these alloys except in C42100, which contains 0.35 wt % Mn. Manganese is a deoxidizer, is a very moderate strengthener, increases electrical resistance, and might be beneficial for welding. Tin is a potent strengthener when it is added to copper. Zinc is also a good strengthener, but considerably higher quantities of zinc are needed to provide the equivalent of the strengthening provided by tin.

One of the important characteristics of these alloys is that their susceptibility to stress corrosion cracking is quite low. In applications in which the strength of a high-zinc brass is needed, but in which applied stresses would make the part subject to stress corrosion failure, an alloy like C42200 or C42500 can be used, and stress corrosion failure is most unlikely.

In the alloys C44300 – C49080, several elements are added to the copper – tin – zinc system for special purposes, depending on fabricating methods and ultimate use of the product. Several alloys have up to 0.25 wt % As, 0.10 wt % Sb, or 0.10 wt % P added to minimize dezincification. Alloys C47600 – C48600 have up to 2.5 wt % lead added for machinability. Small amounts of manganese (0.1 wt %) may be added to improve the hardness, or nickel (0.3 wt %) may be added for higher strength or corrosion resistance.

Table 10 describes the most widely used tin brasses.

6. Wrought Tin Bronzes and Brazing Alloys

In general, the wrought tin bronzes (C50000 series) provide the best combination of strength, electrical conductivity, and resistance to corrosion, as well as great versatility. The wrought tin bronzes can be classified into three subgroups:

1) Alloys containing copper, tin, and phosphorus (C50100 – C52400) usually produced as strip, rod, or wire (small amounts of other elements possibly present)
2) Leaded bronzes (C53200 – C54800) available in plate, sheet, strip, rod, wire, or bar
3) Copper – phosphorus and copper – silver – phosphorus alloys (C55180 – C55284) produced mainly as rod or wire products

6.1. Phosphor Bronzes

The term phosphor bronze usually refers to the alloys C50100 – C52400. The name phosphor bronze might imply a high phosphorus content, but these are copper – tin alloys that contain 0.01 – 0.35 wt % P for deoxidation.

The tin content is of paramount importance in this group of alloys. Figure 4 shows that additions of tin to copper reduce the electrical conductivity substantially. The maximum amount of tin that can be tolerated in wrought alloys is 10 wt % (Fig. 9). Above 10 wt %, as the copper – tin phase diagram indicates, a second phase is formed. This second phase is brittle and can be neither hot-worked nor cold-worked. In practice, alloys containing 5 wt % Sn cannot be hot-worked readily.

Tin is a low-melting-temperature element, and the temperature range between liquidus and solidus is quite large. During the solidification of castings, tin-rich molten metal solidifies last. Capillary action through the dendritic structure and pressures exerted by the solidifying copper-rich metal push the low-melting tin-rich liquid to the casting surface. This is called *inverse segregation.* (Segregation created through solidification is usually near the center of the cast bar.) A gray *delta surface* results, which is quite brittle and must be removed prior to working.

When inverse segregation takes place, the tin content in the interior of the casting is reduced

Table 10. Wrought copper–zinc–tin alloys (wrought tin brasses)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C40500	Cu 94.0–96.0 Sn 0.7–1.3 Zn remainder	Pb 0.05 Fe 0.05	8.83	high-conductivity bronze, electrical conductivity of 41% IACS, hard-temper tensile strength of 440 MPa	electrical meter clips, terminals, fuse clips, contact springs, washers
C41100	Cu 89.0–92.0 Sn 0.30–0.7 Zn remainder	Pb 0.10 Fe 0.05	8.80	lubaloy, electrical conductivity of 32% IACS, hard-temper tensile strength of 460 MPa	bushings, electrical conductors, thrust washers, flexible metal hoses, bearing sleeves
C42200	Cu 86.0–89.0 Sn 0.8–1.4 Zn remainder	Pb 0.05 Fe 0.05	8.80	lubronze, electrical conductivity of 31% IACS, hard-temper tensile strength of 500 MPa	sash chains, electrical terminals, connectors, springs, washers, fuse clips
C42500	Cu 87.0–90.0 Sn 1.5–3.0 Zn remainder	Pb 0.05 Fe 0.05 P 0.35	8.78	lubaloy X, electrical conductivity of 28% IACS, hard-temper tensile strength of 520 MPa	electrical switch springs, terminals, connectors, contacts, fuse clips, pen clips, weather stripping
C44300 (CuZn28-Sn1)	Cu 70.0–73.0 Sn 0.8–1.2 As 0.02–0.06 Zn remainder	Pb 0.07 Fe 0.06	8.53	arsenical admiralty, electrical conductivity of 25% IACS, hard-temper tensile strength of 670 MPa	condenser tubes and plates, heat-exchanger tubing, ferrules, strainers
C46400 (CuZn38-Sn1)	Cu 59.0–62.0 Sn 0.50–1.0 Zn remainder	Pb 0.20 Fe 0.10	8.41	uninhibited naval brass, electrical conductivity of 26% IACS, quarter-hard-temper tensile strength of 480 MPa (hardest temper available), machinability of 30%	condenser plates, welding rods, marine hardware, balls, nuts, bolts
C46500	Cu 59.0–62.0 Sn 0.50–1.0 As 0.02–0.06 Zn remainder	Pb 0.20 Fe 0.10	8.41	arsenical naval brass, electrical and mechanical properties same as C46400	same as C46400 where protection to dezincification is warranted
C48200 (CuZn37-SnPb)	Cu 59.0–62.0 Pb 0.40–1.0 Sn 0.50–1.0 Zn remainder	Fe 0.10	8.44	medium-lead naval brass, electrical and mechanical properties same as C46400, machinability of 50%	same as C46400, especially where better machinability is required
C48500 (CuZn37-Pb25Sn)	Cu 59.0–62.0 Pb 1.3–2.2 Sn 0.50–1.0 Zn remainder	Fe 0.10	8.44	high-lead naval brass, electrical and mechanical properties same as C46400, machinability of 70%	same as C46400, especially where better machinability is required

and may not meet the minimum tin content. Inverse segregation is more prevalent in alloys containing > 5 wt % Sn, leading many manufacturers to cast such alloys continuously in thin slabs (12 mm or 0.5 in. thick). Continuous casting minimizes inverse segregation through rapid and uniform heat removal.

Figure 10 illustrates the effect of increasing the tin content in the copper – tin alloys. Figure 8 shows that increasing the tin content increases the ultimate tensile strength. In addition, the spring-temper fatigue strength increases. All of the mechanical properties, especially the fatigue strength, of alloy C51000 are increased as the tin and phosphorus are raised toward the maximum of the limits specified. However, hot rolling becomes more difficult. Therefore, when the alloy is to meet the most demanding standards, it is usually cast continuously in thin bars, rather than hot rolled, so the tin and phosphorus can be maintained toward the maxima of the specified ranges and the metal toward the mechanical property maxima needed for some uses. When the additional requirement of 20 % IACS minimum electrical conductivity is specified for

C51100, however, the tin and phosphorus contents must be held low.

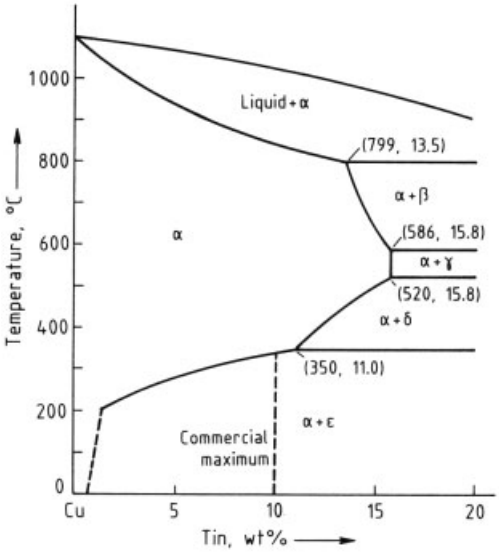

Figure 9. Copper – tin phase diagram
Maximum commercially usable tin content is 10 %, as indicated.

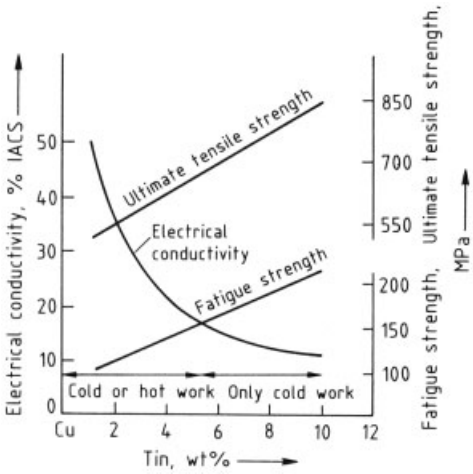

Figure 10. Effect of tin in the C50000 series of alloys

The only other elements alloyed in the phosphor bronzes are either nickel (0.10 – 0.40 wt %) or iron (0.05 – 0.15 wt %) to give extra strength or to act as grain refiners. As in many other alloys, the primary impurity that can be a problem is iron. Iron, being the strong grain refiner that it is, must be controlled to avoid annealing problems. Table 11 describes the most popular phosphor bronzes.

6.2. Leaded Phosphor Bronzes

The alloys C53200 – C54800 are the leaded phosphor bronzes. These alloys are used primarily for bearings. Alloyed with copper, the content is ca. 4 wt % tin (3.5 – 6.0 wt % Sn), 4 wt % lead (0.8 – 6.0 wt % Pb), and 0.2 wt % phosphorus (0.1 – 0.5 wt % P). With this high lead content the metal cannot be hot-worked. The most popular alloy of this group is C54400, which contains, besides the tin and lead, about 4 wt % zinc (1.5 – 4.5 wt %). This "444" alloy has excellent wearing properties, is easily cold-worked, and is readily machined (80 FMB). The zinc increases the hardness. The lead is actually high enough to be a lubricant when the alloy is used as a bearing surface. Table 12 further describes alloy C54400.

6.3. Brazing Alloys

The alloys C55180 – C55284 are copper – phosphorus or copper–silver–phosphorus brazing alloys normally supplied in the form of wire, rolled preforms, or powder. These are specialty alloys, which fit a special niche in the copper alloy system.

The copper – phosphorus alloys (C55180 and C55181) contain 4.8 – 7.5 wt % P. The phosphorus is alloyed not only for deoxidation, but also to provide a low melting temperature and increase the fluidity. This helps the molten alloy flow between the mating surfaces during brazing.

The copper – silver – phosphorus alloys have a similar purpose. The phosphorus content ranges from 4.8 to 7.2 wt % and provides for deoxidation, low melting temperature, and fluidity. The silver content ranges from 1.8 to 15.5 wt %. As the silver content increases, the melting point of the alloy decreases (Fig. 11). This alloy system provides four alloys of different melting temperatures. The silver in the alloy establishes the melting temperature to be used for brazing. If two or more joints must be made in close proximity, but not at the same time, the first braze should be made with an alloy of higher melting

Table 11. Wrought copper–tin alloys (wrought phosphor bronzes)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C50500 (CuSn2)	Sn 1.0–1.7 P 0.03–0.35 Cu remainder	Pb 0.05 Fe 0.10 Zn 0.30	8.89	phosphor bronze, 1.25% E: electrical conductivity of 48% IACS, hard-temper tensile strength of 440 MPa	electrical contacts, connectors, flexible hose, telephone contacts, lead frames
C51000 (CuSn4)	Sn 4.2–5.8 P 0.03–0.35 Cu remainder	Pb 0.05 Fe 0.10 Zn 0.30	8.86	phosphor bronze, 5% A: electrical conductivity of 15% IACS, hard-temper tensile strength of 560 MPa	bridge bearing plates, beater bars, bellows, bourdon tubing, electric and electronic contacts and connectors, springs, brushes, washers
C51100 (CuSn4)	Sn 3.5–4.9 P 0.03–0.35 Cu remainder	Pb 0.05 Fe 0.10 Zn 0.30	8.86	phosphor bronze, 4%: electrical conductivity of 20% IACS, hard-temper tensile strength of 550 MPa	electrical and electronic connectors and contacts, switches, springs
C52100 (CuSn8)	Sn 7.0–9.0 P 0.03–0.35 Cu remainder	Pb 0.05 Fe 0.10 Zn 0.20	8.80	phopshor bronze, 8% C: electrical conductivity of 3% IACS, hard-temper tensile strength of 640 MPa, greater fatigue life than C51000	similar to those for C51000
C52400 (CuSn10)	Sn 9.0–11.0 P 0.03–0.35 Cu remainder	Pb 0.05 Fe 0.10 Zn 0.20	8.78	phosphor bronze, 10% D: electrical conductivity of 11% IACS, hard-temper tensile strength of 690 MPa, greater fatigue life than C51000	similar to those for C51000, bridge compression and expansion plates, springs

Table 12. Wrought copper–tin–lead alloys (wrought leaded phosphor bronzes)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C54400 (CuPb5-Sn5Zn5)	Pb 3.5–4.5 Sn 3.5–4.5 Zn 1.5–4.5 P 0.01–0.50	Fe 0.10	8.89	phosphor bronze, B-2: electrical conductivity of 19% IACS, hard-temper tensile strength of 550 MPa, machinability of 80%	sleeve and thrust bearings, bushings, screw machine products, thrust washers, wear plates

Table 13. Wrought copper – phosphorus and copper – silver – phosphorus alloys (wrought brazing alloys) ∗

UNS (ISO) alloy	Alloying elements, wt%	AWS ∗∗ filler metal designation	Solidus, °C	Liquidus, °C
C55180	P 4.8–5.2 Cu remainder	BCuP-1	710	900
C55181	P 7.0–7.5 Cu remainder	BCuP-2	710	790
C55283	Ag 5.8–6.2 P 7.0–7.5 Cu remainder	BCuP-4	640	725
C55284	Ag 14.5–15.5 P 4.8–5.2 Cu remainder	BCuP-5	640	800

∗ Supplied as rod, wire, and some brazing preforms; not suitable for brazing copper–nickel alloys containing 20% Ni.
∗∗ AWS = American Welding Society.

temperature than the second. Table 13 describes these brazing alloys.

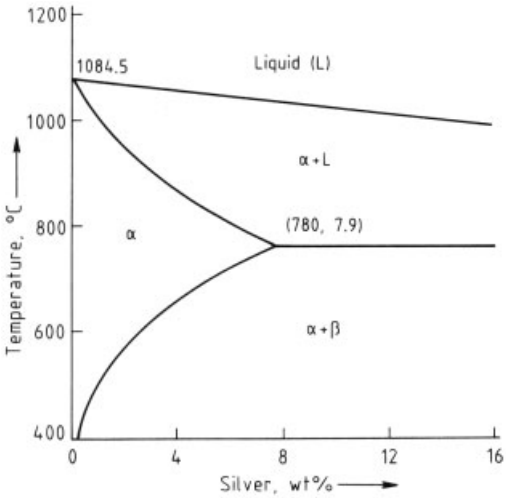

Figure 11. Copper – silver phase diagram [29]

7. Wrought Specialty Alloys

The special alloys (C60000 series) have seen a major expansion by manufacturers seeking stronger alloys that have good forming properties and good conductivity. The series covers basically the copper – aluminum alloys (aluminum bronzes), the copper – silicon alloys (silicon bronzes), and miscellaneous copper – zinc alloys (special brasses).

The search for a "stainless brass" that had the architectural ability of stainless steel to form a protective oxide led to the specialty bronzes. The development of these alloys in the 1960s and 1970s did not produce a stainless brass, but did present some new high-strength alloys with good formability and good tarnish resistance. Many of these new alloys are used extensively in electrical and electronic spring connectors.

When aluminum- or silicon-bearing copper alloys are annealed in atmospheres containing oxygen, the resultant oxide film is quite tenacious and insoluble in the normal cleaning solutions used in most brass mills. Special cleaning solutions and abrasive techniques must be employed to remove these oxide films. When such alloys are soldered or welded, special fluxes are required to dissolve the oxides, thus permitting the solder to wet the metal.

The alloys are used extensively for their desirable attributes. The use of either aluminum or silicon as an alloying element decreases the density substantially below that for most of the copper alloys. Figure 12 illustrates the marked effect of alloying aluminum with copper on the density of the alloy.

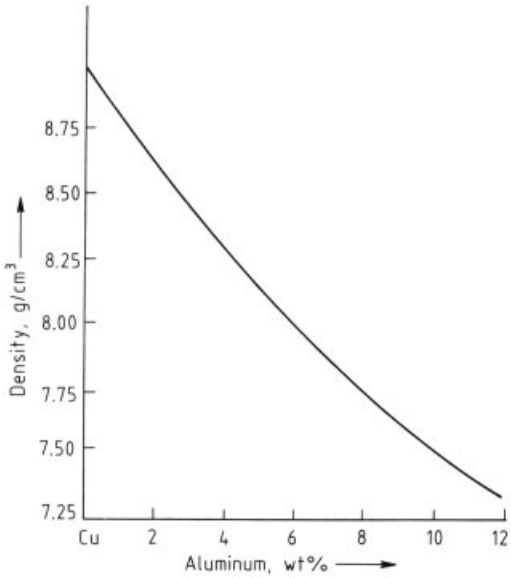

Figure 12. Density of copper – aluminum alloys

Both aluminum and silicon are strengtheners. Alloys with high strength and excellent forming properties permit a fabricator to employ thinner metal without reducing the strength of the fabricated part. The lower density and the reduced thickness permit more pieces to be manufactured from a kilogram of these alloys than could be manufactured from a kilogram of conventional copper alloy.

The specialty copper alloys are subdivided into three categories: (1) copper – aluminum alloys (aluminum bronzes), (2) copper – silicon alloys (silicon bronzes), (3) miscellaneous copper – zinc alloys.

7.1. Copper–Aluminum Alloys (Aluminum Bronzes)

The copper-aluminum alloys (C60000 – C64500) contain between 2 and 15 wt % Al, possibly with other alloying elements. When

cold finished products are to be manufactured, the aluminum content is limited to 9 wt %. The alloy is then an alpha solid solution, which has good cold ductility and may be either coldor hot-worked. These alloys may be supplied as strip or sheet, as well as hot-worked forms.

Above 9 wt % Al, the beta phase is present, which inhibits cold-working but facilitates hot-working. As a result, most alloys containing $\geqq$ 9 wt % Al are supplied as rod, wire, or hot-forgings.

Some alloys that contain 8 – 9 wt % Al may require a special annealing processing to transform the beta phase to alpha to facilitate cold-working. This processing is similar to the diffusion anneal used for the high-zinc alloy.

In addition to copper and aluminum, these alloys may contain one or more other elements.

Aluminum bronze alloys may contain up to 4.5 wt % Fe. The iron enhances corrosion resistance and is a strong grain refiner both for cast grains and for the following anneals. Iron is also a strengthener, increasing both the cold-worked and annealed strengths. As the iron content increases, the magnetic permeability (the ability of the alloy to attract a magnet) increases. This can hinder the use of high-iron alloys in electronics, communications, computers, etc.

Aluminum bronze alloys may contain up to 2 wt % Sn. The addition of tin improves the protection from dealuminification due to oxidizing corrosion of salt water. Tin is also an additional strengthener.

Alloyed in aluminum bronzes up to a maximum of 6 wt %, nickel is used basically to prevent dealuminification, with the additional function of a stabilizer to help control the annealed properties. Manganese is usually alloyed up to 3.5 wt % in the aluminum bronzes, although one alloy may contain up to 14 wt % Mn (C63380). The manganese is a deoxidizer and a strengthener. The high-manganese aluminum bronze alloys (Heusler alloys) have special magnetic properties.

Additions of up to 3 wt % Si may be made to the aluminum bronzes and still maintain good fabricating qualities. Silicon, another element that increases the strength of the alloy, also improves the fluidity for casting and improves weldability. Additions of up to 0.35 wt % As may be made to the aluminum bronzes to improve the oxidation corrosion resistance to reduce dealuminification. Additions of up to 0.55 wt % Co are made to act as strengtheners and additional preventors of dealuminification due to saltwater corrosion.

The impurity levels generally are maintained closely, as are the alloying elements within their designated limits. In many cases, the ranges for the alloying elements may be more closely confined internally than the ranges officially stated in the literature, to aid in processing. Some consequences of exceeding the stated composition limits are as follows:

1) Exceeding the impurity limits of zinc, tin, lead, silicon, or phosphorus may cause cracking during hot-working (hot shortness) and cause problems during subsequent welding.
2) Exceeding the silicon content may cause machining difficulties later.
3) Exceeding the manganese limits may reduce the normal corrosion resistance.
4) Exceeding the aluminum content decreases ductility (leading to ductility cracking during fabrication) and corrosion resistance.

Table 14 discusses the major aluminum bronze alloys.

7.2. Copper – Silicon Alloys (Silicon Bronzes)

Silicon may be alloyed with copper up to 4 wt %, which is the limit for the alpha structure, to give the silicon bronzes (C64700 – C66100). Figure 13 is a partial copper – silicon phase diagram, which clearly illustrates this alloying limitation. Beyond this limit, the alloy has poor working qualities, exhibiting a brittleness due to the presence of the gamma phase. The addition of silicon rapidly increases the work hardening rate, but ductility in the annealed state is high. The addition of silicon decreases the density, the electrical conductivity, and the melting point of the alloys. One important attribute of silicon is that it improves the resistance to stress relaxation. The improved fluidity gives excellent welding qualities to the copper – silicon alloys.

In addition to copper and silicon, the silicon bronzes may contain other alloying elements.

Lead is added in small quantities ($\leqq$ 0.8 wt %) to help improve the machinability of the alloy. Zinc is added in a low-silicon alloy as a

Table 14. Wrought copper–aluminum alloys (wrought aluminum bronzes)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C60600	Al 4.0–7.0 Cu remainder	Fe 0.50	8.17	aluminum bronze, 5%: grade A: electrical conductivity of 17% IACS, hard-temper tensile strength of 410 MPa, available as sheet, strip, and rolled bar	fasteners, deep drawn "gold" decorations and jewelry, parts for corrosion resistance
C61000 (CuAl8)	Al 6.0–8.5 Cu remainder	Pb 0.02 Fe 0.50 Zn 0.20 Si 0.10	7.78	aluminum bronze, 8%: electrical conductivity of 15% IACS, hard-temper tensile strength of 480 MPa, available as rod and wire	bolts, shafts, pump parts and tie rods, welded overlay on steel for wear properties
C61300 (CuAl7-Fe3Sn)	Fe 2.0–3.0 Sn 0.20–0.50 Al 6.0–7.5	Pb 0.01 Zn 0.10 Mn 0.20 Si 0.10 Ni 0.15 P 0.015 for welding also: Cr 0.05 Cd 0.05 Zr 0.05 Zn 0.05	7.95	aluminum bronze, 7%: electrical conductivity of 12% IACS, hard-temper tensile strength of 590 MPa, magnetic permeability of 1.16 μ	fasteners, tube sheets, welded tubes, heat-exchanger tubes, acid-resistant piping, water boxes, other corrosion-resistant vessels
C61400 (CuAl8-Fe3)	Fe 1.5–3.5 Al 6.0–8.0 Cu remainder	Pb 0.01 Zn 0.20 Mn 1.0 P 0.015	7.89	aluminum bronze, 7%: electrical conductivity of 14% IACS, hard-temper tensile strength of 610 MPa, magnetic permeability of 1.16 μ	same as C61300
C61500	Al 7.7–8.3 Ni + Co 1.8–2.2 Cu remainder	Pb 0.015	7.69	lusterloy, gold color, electrical conductivity of 12.6% IACS, hard-temper tensile strength of 860 MPa, available in strip	hardware, decorative metal trim, deep drawn shells, jewelry
C62300 (CuAl10-Fe3)	Fe 2.0–4.0 Al 8.5–10.0	Sn 0.6 Mn 0.50 Si 0.25 Ni + Co 1.0	7.64	aluminum bronze, 9%: electrical conductivity of 12% IACS, typical tensile strength of 610 MPa, available as bar and rod	bearings, bolts, nuts, bushings, gears, cams, pump rods
C62400	Fe 2.0–4.5 Al 10.0–11.5 Cu remainder	Sn 0.20 Mn 0.30 Si 0.25	7.45	aluminum bronze, 11%: electrical conductivity of 12% IACS, typical tensile strength of 655 MPa, magnetic permeability of 1.34 μ, available as rod and bar	gears, wear plates, nuts, bolts, drift pins
C63200 (CuAl9-FeNi4)	Fe 3.5–4.3 Al 8.7–9.5 Mn 1.2–2.0 Ni + Co 4.0–4.8*	Pb 0.02 Si 0.10	7.64	nickel aluminum bronze, electrical conductivity of 7% IACS, hard-temper tensile strength of 640–720 MPa, depending on heat treatment, magnetic permeability of 1.04 μ, available as rod, bar, and forgings	nuts, bolts, shafts, propellers, spark-resistant parts
C63800	Al 2.5–3.1 Si 1.5–2.1 Co 0.25–0.5 Cu remainder	Pb 0.05 Fe 0.20 Zn 0.8 Mn 0.10 Ni + Co 0.20	8.28	coronze, electrical conductivity of 10% IACS, hard-temper tensile strength of 830 MPa, available as strip and sheet	electronic connectors, switches, springs, contacts
C64400	Al 3.5–4.5 Si 0.8–1.3 Ni + Co 4.2–5.0 Cu remainder	Pb 0.03 Fe 0.05 Sn 0.10 Zn 0.20	8.03	C644, electrical conductivity of 15% IACS, hard-temper tensile strength after precipitation heat treatment of 950 MPa, available as strip and sheet	electronic connectors and contacts, switches, springs

* Ni content must exceed Fe content.

strengthener and (possibly) an additional deoxidizer. Tin additions are made to improve corrosion resistance and improve fatigue strength. The addition of manganese to an alloy improves the microstructure through grain refinement. Manganese also improves the welding quality. Nickel, especially in the low-silicon alloys, is a strengthener. (Cobalt in small quantities is difficult to separate from nickel during refining; thus cobalt is counted as nickel.) Chromium is added to aid in grain refinement throughout the processing. Chromium may increase the alloy's resistance to stress relaxation.

For the silicon bronzes, the elements singled out as impurities include lead, iron, tin, zinc, aluminum, manganese, and nickel, although in specific alloys, one or more of these elements may be an alloying element.

Table 15 describes the major silicon bronze alloys.

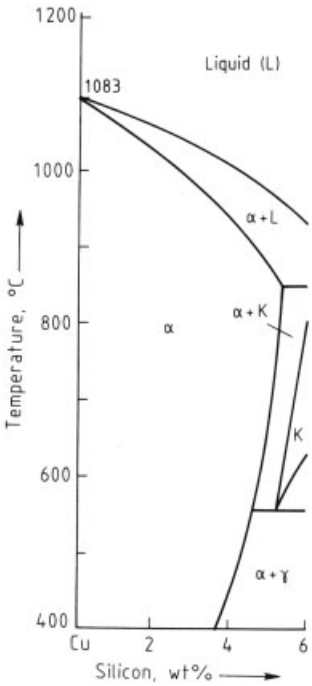

Figure 13. Copper – silicon phase diagram [30, p. 483]

7.3. Miscellaneous Copper – Zinc Alloys

More highly alloyed than the copper – zinc alloys described in Chapter 3, the alloys C66400 – C69999 include alloys with a variety of chemical compositions, mechanical and physical properties, and uses. Zinc is the most popular alloying element, and the strength imparted is generally directly proportional to the zinc content. The addition of three or more alloying elements may drastically change the properties of the alloy.

Addition of up to 3 wt % Pb, most frequently an impurity, sometimes improves machinability. Addition of up to 4 wt % Fe is quite often used as a grain refiner and as a strengthener for special products. Addition of up to 1.5 wt % Sn improves the strength of some alloys. The ductility in the annealed state also is improved by the tin. Addition of up to 2.3 wt % Ni helps to control in-process conditions and to increase strength.

Addition of up to 6 wt % Al imparts greater hardness (strengthener) and, in lower quantities, also imparts greater ductility. Except for three special-purpose alloys (C66900, C69900, and C69910), manganese is limited to 5 %. Several alloys in this series are listed as manganese brasses or manganese bronzes. Manganese is a strong deoxidizer, but it also adds strength to the alloys and is a grain refiner. The three special-purpose alloys (C66900, C69900, and C69910) were initially developed as brazing alloys and sound-deadening alloys, e.g., Incramute. However, production problems and other, more efficient products eliminated their market.

Addition of up to 4.5 wt % Si improves the work hardening rate and, thus, the strength. The resistance to stress relaxation is improved, and in addition, the resistance welding properties are greatly improved.

Where cobalt is added to a copper – zinc – aluminum alloy, it forms a fine precipitate that is an excellent strengthener. Cobalt is added in quantities up to 0.7 %. Small additions of up to 0.8 wt % As are added to this series of alloys because they contain zinc. The arsenic acts as a corrosion inhibitor to minimize dezincification.

Addition of up to 0.06 wt % P ensures total deoxidation, with some strengthening of the alloys. Phosphorus also inhibits dezincification, as does antimony.

Table 16 describes selected alloys.

8. Wrought Copper – Nickel Alloys

The last series of wrought copper alloys is the C70000 alloy series, which basically consists of

Table 15. Wrought copper–silicon alloys (wrought silicon bronzes)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C65100 (CuSi1)	Si 0.8–2.0 Cu remainder	Pb 0.05 Fe 0.8 Zn 1.5 Mn 0.7	8.75	low-silicon bronze B, electrical conductivity of 12% IACS, hard-temper tensile strength of 655 MPa, available as rod, wire, and tubing	hardware, bolts, car screws, nuts, marine hardware, heat-exchanger tubes, welding rod, hydraulic pressure lines, rivets, U-bolts
C65400	Sn 1.2–1.9 Si 2.7–3.4 Cr 0.01–0.12 Cu remainder	Pb 0.05 Zn 0.50	8.55	ultronze, electrical conductivity of 7% IACS, hard-temper tensile strength of 790 MPa, available as sheet and strip, thermal treatment enhances resistance to stress relaxation	electronic contacts, springs, connectors
C65500 (SuSi3-Mn1)	Mn 0.50–1.3 Si 2.8–3.8 Cu remainder	Pb 0.05 Fe 0.8 Zn 1.5 Ni + Co 0.6	8.53	high-silicon bronze A, electrical conductivity of 7% IACS, hard-temper tensile strength of 650 MPa, available as strip, wire, rod, and tubing	hardware, bolts, burrs, clamps, hinges, nails, nuts, screws, bearing plates, bushings, aircraft hydraulic lines, chemical containers

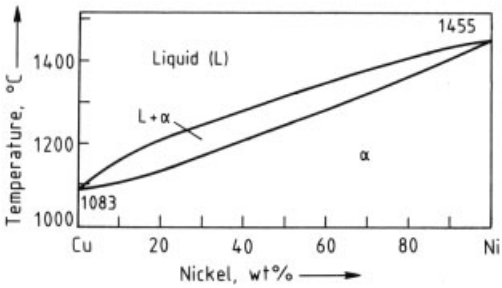

Figure 14. Copper – nickel phase diagram [30, p. 475]

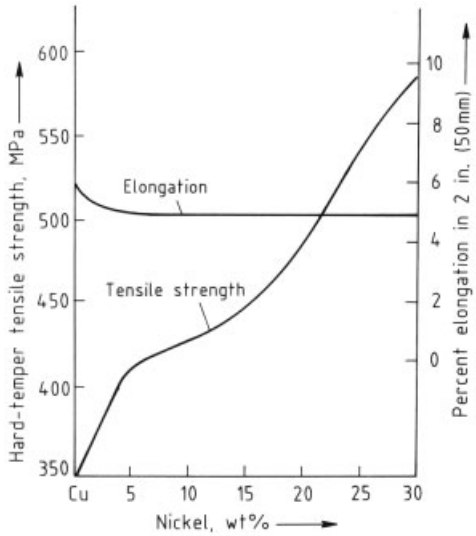

Figure 15. Strengthening of copper by alloying with nickel

the copper – nickel alloys, with other elements added for specific purposes. The alloy series is subdivided into two distinct groups:

1) Copper-nickel alloys (C70100 – C72900)
2) Copper – nickel – zinc alloys (the nickel silvers, C73200 – C79900)

Each of these groups may be divided further into two subgroups, as discussed later.

The copper – nickel equilibrium diagram, Figure 14, shows that the two elements are completely soluble in each other in both the liquid and the solid states. This permits the full range of copper – nickel (copper-rich) alloys and nickel – copper (nickel-rich) alloys.

The color of the copper – nickel alloys depends on the percentage of nickel in the alloy, ranging from the red of pure copper, through rose, to near-white as 30 wt % Ni is approached. If a color match between alloys is required, as

Table 16. Miscellaneous wrought copper–zinc alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C66400	Fe 1.3–1.7 Co 0.30–0.7 Fe + Co 1.8–2.3 Zn 11.0–12.0 Cu remainder	Pb 0.015 Sn 0.05 Ni 0.05 Al 0.05 Mn 0.05 Si 0.05 P 0.02 Ag 0.05	8.74	cobron, electrical conductivity of 30% IACS, hard-temper tensile strength of 610 MPa, available as strip and sheet	electronic contacts, switches, spring washers, fuse clips, connectors, terminals
C66700	Cu + Ag 68.5–71.5 Mn 0.8–1.5 Zn remainder	Pb 0.07 Fe 0.10	8.53	manganese brass, electrical conductivity of 17% IACS, hard-temper tensile strength of 540 MPa, available as strip, sheet, rod, and wire	resistance-weldable brass products
C67500 (CuZn39-AlFeMn)	Cu + Ag 57.0–60.0 Fe 0.8–2.0 Sn 0.50–1.5 Mn 0.05–0.50 Zn remainder	Pb 0.20 Al 0.25	8.36	manganese bronze A, electrical conductivity of 24% IACS, half-hard-temper tensile strength of 410 MPa, available as rod and hot forged parts	automotive clutch disks, pump rods, shafting, pump balls, valve stems and bodies
C68700 (CuZn20-Al2As)	Cu + Ag 76.0–79.0 Al 1.8–2.5 As 0.02–0.06 Zn remainder	Pb 0.07 Fe 0.06	8.33	arsenical aluminum brass, electrical conductivity of 23% IACS, soft-temper tensile strength of 410 MPa, available as soft tubing	condenser, evaporator, and heat-exchanger tubes, especially for salt water
C68800	Zn 21.3–24.1 Al 3.0–3.8 Co 0.25–0.55 Al + Zn 25.1–27.1 Cu + Ag remainder	Pb 0.05 Fe 0.20	8.20	alcoloy, electrical conductivity of 18% IACS, hard-temper tensile strength of 780 MPa or 750 MPa (resistant to stress relaxation), available as strip	springs, switches, electrical and electronic contacts, relays, connectors, plug recepticals
C69000	Cu + Ag 72.0–74.5 Ni + Co 0.50–0.8 Al 3.0–3.8 Zn + Al 25.1–27.1 Zn remainder	Pb 0.025 Fe 0.05	8.20	high-strength copper alloy, electrical conductivity of 18% IACS, hard-temper tensile strength of 780 MPa, available as strip	same as C68800
C69400	Cu + Ag 80.0–83.0 Si 3.5–4.5 Zn remainder	Pb 0.30 Fe 0.20	8.19	silicon red brass, electrical conductivity of 6.2% IACS, 1/8th-hard-temper tensile strength of 690 MPa, available as rod and forged parts	corrosion-resistant valve stems, forged or screw machined parts
C69910	Fe 1.0–1.4 Zn 3.0–5.0 Al 0.25–0.8 Mn 28.0–32.0 Cu + Ag remainder	Pb 0.01	7.7	cobraze, rolled-temper tensile strength of 690 MPa, available in strip (< 1 mm thick)	brazing alloy with good wetability and antigravity flow

in souvenir coinage, the alloy makeup must be carefully controlled.

The 30 wt% Ni of alloy C71500 represents the present maximum nickel content in most copper alloy plants because the high processing temperatures during casting and hot-rolling are generally the maximum operating temperatures in the plants.

As nickel is alloyed to copper in increasing quantities, the tensile strength generally increases. The percent elongation decreases initially, but levels off. These effects are illustrated in Figure 15. Therefore, the strength of the alloy is increased with little loss in ductility, as measured by the percent elongation. However, this relationship is not generally observed in commercial products, since few pure copper–

nickel alloys are used. Usually one or more other elements are alloyed with the copper – nickel, which drastically changes this relationship.

8.1. Cupronickels

The copper – nickel alloys can be further subdivided into two subgroups:

1) Alloys C70100 – C71700, the major additional elements being iron and manganese, with minor amounts of other elements
2) Alloys C71900 – C72900, the other alloying elements becoming more and more effective, generating alloys that have little resemblance to the first group

The alloys C70100 – C71700 are primarily copper – nickel alloys to which iron and/or manganese are added.

Iron is an important element in the copper – nickel alloy system. Addition of up to 2.3 wt % Fe increases the strength with little loss in ductility. Iron also greatly improves the corrosion resistance of the alloys. Where the copper – nickel alloys have been used in saltwater environments without the iron addition, the copper – nickel products rapidly corrode by pitting. Replacement by alloys with the same copper : nickel ratio, but with 1.4 wt % Fe, allowed the product to last for years.

Thermal treatments should be used to ensure that the iron is in solution in the microstructure, especially in alloys C70600 and C71500. Initially, the iron must be placed in solution, and this can be done during the hot-working stage. Subsequently, all annealing operations must be followed by a cold-water quench to ensure that the iron stays in solution. If the iron does precipitate, the corrosion resistance and formability are seriously diminished.

Up to a maximum of 2.5 wt % Mn is added as a deoxidizer and a strengthener. Beryllium is added to one alloy (C71700) to generate a high-strength copper – nickel –beryllium alloy.

In addition, a new alloy (C70320) has been developed, containing no iron or manganese but which has up to 1.2 wt % Al, 1.2 wt % Si, and 0.50 wt % Cr. These elements are strengtheners and deoxidizers, but no published data are presently available to reveal the effects of this interesting combination of elements.

The major copper – nickel alloys in this subgroup are described in Table 17.

The alloys C71900 – C72900 contain some of the most interesting alloys presently available from the copper – nickel system. Several of the alloys are hardened through a spinodal-decomposition hardening mechanism. Spinodal structures may be developed in alloy systems that have a miscibility gap, with one phase separating out of the other during cooling in the gap. The spinodal structure is formed when an alloy in the system (in this case, the copper – nickel – tin system) is heated to a temperature above the miscibility gap, held at that temperature long enough for the metal to reach equilibrium, then rapidly quenched to a temperature within the gap, and held at that temperature long enough for the spinodal decomposition to reach completion.

The spinodal structure, in this case, is attributed to the uniform dispersion of tin-rich regions in the copper-rich matrix. The structure consists of an ordered, three-dimensional array of the tin-rich crystals. Under the high magnification of the electron microscope, the homogenized structure appears as a tweed or tightly woven cloth. Some spinodal-decomposition alloys have been manufactured for commercial use, but little has been published on their properties. The only spinodal alloy now available is C72900. Other alloys in this subgroup are also described in Table 18.

8.2. Copper – Nickel – Zinc Alloys

The alloys commonly called *nickel silvers* were formerly called *German silver.* They contain no silver except as a possible impurity. The history of German silver goes back to the time when the middle classes in Europe were first forming. They wanted eating utensils (table flatware and hollowware) that had the appearance of the silver pieces used by the rich, but at a lower price. These alloys were developed in Germany to meet this need; hence, the term German silver arose.

The copper – nickel – zinc alloys are strong and ductile; therefore, they have many uses. This group may be further divided into two subgroups:

Table 17. Wrought copper–nickel alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C70400 (CuNi5-Fe1Mn)	Fe 1.3–1.7 Ni + Co 4.8–6.2 Mn 0.30–0.8 Cu + Ag remainder	Pb 0.05 Zn 1.0	8.94	copper nickel, 5%: silver-rose color, electrical conductivity of 14% IACS, hard-temper tensile strength of 440 MPa, corrosion resistant, cheaper but less corrosion resistant than C70600, available as strip and tube	condensers, evaporator and heat-exchanger tubes, plates, ferrules
C70600 (CuNi10-Fe1Mn)	Fe 1.0–1.8 Ni + Co 9.0–11.0 Cu + Ag remainder	Mn 1.0 Zn 1.0 Pb 0.05	8.94	copper nickel, 10%: silver-pink color, electrical conductivity of 9% IACS, hard-temper tensile strength of 520 MPa, exceptionally resistant to seawater, available as strip and tube	condenser, distiller tubes, evaporators and heat exchangers, ferrules, seawater boat hulls
C71000 (CuNi20-Mn1Fe)	Ni + Co 19.0–23.0 Cu + Ag remainder	Pb 0.05 Fe 1.0 Zn 1.0 Mn 1.0	8.94	copper nickel, 20%: pale silver color; electrical conductivity of 6.5% IACS, hard-temper tensile strength of 520 MPa, available as strip and tube	communication relays, electrical springs, condenser plates, automotive oil coolers, ferrules, resistors
C71300 (CuNi25)	Ni + Co 23.5–26.5 Cu + Ag remainder	Pb 0.05 Fe 0.20 Zn 1.0 Mn 1.0	8.94	copper nickel, 25%: gray-silver color, electrical conductivity of 6% IACS, hard-temper tensile strength of 530 MPa, available as strip	U.S. coinage: the five-cent piece (the nickel) and the surface of the ten-cent through one-dollar coins
C71500 (Cu-Ni-30-Mn1Fe)	Fe 0.40–1.0 Ni + Co 29.0–33.0 Cu + Ag remainder	Pb 0.05 Zn 1.0 Mn 1.0	8.94	copper nickel, 30%: gray-white color, electrical conductivity of 4.6% IACS, hard-temper tensile strength of 565 MPa, should be annealed before soldering, available as plate, strip, rod, and tube	condensers, evaporators, distiller tubes, ferrules, saltwater pipes, and other saltwater uses

1) Alloys C73200–C77600 (except C76300 and C76390, which contain lead and other alloying elements and, thus, are grouped below) are nonleaded copper–nickel–zinc alloys.
2) Alloys C78200–C79900 are copper–nickel– zinc alloys that also contain lead and additional alloying elements.

The *copper–nickel–zinc alloys* increase in strength as the nickel and zinc additions increase. Figure 16 shows this: At hard temper the tensile strength increases, while the relative elongation remains steady. The plot of the difference between the yield strength and the tensile strength, ΔS, against this combination shows that ΔS follows the tensile strength, indicating that the higher strength does not diminish the formability.

These alloys generally exhibit good corrosion resistance, but they are somewhat susceptible to stress corrosion cracking. There are a few additional elements alloyed.

Up to 0.35 wt % Mn is added in two of the alloys for deoxidation. Manganese may be added many times, but in quantities below the permitted impurity level, as a precaution and aid in deoxidation. Up to 0.25 wt % Si is added at times to improve casting fluidity and strength.

These alloys are described in Table 19.

The *leaded copper–nickel–zinc alloys* have a specific quantity of lead added to improve the machinability of the alloy.

These alloys were used extensively for the manufacture of keys, especially automotive ignition and door keys. The strength and the wearing properties of the keys made from these alloys were superior and permitted long use without

Table 18. Wrought copper–nickel and other element alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Description, properties	Uses
C71900	Ni + Co 28.0–33.0 Mn 0.20–1.0 Cr 2.2–3.0 Zr 0.02–0.35 Ti 0.01–0.20 Cu + Ag remainder	Pb 0.015 Fe 0.50 Zn 0.05 C 0.04 Si 0.25 S 0.015 P 0.02	8.85	copper nickel chromium bearing, electrical conductivity of 4.4% IACS in spinodally decomposed condition, tensile strength of 540 MPa, available in pipe and tube	heat exchangers, water boxes, ferrules
C72200	Fe 0.50–1.0 Ni + Co 15.0–18.0 Cr 0.30–0.7 Cu + Ag remainder	Pb 0.05 Zn 1.0 Mn 1.0 Si 0.03 Ti 0.03	8.94	copper nickel chromium bearing, electrical conductivity of 6.53% IACS, hard-temper tensile strength of 480 MPa, available in pipe and tube	heat exchangers, saltwater boxes
C72400	Ni + Co 11.0–15.0 Al 1.5–2.5 Mg 0.05–0.40 Cu + Ag remainder	Pb 0.05 Fe 0.10 Zn 0.50 Sn 0.05 Mn 1.0	8.59	copper nickel aluminum, electrical conductivity of 10% IACS; solution heat-treated tensile strength of 1000 MPa, available in strip	electrical and electronic contacts, connectors, springs
C72500	Ni + Co 8.5–10.5 Sn 1.8–2.8 Cu + Ag remainder	Pb 0.05 Fe 0.6 Zn 0.50 Mn 0.20	8.89	copper nickel tin bearing, electrical conductivity of 11% IACS, hard-temper tensile strength of 570 MPa, not spinodally decomposable but thermal treatment following finish rolling develops a hardening precipitate that improves formability; available in strip, rod, wire, and tube	electronic connectors and contacts, lead frames, switches, springs, sensing bellows
C72900	Ni + Co 14.5–15.5 Sn 7.5–8.5 Mn 0.05–0.30 Cu + Ag remainder	Pb 0.02 Fe 0.50 Zn 0.50 Nb 0.10 Mg 0.15	8.94	pfinodal, electrical conductivity of 8% IACS, tensile strength of 860 MPa (spinodally decomposed from hard temper), rolled by using powder-metallurgical techniques because a brittle, unremovable Ni–Sn segregate develops in the liquid state; available in strip	electronic connectors, contacts, springs

fear of the key breaking off in the lock. This resulted in the thought that a silver key was a sign of quality. However, these keys were partially replaced by less expensive keys made from nickel-plated leaded brass or aluminum. With the advent of color-anodized aluminum keys, the market was even more rapidly eroded. With the loss of this large market, the supply of alloy has shrunk to that required by a relatively small specialty market.

The alloys are expensive because they require special care in manufacturing. They cannot be hot-rolled, and they work harden rapidly, necessitating a greater number of anneals during processing.

Elements other than nickel and zinc may be added.

Lead may be added to improve the machinability of the alloy. The maximum quantity allowable is 3.5 wt %. Up to 2.5 wt % Mn may be added for deoxidation and as a strengthener. Up to 0.6 wt % Sn may be added to improve the strength of the alloy.

The most important alloy in this group is described in Table 20.

Table 19. Wrought copper–nickel–zinc alloys (wrought nickel silvers)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C73500	Cu 70.5–73.5 Ni + Co 16.5–19.5 Zn remainder	Pb 0.10 Fe 0.25 Mn 0.50	8.83	nickel silver, 18%: electrical conductivity of 8% IACS, hard-temper tensile strength of 545 MPa, normally available as strip in soft condition for deep drawing	deep drawn shells requiring high strength and good corrosion resistance, e.g., ballpoint ink tubes
C74300	Cu 63.0–66.0 Ni + Co 7.0–9.0 Zn remainder	Pb 0.10 Fe 0.25 Mn 0.50	8.65	nickel silver, 8%: electrical conductivity of 10% IACS, hard-temper tensile strength of 640 MPa, available as strip	flatware and hollow-ware for silver plate
C74500 (CuNi10-Zn25)	Cu 63.5–66.5 Ni + Co 9.0–11.0 Zn remainder	Pb 0.10 Fe 0.25 Mn 0.50	8.69	nickel silver, 65–10: electrical conductivity of 9% IACS, hard-temper tensile strength of 590 MPa, available as strip, rod, bar, and wire	rivets, screws, slide fasteners, eyeglass frames, hollowware, nameplates
C75200 (CuNi18-Zn20)	Cu 63.5–66.5 Ni 16.5–19.5 Zn remainder	Pb 0.05 Fe 0.25 Mn 0.50	8.73	nickel silver, 65–18: electrical conductivity of 6% IACS, hard-temper tensile strength of 590 MPa, available as strip, rod, and wire	rivets, screws, flat-ware for silver plat-ing, truss wire, cam-era parts, eyeglasses, jewelry, etching stock, radio dial
C75400 (CuNi15-Zn21)	Cu 63.5–66.5 Ni + Co 14.0–16.0 Zn remainder	Pb 0.10 Fe 0.25 Mn 0.50	8.69	nickel silver 65–10: electrical conductivity of 7% IACS, hard-temper tensile strength of 590 MPa, available as strip, rod, and wire	camera parts, optical parts, etching stock, jewelry
C75700 (CuNi12-Zn24)	Cu 63.5–66.5 Ni + Co 11.0–13.0 Zn remainder	Pb 0.05 Fe 0.25 Mn 0.50	8.69	nickel silver, 65–12: electrical conductivity of 8% IACS, hard-temper tensile strength of 590 MPa, available as strip, bar, rod, and wire	same as C75400
C77000 (CuNi18-Zn21)	Cu 53.5–56.5 Ni + Co 16.5–19.5 Zn remainder	Pb 0.05 Fe 0.25 Mn 0.50	8.70	nickel silver, 55–18: electrical conductivity of 5.5% IACS, hard-temper tensile strength of 690 MPa, available as strip, rod, and wire	optical goods, springs, resistance wire

Table 20. Wrought copper–nickel–zinc–lead alloys (leaded nickel silvers)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Description, properties	Uses
C78200	Cu 63.0–67.0 Pb 1.5–2.5 Ni + Co 7.0–9.0 Zn remainder	Fe 0.35 Mn 0.50	8.69	leaded nickel silver, 65–8: electri-cal conductivity of 10.9% IACS, hard-temper tensile strength of 590 MPa, machinability of 60%, available as strip	key blanks, watch parts

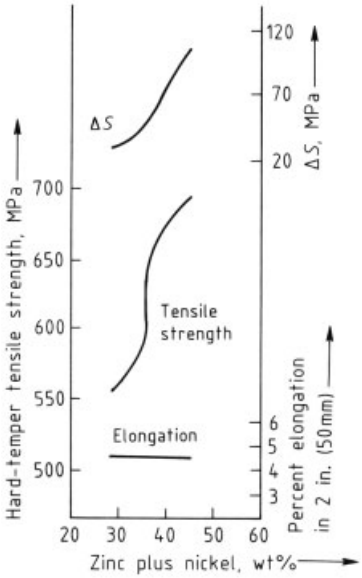

Figure 16. Strengthening of copper by alloying with a combination of zinc and nickel [26]
ΔS is the difference between the yield strength and the tensile strength

9. Casting Alloys

The C80000 and C90000 series of alloys constitute the copper alloys that are supplied as castings. When a product cannot be manufactured from wrought alloy, casting is the next choice. Castings are used generally because the desired configuration cannot be made economically from wrought material either by rolling, stamping, drawing, forming, extruding, or a combination of these metal-forming operations. These cast alloys differ from the wrought alloys in several respects.

1) The cast alloys are not worked easily, either hot or cold; therefore, casting must be near final shape except for simple machining.
2) Although some of the same alloying elements are used in both the wrought and the cast alloys, alloying elements such as lead, tin, iron, and aluminum are generally present in greater amounts in castings than in wrought alloys. The alloying elements are added in casting alloys to improve the metal flow, the cast structure, or the strength. Some alloying elements cannot be used in wrought alloys because of the problems they would cause in processing and end use.
3) The impurity levels in casting alloys can be, and generally are, higher because the material is not worked.
4) The electrical conductivity of cast products is lower than that of wrought products because of the structure, even for pure copper.

In selecting an alloy for a part, many factors must be considered: strength, machinability, corrosion resistance, fatigue resistance, castability, metal cost, dimensions, tolerances, final cast, resistance to wear, lubricity in operation, etc.

9.1. Cast Coppers

The cast coppers C80000 – C81200 are used in the as-cast condition. They are not easily cast, usually having a narrow solidification range requiring extra risers to facilitate casting. At times, chills, insulating sleeves, and topping compounds help and may reduce the required risers.

Cast coppers are soft and easily deformed, especially when compared to higher alloyed casting alloys. Since they are castings, they must be machined to obtain the necessary fit. The low machinability ratings (e.g., 10 % FMB) make the cast coppers difficult to work with. Since the turnings are long, tough, and stringy, a chip breaker is helpful. Sulfur-free coolants should be employed to avoid contamination. If sulfurized oils must be used, parts must be cleaned immediately after machining to avoid staining.

Alloy C80500 contains at least 0.034 wt % Ag. None of the cast coppers can be heat treated to improve the hardness, but the addition of silver, as in the wrought copper alloys, contributes to a higher softening temperature and improved elevated-temperature properties. The alloy is easily soldered and brazed, but has limited weldability. Alloy C80500 generally corrodes like copper.

Alloy C81100 has about the same characteristics as C80500 except, lacking silver, it does not have the higher softening temperature and improved elevated-temperature properties.

Further information on these two cast coppers is found in Table 21.

Table 21. Cast coppers

UNS (ISO) alloy	Alloying elements (min.), wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting*	Description, properties	Uses
C80500	Cu + Ag 99.75 Ag 0.034	B 0.02	8.94	ABDEFG	boron-deoxidized copper, electrical conductivity of 91 % IACS, typical tensile strength of 170 MPa, machinability of 10 %	electrical and thermal conductors, heat sinks, corrosion- and oxidation-resistant uses
C81100	Cu + Ag 99.70	P + Si 0.01	8.94	ABDEFG	copper, electrical conductivity of 92 % IACS, typical tensile strength of 170 MPa, machinability of 10 %	electrical and thermal conductors, heat sinks, corrosion- and oxidation-resistant uses

* Methods of casting: A = centrifugal; B = continuous; C = die; D = investment; E = permanent mold; F = plaster; G = sand; H = other method.

9.2. High-Copper Alloys

The high-copper alloy castings, C81300 – C82900, generally contain ≧ 94 wt % Cu, with additions of beryllium or chromium to permit heat treating to improve the strength of the casting. The general characteristics of the alloys are the same as for the cast coppers, except that the higher alloying contents lower the thermal and electrical conductivities. As with the coppers, machining is difficult and requires special attention. Parts that may be solution annealed and aged are best machined after the solution anneal, with grinding after aging to obtain a proper fit. Care must be maintained during machining and grinding to avoid overheating the part even slightly; overheating can cause hard and/or soft spots to occur.

Alloy C81400 is a low beryllium – chromium copper alloy, and alloy C81500 is a chromium copper alloy. The latter has a much higher electrical – thermal conductivity, since beryllium drastically affects conductivity. Alloy C81500 is the more desirable alloy of the two.

Alloys C82000, C82200, C82500, and C82800 are coppers with varying contents of beryllium, cobalt, and nickel. As total alloying content increases, the strength of the casting increases. Alloys C82500 and C82800 have small additions of silicon that aid in castability, with some additional strengthening of the cast.

Further detailed information on the high-copper casting alloys is located in Table 22.

9.3. Copper – Tin – Zinc and Copper – Tin – Zinc – Lead Alloys

The C83000 – C85900 category of copper alloy castings is varied, covering a broad range of alloys. The lower numbered alloys have lower zinc contents. There are three subgroups: the "red brasses" (C83000 – C83900), the "semi-red brasses" (C84000 – C84900), and the "yellow brasses" (C85000 – C85900).

The alloys between C83000 and C83900 were formerly called red or leaded red brasses. These moderate-strength (240 MPa) alloys are mostly specialty alloys having specialized industrial applications (except for C83600 and C83800). They contain enough lead to greatly improve the machinability (34 % FMB for C83300 to 90 % FMB for C83800), which is important for making the casting useful.

The two most popular alloys, C83600 and C83800, are readily cast with normal quality casting procedures. Heavy cross sections up to 5 cm must be specially handled because of the wide solidification range – approximately 140 °C. A mushy layer forms during solidification and can result in micro or interdentritic shrinkage. However, special casting procedures can prevent this shrinkage. Alloy C83800 has somewhat lower mechanical properties, but its better machinability and lower cost sometimes favor it over C83600. Both alloys are readily joined by using soft soldering and brazing, although external stress should be avoided during brazing, for any stress may cause fractur-

Table 22. High-copper alloys – castings

UNS (ISO) alloy	Alloying elements*, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Method of casting**	Description, properties	Uses
C81400	Cu 98.5 Be 0.02–0.10 Cr 0.6–1.0		8.81	ABDEFG	beryllium-modified chrome copper, electrical conductivity of 70% IACS, typical tensile strength of 365 MPa (heat treated); Be ensures Cr kept under control, allowing production of consistently high-quality chrome copper castings; machinability of 20%	electrical components meeting RWMA class II standards
C81500	Cu 98.0 Cr 0.40–1.5	Si 0.15 Fe 0.10 Al 0.10 Sn 0.10 Pb 0.02 Zn 0.10	8.81	ABDEFG	chrome copper, electrical conductivity of 90% IACS, typical tensile strength of 365 MPa (heat treated), machinability of 20%	electrical or thermal conducts requiring greater strength than C81100
C82000	Cu 95.0 Be 0.45–0.8 Co + Ni 2.4–2.7	Si 0.15 Ni 0.20 Fe 0.10 Al 0.10 Sn 0.10 Pb 0.02 Zn 0.10 Cr 0.10	8.62	ABDEFGH	beryllium copper casting alloy 10C, electrical conductivity of 48% IACS, typical tensile strength of 690 MPa (heat treated), machinability of 40%	resistance welding tips, holders, and arms; circuit-breaker parts, switch-gear parts, soldering-iron tips
C82200	Cu 96.5 Be 0.35–0.8 Ni 1.0–2.0		8.75	ABDEFG	beryllium copper casting alloy 30C, 35C, or 53B, electrical conductivity of 48% IACS, typical tensile strength of 655 MPa (heat treated), machinability of 40%	seam welder electrodes, spot-welding tips, beam welder shapes, water-cooled holders, clutch rings, brake drums, arms bushings for resistance welding
C82500	Cu 95.5 Be 1.90–2.15 Co + Ni 0.35–0.9 Si 0.20–0.35	Ni 0.20 Fe 0.25 Al 0.15 Sn 0.10 Pb 0.02 Zn 0.10 Cr 0.10	8.26	ACDEFGH	beryllium copper 20C, electrical conductivity of 20% IACS, typical tensile strength of 1100 MPa (heat treated), machinability of 30%	mold for forming plastics, die casting plunger tips, safety tools, bearings, gears, wear parts; communication, textile, aerospace; firearms and ordnance parts
C82800	Cu 94.8 Be 2.50–2.75 Co + Ni 0.35–0.7 Si 0.20–0.35	Ni 0.20 Fe 0.25 Al 0.15 Sn 0.10 Pb 0.02 Zn 0.10 Cr 0.10	8.09	ADEFGH	beryllium copper alloy 275C, electrical conductivity of 19% IACS, typical tensile strength of 1140 MPa (heat treated), machinability of 30%	high-fluidity alloy for molds for forming plastics where casting must replicate the details with fidelity; cams, bushings, valve and pump parts; firearms and ordnance parts

* Minimum, except for ranges.
** Method of casting codes are explained in Table 21.

ing. These two alloys constitute the largest tonnage of copper casting alloys. Where corrosion resistance, leak tightness, and high operating temperatures are required, the tin bronze alloys (C90000 – C91900) are preferable.

The alloys between C84000 and C84900 were formerly called semi-red or leaded semi-red brass. The zinc content is greater, while the copper content is lower than for the first subgroup, but these alloys still have a red hue. Higher lead contents improve machinability.

The alloys also have a broad solidification range and the resultant shrinkage problems. In addition, these alloys tend to retain gas, leading to additional unsoundness. To control the gas and improve castability, small quantities of a deoxidizer (e.g., 15 wt % P – Cu) are added to the melt prior to casting. These alloys also suffer from hot shortness. Alloys C84400 and C84800 are the most popular alloys of this subgroup. Generally speaking, these alloys are low in price and readily worked.

The third subgroup of the copper – tin – zinc – lead alloys, C85000 and C85900, were formerly called the yellow and yellow leaded brasses because, having less copper and more zinc, the alloys take on the yellow hue of brass. These alloys are used extensively for die casting. This alloy subgroup is represented by C85200 and C85700. The high zinc content makes the metal cost less than the other C83000 – C85900 alloys. The two are readily cast. Gas absorption is insignificant because of the built-in protection of the evolving zinc vapors. However, the evolving zinc vapors must be rapidly oxidized; otherwise, the zinc severely corrodes the metallic structures, such as the casting dies, leading to a short die life.

When increased fluidity is desired to produce a smoother surface, up to 0.25 wt % Al can be added. All the alloys C83000 – C85900 may be readily nickel or chromium plated. Additional information on these three subgroups of alloys is found in Tables 23, 24, 25.

9.4. Copper – Manganese Alloys

The casting alloys C86000 – C86900 are sometimes called manganese or leaded manganese bronze alloys. Lead improves the machinability but a loss in strength may result when the lead content exceeds 0.1 wt %.

The chemical composition must be maintained during melting, especially the copper and zinc percentages. The foundry practice of "flaring," or heating the melt until zinc distills from the melt, resulting in the loss of zinc, changes the chemical composition and thus the mechanical properties of the casting, and should be avoided. Rather, the proper pouring temperature should be determined and maintained by better foundry practice.

The two most common alloys in this group are C86300 and C86500. The former has high tensile strength (760 MPa), while the latter has low tensile strength (450 MPa). Both have good corrosion resistance, especially to seawater. Alloy C86500 may be soldered or brazed with good results, but alloy C86300 does not lend itself to soldering or brazing because of the beta phase present, although it may be joined by using coated-metal arc welding. Stress relief annealing following joining is recommended for either alloy.

Further information on these two cast alloys is located in Table 26.

9.5. Copper – Zinc – Silicon Alloys

The casting alloys C87000 – C89900 are also called silicon bronze or silicon brass alloys. They have relatively high strength (average 600 MPa tensile strength) with good corrosion and wear resistance. The two alloys C87500 and C87800 have very similar chemical composition, but quite different applications.

Alloy C87500 is a general casting alloy, having excellent foundry properties. It has good machinability (50 % FMB) with moderately high strength (460 MPa). It has good weldability with both oxyacetylene and metallic arc methods, and has excellent soldering and brazing characteristics. Alloy C87500 is also highly resistant to dezincification under most corrosive environments.

Alloy C87800 is a die-casting alloy. Its chemical composition must be carefully controlled, especially with regard to impurities. It is a high melting point alloy – 870 °C – that requires cold-chamber die-casting machines at high pressures and lower speeds than either aluminum or zinc die-casting alloys. The high temperature contributes to die deterioration, a problem that die casters have overcome mainly with refined casting techniques. Joining the alloy can be a problem, since brazing is only fairly successful, while other methods are not recommended. Alloy C87800 has poor machinability (30 % FMB), but this is sufficient for its uses. The alloy is used where the greatest strength and the highest resistance to mechanical wear are needed in a die-cast part.

Table 23. Copper–tin–zinc–lead casting alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Method of casting**	Description, properties	Uses
C83600 (CuPb5-Sn5Zn5)	Cu + Ni 84.0–86.0 Sn 4.0–6.0 Pb 4.0–6.0 Zn 4.0–6.0	Fe 0.30 Sb 0.25 Ni + Co 1.0 S 0.08 P 0.05 Al 0.005 Si 0.005	8.83	ABDG	ounce metal, 85-5-5-5: electrical conductivity of 15% IACS, typical tensile strength of 255 MPa, machinability of 84%	general casting alloy that is sound and machinable, low-pressure valves, pipe fittings, plumber hardware, small gears, gasoline and oil line fittings
C83800	Cu + Ni 82.0–83.8 Sn 3.3–4.2 Pb 5.0–7.0 Zn 5.0–8.0	Fe 0.30 Sb 0.25 Ni + Co 1.0 S 0.08 P 0.03* Al 0.005 Si 0.005	8.6	ABG	hydraulic bronze, 83-4-6-7: electrical conductivity of 15% IACS, typical tensile strength of 240 MPa, machinability of 90%	same as C83600 except higher machinability rating, plumbing fittings, pumps, hardware, carburetors

* For continuous casting, P = 1.5 wt% max.
** Method of casting codes are explained in Table 21.

Table 24. Semi-red brass and leaded semi-red brass casting alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm^3	Method of casting**	Description, properties	Uses
C84400	Cu + Ni 78.0–82.0 Sn 2.3–3.5 Pb 6.0–8.0 Zn 7.0–10.0	Fe 0.40 Sb 0.25 Ni + Co 1.0 S 0.08 P 0.02* Al 0.005 Si 0.005	8.7	ABG	value metal, 81-3-7-9: electrical conductivity of 16.4% IACS, typical tensile strength of 230 MPa, machinability of 90%	low-pressure valves and fittings, general hardware fittings, ornamental castings, plumbing supplies and fixtures
C84800	Cu + Ni 75.0–77.0 Sn 2.0–3.0 Pb 5.5–7.0 Zn 13.0–17.0	Fe 0.40 Sb 0.25 Ni + Co 1.0 S 0.08 P 0.02* Al 0.005 Si 0.005	8.6	AG	plumbing goods brass, 76-2 1/2-6 1/2-5: electrical conductivity of 16.4% IACS, typical tensile strength of 260 MPa, machinability of 90%	plumbing cocks, faucets, stops, and wastes, air and gas line fittings, general hardware fittings, low-pressure valves and fittings

* For continuous casting, P = 1.5 wt% max.
** Method of casting codes are explained in Table 21.

Further information on these two casting alloys is found in Table 27.

9.6. Copper – Tin Alloys

The alloys C90000 – C91900 are commonly called tin bronzes. For some purposes they may have small quantities of zinc ($\leqq 5\,\%$), nickel ($\leqq 2\,\%$), or phosphorus ($\leqq 1.2\,\%$). In general, good foundry practices must be followed since the alloys have a wide solidification range (160 °C) requiring risers, chillers, and other practices to control shrinkage. Topping compounds and the use of a small quantity of phosphor copper as a deoxidizer aid in casting. Care must be taken in brazing the alloys because the brazing temperature is in the hot short range, and any stress applied to the casting during heating or cooling in this region can cause fracturing. The machinability of these alloys is relatively poor (20 – 30 FMB), requiring special attention.

There are three major alloys in this group. The first is C90500, which contains an addition of zinc to increase the strength and hardness of the alloy. The alloy is a moderately strong (320-MPa

Table 25. Yellow brass and leaded yellow brass casting alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Method of casting*	Description, properties	Uses
C85200	Cu 70.0–74.0 Sn 0.7–2.0 Pb 1.5–3.8 Zn 20.0–27.0	Fe 0.6 Sb 0.20 Ni + Co 1.0 S 0.05 P 0.02 Al 0.005 Si 0.05	8.50	AB	high-copper yellow brass, 71-1-3-24: electrical conductivity of 18.6% IACS, typical tensile strength of 260 MPa, machinability of 80%	plumbing fittings and fixtures, ferrules, low-pressure valves, hardware fittings, ornamental brass, chandeliers, irons
C85700 (CuZn33-Pb2)	Cu 58.0–64.0 Sn 0.50–1.5 Pb 0.8–1.5 Zn 32.0–40.0	Fe 0.7 Ni + Co 1.0 Al 0.8 Si 0.05	8.40	AEFG	leaded yellow brass, 63-1-1-35: electrical conductivity of 22% IACS, typical tensile strength of 340 MPa, machinability of 80%	bushings, hardware fittings, ornamental castings, lock hardware

* Method of casting codes are explained in Table 21.

Table 26. Manganese and leaded manganese bronze casting alloys

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Method of casting*	Description, properties	Uses
C86300 (CuZn25-Al6Fe3-Mn3)	Cu 60.0–66.0 Zn 22.0–28.0 Fe 2.0–4.0 Al 5.0–7.5 Mn 2.5–5.0	Sn 0.20 Pb 0.20 Ni + Co 1.0	7.7	ADEFG	manganese bronze (110 000 psi), electrical conductivity of 9% IACS, typical tensile strength of 820 MPa, machinability of 8%	extra-heavy-duty, high-strength alloy for gears, cams, bearings, screw-down nuts, hydraulic cylinder parts
C86500 (CuZn35-AlFeMn)	Cu 55.0–60.0 Zn 36.0–42.0 Fe 0.40–2.0 Al 0.50–1.5 Mn 0.10–1.5	Sn 1.0 Pb 0.40 Ni + Co 1.0	8.3	ADEFG	manganese bronze (65 000 psi), electrical conductivity of 21% IACS, typical tensile strength of 490 MPa, machinability of 26%	parts in contact with fresh or salt water requiring strength and toughness, e.g., propeller blades, hubs, gears, liners

* Method of casting codes are explained in Table 21.

tensile strength) general-purpose bronze casting alloy. It has many uses where there are severe working conditions under heavy pressure, as in bearings and gears.

The second alloy is C90700, which is a tin bronze that has been deoxidized with a trace of phosphorus. The alloy is known for its combination of good strength, ductility, and corrosion resistance. The alloy must be given additional attention during casting, because of shrinkage, and upon chilling, to reduce the grain growth that occurs upon slow cooling. The alloy has no lead, making machining difficult. The alloy is best mated to hardened steel. Machining accuracy is required so that both parts are properly mated. This alloy is especially useful in worm gears.

The third alloy is C91700, which contains, in addition to the copper and tin, a small quantity of nickel as a strengthener and a trace of phosphorus as a deoxidizer. This alloy also has a good combination of strength, ductility, and resistance to wear. Widely used for gears, it was called nickel gear bronze.

Further detailed information on these three casting alloys is found in Table 28.

9.7. Copper – Tin – Lead Alloys

The alloys C92000 – C94500 are essentially an extension of the copper – tin casting alloys, with additions of lead to improve machinability. These alloys have the same casting and joining characteristics as the copper – tin alloys (Sec-

Table 27. Copper – silicon cast alloys (silicon bronzes and brasses)

UNS (ISO) alloy	Alloying elements*, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Method of casting**	Description, properties	Uses
C87500	Zn 12.0–16.0 Si 3.0–5.0 Cu 79.0	Pb 0.50 Al 0.50	8.28	ACDEFG	silicon brass, 82-4-14: electrical conductivity of 6.7% IACS, typical tensile strength of 460 MPa, machinability of 50%	bearing races, brush holders, bearings, gears, impellers, rocker arms, valve stems, small-boat propellers
C87800	Zn 12.0–16.0 Si 3.8–4.2 Cu 80.0	Sn 0.25 Pb 0.15 Fe 0.15 Al 0.15 Mn 0.15 Mg 0.01 Ni + Co 0.20 S 0.05 P 0.01 As 0.05 Sb 0.05	8.3	C	die-cast silicon brass, 500: electrical conductivity of 6.7% IACS, typical tensile strength of 590 MPa, machinability of 40%	high-strength thin-wall castings, lever arms, brush holders, brackets, clamps, hexagonal nuts

* Minimum, except for ranges.
** Method of casting codes are explained in Table 21.

Table 28. Copper – tin cast alloys (tin bronzes)

UNS (ISO) alloy	Alloying elements, wt%	Permissible impurities (max.), wt%	Density, g/cm³	Method of casting**	Description, properties	Uses
C90500 (CuSn10-Zn2)	Cu + Ni 86.0–89.0 Sn 9.0–11.0 Zn 1.0–3.0	Pb 0.30 Fe 0.20 Sb 0.20 Ni + Co 1.0 S 0.05 P 0.05 Al 0.005 Si 0.005	8.72	ABDG	gun metal, 88-10-0-2: electrical conductivity of 11% IACS, typical tensile strength of 310 MPa, machinability of 30%	bushings, bearings, pump impellers, piston rings, valve components, steam fittings, gears
C90700 (CuSn-10P)	Cu 88.0–90.0 Sn 10.0–12.0	Pb 0.50 Zn 0.50 Fe 0.15 Sb 0.20 Ni + Co 0.50 S 0.05 P 0.30* Al 0.005 Si 0.005	8.77	ABDEG	tin bronze, 65: electrical conductivity of 9.6% IACS, typical tensile strength of 380 MPa (permanent mold) or 300 MPa (sand cast), machinability of 20%	worm wheels and gears, bearings for heavy loads at low speeds
C91700 (CuSn12-Ni2)	Cu 84.0–87.0 Sn 11.3–12.5 Ni + Co 1.2–2.0	Pb 0.25 Zn 0.25 Fe 0.20 Sb 0.20 S 0.05 P 0.30* Al 0.005 Si 0.005	8.75	ABDEG	nickel gear bronze, 86 1/2-12-0-0-1 1/2: electrical conductivity of 10% IACS, typical tensile strength of 410 MPa (permanent mold) or 300 MPa (sand cast), machinability of 20%	worm wheels and gears, bearings for heavy loads at low speeds

* For continuous casting, P = 1.5 wt% max.
** Method of casting codes are explained in Table 21.

tion9.6). Zinc, nickel, and phosphorus may be added.

These alloys may be placed into two subgroups. The first has low lead contents (C92000 – C92900), whereas the second has high lead contents (C93000 – C94500). The addition of lead slightly weakens the alloys, possibly by 25 MPa of this tensile strength, but the improved machinability is the driving force for their use.

Three low-leaded tin bronzes from the first subgroup are important. Alloy C92300 contains zinc as a strengthener. The alloy possesses a good combination of strength (275-MPa tensile), good corrosion resistance, and good machinability (42 % FMB). Alloy C92500 contains nickel as a strengthener and is deoxidized with phosphorus. Its characteristics are very similar to C90700 except the lead improves the machinability rating from 20 to 30 % FMB, thus enabling better mating with hardened steel. The strength remains essentially the same – 300 MPa tensile. Alloy C92900 has higher nickel and lead contents than C92500, with a correspondingly higher machinability rating of 40 % FMB.

The higher lead-containing tin bronze casting alloys are represented by four alloys. Alloy C93200 has zinc add for improved strength, hardness, and castability. The lead content improves the machinability rating to 70 % FMB and places the alloy in the area of having excellent anti-friction characteristics with excellent load capacity. Alloy C93500 has zinc and lead additions. The lead content is sufficient to place this alloy in the excellent anti-friction region, ideal for manufacturing bearings and bushings that operate below 230 °C. Alloy C93700 has no strengthening additions to its basic copper –tin– lead alloy makeup. The alloy has exceptionally high casting fluidity; therefore, special precautions must be taken when sand casting to minimize mold-sand penetration. The high machinability rating – 80 FMB – indicates the alloy is easily machined and has excellent anti-friction properties. Alloy C94500 is 19 wt % Pb, a very high lead content. Cast bearings and bushings made from this alloy have superior resistance to scoring and seizure, and exceed the endurance and danger limits of the regular bearing bronzes. Applications involve high speeds, poor lubrication, gritty environments, and heat generating loads at elevated temperatures. The lead is in globular form, thus permitting self-lubrication.

Further information about these alloys is detailed in Tables 29 and 30.

9.8. Copper – Tin – Nickel Alloys

The small group of copper – tin – nickel casting alloys (C94600 – C94900) forms a special niche in the casting alloys. Their use is limited to areas where heat treatment is used to obtain tensile strengths of 450 MPa. They have easy castability, low shrinkage, and a fine grain structure.

Alloy C94700 may contain a small amount of zinc to aid in strength and castability. In the tempered condition, nominal strengths of 450 MPa are reached, but the alloy has a machinability rating of only 30 % FMB. Alloy C94800 is essentially the same but with some lead for improved machinability (45 % FMB). The lead, however, reduces the tensile strength to 410 MPa.

Since heat treatment normally causes distortion, the higher machinability rating is important even with the reduction in the mechanical properties. Both alloys are difficult to weld, but they may be joined by silver or soft soldering.

Further details on these two alloys are located in Table 31.

9.9. Copper – Aluminum – Iron and Copper – Aluminum – Iron – Nickel Alloys

The C95000 – C95900 series of alloys is commonly referred to as the aluminum bronzes; these generally are high-strength alloys. The aluminum bronze alloys are sufficiently interesting to warrant consideration in most cases where a casting alloy is to be employed. Some of their high strength is available in the as-cast condition, and some is derived from heat treatment.

The three basic aluminum bronze alloys are C95300, C95500, and C95800. All have fairly high strength and hardness, excellent corrosion resistance, good wearing characteristics, good machinability, and good fatigue properties. Alloys C95300 and C95500 may be used to 400 °C, while C95800 is restricted to 370 °C.

Alloy C95300 is a heat-treatable casting alloy. When cast, the typical tensile strength is

Table 29. Copper – tin – lead casting alloys (leaded tin bronzes)

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting**	Description, properties	Uses
C92300	Cu + Ni 85.0–89.0 Sn 7.5–9.0 Pb 0.30–1.0 Zn 2.5–5.0	Fe 0.25 Sb 0.25 Ni + Co 1.0 S 0.05 P 0.05* Al 0.005 Si 0.005	8.8	ABG	leaded navy "G" bronze, 87-8-1-4: electrical conductivity of 12% IACS, typical tensile strength of 280 MPa, machinability of 42%	strong general utility structural bronze for severe conditions, e.g., valves, expansion joints, special high-pressure pipe fittings, steam pressure castings
C92500 (CuSn12-Pb2)	Cu 85.0–88.0 Sn 10.0–12.0 Pb 1.0–1.5 Ni + Co 0.8–1.5	Zn 0.50 Fe 0.30 Sb 0.25 S 0.05 P 0.30* Al 0.005 Si 0.005		ABEG	leaded tin bronze, 82-11-1-0-1: typical tensile strength of 300 MPa, machinability of 30%	gears, automotive synchronizer gears and rings
C92900	Cu 82.0–86.0 Sn 9.0–11.0 Pb 2.0–3.2 Ni + Co 2.8–4.0	Zn 0.25 Fe 0.20 Sb 0.25 S 0.05 P 0.50* Al 0.005 Si 0.005	8.87	ABEG	leaded nickel–tin bronze, 84-10-2 1/2-0-3 1/2: electrical conductivity of 9.2% IACS, typical tensile strength of 320 MPa (permanent mold or sand cast), machinability of 40%	gears, wear plates, guides, cams

* For continuous casting, P = 1.5 wt%, max.
** Method of casting codes are explained in Table 21.

520 MPa. After heat treatment, this is raised to 590 MPa. Products from this alloy are 10% lighter than comparable products of tin bronze. Casting the alloy requires experience with careful control and special techniques to maintain sound castings. The alloy is readily joined by most welding, brazing, or soldering techniques.

Alloy C95500 is a heat-treatable casting alloy. In the cast state it typically exhibits a tensile strength of 690 MPa, but after heat treatment this is increased to 830 MPa. It also has a lower density than the tin bronzes. The alloy is more easily cast than C95300, but more care must be used when joining the alloy. The alloy is readily welded, but is rated only fair for brazing. Soldering and oxyacetylene welding are not recommended.

Alloy C95800 is not heat-treatable. A typical tensile strength in the cast state is 655 MPa. The slightly lower aluminum and higher nickel contents degrade the density advantage of the previous two alloys. As with alloy C95300, care must be taken during casting to develop sound castings. The alloy is suitable for joining by soldering and some welding techniques, but brazing is rated only as fair, while carbon arc or oxyacetylene welding should be avoided.

Further information is detailed in Table 32.

9.10. Copper – Nickel – Iron Alloys

The group C96000 – C96900 not only covers casting alloys containing copper, nickel, and iron but also those containing beryllium. The standard copper – nickel – iron alloy is C96400, which gives strong (470 MPa) and ductile castings having excellent corrosion resistance. The iron addition improves the corrosion and erosion resistance in seawater, making the alloy a preferred one for use on oceangoing ships. In addition, the alloy may be joined by brazing, soldering, or welding, except for oxyacetylene or carbon arc welding, which are not recommended.

The beryllium-containing alloys are an outgrowth of alloy C96400, with the beryllium present to permit heat treating. Alloys C96600 and C96700 show the same corrosion attributes of C96400, but after heat treatment, tensile

Table 30. Copper – tin – lead cast alloys (high-lead tin bronzes)

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting**	Description, properties	Uses
C93200 (CuSn7-Pb7Zn3)	Cu + Ni 81.0–85.0 Sn 6.3–7.5 Pb 6.0–8.0 Zn 1.0–4.0	Fe 0.20 Sb 0.35 Ni + Co 1.0 S 0.08 P 0.15* Al 0.005 Si 0.005	8.93	ABEG	high-lead tin bronze, bearing bronze 83-7-7-3: electrical conductivity of 12% IACS, typical tensile strength of 240 MPa, machinability of 70%	general utility bearings and bushings, automotive fittings
C93500 (CuPb9-Sn5)	Cu + Ni 83.0–86.0 Sn 4.3–6.0 Pb 8.0–10.0	Zn 2.0 Fe 0.20 Sb 0.30 Ni + Co 1.0 S 0.08 P 0.05* Al 0.005 Si 0.005	8.87	ABG	high-lead tin bronze, 85-5-9-1: electrical conductivity of 15% IACS, typical tensile strength of 220 MPa, machinability of 70%	small bearings and bushings, bronze backings for babbitt automotive bearings
C93700 (CuPb10-Sn10)	Cu + Ni 78.0–82.0 Sn 9.0–11.0 Pb 8.0–11.0	Zn 0.8 Fe 0.15 Sb 0.55 Ni + Co 1.0 S 0.08 P 0.15* Al 0.005 Si 0.005	8.91	ABEG	high-lead tin bronze, bushing and bearing bronze 80-10-10: electrical conductivity of 10% IACS (varies strongly with temperature), typical tensile strength of 240 MPa, machinability of 80%	bearings for high speed and high pressure, pumps, impellers, corrosion-resistant uses, pressure-tight castings
C94500	Sn 6.0–8.0 Pb 16.0–22.0 Cu + Ni remainder	Zn 1.2 Fe 0.15 Sb 0.8 Ni + Co 1.0 S 0.08 P 0.05* Al 0.005 Si 0.005	9.4	ABG	medium bronze, electrical conductivity of 10% IACS, typical tensile strength of 170 MPa, machinability of 80%	locomotive wearing parts, high-load low-speed uses

* For continuous casting, P = 1.5 wt % max.
** Method of casting codes are explained in Table 21.

strengths of 830 MPa and 860 MPa are obtainable. One basic recommendation is that the machining be carried out after the parts are completely heat treated to avoid problems with fit.

Further details on these alloys are given in Table 33.

9.11. Copper – Nickel – Zinc – Tin – Lead Alloys

These alloys C97000 – C98900 form a group that is used extensively for ornamental and decorative purposes where a white alloy is desired. The alloys are responsive to machining and exhibit good corrosion resistance, but they are not heat treatable to increase strength or hardness. The two alloys discussed from this group vary widely in chemical composition.

The first alloy is C97300. This alloy has good foundry characteristics, castings are readily machined (70 % FMB), and it may be joined using soldering and brazing methods. The high lead content precludes welding. The alloy is not a strong one, with a typical tensile strength of 240 MPa.

The second alloy is C97800. This alloy has only fair castability and fluidity, but with proper techniques and good foundry practices, excellent castings can be produced. The same joining procedures are used for this alloy as for C97300.

Table 31. Copper – tin cast alloys (tin bronzes)

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting*	Description, properties	Uses
C94700	Cu 85.0–90.0 Sn 4.5–6.0 Zn 1.0–2.5 Ni 4.5–6.0	Pb 0.10 Fe 0.25 Sb 0.15 Mn 0.20 S 0.05 P 0.05 Al 0.005 Si 0.005	8.87	ABDEG	constructional bronze; electrical conductivity of 20% IACS, typical tensile strength of 340 MPa (as cast), 450 MPa (tempered), and 590 MPa (heat treated), machinability of 25% FMB in tempered state, 30% FMB as cast	gears, cams, valves and valve stems, feeding mechanisms, roller guides, cylinders for presses, conveyer chains and general construction castings
C94800	Cu 84.0–89.0 Sn 4.5–6.0 Pb 0.30–1.0 Zn 1.0–2.5 Ni 4.5–6.0	Fe 0.25 Sb 0.15 Mn 0.20 S 0.05 P 0.05 Al 0.005 Si 0.005	8.87	EG	leaded nickel tin bronze, electrical conductivity of 14% IACS, typical tensile strength of 310 MPa (as cast) and 410 MPa (tempered), machinability of 45% FMB	essentially same as C94700 except where better machinability is required

* Method of casting codes are explained in Table 21.

Table 32. Copper–aluminum–iron and copper–aluminum–iron–nickel cast alloys (aluminum bronzes)

UNS (ISO) alloy	Alloying elements*, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method* of casting***	Description, properties	Uses
C95300	Fe 0.8–1.5 Al 9.0–11.0 Cu 86.0		7.53	ABEFG	aluminum bronze 9B, 89-1-10: electrical conductivity of 13% IACS, typical tensile strength of 520 MPa (as cast) or 590 MPa (heat treated), machinability of 55%	pickling baskets, nuts, gears, steel mill slippers, marine equipment, non-sparking hardware, welding jaws
C95500 (CuAl10-Fe5Ni5)	Fe 3.0–5.0 Ni 3.0–5.5 Al 10.0–11.5 Cu 78.0	Mn 3.5	7.53	ABEFG	aluminum bronze 9D, 81-4-4-11: electrical conductivity of 8.5% IACS, typical tensile strength of 690 MPa (as cast) or 830 MPa (heat treated), machinability of 50%	valve guides and seats in aircraft engines, corrosion-resistant parts, bushings, gears, worms, pickling hooks and baskets, agitators
C95800 (CuAl10-Fe5Ni5)	Fe 3.5–4.5** Ni 4.0–5.0 Al 8.5–9.5 Mn 0.8–1.5 Cu 79.0	Pb 0.05 Si 0.10 Sn 0.10	7.64	ABEFG	alpha nickel aluminum bronze, 81-5-4-9-1, propeller bronze: electrical conductivity of 7.1% IACS, typical tensile strength of 655 MPa, machinability of 50%	propeller blades and hubs for service in fresh and salt water, fittings, gears, worm wheels, valve guides and seals

* Minimum, except for ranges. ** Ni content must exceed Fe content.
*** Method of casting codes are explained in Table 21.

Table 33. Copper–nickel–iron cast alloys (copper nickels)

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm^3	Method of casting*	Description, properties	Uses
C96400	Cu 65.0–69.0 Fe 0.25–1.5 Ni + Co 28.0–32.0 Nb 0.50–1.5	Pb 0.03 Mn 1.5 Si 0.50 C 0.15	8.94	ABG	70–30 copper nickel, electrical conductivity of 5% IACS, typical tensile strength of 470 MPa, machinability of 20%	valves, pump bodies, flanges, and elbows requiring resistance to seawater
C96600	Fe 0.8–1.1 Ni + Co 29.0–33.0 Be 0.40–0.7 Cu remainder	Pb 0.01 Mn 1.0 Si 0.15	8.80	ABDEG	beryllium cupronickel, electrical conductivity of 4.3% IACS, tensile strength of 830 MPa (heat treated), machinability of 40%	similar to C96400 except where greater strength is required, high-strength constructional parts for seawater, especially for long, unattended submergence, e.g., pump bodies, valve bodies, seawater line fittings
C96700	Be 1.1–1.2 Ni 29.0–33.0 Fe 0.7–1.0 Zr 0.10–0.20 Ti 0.10–0.20 Cu remainder	Mn 0.7 Si 0.15 Pb 0.1	8.60	ABDEG	modified beryllium cupronickel, electrical conductivity of 4.3% IACS, typical tensile strength of 560 MPa (as cast) or 860 MPa (heat treated), better castability than C96600, machinability of 40%	same as C96600, great use for plastic tooling

* Method of casting codes are explained in Table 21.

This alloy is somewhat stronger than C97300, with a typical tensile strength of 380 MPa.

Further detailed information is found in Table 34.

9.12. Special Cast Alloys

The alloys placed in the group C99000 –C99999 do not fit into any previous chemical classification. Each alloy has a special purpose; therefore, commercial usage is limited.

Alloy C99400 is noted for its resistance to dezincification and dealuminification. It is used in those environments where these types of corrosion are a problem or are expected to be a problem. It is somewhat heat treatable, the nominal as-cast tensile strength of 455 MPa being increased to 545 MPa. It has fair-to-good castability, but sometimes requires special foundry techniques.

Alloy C99700 is a low-cost white alloy used for ornamental and decorative purposes instead of C97800. The alloy has a good machinability rating (80 % FMB) and a strength comparable to C97800.

Further information is detailed in Table 35.

10. Master Alloys

Certain elements are to be added to the casting melt that are difficult to dissolve or that evolve as a fume, making exact additions difficult to attain. The use of master alloys aids in the introduction of these troublesome elements. A master alloy is rich in one or more addition elements in a matrix of the major element, in this case copper.

These master alloys may not be true alloys but may be simply mixtures of elements or their compounds. They are produced in various products or configurations:

1) *Pig.* This configuration usually is used where the element to be added is difficult to melt or dissolve. Adding iron to copper is an example. This master alloy is developed at much higher temperatures and times to reach solubilities that are not available in casting the basic alloy.
2) *Pellets.* Essentially the same product as pig, but cast in water to produce the pellets. The pellets dissolve more rapidly than the pig, thus giving more flexibility in the operation.

Table 34. Copper–nickel–zinc casting alloys (nickel silvers)

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting*	Description, properties	Uses
C97300	Cu 53.0–58.0 Sn 1.5–3.0 Pb 8.0–11.0 Zn 17.0–25.0 Ni + Co 11.0–14.0	Fe 1.5 Sb 0.35 S 0.08 P 0.05 Al 0.005 Mn 0.50 Si 0.15	8.95	ADEG	leaded nickel brass, 56-2-10-20-12: electrical conductivity of 5.7% IACS, typical tensile strength of 240 MPa, machinability of 70%	hardware fittings, valves and valve trim, statuary and ornamental castings
C97800	Cu 64.0–67.0 Sn 4.0–5.5 Pb 1.0–2.5 Zn 1.0–4.0 Ni + Co 24.0–27.0	Fe 1.5 Sb 0.20 S 0.08 P 0.05 Al 0.005 Mn 1.0 Si 0.15	8.86	ADEG	leaded nickel bronze, 66-5-2-2-25: electrical conductivity of 4.5% IACS, typical tensile strength of 380 MPa, machinability of 60%	ornamental castings, sanitary fittings, valve bodies, valve seats, musical instrument components

* Method of casting codes are explained in Table 21.

Table 35. Special casting alloys

UNS (ISO) alloy	Alloying elements, wt %	Permissible impurities (max.), wt %	Density, g/cm³	Method of casting**	Description, properties	Uses
C99400	Ni 1.0–3.5 Fe 1.0–3.0 Al 0.50–2.0 Si 0.50–2.0 Zn 0.50–5.0 Cu remainder	Pb 0.25 Mn 0.50	8.30	ABDG	nondezincification alloy, NDZ: electrical conductivity of 16.8% IACS, typical tensile strength of 455 MPa (as cast) or 540 MPa (heat treated), machinability of 50%	valve stems, propeller wheels, electrical parts, gears for mining, marine, or other equipment where resistance to dezincification or dealuminification is required
C99700	Cu 54.0* Ni 4.0–6.0 Al 0.50–3.0 Mn 11.0–15.0 Zn 19.0–25.0	Sn 1.0 Pb 2.0 Fe 1.0	8.19	ADEFG	white manganese brass, electrical conductivity of 3% IACS, typical tensile strength of 380 MPa, machinability of 80%	building hardware, architectural and ornamental fittings, marine hardware, floor drain covers, food handling equipment, swimming pool hardware

* Minimum.
** Method of casting codes are explained in Table 21.

3) *Chunks.* Similar to pig, chunks are not cast in a mold, but in thin friable slabs. The chunks may be added to the melt during charge preparation, added as a ladle addition, or at any convenient point. Adding ferromanganese or phosphorus copper master alloy are examples.
4) *Wire.* This form of a master alloy is quite special and is not an alloy at all. When producing this wire, a tube of the matrix is filled through special compaction of a powder containing a compound of the required element. The tube is then redrawn to a convenient diameter – e.g., 2 mm. The wire contains a known quantity of the element to be added. Adding zirconium to copper is an example using this technology. This method is used for elements that readily fume off the melt.
5) *Briquette.* This configuration readily permits the use of powders to produce the "master alloy." These powders may be pure elements or compounds of a desired element that are compacted. In many cases, the compounds being developed are not element additions, but flux added to purify the melt. This method is widespread in the casting industry, and a prime example is the casting of phosphor bronze alloys.

Generally the producers of master alloys are purchasers and sellers of scrap products. Many of the master alloys are developed at the request of individual customers, but most are made available on a commercial basis. The chemical content ranges are rather broad compared to those of wrought products.

11. Safety and EnvironmentalAspects

Several of the elements associated with copper alloys are toxic. Most are rendered harmless through the use of exhaust systems to remove gaseous effluents, with precipitators and filters to remove the solid particles. The exhaust system removes the fumes generated in melting and transports them rapidly to precipitators to remove the major solid material and then on to bag houses. This procedure removes and collects 99.9 % of the solids.

Usually where casting facilities are present and zinc is an alloying constituent, a thorough exhaust system is used to collect all fumes from the melting and casting area. In the past, workers became ill with "brass founders' ague," "brass chills," or "metal fume fever." Today these conditions do not normally exist.

These collection systems remove such potentially dangerous elements as cadmium, antimony, arsenic, lead, selenium, manganese, tellurium, titanium, and zinc. All of these elemental fumes are rapidly oxidized into removable solids.

Gaseous fumes developed during cleaning or pickling must be removed form the work area. Normally these may be exhausted to the atmosphere, where they are dissipated and rendered harmless. In areas where the general environment cannot tolerate the additional contamination, the fumes may pass through a wet (water spray) collector where a majority of the fumes is removed.

One effluent not always recognized is the metal powder or "dust" often generated in operations. Where any powder is generated, it itself represents an explosive hazard and should be removed from the area. Such products should be vacuumed away to a safe collecting area.

One element not mentioned above is *beryllium.* Beryllium can be highly toxic, not only in the producer's plant, but also in the ensuing user operations.

Grinding and polishing generate dust, but many users have not realized that stamping and forming operations may also form dust particles. Welding – spot, seam, or other types – will produce a fume. Even some soldering operations may produce a fume when soldering beryllium-containing alloys.

Any alloy containing beryllium requires special precautions and ventilation where parts are made by any means that could produce a dust or fume. Most manufacturers of products containing beryllium supply a precautionary note advising that inhalation of excessive concentrations of airborne beryllium can cause pulmonary or respiratory problems.

12. Economic Aspects

New alloys are being developed continuously to fill voids where existing alloys are found wanting. Even though most new alloys do not perform to expectations, the mass of trials does produce breakthroughs that extend the usage of copper alloys. The percentage of copper in alloys is dropping, on the average, but this is offset by expansion of the total copper market and counteracts some of the effect of substitution.

No longer do international borders form barriers to markets. Strip products now have better finishes, closer physical and mechanical tolerances, and better flatness than U.S. producers would have produced without worldwide competition. A new concept of quality control – quality assurance – has emerged to improve product quality through closer in-process inspection and broader use of computerization. The just-in-time delivery of high-quality products to the fabricator smooths out production schedules.

A pervasive change in the copper alloy market has been the continual miniaturization of the parts. One example is the automobile radiator. Brass metal for radiator tube has been reduced in thickness from 0.114 to 0.0064 mm, total usage dropping even more rapidly because radiators are smaller and have less surface area. The tanks and headers, also correspondingly smaller, have undergone a thickness reduction from 0.813 to 0.381 mm. (Also see → Copper, Chap. 7.3.)

One interesting outgrowth of miniaturization is a relatively new concept – shaped wire. The wire is made in cross sections neither round or square. Customer – fabricator interaction leads to cross sections that can be fed into machines that fabricate the contact or connector. Generally, production runs are longer, less scrap is generated, and when necessary the parts can be assembled on a bandolier, which aids in plating and assembly. The final product is lighter, but the consumption of copper alloy is actually increased, because of the expanded market.

13. References

General References

1. *Metals Handbook,* 1948 ed., American Society for Metals, Cleveland, Ohio, 1954, pp. 841 – 942.
2. *Metals Handbook, Properties and Selection of Metals,* 8th ed., **vol. 1,** American Society for Metals, Metals Park 1961, pp. 961 – 1052.
3. *Metals Handbook, Properties and Selection: Non-Ferrous Alloys and Pure Metals,* 9th ed., **vol. 2,** American Society for Metals, Metals Park 1979, pp. 237 –490.
4. *Metals Handbook, Mechanical Testing,* 9th ed., vol. 8, American Society for Metals, Metals Park 1985.
5. *Metals Handbook,* Desk ed., American Society for Metals, Metals Park 1985.
6. J. V. Arcuri, D. L. Potts: *International Metallic Materials Cross Reference,* General Electric Co., Schenectady, N.Y., 1983.
7. S. W. Blakely, J. P. Fleischman, C. A. Isbell, H. F. Nullmeyer: *High Performance Copper Alloy Strip and Wire for Electronic Connectors,* **Part II, vols. 3 and 4,** the UNIMAR Group, Alton 1985.
8. A. Butts: *Copper,* Hafner, New York 1970.
9. C. A. Keyser: *Basic Engineering Metallurgy,* Prentice-Hall, New York 1952.
10. D. McLean: *Mechanical Properties of Metals,* J. Wiley & Sons, New York 1967.
11. T. H. Rogers: *The Marine Corrosion Handbook,* McGraw-Hill, New York 1960.
12. O. O. Miller: *Alloy Digest,* Engineering Alloys Digest, Upper Montclair 1983.
13. C. H. Samens: *Engineering Metals and Their Alloys,* Macmillan, New York 1949.
14. R. A. Wilkins, E. S. Bunn: *Copper and Copper Base Alloys,* McGraw-Hill, New York 1943.
15. Application Data Sheet, Copper Development Assoc., Greenwich, Conn., 1985.
16. *Standards Handbook – Alloy Data/2, Wrought Mill Products,* Copper Development Assoc., Greenwich, Conn., 1973.
17. *Standards Handbook – Alloy Data/7, Cast Products,* Copper Development Assoc., Greenwich, Conn., 1978.
18. "Copper Data," Copper Development Assoc. Publication 12, Greenwich, Conn., 1960.
19. "Bearing Bronzes," Copper Development Assoc. Publication 15, Greenwich, Conn., 1959.
20. "Aluminum Bronze," Copper Development Assoc. Publication 31, Greenwich, Conn., 1960.
21. "Copper and Copper Alloys: Compositions and Mechanical Properties," Copper Development Assoc. Publication 36, Greenwich, Conn., 1958.
22. "Copper and Copper Alloy Springs," Copper Development Assoc. Publication 39, Greenwich, Conn., 1959.
23. "Copper and Its Alloys in Engineering and Technology," Copper Development Assoc. Publication 43, Greenwich, Conn., 1960.
24. "Equilibrium Diagrams of Binary Copper Alloys," Copper Development Assoc. Publication 44, Greenwich, Conn., 1960.
25. "Copper and Copper Alloys, Selected Microstructures and Equilibrium Diagrams," Copper Development Assoc. Publication 64, Greenwich, Conn., 1964.

Specific References

26. Unpublished data, Olin Brass Corp., East Alton, Ill.
27. W. L. Finlay: *Silver-Bearing Copper,* Corinthian, New York 1968, p. 27.
28. J. H. Mendenhall: *Understanding Copper Alloys,* Olin Brass Corp., East Alton 1977.
29. J. C. Chaston: *Metals Handbook,* 8th ed., **vol. 8,** American Society for Metals, Metals Park 1973, p. 253.
30. C. R. Smith in A. Butts: *Copper,* Hafner, New York 1970. Specific References

Copper Compounds

H. Wayne Richardson, CP Chemicals Inc., Sumter, South Carolina 29150, United States

1. Introduction 1
2. The Copper Ions 3
3. Basic Copper Compounds 4
3.1. Copper(I) Oxide 4
3.2. Copper(II) Oxide 5
3.3. Copper(II) Hydroxide 7
3.4. Copper(II) Carbonate Hydroxide . 8
4. Salts and Basic Salts 9
4.1. Copper(I) Chloride 9
4.2. Copper(II) Chloride 11
4.3. Copper(II) Oxychloride 12
4.4. Copper(II) Sulfates 13
4.4.1. Copper(II) Sulfate Pentahydrate 13
4.4.2. Anhydrous Copper Sulfate 17
4.4.3. Copper(II) Sulfate Monohydrate . . . 17
4.4.4. Basic Copper(II) Sulfates 17
5. Compounds and Complexes of Minor Importance 19
5.1. Copper Compounds 19
5.2. Copper Complexes 22
6. Copper Reclamation 23
7. Copper and the Environment 24
8. Economic Aspects 25
9. Toxicology and Occupational Health 26
10. References 27

1. Introduction

The importance of copper and its alloys to today's society is well documented in the articles on → Copper and on → Copper Alloys. Copper compounds, although they represent a small fraction of total copper production, play an important and varied role in industry and agriculture. One of the oldest known fungicides was copper-based and was used extensively in the early part of the century. The last 20 years have brought about a resurgence in the use of copper-based fungicides partly because of a lack of tolerance by fungi to copper and also because of its relatively low toxicity to higher plants and animals (see, however, Chaps. 7 and 9). Although copper is an essential trace element for higher plants and animals, it is acutely toxic in higher doses. Copper compounds are used as nutritional additives in animal feeds and fertilizers, and are found in a variety of dietary supplements. Copper salts are used in the control of algae in lakes and ponds, and the oxides are used in antifouling paints and coatings. Copper acetoarsenite, Paris green, is used as an insecticide, and copper chromium arsenate is an effective alternative to creosote for the preservation of wood.

Copper and its compounds are used catalytically in numerous organic reactions, e.g., polymerization, isomerization, and cracking reactions. They are used in the textile and dye industries in the preparation of rayon and acrylonitrile, as mordants and oxidants in textile dyeing and printing, and in the preparation of azo dyes. Copper compounds are used as pigments in glass, ceramics, porcelains, varnishes, and artificial gems, and in the manufacture of the copper phthalocyanine pigments. Copper salt solutions are used for electroplating, in brazing and burnishing preparations, and as brighteners for aluminum. Solutions of copper(I) complexes are used to selectively absorb carbon monoxide, butadiene, and alkenes from gas streams. In the petroleum industry, copper compounds are used as deodorizing (desulfurization) and purifying agents. Copper(II) carbonate is used in drilling muds to protect against release of poisonous hydrogen sulfide gas, and copper(I) iodide is used in acid muds to bind corrosion inhibitors to the iron drills. More recent applications of copper compounds are in pollution control and solar technology.

The multitude of applications of copper compounds in the biosphere is largely responsible for the extent of academic interest in them. Also, the facile reduction – oxidation of the copper(I) – copper(II) couple, the ease of theoretical treatment of the d^9 copper(II) system, and the varied

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_567

stereochemistries and magnetic behaviors associated with copper ions enhance their theoretical appeal. Since this treatment of copper compounds is primarily from an industrial perspective, the reader whose interest is academic is referred to the classic references [1–4] as well as to more recent materials [5–11] that are attune to the subtle chemistries of copper and its compounds. Compounds of primary industrial importance (Chaps. 3 and 4) are distinguished from compounds of minor importance (Chap. 5) in this article.

2. The Copper Ions

Copper, Cu [*7440-50-8*], M_r 63.546, $[Ar]3d^{10}4s^1$, is a member of the first transition series. It is classified as a transition element in the broader definition because the copper(II) valence state, $[Ar]3d^9$, comprises such a large proportion of the defined chemistry of copper. Copper in its compounds exists primarily in two oxidation states, + 1 and + 2. Although copper(0) and copper(III) compounds have been identified, they are not presently of commercial importance. The stabilities of the various valence states of copper are illustrated by the following standard reduction potentials:

	E_0
$Cu^+ + e \longrightarrow Cu^0$	+ 0.52 V
$Cu^{2+} + 2e \longrightarrow Cu^0$	+ 0.34 V
$Cu^{2+} + e \longrightarrow Cu^+$	+ 0.15 V
$Cu^{3+} + e \longrightarrow Cu^{2+}$	+ 1.80 V

From the above, it is seen that

$$2\,Cu^+ \longrightarrow Cu^0 + Cu^{2+} \quad \Delta E_0 = -0.37\,V$$

with $pK = -5.95$

In other words, the free copper(I) ion does not exist to any appreciable extent in aqueous solution. In the presence of ligands such as ammonia, chloride, or cyanide, solutions of copper(I) can be prepared that are stable with respect to disproportionation. For example, the colorless solution of tetraamminedicopper(I) sulfate is prepared readily by contact of blue tetraamminecopper(II) sulfate with metallic copper in the absence of air. Upon acidification with sulfuric acid, copper powder and a copper(II) ammonium sulfate solution are produced. The insoluble copper(I) chloride can be produced by sulfuric acid acidification of a copper(I) ammine chloride solution. If the solution is acidified with hydrochloric acid, a solution of $[CuCl_2]^-$, $[CuCl_3]^{2-}$, or $[CuCl_4]^{3-}$ species is produced, depending on chloride concentration. Copper(I) chloride is stable to water because of its insolubility, which is a result of the polymeric structure that arises from the chloride's ability to bridge copper. This contrasts with the sulfate's inability to coordinate or bridge strongly. Consequently, copper(I) sulfate can be produced only in nonaqueous media.

The electronic structure of the copper(I) ion is $[Ar]3d^{10}$. The compounds are diamagnetic and colorless except where charge-transfer bands arise. Copper(I) is isoelectronic with zinc(II) and the preferred stereochemistries are similar. As a result of the filled 3*d* level, no ligand field stabilization occurs and electronic distortions are minimized. The stereochemistry around the copper(I) ion is determined mainly by the size of the anions, as well as by the electrostatic and covalent bonding forces. The preferred stereochemistry is tetrahedral, with linear and trigonal planar compounds also being common.

The majority of copper(II) compounds exhibit square planar or distorted octahedral configurations about the copper ion. The $3d^9$ electronic structure gives rise to the classic example of Jahn – Teller distortion in which the four planar metal – ligand distances are smaller than the two axial distances. Copper(II) ions are also found in distorted tetrahedral and various five-coordinate environments.

The copper(II) ion, $[Ar]3d^9$, is predominantly blue or green, and the unpaired 3*d* electron results in magnetic phenomena. In most copper compounds, the unpaired electrons of the copper ions are sufficiently isolated from each other so that the compounds exhibit paramagnetic behavior. However, there are many polynuclear copper compounds in which the spins are coupled, which lowers the magnetic moment. The coupling may be so weak that it must be observed near the absolute zero of temperature, or it may be strong enough to render the compound diamagnetic at room temperature or above.

3. Basic Copper Compounds

3.1. Copper(I) Oxide

Cu_2O [*1317-39-1*], M_r 143.09, *mp* 1235 °C, d_4^{25} 5.8 – 6.2, decomposes above 1800 °C. It occurs in nature as the red or reddish brown mineral cuprite with a cubic or octahedral crystal morphology. Depending on the method of preparation and particle size, the synthetic material is yellow, orange, red, or purple. The yellow material has erroneously been referred to as copper(I) hydroxide, but X-ray diffraction patterns indicate that there are no differences in the crystal structures of the colored forms. Their thermodynamic data are as follows: c_p (298 K) 429.8 J kg^{-1} K^{-1}, c_p (290 – 814 K) 519.2 J kg^{-1} · K^{-1}, c_p (290 – 1223 K) 565.2 J kg^{-1} K^{-1}, $\Delta H°$ (25 °C) – 166.6 kJ/mol. Copper(I) oxide is stable in dry air but slowly oxidizes to copper(II) oxide [*1317-38-0*] in moist air. It is practically insoluble in water but dissolves in aqueous ammonia. In excess hydrochloric acid, soluble copper(I) chloride complexes are formed; however, in dilute sulfuric or nitric acids, disproportionation to the soluble copper(II) salts and copper powder results.

Production. Copper(I) oxide is produced easily by a variety of methods; its instability with respect to oxidation requires careful consideration. Copper(I) oxide produced pyrometallurgically is usually coated with isophthalic acid or pine oil to preserve its integrity [12]. Hydrometallurgically produced material can be stabilized by mixing the particle slurry with glue, gelatin, casein, or dextrin before drying [13–16].

Pyrometallurgical Processes. Copper(I) oxide is formed when copper powder is heated above 1030 °C in air; to prevent further oxidation, it must be cooled quickly in an inert atmosphere. To allow for lower temperature production of copper(I) oxide, carbon can be blended with copper(II) oxide and heated to 750 °C in an inert atmosphere. The material must be stabilized by coating the formed particles with isophthalic acid or pine oil [12]. A more stable copper(I) oxide results when stoichiometric amounts of copper powder and copper(II) oxide are blended, heated to 800 – 900 °C in an inert atmosphere, and allowed to cool. The production can be effected at lower temperature if ammonia or certain ammonium salts are added to the blend [17–19]. The autoclave oxidation of copper metal at 120 °C and about 0.6 MPa gauge pressure with air in the presence of water and small amounts of sulfuric and hydrochloric acids produces a red, pigment-grade product [20]. By varying the pressure and temperature, considerable differences in particle size, coloring, apparent bulk density, and buoyancy have been found.

Hydrometallurgical Processes. Tetraamminedicopper(I) sulfate, $Cu_2(NH_3)_4SO_4$, prepared by leaching an excess of copper with a solution of ammonia and ammonium sulfate, with air as the oxidant, yields a red copper(I) oxide upon acidification to pH 3 – 5 [21]. The less corrosive ammonium carbonate leach system in which $Cu_2(NH_3)_4CO_3$ is produced is more common (see page 5). Upon vacuum distillation, a very stable red Cu_2O product remains [21]. If sodium hydroxide is added to the leach liquor, a yellow microcrystalline powder is precipitated [22]. When the yellow Cu_2O is heated in an excess of sodium hydroxide, it is converted to an orange material of somewhat larger particle size.

Steam stripping of the copper(I) ammine carbonate solution yields a brown, impure product [23] which can be converted to a red material by washing it in an organic acid, e.g., formic or acetic acid [24]. An impure, brown product can also be converted to a red material by boiling it in 20 % sodium hydroxide solution [25].

If a saturated solution of copper(I) ammine carbonate is agitated over copper metal, a layer of red copper(I) oxide is continuously produced which can be broken loose and recovered [26]. When copper salts are leached with chelating agents such as ethylenediaminetetraacetic acid [27] or ammonia [28] under pressure of carbon monoxide or hydrogen, and sodium hydroxide is subsequently added, a relatively stable, yellow copper(I) oxide is obtained; the reaction is catalyzed by an alkali metal iodide [29].

The reduction of a boiling slurry of basic copper(II) sulfate with sulfur dioxide at a pH of about 3 produces a reddish product [30]. Red copper(I) oxide has also been prepared by mixing a slurry of basic copper(II) sulfate with neutral copper(II) sulfate and adding sodium sulfite

to a pH of 5.2. The mixture is then acidified to pH 3.5 – 5 and heated to boiling. The intermediate copper(I) sulfite slurry is decomposed to copper(I) oxide and sulfurous acid. Alkali is subsequently added to maintain a pH of 2.6 – 2.8 [31].

When a solution of copper(I) chloride and sodium chloride is neutralized with sodium hydroxide and then heated to 138 °C under pressure, a red copper(I) oxide is obtained which has an average particle diameter of about 2.5 µm [32]; an orange product (about 1-µm particles) is prepared by neutralizing the solution to pH 8.5 at 60 °C [33]. Simultaneous mixing of copper(I) chloride solutions with sodium chloride and sodium hydroxide solution in the presence of copper(I) oxide seed crystals at a controlled pH of 10.0, 55 °C, and under nitrogen, gives a reddish purple material (average diameter 48 µm). At pH 7.0, a yellow material is obtained with an average particle size of 0.4 µm [34].

The *electrolytic production* of copper(I) oxide between copper electrodes in brine yields a yellow product at room temperature. At higher temperature, an orange or red material is produced.

Uses. The largest commercial use of copper(I) oxide is in antifouling paints for boat and ship bottoms; it is an effective control for barnacles and algae. The yellow or orange copper(I) oxide is used as a seed and crop fungicide, and the red material is used as a pigment in ceramic glazes and glass. Copper(I) oxide is also used in rectifiers and in brazing. Numerous organic reactions are catalyzed by copper(I) oxide, and it is an effective absorbent for carbon monoxide.

Analysis and Specifications. The ASTM approved analysis and specification for pigment-grade copper(I) oxide [35] and the military specification for the pigment grade [36] are listed in Table 1.

Table 1. Specifications for pigment-grade copper(I) oxide

Assay	Mass fraction, %	
	Navy I [36], [37]	Navy II [36]
Copper(I) oxide	97.0	90.0
Total copper (min.)	86.0	80.0
Reducing power (min.)	97.0	90.0
Nitric acid-insolubles (max.)	0.3	0.3
Chloride (max.)	0.4	0.4
Sulfate (max.)	0.1	0.1
Zinc oxide (max.)	–	10.0
Other metals (max.)	0.5	0.5
Acetone-soluble material (max.)	0.5	0.5

3.2. Copper(II) Oxide

CuO [*1317-38-0*], M_r 79.54, *mp* 1330 °C, d_4^{25} 6.48, occurs in nature as the black minerals tenorite (triclinic crystals) and paramelaconite (tetrahedral, cubic crystals). Commercially produced copper(II) oxide is usually black, although a brown product (particle size $< 10^{-6}$ m) can also be produced. Thermodynamic data: c_p (298 K) 531.1 J kg^{-1} K^{-1}, c_p (290 – 1253 K) 682.4 J kg^{-1} · K^{-1}, $\Delta H°$ (25 °C) −155.3 kJ/mol. Copper(II) oxide is stable to air and moisture at room temperature. It is virtually insoluble in water or alcohols. Copper(II) oxide dissolves slowly in ammonia solution but quickly in ammonium carbonate solution; it is dissolved by alkali metal cyanides and by strong acid solutions. Hot formic acid and boiling acetic acid solutions readily dissolve the oxide. Copper(II) oxide is decomposed to copper(I) oxide and oxygen at 1030 °C and atmospheric pressure; the reduction can proceed at lower temperature in a vacuum. Hydrogen and carbon monoxide reduce copper(II) oxide to the metal at 250 °C and to copper(I) oxide at about 150 °C. Ammonia gas reduces copper(II) oxide to copper metal and copper(I) oxide at 425 – 700 °C [17].

Production. Copper(II) oxide can be prepared *pyrometallurgically* by heating copper metal above 300 °C in air; preferably, 800 °C is employed. Molten copper is oxidized to copper(II) oxide when sprayed into an oxygen-containing gas [37]. Ignition of copper(II) nitrate trihydrate [*10031-43-3*] at about 100 – 200 °C produces a black oxide. Basic copper(II) carbonate [*12069-69-1*], when heated above 250 °C, produces a black oxide if a dense carbonate is employed; a brown material is produced when the light and fluffy carbonate is used. An alkali-free oxide can be prepared by ignition of copper(II) carbonate produced from ammonium carbonate and a copper(II) salt solution. Copper(II) hydroxide [*20427-59-2*], when heated above 100 °C, is converted to the oxide.

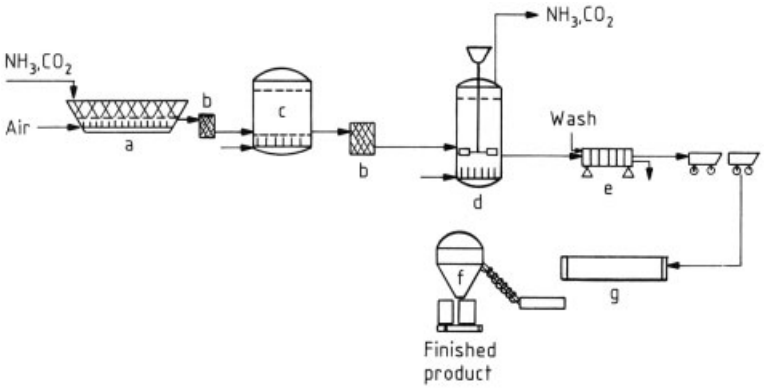

Figure 1. Process flow diagram for production of copper(II) oxide from ammonia – ammonium carbonate leach
a) Leach vat; b) Filter; c) Treatment tank; d) Strip tank; e) Press; f) Bag house; g) Drying kiln

Hydrometallurgy is the most common method for the production of copper(II) oxide. A solution of ammonia and ammonium carbonate in the presence of air effectively leaches metallic copper; the process is represented by the following reactions:

$$2\,Cu + 1/2\,O_2 + 2\,NH_3 + (NH_4)_2CO_3 \longrightarrow [Cu_2(NH_3)_4]CO_3 + H_2O$$

$$2\,NH_3 + (NH_4)_2CO_3 + [Cu_2(NH_3)_4]CO_3 + 1/2\,O_2 \longrightarrow 2\,[Cu(NH_3)_4]CO_3 + H_2O$$

$$Cu + [Cu(NH_3)_4]CO_3 \longrightarrow Cu_2(NH_3)_4CO_3$$

The second and third reactions proceed readily; the first is slow. Consequently, in batch operations the leach is usually begun with a small charge of the copper solution, but continuous operations offer significantly improved rates. The leach liquor is then filtered to remove iron impurities and metallic copper, and is subsequently oxidized by air sparging. If necessary, lead and tin are removed by treatment with strontium, barium, or calcium salts [38–41]. The solution is filtered again and stripped of ammonia and carbon dioxide by steam injection or pressurized boiling to produce a black copper(II) oxide [38]. The ammonia and carbon dioxide are recycled for further use. The process is illustrated in Figure 1. Alternatively, the leach liquor can be treated with strong alkali to precipitate the intermediate copper(II) hydroxide, and then boiled to remove ammonia, with subsequent decomposition of the hydroxide to the black oxide.

The copper(II) ammine sulfate and chloride systems are treated similarly although they are utilized much less frequently because of their highly corrosive nature. However, the copper ammine chloride system, a byproduct of circuit board etching, is recycled out of economic and environmental necessity to copper(II) oxide [42] (see Chap. 6).

Uses. Copper(II) oxide is used as a precursor to a number of copper(II) salts. One of the largest commercial applications is in the production of compounds for wood preservation. Copper(II) oxide is also used extensively as a feed additive and as a pigment in glass, ceramic, and porcelain enamels [43]. In combination with manganese dioxide, it is used as an oxidative catalyst for exhaust gas [44], in the removal of NO_x, CO, and O_3 [45], [46], and in the purification of formaldehyde-containing waste gas [47]. Supported on aluminum phosphate, copper(II) oxide is active in reducing tar and polycyclic hydrocarbons in smoke by absorption and catalytic conversion [48]. Copper(II) oxide is used as a catalyst in the preparation of acrylates [49] and in the production of magnetic storage devices [50]. It has limited application in the petroleum industry as a gas sweetener and is used in welding fluxes for bronze.

Analysis and Specifications. Electrolytic deposition from a solution or iodometric analysis with sodium thiosulfate is standard for copper determination [51]. The technical specification is shown in Table 2.

Table 2. Typical analysis of commercial copper(II) oxide

Assay	Mass fraction, %
Copper	78.5
Iron	0.09
Lead	0.08
Water	0.10
Nitric acid-insolubles	0.10
Zinc	0.05
Surface area, m^2/g	10.0

3.3. Copper(II) Hydroxide

$Cu(OH)_2$ [*20427-59-2*], M_r 97.54, d_4^{25} 3.37, $\Delta H°$ (25 °C) 446.7 kJ/mol, decomposes over 100 °C or over 50 °C in the presence of an excess of alkali. Copper(II) hydroxide is virtually insoluble in water (0.003 mg/L), and decomposes in hot water to the more stable copper(II) oxide and water:

$$\underset{\text{blue}}{Cu(OH)_2} \longrightarrow \underset{\text{green to brown}}{CuO \cdot H_2O} \longrightarrow \underset{\text{brown to black}}{CuO} + H_2O$$

Copper(II) hydroxide is readily soluble in mineral acids and ammonia solution. When freshly precipitated, it is soluble in concentrated alkali, with the formation of $[Cu(OH)_3]^-$ or $[Cu(OH)_4]^{2-}$. Copper(II) hydroxide is inherently unstable but can be kinetically stabilized by a suitable production method.

Production. There are two classes of copper(II) hydroxide. The first is stoichiometrically rather precise, with a copper content as high as 64 %; the theoretical copper content of $Cu(OH)_2$ is 65.14 %. This class is produced by the ammonia process [52–55], which yields a pure product of relatively good stability and large particle size. The best product results from the addition of strong alkali to the soluble copper(II) ammine complex [52], [55]. A relatively large particle-size product, deep blue in color and high in copper content, is precipitated below 35 °C. The resulting material is fairly stable or can be coated with gelatin to enhance its stability [56].

In the copper(II) hydroxide made by the ammonia process, the solubility of the copper(II) ammine complexes provides for crystallite growth. This affords a large particle size, a limited surface area (point of dehydration), and hence a relatively stable product, in contrast with the unstable product (variable assay) that results from the addition of hydroxide solutions to copper(II) salt solutions at 20 °C or above. The reaction with hydroxide is diffusion-controlled, allowing essentially no time for crystallite growth. The product is obtained as a gelatinous, voluminous precipitate with a large surface area, which is quite unstable and difficult to wash free of impurities. If the same reaction, with the same order of addition, is allowed to occur at 0 – 10 °C, a product of defined particle size and measurable surface area results, with greater stability but low assay.

The second class of copper(II) hydroxide, which represents a "stable" product but has lower assay and greater impurity, is produced from an insoluble precursor such as basic copper(II) carbonate or copper(II) phosphate. The first stable copper(II) hydroxide of this kind was made from copper(II) phosphate with alternate additions of copper(II) sulfate and sodium hydroxide solutions [57]. The process is illustrated by the following series of reactions:

$$3\,CuSO_4 + 2\,Na_3PO_4 \longrightarrow Cu_3(PO_4)_2 + 3\,Na_2SO_4$$
$$Cu_3(PO_4)_2 + 6\,NaOH \longrightarrow 3\,Cu(OH)_2 + 2\,Na_3PO_4$$

The alternate copper(II) sulfate – sodium hydroxide addition is continued through 15 or 20 cycles and yields a stable product with 58 – 59 % copper and 3 – 5 % phosphate. The product has a small particle size and a high surface area, and is used as an agricultural fungicide.

Other stable copper(II) hydroxides of high surface area and fine particle size have been produced more recently [58–60]; the processes include the use of a copper(II) oxychloride precursor in the presence of an anionic surfactant [59] or a magnesium sulfate-precipitated precursor [58]. An electrolytically produced material has also been made by using trisodium phosphate as the electrolyte [61].

The classic Bordeaux slurry of copper(II) sulfate and lime in water has been replaced by a powdered stabilized product. This is obtained by mixing copper(II) nitrate solution and lime, adding cellulose pitch liquor (a waste product of the paper industry), and drying to yield a powder which is effective as a fungicide [62]. Sodium carbonate can be used as the precipitating agent instead of lime [63].

Uses. Copper(II) hydroxide is used as an active precursor in the production of copper(II) compounds. Ammonia-processed copper(II) hydroxide is used in the production of copper(II) naphthenate, copper(II) 2-ethylhexanoate, and copper soaps. Ammonia-processed copper(II) hydroxide is also used in the production of rayon (Schweitzer's reagent) and in the stabilization of nylon. Copper(II) hydroxides of the second class are often used as fungicides because of their small particle size. Copper(II) hydroxide is also used as a feed additive, a catalyst in the vulcanization of polysulfide rubber, and an antifouling pigment.

Analysis and Specifications. Copper is determined iodometrically with sodium thiosulfate [51]. The analysis of typical ammonia-processed copper(II) hydroxide (mass fractions in %) is copper 63.0, iron 0.05, zinc 0.05, lead 0.05, naphthenic acid-insoluble material 2.0, sulfate 0.3.

3.4. Copper(II) Carbonate Hydroxide

Copper(II) carbonate hydroxide, also called basic copper(II) carbonate, occurs in nature as the metastable mineral *azurite* [*12070-39-2*], also called chessylite, a blue, monoclinic crystalline or amorphous powder with a formula approximating $2\,CuCO_3 \cdot Cu(OH)_2$, M_r 344.67, d_4^{24} 3.8, $\Delta H°$ (20 °C) − 87.4 kJ/mol, and *malachite* [*12069-69-1*], green, monoclinic crystals with a formula approximating $CuCO_3 \cdot Cu(OH)_2$, M_r 221.12, d_4^{25} 3.9 – 4.0, $\Delta H°$ (20 °C) − 57.7 kJ/mol. The copper(II) carbonate of commerce, malachite, is also known as Bremen green. Pure copper(II) carbonate, $CuCO_3$, has not been isolated. Copper(II) carbonate is virtually insoluble in water but dissolves readily in aqueous ammonia and alkali metal cyanide solutions. Copper(II) carbonate dissolves quickly in mineral acid solutions and warm acetic acid solution, with the formation of the corresponding copper(II) salt. Malachite is much more stable than copper(II) hydroxide but slowly decomposes to the oxide according to the following reaction:

$$CuCO_3 \cdot Cu(OH)_2 \longrightarrow 2\,CuO + H_2O + CO_2$$

Malachite is rapidly decomposed to the oxide above 200 °C.

Production. Two grades of copper(II) carbonate are available commercially, the light and the dense. The light grade is a fluffy product of high surface area. It is precipitated by adding a copper(II) salt solution, usually copper(II) sulfate solution, to a concentrated solution of sodium carbonate at 45 – 65 °C. Azurite is formed initially, and complete conversion to malachite usually occurs within two hours. The conversion is accelerated by the addition of malachite nuclei to the reactor.

A dark green, dense product results when a copper(II) salt solution is added to a solution of sodium hydrogen carbonate at 45 – 65 °C; conversion to malachite requires about one hour in this case. The density is maximized if the reactor is washed with acid prior to the precipitation to prevent premature nucleation on malachite nuclei. (A less dense product would be produced if malachite nuclei are added to the slurry of azurite.) Solutions of copper(II) salt and sodium carbonate can also be added simultaneously at a pH of 6.5 – 7.0 and a temperature between 45 and 65 °C; conversion to malachite is usually complete within one hour.

When a solution of copper(II) ammonium carbonate is boiled, ammonia and carbon dioxide are expelled from the solution, and a deep green, dense copper(II) carbonate precipitates [38].

Uses. Copper(II) carbonate is used as a precursor in the production of copper salts and soaps. It is used in animal feeds as a source of copper, in the sweetening of petroleum, and in electroplating for the control of pH. Copper(II) carbonate is used as a hydrogenation catalyst and as an accelerator in polymerization reactions. The light grade is somewhat effective as a fungicide and is used as a seed protectant.

Analysis and Specifications. Copper is analyzed by iodometric titration with sodium thiosulfate solution [51]. Table 3 gives typical analyses of light and dense technical-grade copper(II) carbonates.

Table 3. Typical analysis of commercially available copper(II) carbonate hydroxides

Assay	Mass fraction, %	
	Light	Dense
Copper	55.5	55.0
Sulfate	0.6	0.6
Iron	0.1	0.1
Zinc	0.01	0.02
Lead	0.003	0.005
Hydrochloric acid-insoluble material	0.05	0.05
Water	1.0	2.0

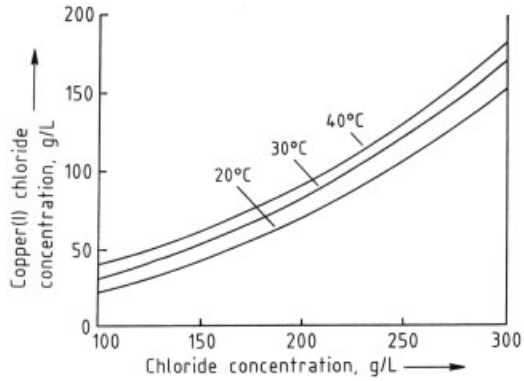

Figure 2. Solubility of copper(I) chloride in excess chloride ion solution at different temperatures

4. Salts and Basic Salts

4.1. Copper(I) Chloride

Copper(I) chloride [*7758-89-6*], CuCl, M_r 99.00, *mp* 422 °C, *bp* 1367 °C, d_4^{25} 4.14, $\Delta H°$ (25 °C) − 134.6 kJ/mol, occurs in nature as the colorless or gray cubic-crystal nantokite. The commercially available product is white to gray to green and of variable purity. Copper(I) chloride is fairly stable in air or light if the relative humidity is less than about 50 %. In the presence of moisture and air, the product is oxidized and hydrolyzed to a green product that approaches copper(II) oxychloride, $CuCl_2 \cdot 3\,Cu(OH)_2$[*12356-86-4*]. In the presence of light and moisture, a brown or blue product is obtained. Copper(I) chloride is slightly soluble to insoluble in water, with values from 0.001 to 0.1 g/L being reported. It is readily hydrolyzed to copper(I) oxide by hot water. Copper(I) chloride is insoluble in dilute sulfuric and nitric acids, ketones, alcohols, and ethers, but it quickly dissolves in hydrochloric acid, alkali halide, or ammonia solutions with the formation of complex compounds that are readily oxidized by air. Copper(I) chloride is soluble in solutions of alkali metal cyanides or thiosulfates and of coordinating amines, pyridines, and nitriles, notably, acetonitrile [64]. The increase in solubility of copper(I) chloride with chloride ion concentration illustrated in Figure 2 [65]. When the chloride concentration is decreased by dilution with water, the pure white copper(I) chloride precipitates.

Production. The direct combination of the elements is the most common method of production. The reaction of copper metal and chlorine is not spontaneous at ambient temperature. Once the metal is heated to red heat in the presence of chlorine, the reaction is self-sustaining and requires external cooling to prevent the metal from melting. A number of similar patents exist [66–69] in which copper metal is reacted with chlorine gas to produce a molten copper(I) chloride that is cast and pulverized. The primary difference in conditions is the use of a shaft furnace [66] as opposed to the use of crucibles. The recommended process temperature varies from 450 to 800 °C [66], [67] or 500 to 700 °C [68], [69]. The conditions required for the high-temperature production of a pure copper(I) chloride can be illustrated by the following:

$$Cu + 1/2\,Cl_2 \longrightarrow CuCl \qquad \Delta H = -134.6\,kJ/mol$$
$$Cu + Cl_2 \longrightarrow CuCl_2 \qquad \Delta H = -247.2\,kJ/mol$$
$$CuCl + 1/2\,Cl_2 \longrightarrow CuCl_2 \qquad \Delta H = -122.6\,kJ/mol$$
$$CuCl_2 + Cu \longrightarrow 2\,CuCl \qquad \Delta H = -22\,kJ/mol$$

Higher temperature and excessive contact with copper metal favor the production of a very pure copper(I) chloride. The lowest possible temperature is obviously 422 °C, the melting point of copper(I) chloride. As the temperature approaches the decomposition temperature for the reaction

$$CuCl_2 \longrightarrow CuCl + 1/2\,Cl_2 \ (993\,°C)$$

a product of higher purity is obtained. Operationally, a temperature between 750 and 900 °C is ideal, and results in a product of > 98 % purity. The process of Degussa [67] illustrates a commonly used commercial process for the high-temperature production of copper(I) chloride; Figure 3 shows a suitable crucible for the production of technical-grade copper(I) chloride. Once the reaction is initiated, chlorine and copper metal (shot, chopped wire, or briquettes) are added continuously. As the molten product is formed on the surface of the upper metal layer, it flows by gravity down through the porous copper bed to effect further reduction of any copper(II) chloride, and out the exit port onto a rotating table where the product is allowed to cool and solidify. The flakes that form are packaged as is or ground to a powder and packaged. Because of the high temperature during the reaction and the volatility of copper(I) chloride, the exit port must be vented to a caustic scrubber. When the molten product is allowed to fall onto a high-speed, horizontally rotating disk constructed of quartz, graphite, or porcelain, small prills of uniform size are produced [70]. The product is spun out onto a water cooled diaphragm and collected. If the copper is contaminated with oxides, hydrogen chloride gas should be added to the chlorine gas stream to prevent the production of basic copper(II) chlorides that would contaminate the product [66].

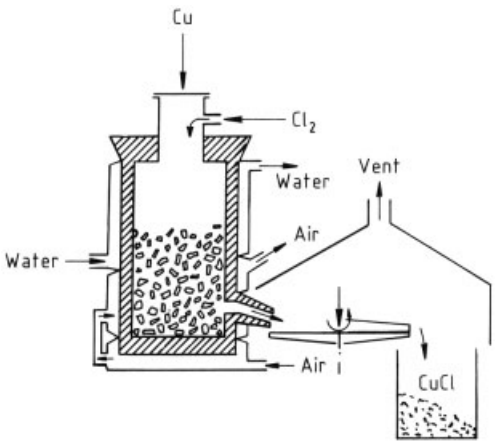

Figure 3. Crucible used in the production of copper(I) chloride

If the product is packaged quickly and sealed properly, no extreme precautions are required. Otherwise, the product must be stored under nitrogen to preserve its integrity or coated with mineral oil as a barrier to moisture.

Copper(I) chloride is also produced hydrometallurgically by the reduction of copper(II) in the presence of chloride ions [71]:

$$2\,CuCl_2 + Na_2SO_3 + H_2O \longrightarrow 2\,CuCl + Na_2SO_4 + 2\,HCl$$

Other reducing agents can be used, such as metallic copper, sulfurous acid, hydroxylamine, hydrazine, or phosphorous acid. The copper(I) chloride solution is produced, for example, by mixing a copper(II) chloride solution with metallic copper in the presence of hydrochloric acid or sodium chloride. The colorless to brown solution is stable only in the absence of air. Continuous preparations of copper(I) chloride solutions have been developed [72], [73]. When they are diluted with water, a white crystalline material precipitates which can be vacuum dried or washed with sulfurous acid, then with alcohol and ether, and carefully dried. Zinc has also been used as a reducing agent in a more recent process [74].

Production of copper(I) chloride by treatment of ores with iron(III) chloride solutions [75], [76] and recovery of the product through chlorination in pit furnaces above 800 °C [77] have also been attempted.

Uses. Copper(I) chloride is used as a precursor in the production of copper(II) oxychloride and copper(I) oxide, as well as fine copper powder [78]. The production of silicone polymers, the vulcanization of ethylene – propene rubbers (EPDM) [79], and acrylonitrile production are other applications. Copper(I) chloride is also used in the purification of carbon monoxide gas [80–83] and the production of phthalocyanine pigments [84], [85]. More recently, copper(I) chloride has been found to be an effective catalyst in the production of dialkyl carbonates [86–88].

Analysis and Specifications. Copper(I) chloride is analyzed according to [51]. Table 4 lists specifications for technical- and reagent-grade material.

Table 4. Specifications for copper(I) chloride

Assay	Mass fraction, %	
	Technical-grade	Reagent-grade
Copper (min.)	97.0	90.0
Acid-insolubles (max.)	0.1	0.02
Iron (max.)	0.01	0.005
Sulfate (max.)	0.3	0.10
Arsenic (max.)	–	0.001
Not precipitated by H_2S as sulfate (max.)	–	0.2

4.2. Copper(II) Chloride

Copper(II) chloride [*7447-39-4*], $CuCl_2$, M_r 134.45, *mp* (extrapolated) 630 °C, d_4^{25} 3.39, begins to decompose to copper(I) chloride and chlorine at about 300 °C. The often reported melting point of 498 °C is actually a melt of a mixture of copper(I) chloride and copper(II) chloride. Decomposition to copper(I) chloride and chlorine is complete at 993 °C. The deliquescent monoclinic crystals are yellow to brown when pure; their thermodynamic data are as follows: c_p (298 K) $-579.2\,J\,kg^{-1}\,K^{-1}$, c_p (288–473 K) $-621.7\,J\,kg^{-1}\,K^{-1}$, c_p (288–773 K) $-661.9\,J\,kg^{-1}\,K^{-1}$, $\Delta H°$ (25 °C) -247.2 kJ/mol. In moist air, the dihydrate is formed. Figure 4 shows the solubility of copper(II) chloride in water and hydrochloric acid at two temperatures [89]. At higher concentrations of hydrogen chloride, $[CuCl_3]^-$ and $[CuCl_4]^{2-}$ complexes are formed. Copper(II) chloride is easily soluble in methanol and ethanol and moderately soluble in acetone.

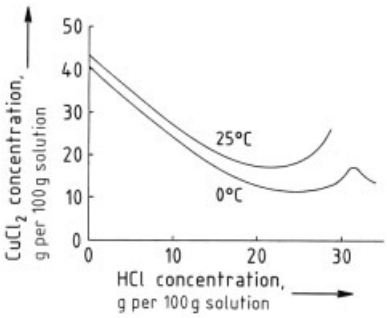

Figure 4. Solubility of copper(II) chloride in hydrochloric acid solutions at different temperatures

The more common commercial form of copper(II) chloride is the dihydrate [*10125-13-0*], $CuCl_2 \cdot 2\,H_2O$, M_r 170.45, *mp* around 100 °C (with decomposition to the anhydrous form). This occurs in nature as blue-green orthorhombic, bipyramidal crystals of eriochalcite, d_4^{25} 2.51. Its solubility characteristics are proportionally similar to those of the anhydrous form. In moist air the dihydrate deliquesces, and in dry air it effloresces.

Production. Because of the relative stabilities of copper(I) chloride and copper(II) chloride at high temperature, it is improbable that a pure anhydrous copper(II) chloride can be prepared by excessive chlorination of copper in a melt, even though such methods have been reported. The most common method for the production of anhydrous copper(II) chloride is by dehydration of the dihydrate at 120 °C. The product must be packaged in air-tight or desiccated containers.

The dihydrate can be prepared by the reaction of copper(II) oxide, copper(II) carbonate, or copper(II) hydroxide with hydrochloric acid and subsequent crystallization. Commercial production of copper(II) chloride dihydrate uses a tower packed with copper. An aqueous solution is circulated through the tower. Sufficient chlorine is passed into the bottom of the tower to oxidize the copper completely [72], [73]; to prevent hydrolysis [precipitation of copper(II) oxychloride] of concentrated copper(II) chloride solutions, they are kept acidic with hydrochloric acid. The tower can be operated batchwise or continuously; Figure 5 (see next page) shows the continuous operation. A hot, concentrated liquor is circulated continuously through the tower, and the overflow from the tower is passed through a crystallizer where the liquor is cooled; the product is then centrifuged, dried, and packaged. The addition of hydrogen chloride is pH controlled; the addition of water is controlled by specific gravity. Copper is added daily or twice daily, as needed.

Uses. Copper(II) chloride dihydrate is used in the preparation of copper(II) oxychloride [90], [91]. It serves as a catalyst in numerous organic chlorination reactions such as the production of vinyl chloride [92] or 1,2-dichloroethane [93]. Copper(II) chloride dihydrate is used in the textile industry as a mordant and in the

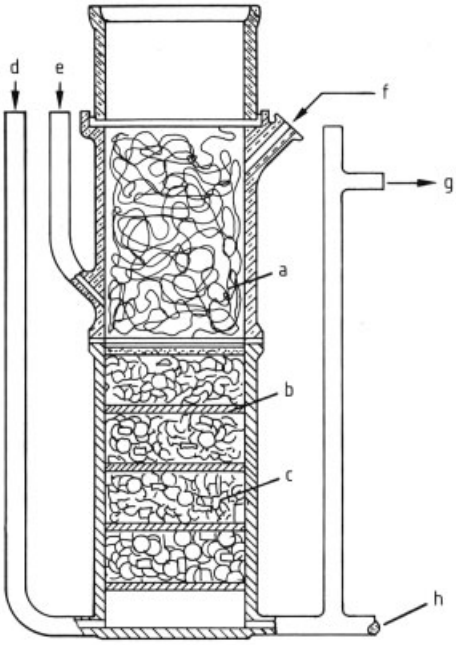

Figure 5. Reactor for the production of copper(II) chloride solutions
a) Copper metal; b) Porous plates; c) Raschig rings; d) Chlorine inlet; e) Steam inlet; f) Solution recycle; g) Solution to crystallizer; h) Drain

petroleum industry to sweeten sulfidic crude oil. Copper(II) chloride solutions are used for plating copper on aluminum, and in tinting baths for tin and germanium. Copper(II) chloride dihydrate is used as a pigment in glass and ceramics, as a wood preservative, and in water treatment.

Analysis and Specifications. Copper is analyzed iodometrically with sodium thiosulfate [51]. A typical analysis of technical-grade copper(II) chloride dihydrate (mass fractions in %) is as follows: copper(II) chloride dihydrate 99.0 (min.), iron 0.02, zinc 0.05, sulfate 0.05, water-insoluble material 0.01, material not precipitated by hydrogen sulfide as sulfate 0.15.

4.3. Copper(II) Oxychloride

Copper(II) oxychloride, $Cu_2Cl(OH)_3$ [*1332-65-6*], M_r 213.56, is usually written as $CuCl_2 \cdot 3\,Cu(OH)_2$. The trade name is copper oxychloride or basic copper chloride; the internationally accepted name (IUPAC) is dicopper chloride trihydroxide. Copper oxychloride is found in nature as the minerals paratacamite, green hexagonal crystals, and atacamite, green rhombic crystals, d_4^{25} 3.72 – 3.76 [94]. It is virtually insoluble in water, dissolves readily in mineral acids or warm acetic acid, and is soluble in ammonia and alkali-metal cyanide solutions. The green oxychloride is converted into blue copper(II) hydroxide in cold sodium hydroxide solution [59] and into the oxide in hot sodium hydroxide solution. With a lime suspension, copper oxychloride is converted into the blue calcium tetracuproxychloride, calcium tetracopper(II) chloride tetrahydroxide, $CaCl_2 \cdot 4\,Cu(OH)_2$ [95]. It is decomposed to the oxide at 200 °C.

Production. Copper(II) oxychloride is most often prepared commercially by air oxidation of copper(I) chloride solutions [90], [91], [96], [97]. For this purpose, a concentrated sodium chloride solution containing about 50 g/L copper(II) is contacted with copper metal to produce a solution containing about 100 g/L copper(I). The copper(I) chloride – sodium chloride solution, with or without the copper metal, is then heated to 60 – 90 °C and aerated to effect oxidation:

$$CuCl_2 + Cu \longrightarrow 2\,CuCl$$

$$6\,CuCl + 3\,H_2O + 3/2\,O_2 \longrightarrow CuCl_2 \cdot 3\,Cu(OH)_2 + 2\,CuCl_2$$

The mother liquor is separated from the precipitate and recycled into the process. The particles that are obtained by the above process are generally less than 4 µm in diameter and are suitable as crop fungicides. The particle size can be reduced further by increasing the agitation during oxidation and by utilizing a lower temperature [96]. Also, spray-drying of the product slurry gives a micronized product [98] as a result of deagglomeration of the material. Copper(II) oxychloride can also be prepared by reaction of a copper(II) chloride solution with sodium hydroxide [94]:

$$4\,CuCl_2 + 6\,NaOH \longrightarrow CuCl_2 \cdot 3\,Cu(OH)_2 + 6\,NaCl$$

or by reaction of a copper(II) chloride solution with freshly precipitated, hydrated copper(II) oxide [99]:

$$CuCl_2 + 2\,NaOH \longrightarrow CuO \cdot H_2O + 2\,NaCl$$

$$CuCl_2 + 3\,CuO \cdot H_2O \longrightarrow CuCl_2 \cdot 3\,Cu(OH)_2$$

Copper oxychloride has also been produced as a byproduct in the electrolytic production of copper(I) chloride [100].

Uses. Copper oxychloride is used primarily as a foliar fungicide [90], [101–105]; it is also used as a pigment.

Analysis and Specifications. A typical analysis of technical-grade copper(II) oxychloride (mass fractions in %) is as follows: copper 56.0, chloride 14.0 – 15.0, sulfate 2.0 – 2.5, hydroxide 22.5 – 23.5, water 3.0 – 6.0. Copper is analyzed by iodometric titration with sodium thiosulfate [51].

4.4. Copper(II) Sulfates

4.4.1. Copper(II) Sulfate Pentahydrate

$CuSO_4 \cdot 5\,H_2O$ [*7758-99-8*], M_r 249.61, d_4^{25} 2.285, c_p (273 – 291 K) 1126 J kg^{-1} K^{-1}, $\Delta H°$ (25 °C) − 850.8 kJ/mol, bluestone, blue vitriol, is found in nature as the mineral chalcanthite, blue trichlinic crystals that can be ground to a light blue powder. Copper(II) sulfate pentahydrate slowly effloresces in dry air or above 30.6 °C with the formation of the trihydrate, $CuSO_4 \cdot 3\,H_2O$. At 88 – 100 °C the trihydrate is produced more quickly. Thermal analysis of the pentahydrate gives the following:

$$CuSO_4 \cdot 5\,H_2O \xrightarrow{88\,°C} CuSO_4 \cdot 3\,H_2O \xrightarrow{114\,°C} CuSO_4 \cdot H_2O$$
$$\xrightarrow{245\,°C} CuSO_4 \xrightarrow{340\,°C} 3\,Cu(OH)_2 \cdot CuSO_4 \xrightarrow{600-650\,°C} CuO$$

Above about 114 °C, the monohydrate is formed, and between about 245 and 340 °C, the anhydrous product $CuSO_4$, results. Figures 6 A and 6 B show solubility of $CuSO_4$ in water and density of the solution as a function of temperature and sulfuric acid concentration [106], [107]. The pentahydrate can be crystallized from the solution either by addition of sulfuric acid or by evaporation. Although the addition of sulfuric acid appears to be more economical, the concentration of the solution by evaporation is preferred because the crystals obtained from a "neutral" (pH 3.5 – 4.0) medium are less prone to hard cake formation than the acid crystals. As the particle size of the pentahydrate decreases, the tendency toward hard cake formation increases, and the necessity to increase the pH during crystallization is enforced. When a free-flowing, commercial product of fine particle size is required, such alkaline additives as calcium oxide or calcium stearate must be incorporated into the final product to assure flowability. The incorporation of excess acid into the product accelerates the in situ dehydration of the pentahydrate and pro-

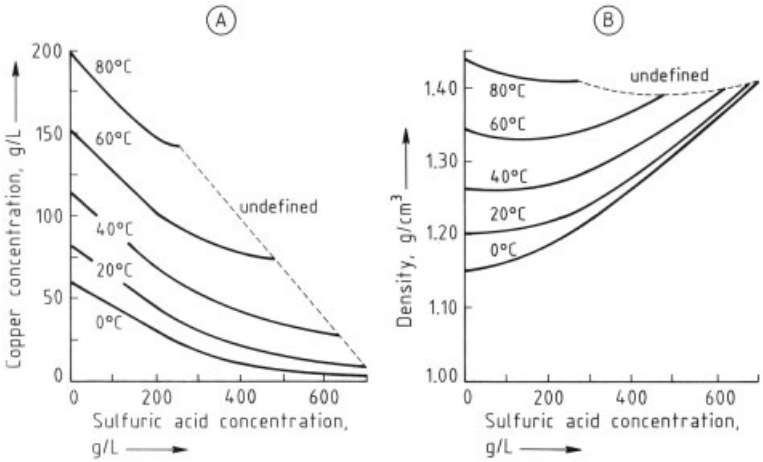

Figure 6. Saturated solutions of copper(II) sulfate
A. Solubility of copper(II) sulfate (as copper) as a function of sulfuric acid concentration and temperature; B. Density of copper(II) sulfate solutions as a function of sulfuric acid concentration and temperature
Reprinted with permission [107]

motes hard cake formation. Lower temperature and lower humidity slow the caking process.

Copper(II) sulfate pentahydrate is soluble in methanol (15.6 g/100 mL solution) but insoluble in ethanol. It readily forms soluble alkaline complexes at sufficiently high concentrations of amines or alkali cyanides (see Chap. 5), but basic sulfates are precipitated from solution by ammonia at an intermediate pH (about 4.2 – 6.8).

Copper(II) sulfate pentahydrate is the most commonly used copper compound because of the economics and the availability of starting materials, the ease of production, and the extent of byproduct utilization (primarily copper electrowinning liquors).

Production. Copper(II) sulfate pentahydrate is prepared most easily by the reaction of a basic copper(II) compound with a sulfuric acid solution (100 – 200 g/L H_2SO_4), e.g.:

$$CuO + H_2SO_4 \longrightarrow CuSO_4(aq) + H_2O$$

Copper metal, sulfuric acid, and air are the most common starting materials for the production of copper sulfate pentahydrate:

$$Cu + H_2SO_4 + 1/2\,O_2 \xrightarrow{65-100\,^{\circ}C} CuSO_4(aq) + H_2O$$

Harike Process. The Harike process [108] is the best commercial example of the preceding reaction (Fig. 7). Blister shot copper up to 25 mm in diameter is added to the reaction tower (2.9 m^2 cross – sectional area) to give a bed of copper metal 2.74-m high. Two different reaction conditions are given in Table 5. Condition A allows for the production of concentrated copper(II) sulfate solutions and subsequent crystallization from acid media. Condition B gives a concentrated neutral solution of copper(II) sulfate that can be crystallized or diluted with water and used for direct production of other copper products. The rate of oxygen consumption by the system is directly proportional to the mass of copper metal

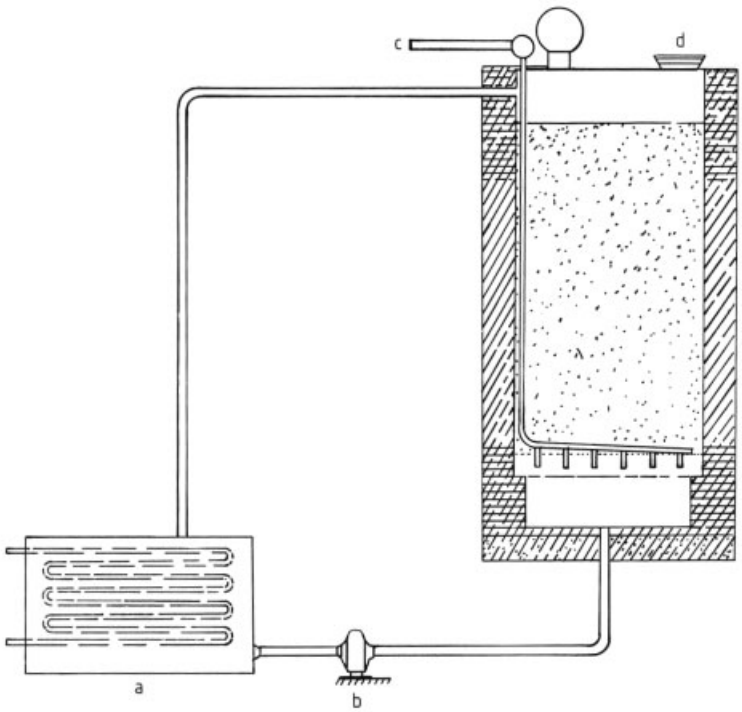

Figure 7. Harike tower process for production of copper(II) sulfate solutions
a) Heat exchanger; b) Circulation pump; c) Air inlet; d) Copper input

dissolved. Air flows of 46 $m^3\,h^{-1}\,m^{-2}$ are used in order to fill with air the voids that are created by the packing of the copper shot. The solution is circulated cocurrently with the air flow to wet the particles continuously and enhance mixing of the solution with the copper metal. If the air flow is decreased and the voids in the copper metal bed are not filled sufficiently, the hot acid solution oxidizes the copper metal in the absence of air with the formation of a copper sulfide film. This film renders the copper metal inert to further oxidation and would lower the efficiency of the tower. On the other hand, an increase in air flow above 46 $m^3\,h^{-1}\,m^{-2}$ decreases the fraction of oxygen consumed and would also result in a lowered efficiency of the tower. If an increased production (greater oxygen consumption) with good tower efficiencies is required, the height of the tower must be increased. This would also be necessary if a copper metal of lower surface area was used such as scrap copper instead of the blister shot copper. The use of cement copper and fine chopped wire creates insufficient voids for adequate air passage; such fine copper must be leached in agitated vessels.

Table 5. Production of copper(II) sulfate by the Harike method

	A∗	B∗
Initial concentration, Cu (g/L)	100	100
H_2SO_4 (g/L)	160	80
Final concentration, Cu (g/L)	160	160
H_2SO_4 (g/L)	80	0
Temperature, °C	85	85
Circulation rate, m^3/h	34–45	34–45
Air to tower, $m^3\,h^{-1}\,m^{-2}$	46	46
Entering oxygen, vol %	20.9	20.9
Exiting oxygen, vol %	2.6	7.3
Oxygen consumed, vol %	90	70
Production rate ($CuSO_4 \cdot 5\,H_2O$)		
t/m^2 of tower cross section	4.65	3.6
t/tower	13.2	10.3

∗ For processes A and B, see text.

Two-Tower Process. Another commercial method for the preparation of copper(II) sulfate pentahydrate utilizes two towers filled with copper shot. One of the towers is filled with a sulfuric acid solution, whereas the other is sparged with air and steam to oxidize the surface of the copper metal and form a layer of copper(II) oxide. The solution is pumped alternately to the other tower to dissolve the layer of oxide while the now drained tower is sparged with air and steam. This process is continued until the desired copper(II) sulfate concentration is reached.

The Trickle method is also used commercially: a copper-filled tower is continuously sprayed from the top with leach liquor from an adjacent reservoir. The liquor drains through the tower and is pumped back to the reservoir. Steam and air are sparged continuously into the tower from the bottom. The process is continued until the copper sulfate solution has the desired concentration.

Solvent Extraction. Copper(II) sulfate pentahydrate is produced from alkaline ammoniacal copper(II) solutions by solvent extraction [107], [109]. The alkaline copper(II) solution is first contacted with an organic extractant that is selective to copper. One of the most common extractants can be represented by the following general formula:

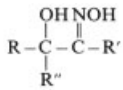

where R and R′ are preferably unsaturated alkyl groups and R″ is preferably hydrogen [110].

$$[Cu(NH_3)_4]^{2+}\,(aq) + 2\,RH \longrightarrow CuR_2 + 2\,NH_3\,(aq) + 2\,NH_4^+\,(aq)$$

The now copper-barren aqueous layer is continuously separated from the copper-containing organic layer which is subsequently contacted with a sulfuric acid solution:

$$CuR_2 + H_2SO_4 \longrightarrow CuSO_4 + 2\,RH$$

The saturated copper(II) sulfate liquor produced is cooled or concentrated by evaporation to crystallize the pentahydrate. The copper-barren ammonia–ammonium salt solution is continuously returned to the copper leach tank. Figure 8 (see next page) shows a flow diagram for a typical solvent extraction ammonia leach circuit utilizing mixer–settlers. With the exception of small quantities of wash water and loss of ammonia, the plant produces no effluent. Also, this process allows for lower overall energy input than acid dissolution, primarily as a result of the more facile oxidation of copper metal in ammoniacal solutions. Ion-exchange resins are

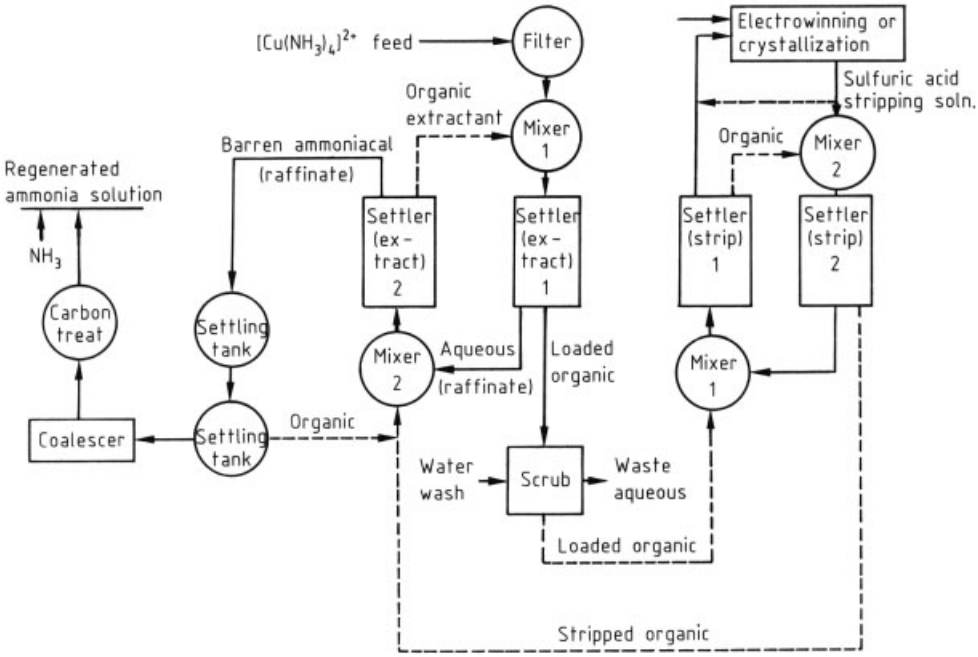

Figure 8. Flow diagram for a solvent extraction plant

used to recover and concentrate copper from the solution [111], [112]. Elution of the resins with sulfuric acid produces copper(II) sulfate solutions.

Byproduct Recovery. About 20 – 30 % of the total copper(II) sulfate pentahydrate produced comes from the electrowinning industry as a byproduct. When the impurity content of the electrowinning liquors is too high to make cathode copper, they are concentrated and crystallized [113]. The product is suitable for most agricultural applications. Acidic copper(II) sulfate etching wastes are neutralized with a copper sludge, filtered, and concentrated to produce an agricultural-grade pentahydrate [114], [115].

Uses. The major uses of copper(II) sulfate pentahydrate have been grouped as follows [113]:

Agriculture (feed supplement, soil nutrient, fungicide)	41 %
Industrial algicides	27 %
Mining (flotation activator)	10 %
Electroplating	5 %

In many instances, the copper(II) sulfate pentahydrate is used as an intermediate, e.g., to produce active foliar fungicides such as Bordeaux mixture, tribasic copper(II) sulfate, or copper(II) hydroxide. A substantial amount of copper(II) sulfate pentahydrate is used in combination with sodium dichromate and arsenic acid for the preservation of wood. Copper(II) sulfate pentahydrate is an effective and economical algicide for lakes and ponds. In the mining industry, it is used as a flotation activator for lead, zinc, and cobalt ores. Copper(II) sulfate solutions are often used in the electroplating industry. The compound is also used as a mordant in textile dyeing, in the preparation of azo and formazan dyes, as a pigment in paints and varnishes, for preserving hides and tanning leather, in pyrotechnic compositions, and in synthetic fire logs.

Analysis and Specifications. Typical analyses for various commercially available crystal sizes are shown in Table 6. Table 7 (both see next page) lists specifications for an electroless copper(II) sulfate pentahydrate. The analytical methods for copper(II) sulfates are given in [51].

Table 6. Typical analysis of copper(II) sulfate pentahydrate

Assay	Mass fraction, %			
	Industrial, granular	Large, medium	Snow, superfine	Powdered
Copper	25.2	25.2	25.2	25.2
Iron	0.008	0.007	0.008	0.17
Silicon dioxide	0.006	–	0.42	–
Water-insolubles	0.008	0.06	0.11	0.11
Calcium oxide	–	–	0.001	0.15

Large crystals:	on 9.5 mm × 19.0 mm through 28.6 mm screen;
Medium crystals:	on 3.2 mm × 12.7 mm through 19.0 mm screen;:
Granular crystals	on U.S. 20 mesh through U.S. 6 mesh;
Industrial crystals:	on U.S. 14 mesh;
Snow crystals:	through U.S. 18 mesh;
Superfine crystals:	through U.S. 25 mesh;
Powdered:	85 % through U.S. 100 mesh

Table 7. Specifications of electroless-grade copper(II) sulfate pentahydrate

Assay	Mass fraction, %
Copper	25.1
Iron	<0.0015
Lead	<0.003
Nickel	<0.001
Manganese	<0.0005
Chromium	<0.0005
Chlorine	nil
Water-insolubles	0.01

4.4.2. Anhydrous Copper Sulfate

$CuSO_4$ [*7758-98-7*], M_r 159.61, occurs in nature as the mineral hydrocyanite [*14567-54-5*]. It is gray to white and has a rhombic crystal morphology. It decomposes to the green basic copper(II) sulfate at 340 °C, and at 600 – 650 °C it decomposes to copper(II) oxide. Some of the properties of the anhydrous salt are as follows: d_4^{25} 3.6, c_p (273 – 291 K) 631.8 J kg^{-1} K^{-1}, c_p (273 – 373 K) 657.3 J kg^{-1} K^{-1}, $\Delta H°$ (25 °C) − 771.6 kJ/mol. The compound is soluble in water (Section 4.4.1), somewhat soluble in methanol (1.1 g/100 mL), but insoluble in ethanol. It readily dissolves in aqueous ammonia and excess alkali metal cyanides, with the formation of complexes. The material is hygroscopic, with conversion to the pentahydrate in moist air below 30 °C.

The anhydrous salt is prepared by careful heating of the pentahydrate salt to 250 °C. It can also be prepared by the reaction of hot concentrated sulfuric acid with copper metal, but purification is rather difficult.

The compound has limited commercial use, but it can be used as a desiccant for removing water from organic solvents and is a sensitive indicator of the presence of moisture in such solvents.

4.4.3. Copper(II) Sulfate Monohydrate

$CuSO_4 \cdot H_2O$ [*10257-54-2*], M_r 175.63, d_4^{20} 3.25, is a whitish powder. The solubility of the monohydrate in water is identical to the pentahydrate on a copper basis. The product is hygroscopic and must be packaged in containers with moisture barriers. It is commercially produced by dehydration of the pentahydrate at 120 – 150 °C. A novel preparation is to triturate stoichiometric ratios of copper(II) oxide and sulfuric acid:

$$CuO + H_2SO_4 \longrightarrow CuSO_4 \cdot H_2O$$

The uses of the monohydrate are analogous to those of the pentahydrate. Although there can be slight economic advantages (primarily freight costs) in using the monohydrate, market acceptance has not been great. Presently, less than 5 % (copper basis) of copper(II) sulfate is marketed in the monohydrate form.

4.4.4. Basic Copper(II) Sulfates

Four distinct basic copper(II) sulfates can be identified by potentiometric titration of a cop-

per(II) sulfate solution with sodium carbonate or sodium hydroxide solution [115], [116]: $CuSO_4 \cdot 3\,Cu(OH)_2 \cdot H_2O$ (langite), $CuSO_4 \cdot 2\,Cu(OH)_2$ (antlerite), $CuSO_4 \cdot 3\,Cu(OH)_2$ (brochantite), and $CuSO_4 \cdot CuO \cdot 2\,Cu(OH)_2 \cdot x\,H_2O$. Their unique crystal morphologies are confirmed by X-ray diffraction.

The most important commercial basic copper(II) sulfate, commonly referred to as *tribasic copper sulfate* [*12068-81-4*], is $CuSO_4 \cdot 3\,Cu(OH)_2$, M_r 452.27, *mp* ca. 380 °C (decomp.), d_4^{25} 3.78, occurs in nature as the green monoclinic mineral brochantite. It is readily soluble in mineral acids, acetic acid solution, and ammonia solution; it is insoluble in water. Above 650 °C it decomposes to copper(II) oxide. If anhydrous copper(II) sulfate is heated cautiously to 650 °C, another naturally occurring mineral is obtained, dolerophane, $CuSO_4 \cdot CuO$, which reacts readily with water at 20 °C to form $CuSO_4 \cdot 3\,Cu(OH)_2 \cdot H_2O$ or at 100 °C to form $CuSO_4 \cdot 2\,Cu(OH)_2$ [117].

Production. Tribasic copper(II) sulfate, the commercially available basic copper(II) sulfate, is most often prepared by the addition of sodium carbonate solution to hot solutions of copper(II) sulfate:

$$4\,CuSO_4 + 3\,Na_2CO_3 + 3\,H_2O \longrightarrow CuSO_4 \cdot 3\,Cu(OH)_2 + 3\,Na_2SO_4 + 3\,CO_2$$

Great variation of particle size can be achieved by control of the precipitation temperature. As the temperature is increased, larger particles with greater bulk density are formed. If a small particle size is desired, lower precipitation temperature should be used. However, as the precipitation temperature is lowered, multiple hydrates and larger amounts of sulfate are incorporated into the product; e.g., around 50 °C, the dried product contains only about 50 % copper. A precipitation temperature around 90 °C is required to make a pure tribasic copper(II) sulfate.

The single largest use of the compound is as a crop fungicide, and small particles are preferred because they give a greater degree of coverage. These small particles can be obtained by high-energy attrition of the product or by carefully controlled precipitation conditions.

Another method of making tribasic copper(II) sulfate is the aeration of a suspension of copper(I) oxide in the presence of stoichiometric quantities of sulfuric acid or copper(II) sulfate [118]:

$$2\,Cu_2O + H_2SO_4 + O_2 + 2\,H_2O \longrightarrow CuSO_4 \cdot 3\,Cu(OH)_2$$

$$6\,Cu_2O + 4\,CuSO_4 + 3\,O_2 + 12\,H_2O \longrightarrow 4\,[CuSO_4 \cdot 3\,Cu(OH)_2]$$

A purified, concentrated solution of copper(II) sulfate containing ammonium sulfate can be neutralized to pH 6.0 – 6.5 with ammonia to give a blue precipitate of approximate stoichiometry $4\,CuSO_4 \cdot 12\,Cu(OH)_2 \cdot 5\,H_2O$. When this is dried, a green tribasic copper(II) sulfate is formed with up to 1 mol of hydration water per mole of tribasic salt [119]; the water content depends on the drying temperature.

A continuous process has been developed in which copper(II) sulfate solutions are neutralized to pH 5.9 at 30 °C by addition of gaseous ammonia to an agitated vessel [120]. A unique, germicidally active, basic copper(II) sulfate whose X-ray diffraction pattern differs from langite and brochantite has been prepared in this way [121].

Reaction of copper(II) sulfate with $Na_3PO_4 \cdot 12\,H_2O$ and sodium hydroxide in aqueous solution gives a fine, blue, basic copper(II) sulfate powder (16 % SO_4, 2.6 % PO_4, and 54.9 % Cu) that is formulated into a germicidal powder.

Tribasic copper(II) sulfate can also be prepared by cautiously heating copper(II) sulfate to 340 °C or by the aeration of a hot copper(II) sulfate solution in contact with copper metal.

Other basic copper(II) sulfates of commercial interest are the classic fungicial mixtures, Bordeaux and Burgundy slurries, in which copper(II) sulfate solution is mixed with lime and soda ash, respectively. They are of variable stoichiometry. More recently, the aqueous Bordeaux suspension has been dried to yield a stable basic copper(II) sulfate powder of variable composition [122].

Uses. As stated earlier, basic copper(II) sulfate is primarily used as a crop fungicide [123], [124]. It has also been utilized as a precursor in the separation of copper from metallic impurities [125].

Analysis and Specifications. The analysis of a typical technical-grade tribasic copper(II) sulfate is (mass fractions in %) as follows: copper 53.5, iron 0.08, sulfate 19.0, carbonate 2.0, water 2.0 – 5.0. This is analyzed according to [51].

5. Compounds and Complexes of Minor Importance

5.1. Copper Compounds

Copper(I) acetate [*598-54-9*], $CuCH_3COO$, M_r 122.6, is obtained as white crystals which are stable when dry but decompose on exposure to water. The product is obtained by reducing an ammoniacal solution of copper(II) acetate; on acidification with acetic acid, the crystals precipitate. Ammoniacal solutions of copper(I) acetate are used commercially to absorb olefins.

Copper(II) acetate monohydrate [*6046-93-1*], neutral verdigris, $Cu(CH_3COO)_2 \cdot H_2O$, M_r 199.65, *mp* 115 °C, d_4^{25} 1.88, decomposes at 240 °C; it forms dark green, monoclinic crystals; its solubility in water at 25 °C is 6.8 g/100 g solution; the compound is slightly soluble in methanol, diethyl ether, and acetone.

Copper(II) acetate monohydrate is produced by the reaction of copper(II) carbonate or copper(II) hydroxide with a solution of acetic acid or by the reaction of copper(II) oxide with hot dilute acetic acid. Alternatively, the material can be made by refluxing aqueous acetic acid in the presence of copper metal and air. The technical product is generally 99 % pure.

Copper(II) acetate monohydrate is used in textile dyeing and as a ceramic pigment. It is used as a fungicide, as a precursor in the production of Paris green [copper(II) acetoarsenite], and as a polymerization catalyst in organic reactions.

Copper(II) acetate, basic [*52503-63-6*], $Cu(CH_3COO)_2 \cdot CuO \cdot 6H_2O$ (variable), M_r 369.26, exists as blue to green salts, depending on the amount of water of hydration. Blue verdigris has the above formula, while green verdigris has fewer molecules of water of hydration. The salts are slightly soluble in water or ethanol, and soluble in dilute mineral acid or aqueous ammonia.

The basic salts are produced by neutralizing copper(II) acetate solutions. They can also be prepared by refluxing acetic acid over copper in the presence of air until the basic salt precipitates.

The basic copper(II) acetates are used as precursors in the manufacture of Paris green [copper(II) acetoarsenite] and as fungicides. They are used as pigments in oil- and water-based paints and in textile dyeing.

Copper(II) acetoarsenite [*12002-03-8*], (acetato)trimetaarsenitodicopper(II), $Cu(CH_3COO)_2 \cdot 3Cu(AsO_2)_2$ (variable), M_r 1013.77, also known as Paris green, is an emerald green, poisonous powder. It is virtually insoluble in water and ethanol but dissolves in dilute mineral acids and aqueous ammonia. Copper(II) acetoarsenite is primarily made by the reaction of a solution of copper(II) sulfate with arsenic(III) oxide, sodium carbonate, and acetic acid

$$4\,CuSO_4 + 3\,As_2O_3 + 4\,Na_2CO_3 + 2\,CH_3COOH \longrightarrow Cu(CH_3COO)_2 \cdot 3\,Cu(AsO_2)_2 + 4\,Na_2SO_4 + H_2O + 4\,CO_2$$

or by the reaction of copper(II) oxide with a hot solution of acetic acid and arsenic(III) oxide. A solution of copper(II) acetate or a suspension of basic copper(II) acetate can also react with arsenic(III) oxide. The product is used as an insecticide, in the preservation of wood, and as an antifouling pigment.

Copper(II) arsenate [*7778-41-8*], $Cu_3(AsO_4)_2 \cdot 4H_2O$ (variable), M_r 540.52, is a blue-green to blue insoluble powder. It is most often prepared by reaction of copper(II) sulfate solutions with arsenic(V) oxide and sodium hydroxide. Copper(II) arsenate is used as an insecticide, fungicide, rodenticide, wood preservative, and antifouling pigment.

Copper(II) arsenite, $CuHAsO_3$ (variable), M_r 187.47, Scheele's green, an insoluble green powder, is prepared by the reaction of a copper(II) sulfate solution with arsenic(III) oxide and sodium hydroxide and is used as pigment and insecticide.

Copper(I) bromide [*7787-70-4*], CuBr, M_r 143.45, *mp* 504 °C, *bp* 1345 °C, d_4^{25} 4.72, forms white cubic crystals that slowly decompose on exposure to light or to moist air; it is soluble in hydrochloric and hydrobromic acids and in aqueous ammonia, but very slightly soluble in water. This compound is prepared pyrometallurgically [see copper(I) chloride, Section 4.1] or by reducing a copper(II) sulfate solution in the presence of sodium bromide, usually with sulfur dioxide or metallic copper. Copper(I) bromide is used as a polymerization catalyst for organic reactions.

Copper(II) bromide [*7789-45-9*], $CuBr_2$, M_r 223.36, *mp* 498 °C, d_4^{25} 4.77, crystallizes from warm solution as black, deliquescent, monoclinic crystals. Below 29 °C, the green tetrahydrate, $CuBr_2 \cdot 4\,H_2O$, results. The anhydrous product is very soluble in water (55.7 g/100 g solution) and soluble in ethanol and acetone. It is insoluble in diethyl ether, benzene, or concentrated sulfuric acid. Copper(II) bromide is most conveniently prepared by dissolving copper(II) oxide in hydrobromic acid; it can also be prepared by the direct action of bromine water on metallic copper. Copper(II) bromide is used as a brominating reagent and catalyst in organic synthesis and as an intensifier in photography.

Copper(II) chromate(VI) [*13548-42-0*], $CuCrO_4$, M_r 179.55, is obtained as reddish brown crystals. Numerous insoluble basic salts can be formed from solutions. Neutral copper(II) chromate is prepared by direct heating of a mixture of copper(II) oxide and chromium(VI) oxide. It decomposes around 400 °C, with the formation of copper(II) chromate(III). Copper(II) chromate(VI) is used in wood preservation and in the weatherproofing of textiles.

Copper(II) chromate(III) [*12018-10-9*], $CuCr_2O_4$, M_r 231.56, forms black, tetragonal crystals which are insoluble in water. It is prepared by heating neutral copper(II) chromate(VI) to 400 °C. Copper(II) chromate(III) is used as a hydrogenation catalyst.

Copper(I) cyanide [*544-92-3*], CuCN, M_r 89.56, *mp* 474 °C (in nitrogen), d_4^{25} 2.92, is usually a white to cream-colored powder, which is practically insoluble in water, cold dilute acids, and ethanol, but dissolves in ammonia and alkali cyanide solutions with the formation of complexes. Copper(I) cyanide is produced by the addition of an alkali cyanide and sodium hydrogensulfite to a solution of copper(II) sulfate:

$$2\,CuSO_4 + 2\,NaCN + NaHSO_3 + H_2O \longrightarrow 2\,CuCN + 3\,NaHSO_4$$

Copper(I) cyanide is very poisonous and must be handled cautiously. It is used in electroplating and in organic reactions as a polymerization catalyst or as a means of introducing the cyanide moiety. Copper(I) cyanide has also been used as an antifouling pigment for marine paints and is an active fungicide and insecticide.

Copper(II) formate [*544-19-4*], $Cu(HCOO)_2$, M_r 153.58, *mp* ca. 200 °C (decomp.), d_4^{25} 1.831, is soluble in water and slightly soluble in ethanol. The reaction of copper(II) oxide, carbonate, or hydroxide with formic acid is used to prepare the product. The most common commercial form, a royal blue material, is prepared by crystallization from water at 75 – 85 °C. A metastable dihydrate is produced by crystallization at 50 – 60 °C. Copper(II) formate tetrahydrate is prepared by crystallization from cool solutions. Copper(II) formate is used to prevent bacterial and mildew growth in cellulosic materials.

Copper(II) gluconate [*527-09-3*], $Cu(C_6H_{11}O_7)_2$, M_r 453.85, forms light blue to blue-green crystals from water; it is soluble in water and slightly soluble in ethanol. Copper(II) gluconate is prepared by the reaction of gluconic acid solutions with basic copper(II) carbonates or copper(II) hydroxide. It is used as a dietary supplement for copper deficiency and in mouth deodorants or breath fresheners.

Copper(I) iodide [*7681-65-4*], CuI, M_r 190.49, *mp* 605 °C, *bp* 1290 °C, d_4^{25} 5.62, occurs in nature as the mineral marshite, white to reddish brown cubic crystals. It is virtually insoluble in water but dissolves in ammonia solution, alkali iodide and cyanide solutions, and dilute hydrochloric acid. Copper(I) iodide is manufactured pyrometallurgically by the reaction of hot copper with iodine vapor. Alternatively, copper(II) salt solutions are reacted with alkali iodides to precipitate copper(I) iodide. Copper(I)

iodide is used as a heat and light stabilizer in polymers, photographic emulsions, and light-sensitive papers, and in oil drilling to aid in corrosion inhibition in highly acid environments. It is also used as a feed additive, in cloud seeding, and as a double salt with mercury(II) iodide as a temperature indicator.

Copper(II) nitrate trihydrate [*10031-43-3*], $Cu(NO_3)_2 \cdot 3\,H_2O$, M_r 241.59, *mp* 114.5 °C, d_4^{25} 2.32, is a blue, deliquescent salt that crystallizes as rhombic plates. It is very soluble in water (77.4 g/100 g solution) and ethanol. When crystallized from solution below the transition point (26.4 °C), the hexahydrate is produced, $Cu(NO_3)_2 \cdot 6\,H_2O$ [*13478-38-1*], M_r 295.64, *mp* 26.4 °C (with loss of 3 mol of water of hydration), d_4^{24} 2.07, blue, deliquescent, prismatic crystals. The anhydrous nitrate is not produced by heating the hydrates; instead, decomposition to the basic copper(II) nitrate [*12158-75-7*], $Cu_2(NO_3)(OH)_3$, begins around 80 °C. Conversion to copper(II) oxide is complete at 180 °C. Copper(II) nitrates are obtained by dissolving copper(II) oxide, basic copper(II) carbonate, or copper(II) hydroxide in nitric acid solution. The various hydrate crystals obtained depend on the conditions of crystallization. Basic copper(II) nitrate can be precipitated directly from solution by neutralization. Copper(II) nitrate salts are often produced by dissolving copper metal in nitric acid solution. The reaction is vigorous, and oxides of nitrogen are evolved:

$$Cu + 4\,HNO_3 \longrightarrow Cu(NO_3)_2 + 2\,H_2O + 2\,NO_2$$

$$3\,Cu + 8\,HNO_3 \longrightarrow 3\,Cu(NO_3)_2 + 4\,H_2O + 2\,NO$$

The second reaction is favored by a lower temperature and dilute acid. Anhydrous copper(II) nitrate [*3251-23-8*] can be produced in ethyl acetate:

$$Cu + N_2O_4 \longrightarrow Cu(NO_3)_2 \cdot N_2O_4 \xrightarrow{90\,°C} Cu(NO_3)_2$$

It is not available commercially.

The hydrates of copper(II) nitrate are used in the textile industry as oxidants and mordants in dyeing. They are used to prepare a copper(II) oxide of high surface area, which is useful as a catalyst in numerous organic reactions. The nitrate salts are used as colorants in ceramics, in the preparation of light-sensitive papers, and in the burnishing of iron, as well as the browning of zinc and brightening of aluminum. They are also used in pyrotechnics and as an oxidative component in solid rocket fuels.

Copper(II) oxalate [*814-91-5*], CuC_2O_4, M_r 151.56, is a blue-white powder, which is very slightly soluble in water. Copper(II) oxalate is prepared by precipitation from a mixture of a copper(II) salt and sodium oxalate solution. The product is used as a stabilizer for acetylated polyformaldehyde and as a catalyst in organic reactions.

Copper(II) phosphate trihydrate [*7798-23-4*], $Cu_3(PO_4)_2 \cdot 3\,H_2O$, M_r 434.61, is a light blue powder which is insoluble in cold water, slightly soluble in hot water, and soluble in ammonia solution and mineral acids. This compound is prepared by reaction of copper(II) sulfate solution with soluble alkali phosphates; it precipitates as a voluminous, almost gelatinous, product. Copper(II) phosphate trihydrate is used as a fungicide and corrosion inhibitor.

Copper(II) diphosphate hydrate [*10102-90-6*], $Cu_2P_2O_7 \cdot x\,H_2O$ (variable), has the following typical analysis: Cu 33 – 36 %, P_2O_7 45 – 49 %. Copper(II) pyrophosphate is a light blue powder which is insoluble in water, but soluble in solutions containing an excess of diphosphate. It is prepared by the precipitation reaction of solutions of copper(II) sulfate and alkali diphosphate. Solutions are used for plating copper on plastics, aluminum, and zinc.

Copper(II) selenide [*1317-41-5*], CuSe, M_r 142.5, *mp* not defined, d_4^{25} 6.0, decomposes at dull red heat; it is insoluble in water but dissolves in hydrochloric acid with evolution of H_2Se. Copper(II) selenide is prepared by reducing copper(II) selenite with hydrazine. It is used as a catalyst in the digestion of organic chemicals by the Kjeldahl method.

Copper(II) soaps are water-insoluble copper salts of long-chain fatty acids. They are usually sold as 6 – 10 % copper solutions in a kerosene diluent. The copper(II) soaps are commonly prepared by direct reaction of the fatty acid with copper(II) hydroxide or basic copper(II) carbonate in an organic diluent, or by precipitation from aqueous media when copper(II)

sulfate solution is mixed with the sodium salt of the respective fatty acid. The common commercially available soaps, concentrations, and uses follow.

Copper (II) naphthenate [*1338-02-9*] contains 8 % copper and is used as a mildewcide in textiles, woods, and paints, and to prevent barnacle growth on ship bottoms.

Copper (II) oleate (9-octadecenoate) [*1120-44-1*], M_r 626.43, contains between 6 and 9 % copper, and is used as a combustion improver in fuel oils, as an emulsifier and dispersant, and as an antifouling coating for fish nets and lines.

Copper (II) stearate (octadecanoate) [*660-60-6*], M_r 630.46, contains 10 % copper, and is used in antifouling paints and in the preservation of wood and textiles.

Copper(I) sulfide [*22205-45-4*], Cu_2S, M_r 159.15, *mp* 1100 °C, d_4^{25} 5.6, occurs in nature as the mineral chalcocite [*21112-20-9*], blue to gray rhombic crystals, also known as copper glance. It has a lustrous metallic look. Copper(I) sulfide is virtually insoluble in water, but is soluble with decomposition in nitric acid and concentrated sulfuric acid, and soluble in alkali cyanide solution through complex formation. It is decomposed to copper(II) oxide, copper(II) sulfate, and sulfur dioxide by heating in air, and to copper(II) sulfide and copper by heating in the absence of air. Copper(I) sulfide is prepared by heating copper and sulfur in a hydrogen atmosphere or by precipitation with hydrogen sulfide from an ammoniacal copper(II) salt solution. It is used in luminous paints, lubricants, solar cells, thermoelements, and semiconductors.

Copper(II) sulfide [*1317-40-4*], CuS, M_r 95.60, d_4^{25} 4.6, decomposes at 220 °C and occurs in nature as the mineral covellite [*19138-68-2*], as blue-black hexagonal or monoclinic crystals. It is virtually insoluble in water, but soluble in alkali cyanides and in ammonia solution with complex ion formation. This compound is decomposed by hot nitric acid. It is stable in dry air, but is slowly oxidized to copper(II) sulfate in moist air. Copper(II) sulfide can be prepared by melting an excess of sulfur with copper(I) sulfide or by precipitation with hydrogen sulfide from a solution of anhydrous copper(II) chloride in anhydrous ethanol. It is used in the dye industry for the preparation of aniline black dyes and as an antifouling pigment.

Copper(II) tetrafluoroborate [*38465-60-0*], $Cu(BF_4)_2$, M_r 237.15, is prepared commercially by the neutralization of tetrafluoroboric acid with copper(II) hydroxide or basic copper(II) carbonate. It is generally produced as a concentrated solution and is used in the production of printed circuits and in copper plating.

Copper(I) thiocyanate [*1111-67-7*], CuSCN, M_r 121.62, *mp* 1084 °C (under nitrogen), d_4^{25} 2.84, is a white to yellow amorphous powder, which is very slightly soluble in water and dilute mineral acids, and soluble in ammonia solution, alkali thiocyanate solutions, and diethyl ether. It is stable to dry air but slowly decomposes in moist air. The material is prepared by reacting a copper(I) chloride solution with an alkali metal thiocyanate at 80 – 90 °C. The recovered product is dried under nitrogen. Copper(I) thiocyanate is used as an antifouling pigment.

5.2. Copper Complexes

Copper(I) and copper(II) ions form many stable complexes with halides, amines, azo compounds, cyanides, and other complexing media. This, in part, accounts for the fact that more X-ray crystal structures have been determined for copper complexes than for any other first-row transition metal ion. Many of these complexes are of great commercial significance. The dissociation constants for a number of copper complexes are given in Table 8 (see next page).

Copper Ammine Complexes. Copper(II) salts form complexes of the type $[Cu(NH_3)_n]^{2+}$ where $n = 1 - 5$. The pentaammine complex is favored only in concentrated ammonia solution (Table 8). The hexaammine complex is formed only in anhydrous ammonia. The tetraammine complex of copper(II) is favored at low ammonia concentrations. For the copper(I) complexes of ammonia, the diammine complex is favored. Tetraammine copper(II) hydroxide, $[Cu(NH_3)_4](OH)_2$, Schweitzer's reagent, is used in the dissolution of cellulose and in the production of rayon (→ Cellulose). Copper(I)

Table 8. Dissociation constants of copper(I) and copper(II) complexes (p$K = -\log K$)

Complex ion	Dissociation products		pK
$[CuCl_3]^-$	CuCl	$+2\,Cl^-$	−2.0
$[CuCl_2]^-$	Cu^+	$+2\,Cl^-$	4.7
$[CuBr_2]^-$	CuBr	$+Br^-$	−3.3 to −2.3
$[CuBr_2]^-$	Cu^+	$+2\,Br^-$	5 to 6
$[Cu(CN)_3]^{2-}$	$[Cu(CN)_2]^-$	$+CN^-$	11.3
$[Cu(CN)_2]^-$	Cu^+	$+2\,CN^-$	16
$[Cu(NH_3)_2]^+$	Cu^+	$+2\,NH_3$	10.8
$[Cu(NH_3)]^{2+}$	Cu^{2+}	$+NH_3$	4.1
$[Cu(NH_3)_2]^{2+}$	Cu^{2+}	$+2\,NH_3$	3.5
$[Cu(NH_3)_3]^{2+}$	Cu^{2+}	$+3\,NH_3$	2.9
$[Cu(NH_3)_4]^{2+}$	Cu^{2+}	$+4\,NH_3$	2.1
$[Cu(NH_3)_5]^{2+}$	Cu^{2+}	$+5\,NH_3$	−0.5
$[CuCl_4]^{2-}$	$[CuCl_3]^-$	$+Cl^-$	0.01
$[CuCl_3]^-$	$CuCl_2$	$+Cl^-$	0.06
$CuCl_2$	$CuCl^+$	$+Cl^-$	0.4
$CuCl^+$	Cu^{2+}	$+Cl^-$	1
$[Cu(OH)_4]^{2-}$	$[Cu(OH)_3]^-$	$+OH^-$	0.9
$[Cu(OH)_3]^-$	$Cu(OH)_2$	$+OH^-$	−5

diammine salt solutions are used in the absorption of olefins and carbon monoxide. Many copper circuit boards are etched with ammoniacal ammonium salt solutions; copper ammine solutions, which can be used in the production of copper compounds, are byproducts of this process.

Copper Chloride Complexes. Copper(I) and copper(II) ions form complexes with hydrochloric acid or soluble metal chlorides. Dilute solutions of copper(II) chloride are blue. As the chloride concentration increases, the color of the solution shifts toward green, intensifying as the concentration of the distorted tetrahedral $[CuCl_4]^{2-}$ ion increases. Copper(II) compounds have been isolated that contain either the $[CuCl_4]^{2-}$ anion or the $[CuCl_3]^-$ anion, e.g., $CsCuCl_4$ and $KCuCl_3$, respectively.

The insoluble copper(I) chloride can be solubilized by chloride-containing solutions, with the formation of the $[CuCl_2]^-$ ion. The benzene-insoluble copper(I) chloride dissolves readily in the presence of anhydrous aluminum chloride, with formation of the $Al[CuCl_4]$ complex. In acetonitrile, copper(I) chloride is solubilized with the formation of the isolable $[Cu(CH_3CN)_4]Cl$ complex.

Copper(I) Cyanide Complexes. Alkali cyanides readily dissolve copper(I) cyanide, with the formation of $[Cu(CN)_2]^-$ and $[Cu(CN)_3]^{2-}$ complexes. Solutions of the complexes are of great importance in the electroplating industry because of their versatility. From saturated copper and sodium cyanide baths, crystals of the double salt $Na_2Cu(CN)_3 \cdot 2\,H_2O$ can be obtained readily.

Other Copper Complexes of Industrial Significance. Solvent extraction of copper(II) ions relies on the ability of copper to form complexes and to break those complexes as a function of pH [110], [126], [127]. The most common reagents used commercially are substituted salicylaldoximes, 1,8-hydroxyquinolines, and α-hydroxyoximes.

Copper(II) is solubilized in alkaline media by complex formation with tartrates for use as Fehling's solution, which is utilized in the analysis of reducing sugars. Copper(II) bis(1,8-dihydroxyquinoline) is used as a textile fungicide and as a pigment. Copper phthalocyanines [128], [129] are exceptionally stable blue and green pigments, as are the azo dye complexes of copper (→ Azo Dyes; → Phthalocyanines).

6. Copper Reclamation

The reclamation or recycling of copper wastes and byproducts is environmentally imperative and, in many cases, economically sound. Better technologies have been developed to produce useful materials from wastes and byproducts,

many of which are now considered hazardous wastes. The establishment of the Resource Conservation and Recovery Act (RCRA) by the EPA defined procedures to assure proper disposal or reclamation of hazardous wastes [130]. However, no distinction is made between waste buried in an approved landfill and waste recovered as useful products. Unfortunately, in many cases this encourages the burial of recyclable waste materials.

The largest sources of reclaimable byproduct copper are the electronics and plating industries, as well as spent electrowinning baths [113]. Electrowinning baths are most often crystallized (Section 4.4.1) or precipitated as tribasic copper(II) sulfate for use in agriculture.

In the manufacture of circuit boards, a copper laminate is selectively dissolved (etched). This dissolution process yields the desired circuit board and a byproduct copper solution, the spent etchant. The earliest etching solutions used iron(III) chloride – hydrogen chloride or chromium trioxide – sulfuric acid solutions which were difficult to recycle. Later, the sulfuric acid etchants that used hydrogen peroxide [131], [132] or persulfate [133–135] as oxidants were developed. Recycling of the copper is achieved through crystallization, precipitation [135], ion exchange [136], or cementation [137], [138]. Alkaline etchants utilized a mixture of ammonia and ammonium sulfate with persulfate as the oxidant [139] or a mixture of ammonia and ammonium chloride with chlorite as the oxidant [140]. The primary etchants used today are the ammonia – ammonium chloride system with air as the oxidant and the hydrochloric acid system with stabilized hydrogen peroxide as the oxidant [141].

Copper is recovered from the acid chloride etchants (1) by neutralization and precipitation of copper(II) oxychloride or copper(II) oxide [42]; (2) by reduction with sodium hydrogensulfite, for example, to precipitate copper(I) chloride which can then be converted to copper(I) oxide by sodium hydroxide treatment [142]; or (3) by cementation with scrap iron [143]. From the alkaline etchants, copper is recovered (1) by distillation to strip ammonia and precipitate copper(II) oxide [144]; (2) by addition of base to precipitate copper(II) hydroxide, followed by steam stripping of the ammonia [145]; (3) by reduction to copper with formaldehyde [146] or $Na_2S_2O_4$ [147]; or (4) by electrolysis [148].

One of the more elegant methods for the recovery of copper from ammonia – ammonium chloride etchants utilizes solvent extraction [149] (Fig. 8). Not only is the copper recovered, but the etchant is simultaneously regenerated:

$$[Cu(NH_3)_4]Cl_2\,(aq) + 2\,RH \longrightarrow \underbrace{CuR_2}_{\text{Fresh etchant}} + 2\,NH_3\,(aq) + 2\,NH_4Cl\,(aq)$$

$$CuR_2 + H_2SO_4\,(aq) \longrightarrow 2\,RH + \underbrace{CuSO_4}_{\text{Reclaimed copper}}\,(aq)$$

where R is an α-hydroxyoxime as shown in page 14.

The copper can be further reclaimed by crystallization or electrowinning.

7. Copper and the Environment

The reclamation of copper byproducts and wastes and the regulation of plant effluents and sanitary waters have resulted in lower indiscriminate release of copper into natural waters [150]. Where large quantities of copper in high concentration exist (see Chap. 6), recovery of the copper can, in many cases, be economical. Regulations have been established to control or limit altogether the introduction of copper into waters [151–155]. The Federal Water Pollution Control Act Amendments of 1972 set up a comprehensive program to "restore and maintain the chemical, physical, and biological integrity of the nation's waters" [151].

Limitations of copper in drinking water have also been established: the World Health Organization recommends 0.05 – 1.5 mg/L [156]. No specific drinking water standards for copper have been established by the Environmental Protection Agency [157].

A variety of technologies presently exist for the removal of copper from wastewaters and drinking water [158], [159]. The most common treatment methods include direct precipitation [151–153], [160–163] or electrolysis [164], [165].

Although copper is toxic in exceedingly low concentrations to certain lower life forms, notably fungi and algae, it is a necessary con-

stituent of higher plants and animals (see, however, Chap. 9). Copper plays a necessary role as an oxidation catalyst and oxygen carrier, probably second only to iron in importance [166]. Copper aids plants in photosynthesis and other oxidative processes. In higher animals, it is responsible for oxidative processes and is present in many proteins such as phenolase, hemocyanin, galactose oxidase, superoxide dismutase, ceruloplasmin, tyrosinase, monoamine oxidase, and dopamine β-hydroxylase [167], [168].

Copper compounds are regularly applied for their nutrient value to agricultural crops in European countries [169–171]. In the United States, more emphasis is now being placed on the value of copper as a nutritional feed additive [169], [172], [173]: it increases the rate of gain and feed efficiencies of the animals. The increased use of copper as a feed additive has caused concern about the environmental impact of high levels of copper in manure [174], [175]. The approved maximum total copper content in animal feeds for Western European countries is shown in Table 9 [169].

Table 9. Maximum copper levels in Western European animal feeds

Use	Total copper, mg/kg
Fattening pigs, up to 16 weeks	175
Fattening pigs, 17th week to 6th months	100
Fattening pigs, over 6 months	50
Breeding pigs	50
Calves, fed milk-based products	30
Calves, other feeds	50
Sheep	20
Other animals	50

The largest single application of copper compounds is as an agricultural fungicide (see Chap. 8). Although copper is a necessary constituent of higher plants and animals, it is highly toxic to certain fungi. In human medicine, the importance of copper as a nutrient in proper development and growth is receiving increased study [167], [176]. Also, copper is being used in the production of new copper antibiotics and copper anti-inflammatory drugs [176]. The recommended daily requirement of copper is as follows [176]:

Humans	2.5 – 5.0 mg
Cattle	50 – 70 mg
Horses	50 – 60 mg
Sheep	10 – 15 mg
Swine	10 – 20 mg

8. Economic Aspects

The world's production of copper compounds represents less than 2 % of the total production of primary metal [65]. In 1974, 7.65×10^6 t of primary copper was produced, with 1.3 % (100 000 t) used in the production of copper compounds. Of these 100 000 t, 80 % is used to prepare copper(II) sulfate pentahydrate.

Historically, the United States and Italy have led the world in producing copper sulfate pentahydrate (Fig. 9) [65]. The U.S. production and valuation of copper sulfate pentahydrate are given in Table 10 [177].

Figure 9. Production of copper(II) sulfate pentahydrate in Italy and the United States (1950 – 1976)

Table 10. Production and valuation of copper(II) sulfate pentahydrate in the United States

Year	Production		Price of copper(II)
	Quantity, t	Copper content, t	sulfate pentahydrate, U.S. $/kg
1978	31 880	8551	1.04
1979	35 005	9286	1.18
1980	31.010	8445	1.23
1981	35 640	9413	1.14
1982	32 230	8385	1.08

Table 11. Annual use of copper in agriculture [169]

Country	Total, t	Fungicide, %	Animal feed, %	Crop nutrient, %
Europe				
France	5 380	83.6	14.9	0.9
Federal Republic of Germany	2 800	35.7	28.6	35.7
Greece	1 410	97.9	1.4	0.7
Italy	6 220	96.5	3.2	0.3
Portugal	1 120	89.3	8.9	1.8
Spain	2 800	89.3	10.7	0.0
Others	15 563	49.8	34.5	15.8
Asia and Australia				
Australia	1 556	19.3	6.4	74.3
India	3 500	0.0	0.0	100.0
Japan	2 040	98.0	1.2	0.7
Others	1 811	78.7	0.6	20.3
Africa				
Algeria	1 750	100.0	0.0	0.0
Kenya	1 000	100.0	0.0	0.0
Tanzania	2 000	100.0	0.0	0.0
Others	2 635	96.4	0.6	3.0
America				
Brazil	7 160	97.8	1.4	0.8
Mexico	1 650	100.0	0.0	0.0
United States	6 400	46.9	9.4	43.8
Others	3 540	40.7	2.8	56.6
Total	70 305	76.4	12.1	11.4

According to estimates about 70 % (70 000 t) of the total world's production of copper compounds is used in agriculture [169]. Table 11 gives a breakdown by region, country, and use. Copper sulfate pentahydrate can be blended directly with feed or fertilizers for animal feeds or crop nutrients, respectively. For application as a foliar fungicide, it must be rendered insoluble to prevent it from being easily washed off the leaf. The first copper fungicide on the market was Bordeaux mixture (1882), a combination of lime and copper sulfate [178]. The inherent phytotoxicity of Bordeaux mixture led to the development of the "fixed coppers" in the 1920s and 1930s. The fixed coppers are grouped into four categories: (1) basic sulfates, (2) basic chlorides, (3) oxides, and (4) miscellaneous – silicates, phosphates, etc. [179], [180]. To enhance their performance, the fixed coppers are usually blended with wetting agents, dispersants, sticking agents, and diluents. Since better coverage and enhanced efficacy are obtained with smaller particles, the basic coppers are usually ground before packaging [181]. Table 12 (see next page) gives a partial list of available fixed copper fungicides.

9. Toxicology and Occupational Health

Copper is an essential trace element in humans. In larger quantities it can be lethal. The salts are usually considered to be more toxic than the metal. The acute oral toxicity in humans, LD_{Lo}, is about 100 mg/kg; however, recovery has occurred after ingestion of up to 600 mg/kg. The symptoms of copper poisoning include nausea, vomiting, gastric disturbances, apathy, anemia, cramps, convulsions, coma, and death.

Many reported cases of illness were once attributed to chronic copper poisoning, but they are

Table 12. Commercially available fixed coppers

Manufacturer/representative	Trade name	Active ingredient	Cu content, %
BASF	BASF-Grünkupfer	copper oxychloride	45
	Kauritil	copper oxychloride	47
Bayer	Cupravit-Spezial	copper oxychloride	45
	Cupravit-Forte	copper oxychloride	50
CP Chemicals	Basic Copper Sulfate	tribasic copper sulfate	53
	Champion WP	copper hydroxide	50
Cuproquim	Hydrox	copper hydroxide	50
Hoechst	Vitigran, Conc.	copper oxychloride	45
Kocide/Griffin	Kocide 101	copper hydroxide	50
Merck	Perenox	copper(I) oxide	50
Norddeutsche-Aff.	Cobre Nordox	copper(I) oxide	50
Phelps-Dodge	Tribasic Copper Sulfate	tribasic copper sulfate	53
Sandoz	Cobre Sandoz	copper(I) oxide	50
Simplot	Blue Shield	copper hydroxide	50
Spiess Urania	Funguran	copper oxychloride	45
	Cuprasol	copper oxychloride	50
Tennessee Copper	Tribasic Copper Sulfate	tribasic copper sulfate	53
Wacker	Wacker-Kupferkalk	copper oxychloride	15 – 18

now thought to be caused by impurities in the refining of copper, e.g., lead, arsenic, and selenium [182], [183]. The question of chronic poisoning is still open to debate. Prolonged exposure to copper dust can cause skin irritation and discoloration of skin and hair. Whether pathological changes occur is uncertain, although there is evidence of accumulation of copper in the liver. Attempts to induce chronic copper poisoning in animals have been unsuccessful.

The inhalation of dusts of copper compounds irritates the upper respiratory tract [184], [185]. Ulceration and perforation of the nasal septum have occurred. Workers exposed to dusts of copper salts complained of metallic tastes and irritation of the oral and nasal mucosa. Smokers complained of an intense sweet taste during inhalation of the smoke. Long-term exposure to copper-containing dust has resulted in atrophic changes in the mucous membranes. Mild nasal discomfort was noted in workplace concentrations as low as 0.08 mg/m^3. A very small fraction of the workers exhibited allergic skin reactions from exposure to copper-containing dust [186], [187].

Copper compounds embedded in the eye produce a pronounced foreign body reaction with discoloration of the ocular tissue. Conjunctivitis, ulceration, and turbidity have been reported [188], [189].

The workplace standard for copper dust concentration in air, MAK and TLV, is 1 mg/m^3 [190–192]. Copper fume standards are 0.2 mg/m^3 [190] and 0.1 mg/m^3 [191], [192]. A number of reviews and books of a more specific nature report detailed animal and workplace studies [193–197].

10. References

1. *Gmelin,*System No. 60 “Kupfer,” Teil A (1955), Teil B (1958 – 66), Teil C (1978), Teil D (1963).
2. P. Pascal in *Nouveau Traité de Chimie Minérale,* **vol. 3,** Masson et cie, Paris 1957, pp. 155 – 421.
3. J. Mellor: *A Comprehensive Treatise on Inorg. and Theoretical Chemistry,* **vol. 3,** Longman’s and Green Co., New York 1923, pp. 49 – 294.
4. A. Butts (ed.): *Copper – The Science and Technology of the Metal, its Alloys and Compounds,ACS Monograph 122,* Reinhold Publ. Co., New York 1954.
5. W. Hatfield, R. Whyman in R. Carlin (ed.): *Transition Metal Chemistry,* **vol. 5,** Marcel Dekker, New York 1969, pp. 47 – 179.
6. M. Kato, H. Jonassen, J. Fanning, *Chem. Rev.* **64** (1964) 99.
7. B. Hathaway, D. Billig, *Coord. Chem. Rev.* **5** (1970) 1 – 43.
8. H. Sigel, *Angew. Chem. Int. Ed. Engl.* **13** (1974) 394.
9. R. Doedens, *Prog. Inorg. Chem.* **21** (1976) 209.
10. M. Bruce, *J. Organomet. Chem.* **44** (1972) 209.
11. F. Jardine, *Adv. Inorg. Chem. Radiochem.* **17** (1975) 116.

12. C. K. Williams Co., US 2 554 319, 1953 (J. Ayers).
13. Rohm and Haas, US 2 184 617, 1939 (L. Hurd).
14. Rohm and Haas, US 2 273 708, 1942 (L. Hurd).
15. K. Hauffe, P. Kofstad, *Z. Elektrochem.* **59** (1955) 399.
16. Merck, US 2 409 413, 1951 (H. Becker).
17. Calumet and Hecla, US 3 466 143, 1969 (H. Day).
18. Glidden, US 2 758 014, 1956 (J. Drapeau, P. Johnson).
19. Glidden, US 2 891 842, 1959 (J. Drapeau, P. Johnson).
20. Norddeutsche Aff., GB 772 846, 1957; DE 1 020 010, 1957 (E. Klumpp).
21. S. Mahalla, US 3 492 115, 1970.
22. Lake Chem. Co., US 2 474 497, 1949 (P. Rowe).
23. G. E. Co., US 3 186 833, 1965 (R. Cech).
24. Lake Chem. Co., US 2 474 533, 1949 (L. Klein).
25. Glidden, US 2 817 579, 1957 (J. Drapeau, P. Johnson).
26. Tennessee Corp., US 3 457 035, 1969 (J. Barker).
27. USA, Secretary of the Interior, US 3 716 615, 1973 (D. Bauer, P. Haskett, R. Lindstrom).
28. USA, Secretary of the Interior, US 3 833 717, 1974 (P. Haskett, D. Bauer, R. Lindstrom, C. Elges).
29. S. Titova, V. Golodov, D. Sokol'skii, SU 488 788, 1975; *Chem. Abstr.* **84** (1976) 124 057 c.
30. Mountain Copper, US 2 665 192, 1954 (P. Rowe).
31. Mountain Copper, US 2 977 195, 1961 (C. Matzinger).
32. Nippon Chem., JP-Kokai 80/71 629, 1980.
33. ICI, GB 936 922, 1963 (A. Campbell, A. Taylor).
34. Nippon Chem., JP-Kokai 78/133 775, 1978; *Chem. Abstr.* **93** (1981) 116 652 p.
35. ASTM D283-52: (reapproved 1978) Philadelphia, pp. 89 – 94.
36. Military Specification: *Pigment, Cuprous Oxide* MIL-P-15169B (Ships), 1963; Amendment – 1 MIL-P-15169B (Ships), 1969.
37. Metallgesellschaft, DE 1 007 307, 1957 (J. Dornauf).
38. W. Kunda, H. Veltman, D. Evans, *Copper Met. Proc. Extr. Met. Div. Symp.* 1970, 27 – 69.
39. Fluor Corp., US 2 927 018, 1960 (C. Redemann).
40. Fluor Corp., US 2 923 618, 1960 (C. Redemann, H. Tschimer).
41. Sherritt Gordon Mines, US 3 127 264, 1964 (H. Tschimer, L. Williams).
42. Ruehl, Erich, Chem. Fabrik, DE 3 115 436, 1982 (W. Jagusch, H. Reichelt).
43. Jap. Bur. Ind. Tech., DE 2 047 372, 1971 (Y. Moriya).
44. Kachita Co., DE 2 028 791, 1970 (T. Tamura, A. Sakamoto, T. Kato).
45. TDK Electronics, JP-Kokai 75/131 668, 1975 (S. Adachi, T. Miyakoshi, M. Hattori).
46. Fordwerke, DE 2 046 180, 1971 (J. Jones, E. Weaver).
47. Mitsui Toatsu Chem., JP-Kokai 74/72 179, 1974 (T. Kobayagawa, Y. Nakajima).
48. Oxy. Catalyst, FR 1 402 087, 1965.
49. Lonza, DE 2 012 250, 1970 (T. Voelker, K. Hering).
50. Philips Gloelampen Fabrik., DE 2 103 538, 1971 (C. Esveldt, N. Slijkerman).
51. C. Freedenthal: *Copper Compounds, Encyclopedia of Industrial Chemical Analysis,* J. Wiley & Sons, New York 1970, pp. 651 – 680.
52. Cellocilk Co., US 1 800 828, 1931 (W. Furness).
53. D. Marsh, B. Marsh, US 2 104 754, 1938.
54. Mountain Copper, US 2 536 096, 1951 (P. Rowe).
55. Lake Chem. Co., US 2 525 242, 1950 (P. Rowe).
56. H. Neville, C. Oswald, *J. Phys. Chem.* **35** (1931) 60 – 72.
57. Copper Research, US 24 324, 1957 (W. Furness).
58. Rohm and Haas, US 3 231 464, 1966 (E. Dettwiler, J. Filliettaz).
59. Cuproquim S.A., US 4 418 056, 1983 (M. Gonzalez).
60. Giulini Aldolfomer Ind., BR 83/01 912, 1983 (J. Giulini, A. Meyer).
61. Kennecott Copper, DE 1 592 441, 1965 (W. Furness); US 3 194 749, 1965.
62. Hoechst, DE 824 199, 1951 (K. Pfaff, A. Voigt).
63. Hoechst, DE 824 200, 1951 (K. Pfaff, A. Voigt).
64. E. Gimsey, D. Muir, A. Parker, *Proc. Australas. Inst. Min. Metall.* **273** (1980) 21 – 25.
65. *Ullmann,* 4th ed., **15,** 564.
66. Degussa, DE 1 000 361, 1955 (F. Bittner).
67. Degussa, FR 2 009 852, 1969.
68. Norddeutsche Affinerie, DE 1 813 891, 1958.

69. Norddeutsche Affinerie, US 3 679 359, 1972 (E. Haberland, W. Perkow).
70. Mitsubishi Metal Corp., JP-Kokai 80/60 003, 1980.
71. R. Keller, H. Wycoff, *Inorg. Synth.* **2** (1946) 1 – 4.
72. Schering, DE 1 080 088, 1958 (H. Niemann, K. Herrmann).
73. Harshaw, US 2 367 153, 1945 (C. Swinehart).
74. Goldschmidt, DE 3 305 545, 1983 (E. Mack, L. Witzke).
75. Cyprus Metallurg., DE 2 607 299, 1976 (D. Goens, P. Kruesi).
76. Cyprus Metallurg., US 3 972 711, 1976 (D. Goens, P. Kruesi).
77. Metallgesellschaft, DE 1 174 996, 1963 (K. Meyer, H. Ransch, H. Pietsch); DE 1 180 946, 1963; DE 1 160 622, 1963.
78. Duisburger Kupferhütte, US 3 353 950, 1967 (H. Junghauss); DE 1 200 545, 1965.
79. Nitto Elec. Ind., JP-Kokai 78/65 348, 1978.
80. Jap. Pure Hydrogen Co., JP-Kokai 70/48 693, 1970 (T. Yamamoto).
81. Phillips Petro. Co., US 3 658 463, 1972 (W. Billigs).
82. Tenneco Chem. DE 2 414 800, 1974 (R. Turnbo, D. Keyworth).
83. H. Allgood, *Fert. Sci. Technol. Ser.* **2** (1974) 289 – 309.
84. DuPont, US 3 230 231, 1966 (H. Burtolo, J. Braun, C. Winter).
85. DuPont, US 4 035 383, 1977 (R. Sweet).
86. Anic S.p.A., DE 2 743 690, 1978 (U. Romano, R. Tessi, G. Cipriani, L. Micucci); US 4 218 391, 1980.
87. Ube Industries, JP-Kokai 79/106 429, 1979 (H. Itatani, S. Dano).
88. Anic S.p.A., DE 3 045 767, 1981 (U. Romano, F. Rivetti, N. Muzio); US 4 318 862, 1983.
89. As reported in A. Siedell (ed.): *Solubilities of Inorganic and Metal Organic Compounds,* **vol. I,** D. Van Nostrand, New York 1940, p. 478.
90. Sudhir Chem. Co., US 3 202 478, 1965 (A. Hindle, S. Raval, S. Damani, H. Damani, K. Damani); GB 912 125, 1962.
91. N. V. Koninklijke Ned. Zoutindustrie, FR 957 457, 1947.
92. Hoechst, DE 1 931 393, 1969 (H. Kreckler, H. Kuckertz).
93. Distillers Co., BE 616 762, 1962; GB 932 130, 1962.
94. W. Feitknecht, K. Maget, *Helv. Chim. Acta* **32** (1949) 1639 – 1653.
95. Wacker, DE 1 083 794, 1960 (H. Baumgartner).
96. Bayer, DE 1 159 914, 1963 (E. Podschus).
97. Instytut Przemyslu Org., PL 55 953, 1968.
98. Bayer, DE 1 161 871, 1963 (E. Podschus, W. Joseph); US 3 230 035, 1965.
99. Ciba, Ltd., CH 243 271, 1946.
100. Sirco, A.G., CH 256 414, 1947.
101. Fisons Pest Control, US 2 907 691, 1959 (G. Hartley, P. Park).
102. Fahlberg-List GmbH, DE 1 920 963, 1970 (G. Guenter, F. Schuelde, K. Schoeyen).
103. Riedel-de-Haen, DE 1 142 724, 1963.
104. P. Ramos, E. Esteban, V. Callao, *Ars. Pharm.* **12** (1971) 471 – 477.
105. W. Feitknecht, K. Maget, *Helv. Chim. Acta* **32** (1949) 1639 – 1653.
106. As reported in W. Linke (ed.): *Solubilities of Inorganic and Metal Org. Compounds,* 4th ed., **vol. I,** Amer. Chem. Soc., Washington, D.C. 1958, pp. 965 – 968.
107. H. Moyer, *AIME Annual Meeting,* New Orleans, Feb. 18 – 22, 1979.
108. Tennessee Copper, US 2 533 245, 1950 (G. Harike).
109. C. Merigold, D. Agers, J. House, *Int. Solvent Extr. Conf.* 1971.
110. General Mills, US 3 224 873 (Swanson).
111. Dow, US 4 031 038, 1977 (R. Grinstead); US 4 098 867, 1978.
112. S. Tataru, RO 67 946, 1979.
113. F. Lowenstein, M. Moran in *Faith, Keyes and Clark's Industrial Chemicals,* 4th ed., John Wiley and Sons, New York 1975, pp. 280 – 283.
114. Hitachi Chem., JP-Kokai 76/03 396, 1976 (T. Senda, T. Umehara); JP-Kokai 76/03 395, 1976.
115. L. Markov, K. Balarev, *Izv. Khim.* **15** (1982) 472 – 481.
116. H. Weiser, W. Milligan, E. Cook, *J. Am. Chem. Soc.* **64** (1942) 503 – 508.
117. A. Binder, *Compt. Rend.* **198** (1934) 653, 2167.
118. Cities Service, US 3 725 535, 1973 (J. Barker).
119. Vereinigte Metallwerke, DE 2 701 253, 1977 (J. Bertha).
120. Compagnie de Saint Gobain, FR 1 302 445, 1962 (M. Provost).
121. Takeda Chem. Ind., FR 1 452 608, 1966.
122. INCRA, US 3 846 545, 1970 (S. Foschi); FR 2 001 428, 1969 (E. Hess, D. Kennedy, C. Whitham).
123. J. Horsfall, J. Heuberger, *Phytopathology* **32** (1942) 226 – 232.

124. J. Horsfall: *Fungicides and their Action,* Chronica Botanica Co., Waltham, Mass. 1945.
125. Duisburger Kupferhütte, DE 2 602 448, 1977 (W. Roever, K. Lippert).
126. A. Biswas, W. Davenport: *Extractive Metallurgy of Copper,* Pergamon Press, Oxford – New York 1976.
127. G. Ritcey, A. Ashbrook: *Solvent Extraction,* Part II, Elsevier, Amsterdam 1979, pp. 197 – 248.
128. DuPont, US 3 230 231, 1966 (H. Bartolo, J. Braun, C. Winter).
129. DuPont, US 4 035 383, 1977 (R. Sweet).
130. USEPA, *Environmental Protecting Agency,* Code of Federal Regulation 40CFR, parts 240 – 299, 1983.
131. Mitsubishi, US 3 597 290, 1971 (A. Naito, Y. Masuda, S. Osawa).
132. Dart Ind., US 4 158 593, 1979 (J. Allan, P. Readio).
133. Denka, US 3 770 530, 1973 (O. Fujimoto).
134. Denka, US 3 936 332, 1976 (A. Matsumoto, K. Itani); US 3 939 089, 1976; Tokai Electro-Chem., US 3 839 534, 1974 (A. Matsumoto, O. Fujimoto).
135. FMC, GB 873 583, 1959 (P. Margulies, J. Kressbach, H. Wehrfritz); DE 1 160 270, 1959.
136. K. Jones, R. Pyper, *J. Met.* **31** (1979) no. 4, 19 – 25.
137. Micro Copper Corp., US 3 679 399, 1972 (D. Linton, A. Zinkl, J. Ballam).
138. Showa Electric Wire and Cable, JP-Kokai 75/84 415, 1975 (Y. Kogane).
139. FMC, US 3 837 945, 1974 (J. Chiang).
140. FMC, US 3 844 857, 1974 (J. Chiang).
141. P. Fintschenko, *Met. Finish. J.* **20** (1974) 138 – 139.
142. Nippon Chem. Ind., JP-Kokai 83/140 321, 1983.
143. A. Celi, DE 3 208 609, 1983.
144. D. Klein, DE 3 204 815, 1983.
145. J. Leroy, Contribution to International Recycling Congress (IRC), Berlin 1982, 592 – 595.
146. Western Electric, US 4 428 773, 1984 (K. Krotz).
147. Hitachi Cable, JP-Kokai 80/159 895, 1980.
148. Western Electric, US 3 843 504, 1974 (B. Nayder).
149. Criterion, US 4 083 758, 1978 (W. Hamby, M. Slade).
150. F. Harrison, *Gov. Rep. Announce. Index (U.S.)* **84** (1984) 42 – 139.
151. USEPA: *Development Document for Effluent Limitations Guidelines and Standards for the Inorganic Chemicals Point Source Category,* Phase II, 440/1-83/007-6 (1983).
152. USEPA: Nonferrous metals manufacturing point source category; *effluent limitations guidelines, pretreatment standards, and new source performance standards, Fed. Regist.* **49** (1984) 8742 – 8831.
153. USEPA: *Clean water;* Inorganic chemicals manufacturing point source category, effluent limitations, guidelines, pretreatment standards and new source performance standards, Fed. Regist. 49 (1984) 33 402 – 33 429.
154. *Wasserhaushaltsgesetz* – WHG – vom 27. 7. 1957 in der Fassung vom 26. 4. 1976, BGBl. I (1976).
155. EG – *Gewässerschutzrichtlinie* vom 4. 5. 1976 – Amtsblatt der EB Nr. L 129/23 vom 18. 5. 1976, hier Anhang Stoffliste II u. a. mit Kupfer.
156. WHO: *International Standards for Drinking Water,* Monograph Series No. 3 (1971) 40.
157. USEPA: "Interim Primary Drinking Water Standards," *Fed. Regist.* **40** (1975) 11 994.
158. J. Hallowell, J. Shea, G. Smithson, A. Tripler, B. Gonser, *Gov. Rep. Announce (U.S.)* **74** (1974) 89 – 161.
159. F. Pierrot, D. Vanel, *Galvano-Organo Trait. Surf.* **52** (1983) 726 – 728.
160. C. Roy, US 3 816 306, 1974.
161. Hoellmueller Maschinenbau, DE 2 257 364, 1974 (E. Lendle, G. Wartenberg).
162. BASF, DE 2 255 402, 1974 (H. Hiller, A. Schuhmacher, H. Schmidt).
163. M. Manafont, ES 397 046, 1974.
164. Nippon Electric, JP-Kokai 84/22 694, 1984.
165. R. Spearot, J. Peck, *Environ. Prog.* **3** (1984) 124 – 128.
166. D. Kertesz, R. Zito, F. Ghiretti in O. Hayaishi (ed.): *Oxygenases,* Academic Press, London 1962.
167. Ciba Foundation Symposium: *Biological Roles of Copper,* no. 79, Excerpta Medica and Elsevier-North Holland, Amsterdam 1980, p. 343.
168. H. Sigel (ed.): *Metal Ions in Biological Systems,* **vol. 12,** Properties of Copper, Marcel Dekker 1981, p. 384.
169. INCRA: V. Shorrocks, Planning Study, *Use of Copper as a Micronutrient for Crops,* Part I, New York, July 1984.
170. W. Ernst, *Schwermetallvegetation der Erde,* G. Fischer Verlag, Stuttgart 1974.
171. *Micronutrients in Agriculture,* Soil Sci. Soc. of Amer. Inc., Madison, Wisc. 1972.

172. G. Cromwell, T. Stahly, W. Williams, *Feedstuffs* **53** (1982) 30, 32, 35 – 36.
173. L. DeGoey, R. Wahlstrom, R. Emerick, *J. Anim. Sci.,* **33** (1971) 52 – 57.
174. INCRA: *Environmental Impact Analysis Report,* Copper Salts in Animal Feed, Int. Copper Res. Assoc. Inc., New York – London 1974.
175. USEPA: "Registration on the Level of Copper in Animal Feed," *Fed. Regist.* **38** (1973) 178.
176. INCRA: *A Critical Review of Copper in Medicine, Report No. 234,* Int. Copper Res. Assoc., Inc., Göteborg – New York 1975.
177. U.S. Dept. of Interior: *Minerals Yearbook 1982,* **vol. I,** U.S. Govt. Printing Office, Washington, D.C., 1983, p. 209.
178. G. Johnson, *Agricult. History* **9** (1935) 67 – 79.
179. D. Frear in *Chemistry of Insecticides, Fungicides, and Herbicides,* 2nd ed., D. Van Nostrand, New York 1948, pp. 211 – 228.
180. E. R. de Ong in *Chemistry and Uses of Pesticides,* 2nd ed., Reinhold Publ. Co., New York 1956, pp. 31 – 47.
181. S. Gunther, *Chem. Ing. Tech.* **31** (1959) 731 – 734.
182. E. Browning: *Toxicity of Industrial Metals,* 2nd ed., Butterworths, London 1969.
183. R. Fabre, R. Truhaut: *Précis de Toxicologie,* Centre Doc. Univ., Paris 1960.
184. A. Askergren, M. Mellgren, Scand. J. Work, *Environ. Health* **1** (1975) 45 – 49.
185. R. Gleason, *Am. Ind. Hyg. Assoc. J.* **29** (1968) 461 – 462.
186. Y. Kambe, Matsushima, T. Kuroume, S. Suzuki, *Proc. Fourth Int. Congr. Rural Med.* Zürich 1970, 55 – 57.
187. E. Saltzer, J. Wilson, *Arch. Dermatol.* **98** (1968) 375 – 376.
188. *Documentation of Threshold Limit Values for Substances in Workroom Air,* 3rd ed., Cincinnati 1974.
189. W. Grant: *Toxicology of the Eye,* 2nd ed., C. C. Thomas, Springfield, Il. 1974.
190. NIOSH: *Manual of Industrial Chemicals,* U.S. Dept. of Labor, September 1978.
191. D. Henschler: *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* Verlag Chemie, Weinheim 1972 /77.
192. DFG: *Maximale Arbeitsplatzkonzentration,* Verlag H. Boldt, Boppard 1976.
193. H. E. Stokinger in *Patty*, vol. 2 A, pp. 1620 – 1630 (1981).
194. S. Cohen, *JOM, J. Occup. Med.* **16** (1974) 621 – 624.
195. Int. Labor Office: *Encyclopedia of Occupational Health and Safety,* McGraw-Hill, New York 1971.
196. N. Sax: *Dangerous Properties of Ind. Materials,* 4th ed., Van Nostrand, New York 1982.
197. W. Dreichmann, H. Gerarde: *Toxicology of Drugs and Chemicals,* Academic Press, New York 1969.

Combinatorial Chemistry

Dominika Tiebes, DECHEMA e.V., Frankfurt am Main, Germany

Stefan Peters, Boehringer Ingelheim Pharma KG, Ingelheim, Germany

1. Introduction 2
2. Concept of Combinatorial Chemistry 3
3. Methods and Techniques of Combinatorial Synthesis 4
3.1. Synthetic Strategies Towards Combinatorial Libraries 4
3.1.1. Split – Pool Synthesis Towards Combinatorial Libraries 4
3.1.1.1. Techniques using Resin Beads . . 5
3.1.1.2. Tea-Bag Method 5
3.1.2. Parallel Synthesis Towards Combinatorial Libraries 6
3.1.2.1. Reaction Apparatus using Resin Beads 8
3.1.2.2. Multipin Technique 8
3.1.2.3. Spatially Addressable Parallel Synthesis on Silica Wafer 8
3.1.3. Reagent Mixture Synthesis Towards Combinatorial Libraries . . 9
3.2. Synthetic Methodology for Organic Library Construction . . 9
3.2.1. Solid-Phase Organic Synthesis . . 10
3.2.1.1. Solid Support, Linker, and Cleavage Strategies 11
3.2.1.2. Reaction Portfolio 12
3.2.2. Synthesis in Solution and Liquid-Phase Synthesis 12
4. Characterization of Combinatorial Libraries 12
4.1. Analytical Characterization . . 12
4.2. Hit Identification in Combinatorial Libraries by High-Throughput Screening 13
4.2.1. Strategies for Libraries of Compound Mixtures 14
4.2.1.1. On-Bead Screening 14
4.2.1.2. Deconvolution 14
4.2.1.2.1. Iterative Deconvolution 14
4.2.1.2.2. Deconvolution by Positional Scanning 16
4.2.1.2.3. Deconvolution by Orthogonal Libraries 16
4.2.1.3. Encoding 17
4.2.1.4. Multiple Cleavable Linkers 18
4.2.2. Strategies for Libraries of Separate Single Compounds 18
4.2.3. Conclusion 19
5. Automation and Data Processing 19
5.1. Synthesis Automation and Data Processing 19
5.2. Automated Purification 21
6. Library Design and Diversity Assessment 22
6.1. Diversity Assessment for Selection of Building Blocks or Compound 22
6.2. Iterative Optimization Methods 23
7. Economic Aspects and Industrial Case Studies 23
8. Further Developments 24
8.1. Combinatorial Chemistry and Catalysis 24
8.2. Combinatorial Chemistry and Material Science 25
9. References 25

1. Introduction

Since the early 1990s combinatorial chemistry has become a major focus of research activity for pharmaceutical, agrochemical, and biotechnology companies.

The ongoing introduction of combinatorial chemistry seems to result in a fundamental change in the way how the drug discovery process for new or improved therapeutics and agrochemicals is carried out. The key issues of combinatorial chemistry are lead finding and lead optimization. A *lead* is a compound with an appreciable degree of affinity for the biological target of interest. Conventionally, drug development involved the individual synthesis and biological evaluation of hundreds to thousands of organic compounds in an attempt to enhance the biological activity, selectivity and bioavailability, while at the same time decreasing its toxicity

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a07_b01

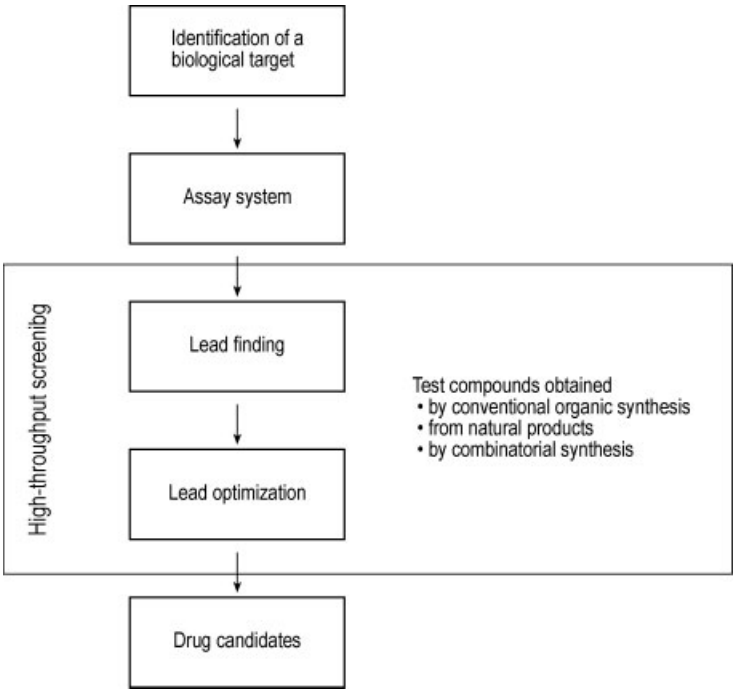

Figure 1. Drug discovery process to new or improved drug candidates in the R & D laboratories of pharmaceutical and agrochemical companies

(see Fig. 1). On an average, such a drug discovery process takes 6.5 years before a suitable drug candidate is found.

Combinatorial synthesis as a high-throughput method enables to rapidly produce hundred to thousand times more compounds than conventional serial organic synthesis in an efficient manner. This in conjunction with high-throughput screening (→ High-Throughput Screening) and the increasing number of new biological targets is expected to accelerate the process for discovering new and improved drug candidates [1–11]. Thus, combinatorial chemistry has attracted attention of companies as a means of cutting the time and high costs associated with the serial drug discovery process.

The underlying concepts used in combinatorial chemistry in the synthesis of organic compounds have their origin in the field of peptide chemistry. Pioneers such as Geysen (multipin method) [12], [13] and Houghten (teabag method) [14] developed various methods for multiple (parallel) synthesis of peptides in 1984 – 1985, based on the techniques of solid-phase peptide synthesis according to Merrifield [15] (→ Peptide Synthesis, Chap. 5.2.) that was presented first in 1963. Further progress was achieved in this field, among others by Frank (spot synthesis) [16] and Fodor (VLSIPS method) [17], [18]. However, combinatorial synthesis of peptides in the literal sense was made possible by the work of Furka (split – pool method) [19–22] in 1988, and this can be regarded as the "birth" of combinatorial chemistry. These studies have clearly demonstrated the power of combinatorial synthesis in medical chemistry. But peptides and oligonucleotides generally have low bioavailability and are rapidly proteolytically degraded and therefore their utility as therapeutic agents is often limited [23]. Due to the more favorable pharmacokinetic properties of many organic compounds (generally with a molecular mass < 750), termed small molecules, combinatorial synthesis of small molecules has rapidly become of particular importance in organic chemistry.

This overview will focus on the methods employed in design, synthesis, and screening of combinatorial libraries of small molecules and

on the interrelationship between these methods. In its optimal form combinatorial chemistry involves robotic instruments for the automated synthesis as well as for the screening of compounds, and computational methods for the library design. Furthermore, industrial case studies will be given in the field of drug discovery and other application areas such as material science and catalysis will be described.

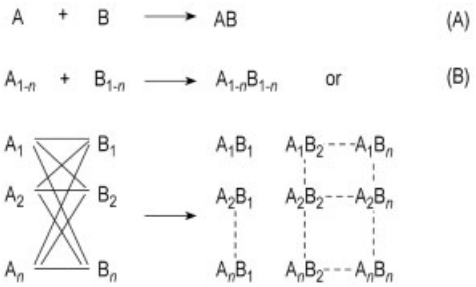

Figure 2. A) In general, in a conventional synthesis one starting material A reacts with one reagent B resulting in one product AB; B) In a combinatorial synthesis different building blocks of type A ($A_1 - A_n$) are treated simultaneously with different building blocks of type B ($B_1 - B_n$) according to combinatorial principles, i.e., each starting material A reacts separately with all reagents B resulting in a combinatorial library A_{1-n} B_{1-n}

2. Concept of Combinatorial Chemistry

The concept of combinatorial chemistry is to prepare a large number of different compounds at the same time – instead of synthesizing compound one by one, as is done conventionally – and then to identify the most promising compound for further development by high-throughput screening.

The salient feature of combinatorial synthesis is that different compounds are generated simultaneously under identical reaction conditions in a systematic manner so that ideally the products of all possible combinations of a given set of starting materials (termed building blocks) will be obtained at once (see Fig. 2). The collection of these finally synthesized compounds is referred to as a *combinatorial library* (The term "library" is not used consistently in the literature. Often this designation is reserved for compound mixtures, whereas collections of separate single substances are called "arrays".)

The combinatorial libraries can be structurally related by a central core structure, termed scaffold, or by a common backbone, as shown in Figure 3. In both cases the accessible dissimilarity of the compounds depends on the building blocks which are used for the construction of the library.

The size of the combinatorial library ranges from a few tens to many hundreds of thousands compounds. In a multistep combinatorial synthesis, generally required for drug relevant compounds, the total number of all synthesized compounds within the combinatorial library is determined by two factors: the number of building blocks used per reaction and the number of reaction steps, in which a new building block is incorporated. The number of all synthesized compounds $\underline{N}$ is given by the equation $N = b^x$, where b is an equal number of building blocks used in each reaction (1, 2, . . ., x) and x is the number of reaction steps, in which a new building block is introduced. If the number of building blocks for each reaction step varies (e.g., a, b, c in a three-step synthesis), then $N = a \cdot b \cdot c$. Thus, a large compound library will be obtained rapidly from only a few building blocks.

Depending on the synthetic strategy used the created combinatorial library can comprise compound mixtures, or separate, single substances. [The term "combinatorial synthesis" was strictly used only when different building blocks are treated in one reaction vessel, and compound mixtures are thus obtained. Since the boundary becomes blurred, herein syntheses of both libraries of compound mixtures and of separate, single substances are included, as long as the procedure can be described as combinatorial (simultaneous synthesis under identical conditions according to combinatorial principles)]. Combinatorial libraries may consist of compounds free in solution, or linked to a solid phase.

3. Methods and Techniques of Combinatorial Synthesis

3.1. Synthetic Strategies Towards Combinatorial Libraries

3.1.1. Split – Pool Synthesis Towards Combinatorial Libraries

FURKA and coworkers pioneered the split – pool synthesis method [19–22] for the synthesis of

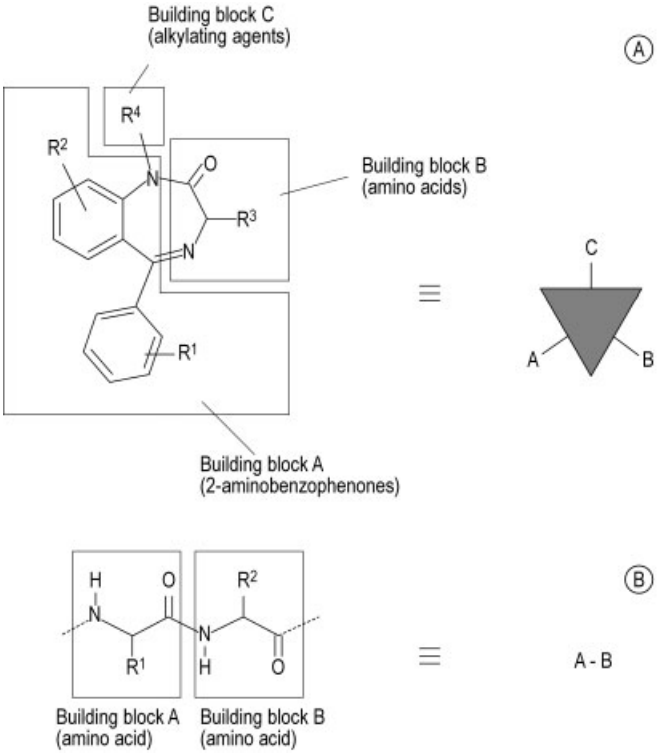

Figure 3. Scaffold-based (A) or backbone-based (B) combinatorial libraries

large peptide libraries in 1988; this approach is termed divide, couple, and recombine synthesis by other workers [24–27].

The split – pool synthesis procedure towards combinatorial libraries, as schematically illustrated in Figure 4, involves the splitting of a quantity of resin beads (a swollen insoluble solid support, see Section 3.2.1.1) into multiple equally sized portions in separate reaction vessels, coupling a single building block to each aliquot and then pooling the resin-bound compounds from all reaction vessels together in one vessel. In this vessel the common steps such as resin washing and deprotection are performed in one batch. By reapportioning the resin-bound compounds into the requisite number of separate reaction vessels the second solid-phase reaction provides compounds that incorporate all of the possible combinations of the two sets of building blocks. These split and pool operations are repeated until the desired combinatorial library has been assembled. Currently this is the most popular method for the synthesis of combinatorial libraries of compound mixtures comprising thousands to hundred thousands of compounds.

Combinatorial libraries resulting from split – pool synthesis are referred to as "one-bead-one-compound" libraries according to Lebl [24–26]. This is based on the fact that due to the nature of splitting and pooling the resin beads, each resin bead in a library contains multiple copies of (ideally) just one single compound. Thus, the library consists of quasi separate compounds as long as they are resin-bound.

A variety of methods have been developed to determine the structure of bioactive compounds within libraries of compound mixtures (see Section 4.2.1).

3.1.1.1. Techniques using Resin Beads

The split – pool protocol is normally carried out on resin beads. Due to the statistical distribution of the solid support at each splitting step

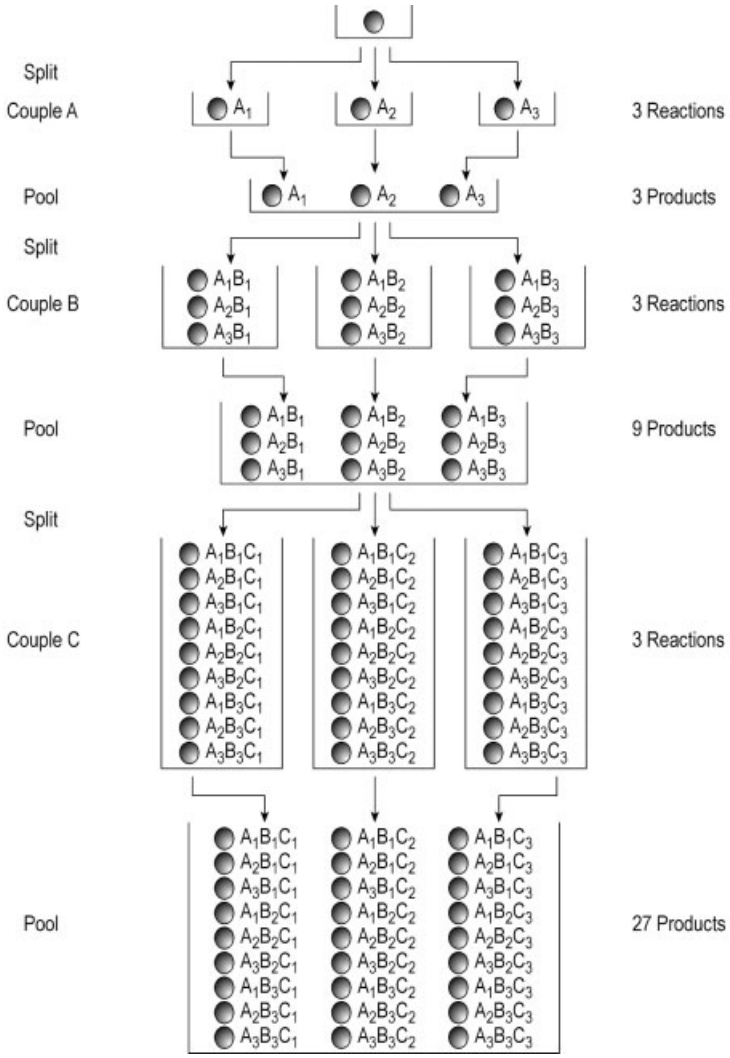

Figure 4. Split – pool synthesis to prepare combinatorial libraries of compound mixtures (Spheres represent resin beads, A, B, C represent the sets of building blocks, borders represent the reaction vessels). In the case of 3 building blocks used in each coupling step, after three stages a total number of 27 different compounds, one on each resin bead, are formed using 9 individual reactions (ignoring deprotection).

the combinatorial synthesis should be carried out with an approximately threefold amount of resin beads (termed *x*-fold, here threefold redundancy), in order to ensure that 95 % of all possible compound members of the combinatorial library are included with a probability greater 99 % [28], [29]. For the commonly used resins (about 100 µm diameter bead), 1 g of the support material corresponds to several million resin beads so that from a statistical point of view, libraries of the order of $> 10^5$ different compounds are possible in practice [28–30]. Depending on the loading capacity of the resin bead, quantities of about 200 pmol (0.1 mg compound, $M_r = 500$) can be obtained per resin bead.

Analytical methods for the quality control of combinatorial libraries comprising resin-bound compound mixtures are described in Section 4.1.

3.1.1.2. Tea-Bag Method

In the so-called tea-bag method, originated by HOUGHTEN and co-workers [14] for multiple

peptide synthesis in 1984, the split – pool protocol occurs batchwise on porous polypropylene mesh packets (termed tea bags) sealed with resin beads for solid-phase synthesis (each tea bag contains resin beads with a different substance). This method offers the advantage that a greater quantity of each compound of the library is available at once (up to 500 µmol), which is sufficient for a complete biological and structural characterization. Furthermore, the structural identity of each synthesized compound can be preserved during the split – pool protocol by labeling the tea bags (see Fig. 5 A).

More recently, NICOLAOU and XIAO [31] as well as MORAN [32] developed a radio frequency encoding system. In practice a glass-encased semiconductor memory microchip capable of receiving, storing, and emitting radio frequency signals [33], [34] is placed in a porous polypropylene capsules along with resin beads. (Glass-encased microchips that are pretuned to emit a unique binary code when pulsed with electromagnetic radiation are widely used to tag and identify laboratory animals via subcutaneous injection.) Taking advantage of the radio frequency encodable microchips each of the capsules (tea bag) can undergo the split – pool protocol in a precise manner, wherein each capsule is radioscanned between solid-phase synthesis rounds and a specific radio frequency signal is recorded on the memory microchip for each reaction step (see Fig. 5 B). In the end, the device can be scanned to record the identity of a compound bound on each batch of resin beads. Depending on the size of the capsule and the resin filling quantities of about 7 – 15 mg compound can be obtained in MicroKan Reactors.

3.1.2. Parallel Synthesis Towards Combinatorial Libraries

Combinatorial libraries can also be prepared by parallel synthesis [35]. Here, compounds are synthesized in parallel using ordered arrays of spatially separated reaction vessels adhering to a traditional "one-vessel-one-compound philosophy" (see Fig. 6). This offers the advantage that each compound, when evaluated for some desired performance (see Section 4.2.2), is substantially "pure" in its local areas, provided that the synthesis has proceeded with high efficiency in each stage. Furthermore, in parallel synthesis the defined location of the compound in the array provides the structure of the compound. A commonly used format for parallel synthesis is the 96-well microtiter plate (standard format in high-throughput screening).

In general, combinatorial libraries comprising hundreds to thousands of compounds are synthesized by parallel synthesis, often in an automated fashion (see Chap. 5). Unlike split – pool synthesis, which requires a solid support, parallel synthesis can be done either on solid phase or in solution.

A number of different solid supports and uniquely designed reaction vessels are adopted for the parallel synthesis of organic compound libraries. The yield of the individual compounds synthesized vary widely from nanomoles to millimoles.

3.1.2.1. Reaction Apparatus using Resin Beads

Various reaction apparatus [36] have been developed for parallel synthesis on the basis of a 96-deep-well polypropylene microtiter plate (e.g., the Flexchem Solid Phase Chemistry System, a modular system of hardware designed for high-throughput parallel synthesis marked by Robbin Scientific Corporation.). The resin beads and reaction solutions are placed into each well of the plate. During the reaction step, the plate is clamped against a Viton (i.e., a fluoroelastomer produced by Du Pont) gasket in order to seal the hole at the bottom of the well. Between reaction steps the reaction solution can be drained and the resin can then be rinsed by removal of the Viton gasket.

In 1993 DEWITT and coworkers described the Diversomer apparatus [37–39], which is one of the first reaction apparatus designed exclusively for the parallel synthesis of small molecules. Pipetting machines dispense reagents and solvents into vials that are located in a custom-made reaction block with 40 positions. The ends of gas dispersion tubes – that contain the resin beads – are placed into the vials allowing the reagents to diffuse into the tubes and contact the solid support. The reaction temperature can be controlled by heating and cooling the reaction block. The apparatus is also enclosed in a

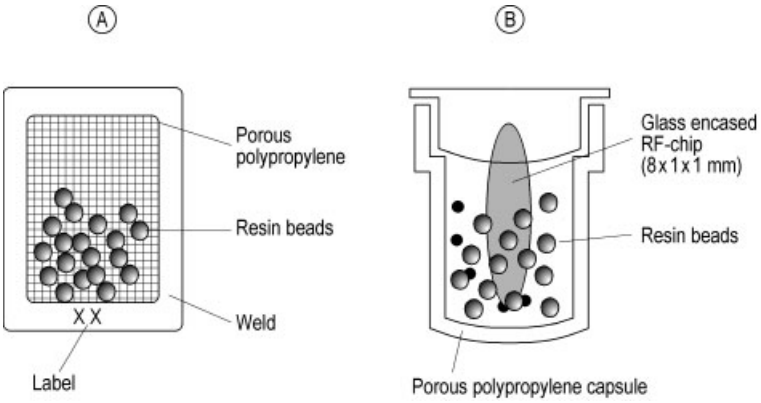

Figure 5. Schematic illustration of labeled polypropylene tea bags (A) and radio-frequency (RF) encoded polypropylene capsules (B)

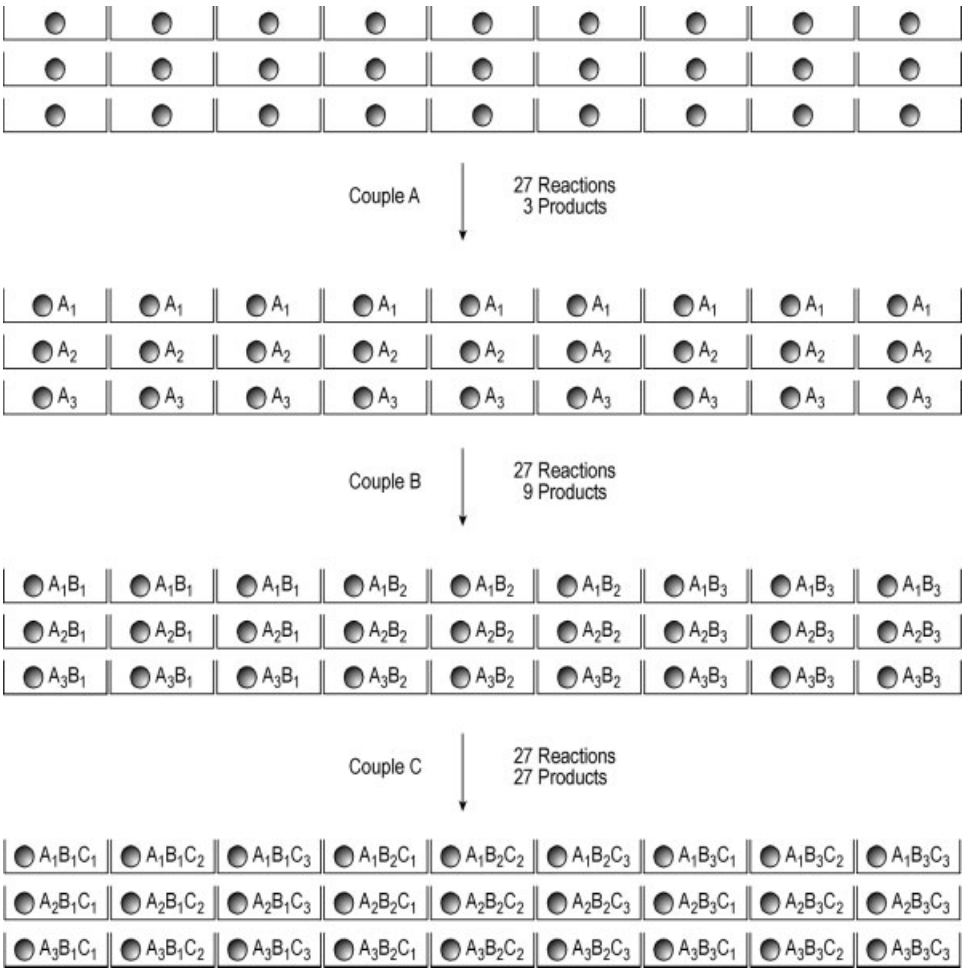

Figure 6. Parallel synthesis to prepare combinatorial libraries of separate single compounds. (Spheres represent resin beads, A, B, C represent the sets of building blocks, borders represent the reaction vessels.) If 3 building blocks are used in each coupling step, after three stages a total of 27 different compounds are formed, one on each resin bead, with 81 individual reactions being used (ignoring deprotection).

manifold with an injectable gasket employed for reagents and solvent additions so that reactions can be maintained under an inert atmosphere. The final compounds are each obtained in quantities larger than 1 mg.

As reaction vessel for resin beads an array of chromatography tubes or syringes fitted with frits and stoppers can also be used. These can be gently agitated on a shaker table or blood tube rotator during the reaction and the resulting compounds can be filtered using a vacuum manifold.

3.1.2.2. Multipin Technique

In 1984, Geysen and coworkers introduced a method for the parallel synthesis of 96 peptide sequences using spatially dispersed polymeric pins [12], [13]. The design of the commercially available Multipin SPOC (Chiron Mimotopes Pty) is based on a modular 8 × 12 matrix of polyacrylamide- or polystyrene-grafted polypropylene pins [40]. The pins are functionalized with a variety of linkers to allow flexibility in solid phase synthesis. The loading levels per pin can range from about 100 nmol to 50 µmol of (ideally) one compound, permitting quantities up to 25 mg compound ($M_r = 500$) to be prepared. Applications are at present in the area of immunology [40].

The execution of the parallel synthesis of up to 96 single compounds by the multipin method (see Fig. 7) involves pipetting the reactants to each well of a standard 96-well microtiter plate. The pin array is then placed on top of the plate and the resin is allowed to incubate with the reactants to perform the coupling step. The reaction temperature can be raised up to 90 °C by placing the reaction block into an incubator. Following each reaction step, the pin array is removed and treated in batch to wash the solid support. These operations are repeated until the desired combinatorial synthesis is completed. The final compounds are cleaved from the pins into a 96-well microtiter plate.

3.1.2.3. Spatially Addressable Parallel Synthesis on Silica Wafer

A special case of parallel synthesis is the spatially addressable synthesis pioneered by Fodor and coworkers [17], [18] in 1991. Here, each member of the library is synthesized at a specific location on a functionalized silica wafer rather than on resin beads in separate reaction vessels. This approach, based initially upon solid-phase peptide synthesis and semiconductor photolithographic techniques by using photolabile amino protecting groups, allows to synthesize combinatorial libraries containing about 50 000 compounds localized to a 50 µm square site on a silica wafer ("library on a chip").

The key points of this method are illustrated in Figure 8. The photolabile protecting groups on the surface of the silica wafer are selectively cleaved at specified locations by site-specific illumination through a mask using instrumentation initially developed for computer microchip construction (→ Imaging Technology, Chap. 4.3.). The entire surface is then exposed to the first set of building blocks bearing also the photolabile protecting group. The coupling reaction occurs only at the sites that were addressed by light in the preceding step. The substrate is illuminated through another mask for deprotection and coupling with a second building block. The steps of illumination and coupling are repeated to build up the members of the compound library. This technique which comprises light-directed spatially addressable parallel chemistry is referred to as very large-scale immobilized polymer synthesis (VLSIPS). It is mostly applied to oligomeric, peptide, and oligonucleotide synthesis.

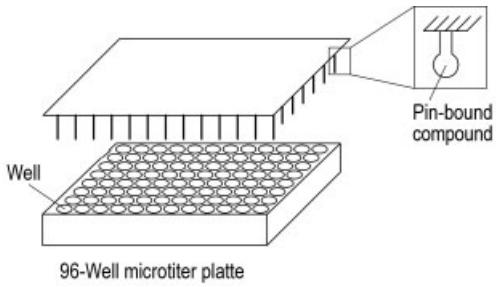

Figure 7. Schematic representation of the Multipin-SPOC-based parallel synthesis

In the light-directed synthesis, the pattern of masks and the sequence of building blocks define the final structure of the compounds synthesized. The equipment required for light-directed peptide synthesis includes a commercial peptide synthesizer, a flow cell to hold the substrate

and deliver fluids from the synthesizer, a translation stage to move among the masks all located on the same large-format glass plate (which is held against the synthesis substrate in a "contact" printing mode) and an arc lamp that can deliver collimated long-wavelength UV light to the mask [41], [42].

3.1.3. Reagent Mixture Synthesis Towards Combinatorial Libraries

In the so-called reagent mixture synthetic approach towards combinatorial libraries of compound mixtures, each reaction step of the combinatorial synthesis is carried out with a building block mixture in one reaction vessel [43]. The use of reagent mixtures requires a thorough knowledge of the mechanism and kinetics involved in the specific reactions being carried out. It is important that the relative reaction rates of the incoming reagents are approximately equal and relatively independent of the resin-bound compounds (i.e., equal nucleophilicity, no significant steric hindrance).

Reasonable results can be obtained for the addition of one set of building blocks. Repetitive cycles using this method can result in large deviations from equimolarity in the final compound mixture, some compounds formed to a lesser extent or not at all.

Combinatorial synthesis towards libraries of compound mixtures can be done either on a solid support or in solution. In both cases an effective deconvolution strategy is required to extract structural information from the results of the biological assay (see Section 4.2.1.2.3).

3.2. Synthetic Methodology for Organic Library Construction

In principle, combinatorial synthesis can be performed both in solution and on solid phase. Although chemistry in solution has the advantage of being familiar and well-established as the method of choice in conventional organic synthesis, to date the majority of the compound libraries have been synthesized on solid phases such as resin beads, pins, or silica wafers. This may be attributed to reasons that include easy parallel work-up procedures (filtration); accelerated reactions with higher yields by employing

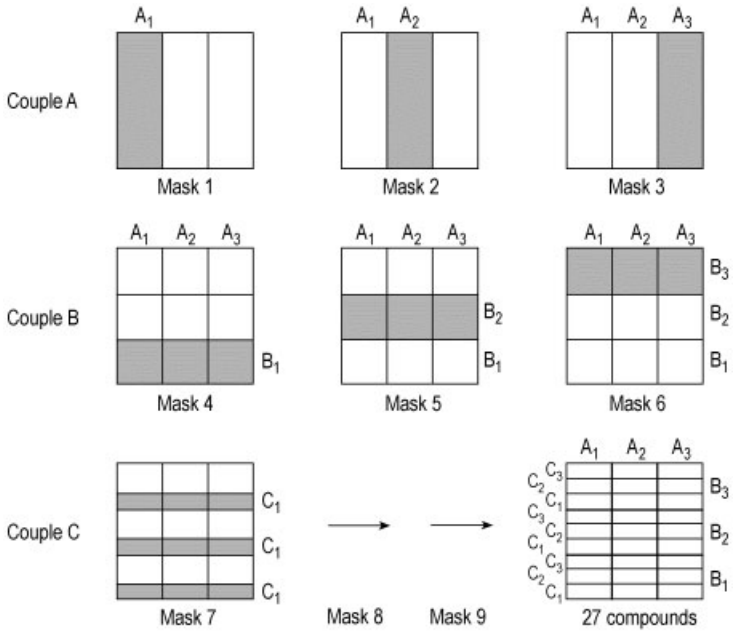

Figure 8. Schematic representation of spatially addressable parallel synthesis on functionalized silica wafer as described in the text (A) and the masking scheme towards combinatorial libraries of a trimer using the generic building blocks A, B, C (B)

excess of reagents; and amenability to automation. Furthermore, it also facilitates the partitioning of compounds into multiple aliquotes in the case of split – pool synthesis.

3.2.1. Solid-Phase Organic Synthesis

Solid-phase organic synthesis is based mainly on the pioneering work of MERRIFIELD [15] in the field of peptide synthesis in 1963 (→ Peptide Synthesis, Chap. 5.2.), and this was followed in the 1970s by investigations on solid-phase synthesis towards organic compounds by LEZNOFF, CAMPS, FRÉCHET, RAPOPORT and others [44–47].

As illustrated in Figure 9, throughout the solid-phase synthesis the compound under construction is covalently attached to a swollen insoluble solid support (usually a resin bead) by a linker that can be cleaved under specific conditions with an appropriate reagent to give the target compound in solution later on (e.g., for assessment of purity, analytical characterization, or biological evaluation). The reactions can be accelerated and driven to completion by using a relatively large molar excess of reagents resulting in reduced reaction time and higher yields. The support matrix in particular facilitates all steps of a synthesis protocol, such as addition of reagent solutions, agitation, filtration, and washing. Thus, solid-phase synthesis enables full automation, even for multistep synthesis, where the building blocks are added repetitively to build up the desired final compound.

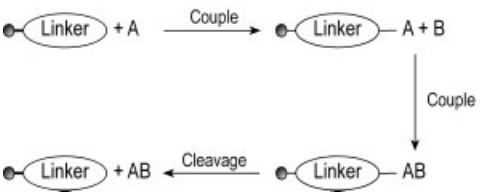

Figure 9. Schematic representation of solid-phase synthesis The sphere represents solid support (usually a resin bead), A and B represent the sets of building blocks, and AB the reaction product

Solid-phase synthesis requires functionalized solid supports and orthogonal chemistry suitable for compound attachment and detachment (orthogonal chemistry means functionalization of a bi-/multifunctional molecule in such a way that only the desired part of the molecule is reacted, leaving the rest of the molecule unaffected), and specialized methods for reaction monitoring in a multistep solid-phase synthesis, including the use of attended orthogonal capping strategies for blocking unreacted substrates. It does not permit the purification of support-bound intermediates. Thus, reactions in solid-phase synthesis must be very selective and efficient. Otherwise, the purification of the final products can present a considerable challenge. The scale of solid phase synthesis is limited and generally restricted by the amount of the solid support and its loading capacity, and the preparation of multimilligram quantities can be cumbersome and expensive for large combinatorial libraries (see Table 1).

Table 1. Comparison of different solid supports

Type of support	Loading capacity	Particle size
Resin beads		
PS[a]-DVB[b] resin	0.8 – 1.2 mmol/g	200 – 400 mesh
PEG[c]-PS[a] resin (TentaGel)	0.2 – 0.3 mmol/g	90 – 130 µm
PEG[c]-PS[a] resin (AgroGel)	0.4 – 0.6 mmol/g	110 µm
High-load resin beads		
PEG[c]-PS[a] resin (TentaGel)	0.4 – 0.6 mmol/g	75 – 110 µm
Macrobeads		
PS[a]-DVB[b] resin	0.8 – 1.2 mmol/g 2 – 5 nmol/bead	160 – 200 µm
PEG[c]-PS[a] resin (TentaGel)	0.25 – 0.35 mmol/g 5 nmol/bead	280 – 320 µm

[a] PS = polystyrene. [b] DVB = divinylbenzene.
[c] PEG = poly(ethylene glycol).

[Example: For a combinatorial library consisting of 10 000 compounds (three-step synthesis at 95 % yield/step) to obtain 50 mg of each compound (M_r = 500) on a PS – DVB resin with a typical loading capacity of 1 mmol/g 1.166 kg solid support is required, i.e., ca. $ 2500 for solid support or ca. $ 6 per compound including chemicals, solvents and disposals.]

3.2.1.1. Solid Support, Linker, and Cleavage Strategies

At present, in combinatorial synthesis of organic compounds two types of polymeric solid support are most commonly used, which are commercially available functionalized with different linkers.

Polystyrene resin cross-linked with 1 – 2 % divinylbenzene (PS – DVB resin, Merrifield resin) is advantageous because of its low price, its high

loading level (up to 2 mmol reactive groups per gram), and its relatively high mechanical stability. One major limitation of the polystyrene resin is that the resin beads are not well solvated in protic solvents, resulting in poor accessibility of reaction sites and thus diminished reaction rate.

Polystyrene – poly(ethylene glycol) copolymers (PEG – PS resin) are obtained by grafting polyoxyethylene chains of controlled lengths onto cross-linked polystyrene-based resin. These polymers are sold, e.g., under the name TentaGel (commercially available from Rapp Polymere, loading capacity in the range of 0.2 – 0.3 mmol/g resin) [48–50] and ArgoGel (commercially available from Argonaut Technologies Inc, loading capacity in the range of 0.4 – 0.5 mmol/g). PEG – PS resins swell extensively in a wide range of solvents and are suitable for some on-bead bioassays (see Section 4.2.1.1), in contrast to PS resins. Disadvantages of PEG – PS relative to PS resins include much higher costs, reduced loading levels, and rather low mechanical stability (stirring or vigorous shaking results in significant loss of resin).

Both resins withstand a wide range of reaction conditions and reagents. New solid supports are being developed to improve the uniformity in size and loading, swelling and mechanical stability properties.

Most linkers are acid or base labile. Since the majority of them originate from peptide synthesis, the reactants are often bound to the support by an ester or amide bond and give carboxylic acids and amides or alcohols or amines after cleavage. A variety of new linkers for combinatorial chemistry is based upon the great wealth of knowledge regarding protection of functional groups. This includes (1) multifunctional linkers which enable the release of different functional groups, depending on the reagent used for cleavage from the linkers and (2) traceless linkers which allow the release of the target compound from the linker without formation of additional functionality, e.g., by the use of cyclization reactions or silicon-based linkers.

An area of special interest are multiple cleavable linkers [51], to release aliquots of the resin-bound compound rather than cleaving all the compound bound to one resin bead at once. When using a multiple release system the compound in solution can be submitted to different bioassays, followed by structural determination of the biologically active substance in compound mixtures (see Section 4.2.1.4). This can be accomplished, e.g., by the use of a three-armed linker that selectively releases compounds at different pH optima [52], [53] or of a photochemically cleavable linker which releases compounds into solution by controlled irradiation [54], [55].

3.2.1.2. Reaction Portfolio

The synthetic repertoire of solid-phase organic chemistry has been growing rapidly in the last years and has expanded its scope in the complexity of compounds synthesized on solid-phase. However, so far only a small percentage of chemistry in solution has been transferred into reliable solid supported protocols [44–47]. This fraction includes multicomponent reactions [56–58]. Two ISIS-based databases have been established covering the solid-phase synthesis literature: solid-phase organic reaction (SPORE) database (from MDL Information Systems and FIZ Chemie Berlin) and Synopsys database (from Synopsys Scientific Systems together with Oxford Diversity).

In combinatorial synthesis the reactions must operate with reliable yield on a structurally broad set of building blocks to provide a multitude of almost pure final compounds under identical conditions. In the most time- and labor-intensive step, selected building blocks are "rehearsed" individually through reactions in the solid phase format, under conditions mimicking those that will be faithfully used in the final combinatorial synthesis. Since it will often be impractical to examine every member of the desired library to confirm its presence, building block combinations that are anticipated to represent worst case scenarios (e.g., with respect to steric and/or electronic factors) are studied and optimized, with problematic building blocks being excluded from the library construction.

In contrast, using an optimized synthesis sequence the combinatorial libraries can be constructed comparatively rapidly and efficiently as long as the building blocks are commercially available or readily accessible.

3.2.2. Synthesis in Solution and Liquid-Phase Synthesis

In the literature few examples exist for combinatorial syntheses in solution [59], most of them single-step parallel syntheses of individual compounds or "reagent-mixture" synthesis of compound mixtures. The use of supported reagents offers opportunities for improving the purity of compounds prepared [60–63]. A complementary approach uses a solid-supported scavenging agent [64–67], nucleophile or electrophile, to remove excess reagents.

An extension of the combinatorial synthesis in solution is achieved by the use of soluble polymeric supports [68], [69] which combines some of the advantages of chemistry in solution and on solid phase. The so-called *liquid-phase combinatorial synthesis* is based on the physical properties of poly(ethylene glycol) monomethyl ether. The polymer is soluble in a variety of aqueous and organic solvents which allows reactions to be conducted in homogeneous phase whereas the propensity to crystallize in appropriate solvents facilitates the isolation and purification of the compound at each step of the combinatorial synthesis.

One of the newer approaches which have yet to be seriously exploited involves the use of fluorinated hydrocarbons [70] to allow further options in the phase separation.

4. Characterization of Combinatorial Libraries

The combinatorial synthesis is followed by the analytical characterization and the biological evaluation of the compound library. Essential elements are the bioassays employed for high-throughput screening (→ High-Throughput Screening) and the methods available for structural elucidation of compounds that emerge from the biological evaluation of these compound libraries, termed as hits. The strategies used depend on the library format – separate single compounds or mixture of compounds, resin-bound compounds or compounds in solution.

4.1. Analytical Characterization

Since in a combinatorial synthesis the amount of compounds bound to a single resin bead is usually of the order of several hundred picomoles, the analytical methods for direct structural determination must be highly sensitive [71]. Efficient methods exist for the structural elucidation of peptides and oligonucleotides in the form of automatic microsequencing procedures (e.g., based on Edman degradation and DNA sequencing) [72]. For minute amounts of low molecular mass organic compounds, mass spectroscopic methods are the method of choice [73], [74], despite progress in the field of IR [75], [76] and NMR [77–80] spectroscopy. The sensitivity of modern mass spectrometry allows the structural determination of compounds in the high femtomolar range. Especially, electrospray mass spectrometry (ES/MS) (→ Mass Spectrometry, Chap. 4.9.) or matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF/MS) techniques (→ Mass Spectrometry, Chap. 4.8.5.) are used for resin-bound compounds.

Characterization of Compound Mixtures. If combinatorial libraries consists of compound mixtures (e.g., after cleavage from the resin beads), a reliable analytical characterization is difficult. In this case the utilization of mass spectrometry for analysis is based on the prediction of mass distribution of the library. Computer-generated distribution profiles can be compared with the actual profile obtained from the compound library [81], [82]. Evaluation of mass distribution detects synthetic problems based on incomplete coupling (shift toward lower molecular masses), incomplete deprotection, or unwanted library modification, such as oxidation, acylation, or alkylation (shift toward higher molecular masses). However, for a larger library of compound mixtures many different compounds will have the same molecular mass. This severely complicates structural determination and even makes it impossible at a certain mixture complexity.

For library analysis of compound mixtures reliable results are obtained by coupling mass spectrometry with HPLC or capillary electrophoresis (CE). In this arrangement, mixtures

of several hundred compounds can be analyzed. A comparison of MS, CE, and NMR for library characterization has been made in [81]. It was concluded that existing analytical techniques can provide sufficient information about prepared libraries.

In addition to the direct methods of structural analysis of libraries of compound mixtures, there are also indirect methods, which are particularly advantageous when working with larger compound libraries made by the split – pool synthesis (see Section 4.2.1).

Characterization of Single Substances. The analysis of separately synthesized single substances is much simpler than that of compound mixtures. Here, the yield and purity can be determined in the traditional way: purification by extraction, crystallization, or chromatography, gravimetrical determination of the yield and confirmation of the purity by elemental analysis, chromatography, or NMR.

4.2. Hit Identification in Combinatorial Libraries by High-Throughput Screening

Organic compounds are usually screened in solution in a 96-well microtiter plate format. Therefore, many bioassays are available, e.g., competitive receptor binding assays with radiolabeled ligands, functional assays, or cell-based assays. All these "solution-phase" bioassays, in principle, can be adapted to high-throughput screening of combinatorial libraries. Besides the growing number of compounds to be tested, the costs for a single bioassay (ca. $ 5 per well) is the main reason for the current trend to miniaturize the assay format towards 384- or 1536-well microtiter plates. (Since the wells of 384- and 1536-well plates are smaller, less substance, solvent, enzyme, etc., is required for one plate, thus reducing the cost per well). In an ultrahigh-throughput screening it should be possible to test more than 100 000 compounds per day.

In the case of combinatorial libraries comprising mixtures of solid-bound compounds the compounds can be screened still covalently attached to the solid support ("on-bead" screening, see Section 4.2.1.1) and, after cleavage of the bond to the support, in solution as compound mixture or as single compounds, when the beads are spatially separated (see Sections 4.2.1.1, 4.2.1.2, 4.2.1.3, 4.2.1.4).

The trend in bioassays is clearly towards examining the test compounds in solution. The results are not influenced by undesired interaction with the solid support. Experiences with "on-bead" assays were sometimes disappointing because even compounds that were known to be biologically active in solution did not bind to a detectable extent to the biological receptor when attached to the bead. Moreover, assaying compounds in solution is probably more pharmacologically relevant. In addition, screening on solid support is not applicable to intracellular targets when a whole cell assay is used.

In the following sections screening strategies for the identification of biologically active compounds (*hits*) in combinatorial libraries are summarized, especially for the most complicated situation when combinatorial libraries of compound mixtures are tested. Here, the exact identity of the compounds has been lost during the synthesis.

4.2.1. Strategies for Libraries of Compound Mixtures

4.2.1.1. On-Bead Screening

Combinatorial libraries can be screened, while the compounds are still attached to the resin beads or onto a surface [17], [18], [83–85]. Thus, the solid support (often TentaGel resin) and its linkers must be soluble in water and for quantitative results (e.g., structure – activity relationships) the beads must be uniform in both size and loading capacity.

The solid-bound library is treated with a labeled soluble biological target. For many studies a fluorescent label has been employed because of the high sensitivity of fluorescent detection. The labeled receptor binds to those resin beads that are derivatized with compounds that have the highest affinity to the biological receptor. The labeled beads are then selected followed by structural elucidation of the support-bound compound (see Section 4.1). The identity of the bioactive substance can be limited to a few alternative structures by mass spectrometric determination of the molecular mass. Very efficient,

automated methods have been developed to isolate the labeled beads, for example, by use of a fluorescence-activated cell sorting instrument [84].

This on-bead screening is particularly useful for libraries of several thousand to million compounds and isolation of a few active compounds from many inactive ones. By using an on-bead assay, the screening of a library of 10^7 resin-bound compounds can routinely be accomplished by one person in one day. One advantage of this method is that, once the library has been prepared and assayed, the remaining compounds may be reused for different biological assays. However, all of the difficulties of screening on the solid support remain.

4.2.1.2. Deconvolution

Three deconvolution strategies can be used for the structural determination of the biologically most active compounds in a combinatorial library comprising mixtures of up to several thousand compounds: (1) iterative deconvolution [27], [86], (2) deconvolution by positional scanning [87–89], or (3) deconvolution by orthogonal libraries [90–92].

4.2.1.2.1. Iterative Deconvolution

The iterative deconvolution method [27], [86] involves preparation of a series of spatially separated sublibraries comprising compound mixtures (termed pools) in which the identity of the building block at least at one position is known and at the remaining positions all combinations of building blocks are incorporated. This can be achieved, e.g., by omitting the final pooling step of the split – pool synthesis. Each pool is screened and the compound mixture with the highest biological activity indicates the importance of the building block at that defined position. On the basis of this result the split – pool synthesis is repeated with the selected building block at the initial defined position in order to prepare pools where the next defined position is introduced. Each pool is evaluated for biological activity in order to select the optimal building block at the additional defined position. The remaining positions are then identified sequentially through an iterative procedure of (re)synthesis and screening, as illustrated schematically in Figure 10 with a generic scaffold.

The iterative deconvolution procedure has been applied successfully to determine the biological active structure directly, but several issues should be considered. Firstly, as the number of compounds in a pool increases, lower concentrations of each compound must often be used in order to maintain the solubility of all the compounds in the pool. Compounds that have modest activity therefore may be not detected. Secondly, the biological activity observed for a given compound mixture depends on both the activity and abundance of the active compounds in each pool, i.e., biological activity is generally a result of the sum of more then one biological active compound. Thus, the compound mixture that shows the greatest biological activity does not necessarily contain the most potent compound(s) [93–96]. Thirdly, the iterative resynthesis and biological evaluation can be a time-consuming and laborious process.

A number of modification have been reported in order to address these complications. JANDA has noted that during the first split – pool procedure, resin can be saved at each cycle immediately prior to resin pooling. This resin can later be used as an intermediate in the iterative resynthesis and deconvolution procedure (recursive deconvolution) [97]. This results in considerable savings in time and effort.

4.2.1.2.2. Deconvolution by Positional Scanning

In the positional scanning approach for deconvolution [87–89], originated by HOUGHTEN and coworkers for peptide libraries, separate sublibraries are prepared, each of which contains a single defined building block at one position and a mixture of building block at the others. The number of sublibraries is equal to the number of variable positions in the substitution pattern. This method resorted to a hybrid of split – pool and parallel synthesis. Each positional sublibrary is then screened to directly determine the building block at each defined position that contributes most to biological activity (see Fig. 11).

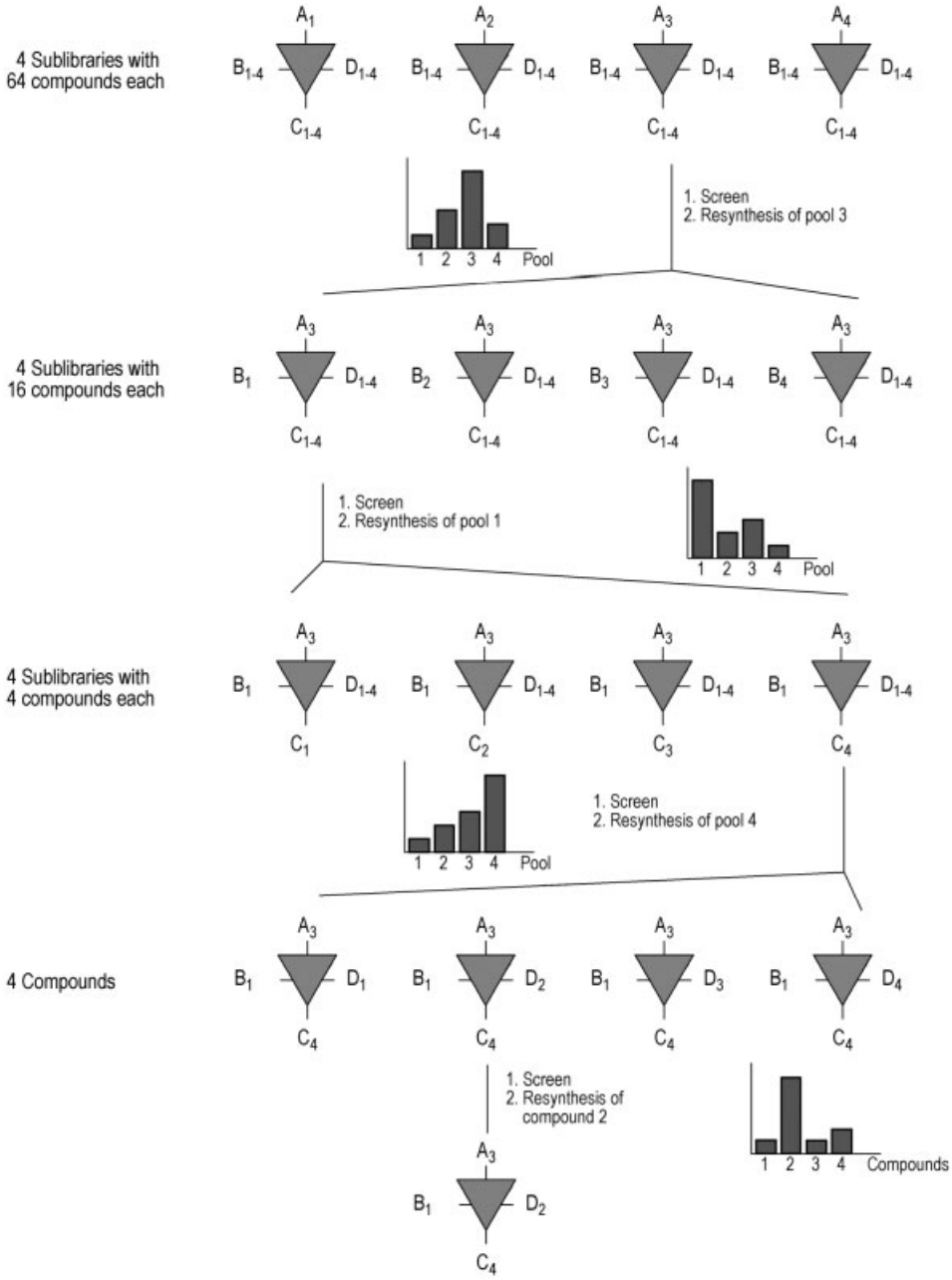

Figure 10. Iterative deconvolution procedure illustrated by generic combinatorial library of a scaffold containing four variable building blocks A, B, C, and D and in total 256 compounds

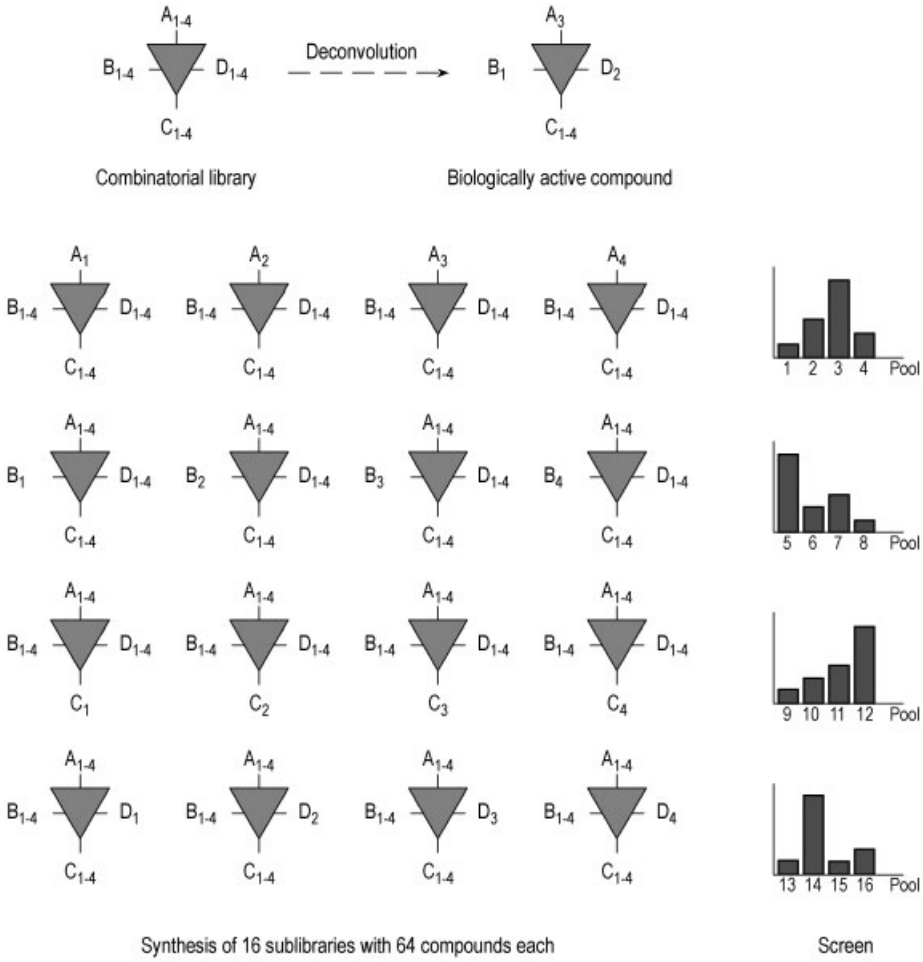

Figure 11. Deconvolution by positional scanning illustrated by a generic combinatorial library of a scaffold containing four variable building blocks A, B, C, and D and in total 256 compounds
Procedure as described in the text.

By combining the positive screening results from all series of sublibraries the whole structure of a highly biological active compound can be deduced directly. In theory, if only one compound was active in the library, activity corresponding to that compound would be found in the one mixture of each sublibrary containing that compound. In reality, the same result is seen, but the activity generally results from the sum of more than one active compound. Anomalous results are seen if the activity is due to the sum of many weakly active compounds.

However, if in one position there is no clear preference for a definitive substituent, all combinations of this substituent must be synthesized in order to find the most active compound. Although active compounds are identified without iterative synthesis using positional scanning, in comparison with iterative deconvolution there is an increased likelihood that the most potent compound(s) will not be identified [98], [99].

The advantage of positional scanning over iterative deconvolution is that the sublibrary syntheses are carried out at once. But the number of split and pool operations that would be required for library synthesis is rather high so that building block mixtures are used in several of the library synthesis steps in order to reduce the number of operations. A thorough examination of the theoretical and experimental aspects of

iterative and positional scanning deconvolution has been published [98].

4.2.1.2.3. Deconvolution by Orthogonal Libraries

A method similar to the deconvolution by positional scanning was developed independently by Pirrung [90], [91] and Déprez [92]. This method has been termed orthogonal library approach or indexed library approach by other authors [100]. With respect to deducing structural information from the screening results this approach involves preparing two series of orthogonal sublibraries consisting of compound mixtures.

The combinatorial synthesis of orthogonal libraries involves the synthesis of two series of sublibraries. In the first series, each of the building blocks A_n reacts separately with a stoichiometric quantity of an equimolar mixture of the building blocks $B_1 - B_x$ and in the second series, each building block B_m reacts separately with an equimolar mixture of the building blocks $A_1 - A_x$. This synthetic approach provides compound mixtures, in which the identity of either the building block A or the building block B is fixed. This synthetic procedure results in each compound being formed exactly twice, and every pair of mixture has exactly one compound in common.

Here, a positive result for any given sublibrary in a biological assay identifies the fixed part of an active compound. By combining the positive results from both series of sublibraries the whole structure of an active component can be deduced directly. Moreover, this form of internal control allows the recognition of false positives in the biological evaluation process. Assignment of activity to a certain structure is possible by simple comparison, that is, without time consuming deconvolution or new synthesis. However, as in the case of the positional scanning method, it seems to be necessary to employ building block mixtures in each of the synthesis steps in order to reduce the number of split – pool operations.

4.2.1.3. Encoding

An alternative strategy with respect to structural determination of active compounds emerging from the biological evaluation of compound mixtures is to label the compound during the split – pool synthesis (encoding or tagging) [9], [101–105]. Here, instead of the structure of the biological active compound the corresponding more accessible (chemical) code is determined.

Chemical Encoding Strategies. In the chemical encoding strategy in general, resin linkers containing two orthogonally protected functional groups are used to allow the concurrent synthesis of both the compound of interest and encoding compounds (so-called tags), on the bead, which upon cleavage are sequenced or otherwise decoded to determine the structure of the compound of interest.

The tags may be added to provide an encoding sequence whereby the structure of the tag encodes for the building block and the location in the sequence encodes for the library synthesis step (oligonucleotides [9], [101–105] and peptides [106–110] as sequenceable tags, Fig. 12 A). The advantage of oligonucleotides and peptides as tags lies in the sensitivity and possibility of sequencing tens or hundreds of samples of nucleic acid codes in parallel (see Section 4.1), However, its application for coding organic libraries is rather limited due to the tendency of DNA and peptide tags to break down under the often very rough conditions of organic synthesis.

Alternatively, each tag may be added individually to the resin [111–117]. In this case the tags must encode for both the building block structure and the step in the library synthesis (haloaromatics as nonsequenceable tags, Fig. 12 B [111–115]).

For each tagging approach, a orthogonal chemistry between the tags and compounds of interest has to be determined, i.e., the compound itself must not be altered during addition of the tag. Those compounds are chosen as encoding compounds that can be easily identified by well-established highly sensitive analytical microtechniques even at very low concentrations.

In all cases, such a coding strategy is, of course, valid only if the relationship between support, compound and the code molecules remains intact. This provides no problem if biological testing is carried out with the support-bound compound. After testing and cleavage the active compound is identified by reading the

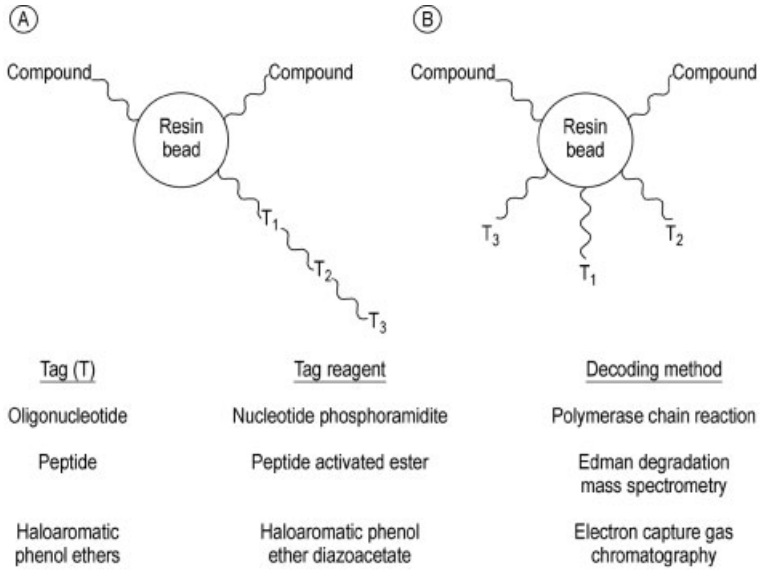

Figure 12. Encoded combinatorial libraries by (A) sequenceable chemical tags (oligonucleotides and peptides) or (B) by nonsequenceable chemical tags (haloaromatic phenol ethers)

code. However, the assay is usually carried out with dissolved compounds. Then, the synthesized compound can be selectively released from the support for the biological assay in solution (see Section 4.2.1.4), while the encoding strand is retained on the bead for later structural elucidation. This procedure requires that the solution containing the compound be spatially addressable to the solid support from which it originated to allow decoding.

Nonchemical encoding strategies, see Section 3.1.1.2.

4.2.1.4. Multiple Cleavable Linkers

Lebl [108], [118–121] and Baldwin [114], [115] first reported a strategy, in which the advantages of performing assays in solution and structural determination of resin-bound compounds are combined. This method is based on multiple cleavable linkers (see Section 3.2.1.2).

The compound library is divided into mixtures of up to several hundred compounds (see Fig. 13). In the first step about one-third of each respective compound bound on a macrobead (about 600 – 800 pmol compound per bead) is split off into solution for biological evaluation. Experiments performed indicate that an optimal mixture complexity is about 20 compounds per mixture [122]. In a second step the collection of beads corresponding to the greatest biological activity is then redistributed in smaller mixtures. Then a further aliquot of the respective compound is released and assaying. Ideally, the beads are rearrayed separately for direct identification of the compound responsible for biological activity ("single bead assay"). The structure of the bioactive compound can then be determined from the remaining resin-bound compound (often only femtomoles) by the analytical method described in Section 4.1.

4.2.2. Strategies for Libraries of Separate Single Compounds

If combinatorial libraries of separate single compounds are prepared by parallel synthesis, elucidating the structure of biological active compounds does not pose a problem, because the site of synthesize is clearly associated with the structure being synthesized (see Section 3.1.2), given that the desired product was build up during synthesis.

One of the major advantages of parallel synthesis is the possibility to compare the bioactiv-

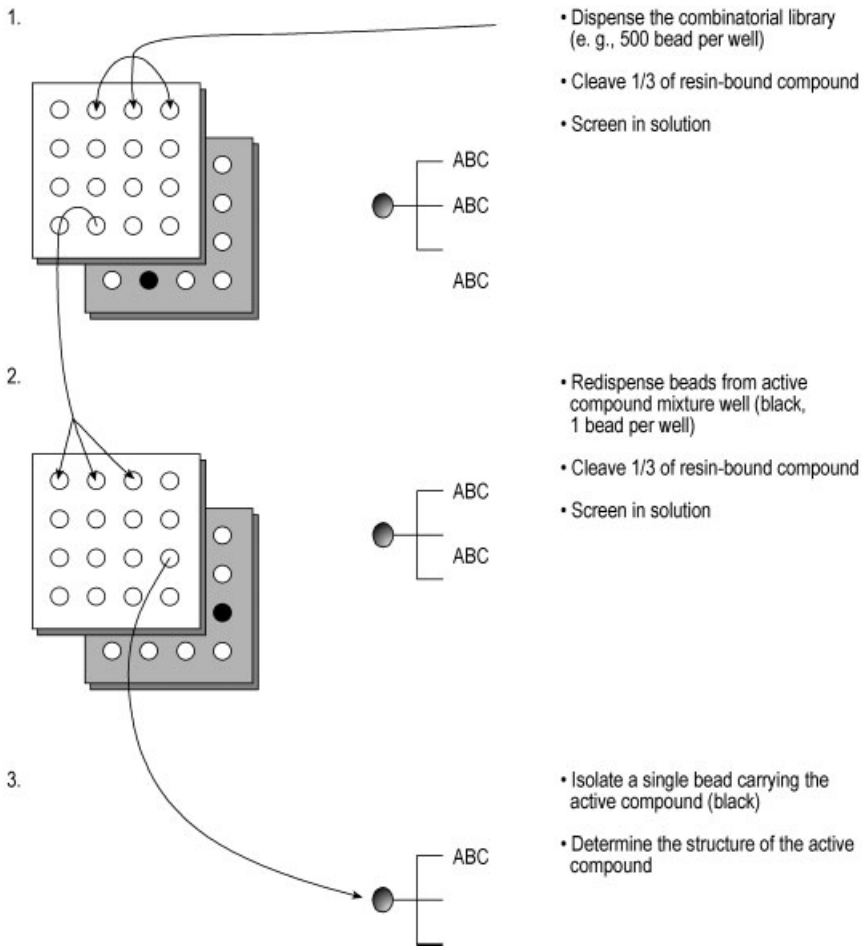

Figure 13. Screening scheme of combinatorial libraries by the use of multiple cleavable linkers: Two-stage screening and then identification of the biologically active compound.

ity of the compounds within the library directly. Thus, the establishment of structure – activity relationships is possible, which is especially important during the drug optimization process. However, the screening capacity may still be a limiting factor, especially during lead optimization (where quite often assays are not automated) or if multiple measurements for dose – response curves are required.

4.2.3. Conclusion

Whether the synthesis and biological evaluation of combinatorial libraries comprising separate compounds or compound mixtures will predominate is difficult to say at present. There are as many examples of both strategies in the literature. It is unlikely that a general strategy will emerge in the future, since the choice of format will always depend on the particular problem in hand.

If compound mixtures are used, significantly more substances can be synthesized and tested per time unit. However, a potential disadvantage is that the initial assays are performed on mixtures of compounds, which may result in some potent library members being overlooked. Although the overall activity of a particular mixture could be poor, it might contain the most active compound in the library. Furthermore, problems with false positives and negatives are known to occur in testing mixtures. This is an important

issue – for which as yet only few guidelines exist – which apparently depends on the particular assay being examined. A common view is that one can afford to miss some active compounds during screening because of the large number of test substances – a sufficient number of interesting compounds for lead discovery will still be detected. Time saved by working with mixtures can, however, be partially or completely lost by the effort required for structural elucidation, that is, new synthesis of less complex sublibraries or single compounds.

5. Automation and Data Processing

5.1. Synthesis Automation and Data Processing

When the size of a combinatorial library reaches thousands of separate single compounds or compound mixtures, manual synthesis and testing against a biological target will hardly be manageable in a acceptable time frame. Therefore, the advent of combinatorial chemistry for the high-throughput synthesis of compounds has driven the advancement of automated methods for synthesis [123], [124] as well as requisite pre- and post-synthesis operations, i.e., resin loading, reagent and resin delivery, compound isolation, purification, and analysis.

Ideally, a fully automated process exists from the combinatorial parallel synthesis to compound isolation, delivering the compounds in a standard 96-well microtiter plate for the following biological high-throughput screening.

Most of the proprietary marketed automated synthesizer derive from instrumentation suitable for multiple peptide synthesis and from pipetting robots. These synthesizers preclude reaction conditions necessary for some organic transformations – especially carbon – carbon coupling reactions using air- or moisture-sensitive reagents that require inert atmosphere – and therefore introduce constraints upon library synthesis. That is why reactor blocks have been elaborated incorporating features most important to synthetic organic chemistry – both solid-liquid-phase synthesis or synthesis in solution. These features include temperatures between – 78 °C and 150 °C and inert atmosphere as well as using a wide variety of organic solvents. Furthermore, reactors for solid-phase synthesis require some mechanism for filtration and washing of the resin-beads (for representative reactors see Section 3.1.2.1).

The architecture of an automated synthesis system can be modular or integrated, open or closed. In a *modular computer-controlled robot system*, all liquid transfers (addition of solvents or reagents and washing solutions) take place in a central module constructed around a robot. The charged reaction vessels are then transferred to special incubation modules, where they remain for the duration of the reaction, while the central module is again available for further operations. In the *closed (nonrobotic) concept,* reagent and solvent are permanently interconnected with the reaction vessels by means of tubes and valves to insure an inert, anhydrous environment. Automation in high-throughput synthesis is still being developed by instrument and robotics companies as well as pharmaceutical companies. Examples are ACT496(MOS) (AdvancedChemTech), Genesis (Tecan), Nautilus(TM)2400 (Argonaut Technologies, Inc.) [125], RAM(TM)Synthesizer (Bohdan Automation, Inc.), Sophas(TM3) (Zinsser Analytic), Syro(TM) (MultiSynTec), and Zymark robotic workstation (Zymark, Inc.).

Automated combinatorial chemistry will demand a computer control of processing instrumentation. A *serial processing system* performs all operations on one sample before proceeding to the next one. On the other hand, automated synthesizers that processes samples *in parallel* perform the same operation on a plurality of samples before proceeding to the next operation. Parallel processing significantly improves the throughput and efficiency of automated synthesis systems. Furthermore, efficient software tools are necessary to program the synthesis run, to retain records of the synthetic operations or the biological testing and to handle the huge amount of compounds and corresponding data with respect to compound searches or data interpretation [126].

Current automated instrumentation for high-throughput synthesis are still below the capacity of the well-established high-throughput screening of compounds. Some instruments can produce 1000 to 2000 compounds per day by parallel synthesis. The time it will take to achieve

an automated synthesis capability that is equal to screening throughput is not yet known. By performing biological binding experiments in microtiter plates and employing both special pooling strategies and automation, nowadays 10 000 – 50 000 compounds can be screened against a given biological receptor in a single day with a good level of reliability (see also Section 4.2).

5.2. Automated Purification

In principle combinatorial synthesis does not create analytically pure compounds even though the applied chemistry might be optimized. With respect to structure – activity relationships the compounds evaluated in lead optimization should be as pure as possible. In high-throughput screening for lead finding the compounds are usually tested without purification. If the purity becomes a severe problem, it must be considered to purify all compounds before screening or to purify only the active molecules.

If *HPLC purification* is necessary, a one-column system can purify about 100 samples in milligram scale in one day. In this case it is advantageous to use an LC – MS system where the fraction collector is driven by the mass spectroscopic information because all fractions are already analyzed and ideally a reformatting step in the 96-well format is not required. HPLC-systems using more than one column in parallel are rare on the market. A commercially available system from Biotage consists of four columns driven by one pump and an individual fraction collector for each column. However, the system is not connected on-line to a MS and all fractions have to be analyzed off-line by MS. This may cause difficulties in sample-tracking and reformatting into the 96-well format, which should be controlled by an appropriate software.

Solid-phase extraction might be an alternative to HPLC separations, especially when large libraries are synthesized for high-throughput screening. Many of the modules for solid phase extraction can handle 96 samples in an microtiter plate array in parallel either manually or automatically, depending on the system. Solid-phase extraction is suitable to eliminate very polar or unpolar byproducts or salts, Examples of equipment include Roliman-96 (EPR Labautomation), Microlute (Zinsser Analytic), Genesis SPE (Tecan), ASPEC XL4 (Gilson, and Abimed Analysen-Technik), VacMaster (ict), and Baker spe∗12G System (J. T. Baker).

6. Library Design and Diversity Assessment

Combinatorial libraries serve two distinct functions in the drug discovery process, lead identification and lead optimization (see Fig. 1), which dictate their size and composition.

A compound library suitable for the discovery of a novel lead, where a priori no specific structure or substructure is obligatory as a starting point, will initially consist of a large number of compounds within a broad structural range (so-called "generic", "unbiased" or "random" library). Thus, the chance to identify a compound of significant affinity for the novel biological target or an alternative lead for a known biological target will be increased in random screening. On the other hand, in the case of lead optimization the compound library will be invariably of a more limited size with close structural resemblance to the original lead structure (so-called "focused", "biased" or "directed" library). Most drug development processes include a series of synthesis cycles within a targeted screening in order to meet a set of predetermined criteria for a drug candidate.

Thus, the library design (compound selection for combinatorial synthesis) as a central element in combinatorial chemistry depends markedly on the intended use of the library (see Table 2). In the case of generic libraries the best overall strategy for library design is rather blurred, and a variety of approaches are under investigation. One approach favors using structural motifs, which have distinguished themselves by appearing frequently in previous potent drugs or drug candidates, so-called "privileged" structures. One of the first examples of this type was provided by Ellman and coworkers [127–130] in the preparation of libraries of benzodiazepines, an important class of heterocycles that have found widespread use in the pharmaceutical industry. Another approach is to incorporate key recognition elements for receptor binding (pharmacophoric patterns: pharmacophores are

those structural features of a molecule required for the particular biological activity) that are relevant to the particular biological target under investigation [131]. These pharmcophoric patterns are used to provide a fingerprint of the specific binding characteristics. A third approach is to select stable compounds upon which only a few or no therapeutic agents or candidates have been based, but which are straightforward to prepare with multiple sites available for the display of different functionalities.

Table 2. Library design in the case of lead finding and lead optimization

	Lead finding	Lead optimization
Library type	generic	focused
Library size	huge ($\gg$ 10 000 compounds, $<$ 1 mg per compound)	modest ($\ll$ 10 000 compounds, $>$ 1 mg per compound)
Structural goal	no specific initial	specific (lead oriented)
Structural diversity	broadest	relatively narrow
Building blocks	any diverse	specific according to retrosynthesis
Synthetic strategy	flexible	well defined
Screening goal	many biological targets	one biological target or target class

6.1. Diversity Assessment for Selection of Building Blocks or Compound

The synthetic effort towards combinatorial libraries of several hundred thousands compounds is considerable. Therefore, computer-based techniques were introduced to reflect and analyze compound libraries to enhance the composition, while constraining their size.

Several groups have validated physicochemical or structural descriptors to quantify the similarity or so-called molecular diversity of compounds and compound libraries [132–136]. Representative approaches are based on physicochemical descriptors i.e., on log P, where P is the octanol-water partition coefficient, or on molar refractivity, 2D descriptors (2D fingerprints, atom pairs), or 3D descriptors in connection with the 3D model of the biological target (CoMFA, 3D fingerprints) (→ Information and Documentation). Examples of diversity profiling using 2D versus 3D descriptors are given in [137], [138]. One of the basic flaws in the concept of a diversity parameter is the lack of evidence that it correlates directly with biological activity of the compounds.

The computer-based library design process includes modules for the selection of the building block sets, the creation of the corresponding virtual compound library, and the selection of the compounds to be synthesized and screened. The selection procedure for the generation of a generic library with a broad range of structurally unrelated compounds by comparing the building blocks or the compounds of the virtual library consists of calculating descriptors for the compounds, weighting them according to an algorithm, and then calculating a similarity coefficient (e.g., Tanimoto coefficient). Part of the selection process is to determine the availability and cost of reagents used as building blocks. The members of a building block set should reflect a broad range of physicochemical properties, functionalities, charges, and conformations. Databases of chemicals are available for this purpose, e.g., the Available Chemicals Directory by MDL Information Systems, a collection of chemical products offered by more than 200 chemical suppliers and totaling well over 400 000 compounds.

Modules in software programs for diversity analysis are, e.g., elector (Tripos), Jarpat (Daylight Chemical Information Systems), C^2-Diversity (Molecular Simulation), and Chem-X (Chemical Design). In the context of management of compound libraries software programs such as Project Library, Legion, or UNITY (all by Tripos) are available.

6.2. Iterative Optimization Methods

The interpretation of the biological data resulting from library screening drives the new (secondary, tertiary, etc.) library design, creating an iterative cycle of combinatorial library synthesis and biological evaluation. One challenge in lead optimization is to capture the data and to build structure – activity relationship (SAR) and quantitative structure – activity relationship (QSAR) models, based on a larger data set. The power of combinatorial chemistry and in this particular case the efficacy of (automated) parallel synthesis is to produce hundreds of analogs in a short time frame. This will contribute to the faster development of a drug candidate with the

desired biological activity. Furthermore, other main tools of traditional drug development including structure- and mechanism-based drug design will be integrated in the library design.

An alternative computer-based approach minimizing the number of compounds to be synthesized is the iterative use of genetic algorithms [139], [140] in library design. Here, in order to find biologically active compounds from a virtual library of hundred thousand compounds only a small number of randomly selected compounds are synthesized, followed by biological evaluation. The most active compounds are selected, which formed the parent generation of computation and synthesis of further compounds. In an example given in [142], this process is carried out with 20 compounds for each of the 20th generations so that in total 400 (0.25 % of the possible 160 000) reaction products were synthesized and tested. The compounds of the 20th generation were 2000 times more active against the biological target (thrombin inhibitor) than those of the first generation.

7. Economic Aspects and Industrial Case Studies

Since the beginning of the 1990s combinatorial chemistry has become a key technology in most pharmaceutical companies. Combinatorial chemistry is changing the drug discovery process (see Fig. 1). The importance of this technology is also indicated by several alliances between major pharmaceutical companies and small "Combi Chem" companies and library suppliers. From 1995 to 1997 alone, more than 60 of these agreements were signed, which typically included up-front payment, milestones, and substantial royalties on the sale of commercialized products [141], [142].

Combinatorial chemistry can contribute to drug discovery on the one hand by improving the finding of biologically active compounds (leads) in combination with high-throughput screening and on the other hand by acceleration of lead optimization resulting in an active drug candidate with a desired pharmacological profile.

Historically, the pharmaceutical compound collections range from several ten thousand to several hundred thousand compounds per company, depending on the size and history of the company. These compounds have long been regarded as the "crown jewels" of their respective R & D organizations. Assuming one newly synthesized compound per researcher per week and using $ 250 000 as the fully burdened cost per researcher per year, a 100 000 compound library approximates a $ 500 million investment in compound synthesis, patenting, and potential screening against biological targets [141]. In this context, combinatorial chemistry raises the opportunity to build up or amplify the number of compounds for high-throughput screening in a shorter period of time. Combinatorial chemistry companies such as Pharmacopeia report the generation of 100 000 compounds per calendar quarter and the use of robotic systems capable of analyzing 100 000 compounds per week.

Another aspect in many companies is the diversity of the compound pool. Typically, it contains many molecules with high structural similarity because they were prepared during lead optimization in which only slight variations on the architecture of the molecules were performed. The design and synthesis of diverse combinatorial libraries according to computational methods described in Chapter 6 may contribute to improve the overall diversity of a compound collection with respect to lead finding.

Thus, combinatorial chemistry is emerging as a time- and cost-efficient tool for drug discovery [143] although there are presently very few compounds in the clinic trails that were discovered utilizing combinatorial chemistry.

Scores of patents already exist in the field of combinatorial chemistry directed to methods of making libraries, devices used in these methodologies, assay methodologies, and on libraries themselves [144].

Combinatorial chemistry will probably not substitute traditional medicinal chemistry, it should be considered as one important tool in the process of drug discovery. Numerous possibilities for structural variations of a given compound, e.g., replacement of atoms in a ring system, change of ring size, rigidization through introduction of bridging units as well as large scale synthesis are still a domain of traditional synthetic chemistry.

8. Further Developments

The vast majority of combinatorial libraries of organic compounds have been designed to accelerate the drug discovery process. Beyond this, the application of combinatorial chemistry to problems of catalysis and material science promises to discover materials of industrial interest with novel catalytic, electronic, magnetic, optical, and mechanical properties. Moreover, the results from such experiments should significantly increase the theoretical understanding of relationship between structures and their properties, enhancing predictive ability even further.

Examples reflecting the current state of development in the field of inorganic solid-state substances with respect to material science and transition metal compounds with respect to (homogeneous) catalysis will be given in the next two sections.

8.1. Combinatorial Chemistry and Catalysis

The challenge in catalysis is to find catalysts with high catalytic activity and selectivity for given chemical transformations and to optimize them. In addition, the catalyst must fulfill technical demands such as satisfactory lifetime and regenerability. Despite extensive studies on structure – activity relationships and underlying mechanisms, many improvements in catalyst design arise from extensive empirical catalyst syntheses and mass screening. In chemical industry the acceleration of the catalyst development can have a large economic and ecological effect, because more than 90 % of current industrial processes involve catalytic steps.

Combinatorial chemistry is not yet widely used to accelerate the development of new catalysts [145–149]. Since 1995 some attempts have been made to apply combinatorial strategies to problems in stereoselective homogeneous catalysis, mainly by U.S. groups. Here, parallel synthesis of ligands, which can complex transition metals, were carried out on the solid phase according to combinatorial principles or at least according to combinatorial related strategies [150], [151]. This resulted in libraries of potential catalytically active transition metal compounds, which were tested to discover homogeneous catalysts for stereoselective synthesis. It is important to mention that the catalytic activity and selectivity could be markedly different, when the catalyst is still attached to the resin bead and after release.

To identify catalysts with desired properties from a compound library it is important to establish techniques for an high-throughput catalyst screening. In this respect the following important question has to be answered: Is it necessary to determine both the catalytic activity and selectivity simultaneously or is the measurement of only one of this characteristics sufficient for the catalytic reaction? In this context it has also to be taken into account that the catalytic activity and selectivity is inherently dependent on physical parameters such as temperature and pressure. A common view in catalysis is that it is possible to obtain selective catalysts in an iterative optimization cycle but that it is more difficult to reach turnover numbers which are sufficient from a technical point of view. Efforts in the field of rapid and parallel catalyst screening so far have focused on for example, the detection of heat liberation in the catalyzed reactions by noninvasive IR thermography [152] or of the catalytic reaction progress by spectrophotometric techniques [153], [154], when chromogenic reagents were used. The analytical results can be quantified and correlated with the catalytic activity.

The concept of developing catalysts combinatorially has great promises, but the proof of the concept has so far been elusive.

8.2. Combinatorial Chemistry and Material Science

Even though the chemistry of extended inorganic solids has been extensively explored, only limited general principles have emerged that allow to predict composition, structure, and reaction pathways for the synthesis of solid state compounds of industrial interest, such as superconducting or magnetoresistive materials. In light of these challenges, it seemed to be that combinatorial chemistry could be a way to the discovery of new solid state materials. However, up to now there are mainly Schultz, Xiang, and coworkers as well as the company Symyx, who have published results in this area of combinatorial chemistry since 1994 [155–162].

Unlike the discovery of new drug candidates, where the assayed properties are based on a molecule – molecule interaction, the property of interest in material science are based on the bulk properties of compounds. Thus, in this case parallel synthesis and characterization of each separate candidate is required.

SCHULTZ, XIANG, and coworkers [155–158] used thin-film deposition methods and physical masking-techniques to synthesize spatially defined libraries of solid-state thin films. The individual element precursors of interest were sequentially sputtered onto the surface of a single-crystal wafer of magnesium oxide or lanthanum aluminate through a series of physical masks. A different masking pattern was used for each element precursor, and the patterns were chosen so that every possible combination of the elements was found somewhere on the wafer (see Section 3.1.2.3). Subsequently, the coated wafer was subjected to thermal and oxidative treatment to form the final library of solid-state thin film.

Rapid detection systems for the physical properties of the thin films are also being developed including contact and rapid scanning probes. The properties of the resulting materials were then evaluated to identify promising candidates for further development.

In addition, improved methods for combinatorial processing are also being developed, in which not only composition and stoichiometry but also film thickness, processing temperature, pressure, and atmosphere can be varied. Multihead ink-jet deposition systems are being tested for generating libraries derived from solution phase precursors. These methods together with theoretical and empirical input may considerably facilitate both the material discovery process and the ability to test theoretical predictions and map phase diagrams.

9. References

1. W. H. Moos, M. R. Pavia, A. D. Elington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997.
2. R. Cortes (ed.): *Combinatorial Libraries – Synthesis, Screening and Application Potential,* Walter de Gruyer, Berlin 1996.
3. I. Chaiken, K. Janda (eds.): *Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery,* ACS, Washington 1996.
4. G. Jung (ed.): *Combinatorial Peptide and Non Peptide Libraries – a Handbook,* VCH Verlagsgesellschaft, Weinheim 1996.
5. R. A. Houghten (ed.): *Combinatorial Libraries: The Theory and Practice of Combinatorial Chemistry,* ESCOM, Leiden 1995.
6. "Combinatorial Chemistry", *Chem. Rev.* **96** (1996) Special thematic issue.
7. "Combinatorial Chemistry", *Acc. Chem. Res.* **29** (1996) Special thematic issue.
8. F. Balkenhohl, C. von dem Bussche-Hünefeld, A. Lansky, C. Zechel, *Angew. Chem. Int. Ed. Eng.* **35** (1996) 2289.
9. M. A. Gallop et al., *J. Med. Chem.* **37** (1994) 1233 – 1251, 1385 – 1401.
10. A dynamic database of articles in the field of library synthesis (both peptide libraries and organic compound libraries) is maintained on the world wide web by the journal *Molecular Diversity,* see also http://www.5z.com
11. A database *Combinatorial Chemistry and Mass Screening* Yellow Pages is maintained on the world wide web http://www.netsci.org/Resources
12. H. M. Geysen, R. H. Meloen, S. J. Barteling, *Proc. Natl. Acad. Sci. USA* **81** (1984) 3998 – 4002.
13. H. M. Geysen, S. J. Barteling, R. H. Meloen, *Proc. Natl. Acad. Sci. USA* **82** (1985) 178 – 182.
14. R. A. Houghten, *Proc. Natl. Acad. Sci. USA* **82** (1985) 5131 – 5135.
15. R. B. Merrifield, *J. Am. Chem. Soc.* **85** (1963) 2149 – 2154.
16. R. Frank, *Tetrahedron* **48** (1992) 9217 – 9232.
17. S. P. A. Fodor et al., *Science* **251** (1991) 767 – 773.
18. Affimax, US 5 143 854, 1992 (M. C. Pirrung, J. L. Read, S. P. A. Fodor, L. Stryer).
19. "Studies on Posssiblities of Systematic Searching for Pharmaceutically Useful Peptides": Á. Furka, notarized Document No. 36237/1982, Budapest, 1982.
20. Á. Furka, F. Sebestyen, M. Asgedom, G. Dibo: *Abstr. 14th Int. Congr. Biochem.* **vol. 5,** Prague, 1988, p. 47.
21. Á. Furka, F. Sebestyen, M. Asgedom, G. Dibo, *Int. J. Pept. Protein Res.* **37** (1991) 487 – 493.
22. F. Sebestyen, G. Dibo, A. Kovacs, Á. Furka, *Biomed. Chem. Lett.* **3** (1993) 413 – 418.

23. R. Hirschmann, *Angew. Chem. Int. Ed. Engl.* **30** (1991) 1278 – 1301.
24. K. S. Lam et al., *Nature* **354** (1991) 82 – 84.
25. M. Stankova et al., *Drug. Dev. Res.* **33** (1994) 146 – 156.
26. K. S. Lam, M. Lebl, V. Krchnák, *Chem. Rev.* **96** (1996) 411 – 488.
27. R. A. Houghten et al., *Nature* **354** (1991) 82 – 86.
28. K. Burgess, A. I. Liaw, N. Wang, *J. Med. Chem.* **19** (1994) 2985 – 2987.
29. P.-L. Zhao et al., *Proc. Natl. Acad. Sci. USA* **92** (1995) 10212 – 10216.
30. J. A. Boutin, A. L. Fauchère, *Trends Pharmac. Sci.* **17** (1996) 8 – 12.
31. K. C. Nicolaou et al., *Angew. Chem. Int. Ed. Engl.* **34** (1995) 2289 – 2291.
32. E. J. Moran et al., *J. Am. Chem. Soc.* **117** (1995) 10787 – 10788.
33. Biomedic Data Sytems, US 5 252 962, 1993.
34. Texas Instruments, US 5 351 052, 1994.
35. M. C. Pirrung, *Chem. Rev.* **97** (1997) 473 – 488.
36. H. V. Meyers et al., *Mol. Diversity* **1** (1995) 13 – 20.
37. S. Hobbes DeWitt, A. W. Czarnik, *Acc. Chem. Res.* **29** (1996) 114 – 122.
38. S. Hobbes DeWitt et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 6909 – 6913.
39. Warner-Lambert-Company, WO 94/08711, 1994.
40. H. M. Geysen, S. J. Rodda, T. J. Mason, G. Triblick, P. G. Schoofs, *J. Immol. Methods* **102** (1987) 259 – 274.
41. T. Kreiner, *Am. Lab.* **2** (1996) 244 – 255.
42. A. C. Pease et al., *Proc. Natl. Acad. Sci. USA* **91** (1994) 5022 – 5026.
43. J. M. Ostresh, J. H. Winkle, V. T. Hamashin, R. A. Houghten, *Biopolymers* **34** (1994) 1681 – 1689.
44. L. A. Thompson, J. A. Ellman, *Chem. Rev.* **96** (1996) 555 – 601.
45. J. S. Früchtel, G. Jung, *Angew. Chem. Int. Ed. Engl.* **51** (1996) 17 – 42.
46. P. H. H. Hermkens, H. C. J. Ottenheijm, D. Rees, *Tetrahedron* **53** (1996) 4527 – 4554.
47. Solid-Phase Chemistry brochures published by Chiron Mimotopes Pty. Ltd.
48. E. Bayer, *Angew. Chem. Int. Ed. Engl.* **30** (1991) 113 – 129.
49. Rapp, DE 3 500 180, 1986 (E. Bayer, W. Rapp).
50. Rapp, US 4 908 405, 1990 (E. Bayer).
51. D. Madden, V. Krchnak, M. Lebl, *Perspect. Drug Discovery Des.* **1** (1995) 269.
52. S. E. Salmon et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 11708 – 11712.
53. Selectide, , WO 95/00165, 1995 (M. Lebl, V. Krchnak, P. Kocis).
54. J. J. Baldwin, J. J. Burbaum, I. Henderson, M. H. J. Ohlmeyer, *J. Am. Chem. Soc.* **117** (1995) 5588 – 5589.
55. C. P. Holmes, D. G. Jones, *J. Org. Chem.* **60** (1995) 2318 – 2319.
56. R. W. Armstrong et al., *Acc. Chem. Res.* **29** (1996) 123 – 131.
57. I. Ugi, A. Dömling, W. Hörl, *Endeavour* **18** (1994) 115 – 123.
58. I. Ugi, S. Lohberger, R. Karl in B. M. Trost, I. Fleming (eds.): *Comprehensive Organic Synthesis,* **vol. 2,** Pergamon, New York 1991, p. 1083 – 1109.
59. R. Storer, *Drug Discov. Today* **1** (1996) 248 – 254.
60. P. Hodge, *Chem. Soc. Rev.* **26** (1997) 417 – 424.
61. S. J. Shuttleworth, *Synthesis* 1997, 1217 – 1239.
62. S. Kobayashi, *Tetrahedron Lett.* **37** (1997) 9221 – 9224.
63. A. Akelah, D. C. Sherrington, *Chem. Rev.* **81** (1981) 557.
64. R. J. Booth, *J. Am. Chem. Soc.* **119** (1997) 4882 – 4886.
65. D. L. Flynn, *J. Am. Chem. Soc.* **119** (1997) 4874 – 81.
66. R. M. Lawrence, *Synthesis* 1997, 553 – 558.
67. S. W. Kaldor et al., *Tetrahedron Lett.* **37** (1996) 7193.
68. D. J. Gravert, K. D. Janda, *Chem. Rev.* **96** (1996) 489 – 509.
69. H. Han, M. M. Wolfe, S. Brenner, K. D. Janda, *Proc. Natl. Acad. Sci. USA* **95** (1995) 6419 – 6423.
70. D. P. Curran, S. J. Hadida, *J. Am. Chem. Soc.* **118** (1996) 2531.
71. W. L. Fitch: "Analytical Methods for the Quality Control of Combinatorial Libraries", in W. H. Moos, M. R. Pavia, A. D. Ellington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997.
72. D. Stoll et al., in H. L. S. Maia (ed.): *Peptides 94, Proc. 23. EPS,* ESCOM, Leiden 1997, p. 401.
73. A. L. Burlingame, R. K. Boyd, S. J. Gaskell, *J. Anal. Chem.* **68** (1996) 599.
74. C. L. Brummel et al., *J. Anal. Chem.* **68** (1996) 237.

75. B. Yan, G. Kumaravel, *Tetrahedron* **52** (1996) 843 – 848.
76. B. Yan, et al., *J. Org. Chem.* **60** (1995) 5736 – 5738.
77. P. A. Keifer et al., *J. Mag. Reson. Ser. A.* **119** (1996) 65 – 75.
78. S. K. Sarkar, R. S. Garigapati, J. L. Adams, P. A. Keifer, *J. Am. Chem. Soc.* **118** (1996) 2305 – 2306.
79. R. C. Anderson et al., *J. Org. Chem.* **60** (1995) 2650 – 2651.
80. G. C. Look, C. P. Holmes, J. P. Chinn, M. A. Gallop, *J. Org. Chem.* **59** (1994) 7588 – 7590.
81. J. A. Boutin et al., *Anal. Biochem.* **234** (1996) 126.
82. J. Metzger, S. Stevanovic, K. H. Wiesmüller, G. Jung, *Methods (San Diego)* **6** (1994) 425.
83. D. A. Campbell, J. C. Bermak, T. S. Burkoth, D. V. Patel, *J. Am. Chem. Soc.* **117** (1995) 5381 – 5382.
84. D. Vetter, E. M. Tate, M. A. Gallop, *Bioconjugate Chem.* **6** (1995) 319 – 322.
85. K. S. Lam, M. Lebl, *Methods, A Companion of Methods in Enzymology* **6** (1994) 372 – 380.
86. H. M. Geysen, S. J. Rodda, T. J. Mason, in R. Porter, J. Whelan (eds.): *Synthetic Peptides as Antigens,* John Wiley and Sons, Chichester 1986, p. 130.
87. C. T. Dooley et al., *Science* **266** (1994) 2019.
88. C. T. Dooley, N. N. Chung, P. W. Schiller, R. A. Houghten, *Proc. Natl. Acad. Sci. USA* **90** (1993) 10811.
89. C. Pinilla, J. R. Appel, P. Blanc, R. A. Houghten, *Biotechniques* **13** (1992) 901.
90. M. C. Pirrung, J. Chen, *J. Am. Chem. Soc.* **117** (1995) 1240.
91. M. C. Pirrung, J. H.-L. Chau, *J. Chem. Biol.* **2** (1995) 621.
92. B. Deprez et al., *J. Am. Chem. Soc.* **117** (1995) 5405.
93. C. Pinilla et al., *Drug Dev. Res.* **33** (1994) 133.
94. J. R. Wyatt et al., *Proc. Natl. Acad. Sci. USA* **91** (1994) 1356.
95. C. T. Dooley, R. A. Houghten, *Life Sci.* **52** (1993) 1509.
96. R. A. Houghten, *Gene* **137** (1993) 7.
97. E. Erb, K. D. Janda, S. Brenner, *Proc. Natl. Acad. Sci. USA* **91** (1994) 11422.
98. L. Wilson-Lingardo et al., *J. Med. Chem.* **39** (1996) 2720.
99. S. M. Freier, D. A. M. Konings, J. R. Wyatt, D. J. Ecker, *J. Med. Chem.* **38** (1995) 344.
100. P. W. Smith et al., *Bioorg. Med. Chem. Lett.* **4** (1994) 2821.
101. K. D. Janda, *Proc. Natl. Acad. Sci. USA* **91** (1994) 10779.
102. J. Nielson, S. Brenner, K. D. Janda, *J. Am. Chem. Soc.* **115** (1993) 9812.
103. S. Brenner, R. A. Lerner, *Proc. Natl. Acad. Sci. USA* **89** (1992) 5381.
104. J. C. Chabala, *Perspectives in Drug Discovery and Design* **2** (1994) 305.
105. M. C. Needels et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 10700.
106. E. R. Felder et al., *Mol. Diversity* **1** (1995) 109.
107. J. M. Kerr, S. C. Banville, R. N. Zuckermann, *J. Am. Chem. Soc.* **115** (1993) 2529.
108. V. Nikolaev et al., *Pept. Res.* **6** (1993) 161.
109. Selectide, WO 94/28028 (M. Lebl et al.).
110. Ciba Geigy, WO 95/16209 (E. R. Felder, H. Rink, I. T. W. Matthews).
111. H. P. Nestler, P. A. Bartlett, W. C. Still, *J. Org. Chem.* **59** (1994) 4723.
112. M. H. J. Ohlmeyer et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 10922.
113. P. Eckes, *Angew. Chem. Int. Ed. Engl.* **33** (1994) 1573.
114. J. J. Baldwin, J. J. Burbaum, I. Henderson, M. H. J. Ohlmeyer, *J. Am. Chem. Soc.* **117** (1995) 5588.
115. J. J. Burbaum et al., *Proc. Natl. Acad. Sci. USA* **92** (1995) 6027.
116. Z. J. Ni, D. Maclean et al., *J. Med. Chem.* **39** (1996) 1601.
117. Z. J. Ni, D. Maclean, C. P. Holmes, M. A. Gallop, in J. N. Abelson (ed.): *Methods in Enzymology,* Academic Press, San Diego 1996, p. 261.
118. M. Lebl, V. Krchnak, S. E. Salmon, K. S. Lam, *Methods (San Diego)* **6** (1994) 381.
119. S. E. Salmon et al., *Proc. Natl. Acad. Sci. USA* **90** (1993) 11708.
120. P. Kocis, V. Krchnak, M. Lebl, *Tetrahedron Lett.* **34** (1993) 7251.
121. Selectide, WO 95/00165 (M. Lebl, V. Krchnak, P. Kocis).
122. N. K. Terret et al., *Bioorg. Med. Chem. Lett.* **5** (1995) 917.
123. S. H. DeWitt: "Automated Synthesis", in W. H. Moos, M. R. Pavia, A. D. Ellington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997, pp. 69 – 77.
124. S. Bormann, *Chem. Eng. News* 1996, 29 – 54.
125. P. D. Hoeprich, *Nature Biotechnology* **14** (1996) 1311 – 1312.
126. J. H. Krieger, *Chem. Eng. News* 1996, 67 – 73.
127. J. A. Ellman, *Acc. Chem. Res.* **29** (1996) 132 – 143.

128. B. A. Bunin, M. J. Plunkett, J. A. Ellman, *Proc. Natl. Acad. Sci. USA* **91** (1994) 4708 – 4712.
129. B. A. Bunin, J. A. Ellman, *J. Am. Chem. Soc.* **114** (1992) 10997 – 10998.
130. The Regents of the University of California, Oakland, US 5 288 514, 1994 (J. A. Ellman).
131. P. L. Myers, J. W. Greene, J. Saunders, S. L. Teig, *Todays Chemists at work* **6** (1997) 47.
132. J. M. Blaney, E. J. Martin, *Curr. Opin. Chem. Biol.* **1** (1997) 54.
133. E. J. Martin et al., *J. Med. Chem.* **38** (1995) 1431 – 1436.
134. W. H. Moos, G. D. Green, M. R. Pavia, *Annu. Rep. Med. Chem.* **28** (1993) 315.
135. D. E. Patterson et al., *J. Med. Chem.* **39** (1996) 3049 – 3059.
136. A. M. Ferguson, D. E. Patterson, C. D. Gary, T. L. Underiner, *J. Biomol. Screening* **1** (1996) 65.
137. S. D. Pickett, J. S. Mason, I. M. McLay, *J. Chem. Inf. Comput. Sci.* **36** (1996) 12.
138. H. Matter, *J. Med. Chem.* **40** (1997) 1219.
139. P. Willett, *Trends Biotechnol.* **13** (1995) 516.
140. L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 2280.
141. M. G. Edwards: "Combinatorial Chemistry Alliances in the 1990s: Review of deal structure", in W. H. Moos, M. R. Pavia, A. D. Ellington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997, pp. 314 – 320.
142. A front-end search engine of ReCap's RCA Database is available on the world wide web http://www.recap.com
143. D. V. Patel: "Applications of Combinatorial Technology to Drug Discovery", in W. H. Moos, M. R. Pavia, A. D. Ellington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997, pp. 78 – 89.
144. K. Bozicevic: "Patent Strategies in Molecular Diversity", in W. H. Moos, M. R. Pavia, A. D. Ellington, B. K. Kay (eds.): *Annual Reports in Combinatorial Chemistry and Molecular Diversity,* **vol. 1,** ESCOM, Leiden 1997, pp. 298 – 313.
145. S. Bormann, *Chem. Eng. News* **74** (1996) 37 – 38.
146. K. Burgess, H.-J. Lim, A. M. Porte, G. A. Sulikowski, *Angew. Chem. Int. Ed. Engl.* **35** (1996) 220.
147. S. R. Gilbertson, S. Wang, *Tetrahedron Lett.* **37** (1996) 6475.
148. Jacobsen, *J. Am. Chem. Soc.* **118** (1996) 8983.
149. G. Liu, J. A. Ellman, *J. Org. Chem.* **60** (1995) 7712.
150. K. D. Shimizu et al., *Angew. Chem.* **109** (1997) 1782 – 1785.
151. B. M. Cole et al., *Angew. Chem. Int. Ed. Engl.* **35** (1996) 1668.
152. F. C. Moates et al., *Ind. Eng. Chem.* **35** (1996) 4801 – 4803.
153. F. M. Menger, A. V. Eleseev, V. A. Migulin, *J. Org. Chem.* **60** (1995) 6666.
154. F. M. Menger, Z. X. Fei, *Angew. Chem.* **106** (1994) 329 – 331.
155. X.-D. Xiang et al., *Science* **268** (1995) 1738.
156. G. Briceno et al., *Science* **270** (1995) 273.
157. L. C. Hsieh-Wilson, X.-D. Xiang, P. G. Schultz, *Acc. Chem. Res.* **29** (1996) 164 – 170.
158. X. D. Sun, C. Gao, J. Wang, X.-D. Xiang, *Appl. Phys. Lett.* **70** (1997) 3353 – 3355.
159. B. E. Baker, N. J. Kline, P. J. Treado, M. J. Natan, *J. Am. Chem. Soc.* **118** (1996) 8721.
160. E. Danielson et al., *Nature* **389** (1997) 944.
161. E. Danielson et al., *Science* **279** (1998) 837.
162. D. E. Akporiaye, I. M. Dahl, A. Karlsson, R. Wendelbo, *Angew. Chem.* **110** (1998) 629 – 631.

Coronary Therapeutics

Martin M. Winbury, Pharmaceutical Research Division, Warner-Lambert Company, Ann Arbor, Michigan, United States; present address: Pharm/Ex, P.O. Box 8335, Ann Arbor, Michigan 48107, United States

1. **Introduction** . . . 1
2. **Anatomy and Physiology of Coronary Circulation** . . . 2
2.1. **Anatomy** . . . 2
2.2. **Physiology** . . . 2
2.3. **Pathophysiology** . . . 2
2.4. **Rationale for Treatment** . . . 3
3. **Nitrates** . . . 4
3.1. **Clinical Factors** . . . 4
3.2. **Pharmacology of Nitrates** . . . 4
3.3. **Mechanism of Relief of Angina** . . . 5
3.4. **Individual Nitrates Use** . . . 5
4. **β-Receptor Blockers** . . . 6
4.1. **Mechanism of Action** . . . 6
4.2. **Classification** . . . 6
5. **Calcium Channel Blockers (Calcium Antagonists)** . . . 7
5.1. **Cardiovascular Effects** . . . 8
5.2. **Therapeutic Use** . . . 9
5.3. **Individual Calcium Antagonists** . . . 9
6. **Large Coronary Artery Dilators** . . . 9
7. **Potent Coronary Arteriolar Vasodilators** . . . 10
8. **References** . . . 10

1. Introduction

The basic problem in coronary heart disease or coronary insufficiency is an imbalance between oxygen supply and tissue requirement. The clinical signs of coronary insufficiency include chest pain, impairment of left ventricular function, and other signs of a compromised central circulation. *Angina pectoris,* the common symptom of coronary insufficiency, is usually associated with stress, emotion, or excess oxygen requirement due to exercise. This form is known as *secondary* or *effort angina.*

There has been interest in a form of angina caused by an absolute reduction in coronary blood flow associated with a spasm of a large coronary artery classified as *primary angina* [1]. This type of angina occurs at rest or during sleep.

Both types of angina are caused by myocardial ischemia or hypoxia, or both. Therapy must be selected on the basis of the type of angina because such drugs as β-adrenergic blocking agents are ineffective in and may actually exacerbate vasospastic angina (primary angina).

Most cases of angina are associated with atherosclerotic narrowing of one or more major coronary arteries, resulting in some degree of stenosis. When the stenosis reaches a critical point, there is exhaustion of compensatory autoregulatory mechanisms involved in maintaining coronary blood flow appropriate to requirement. When the reserve capacity for coronary dilatation is lost, increased demands for oxygen cannot be achieved and coronary insufficiency results as in secondary angina. Coronary anginal episodes can also occur in the presence of nearly normal arteries with minimal stenosis as a result of vasospasm of the artery and, consequently, reduction of flow.

2. Anatomy and Physiology of Coronary Circulation

2.1. Anatomy

The anterior surface of the heart is supplied by the left coronary artery, which originates from the aorta close to the valves. The coronary artery divides into a left anterior descending artery and a circumflex artery to supply the anterior surface of the left ventricle (Fig. 1). Atherosclerosis most frequently develops in these large surface (epicardial) arteries and can ultimately produce significant stenosis. Intramural penetrating branches leave the superficial arteries to penetrate the walls of the left ventricle. These give off branches to supply the superficial portions of the myocardium, the epicardium (a), and the deeper portions, the endocardium (b). Normally the large superficial arteries and penetrating arteries serve as conduits and account for only a small fraction of the resistance to blood flow.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_001

The arterioles, however, account for a high fraction of the resistance to blood flow, and they are involved in autoregulation. However, the large conduit vessels demonstrate vasomotion, and in the presence of stenosis, vasospasm can produce a significant resistance to blood flow. Under the stimulus of chronic coronary insufficiency, collateral vessels develop to supply blood flow to a region in which the normal arterial circulation is blocked by complete stenosis.

2.2. Physiology

The coronary circulation is peculiar because it supplies the organ that generates the pressure for perfusion. During contraction (systole) the myocardial walls throttle circulation through the left ventricle. Because of the periodic compression of the coronary vessels, instantaneous coronary flow varies during each cardiac cycle. In the deeper regions of the heart, the endocardium, flow declines to zero at the height of systolic compression. Accordingly, most of the coronary blood flow occurs during dilatation (diastole) when the heart is in the relaxation phase.

The heart is dependent on flow and extracts 75 % of the oxygen from the arterial blood. An increased requirement for oxygen is normally met by an increase in blood flow through autoregulation. Balance between requirements of the tissues for oxygen and substrate and the blood supply is controlled by the arteriolar resistance vessels. When the oxygen partial pressure is decreased, adenosine is released from the tissue, effecting an increase in blood supply through local arteriolar autoregulation. This mechanism may fail in the case of a stenosis of a large artery. Under those circumstances, resting flow is maintained in regions distal to the stenosis by decreased regional arteriolar resistance, and demand for increased blood flow cannot be met because the arteriolar dilator reserve capacity is exhausted.

The heart has one of the highest oxygen requirements relative to blood flow rate. Oxygen extraction is almost maximal, and increased demands can be met only by increased blood flow. The partial pressure of oxygen in tissue and the concentration of anaerobic enzymes are lower in the subendocardium.

2.3. Pathophysiology

Coronary insufficiency is caused by an imbalance between supply and demand. Disturbances of the fine balance between supply and demand can occur (1) on the supply side by a spasm of a large coronary artery or (2) on the demand side by increased oxygen requirement that cannot be met by increased blood flow. Both conditions are associated with atherosclerotic changes, but this is not a necessary requisite in the case of a vasospastic episode.

The determinants of blood flow and oxygen supply include (1) perfusion pressure, (2) diastolic duration, (3) extravascular compression (systolic wall tension), (4) arteriolar resistance, and (5) large intramural artery tone [3]. The last is an important determination for the distribution of blood flow between the epicardium and endocardium and can be critical when autoregulation becomes exhausted. The extreme vulnerability of the subendocardium to ischemia is a result of mal-distribution. Normally, the resistance of large vessels accounts for 5 % of the total, but during maximal arteriolar dilatation, it can account for 20 % of the total resistance [2].

The requirement for blood flow and oxygen is determined by (1) heart rate, (2) contractile state, and (3) wall tension. The balance of supply and demand must be maintained for adequate oxygenation of the heart tissue.

There are two basic types of angina [1]: *primary angina,* which is caused by a reduction in blood supply due to vasospasm, and *secondary angina,* which results from an increased oxygen demand that cannot be met by an increase in blood flow. This is associated with a fixed obstruction of a major coronary artery. Although coronary blood flow is normal at rest, it cannot increase with the requirements of stress. Approximately 85 % of the patients have secondary angina, while only 15 % have primary angina. Some patients have a combination of both forms and stress may actually induce spasm. *Prinzmetal's angina* is an example of vasospastic angina with a differential electrocardiographic change compared with angina of effort [4]. Vasospastic (primary) angina occurs at rest, and secondary angina is associated with exercise, emotional changes, or stress.

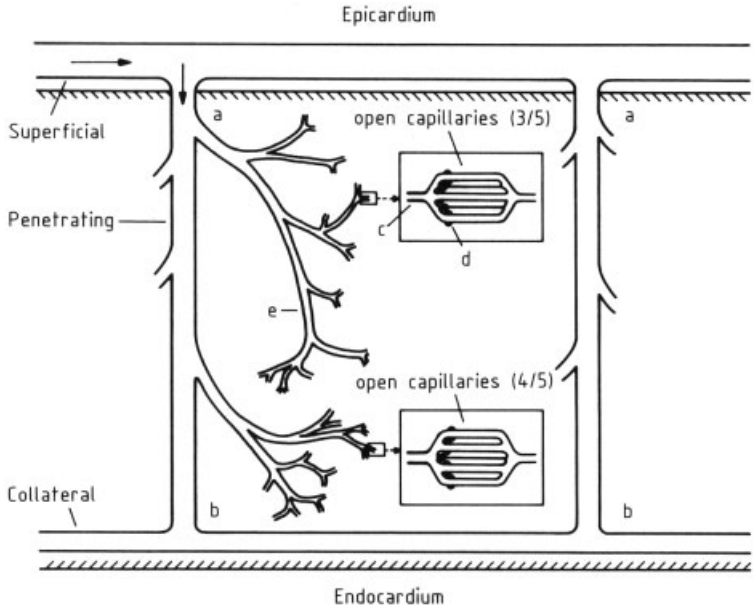

Figure 1. Arterial blood supply through left ventricular wall
a) Subepicardium; b) Subendocardium; c) Arteriole; d) Precapillary sphincter; e) Branch (modified from [2])

2.4. Rationale for Treatment

The primary function of therapy is to restore the proper balance of supply and demand. The approaches to accomplish this depend on the pathogenesis [3]. In the case of *primary angina*, one must increase blood flow by eliminating the spasm. This can be accomplished by nitrates and calcium channel blockers [4]. In the case of acute vasospasm, relief can be achieved by sublingual, intracoronary, or intravenous administration of nitrogylcerin or calcium channel blockers.

Secondary angina is relieved by reducing the oxygen demand by one of the following means:

1) decrease in heart rate and prolongation of diastole (β-adrenergic blockers);
2) decrease in heart work by (a) decreasing in contractility (β-adrenergic blockers, calcium channel blockers), (b) reducing preload by venous pooling (nitrates), or (c) reducing afterload (calcium channel blockers);
3) redistribution of blood flow to favor ischemic endocardial regions (nitrates).

Three types of agents are effective in secondary angina: (1) vasodilators of large coronary arteries, nitrates being the example; (2) β-adrenergic blockers; and (3) calcium channel blockers [3]. Potent coronary arteriolar dilators such as dipyridamole and chromonar are not effective in secondary angina. The capacity to increase blood flow in normal regions of the heart is not associated with a therapeutic effect. In fact, potent coronary arteriolar dilators may actually produce a deleterious effect known as *coronary steal* by diverting the blood flow from ischemic regions to normal regions of the heart.

Improvement of Regional Oxygen Supply. Mechanisms to improve regional oxygen supply include:

1) increase of diastolic pressure gradient by reducing wall tension and/or increasing perfusion pressure (nitrates; β-blockers have the opposite effect);
2) increase of diastolic duration (β-adrenergic blockers);
3) decrease of extravascular compression and wall tension (nitrates and calcium channel blockers; β-blockers may increase);
4) large vessel dilatation and stenosis dilatation (nitrates);
5) redistribution of blood flow to the ischemic regions (nitrates, β-adrenergic blockers, and calcium channel blockers);

6) decrease of spasm (nitrates and calcium channel blockers).

Decrease of Oxygen Requirement. Mechanisms to decrease oxygen requirement are:

1) lowering of wall tension (nitrates; β-blockers increase);
2) decrease in heart rate (β-adrenergic blockers and calcium entry blockers);
3) decrease in contractility (β-adrenergic blockers and calcium channel blockers);
4) decrease of afterload (calcium channel blockers);
5) decrease of preload (nitrates).

3. Nitrates

3.1. Clinical Factors

Nitrates have been the first agents for prophylaxis and treatment of angina pectoris [3]. Over the past 100 years, they have proven to be useful in the treatment of all forms of angina. All nitrates have a similar pharmacological activity [5]. The major differences are of a pharmacokinetic nature, primarily concerning whether or not they are absorbed rapidly by the buccal mucosa and can be used sublingually to relieve an attack.

Nitroglycerin is effective sublingually within 30 s and has a duration of ca. 30 min. Other nitrates are not rapidly active sublingually and are of no value for treatment of an anginal attack. On the other hand, all of the nitrates currently used are effective orally, including nitroglycerin, isosorbide dinitrate, isosorbide-5-mononitrate, pentaerythrityl tetranitrate, and possibly mannitol hexanitrate and erythrityl tetranitrate; the latter two are less important.

In addition to the sublingual or oral route, nitrates are also administered intravenously or they are injected directly into the coronary artery in cases of vasospastic angina (primary angina) or congestive heart failure.

Transdermal preparations of nitroglycerin and isosorbide dinitrate are also available [6]. Percutaneous ointment of nitroglycerin was available many years ago but fell into disrepute. According to more recent studies, percutaneous administration provides protection for at least 4 – 6 h. Patients can be protected during the waking hours on a TID (three times a day) basis. Transdermal devices have been developed that permit absorption over a period of more than 24 h. However, protection over 24 h is questionable. This is not due to lack of constant blood level, but tolerance may develop. Peaks and troughs in blood levels are obviously necessary to provide protection over a prolonged period.

Headache is the major side effect of the nitrates. The existence of a tolerance to the nitrates is controversial. Certainly tolerance to the headache may develop, and there may be some degree of tolerance to the beneficial effect; however, this effect is not of sufficient magnitude to prevent its use.

3.2. Pharmacology of Nitrates

The most prominent pharmacological action of the nitrates is relaxation of vascular smooth muscle. The vasodilator action is possibly caused by bonding to sulfhydryl groups on the nitrate receptors in vascular smooth muscle; this increases the concentration of cyclic guanosin monophosphate, leading to vasodilatation [7].

The nitrate action on the peripheral vascular system appears to be more selective for the venous portion with increased distensibility. Venous dilatation results in a decrease in venous return and preload, enabling a decrease in the size and pressure of the left ventricle. In the coronary system, nitrates are weak transient dilators of small resistive vessels (arterioles), but they have a selective prolonged dilator action on the large coronary arteries [8], [9].

3.3. Mechanism of Relief of Angina

The venous dilator action decreases left ventricular size and oxygen consumption. Another factor that contributes to the beneficial effect is selective dilatation of large coronary arteries and collateral coronary vessels; this redistributes blood flow and increases perfusion of the ischemic endocardium [2]. In addition, the nitrates reverse coronary spasm associated with the large coronary arteries.

In summary, the nitrates relieve angina pectoris by decreasing myocardial oxygen demand,

by decreasing left ventricular wall tension with enhanced perfusion of the subendocardium and, in the case of vasospastic angina, by reversing the coronary spasm of large vessels [10].

3.4. Individual Nitrates Use

The nitrates that are used therapeutically include nitroglycerin, isosorbide dinitrate, isosorbide-5-mononitrate, pentaerythrityl tetranitrate, mannitol hexanitrate, and erythrityl tetranitrate.

Nitroglycerin [*55-63-0*], 1,2,3-propanetriol trinitrate, glyceryl trinitrate, $C_3H_5N_3O_9$, M_r 227.09, *mp* 2.8 °C (labile form) or 13.5 °C (stable form). For its preparation, see [11–16].

Trade Names. Nitro-bid (Marion), Nitrodisk (device; Searle), Nitrodur (device; Key), Nitrostat (Parke-Davis), Transderm-nitro (device, Ciba-Geigy), Deponit (Pharma-Schwarz), Nitrolingual (Pohl).

Isosorbide dinitrate [*87-33-2*] , 1,4 : 3,6-dianhydro-D-glucitol dinitrate, $C_6H_8N_2O_8$, M_r 236.14, *mp* 70 °C, is prepared from sorbitol [17].

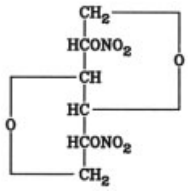

Trade Names. Isoket (Pharma-Schwarz), Iso Mack (Mack), Isordil (Wyeth), Sorbitrate (Stuart).

Isosorbide-5-mononitrate [*16051-77-7*], 1,4 : 3,6-dianhydro-D-glucitol-5-nitrate, $C_6H_9NO_6$, M_r 191.14, *mp* 90 – 91 °C; for its preparation see [18–20].

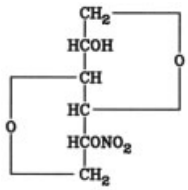

Trade Names. Elantan (Pharma-Schwarz), ISMO (Boehringer-Mannheim), MonoMack (Mack).

Pentaerythrityl tetranitrate [*78-11-5*], 2,2-bis[(nitrooxy)methyl]-1,3-propanediol dinitrate, pentaerythritol tetranitrate, $C_5H_8N_4O_{12}$, M_r 316.15, *mp* 140 °C, is made by nitration of pentaerythritol [21], [22].

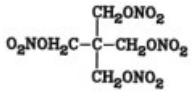

Trade Names. Peritrate (Parke-Davis), Dilcoran (Gödecke).

Mannitol hexanitrate [*15825-70-4*], mannitol nitrate, nitromannite, nitromannitol, $C_6H_8N_6O_{18}$, M_r 452.17, *mp* 106 – 108 °C, is made by nitration of mannitol [22], [23].

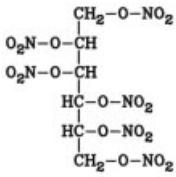

Trade Name. Not approved in the United States; Manexin (Beecham).

Erythrityl tetranitrate [*7297-25-8*], erythritol tetranitrate, erythrol tetranitrate, $C_4H_6N_4O_{12}$, M_r 302.12, *mp* 61 °C, is made by nitration of erythritol [22].

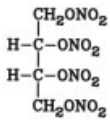

Trade Name. Cardilate (Wellcome).

4. β-Receptor Blockers

In step care therapy for the treatment of ischemic heart disease, the second stage is β-adrenergic blockage [3]. The drugs became available about two decades ago and are an important group in treatment of patients with ischemic heart disease (and hypertension) [24–27]. β-Blockage attenuates the β-adrenergic influences on the heart,

which include increased heart rate and increased contractility, both of which increase myocardial oxygen demand. With a compromised coronary circulation, coronary insufficiency results. β_1-Receptors are found in the heart, and β_2-receptors are found in bronchioles and vascular smooth muscle. During exercise and stress, adrenergic tone is increased, resulting in excessive stimulation of the heart, increase in heart rate, and possible decrease in coronary blood flow as a result of α-adrenergic stimulation.

4.1. Mechanism of Action

Like the nitrates, the β-blockers have multiple actions. They primarily act by blocking the β_1-receptors in the heart. This decreases the heart rate and contractility at rest and during excercise and also prolongs the diastolic perfusion time. The increase in the diastolic perfusion index permits redistribution with greater perfusion of the deeper endocardial regions of the ischemic area of the heart. The reduction in heart rate, contractility, and blood pressure then lowers the oxygen requirement. Therefore, β-adrenergic blockage reduces myocardial oxygen requirement and enhances blood flow to the ischemic region.

However, the mechanism of action comprises some negative aspects. Because of the decreased contractility, β-blockage can lead to some degree of heart failure. In addition, asthmatic patients can encounter bronchoconstriction because of the β_2-blocking action. There may also be an increase in coronary arteriolar tone when the β_2 influence is withdrawn. Likewise, peripheral vascular problems may arise because of vasoconstriction of the extremities.

4.2. Classification

Quantitative differences among the various β-blockers are due to such selective properties as cardioselectivity, intrinsic sympathomimetic activity (ISA), and membrane stabilizing activity. The last is of little consequence in treatment of angina because it refers to antiarrhythmic activity.

Cardioselectivity is determined by the balance between β_1- and β_2-receptor blockage. Cardioselective agents are selective for β_1-receptors and are less liable to produce side effects in patients with chronic lung disease, asthma, or peripheral vascular disease as compared to noncardioselective blockers.

Intrinsic sympathomimetic activity (ISA) or partial agonist activity relates to the β-agonist activity. Theoretically, compounds with ISA are less likely to cause extreme bradycardia or cardiac depression than compounds without ISA. Compounds with ISA are also supposed to be less prone to produce bronchoconstriction. However, the relative importance of this aspect has not been demonstrated in clinical use.

Noncardioselective agents without ISA include propranolol [*318-98-9*], with ISA are less likelytimolol [*91524-16-2*], sotalol [*959-24-0*], and nadolol [*42200-33-9*]; noncardioselective agents with ISA are oxprenolol [*6452-71-7*], alprenolol [*13707-88-5*], and pindolol [*13523-86-9*].

Cardioselective agents without ISA are metoprolol [*37350-58-6*], atenolol [*29122-68-7*], and those with ISA are acebutolol [*37517-30-9*] and practolol [*6673-35-4*]. The physical properties and preparation of these drugs are treated in a separate article (→ β-Receptor Blocking Agents).

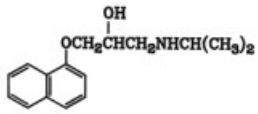

Propranolol

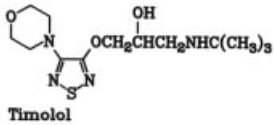

Timolol

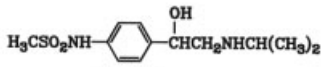

Sotalol

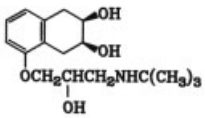

Nadolol

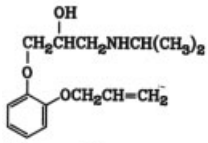

Oxprenolol

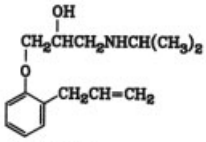

Alprenolol

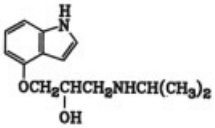

Pindolol

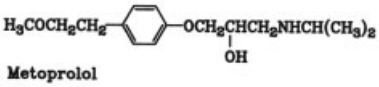

Metoprolol

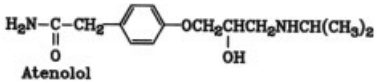

Atenolol

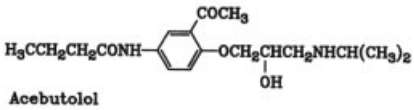

Acebutolol

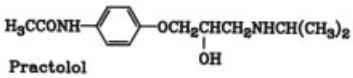

Practolol

The drugs have different durations of action, and those that are considered long-acting include sotalol, nadolol, atenolol, and practolol. The choice of an agent for a particular patient is based primarily on the factors of duration of action and cardioselectivity.

5. Calcium Channel Blockers (Calcium Antagonists)

The calcium channel blockers have been investigated intensively in the past decade. The first was verapamil, which was followed by nifedipine and diltiazem [28]. Originally, verapamil was thought to be a coronary dilator and possibly to have β-blocking action, and nifedipine was classified primarily as a potent coronary dilator. According to the work of Fleckenstein [29], the mechanism of action of these compounds is *calcium antagonism.* Later the term *calcium entry blocker*, or *calcium channel blocker,* was adopted because these compounds inhibit the slow influx of calcium ions into the cardiac muscle and reduce the transcellular passage of calcium ions in vascular smooth muscle. In contrast to the β-adrenergic blockers, these compounds are chemically heterogeneous and have differing pharmacological, electrophysiological, and hemodynamic properties [30]. Their common feature is blockage of calcium entry into the cell.

In both cardiac muscle and vascular smooth muscle, calcium ions are responsible for the coupling of excitation to contraction. Blockage of this coupling causes relaxation of vascular smooth muscle and reduced myocardial contractility. However, uncoupling in the cardiac muscle can finally result in cardiac failure. Some of the agents are unique in the treatment of vasospastic angina (primary angina) and effort angina (secondary angina) because the vascular smooth muscle is very sensitive to their effects [28], [29].

5.1. Cardiovascular Effects

All of the calcium channel blockers exert negative inotropic effects on the isolated myocardium, but this does not appear to be a significant factor in therapeutic use. Lack of a negative inotropic effect in humans is attributed to the reflex sympathetic discharge associated with the vasodilator action of these agents. Furthermore, the dose that depresses the heart is usually ten times that producing coronary dilatation. Myocardial oxygen consumption is usually decreased primarily as a result of decreased afterload (peripheral vasodilatation) and the small negative inotropic effect.

Calcium channel blockers dilate the small and the large coronary arteries [30]. This distinguishes them from the potent coronary dilators, which act primarily on the small vessels, and the nitrates, which act primarily on the larger vessels. In addition, calcium channel blockers reverse the spasm of the large coronary arteries that occurs in vasospastic angina or is induced by ergonovine [4], [28]. Blood flow to the normal and ischemic regions can be increased, indicating that the calcium channel blockers can increase collateral blood flow. Finally, calcium channel blockers have potent peripheral dilator actions and reduce peripheral resistance.

Verapamil has some distinctive electrophysiologic effects; it reduces conduction through the atrioventricular node. As a result, verapamil is exquisitely useful in the treatment of paroxysmal

supraventricular tachycardias [28], [30]. Other calcium channel blockers, such as nifedipine and diltiazem, do not have this property and are not useful in treatment of this arrhythmia. Resting heart rate may be increased by nifedipine, decreased by diltiazem, and scarcely affected by verapamil.

5.2. Therapeutic Use

Calcium channel blockers are recommended for angina pectoris of all types, including primary and secondary angina. Because they prevent or reduce a spontaneously occurring spasm of large coronary arteries, the calcium channel blockers are recommended for the treatment of Prinzmetal's angina and other vasospastic phenomena. A vasospasm may also be involved in angina of effort, but there it primarily acts by reducing myocardial oxygen consumption attributed to reduced afterload and decreased contractility.

5.3. Individual Calcium Antagonists

Many agents show some degree of calcium entry blockage, but the most clinically useful drugs are selective for calcium entry (type A agents) [31–34]. Included in this group are verapamil [*152-11-4*], nifedipine [*21829-25-4*], diltiazem [*33286-22-5*], and other more recently developed compounds such as nimodipine [*66085-59-4*], nitrendipine [*39562-70-4*], nicardipine [*54527-84-3*], and nisoldipine [*63675-72-9*].

Type B calcium channel blockers also block sodium entry; they include prenylamine [*390-64-7*], fendiline [*13042-18-7*], and perhexiline [*6621-47-2*] [35–37]. Lidoflazine [*3416-26-0*] and bepridil [*74764-40-2*] [38–40] may be included as well. The latter two compounds were initially developed as conventional coronary dilators, but later were shown to have some calcium antagonist activity. The physical properties and the preparation of these drugs are treated in a separate article in this Encyclopedia (→ Calcium Antagonists).

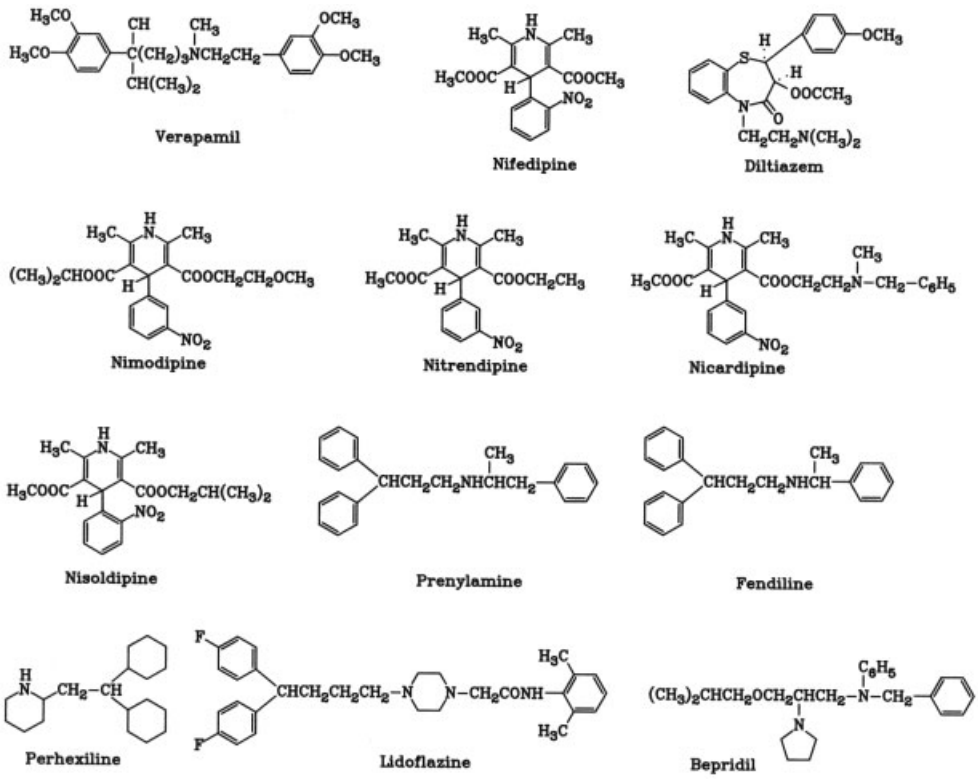

6. Large Coronary Artery Dilators

Several new agents selectively affect the large coronary arteries like the nitrates [41]. The effects on the smaller coronary arteries are modest and probably of little significance in ischemic regions. Molsidomine, trapidil, and nicorandil are successful in treating secondary angina. They have different chemical structures but common pharmacological features that contribute to their antianginal efficacy. All relax the large coronary arteries and collateral blood vessels. Perfusion of the subendocardial region of the ischemic zones is enhanced because of redistribution in favor of the deeper myocardial regions. Like nitroglycerin, these compounds increase venous capacity and reduce myocardial preload and myocardial oxygen consumption [42–50]. Their effectiveness in vasospastic angina has not yet been clearly demonstrated.

Additional effects of nicorandil are (1) a reduction in myocardial infarct size following coronary occlusion, (2) a reduction in the acidosis and lactate content of the ischemic region, and (3) reversal of spasm of the large coronary artery produced by acetylcholine. Trapidil also inhibits platelet aggregation and has a hypolipidemic effect. Molsidomine has been demonstrated to have an antithrombotic effect. These ancillary actions may contribute to the overall antianginal effect, but the major contributions presumably come from the large coronary artery dilatation and redistribution, which favor subendocardial perfusion.

Molsidomine [*25717-80-0*], N-(ethoxy/carbonyl)-3-(4-morpholinyl)sydnone imine, *N*-carboxy-3-morpholinosydnonimine ethyl ester, $C_9H_{14}N_4O_4$, M_r 242.23, *mp* 140 – 141 °C; for its preparation, see [51], [52].

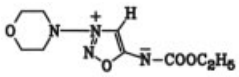

Trade Names. Corvaton (Cassella-Riedel), Molsidolat (Hoechst), Morial (Takeda), Motazomin (Takeda).

Trapidil [*15421-84-8*], *N,N*-diethyl-5-methyl-[1,2,4]triazolo[1,5-*a*]pyrimidin-7-amine, 7-diethylamino-5-methyl-*s*-triazolo[1,5-*a*]pyrimidine, $C_{10}H_{15}N_5$, M_r 205.27, *mp* 98 – 99.4 °C, 102 – 104 °C (from heptane); for its preparation, see [53], [54].

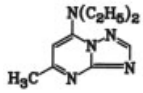

Trade Name. Rocornal (VEB Deutsches Hydrierwerk, Mochida).

Nicorandil [*65141-46-0*], *N*-[(2-nitrooxy)-ethyl]-3-pyridinecarboxamide, *N*-(2-hydroxyethyl)nicotinamide nitrate (ester), $C_8H_9N_3O_4$, M_r 211.17, *mp* 92 – 93 °C; for its preparation, see [55], [56].

Trade Name. Sigmart (Chugai).

7. Potent Coronary Arteriolar Vasodilators

The class of highly effective coronary arteriolar dilators includes dipyridamole, which was the first drug of this kind, and chromonar. Both compounds relax the coronary arterioles (resistance vessels) selectively, which leads to maximum blood flow of the normal regions. However, they do not increase blood flow in the ischemic regions. Because the normal regions are preferred, these drugs can actually cause "coronary steal" by diverting blood flow from the ischemic regions to the normal regions [57], [58]. These compounds are not effective as antianginal agents, but dipyridamole now has a role in preventing platelet aggregation and is used prophylactically for the treatment of cerebral ischemic attacks [59]. Because of the "coronary steal" phenomenon, dipyridamole has been used as a diagnostic agent for patients with coronary artery disease [60].

According to animal studies, the maximum dilatation of the coronary bed with chromonar or dipyridamole prevents normal coronary autoregulation. In addition, the enhancement of

blood flow to the subepicardial region produces a reverse redistribution, which can exacerbate subendocardial ischemia [61].

Dipyridamole [*58-32-2*], 2,2′,2″,2‴-(4,8-dipiperidinopyrimido[5,4-*d*]pyrimidine-2,6-diyldinitrilo)tetraethanol, $C_{24}H_{40}N_8O_4$, M_r 504.62, *mp* 163 °C, is treated in more detail under → Cardiovascular Drugs,

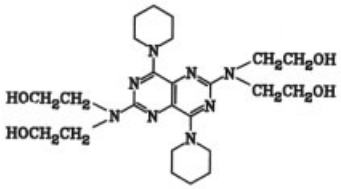

Chromonar [*655-35-6*], [[3-[2-(diethylamino)-ethyl]-4-methyl-2-oxo-2*H*-1-benzopyran-7-yl]-oxy] acetic acid ethyl ester, coumadin, $C_{20}H_{27}NO_5$, M_r 361.42, *mp* 159 – 160 °C; for its preparation, see [62].

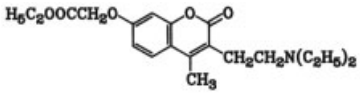

Trade Names (hydrochloride). Antiangor (I.S.O.M., Milano), Intenkordin (Polfa), Intensain (Cassella-Riedel).

8. References

1. A. Maseri in A. Maseri, G. A. Klassen, M. Lesch (eds.): *Preface XIII – XIV in Primary and Secondary Angina Pectoris,* Grune & Stratton, New York 1978.
2. M. M. Winbury, B. B. Howe, M. A. Hefner, *J. Pharmacol. Exp. Ther.* **168** (1969) 70 – 95.
3. E. H. Sonnenblick, *Cardiovas. Rev. Rep.* **2** (1982) 1283 – 1289.
4. J. H. Stein, J. A. Ambrose, B. D. King, M. V. Herman, *Cardiovas. Rev. Rep.* **3** (1982) 1297 – 1314.
5. J. Abrams, *Am. J. Med.* **74** (1983) 85 – 94.
6. V. Elkayam, W. S. Aranow, *Drugs* **23** (1982) 165 – 194.
7. P. Needleman (ed.): *Organic Nitrates,* Springer Verlag, Berlin 1975.
8. B. G. Brown, *Z. Kardiol. Suppl.* **72** (1983) no. 3, 77 – 81.
9. B. G. Brown, E. L. Bolson, R. P. Petersen, C. D. Pierce et al., *Circulation* **64** (1981) 1089 – 1097.
10. P. R. Lichtlen, H. J. Engel, A. Schrey, H. J. C. Swan (eds.): *Nitrates III,* Springer Verlag, Berlin 1981.
11. Sobrero, *Ann. Chem. Pharm.* **64** (1847) 398.
12. Williamson, *Ann. Chem. Pharm.* **92** (1854) 305.
13. Korea Explosives Co., GB 1 567 918, 1980 (K.-S. Hwang, J. Kim, B. D. Kim).
14. J. Meissner GmbH, DE 2 820 890, 1979 (H. J. Gebauer).
15. C. D. Marken, C. E. Kristofferson, M. M. Roland, *Synthesis* 1977, no. 7, 484 – 485.
16. G. S. Biasutti, *ACS Symp. Ser.* 1976, no. 22, 320 – 326.
17. L. A. Silvieri, N. J. DeAngelis in K. Florey (ed.): *Analytical Profiles of Drug Substances,* **vol. 4,** Academic Press, New York 1975, pp. 225 – 244.
18. H. Mack, US 4 371 703, 1983 (P. Stross).
19. Boehringer-Mannheim, US 4 431 829, 1984 (E. Kiegel, K. Lauer).
20. Cassella, US 4 431 830, 1984 (K. Schönafinger).
21. Du Pont, US 2 370 437, 1945 (M. F. Acken, J. C. Vyverberg, Jr.).
22. Du Pont, US 2 545 538, 1951 (R. E. Barnhart, R. M. Cavanaugh).
23. G. Fleury, L. Brissaud, P. Lhoste, *Mem. Poudres* **31** (1949) 107.
24. J. B. Kostis, C. R. Lacy, S. D. Krieger, N. M. Cosgrove, *Am. Heart J.* **108** (1982) 1131 – 1136.
25. D. L. Johnston, D. E. Manyari, W. J. Kostuk, *Can. Med. Assoc. J.* **130** (1984) 1449 – 1455.
26. S. H. Taylor, *Am. J. Cardiol.* **52** (1983) 34 – 42.
27. H. Vik-Mo, P. R. Maroko, L. G. Ribeiro, *J. Am. Coll. Cardiol.* **4** (1984) 735 – 741.
28. A. G. Ellrodt, B. N. Singh, *Pharmacol. Ther.* **23** (1983) 1 – 43.
29. A. Fleckenstein (ed.): *Calcium Antagonism in Heart and Smooth Muscle,* J. Wiley & Sons, New York 1983.
30. P. D. Henry, *Am. J. Cardiol.* **46** (1980) 1047 – 1058.
31. K. R. Karsch, P. Niemozyk, W. Voelker, L. Seibel, *Z. Kardiol.* **73** (1984) 552 – 559.
32. G. DiPasquale, A. M. Lusa, G. L. Manini, P. Dominici et al., *G. Ital. Cardiol.* **14** (1984) 618 – 620.
33. T. Ishikawa, T. Imamura, Y. Koiwaya, K. Tanaka, *Clin. Cardiol.* **7** (1984) 606 – 610.

34. N. S. Khurmi, M. J. Bowles, M. J. O'Hara, V. Subramanian et al., *Am. J. Cardiol.* **54** (1984) 738 – 743.
35. J. Milei, A. Vasquez, N. J. Bolomo, *Cardiology* **69** (1982) 192 – 202.
36. N. Schafer, G. G. Belz, M. Strauch, B. Schneider, *DMW Dtsch. Med. Wochenschr.* **105** (1980) 1253 – 1258.
37. H. D. White, J. B. Lowe, *Int. J. Cardiol.* **3** (1982) 145 – 155.
38. M. D. Jenkins, E. A. Swallow, O. Abdel-Hadi, P. H. Kidner, *Am. J. Cardiol.* **48** (1981) 140 – 146.
39. R. DiBianco, J. Albert, R. J. Katz, J. Spann et al., *Am. J. Cardiol.* **53** (1984) 35 – 41.
40. R. J. Marshall, A. W. Muir, *Br. J. Pharmacol.* **73** (1981) 471 – 479.
41. J. Holtz, M. Giesler, E. Bassenge, *Z. Kardiol.* **72** (1983), Suppl. 3, 98 – 106.
42. W. A. Fach, H. J. Becker, *Z. Kardiol.* **73** (1984) 613 – 622.
43. M. Romano, L. de Caprio, P. Meccariello, S. Cuomo et al., *Int. J. Clin. Pharmacol. Ther. Toxicol.* **22** (1981) 246 – 249.
44. S. Witchitz, H. Kolsky, P. Moisson, *Ann. Cardiol. Angeiol.* **32** (1983) 539 – 543.
45. M. Takahashi, S. Sasayama, M. Nakamura, K. Hirose et al., *Arzneim. Forsch.* **32** (1982) 214 – 218.
46. K. Yamaguchi, K. Suzuki, T. Niho, M. Sato et al., *J. Cardiovasc. Pharmacol.* **5** (1983) 499 – 505.
47. Y. Koretsune, K. Kodama, S. Nanto, K. Ishikawa et al., *Jpn. Circ. J.* **47** (1983) 391 – 399.
48. K. Sakai, Y. Shiraki, H. Nabata, *J. Cardiovasc. Pharmacol.* **3** (1981) 139 – 150.
49. K. Satoh, M. Maruyama, S. Yamashita, N. Taira, *Jpn. Heart J.* **25** (1984) 267 – 273.
50. J. Thormann, M. Schlepper, W. Kramer, W. Gottwik, *Z. Kardiol.* **71** (1982) 743 – 753.
51. Takeda, JP 7 006 265, 1970 (K. Masuda, Y. Imashiro); US 3 769 283, 1973; US 3 812 128, 1974.
52. K. Masuda, T. Kamiya, Y. Imashiro, T. Kaneko, *Chem. Pharm. Bull.* **19** (1971) 72.
53. E. Tenor, H. Fueller, F. Hausschild, DD 55 956, 1967; *Chem. Abstr.* **67** (1967) 90 830 f. E. Tenor, R. Ludwig, *Pharmazie* **26** (1971) 534 – 539.
54. S. Pfeifer, I. Mann, A. Wierer, *Pharmazie* **26** (1971) 539.
55. Chugai, DE 2 714 713, 1977 (H. Nagano, T. Mori, S. Takaku, I. Matsunaga et al.).
56. Chugai, JP 7 809 323, 1978; US 4 200 640, 1980 (T. Kujirai, I. Matsunaga, T. Mori, H. Nagano et al.).
57. P. Lacroix, P. H. Linee, J. B. Le Polles, *J. Pharmacol. Exp. Ther.* **204** (1978) 645 – 654.
58. D. Harris, D. Taylor, B. Condon, D. Ackey et al., *Eur. J. Nucl. Med.* **7** (1982) 1 – 5.
59. M. P. Rivey, M. R. Alexander, J. W. Taylor, *Drug Intell. Clin. Pharm.* **18** (1984) 869 – 880.
60. L. Tavazzi, M. Previtali, J. A. Salerno, M. Chimienti et al., *Cardiology* **69** (1982) 34 – 41.
61. G. J. Gross, J. D. Buck, D. C. Warltier, H. F. Hardman, *J. Pharmacol. Exp. Ther.* **222** (1982) 635 – 640.
62. Cassella, US 3 282 938, 1966 (H. Ritter, R. Beyerle, R. Nitz).

Cough Remedies

Bronchospasmolytics are treated under → Antiasthmatic Agents

Oskar Bub, Knoll AG, Ludwigshafen Federal Republic of Germany

Ludwig Friedrich, Knoll AG, Ludwigshafen Federal Republic of Germany

1. Introduction 1
2. Antitussives 2
2.1. Antitussives Acting on the Central Nervous System 2
2.1.1. Alkaloids and Derivatives 2
2.1.2. Morphinans 3
2.1.3. Arylalkylamines 4
2.1.4. Basic Esters 5
2.1.5. Basic Ethers 7
2.1.6. Piperazine Derivatives 8
2.1.7. Other Compounds 9
2.2. Antitussives Acting Mainly Peripherally 9
3. Expectorants 10
3.1. Secretolytics and Secretomotorics . 10
3.2. Mucolytics 10
3.2.1. Cysteine Derivatives 11
3.2.2. Benzylamines 11
4. References 12

1. Introduction [1–4]

The respiratory tract is an integrated system in terms of anatomy and function. Ailments of the respiratory tract are among the most frequent diseases and tend to affect the whole airway system.

Cough is a normal physiological response to an irritation in the laryngeal–tracheal–bronchial system; it may be caused by a pathogen or by mechanical or chemical stimulation, and is vital in expelling mucus secreted in response to these stimuli. Therefore, normal and productive cough, which is useful to the patient and needs no treatment, is distinguished from nonproductive cough, which can be painful and fatiguing and should be suppressed by antitussive drugs.

Cough remedies are also used as concomitant therapy in a variety of respiratory tract diseases in which cough is a problem. These range from sore throat to more serious complaints such as bronchitis, pneumonia, and asthma.

In the treatment of whooping cough caused by *Bordatella pertussis* infection, antitussives only alleviate paroxysmal coughing.

Symptomatic cough therapy is directed at either suppressing reflex stimulation of the cough center by administering *antitussives* (Chap. 2) or promoting the removal of mucus with *expectorants* (Chap. 3). *Bronchospasmolytics* are used in patients with bronchospasm (→ Antiasthmatic Agents) and *sympathomimetics* help to shrink swollen mucous membranes in cases of inflammation.

The etiology of respiratory tract diseases is often very complex and frequently requires therapy with a combination drug. Expectorants may be combined with bronchospasmolytics, but not with antitussives because this would suppress the cough reflex and the excessive mucus could not be expelled.

Antihistamines are major ingredients of many cough preparations (→ Antiallergic Agents). In addition to their antisecretory effect, antihistamines may have a beneficial effect on allergy-induced coughing; however, they have no antitussive effect of their own.

2. Antitussives

2.1. Antitussives Acting on the Central Nervous System

Antitussives chiefly act on the central nervous system (CNS). They reduce the number and in-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_013

tensity of coughing bouts by inhibiting the cough reflex in the cough center of the brain stem or by blocking sensitive receptors in the bronchi.

Morphine, codeine, and narcotine, isolated from opium (→ Alkaloids), were the first compounds of this group to be used. In addition to their analgesic effect, these drugs also have an antitussive potency.

Morphine is not suited for use as an antitussive because of its adverse effects (respiratory depression, constipation, addictive potential). Its analgesic potency and side effects were attenuated by etherifying its phenolic hydroxyl group (→ Analgesics and Antipyretics), but its antitussive potency was not affected by this substitution.

The methyl ether codeine is the most popular antitussive. Other morphine ethers have also been used as antitussives (e.g., ethylmorphine or pholcodine).

Chemical modification of the morphine nucleus has led to derivatives of greater analgesic and antitussive activity, but these derivatives still have morphine-like side effects (e.g., dihydrocodeine or hydrocodone). Total synthesis of the morphine skeleton by Grewe in 1946 made the morphinans accessible; their (+)-isomers are potent antitussives that do not lead to addiction.

2.1.1. Alkaloids and Derivatives
(→ Alkaloids)

Codeine [*76-57-3*], $C_{18}H_{21}NO_3$, M_r 299.36; monohydrate [*6059-47-8*], M_r 317.38, *mp* 154–156 °C, $[\alpha]_D^{15}$ −136° (*c* = 2; ethanol); hydrochloride dihydrate [*1422-07-7*], $C_{18}H_{22}ClNO_3 \cdot 2\,H_2O$, M_r 371.9, *mp* 280 °C, $[\alpha]_D^{22}$ −108° (H_2O); phosphate [*52-28-8*], $C_{18}H_{24}NO_7P$, M_r 397.38; sulfate [*1420-53-7*], $C_{36}H_{44}N_2O_{10}S$, M_r 696.82.

Codeine occurs naturally in opium (0.3 %), but most of it is prepared by methylation of morphine with trimethylphenylammonium hydroxide [5]:

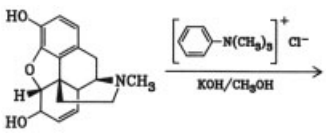

Trade Names. Codicept (Sanol), Codeinfos (Union Quimico-Farmaceutica), Tricodein (Solco, Zyma), Codipertussin (Taeschner); combinations: Codipront (Mack), Tussoretard (Klinge), Codicaps (Thiemann), Empracet (Burroughs Wellcome), Copavin (Lilly), and many others.

Ethylmorphine [*76-58-4*], codethyline, dionin, $C_{19}H_{23}NO_3$, M_r 313.38, *mp* 199–201 °C; hydrochloride dihydrate [*125-30-4*], $C_{19}H_{24}ClNO_3 \cdot 2\,H_2O$, *mp* 123 °C (decomp.).

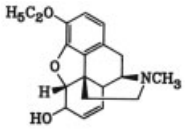

Ethylmorphine is prepared by ethylation of morphine with ethyl benzenesulfonate in the presence of potassium hydroxide [6]. The pharmacological effects of ethylmorphine are very similar to those of codeine; ethylmorphine is slightly less effective as an antitussive [7].

Trade Names. Codethyline (Houdé), Trachyl (Beytout); combinations: Expectorans Solucampher (Delalande), Tussedat (Sagitta).

Pholcodine [*509-67-1*], 3-*O*-(2-morpholinoethyl)morphine, $C_{23}H_{30}N_2O_4$, M_r 398.49, *mp* 91 °C (monohydrate), $[\alpha]_D^{20}$ −95.3° (*c* = 2; ethanol).

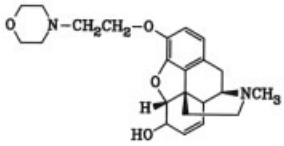

For the synthesis of pholcodine, morphine is alkylated with *N*-(2-chloroethyl)morpholine in the presence of sodium hydroxide [8]. Pholcodine is a centrally acting antitussive of similar potency as codeine; it has minimal side effects and no addictive potential [9].

Trade Names. Ethnine (Allen and Hanburys, Purdue Frederick), Folcodan (L.I.B.S.), Memine (Macfarlan Smith), Tussokon (Pharmacia).

Dihydrocodeine [*125-28-0*], drocode, $C_{18}H_{23}NO_3$, M_r 301.37, *mp* 112–113 °C; hydrogen tartrate [*5965-13-9*], $C_{22}H_{29}NO_9$, M_r 451.5, *mp* 192–193 °C.

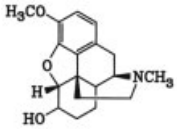

Dihydrocodeine is made by catalytic hydrogenation of codeine over platinum, palladium, or nickel catalysts [10], or by methylation of dihydromorphine with dimethyl sulfate [11]. Dihydrocodeine has antitussive and analgesic properties similar to those of morphine, with fewer side effects [6], [12].

Trade Names. Paracodin (Knoll), Remedacen (MüllerRorer), Rikodeine (Riker), Tiamon (Temmler; combination).

Hydrocodone [*125-29-1*], dihydrocodeinone, $C_{18}H_{21}NO_3$, M_r 299.38, *mp* 198 °C; hydrochloride hydrate [*25968-91-6*], $C_{18}H_{22}ClNO_3 \cdot H_2O$, M_r 353.83, *mp* 185 – 186 °C (decomp.), $[\alpha]_D^{27}$ – 130° (H_2O); bitartrate hemipentahydrate [*34195-34-1*], $C_{22}H_{27}NO_9 \cdot 2.5\,H_2O$, M_r 494.5.

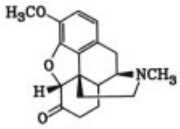

For the synthesis of hydrocodone, codeine is catalytically rearranged by heating with palladium or platinum in alcoholic or aqueous acidic solution [13]. For other methods of preparation, see [3, vol. 1, p. 120]. Hydrocodone is a stronger antitussive and analgesic than codeine, but it also has morphine-like side effects and addictive potential [6].

Trade Names. Dicodid (Knoll), Didrate (Penick), Novocodina (Erba); numerous combinations.

Noscapine [*128-62-1*], narcotine, (3 *S*)-6,7-dimethoxy-3-[(1 *R*)-1,2,3,4-tetrahydro-8-methoxy-2-methyl-6,7-methylenedioxy-1-isoquinolyl]phthalide, $C_{22}H_{23}NO_7$, M_r 413.43, *mp* 176 °C; hydrochloride monohydrate [*912-60-7*], $C_{22}H_{24}ClNO_7 \cdot H_2O$, M_r 467.9, *mp* 200 °C (decomp.).

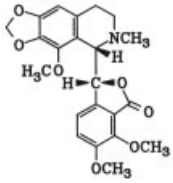

Noscapine is a phthalide alkaloid which occurs in amounts up to about 6 % in opium. It is a centrally acting antitussive of nearly the same potency as codeine, but without morphine-like side effects and with additional bronchodilatory activity [14].

Trade Names. Capval (Dreluso), Lyobex (Lappe), Narcotussin (Biologici, Italia), Tusscapine (Fisons); combinations:Tussoretard (Klinge), Tiamon (Temmler), and many others.

2.1.2. Morphinans

Dextromethorphan [*125-71-3*], (+)-3-methoxy-*N*-methylmorphinan, $C_{18}H_{25}NO$, M_r 271.41; hydrobromide monohydrate [*6700-34-1*], $C_{18}H_{26}BrNO \cdot H_2O$, M_r 370.34, *mp* 122 – 124 °C, $[\alpha]_D^{20}$ +27.6° (c = 1.5; H_2O).

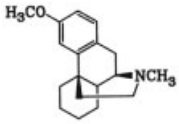

Dextromethorphan is made by resolution of (±)-3-hydroxy-*N*-methylmorphinan (racemorphan; → Analgesics and Antipyretics) with D-tartaric acid and subsequent O-methylation of the (+)-isomer with phenyltrimethylammonium hydroxide [15]. Dextromethorphan has an antitussive effect similar to that of codeine with fewer side effects; it does not lead to addiction [16].

Trade Names. Romilar (Hoffmann-La Roche), Cosylan (Parke-Davis), Dormethan (Dorsey), Husmedin (Toho), Methorcon (Kowa); numerous combinations.

Dimemorfan [*36309-01-0*], 3,17-dimethylmorphinan, $C_{18}H_{25}N$, M_r 255.41, *mp* 90 – 93 °C; phosphate, $C_{18}H_{28}NO_4P$, M_r 353.41, *mp* 267 – 269 °C, $[\alpha]_D^{23}$ +25.7° (c = 0.5; methanol).

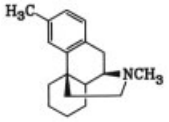

Dimemorfan is prepared by using Grewe's morphinan synthesis [17]. It is 1.5 times as potent an antitussive as dextromethorphan, with some analgesic activity and minimal side effects [18].
Trade Names. Astomin (Yamanouchi), Dastosin (Morrith).

2.1.3. Arylalkylamines

Normethadone [*467-85-6*], 6-(*N,N*-dimethylamino)-4,4-diphenyl-3-hexanone, $C_{20}H_{25}NO$, M_r 295.42, *bp* 164 – 167 °C (0.4 kPa); hydrochloride [*847-84-7*], $C_{20}H_{26}ClNO$, M_r 331.88, *mp* 174 – 175 °C.

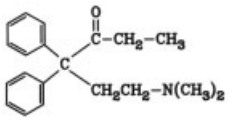

Normethadone is prepared by the same route used for methadone (→ Analgesics and Antipyretics, Chap. 3.1.3.) [19]. Normethadone is twice as potent an antitussive as codeine, but it has reduced analgesic potency compared with methadone, which is also a potent antitussive [20]. Both drugs are rarely used because of their addictive potential.
Trade Name. Ticarda (Hoechst).

Levopropoxyphene [*2338-37-6*], (−)-1-benzyl-3-(*N,N*-dimethylamino)-2-methyl-1-phenylpropylpropionate, $C_{22}H_{29}NO_2$, M_r 339.48, *mp* 75 –76 °C, $[\alpha]_D^{25}$ − 68.2° (*c* = 0.6; chloroform); 2-naphthalenesulfonate monohydrate (napsylate) [*55557-30-7*], $C_{22}H_{29}NO_2 \cdot C_{10}H_8O_3S \cdot H_2O$, M_r 565.72.

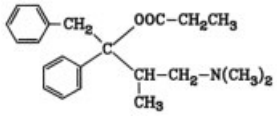

Levopropoxyphene is prepared as described for propoxyphene (→ Analgesics and Antipyretics), using the (−)-form of the intermediates [21]. Levopropoxyphene has useful antitussive properties, whereas analgesic activity is present mainly in dextropropoxyphene [22].
Trade Names. Novrad (Lilly), Sotorni (Ravensberg).

Tipepidine [*5169-78-8*], 3-(di-2-thienylmethylene)-1-methylpiperidine, $C_{15}H_{17}NS_2$, M_r 275.45, *mp* 64 – 65 °C; citrate monohydrate [*5169-77-7*], $C_{21}H_{25}NO_7S_2 \cdot H_2O$, M_r 485.58, *mp* 138 – 139 °C; 4-hydroxybenzophenone-2-carboxylate (hibenzoate) [*31139-87-4*], $C_{29}H_{27}NO_4S_2$, M_r 517.67, *mp* 187 – 190 °C. Grignard reaction of ethyl 1-methyl-3-piperidinecarboxylate (ethyl 1-methylnipecotate) with 2-thienylmagnesium bromide and acid dehydration of the intermediate carbinol yields tipepidine [23]:

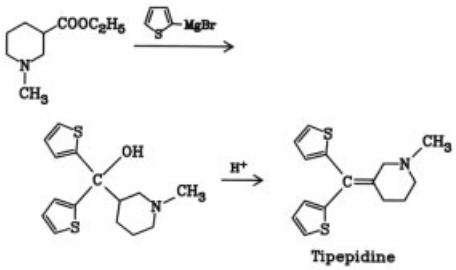

Tipepidine is approximately equipotent to codeine in antitussive activity, but less analgesic and toxic [24].
Trade Names. Asverin (Tanabe), Sotal (G. Ramon).

Clofedanol [*791-35-5*], chlophedianol, 1-(2-chlorophenyl)-1-phenyl-3-dimethylaminopropanol, $C_{17}H_{20}ClNO$, M_r 289.80; hydrochloride [*511-13-7*], $C_{17}H_{21}Cl_2NO$, M_r 326.26, *mp* 190 – 191 °C. For the synthesis of clofedanol, 2-chlorobenzophenone reacts with acetonitrile in the presence of sodium amide; the nitrile group of the product is then catalytically reduced to the primary amine, and the intermediate is converted into clofedanol by reductive *N*-methylation [25]:

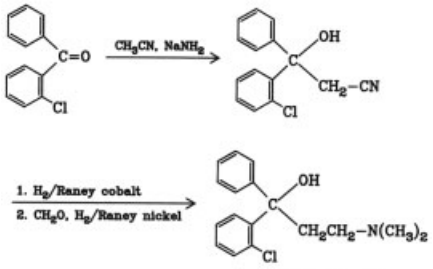

Clofedanol

Clofedanol is a centrally acting antitussive without analgesic activity or morphine-like side effects, but with some of the effects of a local anesthetic [26].

Trade Names. Detigon (Bayer), Pectolitan (Kettelhack Riker), Ulo (Riker), Coldrin (Nippon Shinyaku).

Clobutinol [*14860-49-2*], 2-(4-chlorobenzyl)-4-dimethylamino-3-methyl-2-butanol, $C_{14}H_{22}ClNO$, M_r 255.79, *bp* 179 – 181 °C (1.6 kPa); hydrochloride [*1215-83-4*], $C_{14}H_{23}Cl_2NO$, M_r 292.25, *mp* 169 – 170 °C. Clobutinol is made by Grignard reaction of 4-dimethylamino-3-methyl-2-butanone with 4-chlorobenzylmagnesium chloride [27].

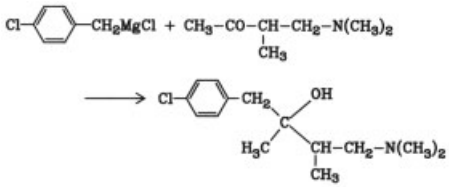

Clobutinol

Clobutinol is an antitussive with CNS activity which is equipotent to codeine and has properties similar to those of clofedanol [28].

Trade Name. Silomat (Thomae, Badrial, Morishita).

Isoaminile [*77-51-0*], 4-dimethylamino-2-isopropyl-2-phenylvaleronitrile, $C_{16}H_{24}N_2$, M_r 244.37, *bp* 138 – 146 °C (0.4 kPa); citrate [*126-10-3*], $C_{22}H_{32}N_2O_7$, M_r 436.62, *mp* 63 – 64 °C. For the synthesis of isoaminile, benzyl cyanide is alkylated in two steps, first by isopropyl bromide and then by 2-(dimethylamino)propyl chloride in the presence of sodium amide [29].

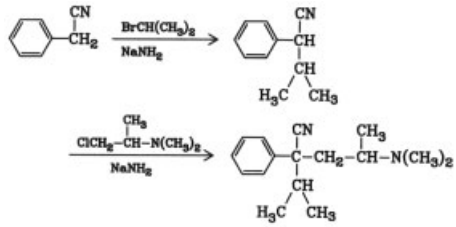

Isoaminile

Isoaminile is similar to codeine, but it has no analgesic activity or morphine-like side effects [30].

Trade Names. Peracon (Kali-Chemie), Dimyril (Fisons).

2.1.4. Basic Esters

Butethamate [*14007-64-8*], 2-diethylaminoethyl 2-phenylbutyrate, $C_{16}H_{25}NO_2$, M_r 263.37, *bp* 167 – 169 °C (1.5 kPa), n_D^{20} 1.4909; citrate [*13900-12-4*], $C_{22}H_{33}NO_9$, M_r 455.51, *mp* 109 – 110 °C. Butethamate is made by alkylation of benzyl cyanide with ethyl bromide in the presence of sodium amide; subsequent hydrolysis of the nitrile group yields 2-phenylbutyric acid, which is then esterified by reaction of the sodium salt with 2-diethylaminoethyl chloride [31].

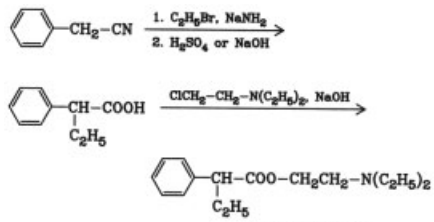

Butethamate

In addition to its antitussive activity, butethamate is active as a bronchodilator [32].

Trade Name. Pertix (Hommel); in many combination drugs.

Butamirate [*18109-80-3*], 2-(2-diethylaminoethoxy)ethyl 2-phenylbutyrate, $C_{18}H_{29}NO_3$, M_r 307.44, *bp* 140 – 155 °C (0.13 kPa); citrate [*18109-81-4*], $C_{24}H_{37}NO_{10}$, M_r 499.57, *mp* 75 °C. For the synthesis of butamirate, 2-phenylbutyric acid (prepared as described for

butethamate) is esterified by reaction of its acid chloride with 2-(2-diethylaminoethoxy)ethanol [33].

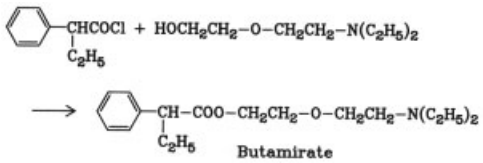

Compared to butethamate, the antitussive activity of butamirate is increased because of the lengthening of the side chain.

Trade Names. Acodeen (Hommel), Sinecod (Hommel, Karlspharma), Intussin (Spofa).

Pentoxyverine [*77-23-6*], carbetapentane, 2-(2-diethylaminoethoxy)ethyl 1-phenyl-1-cyclo-pentanecarboxylate, $C_{20}H_{31}NO_3$, M_r 333.46, *bp* 165 – 170 °C (0.001 kPa); citrate [*23142-01-0*], $C_{26}H_{39}NO_{10}$, M_r 525.60, *mp* 93 °C. Pentoxyverine is made by alkylation of benzyl cyanide with 1,4-dibromobutane in the presence of sodium amide; subsequent hydrolysis of the nitrile group gives 1-phenylcyclopentanecarboxylic acid, which is then esterified in a procedure analogous to that used for butamirate [34].

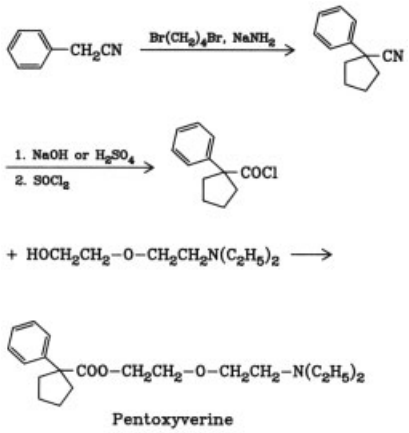

Pentoxyverine is more active than codeine; it has a low toxicity and shows no adverse side effects [35].

Trade Names. Atussil (Squibb), Germapect (Thiemann), Sedotussin (UCB), Toclase (Pfizer, Sumitomo), Tuclase (UCB).

Oxeladin [*468-61-1*], 2-(2-diethylaminoethoxy)ethyl 3-phenyl-3-pentanecarboxylate, $C_{20}H_{33}NO_3$, M_r 335.47, *bp* 140 °C (0.015 kPa); citrate [*16485-39-5*], $C_{26}H_{41}NO_{10}$, M_r 527.61, *mp* 90 – 91 °C. For the synthesis of oxeladin, 2-ethyl-2-phenylbutyric acid (prepared in the same way as phenylbutyric acid, see butamirate) is condensed with bis(2-chloroethyl) ether in the presence of sodium hydroxide, and the product is subsequently treated with diethylamine [36].

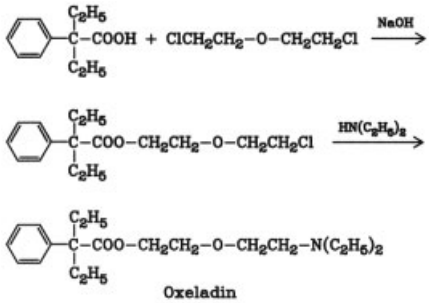

The pharmacological properties of oxeladin are very similar to those of pentoxyverine [37].

Trade Names. Dorex (Woelm), Paxeladine (Beaufour), Pectamol (British Drug Houses, Malesci), Hihustan (Maruko).

Dimethoxanate [*477-93-0*], 2-(2-dimethylaminoethoxy)ethyl phenothiazine-10-carboxylate, $C_{19}H_{22}N_2O_3S$, M_r 358.48; hydrochloride [*518-63-8*], $C_{19}H_{23}ClN_2O_3S$, M_r 394.91, *mp* 161 – 163 °C (decomp.).

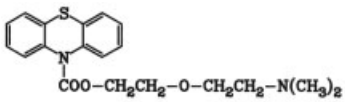

Synthesis: phenothiazine-10-carbonyl chloride is obtained by the reaction of phenothiazine with phosgene and then converted to dimethoxanate with 2-(2-dimethylaminoethoxy)ethanol as described for butamirate [38]. Dimethoxanate is a slightly less active antitussive than codeine and shows additional local anesthetic and mucolytic effects [39].

Trade Names. Cotrane (Clin Midy), Cothera (Ayerst), Perlatos (Farmacologico Mil.).

Pipazethate [*2167-85-3*], 2-(2-piperidinoethoxy)ethyl-10*H*-pyrido[3,2-*b*][1,4]benzo-

thiadiazine-10-carboxylate, pipazetate, $C_{21}H_{25}N_3O_3S$, M_r 399.52; hydrochloride [*6056-11-7*], $C_{21}H_{26}ClN_3O_3S$, M_r 435.98, *mp* 160 – 161 °C.

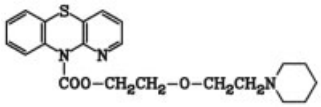

Pipazethate is prepared analogously to dimethoxanate [40]. Pipazethate is predominantly a centrally acting antitussive, similar to dimethoxanate [41].

Trade Names. Selvigon (Homburg, Galenika, Panchemie), Theratuss (Squibb).

2.1.5. Basic Ethers

Benproperine [*2156-27-6*], 1-[1-(2-benzylphenoxy)-2-propyl]piperidine, $C_{21}H_{27}NO$, M_r 309.43, *bp* 159 – 161 °C (0.03 kPa); phosphate [*3563-76-6*], $C_{21}H_{30}NO_5P$, M_r 407.45, *mp* 150 – 152 °C.

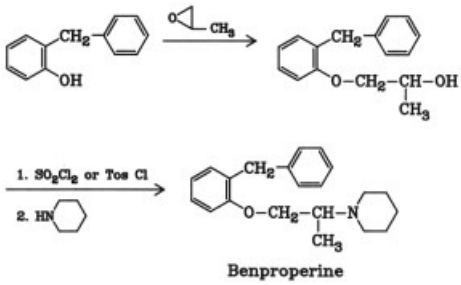

Synthesis: 2-benzylphenol is converted into 1-(2-benzylphenoxy)-2-propanol with propylene oxide; the hydroxyl group of the product is then replaced by chloride or the tosyloxy group and finally by the piperidino group to give benproperine [42].

In addition to its activity on the central nervous system, benproperine also has a peripheral mode of action [43].

Trade Names. Blascorid (Pharmacia, Guidotti), Flaveric (Taito Pfizer), Pectipront (Mack), Tussafug (Robugen, Medipharm).

Cloperastine [*3703-76-2*], 1-[2-[α-(4-chlorophenyl)benzyloxy]ethyl]piperidine, $C_{20}H_{24}ClNO$, M_r 329.88, *bp* 172 – 174 °C (0.01 kPa); hydrochloride, $C_{20}H_{25}Cl_2NO$, M_r 366.31, *mp* 147.9 °C. Synthesis: 4-chlorobenzhydrol is condensed with *N*-(2-chloroethyl)piperidine in the presence of sodium amide [44].

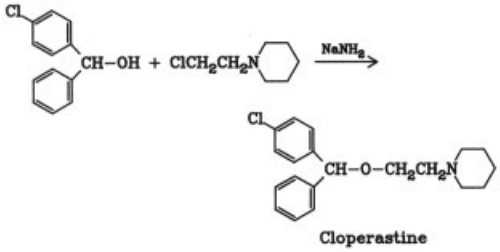

Cloperastine is an antitussive that acts on the central nervous system with nearly the same potency as codeine but with additional antihistaminic and spasmolytic effects [45].

Trade Names. Hustazol (Yoshitomi), Seki (Simes).

Morclofone [*31848-01-8*], 4′-chloro-3,5-dimethoxy-4-(2-morpholinoethoxy)benzophenone, $C_{21}H_{24}ClNO_5$, M_r 405.88, *mp* 91 – 92 °C. Synthesis: 4-chloro-3,5-dimethoxy-4-hydroxybenzophenone is condensed with *N*-(2-chloroethyl)morpholine in the presence of sodium methoxide [46].

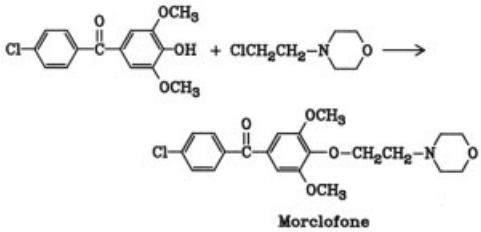

Morclofone is a non-narcotic antitussive with a potency similar to that of codeine [47].

Trade Names. Medicil (Medici), Nitux (Inpharzam), Plausitin (Carlo Erba).

2.1.6. Piperazine Derivatives

Dropropizine [*17692-31-8*], 3-(4-phenyl-1-piperazinyl)-1,2-propanediol, $C_{13}H_{20}N_2O_2$, M_r 236.31, *mp* 105 °C; hydrochloride, $C_{13}H_{21}ClN_2O_2$, M_r 272.78, *mp* 142 °C. Dropropizine is made by reaction of *N*-phenylpiperazine with oxiranylmethanol [48].

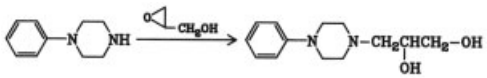

Dropropizine

Dropropizine is nearly as potent an antitussive as codeine and has no adverse side effects [49].

Trade Names. Catabex (Bios-Coutelier), Ribex (Formenti), Larylin (Beiersdorf; combination).

Eprazinone [*10402-90-1*], 3-[4-(β-ethoxy-2-phenethyl)-1-piperazinyl]-2-methylpropiophenone, $C_{24}H_{32}N_2O_2$, M_r 380.53, *mp* 160 °C; dihydrochloride [*10402-53-6*], $C_{24}H_{34}Cl_2N_2O_2$, M_r 453.46. Synthesis: styrene is treated with *tert*-butyl hypobromite in ethanol to give 2-ethoxy-2-phenylethyl bromide, which is then converted into eprazinone by exchange of bromine with piperazine and subsequent Mannich reaction with propiophenone and paraformaldehyde [50].

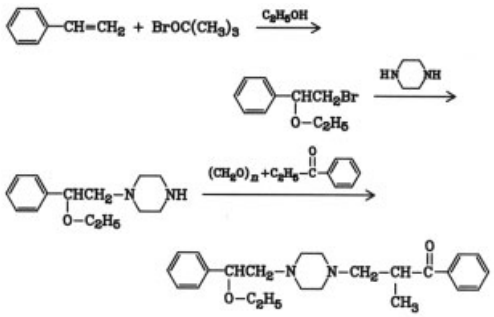

Eprazinone

In addition to its antitussive activity, eprazinone has mucolytic potency [51].

Trade Names. Eftapan (Merckle), Mucitux (Rion, Recordati), Resplen (Chugai).

Zipeprol [*34758-83-3*], 1-methoxy-3-[4-(β-methoxyphenethyl) piperazin-1-yl]-1-phenylpropan-2-ol, $C_{23}H_{32}N_2O_3$, M_r 384.52, *mp* 83 °C; dihydrochloride [*34758-84-4*], $C_{23}H_{34}Cl_2N_2O_3$, M_r 457.45, *mp* 231 °C. Zipeprol is made by the reaction of **1** with **2**, both of which are obtained in the manner described for eprazinone; the starting material for **1** is 3-phenylallyl alcohol [52].

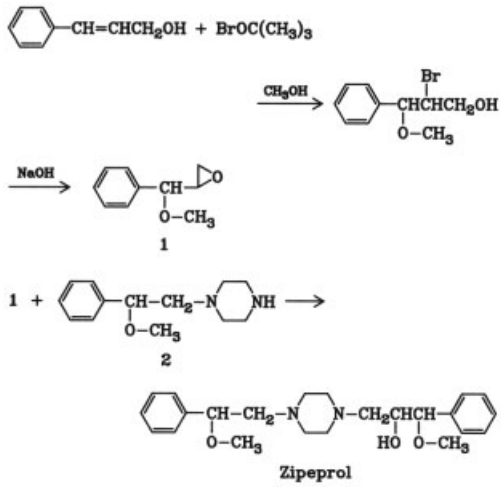

Zipeprol

The pharmacological profile of zipeprol is similar to that of eprazinone [53].

Trade Names. Mirsol (Mepha), Respilene (Rete-France, Organon, Winthrop).

2.1.7. Other Compounds

Fominoben [*18053-31-1*], 3′-chloro-2′-{*N*-methyl-*N*-[(morpholinocarbonyl)methyl]-aminomethyl}benzanilide, $C_{21}H_{24}ClN_3O_3$, M_r 401.89, *mp* 122.5 – 123 °C; hydrochloride [*18053-32-2*], $C_{21}H_{25}Cl_2N_3O_3$, M_r 438.36, *mp* 206 – 208 °C (decomp.). Synthesis: 3-chloro-2-methylaniline is acylated with benzoyl chloride to give the *N,N*-dibenzoylamine, which is then brominated with *N*-bromosuccinimide and subsequently converted with sarcosine morpholide into fominoben with elimination of one benzoyl group [54].

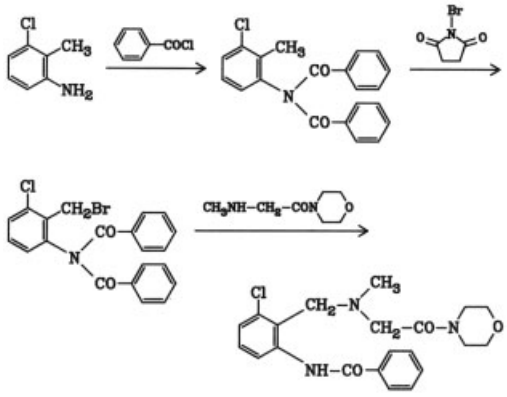

Fominoben

Fominoben has the same antitussive potency as codeine but no analgesic properties or morphine-like side effects. In contrast to the structurally related bromhexine (see Section 3.2.2), it shows no secretolytic or broncholytic activity [55].
Trade Names. Noleptan (Thomae), Deronyl (VEB Radebeul), Terion (Lusofarmaco).

Sodium dibunate is a mixture of the sodium salts of 3,7- and 3,6-di-*tert*-butyl-1-naphthalenesulfonic acid (dibunate, $C_{18}H_{24}O_3S$, M_r 319.45, *mp* 158 °C); the sodium salt [*14992-59-7*], $C_{18}H_{23}NaO_3S \cdot 2\,H_2O$, M_r 360.46, melts at 300 °C (decomp.).

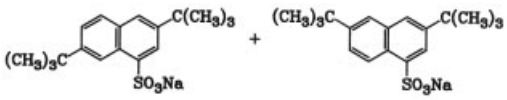

Dibunate is obtained by sulfonation of 2,7-di-*tert*-butylnaphthalene with chlorosulfuric acid. The reaction is accompanied by partial rearrangement of one *tert*-butyl group [56]. Dibunate was originally developed as a surfactant and is structurally unrelated to any other antitussive. It has a safe and effective antitussive action, predominantly on the central nervous system [57].
Trade Names. Becantex (Labaz), Bechisan (Sidus); combinations: Aspekton (Krewel), Makatussin (Makara).

2.2. Antitussives Acting Mainly Peripherally

Several antitussive drugs have been claimed to act predominantly on the peripheral nervous system. Theoretically this could be achieved by selective anesthetic action on either the cough receptors outside the central nervous system or the pulmonary stretch receptors. These receptors, found in the parenchyma of the lungs, are also responsible for regulating respiratory mechanisms (Hering – Breuer reflex). Therefore, the structure of this group of antitussives somewhat resembles that of local anesthetics. In addition to the peripheral activity, some smaller centrally active component may contribute to the overall effect.

Benzonatate [*104-31-4*], 3,6,9,12,15,18,21,24,27-nonaoxaoctacosyl 4-butylaminobenzoate, $C_{30}H_{53}NO_{11}$, M_r 603.73.

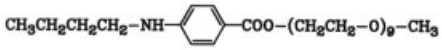

Synthesis: ethyl 4-butylaminobenzoate is transesterified with nonaethyleneglycol monomethyl ether in the presence of catalytic amounts of sodium methoxide [58]. Benzonatate acts mainly on the peripheral nervous system, with a small activity on the central nervous system. The selective anesthetic action on the pulmonary stretch receptors is achieved by combining a local anesthetic moiety, related to tetracaine, with a lipid-soluble side chain [59].
Trade Name. Tessalon (Ciba-Geigy, Endo).

Oxolamine [*959-14-8*], 5-(2-diethylaminoethyl)-3-phenyl-1,2,4-oxadiazole, $C_{14}H_{19}N_3O$, M_r 245.32, *bp* 127 °C (0.05 kPa); hydrochloride, $C_{14}H_{20}ClN_3O$, M_r 281.79, *mp* 153 – 154 °C; citrate [*1949-20-8*], $C_{20}H_{27}N_3O_8$, M_r 437.85. Acylation of benzamidoxime with 3-chloropropionyl chloride and subsequent cyclocondensation of the product with diethylamine yields oxolamine [60].

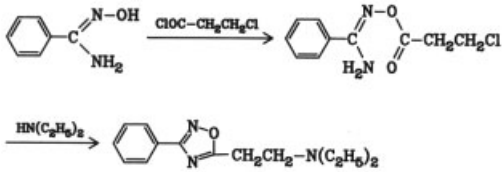

Oxolamine

Oxolamine acts mainly on the peripheral nervous system, with a smaller activity on the central nervous system. Additional analgesic and anti-inflammatory effects may contribute to its overall potency which is comparable to that of codeine [61].
Trade Names. Bredon (Organon), Broncatar (Pulitzer), Perebron (Angelini).

3. Expectorants [62]

Expectorants are substances that aid or accelerate the removal of bronchial secretions from the bronchi and trachea. The group of expectorants

can be divided, though not strictly, into secretolytics, mucolytics, and secretomotorics.

Secretolytics are intended to liquefy the mucus by promoting secretion. Mucolytics modify the physicochemical properties of the mucus, particularly its viscosity. Secretomotorics facilitate the passage and expulsion of secretion.

3.1. Secretolytics and Secretomotorics

Various groups of substances increase bronchial secretion. These consist of inorganic salts (in particular, ammonium chloride, calcium citrate, and potassium iodide); ipecac (dried rhizomes and roots of ipecacuanha) and emetine, which are isolated from ipecacuanha roots; and some saponin-containing drugs, such as senega root, primula rhizome, and licorice root.

Some volatile oils have secretolytic effects and, because they stimulate bronchial ciliary movement, also have secretomotor activity. They include the essential volatile oils of anise, eucalyptus, myrthe, thyme, and turpentine. Terpin hydrate [*2451-01-6*] and pinol hydrate (sobrerol) [*498-71-5*] are derived from turpentine oil.

Phenolic compounds, such as cresol and guaiacol, are also secretomotorics. However, because these compounds are severely irritating their ethers or esters are preferred in therapy. Important examples are guaiacol glycerol ether (guaifenesin), guaiacol phenylacetate, and guaiacol acetylsalicylate (guacetisal [*55482-89-8*]). Sulfoguaiacol [*78247-49-1*] is used in many combination drugs; it is a mixture of the potassium salts of 4- and 5-guaiacolsulfonic acid.

Guaifenesin [*93-14-1*], 3-(2-methoxyphenoxy)-1,2-propanediol, $C_{10}H_{14}O_4$, M_r 198.21, *mp* 78.5 – 79 °C, is made by the reaction of guaiacol with oxiranylmethanol (a) or with 3-chloro-1,2-propanediol (b) in the presence of sodium hydroxide [63].

Guaifenesin increases and promotes the secretions of the respiratory tract by acting directly on the secretory cells [64]. It is also a centrally active muscle relaxant.

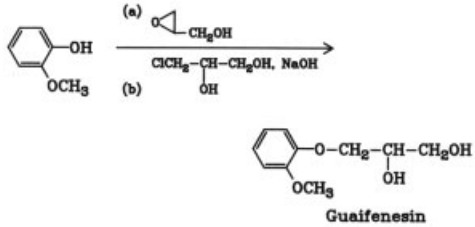

Trade Names. Resyl (Ciba-Geigy), Robitussin (Robins, Scheurich) and many others; numerous combination drugs.

3.2. Mucolytics

Mucolytics consist of several groups of substances with different mechanisms of action. *Surfactants* reduce surface tension; therefore, they lower the viscosity and facilitate the removal of viscid secretions.

The most popular compound of this type is *tyloxapol* [*25301-02-4*], Triton WR 1339, a polymer of 4-isooctylpolyoxyethylene-2-cresyl ether.

Trade Names. Alevaire (Winthrop, Breon, Berger), Tacholiquin (Bene-Chemie).

Cysteine derivatives have the greatest effect on mucus viscosity. Their reactive thiol groups probably cleave the disulfide bridges of mucoproteins.

The *benzylamines* bromhexine and ambroxol (Section 3.2.2) are structurally derived from the alkaloid vasicine. They have both mucolytic and secretolytic effects.

3.2.1. Cysteine Derivatives

Acetylcysteine [*616-91-1*], *N*-acetyl-L-cysteine, $C_5H_9NO_3S$, M_r 163.20, *mp* 109 – 110 °C, is made by acylating L-cysteine hydrochloride hydrate with acetic anhydride in the presence of sodium acetate [65]. Acetylcysteine is an effective mucolytic in vitro and in vivo, especially when administered as an aerosol [60], [66].

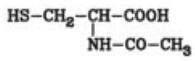

Trade Names. Fluimucil (Arsac, Inpharzam, Zambon), Fabrol (Ciba-Geigy), Mucomyst (Mead Johnson, Allard, Astra).

Carbocysteine [*638-23-3*], *S*-carboxymethyl-L-cysteine, $C_5H_9NO_4S$, M_r 179.21, *mp* 204–207 °C, is made by *S*-alkylation of L-cysteine with chloroacetic acid in the presence of sodium hydroxide [67]. Carbocysteine decreases bronchial secretion and promotes expulsion. It has no free sulfhydryl group and is therefore administered only orally (to induce metabolic activation) and not by aerosol [60], [68].

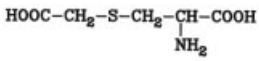

Trade Names. Lisomucil (Lirca), Mucodyne (USV, Berk), Mucopront (Mack), Rhinathiol (Joullié), Transbronchin (Homburg).

Mecysteine hydrochloride [*18598-63-5*], methyl L-cysteinate hydrochloride, $C_4H_{10}ClNO_2S$, M_r 171.66, *mp* 140–141 °C, $[\alpha]_D^{20}$ −2.9° (10 %; methanol), is made either by esterifying L-cysteine hydrochloride monohydrate with methanolic hydrochloric acid or by reducing L-cystine dimethyl ester dihydrochloride catalytically or with zinc and hydrochloric acid [69].

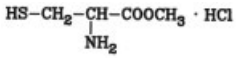

Trade Names. Acthiol (Joullié), Actiol (Lirca), Epecoale (Beppu), Visclair (Sinclair).

Letosteine [*53943-88-7*], 2-[2-(ethoxycarbonylmethylthio)ethyl]-1,3-thiazole-4-carboxylicacid, $C_{10}H_{17}NO_4S_2$, M_r 279.37, *mp* 142 °C, is made by treating ethyl mercaptoacetate with acrolein to give ethyl 3-oxopropylthioacetate, which is then condensed with L-cysteine [70]. For its pharmacological properties, see [71].

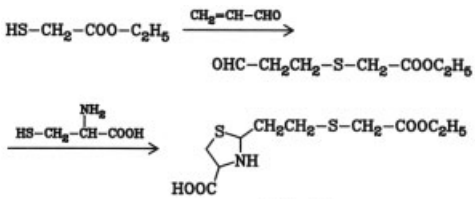

Trade Name. Viscotiol (Erba, ISF).

Mesna [*19767-45-4*], sodium 2-mercaptoethanesulfonate, $C_2H_5NaO_3S_2$, M_r 164.17, is made by S-alkylation of thiourea with 2-bromoethanesulfonic acid and cleavage of the isothiourea group with sodium hydroxide [72]. For its pharmacological properties, see [73].

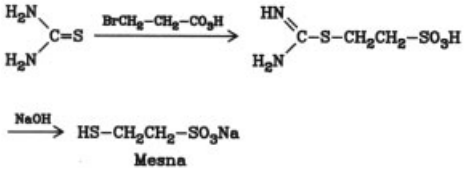

Trade Names. Mistabronco (UCB), Mucofluid (UCB).

3.2.2. Benzylamines

Bromhexine, *N*-cyclohexyl-*N*-methyl-*N*-(2-amino-3,5-dibromobenzyl)amine [*3572-43-8*], $C_{14}H_{20}Br_2N_2$, M_r 376.16; hydrochloride [*611-75-6*], $C_{14}H_{21}Br_2ClN_2$, M_r 416.62, *mp* 237.5–238 °C, is made by alkylating *N*-methylcyclohexylamine with 2-nitrobenzyl bromide to give *N*-methyl-*N*-(2-nitrobenzyl)cyclohexylamine, which is subsequently converted into bromhexine by reduction of the nitro group with hydrazine and Raney nickel, followed by bromination [74]. Bromhexine increases the amount of respiratory tract fluid and reduces its viscosity by activating enzymes that hydrolyze mucopolysaccharides [62], [75].

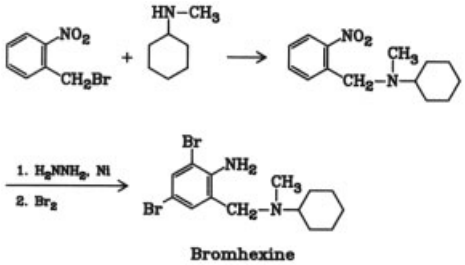

Trade Names. Bisolvon (Thomae, Boehringer Ingelheim, Badrial, Bender, Tanabe), Auxit (Heyden); many combinations.

Ambroxol [*18683-91-5*], *trans*-4-[(2-amino-3,5-dibromobenzyl)amino]cyclohexan-

ol, $C_{13}H_{18}Br_2N_2O$, M_r 378.11; hydrochloride, $C_{13}H_{19}Br_2ClN_2O$, M_r 414.58, *mp* 233 – 234.5 °C (decomp.).

Synthesis: catalytic hydrogenation of 4-acetamidophenol over rhodium catalysts and subsequent hydrolysis with sodium hydroxide yields *trans*-4-aminocyclohexanol; this is then condensed with 2-amino-3,5-dibromobenzaldehyde to give the Schiff base which is finally reduced to ambroxol with sodium tetrahydroborate or formic acid [76].

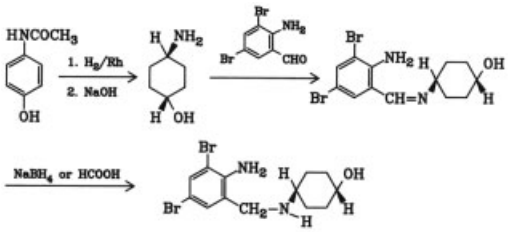

Ambroxol

Ambroxol is the active metabolite of bromhexine and, therefore, has similar pharmacological effects, but with increased bioavailability [77].

Trade Names. Mucosolvan (Thomae, Bender, De Angeli), Fluibron (Chiesi).

4. References

1. D. Miller in M. E. Wolff (ed.): *Burger's Medicinal Chemistry,* 4th ed., Part III, J. Wiley & Sons, New York 1981, pp. 759 – 785.
2. H. Salem, D. M. Aviado (eds.): *Antitussive Agents, International Encyclopedia of Pharmacology and Therapeutics,* Section 27, vols. 1 – 3, Pergamon Press, New York 1970.
3. G. Erhart, H. Ruschig: *Arzneimittel, Entwicklung, Wirkung, Darstellung,* **vol. 3,** Verlag Chemie, Weinheim 1972, pp. 42 – 62.
4. E. Mutschler: *Arzneimittelwirkungen,* 4th ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1981.
5. C. H. Boehringer, DE 247 180, 1912. J. Schwyzer: *Die Fabrikation pharmazeutischer u. chemisch-technischer Produkte,* Springer, Berlin 1931, p. 389.
6. E. Merck AG, DE 131 980, 1902.
7. H. Friebel, C. Reichel, A. v. Graevenitz, *Naunyn-Schmiedebergs Arch. Pharmakol. Exp. Pathol.* **224** (1955) 384.
8. P. Chabrier, R. Guidicelli, J. Thuillier, *Ann. Pharm. Fr.* **8** (1950) 261. Lab. Dausse, US 2 619 485, 1952.
9. A. J. May, J. G. Widdicombe, *Br. J. Pharmacol. Chemother.* **9** (1954) 335. C. E. Heffron, *J. New Drugs* **1** (1961) 217.
10. A. Stein, *Pharmazie* **10** (1955) 180. H. Oldenberg, B. Oldenberg, DE 260 233, 1913.
11. Knoll AG, DE 278 111, 1914.
12. B. Weiss, *Am. J. Pharm.* **31** (1959) 286.
13. Knoll AG, DE 607 931, DE 617 238, DE 623 821 (1935).
14. J. La Barre, H. Plisnier, *Arch. Int. Pharmacodyn. Ther.* **119** (1959) 205. M. S. Segal, *JAMA J. Am. Med. Assoc.* **169** (1959) 1063.
15. O. Schnider, A. Grüssner, *Helv. Chim. Acta* **34** (1951) 2211; Hoffmann-La Roche, US 2 676 177, 1954.
16. B. Pellmont, H. Bächtold, *Schweiz. Med. Wochenschr.* **84** (1954) 1368.
17. Yamanouchi, DE 2 128 607, 1971.
18. M. Murakami, S. Kawahara, N. Inukai, N. Nagano, H. Iwamoto, H. Ida, *Chem. Pharm. Bull.* **20** (1972) 1706.
19. M. Bockmühl, G. Erhart, *Justus Liebigs Ann. Chem.* **561** (1948) 52.
20. K. L. Buchheim, B. Angermann, *Dtsch. Med. Wochenschr.* **76** (1951) 1278.
21. A. Pohland, H. R. Sullivan, *J. Am. Chem. Soc.* **75** (1953) 4458; **77** (1955) 3400.
22. E. B. Robbins, J. A. Miller, *Pharmacologist* **2** (1960) 98.
23. K. Okumara, T. Tanaka, S. Siato, H. Kugita, N. Sugimoto, *Chem. Abstr.* **53** (1959) 10 214 g.
24. K. Higaki, T. Danno, Y. Kowa, N. Sugimoto, *Chem. Abstr.* **54** (1960) 3725 h.
25. Bayer, DE 1 083 277, 1960.
26. R. Gösswald, *Arzneim. Forsch.* **8** (1958) 550.
27. Thomae, DE 1 146 068, DE 1 153 380, 1959.
28. R. Engelhorn, *Arzneim. Forsch.* **10** (1960) 785.
29. Kali Chemie, GB 822 695, 1959.
30. D. Krause, *Arzneim. Forsch.* **8** (1958) 553.
31. Hommels Haematogen, CH 291 375, 1953. Chem. Fabrik Para, CH 292 596, 1953.
32. B. N. Halpern, *Arch. Int. Pharmacodyn. Ther.* **59** (1938) 149.
33. Hommel AG, DE 1 151 515, 1963.
34. H. G. Morren, GB 753 779, 1956.
35. S. Levis, S. Preat, F. Moyeroons, *Arch. Int. Pharmacodyn. Ther.* **103** (1955) 200.
36. V. Petrow, O. Stephenson, A. M. Wild, *J. Pharm. Pharmacol.* **10** (1958) 40.

37. A. David, F. Leith Ross, D. K. Vallance, *J. Pharm. Pharmacol.* **9** (1957) 446.
38. Am. Home Prod. Corp., US 2 778 824, 1957 (C. v. Seemann).
39. C. I. Chappel, M. G. P. Stegen, G. A. Grant, *Can. J. Biochem. Physiol.* **36** (1958) 475.
40. W. A. Schuler, H. Klebe, A. v. Schlichtegroll, *Justus Liebigs Ann. Chem.* **673** (1964) 102.
41. K. J. Hahn, H. Friebel, *Med. Pharmacol. Exp.* **14** (1966) 87.
42. Aktieselskabet Pharmacia, GB 914 008, 1962.
43. K. Yamatsu et al., *Jpn. J. Pharmacol.* **17** (1967) 538.
44. Parke, Davis, GB 670 622, 1952.
45. K. Takagi, T. Yuizono, Y. Kasé, *Chem. Abstr.* **67** (1967) 107 101 u.
46. Carlo Erba, DE 2 016 770, 1971.
47. E. Bosisio, G. Casadonte, G. Sacchetti, *Farmaco Ed. Prat.* **26** (1971) 356.
48. H. G. Morren, BE 601 394, 1961.
49. P. R. B. Noel, *Arzneim. Forsch.* **19** (1969) 1246. K. Cartwright, J. L. Paterson, *J. Pharm. Pharmacol.* **23** (1971) 247.
50. C.E.R.M., US 3 448 192, 1969 (R. Mauvernay).
51. J. Vacher, C. Lakatos, G. Rispat, P. Duchene-Marullaz, *Arch. Int. Pharmacodyn. Ther.* **165** (1967) 1.
52. C.E.R.M., US 3 718 650, 1973 (R. Mauvernay et al.).
53. G. Rispat, H. Burgi, D. Cosnier, P. Duchene-Marullaz, G. Streichenberger, *Arzneim. Forsch.* **26** (1976) 523.
54. G. Krüger, J. Keck, O. Zipp, J. Nickl, H. Machleidt, G. Ohnacker, *Arzneim. Forsch.* **23** (1973) 290.
55. S. Püschmann, R. Engelhorn, *Arzneim. Forsch.* **23** (1973) 296.
56. M. Menard, L. Mitchell, J. Komlossy, A. Wrigley, F. L. Chubb, *Can. J. Chem.* **39** (1961) 729.
57. G. R. de Vleeschhouwer, *Arch. Int. Pharmacodyn. Ther.* **97** (1954) 34.
58. Ciba, US 2 714 608, US 2 714 610, 1955 (M. Matter).
59. K. Bucher, *Schweiz. Med. Wochenschr.* **86** (1956) 94. R. Sell, E. Lindner, H. Jahn, *Naunyn-Schmiedebergs Arch. Pharmacol. Exp. Pathol.* **234** (1958) 164.
60. Angelini Francesco, DE 1 097 998, 1961.
61. B. Silvestrini, C. Pozzati, *Br. J. Pharmacol. Chemother.* **16** (1960) 209.
62. O. P. Schmidt, *Med. Klin. (Munich)* **72** (1977) 145 – 160.
63. H. L. Yale, E. J. Pribyl, W. Braker, F. H. Bergeim, W. A. Lott, *J. Am. Chem. Soc.* **72** (1950) 3710.
64. E. M. Boyd, E. P. Sheppard, C. E. Boyd, *Applied Therapeutics (Toronto)* **9** (1967) 55.
65. T. A. Martin, J. R. Corrigan, C. W. Waller, *J. Org. Chem.* **30** (1965) 2839. Mead Johnson, US 3 184 505, 1965 (T. A. Martin, C. W. Waller).
66. T. A. Martin, D. H. Causey, A. L. Sheffner, A. G. Wheeler, J. R. Corrigan, *J. Med. Chem.* **10** (1967) 1172.
67. M. D. Armstrong, J. D. Lewis, *J. Org. Chem.* **16** (1951) 749.
68. V. N. Huyen, S. Garcet, L. Lakah, *C. R. Seances Soc. Biol. Ses. Fil.* **160** (1966) 1849.
69. M. Bergmann, G. Michalis, *Ber. Dtsch. Chem. Ges.* **63** (1930) 987. L. Zervas, D. M. Theodoropoulos, *J. Am. Chem. Soc.* **78** (1956) 1359.
70. Ferlux-Chimie, DE-OS 2 410 307, 1974 (M. X. Chodkiewicz).
71. M. Neuman, *Med. Actual.* **16** (1980) 109.
72. C. H. Schramm, H. Lemaire, R. H. Karlson, *J. Am. Chem. Soc.* **77** (1955) 6231. UCB, US 3 567 835, 1971 (H. Morren).
73. S. N. Steen, I. Ziment, D. Freeman, J. S. Thomas, *Clin. Pharmacol. Ther. (St. Louis)* **16** (1974) 58.
74. J. Keck, *Justus Liebigs Ann. Chem.* **662** (1963) 171.
75. R. Engelhorn, S. Püschmann, *Arzneim. Forsch.* **13** (1963) 474. H. Eigelreiter, M. Mair, *Arzneim. Forsch.* **17** (1967) 353.
76. J. Keck, *Justus Liebigs Ann. Chem.* **707** (1967) 107.
77. S. Püschmann, R. Engelhorn, *Arzneim. Forsch.* **28** (1978) 889.

Coumarin → Flavors and Fragrances
Coumarone → Furan and Derivatives
Coumarone – Indene Resins → Resins, Synthetic
Cracking → Oil Refining
Crayons → Drawing and Writing Materials

Cresols and Xylenols

HELMUT FIEGE, Bayer AG, Leverkusen, Germany

1.	**Cresols**	1
1.1.	**Physical Properties**	2
1.2.	**Chemical Properties**	4
1.3.	**Formation and Isolation**	5
1.3.1.	Isolation from Coal Tars	5
1.3.2.	Recovery from Spent Refinery Caustics	6
1.4.	**Production**	8
1.4.1.	Alkali Fusion of Toluenesulfonates	8
1.4.2.	Alkaline Chlorotoluene Hydrolysis	10
1.4.3.	Cymene Hydroperoxide Cleavage	11
1.4.4.	Methylation of Phenol	13
1.4.5.	Other Routes	17
1.5.	**Separation of *m*- and *p*-Cresol**	21
1.5.1.	Physicochemical Processes	21
1.5.2.	Separation via Addition Compounds	22
1.5.3.	Separation via Ester or Salt Formation	23
1.5.4.	Separation by Nuclear Substitution	24
1.6.	**Quality Specifications and Analysis**	24
1.7.	**Handling, Storage, and Transportation**	25
1.8.	**Uses**	26
1.9.	**Economic Aspects**	27
2.	**Xylenols**	28
2.1.	**Properties**	28
2.2.	**Isolation**	29
2.3.	**Separation**	30
2.4.	**Production**	31
2.5.	**Quality Specifications and Analysis**	33
2.6.	**Handling, Storage, and Transportation**	34
2.7.	**Uses**	34
2.8.	**Economic Aspects**	35
3.	**Environmental Protection**	35
4.	**Toxicology and Occupational Health**	36
5.	**References**	37

1. Cresols [1–3]

Cresol, methylphenol, C_7H_8O, M_r 108.14, occurs in three isomeric forms:

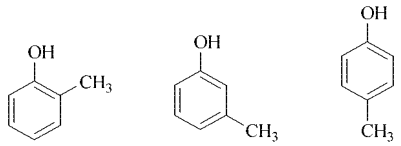

o-Cresol 2-methylphenol [*95-48-7*] | *m*-Cresol 3-methylphenol [*103-39-4*] | *p*-Cresol 4-methylphenol [*106-44-5*]

Mixtures of *m*- and *p*-cresol and of *o-, m-,* and *p*-cresol are occasionally referred to in the technical literature as dicresol and tri- or isocresol, respectively.

Cresylic acids [*1319-77-3*] are mixtures of cresols, xylenols, higher alkylated phenols, and possibly some phenol. *Tar acids* are cresylic acids obtained from tar.

STÄDELER first discovered cresol in cow's urine in 1851 [4]. FAIRLIE and WILLIAMSON found it in coal-tar creosote in 1854. (The word cresol, in which "creosote" and "phenol" are combined, is partly explained by this fact.) Cresol was first synthesized in 1866 by GRIESS, who boiled diazotized toluidine. All three cresol isomers were distinguished for the first time by ENGELHARDT and LATSCHINOFF in 1869.

Cresols and cresol derivatives (ethers and esters) are widely distributed in nature. They are formed as metabolites of various microorganisms and are found in the urine of mammals. Humans eliminate, on average, 87 mg of *p*-cresol per day in the urine [5]. Cresol in its various forms is also detectable in the extracts and water vapor distillates of many plants, e.g., in jasmine flower oil, cassia flower oil, easter lily oil, and ylang ylang oil, in the floral oil of *Yucca gloriosa,* in the scent of *Viola odorata,* in peppermint, eucalyptus, and camphor oil, and in the essential oils of several plants of the genus *Artemisia,* of conifers, of oak wood, and of sandalwood. Small amounts of cresol are also found in certain foods and drinks, e.g., in tomatoes and tomato ketchup, mushrooms, cooked asparagus, milk [6], certain types of cheese, butter oil, red wine, whisky, rum, cognac, and other brandies [7], raw and roasted coffee, black tea, smoked foods, and tobacco and tobacco smoke. Considerable amounts of cresol may be formed and en-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_025

ter surface waters if forests are pyrolyzed by pyroclastic flow from erupting volcanoes, as happened after the eruption of Mount St. Helens in 1980 [8].

Cresols are important chemical raw materials. They were originally obtained only from coal tar; after World War II, they were also obtained from spent refinery caustics. Since the mid-1960s, they have been produced synthetically on an increasingly large scale. "Synthetic cresol" now provides ca. 60 % of the requirements of the United States, Europe, and Japan; only ca. 40 % of the requirements are now met by "natural cresol," i.e., cresol from coal tar and spent refinery caustics.

1.1. Physical Properties [1], [9–12]

As pure substances, *o*- and *p*-cresol are crystalline; *m*-cresol is a viscous oil at room temperature. The cresols have a phenolic odor and are colorless, but become yellow to brown after a time. Because water dissolves freely in cresols, they absorb moisture from the air.

Physical data relating to the cresols have been compiled in Tables 1 and 2. Data concerning the temperature dependence of their thermodynamic properties are found in [9–14]; many additional properties are contained in [11]; and a differential thermal analysis study is described in [15]. Thermophysical properties of *m*-cresol as a function of temperature (303 – 503 K) and pressure (0.1 – 400 MPa) are given in [16]. For the relationship between vapor pressure and association of cresols, see [17].

The solubility of the cresols in phenol and in many organic solvents, e.g., aliphatic alcohols, ethers, chloroform, and glycerol, is high. They are less soluble in water than phenol. The presence of other water-soluble organic compounds (e.g., methanol) raises their solubility in water and reduces the critical solution temperature [19]. Dissolved inorganic salts lower the water solubility of cresols.

The cresols can be distilled with steam and form azeotropes with a number of compounds, e.g., decane, 1-decene, 1-undecene, dodecane, 1,2,4,5-tetramethylbenzene, divinylbenzene, ethylene glycol, diethylene glycol methyl and ethyl ether, diethylene glycol diethyl ether, ethylene glycol diacetate, benzyl formate, phenyl acetate, methyl benzoate, hexane-2,5-dione, 2-octanone, acetophenone, benzylamine, the toluidines, and numerous pyridine derivatives. Vapor – liquid equilibrium data of the cresols with numerous other substances are compiled in [20]; those of *m*-cresol with carbon dioxide and nitrogen under pressure at 190 – 390 °C are given in [21] (2 – 5 MPa) and [22] (2 – 25 MPa), respectively.

1.2. Chemical Properties [2], [23]

Acidity. The cresols are chemically similar to phenol. They are weak acids and dissolve in aqueous alkali to form water-stable salts known as cresolates. Therefore, they can be extracted into sodium hydroxide solution from solvents that are not miscible with water. Nevertheless, their acidity is so low (pK_a 10.1 – 10.3) that even hydrogen sulfide (pK_a 7.2) and carbon dioxide (pK_a 6.4) are able to liberate them from cresolates. Thus, they hardly dissolve at all in solutions of sodium carbonate or sodium hydrogencarbonate.

Etherification and Esterification. The hydroxyl group can be etherified with alkyl halides, dialkyl sulfates, dialkyl carbonates, and toluenesulfonic acid esters, and react with acyl anhydrides or acyl chlorides to give the corresponding tolyl esters, which are normally referred to as "cresyl" esters. Isocyanates react with the OH group to form urethanes.

Substitution of the Hydroxyl Group. Under drastic conditions (420 °C, Al_2O_3), the OH group is replaced by ammonia, and the corresponding toluidine is obtained. If cresols are reacted with sulfur oxytetrafluoride at 150 °C, with diphenylphosphine trichloride at 220 °C, or with phosphorus tribromide at 280 °C, the corresponding fluoro-, chloro-, or bromotoluene is obtained. Butanethiol in hydrochloric acid replaces the phenolic OH group by a thiobutyl group. Distillation with zinc powder gives toluene.

Hydrogenation. Hydrogenation in the vapor phase at 300 – 400 °C under pressure (up to 8 MPa) in the presence of catalysts consisting of transition metals and aluminum oxide

Table 1. Physical properties of pure cresols

Property	*o*-Cresol	*m*-Cresol	*p*-Cresol
Melting point at 101.3 kPa, °C	30.99	12.22	34.69
Depression of *mp* by 1 mol% of impurity, °C	0.596	0.749	0.915
Boiling point at 101.3 kPa, °C	191.00	202.23	201.94
Density at 25 °C (*s* or *l*), g/cm^3	1.135	1.0302	1.154
at 50 °C (*l*), g/cm^3	1.0222	1.0105	1.0116
Viscosity at 50 °C, mPa · s	3.06	4.17	4.48
Refractive index at 50 °C	1.5310	1.5271	1.5269
Dielectric constant at 50 °C	6.00	9.86	10.18
Specific conductance at 50 °C, S/cm	0.43×10^{-10}	63.4×10^{-10}	60.2×10^{-10}
Surface tension at 50 °C, N/m	34.4×10^{-6}	33.4×10^{-6}	34.0×10^{-6}
Dissociation constant at 25 °C	4.8×10^{-11}	8.1×10^{-11}	5.4×10^{-11}
Critical temperature, °C	424.4	432.6	431.4
Critial pressure, MPa	5.01	4.56	5.15
Critical density, kg/m^3	384	346	391
Heat capacity c_{p} at 25 °C (*g*), J mol^{-1} K^{-1}	127.3	124.6	124.6
c_{p} at 25 °C (*s* or *l*), J mol^{-1} K^{-1}	154.7	225.1	150.3
c_{p} at 50 °C (*l*), J mol^{-1} K^{-1}	237.9	235.0	233.3
Enthalpy of fusion ΔH_m at *mp,* 101.3 kPa, kJ/mol	15.830	10.714	12.715
Enthalpy of sublimation ΔH_{sub} at 25 °C, 101.3 kPa, kJ/mol	76.07	–	73.98
Enthalpy of vaporization ΔH_v at 25 °C, 101.3 kPa, kJ/mol	–	61.76	–
ΔH_v at *bp*, 101.3 kPa, kJ/mol	45.222	47.429	47.581
Enthalpy of combustion $\Delta H_c{}^\circ$ at 25 °C, 101.3 kPa, kJ/mol	− 3696	− 3706	− 3701
Enthalpy of formation $\Delta H_f{}^\circ$ at 25 °C (*s* or *l*), 101.3 kPa, kJ/mol	− 204.8	− 194.2	− 199.5
$\Delta H_f{}^\circ$ at 25 °C (*g*), 101.3 kPa, kJ/mol	− 128.7	− 132.4	− 125.5
Free energy of formation $\Delta G_f{}^\circ$ at 25 °C (*s* or *l*), 101.3 kPa, kJ/mol	− 55.7	− 59.2	− 51.0
$\Delta G_f{}^\circ$ at 25 °C (*g*), 101.3 kPa, kJ/mol	− 33.0	− 40.2	− 31.7
Entropy *S* at 25 °C (*s* or *l*), 101.3 kPa, J mol^{-1} K^{-1}	165.5	212.7	167.4
Entropy S° at 25 °C (*g*), 101.3 kPa, J mol^{-1} K^{-1}	352.8	356.3	351.0
Cubic expansion coefficient at 25 °C		0.759×10^{-3}	
Solubility in water at 25 °C, wt %	2.6	2.3	1.9
at 100 °C, wt %	4.8	5.3	4.8
of water in cresol at 25 °C, wt %	13.4	13.1	15.8
at 100 °C, wt %	18.4	20.8	24.9
Critical solution temperature in water, °C	164.8	148.6	143.7
at % cresol, wt %	34	36	34
Distribution coefficient log *P* octanol/water at 25 °C	1.95	–	1.95
Vapor density (air = 1)	3.72	3.72	3.72
Flash point (closed cup), °C	81	86	86
Ignition temperature, °C	555	555	555
Flammability limit, lower, in air at 20 °C, 101.3 kPa, vol%	1.3	1.0	1.0

Table 2. Vapor pressure of pure cresols (in kPa), for $T \geq 260$ °C [9], [18]

Temp., °C	*o*-Cresol,	*m*-Cresol,	*p*-Cresol,
25	0.037	0.019	0.016
50	0.25	0.12	0.11
60	0.49	0.25	0.23
80	1.57	0.86	0.83
100	4.26	2.52	2.47
120	10.10	6.36	6.29
140	21.47	14.26	14.21
160	41.72	29.01	29.07
180	75.27	54.45	54.75
200	127.6	95.52	96.24
220	204.9	158.2	159.6
260	471	378	388
300	940	774	810
340	1704	1423	1521
380	2894	2420	2659

gives toluene (hydrodeoxygenation) [24]. Hydrogenolysis over catalysts at 400 – 500 °C or purely thermally at 500 – 700 °C can be so controlled that cresol is mainly demethylated to phenol (hydrodealkylation) [25]. Hydrogenation over Raney nickel or noble-metal catalysts under less severe conditions gives methylcyclohexanones or methylcyclohexanols.

Oxidation. The cresols are sensitive to oxidation. Depending on the oxidizing agent, reaction conditions, and position of the methyl group, oxidative reactions, which occur by free radical mechanisms, lead to a large number of compounds such as hydroquinones, quinols, quinones, cyclic ketones, furans, dimeric and trimeric cresols, and tolyl ethers. Strongly oxidizing agents break down the phenol ring.

After the hydroxyl group has been protected by esterification or etherification, the methyl group can be selectively mono-, di-, or trichlorinated, or selectively oxidized to a formyl group (e.g., with manganese dioxide or oxygen) or to a carboxyl group (with acid permanganate solution). Unlike *m*-cresol, *o*- and *p*-cresol can be directly oxidized with oxygen to give *o*- and *p*-hydroxybenzaldehyde, e.g., in methanolic sodium hydroxide in the presence of catalytic amounts of iron tetraarylporphines [26]. Alkali fusion of the cresols in the presence of lead oxide or manganese dioxide produces the corresponding hydroxybenzoic acids directly.

Electrophilic Nuclear Substitution. Cresols, like phenol, readily undergo electrophilic substitution. The substituent enters the nucleus mainly in the *o*- and *p*-positions relative to the hydroxyl group. The cresols can therefore be nitrated even with dilute nitric acid. Nitrosation, halogenation, sulfonation, and alkylation occur readily. If *o*- or *p*-cresol is heated with Friedel – Crafts catalysts, such as $AlCl_3$, isomerization to *m*-cresol, thermodynamically the most stable of the three isomers (see page 16), occurs. The heating of dry alkali metal cresolates with CO_2 under pressure produces hydroxy(methyl)benzoic acids (Kolbe – Schmitt reaction). In the presence of alkali, formaldehyde adds to cresols, even at room temperature, to form hydroxy(methyl)benzyl alcohols; under acidic conditions or at elevated temperatures these condense to form high-molecular resins. The main products of the reaction of cresols with chloroform and alkali are *o*-hydroxy(methyl)benzaldehydes (Reimer – Tiemann synthesis). Electrophilic substitutions are occasionally complicated by partial or substantial cessation of the reaction at the primary stage of the cyclohexadienone formed by addition of the electrophile. Thus, if *p*-cresol is heated with tetrachloromethane in the presence of aluminum chloride, the main product is 4-methyl-4-trichloromethyl-2,5-cyclohexadienone (Zincke – Suhl reaction).

Behavior toward Materials [27]. At elevated temperature unalloyed steel is attacked by cresols to a noticeable extent, the severity of the attack depending on the water content. Stress corrosion cracking may occur even at 100 °C. Chromium steels are scarcely more resistant than unalloyed steels. Chromium – nickel stainless steels of the V2A type (AISI 321) are corroded at elevated temperatures at a rate of 0.1 – 1 mm/a, whereas the rate for V4A steels (AISI 316Ti) in boiling cresol is ca. 0.1 mm/a. Nickel is more resistant, its corrosion rate in pure cresols at 190 °C being only ca. 0.001 mm/a. Tantalum is considered to be entirely resistant to cresols.

Aluminum and its alloys are attacked severely at temperatures greater than ca. 120 °C when the water content of the cresol is too low. If the cresol contains less than ca. 0.3 % water, corrosion may proceed explosively. The use of copper and brass alloys should also be avoided.

The corrosive behavior of the cresols can alter significantly, depending on the presence of other compounds. Thus, the simultaneous presence of water-soluble chlorides and basic nitrogen compounds in the distillation of cresol/xylenol fractions has synergistic effects, leading, for example, to very severe corrosion of stainless steels [28].

Extensive experience concerning the behavior both of pure cresols and of cresols mixed with other substances toward metallic and nonmetallic materials is described in [27].

1.3. Formation and Isolation [29], [30]

In addition to phenol, cresols, xylenols, and numerous other phenols are found in the waste-

water of cracking processes [31], tars, and tar-like products (→ Tar and Pitch) formed in the purely thermal cracking (carbonization and coking, → Coal Pyrolysis) [29], [30], oxidizing thermal cracking (→ Gas Production) [29], [30], and hydrogenating thermal cracking (→ Coal Liquefaction [29], [30], Coal Pyrolysis [32]) of such natural materials as bituminous coal [33], lignite, [34], oil shale [35], peat, wood [36], lignin [37], and other biomaterials [38], [39].

The yields of cresols, xylenols, and other phenols and their quantity ratio depend not only on the starting material, but also largely on the process conditions (especially the temperature and the residence time), the type of reactor, and the mode of operation. With bituminous coal, hydrogenation provides the highest yields, low-temperature carbonization gives intermediate amounts, while high-temperature coking produces the lowest yields; gasification of lignite gives significantly more phenol, cresols, and xylenols than that of bituminous coal (Table 3).

Small amounts of cresols, xylenols, and other phenol derivatives are also formed in the catalytic and thermal cracking (especially coking thermal cracking) of petroleum fractions [40].

At the present time cresols and xylenols are isolated from coal tars obtained in the high-temperature coking process, low-temperature carbonization, and Lurgi pressure gasification of coal (see Table 3), and from spent refinery caustics.

1.3.1. Isolation from Coal Tars

A traditional source of tar cresols and xylenols, but one that has been continously declining over the past 40 years (potential ca. 30 000 – 40 000 t/a) is the high-temperature coke-oven tar (composition and work up → Tar and Pitch), obtained in the production of metallurgical coke by horizontal chamber coking of bituminous coal at 900 – 1300 °C [29]. This tar contains, on average, ca. 0.4 – 0.6 wt % phenol, 0.8 – 1 wt % cresols, and 0.2 – 0.5 wt % xylenols. In the United Kingdom, cresols are also produced from low-temperature coal tars obtained in the production of smokeless fuels (semi-coke) according to the Coalite process, where the content of $C_6 - C_8$ phenols is ca. 10 times higher. But by the 1990s these tars had become far less abundantly available. Today the largest source of "natural" cresols and xylenols (35 000 – 55 000 t/a) [41] is provided by the liquid byproducts obtained by Sasol in the pressure gasification of bituminous coal by the Lurgi process to produce synthesis gas for South Africa's three Fischer – Tropsch plants [42], which are similar in composition to low-temperature tars. In the United States the liquid coproducts from the Lurgi gasifiers of the Dakota Gasification Co., which gasifies lignite to synthetic natural gas (SNG), are an essential source of "natural" cresols and xylenols (potential ca. 5000 t/a *o*-cresol, 11 000 t/a *m*/*p*-cresol, and 10 000 t/a xylenols) [43].

When the starting material is high-temperature tar, the phenols are isolated by extraction with sodium hydroxide solution or, in the Lurgi phenoraffin process (→ Tar and Pitch), from the carbolic oil that boils at 180 – 210 °C, from the light oil, and from the filtrate of the naphthalene oil; where other coal tars are concerned the procedures are similar. The hydrocarbons and pyridine bases still present in the crude phenolate caustic are removed by distillation with steam. The crude phenol is then liberated with carbon dioxide. Frequently, the phenolate caustics from coking plant effluents, which contain primarily phenol and cresol and only very small amounts of xylenols [44], are incorporated. The composition of the resulting crude phenol may, therefore, vary extensively. For example, the composition of crude phenol obtained in this way from German high-temperature coke-oven coal tar may be as follows [33]: 15 % water, 30 % phenol, 12 % *o*-cresol, 18 % *m*-cresol, 12 % *p*-cresol, 8 % xylenols, and 5 % trimethylphenols together with higher boiling phenols.

After the alkali content of the crude phenol has been lowered from ca. 2 % to ca. 0.3 % by scrubbing with water, the crude phenol is dehydrated azeotropically and rectified under vacuum into the following fractions: phenol, *o*-cresol, *m*-/*p*-cresol mixture, xylenol, and phenol tar. As the difference between the boiling points of *m*- and *p*-cresol is very small (0.3 °C), these compounds are obtained as a single fraction in which small amounts of 2,6-xylenol and 2-ethylphenol are also present. The ratio of *m*-cresol to *p*-cresol is ca. 60 : 40 for high-temperature tar and ca. 50 : 50 for low-temperature tar.

Table 3. Average yields of phenols in the tars of various cracking processes∗

Phenol	Sump-phase hydrogenation, coal oil	Low-temperature tars	Lurgi gasifier tar	Lignite tar	High-temperature coke-oven tars (1200 – 1300 °C)		
	(ca. 470 °C)	(600 – 650 °C)	(800 – 900 °C, 20 bar)		USA	UK	FRG
Phenol	16	1.4	0.14	6.1	0.20	0.22	0.12
o-Cresol	6	1.4	0.16	1.6	0.08	0.12	0.06
m-Cresol	14	0.9	0.26	1.7	0.14	0.17	0.12
p-Cresol	6	0.8	0.21	1.7	0.09	0.10	0.06
Xylenols	21	6.1	0.78	3.0	0.12	0.18	0.06
Higher phenols	22	12.4	1.67	–	0.27	0.34	0.05
Sum of phenols	85	21.6	3.22	–	0.90	1.13	0.47
Sum of cresols	26	3.1	0.63	5.0	0.31	0.39	0.24

∗ Data in kilogram of phenol per tonne of water-free bituminous coal.

The isolation of cresylic acids from the liquid coproducts obtained by gasification of North Dakota lignite [43] is possible in modifications of known industrial processes [45]. For the further purification of the cresylic acid fractions from neutral oils, tar bases, sulfur compounds, and undesirable phenolic substances (e.g., 2,6-xylenol, *o*-ethylphenol, guaiacol, and other alkoxyphenols), an extractive distillation with diethylene glycol was developed by the Dakota Gasification Co. [46].

Examples of companies that isolate cresols (and xylenols) from coal tar are Sasol Phenolics since 1993 in South Africa, the Merichem Co. in the United States in joint venture with Sasol (named Merisol) and in contract with the Dakota Gasification Co. in the United States (Dakota Gas until now only isolates phenol from the coal tars and markets the remaining crude cresylic fractions for isolation by others), Coalite Chemicals in the United Kingdom (also in contract with Dakota Gas), Rütgers-VfT AG in Germany, DEZA Corporation in the Czech Republic (works up coal tars from Eastern European countries), and in Japan the Nippon Steel Chemical Co., Sumikin Chemical Co., Kansai Distillation Co., and ADCHEMCO Corporation.

1.3.2. Recovery from Spent Refinery Caustics

In the United States, cresols and xylenols are obtained partly from the naphtha fractions produced by catalytic and thermal cracking in the petroleum industry, which contain, on the average, ca. 0.1 % $C_6 - C_8$ phenols [40], [47]. In the scrubbing of the sulfur compounds contained in these fractions (hydrogensulfides, alkyl- and arylthiols) with concentrated alkaline solutions, a process known as "sweetening," the acid cresols and xylenols are also extracted. The composition of the spent cresylate caustics fluctuates widely, and they contain on the average 20 – 25 % $C_6 - C_8$ phenols and 10 – 15 % sulfur compounds. Until around 1990 the caustics were collected by the reprocessing firms Merichem, Northwest Petrochemical, and PMC Specialties (formerly Productol), and reprocessed in central plants by a variety of processes. Because of competition, especially from synthetic *o*-cresol producers, increased environmental restrictions, and an overall lack of raw material, Northwest Petrochemical and PMC Specialties have closed their operations, and Merichem, the biggest of the three, is now the only processor of spent refinery caustics in the United States.

At Northwest Petrochemical in Anacortes, Washington, the thiols in the alkaline solution were first oxidized with air to give disulfides, which were then decanted as an oily layer [48]:

$$RSNa + R'SNa + 0.5\,O_2 + H_2O \longrightarrow RSSR' + 2\,NaOH$$

The phenols were then precipitated from the aqueous alkaline phase in a packed column with a countercurrent stream of carbon dioxide (flue gas) and decanted. The phenols still present in the carbonate – hydrogencarbonate phase (2 – 6 %) were extracted with an organic solvent, from which they are extracted again with an aqueous alkaline phase, which was then returned

to the column. A typical composition of the phenol mixture obtained in this way might be as follows: 20 % phenol, 18 % *o*-cresol, 22 % *m*-cresol, 9 % *p*-cresol, 28 % xylenol (all six isomers plus ethylphenol), and 3 % higher phenols. This mixture was then separated by distillation into phenol, *o*-cresol, *m*-/*p*-cresol mixture, and xylenols, or mixtures in which particular constituents predominate.

The Merichem Co. of Houston, Texas, precipitates the phenols with hydrogen sulfide, a waste product of the refineries in Houston [49], [50], instead of using carbon dioxide. After extraction with a solvent, the sulfide solution obtained in this process is concentrated in a triple-effect evaporator to give sodium sulfide for use in the paper industry and in ore dressing [50].

The water vapor condensate formed at this stage is stripped with natural gas to remove odorous compounds, after which it is passed to a cooling tower. The natural gas is then used as a fuel in the final process air incinerator. Before being introduced into the waste air incinerator, which is operated at ca. 1370 °C, the process air from all of the other parts of the plant is scrubbed with alkaline solution. In addition, odor-neutralizing chemicals are sprayed alongside the plant area. All of the water at the plant (including rain water) is recycled in an integrated system. Water is released into the environment almost completely as water vapor from the cooling towers [50], [51].

In the process operated until 1971 by the Pittsburgh Consolidated Chemical Co. [52], [53], the spent refinery caustics were treated directly with flue gas containing 8 – 14 % carbon dioxide. From the crude phenol thus precipitated, a fraction boiling at 100 – 235 °C was removed by distillation and extracted continuously in a countercurrent of aqueous methanol – light naphtha. After removal of the final thiol and base residues with an ion-exchange resin, the cresylic acids that had entered the aqueous methanolic phase were recovered by fractional distillation.

In 1996, the "cresylics" capacity of Merichem was ca. 55 000 t/a [54]. However, this figure includes phenol, the xylenols, and several other alkylphenols. In addition, Merichem no longer isolates cresylics exclusively from spent refinery caustics, but also works up cresylics from a variety of sources, e.g., from the coal gasification plants of Dakota Gas and Sasol.

The recovery of cresols and xylenols from spent refinery caustics has so far been confined to the United States. Because different refinery techniques are used in Germany, the production of cresol-containing spent caustics is insufficient for economical reprocessing; or no spent caustics are produced at all because desulfurization is carried out differently, e.g., by hydrotreating. In the United States as well, the recovery of cresol from spent refinery caustics is stagnating or even receding because the refineries are likewise changing over either to hydrotreating or to UOP's Merox process, partly because the price of sodium hydroxide has risen. With the Merox process, substantial quantities of cresols remain in the gasoline, and considerably smaller amounts of spent caustics are obtained.

1.4. Production

Since 1965, when the recovery of cresols from coal tar and spent refinery caustics had become insufficient to meet the rising demand, these compounds have been increasingly produced by synthesis. The processes now in use are

- Alkali fusion of toluenesulfonates
- Alkaline chlorotoluene hydrolysis
- Splitting of cymene hydroperoxide
- Methylation of phenol in the vapor phase

The first three processes start from toluene and were developed from the corresponding benzene – phenol syntheses; to some extent they are even carried out in converted plants that formerly produced phenol. In the following descriptions, attention will be primarily paid to the differences between these processes. The methylation of phenol is a process specifically developed to produce cresols and xylenols. Each process gives a different isomer distribution, and all are, therefore, of individual importance.

1.4.1. Alkali Fusion of Toluenesulfonates
[55], [56], [57]

Alkali fusion is used mainly to manufacture *p*-cresol and consists of four reaction steps:

$$2\,CH_3-C_6H_5 + 2\,H_2SO_4 \rightleftarrows 2\,CH_3-C_6H_4-SO_3H + 2\,H_2O \quad (1)$$

$$2\,CH_3-C_6H_4-SO_3H + Na_2SO_3 \longrightarrow 2\,CH_3-C_6H_4-SO_3Na + H_2O + SO_2 \quad (2)$$

$$2\,CH_3-C_6H_4-SO_3Na + 4\,NaOH \longrightarrow 2\,CH_3-C_6H_4-ONa + 2\,Na_2SO_3 + 2\,H_2O \quad (3)$$

$$2\,CH_3-C_6H_4-ONa + SO_2 + H_2O \longrightarrow 2\,CH_3-C_6H_4-OH + Na_2SO_3 \quad (4)$$

$$2\,CH_3-C_6H_5 + 2\,H_2SO_4 + 4\,NaOH \longrightarrow 2\,CH_3-C_6H_4-OH + 2\,Na_2SO_3 + 4\,H_2O$$

The sulfonation of toluene (Eq. 1) is normally carried out with concentrated sulfuric acid at 120 – 130 °C and atmospheric pressure. To ensure that substantially all of the sulfuric acid has reacted, the water of the reaction is entrained with excess toluene vapor. The sulfonic acid mixture is neutralized with sodium sulfite (and/or sodium hydroxide) (Eq. 2) and then fused with excess sodium hydroxide at 330 – 350 °C (Eq. 3). The melt is introduced into water, the sodium sulfite filtered off, and the aqueous phase acidified with sulfur dioxide from reaction step (2) and/or with sulfuric acid (Eq. 4). This gives an aqueous phase containing sodium sulfite, which is returned to the neutralization unit (2), and a crude cresol phase, which is dehydrated by azeotropic distillation. The plant and process are similar to those formerly used in the synthesis of phenol by alkali fusion of sodium benzenesulfonate. The dehydrated crude cresol phase is separated by fractional distillation into phenol, *o*-cresol, *m*-/*p*-cresol mixture, and a residue containing ditolyl sulfones, xylenols, higher phenols, and other compounds. The ditolyl sulfones are formed in the sulfonation; the phenols and xylenols are formed in small quantities in the alkali fusion [58], [59]. See [58] for the mechanism of the fusion reaction.

The distribution of the cresol isomers depends mainly on the sulfonation conditions. Normally a mixture of the following composition is obtained: 6 – 12 % *o*-cresol, 6 – 12 % *m*-cresol, 80 – 85 % *p*-cresol [55], [60]. After the *o*-cresol has been removed by distillation, a *p*-cresol of ca. 90 % purity can be obtained by distillation. From this, *p*-cresol with a *m*-cresol content of only ca. 1 % can be isolated relatively easily by melt crystallization (see Section 1.5.1).

A cresol yield of 80 %, based on toluene, is possible [60]. The yield falls as the *m*-cresol content of the cresol mixture increases [55], [59].

Under kinetically controlled (mild) sulfonation conditions and mild alkali fusion, it is possible to produce cresol mixtures with an even smaller content of *m*-cresol. It has been claimed that sulfonation with chlorosulfuric acid at 33 – 45 °C gives a product free from *m*-cresol at a yield of 90 % (calculated on the basis of toluene) and with an *o*-/*p*-cresol ratio of 15 : 85 [61]. The sulfonation of toluene with sulfur trioxide – sulfur dioxide mixtures at 25 – 50 °C likewise leads to toluenesulfonic acid mixtures almost free of the *m* isomer, and which have an *o*-/*p*-isomer ratio as low as 3 : 97 [62]. In the absence of sulfur dioxide as a solvent, it may even be possible to obtain an *o*- and *m*-isomer content of only 0.5 % when the process is carried out in a loop reactor at 0 °C, and when gaseous SO_3, diluted with an inert gas, is introduced into the reactor at several locations [63]. The inert gas (waste gas) is not necessary if the process is carried out under vacuum [64]. With oleum or sulfuric acid, *m*-free sulfonation to give mainly *p*-toluenesulfonic acid is possible if the reaction is carried out with a short residence time in a rotating film reactor [65].

Thermodynamically controlled sulfonation conditions (high temperature, long reaction time, and the use of a sulfonation mixture with a low water content) lead to an equilibrium mixture containing ca. 5 % *o*-, 54 % *m*-, and 41 % *p*-toluenesulfonic acid [66]. Alkali fusion of such mixtures then gives an isomer distribution of ca. 5 % *o*-cresol, 56 % *m*-cresol, and 39 % *p*-cresol [59].

Technical-grade *m*-cresol can also be produced by toluene sulfonation according to a process developed by the Honshu Chemical Industry Co. [67]. Steam at 165 °C is introduced into an toluenesulfonic acid mixture rich in *m* isomer obtained under thermodynamically controlled conditions as described above (isomerization for several hours at 190 – 200 °C), whereupon *o*- and *p*-toluenesulfonic acid are hydrolyzed back to sulfuric acid and toluene. The toluene is distilled with the steam; *m*-toluenesulfonic

acid with a purity above 90 % remains, together with sulfuric acid. Side reactions during the isomerization and hydrolysis can be suppressed by adding sodium sulfate (5 – 10 mol%, calculated on the basis of sulfuric acid) to the sulfonation batch. In this way, removal of the sulfuric acid that is formed in the hydrolysis as a lower phase at ca. 140 – 150 °C is also possible [68]. It is also possible to neutralize the hydrolysis mixture with concentrated sodium hydroxide solution, filter off the resulting sodium sulfate at 80 °C, introduce the solution of sodium *m*-toluenesulfonate into a sodium hydroxide melt, containing 10 – 15 % potassium hydroxide to facilitate stirring, at an initial temperature of 330 °C, and to fuse it at 340 °C.

The distillation of the crude cresol phase obtained after acidification gives *m*-cresol with a purity of up to 98 % in a yield of ca. 65 %, calculated on the basis of reacted toluene [68].

p-Cresol is produced according to the toluene sulfonation process by PMC (formerly Sherwin-Williams; capacity ca. 18 000 t/a) in the United States, by Inspec Fine Chemicals (formerly Synthetic Chemicals Ltd., now a part of Laporte; capacity ca. 12 000 t/a; sulfonating agent is SO_3) in the United Kingdom, by Beraton (capacity ca. 3000 t/a) in Russia, and in Japan by Konan Chemicals (ca. 4000 t/a). Honshu Chemical Industry Co. in Japan has a standby plant (ca. 2500 t/a) that can produce *m*- or *p*-cresol. A plant designed to operate the Honshu process is also reported to be capable of producing 3000 – 5000 t/a of *p*-cresol or *m*-cresol has been built in India by Gujarat Aromatics.

The toluenesulfonic acid – cresol process is relatively simple as regards the plant required. Its primary drawback is the unavoidable formation of sodium sulfite in aqueous solution. At Inspec Fine Chemicals, the sodium sulfite is upgraded into a product suitable for the paper industry [69]. At PMC, sodium sulfite is oxidized with air to form sodium sulfate [70], but *p*-cresol is still present in the aqueous sulfite solution (3000 – 5000 ppm); to avoid stripping, thereby causing an odor problem in the neighborhood of the plant, the solution is first passed through columns of activated carbon to adsorb the cresol. The activated carbon is regenerated every 24 h with 10 % sodium hydroxide solution. The amount of *p*-cresol recovered in this way is sufficient to pay for the additional installations [71].

1.4.2. Alkaline Chlorotoluene Hydrolysis

The hydrolysis process is important in the production of cresol with a high *m*-cresol content, and is used in Germany by Bayer AG, the world's largest manufacturer of synthetic cresols (capacity more than 30 000 t/a).

In the first reaction step, a mixture of *o*- and *p*-chlorotoluene in a 1 : 1 ratio is produced by chlorinating 1 mol of toluene with 1 mol of chlorine in the presence of iron(III) chloride and disulfur dichloride (→ Chlorinated Hydrocarbons). With $FeCl_3$ and other cocatalysts *o/p* ratios from 3 : 1 [72] to 1 : 2 [73], [74] are possible.

$$CH_3-C_6H_5 + Cl_2 \longrightarrow CH_3-C_6H_4-Cl + HCl$$

In the next reaction step, this mixture is hydrolyzed with excess sodium hydroxide solution (2.5 – 3.5 mol/mol) at 360 – 390 °C and 280 – 300 bar (28 – 30 MPa) [75].

$$CH_3-C_6H_4-Cl + 2\,NaOH \longrightarrow CH_3-C_6H_4-ONa + H_2O + NaCl$$

The exothermic reaction is carried out continuously in high-pressure tubes several hundred meters long; the tubes are made of nickel steel to withstand the corrosive action of the reaction mixture. Separation of the components, which are introduced via a mixing chamber, is prevented by causing the reaction mass to flow at high velocity, with frequent changes in the direction of flow; finally, the reaction mixture is homogeneous.

The cresol in the resulting sodium cresolate solution is set free by neutralization; for this purpose the hydrochloric acid formed during chlorination can also be used. The coproduct sodium chloride can be returned to the chlor-alkali electrolysis.

For economic reasons, it is important that the starting materials are produced at locations sufficiently close to one another that conveyance through pipelines is possible. The technology of the process largely corresponds to that of the Dow – Bayer chlorobenzene – phenol hydrolysis process [76]. In the manufacture of cresols, however, the amounts of the byproducts formed

are larger than in this process: these include ditolyl ethers [bis(methylphenyl) ethers], in particular, tolylcresols [methyl(methylphenyl)phenols], and small amounts of toluene, phenol, benzoic acid, and off-gas from cracking (methane and hydrogen). The removal of small amounts of phenol by azeotropic distillation, e.g., with chlorotoluene, is dealt with in [77].

Whereas it is hardly possible to control the scale on which tolylcresol is formed, the formation of ditolyl ether can be controlled via the temperature, the residence time, and the sodium hydroxide concentration; in addition, the scale on which this byproduct is formed can be minimized by recycling. This possibility arises from the fact that ditolyl ethers are hydrolyzable, albeit less easily than the homologous diphenyl ether. Therefore, if the reaction is optimally controlled, yields of ca. 80 %, based on chlorotoluene, are attainable. However, ditolyl ether has several uses, e.g., as a heat transfer medium (Diphyl DT), as an electrical insulant (Baylectrol 4800; replacement for polychlorinated biphenyl), and as a precursor for tanning agents. Therefore, ditolyl ether is no longer recycled; indeed, the conditions may be adjusted to increase the selectivity for this compound.

The process gives very pure cresols, i.e., products that are almost free from other compounds. With a 1 : 1 *o*/*p*-chlorotoluene mixture, the *o*-, *m*-, and *p*-cresol isomers are present in a ratio of ca. 1 : 2 : 1. This ratio differs from that of the chlorotoluene input because isomerization via aryne intermediates is assumed to occur [78], [79] under the severe conditions of hydrolysis. After the *o*-cresol has been distilled, a *m*-/*p*-cresol mixture containing ca. 70 % *m*-cresol is obtained.

The plant can also be operated on pure *o*- or *p*-chlorotoluene. The hydrolysis of *p*-chlorotoluene then gives a 1 : 1 *m*-/*p*-cresol mixture, and hydrolysis of *o*-chlorotoluene gives a 1 : 1 *o*-/*m*-cresol mixture. Technically pure *m*-cresol is obtainable from the latter by distilling off the *o*-cresol.

Hydrolysis of chlorotoluene without substantial isomerization is possible under milder conditions (200 – 350 °C) with alkali metal hydroxide [79], [80], alkali metal carbonates [81], or ammonium hydroxide [82], or with alkali metal acetate or alkali metal propionate solution [83] in the presence of copper or copper compounds. These processes are not used industrially, among other things because the space – time yield decreases drastically at lower temperatures. Still the best results (nearly 100 % *p*-chlortoluene conversion in 3 – 6 h) are obtained at 250 – 280 °C with alkali metal hydroxide in the presence of copper and/or a copper compound and an additional oxidizing agent like air (O_2) or hydrogen peroxide [84].

1.4.3. Cymene Hydroperoxide Cleavage

The synthesis via cymene hydroperoxide, which is also known as the cymene – cresol process, permits the production of *m*- or *p*-cresol from the corresponding cymenes (isopropyltoluenes). It is unsuitable for the production of *o*-cresol. The process consists of three reaction steps:

1) Toluene propylation and cymene isomerization

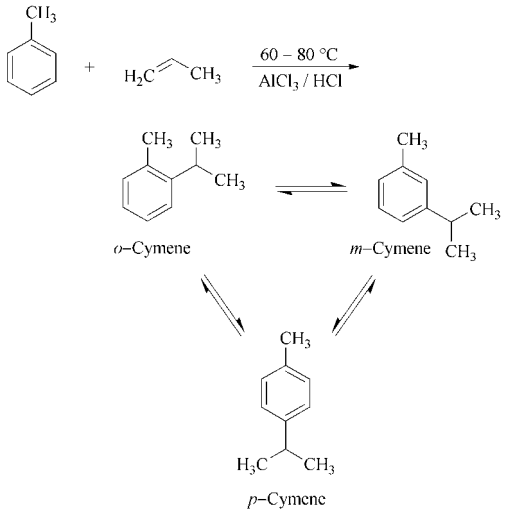

2) Oxidation of cymene to cymene hydroperoxide

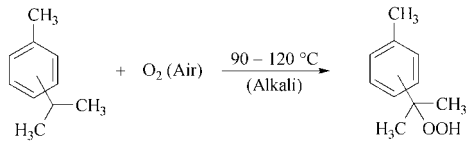

3) Peroxide cleavage

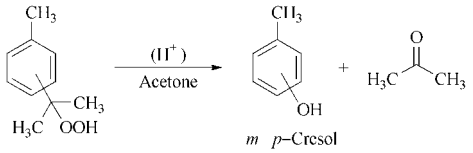

By comparison with the analogous Hock cumene process (→ Acetone), the cymene – cresol process has a lower space – time yield, involves considerably more costly distillation procedures, necessitates more extensive treatment of wastewater, and gives a product yield that is 20 – 30 % lower. The reasons for this are as follows:

1) All three isomers, *o*-, *m*-, and *p*-cymene, are formed in the propylation of toluene. Of the three isomers, *o*-cymene not only is hardly oxidized but also increasingly inhibits oxidation of the other cymenes [85] as its concentration in the isomer mixture rises. For continuous oxidation, the *o*-cymene content of the cymene mixture must be kept below 10 % at all times [86], [87]. Thus, cymene mixtures are needed that contain as little *o*-cymene as possible.
   The lowest *o*-cymene content is that of a cymene mixture in thermodynamic equilibrium [87], [88]. Aluminum chloride must be used to reach this equilibrium as quickly as possible [89–91]. The isomer ratio obtainable with aluminum chloride under practical conditions is ca. 3 % *o*-, 64 % *m*-, and 33 % *p*-cymene. As the *o*-cymene concentration rises while the oxidation is in progress, a large amount of the cymene must be returned to the first stage of the process after the cymene hydroperoxide and other byproducts have been separated, so that the *o*-cymene content can again be lowered to ca. 3 % by isomerization on aluminum chloride. The diisopropyltoluenes formed as byproducts in the alkylation and separated by distillation are also returned to this isomerization step.
2) The oxidation rates of the cymenes are lower than that of cumene. In addition, the cymenes can be oxidized up to a peroxide content of only ca. 20 %, whereas oxidation levels exceeding 30 % are possible with cumene. The proportion of unreacted cymene that must be distilled and returned to the reaction is therefore greater than in the cumene – phenol process. At higher degrees of oxidation, the formation of byproducts from excessive oxidation of the isopropyl group to dimethyl(tolyl)methanol and methylacetophenone is greatly increased.
3) The methyl group of the cymenes is likewise oxidized; for details of the mechanism of *p*-cymene oxidation, see [92]. The ratio at which the methyl group and isopropyl group are subjected to oxidative attack may be as high as 1 : 4 [93]. As the primary peroxide is less stable than the tertiary peroxide, it is more easily converted into secondary products (isopropylbenzyl alcohol, isopropylbenzaldehyde, cuminic acid). A more important drawback is that in the highly exothermic acid cleavage, in addition to isopropylphenol, 1 mol of formaldehyde is formed per mole of primary peroxide still present; because formaldehyde is able to bind up to 2 mol of cresol as a formaldehyde resin, the cresol yield is substantially diminished. Moreover, the presence of resin makes the product more difficult to work up. Various solutions to the problem have been proposed: (1) selective decomposition of the primary peroxide during the actual oxidation by adding catalytic amounts of picoline, pyrrole, thiophene [94], or triphenylphosphine [95]; (2) selective hydrogenation of the primary peroxide over a noble-metal catalyst [96], [97] after the oxidation in a separate stage of the process (before the acid cleavage) or decomposition by passing the primary peroxide over a magnesium aluminate catalyst [98] or treatment with sodium hydroxide and heavy metals [99] or organic quaternary ammonium salts [100]; (3) extraction of the primary peroxide [101–104]. However, it is simpler to dissolve the products resulting from methyl group oxidation continuously and as fully as possible in excess alkali solution present during the actual oxidation itself [85], [86], [105]. The organic acids present in the spent caustics (ca. 0.5 kg of cuminic acid, toluic acid, phthalic acid, and other products per kilogram of cresol) are largely precipitated during the acidification and can be removed by filtration. Addi-

tional removal from the filtrate (wastewater) is possible by extraction [106].
According to another suggestion, hydrogen peroxide should be present during cleavage of cymene hydroperoxide to cresol [107]; alternatively, this cleavage could be effected under mildly hydrogenating conditions [108]. These measures decrease the extent of yield-lowering side reactions undergone either by the primary peroxide or by its secondary product, formaldehyde.

4) The reaction mixture after acidic peroxide cleavage contains not only cresol, acetone, and unreacted cymene, but also isopropenyltoluene, 2-tolylpropanol, isopropylbenzyl alcohol, isopropylbenzaldehyde, methylacetophenone, isopropylphenol, and numerous other compounds. The separation of these byproducts is relatively expensive because of their large number, their similar boiling points, and the formation of azeotropic mixtures [109]. The amount of byproducts can be decreased if first the primary peroxide is selectively hydrogenated over a Pd/C or Pd/titania catalyst (see above) and then the reaction mixture, after concentration by distillation and acidic peroxide cleavage, is again hydrogenated over the same catalyst type under somewhat harsher conditions. The thereby re-formed cymene increases the yield of cresol (based on the cymene consumed) to 80 – 85 % [97].

Cymene – cresol plants, in which cymene mixtures are used and which have a nameplate capacity of 22 000 t/a each, have been operated in Japan by Sumitomo [110] and Mitsui since 1969. The product has a *m*- and *p*-cresol content of more than 99.5 % and a *m/p* ratio of 60 : 40. The acetone coproduct is of high quality [111].

If pure *p*- or *m*-cymene is used, the high cost of recycling and isomerization necessitated by the presence of *o*-cymene is eliminated, and pure *p*- or *m*-cresol can be obtained directly. For example, until 1972 the Hercules Powder Co. operated a plant at Gibbstown in the United States for the manufacture of *p*-cresol from *p*-cymene obtained from natural terpenes [112].

Owing to the close proximity of the boiling points of the cymene isomers (*m*-cymene 175.1 °C, *p*-cymene 177.1 °C, *o*-cymene 178.3 °C, at 101.3 kPa), the *m*-cymene, at most, may possibly be isolated at reasonable cost from the cymene mixture resulting from the propylation of toluene (see above) [91]. However, both pure *m*-cymene and pure *p*-cymene can be isolated according to UOP's Cymex process on *p*-cymene-selective 13X molecular sieves by using toluene as desorption agent [113]. The process corresponds to UOP's Parex process for the isolation of *p*-xylene from C_8 aromatics (→ Xylenes). A cymene – cresol plant, based on the UOP and Hercules Powder technologies, for the manufacture of pure *m*-cresol and *p*-cresol [for the antioxidant 2,6-di-*tert*-butyl-4-methylphenol (BHT)] was started up in 1986 at the Beijing Yanshan Petrochemical Corp. in China (nameplate capacity ca. 20 000 t/a). But it seems the facility is now also used for the production of phenol by the cumene process [114].

If in the Cymex process only one isomer, e.g., *p*-cymene, is required, the *m*-/*o*-cymene mixture leaving the adsorption column, which is impoverished with regard to *p*-cymene, can be reisomerized and returned to the separator [115]. Since it is less important in this case to have the lowest possible *o*-cymene content, the isomerization (and the alkylation) can, for example, be carried out on phosphoric acid – aluminum oxide catalysts at 370 °C instead of using aluminum chloride [115].

In the manufacture of *m*-cresol, *m*-cymene can also be obtained by selective cracking of the *p*-cymene present in cymene mixtures with a low *ortho* isomer content. Cleavage of *p*-cymene (and to a lesser extent that of *o*-cymene) to toluene and propylene is accomplished at 300 – 400 °C on shape-selective ZSM-5 zeolites [116] that are particularly selective in the MgO-modified Li^+-ZSM-5 form [117]. A further, technically less attractive way of isolating *m*-cymene consists of complexing with anhydrous hydrofluoric acid – boron trifluoride at temperatures below − 40 °C [118].

1.4.4. Methylation of Phenol

Synthetic cresol, especially *o*-cresol, and xylenol are now produced largely by methylation of phenol with methanol in the presence

of catalysts. 2,6-Xylenol is produced almost entirely by this method.

The process consists of only one reaction step, but is based on phenol, which is relatively expensive; the separation of products from the reaction mixture is also costly. This is partly because the boiling points of several components are very near one another, and partly because in some cases the purity of the product must meet very strict requirements.

The reaction can be carried out either in the vapor phase or in the liquid phase.

Methylation in the Vapor Phase. Vapor-phase methylation is used mainly to manufacture pure *o*-cresol and/or pure 2,6-xylenol:

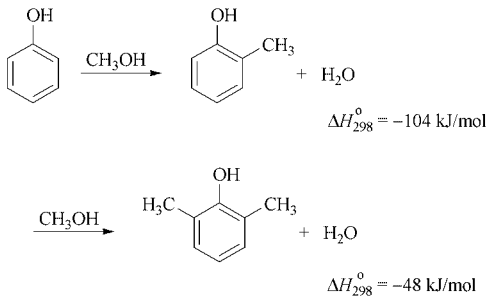

Side reactions:

$$CH_3OH \longrightarrow CO + 2\,H_2O \quad \Delta H^0_{298} = -91\ \text{kJ/mol}$$

$$CO + H_2O \longrightarrow CO_2 + H_2 \quad \Delta H^0_{298} = -41\ \text{kJ/mol}$$

$$CO + 3\,H_2 \longrightarrow CH_4 + H_2O \quad \Delta H^0_{298} = -206\ \text{kJ/mol}$$

At atmospheric or slightly elevated pressure, a superheated mixture of the vapors of phenol, methanol, and water is passed at a liquid hourly space velocity (LHSV) of $1-2\ h^{-1}$ over a metal oxide catalyst fixed in a multitubular reactor of stainless steel at a temperature of 300 – 460 °C [119–121]. The reaction temperature required depends on the nature of the catalyst and on the desired composition of the product. The temperature is permitted to deviate from this level by only a few degrees.

The heat of the reaction is dissipated by boiling organic heat-transfer media [122], by circulating salt melts [120], or through the generation of high-pressure steam [121]. The slight loss of catalytic activity that occurs in the course of time is compensated by raising the temperature accordingly. Carbon deposits are removed from time to time by burning. The presence of water in the feedstock mixture suppresses the decomposition of methanol, extends the regeneration cycle, and prolongs the lifetime of the catalyst [123], [124], which may then endure for several thousand hours.

The product mixture leaving the reactor gives up some of its heat to the feedstock mixture in a heat exchanger, and is then condensed. The uncondensable reaction products formed from the decomposition of methanol (H_2, CH_4, CO_2, and CO) are used as fuel gas in steam production. Aqueous methanol is first distilled from the liquid reaction product and then recycled. Separation into methanol, water containing phenol, and the phenols can be achieved in a single distillation column with sidestream [125]. It is also possible to decant an aqueous methanol phase and to recycle it after adsorbing the contained phenols with an resin adsorbent (desorbing can be achieved with methanol) [126]. Water remaining in the product is then removed either by azeotropic distillation with toluene [122] or as an azeotropic mixture with unconverted phenol [121]. In the latter case, it is necessary to strip the phenol from the wastewater. The dehydrated mixture is then fractionated into phenol ether, phenol, *o*-cresol (≥ 99.5 %), and 2,6-xylenol in an additional series of highly efficient distillation columns operated continuously under vacuum. The phenol ether – phenol fraction is recycled. The distillation residue, which contains, among other compounds, 2,3-, 2,4-, and 2,5-xylenol and 2,4,6- and 2,3,6-trimethylphenol, is usually burned as a fuel oil. If the catalyst has high ortho selectivity, pure (99.5 %) 2,6-xylenol can be obtained directly. If that is not the case or if the reaction conditions are too severe (high temperature, long residence time), the 2,6-xylenol fraction will contain *m*- and *p*-cresol, which must be removed by special methods (Section 2.3). The ratio of *o*-cresol to 2,6-xylenol can be controlled by altering the methanol – phenol ratio, and if the *o*-cresol is recycled, 2,6-xylenol comprises the sole product. Over an extended production time, the yield (selectivity) of *o*-cresol and 2,6-xylenol, calculated on the basis of reacted phenol, is 90 – 98 %.

With regard to the use of a fluidized-bed reactor process and a high-performance silica-supported Fe – V oxide catalyst developed by

Asahi Chem. Ind. for the concurrent production of *o*-cresol and 2,6-xylenol at 320 – 340 °C (e.g., 84 % phenol conversion and 98 % selectivity at 336 °C), see [127], [128–132]. For kinetic data on a promoted Mn oxide catalyst and the use of adiabatic reactors for phenol methylation, see [133]. Vapor – liquid equilibrium data for the alkylation product are given in [134], and details of the distillation in [135–137]. An evaluation of the vapor-phase methylation of phenol to cresols and xylenols over a solid acidic catalyst is reported in [138].

Magnesium oxide catalysts, in particular, are used. They require relatively high operating temperatures (420 – 460 °C), but are so *ortho*-selective that *m*- and *p*-cresol are hardly formed, even at 540 °C [139]. The activity (phenol conversion), as well as the ratio of 2,6-xylenol to *o*-cresol and the *ortho* selectivity depend on the conditions under which the magnesium oxide is produced [140–142], [143]. These properties can be further improved by combining magnesium oxide with other oxides, such as those of the alkali metals [144], manganese [141], [145–147], copper [141], [142], [145], [148], titanium, uranium, zirconium, chromium [149], cerium [150], and silicon [151]. At 440 °C and at a liquid hourly space velocity (LHSV) of 2.0 h^{-1} the continuous reaction of a mixture of phenol, methanol, and water in a 1 : 4 : 2.4 molar ratio over an MgO catalyst thus gives a phenol conversion of ca. 82 %, and a molar ratio of 2,6-xylenol, *o*-cresol, *p*-cresol, 2,4-xylenol, and 2,4,6-trimethylphenol of 69.5 : 23.0 : 0.05 : 0.8 : 6.6. If the catalyst also contains ca. 1 % of copper, the conversion rate is 89 %, and the corresponding molar ratio of the products is 73.3 : 18.9 : 0.06 : 0.07 : 7.0. The waste gas volume is in both cases 0.5 mol per mole of reacted phenol [142]. The mechanism of *ortho* methylation is discussed in [152].

With *manganese oxide* catalysts [153] or *chromium oxide* catalysts [154–156] the methylation can be carried out at ca. 390 °C; and on catalysts based on *vanadium oxide* [157], [158] or *iron oxide* [159–164], [153, 165–168] at ca. 350 – 320 °C, provided that these oxides are combined with smaller amounts of other oxides (e.g., those of Ce, Cr, Mn, Co, Ni, Zn, Al, Ga, In, Si, Ge, Sn, Pb, Bi, Mg, Na) or with one other [127], [130], [131], [169–172]. By addition of, e.g., MgO [168] or Ce – Cr oxide [173] and doping with K_2O, the phenol conversion of the silica-supported V – Fe oxide fluidized-bed catalysts [130], [131] can be enhanced by 10 – 15 % to 93 – 99 % at 320 °C. These mixed oxide catalysts show excellent *ortho* selectivity and give a very high 2,6-xylenol/*o*-cresol ratio and virtually no *p*-cresol. At phenol/methanol molar ratios of ca. 1 : 3 – 6 and liquid hourly space velocities (LHSV) of ca. 1 h^{-1} they permit almost complete conversion of phenol, with 2,6-xylenol yields of up to ca. 97 % (calculated on the basis of phenol). Combination with other metal oxides also extends the catalyst lifetime (especially if alkali metal compounds are added) and diminishes the extensive decomposition of methanol typically seen with manganese, chromium, vanadium, and iron oxide catalysts. In spite of this measure and the addition of water vapor (generally 1 – 2 mol per mole of phenol), 1 – 2 mol of methanol is usually lost by decomposition for every mole of 2,6-xylenol manufactured. The proposed methods for reducing the loss include the addition of sulfur dioxide, hydrogen sulfide, nitric oxide, carbon monoxide, or carbon dioxide to the feedstock mixture [174], recycling a portion of the cracking gas [175], and introducing the methanol [176] or phenol at different parts of the reactor [177]. Relatively little methanol is decomposed on Fe_2O_3 catalysts doped with silicon oxide and magnesium oxide [162]. For the mechanism of the reaction on mixed oxide contacts, see [178], [179].

γ-Aluminum oxide catalyzes the methylation of phenol at a temperature as low as 300 – 320 °C. Under these conditions, at a methanol : phenol molar ratio of 1 : 2, and at a liquid hourly space velocity (LHSV) of 1 h^{-1}, *o*-cresol, 2,6-xylenol, and *m-/p*-cresol mixture are formed in a molar ratio of ca. 82 : 17 : 1, with a total selectivity of 97 % and a phenol conversion of 33 % [180]. Increasing the methanol : phenol ratio and recycling the *o*-cresol enables the *o*-cresol : 2,6-xylenol ratio to be shifted toward 2,6-xylenol [181].

On γ-Al_2O_3, unlike other methylation catalysts, methanol is scarcely decomposed under the conditions described above. In addition, this catalyst must be regenerated only once in ca. 1000 h, and has a working lifetime of several years [182]. Its disadvantage is that to prevent the formation of significant quantities of other di-, tri-, and tetramethylphenols, the phenol con-

version rate must not substantially exceed 60 %, which means that relatively large amounts of the product must be worked up and recycled. In addition, extensive operations, e.g., extractive distillation with diethylene glycol [181], are necessary to free the 2,6-xylenol from the *p*-cresol that is unavoidably formed on this catalyst.

Below 300 °C, increasing amounts of anisole are formed on γ-Al_2O_3; however, because anisole is isomerized to cresol at elevated temperatures [183], [184], only a small amount ($\leq$ 2.5 %) is present in the reaction mixture under the previously mentioned conditions. As the temperature and residence time are increased, the rate of phenol conversion on Al_2O_3 rises, but the *ortho* selectivity falls. This is also basically true of other catalysts. Instead of *o*-cresol and 2,6-xylenol, increasingly large amounts of *m*- and *p*-cresol, other xylenols, multimethylphenols, and methylbenzenes, especially hexamethylbenzene, are formed [185], [186]. Demethylation occurs above 500 °C [187]. Increasing the temperature also results in the formation of larger amounts of gaseous cracking products, and accelerates the carbonization of the catalyst, thus necessitating more frequent regeneration.

The reaction kinetics and the choice and calculation of an optimum multitubular reactor for phenol methylation on γ-Al_2O_3 are discussed in [188–190]. Industrial plants are described in [119] and [181].

Aluminum oxides with strong acid sites [191], *silica – alumina* [152], [192], *zeolites* [192–198] *aluminum phosphates* [186], [199], and *phosphoric acid – kieselguhr* [199] have enhanced isomerization and transmethylation properties. Therefore, they lead to products with a relatively high *m/p* content, especially if the temperature is simultaneously raised to ca. 450 °C. In addition, at a temperature of 250 – 300 °C, shape-selective zeolites [193–195], [200], [201], e.g., of the HKY type [193], and also phosphoric acid – kieselguhr [199], show remarkably high selectivity for *p*-cresol, giving a *m*-/*p*-cresol mixture with *p*-cresol contents of ca. 85 % [193] or 77 % [199], respectively. *Pentasil zeolites* modified, e.g., with Mn [202], Mg, or P [203], also show *para* selectivity.

As alternatives to methanol, methylbenzenes [205–207], dimethyl ether [208], methylamine [209], methane [210], carbon monoxide – hydrogen mixtures [211], and formaldehyde – hydrogen mixtures [212] have been proposed as methylating agents.

Methylation in the vapor phase is utilized by General Electric Co. in the United States (capacity ca. 5000 t/a *o*-cresol and ca. 70 000 t/a 2,6-xylenol) and in the Netherlands (ca. 20 000 t/a 2,6-xylenol); by Inspec Fine Chemicals (formerly Synthetic Chemicals, now a part of Laporte) in the United Kingdom (capacity ca. 10 000 t/a *o*-cresol and 2,6-xylenol, but the plant is also used for the production of 2,4-xylenol by methylation of *p*-cresol and for 2,5-xylenol and 2,3,6-trimethyphenol by methylation of *m*-cresol); Chemopetrol in the Czech Republic (capacity ca. 10 000 t/a *o*-cresol, *m/p*-cresol, 2,6-xylenol and other xylenols); and in Japan by Nippon Crenol (joint venture of Asahi Chemical Industry and Nippon Steel Chemical Company; capacity ca. 8000 t/a *o*-cresol and 15 000 t/a 2,6-xylenol), GEM Polymers (joint venture of General Electric Company, Mitsui Chemicals Co., and Mitsui Toatsu Chemicals; capacity ca. 2500 t/a *o*-cresol and ca. 11 500 t/a 2,6-xylenol), Mitsubishi Gas Chemical Company (capacity ca. 5000 t/a 2,6-xylenol) and Honshu Chemical Industry Co. (capacity ca. 5000 t/a 2,5-xylenol and mainly 2,3,6-trimehyphenol by methylation of *m*-cresol). A vapor-phase methylation plant is said to be in operation in Russia, and activities are also reported from China [213].

Methylation in the Liquid Phase. Results similar to those achieved in the vapor phase are obtained if the oxide catalysts (particularly γ-Al_2O_3) are suspended in a phenol – methanol mixture and heated in an autoclave to 300 – 400 °C [214–216]. The mixture of phenol and methanol [217] or of *o*-cresol and methanol [218] can also be passed as a liquid over a γ-Al_2O_3 catalyst forming a fixed bed in a tube reactor. At 350 °C and a phenol : methanol molar ratio of 0.7 : 1, the pressure is ca. 35 bar [217].

According to a process developed by Chemisches Werk Lowi (now Great Lakes Chemical Corporation) in Germany, phenol and an equimolar amount of methanol are reacted in an autoclave at 350 – 400 °C in the presence of 2 – 3 wt % *aluminum methylate,* with the water of the reaction, together with some of the methanol, being distilled through a small column.

After the catalyst residue has been removed by filtration, the product is worked up by distillation. Methanol, anisole, and phenol are recycled. The yield of cresols and xylenols is ca. 80 % [219], calculated on the basis of converted phenol (ca. 60 %). The main product is *o*-cresol; *m*- and *p*-cresol, 2,6-, 2,4-, and 2,3-xylenol, and trimethylphenols are also formed. The ratio of the products can be adjusted by altering the phenol : methanol ratio, the temperature, and the residence time. The presence of bases (e.g., pyridine) increases the conversion [220], whereas that of acids raises the proportion of *p*-cresol and 2,4-xylenol in the product [221]. Aluminum phenolate can also be used as a catalyst [222]. If necessary, the process can be operated in a multipurpose plant, the capacity of which may be ca. 2000 – 3000 t/a.

Until 1990 Union Rheinische Kraftstoff AG (meanwhile DEA-Mineraloel AG, Werk UK-Wesseling) in Germany operated a process (also known as the Biller process) that used *liquid-phase catalysts* [223], [224], preferably aqueous zinc halide – hydrogen halide solutions [225], at 200 – 240 °C and pressures of ca. 25 bar (2.5 MPa). After the reaction, the organic components were separated from the aqueous catalyst phase by extraction with benzene or cyclohexane, washed with water several times, and worked up by distillation. The catalyst solution was united with the washing water, concentrated, and recycled [226]. The main products were *o*-cresol, *p*-cresol, 2,4-xylenol, and 2,6-xylenol. At a phenol : methanol molar ratio of 1 : 0.6, yields of up to 98 % were obtainable, calculated on the basis of converted phenol. Compared with other processes, this had the advantages not only that *p*-cresol and 2,4-xylenol are obtained in considerably larger quantities, but also that, provided the reaction conditions were carefully harmonized, the *p*-cresol was almost free of *m*-cresol. The extent of *para* substitution increased with the methanol content of the feedstock mixture, but the yield fell. The process required materials that were highly resistant to corrosion, e.g., a reactor lined with tantalum. The plant capacity was ca. 16 000 t/a. The product ratio was controllable within certain limits. As an example, a possible *o*-cresol, *p*-cresol, 2,4-xylenol, and 2,6-xylenol ratio was 7 : 5 : 2 : 1.

Transmethylation and Isomerization. After phenol has been added to the methylphenol residues either from the methylation of phenol or of different origin, or to xylenol for which there is no other use, reaction can occur in the presence of alkylation catalysts to give a product consisting mainly of *o*-cresol, together with *m*-/*p*-cresol. The reaction may proceed either in the vapor phase or in the liquid phase; the temperature is 400 – 500 °C. Catalysts are aluminum oxide [227], chromium oxide – aluminum oxide [228], silica – alumina [229–231], metal oxide – iron oxide [232], or magnesium oxide – tungsten oxide [233]. In the case of 2,6-xylenol, this transmethylation, accompanied by isomerization, can also be accomplished without a catalyst in the vapor phase at 550 – 600 °C or, batchwise, in the liquid phase at 420 – 470 °C [234]. In this way, the *o*-cresol yield from the methylation of phenol can be raised. For cresol isomer mixtures the Chiyoda Chemical Engineering Construction Co. has developed an isomerization – transalkylation process with phenol on silica – alumina catalysts at 400 °C at normal pressure (LHSV 0.5 h^{-1}) which allows the isolation of pure *o*-cresol by recitification of the resulting mixture and recycling of all the other phenolic components [230]. The isolation of pure *m*-cresol (or *p*-cresol) is possible in the same way if the *m*/p rectification cut is additionally passed through a Y zeolite adsorption unit and the unwanted isomer is recycled with the other phenols [231].

Owing to the temporary lack of demand for *o*-cresol, opportunities for isomerizing it to *m*- and *p*-cresol have been investigated. The results formerly obtained with aluminum chloride [235], [236], aluminum fluoride, boron trifluoride, aluminum silicates, or aluminum oxide were unsatisfactory [236], [237] because resinification and disproportionation to phenol, dimethylphenols, etc. occurred in addition to isomerization.

Isomerization without substantial disproportionation is possible, however, on silica – alumina zeolites of the ZSM group [238–240]; at a temperature of ca. 400 °C and *o*-cresol conversions of 40 – 50 %, *m*-/*p*-cresol selectivities of 90 % and an *m* : *p* ratio of 7 : 3 are possible [238].

1.4.5. Other Routes

The rising demand for cresol has stimulated the development of numerous syntheses, especially those that lead selectively to certain cresol isomers. It appears, however, that these processes are not yet being used industrially.

Gulf Oxychlorination. A cresol synthesis analogous to the Raschig phenol synthesis cannot be carried out satisfactorily under industrial conditions: at the temperature needed for oxychlorination in the vapor phase (Eq. 5), toluene is oxidized to a considerably greater extent than benzene, and only low rates of toluene conversion are therefore possible.

$$CH_3-C_6H_5+HCl+0.5\,O_2 \longrightarrow CH_3-C_6H_4-Cl+H_2O \quad (5)$$

$$CH_3-C_6H_4-Cl+H_2O \longrightarrow CH_3-C_6H_4-OH+HCl \quad (6)$$

In a process developed by the Gulf Research Development Co., these disadvantages are avoided by oxychlorinating the toluene with aqueous hydrochloric acid and oxygen at a temperature of ca. 100 °C in the presence of catalytic amounts of nitric acid and a palladium or copper salt [241], [242]. At toluene conversion rates of 80 %, primarily *o*- and *p*-chlorotoluene are obtained in a 2 : 1 molar ratio and with a selectivity of $\geq$ 95 %.

The subsequent hydrolysis of the chlorotoluene mixture is carried out in the vapor phase at 400 – 450 °C, as in the Raschig process, but over improved catalysts, such as lanthanum phosphate that may be activated with copper [243]. At a chlorotoluene conversion of 20 %, the cresol selectivity is ca. 95 %. The isomer ratio differs from that given by the processes already being used industrially. A large amount of *o*-cresol is obtained in addition to *p*-cresol containing some *m*-cresol.

If the hydrochloric acid liberated in the hydrolysis step (Eq. 6) is returned to the oxychlorination reactor, the process becomes equivalent to the oxidation of toluene with oxygen. This fact, together with the high conversion rates and selectivities, makes the process interesting. Commercial exploitation of the process necessitates the extensive recycling and places severe demands on the resistance of the plant to corrosion.

The Gulf oxychlorination process is equally applicable to xylenes, and is therefore also suitable for the manufacture of xylenols.

Oxidative Decarboxylation of Methylbenzoic Acids. As in the Dow process for the manufacture of phenol from benzoic acid, methylbenzoic acids are decarboxylated to cresol when an air – steam mixture is passed through their melt at 200 – 240 °C in the presence of copper and magnesium salts [244].

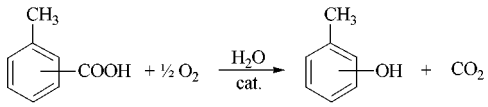

By comparison with the manufacture of phenol, the cresol process is more complex and 10 – 15 % less selective [60]. This is due partly to the fact that cresols have higher boiling points; in consequence, optimal control of the reaction is difficult, and a greater amount of tar is formed [244]. For a means of reducing the tar formation, see [245]. In addition, the simultaneous successive oxidation of the methyl group leads to numerous byproducts [246]. Oxidation in the presence of benzoic acid has been proposed [247] as a mean of preventing the sublimation of methylbenzoic acid.

Better results are achieved when a methylbenzoic acid – steam mixture is oxidized in the vapor phase with oxygen – nitrogen mixtures over mixed catalysts at a temperature of ca. 300 °C [248], [249]. Thus, the reaction of a mixture of *p*-methylbenzoic acid, water, oxygen, and nitrogen (molar ratio of 1 : 18 : 1 : 9) at 300 °C over a Mo – V – Cu – Zn – Na oxide catalyst supported on γ-Al_2O_3 is reported to give a *m*-cresol selectivity of 90 % at a methylbenzoic acid conversion rate of 80 %; as byproducts, 2 % toluene, 5 % bis-(methylphenyl) – bis(methylphenyl) ether mixture, and 3 % of additional carbon dioxide are formed [249].

The process could be of interest in the manufacture of pure *m*-cresol, since only *m*-cresol is formed from both the *ortho* and the *para* forms of methylbenzoic acid [60], [244]. *m*-Methylbenzoic acid gives (in the liquid phase)

o- and *p*-cresol in a 1 : 2 ratio [250], [251]. Regarding the mechanism and kinetics, see [244], [246], and [252], respectively.

The required methylbenzoic acids can be obtained by oxidation of the corresponding xylenes, which are, however, more expensive than toluene.

Baeyer – Villiger Oxidation of *p*- or *o*-Methylbenzaldehyde. This route leads to *o*-cresol or, more interestingly from the practical aspect, specifically to *p*-cresol. Hydrogen peroxide reacts with excess formic acid to give performic acid (Eq. 7), which oxidizes *p*-methylbenzaldehyde to *p*-tolyl formate (Eq. 8); hydrolysis (Eq. 9) then leads to *p*-cresol and formic acid.

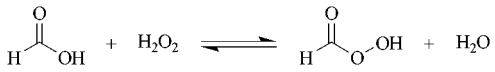

(7)

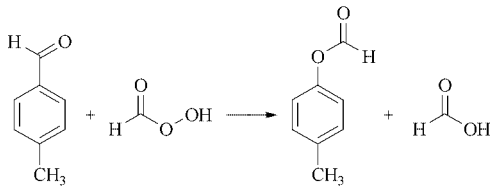

(8)

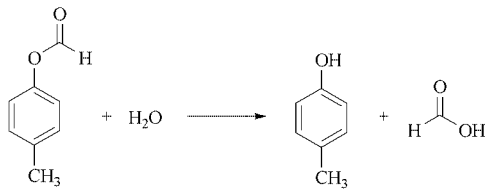

(9)

Several variants of the process have been developed [253–255]. The most favorable process was probably invented by Mitsubishi [256]. Hydrogen peroxide (1.2 mol; 90 %) is added to a mixture consisting of 1 mol of *p*-methylbenzaldehyde, 14 mol of formic acid, and ca. 10 % (4.5 mol) of water. The temperature of the mixture is maintained at 60 – 90 °C; the heat of the reaction is removed by reflux condensation under vacuum. Because of the presence of water, formate is hydrolyzed; after a reaction time of 1 h, the *p*-methylbenzaldehyde conversion rate is 100 % and the *p*-cresol yield is 85 %. If, instead of 90 % hydrogen peroxide, performic acid is used, or if the process is carried out in the presence of a solid acid (zeolite or ion-exchange resin), the yield rises to 90 – 92 %. The product is worked up by distilling the formic acid. Apart from the additional amount formed in the process, the formic acid, which has a concentration of ca. 90 %, is recycled. The *p*-cresol is then isolated from the residue, which still contains 10 – 20 % of high-boiling constituents [256].

The starting material, *p*-methylbenzaldehyde, can be produced from toluene and carbon monoxide by the so-called MGC PTAL process developed by Mitsubishi [257]. It is an intermediate in the manufacture of terephthalic acid.

Nucleus Hydroxylation. The direct hydroxylation of toluene (or xylene), e.g., with oxygen [258–260], in a plasma [261–263], in fuel cells [264], or in the presence of hydrogen [265], with hydrogen peroxide [266–269], inorganic peroxo compounds [270], [271], organic peroxides [272], [273], or iodosyl acetate [274], in the presence of various catalyst systems (Friedel – Crafts catalysts, transition-metal compounds, or irradiation), has been extensively investigated. In contrast to this, hydroxylation of toluene with specific microorganisms gives regioselectively *p*-cresol [275–277].

According to Rhône-Poulenc, the most successful results so far have been obtained with 85 % hydrogen peroxide in the presence of a large amount of trifluoromethanesulfonic acid and a small amount of phosphoric acid (for complexing metal ions) [278].

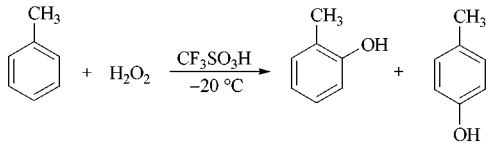

2 : 1

At – 20 to – 15 °C and a molar ratio of H_2O_2, toluene, CF_3SO_3H, and H_3PO_4 of ca. 1 : 10 : 17 : 0.04, *o*- and *p*-cresol are obtained in a 2 : 1 ratio in 80 % yield (calculated on the basis of hydrogen peroxide) at a toluene conversion rate of 10.6 %. One of the prob-

lems is the recycling of the very expensive trifluoromethanesulfonic acid.

Work carried out by UOP indicates that relatively good results are obtainable with hydrogen fluoride, which is less expensive and easier to handle industrially than trifluoromethanesulfonic acid, in the presence of carbon dioxide. With 30 % hydrogen peroxide at 0 °C and at a molar ratio of H_2O_2, toluene, HF, and CO_2 of 1 : 10 : 70 : 10, *o*- and *p*-cresol are obtained in a 2 : 1 ratio in 67 % yield, based on hydrogen peroxide, in addition to a small amount of *m*-cresol. *m*-Xylene gives mainly 2,4-xylenol, *p*-xylene mainly 2,5-xylenol, and *o*-xylene mainly 2,3- and 3,4-xylenol [279]. The effect of the carbon dioxide is not fully understood, but this compound acts to some extent as an internal coolant.

On titanium-modified zeolites, *o*-, *m*-, and *p*-cresols in a 14 : 11 : 75 ratio are formed in a yield of at least 40 %, based on H_2O_2, when a mixture of H_2O_2, water, toluene, methyl isobutyl ketone, and titanium zeolite (pretreated with H_3PO_4) is heated under reflux in a mass ratio of 1 : 7 : 17 : 10 : 2.6 [280]. For results on HY zeolites and titanium silicalite, see [281] and [282].

Cresols can also be obtained by oxidation of toluene with N_2O in the presence of H-ZSM-5 zeolites containing Lewis acidic extra framework Fe, Al [283], or Ga [284]. Thus, toluene with N_2O (molar ratio 1 : 3) at 350 °C/ 1 bar/ WHSV 1 h^{-1} on a H-[Al]ZSM-5 zeolite was oxidized to cresol with 24 % conversion and a selectivity of ca. 27 % (molar ratio *o*- : *m*- : *p*-cresol = 52 : 32 : 16). Besides some higher molecular aromatics the product mixture consisted of benzene, xylene, and phenol [283].

ICI has developed a cyclic process equivalent to the oxidation of toluene with peracetic acid, which gives yields of up to 95 % of an *o*-/*p*-cresol mixture, free of the meta isomer, with a cresol content of 92 %. The "catalyst" in the reaction is iodotoluene. This is oxidized by peracetic acid to iodosyltoluene, which without isolation reacts with toluene and sulfuric acid to form ditolyliodonium hydrogen sulfate. Hydrolysis is effected by diluting the reaction mixture. Cresol is produced; iodotoluene is simultaneously regenerated and recycled [285].

Until now all processes involving nucleus hydroxylation have been unsatisfactory for several reasons; for example, the cresols must be isolated from very dilute solutions, and considerable amounts of auxiliaries and unreacted starting materials must be recycled with as little waste as possible.

Oxidative Methylation of Toluene. At the All-Russian Research Institute for Organic Synthesis Co., Moscow, a pilot plant has been operating in which toluene is oxidized adiabatically in a tube reactor at 700 – 750 °C and with contact time of 0.2 – 0.5 s with oxygen in the presence of methane and steam (molar ratio 1 : 1 – 2 : 10 – 20 : 2 – 6) [286], [287].

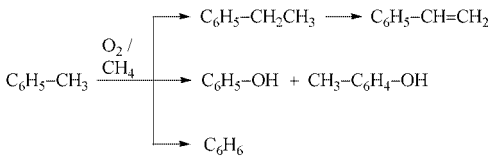

At a toluene conversion per pass of ca. 45 %, 0.52 t of ethylbenzene and styrene, 0.25 t of benzene, ca. 0.11 t of phenol, and ca. 0.06 t of cresol are obtained per tonne of toluene [288], [289]. Thus, if the process were to be used on the scale customary for petrochemical operations, it should be possible to produce considerable amounts of cresol by the "direct oxidation" of toluene.

Hock Reaction of Toluene Derivatives. As in the cymene – cresol process, other toluene hydrocarbons can be peroxidized and cleaved to give cresol and the corresponding carbonyl compound. Particular interest has been shown in toluene derivatives that give *p*-cresol and are more readily accessible than *p*-cymene, e.g., ethylidenebis(*p*-methylbenzene) (coproduct: *p*-methylacetophenone) [290], [291], 4-methylcyclohexylbenzene (coproduct: cyclohexanone) [292], and *p*-xylene (coproduct: formaldehyde) [293], [294]. Oxidation of xylene in the presence of acetic acid gives the corresponding cresyl acetate, which can be hydrolyzed to form cresol [295].

Hydrogenation of *N,N′*-Dialkylaminomethylphenols. The reaction of phenol with paraformaldehyde and, e.g., piperidine in methanol at a 2 : 1 : 1 : 100 molar ratio at ca.

10 °C gives a mixture of piperidinomethylphenols, which without isolation is directly hydrogenated at 120 – 130 °C and 1 MPa over a palladium – carbon catalyst:

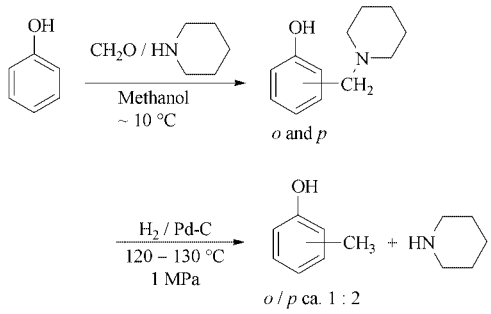

o-Cresol and *p*-cresol in a 1 : 2 molar ratio are obtained in a yield of ca. 90 %, based on converted phenol, or of about 75 %, based on the formaldehyde input. In addition, some 2,4-xylenol is formed (2 – 5 %) without *m*-cresol [296]. The yield is lower (65 – 70 %) if dimethylamine is substituted for piperidine.

Diels – Alder Ring Closure of Isoprene and Vinyl Acetate. The Diels – Alder route leads selectively to *p*-cresol in three reaction steps: (1) Isoprene (1 mol) reacts with vinyl acetate (20 mol) in the presence of 1 % hydroquinone in an autoclave at 180 °C; (2) the reaction product, 1-methylcyclohexen-4-yl acetate, is saponified; and (3) finally, the 1-methylcyclohexen-4-ol is dehydrogenated catalytically to *p*-cresol [297]:

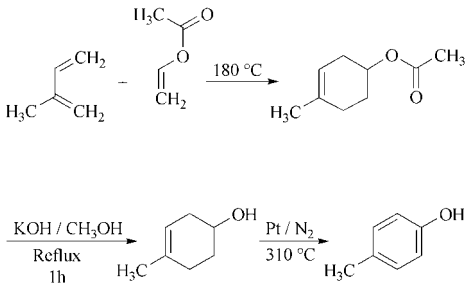

High yields are claimed to be possible when the starting materials are recycled.

Carbonylation. 2-Methallyl chloride, acetylene, and carbon monoxide in acetone as solvent give *m*-cresol in 74 % yield at room temperature.

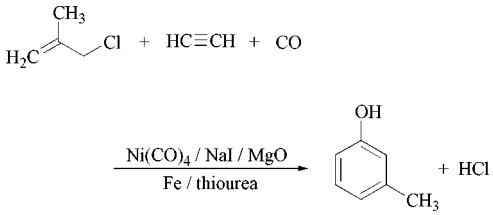

The catalyst consists of tetracarbonylnickel, sodium iodide, iron powder, and thiourea; magnesium oxide neutralizes the hydrochloric acid formed [298].

From Toluidines. Under severe conditions, toluidines in the presence of acidic catalysts (H_3PO_4, HCl, $ZnCl_2$, BF_4) can be hydrolyzed to form the corresponding cresols [299], [300]. Higher yields can be achieved by boiling the diazonium salts obtainable from toluidines [301–303]. Continuous operation of the process is described in [304].

Hydrodealkylation of Xylenols. Xylenols can be demethylated to phenol and cresols with hydrogen under pressure, either thermally at 600 – 700 °C [25] or on metal oxide catalysts at 400 – 500 °C [305]. Simultaneous deoxygenation enables BTX aromatics (BTX = benzene – toluene – xylene) to be produced at the same time. In this way, additional cresols can be produced, if necessary, from the xylenol fractions that are formed on a large scale in the gasification and hydrogenation of coal and other processes.

1.5. Separation of *m*- and *p*-Cresol [30]

Only *o*-cresol can be separated as a pure product by distillation from mixtures of the three cresol isomers. *m*-Cresol and *p*-cresol are obtained as a single fraction because the difference between their boiling points is too small for separation by distillation, even when carried out under vacuum.

1.5.1. Physicochemical Processes

Crystallization. According to the melting point diagram for *m*-/*p*-cresol mixtures (Fig. 1), it is possible to obtain, at normal pressure, pure *p*-cresol from mixtures containing more than

58 % *p*-cresol, and pure *m*-cresol from mixtures containing more than 89 % *m*-cresol. *Meta/para* mixtures containing 42 – 89 % of *m*-cresol cannot be separated by crystallization at normal pressure because *m*- and *p*-cresol exist as a molecular compound in a 2 : 1 molar ratio within this concentration range.

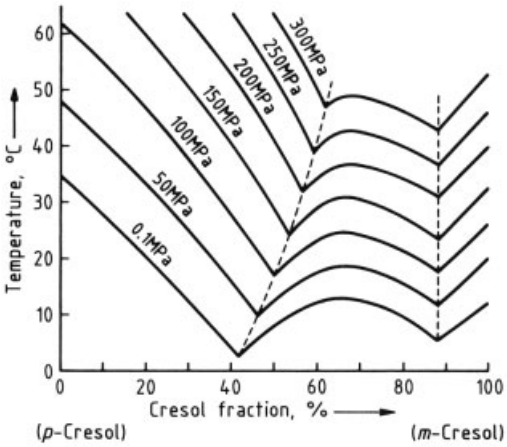

Figure 1. Liquid–solid phase diagram of *m*- and *p*-cresol mixtures under pressure [306]

The range within which *p*-cresol can be isolated by crystallization can be extended by applying pressure (Fig. 1); simultaneously the crystallization temperature and also the *p*-cresol yields, relative to crystallization at normal pressure, are raised [306], [307].

Crystallization is of practical importance as a technique for the separation of pure *p*-cresol from *m*-/*p*-cresol mixtures of high *p*-cresol content obtained from syntheses or other processes. Usually, this method has generally been carried out at normal pressure [308]. In Japan, a high-pressure crystallization technique known as the *Finecry process,* in which a liquid *m*-/*p*-cresol mixture is introduced into a high-pressure vessel of the piston cylinder type and crystallized adiabatically at 200 MPa, has been developed for industrial use by Kobe Steel. After the mother liquor has been drawn off and the plant decompressed, the *p*-cresol emerges as a pure crystalline product. The computer-controlled crystallization – decompression cycle lasts 2 – 5 min [309]. A plant for the separation of 500 t/a *p*-cresol with a purity of ≥ 99.5 % from its ≥ 70 % *p/m* mixtures has been operated since 1986 by the Sumitomo Chemical Co. in Japan [310].

Adsorption. Alkali metal-modified or alkaline earth metal-modified zeolites of type X, A [311], [312], L, or ZMS-5 [313], and also titanium dioxide [314], adsorb *p*-cresol more strongly than *m*-cresol. Thus, one can separate *m*-/*p*-cresol mixtures in an adsorption column and dissolve them again with a suitable desorbing liquid such as an aliphatic alcohol [311], e.g., a mixture of 1-hexanol and 2-heptanol [312], or a ketone [315]. A continuous version of this process has been developed by UOP [113]. It is known as the *Cresex process* and is used industrially in the United States by Merichem. The separating efficiency of the process not only depends on the adsorbent, but is also influenced to a large extent by the desorbent. With 1-hexanol, for example, only one isomer, either *m*- or *p*-cresol, can be isolated at a purity of ca. 99 % in a single operation (1 column); the purity of the other isomer obtained is only ca. 91 %. With 1-pentanol, however, a mixture containing ca. 69 % *m*-cresol and 31 % *p*-cresol can be separated into an extract containing 98.9 % of *p*-cresol and a raffinate containing 99.2 % of *m*-cresol in a single operation [113]. Phenolic impurities present in the cresol feed, e.g., *o*-cresol, *o*-ethylphenol, 2,4-, 2,5-, and 2,6-xylenol, migrate into the *m*-cresol during the adsorptive separation; hence, only pure *p*-cresol can be obtained in one operation if these impurities are present in significant quantities [113]. From mixtures containing additive neutral oils, *o*-cresol, ethylphenol and 2,6-xylenol, *p*- and *m*-cresol can be coextracted in an adsorptive pretreatment step [316].

Other Methods. The separation of *m*-/*p*-cresol mixtures by azeotropic rectification with benzyl alcohol [317] or by countercurrent extraction with selective solvents (aqueous methanol – gasoline) [318] has acquired no practical importance. This also applies to dissociative extraction and rectification processes that exploit the higher dissociation constant (higher acidity) of *m*-cresol in relation to that of *p*-cresol. Dissociative extraction uses systems consisting of aqueous sodium hydroxide solution [319], [320], aqueous trisodium phosphate solution [321], or aqueous aminoethanol solution [319] and an organic solvent, e.g., benzene, hexane, or toluene, by which the *p*-cresol is isolated in a multistep extraction. In dissociative rectifi-

cation, an amount of sodium hydroxide solution equivalent to the *m*-cresol content is added to the mixture and the *p*-cresol is distilled through a column with a large amount of water vapor as an azeotrope [322]. The separation of *m*/*p*-cresol can also be achieved by extractive rectification in the presence of diethanolamine or aniline as reactive entrainers [323].

1.5.2. Separation via Addition Compounds

m-Cresol and *p*-cresol form addition compounds with anhydrous calcium bromide; the compound formed with *p*-cresol has the higher thermodynamic stability. If an *m*-/*p*-cresol mixture in toluene is heated to 100 °C with 1 mol of calcium bromide per mole of *p*-cresol, the precipitate that is filtered off after ca. 8 h consists practically only of the adduct of *p*-cresol, while the *m*-cresol content of the mother liquor is correspondingly higher. If the *p*-cresol adduct is then decomposed, e.g., by heating to 200 °C, *p*-cresol with a purity of ca. 98 % is obtained, together with reusable calcium bromide [324]. Pure *m*-cresol can be isolated from the mother liquor rich in *m* isomer by selective adduct formation, e.g., with sodium acetate [325] or urea [326]. Pure *m*-cresol can also be obtained from the mother liquor after cooling to ca. 0 to − 5 °C and adding calcium bromide. Under these kinetically controlled conditions, the main product is the *m*-cresol – calcium bromide adduct, the subsequent decomposition of which gives *m*-cresol [327]. In adduct formation with excess calcium bromide at 25 – 30 °C, the *m*-cresol content of the mother liquor is directly rendered practically free of *p* isomer. The *p*-cresol from the adduct then contains correspondingly more *m*-cresol [328]. It is not known whether this process, which was developed by the Koppers Co. and has many variants, is being used industrially. Lithium bromide, magnesium bromide, or magnesium chloride can be used in place of calcium bromide, but the results are less satisfactory [324–328].

Many organic nitrogen compounds react with *m*- or *p*-cresol to form crystalline addition compounds that can be separated from the reaction mixture and decomposed to regenerate the nitrogen compound and the respective cresol isomer [30]. Benzidine was formerly considered particularly suitable for the separation of *m*- and *p*-cresol by virtue of the position of, and difference between, the melting points of its adducts; however, not only the carcinogenic properties of this compound but also other considerations now preclude its use. A more suitable and practical process is one in which benzylamine as the auxiliary substance forms with *m*-cresol a 1 : 1 adduct, which has a melting point of 39.5 °C. The adduct is centrifuged and fractionated in two distillation columns operating at different pressures (101.3 and 11.3 kPa) in such a manner that the benzylamine is recovered at the top of column 1 while pure *m*-cresol is obtained at the top of column 2. The mother liquor is distilled in the same way, then pure *p*-cresol is isolated by crystallization from the *m*-/*p*-cresol mixture rich in *p* isomer obtained in this case at the top of column 2. The benzylamine is recycled [329]. An auxilary substance for separation of *p*-cresol as an 1 : 1 addition compound from *m*/*p*-cresol mixtures by melt crystallization in the range of 0 – 50 °C is 2-pyrrolidinone, which thereafter is recovered from the *p*-cresol and the mother liquor by distillation and then recycled [330]. Strategies for the separation of *m*/*p*-cresol by adduct cristallization with ethylendiamine are discussed in [331].

The isolation of *m*-cresol via the 1 : 1 addition compound with urea is a process of some industrial importance. Formerly, the separated adduct was decomposed by heating with water. Now it is decomposed in the organic solvent in which the adduct has been formed. The procedure is as follows:

Urea (1.25 mol per mole of *m*-cresol) is added to a mixture of *m*-/*p*-cresol (65 : 35; 2 parts by weight) and toluene (0.5 parts by weight) at 30 °C. The mixture is cooled to − 10 °C, centrifuged, and washed with toluene (0.2 parts by weight). The product, consisting mainly of the urea adduct, is introduced into toluene (1.5 parts by weight), heated at 80 °C for 10 min, and filtered hot; almost all the urea is thereby held back. Distillation of the filtrate gives toluene, while that of the residue gives ≥ 98 % *m*-cresol in 89 % yield [332].

Advantage may be taken of adduct formation between cresols and certain other phenols with a view to effecting separation. For example, *m*-cresol forms with phenol an adduct with a melting point of 25.9 °C [333], while *p*-cresol forms an adduct with bisphenol A [isopropylidenedi(*p*-

phenol)] [334]. Furthermore, *m*-cresol gives with 1,4-dioxane at 20 °C and 110 – 130 MPa an addition compound that has been proposed for the separation of *m*-cresol from *m*-/*p*-cresol mixtures in which the content of *m*-cresol is high [335]. With α-cyclodextrin under aqueous conditions *p*-cresol preferentially precipitates as an inclusion complex (clathrate), from which *p*-cresol can be isolated in an enriched form by dissolution and extraction with water and ether [336].

m-Cresol can be precipitated as an addition compound with anhydrous sodium acetate from *m*-/*p*-cresol mixtures (e.g., in the ratio 65 : 35) in an organic solvent (gasoline, benzene, or toluene) at 20 – 40 °C. The precipitate is removed from the mother liquor and decomposed in the same solvent by heating to 80 – 95 °C; the sodium acetate is filtered. *m*-Cresol is obtained from the organic solvent as a technically pure product (96 – 99 %). The yield is ca. 70 % [30]. Higher yields (75 %) and improved *m*-cresol purity (99.5 %) are said to be possible if the adduct is decomposed at room temperature with polar solvents such as acetone [337].

1.5.3. Separation via Ester or Salt Formation

The mother liquor enriched with *p*-cresol (e.g., to a *m/p* ratio of 35 : 65) from the precipitation with sodium acetate can react with oxalic acid – with azeotropic dewatering – to give the semiester or diester, which is precipitated on cooling to 20 °C. The ester is separated and then hydrolyzed by heating with water at 85 % in the presence of the respective nonpolar solvent. The oxalic acid may be precipitated on cooling and recycled. The *p*-cresol dissolves in the organic solvent, from which it is isolated by distillation. The alternating use of both processes separates *m*- and *p*-cresol almost entirely [338].

Separation of the cresols via the calcium cresolates, in which advantage is taken either of the relatively greater solubility of the calcium *p*-cresolate or of the easier hydrolyzability of the *m*-cresolate with superheated steam, is unsatisfactory and has no practical importance.

1.5.4. Separation by Nuclear Substitution

Partial separation of cresols is possible by taking advantage either of the lower solubility of *p*-cresolsulfonic acid in sulfuric acid [30], of the more rapid formation of *m*-cresolsulfonic acid during sulfonation [339], or of the fact that the rate of hydrolysis of *m*-cresolsulfonic acid is faster than that of *p*-cresolsulfonic acid [340]. The last-mentioned difference forms the basis of the Bruckner process, the continuous operation of which in a reaction tower is described in [341].

A large-scale industrial process, e.g., used by Bayer in Germany and by Sumitomo and Honshu in Japan, comprises the separation of *m*- and *p*-cresol via the di-*tert*-butyl derivatives. In this technique, the *m*-/*p*-cresol mixture reacts with up to 2 mol of isobutene per mole of cresol in the presence of acidic catalysts (sulfuric acid). The resulting alkylated product consists mainly of 2,6-di-*tert*-butyl-*p*-cresol (*bp* 147 °C at 2.67 kPa) and 2,4-di-*tert*-butyl-*m*-cresol (*bp* 167 °C at 2.67 kPa), which are easily separated by distillation, provided that the catalyst has first been entirely inactivated. The process also yields monobutylated cresols with boiling points of 124 – 129 °C at 2.67 kPa which are returned to the alkylation reactor, and dimers, trimers, and possibly tetramers (*bp* 125 °C at 2.67 kPa) of isobutene, which are removed from the system. *m*-Cresol or *p*-cresol and isobutene are subsequently recovered by dealkylation of the di-*tert*-butylcresols at ca. 200 °C in the presence of such acidic catalysts as sulfuric acid. The individual cresols are then further purified by distillation. Their purity may exceed 99 % [342], [343]. Continuous operation of the butylation – debutylation process is described in [342] and [344–346]. The process is attractive not simply because it gives a very high yield of pure cresol, but also because it enables 2,6-di-*tert*-butyl-*p*-cresol (BHT), a widely used antioxidant, to be produced simultaneously.

If under suitable conditions *m*-/*p*-cresol mixtures are chlorinated with sulfuryl chloride [347], nitrosated with sodium nitrite [348], or condensed with (para)formaldehyde [349], [350], *m*-cresol reacts considerably faster, and *p*-cresol with a purity of 96 – 98 % can be isolated from the reaction mixture as a byproduct. These processes are exploited in practice.

1.6. Quality Specifications and Analysis

Cresols and their mixtures are commercially available in a wide range of grades and purities. They are characterized according to the following criteria [351], [352]:

- *Content of cresols, xylenols, and other phenols,* determined preferably by GC, e.g., according to ASTM 3626-80
- *o-Cresol content* (mixtures), determined via the crystallizing point of the cineole adduct according to [351], PC 17-67; or *m-cresol content* (mixtures), determined via the crystallizing point of the urea adduct according to [351], PC 16-67
- *Crystallizing point*, directly or after drying (pure isomers), e.g., according to [351], PC 10-67
- *Distillation range* (mixtures), e.g., according to [351], PC 5–67, or DIN 52 137 and 52 139
- *Residue on evaporation,* e.g., according to [351], PC 7-67
- *Residue on distillation,* e.g., according to [351], PC 6-67
- *Specific gravity,* e.g., according to [351], PC 2-67
- *Color* (color number), e.g., according to [351], PC 1-67, or ASTM D 3627
- *Water content,* determined either by azeotropic distillation with xylene according to Dean and Stark ([351], PC 3-67) or by the Karl Fischer method (ASTM D 1631-80)
- *Content of neutral oils,* e.g., as determined according to [351], PC 4-67, by distillation after addition of sodium hydroxide solution (cresolate formation)
- *Content of pyridine bases,* determined either according to [351], PC 4-67, by titrating the distillate from the neutral oil determination with hydrochloric acid, or by direct titration with perchloric acid in glacial acetic acid [352]
- *Total sulfur content,* e.g., according to [351], PC 22-67 (reduction method)
- *Acidity or alkalinity,* e.g., according to [351], PC 9-67, by titration against phenol red
- *Solubility in sodium hydroxide solution,* e.g., according to [351], PC 16-67

Specifications laid down in British Standards are as follows: For *o*-cresol (crystallizing point, *fp,* grade A 30.3 – 31.0 °C; grade B 29.3 – 30.3 °C; grade C 28.3 – 29.3 °C), *m*-cresol (*fp* $\geq$ 10.5 °C), and *p*-cresol (*fp* $\geq$ 34.0 °C) in B.S. 522 : 1964; for cresylic acid of specified *o*-cresol content (45 to 90 %) in B.S. 517 : 1964; for cresylic acid of specified *m*-cresol content (60 to 39 %, grades A to E) in B.S. 521 : 1964; and for so-called refined cresylic acid (grades A to E with altogether 0.5, 1.0, 2.0, 3.0, or 5.0 % water, pyridine bases, and neutral oil content) in B.S. 524 : 1964.

Maximal amounts of impurities encountered in commercially available cresols are usually no more than 0.2 % water, 0.15 % pyridine bases, 0.1 % neutral oils, and 0.01 % sulfur. Synthetic cresols are practically free from pyridine bases and sulfur. Pure cresols are colorless but may have a faintly yellowish color. Their mixtures sometimes have a slightly stronger tinge, ranging from yellow to brown.

o-Cresol is produced with a purity of 99.2 – 99.9 % (*fp* $\geq$ 30.7 °C), e.g., for ECN resins (see Section 1.8). Commonly used technical grades have purity levels of 98.2 – 99 % (*fp* $\approx$ 30.3 –30.6 °C), ca. 96 % (*fp* $\approx$ 29.5 °C), and ca. 94 % (*fp* $\approx$ 28.5 °C) and thus correspond to grades A, B, and C of the British Standard. For example, *m*-cresol, is sold at purities of 99 % (*fp* $\approx$ 11.5 °C), 98 % (*fp* $\approx$ 11 °C), and 96.5 % (*fp* $\approx$ 10 °C). *p*-Cresol usually has a purity of 99 % (*fp* $\approx$ 34 °C). Grades with a purity of ca. 95 % (*fp* $\approx$ 31 °C) and ca. 65 % are also available. The main impurity is *m*-cresol.

Technical cresol mixtures are usually classified according to their *m*-cresol content, which is between 20 and 70 %. Within this range, many different grades are supplied, depending on the manufacturer and the customer's requirements; there may also be substantial differences between them on the basis of phenols (phenol itself, *o*-cresol, *p*-cresol, xylenols, 2-ethylphenol, and other alkylphenols) that are additionally present, as well as other impurities (water, pyridine, neutral oils, and sulfur compounds). Thus, there are cresol mixtures containing up to 50 % phenol, or almost 50 % xylenols and other alkylphenols, or which have a total cresol content of only ca. 30 %; normally the latter lies between 60 and 90 %. *m*-/*p*-Cresol mixtures obtained by chlorotoluene hydrolysis or from the cymene – cresol process are distinguished by their high *m*-cresol content (70 % and 60 %, respectively)

and by their high total content of cresol, normally $\geq 99\,\%$.

1.7. Handling, Storage, and Transportation

The cresols are corrosive, and toxic if absorbed through the skin (see Chap. 4). Work with cresols should be carried out in closed apparatus wherever possible. The rooms must be well ventilated, and precautions, such as the use of airlocks and the gas compensation technique, must ensure that cresol vapor or dust cannot be inhaled by personnel. The fundamental criterion of the effectiveness of all such precautions is the MAK value or TLV, which in 1998 was fixed at 5 ppm (22 mg/m^3) for all cresol isomers and isomer mixtures. This level corresponds to the odor threshold. At higher concentrations, e.g., resulting from operational disturbances, suitable respirators must be worn. In addition, persons handling cresol must always have adequate eye protection and, if necessary, body protection, e.g., as provided by polychloroprene or PVC [poly(vinyl chloride)] gloves, long aprons, fully protective suits, etc. Further recommendations on the handling of cresols and xylenols may be found in [353–356] and in the safety data sheets of cresol manufacturers.

Steel vessels are suitable for storage. The discoloration that takes place on storage is promoted by elevated temperatures, light, oxygen, water, and catalysts, e.g., by dissolved iron or copper. If the product is required to remain colorless for as long as possible, it is best to use tanks of stainless steel (V4A; AISI 316Ti). If exceptional color stability is required, the product should also be stored under nitrogen. Access of moisture must be prevented, since cresols absorb water readily. Both *o*-cresol and *p*-cresol, which are solids at room temperature, have to be handled at a temperature higher than their crystallizing points. Cresol – air mixtures reach the lower explosive limit at ca. 76 °C in the case of *o*-cresol and at ca. 83 °C in those of *m*- and *p*-cresol; therefore, if the storage temperature exceeds 70 °C or 75 °C, respectively, the gas space must be rendered inert with nitrogen. Storage practice must comply with local regulations. The storage containers must be tested regularly for soundness and freedom from leaks. As the cresols are classified as water pollutants, measures such as the provision of tank basins must be taken to prevent cresol from entering the soil or drains.

Cresols are transported in heatable tank cars, in tank trucks, or in steel drums with a capacity of ca. 200 L. The container must bear the appropriate hazard warnings in compliance with the regulations. For unloading and sampling, see [355].

Transportation is subject to the following regulations:

International sea transport (IMDG Code) and international air transport (ICAO/IATA-DGR) for *o-, m-,* and *p*-cresol and cresol mixtures: Class 6.1, UN no. 2076, package group II; for cresylic acids: class 6.1, UN no. 2022, package group II. For transport of *o-, m-,* and *p*-cresol, cresol mixtures, and cresylic acids on European roads (ADR), railways (RID), and waterways (ADNR): Class 6.1, no 27b.

National regulations: United States DOT-CFR 49: § 172.101 class 6.1; for *o*-, *m*-, *p*-cresol and cresol mixtures: UN no. 2076; for cresylic acids: UN no. 2022. In Germany for the transport of *o*-, *m*-, and *p*-cresol, cresol mixtures, and cresylic acids on roads (GGVS), railways (GGVE), and waterways (GGVBinSch): Class 6.1, no 27b.

Cresols are liquids that are partly miscible with water and have flash points of 80 – 90 °C. As such, they are, therefore, classified in Germany as belonging to group A, hazard class III, according to the regulations for the storage of flammable liquids (Verordnung über die Lagerung brennbarer Flüssigkeiten, VbF). In the United States, *m*- and *p*-cresol belong to Category 1 of the NFPA 704M fire hazard classification system. Water fog, foam, dry chemical extinguishers, and – for small fires – carbon dioxide are suitable for fire fighting (never use a sharp jet of water). In Germany, VDE Specification 0171 places the cresols in Ignition Group T 1.

1.8. Uses

Most of the *o*-cresol manufactured in Europe is chlorinated to 4-chloro-*o*-cresol (PCOC), the starting material for the chlorophenoxyalkanoic acids 4-chloro-2-methylphenoxyacetic acid

(MCPA), 2-(4-chloro-2-methylphenoxy)-propionic acid (MCPP), and γ-(4-chloro-2-methylphenoxy)butyric acid (MCPB), which are important as selective herbicides. A considerably smaller proportion is nitrated to 4,6-dinitro-*o*-cresol (DNCO), which has both herbicidal and insecticidal properties, and is also used as a polymerization inhibitor for the production and distillation of styrene.

Highly pure *o*-cresol is increasingly processed, especially in Japan, to epoxy – *o*-cresol novolak resins (ECN resins), which are used as sealing materials for integrated circuits (chips). *o*-Cresol (of common quality) is also used to modify traditional phenol–formaldehyde resins.

o-Cresol is also important as a precursor of various dye intermediates, of which the most important in terms of quantity is *o*-cresotinic acid (*o*-hydroxymethylbenzoic acid) produced by the Kolbe synthesis. This acid finds further use in the manufacture of pharmaceuticals, whereas its methyl esters serve as dyeing auxiliaries.

An appreciable amount of *o*-cresol is used as a solvent, either directly or after hydrogenation to 2-methylcyclohexanol or 2-methylcyclohexanone. In the form of its carbonate ester, *o*-cresol constitutes a starting material in the synthesis of coumarin. The alkylation of *o*-cresol with propene gives carvacrol (3-isopropyl-6-methylphenol), which is used as an antiseptic and in fragrances. In addition, small amounts of *o*-cresol are, e.g., alkylated with isobutene and used as starting materials in the production of various antioxidants, for the production of components for thermal recording materials, and for pharmaceuticals (e.g., the muscle relaxant Mephenesin).

m-Cresol, either pure or mixed with *p*-cresol, forms a starting material for such important contact insecticides as *O,O*-dimethyl-*O*-(3-methyl-4-nitrophenyl)thionophosphoric acid (fenitrothion; Folithion, Sumithion) and *O,O*-dimethyl-*O*-(3-methyl-4-methylthiophenyl)thionophosphoric acid ester (fenthion; Baytex, Lebaycid). In addition, *m*-cresol is needed in the synthesis of phenyl *m*-tolyl ether (*m*-phenoxytoluene), which after oxidation to *m*-phenoxybenzaldehyde serves as a building block in the production of insecticides of the pyrethroid type.

Selective methylation of pure *m*-cresol with methanol gives 2,3,6-trimethyphenol, which is an important starting material for vitamin E synthesis; it also serves as a comonomer for the modification of poly(phenylene oxide) resins. Pure *m*-cresol also has considerable importance in the production of fragrance and flavor substances: its isopropylation gives thymol, from which (−)-menthol is obtained by hydrogenation and separation of the isomers. 6-*tert*-Butyl-*m*-cresol, obtained by alkylation of *m*-cresol with isobutene, is the starting material for the perfume fixative musk ambrette. The condensation of 6-*tert*-butyl-*m*-cresol with sulfur dichloride gives 4,4′-thiobis(6-*tert*-butyl-*m*-cresol), an important antioxidant for polyethylene and polypropylene. Further antioxidants are produced from 6-*tert*-butyl-*m*-cresol by condensation with butyraldehyde or crotonaldehyde. Pure *m*-cresol is also used for the production of positive photoresist compositions and for thermal recording materials. In Japan, *m*-toluidine is produced on demand by amination of *m*-cresol.

The 4-chloro-*m*-cresol obtained by selective chlorination of pure *m-cresol or m-/p*-cresol mixtures finds use as a disinfectant and preservative. 2,4,6-Trinitro-*m*-cresol can be used as an explosive.

p-Cresol, pure or mixed with *m*-cresol, is used mainly to produce 2,6-di-*tert*-butyl-*p*-cresol (BHT), a nonstaining and light-resistant antioxidant with a wide range of applications. 2-Alkyl-*p*-cresols obtained by the monoalkylation of *p*-cresol in the *ortho* position, e.g., *tert*-butyl-, cyclohexyl-, methylcyclohexyl-, α-methylbenzyl-, or nonyl-*p*-cresol, can be condensed with formaldehyde or sulfur dichloride to give 2,2′-methylene- or 2,2′-thiodiphenols, which are also important as antioxidants. 2,6-Dicyclopentyl-*p*-cresol, obtainable by dicyclopentylation of *p*-cresol, is a very effective antioxidant. Coupling of 2-*tert*-butyl-*p*-cresol with diazotized 4-chloro-2-nitroaniline yields Tinuvin 326, an absorber of UV light, which finds use, e.g., for polyethylene and polypropylene films and coatings. 2,6-Dinitro-*p*-cresol (DNCP) is used also (see above) as a polymerization inhibitor of styrene. In the fragrance industry, *p*-cresol is used to obtain *p*-cresolcarboxylic acid esters and, particularly, *p*-cresol methyl ether, most of which is used in the production of anisaldehyde, which is also important for the manufacture of 2-ethylhexyl-*p*-methoxycinnamate, a UV adsor-

ber in sunscreens, and for various pharmaceuticals, e.g., the coronary therapeutic Diltiazem and chemotherapeutics like trimethoprim. The direct oxidation of *p*-cresol gives *p*-hydroxybenzaldehyde, which is etherified to give anisaldehyde or is used for the production of the aroma chemical raspberry ketone and raspberry ketone acetate, which is used as a pheromone of the asian melon fly. Like *o*-cresol, *m*- and *p*-cresol are used as components of various dyes; the predominant intermediate here is 2-nitro-*p*-cresol.

An important field of application for technical cresol mixtures is the production of modified phenolic resins by condensation with formaldehyde. The suitability and price of cresol mixtures for this purpose depends on their content of *m*-cresol, the most reactive of the three isomers. Cresol mixtures are highly important as solvents for synthetic resin coatings (wire enamels).

m-/*p*-Cresol mixtures free of the *o* isomer are used to produce neutral phosphoric acid esters (tricresyl phosphate, diphenyl cresyl phosphate), which find use as fire-resistant hydraulic fluids, as additives in lubricants, as air filter oils, and as flame-retardant plasticizers for PVC and other plastics.

The bactericidal and fungicidal properties of cresols enable them to be used as disinfectants in soap. Synthetic tanning agents of commercial importance are obtained by the condensation of formaldehyde with cresolsulfonic acids and by the sulfonation of novolaks obtained from cresols. Crude cresols are used as wood preservatives. Cresol mixtures are also used in ore flotation and fiber treatment, as metal degreasing and cutting oils, as extracting solvents, and as agents for removing carbonization deposits from internal combustion engines. Methylcyclohexanol and methylcyclohexanone produced by hydrogenation of cresols are used in the paint and textile industries.

1.9. Economic Aspects

The combined total production of cresol in the United States, Europe, Japan, and South Africa in 1998 was estimated to be about 175 000 t; ca. 105 000 t of which was synthetic cresol, and ca. 70 000 t natural cresol.

About 42 000 t was produced in the United States, about 50 000 t in Europe, about 58 000 t in Japan, and about 25 000 t in South Africa. About 25 % of the total output is used by the manufacturers themselves. The so-called natural cresol comes from gasification of coal, from coking of coal and from spent refinery caustics in a ratio of about 50 : 35 : 15. Approximately 45 % of the total cresol output is processed in the form of pure *o*-cresol and pure *p*-cresol, and the rest as isomer mixtures. About 35 000 – 40 000 t of pure *m*/*p*-cresol mixture was separated – mostly by the butylation/debutylation-process – into pure *m*-cresol with coproduction of 2,6-di-*tert*-butyl-*p*-cresol (BHT).

The above figure for natural cresol from tars of coal gasification and coking and from spent refinery caustics refers only to the cresol content. It does not include the xylenols contained in the raw material (ca. 30 000 t), which to some extent are sold as though they were cresol.

The total output of cresol has diminshed in the last decade. Stagnating, partly declining markets, the entrance of 2,6-xylenol producers into the *o*-cresol market, the emergence of South Africa in the market for purified cresols, the declining availability of coal tars from coking processes and of spent refinery caustics, and the increasing environmental costs have, especially in the United States and in Europe, led to major industrial changes with restructuring, capacity reductions, and plant closures.

2. Xylenols [1–3]

Xylenol, dimethylphenol, $C_8H_{10}O$, M_r 122.166, occurs in six isomeric forms:

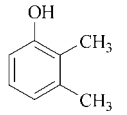

2,3-Xylenol
2,3-dimethylphenol
[*526-53-0*]

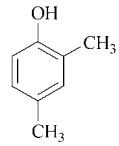

2,4-Xylenol
2,4-dimethylphenol
[*105-67-9*]

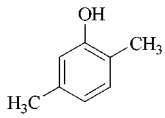

2,5-Xylenol
2,5-dimethylphenol
[*95-87-4*]

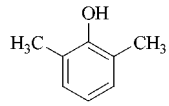

2,6-Xylenol
2,6-dimethylphenol
[*576-26-1*]

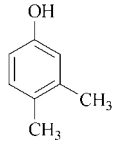

3,4-Xylenol
3,4-dimethylphenol
[*95-65-8*]

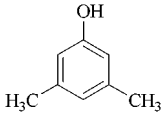

3,5-Xylenol
3,5-dimethylphenol
[*108-68-9*]

The numbering of the substituents normally and correctly begins from the OH group but is occasionally (erroneously) started from the methyl group: 2,4-xylenol can be referred to as 1,3,4-xylenol, 4-hydroxy-*m*-xylene, or 4-hydroxy-1,3-dimethylbenzene.

Xylenols are present, e.g., in the essential oils of various conifers, in tea, in tobacco and tobacco smoke, in roasted coffee, and in various smoked foods. In many cases, they contribute to the flavor of these products.

From the standpoint of economics, the most important isomer is now 2,6-xylenol, of which coal tar contains only a relatively small amount and which is therefore produced almost entirely by synthetic means.

2.1. Properties [1–3]

The xylenols are colorless crystallizing compounds that are soluble in alcohol, acetone, and in many other organic solvents. They are less soluble in water than the cresols. Except for 3,5-xylenol, their acidity is likewise lower than that of the cresols. They are soluble in aqueous sodium hydroxide, but to different extents, depending on the structure. Their chemical behavior is similar to that of the cresols. Physical data are presented in Table 4. For dependence of the thermodynamic data on temperature, see [9], [357–360]. Vapor – liquid equilibrium data of xylenols with numerous other substances are compiled in [20]. For the relationship between vapor pressure and association of xylenols, see [17].

2.2. Isolation

With regard to the formation of xylenols in various thermal cracking processes, see Section 1.3.

Xylenols, with the cresols and other phenols, are isolated from high-temperature, low-temperature, and Lurgi pressure gasifier coal tars (see Section 1.3.1) and from spent refinery caustics (see Section 1.3.2). The resulting crude phenols (cresylic acids) contain all six xylenols in amounts that differ greatly according to the raw material and conditions of formation (see Table 3). Moreover, the ratio of the xylenols themselves differs. In xylenols from spent refinery caustics, 2,4-xylenol generally predominates; besides this, the content of 2,5-xylenol is high, whereas that of 3,5-xylenol is relatively small [361]. Xylenols from bituminous coal tar are characterized by their high content of 3,5-xylenol. The average isomer ratio of bituminous coal tar xylenols is given in [362]; for example, as 31.9 % 3,5-xylenol, 28.1 % 2,4-xylenol, 14.1 % 2,3-xylenol, 13.6 % 2,5-xylenol, 6.9 % 2,6-xylenol, and 6.5 % 3,4-xylenol.

In the distillation of crude phenol, which is normally carried out continuously, the 2,6-xylenol present and some 2-ethylphenol enter the *m*-/*p*-cresol fraction. The other xylenols and the ethylphenols, as well as several C_9 phenols, are separated from the higher boiling phenols and the phenol pitch (→ Tar and Pitch) either together as a crude xylenol fraction or (in the case of xylenols from low-temperature tar) successively in the form of several differently enriched xylenol fractions.

2.3. Separation

For the isolation of individual xylenols, the crude xylenol fraction (see above) is rectified discontinuously into fractions of a narrow boiling temperature range [30].

2,4-Xylenol and 2,5-xylenol are contained in the fraction boiling at 208 – 212 °C at 101.3 kPa. The isomers are usually sold as a mixture. Since the difference between their boiling points (0.2 °C at 101.3 kPa) is too small for separation by distillation, other means of separation must be utilized.

If the fraction is cooled, the greater part of the 2,5-xylenol crystallizes and is then obtained in pure form after recrystallization. After sulfonation of the centrifuged mother liquor with 92 % sulfuric acid and dilution with water, the remaining 2,5-xylenol crystallizes as 2,5-xylenolsulfonic acid. Hot potassium chloride solution is then added to the filtrate, and

Table 4. Physical properties of pure xylenols

Physical property	2,3-Xylenol	2,4-Xylenol	2,5-Xylenol	2,6-Xylenol	3,4-Xylenol	3,5-Xylenol
Melting point (101.3 kPa), °C	72.57	24.54	74.85	45.62	65.11	63.27
Depression by 1 mol% of impurities, °C	0.502	0.576	0.727	0.558	0.727	0.520
Boiling point at 101.3 kPa, °C	216.87	210.93	211.13	201.03	226.95	221.69
at 10.0 kPa, °C	141.63	137.12	136.98	125.48	152.33	147.88
Density at 25 °C, g/cm^3	1.164	1.0160 (*l*)	1.189	1.132	1.138	1.115
at (t, °C), g/cm^3	0.981 (85)	1.0033 (40)	0.958 (85)	0.978 (65)	0.982 (80)	0.971 (75)
Refractive index at (t, °C)	1.5198 (80)	1.5254 (50)	1.5092 (80)	1.5171 (60)	1.5268 (60)	1.5150 (70)
Dielectric constant at 80 °C	4.65	4.98	5.07	4.34	7.72	7.05
Dissociation constant at 25 °C	2.88×10^{-11}	2.51×10^{-11}	3.89×10^{-11}	2.34×10^{-11}	4.34×10^{-11}	6.46×10^{-11}
Critical temperature, °C	449.7	434.4	433.9	427.8	456.7	442.4
Critical pressure, MPa	4.9	4.3	4.9	4.3	5.0	3.6
Critical density, kg/m^3	340	320	350	310	350	250
Heat capacity c_p at 25 °C (*g*), J mol^{-1} K^{-1}	164.2	156.2	157.2	156.1	163.6	152.8
c_p at 80 °C (*l*), J mol^{-1} K^{-1}	274	281	285	246	280	296
Enthalpy of fusion ΔH_m at *mp*, 101.3 kPa, kJ/mol	19.88	12.85	13.91	15.21	13.15	17.42
Enthalpy of sublimation ΔH_{sub} at 25 °C, 101.3 kPa, kJ/mol	84.07	65.90 (*l*)	85.04	75.65	85.79	82.90
Enthalpy of vaporization ΔH_v at *bp*, 101.3 kPa, kJ/mol	47.35	47.18	46.97	44.55	49.70	49.34
Enthalpy of combustion $\Delta H_c{}^\circ$ at 25 °C, 101.3 kPa, kJ/mol	− 4339	− 4351 (*l*)	− 4334	− 4343	− 4338	− 4336
Enthalpy of formation $\Delta H_f{}^\circ$ at 25 °C (*s* or *l*), 101.3 kPa, kJ/mol	− 241.5	− 229.0 (*l*)	− 246.9	− 237.6	− 242.5	− 244.6
$\Delta H_f{}^\circ$ at 25 °C (*g*), 101.3 kPa, kJ/mol	− 157.4	− 163.1	− 161.8	− 161.9	− 156.8	− 161.7
Free energy of formation $\Delta G_f{}^\circ$ at 25 °C (*g*), 101.3 kPa, kJ/mol	− 33.37	− 41.25	− 39.70	− 39.07	− 34.30	− 39.44
Entropy S° at 25 °C (*g*), 101.3 kPa, J mol^{-1} K^{-1}	390.4	398.0	396.0	390.0	391.4	392.0
Cubic expansion coefficient at 25 °C		0.818×10^{-3}				
Solubility in water at 25 °C, wt %	0.47	0.61	0.49	0.64	0.50	0.49
Critical solution temperature in water, °C	210	213	219	237	189	198
at % xylenol, wt %	38	34	42	45	35	35

the temperature is lowered. This precipitates the potassium salt of 2,4-xylenolsulfonic acid, from which, after recrystallization from water and acidification with sulfuric acid, the 2,4-xylenol is released by hydrolysis and stripped with steam.

It is also possible to separate 2,4- and 2,5-xylenol by butylation with isobutene. The process is used industrially in combination with the production of 6-*tert*-butyl-2,4-dimethylphenol. Due to the steric hindrance of the phenolic OH group, the resulting 6-*tert*-butyl-2,4-dimethylphenol cannot be dissolved in 10 % aqueous sodium hydroxide, whereas the 4-*tert*-butyl-2,5-dimethylphenol can. The unwanted of the products separated in this way can be debutylated with a catalytic amount of concentrated sulfuric acid. It has also been proposed that 2,5-xylenol be isolated by exploiting the fact that the two isomers have different butylation and debutylation rates on highly acidic ion-exchange resins [363], depending on temperature. 2,4-Xylenol and 2,5-xylenol can also be separated via adduct formation with phenols or amines [364].

Separation of 2,4-xylenol from its mixture with 2,5-xylenol (and *m*-ethylphenol) can also acchieved with an SrX zeolite as adsorbent and 1-hexanol as desorbent [365].

2,3-Xylenol is isolated from a fraction boiling at 215 – 218 °C by selective dephenolation with dilute aqueous sodium hydroxide, introduction of the free phenols into milk of lime, and stripping with steam; it can then be purified by recrystallization from benzene. It is also obtained in the cooling crystallization of a very narrow distillate fraction. Another method is the cooling crystallization of the distillation cut with *n*-

heptane at − 10 °C, which gives 3,5-xylenol in ≥ 99 % purity [366].

3,5-Xylenol is crystallized by cooling the fraction that boils at 218 – 221 °C, after which it is centrifuged and recrystallized from benzene.

3,4-Xylenol is obtained in the same way as given above from the fraction that boils at 223 – 225 °C; it is likewise centrifuged and recrystallized from benzene.

3,5-Xylenol or 3,4-xylenol can also be separated from other xylenols and from trimethylphenols via adduct formation with anhydrous calcium bromide [367].

Faujasite zeolites modified with copper or silver [368], potassium, barium, or lead [369] cations adsorb 3,5-xylenol less readily than other xylenols, thus enabling 3,5-xylenol to be separated as a raffinate component [368].

2,6-Xylenol has a boiling point ca. 0.9 °C lower than that of *p*-cresol at 101.3 kPa; the difference is ca. 4 °C at 20 kPa [9]. Thus, it can be obtained from the first runnings of the *m*-/*p*-cresol fraction. 2,6-Xylenol can be isolated in numerous ways. In the azeotropic distillation of the *m*-/*p*-cresol fraction with steam, 2,6-xylenol crystallizes from the condensate. 2,6-Xylenol can be distilled from cresolate caustics by virtue of its lower acidity, after which it is obtained from the condensate by extraction with benzene. For this reason, the condensate from the steam distillation of the crude phenolate caustic obtained from coal tar (see Section 1.3.1) is also suitable as a starting material for the isolation of 2,6-xylenol.

In fractions highly enriched with 2,6-xylenol, e.g., those obtained from synthesis by the methylation of phenol (Section 1.4.4), the 2,6-xylenol can be separated from *m*- and *p*-cresol and other phenols by countercurrent extraction with aqueous sodium hydroxide – hydrocarbon mixtures. This "dissociation extraction" technique takes advantage of the lower acidity of 2,6-xylenol. The product can be isolated from the hydrocarbon phase in a purity in excess of 98 % [370], [371]. Very high separation factors were obtained by dissociation extraction using solid anhydrous calcium hydroxide [372]. By comparison with *m*- and *p*-cresol, 2,6-xylenol has less tendency to form hydrogen bonds, and can therefore also be separated from the cresols either by countercurrent extraction with aqueous methanol–octane [373], by extractive distillation with di- or triethylene glycol [135], [181], [374], or alkanolamines [375]. Furthermore, the more polar phenols (cresols) can be distilled from the 2,6-xylenol as azeotropes with $C_{10} - C_{14}$ aliphatics as entrainers [376–378]. Cresols are adsorbed more readily than 2,6-xylenol on zeolites, e.g., of the K – Ba – X type; therefore, 2,6-xylenol can be isolated in high purity as a raffinate stream [379–381]. The isolation of 2,6-xylenol by the fractional crystallization of phenol methylation residues greatly enriched by distillation can be facilitated by adding phenol [382] or water [383], [384]. Good purification is achieved by crystallization under high pressure (e.g., 160 MPa at 75 °C) even in the absence of such additions [385]. Isolation of the 2,6-xylenol by distillation is possible if the *m*- and *p*-cresol impurities are selectively reacted with phosphorus oxide chloride to give phosphoric acid esters [386], with formaldehyde to give diphenylmethanes [387], with isobutene to give *o*-alkylated products [388], or with methanol on a catalyst either of Al_2O_3 [389] or of titanium sulfate–magnesium oxide [390] to give higher boiling xylenols and trimethylphenols.

2.4. Production

Phenol Methylation. Cresols or xylenols are formed in the methylation of phenol with methanol. Details of the catalysis and industrial conditions of the reaction are found in Section 1.4.4.

In the *gas phase, o*-cresol and 2,6-xylenol are the preferred products. The 2,6-xylenol can be made the main product if *ortho*-selective catalysts and high methanol : phenol ratios are used, and/or *o*-cresol is recycled [181]. As required under *ortho*-selective conditions, other products are chiefly 2,4-xylenol by methylation of *p*-cresol [391], and 2,3-xylenol [392] and 2,5-xylenol by methylation of *m*-cresol [393], [394].

For synthesis of 2,6-xylenol by selective *ortho*-methylation in the gas phase, phenol can be replaced by KA oil (a mixture of cyclohexanol and cyclohexanone from the oxidation of cyclohexane) if a magnesium-oxide-supported chromium oxide catalyst is used [395], [396].

With *liquid-phase* catalysts used for methylation (see page 16), it is typical not only that more 2,4-xylenol than 2,6-xylenol is formed, but also

that the purity of this 2,4-xylenol (ca. 90 %) is higher at the outset than that of the 2,4-xylenol obtained, e.g., from coal tars or spent refinery caustics (50 – 75 % purity).

Demethanization of Isophorone. 3,5-Xylenol, which is not normally a product of methylation processes and of which only a limited amount can be isolated in the pure form from natural xylenol mixtures, is produced in the United Kingdom by Inspec Fine Chemicals (formerly Synthetic Chemicals Ltd.) by demethanization of isophorone (3,5,5-trimethyl-2-cyclohexen-1-one) according to a process developed by Shell:

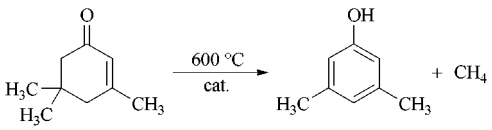

The reaction is carried out in the vapor phase in a multitubular reactor at 600 °C and 0.1 kPa and liquid hourly space velocities (LHSV) around 0.5 h^{-1}. This process, in which the isophorone conversion is practically complete, gives a 3,5-xylenol yield of ca. 70 % on a cobalt – molybdenum – potassium oxide–γ-aluminum oxide catalyst [397], or of ca. 80 % on a new rare-earth metal oxide–α-aluminum oxide carrier catalyst promoted with transition-metal oxides (0.1 % Co) and alkali metal oxides (0.4 % K) [398]. Toluene, xylene, mesitylene, dihydroisophorone, *m*-cresol, 2,4- and 2,5-xylenol, and 2,3,5-trimethylphenol are formed as byproducts. For selectivity improvement by alumina-supported K_2O/Cr_2O_3 or K_2O/V_2O_5 catalysts see [399], [400]. The catalyst has to be regenerated at intervals by burning off accumulated deposits of carbon. The carbon deposits are avoided in the process used industrially by Rütgers-VfT AG in Germany by carrying out the reaction at 520 – 540 °C and 10 – 15 bar in tubes of chromium–nickel steel with a packing of the same material. The yield is stated to be 65 % with an isophorone conversion of ca. 90 % [401]. Increased catalyst performance and throughput and diminished isophorone content of the product (< 3 %) is possible by using reaction zones containing catalyst with graduated reactivity and thermal stabilization (intercooling) of the process stream between the reaction zones [402]. At higher isophorone content the phenols from the condensed reaction product are separated from unconverted isophorone and other nonphenolic byproducts by extraction with sodium hydroxide solution and liberated with carbon dioxide [403]. The crude phenols are then rectified to give pure 3,5-xylenol. For 3,5-xylenol purification from mixtures with isophorone by addition of a tertiary amine (*bp* > 230 °C) followed by vacuum distillation to give azeotrope compositions from which 3,5-xylenol crystallizes, see [404].

Homogeneous catalysis is possible at 570 – 600 °C if ca. 1 wt % of methyl iodide (or another halogenated hydrocarbon) is added to the isophorone. Complete conversion is achieved at a liquid hourly space velocity (LHSV) of 0.5 h^{-1}. After the 3,5-xylenol has been worked up by distillation, it is obtained as a product of ca. 99 % purity in 85 % yield [405]. The iodine content of the product has to be lowered by heating with ca. 0.2 wt % of metallic zinc, magnesium, or iron before the distillation [406].

Isophorone is produced by alkaline trimerization of acetone (→ Ketones). The acetone can also be converted directly on a catalyst of either magnesium oxide [407] or chromium oxide – magnesium oxide – calcium silicate [408] at 470 °C and 0.5 kPa to give a mixture consisting essentially of 3,5-xylenol, isophorone, mesityl oxide, and unconverted acetone. High space – time yields of 3,5-xylenol are only possible if acetone and mesityl oxide are recycled and isophorone is manufactured as a coproduct [408].

Dimethylcumene Hydroperoxide Cleavage. According to the process of the Mitsubishi Gas Chemical Co. in Japan, 3,5-xylenol is obtained by oxidation of 3,5-dimethylcumene to 3,5-dimethylcumene hydroperoxide [409], which is then cleaved [410]. By comparison with the analogous cymene – cresol process (see Section 1.4.3), this method is more expensive, e.g., because the additional methyl group of the dimethylcumene leads to more side reactions (formation of more primary peroxide) and oxidation degrees of only ca. 10 % are economical. The 3,5-dimethylcumene is readily accessible by propylation of *m*-xylene in the presence of catalytic amounts of $AlCl_3$. The ratio of 3,5-

to 2,4-dimethylcumene at isomerization equilibrium is ca. 99 : 1 [411].

Alkaline Chloroxylene Hydrolysis. Under the conditions of the chlorotoluene–cresol process (see Section 1.4.2), chloroxylenes are hydrolyzed to xylenols. From 2,4- or 3,5-dimethylchlorobenzene, for example, one can obtain mixtures of 2,4- and 3,5-xylenol [412] that can be separated by distillation. The hydrolysis of 2,5-dimethylchlorobenzene under pressure leads to 2,5-xylenol. Similarly, the alkaline hydrolysis of 2,3- and 3,4-dimethylbromobenzene gives 2,3- and 3,4-xylenol [413].

Alkali Fusion of Xylenesulfonates. The toluenesulfonic acid – cresol processes (see Section 1.4.1) also give xylenols if xylene is used as the starting material instead of toluene. Thus, the sulfonation of *m*-xylene with 95 % sulfuric acid (4 h at 150 °C) or chlorosulfuric acid (1 h at 50 °C) yields a product consisting primarily of 2,4-dimethylbenzenesulfonic acid, the fusion of which with sodium hydroxide at 320 °C leads to 2,4-xylenol in 79 % yield [414]. If the sulfonation mixture is heated to a temperature of up to 220 °C, the 2,4-dimethylbenzenesulfonic acid isomerizes to the thermodynamically more stable 3,5-form, the fusion of which with excess alkali gives 3,5-xylenol [415]. If the 2,4-/3,5-dimethylbenzenesulfonic acid mixture is heated with water (steam) to 140 – 160 °C, the 2,4-dimethylbenzenesulfonic acid is selectively hydrolyzed; in the following alkali fusion, the residual 3,5-dimethylbenzenesulfonic acid is converted into 3,5-xylenol of 97 – 98 % purity [416] in a yield of ca. 70 %, based on converted *m*-xylene [68].

The sulfonation of *p*-xylene with 98 % sulfuric acid at 140 °C gives 2,5-dimethylbenzenesulfonic acid. Fusion of this with alkali metal hydroxide at 330 – 340 °C gives 2,5-xylenol in 80 % yield, based on *p*-xylene [417], [418].

3,4-Dimethylbenzenesulfonic acid, which is accessible in ca. 98 % yield by sulfonation of *o*-xylene with 96 % sulfuric acid at ca. 150 °C [419], [420], gives 3,4-xylenol in more than 90 % total yield when fused with alkali metal hydroxide at 320 – 330 °C [419]. Under kinetically controlled (mild) sulfonation conditions it is also possible to obtain some 2,3-dimethylbenzenesulfonic acid, and therefrom some 2,3-xylenol.

The alkali fusion of xylenesulfonates is used industrially, e.g., by Inspec Fine Chemicals in the United Kingdom (for 2,5-xylenol and 3,4-xylenol) and by the Taoka Chemical Company in Japan (for 2,3-xylenol and 3,4-xylenol).

Other routes to the xylenols include the Gulf oxychlorination process, nucleus hydroxylation (see Section 1.4.5), and oxidative decarboxylation of dimethylbenzoic acids [244], boiling the diazonium salts of xylidines [421], hydrogen peroxide oxidation of dimethylbenzaldehydes [422], oxidation of 2,4,6-trimethylphenol to 3,5-dimethyl-4-hydroxybenzaldehyde, which is decarbonylated to 2,6-xylenol [423], or hydrogenation of 2-*N,N*-dialkylaminomethyl-4-methylphenol (obtained from *p*-cresol, dialkylamine, and formaldehyde) to give 2,4-xylenol [424].

2,4-Xylenol was also obtained by hydrogenation of bis-[hydroxy(methyl)phenyl]methane, which was obtained by condensation of *o*-cresol with formaldehyde [425].

2,3-Xylenol can be produced by hydrogenation and debutylation of 3,5-di-*tert*-butyl-2-hydroxy-6-methylbenzyl acetate, which is obtained from 4,6-di-*tert*-butyl-3-methylphenol, formaldehyde, and acetic acid [426].

Isomerization. Xylenols for which there is inadequate demand and xylenol fractions from which certain xylenols have already been isolated can be isomerized catalytically. Aluminum oxides or silicates, the activities of which have been further increased with hydrogen fluoride, titanium dioxide, iron oxide, and other metal oxides [427–431], are suitable as catalysts. The equilibrium distribution of the xylenols (2,3 : 2,4 : 2,5 : 2,6 : 3,4 : 3,5-isomer) after their isomerization in the vapor phase at 340 °C on aluminum silicate is given as 9 : 30 : 24 : 11 : 12 : 14 [431]. In the liquid phase at 105 – 130 °C with aluminum chloride in combination with hydrogen chloride as catalyst, 3,5-xylenol is primarily formed [235]; however, uneconomically large amounts of $AlCl_3$ are consumed. With regard to isomerization with $HF - BF_3$, see [432]. If phenol [235] or an aromatic solvent such as benzene is present, transmethylation occurs, with the simultaneous for-

mation of cresol or toluene (see page 16). A recycling process for the preparation of 3,5-xylenol by separating it from the distilled isomer mixture with KX zeolite as adsorbent (*n*-hexanol as desorbent) and isomerizing the residual mixture again with silica – alumina at 400 °C in the presence of phenol, cresol, and trimethylphenol is described in [433].

2.5. Quality Specifications and Analysis

The xylenols are characterized according to the same methods of analysis as are used for cresols (see Section 1.6).

The fraction sold as a 2,4-/2,5-xylenol mixture is a clear yellow liquid that contains ca. 50 – 75 % 2,4-xylenol and ca. 30 – 10 % 2,5-xylenol, depending on its origin. The main impurities are *m*- and *p*-cresol and ethylphenols. 2,4-Xylenol from the methylation of phenol on liquid-phase catalysts (UK-Wesseling) has contained more than 89 % 2,4-xylenol and less than 7.5 % 2,5-xylenol. Pure 2,4-xylenol is offered with a content of ca. 97 % (*fp* 23.0 °C). 2,3-Xylenol and 2,5-xylenol are available with purities of ca. 97 % and ca. 98 %, respectively. 2,6-Xylenol is traded at purities of greater than 99.5 %, 98 – 99 %, 96 – 98 %, and 90 – 96 %. 3,4-Xylenol is available in the grades "technically pure" ($\geq$ 95 %; *fp* $\geq$ 62.2 °C) and "pure" ($\geq$ 98 %; *fp* $\geq$ 64 °C). 3,5-Xylenol is commercially available in purity grades of 95 % (*fp* 60.3 °C) and 99 % (*fp* 62.5 °C).

Xylenol mixtures differ widely in composition, depending on the manufacturer and the requirements of the customer. For example, there are xylenol mixtures (cresylic acids) that contain almost 50 % cresol or 50 % higher alkyl phenols (C_9 and higher). The mixtures may also contain considerable amounts of ethylphenols in addition to the xylenols, depending on their origin. A rough classification is possible according to the distillation range (5 – 95 %), which could be, for example, 200 – 220 °C, 205 – 215 °C, 210 – 225 °C, 220 – 240 °C, 220 – 255 °C, or 230 –240 °C.

2.6. Handling, Storage, and Transportation

The xylenols are similar to the cresols in their physiological behavior. Their handling, storage, and transportation are, therefore, subject to the precautions applicable to cresols (see Section 1.7). This also applies to the precautions taken to prevent discoloration (storage and transportation in vessels of stainless steel, protection from light and moisture, and possible storage under nitrogen). References to the handling of xylenols are made in [353–356].

Xylenols and their mixtures are transported in ca. 200-L steel drums. Large volumes are shipped in liquid form by road or railway tank cars and export containers, which in the case of the pure isomers must be thermostated at a temperature above their melting point. They must be filled and transported under nitrogen. The shipping containers must bear the appropriate hazard warnings in accordance with the regulations.

The transportation of xylenols and xylenol mixtures is subject to the following regulations:

International sea transport (IMDG Code) and international air transport (ICAO/ATA-DGR): Class 6.1, UN no. 2261, packaging group II. Transport on European roads (ADR), railways (RID), and waterways (ADNR): Class 6.1, no. 14b.

National regulations: United States DOT-CFR 49: § 172.101 class 6.1, UN no 2261. With regard to their transportation on roads (GGVS), railways (GGVE), and waterways (GGVBinSch) in Germany, these products belong to class 6.1, no. 14b.

According to "Verordnung über brennbare Flüssigkeiten" (VbF), the properties and flash points ($\geq$ 86 °C) of xylenol mixtures and xylenols place them in group A, hazard class III (inasmuch as they are liquids). Water fog, foam, dry extinguishers, and, for small fires, carbon dioxide are suitable for fire fighting.

2.7. Uses

Xylenol mixtures are used as solvents (e.g., for wire enamels), disinfectants (e.g., in sheep dips), textile auxiliaries, and in ore flotation. Mixtures, particularly if they are rich in 3,5-xylenol and xylenols with two hydrogen atoms in the *o*- and/or *p*-positions, are used in the manufacture of xylenol – formaldehyde resins. Xylenols without an *o*-methyl group are used to produce nontoxic plasticizers and trixylenyl phosphates that serve as fire-resistant hydraulic fluids.

2,3-Xylenol is only used in small amounts, e.g., for the production of a pesticide.

2,4-/2,5-Xylenol mixtures constitute raw materials for antioxidants, especially those intended for gasoline (e.g., 6-*tert*-butyl-2,4-dimethylphenol) and rubber [e.g., 2-methyl-1,1-bis(2-hydroxy-3,5-dimethylphenyl)propane]. Pure 2,5-xylenol is used for the manufacture of pharmaceuticals, e.g., lipid reducers like gemfibrozil.

2,6-Xylenol is used primarily to produce poly(phenylene oxide) (PPO) resins, which are distinguished by high impact resistance, thermal stability, fire resistance, and dimensional stability. Oxidation of 2,6-xylenol gives 2,6-xylenol dimer, a speciality monomer to produce epoxy resins for encapsulating advanced semiconductors. 2,2-Bis(4-hydroxy-3,5-dimethylphenyl)propane (tetramethylbisphenol A), obtained by the condensation of 2,6-xylenol with acetone, serves as an intermediate for polycarbonates whose properties are similar to those of the PPO resins. 2,6-Xylenol is also treated with ammonia to yield 2,6-dimethylaniline, which is used especially in pesticide manufacture, [e.g., the fungizide metalaxyl (Ridomil) and herbizide metazachlor (Butisan S)] but also for the production of pharmaceuticals, e.g., local anesthetics like lidocain. Smaller amounts of 2,6-xylenol are used in the formation of pesticides, and in the manufacture of disinfectants as well as the antioxidant bis(4-hydroxy-3,5-dimethylphenyl)methane.

3,5-Xylenol is used in the production of various pesticides, e.g., 3,5-dimethyl-4-methylthiophenyl methylcarbamate (methiocarb, Mesurol), a broad-spectrum insecticide, acaricide, and molluscicide. 3,5-Xylidine, obtained by amination, is used particularly for perylene pigments. Alkylation with methanol gives 2,3,5-trimethylphenol, which can serve for the production of vitamin E. After chlorination to 4-chloro- and 2,4-dichloro-3,5-dimethylphenol, 3,5-xylenol is used as a disinfectant and industrial preservative. Other xylenols or xylenol mixtures likewise serve as disinfectants after chlorination.

3,4-Xylenol is the starting material for the insecticide 3,4-dimethylphenyl methylcarbamate (MPMC, Meobal).

Pure xylenols, besides the use for pharmaceuticals, are also used in small amounts in the synthesis of dyes, and fragrances.

2.8. Economic Aspects

As a rough estimate, the total output of xylenols in the United States, Western Europe, and Japan in 1998 was ca. 165 000 t. Xylenol mixtures accounted for ca. 30 000 – 40 000 t, 2,6-xylenol for ca. 125 000 t, and the remaining individual isomers (especially 2,4-, 2,5- and 3,5-xylenol), obtained by synthesis or isolated from mixtures, for ca. 10 000 t.

The xylenol mixtures are obtained – mainly in the United States, South Africa and Europe – from spent refinery caustics and coal tars. The production of xylenol by synthesis is accounted for by the United States, Europe, and Japan in the approximate ratio 54 : 24 : 22.

3. Environmental Protection

Cresols and xylenols are readily degradable biologically; the xylenols are degraded somewhat slower than the cresols [434]. Biological treatment plants with suitably adapted bacteria are able to degrade wastewaters with 900 mg of cresol/L and more [435].

In the environment, cresols are degraded at low concentrations, especially by bacteria that are widely distributed in the soil and water (particularly *Pseudomonas* species). Degradation by other organisms, including yeasts, fungi, algae, and higher plants, as well as by photolysis (half-life in air < 1 d), is also known [436]. Accordingly, these compounds do not persist in the environment [437], [438].

However, as cresols and xylenols show high acute toxicity both toward fish (96 h LC_{50} 1 – 10 mg of cresol/L) and to aquatic invertebrates, and as they impair the taste of edible fish and drinking water even in very low concentrations [436], [438], [439], the release of these compounds into surface waters is subject to strict limitations. Therefore, leaked or spilled cresols or xylenols must be prevented from entering surface waters or groundwater.

Liquid products can be adsorbed, e.g., with sawdust, peat, earth, kieselguhr, or other adsor-

bents, and then burned completely in special incineration plants. Solidified products should not be dissolved, but removed mechanically. Final residues can be neutralized with 2 – 3 % aqueous sodium hydroxide and flushed with plenty of water into suitable drains leading to a biological wastewater treatment plant.

Cresols and xylenols can generally be removed from production wastewater by extraction, e.g., with diisopropyl ether, butyl acetate, methyl isobutyl ketone, or toluene [440], to such an extent that the residual contents can be eliminated without difficulty in a biological wastewater treatment plant. The wastewater should enter the biological treatment plant continuously, not intermittently in batches.

The phenolic constituents of wastewater can also be removed to a very large extent with ion-exchange resins or such adsorbents as activated carbon [71], which can be subsequently regenerated, or with oxidizing agents such as hydrogen peroxide [441] or ozone [442], [443]. Stripping with steam is likewise possible, but is less effective.

Cresols and xylenols can be removed from waste gas streams by scrubbing with dilute aqueous sodium hydroxide or by incineration. Another effective method is adsorption on activated carbon, which is able to accumulate up to 0.5 kg of cresols/kg. The adsorbate is subsequently recoverable when the activated carbon is regenerated [444].

In the United States the cresol isomers and cresylic acids are classified as hazardous air pollutants (HAPs) in Title III of the 1990 Clean Air Act Amendements. Therefore, producers and users of cresols and cresylic acids must follow suitable operating procedures and install emission control equipment that follows the maximum achievable control technology (MACT) regulations of the US Environmental Protection Agency (EPA).

4. Toxicology and Occupational Health

Cresols and xylenols burn the skin and mucous membranes through degradation of proteins, similarly to phenols. Initially white, and later brownish black necroses appear on the skin. Certain individuals may show hypersensitivity symptoms in addition to the corrosive effect [353], [354].

Cresols and xylenols, like phenol, are toxic when absorbed through the skin. Their TLV or MAK value (5 ppm) is therefore supplemented by the word "skin" or the letter H (standing for the German word *Haut*). This percutaneous effect is easily overlooked because the initial burning sensation is followed by local anesthesia. The compounds are absorbed particularly rapidly when in the liquid form or dissolved (diluted). Even if only a relatively small area of the skin is exposed, e.g., a single arm or hand, there may be severe injury to health, possibly developing into a lethal paralysis of the central nervous system. Corresponding amounts of cresol or xylenol that are taken up through the respiratory tract, as vapor or as droplets, or through the stomach also cause severe poisoning in addition to corrosion. Systemic damage to liver and kidneys is possible. The initial or consecutive symptoms of poisoning are headache, dizziness, ringing in the ears, nausea, vomiting, muscular twitching, mental confusion, and loss of consciousness, depending on severity.

As phenols are rapidly absorbed through the skin and a causal treatment of the poisoning does not yet exist, rinsing them off the skin (not wiping them off) as soon as possible constitutes a decisively important first aid measure. A mixture of polyethylene glycol and ethanol in a 2 : 1 ratio has thus far proven most suitable for this purpose: it should be kept available for rapid use at the workplace in portable pressure vessels (similar to fire extinguishers, but painted a different color, e.g., green) [353]. Polyethylene glycol 400, or water, is less suitable, particularly in the case of xylenols, which are less soluble and may solidify on the skin. Water in very large quantities should only be used if no other solvents are immediately available. Eyes should be rinsed only with water, but very thoroughly, for at least 15 min. If cresols or xylenols enter the stomach, a large amount of water should be given immediately in small portions to achieve a diluting effect and, in addition, medicinal charcoal (1 tablespoon to 1 glass of water). Medical attention must be given, without delay, in every case. Further information on the prevention of accidents with cresols and xylenols, as well as the steps that should be taken if accidents occur, will be found in [353], [356].

Table 5. Some acute toxicity data for cresols and xylenols (LD_{50}, mg/kg)

	RTECS No.	Oral		Skin	
	[452]	Rat	Mouse	Rat	Rabbit
o-cresol	G06300000	121	344	620	890
m-cresol	G06125000	242	828	1100	2050
p-cresol	G06475000	207	344	750	301
2,4-xylenol	ZE5600000	3200	809	1040	
2,5-xylenol	ZE5775000	444	383		
2,6-xylenol	ZE6125000	296/1750	450/980	2325	1000
3,4-xylenol	ZE6300000		400		
3,5-xylenol	ZE6475000	608/3620	477/620	2400	2000

Prolonged contact with cresols or xylenols at concentrations that, although low, are above the MAK value (5 ppm) may result in absorption through the skin, mucous membranes, and respiratory organs, and cause gastrointestinal disturbances (vomiting, loss of appetite), nervous disorders, headache, dizziness, fainting, and dermatitis.

According to epidemiological investigations [445], cresols are not carcinogenic. A carcinogenic effect would indeed be surprising, since *p*-cresol, for example, is a normal product of human metabolism. Cresols have likewise caused no carcinomas in animal experiments [436], [446]. In the *Salmonella* microsome test, all three cresol isomers have been found to be nonmutagenic toward the most widely differing *Salmonella* strains, both in the presence and in the absence of hamster or rat microsomal fractions [447]. The sister-chromatid exchange mutagenicity tests of the cresol isomers revealed no dose-dependent effects either in vitro in cultured human fibroblast or in vivo in male mice [448]. Toxicological evaluations, of 2,4-xylenol, 2,6-xylenol, and 3,5-xylenol [449], have shown no mutagenic activity, and the available carcinogenicity studies are inconclusive.

Extensive literature compilations on the toxicity of the cresols may be found in [436], [450], [444], which also contain some information on the xylenols, and in [446], [451]. For 2,4-, 2,6-, and 3,5-xylenol, see [449]. Toxicity data for cresols and xylenols are summarized in [452–454]. Some acute toxicity data are compiled in Table 5.

5. References

General References

1. *Beilstein,* **6H,** 349 – 389; **6I,** 169 – 196; **6II,** 322 – 368; **6III,** 1233 – 1341; **6IV,** 1940 – 2093 (Cresols); **6H,** 480 – 494; **6I,** 240 – 245; **6II,** 453 – 466; **6III,** 1722 –1769; **6IV,** 3006 – 3164 (Xylenols).
2. *Houben-Weyl,* **6/1 c,** part 1 and 2.
3. *Ullmann,* 4th ed., **15,** 61 – 77, 5th ed., **A8,** 25 – 59.

Specific References

4. G. Städeler, *Ann. Chem. Pharm.* **77** (1851) 17 – 37.
5. *Geigy Scientific Tables,* **vol. 1,** Ciba-Geigy Ltd., Basel 1984, p. 99.
6. L. Moio, J. Dekimpe, P. Etievant, F. Francesco, *J. Dairy Res.* **60** (1993) 199 – 213.
7. M. Lehtonen, *J. Assoc. Off. Anal. Chem.* **66** (1983) 62 – 70.
8. D. M. Knight et al., *Org. Geochem.* **4** (1982) no. 2, 85 – 92.
9. R. J. L. Andon, D. P. Biddiscombe, J. D. Cox, R. Hanley, D. Harrop, E. F. G. Herington, J. F. Martin, *J. Chem. Soc.* 1960, 5246 – 5254.
10. R. J. L. Andon, J. F. Counsell, E. B. Lees, J. F. Martin, C. J. Mash, *Trans. Faraday Soc.* **63** (1967) 1115 – 1121.
11. A. P. Kudchadker, S. A. Kudschadker, R. C. Wilhoit: *Key Chemicals Data Books, Cresols,* Thermodynamics Res. Center, Texas A & M University, College Station, Texas, 1978.
12. ESDU: *Thermophysical Properties of Industrial Important Fluids on the Saturation Line: Cresols,* item 82 032, Engineering Sciences Data Unit, London 1983.
13. C. L. Yaws, H. M. Ni, P. Y. Chiang, *Chem. Eng. (N.Y.)* **1988,** May 9, 91 – 98.
14. C. L. Yaws, P. Y. Chiang, *Chem. Eng. (N.Y.)* **1988,** Sept. 26, 81 – 88.
15. N. G. Buckman, J. O. Hill, R. J. Magee, *J. Therm. Anal.* **36** (1990) 2555 – 2566.
16. S. L. Randzio, E. A. Lewis, D. J. Eatough, L. D. Hansen, *Int. J. Thermophys.* **16** (1995) 883 – 900.
17. A. Ksiazczak, J. Kosinski, *Fluid Phase Equilib.* **55** (1990) 17 – 37.

18. J. McGarry, *Ind. Eng. Chem. Process Des. Dev.* **22** (1983) 313 – 322.
19. E. Terres, F. Gebert, H. Hülsemann, H. Petereit, H. Toepsch, W. Ruppert, *Brennst. Chem.* **36** (1955) 359 – 372.
20. J. Gmeling, U. Onken, W. Arlt resp. U. Weidlich or J. R. Rarey: *Vapor-liquid equilibrium Data Collection* **vol. I, part 2 b (1978), vol. I, part 1 a (1981), vol. I, part 2 c (1982), vol. I, part 2 d (1982), vol. I, part 2 f (1990),** DECHEMA-Chemistry Data Series.
21. H. M. Sebastian, H.-M. Lin, K.-C. Chao, *J. Chem. Eng. Data* **25** (1980) 381 – 383.
22. H. Kim, W. Wang, H.-M. Lin, K.-C. Chao, *J. Chem. Eng. Data* **28** (1983) 216 – 218.
23. *Rodd's Chemistry of Carbon Compounds,* **vol. III,** part A, 1983, p. 161 – 219; 2nd Supplement to vol. III, part A, 1996, 225 – 376, Elsevier Scientific Publ. Co., Amsterdam-Oxford-New York.
24. R. Wandas, J. Surygala, E. Sliwka, *Fuel* **75** (1996) 687 – 694.
25. Hydrocarbon Res. Inc., US 4 431 850, 1984. *Chem. & Eng. News* **59** (Nov. 1981) 32, 34.
26. Bayer AG, EP 330 036, 1989 (A. Schnatterer, H. Fiege). Bayer AG, EP 451 650, 1991 (A. Schnatterer, H. Fiege, K.-H. Neumann).
27. *DECHEMA-Werkstofftabelle: Kresole,* no. 851 – 852, Verlag Chemie, Weinheim 1959.
28. A. A. Sagues, B. H. Davis, T. Johnson, *Ind. Eng. Chem. Process Des. Dev.* **22** (1983) 15 – 22.
29. M. A. Elliot (ed.): *Chemistry of Coal Utilization,* **2nd suppl. vol.,** J. Wiley & Sons, New York-Chichester-Brisham-Toronto 1981.
30. A. Dierichs, R. Kubička: *Phenole und Basen,* Akademie Verlag, Berlin 1958.
31. J. Schulze, H. Sutter, M. Weiser, *Chem. Ind. (Berlin)* **35** (1983) 448 – 454.
32. R. Cypres, S. Furfari, M. Ghodsi, *Erdöl Kohle* **36** (1983) 471 – 477.
33. H.-G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Berlin-Heidelberg-New York 1968.
34. E. I. Kazakow, E. D. Vilyanskaya, G. A. Markus et al., *Khim. Tverd. Topl. Moscow* **17** (1983) no. 3, 93 – 96 (*Solid Fuel Chem. (Engl. Transl.)* New York (1983) 89 – 92).
35. G. Bett, T. G. Harvey, T. W. Matheson, K. C. Pratt, *Fuel* **62** (1983) 1445 – 1454.
36. R. E. Schirmer, T. R. Pahl, D. C. Elliot, *Fuel* **63** (1984) 368 – 372.
37. H. E. Jegers, M. T. Klein, *Ind. Eng. Chem. Process Des. Dev.* **24** (1985) 173 – 183.
38. J. A. Russel, P. M. Molton, S. D. Landsmann: *Alternative Energy Sources (Washington)* **3** (1980/1983) no. 3, 307 – 322.
39. K. Sipila, E. Kuoppala, L. Fagernas, A. Oasmaa, *Biomass Energy* **14** (1998) 103 – 113.
40. H. L. Lochte, E. R. Littmann: *The Petroleum Acids and Bases,* Chemical Publ. Co., New York 1955, pp. 121 – 125. H. L. Coonradt, B. W. Rope, *Ind. Eng. Chem. Process Des. Dev.* **2** (1963) 317 – 322.
41. Sasol Phenolics, Media Release from May 15, 1997; *Chem. Week,* **1997,** May 21, 19.
42. Sasol. A Corporate Report. Supplement to Financial Mail, June 14, 1985, p. 54; A. Geertsema, *Proc. Annu. Int. Pittsburgh Coal Conf. 7th* **(1990)** 571 – 578.
43. G. Baker, D. Duncan, A. Kuhn, D. Maas, *Proc. Annu. Int. Coal Conf. 7th* **(1990)** 561 – 570.
44. H.-J. Wurm, *Chem. Ing. Tech.* **48** (1976) 840 – 845.
45. A. K. Kuhn, D. H. Duncan, *Proc. Annu. Int. Pittsburgh Coal Conf. 13th* **vol. 1** (1996) 478 – 483.
46. Dakota Gasification Co., EP 545 814, 1992 (D. Duncan, G. Baker, D. Maas, K. Mohl, R. Todd).
47. R. E. Maple: *Symposium on Refining Petroleum for Chemicals,* Am. Chem. Soc. New York City Meeting Sept. 7 – 12, 1969, D 105 – 113.
48. *Chem. Eng.* **69** (1962) August 20, 66 – 77.
49. R. E. Maple, A. R. Price, *Hydrocarbon Process.* **51** (1972) no. 10, 168 – 172.
50. R. E. Maple, A. R. Price, *AIChE Symp. Ser.* **73** (1977) 200 – 206.
51. A. R. Price, *AIChE National Conference on Complete Water Reuse,* New York, April 23 – 24 (1973) pp. 333 – 341.
52. D. C. Jones, J. A. Kohlbeck, M. B. Neuworth, *Ind. Eng. Chem. Prod. Res. Dev.* **2** (1963) 217 – 220.
53. T. P. Forbath (ed.), *Chem. Eng. (N.Y.)* **64** (1957) no. 7, 228 – 231.
54. "Chemical Profile: Cresylics", *Chem. Mark. Rep.,* **23** Dec. 1996, 33.
55. S. W. Eglund, R. S. Aries, D. F. Othmer, *Ind. Eng. Chem.* **45** (1953) 189 – 197.
56. D. McNeil, *Chem. Eng. Monogr.* **15** (1982) 209 – 232.
57. H. Wang, C. Xu, D. Zhang, *Huagong Jinzhan* **17** (1998) no. 3, 25 – 29; *Chem. Abstr.* **129** (1998) 330 492.
58. W. Pritzkow, P. Grothkopf, R. Höring, H. Gross, *Z. Chem.* **5** (1965) 300.
59. Honshu Chem. Ind. Co., JP 70 28 976, 1967; *Chem. Abstr.* **74** (1971) 12 828.

60. G. D. Kharlampovich et al., *Khim. Promst. (Moscow)* **44** (1968) 16 – 20.
61. G. D. Kharlampovich, V. F. Kollegow et al., *Khim. Tekhnol., Resp. Mezhved. Nauchno.-Tekh. Sb.* **5** (1966) 77 – 80.
62. Tennessee Corp., US 2 828 333, 1956; US 2 841 612, 1956.
63. Marni S.A., DE 24 13 344, 1974.
64. Hoechst, DE 28 37 549, 1978.
65. M. Brendel, G. Gárdos, A. Rédey, F. Mikó, *Chem. Tech.* **46** (1994) 20 – 24.
66. A. A. Spryskow, *Zhur. Obsheĭ Khim.* **30** (1960) 2449 – 2453; *Chem. Abstr.* **57** (1962) 16 464.
67. Honshu Chem. Ind. Co., JP 78 21 142, 1976; *Chem. Abstr.* **88** (1978) 190 393.
68. Taoka Chem. Co., EP 80 880, 1982.
69. P. Taffe, *Chem. Age (London)* **9** (May 1980) 18.
70. E. Hurwitz, E. Ciabettari, R. A. Wolff, I. Bernstein, *Ind. Eng. Chem.* **51** (1959) 1301 – 1304.
71. C. D. Baker, E. W. Cark, W. V. Jesernig, C. H. Huether, *AIChE Sympos. Ser.* **70** (1974) 686 – 692.
72. Bayer AG, EP 586 918, 1993 (H. Schrage, H. Fiege).
73. Bayer AG, EP 292 824, 1988 (F.-J. Mais, H. Fiege, K. Röhlk, K. Wedemeyer).
74. Bayer AG, EP 368 063, 1989 (F.-J. Mais, H. Fiege).
75. R. N. Shreve, C. J. Marsel, *Ind. Eng. Chem.* **38** (1946) 254 – 261.
76. O. Lindner, *Chem. Ing. Tech.* **36** (1964) 769 – 774.
77. Bayer AG, DE 29 04 831, 1979.
78. A. T. Bottini, J. D. Roberts, *J. Am. Chem. Soc.* **79** (1957) 1458 – 1462.
79. M. Zoratti, J. F. Bunnet, *J. Org. Chem.* **45** (1980) 1769 – 1776.
80. Dow Chem. Co., US 4 001 340, 1975.
81. Dow Chem. Co., US 1 959 283, 1930; US 2 126 610, 1933.
82. Asahi Chem. Ind., JP 69 17 372, 1967; *Chem. Abstr.* **71** (1969) 112 620.
83. Frontier Chem. Co., US 3 413 341, 1965.
84. Bayer AG, DE 19 633 808, 1996 (D. Bielefeld, M. Jautelat, V. Handschuh).
85. Sumitomo Chem. Co., DE 18 03 036, 1968.
86. Sumitomo Chem. Co., DE 23 25 354, 1973.
87. Hercules Powder, GB 754 872, 1954.
88. R. H. Allen, T. Alfrey, L. D. Yats, *J. Am. Chem. Soc.* **81** (1959) 42 – 46.
89. R. H. Allen, L. D. Yats, *J. Am. Chem. Soc.* **83** (1961) 2799 – 2805.
90. Rütgerswerke & Teerverwertung, DE 12 74 096, 1969.
91. Cosden Oil & Chemical Co., DE 22 47 308, 1972.
92. H. Boardman, *J. Am. Chem. Soc.* **84** (1962) 1376 –1382.
93. G. D. Serif et al., *Can. J. Chem.* **31** (1953) 1229 –1238.
94. Sumitomo Chem. Co., JP 78 82 732, 1976; *Chem. Abstr.* **90** (1979) 6090.
95. Sumitomo Chem. Co., JP 78 82 731, 1976; *Chem. Abstr.* **90** (1979)6091.
96. Mitsui Petroch. Inds., JP 83 198 468, 1982; *Chem. Abstr.* **100** (1984)174 433.
97. Sumitomo Chem. Co., US 5 399 791, 1995; EP 553 874, 1993 (K. Ikimi et al.).
98. Mitsui Petroch. Inds., JP 82 59 861, 1980; *Chem. Abstr.* **97** (1982) 162 579.
99. Hercules Powder, US 2 728 797, 1954.
100. Sumitomo Chem. Co., WO 9 401 387, 1994 (K. Ikimi et al.).
101. Hercules Powder, US 2 779 797, 1953.
102. G. D. Kharlampovich et al., *Zh. Prikl. Khim.(Leningrad)* **46** (1973) 118 – 121.
103. Mitsui Petroch. Inds., JP 73 49 735, 1971; *Chem. Abstr.* **79** (1973) 136 803.
104. Mitsui Petroch. Inds., US 4 408 083, 1981.
105. Mitsui Petroch. Inds., JP 4 036 217, 1974; *Chem. Abstr.* **82** (1975) 170 367.
106. Sumitomo Chem. Co., JP 71 23 723, 1968; *Chem. Abstr.* **75** (1971) 88 329.
107. Mitsui Petrochem. Inds., EP 21 848, 1980.
108. Goodyear Tire & Rubber, EP 77 749, 1982.
109. R. S. Sushko, V. N. Kiva, *Vses. Konf. Teor. Prakt. Rektifikatsii,* [Dokl.] 3rd ed., vol. I, E. G. Melamed (ed.), 1973, pp. 177 – 180; Chem. Abstr. 88 (1978) 63 599.
110. H. Kaminata, *Kagakushi Kenkyu* **25** (1998) 126 – 138; *Chem. Abstr.* **130** (1998) 40 024.
111. K. Ito, *Hydrocarbon Process.* **52** (1973) no. 8, 89 – 90.
112. *Chem. Eng. News* 1953, 246; 1964, 29.
113. R. W. Neuzil, D. H. Rosback, R. H. Jensen, J. R. Teague, A. J. de Rosset, *CHEMTECH* **10** (1980) 498 – 503.
114. "Coproduction of Phenol and Acetone in the m-Cresol Unit of the Yanshan Petrochemical Corp.", *China Chem. Rep.,* (1995) no. 17, 10. Sept., p. 6.
115. UOP Inc., US 4 128 593, 1978.
116. Mobil Oil Corp., EP 12 613, 1979.
117. Sumitomo Chem. Co., EP 127 410, 1984.
118. Société Chimique des Charbonnages Courbevoie, US 3 962 366, 1974.
119. *Chem. Process. (Chicago)* 1966, Mid-November, 113.

120. T. F. Meinhold, J. Bardzik, *Chem. Process. (Chicago)* 1968, April, 31 – 35.
121. Rütgerswerke & Teerverwertung, US 3 347 936, 1965.
122. Y. V. Churkin et al., *Khim. Promst. (Moscow)* 1972, 660 – 666.
123. General Electric Co., US 4 041 085, 1975.
124. General Electric Co., US 3 962 126, 1974.
125. Mitsubishi Petrochemical Co., EP 531 785, 1993 (M. Inaba et al.).
126. Mitsubishi Petrochemical Co., JP 043 46 952, 1992 (M. Inaba, M. Kataoka, M. Abe); *Chem. Abstr.* **119** (1993) 8 492.
127. Asahi Chemical Ind. Co., EP 101 138, 1983.
128. T. Dozono, *Petrotech (Tokyo)* **11** (1988) 776 – 778; *Chem. Abstr.* **110** (1989) 231 179.
129. H. Ohashi, K. Takahashi, *Aromatikkusu* **44** (1992) 350 – 356; *Chem. Abstr.* **119** (1993) 183 212.
130. Asahi Chem. Ind., JP 052 86 880, 1993 (K. Takahashi, H. Murotani); *Chem. Abstr.* **120** (1994) 163 682.
131. Asahi Chem. Ind., JP 061 16 188, 1994 (K. Takahashi, K. Sakai); *Chem. Abstr.* **122** (1995) 105 415.
132. Asahi Chem. Ind., JP 062 98 687, 1994 (K. Takahashi, K. Sakai); *Chem. Abstr.* **121** (1994) 157 288.
133. K. T. Li, J. Wang, K. R. Chang, *Ind. Eng. Chem. Res.* **32** (1993) 1007 – 1011.
134. G. A. Kirichenko, V. N. Kiva, V. S. Bogdanov, *Russ. J. Phys. Chem. (Engl. Transl.)* **52** (1978) 893 – 894.
135. G. A. Kirichenko, V. N. Kiva, N. Z. Baibulatova, *Vses. Konf. Teor. Prakt, Rektifikatsii, [Dokl.], 3rd* **1** (1973) 86 – 89; *Chem. Abstr.* **87** (1977) 22 649.
136. Mitsui Toatsu Chemicals Inc., JP 81 46 829, 1979; *Chem. Abstr.* **95** (1981) 132 485.
137. Mitsui Toatsu Chemicals Inc., JP 81 65 836, 1979; *Chem. Abstr.* **95** (1981) 186 857.
138. H. W. Prengle, Jr., X. Van Tran, K. Moinzadeh, F. A. Bricout, S. Alam; Univ. Houston Texas, USA: Report, NREL/TP-253-4276; Order No. DE 91 002 141, 38pp. (1992); Avail. NTIS; *Chem. Abstr.* **120** (1994) 273 448.
139. General Electric Co., US 3 446 856, 1964.
140. H. Hattori, K. Shimazu, N. Yoshi, K. Tanabe, *Bull. Chem. Soc. Japan* **49** (1976) 969 – 972.
141. General Electric Co., WO 84/01 146, 1982.
142. General Electric Co., EP 127 833, 1984.
143. General Electric Co., US 4 933 509, 1989.
144. BASF, DE 3 524 331, 1985.
145. General Electric Co., DE 3 418 087, 1984.
146. General Electric, US 4 528 407, 1985.
147. China Petrochem Dev., US 4 876 398, 1988.
148. General Electric, EP 785 180, 1997 (G. L. Warner, A. J. Caruso).
149. Conoco Inc., US 4 359 408, 1981.
150. S. Sato, K. Koizumi, F. Nozaki, *J. Catal.* **178** (1998) 264 – 274.
151. General Electric Co., US 4 201 880, 1976.
152. K. Tanabe, T. Nishizaki, *Sixth Internat. Congress on Catalysis,* London, July 12 – 16, 1976, paper B 26, pp. 1 – 8.
153. Mitsubishi Petrochemical Co., JP 60 250 41, 1994 (M. Takiguchi, Y. Fukuhara, K. Yaguchi); *Chem. Abstr.* **120** (1994) 298 241.
154. Mitsui Toatsu Chemicals Inc., US 4 208 857, 1977.
155. Mitsui Toatsu Chemicals Inc., JP 81 55 327, 1979; *Chem. Abstr.* **95** (1981) 186 848.
156. Mitsui Toatsu Chemicals Inc., JP 82 28 018, 1980; *Chem. Abstr.* **96** (1982) 199 288.
157. Asahi Chemical Ind., JP 84 194 834, 1983; *Chem. Abstr.* **102** (1985) 114 797.
158. Asahi-Dow Ltd., JP 81 133 232, 1980; *Chem. Abstr.* **96** (1982) 34 853.
159. Mitsui Toatsu Chemicals Inc., US 4 329 517, 1979.
160. Asahi-Dow Ltd., GB 2 089 343, 1980.
161. Croda Synthetic Chemicals Ltd., EP 50 937, 1981.
162. Bayer AG, EP 10 24 93, 1983 (J. Käsbauer, K. Wedemeyer).
163. Agency of Industrial Science and Technology, JP 78 35 061, 1978; *Chem. Abstr.* **90** (1979) 6086.
164. T. Kotanigawa, M. Yamamoto, K. Shimokawa, Y. Yoshida, *Bull. Chem. Soc. Japan,* **44** (1971) 1961 – 1964.
165. Chemische Werke Hüls AG, DE 3 149 022, 1983 (R. Sigg).
166. Asahi-Dow Ltd., JA 57 091 936, 1982.
167. Chemopetrol, CS 276 514, 1992 (K. Svowoda, J. Kolena, R. Lensky, P. Moravek); *Chem. Abstr.* **120** (1994) 269 834.
168. Mitsubishi Gas Chem. Co., EP 419 045, 1989.
169. Union Rheinische Braunkohlen Kraftstoff AG, GB 2 072 674, 1981.
170. Mitsubishi Gas Chemical Co., US 4 024 195, 1974.
171. Nippon Steel Chem. Co., JP 84 181 232, 1983.
172. Asahi-Dow Ltd., DE 3 147 026, 1981.
173. Asahi Chem. Industry Co., JP 093 08 832, 1996 (H. Ota); *Chem. Abstr.* **128** (1997) 49 818.
174. Mitsui Toatsu Chemicals Inc., JP 81 43 229, 1979; *Chem. Abstr.* **95** (1981) 115 033.
175. Mitsui Toatsu Chemicals Inc., JP 81 45 427, 1979; *Chem. Abstr.* **95** (1981) 97 361.

176. Mitsui Toatsu Chemicals Inc., JP 81 49 330, 1979; *Chem. Abstr.* **95** (1981) 97 358.
177. Mitsui Toatsu Chemicals Inc., JP 83 208 244, 1982; *Chem. Abstr.* **100** (1984) 158 561.
178. T. Kotanigawa, *Bull. Chem. Soc. Japan* **47** (1974) 950 – 953.
179. E. M. Knyazewa, N. S. Khasanova, L. M. Koval, N. N. Sudakova, *Russ. J. Phys. Chem. (Engl. Transl.)* **58** (1984) 592 – 594.
180. Rütgerswerke AG, DE 27 56 461, 1977.
181. Y. V. Churkin, V. N. Khabibullin, G. A. Kirichenko, A. D. Zotov, *Sov. Chem. Ind. (Engl. Transl.)* **4** (1972) 550 – 553.
182. G. D. Kharlampovich, *Proizvod. Pereab. Plastmass Sint. Smol. (Moscow)* **7** (1978) 56 – 58; *Chem. Abstr.* **90** (1979) 199 365.
183. V. N. Vinogradova, G. D. Kharlampovich, *Sov. Chem. Ind. (Engl. Transl.)* **5** (1973) 558 – 560.
184. Conoco Inc., US 4 381 413, 1981.
185. P. S. Landis, W. O. Haag, *J. Org. Chem.* **28** (1963) 585.
186. S. E. Santos, J. M. Marinas, M. P. Alcasar, *React. Kinet. Catal. Lett.* **24** (1984) 247 – 251.
187. General Electric Co., GB 1 034 500, 1965.
188. M. F. Mazitov, A. S. Shelev, B. I. Bykow, V. Churkin, *Sov. Chem. Ind. (Engl. Transl.)* **12** (1980) 995 – 1002.
189. M. Huml, J. Jonaš, *Sb. Pr. Vyzk. Chem. Vyuziti Uhli, Dehtu Ropy (Prague)* **15** (1978) 199 – 210; *Chem. Abstr.* **89** (1978) 59 735.
190. J. Jonaš, CS 214 049, 1980; *Chem. Abstr.* **101** (1984) 54 706.
191. Consolidation Coal Co., US 3 624 163, 1968.
192. M. Janardanarao, G. S. Salvapati, R. Vaideswaran, *Proc. Natl. Symp. Catal., 4th 1978 (Bombay, India)* 1980, 51 – 57.
193. S. Namba, T. Yashima, Y. Itaba, N. Hara, *Stud. Surf. Sci. Catal. (Amsterdam)* **5** (1980) (Catalysis by Zeolites) 105 – 111.
194. S. Balsama, P. Beltrame, P. L. Beltrame et al., *Appl. Catal.* **13** (1984) 161 – 170.
195. Moscow Gubkin Petrochem. SU 1 004 342, 1981; *Chem. Abstr.* **99** (1983) 38 189. Yeda Research and Devel. Co., EP 126 245, 1984.
196. A. A. Agaev, D. B. Tagiev, *Neftekhimiya* **32** (1992) 2, 136 – 140; *Chem. Abstr.* **117** (1992) 173 785.
197. I. Yu. Ponomarenko, L. M. Koval, V. N. Belousova, V. V. Tartynov, *Zh. Fiz. Khim.* **66** (1992) 3231 – 3237; *Chem. Abstr.* **119** (1993) 138 560.
198. O. A. Sinitsyna, B. V. Romanovskii, *Kinet. Katal.* **33** (1992) 1154 – 1161; *Chem. Abstr.* **119** (1993) 162 824.
199. N. D. Limankina, E. A. Vdovtsova, A. S. Sultanov, *Uzb. Khim. Zh.* 1977, no. 6, 34 – 38; *Chem. Abstr.* **89** (1978) 129 157.
200. M. Marczewski, G. Perot, M. Guisnet, Preprints 1. Intern. Symposium Heterogenous Catalysis and Fine Chemicals, Poitiers, France, March 15 – 17, 1998, p. 80 – 89.
201. A. A. Agaev, D. B. Tagiev, Z. M. Pashaev, *Zh. Prikl. Khim. (S.-Petersburg)* **67** (1994) 1038 – 1040; *Chem. Abstr.* **122** (1995) 58 768.
202. B. N. Galich, A. N. Vasiliev, *Neftepererab. Neftekhim. (Kiev)* **44** (1993) 28 – 39; *Chem. Abstr.* **122** (1995) 109 238.
203. O. A. Sinitsyna, B. V. Romanovsky in K. G. Ione (ed.): *Chem. Ecol. Health Proc. Int. Meet. Zeolite Catal. Solution Environ. Probl. 1992,* Nova Science Publishers, Commack, N.Y. 1995, pp. 192 – 135.
204. Coal Tar Research Assoc., GB 1 125 087, 1966.
205. Asahi Chem. Ind., JP 74 028 742, 1970; *Chem. Abstr.* **82** (1975) 111 767. JP 74 043 943, 1970; *Chem. Abstr.* **83** (1975) 27 884. JP 74 043 944, 1970; *Chem. Abstr.* **83** (1975) 27 885.
206. A. A. Agaev, R. I. Mamedova, I. M. Gulieva, *Zh. Prikl. Khim. (S.-Petersburg)* **67** (1994) 464 – 466; *Chem. Abstr.* **121** (1994) 282 608.
207. E. A. Vdovtsova, A. Kochetkova, A. S. Sultanov, *Neftekhimiya* **22** (1982) 200 – 206.
208. O. V. Glukhikh, V. A. Proskuryakow, D. A. Sibarow, *Zh. Prikl. Khim. (Leningrad)* **57** (1984) 658 – 663.
209. Shell Devel. Co., US 2 440 036, 1948.
210. General Electric Co., US 4 225 732, 1977.
211. Ashland Oil Inc., DE 2 251 553, 1972.
212. General Electric Co., US 4 048 239, 1976.
213. *China Chem. Week* **3** (1996) no. 105, 2. Aug., p. 7.
214. Y. V. Churkin, L. A. Rusanova, G. D. Kharlampovich et al., *Tr., Nauch.-Isled. Neftekhim. Proizvod.* **2** (1970) 7 – 13; *Chem. Abstr.* **74** (1971) 76 121.
215. M. Inoue, S. Enomoto, *Chem. Pharm. Bull.* **24** (1976) 2199 – 2203.
216. M. Ehsan, G. S. Salvapati, M. Janardanarao, R. Vaideswaran, *Proc. Catsympo 80, Natl. Catal. Symp., 5th 1980* (pub. 1983) 31 – 37, Hyderabad, India.
217. Continental Oil Co., US 3 994 982, 1975.
218. Continental Oil Co., US 4 022 843, 1975.
219. Chemisches Werk Lowi GmbH, DE 1 265 755, 1965.
220. Chemisches Werk Lowi GmbH, DE 1 543 494, 1966.
221. Chemisches Werk Lowi GmbH, DE 1 817 342, 1968.

222. Coal Tar Res. Assoc., GB 1 060 036, 1965.
223. Union Rheinische Braunkohlen Kraftstoff AG, DE 1 281 448, DE 1 297 110, 1966.
224. Union Rheinische Braunkohlen Kraftstoff AG, GB 1 160 491, 1968.
225. Union Rheinische Braunkohlen Kraftstoff AG, GB 1 077 645, 1966; GB 1 125 077, 1967.
226. Union Rheinische Braunkohlen Kraftstoff AG, US 3 691 238, 1967.
227. Midland-Yorkshire Tar Distillers, GB 1 239 761, 1967.
228. K. K. Tiwari, P. N. Mukherjee, *Indian J. Technol.* **22** (1984) 295 – 300.
229. Consolidation Coal Co., GB 1 238 353, 1968.
230. Chiyoda Chemical Engineering Construction Co., JP 05 032 573, 1993 (S. Asaoka, K. Ikeda, Y. Shirato, K. Shimokawa, M. Shimura); *Chem. Abstr.* **119** (1993) 72 360.
231. Chiyoda Chemical Engineering Construction Co., JP 05 032 574, 1993 (S. Asaoka, K. Ikeda, Y. Shirato, K. Shimokawa, M. Shimura); *Chem. Abstr.* **119** (1993) 72 359.
232. Union Rheinische Braunkohlen Kraftstoff AG, EP 119 421, 1984.
233. Continental Oil Co., US 4 125 736, 1977.
234. Rütgerswerke AG, US 4 149 019, 1976.
235. L. A. Fury, jr., D. E. Pearson, *J. Org. Chem.* **30** (1965) 2301 – 2304.
236. A. Dierichs, R. Kubička: *Phenole und Basen,* Akademie Verlag, Berlin (GDR) 1958, pp. 381 – 385.
237. V. N. Vinogradova, G. D. Kharlampovich, *Khim. Promst. (Moscow)* 1975, 419 – 420.
238. Union Rheinische Braunkohlen Kraftstoff AG, US 4 283 571, 1979.
239. BASF, DE 3 926 290, 1989 (W. Steck, H. Lermer, M. Schwarzmann, T. Dockner).
240. UOP Inc., CA 2 051 949, 1993 (J. W. A. Sachtler, R. J. Lawson).
241. Gulf Research & Devel. Co., US 3 591 644, 1968.
242. Gulf Research & Devel. Co., US 3 591 645, 1968.
243. Gulf Research & Devel. Co., US 3 752 878, 1970.
244. W. W. Kaeding, R. O. Lindblom et al., *Ind. Eng. Chem. Process Des. Dev.* **4** (1965) 97 – 101.
245. DSM NV, US 5 210 331, 1993 (W. Buijs, L. H. B. Frijns, M. R. J. Offermanns).
246. N. L. D'yachenko, G. D. Kharlampovich, L. K. I'llina, *Neftekhimiya* **13** (1973) 870 – 875.
247. Imperial Chemical Ind. Ltd., GB 1 564 567, 1977.
248. Lummus Co., US 4 277 630, 1977.
249. Sumitomo Chemical Co., EP 52 839, 1981.
250. N. L. D'yachenko, G. D. Kharlampovich, L. K. I'llina, *Neftekhimiya* **12** (1972) 762 – 765.
251. M. P. Sharma, J. N. Chatterjea, *J. Chem. Tech. Biotechnol.* **33 A** (1983) 328 – 332.
252. G. D. Kharlampovich, N. L. D'yachenko, *Kinet. Katal.* **8** (1967) 208 – 210.
253. Rhône-Poulenc S.A., DE 2 164 004, DE 21 63 940, 1971.
254. Rhône-Poulenc S.A., DE 2 249 167, DE 2 256 513, 1972.
255. Mitsubishi Gas Chem. Co., DE 2 508 452, 1975.
256. Mitsubishi Gas Chem. Co., EP 74 162, 1982.
257. S. Fujiyama, T. Kasahara, *Hydrocarbon Process.* **57** (1978) no. 11, 147 – 176.
258. Y. Shimamura, H. Misawa et al., *Chem. Lett.* 1983, 1691 – 1694.
259. Mitsui Toatsu Chemicals, JP 07 247 233, 1995; JP 08 157 404, 1996 (F. Matsuda, K. Inoe, K. Kato); *Chem. Abstr.* **124** (1996) 86 574; **125** (1996) 167 558.
260. I. Yamanaka, K. Morimoto, M. Soma, K. Otsuka, *J. Mol. Catal. A* **133** (1998) 251 – 254.
261. M. Ropero, P. Patino, F. Armas, A. Karam, D. Iacocca, *Acta Cient. Venez.* **43** (1992) 280 – 283.
262. P. Patino, M. Ropero, D. Iacocca, *Plasma Chem. Plasma Process* **16** (1996) 563 – 575.
263. M. Tezuka, T. Yajima, *Plasma Chem. Plasma Process* **16** (1996) 327 – 340.
264. Mitsui Toatsu Chemicals, JP 06 072 919, 1994 (K. Ootsuka); *Chem. Abstr.* **121** (1994) 136 596.
265. Tosoh Corp., JP 05 004 935, 1993 (M. Hamada, Y. Sasaki, T. Mijake); *Chem. Abstr.* **119** (1993) 116 938.
266. G. A. Olah, A. P. Fung, T. Keumi, *J. Org. Chem.* **46** (1981) 4305 – 4306.
267. M. A. Brooks, R. Higgins et al., *J. Chem. Soc. Per-kin Trans. II* 1982, 687 – 692.
268. N. R. E. Radwan, M. M. Selim, *Afinidad* **54** (1997) 237 – 240; *Chem. Abstr.* **127** (1997) 50 344.
269. V. B. Athilakshmi, *Indian J. Chem. Sect. B* **37B** (1998) 439 – 442; *Chem. Abstr.* **129** (1998) 175 384.
270. M. K. Eberhardt, *J. Org. Chem.* **42** (1977) 832 – 835. K. Tomizawa, Y. Ogata, *J. Org. Chem.* **46** (1981) 2107 – 2109.
271. G. K. S. Prakash, N. Krass, Q. Wang, G. A. Olah, *Synlett* **1** (1991) 39 – 40.
272. Rhône-Poulenc S.A., DE 1 262 282, 1964.
273. S. Hashimoto, W. Koike, *Bull. Chem. Soc. Japan* **43** (1970) 293.

274. T. Yoneyama, R. H. Crabtree, *J. Mol. Catal. A* **108** (1996) 35 – 40.
275. H. Semba, M. Mukouyama, K. Sakano, *Appl. Microbiol. Biotechnol.* **46** (1996) 432 – 437.
276. Nippon Catalytic Chem. Ind., JP 09 023 893, 1997 (T. Semba, K. Sakano); *Chem. Abstr.* **126** (1997) 211 129.
277. *Biotechnol. News* **7** (1987) no. 13, 9.
278. Rhône-Poulenc Ind., US 4 301 307, 1979.
279. J. A. Vesely, L. Schmerling, *J. Org. Chem.* **35** (1970) 4028 – 4033.
280. Anic S.p.A., DE 31 35 556, 1981.
281. T. Yashima, Y. Kobayashi, T. Komatsu, S. Namba, *Stud. Surf. Sci. Catal.* **75** (1993) 1689 – 1692.
282. W. Yang, X. Liu, *Fenzi Cuihua* **12** (1998) 83 – 84; *Chem. Abstr.* **129** (1998) 95 288.
283. J. L. Motz, H. Heinichen, W. F. Hölderich, *J. Mol. Catal. A* **136** (1998) 175 – 184.
284. Ube Industries, JP 10 025 260, 1996 (T. Atoguchi, T. Sugimoto); *Chem. Abstr.* **128** (1998) 140 517.
285. D. J. Le Count, J. A. W. Reid, *J. Appl. Chem.* **18** (1968) 108 – 110.
286. K. E. Khcheyan, O. M. Revenko et al., *Nov. Metody Sint. Org. Soedin. Osn. Neftekhim. Syr'ya* 1982, 3 – 10; *Chem. Abstr.* **101** (1984) 74 734.
287. NI-Institut Sintetitscheskich Spirtow i Organitscheskich Produktow, US 3 830 853, 1971.
288. K. E. Khcheyan, O. M. Revenko, A. N. Shatalova, *11th World Petroleum Congress,* London 1983, Spec. Paper No. 16, pp. 1 – 7.
289. V. I. Zavorotow, G. L. Bitman, R. M. Gimronova, *Nov. Metody Sint. Org. Soedin. Osn. Neftekhim. Syr'ya* 1982, 11 – 19; *Chem. Abstr.* **101** (1984) 74 735.
290. G. N. Kiričenko, E. G. Mavljutova, T. M. Channon, *Neftekhimiya* **10** (1970) 231 – 235.
291. Sumitomo Chemical Co., JP 80 45 667, 1978; *Chem. Abstr.* **93** (1980) 204 270.
292. Ube Industries, DE 26 03 269, 1976.
293. Rhône-Poulenc S.A., DE 24 41 744, 1974.
294. Mitsubishi Gas Chem. Co., JP 80 51 031, 1978; *Chem. Abstr.* **93** (1980)185 946.
295. M. M. Groshan et al., *Dokl. Akad. Nauk. SSSR Ser. Khim.* **204** (1972) 872 – 878; *Doklady Chemistry (Engl. Transl.)* **204** (1972) 469 – 470.
296. Koppers Co., US 4 215 229, 1978.
297. H. G. Könnecke, K. Kränke, H. Languth, DD 60 047, 1966.
298. L. Cassar, M. Foà, G. P. Chiusolt, *Organomet. Chem. Synth.* **1** (1971) 302 – 304.
299. IG-Farbenindustrie, DE 694 781, 1938.
300. Mitsubishi Chem. Ind., JP 73 103 513, 1973; *Chem. Abstr.* **80** (1974) 70 532.
301. T. Cohen, A. G. Dietz, J. R. Miser, *J. Org. Chem.* **42** (1977) 2053 – 2058.
302. K. Pastalka et al., CS 184 670, 1976; *Chem. Abstr.* **94** (1981) 208 539. CS 193 993, 1977; *Chem. Abstr.* **98** (1983) 71 667.
303. H. Chang, M. Ni, *Fushun Shinyou Xueyuan Xuebao* **16** (1996) no. 4, 26 – 27; *Chem. Abstr.* **126** (1997) 89 104.
304. BASF AG, DE 14 93 492, 1965.
305. F. P. Daly, *J. Catal.* **61** (1980) 528 – 532.
306. M. Moritoki T. Fujikawa, *Process Technol. Proc.* 1984, vol. 2, Ind. Crist, 369 – 372.
307. Sumitomo Chemical Co., DE 2 530 859, 1975 (A. T. Tasaka, H. H. Hosaka, I. A. Dohgane).
308. L. Ma, C. Xu, S. Zhang, H. Ma, J. Pang, *Shiyou Huagong* **26** (1997) 117 – 119; *Chem. Abstr.* **126** (1997) 173 341.
309. M. Moritoki, K. Kitagawa, K. Onoe, K. Kaneko, *Process Technol. Proc.* 1984, vol. 2, Ind. Crist., 377 – 380.
310. M. Moritoko, *Curr. Jpn. Mater. Res.* **15** (1995) 213 – 228; *Chem. Abstr.* **124** (1996) 149 175.
311. Universal Oil Prod. Co., US 3 969 422, 1966.
312. Bayer AG, EP 552 633, 1993 (H. Buysch et al.).
313. Mitsui Petrochem. Ind., JP 82 59 824, 1980; *Chem. Abstr.* **97** (1982) 127 256.
314. Mitsui Petrochem. Ind., JP 80 57 529, 1978; *Chem. Abstr.* **93** (1980) 149 980.
315. Toray Ind. Inc., DE 3 327 146, 1982.
316. UOP Inc., EP 587 949, 1994 (H. A. Zinnen).
317. Pure Oil Co., US 3 031 383, 1959.
318. R. Rigamonti, G. Schiavino, *Chim. Ind. (Milan)* **36** (1954) 611 – 617.
319. M. M. Anwar, M. W. T. Pratt, M. Y. Shaeen, *Int. Solvent Extr. Conf.* [Proc.] vol. 2 (1980), Paper 80-64.
320. L. S. Nikolaeva, A. M. Evseev et al., *Russ. J. Phys. Chem. (Engl. Transl.)* **52** (1978) 1351 – 1353.
321. M. W. Pratt, J. Spokes, *CIM spec.* **vol. 21** (1979) 723 – 738.
322. G. A. Markus, G. I. Mischenko, A. N. Antonova, *Koks Khim.* 1974, no. 8, 37 – 41; *Chem. Abstr.* **82** (1975) 33 129.
323. S. R. Ningoo, V. G. Gaikar, *Sep. Technol.* **4** (1994) 249 – 251.
324. Koppers Co., US 4 267 389, 1979.
325. Koppers Co., US 4 267 390, 1979.
326. Koppers Co., US 4 267 391, 1979.
327. Koppers Co., US 4 267 392, 1979.
328. Koppers Co., US 4 394 526, 1980.

329. Bayer AG, DE 12 15 726, 1964.
330. Bayer AG, DE 19 626 309, 1996 (F-J. Mais, H. Fiege, K. Roehlk, S. Antons).
331. M-J. Lee, P-J. Lien, *Sep. Sci. Technol.* **30** (1995) 3697 – 3713.
332. Sumitomo Chem. Co., DE 2 136 700, 1971.
333. Gesellschaft für Teerverwertung GmbH, DE 1 124 046, 1958.
334. Gesellschaft für Teerverwertung GmbH, DE 1 118 797, 1960. Phillips Petroleum Co., US 4 491 677, 1983.
335. Mitsui Petroch. Ind./Kobe Steel, JP 77 31 039, 1975; *Chem. Abstr.* **87** (1977) 39 110.
336. Kogyo Gijutsuin, Japan; Ensuiko Sugar Refining, JP 06 263 669, 1994 (I. Uemansu, H. Takahashi); *Chem. Abstr.* **122** (1995) 105 414.
337. Mitsui Petrochem. Ind., JP 79 46 729, 1977; *Chem. Abstr.* **91** (1979) 123 551. JP 79 46 730, 1977; *Chem. Abstr.* **91** (1979) 74 334
338. VEB Leuna-Werke, DD 48 625, 1964.
339. S. N. Vyas, S. R. Patwardhan, M. M. Bhave, *Sep. Sci. Technology* **16** (1981) 377 – 384.
340. Y. Muramoto, H. Asakura, *Nippon Kagaku Kaishi* 1975, 672 – 677; *Chem. Abstr.* **83** (1975) 113 274.
341. Koppers Co., US 3 785 776, 1972.
342. W. Weinrich, *Ind. Eng. Chem.* **35** (1943) 264 – 272.
343. D. R. Stevens, *Ind. Eng. Chem.* **35** (1943) 665 – 660.
344. Bayer AG, US 4 144 400, 1976 (H. Fiege, J. Haydn, J. Renner, K. Wedemeyer).
345. Bayer AG, US 4 150 243, 1976 (D. Brueck et al.)
346. Bayer AG, US 4 113 975, 1977 (H. Fiege, J. Haydn, J. Renner, K. Wedemeyer).
347. BIOS 937.
348. Verona Chemical Co., US 1 502 849, 1922.
349. J. Anazawa, K. Uetake, *Kogyo Kagaku Zasshi* **63** (1960) 1963 – 1967.
350. Mitsui Toatsu Chemicals Inc., JP 0253 744, 1990; JP 0253 745, 1990 (T. Miura, K. Watanabe, H. Nakayama, M. Furuya, T. Nagata); *Chem. Abstr.* **113** (1990) 5 927, **113** (1990) 5928.
351. Standardization of Tar Products Test Committee (S.T.P.T.C), P. V. Watkins (ed.): *Standard methods for testing tar and its Products,* 6th ed., Gomersal, Checkheaton, Yorkshire, 1967.
352. C. Zerbe: *Mineralöle,* Part II, Springer Verlag, Berlin-Heidelberg 1969, pp. 494 – 497.
353. Berufsgenossenschaft der Chemischen Industrie; Merkblatt M 018: *Phenol, Kresole und Xylenole,* Heidelberg 1988.
354. Manufacturing Chemists Association, Chemical Safety Data Sheet SD-48: *Cresol,* Washington 1952.
355. ASTM D 3851-83:Sampling and Handling for Phenol and Cresylic Acid, 1983.
356. G. Hommel: *Handbuch der gefährlichen Güter,* Merkblätter **117** (Kresylsäure), **118** (o-Kresol), **118 a** (*m*-cresol), **118 b** (*p*-cresol), 1997; Merkblätter **986** (2,3- und 2,4-Xylenol), **987** (2,5- und 3,5-Xylenol), **988** (2,6- und 3,4-Xylenol), 1991; Ergänzungsband Transport- und Gefahrenklassen, 4th ed., Springer Verlag, Berlin – Heidelberg 1999.
357. A. P. Kudchadker, S. A. Kudchadker: *Key chemicals data books. Xylenols.* Thermodynamics Research Center, Texas A & M University, 1978.
358. S. A. Kudchadker, A. P. Kudchadker, R. C. Wilhoit, B. J. Zwolinski, *Hydrocarbon Process.* **58** (1979) no. 1, 169 – 171.
359. ESDU, Thermophysical properties of in dustrial important fluids on the saturation line: *Xylenols.* Item 83 005, Engineering Science Data Unit, London 1983.
360. T. E. Daubert, R. P. Danner: *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilations,* Hemisphere Publishing Corp., New York – Washington – Philadelphia – London 1989.
361. I. V. Goncharov, V. I. Kulachenko, *Neftekhimiya* **19** (1979) 255 – 258; **18** (1978) 816 – 821.
362. G. A. Makkus, *Koks Khim.* 1975, no. 9, 43 – 47.
363. Conoco Inc., US 4 228 311, 4 247 719, 4 249 026, 4 292 450, 1979.
364. ICI, GB 582 057, 1944. US 2 497 971, 1945.
365. Toray Industries Inc., JP 09 151 147, 1995 (M. Morimoto, A. Miyata, M. Ishikawa); *Chem. Abstr.* **127** (1997) 95 081.
366. Ado Kemuko K.K., JP 04 182 444, 1992 (M. Sekino, Y. Nakai, R. Shimizu); *Chem. Abstr.* **117** (1992) 153 220.
367. Koppers Co., US 4 447 658, 1982.
368. Toray Industries Inc., DE 3 327 146, 1983.
369. Chiyoda Corp., JP 03 020 234, 1989.
370. General Electric Co., US 4 001 341, 1964 (publ. 1974).
371. General Electric Co., US 4 547 596, 1985 (A. K. Mendiratta, J. J. Talley).
372. P. K. Pahari, M. M. Sharma, *Ind. Eng. Chem. Res.* **31** (1992) 2040 – 2042.
373. A. A. Pavlova, *J. Appl. Chem. USSR (Engl. Transl.)* **48** (1975) 1394 – 1396.

374. Consolidation Coal, US 3 331 755, 1964.
375. V. G. Gaikar, G. K. Sau, *Sep. Sci. Technol.* **3** (1996) 1979 – 1988.
376. Consolidation Coal, US 3 337 424, 1965.
377. Koppers Co., DE 19 24 768, 1969.
378. Mitsui Petrochem. Ind., EP 13 133, 1979.
379. Asahi Chem. Ind., DE 2 703 777, 1977.
380. UOP Inc., US 4 386 225, 1982.
381. A. Raychoudhuri, V. G. Gaikar, *Sep. Technol.* **5** (1995) 91 – 96.
382. Rütgerswerke AG, DE 2 531 774, 1975.
383. Asahi-Dow Ltd., JP 82 165 333, 1981; *Chem. Abstr.* **98** (1983) 88 963.
384. Nippon Steel Co., JP 83 52 235, 1981; *Chem. Abstr.* **99** (1983)22 105.
385. Kobe Steel/Mitsui Gas Chem. Co., JP 77 38 545, 1973; *Chem. Abstr.* **87** (1977) 24 943.
386. R. W. Maxwell, US 4 013 520, 1975.
387. CS 230 842, 1986 (J. Ciernik, E. Spousta); *Chem. Abstr.* **105** (1986) 152 704. CS 240 893, 1988 (J. Ciernik et al.); *Chem. Abstr.* **109** (1988) 210 689.
388. B. Chaudhuri, A. A. Padwardhan, M. M. Sharma, *Ind. Eng. Chem. Res.* **29** (1990) 1025 – 1031.
389. Continental Oil, US 3 996 297, 1975.
390. Conoco Inc., US 4 258 220, 1980.
391. Y. V. Churkin et al., *Neftepererab. Neftekhim. (Moscow)* 1974, no. 5, 32 – 33.
392. Kuraray Co., JP 75 71 635, 1973; *Chem. Abstr.* **83** (1975) 205 917.
393. Kuraray Co., JP 75 71 636, 1973; *Chem. Abstr.* **83** (1975) 205 916.
394. Honshu Chem. Ind. Co., JP 73 78 134, 1972; *Chem. Abstr.* **80** (1974) 59 671. JP 84 84 832, 1982; *Chem. Abstr.* **101** (1984) 191 329.
395. F. L. Wang, T. F. Tsai, *Catal. Today* **44** (1998) 259 – 269.
396. US 5 488 173, 1996 (F. L. Wang).
397. Shell Intern. Research, DE 25 29 773, 1975.
398. Shell Intern. Research, EP 80 759, 1982.
399. P. S. S. Prasad, B. D. Raju, K. S. R. Rao, G. S. Salvapathy, P. K. Rao, *J. Mol. Catal.* **78** (1993) no. 2, L19 – L22.
400. P. S. S. Prasad, N. Lingaiah, G. S. Salvapathy, P. K. Rao, *React. Kinet. Catal. Lett.* **59** (1996) no. 1, 151 – 157.
401. Rütgerswerke AG, GB 1 229 359, 1968.
402. Rütgers-VfT AG, EP 875 494, 1998 (W. Bergins, J. Talbiersky).
403. H-G. Frank, J. W. Stadelhofer: *Industrielle Aromatenchemie,* Springer Verlag, Berlin – Heidelberg – New York 1987, pp. 177 – 178.
404. Aristech Chemical Corp., CA 2 067 029, 1992 (W. R. Davie).
405. Shell Oil Co., US 4 086 282, 1976.
406. Shell Intern. Research, DE 2 804 114, 1978.
407. Atlantic Richfield Co., DE 2 316 576, 1973.
408. Rütgerswerke AG, DE 3 015 803, 1980.
409. Mitsubishi Gas Chemical Co., DE 2 521 324, 1974.
410. Mitsubishi Gas Chemical Co., DE 2 650 416, 1976.
411. E. V. Kirkland, O. P. Funderburk, F. T. Wadsworth, *J. Org. Chem.* **23** (1958) 1631 – 1635.
412. Sinclair Research Inc., US 3 352 927, 1964.
413. Bromine Compounds Ltd., US 5 118 877, 1992 (J. Oren, M. Zviely, J. Hermolin).
414. G. D. Kharlampovich, L. Z. Oblasova, N. A. Zhmakina, *Neftepererab. Neftekhim. (Moscow)* 1972, no. 10, 19 – 21.
415. Coalite and Chemical Products Ltd., DE 17 68 679, 1968.
416. Koppers Co., BE 844 866, 1975.
417. L. Z. Oblasova, V. V. Moskvich, G. D. Kharlampovich, A. I. Eliseeva, *Izv. Vyssh. Uchebn. Zaved., Khim. Khim. Tekhnol.* **16** (1973) 1047 – 1050.
418. J. Xiu, P. Ge, F. Chen, Y. Liu, *Jingxi Huagong* **15** (1998) no. 5, 47 – 49; *Chem. Abstr.* **130** (1999) 38 161.
419. G. D. Kharlampovich, L. Z. Oblasova, SU 577 202, 1975; *Chem. Abstr.* **88** (1978) 50 486.
420. V. S. Tarnavskii, I. G. Skorina, A. N. Morozow, *Khim. Tekhnol. Polikondens. Polim.* 1977, 69 – 76.
421. C. Zhang, Z. Liu, H. Chen, Y. Xie, *Huaxue Shiji* **17** (1995) 385 – 386; *Chem. Abstr.* **124** (1996) 316 632.
422. Mitsubishi Gas Chemical Co., JP 05 125 006, 1993 (A. Konishi, O. Kondo, Y. Kuko); *Chem. Abstr.* **119** (1993) 138 872.
423. General Electric Co., EP 686 617, 1995 (A. J. Caruso, J. L. Lee).
424. Mitsui Toatsu Chemicals, JP 07 165 649, JP 07 165 650, 1995 (H. Mizuta, M. Takaoka, T. Nagata); *Chem. Abstr.* **123** (1995) 339 384, **124** (1996) 55 550.
425. RWE Gesellschaft für Forschung und Entwicklung mbH, DE 4 223 979, 1992 (W. Löffler, B. Haas, J. Zimmermann).
426. Bayer AG, US 5 723 694, 1996 (A. Botta, H.-J. Buysch).
427. P. H. Given, *J. Appl. Chem.* **7** (1957) 172 – 193.
428. Union Rheinische Braunkohlen Kraftstoff AG, DE 1 956 383, 1969.

429. K. Kochloefl et al., CS 89 437, 1959.
430. J. Jelinek, CS 95 413, 1960; *Chem. Abstr.* **55** (1961) 7357.
431. I. Pigmen, E. Del Bel, M. B. Neuworth, *J. Am. Chem. Soc.* **76** (1954) 6169 – 6171.
432. V. I. Buraev, I. S. Isaev, V. A. Koptyug, *Zh. Org. Khim.* **15** (1979) 782 – 789.
433. Chiyoda Chemical Engineering Construction Co., JP 09 165 351, 1995 (M. Shimura, K. Ueda, S. Wakamatsu, S. Asaoka); *Chem. Abstr.* **127** (1997) 81 242.
434. P. Pitter, *Water Research* **10** (1976) 231 – 236.
435. W. E. Olive, jr., H. D. Cobb, R. M. Atherton, *Proc. Int. Biodegradation Symp. 3rd.* 1975 (Pub. 1976), 381 – 388.
436. C. T. Helmes, B. Levin, K. McCaleb et al.: *A study of industrial data on candidate chemicals for testing,* Report 1978, EPA 560/5-78-002: Cresols and cresylic acids, p. 3 – 63/3 – 128. Order No. PB-284 950, US Environmental Protection Agency, Washington D.C.
437. R. L. Aaberg, R. A. Peloquin, D. L. Strenge, P. L. Mellinger: *An aquatic pathways model to predict the fate of phenolic compounds,* Report 1983, PNL-4202 App. A-D.; Order No. DE 83 016 661, 1984; Chem. Abstr. 100 (1984) 39 246.
438. R. Koch: *Umweltchemikalien,* 3rd ed., VCH Verlagsgesellschaft mbH, Weinheim 1995, pp. 284 – 287.
439. F. Dietz, J. Traud, *GWF Gas Wasserfach: Wasser/Abwasser* **119** (1978) 318 – 325.
440. S. T. Hwang, *AIChE, Symp. Ser.* **77** (1981) no. 209, 304 – 315.
441. E. J. Keating, R. A. Brown, E. S. Greenberg, *Proc. Ind. Waste Conf. 1976* **33** (publ. 1979) 464 – 470.
442. Y. Zheng, D. O. Hill, C. H. Kuo, *J. Hazard. Mater.* **34** (1993) 245 – 260.
443. *Chem. Eng. (N.Y.)* **102** (1995) no. 4, 25, 27.
444. J. Gordon: *Air Pollution assessment of Cresols,* U.S. NTIS, PB Report 1976, PB-256 737, 66 pp.
445. R. L. Hervin, B. Froneberg, *Health hazard evaluation report* No. 80-020-1054, 1982, 32 pp.
446. DHEW (NIOSH) Publication (U.S.) vol. 78–133: *NIOSH Criteria for a Recommended Standard Occupational Exposure to Cresol.* 1978, 116 pp.
447. S. T. Haworth, K. Lawlor, W. Speck, E. Zeiger, *Environ. Mutagen.* **5** (1983) 3 – 142.
448. M. Cheng, A. D. Klingerman, *Mutat. Res.* **137** (1984) 51 – 55.
449. Berufsgenossenschaft der chemischen Industrie (ed.): Toxikologische Bewertung Nr. **137**: 2,4-Xylenol; Toxikologische Bewertung Nr. **138**: 2,6-Xylenol; Toxikologische Bewertung Nr. **139**: 3,5-Xylenol; Ausgaben 12/95, Heidelberg 1995.
450. D. D. Dietz: "Toxicology studies of cresols", Natl. Toxicol. Program, Research Triangle Park, NC, USA; Report (1992), NTP-TOX-9, NIH/PUB-92-3128; Order No. PB 92-174242, 131 pp.
451. "Toxicological profile for cresols: *o*-cresol; *p*-cresol; *m*-cresol". Clement International Corp., Ruston, USA; Report (1992), ATSDR/TP-91/11; Order No. PB 93-110732, 171 pp.
452. "Registry of Toxic Effects of Chemical Substances (RTECS)", National Institute for Occupational Safety and Health (NIOSH), Cincinnati, USA, 1991.
453. N. I. Sax: *Dangerous Properties of Industrial Materials,* 8th ed., Van Nostrand Reinhold, New York 1992, pp. 959 – 961 (Cresols), pp. 3523 – 3524 (Xylenols).
454. C. L. Yaws, S. D. Sheth, M. Han: *Handbook of Chemical Compound Data Process Safety,* Gulf Publishing, Housten TX, 1997, pp. 136 – 144.

Crop Protection

The following are separate articles: Acaricides; Fungicides, Agricultural; Insect Control; Molluscicides; Nematocides; Plant Growth Regulators; Rodenticides; Weed Control

Hans-Hermann Cramer, Bayer AG, Leverkusen Federal Republic of Germany

1. Agricultural Aspects 1
1.1. Ecological Conditions 1
1.2. Economic Conditions 3
2. World Food Supply 4
3. Crop Protection in Modern Agriculture 5
4. Economic Threshold and Prognosis 9
5. Methods of Crop Protection 10
5.1. Physical and Mechanical Methods . 10
5.2. Chemical Methods 11
5.2.1. Goals 11
5.2.2. Requirements 11
5.2.3. Effects on Farming 13
5.3. Biological Methods 15
5.4. Cultivation 17
5.5. New Methods 18
5.5.1. Biotechnology 18
5.5.2. Research Methods 19
5.5.3. New Indications 19
5.5.4. Integrated Systems 19
6. Safety Aspects 19
6.1. Safety for the User 19
6.2. Safety for the Consumer 20
7. Legal Aspects 21
8. References 22

1. Agricultural Aspects

Crop protection, one of the elements of agricultural production, is an integral feature of complex ecological and economic systems. The goals of crop protection, from the aspects of farm economics and ecology, are to prevent preharvest losses, i.e., to safeguard crop yields; assure quality; replace labor-intensive methods; and contribute to the development of new agricultural production systems in spite of constraints of climate or soil. Crop protection is an essential element of integrated crop production. It plays an important role in environmentally adapted and technologically advanced farming systems [1]. By their very nature, crop protection and nearly all other agricultural disciplines are mutually dependent. This applies to the natural sciences, including ecology, and equally well to the economic and technical aspects of agronomy.

1.1. Ecological Conditions

Agriculture is always characterized by deliberate intervention in ecological systems. Its goal is to replace existing plant societies with crops giving optimum yields. This is possible only if farming techniques are adapted to environmental conditions, ecological influences opposing the agricultural goal are restrained, and ecological conditions themselves are controlled with reference to the agricultural purpose. Agriculture can therefore be seen as ecological management intended to achieve economically optimum yields.

Soil. The most important factor in agricultural production is the soil. A central concern of soil utilization is the long-term maintenance and, if possible, increase of fertility.

Historically, agriculture provided at first for satisfactory yields by cultivation of locations favored by nature. Soil nutrients, soil structure, and terrain were decisive in the separation of land into farming and nonfarming areas. When it was experienced that soil was impoverished by the removal of nutrients through crop harvesting, an attempt was made to compensate by employing the fallow, crop rotation, fertilization with stable manure, and variation in farming systems. Nevertheless, yields in Europe during the 19th century would have sharply deteriorated through soil exploitation if ecochemistry had not achieved decisive breakthroughs. The work of Liebig (1803 – 1873) demonstrated that fertility, balance of soil nutrients, and removal of nutrients by harvesting are all closely related. This knowledge has become the ecological foundation of husbandry. Well-balanced mineral fer-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_061

tilization has since become an established element in good agricultural practice. In addition, it has contributed to ecological equalization: in earlier times, light soils (with a high percentage of sand) were considered less fertile; heavy soils (with a high percentage of loam) were sometimes viewed as difficult although fertile. The addition of nutrients made it possible to compensate for these inequalities and even poorer soils have become suited to the cultivation of wheat.

Fertilization and Cultivation. With regard to crop protection, ecological changes arising from variations of intensity in land use have had important consequences. Increases in yield made possible by the application of mineral fertilizers have generally increased \011 the production of biomass; this biomass usually remains in the field, except for the harvest. Therefore, the proportion of organic substances in the soil, resulting in humus formation, has steadily increased. At the same time, this makes it possible to utilize high-yielding varieties. In addition, quality may be improved, e.g., the baking characteristics of wheat or the brewing quality of barley. Improved soil productivity also makes it possible to manage much denser stands of crops and eliminate the fallow period. Crop rotation is therefore organized less with respect to yield than in accordance with economic and ecological considerations.

All these developments have a direct or indirect effect on plant protection. Only seldom is it possible through plant breeding to combine high yield potential and high resistance to plant diseases in the same variety, and many high-yielding varieties show increased susceptibility to disease. A higher crop density, for example in cereals, leads to changes in the microclimate, which is mostly wetter and somewhat warmer than in a thinner stand. Therefore, the conditions of propagation and infection become more favorable for fungal pathogenic agents, with more serious infestation. On the other hand, densely planted crops compete more vigorously with weeds, and late weed development is suppressed. The elimination of fallow periods and the simplification of the crop rotation cycle, especially with cereals, correspondingly increases the danger of infestation by harmful organisms, such as cereal aphids, cereal powdery mildew, nematodes, and others.

Overall it may seem that the ecological changes which have developed through modifications in agricultural practices, i.e., by increased intensity of soil utilization, have increased the problems of plant protection. However, these changes are not generally related to worsening of damages. In Europe, for example, some very ancient and serious problems have lost practically all significance. Prominent examples are the plagues of locusts, which were very important until the beginning of the 19th century but have now disappeared [2]. The loss in significance of grain ergot (*Claviceps purpurea*) should also be mentioned here, although this has now reappeared with the development of organic farming.

In many cases, a greater demand for plant protection has resulted directly from increased acreage. This applies, for example, to soybeans, of which an average of 28×10^6 ha were grown worldwide during the years 1961 to 1965; in 1984, this acreage, at 52×10^6 ha, had almost doubled.

Migration of Pest Organisms. The transmission of harmful organisms from one continent to another is one of the ecological phenomena that were formerly less significant but have become more important with ever shorter transportation times accompanying the development of international trade. The ecological aspect of these phenomena results from the fact that species introduced into new areas are not confronted with their natural biotic factors of limitation. Consequently their propagation potential can develop fully in their new home, being subject only to climatic conditions. Figuratively speaking, the ecosystem lacks an "immune reaction" in the newly infested areas. Examples of the introduction of harmful organisms with disastrous economic consequences are: the Hessian fly (*Mayetiola destructor*) brought to America from Europe; the migration of the cotton boll weevil (*Anthonomus grandis*) from Mexico to the southern states in the United States; the introduction of the European pine sawfly (*Neodiprion sertifer*) and the gypsy moth (*Lymantria dispar*) into North America from Central Europe; the grape vine louse (*Phylloxera vitifoliae*), the Colorado potato beetle (*Leptinotarsa decemlin-*

eata), and the San-José scale (*Quadraspidiotus perniciosus*) brought to Europe from America; the coffee rust (*Hemileia vastatrix*), introduced into South America from Africa; and many others [3]. In all these cases, basic ecological changes occurred, with direct economic effects on agricultural enterprises.

1.2. Economic Conditions

Agriculture produces about 95 % of the foodstuffs for the world population. The remaining 5 % is provided by fishing, hunting, and gathering. Tillage or cultivation of permanent crops occupies 11 % of the earth's land surface, meadow and pasture land 23 %. In sum about one-third of the land surface above sea level is utilized for agriculture. Employment of the working population of the world in agriculture is about 43 % [4]. These average figures do not provide a complete picture. In the developed countries, only 7.1 % of the working population is engaged in agriculture (e.g., Japan, 8.6 %; France, 7.1 %; Federal Republic of Germany, 3.1 %; and United States, 1.8 %). In the developing countries, the figure is 55.8 % (e.g., Niger, 85.2 %; India, 60.2 %; Egypt, 48.8 %; and Brazil; 35.1 %). In the developed countries, on the average, 7.1 persons work 100 ha of tillable land; in the developing countries, the figure is 68.5. The developed countries produce 35 % of the cereal yield on 23 % of the cereal acreage in the world; the developing countries produce 29 % of the yield on 44 % of the acreage; the centrally planned economies produce 36 % on 33 % of the acreage. The traditional classification of countries as agrarian and industrialized is therefore outdated. It is rather the case that the industrialized countries produce agrarian surpluses with a small labor force whereas the agrarian countries, with a large labor force and large acreage, often do not produce enough to meet their own needs. The reason is that the developed countries have succeeded in rapidly transforming agricultural science into practice. In Germany, for example, this resulted in wheat yields that, between 1800 and 1910, increased on the average from 1.0 t/ha to 2.0 t/ha; by 1950, this figure was 3.0 t/ha; and by 1984, it had reached 6.3 t/ha [2], [4]. These increases were due to the use of mineral fertilizers and plant breeding, and improvements in infrastructure, tillage, mechanization, and plant protection [5]. These influences are all interdependent and the importance of each cannot be measured separately from the others.

By definition, plant protective measures are limited in their effect to the prevention of loss and, therefore, can increase yields only indirectly. This has great economic significance for the farmer, for national import – export planning, and for world trade policies.

Crop Losses. For the individual farmer, crop losses usually occur when fixed costs (buildings, machines, debt service) and variable costs (seed, fertilizer, labor) have already been met; the cost/profit ratio deteriorates accordingly. With a profit margin of 10 %, a crop loss of 5 % reduces the net profit by half. In addition, losses of quality may be crucial (see Section 5.2.3).

Within national economic systems, crop damage may have considerable importance, especially where cash crops are a vital source of foreign exchange. For example, the collapse of the coffee industry in the last century in Ceylon (Sri Lanka) caused by coffee rust (*Hemileia vastatrix*) had far-reaching consequences. In Ghana, Nigeria, and the Ivory Coast Republic during the 1950s, a combination of a virus infection (swollen shoot disease) and a fungal infestation (*Phytophthora palmivora*) caused by cacao bugs (*Miridae*) destroyed 220 000 t of cocoa annualy, more than half of a good harvest. A similar catastrophe, caused by the cotton boll weevil (*Anthonomus grandis*), overtook the cotton industry of the southern United States around the turn of the century. The consecutive epidemics of grapevine powdery mildew (*Uncinula necator*) and grapevine downy mildew (*Plasmopara viticola*), together with the introduction of the grape vine louse (*Phylloxera vitifoliae*), were disastrous for European viniculture in the 19th century. In all such cases, in which whole sectors of the economy are dependent on a single crop, its sudden loss has far-reaching consequences. Destruction of basic foodstuffs by swarms of locusts repeatedly caused famine in medieval Europe [6]. In India, a rice fungus had similar results [5]. More recent examples are the failure of the potato crop in Ireland in 1845 – 1849 and in Germany in 1916 – 1917, caused by the potato blight (*Phytophthora infestans*) [7], [8].

More significant, however, than such catastrophes are the regular hidden losses caused by

insect pests, plant diseases, and competition by weeds. On the average, these amount to a good third of the potential harvest of the 60 crops of international importance [5], [9], constituting thereby a prime factor in the world economy and at the same time an enormous food potential. Even a partial prevention of these losses would substantially extend the generally meager world nutritional base. In comparison to more ambitious projects, this would also lead to a low cost increase in food production, and lower prices. With respect to the limited purchasing power of many developing countries, this would be an important factor.

In addition, harvest fluctuations caused by harmful organisms complicate budgetary calculations.

A good example of international consequences is seen in the variation in the import requirements of the Soviet Union and the corresponding effect on the U.S. agrarian economy.

2. World Food Supply

The world food supply is very unevenly distributed. Agricultural overproduction in some developed countries is in contrast to the deficit in many developing countries that lack the buying power and infrastructure that would make a better distribution of the surplus possible [10]. The problem of world nutrition of course transcends distribution: even with a completely equal distribution of the foodstuffs currently produced, there would be a deficit of 10 – 18 %. Even with good harvests, world grain reserves seldom exceed a one-month supply (measured by market capacity); since the early 1950s reserves have often fallen below a ten-day supply. Seen in this light, world food reserves are marginal. It is of course true that the great flexibility of American agriculture permits production to be increased considerably and quickly to meet shortages in regions with ample purchasing power. On the other hand, world supplies are heavily dependent on a few countries. Export surpluses of grain have so far been produced only by the United States, Canada, Australia, and Argentina. A single harvest failure in the United States caused for example by an epidemic of fungus, would create a world nutritional crisis.

Production. Overproduction is of prime political and economic interest in some developed countries. In its historic context, this development is extraordinary. For the first time in its history, Europe has an oversupply rather than a shortage of food. Since ancient times, the search for agriculturally productive regions has had decisive political consequences, including, for example, the extension of the Roman Empire to North Africa, the migration of Germanic tribes during the 4th century A.D., and the eastward expansion of the German Empire since the Middle Ages. The same is true of past colonial policies, such as those of the French in Algeria, Tunisia, and Morocco. Following the discovery of America, the various waves of westward emigration from Europe were driven to a large extent by shortages of food and land.

Until long after 1945, Europe was dependent on agricultural imports. In the European Community, the current overproduction in some commodities (wheat, sugar, milk products) and the near self-sufficiency in others (cereals other than wheat, potatoes, vegetables, and meat) represent a historical novelty. Problems of supply have been replaced by those of commercial utilization. This situation is aggravated by the serious agrostructural problems of small farms for which it is difficult to find political and social answers. Part of the overproduction results from the importation of fodder such as soybean products. If the production of foodstuffs in the Federal Republic of Germany that depends on foreign fooder is not taken into account, the degree of self-sufficiency in 1983/84 would be reduced from 95 % to 78 % [11].

However, abundance in the industrialized countries tends to stabilize prices, especially those of foodstuffs, which are thus made more available even to countries with limited purchasing power. Producer price indices for U.S. crops rose from 100 to 146 points between 1977 and 1984; at the same time, agricultural production costs rose from 100 to 155 points. In the Federal Republic of Germany, the comparable index for crops rose from 100 in 1977 to 108 in 1984 and for production costs from 100 to 148 [4]. This price pressure forces farmers to increase efficiency further.

Food production has risen considerably, as seen in the following figures in which the aver-

age production of 1974 – 1976 is taken as 100 [4]:

	1972	*1984*
Developed countries	92.1	113.7
Developing countries	90.1	133.0
World	91.3	122.7

As a result of a sharp growth in world population, however, the increase in per capita food supply is smaller (1974 – 1976 = 100):

	1972	*1984*
Developed countries	94.6	106.4
Developing countries	96.5	111.8
World	96.6	104.9

Regional data vary considerably as shown by the following figures for per capita supply of agricultural products for the developing countries of the respective regions (1974 – 1976 = 100):

	1972	*1983*
Africa	102.3	88.7
Latin America	96.7	103.4
Near East	98.7	94.2
Far East	94.3	111.9
Others	99.3	100.6

In contrast to these unfavorable developments in the Near East and particularly in Africa, progress has clearly been made in Southeast Asia and Latin America. This is the consequence of the so-called *green revolution*, with improved yields resulting primarily from successes in the breeding of wheat, rice, and maize. The introduction of more productive varieties better adapted to local conditions has led within a few years to increases in yields that had remained unchanged for centuries; results in some cases were sensational. Initial setbacks showed that the full realization of the genetic potential of the new varieties required new techniques, often including fertilization and flanking plant protection. These methods have been successful in many countries and led to solutions of acute problems of malnutrition. The examples of India, Mexico, and Thailand are given in Table 1 [4].

Clearly, it is difficult to keep agricultural production ahead of a sharply rising population. Nevertheless, the situation can be improved, especially in the developing countries, by modernization of agriculture, principally the choice of variety (plant breeding), fertilization, and plant protection [12].

3. Crop Protection in Modern Agriculture

Crop protection includes all measures of crop maintenance, from sowing to harvest and even to processing and consumption. It extends to all kinds of damage, whether abiotic or biotic. As in human and animal medicine, prevention and hygiene are crucial. Early protective measures are the province of husbandry, which merges by easy stages with plant protection. For example, tillage and the selection of crops for rotation also serve to control weeds, nematodes, and other pests. The choice of varieties and plant nutrition can have considerable influence on the physical condition of crops. Crop protection is not generally thought to include seed purification to prevent weeds, but does include seed dressing for the prevention of fungal infections, as well as quarantine measures against pests or pathogenic organisms.

Plant protection ordinarily applies to entire crops rather than individual plants and is therefore analogous to population medicine. This is particularly true of biological control (see Section 5.3), whereas physical and chemical measures may sometimes be applied to individual plants as well as whole stands of crops.

Crop protection measures are generally classified according to causes of damage or the harmful organisms. Chemical agents are characterized as insecticides, acaricides (against mites = acari), fungicides, herbicides (against weeds and wild grasses), nematocides, rodenticides, and others. Chemical measures have predominant importance for the control of biotic damage. In 1985, world consumption of crop protection products was estimated at $\$ 13\,788 \times 10^6$ (users' prices) of which $\$ 4268 \times 10^6$ represented insecticides, $\$ 2537 \times 10^6$ fungicides, $\$ 6331 \times 10^6$ herbicides, and $\$ 642 \times 10^6$ all others [13]. Therefore, insecticides constitute 31 %, fungicides 18 %, herbicides 46 %, and other products 5 % of consumption. This distribution is subject to regional variation according to climate, crops planted, and the stage of agricultural development. In general, herbicides dominate in the industrialized countries with high labor costs. In 1985 in the United States, the proportion of herbicide consumption was 69 %, insecticides 23 %, and fungicides 8 %,

Table 1. Food production in India, Mexico, and Thailand [4]

	India	Mexico	Thailand
Cereal production, 10^6 t			
1974 – 76	119	16	17
1984	169	26	24
Increase of per capita cereal production *, %	19	29	11
Increase of food production *, %	38	34	25
Increase of per capita food production *, %	14	4	15

* Based on 1974 – 1976 data

whereas in western Europe, the share of herbicides accounted for 47 %, insecticides 17 %, and fungicides 36 %. In 1985 the United States represented 32 % of the world market in crop protection products, Western Europe 22 %, and Eastern Asia including Japan 19 %.

The world pesticide market is shown in Table 2, sales value and distribution of uses in Table 3, and use patterns in Table 4.

Fruits, Vegetables, and Horticulture. Fungicides and insecticides have the greatest importance in the intensively grown crops of fruit, citrus, grapes, vegetables, and garden plants, but the use of herbicides is limited in that case. In fruit cultivation, ground cover is often maintained to conserve soil moisture, with mulching by regular cutting. In growing vegetables and ornamental plants with their numerous varieties weeds are controlled mechanically to avoid problems of varietal compatibility with herbicides. High-quality produce must be free of damage. Although these foodstuffs spoil easily, relatively high returns can be achieved per unit area with special crops allowing a high investment in crop protection. Moreover, in such permanent crops as fruit, citrus, and grapes, characteristic complexes of pathogenic organisms have become established, including scab, powdery and downy mildew, and other fungus diseases. The insect pests include several species of moths (*Tortricidae*), weevils (*Curculionidae*), and others. The soft epidermis, considered desirable in the edible parts of many vegetables, reduces resistance to disease.

Cereals. A completely different pattern is shown with crops grown on a large scale. In maize and the other cereals (wheat, barley, rye, oats), weed control has become largely a domain of agrochemistry. With increasing costs of mechanical weed control, the application of herbicides has become standard practice. Selective products that control weeds in cereals without harming the crop have been developed. In maize, herbicides account for 71 % of all crop protection measures. Insecticides are, however, still important in the control of soil pests, which in many areas result partly from the cultivation of maize without crop rotation. With the traditional cereals, in which chemical weed control figures prominently, insects are of minor significance. In certain regions it may be necessary to control cereal bugs (*Eurygaster* sp.) and occasionally locusts.

In many countries, particularly in Europe, seed dressing is accepted agricultural practice and has almost eliminated many cereal diseases which had devastating effects in earlier times. Smut and bunt diseases have been suppressed. Other cereal diseases have gained in significance with the intensification of crop rotation and an increase in the density of planting. Most important among these are rusts, mildew, and some *Helminthosporium* and *Septoria* species that have been controlled with seed dressings and leaf sprays. Overall, diseases have become more prominent.

Rice. Rice represents special problems arising from farming conditions and climate. It is the classic example of single-crop farming. For the most part, it is cultivated on the same acreage over long periods of time without interruption through crop rotation. In many areas, rice cultivation offers the ecological advantages of regulation of water economy and prevention of soil erosion by terracing. As in every case of single-

Table 2. World pesticide market (1985), %

Crop	Insecticides	Fungicides	Herbicides	Total
Fruits, vegetables, and horticulture	27.4	48.6	7.7	22.0
Corn (maize *)	10.0	4.6	22.0	15.0
Soybeans	3.0	1.1	21.3	11.4
Rice	14.8	11.9	8.0	11.0
Cotton	22.7	0.9	4.5	9.7
Wheat	2.5	13.8	12.5	9.5
Other grains	1.1	5.9	6.6	4.7
Sugar beets	2.3	1.9	4.5	3.3
Sugar cane	1.0	1.0	2.3	1.6
Tobacco	2.5	1.5	0.5	1.3
Groundnuts	1.5	2.2	0.7	1.3
Others	11.2	6.6	9.4	9.2
Total	100.0	100.0	100.0	100.0

* With a global acreage of ca. 125×10^6 ha, maize occupies about 9 % of all arable land.

Table 3. Sales value of pesticides and distribution of uses by crop (1985)

Pesticide	Crop	10^6 $	%
Herbicides	corn	1 395	10.6
Herbicides	soybeans	1 348	10.3
Fungicides	fruits and vegetables	1 232	9.4
Insecticides	fruits and vegetables	1 168	8.9
Insecticides	cotton	969	7.4
Herbicides	wheat	791	6.0
Insecticides	rice	633	4.8
Herbicides	rice	507	3.9
Herbicides	fruits and vegetables	487	3.7
Insecticides	corn	456	3.5
Herbicides	other grains	417	3.2
Fungicides	wheat	351	2.7
Fungicides	rice	302	2.3
Herbicides	cotton	288	2.2
Herbicides	sugar beets	287	2.2
Subtotal (15 indications)		10 631	81.1
All pesticides (other indications)		2 478	18.9
Total		13 109	100.0

Table 4. Use patterns of insecticides, fungicides, and herbicides in different crops, %

Crop	Insecticides	Fungicides	Herbicides
Fruits, vegetables, and horticulture	40	43	17
Corn	23	6	71
Soybeans	8	2	90
Rice	44	21	35
Cotton	76	2	22
Wheat	9	28	63
Other grains	8	24	68
Sugar beets	23	11	66
Sugar cane	19	12	69
Tobacco	61	22	17
Groundnuts	39	33	28
Others	37	13	50
All crops	33	19	48

crop farming, however, rice growing has favored specific pathogenic organisms adapted to the wet and warm irrigation environment. Among the harmful insects are stem borers, plant hoppers, aphids, and many beetle species.

Diseases and weeds also occur. Therefore rice cultivation employs all three classes of crop protective agents, insecticidal, fungicidal, and herbicidal, as is the case with hardly any other crop except peanuts. The rice farming acreage of the earth, at 145×10^6 ha, constitutes 10 % of the arable land area in the world.

Soybeans. Soybeans represent a typical cash crop. Herbicides account for 90 % of the crop protection cost. Large-area cultivation of soybeans would be unmanageable by mechanical weed control.

Cotton. During the entire history of cotton cultivation, insect attack has been a limiting factor [5]. The development of synthetic organic insecticides, chiefly phosphoric esters, has assured the economic viability of this crop in many climatically favorable areas [14].

Although weed control accounts for 22 % of crop protection costs, insecticides account for 76 % and cotton remains the typical insecticide-based crop. All green parts of the plant constitute ideal insect fodder.

Sugar Beet. Sugar beet cultivation is an instructive example of the almost fully mechanized management of an agricultural crop species [2]. The sugar beet yield in the developed countries was raised from an average of 35.5 t/ha in 1965 to about 50 t/ha in 1985. This enormous increase was accomplished with a drastic reduction in labor costs. In 1950, 130 h/ha were still required for tending beet; in 1973 this had been reduced to 30 h, and by 1985 to only 22 h [15].

For several reasons the European sugar beet cultivation faced a serious crisis in the 1950s. These problems have been solved by a succession of innovations. The result is a highly mechanized beet cultivation, less manual labor, higher yields, and more efficiency on acreage controlled by the farmer [16–18].

Following the development of herbicides that selectively eliminate most weeds without harming the beet crop, it turned out that the importance of hoeing had been overestimated in many cases. As with other crops, it eliminates weeds and accomplishes little more [19].

The elimination of manual weed control resulted in a labor saving of 30 – 40 %. However, the separation of beet seedlings continued to require a high degree of manual labor. Ecologically it began to be evident that polyphagous pests, which had formerly been distributed among weeds and crops, now concentrated on the host beets in the fields. This applies particularly to soil pests such as wireworms and cutworms but also to leaf and shoot pests such as pygmy beetles, weevils, diptera, springtails, and aphids. Without doubt, dense vegetation plays a part in limiting pest dispersion.

Hoeing was no longer economical for the separation of seedlings which had to be reduced from ca. 1 million to 70 000 – 80 000 beet plants per hectare. Therefore, plant breeders developed monogermous seeds first by mechanical and later by genetic methods. Consequently, only one seedling developed from each individual grain. Since it had become possible to equalize the number of seed with the final count of beet, mechanical or manual hoeing was no longer necessary. The beet field had therefore become a "blank slate," i.e., during the entire vegetation period, with perfect management, only the beet plants to be harvested remained.

The result, however, was the aggravation of some pest problems. When the number of beets planted correspond to the final count, the failure of each individual plant means an economic loss. The economic threshold is therefore much lower. The beet is the only plant remaining in the entire area as a host for phytophagous insects, and species that formerly had little significance have developed into serious pests.

The resulting problems have been solved partly by coating the seed with insecticides and partly by the application of insecticides to the emerging plants.

4. Economic Threshold and Prognosis

Control measures are necessary when damage is expected to exceed the *economic threshold*, understood as the degree of damage at which control costs are covered by the expected increase

in yield [20]. If control costs are higher than the expected profit, the damage is tolerated for economic and ecological reasons.

Table 5. Critical infestations

Pathogenic organism	Host	Conditions
European corn borer (*Ostrinia nubilalis*)	maize	4 – 8 eggs per 100 plants
Wire worms (*Elateridae*)	maize	1 wireworm per m^2 of soil
Foot rot in cereals (*Pseudocercosporella herpotrichoides*)	wheat	20 % of plants with brownish discoloration at stalk base at onset of sprouting
Colorado potato beetle (*Leptinotarsa decemlineata*)	potatoes	15 larvae per plant
Cabbage gall midge (*Dasyneura brassicae*)	rape	1 midge per plant
Summer fruit tortrix moth (*Adoxophyes orana*)	stone fruit	preflowering, 7 caterpillars per blossom cluster; postflowering, 1 – 3 caterpillars per 100 fruit
Wind-grass (silky bent grass, *Apera spica-venti*)	cereals	10 – 20 plants per m^2
Slender foxtail (black grass, *Alopecurus myosuroides*)	cereals	20 – 30 plants per m^2
Cleavers (*Galium aparine*)	cereals	0.5 plants per m^2

The economic threshold is determined by epidemiological and economic factors. The population density of pathogenic organisms depends on the propagation potential, weather conditions, resistance of the host plants, parasite infestation, soil condition, dispersion of infestation, and other factors. The economic components of this threshold are determined by the control costs, farm prices, yields expected, and the costs which have already been met. Because of the variability of these factors, the economic threshold fluctuates within a wide range. Some influences, such as weather, are notably unpredictable. A more uncertain prognosis requires a wider economic safety margin [21].

As shown in Figure 1, a regular epidemiological development often simplifies the prognosis [20], [22], [23]. In example A, the population density does not normally reach the economic threshold, and savings can be realized in control costs. In example B the pest multiplies under favorable conditions and, uncontrolled, would exceed the economic threshold. The most difficult decision is shown in example C, where the average population density of the pest is just below the economic threshold. In this case control measures became unnecessary because the pest population is reduced by parasites and predatory insects. In example D, the pest population is consistently above the economic threshold, and repeated control measures are indispensable. The decisions of the experienced farmer as to the use of control measures are based on threshold data and on careful observation of the situation under field conditions. Critical infestation figures are given in Table 5 [24].

In the attempt to improve the accuracy of prognosis, procedures have been worked out for apple scab (*Venturia inaequalis*), based on the ability of spores to germinate as a function of temperature and moisture (Fig. 2).

Automatic equipment for recording leaf moisture establishes a data base for apple scab prophylaxis. The prognosis of attack may be based on the number of insects caught in traps (light traps, pheromone traps, and mechanical traps).

Pathogenic organisms in a new area constitute a special case, depending on climatic conditions and on the presence of natural enemies and of suitable hosts. A prognosis requires the help of system analysis.

5. Methods of Crop Protection

5.1. Physical and Mechanical Methods

Mechanical and physical techniques are among the oldest methods of crop protection. Tillage measures such as hoeing, harrowing, and rolling of the soil, prepare the seed bed and help to control weeds. Terrestrial animals are excluded by fences, ditches, and traps. In Central Europe, for example, fencing is required to prevent deer from damaging forest plantations. Ditches protect conifers against pine weevils, cereals against cereal carabides or the broad-nosed weevil (*Otiorrhynchus*), and beets against root weevil (*Bothynoderes*). Mechanical traps are used against such insects as mole crickets (*Gryllotalpa* sp.) and also against mammals such as field mice and the moles (*Talpa europaea*). In some regions, field mice are controlled by temporary flooding, which has also been applied against grubs and other soil pests. Tangle glue protects fruit trees against the nonflying female

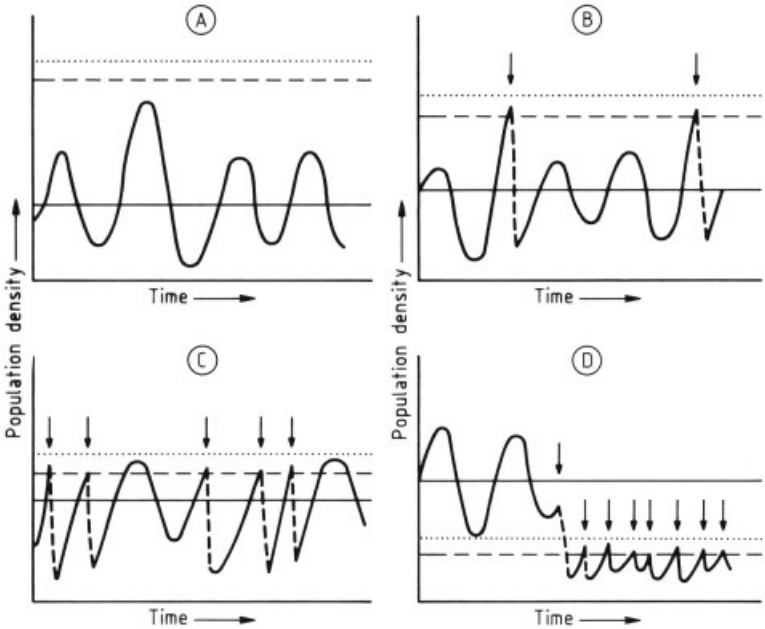

Figure 1. Relation between epidemiological type, economic threshold, and need of control [20], [22], [23]
A) No economic damage, e.g., rice leaf-miner fly (*Agromyza nigrella*)
B) Occasional economic damage, e.g., beet fly (*Pegomyia hyoscyami*)
C) Frequent economic damage, e.g., coddling moth (*Cydia pomonella*)
D) Continuous economic damage, e.g., plant lice as vectors in potato breeding
—— General equilibrium of population
– – – – – Control threshold
· · · · · · · Economic threshold
↓ Control required

of the common winter moth (*Cheimatobia brumata*).

Mechanical measures have even been tried to inhibit the propagation of fungus diseases. A long-term effort has been made to control stem rust in cereals (*Puccinia graminis*) by carefully rooting out the European barberry (*Berberis vulgaris*), the intermediate host on which the aecidia of the fungus develop. More recently the elimination of infected cotoneaster plants has given hope of delaying the introduction of fire blight caused by bacteria (*Erwinia amylovora*) on Rosaceae, including kernel and stone fruit. The trimming of the branches of fruit trees controls fructification and eliminates primary infections such as apple mildew (*Podosphaera leucotricha*).

Among thermal procedures, soil steaming is effective against soil pests, particularly in greenhouses. Cereal smut diseases (*Ustilago* sp.) can be controlled by treating the seed with hot water. The outbreak of some virus diseases can be delayed by heat. Stubble left after harvest is burned to reduce the chances of infestation, but with ecological side effects that may be serious. In some areas it is forbidden by law to burn stubble or weeds.

Crop protection includes refrigeration, e.g., the cold storage of fruit. This delays, but rarely prevents the development of pathogenic agents.

5.2. Chemical Methods

Chemical procedures play a dominant role in crop protection. The reason include their convenient adaptability to farming operations and the wide range of agents available for dealing with specific problems in accordance with legal regulations. In addition, it is fairly easy to calculate risk : utility ratios.

5.2.1. Goals

The goals of optimizing farming methods include (1) increase in yields, (2) improvement of the cost : profit ratio, and (3) accommodation to market conditions, or (4) advances in ecology. Chemical crop protection can make substantial contributions to all these areas [25].

When these goals conflict, it becomes necessary to set priorities and possibly to make compromises, always with a view to long-term cultivation efficiency.

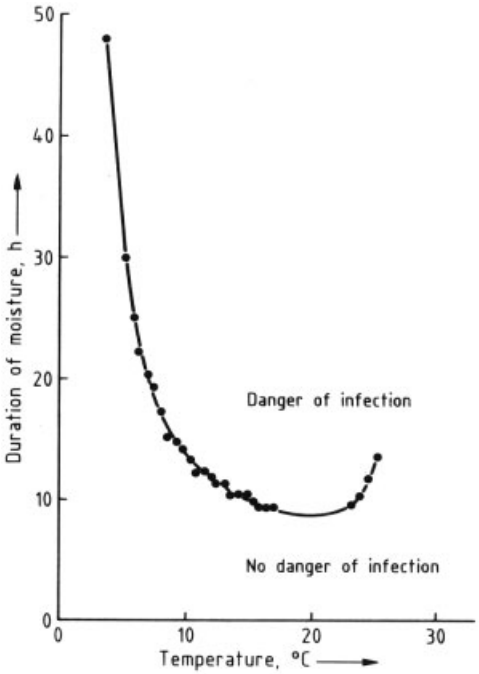

Figure 2. Chances of infection by apple scab (*Venturia inaequalis*) in relation to moisture and temperature [24]

5.2.2. Requirements

The requirements that modern crop protection products must meet include selective action, low persistence, high efficacy, and no intervention with natural equilibrium.

Selective Action. The selectivity requirement for herbicides on the one hand, and insecticides and fungicides on the other, are fundamentally different: a selective herbicide should have the greatest possible effect on all plant species in the field, with the exception of the crop itself. Selective insecticides or fungicides, on the other hand, should be effective against only one species, the target organism, with no effect on any other species in the ecosystem. These requirements encounter the following difficulties:

1) Insecticide compounds may be *monotoxic*, i.e., acting on a specific species, *oligotoxic,* i.e., group-specific, or *polytoxic,* i.e., with a broad spectrum of effect. For ecological reasons higher specificity is generally preferred since a minimal disturbance of the ecosystem, including the beneficial fauna, is expected of monotoxic action. Such specific effects occasionally lead to shifts in the dominance of species because competitors of the eliminated species have utilized the new "ecological niche," thereby taking the place of the target organism.
   An example is the Central European viniculture, where the grape berry moth species *Clysia ambiguella* and *Polychrosis botrana* were first controlled successfully. However, then a species of beetle, the hazal leaf roller, *Bytiscus* sp., became a significant vinicultural pest.
   Herbicides, may also lead to a shift in the dominance of the species within weed societies. The specific action of the growth-hormone herbicides derived from β-indolylacetic acid on dicotyledonous weeds has, in many cases, led to the multiplication of monocotyledonous grasses in cereals, e.g., wind grass (*Apera spica-venti*) or slender foxtail (*Alopecurus myosuroides*).
   Resistance can develop to crop protection products with a specific action. Within the genetic variation of a population chemical intervention selects the resistant individuals, and facilitate their propagation by eliminating the rest of the species; resistance thus builds up. A narrow spectrum of action promotes such developments because the restriction of the effect to specific biochemical processes means that minor deviations from these models cause a loss of efficacy. The great genetic variability of the fungi causing plant diseases imparts considerable significance to this phenomenon.
2) Products with a specific action tend to be more expensive than those with a broad spectrum of effect. For such products research costs have risen disproportionately and the

probability of discovering new pesticides drops with each additional requirement [26]. On the other hand, thorough testing is necessary for the biological validation of selective effects. In addition, the market for products with specific action is naturally smaller than for broad spectrum products.

3) The great number of products available, each with a narrow spectrum of effect, requires careful diagnosis and selection. The demand for higher specificity has resulted in a broad assortment of specialized products.

Low Persistence. The ideal behavior for a crop protection compound is to persist in the environment long enough to accomplish its purpose and then to be degraded rapidly to naturally occurring substances. Evidently, it is not always possible to satisfy both requirements. If the fungal spores are dispersed over a long time or the hatching of insects is delayed, repeated treatment may be necessary. This can be avoided with compounds of higher persistence.

High Efficacy. A modern crop protection product should be applied at the lowest possible dosage of active ingredient per unit of area,that is at high efficacy. This requirement is largely met by new products which are applied in the following dosages [27]:

Active ingredient	*Dosage, g/ha*
Insecticide	
lead arsenate	2000 – 5000
DDT	1000 – 2000
diflubenzuron	125 – 250
deltamethrin	10 – 25
Fungicide	
sulfur	2000 – 5000
dithiocarbamate	1500 – 2000
metalaxyl	125 – 200
triadimefon	50 – 125
Herbicide	
sodium arsenite	1500 – 3000
metribuzin	250 – 500
chlorsulfuron	15 – 20

In all three areas of application, dosages of less than 100 g/ha have been reached, sometimes with a more than 100-fold reduction of dosages of older compounds. Ecological risk is reduced correspondingly. The uniform distribution of such small quantities requires careful formulation, whether as wettable powder, emulsifiable concentrate, granule, or dust. Products of such high efficacy also require careful dosage and even application.

Interference with Natural Equilibrium. The naturel equilibrium should not be disturbed by crop protection, although agriculture itself constitutes intervention into natural processes (see Section 1.1). In view of the complex character of ecosystems, the effects of individual compounds are difficult to assess [28].

5.2.3. Effects on Farming

Crop protection contributes to farming success by safeguarding yields threatened by natural causes. The methods employed range from soil preparation to the protection of the harvest itself and extend to transport and storage. Most importantly, these methods must meet the criterion of cost efficiency.

Soil. Soil treatment protects seed, seedlings, and developing root systems from animal pests, fungi, other microorganisms, and weeds. Active compounds tend to degrade rapidly in soil, although photochemical decomposition is slow. The apparently conflicting goals of persistence and rapid deactivation have been achieved by the development of slow-release granules, from which the active ingredient is gradually liberated by the action of soil moisture. After release, degradation is rapid. Such granules are used against soil insects in maize, beets, grapes, and vegetables.

Soil fungicides are not normally used with crops planted on a large scale, but are restricted to special crops such as ornamentals or hops. Increasing problems with nematodes have made the use of nematocides necessary, especially in crops managed without sufficient rotation. Soil herbicides are used in all developed countries.

Farming Studies. Unique long-term studies on the effect of crop protection on yields have been conducted on two experimental farms in the Rhineland with favorable growth conditions and a normal weed flora (see Tables 6 and 7).

At Höfchen the soil is loam or loess loam, at Laacherhof brown and similar earth. The weed flora in untreated areas is normal on both farms, and the yield reaction to weed control is not excessive. The yield increases in untreated soil,

Table 6. Increase in cereal yield after chemical weed control; long-term averages obtained at experimental farms Höfchen* (H) and Laacherhof** (L), Federal Republic of Germany [29]

Crop	Average of years	Increase, %	
Winter wheat	1966/7 – 1981/2	10	(H)
		10	(L)
Winter barley	1967/8 – 1981/2	14	(H)
		15	(L)
Winter rye	1966/7 – 1981/2	11	(H)
		9	(L)
Winter triticale	1974/5 – 1981/2	17	(H)
Spring wheat	1969 – 1982	7	(H)
soft	1969 – 1982	5	(L)
hard	1969 – 1982	8	(H)
Oats	1976 – 1982	6	(H)
Spring barley	1975 – 1982	8	(H)
	1975 – 1982	2	(L)
Maize	1967 – 1982	19	(H + L)
mechanical control	1967 – 1982	8	

* 205 m above sea level; long-term mean of annual precipitation, 925 mm.
** 40 – 45 m above sea level; long-term mean of annual precipitation, 710 mm.

Table 7. Increase in yield of noncereal crops after chemical weed control; long-term averages obtained at experimental farms Höfchen (H) and Laacherhof (L), Federal Republic of Germany [29]

Crop	Average of years	Product	Compared to the type of control	Increase, %	
Sugar beets	1967 – 1982	roots	mechanical	21	(L)
		sugar	mechanical	24	
Fodderbeets	1970 – 1982	roots	mechanical	20	(H)
		leaf	mechanical	17	(L)
Potatoes	1971 – 1982	tubers	mechanical	24	(H)
			mechanical	9	(L)
			untreated	87	(H)
			untreated	75	(L)
Early potatoes	1970 – 1982	tubers	mechanical	7	(L)
			untreated	50	
Rape	1981 – 1982	seed	untreated	9	(H)
Soybeans	1978 – 1982	seed	untreated	31	(L)
Broad beans	1971 – 1982	beans	mechanical	3	(L)
			untreated	61	
Peas	1971 – 1982	peas	mechanical	13	(L)
			untreated	113	
Tomatoes	1971 – 1982	fruits	untreated	14	(L)
Onions	1970 – 1982	onions	untreated	209	(L)
			mechanical	11	
Strawberries	1970 – 1982	berries	mechanical	75	(H)

calculated from the mean yield surplus of up to 16 years of investigation, may be considered largely representative. For the traditional cereals (Table 6), they show a yield increase from 9 to 17 % in the winter varieties; for all winter cereal species taken together, the result is a mean increase in yield of 12 – 13 % resulting solely from the use of herbicides applied mostly before emergency as in normal practice. There are no basic differences between these two farm sites.

In the summer cereals, with shorter growth periods, the yield effects are less clear, averaging 6 %.

In maize, the 16-year comparison shows that yield more than doubled with herbicides. As hoeing is indispensable if herbicides are not used, the comparison includes mechanical weed control, with results similar to other summer cereals.

The yield per hectare for winter wheat in Höfchen was 5.1 t, increased by treatment to 5.6 t. If a producer price of $ 160/t is assumed,

Table 8. Yield increase by insect control in rice* [9]

	Degree of infestation			
	Low	Medium	High	Mean
Yield, t/ha				
Untreated	3.47	3.42	2.54	3.31
Treated	4.27	4.46	3.90	4.33
Increase	0.80	1.04	1.36	1.03
Increase, %	23	30	53	31

* 843 trials.

Table 9. Relation between yield increase and efficacy of insecticide in rice cultivation [9]

Number of trials	Efficacy, %	Yield increase, t/ha	Yield increase, %
22	1 – 20	0.72	22
58	21 – 40	0.73	22
118	41 – 60	0.89	27
237	61 – 80	0.92	28
123	81 – 90	1.22	37
285	91 – 100	1.20	36
Total 843		Mean 1.03	31

the gross income is increased by $ 90/ha. If single preemergence treatment costs are $ 48 – 60, net income is increased by $ 28 – 40/ha.

The actual gain is, however, higher because weeds in cereals retain moisture and increase drying costs. They also create mechanical difficulties, particularly with the combine harvester if the weeds mat in the cutter bar. The threshing operation also requires thoroughly dried and ripened kernels.

In the case of tubers, seeds, vegetables, and berries, yield comparison between herbicide-treated and -untreated crops is not realistic because these crops cannot be grown economically without mechanical weed control. In Table 7 chemical and mechanical weed control are compared. The yield increases caused by herbicides are of economic importance because of the higher monetary value of most of these crops, and the higher labor costs of mechanical procedures.

The advantage of chemical methods is largely due to the absence of root damage caused by mechanical procedures, and possibly to loss of moisture caused by breaking up the soil.

Comparable trials have not been conducted on insect control, although an evaluation of insecticides in *rice* cultivation has been carried out in Asia and Colombia (see Table 8).

The yield also increases with the efficacy of the insecticide used; however, this effect levels off at about 80 % efficacy (Table 9).

Insect control improves not only yield but also crop quality and storage stability. Freedom from infestation increases the value of the crop.

Similar results are obtained by treating diseases in *wheat* with fungicides [9]. On the average, the yield in the treated plots was 5.83 t/ha compared to 5.29 t/ha in untreated plots, an increase of 0.54 t/ha or 9.3 %. The results of treating mildew in wheat are given in Table 10 [9].

As with rice insect pests, the yield is lower at high infestation even in treated plots. When the efficacy of the fungicide increases, the yield is also increase (Table 11).

Highest efficacies (91 – 100 %) produce a mean yield increase of 0.76 t/ha. In contrast to insect control, there is a considerable yield difference (almost 0.2 t/ha) between the efficacies of 81 – 90 % and those of 91 – 100 %.

As with other crop protection measures, the increased yield is accompanied by improved quality.

Sugar beet yields, for example, are higher in all respects with chemical than with mechanical control (see Table 12).

In addition to yields, chemical crop protection improves crop production systems. To

Table 10. Yield increase by mildew control in wheat* [9]

	Degree of infestation		
	Low	Medium	High
Yield, t/ha			
Untreated	5.61	5.69	4.82
Treated	6.02	6.18	5.54
Increase	0.41	0.49	0.72
Increase, %	7	9	15

* 1914 trials.

this end, the increasing significance of no- or minimum-tillage systems offers an interesting example [1]:

In many parts of the world, traditional tillage systems are being replaced by mechanical and chemical innovations. The steel moldboard plow, introduced in 1830, is stilltoday the basic tool for preparing the seedbed. However, it is energy-intensive and, under certain conditions, suffers from serious disadvantages. It exposes the soil to wind and water erosion and may disrupt the soil – water balance. Plowing can lead to hard plow pans and preserve weed seeds for later germination.

Table 11. Relation between yield increase and efficacy of fungicide in wheat cultivation

Efficacy, %	Yield increase, t/ha
1 – 20	0.3
21 – 40	0.48
41 – 60	0.5
61 – 80	0.53
81 – 90	0.56
91 – 100	0.76

In no-tillage systems, also referred to as zero-tillage, or in chemical tillage systems, crops are planted in unprepared soil by opening a narrow trench just wide and deep enough to provide seed coverage. There is no other soil cultivation. Weeds are controlled by herbicides, and soil pests by insecticides and nematocides. The advantages include better soil moisture retention and reduction of rainfall run-off, water and wind erosion, soil damage by machinery, energy consumption, and production costs.

No-tillage with the application of herbicides, insecticides, and nematocides is a valuable option for regions with serious problems, such as drought, erosion, unstable soils, or unfavorable climatic conditions.

Reduction of Losses. A long-term study at the experimental farm of Laacherhof, Federal Republic of Germany, shows that inadequate crop protection measures reduce apple yields (see Table 13). Similar results have been obtained with other crops [1].

5.3. Biological Methods

Biological pest control is understood as "the use of living organisms (through human intervention) to limit the populations of certain harmful animals or plants" [32]. *Biotechnical procedures* "utilize natural reactions of pests to certain physical or chemical stimuli which achieve an objective other than that associated with their natural significance for the organism"

Table 12. Yield increase of sugar beet with herbicidal over mechanical control* [30]

	Increase, %				
	No. of plants	Beet yield	Sugar concentration	Sugar	
Method				t/ha**	%
Presowing	37	24	2	9.16	26
Preemergence	39	31	5	9.98	37
Postemergence	27	16	2	8.56	18

* 1974 – 1983.
** Yield with mechanical control methods, 7.27 t/ha.

Table 13. Apple yield at Laacherhof, average 1967 – 1981 [31]

Treatment	Yield, t/ha		Yield, %	
	Total	Marketable	Total	Marketable
Complete	24.4	19.3	100	79
No insecticide	17.2	7.6	70	31
No fungicide	14.9	6.6	61	27
None	13.8	4.6	57	19

[32]. Examples include the use of repellents and attractants; pheromones; growth, molting, and development hormones; and optical, acustic, olfactory, and tactile stimuli. The term biological crop protection is often understood to include biotechnical procedures (→ Biological Control).

Weed and Fungus Control. In Chapter 3, it was mentioned that, for chemical crop protection products, 46 % of the world market represents herbicides and 18 % fungicides. In the developed countries, two-thirds of all crop protection expenditures are for herbicides. Attempts to control weeds biologically have employed monophages or oligophages, i.e., insects adapted to certain host plants; phytopathogenic fungi have also been employed. The following problems have arisen:

In actual farming, weed control almost always applies to weed societies rather than individual species, and extends to all species competing with the crop.

In principle, specific competitors can be found to control individual weed species, but the control of weed societies requires competitors for each species; otherwise, ecological niches would open up, to be occupied by the remaining weed species. Polyphagous antagonists cannot be used because they are pronee to attack the crop also.

At present, no practical procedures are available for the biological control of pathogenic fungi; antagonists are being investigated, which may act via metabolic products or competitive behavior. Thus far, insect control remains the principal area for the application of biological procedures.

Insect Control. Biological control procedures have not been established for large-scale cash crops [1], [33]. The annual harvesting of cereals, soybeans, beets, and potatoes has hindered the establishment of a balance between pest and competitor, and in crop rotation has changed biozoenotic conditions fundamentally. Almost half of all successful projects were conducted on islands where dispersion was restricted. Control of pests imported into new areas has been highly successful. In addition, biological procedures are favored whenever the economic threshold is high, e.g., in a forest.

A good example of biological control of insects is provided by parasites and predators against a pest infesting a permanent island crop with a high economic threshold. In the absence of other pests, chemical measures are unnecessary.

Microbiological procedures have also been used for pest control. Uncontrolled propagation of microorganisms with possible unforeseeable ecological effects must be avoided.

5.4. Cultivation

Cultivation techniques can strongly influence the susceptibility ofcrops to infestation. These methods are based on the choice of crop and variety, a sound crop rotation scheme, and tillage intensity. This should give the optimum sowing schedule and technique and the best fertilizing and tending methods.

Choice of Site. Site conditions are determined by climate and soil; tillage under the best conditions is of prime importance. Plant breeding has developed varieties suited to areas beyond the original habitat and differing from it with respect to temperature, precipitation, hours of daylight, and other factors. By avoiding unfavorable site conditions, e.g., those prone to certain infestations, the farmer can improve the crop and reduce the probability of damage. This is important in the propagation of seed or plants. Seed potatoes, for example, are propagated in so-called healthy locations, in which climate does not permit the multiplication of aphids, which are vectors of virus diseases.

Crop Rotation. Crop rotation is among the oldest methods for maintaining soil fertility and preventing damage. In Central Europe, the three-field operation had already come into existence by the 8th century. In this system, the cereals cultivation, if possible with two different species, was followed by a fallow year which regenerated the soils and restrained the growth of weeds and pests. This system was maintained for thousand years; it was superseded only in the 18th and 19th centuries by the improved three-field economy [34], [35], in which the fallow was replaced by red clover. In addition, the introduction of potatoes and sugar beets in Central Europe permitted the inclusion of dicotyledons in the crop rotation practice. This represented considerable progress, with advantages that extended to weed control. Crop rotation usually included two-third cereals and one-third dicotyledons, i.e., clover, potatoes, and beets; 9 – 11 field rotations were practiced [34], [35]. Crop rotation of many individual components is, however, not practicable for economic reasons, but oversimplified crop rotation entails various risks.

Soil Treatment. Soil treatment and fertilization are closely related to the probability of damage. Certain nutritional deficiencies often predispose to infestation. On the other hand, an excess of nutrients, especially nitrogen, may favor other forms of infestation, such as aphids or cereal fungi.

Scheduling. The planting schedule can be so arranged as to avoid late frosts or other critical periods. In Scotland, late planting reduced the population of the cabbage root fly (*Phorbia brassicae*). After the first spring generation had laid its eggs, the initial infestation was depressed during the vegetation period. However, part of the growing season had to be sacrificed [36].

Mechanical Methods. Sowing depth and the seedbed quality influence germination and emergence of the crop; on the other hand, deep plowing and turning of the soil bury the weed seeds, which may preserve them in the ground until the next plowing.

The land surrounding the cultivated area may be developed. Hedges and borders influence local climatic conditions and provide a habitat to birds of prey, insect-eating songbirds, hedgehogs, and predatory and parasitic insects. On the other hand, vegetation may serve as an intermediate host for diseases and pests. The advantages usually tend to outweigh the disadvantages, especially where crop protection measures are employed.

5.5. New Methods

New tendencies in crop protection extend to biotechnology, modern methods of organic synthesis, electronic data processing, and integrated systems in farming [10], [37].

5.5.1. Biotechnology

Biotechnology can be utilized in microbiological, genetic, and diagnostic methods [38]. Both microorganisms and their products of metabolism can be used as pathogens or pathogen antagonists. Metabolism products obtained from microorganisms include antibiotics, such as blasticidin, which was developed against rice blast (*Pyricularia oryzae*). Kasugamycin, validamycin, and polyoxin are related. The use of antibiotics in crop protection has been restricted by the fear of creating resistance, which might even extend to human pathogens.

Antibiotics. A nucleotide antibiotic with an effect like that of an acaricide has been developed under the name niccomycin; high costs, however, have kept it off the market [39], [40]. On the other hand, avermectin, obtained from *Streptomyces avermitilis,* is an effective insecticide and acaricide in such low dosages that it has aroused great interest [38]. The herbicides bialaphos and phosphinotricin are derived from other *Streptomyces* phyla. These are all clearly defined biogenic substances, which can be produced synthetically or biologically, according to costs.

Microorganisms. Microorganisms are applied directly to control phytopathogenic fungi, animal pests, or weeds. Spore preparations are obtained from *Verticillium lecanii*, a fungus that affects aphids and white flies (*Aleurodinae*). The fungus *Trichoderma harzianum* has an antagonistic effect on the phytopathogenic fungus *Stereum purpureum* and is used for sealing

wounds in fruit trees. *Colletotrichum gloeosporiorides* f. sp. *aeschynomene* is used as a spore preparation in soybean and rice against the weed *Aeschynome virginica.* The action is highly specific and is usually supplemented by chemical herbicides [38].

Genetic Engineering. Genetic engineering may provide crop protection by increasing the pathogenicity or antagonistic action of a microorganism. More direct is the genetic alteration of the crop itself; in this case, the line between cultivation and crop protection may become difficult to draw. The important objectives here include the improvement of photosynthetic efficiency and nutrient assimilation and of resistance to drought, disease, and chemical agents. A special case is the creation of herbicide resistance through genetic engineering. Its goal is maximum selectivity and compatibility with crop species. This should reduce the amounts of herbicides used, if postemergence as well as preemergence application becomes possible.

An example of genetic engineering is the herbicide tolerance induced in rape. The utilization of rape as an oleaginous and protein plant was limited by the high content of erucic acids and glucosinolates. Traditional plant breeding created rape varieties free of erucic acids and glucosinolates. This, however, did not guarantee crops of a pure variety because when rape seed is combined, many seeds are shed and remain in the field, retaining viability for 6 – 8 years. Later cultivation mixes the old varieties with the new ones which are free of erucic acids and glucosinolates. This could be prevented if the new varieties were tolerant of a herbicide that destroyed seeds of the old variety in the soil. The "weed rape" could then be eliminated selectively. This practice has reached an advanced stage of development.

Diagnosis. Biotechnology can also be applied to diagnosis, especially in the early stage of a disease, by the use of so-called DNA-probes. A mixture of the DNA of the fungus and of the host plant is prepared by dissolving the DNA in individual filaments, e.g., from plants infested by the fungus. The DNA segments of the fungus, which have previously been cloned in bacteria are added to this mixture. These segments seek out the complementary DNA segments of the fungus. The result is a hybridization which shows that the host plant has been infected. Such "diagnosis kits" are already used in medicine and could advance the early recognition of plant diseases [38].

5.5.2. Research Methods

Chemicals for crop protection are developed by empirical research, research by analogy, and research based an structure – activity relations. The traditional empirical method employs more or less random screening of a large number of compounds to find those that combine efficacy with minimal side effects. Although the statistical chances of success are small (12 000 : 1) the probability of discovering totally new classes of active substances is relatively high.

Research by analogy proceeds from classes of compounds with known biological activity, in the hope of improving the properties by structural changes. This strategy offers a high probability of gradual progress but a smaller probability of scoring a breakthrough. Nevertheless, such important properties as systemic efficacy have been detected by analogy research.

With increasing knowledge of biochemistry and molecular biology research into constitutional effects is receiving more emphasis. The "lock – key principle" is used to bring the active substance to the specific site; compounds are "tailor-made" to accomplish calculated physiological effects [41].

5.5.3. New Indications

New farming ventures will require new functions and new approaches in crop protection. The growing of new types of crops in arid regions or in regions with high salt concentrations will create new problems of crop protection. The same applies to the perspective of "energy farming."

The development of new or modified crops for this purpose, in regions not traditionally accessible to agriculture, will produce new kinds of problems.

5.5.4. Integrated Systems

Crop protection problems cannot be viewed in isolation. The range of possibilities, from mon-

itoring and prognosis to physical, mechanical, chemical, biological, tillage, and biotechnological methods, suggests looking for practical combinations and their optimization. *Integrated crop protection* is defined as "a system in which all justifiable economic, ecological, and toxicological methods, combined with natural factors of limitation, are employed to keep infestations of injurious organisms below the economic threshold. Special emphasis is given to the exploitation of the natural factors of limitation."

Integrated crop production is defined as "a system in which all cultivation methods are coordinated with the soil and the climatic conditions of the site. The goal is a high yield of good quality over a period of many years, with full consideration of ecological and economic effects" [35]. Such methods create an increasing need for products to control pests, diseases, or weeds from the time of sowing until harvest [10]. The development of integrated systems is facilitated by the use of electronic data processing.

6. Safety Aspects

Undesirable side effects of crop protection agents must be kept to a minimum to guarantee safety for the user, the consumer of the agricultural products, and the environment. Safety precautions in the production, transport, and storage of these products are equally necessary.

6.1. Safety for the User

The chance of exposure to crop protection products is greatest during application, and questions of acute and chronic toxicity are of particular concern to the user.

Toxicity data are required by all countries that impose legal restrictions on crop protection, i.e., over 90 % of all the countries in the world. These requirements include regular medical examinations of workers in regular contact with these products. Short-term, chronic, and cumulative effects must be determined as well as carcinogenic, teratogenic, and mutagenic effects.

The toxicological data are reflected in the recommendations for use that accompany the commercial product, and that include provisions forbidding use while eating, drinking, or smoking, or applying the product in a high wind.

Breathing masks and protective clothing are recommended, and adherence to proper dosages and disposal procedures are required, as well as thorough washing after use.

Safety instructions extend to the labelling and closure of the package, and first aid instructions in cases of poisoning. The equipment must be in good condition and provide precise metering of dosages; performance must be checked regularly.

6.2. Safety for the Consumer

Consumer safety requires the absence of residues of crop protection products that might be injurious to health. This rule must be strictly observed with food.

Physical loss and chemical decomposition decrease the concentration of the pesticide more or less rapidly, depending on its stability and the conditions. The concentration is then measured against toxicological standards established by feeding experiments on at least two animal species. The upper limit of tolerance, defined as the amount that creates no pathological changes, is expressed in mg/kg of animal weight and called the *no effect level* (NEL). In extending the results of animal experiments to humans, an additional safety margin is employed according to international custom, usually by establishing one percent of the lowest NEL. The value established by this method is the *acceptable daily intake* (ADI), expressed in mg/kg:

$$\frac{\text{NEL}}{\text{safety margin}} = \text{ADI}$$

This is the base value for the determination of tolerances or maximum limits. These are the highest acceptable residue concentrations of crop protection products established by legislation. In order to establish tolerance, the ADI value is multiplied by the body weight (BW). However, to comply with the contribution of individual foodstuff to the total diet, the product is then divided by a nutritional factor to obtain the toxicologically *permissible level:*

$$\frac{\text{ADI} \times \text{BW}}{\text{nutritional factor}} = \text{permissible level}$$

Accordingly, staple food contributes to the permissible level to a larger extent than specialty

foods. The permissible level assumes a fixed, life-long intake of the residue. For the legislator, this constitutes the upper limit for the establishment of tolerance. The maintenance of the tolerance values in farming practice is still an unsolved problem, because not every lot of produce can be analyzed for residues.

From the known rate of degradation a waiting period or period of nonavailability is established, during which the residue concentration falls below the permissible level. In practice, a prolongation of this waiting period is established if the period between the last application of the pesticide and the harvest can be chosen to be longer than the period of nonavailability. Extending the waiting period then offers additional safety. At the time of harvest, the residue value may have fallen still further below the permissible level. The tolerance values established accordingly are designated as tolerances of *good agricultural practice.* In the Federal Republic of Germany, for example, they constitute the basis for the tolerances established in the "Höchstmengenverordnung Pflanzenschutz" (crop protection regulation governing maximum residue limits).

Differences in safety requirements in different countries have complicated international trade in foodstuffs. Furthermore, crop protection practices and legislation differ from country to country. For example, the growing of fruit in warmer climates requires more insecticides whereas cooler and wetter climates require more fungicide.

In many countries official agencies carry out so-called market basket analyses. A typical selection of foodstuffs is bought, representing the weekly purchases of an average family, and subjected to residue analysis. Thousands of such analyses have given surprisingly similar results: About 80 % of the samples were found free of residues, about 18 % showed residues within legal tolerances, and in 1 – 2 %, tolerances had been exceeded [42]. Considering that the established tolerance values represent legal values rather than toxicological threshold values, these results must be seen as remarkably favorable. They speak well for the farmers who use crop protection products [42].

Furthermore, the situation has been improving from year to year, because more and more highly persistent products are being removed from the market. Examples of test results are given in Table 14.

Residues, when found, are usually far below the tolerance value. For example, in the Federal Republic of Germany from 1977 to 1980, the residues of organic chlorocompounds in milk were 0.2 – 5.0 % of the permissible ADI value [21].

7. Legal Aspects

The varied aspects of crop protection products (sale, handling, residues in foodstuffs and environment, consumer protection) are controlled in many countries by a system of legal regulations. The laws deal with chemicals (e.g., DDT), foodstuffs, toxic substances, and water conservation, as well as with the protection of crops and environment and methods of proper testing of pesticides [43–45].

Legislation is composed of the following elements: protection of crops; protection of the user, the consumer of foodstuffs, and the environment; and registration of crop protection products, i.e., automatic prohibition of all products not registered. Registration usually requires data on metabolism in plants and animals, on leaching action in the soil, and on toxicity to two mammal species, two bird species, and water and soil organisms.

These principles constiute the basis of the Japanese *Agricultural Chemicals Regulation Law* of 1948, as amended by several laws between 1949 and 1984 [46]. The following provision is made:

"The purpose of this law is to conduce to stabilization of agricultural production; protection for health of people; to contribute to environmental conservation of the life of the people, intended for normalization of the quality of agricultural chemicals with insurance of their safe and correct use by establishing a registration system and exercising control over sale and the use concerning agricultural chemicals." (Article 1). "Any manufacturer or importer of agricultural chemicals shall not sell the same unless he has registration granted from the Minister of Agriculture, Forestry, and Fisheries, of each agricultural chemical which he has manufactured, processed, or imported." (Article 2, § 1)

Parallel U.S. regulations are found in *The Federal Insecticide, Fungicide and Rodenticide*

Table 14. Residues in fruits and vegetables in the Federal Republic of Germany [21]

	Samples with no detectable residue, %		Samples with residue within tolerance, %		Samples with residue above tolerance, %	
	1978	1982	1978	1982	1978	1982
Fruits						
domestic	54	74	46	25	0.0	0.3
foreign	58	71	40	28	1.9	1.4
Vegetables						
domestic	83	89	17	10	0.0	0.9
foreign	48	63	50	34	3.5	2.7

Act (FIFRA), of 21 Oct. 1972, as amended in 1975 and 1978 [47]:

"Requirement: Except as otherwise provided by this Act, no person in any State may distribute, sell, offer for sale, hold for sale, ship, deliver for shipment, or receive and (having so received) deliver or offer to deliver, to any person any pesticide which is not registered with the administrator."

The German *Crop Protection Law of 10 May 1968* was modified by the *Third Bill for the Amendment to the Crop Protection Statute* [48], [49]. Its purpose is:

1) "to protect crops from injurious organisms and diseases (crop protection),
2) to protect harvests against injurious organisms (protection of stores),
3) to influence the life processes of plants through the use of substances which are not intended for the nutrition of plants, and
4) to prevent harm which might occur in the use of crop treatment products or other measures of crop protection or protection of stores, particularly where this may involve the health of human beings and animals."

Other paragraphs govern mandatory registration. In the meantime, crop protection legislation has been brought up to date (1986), but this has not yet been promulgated. Essential portions of this amendment will become effective on 1 Jan. 1987, others on 1 Aug. 1988 [50]. A new aspect of this legislation is that crop protection measures must not endanger the "balance of nature." Components of this balance are defined as "soil, water, air, animal and plant species, as well as their systems of interaction." In addition, the principles of integrated crop protection as a component of good agricultural practice are embodied in law. The legislative action to be ex-pected will be of significance for the further development of crop protection.

Furthermore, it was stipulated that international agreements must be observed, in particular, the Code of Conduct adopted in November 1985 by the Food and Agricultural Organization (FAO) of the United Nations. This Code of Conduct has been developed for the guidance of developing countries with inadequate crop protection legislation [51]. It includes testing methods, development of products with high user safety, analysis of the residue effects, equalization of quality standards, packaging specifications, simple directions for use (possibly including pictograms), sales and advertising regulations, and observance of the guiding principles of WHO and FAO. In addition, cooperation between importing and exporting countries is mandated.

8. References

1. C. Ahrens, H. H. Cramer in F. P. W. Winteringham (ed.): "Improvement of Agricultural Production by Pesticides," in: *Environment and Chemicals in Agriculture,* Elsevier, London, New York 1985, pp. 151 – 162.
2. H. H. Cramer in R. Wegler (ed.): "Ökonomisch – ökologische Wechselwirkungen," in: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 3,** Springer Verlag, Berlin, Heidelberg 1976, pp. 39 – 55.
3. H. H. Cramer B. Zeller in R. Wegler (ed.): "Zur Problematik des biologischen Pflanzenschutzes in der Landwirtschaft," in: *Chemie der Pflanzenschutz- und*

Schädlingsbekämpfungsmittel, **vol. 6,** Springer Verlag, Berlin, Heidelberg 1981, pp. 471 – 499.
4. FAO: *Production Yearbook 1983,* Rom 1984, p. 320.
5. H. H. Cramer: "Plant Protection and World Crop Production," *Pflanzenschutz Nachr. (Am. Ed.)* **20** (1967) 1 – 524.
6. F. Curschmann: *Hungersnöte im Mittelalter* – Ein Beitrag zur Wirtschaftsgeschichte des 8. bis 13. Jahrhunderts, Dissertation, Leipzig 1900, p. 32.
7. C. Woodham-Smith: *The Great Hunger, Ireland 1845 – 1849,* Harper and Row, New York, Evanston 1962, p. 510.
8. B. Rademacher: *Der Kampf des Menschen gegen seine Nahrungskonkurrenten,* Rektoratsrede, Verlag Ulmer, Stuttgart, Ludwigsburg 1954, p. 28.
9. C. Ahrens, H. H. Cramer, M. Mogk, H. Peschel: *Economic Impact of Crop Losses,* Proc. of the 10th Int. Congr. of Plant Protection, Brighton 1983, pp. 65 – 73.
10. K. H. Büchel: "Die Zukunft der Chemie in der Landwirtschaft," *Ber. Landwirtsch.* **61** (1983) 382 – 399.
11. Bundesministerium für Ernährung, Landwirtschaft und Forsten (ed.): *Statistisches Jahrbuch über Ernährung, Landwirtschaft und Forsten der Bundesrepublik Deutschland 1985,* Landwirtschaftsverlag GmbH Münster-Hiltrup 1985, p. 172.
12. FAO: *Agriculture: toward 2000,* Rome 1981.
13. "A look at Pesticide Markets," *Farm Chem.* **148** (1985) no. 9, 29 – 33.
14. G. Müller, H. v. Eickstedt, G. Zoebelein: "On the introduction of chemicals for the protection of cotton crops in Latin America and the Middle East," *Pflanzenschutz Nachr. (Am. Ed.)* **22** (1969) 47 – 63.
15. W. Glasow: "Der Arbeitsaufwand beim Vereinzeln von Rüben in Abhängigkeit von Saatgutform, Saatmethode und Vereinzelungsverfahren," in: *Schriftenreihe der Inst. f. landw. Arbeitswiss. und Landtechnik der Max-Planck-Gesellschaft zur Förderung der Wiss.,* Bad Kreuznach, H. 16, Stuttgart 1954, p. 95.
16. M. Hanf: "Von der Mechanik zur Chemie, 35 Jahre Pflanzenschutzentwicklung," in: *Mitteilungen aus der Biol. Bundesanstalt für Land- und Forstw.* H. 146, 1972, pp. 9 – 86.
17. M. Hanf: "Pflanzenschutzentwicklung in Deutschland 1946 bis 1971," in: *BASF – Mitteilungen für den Landbau,* 1972, pp. 1 – 26.
18. M. Hanf: "Geschichte der chemischen Unkrautbekämpfung im Rübenbau," in: *BASF – Mitteilungen für den Landbau,* 3/1975, pp. 1 – 26.
19. K. U. Heyland: "Über die Bedeutung des Herbizideinsatzes für die Entwicklung von Anbauverfahren," *Z. Pflanzenkr. Pflanzenschutz,* Sonderheft VII, 1976, pp. 21 – 28.
20. R. Heitefuß: *Pflanzenschutz, Grundlagen der praktischen Phytomedizin,* Stuttgart 1975, pp. 270.
21. R. Diercks in F. P. W. Winteringham (ed.): "Integrierter Pflanzenschutz: ein Weg zur ökologisch sinnvolleren Nutzung der Pestizide," in: *Environment and Chemicals in Agriculture,* Elsevier, London, New York 1985, pp. 243 – 257.
22. R. Diercks, C. Heye: "Notwendigkeit und Problematik der Ermittlung von Schadensschwellen," *Z. Pflanzenkr. Pflanzenschutz* **77** (1970) 610 – 627.
23. V. M. Stern: Significance of the economic threshold in integrated pest control. In: Proceedings of the FAO Symposium *"Integrated Pest Control"*, vol. II, edited by Food and Agriculture Organization of the United Nations, Rome 1966, pp. 41 – 56.
24. G. M. Hoffmann, F. Nienhaus, F. Schönbeck, H. C. Weltzien, H. Wilbert: *Lehrbuch der Phytomedizin,* 2nd ed., Paul Parey, Berlin, Hamburg 1985, p. 488.
25. G. Haug in R. Wegler (ed.): "Pflanzenschutzforschung der Industrie," in: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 3,** Springer Verlag, Berlin, Heidelberg 1976, pp. 57 –84.
26. H. Hoffmann: "On the probability of discovering new insecticidal compounds. Acyl formula and research prospects." *Pflanzenschutz Nachr. (Am. Ed.)* **29** (1976) 46 – 53.
27. F. W. Kremer: "Moderner Pflanzenschutz. Sicherung der Ernährung der Weltbevölkerung," in: *Schriftenreihe des Fonds der Chem. Industrie,* Heft 21, Chemie und Ernährung, Frankfurt/Main 1984, pp. 7 – 15.
28. G. Schuhmann: "Wie sicher sind unsere Pflanzenschutzmittel?" in: *Schriftenreihe des Fonds der Chem. Industrie,* Heft 21, Chemie und Ernährung, Frankfurt/Main 1984, pp. 25 – 35.

29. W. Kolbe: "Crop production and weed control. A comparative study with reference to long term trials on 40 crops at Höfchen and Laacherhof Experimental Stations," *Pflanzenschutz Nachr. (Am. Ed.)* **36** (1983) 205 – 373.
30. W. Kolbe: "Ten-year trials with Goltix for weed control in sugar and fodder beet (1974 – 1984) with reference to methods of control over the last 200 years," *Pflanzenschutz Nachr. (Am. Ed.)* **37** (1984) 424 – 505.
31. W. Kolbe: "Effect of different crop protection programmes on yield and quality of apples. II (1967 –1981)," *Pflanzenschutz Nachr. (Am. Ed.)* **35** (1982) 125 – 133.
32. J. M. Franz, A. Krieg: *Biologische Schädlingsbekämpfung,* 2nd ed., Paul Parey, Berlin, Hamburg 1976.
33. J. E. Laing, J. Hamai in C. B. Huffacker, P. S. Messenger (eds.): "Biological control of insect pests and weeds by imported parasites, predators and pathogens," in: *Theory and practice of biological control,* Academic Press, New York, San Francisco, London 1976, pp. 673 – 743.
34. R. Diercks: *Alternativen im Landbau. Eine kritische Gesamtbilanz,* Verlag Eugen Ulmer, Stuttgart 1983, p. 379.
35. D. Redlhammer: "Neue Wege im Pflanzenschutz unter besonderer Berücksichtigung des Integrierten Pflanzenbaus," in: *Chemie und Ernährung, Schriftenreihe des Fonds der Chem. Industrie,* Heft 21, Frankfurt/M. 1984, pp. 37 – 43.
36. T. J. Martin, D. B. Morris: "Cabbage root fly, *Phorbia* (*Erioischia*) *brassicae* Bouche, Control with Curaterr, a carbamate insecticide," Pflanzenschutz Nachr. Bayer (Engl. Ed.) **28** (1975) 92 – 136.
37. K. H. Büchel: Political, economic and philosophical aspects of pesticide use for human welfare, *Proc. Fifth Int. Congr. Pesticide Chemistry (IUPAC),* Kyoto 1982, p. 33.
38. F. W. Kremer, P. Kraus, K. A. Lein, H. J. Santel: *Biotechnologie – Alternative oder Ergänzung zum chemischen Pflanzenschutz?* Vortrag 14. 12. 1984, München, 29 msp., unpublished.
39. G. Zoebelein, U. Kniehase: "Laboratory, greenhouse, and field trials on the effect of nikkomycins on insects and mites," *Pflanzenschutz Nachr. (Am. Ed.)* **38** (1985) 203 – 304.
40. W. Crüger, W. Frommer, C. Goelker, J. W. Kaiser, H.-F. Moeschler, O. Salcher, M. Schedel. H. Wehlmann: "New crop protectant active principles from microorganisms: Optimization of manufacturing methods for nikkomycins," *Pflanzenschutz Nachr. (Am. Ed.)* **38** (1985) 305 – 348.
41. K. H. Büchel: "Evolution of crop protection chemicals as exemplified by fungicides," *Pflanzenschutz Nachr. (Am. Ed.)* **35** (1982) 1 – 35.
42. H. H. Cramer: "Zur Problematik des Pflanzenschutzes in der modernen Landwirtschaft," *Chemieunterricht* **9** (1978) 5 – 40.
43. Fonds der Chemischen Industrie: *Pflanzenschutz, Beiheft,* **Folienserie 10,** Fonds der Chemischen Industrie im Verband der Chemischen Industrie, Frankfurt am Main 1985, p. 47.
44. H. H. Cramer, H. J. Jarczyk, *Ullmann,* 4th ed., **6,** 506 – 509.
45. H. J. Jarczyk: "Lysimeter installation of Bayer Agrochemicals Division at Leverkusen," *Pflanzenschutz Nachr. (Am. Ed.)* **36** (1983) 1 – 20.
46. Japanese Society of Agricultural Chemical Industry: *Agricultural Chemicals Laws and Regulations,* Japan (I), Tokyo 1984, p. 38.
47. United States Environmental Protection Agency, Office of Pesticides Programs, Washington DC, Nov. 1978. The Federal Insecticide, Fungicide and Rodenticide Act as Amended by Public Law 95 – 396, Sept. 30, 1978, OPA 17/9, p. 40.
48. Bundesrepublik Deutschland (1978): Pflanzenschutzgesetz vom 10. Mai 1968 (BGBl. I, S. 352), zuletzt geändert durch das Dritte Gesetz zur Änderung des Pflanzenschutzgesetzes vom 16. Juni 1978; Bundesgesetzblatt I, Bundesrepublik Deutschland, 749 ff.
49. Industrieverband Pflanzenschutz e. V.: IPS-Kodex, Frankfurt/Main 1981, p. 217.
50. Industrieverband Pflanzenschutz e. V.: *Das neue Pflanzenschutzgesetz.* Jahresbericht des Industrieverbandes Pflanzenschutz e. V., Frankfurt/Main 1986, pp. 25 – 27.
51. FAO: *International Code of Conduct on the Distribution and Use of Pesticides,* Conference of the Food and Agriculture Organization of the United Nations, Rome, C 85/25, Twenty-third Session, Rome 9 – 28 Nov. 1985, p. 17.

Crotonaldehyde and Crotonic Acid

Rolf P. Schulz, Celanese Chemicals Europe GmbH, Oberhausen, Germany

Jürgen Blumenstein, Celanese Chemicals Europe GmbH, Frankfurt a. M., Germany

Christian Kohlpaintner, Celanese Chemicals Europe GmbH, Oberhausen, Germany

1. Crotonaldehyde 1
1.1. Physical and Chemical Properties . . 1
1.2. Production 2
1.3. Uses 3
2. Crotonic Acid 4
2.1. Physical and Chemical Properties . . 4
2.2. Production 4
2.3. Uses 5
3. Toxicology and Occupational Health . 5
4. References 6

1. Crotonaldehyde

Crotonaldehyde or 2-butenal [*4170-30-3*] exists in two stereoisomeric forms: *cis*-crotonaldehyde [*15798-64-8*] and *trans*-crotonaldehyde [*123-73-9*]. Both isomers are contained in commercially available crotonaldehyde, in which the trans isomer predominates (> 95 %) due to its higher thermodynamic stability. Crotonaldehyde is a colorless liquid with strong lacrimatory properties and a pungent odor. Crotonaldehyde occurs naturally in many plants and plant products, e.g., french beans, rapeseed oil, and soybean oil.

1.1. Physical and Chemical Properties

Some physical properties of crotonaldehyde are summarized in the following [3], [4]: Crotonaldehyde is soluble in water and miscible with many organic solvents. Due to its double bond and the carbonyl group, crotonaldehyde is very reactive. Contamination with alkali or strong acids readily induces exothermic condensation reactions [1]. Crotonaldehyde dimerizes to dicrotonaldehyde (5,6-dihydro-2,6-dimethyl-2*H*-pyran-3-carboxaldehyde) when exposed to strong acids [2].

Boiling point	102.2 °C
Melting point	− 69 °C
Density (20 °C)	0.852 g/cm^3
Refractive index n_D^{20}	1.438
Heat of vaporization	515 J/g
Crotonaldehyde – water azeotrope	24.8 wt % H_2O
Azeotrope *bp*	84 °C
Solubility in water (20 °C)	150 g/L
Solubility of water in crotonaldehyde (20 °C)	9.5 g/100 g
Flash point	13 °C
Auto ignition temperature	155 °C
Explosion limits in air	
lower	2.1 vol %
upper	15.5 vol %

The most important classes of reactions involving crotonaldehyde are hydrogenation, reduction, oxidation, and addition reactions.

Reduction. Both the C – C double bond and the carbonyl group can be reduced. Most economically this can be achieved by catalytic hydrogenation, e.g., in the presence of nickel or copper catalysts [5]. Lower temperatures and pressures lead to selective hydrogenation of the olefinic moiety, which was formerly used to produce *n*-butyraldehyde [6]. Nowadays butyraldehyde is normally made by hydroformylation of propene. The hydrogenation of both the double bond and the carbonyl group leads to *n*-butanol [5], which is today usually produced by hydrogenation of *n*-butyraldehyde. Crotonaldehyde may also be catalytically hydrogenated selectively at the carbonyl group to give crotyl alcohol $CH_3CH{=}CHCH_2OH$ [7–9].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_083

Oxidation of crotonaldehyde leads to crotonic acid (see Chap. 2).

Additions. The ease of addition of a large variety of molecules to the double bond makes crotonaldehyde an important chemical intermediate. For example, alcohols or thiols [10], [11] can be added to the olefinic double bond with a basic catalyst. This reaction is particularly important for the synthesis of 3-alkoxybutyraldehydes. The specialty solvents 3-Methoxybutanol [12], [13] and 3-methoxybutyl acetate can be synthesized by the addition of methanol to crotonaldehyde followed by consecutive hydrogenation and esterification of the intermediary 3-methoxybutyraldehyde. Crotonaldehyde acts as a dienophile in Diels – Alder reactions [14].

1.2. Production

The usual method to produce crotonaldehyde is the aldol reaction of acetaldehyde followed by dehydration of the acetaldol and distillation. Typically the product is about 99 % pure.

$$2\,CH_3CHO \longrightarrow CH_3CHOHCH_2CHO \longrightarrow (E)\text{-}CH_3CH{=}CHCHO + H_2O$$

The aldolization can be catalyzed by all types of basic catalysts, e.g., alkali metal or alkaline earth metal catalysts [15–18], ammonium salts [19], [20], zeolites [21–23], molecular sieves [24] and claylike materials [25]. Acetic acid, mineral acids, or acidic cation-exchange resins are used as catalysts for the dehydration of acetaldol [26], [27].

Major producers of crotonaldehyde are Celanese Chemicals, Daicel, Eastman Chemicals, and Sasol.

A flow diagram of a typical continuous crotonaldehyde production unit is depicted in Figure 1.

Acetaldehyde is aldolized in the water-cooled aldol reactor (a) with aqueous sodium hydroxide as catalyst. The product is neutralized with acetic acid and fed into a stripping column (b). Unconverted acetaldehyde is taken overhead and recycled to the reactor (a). The intermediate acetaldol is fed as a bottom product of column (b) into the azeotrope distillation column (c). The aldol is dehydrated, and a crotonaldehyde/water azeotrope distills and is separated into water and an aqueous crotonaldehyde phase containing ca. 10 % water. The water is fed back into the azeotrope column (c), which it leaves as the bottom product. The aqueous crotonaldehyde stream is fed into the rectification column (d). Again a crotonaldehyde/water azeotrope leaves the column overhead and separates into water, which is recycled into the azeotrope column (c), and aqueous crotonaldehyde, which is recycled into the rectification column. Pure crotonaldehyde is taken as a side cut. Heavy ends are taken from the bottom of column (d) and combined with organic residues from the azeotrope column (c) and incinerated. Wastewater from column (c) is treated in a wastewater-purification plant.

An interesting technology is the oxidation of butadiene to its monoepoxide [28], which is analogous to the Wacker–Hoechst process for the oxidation of ethylene to acetaldehyde, followed by isomerization of the butadiene monoepoxide [29], [30] to crotonaldehyde.

Paraldol as a substitute for acetaldol [31] is more interesting for producing cosmetic-grade 1,3-butylene glycol than for crotonaldehyde synthesis.

The catalytic oxidation of olefins [32–34], the dehydrogenation of allyl alcohol with amines [35], hydratization of acetylene [36], or the enzymatic oxidation [37], [38] of 2-butene are so far not of practical importance.

Crotonaldehyde is also a byproduct in the synthesis of 1,3-butylene glycol [31], [39] and in acetic acid synthesis by methanol carbonylation [40], [41]. As a result of incomplete combustion, it is also contained in automobile exhaust gases.

Storage and Transportation. Any alkaline or acidic impurity may cause dangerously exothermic condensation reactions of the crotonaldehyde and therefore must be thoroughly excluded. The solidification and color development of crotonaldehyde during storage and transportation can be avoided by careful exclusion of oxygen or other strong oxidizers and light by using stainless steel or aluminum containers. This prevents formation of peroxides and of iron salts of crotonic acid, which induce darkening and polymerization. With these precautions, crotonaldehyde can be stored and shipped without the addition of inhibitors; otherwise, water and hydroquinone can be used as stabilizers.

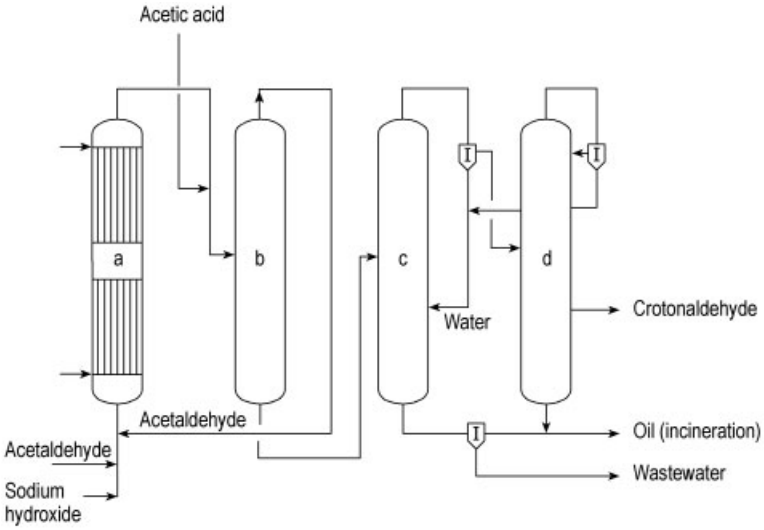

Figure 1. Acetaldehyde condensation to acetaldol and dehydration to crotonaldehyde
a) Aldol reactor; b) Acetaldehyde stripping column; c) Azeotrope distillation column; d) Rectification column

Due to the toxic and lacrimatory properties of crotonaldehyde (Chap. 3) its transport is regulated; mailing and air transport are prohibited [4].

1.3. Uses

Large amounts of crotonaldehyde are used for the synthesis of sorbic acid (*trans*, *trans*-2,4-hexadienoic acid), a food preservative, and Vitamin E. Sorbic acid is made by the condensation of crotonaldehyde with ketene in the presence of organic zinc salts, followed by depolymerization of the resulting polyester either thermally or with a mineral acid [42–45].

α-Tocopherol, one of eight naturally occurring substances with vitamin E activity, is synthesized on an industrial scale by the condensation of phytol with 2,3,6-trimethylhydroquinone. The latter is made in four steps from crotonaldehyde and diethyl ketone (→ Vitamins, Chap. 4.) [46].

3-Methoxybutanol is made by addition of methanol to crotonaldehyde [12], [13]. This product and its acetate ester are used as specialty solvents, for example, for the control of viscosity, drying behavior, and gloss of lacquers and varnishes (→ Solvents, Chap. 6.).

Furthermore, crotonaldehyde is widely used for the production of chemical intermediates, such as crotonic acid, which is made by the oxidation of crotonaldehyde (Chap. 2.2). End uses of crotonaldehyde-based intemediates include pharmaceuticals and biocompatible or medical articles [47], [48], agrochemicals [49], [50], resins, polymetric thickeners [51], paints and coatings [52], dyestuffs [53], rubbers [54] and rubber antioxidants [53], gelatin hardening [53], adhesives [55], [56], leather tanning [53], leather and paper sizing [53], metal brighteners [53], lubricants [57], and corrosion inhibitors [58].

Crotonaldehyde is also employed in flavoring agents [53] and as an odorant and warning agent in fuel gases as well as for the manufacture of tear gas [53], as a solvent for purification of mineral and lubricating oils, as a solvent for poly(vinyl chloride) and in polymer technology as a reaction medium [53].

2. Crotonic Acid

Crotonic acid [*3724-65-0*] is the *trans* isomer of the two stereoisomeric forms of 2-butenoic acid. *cis*-2-Butenoic acid is called isocrotonic acid [*503-64-0*] but is not commercially available in larger quantities. Commercial crotonic

acid is a white to yellow solid with a pungent and suffocating odor. Isocrotonic acid is a colorless liquid.

Crotonic acid occurs naturally, e.g., in crude wood distillate and sponges, and it is a component of defensive substances or secretions of various beetles.

Table 1. Physical properties of crotonic and isocrotonic acid [3], [59], [60]

	Crotonic acid	Isocrotonic acid
Boiling point	189 °C, 81 °C (17 mbar)	169 °C
Melting point	72 °C	15 °C
Density (20 °C)	1.018	1.0267
Refractive index n_D	1.4228 (80 °C)	1.4456 (20 °C)
Solubility in water, g/kg	41.5 (0 °C), 94 (25 °C)	
	656 (40 °C), 1260 (42 °C)	
Flash point	88 °C	
Auto ignition temperature	396 °C	
Vapor pressure	24 Pa (20 °C), 880 Pa (70 °C)	
Specific heat	3.031 J g^{-1} K^{-1} (s), 2.072 J g^{-1} K^{-1} (l)	
Heat of combustion	2.00 MJ/mol	2.03 MJ/mol
Heat of fusion	150.9 J/g	
pK_a	4.817 (25 °C)	

2.1. Physical and Chemical Properties

Some physical properties of crotonic acid and isocrotonic are summarized in Table 1.

Crotonic acid crystallizes as white needles or prisms in the monoclinic system. Its solubilities at 25 °C are: in water 9.4 wt %, ethanol 52 wt %, acetone 53 wt %, ethyl acetate 72 wt %, and toluene 37 wt %. The aqueous solution acts as a weak acid. The azeotrope of crotonic acid and water (96.86 %) boils at 99.7 °C.

Crotonic and isocrotonic acid form an eutectic mixture containing 30 % crotonic acid with a melting point of − 3 °C.

Isocrotonic acid is miscible with water and polar solvents. Heating of either or both acids, or treatment with other acids, bases, or UV radiation leads to isomerization, oligomerization, and polymerization.

Isocrotonic acid isomerizes to crotonic acid. The equilibrium between the two isomers is reached between 140 and 180 °C [3]; 3-butenoic acid is a byproduct of this process. The equilibrium mixture has an isocrotonic to crotonic acid ratio of ca. 0.17 : 1.

Both acids can be hydrogenated to crotyl alcohol and butyric acid or oxidized to the peracid. The most important reactions are polymerizations, additions, and esterifications.

Polymerization. Many types of monomers can be copolymerized with crotonic acid (radical mechanism). Important copolymers are formed with vinyl acetate (see Section 2.3).

Additions. Additions to crotonic acid yield 2- and 2,3-substituted butyric acid. Crotonic acid can be hydrogenated catalytically to butanoic acid. The addition of halogens or hydrogen halides to the olefinic double bond leads to the corresponding 2,3-dihalobutanoic acid or 3-halobutanoic acid, respectively. Treatment with OsO_4 or peracids like peroxybenzoic acid yields 2,3-dihydroxybutanoic acid. The addition of ammonia under suitable reaction conditions gives β-aminobutyric acid. Crotonic acid acts as a dienophile in Diels – Alder additions.

Other Reactions. Crotonic acid can be esterified by conventional methods, although the reaction is often slower than with butyric acid due to the conjugated system. Crotonyl halides can be prepared by the reaction of crotonic acid with a corresponding acyl halide; crotonic anhydride by the reaction of crotonyl chloride with sodium crotonate; or the reaction of crotonic acid with acetic anhydride.

2.2. Production

Crotonic acid is prepared on an industrial scale by the oxidation of crotonaldehyde, for instance, by Daicel and Eastman Chemicals. Peroxocrotonic acid is formed as an intermediate, which comproportionates with crotonaldehyde to yield two molecules of crotonic acid:

$$CH_3CH{=}CHCHO + O_2 \longrightarrow CH_3CH{=}CHC(O)OOH$$

$$CH_3CH{=}CHC(O)OOH + CH_3CH{=}CHCHO \longrightarrow 2\,CH_3CH{=}CHCOOH$$

The reaction is typically conducted at low pressure (ca. 1 – 5 bar) and temperatures between 20 and 45 °C. Higher temperatures result in oxidative decomposition. Catalysts such as manganese, cobalt, copper, or thallium salts result in satisfactory conversions and yields even at 20 °C [3], [61], [62] and prevent the buildup of undesired peroxo compounds. In the former Hoechst

process (manganese salts as catalysts, oxidation at 20 °C) no additional solvents were used. A typical raw product contains about 30 % crotonic acid, 1 – 3 % of formic and acetic acids, water, and only traces of isocrotonic acid. This crude product is fractionated at reduced pressure; unconverted crotonaldehyde is recycled to the oxidation reactor.

Isocrotonic acid is removed from the final product by fractional crystallization, yielding crotonic acid with up to 99.9 % purity [63], [64].

On a nonindustrial scale, crotonic acid can be prepared, e.g., by dehydration of 2-hydroxybutanoic acid, by photochemical oxidation or by photochemical oxidation or oxidative irradiation of crotonaldehyde with ultrasound [65], by the isomerization of vinylacetic acid with sulfuric acid [60], by the condensation of acetaldehyde and malonic acid with pyridine as catalyst (85 % yield) [66], by the oxidation of butene with a heteropolymolybdic acid containing catalyst system [67], by the oxycarbonylation of propene with transition metal complex catalysts [68], by the carbonylation of propylene oxide [69], by the reaction of acetic anhydride with acetaldehyde with basic aluminum acetate as catalyst [70], and by the carbonylation of allyl alcohol in a two-phase system with a nickel-containing phase - transfer catalyst (90 % conversion, 68 % yield) [71] or a palladium catalyst [72].

Isocrotonic acid can be obtained by fractional crystallization, rectification or gel filtration from a mixture of the two 2-butenoic acids. On a laboratory scale it can also be prepared from 2-butanone by bromination followed by a stereospecific Favorskii rearrangement [73]. Other methods to prepare isocrotonic acid are reviewed in [74].

Storage and Transportation. Crotonic acid is usually shipped in paper bags with inner polyethylene bags or polyethylene-lined drums. It should be stored in the dark away from bases, oxidants, and reducing agents to avoid polymerizations or other strongly exothermic reactions.

2.3. Uses

The main use of crotonic acid is in the synthesis of copolymers with a variety of comonomers. Especially important are crotonic acid – vinyl acetate copolymers. Trade names include Cevian, Gelva, Mowilith, and Vinac.

Crotonic acid copolymers have a very broad range of applications: paints and coatings [75–78], hot-melt paints [79]; adhesives [80], [81], and hot-melt adhesives [82]; coatings for paper and textiles; flocculants [83], binders for explosives [84], ceramics [85] and agrochemicals [86]; and drilling additives [87].

Crotonic acid has antimicrobial properties and may like its esters be used in deodorants [88], [89], UV absorbers [90] and metal-effect interference pigments [91]. Fatty alcohol esters of crotonic acid are used in the leather industry [92], [93]; the free acid is used in motor fuels [94], [95], for the etching of metal surfaces [96–98], for electrochemical metal deposition [99], and PVC heat stabilization [100].

Several agrochemicals and pharmaceuticals are based on crotonic acid and its anhydride or chloride.

3. Toxicology and Occupational Health

Crotonaldehyde. Perception of the irritating, pungent, and suffocating odor of crotonaldehyde has been reported at thresholds from 0.035 to 0.12 ppm in air [53]. Therefore, crotonaldehyde is used as an odorant and warning agent in fuel or gas [53].

OSHA, NIOSH, and ACGIH have established a permissible exposure limit and TLV of 2 ppm for an 8-h workday [53]. The German TRGS 900 allows a maximum concentration of 1 mg/m^3, the former MAK classification IIIB has been abandoned [4]. Exposure may occur through inhalation, eye or skin contact and by absorption through the skin or ingestion.

Crotonaldehyde is severely irritating to the eyes, mucous membranes, the upper respiratory system and skin in animals and humans. Sensitization may occur. It is classified as a marine pollutant although its 10-d biodegradibility is $>$ 83 % according to OECD 302 B [4].

Effects on Humans. A 10-min exposure to 4 ppm of crotonaldehyde induces lacrimation; conjunctival irritation is caused by only brief exposure to 45 ppm [101], corneal injuries (exposure severity not specified) could be healed in at least eight cases during 48 h.

Effects on Animals and Microorganisms. The following toxicology and ecotoxicology data have been determined:

LC_{50} (rats)	1500	ppm (30 min)	[53]
LC_{50} (rat, oral)	206 – 300	ppm	[4]
LC_{50} (rat, inhalative)	0.29	mg/L	[4]
LC_{50} (fish/*Pimephales promelas*)	0.85	mg/L (96 h)	[4]
LD_{50} (rabbit, dermal)	128 to 324	mg/kg	[4]
LD_{100} (rat, subcutaneous)	180	mg/kg	[102]
EC_{50} (*Daphnia magna*)	2	mg/L (48 h)	[4]
EC_{50} (algae)	0.88	mg/L (72 h)	[4]
EC_{10} (bacteria)	150	mg/L	[4]

10 ppm of crotonaldehyde caused adverse changes in pulmonary performance in rats during a single 200-min exposure [103]; 1650 ppm (10-min exposure) induces respiratory distress, pulmonary edema, bronchiolar damage, excitation, and terminal convulsions [101].

Chronic effects or oncogenic properties [104], [105] are not linked so far to crotonaldehyde exposure, although crotonaldehyde can induce point mutations in bacteria [106] and carcinogen-induced DNA damage [107] and is therefore classified as mutageneous [4].

Crotonic acid. Crotonic acid is corrosive to the eyes (causing pain, blurred vision, deep burns), skin (causing skin burns and pain), and respiratory tract (causing a burning sensation, cough, headache, nausea, sore throat and shortness of breath; these symptoms may be delayed), and on ingestion (causing pain, burning sensation, sore throat, diarrhea, vomiting).

The following toxicology data have been determined in animal experiments:

LD_{50} (rat, oral)	1 g/kg	[108]
LD_{50} (rat, i.p.)	100 mg/kg	[109]

Crotonic acid is formed by fat metabolization [110], and is rapidly converted by enzymes in the liver or other tissues. No cumulative effects are likely after exposure to nonirritating concentrations [3].

Crotonic acid affects plant growth and the germination of seeds [60].

4. References

Specific References

1. G. Hommel: *Handbuch der gefährlichen Güter,* 4th ed., **vol. 1,** Data sheet No. 65, Springer-Verlag, Berlin – Heidelberg 1987.
2. E. Spaeth, R. Lorenz, E. Freund, *Monatsh. Chem.* **76** (1947) 297.
3. *Ullmann,* 5th ed., **vol. A8.**
4. EG Safety data sheet (MSDS): *Crotonaldehyde,* Celanese, 1998.
5. Hüls AG, EP 470 344, 1994.
6. Buna AG, DD 300 145, 1993.
7. Mitsubishi Petrochemical Co. Ltd., JP 05 301 834, 1994.
8. Rhône-Poulenc Sante, EP 332 522, 1993.
9. Mitsubishi Chem. Corp.; Mitsubishi Petrochemical Co. Ltd., EP 582 277, 1995.
10. AS BELO Bioorg. Chem., SU 1 384 575, 1993.
11. Atochem North America Inc.; Pennwalt Corp., AU 8 776 733, 1993.
12. Hoechst AG, DE 4 315 047, 1994.
13. Daicel Chem. Ind. Ltd., JP 63 196 532, 1993.
14. O. Diels, K. Alder, *Liebigs Ann. Chem.* **470** (1929) 66.
15. Chisso Corp., JP 09 059 201, 1997.
16. Mitsubishi Chemical Industries Co., Ltd., JP 62 195 341, 1987.
17. Mitsubishi Chemical Industries Co., Ltd., JP 49 006 887, 1974.
18. US 3 450 769, 1969 (J. E. Payne, L. D. Hinson).
19. DE 809 808, 1949 (A. M. Paquin).
20. Mitsubishi Gas Chemical Co., Inc., JP 57 018 642, 1982.
21. Mobil Oil Corp., US 5 292 980, 1994.
22. Chisso Corp., JP 09 057 108, 1997.
23. Mobil Oil Corp., US 5 026 919, 1991.
24. Union Carbide Corp., US 4 701 562, 1993.
25. W. Kagunya, W. Jones, *Appl. Clay. Sci.* **10** (1995) nos. 1 – 2, 95.
26. Tokuyama Sekiyu Kagaku K. K., JP 57 163 333, 1982.
27. Mitsubishi Chemical Industries Co., Ltd., JP 48 042 043, 1973.
28. Eastman Chem. Co., Eastman Kodak Co., US 4 942 263, 1995.
29. Eastman Kodak Co., US 4 925 987, 1993.
30. Eastman Kodak Co., US 4 925 986, 1993.
31. Daicel Chemical Industries, US 5 583 270, 1997.
32. Asahi Chemical Industry Co., Ltd., JP 56 059 722, 1981.
33. Asahi Chemical Industry Co., Ltd., JP 54 151 904, 1979.
34. Sun Co. Inc. R & M, EP 527 623, 1994.
35. BASF AG, DE 2 020 865, 1971.
36. All-Union Scientific-Research Institute of Synthetic Rubber, SU 189 826, 1966.
37. Showa Denko KK, JP 01 132 391, 1993.
38. Givaudan Roure Flavors Corporation, US 5 783 429, 1996.

39. Daicel Chemical Industries, Ltd., US 5 345 004, 1997.
40. Hoechst Celanese Corp., WO 9 710 198, 1997.
41. Daicel Chem. Ind. Ltd., EP 696 565, 1996.
42. Ingrasaminte Chim., RO 106 990, 1994.
43. Nippon Synthetic Chem. Ind. Co., JP 05 032 583, 1993.
44. Chisso Corp., DE 4 133 328, 1993.
45. Daicel Chem. Ind. Ltd., JP 62 246 539, 1993.
46. BASF AG, DE 3 636 057, 1993.
47. Medtronic Inc., WO 9 808 551, 1998.
48. Norsk Hydro AS, WO 9 105 817, 1993.
49. BASF AG, WO 9 308 148, 1993.
50. Atochem North America Inc.; Pennwalt Corp., US 4 944 853, 1993.
51. Hoechst AG, DE 3 935 137, 1994.
52. Memtec Ltd., WO 9 601 282, 1996.
53. Occupational safety and health guidelines for Crotonaldehyde, OSHA, http://www.osha-slc.gov/SLTC/ , 1999.
54. NOK Corp., JP 04 247 940, 1992.
55. Aica Kogyo, JP 08 218 046, 1996.
56. Pinchuk NM, SU 686 292, 1996.
57. Texaco Development Corp., EP 201 164, 1986.
58. Dow Chemical Co., US 4 837 323, 1989.
59. International Chemical Safety cards (WHO/IPCS/ILO), *Crotonic acid,* http://siri.uvm.edu/msds/mf/cards/list.html , 1999.
60. Crotonic acid, *Kirk-Othmer,* 3rd ed., **vol. 7,** 1979, pp. 218 – 226.
61. Atlantic Richfield Co., US 4 097 523, 1978.
62. S. P. Boboleva, M. G. Bulygin, V. I. Valov, E. A. Blyumberg, *Neftekhimiya* **30** (1990) no. 2, 239.
63. Hoechst AG, DE 3 245 111, 1984.
64. Hoechst AG, EP 521 499, 1993.
65. Saiden Kagaku KK, JP 08 059 541, 1996.
66. R. A. Letch, R. P. Linstead, *J. Chem. Soc.* (1932) 454.
67. Asahi Chemical Industry Co., Ltd., JP 56 059 722, 1981.
68. Roehm, DE 2 324 132, 1974.
69. Diamond Alkali Co., US 3 024 275, 1962.
70. Daicel Chem., JP 06 087 784, 1994.
71. H. Alper, I. Amer, *J. Mol. Catal.* **54** (1989) no. 2, L33.
72. Quantum Chemical Corp., US 4 801 743, 1989.
73. A. F. Kluge, J. Meinwald, *Org. Synth.* **53** (1973) 123 – 127.
74. M. B. Hocking, *Synthesis* (1971) 482.
75. De Soto, Inc., US 4 132 688, 1979.
76. Mitsubishi Yuka Fine Chemicals Co., Ltd., JP 61 185 571, 1986.
77. Mitsubishi Petrochemical Co., Ltd., DE 2 752 217, 1978.
78. Konica Co., JP 09 294 956, 1997.
79. Mitsubishi Kasei Corp., JP 01 081 862, 1989.
80. Nichiden Kagaku Co., Ltd., JP 56 022 373, 1989.
81. Kuraray Co., Ltd., JP 03 045 679, 1991.
82. Hoechst AG, DE 3 134 998, 1983.
83. Nalco Australia Pty Ltd, AU 668 029, 1996.
84. CFPI, FR 2 701 942, 1994.
85. Hoechst AG, WO 9 407 808, 1994.
86. Nippon Carbide Kogyo Kk, JP 07 258 001, 1995.
87. Nippon Oils and Fats Co., Ltd., JP 61 028 577, 1986.
88. Mitsubishi Materials Corp., JP 09 028 777, 1997.
89. Fritsche Brothers Inc., US 3 074 891, 1963.
90. General Electric Co., EP 373 401, 1990.
91. Daimler Benz AG, Wacker-Chemie GmbH, DE 4 418 076, 1995.
92. Henkel K.G.a.A., DE 4 400 507, 1995.
93. Henkel K.G.a.A., DE 4 223 111, 1994.
94. Faming Zhuanli Shenqing Gongkai Shuominsghu, CN 1 079 243, 1993.
95. Texaco Inc., US 4 869 728, 1989.
96. MEC Co., Ltd., EP 794 269, 1997.
97. MEC Co., Ltd., EP 620 293, 1994.
98. Mitsubishi Chemical Industries Co. Ltd., JP 56 150 593, 1981.
99. T. Al-Sahhaf, M. S. E. Abdo, Altern. Energy Sources, [Proc. Miami Int. Conf.], 8th (1989), Meeting Date 1987, 91 – 101 (CAN 114:131794).
100. Tokuyama Sekisui Kogyo Co. Ltd., JP 74 110 741, 1974.
101. G. J. Hathaway, N. H. Proctor, J. P. Hughes, M. L. Fishman: *Proctor and Hughes' chemical hazards of the workplace,* 3rd ed., Van Nostrand Reinhold, New York, NY, 1991.
102. E. Skog, *Acta Pharmacol. Toxicol.* **6** (1950) 299.
103. ACGIH (ed.): *Documentation of the Threshold Limit Values and Biological Exposure Indices,* 6th ed., American Conference of Governmental Industrial Hygienists, Cincinnati, OH 1991.
104. J. O. Rundell, D. I. Brusick: Litton Bionetics Inc., Project No. 21002, Kensington, MD, USA, 1980.
105. Chem Service, MSDS trans-2-Butenal, 0-228; http://siri.org/msds/ , 1999.
106. T. Neudecker, D. Lutz, E. Eder, D. Henschler, *Mutat. Res.* **91** (1981) 27.
107. American Health Found, US 4 923 813, 1993.

108. H. F. Smyth, C. P. Carparter, *J. Ind. Hyg. Toxicol.* **26** (1944) 269.
109. *Patty's Industrial Hygiene and Toxicology,* 3rd ed., Wiley-Interscience, New York 1982, p. 4953.
110. J. Fruton, S. Simmonds: *General Biochemistry,* 2nd ed., J. Wiley & Sons, New York 1958.

Crown Ethers

EDWIN WEBER, Institut für Organische Chemie, TU-Bergakademie, Freiberg, Germany

1. **Classification and Nomenclature** 1
2. **Physical Properties** 3
3. **Chemical Properties** 3
4. **Crown Ether Complexes** 3
5. **Production** 6
6. **Uses** 6
7. **Toxicology and Occupational Health** . . 9
8. **References** 9

Crown ethers belong to the general group of macrocyclic polyethers. The name crown ethers refers to the crown-like atomic arrangement. The most remarkable property of crown ethers is their ability to yield stable complexes with alkali and alkaline-earth metal ions and ammonium ions. Crown ethers may also act as "second-sphere ligands" of metal ion complexes. Under optimized conditions, crown ethers can be used effectively for ion extractions, but other important uses are, for example, as complexing agents and phase-transfer catalysts in organic syntheses. Besides the characteristics and uses of crown ethers, their production processes and toxicological features are addressed as well.

1. Classification and Nomenclature

Crown ethers [1 – 11] were originally defined as macrocyclic polyethers with the following structure [1]:

where Z = alkylene, arylene, or cycloalkylene

Typical bridging groups are 1,2-ethylene, *o*-phenylene, and 1,2-cyclohexylene, which can also be substituted. In the common crown ethers **1–9** x ranges from 3 to 6. Dioxane ($x = 1$) is not considered a crown ether (→ Dioxane).

Hetero analogues in which oxygen is replaced either partially or completely by other heteroatoms [12], especially sulfur [13, 14] or nitrogen [15, 16], are also known (**10** – **15**) [5, 6, 9]. The heteroatoms can be integrated into heterocycles, as in **15** or can even be a metal center (metallacrown ether) [17, 18]. These hetero crown ethers are also called coronands [5, 19].

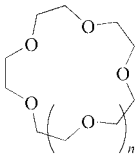

1: $n = 0$ **2**: $n = 1$ **3**: $n = 2$ **4**: $n = 3$

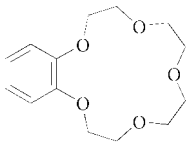

5

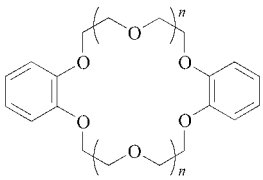

6: $n = 1$ **7**: $n = 2$

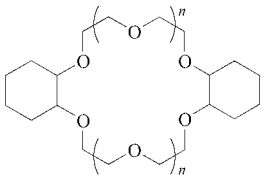

8: $n = 1$ **9**: $n = 2$

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_091.pub2

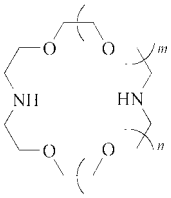

10: $n = 0$, $m = 1$ **11**: $n = 1$, $m = 1$ **12**: $n = 2$, $m = 1$

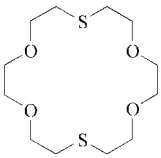

13

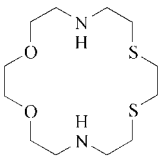

14

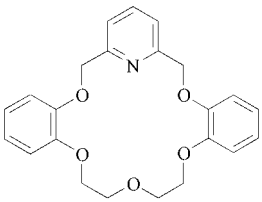

15

Other expanding subclasses in the general group of macrocyclic polyethers are the lariat ethers [20 – 22], in which a crown ether (coronand) ring is attached to one or more donor sidearms, such as **16**, and the spherands [20, 23, 24] having an enforced crown-type cavity, such as **17**. Occasionally, open-chain analogues such as **18** (podands) [25 – 27] and bi- or oligocyclic compounds such as **19** (cryptands) [28 – 30] are also classified as crown ethers [19].

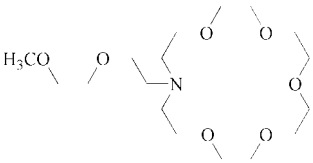

16

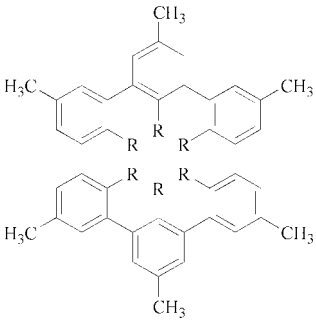

17 R = OCH_3

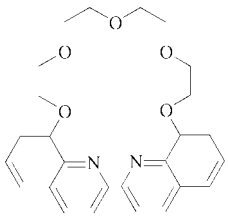

18

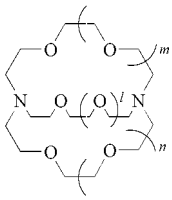

19: $l = m = n = 0 – 2$

A more general term applicable in this field of substances is crown compounds [19].

The name crown ethers refers to the crown-like atomic arrangement [1]. A special system of nomenclature is generally preferred to the more complicated IUPAC system (Table 1) [19]. Thus, compound **2**, which contains 5 endocyclic oxygen atoms and a total of 15 atoms in the ring, is usually called 15-crown-5 (or [15]crown-5). Additional substituents or condensation sites such as "benzo" and "dicyclohexano" (older nomenclature: dicyclohexyl or perhydrodibenzo) appear as a prefix. Acronyms are also used for common crown ethers (Table 1). A more recent terminology enables a more precise indication of the type and sequence of the heteroatoms and bridging units (e.g., **14**: 18 O_2NS_2N-2_6-coronand-6⟩) [31]. The phane nomenclature can be used

Table 1. Nomenclature of crown ethers

Compound no.	IUPAC name	CAS registry no.	Trivial name	Acronym
1	1,4,7,10-Tetraoxacyclododecane	[*294-93-9*]	12-crown-4	12C4
2	1,4,7,10,13-Pentaoxacyclopentadecane	[*33100-27-5*]	15-crown-5	15C5
3	1,4,7,10,13,16-Hexaoxacyclooctadecane	[*17455-13-9*]	18-crown-6	18C6
4	1,4,7,10,13,16,19-Heptaoxacycloheneicosane	[*33089-36-0*]	21-crown-7	21C7
5	2,3,5,6,8,9,11,12-Octahydro-1,4,7,10,13-benzopentaoxacyclopentadecin	[*14098-44-3*]	benzo-15-crown-5	B15C5
6	6,7,9,10,17,18,20,21-Octahydrodibenzo [*b,k*]-[1,4,7,10,13,16]hexaoxacyclooctadecin	[*14187-32-7*]	dibenzo-18-crown-6	DB18C6
7	6,7,9,10,12,13,20,21,23,24,26,27-Dodecahydro-dibenzo [*b,n*] [1,4,7,10,13,16,19,22]-octaoxacyclotetracosin	[*14174-09-5*]	dibenzo-24-crown-8	DB24C8
8	2,5,8,15,18,21-Hexaoxatricyclo [20.4.0.0^{9,14}]-hexacosane	[*16069-36-6*]	dicyclohexano-18-crown-6	DCH18C6
9	2,5,8,11,18,21,24,27-Octaoxatricyclo[26.4.0.0^{12,17}]-dotriacontane	[*17455-23-1*]	dicyclohexano-24-crown-8	DCH24C8
10	1,4,10-Trioxa-7,13-diazacyclopentadecane	[*31249-95-3*]	diaza-15-crown-5	DA15C5
11	1,4,10,13-Tetraoxa-7,16-diazacyclooctadecane	[*23978-55-4*]	diaza-18-crown-6	DA18C6
12	1,4,7,13,16-Pentaoxa-10,19-diazacycloheneicosane	[*23978-10-1*]	diaza-21-crown-7	DA21C7
13	1,4,10,13-Tetraoxa-7,16-dithiacyclooctadecane	[*296-39-9*]	dithia-18-crown-6	DT18C6
14	1,4-Dioxa-10,13-dithia-7,16-diazacyclooctadecane	[*28843-76-7*]	dithiadiaza-18-crown-6	DTDA18C6
15	19,20,22,23-Tetrahydro-12*H*-7,11-nitrilo-6*H*-dibenzo [*b,k*] [1,4,7,10,13] pentaoxacycloeicosin	[*59945-37-8*]	dibenzopyridino-18-crown-6	DBPy18C6

for benzo derivatives (e.g., **6** 1,4,7,14,17,20-hexaoxa[7.7](1,2)benzenophane) [32].

The crown ethers were first described by C. J. Pedersen [1, 33] in 1967; since then they have been extensively studied by many groups [1 – 11].

2. Physical Properties

Nomenclature and physical data for important crown ethers are given in Tables 1 and 2, respectively.

Crown ethers of the ethylene oxide oligomer type (**1** – **4**) are colorless, odorless, viscous liquids or solids with a low melting point. They are strongly hygroscopic and readily soluble in most organic solvents and in water [6, 11].

Crown ethers with condensed aromatic rings (**5** – **7**) are colorless, barely hygroscopic, crystalline compounds. At room temperature they have poor solubility in water, alcohols, and many other common solvents [6]. They are readily soluble only in halogenated hydrocarbons, pyridine, and formic acid.

Crown ethers with alicyclic bridges (**8**, **9**) have much better solubility in hydrocarbons and much poorer solubility in water than the ethylene oxide oligomers [6].

The solubility of crown ethers increases markedly if salts are added [1, 6, 11], e.g., 25-fold for dibenzo-18-crown-6 (**6**) in MeOH after addition of KF [6]. The increase depends on the type of salt used. Conversely, presence of the crown ether increases the solubility of the salt (see Chap. 4).

3. Chemical Properties

Like simple dialkyl ethers, aliphatic and alicyclic crown ethers are chemically stable [6]. Aromatic crown ethers react like anisole or veratrole, i.e., they can be halogenated or nitrated and they react with formaldehyde [1]. Hydrolysis takes place only in special cases [36, 46]. Crown ethers are also thermally stable; dibenzo-18-crown-6 (**6**) can be distilled at 380 °C without decomposition [39]. With hydrogen ions [47] and in the presence of Lewis acids ($AlCl_3$, $TiCl_3$), oxonium compounds are formed [48]. The hetero derivatives (**10** – **15**) are usually more reactive than the classical crown ethers. Aza analogues are strong bases and react with acids to form salts.

4. Crown Ether Complexes

The most remarkable property of crown ethers is their ability to yield stable complexes with alkali and alkaline-earth metal ions and with ammonium ions [5, 11, 49, 50]; complexes with other

Table 2. Physical properties of crown ethers

Compound no.	Trivial name	M_r	*mp*, °C	*bp*, °C	Ref.
1	12-crown-4	176.21	16 [a]	67 – 70 [b]	[34]
2	15-crown-5	220.27	−32 [c, d]	78 [e]	[26, 34]
3	18-crown-6	264.32	36.5 – 38	116 [f]	[27 – 30, 33, 35 – 37]
4	21-crown-7	308.37		oil	[38]
5	benzo-15-crown-5	268.31	79 – 79.5		[33]
6	dibenzo-18-crown-6	360.41	162.5 – 163.5	380	[1, 38]
7	dibenzo-24-crown-8	448.52	113 – 114		[1]
8	dicyclohexano-18-crown-6	372.50	36 – 56 [g]		[1, 39, 40]
9	dicyclohexano-24-crown-8	460.61	<26 [g]		[1]
10	diaza-15-crown-5	218.30	89 – 90		[41]
11	diaza-18-crown-6	262.35	115 – 116		[41]
12	diaza-21-crown-7	306.40	−10 [c, h]		[41]
13	dithia-18-crown-6	296.44	90 – 91		[42]
14	dithiadiaza-18-crown-6	294.48	47		[43, 44]
15	dibenzopyridino-18-crown-6	393.44	135 – 137		[45]

[a] n_D^{20} 1.4621, d_4^{20} 1.109.
[b] At 0.067 kPa.
[c] Freezing point.
[d] n_D^{20} 1.4615, d_4^{20} 1.113.
[e] At 0.067 kPa.
[f] At 0.267 kPa.
[g] Mixture of stereoisomers (commercial product).
[h] d_4^{20} 1.075.

main group metal, transition metal, lanthanide, and actinide ions are also possible [51], as well as complexes with organometallics [52]. Numerous complexes of crown ethers with nonionic organic molecules are also known [53 – 55].

Crystalline complexes of crown ethers and metal ions (Table 3) are obtained by mixing the two components in a common solvent [1, 6]. Both metal complexes and H-bonded complexes that are stable and strongly bound can also be generated in the gas phase from ion molecule reactions [56]. The solid complexes have higher melting points than the free crown ethers (see Table 1). The stoichiometry of the complexes (crown ether: salt = 1 : 1, 2 : 1, or 1 : 2) is largely determined by the fit of crown ether cavity and cation diameter [57 – 59] (see Table 4). Structure determination by means of X-ray diffraction [51, 57, 60 – 63] has confirmed that in the 1 : 1 complexes the cation is located inside the crown ether ring. For the 2 : 1 complexes, sandwich structures were found. In the 1 : 2 complexes, two cations are located inside the crown ether ring. Complexes in which the cation is surrounded by a cage-like structure have also been described. In all cases, the cations are coordinated to the ether oxygen atoms which point into the ring. Hydrogen bonds are effective in complexes with ammonium ions. Frequently there is also contact with the counterion. Moreover, crown ethers may act as "second-sphere ligands" of metal ion complexes [64 – 66].

In solution, the complexes exhibit varying degrees of stability [2, 44, 57, 67] (Table 5). They are most stable under the following conditions:

- Optimum ratio of cation diameter to cavity diameter
- Optimum number of ether oxygens capable of coordination
- Strongly basic ether oxygens
- High charge density at the cation
- Low polarity of the solvent
- Lipophilic, polarizable anion

Nitrogen (e.g., **11**) and, especially, sulfur (e.g., **13**) decrease the stability of alkali and alkaline earth metal ion complexes. Conversely, complexation with Ag^+ or other heavy-metal ions and with ammonium ions is enhanced [2, 67]. The effect of heterosubstitution on the stability of 18-crown-6 complexes is shown in Table 6. Pyridine rings (e.g., **15**) increase selectivity for sodium ions [67]. Recently, seleno-containing crown compounds were shown to complex C_{60} [68].

Table 3. Crystalline complexes of crown ethers with alkali or alkaline-earth metal salts

Crown ether	Salt	Complex	*mp*, °C (complex)	Ref.
5	NaSCN	**5** · NaSCN	162 – 165	[55]
5	KSCN	$(\mathbf{5})_2$ · KSCN	176	[55]
6	KSCN	**6** · KSCN	245 – 246	[55]
6	CsSCN	$(\mathbf{6})_2$ · CsSCN	146 – 147	[55]
6	$Ba(SCN)_2$	**6** · $Ba(SCN)_2$	> 360	[33]
7	KSCN	**7** · $(KSCN)_2$		[57]

Table 4. Diameters of crown ether cavities and of cations [1, 2]

Crown ether	Diameter, pm	Cation	Diameter, pm
1	120 – 150	Li^+	136
2, **5**	170 – 220	Na^+	194
3	260 – 320	Ba^{2+}	268
6	260 – 320	K^+	266
8	260 – 320	Rb^+	294
4	340 – 430	Cs^+	334

Table 5. Stability constants (log K_s) [a] of crown ether – metal ion complexes in water and methanol [6, 67]

Crown ether	Solvent	Na^+	K^+	Cs^+	Sr^{2+}	Ba^{2+}
15C5 (**2**)	water	0.70	0.74	0.8	1.95	1.71
	methanol	3.48	3.77	2.62	2.63	
18C6 (**3**)	water	0.80	2.03	0.99	2.72	3.87
	methanol	4.36	6.06	4.79	> 5.5	7.04
21C7 (**4**)	methanol	1.73	4.22	5.01	1.77	5.44
DB18C6 (**6**)	water	1.64	0.99	0.79	0.94	1.97
	methanol	4.36 [b]	5.00 [b]	3.55 [b]		4.28
DB24C8 (**7**)	methanol	2.25 [b]	3.60 [b]	3.78 [b]		
DCH18C6 (**8**) [c]	water	1.21	2.02	0.96	3.24	3.57
	methanol	4.08 [c]	6.01 [c]	4.61 [c]		

[a] Determined by calorimetry, unless otherwise stated.
[b] Potentiometrically.
[c] Isomer A [67].

Table 6. Stability constants (log K_s) of hetero crown ether complexes [5]

Cation	18-Crown-6 (**3**)	Diaza-18-crown-6 (**11**)	Dithia-18-crown-6 (**13**)
K^+ [a]	6.06	2.04	1.15
Ag^+ [b]	1.60	7.80	4.34

[a] In methanol.
[b] In water.

The stability constants reflect the ion selectivity of the crown ether [49, 57]. A quantitative measure is the ratio K_{M1}/K_{M2} (M1 = cation 1, M2 = cation 2). Ion selectivity depends to a large extent on the solvent; in certain cases it can even be reversed through a change of solvent. For example, in water, 18-crown-6 (**3**) complexes of barium ions are more stable than those of potassium ions (K_{Ba}/K_K = 1.84), whereas in methanol the potassium complexes are more stable (K_{Ba}/K_K = 0.95) [12]. In multiphase systems that are suitable for selective salt extractions — e.g., H_2O, salts/CH_2Cl_2, crown ether — ion selectivity is determined by the relative degree of salt transfer [69]. Effective extraction requires large anions that are strongly polarizable (e.g., picrate), a highly lipophilic crown ether, and a high stability of the complex in the organic solvent.

The thermodynamic characteristics of a large number of crown ether complexes in various solvents have been determined [49, 57, 67]. The

kinetics of crown ether complexation have also been studied [57, 67, 70].

5. Production

The most important methods for the production of crown ethers are analogous to the Williamson ether synthesis [9 – 12, 71 – 73]. The ether rings are formed by reaction of diols with bifunctional components that possess terminal leaving groups such as chloro or tosylate (Eq. 1):

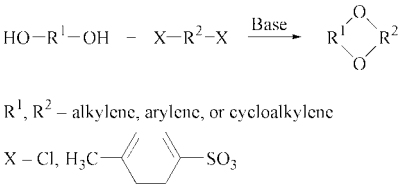

(1)

The reaction is carried out in the presence of a base, e.g., an alkali metal hydroxide or carbonate, cesium fluoride, potassium *tert*-butoxide, or sodium hydride. *n*-Butanol, *tert*-butanol, tetrahydrofuran (THF), and, less frequently, dimethylformamide (DMF) or dimethyl sulfoxide (DMSO) are used as solvents. Cyclization can take place in a single step (Eq. 1) or in several steps:

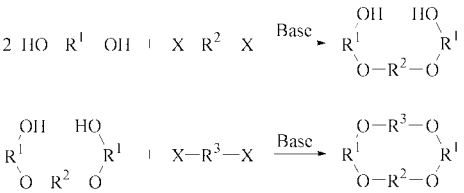

(2)

Protecting groups may be necessary. A one-component method in which the leaving group X is produced from the diol in an intermediate step is also known [6]:

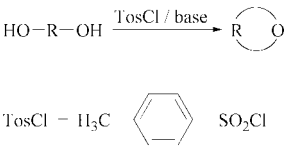

(3)

A template effect, i.e., enhancement of yield and reaction rate by cations that are dimensionally compatible with the crown ether being produced, is often observed [74].

A different production method is based on the cyclooligomerization of oxirane [72, 75]:

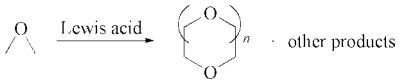

(4)

This also involves template participation [74]. The oxirane method is suitable only for the production of unsubstituted aliphatic crown ethers of a limited range of ring sizes, e.g., 12-crown-4 (**1**), 15-crown-5 (**2**), and 18-crown-6 (**3**). The Williamson procedure is generally applicable and yields both aliphatic and aromatic crown ethers [71 – 73]. Crown ether sulfides are produced similarly (nucleophilic substitution by a thiol group) [13, 14, 71]:

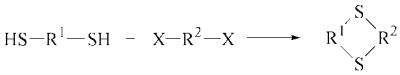

(5)

Synthesis of crown ether amines proceeds via a cyclic diamide or a cyclic bis(tosylamide) [16, 72, 73]. Crown ethers with cyclohexane bridges are obtained through catalytic hydrogenation of the appropriate aromatic compounds [76].

Commercial Products. Most of the crown ethers **1** – **15** are commercially available (Chap. 1); the preparations are usually of > 98 % or > 99 % purity (specification by gas chromatography). Cyclohexano derivatives contain up to 2 % of the corresponding benzo crown ether. The most important suppliers are ABCR, Acros Organics, Alfa Aesar, Avocado, Fluka, ICN, Lancaster Synthesis, Merck, Sigma Aldrich, Strem Chemicals, and TCI [73].

6. Uses

Crown ethers are useful in many fields of application [77 – 79]. Important uses are as complexing agents and phase-transfer catalysts in *organic syntheses* [5, 80 – 84]. They solubilize inorganic salts that are needed as reactants in organic solvents. This amounts to activation of the anion (reagent) [85 – 87]. The main advantages

as compared with phase-transfer catalysts such as quaternary ammonium salts are the greater activating effect on the anion (naked ions) and the absence of foreign ions, the crown ether being neutral. In addition, phase transfer directly from the crystal into the organic phase (solid – liquid phase transfer) is usually possible.

Table 7. Applications of crown ethers in organic and inorganic syntheses

Oxidation (purple benzene)[88]:

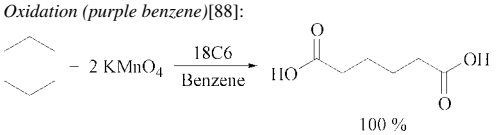

Hydrolysis (naked hydroxyl) [1]:

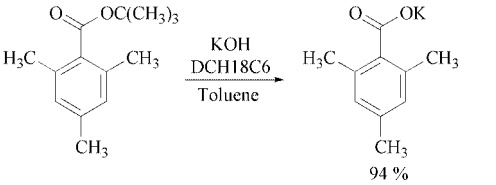

Substitution (naked fluoride) [89]:

$$H_3C(CH_2)_6CH_2Br + KF \xrightarrow[CH_3CN]{DB18C6} H_3C(CH_2)_6CH_2F$$

92 %

Carbene synthesis [90]

$$H_2N\ NH_2 + KOH \xrightarrow[CHCl_3]{18C6} CH_2N_2 \xrightarrow[-N_2]{} CH_2$$

48 %

Formation of carbon – carbon bonds [91]:

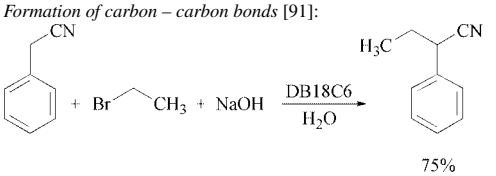

Rearrangement [92]:

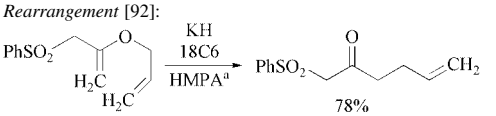

Ligand exchange [93]:

$$[(CH_3CN)WCl_4(NCl)] + KF \xrightarrow{18C6} [K(18C6)]\,[WF_5(NCl)]$$

Pentacoordinate silicon [94]

$$Ph_3SiF - KF \xrightarrow[THF]{18C6} Ph_3SiF_2\ K^+\ (18C6)$$

[a] HMPA = Hexamethylphosphoric triamide

Table 7 illustrates the preparative usefulness of crown ethers. Modern laboratory synthesis, including preparation of natural compounds, would be unthinkable without crown ethers [87]. The industrial application is especially in the manufacture of polymers and polycondensates [95 – 97]. In principle, any reaction involving the participation of ions can be influenced by crown ethers [5, 77].

Another important field of application is in *chemical analysis* [98 – 102]. The ion selectivity of crown ethers permits the concentration, separation, and masking of ions and the determination of ion concentrations. Suitable methods include ion extraction [69], ion chromatography [104 – 106], capillary electrophoresis [104], and membrane transport [99, 107]. It is possible in this way to separate traces of Sr^{2+} from an excess of Ca^{2+}, and even $^{44}Ca^{2+}$ from $^{40}Ca^{2+}$ or other metal ion isotopes [108]. Heterosubstituted crown ethers are mainly used in the analysis of transition-metal ions [100]; they are suitable for the masking and decontamination of environmentally and physiologically damaging heavy-metal ions [109], including separation of cationic fission products from nuclear waste [110]. A microassay of Na^+ and K^+ in blood and urine has also been developed [98, 111]. Crown compounds for the stereognostic coordination of rare metals have been designed [112]. Many potentiometric chemical [111, 113, 114] and optical sensors (luminescent and chromogenic probes) [115 – 117] for metal ions are based on crown compounds.

Recent developments in the area of anion recognition and sensing have produced a variety of receptors for anions, some containing crown subunits, and have been used to separate anions [118]. Heterotopic receptors based on crown compounds with specific cation and anion binding sites that behave as recognition systems for salts (ion pair binding) have been prepared and used [119, 120].

Moreover, crown ethers can be used in developing effective sorbents for separation of organic compounds [53, 55, 121]; chiral derivatives of crown ethers [23, 122, 123] result in racemate splitting (optical resolution) [98, 101].

Aside from these applications in chemical synthesis and analysis, crown ethers can be used for imaging and targeting in the diagnosis of diseases [124], for molecular packing, storage, protection and stabilization of compounds [125], as aids in detergent and surfactant industries, in electroorganic synthesis and galvanizing, in the manufacture of organic conductors and anticorrosives [77], as well as in the preparation of liquid-crystalline phases [126] and Langmuir or Langmuir-Blodgett films [127]. Sensors based

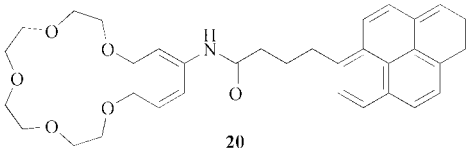

20

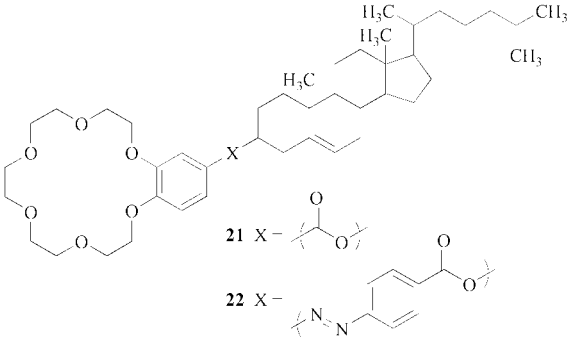

21 X –

22 X –

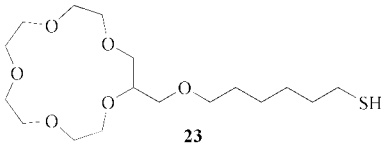

23

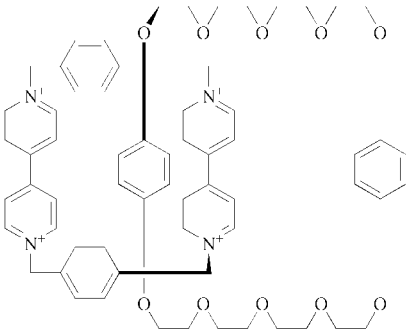

24

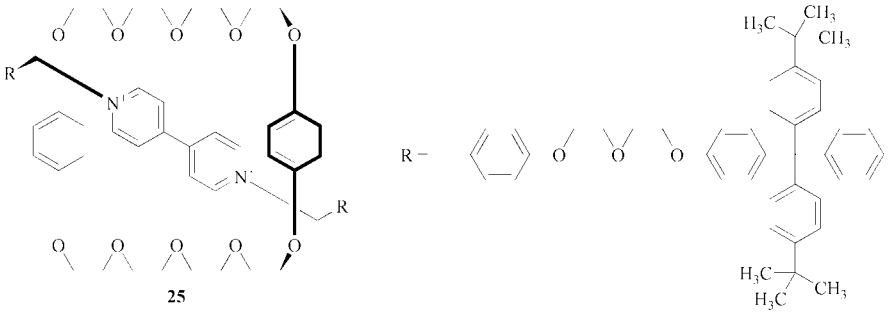

25

Figure 1. Luminescent probe (**20**), liquid crystalline (**21**), organo gelator (**22**), and SAM-forming (**23**) crown-type compounds. Prototype [2]catenane (**24**) and [2]rotaxane (**25**) structures contain crown ethers as an assembly group

Table 8. Toxicological data for crown ethers [6]

Crown	Acute toxicity,		
	oral		dermal
12C4 (**1**)	LD_{50}	3.15 ± 0.06 g/kg [a]	
15C5 (**2**)	LD_{50}	1.02 ± 0.1 g/kg [a]	
18C6 (**3**)	LD_{50}	0.705 ± 0.08 g/kg [a]	LD_{50} 200 mg/kg [b]
DB18C6 (**6**)	LD_{50}	> 300 mg/kg [a]	ALD [e] 7.50 g/kg [c]
	LD_{50}	11 g/kg [b]	
DB24C8 (**7**)	LD_{50}	> 300 mg/kg [a]	
DCH18C6 (**8**)	LD_{50}	> 300 mg/kg [a]	ALD [e] 130 mg/kg [d]
	ALD [e]	300 mg/kg [b]	

[a] Mouse.
[b] Rat.
[c] Guinea pig.
[d] Rabbit.
[e] ALD = approximate lethal dose.

on self-assembled monolayers (SAMs) that contain crown ether subunits [128] or monolayer- and mixed-monolayer-protected clusters (MPCs and MMPCs) involving crown compounds [129] are promising fields of application. Fullerene [130, 131] and dendrimer structures [132, 133] have been combined with crown ethers in order to modify their properties. Moreover, crown-appended gelators have been utilized as building blocks in organogels [134] and as templates for sol-gel transcription [135]. Crown-type compounds have also pioneered new photochemical and photophysical devices [136], including molecular machines [137], e.g., based on catenanes and rotaxanes [138], and are promising in many other technologies emerging from supramolecular chemistry [78]. Some exemplary structures relating to these compound types that involve crown ether subunits are shown in Figure 1.

With reference to life, crown compounds have been explored in a number of biological contexts [139]. They have shown activity in toxicity studies to bacteria [140] and cancer cell lines [141]. Others were found being able to intercalate into DNA and to protect the double helix from enzymatic cleavage [142] or show activity against AIDS related lymphoma cell lines [143]. Certain crown ethers were found to accelerate the reaction rate of lipase enzymes in organic solvents [144] or were shown to form channels in the bilayer of living human embryonic kidney cells and may thus hold potential under the developing concept known as "channel replacement therapy" [145].

7. Toxicology and Occupational Health

The oral and skin toxicity of crown ethers is much greater than that of ordinary ethers [9, 146]. As complexing agents they cause damage to the body's ion balance [109]. The main effects are therefore on the central nervous system [6]. Typical sublethal intoxication phenomena include rapid breathing, nervous tension, lack of coordination, tremor, and lassitude. Many crown ethers also damage the eyes and skin [39]. Lethal toxicity data for common crown ethers can be found in Table 8.

Protective clothing and goggles must be worn when crown ethers are handled [146]. Contact, especially with the eyes, should be avoided. The danger of intoxication by inhalation is greatest for 12-crown-4 [6, 148].

According to the RID and ADR transport regulations, crown ethers fall into class 6.1. Incineration is prescribed for the disposal of residues and waste.

8. References

1. C. J. Pedersen, H. K. Frensdorff, *Angew. Chem.* **84** (1972) 16–26; *Angew. Chem. Int. Ed. Engl.* **11** (1972) 16–26.
2. J. J. Christensen, D. J. Eatough, R. M. Izatt, *Chem. Rev.* **74** (1974) 351–384.
3. R. M. Izatt, J. J. Christensen (eds.): *Synthetic Multidentate Macrocyclic Compounds*, Academic Press, New York–San Francisco–London 1978.

4. R. M. Izatt, J. J. Christensen (eds.): *Progress in Macrocyclic Chemistry*, vols. **1–3**, Wiley, New York 1979, 1981, 1987.
5. E. Weber, F. Vögtle, *Top. Curr. Chem.* **98** (1981) 1–41.
6. M. Hiraoka: *Crown Compounds*, Their Characteristics and Applications, Studies in Organic Chemistry 12, Elsevier, Amsterdam–Oxford–New York 1982.
7. F. Vögtle, E. Weber (eds.): *Host Guest Complex Chemistry — Macrocycles: Synthesis, Structures, Applications*, Springer Verlag, Berlin–Heidelberg–New York–Tokyo 1985.
8. E. Weber, *et al.*, in: S. Patai, Z. Rappoport (eds.): *Crown Ethers and Analogs*, John Wiley & Sons, Chichester 1989.
9. G. W. Gokel: "Crown Ethers and Cryptands," *Monographs in Supramolecular Chemistry*, vol. **3**, The Royal Society of Chemistry, Cambridge 1991.
10. G. W. Gokel (ed.): *Comprehensive Supramolecular Chemistry*, vol. **1**, Elsevier, Oxford 1996.
11. G. W., Gokel, in: J. Atwood, J. Steed (eds.), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York 2004, pp. 326–333.
12. J. S. Bradshaw, R. M. Izatt, A. V. Bordunov, C. Y. Zhu, J. K. Hathaway in [10], pp. 35–95.
13. J. S. Bradshaw, J. Y. K. Hui, *J. Heterocycl. Chem.* **11** (1974) 649–674.
14. J. S. Bradshaw, P. E. Stott, *Tetrahedron* **36** (1989) 461–510.
15. L. F. Lindoy: *The Chemistry of Macrocyclic Ligand Complexes*, Cambridge University Press, Cambridge 1989.
16. J. S. Bradshaw, K. E. Krakowiak, R. M. Izatt: "Aza-Crown Macrocycles," *The Chemistry of Heterocyclic Compounds*, vol. **51**, Wiley, New York 1993.
17. G.-M. Gray, in: *Comments Inorg. Chem.*, vol. **17**, "No. 2", OPA, Amsterdam 1995.
18. I. Haiduc, F. T. Edelmann: *Supramolecular Organometallic Chemistry*, Wiley-VCH, Weinheim 1999.
19. E. Weber, in: J. Atwood, J. Steed (eds.), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York 2004, pp. 261–273.
20. E. Weber in [8], pp. 305–357.
21. G. W. Gokel, O. F. Schall in [10], pp. 97–152.
22. G. W. Gokel, in: J. Atwood, J. Steed (eds.), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York 2004, pp. 782–790.
23. D. J. Cram, J. M. Cram: "Container Molecules and their Guests," *Monographs in Supramolecular Chemistry*, vol. **4**, The Royal Society of Chemistry, Cambridge 1994.
24. E. Maverick, D. J. Cram in [10], pp. 213–243.
25. F. Vögtle, E. Weber, *Angew. Chem.* **91** (1979) 813–837;
    Angew. Chem. Int. Ed. Engl. **18** (1979) 753–776.
26. G. W. Gokel, O. Morillo in [10], pp. 1–33.
27. E. Weber, in: J. Atwood, J. Steed (eds.), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York 2004, pp. 1106–1119.
28. J. M. Lehn, *Acc. Chem. Res.* **11** (1978) 49–75.
29. B. Dietrich in [10], pp. 153–211.
30. B. Dietrich, in: J. Atwood, J. Steed (eds.), *Encyclopedia of Supramolecular Chemistry*, Marcel Dekker, New York 2004, pp. 334–339.
31. E. Weber, F. Vögtle, *Inorg. Chim. Acta* **45** (1980) L65–L67.
32. F. Vögtle, P. Neumann, *Tetrahedron* **26** (1970) 5847–5873.
33. C. J. Pedersen, *J. Am. Chem. Soc.* **89** (1967) 7017–7036.
34. F. L. Cook *et al.*, *Tetrahedron Lett.* **1974**, 4029–4032.
35. DuPont, US 3 562 295, 1971 (C. J. Pedersen).
36. G. Johns, C. J. Ransom, C. B. Reese, *Synthesis* **1976**, 515–516.
37. G. W. Gokel *et al.*, *Org. Synth.* **57** (1977) 30–33.
38. J. Dale, P. O. Kristiansen, *Acta Chem. Scand.* **26** (1972) 1471–1478.
39. C. J. Pedersen, *Org. Synth.* **52** (1972) 66–74;
40. DuPont, US 3 687 978, 1972 (C. J. Pedersen).
41. B. Dietrich, J. M. Lehn, J. P. Sauvage, J. Blanzat, *Tetrahedron* **29** (1973) 1629–1645.
42. J. R. Dann, P. P. Chiesa, J. W. Gates, Jr., *J. Org. Chem.* **26** (1961) 1991–1995.
43. B. Dietrich, J. M. Lehn, J. P. Sauvage, *J. Chem. Soc., Chem. Commun.* **1970**, 1055–1056.
44. Schering Corp., US 3 888 877, 1975 (J. M. Lehn).
45. E. Weber, F. Vögtle, *Chem. Ber.* **109** (1976) 1803–1831.
46. B. Agai *et al.*, *Tetrahedron Lett.* **26** (1985) 2705–2708.
47. G. S. Heo, R. A. Bartsch, *J. Org. Chem.* **47** (1982) 3557–3559.
48. S. G. Bott, U. Kynast, J. L. Atwood, *J. Incl. Phenom.* **4** (1986) 241–246.
49. Y. Inoue, G. W. Gokel (eds.): *Cation Binding by Macrocycles*, Marcel Dekker, New York 1990.
50. G. W. Gokel, A. Abel in [10], pp. 511–535.
51. R. D. Rogers, C. B. Bauer in [10], pp. 315–355.

52. H. G. Richey, jr. in [10], pp. 755–775.
53. F. Vögtle, W. M. Müller, W. H. Watson, *Top. Curr. Chem.* **125** (1984) 131–164.
54. E. Weber in [4], vol. 3 (1987) 337–419.
55. E. Weber, F. Vögtle, in F. Vögtle (ed.): *Comprehensive Supramolecular Chemistry*, vol. **2**, Elsevier, Oxford 1996, pp. 1–28.
56. J. S. Brodbelt in [10], pp. 465–475.
57. F. Vögtle, E. Weber in [8], pp. 207–304.
58. C. J. Pedersen, *J. Am. Chem. Soc.* **92** (1970) 386–391.
59. N. S. Poonia, A. V. Bajaj, *Chem. Rev.* **79** (1979) 389–445.
60. D. E. Fenton, M. Mercer, N. S. Poonia, M. R. Truter, *J. Chem. Soc., Chem. Commun.* **1972**, 66–67.
61. M. Dobler: *Ionophores and their Structures*, John Wiley & Sons, New York 1981.
62. R. Hilgenfeld, W. Saenger, *Top. Curr. Chem.* **101** (1982) 1–82.
63. I. Goldberg in [8], pp. 399–476.
64. H. M. Colquhoun, J. F. Stoddart, D. J. Wiliams, *Angew. Chem.* **98** (1986) 483–503; *Angew. Chem. Int. Ed.* **25** (1986) 487–507.
65. E. Weber in [8], pp. 305–357.
66. S. J. Loeb in [10], pp. 733–753.
67. R. M. Izatt *et al.*, *Chem. Rev.* **85** (1985) 271–339.
68. Y. Liu, J.-R. Han, Y.-I. Zhao, H.-Y. Zhang, Z.-Y. Duan, *J. Incl. Phenom.* **51** (2005) 191–198.
69. B. A. Moyer in [10], pp. 377–416.
70. W. Burgermeister, R. Winkler-Oswatitsch, *Top. Curr. Chem.* **69** (1977) 91–196.
71. G. W. Gokel, S. H. Korzeniowski: *Macrocyclic Polyether Synthesis*, Springer Verlag, Berlin–Heidelberg–New York 1982.
72. B. Dietrich, P. Viout, J. M. Lehn: *Macrocyclic Chemistry*, VCH Verlagsgesellschaft, Weinheim 1993.
73. D. Parker (ed.): *Macrocycle Synthesis*, Oxford University Press, Oxford 1996.
74. N. V. Gerbeleu, V. B. Arion, J. Burgess: *Template Synthesis of Macrocycle Compounds*, Wiley-VCH, Weinheim 1999.
75. CH 601 284, 1977 (J. Dale, K. Daasvatu).
76. C. J. Pedersen, *Org. Synth.* **52** (1972) 66–74.
77. E. Weber, in: *Phase Transfer Catalysts: Properties and Applications*, Merck-Schuchardt, Darmstadt 1987, pp. 33–76.
78. D. N. Reinhoudt (ed.): *Comprehensive Supramolecular Chemistry*, vol. **10**, Elsevier, Oxford 1996.
79. K. Gloe (ed.): *Macrocyclic Chemistry*, Springer, Dordrecht 2005.
80. G. W. Gokel, G. W. Weber: *Phase Transfer Catalysis in Organic Synthesis*, Springer Verlag, Berlin–Heidelberg–New York 1977.
81. E. Weber, F. Vögtle, *Kontakte (Merck)* (1977) no. 3, 36–48.
82. E. Weber, *Kontakte (Merck)* (1981) no. 1, 24–35.
83. F. Montanari, S. Quici, S. Banfi in [78], pp. 389–416.
84. V. K. Ahluwalia, R. Aggarwal: *Organic Synthesis*, Special Techniques, Alpha Science, Oxford 2001.
85. F. Vögtle, E. Weber, *Kontakte (Merck)* (1977) no. 2, 16–28.
86. D. Landini, A. Maia, M. Penso in [10], pp. 417–464.
87. B. Dietrich in [78], pp. 361–387.
88. D. J. Sam, H. E. Simmons, *J. Am. Chem. Soc.* **94** (1972) 4024–4025.
89. C. I. Liotta, H. P. Harris, *J. Am. Chem. Soc.* **96** (1974) 2250–4252.
90. D. T. Sepp, W. P. Weber, *Tetrahedron Lett.* (1974) 2983–2984.
91. M. Makosza, M. Ludwikow, *Angew. Chem.* **86** (1974) 744–745; *Angew. Chem. Int. Ed. Engl.* **13** (1974) 665–666.
92. S. E. Denmark, M. A. Harmata, *J. Am. Chem. Soc.* **104** (1982) 4972–4974.
93. H. Stenger, K. Dehnicke, W. Hiller, *Z. Naturforsch.* **47b** (1992) 1054–1056.
94. R. Damrauer, S. E. Danahey, *Organometallics* **5** (1986) 1490–1494.
95. S. L. Regen, *Angew. Chem.* **91** (1979) 464–472; *Angew. Chem. Int. Ed. Engl.* **18** (1979) 421–429.
96. L. J. Mathias, C. E. Carraher, Jr. (eds.): "Crown Ethers and Phase Transfer Catalysis in Polymer Science," *Polymer Science and Technology*, vol. **24**, Plenum Press, New York–London 1984.
97. S. Bywater, *Prog. Polym. Sci.* **19** (1994) 287–316.
98. E. Weber, *Kontakte (Merck)* (1978) no. 2, 16–29.
99. E. Weber, *Kontakte (Merck)* (1984) no. 1, 26–34.
100. I. M. Kolthoff, *Anal. Chem.* **51** (1979) 1R–22R.
101. Yu. A. Zolotov (ed.): *Macrocyclic Compounds in Analytical Chemistry*, John Wiley & Sons, New York 1997.

102. S. M. Khopkar: *Analytical Chemistry of Macrocyclic and Supramolecular Compounds*, Alpha Science, Oxford 2005.
103. R. M. Izatt, J. S. Bradshaw, R. L. Bruening, B. J. Tarbet, M. L. Bruening in [78], pp. 1–11.
104. J. D. Lamb, R. G. Smith in [78], pp. 79–112.
105. N. A. Pasekova, E. I. Morosanova in [100], pp. 209–283.
106. J. O. Lamb, J. S. Gardner in [78], pp. 349–363.
107. F. d. Jong, H. C. Visser in [78], pp. 13–51.
108. L. Echegoyen, R. C. Lawson in [78], pp. 151–170.
109. R. M. Izatt, *et al.*, in E. J. Ariens (ed.): *Medicinal Chemistry*, vol. **11-VIII**, Drug Design, Academic Press, New York–San Francisco–London 1979, pp. 355–400.
110. B. A. Moyer, J. F. Birdwell, Jr., P. V. Bonnesen, L. H. Delman in [79], pp. 383–405.
111. U. E. Spichiger-Keller: *Chemical Sensors and Biosensors for Medical and Biological Applications*, Wiley-VCH, Weinheim 1998.
112. R. E. Powers, W. L. Fuller, K. N. Raymond in [78], pp. 537–555.
113. Z. Brzozka in [78], pp. 187–212.
114. L. K. Shpigun, R. D. Tsingarelli in [101], pp. 285–357.
115. A. P. de Silva *et al.*, *Chem. Rev.* **97** (1997) 1515–1566.
116. M. Pietraszkiewicz in [78], pp. 225–266.
117. V. A. Bren, *Russ. Chem. Rev.* **70** (2001) 1017–1036.
118. B. A. Moyer, R. P. Singh (eds.): *Fundamentals and Applications of Anion Separations*, Kluwer/Plenum, New York 2004.
119. P. A. Gale, *Coord. Chem. Rev.* **240** (2003) 191–221.
120. J. M. Mahoney, J. P. Davis, B. D. Smith in [118], pp. 115–123.
121. I. V. Pletnev, I. I. Torocheshnikova in [101], pp. 359–404.
122. D. J. Cram, J. M. Cram, *Acc. Chem. Res.* **11** (1978) 8–14.
123. J. F. Stoddart, *Top. Stereochem.* **17** (1987) 207–288.
124. D. Parker in [78], pp. 487–536.
125. D. D. MacNicol, S. J. Rowan in [78], pp. 417–428.
126. E. Dalcanale in [78], pp. 583–635.
127. I. K. Lednev, M. C. Petty, *Adv. Mater.* **8** (1996) 615–630.
128. B. Flink, F. C. J. M. van Veggel, D. N. Reinhoudt, *Adv. Mater.* **12** (2000) 1315–1328.
129. U. Drechsler, B. Erdogan, V. M. Rotello, *Chem. Eur. J.* **10** (2004) 5570–5579.
130. A. Hirsch, M. Brettreich: *Fullerenes*, Wiley-VCH, Weinheim 2005.
131. L. Echegoyen, M. A. Herranz, F. Diederich, C. Thilgen, in F. Toda, R. Bishop (eds.): *Separations and Reactions in Organic Supramolecular Chemistry*, Perspectives in Supramolecular Chemistry, vol. 8, Wiley, Chichester 2004, pp. 137–171.
132. G. R. Newkome, C. N. Moorefield, F. Vögtle: *Dendrimers and Dendrons*, Wiley-VCH, Weinheim 2001.
133. J. M. J. Fréchet, D. A. Tomalia (eds.): *Dendrimers and Other Dendritic Polymers*, Wiley, New York 2001.
134. D. J. Abdallah, R. G. Weiss, *Adv. Mater.* **12** (2000) 1237–1247.
135. K. J. C. van Bommel, A. Friggeri, S. Shinkai, *Angew. Chem.* **115** (2003) 1010–1030; *Angew. Chem. Int. Ed.* **42** (2003) 980–999.
136. V. Balzani, F. Scandola in [78], pp. 687–746.
137. V. Balzani, M. Venturi, A. Credi: *Molecular Devices and Machines*, Wiley-VCH, Weinheim 2003.
138. J.-P. Sauvage, C. O. Dietrich-Buchecker (eds.): *Molecular Catenanes, Rotaxanes and Knots*, Wiley-VCH, Weinheim 1999.
139. G. W. Gokel, W. M. Leevy, M. E. Weber in [79], pp. 253–265.
140. K. Yagi, V. Garcia, M. E. Rivas, J. Sales, A. Camargo, T. Tabata, *J. Incl. Phenom.* **2** (1984) 179–184.
141. M. M. McPee, S. M. Kerwin, *Bioorg. Med. Chem.* **9** (2001) 2809–2818.
142. A. Basak, H. Dugas, *Tetrahedron Lett.* **27** (1986) 3–6.
143. K. Brandt, R. Kruszynski, T. J. Bartczak, I. Porwolik-Czomperlik, *Inorg. Chim. Acta* **322** (2001) 138–144.
144. H. Tsukube, T. Yamada, S. Shinoda, *J. Heterocycl. Chem.* **38** (2001) 1401–1408.
145. G. A. Cook, et.al., *Biophys. J.* **86** (2004) 1424–1435.
146. M. Hiraoka: *Oligomer Handbook*, Kagakukogyo Nippo, 1977, p. 240.
147. E. I. du Pont de Nemours Co., Elastomer Area-Chambers Works, Chemical Hazard Sheet, Chemical No. 148, Sept. 1 (1972).
148. B. K. J. Leong, T. O. T. Ts'o, M. B. Chenoweth, *Toxicol. Appl. Pharmacol.* **27** (1974) 342–354.

Crystal Growth

John C. Brice, Philips Research Laboratories, Redhill, Surrey, United Kingdom
Peter Rudolph, Institut für Kristallzüchtung, Berlin, Federal Republic of Germany

Related Articles Industrial Crystallization is treated in → Crystallization and Precipitation

1. Introduction 2
1.1. Historical Aspects 2
1.2. Uses of Single Crystals 4
1.3. Specification of Crystals 5
1.4. Commercial Aspects 6
1.5. Sources of Information 7
2. Thermodynamic Aspects 8
2.1. Free Energy and Driving Force of Crystallization 8
2.2. Phase Relations for Major Components 9
2.3. Phase Relations for Minor Components 11
2.4. Interfacial Effects 12
2.5. Shortcomings of the Thermodynamic Approach 13
3. Growth Kinetics 13
3.1. Atomic Structure of Growing Interfaces 13
3.2. Rough Faces 14
3.3. Perfect Singular Faces 14
3.4. Imperfect Singular Faces 15
4. Transport Effects 17
4.1. Transport Processes 17
4.2. Boundary Layers 18
4.3. Mass Flow of Major Components . . . 21
4.4. Mass Flow of Minor Components . . . 22
4.5. Heat Flow 23
4.6. Interface Stability 25
5. Practical Considerations 25
5.1. Raw Materials 25
5.2. Temperature Control 26
5.3. Containers and Atmospheres 26
5.4. Selection and Optimization of Methods 27
5.5. Characterization 28
6. Melt Growth Techniques 29
6.1. Crystal Pulling 29
6.2. Bridgman Method 33
6.3. Skull Melting Methods 35
6.4. Zone Melting 35
6.5. Verneuil Process 38
7. Solution Growth Techniques 38
7.1. General Aspects 38
7.2. Low-Temperature Methods 39
7.3. High-Temperature Methods 40
7.4. Liquid-Phase Epitaxy 42
7.5. Electrolytic Methods 43
7.6. Hydrothermal Growth 44
8. Vapor Growth Techniques 44
8.1. Sublimation and Evaporation Methods 44
8.2. Molecular Beam Epitaxy 45
8.3. Chemical Vapor Transport 46
8.4. Vapor-Liquid-Solid Growth 47
9. Solid-Phase Growth Techniques 48
9.1. Strain–Anneal Method 48
9.2. Methods Not Involving Applied Strain 48
9.3. Solid-Phase Epitaxy 49
9.4. Gel Growth 49
10. Trends 49
11. References 50

This article summarizes the theory and practice of the growth of single crystals. Today, the growth of crystals having dimensions of some hundred millimeters in diameter and weights over 200 kg under extreme pure conditions is well-matured in industrial scale. The major user of the crystals is the electronics industry.Growth of crystals demands theoretical knowledge of thermodynamics, kinetics, and transport processes. In practice, some general aspects have to be considered, including choice of raw materials, conditions, methods, and characterization. The industrial and laboratory techniques of crystal growth are discussed. Economical aspects dominate the developments in the future, that is, growing of larger crystals at higher quality to raise yields and device performance. Furthermore, the demand for tighter specifications and increasing range of materials and uses will require more sophisticated methods and an extensive use of automation.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_099.pub2

Symbols:

a	lattice constant, crystal radius
A	radius of fluid, constant
b	Burger's vector
B	magnetic induction
c	concentration
C	specific heat
D	diffusion coefficient
f	growth rate
g	acceleration due to gravity
G	free energy, shear modulus
h	height of nucleus, heat transfer coefficient
H	enthalpy, depth
J	flux
k	segregation coefficient
K	thermal conductivity
L	face size
N_{d}	dislocation density
Q	addition rate of particles
S	entropy, radial spacing
t	time
T	temperature
u	number of nearest neighbors
V_{m}	molar volume
v	velocity of growth face, flow velocity
w	number of nearest neighbors on surface
x	atom fraction
α	expansion coefficient, crystal face parameter, ratio of spacing
β	volume expansion coefficient
γ	surface tension
δ	thickness of boundary layer
η	dynamic viscosity
$\varkappa$	thermal diffusivity
ν	kinematic viscosity
λ	layer spacing
ρ	mass density
σ	relative supersaturation, strain
τ	time constant
ω	rotation rate
∇^2	Laplacian operator $\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)$

1. Introduction

This article summarizes the theory and practice of the growth of single crystals. Consideration is restricted to cases in which the ultimate product exploits *crystallinity; industrial crystallization,* which uses the crystallization process to purify or to obtain a product in an acceptable form (e.g., granulated sugar), is excluded here, but treated in → Crystallization and Precipitation.

1.1. Historical Aspects

Crystals have been grown on a commercially significant scale since about 1950. Before this, almost the only significant activity was the growth of rubies for watch and clock bearings and, after 1940, defense-related activities (e.g., piezoelectric crystals for sonar transducers). However, the origins of the technology occurred much earlier.

According to references [1 – 4] and the sources quoted, the use of crystallization for the purification of salt can be traced back to 2700 B.C., growth from aqueous solution was being studied systematically in 1600 A.D. and high-temperature solution growth in 1800; and by about 1850, melt growth was a process of scientific interest.

Commercial exploitation on a small scale started around 1900 with the "Geneva" rubies grown by the Verneuil process (see Section 6.5), first used for jewelry and later for bearings in watches and similar instruments. Until quite recently, there was a reluctance to use artificially grown crystals as jewels; now, however, about half the rubies used as jewels are synthetic, and materials which do not exist naturally are common; cubic zirconia, for example, is in many ways a better gemstone than diamond.

VERNEUIL's first paper [5] on the method that carries his name was published in 1902; the second, more informative paper [6] appeared in 1904. During the next 44 years (up to the invention of the transistor in 1948), many of the methods discussed in Chapters 6, 7, 8 and 9 were developed. Thus, SPEZIA's work on hydrothermal growth [7 – 9] was published between 1905 and 1908. The foundations of growth from the solid were laid in 1912 [10] and 1918 [11]. Also, in 1918 CZOCHRALSKI's paper [12] on crystal pulling from the melt was published. Before that the growth of crystals from the top into the melt was introduced in 1916 by NACKEN [13]. A similar technique with seed selection was then developed by KYROPOULOS in 1926 [14]. The decomposition of halides was used to grow crystals from the vapor in 1922 [15]. After TAMMANN [16] in 1914 studied for the first time the

unidirectional solidification of metals in small vertical tubes, the crystallization of bulky melts in containers was described in 1925 by Bridgman [17] and Stober [18], and the technique was improved in 1936 by Stockbarger [19] using a two-zone furnace. Zone melting of metals was initiated in 1928 by Kapitza [20].

Over the same period, much of the theoretical base of crystal growth was also laid. Thus, in 1904, Nernst [21] applied boundary layer ideas to crystal growth, and the basic ideas of the kinetics of the growth of perfect crystals were formulated in 1922–1928 by Volmer, Klossel, and Stranski [22 – 24]. Other theoretical aspects were treated by Gibbs, Curie, Wulff, and Kaishev.

The year 1949 can be regarded as a watershed [25]; reference [26] gives an account of the state of knowledge at that time and the motivation of workers in the field, who were mainly engaged in scientific pursuits, i.e., studying crystals and their growth or using crystals in scientific instruments. The synthetic crystals of interest in 1949 were mostly semiconductors (especially germanium). After 1949, most workers were employed directly or indirectly because crystals could be used in commercial products. As a consequence, at this time the Czochralski method, being then as now the most productive bulk crystal growth method (Section 8.1), was extended to large-scale production. In 1950 Teal and Little [27] started the pulling of relatively large germanium crystals for the first transistors at Bell Laboratories. Shortly after, in 1952, Teal and Buehler [28] adapted the Czochralski method for production of silicon crystals. The crystals were real (i.e., dirty, strained, and inhomogeneous as compared to ideal crystals), and these imperfections had to be controlled. For many purposes, one impurity atom per million host lattice atoms is of no importance; however, in a semiconductor device, unwanted impurity atoms at this level can have catastrophic consequences.

Semiconductor devices are made from crystals containing specified amounts of desired impurities (dopants) and preferably no other impurities at a detectable level.

The need for purity led to the invention of zone refining [29] and the floating zone process [30, 31]. Zone refining is a process in which a molten zone is passed along a bar. Impurities are usually more soluble in the melt than in the solid, so the zone tends to carry the impurities with it. Passing a succession of zones can produce extreme purity: levels of 1 impurity atom in 10^{10} host atoms have been claimed, but measurement of impurity contents below 1 in 10^8 is difficult. For details of zone refining, see Section 6.4 and [32].

Another impurity-related problem involves the desired distribution of dopants. In 1953 it was shown [33, 34] that in many situations the amount of impurity incorporated depends on the phase diagram, the growth rate, and the thickness of the boundary layer at the growth face. The general relation between impurity concentration and growth rate is still an unsolved problem (see Chap. 4).

The causes of strain and crystallographic imperfections are now known and, at a price, certain crystals can be grown without dislocations and with small (say $< 10^{-5}$) built-in strain.

Chapters 2, 3, 4 of this article summarize the theoretical problems solved in the last few decades. Chapters 5, 6, 7, 8, 9 discuss practical techniques, which have changed markedly. For example, in 1949 crystal sizes were typically a few millimeters. Current linear dimensions are 100 times larger. In 1949, a temperature of 1000 °C controlled to $\pm$ 1 °C represented good technology; today, control to $\pm$ 0.1 °C is routine, and good technology can improve this by a factor of at least five. In 1949, raw materials containing less than 10–100 ppm of impurities were rare and expensive; today, levels over 1 ppm are unacceptable for many applications. To maintain high purity, clean ambients (e.g., laminar flow cabinets if not full clean room facilities) are common. Meanwhile, the growth of crystals with dimensions of some hundred millimeters in diameter and weights over 200 kg under extreme pure conditions is well-matured on industrial scale. Worldwide, there are nearly hundred companies dealing with production of metal, semiconductor, and dielectric crystals, mostly, combined with subsequent highly precise crystal cut and pieces polishing before sale. Yearly, the advanced semiconductor magazin *III-Vs Review* publishes a special company directory (see, e.g., [35]) with numerous informations on company profiles and product offers.

In order to meet the rapid developments of device dimension reduction and, especially, nanotechnology during the last decade the growth of

crystals with nm and μm dimensions have come to the for. Special target is their multiple arrangement in well-ordered patterns, e.g., by the principle of self-assembly [36]. Beyond the traditional transitions between two phases, the three-phase sequence Vapor-Liquid-Solid (VLS) mechanism (see 8.4) to form nanowhiskers of different materials is more and more applied. Although the VLS mechanism is already known for a long time and was experimentally introduced in 1964 by Wagner and Ellis [37], it has now a renaissance on a much higher technological level. More details are treated in, e.g., (→ Crystallization and Precipitation). History, fundamentals, technological growth aspects, and applicabilities of fiber crystals with diameters in μm dimensions are summarized in the monograph *Fiber Crystal Growth from the Melt* [38].

Finally, the epitaxy techniques for growth of monocrystalline thin films and stacks of quasi two-dimensional thin layers from alternating materials (i.e. multi-quantum wells), showing new physical properties on the nanometer scale, have been developed into the challenging field of quasi-two-dimensional crystal growth with important technological applications in micro- and optoelectronics. Since the first experiments by liquid-phase epitaxy (LPE) [39], chemical vapor deposition (CVD) [40], metal organic chemical vapor deposition (MOCVD) [41] in the 1960s, and molecular beam epitaxy (MBE) [42] in 1975, the capabilities of the methods and device quality have been markedly improved. Even combined methods like MBE with metal organic precursors (MOMBE) [43] are now under investigation. Nowadays, the growth of epitaxial diamond films from an activated vapor phase under low-pressure conditions is possible [44]. An overview on modern epitaxy techniques and applications is given in [3, vols. 3 a, 3 b]. Here only some introductory aspects are referred to in Chapter 8 (see also → Semiconductors, Chap. 5, → Silicon, Chap. 4.2.5, → Thin Films, Chap. 1.4).

1.2. Uses of Single Crystals

An ideal crystal is an infinite array (lattice) of atoms, arranged in patterns that repeat in all three spatial dimensions with uniform repeat distances; these lattice spacings are usually a small number (about 10) of interatomic distances. Lattice spacings for inorganic materials range from about 2.5×10^{-8} to 3.5×10^{-7} cm. Lattice spacings for organic materials involving macromolecules can be much larger (up to 10^{-5} cm for polymers, proteins, and viruses).

Real crystals are finite, and the effects of the surfaces can be detected in some cases. Real crystals also have some unfilled lattice sites (vacancies) and atoms in some sites that are not part of the ideal lattice (interstitials). The number of these point defects depends on the temperature and on the growth conditions or the post-growth conditions in which the crystal comes to equilibrium with its ambient. At low temperatures, the time taken to attain equilibrium can be very long (e.g., years at room temperature). In addition to these thermodynamically predictable point defects, most crystals contain impurity atoms and molecules as well as line defects (dislocations). Except for elemental silicon and germanium, multicomponent bulk crystals (compound semiconductors, oxides, fluorides, etc.) contain dislocations that are mostly inimical to various types of devices due to their surrounding stress field and electrical activity. Additionally, they act as getters for excess and impurity atoms, and this often leads to precipitations (three-dimensional defects). In most cases, characteristic dislocation arrays, such as cellular and grain structures, are formed, which constitute low-angle grain boundaries. Samples containing low-angle boundaries are strictly polycrystalline, but for many purposes, boundaries involving angular changes of about $< 1°$ are acceptable. However, give the increasing demand for nearly perfect highly integrated thin-film devices with high parameter reproducibility, the presence of a network of subgrains, tilted even by lower angles (down to 20 arcsec), is now one of the most serious defects in substrate compound crystals that can significantly impair epitaxial growth and cause scattering of electronic device parameters. Hence, key targets of the crystal grower are the avoidance of subgrain structures and the achievement of homogeneous distribution of the minimum possible number of dislocations. The other two-dimensional sheet defect is a twin plane involving a rotation or reflection of the lattice structure. Samples containing twin planes are not single crystals in the

sense used here, and such samples are unacceptable for many applications.

The existence of long-range order (or the absence of short-range disorder) gives crystals their valuable properties, such as the absence of scattering for long-wavelength photons and particles (e.g., the intrinsic or extrinsic carriers in semiconductors). Here, long means large compared to the interatomic spacing. For photons (e.g., X-rays) and particles (e.g., neutrons and electrons) with wavelengths comparable to the interatomic spacings, systematic interference effects (loosely called diffraction) occur if the wavelength λ and angle θ to the atomic planes fulfill the Bragg condition

$$n\,\lambda = 2d\sin\theta \tag{1}$$

where n is an integer. This property leads to many uses in scientific instruments.

Table 1. Some uses of crystals

Property exploited	Device	Crystals used
Spatial repetition	X-ray prisms	lithium fluoride
	neutron collimators	metals
Absence of	turbine blades	metals
short-range disorder	gramophone styli	sapphire
(giving strength)	bearings	ruby
	tape heads	ferrites
	wire drawing dies	diamond
	abrasives	diamond, sapphire
Uniformity of	lenses, prisms, and	alkali and alkaline
(eliminating scatter	optical windows	earth halides
electromagnetic	lasers	yttrium aluminum
waves)		garnet
	microwave filters	yttrium iron garnet
Uniformity	transistors, diodes,	silicon, gallium
(eliminating scatter of	thyristors, and	arsenide, gallium
charge carriers)	integrated circuits	nitride, silcon carbide
Anisotropy	resonant filters,	quartz, lithium
(allowing exploitation	delay lines	niobate, zinc oxide,
of vector and tensor	Nicol prisms	fluorite
properties)	ultrasonic transducers	Rochelle salt
	gramophone pickups	lithium sulfate
	laser frequency	potassium
	converter	dihydrogenphoshate,
		barium borate

The existence of long-range order which differs with direction leads to anisotropy of crystal properties. Anisotropic properties are useful, for example, in optical polarizers. Even cubic crystals exhibit some anisotropic properties because they are not completely symmetrical (some have no center of symmetry). The absence of a center of symmetry can lead to useful pyroelectric properties (i.e., the appearance of electric charge on some surfaces of crystals that have been subjected to a change in temperature) or piezoelectric effects (the appearance of charge on the surface of crystals subjected to mechanical strain).

Finally, the absence of short-range disorder can give crystals great mechanical strength: no internal sites exist for the nucleation or easy propagation of cracks.

Table 1 contains some examples of the range of applications and typical materials used in these applications. The list is not exhaustive, and interested readers should consult the appropriate contributions in this encyclopedia and standard texts on optics, X-ray and particle diffraction, crystallography, solid-state physics, and the strength of materials. In many applications, crystals offer more than one advantage; in particular, the absence of scattering effects is often a significant advantage of single crystals that are used for some other primary reason.

1.3. Specification of Crystals

Like any other product, crystals should be produced to a specification, but this is not always easy: factors influencing device performance and yield are not always known, and changes in fabrication techniques may call for a material meeting a radically changed specification. Thus, from the producer's point of view, a user's specification "suitable for the application" may be variable and dependent on factors of which he has no knowledge. Similarly, a specification "best available" (implying as pure and as perfect as possible) may also be unsatisfactory: the user's processes might rely on some unspecified impurity or depend on the existence of some imperfections which absorb undesired impurities or modify properties (e.g., diffusion, wettability). The factors that must generally be specified are:

Dimensions and tolerances
Orientation and its accuracy
Gross perfection (low-angle boundaries, twins, inclusions–sizes and volume density)
Crystallographic perfection (dislocation density)
Composition (major component homogeneity)
Dopants (concentration, distribution)

Impurities (maximum concentration of undesired impurities, maximum concentration of all other impurities specified as a total)

In particular cases, other factors may have to be included, and the list will obviously be lengthened if the producer does any post-growth processing such as trimming to size, slicing, or etching. The specification should also include the nature and number of tests to be done. The problems of characterizing crystals are briefly discussed in Section 5.5.

1.4. Commercial Aspects

Three types of organizations are engaged in growing crystals:

1) Organizations that specialize in crystal growth and supply crystals to other user organizations
2) Organizations that grow crystals in-house for their own use
3) Organizations that investigate crystal growth

From the worldwide data available, about 500 organizations can be identified that are interested in crystal growth; of these, perhaps 300 are commercial organizations of categories 1) and 2). Few organizations interested commercially in crystal growth employ more than 50 people who work on crystal growth. Academic teams tend to be small (rarely more than 10). Thus, about 15 000 people in industrial organizations are involved with crystal growth, as well as perhaps another 2000 in academic institutions.

The worldwide output of the industry is about 25 000 t of crystals per year. The major user of these crystals is the electronics industry, which consumes about 16 000 t of silicon (mostly grown by the Czochralski technique), 2500 t of quartz (grown hydrothermally), and 800 t of pentelides (pnictides), such as GaAs, InP,GaP, CdTe, or ZnTe grown by the Czochralski and vertical Bridgman techniques. The current market for ZrO_2 skull-melted crystals, over 98 % of which are fabricated into gemstones, amounts to about 1000 t. The engineering industry uses about 2000 t of slowly cooled, nearly single crystal alloys for highly stressed components and 100 t of nonmetallic crystals for bearings, dies, and tools. About 2000 t of low-quality metal crystals are used in transformer cores and components for electric motors. The optical industry uses about 400 t of assorted halides and some silicon or germanium for windows, prisms, and lenses; other crystals, such as ruby or yttrium aluminum garnet, for lasers; lithium niobate, and ammonium and potassium dihydrogen phosphates for various electrooptic devices. Crystals in common use in smaller amounts are triglycine sulfate (for pyroelectric devices); various chalcogenides, such as ZnS or CdTe, for optical components or $Cd_x\ Hg_{1-x}\ Te$ for infrared detectors; and lithium tantalate for optical and pyroelectric devices. Altogether, about 200 materials are available commercially as single crystals. Most are used at very low rates ($<$ 1 t/a).

When methods are ranked by annual mass produced, the Czochralski technique is first (about 40 %); other melt growth methods (predominantly the Bridgman technique coming to the for even since the last decade as electronically controlled gradient freeze version) produce another 35 % (increasing), solid-state growth accounts for perhaps 12 %, the hydrothermal technique produces about 8 %, and other solution growth techniques account for almost all the remaining output. Perhaps 1 % is grown from the vapor phase.

Melt growth usually has the highest volume growth rate. Quartz can only be produced hydrothermally. Solution growth techniques are usually inexpensive in terms of capital and labor costs. Vapor growth tends to be both slow and expensive. However, both high-temperature solutuion (flux) and vapor growth are increased applied for production of numerous modern semiconductors with extremely high melting points (over 2000 °C), such as SiC, GaN, AlN, being important materials for high-power microelectronics and blue-violett wave regions. For further discussion, see Section 5.4.

When methods are ranked by value, the order of importance is melt growth (about 50 %), solution growth including hydrothermal processes (25 %), vapor growth (15 %), and solid-state growth (10 %). However, this presentation is misleading because a significant fraction (about one-quarter) of the melt-grown crystals are used to provide substrates for the vapor and solution growth of epitaxial layers. The total value of thin layers and their substrates is about 25 % of the value of all the crystals grown.

1.5. Sources of Information

The primary literature (research papers) on crystal growth is extensive: about four papers are published every day on topics relevant to this article, the majority in three international journals (*Journal of Crystal Growth*, *Crystal Growth Design*, and *Crystal Research Technology*, formerly *Kristall und Technik*) and a limited number of conference proceedings. Chemical Abstracts offers *CA Selects Crystal Growth*; this service covers a rather wide field (e.g., industrial crystallization, geological and biological crystallization, as well as the growth of single crystals), allows most of the primary papers to be recognized, covers the patent literature, which is an important source of practical information, and serves as an aid in locating the secondary literature.

Nowadays the abstracts of most current publication and patent citations on crystal growth and epitaxy can be obtained from international databases (e.g., INSPEC with about 4000 journals, Chemical Abstract Service, CRYSTMET, GMELIN, for example) via on-line internet suppliers (e.g., FIZ Karlsruhe).

The secondary literature may be divided into two sections: small area reviews and monographs. Small area reviews are also concentrated; they appear as invited papers at conferences (often in special issues of the *Journal of Crystal Growth*) and in the review journals *Progress in Crystal Growth and Characterization* (Pergamon, Oxford) and *Materials Science & Engineering* (Elsevier). In the 1970s and 1980s numerous reviews were published in a limited number of book series, for example, the series *Crystals* (edited by M. C. Freyhardt) published by Springer (Berlin) or the North Holland (Amsterdam) series *Current Topics in Materials Science* (edited by E. Kaldis).

Fundamental and methodical aspects are reviewed in the *Proceedings of the International Summer Schools on Crystal Growth* (ISSCG) triennially organized by the International Organization of Crystal Growth (e.g., *Theoretical and Technological Aspects of Crystal Growth*, edited by R. Fornari and C. Paorici, Trans Tech Publications, 1998) and *Crystal Growth - from Fundamentals to Technology*, edited by G. Müller, J.-J. Metois and P. Rudolph, Elsevier, 2004). A special chapter on crystal growth of materials treating fundamental and experimental aspects of production of nearly all important substances is included in the *Encyclopedia of Materials: Science and Technology* Elsevier, 2001.

The literature in monograph form is also extensive. There are texts [45, 46] for amateurs or students not specializing in crystal growth. More serious students wanting a broad view should turn to [4, 47 – 51]. Specialists have a wider selection: thus, on low-temperature solution growth, [26] still has much to offer; references [52 – 54] contain much useful material. The last [54] also covers melt growth, as do [55] and [56]; while these three discuss high-temperature solution growth, the standard text is [57]. For vapor growth, consult [58]. Most texts give some account of the theory of the subject, but [59] is the standard text. A number of highly specialized texts contain pertinent information; for example, [60] covers growth in gels (solution growth in the absence of convection), and the growth of polymers (growth controlled largely by the stereochemistry of the molecules) is discussed in [61].

Historical important papers are collected in the book *50 Years Progress in Crystal Growth*, Elsevier, 2004 edited by R.S. Feigelson. In *Bulk Crystal Growth of Electronic, Optical and Optoelectronic Materials*, John Wiley and Sons, 2005, edited by P. Capper, 17 authors summarize the state-of-art of single crystal production and characterization of semiconductors, sapphire, fluorides, scintillator materials, and quartz crystals. Technological aspects of growth and wafer preparation are reviewed in 29 chapters, written by numerous well-known specialists in *Crystal Growth Technology*, Wiley VCH, 2006 edited by H. Scheel and T. Fukuda. An extensive modern compendium comprising three volumes with a total of about 2500 pages was edited during the period of 1993–1994 by D. T. J. Hurle (*Handbook of Crystal Growth*, North-Holland, Amsterdam). Seventy four authors contributed fifty-eight chapters to this almost complete compilation on fundamentals, bulk crystal growth, and epitaxy, which is undoubtedly the leading standard work on these fields at present and will be for some time to come.

Because it is now possible to describe crystal growth processes in a nearly quantitative manner, there is a need for numerical data on physical properties and phase diagrams. The texts

just quoted suggest some sources and others are given in later sections. There are two relevant on-line data banks such as the *Electronic Materials Information Service* (IEE, Station House, Nightingale Road, Hitchin, Herts SG 5 IRJ, United Kingdom) or *EVITHERM Date Centre*, which give relevant physical property data and *MTDS* (National Physical Laboratory, Teddington, Middlesex TW 11 OLW, United Kingdom), which gives thermodynamic data; [62] lists a series of relevant off-line data centers. In 2005, the special book on compound semiconductor material data *Properties of Group-IV, III-V and II-VI Semiconductors* was published by John Wiley and Sons.

2. Thermodynamic Aspects

2.1. Free Energy and Driving Force of Crystallization

The precondition for crystallization of a stable solid phase from a metastable fluid (starting) phase is deviation from thermodynamic equilibrium. Figure 1 plots schematically the free energy versus temperature $G(T)$ of two contacting phases in the neighborhood of a first-order phase transition. Only at the equilibrium point are the Gibbs potentials G_s and G_f equal, but away from equilibrium the potentials are different. The difference ΔG denotes the driving force of crystallization or growth affinity as $\Delta G = G_f - G_s$.

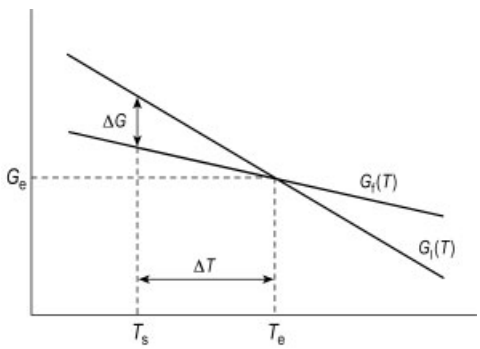

Figure 1. Dependence of free energy G of fluid (f) and solid (s) phases on the temperature T near the thermodynamical equilibrium (e);
ΔG and ΔT denote the driving forces of crystallization and supercooling, respectively

Most crystal growth processes are carried out near equilibrium. Exceptions are vapor growth processes, in particular molecular beam epitaxy, and some techniques involving chemical reactions on the surface of the growing crystal. The driving force for crystallization is usually defined in terms of the free energy released by crystallization. This free energy change ΔG can be related to more easily measured system parameters for supercooled melts, supersaturated solutions, and supersaturated vapors:

1) *Supercooled Melt.* For a melt supercooled by ΔT (at a temperature T_e–ΔT),

$$\Delta G = \Delta H\, \Delta T / T_e \tag{2}$$

where ΔH is the enthalpy released by crystallizing 1 mol of solid and T_e is the temperature at which the solid and liquid are in equilibrium, in this case the melting point. If ΔT is large or ΔH is small, ΔH in Equation (2) must be corrected to allow for the difference in specific heats of the solid and the liquid. Doing this introduces an additional term in ΔT^2 [54, Chap. 3]. Values of ΔT rarely exceed 10 K. Values of $\Delta H/T_e$ usually fall within a range of 10 to 100 J K^{-1} mol^{-1}, and values of ΔT are typically 1 K. Hence, ΔG is about 100 J/mol, and $\Delta G \ll RT$.
For instance, a propagating (111) face of a dislocation-free silicon crystal requires a supercooling of about ΔT = 4 K [64] (maximal 7 K [63]). By inserting this value in Equation (1) together with the material parameters ΔH = 50 kJ mol^{-1} and melting point T_e = 1693 K, a driving force of crystallization at least of ΔG = 118 J mol^{-1} can be deduced.

2) For a *supersaturated solution*

$$\Delta G = n\, RT \ln(1+\Delta c/c_e) \approx n\, RT\, \Delta c/c_e \tag{3}$$

where n is the number of ions formed in the solution by one molecule of the solute, R is the gas constant (8.31441 J mol^{-1} K^{-1}), T is the temperature, and Δc is the amount by which the solute concentration exceeds the equilibrium concentration c_e. The relative supersaturation

$$\sigma = \Delta c/c_e \tag{4}$$

is often used. Typical values of σ are less than 0.1. A driving force of crystallization of about

500 J/mol was found for perfect NaCl crystal growth from aqueous solution [49]. If ΔH is the enthalpy of solution, Equation (2) can also be used to describe growth from a solution cooled by ΔT below its saturation temperature. Some degrees of supercooling were ascertained for the growth of quartz single crystals under hydrothermal conditions [49].

3) For a *supersaturated vapor*

$$\Delta G = RT \ln(1+\Delta p/p_e) \approx RT\,\Delta p/p_e \qquad (5)$$

where Δp is the amount by which the actual partial pressure of the material exceeds the equilibrium value p_e. If ΔH is the enthalpy of vaporization, Equation (2) can be used to express ΔG in terms of cooling the vapor from its equilibrium temperature. Typically $\Delta p/p_e < 0.1$ and ΔG is ca. 1 kJ/mol

The values of T_e, c_e, and p_e are given by the phase diagram which describes equilibria involving pure, unstrained single crystals. To describe the growth of doped or strained crystals, terms involving elastic strain energy must be added [65]. Similarly, any departure from the phase diagram condition (e.g., the application of hydrostatic pressure, electric or magnetic fields) requires minor modifications of the expressions for ΔG. Such effects at most change growth rates by factors of three.

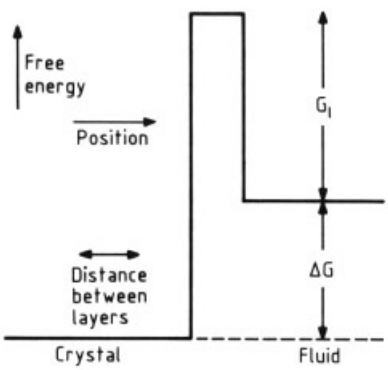

Figure 2. Free energy as a function of position for a simple model of crystal – fluid interface

In the simplest possible free energy–spatial position model of a crystal–fluid interface (Fig. 2), an energy barrier G_I separates the crystal from the fluid. Therefore, the transfer rate from the fluid to the crystal should be proportional to $\exp(-G_I/RT)$, and the rate from crystal to fluid should be proportional to $\exp[-(G_I+\Delta G)/RT]$. Thus, the net rate of transfer to the crystal should depend on $\exp(G_I/RT)\cdot[1-\exp(-\Delta G/RT)]$ which for $\Delta G \ll RT$ reduces to $\Delta G \exp(-G_I/RT)/RT$.

The determination of the concrete value of G_I is somewhat problematical. Approximately it denotes the activation energy for volume diffusion in the fluid phase, that is, viscous flow in the melt (on the order of 5–50 kJ mol^{-1} [66]). At a crystal–solution interface, solvate shells form or disintegrate, and this requires higher activation energies (desolvation or dehydration energy), yielding G_I in aqueous solutions of 50–100 kJ mol^{-1} [49].

2.2. Phase Relations for Major Components

Because crystals are usually grown in near-equilibrium situations, the actual equlibrium states as described by the phase diagrams should serve as a useful guide to the conditions needed. However, drawings lack accuracy and, for exact work, analytical or tabulated numerical data are preferred. In a well-designed crystal growth system, temperatures can be established with errors of less than 1 K; pressures, between 1 and 5 %; and compositions, 0.1 %.

For introductions to the theory of phase diagrams from the point of view of crystal growth, see [59, 67, 68]. Reference [69] gives an excellent account of the determination of phase diagrams. Techniques have since been automated, but the basic problems are unchanged and several researcher's years of work are still required to determine a three-component phase diagram. For further information on the principles involved, see the standard texts [70 – 72]. A literature search is necessary to obtain the latest data on any particular system. A bibliography up to 1956 is given in [73]. Data for many binary systems can be found in [74 – 77]. Oxide systems are covered by [78 – 80]. Reference [81] gives data for glass-containing systems. For data on the solubilities of materials, see [82]. Numerical data for calculation can be obtained from [83, 84], and [85], and from the data banks discussed in Section 1.5. A journal (*Calphad*) is devoted

to phase diagram calculation, and data on actual systems appear in many journals; in particular, diagrams relevant to crystal growth often appear in *Journal of Crystal Growth, Journal of Materials Science, Journal of Solid State Chemistry,* and *Journal of the Electrochemical Society.* The *Phase Diagram Handbook* with its updating service [86] provides contemporary data on binary diagrams.

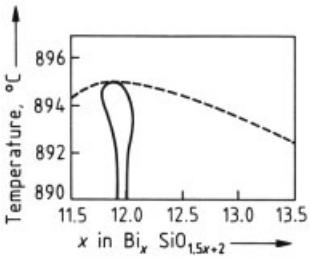

Figure 3. Phase diagram near the maximum melting point of $Bi_x\ SiO_{1.5x\ +\ 2}$ grown in air

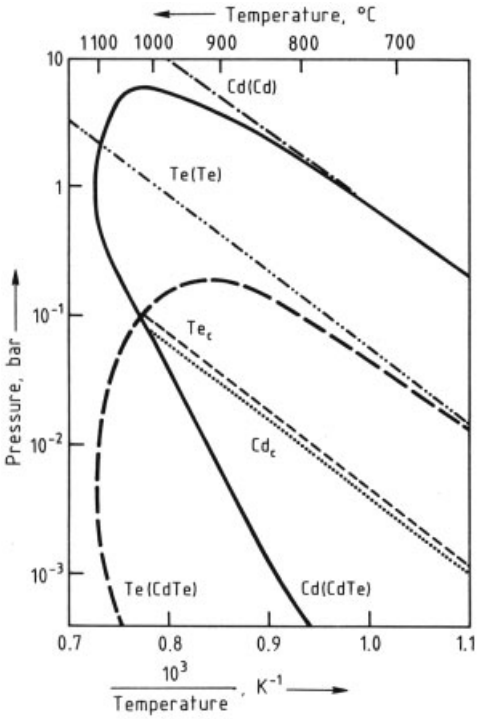

Figure 4. Pressures in the Cd – Te system near the compound CdTe
The convention X (Y), e.g., Te (CdTe), is used to indicate the partial pressure of material X over material Y. The symbols Te_c and Cd_c refer to the partial pressure of Te and Cd for the congruently vaporizing solid

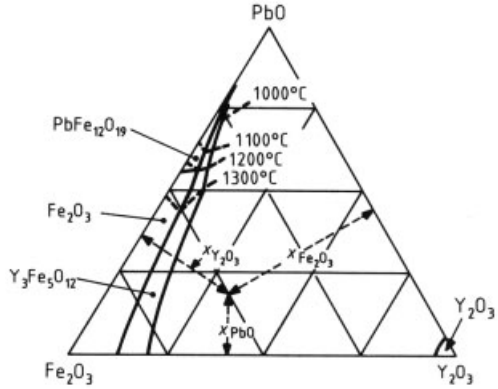

Figure 5. Liquid compositions in the Fe_2O_3–Y_2O_3–PbO system in equilibrium with the indicated solid phase

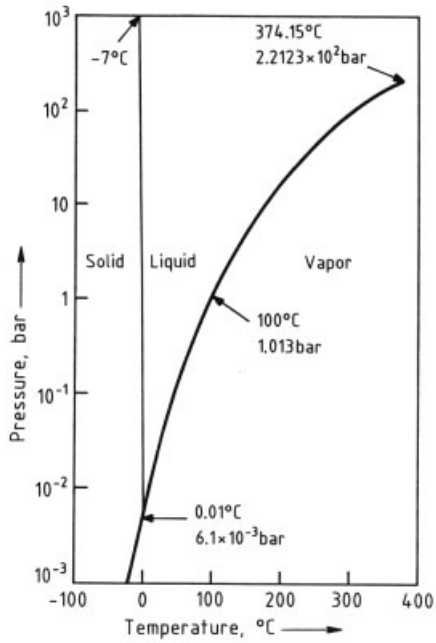

Figure 6. Pressure–temperature diagram for water

Figures 3 – 6 show some typical diagrams; data for these are taken largely from the sources listed above or from [87], in which the phase diagrams of electronic materials are discussed. When phase data are vital, a literature search followed by a critical review of the data should be undertaken. Typical differences found thereby are 1 to 10 K for temperatures, 0.1 to 1 % for compositions, and 5 to 15 % for pressures. Selecting "best available" values is not easy: when

many data are available, median values are probably the best choice, and this has been done in Figure 3 – 6.

Figure 3 shows that part of the phase diagram of the system SiO_2–Bi_2O_3 in air, which is of interest for the growth of the material loosely described as $Bi_{12}SiO_{20}$. The solid line gives the composition of the solid, and the broken line gives the composition of the liquid. At any temperature, the liquid represented by a point on the broken curve is in equilibrium with the solid represented by a point on the part of the solid line nearest to it at the same temperature. The diagram illustrates the wide range of compositions that exist for compounds usually assumed to have a constant composition.

When the growth takes place in pure oxygen, the curves are shifted to the left and the maximum melting point occurs at $x = 11.6$. In argon + 0.5 % oxygen, the curves are shifted to the right and the maximum melting point is at $x = 12.0$.

The formula can differ quite significantly from the idealized value, and at any temperature below the maximum melting point, only two solid compositions can be in equilibrium with a melt.

Figure 4 gives data about vapor pressures of interest in the growth of CdTe. When this formula is written in the form $CdTe_y$, the maximum deviation y lies at 850 °C between about $1 - (1 \times 10^{-4})$ and $1 + (3 \times 10^{-4})$, with a value of $1 + (1.5 \times 10^{-5})$ at the maximum melting point (1092 °C) [88, 89]. Large changes in partial pressures are seen to result from these small changes in composition. The dotted lines represent roughly the partial pressures in equilibrium with the solid which evaporates to give a vapor with the same composition as the solid (called the congruently subliming solid). The distance between the lines has been exaggerated by a factor of about 10^4 to emphasize that they are not equal.

At any temperature, the product $p_{\mathrm{Cd}} p_{Te}^{1/2}$ is a constant. When this condition is observed by annealing at appropriate partial pressures, any composition between the extremes at that temperature can be obtained, but again, only the extreme values can be in equilibrium with a melt.

Figure 5 shows liquid compositions in the system Fe_2O_3–Y_2O_3–PbO which are in equilibrium with the solids indicated in the regions separated by bold lines, the so-called primary crystallization field. The arrows indicate how the molar fractions Fe_2O_3, Y_2O_3, and PbO are represented. For the point shown, the values are 0.50, 0.32, and 0.18, respectively. The broken lines are isotherms. The diagram indicates that, for example, if a liquid containing 50 mol % Fe_2O_3 and 7 mol % Y_2O_3 is cooled from about 1300 °C, the garnet phase should be precipitated. The crystalization field of $Y_3Fe_5O_{12}$ (YIG) is enlarged by the addition of B_2O_3 and PbF_2 to the flux discussed above [90].

Figure 6 shows the phase diagram of water, which is important for hydrothermal growth (Section 7.6). Many other diagrams are of interest to crystal growers; for example, plots of solubility against temperature, which over small ranges can usually be represented as a straight line when the logarithm of solubility is plotted against reciprocal absolute temperature.

2.3. Phase Relations for Minor Components

Usually, the crystal being grown, even if it is a compound, is regarded as contributing one component in a phase diagram and the solute another. Thus, in growing a doped crystal from the melt, the system can be considered a binary one. Similarly in solution growth, three components — the solvent, the crystallizing material, and the solute — are being considered. Because only dilute solid solutions are under consideration, to a first approximation, effects resulting from the concentration of the solute (dopant, impurity) can be neglected, and in equilibrium, is

$$x_{\mathrm{c}} = x_{\mathrm{N}} \exp \frac{-\Delta\mu}{kT} \tag{6}$$

where k is the Boltzmann constant, x_{c} is the atom fraction of the solute in the crystal, and x_{N} is the concentration in the nutrient phase. The parameter $\Delta\mu$ is the chemical potential difference between a solute atom in the crystal and one in the nutrient phase. This difference is almost always positive: it is more difficult to fit a foreign atom into a crystal than into the nutrient (which is usually a fluid). Thus, the segregation coefficient of a solute

$$k_0 = x_{\mathrm{c}}/x_{\mathrm{N}} \tag{7}$$

is usually less than unity. In the general case, $k_0 < 0.1$, but there are many exceptions. It is often

possible to use μ per mole and to replace k in Equation (6 by R (the gas constant).

It is possible to derive Equation (6) rigorously and to show in an ideal case of melt growth that

$$\Delta\mu = \Delta H_2\,(1-T/T_2) - \Delta H_1\,(1-T/T_1) \tag{8}$$

where ΔH_2 is the latent heat of fusion of the crystal which melts at T_2 and ΔH_1 is the latent heat of fusion of the solute which melts at T_1. However, few cases are ideal and it is usually necessary to add extra terms to $\Delta\mu$ which imply the attractive or repulsive forces between the crystal and the solvent atoms, that is, the activities in the solid and fluid phases. For a discussion of these, see [49, 54, 59, 71, 91].

2.4. Interfacial Effects

The surface of a crystal is a gross discontinuity and has a free energy associated with it. The value of this free energy depends on the orientation of the face and the other phase in contact. In surface studies, this phase is usually a vacuum; in the present case, the other phase is often a liquid; for a discussion of crystal–liquid interfaces, see [87]. As a first approximation, the surface free energy is

$$\gamma_{\mathrm{SP}} = (1-w/u)\,\Delta H_{\mathrm{SP}} N_{\mathrm{A}}^{-1/3} \tag{9}$$

where u is the number of nearest neighbors for an atom inside the crystal and w is the number of nearest neighbors in the crystal for an atom on the surface ($w < u$). N_{A} is the Avogadro number and ΔH_{SP} is the energy required to create 1 mol of crystal from the adjacent phase (i.e., the enthalpy of fusion, solution, or vaporization). A plot of γ_{SP} against face orientation gives minima as shown in Figure 7. The cusp-shaped minima give rise to atomically flat (singular) faces. Faces corresponding to less sharp minima may appear to be flat but are not flat on an atomic scale. Typical minimum values of γ_{SP} at crystal–melt and crystal–vapor interfaces are in the rangs 0.1–1.0 and 1.0–10 J/m^2, respectively. In solutions γ_{SP} is somewhat modified by absorption effects of the solvent atoms or molecules at the growing interface and descreases or increases if the solvent is surface active or passive, respectively [92]. Faces that are not singular are atomically rough. The equilibrium shape of a crystal is bounded by the set of singular faces. Real crystals may show nonsingular faces, particularly those corresponding to nonabrupt minima. In the case of small desorientation (some degree) from the singular plane, stepped (vicinal) faces do appear. They consist of lined up singular terraces with equidistant steps of atomic dimensions. Substrates with such surface morphology are of advantages for epitaxial growth in order to guarantee the continuous layer-by-layer growth mode [51].

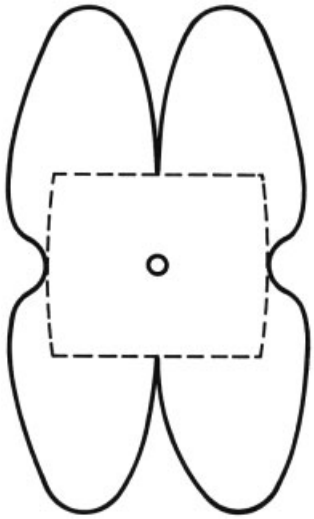

Figure 7. Polar diagram (full line) showing the variation of surface free energy with orientation
The broken line gives the shape of the cross section of the crystal

In general, the state of the surface plays an eminent role in advanced epitaxial processes, especially of superthin films and alternating multilayer structures. First the reconstruction of the substrate crystal surface, which causes a reduction in the free surface energy, may influence the nucleation and growth modes of the layer to be grown onto it [91]. Further, very thin films possess de facto the surface features as a whole. Strained structures and atomic ordering effects may appear within few monolayers of multicomponent materials on a misfitting substrate, modifying markedly their physical properties relative to these of the bulk state [93].

2.5. Shortcomings of the Thermodynamic Approach

Equilibrium thermodynamics cannot describe processes involving heat and mass transport. Modeling by use of nonequilibrium thermodynamics (e.g., employing a maximum entropy principle) is possible but difficult. The usual approach is to model the kinetics and transport processes separately and then to combine the results to give a complete description. This approach is adopted here.

3. Growth Kinetics

3.1. Atomic Structure of Growing Interfaces

The central question to be solved by using the kinetic approach is the mechanism by which the fluid atoms enter the crystal interface, i.e., growth mode and velocity on the atomic scale. This first requires the determination of the atomic structure (morphology) of the faces to be grown. At $T = 0$ all faces would to be atomically smooth because all atoms are at the same level. At $T > 0$, however, some atoms can leave the uppermost atomic plane by thermal oscillation, leaving behind vacancies and unsaturated lateral bonds and giving rise to a degree of roughness. Such faces can grow nearly without strain by the so-called normal growth mode because most approaching atoms from the fluid phase can be docked in available vacancies (Section 3.2). With increasing bond strength of a crystalline material, indicated by the enthalpy of crystallization, the less increasing the temperature is able to promote roughening. Perfect faces of such materials (ionic and some covalent crystals), show a marked resistance to growth, which can only be overcame by activating mechanisms such as nucleus generation (Section 3.3).

The criterion for transition between a smooth and a rough surface α_J [94] has the form

$$\alpha_J = \Delta H \cdot w / RTu \tag{10}$$

where ΔH is the enthalpy of crystallization, w is the number of nearest neighbors of an atom on the face, u is the number of nearest neighbors of an atom in the crystal, R is the gas constant, and T is the absolute temperature. $\Delta H/T$ is the entropy of crystallization: $\Delta H/RT$ for melt growth is often roughly a constant for a given class of materials, e.g., about 1 for most metals. If $\alpha_J < 2$ the face is rough. If $\alpha_J > 2$ the face will be smooth like those of oxide crystals with large ΔH, for example. Note that α_J varies with the orientation of the face. This is especially important for crystals for which $\Delta H/RT \approx 2$, like most semiconductors. In such cases, the w/u ratio in Equation (10) decides whether the given interface tends to be rough or smooth. For instance, silicon crystals have both atomically rough and smooth areas at melt temperature; α_J becomes 2.6 and 1.7 for the $\{111\}$ and $\{100\}$ faces, respectively. Due to the smoothness of the $\{111\}$ faces they require a higher driving force of crystallization (i.e., supercooling, see Section 2.1) than the rough $\{100\}$ faces.

On growth from vapor most crystal orientations of silicon (and of most other materials) grow atomically smooth due to the much higher sublimation enthalpy ΔH (ca. 300 kJ/mol at 1000 °C), causing an α_J of 27. Growth from solution is also characterized by the presence of mostly smooth interfaces because of the markedly lower crystallization temperatures, i.e., high values of $\Delta H/RT$ (α_J is 15 for NaCl in aqueous solution at room temperature). The kinetic growth modes of such atomically smooth faces (Sections 3.3 and 3.4) differ from the normal mode of rough faces, discussed in Section 3.2.

Note that the above analysis is somewhat simplified and valid for thermodynamic equilibrium only because the roughness of the face is restricted to the single-layer solid–fluid model. Especially at higher temperatures, however, one has to allow for a smearing-out of the crystal face at an arbitrary depth [95]. A computer simulation shows that in such cases the transition criterion α_J is 3.2 instead of 2. Further, during the growth (i.e., deviation from thermodynamic equilibrium) the surface is exposed to a driving force (Section 2.1) and a term $-\Delta G/RT$ must be added to the α_J factor that decreases its value as a function of the degree of supercooling (supersaturation). In other words, higher values of ΔG can promote nucleation islands, transforming a former equilibrium smooth face into a roughened one named "kinetic roughening" [96].

3.2. Rough Faces

With reference to Figure 2, if a Boltzmann distribution of energies for particles (atoms, molecules) in both the crystal and the nutrient phases is assumed, the particle flows to and from the crystal can be evaluated. The fraction of particles in the nutrient that can cross the barrier into the crystal is given by $\exp[-(\Delta G+G_{\mathrm{I}})/RT]$ and the fraction in the reverse direction is $\exp[-G_{\mathrm{I}}/RT]$. The actual fluxes are given by multiplying these fractions by a rate constant *kT/h*, where *k* is the Boltzmann constant and *h* the Planck constant $\left(k/h = 2.084 \times 10^{10}\mathrm{S}^{-1}\mathrm{K}^{-1}\right)$ and *kT/h* is an approximation to an atomic vibration frequency. If the spacing between layers is λ, then the growth rate on a rough surface is given by

$$f_{\mathrm{R}} = (kT\,\lambda/h) \times \exp(-G_1 R/T)\cdot[1-\exp(-\Delta G/RT)] \tag{11}$$

ΔG has been evaluated in Equations (2) to (5), and in general $\Delta G/RT$ is small. Thus, Equation (11) can be reduced to

$$f_{\mathrm{R}} \approx \frac{\lambda\, k \Delta G}{h\, R} \exp(-G_1/RT) \tag{12}$$

According to [97], the diffusion constant in the nutrient can be written as

$$D_{\mathrm{S}} = \lambda^2\,(k\,T/h)\exp(-G_{\mathrm{N}}/RT) \tag{13}$$

and the final form can be obtained as:

$$f_{\mathrm{R}} \approx \frac{D_{\mathrm{S}}\,\Delta G}{\lambda\, RT} \exp[-(G_{\mathrm{I}}-G_{\mathrm{N}})/RT] \tag{14}$$

Here, in general, D_{S}, G_{I}, or G_{N} are hardly known, but f_{R} is proportional to the supercooling $\Delta T/T$ or the supersaturation, either $\Delta c/c$ or $\Delta p/p$. Equation (14) applies only to systems near equilibrium. In systems remote from equilibrium (e.g., molecular beam growth systems), Equation (11) does not apply, and the incoming and outgoing fluxes must be evaluated in some other way (see, for example, Section 8.2). Many examples of growth following Equation (11) or (14) are known, but care must be exercised in claiming the validity of these demonstrations because an equation of similar form can be obtained by consideration of transport processes (see Section 4.3).

3.3. Perfect Singular Faces

In equilibrium, a perfect singular face is essentially atomically flat. Depositing an atom or molecule on such a face produces a large increase in free energy. However, a cluster of atoms or molecules can be stable on the face, but the probability of depositing a cluster is much lower than the probability of depositing one particle. Thus, the growth rate f_{PS} on a perfect singular face is expected to be smaller than f_{R}, the growth rate on a rough surface.

The size of a stable cluster may be modeled by a cluster in which every atom makes half the bonds it would if it were inside the crystal. For a crystal with a diamond lattice (e.g., germanium or silicon), the critical nucleus is 3 atoms on $\{111\}$, 2 atoms on $\{110\}$, and 1 atom on $\{100\}$. Thus $\{111\}$ and $\{110\}$ are singular faces, whereas $\{100\}$ and all other faces are rough. Hence, at any supercooling, growth rates on $\{100\}$ greatly exceed those on $\{110\}$ which, in turn, exceed those on $\{111\}$.

Neither consideration of bonding nor the use of α_{J} allows calculation of growth rates. To do this, the critical free energy of a nucleus ΔG^*, must be calculated. As a nucleus grows, the free energy rises to ΔG^* and then decreases so that further growth is rapid. In macroscopic terms, the free energy associated with a cylindrical nucleus is

$$\Delta G_{\mathrm{N}} = \gamma_{\mathrm{CN}}^2\,\pi r h - \Delta \mathrm{G}\,\pi r^2 h \tag{15}$$

where γ_{CN} is the surface free energy per unit area, *r* is the radius of the nucleus, and *h* is its height. The first term is the increase in free energy due to the extra surface created by the nucleus. The second term is the decrease in free energy obtained by solidifying some of the nutrient phase. By differentiating Equation (15), the following is obtained:

$$\Delta G^* = \pi h_{\gamma_{\mathrm{CN}}} \tag{16}$$

and the critical nucleus has a radius $\gamma_{\mathrm{CN}}/\Delta G$. The rate of formation of critical nuclei [49, 98], is proportional to $\exp(-\Delta G^*/RT)$ and to f_{R}.

Two limiting cases can be distinguished. The first occurs when the nuclei expand rapidly (see Eqs. 21 and 22), in which case

$$f_{\mathrm{PS}} = h\mathrm{d}\,N/\mathrm{d}t \tag{17}$$

$$= A_2 \exp\left(-B_2/\Delta G\right) \tag{18}$$

where N is the number of nuclei per unit area, and A_2 and B_2 are constants. The second occurs when the nuclei expand slowly, in which case

$$f_{\mathrm{PS}} = \left[h\pi\, f_{\mathrm{R}}^2 \frac{\mathrm{d}N}{\mathrm{d}t}\right]^{1/3} \tag{19}$$

$$= A_3 (T\,\Delta G)^{2/3} \exp\left(-B_3/T\,\Delta G\right) \tag{20}$$

where A_3 and B_3 are constants. The conditions for rapid and slow lateral expansion of the nuclei depend on the size L of the face. Rapid expansion corresponds to

$$h\, f_{\mathrm{R}} \gg L\, f_{\mathrm{PS}} \tag{21}$$

and slow expansion corresponds to

$$h\, f_{\mathrm{R}} \ll L\, f_{\mathrm{PS}} \tag{22}$$

Intermediate cases ($h f_{\mathrm{R}} \approx L f_{\mathrm{PS}}$) are difficult to express in analytical form.

Thae use of macroscopic quantities is reasonable for large nuclei. However, for small nuclei (a few molecules), macroscopic quantities have no meaning. If there are n^* molecules in the critical nucleus, another approach can be used [99 – 102], to show that with rapid spread,

$$f = A_4\, \Delta G^{1+n^*} \tag{23}$$

and with slow spread,

$$f = A_5\, \Delta G^{1+n^*/3} \tag{24}$$

where A_4 and A_5 are constants.

As can be seen from Equations (18)–(24), the nucleation-related growth rate at atomically smooth faces depends exponentially on the driving force of crystallization ΔG. Thus, so-called exponential growth modes occur.

Examples of all four relations have been discussed; e.g., see [54, Chap. 3].

3.4. Imperfect Singular Faces

Imperfect singular faces have permanent steps. However, creation of a permanent step is difficult. Consider an artificial vicinal face, i.e., one composed of shallow steps with wide singular treads and short rough risers. Figure 8 shows a sequence of interface positions for a system in which $f_{\mathrm{R}} = 2 f_{\mathrm{PS}}$. In real cases, f_{R} is expected to exceed f_{PS} by much larger factors. The argument about bonding suggests that atoms depositing in the corners (ledge sites) between the treads and the risers will more easily satisfy the condition to make half the possible bonds, but it is possible to consider lattices in which this condition is not obeyed. For example, if in a simple cubic lattice, the treads are (100) and the risers (010), an isolated atom in the ledge site (Fig. 9 A) will only make 2 of its possible 6 bonds so that, even at ledge sites, nuclei of 2 atoms will be needed. [This, of course, is less than the 4 atoms arranged in a square needed to form the stable nucleus on a perfect (100) plane.] However, if the treads are (100) and the risers are (011), an atom in the ledge site forms 3 bonds (Fig. 9 B) so that multiatom nuclei are not needed.

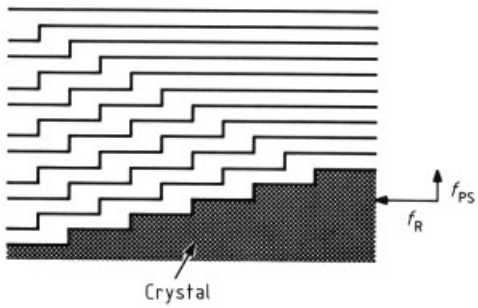

Figure 8. Sequence of interface shapes on a vicinal face The crystal is at the bottom of the figure and the sequence starts from the bottom

Two mechanisms can produce permanent steps. The first is the existence of *multiple twin planes* (Fig. 10). This is usually undesirable but gives a growth rate

$$f_{\mathrm{T}} = \alpha_{\mathrm{TW}}\, f_{\mathrm{R}} \tag{25}$$

where α_{TW} is the ratio of the twin spacing to the interatomic plane spacing normal to the twin planes.

The other mechanism is the existence of *screw dislocations*. A screw dislocation (Fig. 11) is a defect produced by making a cut part way through a block of crystal and displacing the outside edge of the cut by a distance *b,* the Burgers vector of the dislocation. Addition of another layer of atoms does not destroy the step produced. In a growth situation, a spiral forms, which is centered on the dislocation axis shown by an arrow in Figure 11; viewed from above, the dislocated face exhibits a continuous ledge

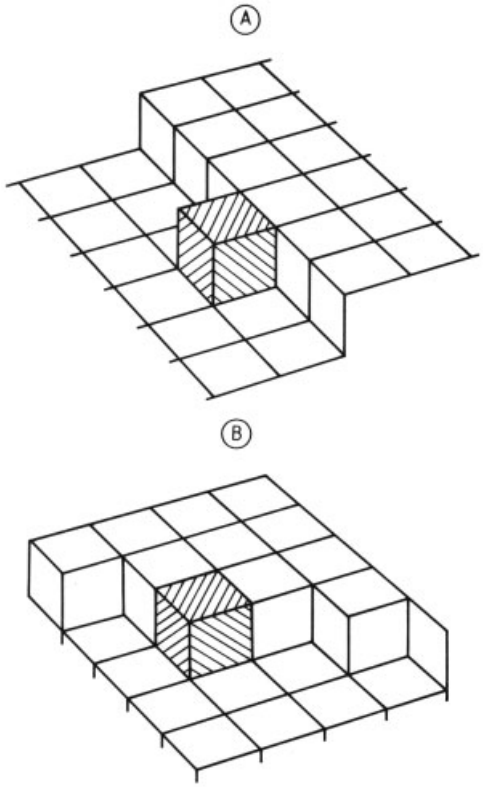

Figure 9. Straight ledge (A) and kinked ledge (B)

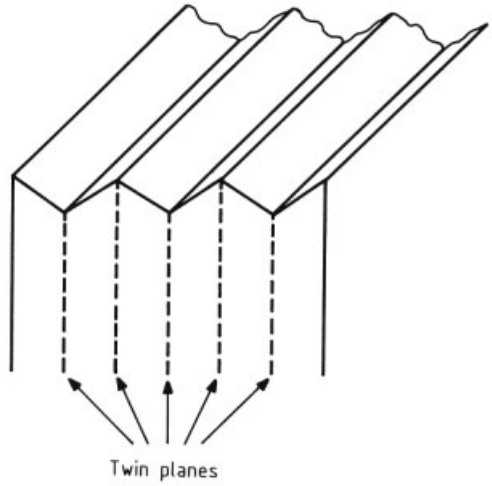

Figure 10. Twinned growth face

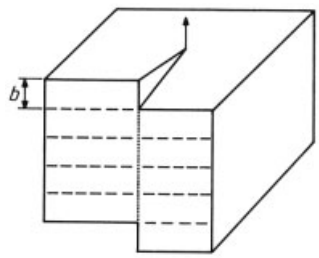

Figure 11. Face intersected by a screw dislocation

as shown in Figure 12. The form of this spiral depends on the symmetry of the crystal lattice. The easiest form to treat is a circular spiral with a radial spacing *S*. It can be shown [54] that

$$S \approx 2_{\gamma_{\mathrm{CN}}} V_{\mathrm{M}} / \Delta G \tag{26}$$

although there is dispute about the numerical factor. The fraction of sites which are ledge sites is

$$\alpha_{\mathrm{L}} = \lambda / 2\pi\, S \tag{27}$$

Hence, assuming that the growth rate is

$$f_{\mathrm{D}} = \alpha_{\mathrm{L}}\, f_{\mathrm{R}} \tag{28}$$

yields

$$f_{\mathrm{D}} = \lambda\, f_{\mathrm{R}}\, \Delta G / 4\pi\, \gamma_{\mathrm{CN}} \tag{29}$$

or

$$f_{\mathrm{D}} = A_6 \Delta G^2 \tag{30}$$

where A_6 is a constant.

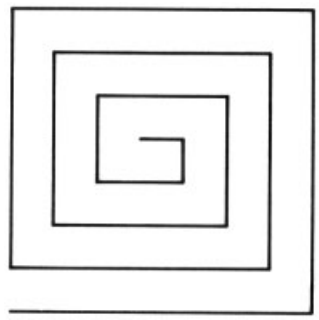

Figure 12. Projection of a face cut by a growth spiral

A more refined treatment [102] of the growth via surface diffusion yields

$$f_{\mathrm{D}} = A_7\, \Delta G^2 \tanh\left(B_7 / \Delta G\right) \tag{31}$$

which reduces to the normal growth mode

$$f_{\mathrm{D}} \sim \Delta G \text{ when } \Delta G \to \infty \tag{32}$$

and to the square growth mode

$$f_{\mathrm{D}} \sim \Delta G^2 \text{ when } \Delta G \to 0 \tag{33}$$

When diffusion in the bulk of the nutrient is a significant limitation, CHERNOV shows that

$$f_{\mathrm{D}} = \frac{A_7\, \Delta G^2 \tanh\left(B_7/\Delta G\right)}{c_1 + c_2 \tanh\left(B_7/\Delta G\right)} \tag{34}$$

where c_1 and c_2 are constants for a given system [103]. Further refinements [104] give

$$f_{\mathrm{D}} \sim \Delta G^n \text{ when } \Delta G \to \infty \tag{35}$$

where n = 3/2 or 5/3. However, in the usually important range (ΔG small), the square law relation still holds.

4. Transport Effects

4.1. Transport Processes

Before the atoms (molecules) pass over from a position in the fluid medium, just beyond the crystal–fluid interface, to their place in the crystal face they must be transported in the fluid over macroscopic distances towards the interface. Such mass transport may proceed by diffusion, forced diffusion (electromigration), buoyancy-free convection, surface-tension-driven free convection, and forced convection (rotation, acceleration, vibration). Further, the heat carried by the species on their way to the crystal by conductive and convective transport must be dissipated in the solid phase by thermal conduction and radiation to maintain a stable propagating interface.

The transport can be fast or slow compared to the kinetics of attachment. Then the rate at which a crystal grows is limited by interfacial kinetics (Chap. 3) or by macroscopic transport, respectively.

In many cases, the growth rate of a crystal can be written in the form

$$f = A_8\, \Delta G^M \tag{36}$$

where A_8 is a temperature-dependent parameter and M can take many values: for example, M = 2/3, 1, 4/3, 3/2, 5/3, 2, 7/3, 8/3, 3 would be expected and are found [48, 54]. Larger values of M might be expected, but more rigorous theory suggests that much larger values of M are improbable and a value of 4 seems to be the largest reported.

In Equation (36) the relevant value of ΔG is that at the growth face. Generally, the value of ΔG can be measured or estimated more easily at some point remote from the growth face. Thus, relating the interface value ΔG_{I} to some other value ΔG_{B} in the nutrient phase, presents a problem. The easiest way to do this is by using boundary layer theory as discussed in Section 4.2. To illustrate the approach, it is better to consider a specific variable (i.e., temperature or concentration) rather than the free energy difference which the variable creates.

Figure 13 shows the type of concentration distribution to be expected in a solution growth system. Equation (36) can be rewritten in the form

$$f = A_9 (c_{\mathrm{I}} - c_{\mathrm{E}})^M \tag{37}$$

where A_9 is a constant and c_{I} and c_{E} are the solute concentrations at the growth face and in equilibrium with the crystal. Also

$$c_{\mathrm{I}} = c_{\mathrm{B}} - \int_0^{\infty} \frac{\mathrm{d}c}{\mathrm{d}x} \mathrm{d}x \tag{38}$$

which, from Figure 13, can be expressed in the form

$$c_{\mathrm{I}} = c_{\mathrm{B}} - \delta_{\mathrm{S}} \frac{\mathrm{d}c}{\mathrm{d}x}\Big|_{x=0} \tag{39}$$

This relation defines δ_{S}. For either a solid or a fluid nutrient phase, the flux of solute at x = 0 is

$$J = D_{\mathrm{S}} \frac{\mathrm{d}c}{\mathrm{d}x}\Big|_{x=0} \tag{40}$$

where D_{S} is the diffusion constant of the solute.

If the nutrient is a solid or an unstirred fluid, it can be shown that δ_{S} is an increasing function of time and so is not useful, but for a stirred, fluid nutrient, δ_{S} is often a constant. The rate of growth of the crystal must be

$$f = V_{\mathrm{m}}\, J \tag{41}$$

where V_{m} is the molar volume of the crystal. The rates of growth given by Equations (37) and (41) must be the same; hence,

$$\frac{f\, \delta_{\mathrm{S}}}{V_{\mathrm{m}} D} + \left[\frac{f}{A_9}\right]^{1/M} = c_{\mathrm{B}} - c_{\mathrm{E}} \tag{42}$$

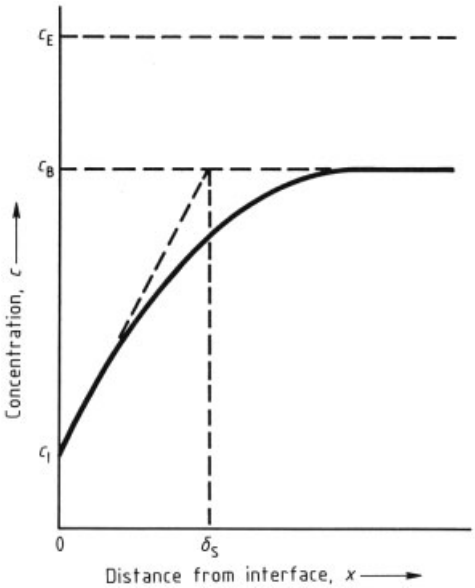

Figure 13. Concentration distribution near a growth face The crystal is in the region $x < 0$

This general relation for growth from solution in a fluid depends only on the assumption that at the growth face, the flow of solute in the fluid is purely diffusive, which is reasonable because the fluid at this location does not move relative to the growth face.

4.2. Boundary Layers

In a stirred fluid, the distributions of velocity v, concentration c, and temperature T near a growth interface often take the form shown in Figure 14, and boundary layers with thicknesses of δ_{v}, δ_{S}, and δ_{T} respectively, can be defined. In general, these layer thicknesses are different. Furthermore, not all authors use the thicknesses defined by Figure 14.

The velocity boundary layer δ_{v} is the most fundamental, and the other two thicknesses are related to this by

$$\delta_{\mathrm{S}} = \delta_v (Sc)^{-n_1} \tag{43}$$

and

$$\delta_T = \delta_v (Pr)^{-n_2} \tag{44}$$

where n_1 and n_2 vary from 1/3 to 1 (see Fig. 15) and the dimensionless numbers Sc and Pr are defined in Table 2, which also defines some other parameters.

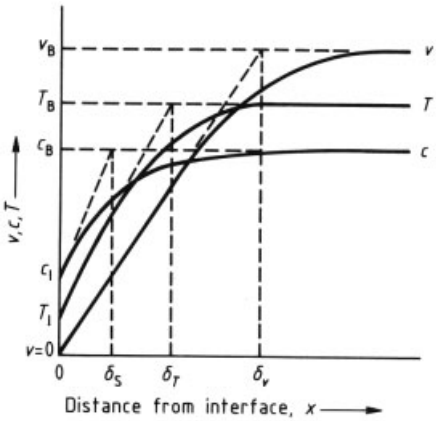

Figure 14. Distribution of temperature (T), concentration (c), and velocity (v) near a growth face, and definitions of δ_{S}, δ_T, δ_v and interface values c_{I}, T_{I}, and v_{I}

Table 2. Some dimensionless numbers

Parameter	Symbol	Definition
Prandtl number	Pr	$C\,\eta/K$
Schmidt number	Sc	$\eta/\rho\, D_{\mathrm{s}}$
Reynolds number	Re	$L\,\rho\, v/\eta$
Grashof number (for heat transfer)	Gr	$g\,L^3\,\beta\,\Delta T\,\rho^2/\eta^2$
Grashof number (for mass transfer)	Gr'	$g\,L^3\,\alpha\,\Delta c\,\rho^2/\eta^2$
Rayleigh number (for heat transfer)	Ra	$C\,g\,\beta\,\Delta T L^3/K\,\eta$
Rayleigh number (for mass transfer)	Ra'	$g\,\alpha\,\Delta c\,L^3/D\,\eta$

[a] C = specific heat; η = dynamic viscosity = $\nu\rho$; ν = kinematic viscosity; K = thermal conductivity; D_{s} = diffusion constant; ρ = density; g = acceleration due to gravity; L = typical system dimension; $\beta = (1/\rho)\,\mathrm{d}p/\mathrm{d}T$ = thermal expansivity; $\alpha = (1/\rho)\,\mathrm{d}\rho/\mathrm{d}c$ = solute expansivity; T = temperature; c = concentration; v = velocity. ΔT and Δc are changes of T and c along L.

Figure 14 has some practical consequences for the crystal grower. In liquid metals, with typically low Prandtl numbers ($N = Pr \ll 1$), the thermal boundary layer δ_{T} extends farther into the fluid than the velocity transition region δ_{v} because heat transport by conduction dominates over that by convection. However, in molten oxides and aqueous solutions with typically low thermal conductivity ($Pr > 1$), δ_{v} is somewhat broader than δ_{T}. Whereas in the last case the

temperature distribution in the melt (i.e., temperature gradient near the interface) is markedly influenced by the mass flow mode, in metallic fluids the temperature field introduced by the heating furnace remains largely unaffected by convection. Compared to multicomponent melts (fluxes) with Schmidt numbers $N = Sc$ of 10 and higher, this value for gases ranges between 0.1 and 1. Thus, in gases δ_c and δ_v extend about the same distance. In liquids, however, δ_v exceeds δ_c considerably, and therefore in front of the interface, the region of major concentration change, the flow velocity is reduced to relatively small values that barely influence the concentration profile.

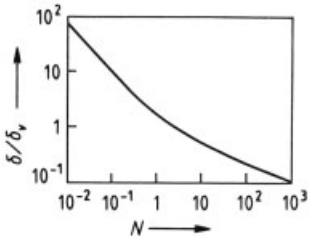

Figure 15. Variation of δ/δ_v with N, where $\delta = \delta_T$, when $N = Pr$, and $\delta = \delta_s$ when $N = Sc$
The data are for a rotating disk but should be approximately valid in most situations. For $N \leq 0.05$, $\delta/\delta_v = 0.701\, N^{-1}$. For $N \geq 200$, $\delta/\delta_v = 1.00\, N^{-1/3}$. For $N = 0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50$, and 100, values of δ/δ_v are 8.06, 4.77, 2.42, 1.56, 1.08, 0.725, 0.546, 0.415, 0.291, and 0.229, respectively. Adapted from [108].

Boundary layer theory cannot be applied indiscriminately. A boundary layer model can only be valid if

1) the Reynolds number

$$Re = Lv/\nu > 1 \tag{45}$$

where L is a dimension of the growth face and v is the velocity of the face relative to the bulk of the fluid which has a kinematic viscosity ν;
2) the dimension of the system normal to the growth face greatly exceeds δ_v; here "greatly" implies a factor of about 5; and
3) the Reynolds number in Equation (45) is such that the flow does not become turbulent. The upper limiting Reynolds number for a rotating disk is about 2000 and for flow along a flat plate about 5000.

A variety of methods are used for stirring in crystal growth systems (Fig. 16) and under further developments (see, e.g., [105 – 107]). These and many other techniques can be described in terms of the following models.

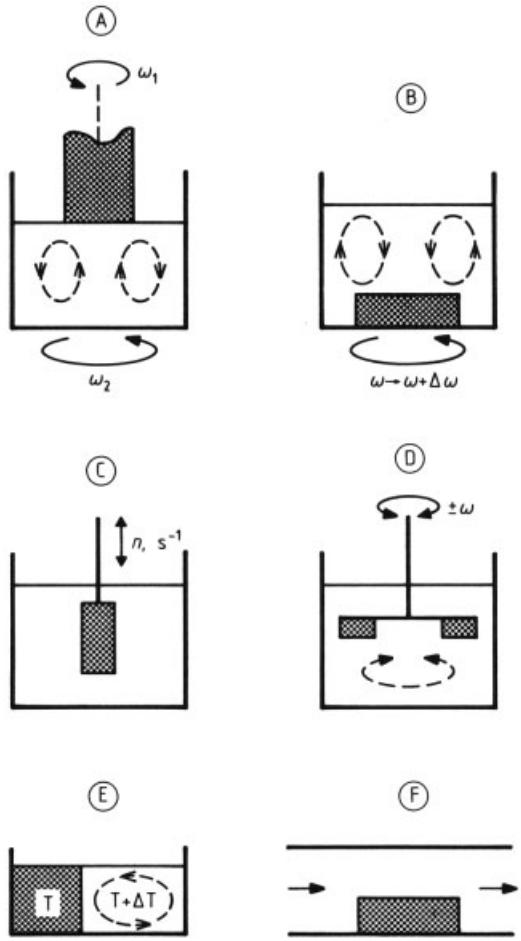

Figure 16. Methods of stirring used in crystal growth
A) Rotation of a pulled crystal; B) Accelerated crucible rotation technique (ACRT); C) Vibration of a plate; D) Reversed rotation; E) Natural convection; F) Forced convection (ω = rotational frequency; n = linear frequency)

Rotating Disk or Cone. For an infinite disk rotating at a rate ω relative to a semi-infinite fluid with the disk on the fluid surface

$$\delta_v = 1.613\,(\nu/\omega)^{1/2} \tag{46}$$

where $\nu = \eta/\rho$ is the kinematic viscosity. The model fits reasonably well, but an error will gen-

erally be introduced by using Equation (46) for a finite disk. Figure 17 gives experimental data for pulled germanium crystals rotated at 90 rpm, which suggest 10 % radial variations in δ_T for a finite crystal with an essentially flat face. For a cone with an included angle 2 θ, δ_v is given by Equation (46) if ω is replaced by $\omega \sin \theta$.

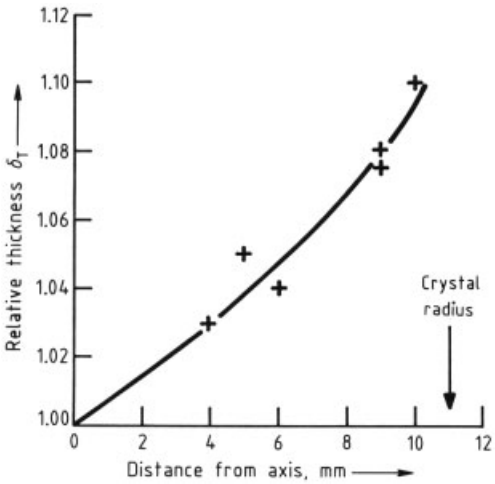

Figure 17. Thickness of a thermal boundary layer relative to the value on the rotation axis as a function of radial position

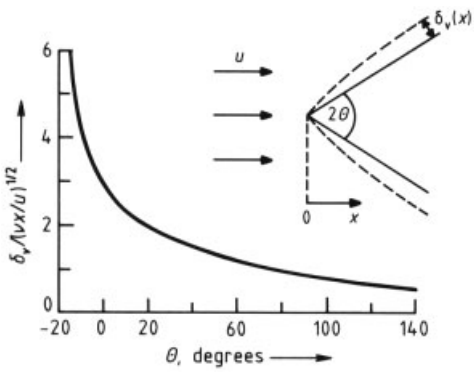

Figure 18. Value of δ_v on a wedge
Note that the origin of x is the point on the wedge furthest to the left. For $\theta \leq 90°$, the origin is the apex. For $\theta \leq 0$, the boundary layer is on the trailing face. For $\theta \geq 90°$, the boundary layer is that within a concave wedge; as $\theta \to 180°$, δ_v tends to the depth of the re-entrant face. Note that x is measured parallel to the flow

Wedge or Plate. Figure 18 gives δ_v for a wedge with included angle 2 θ in a flow with a relative velocity v. The value for a plate (v is the denominator)

$$\delta_v = 2.93\,(\nu x/v)^{1/2} \tag{47}$$

is obtained by making $\theta = 0$. The layer thickness varies with position (x) except for the case $\theta = 90°$; $\theta > 90°$ corresponds to a concave apex, and $\theta < 0$ corresponds to a wedge facing downstream. For $\theta < -17.9°$, the flow breaks away from the surface (i.e., $\delta_v \to \infty$).

Plate Vibrating in Its Plane. A plate vibrating in its plane at n Hz gives a boundary layer with

$$\delta_v = (2\nu/n)^{1/2} \tag{48}$$

but produces appreciable stirring only in a layer with thickness about 2.2 $(\nu/n)^{1/2}$.

Convectively Driven Flow. Convection may occur when there is an adverse density gradient in the fluid phase. Here, adverse is taken to mean that the density increases upwards vertically or in any direction horizontally. This gradient might be caused by a temperature gradient giving the commonly occurring thermal convection. However, growth can also cause density gradients which may give rise to convection. Any horizontal gradient appears to cause convection, but a vertical gradient seems to cause convection only when the Rayleigh number (either Ra or Ra', Table 2) exceeds some critical value which is typically in the range 1000–2000 for systems with low aspect ration h/D with h the height and D the diameter of the melt column. In the case of increasing melt height, like in full-charged Czochralski crucibles or vertical Bridgman containers, the critical Rayleigh value may increase drastically of some orders of magnitude [59]. In fluids with horizontal temperature or solution gradients (e.g., horizontal-boat crystallization techniques), where the Grashof numbers are more representative, even a threshold-free natural convection can take place [59]. If the velocity v of the convective flow is known, δ_v can be deduced from the data given for wedges or plates. Computer programs and experimental models are available for calculation and observation of

v in most crystal growth arrangements (e.g., [109 – 113]). In principle v varies as L^n, where L is the system dimension in the direction of the gradient and $n = 0.5$.

Accelerated Crucible Rotation. Consider the system shown in Figure 16 B rotating at ω radians per second. If a change $\Delta\omega$ is imposed, radial and vertical flows are initiated and these give a boundary layer with a thickness

$$\delta_v = (\nu/\Delta\omega)^{1/2} \tag{49}$$

which is established in a time of order $\Delta\omega^{-1}$ and which decays with a time constant of about

$$\tau = \left[\nu\left(0.618/H^2 + 14.7/A^2\right)\right]^{-1} \tag{50}$$

where H is the depth of the fluid and A is its radius. This method of stirring (the accelerated crucible rotation technique, ACRT [114]) is useful because stirring is induced without a mechanical connection to the inside of the crucible, but changes in the rotation rate at intervals of less than τ are required. In a time t, the velocity distribution change only extends to a distance of approximately

$$h = t\,(\Delta\omega\,\nu)^{1/2} \tag{51}$$

from the growth face.

Nonsteady Magnetic Field. There are various possibilities of arrangement of time-dependent magnetic fields in melt growth systems–rotating, travelling (down- or upward), and alternating (pulsed). Rotating field, causing a torque in the liquid $\tau_{\mathrm{M}}(B^2, \omega_{\mathrm{M}})$, with magnetic induction B and rotating field frequency ω_{M}, can be used for effective promotion of convective flows in electrically conducting melts [115, 116] Traveling and alternating fields can be applied for damping of the convection streams and control of the morphology of the propagating phase boundary [117 – 121]. Nonsteady fields are generated by using certain electromagnetic coil configurations surrounding the melt zone. Then the velocity boundary layer thickness δ_v decreases according to

$$\delta_v \sim \mathrm{e}^{-\tau_{\mathrm{M}}}\left(B^2, \omega_{\mathrm{M}}\right) \tag{51a}$$

This concept, well known from metal casting, has been revived for the melt growth of semiconductors, especially by zone melting, to homogenize the composition and to damp temperature fluctuations by means of well-controlled forced convection. Only very low values of B (a few mT) are necessary for an effective action [122]. For more information, see also [123 – 125].

4.3. Mass Flow of Major Components

In an idealized melt growth system, the composition of the solid with respect to major components is the same as the composition of the melt, and consideration of the mass flows of major components is rarely necessary: the systems can usually be described completely by considering only heat flow (see Section 4.5). In the case of growth of compounds, minor differences usually exist between the compositions of the solid and the melt (i.e., growth near, but not exactly at the maximum melting composition); however, the defect or excess quantities of the components are small and can be treated as impurities by the methods discussed in Section 4.4.

In growth by evaporative methods (e.g., molecular beam epitaxy), the fluxes of the major components can be calculated directly from the parameters of the source and its distance from the growth face (see Sections 8.1 and 8.2).

Strickly speaking, due to the mass balance at the growth interface, i.e., equality of the fluid influxes $j_{\mathrm{f}} = \rho_{\mathrm{f}}\ v_{\mathrm{f}}$ and solid outfluxes $j_{\mathrm{s}} = \rho_{\mathrm{s}} v_{\mathrm{s}}$ (ρ = mass density), the mass transport rate in the fluid perpenticular to the interface v_{f} is related to the incorporation rate in the solid v_{s} by the density ratio $\rho_{\mathrm{s}}/\rho_{\mathrm{f}}$ as

$$v_{\mathrm{f}} = (\rho_{\mathrm{s}}/\rho_{\mathrm{f}})\, v_{\mathrm{s}} \tag{51b}$$

and may, therefore, differ if $\rho_{\mathrm{s}} \neq \rho_{\mathrm{f}}$ leading to a decelerating or accelerating mass flow of the major component in the fluid if ρ_{s} is less or greater than ρ_{f}, respectively. At the melt–solid transition this difference is small ($\rho_{\mathrm{s}} \approx \rho_{\mathrm{f}}$) and the above approximation is valid. In vapor growth systems, however, where $\rho_{\mathrm{s}}/\rho_{\mathrm{f}} \approx 10^3$, a pronounced fluid mass flow of the major component, the so-called advection or Stefan flow, can be induced [4, 59].

In growth from a solid phase (Chap. 9), the two solid phases are usually assumed to have the same composition with respect to major components (and, indeed, often minor components).

4.4. Mass Flow of Minor Components

Figure 19 shows the distribution of a solute near the growth interface for the usual case ($k < 1$) and for $k > 1$, which is unusual but does occur. Two segregation coefficients can be defined:

1) the interface segregation coefficient (for definition of symbols, see Fig. 19)

$$k_{\mathrm{I}} = c_{\mathrm{S}}/c_{\mathrm{I}} \tag{52}$$

which can sometimes be made equal to k_0 (the value deduced from the phase diagram) and

2) the effective segregation coefficient

$$k_{\mathrm{eff}} = c_{\mathrm{S}}/c_{\mathrm{BN}} \tag{53}$$

As is shown in Section 2.1, a propagating interface requires a certain deviation from thermodynamic equilibrium. The incorporation of minor components is also controlled by atomic processes. To account for this, a kinetic segregation coefficient k_{i} is often introduced. Its derivation, however, is somewhat problematical because of the actual concentration c_{I} immediately at the interface is unknown (see Fig. 13). Different theoretical approaches to the formulation of k_{i}, especially in the case of atomically smooth faces, are presented in [4, 49], for example. In general, for growth from a melt involving atomically rough interfaces and near-equilibrium conditions, k_{i} equals k_{I}, and the above equations are justified.

For a macroscopically flat growth interface, a one-dimensional equation

$$D_{\mathrm{S}}\frac{\partial^2 c}{\partial x^2} + f\,\frac{\partial c}{\partial x} = \frac{\partial c}{\partial t} \tag{54}$$

can be solved with the boundary condition

$$D_{\mathrm{S}}\frac{\partial c}{\partial x}\Big|_{x=0} = c_{\mathrm{I}}\,(1-k_{\mathrm{I}})\,f = D_{\mathrm{S}}\,(c_{\mathrm{I}}-c_{\mathrm{BN}})\,/\Delta_{\mathrm{S}} \tag{55}$$

In a steady state, $\partial c/\partial t = 0$ and the solution is [33, 34]

$$k_{\mathrm{eff}} = \frac{k_{\mathrm{I}}}{k_{\mathrm{I}} + (1-k_{\mathrm{I}})\exp\left(-f\,\delta_{\mathrm{S}}/D_{\mathrm{S}}\right)} \tag{56}$$

which, for small values of $f\,\delta_{\mathrm{S}}/D_{\mathrm{S}}$, reduces to

$$k_{\mathrm{eff}} = \frac{k_{\mathrm{I}}}{1-f\,\delta\,(1-k_{\mathrm{I}})\,/D_{\mathrm{S}}} \tag{57}$$

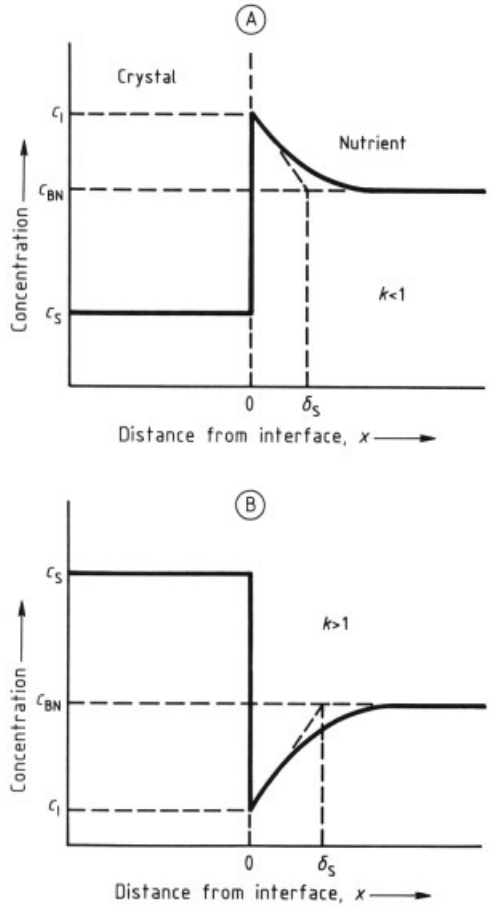

Figure 19. Concentrations of a minor component near a growth face for $k < 1$ (A) and $k > 1$ (B)

These relations have been repeatedly verified and shown to be excellent approximations, provided that δ_{S} is calculated by using Figure 15 rather than one of the limiting formulas and that there is no growth-induced flow. Thus, for example, Equation (54) applies for melt growth if the densities of the solid and liquid are identical. If this condition is not obeyed, then f must be replaced by a parameter $f^* = (1+\Delta)\,f$, where Δ is a constant for the particular system and f^* is the growth rate measured relative to fluid particles near the growth face (i.e., roughly at $x = \delta_{\mathrm{S}}$).

It is permissible to assume that $\partial c/\partial t = 0$ if changes occur at rates that are slow compared to the time (roughly $\Delta_{\mathrm{S}}^2/D_{\mathrm{S}}$) needed to modify the solute distribution in the boundary layer.

This time is usually less than one second so that Equation (54) is generally valid and may be used to describe the effects of changing either the growth rate or the boundary layer thickness δ_S. Such changes can obviously occur as a result of apparatus imperfections which can be mechanical (changes in rotation rates, etc.) or thermal (lack of perfection in temperature control). However, the most intractable problems are caused by thermal oscillations. These are time-dependent variations in the convection currents in the system, which occur in particular ranges of the Rayleigh numbers *Ra* and *Ra′* [59, 111]. Because the Rayleigh numbers depend on L^3 (Table 2), large fluid volumes are more prone to thermal oscillations than small ones and the occurrence of thermal oscillations provides a major limitation on the size of crystal growth systems. The obvious methods of minimizing the problem are (1) to work with minimal density gradients (e.g., keep temperature gradients small), (2) to design the apparatus to have effectively small dimensions (e.g., add baffles, which can be porous), and (3) to maximize viscosity (for conductive fluids, this can be achieved by applying a magnetic field [95]). The critical values of *Ra* and *Ra′* are usually several times the value needed to initiate convection (say > 2500). For *Ra* or *Ra′* greater than this value, both the oscillation amplitude and its frequency increase with *Ra* or *Ra′*.

4.5. Heat Flow

Crystal growth is always accompanied by a latent heat (the enthalpy of crystallization ΔH_C). In most cases, ΔH_C is positive (i.e., heat is evolved), but in a few instances, heat is absorbed, for example, the growth of lithium sulfate from solution in water.

If the subscripts S and N are used to denote the crystal and the nutrient phases, then at any plane growth face in a steady state:

$$K_N \frac{dT_N}{dx}|_{x=0} + \frac{\Delta H_C}{V_m} f = K_S \frac{dT_S}{dx}|_{x=0} \tag{58}$$

where K is the thermal conductivity at the temperature T, V_m is the molar volume, f is the growth rate, and x is a coordinate axis normal to the face with $x = 0$ at the face. The term dT_s/dx denotes the axial temperature gradient in the crystal near to the interface. Numerous efforts have been directed at its exact determination in growth experiments in the past. For instance, it was shown analytically, that in cylindrical crystals with high emissivity ε (metals, semiconductors) pulled from melts (Section 6.1) at high temperatures (> 1000 °C), the dissipation of the heat is dominated by radiation from the crystal surface, and dT_s/dx equals in good approximation $\left(\varepsilon\sigma_P T_e^5 a^{-1}\right)^{1/2}$, where σ_P is Planck's constant, T_e the melt temperature, and a the crystal radius. Computer calculations, including enthalpy of crystallization, thermal conductivity, convection in the fluid phase, and radiation processes, can provide almost realistic knowledge on the course of steady and nonsteady temperature fields in growing crystals ([110, 111], for example). Nowadays, even convective processes such as turbulences in the gas phase within the growth chamber are being included in the models [126].

If the nutrient is a stirred fluid, the first term is $K_N\left(T_{BN} - T_I\right)/\Delta T$ (see Fig. 14). In melt growth, if T_{BN} varies by ΔT at a rate which is slow compared with the response times of the boundary layer $\left(\Delta T^2/\varkappa_N\right)$ and the crystal $(a^2/\varkappa_S)$, then the growth rate changes by

$$\Delta f = (K_N V_m/\Delta H_C)\,(\Delta T_N/\Delta T_S) \tag{59}$$

where $\varkappa$ is the thermal diffusivity ($K/C\rho$) and a is the radius of the crystal. The change ΔT can be caused by a thermal function of the control system or by thermal oscillations. To evaluate the effect of ΔT on the concentration of a solute, $\partial k_{eff}/\partial f$ is calculated from Equation (54) and multiplied by Δf. In solution growth or other methods, the same ideas are used, but the dependence of growth rate on the interface temperature is a complicating factor. (In melt growth, the interface temperature can be regarded as a constant. This is not the case in any other method of growth.)

Dislocations. Uniform heat flow in a growing crystal has no significant effect. A nonuniform flow promotes the dislocation multiplication due to thermomechanical stress such as edge dislocations (Fig. 20). Consider a rod-shaped crystal growing in the z direction with a plane growth face at $z = 0$ and a radius a. The temperature distribution is given by

$$\frac{\partial T}{\partial t}+f\,\frac{\partial T}{\partial z}-\kappa\nabla^2 T=0 \tag{60}$$

where f is the growth rate and κ is the thermal diffusivity. It is assumed that the growth face is the isotherm $T = T_{\mathrm{I}}$ and that heat is lost from the curved surface at a rate $h\,(T–T_0)$, where h is the heat transfer coefficient. Then, provided the crystal is fairly long (say 6 a, although the solution for shorter crystals is known):

$$T-T_0\approx\frac{(T_{\mathrm{I}}-T_0)\left(1-h\,r^2/2\,a\right)}{(1-h\,a/2)}\cdot\exp\left[-\left(\frac{2\,h}{a}\right)^{1/2}z\right] \tag{61}$$

and

$$\frac{\partial T}{\partial r}\approx-\frac{(T_{\mathrm{I}}-T_0)}{(1-h\,a/2)}\,\frac{r}{a}\exp\left[-\left(\frac{2h}{a}\right)^{1/2}z\right] \tag{62}$$

When circles of constant r are considered, it can be seen that this radial gradient will give rise to hoop stresses and if the material is plastic, it will at some magnitude of hoop stress, deform to give dislocations. (If the material is brittle, it will crack at a critical hoop stress.) If τ_{C} is the critical shear stress for plastic deformation and G is the shear modulus, then the dislocation density is

$$N_{\mathrm{d}}=\frac{\alpha}{b}\,\frac{\partial T}{\partial r}-\frac{\pi_{\mathrm{C}}}{G\,ba} \tag{63}$$

where α is the coefficient of thermal expansion. For

$$\frac{\partial T}{\partial r}<\frac{\tau_{\mathrm{C}}}{G\,\alpha\,a} \tag{64}$$

dislocation-free growth is possible. Typically $\tau_{\mathrm{C}}/G\approx10^{-3}$ and $\alpha\approx10^{-5}\ \mathrm{K}^{-1}$ so that if a is measured in centimeters, the critical value of $\partial T/\partial r\approx100/a$ K/cm.

Also the axial temperature distribution in a crystal may create similar magnitudes of stress to the curvature of the isotherms in the radial direction. It was demonstrated by the thermoelastic theory of dislocation generation [127] that nonlinearities of temperature gradients (second derivations) are the sources of stress. Due to the extremely extensive numerical calculations the following relation is helpful in estimating the stress σ in the region behind the growth interface

$$\sigma\approx\alpha\,E\,L^2\partial^2T/\partial z^2 \tag{64a}$$

where E is the elastic modulus and L the characteristic axial length, which is on the order of 20 % of the crystal diameter. Conclusions can then be made about the dislocation density by comparing σ and τ_{c}; no dislocations are generated if $\sigma<\tau_{\mathrm{c}}$. In semiconductors τ_{c} is in the region of 0.7 (GaAs) to 9 MPa (Si), hence, under similar thermal conditions, dislocation-flawed GaAs and dislocation-free Si crystals are grown. Two-dimensional maps of computer-simulated temperature and stress fields in the growing crystal are now routinely plotted [110].

The local dislocation density can be estimated theoretically from the constitutive law linking the plastic shear rate and dislocation density with the applied stress in the course of the cooling-down procedure of the crystal

$$d\varepsilon/dt=\tau bV_o(\tau_{app}-D\sqrt{\tau})^m\cdot\exp\left(-\frac{Q}{kT}\right)$$

where ρ is the density of moving dislocations, b the Burgers vector, V_o a pre-exponential factor, m is a material constant, Q the Peierls potential, D a parameter relating ρ and τ, and τ_{app} is the applied stress. Details of this approach can be found in the profound review [128].

Today, the global numeric modeling of the non-stationary elastic stress and related dislocation development in growing crystals is used by implying the history of the thermal field in the furnace and crystal as well as the convection in the melt and vapour phase [129].

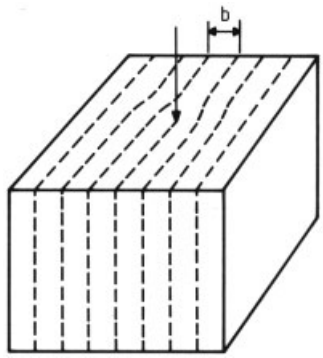

Figure 20. An edge dislocation (arrow)
The broken lines are the intersections of atomic planes with the faces of the crystal. The dislocation axis shown by the arrow is at right angles to the Burger's vector b, which, in this example, is one interplanar spacing

Thermal stresses are not the only source of dislocations; in particular, dislocations in the seed crystal usually propagate into the growing crystal and can nucleate at inclusions in the crystal. Further details and an approach to the dynamic dislocation theory, which considers interactions between dislocations already existing in the crystal during the cooling down process, are given in [128].

4.6. Interface Stability

In Chapter 3, the growth rate f was shown to be an increasing function of ΔG. Thus, if ΔG increases away from a growth face (as, for example, is always the case in solution growth), a projection on the growth face caused by a fluctuation in the growth process may grow more rapidly than the remainder of the face, and the face is unstable. When this happens, the liquid is said to be constitutionally supersaturated (if the instability results from a concentration gradient), constitutionally supercooled (if the instability results from the combination of a concentration and a temperature gradient), or thermally unstable (when a temperature gradient is the sole cause).

Thermal instability can only arise in melt growth. The condition for thermal instability can be derived from Equation (56) and the condition that $\mathrm{d}T_\mathrm{N}/\mathrm{d}x < 0$ (where x is measured normal to the growth face), which together yield an instability condition

$$K_\mathrm{S}\frac{\mathrm{d}T_\mathrm{S}}{\mathrm{d}x}|_{x=0} - \frac{\Delta H_\mathrm{C}\, f}{V_\mathrm{m}} < 0 \tag{65}$$

Elimination of the instability involves practical steps to increase the first term or decrease the second term on the left-hand side of this relation.

The instability condition for constitutional supercooling can be shown to be approximately [130, 131]

$$\frac{\mathrm{d}T_\mathrm{S}}{\mathrm{d}x}|_{x=0} < mc_\mathrm{BN}\ (1-k_\mathrm{I})\ f\, k_\mathrm{eff}/k_\mathrm{I}\, D_\mathrm{S} \tag{66}$$

where c_BN is the concentration of the dopant with distribution coefficients k_I and k_eff and a diffusion coefficient D_S. The parameter m is the slope of the liquidus (i.e., $\partial T_\mathrm{E}/\partial c_\mathrm{BN}$), where T_E is the equilibrium temperature on the phase diagram.

The effective segregation coefficient k_eff represents the influence of melt convection. In the case of complete mixing ($f\delta_\mathrm{s}/D_\mathrm{s} \ll 1$), Equation (54) becomes $k_\mathrm{eff} = k$, and only a small temperature gradient has to be established for preservation of interface stability (if $k < 1$). In motionless melts, $k_\mathrm{eff} = 1$, and this represents the highest danger of instability. Hence, convection may stabilize growing interfaces against morphological instability.

The instability condition for constitutional supersaturation can be shown to be

$$f > \frac{V_\mathrm{m} D_\mathrm{S} c_\mathrm{E}}{T}\,\frac{\partial T_\mathrm{N}}{\partial x} \cdot \left[\frac{c\Delta H_\mathrm{C}}{c_\mathrm{E} RT} - \frac{T\,\partial f/\partial T}{\partial f/\partial\{(c-c_\mathrm{E})/c_\mathrm{E}\}}\right] \tag{67}$$

which points to the role of the temperature gradient in solution growth.

The conditions given above are approximations that ignore some stabilizing forces. In particular, a face growing by nucleation with rapid spread will always be stable and the stabilizing effect of the surface free energy has been ignored. Thus, faces are usually more stable than these relations suggest. For further discussion of stability relations, see [48, 57, 59, 132].

Even if a face is potentially unstable, there is no certainty that the instability will develop. A suitable fluctuation is necessary to initiate unstable growth and it takes time for the instability to develop. For example, in thin-layer growth instabilities are rarely seen, even in systems which must be potentially unstable. However, interface instability does provide a real limit to the maximum obtainable growth rate and, at a given growth rate, to the maximum size of a growth face. (Remember that the maximum value of $\mathrm{d}T_\mathrm{S}/\mathrm{d}x$ falls with face size.)

5. Practical Considerations

5.1. Raw Materials

Many materials with fewer than 10 ppm of impurities are now available. Used with care, such materials yield crystals with about 1 ppm of impurity. Most crystal growers buy their raw materials from specialist suppliers, who supply various grades. Typically, the price paid doubles for each factor of ten reduction in impurity content. Purification is sometimes necessary; thus, most organizations using water as a solvent purify

their own water. A process sequence of filtering, reverse osmosis, deionization, and ultrafiltration is typical. In other cases, raw materials may be transferred to the growth system by distillation, which typically removes more than 90 % of the impurities present. When uncommon materials are required, in-house processes become essential, and crystal growth groups often distill or zone refine some of their raw materials.

5.2. Temperature Control

Spatial variations (temperature distribution) involve both uniformity and controlled nonuniformity. To produce regions with low temperature gradients requires the use of structural components with large effective thermal conductivities. At the lowest temperatures employed for crystal growth, stirred fluid baths are widely used. For higher temperatures (over 100 °C), metals (e.g., aluminum up to about 500 °C, copper to somewhat higher temperatures, and various steels up to about 1200 °C) can be useful. Above about 400 °C, liquid-metal filled heat pipes can be helpful, but with a given heat-transfer fluid, the optimum performance of a heat pipe exists only over a small temperature interval. Obtaining desired gradients over limited distances is more difficult than obtaining a uniform temperature. Essentially, the problem is one of the heat flow that must accompany the gradient. Structural materials with low thermal conductivities (ceramics, mineral fiber wools, and even evacuated spaces) can play a part, but heat transfer in nutrient materials (particularly fluids) and in the crystals being grown can be significant, and considerable ingenuity may be needed to obtain the desired distribution without producing radial gradients that are harmful to the crystal (Section 4.5).

Fully computer controlled heating systems, consisting of cylindrical multizone side heaters and top and bottom heaters, surrounded by insulating materials, are today of increasing interest due to their ability to realize nearly uniaxial heat flow through the growing crystal [133]. Excellent insulative properties are exhibited by modern carbon materials composed of interpenetrating layers of carbon fibers (calcarb). The very good machinability, low thermal conductivity (ca. 10^{-6} K^{-1}), and high temperature stability up to 2000 °C and above make this material very suitable for numerous high-temperature crystal growth processes that require minimal radial temperature gradients.

The availability of sensitive temperature transducers (thermocouples, resistance thermometers, and radiation sensors) together with solid-state temperature controllers has greatly simplified the designer's problems. However, the targets to be reached have also become more difficult: it is now common to demand temperatures of 1000 °C with variations of less than about 0.1 °C, and in the range 40–100 °C variations of less than 0.01 °C are often requested. The state of the art allows a tenfold improvement on these figures by use of multistage processes.

5.3. Containers and Atmospheres

Both the crystal and the nutrient from which it is growing must be in contact directly or indirectly with some solid part of the growth apparatus, and both may be in contact with an atmosphere. Thus, surface contamination and some contamination of the bulk materials, particularly in the slower growth processes, are expected, because most crystal growth processes occur at elevated temperature.

Solids in contact with the crystal and the nutrient must be selected with care. Graphite is a widely used material which is stable in contact with many solid and liquid halides and semiconductors (however, not with silicon), but reacts with many oxides. Silica (vitreous SiO_2) is used extensively but reacts slowly with molten silicon. Glasses, particularly borosilicates, are used widely, but care must be exercised because some components can be leached preferentially. The platinum metals, particularly platinum and iridium, are useful for work with oxides and halides.

For the growth of semiconductor compounds such as GaAs, InP, and GaP, pyrolytic boron nitride (pBN) has became the most widely accepted crucible and container material because of its high purity, thermal stability (up to 3000 °C in inert gas atmosphere), low outgasing, and nonwetting characteristics. In particular, using pBN as alternative to quartz in Czochralski and Bridgman growth methods (Sections 6.1 and 6.2) eliminated the problem of silicon contamination, and the semi-insulating

behavior of GaAs became apparent in production. pBN is also the most widely utilized container material for MBE sources. For HF induction coupling in the region up to 3000 °C, high-purity glassy carbon ampoules and containers have gained in significance.

In choosing solids for containers, care must be exercised in selecting the grades to be used. Materials which are satisfactory in a pure form may be unsatisfactory in technical grades. For example, pure platinum is a satisfactory container for alkaline molten oxide materials, but technical-grade platinum contains some rhodium which is easily leached out by these oxides.

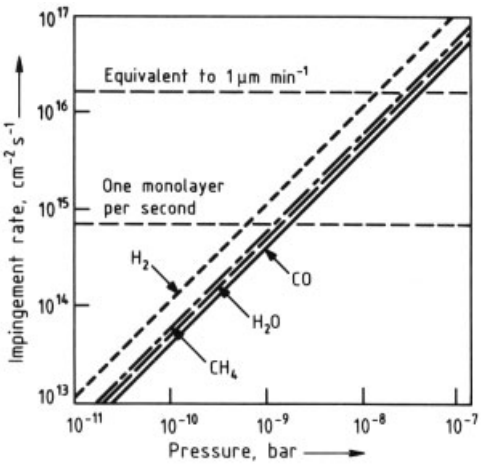

Figure 21. Impingement rates (molecules per unit area per unit time) as a function of partial pressure for typical residual gases in a vacuum system

Similar care is required in choosing gases for ambients. Water is a common contaminant of gases from cylinders (however, many glasses contain significant amounts of water so that if the absence of water is essential, the correct construction materials must be chosen). Hydrogen is widely used because it can be purified readily by diffusion through heated palladium. Vacuum is commonly thought to be a clean "atmosphere." However, typical vacuum systems contain appreciable amounts of H_2O, CO, CH_4, and hydrogen (which is difficult to pump away). Water and carbon monoxide usually come from films absorbed on the inside walls of the vacuum system, and methane can be formed by the cracking of pump oils. For further information about vacuum technology, see [134]. Figure 21 shows that relatively low partial pressures give rise to appreciable impingement rates so that one monolayer can form in one second with a partial pressure of about 10^{-9} bar. A partial pressure of about 10^{-8} bar gives an impingement rate equivalent to a growth rate of about 1 µm/min so that thin-layer growth from a vapor at 1 µm/min requires partial pressures of about 10^{-14} bar to avoid contamination at a level of about 1 ppm. Pressures of 10^{-13} bar are currently the state of the art. Thus, if all atoms impinging are incorporated, a contamination of about 10 ppm must be expected. Fortunately, most of the impinging atoms are desorbed before they are incorporated, and crystals are usually purer, by a factor of ten or more, than these estimates suggest.

5.4. Selection and Optimization of Methods

Selecting and optimizing methods of growth requires information about the methods available, the nature of the material, and its use. For example, if thin layers are needed, cutting thin sheets from bulk crystals might be very expensive, but in some cases, this may be the only approach; for example, no method for producing the α-quartz needed for resonators in thin-sheet form has yet been found. In fact, only a few, rather limited methods exist (Section 6.1) for making thin, unsupported layers. Most thin-layer growth methods make use of a substrate which cannot be removed easily.

Bulk Crystals. For growing bulk crystals, *melt growth* is likely to be the fastest method (volume per unit equipment time), but it is only applicable to materials that melt congruently and have no destructive phase changes. Also, melt growth involves the highest possible temperatures and, therefore, the maximum possibilities of contamination. Of the available melt growth methods, the Czochralski and floating-zone methods can reliably yield dislocation-free crystals, and the floating zone method can yield large crystals only of materials whose melts have a high ratio of surface tension to density. (Typical ratios are in the range 3 to 300 $cm^2\ s^{-2}$ but only materials with ratios exceeding about 100 $cm^2\ s^{-2}$ are suitable for floating zone refining.)

Due to the marked improvement in container materials (pBN, glassy carbon) and computer-controlled temperature programming, the Bridgman method has achieved a noteworthy renaissance for the relatively cheap production of high-quality crystals with excellent diameter accuracy, and dislocation densities in semiconductor compounds of less than 10^3 cm^{-2} are typical [135]). Skull melting is the cheapest method of melt growth but gives high dislocation densities in irregularly shaped crystals.

Vapor growth techniques probably give the second highest growth rates but are applicable only to materials that are volatile or give rise to reversibly decomposable compounds. Due to the low condensation temperature relative to the melting point, low-temperature single-crystalline phases can be grown directly without passing through harmful solid–solid phase transitions. This phase transition is today more and more used for the growth of materials with high melting point, such as nitrides (GaN, AlN) and silicon carbide, which are of increasing importance for advanced micro- and optoelectronics.

Solution growth techniques are usually the slowest processes and are applicable only to materials that are reversibly soluble.

Solid-phase growth techniques are not used commercially for bulk crystal production because yields are low. All the other methods of growth allow some segregation of impurities so that the crystals can be purer than the starting materials. However, for special materials with high melting points or solid phase transitions, solid-state recrystallization by exploiting the phenomenon of grain growth in thermally stressed polycrystalline samples is under investigation.

Thin Layers. For thin-layer processes, crystals can be grown from solution (liquid-phase epitaxy, LPE), from the vapor (vapor-phase epitaxy, VPE, or molecular beam epitaxy, MBE), or from the solid state (solid-state epitaxy, SSE). In the last case, the solid from which growth takes place is likely to be amorphous. Liquid-phase epitaxy is used quite widely. It is almost the only technique available to make thin layers of garnets for magneto-optic devices (bubble memories, display devices) and for single crystalline phosphors for high-intensity cathode-ray tubes. Liquid-phase epitaxy is also used for semiconductors, where it generally yields layers purer than other methods but usually with rather imperfect surfaces. Molecular beam epitaxy is rather slow but can yield complicated structures with many layers and sharp composition changes; the substrate temperature can be very low, e.g., 100 °C. Vapor-phase epitaxy using halides (VPE, CVD) or organometallic compounds (OMCVD or OMLPE) is widely employed for semiconductors. Recently, SSE has been used for semiconductors, but proper evaluation of this method requires more experience with a wider range of materials. Choosing the optimum thin-film method is particularly dependent on the use. Almost the only general statement that can be made is that MBE is unlikely to be economic for thick layers. However, it plays a leading role in the production of quantum-size devices such as quantum wells, wires, and dots.

Optimization. Method optimization usually implies an increase in scale, such as larger volumes for bulk crystals, or larger areas for thin layers. In general, large crystals have to be grown at slower rates than small ones if perfection and uniformity are to be maintained, and costs do not decrease as rapidly as they would if the rates could be maintained. Nutrient costs do not decrease with crystal size and are typically one-third of total costs. A larger apparatus uses energy more efficiently, but the energy costs are low (usually $< 10\,\%$ of total costs). Equipment and labor costs do not rise linearly with crystal dimensions, so some savings are possible. In general, automation is a paying proposition. The extra cost for equipment is usually less than the savings in labor cost, and automated equipment can produce more uniform material with higher yields than manually controlled equipment. Operator errors can easily result in losses that exceed 20 %.

5.5. Characterization

In general, the number of characterization techniques is vast and rapidly growing, so that only the main essentials can be summarized here. With resepect to crystal growth the various characterization techniques can be roughly grouped into three categories:

1) Characterization of composition and purity

2) Structural characterization
3) Characterization with respect to specific physical properties, depending on special material properties and on the application for which the material is intended

Further, analytical methods can be divided into destructive and nondestructive as well as in-situ and post-growth techniques. The development of suitable in-situ techniques is of major importance for advanced crystal growth. Whereas in the case of thin film growth, especially MBE, the in-situ observation and control of the crystallization modes (by RHEED and scanning tunneling microscopy, for example) is already state-of-the-art [136], for bulk growth from the melt this is an as-yet unresolved problem, primarily because of the lack of adequate sensors and analytical techniques. Considerable progress has been achieved in solution and vapor growth by using laser scattering, especially in holographic mode, for the investigation of diffusion boundary layer thicknesses and growth kinetics.

Today it is important that the experimental analytics are combined with modelling of the observed effects. Modelling includes both the simulation of the solid with its various specific structural features and modelling of the effects exerted by a solid with this structure on the analytical probes. Modelling is also necessary for image enhancement.

Skilled personnel can obtain a great deal of data in a nondestructive mode by just looking at a crystal. Visual inspection directly or by microscope (particularly one with an interference, e.g., Nomarski, lens) reveals surface irregularities on natural faces or other grown surfaces. Networks of grooves suggest low-angle grain boundaries. Isolated straight lines (particularly on lightly abraded faces) suggest twins. Pyramids suggest screw dislocations or stacking faults. Sets of parallel lines suggest concentration striae. On etched surfaces, pits probably indicate dislocations or vacancy clusters. For transparent materials, light-scattering centers in the bulk may be inclusions. Many crystals are transparent in the infrared, and infrared microscopy is a powerful tool. Transparent crystals can also be examined by using polarized light, which readily reveals built-in strains. Such strains, indicating large-scale effects resulting from the simultaneous existence of more than one type of growth face, inclusions or dislocations or internal strain arising from dislocation or impurity distribution, can be revealed by X-ray topography (see, for example, [137]). Similar data can also be obtained by electron microscopy. It is often necessary to know the compositions of crystals, especially the concentrations of deliberately added dopants and accidentally occurring impurities. The spatial variations of composition are also of interest. Thus, it is desirable to find methods which work on small samples. Table 3 gives typical data for the methods commonly used. When the sample size is given as an area times a length, the length is normal to the sample surface. Other methods are less generally used. For further data, see [19, 48, 138, 139].

6. Melt Growth Techniques

6.1. Crystal Pulling

The order of the following sections reflects the amount of use that is made of each method; the Czochralski method is used to a greater extent than the other four techniques.

Czochralski Method. The basic sequence of operations in the Czochralski method is shown in Figure 22. A seed crystal is dipped into the melt, and a small amount of the seed is allowed to melt. Pulling is then begun, and the melt temperature is adjusted so that the first portion grown has a diameter just smaller than that of the seed. For dislocation-free growth, the diameter is then reduced to about 2 mm by raising the melt temperature slightly. This diameter is maintained for about 10 mm and then increased to the final target value (dislocations, except those exactly parallel to the seed axis, grow out while this neck is grown). When dislocation-free growth is not required, the diameter can be increased after the first 2 or 3 mm of growth. Decreasing the melt temperature increases the crystal diameter and vice versa. Increasing the rate of pull or the relative rate of rotation of the crystal decreases the diameter. The word relative is used here because many workers may rotate the seed and the crucible in opposite directions. Some workers also raise the crucible to maintain the liquid surface at the same height relative to the apparatus.

Table 3. Methods for determining composition

	Method	Sample size	Destructive?	Detection limit (atom fraction)
a	Gravimetry	20 mm^3	yes	10^{-5}
b	Titrimetry	10 mm^3	yes	10^{-5}
c	Electro-chemical methods	10 mm^3	yes	10^{-4}
d	Isotope dilution	2 mm^3	yes	10^{-4}
e	Atomic absorption spectrometry	20 mm^3	yes	10^{-7}
f	AAS (with graphite furnace)	2 mm^3	yes	10^{-8}
g	Solid-source mass spectrometry	2 mm^3	yes	10^{-8}
h	Neutron activation	0.2 mm^3	yes	10^{-9}
i	Charged-particle activation	2 $mm^2 \times 100\ \mu m$	yes	10^{-9}
j	X-ray fluorescence	25 $mm^2 \times 20\ \mu m$	no	10^{-5}
k	Rutherford backscattering	1 $mm^2 \times 2\ \mu m$	no	10^{-3}
l	Electron-probe (X-ray spectrum)	4 $\mu m^2 \times 2\ \mu m$	no	10^{-3}
m	Secondary-ion mass spectrometry	0.1 $mm^2 \times 10$ nm	yes	10^{-9}
n	Auger spectrometry	100 $\mu m^2 \times 1$ nm	no	10^{-2}
o	X-ray photo-electron spectrometry	10 $mm^2 \times 1$ nm	no	10^{-3}
p	Low-energy ion scattering	1 $mm^2 \times 1$ nm	no	10^{-4}

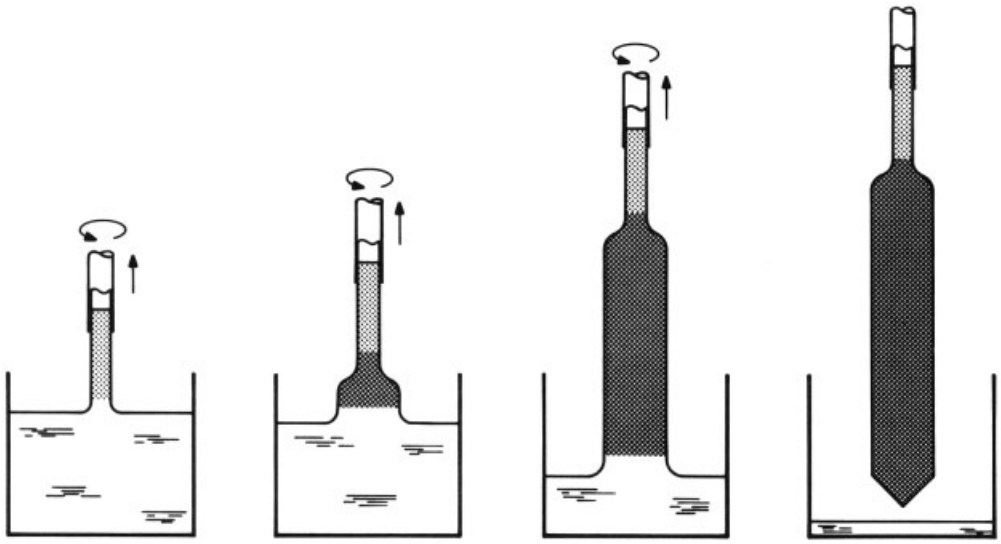

Figure 22. Sequence of operations in the Czochralski method

This is particularly convenient when the system for controlling diameter employs the reflection of a light beam on the curved portion of the melt near the edge of the crystal. The other commonly used method of diameter control involves weighing the crystal or the melt.

The rotation and pulling rates used depend on the material being grown and the diameter of the crystal. Larger crystals need slower pulling and rotation. The maximum pulling rate is roughly inversely proportional to the diameter to the power 1.5. Typical rates for 2-cm diameter crystals range from 3 mm/min (silicon and germanium) to 3 mm/h (oxide crystals with low thermal conductivity). Rates as high as 12 mm/min and as low as 1 mm/h have been reported. Crystal rotation rates tend to be proportional to the reciprocal of the crystal diameter. For 2-cm diameter crystals, rates of 5–50 rpm are typical. Crystals of silicon up to 30 cm diameter are pulled (40-cm technology is under development); alkali metal halides can have much larger diameters.

When electrically conducting crucibles are used, power can be supplied by radio-frequency heating as shown in Figure 23, which also illustrates the use of an encapsulant to prevent decomposition of materials with a volatile com-

ponent. When nonconducting crucibles are used, radio-frequency power may be coupled directly to the melt, but more usually a resistive heater coaxial with the crucible is used. Such heating elements can be made from graphite, silicon carbide, or molybdenum or from resistance wires wound on ceramic formers.

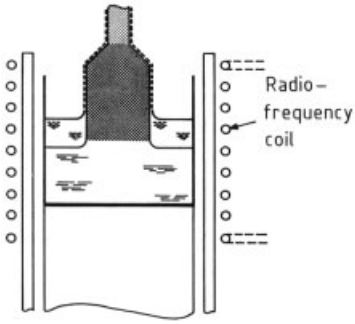

Figure 23. Liquid encapsulated growth by the Czochralski technique (LEC)
This method is used to grow a material with a high partial pressure of one or more component. Use of an encapsulant (usually boric oxide, but occasionally a halide), shown dotted, prevents loss of volatile material from the melt and from the crystal if the pressure in the growth chamber exceeds the sum of partial pressures of the components

Usually, the conventional Czochralski technique is characterized by high temperature gradients and thermal nonlinearities in the growing crystal, which are harmful for materials with low critical resolved shear stress (semiconductor compounds). To decrease the temperature gradients, afterheating systems are used. To prevent dissociative decomposition of the hot crystal surface, in such cases, a closed inner growth chamber with a controlled atmosphere of the volatile component can be used [140].

A comprehensive overview on the Czochralski technique is given in [141].

Kyropoulos Method. The Kyropoulos method uses a seed crystal that is dipped into the melt, but growth is maintained by continuously lowering the crucible temperature (Fig. 24). Linear growth rates are low, but the area of the growth face is large so that volume growth rates are reasonable but significantly lower than those obtained by the Czochralski method. The Kyropoulos method is widely used to produce halide crystals for optical components. Crystal diameters and lengths are usually comparable. Crystals with dimensions greater than 50 cm have been grown. The major practical problem appears to be seed breakage. For this reason, seeds and crucible rotations are sometimes not used. Dislocation densities tend to be high ($> 10^4$ cm^{-2}).

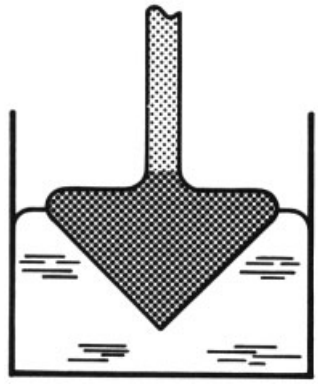

Figure 24. Kyropoulos method

Edge defined film-fed growth and the Stepanov method, which involve pulling crystals through shaped apertures, are shown schematically in Figures 25 and 26. The methods differ in their use of capillary action: in edge-fed growth, the die is wetted by the melt.

Edge defined film-fed growth (EFG) can be used to make single-crystal sapphire pipes but is employed on a large scale to make silicon ribbons for solar cells. Also thin-walled octagonal tubes are commercially produced by the EFG technique; they show reduced thermomechanical stress compared to edged ribbons. Octagons are then cut into wafers 10×10 cm, 300 µm thick with high-speed lasers. Crystalline quality is not usually important in such uses, and the plates, rods, and tubes made this way do contain some low-angle grain boundaries. Crystals grown by the Stepanov method are usually more crystallographically perfect but still tend to have large dislocation densities because of the large temperature gradients near the growth face.

Because the shaped crystals grown by these methods generally have at least one small dimension normal to the growth axis, high linear rates of growth (10 mm/min or more) can often

be maintained. With the Stepanov method, the crystal need not be pulled vertically upward, and equipment in which crystals are pulled vertically downward or horizontally has been described.

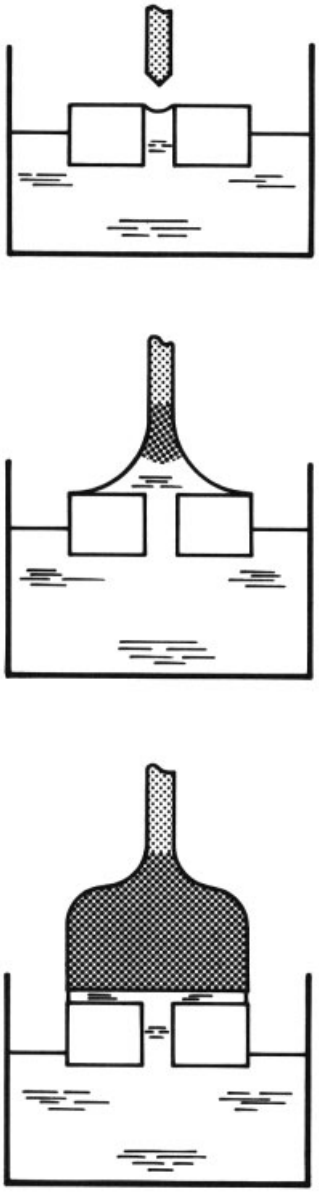

Figure 25. Edge defined film-fed growth

A special branch of high-speed crystalline fiber pulling through wetting and nonwetting dies with micro-orifices has been developed [142]. Small-diameter (50–300 μm) long single-crystalline fibers of numerous oxide materials (e.g., sapphire, $LiNbO_3$, $KNbO_3$) are of considerable interest for miniaturization, especially of multifunctional nonlinear optical devices.

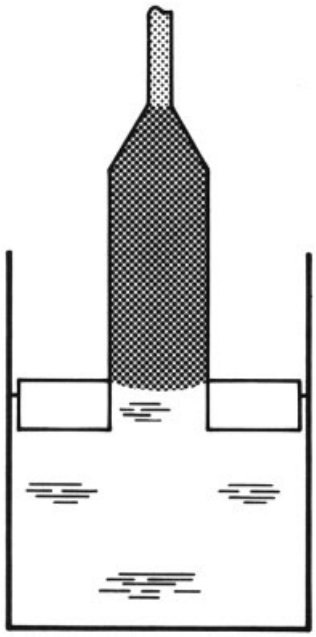

Figure 26. Stepanov method

More details on shaped crystal growth are given in [143, 144].

Dendritic Crystallization. Around 1960, the use of dendritic crystallization was expected to provide the thin sheets of semiconductors needed to make transistors. However, the crystal quality was inadequate: dislocation densities and impurity concentrations were too high. Some workers believed that the material might be useful for solar cells. Certainly, the linear growth rates are attractive (rates of up to a few meters per minute have been reported). The method uses a multiply twinned seed crystal to grow tapelike crystals less than 1 mm thick. In diamond and zinc blende structure semiconductors, the preferred growth direction is <211>. In a typical process, a correctly oriented seed crystal is touched on the surface of a melt about 2 °C above its melting point. The melt temperature is then rapidly reduced by about 20 °C, and

pulling commences. An ingenious modification of the method uses two seeds to pull a so-called web crystal (Fig. 27).

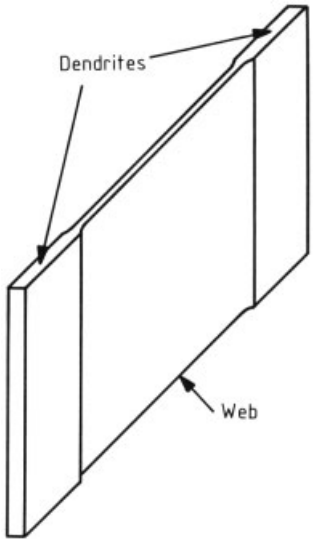

Figure 27. Web growth

6.2. Bridgman Method

A crucible (Fig. 28) is lowered through a temperature gradient which straddles the melting point T_M. If a seed crystal is used and the convex growth face shown can be maintained, a single crystal with the same shape as the crucible should result. The seed crystal can be omitted, and a long, narrow tube can be substituted at the lower end, for example. (If several crystals nucleate, only one will probably emerge from a long, narrow tube.)

In practice, two particularly important problems exist. The first is that it is difficult to find crucibles with the same thermal expansion coefficient as the crystal. If the crucible has a larger coefficient, it will damage the crystal during cooling. If the crucible has a smaller coefficient, damage will occur if the crystal sticks to the crucible.

The other major problem is to obtain and maintain the desired temperature distribution (i.e., the interface shape shown). This requires crucibles with thermal conductivities that are less than those of the materials grown.

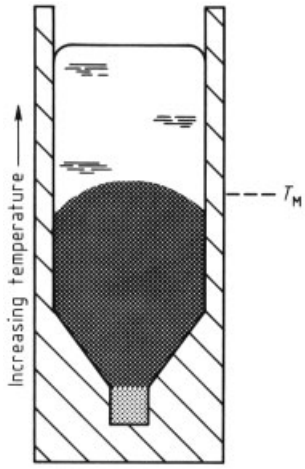

Figure 28. Bridgman crucible
Note the hatched seed crystal

Adding to these needs, the general requirement of no contamination makes crucible selection difficult. Common choices are graphite and silica which is often given a soft carbon coating. However, a variety of materials have been used, e.g., molybdenum for the growth of sapphire, iridium for $FeAl_2O_4$, boron nitride for gold, polyethylene for sodium nitrite, and polyester (e.g., Mylar) for ammonia.

Due to the rapid development of high-temperature construction materials in the 1990s and their application as growth containers (pBN, glassy carbon, etc.; see Section 5.3), the above-mentioned problems were solved for semiconductor compounds (GaAs, InP, GaP). Moreover, the sticking problem was suppressed by using a covering liquid layer (boron oxide) between the crucible wall and growing crystal. These technological improvements, together with the high diameter constancy and low defect density, led to a considerable renaissance of the Bridgman technique, which has become a serious competitor to the Czochralski method in the modern electronics industry [122, 135, 145, 146]. To prevent constitutional supercooling (Eq. 64), the growth rates (1–10 mm/h) are about one order of magnitude lower than in Czochralski growth because of the much lower axial temperature gradients used. Contemporary reviews on Bridgman growth of semiconductor compounds compared with Czochralski method are given in [122, 147 – 149].

Horizontal Boat Method. This method was established in the 1950s for GaAs growth and is still the most widely used production method for Si-doped material for optoelectronics. Pretreated (sand-blasted or etched) quartz and pBN boats are used. Usually, the charge is produced by in-situ synthesis in a separate ampoule, in which the volatile component (As) is evaporated and reacts with gallium held at a suitable temperature. Either the charge or furnace is moved (Fig. 29).

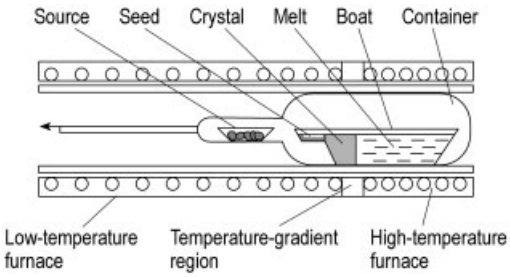

Figure 29. Horizontal boat configuration for the growth of semiconductor compounds

The major advantages are the possibility of stoichiometry control by means of the pressure of the volatile element from an additional source and the free growth of the shoulder region, which results in a low dislocation density of about 10^2 cm^{-2}. Main disadvantages are the noncircular cross section and the limited width of the shoulder. A review on semiconductor compound growth by this method is given in [150]. The succesful horizontal growth of single-crystalline sapphire and YAG plates by applying a restistance tungsten heater is described in [49].

Vertical Bridgman Method. The vertical bridgman method is characterized by the translation of the crucible containing the melt relative to the axial temperature gradient in a vertical furnace [151 – 153]. Whilst it is generally simpler in a mechanical sense to lower the crucible through the gradient, in some cases where vibration of the ampoule is a concern, the crucible is held stationary and the furnace is moved [135]. In the Bridgmann–Stockbarger mode [19], a high-temperature upper zone is followed by an adiabatic middle zone (for locating the position of the growing interface) and a low-temperature lower zone for afterheating and nearly gradient-free cooling. A wide range of materials (metals, semiconductors, oxides) are grown by vertical directional solidification. By connecting a third furnace in series, the stoichiometry of the growing crystal can be controlled by means of the partial vapor pressure of the volatile component, which has been succesfully applied to II-VI growth, for example [154].

Gradient Freeze. In this configuration both the ampoule and the furnace are held stationary, and the freezing isotherm is moved by computer-controlled modification of the furnace temperature (Figure 30). A multifurnace system consisting of 5 to 10 (a maximum of 50 has been described [135]) independently controlled heating zones is used to tailor the temperature field during growth (dynamic gradient freezing). Presently, high-quality semi-insulating GaAs crystals up to 10–15 cm in diameter with low dislocation densities in the range from 500 to 5000 cm^{-2} are grown commercially for high-frequency microelectronics [145, 146]. The reasons for such low dislocation density are the extremely low thermomechanical stress conditions (no sticking, no mechanical vibration, nearly uniaxial heat flow).

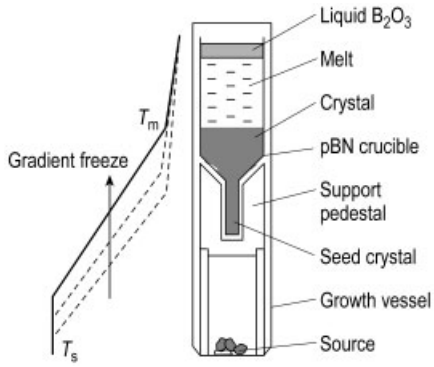

Figure 30. Vertical gradient freeze of semiconductor compounds (GaAs, InP) with stoichiometry control by means of a source of the volatile element (As, P) at T_s

Heat Exchanger Method. This container method, called HEM, is a modified directional solidification technique developed in the 1970s mainly for growth of large sapphire crystals from the melt [155]. But photovoltaic silicon, CdTe and BGO crystals have also been grown by this

method. Large crystals can be grown by HEM since it has a unique feature for independently controlling the heat input and heat extraction, as shown in Figure 31. A crucible with a single-crystal seed centered at the bottom is loaded with sapphire meltstock and placed on a high-temperature, helium-cooled heat exchanger. During melting the seed is cooled by helium gas flowing through the heat exchanger. Then, after melting and homogenization, the growth is initiated and progressed by increasing the helium flow and decreasing the furnace temperature. After complete solidification the furnace temperature and helium flow are reduced to achive in-situ annealing of the boule before the crystal is cooled in a controlled manner.

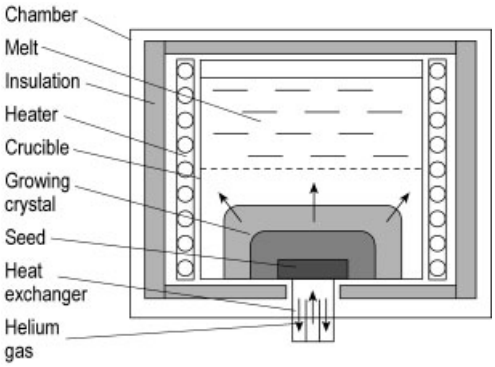

Figure 31. The principle of the heat-exchanger method

To remove trapped ambient gas from the liquid, the melt is overheated above the melting point without affecting the well-cooled seed. Morover, very low temperature gradients can be achieved. These features have allowed production of light-scatter-free sapphire crystals of 34 cm diameter and 65 kg weight. Growth of 50-cm diameter crystals is under investigation. Numerous cylindrical crystals can be prepared from such boules by core drilling.

6.3. Skull Melting Methods

Skull melting processes involve melting a powder charge (usually weighing over 50 kg) in a water-cooled container so that the melt is in contact only with an outer skull of sintered powder. When the system is cooled slowly, crystals nucleate at random and then progressively grow [156]. Figure 32 shows a version of the process using radio-frequency heating. Because the charge usually has an extremely high resistivity at room temperature, this version uses some lumps of the metal whose oxide is being grown to couple the radio-frequency energy to the charge. An alternative process [157] uses an arc to provide the power needed. The process is not very controllable and usually produces crystallographically imperfect crystals. However, for high-melting materials (MgO, CaO, SrO, BaO, and ZrO_2) the method has significant advantages and is used, for example, to grow about 1000 t of cubic zirconia per annum. For instance, Ceres Corporation (USA) produces approximately 20 t per month in 45 inch (117 cm) diameter skull furnaces holding 2.5–3.5 t of charge. The weight of the individual single crystals reaches 3–5 kg. At room temperature, pure ZrO_2 is monoclinic. The cubic form, which is stable at high temperature, can be made stable at lower temperature by doping (e.g., with CaO or La_2O_3). In addition, doping produces almost any color and this, combined with its hardness, large refractive index, and high dispersion, makes zirconia very attractive as a jewel. For further information on the growth, properties, and applications of this and other jewels, see [158].

6.4. Zone Melting

The methods described in Section 6.2 involve melting the entire charge and then progressively freezing it. However, the whole charge need not be molten. A crystal can also be grown by moving a molten zone through the charge [159]. This has the advantage that only part of the system must be at a high temperature at any time.

When the growth process is started with a charge having a uniform concentration c_0 of a solute with an effective distribution coefficient k, then the concentration in the crystal c_S at a distance x from the initial growth face is

$$c_S = k\, c_0 (1 - x/L_0)^{k-1} \tag{68}$$

where L_0 is the length of the charge and the process considered is one in which the whole charge is initially molten and then frozen. However, if a zone process is used with a zone length L, then for $x < L_0 - L$,

$$c_S = c_0 \left[1 - (1-k) \exp\left(-k\, x/L\right)\right] \tag{69}$$

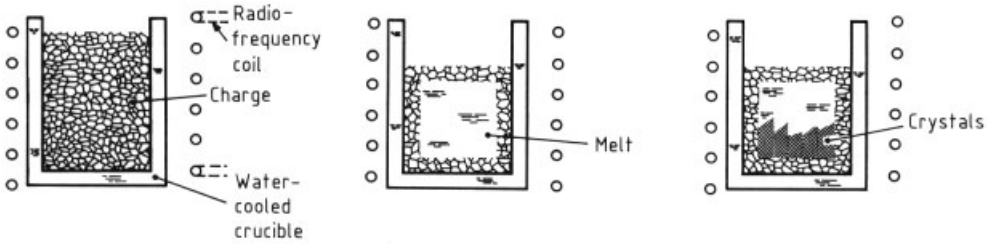

Figure 32. Skull melting. The sequence of operations starts at the left

For $x > L_0 - L$, Equation (65 applies. Figure 33 gives an example of c_S as a function of x. The broken curve is the distribution obtained from Equation (65), i.e., the usual Bridgman case, which is referred to as normal freezing. The full curve labeled $n = 1$ is given by Equation (66) and is usually called a single-pass distribution. If the zone melting process is repeated (i.e., movement of another zone through the same bar), the distribution given by the curve for $n = 2$ is found. Further repetition gives curves for $n = 3$, 4, 5, etc., and an ultimate distribution is obtained which is shown by the line labeled $n = \infty$. These multipass distributions are the basis of zone refining [159], which is a useful method for purification provided the impurities have $k \lessapprox 0.2$. Designing a zone refining process for purification is not easy. A set of impurities with different k values must usually be considered. However, a process for which $n\,L/L_0$ is between 0.5 and 1.5 is a reasonable starting point.

In some processes, the charge is essentially pure, but doped crystals are to be produced by adding a dopant to the zone at the start of the process. If the concentration of dopant is c_0 in the zone, then the concentration in the solid is

$$c_S = k\,c_0 \exp\left(-kx/L\right) \tag{70}$$

If uniformly doped material is needed, the process is designed so that kx/L is small; in this way, crystals with very small gradients of dopant concentration can be obtained. When k is small, these gradients can be much less than those given by normal freezing, as shown when Equations (70) and (68) are compared.

The apparatus for zone melting a charge in a boat is fairly simple to design. All that is needed is a source of power which can heat a short length. Either the source or the boat can be moved.

In some cases, zone melting can be carried out without a boat [154]. A molten zone can exist between two solid rods, as shown in Figure 34. The power to maintain the molten zone can be supplied by electron bombardment, arc image systems, mirror-focused optical or thermal radiation, or laser beams. However, the use of radio-frequency heating as shown is the most popular method.

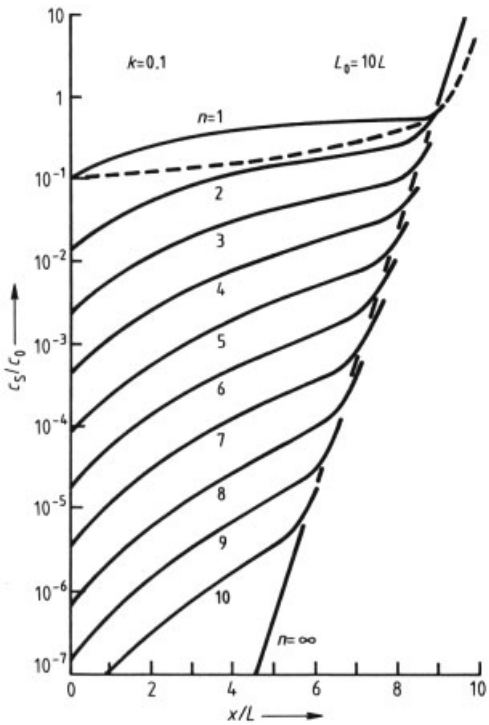

Figure 33. Concentrations (c_S) in the solid as a function of the initial charge concentration (c_0) and position x

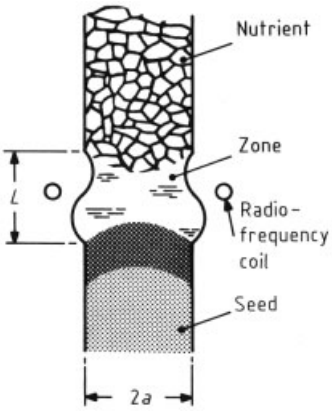

Figure 34. Floating zone
For explanation of symbols, see text

If the zone starts in contact with a seed crystal and is then moved away from the seed, it can melt a polycrystalline nutrient and deposit a single crystal. The advantage of this method is that the molten material is only in contact with its own solid and a vapor phase (which, at least in principle, does not contain harmful impurities). The method has obviously desirable features for the growth of crystals of many reactive materials, but the only large-scale use is for silicon [160]. Oxygen-free silicon in bulk form can be grown only by the floating zone method.

If gravity and surface tension are the only forces affecting the zone size, then the parameter which determines stability is the zone length L (not the rod radius a), based on the assumption that the melt perfectly wets the solid and that the interfaces are flat. If γ is the surface tension (about 0.72 J/m^2 for silicon) and ρ is the density, the maximum zone length is

$$L_{\max} = K(\gamma/\rho g)^{1/2} \tag{71}$$

where K is a constant for which values of 2.5–3.5 have been given [160]. In practice, zone lengths lie between about 0.5 a and 2 a so that, as a first approximation, the maximum zone radius is

$$a_{\max} \approx 3(\gamma/\rho g)^{1/2} \tag{72}$$

For silicon $(\gamma/\rho\, g)^{1/2}$, the capillary constant, is 5.37 mm.

However, these two relations are based on rather doubtful assumptions, and a more realistic model should take account of the extra stabilizing forces resulting from electromagnetic effects when radio-frequency heating is used and the destabilizing force resulting from rotation. In a typical case, the electromagnetic forces exceed the surface tension forces. The pressure resulting from rotation is about

$$p = \rho \int_0^a \omega^2\, r dr$$

If this force should not exceed 10 % of the surface tension force (roughly γ/a), the rotation rate should not exceed

$$\omega_{\max} \approx (\gamma\rho/5\rho)^{1/2} a^{-3/2} \tag{73}$$

which for 10-cm diameter silicon gives a rotation rate of about 6 rpm. In general, Equation (70) appears to give values in the range used experimentally.

Global numerical models show that process optimization is sensitive to the shape of the field coil and the mass flow in the molten zone [161 – 163]. In addition to the flow caused by the relative rotations of the seed and nutrient rods, the large temperature gradients lead to density-gradient and surface tension-gradient (Marangoni) convection. There will also be electromagnetic forces causing flow when radio-frequency heating is used. If the aim is to produce uniformly doped silicon crystals, the best results appear to be obtained with short zones ($L \leq a$), rather low rotation rates (about half the value of $\omega_{\max}$), and growth rates in the range 2–4 mm/min. As with crystal pulling, dislocation-free growth can be obtained by growing a narrow neck from the seed. Under these conditions, about 800 t of silicon up to 15 cm in diameter is produced each year. Float zone material is usually purer (lower oxygen concentration) and tends to be used for high-voltage (power) devices, whereas Czochralski material is used for low-voltage devices, which have a much larger market.

The floating zone method has been used in research for a large number of materials, but apart from silicon, its commercial use appears to be limited to a few metals grown at rates of a few kilograms per year. The optical zone heating technique in mirror furnaces is of special importance in space experiments due to its low power consumption. It is also widely used in fundamental research for growth of dielectric substances [161]. Laser heating arrangements are used for production of oxide fiber crystals by miniature-zone floating [164]. For further data, see [48, 54, 142, 154, 160].

6.5. Verneuil Process

The traditional Verneuil apparatus (shown schematically in Fig. 35) uses an oxygen–hydrogen flame to melt a powder that is sprinkled into one or more of the gas streams. Other flammable gases (acetylene, synthesis gas, natural gas) can be used, and arcs, arc images, and plasmas have been used as the power source. The melt exists as a thin ($\lesssim$ 1 mm) film on the upper surface of the growing crystal. The powder must melt before or immediately after hitting the surface of this film. The process is controlled by the supply rates of heat and powder.

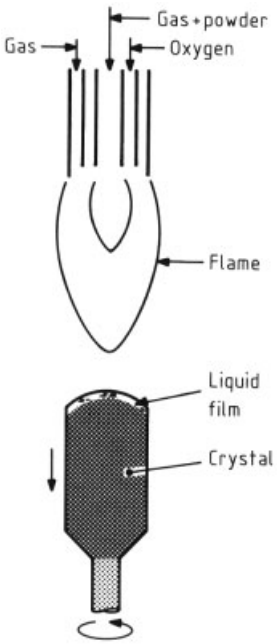

Figure 35. Verneuil process

The axial dopant and impurity distribution in Verneuil crystals follows the zone growth relation of Equation (66). This is due to the constancy of the liquid film volume continuously fed by molten source material.

An obvious refinement of the process is to surround the crystal by a furnace and so limit the temperature gradient in the crystal. This allows growth, for example, of ruby with dislocation densities below 500 cm^{-2}, but usually dislocation densities exceed 10^4 cm^{-2}. Typical growth rates range up to about 20 mm/h, but decrease roughly linearly as the crystal radius increases. Usual seed rotation rates are 5–50 rpm, with larger crystals requiring the slower rates. Diameters are, by modern standards, quite small: 20–50 mm is typical.

Up to the era of electronic clocks and watches, the largest use of the Verneuil process was to produce sapphire (undoped corundum) or preferably the harder ruby (chromium-doped corundum) for bearings in clocks and watches, at rates of up to 700 t/a. Because the hardness of the crystals increases, high dislocation densities could be an advantage for this type of application, although in practice, crystals with high dislocation densities tend to shatter when cut. Some ruby is still used for bearings, but the largest current use is for jewelry (a few hundred tons per year). Many refractory oxides and some other materials (e.g., niobium and silicon) have been grown for research purposes. Commercially available Verneuil-grown crystals include titania, strontium titanate, emerald, and a variety of spinels. The growth of 500 mm long laser crystals with a diameter of 40 mm has been reported [49].

7. Solution Growth Techniques

7.1. General Aspects

The properties of an ideal solvent are as follows:

1) Moderate reversible solubility of the material to be grown
2) Reasonable positive temperature coefficient of solubility
3) Production of crystals with useful habit
4) No formation of solid solutions with the solute
5) Low viscosity
6) Low volatility (except when evaporative methods are used)
7) Easy removal after growth
8) Compatibility with available containers
9) No reactivity with the gas phase
10) Ready availability
11) Inexpensiveness
12) Absence of toxicity

It is rare to find a solvent with all these properties. Many solvents are used for low-temperature

solution growth (Table 4 lists some examples), but the only commercially significant solvent is water. High-temperature solution growth is employed on a much smaller scale than low-temperature processes. Table 5 contains some examples. In commercial processes, lead salts are used widely for the growth of oxidic materials. The other popular solvent is gallium, which is used for the growth of GaAs and similar materials.

Table 4. Non-aqueous solvents for low-temperature solution growth

Solvent	Solutes
Acetone	acridine, trilaurin
Ethanol	hexamethylenetetramine
Carbon tetrachloride	benzophenone
Carbon disulfide	sulfur
Mercury	tin, iron

Table 5. Solvents for high-temperature solution growth

Solvent	Solutes
Lead difluoride, PbF_2	$ZnAl_2O_4$, ZnO
Lead oxide + lead difluoride, PbO + PbF_2	$Y_3Fe_5O_{12}$ and analogs, Al_2O_3, In_2O_3, $LiGaO_2$
Lead oxide, PbO	Al_2BeO_4, ferrites
Lead oxide + boric oxide, PbO + B_2O_3	$Y_3Al_5O_{12}$ and analogs
Bismuth oxide + boric oxide, Bi_2O_3 + B_2O_3	$GaFeO_3$, $Pb_3MgNb_2O_9$, garnets
Bismuth oxide, Bi_2O_3	$Bi_4Ti_3O_{12}$
Lead tungstate, $PbWO_4$	$Be_3Al_2Si_6O_{18}$
Potassium carbonate, K_2CO_3	$KTa_x\,Nb_{1-x}\,O_3$
Sodium tetraborate + boric oxide, $Na_2B_4O_7$ + B_2O_3	TeO_2
Potassium fluoride, KF	$KNbO_3$, $BaTiO_3$
Silicon, Si	SiC
Platinum, Pt	B
Gallium, Ga	GaAs, GaP, GaN, ZnTe
Tin, Sn	GaAs, ZnTe
Iron or iron + nickel, Fe or Fe + Ni	diamond
Tellurium, Te	CdTe

The supersaturation needed to produce crystal growth is usually achieved in one of three ways: cooling, solvent evaporation, or temperature gradient transfer. In this last method, solid nutrient material is placed in contact with the solution at one temperature, while the crystal grows in a region where the solution temperature is lower. The solute concentration c is assumed to be a function of temperature approximated by

$$c = c_0 \exp\left(-\Delta H_S / RT\right) \tag{74}$$

where c_0 is a constant and ΔH_S is the enthalpy of solution for simple materials.

The solvent is actively involved in the growth process. Changing the solvent can have marked effects on the crystals; e.g., the shape may change. In many cases, small amounts of other materials are added to solvents to produce the desired effects; for discussion, see [3] and [26].

7.2. Low-Temperature Methods

Figure 36 shows an apparatus suitable for low-temperature solution growth by solvent evaporation.

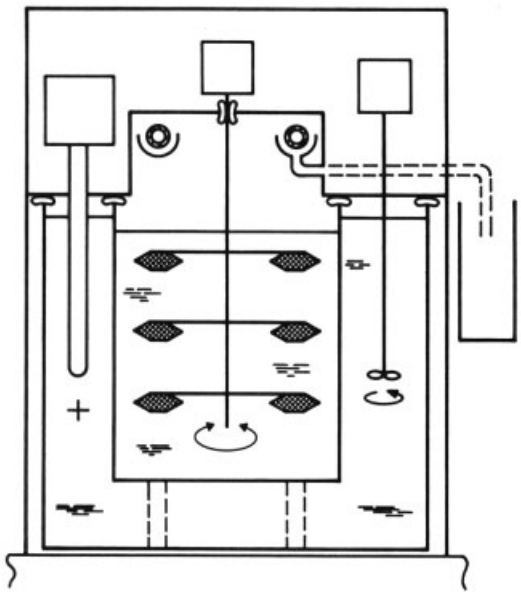

Figure 36. System for low-temperature solution growth by solvent evaporation

The basic components are an outer draught screen, an outer water bath with a propeller stirrer, and an immersion heater. A thermistor (+) below the immersion heater actuates the control system. In the inner solution bath, the crystals grow on seeds cemented to a "spider" which is rotated about the central axis alternately in clockwise and counterclockwise directions (one to four revolutions in each direction). The system has gaskets to seal all the joints; this is particularly necessary with deuterated solvents which are contaminated rapidly by atmospheric moisture.

By omitting the cooling pipe (shown hatched) and the gutter, the same apparatus could be used

for growth by slow cooling. The solution surface is often covered with an immiscible liquid to suppress evaporation; e.g., a layer of a silicone oil on the surface of an aqueous solution allows the use of high growth temperatures which are not practical when evaporation at the free surface is high. In slow cooling methods, the rates of cooling are usually less than 2 K/d. For a given system, the rate depends on the ratio of the area of the growth face to the volume of the solution and on dc/dT.

Figure 37 shows an apparatus for temperature gradient growth.

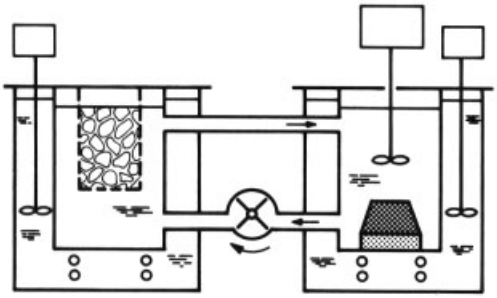

Figure 37. Growth by temperature gradient transfer

The system consists of two vessels in water baths, with appropriate stirring and temperature control. The crystal grows on a seed in the right-hand vessel at a temperature T_g, while in the left-hand vessel solution is saturated at a temperature $T_g + \Delta T$ by solid nutrient in a porous container. Because the centrally located pump can sweep undissolved solute from left to right along the upper pipe, some workers insert an extra tank at a temperature greater than $T_g + \Delta T$ at some point along the upper pipe. Cooling on the upper pipe and heating on the lower one may have to be provided if the temperatures in the two vessels are to be uniform.

In such systems the pump is the critical component: it can easily cause nucleation. Thus any pump should be placed in the flow from the crystallization tank. Pumps should be chosen to produce minimum pressure changes because nuclei tend to form in low-pressure regions.

Typical growth rates are small — usually less than 1 mm/d and occasionally below 0.01 mm/d. Hence, apparatus reliability is vital, and motors and bearings, for example, are usually much more robust than appears necessary for the loads borne. The possibility of power failures must be considered, and backup facilities capable of coping with the longest likely failure must be provided.

Numerous important oxide crystals, especially for applications in conventional and nonlinear optics, piezo- and pyroelectric devices, have long been grown commercially from aqueous solution. By controlling the pH of the solution very large perfect ammonium dihydrogenphosphate (ADP), potassium dihydrogenphosphate (KDP), potassium sodium tartrate (KNT), triglycine sulfate (TGS), and other crystals of dimensions 15 × 15 × 56 cm weighing up to 20 kg (ADP, KDP) can be produced by modified temperature transfer techniques (Fig. 37) in growth periods of up to 6 weeks [4].

Growth from aqueous solutions is becoming increasingly important for the production of crystalline biological macromolecules such as proteins. Commonly, microdialysis cells or liquid–liquid diffusion of a solvent or solute into a macromolecular solution are used. Crystallization results from gradual change in a property of the protein solution, such as pH. Crystals of small dimensions in the range of some hundred micrometers are obtained, suitable for structural X-ray and neutron diffraction analysis. However, the crystallization of biological macromolecules is a complicated multiparameter process, and extensive efforts are still required to understand the interactions of protein molecules in solution that govern nucleation and crystallization. More detailed information about this new important multidisciplinary branch of crystal growth are given in [165, 166].

7.3. High-Temperature Methods

High-temperature solution growth is sometimes called *flux growth* or *growth from a fluxed melt* because the solvents (Table 5) are often the fluxes used to dissolve oxide and other scum from metal surfaces. The principles involved are similar to those in low-temperature solution growth, but the technology differs because of the high temperatures and reactive liquids used. Solvent evaporation has been used in research, but only slow cooling and — to a minor extent — gradient transfer are used commercially.

For seeded growth, equipment is often arranged so that a seed attached to the crucible walls is not in contact with the solution until immediately prior to growth. This allows some superheating of the solution, which is necessary to ensure that no solid particles remain. When the solution is cooled to near its saturation temperature, the crucible is tilted or even rotated by as much as 180° about a horizontal axis to bring the seed into contact with the solution. At the end of growth, the tilting is reversed. To provide stirring, use is made of the accelerated crucible rotation technique (Section 4.2), in which the crucible is rotated alternately clockwise and counterclockwise about a vertical axis (Fig. 38). The construction involves welded joints at the top and bottom of the crucible. With care, the welds can be cut, the crystal removed, fresh nutrient added, a new seed fastened on, and the crucible resealed. Furnaces are made from ceramic tubes and bricks, with outer metal cases and refractory-wool thermal insulation. The heating elements are often silicon carbide rods which should not be operated at temperatures much higher than about 1300 °C.

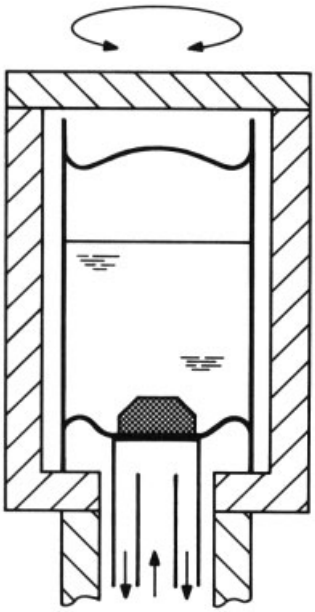

Figure 38. Crucible system for seeded high-temperature solution growth
Gas cooling is used to ensure that the seed is at the coldest point

A spherical platinum crucible, optimal for forced convection arround the vertical axis and separation of the as-grown crystal from the flux by rotating around the horizontal axis, has been used to grow high-perfection yttrium iron garnet (YIG) crystals of about 100 g [90].

Figure 39 shows a gradient transfer system that can be conveniently placed in a heated ceramic tube. In such systems, the bottom of the crucible can easily be well insulated; radiative and convective heat loss from the top of the system then maintains the indicated temperature difference ΔT.

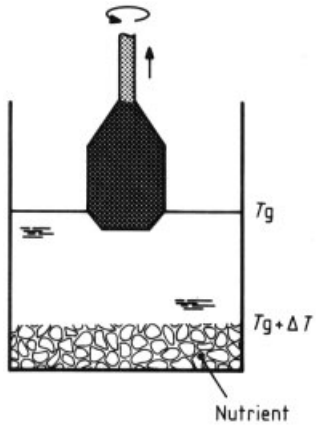

Figure 39. Crucible arrangement for seeded high-temperature solution growth by the gradient transfer method

For fluxes of low volatility, which allow the use of open crucibles, crystals can be grown at the surface of the melt when its temperature is lower than at the bottom. A seed crystal is introduced, as in Czochralski growth, but with a very small temperature gradient. The crucible temperature is continuously lowered at a cooling rate of 0.1–5 K/h. Such top-seeded solution growth (TSSG) is successfully applied to numerous perovskites and garnets.

High-temperature solution growth is widely used to produce thin crystalline plates of high-temperature superconductors like yttrium barium cuprous oxide $YBa_2Cu_3O_{7-x}$ [167]. Flux growth under high pressure is becoming important for the growth of GaN plates from Ga-rich solutions [168]. Specimens of good quality have been grown at 1400–1500 °C in 20 h under high N_2 pressure. They are attractive as homoepi-

taxial substrates for GaN thin-film deposition. However, an increase of the area from presently a few mm^2 to 100 mm^2 is required before any commercial fabrication of blue lasers on GaN substrates could be started.

Small diamond crystals are produced from graphite by dissolving in metallic fluxes (Ni, Co, Mn) at 5–6 GPa and about 1500 °C [4].

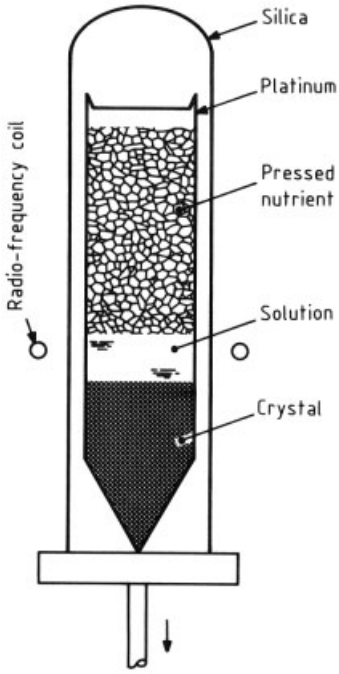

Figure 40. Traveling heater solution growth

Another form of temperature gradient growth is practiced by passing a solution zone along a rodlike bar. This can be done by making an assembly composed of a seed crystal, a disk of material with the initial solution composition, and a nutrient rod (probably of pressed powder) with the composition of the crystal to be grown. If this assembly is placed in a temperature gradient such that the seed is at the coldest point, the solution will migrate up the temperature gradient by dissolving nutrient at the hot face and depositing solid at the cold face. This process has the disadvantage that the growth temperature (i.e., the cold face temperature) rises as growth proceeds. The system shown in Figure 40, known as the travelling heater method (THM), overcomes this problem by having a stable zone length due to the use of a zone heater with a temperature maximum between the dissolving and solidifying interfaces. This method has been applied successfully to growing semiconductor mixed crystals such as (Ga,In)P and (Hg,Cd)Te with reduced dislocation density and homogeneous axial composition distribution [4, 169].

7.4. Liquid-Phase Epitaxy

Liquid-phase epitaxy (LPE) consists of a set of methods used for growing thin layers onto substrates. Epitaxy implies that the film and substrate have identical structure. However, in current usage the word means that the film is a single crystal with some crystallographic relation to the substrate. Often, but not always, the film and substrate have a common symmetry axis. Films on substrates can be strained elastically (Section 2.3). This strain can be so large that brittle materials crack and other materials suffer plastic deformation, giving rise to edge dislocations. These usually have a U shape, with the open end at the substrate surface and the closed end at a distance of about 2 h_c from that surface. It is possible to show that dislocations will not occur until the layer has a thickness h_c such that

$$h_c = \frac{3\,ba}{32\,\Delta a}\,[\ln(h_c/b) + 1] \tag{75}$$

for an abrupt change of the lattice constant by an amount Δa. Complete relief of the elastic strain then suggests a dislocation density

$$N_D = 4(\Delta a/ab)^2 \tag{76}$$

but this neglects the energy needed to create the dislocations, and in general N_D is much lower than predicted by Equation (73). The discrepancy is usually a factor of 10 to 100. A reasonable empirical estimate is

$$N_D \approx 0.1(\Delta a/ab)^2 \tag{77}$$

In practice, abrupt junctions are rare: some interdiffusion will occur and when the strain is $\sigma = \Delta a/a$, then the same geometrical arguments used to derive Equation (73) show that a dislocation density of $N_D = (1/b)\partial\sigma/\partial h$ is obtained, where h is the distance from the substrate surface. Again this value is an overestimate. However, for $\sigma \approx 0.1\,\%$, dislocation densities of 10^9 –10^{10} cm^{-2} have been observed in

regions within a few micrometers of the substrate surface. For further general discussion of epitaxy, see [170 – 172].

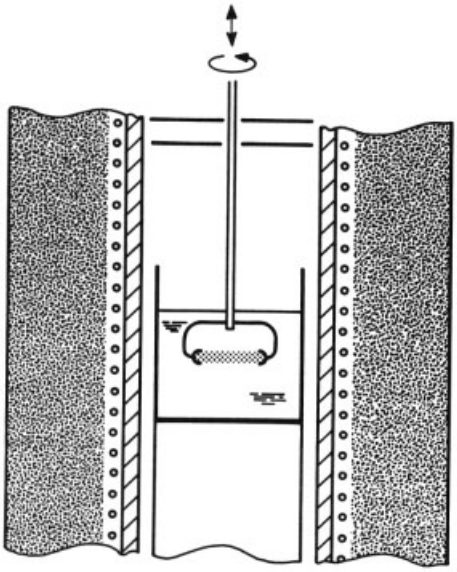

Figure 41. Liquid-phase epitaxy (LPE) by dipping
To produce growth, either the solution is supersaturated before the substrate is dipped, or it is marginally undersaturated before dipping and cooled while the substrate is immersed

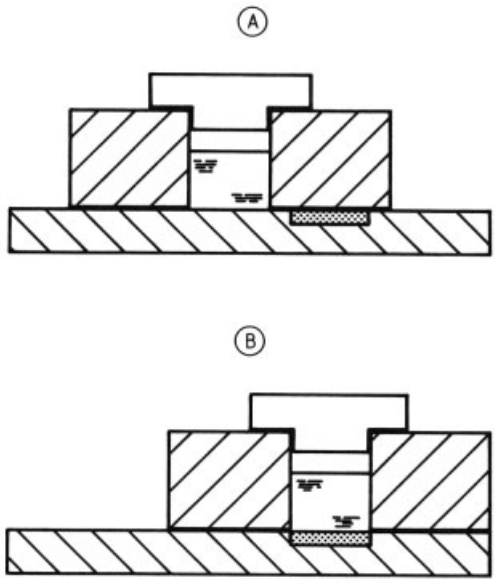

Figure 42. Liquid-phase epitaxy (LPE) in a sliding boat
Growth is produced by either supersaturating the solution before it is moved onto the substrate or cooling while it is over the substrate

A wide variety of LPE methods have been used in research [57]. However, only two are employed on a large scale: the *dipping method* (see Fig. 41 for a representative example) and the *sliding boat process* (see example in Fig. 42). Dipping processes are used mostly for oxidic materials, and sliding boat processes for semiconductors.

Now as before, LPE is the most widley used thin-film deposition technique for large-scale production of light emitting diodes based on III-V semiconductors. The annual world production of LEDs amounts to 500×10^9.

Further information about LPE is given in references [54, 57], and [170 – 172]. Detailed reviews on fundamentals and technology of LPE are given in [173, 174].

7.5. Electrolytic Methods

Electroplating is a familiar process, and by using electrically conducting single-crystal substrates, single-crystal layers can be grown [175]. Bulk crystals can also be grown, but in general the process is rather slow. Thin layers of semiconductors have been grown from metallic solutions by this method; there is some dispute as to whether the principal effect is electroplating or whether growth results from the Peltier cooling which occurs at the semiconductor–metallic liquid interface. Both processes probably occur, and by suitable arrangement of the system, either can be made dominant. Layers can also be grown from nonmetallic solvents [176]. The limitation to conducting materials on conducting seeds, together with the fact that the process offers no particular advantages and some technological problems (e.g., the need to have both conducting and insulating materials in contact with the solution), has restricted the use of growth by electroplating, and there is no sign of its commercial use.

However, electroplating is only one of many electrolytic processes, and the use of electrolysis to change the valences of dissolved ions may be more important than direct electrolytic transport. The processes used include

1) Decomposition of the solvent to yield a supersaturated solution of the product, which is composed of atoms that were originally part of the solvent, e.g., MoO_2 crystals grown by

electrochemical decomposition of $K_2Mo_2O_7$ at 600 °C

2) Decomposition of the solvent to reduce the solubility of a solute, e.g., the growth of $Co_x V_{3-x} O_4$ from solution in $Na_2WO_4 + WO_3$

For further information about these processes which can yield a wide variety of oxides, borides, carbides, phosphides, sulfides, and arsenides, see [177]. None of the processes seems to have been exploited commercially.

7.6. Hydrothermal Growth

Temperature gradient transfer under hydrothermal conditions is used on a large scale for the growth of quartz. Hydrothermal conditions are taken to be temperatures and pressures near the critical values (i.e., 374.2 °C and 221.23 bar) (Fig. 6). Thus, hydrothermal growth could be regarded as intermediate between solution growth and vapor growth. In practice, the solvent is almost invariably a solution of a mineralizer in water; the mineralizer and the solute decrease the saturation vapor pressure and increase the critical temperature. In general, temperatures are less than 400 °C and pressures exceed 250 bar so that the solution is usually a liquid. The mineralizers are usually alkaline (NaOH or, less often, Na_2CO_3 for quartz), but neutral (e.g., alkali metal halides) or acidic (e.g., HI, HBr) mineralizers are also used.

Figure 43 is a schematic sectional view of an autoclave system which exaggerates the horizontal dimensions. Typical autoclaves for pressures over 1000 bar have ratios of height to internal diameter in the range of 16–20, and heights can be 8 m or more. For lower pressures, larger diameters (up to 1 m) and smaller heights (2.5 m) are used. The optimum ratio of external to internal diameter is 1.8–2.5. Thicker walls do not give appreciably higher working pressures.

The autoclave is loaded by putting the seed crystals in suitable racks, loading granular nutrient into the basket, and filling about 75 % of the free volume with the mineralizer solution. The concentration of this solution is usually about 1 mol/L. The autoclave is then sealed and heated to its working temperatures, e.g., 380 °C below the baffle and 350 °C above the baffle. The growth rate rises with the upper temperature, the temperature difference, the degree of fill (which governs the working temperature), the mineralizer concentration, and the amount of opening in the baffle (typically 2–10 %). For most applications of quartz, the growth temperature should be high, the mineralizer concentration should be less than 1 mol/L, and the other conditions are adjusted to limit the growth rate to less than 1 mm/d. Departing from these conditions gives crystals with too large a hydrogen content or too much other impurity to be useful in critical applications. According to the specifications to be met [178], the hydrogen content should not exceed about 200 atomic ppm (based on silicon); at higher hydrogen content, quartz is rather brittle. In current systems, up to 500 kg of quartz is grown in one run. Typical individual crystals weigh from a few hundred grams to several kilograms. For further data, see [48, Chap. 8], [55, Chap. 7.3] and [56, Chaps. 6 and 7]. A review is given in [179].

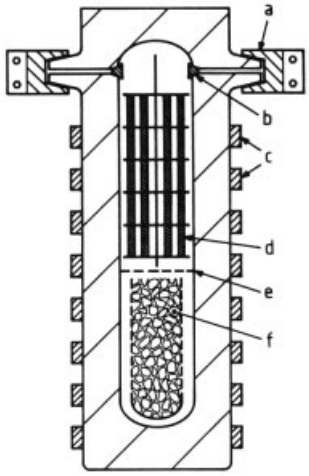

Figure 43. Hydrothermal growth
a) Split clamp; b) Seal ring; c) Heating elements; d) Crystals; e) Baffle; f) Granular nutrient in porous basket

8. Vapor Growth Techniques

8.1. Sublimation and Evaporation Methods

Growth by evaporation and subsequent condensation is used widely to make thin epitaxial layers [134] and bulk crystals of a few materials

(mostly compounds of cadmium or zinc with sulfur, selenium, or tellurium.)

For an element (e.g., zinc), two methods are available. The material can be evaporated in a vacuum and condensed on some region which is on the line of sight from the source, or it can be evaporated into a gas stream which can be conveyed around corners provided it is not cooled until the deposition region is reached. With compounds, the choice of methods is larger. Some compounds evaporate congruently; these can be treated like the elements. However, most compounds do not evaporate congruently. For these, multiple sources (one for each component) can be used. Thus, vacuum evaporation or evaporation with one gas stream per component into gas streams that are combined in the deposition region can be employed. However, only one source can be used for vacuum evaporation if a small volume of the material is vaporized at one time. This process, called *flash evaporation,* can be achieved in several ways. For example, grains of the compound are dropped onto a plate heated to a very high temperature, or a bulk solid is exposed either to a laser beam [180], which is scanned over the solid surface and run in a pulsed mode (*laser-assisted evaporation*), or to an energetic gas plasma (*sputtering*) [181]. Other possibilities also exist: for example, one component can be evaporated into an inert gas stream which in the deposition region reacts with a gas:

$$\mathrm{Cd\,(g) + H_2S\,(g) \rightarrow CdS + H_2} \tag{78}$$

where the CdS is first formed as a supersaturated vapor which then deposits, usually as plates of the solid. For further information, see [58].

Bulk growth from the vapor has become important for two optoelectronic materials: ZnSe and SiC. Both compounds are of special interest for blue and blue-green light emitting diodes and lasers. Moreover, SiC is a suitable substrate for GaN epitaxy as well as for high-temperature high-power microelectronics. ZnSe monocrystals with diameters up to 6 cm and lengths of 3 cm are grown by seeded physical vapor transport [182]. SiC crystals are produced by the sublimation–condensation method in modified Lely reactors made of graphite, as shown in Figure 44 [183].

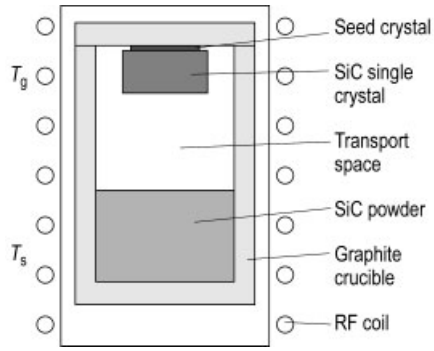

Figure 44. Schematic of a Lely reactor for the vapor growth of SiC single crystals (T_s = sublimation temperature, T_g = growth temperature)

Temperatures of $\geq 2200\ ^\circ$C are required for evaporation. The seed wafer is mounted to directly face the source of SiC powder. Usually, induction heating with frequency of 10 kHz or about 1 MHz and an argon atmosphere with a pressure of 130–5300 Pa are employed during growth at rates of 0.2–2 mm/h. Boules with diameters of 3–5 cm and a length of 2 cm are already produced commercially. Diameters 7.5 cm and 10 cm are under investigation. The main efforts are still directed towards the control of characteristic polytypes, lowering of dislocation density, and preventation of micropipes for growth in the <0001>direction. Such holes, with diameter of several micrometers, are caused by the strain field arround screw dislocations combined with the large value of the Burgers vector in this material. More information is given in [183].

8.2. Molecular Beam Epitaxy

Molecular beam epitaxy (MBE) is a refined evaporation process. Conventional evaporation sources have a rate of evaporation governed by the area of the evaporant and its temperature (or the partial pressure of the evaporant in the source region). Most MBE sources have large evaporant areas but are almost sealed. The rate of efflux from the source is governed by the area of the opening in the source and the partial pressure of the evaporant inside the source. Outside the source, the atmosphere is almost invariably an ultrahigh vacuum (usually less than 10^{-12} bar). Figure 45 shows the working region of a system and a typical source. To minimize the time

to reach the desired vacuum, the substrates and sources are often loaded and unloaded through a vacuum lock so that the main vacuum system is not routinely exposed to the atmosphere. The extreme vacuum used enables MBE films to be grown with very low substrate temperatures (< 100 °C), but in general, somewhat higher temperatures are needed (e.g., 200 °C) to obtain the desired properties. In the MBE growth of binary compounds, the flux of one component usually governs the growth rate, provided the other is available in excess; the excess of noncritical component is not consumed [184].

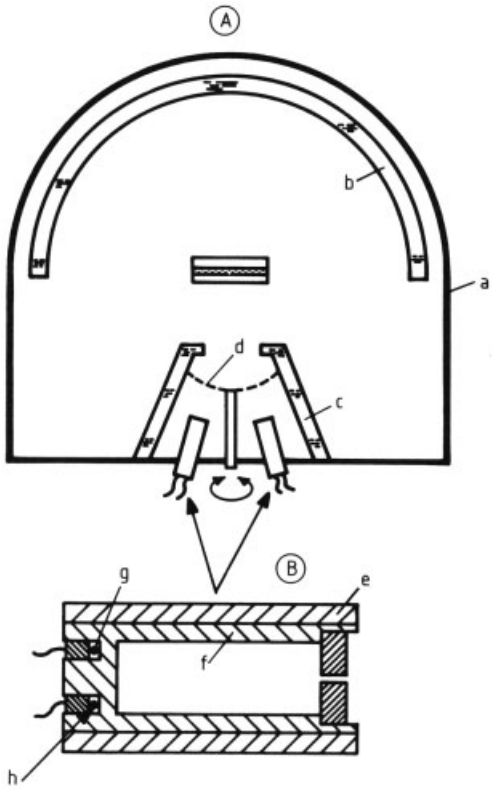

Figure 45. Simple system for molecular beam epitaxy (A) and typical source (B)
a) Outer vacuum jacket; b) Cryogenic screens; c) Source holders; d) Shutters; d) Cylindrical heater; f) Crucible; g) Thermocouples; h) Heater

Molecular beam epitaxy is a powerful method for making multiple-layer structures, such as multiple sandwich structures containing alternate layers of GaAs and AlAs, only two or three atomic layers thick. Such superlattices or quantum well materials have properties that differ radically from either bulk crystal [185]. Very sharp composition changes with dopants can also be obtained. However, MBE is usually a very slow process. Growth rates rarely exceed 1 μm/h so that the method is only viable for very thin layers. Molecular beam epitaxy is used in the production of semiconductor devices to a rapidly increasing extent. For further information, see [172] and [136, 184 – 186].

8.3. Chemical Vapor Transport

The bulk polycrystalline silicon, the nutrient material used in the Czochralski and floating zone processes, is prepared by decomposing volatile silicon compounds on hot silicon filaments by reactions of the type

$$2\,SiHCl_3 + 2\,H_2 \rightarrow 2\,Si + 6\,HCl$$

Such reactions can be used to deposit single crystal films onto single crystal substrates. This type of process is usually described as vapor-phase epitaxy (VPE). Because the reactants are produced outside the growth system, VPE is not strictly identical to chemical vapor deposition (CVD) in the sense described [187, 188], which involves in situ reactions such as

$$CdS\,(s) + I_2\,(g) \rightleftharpoons CdI_2\,(g) + 1/2\,S_2\,(g) \qquad (79)$$

However, all depositions involving chemical reactions are now described as CVD which is taken to include also mixed forms, for example,

$$AsCl_3 + Ga \rightarrow GaCl_3 + 1/2\,As_2 \qquad (80)$$

$$GaCl_3 + 1/2\,As_2 + 3/2\,H_2 \rightarrow GaAs + 3\,HCl \qquad (81)$$

(the reactions are schematic only). The $AsCl_3$ is provided from an external source.

Figures 46 – 48 give examples of these three types of growth. The first and last of these processes are widely used, and in some cases, there is an advantage in using a reduced pressure which increases the diffusion coefficients to give faster growth. For the growth of semiconducting compounds, use of the appropriate organometallic compounds can be a significant advantage [189].

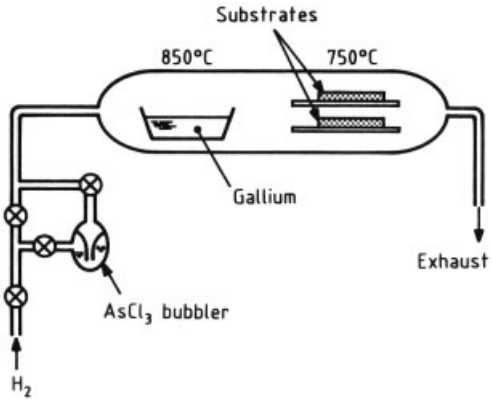

Figure 46. Vapor-phase epitaxy of gallium arsenide

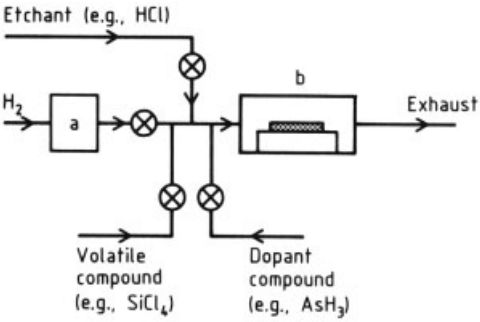

Figure 47. Vapor-phase epitaxy of silicon
a) Purifier; b) Reactor (1100–1200 °C)
Commercial reactors may contain up to 100 substrates; in these large reactors, the susceptor (often heated by radiofrequency) is tilted so that a greater gas flow rate (thinner boundary layer) compensates for the decrease of reactant concentration in the gas after it has passed over many substrates

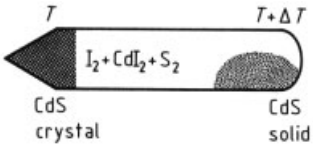

Figure 48. Transport of cadmium sulfide by means of iodine

The use of these materials allows good control over growth so that complex structures composed of very thin layers can be grown. These structures have electronic properties not exhibited by bulk samples [190, 191]. Table 6 gives some examples of the layers that can be grown and the reactants used. For further information on VPE and CVD, see [55, 56, 58, 171, 172, 192 – 198]. Organometallic vapor-phase epitaxy is summarized in [199]

8.4. Vapor-Liquid-Solid Growth

In 1964 [37] the vapour-liquid-solid (VLS) mechanism was proposed for the first time to use such phase transition sequence for the growth of whisker crystals. A review is given by Givargizov, *Rost nitevidnykh i plastinchatykh kristallov iz para.*, Nauka, Moscow 1977. Today, such mechanism is of increasing importance for controllable production of nanowires and nanometer-sized structures–a rapidly expanding area with numerous applications, e.g., in electronics, photonics, biology, and medicine [200]. In the VLS growth mechanism, the reactants are supplied in the vapor phase by MBE, MOCVD, for example, and the diameters of the crystal whiskers are dictated by the size of metallic seed particles acting in molten form as localized solvent on a substrate. The material to be grown translates first from the vapor into the droplet due to the much higher absorption rate at the vapor–liquid surface compared with that at the surrounding solid substrate surface. The crystallization starts at the melt–solid interface between the droplet and substrate. The driving force for crystallization is the supersaturation within the droplet, which is continuously maintained by the material supply diffusing from the vapor–liquid interface to the liquid–solid one. As a result, a crystalline wire is growing under the continuously raising microdroplet (Fig. 49).

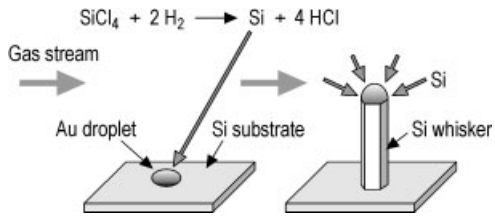

Figure 49. The VLS mechanisms demonstrating the growth of a silicon whisker crystal by using a gold microdroplet as solvent during a CVD or MOCVD process

Table 6. Organometallic chemical vapor deposition

Layer	Substrates	Reactants	Growth temperature, °C
AlAs	GaAs, Al_2O_3	$Al(CH_3)_3 + AsH_3$	700
AlN	Si, Al_2O_3, α-SiC	$Al(CH_3)_3 + NH_3$	1250
CdS	Al_2O_3	$Cd(CH_3)_2 + H_2S$	475
CdSe	Al_2O_3	$Cd(CH_3)_2 + H_2Se$	600
CdTe	Al_2O_3, BeO, GaAs	$Cd(CH_3)_2 + Te(CH_3)_2$	500
GaAs	GaAs, Ge, Al_2O_3	$Ga(CH_3)_3$ or $Ga(C_2H_5)_3 + AsH_3$	650–700
$GaAs_xP_{1-x}$	GaAs, $MgAl_2O_4$	$Ga(CH_3)_3 + PH_3 + AsH_3$	700–725
$Ga_xIn_{1-x}As_yP_{1-y}$	InP	$Ga(CH_3)_3 + In(CH_3)_3 + AsH_3 + PH_3$	670
GaN	Al_2O_3, α-SiC	$Ga(CH_3)_3 + NH_3$	950
GaP	GaAs, Si	$Ga(CH_3)_3 + PH_3$ or $P(C_2H_5)_3$	700 or 500
ZnS	Al_2O_3, BeO	$Zn(C_2H_5)_2 + H_2S$	750
ZnSe	Al_2O_3, $MgAl_2O_4$	$Zn(C_2H_5)_2 + H_2Te$	725
ZnTe	Al_2O_3	$Zn(C_2H_5)_2 + Te(CH_3)_2$	500

9. Solid-Phase Growth Techniques

9.1. Strain–Anneal Method

Before methods for handling materials without contamination were developed, growth from the solid state had very significant advantages. Contamination from surfaces in contact with other solids or with reactive atmospheres penetrated only a small distance because diffusion in the solid state is usually several orders of magnitude slower than in a liquid phase. Also, the apparatus was simple: almost any furnace would serve. At that time, the great interest was in metals or other ductile materials that could safely be strained. Later interest centered on rather brittle materials and, usually, on crystals with great crystallographic perfection. Thus, solid-phase growth techniques which tend to produce heavily dislocated crystals are of interest only when rather imperfect crystals of ductile materials are needed, such as materials with low loss and high permeability for very efficient transformers and motors.

In the strain-anneal method, the sample is first heated to a temperature high enough to ensure that it is not brittle (about half the absolute melting temperature for metals) and then subjected to a strain, usually between 1 and 10 %. The strain employed is critical. If σ_{opt} is the optimum strain, then strains less than $0.9\,\sigma_{\mathrm{opt}}$ or larger than $1.5\,\sigma_{\mathrm{opt}}$ will be less effective by a factor of two or more. The second stage of the process is annealing, which may be carried out to advantage with a temperature gradient across the sample. The annealing temperature is usually at an absolute temperature T which is an appreciable fraction (50–95 %) of the melting temperature on the same scale. If growth is going to occur, it is usually fast: 5–50 mm/h can be expected, but much higher rates (about 1 m/s) have been recorded [201]. The free energy to drive the process is the excess energy attributable to dislocations or small clusters of dislocations and grain boundaries. Ordered arrays of dislocations (small-angle grain boundaries) have lower energies per dislocation than isolated dislocations so that many samples grown by the strain-anneal method contain low-angle boundaries. For further data, see [4, 55, 202].

9.2. Methods Not Involving Applied Strain

In theory, the excess free energy of grain boundaries should be a driving force for crystallization. Although the energy per dislocation in a grain boundary is less than that for isolated dislocations, reducing the area of the boundaries (i.e., increasing the grain size) should decrease the sample free energy. This process occurs in many materials: metal workers exploit it on a large scale to control the properties of their materials. The extend to which it occurs is often related to purity. In particular, precipitated impurities stop grain boundary movements, and a common method for obtaining hard or creep-free metals is to add impurities that migrate to the grain boundaries. Thus, if impurities can be removed, grain growth can be expected. Solid-state electrolysis removes impurities and also heats the samples. To be effective, solid-state electrolysis should involve large current densities (several amperes

or even hundreds of amperes per square centimeter). Subsidiary heating can be applied to raise the sample temperature to a suitable value (55–90 % of the absolute melting temperature is usual). Growth rates can be 3–50 mm/h. For further information, see [203] and [204].

9.3. Solid-Phase Epitaxy

A solid in an amorphous (glassy) state has a higher free energy than the same solid in the form of a single crystal. The recrystallization of silicate glass over the centuries is a familiar process. This process is slow but can be accelerated by heating. It is now possible to create amorphous samples of many materials such as metallic alloy glasses by rapid quenching, and amorphous semiconductors by sputtering, for example, in hydrogen. For semiconductors, selective recrystallization of amorphous layers is now practiced on an increasing scale by the use of layer heating or other energy beam techniques [205 – 213]. Spatial control gives this process an obvious advantage. A less obvious but still significant advantage is that layers with much higher doping densities can be made by this method than by any other. Indeed, dopant concentrations can exceed the maximum equilibrium value.

9.4. Gel Growth

Gels can be regarded as soft pseudosolids. They are actually three-dimensional solid skeletons with liquid-filled interstices. The solid phase is often a network of silicon atoms surrounded by oxygen atoms in a tetrahedral geometry, which is made by hydrolysis of sodium silicate at about pH 8. Other gel-forming materials include gelatin, pectin, or potassium salts of fatty acids. Crystallization in a gel has the advantage of occurring in a strain-free environment, which is useful for some fragile solids that can deform under their own weight. Growth is usually produced by diffusing two solutions (one from each side of the gel) so that a supersaturated solution is created within the gel. With care or by providing a seed, one crystal can be grown. The diffusion rate in the gel is only slightly lower than in the pure fluid phase so the gel process is relatively rapid. The size L of a crystal generally increases with the square root of the time. This gives rapid growth when L is small but rather slow growth as L increases. Typically, L^2 increases at rates measurable in units of square millimeters per day or per week.

Examples of pairs of diffusants are:

1) Aqueous $(NH_4)_2CO_3$ and aqueous $CaCl_2$ to yield $CaCO_3$ crystals
2) CuCl in hydrochloric acid and pure water to yield CuCl crystals
3) Aqueous $AgNO_3$ and sulfuric acid to yield Ag_2SO_4 crystals
4) Pb $(CH_3COO)_2$ in acetic acid and aqueous KI to yield PbI_2 crystals

The technique does not appear to be used commercially. For further information, see [60].

10. Trends

Many methods are obviously available for crystal growth. The extent to which various methods are now used is discussed in Section 1.4, which shows that only few methods are employed to any significant degree. As the range of materials and uses increases, other methods will be employed more often; for example, the use of thin-film techniques is increasing. Table 7 gives production rates and their trends for the more extensively produced materials. Two other trends are also obvious. The first is that for economic reasons, larger crystals will be demanded. Costs of growth and subsequent processing decrease as crystal sizes increase. The second trend is the increasing demand for higher quality because greater purity and perfection, as well as tighter control over dopant distribution, are needed to raise yields and device performance. Decreases in device dimensions and increases in yields may change the quantities of particular materials needed, and some substitution of noncrystalline materials can be expected, e.g., amorphous semiconductors. However, an increasing demand for crystals and for crystals grown to tighter specifications can confidently be expected. To meet these demands at reasonable cost will require better apparatus with more extensive use of automation and better understanding of the mechanisms of crystal growth. Thus, although fewer people may be involved with crystal production, more effort is likely

to be required in process development and research.

Table 7. Annual production rates and trends

Crystal [a]	Production [b] >, t/a	Trends [c]
Silicon	10 000	up
Metals	4000	up [d]
Quartz	2500	constant
III-V compounds [e]	800	constant
Ruby	500	down
Alkali halides	400	down
Germanium	400	constant
Garnets	100	up
Lithium niobate	100	up
Phosphates [f]	20	down
Lithium tantalate	15	constant
Cubic zirconia	2000	up
TGS [g]	10	constant
Diamond	10	up
II-VI compounds [h]	5	up

[a] Crystals of more than 100 other materials can be bought.
[b] Averages; data for Russia, China, etc., are rare so that errors could range from −20 % to + 30 %.
[c] Trends are averaged; some materials show rapid fluctuations, e.g., quartz.
[d] The upward trend for metal crystals is very rapid.
[e] GaAs, GaP, InP, InSb, etc.
[f] Ammonium and potassium dihydrogen phosphates and their analogs.
[g] Triglycine sulfate and similar materials used mainly for pyroelectric devices.
[h] CdS, CdTe, etc.

11. References

1. A. van Hook: *Crystallization*, Reinhold Publ. Co., New York 1961.
2. J. G. Burke: *Origin of the Science of Crystals*, University of California Press, Berkeley 1966.
3. H. Scheel in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **1a**, North Holland Publ., Amsterdam 1993, p. 1.
4. K.-Th. Wilke, J. Bohm: *Kristallzüchtung*, VEB Deutscher Verlag der Wissenschaften, Berlin 1988 und Verlag H. Deutsch, Thun, Frankfurt/M. 1988.
5. A. Verneuil, *Compt. Rend.* **135** (1902) 791.
6. A. Verneuil, *Ann. Chim. Phys.* **3** (1904) 20.
7. G. Spezia, *Accad. Sci. Torino Atti a. Sci. Fis. Mat. Nat.* **40** (1905) 254.
8. G. Spezia, *Accad. Sci. Torino Atti a. Sci. Fis. Mat. Nat.* **41** (1905) 51.
9. G. Spezia, *Accad. Sci. Torino Atti a. Sci. Fis. Mat. Nat.* **44** (1908) 95.
10. A. Sauveur, *Proc. Int. Assoc. Test. Mater. 6th Congress* **11** (1912) Section 1.
11. J. Pintsch, *Grosses Jahrb. Radiotechn. und Elektrotechnik* **15** (1918) 270.
12. J. Czochralski, *Z. Phys. Chem. (Leipzig)* **92** (1918) 184.
13. R. Nacken, *Zs. Instrumentenkunde* **36** (1916) 12.
14. S. Kyropoulos, *Zs. anorg. allg. Chemie* **154** (1926) 308.
15. F. Koref, *Z. Elektrochem.* **28** (1922) 511.
16. G. Tammann: *Lehrbuch der Metallographie*, Leopold Voss, Leipzig 1914.
17. P. W. Bridgman, *Proc. Am. Acad. Arts Sci.* **60** (1925) 303.
18. F. Stober, *Z. Kristallogr.* **61** (1925) 299.
19. D. C. Stockbarger, *Rev. Sci. Instrum.* **7** (1936) 133.
20. P. Kapitza, *Proc. R. Soc. London Ser. A* **119 A** (1928) 358.
21. W. Nernst, *Z. Phys. Chem. (Leipzig)* **47** (1904) 52.
22. M. Volmer, *Z. Phys. Chem. (Leipzig)* **102** (1922) 267.
23. W. Kossel, *Nachr. Ges. Wiss. Goettingen Jahresker. Math. Phys. Klasse* 1927, 135.
24. I. N. Stranski, *Z. Phys. Chem. (Leipzig)* **136** (1928) 259.
25. *Discuss. Faraday Soc.* **5** (1949) .
26. H. E. Buckley: *Crystal Growth*, J. Wiley, New York 1951.
27. G. K. Teal, J. B. Little, *Phys. Rev.* **25** (1950) 16.
28. G. K. Teal, E. Buehler, *Phys. Rev.* **87** (1952) 190.
29. W. G. Pfann, *Trans. AIME* **194** (1952) 747.
30. H. C. Theurer, US 3 060 123, 1952.
31. P. H. Keck, M. J. E. Golay, *Phys. Rev.* **89** (1953) 1297.
32. W. G. Pfann: *Zone Melting*, J. Wiley, New York 1958.
33. J. A. Burton, R. C. Prim, W. P. Slichter, *J. Chem. Phys.* **21** (1953) 1987.
34. J. A. Burton, E. D. Kolb, W. P. Slichter, J. D. Struthers, *J. Chem. Phys.* **21** (1953) 1991.
35. Advanced Semiconductors Buyers' Guide 2003/2004, Special issue of III-Vs Review, Elsevier, 2005.
36. M.-P. Pileni (ed.): *Nanocrystals Forming Mesoscopic Structures*, Wiley-VCH, Weinheim 2005.
37. R. S. Wagner, W. C. Ellis, *Appl. Phys. Letters*, **4** (1964) 89.
38. T. Fukuda, P. Rudolph, S. Uda (eds.): *Fiber Crystal Growth from the Melt*, Springer 2004.
39. H. Nelson, *RCA Review* **24** (1963) 603.
40. B. A. Joyce, *J. Cryst. Growth* **3/4** (1968) 43.
41. H. M. Manasevit, W. I. Simpson, *J. Electrochem. Soc.* **116** (1969) 1725.

42. A. Y. Cho, J. R. Arthur, *Progr. Solid State Chem.* **10** (1975) no. 3, 157.
43. H. Hirayama, H. Asahi in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3a**, North Holland Publ., Amsterdam 1994, p. 184.
44. B. V. Spitsyn in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3a**, North Holland Publ., Amsterdam 1994, p. 401.
45. A. Holden, P. Singer: *Crystals and Crystal Growing*, Heinemann, London 1961.
I. Tarjan, M. Matrei: *Laboratory Manual on Crystal Growth*, Akademiai Kiado, Budapest 1972.
46. P. Kratochvil: *Crystals*, Iliffe, London 1967.
47. B. R. Pamplin: *Crystal Growth*, Pergamon Press, Oxford 1980.
48. J. C. Brice: *Crystal Growth Processes*, Blackie, Glasgow 1986.
49. A. A. Chernov: *Modern Crystallography III*, Springer Verlag, Berlin 1984.
50. W. A. Tiller: *The Science of Crystallization*, Cambridge University Press, Cambridge, 1991 and 1995.
51. I. V. Markov: *Crystal Growth for Beginners*, World Scientific, Singapore 1995.
52. A. van Hook: *Crystallization*, Reinhold Publ. Co., New York 1961.
53. R. F. Strickland-Constable: *Kinetics and Mechanisms of Crystallization*, Academic Press, New York 1968.
54. J. C. Brice: *The Growth of Crystals from Liquids*, North Holland, Amsterdam 1973.
55. R. A. Laudise: *The Growth of Single Crystals*, Prentice-Hall, Englewood Cliffs 1970.
56. P. Hartman: *Crystal Growth*, An Introduction, North Holland, Amsterdam 1973.
57. D. Elwell, H. J. Scheel: *Crystal Growth from High-Temperature Solutions*, Academic Press, New York 1975.
58. M. M. Faktor, I. Garrett: *Growth of Crystals from the Vapour*, Chapman and Hall, London 1974.
59. F. Rosenberger: *Fundamentals of Crystal Growth I*, Springer Verlag, Berlin 1979.
60. H. K. Henisch: *Crystal Growth in Gels*, Pennsylvania State University Press, University Park 1970.
61. B. Wunderlich: *Macromolecular Physics,* vol. 2, Academic Press, New York 1976.
62. H. L. Anderson: *Physics Vademecum*, American Institute of Physics, New York 1981, pp. 54 and 315.
63. K. Fujiwara et al., *J. Crystal Growth* **243** (2002) 275.
64. A. A. Chernov in E. Kaldis (ed.): *Crystal Growth of Electronic Materials*, Elsevier, Amsterdam 1985, p. 87.
65. J. C. Brice, *J. Cryst. Growth* **28** (1975) 249.
J. C. Brice, J. M. Robertson, W. T. Stacy, J. C. Verplanke, *J. Cryst. Growth* **30** (1975) 66.
66. V. M. Glazov, S. N. Chizevskaya, N. N. Glagoleva: *Liquid Semiconductors*, Plenum Press, New York 1969.
67. C. H. P. Lupis: *Chemical Thermodynamics of Materials*, North Holland Publ., New York 1983.
68. A. D. Pelton in P. Haasen (ed.), *Mat. Sci. and Technol.*, vol. **5**, VCH Verlagsgesellschaft, Weinheim, Germany 1991.
69. W. Hume-Rothery, J. W. Christian, W. B. Pearson: *Metallurgical Equilibrium Diagrams*, Chapman and Hall, London 1952.
70. P. Gordon: *Principles of Phase Diagrams in Materials Systems*, McGraw-Hill, New York 1968.
71. F. A. Kroger: *Chemistry of Imperfect Crystals*, North Holland, Amsterdam 1973.
72. A. M. Alper: *Phase Diagrams: Materials Science and Technology*, Academic Press, London 1970.
73. A. Prince: *The Constitutional Diagrams of Alloys*, Institute of Metals, London 1956.
74. M. Hansen: *Constitution of Binary Alloys*, McGraw-Hill, New York 1958.
75. R. P. Elliott: *Constitution of Binary Alloys*, First Supplement, McGraw-Hill, New York 1965.
76. F. A. Shunk: *Constitution of Binary Alloys*, Second Supplement, McGraw-Hill, New York 1969.
77. E. A. Brandes: *Smithells Metals Reference Book*, Butterworth, London 1983.
78. G. V. Samsonov: *The Oxide Handbook*, Plenum Publishing, New York 1973.
79. E. H. Levin, C. R. Robbins, H. F. McMurdie: *Phase Diagrams for Ceramists*, American Ceramic Society, Columbus 1964 and 1969 Supplement.
80. E. H. Levin, H. F. McMurdie: *Phase Diagrams for Ceramists*, 1975 Supplement, American Ceramic Society, Columbus 1975.
81. R. H. Doremus: *Glass Science*, J. Wiley, New York 1973.
82. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds*, Pergamon Press, Oxford 1963.
83. F. D. Rossini: *Selected Values of Chemical Thermodynamic Properties*, NBS, Washington 1952.

84. JANAF: *Thermochemical Tables*, NSRDS-NBS, Washington 1971.
85. D. E. Gray: *American Institute of Physics Handbook*, McGraw-Hill, New York 1972.
86. W. G. Moffatt: *Handbook of Binary Phase Diagrams*, General Electric Company, Schenectady, continuing.
87. D. P. Woodruff, *The Solid-Liquid Interface*, Cambridge University Press, 1973.
88. J. H. Greenberg, *J. Cryst. Growth* **161** (1996) 1.
89. P. Rudolph, *Prog. Cryst. Growth Charact.* **29** (1994) 275.
90. W. Tolksdorf in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 563.
91. H. H. Farrell, J. P. Harbison, L. D. Peterson, *J. Vac. Sci. Technol.* **B5** (1987) 1482.
92. P. Rudolph in R. Fornari, C. Paorici (eds.): *Theoretical and Technological Aspects of Crystal Growth*, Trans Tech Publ., Zürich 1998, p. 1.
93. A. Zunger, *MRS Bulletin* **22** (1997) 20; G. B. Stringfellow, *MRS Bulletin* **22** (1997) 27.
94. K. A. Jackson: *Liquid Metals and Solidification*, Amer. Soc. Metals, Novelty, Ohio, 1958, p. 174.
95. D. E. Temkin in N. N. Sirota (ed.): *Mechanism and Kinetics of Crystallization*, Nauka i Technika, Minsk 1964, p. 68.
96. P. Bennema in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, North Holland, Amsterdam 1993, p. 477.
97. S. Glasstone, K. J. Laidler, H. Eyring: *The Theory of Rate Processes*, McGraw-Hill, New York 1941.
98. W. B. Hillig in W. Kingery (ed.): *Kinetics of High Temperature Processes*, J. Wiley, New York 1959, p. 627.
99. J. C. Brice, *J. Cryst. Growth* **1** (1967) 218.
100. D. Walton, *J. Chem. Phys.* **37** (1962) 2182.
101. D. Walton, *Philos. Mag.* **7** (1962) 1691.
102. W. K. Burton, N. Cabrera, F. C. Frank, *Philos. Trans. R. Soc. London Ser. A* **243**, (1951) 299.
103. A. A. Chernov, *Sov. Phys. Usp. (Engl. Transl.)* **4** (1961) 776.
104. P. Bennema, G. H. Gilmer in *Crystal Growth*, An Introduction, North Holland, Amsterdam 1973 , Chap. 10.
105. H. J. Scheel, *J. Crystal Growth* **287** (2006) 214.
106. K. T. Zawilski, M. Claudia C. Custodio, R. C. DeMattei, R. S. Feigelson, *J. Crystal Growth*, **282** (2005) 236.
107. G. N. Kozhemyakin, *J. Crystal Growth* **257** (2003) 237.
108. E. W. Sparrow, J. L. Gregg, *J. Heat Transfer* **81** (1959) 249.
109. G. Müller, A. Ostrogorsky in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2b**, North Holland, Amsterdam 1994, p. 709.
110. F. Dupret, N. van den Bogaert in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2b**, North Holland, Amsterdam 1994, p. 875.
111. J. J. Derby et al. in R. Fornari, C. Paorici (eds.): *Theoretical and Technological Aspects of Crystal Growth*, Trans Tech Publ., Zürich 1998, p. 119.
112. J. J. Derby, A. Yeckel in G. Müller, M J.-J. Metois and P. Rudolph (eds.): *Crystal Growth — from Fundamentals to Technology*, Elsevier, Amsterdam 2004, p. 143.
113. K. Kakimoto in G. Müller, M J.-J. Metois and P. Rudolph (eds.): *Crystal Growth — from Fundamentals to Technology*, Elsevier, Amsterdam 2004, p. 169.
114. H. J. Scheel, *J. Cryst. Growth* **13/14** (1972) 560.
115. W. C. Johnston, W. A. Tiller, *Trans. AIME* **221** (1961), 331, **224** (1962) 214.
116. O. Pätzold, I. Grants, U. Wunderwald, K. Jenker, A.- Cröll, G. Gerbeth, *J. Crystal Growth* **245** (2002) 237.
117. A. Krauze et al., *J. Crystal Growth* **265** (2004) 14.
118. T. P. Lyubimova, A. Croell, P. Dold, O. A. Khlybov, I. S. Fayzrakhmanova, *J. Crystal Growth* **266** (2004) 404.
119. P. Schwesig, M. Hainke, J. Friedrich, G. Mueller, *J. Crystal Growth* **266** (2004) 224.
120. S. Yesilyurt, S. Motakef, R. Grugel, K. Mazuruk, *J. Crystal Growth* **263** (2004) 80.
121. A. Mitric et al., *J. Crystal Growth* **287** (2006) 224.
122. G. Müller in R. Fornaci, C. Paorici (eds.): *Theoretical and Technological Aspects of Crystal Growth*, Trans Tech Publ., Zürich 1998, p. 87.
123. M. Abricka, J. Krümins, Yu. Gelfgat, *J. Cryst. Growth* **180** (1997) 388.
124. A. Cröll, P. Dold, Th. Kaiser, F. R. Szofran, K. W. Benz in H. R. Huff, H. Tsuyga, U. Gösele (eds.): *Silicon Materials Science and Technology*, **Proc. vol. 98–1**, The Electrochem. Soc. Inc., Pennington NJ 1998, p. 429.
125. D. T. J. Hurle, R. W. Series in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 259.

126. J. Fainberg, H.-J. Leister, G. Müller, *J. Cryst. Growth* **180** (1997) 517.
127. V. L. Indenbom, *Kristall und Technik* **14** (1979) 493.
128. J. Völkl in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2b**, North Holland, Amsterdam 1994, p. 821.
129. S. Pendurti, V. Prasad, H. Zhang, *Modelling Simul. Mater. Sci. Eng.* **13** (2005) 249 and 267.
130. W. A. Tiller, K. A. Jackson, J. W. Rutter, B. Chalmers, *Acta Metallurg.* **1** (1953) 428.
131. J. R. Carruthers in N. Hannay (ed.): *Treatise on Solid State Chemistry*, vol. **5**, Plenum Press, New York 1975, p. 325.
132. A. A. Chernov, *J. Cryst. Growth* **24/25** (1974) 11.
133. M. Althaus, K. Sonnenberg, E. Küssel, R. Naeven, *J. Cryst. Growth* **166** (1996) 566.
134. R. V. Stuart: *Vacuum Technology*, Thin Films and Sputtering, Academic Press, London 1983.
135. E. Monberg in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 51.
136. C. T. Foxon in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3a**, North Holland, Amsterdam 1994, p. 93.
137. B. K. Tanner: *X-ray Diffraction Topography*, Pergamon Press, Oxford 1976.
138. D. Betteridge, H. E. Hallam: *Modern Analytical Methods*, The Chemical Society, London 1974.
139. E. J. Millett, *J. Cryst. Growth* **48** (1980) 666.
140. P. Rudolph, M. Neubert, S. Arulkumaran, M. Seifert, *Cryst. Res. Technol.* **32** (1997) 35.
141. D. T. J. Hurle, B. Cockayne in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 99.
142. P. Rudolph, T. Fukuda, *Cryst. Res. Technol.* **34** (1999) 3.
143. P. Rudolph: *Profilzüchtung von Eiskristallen*, Akademie-Verlag, Berlin 1982.
144. V. A. Tatarchenko in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2b**, North Holland, Amsterdam 1994, p. 1011.
145. X. Liu, *III-Vs Review* **8** (1995) 14.
146. M. Tatsumi, K. Fujita in M. Isshiki (ed.): *Recent Development of Bulk Crystal Growth*, Research Signpost, Trivandrum 1998, p. 47.
147. J. B. Mullin in K. A. Jackson (ed.): *Compound Semiconductor Devices. Structures and Processing*, Wiley-VCH, Weinheim, Germany 1998, p. 1.
148. P. Rudolph, M. Jurisch, *J. Cryst. Growth* **198/199** (1999) 325.
149. M. Jurisch et al., *J. Crystal Growth* **275** (2005) 283.
150. P. Rudolph, F. Kiesling, *Crystal Res. Technol.* **23** (1988) 1207.
151. M. J. Reiner, K. E. Taylor, EP-A 092 496, 1983.
152. D. N. Duhl, M. L. Gell, GB 2 112 812, 1982.
153. GB 2 112 309, 1982 (G. J. S. Higginbotham, J. R. Marjoram, F. J. Horrocks).
154. P. Rudolph in M. Isshiki (ed.): *Recent Development of Bulk Crystal Growth*, Research Signpost, Trivandrum 1998, p. 127; *Prog. Crystal Growth and Charact.* **29** (1994) 275.
155. F. Schmid, C. P. Khattak, D. M. Felt, *Am. Ceramic Soc. Bull.* **73** (1994) 39.
156. A. Rabenau, *Chem. Ing. Tech.* **36** (1964) 542.
157. C. T. Butler, B. J. Sturm, R. B. Quincy, *J. Cryst. Growth* **8** (1971) 187.
158. M. O'Donoghue: *Identifying Man-Made Gems*, NAG Press, London 1983.
159. W. G. Pfann: *Zone Melting*, J. Wiley, New York 1958.
160. W. Keller, A. Muhlbauer: *Floating Zone Silicon*, Dekker, New York 1981.
161. J. Bohm, A. Lüdge, W. Schröder in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 213.
162. H. Riemann et al., *J. Electrochem. Soc.* **142** (1995) 1007.
163. A. Mühlbauer, A. Muiznieks, J. Virbulis, A. Lüdge, H. Riemann, *J. Crystal Growth* **151** (1995) 66.
164. R. S. Feigelson, *J. Crystal Growth* **79** (1986) 669.
165. A. McPherson in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 417.
166. F. Rosenberger in R. Fornari, C. Paorici (eds.): *Theoretical and Technological Aspects of Crystal Growth*, Trans Tech Publ., Zürich 1998, p. 241.
167. H. J. Scheel in R. Fornari, C. Paorici (eds.): *Theoretical and Technological Aspects of Crystal Growth*, Trans Tech Publ., Zürich 1998, p. 201.
168. S. Porowski, *J. Crystal Growth* **166** (1996) 583.
169. R. Triboulet, *Progress Cryst. Growth and Charact.* **28** (1994) 85.
170. G. B. Stingfellow, *Rep. Prog. Phys.* **45** (1982) 469.
171. B. A. Joyce, *Rep. Prog. Phys.* **37** (1974) 363.
172. J. W. Matthews: *Epitaxial Growth*, **Parts A and B**, Academic Press, New York 1975.

173. M. B. Small, E. A. Giess, R. Ghez in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3a**, North Holland, Amsterdam 1994, p. 223.
174. E. Bauser in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3b**, North Holland, Amsterdam 1994, p. 879.
175. J. G. Wright in J. W. Matthews: *Epitaxial Growth*, Academic Press, New York 1975, Chap. 2.3.
176. D. Elwell, in E. Kaldis, H. J. Scheel (eds.): *Current Topics in Materials Science*, vol. **2**, North Holland, Amsterdam 1977, Chap. 11.6.
177. W. Kunnmann in R. A. Lefever (ed.): *Preparation and Properties of Solid State Materials*, vol. **1**, Dekker, New York 1971, Chap. 1.
178. J. C. Brice, *Rep. Prog. Phys.* **40** (1977) 567.
179. K. Byrappa in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **2a**, North Holland, Amsterdam 1994, p. 465.
180. J. T. Cheung, T. Magee, *J. Vac. Sci. Technol. A* **1** (1983) 1604.
181. S. A. Barnett, G. Bajor, J. E. Greene, *Appl. Phys. Lett.* **37** (1980) 734.
182. G. Cantwell et al., *J. Appl. Phys.* **71** (1992) 2931.
183. N. Ohtani et al. in M. Isshiki (ed.): *Recent Development of Bulk Crystal Growth*, Research Signpost, Trivandrum 1998, p. 231.
184. C. T. Foxon, B. A. Joyce in E. Kaldis (ed.): *Current Topics in Materials Science*, vol. **7**, North Holland, Amsterdam 1981, Chap. 1.
185. A. C. Gossard: *Treatise on Materials Science and Technology*, vol. **24**, Academic Press, New York 1982.
186. B. A. Joyce, *Phys. Educ.* **18** (1981) 328.
187. H. Schäfer in *Crystal Growth*, An Introduction, North Holland, Amsterdam 1973 , p. 143.
188. H. Schäfer: *Chemical Transport Reactions*, Academic Press, New York 1964.
189. G. B. Stringfellow, *J. Cryst. Growth* **62** (1983) 225.
190. J. C. Portal et al., *Semicond. Sci. Technol.* **1** (1986) 3.
191. M. S. Skolnick et al., *Semicond. Sci. Technol.* **1** (1986) 29.
192. J. Bloem, W. A. P. Claassen, *Philips Tech. Rev.* **41** (1983) 60.
193. J. Bloem, Y. S. Oei, H. H. de Moor, J. H. L. Hanssen, L. J. Giling, *J. Cryst. Growth* **65** (1983) 399.
194. L. J. Giling, *J. Electrochem. Soc.* **129** (1982) 634.
195. J. H. L. Hanssen, A. A. Saaman, H. H. de Moor, L. J. Giling, J. Bloem, *J. Cryst. Growth* **65** (1983) 406.
196. G. R. Srinivasan in D. G. Gupta (ed.): "Silicon Processing," *ASTM STP 804*, ASTM, New York 1983, p. 151.
197. G. B. Stringfellow, *J. Cryst. Growth* **62** (1983) 225.
198. G. B. Stringfellow, *Rep. Prog. Phys.* **45** (1982) 469.
199. D. W. Kisker, T. F. Kuech in D. T. J. Hurle (ed.): *Handbook of Crystal Growth*, vol. **3a**, North Holland, Amsterdam 1994, p. 93.
200. A. I. Persson, M. W. Larsson, S. Stenström, B. J. Ohlsson, L. Samuelson, L. R. Wallenberg, *Nature Materials* **3** (2004) 677.
201. A. Mineo, A. Matsuda, T. Kurosu, M. Kikuchi, *Solid State Commun.* **13** (1973) 329.
202. K. T. Aust, *J. Cryst. Growth* **13/14** (1972) 57.
203. R. C. Jordan, D. W. Jones, V. H. Hems, *J. Less. Common Met.* **42** (1975) 101.
204. C. M. Muirhead, D. W. Jones, *J. Less Common Met.* **50** (1976) 73.
205. J. S. Williams et al., *Appl. Phys. Lett.* **33** (1978) 542.
206. D. H. Auston, J. A. Golovchenko, P. R. Smith, C. M. Surko, *Appl. Phys. Lett.* **33** (1978) 539.
207. E. F. Kennedy, L. Csepregi, J. W. Mayer, *J. Appl. Phys.* **48** (1977) 4241.
208. L. N. Aleksandrov: *Growth of Crystalline*, Semiconductor Materials on Crystal Surfaces, Elsevier, Amsterdam 1983.
209. L. N. Aleksandrov, *Prog. Cryst. Growth Charact.* **9** (1984) 227.
210. J. C. Fan, N. M. Johnson: *Energy Beam Solid Interactions and Transient Thermal Processing*, North Holland, Amsterdam 1984.
211. J. M. Poate, G. Foti, D. C. Jacobson: *Surface Modification and Alloying*, Plenum, New York 1983.
212. F. A. Smidt: *Ion Implantation for Materials Processing*, Noyes Data Corp., Park Ridge, N.J., 1983.
213. C. W. White, S. P. Pearcy: *Laser Beam and Electron Beam Processing of Materials*, Academic Press, New York 1980.

Cyanamides

THOMAS GÜTHNER, Degussa AG, Trostberg, Germany
BERND MERTSCHENK, Degussa AG, Trostberg, Germany (Chap. 5)

1. Calcium Cyanamide 1
1.1. Physical Properties 2
1.2. Chemical Properties 2
1.3. Production 2
1.3.1. Overall Process 2
1.3.2. Manufacture 3
1.3.3. Processing of Technical Calcium Cyanamide 5
1.4. Quality Specifications 5
1.5. Storage and Transportation 5
1.6. Uses 5
1.7. Legal Aspects 6
2. Cyanamide 6
2.1. Physical Properties 6
2.2. Chemical Properties 7
2.3. Production 10
2.4. Analysis 11
2.5. Storage and Transportation 11
2.6. Uses 11
2.7. Legal Aspects 12
3. Dicyandiamide 12
3.1. Physical Properties 13
3.2. Chemical Properties 13
3.3. Production 16
3.4. Quality Specifications and Analysis 17
3.5. Storage and Transportation 17
3.6. Uses 17
3.7. Legal Aspects 19
4. Substituted Cyanamides 19
5. Toxicology and Occupational Health 19
6. References 20

Calcium cyanamide is produced from calcium carbide and nitrogen gas by an exothermic solid nitrogenation at 1000°C. It is used as a slow release nitrogen fertilizer with beneficial side effects against soilborne deseases. In chemical industry, it is the base for cyanamide and thiourea chemistry.Cyanamide is a fairly stable, toxic substance obtained from calcium cyanamide by reaction with water and carbon dioxide. It has vast applications in industrial nitrogen organic chemistry, and is also directly used as a plant protection agent.Dicyandiamide is derived from cyanamide by dimerization in presence of alkali. The nontoxic compound has found widespread industrial applications in resins, plastics, water treatment, and pharmaceuticals.

1. Calcium Cyanamide

Calcium cyanamide was probably first obtained in the laboratory in 1877 by heating calcium carbamate to red heat [3]. In 1889 it was prepared in larger quantities by heating thoroughly mixed, finely pulverized urea and calcium oxide [4]. A commercial process for the nitrogenation of calcium carbide, the Frank – Caro process, was patented in Germany in 1895 [5]. A calcium cyanamide plant using a batch oven furnace was erected at Piano d'Orta, Italy, in 1905. At about the same time, the Polzeniusz – Krauss channel furnace was put into service. By 1910, plants were established in Germany (Bayerische Kalkstickstoffwerke; AG für Stickstoffdünger), France, Japan, Sweden, Switzerland, and the United States (American Cyanamid).

Calcium cyanamide [*156-62-7*], CaNC≡N, M_r 80.10, is the neutral salt of cyanamide, NH_2CN (see Chap. 2); it is also known as lime nitrogen, nitrolime, and kalkstickstoff.

Industrial-grade calcium cyanamide contains, in addition to $CaCN_2$, ca. 20 % CaO and 10 – 12 % free carbon, which gives the product its gray-black color. It also contains a small amount of nitrides formed from silica and alumina. The total nitrogen content varies from 22 to 25 %, depending on the raw materials used. Of the total nitrogen, 92 – 95 % is present as cyanamide and 0.1 – 0.4 % as dicyandiamide; the remainder is present as nitrides.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_139.pub2

1.1. Physical Properties [6]

Pure calcium cyanamide is a colorless hygroscopic salt, which forms rhombohedral crystals. In a nitrogen atmosphere, it melts with decomposition at 1300 °C. Other physical properties are given below.

Density at 25 °C	2.36 g/cm^3
Heat of fusion	54 kJ/kg
Specific heat, 20 – 100 °C	909 J/kg
Enthalpy of formation ΔH_{298}	− 348 kJ/mol
Free energy of formation ΔF_{298}	− 303 kJ/mol
Normal entropy S	87.1 J mol^{-1} K^{-1}

1.2. Chemical Properties [6]

Calcium cyanamide decomposes above 1000 °C in a way that depends on the temperature, the partial pressure of nitrogen, and impurities present. The principal decomposition products formed by heating under vacuum or in an inert gas are calcium carbide, calcium metal, and nitrogen:

$$2\,CaCN_2 \rightarrow CaC_2 + Ca + 2\,N_2$$

Decomposition at higher nitrogen pressure gives primarily cyanide products. Calcium nitride and carbon are always formed.

Calcium cyanamide reacts with oxygen and carbon dioxide, starting at ca. 475 °C, with formation of nitrogen and calcium carbonate; calcium oxide forms above 850 – 900 °C. Carbon present as an impurity cannot be removed by oxidation, which attacks calcium cyanamide preferentially.

Carbon monoxide and calcium cyanamide react above 1000 °C to form CaO and calcium carbide.

The reactions of calcium cyanamide in aqueous solution are determined primarily by temperature and pH [7]. At room temperature, monocalcium cyanamide $Ca(HCN_2)_2$ forms; when heated at pH 9 – 10, this is converted to calcium hydroxide and dicyandiamide $(NH_2)_2CNCN$ (see Chap. 3). Below 40 °C, cyanamide is obtained at pH 6 – 8; lime is precipitated with carbon dioxide.

Urea is formed in the presence of acid and catalysts, and thiourea in the presence of sulfides.

At 200 °C under pressure, calcium cyanamide is hydrolyzed to ammonia and calcium carbonate in an alkaline medium. The reaction was employed to produce ammonia in the early1900s.

1.3. Production

1.3.1. Overall Process

Calcium cyanamide is manufactured in three steps. First, lime is made by heating high-grade limestone:

$$CaCO_3 \rightarrow CaO + CO_2$$

Second, calcium carbide is synthesized from lime and coke or coal in an electric furnace:

$$3\,C + CaO \rightarrow CaC_2 + CO$$

Electric resistance heating is used to melt the lime and effect the endothermic reaction with the coke (→ Calcium Carbide).

Third, calcium cyanamide is synthesized from calcium carbide and nitrogen. This reaction is exothermic but requires heating of a portion of the reaction mixture to the initiation temperature of 900 – 1000 °C:

$$CaC_2 + N_2 \rightarrow CaCN_2 + C$$

The heat source is then removed [7]. The reaction continues by controlled addition of nitrogen; it produces 286.6 kJ/mol at 1100 °C, and 295 kJ/mol at 0 °C.

Mechanism. It is believed that calcium cyanamide is formed through a number of intermediates, such as $Ca(CN)_2$, CaC_2N_2, CaC, and Ca_2N_2 [8 – 10]. However, in the nitrogenation of calcium carbide, the main products are calcium cyanamide and carbon. Above 1000 °C, cyanide also forms and is in equilibrium with cyanamide and carbon:

$$CaCN_2 + C \rightleftharpoons Ca(CN)_2 \quad \Delta H_{298} = 163\,\mathrm{kJ/mol}$$

This reaction is endothermic, whereas the nitrogenation is exothermic.

Above 1160 °C, the system $CaCN_2$ – C – $Ca(CN)_2$ melts, with over 60 % cyanamide present at equilibrium.

Small amounts of cyanide present at the usual reaction temperatures of 1000 – 1100 °C are

rapidly converted to calcium cyanamide by slow cooling; the cooled product is practically free of cyanide.

The cyanamide – cyanide equilibrium is shifted towards cyanide by addition of alkali metal halides. The yield of cyanide can be improved by quick cooling. This was the basis of the fusion cyanide process to produce crude sodium cyanide for metal leaching. A recent investigation showed, that a calcium cyanamide conversion efficiency of up to 95 % can be reached. Most cyanide used for ore leaching however is made from hydrocyanic acid. [182]

Nitrogenation Byproducts. Technical-grade calcium carbide contains impurities that, during nitrogenation, lower the yield of desired product. This is because silica and alumina, which are only partially reduced in the carbide furnace, are present in the carbide raw materials. In the nitrogenation, Al_2O_3 and SiO_2 react with $CaCN_2$ to form nitrides.

$$3\,CaCN_2 + Al_2O_3 \rightarrow 3\,CaO + 2\,AlN + 3\,C + 2\,N_2$$

$$3\,CaCN_2 + SiO_2 \rightarrow 2\,CaO + CaSiN_2 + 3\,C + 2\,N_2$$

Industrial-grade calcium carbide also contains calcium hydroxide and calcium carbonate. On heating, water and carbon dioxide are liberated; reaction with carbide gives acetylene. As a result, hydrogen, formed by decomposition of acetylene, is always found in the exhaust gas of the nitrogenation furnaces.

These side reactions and the inadvertent introduction of some oxygen during the nitrogenation reduce the overall yield by ca. 10 %. The purest possible raw materials give the most favorable results economically.

Catalysts. Various catalysts that act as fluxes are used to accelerate the reaction or lower the required temperature. The most commonly used are calcium chloride and calcium fluoride. Their function is not completely clear, but they may provide a liquid phase in which the reaction can occur. The product is in the form of a well-sintered pig, indicating the presence of a liquid phase at some stage of the reaction.

Rate studies have shown that calcium fluoride reduces the temperature of optimum reactivity and increases the reaction rate by a factor of 4.5 at 1000 °C [11].

Reaction Kinetics. The conversion rate depends on temperature, nitrogen partial pressure, carbide purity, and the presence of additives. The crystallite size also plays a part. High-grade carbide consisting of coarse crystals reacts more slowly than lower-grade carbide of lamellar structure. As indicated above, metal halides accelerate the reaction.

In the nitrogenation of calcium carbide, the reaction proceeds inward from the grain surface. The rate of nitrogen transport through the porous layer is a determining factor at lower temperature, whereas the chemical reaction at the boundary layer is decisive at higher temperature and in the presence of additives. The dependence of the rate constant on temperature, additives, and nitrogen pressure has been reported [12]. Figure 1 shows the effect of nitrogen pressure.

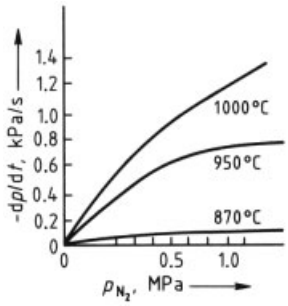

Figure 1. Rate of calcium carbide nitrogenation as a function of nitrogen pressure [12]
Powdered CaC_2 (67 %) containing 1.2 % CaF_2 used.

1.3.2. Manufacture

Both batch and continuous processes have been used to prepare calcium cyanamide [13]. The more important processes are the Frank – Caro batch furnace process and the Trostberg rotary furnace process. In these processes, the exothermal reaction between carbide and nitrogen takes place between 1000 and 1150 °C. In the continuous rotary furnace process, the reactants, except at the beginning, are heated exclusively by the heat of reaction, whereas the batch furnace process requires ignition for each batch.

Frank – Caro Batch Oven Process [11]. The batch oven process was widely used in the first half of the 20th century. Today, it has much less importance due to unfavorable nitrogenation yields, higher labor costs and lower throughput. It employs a batch reactor filled with up to 10 t of ground calcium carbide. After initial ignition by electric heating or a pyrotechnic mixture, the reaction proceeds spontaneously. Nitrogen gas is introduced through inlets in the furnace shell. The total reaction lasts several days and leads to a sintered block of calcium cyanamide, which is broken into pieces and milled.

Trostberg Rotary Furnace Process. In this continuous nitrogenation process, ground lime nitrogen (crude calcium cyanamide) is metered onto a material bed in a broadened part of the rotary furnace in such a way as to keep the carbide concentration in the bed as low as possible. The process is controlled by varying the carbide content.

The rotary furnace developed by SKW Trostberg (now Degussa AG) is ca. 20 m long and includes the broadened furnace head, where the main reaction takes place [14]. The furnace is lined with fireclay.

Ground carbide with a CaC_2 content of 55 – 60 % is continuously mixed with calcium fluoride and ground lime nitrogen and is blown into the furnace with nitrogen. The average residence time of the solid is between 5 and 6 h. The reaction product is a granular or powdered lime nitrogen, which is transferred from the rotary furnace to a cooling drum. The heat released by the reaction is sufficient to bring the starting materials to the desired temperature of 1000 – 1100 °C. After the furnace has been started, it remains in operation without external heating for many months. The capacity of a unit is about 25 t/d of fixed nitrogen (Fig. 2).

Other Processes. Because large amounts of energy are consumed, particularly in the preparation of calcium carbide, less energy intensive methods for manufacturing calcium cyanamide have been examined. Between 600 and ca. 1000 °C many reactions of lime with nitrogen compounds containing hydrogen or carbon lead to calcium cyanamide. Compounds such as hydrocyanic acid, dicyanogen, urea, and dicyandiamide form calcium cyanamide with lime.

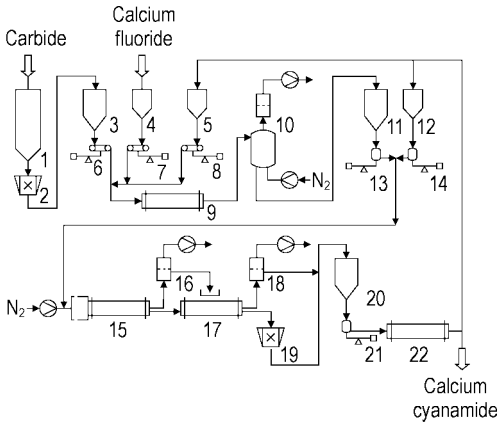

Figure 2. Trostberg rotary furnace process
1) Calcium carbide lump silo; 2) Crusher; 3) Crushed carbide silo; 4) Calcium fluoride silo; 5) Calcium cyanamide silo; 6)- 8) Weigh feeders; 9) Tube mill; 10) Homogenizing; 11) Milled carbide silo; 12) Calcium cyanamide silo; 13), 14) Weigh feeders; 15) Rotary furnace; 16), 18) Dust filters; 17) Rotary cooling tube; 19) Crusher; 20) Silo; 21) Weigh feeder; 22) Tube mill

Lime and urea form calcium cyanate [15], which on heating is converted to calcium cyanurate and finally to calcium cyanamide:

$$CaO + 2\,H_2NCONH_2 \xrightarrow{200\,°C} Ca(OCN)_2 + H_2O + 2\,NH_3$$

$$3\,Ca(OCN)_2 \xrightarrow{400\,°C} 2$$

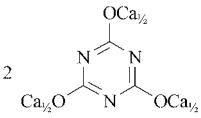

$$\xrightarrow{700\,°C} 3\,CaCN_2 + 3\,CO_2$$

Lime reacts with hydrogen cyanide [16] at 750 – 850 °C:

$$CaO + 2\,HCN \rightarrow CaCN_2 + CO + H_2$$

The reaction of lime with ammonia and carbon monoxide at 700 – 900 °C produces 99 % calcium cyanamide:

$$CaO + 2\,NH_3 + 3\,CO \rightarrow CaCN_2 + 2\,CO_2 + 3\,H_2$$

The above processes give white calcium cyanamide, whereas the product obtained from limestone and coal always contains carbonaceous impurities. None of these processes have yet been commercialized because of cost or poor yield [13].

1.3.3. Processing of Technical Calcium Cyanamide

Crude calcium cyanamide is reduced in size and ground in tube mills to allow passage through a 0.2-mm screen. If calcium cyanamide is sold in granular form, grinding is omitted and the desired grain size is separated by screening.

When the carbide content is above 0.1 %, as in the Frank – Caro process, calcium cyanamide is degassed by treatment with water to convert the carbide to acetylene and calcium hydroxide. For environmental and safety reasons, the acetylene-containing off-gas must be incinerated and scrubbed to remove acidic byproducts.

Granulation. Because of the dust problems associated with finely ground calcium cyanamide in fertilizer applications, the product is oiled. Alternatively, it is converted to a more compact form by granulation or compression. Granulation with calcium nitrate solutions gives a fine, finished calcium cyanamide bead. This form affords a nitrate nitrogen system that can be used directly in the plant as starter nitrogen. In addition, it affords a slowly acting cyanamide nitrogen source. This product has an almost unlimited storage life, because the free calcium oxide is completely hydrated and no expansion can occur. In Europe, Degussa AG is the main producer of beaded calcium cyanamide.

1.4. Quality Specifications

Commercial-grade calcium cyanamide has approximately the following composition:

Total nitrogen	19 – 20 %
Cyanamide nitrogen	ca. 15 %
Nitrate nitrogen	ca. 2 %
Dicyandiamide nitrogen	0.4 – 0.7 %
Other nitrogen	2.0 – 2.5 %
Total calcium oxide	53 – 55 %
Water (chemically bound)	7 – 8 %

Total nitrogen is determined by the Kjeldahl or Dumas method.

1.5. Storage and Transportation

Calcium cyanamide is stored in warehouses or silos, but storage life with free calcium oxide is limited because the latter is expanded by moisture.

Bags used for ground and granular calcium cyanamide must be moisture-tight to prevent grain breakdown and expansion. Polyethylene valve bags in combination with multilayer paper bags are excellent for granular calcium cyanamide. Paper and plastic bags are used for the less sensitive beaded product.

Calcium cyanamide with a calcium carbide content below 0.1 % is not classified as dangerous goods according to transport regulations. If the calcium carbide content exceeds 0.1 %, calcium cyanamide has to be classified as UN class 4.3 (substances that emit flammable gases upon contact with water), UN no. 1403, packaging group III, hazard label 4.3 according to actual UN regulations on transport of dangerous goods.

1.6. Uses

Calcium cyanamide is used in agriculture and horticulture predominantly in Europe and Asia as a slow-release nitrogen fertilizer which exhibits beneficial side effects against soilborne plant diseases, slugs, and germinating weeds. The fertilizer is broken down by soil moisture into highly reactive lime and free cyanamide. The latter is responsible for the fertilizer's fungicidal and herbicidal side effects. Within a few days, soil microbes convert the cyanamide to urea and then to ammonia. A second known pathway leads to dicyandiamide, which acts as a nitrification inhibitor in the soil. In this way, the conversion of ammonia nitrogen into nitrate is slowed down for several weeks, and leaching of fertilizer nitrogen is prevented (see also → Fertilizers). Calcium cyanamide is particularly valuable for intensively used soils that are highly infested with soilborne pathogens which cause root and stem rot [17, 18].

Recent developments in urea fertilizers use a calcium cyanamide core within urea granules to stabilize nitrogen against nitrification [19, 20].

Calcium cyanamide is also used to control animal pests on pasture. When applied to grassland, it kills the dwarf water snail, which is the intermediate host of the liver fluke. It also kills eggs and larvae of gastric and intestinal parasites in the grass and controls salmonella in liquid sewage.

Over 99 % of nitrogen oxides are effectively removed from waste gases by scrubbing with calcium cyanamide [21]. In Portland cement, cyanamide and calcium cyanamide improve setting characteristics, increase compressive strength, and reduce freeze – thaw damage [22].

Pharmaceutical-grade calcium cyanamide is used to treat alcoholism. A small pill taken once a day subjects the alcohol abuser to an unpleasant flush (see Chap. 5), which discourages drinking [23]. Calcium cyanamide is used for steel nitridation [24] and, to some extent, for desulfurization [25]. It has various uses in the chemical industry for the production of cyanamide, dicyandiamide, and thiourea.

1.7. Legal Aspects

Calcium cyanamide is listed in global inventories of existing chemicals, e.g., EINECS (European Union), TSCA Inventory (United States), DSL (Canada), AICS (Australia), IECSC (China) and MITI (Japan). Within the European Union, cyanamide is classified as harmful (danger symbol X_n), R41(risk of serious damage to eyes), S22 (do not breathe dust) and S36/37/39 (wear suitable protective clothing, gloves and eye/face protection). In Germany, calcium cyanamide is classified in water hazard class 2 (hazardous for waters).

2. Cyanamide

Cyanamide [*420-04-2*], $NH_2C{\equiv}N$, M_r 42.04, was first obtained in 1838 by passing gaseous ammonia over cyanogen chloride [26]. Another pathway was the desulfurization of thiourea by means of lead or mercury compounds [27]. Just after the discovery of the calcium cyanamide process, the laboratory synthesis of cyanamide became more convenient [28, 29]. However, the latent instability of cyanamide prevented its industrial-scale production and usage until processes were developed around 1965 to purify and stabilize cyanamide in aqueous solution and in crystalline form.

Recently, cyanamide has been found in nature. Hairy vetch (*Vicia villosa*) produces free cyanamide from inorganic nitrate, obviously as an allelochemical to compete against neighboring plants [30, 31]. Cyanamide is regarded as a key intermediate within chemical evolution on earth billions of years ago [32]. It is therefore not astonishing that cyanamide metabolism is common in nature.

2.1. Physical Properties [6]

Cyanamide forms colorless, deliquescent orthorhombic crystals or long transparent needles (from water) [33], large platelike crystals when allowed to crystallize spontaneously from water, and crystals from dimethyl phthalate solution [34]. The crystal structure of solid cyanamide has been thoroughly investigated [35]. Physical properties are listed below.

Melting point	46 °C
Vapor pressure at 20 °C	5×10^{-3} mbar
Crystal density	1.282 g/cm^3
Melt density at 57 °C	1.069 g/cm^3
Refractive index n_D^{48}	1.4418
Enthalpy of formation at 25 °C	+ 61.3 kJ/mol
Specific heat (0 – 39 °C)	2.291 J/g·K
Latent heat of fusion	8.76 kJ/mol
Latent heat of vaporization	70.5 kJ/mol
Heat of solution (in water)	+ 15.1 kJ/mol
Enthalpy of dimerization to dicyandiamide at 25 °C	− 48.8 kJ/mol
Enthalpy of hydration to turea at 25 °C	− 109 kJ/mol
Dissociation constant pK_a	10.3

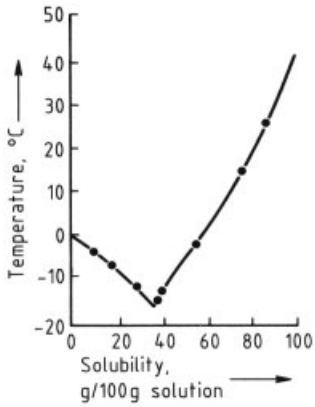

Figure 3. Solubility of cyanamide in water

Cyanamide is a weak acid that is very highly soluble in water. It is completely soluble at 43 °C, and has a minimum solubility at − 15 °C (Fig. 3). It is highly soluble in polar organic solvents, such as alcohols, ethers, esters, nitriles

and amides, but less soluble in nonpolar solvents like saturated, halogenated, or aromatic hydrocarbons. Ketones are not recommended as solvents for cyanamide, because they tend to react under certain circumstances.

The IR spectrum of cyanamide exhibits an intense, poorly resolved doublet at 2225 – 2260 cm^{-1}. The UV spectrum shows only weak absorption below 230 nm, with no maximum observable above 208 nm. The infrared and Raman spectra support the *N*-cyanoamine structure, $NH_2-C\equiv N$ [22].

2.2. Chemical Properties

Cyanamide can be regarded as the amide of cyanic acid or as the nitrile of carbamic acid. The difunctional molecule is capable of reacting as a nucleophile at the amino group or as an electrophile at the cyano group. Many reactions with polyfunctional compounds involve both reactive sites, thus forming heterocyclic structures.

Cyanamide exists in two tautomeric forms that cannot be isolated in pure form:

$$\underset{\text{Nitrile form}}{H_2N-C\equiv N} \rightleftarrows \underset{\text{Carbodiimide form}}{HN=C=NH}$$

Spectroscopic evidence and reactivity clearly show that the cyanamide form predominates by far. However, in some special reactions, e.g., silylation and stannylation, cyanamide reacts solely as the carbodiimide form.

Oligomerization. One of the most important reactions of cyanamide is *dimerization* to cyanoguanidine (dicyandiamide). Unless kept cool and stabilized by weak acids, crystalline cyanamide may polymerize violently. In aqueous solutions, polymerization becomes significant above pH 5, with a maximum reactivity between pH 8 and 10

Melamine [*108-78-1*] was produced until around 1975 from dicyandiamide at elevated temperatures and pressure. Liquid ammonia, with or without an organic solvent, is a suitable reaction medium [36]. The mechanism of *trimerization* has not been established, but it is believed to involve either the reaction of a molecule of cyanamide with a molecule of dicyandiamide or the reaction of two molecules of dicyandiamide to form guanylmelamine, which breaks down into melamine and cyanamide (→ Melamine and Guanamines).

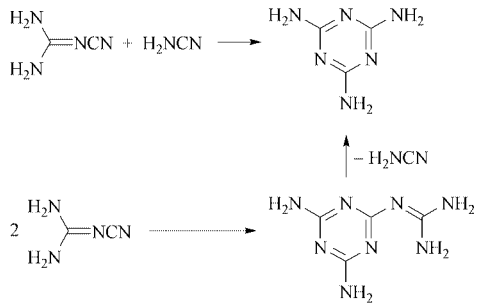

Reaction with Water. The hydrolysis of cyanamide to urea is catalyzed by acids. In aqueous solution cyanamide has the highest stability between pH 4.0 and 4.5. At lower pH, especially in the presence of strong mineral acids, the hydration reaction may become violent and can lead to vapor explosions if the heat is not dissipated [37].

Reaction with Hydrogen Sulfide or Hydrogen Selenide. Thiourea [*62-56-6*] can be prepared by the action of hydrogen sulfide on an aqueous solution of cyanamide with ammonia as catalyst. The optimum conditions are pH 8 – 9 and a temperature above 50 °C [38].

$$H_2NCN + H_2S \rightarrow H_2NCSNH_2$$

Selenourea [*630-10-4*] is obtained similarly from hydrogen selenide [39].

$$H_2NCN + H_2Se \rightarrow H_2NCSeNH_2$$

Reaction with Thiols and Thiophenols. Cyanamide reacts with thiols and thiophenols more readily than with alcohols and phenols; the products are pseudothioureas. The reaction is conveniently conducted in diethyl ether.

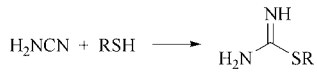

Reaction with Alcohols and Phenols. Under suitable conditions, cyanamide reacts with alcohols and phenols to give *O*-alkyl- and *O*-arylisoureas (also called pseudoureas). Alkylisoureas are usually prepared as hydrochlorides by treating a solution of cyanamide in the appropriate anhydrous alcohol with dry hydrogen chloride; sulfuric acid has also been recommended [40].

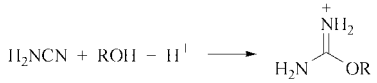

The alkylisourea bases, which have basic strengths K_b of around 1×10^{-4} [41], can be liberated from their salts by treatment with moist potassium hydroxide under ether or with sodium alkoxides in anhydrous alcohols. They are fairly stable under normal conditions but are sensitive to moisture and dissociate when heated.

O-Alkylisourea compounds exhibit a significant alkylating power. The hydrohalide or carboxylate salts decompose to urea and alkyl halides or alkyl carboxylates upon heating to temperatures above 100°C. Because of their potential mutagenic properties, these compounds should be handled with care.

The synthetic potential of *O*-alkylisourea compounds is based on its ease of reaction with simple amines to form guanidines, thus acting as an activated form of cyanamide in a guanylation (or amidination) reaction. In addition, *O*-alkylisoureas are the ideal reagents to build up 2-alkoxypyrimidines or 2-alkoxy-1,3–5-triazines.

O-Methylisourea can be nitrated to *O*-methyl-*N*-nitroisourea, useful as a nitroguanylation agent [42].

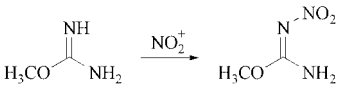

Reaction with Ammonia and Amines. Ammonia adds readily to the nitrile group of cyanamide to form guanidine. Weakly acidic conditions help to avoid dimerization and, in aqueous systems, hydrolysis to urea.

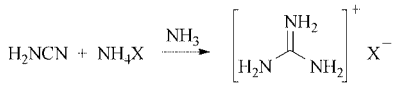

where X = halogen.

Primary and secondary amines react with cyanamide to form substituted guanidines under the above conditions.

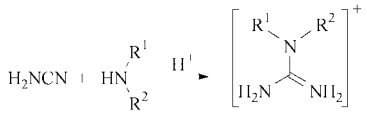

Reaction with Amino Acids. Cyanamide adds to amino acids to form carboxyalkylguanidines. The conversion of glycine to glycocyamine was the first such reaction reported [43].

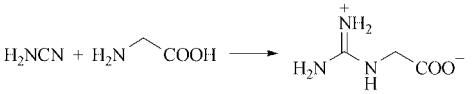

Sarcosine reacts with cyanamide to form creatine [44, 45], which can be further processed to creatinine at higher temperatures [46].

Cyanamide and dialkylcyanamides can be used as coupling agents to induce the formation of peptide bonds when heated with *N*-benzyloxycarbonyl amino acids and amino acid esters [47]. This dehydrating effect is ascribed to the intermediate formation of an *O*-acylpseudourea or amino anhydride, which reacts further with the amino ester [48].

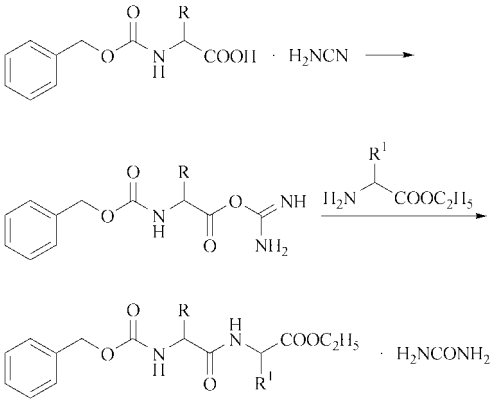

Reaction with Hydrazines and Hydrazones. Cyanamide reacts with hydrazines to form aminoguanidines, which are versatile intermediates for the synthesis of nitrogen-rich compounds, particularly heterocycles. Aminoguanidine [*79-17-4*], itself unstable, is conveniently isolated as the relatively insoluble hydrogen carbonate after reaction of aqueous hydrazine with cyanamide at around pH 6 [49, 50].

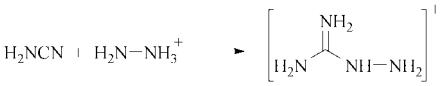

Aminoguanidine salts react further with hydrazine in aqueous or alcoholic solution to give salts of 1,2,3-triaminoguanidine [*2203-24-9*] [51].

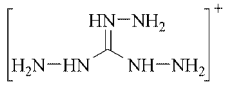

The reaction of cyanamide with phenylhydrazine hydrochloride in boiling alcohol produces two isomeric aminoguanidines [52].

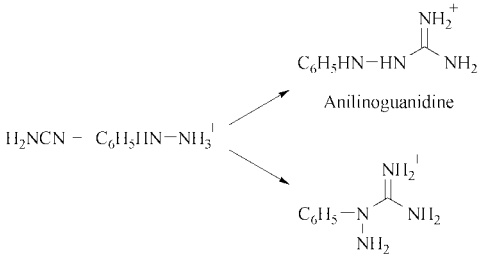

Reaction with Hydroxylamines. Hydroxylamine hydrochloride reacts slowly with aqueous cyanamide to give hydroxyguanidine [*13115-21-4*] [53].

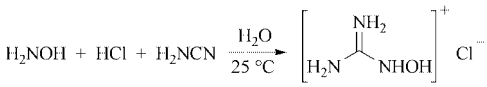

Cyanide, also that present in crude calcium cyanamide (see "Black Cyanide" in Chapter 1.3.1) can react with cyanamide via unisolable intermediates forming oxalate upon hydrolysis [54]. 1,3,5-Triazine carboxylates can be found as intermediates or side-products [55].

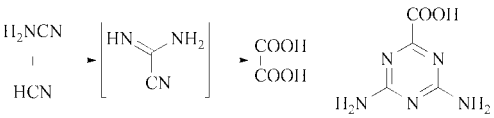

Alkylation with Alkyl Halides. Cyanamide salts can be dialkylated, especially with allyl and benzyl chlorides in water or alcohol at 60 °C [56].

$$2\,CH_2{=}CHCH_2Cl + NaHNCN + NaOH \rightarrow (CH_2{=}CHCH_2)_2NCN + 2\,NaCl + H_2O$$

Contrary to alkylations, cyanamide reacts upon silylations and stannylations as the carbodiimide tautomer. Bis-trimethylsilylcarbodiimide can be made from hexamethyldisilazane [57]. It is a versatile reagent to convert carbonyl functions to *N*-cyanoimines.

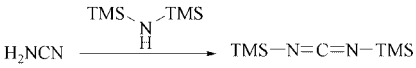

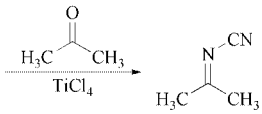

Acylation. Cyanamide can be acylated by reaction with carboxylic esters, acid chlorides, or anhydrides. The resulting *N*-cyanoamides are relatively strong acids that are best isolated as alkali metal salts [58]. Starting from alkyl chloroformates or dialkyl carbonates, alkoxycarbonylcyanamides are formed [59, 60]. Methoxycarbonylcyanamide is used to produce the benzimidazole fungicide Carbendazime [61].

N-Cyanoimines. The reaction of orthoesters, imidates, or amidines with cyanamide gives compounds such as *N*-cyanoimidates or *N*-cyanoamidines [62, 63]. These compounds are useful intermediates for the synthesis of 2-amino-1,3,5-triazines. Dimethyl-*N*-cyanimidocarbonate [64] (accessible from dimethylimidocarbonate) and dimethyl-*N*-cyanoimidodithiocarbonate [65] (from CS_2, cyanamide, and a methylating agent) are convenient starting materials for *N′*,*N″*-disubstituted *N*-cyanoguanidines such as Cimetidine or Tiotidine [66 – 68].

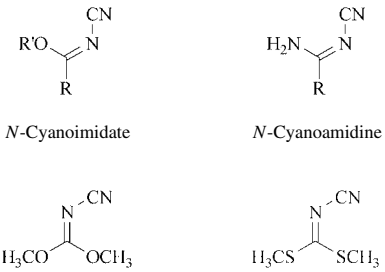

N-Heterocycles. A vast variety of *N*-heterocycles like imidazoles, 1,2,4-triazoles,

tetrazoles, pyrimidines and triazines can be made from cyanamide, often via isolable intermediates like guanidines, isourea compounds, or *N*-cyanoimines. Typical reactions are the synthesis of 2-aminoimidazoles [69, 70], 3-amino-1,2,4-triazole [71], 5-aminotetrazole [72], 2-aminopyrimidines [73], 2-halogenopyrimidines [74], 2-amino-1,3,5-triazines [62, 75], and 3-amino-1,2,4-benzotriazine-1-oxides [76].

2.3. Production

Cyanamide is manufactured from calcium cyanamide by continuous carbonation in an aqueous medium [77].

$$CaNCN + H_2O + CO_2 \rightarrow NH_2CN + CaCO_3$$

The following reactions take place:

$$2\,CaCN_2 + 2\,H_2O \rightarrow Ca(NHCN)_2 + Ca(OH)_2$$

$$Ca(OH)_2 + 2\,CO_2 \rightarrow Ca(HCO_3)_2 \rightarrow CaCO_3 + H_2O + CO_2$$

$$Ca(HNCN)_2 + CO_2 + H_2O \rightarrow 2\,NH_2CN + CaCO_3$$

In the continuous cyanamide process, the amount of calcium cyanamide and water added is equivalent to the aqueous cyanamide removed (Fig. 4) [78].

A carbonated slurry of cyanamide solution, solid calcium carbonate, and graphite is cooled to remove the heat of reaction (d). Part of the slurry is recycled to facilitate temperature control; the remainder is filtered, yielding a cyanamide solution and a cake of calcium carbonate and graphite. The final concentration of cyanamide is normally 25 %.

The calcium cyanamide feed is well mixed with the recycled slurry and filtrate in a feed vessel (b) and added at a rate to maintain a pH of 6.0 – 6.5 in the cooling tank. After carbonation, the slurry is held at 30 – 40 °C to complete the formation of calcium carbonate. The equipment is preferably stainless steel. This process is somewhat modified to produce a commercial 50 % solution and recover crystalline cyanamide. Calcium and iron are removed by ion exchange, if desired. Alternatively, the cyanamide solution can be further purified by membrane filtration [79].

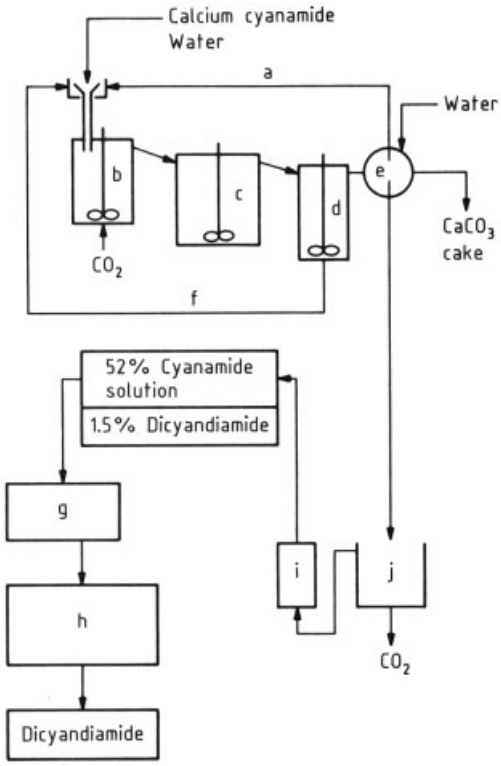

Figure 4. Continuous manufacture of cyanamide and dicyandiamide from calcium cyanamide
a) Liquid return line; b) Reactor; c) Reprecipitation tank; d) Cooling tank; e) Filtration equipment; f) Slurry return line; g) Cooling and crystallization of cyanamide; h) Solution with less than 38 % cyanamide; i) Ion exchange; j) Storage tank

Commercial Forms. The two main producers of cyanamide are Degussa AG in Trostberg Germany and Nippon Carbide in Japan. Both offer two grades:

1) Cyanamide in 50 % aqueous solution (trade names Cyanamid L500 and NC Cy-50), a clear slightly yellowish solution that can change reversibly to blue under the influence of light.
   Specifications:

Cyanamide	50 ± 1 %
Dicyandiamide and related compounds	max. 5 %
Iron	max. 500 mg/L
Calcium	max. 300 mg/L
Stabilizer	max. 2 %

The maximum storage life is 3 months if stored cool (below 20 °C) and under controlled conditions. It is shipped in polyethylene cans or drums up to 200 liters or as bulk

in specialized ISO containers equipped with a pressure relief system.

2) Pure, crystalline cyanamide (trade names Cyanamid F1000 and NC Cy-100): colorless crystals that are slightly hygroscopic and tend to cake under pressure.
Specifications:

Cyanamide	min. 99 %
Dicyandiamide	max. 1 %
Water	max. 1 %

The storage life is at least 3 months at temperatures below 20°C. The product is packed in steel drums with polyethylene inliners. The packaging size is limited to 50 kg.

2.4. Analysis

Cyanamide is determined by potentiometric titration with silver nitrate in ammoniacal medium [80]. It can also be determined by precipitation as the silver salt in ammoniacal solution and back-titration of the excess of silver ions [81] or by determination of nitrogen in the silver cyanamide precipitate [82]. Cyanamide can also be determined by direct HPLC with UV detection at 200 nm [83].

Traces of cyanamide can be analyzed spectrophotometrically after derivatization with pentacyanoammineferrate(II) [84]. Due to its low UV absorption, direct HPLC determination is not generally suitable. However, after derivatization with 1,2-naphthoquinone-4-sulfonate or with acetic anhydride, it can be detected by HPLC [85].

Dicyandiamide, the main impurity in cyanamide is routinely determined by reverse-phase HPLC (see Section 3.4).

2.5. Storage and Transportation

The commercial 50 % solution produced by Degussa AG is stabilized by a special buffer system to a pH between 4.0 and 4.6. Typically, it contains less than 1.5 % dicyandiamide, less than 0.5 % urea, and less than 0.1 % thiourea. This solution is stable for at least 3 months if stored at a temperature below 20 °C. Prolonged storage leads to an increase in the pH and thus to instability of the solution. If a pH of 5.2 has been reached, the product should be used quickly, above pH 5.5 it should be discarded immediately by mixing it with at least the double volume of water. Otherwise, a thermal runaway reaction leading to temperatures above 100 °C and potential bursting of the container may occur.

Aequous cyanamide solutions can be stored in stainless steel (e.g. DIN 1.4541, AISI 321 or DIN1.4571, AISI 316Ti), glass or glass-lined vessels, polyethylene, polypropylene, or poly(vinyl chloride) containers. Iron, copper, zinc, and brass unsuitable for storing cyanamide.

The commercial crystalline Cyanamide F1000 from Degussa AG contains a special stabilizer system to ensure safe handling and transportation. It can be stored at least for 3 months at or below 20 °C. Temperatures above 40 °C must be avoided in any case.

The following actions should never be done with commercial cyanamide: Addition of strong acids, addition of bases or basic substances, evaporation of aqueous cyanamide solution, recrystallization or distillation of cyanamide.

According to the valid regulations on transport of dangerous goods, a stabilized 50 % aqueous cyanamide solution is classified in class 8 (corrosive substances), UN number 2922, packaging group III, hazard label 8 and 6.1. As far as bulk containers are used for transport, the designated tanks (24 m^3) must be thermally insulated and equipped with a pressure relief system. Two bursting disks are required, with a diameter of 25 cm each, responding at an overpressure level of 6 bars. In addition, the maximum filling temperature must not exceed 18°C [86]. Cyanamide solutions shipped in drums of up to 200 liters do not require special precautions.

2.6. Uses

Due to its unique structure and reactivity, cyanamide has found very widespread application in many different areas like agrochemicals, pharmaceuticals, and specialty and industrial chemistry.

Agrochemistry. Cyanamide is used as a plant growth regulator and budbreaking agent (Dormex) to compensate for the lack of frost and to improve the budbreak of grapes and other fruit crops in areas with mild winters [87].

Cyanamide has a significant fungicidal and herbicidal potential [88] and is used in pig farming to overcome sanitary problems related with the pig manure. Because of rapid degradation, cyanamide does not cause any residue problems, but contributes to the overall nitrogen supply of the crops.

Cyanamide is the key raw material for many herbicides, fungicides, and insecticides. It is used to produce the herbicides Amitrol [71] and Hexazinone [89] and builds up the pyrimidine or 1,3,5-triazine subunit for a vast number of sulfonylurea herbicides like Bensulfuron or Metsulfuron methyl [90, 91]. The benzimidazole fungicides Carbendazim and Benomyl and the pyrimidine fungicides Cyprodinil, Pyrimethanil, and Mepanipyrim [92, 93] are based on cyanamide, as well as the pyrimidine insecticides Bupyrimate and Dimethirimol [94, 95] and neonicotinoids like Acetamiprid [96].

Many fungicidal and biocidal guanidine and biguanide compounds like Dodine [97], Guazatine and Iminooctadine [98] are based on cyanamide chemistry.

Pharmaceuticals. Cyanamide is the key raw material for the anti-ulcer drug Cimetidine [99]. The majority of antiviral and anticancer drugs with pyrimidine or purine structures are based on cyanamide and guanidine chemistry.

Some new anticancer drugs like Gleevec [100] and TMC-125 [101] exhibit an aminopyrimidine core derived from cyanamide.

The antihypertonic Minoxidil [102] has now found a second application as a hair growth stimulant. The 2-aminobenzimidazole unit in anthelmintics for human and veterinary application, such as Albendazole, Flubendazole, and Mebendazole [103], is derived from cyanamide. Fattyl alkyl guanidine salts are used in cosmetics for hair repair [104].

Industrial Chemistry. Cyanamide reacts with hydrogen peroxide to activate the peroxide bond. This has been used to improve bleaching of cellulosic pulp and textiles [37, 105 – 108]. Cyanamide salts of lead and zinc are potent corrosion inhibitors that are used for metal coatings [109]. Cyanamide is used to modify dyes for higher fiber reactivity [110]. In polymers, cyanamide has been used to cure epoxy resins [111] and to modify polyurethane dispersions [112, 113].

2.7. Legal Aspects

Cyanamide is listed in global inventories of existing chemicals, e.g., EINECS (European Union), TSCA Inventory (United States), DSL (Canada), AICS (Australia), IECSC (China) and MITI (Japan). Within the European Union, the chemical substance cyanamide , i.e., the solid compound, is classified as toxic (danger symbol T) and must be labeled with R21 (harmful in contact with skin), R25 (toxic if swallowed), R36/38 (irritating to eyes and skin) and R43 (may cause sensitization upon skin contact).

Following specific legal EU definitions, a 50 % aqueous cyanamide solution is to be considered as a preparation that was examined per se for acute oral toxicity (see Chapter 5). Consequently, this preparation has to be classified as harmful (danger symbol X_n) and R22 (harmful if swallowed) instead of R25. The other R phrases mentioned above are applicable for 50 % cyanamide solutions too. In Germany, cyanamide is classified in water hazard class 2 (hazardous for waters).

3. Dicyandiamide

The dimer of cyanamide is known as dicyandiamide or cyanoguanidine. The occurrence of dimerization on evaporation of an aqueous solution was known as early as 1851; the product was later named param (from paracyanamide). The name dicyandiamide was coined in 1862 [114] and, together with the abbreviation *dicy*, is now firmly established. The name dicyandiamide is not as descriptive as the name cyanoguanidine, which is preferred by Chemical Abstracts; however, the former is used in this article.

Raman spectra support the cyanoguanidine structure (**1**) [115]; the crystalline form apparently has the tautomeric form (**2**) [116].

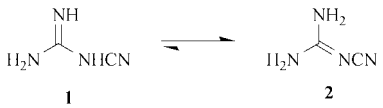

Tautomerism undoubtedly exists between two or more forms of dicyandiamide. The preferred formulation appears to be tautomer **2**.

Dicyandiamide has achieved industrial importance as an intermediate for the production of guanidine salts (→ Guanidine and Derivatives), guanamines, and formerly melamine (→ Melamine and Guanamines).

3.1. Physical Properties [6]

Dicyandiamide [*461-58-5*], M_r 84.084, is an odorless, colorless, nonvolatile powder with a monoclinic prismatic crystal structure [117]. Solubilities are shown in Table 1 and Figure 5A; some other physical properties follow:

Melting point	210 – 212 °C
Crystal density	1.404 g/cm^3
Enthalpy of formation at 25 °C	+18.4 kJ/mol
Dissociation constants	
pK_a	14.8
pK_b	14.4

In some solvent mixtures, especially with strong hydrogen bonding capacity, dicyandiamide exhibits a high cosolvent effect with increased solubility (Fig. 5/>B).

3.2. Chemical Properties

Dicyandiamide reacts with a variety of reagents. It is the simplest organic compound containing the C−N, C=N, and C≡N groupings. Reaction may occur at one or more of these sites. The literature on the chemistry of dicyandiamide is extensive, and many more reactions than those indicated below have been reported.

Formation. Dicyandiamide is formed from ammonium dicyanoamide, equivalent to the Woehler urea synthesis [118].

Decomposition Reactions. Dicyandiamide is thermally stable up to about 170 °C, above which decomposition reactions occur that become violent around the melting point. Therefore, the measured melting point strongly depends on the heating rate. Decomposition products are melamine, ammonia, and higher condensed triazines like melam, melem, and melone [119].

Salt Formation. Dicyandiamide is amphoteric and forms salts with organic and inorganic acids and with alkali in an anhydrous medium. In very concentrated aqueous acid, the corresponding salt precipitates. Such salts are almost completely hydrolyzed in dilute solution.

With alkali and alkaline earth metal hydroxides, dicyandiamide forms salts [120], which are stable in concentrated aqueous solutions but hydrolyze in dilute solutions. The sodium salt is 94.5 % hydrolyzed in 0.1 M solution.

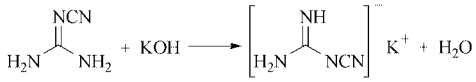

Hydrolysis. At elevated temperature in the presence of an equivalent amount of a mineral acid, dicyandiamide yields guanylurea salts. This reaction is quantitative and can be used as a basis for the determination of dicyandiamide [121]:

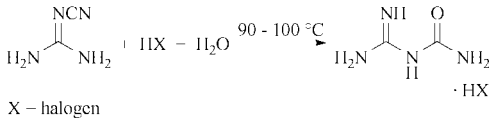

At atmospheric pressure and moderate temperature in the presence of aqueous ammonia, reaction of dicyandiamide with hydrogen sulfide gives guanylthiourea. With increased pressure, an additional atom of sulfur may be introduced to yield dithiobiuret [122]. Unlike dicyandiamide, guanylurea is basic and readily forms salts with organic and inorganic acids.

The reaction of dicyandiamide with concentrated sulfuric acid or glacial acetic acid containing a small amount of water yields guanidine and ammonium salts instead of guany-

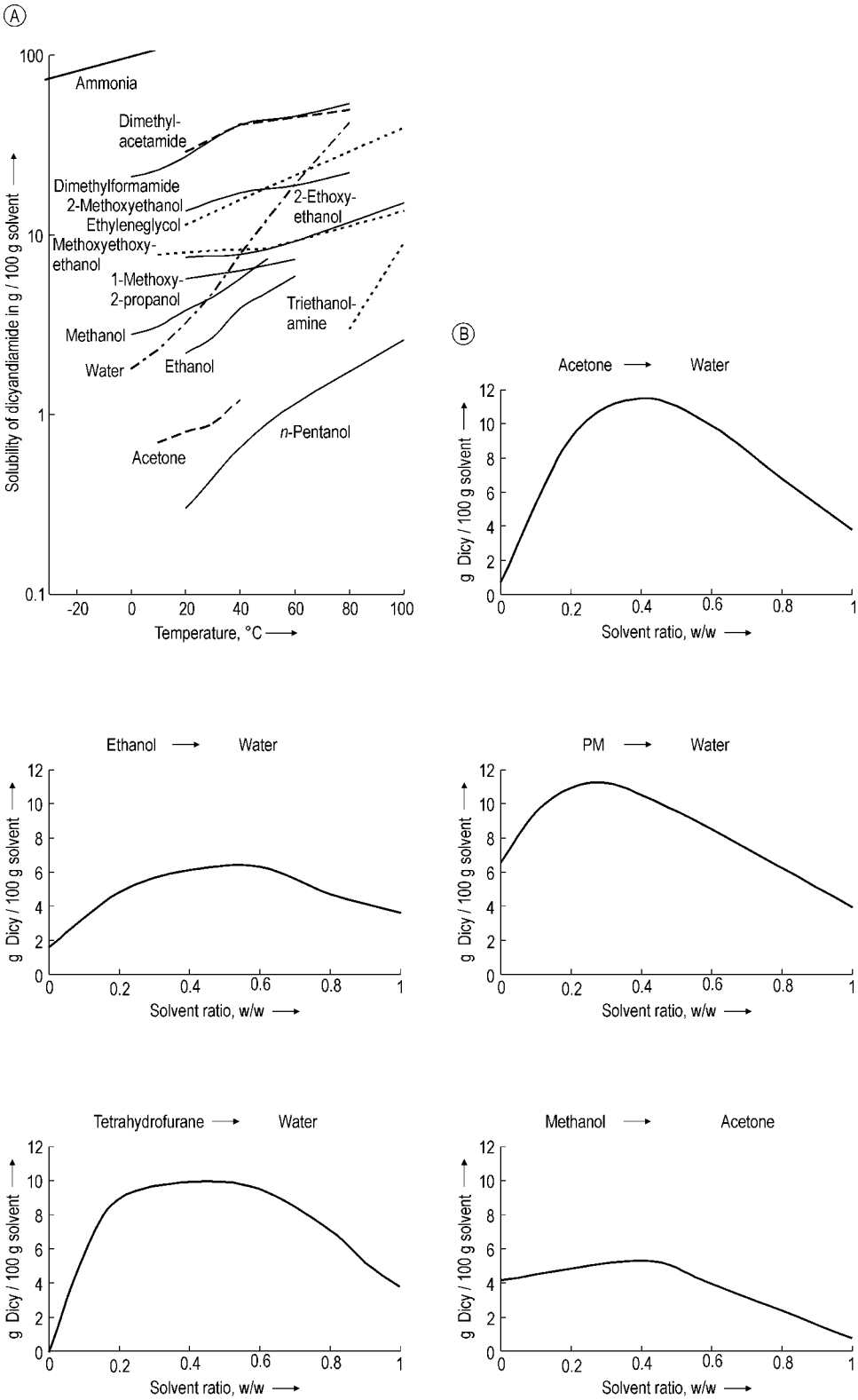

Figure 5. A) Solubility of dicyandiamide in various solvents (g/100 g solvent); B) Solubilities of dicyandiamide (dicy) in some solvent mixtures (g/100 g solvent, solvent ratio by weight)

Table 1. Solubility of dicyandiamide in various solvents (g/100 g solvent)

Temperature	−33	0	10	20	30	40	50	60	80	100
Ethanol				2.2	2.7	3.9	4.8	5.9		
2-Methoxyethanol				13.6		17.1		18.8	22.1	
2-Ethoxyethanol				7.5			8.3			15.1
1-Methoxy-2-propanol				5.7		6.3		7.3		
Acetone			0.7	0.8	0.9	1.2				
Dimethylformamide		21	23	27.2		40.9		45.9	53.7	
n-Pentanol				0.3			0.9			2.6
Ammonia	72.0				131					
Dimethylacetamide				29.0		41.1		44.8	49.7	
Triethanolamine									3.0	9.0
Methoxyethoxyethanol			7.8		8.2		8.4			13.6
Ethylene glycol				11.3			18.3			39.6
Water		1.8	2.3	3.2	4.7	7.8	12.3	19	42	
Methane		2.8	3.1	3.8	4.5	5.7	7.4			

lurea. Treatment of dicyandiamide with nitric acid gives nitroguanylurea [123].

The imino group of dicyandiamide is hydrolyzed by barium hydroxide to give cyanourea [*2208-89-1*].

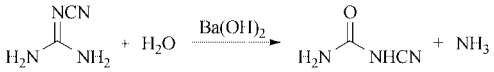

Reaction with Amines. Heating dicyandiamide with aromatic amines in water in the presence of an equivalent amount of mineral acid gives high yields of aryl biguanide salts [124].

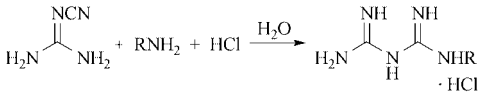

R = aryl

Yields of biguanide salts from the reaction of dicyandiamide with ammonium salts are poor. The biguanide obtained reacts further with the ammonium salt, and the product forms the guanidine salt. This reaction is used for the commercial preparation of guanidine nitrate [*506-93-4*] [125]:

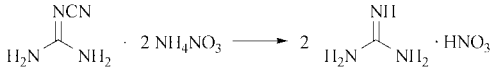

Reaction with Formaldehyde. Treatment with formaldehyde [126] produces resinous compositions of varying properties (→ Amino Resins) under acidic or alkaline conditions. The reaction can be controlled to give mainly the water-soluble hydroxymethyldicyandiamide, or carried further to yield amorphous, infusible solids.

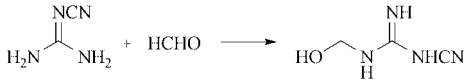

Heterocyclic Compounds. Under pressure and in the presence of ammonia, dicyandiamide cyclizes to melamine, of which large amounts have been made in this manner. At present, however, melamine is produced primarily by the urea process [127].

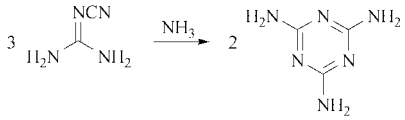

Excellent yields of guanamines are obtained when dicyandiamide is heated with alkyl or aryl nitriles in the presence of a small amount of alkali [128].

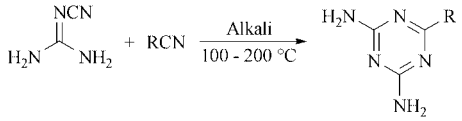

In some cases, carboxylic esters can be treated with dicyandiamide in the presence of an alcohol and a metal salt to form substituted 1,3,5-triazines [129].

Reaction with urea affords ammeline [*645-92-1*].

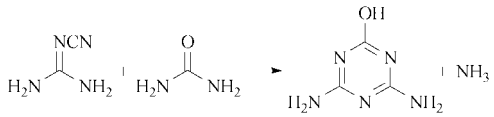

Dicyandiamide reacts with a wide variety of compounds to yield substituted pyrimidines. Acetoacetic ester reacts with dicyandiamide in the presence of sodium ethoxide and ethanol to give 2-cyanoamino-4-hydroxy-6-methylpyrimidine [130]. Derivatives of barbituric acid are synthesized by condensation of dicyandiamide with disubstituted malonic or cyanoacetic acid; their ester, chloride, or nitrile derivatives yield cyaniminopyrimidines [131]. These can be alkylated in the 3-position and hydrolyzed to *N*-monoalkyl-5,5-disubstituted barbituric acids [123].

3.3. Production

Dicyandiamide is formed in almost quantitative yield by the dimerization of cyanamide in alkaline solution. The reaction occurs most readily at pH 8 – 10 by the addition of the anionic cyanamide species to the nitrile group. The product is a weaker acid than cyanamide; it is protonated immediately with the generation of a new cyanamide anion, which continues the process.

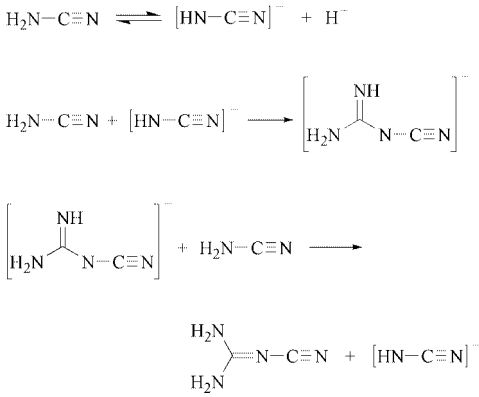

In the industrial production of dicyandiamide, cyanamide is first prepared from calcium cyanamide by continuous carbonation in an aqueous medium (see Chap. 2) at pH 7 – 8 and is then dimerized at pH 8 – 10.

$$CaNCN + H_2O + CO_2 \rightarrow H_2NCN + CaCO_3$$

$$2\,H_2NCN \rightarrow (NH_2)_2C{=}NCN$$

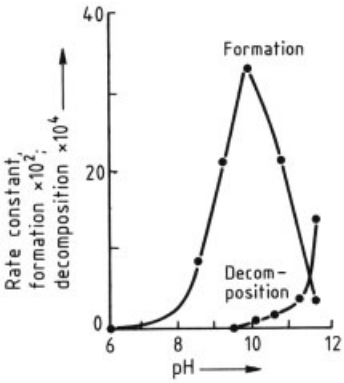

Figure 6. Effect of pH on rates of formation and decomposition of dicyandiamide

To obtain a clean product, dimerization is carried out after separation of the precipitated calcium carbonate and other insoluble matter such as graphite [132]. The filtered solution containing ca. 25 % cyanamide is adjusted to ca. pH 9 and held at ca. 80 °C for 2 h to effect the desired conversion. The resulting hot liquor is filtered and cooled in a vacuum crystallizer. The dicyandiamide crystals are centrifuged and transferred to a rotary drier. The pH of the reaction must be controlled. An increase in alkalinity accelerates decomposition by hydrolysis (Fig. 6).

The dried product is stored in bulk or packaged in multiwall bags.

3.4. Quality Specifications and Analysis

Dicyandiamide is determined by reverse-phase HPLC with water or water/methanol (99/1) as mobile phase and detection at 220 nm [133, 134]. Alternatively, thin layer chromatography on silica plates with acetonitrile/water/acetic acid (90/10/1) as mobile phase can be used. The dicyandiamide spots are visualized with elemental chlorine and potasssium iodide/starch [135]. After derivatization with 3-hydroxy-2-butanone, dicyandiamide can be analyzed by HPLC and detected at 260 nm [136]. For resin applications, the color of a solution of dicyandiamide in aqueous formaldehyde is determined, e.g,. according to the APHA scale.

A typical analysis of commercial dicyandiamide is:

Dicyandiamide	min. 99 %
Water	max. 0.2 %
Water insolubles	max. 0.1 %
Melting point	209 – 212 °C

Dicyandiamide is also sold in micronized form or as solution for special applications.

3.5. Storage and Transportation

Dry crystalline dicyandiamide is stable at normal temperatures. Dicyandiamide is stable in aqueous solution up to 80 °C; above 80 °C ammonia is liberated. Dicyandiamide is not hygroscopic, flammable, or corrosive.

Dicyandiamide is packed in paper bags of 25 kg net weight or in big bags of up to 1000 kg.

3.6. Uses

Dicyandiamide is an intermediate for a number of organic nitrogen compounds, such as guanamines, biguanides, and guanidine salts. Guanidine phosphate is used as a fire-retardant in water-based systems; guanamines are employed as copolymers in resin compositions. Melamine is used primarily as a cross-linking agent in coatings and laminates. Substituted biguanides, prepared from dicyandiamide and arylamines, are used as stabilizers and antioxidants for soaps and other fatty materials.

Dicyanamide condensates. The condensation of dicyandiamide with formaldehyde gives resins with broad industrial applications. The reaction can be controlled to give a water-soluble intermediate for impregnation of paper, wood, leather, and textiles [137]. These resins are effective as direct-dye fixing agents for textiles and leather products. Such agents allow dyeing and printing of fabrics with direct dyes, rather than with more expensive vat dyes. They also improve wash- and lightfastness. Dicyandiamide itself prevents odor formation in resin-treated fabrics.

Melamine resins can be flexibilized by modification with dicyandiamide, thus giving postforming properties to high-pressure laminates [138].

Dicyandiamide – formaldehyde condensates modified with cationic polyaminopolyamides have been used in paper manufacture [139]. Avoiding the use of formaldehyde, dicyandiamide can be condensed with polyamines with cleavage of ammonia. These condensates are used as textile auxiliaries [140].

Under certain conditions, formaldehyde and mixtures of dicyandiamide with casein or other proteins, urea, melamine, or aromatic amines give resinous compositions varying from white powders to creamy pastes or clear, glasslike products.

Dicyandiamide reduces the viscosity of certain colloidal solutions [141]. This property is commercially significant in the manufacture of glues and adhesives, in the coating and sizing of paper and textiles, and in the conditioning of phosphate drilling muds. This action could be useful in other applications requiring viscosity control.

Dicyandiamide fluidizes adhesives and glues that are based on converted starches, casein, α-protein (soya protein), and hide proteins or collagens, thus extending the useful life of the adhesive. Dicyandiamide is 2 – 5 times more effective than urea on the basis of weight.

Flame retardants. Dicyandiamide, in combination with phosphoric acid, is used as a flame-retardant for cellulosic materials [142]. It is used

commercially as a fire-retardant for wood, particularly shingles [143, 144].

Dicyandiamide is stable at room temperature and is used in coating formulations as a flame-retardant. When exposed to a flame, the coating composition bubbles up or intumesces to a foam that protects the substrate from ignition.

An intumescent formulation including dicyandiamide, pentaerythritol, monoammonium phosphate, titanium dioxide, latex binder, and water gives good foam volume at low coating weights with reasonable stability of the foam structure under ignition [145, 146].

Smoke- or gas-generating pyrotechnic compositions contain dicyandiamide as fuel with high nitrogen production. Such formulations are used for car air-bags [147] and for agrochemical fumigants [148].

Plastic additives. Dicyandiamide can be used as a component in copolymers and co-condensates [149]. It is used for the thermal stabilization of poly(vinyl chloride) (PVC), for which it is introduced in the solid form at 170 °C into the PVC melt or as an aqueous solution to a PVC emulsion [150]. It stabilizes cellulose and homopolymers and copolymers of formaldehyde.

Epoxy curing. A major industrial application for dicyandiamide is curing of epoxy resins. At temperatures above 180°C, it reacts in a highly exothermic, autocatalytic reaction to form a rigid polymer network. Although many mechanistic investigations are reported [151 – 154], the reaction pathways and the final structure of the cross-linked network is still uncertain. However, it is obvious, that dicyandiamide not only reacts at the two amino groups, but involves the cyano group reacting with α β-hydroxy group formed from the epoxy resin to form oxazolidinone ring structures [155]. The maximum cross-link density is obtained at a molar ratio dicyandiamide-to-epoxy near 1:7, which means to a hydrogen equivalent weight (HEW) for dicyandiamide of 12 g mol^{-1}.

For industrial applications, the curing temperature of pure dicyanamide/epoxy systems is generally too high. Therefore, accelerator compounds are used for most applications. Imidazole compounds like 2-*methylimidazole* [*693-98-1*] can decrease the curing temperature to 100–160°C depending on the accelerator amount and resin type, but signifcantly decrease the storage life of the resin mixture. *N′*,*N′*-Dimethyl-*N*-arylurea compounds (often called "urones") like Diuron [*330-54-1*] or 2,4-toluylene-bis(*N′*,*N′*-dimethylurea) [*17526-94-2*] are generally used as latent accelerator compounds providing low curing temperatures (120–160°C) combined with a long storage lifetime of the formulated material [155 – 157]. Amine adducts with, e.g., polyphenols are also used as latent accelerators [158].

Main industrial application fields for dicyandiamide-cured epoxy resins are in electro laminates (printed circuit boards), fiber-reinforced plastics (glass and carbon fiber composites, e.g., for ski equipment, tennis rackets, fishing rods, automotive and aerospace structural applications), heat-curable adhesives (mainly for structural bonding in the automotive and aerospace field) and decorative and functional powder coatings.

For printed circuit boards, dicyandiamide is generally used as a solution in various solvents, methyl cellosolve and dimethylformamide being preferred because of their good solubility properties (see Section 1.1). Imidazoles (usually 0.05 to 0.1 g per 100 g resin) are used as accelerators to catalyze the reaction with mainly brominated bisphenol A epoxy resins. Silane-coated glass fabrics are impregnated with the resin varnish, dried, and prereacted to form a prepreg. Those are pressed in several layers together with copper foil and cured at about 180°C [159] to form a rigid printed circuit board. It is known, that copper interferes with the dicy/epoxy reaction thus influencing adhesion and the polymer cross-link structure [160].

For composite materials and structural adhesives, dicyandiamide is used in a micronized form (particle size < 10 μm or below) in combination with urone-type accelerators [156, 157]. Typical formulations contain 100 parts of a standard liquid bisphenol A resin, 7 parts dicyanamide and 1 to 5 parts accelerator. The pasteous mixture is rolled onto fiber filaments or fabrics and prepregged by short period heating. The prepregs are pressed under vacuum and heat to form the specimen. Resins transfer molding is a newer development demanding for even smaller particle size dicyandiamide because of filtering effects.

Adhesives are based on the same resin/dicyanamide/accelerator mixtures, typically for-

mulated including mineral fillers and fumed silica to provide thixotropy. Such adhesives are mainly used for metal bonding, because dicyandiamide strongly increases adhesion to metals by altering the surface network structure [161].

For decorative powder coatings, dicyanamide/epoxy systems have lost some commercial significance in the past years, in favour of cheaper epoxy/polyester hybrids. However, functional coatings with high chemical resistance, good adhesion, and high corrosion resistance (e.g., for concrete rebar coating, pipe coating, and gas or water fittings) are mainly based on dicyandiamide-cured epoxy systems, often accelerated by imidazole/epoxy adducts.

Agriculture. Dicyanamide is used as a fertilizer additive to prolong the shelf life of nitrogen nutritients in soil [162 – 164]. This is based on the ability of dicyandiamide to inhibit the urease and nitrification activity of soil microorganisms [165].

Pharmaceuticals. Dicyandiamide is the raw material for antidiabetic biguanides like Metformin [*657-24-9*] and Phenformin [*114-86-3*], the antimalarial Proguanil [*500-92-5*], and the antiviral Moroxidine [*3731-59-7*].

3.7. Legal Aspects

Dicyandiamide is listed in global inventories of existing chemicals, e.g., EINECS (European Union), TSCA Inventory (United States), DSL (Canada), AICS (Australia), IECSC (China), and MITI (Japan). It does not need to be classified and labeled according to EU Directives on dangerous substances. In Germany, dicyandiamide is classified in water hazard class 1 (slightly hazardous for waters).

4. Substituted Cyanamides

Substituted cyanamides have found much less technical application. In general, they are not produced from cyanamide itself, but from cyanogen chloride by reaction with amines.

$$R_2NH + ClC{\equiv}N \rightarrow R_2NC{\equiv}N + HCl$$

Alkyl- and arylcyanamides are intermediates in the production of *N*,*N′*-disubstituted guanidines; arylcyanamides have been proposed as curable resins.

Dicyanamide is an instable, strongly acidic substance. The stable sodium dicyanamide [*1934-75-4*] is produced from cyanamide and cyanogen chloride on a commercial scale. It is used to produce biguanide disinfectants like PHMBG and Chlorhexidine [166]. It is also is a valuable reagent to synthesise certain 1,3,5-triazines [167, 168].

5. Toxicology and Occupational Health

The production of calcium cyanamide and cyanamide does not present any serious health hazard [169, 170]. However, exposure to these substances in combination with the consumption of alcoholic beverages may cause a vasomotor reaction known as cyanamide flush or mal rouge. Typical symptoms are hyperemia of the upper half of the body, tachycardia, dyspnoea, headache, and nausea [171]. Cyanamide inhibits the enzyme aldehyde dehydrogenase and thus interferes with the oxidation of alcohol. This results in the accumulation of acetaldehyde in the blood, an effect, which may account for these unpleasant, but transient symptoms.

Calcium Cyanamide. The acute oral LD_{50} of technical-grade calcium cyanamide is 765 mg/kg in rats [172] and 1800 mg/kg in mice [173]. The dermal LD_{50} (rabbit) is greater than 2000 mg/kg [173]. From a study on the acute inhalation toxicity (rat) it can be concluded that the 4-h LC_{50} of the technical-grade material is greater than 155 mg/m^3. No mortalities were observed at this maximum attainable concentration [172]. A single semi-occlusive application of calcium cyanamide (technical grade) to intact rabbit skin for 4 h elicited no dermal irritation [173]. Introduction of this test material into the rabbit eye produced severe ocular lesions [174]. A battery of mutagenicity tests (in vitro, in vivo) with technical-grade calcium cyanamide did not reveal indications for mutagenic activity [174]. According to a bioassay of calcium cyanamide for possible carcinogenicity, the test formulation was not carcinogenic for F344 rats or B6C3F1 mice of either sex [175].

Cyanamide. A comprehensive overview on the toxicity of cyanamide is given in [176]. The acute oral LD_{50} (rat) is 125 mg/kg. When applied dermally to rabbits, the LD_{50} was found to be 848 mg/kg. From an acute inhalation toxicity study in rats (4-h exposure) it appeared that the LC_{50} is greater than 1000 mg/m^3. No mortality or severe injury occurred at this highest possible concentration. All these values refer to the active ingredient cyanamide. The corresponding data for a 50 % aqueous solution are 300 mg/kg (acute oral LD_{50}, rat), 1700 mg/kg (acute dermal LD_{50}, rabbit) and $>$ 2000 mg/m^3 (LC_{50} inhalation, rat). Cyanamide (solid or 50 % aqueous solution) has distinct irritating properties to skin and eyes and is regarded as sensitizing in the guinea pig (Magnusson – Kligman method). Cyanamide did not show mutagenic effects in several experiments such as the Ames Test.

Dicyandiamide is essentially nontoxic [177]. The acute oral LD_{50} (rat) is higher than 10 000 mg/kg. In rabbits, the dermal LD_{50} is $>$ 2000 mg/kg. After inhalation of dust for 4 h, the LC_{50} in rats was greater than 259 mg/m^3. No mortality or signs of intoxication were seen at this maximum attainable concentration. The substance is not irritating to skin and eyes and does not have a sensitizing effect in the guinea pig. A battery of in vitro tests did not reveal mutagenic properties. The compound was not carcinogenic in an oral long-term study with rats. From human experience it appears that product dust may cause temporary mechanical eye irritation. An individual case of allergic contact dermatitis was observed in a worker employed in the production of flame retardants [178]. Patch tests with the dry powdered material on 200 human subjects did not indicate sensitizing or primary irritant effects [179].

Exposure Limits. Regarding the maximum allowable concentration at workplace, no such value has been set for dicyandiamide [177]. In the United States, recommended exposure limits are existing for calcium cyanamide (0.5 mg/m^3) and cyanamide (2 mg/m^3) [180]. The occupational exposure limit in Germany is 1 mg/m^3 for calcium cyanamide as well as for cyanamide [181].

6. References

General References

1. *Houben-Weyl*, vol. 8, 1952.
2. *Houben-Weyl*, vol. E4, 1983.

Specific References

3. E. Drechsel, *J. Prakt. Chem.* **16** (1877) 180 – 200.
4. F. Emich, *Monatsh. Chem.* **10** (1889) 321 – 352.
5. DE 88 363, 1895 (N. Caro, A. Frank).
6. *Gmelin*, system no. 28, Calcium, Parts B 1 – 2 (1956) pp. 179 – 195;
   Gmelin, system no. 14, Carbon, Part D 1 (1971) pp. 258 – 297.
7. *Gmelin*, system no. 28, Calcium, Part B 3 (1961) pp. 962 – 971.
8. T. Aono, *Bull. Chem. Soc. Jpn.* **16** (1941) 92 – 98.
9. V. Ehrlich, *Z. Elektrochem.* **32** (1926) 187 – 188.
10. F. E. Polzeniusz, *Chem. Ztg.* **31** (1907) 958.
11. M. L. Kastens, W. G. McBurney, *Ind. Eng. Chem.* **43** (1951) 1020 – 1033.
12. H. Roech, *Chem. Ztg. Chem. Appar.* **88** (1964) 191 – 199.
13. *Gmelin*, system no. 28, Calcium, Part B 1 – 2, 1956 – 1957, pp. 179 – 196.
14. Südd. Kalkstickstoff-Werke, DE 917 543, 1952 (F. Kaess et al.).
15. V. G. Golov et al., *Tr. N.-I i Proekt. In-Ta Azot. Prom.-Sti i Produktov Organ. Sinteza* **28** (1974) 49 – 52;
    Chem. Abstr. **84** (1975) 159 006 n.
16. "Cyanamide by Cyanamid," Amer. Cyanamid, Wayne, N.J., 1961;
    O. I. Polyanshikov et al., *Vopr. Khim. Tekhnol.* **39** (1975) 136.
17. K. Rathsack, *Landwirtsch. Forsch. 6 (Special ed.)* (1954) 116 – 132.
18. I. S. Cornforth, "Calcium Cyanamide in Agriculture", *Soil Fertilizers* **34** (1971) 463.
19. R. O. Hartman, T. G. Holt, A. R. Shirley, US 2004/16275.
20. Denki K.K., JP 2005/104766 (S. Sakashita, H. Naokawa).
21. Nukem, DE 2 926 107, 1981 (H. Qui Umann).
22. Amer. Cyanamid, US 3 503 766, 1970 (F. De Maria).
    R. E. Dwyer, US 4 049 465, 1977.
23. K. Arikawa, K. Inanaga, *Folia Psychiat. Neurol. Jpn.* **27** (1973) no. 1, 9.
24. Goerig Co., DE 1 771 827, 1973 (P. Birk, K. Wohlgemuth).
    K. Deutzmann, DE 2 136 450, 1973 (W. Blank, O. Vorbach).

25. Südd. Kalkstickstoff-Werke, DE 2 252 795, DE 2 252 796, 1974 (W. Meichsner).
26. A. Bineau, *Ann. Chim. Paris* **67** (1838) 225 – 272.
27. E. Winterstein, A. Küng, *Hoppe-Seylers Z. Phys. Chem.* **59** (1909) 144.
28. F. Baum, *Bio. Z.* **26** (1910) 325.
29. F. Baum, *Org. Synth. Coll. Vol.* **1** (1948) 203.
30. T. Kamo, S. Hiradate, Y. Fujii, *J. Chemical Ecology* **29** (2003) 275.
31. T. Kamo, K. Kato, S. Hiradate, Y. Fujii, M. Hirota, IPGSA Conference, Canberra, 22. Sept. 2004.
32. A. Ingar, R. W. A. Luke, B. R. Hayter, J. D. Sutherland, *ChemBioChem* (2003) 504.
33. M. Frund, A. Schander, *Ber. Dtsch. Chem. Ges.* **29** (1896) 2500 – 2505.
34. D. Costa, C. Bolis-Cannella, *Ann. Chim. (Rome)* **43** (1954) 769 – 778.
35. L. Denner, P. Luger, J. Buschmann, *Acta Cryst. C* **44** (1988) 1979.
36. M. M. Hooves, *Chem. Eng.* **57** (1950) 132 – 134.
37. H. Michaud et al., *Chemiker-Zeitung* **112** (1988) 287.
38. S. A. Miller, B. Bann, *J. Appl. Chem.* **6** (1956) 89 – 93.
39. H. Hope, *Acta Chem. Scand.* **18** (1964) 1800.
40. M. L. Kilpatrick, *J. Am. Chem. Soc.* **69** (1947) 40 – 46.
41. S. Basterfield, J. Tomecko, *Can. J. Research* **8** (1933) 458 – 462.
42. Takeda Chemical Ind., WO 97/00867 (H. Uneme, M. Konobe, H. Ishizuka, Y. Kamiya).
43. International Minerals & Chemicals Corp., US 2 620 354, 1952 (B. Vassel, W. D. Janssens).
44. H. B. Howard, J. K. Espenan, *Enzymologia* **8** (1939) 4.
45. SKW Trostberg, US 5 719 319, 1998 (S. Weiss, H. Krommer).
46. G. Edgar, W. S. Hinegardner, *Org. Synth. Coll. Vol.* **1** (1941) 172.
47. G. Losse, H. Weddige, *Angew. Chem.* **72** (1960) 323.
48. G. Losse, H. Weddige, *Justus Liebigs Ann. Chem.* **636** (1960) 14 4 – 149.
49. J. T. Thurston, L. P. Ferris: *Inorganic Synthesis*, vol. **3**, McGraw-Hill, New York 1950, pp. 45 – 47.
50. Bayerische Stickstoff-Werke, DE 730 331, 1942 (K. Zieke, G. H. Hoffmann).
51. W. Sauermilch, *Explosivstoffe* **12** (1965) no. 9, 197 – 199.
52. P. Fanti, H. Silbermann, *Justus Liebigs Ann. Chem.* **467** (1928) 274 – 284.
53. P. Adams et al., *J. Org. Chem.* **18** (1953) 934 – 940.
54. G. Barsky, *News Edit. Am. Chem. Soc.* **18** (1940) 759.
55. T. Guethner, ImSat-6 Conference, Rechenberg, Sept. 2003.
56. E. B. Vliet: *Organic Synthesis*, Collective volume 1, J. Wiley & Sons, New York 1941, pp. 203 – 204.
57. SKW Trostberg AG, DE 19822281, 1999 (T. Guethner, H. Krommer).
58. SKW Trostberg, DE 2 757 586, 1979 (S. Weiss).
59. DuPont, DE 1 795 849, 1977 (C. de Witt Adams, R. Schlatter).
60. SKW Trostberg, DE 2 747 472, 1979 (S. Weiss).
61. DuPont, US 5 943 84, 1966 (C. de Witt Adams, R. Schlatter).
62. K. R. Huffmann, F. C. Schaefer, *J. Org. Chem.* **28** (1963) 1816.
63. H. Schäfer, K. Gewald, *J. Prakt. Chem.* **318** (1976) 347.
64. Bayer Cropscience AG, WO 2004/101501 (H. Weintritt, M. Bauer, A. Heyn).
65. L. K. Marble, *Organic Preparations and Procedures International* **30** (1998) 473.
66. Smith Kline & French, DE 2 344 779, 1973 (G. J Durant, J. C. Emmet, C. R. Ganellin).
67. G. E. Robinson, *Chem. Ind.* (1983) no. 5, 349.
68. J. de Schrijver, *Chimica Oggi* **11/12** (1995) 21.
69. Lepetit S. p. A., US 3 450 709, 1969 (G. Lancini, E. Lazzari).
70. A. Kreuzberger, R. Schücker, *Arch. Pharm.* **306** (1973) 169.
71. Atochem, EP 168 296, 1985 (M. Alicot).
72. Nippon Carbide K. K., US 5 594 146, 1997 (M. Murotani, H. Mura, M. Takeda, H. Shibafuchi).
73. DuPont, US 4 235 802, 1980 (J. J. Fuchs).
74. Lonza, EP 582 288, 1993 (F. Previdoli, R. Imwinkelried).
75. SKW Trostberg AG, DE 3411202, 1985 (S. Weiss, H. Krommer).
76. Boehringer Mannheim, EP 675 881, 1993 (W.-G. Friebe, W. Scheuer, U. Tibes).
77. Amer. Cyanamid, US 3 300 281, 1967 (D. R. May).
78. Amer. Cyanamid, DE 1 568 032, 1966 (D. R. May).
79. Degussa AG, WO 03/089372 (T. Guethner, J. Fasenacht).
80. SKW Trostberg AG, Method-SOP 018–018/1, 1999 (M. Wildenauer).

81. SKW Trostberg AG, Method R85001, 1985 (U. Rust).
82. "VDLUFA Method 3.6.1", *Methodenbuch*, 4th ed., vol. **II** VDLUFA Verlag, Darmstadt 1995.
83. M. Turowski, B. Deshmukh, *Analytical Letters* **37** (2004) 1981.
84. D. A. Buyske, V. Downing, *Anal. Chem.* **32** (1960) 1798.
85. SKW Trostberg, DE 3 625 205, 1990 (U. Rust).
86. BAM Bundesanstalt für Materialprüfung, Report on testing cyanamide formulations and expert opinion on transport classification, Log No. II.2-776/03, Berlin May 26, 2004.
87. Y. Shulman, G. Nir, L. Fanberstein, S. Lavee, *Sci. Hortic. (Amsterdam)* **19** (1983) 97 – 104.
88. Südd. Kalkstickstoff-Werke, FR 1 492 380, 1966.
89. DuPont, US 3 850 924, 1974 (J. J. Fuchs, J. B. Wommack).
90. DuPont, US 4 420 325, 1983 (R. F. Sauers).
91. DuPont, US 4 370 480, 1983 (G. Levitt, W. C. Petersen).
92. Ihara, EP 224 339, 1987 (S. Ito, K. Masuda, S. Kusano, T. Nagata, Y. Kojima, N. Sawai, S. Maeno).
93. Ciba-Geigy, EP 3 105 50, 1989 (A. Hubele).
94. ICI, DE 2 265 312, 1977 (A. M. Cole, B. K. Kenneth, J. A. W. Turner).
95. ICI, DE 1 795 835, 1976 (B. K. Snell, R. S. Elias, P. F. H. Freeman).
96. Nippon Soda, WO 91 04 965, 1991 (K. Ishimitsu, J. Suzuki, H. Ohishi, T. Yamada, R. Hatano, N. Takasuka, J. Mitsui).
97. A.P.A. Spa, DE 3 248 115, 1982 (H. Schwienbacher).
98. Dainippon Ink, EP 155 509, 1985 (N. Yoshioka, Y. Okunishi, Y. Miura, Y. Mori, E. Adachi, Y. Kataoka).
99. SKF, DE 2 344 779, 1974; US 3 876 647, 1975 (G. J. Durant et al.).
100. J. Zimmermann, *Nachrichten aus der Chemie* **50** (2002) 1084.
101. D. W. Ludovici et. al., 226th ACS National Meeting Washington D.C., September 2003.
102. Upjohn, DE 2 114 887, 1971 (W. C. Anthony).
103. Janssen, DE 2 029 637, 1971 (J. L. Van Gelder, A. H. Raeymaekers, L. F. Roevens).
104. Goldschmidt AG, DE 10327871, 2005 (M. Pascaly et al.).
105. SKW Trostberg, DE 3 544 398, 1987 (B. Hammer, H. Michaud, S. Weiss).
106. J. G. Kuchler, W. Sturm, H. E. Teichmann, *Papier (Darmstadt)* **47** (1993) 53.
107. *PTS Forschungsbericht* **14/98** (1998) A-C, i, 1–51.
108. O. Kordsachia, R. Patt, W. Sturm, *Wochenblatt für Papierfabrikation* **129** (2001) 941.
109. Wayne Pigment Corp., US 5 176 894, 1993; US 5 487 779, 1996 (J. Sinko).
110. Hoechst, EP 542 214, 1992 (J. Dannheim, W. H. Russ).
111. Amer. Cyanamid, US 4 273 686, 1981 (J. Noland, M. Hajek).
112. Bayer, EP 312 836, 1988 (W. Schäfer, H. P. Müller, R. Dhein).
113. Bayer, EP 641 772, 1995 (H. Täubel, T. Münzmay, T. Hassel, W. Krümmel, H. Ehlert).
114. J. Haag, *Justus Liebigs Ann. Chem.* **122** (1862) 22.
115. L. Kahovec, K. W. Kohlraush, *Z. Phys. Chem. (Leipzig)* **193** (1944) 188 – 195.
116. E. W. Hughes, *J. Am. Chem. Soc.* **62** (1940) 1258 – 1267.
117. F. L. Hirshfeld, H. Hope, *Acta Cryst. B* **36** (1980) 406.
118. B. V. Lotsch, J. Senker, W. Schnick, *Inorganic Chemistry* **43** (2004) 895.
119. B. Juergens et al. *J. Am. Chem. Soc.* **125** (2003) 10288.
120. Amer. Cyanamid Co., US 2 357 261, 1944 (D. W. Kaiser).
121. Du Pont, GB 434 961, 1935.
122. Amer. Cyanamid, US 2 371 112, 1945.
123. J. Thiele, E. Uhfelder, *Justus Liebigs Ann. Chem.* **303** (1898) 108.
124. ICI, GB 577 843, GB 581 346, 1946 (F. H. S. Curd, F. L. Rose).
125. J. H. Paden, K. C. Martin, R. C. Swain, *Ind. Eng. Chem.* **39** (1947) 952.
126. C. Hasegawa, *J. Soc. Chem. Ind. (Jpn.)* **45** (1942) 416.
127. T. C. Ponder, *Hydrocarbon Process.* **48** (1969) no. 11, 200.
128. P. Ostrogovich, *Atti Accad. Lincei* **20** (1911) no. 1, 249; W. Zerwick, W. Brunner (Alien Property Custodian), US 2 302 162, 1942.
129. BASF, DE 4 335 497, 1995 (B. Schäfer, H. Mayer).
130. F. Pohl, *J. Prakt. Chem.* **77** (1908) no. 2, 533.
131. I.G. Farbenind., DE 590 175, 1933 (L. Taub, W. Kropp).
132. Südd. Kalkstickstoff-Werke, DE 1 758 250, 1968.
133. K. Vilsmeyer, *Z. Pflanzenernähr. Bodenk.* **147** (1984) 264.
134. Degussa AG, Method-SOP 004–012/2, 1999 (C. Riefer).

135. Degussa AG, SOP 712.1, 1992 (U. Rust).
136. SKW Trostberg AG, DE 3738675, 1988 (U. Rust).
137. F. Wolf, D. Spiethoff, *Melliand Textilber.* **48** (1967) no. 12, 1456 – 1460.
138. Agrolinz Melamin G.m.b.H, WO 98/02474 (N. Corte, F. Heger).
139. Hercules, US 4 380 603, 1981; US 4 383 077, 1983 (R. A. Bankert).
140. Sandoz, DE 3 216 913, 1982 (J. R. Runyon, S. Valenti).
141. Le Page Glue Co., US 2 282 364, 1942 (W. G. Kunze, R. B. Evans).
Amer. Cyanamid, US 2 581 111, 1951 (C. G. Landers, J. Studeny).
142. J. W. Lyons: *The Chemistry and Uses of Fire Retardants*, J. Wiley & Sons, New York 1970, p. 136.
Amer. Cyanamid, US 3 625 753, 1969 (S. J. O'Brien, R. G. Weyker).
143. Koppers Co., US 3 159 503, 1964 (I. S. Goldstein, W. A. Dreher).
144. Ciba-Geigy, DE 2 942 788, 1980 (P. Rohringer, H. Wegmueller).
145. I. J. Cummings, *Ind. Eng. Chem.* **46** (1954) 1985.
146. Dixon International Ltd., US 4 210 725, 1980 (C. A. Redfard).
147. SNPE Materiaux Energetiques, EP 1496036, 2005 (G. Chounet, B. Thibieroz).
148. LCB, FR 2845378, 2004 (P. Rocquet).
149. Ciba-Geigy, DE 2 017 114, 1970 (K. T. Shen, G. W. Jarvis).
150. R. Schlimper, *Plaste Kautsch.* **10** (1963) 19 – 21.
151. M. D. Gilbert, N. S. Schneider, W. X. Zukas, W. J. McKnight, *Polym. Mater. Sci Eng.* **56** (1987) 351.
152. J. Galy, D. Gulino, J-P. Pascault, *Makromol. Chem.* **188** (1987) 7.
153. A. Pfitzmann, K. Schlothauer, M. Fedtke, *Polym. Bull.* **27** (1991) 59.
154. A. Pfitzmann, M. Fedtke, *Kautschuk Gummi Kunststoffe* **46** (1993) 596.
155. N. Poisson, A. Maazouz, H. Sautereau, M. Taha, X. Gambert, *J. Appl. Polymer Sci.* **69** (1998) 2487.
156. T. Güthner, B. Hammer, *J. Appl. Polym. Sci.* **50** (1993) 1453.
157. H.-S. Kim, W.-Y. Kim, Y.-J. Kim, *Kongop Hwahak* **7** (1996) 963, CA 125:330432.
158. H. Sanftenberg, M. Fedtke, *Angew. Makromol. Chem.* **225** (1995) 99.
159. S.-G. Hong, C.-S. Wu, *Thermochimica Acta* **316** (1998) 167.
160. M. Gundjian, K. C. Cole, *J. Appl. Polymer Sci.* **75** (2000) 1458.
161. M. Vrana, J. G. Dillard, T. C. Ward, M. D. Rakestraw, D. A. Dillard, *J. Adhesion* **55** (1995) 31.
162. SKW Trostberg, DE 2 531 962, 1977 (K. Büchler).
163. H. J. Di, K. C. Cameron, Controlling Nitrogen Flows and Losses, 12th Nitrogen Workshop, Exeter, UK, Sept 21-24, 2003, 431.
164. SKW Stickstoffwreke Piesteritz G.m.b.H., DE 10342551, 2005 (U. Radics et al.).
165. L. M. Estermaier, H. H. Sieber, F. Lottspeich, D. H. M. Matern, G. R. Hartmann, *Angew. Chem.* **104** (1992) 655.
166. Degussa, DE 3 729 530, 1990 (K. Huthmacher, A. Kleemann, H. Bethge, R. Braun).
167. DD 71 768, 1968 (K-C. Röber, G. Rembarz, E. Fischer, R. Ohff, H. Cramer).
168. DuPont, US 4 933 450, 1990 (G. C. Chiang).
169. B. Mertschenk et al. *Zbl. Arbeitsmed.* **41** (1991) 107.
170. B. Mertschenk et al. *Zbl. Arbeitsmed.* **43** (1993) 254.
171. DFG: *Occupational Toxicants. Critical Data Evaluation for MAK Values and Classification of Carcinogens*, vol. **5**, VCH, Weinheim 1993,
172. Central Institute for Nutrition and Food Research TNO, Zeist (The Netherlands), Reports to SKW Trostberg AG, 1976 – 1978.
173. Huntingdon Research Centre Ltd., Huntingdon, Cambridgeshire (England), Reports to SKW Trostberg AG, 1989 – 1992.
174. DFG: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*, Wiley-VCH, Weinheim 1997, 25. Lieferung, Calciumcyanamid.
175. National Cancer Institute, Technical Report Series No. 163, US Department of Health, Education, and Welfare, Bethesda, MD, USA, 1979.
176. DFG: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begruendungen von MAK-Werten*. Wiley-VCH, Weinheim 2002, 35. Lieferung, Cyanamid.
177. DFG: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begruendungen von MAK-Werten*. Wiley-VCH, Weinheim 2003, 37. Lieferung, Dicyandiamid.
178. H. Senff et al. *Derm. Beruf Umwelt* **36** (1988) 99.

179. American Cyanamid Company, company report “Aero Dicyandiamide”, p. 3, 1959.
180. NIOSH, Pocket Guide to Chemical Hazards, January 2003.
181. Technische Regeln für Gefahrstoffe (TRGS), TRGS 900: Grenzwerte in der Luft am Arbeitsplatz “Luftgrenzwerte”, Fassung Oktober 2000, zuletzt geändert im Juli 2004.
182. C. A. Pickles, J. M. Toguri, R. A. Bergmann, *CIM Bulletin* **95** (2002) 81.

Cyanates, Inorganic Salts

PETER M. SCHALKE, INEOS Paraform GmbH & Co KG, Mainz, Germany

1. **Introduction** 1
2. **Properties** 1
3. **Production** 2
4. **Uses** 3
5. **References** 3

1. Introduction

Cyanates are salts of an acid having the empirical formula HNCO, M_r 43.03, which exists in two tautomeric forms [1 – 3]:

H−O−C≡N	⇄	H−N=C=O
Cyanic acid		Isocyanic acid
[*420-05-3*]		[*75-13-8*]

At ambient temperature, isocyanic acid clearly predominates. However, esters of both isomers exist, e.g., ethyl cyanate, C_2H_5OCN, and ethyl isocyanate, C_2H_5NCO. Isocyanic acid can be represented by three mesomeric structures:

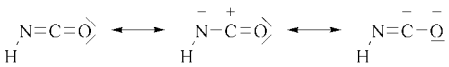

Likewise, the anion of the inorganic salts is a resonance hybrid and may be derived from either tautomeric form of the acid.

Cyanate anion ↔ Isocyanate anion

Only sodium and potassium cyanate are manufactured commercially. There is no commercial interest in any other cyanate salt. Free cyanic acid HOCN is unstable and cannot be isolated.

2. Properties

In aqueous solution, sodium and potassium cyanate undergo hydrolysis according to Equations (1) and (2) [8].

$$OCN^- + 2\,H_2O \rightarrow CO_3^{2-} + NH_4^+ \quad (1)$$

$$OCN^- + NH_4^+ \rightarrow CO(NH_2)_2 \quad (2)$$

The rate of hydrolysis depends strongly on temperature, pH and concentration. In dilute strong acids, the reaction proceeds according to Equation (1). The reaction is rapid and complete within a short time. It is therefore used in the chemical analysis of cyanates. In concentrated hydrochloric acid, the trimerization product cyanuric acid can be observed as a byproduct. In alkaline solution, the rate of hydrolysis is very slow. Some cyanate decomposes with the formation of urea according to Equation (2).

The formation of cyanate ions is the first step in the detoxification of cyanide ions in water using hydrogen peroxide [9], followed by the further decomposition of the cyanate according to Equation (1) or (2).

In the dry state, cyanate salts decompose above 450 °C to form cyanides. The decomposition is influenced by certain metals. Both salts are hygroscopic, potassium cyanate much more so than sodium cyanate. When cyanate salts are heated to around 300 °C in moist air, they undergo hydrolysis to the corresponding carbonate with liberation of ammonia.

Sodium and potassium cyanate dissolve only in water. Solubility in organic solvents is very low. Some physical properties are listed in Table 1.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_157.pub2

Table 1. Physical properties of sodium and potassium cyanate

	Sodium cyanate	Potassium cyanate
CAS registry number	[*917-61-3*]	[*590-28-3*]
M_r	65.01	81.12
mp, °C	550	315
ϱ, g/cm^3	1.893	2.056
Solubility (g/100 g solvent)		
in water	10.68 at 16 °C	75 at 25 °C
in ethanol	0.5 at *bp*	0.53 at *bp*
in benzene	0.13 at *bp*	0.18 at *bp*
in liquid ammonia	1.72 at − 19.8 °C 0.72 at 45 °C	1.70 at 25 °C

Both potassium and sodium cyanate are nonflammable. Strong irritation occurs when the dust is inhaled or gets into the eyes. According to transportation regulations, cyanates are not classified as dangerous goods.

3. Production

Sodium Cyanate, Technical-Grade. In production, sodium carbonate is used as sodium source [4]. Sodium hydroxide [5] is only useful for laboratory synthesis. Under production conditions, safety considerations make sodium hydroxide unpractical. Urea is the preferred source for the cyanate part of the molecule, giving rise to ammonium carbonate as byproduct. Sodium allophanate has been postulated as a possible reaction intermediate [6]. The appearance and disappearance of such an intermediate can be observed by HPLC analysis. The reaction equation is therefore postulated as follows (Eqs. 3 – 5).

$$H_2NCONH_2 \rightarrow NH_4^+OCN^- \quad (3)$$

$$Na_2CO_3 + NH_4^+OCN^- + H_2NCONH_2 \rightarrow NaOCN + NH_2CONHCOONa^+ + NH_3 + H_2O \quad (4)$$

$$NH_2CONHCOONa \rightarrow NaOCN + NH_3 + CO_2 \quad (5)$$

The overall reaction is the formation of two moles of sodium cyanate and one mole of ammonium carbonate from one mole of sodium carbonate and two moles of urea (Eq. 6).

$$Na_2CO_3 + 2\,H_2NCONH_2 \rightarrow 2\,NaOCN + (NH_4)_2CO_3 \quad (6)$$

The reaction is endothermic and requires strong heating above 200 °C for a prolonged period of time to reach completion. The reaction takes place in a urea melt with a high content of dispersed solids.

The product from this process has a purity of 89 – 92 %. The main impurity is excess sodium carbonate with small amounts of sodium allophanate, cyanuric acid, biuret, and urea.

Sodium Cyanate, High-Grade. Sodium cyanate with at least 95 % purity can be obtained by using technical-grade sodium cyanate and urea as starting materials. The process proceeds in the molten state, as described for potassium cyanate [7]. Because of the high melting point of the sodium cyanate, the reaction temperature must be raised to 550 – 600 °C. At these high temperatures, some cyanate decomposes to sodium cyanide, which can lead to formation of undesired byproducts and requires treatment of the wastewater. Also, the byproduct ammonia may ignite (flash point 600 °C). For these reasons, production is not without risk.

Potassium Cyanate, High-Grade. The production of potassium cyanate from potassium carbonate and urea requires reaction temperatures of 400 °C or above. The product is obtained from the reactor as a liquid, which must be solidified by cooling and ground after solidification. Since the crystalline structure of the potassium carbonate is destroyed in the melting process, the urea can reach nearly all potassium ions and convert them to potassium cyanate at a much higher rate than in the case of the technical-grade sodium salt. It is therefore easy to reach high purities above 95 %. While possible impurities like urea, biuret, cyanuric acid, and potassium allophanate are unstable at the reaction temperature and therefore are easily eliminated, the temperature is still too low to allow for significant formation of potassium cyanide.

Treatment of Byproducts. To facilitate isolation of the ammonium carbonate byproduct during the reaction, a continuous process for cyanate production is preferred to a batch process. If undesired, the ammonia in the off-gases can be burned. It is more economic to trap the ammonia by washing the gases with cold water or acid and use the resulting aqueous solution otherwise, for example, reaction with formaldehyde to produce hexamethylenetetramine.

4. Uses

A review of the chemical reactions involving cyanate salts has been published [10].

Sodium cyanate is used as such in steel hardening; potassium cyanate is used as herbicide in the onion crop. All other applications require chemical synthesis, for which sodium cyanate is often the salt of choice. In most reactions, it can be regarded as a safe alternative to phosgene chemistry. The following reactions are or have been of commercial value.

Reactions with primary or secondary amines afford monosubstituted ureas [11] (Eq. 7).

$$RNH_2 \cdot HCl + MOCN \rightarrow RNHCONH_2 + MCl \quad (7)$$

where, e.g., R = ethyl, *tert*-butyl, cyclohexyl, phenyl

Sodium and potassium cyanate behave alike. Potassium cyanate is sometimes preferred to sodium cyanate because of its higher purity and its much higher solubility in water, a popular solvent for this type of reaction. Another solvent used for this reaction is ethyl acetate. Well-known pharmaceuticals produced by this type of reaction are the antiepileptic carbamazepine and the dermatotherapeutic agent hydroxyurea. The fungicides cymoxanil, lenacil and terbacil are also produced by this type of reaction, as well as the herbicides isoproturon and diuron. The reaction of aniline and sodium cyanate yields phenylurea, an intermediate for the manufacture of certain dyestuffs.

In alkylation reactions, only sodium cyanate gives satisfactory yields (Eq. 8); potassium cyanate cannot be used.

$$RX + NaOCN \rightarrow RNCO + NaX \quad (8)$$

Suitable alkylating agents are alkyl halides or sulfates. For example, allyl isocyanate is obtained from the reaction of sodium cyanate with allyl chloride. The product trimerizes readily to form triallyl isocyanurate [12]. The reaction of dimethyl sulfate with sodium cyanate results in the formation of methyl isocyanate, an intermediate for a number of insecticides [13] mainly used for rice and sugar cane crops.

Aromatic acid chlorides can react with sodium cyanate to yield benzoylurea derivatives, a novel class of herbizides [14]. Similarly, some sulfonylureas are antidiabetic agents, and sodium cyanate can be used for their synthesis. The reaction of sodium cyanate with sulfonyl chloride gives chlorosulfonylisocyanate in situ, which can be treated further in a one-pot synthesis (Eq. 9) [15]. To obtain high yields, a sodium cyanate with a low sodium carbonate content must be used.

$$NaOCN + SO_2Cl_2 \rightarrow ClSO_2NCO + NaCl \quad (9)$$

Sodium cyanate is also used as a building block in the multistep synthesis of the herbicides carfentrazone and sulfentrazone [16].

Unique applications for potassium cyanate are rare. The major application of potassium cyanate is as an ingredient in certain steel-hardening salts for the Tufftride/Tenifer process [17].

Potassium cyanate is also used in the synthesis of Carbidopa, a drug against Parkinson's disease. The reaction of amidosulfuric acid with potassium cyanate yields potassium hydroxyureasulfonate, which has found application as a polymerization catalyst for acrylonitrile [18].

The worldwide demand for cyanate salts is in the order of 8000–10 000 t/a. Production plants exist in Japan, South Korea, India, China, and Germany. Although there are some new applications in the agrochemical field, most agrochemicals which use cyanate salts for their synthesis are being slowly replaced in Europe by other agents or losing importance because of the increased use of genetically modified crops. In India and China, however, where there is a considerable increase in chemical production for agriculture, medicine, and steelmaking, new capacities have been installed.

5. References

1. J. Goubeau, *Ber. Dtsch. Chem. Ges.* **68** (1935) 912 – 919.
2. J. C. Kauer, W. W. Henderson, *J. Am. Chem. Soc.* **86** (1964) 4723.
3. *Gmelin*, **Syst. No. 14 "carbon", part DI**, (1971) p. 327.
4. DuPont, GB 1 145 777, 1969.
   Sakae Food Suff Industry Co., DE 1 205 065, 1966 (I. Tomiro).
   Degussa, GB 339 220, 1930.
   DuPont/Degussa, US 1 915 425, 1933 (H. Kloepfer).
   FMC, EP 0 122 031 B1, 1987 (W. B. Dodge, M. Halfon).

5. Asahi Chemical Industry Co. Ltd., JP 49 042 800, 1966 (Y. Nakayama et al.).
Stamicarbon B. V. Neth., DE-OS 2 431 205, 1975 (J. Verstegen).
6. VEB Agrochemie Piesteritz, DD 221 449, 1985 (W. Schunk et al.).
7. US 2 690 957, 1954 (W. P. Horst).
8. J. A. Kemp, G. Kohnstam, *J. Chem. Soc.* **1956** no. 11, 900.
9. Degussa, DE-OS 2 352 856, 1976 (J. Fischer, H. Knorre).
Degussa, DRP 742 074, 1943 (H. Beier).
H. Knorre, *Galvanotechnik* **66** (1975) no. 5, 374 – 383.
10. *Houben-Weyl*, 4th ed., vol. **VIII**, pp. 89, 125;
Ethyl Corp., US 2 866 801, 1958 (C. H. Himel, L. Richards).
H. Böhme, W. Pasche, *Arch. Pharm.* **302** (1969) 335.
K. A. Jensen, M. Due, A. Holm, *Acta Chem. Synd.* **19** (1965) 438.
11. DuPont, US 3 235 357, 1966 (H. Loux).
DuPont, US 3 235 360, 1966 (E. Soboczenski).
DuPont, US 3 352 862, 1965 (E. Soboczenski).
12. Nippon Kasei Chemical Co., DE-OS 2 839 084, 1979 (T. Nakamuro et al.).
13. Degussa, US 4 206 136, 1980 (G. Gieselmann et al.).
Degussa, DE 2 828 259 C2, 1980 (G. Gieselmann, K. Günther).
14. Ciba-Geigy Res. Discl. 351, 442–7;
Chem Abstr. 119 : 271113.
15. Hoechst Agrevo, EP 560 178, 1996 (G. Schlegel);
Hoechst Agrevo, EP 507 093, 1996 (G. Schlegel et al.).
16. FMC, US 5 256 793, 1993 (A. R. Bailey, M. Halfon, E. W. Sortore).
17. Daimler-Benz, DE-OS 2 602 754, 1976 (C. Lovasz et al.),
Degussa, DE 1 191 655 OT, 1965 (J. Müller).
Degussa, DE 1 149 035 OT, 1963 (J. Müller, C. Albrecht).
Degussa, DE 1 234 873 OT, 1967 (J. Müller).
Degussa, DE 1 280 018 OT, 1968 (J. Müller).
18. Asahi Kasei Kogyo, DE 1 720 202, 1973 (T. Ohfuka, K. Shirode, Y. Ichikawa).

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Cyano Compounds, Inorganic

Standard Article

Ernst Gail[1], Stephen Gos[2], Rupprecht Kulzer[3], Jürgen Lorösch[4], Andreas Rubo[5], Manfred Sauer[6]

[1]Degussa AG, Düsseldorf, Federal Republic of Germany
[2]CyPlus GmbH, Hanau, Federal Republic of Germany
[3]Degussa AG, Düsseldorf, Federal Republic of Germany
[4]Degussa AG, Düsseldorf, Federal Republic of Germany
[5]CyPlus GmbH, Hanau, Federal Republic of Germany
[6]Degussa AG, Düsseldorf, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a08_159.pub2
Article Online Posting Date: January 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Hydrogen Cyanide
1.1. Properties
1.2. Production
1.2.1. Andrussow Process
1.2.2. Methane - Ammonia (BMA) Process
1.2.3. Shawinigan Process
1.3. Storage and Transportation
1.4. Economic Aspects and Uses
2. Metal Cyanides
2.1. Alkali Metal Cyanides
2.1.1. Properties
2.1.2. Production
2.1.3. Commercial Forms, Specifications, and Packaging
2.1.4. Transportation, Storage, and Handling
2.1.5. Economic Aspects and Uses
2.2. Alkaline Earth Metal Cyanides
2.3. Heavy-Metal Cyanides
2.3.1. Iron Cyanides
2.3.1.1. Properties
2.3.1.2. Production
2.3.1.3. Commercial Forms, Specifications, and Packaging
2.3.2. Cyanides of Copper, Zinc, and Cadmium
2.3.3. Cyanides of Mercury, Lead, Cobalt, and Nickel
2.3.4. Cyanides of Precious Metals
2.4. Cyanide Analysis
3. Detoxification of Cyanide-Containing Wastes
3.1. Wastewater Treatment
3.2. Solid Wastes
4. Cyanogen Halides
4.1. Properties
4.2. Production
4.3. Storage and Transportation
4.4. Uses
5. Cyanogen

5.1. Properties
5.2. Production
5.3. Storage, Transportation, and Uses
6. Toxicology and Occupational Health
7. Responsible Care and Product Stewardship

[Top of Page]

1. Hydrogen Cyanide

Hydrogen cyanide [74-90-8], hydrocyanic acid, prussic acid, formonitrile, HCN, M_r 27.03, is a colorless liquid with the characteristic odor of bitter almonds. Hydrogen cyanide in aqueous solution was first prepared by SCHEELE in 1782 [18]. The acid occurs naturally in combination with some glucosides, such as amygdalin.

Hydrogen cyanide is currently produced by direct reaction of alkanes with ammonia, and indirectly as a byproduct of the manufacture of acrylonitrile (by ammoxidation of propene ⟶ Acrylonitrile). Major end uses include the production of adiponitrile, methyl methacrylate, cyanuric chloride, chelating agents, sodium cyanide, and methionine and its hydroxy analogues.

1.1. Properties

Physical properties [1], [19-22]:

Melting point	–13.24 °C
Boiling point	25.70 °C
Vapor pressure	
at 0 °C	35 kPa
at 20 °C	83 kPa
at 50 °C	250 kPa
Enthalpy of formation (HCN liquid), ΔH	3910 kJ/kg
Critical temperature	183.5 °C
Critical density	0.20 g/cm^3
Critical pressure	5 MPa (50 bar)
Density (20 °C)	0.687 g/cm^3
Specific heat, liquid (20 °C)	2630 J kg^{-1} K^{-1}
Specific heat, gas (25 °C)	1630 J kg^{-1} K^{-1}
Heat of fusion	310 kJ/kg
Heat of vaporization	935 kJ/kg
Heat of polymerization	–1580 kJ/kg
Explosive range in air	5.5 – 46.5 vol %
Flash point	–17.8 °C
Ignition temperature	535 °C
Dynamic viscosity, η (20 °C)	0.192 mPa · s
Surface tension (20 °C)	18.33 mN/m
Dielectric constant	
at 0 °C	158.1
at 20 °C	114.9
Dissociation constant	
aqueous solution, pK (20 °C)	9.36

As a result of its high dielectric constant, HCN has acquired some importance in preparative chemistry as a nonaqueous, ionizing solvent [23].

Chemical Properties. Some comprehensive reviews on hydrogen cyanide and cyanogen compounds have been published [1-3]. The acid occurs only in the nitrile form. Although isomeric isonitrile HNC has been detected in interstellar space, all efforts to isolate this compound have failed. As the nitrile of formic acid, HCN undergoes many typical nitrile reactions. For example, hydrogen cyanide can be hydrolyzed to formic acid by aqueous sulfuric acid or hydrogenated to methylamine. Hydrogen cyanide adds to carbon – carbon double bonds and forms cyanohydrins with carbonyl groups of aldehydes or ketones. The most important uses of this type are in the manufacture of acetone cyanohydrin (an intermediate in the production of methyl methacrylate) and in the production of adiponitrile from butadiene and hydrogen cyanide. Another example is the multistep synthesis of amino acids via hydantoins. Hydrogen cyanide can be oxidized by air over silver or gold catalysts at 300 – 650 °C to yield cyanic acid (HOCN) and cyanogen $(CN)_2$ in an approximate 2 / 1 ratio.

The reaction of hydrogen cyanide with chlorine gives cyanogen chloride (see Section Production). For industrial purposes, the latter compound is usually directly trimerized to cyanuric chloride, the starting material for the chemistry of *s*-triazines.

In the presence of oxygen or air, hydrogen cyanide burns with a very hot flame. For the reaction

$$2\,HCN + 1.5\,O_2 \rightarrow N_2 + 2\,CO + H_2O$$

the heat of formation is calculated to be –723.8 kJ/mol and the adiabatic flame temperature is 2780 °C. Pure liquid or gaseous HCN is inert to most metals and alloys such as aluminum, copper, silver, zinc, and brass. At higher temperatures (> 600 °C), the acid reacts with metals that can form carbides and nitrides (titanium, zirconium, molybdenum, and tungsten).

Pure liquid hydrogen cyanide has a tendency to polymerize to brown-black, amorphous polymers, commonly called azulmic acid [26746-21-4]. The reaction is accelerated by basic conditions, higher temperature, UV light, and the presence of radicals. Since the decomposition is exothermic, the polymerization reaction is autocatalytic and can proceed with explosive violence to form the HCN dimer iminoacetonitrile [1726-32-5] and the HCN tetramer diaminomaleonitrile [1187-42-4] as intermediates [24], [25]. Both compounds are presumed to be important in the evolution of life, and cyano compounds may play a role in prebiotic syntheses [26], [27]. In the liquid phase, hydrogen cyanide is stabilized by the presence of small amounts of acids (0.1 wt % H_3PO_4, 1 – 5 % HCOOH or CH_3COOH) or 0.2 wt % SO_2 in the gas phase.

The HCN – H_2O System. At 25 °C, hydrogen cyanide is miscible with water in all ratios. The solution is a weak acid, with a dissociation constant of the same order of magnitude as amino acids. The relation between total cyanide concentration ($c_{HCN} + c_{CN^-}$) and dissociated cyanide (c_{CN^-}) in a dilute aqueous solution, as a function of pH, is illustrated in Figure 1.

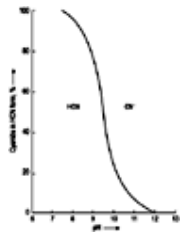

Figure 1. HCN/CN^- equilibrium as a function of pH

Figure 2 shows the liquid – vapor equilibrium diagram at atmospheric pressure. Because of the high HCN partial vapor pressure it is difficult to separate the acid from a gas mixture by absorption in water, which is an important consideration for industrial practice [28].

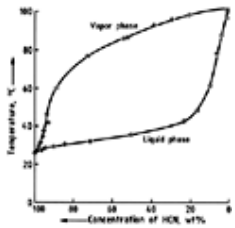

Figure 2. Liquid – vapor diagram of the system H_2O – HCN at atmospheric pressure

The stability of hydrogen cyanide solutions depends on the degree of dilution: at concentrations below 0.1 mol/L HCN, the acid is stable; addition of traces of acid prevents decomposition. Like pure HCN, the exothermic polymerization of an aqueous solution is accelerated by the presence of alkali. In addition to azulmic derivatives, small amounts of amino acids and purine bases are formed. This fact is of some importance to biological chemistry. The C≡N triple bond is hydrolyzed by strong alkali or acid to give formic acid and ammonia. Higher temperatures or hydrothermal conditions favor these reactions.

1.2. Production [1], [29], [30]

Hydrogen cyanide can be produced when sufficient energy is supplied to any system containing the elements hydrogen, nitrogen, and carbon. Generally, only processes starting from hydrocarbons and ammonia are of economic importance today; however, the production of hydrogen cyanide from formamide [31] is carried out in a plant at Ludwigshafen, Germany and since 1997 in a plant in Niigata, Japan. Alternative raw materials such as methanol, carbon, or carbon monoxide [32-41] are of interest since the 1980s. The ammoxidation of methanol to HCN is performed in Japan in a 10 000 – 15 000 t/a plant.

These routes can compete with conventional technology given specific site conditions. The reaction of hydrocarbons with ammonia

$$C_xH_{2x+2} + xNH_3 \rightarrow xHCN + 2\cdot(2x+1)H_2$$

is highly endothermic and needs a continuous heat supply. The means of providing this energy requirement are manifold and characteristic of the different processes [43]. Direct microwave heating of the catalyst has been developed for small-scale local synthesis of HCN [44]. The main processes currently used to make hydrogen cyanide: the Andrussow ammoxidation process, which involves the reaction of ammonia, methane, and air over a catalyst gauze; and the two ammonia dehydration routes, the methane – ammonia (BMA) and the Shawinigan processes, developed by Degussa and Gulf Oil Company, respectively. The latter are performed in externally heated ceramic tubes or in an electrically heated fluidized coke bed. Twenty-five percent of the hydrogen cyanide in the United States and 20 % in Western Europe is obtained as a byproduct in

the manufacture of acrylonitrile by the oxidation of propene in the presence of ammonia (Sohio technology).

1.2.1. *Andrussow Process*

The Andrussow process was developed around 1930 by L. ANDRUSSOW of I.G. Farben [45-47] and is the most widely used method for direct synthesis of hydrogen cyanide. The average capacity of single commercial installations is (5 – 30) × 10^3 t/a; large plants are operated by DuPont, Rohm and Haas, Novartis and Cyanco in the United States; ICI, Butachemie, Röhm and Elf-Atochem in Western Europe; and Mitsubishi Gas in Japan.

Figure 3 is a flow diagram of the Andrussow process. Natural gas, essentially sulfur-free methane, is mixed with ammonia. Compressed air is added in a volume ratio that corresponds closely to the theoretical reaction [48]:

$$CH_4 + NH_3 + 1.5\,O_2 \rightarrow HCN + 3\,H_2O \quad \Delta H = -474\ \mathrm{kJ/mol}$$

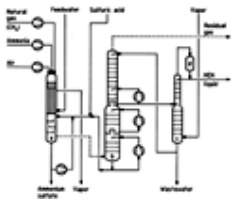

Figure 3. Simplified diagram of the Andrussow process

a) Reactor and ammonia scrubber; b) HCN absorption tower; c) HCN rectifier; d) Condenser

The mixture is passed over a platinum – rhodium or platinum – iridium gauze [49] catalyst; temperature and upper flammability limit should be monitored carefully [50], [51]. The reaction takes place at > 1000 °C, at around atmospheric pressure, and with a gas velocity through the catalyst zone of about 3 m/s. To avoid decomposition of HCN, the effluent gas from the reactor is quickly cooled in a waste-heat boiler, which produces the steam used in the process.

After the waste-heat boiler, the gas is washed with dilute sulfuric acid to remove unreacted ammonia; this is necessary to prevent polymerization of HCN. Because disposal of the resulting ammonium sulfate solution is expensive, other systems have been patented [52-54]. Alternatively, the off-gas from the reactor is passed through a monoammonium phosphate solution [55], [56], which converts the ammonia to diammonium phosphate. To effect thermal reversal of the phosphate equilibrium, the absorption solution is boiled in a stripper by injection of steam. The released ammonia is condensed and recycled to the reactor, while the regenerated monoammonium phosphate solution is pumped back to the absorber.

After the ammonia scrubber, the gas is passed through a countercurrent column in which the hydrogen cyanide is absorbed in cold water and the resulting solution is stabilized by adding acid (ca. 0.1 %). The hydrogen cyanide is stripped from the aqueous solution in a rectifier and condensed. The end product is highly pure and has a water content of less than 0.5 %. The aqueous absorber solution, containing traces of HCN, is cooled and fed back to the absorption tower. The residual gases, H_2, CO, and N_2, can be used for heating or methanated in a separate unit and recycled as feedstock for HCN manufacture [57].

The advantages of the Andrussow process include (1) long catalyst life, up to 10 000 h; (2) well-tested technology, with a simple and safe reaction system; and (3) high-purity HCN. A disadvantage is that the process is dependent on pure methane as a raw material to avoid carburization of the platinum catalyst; for example, a small percentage of higher hydrocarbon impurities rapidly causes problems with the catalyst system and reduces conversion rates. Other important poisons for platinum are sulfur and phosphorus compounds. A further problem is the relatively low yield based on methane (60 – 70 %) and ammonia (70 %), as well as the low hydrogen cyanide concentration in the product gas, so that recovery equipment must handle large volumes of gas.

The compositions of reaction and residual gases in the Andrussow and BMA processes are compared in Table 1.

Table 1. Composition of off-gases and residual gases (volume fractions, %) on the basis of pure methane and ammonia in the BMA and Andrussow processes

Compound	BMA		Andrussow *	
	After reaction	Residual	After reaction	Residual
HCN	22.9	$<10^{-2}$	7.6	$<10^{-2}$
NH_3	2.5	$<10^{-2}$	2.3	> 0
H_2	71.8	96.2	13.3	14.7
N_2	1.1	1.5	49.2	54.6
CH_4	1.7	2.3	0.3	0.4
CO			3.8	4.2
H_2O			23.1	25.7
CO_2			0.4	0.4

* Calculated because no consumption figures are available.

1.2.2. *Methane – Ammonia (BMA) Process*

The basis of the Degussa BMA process is the formation of hydrogen cyanide in the absence of oxygen [58-62]. The reaction

$$CH_4 + NH_3 \rightarrow HCN + 3\,H_2\,\Delta H = +252\,kJ/mol$$

is endothermic, requires temperatures above 1200 °C, and is performed in externally heated, alumina tube bundles, which are coated with a thin layer of a special platinum catalyst [63], [64]. Several of these bundles are fixed in a reaction furnace unit. A mixture of ammonia and methane (natural or refined gas with a content of 50 – 100 vol % methane) is passed through the tubes and quickly heated to 1300 °C at normal pressure. To avoid the formation of any disturbing deposits of carbon black, the NH_3/CH_4 ratio is kept between 1.01 and 1.08. After leaving the reaction tubes, the product gas is cooled to 300 °C by passage through a water-cooled aluminum chamber. A kinetic study has shown that a particular temperature profile is essential for this process [65].

The subsequent reaction steps, ammonia absorption and hydrogen cyanide isolation, are similar to those of the Andrussow process. A distinct advantage is the higher HCN content (Table 1) of the product gas, so that the number of steps and the size and cost of recovery equipment, are greatly reduced. The tail gas consists mainly of pure hydrogen. If this is not needed for other syntheses, it can be used as fuel for heating the furnace. About 80 – 87 % of the ammonia and 90 – 94 % of the methane are converted to hydrogen cyanide. The specific energy consumption of ca. 4×10^6 kJ/100 kg HCN reported thus far has been considerably decreased by more recent developments [66], [67]. A large part of the heating energy is recovered and used in the air preheater or steam generator.

If the methane supply is limited, the process can be carried out directly with liquefied hydrocarbons or ethanol, or in a three-step reaction starting from methanol [68-70]. Both Degussa and Lonza utilize the BMA route to produce hydrogen cyanide.

1.2.3. *Shawinigan Process*

In the Shawinigan process, hydrocarbon gases are reacted with ammonia in an electrically heated, fluidized bed of coke. The process, sometimes called the Fluohmic process, was developed in 1960 by Shawinigan Chemicals [71-74], now a division of Gulf Oil Canada.

In a circular reaction cavity constructed from alumina and silicon carbide, the mixture of ammonia and hydrocarbon (N/C ratio slightly > 1) passes through a fluidized bed of coke, heated by electrodes immersed in the bed. The chemical reaction is similar to the methane – ammonia process, but no catalyst is required and temperatures are kept above 1500 °C. Other carbon compounds, such as naphtha or lighter hydrocarbons, can also be converted. Propane is usually the main feedstock. The reaction can be described as

$$3\,NH_3 + C_3H_8 \rightarrow 3\,HCN + 7\,H_2\,\Delta H = +634\,kJ/mol$$

Unreacted feed gas is almost completely decomposed to the elements. This reduces the quantity of ammonia to be removed from the product gas and leads to the formation of coke particles. The control of coke particle size in the bed is an important operating parameter.

The reactor effluent contains up to 25 vol % HCN, 72 vol % H_2, 3 vol % N_2, and only 0.005 vol % NH_3. Coke is removed in a water-cooled, cyclone-entrained bed. The gas is further cooled and enters the absorption equipment where HCN is removed. The residual gas, nearly pure hydrogen, can be used for other chemical processes. Some of the hydrogen is recycled to the reaction unit to inhibit the formation of soot.

Coke from the cyclone is screened, and three fractions are separated, stored, and then fed back to the reactor system in the desired proportions to control the particle size distribution. By regulating the rate of coke recycling, the level of the fluidized bed and the reaction temperature can be controlled.

In practice, at least 85 % of the ammonia and up to 94 % of the hydrocarbon are converted to hydrogen cyanide. Because of the high electric power consumption (6.5 kWh per kilogram of HCN) the Shawinigan process would probably only be attractive where low-cost electricity is available.

This process is employed by Polifin in South Africa, by Arogenesas in Spain, and by Ticor in Australia.

1.3. Storage and Transportation [75], [76]

Handling, storage, and transportation of hydrogen cyanide are determined by its low boiling point, high toxicity, and instability in the presence of moisture, bases, or other impurities. The liquid acid is relatively uncorrosive. Materials compatible with HCN at normal temperatures are stainless steel, Hastelloy, and Monel. To prevent polymerization, stabilizing agents, such as sulfuric acid, phosphoric acid, oxalic or acetic acid, and sulfur dioxide are used. The type and quantity of stabilizer (usually < 0.5 %) depend on storage capacity, temperature, and residence time in a container. A combination of H_2SO_4 and SO_2 prevents the decomposition of HCN in the liquid and vapor phases. Larger quantities of hydrogen cyanide are stored at a maximum temperature of 5 °C and must be permanently recirculated.

Additionally, the color of the liquid is monitored and should not exceed APHA 20. To keep the concentration of gas below the danger level, good ventilation of buildings in which HCN is stored and handled is of primary importance. Hydrogen cyanide is usually classified by governmental authorities as very poisonous, requiring special packaging and transportation regulations. Similar strict procedures exist for solutions with an HCN content of 5 % or more. Smaller quantities of the stabilized acid are

transported in metal cylinders of up to 56 kg nominal water capacity in the United States and up to 60 kg in Germany. Cylinders cannot be charged with more than 0.55 – 0.60 kg of liquid HCN per 1 L bottle, and resistance to deformation must be tested up to 10 MPa before first filling. The water content should not exceed 3 %, and storage time should be less than one year. According to national regulations, transportation of quantities > 100 kg requires special permission. Procedures covering details of tank car size (up to 50 t), shipping, loading, and handling must be obeyed.

Hydrogen cyanide is mainly shipped by rail under the provisions of the applicable transport regulations. Transport in railcars is permitted only by approval of the competent authorities. The transport classifications of the most common transport regulations are as follows:

RID/ADR: Class 6.1, 1°; (a).

CFR 49: Class 6.1, UN 1051, PG I, “Poison Inhalation Hazard, Hazard Zone A”.

1.4. Economic Aspects and Uses

In 1998 the production capacity of synthetic hydrogen cyanide was about 650 × 10^3 t/a in the United States, 400 × 10^3 t/a in Europe, and 35 × 10^3 t/a in Japan. Without detailed knowledge of the type and age of catalyst used in various acrylonitrile plants, the additional amount of byproduct HCN can only be estimated. With an estimated HCN formation in the acrylonitrile processes of 10 %, an additional 180 × 10^3 t/a (United States), 130 × 10^3 t/a (Europe), and 60 × 10^3 t/a (Japan) is available. Currently, between 60 and 80 % of the existing production capacity is being utilized. The variety of uses has changed enormously in the past 15 years. Acrylonitrile production starting from acetylene was formerly the major consumer of hydrogen cyanide; in contrast, the Sohio process now in use yields HCN as a byproduct.

Today, the main outlet for HCN is the manufacture of methyl methacrylate. Acetone is treated with hydrogen cyanide to produce acetone cyanohydrin, which is treated with sulfuric acid and methanol to form methyl methacrylate. HCN is used to a similar extent in adiponitrile production, in which 2 mol HCN are added to 1 mol butadiene in a two-stage process. Virtually all the adiponitrile is then converted to hexamethylenediamine, a nylon precursor. Hydrogen cyanide reacts with chlorine to form cyanogen chloride (see Section Production) which is usually directly trimerized to cyanuric chloride. Herbicides based on cyanuric chloride have been successfully employed in recent years.

D,L-Methionine, one of the largest volume amino acids produced commercially, is manufactured in a complex, multistep synthesis which employs hydrogen cyanide as one of the raw materials. Other uses of HCN are in the production of chelating agents such as ethylenediaminetetraacetic acid (EDTA) or nitrilotriacetic acid (NTA), starting from aldehydes, amines, and hydrogen cyanide. A wide variety of organic intermediates can also be made from HCN and have considerable potential utility. The most important derivatives of hydrogen cyanide in inorganic chemistry are alkali metal cyanides and cyanide complexes of iron.

[Top of Page]

2. Metal Cyanides

Metal cyanides are compounds of metals in which one or more cyanide ions act as monodentate ligands with carbon as the preferred donor atom. Simple metal cyanides are represented by the formula M $(CN)_x$. Depending on the metal, simple cyanides dissolve more or less readily in water, forming metal ions and cyanide ions:

$$M(CN)_x \rightleftharpoons M^{x+} + xCN^-$$

The solubility is influenced by pH and temperature, especially because hydrogen cyanide is formed by the hydrolysis of cyanide:

$$CN^- + H_2O \rightleftharpoons HCN + OH^-$$

The cyanide complexes can generally be described by the formula A_y [M $(CN)_x$], where A is an alkali, alkaline earth, or heavy metal, and M is normally a transition metal. Most cyanide complexes in which A is an alkali or alkaline earth metal are highly soluble in water and form alkali or alkaline earth metal ions and complex transition metal cyanide anions, e.g.,

$$A_y[M(CN)_x] \rightleftharpoons yA^+ + [M(CN)_x]^{y-}$$

The complex cyanide anion may then undergo further dissociation and release cyanide ions:

$$[M(CN)_x]^{y-} \rightleftharpoons zCN^- + [M(CN)_{x-z}]^{(y-z)-}$$

The dissociation of the cyanide complex depends strongly on the type and valency of the complexed metal ion, as well as on the pH and concentration of the solution. Insoluble or slightly soluble cyanide complexes are formed when the alkali or alkaline earth metal anion A is replaced by a heavy metal ion. In this case, more complicated cyanide complexes are often formed, which cannot be described by the above-mentioned formula. Furthermore, most heavy metal ions allow the formation of mixed cyanide complexes in which one or more of the cyanide groups can be replaced by other ligands such as halides, pseudohalides, nitrogen oxides, sulfur compounds, water, or ammonia [4]. The properties and toxicities of these compounds depend on the structures of the metal cyanides.

2.1. Alkali Metal Cyanides

Since the introduction of the Castner process, sodium cyanide [143-33-9] has been the metal cyanide with the greatest commercial importance. Until about 1900, potassium cyanide [151-50-8] was the more common compound, because it could

be produced more easily by melting potassium carbonate with potassium hexacyanoferrate(II), one of the oldest known cyano compounds. The properties and reactions of both alkali cyanides are very similar.

2.1.1. *Properties*

Sodium and potassium cyanide are colorless, hygroscopic salts with a slight odor of hydrogen cyanide and ammonia in moist air. They are fairly soluble in water (Fig. 4), and the sodium salt forms two hydrates: NaCN · 2 H_2O below 35 °C and NaCN · 1/2 H_2O at higher temperatures. Further physical properties are listed in Table 2. In the absence of air, carbon dioxide, and moisture, the alkali metal cyanides are stable, even at fairly high temperature, and can be stored indefinitely [2].

Table 2. Physical properties of NaCN [143-33-9] and KCN [151-50-8]

Physical quantity	NaCN	KCN
M_r	49.015	65.119
mp	561.7 °C (98 wt %)	634.5 °C
bp	1500 ± 10 °C	
ϱ, density	1.620 g/cm^3 (6 °C, rhombic)	1.553 g/cm^3 (20 °C, cubic)
	1.595 g/cm^3 (20 °C, cubic)	1.56 g/cm^3 (25 °C, cubic)
	1.19 g/cm^3 (850 °C, fused)	
c_p, specific heat capacity	1.667 kJ kg^{-1} K^{-1} (273.1 K)	1.00 kJ kg^{-1} K^{-1} (25 – 72 °C)
	17.950 kJ kg^{-1} K^{-1} (288.0 K)	
	1.402 kJ kg^{-1} K^{-1} (298.6 K)	
H, enthalpy (25 °C)	–89.9 kJ/mol	–112.63 kJ/mol
ΔH_f, heat of fusion	314 kJ/kg	225 kJ/kg
ΔH_v, heat of vaporization	3185 kJ/kg	
p_v, vapor pressure	0.10 kPa (800 °C)	
	1.65 kPa (1000 °C)	
	11.98 kPa (1200 °C)	
	39.10 kPa (1350 °C)	
Solubility in 100 g of		
Water	see Figure 4	see Figure 4
Ethanol, 100 %	1.235 g (25 °C)	0.57 g (19.5 °C)
Ethanol, 95 %	2.445 g (25 °C)	
Methanol	6.44 g (15 °C)	4.91 g (19.5 °C)
	4.58 g (25 °C)	
	4.10 g (67.4 °C)	
NH_3, liquid	58 g (–31 °C)	4.55 g (–33.9 °C)

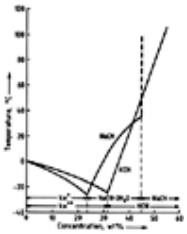

Figure 4. Solubility of NaCN and KCN in water

* Solid phase: ice–NaCN · 2 H_2O or ice–NaCN

** Solid phase: ice–KCN

Dry CO_2 does not react with dry alkali metal cyanides; however, in moist air slow decomposition takes place, even at normal temperature, releasing HCN:

$$2\,NaCN + CO_2 + H_2O \rightarrow Na_2CO_3 + 2\,HCN$$

When this occurs, the salt sometimes becomes brownish because of the formation of polymerization products of HCN (azulmic acid). Alkali metal cyanides are totally decomposed to HCN by the action of strong acids, e.g.,

$$2\,NaCN + H_2SO_4 \rightarrow 2\,HCN + Na_2SO_4$$

At elevated temperatures, oxygen reacts with alkali metal cyanides, to form the metal cyanate and carbonate, nitrogen, and

carbon dioxide [79]:

$$2\,NaCN + O_2 \rightarrow 2\,NaOCN$$

$$2\,NaOCN + 3/2O_2 \rightarrow Na_2CO_3 + N_2 + CO_2$$

The oxidation of alkali metal cyanides to cyanates by bubbling an air stream through an alkaline melt is used for the bath nitriding of steel [80], [81]. Oxides of lead, tin, copper, nickel, and iron react with alkali metal cyanides above 560 °C to form the corresponding cyanate and carbonate, nitrogen, carbon dioxide, and the corresponding metal [82], [83]. Violent oxidation of alkali metal cyanides can result from the addition of strong oxidants such as nitrate, nitrite, or chlorate to a melt:

$$NO_3^- + CN^- \rightarrow N_2 + CO_3^{2-}$$

$$5\,NO_2^- + 3\,CN^- \rightarrow 4\,N_2 + O^{2-} + 3\,CO_3^{2-}$$

These reactions have been utilized for the destruction of waste salts from hardening shops [84], [85].

When alkali metal cyanides are dissolved in water, a pH-dependent, reversible equilibrium is established between hydrocyanic acid and alkali-metal hydroxide [86], [87]:

$$NaCN + H_2O \rightleftharpoons HCN + NaOH$$

For example, in an aqueous solution of NaCN at pH 9.4, half of the total cyanide is present as HCN [88] (cf. Fig. 1). As a result of this hydrolysis, solutions of alkali-metal cyanides in water are always strongly alkaline. Furthermore, commercial products always contain small amounts of alkali metal hydroxide to enhance stability; hence, the actual pH values of solutions are usually higher than those calculated on the basis of the cyanide concentration and the hydrolysis reaction (Table 3).

Table 3. pH values of aqueous solutions with varying concentrations of NaCN and KCN [72]

	Concentration, mol/L					
	1	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
NaCN	11.64	11.15	10.67	10.15	9.6	8.9
KCN		11.37	10.91	10.35		

Nevertheless, when alkali metal cyanides are dissolved in tap water, discoloration and precipitation of brownish-black amorphous polymerization products of HCN (azulmic acid) may occur when the water contains too much CO_2 or other acidic components. When solutions of alkali-metal cyanides are stored for a long time or heated, slow hydrolysis of the C≡N bond takes place to yield the alkali metal formate and ammonia [89], [90]:

$$NaCN + 2\,H_2O \rightarrow HCOONa + NH_3$$

The hydrolytic decomposition of concentrated NaCN solutions can be suppressed, even at 60 – 80 °C, by the addition of 0.5 – 2 % NaOH [91]. However, the hydrolysis is fast and complete when the reaction is carried out at 170 °C under pressure [92], [93]. This process has been investigated for the destruction of cyanide wastes from metal hardening plants [94]. Atmospheric oxygen achieves only partial oxidation of cyanide in aqueous solution at elevated temperature [95]; however, the reaction can be enhanced by the use of catalysts such as activated carbon [96], [97]. A fast conversion of cyanide to cyanate is brought about by ozone at pH 10 – 12; this is followed by a slower oxidation to nitrogen and carbonate [98], [99].

Cyanide is also oxidized to cyanate by hydrogen peroxide [100] and by peroxomonosulfate, peroxodisulfate [101], and permanganate ions [102]. The anodic oxidation of cyanide leads, via cyanogen and the cyanate, to the carbonate, nitrogen, ammonia, and urea [103]. Halogens react with cyanide in aqueous solution to give cyanogen halides, XCN (X = Cl, Br, I), which are hydrolyzed more or less quickly, depending on the pH and temperature of the solution [104] (see Section Properties).

Reaction with chlorine or hypochlorite above pH 11.5 is the most widely used process for detoxification of solutions and wastewaters containing cyanide at concentrations below 1 g/L.

The reaction of alkali cyanides with sulfur or polysulfides in aqueous solution at elevated temperature is used for the production of alkali metal thiocyanates [105]. The reaction with thiosulfate to give thiocyanate is used therapeutically in the treatment of cyanide poisoning [106]. Heavy metal ions react in aqueous solution with cyanide to yield insoluble cyanides, all of which dissolve in the presence of excess cyanide to give stable complex salts [107]. These compounds play an important role in metallurgy and electroplating [108]. Some base metals (e.g., zinc or nickel) and, in the presence of oxygen or oxidizing agents, even precious metals (e.g., gold or silver) are dissolved by aqueous solutions of alkali metal cyanides [109]. This property of alkali metal cyanides has been utilized for about 100 years in the leaching of precious metals ores:

$$4\,NaCN + 2\,Au + 1/2\,O_2 + H_2O \rightarrow 2\,Na[Au(CN)_2] + 2\,NaOH$$

Tantalum, titanium, and tungsten are not attacked by alkali metal cyanide solutions at room temperature.

Alkali metal cyanides react with carbonyl groups of organic compounds in aqueous solution to form cyanohydrins [111], [112]:

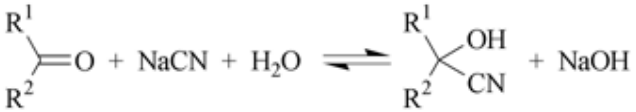

Organic compounds with labile halogen atoms can be converted by reaction with alkali metal cyanides in aqueous solution to nitriles, which can be further processed to give carboxylic acids, amines, etc.:

$$RCl + NaCN \rightarrow RCN + NaCl$$

2.1.2. *Production*

The large-scale production of alkali metal cyanides began in the second half of the 19th century, when J. S. MAC ARTHUR and the FORREST Brothers patented their process for the extraction of gold and silver from ores. The manufacturing method was based on work of F. and E. RODGERS [113], who fused potassium or sodium carbonate with potassium hexacyanoferrate (II), for which animal blood and waste were used as raw materials. A large number of modified processes were patented up to 1900, but the production of cyanides on a major industrial scale finally began with the development of a process based on a patent of H. J. CASTNER, which employed the reaction of sodium metal with charcoal and ammonia [114]:

$$Na + C + NH_3 \rightarrow NaCN + 3/2\,H_2$$

The Castner process was originally a two-stage process, which was developed mainly at Degussa [115], [116] and used there until 1971. In the 1960s sodium cyanide and potassium cyanide are produced exclusively by the neutralization process, which is based on the stoichiometric reaction of a solution of sodium or potassium hydroxide with liquid or gaseous hydrogen cyanide [117-121].

$$MOH + HCN \rightarrow MCN + H_2O$$

Although the process appears very simple, it involves a number of technical difficulties, which result from the tendency of hydrogen cyanide to polymerize, as well as from the hydrolysis of the alkali metal cyanides and their reactivity with carbon dioxide and air. High-grade products can be obtained only when the production process is controlled very closely. In principle, all plants employ the following process steps, which differ only in some details.

Gaseous or liquid HCN is first introduced continuously into concentrated solutions of NaOH or KOH at elevated temperature. The heat of neutralization is used for the subsequent vaporization of water. The use of special mixing devices and the addition of 0.2 – 3 % excess NaOH prevents polymerization [122-124]. Simultaneously or in a second step, water is vaporized under reduced pressure and the alkali metal cyanide is precipitated below 100 °C. The shorter the residence time of the cyanide in solution, the cleaner the product and the lower the content of hydrolysis products and iron impurities [125], [126].

The cyanide crystals are separated by filtration or centrifugation, and various methods are used to dry the wet salt, which still contains about 12 % water. A two-step drying process with hot, CO_2-free air at 200 – 430 °C is advantageous because the product remains free flowing [127], [128]. Sometimes fluidized-bed dryers are utilized [129]. In any case, the used air must be washed and the wastewater detoxified to avoid pollution problems. Standard equipment can be employed for further processing of the salt, which still contains up to 1 % moisture at this stage [130]. The final commercial forms are powder, granulate, briquettes, or tablets [131], [132].

The production of NaCN from nitrogenous waste from the manufacture of sugar [133] is the precursor of the “neutralization process.” This process no longer has any commercial importance.

2.1.3. *Commercial Forms, Specifications, and Packaging*

Most alkali metal cyanides are supplied in solid form. Common commercial forms for NaCN and KCN are granular products composed of irregularly shaped particles ca. 0.1 – 5 mm in size, nearly free of dust; molded briquettes weighing 15 – 40 g; and cylindrical tablets ca. 25 – 40 mm in diameter and weighing 20 – 40 g.

Briquettes and tablets are the preferred product forms, especially for gold extraction by cyanidation because of their dust-free nature. Smaller amounts of NaCN and KCN are used as granules or as coarse-grained powders for the preparation of mixtures or for direct application in the electroplating industry or in organic synthesis.

The specifications according to the German standards [134] for the electroplating industry (DIN 50 971) for NaCN and KCN are listed in Table 4. The quality of commercial products normally exceeds the requirements of DIN 50 971.

Table 4. Specifications and average analysis of NaCN and KCN (data given in wt % or ppm = mg/kg; M = Na or K)

	NaCN		KCN	
	DIN 50 971	Average analysis	DIN 50 971	Average analysis
MCN, %	> 98.0		98 – 99	

CN^-, %	51.5	> 52.0	39.0	> 39.0
MOH, %		≈ 0.5		≈ 0.1
M_2CO_3, %		≈ 0.5		≈ 1.0
Cl^-, %		≈ 0.02		≈ 0.0015
SO_4^{2-}, %		≈ 0.05		< 0.01
SO_3^{2-}, %		< 0.05		< 0.01
S^{2-}, ppm		< 1		< 1
MOCN, %		≈ 0.1		≈ 0.1
SCN^-, %		< 0.01		< 0.01
HCOOM, %		≈ 0.5		≈ 0.3
H_2O, %		≈ 0.2		≈ 0.2
Sb, ppm	< 5	< 5	< 5	< 5
As, ppm	< 5	< 5	< 5	< 5
Pb, ppm	< 50	< 50	< 50	< 50
Cd, ppm	< 10	< 10	< 10	< 10
Fe, ppm	< 50	< 50	< 50	< 50
Cu, ppm	< 10	< 10	< 10	< 10
Ni, ppm	< 50	< 50	< 50	< 50
Zn, ppm	< 20	< 20	< 20	< 20
Sn, ppm	< 80	< 80	< 30	< 30
Na, %		< 48	< 0.5	< 0.3
Insolubles, %	< 0.005	< 0.001	< 0.005	< 0.001

Today, the preferred packaging forms for NaCN and KCN are flexible intermediate bulk containers (FIBCs) in a plywood box, especially for the mining industry. For further protection and to make the packaging air- and watertight, an additional polyethylene inliner is used inside the plywood box. This type of package usually contains 1 t. If the customer is located in the vicinity of a production plant, large quantities of cyanide can also be supplied in returnable metal IBCs (Intermediate Bulk Container), generally containing 1 t of product.

For smaller customers, for example, the electroplating industry, solid alkali metal cyanides are also supplied in plated steel drums containing 50 or 100 kg of cyanide. The lid of the drum is fixed with a locking ring and hermetically sealed.

All types of packaging must meet the official requirements for the transportation of toxic substances (UN approval). Furthermore, the packagings must be labeled according to the transport regulations including the proper shipping name, the UN product number, labeling as a toxic product (skull and crossbones, marine pollutant labels) and the country of origin, as well as recommendations on safe handling and first aid procedures.

Sodium and potassium cyanides can be supplied as stable aqueous solutions [91] with about 30 wt % NaCN or 40 wt % KCN in containers with a capacity of at least 20 m^3, in railcars or tank trucks. However, liquid cyanide supply is only economically feasible within a limited radius of the production plant and is not permitted by sea freight. The concentration of the solution can be adjusted to ensure that, even in winter, no precipitation occurs. Liquid cyanide is easier to handle and is the preferred form when large quantities of cyanides are required and aqueous solutions can be used. Major consumers are the mining industry and the organic chemical industry.

2.1.4. *Transportation, Storage, and Handling*

Transportation. Alkali metal cyanides can be shipped by road, rail, or sea freight under the provisions of the applicable transport regulations. The transport classifications of the most common transport regulations are as follows:

Alkali metal cyanide solid: RID/ADR (European land transport) and ADNR (European inland navigation): Class 6.1, 41° a)

IMDG Code (marine transport) and CFR 49 (U.S. domestic transport): Class 6.1, UN 1689 (for NaCN), UN 1680 (for KCN), PG I

Alkali metal cyanide solutions: RID/ADR (European landtransport) and ADNR (European inland navigation): Class 6.1, 41° a), b), or c) (depending on the concentration)

IDMG Code (marine transport) and CFR 49 (U.S. domestic transport): Class 6.1, UN 1935, PG I, II, or III (depending on the concentration)

They must never be stored or transported in the same transport unit together with acids or acid salts, foods, feedstuffs, or anything else intended for human or animal consumption.

Carriers should be selected on the basis of their safety commitment and record as well as commercial factors. The Safety and Quality Assessment System (CEFIC or CMA) is an objective means of carrying out this selection for road transportation.

Hauliers should be experienced in handling cyanides and know how to act in case of emergencies.

As well as the general provisions of ADR, driver training should include awareness of the specific hazards of cyanides.

Appropriate personal protective equipment must be carried.

Loads containing cyanides must only be left unattended in secure areas away from the general public. The vehicle shall be locked.

All incidents and accidents must be reported, investigated, and corrective action taken to prevent recurrence.

When severe weather conditions are experienced during transport, for example, icy roads, snow, or poor visibility caused by fog, the delivery should be stopped at the next suitable parking place. Delivery should not resume until weather conditions improve.

All delays during transport caused be severe weather conditions, breakdowns, or other reasons must be reported to the consignor immediately.

Transport accidents must also be reported to the consignor immediately. An on-board telephone is highly recommended.

If emergency action must be taken by drivers when leaks, spills, or fire occur during transport, the instructions given in Transport Emergency Card must be followed.

Drums, boxes, or intermediate bulk containers must be properly secured to minimize the risk of movement during transit or loss of containment in an incident.

For transportation by truck or rail the use of tilts is recommended as the solid sides minimize lateral movement during transit. Where possible, material should be palletized and the pallet securely fixed to the vehicle.

Free space in all means of transport should be minimized. This can be achieved by efficient packing and/or use of appropriate void fillers. Particular attention should be paid to securing the load adjacent to the container doors to minimize the risk of drums falling out when the doors are opened.

Small loads of a few drums, often as part of a mixed load, must be segregated from noncompatible cargo and securely fixed. A storage box fixed to the vehicle which provides secondary containment in the event of spillage is suitable.

Sodium cyanide solution should only be transported on defined routes. The route to be followed must be selected carefully and should be known to both the haulier and the consignor or supplier. As far as possible, the route should utilise motorways/highways and avoid major population centers.

Storage and Handling. Any type of packaging that contains cyanide must be stored in a dry place and protected against corrosion. Rooms in which cyanides are stored and processed must be well ventilated. A full mask with a type B filter (identifying color gray, shelf life unused max. five years), must be worn when HCN might be present. The cyanide storage area must be clearly marked and unauthorised persons must never be allowed access to rooms in which cyanides are stored or processed. Therefore, a cyanide area must be separated from other areas, e.g., by fences or walls. In case of an emergency, powder fire extinguishers, safety showers, eye washes, and antidotes should be available in the immediate vicinity.

If the original cyanide drums are used as storage containers, they must be tightly closed immediately after removal of cyanide and should not be allowed to stand open. Stainless steel should be used as construction material for equipment that comes in contact with cyanide solutions up to 100 t [109]. However, concentrated cyanide solutions must be stored in steel or stainless steel containers which are protected against leakage by a collecting basin constructed of alkali-resistant concrete.

Spilled solid cyanide must be swept up immediately with a dustpan and brush and returned to the container. If it has become dirty and unusable, the cyanide must be destroyed or disposed of in accordance with legal regulations. In addition, spilled cyanide solutions and wastewater that contains cyanide must be detoxified (see Section Wastewater Treatment) before discharge into public sewage system or open water. Containers, equipment, and any objects or floor areas that have become contaminated with cyanide must be washed thoroughly with water, which must then also be detoxified.

Work with cyanide requires the greatest possible caution and strictest adherence to all safety regulations. Never work alone when handling cyanide, do not eat, drink, smoke, use snuff, or chew gum or tobacco. Also do not store any food, beverages, or smoking material in the cyanide area. Prevent any contact of cyanide with the skin or eyes, as well as the respiration of cyanide dust or hydrogen cyanide gas. If a danger of cyanide contamination exists, rubber gloves, rubber boots, a rubber apron, and safety goggles or protection for the face and head must be worn. Safety showers and eye washes should be located in the immediate vincinity of each cyanide area. Prevent the development of cyanide dust. If cyanide dust is produced, wear approved respiration protection, preferably a full-face mask. Filters must be replaced after each use. Never bring acids, acid salts, or oxidants into contact with cyanide or allow combined storage, because any direct contact would generate highly poisonous gaseous hydrogen cyanide. Never handle cyanide at a pH which may produce hydrogen cyanide gas.

After handling cyanide, wash any contaminated parts of the body, especially the hands thoroughly with copious water. Contaminated clothes should be washed at work and never at home.

Further safety information is compiled in the Chemical Safety Data Sheets of the Manufacturers Association of the Chemical

Industry [135], [136].

2.1.5. *Economic Aspects and Uses*

In gold extraction by cyanidation the countercation has no effect on the leaching process. Therefore, the cheaper cyanides are used for this application. As potassium cyanide is more expensive than sodium cyanide, the latter is used exclusively in this industry. In other applications, the countercation is of importance, and potassium cyanide has some advantages, e.g., in electroplating, because the potassium ion has greater mobility than the sodium ion.

Gold and Silver Leaching. The leaching of precious metals such as gold and silver from ores is the main application of NaCN. As a result of its unique chemical affinity, especially for gold, cyanide forms the most stable complex of all lixiviants available, which is important for the complete recovery of the accessible gold in the ore. The high affinity for gold also enables operation at relatively low cyanide concentration to nevertheless achieve a high recovery.

As cyanidation has been known and applied for more than 100 years in all five continents, the potential effect of cyanide on the environment has been studied in detail, and suitable technology for effective cyanide detoxification is available and can be used to avoid environmental impact of cyanide.

All efforts to replace cyanide with other less toxic complexing agents have remained unsuccessful. All other lixiviants, such as the systems chloride/chlorine, bromide/bromine, hydrogen sulfide, and thiourea, require much higher concentrations, more complicated processes, and higher costs compared to cyanidation technology. Furthermore, these reagents are also toxic, and their impact on the environment as well as suitable detoxification techniques have not yet been studied in detail.

Therefore, cyanide will remain the technology of choice in the future for gold extraction, especially because developments in cyanide recovery and cyanide detoxification of the tailings are constantly improving the environmental suitability of the cyanidation process.

Base Metal Flotation. Cyanide is used as a depressant in the selective flotation of zinc, lead, copper, and iron [138]. This application is, however, declining because other, higher performance and more economical compounds have become increasingly popular over the last few years.

Electroplating (⟶ Electrochemical and Chemical Deposition – Specific Deposits and Plating Solutions***).*** Large quantities of alkali metal cyanides, as well as the complex cyanides of copper, zinc, and cadmium manufactured therefrom, are used in electroplating processes. In particular, the plating of iron, steel, and zinc is carried out with cyanide-containing electrolytes. Very good corrosion protection resulting from the even deposition of layers, ease of use, and low sensitivity of the bath to impurities are the main advantages of these electrolytes. In this application, potassium cyanide is preferred, patricularly for the deposition of gold and silver. Generally, electrolytes based on potassium salts have the advantages of better performance and lower sensitivity to impurities [139].

When the highest quality is required, cyanide pretreatment and degreasing baths are still in use today, especially since attempts to employ other strong complexing agents have failed because of insurmountable wastewater problems.

Metal-Surface Hardening in Molten Salts. Sodium and potassium cyanides are important because they release nitrogen or carbon or both onto the surface of the material being treated at relatively low temperatures. The Tufftride process is employed primarily for the hardening of steel parts used as tools or in engines where they are highly stressed [140-143]. Particularly sodium cyanide is used in this application, because of its lower price.

Organic Synthesis [111]. Large quantities of sodium cyanide are used to introduce cyano groups into organic compounds. This is the first step in the synthesis of a variety of well-known compounds used in pharmacy, complex chemistry, and polymer chemistry.

In particular, organic halogen compounds react with sodium cyanide in water or polar aprotic solvents to form nitriles:

$$\mathrm{NaCN + RCl \rightarrow RCN + NaCl}$$

The nitriles can be converted to carboxylic acids, amides, esters, and amines. The sodium salt of cyanoacetic acid, for example, is manufactured in this way and is raw material for the manufacture of malonic acid, caffeine, and numerous other important organic products.

Inorganic Chemistry. An important application of sodium and potassium cyanides is the production of heavy metal cyanides, particularly complex iron cyanides (see Section Heavy-Metal Cyanides).

2.2. Alkaline Earth Metal Cyanides

Properties and Uses. The properties of the alkaline earth metal cyanides differ from those of the alkali metal cyanides. They are less stable and decompose at elevated temperatures to release HCN. Even when alkaline earth metal cyanides are in contact with moist air or dissolved in water, hydrolysis takes place, e.g.,

$$\mathrm{Ca(CN)_2 + 2\,H_2O \rightarrow Ca(OH)_2 + 2\,HCN}$$

As a consequence, pure solid calcium cyanide is unstable, and the commercial product can only be manufactured in amorphous form with a content of about 45 – 50 % NaCN equivalent, known as black cyanide. As a solution, calcium cyanide is only sufficiently stable for commercial use at concentrations below 17 %.

Liquid and solid calcium cyanide are only produced in South Africa, where they are used for the leaching of precious metals from ores [145].

Apart from this compound, only barium cyanide has any commercial importance as a result of its use in electroplating to separate carbonate from cyanides containing electrolytes [139]. Some of the physical properties of these compounds are listed in Table 5.

Table 5. Physical properties of the alkaline-earth-metal cyanides

Property	$Mg(CN)_2$ [4100-56-5]	$Ca(CN)_2$ [592-01-8]	$Sr(CN)_2$ [52870-08-3]	$Ba(CN)_2$ [542-62-1]
M_r	76.37	92.12	139.67	189.40
mp, °C		> 350		≈ 600
Decomp., °C	> 300	≈ 640	> 500	> 600
Other crystal forms	$Mg(CN)_2 \cdot 2 NH_3$	$Ca(CN)_2 \cdot 2 NH_3$	$Sr(CN)_2 \cdot 4 H_2O$	$Ba(CN)_2 \cdot 2 H_2O$

Production. The alkaline earth metal cyanides can be obtained by reaction of the corresponding metal hydroxides with hydrogen cyanide [5]. However, only liquid calcium cyanide is produced by this method. The resulting product has a brownish color due to polymerization of HCN, liberated by hydrolysis.

The production process of the commercially available solid calcium cyanide is different. Black cyanide is produced by treating calcium carbide with nitrogen gas to give calcium cyanamide. The crude cyanamide is then mixed with sodium chloride and heated in electric furnaces above 1000 °C [146]. The cyanamide is converted into calcium cyanide, which is quenched to produce amorphous flakes. The reaction product is gray-black because of the presence of 1 – 2 % carbon. The cyanide content of the amorphous black cyanide is ca. 45 – 50 % NaCN equivalent. Other constituents are 32 % NaCN, 12 % CaO, 2 – 3 % $CaCl_2$, and 2 – 3 % CaNCN.

2.3. Heavy-Metal Cyanides

Heavy metals usually form stable, insoluble or sparingly soluble simple cyanides $M(CN)_x$, which dissolve in the presence of excess alkali metal cyanide to form stable complex salts [107]:

$$M(CN)_x + yCN^- \rightleftharpoons \left[M(CN)_{x+y}\right]^{y-}$$

where $x = 1 – 3$; $y = 1 – 4$

When dissolved in water, the complex cyanide anions tend to dissociate, and various dissociation equilibria are established. The dissociation properties of the most important heavy-metal cyanides are shown in Table 6.

Table 6. Stability of heavy-metal cyanide complexes [4], [107]

Dissociation equilibria	$pK_{dissociation}$
$[Pb(CN)_4]^{2-} \rightleftharpoons Pb^{2+} + 4 CN^-$	10.3
$[Cd(CN)_4]^{2-} \rightleftharpoons [Cd(CN)_3]^- + CN^-$	2.5
$[Cd(CN)_3]^- \rightleftharpoons Cd^{2+} + 3 CN^-$	14.7
Σ	17.2
$[Zn(CN)_4]^{2-} \rightleftharpoons [Zn(CN)_3]^- + CN^-$	1
$[Zn(CN)_3]^- \rightleftharpoons Zn^{2+} + 3 CN^-$	17.9
Σ	18.9
$[Ag(CN)_2]^- \rightleftharpoons Ag^+ + 2 CN^-$	20.9
$[Ni(CN)_4]^{2-} \rightleftharpoons Ni^{2+} + 4 CN^-$	≈ 22
$[Cu(CN)_4]^{3-} \rightleftharpoons [Cu(CN)_3]^{2-} + CN^-$	1.5
$[Cu(CN)_3]^{2-} \rightleftharpoons [Cu(CN)_2]^- + CN^-$	5.3
$[Cu(CN)_2]^- \rightleftharpoons Cu^+ + 2 CN^-$	23.9

	Σ	30.7
$[Fe(CN)_6]^{3-} \rightleftarrows Fe^{3+} + 6\ CN^-$	≈	36
$[Au(CN)_2]^- \rightleftarrows Au^+ + 2\ CN^-$	≈	37
$[Hg(CN)_4]^{2-} \rightleftarrows Hg^{2+} + 4\ CN^-$		40.5
$[Fe(CN)_6]^{4-} \rightleftarrows Fe^{2+} + 6\ CN^-$	≈	42
$[Co(CN)_6]^{4-} \rightleftarrows Co^{2+} + 6\ CN^-$	≈	64
$[Pt(CN)_4]^{2-} \rightleftarrows Pt^{2+} + 4\ CN^-$	≈	40
$[Pd(CN)_4]^{2-} \rightleftarrows Pd^{2+} + 4\ CN^-$	≈	42

2.3.1. *Iron Cyanides*

2.3.1.1. *Properties*

The most important iron cyanides are the hexacyanoferrate(II), ferrocyanide, $[Fe(CN)_6]^{4-}$, and the hexacyanoferrate(III), ferricyanide, $[Fe(CN)_6]^{3-}$, complexes, which have an octahedral configuration. They are among the most stable inorganic complexes and are practically nontoxic because of the strong bonding between iron and cyanide (for LD_{50} values, see Toxicology and Occupational Health). Practically none of the known cyanide reactions can be observed when these compounds are dissolved in water. Only when irradiated with UV light does the hexacyanoferrate(II) ion decompose slowly at normal temperature to form cyanide ions [148]:

$$[Fe(CN)_6]^{4-} + H_2O \underset{\text{dark}}{\overset{\text{UV}}{\rightleftharpoons}} [Fe(CN)_5H_2O]^{3-} + CN^-$$

In the absence of UV light, only very strong oxidants like ozone are able to oxidize the cyanide and destroy the hexacyanoferrate(II) complex slowly at elevated temperature [149]:

$$2[Fe(CN)_6]^{4-} + 13\,O_3 + 4\,OH^- + H_2O \rightarrow 2\,Fe(OH)_3 + 12\,OCN^- + 13\,O_2$$

Permanganate, peroxodisulfate, hypochlorite, chlorine, iodine, cesium(IV), and other oxidants oxidize hexacyanoferrate(II) to hexacyanoferrate(III) at moderate temperatures [150], e.g.,

$$5\,[Fe(CN)_6]^{4-} + MnO_4^- + 8\,H^+ \rightarrow 5[Fe(CN)_6]^{3-} + Mn^{2+} + 4\,H_2O$$

This reaction is used for the determination of hexacyanoferrate (II) in water. Hydrogen peroxide oxidizes hexacyanoferrate(II) to hexacyanoferrate(III) below pH 3 and reduces hexacyanoferrate(III) to hexacyanoferrate(II) above pH 7:

$$2[Fe(CN)_6]^{4-} + H_2O_2 + 2\,H^+ \rightarrow 2[Fe(CN)_6]^{3-} + 2\,H_2O$$

$$2[Fe(CN)_6]^{3-} + H_2O_2 \rightarrow 2[Fe(CN)_6]^{4-} + 2\,H^+ + O_2$$

When acidified with mineral acids, solutions of hexacyanoferrates decompose slowly at room temperature and more rapidly when heated, releasing HCN and forming precipitates of iron(II) hexacyanoferrate(II) [14460-02-7] [150]:

$$3[Fe(CN)_6]^{4-} + 12\,H^+ \rightarrow 3\,H_4[Fe(CN)_6] \rightarrow 12\,HCN + Fe_2[Fe(CN)_6]$$

Insoluble or sparingly soluble metal hexacyanoferrates(II) are formed by the addition of aqueous solutions to neutral or weakly acidic solutions of metal salts, particularly heavy metal salts [150]:

$$4\,M^+ + [Fe(CN)_6]^{4-} \rightarrow M_4[Fe(CN)_6]\,(\text{sparingly soluble})$$

$$2\,M^{2+} + [Fe(CN)_6]^{4-} \rightarrow M_2[Fe(CN)_6]\,(\text{sparingly soluble})$$

These reactions are used to separate heavy metals from solutions for the purpose of refining and for the production of colored pigments, e.g., Prussian blue pigments, which contain the hexacyanoferrate(III) anion:

$$M^+ + Fe^{2+} + [Fe(CN)_6]^{3-} \rightarrow MFe[Fe(CN)_6]$$

M = K, Na, NH_4, Cu

Multinuclear complexes are formed by the thermolysis of hexacyanoferrate(II) complexes with complex cations [5]. Many anionic and neutral ligands can be introduced into the hexacyanoferrate anion by substitution of one CN group, to form pentacyano complexes [4], [5].

$$[Fe(CN)_6]^{3-/4-} + X \rightarrow [Fe(CN)_5X]^{2-/3-} + CN^-$$

where X = H_2O, NH_3, NO, NO_2, NO_5, CO

$$[Fe(CN)_6]^{3-/4-} + X^- \rightarrow [Fe(CN)_5X]^{3-/4-} + CN^-$$

where X = SCN, NCS.

Some of these compounds are used in analytical chemistry. Hexacyanoferrate(II) reacts with a variety of organic amino compounds to form insoluble or slightly soluble salts of general composition BA, B_3A_2, B_2A, B_3A, and B_4A, where B is the cationic nitrogen compound and A the anionic hexacyanoferrate(II) [4]. These reactions are used in analytical and pharmaceutical chemistry. Hexacyanoferrate(III) is a strong oxidant which can dissolve metals; it is used in photographic processes [151] and etching techniques [152]:

$$[Fe(CN)_6]^{3-} + M \rightarrow [Fe(CN)_6]^{4-} + M^+$$

The hexacyanoferrates(II) of sodium, potassium, and calcium crystallize with different amounts of water of crystallization, whereas potassium hexacyanoferrate(III) is obtained in anhydrous form. The hexacyanoferrates(II) start to lose water of crystallization above 30 °C and become anhydrous white powders above 80 °C. All hexacyanoferrates decompose above 400 °C to form alkali metal or calcium cyanide, elemental iron, carbon, and nitrogen, e.g.,

$$Na_4[Fe(CN)_6] \rightarrow 4\,NaCN + Fe + 2\,C + N_2$$

The alkali metal hexacyanoferrates are readily soluble in water; the solubility of the calcium salt is more than twice that of the sodium salt (Table 7). Double salts of low solubility may be formed in the presence of Ca^{2+} and K^+ or NH_4^+ ions.

Table 7. Solubilities of hexacyanoferrates in water (wt % of anhydrous salt)

Compound	Temperature, °C		
	20	50	80
$Na_4[Fe(CN)_6]$	16	26	38
$K_4[Fe(CN)_6]$	22	32	40
$Ca_2[Fe(CN)_6]$	36	42	44
$K_3[Fe(CN)_6]$	31	39	45

2.3.1.2. *Production*

Historically, the preparation of hexacyanoferrates(II) was based on the fusion of potash with iron compounds and animal residues such as hide, horns, or dried blood, which led to the German name *gelbes Blutlaugensalz* (yellow salt of blood lye). Later, the absorption of hydrogen cyanide from coal gas on iron hydroxide was used to produce hexacyanoferrates. Today, synthetic hydrogen cyanide, iron(II) chloride, and alkali metal or calcium hydroxide are the raw materials for the large-scale production of hexacyanoferrates(II) on a large scale.

Calcium hexacyanoferrate(II) is usually the primary product, which is subsequently converted to the potassium and sodium salts. Liquid hydrogen cyanide and an aqueous solution of iron(II) chloride are mixed with calcium hydroxide solution in stoichiometric amounts in a stirred reactor:

$$3\,Ca(OH)_2 + FeCl_2 + 6\,HCN \rightarrow Ca_2[Fe(CN)_6] + CaCl_2 + 6\,H_2O$$

After filtration, the solution is concentrated by evaporation of water under reduced pressure, and the calcium hexacyanoferrate(II) crystallizes with eleven molecules of water of crystallization. The relatively coarse-grained salt is then separated by filtration and generally used without drying. For the conversion of calcium hexacyanoferrate(II) to the potassium or sodium salt, two methods are used.

In the first, a stoichiometric amount of potassium chloride is added to the filtered calcium hexacyanoferrate(II) solution and the hexacyanoferrate(II) precipitates as a sparingly soluble potassium calcium double salt [153]

$$Ca_2[Fe(CN)_6] + 2\,KCl \rightarrow K_2Ca[Fe(CN)_6] + CaCl_2$$

The double salt is separated by filtration, redispersed in water, and then converted to the soluble potassium hexacyanoferrate(II) by the addition of potassium carbonate:

$$K_2Ca[Fe(CN)_6] + K_2CO_3 \rightarrow K_4[Fe(CN)_6] + CaCO_3$$

After separation of the insoluble calcium carbonate, the potassium hexacyanoferrate (II) solution is concentrated by evaporation, and the potassium salt crystallizes with three molecules of water of crystallization. After filtration, the salt is

carefully dried and packaged.

A similar method can be used for the production of sodium hexacyanoferrate(II): sodium carbonate is added to the calcium hexacyanoferrate(II) solution and the precipitated calcium carbonate is separated:

$$Ca_2[Fe(CN)_6] + 2\,Na_2CO_3 \rightarrow Na_4[Fe(CN)_6] + 2\,CaCO_3$$

Usually, the sodium hexacyanoferrate(II) is synthesized directly from solutions of sodium cyanide and iron(II) chloride [154]:

$$6\,NaCN + FeCl_2 \rightarrow Na_4[Fe(CN)_6] + 2\,NaCl$$

The crystalline Na_4 [Fe $(CN)_6$] · 10 H_2O can be obtained when the solution is concentrated; it must be dried very carefully to avoid loss of water of crystallization, which would influence its solubility properties.

The disadvantages of the double-salt process are the necessity of recycling the various mother liquors and the formation of large quantities of solid waste products.

In the second conversion method, the hexacyanoferrate(II) is precipitated from a solution of the calcium salt as Berlin white by the addition of iron(II) chloride solution:

$$Ca_2[Fe(CN)_6] + 2\,FeCl_2 \rightarrow Fe_2[Fe(CN)_6] + 2\,CaCl_2$$

This iron(II) hexacyanoferrate(II) is separated by filtration, redispersed in water, and converted to sodium or potassium hexacyanoferrate(II) by the addition of stoichiometric amounts of sodium or potassium cyanide:

$$Fe_2[Fe(CN)_6] + 12\,KCN \rightarrow 3\,K_4[Fe(CN)_6]$$

Potassium hexacyanoferrate(III) can be obtained by the oxidation of potassium hexacyanoferrate(II). Usually, anodic oxidation is applied by means of nickel electrodes [155]:

$$2\,K_4[Fe(CN)_6] + 2\,H_2O \rightarrow 2\,K_3[Fe(CN)_6] + 2\,KOH + H_2$$

Alternatively, hydrogen peroxide can be used for the oxidation of hexacyanoferrate(II) when the reaction is carried out below pH 3 [156]:

$$3\,K_4[Fe(CN)_6] + H_4[Fe(CN)_6] + 2\,H_2O_2 \rightarrow 4\,K_3[Fe(CN)_6] + 4\,H_2O$$

The dry, ruby-red potassium hexacyanoferrate(III) is obtained by concentration of the solution at reduced pressure and crystallization at ca. pH 7. The fact that H_2O_2 reduces hexacyanoferrate(III) to hexacyanoferrate(II) at higher pH must also be considered.

2.3.1.3. *Commercial Forms, Specifications, and Packaging*

The following commercial products are most frequently used:

Sodium hexacyanoferrate(II) [69043-75-0], sodium ferrocyanide, yellow prussiate of soda (YPS), Na_4 [Fe $(CN)_6$] · 10 H_2O, M_r 484.108, ϱ 1.46 g/cm^3, light yellow, monoclinic crystals, readily soluble in water (33.7 g in 100 g H_2O at 20 °C).

Potassium hexacyanoferrate(II) [14459-95-1], potassium ferrocyanide, yellow prussiate of potash (YPP), K_4 [Fe (CN) $_6$] · 3 H_2O, M_r 422.39, ϱ 1.85 g/cm^3, lemon yellow, monoclinic crystals, readily soluble in water (33.7 g in 100 g H_2O at 20 °C).

Calcium hexacyanoferrate(II) [13821-08-4], calcium ferrocyanide, yellow prussiate of calcium (YPC), Ca_2 [Fe (CN) $_6$] · 11 H_2O, M_r 490.296, ϱ 1.68 g/cm^3, yellow monoclinic crystals, highly soluble in water (148.4 g in 100 g H_2O at 20 °C).

Potassium hexacyanoferrate(III) [13746-66-2], potassium ferricyanide, K_3 [Fe $(CN)_6$], M_r 329.25, ϱ 1.858 g/cm^3, orange to ruby red, monoclinic crystals, readily soluble in water (46.4 g in 100 g H_2O at 20 °C).

The commercial products are fine, crystalline, free-flowing powders. Normally, they are packaged in polyethylene bags of 25-kg net weight, and must be stored under cool, dry conditions and kept away from acids. Hexacyanoferrates are not dangerous goods according to appendix C of the EVO, GGVS, RID, ADR, and IMCO Code. The specifications and average analyses for the commercially used hexacyanoferrates are listed in Table 8.

Table 8. Specifications of hexacyanoferrates

	Na_4[Fe (CN) $_6$] · 10 H_2O (anticaking of NaCl)	**K_4[Fe (CN) $_6$] · 3 H_2O (refining of wine)**	**Ca_2[Fe (CN) $_6$] · 11 H_2O (anticaking of NaCl)**	**K_3[Fe $(CN)_6$] (photographic industry [*])**

Assay, %	> 99.0	> 99.0	> 99.5	> 98.5
$[Fe(CN)_6]^{4-}$, %	> 43.3	> 49.7	> 43.0	< 0.5
H_2O, %	< 37.5 *	< 12.0	< 41 *	< 0.1
Ca + Mg, mval		< 60		
Cl^-, %	< 0.2	< 0.02	< 0.6 *	< 0.25
SO_4^{2-}, %	< 0.07	< 0.005	< 0.01 *	< 0.1
CO_3^{2-}, %	< 0.15 *	< 0.1 *	< 0.05 *	
S^{2-}, ppm		< 100		
NH_3, ppm		< 100		
Heavy metals, as				
Pb, ppm	< 10	< 10	< 10	
Zn, ppm	< 25	< 10	< 25	
As, ppm	< 3	< 1	< 3	
HCN, ppm		< 100		
Insolubles, %	< 0.03	< 0.01	< 0.1 *	< 0.05
pH (10 % solution)	10 – 11 *	10 – 11 *	10 – 11 *	6 – 8

* Average analysis.

Uses and Economic Aspects. Sodium, potassium, and calcium hexacyanoferrates(II) are equally suitable for use in most applications, but sometimes one product may be preferred. For the production of blue pigments, the most important application of hexacyanoferrates, the potassium salt is used in Europe, whereas the sodium salt is preferred in the United States. Generally, blue hexacyanoferrate(II) compounds are formed by the reaction of iron(III) salts with hexacyanoferrates (II):

$$FeCl_3 + K_4[Fe(CN)_6] \rightarrow FeCl_2 + K_3[Fe(CN)_6] + KCl \rightarrow KFe(III)[Fe(II)(CN)_6] + 3\,KCl$$

The production of blue pigments is more complicated, however, and is based primarily on the oxidation of iron(II) hexacyanoferrate(II) precipitates in the presence of alkali metal or ammonium salts [2]. The properties of blue pigments are not dependent only on the composition; the precipitation method and the finishing of the product are also very important. The use of hexacyanoferrates(II) for the production of classified documents is also based on this reaction. Large amounts of all three types of hexacyanoferrates (II) are used as additives for rock salt to prevent caking [157]; hexacyanoferrates(II) act as crystal growth inhibitors in this application. The ability of hexacyanoferrates(II) to form insoluble or almost insoluble heavy metal compounds is used, e.g., in the production of citric acid by fermentation [158], in the refining of wine [159], in the electroplating of tin [160], and in analytical chemistry. Potassium hexacyanoferrate(III) and alkali hexacyanoferrate(II) combined with peroxodisulfates play an important role as bleaching agents in color photography [151]. Both hexacyanoferrates are also important depressants for the separation of molybdenum from copper by flotation [161].

2.3.2. *Cyanides of Copper, Zinc, and Cadmium*

Properties.

Copper(I) cyanide [544-92-3], CuCN, M_r 89.56, *mp* 473 °C, forms white monoclinic prismatic crystals, ϱ 2.92 g/cm^3, which are insoluble in water, dilute mineral acids, and organic solvents; CuCN is soluble in ammonia and in alkali metal cyanide solution because equilibria are established with highly soluble cyano complexes:

$$CuCN + 2\,NaCN \rightleftharpoons Na_2[Cu(CN)_3] + NaCN$$

(slightly soluble) (slightly soluble)

$$Na[Cu(CN)_2] + NaCN \rightleftharpoons Na_2[Cu(CN)_3] + NaCN \rightleftharpoons Na_3[Cu(CN)_4]$$

(slightly soluble)

These complexes can also be obtained as solids. In the absence of air and moisture, pure CuCN can be stored indefinitely without deterioration.

Zinc cyanide [557-21-1], $Zn(CN)_2$, M_r 117.42, *mp* 800 °C (decomp.), forms a white crystalline powder with a slight HCN odor, ϱ 1.852 g/cm^3 (25 °C), which is insoluble in water but soluble in aqueous solutions of alkali metal hydroxides or cyanides by formation of hydroxy ($[Zn(OH)_4]^{2-}$) or cyano complexes ($[Zn(CN)_4]^{2-}$). In the absence of air, carbon dioxide, and moisture, zinc cyanide can be stored indefinitely without deterioration.

Cadmium cyanide [542-83-6], $Cd(CN)_2$, M_r 164.45, *mp* 200 °C (decomp.), forms white irregular octahedral crystals, ϱ

2.23 g/cm^3 (25 °C), which are soluble in water (1.7 g in 100 g H_2O at 15 °C). Cyano complexes are formed when cadmium cyanide is dissolved in aqueous solutions of alkali metal cyanides: $[Cd(CN)_3]^-$, $[Cd(CN)_4]^{2-}$. In the absence of air and moisture, pure $Cd(CN)_2$ is stable and can be stored indefinitely; in moist air, it decomposes slowly.

Production. The preferred method for the production of copper cyanide is based on the reaction of copper(II) salts, metallic copper, and hydrogen cyanide in hydrochloric acid solution [162], [163]. The reaction solution can be recycled, which avoids environmental problems. The $Cu(CN)_2$ formed when copper(II) salts react directly with alkali cyanide or HCN is unstable and decomposes to CuCN and cyanogen, $(CN)_2$ (see Section Production).

Zinc cyanide can be obtained from the reaction of zinc salts with alkali cyanide in aqueous solution. However, the reaction of zinc oxide with hydrogen cyanide in acetic acid has the advantage that no additional salt is produced [164].

Cadmium cyanide can be produced by evaporation of a solution of cadmium hydroxide in aqueous hydrogen cyanide [165] or by precipitation from a cadmium salt solution with alkali cyanide [166].

Commercial Forms, Specifications, and Packaging. Commercially, CuCN and $Zn(CN)_2$ are supplied as white, fine crystalline powders, packaged in steel drums of 10 kg, 25 kg, and 50 kg net weight. Also 1 t packages, as described for the alkali metal cyanides are available. The specifications and the average analyses are shown in Table 9.

Table 9. Specifications (DIN 50971) and average analysis of the cyanides of copper, zinc, and cadmium

	CuCN		$Zn(CN)_2$		$Cd(CN)_2$	
	DIN 50 971	Average analysis	DIN 50 971	Average analysis	DIN 50 971	Average analysis
Cu^+, %	> 70.0	> 70.0	< 0.001	< 0.001	< 0.001	< 0.001
Zn^{2+}, %	< 0.002	< 0.002	> 55.0	> 55	< 0.005	< 0.005
Cd^{2+}, %	< 0.001	< 0.001	< 0.005	< 0.005	> 67.0	> 67.0
CN^-, %	> 28.8	> 28.8	> 42.0	> 43	> 30.0	> 30.0
OH^-, %		< 0.05		< 0.05		< 0.05
CO_3^{2-}, %		< 0.1		< 0.5		
Cl^-, %		< 0.1		< 0.1		< 0.1
SO_4^{2-}, %		< 0.1		< 0.3		< 0.5
H_2O, %		< 0.1		< 0.5		< 0.5
Sb, ppm	< 10	< 10	< 10	< 10	< 100	< 100
As, ppm	< 10	< 10	< 10	< 10	< 100	< 100
Pb, ppm	< 20	< 20	< 50	< 50	< 50	< 50
Fe, ppm	< 100	< 100	< 100	< 100	< 500	< 500
Ni, ppm	< 20	< 20	< 50	< 50	< 50	< 50
Sn, ppm	< 10	< 10	< 10	< 10	< 100	< 100
Insoluble in NaCN solution, %	< 0.02	< 0.02	< 0.1	< 0.1	< 0.1	< 0.1

Water-soluble, complex copper and zinc cyanides, as well as mixtures of CuCN and $Zn(CN)_2$ with alkali metal cyanides (containing ca. 23 or 29 % Cu or 23 % Zn), are available for use in electroplating processes.

For transportation, storage, and handling, the same precautions must be observed as for alkali metal cyanides.

Uses and Economic Aspects. The simple cyanides of copper and zinc are used primarily in combination with alkali metal cyanides in bath formulations for the electroplating industry. The main applications are copper, zinc, and brass plating.

Copper(I) cyanide is an important reagent for the introduction of cyano groups into aromatic rings to produce benzonitrile derivatives (Sandmeyer reaction). Furthermore, it is used in the synthesis of insecticides and fungicides [167], in the preparation of special paints for ships, as a catalyst for the polymerization of organic compounds, and for the synthesis of anticancer drugs. Copper and zinc cyanide are used in the production of phthalocyanine dyes.

2.3.3. *Cyanides of Mercury, Lead, Cobalt, and Nickel*

The cyanides of mercury, lead, cobalt, and nickel do not have any major industrial importance. Water-soluble complex cyanides can be obtained by the reaction of the simple cyanides with alkali metal cyanides.

Properties, Preparation, and Uses.

Mercury(II) cyanide [3021-39-4], Hg $(CN)_2$, M_r 252.65, ϱ 4.0 g/cm^3, forms white crystals, which are very soluble in water; it is highly toxic. The substance can be prepared by the reaction of HgO with aqueous HCN. When heated, it decomposes to Hg and $(CN)_2$, and may explode if detonated [165], [168]. The complex salts Na [Hg $(CN)_3$], M_r 301.66, and K_2 [Hg $(CN)_4$] [591-89-9], M_r 382.87, are white crystalline powders which are very soluble in water. Mercury (II) oxide cyanide [2040-54-2], Hg $(CN)_2 \cdot$ HgO, M_r 469.26, is less stable and more sensitive to shock than picric acid [169].

Lead cyanide [13453-58-2], $Pb(CN)_2$, M_r 259.25, can be obtained by the reaction of aqueous solutions of lead salts and alkali metal cyanides in the presence of HCN [2]. In the absence of HCN, oxide cyanides are formed, e.g., $Pb(CN)_2 \cdot 2$ PbO.

Cobalt(II) cyanide [542-84-7], Co $(CN)_2 \cdot 2\ H_2O$, M_r 110.95, *mp* 280 °C (decomp.), ϱ 1.872 g/cm^3, forms red-brown needles which are insoluble in water. It can be obtained by the reaction of aqueous solutions of cobalt salts and alkali metal cyanides, and can be converted to soluble complex compounds with ammonia or excess alkali metal cyanide.

Complex cobalt(III) cyanides are formed when solutions of complex cobalt(II) cyanides are heated in the presence of oxygen:

$$2\,K_4[Co(CN)_6] + H_2O + 1/2\,O_2 \rightarrow 2\,K_3[Co(CN)_6] + 2\,KOH$$

Cobalt cyanides are among the most stable complexes; therefore, cobalt compounds have been used as antidotes for cyanide poisoning [170].

Nickel cyanide [557-19-7], Ni $(CN)_2$, M_r 110.74, *mp* > 200 °C (decomp.), ϱ 2.393 g/cm^3, forms brownish-yellow crystals which are insoluble in water. It can be obtained by the dehydration of Ni $(CN)_2 \cdot 4\ H_2O$, which is precipitated when alkali metal cyanide is added to a solution of a nickel salt.

Potassium tetracyanonickelate(II) monohydrate [14323-41-2], K_2 [Ni $(CN)_4$] $\cdot\ H_2O$, M_r 258.97, forms orange-red crystals which are very soluble in water. The complex salt is formed when excess KCN is employed in the preparation; K_2 [Ni $(CN)_4$] is used as an additive to electrolytes in the plating of gold, silver, and zinc, to improve the quality of the deposits.

2.3.4. *Cyanides of Precious Metals*

The precious metal cyanides can be regarded as the historical starting point of inorganic cyanide chemistry. Their importance for the recovery, refining, and electroplating of precious metals is as great today as it was 100 years ago.

Gold(I) cyanide [506-65-0], AuCN, M_r 222.98, ϱ 7.12 g/cm^3, forms yellow crystals which are insoluble in water and hot dilute acids, but soluble in ether, alkali metal hydroxide, and thiosulfate solutions. Water-soluble [Au $(CN)_2]^-$ complexes are formed when AuCN is dissolved in alkali metal cyanide solution or when gold metal is attacked by alkali metal cyanide in the presence of oxygen or oxidizing agents.

Silver(I) cyanide [506-64-9], AgCN, M_r 133.89, *mp* 320 °C (decomp.), ϱ 3.95 g/cm^3, forms white hexagonal crystals which are insoluble in water but soluble in aqueous alkali metal cyanide solution, forming complex [Ag $(CN)_2]^-$ compounds. Like AgCl, AgCN becomes brown if exposed to light. The production of AgCN is based on the reaction of $AgNO_3$ with alkali metal cyanide in aqueous solution. Its main application is in electroplating.

Palladium(II) cyanide [2035-66-7], $Pd(CN)_2$, M_r 158.7, *mp* 210 °C (decomp.), is soluble in water.

Platinum(II) cyanide [592-06-3], $Pt(CN)_2$, M_r 247.27, is insoluble in water and dilute acids. Both palladium(II) cyanide and platinum(II) cyanide can be prepared from the corresponding chlorides by precipitation with Hg $(CN)_2$. Tetracyano complexes are formed with excess alkali metal cyanide: K_2 [Pd $(CN)_4$] $\cdot$ 3 H_2O forms white rhombohedral fluorescent crystals, and K_2 [Pt $(CN)_4$] $\cdot$ 3 H_2O forms rhombic yellow prisms with blue fluorescence. The barium salt Ba [Pt $CN)_4$] $\cdot$ 4 H_2O is used to prepare the fluorescent coating of X-ray screens; the alkali metal tetracyanoplatinates are used in electroplating. The unstable Pd $(CN)_4$ is also known.

2.4. Cyanide Analysis

The cyanide content of solid alkali metal cyanides, alkaline earth metal cyanides, and dissociating heavy metal cyano complexes such as $[Zn(CN)_4]^{2-}$ may be determined by the argentometric titration of a dilute aqueous solution of the cyanide, according to the method of Liebig – Denigés, in which potassium iodide is used as an indicator [171]:

$$2\,CN^- + Ag^+ \rightarrow [Ag(CN)_2]^-$$

A very accurate indication of the end point of this reaction can be obtained potentiometrically by using a silver – calomel or silver – Thalamide (Schott Glaswerke, Mainz) electrode pair. This method of determination is called the determination of free cyanide and is also an important parameter in the leaching of gold and silver ores.

For more stable cyano complexes or insoluble cyanide compounds [e.g., iron(II) hexacyanoferrate(II)] or in the presence of substances that disturb the argentometric titration, liberation and separation of the cyanide must be carried out before

determination. This can be done by acidifying and boiling the cyanide solution or suspension, and absorbing the released hydrogen cyanide in sodium hydroxide solution [172]. Depending on the concentration, the cyanide content of the absorption solution may be analyzed argentometrically or colorimetrically, e.g., by using pyridine-barbituric acid reagents [173]. This method of determination is called the determination of total cyanide, because all except extremely stable cyanide complexes, such as that of cobalt, liberate cyanide for determination. There is an ISO standard for total cyanide determination (ISO/DIS 6703/1) [174]. The ASTM determination method for total cyanide is also used internationally [175].

To evaluate the toxicity of wastewaters and effluents that contain cyanide, readily liberable cyanide (ISO/DIS 6703/1) is a reliable method for determining less stable complexes of cyanide. Hydrogen cyanide, alkali and alkaline earth metal cyanides, and the cyanides of zinc, cadmium, silver, mercury, copper, and nickel are determined by this method. The complex cyanides of iron, cobalt, gold, and nitriles [174] are not amenable to this method. In this method, hydrogen cyanide is liberated at pH 4 in the presence of zinc metal and ethylenediaminetetraacetic acid. The HCN is transferred to an absorption vessel and analyzed argentometrically or colorimetrically. The ASTM method for the determination of less stable complexes of cyanide is called the weak acid-dissociable cyanide method [175]. Another fast determination method of cyanide analysis, similar to the above two distillation methods, is the colorimetric picric acid method [176]. This method has gained in popularity since the 1970s, it yields very similar results to the above two standard methods but is much quicker. This is important, where effluents are discharged continuously (e.g., gold mine effluent).

Other cyanide determination methods such as the colorimetric method with pyridine – barbituric acid reagents [173] and ion-selective electrode [177] methods for free cyanide are prone to interference by various substances. The ion-chromatographic method has been a subject of research since the 1980s, and techniques have been developed to determine several cyanide species selectively, especially at low concentrations. However, these techniques and instrumentation have not yet achieved widespread application [177].

Quick tests to check the concentration of cyanides in wastewater are based on various colorimetric methods [173], [176]. Generally, the reactions permit detection of very small amounts of free cyanide, but they can be disturbed by various substances, e.g., reducing agents or thiocyanate, sulfide, and nitrite, which all produce similar colors.

[Top of Page]

3. Detoxification of Cyanide-Containing Wastes

Whenever cyanides are manufactured or used, effluents and wastes containing various amounts of cyanide are produced. Because of the high toxicity of cyanide to all forms of life (see Chap. Toxicology and Occupational Health), the effluents and wastes must be treated to reduce the cyanide content to concentrations that are acceptable with the regard to the particular environmental conditions [178-182]. Depending on the quantity and type of cyanidic waste, various detoxification methods are used [183-187]. In addition to effectiveness and cost of treatment, the formation of undesirable byproducts and additional salting of the wastewater are factors of growing importance in choosing an effluent-treatment method.

Heavy metals may be precipitated and separated as hydroxides, carbonates, or sulfides, when complex heavy metal cyanides are oxidized. Wastewater treatment can be carried out batchwise or continuously, and the process may be monitored and controlled automatically.

3.1. Wastewater Treatment

Chlorination. Alkaline chlorination is the most frequently used process for the treatment of effluents containing < 1 g CN^- per liter [188]. In principle, this method allows the destruction of all commercially used simple and complex cyanides, with the exception of complex iron cyanides, which are only attacked above 80 °C. The treatment can be carried out with chlorine and alkali [NaOH, $Ca(OH)_2$] or with readymade hypochlorite solutions that contain about 12 % NaOCl. At first, toxic cyanogen chloride is formed, which is hydrolyzed quickly to cyanate and chloride at pH > 11 [189]:

$$CN^- + OCl^- + H_2O \rightarrow ClCN + 2\,OH^-$$

$$ClCN + 2\,OH^- \rightarrow OCN^- + Cl^- + H_2O$$

Additional salting of the water is the main disadvantage of alkaline chlorination: at least 5 kg of NaCl is produced per kilogram of cyanide for the oxidation of cyanate, and an additional 6.5 kg of NaCl per kilogram of cyanide results from total oxidation [190]. Alkaline chlorination is frequently used for the treatment of smaller quantities of effluent and less for large continuous operations. It is frequently used, for example, in the detoxification of electroplating effluents, when permissible. Lower chloride limits in the treated effluent make an alternative process necessary. Hypochlorite solutions tend to lose their active oxygen content relatively fast.

Hydrogen Peroxide Oxidation. The use of hydrogen peroxide for cyanide detoxification has increased because, in this case, the oxidation of cyanide leads directly to cyanate without the formation of toxic intermediates and byproducts, and the oxidant does not cause additional salting [190], [191]:

$$CN^- + H_2O_2 \rightarrow OCN^- + H_2O$$

The H_2O_2 process is used mainly for the batchwise treatment of effluents from organic nitrile synthesis and metal hardening plants, and can be monitored and controlled automatically [192], [193]. One process allows delivery of accurate amounts of H_2O_2 to continuous effluent streams, which is controlled by on-line measurement of the total oxygen demand [194]; this process is used, for example, in the treatment of tailings from gold mines [177], [195]. More recently, the H_2O_2 process has increasingly found applications in the seasonal detoxification of effluents from gold mines, because of the ease of operation and relatively low investment costs. The H_2O_2 process has been optimized by the addition of compounds that reduce the

quantitiy of H_2O_2 required [196].

Other Methods of Treatment. A method from 1982 is based on the treatment of effluents with a mixture of sulfur dioxide (2.5 %) and air in the presence of small amounts of copper salts (> 50 mg Cu per liter) as catalyst [202]:

$$CN^- + O_2 + SO_2 + H_2O \rightarrow OCN^- + H_2SO_4$$

This is called the Inco process. The detoxification of cyanide-containing pulps is becoming more and more important. In addition to the Inco process method above, the Caro's acid process is also used to create cyanide-containing pulps according to the reaction:

$$CN^- + H_2SO_5 \rightarrow OCN^- + H_2SO_4$$

A further development uses the synergies between the Inco process and the Caro's acid or hydrogen peroxide process to detoxify cyanide-containing slurries more efficiently. This process is called the CombiOx process [203].

In the case of highly concentrated cyanide solutions, the volatilization and catalytic oxidation of HCN after acidification of the solutions can be used [186]. Sometimes, the hydrogen cyanide can be reused by absorption in alkali metal hydroxide solution and conversion to alkali metal cyanide. Atmospheric oxygen is occasionally used for the destruction of cyanide in weak solutions, in combination with activated carbon [97], [198] or microorganisms [199]. Removal of cyanides by ion exchange and reverse osmosis does not solve the problem entirely because the concentrates require further treatment [200], [201].

The hydrolysis of cyanide at 180 – 230 °C under pressure can be used for the destruction of simple and complex cyanides, even in high concentrations [94]. The process does not require chemicals, but the investment for the plant is high.

The conversion of cyanide to thiocyanate by the addition of sulfur is a process that used to be employed but is now regarded more critically [183]. This process has the advantage of low reagent costs, but the investment for plant can be high. It is less suitable for seasonal detoxification, since start up and decommissioning can take time.

Very effective oxidation of cyanide to cyanate, carbonate, and nitrogen takes place with ozone [94], [197]; however, this requires high capital investment and causes problems with the adaptation to varying oxygen demand.

The conversion of cyanide to hexacyanoferrate(II) by reaction with iron(II) salts in alkaline solution followed by precipitation of hexacyanoferrate(II) as iron(II) hexacyanoferrate(II) at pH 3 – 4 is one of the oldest methods for cyanide removal [204]. However, this should be used only for the removal of complex iron cyanides, if at all, because the sludge must be disposed of in such a manner that water resources are protected. Furthermore, the filtrates must be treated again to destroy the remaining traces of cyanide.

In many cases, low concentrations of complex iron cyanides (e.g., < 20 mg of cyanide per liter) in effluents discharged to sewage systems are tolerated because they are precipitated in contact with metal salts (e.g., the iron salts present in domestic wastewater) and are separated with the sludge of municipal sewage treatment plants.

3.2. Solid Wastes

For a long time, disposal of the solid wastes from metal hardening plants, which contain alkali metal cyanides, cyanates, nitrites, and nitrates in addition to other inorganic salts, was a special problem [84], [85]. New processes have been developed that, in many cases, allow regeneration of waste salts. In any case, solid wastes should not be treated or destroyed by a wet chemical process. The safest and most practical solution to the disposal problem at the present time is to deposit them in disused salt mines.

[Top of Page]

4. Cyanogen Halides

Halogen cyanides, usually known as cyanogen halides, have been the subject of considerable interest, and some aspects of their chemistry have been reviewed in detail [2], [6-11]. These compounds, except for cyanogen fluoride, have been known for a long time [205], [206]. Cyanogen fluoride was prepared in 1981 [207], [208]. Depending upon the electronegativity of the halogens relative to the cyanide group, the cyanogen halides can act as halogenating or cyanating agents. The cyanogen halides can be regarded as pseudohalogen or interhalogen compounds with a definite nitrile structure (X–C≡N) [209]. The most important cyanogen halide is cyanogen chloride, which is used for the commercial production of cyanuric chloride, a starting material for pesticides, dyes, and drugs (⟶).

4.1. Properties

Physical Properties. The pure cyanogen halides are colorless, highly volatile, and poisonous compounds. Some physical properties are given in Table 10.

Table 10. Physical properties of the cyanogen halides [10], [11], [209]

	FCN [1495-50-7]	ClCN [506-77-4]	BrCN [506-68-3]	ICN [506-78-5]
	45.02	61.47	105.92	152.92

M_r				
mp, °C	– 82	– 6.5	52	148 *
bp, °C	– 46.2	12.5	61	subl. > 45 *
Density ϱ, g/cm^3		1.24 (0 °C)	2.01 (18 °C)	2.84 (18 °C)
		1.19 (15 °C)	1.18 (54 °C)	
Bond length, pm				
C–Hal	126.0	162.9	179.0	199.5
C–N	116.5	116.3	115.8	115.9
Enthalpy of vaporization ΔH_v, kJ/mol	22.39	26.75	33.82	40.02
Enthalpy of fusion ΔH_f, kJ/mol		11.39		19.88
Enthalpy of sublimation ΔH_s, kJ/mol			45.21	59.90
Dipole moment μ, D	2.17	2.80	2.94	3.71

* Triple point, *p* = 132.1 kPa; subl. = sublimation.

Chemical Properties. Pure cyanogen halides are stable at normal temperature. Impurities catalyze the exothermic trimerization to the corresponding cyanuric halides. Catalysts are proton acids, Lewis acids (e.g., heavy metal salts), and bases. Relatively little is known about the preparative use of FCN [208], which explodes at – 41 °C on ignition. Mixtures of FCN and air are more explosive than mixtures of acetylene and air [210]. In a catalytic reaction, ClCN forms not only the trimer (cyanuric chloride [108-77-0]) but also a stable tetramer, 2,4-dichloro-6-isocyanodichloro-*s*-triazine [877-83-8] [211].

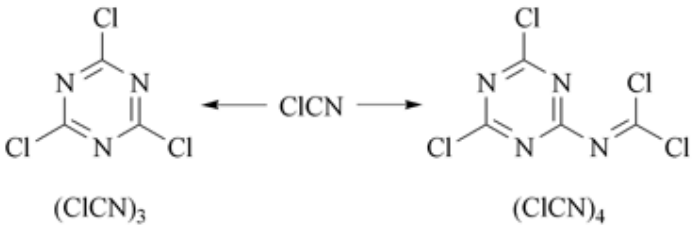

Cyanogen bromide polymerizes more readily to cyanuric bromide [14921-00-7] than does ClCN to cyanuric chloride [212].

When heated to 130 °C in a sealed tube, cyanogen iodide does not polymerize [213]. Above 150 °C, ICN eliminates iodine:

$$2\,ICN \rightarrow (CN)_2 + I_2$$

The reaction is not reversible [214].

Cyanogen chloride is an intermediate in the detoxification of cyanide-containing wastewater with sodium hypochlorite solutions; therefore, the cyanogen halide must be removed immediately by appropriate hydrolysis [215]. Alkaline hydrolysis of cyanogen halides proceeds in different ways:

$$XCN + 2\,NaOH \rightarrow NaOCN + NaX + H_2O (X = Cl\,, Br)$$

$$ICN + 2\,NaOH \rightarrow [NaOI] + NaCN + H_2O$$

Thus, ClCN and BrCN are hydrolyzed to the cyanate and halide salts; ICN, however, forms hypoiodite, which disproportionates to iodide and iodate. This is due to the more electropositive character of iodine relative to chlorine [11]. Therefore, ClCN can be used for the electrophilic introduction of a cyano group into a substrate; use of ICN results in electrophilic iodination. The reaction pathway of BrCN often depends on the reaction conditions and on the nature of the substrate [216]. Cyanogen halides add to olefins and acetylenes in the presence of acid catalysts [217], [218]. Aromatic ring systems can be cyanated by ClCN or BrCN under Friedel – Crafts conditions [209], [220]. The reactions of cyanogen halides, particularly ClCN and BrCN, with nitrogen, oxygen, and sulfur nucleophiles lead to cyanamides [221-223] cyanates [224], [225], and thiocyanates [226].

In some cases, the newly formed compound undergoes further reactions at the C≡N triple bond to form guanidines [221], [222], ureas, or heterocyclic ring systems [227]. Cyanogen bromide reacts with tertiary alkylamines to form ammonium bromides, which decompose to give dialkyl cyanamides and alkyl bromides (von Braun reaction) [228], [229]. Reaction of sulfur trioxide with ClCN leads to the very reactive chlorosulfonyl isocyanate [230], [231]. Cyanogen halides undergo radical reactions at elevated temperatures on irradiation with UV light or in the presence of peroxide promoters. Thus, cyanogen chloride and acetonitrile are converted to malonodinitrile [232-235]. In the gas phase, aromatic methyl compounds and cyanogen chloride react to give arylacetonitriles [236].

4.2. Production

Cyanogen fluoride is prepared by the reaction of tetracyanomethane with cesium fluoride [237]:

$$C(CN)_4 + CsF \rightarrow FCN + CsC(CN)_3$$

The pyrolysis of cyanuric fluoride at 1300 °C under reduced pressure gives FCN in ca. 50 % yield [207], [208], [210].

The preparation of the industrially important ClCN has been developed extensively; the synthetic pathways are as follows:

1. Electrolysis of an aqueous solution of HCN and NH_4Cl [238]
2. Reaction of complex cyanide salts (e.g., Na_2 [Zn $(CN)_4$]) with chlorine below 20 °C [239]
3. Formation of ClCN from cyanide salts (mostly NaCN) and chlorine in an exothermic reaction; in a continuous process, sprayed aqueous NaCN solution is contacted with chlorine, and the reaction heat evaporates the ClCN [240].
4. Processes involving hydrogen cyanide and chlorine as the most convenient starting materials; the reaction is carried out in aqueous solution [241-246] (for a detailed procedure, see also ⟶) in organic [247] and inorganic solvents [248], and in the gas phase [249], [250]. To avoid the byproduct HCl, attempts have been made to reoxidize HCl to chlorine with oxygen [245] and hydrogen peroxide [251], [252] in catalyzed reactions.
5. Chlorinolysis of cyanogen in the gas phase at 300 – 600 °C in the presence of a catalyst [253], [254]
6. Pyrolysis of cyanuric chloride at 600 – 900 °C in the presence of a charcoal catalyst [255]
7. High-temperature syntheses based on elemental chlorine, nitrogen, and carbon [256]

Usually, these processes lead to impure ClCN; therefore, in many cases the product must be purified before further use. Water can be removed by treatment with calcium salts or molecular sieves and by fractional distillation [257]. Chlorine is stripped with water (mixed with HCN and HCl) [258] or with solutions containing a soluble iron(II) salt and an insoluble carbonate [259]; chlorine may also be eliminated by treatment with γ-aluminum oxide [260]. Cyanogen chloride, free of HCl, is obtained by treatment with water [258] or with organic solvents [261], [262]. Hydrogen cyanide is removed by distillation with chlorine [263] or by washing with $FeSO_4$ solutions [264].

Cyanogen bromide can be prepared from bromine and cyanide salts or hydrogen cyanide [265], [266]. Cyanogen iodide is synthesized analogously [267], [268].

4.3. Storage and Transportation

In most cases, ClCN is used in the gas phase immediately after preparation. Relatively small amounts are condensed and stored in gas containers as liquids. Steel cylinders, which must meet specific requirements in each country, can be used for shipment. The condensed and bottled ClCN must be very pure; moreover, it must be mixed with a stabilizing agent (generally sodium pyrophosphate) to inhibit exothermic polymerization caused by impurities [269-271]. Impure BrCN can polymerize during storage [272]; the danger of explosion exists with closed BrCN bottles [273]. Cyanogen halides should be handled carefully because they are very toxic (similar to HCN) and have a strong lacrimatory effect.

4.4. Uses

Most of the cyanogen chloride is used for the production of cyanuric chloride. The reaction with amines leads to diphenylguanidines (vulcanization accelerators) [274]. Cyanogen chloride and bisphenols react to give cyanate esters, which can polymerize to polytriazine resins [224], [225].

Halogen cyanides and cyanamide react to form dicyanamide [504-66-5] [275], [276], which is used for the preparation of pharmaceutical bisbiguanides. The industrial production of malonodinitrile [109-77-3] [232-235], and chlorosulfonyl isocyanate [1189-71-5] [230], [231] based on cyanogen chloride has become increasingly important. Cyanogen chloride is the most economically significant of all cyanogen halides. Its worldwide production capacity is more than 230 000 t/a (mainly in Europe and the United States).

Cyanogen bromide and cyanogen iodide are commonly used in the laboratory, e.g., for the dealkylation of tertiary amines [228], [229], the selective cleavage of peptides [277], and the preparation of heterocyclic compounds.

[Top of Page]

5. Cyanogen

The first synthesis of cyanogen [460-19-5], dicyanogen, oxalonitrile, $(CN)_2$, M_r 52.05, was carried out by GAY-LUSSAC in 1815 by pyrolysis of silver cyanide [264]. The chemistry of this reactive compound is discussed in [1], [12-17]. In many respects, cyanogen can be compared to the halogens. In addition, many of its properties are a result of the reactivity of the cyano groups. Although various procedures for the synthesis of cyanogen are known, commercial interest in cyanogen is limited.

5.1. Properties [1], [17]

Physical Properties. At room temperature, cyanogen is a colorless, flammable, and very poisonous gas with a pungent odor. Some important physical properties are listed below:

Melting point	–27.98 °C
Boiling point	–21.15 °C
Critical pressure, $p_{crit.}$	60.79 bar
Critical temperature, $t_{crit.}$	128.3 °C
Density	2.321 g/L at 0 °C
	1.25 g/cm^3 at –95 °C

0.954 g/cm^3 at –21.17 °C

Vapor pressure

t, °C	–21.15	–4.88	20.88	44.43	72.40
p, bar	1	2	5	10	20

Solubility at 20 °C: water dissolves 4.5 times its own volume of cyanogen; diethyl ether, 5.0 times; and ethanol, 23.0 times.

Bond length:	
C–C	138 pm
C–N	113 pm
Bond angle: C–C–N	179° 38′
Enthalpy of fusion, ΔH_f	8.112 kJ/mol
Enthalpy of vaporization	23.341 kJ/mol
MAK	10 ppm

The values reported for the ignition limits of cyanogen – air mixtures vary widely, depending on experimental conditions, e.g., with humid air (1.7 % H_2O); the reported values are 6.4 – 7.25 vol % for the lower limit, and 26.2 – 30.6 vol % for the upper limit.

Chemical Properties. Cyanogen has a high thermal stability; nevertheless at high temperatures (and when irradiated with UV light), polymerization to paracyanogen occurs [17], [279], [280]. With stoichiometric amounts of air or oxygen, cyanogen burns with the hottest flame (ca. 5000 K) ever observed in chemical reactions [281], [282].

Cyanogen undergoes reactions which depend either on its properties as a pseudohalogen or on the reactivity of the C≡N triple bond [1], [15], [17]:

1. Hydrogenation of cyanogen at 550 – 675 °C gives hydrogen cyanide [283].
2. At 300 – 500 °C, chlorine and cyanogen react to give cyanogen chloride and, subsequently, cyanuric chloride in the presence of an active charcoal catalyst [284], [285].
3. The direct introduction of a cyano group into an aryl or alkyl compound proceeds via a radical- or heat-initiated reaction [286-289].
4. Two moles of hydrogen cyanide add to cyanogen to form diiminosuccinonitrile [290].
5. Cycloadditions (Diels – Alder) with organic dienes give cyanopyridines [291].
6. Alkaline hydrolysis leads to cyanide and cyanate [292].
7. The acid-catalyzed hydrolysis of cyanogen first gives cyanoformamide [293], [294] and then oxamide and oxalic acid. Oxamide [471-46-5] is a nitrogen fertilizer with good depot activity. Its industrial preparation has been published in several papers; most of the synthetic paths use the isolated $(CN)_2$ as an intermediate [295-300], but in situ procedures are also known [301].

Heterocyclic compounds are formed by the reaction of cyanogen and disulfur dichloride [302] or sulfur [303]:

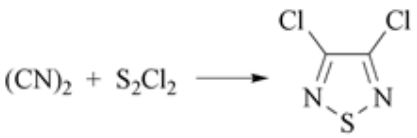

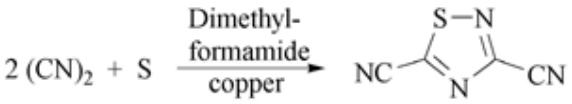

5.2. Production

Numerous synthetic pathways to cyanogen have been devised [1], [13], [15], [17]. In the laboratory, the compound is normally prepared via cyanide salts; in this case, the reaction of alkali metal cyanides with copper(II) salts is more common than the thermal decomposition of $Hg(CN)_2$ or AgCN [17].

$$CuSO_4 + 2\,KCN \rightarrow Cu(CN)_2 + K_2SO_4$$

$$2\,Cu(Cu)_2 \rightarrow (CN)_2 + 2\,CuCN$$

However, the oxidation of hydrogen cyanide in the presence of a catalyst has greater industrial interest. These reactions are based on the principle of JACQUEMIN [304], that is the conversion of cyanides, by treatment with copper salts, to cyanogen in aqueous solution. In a continuous process, the copper(I) salt formed must be reoxidized rapidly to recycle the catalyst. Reoxidation of copper(I) to copper(II) can be carried out directly with air or oxygen [305], [306] or with NO_2 [307] the NO formed in the second case can then be reoxidized with air or oxygen [308], [309]. The oxidation of hydrogen cyanide to cyanogen in nonaqueous solutions has also been achieved [310], [311]. Hydrogen peroxide is another reoxidation agent [312-315].

The continuous production of $(CN)_2$ from HCN and O_2 or H_2O_2 is utilized industrially on a small scale.

In addition to the liquid-phase oxidation of HCN, various gas-phase oxidations have been investigated [316-321].

Cyanogen iodide decomposes to cyanogen and iodine (Chap. Cyanogen Halides). In 1984 the reaction of cyanogen halides with trimethylsilyl cyanide in the presence of Lewis acids has been described [322].

Cyanogen prepared by the above procedures contains impurities such as O_2, N_2, NO, CO_2, or ClCN. Its purification can be achieved by fractional vaporization of the crude material [323] or by scrubbing the gas with aqueous H_2O_2 [324].

5.3. Storage, Transportation, and Uses

Storage and Transportation. The flammability and toxicity of cyanogen necessitate special storing and shipping conditions. Cyanogen should be stored in cool, well-ventilated locations, not near to flammable goods or open flames. Dry and pure $(CN)_2$ can be stored in steel cylinders under pressure and at normal temperatures [325]. Comprehensive studies have been made on the stability of pure cyanogen to heat, pressure, chemical additions, and severe mechanical shock [326]. Shipment must be in accordance with governmental regulations in each country.

Uses and Economic Value. Cyanogen is used mainly in the laboratory. In small quantities, cyanogen is used industrially for drug synthesis. The preparation of oxamide (a slow-release fertilizer) has been carried out on a small scale [285]. Large-scale industrial applications are not yet known.

[Top of Page]

6. Toxicology and Occupational Health

Toxicity and Cyano Compounds. Hydrogen cyanide and other cyano compounds that can form HCN or free cyanide ions are highly toxic to most forms of life. The lethal dose depends on a number of factors, and differs from species to species. A useful survey on this subject can be found in [181]. As shown in this report, the toxicity of metal cyanide complexes is a function of the bond strength between the metal atoms and the cyanide ligands. Toxicity decreases with increasing bond strength, and complexes such as hexacyanocobalt(II) are nontoxic. Hexacyanoferrates(II) and (III) also have very low toxicities [327]. These differences in toxicity are particularly important when the acceptable concentrations of cyanides in wastewater are being considered. Some of the lethal concentrations are shown in Table 11. The LD_{50} values of complex hexacyanoferrates (rat, oral) are as follows:

$Ca_2\,[Fe\,(CN)_6\,] \cdot 11\,H_2O$ > 5110 mg/kg

$Na_4\,[Fe\,(CN)_6\,] \cdot 10\,H_2O$ > 5110 mg/kg

$K_4\,[Fe\,(CN)_6\,] \cdot 3\,H_2O$ 3613 mg/kg

$K_3\,[Fe\,(CN)_6\,]$ 3503 mg/kg

Table 11. Lethal concentrations of some cyanides (in mg/L)

Substance	Species		
	Fish	*Daphnia magna*	*Escherichia coli*
NaCN	0.05 (5 d) [328]	3.4 (48 h) [329]	0.0004 – 0.1 [330]
NaCN	0.3 – 0.7 (96 h) [331]		
NaCN	1.0 (20 months) [332], [333]		
$K_3[Cu\,(CN)_4\,]$ [334]	1.0	0.8	2500
CuCN	2.2 [179]		
$K_2[Zn\,(CN)_4]$ [334]	0.3	13.5	625

$Zn(CN)_2$	0.2 – 0.3		
$K_2[Cd(CN)_4]$ [334]	0.75	0.5	250
$Cd(CN)_2$	0.17 [179]		
$K_2[Ni(CN)_6]$ [334]	30	75	5000
$Na_4[Fe(CN)_6]$	500 [179]	> 600 [336]	> 1000 [335]

Poisoning by Hydrogen Cyanide and Alkali Metal Cyanides. Hydrogen cyanide is easily absorbed, even through intact skin and mucous membranes. Not even full respiratory protection can prevent hydrogen cyanide from penetrating into the organism. The speed of absorption of orally ingested cyanide depends on the contents of the stomach and gastric pH; empty stomach and low pH lead to rapid absorption of cyanide. Symptoms [337] of the poisoning may vary somewhat, depending on the quantity of hydrogen cyanide absorbed or the quantity of cyanide ingested. Low concentrations of hydrogen cyanide irritate the nasopharyngeal cavity; occasional complaints involve headache, anxiety, and nausea. Rapid inhalation of the lethal dose (ca. 50 – 100 mg) of hydrogen cyanide results in collapse of the individual. This apoplectiform syndrome, characterized by a short spasmodic stage with subsequent respiratory paralysis, results in immediate death. Poisoning symptoms following oral ingestion of cyanides, on the other hand, exhibit a slow progress, even with large doses (150 – 250 mg), because of the slow release of hydrogen cyanide. Case histories described in the literature involve poisoning durations up to several hours. Cyanide affects many enzymes. The inhibition of cytochrome c oxidase is responsible for the acute effects and the marked metabolic changes (lactic acidosis). As in other intoxication, treatment includes the support of vital signs (supportive therapy), prevention of further poison absorption, and enhancement of poison elimination. The administration of antidotes in severe cyanide poisoning is mandatory, although this opinion has been contradicted.

The following concentrations of hydrogen cyanide and other cyanides in air are considered acceptable at the workplace:

TLV–TWA	(hydrogen cyanide)	10 mg/m^3 [338]
MAK	(hydrogen cyanide)	11 mg/m^3 [339]
TLV–TWA	(cyanide dust, as CN)	5 mg/m^3 [338]
MAK	(cyanide dust, as CN)	5 mg/m^3 [339]

Therapy for Poisoning by Hydrogen Cyanide and Other Cyanides. Different antidotes or combination of antidotes are preferred in different countries. The mechanisms of antidotal action follow different principles:

1. Direct complexation of the cyanide and excretion as cyanide – antidote complex (hydroxocobalamin, dicobalt edetate)
2. Methemoglobin-forming compounds which shift the cyanide from the intracellular to the blood compartment [amyl nitrite, sodium nitrite, 4-dimethylaminophenol (4-DMAP)]
3. Sulfate donors, which enhance the natural pathway of detoxification (thiosulfate).

The first aid given to victims of hydrogen cyanide poisoning is not fundamentally different from the measures taken in other acute emergencies. The success of the treatment, however, depends on the immediate start of antidote therapy. Seconds may determine the success or failure of the treatment of cyanide poisoning. The therapeutic measures required depend on the particular poisoning syndrome encountered. The initial treatment is of supreme importance and must be aimed at safeguarding vital functions such as respiration and circulation.

Cyanogen Chloride, Cyanogen Bromide [337]. Because of their high volatility, cyanogen chloride and cyanogen bromide are strong irritants to mucous membranes (eyes, lungs). Their effects are essentially identical to those produced by hydrogen cyanide; however, they are rarely encountered because the irritating effects serve as a warning of potential danger. When handling liquid cyanogen chloride, attention must be paid to the potential for skin absorption. The exposure limit of cyanogen chloride (TLV – TWA) is 0.6 mg/m^3 [338]. The therapy is analogous to that for hydrogen cyanide poisoning.

Cyanogen [337]. Cyanogen irritates mucous membranes and is probably hydrolyzed in the organism to hydrogen cyanide and cyanic acid. The symptoms are similar to those produced by hydrogen cyanide poisoning.

TLV – TWA	20 mg/m^3 [338]
MAK	22 mg/m^3 [339]

The therapy is analogous to that for hydrogen cyanide poisoning.

[Top of Page]

7. Responsible Care and Product Stewardship

Responsible application is achieved by minimizing losses of reagent and avoiding negative impact on the environment as well as adopting a Responsible Care philosophy to the application of cyanide. This is a complex system with a greater commitment than simply avoiding reagent losses for economic purposes. The purpose of Responsible Care is to increase the

awareness of people involved in cyanide production, transportation, storage, and application to prevent any incidents/accidents. The economic losses which can arise by not adopting a suitable safety program have long been realized by practically all industries. Important factors influencing safety behavior are

- Responsible Care philosophy
- Internal safety program
- Legislation
- Common sense
- Individual behavior

These factors can interact and overlap, and one factor may overrule another, for example, a company safety program may be more stringent than local legislation.

Risks involved in handling cyanide can be categorized into intentional and unintentional misuse. Intentional misuse (e.g., the use of cyanide in fishing) can only be prevented by strict control of the end user or by implementing strict laws. Unintentional misuse, meaning any accident or incident occurring due to the lack of knowledge about how to handle cyanides in a proper manner, or even more important due to an insufficient overview of possible risk situations that might arise, can be avoided by education and increasing the awareness of how to handle this toxic chemical. For example, the formation of hydrocyanic acid is a potential risk in all cyanide-using industries if cyanides are transported together with acids on a truck.

All cyanides and cyanide complexes should be treated with caution. Even stable cyanide complexes like ferro- and ferricyanides, which are relatively nontoxic, can potentially lead to a generation of hydrocyanic acid if they come into contact with strong acids, or even weaker acids over a long time period.

Concentration is also very important, that is, it is important to differentiate between an absolute amount of cyanide that can cause a serious incident and amounts that do not potentially lead to a serious incident.

The risk of a cyanide accident with acids in the gold mining industry is relatively low compared to other industries, since there is a lower risk of acid contact with cyanides. In the electroplating industry it is common practice to use acid baths in combination with cyanide baths. These electroplating lines are sometimes in the same area, without separation (e.g., by a wall). Even more important is that cyanides and acids might be transported to the site on the same truck, and in the case of a road accident HCN could be formed. In the metal-hardening industry, high temperatures are applied, for example, in the Durferrit process. The higher temperatures also represent a potentially higher risk.

A safety cycle describes how the responsibilities in the use of cyanide can be shared between the cyanide user and the cyanide manufacturer. Safe and environmentally friendly production, packaging, and transport are the responsibilities of the cyanide manufacturer. The cyanide user should also take responsibility for that part of the transport which is no longer in the hands or control of the cyanide manufacturer. In addition, the cyanide user should provide safe storage and handling areas and for responsible disposal of the cyanide and of the packing. The cyanide manufacturer should provide the user with the necessary information to enable him to implement his share of the responsibility. Cyanide manufacturers should perform site survey and safety training for their customers to guarantee a dynamic process and eliminate any weak points with regard to Responsible Care. The following are the main points of interest in such a site survey or safety training:

- Storage
- Safe handling
- Environment
- Transport
- Emergency procedure
- Detoxification; responsible disposal

The management of the cyanide user should be open to any safety improvement feedback from the cyanide personnel and also provide regular safety training. An intact pre-emergency concept must be guaranteed, and the functioning of this concept tried out regularly to eliminate any weak points in communication. The most important point is that a medical doctor should be available immediately in the event of an incident.

The cyanide manufacturer must provide the user with the necessary safety information, so that the user can implement safety measures and carry out his own internal safety training and surveys.

It is recommended that the following material is provided to the responsible personnel:

- Pocket cards with instructions for safe handling
- Medical brochures with recommendations for the medical practitioners
- Placards and signs for cyanide areas
- Video with recommendations on safe handling
- Transport Cards for transport personnel

[Top of Page]

References

General Reference

1. *Gmelin*, **System no. 14, D 1, Kohlenstoff** (1971).

2. H. E. Williams: *Cyanogen Compounds*, 2nd ed., Arnold & Co., London 1948.
3. M. H. Ford-Smith: *The Chemistry of Complex Cyanides*, Dep. of Scientific and Industrial Research, Her Majesty's Stationery Office, London 1964.
4. A. G. Sharpe: *The Chemistry of Cyano Complexes of the Transition Metals*, Academic Press, London 1976.
5. A. M. Golub *et al.*: *Chemie der Pseudohalogenide*, Hüthig Verlag, Heidelberg 1979.
6. V. Migrdichian: *The Chemistry of Organic Cyanogen Compounds*, Reinhold Publ. Co., New York 1947, pp. 97 – 124.
7. *Houben-Weyl*, **8**, 90 – 93.
8. G. Brauer: *Handbuch der präparativen anorganischen Chemie*, 2nd ed., vol. 1, Enke Verlag, Stuttgart 1960.
9. B. S. Thyagarajan: *The Chemistry of the Cyanogen Halides*, Intra-Science Chemistry Reports, vol. 2, no. 1, Santa Monica, California 1968, pp. 1 – 89.
10. J. MacCordick, *Bull. Soc. Chim. Fr.* 1973, 50 – 70.
11. *Gmelin*, **System no. 14, D3, Kohlenstoff** (1976) 149 – 266
12. *Houben-Weyl*, **8**, 256 – 265.
13. T. K. Brotherton, J. W. Lynn, *Chem. Rev.* **59** (1959) 841 – 883. Links
14. S. Nakamura, *Yuki Gosei Kagaku Kyokai shi* **26** (1968) 23 – 26.
15. R. H. Fahnenstich, W. Th. Heimberger, F. M. Theissen, W. M. Weigert, *Chem. Ztg.* **96** (1972) 388 – 396.
16. G. Brauer: *Handbuch der präparativen anorganischen Chemie*, 3rd ed., vol. **2**, Enke Verlag, Stuttgart 1978.
17. H. W. Roesky, H. Hofmann, *Chem. Ztg.* **108** (1984) 231 – 238.

Specific Reference

18. H. Bauer, *Naturwissenschaften* **67** (1980) 1 – 6. Links
19. W. F. Giauque, R. A. Ruehrwein, *J. Am. Chem. Soc.* 1939, 2626 – 2633. Links
20. G. E. Coates, J. E. Coates, *J. Chem. Soc.* 1944, 77 – 81.
21. G. Bredig, L. Teichmann, *Z. Elektrochem.* **31** (1925) no. 8, 449 – 454.
22. G. Kortüm, H. Reber, *Z. Elektrochem.* **65** (1961) no. 9, 809 – 812.
23. B. Grüttner, M. F. A. Dove, A. F. Clifford: *Chemistry in Anhydrous, Prototropic Inorganic Solvents*, vol. **II**, part 1, Pergamon Press, Oxford, London, Edinburgh, New York, Toronto, 1971.
24. T. Völker, *Angew. Chem.* **72** (1960) 379 – 384.
25. Röhm & Haas, DE 1 020 320, 1956 (T. Völker, H. Zima).
26. W. Ruske, *Wiss. Z. Humboldt Univ. Berlin. Math. Naturwiss. Reihe* **8** (1958) 557 – 572.
27. J. P. Ferris, W. J. Hagan, *Tetrahedron* **40** (1984) no. 7, 1093 – 1120. Links
28. *Ullmann*, 3rd ed., **5**, 629
29. H. Schaefer, *Chem.-Tech. Heidelberg* **7** (1978) 231 – 237.
30. F. Götzen, *Hüttenwesen Symposium Int. Gas Union Karlsruhe* **12** (1972) 1 – 15; *Chem. Abstr.* **81** (1974) 65 808 m.
31. *Ullmann*, 3rd ed., **5**, 631 – 633.
32. Standard Oil, US 4 423 023, 1983 (L. J. Velenyi, H. F. Hardman, F. A. Pesa).
33. DuPont, US 4 164 552, 1979 (F. J. Weigert).
34. V. D. Sokolovskii, Z. G. Osipova, S. Y. Burylin, *React. Kinet. Catal. Lett.* **15** (1980) no. 3, 311 – 314.
35. Nitto Chem. Ind., EP-A 0 089 118, 1983 (Y. Sasaki, H. Utsumi).
36. Monsanto, US 4 457 904, 1984 (J. W. Gambell, S. R. Anvie).
37. Standard Oil, US 4 423 023, 1983 (J. L. Velenyi, H. F. Hardman, F. A. Pesa).
38. Monsanto, US 4 425 260, 1984 (J. R. Ebner).
39. BASF, DE 1 143 497, 1963 (G. Schulze, G. Weiß).
40. Sumitomo, DE-OS 2 350 212, 1974 (T. Shiraishi, H. Ichihashi, F. Kato, E. Niihama).
41. Distillers Ltd., GB 913 836, 1962.
42. Proceedings of TOCAT 3 in press.
43. H. Bockhorn, R. Coy, F. Fetting, W. Prätorius, *Chem. Ing. Tech.* **49** (1977) no. 11, 883 – 895.
44. DuPont, WO 95/21126, 1995 (R. J. Riegert, M. Mehdizadeh).
45. L. Andrussow, *Angew. Chem.* **48** (1935) no. 37, 593 – 604.
46. C. T. Kautter, W. Leitenberger, *Chem. Ing. Tech.* **25** (1953) no. 12, 697 – 701.
47. L. Andrussow, *Chem. Ing. Tech.* **27** (1955) 469 – 474.
48. DuPont, US 4 128 622, 1978 (K. D. Loos, K. C. McCullough, D. Q. Whitworth).
49. Johnson Matthey & Co., US 4 469 666, 1984 (D. J. Stephenson, A. E. Heywood, G. L. Selman).
50. W. R. Rolingson, J. McPherson, P. D. Montgomery, B. L. Williams, *J. Chem. Eng. Data* **5** (1960) no. 3, 349 – 351. Links
51. N. Kalkert, H.-G. Schecker, *Chem. Ing. Tech.* **51** (1979) no. 9, 895.
52. Freeport Sulphur, US 2 590 146, 1952 (G. Barsky).
53. DOW Chemical, US 2 899 274, 1959 (R. A. Smith, H. H. McClure, G. C. Bond).
54. Toyo Koatsu Industries, US 3 112 177, 1968 (S. Fujise, N. Nagai, T. Numata).
55. DuPont, US 2 797 148, 1957 (H. C. Carlson).
56. DuPont, US 3 914 386, 1974 (N. R. Wilson).

57. DuPont, EP-A 0 030 141, 1983 (G. Levitt).
58. F. Endter, *DECHEMA Monogr.* **33** (1959).
59. F. Endter, *Chem. Ing. Tech.* **30** (1958) no. 5, 305 – 310.
60. DEGUSSA, DE 959 364, 1954 (F. Endter).
61. DEGUSSA, DE 882 985, 1953 (R. Wendlandt, G. Hoffmann, E. Kokert).
62. DEGUSSA, DE 1 007 300, 1957 (F. Endter).
63. DEGUSSA, DE 3 923 034, 1989 (F. Bittner, C. Hecht, P. Lok-Herber, P. Panster).
64. DEGUSSA, DE 19 617 040, 1998 (L. v. Hippel, C. Bussek, J. Sauer, M. Sauer, D. Arntz).
65. E. Koberstein, *Ind. Eng. Chem. Process. Des. Dev.* **12** (1973) no. 4, 444 – 448.
66. DEGUSSA, DE-OS 33 09 394, 1983 (R. Manner, H. Schaefer, C. Voigt, W. D. Pfeifer).
67. DEGUSSA, DE 31 34 851, 1982 (H. Bruck).
68. DEGUSSA, EP-A 0 072 416, 1981 (C. Voigt, P. Kleinschmit).
69. DEGUSSA, US 4 289 741, 1981 (C. Voigt, P. Kleinschmit, E. Walter).
70. DEGUSSA, US 4 387 081, 1983 (C. Voigt, P. Kleinschmit, G. Schreyer, G. Sperka).
71. N. B. Shine, *Chem. Eng. Prog.* **67** (1971) Links
72. Shawinigan Chem., CA 573 348, 1959 (H. S. Johnson, A. H. Andersen).
73. Shawinigan Chem., US 3 097 921, 1963 (D. J. Kennedy, N. B. Shine).
74. Shawinigan Chem., US 3 032 396, 1962 (D. J. Kennedy).
75. *Hommel-Handbuch der gefährlichen Güter*, 4. Lfg., "Merkblatt 42", Springer Verlag, Berlin, Heidelberg, New York 1980.
76. R. Baxter, *Ind. Finish. Wheaton Ill.* **53** (1977) no. 12, 38 – 41.
77. J. E. Johnson, E. R. Poor, US 2 753 248, 1956.
78. Compur Electronic GmbH, Steinerstraße 15, 8000 München 70, FRG.
79. H. B. Northrup, *Trans. Am. Electrochem. Soc.* **60** (1931) 49 – 53.
80. DEGUSSA, DE 1 149 035, 1963 (C. Albrecht, J. Müller).
81. DEGUSSA, DE 1 178 668, 1964.
82. L. Hackspill, R. Grandadam, *C.R. Acad. Sci.* **180** (1925) 931.
83. L. Hackspill, R. Grandadam, *Ann. Chim. (Paris)* **5** (1926) no. 10, 237.
84. W. Müller, L. Witzke, *Chem. Ing. Tech.* **45** (1973) no. 22, 1285 – 1289.
85. H. Beyer, G. Dittrich, H. Kunst, F. Schmidt, *Chem. Anlagen + Verfahren* 1973, no. 11, 69 – 75.
86. R. W. E. B. Harman, F. P. Worky, *Trans. Faraday. Soc.* **20** (1925) 502; *Chem. Abstr.* **19** (1925) 923.
87. B. Ricca, G. D'Amore, *Gazz. Chim. Ital.* **79** (1949) 308 – 322; *Chem. Abstr.* **43** (1949) 8821.
88. E. Deltombe, M. Pourbaix, *Proc. Meet. Int. Comm. Electrochem. Thermodyn. Kinet. 6th* 1955, 138 – 142; *Chem Abstr.* 1956, 6896.
89. J. Zawidski, T. Witkowski, *Rocz. Chem.* **5** (1925) 515 – 528; *Chem. Abstr.* 1926, 3258.
90. G. W. Wiegand, M. Tremelling, *J. Org. Chem.* **37** (1972) 914 – 916. Links
91. DuPont, US 2 773 752, 1953 (V. W. Kremer, C. H. Lemke).
92. H. Sulzer, *Z. Angew. Chem.* **25** (1912) 1268 – 1273.
93. G. W. Heisse, H. E. Foote, *Ind. Eng. Chem.* **12** (1920) 331 – 336.
94. J. Hoerth, U. Schindewolf, W. Zbinden, *Chem. Ing. Tech.* **45** (1973) 641 – 646.
95. E. v. Papp, J. Pogany, *Z. Angew. Chem.* **54** (1941) 55.
96. W. Bucksteeg, *Purdue Univ. Eng. Bull. Ext. Ser.* **121** (1966) 688 – 695.
97. D. C. Hoffmann, *Plating (East Orange, NJ)* **60** (1973) 157 – 161.
98. M. R. Lowndes, *Chem. Ind.* **34** (1971) 951 – 956.
99. C. Fabjan, *Galvanotechnik* **66** (1975) 100 – 107.
100. DEGUSSA, DE 742 074, 1943 (H. Beier).
101. J. P. Zumbrunn, *Inf. Chim.* **124** (1973) 189 – 194.
102. H. Stamm, *Z. Angew. Chem.* **47** (1934) 579, 791 – 795.
103. A. T. Kuhn, *J. Appl. Chem. Biotechnol.* **21** (1971) no. 2, 29 – 34.
104. E. Schulek, P. Endröi, *Anal. Chim. Acta* **5** (1951) 252. Links
105. P. D. Bartlett, R. E. Davis, *J. Am. Chem. Soc.* **80** (1958) 2513 – 2516. Links
106. K. R. Leininger, J. Westley, *J. Biol. Chem.* **243** (1968) 1892 – 1899. Links
107. L. G. Sillen, A. E. Mostell: *Stability Constants of Metal-Ion Complexes*, HMSO, London 1964.
108. J. Bailar: *The Chemistry of the Coordination Compounds*, Reinhold Publ. Co., New York 1956.
109. E. Rabald: *Dechema-Werkstoff-Tabelle*, 3rd ed., part 12, 1963.
110. *Ullmann*, 4 th ed., **8**, 253.
111. V. Migrdichian: *The Chemistry of Organic Cyanogen Compounds*, Reinhold Publ. Co., New York 1947.
112. *Houben-Weyl*, **8**, 247 – 358.
113. F. and E. Rodgers, *Justus Liebigs Ann. Chem.* **41** (1942) 285.
114. GB 12 219, 1894 (H. J. Castner); GB 21 732, 1894; DE 90 999, 1894.

115. DEGUSSA, DE 124 977, 1900; DE 126 241, 1900; DE 148 045, 1901; DE 148 046, 1901; DE 149 678, 1901; GB 265 639, 1926 (H. Freudenberg); DE 851 056, 1943 (W. Stephan).
116. *Ullmann*, 3rd ed., **5**, 643 – 644.
117. DEGUSSA, US 716 350, 1902 (F. Roessler).
118. BASF, DE 1 064 934, 1959 (L. Zimmermann, H. G. Schwarz).
119. DuPont, US 2 993 754, 1961 (W. R. Jenks, J. S. Linder).
120. Monsanto, US 2 708 151, 1955 (T. D. McMinn).
121. Monsanto, US 2 726 139, 1955 (G. D. Oliver).
122. DuPont, US 2 993 754, 1961 (W. R. Jenks, J. S. Linder).
123. DuPont, DE 1 103 905, 1959 (W. R. Jenks, J. S. Linder).
124. DuPont, US 2 876 066, 1959 (B. N. Inman); US 2 949 341, 1960 (C. P. Green); US 3 015 539, 1962 (J. M. Snyder).
125. DuPont, DE 1 021 340, 1956; US 2 773 752, 1956 (V. W. Kremer, C. H. Lemke).
126. Soc. Elec. Chimie, FR 1 338 413, 1962; FR 1 338 450, 1962.
127. DuPont, DE 1 150 367, 1958 (C. P. Green, W. R. Jenks).
128. DEGUSSA, DE 1 089 738, 1960 (H. Klebe, K. Rinn).
129. BASF, DE 1 144 246, 1960 (E. Hartert, H. G. Schwarz).
130. Shawinigan, US 3 207 574, 1965 (F. B. Popper).
131. DuPont, DE 1 120 437, 1959 (C. P. Green); GB 842 077, 1959.
132. I. G. Farben, DE 710 757, 1941 (K. Gabel).
133. DEGUSSA, DE 111 154, 1898; DE 456 350, 1924 (C. Andrich).
134. DIN 50 971, 30, Dec. 1984.
135. Chemical Safety Data Sheet SD–30 (1967), Manufacturing Chemists Association, Safety and Fire Protection Committee, 1825 Connecticut Avenue, N.W. Washington, D.C. 20 009.
136. Guidelines for Storage, Handling and Distribution of Alkali Cyanides, CEFIC-Cyanides Sector Group Avenue Van Nieuwenhuyse 4/2 B-11610 Brussels/Belgium.
137. A. Bergmann, A. Guzman, G. Potter: *Regeneration of Cyanide*, A Case Study, Conference of Cyanide and the Environment, University of Arizona, Tucson, Dec. 1984.
138. US 1 421 585, 1922 (G. E. Sheriwan, G. G. Griswold).
139. *Praktische Galvanotechnik*, 2nd ed., Leuze Verlag, Saulgau 1970, p. 291.
140. L. Marshall, S. S. Meusel, *Engineering (London)* **181** (1956) 425 – 428.
141. DEGUSSA, US 2 875 095, 1959 (J. Müller); DE 1 108 246, 1959 (J. Müller); DE 1 149 035, 1963 (C. Albrecht, J. Müller).
142. Kolene Corp., US 3 317 357, 1967 (J. Müller).
143. H. Kunst: *Aufkohlen in Salzschmelzen. Wärmebehandlung der Bau- und Werkzeugstähle*, 3rd ed., BAZ Buchverlag, Basel 1978.
144. I. G. Farben, DE 529 506, 1928 (C. Schumann, R. Fick); DE 529 575, 1928 (R. Fick).
145. G. H. Buchanan, *Trans. Am. Electrochem. Soc.* **60** (1931) 93.
146. G. H. Buchanan, *Trans. Electrochem. Soc.* **60** (1931) 93 – 112.
147. H. H. Franck, W. Burg, *Z. Electrochem. Angew. Phys. Chem.* **40** (1934) 686 – 692.
148. Shin-inchi Ohno, *Bull. Chem. Soc. Jpn.* **40** (1969) 1765 – 1769.
149. H. Bauer, *Münchener Beiträge z. Abwasser, Fisch. u. Flußbiol.* **28** (1979) 181 – 192.
150. *The Chemistry of the Ferrocyanides*, vol. VII, American Cyanamid Co., New York 1953.
151. *Ullmann*, 4th ed., **18**, 463 – 467. B. A. Hutchins, L. E. West, *J SMPTE*, **66** (1957) 764 – 768.
152. J. Jostan, H. Mietz, DE 2 534 213, 1975 (J. Jostan, H. Miez).
153. *Ullmann*, 3rd ed., **5**, 659.
154. American Cyanamid, US 3 695 833, 1972 (O. F. Wiedeman, K. W. Saunders, M. N. O'Connor).
155. G. Grube: *Die elektrolytische Darstellung des Ferricyankaliums*, Enke Verlag, Stuttgart 1913.
156. DEGUSSA, DE 2 016 848, 1972 (H. Reinhardt, K. Treibinger, G. Kallrath).
157. Ch. E. MacKinnon, *Symp. Salt, 2nd*, **1** (1965) 356 – 364.
158. R. Noyes: “Citric Acid Production Processes,” *Chem. Process. (Chicago)* **37** (1969) 70 – 120.
159. A. Dinsmoor Webb: “Chemistry of Winemaking,” *Adv. Chem. Ser.* **137** (1974) 278 – 305. Links
160. DuPont, US 2 512 719, 1950 (R. O. Hull); US 2 585 902, 1952 (A. G. Gray).
161. R. D. Crozier, *Mining Magazine*, February 1979, p. 174.
162. DuPont, DE-OS 1 951 260, 1970 (W. T. Hess).
163. H. J. Barber, *J. Chem. Soc.* 1943, 79.
164. DuPont, DE-OS 2 012 444, 1970 (J. D. Rushmere).
165. W. Biltz, *Z. Anorg. Chem.* **170** (1928) 161 – 183.
166. K. Masaki, *Bull. Chem. Soc. Jap.* **6** (1931) 143 – 147.
167. BASF, FR 1 539 598, 1968.
168. L. Wöhler, J. F. Roth, *Chem. Ztg.* **50** (1926) 761 – 763 and 781 – 782.

169. H. Karst, *Z. Gesamte Schiess Sprengstoffwes.* **77** (1922) 116 – 117.
170. G. Paulèt, *Presse Méd.* **66** (1958) 1435.
171. G. Denigés, *Ann. Chim. Phys.* **6** (1895) 381.
172. W. Leithe: *Die Analyse der org. Verunreinigungen in Trink-*, Brauch- und Abwässern, Wissenschaftl. Verlags GmbH, Stuttgart 1972.
173. E. Asmus, H. Garschagen, *Z. Anal. Chem.* **138** (1953) 414 – 422.
174. *Annual Book ASTM Standards*, American Society for Testing and Materials, Philadelphia 1983, D 2036 – 82.
175. *German Standard Methods for the Analysis of Water, Waste Water and Sludge*, DIN 38 405, part 13, DIN Deutsches Institut für Normung e.V., D-1000 Berlin 30.
176. D. J. Barkley, J. C. Ingles, Report 221, Canmet, February 1970.
177. A. Smith, T. Mudder: *Chemistry and Treatment of Cyanidation Wastes*, Mining Journal Books Ltd., London 1991, p. 36.
178. W. S. Spector: *Handbook of Toxicology*, vol. **I**, Saunders, Philadelphia, London.
179. P. Doudoroff: *Toxicity to Fish of Cyanides and Related Compounds, A Review.* Environmental Protection Agency Report No. EPA–600/3–76–038 (1976).
180. P. Doudoroff: *A Critical Review of Recent Literature on the Toxicity of Cyanides to Fish*, American Petroleum Institute, Washington, D.C. 1980.
181. J. L. Huiatt, J. E. Kerrigan, F. A. Olson, G. L. Potter: *Cyanide from Mineral Processing*, Proceedings of a Workshop, Utah Mining and Mineral Resources Research Institute, College of Mines and Mineral Industries, 209 W.C. Browning Building, Salt Lake City, Utah, 1983.
182. USEPA: *Ambient Water Quality Criteria for Cyanide*, United States Environmental Protection Agency. EPA-440/5–80–00 (1980).
183. B. F. Dodge, D. C. Reams: "A Critical Review of the Literature Pertaining to the Disposal of Waste Cyanide Solutions, part 2," *Plating (East Orange, N.J.)* **36** (1949) 571 – 577, 664.
184. A. K. Reed *et al.*: *An Investigation of Techniques for Removal of Cyanide from Electroplating Waste*, U.S. Environ. Protection Agency, Water Pollution Control Series, Program No. 120 10 EIE 11/71 (1971).
185. J. C. Ingles, J. S. Scott: *Overview of Cyanide Treatment Methods*, Cyanide in Gold Mining Seminar, Ottawa, Ontario, 1981.
186. J. Conrad, M. Jola: "Cyanidentgiftung heute," *Oberfläche Surf.* **13** (1972) no. 7, 143 – 148.
187. L. E. Lancy, W. Nohse, D. Wystrach: "Practical and Economical Comparison of the Most Common Metal Finishing Waste Treatment Systems," *Plating (East. Orange, N.J.)* **59** (1972) 126 – 130.
188. G. C. White: *Handbook of Chlorination*, Van Nostrand Reinhold Co., New York 1972.
189. W. Stumm, H. Woker, H. W. Fischer, *Z. Hydrol.* **16** (1954) 1 – 21.
190. H. Knorre, *Galvanotechnik* **66** (1975) 374 – 383.
191. J. P. Zumbrunn, *Inf. Chim.* **124** (1973) 189 – 194.
192. DEGUSSA, DE 2 352 856, 1973 (J. Fischer, H. Knorre, G. Pohl); US 3 970 554, 1976 (J. Fischer, H. Knorre, G. Pohl).
193. DEGUSSA, DE 2 917 714, 1983 (K. Eiermann, R. Goedecke, H. Knorre, R. Möller).
194. DEGUSSA, DE 3 125 452, 1981 (H. Knorre, J. Fischer, K. Stutzel); US 4 416 786, 1983 (H. Knorre, J. Fischer, K. Stutzel).
195. H. Knorre, A. Griffiths, *Conference on Cyanide and the Environment*, University of Arizona, Tucson, Dec. 1984.
196. S. Gos, F. Ladwig, *Proceedings Randol Gold Forum '92*, Vancouver 1992.
197. C. A. Walker, W. Zabban, *Plating (East Orange, N.J.)* **40** (1953) 777 – 780.
198. G. Wysocki, B. Höke, *Wasser Luft Betr.* **18** (1974) 311 – 314.
199. R. S. Murphy, J. B. Nesbitt, *Eng. Res. Bull. P. State Univ. Coll. Eng.* **B 88** (1964); *Chem. Abstr.* **61** (1964) 11 740 d.
200. K. Marquardt, *Metalloberfläche* **23** (1969) no. 8, 231 – 236.
201. W. Götzelmann, *Galvanotechnik* **64** (1973) 588 – 600.
202. E. Deruyst, V. A. Ettel, G. J. Barbely, J. Roy Gordon Research Laboratory, Sheridan Park, Mississauga, Ontario, presented at the 1982 AIME Annual Meeting, Dallas, Texas.
203. Presentation, E. Deruyst, Indaba Conference, 19th Feb. 2003, www.cyplus.com.
204. W. Götzelmann, G. Spanier, *Galvanotechnik* **54** (1963) 265.
205. C. L. Berthollet, *Ann. Chim. Phys.* **1** (1789) 30 – 35.
206. L. J. Gay-Lussac, *Ann. Chim. Phys.* **95** (1815) 200 – 213.
207. F. S. Fawcett, R. D. Lipscomb, *J. Am. Chem. Soc.* **82** (1960) 1509 – 1510. Links
208. H. Schachner, W. Sundermeyer, *J. Fluorine Chem.* **18** (1981) 259 – 268. Links
209. R. Ch. Paul, R. L. Chauhan, R. Parkash, *J. Sci. Ind. Res.* **33** (1974) 31 – 38.
210. F. S. Fawcett, R. S. Lipscomb, *J. Am. Chem. Soc.* **86** (1964) 2576 – 2579. Links
211. Geigy, GB 942 545, 1983.
212. A. A. Woolf, *J. Chem. Soc.* 1954, 252 – 265.
213. E. Mulder, *Recl. Trav. Chim. Pays-Bas* **5** (1880) 84 – 98.
214. D. M. Yost, W. E. Stone, *J. Am. Chem. Soc.* **55** (1933) 1889 – 1895. Links
215. P. L. Bailey, E. Bishop, *Analyst (London)* **97** (1972) 691 – 695.
216. D. Martin, H.-J. Niclas, *Chem. Ber.* **100** (1967) 187 – 195.
217. J. V. Bodrikov, B. V. Danova, *Zh. Org. Khim* **8** (1972) no. 12, 2462 – 2467.

218. J. Iwai, T. Iwashige, Y. Yuma, N. Nakamune *et al.*, *Chem. Pharm. Bull.* **12** (1964) 1446 – 1451. Links
219. Mobil Oil, US 3 433 821, 1969 (L. A. Hamilton, P. S. Landis).
220. E. H. Bartlett, C. Eaborn, D. R. M. Walton, *J. Organomet. Chem.* **46** (1972) , no. 2, 267 – 269. Links
221. Nippon Carbide Industries Co., JP-Kokai 80 133 352, 1980.
222. Bayer, DE 1 958 095, 1971 (E. Enders, K. Wedemeyer, D. Lesch).
223. Bayer, DE 3 538 128, 1985 (R. Fauss, H. J. Riebel).
224. Bayer, DE 2 507 671, 1976 (G. Rottloff, R. Sundermann, E. Grigat, R. Putter).
225. Bayer, DE 2 507 705, 1976 (G. Rottloff, R. Sundermann, E. Grigat, R. Putter).
226. Bayer, DE 1 183 903, 1964 (K. Goliasch, E. Grigat, R. Putter).
227. K. Lempert, G. Doleschall, *Tetrahedron Lett.* 1963, 781. Links
228. J. v. Braun, *Chem. Ber.* **33** (1900) 1438 – 1452.
229. G. Fodor, Sh. Abidi, *Tetrahedron Lett.* 1971, no. 18, 1369 – 1372. Links
230. R. Graf, *Chem. Ber.* **89** (1956) 1071 – 1079.
231. Lonza, CH 680 292, 1990 (T. Cieslewski, R. Chuck, M. Gross, M. P. Galli).
232. DEGUSSA, DE 1 911 174, 1979 (W. Weigert, T. Lussling, F. Theissen).
233. DEGUSSA, DE 1 768 154, 1977 (T. Lussling, F. Theissen, W. Weigert).
234. DEGUSSA, DE-OS 3 006 492, 1981 (A. Kleemann, P. Schalke).
235. Lonza, DE-OS 2 449 013, 1975 (A. Egger, E. Widmer, A. Faucci, R. Gregorin).
236. DEGUSSA, DE-OS 3 006 424, 1981 (A. Kleemann, P. Schalke)
237. E. Mayer, *Angew. Chem.* **81** (1969) 627.
238. Standard Oil, US 3 105 023, 1963 (R. W. Foreman, F. Veatch).
239. H. Schröder, *Z. Anorg. Allg. Chem.* **297** (1958) 296 – 299.
240. Nilok Chemicals Inc., GB 967 458, 1960.
241. DEGUSSA, DE 833 490, 1949 (H. Huemer).
242. DEGUSSA, DE 842 067, 1950 (H. Huemer, H. Schulz).
243. DEGUSSA, DE 827 358, 1949 (1952) (H. Huemer, H. Schulz, W. Pohl).
244. DEGUSSA, DE-OS 2 521 580, 1976 (L. Devlies, R. Hendricx, K. Henkel, N. Kriebitzsch, M. Petzold).
245. Ciba-Geigy Corp., US 4 100 263, 1978 (R. Miller).
246. Bayer, DE-OS 3 117 054, 1982 (H. Königshofen, D. Bruck, A. Nierth, M. Zlokarnik, H. J. Uhlmann).
247. Agripat SS.A, DE-OS 1 801 311, 1969 (W. S. Durrell, R. J. Eckert, Jr.).
248. SKW Trostberg, DE-OS 2 838 016, 1980 (G. Buchreiter, P. Kniep, K. Scheinost, H. R. Vollbrecht).
249. DEGUSSA, DE-OS 2 154 721, 1973 (F. Geiger, W. Weigert).
250. Bayer, DE-OS 2 157 973, 1973 (Z. Kricsfalussy, K. Blöcker, J. Pawlowski, B. Scherhag, R. Weiler).
251. DEGUSSA, DE-AS 2 521 581, 1977 (W. Heimberger, G. Schreyer).
252. DEGUSSA, DE-AS 2 521 582, 1979 (W. Heimberger, G. Schreyer).
253. Röhm, DE-AS 2 442 161, 1975 (W. Gruber, F. Schnierle, G. Schröder).
254. Asahi Chemical Ind., JP 5 5020–224, 1980.
255. SKW-Trostberg, DE-OS 3 103 963, 1982 (U. Kriele, H. Michand, A. Michlbauer).
256. DuPont, US 3 023 077, 1959 (S. V. R. Mastrangelo).
257. DEGUSSA, DE 2 931 353, 1982 (F. Bittner, W. Heimberger, K. Henkel, N. Kriebitzsch, M. Petzold).
258. Agripat, FR 1 555 981, 1969 (W. J. Evers).
259. Ciba-Geigy Corp., US 4 255 167, 1981 (R. M. Babb, M. J. Guillory).
260. DEGUSSA, DE-OS 2 363 867, 1975 (F. Geiger, W. Heimberger, H. Schmitt, G. Schreyer).
261. DEGUSSA, DE-OS 1 809 607, 1970 (F. Geiger, W. Weigert, T. Lussling).
262. DEGUSSA, DE-OS 2 154 721, 1973 (F. Geiger, W. Weigert).
263. T. S. Price, S. J. Green, *Chem. News J. Phys. Sci.* **120** (1920) 101 – 102.
264. American Cyanamid, US 3 226 182, 1965 (J. F. Martino).
265. W. W. Hartmann, E. E. Dreger, *Org. Synth. Coll.* **2** (1943) 150 – 151.
266. E. Schulek, V. Gervay, *Z. Anal. Chem.* **92** (1933) 406 – 417.
267. B. Bak, H. Hillebert, *Org. Synth.* **32** (1952) 29 –31. Links
268. C. A. Goy, D. H. Shaw, H. O. Pritchard, *J. Phys. Chem.* **69** (1965) 1504 – 1507.
269. M. S. Kharasch, A. R. Stiles, E. V. Jensen, A. W. Dewis, *Ind. Eng. Chem.* **41** (1949) 2840 – 2842.
270. A. B. Van Cleave, R. L. Eager, *Can. J. Res. Sect. F* **25** (1947) 284 – 290.
271. A. B. Van Cleave, H. E. Mitton, *Can. J. Res. Sect. B* **25** (1947) 430 – 439.
272. A. B. Van Cleave, V. C. Haskell, J. H. Hudson, A. M. Kristjanson, *Can. J. Res. Sect. B* **27** (1949) 266 – 279.
273. M. J. Grossman, *Chem. Eng. News* **58** (1980) no. 35, 43. Links
274. American Cyanamid, US 1 727 093, 1929 (G. Barsky).
275. W. Mandelung, E. Kern, *Justus Liebigs Ann. Chem.* **427** (1922) 1 – 34.
276. Nilok Chemicals, US 3 429 658, 1969 (G. V. Vosseller).

277. B. Witkop, *Adv. Protein. Chem.* **16** (1961) 221 – 321. Links
278. L. J. Gay-Lussac, *Ann. Chim. (Paris)* **95** (1815) 172 – 199.
279. H. J. Rodewald, *Chemiker-Ztg.* **89** (1965) 522 – 525.
280. R. D. Dresdner, J. Merrit, J. P. Royal, *Inorg. Chem.* **4** (1965) 1228 – 1230. Links
281. J. B. Conway, R. H. Wilson, A. V. Grosse, *J. Am. Chem. Soc.* **75** (1953) 499. Links
282. J. Janin, A. Bouvier, *Spectrochim. Acta* **20** (1964) 1787 – 1798.
283. N. C. Robertson, R. N. Pease, *J. Am. Chem. Soc.* **64** (1942) 1880 – 1886. Links
284. Asahi Chemicals Ind., JP-Kokai 80 20 224, 1980.
285. Asahi Chemicals Ind., JP-Kokai 80 422 242, 1980.
286. N. B. H. Henis, Y.-H. So, L. L. Miller, *J. Am. Chem. Soc.* **103** (1981) , 4632 – 4633. Links
287. N. B. H. Henis, L. L. Miller, *J. Am. Chem. Soc.* **104** (1982) 2526 – 2529. Links
288. Y.-H. So, L. L. Miller, *J. Am. Chem. Soc.* **103** (1981) 4202 – 4209.
289. H. Bock, J. Wittmann, H. J. Arpe, *Chem. Ber.* **115** (1982) 2326 – 2337.
290. DuPont, US 3 564 039, 1971 (O. W. Webster).
291. G. J. Janz, *1,4 [One,Four] Cycloaddition React.* 1967, 97.
292. G. J. Janz, *Inorg. Synthesis* **5** (1957) 43 – 48.
293. Hoechst, DE-OS 2 460 779, 1976 (W. Riemenschneider, P. Wegener).
294. Asahi Chem. Ind., JP-Kokai 74 13 3326, 1970.
295. DEGUSSA, FR 2 003 105, 1969.
296. Asahi Chem. Ind., JP 74 13 110, 1974.
297. Asahi Kasei Kogyo, DE 2 456 035, 1974 (S. Kamada, T. Yamashita, T. Ide, R. Ajiki, T. Kitamura).
298. Asahi Chem. Ind., JP 76 29 429, 1976.
299. W. Riemenschneider, *Chem. Ing. Tech.* **50** (1978) 55.
300. V. D. Parkhomenko, V. K. Steba, A. A. Pivovarov, B. J. Mel'nikov, *Khim. Tekhnol.* **4** (1981) 3 – 4; *Chem. Abstr.* **95** (1981) 149 178 h.
301. Hoechst, DE 2 427 269, 1982 (W. Riemenschneider, P. Wegener).
302. L. M. Weinstock, P. Davis, B. Handelsman, R. Tull, *J. Org. Chem.* **32** (1967) 2823 – 2829. Links
303. H. W. Roesky, K. Keller, J. W. Bats, *Angew. Chem.* **95** (1983) 904.
304. G. Jacquemin, *Ann. Chim. Phys.* **6** (1885) 140.
305. E. Pfeil, DE 1 163 302, 1965 (E. Pfeil, K. H. Schäfer, R. Nebeling).
306. DEGUSSA, DE-AS 2 461 204, 1977 (W. Heimberger, G. Schreyer, H. Schmitt).
307. Sagami Chemical Research Center, Tokio, DE-AS 1 297 589, 1969 (S. Nakamura).
308. Asahi Kasei Kogyo K. K., DE-AS 2 355 040, 1977 (T. Yamashita, T. Ide, S. Kamada, T. Kitamura).
309. Hoechst, DE-AS 2 427 268, 1979 (W. Riemenschneider, P. Wegener).
310. Union Oil Comp. of California, US 3 630 671, 1971 (M. J. Block).
311. Union Oil Comp. of California, US 3 769 388, 1973 (K. L. Olivier, M. J. Block).
312. DEGUSSA, DE-OS 2 012 509, 1971 (W. Heimberger, T. Lussling, W. Weigert).
313. DEGUSSA, DE-OS 2 022 454, 1971 (J. Heilos, W. Heimberger, T. Lussling, W. Weigert).
314. DEGUSSA, DE-OS 2 022 455, 1971 (J. Heilos, W. Heimberger, T. Lussling, W. Weigert).
315. DEGUSSA, DE-AS 2 118 819, 1974 (J. Heilos, W. Heimberger, T. Lussling, W. Weigert).
316. Pure Oil Comp., US 2 884 308, 1956 (W. L. Fierce, W. J. Sandner).
317. Röhm, DE-OS 2 341 370, 1973 (W. Gruber, G. Schröder).
318. Mitsubishi, DE-OS 2 532 307, 1976 (T. Onoda, T. Yokohama, K. Shinpei).
319. DEGUSSA, DE-OS 1 927 847, 1970 (F. Geiger, T. Lussling, W. Weigert).
320. American Cyanamid, US 2 730 430, 1956 (J. K. Dixon, J. E. Longfield).
321. DEGUSSA, DE 1 283 815, 1967 (R. Fahnenstich, H. Schulz).
322. Bayer, DE-OS 3 229 415, 1984 (R. Fauß, K. H. Linker, K. Findeisen).
323. W. Tsang, S. H. Bauer, M. Cowperthwaite, *J. Chem. Phys.* **36** (1962) 1768 – 1775. Links
324. Asahi Chemical Industries, JP-Kokai 75 93 893, 1975.
325. J. B. Conway, A. V. Grosse, *J. Am. Chem. Soc.* **80** (1958) 2972 – 2976. Links
326. R. P. Welcher, D. J. Berets, L. E. Sentz, *Ind. Eng. Chem.* **49** (1957) 1755 – 1758.
327. R. Dailey and A. Weissler: *Monograph on Ferrocyanide Salts*, Food and Drug Administration Report No. FDA/BF–79/28 Washington, D.C., 1978.
328. A. Harsten, Blackhill, *Engineer (London)* **22** (1934) 145 – 174.
329. B. G. Anderson, *Sewage Works J.* **16** (1944) 1156.
330. G. Bringmann, R. Kühn, *Z. Wasser Abwasser Forsch.* **10** (1977) 87 – 98.
331. C. Henderson, Q. H. Pickering, A. E. Lencke, *Eng. Bull. Purdue Univ. Ext. Ser.* 1960, no. 106, 120 – 130.
332. W. P. Bridges: *Sodium Cyanide as a Fish Poison*, Special Scientific Report Fisheries No. 253, US Dept. Interior (1958).
333. J. R. Hawley: *The Use, Characteristics and Toxicity of Mine-Mill Reagents in the Province of Ontario*, Ontario Ministry of Environment, 135 St. Clair Avenue West, Toronto 1972.

334. W. Bucksteeg, *Schweiz. Z. Hydrol.* **22** (1960) 407.
335. G. Bringmann, R. Kuhn: “The Toxic Effect of Waste Water on Aquatic Bacteria, Algae, and Small Crustaceans,” *Gesund. Ing.* **80** (1959) 115 – 120.
336. B. G. Anderson: “The Toxicity Thresholds of Various Sodium Salt Determined by the Use of Daphnia magna,” *Sewage Works J.* **18** (1946) 82.
337. *Patty's Industrial Hygiene and Toxicology*, 3rd ed., J. Wiley & Sons, New York 1978, p. 4845.
338. TLVs, *Threshold Limit Values for Chemical Substances in the Work Environment*, (Adopted by ACGiH for 1984 – 85) American Conference of Governmental Industrial Hygienists.
339. MAK. Mitteilung XX der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, 1984.
340. F. Brivet, J. F. Delfraissy, M. Duche, P. Bertrand, J. Dormont, “Acute Cyanide Poisoning; Recovery with Nonspecific Supportive Therapy”, *Intensive Care med.* **9** (1983) 33. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Cyanuric Acid and Cyanuric Chloride

KLAUS HUTHMACHER, Degussa AG, Hanau-Wolfgang, Federal Republic of Germany

DIETER MOST, Degussa AG, Hanau-Wolfgang, Federal Republic of Germany

1. Introduction 1
2. Cyanuric Acid 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.3. Production 3
2.4. Environmental Protection 4
2.5. Quality Specifications and Chemical Analysis 4
2.6. Storage and Transportation 5
2.7. Uses 5
2.8. Economic Aspects 5
2.9. Toxicology and Occupational Health . 11
3. Cyanuric Chloride 11
3.1. Physical Properties 11
3.2. Chemical Properties 11
3.3. Production 14
3.4. Environmental Protection 14
3.5. Quality Specifications and Analysis . 15
3.6. Storage and Transportation 15
3.7. Uses 15
3.8. Economic Aspects 16
3.9. Toxicology and Occupational Health . 16
4. References 19

1. Introduction

Many acyclic compounds containing C – N double or triple bonds tend to condense to form six-membered rings with alternating C and N atoms:

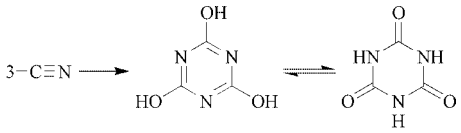

These derivatives are called *s*-triazines or 1,3,5-triazines. Although the ring structure is not quite a regular hexagon, spectroscopic data indicate an aromatic character of the ring system [7]. The aromaticity of the triazine ring is discussed in [6, p. 466]. A large number of their reactions are described in the literature, mostly substitution reactions with ring retention.

Compound **1** is commonly called cyanuric acid, even though the oxo (keto) structure **1b** is dominant over the hydroxy (enol) structure **1a**. Reactions with electrophiles generally give derivatives with structure **1b** (isocyanurates, e.g., **1** – **28**). Triazine compounds of structure **1a** (cyanurates, e.g., **30** – **55**) are made from cyanuric chloride (**29**). Compounds of industrial importance of both types are shown in Tables 1 and 2.

Compounds that are mixtures of tautomers are known, although the keto structure is generally more stable than the hydroxy structure. Some alkyl cyanurates rearrange to give the corresponding isocyanurates [1], [8], [9]. The rearrangement of triallyl cyanurate (TAC) (**39**) to triallyl isocyanuarate (TAIC) (**11**), which can be carried out thermally or catalytically is of some industrial interest.

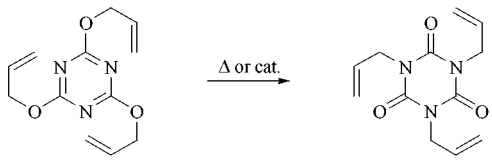

Cyanuric acid and cyanuric chloride are readily interconvertible [10], [11]:

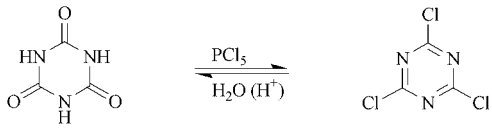

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_191

Although cyanuric acid and cyanuric chloride have been known since 1776 and 1827, respectively, they only gained industrial importance in the 1950s.

2. Cyanuric Acid

2.1. Physical Properties

Cyanuric acid [*108-80-5*], isocyanuric acid, 1,3,5-triazine-2,4,6(1*H*,3*H*,5*H*)-trione (**1**), $C_3H_3N_3O_3$, M_r 129.8, *mp* 320–330 °C (decomp.), is an odorless white solid, $d_0^4 = 1.8$. Its solubility in water is 0.2 wt % at 10 °C and 2.6 wt % at 90 °C. At room temperature cyanuric acid is only slightly soluble in common organic solvents such as acetone, benzene, ether, ethanol, hexane, and isopropyl alcohol (0.1 wt %). Cyanuric acid has significant solubility in solvents with large dipole moments. The solubility (wt %) of cyanuric acid at 25 °C is as follows:

Dimethyl sulfoxide	17.4
Sulfuric acid (98 wt %)	14.1
Dimethylformamide	7.2
N-methyl-2-pyrrolidone	6.3
Dimethylacetamide	3.0
Pyridine	2.2

Cyanuric acid crystallizes from water as a colorless crystalline dihydrate, which loses water in dry air. Cyanuric acid is a weak tribasic acid (K_a values ca. 10^{-7}, 10^{-11}, and 10^{-14} [3, D1]). The solubilities of the mono-, di-, and trisodium salts in water are 0.9, 5.7, and 14.1 g per 100 mL of solution, respectively [12]. For other physical data, see [1] and [3, D1].

2.2. Chemical Properties

Only a few reactions of cyanuric acid occur by replacement of hydrogen at oxygen: in alkaline solution the hydroxy structure is favored over the keto structure. Salts are formed with inorganic and organic bases [1].

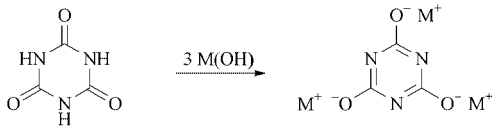

M = alkali or alkaline earth metal

The reaction of cyanuric acid with ammonia under pressure in the presence of a catalyst at 350–400 °C produces melamine [*108-78-1*] [13], [14], [15].

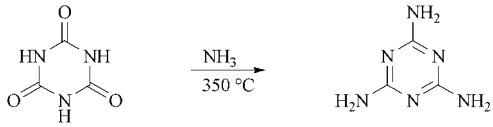

In the presence of alkaline catalysts, ketene reacts with cyanuric acid to afford triacetyl cyanurate [*13483-16-4*] in high yield [16].

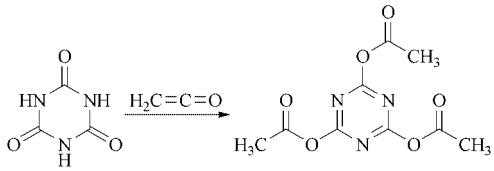

The typical reactions of cyanuric acid are those of a cyclic imide, in which the nucleophilic isocyanurate nitrogen atom, mostly in the form of an anion, attacks a positively polarized carbon atom of the reaction partner. This type of reaction usually results in trisubstitution. Epoxides undergo addition to form hydroxyalkyl isocyanurates when heated in DMF. Ethylene oxide reacts nearly quantitatively at 100 °C to form tris(2-hydroxyethyl) isocyanurate (**8**). Similarly, propylene oxide gives the corresponding hydroxypropyl analogue [17], [18]. Epichlorohydrin forms tris(3-chloro-2-hydroxypropyl) isocyanurate (**9**) in dioxane at 110 °C in the presence of basic catalysts [19]; its dehydrochlorination in aqueous alkali produces the commercial product tris(2,3-epoxypropyl) isocyanurate (**10**) [20]. A typical addition reaction of cyanuric acid to a double bond is the formation of bis(2-cyanoethyl) isocyanurate [*2904-27-0*] and tris(2-cyanoethyl) isocyanurate (**27**) by reaction with acrylonitrile in DMF containing Triton B (a surfactant) at 130 °C [18]. Saponification of tris(2-cyanoethyl) isocyanurate gives tris(2-carboxyethyl) isocyanurate (**22**). Esterification of **22** with ethanol gives tris(2-carbethoxyethyl) isocyanurate (**23**), which is a starting material for the commercial light stabilizer and antioxidant **24**. Cyanuric acid dissolves readily in aqueous formaldehyde [12], [21] with formation of tris(hydroxymethyl)

isocyanurate (**18**). Alkyl halides react with cyanuric acid at elevated temperature in the presence of a proton acceptor with elimination of hydrogen halide. Thus, allyl chloride forms the commercial product triallyl isocyanurate (**11**) in dichlorobenzene/triethylamine at 130 °C [22], [23]; 2-chloroethanol yields tris(2-hydroxyethyl) isocyanurate (**8**) in aqueous sodium hydroxide [24]. The trisubstituted isocyanurates (**11**) – (**17**), (**21**), and (**26**), as well as the commercial products (**19**) and (**20**) can be prepared similarly. Chlorine reacts with cyanuric acid in water at controlled pH to afford the *N,N′*-dichloro and *N,N′,N″*-trichloro compounds (**3**) – (**7**) (→ Chloroamines). Which compound is formed depends upon the ratio of sodium hydroxide to cyanuric acid. If a 2:1 molar ratio is used, mainly *N,N′*-dichloroisocyanuric acid (DCCA) is obtained. If a 3:1 molar ratio is used, mainly *N,N′,N″*-trichloroisocyanuric acid (TCCA) is obtained in 90 % yield.

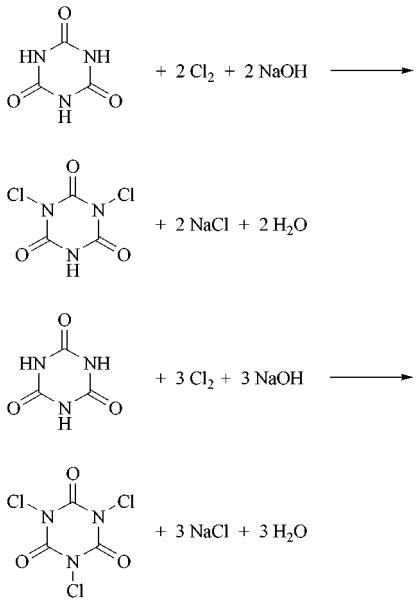

Cyanuric acid decomposes slowly above 200 °C and rapidly above its melting point (320 – 330 °C). The main product is isocyanic acid [*75-13-8*] [25], [26]. At higher temperature, some reactions are likely to proceed via a cyanic acid intermediate, that is, they involve ring cleavage.

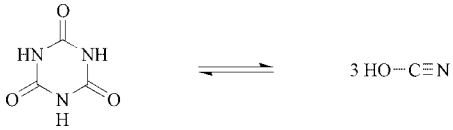

Another example of ring cleavage is the conversion of long-chain fatty acids into nitriles above 250 °C [27].

2.3. Production

Laboratory quantities of pure cyanuric acid are produced by hydrolysis of cyanuric chloride [11] or melamine [28]. The acid is further purified by recrystallization from dimethylformamide or via the sodium or ammonium salts.

On a commercial scale cyanuric acid is manufactured by pyrolysis of urea between 200 and 300 °C to give ammonia and cyanuric acid.

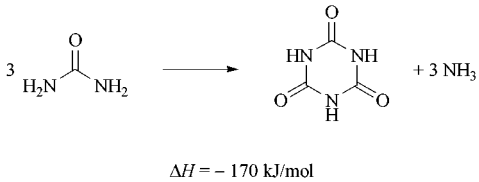

In some processes, urea is heated in a kiln for several hours. The process starts with molten raw material and ends with solid cyanuric acid. Numerous patents deal with methods to improve heat and mass transfer with specially designed reactors or with a modified reaction mixture. The reaction can be carried out in heated screws, revolving drums [29], steel conveyor belts [30], rotary-tube furnaces [31], or ball mills [32]. Pyrolysis in a fluidized bed of AlF_3 particles [33] or cyanuric acid crystals [34] and heating of molten urea by high-frequency induction have also been reported [35]. The caking problem is solved by premixed powders with definite urea to cyanuric acid ratios [31], or preformed urea cyanurate [34], [36], or by recycling the crude product [31]. Other methods use molten metals [37], [38] or salts [39] as the heat-transfer medium. Urea is fed into the bottom of a crucible filled with agitated molten tin or tin alloy, and the cyanuric acid collects at the metal surface. Other methods employ high-boiling solvents such as polyglycol ethers [40] and sulfolanes [41], which dissolve urea but not cyanuric acid. These methods avoid

localized overheating, which reduces yields by formation of isocyanic acid. Other solvents have also been reported [3, D1, p. 369]. Byproduct formation is suppressed by rapidly removing the ammonia at reduced pressure, sparging with inert gas, or codistillation.

Crude cyanuric acid contains up to 30 % impurities, consisting mainly of melamine and its precursors:

Biuret [*108-19-0*]

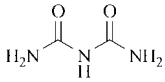

Triuret [*556-99-0*]

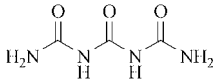

Ammelide [*645-93-2*]

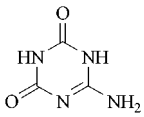

Ammeline [*645-93-1*]

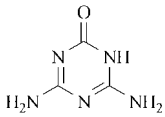

Melamine [*108-78-1*]

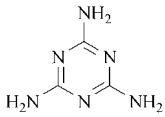

Dilute mineral acids convert the impurities to ammonia and carbon dioxide or cyanuric acid [31], [42]. Another method is to dissolve the crude acid in NaOH or dilute aqueous ammonia, to filter the organic impurities off, and to precipitate the acid by adding mineral acids [43], [44].

2.4. Environmental Protection

Cyanuric acid is biodegraded readily under anaerobic conditions [45], in the soil [46], and in aerated systems [47], [48]. The principal environmental impact comes from salt in the effluents resulting from the manufacture of cyanuric acid and its derivatives.

2.5. Quality Specifications and Chemical Analysis

Cyanuric acid is available as solid in powdered form or mainly in coarse granular form, ≥ 85 % 10 – 100 mesh. Manufacturers' specifications are in the following range:

Cyanuric acid	≥ 98.5
Ammelide + ammeline	≤ 1 %
Color (APHA)	≤ 50
Moisture	≤ 0.6 %
Ash	50 – 1000 ppm
pH	4 – 5 (0.25 % solution or 10 % slurry)

Thin-layer chromatography on cellulose is suitable for quality control. Cyanuric acid is determined gravimetrically as its melamine salt [49] or by potentiometric titration of a solution in DMSO/benzene with tetrabutylammonium hydroxide [50], [51]. In wastewater cyanuric acid can be determined by amperometric titration with $Hg(NO_3)_2$ [52] or by differential pulse polarography [53]. Solid cyanuric acid can be readily identified by IR spectroscopy [54] or X-ray diffraction. Cyanuric acid and the isocyanurates can also be determined by nitrogen analysis according Kjeldahl or Dumas. Impurities in cyanuric acid are determined as follows:

- Liberation, distillation, and determination of ammonia (Kjeldahl, Dumas, or ion-selective electrodes [55])
- Determination of urea by urease acidimetry
- Determination of biuret by UV spectrophotometry

2.6. Storage and Transportation

Cyanuric acid should be stored in a dry place. Silos and pneumatic conveyors should be grounded to prevent electrostatic discharges. The loose bulk density is 800 – 900 g/L. The product is packed in polyethylene-lined fiber drums or paper bags. The shipping classification in the United States is "Acid, NOIBN, Dry" (NOIBN, not dangerous according to DOT CFR Title 49).

2.7. Uses

Cyanuric acid is widely used for the manufacture of *N*-chlorinated isocyanurates (**3**) – (**7**), which are employed most importantly in swimming pool disinfectants, household bleaches, industrial and institutional cleaners, dishwasher detergents, and general sanitizers (→ Chloroamines). Cyanuric acid serves as stabilizer of available chlorine in swimming pools. In concentrations of 25 – 50 ppm, it reduces the deactivation rate of chlorine by a factor of 5 – 10 [56]. Cyanuric acid is approved as a source of nitrogen in ruminant feed [57], [58]. The use of cyanuric acid as a source of isocyanic acid by heating to 450 – 650 °C has been claimed [59]. Small quantities of cyanuric acid are used for reducing nitrogen oxides (NO_x) in stationary diesel engine exhaust gas or in coal-, oil-, or gas-fired boilers. The isocyanurates **8** – **12**, **22**, and **23** are used in the plastics industry [12], [60]. Triallyl isocyanurate (TAIC; **11**) can be homopolymerized and copolymerized. It serves as cross-linking agent in polyethylene and poly(vinyl chloride) and in laminate formulations [61], [62]. Tris(2-hydroxyethyl) isocyanurate (THEIC; **8**) is a cross-linking agent for polyurethanes, polyesters, and alkyd resins used for wire enamels and electrical varnishes [63]. Triglycidyl isocyanurate (TGIC; **10**) is a cross-linking agent in epoxy resins [19], [64] and a curing agent for weather-resistant powder coatings. Melamin cyanurate (**28**) has achieved commercial importance as halogen-free flame retardant in polyamide resins and compositions [65]. It is also useful as a solid lubricant [66]. Tris(2-carboxyethyl) isocyanurate (TCEIC; **22**) is used in making water-soluble alkyd resins; the esters serve as plasticizers for poly(vinyl chloride) and lubricants [67], [68]. An overview of important products based on cyanuric acid is given in Table 1.

2.8. Economic Aspects

World capacity for cyanuric acid is estimated to be 80 000 t/a (1997), more than 90 % of which is used to make *N*-chlorinated isocyanurates. Monsanto, FMC and Olin (all USA) are the world's largest bulk suppliers of cyanuric acid and chlorinated isocyanurates. In Dezember 1992 Occidental Chemical entered the US market by purchasing Monsanto's ACL business (chlorinated isocyanurates). The plants will continue to be operated by Monsanto. Important producers of cyanuric acid are listed in the following:

United States	FMC, Monsanto, Occidental, Olin
Japan	Shikoku Kasei Kogyo, Nissan Chemical Industries, Nippon Soda
France	Atochem
Spain	Delsa
Italy	Sigma
Germany	BASF
Taiwan	Ferrel Int., Kuang Ming Enterprise, Tai-Yu Chem. Ind.

There are also smaller producers in China, Korea and the former Soviet Union. All major producers also make the chlorinated derivatives.

2.9. Toxicology and Occupational Health

Cyanuric acid has been classified as essentially nontoxic (acute oral and dermal), nonirritating to the skin, and essentially nonirritating to the eye [12], [69].

Acute toxicity	
LD_{50} (rat, oral)	> 5000 mg/kg [70]
LD_{50} (rabbit, dermal)	> 5000 mg/kg
LC_{50} (fish)	> 2000 mg/L (24h) [71]
EC_{50} (Daphnia)	> 2000 mg/L [71]
Subacute toxicity	
NOEL (fish)	1000 mg/L [71]
NOEL (algae)	2500 mg/L [71]
Inhib. of reprod. (Daphnia)	1000 mg/L [71]

Inhalation of the dust, contact with the eye, and ingestion should be avoided. Above 200 °C, cyanuric acid forms toxic isocyanic acid.

3. Cyanuric Chloride

3.1. Physical Properties

Cyanuric chloride, 2,4,6-trichloro-1,3,5-triazine [*108-77-0*] (**29**), $C_3N_3Cl_3$, M_r 184.5, triple point 145.7 °C at 255 kPa; *bp* 194 °C, d_4^{20} 1.92 (solid), d_4^{150} 1.48 (melt), forms colorless monoclinic crystals of pungent odor, reminiscent of acetamide and acid chlorides. The compound is soluble in acetonitrile, ether, ketones, and chlorinated hydrocarbons, but insoluble in water. The

Table 1. Selected triazine derivatves based on cyanuric acid

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Cyanuric acid (CA)	starting material for chloroisocyanurates and for a variety of organic compounds	[*108-80-5*]	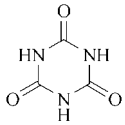	**1**
Trisodium cyanurate	starting material for chloroisocyanurates and for a variety of organic compounds	[*3047-33-4*]	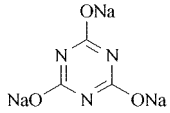	**2**
Sodium dichloroisocyanurate	swimming pool sanitizer, disinfectant	[*2893-78-9*]	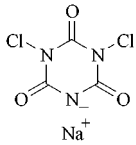	**3**
Sodium dichloroisocyanurate dihydrate	swimming pool disinfectant, dishwasher detergents, scouring cleaners, and sanitizers	[*51580-86-0*]	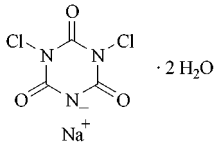	**4**
Potassium dichloroisocyanurate	dishwasher detergents, scouring cleaners and sanitizers, swimming pool disinfectant	[*2244-21-5*]	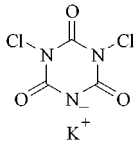	**5**
Trichloroisocyanurate (TCCA)	swimming pool sanitizer, disinfectant	[*87-90-1*]	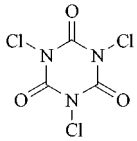	**6**
Trichloroisocyanurate tetrakis (potassium dichloroisocyanurate)	scouring powders, household bleaches, institutional and industrial cleaners, dishwasher detergents	[*64474-05-1*]	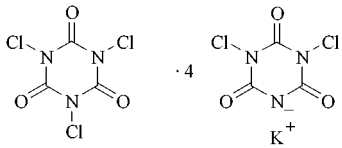	**7**

Table 1. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Tris (2-hydroxyethyl) isocyanurate (THEIC)	modifier for polyesters, polyester imides, and alkyd resins	[*839-90-7*]	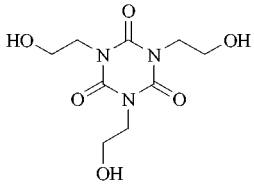	**8**
Tris (3-chloro-2-hydoxypropyl) isocyanurate	photohardening resin compounds, intermediate	[*7423-53-2*]	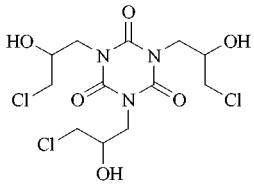	**9**
Tris (2,3-epoxypropyl) isocyanurate (triglycidyl isocyanurate, TGIC)	epoxy resins	[*2451-62-9*]	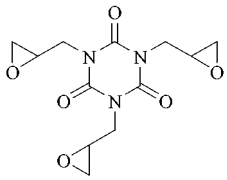	**10**
Triallyl isocyanurate (TAIC)	cross-linking agent for vinyl polymers	[*1025-15-6*]	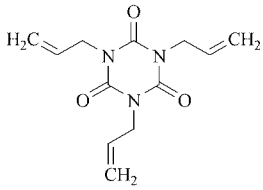	**11**
Trimethallyl isocyanurate	cross-linking agent for vinyl polymers vulcanizing agent	[*6291-95-8*]	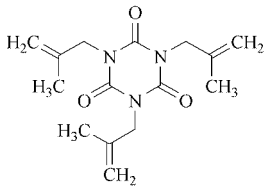	**12**

Table 1. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Tris (2-ketopropyl) isocyanurate	intermediate	[*61050-97-3*]	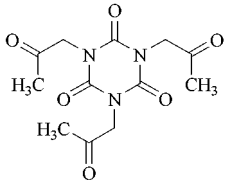	**13**
Tris (carboxymethyl) isocyanurate	water-soluble rust inhibitor	[*1968-52-1*]	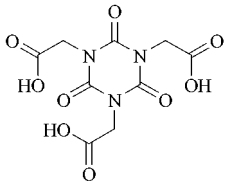	**14**
Tribenzyl isocyanurate	insecticidal activity, processing aids for extrusion of synthetic rubber	[*606-03-1*]	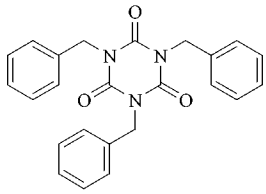	**15**
Tri-*n*-hexyl isocyanurate	light stabilizer	[*36761-61-2*]	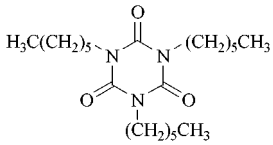	**16**
Trimethyl isocyanurate	antifouling agent, absorbent for sulfur dioxide removal from waste gases, fireproofing agents	[*827-16-7*]	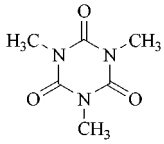	**17**
Tris (hydroxymethyl) isocyanurate	intermediate for light stabilizers	[*10471-40-6*]	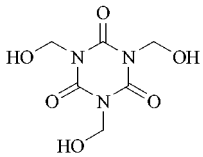	**18**

Table 1. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
1,3,5-Tris (4-hydoxy-3,5-di-tert-butylbenzyl) isocyanurate	light stabilizer, antioxidant	[*27676-62-6*]	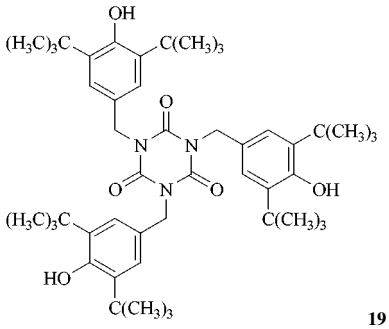	**19**
1,3,5-Tris (2,6-dimethyl-3-hydroxy-4-*tert*-butylbenzyl) isocyanurate	light stabilizer, antioxidant	[*40601-76-1*]	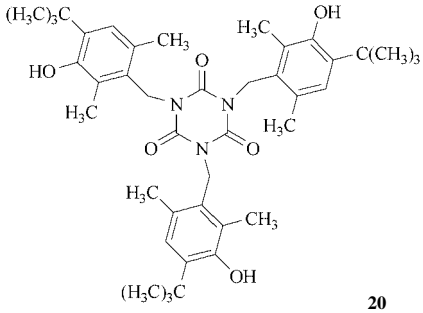	**20**
Tris (ethoxycarbonylmethyl) isocyanurate	intermediate for light stabilizers	[*69455-18-1*]	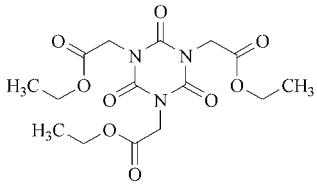	**21**
Tris (2-carboxyethyl) isocyanurate (TCEIC)	water-soluble alkyd resins, plasticizer for PVC	[*2904-41-8*]	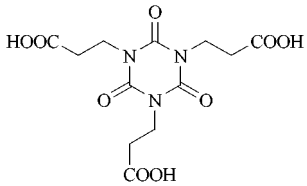	**22**
Tris (2-carbethoxyethyl) isocyanurate	intermediate for light stabilizers	[*2904-39-4*]	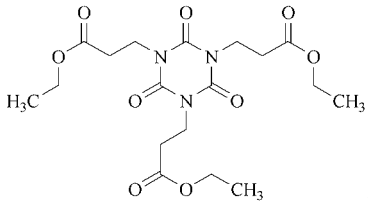	**23**

Table 1. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Tris [2-[3-(3,5-di-*tert*-butyl-4-hydroxyphenyl)propionyl-oxy]ethyl] isocyanurate	light stabilizer, antioxidant	[*34137-09-2*]	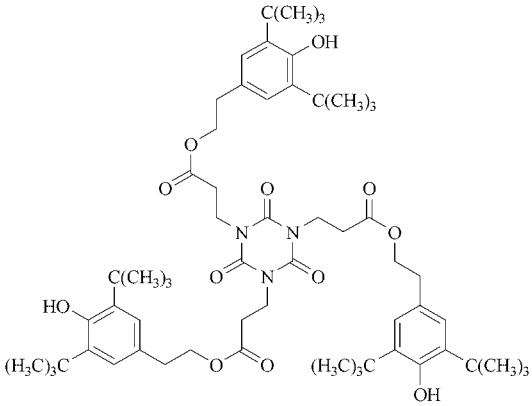	**24**
Tris (triphenylstannyl) isocyanurate	intermediate	[*752-74-9*]	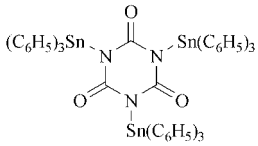	**25**
Tris (carbamoylmethyl) isocyanurate	intermediate, flame retardant	[*1843-48-7*]	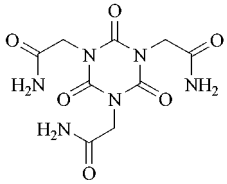	**26**
Tris (2-cyanoethyl) isocyanurate	flame retardant	[*839-90-7*]	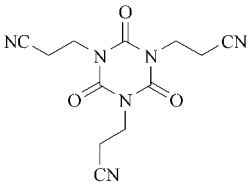	**27**
Melamine cyanurate	halogen-free flame retardant	[*37640-57-6*]	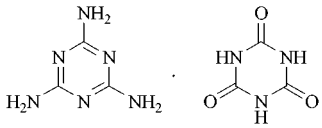	**28**

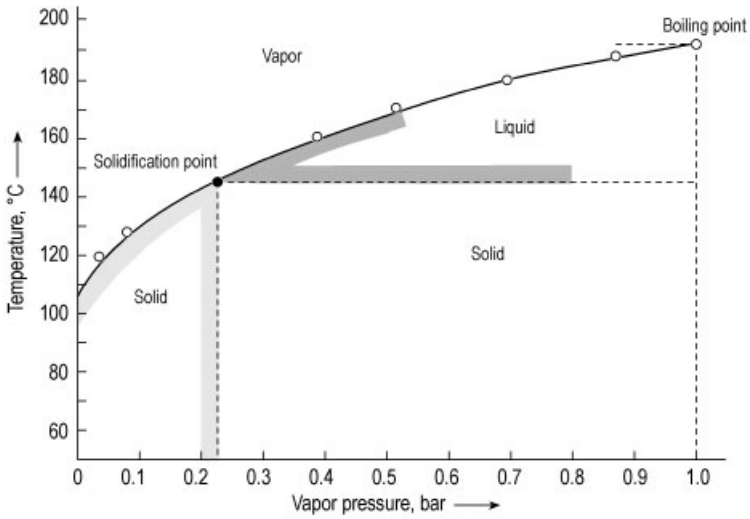

Figure 1. Vapor pressure of liquid cyanuric chloride

solubility (in wt %) of cyanuric chloride in various solvents at 25 °C is as follows:

Acetone	25.0
Acetonitrile	21.0
Acrylonitrile	19.0
Benzene	19.0
Chlorobenzene	16.0
Tetrachloromethane	7.5
Diethyl ether	14.0
Dioxane	55.0
Ethyl acetate	21.0
Methyl vinyl ketone	27.0
Nitrobenzene	18.0
Tetrahydrofuran	43.0
Trichloromethane	20.0

The solubility in toluene at various temperatures is as follows:

at 20 °C	17.2 wt %
at 30 °C	19.8 wt %
at 40 °C	25.0 wt %
at 50 °C	29.7 wt %

Heat capacity c_p is $0.99\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ at 150 °C, heat of fusion is 123 kJ/kg, and heat of vaporization is 256 kJ/kg. The vapor pressure of cyanuric chloride as a function of temperature is shown in Figure 1.

The vapor pressure (kPa) of the liquid as a function of temperature (°C) is given by

$$\ln p/\mathrm{mmHg} = -0.33480 - 742.7311/(T/\mathrm{K}) + 11.15147 \ln (T/\mathrm{K})$$

Cyanuric chloride sublimes before reaching the boiling point; the vapor pressure (in mbar) of the solid is:

At 20 °C	0.025
At 62 °C	1.040
At 70 °C	2.700
At 114 °C	34.900

3.2. Chemical Properties

Under anhydrous conditions, cyanuric chloride can act as a chlorinating agent. Alcohols and tertiary amines are converted to alkyl chlorides [72], [73]. Carboxylic acids form acid chlorides in anhydrous acetone in the presence of triethylamine. The acid chloride can be isolated or treated in situ to give the ester, amide, etc. [74]. In both cases cyanuric acid is formed as byproduct.

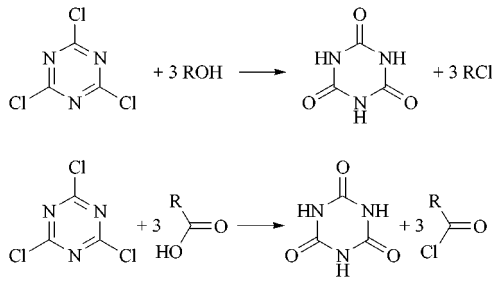

Cyanuric chloride can be used as a condensation (water-withdrawing) reagent in chemical synthesis. Amides and thioamides can be converted to nitriles.

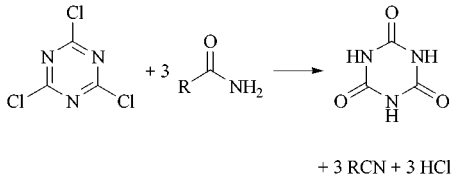

Table 2. Cyanuric acid as condensation/halogenation/oxidation agent in organic synthesis [74–85]

Starting material	Product	Refs.
$RCONH_2$ Carboxylic acid amides	RCN Carboxylic acid nitriles	[75], [76]
Aldoximes	R-CN Carboxylic acid nitriles	[75]
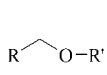 Ethers	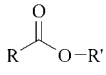 Carboxylic esters	[77]
Carboxylic acids	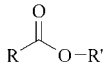 Carboxylic acid esters	[78]
Hydroxycarboxylic acids	Cyclic esters (lactones)	[79]
Maleic acid derivatives	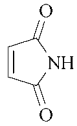 Maleimide	[80]
RCOOH Carboxylic acids	$RCONH_2$ Carboxylic acid amides (peptides)	[74]
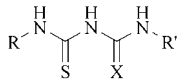 X = O, N, S	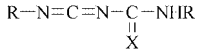 Carbodiimides	[81], [82]
Substituted thioureas		
ROH Alcohols	RY Y = Cl, Br Alkyl halides	[83], [84] [85]
RCOOH Carboxylic acids	RCOCl Carboxylic acid chlorides	[74]

Aldoximes react in an analogous way. A selection of important reactions in which cyanuric chloride acts as chlorination/condensation reagent are listed in Table 2.

Under anhydrous conditions tertiary aliphatic amines are converted to substituted melamine derivatives. At the same time one alkyl group is removed and converted to the corresponding alkyl chloride [73].

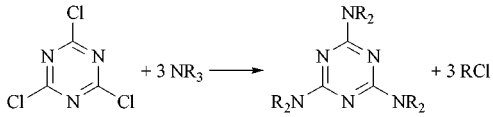

Cyanuric chloride usually reacts like an acid chloride with formation of hydrogen chloride.

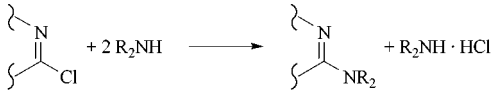

According to Fierz – David [86] a suspension of cyanuric chloride in ice water (0 °C) is fairly stable for nearly 12 h. Above 10 °C hydrolysis occurs rapidly; for example, at 30 °C more than 40 % of the cyanuric chloride is hydrolyzed within 1 h. The final hydrolysis product is cyanuric acid. Figure 2 shows the dependence of the hydrolysis of cyanuric chloride ($c = 50$ g/L) in water as a function of hydrolysis time at different temperatures.

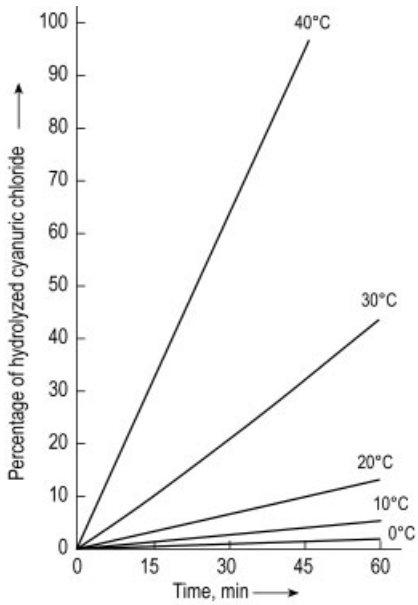

Figure 2. Hydrolysis of cyanuric chloride as a function of temperature

In an alkaline aqueous medium, e.g., in the presence of NaOH or $NaHCO_3$, at 40 °C, 2,4-dichloro-6-hydroxy-1,3,5-triazine is formed. With an excess of base 2,4-dihydroxy-6-chloro-1,3,5-triazine is formed. Trisodium cyanurate is formed only above 125 °C. A number of processes take advantage of the temporary stability of cyanuric chloride in water by reacting a suspension of finely divided cyanuric chloride in ice water with a strong nucleophilic agent such as an alcohol, a thiol, or a primary or secondary amine [87]. These reactions often proceed stepwise to trisubstitution at a well-defined temperature. An empirical rule for amine substituents states that the first chlorine substituent is replaced at 0 – 5 °C, the second at 30 – 50 °C, and the third at 70 – 100 °C [86]. Sodium hydroxide, sodium hydrogencarbonate, disodium hydrogenphosphate, and tertiary amines can be used as HCl scavengers (see also [6, p. 483]).

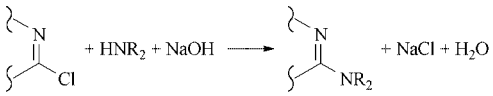

The triazine product is usually isolated by filtration or centrifugation. Solvents such as acetone, methyl ethyl ketone, or toluene are often used for these reactions. Cyanuric chloride, melt or powder, is dissolved or suspended in the solvent with the nucleophile and treated with aqueous NaOH. The product is filtered or is isolated from the organic layer by evaporation. Both methods are used to introduce different substituents successively at different temperatures. Replacement of the third chlorine atom is sometimes carried out by employing high-boiling solvents or simply the molten reactant. Care must be taken to neutralize the liberated hydrochloric acid and to dissipate the heat of reaction. Otherwise, the exothermal hydrolysis of cyanuric chloride ($\Delta H = -2164$ kJ/kg) can become uncontrollable, especially in the manufacture of cyanurates from alcohols [88]. Hydrolysis is accelerated by water-miscible solvents. Storage of solutions of cyanuric chloride in aqueous acetone without heat removal can result in a runaway reaction, which can even start at room temperature. Cyanuric chloride reacts with DMF to form Gold's salt [*20353-93-9*] [89].

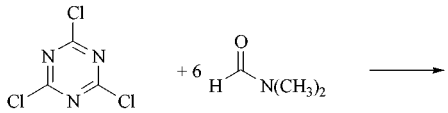

$$3\left[(CH_3)_2N{=}CH{-}N{=}CH{-}N(CH_3)_2\right]^+ Cl^- + 3\,CO_2$$

The reactions of cyanuric chloride are surveyed in [90], [91].

3.3. Production

Most cyanuric chloride facilities are found at plants producing hydrocyanic acid, the key starting material. Sodium cyanide is a minor source.

The *hydrocyanic acid* is chlorinated to cyanogen chloride, which is trimerized directly to cyanuric chloride:

$$HCN + Cl_2 \longrightarrow ClCN + HCl \quad \Delta H = -89\ \text{kJ/mol}$$

$$3\,ClCN\ (g) \longrightarrow (ClCN)_3\ (g) \quad \Delta H = -233\ \text{kJ/mol}$$

In a typical process, hydrocyanic acid and chlorine are fed to a loop reactor at 20 – 40 °C (see Fig. 3). Cyanogen chloride saturates the aqueous reaction medium and leaves the loop (a and b) [92]. It is washed with water (c) and dried (d) and is then ready for trimerization.

The wash water from c dissolves the hydrogen chloride and delivers it to a stripper (e), which recycles dissolved cyanogen chloride and releases hydrochloric acid. Sections e, a, and c can also be combined in a single apparatus [93].

The dry cyanogen chloride is trimerized at ≥ 300 °C on activated carbon (f) [94], [95]. The cyanuric chloride vapors are condensed to molten or solid product (g) [96], [97], which is dissolved in a solvent for captive use or filled from a hopper (h) into containers (i). Tail gases are scrubbed and recycled (k).

The ClCN yield in this process exceeds 95 %; the $(ClCN)_3$ yield exceeds 90 %.

A number of processes eliminate the formation of dilute hydrochloric acid during cyanogen chloride manufacture: a one-step gas-phase reaction of hydrocyanic acid and chlorine on charcoal [98], [99] or in tetrachloroethane [100], and a one-step reaction of cyanogen and chlorine on charcoal [101].

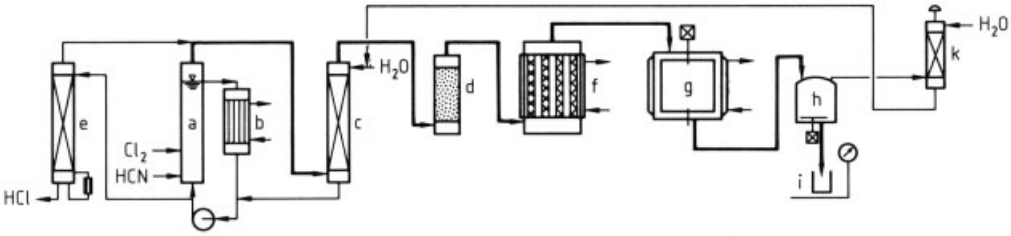

Figure 3. Production of cyanuric chloride from chlorine and hydrocyanic acid
a) ClCN reactor; b) Heat exchanger; c) Scrubber; d) Dryer; e) Effluent stripper; f) Trimerizer; g) Condenser; h) Storage; i) Drum; k) Tail gas scrubber

Catalysts for cyanogen chloride trimerization other than activated carbon include fused cyanuric chloride [102], [103], molten aluminum chloride [104], eutectic mixtures of tetrachloroaluminates [105], aluminum silicates [106], and zeolites [107] doped with metal oxides or sulfides [108].

In some of these processes, "tetrameric cyanogen chloride" (2,4-dichloro-6-isocyanodichloro-5-triazine) [*877-83-8*] is a byproduct; its conversion into cyanuric chloride is catalyzed by iron oxide [109], [110].

In the laboratory, ClCN in benzene or chloroform solution trimerizes in the presence of hydrogen chloride [111]. The use of the hydrogen chloride – dimethyl ether azeotrope on a commercial scale has been patented [112].

Equipment for cyanogen chloride production is lined with glass or fluoro polymers, or made of graphite or special resins. For cyanuric chloride, nickel, stainless steels, or aluminum is used. Moisture causes severe corrosion.

3.4. Environmental Protection

Normally the only effluent is dilute hydrochloric acid from cyanogen chloride production. This acid is usually kept below the concentration of the H_2O – HCl azeotrope, since the unwanted hydrolysis of ClCN depends on the concentration of the acid. Uses for this weak acid are very limited; appreciable amounts are neutralized with caustic soda or limestone, or in ion exchangers.

Residual gases from the trimerization are washed or condensed to recycle excess chlorine and unreacted ClCN. The tail gases consist of nitrogen and carbon dioxide; when they contain chlorine, they can be used to make bleach (NaOCl).

Spilled molten cyanuric chloride is cooled by covering it with foam to prevent sublimation. The solid is collected and hydrolyzed to cyanuric acid. Fumes from venting of tanks containing molten cyanuric chloride are scrubbed with the anhydrous solvent used in the consumer's process.

3.5. Quality Specifications and Analysis

The industrial product contains 99 % cyanuric chloride (min.) and less than 1 % insolubles. In many applications, cyanuric chloride is in suspension, so particle size is important. To facilitate reaction and avoid dust formation, a particle size between 10 and 160 µm (> 80 mesh Tyler) is optimal. Different sieving methods give different results because of abrasion or agglomeration.

To test reactivity, cyanuric chloride and an amine are reacted under standard conditions and the unchanged material is determined.

Hydrolysis caused by improper handling may result in contamination with cyanuric acid, which is found as insoluble matter in dried solvents such as toluene.

Cyanuric chloride can be assayed via the melamine complex by IR spectroscopy after hydrolysis of a chloroform extract [113], or by titration of excess reagent after reaction with morpholine [114].

3.6. Storage and Transportation

The solid is shipped in drums, with or without polyethylene liners, or in bulk bins. Free-flowing

cyanuric chloride is also transported in fluidized bulk trailers; molten product is shipped in heated tank trailers.

Cyanuric chloride should be stored in strictly dry conditions. The solid should be stored in a cool place to avoid an increase in particle size. Tank farms for molten product must have a well-designed heat-tracing system that includes the vent and relief lines.

The following transportation regulations apply:

Road (GGVS/ADR)	Class 8, 39b
Rail (GGVE/RID)	Class 8, 39b
Sea (IMDG)	Class 8; UN no. 2670, PG II
Air* (ICAO-TI/IATA/-DGR)	Class 8; UN no. 2670, PG II

* Only for small samples.

3.7. Uses

The most important cyanuric chloride derivatives are the aminotriazines. Alkylaminotriazines are used as pesticides and herbicides (**40**–**42**). The triazines are still among the largest selling herbicides; most of them are 2-chloro- or 2-methylthio-4,6-dialkylamino-1,3,5-triazines. Simazine (**42**) is one of the oldest. The most important triazine-based herbicide is atrazine (**40**).

Reaction products of aminostilbenes with substituted triazines are used as brightening agents for fabrics and paper (→ Optical Brighteners), especially bis(triazinylamino)stilbenedisulfonic acids (**43**).

Dye derivatives prepared from cyanuric chloride react chemically with fabrics (→ Reactive Dyes). The Procion dyes (**45**) were among the first.

Cyanuric chloride (**29**) and cyanurates are used as cross-linking agents. Cyanuric chloride (**29**) and especially 2,4-dichloro-6-hydroxy-1,3,5-triazine (**30**) and 2,4-dihydroxy-6-chloro-1,3,5-triazine (**31**), both in the form of their sodium salts, are used in gelatin and glues [115]; triallyl cyanurate (**39**) is used in the rubber and plastic industry (→ Allyl Compounds, Chap. 3.3.1.) [116]. Other uses of triazines include modifiers, accelerators, UV stabilizers, flame retardants, pharmaceuticals, antioxidants, antiozonants, and heavy metal scavengers.

An overview of important products based on cyanuric chloride is given in Table 3. The important class of triazine-based herbicides is surveyed in [117].

3.8. Economic Aspects

Worldwide annual production of cyanuric chloride exceeds 100 000 t/a. The available capacities are well above demand. Producers are listed in Table 4. The capacity of production units ranges from hundreds to several thousand tonnes per month.

Table 4. Manufacturers of cyanuric chloride

Country	Company
Belgium	Degussa
Federal Republic of Germany	Degussa
	SKW-Trostberg
Former Soviet Union *	
Switzerland	Lonza
United States	Novartis *
	Degussa

* Captive use only.

More than 70 % of production is used for pesticides and especially herbicides, with about 70 % of that volume for atrazine (**40**). Triazine herbicides are produced all over the world. Less than 30 % of demand is accounted for by industrial applications, predominantly optical brighteners, UV stabilizers, reactive dyes, and certain anthraquinone dyes. Applications based on cross-linking chemicals such as triallyl cyanurate (**39**) are of increasing importance.

3.9. Toxicology and Occupational Health

Cyanuric chloride strongly irritates the skin, mucous membranes, including those of the eye, and the respiratory and gastrointestinal tracts [70, pp. 2763–2765]. The 1-min irritation threshold for the mucous membranes is 0.3 mg/m^3. Allergic reactions are possible.

Direct contact with cyanuric chloride should be avoided; ventilation in the work place is necessary. A full-face gas mask with an active-carbon canister or a self-contained breathing apparatus must be used during exposure to vapor and dust. Cyanuric chloride is labeled with the hazard symbol X_i and assigned to the groups R

Table 3. Selected triazine derivatives based on cyanuric chloride

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Cyanuric chloride (CC)	starting material for herbicides, optical brighteners, reactive dyes, cross-linking agents	[*108-77-0*]	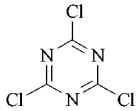	**29**
2,4-Dichloro-6-hydoxy-1,3,5-triazine sodium salt	cross-linking agent for gelatin	[*2736-18-7*]	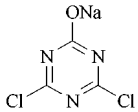	**30**
2,4-Dihydroxy-6-chloro-1,3,5-triazine disodium salt	cross-linking agent for gelatin	[*67410-55-3*]	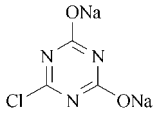	**31**
2,4-Dichloro-6-methoxy-1,3,5-triazine	pesticide intermediate	[*3638-04-8*]	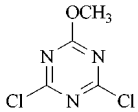	**32**
2-Chloro-4,6-dimethoxy-1,3,5-triazine	intermediate for UV-stabilizers and optical brighteners	[*3140-73-6*]	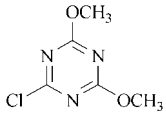	**33**
2-Methylthio-4,6-dichloro-1,3,5-triazine	pesticide intermediate	[*13705-05-0*]	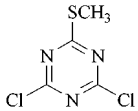	**34**
2-Amino-4,6-dichloro-1,3,5-triazine	gelatin hardener for photographic paper	[*933-20-0*]	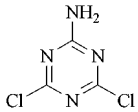	**35**
4,6-Dichloro-2-(2'-chloroanilino)-1,3,5-triazine	nonsystemic foliar fungicide with protective action	[*101-05-3*]	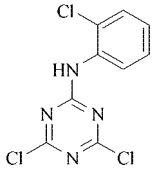	**36**

Table 3. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
2-Chloro-4,6-diamino-1,3,5-triazine	adhesion promoter for rubber and metal composites	[*3397-62-4*]	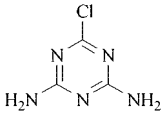	**37**
2,4,6-Tris(hydroxy-ethylamino)-1,3,5-triazine (THEM)	modifier for melamine resins	[*4403-07-0*]	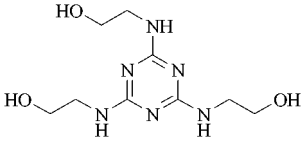	**38**
Triallyl cyanurate (TAC)	cross-linking agent	[*101-37-1*]	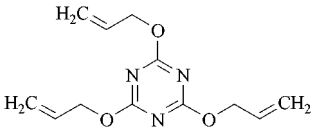	**39**
2-Chloro-4-(ethylamino)-6-(isopropylamino) triazine	herbicide	[*1912-24-9*]	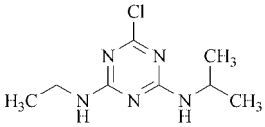	**40**
2-[[4-chloro-6-(ethylamino)-*s*-triazin-2-yl]amino]-2-methylpropionitrile	herbicide	[*21725-46-2*]	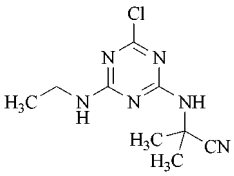	**41**
Chlorobis(ethylamino)triazine	herbicide	[*122-34-9*]	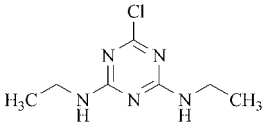	**42**
Bis(triazinylamino)stilbensulfonic acids	optical brighteners		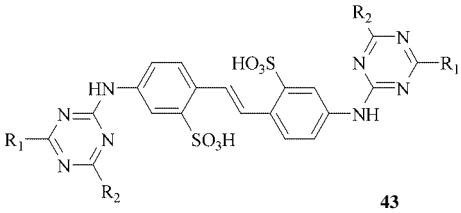	**43**

Table 3. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
2,4-Dimethoxy-6-(1-pyrenyl)-1,3,5-triazine	optical brightener for polyester fibers	[*3271-22-5*]	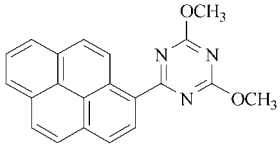	**44**
Procion Brilliant Orange M-G	reactive dye	[*6522-74-3*]	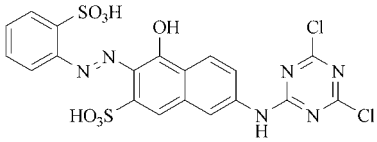	**45**
2,4,6-Trismercapto-1,3,5-triazine sodium salt (TMT)	heavy metal scavenger	[*17766-26-6*]	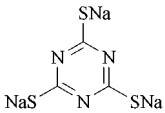	**46**
2,4-Diamino-6-(cyclopropylamino)-1,3,5-triazine	insect growth regulator for leaf miner control	[*66215-27-8*]	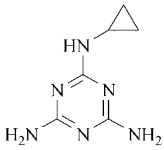	**47**
2,4,6-Tris(1-aziridinyl)-1,3,5-triazine	cross-linking agent; thermal-transfer printing materials	[*51-18-3*]	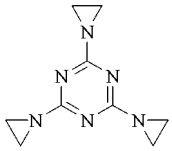	**48**
2,4,6-Tris-(dimethylamino)-1,3,5-triazine	antitumor agent	[*645-05-6*]	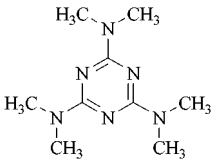	**49**

36/37/38; S 28. The following R and S phrases should also be applied:

R 22	harmful if swallowed
R 26	very toxic by inhalation
R 41/43	may cause sensitization by inhalation and skin contact
S 36/37/39	wear suitable protective clothing, gloves and eye/face protection
S 45	in case of accident or if you feel unwell, seek medical advice immediately

Funnels for adding cyanuric chloride powder to flammable solvents must be grounded. Cyanuric chloride should only be handled by qualified personnel.

Table 3. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
Poly-{[6-[(1,1,3,3-tetramethylbutyl)-amino]-1,3,5-triazine-2,4-diyl][1,6-bis(2,2,6,6-tetramethyl-piperidylamino)hexamethylene]}	UV stabilizer	[*70624-18-9*]	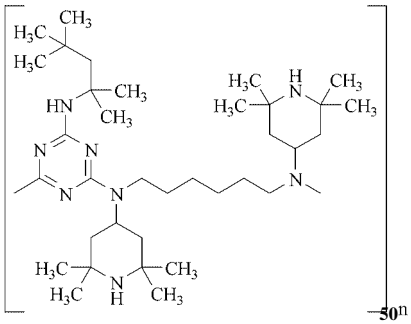	**50**
Poly-{[6-morpholino-1,3,5-triazine-2,4-diyl][1,6-bis(2,2,6,6-tetramethyl-4-piperidylamino)hexamethylene]}	UV stabilizer	[*82451-48-7*]	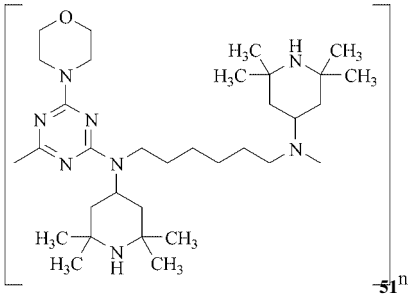	**51**
1,5,8,12-Tetrakis[4,6-bis(*N*-butyl-*N*-1,2,2,6,6-pentamethyl-4-piperidylamino-1,3,5-triazin-2-yl]-1,5,8,12-tetrazadodecane	UV stabilizer	[*106990-43-6*]	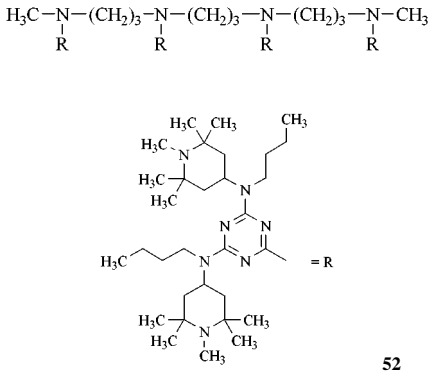	**52**

Table 3. (Continued)

Compound (abbreviation)	Application	CAS registry no.	Structure	Number
2,4,6-Tris[anilino-4-(carbo'-2'-ethyl-hexyl-1'-oxyl)]-1,3,5-triazine	UV stabilizer for cosmetics	[*88122-99-0*]	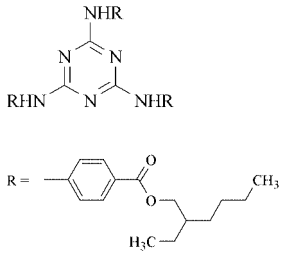	**53**
2,4,6-Tris(2,4,6-tribromo-phenoxy)-1,3,5-triazine	flame retardant	[*25713-60-4*]	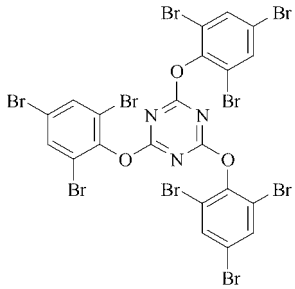	**54**
2,4-Bis(*n*-octylthio)-6-(4-hydroxy-3,5-di-*tert*-butyl-anilino)-1,3,5-triazine	heat stabilizer, antioxidant	[*991-84-4*]	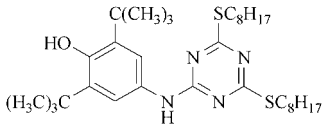	**55**

4. References

General References

1. E. M. Smolin, L. Rapoport in A. Weissberger (ed.): *The Chemistry of Heterocyclic Compounds,* **vol. 13,** Interscience, New York 1959.
2. E. J. Modest in R. C. Elderfield (ed.): *Heterocyclic Compounds,* **vol. 7,** J. Wiley & Sons, New York 1961, pp. 627 – 719.
3. *Gmelin,* **system no. 14, Carbon, D 1,** 366 – 381; **D 3,** 272 – 287.
4. *Kirk-Othmer,* **7,** 397 – 410.
5. *Beilstein,* **E III/II, 26,** 632 ff.
6. J. M. E. Quirke in A. J. Boulton, A. McKillop (eds.): *Comprehensive Heterocyclic Chemistry,* **vol. 3,** Pergamon Press, Oxford 1984, pp. 457 – 530.

Specific References

7. P. J. Wheatleg, *Acta Crystallogr.* **8** (1955) 224.
8. M. L. Tosato, *J. Chem. Soc. Perkin Trans. 2* 1979, 1371.
9. L. Paoloni, M. L. Tosato, M. Cignitti, *J. Heterocycl. Chem.* **5** (1968) 533.
10. F. Beilstein, *Justus Liebigs Ann. Chem.* **116** (1860) 357 – 358.
11. F. Zobrist, H. Schinz, *Helv. Chim. Acta* **35** (1952) 2387.
12. FMC Chemicals: *Cyanuric Acid,* Product Promotion Bulletin 10 B, 1965.
13. American Cyanamid, GB 598 175, 1948.
14. *Kirk-Othmer,* **7,** 1979, p. 397.
15. H. Verbrueggen, K. Krolikiewicz, *Chem. Ber.* **117** (1984) 1523.
16. FMC, US 3 318 888, 1967 (J. H. Blumbergs, D. G. MacKellar).

17. R. W. Cummins, *J. Org. Chem.* **28** (1963) 85 – 89.
18. Th. C. Frazier, E. D. Little, B. E. Lloyd, *J. Org. Chem.* **25** (1960) 1944 – 1946.
19. Devoe & Raynolds, US 2 809 942, 1957 (H. G. Cooke).
20. Ciba-Geigy, US 3 547 918, 1967 (D. Porret, K. Metzger, A. Heer).
21. Z. N. Pasenko, L. I. Chovnik, *Ukr. Khim. Zh.* **30** (1963) no. 2, 195 – 198; *Chem. Abstr.* **61** (1964) 1866.
22. Allied: *Triallyl Isocyanurate,* Product Bulletin.
23. FMC, US 3 075 979, 1963 (J. J. Tazuma, R. Miller).
24. A. A. Sayigh, H. Ulrich, *J. Chem. Soc.* 1961, 3148 –3149.
25. F. W. Hoover, H. B. Stevenson, H. S. Rothrock, *J. Org. Chem.* **28** (1963) 1825 – 1830.
26. M. Linhard, *Z. Anorg. Allg. Chem.* **236** (1938) 200 – 208.
27. Sun Chem., US 2 444 828, 1948 (W. Kaplan).
28. G. Ostrogovich, R. Bacaloglu, E. Fliegl, *Tetrahedron* **24** (1968) 2701 – 2705.
29. Allied, US 3 093 641, 1963 (R. L. Formaini).
30. BASF, GB 1 042 174, 1962 (O. B. Claren, L. Unterstenhoefer).
31. FMC, US 2 943 088, 1959 (R. H. Westfall).
32. Toa Gosei Chem. Ind. Co., JP 27 182, 1963; *Chem. Abstr.* **64** (1966) 9748.
33. Lentia, DE-AS 1 179 215, 1961 (A. Schmidt, F. Weinrotter, W. Muller).
34. Allied, US 3 394 136, 1966 (W. P. Moore, D. E. Elliott).
35. Nippon Soda Co., EP 0 012 773, 1978 (H. Kizawa, R. Ichihara, T. Yao, I. Kikuchi).
36. Monsanto, US 3 154 545, 1964 (W. F. Symes, S. Varopolus).
37. BASF, DE-AS 1 065 420, 1958 (O. B. Claren).
38. Shikoku Kasei Kogyo Co., US 3 275 631, 1966 (H. Yanagizawa).
39. Stamicarbon, US 4 112 232, 1978 (M. G. R. T. de Cooker).
40. Dow Chemical, US 3 810 891, 1974 (J. M. Lee).
41. FMC, US 3 563 987, 1971 (S. Berkowitz).
42. Allied, US 3 296 262, 1967 (H. F. Scott).
43. Monsanto, US 3 357 979, 1967 (E. C. Sobocinski, W. F. Symes).
44. Allied, US 3 172 886, 1965 (I. Christoffel, D. P. Schulz).
45. J. Saldick, *Appl. Microbiol.* **28** (1974) 6, 1004 – 1008.
46. D. C. Wolf, J. P. Martin, *J. Environ. Qual.* **4** (1975) 1, 134 – 139.
47. H. L. Jensen, A. S. Abdel-Ghaffar, *Arch. Microbiol.* **67** (1969) 1 – 5.
48. Ciba-Geigy, US 4 274 955, 1978 (J. Zeyer, R. Hutter, P. Mayer).
49. L. Nebbia, F. Guerrieri, B. Pagani, *Chim. Ind. (Milan)* **39** (1957) 2, 81 – 83 *Abstr.* **51** (1957) 7241.
50. G. A. Harlow, C. M. Noble, G. E. A. Wyld, *Anal. Chem.* **28** (1956) 784 – 791.
51. R. C. Cundiff, P. C. Markunas, *Anal. Chem.* **28** (1956) 792 – 797.
52. P. Beilstein, R. Huettler, A. M. Cook, *J. Agric. Food Chem.* **29** (1981) 1132. T. Sugita et al., *Bull. Environ. Contam. Toxicol.* **44** (1990) 567. A. Ghiorghis, A. H. Talebian, *J. Liq. Chromatogr.* **14** (1991) 2331.
53. J. Struys, P. M. Wolfs, *Anal. Chim. Acta* **199** (1987) 173.
54. C. J. Pouchert: *The Aldrich Library of Infrared Spectra,* 3rd ed., Aldrich Chemical Co. Inc., 1989.
55. Z. A. Malumyan et al., *Prom-st. Arm.* **43** (1985).
56. FMC, US 2 988 471, 1961 (R. J. Fuchs, J. A. Lichtman).
57. Food Service Corp., US 2 808 332, 1957 (P. C. Anderson, J. L. C. Rapp).
58. J. Kamlet, US 2 768 895, 1956 (J. Kamlet).
59. Olin Mathieson, FR 1 328 696, 1962 (R. B. East).
60. Monsanto: *Cyanuric Acid,* Bulletin no. IC/505/233.
61. Boing Airplane Co., US 3 108 902, 1963 (J. R. Galli, R. B. Johnson).
62. American Cyanamid, US 3 044 913, 1962 (L. A. Lundberg).
C. E. Schildknecht in J. L. Koschwitz, (ed.): *Encyclopedia of Polymer Science and Engineering,* **vol. 4,** John Wiley & Sons, Inc, New York, 1986, pp. 802 – 811.
Triallyl isocyanurate, product bulletin, Allied Chemical Corp., Morristown, N. J., 1973.
63. Allied, US 3 088 948, 1963 (E. D. Little, B. T. Poon).
64. Ciba-Geigy: *Merkblatt Araldit* PT 810, Publ. No. 24 384, Chem. Mark. Rep. 5 (June 20, 1977).
65. Mitsubishi Electric Industries Ltd., US 4 298 518, 1981 (O. Yasuhiro, M. Yukinobu, H. Rioji). E. I. DuPont de Nemours & Co. Inc., US 4 360 616, 1983 (R. U. Pagilagan). Occidental Chemical Corp., US 4 504 611, 1985 (C. S. Ilarda, J. J. Duffy).

66. T. Wang, D. Zhang, *Guti Runhua* **9** (1989) 137. Kyodo Fat and Oil Co. Ltd., JP 02 016 193, 1990 (Y. Nakanishi et al.).
67. Spencer Chemical Co., GB 912 563, 1961.
68. H. Yanagizawa, *Plast. Ind. News* **1968,** 16 – 18.
69. E. Canelle, *Am. J. Pub. Health* **64** (1974) 155 – 162.
70. C. F. Reinhardt, M. R. Britelli in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industriel Hygiene and Toxicology,* 3rd ed., **vol. 2A,** Wiley-Interscience, New York 1981, pp. 2765 – 2769.
71. MSD, Cyanursäure, Version 1.01, Elf-Atochem Deutschland GmbH.
72. S. R. Sandler, *J. Org. Chem.* **35** (1970) 3967 – 3968.
73. E. Korber, R. Raetz, *J. Org. Chem.* **27** (1962) 2509.
74. T. Venkataraman, D. R. Wagle, *Tetrahedron Lett.* (1979) no. 32, 3037.
75. *J. C. S. Chem. Comm.* (1972) 1226.
76. *Synthesis* (1980) 657.
77. *Tetrahedron Lett.* (1968) no. 55, 5819.
78. *Tetrahedron Lett.* (1985) no. 26, 2901.
79. *Tetrahedron Lett.* (1980) no. 21, 1893.
80. Katayama Chemical Works Co. Ltd., JP 63 255 261, 1988 (A. Tsuda, K. Tsuji).
81. *Tetrahedron Lett.* (1971) no. 29, 2765.
82. *Chem. Ber.* **121** (1988) 2, 271.
83. *J. Org. Chem.* **35** (1979) 3967.
84. *Chemiker Ztg.* **2** (1972) 96.
85. *Chem. Ind. (London)* (1971) 1516.
86. H. Fierz-David, M. Matter, *J. Soc. Dyers Colour.* **53** (1937) 424.
87. Degussa: *Cyanuric Chloride,* Product Bulletin, 1985.
88. Institution of Chemical Engineers: *Loss Prevention Bulletin,* 1979, no. 025.
89. J. T. Cupton et al., *J. Org. Chem.* **45** (1980) 4522.
90. V. L. Mur, *Russ. Chem. Rev. (Engl. Transl.)* **33** (1964) 92 – 103.
91. *Houben-Weyl,* **vol. E9c,** p. 667 ff.
92. Degussa, US 2 672 398, 1950 (H. Huemer, H. Schulz, W. Pohl).
93. Geigy Chem., US 3 197 273, 1961 (E. B. Trickey).
94. American Cyanamid, US 2 491 459, 1945 (J. T. Thurston).
95. Degussa, DE 842 067, 1950 (H. Huemer, H. Schulz).
96. Degussa, US 2 734 058, 1956 (H. Schulz, H. Huemer).
97. Degussa, US 4 245 090, 1981 (R. Goedecke, M. Liebert, W. Nischk, W. Plötz, K. Puschner).
98. Nitto, JP 2282-3, 1959.
99. Monsanto, US 2 965 642, 1960 (E. E. Drott, G. D. Oliver).
100. Saint Gobain, FR 1 251 359, 1959 (A. J. Courtier, H. Jean).
101. Röhm & Haas, DE-AS 1 068 265, 1958 (H. Zima).
102. Krupp, DE-AS 1 291 745, 1965 (W. Kirchhof, K. Schauerte). *Chem. Abstr.* **70** (1969) 106 573.
103. Bayer, GB 718 806, 1954.
104. ICI, GB 566 827, 1945 (T. P. Metcalfe).
105. Toyama, GB 1 086 706, 1967.
106. Nitto, JP 23 509, 1962 (N. Umehara, J. Ishikura).
107. Nitto, JP 9 136, 1958.
108. Bayer, DE 1 179 213, 1964 (A. v. Friedrich).
109. Bayer, US 3 080 367, 1959 (A. v. Friedrich, P. Schmitz).
110. Lonza, CH 396 020, 1966 (C. Zinsstag, R. Gentili).
111. O. Diels, *Ber. Dtsch. Chem. Ges.* **32** (1899) 691.
112. Lonza, US 2 838 512, 1958 (M. Teysseire, H. P. Sieber).
113. R. Bacaloglu, E. Fliegl, G. Ostrogovich, Fresenius'Z. *Anal. Chem.* **257** (1971) 273 – 274; *Chem. Abstr.* **76** (1972) 67940.
114. Y. V. Lyande, V. V. Savenkova, A. A. Cherkasskii, *Zh. Anal. Khim.* **30** (1975) 9, 1817 – 1821; *Chem. Abstr.* **84** (1976) 53 629.
115. E. Zerlotti, *Nature (London)* **214** (1967) 1304 – 1306.
116. Degussa, *Triallylcyanurat,* Product Bulletin, 1980.
117. *Ullmann,* 4th ed., **vol. 12,** 1976, p. 597 ff.

Cyclododecanol, Cyclododecanone, and Laurolactam

THOMAS SCHIFFER, Degussa-Hüls AG, Marl, Federal Republic of Germany

GEORG OENBRINK, Degussa-Hüls AG, Marl, Federal Republic of Germany

1. Cyclododecanol and Cyclododecanone 1
1.1. Properties 1
1.2. Production 2
1.3. Uses 3
2. Laurolactam 3
2.1. Properties 3
2.2. Production 3
2.3. Uses 5
3. Toxicology 5
4. References 5

1. Cyclododecanol and Cyclododecanone

Commercial production of cyclododecanol [*1724-39-6*] (**3**) and cyclododecanone [*830-13-7*] (**4**) began when WILKE et al. [1], [2] found an efficient synthesis for 1,5,9-cyclododecatriene [*4904-61-4*] (**1**)by cyclotrimerization of butadiene [*106-99-0*]. First made byDuPont in the United States and Hüls in Europe, world production capacityhas since risen to ca. 5×10^4 t/a. Cyclododecanol and cyclododecanone are mainly used to synthesize dodecanedioic acid [*693-23-2*] (**5**) and azacyclotridecane-2-one [*947-04-6*] (laurolactam, **8**).

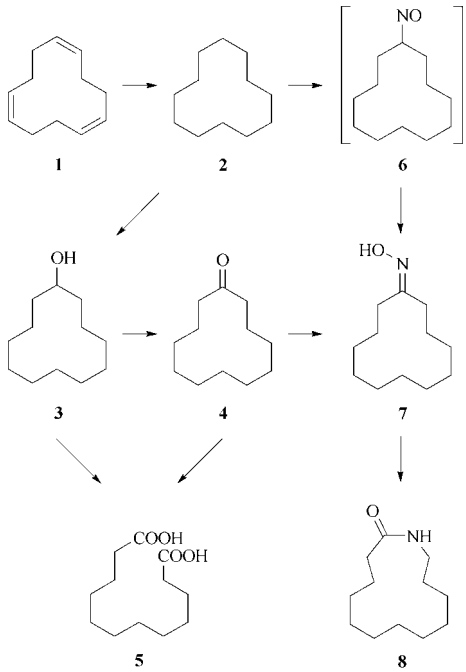

1.1. Properties

At room temperature, both cyclododecanol, $C_{12}H_{24}O$, and cyclododecanone, $C_{12}H_{22}O$, are colorless crystals that are soluble in organic solvents but only slightly soluble in water (0.004 wt % at 298 K). Table 1 lists further physical properties for both.

Cyclododecanol reacts like an aliphatic secondary alcohol. Dehydration of cyclododecanol in dilute sulfuric acid or in the gas phase with catalysts yields cyclododecene [*1501-82-2*]. When treated with organic acids or anhydrides, cyclododecanol forms esters. Oxidation of cyclododecanol with nitric acid gives dodecanedioic acid (**5**). Dehydrogenation of cyclododecanol at about 230 °C yields cyclododecanone, and hydrogenation of cyclododecanol produces cyclododecane [*294-62-2*] (**2**).

In most cases, cyclododecanone reacts like an aliphatic ketone [3]. Hydroxylamine converts cyclododecanone into cyclododecanone oxime [*9466-89-4*] (**7**), Beckmann rearrangement of which yields laurolactam (**8**). Addition of hydrocyanic acid to cyclododecanone produces cyanohydrin (95 % yield) [4]. Treatment of cyclododecanone with elemental bromine yields 2-bromocyclododecanone and 2,11-dibromocyclododecanone [5], which are used as precursors for fragrances and musks.[6] Aldol condensation [7] and Wittig reaction [8] are also known for cyclododecanone.

Although cyclododecanone gives pure cyclododecanol when hydrogenated under pressure in the presence of a copper – chromium cat-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_201

Table 1. Physical properties

Property	Cyclododecanol	Cyclododecanone	Laurolactam
Molecular mass M_r	184.3	182.3	197.3
Melting point, °C	78.4	60.8	151.8
Boiling point, °C			
at 101.3 kPa	278	274.8	
at 8.0 kPa	190.8	178.4	
at 1.3 kPa	146.2	130.4	
Vapor pressure at 180 °C, kPa			0.35
Density ϱ, g/cm^3			
at 100 °C	0.909	0.888	
at 160 °C			0.90
at 200 °C	0.817	0.819	
Refractive index, D line, at 80 °C	1.4692	1.4581	
Heat of fusion, kJ/mol	15.1	16.6	16.1
Heat of vaporization, kJ/mol	65.1	56.2	91.4
Heat of formation, kJ/mol	−557	−482	
Specific heat, J g^{-1} K^{-1}			
at 100 °C	2.92	2.39	
at 200 °C	3.02	2.81	
Thermal conductivity, kJ m^{-1} h^{-1} K^{-1}	0.741	0.477	
Viscosity, mPa s			
at 100 °C	6.24	2.28	
at 150 °C	1.60	1.10	
at 200 °C	0.70	0.70	
Flash point, °C	137	119	192
Ignition point, °C	380	280	320 – 330

alyst, more severe conditions produce cyclododecane. Upon oxidation with nitric acid, cyclododecanone, like cyclododecanol, yields dodecanedioic acid; however, the oxidation of cyclododecanol is more efficient.

1.2. Production

Cyclododecane (**2**), which is obtained by catalytic hydrogenation of 1,5,9-cyclododecatriene (**1**; → Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene), is the starting material for producing cyclododecanol and cyclododecanone.

Oxidation of cyclododecane with air or oxygen gives cyclododecanol and cyclododecanone [9]. The standard enthalpies of reaction are − 205 and − 395 kJ/mol, respectively. Cyclododecane hydroperoxide [*2698-67-1*] is involved in the oxidation as intermediate, but in general it is decomposed directly by a catalyst. Boric acid has found broad industrial application as catalyst [10–13].

Continuous or batch air oxidation (boric acid catalyst, atmospheric pressure, 160 – 180 °C) converts 25 – 30 % of the cyclododecane. Because boric acid esterifies the cyclododecanol product, the reaction mixture (cyclododecane, cyclododecyl borate, cyclododecanone) is treated with water and alkali at 90 – 100°C to hydrolyse the ester. After separation of the aqueous and organic phases, cyclododecane is distilled from the organic phase under reduced pressure and recycled for further oxidation. The aqueous phase is concentrated, and boric acid can be crystallized as orthoboric acid, dried, and also returned to the oxidation section. The selectivity of the overall reaction producing a mixture of cyclododecanol and cyclododecanone from cyclododecane is 80 – 82 mol % with a molar ratio of 10/1 (alcohol/ketone).

Pure cyclododecanol is made by hydrogenating the mixture in the presence of a copper – chromium catalyst at 30 MPa and 160°C. Pure cyclododecanone is obtained by dehydrogenating the mixture over copper or copper – chromium catalysts on an active support at 230 – 245 °C and atmospheric pressure [19].

After dehydrogenation (endothermic, 75.4 kJ/mol) the reaction mixture is distilled to give pure cyclododecanone, cyclododecanol, and a small amount of residue.

Figure 1 shows a flow diagram for the process.

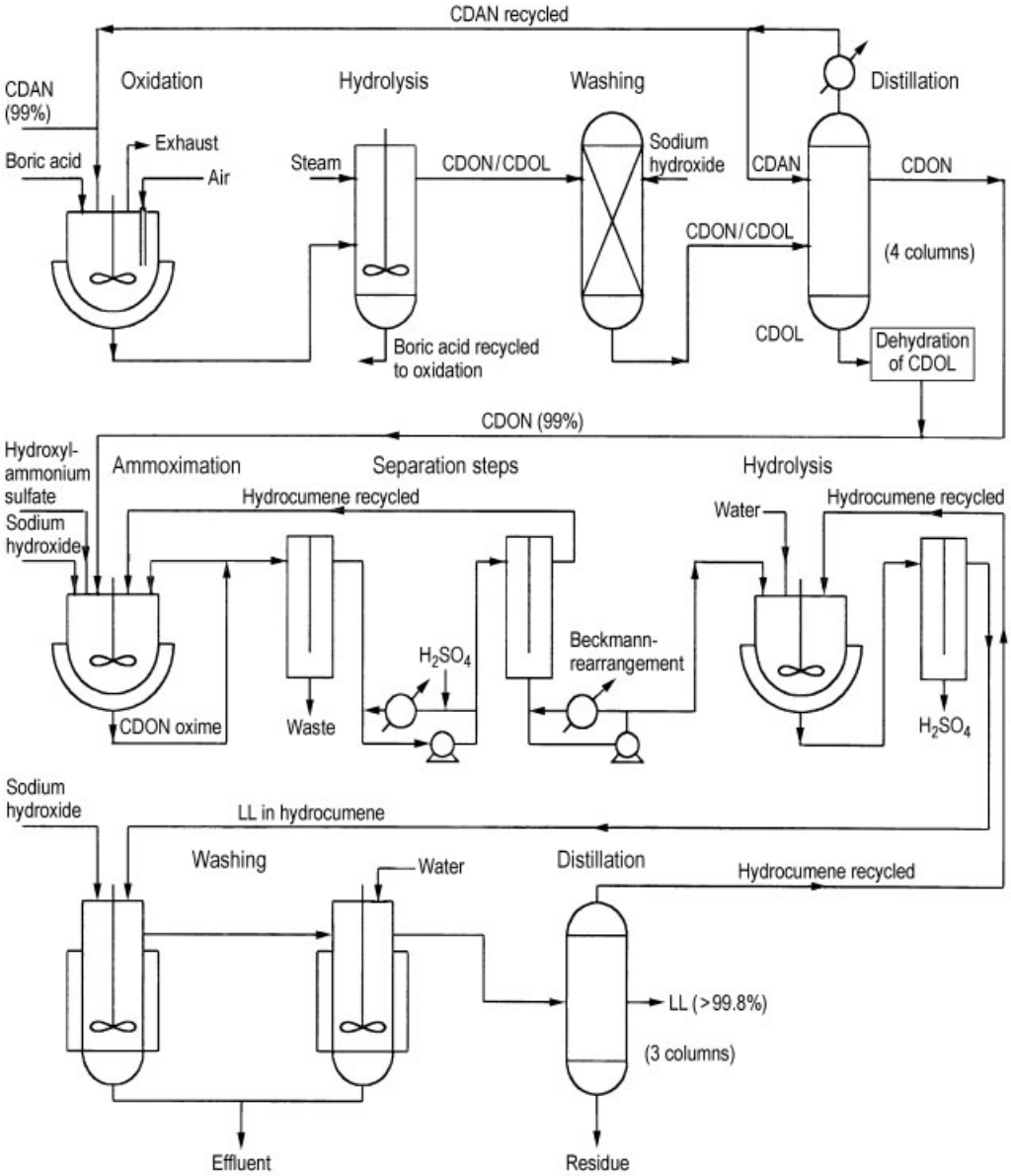

Figure 1. Process flow diagram for the production of cyclododecanol, cyclododecanone, and laurolactam (Hüls process) CDAN = cyclodecane, CDOL = cyclododecanol; CDON = cyclododecanone, LL = laurolactam

Storage and Transportation. Cyclododecanol and cyclododecanone are preferably stored and transported in stainless steel containers under a dry inert atmosphere. To prevent crystallisation, vessels or tanks must be heated.

Analysis. Although the melting point generally indicates the purity of cyclododecanone, gas chromatography is the best method for evaluating the purity of cyclododecanone or cyclododecanol.

1.3. Uses

The cyclododecanonol – cyclododecanone mixture from boric acid catalyzed oxidation is the starting material for dodecanedioic acid (**5**) which is used to prepare polyamides like polyamide 612, copolyamides, polyesters, and esters. More recently, dodecanedioic acid has found broad application in coatings [15–17]. Pure cyclododecanone is starting material for the production of laurolactam (**8**). Cyclododecanone and cyclododecanol are intermediates in organic syntheses. The esters and ethers of cyclododecanol are used in fragrances.

2. Laurolactam

Made by Degussa-Hüls, Ems-Ube, and Elf-Atochem, production capacity of azacyclotridecan-2-one [*947-04-6*] (laurolactam, **8**) is about 4×10^4 t/a.Laurolactam is mainly used as monomer in engineering plastics like polyamide 12 and copolyamides.

2.1. Properties

At room temperature, laurolactam, $C_{12}H_{23}NO$ forms colorless crystals that are soluble in organic solvents but only slightly soluble in water (0.03 wt % at 20 °C). Table 1 lists the physical properties of laurolactam.

2.2. Production

There are two main routes for commercial production of laurolactam. Degussa-Hüls and Ems-Ube use cyclododecanone (**4**) as starting material. Ammoximation of pure cyclododecanone with hydroxylammonium sulfate provides cyclododecanone oxime [*946-89-4*] (**7**), which undergoes Beckmann rearrangement in the presence of strongmineral acids at elevated temperatures (115 °C). Figure 1 shows a process flow diagram [18].

Because the oxime melts at 135 °C but begins to decompose at 110 °C, the cyclododecanone (purity 99 %), which is dissolved in a hydrocarbon like hydrocumene, is treated at 90 – 100 °C with an aqueous solution of hydroxylamine sulfate. After removal of the organic phase, the oxime is extracted from the organic phase with cold concentrated sulfuric acid.

Heating the acid solution of the oxime to 100 – 120 °C produces laurolactam. The enthalpy of the reaction (cyclododecanone oxime → laurolactam) is 166 kJ/mol. Water is added, and the crude lactam is extracted into an organic solvent, which is distilled to give the pure laurolactam (95 % yield).

Elf Atochem (France) starts from cyclododecane (**2**) and nitrosyl chloride. In a photoreaction, a nitrosyl radical substitutes a hydrogen atom of cyclododecane to give the unstable nitrosylcyclododecane (**6**), which undergoes tautomeric transformation into cyclododecanone oxime (**7**). Beckmann rearrangement follows, as described above [14].

Storage and Transportation. Laurolactam is stored and transported in molten form in pipelines or stainless steel containers under a dry inert atmosphere. To prevent crystallization containers must be heated. Crystalline laurolactam is handled as powder or granulate in bags.

Analysis. The melting point generally indicates the purity of laurolactam. Gas chromatography is used for further evaluation.

2.3. Uses

Polymerization of laurolactam provides polyamide 12. Hydrolytic polymerization of laurolactam has found broad industrial application, though activated cationic and anionic polymerisation are also well known.

3. Toxicology

The oral LD_{50} of both cyclododecanol and cyclododecanone for rats is > 2.5 g/kg. Contact of cyclododecanol with the skin, eyes, or mucous membranes causes no irritation. Although cyclododecanone has no effect on the eyes or mucous membranes, it does irritate the skin slightly [20].

The oral LD_{50} of laurolactam for rats is 2.33 g/kg; laurolactam does not irritate the skin or mucous membranes.

4. References

1. G. Wilke, *Angew. Chem.* **69** (1957) 397, **75** (1963) 10; *Angew. Chem. Int. Ed. Engl.* **2** (1963) 105.
2. H. Breil, P. Heimbach, M. Kröner, H. Müller, G. Wilke, *Makromol. Chem.* **69** (1963) 18.
3. K. Kosswig, *Chem. Ztg.* **96** (1972) 373.
4. W. Kirchhoff, W. Stumpf, W. Franke, *Justus Liebigs Ann. Chem.* **681** (1965) 32.
5. W. Ziegenbein, *Chem. Ber.* **94** (1961) 2989.
6. Firmenich, CH 589 642, 1977 (J. J. Becker).
7. Studiengesellschaft Kohle m.b.H., DE 2 417 357, 1975 (R. Koester, A.-A. Pourzal).
8. Firmenich, EP 0 032 713, 1981 (A. Uijttewaal, R. Snowden).
9. F. Broich, H. Grasemann, *Erdöl Kohle Erdgas Petrochem.* **18** (1965) 360.
10. Bayer, DE 2 260 114, 1974 (H. Rosenbaum, R. Thiel).
11. Chemische Werke Hüls, DE 1 643 825, 1971 (H. Grasemann).
12. Halcon, NL 6 600 797, 1966.
13. Toa Gosei Chem. Ind., DE 2 101 633, 1971 (H. Kato, T. Matsubara).
14. Chemische Werke Hüls, DE 1 248 650, 1964 (M. zur Hausen, W. Knepper).
15. BASF Lacke und Farben, EP 0 826 018, 1995 (D. Möller, J. Woltering).
16. BASF Lacke und Farben, EP 0 858 471, 1996 (E. Sapper, J. Woltering).
17. Kansai Paint Co., US 5 763 535, 1997 (T. Adachi, N. Numa).
18. F.-E. Baumann in *Kunststoff Handbuch, Vol. 3/4 "Polyamide"* , Chaps. 4.1 – 4.4 , Hanser Verlag, München, 1998.
19. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 4th ed., VCH, Weinheim 1994, p. 284.
20. Hüls AG, unpublished results.

Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene

THOMAS SCHIFFER, Degussa-Hüls AG, Marl, Federal Republic of Germany

GEORG OENBRINK, Degussa-Hüls AG, Marl, Federal Republic of Germany

1. Cyclododecatriene 1
1.1. Properties 1
1.2. Production; Storage and Transportation; Analysis 2
1.3. Uses 2
2. Cyclooctadiene and 4-Vinylcyclohexene 2
2.1. Properties 3
2.2. Production; Storage and Transportation; Analysis 3
2.3. Uses 3
3. Toxicology 3
4. References 3

1. Cyclododecatriene

Commercial production of 1,5,9-cyclododecatriene [*4904-61-4*]began when WILKE et al. [1], [2] found an efficient synthesis by cyclotrimerization of butadiene [*106-99-0*]. Since the first plants were brought onstream by Hüls in 1965 and by DuPont in 1969, world production capacityof cyclododecatriene has risen to 10^5 t/a. Today the main producers of cyclododecatriene are Degussa-Hüls and Shell Chemical in Europe, Du Pont in the United States, and Mitsubishi Chemicals in Asia.

1.1. Properties

While the all-*trans* isomer of 1,5,9-cyclododecatriene melts at 34 °C, the other three isomers have melting points below room temperature. Melting points of technical mixtures depend on the catalyst used and the ratio of isomers that it provides. Further physical properties of the four isomers of 1,5,9-cyclododecatriene, $C_{12}H_{18}$, M_r 108.18, and their reduction products are listed in Table 1. All isomers are colorless and have a typical terpene-like odor.

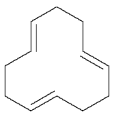

trans,trans,trans-

cis,trans,trans-

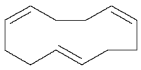

cis,cis,trans-

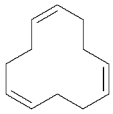

cis,cis,cis–

1,5,9-Cyclododecatriene

The two main isomers, *cis,trans,trans-* and all-*trans*-1,5,9-cyclododecatriene have been studied extensively [3]. They are generally characterized by their tendency to form complexes with transition metals and to undergo transannular reactions and isomerization.

All isomers behave like typical olefins. Complete bromination forms hexabromocyclododecane [*3194-55-6*]. Selective hydrogenation of cyclododecatriene provides cyclododecene [*1501-82-2*], the remaining double bonds of which exhibit a decline in reactivity . This declining reactivity also leads to good yields of many classical double bond reactions, such as epoxidation, nitrosation [4], hydrocarboxylation, and carbonylation [5]. At higher temperature and pressure complete hydrogenation takes place to yield cyclododecane [*294-62-2*].

1.2. Production; Storage and Transportation; Analysis

1,5,9-Cyclododecatriene is produced by the cyclotrimerization of butadiene in the presence of catalysts based on titanium, chromium, or nickel [6]. The catalysts are prepared by reducing an appropriate salt in a hydrocarbon solvent with an aluminum alkyl. Titanium catalysts [7] produce predominantly the *cis, trans,trans* isomer, whereas nickel and chromium catalysts provide the all-*trans* isomer. In general, the yield of cyclododecatriene is greater than 80 % with some dimerization byproducts and butadiene oligomers. Before distillation of the reaction

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_205

Table 1. Physical properties of cyclododecatriene, cyclooctadiene, and derivatives

Component	CAS registry no.	*bp*, °C 2.7 kPa	*bp*, °C 101.3 kPa	*mp*, °C	Relative density d_4^{20}	Refractive index n_D^{20}
trans,trans,trans-1,5,9-Cyclododecatriene	[*676-22-1*]	110	234	+34	0.864 (40 °C)	1.500
cis,trans,trans-1,5,9-Cyclododecatriene	[*706-31-0*]	116	240	−17	0.892	1.508
cis,cis,trans-1,5,9-Cyclododecatriene	[*2765-29-9*]	117	244	−8		1.513
cis,cis,cis-1,5,9-Cyclododecatriene	[*4736-48-5*]	110		−1		1.510 (25 °C)
trans-Cyclododecene	[*1486-75-5*]	118			0.872	1.485
cis-Cyclododecene	[*1129-89-1*]	121			0.875	1.486
Cyclododecane	[*294-62-2*]	122	243	+61	0.830 (65 °C)	1.455 (65 °C)
cis,trans-1,5-Cyclooctadiene	[*5259-71-2*]			−62	0.873	1.493
cis,cis-1,5-Cyclooctadiene	[*1552-12-1*]	50	151	−70	0.881	1.494
trans-Cyclooctene	[*931-89-5*]	42	145	−59	0.848	1.476
cis-Cyclooctene	[*931-87-3*]	43	145	−12	0.847	1.476
Cyclooctane	[*292-64-8*]	48	151	+14	0.835	1.459
4-Vinylcyclohexene	[*100-40-3*]		129	−101	0.831	

product, the catalyst must be deactivated, generally by an aqueous caustic wash.

All isomers of cyclododecatriene and their hydrogenation products are preferably stored in tanks and transported in pipelines or stainless steel or aluminum containers as a liquid. If the all-*trans* product is a major isomer, tanks, pipelines and containers must be heated slightly. Generally, the unsaturated compounds are covered with an inert gas to avoid oxidation reactions, and should contain an inhibitor unless they are to be processed immediately.

Gas chromatography is the best method for evaluating the purity of cyclododecatriene and side products.

1.3. Uses

Cyclododecatriene is an important precursor in organic chemistry. Bromination provides hexabromocyclododecane, which has found industrial application as a flame retardant [8]. Hydrogenation of cyclododecatriene is the most significant reaction, providing cyclododecane as an important intermediate in the synthesis of laurolactam [*947-04-6*], 12-aminododecanoicacid [*693-23-2*], and dodecanedioic acid [*693-23-2*]. These are important monomers for nylon 12, nylon 612, copolyamides [9], polyesters, and coating applications [10].

2. Cyclooctadiene and 4-Vinylcyclohexene

Cyclodimerization of 1,3-butadiene provides 1,5-cyclooctadiene and 4-vinylcyclohex-1-ene as main products. Reaction process and kinetics were examined intensively by WILKE et al. [11] in the late 1960s, but it was only ten years later that commercial production started. World capacity of butadiene cyclodimerization products is estimated at ca. 10 000 t/a, mainly produced by Degussa-Hüls and Shell.

2.1. Properties

1,5-Cyclooctadiene [*111-78-4*] and 4-vinylcyclohexene [*100-40-3*], both C_8H_{12}, M_r 108.18 are colorless liquids with specific odors. Physical properties are listed in Table 1.

cis,cis-1,5-Cyclooctadiene

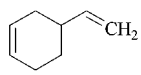

4-Vinylcyclohexene

Cyclooctadiene and 4-vinylcyclohexene both react like typical olefins. Complete halogenation provides tetrahalogenated compounds. Due to declining double bond activity, selective hydrogenation of cyclooctadiene leads to cyclooctene [*931-88-4*]. At high temperature and pressure hydrogenation provides cyclooctane [*292-64-8*].

Hydrogenation of vinylcyclohexene causes disproportionation to give a mixture of ethylcyclohexane [*1678-91-7*] and ethylbenzene [*100-41-4*] [12]. Dehydrogenation of vinylcyclohexene provides styrene [*100-42-5*] [13].

Further typical olefin reactions such as epoxidation, hydrocarboxylation, and carbonylation are known for both compounds.

2.2. Production; Storage and Transportation; Analysis

When nickel catalysts are modified with appropriate ligands, preferably organophosphorus compounds, cyclodimerization of butadiene takes place instead of cyclotrimerization, and the stable *cis, cis* isomer of cyclooctadiene is obtained in more than 90 % yield with 4-vinylcyclohexene and cyclododecatriene as byproducts. Generally the reaction takes place at about 80 °C and 1 – 5 bar, though higher pressure and temperature are also used. The products are separated and purified by distillation.

When iron complexes [14], [15] or copper catalysts [16] are used instead of nickel, 4-vinylcyclohexene becomes the main dimerization product, which provides a benzene-free route to styrene [17].

Butadiene dimerization products are preferably stored and transported in stainless steel or aluminum containers in a dry, inert atmosphere.

Refractive index and gas chromatography are good methods for evaluating the purity of butadiene dimerization products.

2.3. Uses

Ring opening polymerization of cyclooctene, obtained by the selective hydrogenation of cyclooctadiene, is the most significant industrial application. The macrocyclic polymer, *trans*-polyoctenamer (trade name Vestenamer) with a molecular mass of up to 100 000 is used as modifier in rubbers and thermoplastics.

Small amounts of the monomer are used as a precursors in organic synthesis of fragrances, pharmaceuticals, etc. Halogenated cyclooctadienes and vinylcyclohexenes have found application as flame retardants.

3. Toxicology

Contact with cyclododecatrienes, cyclooctadiene, and vinylcyclohexene slightly irritates the skin. Oral LD_{50} values (rat) follow: cyclooctadiene 1.9 g/kg, 4-vinylcyclohexene 4.55 g/kg, cyclododecatriene 4.4 g/kg. Due to the tendency to form epoxides, the cycloolefins are suspected of causing cancer, especially vinylcyclohexene. Therefore, in 1997 vinylcyclohexene was classified in Category III A2 [18].

4. References

1. G. Wilke, *Angew. Chem.* **75** (1963) 10 – 20; *Angew. Chem. Int. Ed. Engl.* **2** (1963) 105 – 16.
2. B. Bogdanovic, P. Heimbach, M. Kröner, G. Wilke, E. G. Hofmann, J. Brandt, *Liebigs Ann. Chem.* **727** (1969) 143 – 16.
3. K. Kosswig, *Chem. Ztg.* **96** (1972) 373 – 38.
4. Studienges. Kohle, DE 1 058 987, 1957 (G. Wilke).
5. K. Bittler et al., *Angew. Chem.* **80** (1968) 352; *Angew. Chem. Int. Ed. Engl.* **7** (1968) no. 5, 329 – 335. J. Graefe et al., *Tetrahedron* **26** (1970) 2677.
6. C. W. Bird: *Oligomerization of Olefins in Transition Metal Intermediates in Organic Synthesis,* Logos Press – Academic Press, London 1966. P. Heimbach, R. Traummüller: *Ring Synthesis with Nickel Catalysts in Chemistry of the Metal-Olefin Complexes,* Verlag Chemie, Weinheim 1970. P. Heimbach, P. W. Jolly, G. Wilke: "Pi-Allyl-Nickel Intermediates in Organic Synthesis," in E. G. A. Stone, R. West (eds.): Advances in Organometallic Chemistry, vol. 8, Academic Press, New York–London 1970.
7. H. Weber, W. Ring, U. Hochmuth, W. Franke, *Liebigs Ann. Chem.* **681** (1965) 10 – 2.

8. BASF, EP 0 004 050, 1979 (K. Haan et al.). Hüls, DE 2 906 336, 1980 (J. K. Rigler et al.).
9. L. Bottenbruch, R. Binsack (eds.): *Kunststoff Handbuch,* **vol. 3/4, "Polyamide"** , Chaps. 4.1 – 4.3, Hanser Verlag, München, 1998.
10. BASF, DE 4 341 235, 1995 (C. Hilger, J. Woltering, K. Cibura). PPG Industries, US 5 407 707, 1995 (R. L. Simeone et al.). Nippon Paint Co., EP 0 653 466, 1995 (A. Uenaka et al.).
11. W. Brenner, P. Heimbach, H. Hey, E. W. Müller, G. Wilke, *Liebigs Ann. Chem.* **727** (1969) 161 – 18.
12. Phillips Petroleum, US 3 855 327, 1974 (W. G. Billings).
13. Texaco, US 4 163 761, 1979 (J. A. Patterson, W. C. Crawford, J. R. Wilson). Mitsubishi Chemical, DE 2 824 649, 1978 (T. Okano, T. Masuyama, T. Yokoyama).
14. A. Yamamoto, K. Morifuji, S. Ikeda, T. Saito, Y. Uchida, A. Minoso, *J. Am. Chem. Soc.* (1968) 1878 – 188.
15. D. Huchette, B. Thery, F. Petit, *J. Mol. Catal.* **4** (1978) 433 – 442. DSM, EP 0 667 841, 1993 (H. A. M. Duister, J. G. D. Haenen).
16. Dow Chemical, US 5 329 057, 1994 (R. W. Diesen et al.). Phillips Petroleum, US 5 625 101, 1997 (T.-T. P. Cheung, M. M. Johnson).
17. Chem Systems, *Styrene from Butadiene*, New York, 1995.
18. Deutsche Forschungsgemeinschaft: *MAK und BAT-Werte Liste 1997,* Wiley-VCH, Weinheim, 1997.

Cyclohexane

M. Larry Campbell, Exxon Chemical Co., Florham Park, New Jersey 07932, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 2
4.1. **Hydrogenation of Benzene** 3
4.2. **Separation from Naphtha** 5
4.3. **Cost Considerations** 5
5. **Quality Specifications and Tests** 7
6. **Uses, Economic Aspects** 7
7. **Homologues** 7
7.1. **Cyclohexene** 7
7.2. **Methylcyclohexane** 7
7.3. **Methylcyclopentane** 7
8. **Toxicology and Occupational Health** . 8
9. **References** 8

1. Introduction

Cyclohexane [*110-82-7*], C_6H_{12}, is a cycloalkane, which was synthesized by Baeyer in1893 and discovered in Caucasian crude oil by Markovnikov soon afterward. Its presence in U.S. crude oils was reported in 1931 [1]. Cyclohexane was first synthesized by hydrogenation of benzene in 1898 [2]. The predominant use of cyclohexane is in the manufacture of nylon.

2. Physical Properties

Like all cycloalkanes and related cycloolefins, cyclohexane is a clear liquid with a pungent petroleum-like odor. It is essentially insoluble in water but miscible in most organic liquids, and is noncorrosive, easily flammable, and considered to be much less toxic than benzene.

Cyclohexane can exist in two conformations, the chair and boat forms:

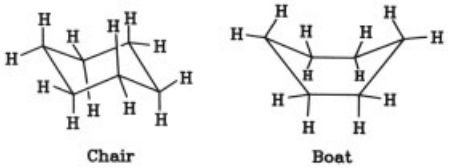

Transformation from one isomer to the other involves rotation about carbon – carbon single bonds. The energy barriers associated with the transformation are low, but the boat form is estimated to have a potential energy 21 – 25 kJ/mol higher than the chair form [3]. As a result, cyclohexane is almost exclusively in the chair form at room temperature. Stereochemistry has no influence on its use as a raw material for nylon manufacture or in other applications.

Table 1 lists the important physical properties of cyclohexane; two closely related naphthenes, methylcyclopentane and methylcyclohexane; and the olefinic homologue, cyclohexene. Table 2 compares the key properties of some alkylcyclohexanes with cyclohexane. Tables 3, 4, and 5 give boiling points and compositions for azeotropes of cyclohexane with various compounds.

Although cyclohexane does not ignite spontaneously at ambient temperature, it must be kept away from an open flame. Information on the explosive limits and ignition properties for cyclohexane and related compounds is given in Table 6.

3. Chemical Properties

Cyclohexane is a relatively stable, nonpolar ring compound. Therefore, with thermal treatment, little isomerization to methylcyclopentane or ring-opening occurs, but benzene is formed by dehydrogenation. At high temperature (700 – 800 °C) cyclohexane decomposes to butadiene and other products. Aluminum chloride readily isomerizes cyclohexane to methylcyclopentane [7].

The most important commercial reaction of cyclohexane is its liquid-phase *oxidation* with air in the presence of soluble cobalt or boric acid catalysts to produce a mixture of cyclohexanol and cyclohexanone (KA oil) (→ Cyclohexanol

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_209

Table 1. Physical properties of cyclohexane, methylcyclohexane, cyclohexene, and methylcyclopentane

	Cyclohexane	Methylcyclo-hexane	Cyclohexene	Methylcyclo-pentane
CAS registry number	[*110-82-7*]	[*108-87-2*]	[*110-83-8*]	[*96-37-7*]
Formula	C_6H_{12}	C_7H_{14}	C_6H_{10}	C_6H_{12}
M_r	84.157	98.183	82.141	84.157
bp, °C	80.738	100.934	82.979	71.812
fp, °C	6.554	−126.593	−103.512	−142.455
ϱ_4^{20}, g/cm^3	0.778 55	0.769 39	0.810 96	0.74864
n_D^{20}	1.426 23	1.423 12	1.446 54	1.40970
Heat of fusion, kJ/mol	2.6787	6.7550	3.2954	6.9333
Entropy of fusion, J mol^{-1} K^{-1}	9.575	46.088		53.046
Heat of vaporization, kJ/mol				
at 25 °C	33.059	35.383		31.610
at *bp*	29.977	31.150		29.098
Entropy of vaporization, J mol^{-1} K^{-1}				
at 25 °C	110.879	118.675		106.022
at *bp*	84.70	83.275		84.351
Heat of combustion [a] at 25 °C, kJ/mol	3 922.45	4 568.34	3 752.92 [b]	3 940.36
Heat of formation [c], kJ/mol	−123.22	−154.870	−5.35	−106.76
Entropy at 25 °C J mol^{-1} K^{-1}	298.435	343.569		340.136
ΔG°, kJ/mol	31.778	27.298	106.86	35.80
Critical temperature, °C	280.3	298.97	287.4	259.58
Critical pressure, MPa	4.07	3.469	4.347	3.785
Critical density, g/cm^3	0.273	0.267	0.281	0.264
Dynamic viscosity at 20 °C, mPa · s	0.977	0.732	0.980	0.505
Kinematic viscosity at 20 °C, mm^2/s	1.259			
c_p at 300 K J mol^{-1} K^{-1}	107.098	136.11	105.84	110.66
at 800 K J mol^{-1} K^{-1}	279.51	329.67	249.07	267.12
Surface tension, mN/m	24.98	23.70	26.56	22.19
Constants for Antoine equation [d], kPa [4]				
A	13.7377	196.174	13.8093	13.7873
B	2766.3	146.6	2813.53	2731.00
C	−50.50	374.1	−49.98	−47.11
Dielectric constant at 20 °C and 10^5 Hz	2.021			
Cryoscopic constant[e], mole fraction/ °C	0.00411			

[a] High heating value.
[b] At 30 °C.
[c] Gaseous.
[d] $\ln p = A - B/(C + T)$.
[e] For use in calculating mol % purity, *p*, by using the equation $\log p = 2.00000 - (A/2.302\,59)\,(T' - T)$ where T' is the freezing point of a given sample [5], [6].

and Cyclohexanone, Chap. 3.2.). Cyclohexanol is oxidized to cyclohexanone, which is used to produce caprolactam, a monomer for nylon-6 production (→ Caprolactam, Chap. 4.; → 4. Fibers, Synthetic Organic). Alternatively, KA oil can be converted to adipic acid and hexamethylenediamine, the monomers for nylon-66 production (→ Adipic Acid; → Hexamethylenediamine) [8].

4. Production

Virtually all cyclohexane is produced commercially by hydrogenation of benzene (→ Benzene, Chap. 10.). A small amount is produced by superfractionation of the naphtha fraction from crude oil. Naturally occurring cyclohexane can be supplemented by fractionating methylcyclopentane from naphtha and isomerizing it to cyclohexane.

4.1. Hydrogenation of Benzene

Benzene can be hydrogenated catalytically to cyclohexane in either the liquid or the vapor phase in the presence of hydrogen. Several cyclohexane processes, which use nickel, platinum, or palladium as the *catalyst*, have been developed. Usually, the catalyst is supported, e.g., on alumina, but at least one commercial process utilizes Raney nickel.

Table 2. Some physical properties of alkylcyclohexanes

Compound, CAS registry number	*bp* ∗, °C	*fp*, °C	d_4^{20}	$n_D{}^{20}$
Cyclohexane [*110-82-7*]	80.738	+6.554	0.77855	1.42623
Methylcyclohexane [*108-87-2*]	100.934	− 126.593	0.76939	1.42312
1,1-Dimethylcyclohexane [*590-66-9*]	119.543	− 33.495	0.78094	1.42900
cis-1,2-Dimethylcyclohexane [*2207-01-4*]	129.728	− 50.023	0.79627	1.43596
trans-1,2-Dimethylcyclohexane [*6876-23-9*]	123.419	− 88.194	0.77601	1.42695
cis-1,3-Dimethylcyclohexane [*638-04-0*]	120.088	− 75.573	0.76603	1.42294
trans-1,3-Dimethylcyclohexane [*2207-03-6*]	124.450	− 90.108	0.78472	1.43085
cis-1,4-Dimethylcyclohexane [*624-29-3*]	124.321	− 87.436	0.78285	1.42966
trans-1,4-Dimethylcyclohexane [*2207-04-7*]	119.351	− 36.962	0.76255	1.42090
1,1,3-Trimethylcyclohexane [*3073-66-3*]	136.626	− 65.750	0.77883	1.42955
1,1,2-Trimethylcyclohexane [*7094-26-0*]	145.2	− 29	0.8000	1.4382
1,1,4-Trimethylcyclohexane [*7094-27-1*]	135		0.7722	1.4251
Ethylcyclohexane [*1678-91-7*]	131.783	− 111.323	0.78792	1.43304
n-Propylcyclohexane [*1678-92-8*]	156.724	− 94.900	0.79360	1.43705
Isopropylcyclohexane [*696-29-7*]	154.762	− 89.390	0.80221	1.44087
n-Butylcyclohexane [*1678-93-9*]	180.947	− 74.725	0.79918	1.44075
Isobutylcyclohexane [*1678-98-4*]	171.221	− 94.785	0.79521	1.43861
sec-Butylcyclohexane [*52993-54-1*]	179.335		0.81314	1.44673
tert-Butylcyclohexane [*3178-22-1*]	171.591	− 41.158	0.81267	1.44694
n-Decylcyclohexane [*1795-16-0*]	297.589	− 1.726	0.81858	1.45338
cis-1-Methyl-4-isopropylcyclohexane [*6069-98-3*]	172.7	− 89.80		
trans-1-Methyl-4-isopropylcyclohexane [*1678-82-6*]	170.5	− 86.35		

∗ At 101.3 kPa.

Table 3. Binary azeotropes with cyclohexane

Component	Component *bp*, °C	Cyclohexane, *wt %*	Azeotropic, *bp*, °C
Water	100.0	91.6	69.5
2,4-Dimethylpentane	80.5	48.6	80.2
2,2,3-Trimethylbutane	80.9	46.6	80.0
Benzene	80.1	48.8	77.6
		48.1	77.56
Ethanol	78.3	68.7	64.8
1-Propanol	97.2	81.5	74.7
2-Propanol	82.3	68.0	69.4
1-Butanol	117.7	90.5	79.8
Allyl alcohol	97.0	42.0	74.0
Acetone	56.3	32.5	53.0
2-Butanone	79.7	54.6	71.6
		56.0	71.5
Methyl acetate	57.0	17.0	54.9
Ethyl acetate	77.1	44.0	71.6
Vinyl acetate	72.2	37.7	67.4
2-Methoxyethanol	125.0	85.0	77.5
Thiophene	84.2	58.8 ∗	77.9
Isopropyl methyl sulfide	84.8	70.0 ∗	79.8
Nitromethane	100.8	73.5 ∗	69.5

∗ vol %.

Hydrogenation proceeds readily and is highly exothermic ($\Delta H_{500\,\mathrm{K}} = -216.37$ kJ/mol). From an equilibrium standpoint, the *reaction temperature* should not exceed 300 °C. Above this, the equilibrium begins to shift in favor of benzene so that high-purity cyclohexane cannot be produced. As a result of these thermodynamic considerations, temperature control of the reaction is critical to obtaining essentially complete conversion of benzene to cyclohexane.

Temperature control requires economic and efficient heat removal. This has been addressed in a number of ways by commercial processes. The earlier vapor-phase processes used multistage reactors with recycle of cyclohexane as a diluent to provide a heat sink, staged injection

Table 4. Ternary azeotropes with cyclohexane

Component		Component, wt %			Azeotropic
A	B	A	B	C ∗	*bp*, °C
Water	ethanol	4.7	19.7	75.5	62.6
Water	ethanol	7.0	17.0	76.0	62.1
Water	1-propanol	8.5	10.0	81.5	66.6
Water	2-butanone	5.0	35.0	60.0	63.6
Ethanol	benzene	30.4	10.8	58.8	63.1
1-Propanol	benzene	18.0	28.0	54.0	73.8
1-Butanol	benzene	4.0	48.0	48.0	77.4
2-Butanol	benzene	8.0	42.0	50.0	76.7
2-Methoxyethanol	benzene	9.0	39.1	51.9	73.0

∗ C = cyclohexane.

Table 5. Quaternary azeotrope

Component, wt %				Azeotropic *bp*, °C
Water	Ethanol	Benzene	Cyclohexane	
7.1	17.4	21.5	54.0	62.19
6.1	19.2	20.4	54.3	62.14

Table 6. Explosive limits and ignition properties

	Explosive limits				Autoignition	Flash point,
	Lower, vol %	Upper, vol %	Lower, g/m^3	Upper, g/m^3	temp, °C	°C
Ethylcyclohexane	0.9	6.6	42	310	260	−21
Benzene	1.2	8.0	39	270	555	−11
Cyclohexane	1.2	8.3	40	290	260	−18
Cyclohexene					310	< −20
Methylcyclohexane	1.1		45		260	−4

of benzene feed between reactors, and interstage steam generators to absorb the exothermic heat of hydrogenation [9], [10]. In the 1970s processes have been developed that use only one reactor [11], [12] or a combination of a liquid- and a vapor-phase reactor [13]. The objectives of the later processes were to reduce capital cost and improve energy utilization. However, all of the commercial processes have comparably low capital cost and good energy efficiency.

A generalized flow scheme for a *vapor-phase process with multistage reactors* in series is shown in Figure 1. Benzene feed is divided and fed to each of the first two reactors. Recycled cyclohexane is introduced to the first reactor along with hydrogen. The recycled cyclohexane enables higher conversion in the reactors by absorbing part of the heat of hydrogenation. Steam generators between the reactors remove the heat of hydrogenation. The outlet temperature of the last reactor is controlled to achieve essentially 100 % conversion of benzene to cyclohexane. The effluent from the last reactor is cooled, and the vapor and liquid are separated. Part of the hydrogen-rich vapor is recycled to the first reactor, and the rest is purged to fuel gas or hydrogen recovery facilities. The liquid from the separator goes to a stabilizer where the overhead gas is sent to fuel gas; the remaining material is cyclohexane product, part of which is recycled to the first reactor.

A simplified flow scheme for a *process with liquid- and vapor-phase reactors* is provided in Figure 2. Benzene and hydrogen are fed to the liquid-phase reactor, which contains a slurry of finely divided Raney nickel. Temperature is maintained at 180 – 190 °C by pumping the slurry through a steam generator and by vaporization in the reactor. Roughly 95 % of the benzene is converted in this reactor. The vapor is fed to a fixed-bed reactor where the conversion of benzene is completed. The effluent from the

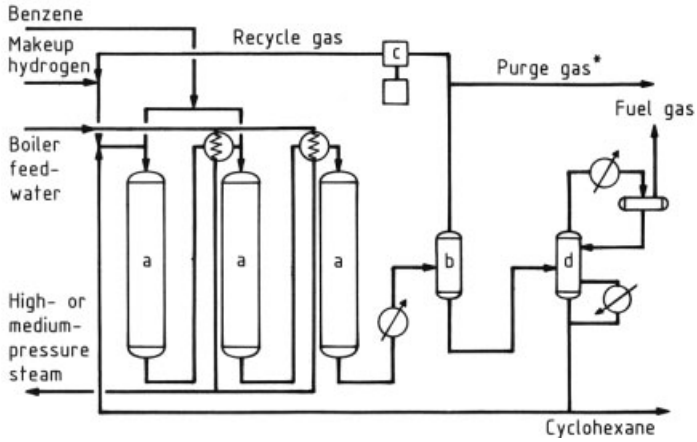

Figure 1. Multistage reactor, vapor-phase process
a) Reactors; b) Separator drum; c) Compressor; d) Stabilizer
∗ To hydrogen recovery (optional).

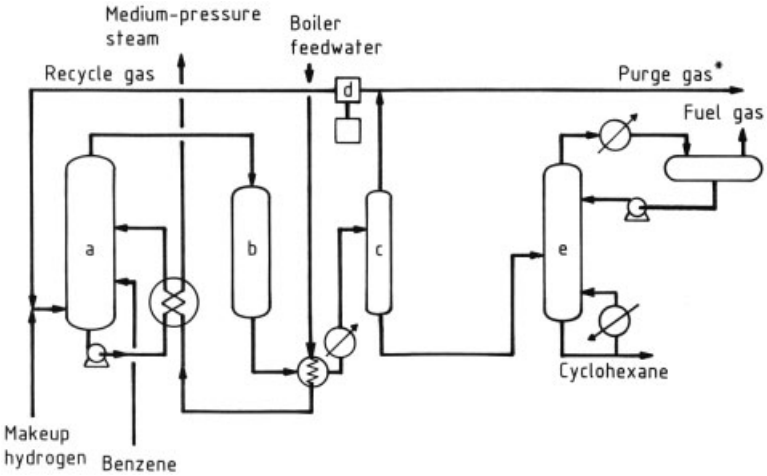

Figure 2. Liquid- and vapor-phase reactor process
a) Liquid-phase reactor; b) Vapor-phase reactor; c) Separator drum; d) Compressor; e) Stabilizer
∗ To hydrogen recovery (optional).

fixed-bed reactor is processed as described previously for the vapor-phase process.

Benzene hydrogenation is done typically at 20 – 30 MPa. The maximum reactor temperature is limited to ca. 300 °C so that a typical specification of < 500 mg/kg benzene and < 200 mg/kg methylcyclopentane in the product can be achieved. This is necessary because of the thermodynamic equilibrium between cyclohexane – benzene and cyclohexane – methylcyclopentane. Actually, equilibrium strongly favors methylcyclopentane, but the isomerization reaction is slow enough with the catalysts employed to avoid a problem if the temperature is controlled.

The *hydrogen content* of the makeup hydrogen has no effect on product purity but it does determine the makeup, recycle, and purge gas rates. Streams with as low as 65 vol % hydrogen can be used.

Carbon monoxide and sulfur compounds are *catalyst deactivators.* Both can be present in the hydrogen from catalytic naphtha reformers or ethylene units, which are typical sources of makeup hydrogen. Therefore, the hydrogen-containing stream is usually passed

through a methanator to convert carbon monoxide to methane and water. Prior to methanation, hydrogen-containing gas can be scrubbed with caustic to remove sulfur compounds. Commercial benzene contains less than 1 mg/kg sulfur. In some cases, the recycle gas is also scrubbed with caustic to prevent buildup of hydrogen sulfide from the small amount of sulfur in the benzene. With properly treated hydrogen and specification benzene, a catalyst life in excess of three years can be achieved easily in fixed-bed reactors that use noble-metal catalysts supported on a base. The catalyst in the process that uses Raney nickel in suspension is reported to have a typical life of about six months before it must be replaced [14].

4.2. Separation from Naphtha

The cyclohexane content of the naphtha fraction of crude oil can vary from 0.5 to 5.0 vol %. However, *n*-hexane, isohexanes, methylcyclopentane, benzene, and dimethylpentanes have normal boiling points very close to cyclohexane; and the freezing points of cyclohexane and benzene are separated by $< 2\,°C$. Consequently, recovery of cyclohexane from naphtha by distillation or crystallization is difficult and uneconomic.

Cyclohexane can be separated from benzene by solvent extraction or extractive distillation, but no practical solvent system has been developed for separating cyclohexane from hexanes or methylcyclopentane. Trace quantities of benzene and *n*-hexane can be removed with zeolite X and zeolite A sieves, respectively. The difficulty of separating cyclohexane from naphtha is underscored by the fact that only one company, Phillips Petroleum, produces cyclohexane by distillation (ca. 10 % of U.S. capacity). This route is not used commercially elsewhere in the world.

4.3. Cost Considerations

The investment for a benzene hydrogenation unit that produces 100 000 t/a of cyclohexane was ca. $(5-7) \times 10^6$ \$ for a unit completed in 1985. This excludes investment for hydrogen recovery or manufacture and associated facilities, such as utility systems and tankage.

The net operating cost is low because of recovery of the heat of hydrogenation by steam generation and heat integration. The major costs are maintenance expenses, labor, capital recovery charges for the plant and working capital,benzene feedstock, and hydrogen (→ Hydrogen). The cost of hydrogen can vary widely and can affect cyclohexane manufacturing cost by as much as \$ 40 /t.

The selling price of cyclohexane is tied to the selling price of benzene. In the early 1980s, its price was ca. \$ 100 /t above the benzene price.

5. Quality Specifications and Tests

The typical purity specification for cyclohexane is 99.9 wt % min. Impurities are unconverted benzene, aliphatic hydrocarbons in the benzene feed to the hydrogenation unit, methylcyclopentane formed by isomerization of cyclohexane, and methylcyclohexane formed by hydrogenation of the small amount of toluene in benzene. Freezing point, relative density, and a distillation range of 1 °C, with 80.7 °C in the range, are generally specified. A sulfur level of less than 1 mg/kg as measured by ASTM method D-3120 is also specified. In addition, specifications of phenol, chloride, and basic nitrogen content exist, but they are of little importance because the levels of these in product cyclohexane are generally less than 0.1 mg/kg. Although these specifications, with the appropriate ASTM test method, are used in contractual agreements between cyclohexane producers and their customers, gas chromatography is generally employed for quality control in manufacturing because of its precision and reproducibility.

6. Uses, Economic Aspects

Over 98 % of the cyclohexane produced is used to make nylon intermediates: adipic acid, caprolactam, and hexamethylenediamine, with the first two consuming ca. 95 % of the cyclohexane used in nylon manufacture. Minor miscellaneous uses, such as solvents and polymer reaction diluents, consume the remainder of the cyclohexane produced.

Estimated worldwide production capacity and consumption of cyclohexane are given in Table 7. Consumption dropped significantly in

Table 7. Estimated production capacity and consumption of cyclohexane, in 10^3 t/a

Year	Capacity *				Consumption			
	United States	Western Europe	Japan	World	United States	Western Europe	Japan	World
1979					1000	750	510	3150
1980					785	600	500	2850
1981					720	650	480	2800
1982	1480	940	580	4250	470	600	430	2465
1983	1510	940	580	4280	665	725	460	2890
1984	1330	940	530	4050	860	710	480	3140
1995	1400	1100	740	5100				

* At end of year.

the early 1980s because of the generally depressed economy. The future growth in cyclohexane consumption is forecast to be 2 – 3 % /a. Its growth is limited because nylon is a mature product and other routes to nylon intermediates, which use butadiene, acrylonitrile, or phenol, have been developed. The utilization of cyclohexane recovered to slightly over 75 % of capacity in 1984 because consumption increased and several units were shut down.

7. Homologues

7.1. Cyclohexene

Cyclohexene can be used to synthesize many interesting products of potential commercial value. However, no viable process has been developed for producing it in large volume. The reason for this is the difficulty in hydrogenating benzene to cyclohexene because the ΔG° (298 K) for converting benzene to cyclohexene is -22.8 kJ/mol, which is considerably less than that for benzene to cyclohexane [ΔG° (298 K) = -97.9 kJ/mol]. Consequently, the hydrogenation reaction tends to go all the way to cyclohexane.

Despite the thermodynamic difficulty, considerable work on heterogeneous catalyst systems has been reported [15–18]. Ruthenium-based systems appear to be the best, but selectivities for cyclohexene are < 30 %; the rest of the product is cyclohexane.

7.2. Methylcyclohexane

The primary use of methylcyclohexane is as a component in the feed to catalytic naphtha reformers where it is readily dehydrogenated to toluene, a high-octane motor gasoline component. Some methylcyclohexane is used for solvent applications in which it is one component of a narrow boiling range fraction of naphtha. Like cyclohexane, it can be produced easily by hydrogenation of toluene, but no current market exists for large-scale quantities.

7.3. Methylcyclopentane

The primary use of methylcyclopentane, like methylcyclohexane, is as a component in the feed to catalytic naphtha reformers where it is isomerized to cyclohexane and, in turn, dehydrogenated to benzene. Methylcyclopentane is a minor component in some light aliphatic solvents, but it is not produced as a pure component in significant volume.

8. Toxicology and Occupational Health

Cyclohexane is mildly irritating to the skin. Both the TLV – TWA and MAK are 200 ppm, 700 mg/m^3; the TLV – STEL was listed in 1986 – 1987 as (375 ppm), (1300 mg/m^3), with a note of intention to delete this value [19], [20]. At levels above the TWA, cyclohexane can cause irritation of the eyes and mucous membranes. When high concentrations are inhaled, narcosis may occur. In contrast to benzene, little or no effect on the blood, even after long exposure, has been reported.

Methylcyclohexane is somewhat less narcotic than cyclohexane so a higher TLV – TWA of 400 ppm, 1600 mg/m^3 has been established; the MAK is 200 ppm, 810 mg/m^3; the TLV – STEL was (500 ppm), (2000 mg/m^3)

in 1986 – 1987, with the same notation as for cyclohexane. The toxic properties of cyclohexene (TLV – TWA and MAK of 300 ppm, 1015 mg/m^3) and methylcyclopentane (TLV and MAK have not been established) are thought to be similar to those of cyclohexane.

9. References

1. J. H. Bruun, M. M. Hicks-Bruun, *J. Res. Natl. Bur. Stand., Sect. A* **7** (1931) 607.
2. P. Sabatier, *Ind. Eng. Chem.* **18** (1926) 1005.
3. W. C. Dauber, K. S. Pitzer in M. S. Newman (ed.): *Steric Effects in Organic Chemistry,* J. Wiley & Sons, New York 1956, pp. 13 – 15.
4. R. C. Reid, J. M. Prausnitz, J. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977, App. A.
5. *Selected Values of Hydrocarbons and Related Compounds,* A.P.I. Research Project 44, Carnegie Press, Pittsburgh, Pa., 1960, p. 1.
6. P. R. Dreisbach, *Adv. Chem. Ser.* **15** (1955) 441.
7. N. I. Shuikin et al., *Neftekhimiya* **1** (1961) 756; *Chem. Abstr.* **57** (1962) 8452.
8. G. Nawata, *Chem. Econ. Eng. Rev.* **4** (1972) no. 5, 47 – 61.
9. UOP, US 3 054 833, 1962 (G. R. Donaldson, V. Haensel).
10. UOP, US 2 755 317, 1956 (L. S. Kasel).
11. Texaco, US 3 767 719, 1973 (J. H. Colvert, E. F. Jones, H. C. Kaufman, R. Smith).
12. S. Field, M. H. Dolson, *Hydrocarbon Process.* **46** (1967) no. 5, 169 – 174.
13. Institut Français de Pétrole, US 3 597 489, 1971 (O. D. Vu, R. Odello).
14. B. Cha, J. Cosyns, M. Derrien, A. Forge, *AIChE Symp. Ser.* **70** (1974) no. 142, 128 – 130.
15. Stamicarbon, US 3 391 206, 1966 (F. Hartog).
16. Du Pont, US 3 767 720, 1973 (W. C. Drinkard).
17. M. M. Johnson, G. P. Nowak, *J. Catal.* **38** (1975) 518.
18. C. U. I. Oldenbrand, S. T. Lundin, *J. Chem. Technol. Biotechnol.* **30** (1980) 677 – 687; **31** (1981) 660 – 669; **32** (1982) 365 – 375.
19. ACGIH (ed.): *Threshold Limit Values (TLV),* ACGIH, Cincinnati, Ohio, 1986 – 1987.
20. DFG (ed.): *MAK- und BAT-Werte-Liste 2001,* Wiley-VCH, Weinheim 2001.

Cyclohexanol and Cyclohexanone

MICHAEL TUTTLE MUSSER, E. I. Du Pont de Nemours & Co., Sabine River Laboratory, Orange, Texas 77631, United States

1. Introduction 1
2. Physical and Chemical Properties . . 1
3. Production 2
3.1. Hydrogenation of Phenol 2
3.2. Liquid-Phase Oxidation of Cyclohexane 3
3.3. Boric Acid Modified Oxidation of Cyclohexane 4
3.4. Production and Hydration of Cyclohexene 4
3.5. Dehydrogenation of Cyclohexanol . . 6
4. Plant Safety 6
5. Quality Specifications and Chemical Analysis 7
6. Storage and Transportation 7
7. Uses and Trade Names 7
8. Derivatives 7
8.1. Esters of Cyclohexanol 7
8.2. Cyclohexanone Oxime 8
8.3. Methylcyclohexanols and Methylcyclohexanones 8
8.4. Trimethylcyclohexanols and Trimethylcyclohexanones 8
9. Economic Aspects 10
10. Toxicology and Occupational Health . 10
11. References 11

1. Introduction

Cyclohexanol [*108-93-0*], $C_6H_{11}OH$, M_r 100.16, is a colorless substance forming hygroscopic crystalline needles having a camphor-like odor. It is very soluble in alcohols, ketones, ethers, and alicyclic and aromatic hydrocarbons. In its pure state, it melts near room temperature, 25.15 °C. Cyclohexanone [*108-94-1*], $C_6H_{10}O$, M_r98.15, is a colorless to pale yellow liquid having an odor similar to that of acetone. It is also very soluble in most organic solvents.

Both cyclohexanol and cyclohexanone are produced on a large commercial scale. The vast majority of these compounds and their mixtures is consumed in the production of caprolactam and adipic acid, intermediates in the manufacture of nylon 6 and nylon 66 (→ Adipic Acid; → Caprolactam; → 4. Fibers, Synthetic Organic).

2. Physical and Chemical Properties

Physical Properties. The most important physical properties of cyclohexanol and cyclohexanone are shown in Table 1.

Chemical Properties. *Cyclohexanol* undergoes chemical reactions typical of secondary alcohols. It can esterify most organic acids or acyl halides, whereas halogen acids convert the alcohol to the corresponding cyclohexyl halide. Dehydration with sulfuric acid or vapor-phase treatment with alumina at 300 – 400 °C leads to cyclohexene. Mild oxidation or catalytic dehydrogenation gives cyclohexanone. Stronger oxidizing agents, such as nitric acid, give a good yield of adipic acid.

Cyclohexanone undergoes most of the reactions expected of aliphatic ketones. At room temperature, 0.02 % exists in the enol form [1]. Cyclohexanone reacts with bisulfite to give the addition product; it also reacts with hydrogen cyanide to give the cyanohydrin. It can be hydrogenated to cyclohexanol and under more vigorous conditions is further reduced to cyclohexane. Reactions with chlorine or bromine lead to substitution at either or both of the 2- and 6-positions. It readily undergoes base-catalyzed Aldol condensation with itself or with other aldehydes or ketones. The most important commercial reactions of cyclohexanone are with hydroxylamine to give cyclohexanone oxime (an intermediate on the way to caprolactam) and its oxidation with nitric acid to give adipic acid.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_217

3. Production

Cyclohexanol and cyclohexanone were first produced on an industrial scale by the hydrogenation of phenol. This was followed in the 1940s by transition metal-catalyzed liquid-phase air oxidation of cyclohexane, which gave a mixture of cyclohexanol and cyclohexanone, the ratio of which could be partially controlled by the choice of metal catalyst. In the 1950s, Scientific Design developed a boric acid modified liquid-phase air oxidation of cyclohexane. This process, in which the intermediate cyclohexyl hydroperoxide was trapped as the peroxoborate ester, led to a higher yield and a higher ratio of cyclohexanol to cyclohexanone. Around 1990, Asahi commercialized the production of cyclohexanol from cyclohexene, an intermediate obtained from the partial hydrogenation of benzene.

Table 1. Physical properties of cyclohexanol and cyclohexanone

Property	Cyclo-hexanol	Cyclo-hexanone
mp, °C	25.15	−47
bp, °C	161.1	156.4
Vapor pressure, kPa	0.15 (20 °C) 0.48 (35 °C)	0.52 (20 °C)
d_4^{20}	0.9493	0.9455
n_D^{20}	1.4656	1.4552
Expansion coefficient	8.25×10^{-4}	9.14×10^{-4}
Dynamic viscosity, mPa · s	41.07 (30 °C) 17.20 (45 °C)	1.803 (30 °C)
Specific heat, J/g	1.747	1.811
Surface tension, mN/m	33.47 (30 °C)	33.51 (30 °C)
Heat of fusion, kJ/mol	1.791	1.501
Heat of combustion, MJ/mol	−3.722	−3.521
Heat of vaporization, kJ/mol	45.51	44.92
Heat of formation, kJ/mol	−352	−272
Flash point, °C	68	54
Autoignition temp., °C	300	420
Solubility at 20 °C, g of compound/100 g of water	3.6	9.0
g of water/100 g of compound	12.6	5.7
Azeotropes		
With water: bp, °C	97.8	96.6
composition, water : compound	80 : 20	56 : 44
With phenol: bp, °C	183	
composition, phenol : compound	87 : 13	

Cyclohexanone can be prepared by dehydrogenation of cyclohexanol; the reverse reaction, hydrogenation of cyclohexanone to cyclohexanol, is also practical although not commercially practiced. Most large-scale cyclohexanol/cyclohexanone facilities employ some form of cyclohexane oxidation technology. The only remaining large phenol-based plants belong to DSM in Holland and Allied in the United States. Asahi has the only commercial cyclohexene-based plant, and there are reports of planned expansions at that facility in Japan as well as building a new plant in China.

3.1. Hydrogenation of Phenol

The product of phenol hydrogenation, i.e., cyclohexanol, cyclohexanone, or a mixture of the two, is determined by the metal catalyst. *Cyclohexanol* can be prepared in high yield by either vapor- or liquid-phase hydrogenation of phenol. Vapor-phase hydrogenation can be carried out by using a supported nickel catalyst or a nickel catalyst containing copper, cobalt, or manganese. The support is usually alumina or silicic acid. Yields as high as 98 % have been reported [2–5]. A flow diagram of the vapor-phase hydrogenation is shown in Figure 1.

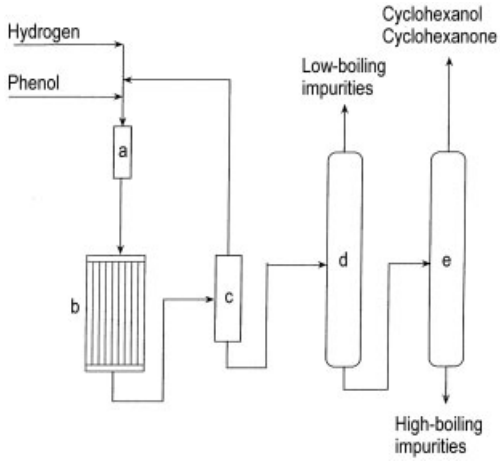

Figure 1. Vapor-phase phenol hydrogenation process
a) Phenol evaporator; b) Hydrogenation reactor; c) Condenser; d) Low-boiler removal column; e) Cyclohexanol/cyclohexanone recovery column

Hydrogenation of phenol to *cyclohexanone* can be carried out in the vapor phase with a large variety of supported noble metal catalysts, including palladium, platinum, iridium, ruthenium, rubidium, and osmium. The hydrogenation is normally carried out at 140 – 170 °C and atmospheric pressure. Yields as high as 95 % at 100 % conversion can be obtained [6], [7].

Commercial liquid-phase hydrogenation of phenol under mild conditions with a palladium on carbon catalyst gives greater than 99 % yield of cyclohexanone at 90 % conversion [8]. By adjusting the operating conditions and changing to a Raney nickel catalyst, 99.9 % selectivity for cyclohexanol can be achieved [9]. Hydrogenation conditions can be adjusted to produce the mixture of cyclohexanol and cyclohexanone that is preferred by adipic acid producers.

3.2. Liquid-Phase Oxidation of Cyclohexane

The liquid-phase air oxidation of cyclohexane [*110-82-7*] to cyclohexanol and cyclohexanone was developed in the 1940s. The oxidation is usually carried out, either uncatalyzed or with a soluble cobalt catalyst, in a series of agitated reactors at 140 – 180 °C and at 0.8 – 2 MPa, although a single tower oxidizer can be used [10], [11]. The total residence time in the oxidizers is 15 – 60 min. Some of the initially formed cyclohexyl hydroperoxide [*766-07-4*](Eq. 1) is converted in the oxidizers to cyclohexanol, cyclohexanone, and byproducts (Eq. 2).

$$C_6H_{12} + O_2 \longrightarrow C_6H_{11}OOH \quad (1)$$

$$3\,C_6H_{11}OOH \longrightarrow 2\,C_6H_{11}OH + C_6H_{10}O + H_2O + O_2 + \text{byproducts} \quad (2)$$

The cyclohexanol to cyclohexanone ratio exiting the air oxidizers can be altered by the choice of catalyst used in the air oxidizer. The normal cobalt-catalyzed air oxidation and ″deperoxidation″ gives a cyclohexanol to cyclohexanone ratio of about 3.5. If chromium(III) is added to the air oxidizer, it appears to promote dehydration of cyclohexyl hydroperoxide to cyclohexanone and water, as opposed to the free-radical decomposition mechanism of cobalt, leading to a lower cyclohexanol to cyclohexanone ratio [12–14].

Because the intermediate cyclohexyl hydroperoxide and the products, cyclohexanol and cyclohexanone, are more readily oxidized than cyclohexane, the conversion of cyclohexane in the air oxidizers must be kept low (usually under 6 %) to maximize yield. Depending on the conversion, the total yield of alcohol, ketone, and hydroperoxide varies from 70 to 90 %. Byproducts of the oxidation include a wide range of mono- and dicarboxylic acids, esters, aldehydes, and other oxygenated materials.

To maximize overall yield of cyclohexane to the alcohol and ketone, most companies try to maintain a high percentage of the oxidized cyclohexane leaving the oxidizers as cyclohexyl hydroperoxide. To maximize the amount of cyclohexyl hydroperoxide leaving the air oxidizers, the oxidation can be carried out uncatalyzed or with a modified catalyst that minimizes hydroperoxide decomposition in the oxidizer [15–20]. This permits cyclohexyl hydroperoxide to be decomposed into cyclohexanol and/or cyclohexanone under milder conditions and in higher yields than if it had occurred in the air oxidizer.

For example, the air oxidation product may or may not be washed with water followed by conversion of cyclohexyl hydroperoxide to cyclohexanol and cyclohexanone with a homogeneous or heterogeneous metal catalyst [21–35]. Another method of decomposing the hydroperoxide is to contact it with an aqueous caustic phase containing a few parts per million of cobalt, which decomposes the hydroperoxide in high yield to the ketone and alcohol [36–41]. This caustic decomposition process can result in a lower alcohol to ketone ratio in the final product. A flow diagram of this caustic process is shown in Figure 2.

Table 2. Ratio of cyclohexanol to cyclohexanone produced from cyclohexyl hydroperoxide

Solution ∗ – Modification	Catalyst	Temperature, °C	Ratio ∗∗
A	none	155	0.8
A	Cr	155	0.4
A	Ni	155	1.4
A	V	155	1.4
A	Co	155	3.0
A	Mn	155	3.3
B	Co	160	2.5
A – hydrogenation of **1**	Pd	150	5 – 10
A – separate caustic aqueous phase	Co	120	0.2

∗ A = cyclohexyl hydroperoxide (**1**) in cyclohexane; B = **1** in air oxidizer tails.
∗∗ Cyclohexanol : cyclohexanone.

Several patents report that cyclohexyl hydroperoxide in the oxidizer effluent can be hydrogenated in high yield to cyclohexanol [42–47].

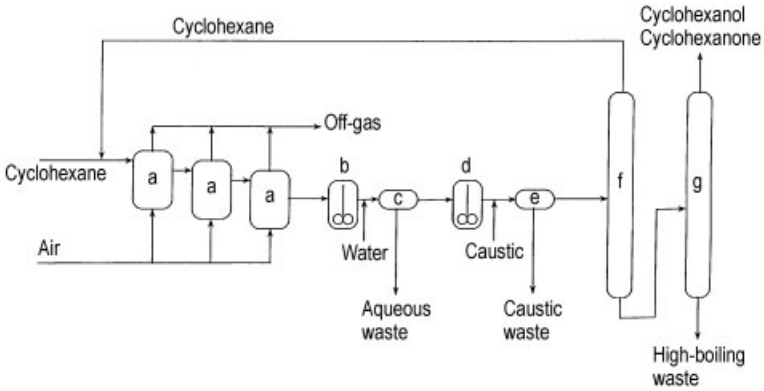

Figure 2. Cyclohexane oxidation/caustic decomposition process
a) Air oxidizers; b) Water wash decanter; c) Decantation; d) Caustic decomposition reactor; e) Caustic decantating; f) Cyclohexane recovery column; g) Cyclohexanol/cyclohexanone recovery column

Table 2 shows the cyclohexanol to cyclohexanone ratio that can be obtained by using a range of conditions and a variety of hydroperoxide decomposition methods.

3.3. Boric Acid Modified Oxidation of Cyclohexane

In the 1950s, Scientific Design (now Halcon International) developed a process in which anhydrous metaboric acid [*13460-50-9*]was added as a slurry to the first of several staged air oxidation vessels (see Figure 3) [48], [49]. The cyclohexyl hydroperoxide formed is trapped as the cyclohexyl perborate ester (Eq. 3). The perborate ester then reacts with additional cyclohexane in the air oxidizer to yield the borate ester and additional cyclohexanol (Eq. 4). This borate ester is relatively stable in the air oxidizer and is protected from over oxidation to ring-opened products. The ester is subsequently hydrolyzed to cyclohexanol and boric acid [*10043-35-3*](Eq. 5). The boric acid is dehydrated to metaboric acid and is recycled to the air oxidizer (Eq. 6).

$$18\,C_6H_{12} + 9\,O_2 + 2\,H_3B_3O_6 \longrightarrow 6\,B(OOC_6H_{11})_3 + 12\,H_2O \quad (3)$$

$$B(OOC_6H_{11})_3 + 3\,C_6H_{12} \longrightarrow B(OC_6H_{11})_3 + 3\,C_6H_{11}OH \quad (4)$$

$$B(OC_6H_{11})_3 + 3\,H_2O \longrightarrow 3\,C_6H_{11}OH + H_3BO_3 \quad (5)$$

$$3\,H_3BO_3 \longrightarrow H_3B_3O_6 + 3\,H_2O \quad (6)$$

The chemistry of this process is very similar to the metal-catalyzed oxidation except that the cyclohexyl rings are protected from further attack. Compared to other commercial processes, the boric acid process is characterized not only by a higher investment and a higher operating cost (to recover and recycle the boric acid), but also by a higher yield of cyclohexanol and cyclohexanone (up to 88 %) and an alcohol to ketone ratio as high as 10 : 1. The technology was further developed by Institut Français du Pétrole [50], [51]. The Halcon process has been licensed to several companies around the world. The major producers using this technology today are Solutia, DuPont (UK), Bayer, and Mitsubishi. The original plant at Rhône-Poulenc, based on this technology, was now shut down.

3.4. Production and Hydration of Cyclohexene

In 1972 the first patent covering the selective hydrogenation of benzene [*71-43-2*] to cyclohexene [*110-83-8*] appeared [52]. Shortly thereafter, two patents described the hydration of cyclohexene to cyclohexanol by using various catalysts [53], [54]. Since then, a large number of patents have appeared worldwide. In 1990, Asahi commercialized this process with a 100 t/a

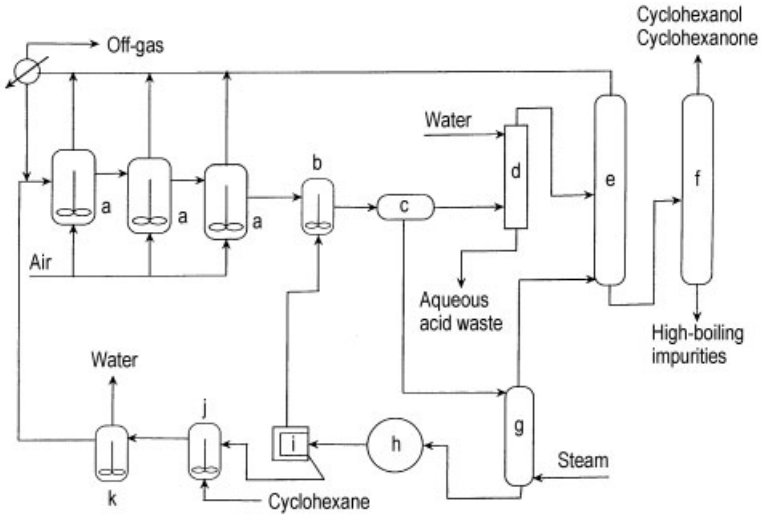

Figure 3. Boric acid modified oxidation of cyclohexane
a) Air oxidizers; b) Ester hydrolysis vessel; c) Decanter; d) Extraction column; e) Cyclohexane recovery column; f) Refining column; g) Stripping column; h) Vaccuum crystallizer; i) Boric acid centrifuge; j) Boric acid slurry tank; k) Dehydrator

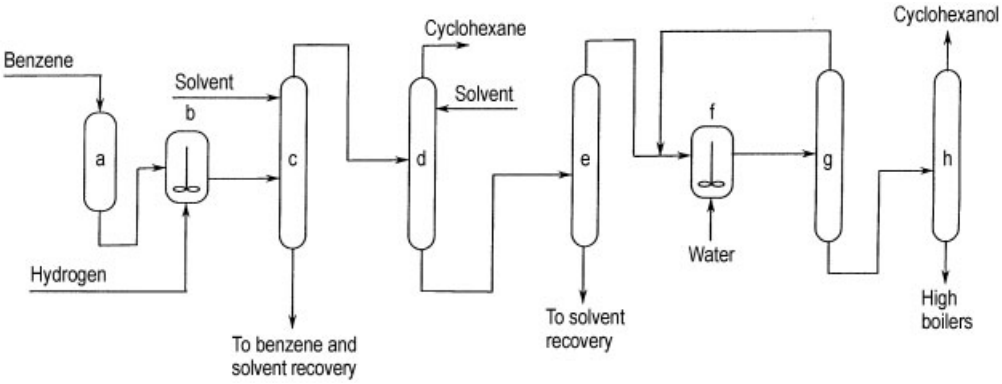

Figure 4. Partial hydrogenation of benzene to cyclohexene, followed by hydration to cyclohexanol
a) Benzene purification; b) Partial hydrogenation; c) Separation of unconverted benzene from cyclohexene and cyclohexane; d) Separation of cyclohexene from cyclohexane; e) Separation of cyclohexene from extractive solvent; f) Hydration of cyclohexene; g) Recycle of unconverted cyclohexene; h) Purification of cyclohexanol

plant in Japan. The process involves three primary steps: the selective hydrogenation of benzene; the separation of the cyclohexene from the unconverted benzene and the over-hydrogenation product cyclohexane; and the hydration of the cyclohexene (Fig. 4).

The hydrogenation is carried out with a specially prepared heterogeneous nickel catalyst. The reaction is run to about 50 % conversion, at which point the product consists of about 35 % cyclohexene, 15 % cyclohexane, and 50 % unconverted benzene. The selectivity is very sensitive to impurities such as sulfur and iron. For this reason the benzene must be specially purified and the hydrogenation vessel lined with an inert material [55–58].

The mixture of cyclohexane, cyclohexene, and benzene can not be distilled but must go through two successive extractive distillations. The best solvent to use in these distillations has been the subject of numerous patents [59–62]. The recovered cyclohexane could be converted back to benzene, but is probably sold to other caprolactam and adipic acid manufacturers.

The hydration of cyclohexene is also performed with a heterogeneous catalyst in a slurry

reactor. The extent of hydration is limited by the equilibrium constant, which would restrict conversion to about 14 %. The use of several additives to increase the conversion of cyclohexene to cyclohexanol has also been patented [63–68].

The yield of this process from benzene to cyclohexanol is extremely high, probably greater than 95 %. Although byproducts are not a concern, the investment is significantly higher than that of a conventional cyclohexane oxidation plant. The fact that the product is cyclohexanol, not cyclohexanone, could make it unattractive as the basis for a caprolactam plant.

3.5. Dehydrogenation of Cyclohexanol

Cyclohexanol can be dehydrogenated to cyclohexanone (the desired starting material for caprolactam manufacture) without catalyst by passing the vapors through a tube furnace at 400 – 450 °C. The vapors are cooled, the hydrogen separated, and the cyclohexanone, which contains cyclohexene and water, is purified in a distillation column. The purity of the resulting cyclohexanone is 98 – 99 % [69–72].

The catalytic dehydrogenation, which is generally practiced, can be carried out under milder conditions and in better yield over a variety of catalysts, including chromium oxide – copper [73], copper chromate [74], nickel [75], [76], zinc sulfide, and zinc – iron catalyst [77–79], cobalt carbonate [80], and other metals [81].

4. Plant Safety

The major industrial hazard associated with producing cyclohexanol and cyclohexanone is oxidation of cyclohexane at elevated temperature and pressure. Inside the air oxidation unit, the vapor is usually too rich in fuel (cyclohexane) to be explosive. However, in the air inlet piping, an explosive mixture can result if hot cyclohexane backs up into this system. Some of the solids formed in the oxidation, usually called ″coffee grounds,″ can be pyrophoric. At least one incident occurred in which the air sparger, entering the oxidizer, was observed to be ″glowing red,″ indicating a fire inside the sparger. An incident of a fire inside an oxidizer at the sparger nozzles has been reported by ICI.

A sudden release of the hot cyclohexane to the atmosphere, (e.g., from the rupture of a process line or a break in the air oxidation vessel itself) results in an extremely hazardous situation. An incident of this type occurred at a Monsanto plant at Pensacola, Florida in 1971. Fortunately, the resulting vapor cloud was dispersed by the prevailing winds before it could ignite.

In 1974, a much more serious incident occurred at a Nipro plant in Flixborough, England. A section of 50-cm pipe connecting two air oxidation vessels ruptured, resulting in the release of a large cloud of cyclohexane vapor. About 30 s after the break occurred, the vapor cloud ignited, and the resulting explosion destroyed the plant. A total of 28 lives were lost.

5. Quality Specifications and Chemical Analysis

High-grade commercial cyclohexanol, usually obtained from hydrogenation of phenol, has the following specifications:

mp	19 – 23 °C
bp	159 – 162 °C
Water content	0.1 %
Phenol content	0.0001 %

The best analytical method for cyclohexanol is gas chromatography with Carbowax 20M on Chromosorb. Common impurities, such as cyclohexane, cyclohexanone, phenol, pentanol, and benzene, do not interfere. Cyclohexanol can be characterized as the 4-nitrobenzoyl derivative, *mp* 52 °C. It can be determined colorimetrically by using a solution of 4-hydroxybenzaldehyde in sulfuric acid [82]. The preferred acylating agent for microscale quantities is acetic anhydride [83].

Commercial cyclohexanone is available in several grades. A typical material is colorless to pale yellow, contains 98 – 99 % cyclohexanone and 0.2 % water, and has a boiling range of 151 – 157 °C.

Cyclohexanone, like cyclohexanol, is best analyzed by using a Carbowax 20M on Chromosorb gas chromatography column. Cyclohexanone can be characterized as the oxime (*mp* 90.5 °C), the semicarbazone (*mp* 167 °C), or the nitrophenylhydrazone (*mp* 146 °C). It can be

quantitatively determined by reaction with hydroxylamine hydrochloride, followed by basic titration of the hydrogen chloride that is liberated [84], [85].

6. Storage and Transportation

Cyclohexanol and cyclohexanone are not corrosive to iron or steel and can be stored or transported in drums, tank trucks, or tank cars. An inert gas (nitrogen) blanket is not required. The U.S. DOT classifies both cyclohexanol and cyclohexanone, or mixtures thereof, as hazardous flammable liquids. For transportation, they must be labeled flammable and marked Hazard Level III [86]

7. Uses and Trade Names

Cyclohexanol has numerous industrial applications, the major two being as an intermediate in the production of adipic acid (for use in nylon 66) and cyclohexanone, which is then converted to caprolactam (the monomer for nylon 6). The U.S. market for refined cyclohexanol, other than as a nylon intermediate was 10×10^3 t in 1992 [87]. The major uses are:

1) As a solvent for lacquers, varnishes, oils, alkyd resins, gums, shellacs, ethyl cellulose, acid dyes, and natural resins
2) In the preparation of esters for plasticizers, e.g., dicyclohexyl phthalate
3) In the laundry industry and in soaps and synthetic detergents as a homogenizer and stabilizer
4) In paint and varnish removers
5) In the insecticide, fragrance, polish, and rubber cement industries

Trade names: Cyclohexyl Alcohol, Hexahydrophenol, Hydrophenol, Hexalin, and A.

Cyclohexanone. Over 97 % of all cyclohexanone produced is oxidized to adipic acid or converted to cyclohexanone oxime, which is then rearranged with sulfuric acid to caprolactam. Other uses for cyclohexanone include thinner or solvent for synthetic resins, polymers, and lacquers. It is also used as a starting material in the synthesis of many insecticides, herbicides, and pharmaceuticals, as well as being a good solvent for many insecticides.

Trade names: Hexanon, Hydrol-O, Sextone, Pimelic Ketone, Ketohexamethylene, and K.

8. Derivatives

8.1. Esters of Cyclohexanol

Numerous esters of cyclohexanol are of commercial significance, the more important being dicyclohexyl phthalate, dicyclohexyl adipate, and cyclohexyl acrylate.

Dicyclohexyl phthalate [*84-61-7*], $C_{20}H_{26}O_4$, M_r 330.45, *mp* 65 °C, *bp* 218 °C (0.6 kPa), d^{20} 1.148. This diester is normally prepared from the reaction of phthalic anhydride with cyclohexanol in an inert solvent like toluene at ca. 130 °C.

Dicyclohexyl phthalate is used as a plasticizer to modify the properties of synthetic resins, in alkyd resins and cellulose nitrate to increase their stability to light and weathering, and to improve the chemical and physical properties of plastics by preventing creep. It is very stable to heat and light and imparts a glossy finish to extruded materials. It is also used in paper finishes and makes printers ink water-resistant.

Dicyclohexyl adipate [*849-99-0*], $C_{18}H_{30}O_4$, M_r 294.50, *mp* 35.5 °C, *bp* 324 °C, d^{20} 1.037, n^{20} 1.4720.

This diester can be prepared by esterification of adipic acid, transesterification of dimethyl adipate, or direct reaction of adipic acid with cyclohexene [65]. It is used as a plasticizer in vinyl copolymers, polystyrene, and paints. A patent indicates that it can be oxidized in high yield to adipic acid [88].

8.2. Cyclohexanone Oxime

Cyclohexanone oxime [*100-64-1*], $C_6H_{10}NOH$, M_r 113.16 , *mp* 90.5 °C, *bp* 204 °C, sublimes at room temperature (0.6 kPa); it is a white crystalline solid that can be crystallized as prisms from hydrocarbon solvents. This compound is

the most important commercial derivative of cyclohexanone. It can be prepared by warming cyclohexanone with an aqueous mixture of hydroxylamine hydrochloride and sodium bicarbonate. The oxime undergoes a Beckmann rearrangement in the presence of sulfuric acid to give caprolactam.

Large commercial quantities of cyclohexanone oxime are used in the preparation of caprolactam, the intermediate for the preparation of nylon 6. The cyclohexanone oxime can also be reduced to a mixture of cyclohexylamine and dicyclohexylamine or it can hydrolyzed to cyclohexanone.

8.3. Methylcyclohexanols and Methylcyclohexanones

Methylcyclohexanol, $CH_3C_6H_{10}OH$, M_r 114.19, is usually available as a mixture of the cis and trans isomers of 2-, 3-, and 4-methylcyclohexanol. These derivatives are prepared by hydrogenation of individual or mixed cresols. Table 3 gives the physical properties of the individual isomers and of a commercial mixture.

The commercial mixture is a colorless liquid and is used as a solvent for resins, oils, and waxes, as a blending agent in soaps, and as an antioxidant in lubricants [89], [90].

Methylcyclohexanone is normally sold as a mixture of three isomers (2-, 3-, and 4-methylcyclohexanone). The individual isomers can be isolated; their physical properties, along with those of a commercial mixture, are listed in Table 3.

Methylcyclohexanones can be prepared by mild oxidation of the corresponding alcohols or by direct hydrogenation of cresol in a process similar to the hydrogenation of phenol to cyclohexanone.

The technical-grade commercial mixture contains all three isomers. At 20 °C its solubility in water is 1.5 %, and 3 % water is soluble in the methylcyclohexanones. Methylcyclohexanones are used as solvents in the dye and resin industries.

8.4. Trimethylcyclohexanols and Trimethylcyclohexanones

3,3,5-Trimethylcyclohexanol (**2**), $C_6H_{18}O$, M_r 142.23, is a colorless, crystalline material, which is commercially produced as a mixture of the *cis* and *trans* isomers. The individual isomers can be separated by careful distillation or crystallization.

The physical properties of the two isomers are: *cis* isomer [*933-48-2*], *mp* 37.3 °C, n^{60} 1.4390, d^{60}0.860, vapor pressure 1.3 kPa (at 86 °C), flash point 88 °C, solubility in water 0.19 % (at 60 °C), water solubility in alcohol 4.0 % (at 60 °C); *trans* isomer [*767-54-4*], *mp* 57.3 °C, n^{60} 1.4390, d^{60}0.862, vapor pressure 1.3 kPa (at 76 °C), flash point 76 °C, solubility in water 0.17 % (at 60 °C), water solubility in alcohol 2.6 % (at 60 °C).

Trimethylcyclohexanol (**2**) is miscible in all proportions with most organic solvents. It is a cyclic alcohol with similar properties to cyclohexanol. Oxidation of trimethylcyclohexanol with nitric acid gives a mixture of 2,2,4-trimethyladipic acid and 2,4,4-trimethyladipic acid. Trimethylcyclohexanol is prepared by hydrogenating 3,5,5-trimethyl-2-cyclohexen-1-one [*78-59-1*] (**1**), isophorone (a condensation-product of acetone).

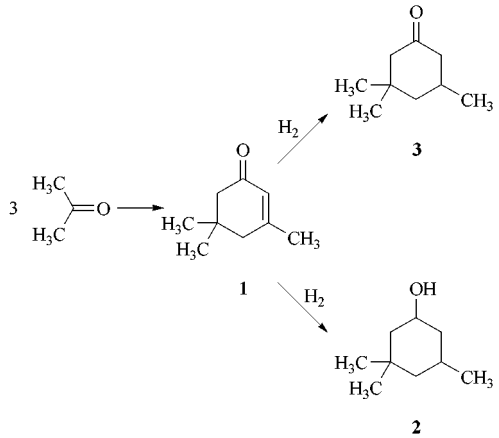

The best catalyst for the hydrogenation is nickel or a mixture of nickel and copper on either silica gel or alumina. The hydrogenation is carried out in high yield at 120 – 200 °C and 1 – 2 MPa. The byproducts of the reaction are small amounts of

Table 3. Physical properties of methylcyclohexanols and methylcyclohexanones

Isomer	CAS registry number	*mp*, °C	*bp*, °C	d^{20}	*n*	Flash point, °C
Methylcyclohexanols						
DL-*cis*-2-Methylcyclohexanol	[*7443-70-1*]	7	165.2	0.934		
DL-*trans*-2-Methylcyclohexanol	[*7443-52-9*]	−4	166.5	0.924		
D-*cis*-3-Methylcyclohexanol	[*5454-79-5*]	−4.7	174.0	0.917		
L-*trans*-3-Methylcyclohexanol	[*7443-55-2*]	−1	175.0	0.915		
cis-4-Methylcyclohexanol	[*7731-28-4*]	−9.2	174.0	0.912		
trans-4-Methylcyclohexanol	[*7731-29-5*]		173.0	0.916		
Commercial mixture	[*25639-42-3*]	−50	173 – 175	0.913		154
Methylcyclohexanones						
DL-2-Methylcyclohexanone	[*583-60-8*]	−13.9	165.1	0.925	1.4440 (25 °C)	48
DL-3-Methylcyclohexanone	[*591-24-2*]	−73.5	170.0	0.920	1.4449 (20 °C)	48
4-Methylcyclohexanone	[*589-92-4*]	−40.6	171.3	0.916	1.4451 (20 °C)	48
Commercial mixture	[*1331-22-2*]		165 – 171	0.918 – 0.925		

trimethylcyclohexanone, trimethylcyclohexane, and trimethylcyclohexene.

The most important use for the alcohol is as a feedstock for trimethyladipic acid [93–97]. This mixture of 2,2,4-trimethyladipic acid [*3586-39-8*]and 2,4,4-trimethyladipic acid [*3937-59-5*], is used in the manufacture of plasticizers and lubricants, and can be converted to dinitriles, diamines, or glycols. Other uses for trimethylcyclohexanol are as replacement for cyclohexanol in polishes and waxes. Because of its lower vapor pressure and unique structure, trimethylcyclohexanol is better than cyclohexanol in varnishes and shellacs.

The trimethylcyclohexanol that is sold commercially is a mixture of the cis and trans isomers. Analysis of a typical commercial product indicates 98 % trimethylcyclohexanols, 1 % trimethylcyclohexanone, 0.4 % trimethylcyclohexane, 0.3 % trimethylcyclohexene, 0.1 % isophorone, and 0.2 % water.

Trimethylcyclohexanol can be stored and transported in iron or steel drums, containers, tank cars, or tank trucks. It is stable indefinitely and does not need to be blanketed with an inert gas.

3,3,5-Trimethylcyclohexanone (**3**) [*873-94-9*], $C_6H_{16}O$, M_r 140.24, is a colorless liquid having properties similar to those of cyclohexanone. It has the following physical properties: *mp* −10 °C, *bp* 188.8 °C, d^{20} 0.888, n^{20} 1.4455, flash point 72 °C, solubility (at 20 °C) of ketone in water 0.3 %, water in ketone 1.4 %. It is miscible in all proportions with most organic solvents.

Trimethylcyclohexanone (**3**) is also prepared by hydrogenating trimethylcyclohexenone (**2**). However, the olefin hydrogenation can be carried out without hydrogenating the ketone by using a palladium-supported catalyst at 110 – 170 °C and 0.2 – 6 MPa [98]. The combination of catalyst, reaction conditions, and amount of hydrogen determines whether the ketone or alcohol is produced.

Trimethylcyclohexanone is a good solvent for low molecular mass poly(vinyl chloride)-type alkyd resins and for nitrocellulose. It can be used in place of methylcyclohexanone, where its higher boiling point results in superior performance in some applications.

9. Economic Aspects

The quoted price for technical-grade cyclohexanol in February 1999 was $ 1.83/kg [99] as compared to $ 1.32/kg in 1986 [100]. U.S. sales of cyclohexanol in 1992 were about 10×10^3 t [87], up from 4.3×10^3 t in 1971 [101]. The remaining Western world sales were believed to be about half of the U.S. sales.

In 1992, U.S. non-adipic acid, non-caprolactam sales of cyclohexanone were about 20×10^3 t, up from 15×10^3 t in 1970 [101]. The early 1999 cost of technical-grade cyclohexanone was $ 1.61/kg [99], whereas the quote

for the same material in November 1986 was $ 1.25/kg [100].

This sales volume for either chemical does not represent capacity, as the vast majority of both are consumed within the manufacturing companies to make adipic acid or caprolactam. In 1992, the worldwide capacity for cyclohexanol-cyclohexanone manufacture was about 3.2×10^6t. Of that total, about 58 % was used for caprolactam manufacture, 40 % for adipic acid manufacture, and the remainder was sold for other uses.

10. Toxicology and Occupational Health

Cyclohexanol and cyclohexanone are moderately toxic [102], but have no apparent carcinogenicity. For cyclohexanol the OSHA TLV for exposure to vapor for 8 h is 50 ppm [103]. The ACGIH has established TLV-TWA values for exposure to cyclohexanol vapor of 50 ppm, 200 mg/m^3 [104]. The TLV-TWA for cyclohexanone is 25 ppm, 100 mg/m^3. The MAKs for both compounds are 50 ppm, 200 mg/m^3.

Excessive exposure to vapor of either chemical results in irritation to the eyes, nose, and throat. Exposure of animals to a high concentration of cyclohexanol vapor resulted in irritation, lacrimation, salivation, lethargy, incoordination, narcosis, mild convulsions, and, in some cases, death. Degenerative changes were noted in the brain, kidney, and liver of rabbits exposed repeatedly to high vapor concentrations of cyclohexanol [102].

Exposure of animals to high concentrations of cyclohexanone vapor also resulted in irritation, lacrimation, and salivation. Decreased heart rate and depression of the CNS were also noted [105]. Degenerative changes were noted in the kidney and liver of monkeys and rabbits exposed repeatedly to high vapor concentrations of cyclohexanone.

For single dosages of cyclohexanol, the oral LD_{50} for rats is 2.06 g/kg [106] and the minimum lethal dose for rabbits is 2.4 g/kg [92]. With cyclohexanone, the oral LD_{50} for rats is 1.62 g/kg [107] and the minimum lethal dose for rabbits is 1.6 – 1.9 g/kg [92].

Cyclohexanol and cyclohexanone are moderately irritating to the eye and can cause corneal injury [107], [108]. Cyclohexanol can be absorbed through the skin in toxic amounts, and extensive exposure can cause tremors, narcosis, hypothermia, and death [92]. Cyclohexanone is only moderately irritating to the skin. Defatting and dermatitis of the skin occurred when animals were exposed for long periods of time [109].

Derivatives. Commercial *methylcyclohexanol* is low in toxicity, and the ACGIH has set TLVs of 75 ppm, 350 mg/m^3 (STEL) and 50 ppm, 235 mg/m^3 (8-h TWA). The MAK value is the same as the TLV/TWA value [91]. Excessive exposure causes headaches and eye and nose irritation.

Methylcyclohexanones are moderately toxic to laboratory animals and cause eye, nose, and throat irritation; they have produced corneal damage to the eyes of rabbits. The single-dose oral LD_{50} for rats is ca. 2 g/kg [92]. 2-Methylcyclohexanone may be absorbed through the skin. Its TLV/TWA and MAK are 50 ppm, 230 mg/m^3; the TLV/STEL is 75 ppm, 345 mg/m$^{3\cdot}$.

Trimethylcyclohexanol has the same toxicological properties as methylcyclohexanol, but because of its lower vapor pressure, it presents fewer hazards. Neither TLV nor MAK values have been adopted.

The toxicity of *trimethylcyclohexanone* is similar to that of methylcyclohexanone, but its lower vapor pressure makes it less dangerous to handle.

11. References

1. G. Schwarzenbach, E. Felder, *Helv. Chim. Acta* **27** (1944) 1044. G. Schwarzenbach, C. Witwer, *Helv. Chim. Acta* **30** (1947) 659, 669.
2. BASF, DE 352 439, 1913.
3. Tetralin Ges., DE 299 012, 1916.
4. *Chem. Tech. (Leipzig)* **18** (1966) no. 10, 608 – 613.
5. Stamicarbon, US 3 305 586, 1967 (P. Bernard).
6. Stamicarbon, GB 890 095, 1960.
7. G. D. Lyubarskii, G. K. Kervalishvili, *Khim. Promst. (Moscow)* **7** (1972) 491.
8. Allied Chemical Corp., US 4 092 360, 1978 (J. F. van Peppen, W. B. Fisher).
9. Allied Chemical Corporation, US 5 015 787, 1991 (J. F. van Peppen).

10. Du Pont, US 3 530 185, 1970 (K. Pugi).
11. Du Pont, US 3 957 876, 1976 (M. Rapoport, J. O. White).
12. Union Carbide, US 3 404 185, 1968 (T. Adams).
13. Celanese, US 3 598 869, 1971 (P. J. Volpe, W. J. Humphrey).
14. Du Pont, US 3 987 100, 1976 (W. J. Barnett, D. L. Schmitt, J. O. White).
15. Rhône-Poulenc, US 3 510 526, 1970 (J. P. M. Bonnart, Y. Bonnet, P. P. M. Rey).
16. Arco, US 3 949 004, 1976 (H. A. Sorgenti, S. N. Rudnick).
17. Rhône-Poulenc, BE 731 125, 1969.
18. Arco, GB 1 335 296, 1973 (H. A. Sorgenti, S. N. Rudnick).
19. Du Pont, US 4 675 450, 1987 (J. B. Lyon, G. T. Stowe).
20. Monsanto, US 3 260 742, 1966.
21. Du Pont, US 2 851 496, 1958 (H. L. Cates, R. W. Wheatcroft, A. B. Stiles, J. O. Punderson).
22. Rhône-Poulenc, US 3 925 316, 1975 (J-C. Brunie, N. Crenne, F. Maurel).
23. Rhône-Poulenc, US 3 923 895, 1975 (M. Constantini, N. Crenne, M. Jouffret, J. Nouvel).
24. Rhône-Poulenc, US 3 928 452, 1975 (J-C. Nouvel, M. Costantini, N. Crenne, M. Jouffret).
25. Rhône-Poulenc, US 3 927 105, 1975 (J-C. Brunie, N. Crenne).
26. Stamicarbon (DSM), US 3 941 845, 1976 (W. Vvoskuil, J. J. M. van der Donck).
27. Stamicarbon (DSM), US 3 987 101, 1976 (J. Wolters, J. L. J. P. Hennekens).
28. Stamicarbon (DSM), US 4 042 630, 1977 (J. Wolters, J. L. J. P. Hennekens).
29. Du Pont, GB 777 087, 1957.
30. Rhône-Poulenc, BE 715 662, 1968.
31. Rhône-Poulenc, GB 1 229 734, 1971 (J-C. Brunie, N. Crenne).
32. Rhône-Poulenc, US 3 923 895, 1975 (M. Constantini, N. Crenne, M. Jouffret, J. Nouvel).
33. Du Pont, US 4 326 084, 1981 (J. D. Druliner, P. J. Krusic, S. D. Ittel, C. Tolman).
34. Du Pont, US 4 503 257, 1985 (J. D. Druliner, J. Hermolin).
35. DSM, US 5 298 665, 1992 (H. A. C. Baur, Hoogendoorn, Janssen, Ragten).
36. Phillips, US 2 931 834, 1960 (W. W. Crouch, J. C. Hillyer).
37. BASF, GB 1 382 849, 1975 (O. A. Grosskinsky, G. Herrmann, R. Kaiser, A. Kuessner).
38. Stamicarbon (DSM), US 4 238 415, 1980 (W. O. Bryan).
39. Stamicarbon (DSM), EP 0 004 105, 1979 (W. O. Bryan).
40. DSM, CA 1 244 475, 1983 (Housmans, Vannispen, Plantema).
41. DSM, EP 659 776, 1995 (H. A. C. Baur, U. F. Kragten).
42. Rhône-Poulenc, US 3 557 215, 1971 (J. P. M. Bonnart, Y. Bonnet, P. P. M. Rey).
43. Rhône-Poulenc, US 3 694 511, 1972 (J. Nouvel).
44. Stamicarbon (DSM), US 3 927 108, 1975 (C. G. M. van der Moesdijk, A. M. J. Thomas).
45. Stamicarbon (DSM), US 3 937 735, 1976 (P. L. M. Dois).
46. Rhône-Poulenc, GB 1 221 629, 1971 (J. Nouvel).
47. DuPont, US 4 720 592, 1988 (U. N. Besmar, J. B. Lyon, F. J. Miller, M. T. Musser).
48. Halcon, US 3 317 614, 1967 (W. C. Long).
49. Halcon, US 3 665 028, 1972 (J. L. Russell).
50. Institut Français du Pétrole, FR 1 556 979, 1969 (J. Alagy, A. Zuech, B. Cha).
51. Institut Français du Pétrole, FR 1 573 834, 1969 (J. Alagy, L. Asselineau, C. Busson, B. Cha).
52. Du Pont, BE 782 721, 1972 (W. C. Drinkard).
53. BASF, DE 2 124 590, 1972 (R. Platz, W. Fuchs, C. Dudeck).
54. Ube Industries, JP 7 809 746, 1978 (T. Kawahito, M. Tamure).
55. Asahi, JP 60 255 738, 1985 (H. Nagahara, Y. Fukuoka).
56. Asahi, JP 90 016 736B, 1990.
57. Bayer, DE 4 414 328, 1995 (H. Fiege, H. Haerle, A. Klausener).
58. Sumitomo, US 4 665 274, 1987.
59. Asahi, JP 61 221 134, 1985 (Y. Fukuoka, O. Mitsui).
60. DSM, US 4 927 974, 1989 (van den Brink, van Geem).
61. Sumitomo, JP 62 126 139, 1987.
62. Sumitomo, JP 62 123 136, 1987.
63. Asahi, EP 162 475, 1985 (M. Tojo, O. Mitsui).
64. Asahi, JP 63 154 636, 1988 (O. Mitsui, H. Goto).
65. Du Pont, US 5 321 155, 1994 (W. C. Drinkard, G. C. Grunewald).
66. Sumitomo, US 4 691 064, 1987.
67. Sumitomo, US 4 716 253, 1987 (T. Shirafuji, K. Sakai, K. Hirose).
68. Sumitomo, US 4 670 612, 1986.
69. F. Andreas, *Chem. Tech. (Leipzig)* **18** (1966) no. 10, 608.

70. J. W. Bruce, *Ind. Chem.* 1963, no. 3, 121.
71. Scientific Design, BE 598 094, 1960.
72. Monsanto, US 2 970 172, 1959 (D. R. Cova).
73. Wingfoot Corp., US 2 218 457, 1937 (C. F. Winans).
74. Du Pont, US 2 163 284, 1937 (W. A. Lazier).
75. Monsanto, US 2 970 172, 1959 (D. R. Cova).
76. L. K. Filippenko, K. N. Belonogov, V. P. Gostikin, *Izv. Vyssh. Uchebn. Zaved. Khim. Tekhnol.* **13** (1970) 3.
77. F. Laucht, US 2 338 445, 1939.
78. I. G. Farben, DE 743 004, 1940 (F. Laucht, O. Klopfer).
79. K. Smeykal, H.-J. Naumann, *Chem. Tech. (Leipzig)* **13** (1961) no. 3, 132.
80. Allied Signal Corporation, US 3 974 221, 1976 (J. Duggan).
81. BASF, US 3 149 166, 1964 (G. Poehler, A. Wegerich, H. Giehne, O. Goerhre).
82. S. D. Nogare, J. Mitchell, Jr., *Anal. Chem.* **25** (1953) 1376.
83. N. A. Cheronics, T. S. Ma: *Organic Functional Group Analysis by Micro and Semimicro Methods,* Interscience, New York 1964, pp. 493 – 496.
84. J. Mitchell, Jr., D. M. Smith, W. M. D. Bryant, *J. Am. Chem. Soc.* **63** (1941) 573.
85. W. M. D. Bryant, D. M. Smith, *J. Am. Chem. Soc.* **57** (1935) 57.
86. United States Government, "Code of Federal Regulations", 49CFR172.101, Oct. 1, 1998.
87. *Kirk-Othmer,* 4th ed., **49** (1993) 885.
88. Toray, GB 1 402 480, 1975.
89. A. K. Doolittle: *The Technology of Solvents and Plasticizers,* J. Wiley & Sons, New York 1954.
90. I. Mellon: *Industrial Solvents,* Reinhold Publ., New York, 1950, pp. 521 – 522.
91. DFG (ed.): *Maximum Concentrations at the Workplace (MAK),* VCH Verlagsgesellschaft, Weinheim 1986.
92. J. F. Treon, W. E. Crutchfield, Jr., K. V. Kitzmiller, *J. Ind. Hyg. Toxicol.* **25** (1943) 199.
93. VEBA, DE 1 111 163, 1959.
94. VEBA, DE 1 229 510, 1963 (K. Schmitt, H. Heumann, W. Pollack).
95. VEBA, DE 1 418 067, 1959.
96. VEBA, DE 1 418 074, 1959 (E. Rindtorff, K. Schmitt, H. Heumann).
97. VEBA, DE 1 468 693, 1963 (K. Schmitt, H. Neumann).
98. VEBA, US 3 361 822, 1968 (K. Schmitt, J. Disteldorf, H. Schnurbusch, W. Hilt).
99. *Chemical Market Reporter*, 1999 (Feb.15), Schnell Publishing Co., New York.
100. *Chemical Market Reporter*, 1986 (Nov.10), Schnell Publishing Co., New York.
101. C. S. Hughes (compiler): *Chemical Economics Handbook*, SRI International, Menlo Park, CA., Apr. 1983, 638.7020-A.
102. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., Wiley-Interscience, New York 1982, pp. 4643 – 4649.
103. U.S. Government, CFR 29, Washington, D.C., Jul. 1, 1978, Section 1910.1000.
104. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* ACGIH, Cincinnati, Ohio, 1986/1987.
105. H. Specht, J. W. Miller, P. J. Valaer, R. R. Sayers: "Acute Response of Guinea Pigs to the Inhalation of Ketone Vapors," U.S. Public Health Service, NIH Bulletin, no. 176, Division of Industrial Hygiene, 1940, pp. 1 – 66.
106. F. Bar, F. Griepentrog, *Med. Ernaehr.* **8** (1967) 244.
107. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, U. C. Pozzani et al.: "Range Finding Toxicity Data VII," *Am. Ind. Hyg. Assoc. J.* **30** (1969) 470 – 476.
108. J. Pohl, *Zentralbl. Gewerbehyg. Unfallverhuet.* **12** (1925) 91.
109. P. K. Gupta, W. H. Lawrence, J. E. Turner, J. Autian: "Toxicological Aspects of Cyclohexanone," *Toxicol. Appl. Pharmacol.* **49** (1979) 525 – 533.

Cyclooctadiene → Cyclododecatriene and Cyclooctadiene

Cyclopentadiene and Cyclopentene

DIETER HÖNICKE, Technische Universität Chemnitz, Chemnitz, Federal Republic of Germany (Chaps. 1 and 2)

RINGO FÖDISCH, Technische Universität Chemnitz, Chemnitz, Federal Republic of Germany (Chaps. 1 and 2)

PETER CLAUS, Institut für Angewandte Chemie Adlershof, Berlin, Federal Republic of Germany (Chaps. 1 and 2)

MICHAEL OLSON, General Motors Research Laboratories, Warren, Michigan 48090, United States (Chap. 3)

1. Cyclopentadiene and Dicyclopentadiene 1
1.1. Physical Properties 1
1.2. Chemical Properties 2
1.3. Production 5
1.4. Uses 5
2. Cyclopentene 7
2.1. Physical Properties 7
2.2. Chemical Properties 7
2.3. Production 8
2.4. Uses 9
3. Toxicology 10
4. References 11

1. Cyclopentadiene and Dicyclopentadiene

The physical and chemical properties of cyclopentadiene and dicyclopentadiene have been reviewed elsewhere [1, pp. 417 – 429.], [2], [3], [4], [5].

1.1. Physical Properties

Cyclopentadiene [*542-92-7*], C_5H_6, M_r 66.10, is a colorless liquid having a sweet terpenic odor. It is also known as pentole and pyropentylene. It is soluble in most organic solvents, including ethanol, diethyl ether, acetone, petroleum ether, benzene, and toluene, but is relatively insoluble in water.

The boiling points and compositions of three hydrocarbon azeotropes with cyclopentadiene are listed below:

Second component	*bp* (0.1 MPa), °C	Cyclopentadiene, mole fraction
n-Pentane	35.3	0.31
2-Methyl-2-butene	38.0	0.37
cis-2-Pentene	36.9	0.03

Dicyclopentadiene [*77-73-6*], $C_{10}H_{12}$, M_r 132.20, colorless crystals, has a camphoraceous odor and is soluble in ethanol, diethyl ether, and acetic acid.

Table 1. Physical properties of cyclopentadiene and dicyclopentadiene

Physical property	Cyclopentadiene	Dicyclopentadiene
Boiling point at 101.3 kPa, °C	41.5	169.8
Freezing point, °C	− 85	33.9
Density, g/cm^3	0.8021 (20 °C)	0.9770 (35 °C)
Refractive index n_D	1.440 (20 °C)	1.5050 (35 °C)
Dielectric constant at 40 °C	2.43	
Dipole moment, vapor, D	0.416	
Surface tension at 40 °C, N/m	3.15×10^{-2}	
Liquid molar volume, m^3/kmol	0.082925	0.138585
Solubility parameter, (J/m^3)$^{1/2}$	16.961×10^3	17.445×10^3
van der Waals volume, m^3/kmol	0.04368	0.08146
van der Waals area, m^2/kmol	5.67×10^8	9.3×10^8
Critical pressure, MPa	5.15	3.06
Critical temperature, °C	233.8	386.8
Critical volume, cm^3/mol	225	445
Critical compression factor	0.275	0.248

The physical properties of cyclopentadiene and dicyclopentadiene are listed in Table 1; thermal and thermodynamic properties are summarized in Table 2. The vapor pressure of cyclopentadiene (MPa) over the range 220 – 320 K is represented by

$$\log p = 3.8892 - 1530.49/T$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_227

The vapor pressure of dicyclopentadiene (MPa) over the range 305 – 450 K is represented by

$$\log p = 3.6172 - 2056.49/T$$

Table 2. Thermal and thermodynamic properties of cyclopentadiene and dicyclopentadiene

Property	Cyclo pentadiene	Dicyclo-pentadiene
Autoignition temperature. °C	640	510
Flash point, °C	< − 46.15	32.2
Flammability limits, vol %	1.7 – 14.6	1.0 – 8.3
Heat of combustion, kJ/kg	53 649	41 870
Heat of vaporization, kJ/kg	415.7 (41.5 °C)	325.7 (169.8 °C)
Heat of fusion at 32.5 °C, kJ/kg		95 385
Heat of dimerization at 25 °C, kJ/kg	584 ± 20	
Specific heat capacity , kJ $kg^{-1}K^{-1}$		
liquid	1.780	1.311
gas, 0.1 MPa	1.074 (25 °C)	0.935 (33 °C)
Vapor pressure at 20 °C, kPa	45.1	0.04
Absolute entropy,∗ J mol^{-1} K^{-1}	280.3	240
Enthalpy of formation,∗ kJ/mol	130.8	196.1
Gibbs free energy of formation,∗ kJ/mol	172.6	375.0

∗ Ideal gas at 25 °C

Additional properties are available [4–6]. The IR [7], ^{1}H NMR [8], and mass spectra [9] of cyclopentadiene and the ^{1}H NMR spectrum [10] of dicyclopentadiene have been reported.

1.2. Chemical Properties

Cyclopentadiene (**1**) is an extremely reactive compound. Neat, it can only be preserved at − 80 °C or below, and then only for a limited amount of time. At or above room temperature, it dimerizes spontaneously to give dicyclopentadiene (**2**).

1 **2**

The dimerization rate r_D is a function of temperature [1, p. 423]:

T, °C	−20	0	10	15	20	25	30	40
r_D, mol %/h	0.05	0.5	1	1.5	2.5	3.5	6	15

Above 160 °C, dicyclopentadiene dissociates into cyclopentadiene again. Therefore, cyclopentadiene is stored and handled as its stable dimer. The dimerization of cyclopentadiene is exothermic and may get out of control, leading to polymerization and even coking in sealed vessels that lack effective heat removal.

Reactions of cyclopentadiene with aggressive reagents, such as concentrated sulfuric acid, fuming nitric acid, or ozone, can lead to decomposition or even explosion, depending on the conditions, particularly the reagent/substrate ratio. If proper care is taken, however, selective reaction can be achieved with most reagents. Major reaction types are cycloadditions and additions to the double bonds, oligo- and polymerizations, substitutions at the reactive methylene group, and oxidations.

A variety of *cycloaddition reactions* have been realized (Fig. 1). The formation of **3** [11] and **4** [12] are examples of [2 + 2] and [2 + 3] cycloadditions, respectively. By far the most important and most versatile types are catalytic and noncatalytic [4 + 2] cycloadditions, the Diels – Alder reactions. They afford the ring systems **5**, **6**, **7**, **8**, and **9** by reaction of cyclopentadiene with monoolefins, olefinic esters [13], acetylenes [14], [15], conjugated diolefins [16], [17], [18], cumulated diolefins [19], aldehydes [20], and activated ketones [21]. The cycloaddition of glyoxylic acid with cyclopentadiene in aqueous solution leads to α-hydroxy-γ-lactones after rearrangement of the Diels – Alder adduct [22]. Diels – Alder reactions with monoolefins and acetylenes are particularly facile with substrates bearing one (R^1) or two (R^1 and R^2) electron-withdrawing groups, such as carbonyl-, cyano-, or halomethane moieties. With chiral boron catalysts, highly effective enantioselective Diels – Alder reactions of cyclopentadiene with acetylenic aldehydes can be carried out to give products of type **6** [23]. Using chiral monoalcohols such as isosorbide as starting materials to provide acrylate derivatives for Diels – Alder reactions with cyclopentadiene provides exclusively endo adducts with high diastereoselectivity [24]. Diels – Alder additions of cyclopentadiene with cyclic substrates afford tricyclic ring systems that exhibit predominantly endo geometry, as in the case of dicyclopentadiene (**12**) and the Diels – Alder adduct **13** of cyclopentadiene with maleic anhydride.

12

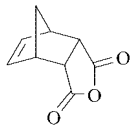

13

Compounds **10** [25] and **11** [26] are examples of [4 + 3] and [4 + 4] cycloaddition products of cyclopentadiene, respectively.

Addition reactions have been realized with a variety of reagents. Examples are the selective hydrogenation to give cyclopentene, hydrohalogenations to give 3-halocyclopentenes, catalyzed hydrocyanation to give cyanocyclopentene and dicyanocyclopentane [27], and halogenations to give 3,5-dihalocyclopentenes and 1,2,3,4-tetrahalocyclopentanes. Tetrachlorocyclopentane can be readily converted into hexachlorocyclopentadiene: catalytic chlorination of the tetrachlorocyclopentane gives octachlorocyclopentane, which is dehydrochlorinated.

Oligo- and polymerizations can be carried out thermally or catalytically. Thermal reactions have been reported to occur by a series of consecutive cycloadditions of cyclopentadiene with Diels – Alder adducts like **14** to give products of structure **15** [28], [29]. A special case is the thermally induced homopolymerization of cyclopentadiene via dicyclopentadiene (**12**) to give crystalline oligomers or amorphous polymers, depending on the reaction temperature and time. Polymerization of cyclopentadiene with cationic catalysts gives cross-linked elastomeric materials [30]; with Ziegler-type catalysts, these polymerizations give elastomers. In addition, numerous copolymers of cyclopentadiene, e.g., with aromatics and/or olefins, drying oils, and indene – coumarone resins, have been reported.

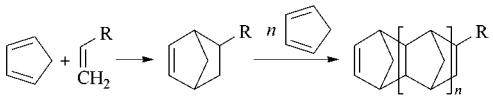

14 **15**

Substitution reactions at the reactive methylene group occur predominantly by initial deprotonation, giving the resonance-stabilized cyclopentadienyl anion, which exhibits quasiaromatic character. Products derived from it are organometallics, such as alkali-metal and Grignard compounds and especially transition-metal complexes. A variety of the latter have been synthesized following the discovery of ferrocene. Other notable substitution reactions at the methylene group are mono- and dialkylations with alkyl halides, base-catalyzed carboxylation with carbon dioxide, and base-catalyzed condensations with aldehydes and ketones to give fulvenes. The last has been used to prepare a wide variety of fulvenes by starting from acyclic, carbocyclic, and heterocyclic [31] carbonyl compounds in efficient, high-yield syntheses [32].

Oxidation reactions with molecular oxygen occur spontaneously and provide intractable peroxidic materials. Controlled photochemical oxidation with oxygen gives the cyclic peroxide **16** [33], which can be converted into **17**, **18**, or **19** by reduction. Reaction with oxygen in the gas phase catalyzed heterogeneously by various mixed metal oxides has been reported to produce maleic anhydride [34], [35] and in one case a mixture of maleic and phthalic anhydride [36]. Heterogeneously catalyzed ammoxidation of cyclopentadiene affords pyridine [37].

16

17

18

19

1.3. Production

Cyclopentadiene is available from coal tar as dicyclopentadiene in quantities of ca. 10 – 20 g per ton of coal and can be readily isolated [38]. Considerably larger quantities are available from steam cracking operations, e.g., of naphtha, which affords ca. 14 kg of dicyclopentadiene per ton of feed material [38–40]. In this case, the C_5 fraction of pyrolysis gasoline contains

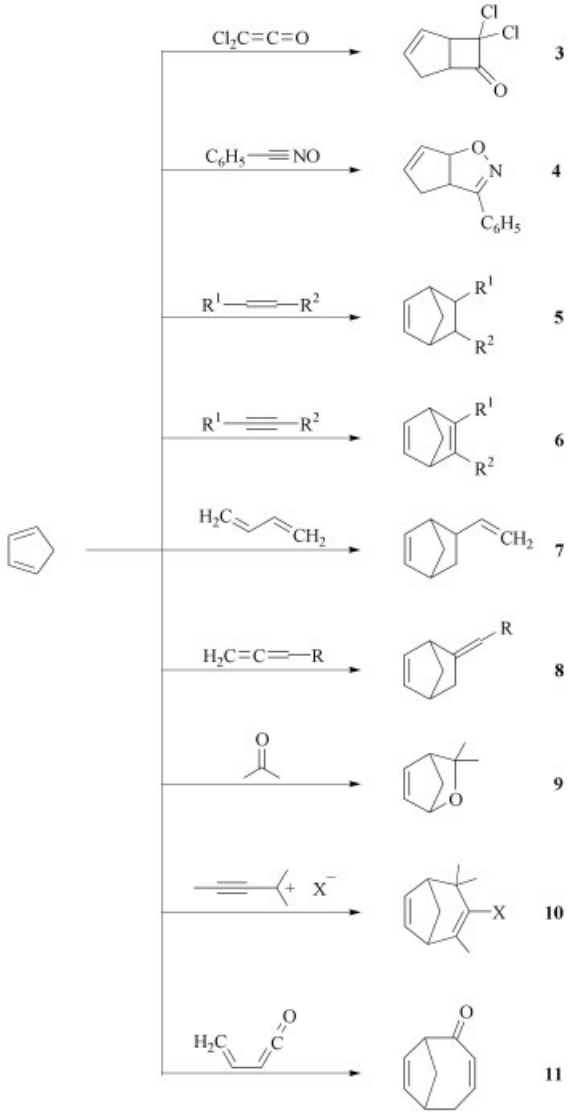

Figure 1. Cycloaddition reactions of cyclopentadiene

ca. 25 wt % cyclopentadiene with approximately equal proportions of isoprene and of *n*-pentane, along with other C_5 hydrocarbons.

From the C_5 fraction, cyclopentadiene can be isolated by the sequence of steps shown in Fig. 2.

In the first step, cyclopentadiene is dimerized to give dicyclopentadiene by heat soaking the entire C_5 fraction, either at normal pressure and 30 – 100 °C over 5 – 24 h or at elevated pressure and 140 – 150 °C [38], [41]. In the second step, the remaining components of the original C_5 fraction, which have boiling points of 28 – 50 °C, are distilled overhead, and crude dicyclopentadiene (*bp* 172.8 °C) is obtained in 85 – 90 % purity [38] at the bottom of the column. In the third step, the crude dicyclopentadiene is monomerized, either in the liquid phase at 170 – 200 °C [39] or in the gas phase at 300 – 400 °C in the presence of diluents such as steam, hydrogen, nitrogen, or methane [1, p. 424], to give cyclopentadiene in ca. 95 % purity [38].

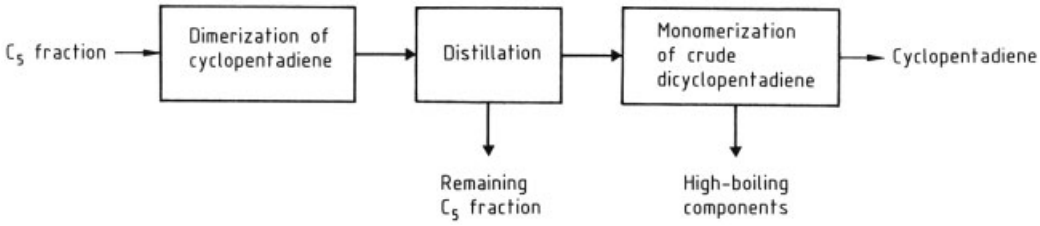

Figure 2. Block diagram for the separation of cyclopentadiene from a C_5 fraction

Cyclopentadiene can be further purified by sequential dimerizations under carefully controlled slow heat soaking [38] and subsequent monomerizations. This procedure largely avoids the codimerization of cyclopentadiene with residual unsaturated C_5 impurities, allowing the latter to be removed from dicyclopentadiene by stripping.

Alternative modes of formation of cyclopentadiene are heterogeneously catalyzed monomerizations of dicyclopentadiene [42], [43]; catalyzed dehydrocyclization of piperylene [44], [45]; catalyzed dehydrogenations of cyclopentene [46], [47] and cyclopentane [48]; catalyzed hydrodenitrification of pyridine, quinoline [49], and other aromatic amines [50]; the pyrolysis of phenol [51], [52]; and the pyrolysis of plastics and rubbers [53]. Another possibility for the isolation of cyclopentadiene from the C_5 fraction of pyrolysis gasoline is the selective formation of a Diels – Alder adduct by introducing the C_5 fraction into aqueous maleic acid. The pure Diels – Alder adduct precipitates, whereas the remaining dienes of the C_5 fraction do not react [54–56].

1.4. Uses

Cyclopentadiene is the single major precursor for the production of cyclopentene by selective hydrogenation. It is used as a prepolymer to modify the properties of Ziegler – Natta catalysts [57]. Ethylidenenorbornene (**8**; R = CH_3), which can be produced by Diels – Alder addition of cyclopentadiene with 1,2-butadiene or with 1,3-butadiene followed by isomerization of the ensuing adduct **7**, serves as a termonomer in the production of ethylene – propylene terpolymer elastomers. Polymerization of **8** occurs through the more reactive endocyclic double bond, whereas the exocyclic double bond imparts the active site for vulcanization. The facile reversibility of the Diels – Alder reaction has been used for the transient protection of α, β-unsaturated carbonyl compounds from polymerization while the compounds were chemically modified at or near the carbonyl functions [58]. By this method, isoprenic alcohols such as linalool and nerolidol have been synthesized in high yields starting from methyl vinyl ketone [59]. Substituted pentacycles such as **21**, which are readily obtained from Diels – Alder adducts **20** of cyclopentadiene with substituted paraquinones, have been shown to undergo facile acid-catalyzed ring opening to the tricyclic system **22**, which in turn is completely reconverted to **21** in the presence of light. Thus, the system bears prospects for light – energy conversion by storing energy in the strained pentacyclic system **21** [60]. Diels – Alder adducts of fulvenes with activated olefins have been converted into bicyclic ketones such as **23**, which are useful in perfume formulations. Complexes of the cyclopentadienyl anion with transition metals (metallocenes) such Ti, Zr, and Hf act as polymerization catalysts [61]. Several complexes with Ti, V, Nb, or Mo as central atom [62] and the Diels – Alder adducts of cyclopentadiene with anthraquinone derivatives [63] show cancerostatic activity.

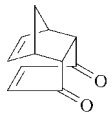

20

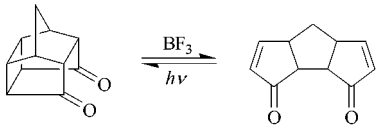

21 **22**

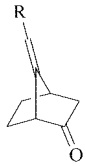

23

Homopolymers of cyclopentadiene have been reported to be useful as modifiers for high molecular mass polymers as base materials for coatings, adhesives, paper-sizing agents, and printing inks. The Diels – Alder adduct **13** of cyclopentadiene and maleic anhydride is useful for the production of modified polyester resins [38].

Hexachlorocyclopentadiene, which was a substantial commercial product derived from cyclopentadiene, has decreased in importance considerably. Major products derived from it were the chlorinated pesticides aldrin, dieldrin, chlordane, and heptachlor, but their application has been drastically restricted due to their toxicity. Hexachlorocyclopentadiene has also been used for the preparation of flame- and fire-retardant chemicals. In particular, the Diels – Alder adduct with maleic anhydride has been used as a comonomer in the synthesis of fire-resistant polyesters. Other cyclopentadiene-containing resins are used, e.g., as serum and plasma separating layers for blood-testing containers [64] and for the production of synthetic lubricants [65].

Dicyclopentadiene has attracted attention as a building block for the production of modified hydrocarbon resins, which show increased reactivity in copolymerizations with drying oils and produce paint resins with improved drying rate, gloss, and hardness. Modifications of unsaturated polyester resins with dicyclopentadiene have also been reported [38]. Hydrogenation of dicyclopentadiene and subsequent acid-catalyzed isomerization of the ensuing saturated product **24** has opened a facile synthetic route to adamantane **25**.

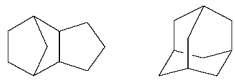

24 **25**

2. Cyclopentene

The physical and chemical properties of cyclopentene have been reviewed elsewhere [66, pp. 593 – 605], [3].

Table 3. Physical properties of cyclopentene

Boiling point at 101.3 kPa, °C	44.2
Freezing point, °C	−135.1
Density at 20 °C, g/cm^3	0.7720
Dynamic viscosity at 13.5 °C, Pa s	4.29
Refractive index n_D^{20}	1.4194
Effect of temperature on n_D over the range 10 – 25 °C, K^{-1}	−0.0006
Dielectric constant at 20 °C	2.095
Dipole moment in CCl_4, D	0.93
Surface tension at 13.5 °C, N/m	2.356×10^{-2}
Liquid molar volume, m^3/kmol	0.088841
Solubility parameter, (J/m^3)$^{1/2}$	17.156×10^3
van der Waals volume, m^3/kmol	0.04681
van der Waals area, m^2/kmol	0.605×10^9
Critical pressure, MPa	4.79
Critical temperature, °C	233.8
Critical volume, cm^3/mol	240
Critical compression factor	0.273

2.1. Physical Properties

Cyclopentene [*142-29-0*], C_5H_8, M_r 68.12, is a colorless, pungent liquid. It is soluble in most organic solvents, including ethanol, diethyl ether, petroleum ether, and benzene. However, at 23 °C only 160 ppm of cyclopentene is soluble in water and 400 ppm of water is soluble in cyclopentene. Two typical organic azeotropes with cyclopentene are listed below:

Second component	*bp* (0.1 MPa), °C	Cyclopentene, mole fraction
Methanol	37	ca. 0.68
Methyl acetate	41.7	0.74

The physical properties of cyclopentene are listed in Table 3; thermal and thermodynamic properties are summarized in Table 4. The vapor pressure of cyclopentene (MPa) over the range 215 – 350 K is represented by

$$\log p = 3.6594 - 1482.13/T$$

Table 4. Thermal and thermodynamic properties of cyclopentene

Autoignition temperature, °C	395
Flash point, °C	− 28.9
Explosive limits in air at 0.1 MPa and 20 °C,	
vol %	3.4 – 8.5
g/m^3	103 – 258
Heat of combustion	
at 0.1 MPa and 30 °C, kJ/kg	45 756
Heat of vaporization at 44.2 °C, kJ/kg	394.77
Heat of fusion at − 135.1 °C, kJ/kg	49.37
Specific heat capacity, kJ kg^{-1}K^{-1}	
liquid	1.799
gas (25 °C, 0.1 MPa)	1.103
Vapor pressure at 25 °C, kPa	48.9
Absolute entropy,∗ J mol^{-1} K^{-1}	289.4
Enthalpy of formation,∗ kJ/mol	32.93
Gibbs free energy of formation,∗ kJ/mol	110.8

∗ Ideal gas at 25 °C

Additional properties can be found in [6]. The IR [7], ^{1}H NMR [8], and mass spectra [9] of cyclopentene have been reported.

2.2. Chemical Properties

Cyclopentene undergoes the usual reactions of monoolefins, i.e., cycloadditions and additions to the double bond, substitutions in the allylic positions, and oxidative as well as metathetic cleavage of the ring system.

Cycloaddition reactions occur with allene [67] and butynone [68] in a [2 + 2] fashion and with conjugated dienes such as cyclopentadiene and 1,3-cyclohexadiene. Photochemically induced cycloadditions with aromatic substrates such as benzene [69] or toluene [70] occur by *meta* addition. Palladium-complex-catalyzed reaction of cyclopentene with dimethyl acetylenedicarboxylate results in a cocyclization reaction [71].

Addition reactions are numerous, such as hydrogenation to give cyclopentane, halogenations to give *trans*-1,2-dihalocyclopentanes, formylation to give formylcyclopentane, and hydrocarboxylation (Koch reaction) to give cyclopentanecarboxylic acid. Acid-catalyzed reactions with substituted benzenes lead to alkylation products, whereas the joint reaction of cyclopentene, benzene, and carbon monoxide leads to simultaneous alkylation and formylation to give 4-cyclopentylbenzaldehyde [72]. Epoxidation of cyclopentene has been carried out with peracids, hydrogen peroxide [73], [74], and oxygen by cooxidation with acetaldehyde [75]. Direct hydroxylation of cyclopentene to give 1,2-dihydroxycyclopentane has been achieved with modified permanganate [76] and with hydrogen peroxide – formic acid [77]. Cyclopentanone is obtained either by Pd-catalyzed oxidation with oxygen in a microemulsion [65], catalytically in the gas phase [78], or by catalytic hydration of cyclopentene and subsequent oxidative dehydrogenation of the intermediate cyclopentanol. Cyclopentanol can also be obtained by hydroboration of cyclopentene and subsequent oxidation [79]. Cyclopentene can be dimerized by homogeneous [80] and oligomerized by heterogeneous catalysis [81].

Substitution reactions occur predominantly in the allylic positions. Reactions with molecular oxygen lead to **26** as the primary product, both under autoxidation [82] and photooxidation [83]. Secondary reaction products are epoxides and alcohols [84]. Catalyzed photooxidations of cyclopentene lead to **27** [85] or **28** [86], depending on the catalyst used. Catalyzed cooxidation of cyclopentene and acetic acid affords **29** [87].

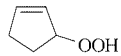

26

27

28

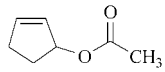

29

Ring cleavage and ring enlargement reactions occur under drastic oxidation or by metathesis. Ozonolysis in methanol can be guided by selective acid catalysis such that **30a** [88], **30b**, or **30c** [89] is formed. Catalyzed oxidations of cyclopentene with hydrogen peroxide and noncatalyzed oxidation with oxygen yield **30d** [90], and oxidation with nitric acid at elevated temperature and pressure gives **30e**. Oxidation of cyclopentene with oxygen on heterogeneous V_2O_5-based catalysts yields maleic anhydride [91], and with special catalysts, mixtures of maleic and phthalic anhydride result [92]. Heterogeneously catalyzed ammonoxidation yields pyridine and acyclic nitriles [93].

Metathesis of cyclopentene catalyzed homogeneously produces a polymer of structure **31**, called polypentenamer [40].

$R–(CH)_3–R'$
30

	a	b	c
R	CH_3OOC	CH_3OOC	O=CH
R′	CH_3OOC	$(CH_3O)_2CH$	$(CH_3O)_2CH$

	d	e
R	O=CH	HOOC
R′	O=CH	HOOC

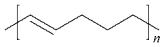

31

2.3. Production

Cyclopentene is formed in proportions of ca. 2.2 kg per ton of feed in the steam cracking of naphtha and is present in ca. 4 wt % in the C_5 fraction of pyrolysis gasoline [38], [39]. Such a low proportion is insufficient for industrial production, and hence, cyclopentene is effectively produced solely by selective hydrogenation of cyclopentadiene, generally by heterogeneous catalysis of gas-phase reactions with a wide variety of supported metal catalysts [94–99] or by heterogeneous [100–103] or homogeneous [104–106] catalysis of liquid-phase reactions. Feed materials are either crude dicyclopentadiene or the entire C_5 fraction.

Figure 3 shows the flow scheme of the Bayer process for the production of cyclopentene from crude dicyclopentadiene derived either from coal tar or steam-cracking operations (Section 1.3).

Crude dicyclopentadiene is monomerized in cracking column (a) at 170 – 190 °C, whereby codimers function as a diluent for the feed and are continuously removed from the bottom of column (a). Cyclopentadiene leaves the top of column (a), is mixed with hydrogen, and hydrogenated on a palladium catalyst [107] at 70 – 80 °C under moderate pressure in reactor (b). The effluent from (b) passes into (c), where it is separated into residual gas and liquid products. The liquid fraction passes flash column (d) to remove the major part of high-boiling components and is subsequently fractionated in column (e) to give cyclopentene. The yield of cyclopentene is 85 – 90 %, based on monomeric cyclopentadiene; the selectivity is 90 – 95 %.

Figure 4 shows the flow scheme of the NMP process for the production of cyclopentene from a C_5 fraction.

The C_5 fraction is introduced to reactor (a), where cyclopentadiene is thermally dimerized. In column (b) the remaining diolefins of the C_5 fraction (mostly isoprene and piperylene) and cyclopentene are selectively extracted with *N*-methylpyrrolidone (NMP). The raffinate is distilled in column (c) to give the lower boiling olefinic and paraffinic C_5 hydrocarbons as distillates and crude dicyclopentadiene as bottoms. Crude dicyclopentadiene is monomerized in cracking column (d) to give cyclopentadiene at the top and high boilers at the bottom. Subsequently, cyclopentadiene is mixed with hydrogen and selectively hydrogenated in reactor (e). In drum (f), the effluent from (e) is separated into residual gas and liquid crude cyclopentene. The latter is combined with the mixture of C_5 diolefins and cyclopentene obtained after extraction in column (b). The combined stream is processed by extractive distillation with NMP in column (g) to give piperylene at the bottom and a mixture of cyclopentene and isoprene at the top. This mixture is subsequently separated by fractional distillation in column (h).

This process allows the simultaneous production of isoprene and cyclopentene as well as the isolation of the cyclopentene present in the C_5 fraction.

Alternative modes of formation of cyclopentene are catalytic cyclization of piperylene [108], metathesis of 1,6-cyclodecadiene and polypentenamer, thermal isomerization of vinylcyclopropane, and pyrolysis of polymers [65].

2.4. Uses

Cyclopentene does not find substantial use in chemical syntheses. The broadest potential is offered by the production of polypentenamer, which has been tested extensively as an elastomer. Of potential use are also the methods for the selective oxidative cleavage of cyclopentene to give the various bifunctional products of structures **30a** – **30e**. Alkylcyclopentylbenzenes have low pour points and high flash points

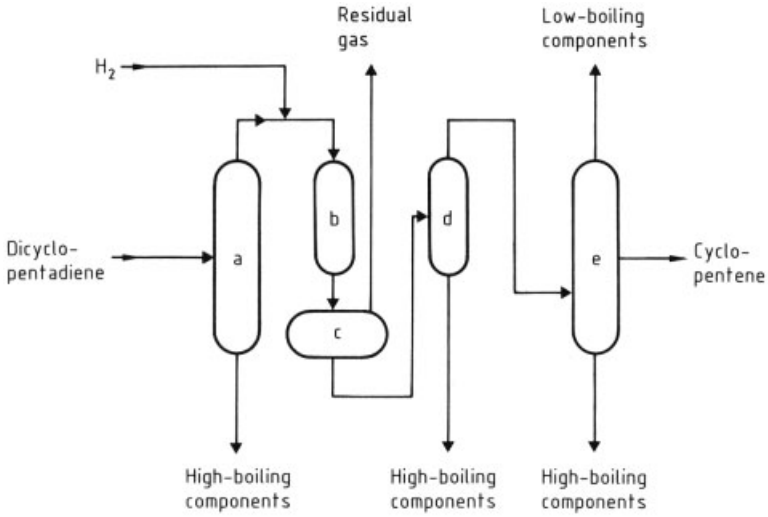

Figure 3. Bayer process for the production of cyclopentene from crude dicyclopentadiene
a) Cracking column; b) Hydrogenation reactor; c) Gas – liquid separator; d) Flash column; e) Distillation column

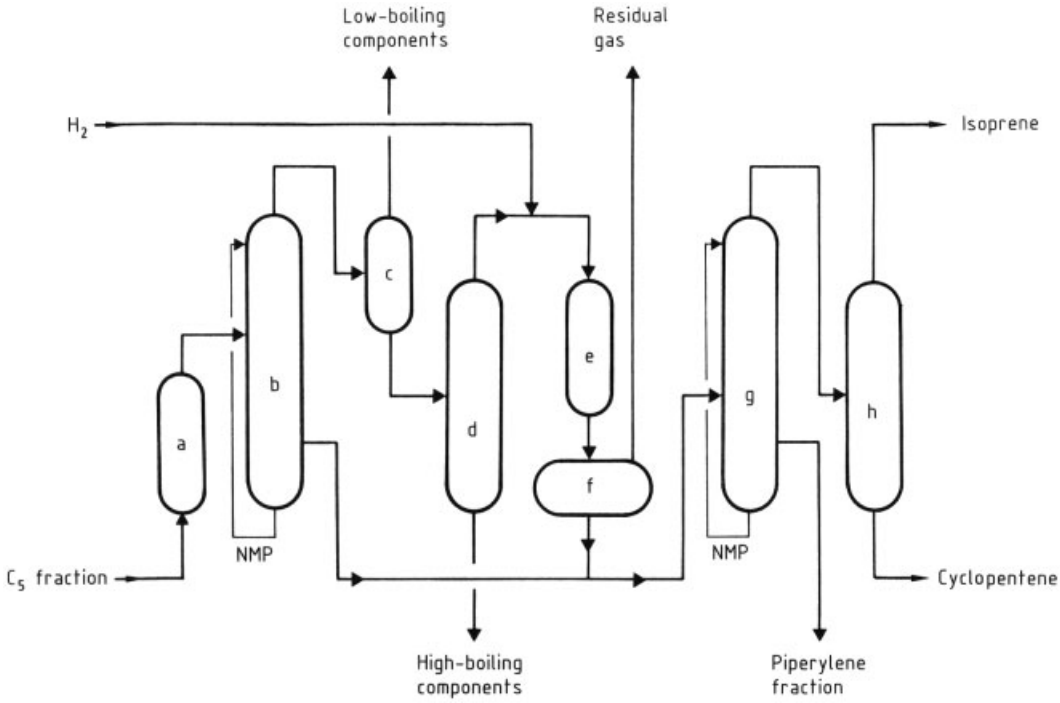

Figure 4. NMP process for the production of cyclopentene and isoprene from a C_5 fraction
a) Dimerization reactor; b) Liquid – liquid extractor; c) Distillation column; d) Cracking column; e) Hydrogenation reactor; f) Gas – liquid separator; g) Column for extractive distillation; h) Column for fractionated distillation

and are useful as working fluids, and cyclopentylphenols have been tested as disinfectants, antioxidants, and light stabilizers for polymers. Polycyclopentene is used as a nucleating agent in the polymerization of propylene, whereby the properties of the resulting polypropylene are modified [109]. Epoxycyclopentanol (**28**) is of potential use for the synthesis of prostaglandins.

3. Toxicology

Among the three chemicals, acute toxicity increases in the order cyclopentene, cyclopentadiene, dicyclopentadiene. However, cyclopentene is most readily absorbed through the skin.

Cyclopentadiene forms a dimer spontaneously, and the reported toxicity probably is that for the dimeric form. The oral LD_{50} in rats is 0.82 g/kg [110]. The dermal LD_{50} in rabbits is 6.72 mL/kg [110]. Subcutaneous injection of 0.5 – 1 mL of cyclopentadiene monomer had no detectable effects on rabbits; 3 mL produced narcosis and fatal convulsions [111]. Mild kidney and liver damage occurred after exposure of rats to 500 ppm for 35 daily exposures each of 7 h [112]. No toxicological effects were observed in rats, guinea pigs, rabbits, or dogs exposed to 250 ppm for 135 d, 7 h/d [112]. Dogs exposed to 400 ppm cyclopentadiene for 39 d, 6 h/d, then to 800 ppm for 16 d, 6 h/d, showed no abnormalities in hematological parameters, liver function, serum enzyme levels, electrocardiogram, or microscopic appearance of tissues [112]. There is no evidence to suggest mutagenic, carcinogenic, or teratogenic potential of cyclopentadiene. The odor of cyclopentadiene at 250 ppm is unpleasant; thus, human workplace exposure limits have been placed below this level. The ACGIH has adopted a TLV value for cyclopentadiene of 75 ppm and an STEL limit of 150 ppm [113]. The MAK value is 75 ppm [114].

Dicyclopentadiene, the tricyclic form of cyclopentadiene, is the most toxic of the cycloolefin compounds described under this keyword. Oral and intraperitoneal LD_{50} values in rats are 0.35 mL/kg and 0.31 mL/kg, respectively. The dermal LD_{50} value for rabbits is 5.08 mL/kg [115]. Exposure of rats to saturated dicyclopentadiene vapor for 1 h produced 100 % mortality [115]. LC_{50} values for 4 h exposure to dicyclopentadiene vapor range from 146 ppm in mice to ca. 375 ppm in rats and 770 ppm in guinea pigs and rabbits [115]. Rats exposed to 0, 72, 146, and 332 ppm (7 h/d, 5 d/week, 10 d) exhibited treatment-related mortality only in the highest concentration group and no significant toxicological signs in the other groups [115]. Exposure to dicyclopentadiene at concentrations up to 74 ppm, 7 h/d, 5 d/week for 89 d, caused increased liver and kidney weights and renal tubular degeneration in male rats [115]. Dogs exposed to 0, 9, 23, and 32 ppm for 89 d showed no dose-related pathological changes and only minimal changes in serum enzyme levels [115]. Dicyclopentadiene is irritating to the eyes and respiratory tract of humans at concentrations as low as 1 ppm for 7 min. [115]. The olfactory threshold is ca. 0.003 ppm [112]. ACGIH has adopted a TLV of 5 ppm for dicyclopentadiene, but no STEL value has been proposed [113]. No MAK value has been adopted, but Soviet and Bulgarian workplace limits are 1 mg/m^3 or 0.185 ppm [114].

Cyclopentene is moderately toxic. In rats, the oral LD_{50} is 2.14 mL/kg [116], and the 4-h LC_{50} for cyclopentene vapor exceeds 8110 ppm [117]. However, concentrated vapors are lethal and exposure to 16 000 ppm for 4 h caused death in 4/6 rats [116]. Transdermal absorption occurs readily; the dermal LD_{50} in rabbits is 1.59 mL/kg [116]. Subacute exposure (6 h/d, 5 d/week, 3 weeks) of rats to 0, 870, and 3100 ppm cyclopentene vapor produced no toxicologically significant effects [117]. Chronic inhalation (6 h/d, 5 d/week, 12 weeks) of up to 1139 ppm did not result in hematological alterations, changes in serum chemistry, or organ pathology in rats [117]. Cyclopentene has not been suggested to be mutagenic or carcinogenic and reproductive hazards are unknown. No specific human toxicity attributed to cyclopentene has been described; however, headache and unpleasant sensations were reported by humans exposed to 32 ppm for 10 min. [117]. The threshold for detection of the odor of cyclopentene is ca. 13 ppm [117]. No MAK value for cyclopentene has been officially adopted [114]; however, values in the range 10 – 15 ppm are suggested based on subjective effects [117]. No TLV or STEL limits for cyclopentene are in effect currently in the United States [113].

These chemicals are listed in the U.S. Environmental Protection Agency's inventory for the TSCA [118]. Little is known about the hazards of chronic exposure; however, these chemicals are not presently listed as candidates for genetic toxicity or carcinogenicity testing by the U.S. Public Health Service [119].

Signs and symptoms of human intoxication with these cycloolefins include irritation of the eyes, nose, throat, and skin, excitation, loss of coordination and equilibrium, hypothermia, dyspnea, central nervous system depression, stupor, and coma [120]. Toxic and irritant properties and the possibility of dermal absorption suggest strongly that prudent handling practices, including adequate ventilation, protective clothing, neoprene gloves, chemical goggles, and respirators (self-contained, full face-mask, organic vapor cartridge) be used when working with these chemicals.

4. References

1. *Kirk-Othmer* , 4th ed., **vol. 7** .
2. F. Asinger: *Die Petrolchemische Industrie,* Akademie-Verlag, Berlin 1971, pp. 447 – 451.
3. C. T. Lin, F. K. Young, M. R. Brulé, L. L. Lee, K. E. Starling, J. Chao, *Hydrocarbon Process.* **59** (1980) 117 – 123.
4. *Verfahrenstechnische Berechnungsmethoden, Teil 7, Stoffwerte,* VCH Verlagsgesellschaft, Weinheim 1986.
5. T. E. Taubert, R. P. Danner (eds.): *Data compilation, Part 2,* Design Institute for Physical Property Data, American Institute of Chemical Engineers, et al., Taylor & Francis, London 1989.
6. D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds,* J. Wiley and Sons, New York 1969, p. 346.
7. API 41 Selected Infrared Spectral Data vol. II, Thermodynamics Research Center, Chem. Eng. Division of Texas Engineering Experiment Station, College Station, Texas 1983.
8. W. Brügel: *Handbook of NMR Spectral Parameters,* vol. 1, Heyden and Son Ltd., London 1979.
9. A. Cornu, R. Massot: *Compilation of Mass Spectral Data,* **vol. 2,** Heyden, London-New York-Rheine 1975.
10. K. C. Ramey, D.C. Lini, *J. Magn. Reson.* **3** (1970) 94.
11. T. Teiter, R. L. N. Harris, *Aust. J. Chem.* **32** (1979) 1329-1337.
12. G. Bailo, P. Caramella, G. Cellerino, A. G. Invernizza, P. Gruenanger, *Gazz. Chim. Ital.* **103** (1973) 47 – 59.
13. M. B. McGinnis et al., *J. Org. Chem.* **61** (1996) 3496 – 3500.
14. Esso Research and Engineering Co., DE-OS 2 234 087, 1972 (F. Baldwin, G. Sartori, J. Lefebvre).
15. G. Maier, W. A. Jung, *Chem. Ber.* **115** (1982) 804 – 807.
16. Asahi Chem. Industry Co., JP-Kokai 92 353, 1973 (S. Miyata).
17. Mitsui Petrochem. Ind., JP 25 262, 1974 (M. Ogawa).
18. ICI, DE-OS 2 441 433, 1975 (P. Jhawar).
19. Japan. Geon Co., JP-Kokai 26 748, 1973 (R. Kita).
20. Z. Zhu, J. H. Espenson, *J. Am. Chem. Soc.* **119** (1997) 3507 – 3512.
21. B. F. Martynov, L. T. Lantseva, B. L. Dyatkin, *Zh. Org. Khim.* **11** (1975) 2282 – 2285.
22. A. Lubineau, J. Augé, E. Grand, N. Lubin, *Tetrahedron* **50** (1994) no. 34, 10265 – 10276.
23. K. Ishihara, S. Kondo, H. Kurihara, H. Yamamoto, *J. Org. Chem.* **62** (1997) 3026 – 3027.
24. A. Loupy, D. Monteux, *Tetrahedron Lett.* **37** (1996) no. 39, 7023 – 7026.
25. H. Mayr, I. Halberstadt, *Angew. Chem.* **92** (1980) 840 – 841; *Angew. Chem. Int. Ed. Engl.* **19** (1980) 814.
26. R. L. Danheiser, S. K. Gee, H. Sard, *J. Am. Chem. Soc.* **104** (1982) 7670 – 7672.
27. C. Y. Wu, H. E. Swift, *Prepr. Am. Chem. Soc. Div. Petr. Chem.* **25** (1980) no. 2, 372.
28. Hitachi Chem, Co., JP-Kokai 31 970, 1972 (H. Kouchi, T. Akima).
29. Daichi Kogyo Seiyaku Co., JP-Kokai 39 689, 1974 (I. Morita, H. Shimoda, I. Nishizawa).
30. G. Heublein, B. Adelt, *Plaste Kautsch.* **19** (1972) 728 – 730.
31. D. B. Knight, R. W. Hall, D. G. Cleary, *J. Heterocycl. Chem.* **18** (1981) 1649 – 1650.
32. K. J. Stone, R. D. Little, *J. Org. Chem.* **49** (1984) 1849 – 1853.
33. W. Adam, H. J. Eggelte, *J. Org. Chem.* **42** (1977) 3987 – 3988.
34. Hitachi Chem. Co., JP 21 093, 1973 (T. Fujiki, H. Tsukumi, Y. Yamazaki).
35. Ube Ind., JP-Kokai 13 113, 1974 (S. Umemura, F. Sakai).

36. D. Hönicke, R. Bührer, A. Newrzella, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) 569.
37. Hoechst, DE-OS 3 244 032, 1984 (R. Bicker).
38. W. Meyer, *Hydrocarbon Process.* **55** (1976) no. 9, 235 – 238.
39. *Winnacker-Küchler,* **5,** 191, 247.
40. W. Graulich, W. Swodenk, D. Theisen, *Hydrocarbon Process.* **51** (1972) no. 12, 71 – 75.
41. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie,* Verlag Chemie, Weinheim-New York 1976, p. 98.
42. Institute of Petrochemical Processes, SU 1 109 371, 1983 (M. R. Musaev, S. M. Sharifova, S. M. Mirzoev, T. A. Samedova, M. A. Aliev, Y. G. Mamedaliev).
43. M. M. Ermilova, N. L. Basov, V. S. Smirnov, A. N. Rumyantsev, V. M. Gryaznov, *Izv. Akad. Nauk SSSR, Ser. Kim.* **8** (1979) 1773 – 1775; *Chem. Abstr.* **92** (1980) 22 104 p.
44. G. V. Jsagulyants, K. M. Gitis, M. I. Rozengart, V. N. Kornyshev, G. L. Markaryan, *Geterog. Katal. 5th.* 1983, 511 – 516; *Chem. Abstr.* **100** (1984) 191 200 f.
45. Acad. Sc. Belorussian SSR, SU 1 081 153, 1982 (Y. G. Egiazarov, M. G. Savon'kina, N. S. Bobchenok).
46. R. W. Carr, J. O. Shoemaker, *Chem. Eng. Commun.* **19** (1982) no. 1 – 3, 91 – 98.
47. M. Tokuda, H. Suginome, L. L. Miller, *Tetrahedron Lett.* **23** (1982) 4573 – 4576.
48. P. Ausloos, S. G. Lias, R. E. Rebbert, *J. Phys. Chem.* **85** (1981) no. 2, 2322 – 2328.
49. Z. Kafka, M. Kuras, L. Vodicka, *Sb. Vys. Sk. Chem. Technol. Praze Technol. Paliv* **D 47** (1983) 199 – 209; *Chem. Abstr.* **100** (1984) 141 776 c.
50. M. J. Ledoux, *Appl. Catal.* **9** (1984) no. 1, 31 – 40.
51. E. Weckman, K. Brezinsky, J. DeMay, I. Glassmann, *Chem. Phys. Processes Combust,* 1983, no. 35; *Chem. Abstr.* **101** (1984) 93 818 q.
52. R. Cyprés, *Compend. Dtsch. Ges. Mineraloelwiss. Kohlechem. 1976/77* (1976) no. 1, 435 – 455.
53. W. Kaminsky, *Angew. Makromol. Chem.* **232** (1995) 151 – 165.
54. R. Kubicka, M. Goppoldova, CS 178 509, 1973.
55. M. Goppoldova, M. Goppold, *Chem. Prum.* **30** (1980) no. 7, 362 – 366; *Chem. Abstr.* **94** (1981) 102 899 h.
56. M. Goppoldova, *Chem. Prum.* **30** (1980) no. 8, 408 – 413; *Chem. Abstr.* **94** (1981) 102 872 u.
57. M. A. S. Costa, F. M. B. Coutinho, *Polym. React. Eng.* **2** (1994) 241 – 250.
58. Kohijin Co., DE-OS 2 217 623, 1972 (A. Oshima, K. Tsuboshima, N. Takahashi). JP-Kokai 45 011, 1974 (N. Takahashi, M. Nonaka).
59. R. Bloch, *Tetrahedron* **39** (1983) 639 – 643.
60. G. Mehta, D. S. Reddy, A. V. Reddy, *Tetrahedron Lett.* **25** (1984) 2275 – 2278.
61. W. Kaminsky, A. Bark, *Polym. Int.* **28** (1992) 251 – 253.
62. P. Köpf-Maier, H. Köpf, *J. Organomet. Chem.* **342** (1988) 167 – 176.
63. T. Chiba, *Akita Kogyo Koto Senmon Gakko Kenkyu Kiyo* **32** (1996) 42 – 45.
64. Sekisui Chemical Co. Ltd., EP 0 566 794 A1, 1993 (H. Isogawa, H. Anraku).
65. Pennzoil Products Company, US 4 849 566, 1989 (C. G. Venier, E. W. Casserly).
66. *Kirk-Othmer* , 4th ed., **vol. 8** .
67. J. H. Lukas, A. P. Kouwenhoven, F. Baardman, *Angew. Chem.* **87** (1975) 740 – 741; *Angew. Chem. Int. Ed* . Engl. 14 (1975) 709.
68. B. B. Snider, L. A. Brown, R. S. Eichen Conn, T. A. Killinger, *Tetrahedron Lett.* 1977, 2831 – 2832.
69. D. Bryce-Smith, B. Foulger, J. Forrester, A. Gilbert, B. H. Orger, H. M, Tyrrell, *J. Chem. Soc. Perkin Trans I* 1980, 55–71.
70. D. Bryce-Smith, W. M. Dadson, A. Gilbert, B. H. Orger, H. M. Tyrrell, *Tetrahedron Lett.* 1978, 1093 – 1096.
71. L. D. Brown, K. Itoh, H. Suzuki, K. Hirai, J. Ibers, *J. Am. Chem. Soc.* **100** (1978) 8232 – 8238.
72. B. L. Booth, T. A. El Fekky, G. F. M. Noori, *J. Chem. Soc. Perkin Trans I* 1980, 181 – 186.
73. J. Rebek, R. Mc. Cready, S. Wolf, A. Mossman, *J. Org. Chem.* **44** (1979) 1485 – 1493.
74. J. Rebek, R. Mc. Cready, *Tetrahedron Lett.* 1979, 1001 – 1002.
75. H. Kropf, M. R. Yazdanbakhch, *Synthesis* 1977, 711 – 713.
76. W. Reischl, E. Zbiral, *Tetrahedron Lett.* 1979, 1109 – 1110.
77. Degussa, DE-OS 2 937 810, 1981 (G. Kaebisch, H. Malitius, S. Raupach, R. Truebe, H. Wittmann).
78. H. Ogawa, H. Fujinami, K. Toya, *J. Chem. Soc. Chem. Commun.* 1981, 1274 – 1275.
79. H. C. Brown, A. K. Mandal, *Synthesis* 1980, 153 – 155.
80. B. F. Goodrich, US 3 808 283, 1974.

81. Mobil Oil Corp., US 4 255 600, 1981 (L. Young).
82. F. R. Mayo, P. S. Fredricks, T. Mill, J. K. Castleman, T. Delaney, *J. Org. Chem.* **39** (1974) 885 – 889.
83. A. J. Bloodworth, H. J. Eggelte, *J. Chem. Soc. Perkin Trans I,* 1981, 1375 – 1382.
84. K. Blau, U. Müller, W. Pritzkow, W. Schmitt-Renner, Z. Sedshaw, *J. Prakt. Chem.* **322** (1980) 915 – 932.
85. E. O. Mihelich, D. J. Eickhoff, *J. Org. Chem.* **48** (1983) 4135 – 4137.
86. Procter and Gamble Co., US 4 345 984, 1982 (E. Mihelich).
87. A. Heumann, B. Akermark, *Angew. Chem.* **96** (1984) 443 – 444; *Angew. Chem. Int. Ed. Engl.* **23** (1984) 453.
88. K. Griesbaum, J. Neumeister, M. P. Saxena, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) 252 – 257.
89. S. L. Schreiber, R. E. Claus, J. Reagan, *Tetrahedron Lett.* **23** (1982) , 3867 – 3870.
90. Bayer, DE-OS 2 201 456, 1973 (H. Waldmann, W. Schwerdtel, W. Swodenk). DE-OS 2 252 674, 1974 (H. Waldmann, W. Schwerdtel, W. Swodenk). DE-OS 2 261 657, 1974 (K. Wedemeyer, A. Klein).
91. Nippon Zeon Co., JP 29 165, 1974.
92. D. Hönicke, K. Griesbaum, *Chem. Ing. Tech.* **54** (1982) 497.
93. Nissan Chem. Ind., JP-Kokai 64 020, 1973 (A. Murata, H. Suzuki).
94. Gulf Research and Dev. Co., US 4 062 902, 1976 (A. Montagna).
95. Anic S.p.A., DE-OS 2 622 917, 1975 (D. Sanfilippo, M. Morelli).
96. A. V. Topchiev, DE-OS 3 100 631, 1981.
97. M. B. Taghavi, G. Pajonk, S. J. Teichner, *Bull. Soc. Chim. Fr.* 1978, nos. 7 – 8, part 1, 285 – 293.
98. Yu. S. Rozov, B. A. Grigorovich, V. Sh. Fel'dblyum, A. M. Kutin, I. M. Galperin, K. S. Solovev, V. S. Elizarov, SU 535 099, 1975.
99. A. V. Topchiev, NL-A 7 408 819, 1974.
100. L. Cerveny, J. Vopatova, V. Ruzicka, *React. Kinet. Catal. Lett.* **19** (1982) no. 1 – 2, 223 – 226; *Chem. Abstr.* **97** (1982) 40 091 q.
101. Goodyear Tire and Rubber Co., US 4 131 627, 1977 (L. G. Wideman).
102. Goodyear Tire and Rubber Co., US 4 131 629, 1977 (L. G. Wideman, E. A. Ofstead).
103. Y. Nakamura, H. Hirai, *Chem. Lett.* 1976, no. 11, 1197 – 1202.
104. L. I. Gvinter, V. Sh. Fel'dblyum, L. Kh. Freidlin, L. N. Suvorova, *Neftekhimiya* **23** (1981) no. 1, 41 – 44; *Chem. Abstr.* **98** (1983) 179 944 j.
105. Goodyear Tire and Rubber Co., US 4 188 348, 1978 (H. R. Menapace).
106. SABA, EP-A 9 035, 1978 (H. R. Menapace).
107. Bayer, DE-OS 2 025 411, 1970 (W. Schwerdtel, W. Swodenk, P. Woernerle).
108. I. M. Kolesnikov, N. N. Belov, *Zh. Fiz. Khim,* **52** (1978) no. 10, 2712; *Chem. Abstr.* **90** (1979) 86 800 d.
109. D.-H. Lee, K.-B. Yoon, *J. Appl. Polym. Sci.* **54** (1994) 1507 – 1511.
110. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, U. C. Pozzani, *AMA Arch. Ind. Hyg. Occup. Med.* **10** (1954) 61 – 68.
111. W. F. von Oettingen, *U.S. Public Health Bulletin* no. 255 (1940) 40 – 41.
112. Dow Chemical Company, Biochemical Laboratory, unpublished results (1964), cited in [113]
113. American Conference of Governmental Industrial Hygenists: *Documentation of the Threshold Limit Values,* 4th ed., ACGIH, Cincinnati 1980.
114. International Labor Office: *Occupational Safety and Health Series no. 37,* "Occupational Exposure Limits for Airborne Toxic Substances," 2nd ed., International Labor Office, Geneva 1980.
115. E. R. Kinkead, U. C. Pozzani, D. L. Geary, C. P. Carpenter, *Toxicol. Appl. Pharmacol.* **20** (1971) 552 –561.
116. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, U. C. Pozzani et al., *Am. Ind. Hyg. Assoc. J.* **30** (1969) 470 – 476.
117. G. Kimmerle, J. Thyssen, *Int. Arch. Arbeitsmed.* **34** (1975) 177 – 184.
118. U.S. Environmental Protection Agency, Office of Toxic Substances: *Candidate List of Chemical Substances,* Toxic Substances Control Act, U.S. EPA, Washington 1983.
119. U.S. Department of Health and Human Services, Public Health Service: *National Toxicology Program,* Fiscal Year 1985 Annual Plan, NTP, Research Triangle Park 1985.
120. E. E. Sandmeyer in *Patty's Industrial Hygiene and Toxicology* **vol. 2 B,** Wiley Interscience, New York 1981, pp. 3221 – 3251.

Cystine, Cysteine → Amino Acids
Cytokinins → Plant Growth Regulators

Cytostatics → Cancer Chemotherapy
Dairy Products → Milk and Dairy Products

Dairy Products, Imitation

PETER P. NOZNICK, Beatrice Companies, Chicago, Illinois 60602, United States

1. **Introduction** 1
2. **Definition of Imitation Milk** 1
3. **Ingredients** 2
4. **Manufacture** 4
5. **Products** 4
6. **Standards and Quality Control** 9
7. **Nutrition** 9
8. **Toxicology** 9
9. **Economic Aspects** 10
10. **References** 10

1. Introduction

In 1869, the French pharmacist HIPPOLYTE MEGE-MOURIEZ invented *olemargarine* as a substitute for butter; this was probably one of the earliest examples of an imitation dairy product [1]. *Margarine* was offered as a cheaper spread than butter, and almost 72 years passed before it was accepted as a food in its own right. In 1941 the FDA promulgated a federal standard for margarine [2], and in 1957 sales of 3.9 kg of margarine per capita surpassed the sales of butter [3].

During this growth period, the United States Congress passed the Filled Cheese Act of 1896 [4] and the Filled Milk Act of 1923 [5], which prohibited the development of substitute milk and cheese products. *Filled milk* (or *filled cheese*) is a substitute dairy product in which all or part of the butterfat is replaced with other fats or oils.

In the late 1940 s and the early 1950 s, sodium caseinate [*9005-46-3*], which was considered a chemical by the U.S. government, was found to be an excellent and inexpensive substitute for skim milk. The substitution of sodium caseinate opened the way for the development of imitation dairy products.

United States government actions, such as the repeal of the Filled Cheese Act and a federal court decision that the Filled Milk Act was unconstitutional, encouraged further development and marketing of imitation dairy products [6–8].

In 1987, there was no federal legal bars to the production or sale of imitation dairy products in the United States. Elsewhere in North America, the sale of some imitation dairy products is permitted in Mexico, but none are permitted in Canada.

The EEC countries allow the sale of nondairy coffee creamers and nondairy whipped toppings, provided they are not advertised as replacements of the dairy products simulated. There is concern in the EEC countries of the increasing sales of the nondairy products, and alternative choices for control are being considered by the Food and Drink Federation in the United Kingdom.

In Southeast Asia, canned imitation evaporated milk is legal in the Philippines, and imitation sweetened condensed milk is legal in Malaysia and Singapore.

2. Definition of Imitation Milk

The FDA defines a food as an imitation if it is a substitute for or resembles another food, but is nutritionally inferior to that food [9]. Nutritional inferiority includes any reduction in the content of an essential nutrient present in a measurable amount. A measurable amount per serving is 2 % or more of the recommended daily allowance [9], [10].

To paraphrase the FAO definition, an imitation milk product is prepared when one or more milk constituents have been replaced with other ingredients and the resulting food resembles the original product [11].

The term "imitation" refers to all simulated dairy products, except for nondairy creamer and nondairy whipped topping. Both were well-recognized market products prior to the estab-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_239

lishment of the present FDA definition of an imitation food. These names, commonly used in the United States, are also used in this article.

3. Ingredients

Skim milk is the portion of milk that remains after milk fat has been removed by cream separation [12], [13]. Skim milk typically contains 90.8 wt % moisture, 3.41 wt % protein, 0.18 wt % fat, 4.85 wt % lactose [*63-42-3*], 0.76 wt % ash, and 0.12 wt %calcium, and 0.10 wt % phosphorus, as contained in the ash.

Nonfat dry milk is skim milk with only the fat and water removed. Standards in the United States and those proposed by the FAO [11] allow a maximum of 5 % moisture and less than 1 % fat [14].

Casein [*9000-71-9*] is prepared from skim milk by precipitation with acid at pH 4.6 or by the action of the enzyme rennin [*9001-98-3*] [15–19]. Edible dry casein is graded according to USDA standards as Extra or Standard [20].

Sodium Caseinate [*9005-46-3*] is made by solubilization of isoelectric casein with food-grade caustic soda and subsequent spray-drying. Caseinates are excellent food ingredients and are used in at least 26 food items [16].

Soy protein isolate is prepared from defatted soy flakes by separating and spray-drying the soluble protein fraction [21]. The dry isolate contains 90 – 95 % protein.

Whey protein concentrate made by ultrafiltration is a demineralized product rich in protein and with much of the lactose removed [22]. It has the advantage of being soluble over a pH range of 2 – 10 unless heated above 70 °C.

Vegetable fats constitute an important ingredient in imitation dairy products [23]. They contribute organoleptic properties such as richness and "mouthfeel," provide body or viscosity, play an important role in the air-incorporating capability (in whipped topping), increase whitening capacity (in nondairy creamers), and add tenderness (in imitation cheese).

The substitution of a vegetable fat for butterfat is complicated because available natural fats have fatty acid profiles which differ from that of butterfat [24–27]; see also → Fats and Fatty Oils. Fatty acids largely determine the physical properties of fats (triglycerides), which are esters of glycerol and fatty acids.

Table 1 compares the fatty acid composition of several functional vegetable fats with that of butterfat. Butterfat contains a variety of fatty acids, and some 70 % of them are saturated [24]. By contrast, domestic vegetable fats have fewer types of fatty acids, and these are of longer chain length, i.e., C_{16} and C_{18} and are saturated and unsaturated [25–27]. The unsaturated fatty acids listed in Table 1 are called palmitoleic and oleic if they have one unsaturated double bond; linoleic, if they have two; and linolenic, if they have three double bonds. Coconut fat, a lauric fat, contains approximately 95 % of a mixture of saturated fatty acids with short, intermediate, and long chain lengths. It is especially high in lauric acid [*143-07-7*] [25].

Every fat consists of a physical mixture of microscopic crystals of solid fat that holds the liquid oil. The *Solid Fat Index* (SFI) defines the ratio of solid fat to liquid oil at a given temperature. This measure indicates both the physical properties and the eating quality of a given fat [28]. The SFI is determined by dilatometry [28] and pulsed ^{1}H-NMR spectrometry [29].

According to the SFI profiles in Figure 1, the fat component of margarine (a) compares very closely with that of butterfat (b) [25], [30]; however, as shown in Table 1, it has a very different combination of fatty acids [24], [26]. Both have the moderate range of plasticity required for good spreadability at 25 °C [28]. By contrast, a hydrogenated coconut fat and hydrogenated palm kernel fat (lauric hard butter), as shown in Figure 1 by curves (c) and (d), respectively, contain relatively large amounts of solid fats [25], [31] at 15.5 – 21.1 °C and almost none at body temperature [25], [31]. The high content of solid fats in the fat mixture at room temperature adds body and firmness to the finished food and lessens the oil migration that causes greasiness. However, the rapid drop in the content of solid fat between room temperature and the final melting temperature in the mouth is responsible for the agreeable cooling effect of the product when tasted. These characteristics distinguish the lauric fats such as coconut and palm kernel oil from domestic oils and fats, and are preferred in whipped toppings and other imitation dairy products. Hard butters that closely resemble lauric fats, as shown by curve (e) in Figure 1, have been developed from domestic oil [32]. Be-

Table 1. Fatty acid composition and iodine values of four fats ∗∗

Fatty acid	Butterfat, anhydrous, wt % [24]	Hydrogenated coconut oil, wt % [25] ∗	Regular hard margarine, wt % [26]	Hard butter, wt % [27]
Butyric [*107-92-6*]	3.23			
Caproic [*142-62-1*]	1.91	0.1		
Caprylic [*124-07-2*]	1.11	6.6		
Capric [*334-48-5*]	2.50	6.0		
Lauric [*143-07-7*]	2.79	48.0		
Myristic [*544-63-8*]	10.0	17.8		
Palmitic [*57-10-3*]	26.17	8.4	11.0	2
Stearic [*57-11-4*]	12.06	8.4	7.0	6
Palmitoleic [*10030-73-6*]	2.23			
Oleic [*112-80-1*]	25.03	3.6	59.0	83
Linoleic [*60-33-3*]	2.25	0.6	17.0	5
Linolenic [*463-40-1*]	1.45		1.0	
Iodine value		3	83	80

∗ Melting point 33 °C;
∗∗ wt % values based on USDA data on fatty acid composition expressed as grams fatty acids in 100 grams of fat.

cause of their lower cost, these products are now used as substitutes for the substitutes.

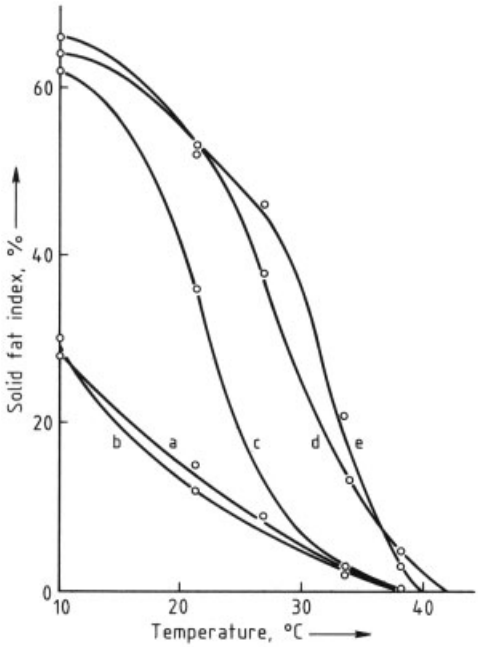

Figure 1. Solid fat index profile for several fats
a) Table margarine [30]; b) Butter oil [25]; c) Hydrogenated coconut fat [25]; d) Lauric hard butter [31]; e) Nondairy specialty fat [32]

Emulsifiers. The emulsifiers used in imitation dairy product formulations are esters of edible fatty acids of vegetable or animal origin and polyvalent alcohols such as glycerol [*56-81-5*], propylene glycol [*57-55-6*], and sorbitol [*50-70-4*] [33–38]. These emulsifiers may also be esterified with organic acids such as lactic, tartaric, and acetic acid or also reacted with ethylene oxide [35], [39–42]. Another group of emulsifiers is made by esterification of edible fatty acids with polyglycerols [43–46]. Lecithin [*8002-43-5*], a natural emulsifier, is extracted commercially from soybean oil [47], [48]. Many emulsifiers also contain small amounts of sodium or potassium salts of fatty acids, e.g., sodium oleate [*143-19-1*] [49].

Food emulsifiers reduce the surface tension of the oil – water interface; they adhere to the surface of the oil droplets and thus prevent their coalescence (see → Emulsions). Emulsifiers also stabilize the basic sodium caseinate – vegetable oil emulsion found in many imitation dairy products [50].

Carbohydrates. Carbohydrates are used as sweetening agents; they constitute a significant fraction of the total solids in imitation dairy products. The major sugars used are cane sugar and corn syrup solids at a sweetener level of 42 dextrose equivalents (DE). Corn syrup solids at DE levels of 20 – 25 are employed as carriers and encapsulating agents in spray-dried whip toppings and nondairy creamers [51]. Lactose [*63-42-3*] is used as an imitation ingredient for lactic acid fermentation, e.g., in imitation sour cream, salad dressings, and dips [52]. Whey is often used as a source of lactose.

Stabilizers. Colloidal polysaccharides (gums), such as carrageenan, locust bean gum, sodium alginate, guar gum, or carboxymethyl cellulose, thicken or increase viscosity [53–57]. Some polysaccharides have the following specific effects:

1) improvement of freeze – thaw stability, e.g., xanthan gum [*11138-66-2*] [58];
2) retardation of the growth of ice crystals, e.g., carboxymethyl cellulose [*9004-32-4*] or sodium alginate [*9005-38-3*] [55], [57];
3) prevention of syneresis, e.g., carboxymethyl cellulose, guar gum, xanthan gum, or microcrystalline cellulose [56–59];
4) stabilization of dispersions, e.g., cocoa powder (carrageenan [*9000-07-1*]) [53];
5) stabilization of oil-in-water emulsions, e.g., methyl cellulose [*9004-67-5*], hydroxypropyl methyl cellulose [*9004-65-3*], or hydroxypropyl cellulose [*9004-64-2*] [60–62]; and
6) improvement of cheese meltdown, e.g., locust bean gum [*9000-40-2*] [54].

Additives. Food-grade salt is used as a flavorant [63]. Food-grade disodium hydrogen phosphate [*7558-79-4*] and sodium citrate [*68-04-2*] are used as buffers for pH control in nondairy creamers and as emulsifiers in imitation cheese. Usage of flavors and coloring agents varies locally.

4. Manufacture

Imitation dairy products may be formulated with ingredients of dairy origin but are not limited to them. They resemble and substitute for dairy products. These products are processed largely with standard dairy equipment by using procedures similar to those of the dairy industry [64], [65]. Homogenization is used to disperse the fat uniformly and to control the size of the globule, which ranges from 1 to 3 μm in the finished emulsion.

Pasteurization can be carried out by either the batch or the vat method, or the high-temperature short-time method (HTST). Batch pasteurization is carried out at 63 °C for a minimum of 30 min. In practice, values of 71 °C for 30 min are used to adjust for higher fat and sugar levels. The HTST method is a continuous process that pasteurizes milk at a temperature of 74 °C for 15 s, although temperatures of 80 °C are not uncommon. This method can be substituted for the batch system of pasteurization in Figure 2. It is recommended when production exceeds 907 kg/h for a minimum of 3 h.

The manufacture of imitation dairy products involve a sequence of several consecutive but different unit operations, such as blending, or homogenization (Fig. 2). Only procedures that deviate from the common procedure illustrated in Figure 2, are described in detail when the various imitation dairy products are treated.

5. Products

Imitation Milk. Imitation milk is manufactured and sold in areas where fresh milk is unavailable, such as the islands of the Pacific Ocean, tropical areas of Latin America, Southeast Asia, and military installations. The source of protein is usually nonfat dry milk or, less frequently, fluid skim milk. Sodium caseinate may be used with concentrated whey protein and corn syrup solids at a DE of 42. Lauric- type fats, such as hydrogenated coconut or palm kernel fats, are generally used.

Imitation milk typically contains 3.25 % vegetable fat, 0.15 – 0.25 % mono- and diglycerides [66], 9 % nonfat dry milk solids, coloring agents, flavor ingredients, and vitamins, all in water. Imitation milk is made according to the basic procedure outlined in Figure 2; the homogenization pressure is 13.7 MPa in the first stage and 3.4 MPa in the second.

Soy Milk. Soy milk originated in the Orient two millenia ago and was introduced to the West around 1897 [67]. Soy milk is manufactured and sold in the Americas to religious groups, vegetarians, those (including infants) unable to consume cow's milk, and the general consuming public. If soy milk is sold in the United States, it must be labeled Imitation Soy Milk according to FDA standards; otherwise, the word "milk" must be avoided [68] and a new fanciful name used. Soy milk is prepared traditionally by grinding soaked soybeans with cold water, cooking them, and extracting the milk. A representative example contains 11.7 % solids, 3 – 3.2 % pro-

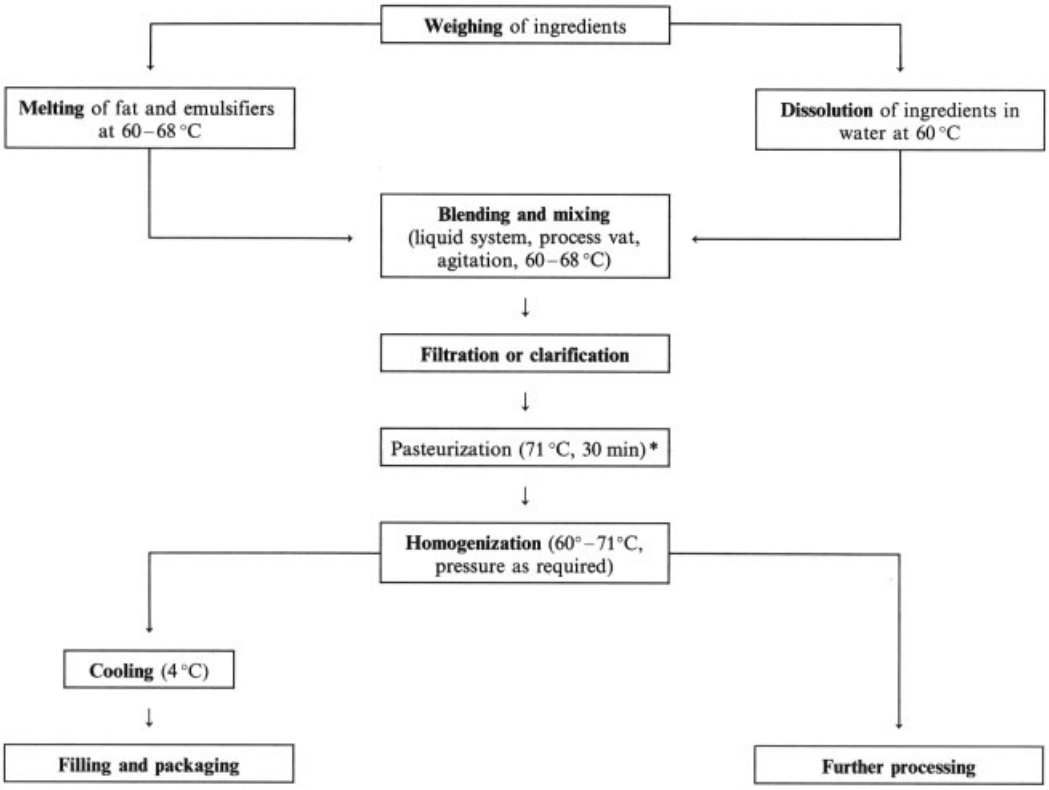

Figure 2. General processing steps for the manufacture of imitation dairy products
∗ except for sweetened condensed milk [80–90 °C; 0.5–2 min (HTST) or 10 min (batch)]

tein, 4.1 % oils or fat, 4.0 % sugars, and 0.4 % ash.

Imitation Sour Cream. Imitation sour cream is made from a simulated or imitation cream formulated of 17.5 % vegetable fat, 9 % nonfat milk solids, 0.5 % mono- and diglycerides, suitable coloring agents and flavor ingredients, all made up to 100 % with water.

The preparation process is carried out according to the basic procedure outlined in Figure 2; the homogenization pressure is 13.7 MPa in the first stage and 3.4 MPa in the second. Sour imitation cream is made by inoculating imitation cream at 22 °C with a starter culture and allowing the lactic acid flavor to develop. It is then packaged and allowed to set 24–28 h at 4.4 °C [64]. Direct acidification with lactic acid, citric acid, and other acids may also be used to make sour imitation cream, which circumvents the lengthy fermentation time, and assures a predictable flavor and product texture.

Dehydrated imitation sour cream is prepared by spray-drying with either a box-type or a tower spray dryer [65]. It is used as an ingredient in such convenience foods as dehydrated sauces and dehydrated liquid salad dressings [61].

Nondairy Creamers. Nondairy creamers are simulated creams used to whiten coffee. They have gained wide consumer acceptance because of lower price, convenience, and good shelf life. They are marketed in liquid, frozen, and powdered forms. The liquid form is used in small single-service containers where product turnover is rapid. The frozen variety has good shelf stability and is sold through retail and institutional channels. The dehydrated form has excellent shelf stability, does not require refrigeration, and is convenient to use. It is usually added directly to coffee in the powdered form, although it may be reconstituted.

Nondairy creamers must show good whitening power and be free of the defect called feath-

ering, or oiling-off, shown by the appearance of white coagulum on the surface of the coffee. Stabilizer salts (dipotassium hydrogen phosphate and sodium citrate) are essential to prevent feathering, which is caused by the pH of the water, the concentration of carbonates in the water, and the inherent acidity of coffee itself. Brews of roasted coffee vary in pH from 4.1 to 5.3, according to the area of origin. Per serving, well-stabilized nondairy creamer will raise the pH of coffee from 4.8 to 5.6, well above the coagulation pH of sodium caseinate (pH 4.6).

A typical liquid nondairy creamer contains 10.5 % corn syrup solids (25 DE), 10 % partially hydrogenated coconut oil or hydrogenated soy oil; 1.0 % sodium caseinate; 0.4 % emulsifier – a mixture of glycerol monostearate [*31566-31-1*] and distearate, sodium stearyl-2-lactylate [*25383-99-7*], and polyoxyethylene(20) sorbitan monostearate (Polysorbate 60); 0.10 – 0.15 % dipotassium hydrogen phosphate or disodium hydrogen phosphate; 0.1 – 0.2 % of a mixture of xanthan gum and guar gum; flavoring and coloring (β-carotene [*7235-40-7*]) agents as required; and water to 100 %.

Soy protein isolate can be substituted for sodium caseinate, and other emulsifiers, such as sorbitan monostearate and diacetyltartaric acid esters of monoglycerides, may be used.

Processing is outlined in Figure 2; the homogenization pressure is 13.7 MPa in the first stage and 3.4 MPa in the second. Homogenization is followed by cooling to 4.0 °C, and the product is packaged and stored at low temperature [64].

Frozen nondairy creamers are similar in formulation to the liquid form, except that greater attention is paid to the selection and concentration of the gums included.

Spray-Dried Nondairy Creamers. The basic ingredients are vegetable fat, emulsifier, protein (usually sodium caseinate), and a carbohydrate sweetener. The formulation of the powder is designed to provide sufficient fat to whiten the coffee when added in powdered form, and sufficient protein to encapsulate the oil during drying. The powder then flows freely and will not oil off in coffee. A typical spray-dried nondairy creamer contains 64.0 % corn syrup solids (25 DE), 25.75 % hydrogenated coconut or soybean oil, 4.0 % sodium caseinate, 1.3 % glycerol mono- or distearate, 2.0 % dipotassium hydrogen phosphate, 1.45 % anticaking agent (sodium silicoaluminate), 1.0 % moisture, and 0.25 % coloring and flavoring agents. This creamer is processed (as outlined in Fig. 2, except that the anticaking agent is withheld) by mixing the ingredients with water; the mix is adjusted to 45 % total solids at 60 °C, pasteurized at 71 °C for 30 min, homogenized at 13.7 MPa in the first stage and 3.4 MPa in the second, spray-dried in a vertical dryer, mixed with the anticaking agent, and agglomerated. Large particles improve the wettability of the powder in hot coffee [69], [70].

Evaporated Imitation Milk. Evaporated imitation milk is produced in the United States and in areas such as Southeast Asia. The formula used in the United States includes 6.0 % vegetable fat, 20 % skim milk solids (26 % total solids). The fats used are partially hydrogenated soy and cottonseed, and palm and coconut fats in the Asian area. Evaporated imitation milk is manufactured by the conventional method of water removal by vacuum pan evaporation [70] or by recombination of nonfat dry milk solids with water and vegetable fat to achieve the required solids level [71]. The addition of the proper amount of disodium hydrogen phosphate to the concentrated product before sterilization is essential to both processes.

In the *conventional method,* the fluid skim milk is prewarmed by passage through a single-pass tubular heater and concentrated in a vacuum pan. The concentrate is dropped, mixed with vegetable fat and calcium carrageenan, homogenized, standardized, treated with disodium hydrogen phosphate, fortified according to the local health code, packaged, and sterilized at 117 °C for 15 min or 137 °C for 2 min [65], [70], [72].

Evaporated imitation milk is manufactured largely by *recombination and reconstitution* as practiced in Southeast Asia, which precludes the need for an evaporation step. High-heat nonfat dry milk with a whey protein nitrogen index of 1.5 mg/g of nitrogen is recommended [71].

The procedure deviates from that shown in Figure 2, because after blending and filtration, the mix is preheated to 60 °C and homogenized in two stages at 14 MPa and 3.4 MPa. After homogenization, the viscosity of the prod-

uct is determined, the solids are adjusted, and monosodium and disodium hydrogen phosphate are added as required for stability. Furthermore the product is fortified with vitamins A and D as required by local code, canned, and sterilized (117 °C for 15 min) [71], [72].

Sweetened Condensed Imitation Milk. Sweetened condensed imitation milk is concentrated imitation milk preserved with a high concentration of sucrose. It has an excellent shelf life because the high osmotic pressure caused by the increased concentration of dissolved sugar inhibits the growth of microorganisms [70]. Sweetened condensed imitation milk is manufactured by vacuum pan concentration, to a limited extent, in Southeast Asia [73]. However, the trend has been toward recombination – reconstitution.

The selection of nonfat dry milk is important to avoid quality problems, such as excess viscosity in the finished product, especially on storage [74]. Pilot-scale evaluation of the effect of nonfat dry milk on the finished product is more important than the heat-treatment history of nonfat dry milk ingredients [75]. High grade crystalline sugar and partially hydrogenated lauric or palm fats are used.

Manufacturing procedures for sweetened condensed milk vary in formula, although two are used most frequently: (1) the U.S. Standard, which consists of 20 % nonfat solids, 8 % fat, 45.4 % sucrose with total solids of 73.4 %, and (2) the EEC formula, which consists of 22 % nonfat dry milk, 9 % fat, 43.5 % sucrose with total solids of 74.5 %. The order of mixing ingredients is important because nonfat dry milk must be hydrated before sugar is added. The mixing system is critical because high shearing is necessary to hydrate the product in the very viscous mixture of nonfat dry milk and crystalline sugar in a very limited amount of water; the Cowles dissolver is generally used [76], [77]. To expedite the solubilizing of nonfat dry milk, a small excess (ca. 10 %) of water is added and subsequently removed.

The process differs in some details from the procedure outlined in Figure 2. Thus, the dry ingredients are dissolved sequentially in water at 50 – 60 °C, first the nonfat dry milk, and then the sugar. The dry milk must be dissolved completely before the sugar is added. The fat is added last. The blend is filtered, homogenized in a single stage within a pressure range of 3.5 – 6.8 MPa, pasteurized at 80 – 90 °C for 0.5 – 2 min (HTST) or 10 min (batch), flashed into a large vacuum cooler under aseptic conditions, where it is cooled to 25 – 30 °C, seeded with sterile lactose, filled, and canned. Flash cooling removes the excess water [77].

Nondairy Whipped Toppings. Nondairy whipped toppings are prepared as oil-in-water emulsions. They contain the basic ingredients fat, protein, sucrose, corn syrup solids, stabilizer, and emulsifiers. Protein-less topping is an exception; it contains a cellulose derivative in place of protein. These ingredients in liquid toppings are used in the following proportions: 15 – 35 % vegetable fat, 1 – 6 % protein, 6 – 12 % sucrose, 2 – 20 % corn syrup solids, 0.1 – 0.7 % stabilizer, 0.4 – 4.0 % emulsifier, salt, and flavoring and coloring agents and water as needed.

A formulation for a simple nondairy liquid whip topping at a moderate fat level consists of 25.6 % partially hydrogenated coconut oil, 6.8 % sucrose, 2.3 % sodium caseinate, 0.7 % lactic acid ester of glycerol monopalmitate, 0.30 % sodium stearyl-2-lactylate (SSL), 0.12 % Polysorbate 60, 0.073 % sodium alginate, 0.05 % dipotassium hydrogen phosphate, and water to 100 %.

The preparation of the emulsion follows the general procedure illustrated in Figure 2, except that homogenization is carried out in two stages at a pressure of 10.3 MPa in the first stage and 3.4 MPa in the second. After homogenization, quick cooling of the emulsion is important to induce rapid fat crystallization.

These toppings are manufactured and marketed in several different forms. Nondairy liquid toppings are distributed as refrigerated items to food processors such as bakeries, confectioneries, ice cream manufacturers, and restaurants. Distributed initially in small milk cans on a returnable basis, products are now sold in 1-Gallon (3.79-L) gable top paper milk cartons. These toppings are also available in the concentrated and frozen forms.

Other emulsifiers used in formulating nondairy liquid whip toppings are, for example, lecithin, polyglycerol esters of fatty acids, and hydroxypropyl cellulose which is used as a substitute for a protein stabilizer. This cellulose in-

gredient and other related cellulose derivatives are often added to liquid whips that are to be sold in the frozen state.

Well accepted as a whipped topping, *ultrapasteurized aerosol nondairy topping* is packed under pressurized nitrous oxide gas in strong metal containers fitted with a product release valve. When released, the pressurized gas forces the liquid topping through the valve, creating the whip as it expands through the nozzle.

Ultrapasteurization, according to the Public Health Service of the FDA, means that such products must have been thermally processed at or above 138 °C for at least 2 s, either before or after packaging, to produce a product which has an extended shelf life under refrigeration [66]. The ultrapasteurized mix is processed outside the can, cooled to 4 °C, canned, capped, and charged with nitrous oxide gas.

Spray-dried whip topping was developed in the 1950 s and early 1960 s. This form of topping is reconstituted by the consumer with cool water, milk, or skim milk, and mechanically whipped in a few minutes [78], [79]. It has excellent foam stiffness and aeration. The powder product is very stable.

An example of a spray-dried topping is formulated to contain 45 % partially hydrogenated coconut fat or palm kernel fat, 10 % lactic acid ester of monopalmitate, 8 % sodium caseinate, 26 % corn syrup solids (25 DE), 10 % cane sugar, 0.06 % sodium alginate, and 0.94 % moisture. The mixture is adjusted to 45 % total solids or higher, processed as outlined in Figure 2, homogenized in two stages at 10.3 MPa and 3.4 MPa, spray-dried under pressure atomization into a vertical dryer or box dryer, cooled to 7 – 13 °C, tempered (i.e., stored to dissipate the heat of crystallization), dry-blended with additional sugar and flavoring as required, and packaged.

Other lipophilic emulsifiers that perform well in dehydrated whip toppings are propylene glycol monostearate and acetylated monoglycerides.

Protein-less nondairy whip topping is prepared by substituting methyl cellulose, hydroxypropyl methyl cellulose, or hydroxypropyl cellulose for sodium caseinate or soy protein isolate. These ingredients form stable films around fat globules and are acid-stable. These properties have been utilized in developing stable toppings made with acidic fruits [80], [81].

Prewhipped frozen toppings are the most popular form marketed in the United States. Prewhipped by the manufacturer, they are sold in the frozen form and usually packed in round, widemouthed plastic containers. They are ready to be used upon thawing.

One method of manufacture consists of pumping the homogenized emulsion, at a temperature of 2 – 6 °C, through a rotary whipping machine and on to a filling machine. After packaging, the product is quickly frozen to – 29 °C.

Imitation Cheese. Two types of imitation cheese are manufactured in the United States. One is made with casein, vegetable fat, and water. The other is made with skim milk and vegetable fat, and then fermented. The ingredients required are skim milk or nonfat milk, acid or rennet casein (preferably rennet), partially hydrogenated soybean oil, salt, a coloring agent, and disodium hydrogen phosphate. The flavors used are prepared by lipase and protease hydrolysis of cheese trimmings.

Imitation cheese made with casein and vegetable fat is prepared by mixing 27 % casein and 25 % fat with salt, disodium hydrogen phosphate, flavoring, coloring, and water to bring the total to 100 %. The mix is plasticized at a temperature range of 82 – 92 °C in an auger-screw cooker heated by steam injection [82]. After being cooked, the imitation cheese is extruded, and packaged. The pH of imitation cheese is adjusted to 5.6 – 5.8 for improved texture and stringiness when heated.

Imitation cheese made with skim milk and vegetable fat utilizes conventional cheese-making techniques. Imitation milk is prepared by processing skim milk and vegetable fat as described in Figure 2. The milk, at 22 °C, is inoculated with 2 – 5 % lactic starter culture, which is kept at 22 °C until the curd is set at pH 4.6. The curd is then cut, cooked, and finished to the required degree by adding hydrolyzed cheese flavor (→ Cheese, Processed Cheese, and Whey). Imitation cheese produced by fermentation is usually made into processed imitation cheese. Imitation casein vegetable fat cheese is used primarily to manufacture imitation mozzarella.

6. Standards and Quality Control

Sanitary quality standards are regulated by sanitary codes at the state and local levels in the United States. However, the United States has no federal standards of identity for imitation dairy products, except for margarine [83] and mellorine [84], previously considered imitation dairy products but now accepted as standard foods. Analytical procedures for quality control of ingredients and finished imitation dairy products are published by several trade associations [72], [85], [86] and technical societies [87], [88].

7. Nutrition

The use of imitation dairy products should not be viewed as a potential nutritional hazard, unless they are substituted indiscriminately in infant formulas [89]. However, proper labeling and suitable levels of nutritional equivalence should be observed for fluid imitation milk. A nutritional study of imitation milks carried out on monkeys and rats showed no differences in overall appearance, milk intake, and body weight compared to those fed cow's milk [90]. A review of nutritional studies of coconut oil suggests that some negative findings are consequences of abnormal nutrition, rather than inherent defects in coconut oil [91]. Coconut oil and palm kernel oil are the two most important lauric oils. Epidemiological studies of atherosclerosis in populations consuming large amounts of coconut oil are cited to show that coconut oil in a natural diet is not disadvantageous and may even be of advantage.

A nutritional study of severely malnourished infants indicated that "milks based on a high quality ISP (isolated soy protein) appropriately enriched with D,L-methionine are the practical equivalent of the best modified cow's milk products as the main or only source of protein in the diet of rapidly growing infants and small children" [92–96].

A review from 1983 contains recommendations for and against soy protein formulas with regard to infant feeding [97].

8. Toxicology

Most additives used in the food industry in the United States are considered GRAS (generally recognized as safe) by the FDA [98]. The FDA has set maximum usage limits on a few food emulsifiers used in nondairy creamers and nondairy whipped toppings [99]. The highest limited amount of emulsifier is 0.4 wt %; the lowest, 0.3 wt %, based on the finished product. The FAO – WHO has published a list of ADI (acceptable daily intake) values in which 16 emulsifiers are compared [33], [103]. The highest listed ADI value is 50, the lowest is $0-2.5\,\mathrm{mg\,kg^{-1}d^{-1}}$. Other ingredients with ADI imposed limits by FAO – WHO are xanthan gum with a limit of $10\,\mathrm{mg\,kg^{-1}d^{-1}}$, as well as several food colors whose highest limit is $0-25\,\mathrm{mg\,kg^{-1}d^{-1}}$ and lowest is $0-0.1\,\mathrm{mg\,kg^{-1}d^{-1}}$ [100], [101].

9. Economic Aspects

The difference in cost between butterfat and vegetable fat has provided a major incentive for the development of imitation dairy products. For example, in early 1986 butterfat was six times as costly as hydrogenated coconut fat. Sales of imitation dairy products in the United States amounted to about 628×10^6 \$ in supermarket retail sales for 1985. Whipped and frozen toppings and powdered nondairy creamers account for more than three-quarters of these sales [102].

For the period 1980 – 1984 sales of both frozen nondairy toppings and frozen nondairy creamers rose slightly; all other categories declined (no data are available for imitation cheese).

10. References

1. S. F. Riepma: *The Story of Margarine,* Public Affairs Press, Washington, D.C., 1970.
2. CFR, FDA: *TITLE 21,* **section 166.4 to 166.110, Margarine.** Office of Federal Register, Washington, D.C., 1981.
3. J. J. Massiello, *J. Am. Oil Chem. Soc.* **55** (1978) 262 – 265.
4. U.S. Congress: *Filled Cheese Act,* **vol. 29,** Chap. 337, Government Printing Office, Washington, D.C., 1897, pp. 253 – 256.

5. U.S. Congress: *Filled Milk Act,* **vol. 42,** part 1, Chap. 262, Government Printing Office, Washington, D.C., 1923, pp. 1486 – 1487.
6. FDA Announcement, Aug. 2, 1973, *Fed. Reg.* **38,** no. 48, pp. 2.0748 – 2.0749.
7. U.S. Federal Court Action 1972-Milnot vs. Richardson 350 F, Supplement 222. U.S. District Court, Springfield, Ill.
8. Repeal of Filled Cheese Act. United States Statute at Large, vol. 88. part 2, Washington, D.C., 1974, pp. 1466 – 1468.
9. CFR, FDA: *TITLE 21,* **chem. 1, Section 101 3 (c) and 4,** Office of Federal Register, Washington, D.C. 1981.
10. CFR, FDA: *TITLE 21,* **Chan. 1, Section 101.9.** Nutrition of Foods Office of Federal Register, Washington, D.C., 1981.
11. FAO, *Code of Principles concerning Milk and Milk products and Associated Standards,* 7th ed, FAO, Rome 1973.
12. U.S.D.A.: *Handbook No. 8 – 1,* ARS. Item no. 01 – 085. Washington, D.C. 1976.
13. CFR, FDA: *TITLE 21,* 131.143, Office of Federal Register, Washington, D.C. 1981.
14. American Dry Milk Inst., *Inc. Bulletin 916,* Revised Standards of Dry Milk. Chicago, Ill. 1971.
15. S. C. Hugunin, N. S. Ewing: *Dairy Based Ingredients for Food Products,* Dairy Research Inc., Rosemont, Ill., pp. 17, 18.
16. L. L. Muller, *Dairy Sci. Abstr.* **33** (1971) , no. 9, 659.
17. A. van Krevald, Voeding – 30e jaargang no. 5, 1969.
18. E. J. Mann, *J. Soc. Dairy Technol.* **24** (1971) , no. 4, 145 – 153.
19. J. H. Jonas, *J. Milk Food Technol.* **36** (1973) no. 6, 323 – 332.
20. Federal Register (U.S.) July 20, 1968.
21. Ralston Purina Co. St. Louis Mo. Supro 710; Staley Continental Inc., Decatur, Ill., Miro-Pro 111.
22. K. J. Burgess: *Food: Flavoring, Ingredients, Packaging, and Processing,* **vol. 3,** no. 3, United Trade Press House, London, U.K., 1981, pp. 22, 23, 27, 31, 41.
23. O. L. Brekke: *Handbook of Soy Oil Processing & Utilization,* Chap. 19, American Soybean Assoc. & American Oil Chemist's Society 1981, pp. 396 – 398.
24. U.S. Department of Agriculture, adopted from *Agriculture Handbook,* no. 8 – 1: "Composition of Food," Dairy & Egg Products, Raw, Processed, Prepared, Item N001 – 003 (1976).
25. J. R. Ryberg: "Choosing a fat system for a filled or imitation dairy cheese," *Food Prod. Dev.* **2** (Aug. – Sept. 1968) 60, 62, 64, 66.
26. U.S. Department of Agriculture, (1979) adapted from *Agriculture Handbook* no. 8 – 4: "Composition of Foods: Fats and Oils, Raw, Processed, Prepared," contained fractional percentage palmitic, stearic, and archidic acids. Margarine fat prepared from partially hydrogenated soy and cottonseed oils.
27. F. R. Paulicka, *J. Am. Oil Chem. Soc.* **53** (1976) 421 – 424 (Hard butter prepared from partially hydrogenated soybean oil).
28. L. H. Wiederman, *J. Am. Oil Chem. Soc.* **55** (1978) 823 – 829.
29. B. L. Madison, R. C. Hill, *J. Am. Oil Chem. Soc.* **55** (1978) 328 – 331.
30. J. B. Carden, D. F. Meisner, *Baker's Dig.* **44** (1970) no. 2, 58 – 60.
31. Paramount *C. Durkee Product Data Sheet,* Durkee Industrial FOODS, Non Dairy Specialty Fat. Cleveland, Ohio, 1981.
32. Tem-Wip, Non-dairy Special Fat Product Information Fact Sheet, Bunge Edible Oil Corp., Bradley, Ill.
33. N. Krog, J. B. Lauridsen: *Food Emulsions,* Chap. 6 Food Emulsifiers, Marcel Dekker, New York, Basel 1976.
34. N. H. Nash, L. M. Brickman, *J. Am. Oil Chem. Soc.* **49** (2972) 457 – 461.
35. J. B. Lauridsen: *Food Surfactants and Their Structure and Polymorphism,* Am. Oil Chem. Soc. Short Courses, Hawaii, May 10 – 14, 1986.
36. Mono- and diglycerides of fatty acids, Durem 117 Durkee Foods, Div. of SCM Corp., Cleveland, Ohio 4415.
37. Promodan SPV: *Distilled propyleneglycol esters of hydrogenated vegetable fatty acids,* Grinsted Products, Inc., 201 Industrial Parkway, Industrial Airport, Kansas 66 031.
38. Span 60, *Sorbitan Monostearate* NF ICI Americas, Inc., Chemical Div. Wilmington, Del. 19 897.
39. Durlac 100, *Lactic acid esters of monoglycerides,* Durkee Foods Div of SCM Corp. Cleveland, Ohio 4415.
40. SETODAN, *Acetic acid esters of monoglycerides,* Grindsted Products. 38 Edwin Rahrs Vej. DK-8220 Brabrand, Denmark.
41. Panodan, *Diacetyl tartaric acid esters of monoglycerides.* Grindsted Products, Inc., 201 Industrial Parkway, Industrial Airport, Kansas 66 031, United States.

42. Polysorbate 60, *Polyoxyethylene(20) sorbitan monostearate,* ICI Americas Inc., Chemicals Division, Wilmington, Del. 19 897.
43. V. K. Babayan, US 3 637 774, 1963.
44. V. K. Babayan, H. Kaunitz, C. A. Slanetz, *J. Am. Oil Chem. Soc.* **41** (1964) 434 – 437.
45. Caprol ET, *Polyglycerol esters of fatty acids,* Capitol City Products Company, Chemical Specialties Group, Columbus, Ohio 43 216.
46. R. McIntyre, *J. Am. Oil Chem. Soc.* **56** (1979) 835 A – 840 A.
47. G. B. Ansell, J. N. Hawthorne: *Phospholopids,* Elsevier, New York 1964.
48. Central Soy Chem. Div.: *Lecithin,* P.O. Box 1400, Fort Wayne, In., United States.
49. P. Becher: *Emulsions. Theory and Practice,* ACS Monographs series no. 162, Robert E. Kreiger Publ., Huntington, N.J., 1977.
50. N. Krog, *J. Am. Oil Chem. Soc.* **54** (1977) 124 – 131.
51. H. Pancoast, R. Junk: *Handbook of Sugar,* Avi Publishers. Westport, Conn. 1980.
52. T. A. Nickerson: "Lactose," in B. H. Webb, A. H. Johnson, J. A. Alford (eds.): *Fundamentals of Dairy Chemistry,* 2nd ed., Avi Publishing Co., Westport, Conn., 1980.
53. G. A. Towle: *Industrial Gums,* Chap. 5, Academic Press, New York 1973, pp. 83, 114.
54. F. Roll: *Industrial Gums,* Chap. 15, Academic Press, New York 1973, pp. 323 – 337.
55. W. H. McNeely, D. J. Pettitt: *Industrial Gums,* Chap. 4, Academic Press, New York 1973, pp. 49 – 81.
56. A. M. Goldstein, E. N. Alter, J. K. Seaman: Industrial Gums, Chap. 14, Academic Press, New York 1973, pp. 303 – 321.
57. J. B. Batdorf, J. M. Rossman: *Industrial Dums,* Chap. 31, Academic Press, New York 1973, pp. 695 – 729.
58. W. H. McNeely, K. S. Kang: *Industrial Gums,* Chap. 21, Academic Press, New York 1973, pp. 473 – 497.
59. Avicel Application Bulletin Noc-29, copyright 1982. FMC Corporation, Food and Pharmaceutical Products Division, Philadelphia, Penn. 19 103.
60. R. E. Rich, US 2 868 653, 1958 (H. H. Diamond, E. L. Powell).
61. G. K. Greminger, Jr., A. B. Savage: *Industrial Gums,* Chap. 28, Academic Press, New York 1973, pp. 619 – 647.
62. A. J. Desmarais, *Industrial Gums,* Chap. 29, Academic Press, New York 1973, pp. 649 – 672.
63. D. W. Kaufman: *Sodium Chloride,* ACS Monograph Series no. 145, Reinhold Publ., New York 1960.
64. Milk Industry Foundation: *Manual for Milk Plant Operators,* Washington, D.C. 1967.
65. A. W. Farrell: *Food Engineering Systems,* **vol. 1, 2,** Avi Publishing Inc., Westport, Conn. 1976.
66. *Grade A Pasteurized Milk Ordinance Ultrapasteurization,* U.S. Department of Health, Education & Welfare, Public Health Service, FDA, Washington, D.C., 1978.
67. W. Shurtleff, A. Aoyogi: Tofu and Soy Milk Production, *The Book of Tofu,* **vol. 11,** New Age Foods, Study Center, Lafayette, In. 1979, pp. 197, 199, 200.
68. CFR, FDA: *TITLE 21,* 131.110. Standard of Identity, Milk. Office of Federal Register, Washington, D.C., 1981.
69. W. B. Van Arsdel, M. J. Copley, A. J. Morgan: *Food Dehydration,* **vol. 2,** Chap. 15, Dry Milk Products, Avi Publishing Co., Westport, Conn., 1973, pp. 290 – 346.
70. O. F. Hunziker: *Condensed Milk, Powdered Milk,* Published by O. F. Hunziker, La Grange, Ill., 1949.
71. F. G. Kieseker, Proceedings of IDF Seminar, Singapore 1980, Paper 18. Document 142, 1982. General Secretariat, B-1040 Brussels 1980.
72. National Food Processors Assoc: *Laboratory Manual for Food Canners and Processors,* **vol. 1,** 2, Avi Publishing Company, Westport, Conn., 1978.
73. K. Eggers: Proceedings of IDF Seminar, Singapore 1980; Paper 14, Document 142.1982.
74. F. G. Kieseker, P. Southby, *Aust. J. Dairy Technol.* **20** (1965) 136.
75. F. G. Kieseker, P. Southby: *Winter School on Spray Drying.* Aust. Soc. Dairy Tech. Highett, Vict. Aust. General Secretariat, B-1040 Brussels, 1980.
76. D. F. Newstead, *Proceedings of IDF Seminar,* Singapore, Paper 13, Document 142 (1982) ; General Secretariat, B-1040 Brussels.
77. Beatrice Companies, Inc., US 3 126 283, 1964 (P. P. Noznick, R. H. Bundus, I. B. Eggen).
78. Beatrice Companies, Inc., US 3 098 748, 1963 (P. P. Noznick, C. W. Tatter).
79. General Foods Corp., US 3 246 996, 1966 (P. P. Noznick, C. W. Tatter).
80. Methocel Premium Gums, Applications Bulletin, Dow Chemical U.S.A., Special Chemicals Dept. Midland, Mich. 48 674.

81. Klucel – Hydroxypropyl cellulose Bulletin, "FF-314" Bulletin Klucel (hydroxypropyl cellulose), Hercules Inc., Wilmington, Del. 19 899.
82. Damrow Company. Fond Du Lac, Wi. Cheese Cooker.
83. CFR, FDA: CFR, FDA: *TITLE 21,* Section 166.40, 166.110. Office of Federal Register, Washington D.C. 1981.
84. CFR, FDA: *TITLE 21,* Section 135.130. Office of Federal Register, Washington D.C. 1981.
85. APHA. American Public Health Assoc.: *Compendium of Methods of the Microbiological Examination of Foods,* 1750 Broadway, New York, N.Y. 1984.
86. Milk Industry Foundation: *Laboratory Manual,* 910 Seventeenth Street, N.S. Washington, D.C.
87. AOAC: *Official Methods of Analysis of the Association of the Official Analytical Chemicals,* 14th ed., Published by the AOAC, Arlington, Va. 1984.
88. AOCS: *Official and Tentative Methods of the American Oil Chemist Assoc.,* 3rd ed. Published by AOCS, Champaign, Ill., 1974.
89. AAP. American Academy of Pediatrics, *Comm. Nutrition,* **49** (1972) no. 5, 770, 775. (Codex alimentarius commission of FAO/WHO).
90. H. Kaunitz, J. Sawyer, *Med. Ernähr.* **7** (1970) no. 11, 158, 161.
91. H. Kaunitz, *J. Am. Oil. Chem. Soc.* **47** (1970) no. 10, 462 A, 464 A, 482 A, 485 A.
92. G. G. Graham et al., *Dis. Child* **120** (1970) 419 – 423.
93. Pro Sobee, *Isolated Soy Protein Imitation Milk,* Mead Johnson & Co., Evansville, In. 47721.
94. A. Meister: *Biochemistry of the Amino Acids,* 2nd ed., **2 vols.,** Academic Press, New York 1965.
95. Similac *Modified Cow's Milk,* Ross Laboratories, Columbus, Ohio.
96. Casec *Casein,* Research Center, Mead Johnson & Co., Evansville, In. 47 721.
97. American Academy of Pediatrics, Committee on Nutrition, M. Mauer, (chair) 72. 359 – 363 *Pediatrics,* 1983.
98. CFR, FDA: *TITLE 21,* Section 121.01 Gras list., Sections 182, 184, 181, 180, Office of Federal Register, Washington, D.C., 1981.
99. CFR, FDA: *TITLE 21,* Section 172.842, 172.836, 172.838, 172.840, 172.844, 172.846, 172.852, 172.828, 172.828, and 172.856. Office of Federal Register, Washington, D.C., 1981.
100. *Toxicological Evaluation of Certain Food Additives,* 17th & 18th Reports of the Joint FAO/WHO Expert Committee on Food Additives, WHO Technical Report Series No. 539 & 557, Geneva 1974.
101. *Codex Alimentarius* CAC/FAL 1 – 1973 (Suppl. 1); first-series supplement to list of additives, issued by Secretariat of joint FAO/WHO committee, Rome 1974.
102. SAMI Report, Rockefeller Center, Time – Life Bldg., New York 1986.
103. *Food Chemical Codex,* National Academy Press, Washington, D.C.

DDT → Insect Control
Decalin → Naphthalene and Hydronaphthalenes
Decanol → Fatty Alcohols
Defoaming Agents → Foams and Foam Control
Degreasing → Metals, Surface Treatment
Dehydrogenation → Hydrogenation and Dehydrogenation

Dental Materials

Robert G. Craig, formerly The University of Michigan, Ann Arbor, MI, United States (Chap. 1)

Dieter Welker, formerly Klinikum der Friedrich-Schiller-Universität Jena, Jena, Federal Republic of Germany (Chap. 1 and 6)

Josef Rothaut, Wieland Dental & Technik GmbH & Co. KG, Pforzheim, Federal Republic of Germany (Chap. 2)

Klaus Georg Krumbholz, DeguDent GmbH, Hanau-Wolfgang, Federal Republic of Germany (Chap. 3)

Klaus-Peter Stefan, ESPE Dental AG, Seefeld, Federal Republic of Germany (Chap. 4)

Klaus Dermann, formerly Degussa-Hüls AG, Hanau-Wolfgang, Federal Republic of Germany (Chap. 5)

Hans-Joachim Rehberg, formerly Bayer AG, Leverkusen, Germany (Chap. 6, 7 and 8)

Gertraute Franz, formerly Universitätskrankenhaus Eppendorf, Hamburg, Federal Republic of Germany (Chap. 9)

Klaus Martin Lehmann, formerly Klinikum der Philipps-Universität Marburg, Marburg, Federal Republic of Germany (Chap. 9)

Matthias Borchert, BIOS Dental GmbH, Bohmte, Federal Republic of Germany (Chap. 10)

1. Introduction 2
1.1. Direct Materials 2
1.2. Indirect Materials 4
1.3. Auxiliary Materials 6
2. Metals 7
2.1. Introduction 7
2.2. Metallic Materials for Fillings 8
2.2.1. Pure Gold 8
2.2.2. Silver Amalgams 8
2.2.3. Copper Amalgams 10
2.2.4. Gallium Alloys 10
2.3. Metallic Materials for Prosthetic Restorations 10
2.3.1. Precious Metal Casting Alloys 11
2.3.2. Nonprecious Metal Casting Alloys . . 16
2.3.3. Casting Alloys for Removable Dentures 17
2.4. Brazing Alloys 18
2.5. Metallic Materials for Implants . . . 19
2.6. Metals and Alloys for Other Purposes 20
3. Dental Ceramic Materials 20
3.1. Ceramic Teeth 20
3.1.1. Raw Materials 21
3.1.2. Fabrication 21
3.1.3. Quality Requirements 22
3.2. Dental Ceramic Fused to Metal Restorative Materials 22
3.2.1. Thermal Expansion 23
3.2.2. Raw Materials 23
3.2.3. Fabrication 23
3.2.4. Quality Requirements 24
3.2.5. Ceramic Fused to Titanium 24
3.3. Dental Ceramics for Fixed Restorations 24
3.3.1. Jacket Crowns 24
3.3.2. Newer Developments 25
4. Cements 27
4.1. Cements Setting by Neutralization Reaction 27
4.1.1. Zinc Phosphate Cements 27
4.1.2. Silicate Cements 28
4.1.3. Silicophosphate Cements 28
4.1.4. Zinc Polycarboxylate Cements 28
4.1.5. Glass Ionomers 28
4.1.6. Resin-Modified Glass Ionomers 30
4.1.7. Polyacid-Modified Composites 30
4.2. Cements Based on Complex Formation 30
4.2.1. Zinc Oxide – Eugenol Cements 30
4.2.2. Calcium Hydroxide Cement 31
4.2.3. Other Chelating Cements 31
4.3. Standards 31
5. Synthetic Resins 31
5.1. Denture Base Resins 31
5.1.1. Heat-Cured Polymers 32
5.1.2. Cold-Cured Polymers 32
5.1.3. Light-Cured Polymers 33
5.1.4. Microwave-Cured Polymers 33
5.1.5. Other Denture Resins 33

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_251.pub2

5.2. Acrylic Teeth 34
5.3. Crown and Bridge Veneer Resins . 34
5.3.1. Heat-Cured Resins 34
5.3.2. Light-Cured Resins 34
5.4. Resin – Metal Bonding 35
5.5. Resins for Orthodontics 35
5.6. Restorative Resins 35
5.6.1. Paste – Paste Systems 36
5.6.2. Light-Cured Polymers 36
5.6.3. Ormocer-Based Filling Materials . . . 37
5.6.4. Sealant Resins 37
5.6.5. Resin – Tooth Bonding 37
5.6.6. Toxicology 38
5.7. Standards 38
6. Impression Materials 38
6.1. Elastic Materials 39
6.1.1. Alginates 39
6.1.2. Elastomers 39
6.2. Rigid Materials 41
6.2.1. Plaster of Paris 41
6.2.2. Zinc Oxide – Eugenol Pastes 42
6.2.3. Resins 42
6.2.4. Thermoplastic Impression Materials . 42
7. Dental Waxes 42
8. Other Model and Die Materials . . . 43
9. Dental Gypsum Products 44
10. Dental Investments 47
11. References 49

1. Introduction

Dental materials are used for the replacement of destroyed or lost structures (hard tooth substance, teeth, soft tissues of the mouth, tissues of the head) and for the restoration of disturbed functions (masticatory, phonetic, esthetic, physiognomic) of the orofacial organ. They are involved in the morphological, physiological, and psychological reconstitution of the patient (curative effect). At the same time they prevent structural decay and functional failure (preventive effect). Their use takes into account biological and technical aspects. This is documented by the large number of international standards (ISO). There is a requirement for chemical biocompatibility (lack of toxicity, allergenicity, mutagenicity, cancerogenicity), functional biocompatibility (strength, environmental resistance), and technical functionality (handling, ergonomics, economy). Some dental materials are incorporated permanently or temporarily in the orofacial organ; legally, they are similar to medicinal products. Other materials are used for the production of tooth restorations and do not come into contact with the patient's tissues. From the standpoint of industrial hygiene they must be completely harmless to the user (dentist, dental technician, dental assistant).

There are may uses for dental materials:

- Reconstruction of decayed tooth structure (enamel, dentin) by fillings and crowns
- Replacement of missing teeth by partial or full dentures
- Surgical prostheses and epitheses in orofacial prosthetics

Some dental materials are primarily used for preventive purposes (sealants for developmental pits and grooves in the biting surfaces of teeth and athletic mouth guards to reduce forces applied to teeth during contact sports).

According to their handling dental material can be grouped in the following way:

- Direct restorative materials, which are processed in the mouth
- Indirect restorative materials, which are fabricated outside the mouth

1.1. Direct Materials

Pure Gold. Gold foil can be formed into solid gold of nearly theoretical density by mechanically condensing it directly into a retentive cavity preparation. By means of small-diameter condenser tips, the gold is work-hardened to have properties comparable to a 90 % gold – 10 % copper cast alloy. Electrolytically produced powdered gold containing 0.5 % calcium may be substituted for gold foil. These restorations have low mechanical properties and are restricted to small non-stress-bearing applications in posterior teeth because of aesthetics.

Silver Amalgam. Silver amalgam is supplied as a powder consisting mainly of silver, with a balancing amount of tin, copper, and minor elements. These amalgams are classified as low-copper (< 6 %) and high-copper (13 – 30 %) types. The powder is mixed with mercury to form a plastic mass which is condensed into a retentive cavity preparation. The surface

is carved shortly after placement; depending on the product, the surface may be finished and polished immediately or at a second appointment. Amalgam is used in posterior teeth because of its gray appearance and is successful in stress-bearing restorations. In badly decayed teeth, amalgam may be retained by threaded metallic pins placed into dentin, which extend into the amalgam. It can be used to build up the crown of a tooth (a core) on which a cast restoration is placed.

Polymers and Composites. Difunctional oligomers and monomers, usually mixed with fillers, are used to fill developmental pits and fissures in the occlusal surfaces of enamel to prevent caries. Before placement, the enamel is etched with acid, rinsed, and dried. Etching provides for mechanical retention of the sealant. The sealants are formulated to be polymerized in situ by a chemical-activated or light-activated free-radical mechanism. The treatment is most successful on the teeth of children soon after eruption.

Anterior teeth with small to moderate fractures or decay can be restored with composites. They contain an inorganic filler of various particle sizes, as well as difunctional acrylate oligomers and monomers. Composites are today only rarely available as two pastes for the chemically activated type and mostly as a single paste for the light-activated type. After cavity preparation, an adhesive is applied on enamel and dentin and polymerized. The peripheral enamel is etched with acid, rinsed, and dried. Equal volumes of the chemically activated type are mixed, syringed or packed into the preparation, and allowed to polymerize. For the light-activated type, the paste is placed into the cavity preparation and polymerized by exposure to light. Setting occurs promptly, and final finishing to proper contour is done. When the proper shade is used, the boundary between the restoration and the tooth is difficult to detect.

Cervical caries or eroded areas of teeth are generally restored with glass ionomers. They consist of a powder and a liquid (an acid-soluble glass and an aqueous solution of polyacrylic acids). The exposed dentin is cleaned, and an adhesive is applied. Any peripheral enamel is etched with acid. The components are spatulated to a thick paste and packed into the prepared area. Leached metal ions from the powder and the polymer acid react; a similar reaction may occur between the calcium from the tooth and the polymer acid to form an adhesive bond. Finishing is delayed for 24 h because of the slow completion of the reaction. This method is most successful for cervical restoration; its functions are to restore the contour of the tooth, thus increasing periodontal health, and to reduce the sensitivity of the tooth.

Composites have been used in posterior teeth, but the lack of wear resistance has restricted their application to situations in which aesthetics is essential. Composites with higher filler content and wear resistance have been developed, and restoration of posterior teeth with composites has increased.

Agents are available to bond enamel and dentin with composites. The bonding agents are organic compounds with a variety of functional groups. They are applied to the cavity preparation before the composite, but the adhesive bond is weak.

Composites are used as cores in severely broken-down teeth and, in combination with the acid-etch technique, as veneers on severely discolored or deformed anterior teeth.

Cements. Cements are classed as *luting cements* (attaching a cast gold restoration to a tooth) or *base cements* (support for an amalgam restoration). Chemically, they are classified as (1) zinc phosphate, (2) zinc oxide – eugenol and modifications, (3) zinc polyacrylate, (4) glass ionomer, or (5) resin and composite cements. The applications of cements are listed in Table 1.

Zinc phosphate cement is supplied as a powder (mainly ZnO) and a liquid (mainly H_3PO_4 in water). The ratio of powder to liquid is higher for a base than for a luting cement. The acid – base reaction is exothermic, and the cement is mixed on a cool glass slab to control the temperature. The mix is then placed on the restoration and tooth, and the excess is forced out when the restoration is seated. A base cement is stiffer and is packed into the bottom of the cavity preparation and contoured. It is acidic when placed, and the pulp is protected with a varnish on the dentin. Cements have low thermal conductivity and high stiffness and strength; they provide thermal insulation for the pulp and mechanical support for

the restoration. Zinc phosphate cement can also serve as a temporary restoration while a permanent restoration is being made.

Table 1. Applications of cement types

Cement type	Applications
Zinc phosphate	retention of restorations and orthodontic bands; high-strength bases; temporary restorations
Zinc oxide – eugenol	temporary and permanent retention of restorations; low- and high-strength bases; temporary restorations; root canal sealers; gingival surgical tissue pack
Zinc polyacrylate	retention of restorations and orthodontic bands; temporary restorations; high-strength bases
Glass ionomer	retention of restorations
Resin and composite	retention of crowns, bridges, and orthodontic brackets; retention of acid-etched bridges

Zinc oxide – eugenol (ZOE) cements have a variety of applications (Table 1). Unmodified ZOE consists of a powder or paste (mainly ZnO) and a liquid or second paste (mainly eugenol). No significant heat is produced during the mixing or the formation of the chelate, zinc eugenolate. Unmodified ZOE has low strength and is used in a thin layer as a liner, as a root canal sealer (to fill a root when the pulp has been extirpated), or as a surgical pack for areas around the teeth where gingival tissue has been removed for periodontal reasons. The cement has an obtundent effect on the pulp.

Polymers added to ZOE improve its strength and allow it to be used for temporary restorations or for the temporary retention of crowns and bridges. When 2-ethoxybenzoic acid is substituted for part of the eugenol and alumina is added, a high-strength cement results that can be used for permanent cementation.

Zinc polyacrylate cements are usually powders (mainly ZnO) and liquids (mainly aqueous solutions of copolymers of acrylic acid). They are spatulated, with thin mixes used to retain restorations or orthodontic bands, and thick mixes used as bases. Although the mixes are acidic, they have a much smaller effect on the pulp than phosphoric acid.

Glass ionomer cements are powders (acid-soluble glass containing high levels of fluorine) and liquids (aqueous solution of acrylic acid copolymers). They are spatulated, and the cementing consistency is used to retain cast metal restorations. The cement sets promptly, but the chemical reaction is completed slowly and the cement margin should be protected with varnish from saliva for a day.

Resin cements are powders (acrylic polymers and sometimes fillers) and liquids (acrylic monomers), whereas *composite cements* are usually two pastes containing diacrylate oligomers, diacrylate monomers, fillers, and polymerization initiator systems. They are used for the retention of crowns and bridges. In combination with the etching of enamel, these cements are used to attach orthodontic brackets to teeth without bands. The acrylic cements are better with plastic brackets, whereas the resin composites are better for metal brackets. The latter are also used to attach bridges to acid-etched enamel and acid-etched base metals.

Liners. When dental decay in dentin extends near or to the pulp, a liner is needed. If a thin layer of sound dentin exists, a paste – paste zinc oxide – eugenol liner is used as an obtundent and to protect the pulp. If decay extends to the pulp, a calcium hydroxide liner is needed to stimulate the deposition of reparative dentin in the area of exposure. The calcium hydroxide materials are usually supplied as two pastes which are mixed together and set by the reaction of a disalicylate and calcium. Liners are used in thin layers, have very low strengths, and have therapeutic value. High-strength bases are placed on top of them to give mechanical support.

Varnish. Dilute solutions of natural or synthetic polymers in organic solvents are used as varnishes. They form a thin plastic film over freshly cut dentin and reduce the diffusion of material components into the tubules to the pulp.

1.2. Indirect Materials

Cast Restorations. Alloy classes are gold – copper, palladium – silver, nickel – chromium, cobalt – chromium, and cobalt – chromium – nickel. They are fabricated by the lost-wax method: a wax or plastic pattern of the restoration is placed in a refractory investment (see Sec-

tion Auxiliary Materials). After the investment has hardened, it is heated to remove the wax and to expand the mold. Molten alloy is forced into the space occupied by the pattern, usually by centrifugal force. The investment and metal are then cooled, and the restoration is finished.

Gold casting alloys are of types 1 to 4 with increasing copper and other non-noble metal content. Type 1 is soft and used for non-stress-bearing restorations; type 2 is stronger and used for simple posterior stress-bearing restorations; type 3 is still harder and stronger, and is used for complex posterior crown and bridge restorations; and type 4, the strongest, can be used in removable partial dentures. Types 3 and 4 contain enough copper for their as-cast hardness to be increased by heat treatment.

The price of gold has increased the use of palladium – silver, cobalt – chromium, and nickel – chromium alloys for crowns and bridges. They are generally harder and stronger than type 3 gold alloys. They have higher melting points than gold alloys and must be cast in high-temperature investments.

Removable partial denture frameworks are almost always cast from nickel – chromium, cobalt – chromium or cobalt – chromium – nickel alloys because of their lower cost and higher stiffness (about twice that of gold alloys). Partial denture frameworks are so large that free-standing patterns distort during investing. Patterns are formed on refractory investment models of the upper or lower teeth. The pattern and the model are then invested to produce the mold used in the casting process. Plastic base and artificial teeth (Section Auxiliary Materials) are attached to the framework to produce the denture. These alloys have higher melting points and hardness, and lower density, than gold alloys and are more difficult to finish.

Wrought Metals. Wrought wires of 18 – 8 stainless steel, cobalt – chromium – nickel, nickel – titanium, and titanium – molybdenum are used in orthodontics in biomechanical devices to move teeth by the application of forces. After treatment, these appliances are removed. The stiffness of 18 – 8 stainless steel and cobalt – chromium – nickel is the highest, followed by titanium – molybdenum and then nickel – titanium. Fabrication of the appliances often involves soldering or spot-welding, which must be done carefully so that minimal damage to the wrought metal occurs.

Stainless steel is also employed in the manufacture of small tapered files and reamers used to clean out the root canals of infected teeth during endodontic treatment.

Preformed wrought shell crowns of stainless steel, nickel – chromium, or tin- and aluminum-based alloys are used for semipermanent and temporary restoration of children's primary teeth. They are supplied in a multitude of sizes for different teeth. The best size is selected, trimmed, and adapted to the correct shape. It is then cemented with a permanent or temporary cement.

Feldspathic, aluminous, and *crystallizable* glass ceramics are used for the preparation of full anterior crowns. In the first two, the crowns are built up on a die of the prepared tooth covered with platinum foil. Ceramic powder is mixed with water, and the slurry is painted on the matrix. Water is removed by vibration, blotting, baking, and firing. The crown is built up initially with more opaque ceramics; then more translucent ceramics are used toward the surface. Characterization results in a restoration that matches the appearance of natural teeth. Aluminous ceramics have alumina added to the feldspar to improve resistance to fracture.

Ceramic crowns are retained with a dental cement. Because they are brittle and lack high strength, ceramic crowns are used mainly on anterior teeth where appearance is primary.

Ceramics are used to manufacture artificial teeth for dentures. These teeth are held in position by acrylic polymers and by mechanical retention of pins in anterior, and diatoric holes in posterior, teeth. The teeth are available in a variety of shades, sizes, types, and shapes.

Ceramics and Metal – Ceramic Combinations. A thin metal coping is cast to fit the tooth. Special ceramics that bond to the metal are applied to the outside surface. The completed crown has a strong metallic substructure covered with an aesthetic outer layer. These combinations have superior strength compared to ceramic restorations and can be used for posterior as well as anterior crowns and bridges.

Metals that have been used are gold-, nickel-, and palladium-based alloys. Feldspathic ceramics with matching thermal expansion properties

are essential. They are also retained in the mouth with dental cements.

Full ceramic crowns and small bridges can be made using press ceramics. In this process a ceramic pellet is heated to a soft stage and pressed into a form inside of an investment material similar to casting a metal crown. The material is a glass ceramic reinforced by leucite crystals.

Another way to produce full ceramic crowns and small bridges is to use aluminum oxide or an aluminum oxide/zirconium oxide ceramic to build up the crown on a refractory die, loosely sinter it so that it does not shrink, and infiltrate it with a glass. For a more toothlike appearance it must be veneered with a veneering ceramic like a metal framework.

CAD/CAM techniques can also be used to prepare full ceramic restorations. When this technique is used to prepare frameworks from aluminum oxide or yttrium-stabilized zirconium oxide, even wide-span bridges can be made out of full ceramic. The framework must be veneered with a veneering ceramic which has a thermal expansion coefficient that fits to the framework. Since milling of a framework out of densely sintered zirconium oxide ceramic is time-consuming, a computer technique is used to precalculate sinter shrinkage and produce an enlarged framework out of presintered material. The enlarged framework is then dense sintered to a fitting framework.

Ceramics are today only rarely used to manufacture artificial teeth for dentures. These teeth are held in position by acrylic polymers and by mechanical retention of pins in anterior, and diatoric holes in posterior, teeth. The teeth are available in a variety of shades, sizes, types, and shapes.

Polymers. Acrylic polymers and copolymers are used for partial and complete denture bases, soft and hard denture liners, artificial denture teeth, crown and bridge facings, and temporary materials. They are usually processed by a dough-molding technique, and free-radical polymerization is initiated either by heat or by a chemical. Shrinkage is controlled by mixing three parts of polymer and one part of monomer (usually methyl methacrylate) to form a dough. The mixture is packed or pressed into a sectional mold formed by the lost-wax technique. The mold is formed by investing a denture where the artificial teeth are supported by a wax denture base. Pressure is maintained on the mold until the monomer has polymerized. The flask is then removed from the appliance and the plastic base is finished. A pour processing technique is sometimes used, in which a mixture of polymer and monomer is poured into the sectional mold.

The tissue-bearing surface of a denture may be made of a soft polymer such as plasticized acrylic or a silicone rubber for patients who have special problems wearing dentures.

The facial surfaces of anterior metallic crown copings may be coated with tooth-colored acrylics for aesthetic reasons. The polymer is mixed or delivered as a paste with diacrylate monomers having a low vapor pressure. They can be processed in an oven or cured by light or microwaves without investing because of the low volatility of the monomers. Acrylic denture teeth are also manufactured by means of dough molding. They are available in a wide range of shades, types, and geometries. They are softer and less wear resistant than ceramic teeth but can be chemically bonded to the acrylic denture base. Temporary restorations are made directly in the mouth from polymer – monomer mixes, with the aid of an impression. They are then retained with a temporary cement until the permanent restoration is made.

Athletic mouth protectors are made by forming in a vacuum, a sheet (3 mm) of polyethylene – poly(vinyl acetate) copolymer over a model of the athlete's upper teeth. After cooling, the excess is cut away with scissors and the edges are flame-polished. When worn, these protectors nearly eliminate oral injuries in contact sports.

1.3. Auxiliary Materials

Impression Materials. Since diagnostic and therapeutic measures in the oral cavity are restricted, a working model must be created outside the mouth. This is done by physically transferring information from the patient to the dental laboratory by a two-step inversion process:

1) Impression taking: a plastic material is brought into contact with the object to be reproduced and removed as an impression (negative) after hardening by cooling or chemical reaction
2) Model production: the impression is filled with a die material to give a model (positive)

The model must correspond to the shape and dimensions of the morphological situation in the mouth and, if necessary, give an instantaneous picture of functional sequences. Impression and die materials form an integrated whole. For the application of the impression material to the object to be reproduced, rigid impression trays are necessary. They are available prefabricated from metal or resin or custom made in the dental laboratory from light-polymerizable filler-containing polymers (composite materials).

Models and Dies. Models and dies are usually made from one of three types of gypsum material and, less frequently, from plastics (usually epoxy) and metals (electroformed silver or copper). Gypsum materials are compatible with all impression materials, whereas epoxies are compatible with some rubber impressions. Dental compound and addition silicones are compatible with copper forming, whereas most rubbers are compatible with silver forming.

Plaster, stone, and *improved stone gypsum* products are supplied as powders that are mixed with water to a heavy slurry which is vibrated into impressions. The strength, accuracy, and cost increase from plaster to improved stone.

Epoxy materials are supplied as an epoxide paste and a liquid amine catalyst. The fluid mix is poured into the rubber impression. They set slowly compared to gypsum materials.

Electroforming of rubber impressions is done by painting the impression with finely divided metal powder; this surface is then made the cathode in an electroplating system. Pure copper or silver is used as the anode, and electroforming is carried out by using the appropriate electrolyte and direct current (d.c.). Plating is continued until more than 1 mm has been deposited; the remaining space is then filled with acrylic plastic or improved stone.

Pattern Materials. Patterns to be used in the lost-wax technique are most often made from wax and occasionally from acrylic plastics, supplied as a powder and a liquid. Inlay, crown, and bridge patterns are made from harder waxes which are built up in increments on the model or die. Accurate wax patterns are possible when the thermal expansion and stress relaxation properties of the wax are taken into consideration. Denture bases are prepared from softer waxes, which keep the artificial teeth in the correct relationship before the base is invested and processed in acrylic plastic.

Dental Casting. The lost-wax technique in combination with centrifugal casting is the most common way to make dental castings. The free-standing wax pattern or the wax pattern on a refractory investment model is invested in a ring by using gypsum-, phosphate-, or silicate-bonded investment. The latter two are used with high-melting alloys. In addition to the binder, the investment consists of silica materials. It is mixed with water or special liquids and poured around the pattern in the ring. After setting, the wax or plastic pattern is burned out (480 – 980 °C, depending on the technique). The alloy is melted and then cast into the mold. The shrinkage of the wax pattern during mold formation and of the alloy on cooling is compensated for by careful control of the setting and thermal expansion of the investment.

2. Metals

2.1. Introduction

Metals and alloys are widely used materials in modern dentistry. In restorative dentistry metallic materials are used for fillings for periodontal and endodontic treatment; prosthetic dentistry uses metals for crowns and bridges, partial dentures and implant suprastructures; and for orthodontic treatment, the elastic and shape-memory effects of metals and alloys are utilized. Finally in dental implantology, titanium metal is almost exclusively used for its osseointegration capabilities.

Depending on the specific application, the required physical properties and processing characteristics vary widely. For instance, an amalgam filling material must first be a plastic mass, which allows the dentist to pack it into the dental cavity; it must then harden so that the filling can be contoured with a sharp instrument shortly after placement; finally, it must harden rapidly, so the patient can chew an hour or two after the filling has been placed. Root canal posts and implants must be made of rigid, strong materials with a high elastic limit. They are mostly prefabricated because of requirements for precision, surface properties, and strength. Other requirements for certain applications (indications) may

be a certain coefficient of thermal expansion or melting range if the metal framework is to be veneered with a tooth-colored ceramic material.

General requirements for all dental materials result from the requirements to be met by a finished dental restoration in the oral cavity and from interactions with the biological and chemical processes of the patient. The quality of the restoration, which ultimately guarantees long-term success in the mouth, is determined not only by the material but also by the quality of its dental and technical processing. These three areas — medicine, processing technology, and materials science — cannot be regarded independently of each other.

The most important medical requirements on dental alloys are:

- Corrosion and tarnish resistance in the mouth
- Biocompatibility
- Esthetics

No material should be introduced into the oral cavity that can corrode under the sometimes-aggressive conditions that prevail there, that is, dissolve with the formation of considerable amounts of metal ions. Tarnished layers, which can develop by the reaction of metal ions with constituents of the saliva, are also undesirable because they lead to esthetically unsatisfactory results.

The requirement for a good biocompatibility of a material is closely related with the requirement for corrosion resistance. This means that materials which come into contact with the body of the processor or of the patient must not cause any harm to health. When a new alloy is formulated, a high corrosion resistance must be ensured, which in addition to the composition can be improved further by a homogeneous, fine-grain microstructure. An appropriate processing procedure in the laboratory also plays an important role here. However, since even highly corrosion resistant alloys release minimal amounts of ions into solution, elements that have a proven high allergenic or toxic potential, such as Cd, Be, Pb, and Ni, must not be used.

2.2. Metallic Materials for Fillings

2.2.1. Pure Gold

In the past, the filling of tooth cavities by cold-weldable pure gold was rather popular. Today this technique is less important, because it is time consuming and inconvenient to the patient.

Gold for fillings is available in different shapes: gold foil of 1 μm thickness or rolled into cylinders (Gold Cylinder, Williams Gold Ltd.); electrolytically deposited mat gold in strips partially wrapped in gold foil; and gold powder encased in gold foil (Williams Gold Ltd.); partially sintered chemically deposited gold powder [12, 13]. One type contains a small addition of calcium to improve work hardening (Electraloy RV, Williams Gold Ltd.). The pieces must be annealed and allowed to cool; they are then inserted immediately into the cavity and condensed by malleting with hand instruments or electric condensers [13]. Direct plastic gold fillings are advisable only for small cavities on the occlusal surfaces or near the neck of the tooth. Larger fillings with two or more surfaces must be made by the inlay casting technique.

2.2.2. Silver Amalgams

Amalgams are alloys of mercury with other metals. Silver amalgams were introduced to dentistry at the beginning of the last century. The dentist mixes an alloy powder with mercury to obtain a plastic mass. This plastic amalgam is then packed into the prepared tooth cavity. It begins to harden after a few minutes; the surface can be shaped by carving after 10 min or less; and after 1 – 2 d, the amalgam is completely hardened, and can be polished.

Properties. Today many different types of alloys are available for amalgams. They consist of finely divided Ag – Sn – Cu alloys, sometimes with small additions of zinc, mercury, indium, or noble metals. They are distinguished by the following criteria

Silver content:	>65 % or 40 – 60 %
Copper content:	10 – 30 % or <6 %
Grain size:	<150 μm or <60 μm
Shape:	cut filings or atomized spherical particles
Number of alloys:	single alloy or particles from different alloys

Until around 1970, only alloys with high silver content and low copper content were used. These conventional alloys mainly consist of the intermetallic phase Ag_3Sn (γ), which reacts with mercury to form two new intermetallic phases (γ_1 and γ_2):

$$Ag_3Sn\,(\gamma) + Hg \rightarrow Ag_3Hg_4\,(\gamma_1) + Sn_8Hg\,(\gamma_2) + Ag_3Sn$$

An excess of Ag_3Sn remains in the hardened amalgam; its γ_2 phase is the weakest and most prone to corrosion. Corrosion products are tin oxides and tin oxide chlorides. The mercury from the corroding γ_2 phase diffuses back into the amalgam and causes mercuroscopic expansion [14] which damages the margins of the filling.

Since 1965, alloys with higher copper content have been developed. Instead of Sn_8Hg, the mercury-free phase Cu_6Sn_5 (η') is formed, which is less prone to corrosion; it has the additional advantage of making the matrix phase Ag_3Hg_4 more creep resistant.

The first amalgams that were free of the γ_2 phase were the dispersion-type alloys, which contain about 30 % of spheres (< 35 µm) of the Ag – Cu eutectic (28 % Cu) in addition to the conventional Ag_3Sn filings [15]. The initial reaction with mercury is similar to that of the conventional alloys, which is followed by

$$Sn_8Hg + Ag/Cu \rightarrow Cu_6Sn_5 + Ag_3Hg_4$$

This diffusion-controlled reaction takes place within 1 – 7 d.

Atomized alloys consisting of Ag_3Sn and Cu_3Sn with an extremely fine duplex structure were later developed to eliminate the γ_2 phase in dental amalgams. The silver content is 40 – 60 %, and the copper content 13 – 29 % [16, 17]. The reaction with mercury is:

$$Ag_3Sn/Cu_3Sn + Hg \rightarrow Ag_3Hg_4 + Cu_6Sn_5$$

An excess of unchanged particles remains in the completely hardened amalgam. The non-gamma-2 amalgams with higher copper content yield much more stable margins and brighter polished fillings than the conventional low-copper amalgams.

Some of the cut alloys are preamalgamated to accelerate the wetting and reaction with mercury. For this purpose, the alloy powder is treated with a mercury salt solution, which precipitates mercury at the surface of the particles. Zinc is also added as a deoxidizer; it slightly increases the strength of the amalgam but, if contaminated with moisture, has the disadvantage of producing hydrogen, which causes delayed expansion or blisters.

Production. Filings are produced by cutting annealed ingots of alloy (10 – 20 kg) in a machine with a rotating milling cylinder and water cooling. They are ground, sieved, annealed under protective gas, etched with acid or preamalgamated (see above), rinsed, and dried (Fig. 1).

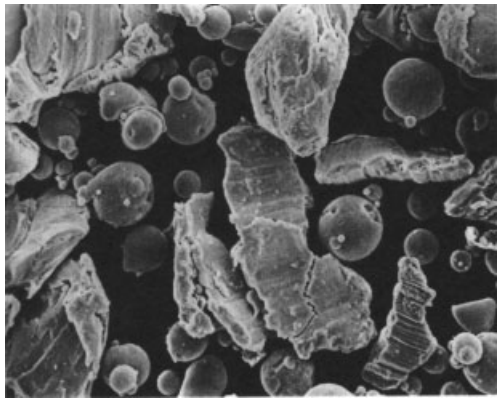

Figure 1. Remission electron microscopy (REM) photograph of Luxalloy-non-gamma 2

Spherical alloy particles are produced directly from the melt by pouring it through a hole in the bottom of the crucible and atomizing the melt stream with high-pressure nitrogen. The atomized droplets solidify while falling or are quenched in water, forming small spheres or irregular particles. The powder is sieved and annealed in a reducing atmosphere or etched with acid.

Cut filings or atomized particles are either used as a single-alloy powder or mixed according to different manufacturing prescriptions. By annealing, the rate of reaction with mercury (setting time, working time for the dentist) and the setting expansion of the alloy can be influenced and optimized.

Standards. Amalgams were the first dental materials to be standardized. The following international standards exist: ISO 24234 (Den-

tistry – Mercury and Alloys for Dental Amalgam).

Handling. In earlier times, the alloy powder and mercury were mixed in a glass mortar with pestle. Today, electric mixing devices produce the amalgam paste in 6 – 30 s. The recommended ratio of mercury to alloy depends on the type of particles. The range is 0.9 – 1.2:1, with less mercury used for spherical powders and more for filings. An excess is expressed by the dentist during compression into the cavity. Ratio, mixing frequency, and mixing time influence the properties of the amalgam before and after hardening. A low ratio causes a dry amalgam paste and porosities in the filling. An extended mixing time may change the setting expansion into a contraction.

Toxicology and Occupational Health. The mercury release from amalgam into water or saliva was found to be very low: 2.4 – 3.8 $\mu g/cm^2$ during the first day from fresh amalgam and 0.1 – 0.3 $\mu g/cm^3$ per day from completely hardened amalgam. Tooth brushing twice a day increases this amount to 0.5 – 0.8 $\mu g/cm^2$ per day. The daily uptake of mercury from normal food is 7 – 12 µg, and can be as high as 40 – 50 µg with one meal of fish [18]. The daily dose from amalgam fillings is below 10 µg [80].

Spilled mercury and amalgam waste can be a source of mercury vapor and therefore hazardous for the dentist and his staff. Spillage must be avoided by using modern dispensing and mixing devices, and by keeping wastes in closed boxes under thiosulfate solution. By this means, the mercury concentration in dental offices can be kept below 10 $\mu g/m^3$. The TLV and MAK of mercury are 50 and 100 $\mu g/m^3$, respectively.

In the mouth, a protective layer of saliva suppresses the evaporation of mercury. The uptake of mercury by breathing through the mouth was found to be less than 0.1 $\mu g/m^3$ for an amalgam surface of 1 cm^2 [19]. Contact between amalgam fillings and gold restorations in the mouth should be avoided, because the different potentials of these materials can lead to the following disadvantages: elevated corrosion, discoloration, and metallic taste. Allergic reactions to mercury in amalgam fillings are observed only in very rare cases.

2.2.3. Copper Amalgams

In former times, copper amalgams were used in childrens' temporary teeth. They were supplied in the form of tablets with about 35 % copper and 65 % mercury. They had to be plasticized by heating above 128 °C, and they hardened on cooling to form the phase Cu_7Hg_6 [20]. The heating process is a dangerous source of mercury vapor. The corrosion resistance of copper amalgams is poor.

2.2.4. Gallium Alloys

Gallium has a melting temperature of 30 °C. A eutectic alloy with indium and tin is liquid at room temperature and can be used to form a paste similar to the amalgam paste. When it is mixed with an alloy similar in composition to an amalgam alloy with silver, copper and tin and in one case 9 % palladium, it hardens in the same way. Since the vapor pressure of liquid gallium at mouth temperature is close to that of solid iron at this temperature, evaporation of gallium is of no concern.

The problem with this type of filling material is that gallium is not as noble as mercury. This means corrosion rates are higher than those for amalgam. The alloy tends to undergo dark discoloration in the mouth and the corrosion rate of gallium from this type of alloy is two orders of magnitude higher than that of mercury from amalgam. Therefore biocompatibility of this type of alloy is by no means better than that of amalgam.

2.3. Metallic Materials for Prosthetic Restorations

The wide span of qualities required in materials for restorative devices has led to a variety of metallic dental materials. Castings are used for rebuilding partly damaged teeth (inlays, onlays, crowns) and for replacing missing teeth (bridges, partial dentures). Cast fillings, crowns, and bridges are permanently fixed to the tooth or abutment with a dental cement; removable partial dentures are usually held in position by clasps or telescoping parts attached to the adjacent teeth. Prefabricated precision attachments

are also used as interlocking devices for the reversible fixation of partial dentures. Brazing is the most commonly used technique when metallic parts of a dental restoration must be joined, but other techniques, such as laser welding, are becoming increasingly popular.

To achieve a lifelike appearance of the artificial dentures, tooth-colored polymeric and ceramic coatings (veneers) are applied to the metal substructure. Depending on the veneering material and the technique of application, the substrate alloys must have very specific properties. Since the ceramic veneer is fired onto the metallic dental framework, alloys suitable for this technique must fulfill specific requirements concerning thermal expansion, melting range, and bonding characteristics to the ceramic material. Alloys for this porcelainfused-to-metal technique are usually referred to as PFM alloys or simply porcelain alloys. Metallic dental restorations are usually manufactured by a specific dental casting technique. Other means of manufacturing inlays, crowns, and bridges are milling, sintering, or electroforming [81 – 83].

2.3.1. Precious Metal Casting Alloys

Due to their high corrosion resistance, precious metals are well suited for use in the oral cavity, where the metals come in contact with saliva, which contains hydrogen ions H^+ (pH value of up to 3) and oxygen.

A measure for the stability of a metal is its electrochemical standard potential (relative to the standard hydrogen electrode):

Al	– 1.7	Co	– 0.3
Ti	– 1.6	Cu	+ 0.4
Zn	– 0.8	Hg	+ 0.8
Cr	– 07	Ag	+ 0.8
Fe	– 0.4	Pd	+ 1.0
Ni	– 0.3	Au	+ 1.4

The more positive the potential, the more corrosion-resistant the element. Due to their very high values, precious metals are not attacked by hydrogen ions. As shown in Figure 2, the high electrochemical potential of gold and palladium is retained in binary alloys provided the Au or Pd content exceeds 50 atom % in all phases of the alloy.

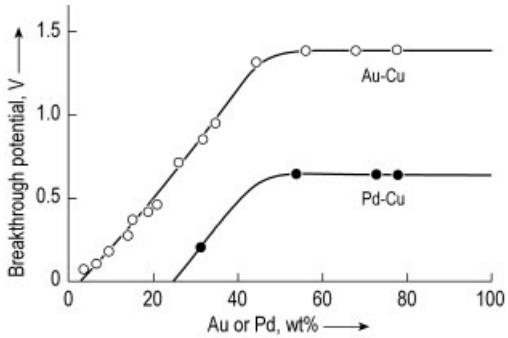

Figure 2. Corrosion resistance of binary gold and palladium alloys

Gold has a melting point of 1063 °C and a coefficient of thermal expansion of ca. 14.6×10^{-6} K^{-1} and therefore offers ideal properties as a dental material. The warm yellow color is highly esteemed due to its positive influence on the esthetics in the oral cavity. The platinum group metals all have a significantly higher melting point, so that, with the exception of palladium, none of these elements can be considered as a base for dental alloys.

In comparison with gold, palladium has a lower density and a lower price, so that material costs for restorations with palladium-base alloys are considerably lower. Palladium, however, is white. Since the coefficient of thermal expansion and melting point are outside the range required for ceramic alloys, considerable amounts of alloying elements have to be used to meet the requirement profile for PFM alloys.

Platinum is used as alloying element, especially in high-gold alloys, to increase the strength and to raise the melting point. Ruthenium and iridium are used as additives for grain refinement.

Tables 2 and 3 give an overview of the various base systems used as dental alloys with their main alloying elements. A differentiation is made between alloys suitable for ceramic veneering (PFM alloys) and alloys mainly used for unveneered restorations or for veneering with dental acrylics (C + B alloys).

High-Gold Alloys. Alloys on the base of Au – Ag – Cu have been used for the manufacturing of dental restorations since the introduction of the casting technique. The Au – Ag – Cu ternary system is shown in Figure 3. While a complete miscibility of the components is possible in the

Table 2. Composition of typical precious metal dental alloys

Alloy	Composition, wt %									
	Au	Pt	Pd	Ag	Cu	Sn	Zn	In	Ga	Others
Gold-alloys for inlays, crowns, and bridges (not for ceramic veneering)										
Degulor A	87.5	–	1.0	11.5	–	–	–	–	–	Ir
Degulor C	74.0	2.4	2.0	13.5	7.0	–	1.0	–	–	Ir
Degulor M	70.0	4.4	2.0	13.5	8.8	–	1.2	–	–	Ir
Stabilor G	58.0	0.1	5.5	23.3	12.0	–	1.0	–	–	Ir
Stabilor NF IV	55.0	–	9.9	29.0	–	1.0	1.0	4.0	–	Ir
Silver-alloys for inlays, crowns, and bridges (not for ceramic veneering)										
Palliag M	2.0	–	27.4	58.5	10.5	–	1.5	–	–	Ir
Pallatop (Clasp)	12.0	–	20.0	51.0	15.6	–	?	?	?	Re
Gold alloys for ceramic veneering										
Degudent G	86.0	10.4	–	–	–	–	–	1.5	–	Ta, Rh
BiOcclus 4	85.8	11.0	–	–	–	–	0.5	1.7	–	Ta, Rh
Degudent H	84.4	8.0	5.0	–	–	–	–	2.5	–	Ta
Degutan	80.2	4.0	13.5	–	–	2.1	–	–	–	Ir
Degudent U	77.3	9.8	8.9	1.2	0.3	0.5	–	1.5	–	Ir, Re, Fe
Degunorm	73.8	9.0	–	9.2	4.4	–	2.0	1.5	–	Ir
Gold-Palladium alloys for ceramic veneering										
Deva 4	51.1	–	38.5	–	–	–	–	9.0	1.2	Ir
Degubond 4	49.6	–	29.0	17.5	–	3.0	–	–	0.5	Ta, Ir, Re
Palladium and alloys for ceramic veneering										
Degupal G	4.5	–	77.3	7.2	–	4.0	–	–	6.0	Ge, Ru
Bond-on 4	1.0	1.0	79.7	–	5.0	6.5	–	–	6.0	Ru
Pors-on 4	–	–	57.8	30.0	–	6.0	2.0	4.0	–	Ru
Palliag LTG	–	–	36.9	56.0	–	3.0	4.0	–	–	Ir

* Coefficient of thermal expansion

Au – Cu and Au – Ag systems over the entire range of concentration, the solubility of silver in copper and of copper in silver is only limited. With decreasing temperature, this solubility gap increasingly extends in the direction of the gold corner. Alloys with a composition in the range of the solubility gap can be hardened by heat treatment [86]. During this process, precipitates that are rich in copper or silver are formed and increase the hardness and strength of the alloy. Figure 4 shows the structure of the Degulor M alloy in the soft and hardened condition.

By means of slight variations in the silver and copper contents, the extent of precipitate and thus the mechanical properties of these alloys can be varied. Consequently, it is possible to design optimum alloys for the various clinical indications. A further improvement of these alloys is achieved by adding platinum and iridium. Due to the grain-refining effect of these elements, dendritic solidification and thus an inhomogeneous element concentration within the grain (segregation) can be avoided. The fine-grain microstructure also leads to an improved corrosion and tarnish resistance; ductility and strength are increased. Table 2 lists the composition and Table 3 the properties of Degulor alloys for various indications, together with a selection of other DeguDent alloys.

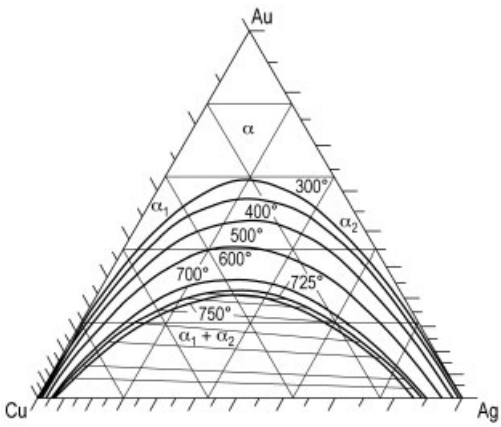

Figure 3. Au – Ag – Cu ternary system

In the past high-gold alloys based on Au – Ag – Cu could be veneered with resin materials

Table 3. Physical properties of typical precious metal dental alloys

Alloy	CTE *, 10^{-6} K^{-1}	Melting range, °C	0.2 % proof stress, MPa	
			s	h
Gold-alloys for inlays, crowns, and bridges (not for ceramic veneering)				
Degulor A		1030 – 1080	80	–
Degulor C		900 – 975	330	350
Degulor M		900 – 970	400	620
Stabilor G		860 – 940	400	830
Stabilor NF IV		940 – 1065	310	555
Silver-alloys for inlays, crowns, and bridges (not for ceramic veneering)				
Palliag M		950 – 1040	320	940
Pallatop (Clasp)		950 – 965	530	863
Gold alloys for ceramic veneering				
Degudent G	14.7	1045 – 1140	420	470
BiOcclus 4	14.5	1040 – 1140	360	530
Degudent H	14.4	1100 – 1210	480	600
Degutan	14.0	1190 – 1280	545	635
Degudent U	14.0	1150 – 1260	470	580
Degunorm	16.8	900 – 990	340	500
Gold-Palladium alloys for ceramic veneering				
Deva 4	13.8	1235 – 1315	425	545
Degubond 4	14.7	1160 – 1280	390	550
Palladium and alloys for ceramic veneering				
Degupal G	14.3	1170 – 1295	585	585
Bond-on 4	14.0	1155 – 1290	575	575
Pors-on 4	15.2	1175 – 1275	340	650
Palliag LTG	16.5	1050 – 1130	330	510

* Coefficient of thermal expansion

based on poly(methyl methacrylate) (PMMA) to give the restoration a tooth-colored appearance. Esthetically higher grade restorations, however, could only be obtained by veneering with ceramic materials that are fused to the alloy at ca. 980 °C. Since the solidus temperature of the alloys based on Au – Ag – Cu lies below this temperature, these alloys could not be used for this purpose. Newly developed low-fusing (hydrothermal) ceramic materials allow for ceramic veneering of this type of alloys.

Since the difference between the firing temperature of the ceramic and the solidus temperature of the alloy should be at least 100 °C, the solidus temperature of such an alloy must be raised above the melting point of gold (1063 °C). Therefore, platinum and palladium are used as alloying elements (Fig. 5). However, the addition is limited by the requirement for a maximum liquidus temperature of 1300 °C.

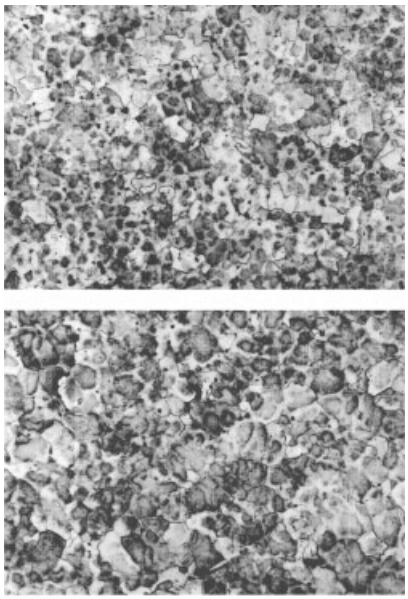

Figure 4. Microstructure of Degulor M in the hardened and soft states
A) Hardened, soft + 15 min/400 °C/air; B) Soft, 15 min/750 °C/Ar, H_2O

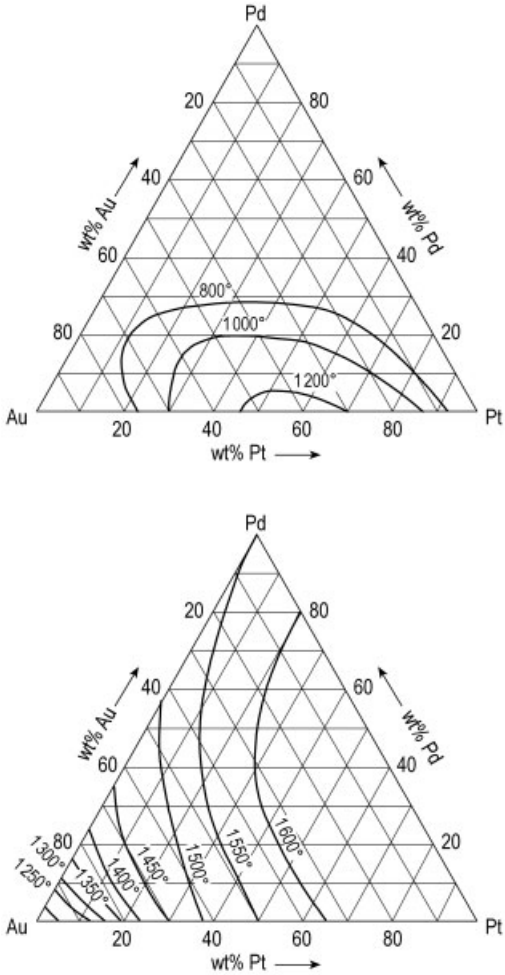

Figure 5. Ternary Au – Pt – Pd phase diagram with isotherms of the solubility gap and some liquidus temperatures

Palladium as an alloying element can be used very effectively to raise the melting range because, in contrast to platinum, it widens the melting range considerably less. A very wide melting range increases the tendency to develop segregations during solidification of the alloy. However, because of the unlimited miscibility of Au and Pd only a solid-solution hardening is possible that does not result in an adequate strength of these alloys. Moreover, palladium adversely affects the gold color. In contrast, platinum permits precipitate hardening at Pt contents exceeding about 6 wt % (Fig. 5) [87]. Therefore, palladium, which raises the melting range, and platinum, which increases the melting range and hardenability, are usually added to high-gold PFM alloys. Moreover, since alloys with high strength are required, especially for long-span bridges, nonprecious metals such as copper, tin, indium, iron, etc. are added to these alloys in concentrations up to 2 wt %. Since in systems with such complex structures and properties a theoretical description with phase diagrams is not possible, the development of dental alloys usually requires considerable experimentation with different alloy formulations. Results of such development efforts are often reflected in patents [88 – 90].

Alloying elements such as tin, indium, iron, etc., efficiently harden high-gold PFM alloys but they also lower the melting range. For compensation, palladium is added. Due to the decolorizing effect of palladium, gold-base high-strength ceramic alloys are therefore nearly white. Table 3 lists the properties of high-gold Degudent alloys. Aside from increasing the strength, the nonprecious metals are significant for a good bond between alloy and veneering ceramics. During oxidation and veneering, these additives are oxidized near and at the surface and support the wetting of the alloy by the glass phase of the dental ceramics [91].

The development of new low-fusing dental ceramic materials such as Duceragold (firing temperature ca. 740 – 780°C) made it possible for the first time to develop a golden-yellow, high-strength alloy (Degunorm) which is suitable even for large-span bridges and partial dentures. Degunorm and Duceragold have been marketed as the Golden Gate System since 1993. The alloy is mainly derived from Degulor M, which has been known since the mid-1950s. By lowering the Cu content to 4.4 % and removing the Pd the veneering characteristics and color were optimized (see Tables 2 and 3).

Gold-Reduced Alloys. With the sharp increase of the gold price in 1982, high-gold alloys became increasingly unaffordable for many patients and for health insurance systems, leading to an increased consumption and development of alloys with a reduced gold content. The Au – Ag – Cu alloy base is similar to that of high-gold alloys but the gold content is reduced to ca. 60 % and the silver content is increased by more than 20 %. The palladium content is ca. 5 %. Due to the low atomic weight of palladium

compared with gold, it can replace almost twice the amount of gold and platinum in these alloys, so that the corrosion resistance is also ensured. Due to their Pd and lower gold content these alloys no longer display the intensive yellow color of the Degulor group alloys. Typical examples are the Stabilor alloys (Tables 2 and 3).

Improper handling of these alloys can lead to porosities on the surface of dental castings. These porosities can lead to the formation of so-called aeration elements. Due to the different oxygen contents on the surface of the alloy and inside the shrinkage cavities, a difference in the electrochemical potential results between the two areas. In copper-containing alloys, copper ions go into solution inside the pore, migrate out of the pore, and react with anions in the saliva to form a dark discoloration around the shrinkage cavity (tarnish).

This potential problem provided the impetus for the development of copper-free, gold-base alloys, based on the ternary system Au – Ag – Pd. In contrast to the Au – Ag – Cu system, however, all binary systems and the ternary system have no miscibility gaps. Precipitate hardening is therefore not possible.

The ternary system Au – Pd – In has a solubility gap like the binary-edge systems Au – In and Pd – In. It should therefore be possible to achieve an increase in strength by alloying an adequate amount of indium to the Au – Pd – Ag base. The range of the solid solution of In in this ternary basis is the narrowest at medium gold contents, so that only little indium has to be alloyed to obtain precipitate hardening.

The C & B alloys are not suitable for ceramic veneering. Since there was also a demand for economical alloys for ceramic veneering, it was necessary to reduce the platinum content in addition to the gold content. Therefore, Au – Pd was chosen as the base system. To achieve a noticeable cost reduction, the gold content was reduced to 50 – 60 % and the palladium content was raised to 35 – 40 %. Due to the strong decolorizing effect of palladium all these alloys are white.

The high proportion of palladium results in a considerable rise of the melting range to values far above 1400 °C. To lower the melting range, elements such as indium, zinc, tin and to a slight extent also gallium have to be alloyed to considerably lower the melting range. At the same time the severe lattice distortions introduced by these alloying elements cause strong solid-solution hardening. A typical product is Deva 4 (Tables 2 and 3). The high precious metal content of this alloy (ca. 90 wt %) ensures good corrosion resistance in the mouth. The higher content of nonprecious elements in comparison with high-gold alloys, however, can make these alloys more sensitive to processing errors.

Palladium Alloys. Attempts to develop low-gold (or even gold-free) precious metal dental alloys date back to the 1930s, where following the depression Ag – Pd alloys were introduced as dental casting alloys. When manufactured properly, dental restorations made from these alloys show reasonable clinical behavior, but a slight darkening may occur over time as a result of the high silver content of these alloys and the tendency of silver to tarnish if it comes into contact with sulfur. Silver-base alloys that contain up to 5 % gold are still widely used, e.g., Palliag M (see Tables 2 and 3).

In the field of PFM alloys the development of low-gold Pd alloys dates back to the early 1970s, when Pd – Ag alloys were first introduced for veneering with dental ceramics. However, due to the tendency of Ag-containing alloys to frequently cause greenish discolorations in the dental ceramic materials available at that time, a higher acceptance of Pd alloys for ceramic veneering came only when first Ag-free versions were developed in the late 1970s as a response to soaring gold prices and with the introduction of nongreening ceramic in the early 1980s (e.g., Duceram).

To obtain Pd-base alloys suitable for dental ceramic veneering materials and for melting and casting with standard laboratory equipment the melting range of palladium has to be lowered by at least 250 °C, while the coefficient of thermal expansion has to be increased to a minimum of $14.0 \times 10^{-6}\,K^{-1}$. For this purpose, low-melting metals are used for alloying. The element with the strongest influence in lowering the melting temperature is gallium (Fig. 6). Tin, copper, and cobalt also lower the melting temperature significantly, whereas indium and silver have only a minor effect. Silver, copper and cobalt are completely soluble in palladium, whereas gallium, tin, and indium have solubility limits in palla-

dium of 10, 17, and 18 wt %, respectively, at 700 °C [92].

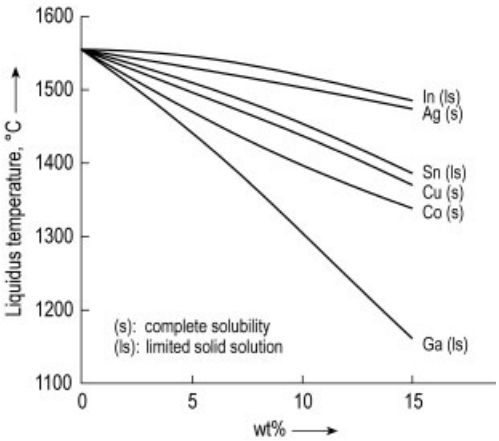

Figure 6. Influence of various metals on the liquidus temperature of binary Pd alloys

In contrast to high-gold alloys, in which a finely dispersed second phase with a high content of precious metal is required to increase the strength, second phases are undesired in palladium-base alloys. To obtain a single-phase alloy, the amount of gallium and copper have to be reduced distinctly (Fig. 7). The ternary phase diagram shows that with decreasing copper content the area of the solid solution crystal also becomes wider, but even outside the two-phase region the formation of a second phase is possible as a result of additional alloying elements. A further reduction in the risk of undesired segregation and the formation of second phases can be achieved by grain refinement.

Palladium and silver, similar to palladium and copper, are completely miscible. Silver, however, can be alloyed to a high degree since it does not lead to any formation of oxide during the firing of veneering ceramics. Typical Pd – Ag alloys contain ca. 30 wt % silver. As additional alloying elements, tin, indium, and zinc are preferably used to provide the required mechanical characteristics and CTE values by solid-solution and precipitate hardening. They also lower the melting range.

At 1000 °C silver has a vapor pressure of ca. 100 mbar. During the firing of ceramic materials, which is performed under vacuum (about 50 mbar), silver evaporates from the surface of the alloy and is partly deposited on the surface of the ceramic, often resulting in a greenish discoloration. New types of ceramics which are insensitive to discoloration by silver have been developed (e.g., Duceram).

2.3.2. Nonprecious Metal Casting Alloys

Nickel-Base Alloys. Although Co – Cr and Ni – Cr alloys with adequate tarnish and corrosion resistance have been available since about 1930, noble-metal alloys are still the materials of choice for cast inlays, crowns, and bridges. Efforts to fuse dental ceramics to nonprecious alloys were first successful only around 1970, when Ni – Cr alloys were used that are similar to creep- and sag-resistant superalloys for high-temperature applications such as furnaces and

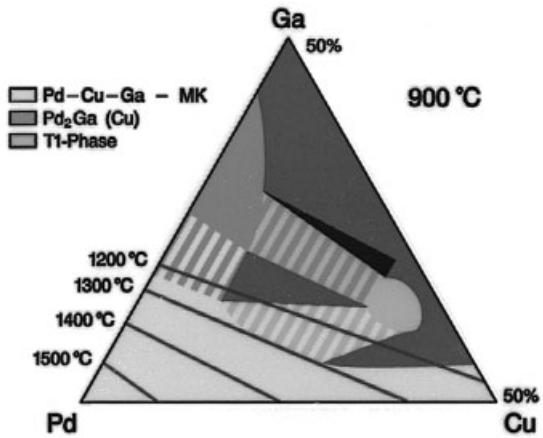

Figure 7. Ternary phase diagram Pd – Cu – Ga

turbines. Chromium forms a passive layer on these alloys, which prevents further corrosion. The basic composition is usually 80 % nickel and 20 % chromium. Specific additives (e.g., Mo, Nb, Ta, and Al) are necessary to obtain the desired properties and control carbide formation. Manganese provides sulfidation resistance, while boron and silicon lower the melting range, and beryllium improves castability as well as the alloy – ceramic bond.

Advantages of CFM alloys based on the Ni – Cr system are the low density, lower thermal conductivity, the high yield strength, and the modulus of elasticity, which is almost twice that of precious metal alloys. However, the handling of these tough alloys is not as easy as that of precious metal alloys, and high-temperature laboratory equipment is required for processing on a consistently successful basis. The technical procedure is extremely sensitive, and technicians must be specially trained. Common problems include miscasts, poor fit, porcelain fracture, weak joints when brazed after ceramic firing, and extended finishing time.

The biocompatibility of nickel alloys is another point of concern. Nickel can cause allergic dermatitis. Therefore, the use of nickel-containing alloys is contraindicated in nickel-sensitive patients. Technicians must also be aware of this potential risk. Moreover, nickel dust is considered a carcinogen. Cancer of the respiratory tract has been particularly well documented in persons who are occupationally exposed to nickel; therefore, precautions must be taken to avoid inhalation of nickel dust during processing. Beryllium is of concern mainly to the dental technician because it may provoke beryllosis.

Cobalt-Base Alloys. Primarily because of the hazards associated with nickel and beryllium, alloys have been developed that do not contain these elements. The Co – Cr – Mo casting alloys used for removable dentures (Section 2.3.3) are not useful for the CFM technique, because they are not scale resistant, their oxides react unfavorably with the ceramic veneer, and their thermal expansion does not match that of the common brands of dental ceramics. Special alloys have been developed to overcome these problems. Some Co – Cr alloys are on the market, with tungsten instead of molybdenum as a main constituent or with titanium partially replacing chromium. The latter alloys are especially prone to severe oxidation at elevated temperature and must be melted and cast under vacuum. Strength, melting range, and thermal expansion of these Co – Cr CFM alloys are higher than those of the Ni – Cr alloys. A completely modified processing technique is necessary to obtain acceptable results [5 – 7].

Titanium and Titanium Alloys. Formerly, titanium and titanium alloys were only used for implants, but pure titanium has become increasingly popular as casting alloy. Advantages of titanium are biocompatibility, low density, and low thermal conductivity. A disadvantage is that titanium undergoes a phase transformation during cooling from solidification temperature to room temperature. The high-temperature α-phase is stabilized by impurities, especially oxygen. Contamination of the surface creates a brittle surface layer called α-case. Since titanium is a very reactive metal it must be cast under an atmosphere with a very low oxygen content, and the investment material must be chemically more stable than for other dental castings. The low density is not only an advantage but also a disadvantage, since higher casting forces are required to obtain void-free castings. In this context the low X-ray opacity of titanium is an advantage because radiographs can be used to determine whether the casting is free of voids.

Recently, ceramic materials for veneering titanium have been introduced into the market (e.g., Duceratin). Clinical observations have indicated that titanium veneered with these dental ceramic materials still has a higher risk of fracture and failure of the metal – ceramic bond. [81, 84]. However, acrylic resin veneers give better results on titanium than on precious-metal or other nonprecious-metal dental alloys [85].

2.3.3. Casting Alloys for Removable Dentures

Removable partial dentures are commonly cast of suitable alloys, whereas complete dentures are now rarely made of metal; for orthodontic appliances, wires made of stainless steel or nickel – titanium alloys are generally used. The removable partial denture, which is primarily

Table 4. Composition of typical base metal dental alloys

Alloy	Composition, wt %									
	Co	Cr	Ni	Mo	Mn	Fe	Nb	W	Si	Others
Biosil F	64.8	28.5	–	5.3	0.5	–	–	–	0.5	C
Duceranium U	0.5	21.5	59.0	4.5	0.4	3.5	3.2	5.0	0.8	C, Cu
Duceralloy U	59.0	25.0	–	–	0.3	2.0	2.0	10.0	0.8	C, V, Al
NPG			4.3		1.7	3.0				**Cu**, Al, Zn
Rexillium III	0.5	13.0	76.0	3.0	–	–	–	–	1.0	Ti, **Be**, Ga
Remanium GM 800	63.5	30.0	–	5.0	0.2	–	–	–	1.0	C
Remanium 2000	61.5	25.0	–	7.0	–	–	–	5.0	1.5	

Table 5. Physical properties of typical precious metal dental alloys

Alloy	CTE *, 10^{-6} K^{-1}	Melting range, °C	0.2 % proof stress, MPa	
			s	h
Biosil F		1320 – 1380		700
Duceranium U	14.5	1325 – 1328	325	335
Duceralloy U	14.4	1280 – 1330	550	550
NPG		1012 – 1068		210
Rexillium III	14.2	1232 – 1288	510	580
Remanium GM 800		1240 – 1410		710
Remanium 2000	13.8	1290 – 1415		600

supported by the mucosa, is usually held in position by clasps around the remaining natural teeth. The clasps and the denture base are often cast simultaneously as one piece of a single alloy by using a specific investment casting technique. Co – Cr alloys are mainly used for this purpose (e.g., Biosil F; Tables 4 and 5). Esthetically more pleasing devices to attach a removable partial denture to natural teeth or to a fixed partial denture are telescoping crowns or industrially prefabricated attachments (see Section 2.6).

2.4. Brazing Alloys

Brazing, commonly called soldering in the dental profession, is the operation most extensively used in dental laboratories for joining two pieces of a metallic dental appliance together. Spot welding is used occasionally, and laser welding is becoming increasingly popular.

Brazing alloys are used in the form of strips, rods, or ingots. Brazing is frequently conducted in electrical furnaces or by means of a gas – oxygen torch between 750 and 1150 °C. A brazing flux dissolves or removes oxide and has a positive effect in promoting rapid wetting and spreading of the brazing filler metal. Because brazing must be conducted at temperatures that are 20 – 150°C above the liquidus temperature of the filler metal without fusing the parent metal, a suitable brazing alloy has to be chosen for the particular combination of metals to be joined.

Noble-Metal Alloys. Because brazed joints must be corrosion resistant in the oral environment, noble-metal alloys are the filler materials of choice. Brazing alloys based on the Au – Ag – Cu system are called *gold solders*. They usually contain 50 – 75 % gold, up to 10 % platinum group metals, balance silver and copper, plus up to 15 % zinc and, occasionally, minor additions of indium or tin to lower the melting range. These yellow-colored gold solders are suitable for all joining purposes in the fabrication of fixed restorations, except for joining prior to firing the ceramic veneer. For this application, high-melting Au – Ag alloys are used, which contain 60 – 80 % gold and up to 10 % platinum metals, but only trace amounts of base metals.

When the joint should not differ in color from white parent alloys, *Au – Pd – Ag brazing al-*

loys are available. Although they contain up to 45 % copper, zinc, tin, and indium, some of these have acceptable corrosion resistance if the noble metal content is not less than 40 %. These white brazing alloys are preferred in the fabrication of orthodontic appliances, removable partial dentures, and ceramic-veneered restorations made of base-metal casting alloys.

Brazing alloys of Ag – Cu – Zn, sometimes used for joints in removable appliances, cannot be accepted for long-term use, because of their low corrosion resistance. Furthermore, the brazing operation itself reduces the passivation of some base-metal parent alloys, which can result in mechanical breakdown of the appliance because of corrosion of the joint.

Alloys of Au – Ni with about 80 % gold, are used particularly for direct brazing of a clasp wire to denture bases cast from Co – Cr alloy. Fusion of a high-melting *Au – Ni brazing alloy* to one part of a restoration made of a base-metal alloy, prior to brazing it with a gold solder to another part made of a noble-metal alloy [1, 3, 4], is also common.

Base-Metal Alloys. So far the only base-metal brazing alloys commonly used in joining base-metal CFM alloys prior to firing the ceramic veneer are Ni – Cr – Si – B alloys, e.g., BNi74CrFeSiB 980 – 1080 (ISO 3677 designation for nickel-based solder containing 14 % Cr, 4.54 % Fe, 4.5 % Si, 13 % B, and having a working range of 980 – 1080 °C).

2.5. Metallic Materials for Implants

Since the early 1970s alloplastic implants have become increasingly important. They are used as replacements for missing teeth and for retaining full dentures. All implant materials must be highly corrosion resistant in the oral cavity as well as in the bone structure. They must also be strong enough to withstand mastication forces. Implants are either supported by the jaw bone (subperiosteal implants) or fixed in the jaw bone (endo-osseous and endodontic implants). *Subperiosteal implants* are fabricated, according to an impression of the jaw bone, in a dental laboratory by using the lost-wax method with a Co – Cr – Mo casting alloy. Individually shaped implant frameworks are obtained, which allow a prosthesis to be fixed on the posts passing through the gum. Nowadays this kind of implant is not very popular anymore because of problems with the fixation of the implant on the bone.

For *endo-osseous and endodontic* (transfixation) implants, prefabricated parts of various alloys are available from the dental industry. They are fixed in the jawbone. Titanium and titanium-based alloys are most important because of their favorable chemical and physical properties and their excellent biocompatibility (e.g., Ti with 6 % Al and 4 % V, or Ti with 5 % Al and 2.5 % Fe). Tantalum is also highly biocompatible, but because of its high density and poor mechanical properties, it is not yet used as an implant material on a wide basis. Wrought cobalt-based alloys, such as Co – Cr – W – Ni and Co – Ni – Cr – Mo, have also been used as implant materials. Formerly, stainless steel with ca. 18 % Cr, 14 % Ni, and 3 % Mo was used as implant material, but its corrosion resistance and hence its biocompatibility are not as good as of other alloys [8].

Good incorporation of implants into the jaw bone requires a rough surface, while the surface of the transmucosal part has to be smooth. Roughening of the surface is done by spray coating of titanium with titanium or by sandblasting or etching. The difficulty in all-cases is to keep the surface free from foreign material which may reduce biocompatibility.

To enhance the healing process and to provide a bioactive surface some implants are provided with a surface layer of hydroxyapatite or similar calcium phosphate compounds. The problem with these layers is that the adherence to the metal surface is too weak to withstand the shear forces during the insertion process. Chipped-off parts from the surface may create problems during the healing process.

Implants are also available completely made from ceramic material, mainly aluminum oxide, or from carbon glass or carbon fiber. The disadvantage of the carbon implants is the black color, which is visible through the gum.

A survey of materials approved in general implant surgery is given in the ISO Standard Series 5832.

Implant systems are most successful when prefabricated endodontic and endo-osseous implants are provided along with the corresponding root canal and bone-cutting instruments, respectively.

Because of the different techniques of placing implants, various shapes of implants are available. Blades, screws, and hollow cylinders are used mainly for endo-osseous implants [9, 10]. After a root resection, endodontic implants must be inserted through the root canal into the jawbone. The usual implant shapes are therefore limited to needles, pins, and screws [11].

2.6. Metals and Alloys for Other Purposes

For other applications and operations in dentistry besides casting and brazing, metals and alloys are used in the form of either prefabricated parts or semifinished materials.

Root Canal Posts. Because of the severely corrosive conditions in crevices, highly corrosion-resistant materials must be used for root canal posts which are inserted into the root of a nonvital tooth as a retention for the artificial die. Some prefabricated root canal posts are made solely of noble metals (e.g., 80 % Pt – 20 % Ir or 60 % Au – 20 % Pd – 20 % Pt), and others from titanium or a titanium alloy with 6 % Al and 4 % V. Stainless steel is not well suited for root canal posts, even if the posts are gold plated, because of its low corrosion resistance. Prefabricated posts come in different shapes and dimensions for different root sizes and purposes.

Precision Attachments. An important application of noble-metal alloys is in prefabricated precision attachments. Various anchors, hinges, bar joints, etc., are available, which are based largely on the male – female principle. They are used to couple different parts of a prosthesis (e.g., a removable partial to a fixed partial denture) and to achieve fastening and accurate location of the prosthesis. The advantage of these attachments is that the prosthesis or parts of it can be removed for inspection and repair, for cleaning, and for care of the adjacent oral tissue. Depending on the construction, indication, and intended joining technique (brazing, casting onto), precision attachments are made of high-melting, nonoxidizing alloys (e.g., Pt with 20 % Ir or Au with 15 – 25 % Pt and 15 – 25 % Pd) or of high-strength alloys based on the Au – Ag – PGM system or the Ag – Pd – Cu system.

Platinum Foils. A particular application of platinum foils (about 25 μm thick) is in the fabrication of full ceramic crowns (jacket crowns), in which a foil cap must conform to the die onto which a particular alumina porcelain is fired. After the firing operation, the platinum foil is removed, leaving a gap for cementation.

An alternative to this full ceramic crown is the platinum-foil-reinforced ceramic crown, in which the cap made from the foil is conditioned in such a way that the dental ceramics commonly used for veneering cast metal substrates can be fused to it, and the composite crown then cemented. Successful conditioning techniques are tin plating and oxidizing, fusing of gold bonding agents, or covering with a gold brazing alloy [1, 3, 4].

Tools, Sheets, Bands, and Wires. Stainless steel is the material from which not only instruments and tools for dentistry are fabricated, but sheets, bands, wires, etc., for intraoral application are made as well. Usually the 18 – 8 variety of austenitic steel is employed, whose corrosion resistance depends on proper processing, e.g., heat treatment or cold working. Occasionally, Ni – Cr and Co – Cr alloys are also used.

Orthodontic Materials. Arch bars, springs, and bands in orthodontic appliances are generally made of stainless steel and are constructed and joined by spot welding or, occasionally, by brazing with silver solder. Both these joining techniques can reduce corrosion resistance, but this is acceptable because such appliances are used temporarily. Nickel – titanium alloys have also become popular for use in orthodontic appliances. Stainless steel is also employed for prefabricated screws and brackets in orthodontic appliances; for matrix bands to support the filling material packed into the tooth cavity; and for pins to improve the retention of large amalgam or resin restorations.

3. Dental Ceramic Materials [21 – 23]

3.1. Ceramic Teeth

As early as 1774, Duchâteaux and Chémant reported the use of ceramic powders for fabrication of teeth. Fonzi, in 1808, was the first to fire platinum posts into teeth as anchoring elements.

Ceramic teeth were manufactured in 1838 (Ash, England) and 1844 (S. S. White, United States). Since 1950, demand for ceramic teeth has diminished continuously. Today, artificial teeth fabricated from resin materials are preferred due to their lower price and easier processibility.

3.1.1. Raw Materials

The formula for table china was the starting point for the development of ceramic powders for teeth; as a rule, this porcelain is composed of 50 % kaolin, 25 % feldspar, and 25 % quartz. Teeth made from this material, while functionally acceptable, were faulty aesthetically because of their excessively white opaqueness. The high content of kaolin is responsible for this white opacity because of the highly refractive, highly dispersed mullite that forms from kaolin during firing. For this reason, the proportion of kaolin in the formula was greatly decreased or eliminated. However, the mechanical properties of dental porcelain suffered as a result.

The formulas for today's dental porcelain powders usually include more than 90 % of a mixture of feldspar and quartz, the quartz component being 15 – 25 % of the mix. For this reason, the term "porcelain tooth" has been replaced in the newer literature by "ceramic tooth."

Feldspar. Feldspar is the most important raw material, in ceramic powders. High demands are made on its quality: it should be largely free of foreign phases and low in iron content because of the danger of discoloration. Norwegian and Swedish feldspars are well suited because they consist largely of orthoclase ($K_2O \cdot Al_2O_3 \cdot 6\ SiO_2$), with the following typical composition (mass fractions in percent):

SiO_2	65.0
Al_2O_3	18.5
K_2O	12.5
Na_2O	2.5
CaO	0.2
Fe_2O_3	0.02

Quartz. The quartz used must have a similar purity. Quartz decreases the melting range of the feldspar – quartz mixture and provides sufficient rigidity to the ceramic mass during firing by increasing its viscosity. This assures that the formed contours of the tooth are not lost during firing. Furthermore, the quartz content affects several important properties of the ceramic material: flexural strength, thermal expansion, resistance to thermal shock, and opacity. These properties depend not only on the proportion of quartz but, significantly, on how much of the quartz passes into the glassy phase and how much remains in the crystalline phase.

Occasionally, minor amounts of opacity-inducing materials such as kaolin, mullite, or carborundum are added to the mixture. Addition of flux powders, such as soda or calcite, can reduce the melting temperature.

3.1.2. Fabrication

The raw materials are crushed first in a jaw crusher or chaser mill and then ground in ball mills before being mixed in accordance with the prescribed formula (maximum particle size 150 μm). To achieve better results, the mixtures are customarily fritted first. For this step, the raw materials are placed in fireclay containers and fired between 1200 and 1400 °C until the feldspar has melted completely and the quartz has largely dissolved in the mass. The glass blocks that result from this process are again crushed and ground to a fine powder (maximum particle size 100 μm).

The starting material for tooth fabrication thus prepared must now be given a *color.* A large number of differently colored ceramic powders are needed, because manufacturers offer approximately 20 different shades of teeth. Furthermore, every tooth is built up of several differently opaque and variously shaded layers. The natural appearance required cannot be attained in any other way. Shading is provided by the pigments ordinarily used in ceramics, e.g., red-brown spinels (Fe, Zn, Cr, Al oxides) and yellow vanadium-containing ZrO_2. Precise control of the color values and the transparency of the powders are prerequisite to achieve the desired tooth colors. The fluorescence of natural teeth is imitated by using special fluorescing pigments or by adding lanthanide oxides to the starting frit. The colored powders are converted to plastic masses by use of appropriate organic additives, e.g., starch paste or oils.

Shaping teeth requires much manual effort. Two-part metal molds are generally used. The

labial layers of the tooth are placed in one-half of the mold and the lingual layers in the other. Posterior teeth are formed from occlusal and basal halves. To permit retention of the completed teeth in the denture base material, precious-metal shells (largely palladium) are mounted on the back of the anterior teeth. After firing, gold-plated posts are soldered to these shells. For posterior teeth, retention is provided by dovetail-shaped undercuts in the base region.

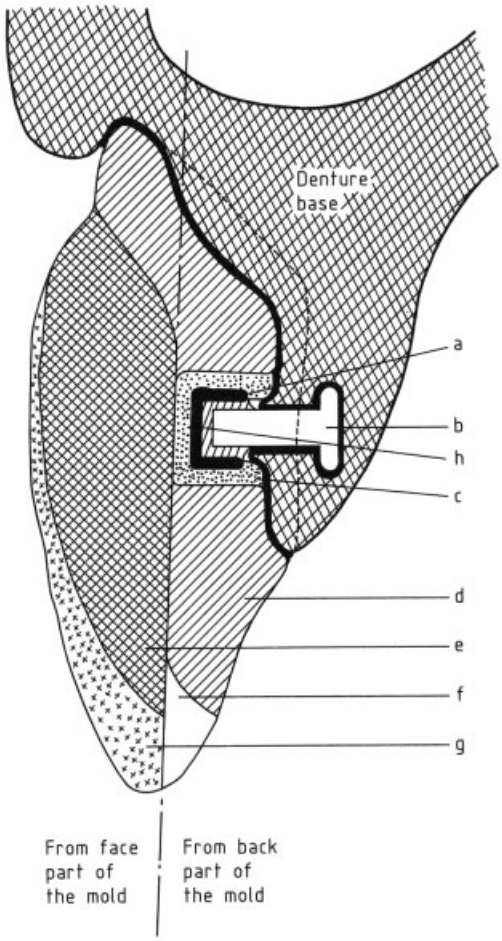

Figure 8. Cross section of a ceramic tooth and embedding in denture base
a) Precious-metal shell; b) Soldered gold-coated pin; c) Core material; d) Back material; e) Dentin material; f) Translucent material; g) Enamel material; h) Solder

The filled halves of the form are pressed together for joining. When they are heated to 150 – 200 °C, the organic components lose their plasticizing effect and serve as binders. The formed teeth are then placed on ceramic firing plates and heated to 800 °C to volatilize the organic components. The tooth is finally fired at 1200 – 1350 °C for 8 – 30 min, depending on the manufacturer. Today, the vacuum-firing technique, introduced in 1949 by Gatzka, is used primarily. This produces teeth essentially free of pores, which is beneficial in terms of strength and polishability of the teeth and — by increasing the transparency — particularly desirable in terms of aesthetic effects. In this procedure, the furnace is evacuated to approximately 1 kPa. When the tooth surfaces have become sintered, firing is continued at atmospheric pressure; firing is complete when the tooth surfaces are smooth while the tooth contours are still unchanged. After soldering of the retentions and the required quality checks, the teeth are collected in sets according to color and shape (Fig. 8).

3.1.3. Quality Requirements

Quality requirements are defined in ISO Standard 22112 (Dentistry – Artificial teeth for dental prostheses). Porosity, color matching with the shade guide, and dimensional stability are particularly stressed. Additional requirements are tissue compatibility of the material, and observance of laws and regulations relating to protection from radiation. As a result, uranium salts, formerly added to achieve fluorescence, are no longer used.

3.2. Dental Ceramic Fused to Metal Restorative Materials

In the past, partial dental restorations such as crowns and bridges were fabricated entirely of precious metal. Today, standard practice involves casting the framework in metal and veneering this with a ceramic layer. After insertion, such a prosthesis can hardly be distinguished from natural teeth.

The first attempts of this sort were made in 1728 by Fauchard, who attempted to enamel a gold denture base. However, only in 1962, was Weinstein, an American, able to develop frits that achieved sufficient adhesion between metal and ceramic [24]. The bonding material

produced is related to enamel. The metal provides sufficient tensile strength, while the ceramic, glaze-like veneer affords great abrasion resistance, good tissue compatibility, and an esthetic appearance.

In the past, mechanical retention of the ceramic layer on the rough metal surface was thought to be a prerequisite for good adhesion between the two components. Today, more significance is attributed to adhesion through chemical bonding between the materials. Bond formation generally is ascribed to the presence of easily oxidized components, such as tin and indium, in the alloy (see Chap. 2). Presumably, the metal oxide layer between the metal and the ceramic, enriched through a so-called oxide firing step, provides the chemical bridge between the components [25].

3.2.1. Thermal Expansion

The durability of the bond between the metal and the ceramic depends particularly on the thermal expansion of the components [26]. If the difference in this parameter is too great, the ceramic inevitably will split away from the metal during cooling. Under optimal condition, the ceramic shrinks a little less than the metal during cooling; i.e., the ceramic has a smaller coefficient of thermal expansion than the metal. It is then under slight compressive strain, which avoids tensile strain that is absorbed less well by the ceramic. Ordinarily, metals have a higher coefficient of thermal expansion than glasses; the values of metals used for casting partial dentures are in the range of $130 - 150 \times 10^{-7}\ K^{-1}$ between 25 and 500 °C. Glasses with sufficient chemical resistance, however, have a significantly lower coefficient of thermal expansion. The required high coefficient is attained through the precipitation of a finely divided, dendritic form of leucite ($K_2O \cdot Al_2O_3 \cdot 4\ SiO_2$) as a crystalline phase in the glass during firing [27]. Leucite is transformed from the tetragonal to the cubic form in the range from room temperature to 620 °C. This transformation is accompanied by a 5 % increase in volume and thus is associated with a high coefficient of thermal expansion. A concentration of 20 – 30 % of leucite suffices to provide an adequate coefficient of thermal expansion to the ceramic. In theory, any frit that will precipitate the required amount of leucite through appropriate temperature handling suffices [28]. In practice, however, several frits of different composition are used, which permit better control of the properties of the ceramic and a broadening of the melting range; this simplifies subsequent firing of the ceramic to the bridge framework.

3.2.2. Raw Materials

Orthoclase (potassium feldspar) is the most important raw material for a ceramic that is to be fused to metal. To make a frit that easily segregates leucite, the potassium content must be increased, e.g., by adding potassium carbonate, K_2CO_3. Some manufacturers prefer to replace natural feldspar with a mixture of other raw materials, e.g., quartz flour, alumina, and potassium carbonate. Furthermore, the addition of flux powders, such as alkalis, alkaline earths, or boric acid, is necessary to lower the fusion range. The ceramic must be able to be applied at a temperature significantly below the softening temperature of the metal. Firing temperatures between 900 and 980 °C are customary, but temperatures as low as 750 °C are becoming increasingly common. The attainable decrease in the firing temperature is limited by the increasing chemical solubility of the ceramic.

3.2.3. Fabrication

The raw materials are first sintered into frits. The composition of ceramic materials varies greatly among manufacturers. This is caused, in part, by the practice of mixing several different frits. The main constituents of most of the relevant ceramic materials are present in these following ranges (mass fractions in percent):

SiO_2	58 – 66
Al_2O_3	15 – 17
K_2O	10 – 14
Na_2O	5 – 8
Li_2O	0 – 1
CaO	0 – 2
B_2O_3	0 – 3

A ceramic of this sort is made into ready-to-use powders by adding opacifiers and pigments to the frit, so that the final veneer corresponds in color to the shade guides provided.

Every veneer requires three different powders: the opaque, dentin, and enamel powders. The opaque ceramic provides adhesion and simultaneously prevents the metal from showing. To meet this function, it contains up to 12 % of opacifiers such as TiO_2, SnO_2, CeO_2, ZrO_2, or $ZrSiO_4$. Dentin and enamel powders make a toothlike appearance possible through their coordinated shades and translucence.

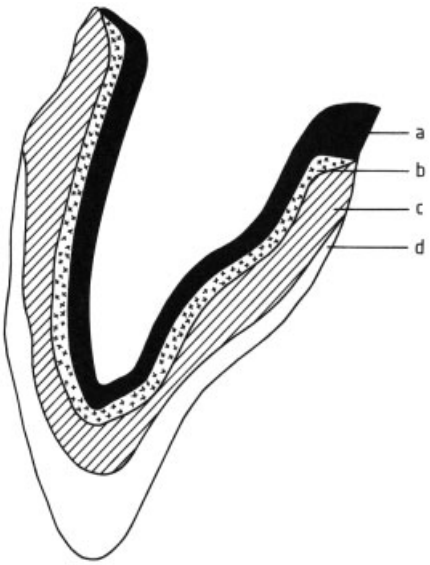

Figure 9. Cross section of a ceramic-fused-to-metal crown a) Metal substructure (0.2 – 0.4 mm); b) Opaque layer (0.2 mm); c) Dentin layer (0.6 mm); d) Enamel

The dental technician mixes the ceramic powders to a creamy mass using water or a special modeling liquid. Each of the three ceramic layers is applied and fired separately. Appropriate distribution of grain size of the ceramic powders assures sufficient contour stability and condensability of the ceramic pastes. The grain size is in the range of 2 – 80 µm, with a maximum between 20 and 30 µm. Each firing lasts about 5 min. The vacuum-firing procedure, described for tooth firing, has been widely adopted. If the veneer must be ground later for adjustment, glazing and refiring provide the necessary shiny appearance. Unpigmented, finely ground glass frits are used as glazes. They contain sufficient flux and can be fired at lower temperatures than those required for veneering. Strongly pigmented glazes are used to provide specific colors. For example, tooth necks can be accentuated with color, and superficial anomalies, such as enamel tears and white spots, can be copied. Figure 9 shows a cross section of a crown.

3.2.4. Quality Requirements

The requirements and testing procedures are defined in the standard ISO 9693. Important properties that must be checked constantly by the manufacturer include thermal expansion, adhesion to the metal, bending strength, firing shrinkage, chemical durability, and transparency and color of the fired ceramic materials.

3.2.5. Ceramic Fused to Titanium

Since about 1985, titanium has been used to a small but increasing extent as a framework material. Titanium has a lower thermal expansion (92×10^{-7} K $^{-1}$) than the traditional "ceramic alloys". Therefore, the ceramic does not require leucite to enhance its thermal expansion. The firing temperature of the ceramic must be below 850 °C, due to the high reactivity of titanium. The veneer ceramic consists predominantly of a leucite-free glass with an increased content of alkali metal oxide and boric acid.

3.3. Dental Ceramics for Fixed Restorations

3.3.1. Jacket Crowns

Long before the combination of metal and ceramic was accepted, damaged teeth were repaired exclusively with ceramic materials. The first success was achieved by H. Land, who introduced his jacket crowns in 1895 in the United States. In Europe they became popular only after 1918.

The following technique is used to fabricate a jacket crown: A model is made of a die of the prepared tooth. The model is covered with platinum foil; the foil serves as a base for the layers of ceramic to be placed on it. The foil is removed after the crown has been completed. As with ceramic-fused-to-metal restorations, several layers of ceramic, each with a specific transparency and shade, are fired. The core material

is covered with a dentin layer and then with an enamel layer.

The demands made on the ceramic's physical properties are different from those required of ceramic powders for application to metal. Because no metal base is present, the thermal expansion of the ceramic does not have to be increased, and the firing temperature does not have to be reduced below 1000 °C. Nor need the core material be extremely opaque. However, the absence of a support for the crown, otherwise provided by the metal framework, must be compensated by increased strength, at least of the core material. The flexural strength of this material should be at least 100 MPa, whereas 50 MPa is sufficient for the opaque material in the metal – ceramic technique.

Jacket crowns were fabricated in the past from ceramic powders quite similar to those used for teeth. This required that they had to be fired at temperatures no lower than 1250 °C. Today, ceramics are available which, through the addition of boric acid and other fluxes, can be fired at temperatures below 1100 °C. The necessary increase in bending strength of the core material is achieved by adding as much as 50 % alumina, usually in the form of carborundum [29]. Prerequisites to the increased flexural strength elicited by the carborundum are good binding with the surrounding glass matrix and a controlled difference in thermal expansion between the carborundum and the glass phase.

Because of the better mechanical properties of metal – ceramic crowns and because of their more readily attainable exactness of fit, they have largely replaced jacket crowns. But for patients with allergic reactions to precious metals, jacket crowns still find some use. Additionally, the lack of the opaque metal layer allows for a more natural appearance of the restoration.

3.3.2. Newer Developments

While the importance of traditional jacket crowns approaches zero, extensive efforts have been made to replace the metal frame by newly developed high-strength ceramic. Suitable ceramic materials, superior to the jacket crown material regarding their strength are readily available. Glass ceramics, for example, have flexural strengths up to 300 MPa, and oxide ceramics such as Al_2O_3 and ZrO_2 even up to 1000 MPa. While some glass ceramics have a tooth-like appearance, oxide ceramics can serve only as core materials because of their opacity and have to be veneered with more transparent ceramics. There are no generally accepted requirements for the strength of ceramics for dental restorations. The standard ISO 6872 requires a minimum flexural strength of 100 MPa for ceramic materials used for the fabrication of the supporting structure of crowns, veneers, inlays, and onlays. However, ceramics for crowning posterior teeth are only acceptable if they considerably exceed this minimum, and materials for bridges are expected to have flexural strengths above 300 MPa. Besides the dimensions of the restoration and the expected masticatory pressure, the durability also depends on the manner in which it is fixed to the prepared tooth. The main problem lies in manufacturing dental restorations from these ceramic materials at an acceptable price. The fact that each restoration has its individual dimensions which must be observed exactly exacerbates the problem.

There are several methods for shaping the ceramic material.

Sintering of ceramic powder has the problem that shrinkage during the sintering process means that the shape before and after firing differ considerably. In the case of metal – ceramic and jacket crowns, the inner surface of the restoration, which has to fit exactly to the prepared tooth, is determined by the metal frame or the platinum foil. During the firing process, the ceramic powder shrinks towards the metallic base. The form, which the dental technicians build up with the ceramic powder slip, must be oversized to compensate for the shrinkage.

Normally, glass ceramic is produced by casting of the glass melt into the desired form, followed by a tempering process, in which the glass partly crystallizes. Distribution, quantity, and size of the crystals are responsible for the quality and the properties of the material. But if the glass ceramic is sintered together from fine particles, it is hardly possible to achieve an optimal crystallization process, and the resulting material is inferior. Thus, for the production of glass ceramic restorations dense ingots are used, which are cast or pressed in a prepared form.

Dentsply introduced the first glass ceramic dental system Dicor in 1984. It was developed

jointly with Corning Glass Works. Its basis is a mica – glass ceramic [31]. The tooth crown was cast of a special glass at 1370 °C. A centrifugal casting procedure similar to that used for casting metal denture restorations was used. In the subsequent ceramic firing (14 h at 1070 °C), approximately 55 % of the glass crystallizes. The structure of the resulting translucent glass ceramic is characterized by platelike, interlaced crystals of a fluoride-containing mica with the formula $K_2Mg_5Si_8O_{20}F_4$. These crystals, approximately 1 µm in size, should lend strength to the material and retard crazing. A colored glaze is fired onto the structure to provide the desired shade [32]. The long–term success of this material was not sufficient; therefore it has been withdrawn from the market.

In 1990, Ivoclar introduced their Empress system [33]. This is also a glass ceramic system, but with some important differences compared to Dicor. The ingots are not cast but pressed into the form by a special hot-pressing device. For this reason, they have to be heated up only to the plastic state, and the reinforcing crystals present in the ingots do not dissolve during the hot-pressing process. This has the advantage that the time-consuming subsequent ceramic firing is not necessary. The crystal phase in Empress is leucite, but in contrast to metal – ceramics the crystallization process is optimized to give a flexural strength of about 150 MPa. Similar to Dicor, a colored glaze gives the desired shade. Similar systems from other manufactures are nowadays common.

Another way to overcome the problems associated with the sintering shrinkage is represented by In-Ceram, a product marketed by Vita Zahnfabrik since 1989 [34]. The model of the prepared tooth is duplicated with a special gypsum. An alumina powder slip is applied to this die and fired. Firing has to be timed in such a way that the alumina does not shrink and a porous, partially sintered preform is created. This preform is infiltrated with a glass melt in a second firing step. The resulting dense core has a flexural strength of more than 400 MPa. Therefore, this material is useful not only for inlays and crowns, but also for bridges. A toothlike appearance can be achieved by veneering the core with a more transparent ceramic.

In 1988, the first CAD/CAM system for manufacturing dental restorations was offered by Siemens. This CEREC-System was developed by Mörmann and Brandestini. It was suitable for the production of inlays only, but today, systems are available for the production of crowns and even bridges [35]. All these systems use industrially produced dense ceramic blocks as starting materials, they are therefore not confronted with the problem of shrinkage. The process can be divided into three steps:

1) Three-dimensional scanning of the surface of the prepared cavity or tooth stump by an optical or mechanical device and transmission of the data to a computer. Scanning can be performed either intra-orally or extra-orally after manufacture of a model from an impression. Instead of the cavity or the stump, a preformed model of the desired restoration can be used for the scanning process.
2) Design of the restoration using the three-dimensional computer model derived from the scanning device. The operator can either complete or change the restoration on the screen. The outer form with the occlusal surface can be designed with the assistance of different tooth forms stored in the computer's database.
3) Fabrication of the restoration from the ceramic block by a computer-controlled precision milling machine.

Many types of ceramic are applicable for this purpose, but ceramics with a glass-matrix are preferred because of their toothlike appearance and ease of grinding. The strength of these ceramics is sufficient for inlays and crowns, but for bridges oxide-ceramics have to be chosen. Most CAD/CAM systems can also be used for the milling of metal preforms.

The CAD/CAM systems are expensive but advantageous in that they allow the production and insertion of restorations at the chairside during a single visit.

CAD/CAM systems are also used to produce frameworks for full ceramic bridges even for the posterior region. Yttrium–stabilized zirconium oxide ceramic is mainly used for this purpose. Beside the main constituent zirconium oxide the ceramic contains 5 % yttrium oxide and a small amount of aluminum oxide. The main impurity is hafnium oxide, which is contained in an amount of approximately 2 %. The addition of yttrium stabilizes the tetragonal high–temper-

Table 6. Properties of cements setting by neutralization reaction

Property	Zinc phosphatecements	Silicate cements	Zinc polycarboxylate cements	Glass ionomers	Resin-modified glass ionomers	Polyacid-modified composites
Strength	low	medium	low	high	higher	highest
Wear resistance	low	low	low	high	medium	high
Solubility	medium	high	low	very low	low	very low
Shrinkage	no	no	no	no	low	high
Swelling	no	no	no	no	high	high
Bond	mechanical interlocking	mechanical interlocking	dynamic, chemical bond	dynamic, chemical bond	dynamic, chemical bond	with bonding systems
Fluoride release	addition of fluorides	high	addition of fluorides	high	high	low
Pulp sensitivity	medium	high	very low	medium	medium	medium
Biocompatibility	under discussion (H_3PO_4)	under discussion (H_3PO_4)	good	very good	under discussion (residual mono-mers; HEMA)	under discussion (residual monomers)
Translucency	low	medium	low	medium	high	high
Esthetics	low	medium	low	medium	good	very good

ature phase of zirconium oxide down to room temperature. Under the stress of a propagating crack the stable monoclinic phase is produced, associated with expansion. The pressure from this expansion stops further growth of the crack. This gives this type of material its high strength and durability.

Originally, frameworks of zirconium oxide ceramic were made from dense sintered material. Because of the rigidity of the material the milling process took a long time. Nowadays, mainly presintered material is used. It has a chalklike consistency and can be easily milled. To get a fitting restoration the computer calculates an oversized framework to be milled by taking into account the firing shrinkage. After dense sintering of the material the framework is veneered with a special veneering ceramic which is adjusted to the low expansion coefficient of zirconium oxide.

4. Cements

In dentistry the term "cement" is often used with two different meanings. Dentists use "cement" for all luting materials for the fixation of crowns, bridges, inlays, or onlay, even if the material is resin-based. Chemically speaking, the term "cement" describes dental materials that contain reactive fillers as one component of the setting system. Only these materials are treated in this chapter.

These cements can be classified according to their type of setting reaction:

– Neutralization reaction
– Complex formation reaction

4.1. Cements Setting by Neutralization Reaction

The development of amalgam, gold alloys, and dental porcelain in the 1800s led to the search for luting and more esthetic restorative materials. Cements composed of calcium oxide and phosphoric acid, magnesium oxide and solutions of magnesium chloride (magnesium oxychloride cement), or zinc oxide and zinc chloride solutions were developed. The first dental cement with satisfactory properties and relatively low pulp irritation was zinc phosphate cement. The most important properties of cements setting by neutralization reaction are listed in Table 6.

4.1.1. Zinc Phosphate Cements

Zinc phosphate cement (phosphate cement) was introduced in 1879 [93]. The powder contains zinc oxide and small amounts of magnesium oxide. The liquid is an aqueous solution of orthophosphoric acid (ca. 50 – 60 %) and aluminum phosphate. Both components may contain minor amounts of additives to improve various characteristics. During setting, the oxides are neutralized by phosphoric acid. Set cements

consist of unchanged zinc oxide particles embedded in an amorphous matrix of zinc phosphate and zinc aluminophosphate [94]. Due to the relatively low strength and the opaque appearance, zinc phosphate cements are used exclusively as luting materials for restorations.

Trade Names: Fleck's cement (Mizzy); Harvard Cement (Richter & Hoffmann).

4.1.2. Silicate Cements

Replacing the zinc oxide of zinc phosphate cements by ion-leachable glasses gives the silicate cements, first developed in 1902 [95]. Glasses are melted from silica, alumina, sodium fluoride, cryolite, and calcium sources such as calcium fluoride, calcium oxide, calcium carbonate, or calcium phosphate [94]. The liquid is similar to that used with zinc phosphate cements. In the setting reaction, the glass is attacked by acid, and metal ions are released and precipitated as phosphates. Unconsumed glass particles surrounded by silica gel are embedded in the phosphate matrix. Due to the fluoride content of the glass, silicate cements release fluoride. Strength and esthetics are better than those of zinc phosphate cements, but solubility is higher [94]. Old silicate cements led to some extent to pulp necrosis, an effect that was explained by the relatively high content of arsenic impurities in their glasses. But even cements with practically arsenic free glasses led to pulp sensitivity [96]. Possibly the slow increase in pH leading to a prolonged contact of the pulp with an acidic environment can explain these findings [94, 96]. The major use of silicate cements is in anterior restorations.

Trade Name: Silicap (Vivadent).

4.1.3. Silicophosphate Cements

Silicophosphate cements (stone cements) are a combination of zinc phosphate and silicate cements. The powder contains both zinc oxide and ion-leachable glass. Their properties are in between those of zinc phosphate and silicate cements [94]. They are used as filling materials, like silicate cements.

Trade Name: Drala Steinzement (Detax).

4.1.4. Zinc Polycarboxylate Cements

Zinc polycarboxylate cements were developed by D. C. Smith in 1964 and became commercially available in 1968 [93]. The phosphoric acid of phosphate cements was replaced by an aqueous solution of polyacrylic acid (20 – 60 %). Different additives to the powder (e.g., fluorides) are used to improve cement properties. Some products contain dried polyacrylic acid in the powder and are mixed only with water. The major new feature of these cements was their chemical adhesion to tooth tissue [93]. This property can be explained by the fact that the carboxylic groups of the polyacid react not only with the zinc ions released from the powder (setting reaction), but also with calcium ions of the tooth structure [94]. Zinc polycarboxylate cements have relatively low mechanical strength but are well tolerated by the pulp. For that reason they are used mainly as luting materials for metallic restorations and base materials.

Trade Names: Durelon (ESPE); Poly F Plus (Dentsply).

4.1.5. Glass Ionomers

History. A. D. Wilson investigated the use of ion-leachable glasses with polyacrylic acid. By increasing the basicity of the glasses, systems with acceptable reactivity were obtained in 1969 [97]. That class of materials was named glass ionomer (glass + ionic polymer) or glass polyalkenoate. The first glass ionomer ASPA I (**a**luminum **s**ilicate **p**oly**a**crylate) was used clinically in 1970 by J. W. McLean. Due to a short working time and a very slow setting, the workability was restricted [98].

Composition. All glass ionomers contain three essential components: an ion-leachable glass, a polycarboxylic acid, and water.

The first glasses were calcium aluminum fluorosilicate glasses (with small amounts of sodium and phosphorus), melted from the same raw materials as the glasses for silicate cements. A typical composition (G 200) is: SiO_2 30.1 %, Al_2O_3 19.9 %, AlF_3 2.6 %, CaF_2 34.5 %, NaF 3.7 %, $AlPO_4$ 10.0 % [2]. After introduction of the sandwich technique (1985), in which dentin is replaced by glass ionomer, and enamel by composite resins, the need for radio-opaque

glass ionomers became evident [98]. Modern glasses for glass ionomers contain heavy elements such as strontium, barium, or lanthanum. For the reactivity of glasses the content of aluminum and the formation of a glass matrix containing Si–O–Al bonds is important [99].

For extended use in dentistry an adjusted working and setting time is needed. Since these requirements are only partially realizable with the glass structure, the setting characteristic has to be adjusted by different techniques. Temperature treatment of the glass powders prolongs the working time [98], as does the reduction of reactive ions on the particle surface by a pretreatment with dilute acids [100]. Addition of tartaric acid as chelating agent prolongs the working time and accelerates the setting reaction [101]. Most modern glass ionomers contain tartaric acid as a setting modifier.

In the first glass ionomers, an aqueous solution of polyacrylic acid was used. Many commercially available products still contain such solutions, even though they tend to form gels at lower temperatures. Copolymers of acrylic and itaconic acid [98] or of acrylic and maleic acid [102] are used to overcome this problem and to increase reactivity. In some products the high ratio of maleic acid in the copolymer leads to a higher degree of cross-linking in the material, resulting in high mechanical strength. The use of polyvinylphosphonic acid [103] or copolymers of acrylic and vinylphosphonic acid [104] has also been reported. The polycarboxylic acids used in glass ionomer formulations have a molecular mass in the range of 10 000 – 100 000 and are used in concentrations of about 20 – 60 % (depending on the different products). Depending on the indication the powder : liquid ratio can vary in a wide range of about 1.5 : 1 to 4 : 1. All glass ionomers contain water, which is the reaction medium and the only liquid component for paste formation.

Setting Reaction. After the attack of the polyacid on the glass, cross-linking of polycarboxylate with metal ions released by the glass takes place. A calcium aluminum polycarboxylate matrix is formed, in which glass particles covered by silica gel and water molecules are embedded. As a result of ion hydration, the content of evaporable water decreases with time [105], while the cross-linking reaction still slowly continues. Both effects lead to an increase in mechanical strength and Young's modulus [106] and reduced wear [107] in the course of time.

Further Developments. In a search for improved wear resistance the cermet (**cer**amic **met**al) ionomers were developed in 1985 [108]. In these materials silver is sintered with the glass particles. In other products silver is added as a powder component.

By the use of dried polyacid in the powder setting time was prolonged. An improvement in handling characteristics was obtained by the introduction of capsules containing predosed components which are mixed mechanically.

Recently, highly viscous glass ionomers were developed. These materials have high strength, high wear resistance, and low solubility [109 – 113].

Properties. Glass ionomers adhere chemically to the tooth structure by the same mechanisms as zinc polycarboxylate cements. Due to the possibility of reformation of broken bonds during the remodeling of dentin, the bond has a dynamic character. In a humid environment neither shrinkage nor swelling is observed. Fluoride release [114 – 116] and recharge (e.g., from toothpastes) [117] explain why secondary caries are only rarely observed near glass ionomer fillings. Fluoride and calcium release promote dentin remineralization [118]. The very good biocompatibility of glass ionomers, demonstrated in many tests [119], led to their use as bone cements and bone replacement materials in orthopedic and ear – nose – throat surgery with good results [120 – 123].

Uses. Beside the major use as filling, luting, and base or core buildup materials [98], further indications are endodontic filling [124], fissure sealant [125], orthodontic luting material [126], filling material for open furcations [127], and retrograde root-filling material [128].

Trade Names: Ketac-Fil Plus (ESPE), Ketac-Molar (ESPE), Fuji II (GC), Fuji IX (GC) (filling materials); Ketac-Cem (ESPE), Fuji I (GC) (luting materials); Ketac-Silver (ESPE), Miracle Mix (GC); HiDense (Shofu) (Cermets); Ketac-Bond (ESPE) (base material); Ketac-Endo (ESPE), Endion (Voco) (root-filling materials).

4.1.6. Resin-Modified Glass Ionomers

In an attempt to combine the advantages of glass ionomers with those of composites, monomers and initiators were added to glass ionomers. Thus, parallel to the glass ionomer reaction a polymerization reaction can be started by photochemical or redox initiation [129]. Methacrylated polycarboxylic acids are also used [130]. Often these materials are called light-curing glass ionomers; the general term resin-modified glass ionomer cements has been proposed for this class of materials [131]. An interpenetrating network of ion-cross-linked polycarboxylate and poly(meth)acrylates is formed. To obtain compatibility of water-based glass ionomers and resins, all materials contain hydroxyethylmethacrylate and other hydrophylic monomers [93]. Due to the hydrophilicity, a relatively high swelling is observed. For this reason resin-modified glass ionomer cements are not recommended for cementation of all-ceramic restorations (cracking of such restorations was observed) [132 – 134]. Resin-modified glass ionomers have higher strength and better esthetics than conventional glass ionomers but a relatively low wear resistance. Depending on the amount of resins in the structure, these materials can be self-adhesive. Due to the ease of handling they are largely used in pediatric dentistry.

Trade Names: Photac-Fil Quick (ESPE), Fuji II LC (GC), Vitremer (3M) (filling materials); Fuji Plus (GC), Vitremer Luting (3M) (luting materials)

4.1.7. Polyacid-Modified Composites

Conventional and resin-modified glass ionomers contain all three essential components of a glass ionomer, which react spontaneously when they come into contact. In search of light curing "one-component-glass ionomers", water-free systems were formulated. The glass ionomer reaction takes place only on water absorption in the oral environment [135, 136]. These materials are known as polyacid-modified composites [131] or compomers (**comp**osite ion**omers**). These materials are composite resins containing ion-leachable glass fillers, polymerizable acids, polycarboxylic acids, and hydrophilic monomers that are compatible with the acids and facilitate water absorption for the glass ionomer reaction [93]. An important property is the ease of handling. Strength and esthetics are better than those of resin-modified glass ionomers and comparable to those of composites. However, they have no self-adhesion and only low fluoride release compared to glass ionomers. They shrink comparably to composites and show the same tendency to swell as resin-modified glass ionomers [132 – 134]. Resin-modified glass ionomers and polyacid-modified composites are often referred to by the collective term "glass ionomer composite hybrids" (or simply hybrids). For both classes the most important indications are in fillings, the luting of metallic restorations, or as a liner or base [94].

Trade Names: Dyract (Dentsply), Compoglass (Vivadent), Hytac (ESPE) (filling materials); Dyract Cem (Dentsply) (luting material).

4.2. Cements Based on Complex Formation

These cements harden by the reaction of a metal oxide, hydroxide or salt with a chelating agent. Chelates form a matrix in which the inorganic particles are embedded. These materials generally have low mechanical strengths and therefore their use is limited to temporary fillings or lutings, pulp capping materials or root fillings.

4.2.1. Zinc Oxide – Eugenol Cements

Zinc oxide – eugenol cements (ZOE cements) contain zinc oxide and eugenol (the principal component of oil of cloves) as essential components. In two-paste systems, zinc oxide is suspended in a vegetable or mineral oil, and eugenol is thickened with resins. Small amounts of water or acetic acid are added to initiate and accelerate the reaction. The strength of these materials is low, and their solubility is high compared to other cements. Depending on the powder/liquid ratio or paste formulation, temporary restorations, luting cements, or root canal sealers can be prepared. Similar systems are used as impression materials [94]. In some formulations part of the eugenol is replaced by *o*-ethoxybenzoic acid. These EBA cements are stronger. They are used for the same indications as ZOE cements [94].

Due to the phenolic character of eugenol, the use of certain resin-based luting materials after temporary cementation with ZOE or EBA cements is problematic (inhibition of polymerization by phenols).

Trade Names: Temp Bond (Kerr), Scutabond NF (ESPE).

4.2.2. Calcium Hydroxide Cement

First attempts to use calcium hydroxide in dentistry are dated from the late 1700s. Calcium hydroxide has an antimicrobial effect and induces the formation of secondary dentin. The calcium hydroxide has been used in pulp capping since 1859 [137]. Cements contain salicylate chelating agents that react with calcium hydroxide (e.g., methyl salicylate or butanediol-disalicylate) [138, 139].

Trade Names: Dycal (Dentsply), Alkaliner (ESPE), Life (Kerr).

4.2.3. Other Chelating Cements

Various chelating agents are used in dentistry, especially together with zinc oxide. Root canal filling materials containing a 1,3-diketone and eugenol-free temporary cements containing liquid carboxylic acids are known.

Trade Names: Diaket (ESPE) (root filling material); Temp Bond NE (Kerr), Freegenol (GC), Procem (ESPE) (temporary cements).

4.3. Standards

The international standards for dental cements are:

- ISO 3107, Dentistry – Zinc oxide/eugenol and zinc oxide/non eugenol cements
- ISO 6876, Dental root canal sealing materials
- ISO 9917-1, Dentistry – Water based cements – Part 1 Powder/liquid acid base cements
- ISO 9917–2, Dentistry – Water based cements – Part 2 Light activated cements

5. Synthetic Resins

Synthetic resins are of great significance in dentistry. Polymers of the polymethacrylate group have been in use since around 1934, and they are still used today [36].

5.1. Denture Base Resins

The first organic base material for dentures was hard rubber. Pink India rubber was vulcanized in a water-filled autoclave at ca. 140 °C. This denture base material was used in combination with porcelain teeth. The dental prostheses produced in this way were functional, but not esthetically pleasing. The India rubber procedure is of only historical interest today.

The basic process of denture fabrication has remained unchanged. A stone (gypsum) model of the jaw is prepared from an impression of the patient's oral cavity. The original shape of the denture base is molded in wax on the stone model. The artificial teeth, either of porcelain but today mainly of resin, are mounted on this wax rim. The wax denture on the stone model is then invested with gypsum in the lower half of a denture flask. Subsequently, the other half of the flask, open on the upper side, is attached and the wax model is embedded completely. The gypsum on top of the upper half of the mold is planed by pressing with a lid. After the plaster has set, the two halves are separated and the wax is removed by melting. The artificial teeth remain fixed in the upper half of the flask. In this way, a precisely fitting cavity is obtained in which raw, unvulcanized rubber or the resin mixture to be polymerized is pressed and hardened.

Insulation of the stone plays a decisive role in this process, because the resin must be kept separate from the plaster. Initially, dental technicians used ductile tinfoil, which could be adapted with perfect precision to the teeth on the free plaster surface of the model. Tinfoil also seals the doughy mass to be polymerized from the steam that can evolve from hot plaster. More conveniently, a coating consisting of a ca. 3 % solution of sodium alginate in water can be applied to the plaster surface. It reacts with the plaster surface to give a water-insoluble calcium alginate film.

The decisive breakthrough in denture base resins was the development of poly(methyl methacrylate) by Röhm. Roth, a dental technician, mixed ground poly(methyl methacrylate) with the methyl methacrylate monomer; this gave a dough that could be used to make a functional and aesthetically satisfactory dental prosthesis, as described above [37]. The polymer – monomer mixture solved two problems of methyl methacrylate polymerization simultaneously: (1) the pure monomer shrinks by ca. 20 % during polymerization and (2) it simultaneously releases a large amount of heat (56.6 kJ/mol).

Poly(methyl methacrylate) [*9011-14-7*] has a dominant position in dental resins; it meets best the following requirements of a denture material:

- Absence of taste or odor, even after long periods of wear
- Good tissue compatibility
- Good mechanical properties
- Exact reproduction of details of the model
- Dimensional stability
- Low water sorption
- Color stability
- Good aesthetic appearance
- Simple manipulation (no need for expensive equipment)
- Good reparability
- Ease of cleaning

Methacrylates are available as powder and liquid. The polymer powder, today usually bead polymer, is mixed with the liquid monomer in a weight ratio of (2.5 – 3):1. After a certain swelling time, doughs are obtained which can easily be manipulated by pressing, pouring, or molding.

Mineral pigments are the principal coloring agents used; organic dyes are less important. To avoid formation of stress cracks and improve mechanical properties, the liquid monomer generally contains up to 10 % of a cross-linking agent. Ethylene glycol dimethacrylate [*97-90-5*] is preferred. Depending on the polymerization initiator used, four types of materials can be distinguished: heat-, cold-, light-, and microwave-cured polymers.

5.1.1. Heat-Cured Polymers

The procedure developed by Roth is called the "*Paladon technique*" in dental laboratory technology. In heat-cured polymers, 0.1 – 0.5 % dibenzoyl peroxide [*94-36-0*] is generally used as the free-radical initiator.

After the resin dough has been added to the insulated plaster form and pressed, the form is placed in a hot water bath. Dibenzoyl peroxide decomposes between 80 and 100 °C and initiates the radical chain reaction. If heating is too rapid, polymerization proceeds too vigorously and a porous product is formed due to boiling of the monomer (*bp* 100.2 °C at 101.3 kPa). Therefore, the initial heating process should be held at 70 – 75 °C for 30 min; then the water bath is allowed to boil for an additional 30 min.

The resin can be added to the hollow mold by injection or by packing the negative mold and applying pressure with a pneumatic press. In the packing and pressing procedure, excess resin is placed in the denture flask and the two molds are joined under pressure; excess material is squeezed out between the casts. The flask is then opened, and squeezed out excess material is cut off. The closed flask is then placed in the water bath. Better fit can be obtained by using the injection technique [38]; in this case, the resin is injected under high pressure into the closed flask. The flask is then placed in the water bath and held under pressure until polymerization is complete.

Trade names: Paladon 65 (Heraeus Kulzer, Germany); SR 3/60, SR Ivocap (Ivoclar, Liechtenstein); Lucitone (Dentsply, USA).

5.1.2. Cold-Cured Polymers

In addition to the heat-cured polymers described above, cold-cured polymers (also called autopolymers) have increased in significance [36]. In cold-cured polymers, the polymerization is initiated by a two-component redox system, one component of which is in the powder and the other in the liquid. Such systems are usually composed of tertiary amines, e.g., *N*,*N*-dimethyl-*p*-toluidine [*99-97-8*] (1 – 2 %), and peroxides, particularly dibenzoyl peroxide (0.5 – 2 %) which has been used worldwide since about 1940 [39]. The addition of UV stabilizers is customary to reduce discoloration.

When the components of the redox system are combined, numerous radicals form which activate polymerization. However, the formation of inactive byproducts leads to a yellow-brown discoloration of the resin. Initiator systems that do not produce discoloration were developed in 1956 [40]; they consist of a 5-substituted barbituric acid combined with dibenzoyl peroxide or heavy-metal salts. The color stability of cold-cured polymers having this type of initiator system is so high that the quality of the polymer reaches that of heat-cured materials.

Cold-cured polymers are manipulated like the heat-cured polymers, but the ratio of monomer to polymer is higher (10 : 4 to 10 : 7), and polymerization takes place in the closed flask at room temperature in 10 – 30 min. If no injection pressure is used, i.e., in the completion of cast dentures with resin (partial dentures) and in the fabrication of complete cast dentures, polymerization is carried out in a pressure chamber at 0.2 – 0.3 MPa and 45 – 55 °C to avoid porosity of the product.

Cold-cured polymers are preferred today for repairing fractured prostheses and for relining poorly fitting ones. The cold-cured resins are also used almost exclusively for the completion of cast partial dentures.

For complete dentures, cold-cured polymers are injected. While heat-cured polymers show certain strain after cooling, which arises from differences in expansion coefficients between plaster and resin, cold-cured resins polymerize practically free of strain. This results in a better fit of prostheses fabricated from cold-cured resins.

Trade names: Palapress, Paladur (Heraeus Kulzer, Germany); SR 3/60 Quick (Ivoclar, Liechtenstein), Degupress (Degussa-Hüls, Germany).

5.1.3. Light-Cured Polymers

The first denture resin that was polymerized by the action of light from a halogen source was introduced in 1984 (Triad, Dentsply). The resin material is supplied in a ready-to-use form which eliminates the need for mixing in the dental laboratory.

Instead of methyl methacrylate, diesters of methacrylic acid with high molecular mass diols similar to bis-GMA (see Section 5.6) or high molecular mass urethane dimethacrylates are used as monomer.

The handling technique has been adjusted because complete investment in plaster is no longer possible, since the light must have access to the resin from all sides. For this reason, for example, pontics are provided in prefabricated standard sizes. A more detailed description of light curing is given in Section 5.6.2 on resin-based filling materials.

Trade name: Triad VLC, Eclipse (Dentsply, USA).

5.1.4. Microwave-Cured Polymers

The reaction of microwave-cured polymers is almost the same as for heat-cured materials. The main difference is that the flask is not placed in a water bath and heated to boiling temperature. Instead, the water contained in the gypsum in the flask is boiled by microwave energy. Since metal flasks are impermeable to microwaves, flasks for microwave-cured polymers are made from fiber-reinforced polyester. The initiator system must be adjusted to the polymerization process. Therefore only materials specifically designed for microwave curing can be used for this polymerization procedure.

Trade name: Acron MC (GC, Japan).

5.1.5. Other Denture Resins

Soft resins are used as liners of dentures. In the past, plasticized acrylic or methacrylic acid esters or esters of higher alcohols were used; however, they are not sufficiently durable. If made of silicones, the liners have better durability but insufficient bonding to the denture base material.

Many attempts have been made to replace methyl methacrylate as the denture base material by other plastics [41, 42]. *Polyamides* and *polyurethanes,* practically indestructible resins in the fabrication of dental prostheses, have failed because of their insufficient rigidity and their excessive water sorption. At one time, *polycarbonates* (e.g., Thermpont, Andoran, and Copodan) appeared to be alternative resins because of their high impact strength. However, their dimensional stability did not meet clinical

requirements [41]. The high sensitivity of these materials to the steam evolved during injection into the mold, was responsible for alterations in shape and color. Furthermore, polycarbonates are sensitive to alkali which is present in many denture cleansers and leads to the formation of strain, tears, and denture fractures. The special injection techniques required and the associated expensive devices have also contributed to the elimination of polycarbonates from the denture base market.

Resins based on *poly*(*vinyl chloride*) [*9002-86-2*] appear to have a better chance for use as denture materials. Copolymers of vinyl chloride, vinyl acetate, and methyl methacrylates (e.g., Luxene and Virlene) are used because of their constancy of shape, low water sorption, and high fracture resistance. They can only be handled in licensed laboratories and require relatively complicated and expensive special apparatus; their market share has, therefore, remained minor.

5.2. Acrylic Teeth [43]

Most artificial teeth set up in a denture base resin are acrylic teeth, made from poly(methyl methacrylate) and methyl methacrylate. The early problems of poor shade composition, shade durability, crazing resistance, and particularly, abrasion resistance have been solved. Today, resin teeth can replicate natural teeth perfectly. The resin tooth is built up from cervical, dentin, and incisal resins, each of which is colored to correspond to the appropriate tooth shade.

The layers are pressed into metal molds, one after the other, and heat-polymerized under pressure. To give greater hardness to the outmost (incisal) layer for greater abrasion resistance, ca. 10 % of a cross-linking agent, preferably ethylene glycol dimethacrylate [*97-90-5*], is added. The cervical material is not or only weakly cross-linked, because it then absorbs the monomer of the denture base resin dough more readily, which provides for better bonding of the acrylic tooth to the denture base.

Despite its high abrasion resistance, poly(methyl methacrylate) is not entirely satisfactory for teeth in the posterior region. The mechanical properties and the wear resistance of the resin can be increased enormously when it is filled with 30 – 60 % of very small (mean grain size 0.04 μm) particles of hydrophobic (fumed) silica (e.g., Aerosil, Degussa).

Trade names: SR-Orthosit, SR-Vivodent PE (Ivoclar, Liechtenstein) Vitapan (Vita Zahnfabrik, Germany).

5.3. Crown and Bridge Veneer Resins [44]

Resins are also used to veneer the visible parts of metal crowns or bridges. Today, polymers of high molecular mass methacrylic acid diesters are used almost exclusively for this purpose and for jacket crowns. The abrasion resistance of the methacrylates can be increased by adding inorganic fillers such as very fine grained silica, and fine grained glass or quartz (particle size < 1 μm)

5.3.1. Heat-Cured Resins

Heat-cured resins provide better results than cold-cured polymers for crowns and bridges. They are available as combinations of powder and liquid or as ready-to-use pastes. The resin pastes are molded directly on the metal framework by the dental technician and are then polymerized in a pressure vessel at approximately 0.6 MPa and 110 °C. Steam, water, or glycerol is used as the heat-transfer medium.

Trade names: SR-Chromasit (Ivoclar, Liechtenstein); Biodent K+B Paste (Dentsply, Germany); Vita Zeta (Vita Zahnfabrik, Germany).

5.3.2. Light-Cured Resins

A breakthrough, particularly in processing, occurred when light-cured crown and bridge veneer resins were introduced [45]. These materials consist of ready-to-use single-paste resins, which eliminates the need for mixing. Polymerization is induced by irradiation with light (wavelength 300 – 500 nm). An aromatic ketone or camphoroquinone is employed as photoinitiator; tertiary amines are frequently used as reaction promoters. Light-initiated polymerization can provide resins that are fully equal in quality to heat-cured ones.

Trade names: Signum + (Heraeus-Kulzer, Germany); Sinfony (Espe, Germany); Cristobal + (Dentsply, USA).

5.4. Resin – Metal Bonding

The question of adhesion between metal and resin is of great significance for denture resins and for crown and bridge veneer resins. Unfortunately, without special bonding processes, gaps form between metal and resin in cast prostheses after short periods of use. Bacteria and food debris can settle in these gaps, leading to discoloration at the metal – resin interface. The same problem arises in crowns and bridges. Formerly, the resin could be held to the metal framework only with mechanical devices (metal beads or wires), which may be visible through the resin.

In 1984, the *Silicoater technique* was developed, which solved the problem of adhesion between metal and resin for all precious and nonprecious alloys. In this procedure, a silane is pyrolyzed in a flame and the products are deposited on the metal framework, where an SiO_xC layer of 0.1 µm thickness forms. On this surface, another silane, such as 3-methacryloxypropyltrimethoxysilane [*2530-85-0*], can simply be painted. If the methacrylate monomer is then applied, it reacts with the silane. The metal-to-resin bond achieved in this manner is stronger than the self-tear resistance of the resin. The bond remains stable in the oral milieu [47, 48].

Another process of silane coupling between metal and resin is the Rocatec technique. In this process, introduced in 1989, the surface energy of the metal is increased by two-step sandblasting with aluminum oxide. The first step uses pure aluminum oxide, and the second step with aluminum oxide coated with a special glass, part of which remains on the metal surface after impact. A silane is then applied to the treated surface.

Other systems apply hydrophobic fluorinated methacrylates to the surface (Sebond-MKV), use a molecule containing an acidic group (SR-Isosit-Spectra-Link) or use tin plating and oxidation of the tin layer on the surface (OVS). Another method is bonding with 4-methacryloxyethyl-trimellitone-hydrid (4-META).

Trade names: Silicoater (Heraeus-Kulzer, Germany). Rocatec (Espe, Germany); Sebond-MKV (Schütz Dental, Germany); SR-Isosit-Spectra-Link (Ivoclar, Liechtenstein)

5.5. Resins for Orthodontics

Orthodontic treatment is used to correct the position of teeth. For removable orthodontic devices, cold-cured polymers, such as those described in Section 5.1.2, are used. For fixed orthodontic appliances, resins very similar in chemical composition to those used in cold-cured restorative materials are employed as adhesives.

Trade names: Orthocryl (Dentaurum, Germany); Super C Ortho (Amco, USA).

5.6. Restorative Resins

In the past, only silicate restorations and expensive fired porcelain inlays were available to provide "invisible" tooth restoration. Silicate cement fillings have insufficient durability because their surfaces wash out and they do not esthetically satisfy.

Soon after their discovery, cold-cured poly(methyl methacrylate) resins (Section 5.1.2) were also used for restorative fillings. However, the presence of amine – peroxide catalysts in the resins leads to a bright yellow discoloration after service periods of only a few months. Similarly, the abrasion resistance of the restoration is not sufficient. In addition, the high concentration of unconverted monomer remaining in the resin can provoke pulp irritation. Therefore, these resins have been replaced by the new filler resins described in the following paragraphs.

Bowen transferred the structural principle of an epoxy resin to a methacrylate material and thus gave the latter the excellent material properties of filled epoxies [36]. The key chemical in this method is the so-called Bowen monomer (bis-GMA)

which is made from bisphenol A di(2,3-epoxypropyl) ether [*1675-54-3*] and acrylic acid [*79-10-7*]: Bowen also proposed using a mixture of 70 % ground quartz glass (average particle size approximately 15 µm) and 30 % of bis-GMA [36, 49] . Not only is the composite made in this way much more rigid and harder than polymers of methyl methacrylate, but it also shrinks less during polymerization because of

the greater molecular mass per double bond. Today monomer mixtures of bis-GMA and triethylene glycol dimethacrylate [*109-16-0*], as well as other dimethacrylates which act as diluents and certain urethane dimethacrylates, are used. In addition, the high content of inorganic fillers greatly decreases the thermal expansion coefficient and thus approaches that of the natural tooth.

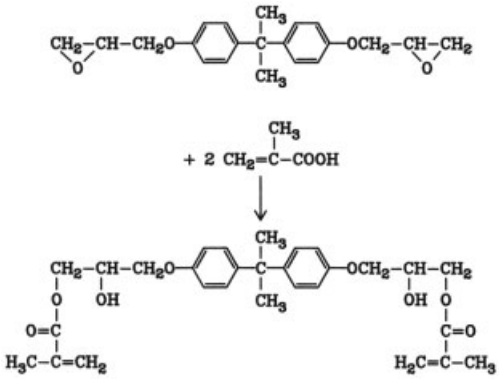

Bis-GMA [*1565-94-2*]

The important problem of matrix-to-filler bonding is solved by treating the inorganic filler material with special silanes [50, 51], e.g.:

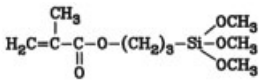

3-Methacryloxypropyltrimethoxysilane [*2530-85-0*]

The silanol groups of the hydrolyzed silane react with the hydroxyl groups of the silica filler, and its methacryl groups enable copolymerization with the matrix resins to occur, thus providing a chemical bond between matrix and filler.

The structure and particle size of the inorganic filler material have also been modified. The relatively coarse-grained fillers of the first composite generations (especially quartz with a particle size of 15 – 30 µm) did not permit glaze polishing of the external surface of the restoration. The use of fumed silica with a particle size of 0.01 – 0.04 µm led to restorative materials whose surfaces were homogeneous and amenable to glaze polishing. These "microfilled" composites contain about 50 vol % fillers.

In a more recent development, glass particles (< 1 µm; barium or strontium silicate glass) are mixed with the fumed silica. Such composites, called hybrids, are characterized by particularly good mechanical properties, homogeneous surfaces, and good polishability. These materials contain ca. 10 wt % fumed silica and ca. 70 % glass.

Composite restorative materials are no longer limited to the anterior region. Nowadays they have the required strength and abrasion resistance for use on occlusal surfaces.

5.6.1. Paste – Paste Systems

The composites whose chemical composition is described in the previous section have been available as two-paste systems: base paste and catalyst paste [36, 50, 51]. Nowadays there are only a few materials of this kind available on the market.

The tooth-colored base paste contains a tertiary amine (1 – 2 %), and the catalyst paste contains dibenzoyl peroxide (1 – 1.5 %). The tertiary amine usually is *N,N*-bis(2-hydroxyethyl)-*p*-toluidine [*3077-12-1*]; it does not have as great a tendency for discoloration as the *N,N*-dimethyl-*p*-toluidine used earlier. However, UV stabilizers must still be added. The catalyst pastes have a strictly limited shelf life of approximately one year and must be stored in a refrigerator.

Before placement, the base and catalyst pastes are mixed in a 1 : 1 ratio. Depending on the particular product, the working time ranges from 1 to 2 min and the setting time from 3 to 5 min. Disadvantages of the paste – paste system are the severely limited shelf life of the peroxide paste and the still-unsatisfactory color stability. Furthermore, air bubbles are generated in the pastes during mixing which severely affect the quality of the restoration.

5.6.2. Light-Cured Polymers

Light-cured polymers do not have the disadvantages of the paste – paste systems described in the previous section. The light-cured composites are almost identical in composition to those materials described in Section 5.6. However, they consist of only one component that contains a photoinitiator system.

The advantages of light-cured composites over the paste – paste systems are

Immediate readiness for use
Fewer air bubbles because mixing is not required
Prolonged working time
Better color stability
Much longer shelf lives (three years)
No need for cool storage
Much higher reaction rate

The first radiation-cured composites were polymerized with UV radiation (320 – 440 nm) in the presence of an aromatic ketone initiator such as 1,2-diphenyl-2,2-dimethoxyethanone [*24650-42-8*]. At an irradiation time of 20 s, a maximum depth of cure of only 1.5 mm was attainable. Nowadays, visible light (380 – 500 nm) is used with camphoroquinone [*10373-78-1*] as initiator and certain aliphatic and aromatic amines as polymerization promoters, and only common halogen lamps (150 – 250 W) or blue LED lights are required. In addition, the risk of UV radiation to the dentist's and patient's eyes disappeared. The most important advantage of polymerization with visible light, however, is the much greater cure depth attained. Depending on the composite and the light source, curing depths of 4 – 8 mm can be attained after 20 – 40 s irradiation time. Fiberoptics conduct the light to the site at which radiation is required.

Trade names: Charisma, Soltaire 2 (Heraeus-Kulzer, Germany); Artemis, Tetric Ceram (Vivadent, Liechtenstein); Prisma TPH, Surefil (Dentsply, USA); Beautifil (Shofu, Japan); Clearfil posterior (Kuraray, Japan); Prodigy (Kerr, USA); Filtek (3M Espe, USA).

5.6.3. Ormocer-Based Filling Materials

In this type of polymer-based filling material, the main component of the monomer matrix is an organically-modified-ceramic (ormocer). This is a cross-linked silicone oligomer bearing methacrylate groups. Dimethacrylates are used as diluents. Initiators and fillers are like in the light-cured polymers (Section 5.6.2). Advantages of this type of material are:

- Lower polymerization shrinkage because of the size of the oligomer molecule
- Lower dissolution rate of residual monomers because of the hydrophobic nature of the molecule, and a higher probability of each monomer molecule being incorporated in the polymer network because of the multifunctionality
- Higher abrasion resistance compared to materials with the same type and size of filler particles

Trade names: Definite (Dentsply De Trey, Germany); Admira (Voco, Germany).

5.6.4. Sealant Resins

Fissures of the posterior teeth are potential sites at which carious lesions can arise. However, when these fissures are sealed with thin layers of bis-GMA or related compounds, the occlusal surfaces can largely be protected from caries. The sealant resins, like restorative resins, are available as cold-cured and light-cured systems.

Trade names: Estiseal F (Heraeus-Kulzer, Germany); Delton (Dentsply, USA).

5.6.5. Resin – Tooth Bonding

Adhesion of the resin to the tooth is very important because otherwise restorations have short lifetimes, and marginal gaps are formed between the restorative resin and the tooth [49, 51, 52]. These gaps can be sites of secondary caries. In general, no bonding occurs between tooth and resin, and the dentist must provide mechanical anchorage to ensure retention. However, when mechanical retention is used, a relatively large amount of sound tooth tissue is lost.

The *acid-etch technique* has been used to provide bonding of *enamel* to composite resins. The dental enamel surface was first roughened with an approximately 35 % solution of phosphoric acid so that microretentions result. A thin layer of resin of the bis-GMA type was then applied to the etched surface; the monomer penetrates the pores etched into the enamel and is firmly anchored there. Finally, the restorative resin was applied, and a mechanical bond formed between dental enamel and resin.

Adhesion to dentin can not so easily be achieved. It took a while until it was realized that chemical bonding is not the way to solve the

problem. With modern materials a smear layer that is formed on the dentin surface during cavity preparation is dissolved in an acid. Part of the inorganic material of the dentin is dissolved, too. Collagen fibers extending into sound dentin remain on the surface. The dissolved material is washed away or remains in the adhesive. The adhesive penetrates the collagen and a part of the dentin tubuli to form a hybrid layer. The methacrylate groups of the monomers in the adhesive provide chemical bonding to the filling material. Since moisture is always present at the dentin, surface monomers in the adhesive have to be hydrophilic.

Trade names: Scotchbond (3M Espe, USA); Clearfil bond (Kuraray, Japan); Dentin Adhäsit (Ivoclar, Liechtenstein); Syntac (Vivadent, Liechtenstein) Prime & Bond NT (Dentsply, USA);

5.6.6. Toxicology

The toxicology of restorative resins is an important international issue, especially because of the relatively high pulp toxicity of methyl methacrylate. For this reason, new monomers with high molecular mass have been developed, e.g., bis-GMA (LD_{50} 14.15 mL/kg) which is not orally toxic and does not irritate skin or mucous membranes.

Studies on the biocompatibility of restorative resins have increased enormously in the past few years. A series of important components of restorative resins has been examined for tissue compatibility, and none of the substances examined caused irritation [46].

5.7. Standards

The following international standards on polymer based materials apply:

- ISO 1567: Dentistry — Denture Base Polymers
- ISO 4049: Dentistry — Polymer-based filling, restorative and luting materials
- ISO 6874: Dentistry — Polymer–Based Pit and Fissure Sealants
- ISO 6876: Dental Root Canal Sealing Materials
- ISO 7491: Dental Materials — Determination of Color Stability
- ISO 10139–1: Dentistry — Soft lining materials for removable dentures — Part 1: Materials for short-term use
- ISO 10139–2: Dentistry — Soft lining materials for removable dentures — Part 2: Materials for long-term use
- ISO 10477: Dentistry — Polymer-Based Crown and Bridge Materials
- ISO 22112 Dentistry — Artificial teeth for dental prostheses

6. Impression Materials [53 – 61]

According to their handling (chemoplastic or thermoplastic) and their properties after setting (elastic or hard, reversible or irreversible), impression materials can be grouped as follows:

Chemoplastic
- *Elastic, irreversible*
  - alginate
  - elastomers
- *Hard, irreversible*
  - impression gypsum
  - ZOE paste
  - polymer

Thermoplastic
- *Elastic, reversible*
  - Hydrocolloid
- *Hard, reversible*
  - impression compound
  - wax – resin
  - gutta-percha

The oldest known impression material is plaster of paris (Dwinelle 1840). Dental compound (Stent 1840) and gutta-percha (1850) were developed for use as impression materials in the mid-1800s. Hydrocolloids (agar) were introduced in 1927 (Poller), zinc oxide and eugenol pastes in 1935 (Kelly); alginates in 1940 (Wilding); polysulfides and condensation silicone resins in 1955; polyether resin in 1964 (Espe); and addition silicone resins in 1976.

The principal requirements for impression materials are:

- Biocompatibility for patient and user
- Physical and chemical properties in accordance with the purpose of impression taking

(precision of the dies, recording of the function): Flowability during impression taking, detail reproduction, mechanical strength, dimensional stability of the impression, compatibility with the die material
– Efficient workability to optimal quality (lot consistency, storability, ease of dosing and mixing), hygienic aspects (disinfectability), ergonomically favorable methodology, economy

The dentist needs a large choice of impression materials, because they fulfill the requirements to different degrees. For each task a certain impression material is best suited. The choice of an impression material is mainly governed by the required precision of the die (documentation, type of dental restoration), the object to be replicated (shape and condition), and the method of impression taking. When equally good results can be obtained with different impression materials, ergonomic and economic aspects are decisive.

6.1. Elastic Materials

An important advantage of elastic materials is that it is possible to remove them from the object from which the impression is taken without breakage or extreme deformation, even when they fill undercuts.

6.1.1. Alginates

Alginates (see also → Polysaccharides) are derived from alginic acid, a polyuronic acid composed of up to 750 units of D-manonuronic acid and L-guluronic acid 1,4-glycosidicly linked. It is contained in the marine algae in the amount of 40 % (dry substance). It is almost insoluble in water but has a high ability to swell by the absorption of water.

Dental alginates are powder mixtures with the following composition: sodium alginate (10 – 20 wt %), calcium sulfate (10 wt %), trisodium phosphate (1 – 4 wt %), diatomite (70 – 80 wt %), color and flavor (1 – 2 wt %). Mixing with water gives a plastic mass which sets by chemical reaction to give an elastic hydrogel. Monovalent soluble salts of alginic acid (Na, K, NH_4 alginate) are precipitated as sparingly soluble Ca salts. Since 1980 no lead compounds are contained in alginate, and low-dust and no-dust versions are available. The reaction is prolonged to a practical duration by a retarder (trisodium phosphate) which initially reacts with the Ca ions. Fast- and normal-setting preparations are available. The filler (diatomite) influences the plasticity of the mixture as well as the strength and elasticity of the impression. The plasticity of the mixture can be influenced by the powder/liquid ratio, and the working time by the water temperature. Dosing is performed manually, and mixing manually or in a mixer. The hydrogel ages (syneresis) and is sensitive to environmental conditions. Storage in air leads to a significant shrinkage of the impression. Disinfection is possible if well-proven recommendations are followed. The die must be manufactured as soon as possible, preferably 15 min after removal of the impression from the mouth and not later then 60 min after removal if stored in a humidor.

Plastic deformations result from pulling over strongly curved parts of the impression. Thin bridges of the impression material tear off. Detail reproduction requires pressure during impression taking and is not as good as with reversible hydrocolloids or elastomers.

Handling is simple and cheap. Alginates are used in many ways: to obtain dies for documentation, for therapy planning, removable dentures and removable orthodontic devices. They are not suited for fixed restorations (crowns and bridges).

A 3 % aqueous solution of alginate is used to produce insulating layer on gypsum dies in dental polymer technology.

6.1.2. Elastomers

Silicones. Impression silicones can be differentiated into condensation cross-linked silicones (C-silicones) and addition cross-linked silicones (A-silicones).

In the case of *C-silicones*, organic polysilanols and polydimethylsilanediols are condensed to give macromolecules by cross-linking agents (tri- and tetrafunctional alkoxysilanes) and accelerators (e.g., tin octanoate and dibutyltin dilaurate) with cleavage of volatile byproducts (alcohols). Fillers (10 – 80 wt %) are used to adjust the consistency: extra fine, fine,

medium, heavy, and putty. Dosing and mixing of the base paste and hardener (as liquid or paste) are performed manually. The addition of colorants makes homogeneous mixing easier and allows materials of different consistency to be distinguished when combinations of materials are used.

In the case of *A-Silicones*, polyvinylsiloxanes and siloxanes with terminal SiH groups are additively coupled by a platinum(0) divinyltetramethyldisiloxane complex (Karstedt catalyst). Reactive oligomers are formed as discrete intermediates by transfer of H atoms to give dimethylene bridges without formation of byproducts (hydrosilylation). The polymerization reaction can be disturbed by catalyst poisons (local anesthetics, astringents, traces of sulfur from vulcanized latex gloves, accelerators and cross-linking agents of C-silicones). The consistency can be adjusted from low viscosity to putty by fillers (40 – 70 wt %) and softening agents or paraffin oil (5 – 10 wt %). The products consists of a base paste and a catalyst paste, which are

- Dosed and mixed manually
- Pressed out of a double cartridge by a static mixer (homogeneous mixing)
- Dosed and homogeneously mixed by a machine with a dynamic mixer

Stability of the flowable silicones during application results from their thixotropy (reversible sol – gel transformation).

By addition of small amounts of nonionic tensides, which lower the interfacial tension with respect to water, the wettability of the hydrophobic A-silicones is enhanced. During impression taking, A-silicones with added tensides remain mainly hydrophobic, since the tenside micelles align only during contact with the moist impression object. However, contact to the hydrophilic gypsum during die production is enhanced, and this avoids voids (bubbles) on fine edges of the die. This can be achieved by the treatment of nonhydrophilic silicone with a wetting liquid (tenside in ca. 40 wt % isopropyl alcohol).

A comparison of C- and A-silicones is given in Table 7.

A-silicones can be used for all kinds of impressions in dentistry because of excellent detail reproduction, high elasticity, and superior dimensional stability. They are neutral in taste and odor. They are especially suitable where high precision is needed (fixed restorations). In cases where other, cheaper impression materials fulfil the requirements for precision, A-silicones are not used.

Table 7. Differences between condensation cross-linking silicones (C-silicones) and addition cross-linking silicones (A-silicones)

Property	C-silicones	A-silicones
Flowability/plasticity	depends on consistency	depends on consistency
Detail reproduction	shift of details on shrinkage	high
Plastic/elastic transition	slow	easier to recognize
Resistance to deformation	depends on consistency	depends on consistency
Residual plastic deformation	higher	very low, even under high compression
Change in volume	contraction, depending on consistency, thickness of layer, and storage time	very low contraction
Storage of the impression	limited by shrinkage	indefinite, several die casts possible
Die material	galvanic die production limited	usable for all types
Disinfection	dimensional changes possible	without problems
Costs	low	higher

For silicones different impression methods exist. The highest precision is achieved with a combination of a high-viscosity material (putty) with a low-viscosity material (double-mixing impression, corrective impression, sandwich impression). Pressure during application and a dry impression object are important requirements for a successful impression of the prepared tooth substances. For functional impressions (full dentures) slowly reacting materials are advantageous.

Further uses of A-silicones are:

- Evaluation of the positional jaw relation of the patient (materials with a high Shore hardness)
- Duplication of dies in dental technology with low viscosity materials
- Rebasing of dentures with a soft A-silicone prosthetic material
- Soft A-silicone material in maxillofacial prosthetics.

Polyethers. A poly(ethylene oxide)/poly(tetramethylene oxide) copolymer with terminal aziridino groups (content in the base paste ca. 60 wt %) are linked additively by an organic sulfonium fluoroborate as starter (content in the reactor paste ca. 30 wt %). The consistency of

the pastes is adjusted by fillers and softening agents (triglycerides). Mechanical mixing is the method of choice. Cooling of the pastes results in substantial increases in viscosity and makes mixing difficult. Before use pastes stored in the refrigerator have to be brought to room temperature. Detail reproduction, plastic/elastic transition, permanent deformation, and dimensional stability are the same as those of A-silicones. Resistance to deformation during removal of the impression from the teeth and from the die is higher. Polyethers are hydrophilic. They can be stored in air indefinitely. In aqueous media they exhibit a small time-dependant swelling. The time available for disinfection is limited. Short-term taste disturbance may occur with some patients. The main use is for fixed restorations. Mono-phase impressions and one-time two-phase impressions with individual impression trays are preferred. The die material is gypsum. Galvanic die fabrication by silver plating is possible; copper plating is not recommended. A special polyether material for evaluation of the positional jaw relation is available.

Polysulfides. Polysulfide oligomers with terminal and side-chain SH groups are three-dimensionally linked at the SH groups by lead dioxide or manganese dioxide. An elastomer is created by polycondensation. The base paste consists of 50 – 90 wt % polysulfide oligomer and, depending on the consistency, 10 – 50 wt % filler (TiO_2) and softening agents (neutral oils). The reaction paste normally consists of 77 wt % PbO_2, 3 wt % sulfur, and 20 wt % filler, and softening agents. Newer products contain organic peroxides instead of PbO_2. Common are one-time impressions with two different consistencies and individual impression trays made of resin material. Handling, indication and properties are, at their best, comparable to those of C-silicones. The materials are sticky, biocompatibility is limited (Pb content), and taste and odor are unpleasant. The market share of this impression material is very small nowadays.

Agar hydro gels consisting of 8 – 15 wt % agar (a heteropolysaccharide from brown algae) and 75 wt % water are converted by heat to a flowable hydro sol and set again by cooling. Viscosity, plasticity, and elasticity are reversibly dependent on the temperature. Addition of glycerol (7 wt %) and borax (0.2 wt %) enhance elasticity and strength; fillers (7 wt % kaolin) affect the consistency. Potassium sulfate (0.1 wt %) compensates for the negative influence of borax on the setting reaction of gypsum. Antiseptic additives (thymol) prevent microbial contamination.

Hydrocolloid materials are distributed in tubes (tray material) or in bars (syringe material). They have to be conditioned by stepwise heat treatment (water bath with three tanks): liquefaction of the gel to a sol in boiling water, tempering for storage (62 – 69 °C); cooling to impression-taking temperature (45 – 45 °C). Low-viscosity hydrocolloid is placed on the impression object from a syringe, and then high-viscosity material is applied on top with an impression tray. Special impression trays are necessary that provide mechanical retention for the impression material and are equipped with channels for water cooling.

Detail reproduction of the hydrophilic material is very good, even under low pressure. Permanent plastic deformation is higher than with elastomers, and tear resistance is much lower. The field of use is impressions for fixed restorations (inlays, crowns, bridges). Results equal to those of elastomers are reached with correct handling. Disinfection is possible only with short contact time. Hydrocolloid impressions can not be stored; immediate casting of a die (gypsum) is necessary.

6.2. Rigid Materials

6.2.1. Plaster of Paris

A powder of $CaSO_4$ β-hemihydrate (90 wt %), fillers (alumina, $CaCO_3$, diatomite; 7 wt %), accelerators (K_2SO_4, NaCl; 2 wt %), color, and flavor is mixed with water (1.5 to 2 parts of powder, 1 part of water) and hardens to an irreversible rigid impression within 3 – 5 min. Flowability and plasticity can be satisfactorily adjusted by means of the water content and depend on the working temperature. The reaction is exothermic. Plaster expands slightly (max. 0.15 % linear). Detail reproduction is very good as long as no small parts of the impression break off and get lost. In the case of undercuts, the hardened material must be broken and put together again. Plaster impressions can be stored indefinitely. The material for die casting is also

gypsum. Before pouring in the die gypsum, the impression must be disinfected, watered, and treated with a release agent. The separation of impression and die gypsum is difficult, and the surface of the die can be damaged during this procedure.

Plaster of Paris is used for functional and anatomic impression taking of the edentulous jaw and fixation impression of restorations. It is rarely used for impressions since the introduction of alginate and elastomers. The procedure is unpleasant for the patient. Another use is fixation during determination of the positional jaw relation.

6.2.2. Zinc Oxide – Eugenol Pastes

The reactive components ZnO and eugenol (80 – 96 wt %) are supplied as two pastes, which are adjusted in consistency by oils (olive oil, mineral oil) and fillers (talc, kaolin, rosin, vaseline). Zinc acetate and magnesium chloride are used as accelerators for the formation of zinc eugenolate. Addition of a few drops of water during mixing enhances setting. Irritation of tissue by eugenol is possible. Impressions are taken from the wet mucosa with individual impression trays (functional impression with thin layers of paste; unwanted high-detail reproduction to avoid micromechanical influences on the mucosa by the prosthesis since the details especially those of the oral mucosa change rapidly). During removal from undercuts, plastic deformation of the impression occurs. Impressions can be disinfected and stored for a long time. The die material is gypsum. After setting of the die gypsum, the paste can be softened by immersion in water (70 °C) and pulled off. Zinc oxide pastes are rarely used nowadays (compression impression of the mucosa).

6.2.3. Resins

Self-hardening methyl methacrylate/poly(methyl methacrylate) systems are no longer used today for biological and methodological reasons. Polymeric materials for temporary rebasing of dentures can be used as impression materials to give a working model for a new denture. These materials consist of a polymer bead powder of higher acrylates [poly(ethyl methacrylate)], which is swelled for a few minutes in a mixture of phthalate ester and alcohol to give first a flowable and then a stable gel, without a setting reaction taking place. The mass reacts on short term loading with rebound (elasticity), and on long term loading with flow (plasticity).

6.2.4. Thermoplastic Impression Materials

Thermoplastic impression materials were formerly composed of natural materials (copal, shellac, colophonium, mastics, sandarac resin, carnauba wax, raw rubber). Today mainly synthetic resins, waxes, and fatty acids (stearic, palmitic, and oil acids) are used. Added to this are fillers and colors. The masses are reversibly plastified in a water bath or by a flame. They can be disinfected. Gypsum is used as die material. Thermoplastic materials belong to the oldest impression materials. Today their market share is low.

Thermoplastic impression compounds consist of resin, fatty acid, and filler in a ratio of 2/1/3. The usual filler is magnesium silicate (talc). The necessary plasticity for impression taking is reached at 50 – 45 °C. Hardening takes place above 37 °C. At room temperature the materials remain rigid, even under pressure. Today these materials are used in dentistry only as accessories during impression taking (individualization of impression trays).

Wax/resin materials contain 70 – 80 wt % synthetic resins, 15 – 25 wt % hard paraffin, paraffin oil, and fillers. Some products contain metal powders to enhance heat capacity. At mouth temperature the materials are flowable under the pressure from the soft tissues of the mouth (functional impression), and rigid at room temperature and pressureless storage.

Gutta-percha is the *trans* isomer of natural rubber. Impression materials contain 60 wt % gutta-percha, resins (25 wt %), and fillers (15 wt %). Gutta-percha is occasionally used for special impression tasks for removable prostheses and in maxillofacial prosthetics.

7. Dental Waxes

Waxes are used in rather large quantities in various dental laboratory and clinical procedures.

Different types of waxes are required for various applications. The differences relate chiefly to softening behavior and hardness, and the characteristics desired are usually obtained by blending natural and synthetic waxes. Waxes are not chemically uniform materials. They are defined only by their similar properties.

Natural waxes are obtained either from animals, mostly insects, or from minerals. In the dental literature, the following natural components are often cited: paraffin, ceresin, carnauba wax, candelilla wax, and beeswax.

Synthetic waxes, waxlike materials, and additives are often added to the natural waxes. Among these, the most important synthetic waxes are polyethylene wax, polyoxyethylene glycol wax, and esters of long-chain aliphatic alcohols and organic acids.

Various additives, e.g., resins, rosin, dammar, myricyl palmitate, or stearates, are added to special-purpose waxes. Coloring agents, usually fat-soluble dyes, are dissolved in dental waxes to differentiate the various types and to approximate the natural color of the gingiva and the teeth.

Most of the waxes and waxlike substances can be melted together to form mixtures that do not separate when cooled. In this way, any special dental wax required for specific dental laboratory and clinical applications can be made.

Dental waxes are generally used after being heated to a softened or liquid state. Below the melting point, a solid – solid change occurs in the crystal lattice; a wax should be molded above this temperature to avoid internal stress which can result in distortions of the pattern. The *advantages* of waxes include easy workability, good polishability, broad softening range, low viscosity when melted, hydrophobic behavior, and clean burnout. Among the *disadvantages* of waxes are high thermal expansion, formation of internal stresses, solid – solid transition, and low hardness at room temperature.

Some dental waxes are used as material for patterns in the lost-wax technique. This technique, also called the indirect technique, involves fabrication of a wax pattern, embedding it in an investment, burnout of the wax, and casting a molten alloy in the hollow space. Inlays, crowns, bridges, and the metal framework of partial dentures are constructed in this manner. Other types of waxes are used for bite registration, construction of trial baseplates for complete dentures, and temporary connection of fragments of a broken denture before a repair procedure.

Many types of dental wax are marketed. Their names indicate the uses for which they are intended:

modeling wax
casting wax, burnout wax, inlay wax
carding wax, setup wax
bite registration wax
blockout wax
baseplate wax
correction wax
sticky wax, adhesive wax.

Because of their particular importance, dental inlay casting wax and baseplate wax are treated in more detail in the following paragraphs.

Dental Inlay Casting Wax. Inlay wax is used to produce cast inlays. It is available in two slightly different types: type I for direct modeling in the mouth and type II for indirect modeling on the die. Both wax qualities must (1) burn out without residue, (2) have minimum thermal contraction, (3) be carved without chipping and flaking, and (4) be intensely colored. Type I must not solidify at 37 °C but should flow easily at 43 °C, whereas type II must still not soften at 34 °C and should flow easily at 40 °C. Inlay waxes contain a larger fraction of hard waxes, such as carnauba wax, than the related baseplate waxes.

Dental Baseplate Wax (Modeling Wax). Baseplate wax is preferred in the production of dentures to shape the denture base that is later made from resin. The waxes are mixtures of paraffin and beeswax with minor additions of carnauba wax. The modeling waxes are classified into three types which differ slightly in their softening range and hardness. The following requirements must be fulfilled: easy removal with boiling detergent solution, low thermal contraction, easy trimming and carving, and glossy surface after gentle flaming.

8. Other Model and Die Materials

The most important model material is dental stone (model gypsum); this is treated in detail

in Chapter 9. In this chapter, only model materials other than gypsum are treated briefly; these materials, however, are much less important than gypsum.

Die Cements. The composition of die cements is similar to that of the silicophosphate and stone cements described in Chapter 4. As a result of their poor physical properties, these materials are rarely used today.

Only small amounts of powder and liquid (acid) can be mixed to form a thick, homogeneous, plastic mix. For this reason, only small models or single tooth dies can be fabricated. The advantages of die cements are their high rigidity and good scratch resistance. Their disadvantages are poor flow rate, contraction during setting and storage, and the fact that they can be used only with compound impression materials.

Resins. Various resins which cure at room temperature are used as binders for chemically inert fillers. Methyl methacrylates, polyesters, epoxies, and ethylenimine resins are available commercially. Very fine quartz powder and metal powders are used as fillers; their chief function is to reduce shrinkage during setting. The materials are available as powder and liquid or as two-paste systems.

The materials are rarely used because of their disadvantages. Resins are very rigid and provide good reproduction of details. However, they require a long setting time, some products are sticky after mixing is complete, they are not compatible with all impression materials, and they contract during setting.

Metallic Materials. The *amalgams* used as die materials are identical with the silver amalgams described in Chapter 2. They cannot be used with elastic impressions because they must be placed by plugging. Hardness and absence of shrinkage are good; detail reproduction is satisfactory. Their use is restricted to small models or dies.

Electrolytic plating of the impressions with copper or silver and, more rarely, with nickel produces very good models and dies. However, plaster, alginate, and hydrocolloid impression materials cannot be electroplated. The impression, made conductive by dusting with metal or graphite powder (dental compound), is suspended in an aqueous electrolyte as the cathode. The main components of the electrolyte are copper sulfate or silver cyanide, respectively. Pure copper or silver is used as the anode. Appropriate selection of the bath and the electrical current leads to development of a thin, but smooth and homogeneous, coating of the impression. Detail reproduction, precision, and rigidity of the models are very good. To prevent deformation during manipulation, dental stone is cast into the plated impression. The only disadvantage is the time required for the plating procedure (up to 10 h).

9. Dental Gypsum Products

Gypsum materials have a wide spectrum of use in dentistry. In addition to their use as model materials, they also serve as impression plasters (Section 6.1.1), as a component of alginate impression materials (Section 6.2.2), and as a component of investment materials (Chap. 10). History, raw materials, phases of the $CaSO_4$–H_2O system, natural and byproduct gypsum, and production of calcium sulfate hemihydrate are described elsewhere (→ Calcium Sulfate) [62]. Dental gypsum products are classified by international standards:

Type 1: Impression plaster
Type 2: Dental plaster
Type 3: Dental stone
Type 4: Dental stone, high strength

Use as Model Material. A dental impression is a negative reproduction of the teeth and the oral tissues. By pouring a suitable model material into this negative form, the dental technician obtains a positive cast of the form and relationship of the teeth and oral tissues. Dental stone and high-strength dental stone are ideally suited for high-precision mastercasts. Dental plasters are suitable for study, repair, and reference casts.

Requirements. The requirements for such a material depend on the purpose for which it is used: *type 2 dental plaster* should be easy to manipulate and show a smooth surface; *type 3 dental stone* and *type 4 dental stone, high strength* should be easy to manipulate, be compatible with common impression materials, have a very good fluidity in the plastic phase thus needing a short time for vibration, and have a short setting time (max. 30 min). On the other hand, the

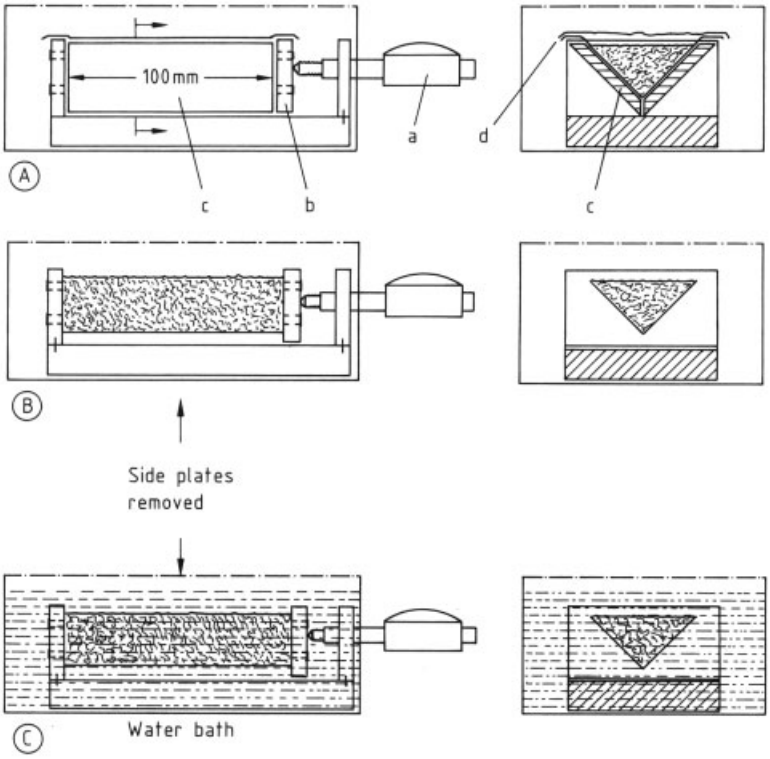

Figure 10. Measurement of dimensional change of dental stone
A: In a closed extensometer at 23 °C
B: With side plates removed (23 °C, 50 % relative humidity) 2 h after start of mixing
C: Immersion in water bath for 5 min after 24 h
a) Dial gauge; b) Movable front plate; c) Removable side plates; d) Polytetrafluoroethylene (PTFE) sheet

setting time must be long enough to facilitate pouring of the impression [62 – 64].

In addition to the requirements mentioned above, the final model should show the finest details; have a smooth surface without pores; have sufficient strength, hardness, and wear resistance; and have a constant volume (small expansion is better than even minimal shrinkage) [63, 64].

The requirements for hardness, compressive strength, and the maximum linear setting expansion are different for all four types of gypsum. None of the model materials of type 2 – 4 can fulfill perfectly all the requirements mentioned. Dental stone high strength fulfills them best and is therefore the most common model material for casts used for the construction of inlays, crowns, bridges and removable partial or full dentures [63, 64].

Chemical Composition. Calcium sulfate hemihydrate occurs in two phases: α and β (for more details, see → Calcium Sulfate). The composition of dental gypsum products are as follows:

- Type 1 impression plaster (see Section 6.2.1).
- Type 2 dental plaster consists of β-$CaSO_4$ hemihydrate and modifiers.
- Type 3 dental stone consists of α-$CaSO_4$ hemihydrate crystals with a small specific surface and modifiers.
- Type 4 dental stone high strength consists of an α-$CaSO_4$ hemihydrate that has been dehydrated in an autoclave with special solutions [65]. This type also contains modifiers, added by the manufacturer to obtain the properties of the final product required by national and international specifications.

Typical ionic additives are potassium sulfate, potassium citrate and sodium potassium tartrate.

Water-soluble melanine resins are used as nonionic additives to increase the fluidity of the water – hemihydrate mixture. All these modifiers influence the setting expansion, the setting time, the hardness of the dental stone, and the morphology of the dihydrate crystals. Potassium-containing additives can produce crystals of syngenite (potassium calcium sulfate), which can appear as efflorescences on the surface of dental casts [66 – 68].

The *setting time* [69 – 72] can be shortened by (1) addition of inorganic acids and their salts, such as calcium sulfate dihydrate, 3 % sodium sulfate; 2.5 % sodium chloride, or 2 – 3 % potassium sulfate; (2) intensive and extended spatulation (excessive spatulation destroys the crystal structure and decreases strength); and (3) low water-to-powder ratio.

The setting time can be prolonged by (1) addition of organic acids and their salts as well as the presence of organic colloids such as dried blood or gelatin; (2) addition of salts of phosphoric and boric acids; (3) addition of more than 3 % of sodium chloride; (4) presence of nitrate or citrate; and (5) high water-to-powder ratio [62].

The setting reaction does not completely convert the calcium sulfate hemihydrate to dihydrate. Residual hemihydrate is present in the structure of fully set casts [66].

Physical Properties. The *strength* of hardened calcium sulfate dihydrate is directly proportional to its density; therefore, the strength depends on the ratio of water to hemihydrate powder and on the porosity of the hardened plaster or stone. Dental α-$CaSO_4$ hemihydrates (stones) need small amounts of water for mixing and fluidizing and therefore have a high density and great strength; however, they are brittle. The strength of calcium sulfate dihydrate also depends on the humidity of the atmosphere during storage.

The *hardness* of rehydrated type 3 dental stone and type 4 dental stone high strength increases with the time of storage. Chemically produced gypsum reaches its final hardness after 4 d, whereas natural gypsum needs 6 d. Due to sedimentation, the hardness of a cast is greater on the bottom than on the surface, where more pores are located [62, 73 – 75].

The hardness can also be influenced by contact with some impression materials. Contact with hydrocolloids, condensation-cured silicones, and polyethers lowers the hardness. This tendency can be reduced if the impression with the mixed hemihydrate is stored at 95 % R.H. for 30 min or during the entire setting period [62].

Volume Change. The absolute volume of a mixture of $CaSO_4$ hemihydrate and water first decreases when $CaSO_4$ dihydrate is formed. The contraction can only be measured for a short time after mixing [71, 76 – 79]. As soon as the rehydrated gypsum hardens, expansion starts. The thermal expansion during setting is reversible, but the setting expansion is almost irreversible. This setting expansion is influenced by (1) the water-to-powder ratio: the lower the water content, the greater is the expansion; (2) the presence of chemicals: for example, potassium sulfate lowers the expansion but also the strength; (3) the climate: storage of a model of type 4 dental stone high strength after the removal of the impression at 23 °C and 50 % relative humidity results in a contraction of about 0.05 % of its highest setting expansion within 7 d (Figs. 10 and 11); and (4) spatulation: high intensity and long time of spatulation of the water – hemihydrate mixture increase the expansion of type 3 and type 4 dental stone.

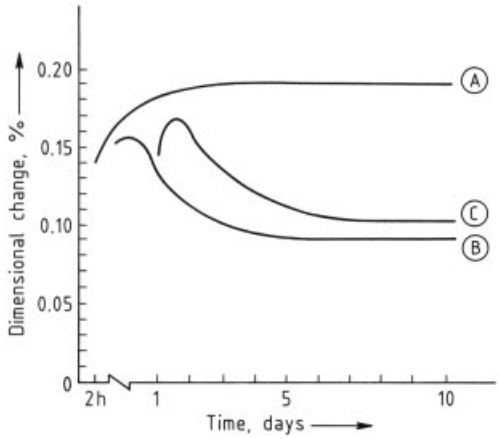

Figure 11. Dimensional change of dental stone model (at 23 °C and 50 % rel. humidity) as a function of storage conditions [62, 76]
A: Storage in a closed extensometer
B: Storage at 23 °C and 50 % relative humidity
C: Immersion in water bath for 5 min and subsequent storage as in B
For experimental conditions of A, B, C, see Figure 10.

The results of the setting expansion depend on the method and time of measurement (Figs. 10 and 11). The dental gypsum products are measured in an extensometer with a length of 100 mm [64]. The material of the surface of the extensometer in contact with the water – dihydrate mixture can influence the setting expansion [97].

Quality Standards. The Technical Committee 106 of the ISO has published ISO Standard 6873 (1998) Dental Gypsum Products. This standard also specifies the corresponding test methods.

Table 8 shows the ISO requirements for four different types of dental gypsum products. Table 9 compares the results obtained for various commercial products.

Storage and Packaging. Types 3 and 4 stones should be packed and stored in protective, moistureproof, and sealed containers. Each container must at least be marked clearly with the following information: name and/or trademark of the manufacturer, type, lot number, date of packaging, and date beyond which the material should not be used without retesting when stored under the recommended conditions: ca. 23 °C and not more than 50 % relative humidity.

Trade names: Type 2: Snow White Plaster No. 1 (Kerr); Alabaster (Dr. Böhme); Laboratory Plaster (Whip Mix).

Type 3: Castone, R & R Lab Stone (Dentsply); Coecal Buff, White (Coe Lab.); Octa-Mol (Heraeus-Kulzer); Modelit (Dr. Böhme); Microstone, Quickstone (Whip Mix).

Type 4: Duralit (DeguDent); Duroc (Dentsply); Begodur (Bego); Octa-Stone (Heraeus-Kulzer); Gilstone (Giulini); Silky-Rock (Whip Mix); Super-Cal Dental Stone (Coe Lab.); Tewestone (Kettenbach); Vel-Mix (Kerr).

10. Dental Investments

An investment is generally a powder mixture which can be mixed with a liquid to give a flowable paste that sets in a given time. The refractory material allows heating above 1100 °C. These properties make the material suitable as a mold material for alloy casting.

In dental technology a model is made from an impression that reproduces the teeth. The wax pattern (or resin, adapter cap) is shaped on this model. The pattern is removed from the model and provided with sprues on a base former and a casting ring. The mixed investment material is then poured into the casting ring. After hardening of the investment, the mold is heated to burn out the wax model. The casting mold is brought to the final temperature required by the particular alloy. After casting the object is divested and is ready to be worked on.

Classification. Investments can be classified according to the nature of the bonding material (gypsum- or phosphate-bonded) or the purpose for which they are produced (Co – Cr castings, precious metal castings, soldering of metallic parts, ceramic pressure technique).

Requirements. Investment casting materials should fulfill the following requirements:

- Expansion control: diluting the liquid component of phosphate-bonded investment materials with distilled water allows the expansion of an investment to be controlled in a way that the contraction of the particular casting alloy is compensated
- A suitable setting time
- Creamy consistency during investment
- Production of a smooth surface and reproduction of fine details
- No cracking during heating and on impingement of the molten alloy
- Compatibility with the alloys
- Keeping the high thermal expansion during the entire time of casting and reversible thermal expansion on cooling
- Easy removal from the cast object
- A possible additional requirement is suitability for the rapid-burnout method.

Chemical Composition. Casting investment materials consist of bonding and refractory materials. The refractory materials such as silica, usually quartz or cristobalite or a mixture thereof are responsible for the thermal expansion. α-Cristobalite changes reversibly into β-cristobalite, and α-quartz into β-quartz. These two crystal structure transformations result in the reversible thermal expansion of an investment. Standard investments need time for these

Table 8. Requirements and limits for types 1 – 4 dental gypsum products according to ISO Standard 6873. Working time and setting time are measured from the first contact of hemihydrate and water; the setting time must be within ± 20 % of the time given by the manufacturer.

Type	Testing consistency, mm	Pouring time, min	Setting time, min		Setting expansion, %	Compressive strength, N/mm^2	
		min.	min.	max.	max.	min.	max.
1	80 ± 4	1.25	2.5	5.0	0.15	4.0	8.0
2	75 ± 4	2.5	6.0	30.0	0.30	9.0	
3	30 ± 3	3.0	6.0	30.0	0.20	20.0	
4	30 ± 3	3.0	6.0	30.0	0.15	35.0	

Table 9. Physical properties of some commercial type 4 dental stones, high strength

Product	Water-to-powder ratio, cm^3/g	Linear setting expansion, % (ISO 6873)	Compressive strength, N/mm^2 (ISO 6873)	Hardness, N/mm^2 (DIN 13 911) (1976)
Begodur	0.23	0.13	40	160
Fujirock	0.20	0.10	40	240
Geostone	0.23	0.10	35	190
Gilstone	0.23	0.13	43	185
Silky Rock	0.22	0.11	40	190
Tewestone	0.23	0.10	43	200
Vel-Mix Stone	0.23	0.14	40	185

transformations to take place without formation of cracks; hence, holding times at 270 and 573 °C are needed during the heating process.

Gypsum-Bonded Investments. The bonding material is α-calcium sulfate hemihydrate $CaSO_4 \cdot 0.5\,H_2O$, which is converted to calcium sulfate dihydrate $CaSO_4 \cdot 2\,H_2O$ by mixing with water. The excess water is vaporized during initial heating to about 100 °C, and above this temperature water of crystallization is driven out of the investment. After reaching a temperature of 150 °C, gypsum-bonded investments either show a slight thermal contraction or remain unchanged up to 225 °C. At higher temperatures, the presence of silica dominates the thermal expansion. Gypsum-bonded investments should not be heated to temperatures exceeding 700 °C. Beyond that point, gypsum decomposes in the presence of silica to liberate sulfur dioxide, which causes porosity and corrosion of the casting.

$$2\,CaSO_4 \rightarrow 2\,CaO + 2\,SO_2 + O_2$$

Expansion control of gypsum bonded investments is achieved by varying the mixing ratio with water.

Phosphate-Bonded Investments. High-temperature investments require a bonding material other than gypsum. This bonding material consists of magnesium ammonium phosphate or a soluble acid phosphate such as ammonium dihydrogenphosphate and a metal oxide such as magnesia. These materials are mixed with an aqueous silica sol. The following simplified chemical reaction occurs during setting at room temperature and gives the investment the green strength before heating:

$$(NH_4)\,H_2PO_4 + MgO + 5\,H_2O \rightarrow MgNH_4PO_4 \cdot 6\,H_2O$$

The next reaction takes place during heating, water and ammonia are evolved above 200 and 300 °C, respectively.

$$2\,(MgNH_4PO_4 \cdot 6\,H_2O) \rightarrow Mg_2P_2O_7 + 13\,H_2O + 2\,NH_3$$

The silica-catalyzed decomposition into magnesium diphosphate increases the strength and hardness of the phosphate-bonded investment. By heating above 1000 °C with MgO surplus the following reaction takes place:

$$Mg_2P_2O_7 + MgO \rightarrow Mg_3(PO_4)_2$$

Physical Properties. The most important physical property of a dental investment material is expansion. The expansion is the sum of setting and thermal expansions and is responsible for the outstanding accuracy of dental restorations.

Setting Expansion. The setting of the phosphate-bonded investments is accompanied by the evolution of heat. At the end of the reaction process, the temperature reaches a maximum of 55 – 80 °C. This temperature increase affects the setting expansion. A standardized method for measurement of the setting expansion is presently under discussion. The setting expansion constitutes about 30 – 50 % of the thermal expansion and is caused by the reaction of the bonder. The bonder transformation separates the quartz and cristobalite grains and the result is an increase in volume. This process is intensified by mixing the powder with a special liquid and causes aqueous silica particles to crystallize to long needles.

The setting expansion of gypsum-bonded investments is based on the same principles as for gypsum products in Chapter 9.

Thermal Expansion. When dental alloys are cast, they contract on cooling. For this reason, thermal expansion is very important for dental investments and should compensate for the shrinkage of the casting alloys; however, the two materials have different thermal contraction curves. The thermal expansion of the investment materials depends on their composition and, in the case of phosphate-bonded investments, on the dilution of the liquid component and the particle size of the silica. The thermal expansion curves for quartz and cristobalite are very much different. Moreover, both silica modifications show a phase transition, each at a different temperature: quartz at 573 °C and cristobalite at 200 – 270 °C. All modifications of silica show a break in their expansion curves. Therefore, the thermal expansion curves for investment materials are very complicated.

Setting time of phosphate bonded investment materials is dependent on the reaction time of the two bonding materials and consequently on their mixing ratio and their grain size. Gypsum-bonded investment materials include small amounts of chemical additives which influence the setting time.

Hardness. All dental investments need enough hardness to resist the high temperatures and the casting impact.

Quality Standards. The classification and requirements for investment materials together with the corresponding test methods are described in the standards ISO 9694 (Dental Phosphate-Bonded Casting Investments, 1996) and ISO 7490 (Dental Gypsum-Bonded Casting Investments 2000).

Application. In general, the instructions of the manufacturer should be strictly followed for optimum results. Gypsum-bonded investment materials should not come into contact with phosphate-bonded investment materials. For gypsum- and phosphate-bonded investment materials, the use of separate mixing bowls and spatulas is necessary. Standard investment materials need preheating with holding times to withstand the quartz and cristobalite phase transitions without cracking. Rapid-burnout investment materials can be preheated in an oven at the final temperature. Exact ratios of the bonding and the silica materials and the use of special liquid components are the reasons for these robust properties. The advantages are that they do not require preheating in a controlled oven, have shorter heating times, and save energy.

11. References

1. J. F. Bates, A. G. Knapton: "Metals and Alloys in Dentistry," *Int. Met. Rev.* **22** (1977) no. 3, 39 – 60.
2. G. Tammann, *Z. Anorg. Allg. Chem.* **107** (1919) 1 – 240.
3. H. Knosp, M. Nawaz, M. Stümke, *Gold Bull.* **14** (1981) no. 2, 57 – 64.
4. R. M. German, *Int. Met. Rev.* **27** (1982) no. 5, 260 – 288.
5. R. L. Bertolotti, *J. Am. Dent. Assoc.* **108** (1984) no. 6, 959 – 966.
6. J. R. Kelly, T. C. Rose, *J. Prosthet. Dent.* **49** (1983) no. 3, 363 – 370.
7. *Cobalt, Monograph*, edited by centre d'information du cobalt, Bruxelles 1960.
8. H. Newesely, *Naturwissenschaften* **70** (1983) 324 –331.
9. W. Wagner, P. Tetsch: *Deutscher Zahnärzte-Kalender 1986*, Hanser Verlag, München 1986, pp. 45 – 58.

10. P. J. Branemark, G. A. Zarb, T. Albrektsson: *Tissue Integrated Prostheses Osseointegration in Clinical Dentistry*, Quintessenz-Verlag, Berlin 1985.
11. J. Wirz: *Deutscher Zahnärzte-Kalender*, Hanser Verlag, München, 1983, pp. 52 – 66.
12. R. V. Williams, C. E. Ingersoll, *Phillip Journal für restaurative Zahnmedizin* **1** (1984) 33 – 39.
13. W. I. Ferrier: *Gold Foil Operations*, University of Washington Press, Seattle 1959.
14. K. Dreyer Jørgensen: *Amalgame in der Zahnheilkunde*, Hanser Verlag, München-Wien 1977.
15. D. B. K. Innes, W. V. Youdelis, *J. Can. Dent. Assoc.* **29** (1963) 587 – 593.
16. Special Metals Corp., US 3 980 472, 1975 (K. Asgar, S. H. Reichman).
17. Sybron Corporation, US 3 985 558, 1975 (H. H. Simpson).
18. R. Kropp, H. Haußelt, *Quintessenz* **34** (1983) 1027 – 1031.
19. A. Knappwost, E. Gura, D. Fuhrmann, A. Enginalev, *ZWR* **94** (1985) 131 – 139.
20. E. Lugscheider, G. Jangg: *Z. Metallkd.* **62** (1971) 548 – 551.
21. P. A. Steinberg, K. H. Schmitz: *Grundriß der Dental-Keramik*, Verlag Neuer Merkur, München 1968.
22. Ch. Hahn, D. Rothemund: "Dentalkeramik," in: *Handbuch der Keramik,* Verlag Schmidt GmbH, Freiburg i. Br. 1977 and 1981, Gruppe II L, pp. 1 – 12.
23. J. W. McLean (ed.): *Dental-Keramik*, Quintessenz Verlags-GmbH, Berlin 1984.
24. Permadent Products Corporation, US 3 052 982 (= DE 1 441 336), 1962 (M. Weinstein, S. Katz, A. B. Weinstein).
25. J. Jach, A. Joshi, D. Sengupta, *Semicond. Insul.* **5** (1980) 111 – 122.
26. P. Dorsch, Dental Labor XXXI (1983) 1237 – 1242.
27. K. Krumbholz, *Quintessenz Zahntech.* **9** (1983) 1175 – 1179.
28. Johnson & Johnson Dental Products Company, EP-A 0 124 298, 1984 (C. Panzera).
29. J. W. McLean, T. H. Hughes, *Br. Dent. J.* **119** (1965) .
30. J. Schmidseder, "Hi-Ceram," *Phillip Journal für restaurative Zahnmedizin* **3** (1986) 226.
31. Corning Glass Works, US 3 839 055, 1974 (D. G. Grossmann).
32. Dentsply International Inc., EP-A 0 154 137, 1985 (J. J. Blair, R. C. Shue).
33. A. Wohlwend, P. Schärer, "Die Empress-Technik. Ein neues Verfahren zur Herstellung von vollkeramischen Kronen, Inlays und Facetten," *Quintessenz Zahntech.* **16** (1990) 996.
34. H. Claus, "Vita In-Ceram, ein neues Verfahren zur Herstellung oxid-keramischer Gerüste für Kronen und Brücken," *Quintessenz Zahntech.* **16** (1990) 35.
35. J. Becker, D. Heidemann, "Entwicklungsstand und Probleme des Einsatzes von CAD/CAM-Systemen," *Dtsch. Zahnaerztl. Z.* **48** (1993) 611.
36. A. Groß, *Chem. Uns. Z.* **13** (1979) 142.
37. Kulzer & Co GmbH, DE 737 058, 1983.
38. W. Westermann, H. Meiners, *Zahnärztl. Welt* **84** (1975) 531.
39. Degussa, DE 975 072, 1941 (A. E. Czapp, E. Schnebel, A. Gölz).
40. H. Bredereck, et al., *Chem. Ber.* **87** (1954) 129; **88** (1955) 438; **89** (1956) 731.
41. H. Marx: "Prothesenkunststoffe und ihre Verarbeitungsverfahren," in *Fortschritte in der zahnärztlichen Prothetik und Werkstoffkunde*, vol. **1**, Hanser Verlag, München 1980, pp. 344 – 356.
42. A. Schmidt, "Die Geschichte der Methacrylate in der Stomatologie," *Zahntechnik (DDR)* **19** (1978) 436.
43. H. Marx: "Kunststoffzähne," in *Fortschritte in der zahnärztlichen Prothetik und Werkstoffkunde*, vol. **1**, Hanser Verlag, München 1980, pp. 358 – 360.
44. H. Marx: *Fortschritte in der zahnärztlichen Prothetik und Werkstoffkunde,* vol. 1, 1980, H. Marx: "Verblendkunststoffe und ihre Verarbeitungsverfahren," in *Fortschritte in der zahnärztlichen Prothetik und Werkstoffkunde*, vol. **1, vol. 2**, Hanser Verlag, München 1980, 1984; pp. 362 – 367 (vol. 1); pp. 252 –259 (vol. 2).
45. M. Hofmann, J. Hofmann, "Eine neue Kunststoffverblendtechnik," *Quintessenz Zahntechnik* **7** (1981) 1101.
46. H. R. Stanley, R. L. Bowen, J. Folio, "Compatibility of various materials with oral tissues" II: "Pulp responses to composite ingredients," *J. Dent. Res.* **58** (1979) 1507.
47. R. Musil, H.-J. Tiller, "Die molekulare Kopplung der Kunststoff-Verblendung an die Legierungsoberfläche," *dentallabor* **32** (1984) 1155.

48. Kulzer, US 4 600 390, 1986 (R. Göbel, R. Musil, H.-J. Tiller, St. Oppowsky, A. Schmidt, R. Janda).
Bard Ltd., DE 3 403 874, 1986 (D. E. Cross).
49. M. Atsuta, A. K. Abell, D. T. Turner, N. Nakabayashi, M. Takeyama, "A new coupling agent for composite materials 4-methacryloxyethyltrimellitic anhydride," *J. Biomed. Mater. Res.* **16** (1982) 619.
50. H. W. R. Schaefer, "Fortschritte auf dem Gebiet der Kunststoff-Füllungen-Füllungsmaterialien mit feinstteiligem Siliciumdioxid als anorganischem Füllstoff," *Quintessenz* **32** (1981) 1.
51. H. Newesely, "Zemente, Composites, Adhäsive und Kleber," in *Fortschritte in der zahnärztlichen Prothetik und Werkstoffkunde,* vol. 1, Hanser-Verlag, München vol. 1, 1981, pp. 392 – 411. H. Newesely, R. Janda: vol. 2, 1984, pp. 284 – 303.
52. R. L. Bowen, US 3 066 112, 1962;
US 3 179 623, 1965;
3 194 783, 3 194 784, 1965.
53. J. Anderson et al.: *Applied Dental Materials*, Blackwell Science Ltd., Oxford 1998.
54. R. G. Craig: *Restorative Dental Materials*, 10th ed., C. V. Mosby Co., St. Louis 1996.
55. E. H. Greener, J. K. Harcourt, E. P. Lautenschlager: *Materials Science in Dentistry*, Williams & Wilkins Co., Baltimore 1972.
56. F. J. Harty, D. H. Roberts: *Restorative procedures for the Practicing Dentist*, J. Wright and Sons Ltd., Bristol 1974.
57. K. Körber, K. Ludwig: *Zahnärztliche Werkstoffkunde und Technologie*, 2nd ed., Thieme-Verlag, Stuttgart 1993.
58. R. W. Phillips: *Skinner's Science of Dental Materials*, 9th ed., Saunders, Philadelphia 1991.
59. J. Viohl: "Abformwerkstoffe," in K. Eichner, H. F. Kappert (eds.): *Zahnärztliche Werkstoffe und ihre Verarbeitung*, 6th ed., vol. **1**, Hüthig Verlag, Heidelberg 1996.
60. B. Wöstmann: "Abformmaterialien," in H. Meiners, K. M. Lehmann (eds.): *Klinische Materialkunde für Zahnärzte*, Hanser Verlag, München – Wien 1998.
61. Deutsches Institut für Normung e.V. (ed.): *DIN-Taschenbuch 267 Zahnheilkunde; Werkstoffe. Normen, Gesetze, Richtlinien*, 1st ed., Beuth Verlag, Berlin – Wien – Zürich 1997.
62. G. Franz: *Dentalgipse*, Hanser Verlag, MünchenWien 1981.
63. Deutsches Institut für Normung (DIN): DIN 13 911 (Dental Gypsum).
64. International Organization for Standardization (ISO): ISO 6 873 (Dental Gypsum Products).
65. G. Giulini, DE 1 157 128, 1968.
66. U. Sondermann, K. M. Lehmann (eds.): *Dentalgipse und ihre Anwendung*, Philipps Universität, Marburg 1998.
67. J. Wolf, U. Sondermann, K. M. Lehmann: "Der Nachweis verschiedener Kristallphasen an Gipsoberflächen mittels Röntgendiffraktometrie," *Interdiszipl. J. Proth. Zahnheilk.* **1** (1998) 175 – 178.
68. K. M. Lehmann, J. Wolf: "Der Einfluß von Stellmitteln auf das Abbinden eines α-Calciumsulfat-Subhydrates," *Interdiszipl. J. Proth. Zahnheilk.* **1** (1998) 425 – 430.
69. K. Dreyer Jørgensen, A. S. Posner, *J. Dent. Res.* **38** (1959) 491.
70. G. Franz: "Gips," in K. Eichner (ed.): *Zahnärztliche Werkstoffe und ihre Verarbeitung*, 4th ed., Hüthig Verlag, Heidelberg 1981.
71. J. K. Harcourt, E. P. Lautenschlager, *J. Dent. Res.* **49** (1970) no. 3, 502.
72. D. C. Smith, *J. Dent. Res* **41** (1962) 1256.
73. W. Finger, *Zahnärztl. Welt* **81** (1972) 747.
74. F. A. Peyton, J. P. Leibold, G. Ridgley, *J. Prosthet. Dent.* **2** (1952) 381.
75. H.-J. Rehberg, G. Becherer, *Dtsch. zahnärztl. Z.* **9** (1954) 488.
76. G. Franz: *Untersuchungen über das Dimensionsverhalten zahnärztlicher Hartgipse*, Hanser Verlag, München-Wien 1978.
77. D. B. Mahler, *Int. Dent. J.* **5** (1955) 241.
78. H. Marx, *Dtsch. Zahnärztl. Z.* **31** (1976) 694.
79. H. H. van't Hoff, E. E. Armstrong, S. Hinrichsen, F. Weigert, G. Just, *Z. Phys. Chem. Stoechiom. Verwandschaftsl.* **45** (1903) 257.
80. S. Halbach, *Human & Experimental Toxicology* **13** (1994) 496 – 501.
81. M. Walter, P. D. Reppel, K. Böning, W. B. Freesmeyer, *J. Oral Rehabil.* **26** (1999) 91 – 96.
82. A. H. Sherif, *Egypt Dent. J* **41** (1995) 1067 – 1076.
83. B. S. Vence, *J Prothet. Dent.* **77** (1997) 444 – 449.
84. L. Pröbster, J. Geis-Gerstorfer, A. Simonis, J. Setz, H. Weber, *Dent. Labor (München)* **39** (1991) 1073 – 1078.

85. P. Rammelsberg, R. van Kaldenkerken, P. Pospiech, *Dtsch Zahnärztl Z.* **53** (1998) 844 – 848.
86. E. Wagner: *Werkstoffkunde der Dental-Edelmetallegierungen*, Verlag Neuer Merkur, Munich 1980.
87. E. M. Wise: *Palladium Recovery and Uses*, Academic Press, New York 1968.
88. Degussa AG, DE 11 83 247, 1962 (Wagner).
89. Degussa AG, DE 15 33 233, 1966 (Wagner).
90. Degussa AG, DE 27 55 913, 1977 (Knosp).
91. W. Groll, Thesis, Universität Erlangen, 1987.
92. E. Schmid, Thesis, Universität Stuttgart, 1987.
93. D. C. Smith, *Biomaterials* **19** (1988) 467 – 478.
94. K. J. Anusavice: *Phillips' Science of Dental Materials*, 10th ed., W. B. Saunders Comp., Philadelphia PA, 1996, p. 525 ff.
95. DE 174 558, 1903 (P. Steenbock).
96. K. Falck: *Einführung in die Werkstoffkunde für Zahnärzte*, Carl Hanser Verlag, München 1948.
97. NRDC, DE 2 061 513, (Priority GB: 1969) (A. D. Wilson, B. E. Kent).
98. A. D. Wilson, J. W. McLean: *Glasionomerzement*, Quintessenz Verlags-GmbH, Berlin 1988.
99. R. G. Hill, A. D. Wilson, *Glass Technol.* **29** (1988) 150 – 158.
100. ESPE, EP 0 023 013, (Priority: 1979, DE) (W. Schmitt, R. Purrmann, P. Jochum, O. Gasser).
101. S. Crisp, A. D. Wilson, *J. Dent. Res.* **55** (1976) 1023 – 1031.
102. ESPE, EP 0 024 056, (Priority: 1979, DE) (W. Schmitt, R. Purrmann, P. Jochum, O. Gasser).
103. NRDC, EP 0 340 016, (Priority: 1988, GB) (A. D. Wilson, J. Ellis).
104. Albright & Wilson, GB 2 291 060, (1994) (C. E. Wilde, M. J. Williams).
105. A. D. Wilson, J. M. Paddon, S. Crisp, *J. Dent. Res.* **58** (1979) 1065 – 1071.
106. R. G. Hill, *J. Mater. Sci.* **28** (1993) 3851 – 3858.
107. A. J. de Gee, R. N. B. van Duinen, A. Werner, C. L. Davidson, *J. Dent. Res.* **75** (1996) 1613 – 1619.
108. ESPE, DE 3 248 357, 1982 (J. W. McLean, O. Gasser).
109. R. Hickel, *Dtsch. Zahnärztl. Z.* **52** (1997) 572 – 585.
110. R. Frankenberger, J. Sindel, N. Krämer, *Quintess. Int.* **28** (1997) 667 – 676.
111. R. Guggenberger, R. May, K.-P. Stefan, *Biomaterials* **19** (1998) 479 – 483.
112. N. Krämer, M. Pelka, P. Kautetzky, J. Sindel, A. Petschelt, *Deutsch. Zahnärztl. Z.* **52** (1997) 186 – 189.
113. K.-P. Stefan, *J. Dent. Res.* **77** (1998) Special Issue B, 688 (Abstract Nr. 454).
114. L. Forsten, *Scand. J. Dent. Res.* **98** (1990) 179 – 185.
115. A. Stassinakis, J. Gujer, B. Hugo, P. Hotz, *Acta Med. Dent. Helv.* **1** (1996) 244 – 249.
116. E. Glockmann, C. Gehroldt, K. Triemer, *Dtsch. Zahnärztl. Z.* **52** (1997) 668 – 672.
117. L. Forsten, *Scand. J. Dent. Res.* **99** (1991) 241 – 245.
118. H. Ngo, V. Marino, G. J. Mount, *J. Dent. Res.* **77** (1998) Special Issue B, 641 (Abstract Nr. 75).
119. G. Schmalz, K.-A. Hiller, F. Dörter-Aslan, *J. Mat. Sci.: Mat. In Med.* **5** (1994) 43 – 51.
120. L. M. Jonck, C. J. Grobbelaar, H. Strating, *Clin. Mater.* **4** (1989) 85 – 107.
121. L. M. Jonck, C. J. Grobbelaar, H. Strating, *Clin. Mater.* **4** (1989) 201 – 224.
122. G. Geyer, J. Helms, *Adv. Biomater.* **10** (1990) 529 – 535.
123. I. M. Brook, P. V. Hatton, *Biomaterials* **19** (1998) 565 – 571.
124. ESPE, EP 0 469 573, (Priority: 1990, DE) (O. Gasser, R. Guggenberger).
125. J. W. McLean, A. D. Wilson, *Br. Dent. J.* **136** (1974) 269 – 276.
126. H. W. Seeholzer, *Schweiz. Monatsschr. Zahnmed.* **97** (1987) 344 – 347.
127. C. Hüskens, C. Matter-Grütter, F. Lutz, *Schweiz. Monatsschr. Zahnmed.* **105** (1995) 216 – 221.
128. F. Khoury, H. J. Staehle, *Dtsch. Z. Mund Kiefer Gesichtschir.* **11** (1987) 351 – 355.
129. R. Mathis, I. L. Ferracane, *J. Dent. Res.* **66** (1987) 113 (Abstr. 51).
130. 3M, US 5 130 347, 1991 (S. B. Mitra).
131. J. W. McLean, J. W. Nicholson, A. D. Wilson, *Quiness. Int.* **25** (1994) 587 – 589.
132. M. B. Miller, *Phillip J.* **14** (1997) 385 – 386.
133. *The Dental Advisor* **14** (1997) Juni, 2.
134. J. Sindel, N. Krämer, R. Frankenberger, S. Bischoff, *J. Dent. Res.* **77** (1998) Special Issue B, 928 (Abstract 2375).
135. Ernst Mühlbauer KG, EP 0 219 058, (Priority: 1985, DE) (J. Engelbrecht).
136. Dentsply, EP 0 621 028, (Priority: 1993, US) (C.-T. Hunag, P. Hammersfahr).
137. H. J. Staehle: *Calciumhydroxide in der Zahnheilkunde*, Carl Hanser Verlag, München 1990.
138. Hoechst, DE 1 006 125, 1954 (O. Fuchs, G. Wegner).
139. ESPE, EP 0 094 049, (Priority: 1982, DE) (P. Jochum, O. Gasser).

Dermatotherapeutic Agents

Gerard J. Gendimenico, Johnson & Johnson Pharmaceutical Research and Development, L.L.C., Raritan, New Jersey, United States

1. **Functions and Structure of the Skin** . 1
2. **Topical Therapy** 2
2.1. **Topical Dosage Forms** 2
2.2. **Percutaneous Absorption** 4
2.3. **Vehicles** 4
3. **Classification of Drugs** 5
3.1. **Anti-Inflammatory Agents** 5
3.2. **Antiacne Agents** 9
3.3. **Antipsoriatics** 11
3.4. **Antifungals** 14
3.5. **Antiaging** 16
4. **References** 16

1. Functions and Structure of the Skin

Functions. The skin is a three-compartment organ that serves several essential roles for the human body. As the major protective barrier between the body and the environment, the skin minimizes the potentially harmful effects of ultraviolet radiation, temperature extremes, microbial infection, and mechanical, chemical, or electrical trauma.

The skin is also an important sensory organ, containing nerve endings receptive to pressure, pain, and temperature. It protects against changes in body fluid status and provides a major means of thermoregulation through changes in the vascular system and evaporation of sweat.

Elimination of biochemical wastes and xenobiotics is accomplished in the skin by metabolic and excretory pathways, including sweating. Various synthetic activities may also occur, including sunlight-mediated production of vitamin D from precursors present in the epidermis.

Structure. The outer protective layer of the skin, the epidermis, is a thin, stratified squamous epithelium composed primarily of keratinocytes, melanocytes, and Langerhans cells. The epidermis (ca. 0.07 – 0.1 mm thick) lacks blood vessels. Keratinocytes, the predominant epidermal cell type, synthesize keratin, an insoluble fibrous protein, which is responsible for the protective qualities of the epidermis and is the structural component of hair and nails.

Keratinocytes originate in the innermost or basal layer of the epidermis and undergo a multistage differentiation process. This maturation process, termed keratinization, occurs as the cells migrate toward the surface and culminates in formation of a flattened, dehydrated, compact outer layer (0.01 mm thick) of nonviable horny cells. This layer, the stratum corneum, plays the critical role in the barrier functions of the skin and has significant water-retaining properties. It is composed primarily of keratin and lipids. Continuous shedding of surface cornified cells, i.e., desquamation, results in regeneration of stratum corneum approximately every 14 days. Complete epidermal turnover takes ca. 28 days.

Melanocytes originate from the neural crest and produce melanin, a pigment that largely determines normal skin color. Together with keratin, melanin protects the underlying basal layer against ultraviolet radiation damage. Located in the basal layer of the epidermis, melanocytes synthesize pigmented granules called melanosomes that are released from the dendrites of the melanocytes and transferred to adjacent keratinocytes through a phagocytic process. The number and size of melanosomes determine normal skin complexion and the degree of protection against UV-B radiation (280 – 320 nm) penetration. Sun-induced tanning is a delayed response that results from increased production of melanin.

Langerhans cells are dendritic cells involved in immune functions of the skin. These cells are responsible for uptake and transfer of antigens from the epidermal surface to T-cell lymphocytes, thereby initiating delayed hypersensitivity reactions.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_301.pub2

The *dermis,* which forms the 2 – 3-mm-thick middle layer of the skin, consists of dense connective tissue with collagenous and elastic fibers dispersed in an amorphous ground substance composed primarily of glycosaminoglycans.

Collagen is a long, thin fibrous protein that accounts for the bulk of the dermis and lends strength and flexibility to the skin. Elastic fibers, consisting of protein fibrils and an amorphous protein, elastin, are primarily responsible for maintaining the skin's elasticity.

The dermis is rich in smaller blood vessels, lymphatic vessels, and nerves. The major cells of the dermis include fibroblasts, which are necessary for tissue repair, phagocytic histiocytes, and mast cells, which contain histamine and other chemical mediators of inflammation.

Underlying the dermis is the *subcutis,* a thick inner layer of protective connective tissue. Subcutaneous tissue contains large blood vessels and nerves, as well as numerous adipocytes, which manufacture and store fat. The subcutis also serves as a cushion to external impact and is an effective heat insulator.

Sebaceous glands, apocrine and eccrine sweat glands, and hair follicles may originate in the dermis or subcutaneous tissue and extend upward to the skin surface.

The structural and physiological properties of the skin account for its intrinsic impermeability and enable it to effectively perform its many protective functions. The same qualities, however, contribute to the difficulties associated with topical drug therapy. The factors that influence drug penetration through the skin, i.e., percutaneous absorption, are briefly discussed in Sections 2.2 and 2.3.

2. Topical Therapy

Topical application of drugs to the skin is often desirable in treating dermatological and, less commonly, systemic disorders.

Systemic administration, although advantageous in some skin diseases, is frequently limited by (1) failure to achieve therapeutic concentrations of many drugs, particularly in avascular regions of the skin and (2) unwanted pharmacologic effects outside the skin. Topical therapy allows for selective treatment of skin disease while the possibility of systemic adverse effects is minimized. Topical therapy generally does not result in appreciable levels of drug in the systemic circulation; however, some therapeutic agents, e.g., nitroglycerin [*55-63-0*] and scopolamine [*51-34-3*], have been successfully delivered transdermally.

2.1. Topical Dosage Forms

A wide variety of dosage forms exists for the topical delivery of medicinal compounds to the skin. Although systemic absorption can be either a deliberate or unintentional consequence of topical application, the action of topically applied drugs is most often limited to the site of application. Local therapeutic action is enhanced by formulations that promote prolonged contact with the target area while reducing the likelihood of absorption. Semisolid preparations, such as ointments, pastes, creams, and gels, fulfill these criteria to varying degrees. Unlike hydroalcoholic solutions, powders, and suspensions, which may be used to deliver drugs topically as well as by other routes of administration, semisolids are used exclusively for topical drug delivery.

Semisolid ointment bases are used primarily as vehicles to prepare pharmaceutical ointments, creams, and pastes. Ointment bases vary in viscosity and solubility and are classified as oleaginous, emulsifiable, emulsified, and water-soluble.

Ointments are formed by dissolving or suspending a drug in a semisolid base. Although the term "ointment" is sometimes used specifically for drug-containing oleaginous bases, emulsion bases (creams), water-soluble bases, and pastes are also considered ointments. *Pastes,* which contain large quantities of insoluble solids such as zinc oxide or talc, are stiffer than other types of ointments and are used often as protective barriers or absorbents.

Hydrocarbon-based ointments are typically hydrophobic, greasy, and occlusive, and inhibit the evaporation of water from the skin. In addition to this moisturizing effect, oleaginous ointments have emollient and protective qualities. When applied to intact skin, ointments form a thin film that is maintained at the application site for an extended period, providing prolonged delivery of the active ingredient.

Oleaginous ointment bases include yellow and white petrolatums, which are mixtures of liquid and solid hydrocarbons derived from petroleum (→ Waxes, Chap. 4.2). Natural waxes (e.g., beeswax) or synthetic waxes (e.g., cetyl esters wax) may be added to oleaginous bases to enhance stiffness; vegetable oils (e.g., olive oil and cottonseed oil) can be incorporated as softeners. Less commonly, vegetable oils and animal fats are themselves used as ointment bases.

Fusion of solid waxes or polyethylene resin with mineral oil produces hydrocarbon ointments that allow for better diffusion and release of drug. Ointments containing silicone oils are excellent emollients but, like hydrocarbon oils, are poor solvents for many drugs.

Emulsifiable bases share many of the qualities of oleaginous bases, but differ in their capacity to absorb water. Emulsifiable bases incorporate large quantities of water, forming water-in-oil (w/o) emulsions whose external phase retains the characteristics of an oleaginous film. Absorption bases of this type, such as hydrophilic petrolatum and anhydrous lanolin, contain emulsifying agents (e.g., cholesterol) and long-chain alcohols (e.g., stearyl alcohol and lanolin alcohols), which aid the hydration process.

Creams are semisolid emulsions that are more fluid and more easily spread than ointments. Emulsion bases include w/o emulsions, which absorb small quantities of water, and water-washable oil-in-water (o/w) emulsions.

Like oleaginous and emulsifiable bases, w/o emulsions are greasy and difficult to remove from the skin. They are effective emollients, but somewhat less occlusive than oleaginous bases. Examples of w/o emulsion bases are lanolin and petrolatum – rose water ointment. The latter is sometimes called "cold cream" because of the cooling effect produced by release and evaporation of the aqueous external phase when the cream is applied to the skin.

Water-washable o/w emulsions are commonly referred to simply as "creams." Because they are nonocclusive, nongreasy, and easily washed from the skin, these creams have relatively high esthetic appeal and patient acceptance.

Typically, creams have (1) an oleaginous internal phase composed of petrolatum-based hydrocarbons and high molecular mass alcohols or fatty acids, (2) an aqueous external phase often containing preservatives, humectants, and buffers, and (3) one or more emulsifying agents.

Commonly used emulsifiers include anionic surfactants (e.g., sodium lauryl sulfate and triethanolamine stearate) and nonionic surfactants (e.g., ethylene oxide derivatives). Cationic detergents, such as quaternary ammonium salts, are more irritating and somewhat less useful.

Because most of the aqueous phase evaporates when a cream is applied to the skin, a nonvolatile cosolvent, such as propylene glycol, is usually incorporated to assure drug solubilization in the oleaginous film.

Drug solubility may be further enhanced by water-soluble ointment bases containing liquid and waxy solid poly(ethylene glycol). These bases are nonocclusive, nonirritating, and washable from the skin. Nonetheless, bioavailability from water-soluble bases may not be improved over hydrocarbon bases because maximization of drug solubility can be offset by insufficient stratum corneum hydration.

Aqueous gels or jellies represent another class of water-soluble ointment bases in which a liquid internal phase is immobilized within a three-dimensional matrix. This system is formed by colloidal dispersions of small inorganic or large organic molecules in aqueous media. Gels can also be formed by semisolid w/o or o/w microemulsions, which may provide better solubilization of drug after application to the skin. Gelling agents include nonionic or anionic cellulose derivatives (e.g., sodium carboxymethylcellulose), acidic carboxy vinyl polymers, and other hydrophilic colloids, such as magnesium aluminum silicate, sodium alginate, and tragacanth. Some gelling agents can also be used with nonpolar solvents.

Although semisolids are an invaluable class of topical delivery system, liquid and solid dosage forms are also widely used for topical therapy. Liquid preparations applied to the skin include lotions and solutions. *Lotions* are generally aqueous preparations that spread easily on the skin. They may contain suspended solids or emulsified liquids. Topical *solutions* often contain aqueous vehicles with cosolvents, e.g., alcohol, acetone, and poly(ethylene glycol), as required for drug solubilization. Propylene glycol can also be used as a solvent, particularly in topical solutions of corticosteroids.

Topical agents are also formulated into *powders.* For example, antifungal agents, such as miconazole nitrate [*22832-87-7*], are widely used in powder form as an adjunct to treatment with antifungal solutions, creams, and ointments.

2.2. Percutaneous Absorption

Topically applied drugs often cannot reach disease sites in the epidermis or dermis in sufficiently high concentrations to exert a therapeutic effect [1, 2]. The stratum corneum is the major barrier to skin penetration. The interaction of the stratum corneum with the applied drug and vehicle largely determines the rate and extent of percutaneous absorption.

The two major processes in percutaneous absorption are *intraphase diffusion* and *interphase partitioning* [3]. These occur consecutively and repeatedly as the drug passively leaves the vehicle and permeates the various layers of the skin. The higher the diffusion and partition coefficients of a drug, the greater is the drug's ability to penetrate the skin. Drugs move from the vehicle to the viable epidermis primarily through the stratum corneum, but may also gain entry through pilosebaceous follicles or sweat ducts.

The rate of penetration is also directly proportional to the rate of vehicle release of drug, concentration of applied drug, surface area, duration of application, and degree of stratum corneum hydration. Percutaneous absorption is inversely proportional to stratum corneum thickness and, therefore, may vary, depending on anatomical site.

Transfer of drug from the vehicle to the skin surface, the initial step in percutaneous absorption, is determined by drug diffusion through the vehicle and partitioning into surface skin tissue (e.g., stratum corneum or sebum).

Solubility is a crucial factor in release of drug from the vehicle. First, dissolution of the drug in the vehicle is a prerequisite for its diffusion through the vehicle to the skin surface. Second, the relative solubility of the drug in the vehicle and skin surface is a major determinant of the partition coefficient between the two phases.

The vehicle – skin surface partition coefficient is an important factor in percutaneous absorption [4]. Increasing the lipid solubility of a drug increases its penetration into the stratum corneum and often enhances percutaneous absorption.

After reaching the skin surface, the drug must diffuse across the stratum corneum (or through the follicle) before moving relatively easily through underlying tissue. Drug diffusion through the stratum corneum is generally slow and represents the rate-limiting step in percutaneous absorption. Large lipophilic drugs, such as the corticosteroids, may be preferentially absorbed by the transfollicular route because of their relatively high diffusion rates in the sebum compared to that in the stratum corneum.

Drug transport is driven by a concentration gradient that is maintained by rapid removal of drug by the systemic circulation via dermal capillaries. The systemic circulation is often referred to as a "sink" for the drug.

2.3. Vehicles

A drug's formulation may have a significant impact on its topical potency, given the critical role of vehicle release of drug and the stratum corneum – vehicle partition coefficient in percutaneous absorption. Vehicles may enhance drug penetration by influencing solubility and release or by altering stratum corneum resistance [5].

Oleaginous vehicles, including petrolatum-based ointments and, to a lesser extent, water-in-oil emulsions, form an occlusive barrier on the skin, preventing the evaporation of sweat and insensible water loss and thereby promoting hydration of the stratum corneum. Hydration softens the stratum corneum, expanding the protein matrix by a "sponge" effect and increasing drug penetration.

Release of a drug from the vehicle is maximal from a saturated solution. This assures that a highly soluble drug is not preferentially retained in the vehicle. Among the vehicle factors that may affect solubility are pH, particle size, and choice of solvent.

The important role of drug solubility, and, therefore, choice of solvent, in determining drug release, penetrability, and bioavailability was demonstrated by numerous studies of topical corticosteroids. The solubility of corticosteroids is greatly enhanced by adding cosolvents, such as propylene glycol or dimethylsulfoxide (DMSO), to aqueous vehicles. For example, the

release of fluocinonide (0.025 wt %) from propylene glycol – water gels is greatest from vehicles containing exactly enough propylene glycol to completely dissolve the drug [6]. Excess propylene glycol reduces the bioavailability of fluocinonide by decreasing the partition coefficient of the drug between the skin and the vehicle. Similarly, maximal penetration cannot be attained from vehicles containing insufficient solubilizer.

Organic solvents, such as acetone or ethanol can also enhance the penetration of corticosteroids from topical solutions; however, the irritation caused by these solvents limits their use in topical drug therapy. Oil-in-water emulsions that contain castor oil also favor the absorption of corticosteroids.

Modifications in the vehicle after application to the skin can affect drug bioavailability. For example, adding a volatile solvent such as isopropanol to the vehicle substantially increases the penetration of corticosteroids from nonvolatile solvents, such as propylene glycol or isopropyl myristate [7]. Rapid evaporation of isopropanol after application to the skin increases the drug concentration in the nonvolatile component.

In addition to their effect on solubility, solvents may also influence the barrier properties of the skin. Propylene glycol and aprotic solvents, such as DMSO, dimethylacetamide (DMA), and dimethylformamide (DMF), interact with the stratum corneum to increase the permeability of the skin to various compounds, including glucocorticosteroids and antibiotics.

Although the mechanism of their effect is not clearly defined, studies suggest that DMSO and related compounds induce reversible changes in the otherwise impenetrable protein matrix of the stratum corneum [8]. Expansion of the stratum corneum and creation of channels by incorporation of large quantities of solvent may contribute to enhanced drug permeation. Among the drawbacks of these agents are irritation, odor, and a potential for systemic absorption and toxicity.

Relatively low concentrations of the odorless, nonirritating solvent 1-dodecylazacycloheptan-2-one [*59227-89-3*] (N-0252, Azone) have been shown to increase the penetration of antibiotics, corticosteroids, and fluorouracil [9].

3. Classification of Drugs

A broad range of pharmacological agents are employed in the treatment of skin diseases. Some drugs, such as the antiacne agent tretinoin, have been developed specifically for dermatological disorders. Glucocorticosteroids and other classes of drugs are widely used in systemic diseases, but have a unique role in cutaneous disease. This chapter focuses on the major categories of dermatotherapeutic agents to provide an overview of the most common cutaneous disorders.

3.1. Anti-Inflammatory Agents

Inflammation is a complex host response to cellular injury associated with many cutaneous disorders (→ Anti-inflammatory–Antirheumatic Drugs). Cutaneous inflammation is often mediated by immunologic mechanisms, including antigen-dependent mast cell release of histamine, leukotrienes, chemotactic factors, and other vasoactive and smooth muscle contracting substances. Inflammation can also be produced by prostaglandins and kinins [10].

Eczematous dermatitis, for example, represents a group of inflammatory skin disorders that develop in response to various exogenous or endogenous stimuli. The types of eczema include seborrheic, atopic, stasis, infectious eczematoid, and allergic or irritant contact dermatitis.

Eczema is often mediated by immune mechanisms. *Allergic contact dermatitis* is a delayed hypersensitivity reaction to an external substance (e.g., nickel, rubber, or poison ivy) in previously sensitized individuals. *Photoallergic contact dermatitis,* in which photoactivation of the compound or hapten (e.g., halogenated salicylanilides) is followed by conjugation with skin proteins and results in the formation of a complete antigen, is associated with a cell-mediated immune response similar to that in contact allergy. In contrast, *nonallergic contact dermatitis* (irritation) may be elicited immediately after initial exposure to primary chemical or mechanical irritants.

Atopic dermatitis is a common form of chronic eczema that often develops in genetically predisposed children in conjunction with hay fever or asthma. The underlying cause is

not known, but it may be related to alteration in cellular and humoral immunity mediated by cyclic nucleotides or to autonomic nervous system function [11].

Immediate-type hypersensitivity reactions may also lead to inflammation. *Urticaria* or hives, characterized by pruritic, erythematous areas of localized edema known as wheals, may result from IgE-dependent mast cell degranulation on exposure to certain foods, drugs (e.g., penicillin), or inhaled allergens, or to physical agents, such as cold, sunlight, or pressure.

Although the pathophysiologic characteristics of eczematous dermatoses may vary, the *clinical features* are similar. These include erythema, blisters, crusting, and pruritus in acute conditions and epidermal thickening, accentuated skin lines, and dryness in chronic stages. Thus, *symptomatic treatment* of eczema is relatively uniform and often includes systemic antihistamines for pruritus and topical or systemic anti-inflammatory drugs. Adjunctive therapy may include topical antipruritics, occlusive emollients, and baths. Many of the same treatment modalities are employed in urticaria.

The following four classes of drugs are frequently used in the symptomatic treatment of inflammatory disorders and, as indicated, may also benefit various other cutaneous disorders.

Antihistamines (→ Antiallergic Agents, Chap. 2). Histamine [*51-45-6*] is a substituted imidazole formed by enzymatic decarboxylation of the amino acid histidine. It is primarily synthesized and stored in mast cells and basophils. One of several chemical inflammatory mediators of immediate hypersensitivity reactions, histamine is actively secreted from IgE antibody-bound mast cells in response to a specific antigen. Histamine release is triggered by influx of calcium ions, which initiates a series of intracellular changes mediated by membrane and cytosolic signal transducing pathways [12, 13].

Histamine has a wide range of *physiologic effects* mediated through four distinct G protein-coupled receptors, H-1, H-2, H-3, and H-4 [14]. The H-1 and H-2 receptor subtypes are found in skin. Both of these receptors mediate the vasodilatory effects of histamine. H-1 receptors participate in histamine-induced increases in capillary permeability and edema. These responses are manifested in the skin as localized erythema and whealing (urticaria).

Histamine-induced stimulation of sensory nerve endings, predominantly an H-1 effect, completes the so-called "triple response" to histamine. This action causes axon-reflex vasodilation characterized by an erythematous flare. Nerve stimulation also accounts for histamine-induced pruritus and pain.

Widespread release of histamine in the body can cause extensive vasodilation and a marked fall in blood pressure. This response, combined with vascular fluid leakage, can lead to a condition resembling surgical or traumatic shock.

Extravascular smooth muscle contraction is H-1-mediated and leads to bronchoconstriction and increased force of heart contraction. In some cases, activating H-2 receptors will relax smooth muscle. In the heart, almost all of the effects of histamine on cardiac muscle, such as increases in heart rate and force of heart contraction, are through H-2 receptors. Gastric secretion is stimulated by histamine at H-2 receptors and can be attenuated by H-2 receptor antagonists [15]. H-4 receptors, present on bone marrow-derived secretory cells, play a role in inflammation and pruritus [16].

Systemic H-1 antihistamines competitively and reversively bind to H-1 receptors. These antihistamines belong to six chemical classes and all share in common a substituted tertiary amino group. The older H-1 compounds are effective histamine blockers but they also possess varying degrees of anticholinergic, antiemetic, and sedative properties. More recent H-1 blockers have been designed so as not to penetrate the blood-brain barrier and thus cause little or no sedation. These newer compounds also lack anticholinergic activity.

The original H-1 antihistamines, such as tripelennamine [*91-81-6*], pyrilamine [*91-84-9*], diphenyhydramine [*58-73-1*] and chlorpheniramine [*132-22-9*], have been widely used for symptomatic relief of histamine-induced pruritus and urticaria [17]. Antihistamines are generally not used topically because of their potential to induce contact or photocontact dermatitis. The antipruritic effects of oral H-1 antihistamines in eczematous dermatitis (e.g., atopic dermatitis) or nonallergic pruritic disorders may be primarily mediated by a sedative action. The nonsedating H-1 antihistamines, including lo-

ratadine [*79794-75-5*] and fexofenadine [*83799-24-0*], are useful for allergic rhinitis and chronic urticaria, but their effectiveness in dermatitic and pruritic conditions continues to be investigated.

Studies of H-2 blockers in the treatment of pruritus have produced mixed results. Hydroxyzine [*68-88-2*], a piperazine H-1 blocker, is a frequent choice for systemic treatment of pruritus and appears to be one of the most effective antihistamines for suppression of whealing. With the exception of cyproheptadine [*129-03-3*], H-1 antihistamines appear less effective in physical urticarias than in other forms of urticaria.

The H-2 blocker cimetidine [*51481-61-9*] has been used together with H-1 antihistamines for treatment of chronic urticaria unresponsive to H-1 therapy. Evidence exists that such combinations are more effective than the use of H-1 blockers alone, although cimetidine may worsen symptoms by blocking feedback inhibition of histamine release. Likewise, diminished T-cell suppression by cimetidine may cause exacerbation of delayed hypersensitivity.

Cromolyn sodium [*15826-37-6*] (→ Antiallergic Agents, Chap. 3), a water-soluble bis(chromone) derivative, antagonizes histamine through stabilization of the mast cell membrane rather than receptor blockade. This agent may be used orally in mastocytosis, food allergies, and dermatitis herpetiformis. Topical formulations may be beneficial in atopic dermatitis, atopic keratoconjunctivitis, and various ulcerations.

Doxepin [*1668-19-5*] (→ Psychopharmacological Agents), a tricyclic antidepressant, is a potent H-1 antagonist with mild H-2 blocking activity. Intradermally administered doxepin suppressed whealing in human skin, and topical applications were shown to be possibly effective for the control of pruritus [18]. In a double-blind study, oral doxepin was significantly more effective and less sedative than diphenhydramine [*58-73-1*] in the treatment of chronic urticaria. When tested topically, doxepin was more effective than placebo in reducing pruritus in patients with atopic dermatitis [19]. Thus, compounds in this class may possibly prove useful for controlling allergic skin reactions.

Nonsteroidal Anti-Inflammatory Drugs. These drugs are widely used for treating internal inflammation and pain but may also have some utility for treating cutaneous inflammation. Drugs in this class inhibit the production of eicosanoids, inflammatory mediators derived primarily from arachidonic acid [20]. The major eicosanoid-producing enzymes are cyclooygenases (COX) and lipoxygenases (LOX). Two enzymes, COX-1 and COX-2, produce prostaglandins. Most traditional NSAIDs, such as acetylsalicylic acid [*50-78-2*] and indomethacin [*53-86-1*], are non-selective and inhibit both COX-1 and COX-2. Gastrointestinal and renal toxicity by NSAIDS is the result of inhibiting COX-1, an enzyme needed for normal physiological processes in these two organ systems. COX-2-selective inhibitors were developed to decrease the risk of gastrointestinal ulceration.

Lipoxygenases produce several forms of lipid hydroxy acids, most notably hydroxyeicosatetraenoic acids (HETES) and leukotrienes. Inhibitors of 5-lipoxygenase, such as zilueton [*111406-87-2*], prevent the synthesis of leukotrienes. The leukotriene antagonists zafirlukast [*107753-78-6*] and montelukast [*158966-92-8*] block the action of leukotrienes at their receptors.

Studies in animals and humans show that prostaglandins and leukotrienes are produced in various skin inflammatory conditions, including ultraviolet radiation-induced erythema, contact dermatitis and psoriasis. Indomethacin has been shown to partially inhibit erythema induced by UV-B radiation (290 – 320 nm) when administered topically [21 – 23]. In contrast, topical corticosteroids were found to be inactive. Topical bufexamac [2438–72–4] has demonstrated some efficacy in steroid-responsive dermatoses other than psoriasis.

Glucocorticosteroids (→ Hormones, Chap. 4). Systemic and topical glucocorticosteroids are widely used to relieve inflammatory and pruritic manifestations of various acute and chronic skin diseases.

Glucocorticoids are 21-carbon pregnane derivatives that are synthesized in the adrenal cortex. Biological activity is dependent on the presence of a double bond at the 4 – 5 position, two keto groups (positions 3 and 20), and three hydroxy groups (positions 11, 17, and 21). Structural modification of the steroid nucleus greatly influences therapeutic potency as a result

of changes in pharmacokinetic parameters, such as absorption, protein binding, and metabolism. For example, fluorination at the 9-position, addition of a 1,2-double bond, esterification of the 17- or 21-hydroxyl groups, and inclusion of an acetonide group at positions 16 and 17 greatly increase anti-inflammatory activity [24, 25]. Replacement of the 21-hydroxyl group with chlorine results in a highly effective class of corticosteroids; these synthetic steroids, which have two or more halogen groups, are the most potent of the synthetic glucocorticoids [26].

Other structural modifications of the steroid nucleus has led to development of moderately potent glucocorticoids with less propensity to cause adverse effects such as skin atrophy. These include mometasone furoate [*83919-23-7*], a 21-chloro compound with a 17-furoate moiety, and prednicarbate [*73771-04-7*], which is a non-halogenated, 17,21-diester. These agents either have low rates of transdermal absorption or are rapidly metabolized to compounds with minimal glucocorticoid activity.

Glucocorticoids inhibit inflammation by various mechanisms. Like other steroid hormones, they initiate their biological effects by binding to intracellular receptors. Once bound to the receptor, the glucocorticoid – receptor complex translocates to the nucleus. This complex then interacts alone or with other proteins to attenuate or enhance expression of numerous genes. Consequently, almost all aspects of the inflammation process are diminished by glucocorticoids. These include reduced leukocyte infiltration into tissues, reduced expression of cytokines, decreased expression of eicosanoid-generating enzymes (phospholipases, cyclooxygenases) and proteolytic enzymes. This in part accounts for the ability of glucocorticoids to suppress cell-mediated immune responses. In skin, glucocorticoids cause vasoconstriction and inhibit proliferation of keratinocytes.

Systemically administered glucocorticoids decrease the number of polymorphonuclear leukocytes (PMNS) at sites of tissue inflammation, as reflected by an increase in circulating PMNS. These changes are accompanied by decreases in circulating lymphocytes, monocytes, and eosinophils [27, 28].

Corticosteroids also exert their anti-inflammatory effect by inhibiting T-lymphocyte and B-lymphocyte function and reduce the immune responsiveness and microbicidal activity of macrophages. Although corticosteroids inhibit antibody synthesis and related B-cell functions, they have a more pronounced effect on T-cell-mediated, delayed hypersensitivity responses. In addition, corticosteroids depress antigen-induced proliferation of T-cells and cytokine release and interfere with cytokine activities.

Topical glucocorticosteroids are particularly beneficial in inflammatory skin disorders, such as atopic and contact dermatitis, eczema, seborrheic dermatitis, and psoriasis [29]. Daily applications are generally employed to limit the possibility of tachyphylaxis. Moreover, steroids apparently are slowly released from the stratum corneum for several days after a single application.

Choice of vehicle is determined by lesion type and drug solubility and is an important factor in epidermal penetration of glucocorticoids [30]. Hydration of the stratum corneum by ointment bases may enhance steroid penetration and may be useful in dry, scaling lesions. Occlusion of the skin with an impermeable plastic film has been shown to increase steroid penetration by ca. 10-fold. Glucocorticosteroids are highly soluble in gel vehicles and are equally effective in gel or ointment bases.

The anti-inflammatory potency of glucocorticosteroids closely parallels their capacity to induce gluconeogenesis and protein catabolism. Thus, skin atrophy may result from either topical or systemic corticosteroid administration and appears to be related to inhibition of collagen synthesis. In addition to epidermal thinning and stria formation, possible adverse effects of potent halogenated topical steroids include stellate pseudoscarring, purpura, telangiectasia, acneiform and rosacea-like eruptions, and perioral dermatitis.

Glucocorticosteroids are also administered by intralesional or intramuscular injection. Among the skin conditions responsive to intralesional steroids are acne vulgaris, lichen planus, psoriasis, localized eczema, keloids, and sarcoidosis. The intramuscular route is used extensively for various acute and chronic dermatoses, but may be associated with increased risk of steroid toxicity, especially alteration of the hypothalamic – pituitary – adrenal axis.

Similarly, systemic glucocorticosteroids are important in a broad range of acute and chronic disorders (e.g., allergic contact dermatitis, atopic dermatitis, lichen planus, and herpes zoster) [31]. Their use is limited by the risk of hypothalamic – pituitary – adrenal axis suppression, cushingoid symptoms, and a host of other potential metabolic, hematologic, and infectious complications.

Steroids may be administered in single daily doses or on alternate days to minimize the potential toxicity of long-term maintenance therapy. Although single-dose therapy may be an alternative to divided doses in many acute dermatoses, achieving adequate initial control of inflammation by using alternate-day regimens may be difficult. Likewise, high-potency topical steroids are generally used for initial therapy and are replaced by less potent agents when appropriate. Cessation of steroid therapy or conversion from daily to alternate-day therapy requires careful modulation of dosage to avoid rebound flare of disease or adrenal insufficiency.

Calcineurin Inhibitors. Calcineurin, a calcium-dependent phosphatase enzyme, is activated by elevated intracellular calcium. Activated calcineurin triggers the synthesis and release of the T-lymphocyte growth stimulator, interleukin-2. Compounds with binding affinity for immunophilin proteins form a complex that binds to calcineurin and inhibits its activation. Calcineurin inhibitors thus inhibit IL-2 production and T-lymphocyte proliferation, and cause immune suppression [32].

The topically applied calcineurin inhibitors tacrolimus [*109581-93-3*] and pimecrolimus [*137071-32-0*] are used in patients with moderate to severe cases of atopic dermatitis who are unresponsive to conventional treatments [33]. Unlike topical glucocorticoids, these agents do not cause skin thinning. Short-term side effects of calcineurin inhibitors are mainly local, consisting of burning and irritation at application sites. Practitioners have been alerted to possible long-term effects of cancer, however these warnings are based on the compounds' immunosuppressive properties and findings from animal toxicity studies.

Topical Antipruritics, Emollients, and Baths. Lotions containing antipruritics, such as menthol [*1490-04-6*], phenol [*108-95-2*], camphor [*76-22-2*], and astringents, such as zinc oxide [*1314-13-2*] and calamine [*8011-96-9*], are often used for symptomatic treatment of acute eczema and urticaria. Wet dressings containing aluminum acetate [*139-12-8*] solution or bath preparations containing colloidal solids, such as oatmeal or *sulfur* [*7704-34-9*], may also be beneficial.

Antipruritic tar baths and emollient oil baths are helpful for chronic lesions. Occlusive emollient creams or ointment bases, including anhydrous lanolin [*8006-54-0*] and petrolatum [*8009-03-8*], decrease evaporative loss of water from the skin surface and may be applied to relieve dryness. Urea [*57-13-6*] and lactic acid [*50-21-5*] are frequently incorporated into these emollient bases to enhance moisture retention and soften the stratum corneum.

Local anesthetics, including the ester derivative benzocaine [*94-09-7*], depress the conduction of nerve impulses and may provide relief of pain and itching. Their use in eczema is limited because they may cause contact sensitization.

3.2. Antiacne Agents

Acne vulgaris is a common dermatological disorder characterized by noninflammatory comedones (open or closed) and inflammatory papules, pustules, and cysts.

Therapy may be directed at one or more of the following pathophysiologic processes associated with acne: increased sebum production, hyperkeratinization, *Propionibacterium acnes* (*P. acnes*) proliferation, and inflammation.

The initial *stimulus for acne development* appears to be an alteration in the normal pattern of follicular epithelial cell differentiation known as abnormal follicular keratinization [34]. The cause of this defect is unknown, but may relate to hormonal changes, increased sebum production, or bacterial colonization of the hair follicles. The altered keratinization process is characterized by production of a mass of tightly coherent keratinocytes and results in blockage of sebum flow, overgrowth of *P. acnes*, and follicular dilation to form microcomedones. Breakdown of the follicular wall initiates the inflammatory phase of acne, which is mediated by bacteria-derived enzymes and chemotactic factors [35].

Consequently, therapeutic agents that reverse the underlying keratinization defect are particularly effective in clearing microcomedones and preventing their formation. They also indirectly prevent inflammatory lesions. Antibacterial, anti-inflammatory, and sebostatic agents have a pronounced effect on inflammatory lesions, but do not correct the underlying abnormal follicular keratinization. Topical keratolytic or peeling agents, including sulfur, resorcinol [*108-46-3*], and salicylic acid [*69-72-7*], do not prevent lesion formation and, therefore, provide only temporary clearing.

Comedolytic Agents. The *retinoids,* including vitamin A and its derivatives, modify the altered epithelial replication process characteristic of acne, psoriasis, ichthyosis, and other keratinizing dermatoses [36].

In a study comparing the antikeratinizing effects of topical antiacne agents in rhino mouse skin, only the retinoids produced substantial reduction in the size of utriculi (keratinized follicles) [37]. Of the retinoids, all-*trans*-retinoic acid [*302-79-4*] (tretinoin, → Vitamins, Chap. 2.1) exhibited the greatest comedolytic potency. Tretinoin also gave greater reduction in comedones than motretinide [*56281-36-8*] in clinical studies [38]. Orally administered retinoids showed less comedolytic activity than topical retinoids in an earlier study using the rhino mouse model [39]. Oral tretinoin had more than twice the potency of oral 13-*cis*-retinoic acid [*4759-48-2*] (isotretinoin).

Tretinoin is a natural retinoid metabolite of vitamin A that binds to cellular retinoic acid binding proteins in the cytoplasm and receptors in the nucleus. The nuclear receptors are mechanistically considered the site of action of tretinoin. Tretinoin activates the three members of the retinoic acid (RAR) nuclear receptors (RARα, RARβ, RARγ) which act to modify gene expression, and subsequent protein synthesis, to account for its beneficial clinical effects [40].

Tretinoin is perhaps the most effective topical agent for treatment and prevention of comedonal and inflammatory acne. Tretinoin decreases horny cell cohesiveness and stimulates epidermal cell turnover, thereby removing existing microcomedones and preventing their reformation. Tretinoin may enhance resolution of inflammatory lesions by increasing blood flow. Moreover, its ability to reduce stratum corneum thickness and to potentiate the absorption of other topical compounds makes tretinoin a potentially valuable component in combination acne therapy.

Synthetic receptor-selective topical retinoids are also available to treat acne vulgaris. Adapalene [*106685-40-9*], a naphthoic acid derivative, and tazarotene [*118292-40-3*], an acetylenic retinoid, are selective for RARβ and RARγ [41, 42]. Tazarotene does not bind to the RARs but is metabolized by an esterase to the active form, tazarotenic acid, which binds to RARs and activates them.

Antibacterial Agents. The oxidizing agent benzoyl peroxide [*94-36-0*] is commonly used in the topical treatment of acne. Benzoyl peroxide possesses both comedolytic and antibacterial properties, but the latter is believed to be primarily responsible for its therapeutic efficacy. Release of free-radical oxygen by benzoyl peroxide decreases the levels of *P. acnes* and skin surface lipids, leading to reduction in inflammatory lesions. Benzoyl peroxide also reduces inflammatory lesions via stimulation of blood flow and may have a sebostatic effect. Acetone or water-based gel formulations may allow for enhanced penetration and efficacy of benzoyl peroxide. The combination of benzoyl peroxide with erythromycin [*114-07-8*], a macrolide antibiotic (→ Antibiotics, Chap. 3.5.2), is more effective than either ingredient alone in reducing inflammatory lesions.

Oral tetracyclines or erythromycin are the drugs of choice for treatment of moderate inflammatory acne, with cotrimoxazole (trimethoprim [*738-70-5*] plus sulfamethoxazole [*723-46-6*]) occasionally used in refractory patients. Topical antibiotics, including erythromycin [*114-07-8*], clindamycin [*18323-44-9*] (→ Antibiotics, Chap. 4), tetracycline [*60-54-8*] (→ Antibiotics, Chap. 4), and meclocycline [*2013-58-3*], are generally recognized as safe and effective alternatives to long-term systemic therapy [43, 44].

The mechanism of action of oral and topical antibiotics in acne is not fully understood. Previously, oral antibiotics were shown to suppress follicular bacterial colonization and production of free fatty acids. Later studies suggested that

inhibition of bacterial metabolism (e.g., lipase production) may be the primary mode of action of topical antibiotics [45]. Topical antibiotics also possess anti-inflammatory properties mediated in part by inhibition of neutrophil chemotaxis [46].

Azelaic acid [*123-99-9*] (→ Dicarboxylic Acids, Aliphatic), a naturally occurring saturated dicarboxylic acid, is topically effective in treating mild to moderate inflammatory acne vulgaris. The exact mechanism is not known but it has antimicrobial activity against *P. acnes* and *S. epidermidis* with normalization of keratinization.

Systemic Retinoids. Oral isotretinoin provides excellent therapeutic results and long-term remissions in patients with severe cystic acne [47]. Its use is limited by a high incidence of mucocutaneous (xerosis, cheilitis) and systemic adverse effects (vertebral hyperostoses, hypertriglyceridemia) and because it is teratogenic in humans. Women of child-bearing age must have a pretreatment pregnancy test and be ensured of adequate contraception for the duration of therapy.

The effects of isotretinoin in acne are related primarily to marked suppression in sebum production that results from a direct effect on the sebaceous gland [48]. The decrease in sebum production is accompanied by reduction in bacterial flora [49]. Less clearly characterized is the inhibitory effect of isotretinoin on follicular hyperkeratinization [50]. Isotretinoin may also exert an anti-inflammatory effect with inhibition of neutrophil function.

Hormones. Elevated levels of androgens are known to cause acne in some women. High levels of androgens lead to increased sebum production and follicular hyperkeratosis. Oral contraceptives can reduce androgen levels and are effective in treating acne in women. This therapy for acne targets sex hormone binding globulin (SHBG) and free testosterone [*58-22-0*]. A low dose oral contraceptive containing norgestimate [*35189-28-7*] and ethinyl estradiol [*57-63-6*] (→ Contraception, Chap. 4.1) is a safe and effective treatment of acne with efficacy similar to benzoyl peroxide and topical or systemic antibiotics [51]. This combination increases SHBG and decreases free testosterone in healthy women. Norgestimate binds to progestin receptors and has low affinity for androgen receptors and SHBG. Ethinyl estradiol may have antiandrogenic effects by suppressing the secretion of gonadotropins and inhibiting the production of ovarian androgens.

Spironolactone is an antiandrogen that blocks the binding of androgen to the androgen receptor and is useful in treating recalcitrant acne in women. Combinations of synthetic estrogens with nonandrogenic progestogens or the synthetic antiandrogen cyproterone acetate [*427-51-0*] (→ Hormones, Chap. 3.6.3) have been used successfully to treat female acne patients with acne conglobata.

Anti-Inflammatory Agents. Short-term therapy with oral corticosteroids (e.g., prednisone [*53-03-2*]) in conjunction with maintenance therapy with sulfones (e.g., dapsone [*80-08-0*]) is sometimes used in patients with severe acne conglobata. Dapsone has been shown to stabilize lysosomes and interfere with neutrophil function. Intralesional injections of triamcinolone acetonide [*76-25-5*] are frequently used for patients with severe scarring. Ibuprofen [*15687-27-1*] has also shown some efficacy in diminishing inflammatory acne.

3.3. Antipsoriatics

Psoriasis is a common, chronic skin disorder of variable and unpredictable course. It is characterized by accelerated epidermal cell turnover and inflammation to produce elevated, erythematous plaques covered with silver scale. The underlying cause of the disease is unknown, but various environmental, genetic, and immunologic factors influence its expression. A familial predisposition to psoriasis is evidence of one or several genes accounting for its expression. A dysregulation in the immune system, primarily with T-lymphocytes, is considered to be the fundamental mediator of the epidermal and inflammatory changes in the psoriatic lesion [52]. External stimuli, such as infectious disease, skin trauma, or psychological stress are known triggers of psoriatic plaque formation.

Topical therapy is most effective for short-term use in mild psoriasis and may be combined with ultraviolet radiation or systemic agents in extensive or resistant cases. Many of the ther-

apeutic modalities for severe psoriasis (photochemotherapy, phototherapy, and cytotoxic therapy) directly suppress immune cell activity. Some of these agents also diminish epidermal hyperplasia by inhibiting DNA synthesis and mitosis.

Topical Agents. Occlusive emollients, such as petrolatum and mineral oil [*8012-95-1*], are frequently used alone or in combination with keratolytic agents, such as salicylic acid [*69-72-7*] and urea [*57-13-6*], to facilitate hydration and removal of hyperkeratotic plaques.

The topical synthetic vitamin D analog, calcipotriene [*112965-21-6*] (calcipotriol, → Vitamins, Chap. 3.3) is frequently used for treating moderate plaque psoriasis. This agent acts to enhance keratinocyte differentiation and also exhibits immunomodulatory activity [53]. The major local side effect of calcipotriene is skin irritation; unlike corticosteroids, calcipotriene does not cause skin atrophy. Adverse systemic effects of topical calcipotriene are limited due to its rapid metabolic inactivation within 24 h after it is applied.

Potent topical corticosteroids (e.g., clobetasol propionate [*25122-46-7*] and halobetasol propionate [*66852-54-8*]) are frequently administered intermittently in more severe cases of psoriasis because of their vasoconstrictive, antimitotic, and anti-inflammatory effects. Glucocorticoids also have significant inhibitory effects on immune cell activation. Intralesional steroids are also used in psoriasis. Long-term use of corticosteroids results in a loss of effectiveness and an increased risk of toxicity.

Coal tar preparations have long been used safely and effectively in psoriasis. Coal tar is a complex and variable mixture of at least 10 000 compounds; it contains many phenols, bases, and polyaromatic hydrocarbons, including several potent photosensitizers and carcinogens. Although its mechanism of action in psoriasis is not well-defined, coal tar inhibits DNA synthesis and may also have antipruritic and vasoconstrictive properties. Coal tar is commonly used in conjunction with ultraviolet radiation as described below (see page 12).

Anthralin [*480-22-8*] (dithranol), a synthetic derivative of the natural antipsoriatic agent chrysarobin, depresses DNA synthesis and epidermal replication by interfering with polyamine production. It may also exert anti-inflammatory and immunosuppressive effects. Like coal tar, anthralin may cause skin irritation and staining.

A synthetic retinoid, tazarotene, is topically effective for moderate psoriasis. Retinoic acid [*302-79-4*] has also been applied topically in psoriasis, as have several cytotoxic agents, including the sulfur and nitrogen mustards (e.g., mechlorethamine hydrochloride [*55-86-7*]).

Phototherapy. Ultraviolet radiation is well-known for its therapeutic value in psoriasis. Its biologic effects, thought to account for its efficacy in psoriasis, are wide-ranging; it can suppress the immune system and inhibit mitosis. These properties are primarily attributable to the UV-B wavebands (290 – 320 nm), although relatively high doses of UV-A (320 – 400 nm) may also be beneficial. UVB phototherapy is often used for patients who either do not respond to topical antipsoriatic drugs or for those who find topical treatment inconvenient.

Ultraviolet radiation is widely used in combination with anthralin and/or coal tar in the treatment of psoriasis. The *Goeckerman regimen* generally consists of multiple daily applications of crude coal tar ointment, a daily tar bath, and subsequent exposure to UV-B and UV-A radiation.

Whether or not the photosensitizing action of coal tar is of therapeutic relevance in psoriasis is unclear. Although the action spectrum of coal tar phototoxicity is in the UV-A range and the inhibitory action of coal tar on DNA synthesis is enhanced by UV-A radiation, coal tar is not photosensitive at commonly used doses of UV-A. Moreover, UV-B radiation–coal tar appears to be more effective than UV-A radiation–coal tar in psoriasis. Other studies suggest that the contribution of coal tar to the efficacy of Goeckerman therapy is minimal [54]. The *Ingram regimen* is a variation of the Goeckerman regimen in which an anthralin – salicylic acid paste is applied after tar baths and UV radiation.

In addition to tar and anthralin, phototherapy (UVB) effectiveness can be enhanced when combined with other topical and systemic drug treatments. These include corticosteroids, calcipotriene, methotrexate [*59-05-2*] and retinoids.

Systemic Therapy. Use of systemic drugs is reserved for cases of severe or widespread psoriasis. These drugs fall into three main categories: cytotoxic agents, retinoids, and immunomodulators.

Cytotoxic drugs exert a suppressive effect on rapidly proliferating tissue by blocking nucleic acid synthesis. They are used primarily in cancer chemotherapy, but may be used to treat severe, recalcitrant psoriasis (→ Cancer Chemotherapy). Methotrexate [*59-05-2*], a folic acid antagonist (→ Cancer Chemotherapy, Chap. 2.1), is considered the most effective chemotherapeutic agent for psoriasis [55]. Hepatotoxicity and immunosuppression are major drawbacks to long-term therapy with methotrexate.

Hydroxyurea [*127-07-1*], 6-thioguanine [*154-42-7*] (→ Cancer Chemotherapy, Chap. 2.7), azathioprine [*446-86-6*], and mycophenolic acid [*24280-93-1*] are other antimitotics that have been used to treat psoriasis. Like methotrexate, their use is limited by severe toxicity.

Oral retinoids, including the aromatic retinoic acid derivative etretinate [*54350-48-0*] and its corresponding carboxylic acid form, acitretin [*55079-83-9*], are highly effective in selected forms of psoriasis. The pustular and exfoliative subtypes of psoriasis are more responsive to these retinoids than plaque psoriasis. Acitretin has a more favorable elimination profile from the body than etretinate, with a half-life of about 50 hours for the acid versus 120 days for the ester. Etretinate may persist for several years because of its tendency to be stored in body fat. Systemic retinoids are sometimes combined with psoralen photochemotherapy [56].

Immunomodulatory compounds directly prevent T-lymphocyte activation or prevent cytokines from acting on their target cells. Cytokines released by activated T-cells are considered the primary inducers of the accelerated keratinocyte proliferation and inflammation so characteristic to psoriasis.

Cyclosporine (cyclosporin A) [*59865-13-3*], an orally administered calcineurin inhibitor, suppresses the immune system by preventing the production of IL-2 release by T-lymphocytes [57]. Cyclosporine is highly effective for all forms of psoriasis, especially those exhibiting severe inflammation. It is commonly used to bring severe psoriasis under control, followed by a switch to other systemic or topical drugs for maintenance therapy. Long-term use of cyclosporine is limited because of its side effects of renal toxicity and hypertension.

Biologic agents, which are large protein molecules synthesized by living organisms, have been designed to specifically interfere with abnormal T-lymphocyte or cytokine functions [33]. These agents are highly effective therapeutic regimens for moderate to severe psoriasis. Because of their immunosuppressive properties, biologic treatments could potentially augment the rates of infections and cancer in susceptible patient populations.

Alefacept, a dimeric human fusion protein that binds to the CD2 lymphocyte receptor, is given by intravenous or intramuscular routes. The interaction of alefacept with CD2 causes inhibition or enhancement of T-lymphocyte binding to other immune cells. Clearing of psoriasis by alefacept is accompanied by reduced numbers of T-lymphocytes and lowered cytokine levels in lesional skin [58].

The T-lymphocyte is also the target of *efalizumab*, a humanized monoclonal antibody administered subcutaneously. Efalizumab binds to CD11a, the alpha-subunit of lymphocyte function-associated antigen-1 (LFA-1). This binding prevents LFA-1 on T-cells from interacting with intercellular adhesion molecules on endothelial cells and antigen presenting cells, thereby blunting the inflammatory process.

A key pathogenic cytokine in psoriasis is tumor necrosis factor-α (TNF-α). *Etanercept* is a subcutaneously administered dimeric fusion protein that contains the ligand-binding portion of the TNF-α receptor. By acting as a soluble TNF-α receptor, etanercept binds to free TNF-α and prevents it from interacting with the TNF-α receptors on immune and skin cells.

Photochemotherapy. Although the effects of systemic chemotherapy are nonselective, the combination of a photosensitive compound and UV radiation restricts the therapeutic action of a drug to areas of epidermal hyperplasia. An ideal photosensitizer is activated by relatively low doses of UV radiation and is otherwise inert.

Psoralens are natural furocoumarins well-known for their photosensitizing potential. Porphyrins, including hematoporphyrin derivatives (HPD), are also potent photosensitizers. Both

psoralens and porphyrins have an action spectrum in the UV-A range. Porphyrins are also activated by visible light.

Photoactivated psoralens covalently bind to DNA pyrimidines, producing monoadducts and cross-links, thereby inhibiting DNA synthesis and cell proliferation. Suppressive effects of PUVA (psoralen and UV-A) on the immune system, particularly T-lymphocytes, have been demonstrated in animals and humans. The mechanism of action of PUVA therapy in psoriasis is presumably related to its antiproliferative and immunosuppressive effects [59].

In PUVA therapy, 8-methoxypsoralen [*298-81-7*] is orally administered 2 h before exposure to increasing doses of UV-A. An average of 20 treatments over a 3-month period has been shown to be highly effective as initial therapy for psoriasis. In addition, PUVA therapy is valuable in treating other skin disorders, including vitiligo, mycosis fungoides, and atopic dermatitis.

Long-term side effects of PUVA primarily affect skin, immunologic and ocular systems. With PUVA, premature aging of skin is common but the major concern is increased risk of cutaneous malignancies. The risk of squamous cell carcinoma is definitely increased in patients treated over many years with PUVA [60]. It is also possible that the immunosuppressive effects associated with PUVA therapy can contribute to development of skin neoplasia. Eye protection during PUVA therapy is essential to prevent cataracts.

3.4. Antifungals (→ Antimycotics)

Superficial fungal infections are frequently caused by dermatophytes of the genera *Microsporum, Epidermophyton,* and *Trichophyton.* The most common dermatophyte infections are tinea cruris, tinea corporis, tinea pedis, tinea unguium, and tinea capitis. Superficial mycoses are also caused by the yeasts *Malassezia ovalis*, the etiological agent in tinea versicolor, and *Candida*.

The causative organisms are predominantly localized in keratinized areas of the skin, hair, and nails.

Topical agents are generally satisfactory for the treatment of dermatophytosis, cutaneous candidiasis, and tinea versicolor, but are often ineffective for infections of the nail and scalp. Oral agents are usually required for extensive or refractory dermatophytosis or tinea versicolor.

The following drugs are commonly employed in the treatment of superficial fungal or yeast infections. Some of these agents are also used to treat systemic mycoses.

Azoles (→ Antimycotics, Chap. 2). The broad-spectrum azoles are a large family of antifungal agents classified as belonging chemically to either imidazole or triazole classes [61], and are active against dermatophytes, yeasts, dimorphic fungi, some gram-positive bacteria, and certain protozoa. They exert their antifungal effect primarily by blocking the synthesis of ergosterol, an important regulator of cell membrane permeability [62, 63].

Topical azoles, such as clotrimazole [*23593-75-1*] and econazole [*27220-47-9*], are effective in treating cutaneous candidiasis and most tinea infections [64 – 67]. They are generally considered superior to such agents as tolnaftate, haloprogin, and undecylenic acid. No significant difference in the safety or efficacy of various azoles has been established.

Combinations of topical azoles and corticosteroids (e.g., miconazole plus hydrocortisone; clotrimazole plus betamethasone dipropionate [*5593-20-4*]) are popular for acute inflammatory mycoses. The steroid component appears to provide rapid symptomatic relief without interfering with antifungal efficacy [68]. Also, some topical azoles have been reported to exert an anti-inflammatory effect [69].

itraconazole, and fluconazole are orally active agents. Ketoconazole and itraconazole are highly effective in all superficial fungal infections. Oral ketoconazole is generally used to treat deep mycoses (e.g., paracoccidioido mycosis and histoplasmosis) or resistant superficial mycoses (e.g., chronic mucocutaneous candidiasis) and is the drug of choice in tinea versicolor. Itraconazole and fluconazole are effective in the treatment of onychomycosis, a common fungal infection of the fingernail and toenail as well as systemic infections [70, 71]. Fluconazole primarily focuses on yeasts (*Candida albicans*).

Allylamines. Terbinafine hydrochloride [*78628-80-5*] is a second generation allylamine structurally related to naftifine. Terbinafine is 10 to 100 times more potent than naftifine in vitro.

Both agents inhibit squalene oxidase and decrease ergosterol, an essential component of cell membranes in fungi. Topical terbinafine is indicated for tinea infections while oral terbinafine is indicated for the treatment of onychomycosis of the toenail or fingernail due to dermaytophytes.

Butenafine hydrochloride [*101827-46-7*] is a benzylamine, structurally related to the allylamines, and is indicated for tinea pedis. It is active against *E. floccosum, T. mentagrophytes and T. rubrum.*

Ciclopirox. Ciclopirox olamine [*41621-49-2*] (→ Antimycotics, Chap. 6), a pyridone–ethanolamine salt, is structurally dissimilar to other antifungal agents. Its primary mode of action appears to be inhibition of transmembrane transport of substrates and ions essential for cell metabolism and energy production. Ciclopirox is similar to the imidazoles in antimicrobial spectrum, therapeutic uses, and efficacy. Limited studies suggest superior penetration of topical ciclopirox into horny material and nails.

Acidic Compounds. Acidic compounds reduce the local pH to levels unfavorable for dermatophyte growth. Undecylenic acid [*112-38-9*], an 11-carbon unsaturated fatty acid, is the most widely used compound in this class; however, it is effective in only 50 % of the patients. Zinc or calcium salts of undecylenic acid are also used.

Similarly, the fungistatic activity of triacetin [*102-76-1*] (glyceryl triacetate) results from its hydrolysis by fungal esterases to acetic acid.

Tolnaftate. Tolnaftate [*2398-96-1*] is a thiocarbamate derivative similar in spectrum of action and efficacy to undecylenic acid. It is most often used in dermatophyte and tinea versicolor infections.

Haloprogin. An iodinated trichlorophenol, haloprogin [*777-11-7*] has a broad spectrum of activity against dermatophytes, *Candida* and *Pityrosporon.* It is similar to tolnaftate in efficacy for tinea cruris, tinea corporis, and tinea pedis.

Morpholine Derivatives. Amorolfine [*67467-83-8*] is a morpholine derivative which possesses both fungistatic and fungicidal activity in vitro [72]. Amorolfine is an effective antifungal agent for patients with onychomycosis and dermatomycosis caused by dermatophytes, yeasts, and sensitive molds.

Griseofulvin (→ Antimycotics, Chap. 5)**.** Griseofulvin [*126-07-8*] is a safe, effective, orally administered fungistatic antibiotic that inhibits cell mitosis by binding to microtubules during metaphase. Griseofulvin is active only against dermatophytes and concentrates in the keratinized outer layer of the stratum corneum. Absorption of griseofulvin is increased by fat intake, reduction in particle size, and multiple-dose regimens.

Griseofulvin is effective for superficial dermatophyte infections of the skin, hair, and nails and is generally reserved for infections that have failed to respond to topical therapy. It is considered the drug of choice for tinea capitis. Griseofulvin has no action against *C. albicans* or tinea versicolor.

Polyenes. Nystatin [*1400-61-9*] (→ Antimycotics, Chap. 3) is a polyene macrolide antibiotic that binds to sterol constituents in the yeast cell membrane, thereby altering its permeability. Nystatin is active against various yeast and fungi, but is most often used topically in treating cutaneous and mucocutaneous candidiasis. Nystatin is ineffective in dermatophyte infections.

Like nystatin, amphotericin B [*1397-89-3*] (→ Antimycotics, Chap. 3.1) exerts its antifungal activity by an effect on cell membrane permeability. It may also increase T-cell and B-cell responsiveness. Topical amphotericin is used primarily for cutaneous and mucocutaneous candidiasis and has been shown to be as effective as nystatin in intertriginous candidiasis.

Amphotericin is commonly administered intravenously for the management of various deep mycoses. It has been associated with significant toxicity, including impaired renal function.

Miscellaneous Topicals. Diiodohydroxyquinoline [*83-73-8*] (iodoquinol) and iodochlorhydroxyquinoline [*130-26-7*] (clioquinol) are antimicrobial agents used alone or in combination with cortisol (hydrocortisone) for treating *Candida* and dermatophyte infections, although their efficacy has not been definitely established.

Gentian violet [*548-62-9*] is a mixture of methylpararosanilines used for cutaneous or mucocutaneous candidiasis. Castellani's paint

[*8052-17-3*] contains basic fuchsin, boric acid, phenol, and resorcinol in a water–ethanol–acetone solution and is of benefit in subacute and chronic dermatophytosis. Whitfield's ointment, a 2 : 1 combination of the fungistatic agent benzoic acid [*65-85-0*] and the keratolytic salicylic acid, is helpful in the treatment of chronic dermatophytosis of the palms and soles and in tinea versicolor. Sodium thiosulfate [*10102-17-7*], selenium sulfide [*7488-56-4*], and acrisorcin [*7527-91-5*] are primarily effective in tinea versicolor. Zinc pyrithione [*13463-41-7*] is effective as an anti-*Pityrosporum* agent.

3.5. Antiaging

Skin undergoes aging by two separate and distinct processes: intrinsic aging and photoaging. Chronic sun exposure of the skin leads to photoaging, defined as photodamage superimposed on the natural (intrinsic) aging process. In contrast to intrinsically aged skin, photoaging is manifested by coarse wrinkles, reduced elasticity, darkened and uneven pigmentation and roughness. The main histologic features of intrinsic aging are epidermal and dermal atrophy while the hallmark of photoaging is dermal elastosis. Photodamage is likely mediated by destruction of extracellular matrix (collagen and elastin) of the dermis by matrix degrading metalloproteinases.

Topical tretinoin is a pharmacologic agent shown to have proven efficacy in treating photoaged skin [73, 74]. Tretinoin can enhance the appearance of photodamaged skin by effacing fine wrinkles, lightening hyperpigmented age spots, reducing roughness and improving the overall texture of the skin surface.

Botulinum toxin A, a protein produced by *Clostridium botulinum*, is indicated for the effacement of large wrinkles between the eyes (glabellar lines). The toxin, when injected subcutaneously, blocks the contractile function of subcutaneous muscles, thereby lessening their ability to furrow the skin into a creased pattern [75]. Botulinum toxin type A is taken up by nerve terminals at neuromuscular junctions and subsequently inhibits the release of acetylcholine from synaptic vesicles. Due to its endopeptidase activity, the toxin specifically cleaves and inactivates SNAP-25 protein, which is required for the fusion of synaptic vesicles with the plasma membrane [76].

4. References

1. R. J. Scheuplein, *Curr. Probl. Dermatol.* **7** (1978) 172–186.
2. B. Idson, *J. Pharm. Sci.* **64** (1975) no. 6, 901 – 924.
3. J. Ostrenga, C. Steimetz, B. J. Poulsen, *J. Pharm. Sci.* **60** (1971) no. 8, 1175 – 1183.
4. T. Higuchi, *J. Soc. Cosmet. Chem.* **11** (1960) 85 – 97.
5. M. F. Coldman, B. J. Poulsen, T. Higuchi, *J. Pharm. Sci.* **58** (1969) 1098 – 1102.
6. J. Haleblian, B. J. Poulsen, K. H. Burdick, *Curr. Ther. Res.* **22** (1977) no. 5, 713 – 721.
7. M. F. Coldman, B. J. Poulsen, T. Higuchi, *J. Pharm. Sci.* **58** (1969) no. 9, 1099 – 1102.
8. A. M. Kligman, *JAMA* **193** (1965) no. 10, 140 – 148.
9. R. B. Stoughton, *Arch. Dermatol.* **118** (1982) 474 – 477.
10. E. M. Smyth, A. Burke G. A. FitzGerald: "Lipid-Derived Autacoids: Eicosanoids and Platelet-Activating Factor," in L. L. Brunton, J. S. Lazo, K. L. Parker (eds.): *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th ed., McGraw-Hill, New York 2006, pp. 653 – 670.
11. J. M. Hanifin, S. C. Chan, *J. Invest. Dermatol.* **105** (1995) suppl. 1, 84S – 88S.
12. M. J. Fischer, J. J. Paulussen, R. Roozendaal et al., *Inflamm. Res.* **45** (1996) 564 – 573.
13. M. M. Hamawy, S. E. Mergenhagen, R. P. Siraganian, *Cell. Signal.* **7** (1995) 535 – 544.
14. R. A. Skidgel, E. G. Erdös: "Histamine, Bradykinin and Their Antagonists," in L. L. Brunton, J. S. Lazo, K. L. Parker (eds.): *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th ed., McGraw-Hill, New York 2006, pp. 629 – 651.
15. K. Pattichis, L. L. Louca, *Drug Metabol. Drug Interact.* **12** (1995) 1 – 36.
16. I. J. de Esch et al., *Trends Pharmacol. Sci.* **26** (2005) 462 – 469.
17. O. B. Christensen, H. I. Maibach, *Semin. Dermatol.* **2** (1983) no. 4, 270 – 280.
18. J. E. Bernstein, D. H. Whitney, K. Soltani, *J. Am. Acad. Dermatol.* **5** (1981) 582 – 585.
19. L. A. Drake, J. D. Fallon, A. Sober, *J. Am. Acad. Dermatol.* **31** (1994) 613 – 616.

20. A. Burke, E. Smyth, G. A. FitzGerald: "Analgesic-Antipyretic Agents, Pharmacotherapy of Gout," in L. L. Brunton, J. S. Lazo, K. L. Parker (eds.): *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th ed., McGraw-Hill, New York 2006, pp. 671 – 715.
21. M. E. Goldyne, *J. Invest. Dermatol.* **64** (1975) 377 – 385.
22. D. S. Snyder, W. H. Eaglstein, *Br. J. Dermatol.* **90** (1974) 91 – 93.
23. W. H. Eaglstein, L. D. Ginsberg, P. M. Mertz, *Arch. Dermatol.* **115** (1979) 1421 – 1423.
24. J. Elks, *Br. J. Dermatol.* **94** (1976) suppl. 12, 3 – 13.
25. R. G. Dluhy, D. P. Lauler, G. W. Thorn, *Med. Clin. North Am.* **57** (1973) 1155 – 1167.
26. S. Yawalkar et al., *J. Am. Acad. Dermatol.* **25** (1991) 1137 – 1144.
27. J. Thompson, R. Van Furth: "The Effect of Glucocorticosteroids on the Kinetics of Monocytes and Peritoneal Macrophages," in Van Furth (ed.): *Mononuclear Phagocytes*, Blackwell Scientific Publ., Oxford 1970, pp. 255 – 281.
28. T. F. Dougherty: "Effects of Corticosteroid Hormones on Lymphatic Tissues," in J. W. Rebuck (ed.): *The Lymphocytes and Lymphocytic Tissue*, Hoeber, New York 1960, pp. 112 – 124.
29. R. C. Cornell, H. I. Maibach, "Clinical Indications: Real and Assumed," in H. I. Maibach, C. Surber (eds.): *Topical Corticosteroids*, Karger, Basel 1992, pp. 154 – 162.
30. R. B. Stoughton, *Arch. Dermatol.* **106** (1972) 825 – 827.
31. R. M. Fine, *Semin. Dermatol.* **2** (1983) no. 4, 250 – 256.
32. E. Bornhovd, W. H. Burgdorf, A. Wollenberg, *J. Am. Acad. Dermatol.* **45** (2001) 736 – 743.
33. A. B. Gottlieb, *J. Am. Acad. Dermatol.* **53** (2005) S3 – S16.
34. J. J. Leyden, A. M. Kligman: "The Role of Bacteria in Acne Vulgaris," in R. Fleischmajer (ed.): *Progress in Diseases of the Skin*, vol. **2**, Grune & Stratton, Orlando 1984, pp. 21 – 29.
35. J. Knop, T. Bossecker, P. M. Kovary, *Arch. Dermatol. Res.* **274** (1982) 267 – 275.
36. G. L. Peck: "Retinoids in Clinical Dermatology," in R. Fleischmajer (ed.): *Progress in Diseases of the Skin*, vol. **1**, Grune & Stratton, Orlando 1981, pp. 227 – 269.
37. J. A. Mezick, M. C. Bhatia, L. M. Shea, E. G. Thorne et al.: "Anti-Acne Activity of Retinoids in the Rhino Mouse," in H. I. Maibach, N. J. Lowe (eds.): *Models in Dermatology*, vol. **2**, Karger, Basel, Switzerland 1985, pp. 59 – 63.
38. K. Nordin, T. Fredricksson, C. Rylander, *Dermatologica* **162** (1981) 104 – 111.
39. J. A. Mezick, M. C. Bhatia, R. J. Capetola, *J. Invest. Dermatol.* **83** (1984) 110 – 113.
40. G. J. Fisher, J. J. Voorhees, *FASEB J.* **10** (1996) 1002 – 1013.
41. B. Shroot, S. Michel, *J. Am. Acad. Dermatol.* **36** (1997) part 2, S96 – S103.
42. R. A. S. Chandraratna, *J. Am. Acad. Dermatol.* **37** (1997) part 3, S12 – S17.
43. E. A. Eady, K. T. Holland, W. J. Cunliffe, *J. Am. Acad. Dermatol.* **5** (1981) 455 – 489.
44. R. B. Stoughton, *Arch. Dermatol.* **115** (1979) 486 – 489.
45. S. J. Adams, E. M. Cooke, W. J. Cunliffe, *J. Antimicrob. Chemother.* **7** (1981) suppl. A, 75 – 80.
46. N. B. Esterly, N. L. Furey, L. E. Flanagan, *J. Invest. Dermatol.* **70** (1978) 51 – 55.
47. G. L. Peck, T. G. Olsen, D. Butkus, *J. Am. Acad. Dermatol.* **6** (1982) 735 – 745.
48. J. S. Strauss, A. M. Stranieri, L. N. Farrell et al., *J. Invest. Dermatol.* **74** (1980) 66 – 67.
49. G. Plewig, J. Nikolowski, H. H. Wolff, *J. Am. Acad. Dermatol.* **6** (1985) 766 – 785.
50. P. M. Elias, P. Fritsch, M. A. Lampe et al., *Clin. Res.* **28** (1980) 248.
51. A. W. Lucky et al., *J. Am. Acad. Dermatol.* **37** (1997) 746 – 754.
52. E. Christophers, *Int. Arch. Allergy Immunol.* **110** (1996) 199 – 206.
53. K. Kragballe, *J. Am. Acad. Dermatol.* **37** (1997) part 2, S72 – S76.
54. M. J. LeVine, H. A. D. White, J. A. Parrish, *J. Invest. Dermatol.* **73** (1979) 170 – 173.
55. G. D. Weinstein, *Ann. Intern. Med.* **86** (1977) 199 – 204.
56. C. W. Ehmann, J. J. Voorhees, *J. Am. Acad. Dermatol.* **6** (1982) 692 – 696.
57. M. A. de Rie, J. D. Bos, *Clin. Dermatol.* **15** (1997) 811 – 821.
58. F. Chamian et al., *Proc. Natl. Acad. Sci. U.S.A.* **102** (2005) 2075 – 2080.
59. J. Lauharanta, *Clin. Dermatol.* **15** (1997) 769 – 780.
60. R. S. Stern et al., *N. Engl. J. Med.* **310** (1984) 1156 – 1161.
61. J. E. Bennett: "Antimicrobial Agents: Antifungal Agents," in L. L. Brunton, J. S. Lazo, K. L. Parker (eds.): *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 11th ed., McGraw-Hill, New York 2006, pp. 1225 – 1241.

62. M. S. Marriott, *J. Gen. Microbiol.* **117** (1980) 253 – 255.
63. H. Van Den Bossche, G. Willemsens, W. Cools, W. F. J. Lauwers et al., *Chem.-Biol. Interact.* **21** (1978) 59 – 78.
64. J. E. Fulton, *Arch. Dermatol.* **111** (1975) 596 – 598.
65. P. H. Spiekermann, M. D. Young, *Arch. Dermatol.* **112** (1976) 350 – 352.
66. O. East, J. T. Henderson, *Dermatologica* **166** (1983) suppl. 1, 1 – 14.
67. R. C. Heel, R. N. Brogden, T. M. Speight et al., *Drugs* **16** (1978) 177 – 201.
68. C. Samson, E. Peets, R. Winter-Sperry, H. Wolkoff: "Clotrimazole/Betamethasone Dipropionate – Lotrisone® – An Antifungal/Steroid Combination for Treatment of Cutaneous Dermatophyte Infections," in H. I. Maibach, C. Surber (eds.): *Topical Corticosteroids*, Karger, Basel 1992, pp. 318 – 334.
69. J. V. Cutsem et al., *J. Am. Acad. Dermatol.* **25** (1991) 257 – 261.
70. M. Brautigam, *J. Am. Acad. Dermatol.* **38** (1998) S53 – S56.
71. L. Drake et al., *J. Am. Acad. Dermatol.* **38** (1998) S87 – S94.
72. A. Polak, P. G. Hartman: "Preclinical Data of Amorolfine," in J. W. Rippon, R. A. Fromtling (eds.): *Cutaneous Antifungal Agents: Selected Compounds in Clinical Practice and Development*, Marcel Dekker, New York 1993, pp. 13 – 26.
73. A. M. Kligman et al., *J. Am. Acad. Dermatol.* **15** (1986) 836 – 859.
74. E. A. Olsen et al., *J. Am. Acad. Dermatol.* **26** (1992) 215 – 224.
75. J. A. Carruthers et al., *J. Am. Acad. Dermatol.* **46** (2002) 840 – 849.
76. S. Kalandakanond, J. A. Coffield, *J. Pharmacol. Exp. Ther.* **296** (2001) 980 – 986.

Laundry Detergents

Eduard Smulders, Henkel KGaA, Düsseldorf, Germany (Chap. 1, 3, 4, 5, 7, 8, 9, 12 and 13)

Wolfgang Rybinski, Henkel KGaA, Düsseldorf, Germany (Chap. 2)

Eric Sung, Henkel KGaA, Düsseldorf, Germany (Chap. 3, 4 and 5)

Wilfried Rähse, Henkel KGaA, Düsseldorf, Germany (Chap. 6)

Josef Steber, Henkel KGaA, Düsseldorf, Germany (Chap. 10)

Frederike Wiebel, Henkel KGaA, Düsseldorf, Germany (Chap. 11)

Anette Nordskog, Henkel KGaA, Düsseldorf, Germany (Chap. 2)

Related Articles → Cleansing Agentsand → Surfactantsare separate keywords

1. Historical Review 3
2. Physical Chemistry of the Washing Process 5
2.1. Introduction to the Washing Process 5
2.2. Components 6
2.2.1. Substrates 6
2.2.2. Washing Liquid/Water 9
2.2.3. Soils 10
2.3. Wetting/Adsorption on Surfaces 10
2.3.1. Wetting 10
2.3.2. Drainage 11
2.3.3. Adsorption Layer 12
2.4. Soil Removal and Stabilization 15
2.4.1. Builders 15
2.4.1.1. Overview on Builder Systems . . 15
2.4.1.2. Precipitation Builders 15
2.4.1.3. Complexing Agents 16
2.4.1.4. Ion Exchangers 17
2.4.1.5. Cobuilders 18
2.4.1.6. Soluble Builders 19
2.4.2. Oily Soil 20
2.4.3. Particulate Soil 25
2.4.4. Chemical Reactions 29
2.4.4.1. Bleachable Stains 29
2.4.4.2. Enzymatic Soil Removal 32
2.4.5. Soil-Repelling Effects 32
2.5. Formulation 34
2.5.1. Foam 34
2.5.2. Microemulsions 36
2.5.3. Phase Stability 36
2.5.4. Softeners 37
2.6. Concluding Remarks 37
3. Detergent Ingredients 37
3.1. Surfactants 37
3.1.1. Anionic Surfactants 39
3.1.2. Nonionic Surfactants 46
3.1.2.1. Alcohol Ethoxylates (AE) 46
3.1.2.2. Alkylphenol Ethoxylates (APE) . 47
3.1.2.3. Fatty Acid Alkanolamides (FAA) 47
3.1.2.4. Alkylamine Oxides 48
3.1.2.5. *N*-Methylglucamides (NMG) . . 48
3.1.2.6. Alkylpolyglycosides (APG) . . . 48
3.1.3. Cationic Surfactants 49
3.1.4. Amphoteric Surfactants 51
3.2. Builders 51
3.2.1. Alkalies 52
3.2.2. Complexing Agents 52
3.2.3. Ion Exchangers 55
3.2.4. Builder Combinations 59
3.3. Bleaches 59
3.3.1. Bleach-Active Compounds 60
3.3.2. Bleach Activators 63
3.3.3. Bleach Catalysts 65
3.3.4. Bleach Stabilizers 65
3.4. Further Detergent Ingredients 66
3.4.1. Enzymes 66
3.4.2. Soil Antiredeposition Agents, Soil Repellent/Soil Release Agents 68
3.4.3. Foam Regulators 69
3.4.4. Corrosion Inhibitors 71
3.4.5. Fluorescent Whitening Agents . 71
3.4.6. Dye Transfer Inhibitors 72
3.4.7. Fragrances 74
3.4.8. Dyes 74
3.4.9. Fillers and Formulation Aids . . 74
4. Laundry Products 75
4.1. Heavy-Duty Detergents 76
4.1.1. Conventional Powder Heavy-Duty Detergents 76
4.1.2. Compact and Supercompact Heavy-Duty Detergents 78
4.1.3. Extruded Heavy Duty Detergents 80
4.1.4. Heavy-Duty Detergent Tablets . 80

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_315.pub2

4.1.5. Color Heavy-Duty Detergents . . 81
4.1.6. Liquid Heavy-Duty Detergents . 82
4.2. Specialty Detergents 83
4.2.1. Powder Specialty Detergents . . 83
4.2.2. Liquid Specialty Detergents . . . 83
4.3. Laundry Aids 84
4.3.1. Pretreatment Aids 84
4.3.2. Boosters 85
4.3.3. Aftertreatment Aids 86
4.3.3.1. Fabric Softeners 86
4.3.3.2. Stiffeners 87
4.3.3.3. Laundry Dryer Aids 88
4.3.4. Other Laundry Aids 89
4.3.4.1. Refreshing Products for Dryer Application 89
4.3.4.2. Odor Removers for Washer Application 89
5. Industrial and Institutional Detergents 89
6. Production of Powder Detergents 90
6.1. Technology Overview 91
6.2. Manufacturing Processes 92
6.2.1. Traditional Spray-Drying Process 92
6.2.2. Superheated Steam Drying 93
6.2.3. Nontower Agglomeration Process 94
6.2.4. Nontower Compound Technology 95
6.3. Densification Processes 95
6.3.1. Dry Densification in a Mixer 95
6.3.2. Dry Densification in a Spheronizer 96
6.3.3. Dry Densification in a Roller Press 96
6.3.4. Wet Granulation 96
6.3.5. Spaghetti Extrusion 97
6.3.6. Postaddition Process 99
6.3.7. Dry Densification in a Tablet Press 99
6.4. Raw Materials 100
6.4.1. Anionic Surfactants 100
6.4.2. Nonionic Surfactants 102
6.4.3. Builders 103
6.4.4. Peroxygen Bleaches 103
6.4.5. Enzymes 104
7. Analysis of the Composition . . 105
7.1. Detergent Ingredients: 106
7.2. Purposes of Detergent Analysis: 106
7.3. Sample Preparation: 106
7.4. Analytical Methods 106
7.4.1. Qualitative Analysis: 106
7.4.2. Sample Preparation 107
7.4.3. Quantitative Analysis 107
7.4.4. Separation Methods: 108
7.4.5. Structure Determination 109
7.4.6. Determination of Characteristic Values 109
7.4.7. Analysis Automation: 109
7.5. Sources of Information: 110
8. Test Methods for Laundry Detergents 110
8.1. Laboratory Methods 110
8.2. Practical Evaluation 111
8.3. Consumer Tests 112
9. Economic Aspects 112
9.1. Detergent Components 113
9.1.1. Surfactants 113
9.1.2. Builders 113
9.2. Laundry Detergents 115
9.3. Fabric Softeners 115
9.4. Other Laundry Aids 117
10. Ecology 117
10.1. Laundry, Wastewater, and the Environment 117
10.2. Contribution of Laundry to the Sewage Load 118
10.3. Detergent Laws 118
10.3.1. Development of the European Detergent Legislation 118
10.3.2. Regulatory Limitations on Anionic and Nonionic Surfactants . 120
10.3.3. Primary Biodegradation Test Procedures 121
10.3.4. Regulation of Maximum Phosphate Content in Detergents . . . 122
10.4. General Criteria for the Ecological Evaluation of Detergent Chemicals 123
10.4.1. Concept 123
10.4.2. Environmental Exposure Assessment 124
10.4.2.1. Biodegradation 124
10.4.2.2. Biodegradability Standard Test Methods 125
10.4.2.3. Supplementary Biodegradation Test Methods 126
10.4.2.4. Exposure Analysis 127
10.4.3. Assessment of Environmental Effects 128
10.4.3.1. Basic Ecotoxicity Tests 128
10.4.3.2. Subchronic and Chronic Ecotoxicity Tests 130
10.4.3.3. Biocenotic Ecotoxicity Tests . . . 130
10.4.3.4. Bioaccumulation 131
10.4.4. Process of Environmental Risk Assessment 131

10.5. Ecological Characterization of Main Detergent Ingredients . . 132
10.5.1. Surfactants 132
10.5.1.1. Anionic Surfactants 132
10.5.1.2. Nonionic Surfactants 135
10.5.1.3. Cationic Surfactants 136
10.5.2. Builders 137
10.5.2.1. Zeolites 137
10.5.2.2. Polycarboxylates 138
10.5.2.3. Citrates 138
10.5.2.4. Sodium Carbonate (Soda Ash) . 139
10.5.2.5. Nitrilotriacetate (NTA) 139
10.5.3. Bleaching Agents 139
10.5.3.1. Sodium Perborate 139
10.5.3.2. Sodium Percarbonate 140
10.5.3.3. Tetraacetylethylenediamine (TAED) 140
10.5.4. Auxiliary Agents 140
10.5.4.1. Phosphonates 140
10.5.4.2. EDTA 141
10.5.4.3. Enzymes 142
10.5.4.4. Optical Brighteners 142
10.5.4.5. Carboxymethyl Cellulose 142
10.5.4.6. Dye Transfer Inhibitors 143
10.5.4.7. Fragrances 143
10.5.4.8. Foam Regulators 143
10.5.4.9. Soil Repellents 143
10.5.4.10. Dyes 144
10.5.4.11. Sodium Sulfate 144
11. Toxicology 144
11.1. Detergent Ingredients 144
11.1.1. Surfactants 144
11.1.2. Builders 146
11.1.3. Bleach-Active Compounds 146
11.1.4. Auxiliary Agents 146
11.2. Finished Detergents 147
11.3. Conclusions 147
12. Textiles 148
13. Washing Machines and Wash Programs (Cycles) 155
13.1. Household Washing Machines 155
13.1.1. Classification 155
13.1.2. Operational Parameters 157
13.1.3. Wash Programs 159
13.1.3.1. Japanese Washing Machines and Washing Conditions 159
13.1.3.2. North American Washing Machines and Washing Conditions . 160
13.1.3.3. European Washing Machines and Washing Conditions 160
13.1.4. Energy and Water Consumption 162
13.1.5. Construction Materials Used in Washing Machines 162
13.1.6. The Market for Washing Machines 163
13.2. Laundry Dryers 164
13.3. Washing Machines for Institutional Use 165
13.3.1. Batch-Type Machines 165
13.3.2. Continuous Batch Washers 166
14. References 166

1. Historical Review

The symbol used by the ancient Egyptians to represent a launderer was a pair of legs immersed in water. This choice was logical, because at that time the standard way to wash clothes was to tread on them. The "fullones" of ancient Rome earned their bread, too, by washing clothes with their feet. The washing process then was very simple: laundry of every kind was subjected to purely mechanical treatment consisting of beating, treading, rubbing, and similar procedures. It has long been known, however, that the washing power of water can be increased in various ways. Rainwater, for example, was found to be more suitable for washing than well water. Hot water also was found to have more washing power than cold, and certain additives seemed to improve any water's effectiveness.

Even the ancient Egyptians used soda ash as a wash additive. This was later supplemented with sodium silicate to make the water softer. These two substances also formed the basis for the first commercial detergent brand to appear on the German market, Henkel's "Bleichsoda", introduced in 1878. Its water-softening effect was a result of the precipitation of calcium and magnesium ions, and it simultaneously eliminated iron salts, which had a tendency to turn laundry yellow. Used along with soap, which had also been known since antiquity, this product prevented the formation of inactive material known as "lime soaps", and the laundry no longer suffered from a buildup of insoluble soap residues.

Soap is the oldest of the surfactants. It was known to the Sumerians by ca. 2500 B.C., although the Gauls were long credited with its discovery. For more than 3000 years, soap was regarded strictly as a cosmetic — in particular

a hair pomade — and as a remedy. Only in the last 1000 years has it come to be used as a general purpose washing and laundering agent. Soap remained a luxury until practical means were discovered for producing soda ash required for saponification of fats. With the beginning of the 20th century and the introduction in Germany of the first self-acting laundry detergent (Persil, 1907), soap took its place as one ingredient in multicomponent systems for the routine washing of textiles. In these, soap was combined with so-called builders, usually sodium carbonate, sodium silicate, and sodium perborate. The new detergents were capable of sparing people the weather-dependent drudgery of bleaching their clothes on the lawn, and the enhanced performance of these new agents substantially reduced the work entailed in doing laundry by hand.

The next important development was the transition brought about by technology from the highly labor-intensive manual way of doing laundry to machine washing. This change in turn led to a need for appropriate changes in the formulation of laundry detergents. Soap, notorious for its sensitivity to water hardness, was gradually replaced by so-called synthetic surfactants with their more favorable characteristics. The term "synthetic surfactants" is erroneous when used to distinguish from soap, which itself is made through synthesis as well. The first practical substitutes for soap were fatty alcohol sulfates, discovered in Germany by Bertsch and coworkers in 1928 [43]. The availability of synthetic alkyl sulfates based on natural fats and oils made possible the first neutral detergent for delicate fabrics: Fewa, introduced in Germany in 1932. This was followed in 1933 on the U.S. market by Dreft, a similarly conceived product. Fatty alcohol sulfates and their derivatives (alkyl ether sulfates, obtained by treating fatty alcohols with ethylene oxide and subsequent sulfation) still retain their importance in many applications, particularly in heavy-duty detergents, specialty detergents, dishwashing agents, cosmetics, and toiletries. The general acceptance worldwide of synthetic surfactants, with their reduced sensitivity to water hardness relative to soap, is a development of the 1940s. Procter & Gamble introduced the synthetic detergent Tide in the USA in 1946. By the 1950s the widespread availability of tetrapropylenebenzenesulfonate (TPS), a product of the petrochemical industry, had largely displaced soap as the key surfactant from the detergent market in the industrialized nations. The only remaining role of soap in the industrialized countries became that of a foam regulator. The favorable economics associated with TPS, along with its desirable properties, caused this branched-chain synthetic surfactant to capture ca. 65 % of the total synthetic surfactant demand in the Western world in the late 1950s.

However, a new criterion soon appeared for surfactants, a criterion that had long been ignored: the biodegradability of the products. As a result of several unusually dry summers in Germany, especially that of 1959, water flow in many streams and rivers was severely restricted. Great masses of stable foam began to build up in the vicinity of weirs, locks, and other constructions in waterways. These were caused by insufficiently biodegraded TPS and nonylphenol ethoxylates, another class of surfactants that had been introduced at that time. The discovery that many surfactants could emerge unchanged even from a modern sewage treatment plant and thus enter surface waters led to adoption of the first German Detergent Law in 1961 [44], whose provisions took effect in 1964 [45]. Manufacturers were subsequently enjoined from marketing any detergents or cleansing agents whose biodegradability fell below 80 % in a test devised by the Detergent Commission. Initially only anionic surfactants were affected, but these were later joined by nonionic surfactants. The German precedent was soon followed by the enactment of similar legislation in such countries as France, Italy, and Japan. In the UK and the USA, the transition to biodegradable surfactants occurred as a result of voluntary agreements between industry and government.

For a long time branched alkylbenzenesulfonates and nonylphenol ethoxylates have been replaced in most countries by the much more rapidly and effectively degradable linear alkylbenzenesulfonates and long-chain alcohol ethoxylates.

Another important step in the development of laundry detergents was replacement of builders such as sodium carbonate by complexing agents. The first complexing agents that were used were of the sodium diphosphate type, but these were replaced after World War II by the more ef-

fective sodium triphosphate. Inorganic ion exchangers such as zeolite A have meanwhile replaced sodium triphosphate in many countries, particularly in those where phosphate legislation has been enacted. Table 1 summarizes the performance characteristics of detergents containing various builders and surfactants [46].

The last decades also saw the introduction of other ingredients for improving detergency performance, and their presence in formulations has been state of the art for modern laundry detergents. Key ingredients among these are the following:

Ion exchangers (zeolites)
Soil antiredeposition agents
Enzymes
Fluorescent whitening agents
Foam regulators
Bleach activators
Soil repellents
Polycarboxylate cobuilders

Table 2 shows highlights of the historical development of detergent ingredients and detergents with concurrent developments in textile fibers and washing machines. Comparison of the changes clarifies how closely the various participants in the washing process are tied to one another. The laundry occupies the center of the stage, and it has consistently kept pace both with detergents and with available washing processes and machines, at least to the extent that the newly developed textiles have been found to have value and market appeal. Thus, close cooperation between all manufacturers involved has become essential for the development of more and more optimized washing processes.

2. Physical Chemistry of the Washing Process

2.1. Introduction to the Washing Process

Washing and cleaning in aqueous wash liquor is a complex process involving the cooperative interaction of numerous physical and chemical influences, and occurs as a result of nonspecific adsorption of surfactants on the various interfaces present [47] and through specific adsorption of chelating agents on certain polar soil components [48]. The nature of the soil as well as the water quality and the nature of the fibers to be cleaned also play a major role in the process. In the broadest sense, washing can be defined as both the removal by water or aqueous surfactant solution of poorly soluble matter and the dissolution of water-soluble impurities from textile surfaces.

The three important components of the washing process are soil, substrate, and water or washing liquid. Different kinds of soils can undergo interactions with one another, and in most cases the soils are mixed soils consisting of several different components. The detergent components are dissolved and transported through the liquid, and interactions occur with all the main components: soil, water, substrate, and detergent components. In this chapter the interactions are explained with respect to the different steps of the soil removal process (Fig. 1).

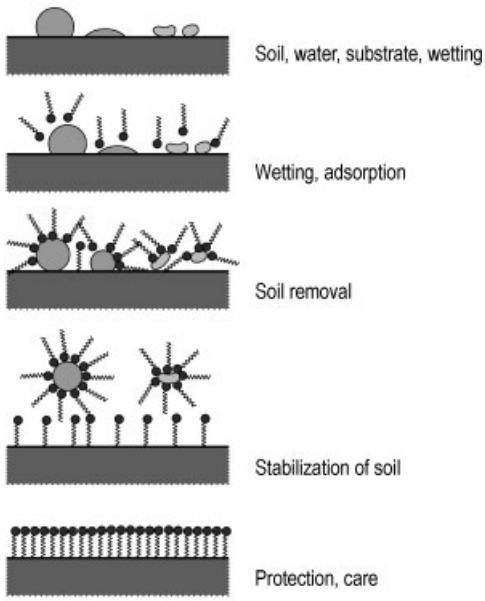

Figure 1. Different steps of the soil removal process

The fabrics are wetted by the soil and the washing liquid. Wetting by the washing liquid is enhanced by the surfactants of the detergent. The surfactants are adsorbed on the soil and on the fabric surface and, via a rolling-up process, they contribute to the removal of fatty and oily soils. Adsorption of surfactants on the surface

Table 1. Properties of detergents based on different surfactants and builders

Detergent	Single wash cycle performance	Soil antiredeposition capability	Deposition on fabrics and washing machines	Yellowing and bad odor
Soap–sodium carbonate–sodium silicate	poor	poor	very heavy	heavy
Synthetic nonionic surfactants–sodium carbonate–sodium silicate	fair to good	fair	heavy	none
Synthetic surfactants–sodium diphosphate	good	fair to good	heavy	none
Synthetic surfactants–sodium triphosphate	good	good	weak	none
Synthetic surfactants–sodium triphosphate–zeolite 4 A	good	good	weak	none
Synthetic surfactants–zeolite 4 A–polycarboxylates–sodium carbonate	good	good	weak	none

of particular soils and fabrics also plays a major role in the removal of these particles. Another important parameter in the soil removal process is the presence of ions and electrically charged molecules. The release of calcium ions from soil deposits and fibers loosens the structure of the residue. Compression by electrolytes of the electrical double layer at interfaces is also significant [47].

When removed from the surface the soil must be stabilized in the washing liquid to prevent redeposition, which in most cases leads to a graying effect of the fibers. Stabilization is obtained through emulsification and dispersion of the soils in the washing liquid. The terms secondary wash performance or multiple wash cycle are used to describe these phenomena, in contrast to the soil removal mentioned above, which is often described in terms of primary wash performance. Other physicochemical processes important for the liquid–liquid interface are the interfacial tension and viscosity, as well as the electric charge and the penetration of the active ingredient.

As a last step in the washing process one can consider adsorption of surface-active substances on the fibers. This may lead to an increased softening effect or to weaker interactions between soil and fabrics for future soiling. Hence the soil will be more easily removed during the next washing cycles. These effects are described in Sections 2.4.5 and 2.5.4.

Another way to organize the physicochemical phenomena of the washing process is related to the different interfaces in the washing system, that is, the air–water, liquid–liquid, solid–liquid, and solid–solid interfaces. Important parameters concerning the air–water surface are the surface tension of the liquid, which should be as low as possible to achieve good wetting and solubility properties. Also foam generation and film elasticity and viscosity are decisive factors. Concerning the solid–liquid interface important parameters are the disjoining pressure, suspension stability, and electric charge, whereby for solid–solid interfaces adhesion, flocculation, heterocoagulation, and sedimentation must be considered.

The interfacial chemical properties undergo changes as detergent compounds are adsorbed, and these changes are prerequisite to effective soil removal. The significance of a given change in respect to the overall process can vary, depending on the system involved. However, in general, the greater the equilibrium adsorption of washing-active substances, and the more favorable their adsorption kinetics, the better their detergency performance [47].

2.2. Components

2.2.1. Substrates

The type of fiber generally has a dramatic influence on the degree of hydrophobicity/hydrophilicity, the wettability, and the extent of soil removal. The active groups of cotton, the most common fiber material in Western Europe, are hydrophilic and have a negative surface charge when placed in water. Because of the negative surface charge, cotton has specific characteristics and can interact with soil in a different way than other synthetic fibers. Cotton also has a large amount of calcium at its surface, which makes it difficult to remove the soil. This is a characteristic difference to synthetic fibers. Through different kinds of surface treatment nature and characteristics of the fiber surface can be changed and hence soil removal improved.

The different effects that complexing agents and surfactants have on a series of fibers are clearly demonstrated in terms of soil removal in

Table 2. Development of washing equipment from 1876 to 2000

Year	Detergent ingredients	Detergents	Textile fibers	Washing equipment
1876	sodium silicate soap starch		cotton linen wool	boiler
1878	sodium carbonate sodium silicate	prewashing product and laundry softener (Henko, Henkel, Germany)		
1890			cupro	
1907	soap sodium carbonate sodium perborate sodium silicate	heavy-duty detergent (Persil, Henkel, Germany)	rayon acetate silk	wooden vat machine
1913	proteases (pancreatic enzymes)	prewashing product (Burnus, Röhm & Haas, Germany)		
1920			viscose staple fiber	metal tub agitator washing machine
1932/33	synthetic surfactants (fatty alcohol sulfates)	specialty detergent (Fewa, Henkel, Germany; Dreft, P & G, USA)		
1933	sodium diphosphate magnesium silicate	heavy-duty detergent (Persil, Henkel, Germany)		
1940	alkylsulfonates (Mersolat) antiredeposition agents (CMC[a])	heavy-duty detergent (Henkel, Unilever, Germany)		automatic agitator washing machine (Blackstone, USA)
1946	fatty alcohol sulfates alkylbenzenesulfonates sodium triphosphate	heavy-duty detergent (Tide, P & G, USA)	polyamide	automatic drum-type machine (Bendix, USA)
1948	nonionic surfactants	heavy-duty detergent (All, Monsanto, USA)		
1949	fluorescent whitening agents (optical brighteners)	rinsing additive (Sil, Henkel, Germany)		
1950	fragrances cationic surfactants	heavy-duty detergent (Dial, Armour Dial, USA) fabric softener (CPC International, USA)		
1954	anionic–nonionic combinations foam-regulators (soap)	 heavy-duty detergent (Dash, P & G, USA)	polyacrylonitrile	semiautomatic drum-type machine (washing automatically, rinsing by hand, spinning separately, Europe)
1957			polyester resin-finished cotton	automatic drum-type machine (washing and rinsing automatically, spinning separately, Europe)
1959	proteases (enzymes)	prewashing detergent (Bio 40, Gebr. Schnyder, Switzerland)		
1962	foam-regulators (behenate soap)	heavy-duty detergent (Dash, P & G, Germany)	polyester–cotton blend	fully automatic drum-type machine (washing, rinsing, and spinning automatically, Europe)
1965			polyurethane wool with reduced felting	wash dryer (washing, rinsing, spinning, and tumble drying automatically in one machine)
1966			nonwovens	
1970	fatty acid amine condensation product	detergent for delicates with fabric softening effect (Perwoll, Henkel, Germany)	resin-finished linen	
1972	bleach activators (TAGU) [b] NTA [c]	heavy-duty detergent (Cid, Henkel, Germany) heavy-duty detergent (various brands, Canada)		

Table 2. Continued

1973	amylases (enzymes)	heavy-duty detergent (Mustang, Henkel, Germany)		
1974	sodium percarbonate	heavy-duty detergent (Dixan, Henkel, Cyprus; Persil, Lever, UK)		
1975	sodium citrate	heavy-duty liquid detergent (Wisk, Lever, USA)		
1976	zeolite 4 A	heavy-duty detergent (prodixan, Henkel, Germany; Tide, P & G, USA)		
	foam-regulators (silicone oils)	heavy-duty detergent (Mustang, Henkel, Germany) heavy-duty liquid detergent without builders (Era, P & G, USA)		
1977	layered silicates–cationic surfactants	dual-function detergent with fabric-softening effect (Bold 3, P & G, USA)		
1978	bleach activator	heavy-duty detergent	Dunova (modified polyacrylonitrile fiber)	microprocessor operation, electronic sensing
	TAED[d]	(Skip, Lever, France)		
1981		liquid heavy-duty detergent in Europe (Liz, Henkel, Germany; Vizir, P & G, Germany		
	esterquats	fabric softeners (Minidou, Lesieur Cotelle, France)		
1982	zeolite 4 A–NTA builder systems	nonphosphate heavy-duty powder detergent, (Dixan, Henkel, Switzerland)	PTFE[e]-fiber Goretex	
1984	poly(acrylic acid), poly(acrylic acid-*co*-maleic acid)	heavy-duty detergents		
1986			PES[f]-polyether fiber Sympatex	controlled water intake (Europe)
1987		compact heavy-duty detergent (Attack, Kao, Japan)		
	cellulase (enzyme)	compact heavy-duty detergent (Attack, Kao, Japan)		
	soil repellent	heavy-duty detergents, USA		
1988	lipase (enzyme)	compact heavy-duty detergent (Hi Top, Lion, Japan)		
1990	alkylpolyglycosides (APG)	liquid heavy-duty detergent (Persil, Henkel, Germany)		
	foam regulators (paraffins)	heavy-duty detergent (Persil, Henkel, Germany)		
1991	dye transfer inhibitor PVP[g]	heavy-duty color detergent (Persil Color, Henkel, Germany)		
1992		extruded supercompact heavy-duty detergent (Persil Megaperls, Henkel, The Netherlands, Belgium, Austria)		
1994	*N*-methylglucamide	heavy-duty detergents (Ariel, P & G, Germany; Tide, P & G, USA)		
	bleach catalyst Mn-TACN[h]	heavy-duty detergent (Omo Power, Lever, The Netherlands; Persil Power, Lever, UK)		
1995				fuzzy logic control
1996		gel-structured liquid heavy-duty detergents (Persil Gel and Dixan Gel, Henkel, Europe)		
1997–1998		heavy-duty detergent tablets (Europe: Colon, Benckiser, Spain; Persil, Lever, UK; Le Chat, Henkel, France; Dixan, Henkel, Italy; Persil, Henkel, Germany)		

[a] CMC = carboxymethyl cellulose.
[b] TAGU = tetraacetylglycoluril.
[c] NTA = nitrilotriacetic acid.
[d] TAED = tetraacetylethylenediamine
[e] PTFE = polytetrafluoroethylene
[f] PES = poly(ethylene glycol terephthalate)
[g] PVP = poly(*N*-vinylpyrrolidone)
[h] Mn–TACN = manganese–tetraazacylononane complex

Figure 2. The anionic surfactant selected in this case (alkylbenzenesulfonate) has virtually no effect on cotton, whereas even by itself sodium triphosphate, as an example of a complexing agent, brings about substantially improved soil removal. Permanent press cotton also shows a greater effect for sodium triphosphate than for surfactants alone. The effect of sodium triphosphate is particularly pronounced with the relatively hydrophilic synthetic fibers polyamide and polyacrylonitrile. This behavior changes when more hydrophobic textile fibers are used. An effect is still apparent with polyester/cotton and polyester, but it is not greater than that of surfactants alone. At the same time, overall wash performance decreases. In the case of the very hydrophobic polyolefin fibers, the wash effectiveness of the surfactant is substantially greater than that of the complexing agent. These examples show the synergistic way in which surfactants and complexing agents or ion exchangers complement each other, not only in the case of mixed soils, but also with regard to removing soil from different fibers.

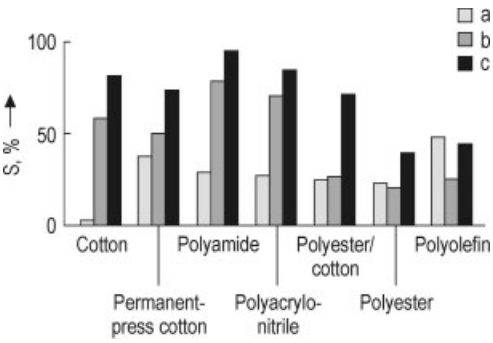

Figure 2. Influence of textile fibers on soil removal *S* [46] Detergents: a) 1 g/L Alkylbenzenesulfonate + 2 g/L sodium sulfate; b) 2 g/L sodium triphosphate; c) 1 g/L alkylbenzenesulfonate + 2 g/L sodium triphosphate

2.2.2. Washing Liquid/Water

The most obvious role of water is to serve as a solvent, both for the detergent and for soluble salts within the soil. Water is also the transport medium for dispersed and colloidal soil components, however. The washing process begins with wetting and penetration of the soiled laundry by the detergent solution.

The degree of hardness is one of the most important parameters of the water itself. Water hardness is defined in terms of amount of calcium and magnesium salts solved in the water. In different regions and countries the water hardness can vary. In most European regions the water is hard. In North America, Brazil, and Japan, the water is comparatively soft. The ions in the washing water can have different effects on the washing process. For example ion exchange with the counterions of the soaps and surfactants of the detergent can take place, and insoluble calcium or magnesium compounds are formed instead of soluble sodium compounds. This binds the surfactant and hinders it in contributing to the soil removal process. The precipitates thus formed may also have an impact on the washing machine and the fabrics to be washed, because they form residues which are difficult to remove. Moreover, high calcium content in the washing water may have an influence on the removal of particulate soil.

The presence of trace amounts of iron, copper, or manganese ions in water may also be detrimental to washing. These ions can catalyze decomposition of bleaching agents during the washing process. Complexing agents or ion exchangers are often found in detergents, and one of their functions is to bind multivalent alkaline earth and heavy-metal ions through chelation or ion exchange.

Table 3. Surface tension of different solvents

Solvent	Surface tension, mN/m
Water	72.5
Hexane	19.5
Decane	23.9
Benzene	28.4
Toluene	29.3
Chlorobenzene	33.2
Tetrachloromethane	25.7
Cyclohexane	24,7

One of the important functions of water, besides being a transport medium for the active substances of the detergent and for the soil after removal, is to wet and penetrate into the fabrics. Water has a very high surface tension (72 mN/m). To obtain a more effective wetting behavior the surface tension of water must be reduced to the usual surface tension level of organic solvents. As a comparison the surface tensions for water and some organic solvents is shown in Table 3 [49]. The reduction of sur-

face tension can be achieved through the surface-active properties of the surfactants, which effectively reduce the surface tension of water to values of 30 mN/m and below. The surfactants thus become key components of any detergent. This will be further explained in Section 2.4.2.

Other important parameters which may influence the washing process are mechanical input, washing time, and temperature.

2.2.3. Soils

One way to categorize soils is by their origin, i.e., residues from domestic activities like cooking or gardening, impurities from industrial or commercial activities, bodily excretions, and dust from the atmosphere.

From a detergency standpoint, however, it is more appropriate to regard the principal types of soil in terms of their chemical properties, for example, water-soluble materials such as sugar, inorganic salts, urea, and perspiration, and solid particles such as rust/metal oxides, soot (carbon black), carbonates, silicates, humus etc. Another big group of substances are fats and oils such as animal fat, vegetable fat and oil, sebum, mineral oil, and grease. These are very common soils resulting from domestic and industrial activities and can in principle be removed with surfactants, which are one of the main components of detergents. Proteins such as blood, grass, egg, milk, and keratin from skin are another group of soils. These require enzymes to cleave them into smaller parts before they can be removed by the surfactants. Another group of stains can only be removed by chemical reactions. These are typically wine, coffee, tea, fruit juices, and vegetable stains.

Often, these stains are not in pure form. Most stains are mixtures of two or more components, for example, a pigment and a fat or oil. Moreover, interactions between the soils can make them even more difficult to remove. Hence, removing these soils is the combined work of several detergent components.

The removal of common stains like solid particles and fat or oil takes place in pure interfacial processes through the action of surfactants, that is, without any chemical change of the soils, whereas other types of stains can only be removed by chemical changes of the colored molecules. One example of the latter is the removal of bleachable stains, like the above-mentioned wine, tea, and coffee, which in most cases is an oxidation reaction leading to changes in the properties of the colored substance that turn it into a colorless substance. In some cases the colorless residue stays on the fabric, and sometimes the chemical change may result in easier removal from the fabric. The same is true for proteins. Through the action of enzymes they go through a chemical change and are cleaved into smaller molecules which can then be removed more easily. Among the most difficult soils to remove from fabrics are pigments, such as carbon black, inorganic oxides, carbonates, and silicates. Other problematic soils include fats, waxes, higher hydrocarbons, denatured protein, and certain natural dyes. Mostly these stains are present on the fibers in the form of mixed soils.

Apart from surfactants, enzymes, and bleach, other components taking part in the soil removal process are water-soluble complexing agents and water-insoluble ion exchangers. In the following chapters, the physicochemical effects of these different soil removal processes will be explained.

2.3. Wetting/Adsorption on Surfaces

2.3.1. Wetting

In the first phase of washing, textile fibers and soil must be wetted as thoroughly as possible by the wash liquor [50]. Wetting is a complex process which is determined by the interaction of the different interfacial tensions between the solid surface, the liquid and the gas phase. A contact angle θ between the solid and a drop of a liquid applied to its surface is formed and this can be taken as a measure of wetting (Fig. 3). The contact angle depends on the structure of the surface and the surface tension of the liquid.

To a first approximation, wetting can be described by the Young equation [47, 51]:

$$\gamma_S = \gamma_{SL} + \gamma_L \cos\theta \quad (1)$$

γ_S = interfacial tension substrate/air (mN/m)
γ_{SL} = interfacial tension substrate/liquid (mN/m)

γ_L = interfacial tension liquid/air (mN/m)
θ = contact angle of the wetting liquid

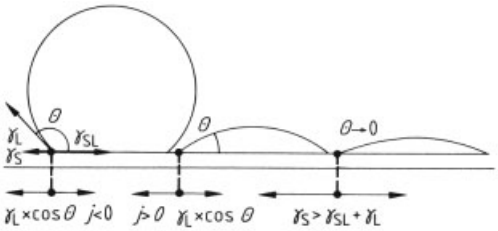

Figure 3. Schematic representation of wetting of a solid surface

As the interfacial tension liquid–air and the contact angle are easily accessible by experiments, sometimes the so-called wetting tension is defined:

$$j = \gamma_S - \gamma_{SL} = \gamma_L \cos\theta \quad (2)$$

Total wetting of a solid is possible only if the liquid drop spreads spontaneously over the solid surface, e.g., when $\theta = 0$ and $\cos\theta = 1$. Strictly speaking, the limiting case described above, that of total wetting (Eq. 2), is only applicable if γ_{SL} is so reduced through adsorption that it approaches zero. With a given solid surface possessing a low surface energy, for various liquids a linear relationship normally exists between $\cos\theta$ and surface tension. The limiting value for $\cos\theta = 1$ is a constant of the solid and is referred to as its critical surface tension γ_c. This means that only those liquids with surface tension equal to or less than the critical surface tension of a given solid spread spontaneously, thus causing thorough wetting. Table 4 collects critical surface tension data for various synthetic materials.

Table 4. Critical surface tensions of several typical plastics [52]

Polymer	γ_c at 20 °C, mN/m
Polytetrafluoroethylene	18
Polytrifluoroethylene	22
Poly(vinyl fluoride)	28
Polyethylene	31
Polystyrene	33
Poly(vinyl alcohol)	37
Poly(vinyl chloride)	39
Poly(ethylene terephthalate)	43
Poly(hexamethylene adipamide)	46

Polyamide, for example, which has a γ_c of ca. 46 mN/m, is relatively easy to wet with standard commercial surfactants, whereas polytetrafluoroethylene, which has a γ_c of about 18 mN/m, is wettable only with special fluoro surfactants. This relationship simplifies the choice of a proper surfactant for a given wetting problem. More recently, special attention has been given to ultrahydrophobic surfaces, which in addition to a very high contact angle have surface structures which diminish further the contact area between liquid and solid surface. Surfaces with such properties are known to have self-cleaning properties, although applications for washing are not known until now.

To better understand and predict the mechanism of droplet spreading on surfaces, new modeling methods have been developed. The spreading of an aqueous drop on a hydrophobic surface and the adsorption of surfactants from the drop on the air–water and water–hydrophobic solid interfaces was studied by using molecular dynamics. An example of a simulated time sequence for a drop with a linear, flexible-chain surfactant is shown in Figure 4.

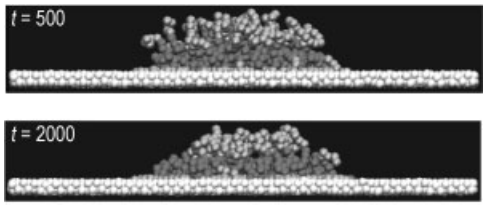

Figure 4. Time sequence of a drop (solvent not shown) with a linear, flexible-chain surfactant adsorbed on the interfaces of the spreading drop [53]

2.3.2. Drainage

Drainage is closely connected with wetting phenomena. Drainage is of specific importance in the rinsing step, especially for hard surfaces. The washing or cleaning liquid should completely wet the surface and rinse off in a uniform and thin film. Only few studies are related to the physicochemical effects of drainage from solid surfaces in presence of surfactants [54]. Figure 5 shows a set of drainage curves for surfactant solutions, obtained with a gravimetric method.

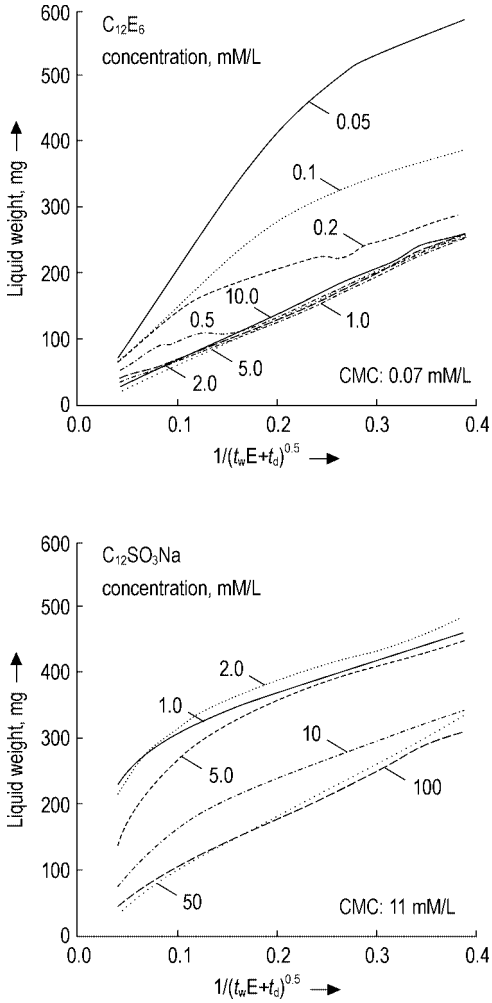

Figure 5. Weight of liquid film draining from a cylinder as a function of $1/(t_w + t_d)^{0.5}$ for different concentrations of a nonionic surfactant $C_{12}E_6$ and an anionic surfactant $C_{12}SO_3Na$ [54]

The drainage curve of a pure liquid is known be a linear function of $1/(t_w + t_d)^{0.5}$. With addition of surfactant the amounts of liquid remaining on the surface increase, reach a maximum, and then decrease, so that at high concentrations the drainage behavior of a pure liquid is reached again. The Gibbs–Marangoni effect, which retards the liquid flow in the interface, is the reason for this behavior. At high concentration this effect is compensated by diffusion of the surfactants. Therefore, it is important to have a high enough surfactant concentration to guarantee both complete wetting and small liquid amounts during drainage.

2.3.3. Adsorption Layer

The adsorption of components, especially surface-active substances, at interfaces in the washing process is an important mechanism which was already shown for wetting and drainage. This regards the gas–liquid, liquid–liquid, and liquid–solid interfaces. The adsorption at the gas–liquid interface Γ is related to the surface tension γ by the Gibbs equation 3

$$\Gamma = -\frac{1}{RT}\frac{\mathrm{d}\gamma}{\mathrm{d}\ln c} \tag{3}$$

Therefore surface tension–concentration curves are a useful tool to characterize this interface. Figure 6 shows the surface tension as a function of concentration for surfactants with the same alkyl chain length and different hydrophilic groups. All curves have a breakpoint which is characteristic for the critical micelle concentration (CMC) of surfactants, with the nonionic surfactant having the lowest CMC and surface tension above the CMC, whereas anionic and cationic surfactants are quite similar.

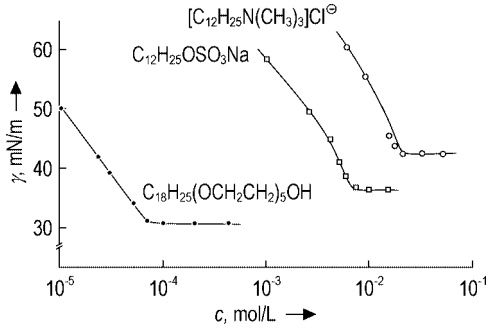

Figure 6. Surface tension as a function of concentration for different surfactants [55]

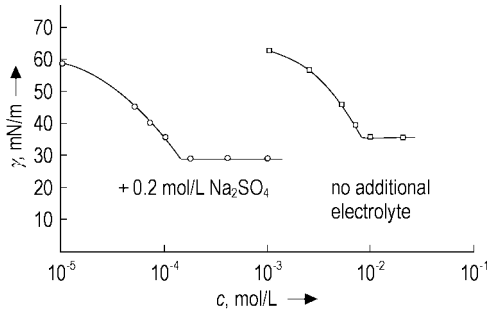

Figure 7. Influence of counterions on the surface activity of a typical anionic surfactant as a function of surfactant concentration [55]

The surface tension is dependent on alkyl chain length and electrolyte concentration (Figs. 7 and 8). Electrolytes strongly decrease the CMC of the surfactants. The same effect is observed for increasing chain length of the surfactant.

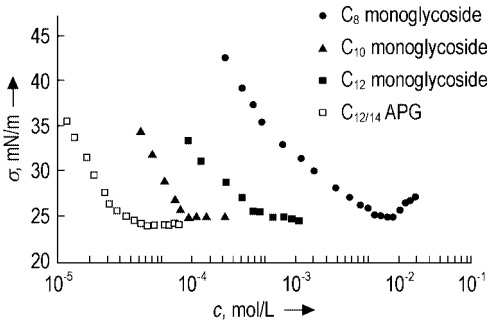

Figure 8. Surface tension of alkyl monoglycosides and alkyl polyglycosides (AGP®) [56]

The kinetics of adsorption of surfactants at the liquid–gas interface is an important aspect for applications like detergency, as in many processes the state of equilibrium is not reached. Figure 9 shows the dynamic surface tension of an anionic and a nonionic surfactant as a function of time for different concentrations. The two surfactants strongly differ. Whereas the nonionic surfactant shows strong time dependence even for high concentrations, the surface tension of the anionic surfactant only slightly changes above the CMC.

The adsorption effects at the liquid/liquid interface which are important for the removal of oily soils are quite similar to adsorption at the gas–liquid interface [58]. Table 5 lists interfacial tensions of pure oils in water [59]. The interfacial tension decreases with increasing surfactant concentration until a breakpoint is reached. This breakpoint corresponds to the CMC for the nonpolar alkane, whereas it is shifted to higher concentrations for the more polar oils. This is due to the higher solubility of the surfactant in these oils.

Table 5. Interfacial tension between water and organic solvents

System	Interfacial tension, mN/m
Water/decane	48.7
Water/isopropyl myristate	27.9
Water/2-octyldodecanol	24.2

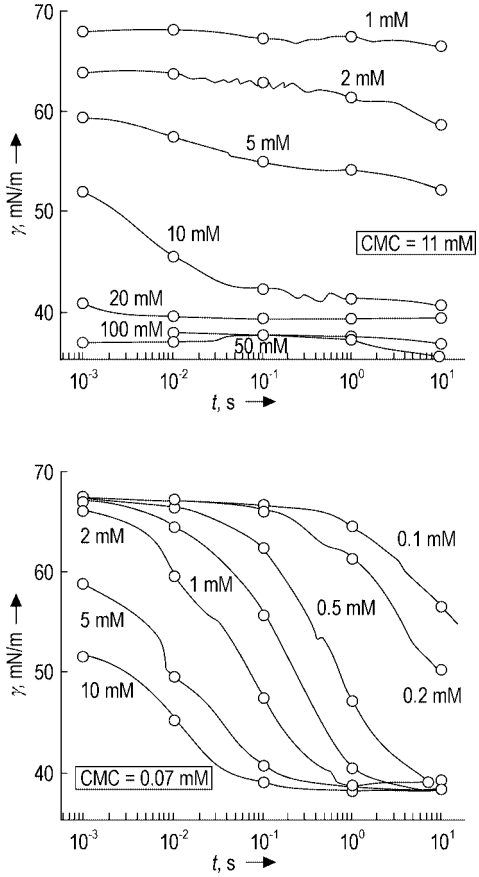

Figure 9. Dynamic surface tension of $C_{12}SO_3Na$ (top) and $C_{12}E_6$ (bottom) as a function of concentration at 40°C [57]

Figure 10 shows the interfacial tension of different surfactants for three oils as a function of concentration. The adsorption of surfactants and ions both on the fiber and soil surface is a prerequisite for the removal of pigment soil from solid surfaces. Figure 11 shows the adsorption of a nonionic and an anionic surfactant on a typical hydrophobic pigment. In contrast to anionic and cationic surfactants, nonionic surfactants have no or little effect on surface charge [60]. Therefore, their mode of action cannot be ascribed to changes in electrical charge on pigments and fibers. Instead, the effect is related exclusively to properties of the adsorption layer. Compared to anionic surfactants, the adsorption of nonionic surfactants is quite strong at hydrophobic, weakly polar interfaces. This is illustrated in Figure 11 by the relative adsorption of dode-

cyl dodecaglycol ether and *n*-dodecyl sulfate on activated carbon black [61].

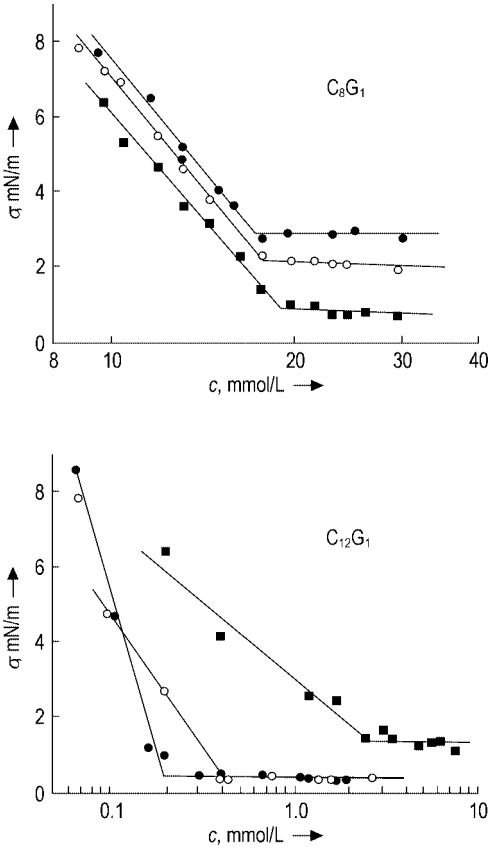

Figure 10. Effect of alkylglycoside surfactants C_nG_1 on the oil/water interfacial tension σ for three different oils: ● n-decane; ○ isopropyl myristate; ■ 2-octyldodecanol as a function of concentration c_1 at 60 °C. a) C_8G_1; b) $C_{12}G_1$ [59]

Figure 12 shows the adsorption of nonionic surfactants of the glycoside type on a hydrophilic substrate as a function of the concentration relative to the CMC [58]. The adsorbed amounts are smaller than on a hydrophobic surface by a factor of about 100. Surface coverage of a close-packed monolayer is not reached in the studied concentration range, which indicates the adsorption of isolated individual molecules.

For the soil removal of pigments and the repulsion between removed soil and the surface of the substrate the adsorption on fibers has a strong influence. As an example for this effect the adsorption of nonionic surfactants on polyester fibers is shown in Figure 13 [62]. The adsorbed amounts reach a plateau value at about the CMC and are lower for the nonionic surfactant with the longer hydrophilic group at low surfactant concentration. Considering the adsorption effects on the soils and different fibers it is important to choose the right surfactant structure for optimum adsorption effects and therefore optimum soil removal.

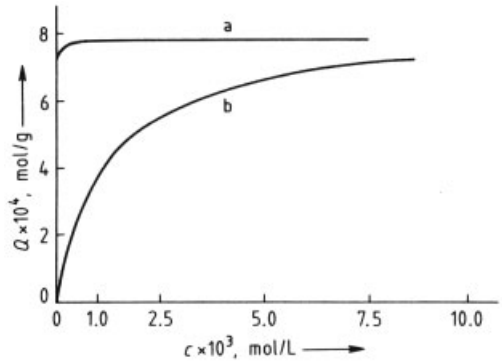

Figure 11. Adsorption isotherms for surfactants on activated carbon black [61]
25 °C, surface area 1150 m^2/g (from BET measurements)

1) $C_{12}H_{25}O\,(CH_2CH_2O)_{12}H$;
2) $C_{12}H_{25}OSO_3Na$

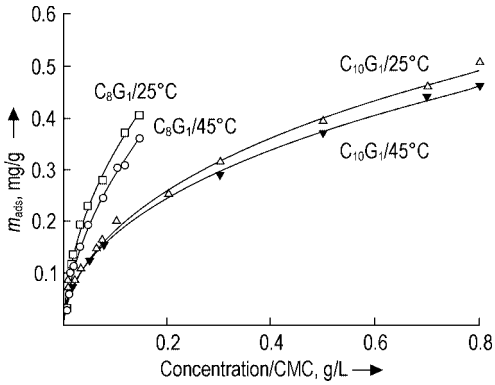

Figure 12. Adsorbed amounts m_{ads} of C_nG_1 on hydrophilic glass CPG-10 as a function of concentration relative to the CMC [58]

The adsorption of inorganic ions on fibers and soil is an effect which is often underestimated in soil removal. These ions can come from the soil, from the water (hardness ions), or from the detergent. The adsorption of inorganic ions can modify the surface charge of the solids and compete

with or enhance the adsorption of surfactants. An example of the adsorption of inorganic ions is given in Figure 14 [63]. The adsorption isotherm of Ca^{2+} ions on SiO_2 shows strong adsorption with an increase up to high concentrations. By this adsorption the negative surface charge of SiO_2 can be decreased and even reversed, so that anionic surfactants can be adsorbed on this surface [63].

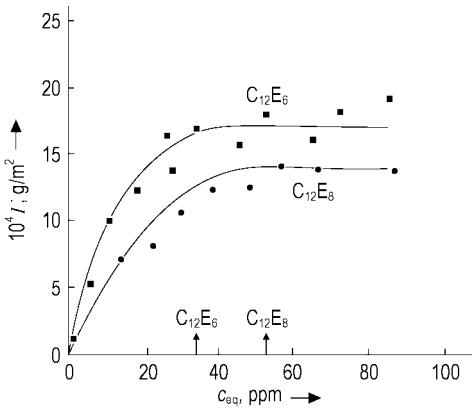

Figure 13. Surface excess Γ of $C_{12}E_6$ and $C_{12}E_8$ on polyester fibers as a function of the equilibrium concentration at 25 °C [62]

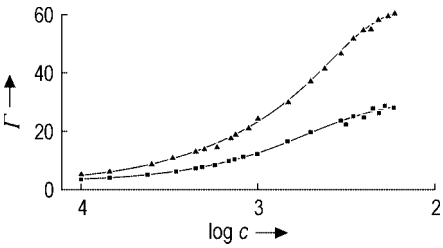

Figure 14. Adsorbed amounts $\Gamma/10^{-8}$ mol g^{-1} of Ca^{2+} (▲) and Fe^{3+} (■) ions on SiO_2 as a function of log (c/mol L^{-1}) [63]

2.4. Soil Removal and Stabilization

2.4.1. Builders

2.4.1.1. Overview on Builder Systems

Besides surfactants builders play an important role in the washing process [61]. They have mainly two important functions: supporting soil removal and water softening by removing the calcium and magnesium ions resulting from water and soil. Salts of multivalent cations are almost always present in soils and on textile fiber surfaces. Examples are calcium carbonate, calcium phosphate, and calcium stearate. Cationic bridges, which are responsible for binding soil components chemically to fibers, also frequently form. This type of linkage can be due to the carboxyl groups, for example, commonly found in cotton as a result of oxidation, but such a linkage may also arise from the presence of reactive centers associated with metal oxides or from soaps derived from sebum. Problems stem principally from poorly soluble calcium salts, whose solubility is further diminished as the water hardness of the wash liquor increases. On the other hand, their solubility in distilled water is higher because of displacement of the solubility equilibrium. When calcium salts are dissolved from a multilayered soil deposit, cavities remain in the structure. These cavities loosen the deposit and facilitate its removal from the surface. Thus, one task of detergent components is to create the highest possible calcium ion concentration gradient between soil and aqueous phase during the washing process.

The substances used as builders can be very different: precipitating alkaline materials such as sodium carbonate, sodium silicate, or soaps; complexing agents like sodium triphosphate or nitrilotriacetic acid (NTA); ion exchangers such as zeolites; and polycarboxylic acids.

The choice of builder strongly depends on the type of detergent, washing conditions, and the region where the detergents are used. The physical chemistry of the builder substances is quite different and will be briefly summarized in the following.

2.4.1.2. Precipitation Builders

The performance of precipitation builders is mainly determined by the solubility product of their calcium and magnesium salts and a low affinity of these salts to the fiber. The latter effect is especially important, as the insoluble salts should not adhere to the fiber to cause residues on the surface. This is mainly influenced by additives (see Section 2.4.1.5). Sodium carbonate is the most often used precipitation builder. Calcium carbonate has a very low solubility and

therefore removes Ca^{2+} ions to low residual concentrations. Fatty acids are another possibility for the removal of calcium ions, as these substances strongly react with Ca^{2+} ions in the alkaline pH range. The solubility product is shown in Table 6 for fatty acids of different chain lengths [64]. The solubility of calcium salts of long-chain fatty acids is very low.

Table 6. Solubility products of calcium carboxylates at 23 °C [64]

Carbon no.	Solubility product
8	2.7×10^{-7}
9	8.0×10^{-9}
10	3.8×10^{-10}
11	2.2×10^{-11}
12	8.0×10^{-13}
14	1.0×10^{-15}
16	1.6×10^{-16}
18	4.0×10^{-18}

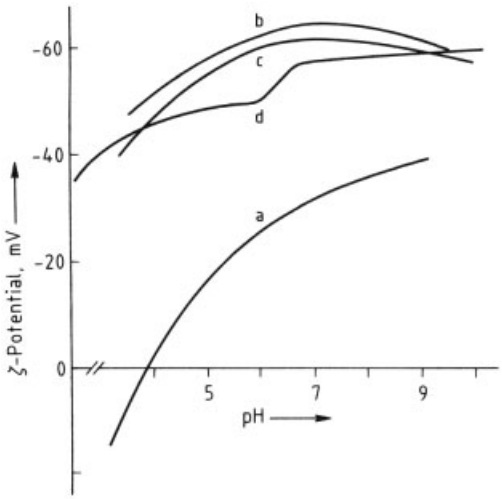

Figure 15. ζ Potential of hematite as a function of pH at 25 °C in the presence of sodium chloride (a), sodium triphosphate (b), benzenehexacarboxylic acid (c), and 1-hydroxyethane-1,1-diphosphonic acid (d) [48]
Anion concentration: 2.5×10^{-3} mol/L

2.4.1.3. Complexing Agents

Whereas surfactants are adsorbed nonspecifically at all hydrophobic surfaces, complexing agents can undergo specific attraction to surfaces that have distinct localized charges. The main process is chemisorption and is especially characteristic of metal oxides and certain fibers [48]. As shown in Figure 15, the adsorption of a complexing agent produces an effect similar to that of an anionic surfactant. The change in ζ potential for hematite is taken as illustrative.

The adsorption of the complexing agents has a significant impact on the dispersion properties. This can be shown for the sedimentation of graphitized carbon black and kaolinite in solutions of sodium triphosphate (Fig. 16) [61], but can also be observed for polar pigments such as metal oxides [65]. Whereas the settling volume V_S of graphitized carbon black hardly changes with increasing concentration of the complexing agent, kaolinite is dispersed much better at high concentration of the complexing agent, as is indicated by the lower settling volume.

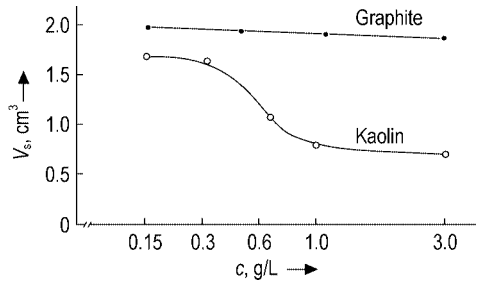

Figure 16. Settling volume of graphitized carbon black and kaolinite in sodium triphosphate solutions at 16°dH water hardness, 0.30 g/10 mL graphitized carbon black, 0.50 g/10 mL kaolinite [61]

The specificity of adsorption of complexing agents with respect to metal oxides is so great that even displacement of anionic surfactants from surfaces with lower adsorption energies is possible [48].

Complexing agents suppress the adsorption of anionic surfactants on metal oxides. However, adsorption is enhanced on materials such as carbon black or synthetic fibers. This effect is due to the electrolyte character of the complexing agent. The washing process generally involves removing mixed soils that consist of both hydrophilic and hydrophobic matter from fiber surfaces. For this reason, the different specificities of complexing agents and surfactants give complementary functions to these two types of material.

Figure 17 illustrates how surfactants, salts, and complexing agents complement each other in the removal of soil. The negative influence of calcium is eliminated by sodium triphosphate or

other complexing agents. The magnitude of the indirect counterion effect of sodium ions is made clear when sodium sulfate is added to alkylbenzenesulfonate (curve b).

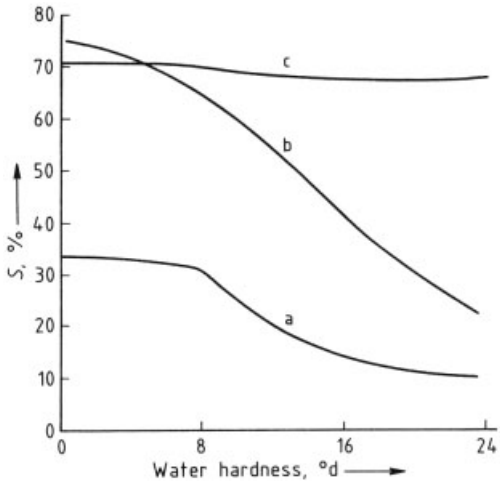

Figure 17. Soil removal *S* from wool as a function of water hardness at 30 °C [48]
a) Sodium alkylbenzenesulfonate (0.5 g/L); b) Sodium alkylbenzenesulfonate (0.5 g/L) together with sodium sulfate (1.5 g/L); c) Sodium alkylbenzenesulfonate (0.5 g/L) together with sodium triphosphate (1.5 g/L)

Before slightly soluble cations can be dissolved from soil and fibers, adsorption of the complexing agent takes place on the surface, particularly in those areas that contain multivalent cations. In the course of subsequent desorption of water-soluble multivalent cation complexes, many of the soil–fiber bonds are broken, leading to a marked enhancement of the washing effect. Removal of cations from soil and fibers by adsorption–desorption processes and displacement of solubility equilibria are the most important phenomena that accompany the use of complexing agents and ion exchangers in the washing process.

2.4.1.4. Ion Exchangers

Zeolites (→ Zeolites) are used to replace phosphate in many detergents to prevent eutrophication of stagnant and slowly flowing surface waters. The main type of zeolite used in detergents is zeolite A. These water-insoluble, finely dispersed ion exchangers differ in their properties from water-soluble complexing agents.

The main properties of zeolites in the washing process besides ion exchange of the calcium and magnesium ions of the water hardness are:

- Adsorption of water-soluble substances such as dyes on the zeolite particles
- Heterocoagulation of pigments and solid fats with zeolite
- Action as crystallization nuclei for sparingly soluble salts

The most characteristic feature of zeolites is the ion exchange of the sodium ions in the crystal structure by calcium and magnesium ions. Figure 18 shows the ion-exchange kinetics of zeolites A and X for calcium and magnesium ions [66]. Calcium ions diffuse at a high rate into both types of zeolite with a slight preference for the wider pore zeolite X. These differences are only evident for short times which are of not practical importance for the washing process. The ion-exchange kinetics are more strongly dependent on the pore size of the ion exchanger for magnesium ions. Despite the smaller ion radius at 25 °C the magnesium ion has a more stable and larger hydration shell than the calcium ion and therefore more slowly penetrates into the pore system of the zeolite.

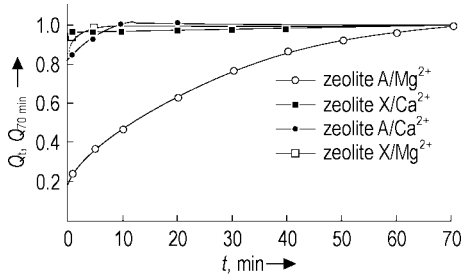

Figure 18. Kinetics of ion exchange of calcium and magnesium ions for zeolite A and zeolite X (25 °C, ion concentration 5.36×10^{-3} mol/L, zeolite concentration 1 g/L) [66]

A comparison between the decrease of water hardness by ion exchange and washing performance is given in Figure 19 [67]. A decrease in water hardness from 16°dH down to 3–4°dH only slightly influences the detergency. Only a further decrease of the calcium ion concentration leads to a significant increase of soil removal from the fabric. Because zeolite A is an ion

exchanger, calcium ion exchange is decreased by a high concentration of sodium ions despite the high selectivity of the ion-exchange process. Thus detergency in the presence of sodium ions slightly decreases.

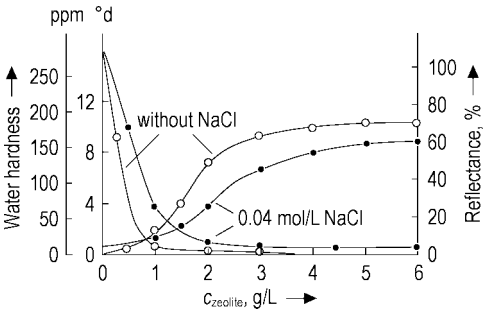

Figure 19. Influence of NaCl on the water-softening effect and the washing performance of zeolite A. Water-softening effect at 90 °C after 15 min and washing performance at 90 °C and 285 ppm water hardness on cotton soiled with particulate sebum [67]

The ion exchange of the zeolite can be described by Equation 4.

$$Q_{Ca^{2+}} = \frac{Q_m c_{Ca^{2+}}}{c_{Ca^{2+}} + 2\frac{b_2}{b_1}\left(c'_{Na^+} + 2\,Q_{Ca^{2+}}\right)} \quad (4)$$

$Q_{Ca^{2+}}$ = exchanged amount of calcium ions
Q_m = maximum exchanged amount of calcium ions
$c_{Ca^{2+}}$ = equilibrium concentration of calcium ions
c'_{Na^+} = initial concentration of sodium ions
b_1 and b_2 = constants

Zeolites show significant adsorption properties in the washing process and under the conditions of the wastewater [68]. Figure 20 demonstrates the adsorption of a cationic dye (methylene blue) and an anionic dye (benzopurpurin) onto zeolite A [66]. The cationic dye is strongly adsorbed on the negatively charged surface of zeolite A, whereas the anionic dye is only adsorbed on zeolite A which was extracted from a detergent formulation produced on a technical scale. This is due to hydrophobization of the zeolite surface in the production process, which increases the interaction of the dye and the zeolite surface.

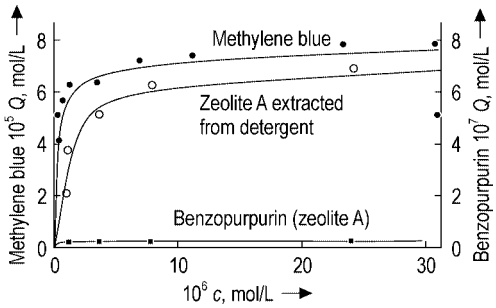

Figure 20. Adsorption of dyes on zeolite A at 23 °C [66]

The concentration of complexing agent and the temperature are generally the decisive factors in removing multivalent metal ions by a water-soluble complexing agent; the binding ability diminishes with increasing temperature. Essentially the same relationships also apply when using a water-insoluble ion exchanger, although the temperature effect is usually reversed [69] (Tab. 7).

Table 7. Binding capacity for calcium and magnesium ions of zeolites P and A [69]

	Zeolite P	Zeolite A
Ca binding capacity at 25 °C	140–165 mg CaO/g	160–175 mg CaO/g
Ca binding capacity at 60 °C	180 mg CaO/g	180 mg CaO/g
Mg binding capacity at 25 °C	20–40 mg MgO/g	20–40 mg MgO/g

2.4.1.5. Cobuilders

The mechanisms of action for complexing agents and water-insoluble ion exchangers are different. The two complement each other in their respective roles. Figure 21 shows the way in which a small amount of a water-soluble complexing agent can increase the washing effectiveness of the water-insoluble ion exchanger zeolite A. The effect results from an increase in the rate of dissolution of divalent ions from soil and fibers. The mode of action is depicted schematically in Figure 22. The water-soluble complexing agent serves as a carrier that transports calcium from the precipitate into the water-insoluble ion exchanger. The process is based on successive adsorption, desorption, and dissociation, and it accelerates the delivery of free calcium ions into solution.

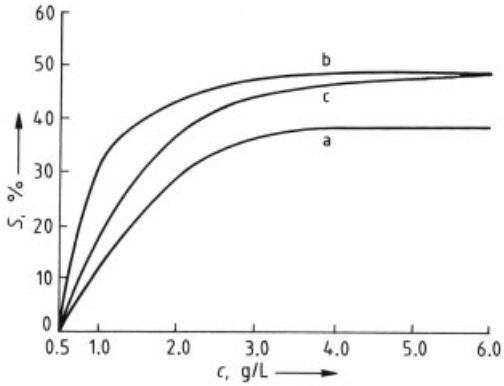

Figure 21. Comparison of soil removal S of zeolite A (a), sodium triphosphate (b), and a mixture of the two builders in the ratio zeolite A: sodium triphosphate = 9 : 1 (c); results obtained for non-resin-finished cotton tested in a Launderometer [70]
Wash time: 30 min with heating; temperature: 90 °C; water hardness: 16 °d (285 ppm)

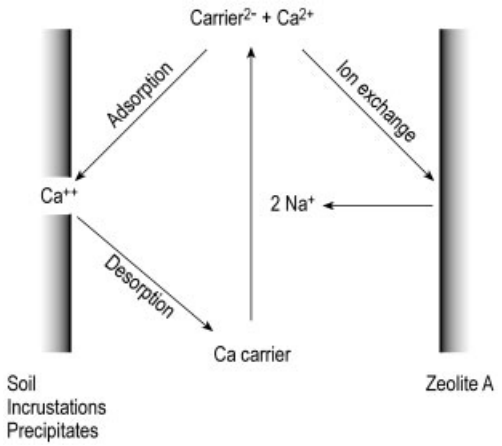

Figure 22. Mechanistic scheme for the carrier effect [71]

Polycarboxylates are widely used in detergents as cobuilders [72 – 75]. They are able to retard precipitation of sparingly soluble calcium salts such as calcium carbonate and calcium phosphate (*threshold effect*) when used in small concentrations. As anionic polyelectrolytes they bind cations (counterion condensation), whereby multivalent cations are strongly preferred. Polycarboxylates can also disperse solids in aqueous solution. Both dispersion and threshold effect result from adsorption of the polymer on the surface of soil and $CaCO_3$ particles, respectively. Stabilization of sparingly soluble salts such as $CaCO_3$ in a colloidal state by polycarboxylates occurs at substoichiometric concentrations of the cobuilder in the washing liquor, which is an advantage compared to ion exchange or complexation. Thus, small amounts of threshold-active compounds can be used as cobuilders even in laundry detergents with a high content of sodium carbonate. The effect, however, is strongly dependent on the washing conditions, such as temperature, soda ash, and cobuilder concentration. Figure 23 illustrates the range of effectiveness (grey areas) of a polycarboxylate (AC) in a carbonate-containing system for typical European conditions of water hardness (3.04×10^{-3} mol/L Ca^{2+}) [76]. The appearance of $CaCO_3$ particles larger than ca. 0.2 μm within 30 min was taken as an indicator of the threshold effect. The results show that polycarboxylates are no longer threshold-active above 40 °C. This holds even more for higher carbonate concentrations, i.e., with detergents containing soda ash as the sole builder.

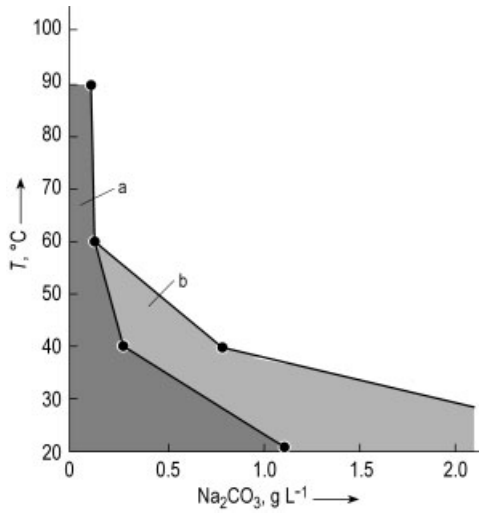

Figure 23. Precipitation inhibition of $CaCO_3$ by AC (threshold effect) as a function of temperature and soda ash concentration (3.04×10^{-3} mol/L Ca^{2+})
a) 105 mg/L AC; b) 210 mg/L AC

2.4.1.6. Soluble Builders

Different types of silicates are described as alternatives for zeolites as builders in detergents [69]. The common property of these silicates is their

solubility. They are either crystalline layered silicates, for example, δ-sodium silicate, or certain amorphous silicates. The sodium ions of these silicates are accessible and can be exchanged for calcium ions. This is shown in the isotherms of Figure 24 for δ-sodium silicate [77].

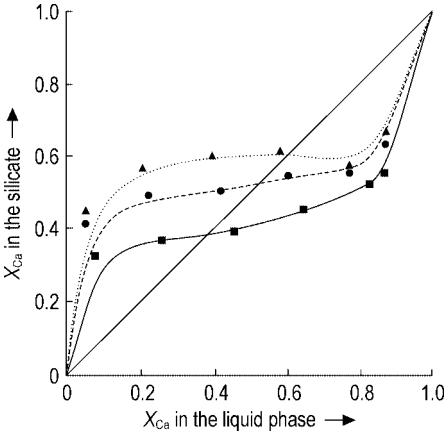

Figure 24. Equilibrium isotherms of the Na^+/Ca^{2+} system at 293 (■), 313 (●) and 333 K (▲) and total ion concentration of 0.02 N [77]

The calcium binding capacity is temperature-dependent and increases at higher temperatures. Table 8 shows the calcium- and magnesium-binding capacity for a crystalline layered silicate and an amorphous type. The calcium-binding capacity for both types is in the same range. It is interesting that the magnesium binding capacity for the two silicates is much higher than that of zeolite A. In addition, the binding capacity increases with increasing pH. The solubility of the amorphous type is higher than that of crystalline layered silicate. These silicates differ from conventional sodium silicates in their solubility, which is much lower than that of conventional silicates. Therefore the calcium ions of the washing liquor react with nonmolecularly dissolved sodium silicate to form insoluble calcium and magnesium silicates. These solids are well dispersed as fine particles and do not precipitate onto fabrics.

Table 8. Binding capacity of sodium silicates [69]

Sodium silicate	Ca binding capacity (mg CaO/g)	Mg binding capacity (mg MgO/g)
crystalline	90–110 (25°C)	100–125 (25°C)
amorphous	120–140 (40°C)	150–170 (40°C)

2.4.2. Oily Soil

In typical washing and cleansing operations the wetting situation as described in Section 2.3.1 is much more complicated, since the solid surfaces that are present tend to be irregularly covered with oily or greasy soils. Thus, the wash liquor must compete with soil in wetting the surface. Figure 25 illustrates the problem schematically.

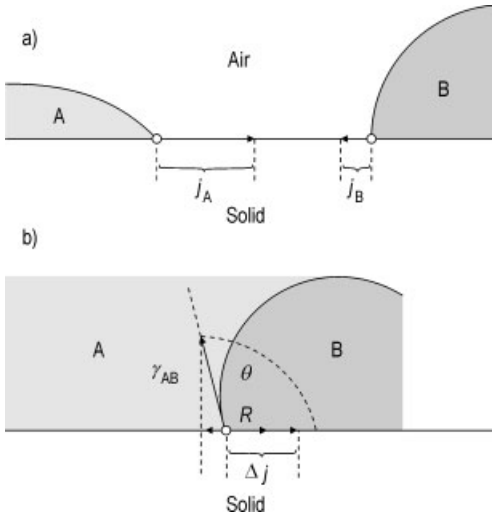

Figure 25. Two liquids on a solid surface
a) Separated; b) Overlapping
A) Wash liquor; B) Oily dirt

If two drops of different liquids (e.g., wash liquor A and an oily residue B) are placed next to each other on a solid surface S, two different wetting tensions j_A and j_B act on the surface. If the two liquids come into direct contact and form a common interface, then the difference in their wetting tensions Δj, the so-called oil-displacement tension, is effective along the line of contact. In addition, a portion of the interfacial tension γ_{AB} remains effective along the substrate surface, but with reversed sign and a magnitude of $\gamma_{AB} \cos \theta$, where θ is the contact angle in the oily phase B. The total force that results is the contact tension R, given by Equation 5.

$$R = \Delta j + \gamma_{AB} \cos\theta \quad (5)$$

The contact tension acts in such a direction as to constrict the oil drop. As a result of adsorption of surfactants from phase A, the value

of Δj is increased and that of γ_{AB} is decreased. The value of γ_{AB} cos θ is negative for obtuse contact angles. From Equation 5, it can be seen that both quantities contribute to an increase in the contact tension and thus to better penetration of the oil drop. This complex process, in which a surface undergoes wetting first by oil and then by water, is known as rollup. For many cases of practical interest, rollup is not a spontaneous phenomenon. More often, total constriction of an oil drop can be achieved only if mechanical energy is applied to the system. The required energy is directly proportional to the interfacial tension γ_{AB} and decreases with increasing surfactant concentration. The process is shown in Figure 26 [78].

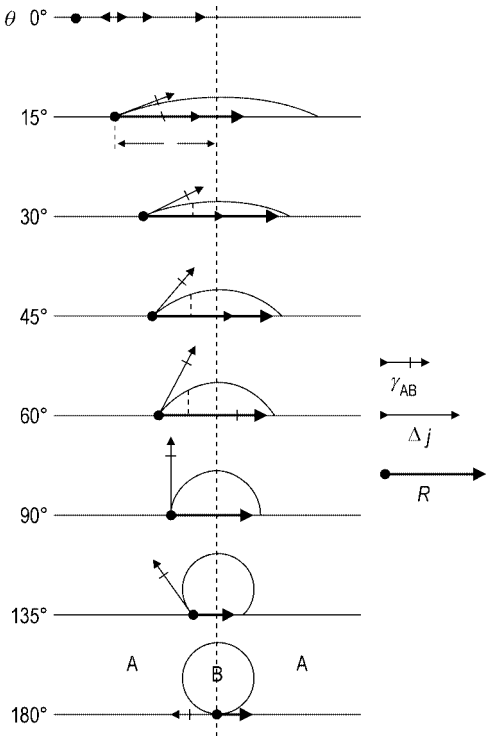

Figure 26. Schematic view of the phases of displacement of an oily drop B by a detergent A

All the relationships presented thus far confirm the premise that interfacial tension is the primary force resisting the removal of liquid soils, and this force must be minimized if the washing process is to be effective.

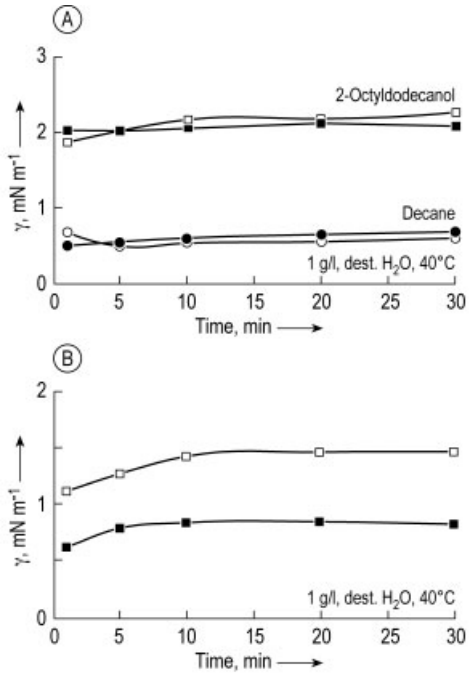

Figure 27. Dynamic interfacial tension of $C_{12/14}$ fatty alkyl sulfate (FAS) and linear alkylbenzenesulfonate (LAS) as a function of time for decane and 2-octyldodecanol (top) and isopropyl myristate (IPM) (bottom)

Findings such as these are independent of the nature of the fiber and apply to all soils containing hydrophobic particles and oily material [79]. The oil specificity of surfactants is very important for the use in different applications [80]. Figure 27 shows the interfacial tension of (linear) alkylbenzenesulfonate (LAS) and of a fatty alcohol sulfate ($C_{12/14}$ FAS) against three different oils [81]. Both surfactants show nearly identical interfacial tension against decane and 2-octyldodecanol. In contrast the interfacial tension of FAS against isopropyl myristate is significantly lower than that of LAS. As a result, most commercial detergents contain mixtures of surfactants in carefully determined proportions. The interfacial tension in detergents is a measure of their product performance. This is shown by way of example for five different heavy-duty detergents in Figure 28 [80].

Oil-removal values also reach their upper limit at the critical micelle concentration of the surfactant employed [82]. This is illustrated in Figure 29 for the example of removing olive oil from wool.

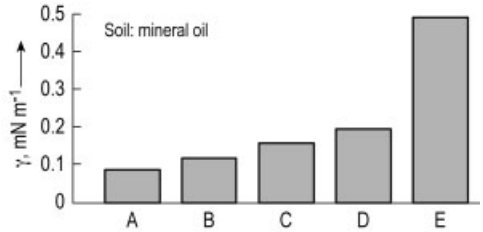

Figure 28. Interfacial tension of different detergents (A to E) against mineral oil [80]

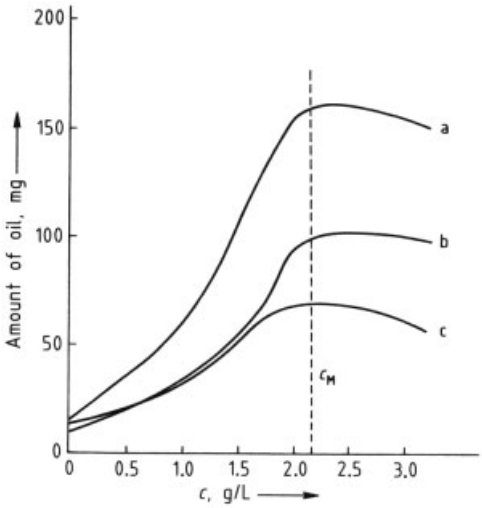

Figure 29. Removability of an olive oil–oleic acid mixture from wool, shown as a function of sodium *n*-dodecyl sulfate concentration for various degrees of oil coating (c_M=critical micelle concentration) [82]
a) 4.6 % Oil coat; b) 3.3 % Oil coat; c) 2.2 % Oil coat

From this behavior it can be concluded that effective washing is a result of the properties of individual surface-active ions, not of micelles. Just as with emulsification, however, micelles are able to solubilize water-insoluble materials and thereby prevent redeposition of previously removed soil in the later stages of the washing process. Values reported in the literature for critical micelle concentrations are based on pure solutions of surfactant. However, in an actual washing process, surfactants are invariably adsorbed onto a variety of surfaces, which means that the true surfactant concentration in the wash liquor is correspondingly diminished. The true surfactant concentration in solution is the only significant measure.

One way to reduce the interfacial tension is to create appropriate mixed adsorption layers comprised of surfactants of differing constitution. For example, Figure 30 shows the interfacial tension of the system water–toluene as a function of composition for a surfactant mixture containing the anionic surfactant sodium *n*-dodecyl sulfate (SDS) and the nonionic surfactant hexadecyl triglycol ether ($C_{16}E_3$), whereby the total surfactant concentration is kept constant [83].

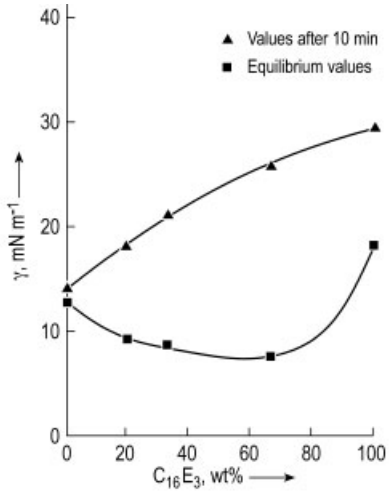

Figure 30. Interfacial tension after 10 min and after equilibrium vs. composition of the mixture (organic phase toluene, concentration 0.5 g/L, pH = 10, T = 298 K) [83]

Even small additions of one surfactant to another may cause a significant reduction in the interfacial tension. In a broad concentration range mixtures of these surfactants show lower values of interfacial tension than the two surfactants alone. The presence of a small amount of nonionic surfactant in the surface layer reduces the mutual repulsion of negatively charged groups on the anionic surfactant, as a result of which adsorption increases. A quantitative relationship for the surface activity and the CMC for surfactant mixtures was derived by Rosen [84]. The CMC of a mixture can be described by the following equations with β as the interaction parameter:

$$\frac{1}{\mathrm{CMC_{mix}}} = \frac{\alpha}{f_1\,\mathrm{CMC_1}} + \frac{1-\alpha}{f_2\,\mathrm{CMC_2}} \tag{6}$$

$$f_1 = \exp \beta[1-x_1]^2 \tag{7}$$

$$f_2 = \exp \beta\, x_1^2 \tag{8}$$

For an attractive (synergistic) interaction the parameter β is negative, and for repulsion β is positive. The attractive interaction leads to a lower CMC of a specific surfactant mixture compared to the ideal case (Fig. 31).

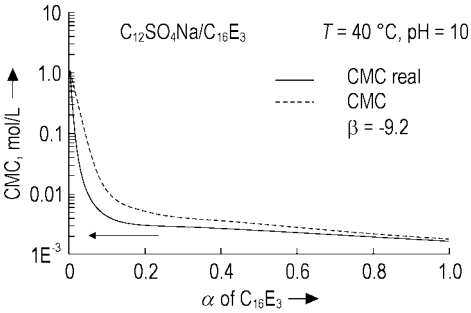

Figure 31. Critical micelle concentration of mixtures of an anionic surfactant $C_{12}SO_4Na$ and a nonionic surfactant $C_{16}E_3$ as a function of the mole fraction α of nonionic surfactant in comparison to ideal behavior of the mixtures [83]

Phase Behavior Phase behavior can have a significant impact on detergency [85]. If there is no phase change for the surfactant water system, a linear dependence of the cleaning efficiency on temperature is observed, which is shown as an example for household detergents in Figure 32. The surfactant is in an isotropic micellar solution at all temperatures. The cloud point of the surfactant used here is 85 °C at the given concentration (2 g/L), i.e., above the highest washing temperature.

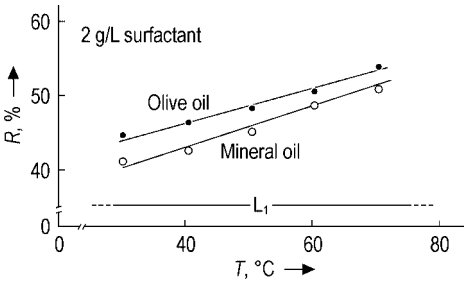

Figure 32. Phase behavior of $C_{12}E_9$ and household detergency [85]

Tests with other pure ethoxylated surfactants have revealed that a discontinuity is observed with respect to oil removal versus temperature in cases where dispersions of liquid crystals exist in the binary system water–surfactant. Figure 33 shows that the detergency values for mineral oil and olive oil, i.e., two oils with significantly different polarities, are at different levels. It also demonstrates that in both cases a similar reflectance versus temperature curve exists. In the region of the liquid-crystal dispersion, i.e., between 20 and 40 °C, oil removal increases significantly. Above the phase transition $W + L_\alpha \rightarrow W + L_3$, between 40 and 70 °C, no further increase in oil removal takes place. For olive oil, a small decrease in detergent performance is observed. The interfacial tensions between aqueous solutions of $C_{12}E_3$ and mineral oil lie at about 5 mN/m at 30 and 50 °C. These relatively high values indicate that in this system the interfacial activity is not the decisive factor in oil removal from fabrics. The macroscopic properties of the liquid-crystal dispersion seem to be responsible for the strong temperature dependence. It can be assumed that fragments of liquid crystals are adsorbed onto fabric and oily soil in the $W + L_\alpha$ range during washing. The local surfactant concentration is therefore substantially higher in comparison to the molecular surfactant layer that forms when surfactant monomers adsorb. As the viscosity of liquid crystals in the single-phase range is strongly temperature dependent, it can be assumed that the viscosity of a fragment of a liquid crystal deposited on a fabric also significantly decreases with increasing temperature. Thus, the penetration of surfactant into the oil phase and removal of oily soil is promoted.

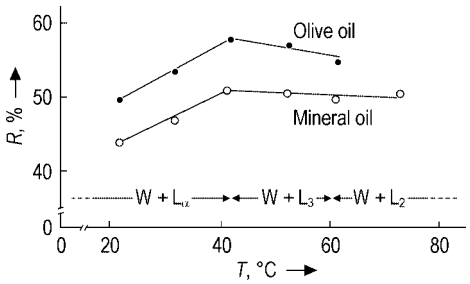

Figure 33. Phase behavior of the polyoxyethylene alcohol $C_{12}E_3$ and detergency (2 g/L surfactant) [85]

Technical-grade surfactants are of specific interest for applications. As in the case of pure nonionic surfactants, definite ranges for technical-grade surfactants exist in which there is only a

slight dependence of oil removal on temperature (Fig. 34). For $C_{12/18}E_5$, this is in the range of the two co-existing liquid phases (W + L_1), and for $C_{12/18}E_4$ it is in the range of the surfactant liquid phase (W + L_2). An unusually strong increase of oil removal with increasing temperature occurs in the region of the liquid-crystal dispersion (W + L_α). At 30 and 50°C the interfacial tensions between aqueous surfactant solutions and mineral oil and the contact angles on glass and polyester were determined for $C_{12/18}E_4$. Whereas the values of interfacial tensions are practically identical (ca. 0.1 mN/m) the contact angles on both substrates are slightly less advantageous at higher temperatures. Hence, the increased oil removal between 30 and 50 °C cannot be attributed to an increase in the adsorbed amounts of surfactants. Rather in both cases, the decisive part is probably played by the macroscopic properties of the liquid crystal dispersion and their temperature dependence.

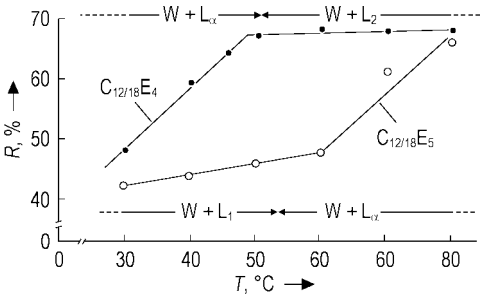

Figure 34. Phase behavior of the polyoxyethylene alcohols $C_{12/18}E_4$ and $C_{12/18}E_5$ and detergency [85]

During oil removal from hard surfaces or fabrics ternary systems occur in which three phases coexist in equilibrium. These systems are also referred as three-phase microemulsions. These effects were studied in detail for alkyl polyglycol ethers [86]. Depending on temperature different phases exist, having a three-phase region between the temperatures T_l and T_u. When these three phases are formed, extremely low interfacial tensions between two phases are observed. Figure 35 (bottom) represents the three-phase temperature intervals for $C_{12}E_4$ and $C_{12}E_5$ versus the number *n* of carbon atoms of *n*-alkanes. The top part of Figure 35 shows the detergency of these surfactants for hexadecane. Both parts of Figure 35 indicate that maximum oil removal occurs in the three-phase interval of the oil used (*n*-hexadecane) [87].

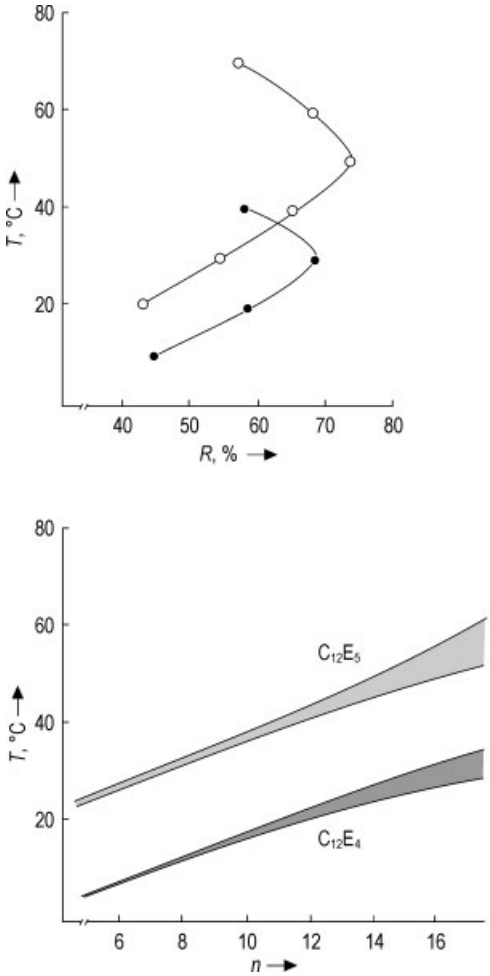

Figure 35. Detergency of $C_{12}E_4$ and $C_{12}E_5$ against hexadecane as a function of temperature (left) and the corresponding three-phase ranges for these surfactants as a function of the number *n* of carbon atoms of alkanes [87]

This means that not only the solubilization capacity of the concentrated surfactant phase, but probably also the minimum interfacial tension existing in the range of the three-phase body are responsible for maximum oil removal. Further details about the influence of the polarity of the oil, the type of surfactant, and the addition of salt are summarized in [88].

Studies of diffusional phenomena have direct relevance to detergency processes. Experiments are reported which investigate the effects

of changes in temperature on the dynamic phenomena which occur when aqueous solutions of pure nonionic surfactants contact hydrocarbons such as tetradecane and hexadecane. These oils can be considered to be models of nonpolar soils such as lubricating oils. The dynamic contact phenomena, at least immediately after contact, are representative of those which occur when a cleaner solution contacts an oily soil on a polymer surface. With $C_{12}E_5$ as the nonionic surfactant at 1 wt % in water, quite different phenomena were observed below, above, and well above the cloud point when tetradecane or hexadecane was carefully layered on top of the aqueous solution. Below the cloud point temperature of 31 °C, very slow solubilization of oil into the one-phase micellar solution occurred. At 35 °C, which is just above the cloud point, a much different behavior was observed. The surfactant-rich L_1 phase separated to the top of the aqueous phase prior to the addition of hexadecane. On addition of the oil, the L_1 phase rapidly solubilized the hydrocarbon to form an oil-in-water microemulsion containing an appreciable amount of the nonpolar oil. After depletion of the larger surfactant-containing drops, a front developed as smaller drops were incorporated into the microemulsion phase.

Unlike the experiments carried out below the cloud point temperature, appreciable solubilization of oil was observed in the time frame of the study, as indicated by upward movement of the oil–microemulsion interface. Similar phenomena were observed with both tetradecane and hexadecane as the oil phases. When the temperature of the system was raised to just below the phase inversion temperatures of the hydrocarbons with $C_{12}E_5$ (45 °C for tetradecane and 50 °C for hexadecane), two intermediate phases formed when the initial dispersion of L_1 drops in water contacted the oil. One was the lamellar liquid-crystalline phase L_α (probably containing some dispersed water). Above it was a middle-phase microemulsion. In contrast to the studies below the cloud point temperature, there was appreciable solubilization of hydrocarbon into the two intermediate phases. A similar progression of phases was found at 35 °C with *n*-decane as hydrocarbon. At this temperature, which is near the phase-inversion temperature of the water–$C_{12}E_5$–decane system, the existence of a two-phase dispersion of L_α and water below the middle-phase microemulsion was clearly evident. These results can be utilized to optimize surfactant systems in detergents, and in particular to improve the removal of oily soils. The formation of microemulsions is also described in the context of the pretreatment of oil-stained textiles with a mixture of water, surfactants, and cosurfactants.

Besides cleaning efficiency the liquid-crystalline phases of surfactant systems at higher concentrations are of crucial importance for the processing of concentrated surfactant systems and the formulation as well as the application of liquid products. Studies of phase behavior can be supported by computer simulations, which can predict the liquid-crystalline phases [89].

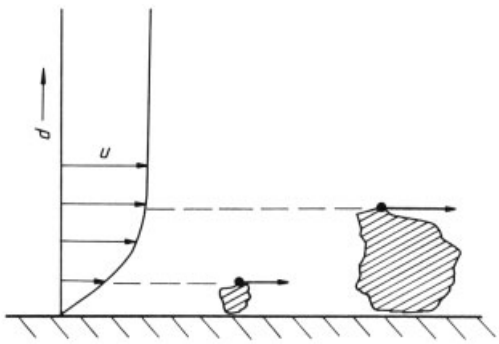

Figure 36. Schematic drawing of the effect of liquid flow on adhering particles of different size
u Flow velocity; *d* Distance from the solid surface

2.4.3. Particulate Soil

For solid soil particles above a certain size hydrodynamic effects are advantageous for soil removal [90]. The significance of hydrodynamic effects in the removal of particulate soil from fibers decreases with decreasing particle size. Even with vigorous mechanical action, a laminar film of finite thickness in which no flow takes place is present on every surface. The flow velocity gradient increases with increasing distance from the surface, as shown schematically in Figure 36.

Thus, an increase in mechanical force has a significant soil-removing effect on larger particles. A high flow velocity gradient in the immediate vicinity of a surface is excluded for hydrodynamic reasons. Therefore, washing machines

employ abrupt changes in direction of operation to achieve adequate turbulence near substrate surfaces. Even so, particles smaller than 0.1 μm cannot be displaced by mechanical means alone.

For the effective removal of solid particles below this size, adsorption of surfactants on the particle surface and stabilization of the particles in the washing solution are the most important steps. Concerning the stability of dispersed solid particles, charge effects play a major role. Different surfaces may have different surface charges. Through the adsorption of surfactants on the surfaces the charge effects may change and the particle are stabilized by electrostatic effects.

The electrostatic stabilization of such a dispersion is not only dependent on surfactant adsorption, but also on salt concentration and the pH of the washing solution. Dispersions can also be stabilized by other mechanisms, e.g., steric stabilization or stabilization through depletion. However, these mechanisms are not treated here.

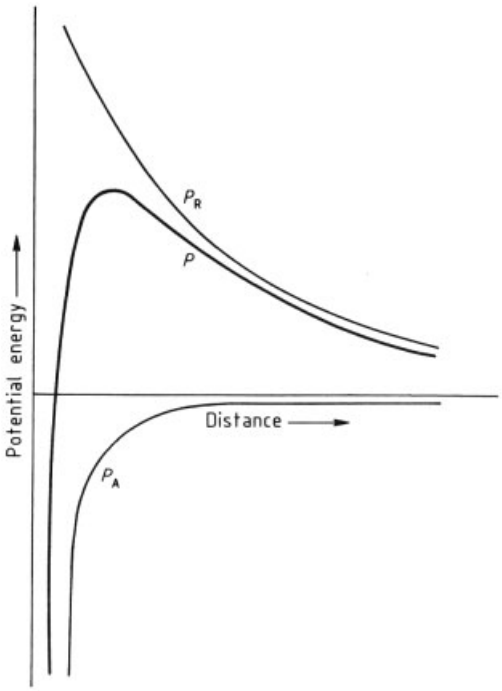

Figure 37. Calculated potential energy of attraction P_A and repulsion P_R as a function of the distance of a particle from a fabric, along with the resultant potential P; predictions based on the DLVO theory
DLVO computational parameter $z = 4$ [93]

Derjaguin–Landau–Verwey–Overbeek theory (DLVO). A theoretical understanding of the interaction forces causing a solid particle to adhere to a more or less smooth surface is based on the Derjaguin–Landau–Verwey–Overbeek (DLVO) theory [91, 92]. Since this theory was developed to explain the phenomena of flocculation and coagulation, however, it can be applied to the washing process only in modified form [93]. A plot of potential energy as a function of the distance of a particle from a fabric shows that the potential energy passes through a maximum (Fig. 37).

The minimum in the potential energy curve corresponds to the closest possible approach, i.e., to the minimum distance that can be established between the particle and the fiber. The maximum is a measure of the potential barrier that must be overcome if the particle either is removed from the fiber or approaches the fiber from a distance. Adhering particles are more easily removed if the potential barrier is small. Conversely, a soil particle already in the wash liquor is less likely to establish renewed contact with the fabric if the potential barrier is large.

With respect to the washing process, if a particle is bound to a fabric, only a single common electrical double layer located at the overall external surface exists initially. None is present within the zone of contact. During the washing process, new diffuse double layers are created which cause a reduction in the free energy of the system. The free energy of an electrical double layer is a function of distance and diminishes asymptotically to a limiting value that corresponds to a condition of no interaction between two double layers. Twice as much effort must be expended to bring a particle into contact with a substrate because of the presence on both surfaces of a double layer (curve 2 F in Fig. 38). The separation of two adhering surfaces is characterized initially only by van der Waals–London attractive forces P_A and Born repulsion forces P_B, since at this point no electrical double layer exists [93]. In Figure 38, the equilibrium condition corresponding to the potential energy minimum has been taken as the zero point on the abscissa. With increasing distance between the particle and the contact surface, a diffuse double layer arises, which assists in the separation process by establishing an element of repulsion. Thus, the true potential curve P for the sepa-

ration of an adhering particle in an electrolyte solution results from a combination of the van der Waals–Born potential and the free energy of formation of the electrical double layer. The important conclusion from Figures 37 and 38 is that an increase in the potential of the electrical double layer increases the energy barrier for particle deposition but decreases that for particle removal.

A schematic representation of the adsorption layers on substrate and soil particles is shown in Figure 39. As can be seen from the diagram, both surfactant layers advance to the point of particle–surface contact. One consequence is the development of a disjoining pressure, which leads to separation of the soil particle from the surface. This effect is obviously present with anionic surfactants as well. However, this pressure is the decisive factor with nonionic surfactants, due to the absence of any repulsive components of electrostatic origin.

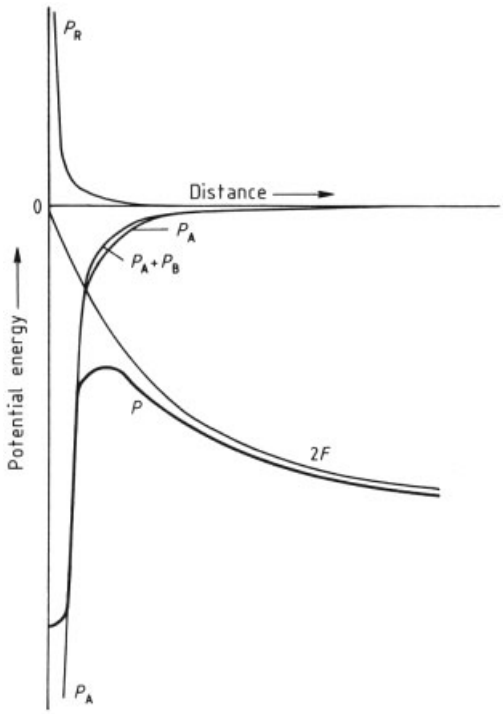

Figure 38. Potential energy diagram for the removal of an adhering particle [61]
P_B Born repulsion; P_A van der Waals attraction; F Free energy of the double layer; P Resulting potential curve DLVO computational parameter $z = 4$

Figure 40 shows the effect of several aqueous surfactant solutions on the surface potential of carbon black [61]. The surfactants chosen for the study all share the same hydrophobic component, but their hydrophilic groups vary. Electrophoretic mobility (EM) is taken as a measure of surface potential in this case.

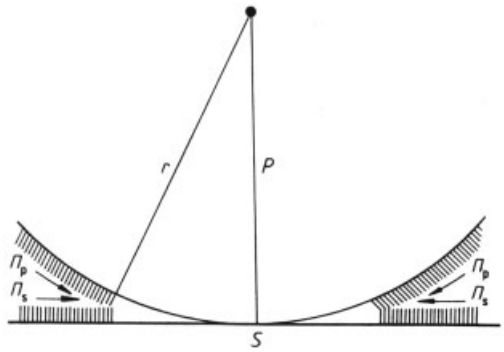

Figure 39. Schematic representation of adsorption-induced separation of a spherical particle from a hard surface
S Surface; P Particle; Π_S Splitting pressure of the surfactant layer on the surface; Π_P Splitting pressure of the surfactant layer on the particle [93]

Carbon black acquires a negative charge in water. The negative charge of pigments and fibers is further increased by adsorption of anionic surfactants. The corresponding increase in mutual repulsion is responsible for an increase in the washing effect. Dispersing power for pigments also increases for the same reason, whereas the redeposition tendency of removed soil is diminished.

In contrast to anionic surfactants, cationic surfactants reduce the magnitude of any negative surface charge. Consequently, electrostatic attraction between soil and fabric is increased. For this reason, cationic surfactants may cause a decrease in washing effect below that observed with pure water. Significant soil removal then occurs only at high surfactant concentrations, at which a complete charge reversal takes place on both fabric and soil. When the laundry is rinsed, however, a second charge reversal takes place on the fabric. At the point of electrical neutrality, redeposition of previously removed pigment soil is observed. This is why cationic surfactants are less suited for use in detergents than anionic surfactants.

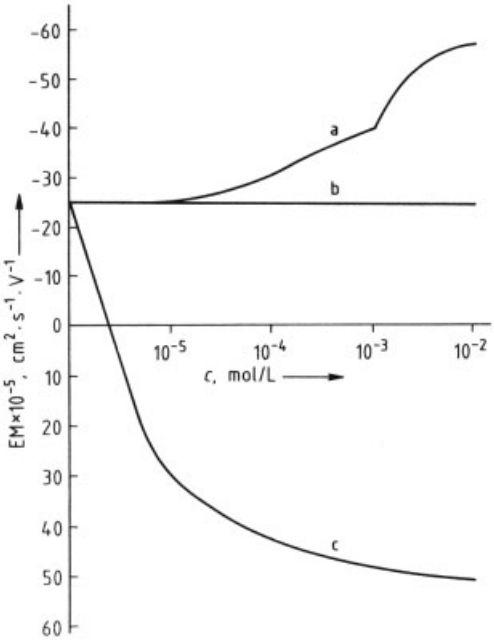

Figure 40. Electrophoretic mobility (EM) of carbon black in solutions containing various surfactants at 35 °C [61]

a) $C_{14}H_{29}OSO_3Na$ (anionic);
b) $C_{14}H_{29}O\ (CH_2CH_2O)_9H$ (nonionic);
c) $C_{14}H_{29}N\ (CH_3)_3Cl$ (cationic)

The negative influence exerted by calcium ions originating from water hardness can also be explained with the help of potential theory. According to the Schulze–Hardy rule, compression of an electrical double layer increases rapidly as the valence of a cation increases. Therefore, high concentrations of calcium might cause attractive forces to become the dominant factor, leading to significantly lower washing efficiency than would be achieved in distilled water.

Addition of an electrolyte generally leads to compression of the electrical double layer between two hard surfaces. Because the surface potentials cannot be measured directly, the zeta potential or electrophoretic mobility of a particle is used as a measure of its surface charge. The zeta potential can be defined as the potential difference in the electrical double layer at a shear plane between the fairly immobile layer on the surface of the particle and the diffuse layer outside the slipping plane when the solid and liquid phases are moved tangentially to each other. The zeta potential is a good indicator of the magnitude of the repulsive interaction between colloidal particles and is very strongly dependent on the electrolyte concentration in the solution. As the indifferent-electrolyte concentration increases, the zeta potential decreases and a sharply defined coagulation concentration can be identified. This is the point at which the potential energy barrier opposing coagulation disappears, and hence coagulation takes place (critical coagulation concentration, CCC). To prevent agglomeration a sufficient potential (Fig. 41) should exist, and the concentration of indifferent electrolyte should not exceed a certain level [94].

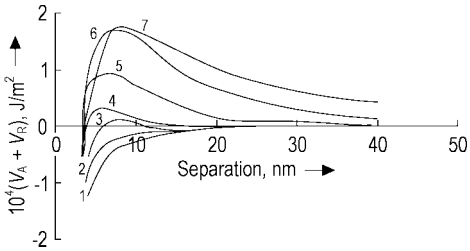

Figure 41. Total potential energy of interaction between flat plates of kaolinite as a function of separation at pH 10.4 for various electrolyte concentrations [95] (cited in [94])

Generally, fibers and pigments in an aqueous medium acquire negative charges, whereby the extent of charge increases with increasing pH. This is illustrated in Figure 42, in which the zeta potential of various fibers is taken as a measure of electrical charge and is plotted against pH [90].

Essentially similar results are obtained for all major particulate soil components. This is one of the reasons for enhancement of wash performance by mere introduction of alkali. However, repulsive forces between soil and fibers alone are insufficient to produce satisfactory washing even at high pH.

The dependence of the surface charge of TiO_2 particles on pH in salt solutions of various concentrations is shown in Figure 43 [94]. At pH 5.8 the surface has zero charge. Below pH 5.8, the particles have positive surface charge, while they have a negative surface charge above pH 5.8. The point of zero charge is characteristic for the particular metal oxide. The influence of the electrolyte concentration on the absolute value of the surface charge is also obvious from Figure 43.

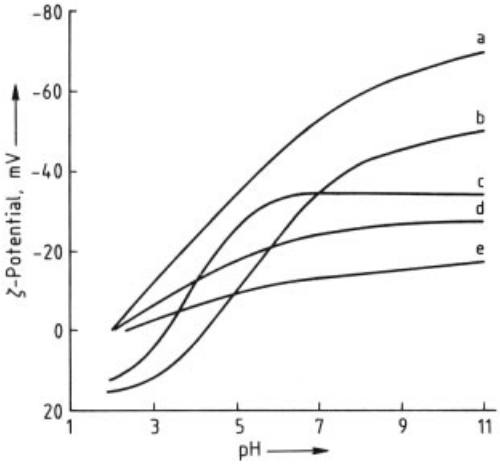

Figure 42. ζ Potential of various fibers as a function of pH [90]
a) Wool; b) Nylon; c) Silk; d) Cotton; e) Viscose

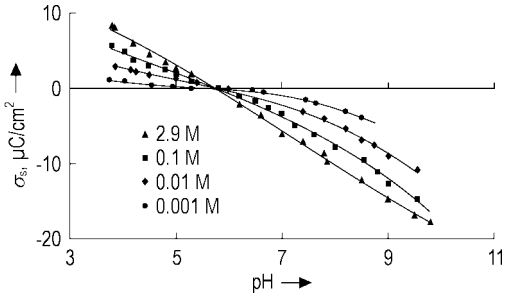

Figure 43. Typical examples of surface charge/pH behavior for TiO_2 in KNO_3 solutions in an inert atmosphere at 25 °C [96] (cited in [94])

From the theories of double layers it follows that the distribution of the countercharge also changes when the salt concentration is increased, to promote accumulation of charge in the region close to the surface. This explains the increase in positive charge with increasing salt concentration at low pH values. If specific adsorption of one of the ions of the electrolyte takes place on the particle surface, the picture is basically changed. Such ion adsorption may influence the position of the point of zero charge. Strong specific adsorption of ions on the particle surface leads to increased shift of the point of zero charge. Hence, salt concentration, pH value, and the resulting degree of surface charge of the particles play a major role in the electrostatic stabilization of particulate soil.

2.4.4. Chemical Reactions

2.4.4.1. Bleachable Stains

For the effective removal of colored stains like tea, coffee, red wine, and fruit juices chemical changes of the colored substances are necessary. This can be achieved by oxidation of the active groups or double bonds of the colored substances. This kind of oxidative bleaching will be performed with chlorine or oxygen bleaches. Oxygen bleach is the most common in Western Europe because the only byproduct after bleaching is water.

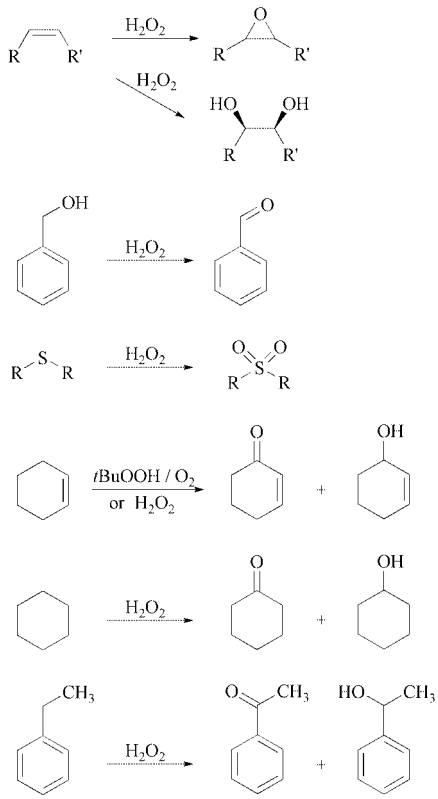

Figure 44. Schematic representation of possible chemical changes of substrates upon oxidation with H_2O_2 and an oxidation catalyst [97]

Most colored stains are relatively large conjugated molecules. Epoxidation of a double bond is one of the most typical chemical changes on oxidative bleaching. An epoxide is formed which can transform into an alcohol (Fig. 44).

Thus, the conjugated systems are destroyed and the molecule loses its color, although in many cases it still adheres to the fabric.

Figure 44 shows some of the chemical changes which can be expected to occur upon oxidative bleaching of colored stains [97].

H_2O_2. The most common source of active oxygen for oxygen bleach is H_2O_2, usually in the form of perborate or percarbonate. These can set free H_2O_2 spontaneously at higher temperatures during the washing process. The H_2O_2-releasing mechanism of percarbonate is illustrated below.

$$2\,Na_2CO_3 \cdot 3\,H_2O_2 \rightarrow 2\,Na_2CO_3 + 3\,H_2O_2 \qquad (9)$$

There are many theories about the active species of H_2O_2 bleaching. Mostly the HOO^- anion at higher pH values [98] or the superoxide radical $O^{\cdot 2-}$ [99, 100] are held responsible for the bleaching reaction. However, it is well known that the H_2O_2 bleaching process is most effective at temperatures of 60 °C and higher.

The mechanism for formation of radicals upon bleaching with H_2O_2 is suggested to be [100]:

$$HO_2^- + H_2O_2 \rightarrow HO_2^{\cdot} + HO^- + HO^{\cdot} \qquad (10)$$

The radicals are said to be responsible for the bleaching reaction. The pH dependence of the bleaching as percentage stain removal is shown in Figure 45. The bleaching efficiency reaches its maximum at a pH value of about 11. This corresponds to the pK_a value of H_2O_2, which is 11.7 at 25 °C.

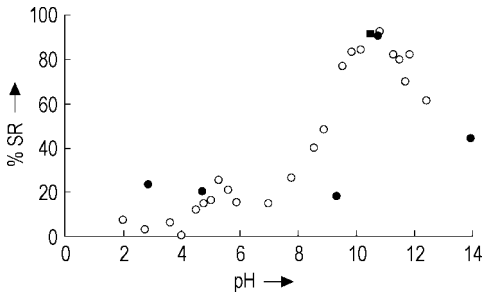

Figure 45. Relative stain removal (% SR) after 60 min of bleaching with H_2O_2 at 60 °C and different pH values [100] ○ = 6.5 × 10^{-3} mol/L H_2O_2 in buffer solution. Scavenging test with ■ = 10^{-3} mol/L 4-nitroso-N,N-dimethylaniline, ● = 10^{-3} mol/L potassium hexacyanoferrate (11)

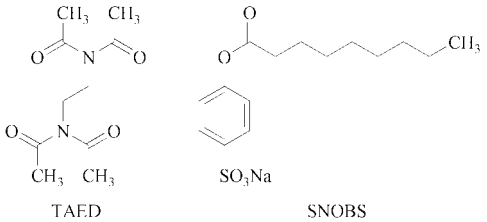

Figure 46. The two most common bleach activators TAED and SNOBS

Activators. To achieve higher bleaching efficiency at lower temperatures a bleaching activator like tetraacetylethylenediamine (TAED, Europe) or *p*-nonanoyloxybenzolsulfonate (SNOBS, USA; Fig. 46) can be added in combination with percarbonate or perborate [101].

By reaction with H_2O_2, tetraacetylethylenediamine forms two moles of peracetic acid per mole of TAED. The peracetic acid is active at lower temperatures than the H_2O_2 itself. The combination of H_2O_2 and peracetic acid enables effective bleaching over a broad range of temperatures. SNOBS reacts with H_2O_2 to form pernonanoic acid, which has similar advantages to peracetic acid.

The reaction with H_2O_2 to form a peracid under alkaline conditions is shown in Figure 47 [102]. The rate of peracetic acid formation increases with increasing pH value of the solution.

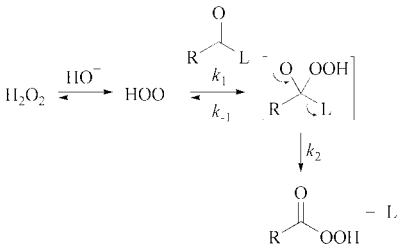

Figure 47. Mechanism for the transformation of H_2O_2 into a peracid by using an activator [101]

The kinetics of the transformation of H_2O_2 into peracetic acid is shown in Figure 48 [102].

The mechanism of bleaching with a peracid can occur by a nucleophilic mechanism at pH values above the pK_a of the peracid, and by an electrophilic mechanism at pH values below the pK_a value of the peracid. This is shown schematically in Figure 49 [103].

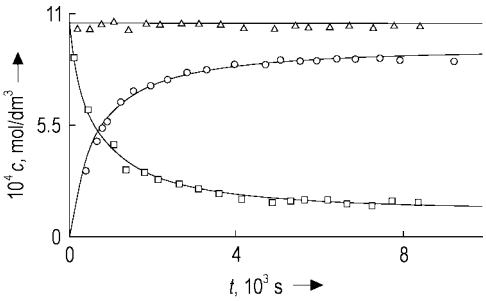

Figure 48. Time course of the reaction of TAED and H_2O_2, showing the formation of peracetic acid as a function of time at pH 9.6. ○ = peracetic acid, □ = H_2O_2, △ = peracetic acid plus H_2O_2 [102].

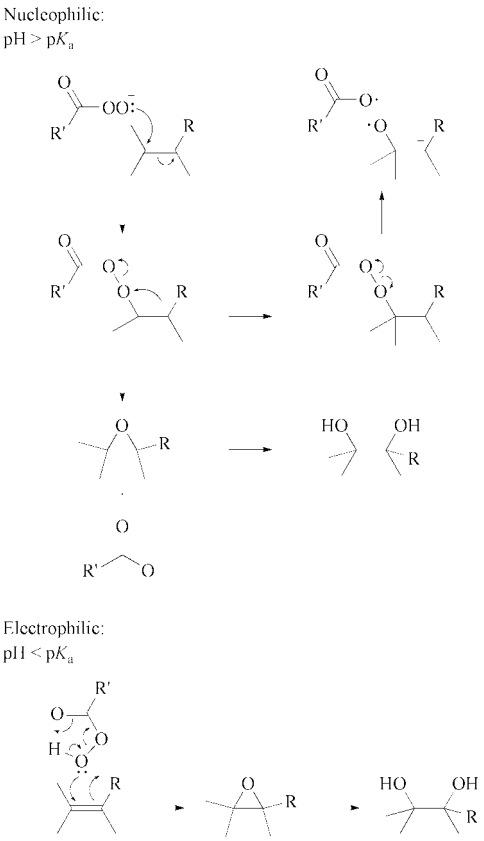

Figure 49. Reaction mechanism for bleaching with a peracid at pH values above and below the pK_a value of the peracid [103]

The bleaching efficiency of a peracid as a function of the pH value of the washing liquid is shown in Figure 50. The bleaching efficiency reaches its maximum at pH values close to or slightly above the pK_a value of the peracid, and collapses at pH values higher than this. The pK_a value of peracetic acid is 8.2, which corresponds to its maximum bleaching performance.

Peracids need not necessarily be formed in situ. They can also be directly included in the formula without extra addition of H_2O_2. However, few effective and stable peracids are commercially available, and hence the application of peracids in detergent formulations is not common.

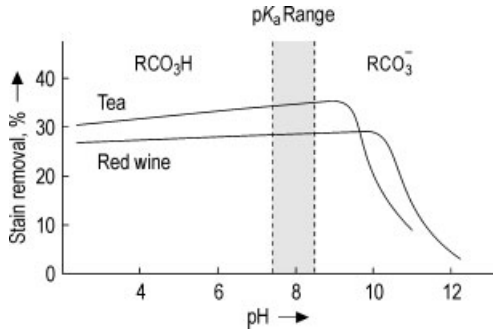

Figure 50. Bleaching performance of peracid in dependence on the pH value for tea and red wine stains [104]

Decomposition of H_2O_2 or Peracid. The decomposition of hydrogen peroxide can proceed spontaneously or catalytically.

Autoxidation of H_2O_2 to water and oxygen (Eq. 11) is an exothermic process. At higher temperatures this reaction takes place quickly, but at room temperature it is relatively slow, and hence H_2O_2 can be regarded as stable at room temperature.

$$2\,H_2O_2 \rightarrow 2\,H_2O + O_2 \tag{11}$$

pH, light, and traces of metal ions as catalysts can influence the inefficient decomposition of hydrogen peroxide. In this case decomposition proceeds by a radical mechanism, in which the free radicals from Equation 10 function as chain propagators [100].

$$HO^{\cdot} + H_2O_2 \rightarrow H_2O + HO_2^{\cdot} \tag{12}$$

$$H_2O_2 + HO_2^{\cdot} \rightarrow H_2O + O_2 + HO^{\cdot} \tag{13}$$

Decomposition of peracids, however, leads to the formation of the corresponding carboxylic acids, which are not bleach-active.

Catalytic Bleaching with H_2O_2. Using a metal catalyst in addition to H_2O_2 can lead to even better bleaching efficiency at lower temperatures. Although the mechanism of the catalyzed bleaching reaction is not fully understood, the most common opinion in the literature is that the transition metal complex forms a metal oxo intermediate with the active oxygen species (H_2O_2) [103]. The metal is hereby oxidized to a higher state. On oxidation of a substrate by the metal oxo intermediate, the metal returns to its original oxidation state. Cleavage of the metal–oxo bond can take place via a homolytic or a heterolytic path. In the case of homolytic cleavage radicals are formed and the further mechanism is radical. In the case of heterolytic cleavage a nonradical mechanism takes place.

Stability to hydrolysis in alkaline solution is one of the main problems concerning most of the catalysts suggested in the literature. This is a problem especially for manganese and iron catalysts. On hydrolysis, water-insoluble MnO_2 and Fe_2O_3 are formed, which are inert in the bleaching reaction.

Fiber care is another important problem concerning catalytic bleaching with H_2O_2. Until now, high catalyst performance has been associated with high fiber damage. This has been a restriction for the general use of catalysts in detergents until now.

2.4.4.2. Enzymatic Soil Removal

Detergent compositions usually contain several different kinds of enzymes. Generally, the action of the enzymes is to cleave large biological polymers such as proteins in natural stains and to make their subsequent removal by other detergent components easier. Enzymes are very selective biological catalysts, and hence one particular class of enzyme can cleave one particular class of biopolymer.

One of the most common enzymes in detergent formulations is protease, which is especially active in the degradation of proteins and polypeptides. Cellulases are responsible for the cleavage of amorphous cellulose, as in vegetable- and fruit-based stains, whereas amylase degrades starch and helps to remove starch-based soils like flour- or potato-based stains. Lipase hydrolyzes the ester bonds of natural animal and plant fats [105, 106].

Enzymes typically show a maximum efficiency at a certain pH value and a certain temperature. To optimize the efficiency it is important that the maximum performance is within the range of the normal detergent parameters. The typical performance maximum in dependence on the pH value of an amylase is shown in Figure 51 [107].

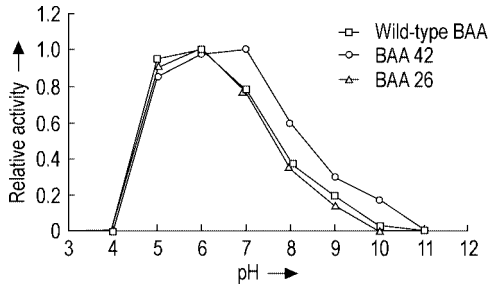

Figure 51. pH–activity profiles for amylase [107]

As the subject of this article is the physical chemistry of the washing process, the mechanism of enzyme activity will not be explained further here. More detailed information about the enzyme mechanism in detergents can be found elsewhere [105 – 108] (see also → Enzymes, Section 5.2.1).

2.4.5. Soil-Repelling Effects

Generally it is desired that fat and pigment soils which have been removed from the textile surface are not redeposited on the textile surface during the washing process. As a result of redeposition the textiles may appear somewhat gray after several wash cycles. To prevent the redeposition of such soils several effects play a role. Once removed from the surface the soil must be stabilized in solution. At the same time the surface can be treated to prevent soil from adhering to the surface again. This may lead to the long-term effect that soil is generally less firmly fixed to the surface and can be more easily removed during the next washing cycles. These effects are also known as soil-repellent effects. The easiest way to achieve soil-repellent performance is to add a soil-repelling agent to the

washing formulation. These are mostly polymers, and may work by one of many mechanisms. The most important substances increase the electrostatic repulsion of soil molecules or bind the molecules into a polymer matrix in the solution [109]. Other possibilities to achieve a soil-repellent effect are different kinds of textile pretreatment. The common goal for both pretreatment and most detergent-delivered soil-release agents is a polymeric molecular layer which covers the fibers of the textile and modifies its surface characteristics.

In a surfactant-free environment, the adsorption of most macromolecules onto a solid surface is effectively an irreversible process. The reason for this irreversibility is the vast number of points of contact that exist between macromolecule and substrate. From a statistical standpoint, the number of bonds between the two is so large that adherence is ensured regardless of the strength of the bonds. Most of the observed binding results from weak hydrophobic interactions.

In a multicomponent system containing both surfactants and macromolecules, competitive adsorption is possible at the substrate surface. This permits the surfactant to successively destroy the individual points of contact binding the macromolecule, thereby displacing it from the interface. This mechanism is most often observed for nonionic surfactants. Anionic surfactants are often capable of forming polymer–surfactant complexes, in the course of which the conformation of the macromolecule is altered in such a way as to reduce the extent of its attraction to the interface.

Cotton and synthetic fibers such as polyester have very different surface characteristics. Whereas cotton is hydrophilic and negatively charged in water at alkaline pH values, synthetic fibers are comparatively neutral. The affinity of hydrophobic oils and greases is therefore higher to the hydrophobic polyester surface then to the hydrophilic cotton surface. Oil and grease wet the synthetic textile surface comparably well and make the rolling up process of oil removal more difficult. Through adsorption of a soil-release agent the surface characteristics are changed to more hydrophilic, and hence the soil removal process is made easier. On a more hydrophilic surface the affinity of hydrophobic soils is less strong and water can penetrate better into the fabrics and between the soil and the modified surface. Because of this, different soil-release agents should be chosen for cotton and for synthetic fibers. In fact up to now there are more effective soil-release agents for synthetic fibers than for cotton. The absolute amount of polymer adsorbed on the surface is therefore mostly not the determining parameter [110].

Polymers or nanoparticles can be added to the surface of textiles during production and lead to soil-repellent qualities which are said to be retained for several washes. For example, nanosized whiskers aligned by proprietary spines, designed to repel hydrophobic or hydrophilic liquids, can be attached to the fabrics during production [111]. Another example of soil repellency through pretreatment is the use of plasma technology to modify or to apply a thin polymer layer to the cotton surface [112, 113]. By using radio-frequency plasma, chemicals such as fluorocarbons can be activated and polymerized on the surface of the textile. Ar, O_2, air, and NH_3 gas plasma treatments have been shown to increase the wettability of the cotton surfaces, whereas fluorinated gas plasma leads to a decrease in water adsorption. This treatment generates more hydrophobic surfaces [113].

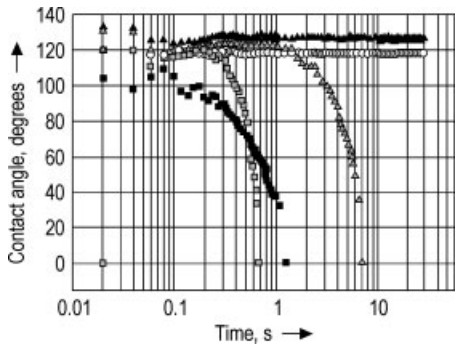

Figure 52. Dynamic contact angles for knitted fabric after soiling and washing with and without pretreatment with a soil-release polymer. (○) untreated; (●) after treatment with a soil-release polymer, only one value recorded; (▥) stained after treatment; (■) washed after treatment and staining; (△) stained without treatment; (▲) washed after staining without treatment [114]

A new method to visualize the soil-release polymer on the textile surface and its influence on the degree of soiling has been developed [114, 115]. The visualization of the transparent

film was achieved with an imaging instrument operating on the principle of chromatic aberration. Water contact angles before and after impregnation of different polyester fabrics with a polyether—polyester soil repellent polymer are shown in Figure 52 [114]. The contact angle changes to small values after impregnation, and this enables water to penetrate into the textile.

The residual soil degree could be evaluated and the relative cleaning performance calculated. If the fabrics are washed after impregnation prior to staining, the effect is much less. It is suggested that the polymer film shields the fabric against soil, and that the soil protects the impregnation from being washed out.

Some further examples of soil-release polymer are acrylate polymers and terephthalate polyesters. Polyacrylate films may be deposited onto the surface and then cured in the presence of a diepoxy compound. This leads to a cross-linked polymer film which inhibits soil deposition. When a terephthalate polyester is used as soil-release polymer the aromatic/glycol groups orient themselves to the polyester fabric surface while the ethylene oxide blocks of the polymer are oriented into the solution and hinder the soil from coming into contact with the fiber surface. Adsorbed polyester can form a hydrophilic film which remains on the fabric after it is removed from the wash solution and dried. Upon washing with detergent compositions containing the soil-release polymer the film is renewed. Hence, the soil-release effect of this polymer class increases after several washing cycles [109].

For cotton other soil-release polymers, e.g., carboxymethyl cellulose, are more effective. To ensure satisfactory results with mixed laundry, combinations of several antiredeposition agents often must be used.

Figure 53 reveals clear differences in the time-dependent increase in weight when prewashed polyester is immersed in water [80]. The polyester fibers had previously been washed with different detergent formulations. Table 9 shows that carboxymethyl cellulose, a frequently used antiredeposition agent for cotton, has no effect on polyester. By contrast, methylhydroxypropyl cellulose and soil repellents cause the polyester surface to become considerably more hydrophilic. It is noteworthy that the effects are retained with both detergents, albeit to a somewhat reduced extent. Only with these formulations is a significant increase in the soil antiredeposition effect observed, as evidenced by the changes in percentage remission. Hydroxyethyl cellulose can be regarded as a representative of numerous polymers which, though they are readily adsorbed from aqueous solutions and are capable of showing considerable antiredeposition activity in pure water, nonetheless loose most of their effectiveness in a detergent solution as a result of competitive adsorption and displacement by surfactants.

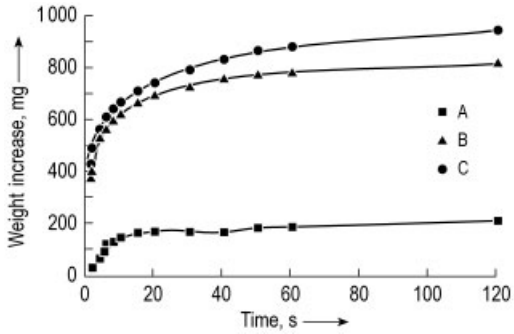

Figure 53. Kinetics of wetting of polyester fibers by three different detergents. The increase in weight of the fibers during wetting by pure water is plotted as a function of the wetting time [80]

2.5. Formulation

2.5.1. Foam

The control and adjustment of the foam properties is essential for all washing and cleaning processes. Depending on the product either the formation of high and stable foam is desired or low foam is absolutely necessary. From a physicochemical point of view the generation and stabilization of foam depends on several parameters (Fig. 54).

The diffusion of the surfactants into the newly formed interface is an important parameter for foam formation. A faster diffusion of the surfactant into the surface and the incorporation into the surface layer facilitates the stabilization of the foam bubbles. The dynamic surface tension is a suitable parameter to characterize this effect. This was studied for different types of surfactants [54]. A correlation was derived between

Table 9. Comparison of changes in polyester wetting tensions with changes in remission for a heavy-duty detergent (I) and a low-temperature heavy-duty detergent (II) for polyester

		Antiredeposition agent (in % detergent charge)			Change in wetting tension*		Change in remission**	
		Sodium carboxy-methyl cellulose	Methylhy-droxy-propyl cellulose	Hydroxyethyl cellulose	I	II	I	II
Detergent I, g/L	Detergent II, g/L				Δj_v, mN/m	Δj_v, mN/m	ΔR, %	ΔR, %
7.4	4.5	0	0	0	0	−1		
0	0	0.5	0	0	1	1		
7.4	4.5	0.5	0	0	2	0	3	−1
0	0	0	0.5	0	22	22		
7.4	4.5	0	0.5	0	14	14	17	16
0	0	0	0	0.5	22	21		
7.4	4.5	0	0	0.5	3	2	5	0

* Δj_v = calculated for advancing contact angle conditions.
** ΔR = measured after three washes.

the relative dynamic surface pressure $\Pi(t)$ according to Equation 14 and the foam formation for standard conditions (Fig. 55).

$$\Pi\,(t)\,/\,\Pi_{\mathrm{eq}} = [\sigma_{\mathrm{water}} - \sigma\,(t)]\,/\,(\sigma_{\mathrm{water}} - \sigma_{\mathrm{eq}}) \qquad (14)$$

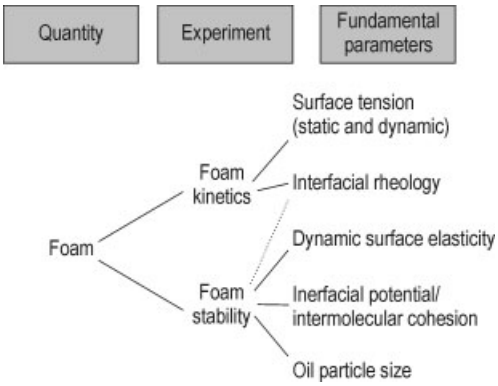

Figure 54. Foaming of surfactant–water systems: correlation of experimental data with physical chemical parameters

The diagram clearly shows the importance of dynamic effects of the surfactant solution regarding diffusion behavior for the formation and stabilization of newly generated foam bubbles.

Foam control and defoaming effects are specific aspects of detergents especially important for washing machines. A schematic of the defoaming mechanisms is shown in Figure 56. For a liquid defoamer principally two mechanisms are possible. Either the defoamer spreads on the surface or it forms a droplet. In the case of spreading, the foam lamella becomes thinner and finally disrupts. A defoamer droplet forms a bridge within the foam lamella, which also leads to film rupture. In a similar way hydrophobic defoamer particles perform.

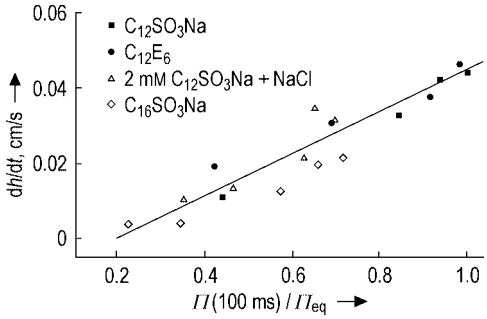

Figure 55. Correlation between relative dynamic surface pressure and foam formation dh/dt for different surfactants [57]

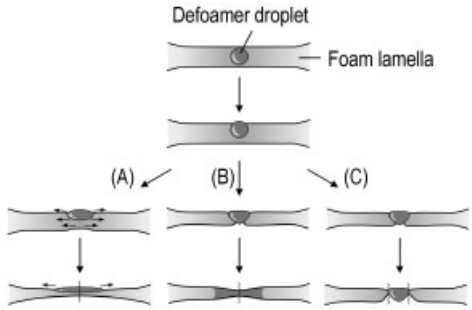

Figure 56. Schematic of defoaming mechanisms

2.5.2. Microemulsions

The spontaneous formation of microemulsions from aqueous surfactant systems and oily soils during the washing process was already mentioned in Section 2.4.2, but microemulsions can also be used as a washing liquid [116]. Solans et al. [117] published studies of textile cleaning with microemulsions. At washing temperatures between 23 and 34 °C, homogeneous microemulsions obtained from the system water/C_{12} - E_4/n-hexadecane and systems with technical nonionic surfactant mixtures removed 1.5 to 2 times more soil from wool, cotton, and cotton–polyester blended fabrics stained with oily and particulate soils than a highly concentrated liquid detergent (Fig. 57). Soil removal by the microemulsions was increased by 20 to 25 % by adding 0.05 M of the electrolytes sodium triphosphate and sodium citrate, which act as builders. The microemulsions also proved superior to the liquid detergent, in that they could be used seven times without losing any of their cleaning effectiveness. Förster et al. [116] systematically describe the phase behavior of quaternary systems, consisting of water, nonionic surfactants, a cosurfactant, and a hydrocarbon, also with regard to possible applications in textile cleaning.

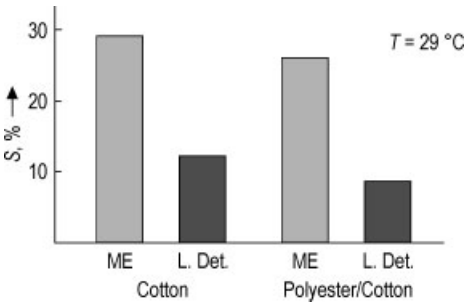

Figure 57. Soil removal *S* by a surfactant phase microemulsion (ME) and by a 1 % aqueous liquid detergent solution (L. Det) from different fabrics [117]

2.5.3. Phase Stability

The phase behavior of surfactant systems not only has an influence on the soil removal process, but also on the viscosity and stability of the detergent formulation. The phase behavior of a binary mixture of an anionic surfactant in water is dependent on the concentration of the surfactant and the temperature and is shown for a C_{14} fatty alcohol sulfate in Figure 58 [81]. At low concentrations a micellar phase appears. With increasing concentration it changes into a hexagonal phase and at even higher concentrations into a lamellar phase. With increasing alkyl chain length, the solubility of the surfactant shifts to higher temperatures, while the hexagonal phase becomes broader. The hexagonal phase is characterized by a maximum in viscosity and yield value, as shown in Figure 59 [81]. The phase behavior shown in Figure 58 is typical for anionic surfactants and plays an important role in the efficiency of soil removal and in the yield value and phase stability of liquid detergent formulations.

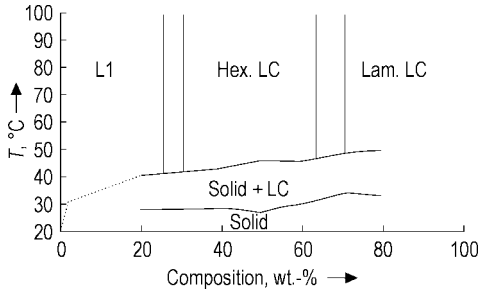

Figure 58. Phase diagram for C_{14} fatty alcohol sulfate in water at pH 11.5. L1 is the micellar phase, Hex LC the hexagonal, and Lam LC the lamellar phase [81]

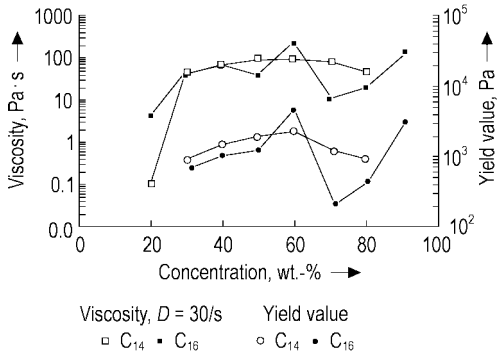

Figure 59. Viscosity and the yield value for two fatty alcohol sulfates with different alkyl chain length [81]

2.5.4. Softeners

The softening effect of cationic substances adsorbing on the textile surface is a well-known process. Mostly the substances responsible for the effect are quaternary ammonium compounds like ditallow alkyl dimethyl ammonium chloride or *N*-methyl-*N*,*N*,*N*-triethanolamine ditallow ester. Because of their positive charge they are adsorbed selectively on the negatively charged textile surface. The adsorbed amounts of quaternary compounds are very different for the various textile types. On cotton and wool the adsorption is highest, whereas it decreases for the synthetic fibers polyamide, polyacrylonitrile, and polyester, as shown in Figure 60 [118]. Because of interactions with the anionic surfactants of the detergent formulation the quaternary cationic compounds cannot be added directly to the detergent formulation. Due to the insufficient biodegradability of compounds like ditallow alkyl dimethyl ammonium chloride, they were changed into the more readily biodegradable ester quaternary ammonium compounds mentioned above. The adsorption behavior, however, remains the same for both compound classes.

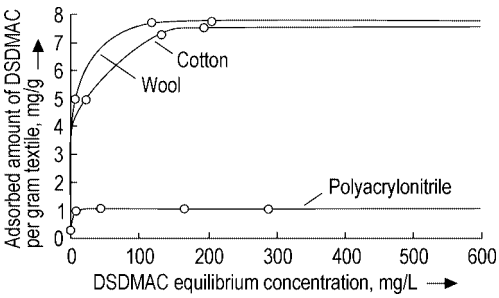

Figure 60. Adsorption of ditallow alkyl dimethyl ammonium chloride on different textiles [118]

2.6. Concluding Remarks

To simplify a physicochemical approach to the washing process, it has been necessary to treat separately each of the several phenomena involved and to isolate them from one another. In any real washing process, the various mechanisms are all at work more or less simultaneously. Thus, these mechanisms affect one another in a mutually supportive and additive way. Investigations on the physical chemistry of washing have made possible a rather thorough understanding of the process. Despite the complexity of the washing phenomenon and the continued presence of certain unanswered questions, increasing knowledge of physicochemical correlations has exerted a major influence on product development.

3. Detergent Ingredients

Detergents for household and institutional use are complex formulations containing up to more than 25 different ingredients. These can be categorized into the following major groups:

Surfactants
Builders
Bleaching agents
Auxiliary agents (additives)

Each individual component of a detergent has its own very specific functions in the washing process. To some extent they have synergistic effects on one another. In addition to the above ingredients, certain additives are made necessary for production reasons, whereas other materials may be added to improve product appearance.

3.1. Surfactants

Surfactants constitute the most important group of detergent components, and they are present in all types of detergents. Generally, surfactants are water-soluble surface-active agents comprised of a hydrophobic portion (usually a long alkyl chain) attached to hydrophilic or solubility-enhancing functional groups.

A surfactant can be grouped in one of four classes, depending on what charge is present in the chain-carrying portion of the molecule after dissociation in aqueous solution:

Anionic surfactants
Nonionic surfactants
Cationic surfactants
Amphoteric surfactants

Table 10 [125] provides an overview of these classes.

In general, both adsorption and detergency performance increase with increasing chain length. For example, ionic surfactants bearing *n*-alkyl groups show a linear relationship between the number of carbon atoms in the surfactant molecule and the logarithm of the amount of surfactant adsorbed on activated carbon or kaolin clay.

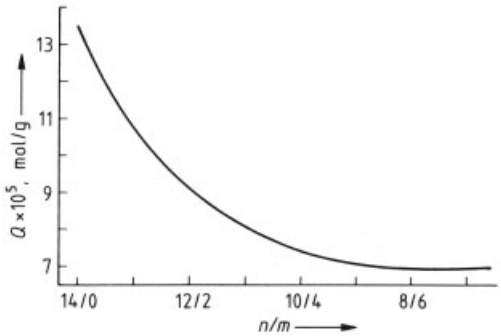

Figure 61. Decrease in adsorbed amount Q at equilibrium with increased branching of the hydrophobic residue [126] Adsorbent: activated carbon M; amount of adsorbent: 0.050 g; particle diameter: 0.084 cm; surfactant:

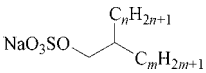

surfactant concentration: 1×10^{-4} mol/L; temperature: 25 °C

The structure of the hydrophobic residue also has a significant effect on surfactant properties. Surfactants with little branching in their alkyl chains generally show a good cleaning effect but relatively poor wetting characteristics, whereas more highly branched surfactants are good wetting agents but have unsatisfactory detergency performance. For compounds containing an equal number of carbon atoms in their hydrophobic chains, wetting power increases markedly as the hydrophilic groups move to the center of the chain or as branching increases. Simultaneously a decrease in adsorption and detergency performance occurs (Figs. 61 and 62).

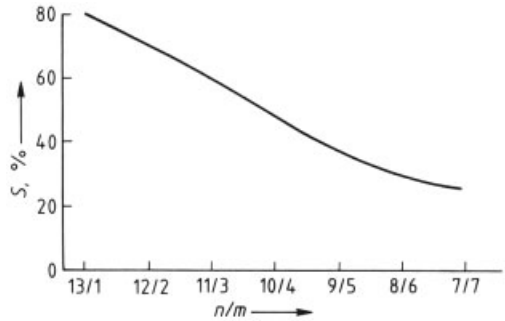

Figure 62. Decrease in soil removal S from soiled cotton as a function of increased branching in the hydrophobic residue [126]

Surfactant:

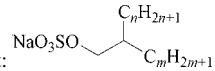

temperature: 90 °C; bath ratio: 1 : 12.5; water hardness: 16 °d (285 ppm); surfactant concentration: 2.91×10^{-3} mol/L

The changes with respect to adsorption, wetting, and detergency performance that result from varying the degree of branching are far more significant for ionic surfactants than for nonionic surfactants. In the case of anionic surfactants, losses in detergency performance caused by increased branching can be compensated for to some extent, provided the overall number of carbon atoms is increased appropriately.

Table 10. Surfactants of various ionic nature [125]

Surfactant	Formula	Electrolytic dissociation	Ionic nature
Alkyl poly(ethylene glycol) ethers	$RO-(CH_2-CH_2-O)_nH$	no	nonionic
Alkylsulfonates	$R-SO_3^-\ Na^+$	yes	anionic
Dialkyldimethylammonium chlorides	$\left[H_3C-N^+(R)_2-CH_3\right] Cl^-$	yes	cationic
Betaines	$R-N^+(CH_3)_2-CH_2-C(=O)O^-$		amphoteric

Household washing of textiles normally poses few situations requiring extraordinary wetting power. If problems arise, they can usually be overcome by increasing the wash time or the amount of detergent used. Most important is the effectiveness of the rolling-up process. Optical microscopy has revealed that oily and greasy soil tends to reside in more or less evenly distributed layers on the surface of fibers. These layers are gradually constricted by the action of a surfactant and its associated spreading pressure, with the layers ultimately being reduced to drops resting loosely on the fibers, which are then easily detached. Thus, oily residue as a fiber wetting agent is replaced by aqueous wash solution.

The number of types of surfactants suitable for use in laundry products has increased considerably in the past 50 years. The principal criteria for judging surfactant suitability apart from performance are toxicological and ecological characteristics. Cationic and nonionic surfactants have come to play an increasingly important role along with their anionic counterparts (cf. Table 11). Despite the wide choice of possibilities, however, only few surfactants account for the major share of the market, partly as a result of economic factors [120], [127].

Anionic surfactants are the most common ingredients in detergents designed for laundry, dishwashing, and general cleansing. Nonionic surfactants such as alcohol ethoxylates have acquired great importance during the last decades. Cationic surfactant use is largely restricted to fabric softeners because of the fundamental incompatibility of these materials with anionic surfactants and their poor cleaning efficiency. Amphoteric surfactants still lack a significant place in the market. Around the world a remarkable variability in the types of surfactants employed in products for similar purposes can be seen. The reasons are to be found in the variations in the kinds of fabric encountered worldwide, the diversity in washing machine technology, and different regional habits for fabric use and care (Fig. 63).

Wash technology has been the subject of major changes and developments during the past 50 years. The textile market has also changed, with synthetic fibers playing an increasingly large role. The wide variety of textile fibers and fiber blends and their differing characteristics have forced manufacturers to devise surfactants with a broad performance range. No single surfactant is capable of fulfilling all demands in an optimal way; consequently, the trend has been increasingly toward using surfactant blends, in which the characteristics of each component are intended to supplement those of the others. The widening scope of the demand for surfactants used in detergents not only relates to performance, but encompasses toxicological, ecological, and economic considerations as well. Surfactants suitable for detergent use are expected to demonstrate the following characteristics [129, 130]:

Specific adsorption
Excellent soil removal
Low sensitivity to water hardness
Good dispersion properties
Good soil antiredeposition capability
High solubility
Sufficient wetting power
Desirable foam characteristics
Neutral odor
Low intrinsic color
Sufficient storage stability
Good handling characteristics
Low toxicity to humans
Favorable environmental behavior
Assured raw material supply
Competitive costs

3.1.1. Anionic Surfactants

Most detergents contain larger amounts of anionic surfactants than nonionic surfactants. The following discussion deals with anionic surfactants that are in widespread use and/or have favorable characteristics. Anionic surfactants and their manufacturing processes are amply described in the literature [273].

Soap. Soap is no longer as important in many parts of the world as it was before the existence of mass-produced synthetic surfactants. Although soap powders for washing once contained as much as 40 % soap as their sole surfactant, powder detergents have since the 1950s been formulated with mixtures of far more effective surfactants in considerably smaller proportion. A further reason for the decreasing use

Table 11. Key surfactants

Structure		Chemical name	Acronym
Anionic surfactants			
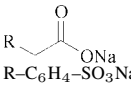	$R = C_{10-16}$	soaps	
$R–C_6H_4–SO_3Na$	$R = C_{10-13}$	alkylbenzenesulfonates	LAS
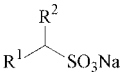	$R^1 + R^2 = C_{11-17}$	alkanesulfonates	SAS
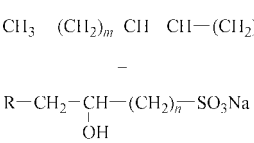	$n + m = 9–15$		
	$n = 0,1,2\ldots m = 1,2,3\ldots$		
	$R = C_{7-13}$ $x = 1,2,3$	α-olefinsulfonates	AOS
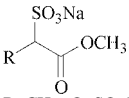	$R = C_{14-16}$	α-sulfo fatty acid methyl esters	MES
$R–CH_2–O–SO_3Na$	$R = C_{11-17}$	alcohol sulfates, alkyl sulfates (linear and branched)	AS
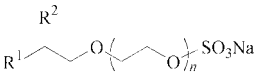	a) $R^1 = H$	alkyl ether sulfates	AES
	$R^2 = C_{10-12}$	a) fatty alcohol ether sulfates	
	b) $R^1 + R^2 = C_{11-13}$	b) oxo alcohol ether sulfates	
	$R^1 = H, C_1, C_2\ldots$		
	$n = 1–4$		
Cationic surfactants			
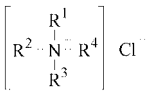	$R^1, R^2 = C_{16-18}$	quaternary ammonium compounds	
	$R^3, R^4 = C_1$	tetraalkylammonium chloride	
Nonionic surfactants			
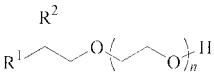		alcohol ethoxylates	AE
	a) $R^1 = H$ $R^2 = C_{6-16}$	a) fatty alcohol ethoxylates	
	b) $R^1 + R^2 = C_{7-13}$		
	$R^1 = H, C_1, C_2\ldots$	b) oxo alcohol ethoxylates	
	$n = 3–15$		
$R–C_6H_4–O–(CH_2–CH_2–O)_nH$	$R = C_{8-12}$ $n = 5–10$	alkylphenol ethoxylates	APE
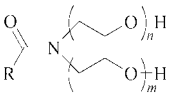	$R = C_{11-17}$		
	$n = 1, 2$ $m = 0, 1$	fatty acid alkanolamides	FAA
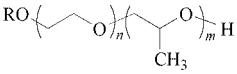	$R = C_{8-18}$		
	$n = 3–6$ $m = 3–6$	alcohol alkoxylates (EO/PO adducts)	
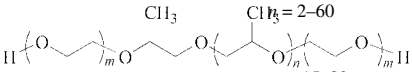	$n = 2–60$	ethylene oxide–propylene oxide	EPE
	$m = 15–80$	block polymers	
	$R = C_{12-18}$	alkyldimethylamine oxides	

Table 11. Continued

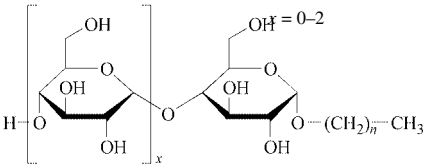	$x = 0–2$		
	$n = 11–17$	alkyl polyglycosides	APG
	$n = 10–16$	fatty acid *N*-methylglucamides	NMG
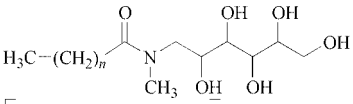			
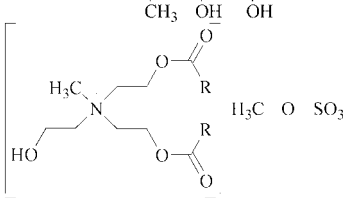		*N*-methyl-*N*,*N*-bis[2-($C_{16/18}$-acyloxy)ethyl]-*N*-(2-hydroxy-ethyl)ammonium-methosulfate	EQ
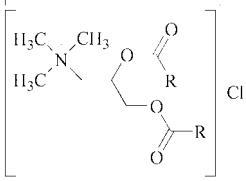		*N*,*N*,*N*-trimethyl-*N*-[1,2-dl($C_{16/18}$-acyloxy)propyl]-ammonium-chloride	DEQ
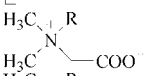	$R = C_{12-18}$	alkyl betaines	
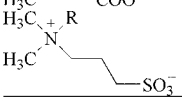	$R = C_{12-18}$	alkylsulfobetaines	

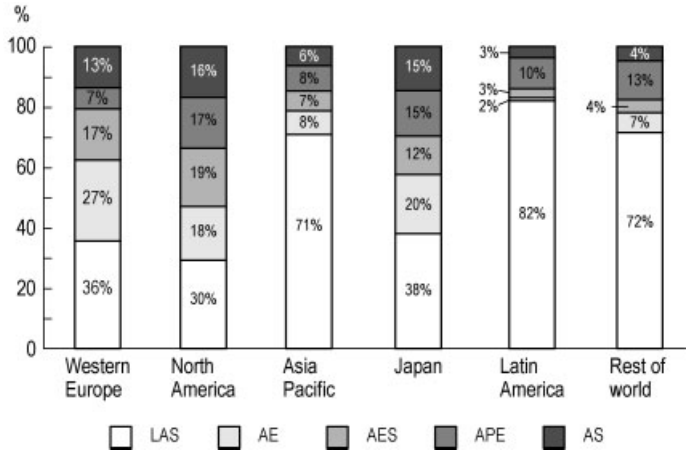

Figure 63. 1995 Global use pattern of major surfactants [128]

of soap in laundry detergents is its sensitivity to water hardness, manifested through inactivation due to the formation of lime soap, which tends to accumulate on fabrics and washing machine parts. Such accumulation reduces the absorbency of fabrics and their permeability to air, and eventually through "aging" causes laundry to become discolored and to develop malodors.

The primary function that remains for soap currently is as a foam regulator in laundry detergents in Europe. Nonetheless, soap has remained worldwide the largest surfactant by volume. It is still the surfactant of choice in many countries with low gross national product (cf. Chap. 9 and cf. Section 3.4.3).

Alkylbenzenesulfonates (LAS and TPS). Until the mid-1960s, this largest class of synthetic surfactants was most prominently represented by tetrapropylenebenzenesulfonate (TPS):

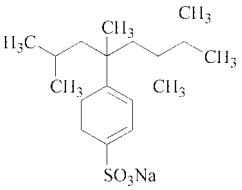

In the 1950s TPS had largely replaced soap as an active ingredient in laundry detergents in Europe, the USA, and Japan. It was later found, however, that the branched side-chain present in TPS prevents the compound from undergoing sufficient biodegradation. Thus, means were developed to replace it by more readily biodegradable straight-chain homologues. Since that time, favorable economic circumstances and good performance characteristics have permitted straight-chain or linear alkylbenzenesulfonates (LAS) to take the lead among laundry detergent surfactants in Europe, the Americas, and Asia. Nevertheless, a few countries remain in which TPS continues to be used in detergents (see Section 10.3.1). The large-scale manufacturing methods for LAS have been reviewed [137].

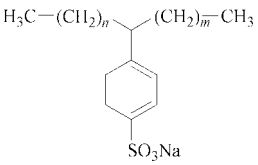

$n + m = 7–10$ LAS

Apart from their very good detergency performance, LAS have interesting foaming characteristics, which are of great significance to their use in detergents. Their foaming ability is high, and the foam that is produced is readily stabilized by foam stabilizers, an important factor in many parts of the world. At the same time, however, LAS can be controlled easily by foam regulators, and this is significant with respect to detergents for the European market, where horizontal-axis drum-type washing machines are common.

As a result of their high solubility, LAS are also frequently employed in formulations for liquid detergents. However, LAS are sensitive to water hardness: the detergency performance of LAS diminishes as the hardness of the water increases. The relationship between water hardness and performance for a series of surfactants in the absence of builders or sequestrants is well demonstrated by soil removal from wool, as illustrated in Figure 64.

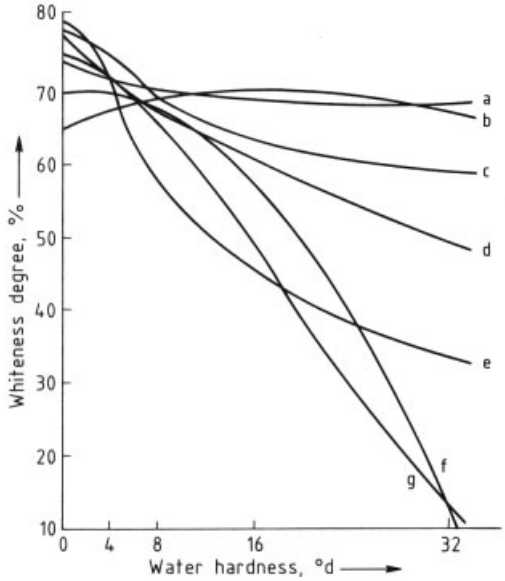

Figure 64. Detergency performance on wool by various surfactants as a function of water hardness [131]
Time: 15 min; temperature: 30 °C; bath ratio: 1 : 50; concentration: 0.5 g/L surfactant + 1.5 g/L sodium sulfate
a) Nonylphenol 9 EO; b) C_{12-14} Alcohol 2 EO sulfates; c) C_{15-18} α-Olefinsulfonates; d) C_{16-18} α-Sulfo fatty acid esters; e) C_{12-18} Alcohol sulfates; f) C_{10-13} Alkylbenzenesulfonates; g) C_{13-18} Alkanesulfonates

The decline in detergency performance with increasing water hardness is most dramatic with soap. Sensitivity to water hardness largely disappears in built formulations (zeolite A, triphosphate, soda ash, etc.) because of sequestration,

binding, or ion exchange of the water hardness ions (cf. Chap. 2, Fig. 17). Figure 65 illustrates wool wash performance for various readily accessible anionic surfactants in the absence of complexing agents, plotted as a function of surfactant concentration. It can be seen that products with a lower sensitivity to hardness display only a slight advantage.

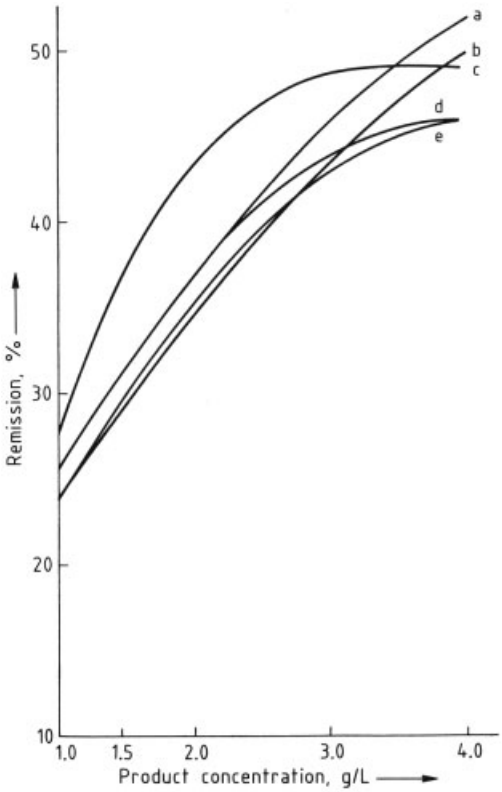

Figure 65. Detergency performance on wool of various synthetic anionic surfactants as a function of concentration [129]
Product: 25 % surfactant + 75 % sodium sulfate; temperature: 30 °C; wash time: 15 min; bath ratio: 1 : 30; water hardness: 16 °d (285 ppm $CaCO_3$)
a) C_{12-14} Alcohol 2 EO sulfates; b) C_{15-18} α-Olefinsulfonates; c) C_{16-18} α-Sulfo fatty acid esters; d) C_{13-18} Alkanesulfonates; e) Alkylbenzenesulfonates

Secondary Alkanesulfonates (SAS).

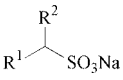

$R^1 + R^2 = C_{11-17}$

Sodium alkanesulfonates (SAS) have been known as commercial surfactants since the 1940s. Their large-scale production began in the late 1960s. Sodium alkanesulfonates are still valued as specialty anionic surfactants for consumer products. Their 1998 demand amounted to only 60×10^3 t. SAS feature high solubility, fast wetting properties, chemical stability to alkali, acids, and strong oxidants including chlorine. Sodium alkanesulfonates are produced by photochemical sulfoxidation or sulfochlorination of suitable C_{12}–C_{18} paraffin cuts. Manufacturing, properties, applications, and characteristics of SAS have been reviewed [132 – 135]. Sodium alkanesulfonates are the compounds that closely resemble LAS in terms of solubility, solubilizing properties, wetting power and other detergency properties. Therefore, SAS can largely be substituted for LAS in most formulations.

Alkanesulfonates are completely insensitive to hydrolysis even at extreme pH values, a result of the presence of the stable carbon–sulfur bond.

The water hardness sensitivity and foaming characteristics of SAS largely resemble those of LAS, as discussed above.

α-Olefinsulfonates (AOS).

R^1–CH_2–CH=CH–$(CH_2)_n$–SO_3Na Alkenesulfonates

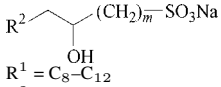

Hydroxyalkanesulfonates

$R^1 = C_8$–C_{12} $n = 1, 2, 3$
$R^2 = C_7$–C_{13} $m = 1, 2, 3$

α-Olefinsulfonates (AOS) are produced commercially starting from α-olefins. Alkaline hydrolysis of the sultone intermediate results in ca. 60–65 % alkenesulfonates and ca. 35–40 % hydroxyalkanesulfonates. Because of the use of olefinic precursors, these mixtures are customarily called α-olefinsulfonates.

In contrast to LAS and SAS, AOS show little sensitivity to water hardness (cf. Fig. 64). This apparent advantage is only of significance in a few very special applications, however. Depending on chain length, AOS may cause foaming problems in drum-type washing machines, which requires the addition of special foam regulators. Since this problem does not arise with washing machines of the type used in Japan, AOS have long been an important component of some Japanese detergents.

In addition to α-olefinsulfonates, sulfonates prepared from olefins with internal or central double bonds and from vinylidene olefins also exist. These types of olefinsulfonates are unsuitable for detergent use, however, because of their poor performance characteristics. In Figure 66, a C_{15-18} α-olefinsulfonate is compared with a C_{15-18} internal olefinsulfonate, the inferior performance of which is clearly apparent. The reason for the difference is the fact that AOS have their hydrophilic group in a terminal position, whereas with the internal olefinsulfonates, additional isomers can arise in which hydrophilic groups are distributed throughout the entire hydrophobic chain. This has the same consequences as branching (cf. Section 3.1). However, olefinsulfonates prepared from internal olefins generally show very good textile wetting characteristics.

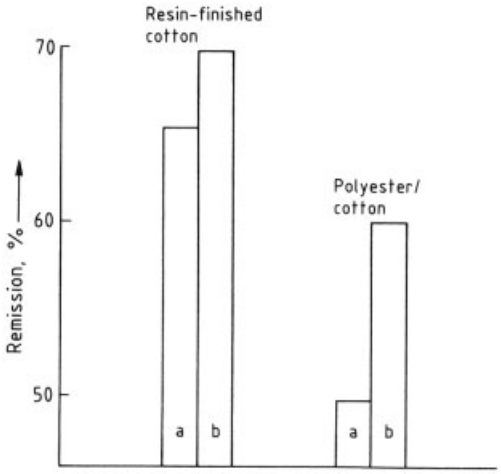

Figure 66. Performance of various olefinsulfonates [136] Water hardness: 16 °d (285 ppm $CaCO_3$); wash time: 30 min; wash temperature: 60 °C
a) 0.75 g/L C_{15-18} Internal olefinsulfonates + 2.0 g/L sodium triphosphate; b) 0.75 g/L C_{15-18} α-Olefinsulfonates + 2.0 g/L sodium triphosphate

α-Sulfo Fatty Acid Esters (MES)

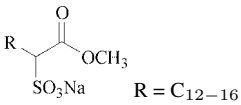

α-Sulfo fatty acid methyl esters

Another class of anionic surfactants is the α-sulfo fatty acid esters, particularly the methyl derivatives. This group of surfactants is also called methyl ester sulfonates (MES). Methyl ester sulfonates are derived from a variety of methyl ester feedstocks from renewable resources, such as coconut, palm kernel, palm stearin, beef tallow, lard, and soy. Good detergency performance is attained with products having rather long hydrophobic residues (e.g., palmitic and stearic acid derivatives) [137]. The sensitivity of MES to water hardness is small relative to that of LAS and SAS, more closely resembling that of AOS. One of the interesting detergency properties of α-sulfo fatty acid methyl esters is their exceptional dispersion power with respect to lime soap.

Special attention is to be paid to the manufacturing of finished laundry products based on MES. Insufficient storage stability of MES has so far impeded their large-scale use in detergents; they have only been used in a few Japanese detergents [138].

Alkyl Sulfates (AS).

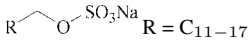

Alkyl sulfates (AS), also known as alcohol sulfates, achieved prominence as early as the 1930s in detergents for delicate fabrics in Germany and the USA and as components of textile auxiliaries. Their availability resulted from the development by Schrauth to produce primary fatty alcohols by high-pressure hydrogenation of fatty acids and their methyl esters. Alkyl sulfates are produced either from natural fatty alcohols derived from palm oil, palm kernel oil, or coconut oil, or from oxo alcohols, i.e., of petrochemical origin [127]. They are characterized by desirable detergency properties, and have found increasing application not only in specialty products, but also in heavy-duty detergents, especially in those concentrated products produced by nontower technology (cf. Section 6.2.3).

Alcohol sulfates derived from natural, renewable feedstocks have been playing a steadily increasing role as surfactants in laundry detergents since the 1980s. They have found increasing application at the expense of LAS [119, 127, 139 – 142].

Alkyl Ether Sulfates (AESs).

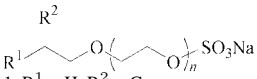

Alkyl ether sulfates

1. $R^1 = H, R^2 = C_{10-12}$ — Fatty alcohol ether sulfates
2. $R^1 + R^2 = C_{11-13}$ — Oxo alcohol ether sulfates
   $R^1 = H, C_1, C_2 \ldots$

$n = 1–2$

Alkyl ether sulfates (AES), also known as alcohol ether sulfates, are obtained by ethoxylation and subsequent sulfation of alcohols derived from natural feedstocks or of synthetic alcohols. They exhibit the following unique characteristics relative to alkyl sulfates:

Low sensitivity to water hardness (cf. Fig. 64)
High solubility
Good storage stability at low temperature in liquid formulations

Commercial AESs consist of alkyl ether sulfates and alkyl sulfates as the main components. Unsulfated alcohols, alcohol ethoxylates, inorganic salts, and poly(ethylene oxide) sulfates are contained as byproducts [127]. Manufacturing, properties, and analysis of AES have been reviewed [143].

Alcohol ether sulfates that are least sensitive to water hardness, e.g., sodium C_{12-14} *n*-alkyl diethylene glycol ether sulfates, actually demonstrate increased detergency performance (e.g., on wool) as the hardness increases. This is a result of the positive electrolyte effects attributable to calcium/magnesium ions. Addition of sodium sulfate produces a slight improvement of detergency performance only in regions of low water hardness. However, detergency performance declines in the presence of sodium triphosphate due to calcium/magnesium sequestration (cf. Fig. 67).

Binding of alkaline earth ions can occur not only through complexation, but also as a result of ion exchange. Calcium ions can be exchanged particularly effectively and with favorable kinetics by using zeolite A. The properties discussed above in the context of calcium sequestration can be equally well described in terms of ion exchange.

From the foregoing, it follows that if one were to undertake the development of, for example, a wool detergent, an important consideration would be the careful choice of supplemental salts; i.e., the presence of complexing agents or ion exchangers is not always advantageous in a detergent formulation containing surfactants if these are insensitive to water hardness. On the other hand, a surfactant that is sensitive to water hardness is not necessarily inferior, provided it is combined with the proper complexing agents or ion exchangers. The situation can be somewhat more complicated, however, if a detergent has to be developed for fibers whose nature requires the presence of complexing agents for proper washing.

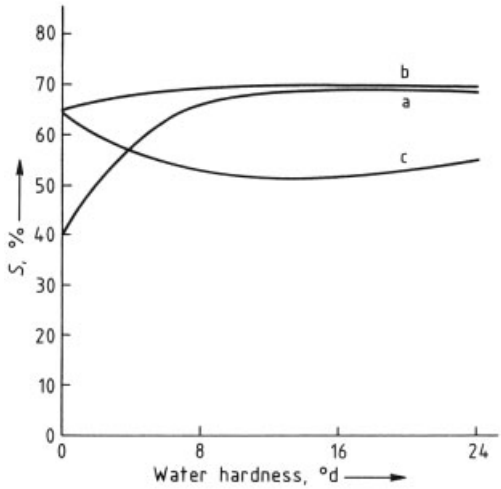

Figure 67. Soil removal *S* from wool as a function of water hardness at 30 °C with 0.5 g/L sodium C_{12-14} *n*-alkyl diethylene glycol ether sulfates containing no added electrolyte

a) with 1.5 g/L sodium sulfate
b) and with 1.5 g/L sodium triphosphate
c) [46]

Alkyl ether sulfates are very intensively foaming compounds that are well suited for usein high-foam detergents for vertical-axis washing machines, but are less directly applicable to detergents for horizontal-axis drum-type machines. Because of their specific properties, alkyl ether sulfates are preferred constituents of detergents for delicate or wool washables, as well as foam baths, hair shampoos, and manual dishwashing agents. The optimal carbon chain length has been established to be C_{12-14} with ca. 2 mol of ethylene oxide.

Alcohol ether sulfates have achieved some importance in the U.S. and Japanese markets.

This is because their critical micelle concentration is considerably lower than that for LAS, resulting in very satisfactory washing power even at the low detergent concentrations typical in these countries. In Europe, the use of alcohol ether sulfates has so far been largely restricted to specialty detergents.

Alkyl ether sulfates were formerly prepared exclusively by ethoxylation and sulfation of native fatty alcohols. For a long time, synthetic alcohols have also been used, particularly the partially branched oxo alcohols. Oxo alcohol ether sulfates have good wetting properties comparable to those of straight-chain native fatty alcohol ether sulfates, whereas their detergency performance is inferior to the latter [144].

3.1.2. Nonionic Surfactants

3.1.2.1. Alcohol Ethoxylates (AE)

Nonionic surfactants of the alcohol ethoxylate (AE) type do not dissociate in aqueous solution. Some of their properties can be singled out for special attention:

The absence of electrostatic interactions
Behavior with respect to electrolytes
The possibility of favorable adjustment of hydrophilic–lipophilic balance values (HLB)
Anomalous solubility in water

Adsorption phenomena involving nonionic surfactants can be explained on the basis of hydrophobic interactions, in some cases coupled with steric effects. Electrolytes have no direct influence on adsorption with nonionic surfactants. Nonionic surfactant detergency performance is adversely affected by the presence of polyvalent cations. These cations cause reduced negative ζ-potentials of fiber surface and soil, leading to reduced repulsion and correspondingly poor soil removal.

An important advantage of nonionic surfactants of the alcohol ethoxylate type is the fact that a proper balance (HLB) can be easily achieved between the hydrophobic and hydrophilic portions of the nonionic surfactants. For example, the hydrophilic portion of the molecule can be extended gradually by stepwise addition of ethylene oxide moieties. This leads to a stepwise increase in hydration and corresponding successive increase in solubility. On the other hand, with ionic surfactants, the presence of even one ionic group makes such a strong contribution to hydrophilic character that the introduction of further ionic groups totally eliminates the possibility of any equilibrium relationship with respect to the hydrophobic portion. The result is a rapid deterioration of typical surfactant characteristics. For a surfactant bearing two strong ionic hydrophilic groups to show washing activity, a considerably longer alkyl chain ($> C_{20}$) would be necessary.

Nonionic surfactants with a given hydrophobic residue can be adjusted to have optimal properties for various substrates with respect to adsorption and detergency performance simply by changing the degree of ethoxylation. Performance shows an initial increase with an increasing degree of ethoxylation, but a point is then reached after which it declines markedly (Fig. 68). Textile wetting power often decreases only at very high degrees of ethoxylation, whereas the wetting power for hard hydrophobic surfaces continues to climb as the number of ethoxy groups increases.

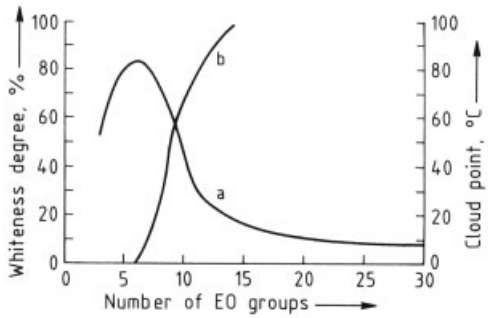

Figure 68. Detergency performance on wool and cloud point of C_{13-15} oxo alcohols as a function of degree of ethoxylation [129]
a) Detergency performance on wool (temperature: 30 °C; water hardness: 16 °d (285 ppm $CaCO_3$); bath ratio: 1 : 30; time: 15 min; concentration: 0.75 g/L surfactant +2.25 g/L sodium sulfate); b) Cloud point (surfactant concentration 10 g/L)

Alcohol ethoxylates show a solubility anomaly: when they are heated in aqueous solution, suddenly turbidity appears, usually at a relatively precise temperature. This is caused by a separation of the alcohol ethoxylate solution into two phases, one with a high water content and one with a low water content. The

corresponding characteristic temperature for a given surfactant is known as its cloud point. The cloud point moves to a higher temperature as the number of ethoxy groups increases. If the cloud point is not greatly exceeded, then the largely aqueous phases and largely surfactant phases form an emulsion. For a given surfactant, adsorption (hence, detergency performance) decreases when the cloud point is surpassed to a significant extent. The main reason for this behavior is the reduced solubility of the alcohol ethoxylates, which are then expelled from the aqueous phase. Nevertheless, nonionic surfactants with a cloud point somewhat below the application temperature commonly show better performance than those whose cloud point is higher (Fig. 68). Thus, the wash temperature is a significant factor in determining an optimal degree of ethoxylation (Table 12). The table shows that the best detergency performance is obtained when the temperature is maintained near the cloud point.

The cloud point can be greatly reduced by addition of several grams of electrolytes per liter, depending on the surfactant. However, everything said in this section is strictly applicable only to systems comprised of pure nonionic surfactants. For binary mixtures of nonionic and ionic surfactants, it is important to recognize that even a small amount of ionic surfactant can cause the cloud point to rise more or less dramatically.

The share of nonionic surfactants in overall surfactant production and use has been increasing since the 1970s. The major contributors to this increase have been fatty alcohol, oxo alcohol, and secondary alcohol ethoxylates, all of which are obtained by reaction of the corresponding alcohols with ethylene oxide. By varying the carbon chain length and the degree of ethoxylation, these nonionic surfactants can be tailor-made with respect to the washing temperature. For these reasons, the increased use of nonionic surfactants has been partly concomitant with the trend to wash at lower temperature and with changes in the production shares of different fibers [119, 145].

The reasons for the increased use of nonionic surfactants are found in their favorable detergency properties, particularly with respect to synthetics, and the decrease in wash temperatures observed since the 1970s.

Favorable detergency properties of nonionic surfactants derive largely from the following factors:

Low critical micelle concentration (c_M)
Very good detergency performance
Excellent soil antiredeposition characteristics with synthetic fibers

The low c_M values of nonionic surfactants indicate that they display high detergency performance even at relatively low concentrations. Table 13 provides data illustrating the low c_M of these compounds relative to that of anionic surfactants.

Table 13. Critical micelle concentration c_M of various surfactants [131]

Surfactant	c_M, g/L
LAS (C_{10-13} alkyl)	0.65
C_{12-17} Alkanesulfonates	0.35
C_{15-18} α-Olefinsulfonates	0.30
C_{12-14} Fatty alcohol 2 EO sulfates	0.30
Nonylphenol 9 EO	0.049
Oleyl/cetyl alcohol 10 EO *	0.035

* Iodine number 45.

3.1.2.2. Alkylphenol Ethoxylates (APE)

Alkylphenol ethoxylates (APE), are based on *p*-octyl-, nonyl-, and dodecylphenol poly(ethylene glycol) ethers. They achieved early success due to their exceptional detergency properties, particularly their oil and fat removal characteristics.

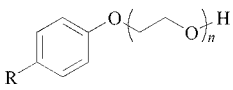

$R = C_{8-12}$ $n = 5–10$

The usage of APE has greatly declined, however, especially in Europe since 1986, due to their environmental characteristics, particularly the extent of their biodegradability and the fish toxicity of certain metabolites resulting from partial biodegradation (see Chap. 10).

3.1.2.3. Fatty Acid Alkanolamides (FAA)

Fatty acid alkanolamides are ethanolamides of fatty acids with the following structure:

Table 12. Detergency performance (in % remission) of alcohol ethoxylates with comparable cloud points [131]

Surfactant *	Cloud point, °C	Cotton 60 °C	Cotton 90 °C	Permanent-press cotton 60 °C	Permanent-press cotton 90 °C	Polyester/cotton 60 °C	Polyester/cotton 90 °C
C_{11-15} *sec*-Alcohol 9 EO	59	71	53	69	55	65	48
C_{9-11} Oxo alcohol 7 EO	61	57	44	66	52	53	43
Oleyl/cetyl alcohol 10 EO **	89	60	68	70	72	57	67
C_{13-15} Oxo alcohol 11 EO	88	54	69	67	69	53	58

* Surfactant concentration: 0.75 g/L; water hardness: 16 °d (285 ppm); wash time: 30 min.
** Iodine number 45.

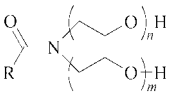

$R = C_{11-17}$ $n = 1, 2$ $m = 0, 1$

Fatty acid alkanolamides find only little application in laundry detergents. Their most important feature is foam boosting, i.e., adding desired stability to the foam produced by detergents prone to heavy foaming. This property is not desirable for horizontal-axis drum-type washing machines employed, e.g., in Europe. Nevertheless, small amounts of FAA as cosurfactants are capable of enhancing the soil removal properties of the classical detergent components at low washing temperatures.

3.1.2.4. Alkylamine Oxides

The alkylamine oxides are produced by oxidation of tertiary amines with hydrogen peroxide. They are compounds that exhibit cationic behavior in acidic conditions (pH < 3), but they behave as nonionic surfactants under neutral or alkaline conditions. For this reason they are included in the nonionic surfactant category.

$R = C_{12-16}$

Compounds in this class have been known since 1934 and were described as detergent components in a patent issued to IG-Farbenfabriken. Combinations of alkylbenzenesulfonates and specific amine oxides are reputed to be especially gentle to the skin. Despite good detergency properties, however, they are rarely included in laundry detergent fromulations. The reasons for this include high cost, low thermal stability, and high foam stability.

3.1.2.5. *N*-Methylglucamides (NMG)

N-Methylglucamides are a new type of nonionic surfactants, that has been introduced in detergents in the 1990s. *N*-Methylglucamides derive their hydrophilic character from the hydroxyl groups of sugar or starch glucose moieties. These surfactants are obtained by reacting sugars with methylamine and subsequent acylation with fatty acid. They have increasingly been used as co-surfactants in powder and liquid detergent formulations.

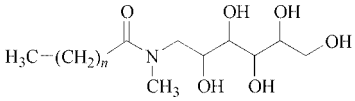

3.1.2.6. Alkylpolyglycosides (APG)

Alkylpolyglycosides are produced by the dehydration–condensation reaction of alcohols with glucose. Alkylpolyglycosides have distinct lathering characteristics, especially in combination with anionic surfactants. Due to their good foaming properties APG are predominantly used in dishwashing detergents, liquid detergents, and specialty detergents for fine fabrics.

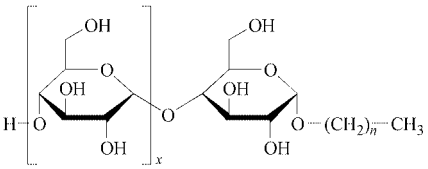

Excellent skin compatibility is one of the outstanding characteristics of alkylpolyglycosides. Many APG compounds with special properties can be synthesized by varying the alkyl chain length and the degree of glucose polymerization [146]. Since APG are completely based on natural resources, they biodegrade easily and rapidly to carbon dioxide and water only.

3.1.3. Cationic Surfactants

Long-chain cationic surfactants such as distearyldimethylammonium chloride (DSDMAC) exhibit extraordinarily high sorption power with respect to a wide variety of surfaces [147, 148, 151, 152]. They are very strongly absorbed to the surface of natural fibers, such as cotton, wool, and linen. Adsorption on synthetic fibers is much less pronounced. Figure 69 shows this behavior for several textile fibers. Adsorption rises steeply at low surfactant concentrations, followed by rapid saturation as the concentration increases. This behavior suggests complete coverage of boundary surfaces.

At the same time, cationic surfactants display behavior opposite to that of anionic surfactants as regards charge relationships on solids. Since the surfactant molecules bear a positive charge, their adsorption reduces the negative ζ-potential of solids present in aqueous solution, thereby reducing mutual repulsions, including that between soil and fibers. Use of higher surfactant concentrations causes charge reversal; thus, solid particles become positively charged, resulting again in repulsion. Soil removal can be achieved if adequate amounts of cationic surfactants are present and if their alkyl chains are somewhat longer than those of comparable anionic surfactants. This fact has little practical significance, however, since the subsequent rinse and dilution processes cause charge reversal in the direction of negative ζ-potentials, whereby a large fraction of the previously removed soil is once more attracted to the fibers. Therefore, cationic surfactants are employed only for the purpose of achieving certain special effects, which include applications in rinse-cycle fabric softeners, antistatic agents, and microbicides.

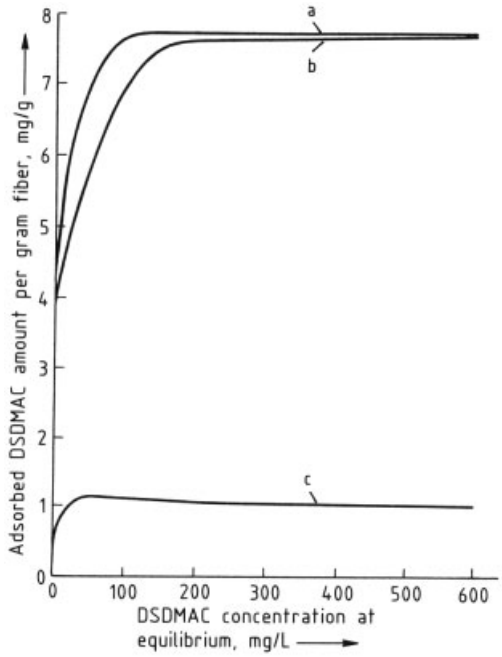

Figure 69. Adsorption isotherms of distearyldimethylammonium chloride (DSDMAC) on wool (a), cotton (b), and polyacrylonitrile (c) [147]
Temperature: 23 °C; time: 20 min; bath ratio: 1 : 10

Mixtures made up of equivalent amounts of anionic surfactant and cationic surfactant remain virtually unadsorbed on surfaces and thus display no washing effect. Reactions between anionic and cationic surfactants produce neutral salts with extremely low water solubility. Regarding the washing process, these salts behave like an additional load of greasy soil. On the other hand, addition of small amounts of certain specific cationic surfactants to an anionic surfactant — or even a nonionic surfactant — can enhance detergency performance.

Nonionic surfactants are more tolerant of the presence of cationic surfactants than anionic surfactants. Mixtures of the two are sometimes used in specialty detergents intended to have a wash-cycle fabric-softening effect. In such cases, it must be taken into account that adsorption of the cationic surfactant can be greatly reduced by the

presence of the nonionic surfactant, depending on the concentration of the latter, a phenomenon that can have a negative influence on the fabric-softening characteristics (Fig. 70).

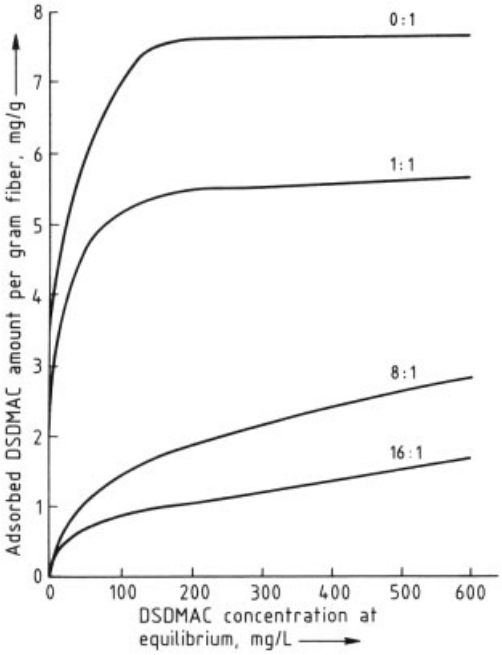

Figure 70. Adsorption isotherms of distearyldimethylammonium chloride (DSDMAC) on cotton as a function of the alkyl polyglycol ether : DSDMAC ratio at equilibrium [147] Temperature: 23 °C; time: 20 min; bath ratio: 1 : 10

Dialkyldimethylammonium Chlorides. The first surfactant developed in this category was distearyldimethylammonium chloride (DSDMAC), introduced in 1949 as a fabric softener for cotton diapers and presented to the U.S. market a year later as a laundry rinse-cycle fabric softener.

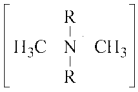

R = C_{16-18}

Only in the mid-1960s did these surfactants begin to have a major impact as laundry aftertreatment aids on the U.S. and Western European markets (especially in the Federal Republic of Germany).

Imidazolinium salts such as 1-(alkylamidoethyl)-2-alkyl-3-methylimidazolinium methylsulfate have achieved a place as rinse-cycle softening agents, although not nearly as significant as that of DSDMAC.

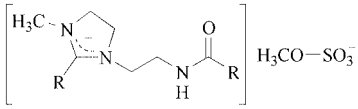

R = C_{16-18}

Alkyldimethylbenzylammonium Chlorides. Compounds of the alkyldimethylbenzylammonium chloride type show only limited fabric softening character, but they have been used in laundry disinfecting agents as a result of their activity toward gram-positive and gram-negative bacteria.

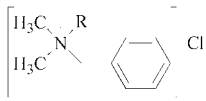

R = C_{8-18}

The high adsorption capability of alkyldimethylbenzylammonium chloride also leads to applications as antistatic agents in laundry aftertreatment products.

Esterquats (EQ). In the new generation of fabric softeners that came up in the 1980s/1990s, DSDMAC has been largely replaced by esterquats, e.g., products with the following structures [149, 150]:

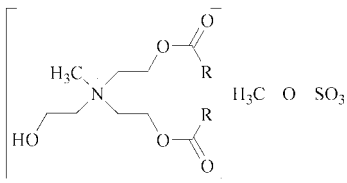

N-Methyl-*N*,*N*-bis[2-($C_{16/18}$-acyloxy)ethyl]-*N*-(2-hydroxyethyl)ammonium-methosulfate EQ (**E**ster**q**uat)

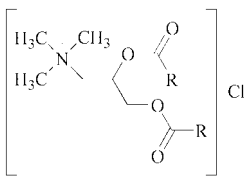

N,*N*,*N*-Trimethyl-*N*-[1,2-

di($C_{16/18}$-acyloxy)propyl] ammonium chloride DEQ (**Die**ster**q**uat)

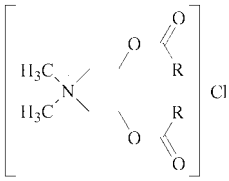

Diethyl**e**ster **dim**ethyl **a**mmonium **c**hloride (DEEDMAC)R = Tallow fatty acid

Due to their ester bonds, which are potential breaking points, esterquats are readily biodegradable in contrast to DSDMAC, which is only poorly biodegradable (see Chap. 10). Esterquats have favorable ecotoxicological and toxicological properties. All available data proves that esterquats present no hazard for living organisms.

Esterquats have similar physicochemical and chemical properties as DSDMAC [149, 150]. It may be assumed that esterquats exhibit adsorption characteristics that are very similar to those of DSDMAC, as shown in Figure 69.

3.1.4. Amphoteric Surfactants

Compounds of the alkylbetaine or alkylsulfobetaine type possess both anionic and cationic groups in the same molecule even in aqueous solution. Despite what could be seen in some respects as excellent detergency properties, these surfactants are only rarely employed in laundry detergents, primarily for cost reasons. They are mainly applied in manual dishwashing products.

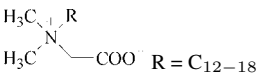

R = C_{12-18} Alkylbetaines

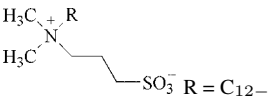

R = C_{12-18} Alkylsulfobetaines

3.2. Builders

Detergent builders play a central role in the course of the washing process [153]. Their function is largely that of supporting detergent action and of water softening, i.e., eliminating calcium and magnesium ions, which arise from the water and from soil.

The category of builders is predomintantly comprised of several types of materials: specific precipitating alkaline materials such as sodium carbonate and sodium silicate; complexing agents like sodium triphosphate or nitrilotriacetic acid (NTA); and ion exchangers, such as water-soluble polycarboxylic acids and zeolites (e.g., zeolite A).

Detergent builders must fulfill a number of criteria [46]:

1) Elimination of alkaline earth ions originating from
   - water
   - textiles
   - soil
2) Soil and stain removal
   - high specific detergency performance for particulate soil and fats
   - distinct detergency performance on specific textile fibers
   - enhancement of surfactant properties
   - dispersion of soil in detergent solutions
   - favorable influence on foam characteristics
3) Multiple wash cycle performance
   - good soil antiredeposition capability
   - prevention of lime deposits on textiles and washing machine parts
   - anticorrosion properties
4) Handling properties
   - sufficient chemical stability
   - no hygroscopicity
   - good color and odor qualities
   - compatibility with other detergent ingredients
   - good storage stability
   - assured raw material basis
5) Human toxicological safety assurance
6) Environmental properties
   - response to deactivation by biological degradation, adsorption, or other mechanisms

no adverse effects on the biological systems found in sewage plants and surface water
no uncontrolled accumulation
no heavy-metal ion remobilization
no eutrophication
no detrimental effects on drinking water quality

7) Cost effectiveness

3.2.1. Alkalies

Alkalies such as potash and soda ash have been used to enhance the washing effectiveness of water since antiquity. Their activity is based on the fact that soil and fibers become more negatively charged as the pH value increases, resulting in increased mutual repulsion. Alkali also precipitates ions that contribute to water hardness.

At the beginning of the 20th century, the principal ingredients (apart from soap) of all detergents were soda ash and silicate, which often made up nearly 50 % of the formulation of a powder detergent. These detergent ingredients were partly replaced during the 1930s by sodium monophosphate and diphosphate. All these builders removed water hardness by precipitation.

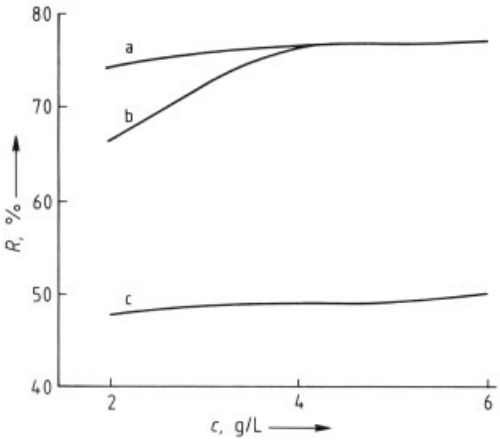

Figure 71. Comparison of soil removal (given as remission *R* in %) on permanent-press cotton at 0 °d and 90 °C for sodium triphosphate- and sodium carbonate-containing detergents, formulated on the basis of
a) 40 % Sodium triphosphate; b) 20 % Sodium triphosphate + 20 % sodium sulfate; c) 40 % Sodium carbonate

Modern builders no longer precipitate hardness; instead, water hardness is removed by complexation (sequestration) or ion exchange. Even commercial laundries using soft water have converted to detergents that are low in soda ash and contain the complexing agent sodium triphosphate. The latter has many valuable properties, among which are better detergency performance than soda ash (Fig. 71) [154].

3.2.2. Complexing Agents

Soda ash induces precipitation of calcium and magnesium salts from wash water. This can lead to formation of lime deposits on both laundry and washing machine parts. By contrast, sequestering agents form stable, water-soluble complexes with alkaline-earth ions, as well as with traces of heavy-metal present in water. Often the resulting complexes are chelates (Fig. 72). Sodium triphosphate forms a stable water-soluble complex with calcium when their stoichiometric ratio is 1 : 1 (Fig. 72). However, when a substoichiometric quantity of sodium triphosphate is present, the water-insoluble dicalcium triphosphate is formed, which precipitates on textile fibers and washing machine parts. Its precipitation can be impeded by adding small amounts of hydroxyethanediphosphonate and/or special polycarboxylates to the detergent formulation [167].

Temperature and complexing agent concentration are generally the decisive factors in successful elimination of polyvalent metal ions. Table 14 shows the calcium binding capacity of various sequestering agents as a function of temperature. Calcium binding capacity is here to be viewed as a quantitative measure of the stoichiometry of the resulting complexes. For most sequestering agents, this capacity decreases markedly with increasing temperature.

Regarding stability, the data presented show only that the stability constants (which are a function of the method of study employed, e.g., the dissolving power for freshly precipitated calcium carbonate) exceed a specific value that is dependent on the solubility of the calcium carbonate. Extremely high stability constants or stabilities, which would result in very low calcium ion concentrations, are not required; indeed, they are normally undesirable. It is important that

Table 14. Calcium binding capacity of selected sequestering agents [46]

Structure	Chemical name	Calcium binding capacity, mg CaO/g	
		20 °C	90 °C
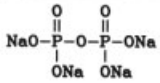	sodium diphosphate	114	28
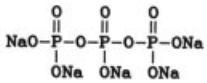	sodium triphosphate	158	113
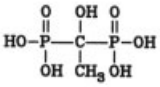	1-hydroxyethane-1,1-diphosphonic acid	394	378
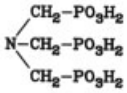	nitrilotrimethylenephosphonic acid	224	224
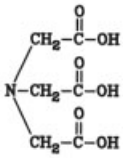	nitrilotriacetic acid	285	202
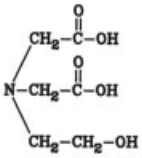	*N*-(2-hydroxyethyl)iminodiacetic acid	145	91
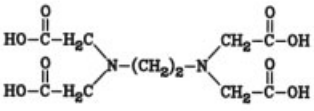	ethylenediaminetetraacetic acid	219	154
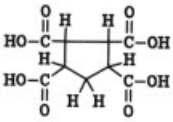	1,2,3,4-cyclopentanetetracarboxylic acid	280	235
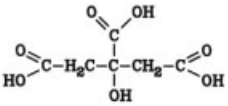	citric acid	195	30
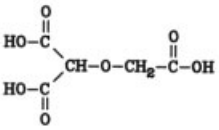	*O*-carboxymethyltartronic acid	247	123
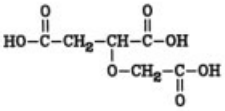	*O*-carboxymethyloxysuccinic acid	368	54

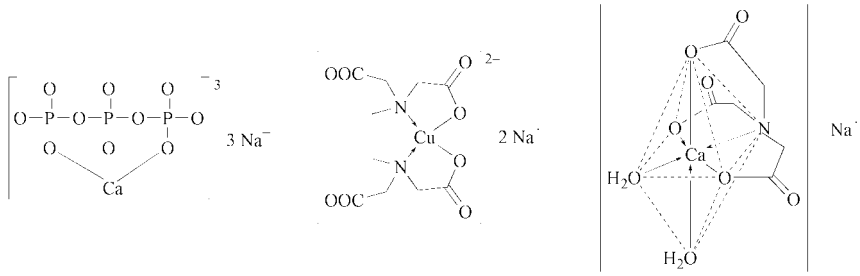

Figure 72. Metal complexes

those salts present to the greatest extent, such as calcium carbonate or others with even higher solubility products, be prevented from precipitating during the washing process. Less soluble calcium salts normally play a minor role.

Relative to other polyvalent ions, sequestration of alkaline earth ions is of primary concern because these ions are likely present in high concentration in tap water. Nevertheless, heavy-metal ions must also be eliminated because even in trace amounts their presence can have a negative effect on the washing process. For this reason, low concentrations of selective complexing agents are generally added as well, usually specific phosphonic acids, e.g., sodium hydroxyethanediphosphonate or sodium diethylenetriaminepentakis(methylenephosphonate).

If a sequestering agent is present in less than stoichiometric amounts relative to polyvalent metal ions, precipitation of carbonates and insoluble salts of the sequestering agent with the ions causing the water hardness usually results. Even with adequate amounts of sequestering agent, this effect is important because dilution during the rinse cycle can cause sequestrant concentration to drop below the necessary value, thereby permitting undesirable precipitates to form. These can build up on both fabric and washing machine components, leading eventually to serious accumulations. The problem can be particularly severe if conditions permit large crystals to form as a result of seed crystals located on fabric or machine components.

There are a number of compounds that even in substoichiometric amounts are capable of retarding, hindering, or otherwise interfering with precipitation of insoluble salts. In some cases their action induces salts to precipitate in amorphous form, thereby strongly reducing the tendency toward formation of crystals such as calcite, whose sharp edges can be damaging to fabrics. Sodium polycarboxylates and sodium hydroxyethanediphosphonate are frequently used additives in modern detergents, which strongly exhibit the latter property even at low concentrations; this is known as a *threshold effect* [167].

Despite the many desirable properties shown by sodium triphosphate in the washing process, its continued use has been the subject of an international debate between industry, governments, and water authorities in Europe, the USA, Japan, and other regions for many years. The problem has been that sodium triphosphate is a contributor to eutrophication of standing or slowly flowing surface waters; that is, it may lead to overfertilization, which in turn encourages extreme algal growth and adversely affects marine organisms. Recognition of the problem led to an intense worldwide search for suitable replacements in the 1960s and 1970s (see also Chap. 10). Developments have been concentrated not only on sequestering agents, but also on ion exchangers, since these are capable of binding polyvalent metal ions [155 – 157]. Most of the promising substitutes examined so far were organic compounds, primarily those derived from raw materials produced by the petrochemical industry. However, few of these substances were available in large quantity. The cost for commercial production on the necessary scale of $> 10^6$ t/a was in many cases prohibitive.

Apart from eliminating cations and achieving good soil and stain removal, other important factors in the washing process are dispersion of soil and prevention of soil redeposition (cf. Chap. 2). Because of the presence of distinct

localized charges, sequestering agents are readily adsorbed onto particulate soil. As a result, these compounds often act as effective dispersing agents for such soils.

Figures 73 and 74 show the relationship between dispersing agent concentration and sediment volume for two structurally different soil pigments where the sediment volume is understood as an indicator of the effectiveness of the dispersion process. Voluminous sediments are formed in the presence of poor dispersing agents; adsorption is insufficient to produce an adequate negative surface charge, and the result is coagulation. Figure 73 shows that the extent of dispersion, in this case of kaolin clay, depends on the specificity with which a sequestering agent is adsorbed. Virtually no effect is produced by sodium sulfate (a) and a nonionic surfactant (b). Anionic surfactants begin to show stabilization only at high concentrations that are rarely used in practice (f).

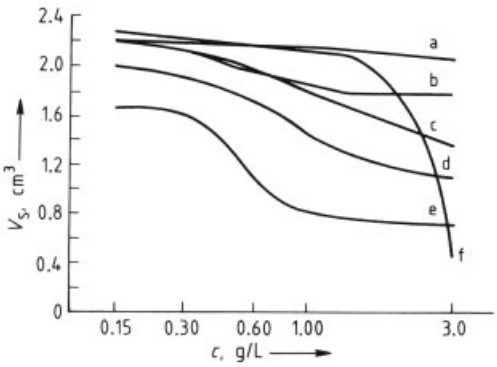

Figure 73. Sediment volume V_s of kaolin clay as a function of active ingredient concentration for water of hardness 16 °d (285 ppm $CaCO_3$) [46]
a) Sodium sulfate; b) C_{16-18} *n*-Alkyl decaglycol ethers; c) Hydroxyethyliminodiacetic acid, sodium salt; d) Nitrilotriacetic acid, sodium salt; e) Sodium triphosphate; f) Sodium *n*-dodecyl sulfate

With graphite, the results are exactly opposite to those observed with kaolin clay (Fig. 74). Sodium triphosphate alone shows no effect and is thus analogous to sodium sulfate (curves a and b), whereas sodium *n*-dodecyl sulfate has a very significant stabilizing effect. Because *n*-dodecyl sulfate is sensitive to water hardness, the dispersing effect increases with diminishing hardness; hence, the sediment volume decreases (both curves c: 16 and 0 °d). Although neither electrolytes nor sequestering agents alone are capable of stabilizing the particulate soil graphite, they are capable of exerting a positive influence, either indirectly through an electrolyte effect (curve d) or in hard water as a result of fixation of alkaline-earth ions (curve e). With sodium triphosphate, sequestration is accompanied not only by an electrolyte effect, but also by a pH effect. An increase in the hydroxide ion concentration and accompanying hydroxide ion adsorption causes increased electrostatic repulsion and, hence, better dispersion (reduced sediment volume). The indirect effect of sequestering agents on the dispersion of hydrophobic particulate soil applies also in principle to emulsification of water-insoluble greasy soil.

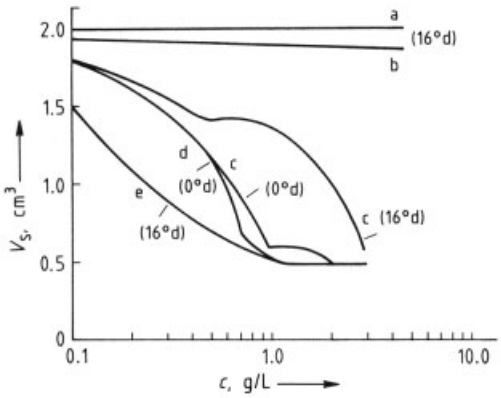

Figure 74. Sediment volume V_s of graphite as a function of active ingredient concentration [46]
a) Sodium sulfate; b) Sodium triphosphate; c) Sodium *n*-dodecyl sulfate; d) Sodium *n*-dodecyl sulfate + 1.5 g/L sodium sulfate; e) Sodium *n*-dodecyl sulfate + 1.5 g/L sodium triphosphate

A number of additional factors must be considered when selecting a sequestering agent for detergent use. Several important criteria have been listed in Section 3.2. The extent to which evaluation of a sequestering agent can vary depends on how the several criteria are weighted, as is apparent from Table 15.

3.2.3. Ion Exchangers

The disruptive effect of polyvalent metal ions can be reduced not only by the use of low molecular mass sequestering agents, but also by ion

Table 15. Evaluation of sequestering agents according to various criteria

Ingredient (as sodium salt)	Calcium sequestration	Detergency performance	Incrustation on fabrics and washing machines	Hygroscopicity
Sodium diphosphate	low	very good	very heavy	insignificant
Sodium triphosphate	adequate	very good	little	insignificant
1-Hydroxyethane-1,1-diphosphonic acid	very high	very good	very little	hygroscopic
Nitrilotrimethylenephosphonic acid	high	good	little	very hygroscopic
Nitrilotriacetic acid	high	very good	very little	hygroscopic
N-(2-hydroxyethyl)iminodiacetic acid	less adequate	good		very hygroscopic
Ethylenediaminetetraacetic acid	high	good	very little	very hygroscopic
1,2,3,4-Cyclopentanetetracarboxylic acid	high	fair	little	hygroscopic
Citric acid	low	fairly low	little	insignificant
O-Carboxymethyltartronic acid	high	good	little	hygroscopic
O-Carboxymethyloxysuccinic acid	less adequate	fair		hygroscopic

exchangers [156]. Table 16 shows that ion exchangers generally have a high binding capacity for calcium but that this usually decreases with increasing temperature.

In the early years the idea of introducing water-insoluble substances into detergents had never been seriously investigated. This was primarily because the known materials lacked sufficient calcium binding ability, were available only in forms with unsuitable particle structure, and were deemed impractical for economic reasons. Success was first achieved through research in the field of sodium aluminum silicates. Systematic investigation revealed that, among the many known types of sodium aluminum silicates, those with a regular crystalline form were appropriate for use in the washing process. One particular modification proved especially applicable and economically interesting: zeolite A [121 – 123, 159 – 163], manufactured under the registered name "Sasil" (**s**odium **a**luminum **sil**icate) [155 – 157]. The ion-exchange behavior of this particular water-insoluble crystalline sodium aluminum silicate (Table 16) depends on ionic size and on the state of hydration of the ions. In addition to Ca and Mg, exchange also takes place with, e.g., Pb, Cu, Ag, Cd, Zn, and Hg ions. The elimination of calcium ions — and to a lesser extent magnesium ions — is of greatest importance for the washing process, but ion exchange capacity for heavy metal ions is also important from an ecological standpoint.

Ion exchange is dependent not only on ionic size, but also on concentration, time, temperature, and pH value. Figure 75 shows the exchange kinetics as a function of concentration and time. Calcium ions are exchanged very rapidly. The process occurs somewhat more slowly with magnesium, although exchange becomes more rapid at higher temperature. These results can be explained by the larger hydration shell of magnesium ions that impedes exchange at low temperature but is largely destroyed at higher temperature.

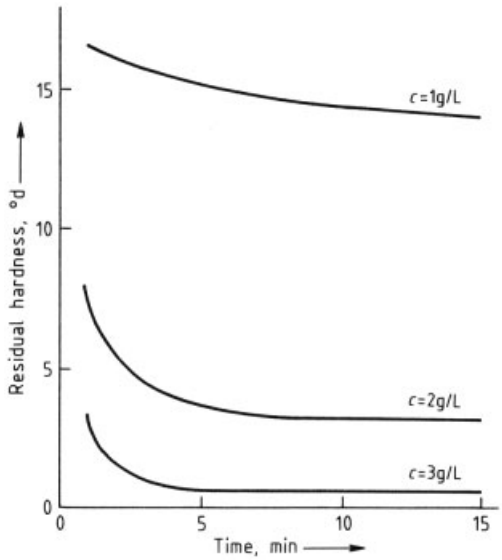

Figure 75. Kinetics of calcium binding to zeolite A [164] Initial hardness: 30 °d (530 ppm $CaCO_3$); temperature: 25 °C

The fact that ion exchange occurs in a heterogeneous phase and that adsorption–desorption phenomena are absent means it is advantageous to employ combinations of sodium aluminum silicates with water-soluble sequestering agents. The latter have the ability to remove polyva-

Table 16. Calcium binding capacity of selected ion exchangers [158]

Formula	Chemical name	Water solubility*	Calcium binding capacity**, mg CaO/g	
			20 °C	90 °C
$\left.-CH_2-\underset{COOH}{CH}-\right\|_n$	Poly(acrylic acid)	s	310	260
$\left(-CH_2-\underset{COOH}{CH}-\right)_x \left(-CH_2-\underset{H_2C\ OH}{CH}-\right)_y{}_n$	Poly(acrylic acid-*co*-allyl alcohol)	s	250	140
$\left(-CH_2-\underset{COOH}{CH}-\right)_x \left(-\underset{HOOC}{CH}-\underset{COOH}{CH}-\right)_y{}_n$	Poly(acrylic acid-*co*-maleic acid)	s	330	260
$-CH_2-\overset{OH}{\underset{COOH}{C}}-{}_n$	Poly(α-hydroxyacrylic acid)	s	300	240
$\left[-CH_2-CH_2-\underset{HOOC}{CH}-\underset{COOH}{CH}-\right]_n$	Poly(tetramethylene-1,2-dicarboxylic acid)	s	240	240
$\left[-\underset{OCH_3}{CH}-CH_2-\underset{HOOC}{CH}-\underset{COOH}{CH}-\right]_n$	Poly(4-methoxytetramethylene-1,2-dicarboxylic acid)	s	430	330
$xNa_2O \cdot Al_2O_3 \cdot ySiO_2 \cdot zH_2O$	Sodium aluminium silicate (Zeolite A)	i	165	190

* s = soluble, i = insoluble.
** Sodium salt.

lent cations, especially calcium and magnesium, from a solid surface, transport them through an aqueous medium, and then release them to the ion exchanger. This property is best described as a carrier effect consisting of the following steps (cf. also Fig. 22):

1) Sorption of the carrier onto the fiber–soil interface
2) Complexation of calcium and magnesium ions
3) Transfer from the solid fiber–soil interface through the wash liquor to the sodium aluminum silicate–water interface
4) Dissociation of the complex into carrier and calcium and magnesium ions
5) Ion exchange of calcium and magnesium ions for sodium ions in the sodium aluminum silicate crystalites
6) Renewed sorption of the carrier on the fiber–soil interface and further complexation with calcium and magnesium ions

If the properties of sodium triphosphate (a water-soluble sequestering agent) are compared with those of sodium aluminum silicate (zeolite A), a number of similarities become apparent, as do typical differences [156, 164]:

Characteristics of sodium triphosphate:
Characteristics of sodium triphosphate:

1) a) complex formation with polyvalent ions
   b) alkaline reaction
   c) specific adsorption on particulate soil and fibers
   d) specific electrostatic charging of particulate soil and fibers, dissolution of polyvalent ions from soil and fibers
   e) nonspecific electrolyte effect

Characteristics of sodium aluminum silicate:
2
a) binding capability for polyvalent ions through ion exchange
b) alkaline reaction
c) adsorption of molecularly dispersed substances
d) heterocoagulation with particulate soil
e) crystallization surface for poorly soluble compounds

The different properties of sodium triphosphate and sodium aluminum silicates can be explained on the basis of their differing solubilities and the divergent ways they eliminate ions contributing to water hardness. In the case of sodium triphosphate, calcium binding occurs by

chelation, whereas with sodium aluminum silicates, calcium binding is a result of ion exchange [164].

Ca^{2+} binding by chelation (sodium triphosphate):

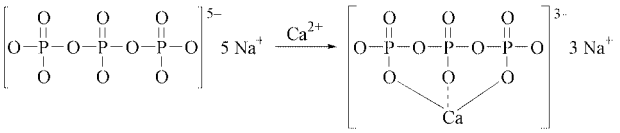

Ca^{2+} binding by ion exchange (zeolite A):

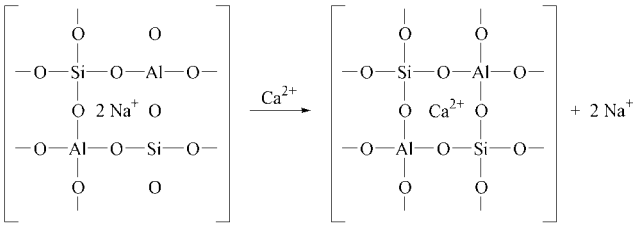

Additional differences result from the exceptional adsorption capability of the triphosphate anion.

The binary builder system sodium triphosphate–zeolite A assures outstanding multiple wash cycle performance; that is, incrustations resulting from precipitation of calcium and magnesium phosphates are largely avoided. Replacement of sodium triphosphate by zeolite A in a detergent leads to some deterioration in removal of some soils and stains. By reoptimization of the detergent formulation with respect to surfactants, bleach system, enzymes, etc., this effect can be easily compensated for.

Alternative ion exchangers like certain water-soluble polycarboxylic acids have also achieved great importance for phosphate-free and low-phosphate detergents (see Section 3.2.4).

Zeolite A, as it is mass produced for use in detergents, appears as cubic crystals with rounded corners and edges. These crystalline particles show a tendency to form agglomerates [156, 160] (Fig. 76).

The average diameter of a zeolite A particle is ca. 4 µm. A diameter distribution curve obtained by differentiation of a residue summation curve (determined by using a Coulter counter) shows quite a sharp maximum at the corresponding abscissa value (Fig. 77).

Zeolite A is specially optimized for use in detergents. Its unique particle form, cubes with rounded corners, protects fibers against damage, and the tiny particles with their narrow range of particle sizes ensures that deposition on laundry is practically nonexistent [156].

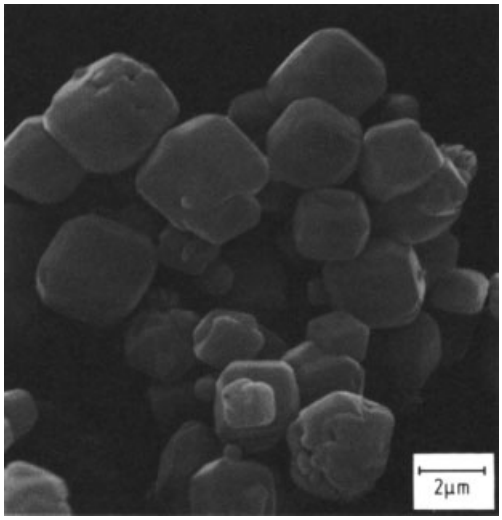

Figure 76. Scanning electron micrograph of zeolite A [160]

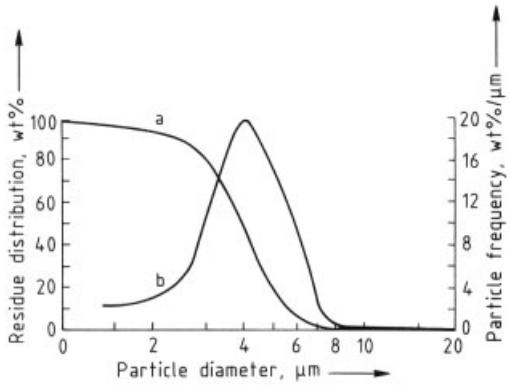

Figure 77. Residue summation curve and particle diameter distribution curve for zeolite A [160]
Coulter counter, 50 µm cell
a) Residue distribution; b) Particle frequency

Table 17 summarizes a number of important criteria for the evaluation of water-soluble and water-insoluble ion exchangers.

Table 17. Comparison of various ion exchangers [46]

Substance (as sodium salt)	Calcium sequestration	Detergency performance	Incrustation on fabrics and washing machines	Hygroscopicity	Biodegradability*
Poly(acrylic acid)	very high	good	little	very hygroscopic	minimal
Poly(α-hydroxyacrylic acid)	high	very good	little	hygroscopic	minimal
Poly(acrylic acid-*co*-allyl alcohol)	high	fairly good	little	very hygroscopic	minimal
Poly(4-methoxytetramethylene-1,2-dicarboxylic acid)	high	good	little	very hygroscopic	minimal
Poly(tetramethylene-1,2-dicarboxylic acid)	high	good	little	very hygroscopic	minimal
Poly(acrylic acid-*co*-maleic acid)	high	good	little	hygroscopic	minimal
Sodium aluminum silicate (zeolite A)	high	good **	little	minimal	irrelevant

* Under the conditions of the Closed Bottle Test [338].
** In combination with appropriate soluble sequestering agents or soluble ion exchangers.

3.2.4. Builder Combinations

Since the beginning of the 1980s, the ternary builder combination zeolite A–polycarboxylate–soda ash has been increasingly used in nonphosphate detergents all over the world [165]. Polycarboxylates — polyacrylate and poly(acrylate-*co*-maleinate) — have an outstanding function in this ternary builder system in that they effectively prevent lime and soil depositions on laundry and washing machines. The mechanism of action is based on the *threshold effect* (see Section 3.2.2) [166, 167]. Further ternary builder systems such as sodium triphosphate–sodium carbonate–sodium polycarboxylate are also frequently applied, of course only in those global regions where phosphate-built laundry detergents are common. The use of substoichiometric quantities of high molecular mass polymers or copolymers of acrylic and maleic acid in these formulations enables a reduction in the content of sodium triphosphate and its partial substitution by the more cost effective sodium carbonate. This is due to the ability of the polycarboxylates to retard or inhibit crystal growth of sparingly soluble salts, such as calcium phosphate and calcium carbonate. Increasing charge density of the polycarboxylic acid (i.e., maleate content in the copolymer) increases its sequestration capacity [167, 168].

Special crystalline and amorphous silicates, sodium carbonate–silicate compounds, and zeolite P are also used as alternatives to zeolite A. These alternatives are found in some European, U.S., and Japanese nonphosphate detergent formulations [169, 170]. Zeolite P was introduced as a builder in laundry detergents in 1994 [171, 172]. It has the chemical formula $Na_2O \cdot Al_2O_3 \cdot (2–5)\,SiO_2 \cdot x\,H_2O$ and is also known under the trade names zeolite MAP, Doucil A 24, and Wessalith NaP. Zeolite P has the advantage of substantially stronger calcium binding than zeolite A. Furthermore, zeolite P exhibits a superior uptake for nonionic surfactants due to is smaller average particle size and, hence, larger surface per unit weight than zeolite A [256, 288 – 290]. This property can be essential for special detergent manufacturing technologies.

Unlike zeolites, silicate-based builders are water-soluble and contribute to the alkalinity of the washing liquor. They are generally better compatible with sensitive bleach components such as percarbonate. Crystalline layered disilicate $Na_2Si_2O_5$ and sodium carbonate–sodium silicate are most suitable for use as detergent builders. They are commercially available under the tradenames SKS-6 and Nabion 15, respectively. Due to their superior efficiency in preventing buildup of lime deposits, they can replace zeolite A in nonphosphate detergents [169, 170, 282].

3.3. Bleaches

The term bleach can be taken in the widest sense to include the induction of any change toward a lighter shade in the color of an object. Physically, this implies an increase in the reflectance of visible light at the expense of absorption.

Bleaching effects can occur through mechanical, physical, and/or chemical means, specifi-

cally through change or removal of dyes and soil adhering to the bleached object. In the washing process, all of these processes occur in parallel, but to varying extents. The relative significance of each is determined in part by the nature of the soil and dyes present. Mechanical/physical mechanisms are effective primarily for the removal of particulate and greasy soil. Chemical bleaching is employed for the removal of colored nonwashable soils and stains adhering to fibers and is accomplished by oxidative or reductive decomposition of chromophoric systems. Only oxidative bleaches are used in laundry products to a great extent; many soils commonly encountered in everyday laundry contain compounds which, if bleached reductively, become colorless but may later return to their colored forms as a result of subsequent atmospheric oxidation. Nevertheless, this generalization does not rule out the use of special reductive bleaches (e.g., $NaHSO_3$, $Na_2S_2O_4$) to treat specific types of discoloration occurring in either household or institutional settings.

The extent of the bleaching effect that can be achieved is dependent on a number of factors, including the type of bleach, its oxidation potential and concentration, as well as the residence time in the washing or rinsing process, the wash temperature, the type of soil to be bleached, and the nature of the fabric.

Bleachable soils encountered in household and institutional laundry consist of a broad spectrum of diverse materials which are generally of vegetable origin and contain primarily polyphenolic compounds. These include the anthocyanin dyes derived from, e.g., cherries, blueberries, and currants, and curcuma dyes from curry and mustard. The brown tannins found in, e.g., fruit, tea, and wine stains arise from condensation of polyphenols with proteins. Other brown organic polymers include the humic acids present, for example, in coffee, tea, and cocoa. The green dye chlorophyll and the red betanin from beets are pyrrole derivatives, as are the urobilin and urobilinogen dyes derived from degradation of hemoglobin and discharged in urine. Carrot and tomato stains contain carotenoid dyes. Dyes of commercial origin such as those found in cosmetics, hair coloring agents, and ink are also important. Blood is also a bleachable soil, but its removal can sometimes present problems.

Two procedures have attained major importance in oxidative bleaching during the washing and rinsing processes: peroxide bleaching and hypochlorite bleaching. The relative extent of their application varies, depending heavily on laundering habits in the different global regions.

3.3.1. Bleach-Active Compounds

The dominant bleaches in Europe and many other regions of the world are of the *peroxide* variety. Hydrogen peroxide is converted by alkaline medium to the active intermediate hydrogen peroxide anion according to the following equation:

$$H_2O_2 + OH^- \rightleftharpoons H_2O + HO_2^-$$

The perhydroxyl anion oxidizes bleachable soils and stains. The usual sources of hydrogen peroxide are inorganic peroxides and peroxohydrates. The most frequently encountered source is sodium perborate (sodium peroxoborate tetrahydrate, $NaBO_3 \cdot 4\ H_2O$) [173], which in crystalline form contains the peroxodiborate anion:

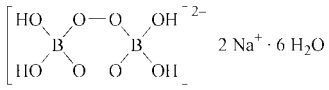

Peroxodiborate is hydrolyzed in water to form hydrogen peroxide. Sodium perborate monohydrate and sodium percarbonate, which dissolve more quickly than perborate tetrahydrate, have been increasingly used either as separate color-safe bleach or in bleach-containing detergents at the expense of sodium perborate tetrahydrate. The monohydrate version has an improved storage stability as compared with the tetrahydrate version. Sodium perborate monohydrate is the preferred ingredient for laundry products in countries having high ambient temperatures because, contrary to perborate tetrahydrate, the monhydrate does not cause caking of powder products upon storing at elevated temperatures for prolonged periods. In contrast to what its name would suggest, sodium perborate monohydrate contains no water of crystallization. Its molecular structure is [174]:

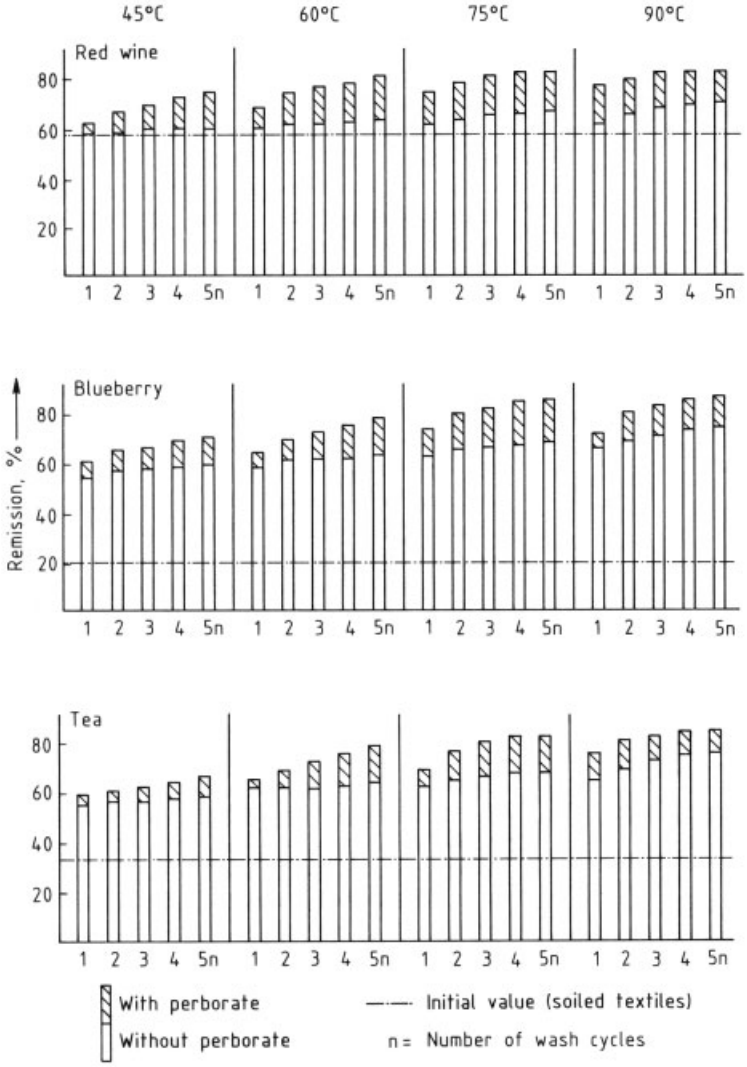

Figure 78. Bleach performance of sodium perborate vs. temperature [178]
Initial perborate tetrahydrate concentration 1.5 g/L = 150 mg of active oxygen/L

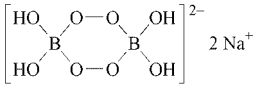

On a per weight basis the monohydrate has a considerably higher content of active oxygen than the tetrahydrate. For this reason, the monohydrate has become the favorite form of perborate in compact detergents.

The use of sodium percarbonate has gained importance in those countries in which boron is either banned or restricted for environmental and regulatory reasons, or negatively discussed. In contrast to sodium perborate, sodium percarbonate is a true peroxohydrate ($Na_2CO_3 \cdot 1.5\ H_2O_2$). In order to achieve a good storage stability for the bleach, coated or stabilized percarbonate must be used.

The salts of peroxomono- and peroxodisulfuric acid and peroxomono- and peroxodiphosphoric acid are not significant as detergent bleaches. This results largely from their insufficient bleaching power in wash liquor, either because they are insufficiently hydrolyzed to hydrogen peroxide in an alkaline medium or because their oxidation potential is too low.

Comparison shows that sodium perborate and sodium percarbonate have the most preferred prerequisites for use as a detergent bleach additive. Oxygen bleaches have been reviewed [174 – 177].

The concentration of bleach-active hydrogen peroxide anion increases with pH value and temperature. Sodium perborate exhibits significantly less bleaching efficiency at temperatures below 60 °C. Even at low temperatures (i.e., where the reaction equilibrium is unfavorable), hydrogen peroxide anions are present in the wash water, but show only modest bleaching power (Fig. 78). The bleaching effect also increases markedly with increasing perborate concentration (Fig. 79) and time.

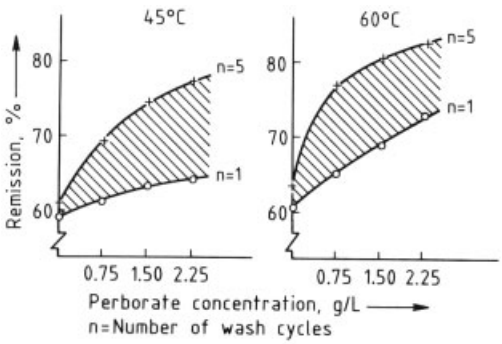

Figure 79. Bleach performance of sodium perborate vs. concentration (red wine soil) [178]

Many attempts were made in the 1980s to improve the performance of a detergent by using organic peroxy acids, for example, monoperoxyphthalic acid and diperoxydodecanedioic acid (DPDDA) salts, as bleach components:

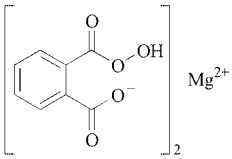

Monoperoxyphthalic acid monomagnesium salt

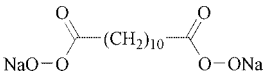

Diperoxydodecanedioic acid sodium salt

With these it is possible to obtain significant bleaching at a temperature as low as 30 °C. However, for several reasons (dye damage and cost effectiveness) a technical and commercial breakthrough with these peroxyacids has not been achieved so far.

Hypochlorite is used for bleaching in many global regions where laundry habits, such as cold water washing, cause sodium perborate to be less effective. In an alkaline medium, hypochlorite bleaches are converted to the hypochlorite anion:

$$HOCl + OH^- \rightleftharpoons ClO^- + H_2O$$

Hypochlorite bleach chemistry and applications have been reviewed [179 – 181].

Hypochlorite can be used in either the wash or the rinse cycle at concentrations between ca. 50 and 400 mg/L active chlorine (Fig. 80) and a wash liquor ratio (bath ratio) between 1 : 15 and 1 : 30.

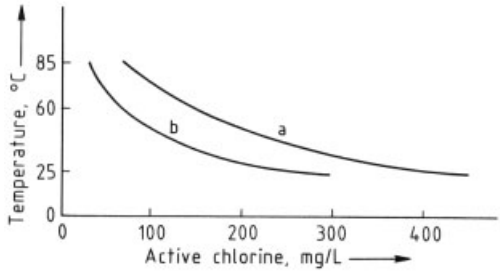

Figure 80. Recommended concentrations of active chlorine as a function of wash temperature [182]
a) Maximum concentration of application; b) Minimum concentration of application
"Active or available chlorine" is calculated as twice the actual weight percent of chlorine in the hypochlorite molecule

Normally, an aqueous solution of sodium hypochlorite (NaOCl) is used as a source of active chlorine. Organic chlorine carriers (e.g., sodium dichloroisocyanurate), which are hydrolyzed to hypochlorite in an alkaline medium, are only common in automatic dishwashing detergents. Figure 81 shows a comparison of the bleaching properties of sodium hypochlorite and sodium perborate as a function of temperature and pH value.

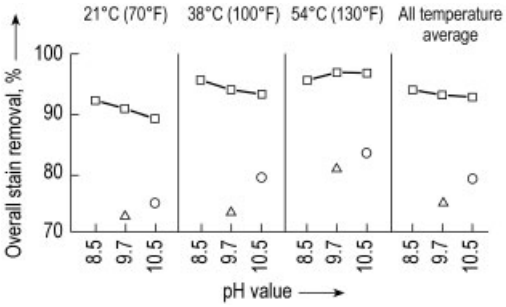

Figure 81. Stain removal comparisons
□ Hypochlorite (200 ppm active chlorine) + 0.15 % detergent; ○ Perborate bleach (28 ppm active oxygen) with protease enzymes + 0.15 % detergent; △ 0.15 % Detergent only[183]

One of the major advantages of powdery sodium perborate over liquid sodium hypochlorite is the fact that whereas the latter must be added separately in either the wash or the rinse cycle, perborate can be incorporated directly into a powder laundry product. This results in a mild color-safe bleach. By contrast, successful addition of chlorine bleach solution, whether in a household or an institutional machine, is heavily dependent on experience of the user and on observing the manufacturer's recommendations. Incorrect dosage of sodium hypochlorite may easily occur, and this may cause significant damage to laundry and colors. A further advantage of sodium perborate is its long shelf life, whereas sodium hypochlorite solutions have limited storage stability, which depends very much on the content of impurities, particularly traces of heavy metal ions. On the other hand, hypochlorite bleaches can be used in both wash and rinse cycles independent of temperature. They provide effective bleaching and disinfection (hygiene) even at very low temperature. However, because of its great reactivity and extraordinarily high oxidation potential, sodium hypochlorite, in contrast to sodium perborate, may cause problems with textile dyes and most fluorescent whitening agents, both of which often show poor stability in the presence of chlorine. Studies of washing and bleaching habits show that peroxide bleach use dominates in Europe, chlorine bleach being used predominantly in Mediterranean countries. Hypochlorite bleach in either the wash or the rinse cycle is still the preferred bleaching agent in a large part of the world (cf. Table 18). The use of peroxide ("oxygen") bleach in laundry products has increased significantly in the USA, Japan, and other regions of the world during the 1980s and 1990s.

3.3.2. Bleach Activators

To achieve satisfactory bleaching with sodium perborate and sodium percarbonate at temperatures $\leq 60\ ^{\circ}C$, so-called bleach activators are commonly utilized. These are mainly acylating agents incorporated in laundry products. When present in a wash liquor of pH value 9–12, these activators preferentially react with hydrogen peroxide (perhydrolysis) to form organic peroxy acids in situ. As a result of their higher oxidation potentials relative to hydrogen peroxide, these intermediates demonstrate effective low-temperature bleaching properties. Due to the low concentration, resulting from the in situ generation, these peroxyacids are much less aggressive to fabric dyes and fluorescent whitening agents than sodium hypochlorite. However, activated peroxy bleach systems may adversely affect the color fastness of fibers treated with some specific dyes upon multiple washing [184]. Among the wide variety of bleach activators investigated [185], only the following compounds have been incorporated on a large scale in laundry products worldwide:

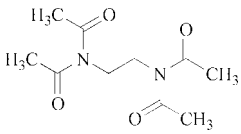

Tetraacetylethylenediamine (TAED) [186].

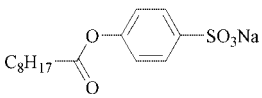

Sodium *p*-nonanoyloxybenzenesulfonate (NOBS) [193]

The first activator reacts with hydrogen peroxide according to the following scheme to produce peroxyacetic acid:

Table 18. Washing habits in different regions

Washing conditions	United States	Japan	Europe
Washing machine	vertical axis/agitator type	vertical axis/pulsator type	horizontal axis drum type
Heating equipment	no	no	yes
Fabric load, kg	23–5	4–8	2–5
Amount of water, L	medium loads: ca. 50–60 large loads: ca. 65–75	low: 30 high: 65	low: 10–16 high: 20–25
Total water consumption, L (regular cycle)	120	120–150	55–90
Wash liquor ratio	ca. 1 : 25	ca. 1 : 10–1 : 15	ca: 1 : 4–1 : 10
Washing time, min	8–18	5–15	50–60 (90 °C) 20–30 (30 °C)
Washing, rinsing, and spinning time, min	20–35	15–35	100–120 (90 °C) 50–70 (30 °C)
Washing temperature, °C	hot: 55 (130 °F) warm: 30–45 (90–110 °F) cold: 10–25 (50–80 °F)	5–25	90 60 40 30
Water hardness (average), ppm $CaCO_3$	relatively low, 100	very low, 30	relatively high, 250
Detergent dispenser	mostly no	mostly no	yes
Recommended detergent dosage * ,g/L	1.0–1.2	0.5–0.8	4.5–8.0
g/kg laundry	15–30	5–7	15–35
Peroxide bleach	mostly added separately	mostly dosed separately	mostly incorporated in detergent
Chlorine bleach	added separately	dosed separately	predominantly in mediterranean countries

* In the United States and Japan mostly without bleaching components.

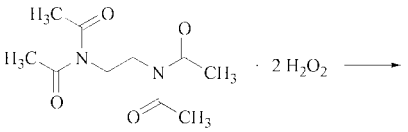

TAED

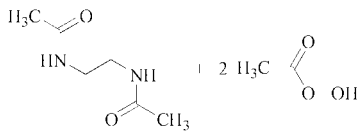

DAED

Under the conditions of the washing process the reaction activates only two acetyl groups. TAED has been used in laundry products in Europe since 1977 and has became state-of-the-art in many laundry detergents and bleaches. TAED has remained the bleach activator of choice in Europe, whereas NOBS has been predominantly used in the Americas and Asia. The broad spectrum of effectiveness of TAED has been reviewed [187 – 190]. Sodium *p*-nonanoyloxybenzenesulfonate (NOBS) reacts with hydrogen peroxide (is perhydrolyzed) to produce peroxynonanoic acid. Aside from their efficiency regarding improvement of low-temperature bleaching, these bleach activators markedly contribute to enhancing the hygiene of laundry due to the biocidal effect of the peracids generated in situ [191, 192]. The effectiveness of the system sodium perborate–TAED on tea stains as a function of concentration and temperature is shown in Figures 82 and 83.

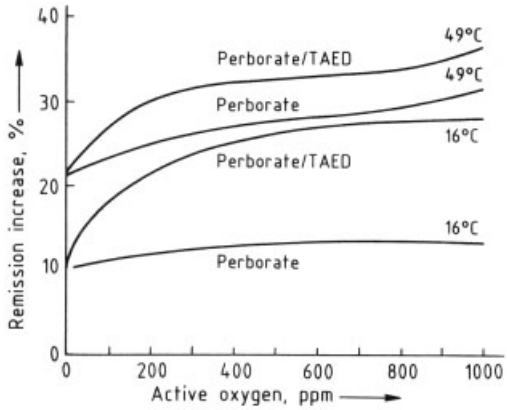

Figure 82. Bleaching effect as a function of bleach concentration [194]

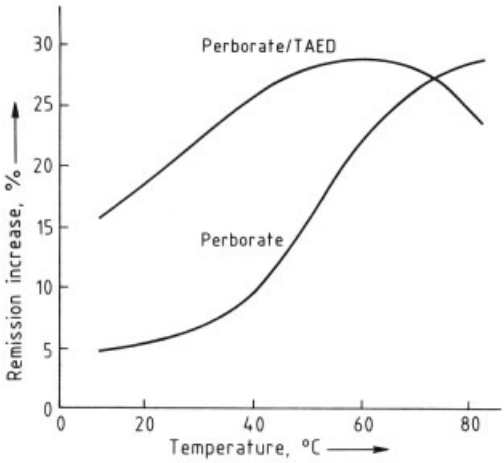

Figure 83. Bleaching effect as a function of temperature [194]

Perborate–TAED bleach systems are not sufficiently effective at cold washing temperatures used in large regions of the world. However, soaking of the laundry in a detergent containing activated perborate prior to washing can greatly enhance the removal of bleachable stains due to the prolonged time of soaking [195]. The problem of graying of white textiles caused by dye transfer during soaking can be reduced by the addition of perborate–TAED to a detergent.

3.3.3. Bleach Catalysts

Numerous attempts have been made to introduce into sodium perborate and/or percarbonate based laundry products small amounts of catalysts that would increase their bleaching performance, especially at low temperature. In most cases, traces of heavy metal ions have been used. While addition of heavy metal ions alone can indeed cause decomposition of sodium perborate, it does not lead to a better bleach. In fact, bleaching action is usually diminished, and serious fiber and color damage may occur. The patent literature describes a number of examples in which the bleaching power of sodium perborate can be increased by heavy metal ion chelates. However, the effectiveness of such systems remains controversial, and catalysts have not been accepted as a part of the washing process. Excellent catalytic bleach enhancement by manganese complexes derived from 1,4,7-trimethyl-1,4,7-triazacyclononane and related ligands has been reported [196]. However, under certain conditions severe dye and fiber damage has been observed in connection with this catalyst system. For this reason detergents using manganese catalysts introduced on the European markets in 1994 were later withdrawn.

Photobleaching with aluminum or zinc phthalocyaninetetrasulfonate represents another form of bleach catalysis. With photobleaching, oxygen from the atmosphere is catalytically activated by the phthalocyanine compound, and the active oxygen generated in turn bleaches oxidizable stains. Suitable conditions for the few detergents that contain photobleaches include — apart from soaking in sunlight — slow drying of the laundry under conditions of high humidity and high light intensity. Therefore, photobleach catalysts are of interest and have been employed in laundry products only in regions with intense solar radiation.

3.3.4. Bleach Stabilizers

Traces of ions such as copper, manganese, and iron catalyze the release of oxygen from bleach systems. This reduces bleach effectiveness and at the same time causes damage to fabrics and dyes. The addition of small amounts of finely divided magnesium silicate largely suppresses this catalysis as a result of absorption of the heavy metal ions. Therefore, magnesium silicate was used in perborate based laundry detergents until the 1980s.

Another possibility for elimination of trace heavy metal ions is the addition of selective complexing agents. Sodium triphosphate has only a modest complexing effect with heavy metal ions and is incapable of causing sufficient bleach stabilization in their presence. By contrast, complexing agents such as sodium hydroxyethanediphosphonate (HEDP) or sodium diethylenetriamine pentakis(methylenephosphonate), as well as other phosphonic acids, exhibit a marked stabilizing effect. The use of these phosphonates is state of the art in many modern laundry products. Amounts of much less than 1 wt % of these strong, selective complexing agents suffice in detergent formulations.

3.4. Further Detergent Ingredients

Surfactants, builders, and bleaches are quantitatively the major components of modern detergents; the auxiliary agents — also called additives — discussed in this section are introduced only in small amounts, each to accomplish its own specific purpose. Today, their absence from detergent formulations is difficult to imagine. They have greatly contributed to making laundry detergents more multifunctional during the last 30 years.

3.4.1. Enzymes

Stubborn proteinaceous stains derived from sources such as milk, cocoa, blood, egg yolk, and grass are resistant to removal from fibers by enzyme-free detergents, particularly when stains have dried-on. So are chocolate and starch-based food stains as well as greasy/fatty stains, particularly in low-temperature washing. Proteolytic, amylolytic, and lipolytic enzymes are usually capable of eliminating such soil without major problems during washing.

Use of enzymes in detergents was first described by Otto Röhm in a 1913 patent application. Enzyme-containing detergents failed to play a major role in the following decades because the only available proteolytic enzyme preparation, a pancreatic extract obtained from slaughtered animals, was too sensitive to the alkaline and oxidative components of detergents. Only in 1959 were Röhm's ideas realized by the use of proteolytic enzymes from fermentation with specific strains of bacteria (*Bacillus subtilis*, later on *Bacillus licheniformis*). These *protease* enzymes were highly resistant to alkali and showed adequate stability at temperatures as high as ca. 65 °C for the time period required by normal washing processes. The first successful detergents on the market containing these proteases were Bio 40 from Gebr. Schnyder in Switzerland in 1959, followed by Biotex produced by Kortman & Schulte in The Netherlands in 1963. Commercial production of detergent enzymes experienced rapid expansion in the years that followed. For example, by 1969 nearly 80 % of the detergents marketed in the Federal Republic of Germany contained proteases as enzyme additives. Industry reduced this proportion to < 50 % as a result of public discussion of certain toxicological concerns. The reservations regarding detergent enzymes have since been addressed and overcome by technical modifications, particularly those involving the delivery form in which enzymes are introduced into detergents: enzymes have been manufactured since 1971 as dust-free granulates, prills, extrudates, or pellets. Today's detergent enzymes are perfectly safe. As a result, nearly all detergents produced in Europe, North America, Japan and many other countries worldwide contain enzymes today. Table 19 provides a brief historical review of the production and application of detergent enzymes.

The effectiveness of proteolytic, amylolytic, and lipolytic detergent enzymes is based on enzymatic hydrolysis of peptide, glucosidic, or ester linkages, respectively. To be well suited for use in detergents, enzymes must exhibit the following properties [197]:

- Activity optimum at alkaline pH
- Efficacy at low wash temperatures of 20–40 °C
- Stability at wash temperatures up to 60 °C
- Stability in the presence of other detergent ingredients, such as surfactants, builders, and activated bleach, both during storage and use
- Low specificity to soils, i.e., a specificity broad enough to enable the degradation of a large variety of proteins, starches, and triglycerides

The activity of enzymes depends on temperature and pH value of the wash bath. Figures 84 and 85 show these respective dependencies of a typical commercially available protease (Savinase) [199].

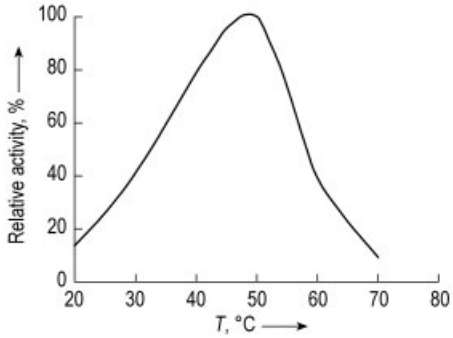

Figure 84. Activity of Savinase at different temperatures [199]

Table 19. Historical review of the preparation and use of detergent enzymes [197, 198]

Year	Enzyme	Enzyme-containing detergents
1913	Otto Röhm claims the use of tryptic enzymes for detergents	
1931		detergents containing pancreatic enzymes (Burnus, Germany)
1959	Bacterial protease	Bio 49 (Gebr. Schnyder, Switzerland)
1960	Alcalase and Maxatase, microbial proteases made available on a commercial scale by Novo Industri, Bagsvaerd, Denmark and Gist-Brocades, Delft, The Netherlands, respectively	
1963		prewash detergent Biotex (Kortman & Schulte, The Netherlands)
1966		first heavy-duty detergent with microbial proteases
1969		microbial proteases contained in 80 % of all detergents in Germany
1970		severe setback of production of microbial proteases due to public criticism (the "allergy debate")
1972	Additional microbial enzymes made commercially available for use in detergents (amylases)	enzymes in detergents declared to be safe by the German Federal Health Agency
1973	Amylase introduced in Germany	presoak and prewash detergent (Mustang, Henkel, Germany)
1975		market share of enzyme-containing detergents in Germany is 80 %
1975	Liquid proteases for HDL (Maxatase LS, Alcalase L)	
1980s	High-alkaline proteases become available (Savinase, Novo; Maxacal, IBIS; BLAP, Henkel; Purafect, Genencor)	improved efficacy and storage stability for enzymes in heavy-duty detergents
1986	Cellulase (Celluzyme, Novo)	
1987		heavy-duty detergent containing cellulase KAC (Attack, Kao, Japan)
1988	Lipase (Lipolase, Novo)	heavy-duty detergent containing lipase (Hi Top, Lion, Japan)
1988–1989	Bleach-resistant proteases (Maxapem, Gist-Brocades; Durazym, Novo)	
1990s	Enzymes spreading out in detergents worldwide (China, India, South America)	roll-out of multifunctional detergents with more than one enzyme in Europe, Japan, and the USA

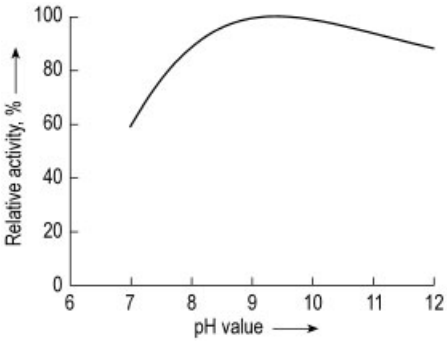

Figure 85. Activity of Savinase at different pH values [199]

Ever since detergent-grade enzymes were introduced in the 1960s there has been made considerable progress in terms of

Delivery form
Reduced dust content
Improved alkali resistance
Increased low-temperature activity
Increased activity per unit weight of finished product
Storage stability
Resistance to active oxygen
Effectiveness, functionality
Genetic and enzyme engineering
Decreasing market prices

Enzymes in detergents have therefore meanwhile become state-of-the-art worldwide. Their use spread dramatically in Western Europe, North America, and Japan in the 1970s and 1980s. The 1990s saw also the opening up of the vast detergent markets of China, South America, and India for enzymes [198]. Figure 86 [200] shows the geographical use and value of enzymes in the world. Penetration of enzymes in the U.S. market is 75 %, compared with 95 % for both Europe and Japan. The mainstay of the market have been the protease types.

Only above a certain temperature of the wash bath (mostly 55 °C) does the activity of enzymes decrease, and then very rapidly. Other modes of decomposition present little problems.

For example, intensive research has shown that detergent-grade enzymes currently available are not subject to significant damage during storage or in the wash liquor, provided detergents containing them are properly formulated. Stability problems with respect to anionic surfactants and active oxygen were once the subject of debate, but usually these are readily solved by the proper choice of enzyme type and the appropriate formulation.

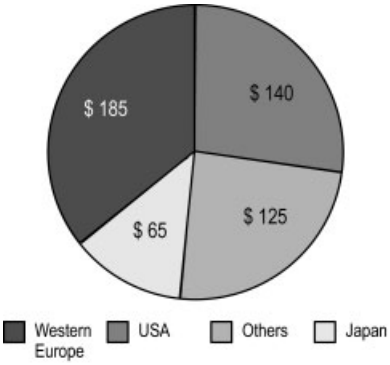

Figure 86. Regional detergent enzyme market values (1995, 10^3 $)

Especially in the U.S. and European markets, *amylases* have been added to detergents along with proteases since 1973 to capitalize on the activity of the amylases toward starch-containing soils. Of the different amylases available, only α-amylases are used for detergents. They are able to catalyze the hydrolysis of the amylose and amylopectin fractions of starch, i.e., cleavage of the α-1,4-glycosidic bonds of the starch chain [201]. This eases the removal of starch-based stains by the detergent.

In this context, the *lipases* which, due to their substrate specificity, ease the removal of triglycerides-containing soils are also worth mentioning. The activity of lipases is highly dependent on temperature and concentration. Studies on model greasy soils have shown that removal of solid fats at very low temperatures (e.g., 20 °C) is largely due to the action of added lipases [202]. Lipases produced economically by applying genetic engineering techniques have been commercially available since 1988. The removal of triglyceride-based fatty stains is mostly evident after several washings and drying cycles. At present lipases are contained as further enzymes in premium branded detergents.

Cellulases are capable of degrading the structure of damaged (amorphous) cellulose fibrils, which exist mostly at the surface of cotton fibers after multicycle washing and using. Cotton textiles treated with a cellulase-containing detergent have a smooth surface. For these reasons cellulases are preferably applied in specialty detergents for delicate fabrics or in color heavy-duty detergents for colored textiles. Cellulases also help prevent redeposition of soil on cotton fibers. As such, garments treated with cellulase-containing detergents look new for a longer time, i.e., whites and coloreds do not look dull or dingy after multicycle washing. Today many detergents contain blends of two, three, or even four enzymes [203].

3.4.2. Soil Antiredeposition Agents, Soil Repellent/Soil Release Agents

The principal characteristic expected of a detergent is that it will cause soil to be removed from textile fibers during the washing process. Removed soil is normally finely dispersed, and if a suboptimal detergent formulation is employed, or too little detergent is used, some may return to the fibers. This is termed a wash liquor showing "insufficient soil antiredeposition capability". The problem becomes especially apparent after multicycle washing as a distinct graying of the laundry, which then looks dull and dingy.

Redeposition of displaced soil can be largely prevented by carefully choosing the various detergent components (surfactants and builders), but addition of special antiredeposition agents is also helpful. Such agents act by becoming adsorbed irreversibly, i.e., in a way that prevents their removal by water, on both textile fibers and soil particles. Adherence of the soil to fibers is thereby hindered. This phenomenon is called soil repellent or soil release effect. Classical antiredeposition agents are carboxymethyl cellulose (CMC) derivatives bearing relatively few substituents. Analogous derivatives of carboxymethyl starch (CMS) have played a similar role. These substances are effective only with cellulose-containing fibers such as cotton and blends of cotton and synthetic fibers. With the increasing replacement of these natural fibers by

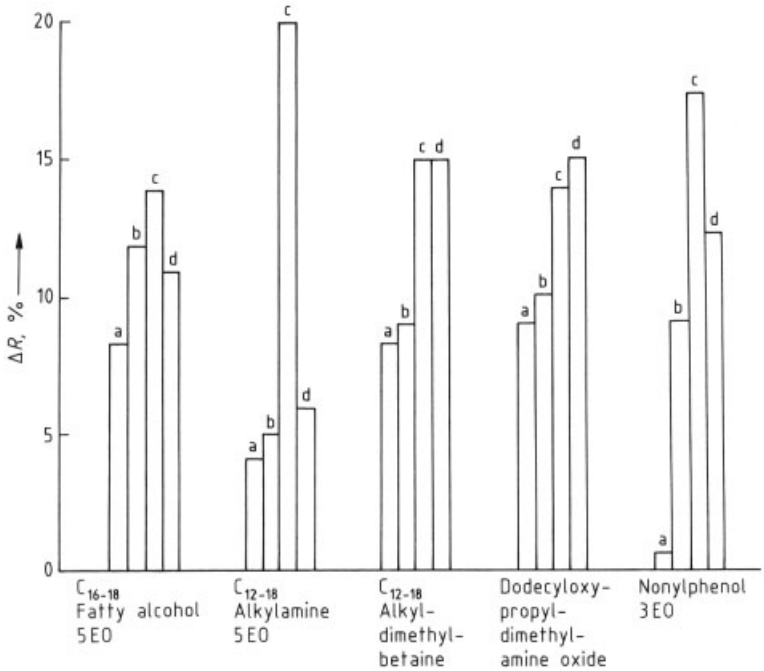

Figure 87. Influence of specific surfactants on the soil antiredeposition capability of detergents [129]
ΔR indicates the improvement in the soil antiredeposition capability conferred by a given additive
a) Cotton, 95 °C; b) Permanent-press cotton, 95 °C; c) Polyester, 60 °C; d) Polyester/cotton, 60 °C
5 g/L Detergent; 0.2 g/L surfactant; number of wash cycles: 3; time: 30 min

synthetics, on which CMC has virtually no effect, the need arose to develop other effective antiredeposition agents and soil repellents. Some surfactants have been found to be well-suited to the purpose (Fig. 87), as are such nonionic cellulose ethers [204, 205] as the following:

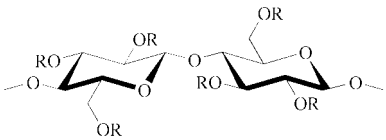

R: CH_3, CH_2CH_3, CH_2CH_2OH, $-CH_2CHOHCH_3$, $-CH_2CH_2CHOHCH_3$

Also anionic derivatives of polymers from terephthalic acid and polyethylene glycol have proved to be very effective soil repellents, particularly on polyester fibers and polyester–cotton blends. They impart hydrophilic properties to these fibers and thus strongly repel oily/greasy soil [206 – 208].

Because of the multitude of fabrics currently on the market, manufacturers supplement most modern detergents with mixtures of anionic and nonionic polymers (e.g., carboxymethyl cellulose–methyl cellulose) and polymers from polyethylene glycol and terephthalic acid and their anionically modified derivatives.

Polymer cobuilders, such as sodium polycarboxylates, and cellulose enzymes have also proved to exert a very good soil antiredeposition effect with cotton fibers.

3.4.3. Foam Regulators

For soap detergents, which are very popular in regions with low per capita income worldwide, foam is understood as an important measure for washing performance. With detergents based on synthetic surfactants, however, soap has lost virtually all of its former significance in the industrialized countries. Nonetheless, most consumers — apart from those using horizontal

axis drum-type washing machines — still expect their detergent to produce voluminous and dense foam. The reason for this seems to be largely psychological (i.e., foam provides evidence of detergent activity and it hides the soil). Consequently, detergents designed for use in vertical-axis washing machines (i.e., products mainly for the non-European market) are quite frequently given the desired foam characteristics by incorporating small amounts of foam boosters. Compound types suited to the purpose include

Fatty acid amides
Fatty acid alkanolamides
Betaines
Sulfobetaines
Amine oxides

In Europe, horizontal axis drum-type washing machines are very common; with such machines, only weakly or moderately foaming detergents are permissible. Thus, foam boosters have lost their former significance in the European market, with the exception of the role they retain in certain specialty detergents (e.g., detergents for hand washables). Especially at high temperature, heavy foaming may cause overfoaming in drum-type machines, often accompanied by considerable loss of active ingredients. Furthermore, large amounts of foam reduce the mechanical action to which laundry is subjected in horizontal axis machines and hence the washing performance. Moreover, excessive foam may result in poor rinsing, spinning, and draining of the washing machine. On the other hand, too little foam must be avoided because consumers have the perception that low foam levels are a sign of poor cleaning performance. A certain level of foam can be beneficial for delicate fabrics to reduce fabric damage. For these reasons, foam regulators — often somewhat incorrectly described as "foam inhibitors" — are commonly added to minimize detergent foaming tendencies.

Ensuring effective foam regulation requires that the regulating system be precisely matched to the other detergent components present. The most difficult cases are surfactants with exceptionally high foam stability, such as alcohol sulfates, alcohol ether sulfates, alkylpolyglycosides, and alkylglucamides. By contrast, an unstable foam, even if present in large amounts, like that produced by LAS or SAS, seldom causes a problem. The principal requirements for a substance to show foam regulating capability are extremely low water solubility and high surface spreading pressure. Typically, an antifoam is a dispersion of hydrophobic solid particles in a hydrophobic liquid (→ Foams and Foam Control). Foam regulators show a wide range of physicochemical properties, but their mechanism of action can usually be explained by assuming that they either force surfactant molecules away from interfaces or they penetrate interfaces that are already occupied by surfactants, thereby creating defects. These defects weaken the mechanical strength of the foam lamellae and cause their rupture.

A great many foam regulators are described in the patent literature, but relatively few have had any real impact. The detergents on the market that are based on LAS/alcohol ethoxylates can be effectively controlled by soaps with a broad chain length spectrum (C_{12-22}). Such soaps have limitations, however, for the following reasons:

1) The foam-depressing activity of soaps is largely due to calcium salts that form during the washing process, with the calcium originating either from hard water or from calcium-containing soil. In other words, satisfactory foam regulation occurs only when a sufficiently high calcium ion content is assured. It is absent when exceptionally soft water is used and when the laundry is only lightly soiled, and foaming problems may result.
2) The foam regulating power of soap is substantially lower with detergents based on anionic surfactants other than alkylbenzenesulfonates (e.g., α-olefinsulfonates, alcohol sulfates, α-sulfo fatty acid esters).
3) When soap is used as a foam regulator, the only complexing agents that can be added as builders are those with limited stability constants; i.e., even in the presence of builders, the calcium ion concentration must remain high enough to permit formation of sufficient lime soap able to act as a foam regulator. An advantage of sodium triphosphate, widely used as a complexing agent in detergents, is that stability constant values for calcium triphosphate complexes are quite low. Other complexing agents (e.g., nitrilo-

triacetic acid, NTA), that have higher stability constants, prevent the formation of foam-regulating lime soaps (Fig. 88).

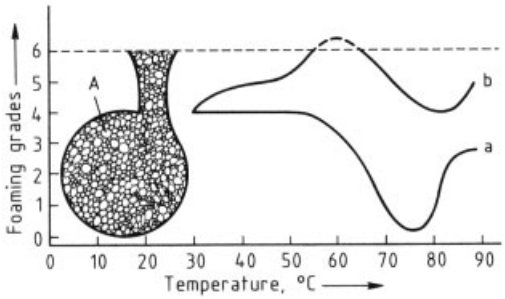

Figure 88. Foaming behavior in a drum-type washing machine [46]
A) Schematic drawing of the inner drum of a drum-type washing machine, including dispenser
a) 7.5 g/L Detergent based on 40 % sodium triphosphate; b) 7.5 g/L Detergent based on 40 % NTA
Foam grading: 0, no foam; 4, foam at the upper edge of the front door; 5, foam in the dispenser; 6, foam overflow

Intensive investigations have shown that specific silica–silicone mixtures or paraffin oil systems are considerably more universal in their applicability and that their effectiveness is independent of both water hardness and the nature of the surfactant–builder system employed [209 – 212]. Therefore, most heavy-duty detergents in Europe have silicone oil and/or paraffins as foam depressors. Soap has almost lost its importance as foam regulator. Silica–silicone systems, frequently called silicone antifoams, are usually commercially available as concentrated powders. They are robust and effective over a wide range of washing conditions for a wide range of laundry products at quite low levels in the formulation (typically 0.1–0.4 %). The key silicone oils used for antifoams are dimethylpolysiloxanes [213]. Silicone antifoam powders are easily processed as postaddition components for detergent manufacturing. It is not recommended to process them in the detergent slurry during production, because the antifoam performance is largely destroyed during spray drying.

3.4.4. Corrosion Inhibitors

Washing machines currently on the market are constructed almost exclusively with drums and laundry tubs of corrosion-resistant stainless steel or with an enameled finish that is inert to alkaline wash liquors or plastic. Nevertheless, various machine components are made of less detergent resistant metals or alloys. To prevent corrosion of these parts, modern detergents contain corrosion inhibitors in the form of sodium silicate. The colloidal silicate that is present deposits as a thin, inert layer on metallic surfaces, thereby protecting them from attack by alkali.

3.4.5. Fluorescent Whitening Agents (see also → Optical Brighteners)

Properly washed and bleached white laundry, even when clean, actually has a slight yellowish tinge. For this reason, as early as the middle of the 19th century, people began treating laundry with a trace of blue dye (blueing agents, e.g., ultramarine blue) so that the color was modified slightly and a more intense visual sensation of improved whiteness was produced. Blueing agents are still popular in some regions of the world. Modern detergents contain fluorescent whitening agents (FWA), also known as *optical brighteners* (→ Optical Brighteners), to accomplish the same purpose. Fluorescent whitening agents are organic compounds that convert a portion of the invisible ultraviolet light into longer wavelength visible blue light. It is well known that the yellowish cast of freshly washed and bleached laundry is a result of partial absorption of the blue radiation reaching it, resulting in reflected light that is partially deficient in the blue region of its spectrum. The radiation emitted by optical brighteners makes up for this deficiency, so that the laundry becomes both brighter and whiter (Fig. 89).

This process of compensating for differences in yellow and blue radiation is fundamentally different from treatment with a blueing agent (washing blue). The latter entails subtractive absorption of yellow light, which results in an overall reduction in brightness (Fig. 90).

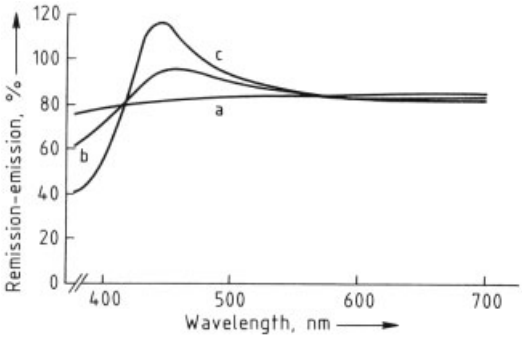

Figure 89. Remission–emission curve of a bleached fabric after application of a fluorescent whitening agent [125]
a) Bleached fabric; b) Fabric treated with small amounts of whitener; c) Fabric treated with large amounts of whitener

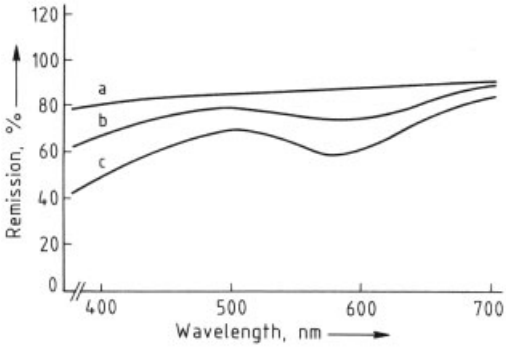

Figure 90. Remission curve of a bleached fabric after application of ultramarine blue [125]
a) Bleached fabric; b) Weakly blued fabric; c) Intensely blued fabric

Limits exist for the extent of whitening that can be achieved by fluorescent whitening agents. This is due in part to the fact that these agents are themselves dyes that exhibit a certain extent of reflectance in the visible region in addition to their emission. Thus, any agglomeration or buildup of fluorescent whitening agent on the fibers eventually becomes apparent. Furthermore, the absorption edge gradually shifts into the visible region of the spectrum, leading to emission at higher wavelengths. These factors can work in combination with the innate color of the fabric, however, and eventually result in a visible shift of the hue (discoloration).

The major commercially available optical brighteners for laundry products are based on four basic structural frameworks: distyrylbiphenyl, stilbene, coumarin, and bis(benzoxazole). Table 20 provides an overview of the most important optical brighteners. Those mentioned here are estimated to cover more than 95 % of optical brighteners for laundry products on the market.

The application of FWA in the wash liquor is essentially a dyeing process (FWA are actually dyes). In the case of cotton and chlorine-resistant FWA, binding occurs through the formation of hydrogen bonds to the fibers. Permanent-press easy-care cotton is generally less susceptible to the effects of FWA. Whitening effects achieved with polyamide whiteners, and particularly polyester whiteners (which tend to be less adsorbed), are due largely to the diffusing power of whitening agent molecules present at the fiber surfaces.

Fluorescent whitening agents are evaluated not only on the basis of their affinity for the relevant fibers, but also on their stability and fastness. "Stability" means their resistance to chemical change during the washing process prior to their adsorption on fibers. "Fastness" refers to their chemical resistance after adsorption. Of primary concern is fastness against light and oxygen, as well as good oxygen stability and, in countries where chlorine bleach is used, fastness and stability to hypochlorite bleaches.

Most detergents currently on the market contain fluorescent whitening agents. A concern to detergent formulators is the difficulty to define optimal FWA type(s) and concentration in laundry detergents for sufficient whiteness maintenance because of the wide variety of washing habits and fiber types used for white textiles [215]. The use of FWA-containing detergents on certain pastel fabrics can cause unwanted color shifts; for this reason special FWA-free detergents are very popular [e.g., color detergents (see Section 4.1.5) and detergents for woolens)].

3.4.6. Dye Transfer Inhibitors

Frequently used dye transfer inhibitors are poly(*N*-vinylpyrrolidone) and poly(vinylpyridine *N*-oxide). These inhibitors form stable complexes with dye molecules that detach from colored textiles and diffuse into the wash bath. As such, the dye transfer inhibitors impede the re-adsorption or redeposition of the dyestuff on other textiles. The result is that white and colored garments look new longer after multicycle

Table 20. Summary of key fluorescent whitening agents for laundry products [214]

Structure	Chemical Name
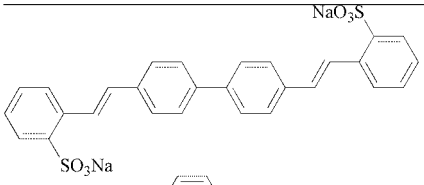	4,4′-di(2-sulfostyryl)biphenyl disodium salt
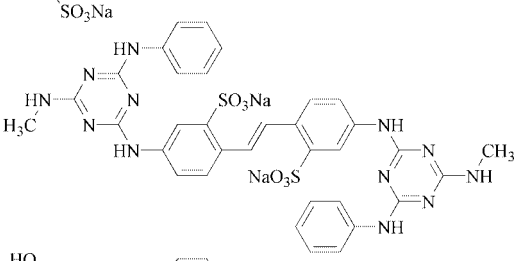	4,4′-Bis[[6-anilino-4-(methylamino)-1,3,5-triazin-2-yl]amino]stilbene-2,2′-disulphonic acid, disodium salt
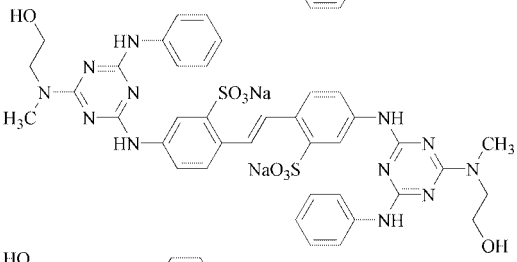	4,4′-bis[[6-anilino-4-[(2-hydroxyethyl)methylamino]-1,3,5-triazin-2-yl]amino]stilbene-2,2′-disulphonic acid, disodium salt
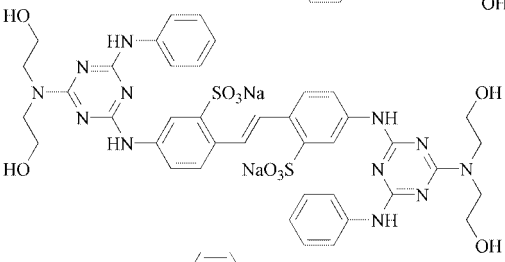	4,4′-Bis[[2-anilino-4-[bis(2-hydroxyethyl)amino]-1,3,5-triazin-6-yl]amino]stilbene-2,2′-disulfonic acid, disodium salt
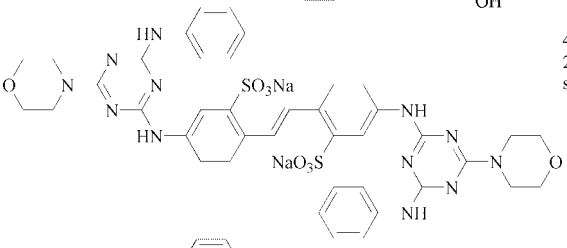	4,4′-Bis[(2-anilino-4-morpholino-1,3,5-triazin-6-yl)amino]-2,2′-stilbenedisulfonic acid, disodium salt
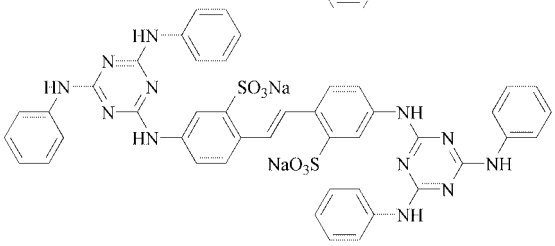	4,4′-bis[(4,6-dianilino-1,3,5-triazin-2-yl)amino]stilbene-2,2′-disulphonic acid, disodium salt
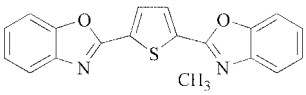	2,5-bis(benzoxazol-2-yl)thiophene
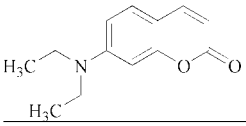	7-diethylamino-4-methylcoumarin

washing. A dull and dingy look of the textiles is avoided.

3.4.7. Fragrances

Fragrances were first added to detergents in the 1950s. Their presence is more than simply a fad or a matter of fashion. Thus, apart from their role in providing detergents with a pleasant odor, an important function of fragrances is to mask certain odors arising from the wash liquor during washing. This is particularly important because most washing machines are installed in the living areas of homes. Fragrances are also intended to confer a fresh, pleasant odor on the laundry itself. For this, long-lasting fragrances on dry laundry, resulting either from detergents or from fabric softeners have become a more and more important factor in the 1990s.

Detergent fragrances are generally present only in very low concentrations, usually $< 1\,\%$ in powder products. They are all complex mixtures of many individual ingredients (Table 21).

Table 21. Hypothetical formulation of a fragrance blend for detergent use

Ingredient	Quantity, g
Cyclohexyl salicylate	200.0
Iso E Super	100.0
Vertofix Coeur	80.0
Boisambrene Forte	50.0
Sandelice	40.0
Dihydromyrcenol	70.0
Benzyl acetone	35.0
Verdyl propionate	30.0
Verdyl acetate	20.0
Isoraldein 70	75.0
Floramat	25.0
O-tert-butylcyclohexyl acetate	20.0
Cironellol	20.0
Geraniol	20.0
Linalyl acetate	50.0
Hedione	50.0
Neroli phase oil	40.0
Linalool	20.0
Diphenyl ether	10.0
Anisaldehyde	5.0
Cyclovertal	3.0
Styrolyl acetate	5.0
Isobornyl acetate	15.0
Ambroxan	2.0
Herbavert	15.0

Several factors are involved in establishing the composition of a fragrance mixture apart from odor and the cost of the often expensive individual components. In particular, the detergent formulation must be taken into careful account, as must the properties of the fabrics to be washed. Chemical stability relative to other detergent ingredients is especially important, as is the limited volatility of the individual fragrances. Temperatures to which the detergent will be exposed during storage must also be considered.

The fragrance industry has at its disposal for the preparation of detergent fragrance oils not only a large selection of synthetic compounds, but also a wide range of natural fragrances. The overall selection is, nevertheless, restricted considerably due to the instability of many fragrances and the fact that some have a tendency to discolor detergents to which they are added. There are also restrictions for toxicological and ecological reasons (cf. Chaps. 10 and 11).

3.4.8. Dyes

Until the 1950s, powder detergents were more or less white, consistent with the color of their components. Thereafter, products were commonly encountered in which colored granules were present along with the basically white powder: certain components had been deliberately dyed to make the finished detergent more distinctive, featuring speckles. Uniformly colored detergents have also appeared on the market, and the idea of introducing coloring agents has become quite common. The preferred colors for both powders and liquids are blue, green, and pink.

There are two important criteria for selecting a coloring agent:

1) Good storage stability with respect to other detergent components and to light and temperature
2) No significant tendency to adversely affect textile fibers through adsorption after multiple washing

3.4.9. Fillers and Formulation Aids

The usual *fillers* for powder detergents are inorganic salts, especially sodium sulfate. Their purpose is to confer the following properties on a detergent:

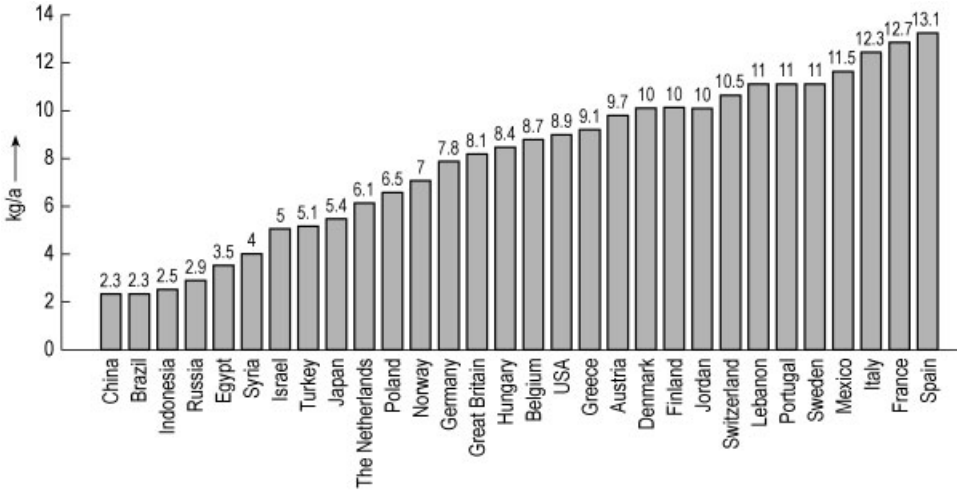

Figure 91. Per capita consumption of detergents in 1996 (kg/a)

Flowability
Good flushing properties
High solubility
No caking of the powder even under highly humid conditions
No dusting

Compact detergents usually contain no fillers.

Formulation aids are materials required in the manufacturing of liquid detergents. The most important of these have the assignment of ensuring through their own hydrotropic characteristics that the other detergent components can be combined in a stable way in an aqueous environment. Above all, these ingredients must prevent phase separation occurring with time, and precipitation occurring as a result of shifts in temperature. Commonly used materials include short-chain alkylbenzenesulfonates (e.g., toluenesulfonate and cumenesulfonate) and, less frequently, urea, as well as low-molecular mass alcohols (ethanol, 2-propanol, glycerol) and polyglycol ethers [poly(ethylene glycols)]. These alcohols are frequently called solubilizers.

4. Laundry Products

Laundry products currently on the market in various parts of the world can be classified into the following groups from a product standpoint:

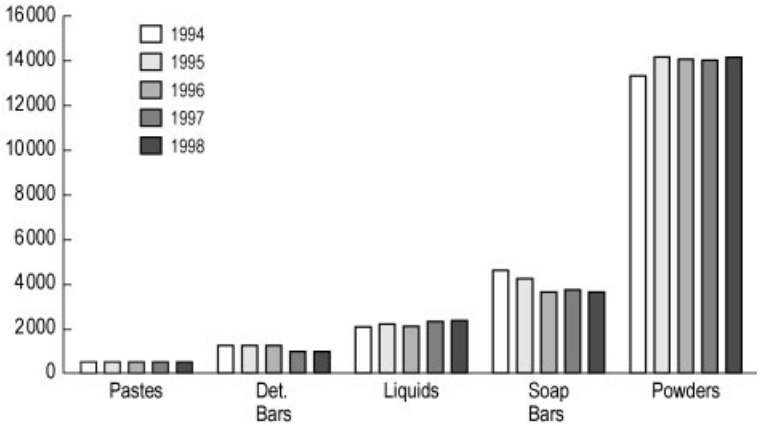

Figure 92. World production of individual heavy-duty detergents (in 10^3 t)
(source: Ciba Speciality Chemicals Ltd., Switzerland)

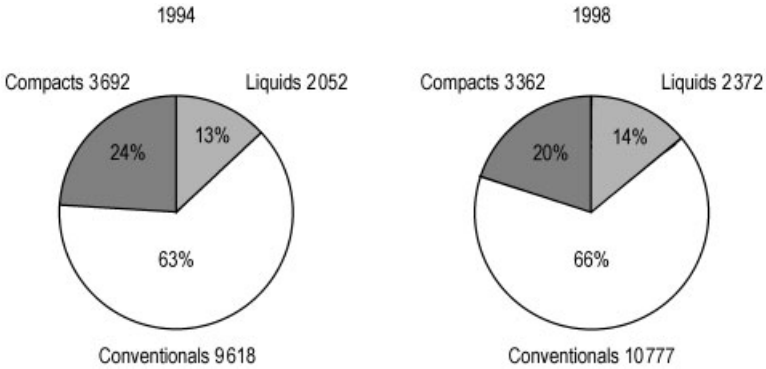

Figure 93. Detergent production, world total (10^3 t)
(source: Ciba Speciality Chemicals Ltd., Switzerland)

Heavy-duty detergents
Specialty detergents
Laundry aids

Quite large differences exist in the per capita consumption of detergents in various countries of the world. As depicted in Figure 91, it amounted to 2–3 kg/a in countries such as Brazil, China, and Russia and to more than 10 kg/a in Mexico and some European countries in 1997 [216, 217].

4.1. Heavy-Duty Detergents

The category of heavy-duty detergents includes those detergent products suited to all types of laundry and to all wash temperatures. They are offered in the delivery forms of conventional powders, extrudates, tablets, bars, and liquids and pastes. Figure 92 indicates that worldwide powders have been the dominant delivery form of detergents over the years. They account for ca. 65 % of the world production of detergents. Compact powders worldwide accounted for 24 % of total detergent powder production in 1998. Depending on quality, product concept, manufacturing process, local standards, local regulations, and voluntary agreements, great differences are found from one formulation to another. For example, detergent manufacturers market laundry detergents with and without phosphate in various regions and countries for ecological and regulatory reasons. Figure 93 provides an overview of global production shares of individual heavy-duty detergent types in 1994 and 1998.

The world production of laundry detergents amounted to 21.5×10^6 t in 1998. Table 22 shows the proportions of the tonnages of individual delivery forms of detergents in 1994 and 1998. As can be seen, the total produced volume remained nearly constant in the 1990s worldwide.

4.1.1. Conventional Powder Heavy-Duty Detergents

Significant differences in composition exist among powder heavy-duty detergents around

Table 22. Worldwide production Detergent Industry

Production, 10^3 t	1994	1998
Soap bars	4579	3651
Detergent bars	1251	968
Powders	13 310	14 137
Pastes	496	505
Liquids	2052	2372
Total	**21 688**	**21 633**
World Average per capita consumption, kg/a	4.5	3.9

(Source: Ciba Speciality Chemicals Ltd., Switzerland)

Table 23. Powder heavy-duty detergent formulations around the world in 1999[a]

Ingredients	Examples	Composition, %				
		USA	South America	Europe	China, India	Japan
Surfactants	alkylbenzenesulfonate, alcohol sulfate, alcohol ethoxylate	10–25	18–25	10–25	8–18	25–40
Builders	zeolite, sodium tiphosphate, Na-citrate, Na-silicate, Na-carbonate/Na-bicarbonate	40–65	40–55	30–55	30–50	40–55
Cobuilders	sodium polycarboxylate	0–5	–	3–8	–	0–5
Bleaching agents	sodium perborate, sodium percarbonate	0–10	–	8–15	–	0–6
Bleach activators	TAED, NOBS	0–3[b]	–	1–7[c]	–	0–3[b]
Antiredeposition agents	carboxymethyl cellulose, cellulose ethers	1–2	0.5–1	0–1	0.5–1	0.5–1
Stabilizers	phosphonates	–	–	0–1	–	–
Foam regulators	soap, silicone oil and/or paraffins	–	–	0.1–4	–	–
Enzymes	protease, cellulase, amylase, lipase	0.3–2	0.3–0.8	0.3–2	0.3–0.8	0.3–1.5
Optical brighteners	stilbene-, biphenyldistyryl derivatives	0.1–0.3	0.1–0.3	0.1–0.3	0.1–0.2	0.1–0.2
Soil repellents	poly(ethylene glycol terephthalate) derivatives	0–1	–	0–1.5	–	–
Fillers/processing aids	sodium sulfate	5–30	20–35	0–30	20–40	5–15
Minors	Fragrance	+/−	+	+/−	+	+
Water		5–15	5–15	5–15	5–15	5–15

[a] All figures expressed as 100 % active material, except for enzymes for which figures relate to % granulate.
[b] Nonanoyloxybenzenesulfonate (NOBS).
[c] Tetraacetylethylenediamine.

the world. An approximate breakdown of detergent formulations for various global regions is provided in Table 23.

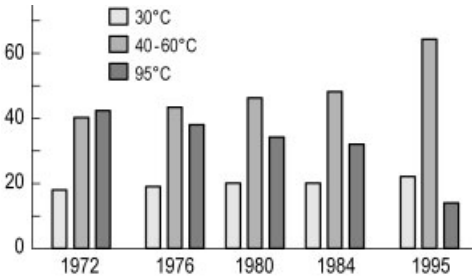

Figure 94. Percentage of wash loads versus washing temperature in Germany (source: Leading German Market Research Institute)

Recommended detergent dosages in Europe generally lie between 3 and 8 g detergent per liter of wash water, whereas values of 0.5–1.5 g/L are usual in the Americas and Eastern and Southern Asia. There are a number of reasons for the latter being so much lower; for example, differences in washing habits (different washing machines, handwashing), softer water, higher bath ratios, separate addition of bleach, e.g., hypochlorite solution or sodium perborate, and different detergent formulations. Detergents in North America, Japan, and Brazil are designed for agitator and impeller type washing machines (cf. Chap. 13). These detergents lack foam suppressors. They sometimes even contain foam boosters instead.

The rise in popularity of ever more colored and easy-care fabrics in the last 30 years, along with the trend toward energy conservation, has resulted in a large decline in the once common European practice of washing at 95 °C, while the 30–60 °C wash temperatures have significantly gained favor. The change is illustrated in Figure 94 for the time period 1972 to 1995 in Germany. A similar trend from hot ($\geq$ 50 °C) to warm (27–43 °C) and cold washing is apparent in the USA (Fig. 95). Detergent manufacturers have responded to these changes by increasing

the content of surfactants that are effective in low-temperature washing, as well as by adding various enzymes, bleach activators, soil repellents, and phosphonates. Detergents have become more and more multifunctional over the last decades, particularly those for use at low washing temperatures.

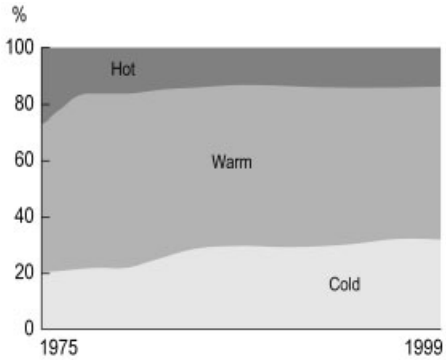

Figure 95. Home laundry washing temperature trends in the USA

4.1.2. Compact and Supercompact Heavy-Duty Detergents

Apart from a few minor and temporary attempts in Japan, Germany, and the USA in the 1970s and early 1980s [145], this heavy-duty detergent category was introduced on a large scale for the first time in Japan in 1987. Within a couple of years it almost completely replaced the conventional powders. The wave of compact detergents reached Europe in 1989. North America followed shortly afterwards [220]. The global introduction of compact and supercompact powder detergents in the late 1980s and early 1990s has had a major impact on detergent categories of large markets such as Japan, Europe, and the USA, as well as on composition and manufacturing processes of heavy-duty detergents [220]. Production processes of concentrated detergents are described in Chapter 6. A comparison of the composition of conventional versus compact heavy-duty detergents in Europe is depicted in Table 24 [216, 217]. For reasons of significantly reduced recommended dosage along with increased detergency performance compact detergents have markedly higher contents of surfactants, active oxygen for bleach, bleach activators, and enzymes than conventional detergents. On the other hand, the total level of alkaline builders has been reduced and the conventional filler/processing aid sodium sulfate has been omitted to a large extent (Fig. 96 [216, 217]). The apparent density of most compact detergents is above 0.75 kg/L. Most of the first-generation compact detergents were produced by the method of downstream compaction of tower powder.

Table 24. Composition of conventional versus compact heavy-duty detergents in Europe, % [216]

Ingredients	Conventional	Compact
Anionic and nonionic surfactants	10–15	10–25
Builders	25–50	25–40
Cobuilders	3–5	3–8
Bleaching agents	10–25	10–20
Bleach activators	1–3	3–8
Antiredeposition agents	0–1	0–1
Corrosion inhibitors	2–6	2–6
Stabilizers	0–1	0–1
Foam regulators	0.1–4.0	0.1–2.0
Enzymes	0.3–0.8	0.5–2.0
Optical brighteners	0.1–0.3	0.1–0.3
Soil repellents	+/−	+/−
Fillers/processing aids	5–30	–
Minors, water	balance	balance
Bulk density, g/L	500–650	600–900

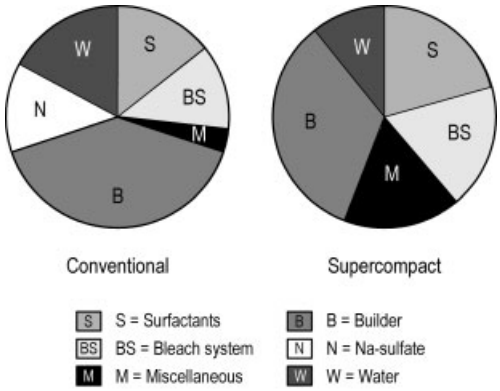

Figure 96. Formulations of heavy-duty detergents [216]

Heavy-duty detergents featuring bulk densities of from 0.8 to 1.0 kg/L are called supercompact detergents or second-generation compact detergents. The prevailing manufacturing methods used are wet granulation/compounding

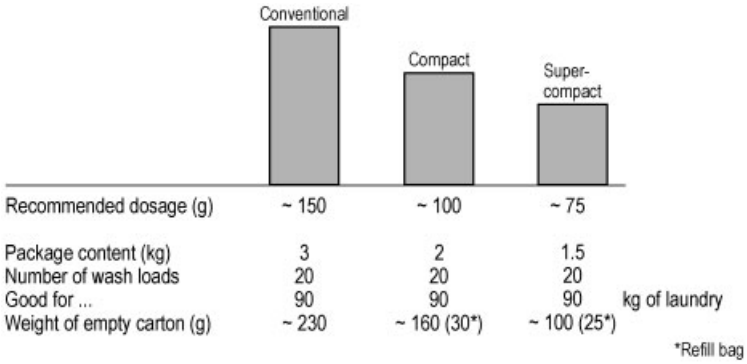

Figure 97. Consumer-relevant characteristics of heavy-duty detergents

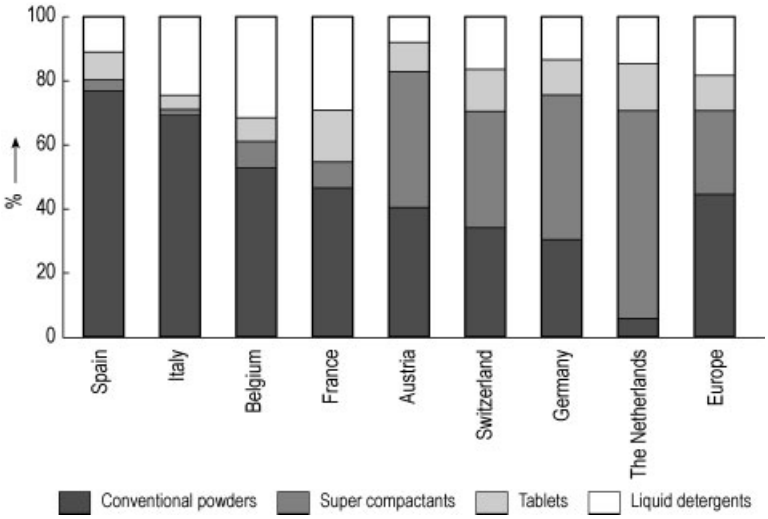

Figure 98. Individual heavy-duty detergents in Europe, market shares (value) 2000 in % [218]

and extrusion. Supercompact detergents have made their way worldwide since 1992 and are meanwhile state of the art in many countries. Consumer relevant properties of (super)compact vs. conventional detergents are depicted in Figure 97 [216, 218]. Market shares of compact/supercompact powder heavy-duty detergents in Europe as compared with conventional powders and liquid detergents are indicated in Figure 98 [218] for 2000. The market shares significantly vary from country to country and from region to region. In the USA compact powders accounted for 80 % of total detergent powders, whereas their share in Japan was even 92 % in the total powder category in 1998.

Sodium triphosphate has been increasingly losing its importance as a key ingredient in detergents in Europe, the USA, and some East-Asian countries since the mid-1980s. 1998 market shares of nonphosphate powder detergents in most European countries are shown in Figure 99 [220]. Triphosphate has been replaced mainly by zeolite A, special silicates, sodium carbonate, specific polycarboxylates, and citrate (see Section 3.2). A key driving force for the increasing market shares of nonphosphate powder heavy-duty detergents has been the proliferation of supercompact detergents [216, 217]. Generally, the use of (super)compact detergents has contributed to decreasing their influx into wastewaters, much to the benefit of the environment [226].

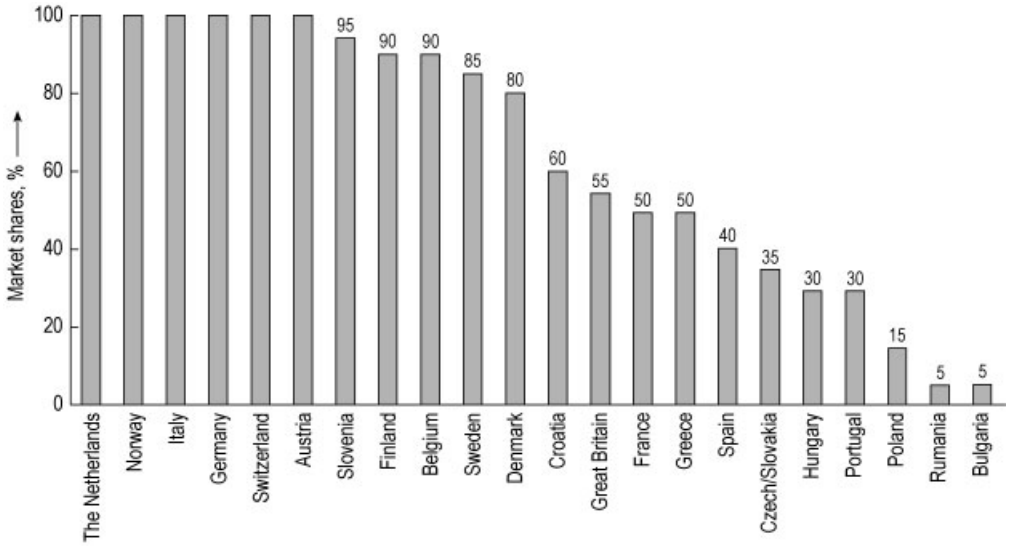

Figure 99. Market shares of nonphosphate HDD powders in Europe in 1998 [220]

4.1.3. Extruded Heavy Duty Detergents [225]

Extruded detergents have been an innovative and unconventional delivery form of second-generation compact detergents. The large-scale extrusion process for detergents was applied for the first time in 1992 and is described in Chapter 6 [227 – 230]. Extruded detergents have since then been marketed under the registered trademark of "Megaperls" by Henkel throughout Europe.

Extruded detergents feature spherical particles of uniform particle size (for example 1.4 mm) whereas conventional spray-dried detergent powders and granulated supercompact detergents are characterized by a broad particle size distribution. The density of an extruded particle is approximately 1.4 kg/L. Extruded detergents have one of the highest densities achieved in detergent manufacturing.

Apart from being an extraordinary basis for manufacturing supercompact products, extruded detergents feature additional advantages such as complete absence of dust particles, very high homogeneity, no segregation of particles, and excellent free-flowing characteristics. Extruded detergents allow anionic surfactant contents of more than 20 % along with very high densities.

4.1.4. Heavy-Duty Detergent Tablets [231 – 234]

Busier consumer lifestyles have increased demand for laundry products that emphasize convenience and ease of use, as exemplified by detergent tablets. Apart from very few regional distributions in the past, detergent tablets were first introduced on a large scale in Europe in late 1997. Ever since they have spread out on the European market and held a share of 10 % in the heavy-duty detergent category by mid-2000. The Japanese, Canadian, and U.S. markets saw their introduction during the year 2000. In the wide field of laundry products, tablets meanwhile also cover the categories of bleach boosters and water softeners in Europe.

Consumer-relevant characteristics of tablets are ease of dispensing and convenience in handling, i.e., no dosing and dispensing aids are needed. Other advantages include precise dosing, smaller packages as compared with powder products due to the highly concentrated form, extra portability, and more accurate sense of how many washes remain in the detergent box.

Tablets are the most compact delivery form of nonliquid detergents. They have densities of 1 to 1.3 kg/L. This results in benefits in terms of lower packaging volume, ease of transportation and storage, and reduced shelf space. Detergent

tablets therefore belong to the supercompact detergent category (see Section 4.1.2).

One of the key requirements of laundry detergent tablets is fast disintegration. Upon their first contact with water, be it in the washing machine dispenser of a European drum-type washer or in the basket of top-loading agitator (North America) or pulsator machines (East Asia), tablets must disintegrate within seconds or at least within a minute or so. This is a necessary prerequisite for the next step of dissolving, particularly with regard to the short washing times used in Japan, the USA and, recently, Europe. Special precautions must therefore be taken to avoid intermediate formation of viscous liquid-crystalline phases (gelling effect) of the surfactants in the tablet detergent during contact with water. To enable fast disintegration laundry detergent tablets generally contain special disintegrants. These can be classified into four groups:

1) Effervescents such as carbonate/hydrogencarbonate/citric acid
2) Swelling agents like cellulose, carboxymethyl cellulose, and cross-linked poly(*N*-vinylpyrrolidone)
3) Quickly dissolving materials such as Na (K) acetate or Na (K) citrate
4) Rapidly dissolving water-soluble rigid coating such as dicarboxylic acids

or combinations therof.

Another specific demand made on heavy-duty detergent tablets is sufficient hardness to enable handling during packaging, transportation, and in-home use. The conflicting desired properties of sufficient hardness and fast disintegration have to be well balanced. Usually, laundry detergent tablets have weights of ca. 35 to 45 g and diameters in the range of 40 to 45 mm. The most common recommended dosage in Europe is two tablets per wash cycle. One- and two-phase (colored) tablets are available on the market. Two-phase tablets allow separation of certain detergent ingredients that otherwise might adversely affect each other during storage, for example, enzymes and activated bleach.

In terms of their composition, heavy-duty detergent tablets differ from other supercompact detergents mostly with regard to the disintegrants. Table 25 lists the composition of European detergent tablets. All compact heavy-duty detergents in Europe have been nonphosphate zeolite-based thus far [216]. The reappearance of phosphate in the detergent tablet category in Europe (see Table 25) may therefore be considered a remarkable development [231].

Table 25. Composition of heavy-duty detergent tablets on the European market [231]

Ingredients	Zeolite-based, %	Phosphate-based, %
Surfactants	13–18	15–18
Bleaching agents	13–15	12–16
TAED	3–7	4–7
Zeolite	11–25	
Sodium triphosphate		25–47
Layered silicate	0–9	–
Sodium polycarboxylate	2–3	2
Disintegrants	5–17	0–12
Enzymes	2–4	2–3

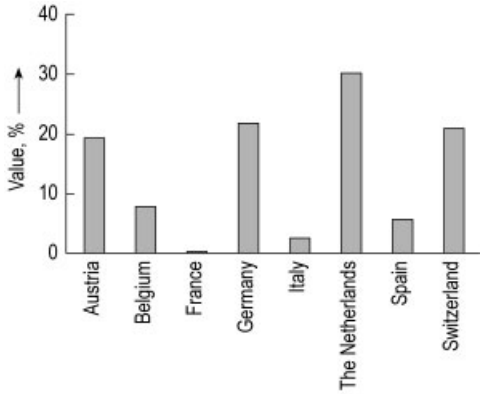

Figure 100. Market Shares (value) of Color Heavy-Duty Detergents in the Total HDD Category (source: Leading International Market Research Institute)

4.1.5. Color Heavy-Duty Detergents

In 1991, a new category of heavy-duty detergents appeared on the West European market, for which the name color detergents was coined. So far, this type of compact detergent has mainly spread out on the market in Western Europe. 1999 market shares of color detergents in some countries on the European continent are shown in Figure 100. Color detergents differ from conventional European heavy-duty detergents in that they contain neither bleach nor optical brighteners and in that they feature specific dye transfer inhibitors such as poly(*N*-vinylpyrrolidone) or poly(vinylpyridine *N*-oxide).

Color detergents are recommended and used for washing colored laundry, for which improved color care is claimed.

4.1.6. Liquid Heavy-Duty Detergents

Liquid detergents have been common in the USA since the 1970s for a variety of reasons. Today, the USA is the largest single market of liquid heavy-duty detergents (HDL) in the world. The market share of HDL in the USA was some 50 % in 1999. Until 1987, such products were insignificant in Europe, although liquid heavy-duty detergents had been available on the European market since 1981. In 1998, liquid heavy-duty detergents represented 12 % of the market in Europe, and their worldwide share was 14 % of total detergent production.

Liquid heavy-duty detergents are distinctive because of their usually high surfactant content (up to ca. 50 %). They rarely contain builders such as zeolite or triphosphate and are generally devoid of bleaching agents, the latter for reasons of loss of active oxygen and incompatibility with enzymes during storage. They are most effective in removing greasy and oily soil, especially at wash temperatures below 60 °C. A comparison of the detergency characteristics of European liquid and powder heavy-duty detergents is depicted in Table 26 [119]. From the characteristics of the two types of detergent it is obvious that both products have different but distinct fields of optimal performance. Depending on washing habits, HDL are alternatives to powders especially for washing in the lower temperature range [235].

HDL are marketed both with and without added builders. Formulations for liquid heavy-duty detergents with (structured) and without builders (unstructured) in Europe are given in Table 27 [216]. Liquid detergents with builders typically contain 15–30 % surfactants; the total amount of surfactants in products without builders can be as high as 50 %. The unstructured liquid detergents have continued to dominate the market. Besides having a high content of synthetic surfactants, they include a large proportion of soap with a defined carbon chain length as a water-softening component, as well as ingredients needed as formulation aids. The total content of washing active components can be as high as 60 %. The liquid products still contain water and are free of bleaching agents. Although many patents have been filed for liquid detergents that contain bleaching agents, none have yet been marketed successfully. In common with the increasing multifunctionality of powders, liquid heavy-duty detergents show a steady trend towards a higher enzyme content and incorporation of more types of enzymes.

Table 26. Performance of powder versus liquid heavy-duty detergents

Product characteristics	Powder 60–90 °C	Liquid 30–60 °C
Removal of		
Fatty/oil soil	++	+++
Particulate soil	++	+
Bleachable stains	+++	++
Proteinaceous stains	++	++
Detergency performance on		
Cotton	++	+
Cotton/polyester blends	++	++
Synthetics	+	++
Wool	+	++
Antiredeposition/whitening		
30 to 60 °C	++	++
90 °C	++	+
Prevention of inorganic salt deposits	+	++
Fabric care	+	++
Solubility/dispersability	+	+++

+ = excellent;
++ = very good;
+++ = good

Table 27. Composition of liquid heavy-duty detergents in Europe (2000), %

Ingredients	Unstructured	Structured
Anionic surfactants	7–18	10–25
Nonionic surfactants	15–30	6–10
Soaps	10–22	4–6
Builders	0–8	15–30
Solubilizers	0–12	0–5
Alcohols	8–12	0–5
Enzymes	0–2.5	0–1.5
Optical brighteners	0.05–0.25	0.05–0.25
Stabilizers	+/–	+/–
Fragrance	+	+
Minors, water	30–50	30–50

Increasing concentrations of total active ingedients have been introduced in liquid detergents in the first half of the 1990s, although the degree of compaction has not been as drastic as with powdered products. Conventional liquid heavy-

duty detergents still dominate the market in some European countries.

The physical appearance of liquid laundry detergents remained unchanged for a long time. They were marketed as low-viscosity (250–300 mPa · s) products with a Newtonian flow behavior. This changed with the launch of detergent gels in Europe. Gel-like pourable heavy-duty detergent concentrates that feature a specific rheological behavior represent a new delivery form in the liquid segment in Europe [236]. These gel detergents have much higher viscosities than conventional HDL (ca. 1500–3000 mPa · s). Gels were introduced in almost all European markets in 1997. Because of their high viscosity and rheological properties gels show improved detergency performance in pretreating stains as compared with conventional HDL [236]. Due to the continuous increase in the use of liquid concentrates and the reduction of the recommended dosage, it has been possible to reduce package sizes without decreasing the number of wash loads per package. This has resulted in large savings on packaging material and transportation costs.

4.2. Specialty Detergents

Specialty detergents play a relatively minor role outside Europe, but they are quite important in Europe.

Specialty detergents are products developed for washing specific types of laundry. Such detergents are generally used in washing machines but also for hand washing. They usually require the use of special washing programs (e.g., a wool cycle to prevent felting or a gentle cycle to prevent wrinkling). They can be characterized by the fact that fabric care ranks high and is a more important feature than with heavy-duty detergent for which soil and stain removal is the key characteristic.

4.2.1. Powder Specialty Detergents

Distinction is made among products of the following types:

Detergents for delicate and colored laundry
Detergents for woolens
Detergents for curtains
Detergents for washing by hand

Many of these products are appropriate for either machine or handwashing. Most of them are marketed with an average bulk density of 0.4–0.5 kg/L.

Detergents specially designed for *delicate and colored laundry* contain neither bleach nor fluorescent whitening agent which may adversely affect sensitive dyes used with some of these fabrics. These detergents are particularly useful for laundry colored with dyes sensitive to oxidation or for pastel fabrics that otherwise might experience color shifts if treated with fluorescent whitening agents. Most products today contain cellulases, which help fabrics look new longer and keep colors bright. Dye transfer inhibitors are added to some specialty detergents.

Detergents for *wool* are primarily intended for use in washing machines. Particular care must be exercised to prevent damage to the sensitive fibers that constitute natural wool, including applying low-temperature washing, short washing times, high bath ratios, and avoiding vigorous mechanical action.

High-foaming detergents for *washing by hand* are intended for washing small amounts of laundry in the sink or in a bowl. Due to the increased number of household washing machines, the use of specialty handwashing detergents has been gradually declining in Europe through the last decades. Their share was ca. 3 % in the entire detergent category in 1999.

Table 28 provides the compositions of various types of specialty detergents.

4.2.2. Liquid Specialty Detergents

Liquid specialty detergents have been on the market in Europe for a long time. Some liquid specialty detergents are intended for handwashing. However, most specialty liquid products are intended for washing machine application, e.g., detergents for woolens. The latter may be free of anionic surfactants, in which case they usually contain mixtures of cationic and nonionic surface active agents. The cationic agents act as fabric softeners to help keep wool soft and fluffy. Table 29 describes typical formulations for liquid specialty detergents.

Table 28. General formulations of West European powder specialty detergents

Ingredients	Composition, %			
	Detergents for delicate textiles	Detergents for woolens	Detergents for curtains	Handwashing detergents
Anionic surfactants (alkylbenzenesulfonates, fatty alcohol ether sulfates)	5–20	10–20	10–16	12–25
Nonionic surfactants (alcohol ethoxylates)	2–10	2–20	2–7	1–4
Soaps	1–5	0–5	2–6	0–5
Cationic surfactants				0–5
Zeolite A	20–40	5–40	25–40	7–30 *
Sodium perborate			0–15	
Sodium silicate	0–7	0–3	2–5	2–9
Sodium polycarboxylate	2–5	2–5	2–5	2–5
Antiredeposition agents	0.2–1.5	0–1.5	0.5–1.5	0.5–1.5
Enzymes	0–1			0–1
Optical brighteners	0–0.2		0.1–0.2	0–0.1
Dye Transfer Inhibitors	0–0.5			
Fragrances	+	+	+	+
Fillers and water	balance	balance	balance	balance

* Some hand washing detergents contain 15–30 % sodium triphosphate instead.

Table 29. Frame formulations of European liquid detergents for woolens

Ingredients	Composition, %	
	With incorporated softeners	Without incorporated softeners
Anionic surfactants (alkylbenzenesulfonates, fatty alcohol ether sulfates)	0–1	10–15
Nonionic surfactants (alcohol ethoxylates, fatty acid amides)	15–20	0–2
Cationic surfactants (esterquats)	2–4	–
Solubilizers (ethanol, propylene glycol)	0–5	0–5
Hydrotropes (toluenesulfonates, xylenesulfonates, cumenesulfonates)		0–3
Fragrances, dyes	+	+
Water	60–70	60–80

4.3. Laundry Aids

Laundry aids are products developed to meet the varying needs of often widely divergent washing habits and practices in different parts of the world. For example, most powdered heavy-duty detergents in Europe contain oxygen bleach. Most in the Americas, Africa, and Asia do not, and hence separate addition of bleach may be required. Depending on their use, laundry aids can be divided into the following categories:

Pretreatment aids (prespotters, water softeners)
Bleaches
Detergent boosters
Fabric softeners
Stiffeners
Laundry dryer aids
Refreshing products for dryer application
Odor removers

The global laundry aids market was worth \$ 38.7×10^9 in 1998. Western Europe accounted for 30.4 %, while the USA ranked second with 20.8 % of global sales [237].

4.3.1. Pretreatment Aids

The most important pretreatment aids are soil and stain removers and water softeners. Special presoaking agents that were frequently applied in the past have gradually lost their market importance in the industrialized countries.

Table 30. Frame formulations of soil and stain removers in Europe and the USA

Components	Composition, %				
	Europe		USA		
	Pastes	Sprays	Liquid/sprays	Sticks	Gels
Anionic surfactants (alkylbenzenesulfonates, alcohol sulfates, soaps)	15–30		0–3	10–15	0–0.5
Nonionic surfactants (nonylphenol ethoxylates, alcohol ethoxylates, amine oxides)	3–10	15–40	1–15	25–35	8–45
Hydrotropes (sodium xylenesulfonate)			0–10	0–5	0–5
Builders (sodium citrate, sodium carbonate)			0–3.5	1–2	0–2
Oxidants (hydrogen peroxide)			0–3.5	–	–
Bleach activator			0–0.5	–	–
Solvents, solubilizers		30–80	0–12	25–35	0–35
Enzymes			0.2–1	2–3	0–2
Dyes, fragrance, stabilizers, water	balance	balance	balance	balance	balance

Soil and stain removers , also called prespotters or pretreatment aids, are frequently products with high surfactant contents. Their direct application helps removing greasy and/or bleachable stains (Table 30). Such products are applied to soiled areas prior to washing. They are supplied as pastes, sticks, or aerosols or in trigger spray bottles. The major targets of such products are fabrics that require low-temperature washing. The regular use of soil and stain removers can also help prevent the buildup of soil on problem areas such as shirt collars and cuffs after multicycle wear/washing.

Spray stain removers mainly consist of mixtures of solvents and surfactants. Solvents similar to those employed in dry cleaning loosen grease so that it can subsequently be emulsified by appropriate surfactant blends. Sprays generally work more rapidly than paste products. Proper application of such products prevents formation of undesirable rings around stains, a consequence that is nearly unavoidable when stains are subjected to local treatment with pure solvents instead.

Water softeners are used both in the USA and in Europe. European manufacturers recommend that these softeners be used in combination with detergents for applications involving water of medium to high hardness. However, the great majority of heavy-duty laundry detergents contain sufficient builders and cobuilders to prevent build-up of lime deposits upon multicycle washing even in hard water. European water softeners generally contain ion exchangers, usually zeolite A and polymeric carboxylic acids (Table 31).

Table 31. Frame formulation of water softeners in Europe

Ingredients	Composition, %
Zeolite A	45–60
Sodium polycarboxylate	10–15
Sodium silicate	10–15
Sodium sulfate, minors, water	balance

4.3.2. Boosters

Boosters are products that can be added separately to detergents to exert specific influences on the washing process and thereby improve its effectiveness. The principal types of boosters offered on the market are bleaching agents and laundry boosters.

Bleaching agents are found worldwide on the market. They have been common in the Americas and Asia for a long time, where they are available in both powder and liquid

form. Powder bleaching agents generally contain sodium perborate or sodium percarbonate (Table 32), whereas most liquid bleaching agents are dilute solutions of sodium hypochlorite or hydrogen peroxide. Bleaches are usually applied optionally when needed along with detergent. Sodium hypochlorite is to be used for whites only, whereas hydrogen peroxide solutions and powder and tablet peroxygen bleaches (perborate/percarbonate) are considered color-safe products.

Table 32. Frame formulations of all-fabric bleaches in the USA

Ingredients	Composition, %	
	Powders	Liquids
Builders (sodium carbonate, zeolite, sodium citrate, sodium silicate)	60–75	–
Anionic surfactants (alkylbenzenesulfonates, alcohol sulfates, alcohol ether sulfates)	5–15	2–10
Nonionic surfactants (alcohol ethoxylates, amine oxides)	0–trace	0.3–5
Bleach (sodium perborate, sodium percarbonate, hydrogen peroxide)	3–15	3–4 *
Fillers (sodium sulfate, sodium chloride)	5–20	0–2
Fluorescent whitening agents	0.2–0.5	0.1–0.2
Enzymes	0–1	–
Dyes, fragrances, pH-regulators	+	+
Water	balance	balance

* Hydrogen peroxide.

Table 33. Frame formulations of laundry boosters in the USA

Ingredients	Composition, %	
	Powders	Liquids
Builders		
Sodium carbonate	30–40	–
Sodium borate	0–50	0–1
Sodium silicate	0–5	–
Anionic surfactants (alkylbenzenesulfonates, alcohol sulfates, alcohol ether sulfates)	0–10	0–5
Nonionic surfactants (alcohol ethoxylates, amine oxides)	–	0–1
Hydrotropes	–	+
Fluorescent whitening agents	0–1	0–0.5
Enzymes	0–2	0–1
Dyes, fragrances, neutralizing agents	+	+
Water	balance	balance

Laundry boosters are marketed largely in the USA; various formulations exist. They contain either sodium silicate, sodium citrate, sodium borate, or sodium carbonate, usually in combination with surfactants. Enzymes are often present as well (Table 33).

4.3.3. Aftertreatment Aids

After the washing process has been completed and soil removal has been accomplished, fabrics are sometimes subjected to some type of aftertreatment. The goal is to increase the usefulness of laundry by restoring textile characteristics that have suffered in the course of the wash. Needs in this respect can vary considerably, depending on the fabric involved. Thus, products in this category may be called on to provide elastic stiffness, improved fit and body (shirts and blouses); smoothness, sheen; good drape (curtains); fluffiness and softness (underwear, towels, bath gowns); or antistatic properties (garments, underwear, hosiery, and other easy-care articles made from synthetic fibers). To achieve such effects, the following product groups are marketed:

Fabric softeners
Stiffeners
Laundry dryer aids

4.3.3.1. Fabric Softeners

Textiles washed by machine are subjected to greater mechanical stress than those washed by hand. Indeed, machine-washed laundry may be so severely jammed that the pile of the fibers at the fabric surface is reduced to an extreme state of disarray, especially in the case of natural fibers such as cotton and wool. During subsequent drying in static air (e.g., when laundry is dried hanging indoors), this condition tends to become fixed in the fabrics, and the laundry acquires a harsh feel. Addition of a liquid fabric softener in the final rinse (rinse-cycle softener) results in fabrics that feel softer.

The conditions described above are frequently found in Europe, Japan, and other regions of the world. Fabric softeners play another role in the USA, where laundry is mostly dried in mechanical dryers. The tumbling of the laundry in the dryer accomplishes its own softening effect. The chief task of household fabric softeners in the USA is rather to impart antistatic properties and a pleasant odor to the laundry. Therefore,

Table 34. Sorption of distearyldimethylammonium chloride * [147]

Textiles	Adsorbed amount **			
	mg/g	mol/g	mg/m^2	%
Wool	1.20	2.06	301	100
Resin-finished cotton	1.18	2.01	133	98
Cotton	1.17	2.00	169	96
Polyester/cotton	1.17	2.00	110	98
Polyamide	0.96	1.63	66	79
Polyacrylonitrile	0.90	1.53	68	74
Polyester	0.57	0.97	109	47

* Equilibrium conditions: time = 60 min, t = 23 °C, bath ratio = 1 : 10; initial concentration: 120 mg/L.
** Based on the mass and geometric surface of the textiles.

fabric softeners in the USA are frequently applied as sheets impregnated with active material which are added to the moist laundry at the beginning of the dryer cycle (see Section 4.3.3.3)

The principal active ingredients in commercially available rinse-cycle softeners are usually cationic surfactants of the quaternary ammonium type [149 – 152, 244, 245]. When applied in appropriate concentrations, cationic surfactants are adsorbed nearly quantitatively onto natural fibers, in contrast to their behavior with synthetic fibers (Table 34) [221 – 223].

To prevent undesired interactions between the anionic surfactants of a detergent and the cationic surfactants of a fabric softener, the latter must be introduced only in the last rinse cycle. Overuse of fabric softener must be avoided. Otherwise, the absorbency characteristics of the textiles decrease which in turn adversely affects the function of towels.

Conventional softeners, which contain on average 4–8 % active material, have been partially replaced in many countries by softener concentrates having some 12–30 % active material. This development is an answer to the increasing public environmental criticism centered on the bulky plastic packaging of conventional softeners. Bottles with volumes of 4–6 L of conventional fabric softener need large amounts of plastic for their manufacture, and also present waste disposal problems.

Due to its poor biodegradability the formerly widespread distearyldimethylammonium chloride (DSDMAC) [149], the former active material of most softeners in Europe, the USA, and Japan, has been replaced by the readily biodegradable (see Section 10.5.1) esterquats in the 1980s and 1990s. Frame formulations of liquid fabric softeners are given in Table 35.

The formulation of products with higher amounts of active material requires a well-balanced system of selected emulsifiers to maintain good dispersion stability upon storage.

Fabric softeners have additional benefits. They impart good antistatic properties on fabrics. They prevent the build-up of electrostatic charges on synthetic fibers, which in turn eliminates such unpleasant phenomena as fabric cling during handling and wearing, crackling noises, and dust attraction. Also, fabric softeners make fabrics easier to iron and help reduce wrinkles in garments. In addition, they reduce drying times so that energy is saved when softened laundry is tumble-dried. Last but not least, they also impart a pleasant fragrance to the laundry. That is why branded softeners are frequently offered with different fragrances.

4.3.3.2. Stiffeners

If stiffness and body are desired rather than soft and fluffy laundry, stiffeners can be added as aftertreatment. Usual agents for this purpose include natural starch derived from rice, corn, or potato, which can be used to impart extreme stiffness to fabrics. Synthetic polymeric stiffeners are good alternatives. They impart a more modest degree of stiffness than natural starch does, which is more closely attuned to contemporary taste. Products of the latter type are generally liquid and are easier to apply than natural starch. Stiffeners are supplied as dispersions and contain, in addition to a small amount of starch, substances such as poly(vinyl acetate), which has been partially hydrolyzed to poly(vinyl alcohol). Table 36 shows a frame formulation for typical European liquid stiffeners. Such products are offered not only as dispersions, but also

Table 35. Formulations of fabric softeners in Europe, the USA, and Japan

Ingredients	Composition, %						
	Europe			USA			Japan
	Regular	Concentrate	Conditioner	Regular	Concentrate	Conditioner	
Quaternary ammonium compounds[a]	3–11	11–20	11–26	3–11	10–30	15–35	0–20
Aminoamide	0–2	0–2	0–2	0–2	0–2	0–2	0–5
Nonionic surfactants	0–0.8	0–3	0–3	0–0.8	0–3	0–3	0–5
Additional softening compounds[b]	0–1	0–4	0–4	0–1	0–4	0–4	0–1
Preservatives, antibacterial agents	0–0.5	0–0.3	0–0.1	0–0.5	0–0.3	0–0.1	0–2
Dye fixatives	0–0.2	0–0.2	0–0.2	0–0.2	0–0.2	0–0.2	
Ironing aids[c]	–	–	0–6	–	–	0–6	0–5
Dyes, fragrances	+	+	+	+	+	+	+
Ethanol/isopropanol solvents	0.2–2.5	0.5–3	0.5–4	0.2–2.5	0.5–3	0.5–4	0–3
Diol solvents	–	–	0–12	–	–	0–15	0–10
Viscosity regulators[d]	0–0.2	0–0.6	0–0.7	0–0.2	0–0.6	0–0.7	0–0.5
Water	balance	balance	balance	balance	balance	balance	balance

[a] Esterquat type: triethanol amine, diethanol amine, or epichlorohydrine esterified with tallow-based acids or oil-based acids are preferred in Europe. Diesters are preferred due to superior softening performance. Aside from esterquats, distearyl dimethyl ammonium chloride (DSDMAC) types, imidazolines, and amino ester salts are used in some countries.
[b] Fatty alcohols, fatty acids, triglycerides.
[c] Silicone oils, dispersed polyethylene.
[d] Magnesium chloride, calcium chloride, sodium chloride, sodium acetate.

as aerosol sprays. The latter allow local treatment of garments, for example, collars and cuffs.

Table 36. Formulation of liquid stiffeners in Europe

Ingredients	Composition, %
Alkyl/alkylaryl polyglycol ethers	0.1–2
Poly(vinyl acetates) (partially hydrolized)	10–30
Starch	0–10
Poly(ethylene glycols)	0.5–1.5
Fluorescent whitening agents	+/-
Dyes	+/-
Water	balance

Synthetic polymeric stiffening agents are called permanent stiffeners because, unlike starch, their effectiveness endures through several wash cycles. This property also has its negative aspects, however, since poly(vinyl acetate) film on a fabric can attract both soil and dyes, thereby leading to discoloration.

4.3.3.3. Laundry Dryer Aids

Laundry dryers are far more widespread in the USA than in Europe. They have held a significant place in the U.S. market since the early 1970s. Dryer aids are introduced into the dryer along with the damp, spin-dried laundry. During the drying process, they impart to the laundry both softness and a pleasant odor. Most importantly, however, they prevent static buildup on the fabric. The latter point is particularly significant in the USA, where synthetic fabrics are quite frequently used. Synthetic fabrics are less popular in Europe and elsewhere.

Laundry dryer aids are almost exclusively applied as sheets, which serve as carrier. The sheets are made from nonwoven material and impregnated with both fabric softeners and temperature-resistant fragrance oils. During the drying process, active ingredients from the dryer sheet are transferred to the laundry through frictional contact. The sheet materials are designed for single use and are discarded at the end of the drying cycle.

An additional reason for the popularity of dryer aids in the USA is the fact that washing machines in the USA often lack dispensers for automatic addition of liquid fabric softeners to the rinse cycle.

4.3.4. Other Laundry Aids

4.3.4.1. Refreshing Products for Dryer Application

Refreshing products form a new subsector in the laundry aids category. They were introduced in the USA and Europe in the late 1990s. These products are used for dry-clean and handwashable clothes. They do not replace professional dry cleaning, but they allow the consumers to refresh their garments any time at home. The custom products have two main features: a refreshing and a spot-cleaning action. The refreshing action is achieved inside the tumble dryer by using a moist impregnated sheet along with the item to be treated. Some systems use a plastic bag for producing a damp microclimate inside the dryer, whereas other systems work without a bag. In both cases the water vapor removes unpleasant odors from the garments. For spot cleaning either the moist sheet or a separate stain remover liquid is used in conjunction with an absorbent layer of cloth or paper placed underneath the stained garment to pick up dissolved soil.

4.3.4.2. Odor Removers for Washer Application

Laundry odor removers are liquid products. They were introduced on the U.S. market in the year 2000. Their formulation and technology allows the removal of (unpleasant) odors from laundry. The product is dispensed directly into the washer along with the detergent at the beginning of the wash cycle. Their main active component is based on cyclodextrins.

5. Industrial and Institutional Detergents

Even though the principles that determine the effectiveness of detergents for household and professional laundries are the same, detergents for large-scale institutional use generally differ insofar as they must be designed to meet the special circumstances associated with laundry on an industrial scale [224]. In contrary to home laundering, professional laundries have to deal with large volumes of textile garments and require therefore completely automatic processing with microprocessor-controlled machines and dosing units. They also need data retrieval systems for complete process control. The washing process is in general much shorter and is operated with soft water. For environmental and economic reasons energy and water recycling is essential. Since soil levels are commonly much higher in professional than in household laundry and the variety of textiles is very large, concentrated products, special bleaches, and stain-removal processes are often applied. In many cases disinfection of the laundry load is necessary. Specific packaging for laundry products in large quantities and safe storage systems are frequently required. On the whole, a professional laundry process is a thorough balance between economy, ecology, quality, and safety.

A large variety of products for professional laundries is available today. These products belong to one of the following three major categories:

Base detergents
Detergent additives
Specialty detergents

Table 37 shows the formulation of these products for institutional use.

Base Detergents. In principle, such products are completely built detergents and they are primarily responsible for a good basic detergency. Base detergents are usually highly alkaline and contain no bleach. The major ingredients are surfactants, builders (either phosphate or zeolite with polymer cobuilder), alkalies, fluorescent whitening agents, complexing agents, antiredeposition agents, and enzymes. Currently, they are used quite commonly in continuous batch washing machines with counterflow systems. Oxygen or chlorine bleach can be added depending on the degree of soiling. The addition of other detergent additives such as disinfecting agents or enzyme boosters is optional. Base detergents are available in form of powders, liquids, and pastes.

Detergent Additives. In order to improve the removal of heavy soilings and tough stains, which are quite common in professional laundries, detergent additives are frequently used. Special products like surfactant,

Table 37. Formulations of various types of detergents for institutional use

Components	Detergents		Partially built products		
	Base	Specialty	Surfactant boosters	Bleaching agents	Enzyme boosters
Surfactants	×	×	×		×
Sodium triphosphate or zeolite/polycarboxylate	×	×			
Alkalies (soda ash, metasilicate)	×	×	×		
Bleaching agents				×	
Fluorescent whitening agents	×		×	×	
Enzymes	×	×			
Complexing agents (phosphonates)	×			×	
Antiredeposition agents	×	×			×

alkaline, bleach, and enzyme boosters are commonly available. One of the most important additives is the disinfectant for hospital use. Such disinfectant additives are specifically designed to handle laundry originating from isolation wards and therefore require thermal and chemothermal sterilization. Common disinfectants for this purpose include peracetic acid–hydrogen peroxide mixture, sodium *N*-chloro-*p*-toluenesulfonamide (chloramine T), dichlorodimethylhydantoin, quaternary ammonium compounds, and other organic chlorine carriers which liberate active chlorine in the presence of water. Disinfecting procedures of this kind are mainly common in Europe. Laundry sterilization in the USA is accomplished either thermally or by active chlorine bleaches.

Table 38. Formulation of a perborate-containing detergent for institutional use

Components	Examples	Composition, %
Surfactants	alkylbenzenesulfonates, soaps, alcohol ethoxylates	10–15
Sequestering agents	pentasodium triphosphate, zeolite, polymer cobuilder	20–30
Alkalies	sodium carbonate/silicates	10–20
Bleaching agents	sodium perborate, TAED	15–30
Fluorescent whitening agents	stilbene derivatives	0.1–0.3
Antiredeposition agents	carboxymethyl cellulose, polymers	0.5–2
Corrosion inhibitors	sodium silicate	3–5
Foam iregulators	silicone/paraffin	0.1–0.3
Stabilizers	phosphonates	0.2–2
Fragrance oils		0.1–0.2
Enzymes	protease, amylase	0.5–1.0

Specialty detergents are products formulated in such a way as to meet the demands of particular kinds of laundry or particular laundering processes. These include, for example, detergents without whitening agents for easy-care and colored fabrics, detergents with high contents of nonionics for heavily soiled working clothes, and enzyme-containing products for proteinaceous soils.

In many semiprofessional laundries like in small hotels and restaurants, detergents which are more or less comparable to those for the household sector are required. For example, perborate-containing detergents have been very popular especially in Europe. The formulation of a perborate containing detergent is given in Table 38.

6. Production of Powder Detergents

Until 1987 the term powder detergent worldwide was synonymous to what today is called a conventional detergent. Powder detergents were manufactured mainly by hot spray drying and had densities of some 300 to 550 g/L. Triggered by the rapid market growth of compact detergents in Japan in the late 1980s, a significant change of the technology to manufacture high-density (> 600 g/L) heavy-duty detergents also took place in Europe and the USA, following the fast growing demands of the respective markets. In major parts of Europe and the vast majority of the USA, compact detergents became state of the art during the 1990s. In addition, detergent tablets have made their way in Europe since 1998, whereas their market introduction in North America began in 2000.

For compact powder detergents, the traditional manufacturing process comprising slurry drying of most detergent components in a spray

tower and postaddition of temperature-sensitive components had to be extended by an additional densifying step.

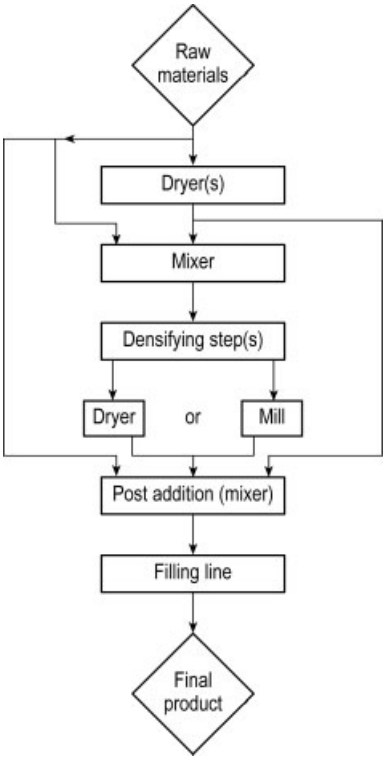

Figure 101. Manufacturing of compact powder detergents, schematic block flow diagram

In addition, quite a different development took place in the 1990s, namely, the use of nontower technology. This technology encompasses processes in which the drying step in the spray tower is replaced by a process which uses an increased number of dried raw materials and compounds (e.g., FAS on zeolite) than before. Initially, the bulk density of the tower powder was increased by dry mixing/milling the powder in fast-revolving, continuous, preferably high-speed mixers. In the 1990s, the European detergent industry developed new and more effective densifying processes which are based on tower powder, nontower granules, and dried raw materials or compounds. The achievable bulk densities depend on the detergent composition and the densifying process used (see Table 39).

Table 39. Densifying processes

Densifying process	Bulk density
Dry mixing/milling	650–720 g/L
Wet granulation/drying	750–900 g/L
Nonionic granulation	≈ 800 g/L
Roller compaction/milling	700–900 g/L
Bar extrusion/milling	700–900 g/L
"Spaghetti" extrusion	700–900 g/L
Tableting	> 1200 g/L

For each process many variants exist, e.g., wet granulation using

Water
Water-based binder, or
Pastes (surfactants)

in a

Dryer (fluidized-bed, linear, or round)
Mixer (Lödige, Fukae, Drais, Ballestra, Eirich, Schugi) [246], or
Combination of various mixers/dryers

A different technology has gained increasing importance as well, i.e., the process of mixing of compounds (see Section 6.2.4). In this process each raw material or part of it is separately agglomerated to dense particles. Therefore, anionic surfactant compounds, nonionic surfactants, builders, etc., can be easily varied in their content by dosing more or less of each compound. Compounds are manufactured by the detergent industry and — increasingly — by raw material suppliers, e.g., defoamer or tetraacetylethylenediamine granulates.

6.1. Technology Overview

Figure 101 depicts currently used detergent production technologies in a simplified way. Table 40 illustrates the main process alternatives, starting with drying/mixing, followed by a densification/shaping step. The processes are terminated with postaddition of thermally and chemically sensitive raw materials. Figure 102 additionally shows these alternatives as a chain.

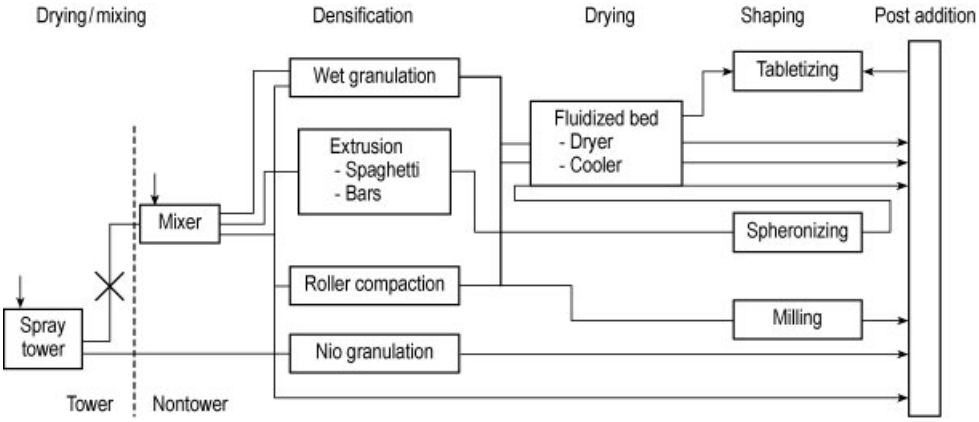

Figure 102. Chains of different processes

Table 40. Detergents manufacturing process alternatives

Process	Drying step	Densifying step
Tower process	Spray drying	Dry mixing/milling

Nontower (mixer(s))-process	– fluidized bed
	– chemical binding with e.g. zeolite, soda ash
	– neutralization with soda ash and chemical binding
		– integrated in the high granulation step or
		– mixture is
		  – extruded
		  – tableted
		  – roller compacted

Combination of tower/nontower "premix-process"	– one part is spray dried
	– another part is granulated with water and/or surfactants
	– third part is dry mixed
		– without densifying
		– integrated in the high shear granulation step, or
		– mixture is
		  – extruded
		  – tableted
		  – roller compacted

Compound process (mixer(s))	Compounds are – fluidized bed dried – roller compacted – granulated with surfactants	– integrated in the high shear granulation/extrusion/roller compaction step

6.2. Manufacturing Processes

6.2.1. Traditional Spray-Drying Process [247]

In the first step of the traditional spray-drying process (Fig. 103) a slurry of thermally stable and chemically compatible ingredients of the detergent is prepared. Solid and liquid raw materials are drawn off from silos or tanks and introduced batchwise into scales. Water is added as required to maintain a manageable viscosity. The liquid mixture and the solids are taken from the scales and mixed to form a slurry in a crutcher or by using some other type of forced mixer. The slurry is transferred to a stirred storage vessel from where the process runs continuously.

The slurry is transported by a low- or medium-pressure pump (up to 80 bar). Changes in pressure in the high-pressure portions of the system are compensated for by using an air vessel. The slurry is sprayed into the tower by two spraying levels connecting a series of nozzles. The number and type of nozzles must be designed such that overlapping of spray cones is avoided. Dried tower powder flows off from the tower at a temperature of 90–100 °C. To prevent lumping, an airlift is used for cooling.

The spray-drying process can be run cocurrently or, most common, countercurrently [248]. If particles and air move cocurrently and the slurry enters the hottest zone of the tower, a rapid evaporation of water occurs, and blown-up, relatively light particles (beads) result. In the countercurrent process, drying starts in an area of high humidity and lower temperature, and beads are obtained that are characterized by thicker walls. The lower free-fall speed due to the countercurrent airstream results in a higher number of particles in the tower at any time. Agglomeration products resulting from a countercurrent system are heavier and coarser than

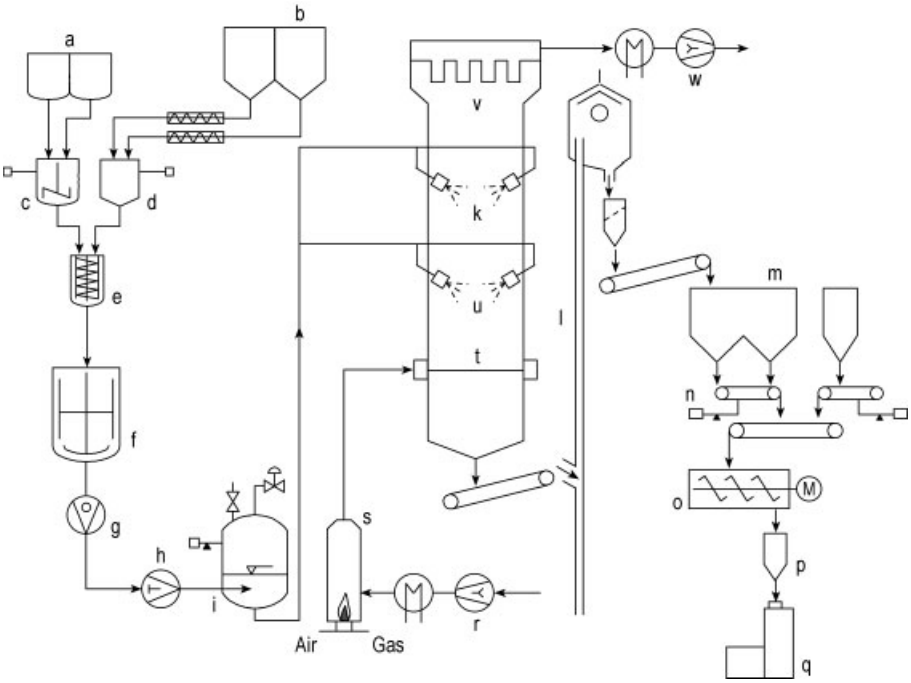

Figure 103. Traditional spray-drying process
a) Storage tanks for liquid raw materials; b) Storage silos for solid raw materials; c) Liquids weighing vessel; d) Solids weighing vessel; e) Mixing vessel; f) Intermediate tank; g) Booster pump; h) High-pressure pump; i) Air vessel; k) Nozzles; l) Airlift; m) Storage bunker; n) Belt conveyor scales; o) Powder mixer; p) Sieve; q) Packaging machine; r) Air inlet fan; s) Burner; t) Ring channel; u) Spray tower; v) Top filter; w) Exhaust

those resulting from cocurrent spray drying. Figure 103 shows a countercurrent operation. The air is directly heated by exhaust gases from an oil- or gas-fueled burner and led into the tower at about 300 °C through a ring channel. Appropriate design of both the air inlet zone above the cone and the air outlet at the tower dome is important. Air entering the tower is "swirled", i.e., it is accelerated tangentially and vertically. This causes homogeneous heat transfer between air and product. Water vapor, combustion gases, and drying air are withdrawn from the tower by means of a ventilating fan.

Fine particles are drawn off from the tower along with the exhaust gas and collected in a filter. In modern plants this filter is installed on top of the tower so that the separated fines and dust fall down into the agglomeration zone of the tower [249]. The resulting tower powder has a low content of fine particles. Heat exchangers are installed to use the exhaust gas energy for preheating the burner air. Another way is partially recycling the exhaust air (about 50 %) and mixing it with fresh air from the atmosphere [250]. In doing so, energy savings of about 10–20 % can be achieved. Plant control is possible by installation of an expert system [251].

6.2.2. Superheated Steam Drying

Environmental issues of spray drying using hot air are:

Emission of dust, normally < 10 mg/m^3
Emission of odor and volatile organic matter in very small amounts, normally < 50 mg/m^3
Potential dust-explosion and fire hazard
Loss of vaporization heat

These problems can be solved by switching over to drying with superheated steam instead of air [252]. The usage of superheated steam allows the heat of vaporization to be recovered [253]; the system is closed and no exhaust gas escapes from the tower [254]. Presently, this technology is being adapted in a pilot project [255].

The bulk density of steam-dried products is significantly higher and the water content is lower compared to hot-air spray drying.

6.2.3. Nontower Agglomeration Process

The nontower agglomeration process is carried out in a continuous mixer such as the Ballestra Kettemix Reactor or the Lödige CB-Mixer or in a combination of high- and low-shear mixers. Zeolite is applied in powdered form to bind surplus water, especially as zeolite P or X or as superdried zeolite A plus soda ash. In the first step the powder components are continuously added. Metering is carefully controlled at the entrance of the reactor. In this zone all liquids, such as nonionic surfactants, polymer solutions, fatty acids, sodium silicates, and linear alkylbenzenesulfonic acid, are separately dosed [256].

Due to the high energy input granules are formed and at the same time neutralization with sodium carbonate or sodium hydroxide takes place. The granules can take up limited amounts of nonionic surfactants. Optionally the granules are surface-treated with zeolite ("powdered") in an additional step to reduce product stickiness, which is mainly due to the nonionics. If the water binding capacity of zeolite and soda ash is not sufficient, e.g., in the case of neutralization with sodium hydroxide solution, added water and reaction water have to be removed from the product in a further step. This is done in a fluidized-bed dryer, where the granules are dried with hot air (140 °C) in the entrance zone. In the outlet zone the granules are cooled down, normally using cold air (5–15 °C). The drying process is controlled by measuring the product water content and temperature. Depending on the formula the product temperature has to be maintained between 50 and 110 °C (preferably between 60 and 80 °C).

If the water balance is fitted by using small amounts of water and/or by neutralization with carbonates the granules are only cooled in the fluidized bed and not dried. The fluidization air is transported in a cycle with a cooling/heating step to ensure a constant water content of the air.

After sieving and milling, the granules are transported to the postaddition step.

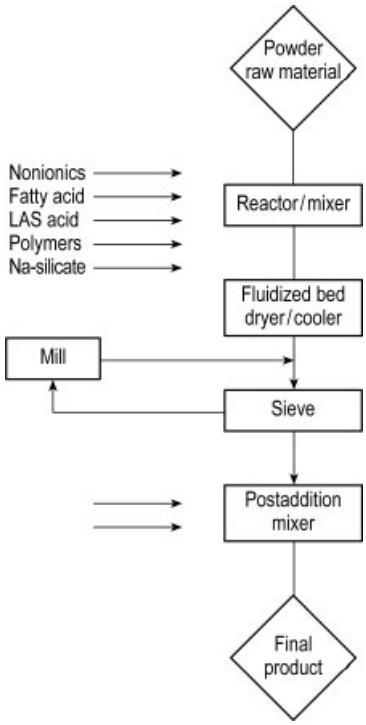

Figure 104. Nontower agglomeration process

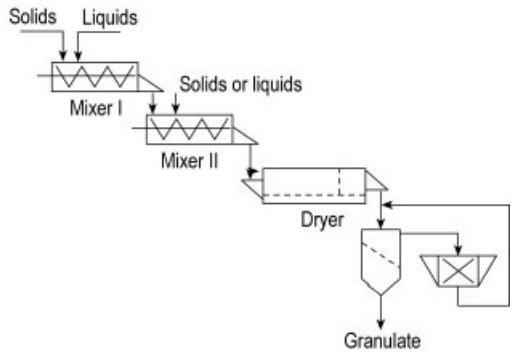

Figure 105. Flow sheet diagram of a typical two-mixer nontower process

Quite a few limitations with respect to the nontower agglomeration process exist. Firstly, the neutralization heat increases the granulation temperature. Therefore, the amount of acids in the feed material is strongly limited. Secondly, the water level must be lower than 14 %. The nonionic surfactant content depends on the detergent composition. At higher nonionics contents the granules become too soft and

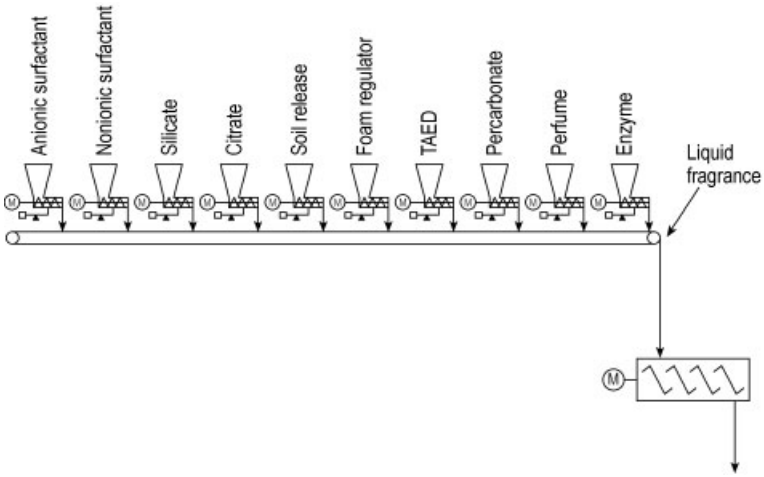

Figure 106. Dosing and mixing of different compounds

sticky. A typical nontower agglomeration process with one mixer and a fluidized-bed dryer is shown in Figure 104 as a block diagram. The more advanced variant with two mixers and a fluidized-bed cooler is illustrated in Figure 105 as flow sheet diagram. The process uses a high-shear/low-shear mixer combination. Neutralization is effected by using 50 % caustic soda.

6.2.4. Nontower Compound Technology

Another way of realizing a nontower process is application of compound technology. A finished laundry detergent is produced by blending the following components [257]:

Nonionic surfactant agglomerate	13.4 wt %
Anionic surfactant agglomerate	32.5 wt %
Layered silicate compacted granules	10.1 wt %
Granular percarbonate	22.7 wt %
Tetraacetylethylenediamine agglomerate	7.8 wt %
Suds suppressor agglomerate	6.5 wt %
Perfume encapsulate	0.1 wt %
Granular soil release polymer	0.4 wt %
Granular sodium citrate dihydrate	3.5 wt %
Enzymes	2.0 wt %

Each component is produced separately, in most cases by using the granulation technology. By applying different densification technologies an inhomogeneous product aspect results after blending of the compounds. Compound blending is presented in Figure 106.

6.3. Densification Processes

To increase the bulk density of the final product to values higher than about 650 g/L, one or more densifying steps must be used in detergent production and/or denser raw materials/compounds have to be applied. Usually, the particle size distribution of the densified product is adjusted by sieving/milling. In most cases the size distribution is similar to that of tower powder or standard products (0.2–0.8 mm). Larger and denser particles may retard dissolution of the finished product. Therefore, detergent manufacturers try to minimize the quantity of oversize particles.

Many process alternatives for densification exist. Typical process technologies are described in the following.

6.3.1. Dry Densification in a Mixer

A very simple and cheap method for achieving a limited increase in detergent density is mixing/milling/spheronizing of tower powder in a continuous mixer such as the Lödige CB-mixer (Fig. 107). The densification yields an increase in bulk density of 50–100 g/L [258, 259]. Usually, a liquid (nonionic surfactant) is added to the CB-mixer to reduce stickiness, especially at the mixer walls and the tools.

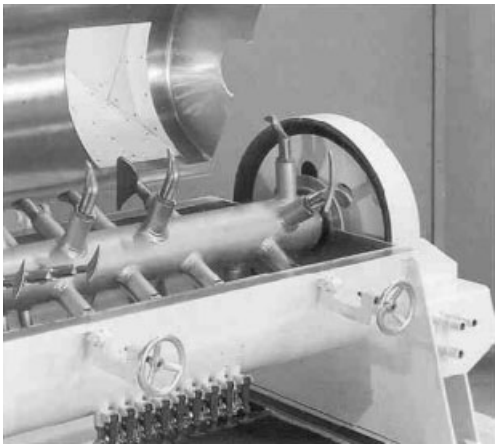

Figure 107. Lödige CB-mixer for dry densification of detergent tower powder (view of the internals)

6.3.2. Dry Densification in a Spheronizer

The bulk density of some products can be increased slightly by rounding off the particles in a spheronizer [247]. After entering the spheronizer the particles are accelerated by a high-speed rotating disk. They are driven against the vessel wall by centrifugal forces. There they roll against each other in a particle cloud. As a consequence, edges and surfaces of the particles are polished. In this way, the form of the particles is improved.

6.3.3. Dry Densification in a Roller Press

Compacting of raw materials or compounds in a roller press is a very simple process, that is followed by a relatively complex milling/sieving step (Fig. 108). Compacting in a roller press is accompanied by a dust problem, especially when using products with a high content of organic matter. The resulting product is very dense, but the particles are irregular in shape (splinters). The latest development in this field is a closed device in which rolls, mills, sieves, and recycling tubes are installed in one unit.

For "dry" densification an extruder can be applied as well. In this case, a binder/lubricant must be added to the powder mixture. The product is extruded to form bars. The final form can be achieved by milling/sieving, as shown in Figure 108.

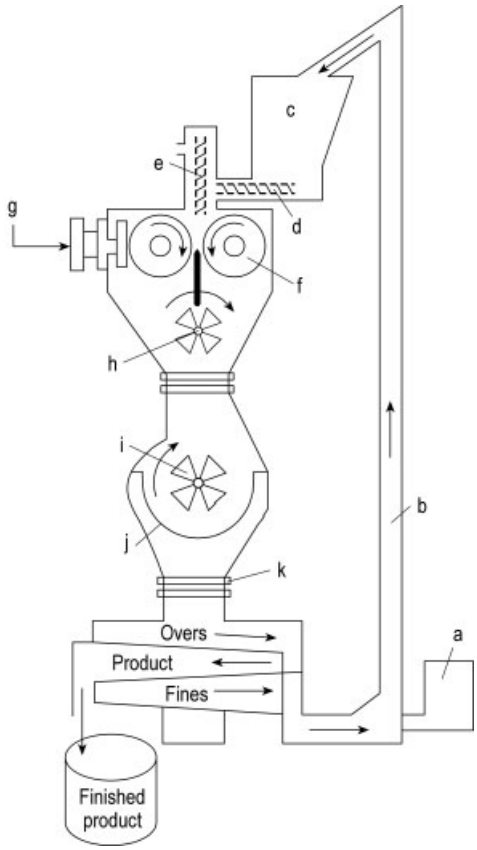

Figure 108. Typical compaction/granulation systems with screening and recycling of overs and fines
a) Original powder feeder (lower hopper), b) Recycle system, c) Upper feed hopper, d) Horizontal feed screw, e) Vertical screw, f) Compaction rolls, g) Pressure applied, h) Prebreak, i) Granulator, j) Screen, k) Screener

6.3.4. Wet Granulation

Wet granulation is the most common process for densifying tower powders, raw materials, or compounds. Wet granulation is performed by adding water, polymer solutions, anionic surfactant pastes, or surfactant gels at high shear rates [260 – 263]. The wet granulation process consists of the following steps:

Step 1: grinding in a high-speed mixer or in a mill; addition of polymer solution (7–12 %) via spray nozzles to the powder in the mixer (pregranulation)

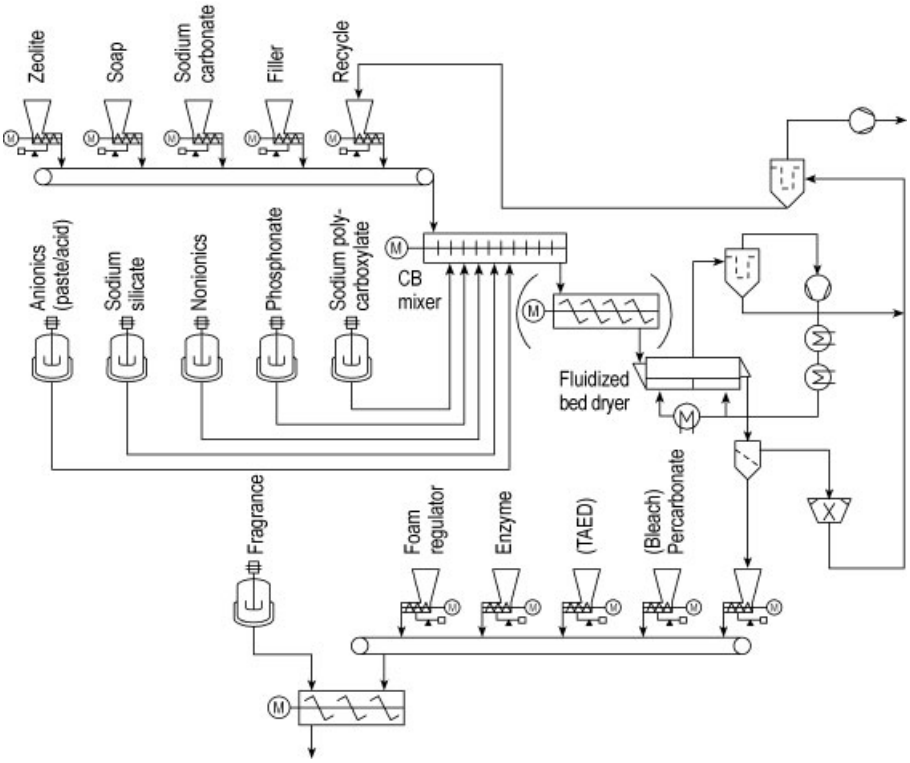

Figure 109. Flow sheet of a typical nontower granulation process

Step 2: granulation and conditioning in a high- or medium-shear-rate mixer
Step 3: evaporation of added water in a fluidized bed and/or cooling down
Step 4: sieving of coarse particles and milling; usually, fine particles remain in the product

Figure 109 illustrates one of several possible designs of a typical wet granulation process [264]. The use of two or more mixers of different types in a cascade increases process flexibility. Energy input, granulation temperature and time, dosage of liquids, powdering of the particles, and types of mixers represent the know-how of the detergent manufacturers.

A very promising new granulation technology is foam granulation [265, 330]. Usually, the granulation liquid is sprayed on the solids or dosed in a jet. The addition of a foam enables a more flexible granulation with less recycled material. A comparison of different granulation techniques is shown in Figure 110.

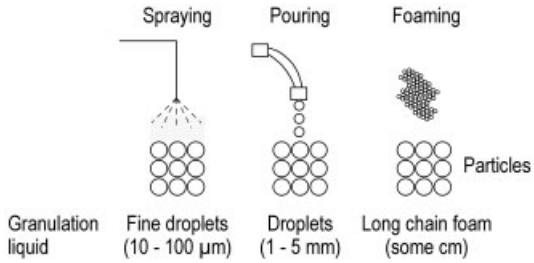

Figure 110. Wet granulation

6.3.5. Spaghetti Extrusion

The spaghetti extrusion process represent a new approach to densifying detergents. Powders, raw materials, and compounds are mixed while a lubricant is added. The mixture is plastified in an extruder and pressed through many small holes. The resulting spaghetti-shaped detergent strands are cut in such a way that the length-to-diameter ratio of the resulting tiny cylinders is unity. The cylinders are rounded off in a spheronizer [227,

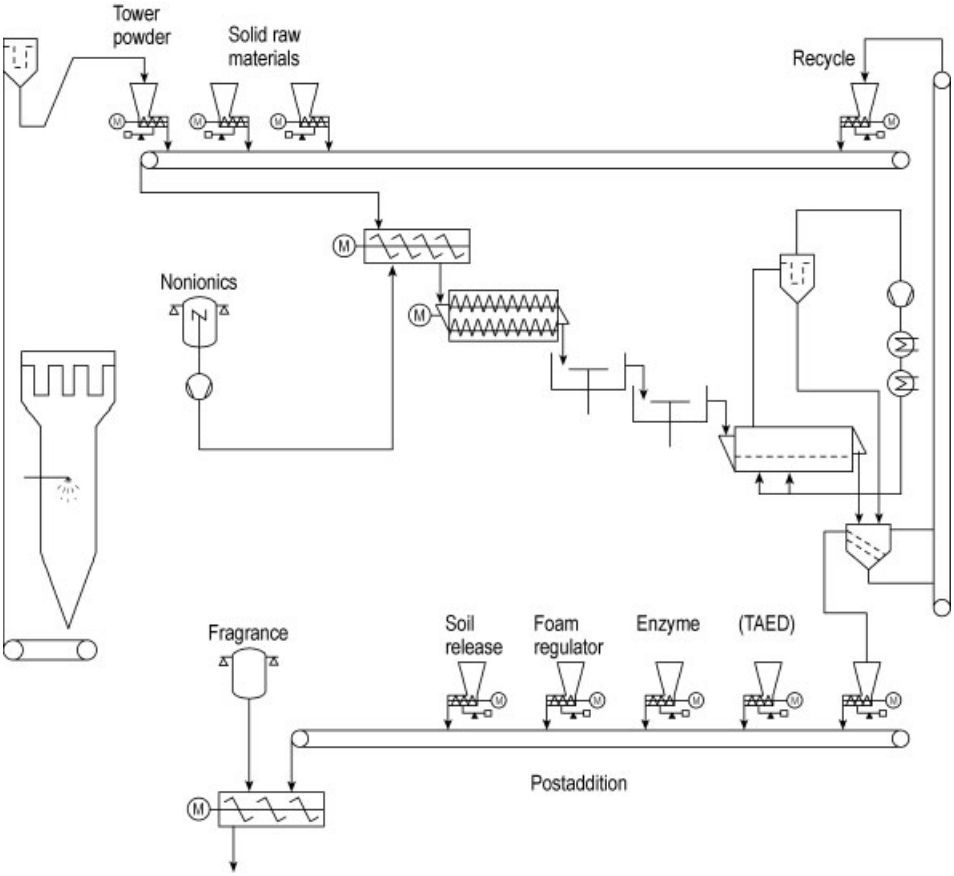

Figure 111. Flow sheet diagram of the extrusion ("Megaperls") process

266]. To harden the product an additional cooling step can be necessary. This densification process has the advantage that a dust-free product is obtained with spherically shaped particles of uniform size, whereas all other processes deliver powders, granulates, splinters, or mixtures thereof, that have a more or less broad spectrum of particle sizes.

Densification by extrusion occurs before the postaddition process (Fig. 111). If the raw materials are used as they are, no further granulation step (e.g., tower drying or wet granulation) is necessary. Both densification and particle shaping occurs inside the extruder by means of a specially constructed extruder head and knife.

The detergent mass is plastified in a cooled twin-screw extruder. In the extruder head a mass temperature of 60 to 80 °C prevails under a pressure of 70 to 110 bar. The mass is pressed through bore holes of 1.4 mm in diameter. There are about one hundred bore holes per plate and more than hundred plates per extrusion die.

The extruded particles differ significantly from other detergent particles, they are much larger, spherical, and nearly uniform (see Fig. 112).

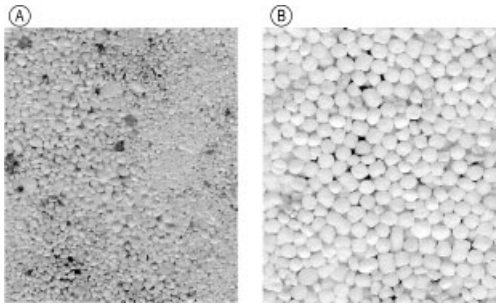

Figure 112. Appearance of market detergent products
A) Particles from granulation/compounding; B) Particles from extrusion

6.3.6. Postaddition Process

The finishing process of granulated or extruded detergents includes postaddition of thermally sensitive raw materials and/or chemically incompatible ingredients such as bleaching components (sodium perborate, sodium percarbonate), bleach activators, enzymes, soil repellents, foam inhibitor concentrate, fragrances, and, occasionally, colored granules. After the postaddition step the finished product is ready for sieving and packaging.

Postadded granules or powders generally have to be adjusted in their particle size distribution to match the particles of the base powder. Smaller postadded particles, particularly when they have higher densities, tend to segregate in the package or in the process before.

6.3.7. Dry Densification in a Tablet Press

Tablets are a convenient application form for laundry products. Tablets may consist of one, two, or more layers. In Europe, usually two tablets are recommended to clean a 4.5 kg laundry load in the washing machine. To achieve maximum performance all ingredients need to be in a dissolved state from the very beginning of the wash cycle [267 – 271]. Two main problems must be solved:

1) The low surface-to-volume ratio of detergent tablets adversely affects the dissolution rate.
2) In the presence of surfactants, especially nonionics, gel phases may form upon first contact of the tablets with water. Gel formation impedes fast dissolution.

The different concepts that have been developed to formulate and to produce tablets that quickly disintegrate are presented in Section 4.1.4.

Independent of the product concept many physical parameters are identical for the finished tablet and its manufacturing (Table 41). All manufacturers use rotary die presses for production.

The critical points of tableting detergents are the strong reciprocal dependency of tablet hardness and dissolution or disintegration time. The tableting pressure depends on the mechanical stability of the tablet and on its disintegration time. Tablets produced under high pressure are very stable but have too long disintegration times. On the other hand, tablets produced under low pressure are friable. Only a narrow range of tableting pressures can be applied for producing stable, nonfriable tablets with a short disintegration time. The so-called operation area for the tableting press is not fixed, the values depend on formula, size and distribution of the particles as well as on the bulk density of the premixture.

Table 41. Parameters affecting tablet quality and process efficiency

Product Assumptions	Process Conditions	Tableting Machine Arrangement
• Formula (formula splitting into two phases) • Content of liquids • Tableting aids • Particle size and distribution • Bulk density • Stickiness • Particle shape • Moisture content • Disintegrant in the tablet and/or coating	• Speed of tableting • Climatic conditions • Hardness of particles and tablets • Homogeneity of mixtures • Compressibility • Pressure of tableting • Temperature of the tablets	• Surface of the punches • Dosing systems • Tablet shape, weight, height and type • Control systems • One, two or three different phases • Tablets with an inscription

As described in Section 6.3.6, the particle size distribution of all formula components has to be adjusted.

The ingredients of the tablet can be separated into two layers. In this way, for example, enzymes and bleach can be compacted together in one tablet but in different layers without interaction during storage. The manufacturing process for a single-layer tablet is given below. For the second and third layer, the same procedure is applied in separate equipments to components with different composition and color:

Step 1: Granulation of the main components and drying
Step 2: Sieving of the granules in a range of 0.2 to 2 mm or a smaller range
Step 3: Postaddition/mixing/sieving (bleach, enzymes, salts, optical brighteners, polymer compounds, fragrances, suds and pH regulators, recycled materials, dyes)
Step 4: Tableting
Step 5: Flow wrapping of the tablets
Step 6: Packaging

The whole process must be temperature- and moisture-controlled. To prevent picking and sticking of granules the water content is limited to 6–9 %, depending on the formula. To ensure maximum action of the disintegrant, wrapping of the tablet in a so-called flow packer is necessary. Each flow pack usually contains two tablets for one wash cycle.

Figure 113 shows the dependency of the tableting pressure and dissolution rate on the particle size. Very fine particles need low pressures to form stable tablets. Since a disintegration agent with a small particle size develops only low forces, a minimum size of the disintegrant is needed, optimally adjusted to the particle size distribution of the granules. An interesting range of particle size may be, for example, 0.4 to 1.2 mm.

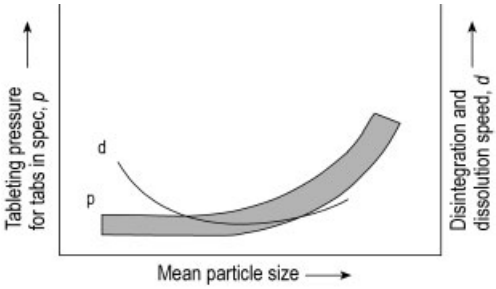

Figure 113. Dependence of the tableting pressure and dissolution rate on the particle size

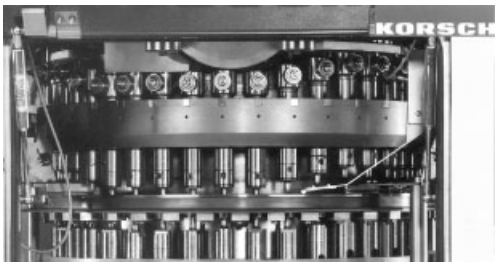

Figure 114. Punches of a double-layer tableting machine [272]

A contrary approach is the manufacturing of a friable, unstable detergent tablet with low tableting pressures. This kind of tablets already break by simply touching them. To stabilize the product it has to be coated with materials that form a very hard, easily melting layer or shell. A preferred technology may be using an enrober with single or double curtains. Suitable raw materials for enrobers are C_{12-14} dicarboxylic acids combined with a disintegrant.

Rotary die presses are used for tableting detergents [246]. Two-layer-tablet presses are equipped with an upper and lower punch and two filling stations. The residence time under maximum pressure and the press forces determine tablet stability and disintegration time. Figure 114 shows the punches.

6.4. Raw Materials

Each of the key detergent ingredients (see Chap. 3) consists of chemically and physically very different raw materials, also called systems, e.g., bleaching systems with sodium perborate monohydrate or percarbonate and TAED. The ingredients influence the choice of production technology.

6.4.1. Anionic Surfactants

For laundry detergents usually linear alkylbenzenesulfonates and alcohol sulfates, linear or branched, are used as anionic surfactants. As LAS do not crystallize, they are only available on the market as an aqueous paste or a dusty powder or powder compound. Alcohol sulfates crystallize more or less well, depending on their carbon chain distribution. Quite a few manufacturing processes exist for the production of AS compounds with up to 93 % active matter. Anionic surfactants and their manufacturing processes are amply described in the literature [273].

The starting materials for alcohol sulfates, alcohol sulfuric acid semiesters, are derived from sulfation of long-chain alcohols (C_{12}–C_{18}) with SO_3. The semiesters are not stable and have to be instantaneously neutralized with sodium hydroxide. One process is "low-water" neutralization with highly concentrated sodium hydroxide at elevated temperatures in a melt in a continuous neutralization loop [274]. Another technology applies a spray neutralization process, in which the alcohol sulfuric acid semiester and sodium hydroxide solution are mixed and sprayed into the tower using a special nozzle. The heat of neutralization evaporates the water, so that no further energy input is required for drying [275].

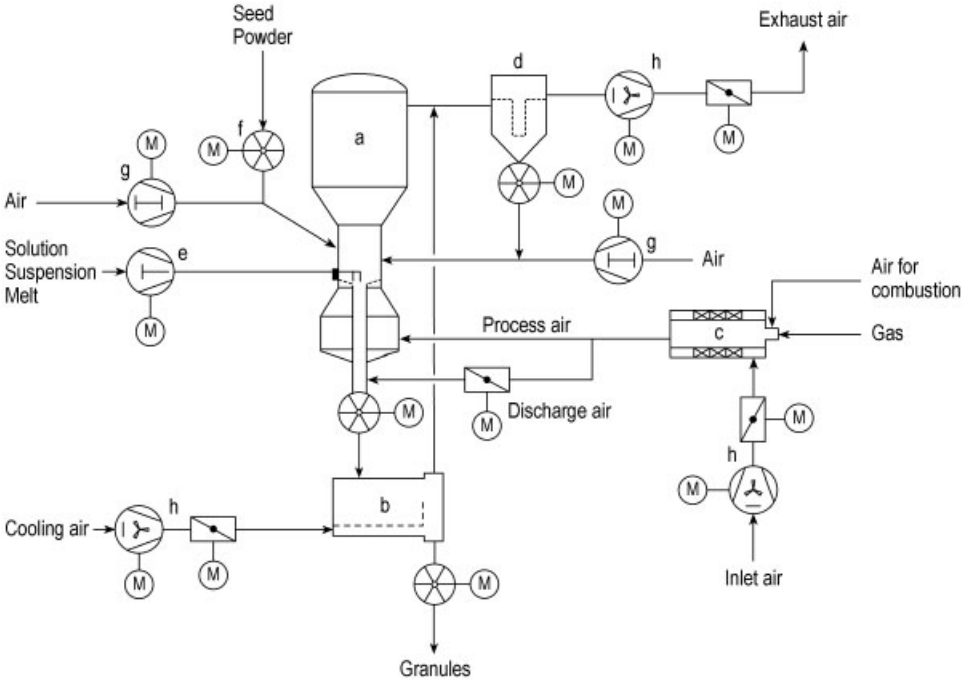

Figure 115. Flow sheet of the one-step agglomeration/drying process[279]
a) Fluidized-bed dryer; b) Fluidized-bed cooler; c) Heater for process air; d) Deduster; e) Dosing pump; f) Metering rotary valve; g) Conveying air fan; h) Fan

The resulting powder must be granulated under densifying conditions [276]. Another method embodies drying in a high-shear-rate dryer, in which the paste is dried in a thin film along the wall of the dryer [277].

A further process recommends applying vacuum neutralization in a special film reactor with a high-shear mixer on the top [256]. The reaction rate is controlled by continuously measuring the pH value of the product. Alcohol sulfates are most frequently used in a paste delivery form. The paste can be dried in a cocurrent spray-drying tower under inert atmosphere. The resulting powder is very light and fine and must be densified, e.g., by extrusion as noodles. The produced shape is shown in Figure 116.

The Ballestra company offers the Dryex process for the production of dry anionic surfactants, especially for AS. The neutralized and preheated (80–85 °C) pastelike surfactant is fed into the top of a wiped film evaporator. Rotating blades form a thin film over the evaporator surface and remove the dried product from the wall. Drying is supported by using a slight underpressure. The product can be flaked and milled in-line. Similar processes are available from various manufacturers of thin-film evaporators.

In another more elegant and economic method the paste is dried in a fluidized-bed dryer/agglomerator. This process, shown in Figure 115, performs drying, densifying, and sieving in one apparatus [278]. The paste is sprayed onto smaller particles, for example, product dust which is fluidized in the bed. The product is sprayed and dried shell by shell. Rotation and the method of drying provide products of high bulk density. The stickiness can be reduced by compounding with carrier materials such as zeolite, starch, and inorganic or organic salts.

The drying of AS paste is accompanied by emissions resulting from unavoidable small amounts of free fatty alcohols in the exhaust air. The problem may be partially resolved by reducing the inlet temperature.

Shape and particle size of the product depend strongly on the process type, as shown in Figure 116.

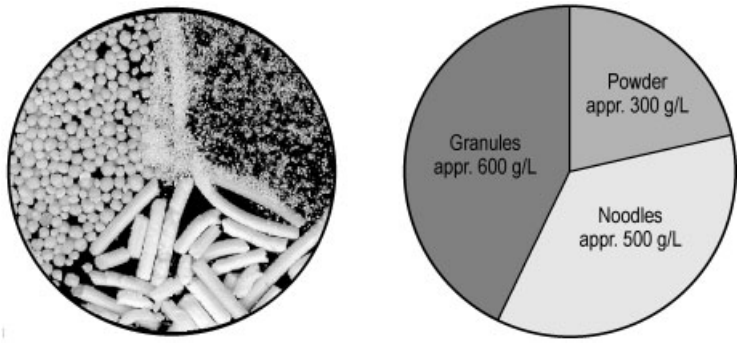

Figure 116. Particle shapes of dried anionic surfactant salts

6.4.2. Nonionic Surfactants

Nonionic surfactants are derived either from natural renewable feedstocks or from petrochemicals. As *renewable raw materials* for surfactants, particularly tallow, coconut and palm kernel oil (laurics), and soybean oil are of paramount importance. The oils are purified in a first step with sulfuric acid in separators or decanters [43, 280]. The triglycerides are either split with steam at elevated temperatures and pressures to form fatty acids and glycerol, or transesterified with methanol to produce fatty acid methyl esters in a one-step reaction. The methyl ester are subsequently hydrogenated to form fatty alcohols (→ Fatty Alcohols).

For the production of fatty alcohols from *petrochemicals* (paraffins and ethylene) a range of processes exist [281]. The most well-known processes start with the catalyzed oligomerization of ethylene either according to the Shell Higher Olefins Process (SHOP; → Hydrocarbons, Section 2.1.2.2) or the Ziegler trialkylaluminum process (→ Alcohols, Aliphatic, Section 2.3.5). The SHOP olefins are hydroformylated with carbon monoxide and hydrogen to form aldehydes (→ Oxo Synthesis) and in a further step hydrogenated to low-branched alcohols (→ Alcohols, Aliphatic, Section 2.3.2). In the Ziegler process trialkylaluminum is oxygenated and in a further step hydrolyzed with sulfuric acid, which converts it to long-chain linear alcohols.

The manufacture of nonionic surfactants starts in most cases with chemical modifcation of fatty alcohols, fatty acids, or fatty acid methyl esters. Fatty alcohols are converted to alcohol ethoxylates and alkylpolyglycosides; fatty acids to fatty acid ethoxylates and alkanolamides; and the methyl esters to the corresponding ethoxylates and glucamides. Important nonionic surfactants are shown below:

Fatty alcohol ethoxylates $R–(OCH_2–CH_2)_nOH$

Fatty acid methylester ethoxylates

Fatty acid ethoxylates

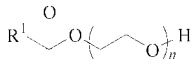

N-Methylglucamide

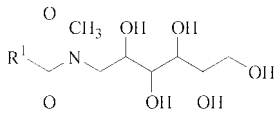

Alkylpolyglycosides (APG)

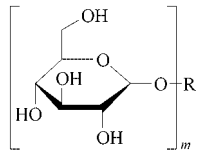

$R = C_8–C_{18}$
$R_1 = C_7–C_{17}$
$n = 2–10$
$m = 1–3$

Ethoxylated fatty alcohols are the most important nonionic surfactants. Also, ethoxylated fatty acids and fatty acid methyl esters [266] are valuable products for the detergent industry. Therefore, ethoxylation is a key technology for manufacturing nonionic surfactants. The ethoxylation process is conducted in batch reactors. A description of the ethoxylation process in given in → Surfactants, Section 7.2.

The ethoxylated substrates are chemically inhomogeneous because, firstly, the raw material has a chain length distribution (e.g., $C_{12}–C_{14}$ alcohols) and, secondly, the addition of ethylene oxide is carried out in a broad concentration range. The designation "fatty alcohol + 7 EO"

indicates a mean value of seven added molecules of ethylene oxide per molecule of alcohol. In reality a mixture of alcohol ethoxylates with zero to eighteen EO moieties (broad-range ethoxylates with KOH catalyst) is obtained. When special catalysts are used, narrow-range ethoxylates are formed. These process parameters affect the product quality significantly and are part of the manufacturer's know-how.

6.4.3. Builders

Quite a few builders are applied in the detergent industry today. Worldwide, sodium triphosphate is a key builder used on a large scale. Since the early 1980s sodium aluminum silicates (zeolites) have become more and more important, particularly in Europe, the USA, and Japan. Today, the field of builders/cobuilders is more complex than ever. Apart from sodium triphosphate and zeolites, special layered silicates, amorphous silicates, and compounds made from sodium carbonate and amorphous silicate are applied [169, 170, 282]. Besides zeolite A, other zeolite types have gained importance, e.g., zeolite P and zeolite X. Production processes for zeolite A have been described [247].

Layered Silicates. Several methods have been established for manufacturing layered silicates. One method uses the crystallization of amorphous sodium silicates (SiO_2:Na_2O ratios from 1.9 to 3.5), evaporation of the water, and tempering at above 450 °C but below the melting point, to achieve complete crystallization. From this process α-, β-, γ-, and δ-silicate ($Na_2Si_2O_5$) modifications are available [169, 283, 284]. Another process comprises spray drying of a water glass solution, milling of the product, and tempering in a rotary kiln at 400 to 800 °C for 1 h to crystallize the sodium disilicate [285]. Subsequent densification can be carried out by roller compaction, if necessary.

Amorphous Silicates. A simpler process is the manufacture of amorphous silicates. A spray-dried silicate solution having an SiO_2 : Na_2O ratio of between 2 and 3.5, which is derived from the hydrothermal or water glass process, is compacted and partially loaded with a selected liquid, e.g., nonionics [286]. A variation of the densifying process is the granulation of milled silicate particles [274, 287].

In the 1990s, some raw material suppliers developed manufacturing processes for detergent-grade zeolite types, such as zeolite P and zeolite X, which feature increased loading capacity for liquids [256, 288 – 290]. The manufacturing processes differ from the zeolite A manufacturing technologies in the crystallization conditions.

6.4.4. Peroxygen Bleaches

The detergent industry has used sodium perborate tetrahydrate or monohydrate as a source of active oxygen since 1907. In addition, sodium percarbonate has gained increasing importance in the last decades.

Sodium Perborate. As raw materials for the sodium perborate tetrahydrate process boron minerals, e.g., kernite or tincal with B_2O_3 contents $> 47\,\%$, are utilized. In the first step the boron compounds are dissolved in an alkaline solution at about 90 °C to form sodium metaborate:

$$Na_2B_4O_7 + 2\,NaOH \rightarrow 4\,NaBO_2 + H_2O$$

After filtering and diluting the solution to 75 g/L, hydrogen peroxide is added with cooling (< 30 °C) in a crystallizer:

$$NaBO_2 + H_2O_2 + 3\,H_2O \rightarrow NaBO_3 \cdot 4H_2O$$

The sodium perborate tetrahydrate crystals are separated from the liquid phase in a centrifuge. They contain 3 to 10 % of surface-adsorbed water which is removed in a fluidized-bed dryer with product temperatures below 60 °C. Above this temperature (65.5 °C) the product melts in its water of crystallization. Sodium perborate monohydrate is produced from the tetrahydrate by dehydration in a fluidized bed, conditioning, and drying with a steam–air mixture at temperatures not exceeding 100 °C. In its crystalline state sodium perborate forms a stabilizing anion (see Section 3.3.1) [174, 291]:

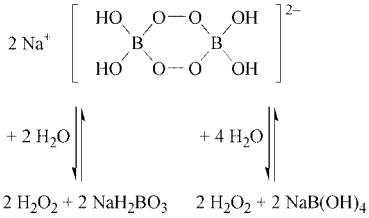

Sodium percarbonate ($Na_2CO_3 \cdot 1.5\ H_2O_2$) is easily manufactured starting from a crystallization process or from spraying a 35 % aqueous hydrogen peroxide solution onto sodium carbonate in a mixer or, preferably, in a fluidized bed. Percarbonate stability depends on the degree of purity of the sodium carbonate used. Especially heavy metal ions adversely affect the stability by catalyzing the decomposition of hydrogen peroxide.

To enable storage of sodium percarbonate in silos it has to be protected by a coating layer. It is common to use inorganic materials as a coating layer, e.g., sodium carbonate, sodium perborate monohydrate, or sodium sulfate. Coating is performed in batchwise or continuously operating fluidized beds.

Stability. All bleaching agents mentioned above are storage stable up to 60 °C. The perborates are stable at this temperature even when they are loaded with nonionic surfactants. Dust and particles of percarbonate in direct contact with even small amounts of organic material may result in spontaneous ignition of the organic material. Caution measures are to be taken in case of a damaged coating layer or with broken particles.

6.4.5. Enzymes

Detergents may contain up to four different types of enzymes: proteases, amylases, cellulases, and lipases. Enzymes are effective at low temperatures. Even small amounts of these biocatalysts guarantee major effects.

Enzymes are produced in fermenters by microorganisms such as bacteria, yeasts, and fungi. For production of proteases and amylases *Bacillus* species are used, which today are usually genetically engineered. A flow sheet of production of enzymes for detergents is show in Figure 117. Production of enzymes is described in more detail in → Enzymes, Section 3.1.

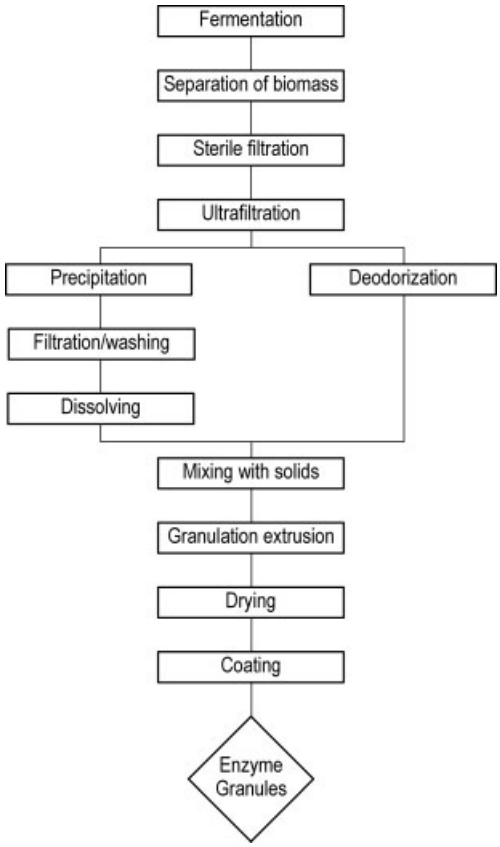

Figure 117. Downstream processing in enzyme production

In the fermenter, the extracellularly produced enzymes are dissolved and highly diluted in the fermentation broth. The first step of downstream processing is the separation of the biomass and other solids originating from the nutrient medium (Fig. 117).

Three alternatives are in use:

Centrifuges with extreme centrifugal forces (separators)
Precipitation with salts
Microfiltration

This separation step has proved difficult in practice because the bacterial cells that produce the enzymes are very small (0.2 to 0.5 µm) and their density is close to that of the fermenter broth. Moreover, the fermenter broth has a relatively

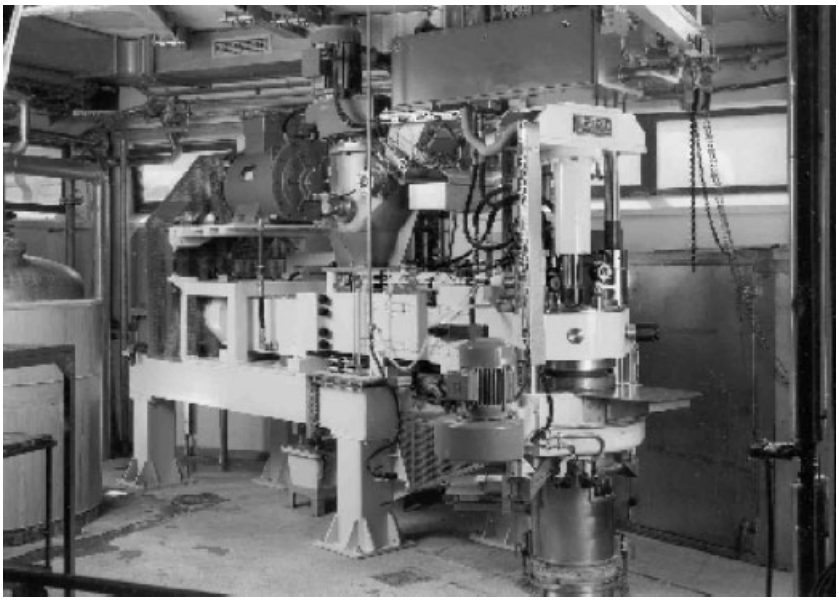

Figure 118. Extrusion process [293]

high viscosity, frequently combined with a non-Newtonian flow behavior. A further limitation is imposed by the poor temperature stability of the enzymes, so that, in many cases, the working temperature may not exceed 30 °C [292]. In addition, slimes can impede work-up.

To guarantee a microorganism-free product, a sterile filtration in a microfiltration plant follows. In the next step the highly diluted solution is ultrafiltered to achieve 15- to 25-fold concentration of the large protein molecules. For the next step two alternatives exist:

1) Precipitation of the enzymes, filtering the suspension, washing and dissolution of the cake.
2) Deodorization of the concentrated enzyme solution to remove unpleasant odors, generated during fermentation and sterilization. This is conducted in vacuum with superheated steam.

The concentrated enzyme solution is finally sprayed on solid carriers, granulated or extruded (Fig. 118), dried, and coated. Coating is performed in a fluidized bed by spraying a polymer solution on the particles under drying conditions. The coating layer must be perfect to prevent human contact with the particles and to make them dust-free.

7. Analysis of the Composition

The fact that detergents can consist of a large number of individual components whose structures vary greatly is apparent from Chapters 3 and 4. The dramatic industrial advances of the past 50 years together with ecological and economic issues have resulted in steady modifications of detergent formulas. These modifications have required adjustments of analytical methods, with routine procedures frequently becoming outdated.

The multitude of products and a constantly growing variety of raw materials, as well as the complexity of their constitution, makes a summary of the methods used for detergent analysis difficult, especially because the situation is complicated even further by the wealth of analytical techniques available.

The most important factor in choosing an analytical method is the nature of the question posed. In addition, the value of an analysis is highly dependent on the use of proper sampling techniques. Often several methods of analysis are employed in the search for a single answer.

The following procedure can be regarded as a typical example of how a general analysis of a powder detergent might be conducted:

First, the product is separated by extraction into ethanol-soluble and ethanol-insoluble fractions. The ethanol extract can usually be regarded as containing all of the surfactants. The various surfactants are then determined: anionic surfactants by two-phase titration, nonionic surfactants by passing the ethanol extract through both anionic and cationic ion-exchange resins, and the soaps either by titration of the alkalinity of the ethanol extract or by a second two-phase titration in alkaline medium. After suitable sample preparation, such major inorganic components as carbonates, silicates, phosphates, borates, and sulfates are determined either instrumentally or by classical wet methods. All of the more specialized auxiliary ingredients (e.g., soil antiredeposition agents, fluorescent whitening agents, enzymes, chelating agents, bleach activators, etc.) are present in relatively small quantities, and their qualitative and quantitative analysis requires considerable knowledge and experience. Thus, they would fall outside the bounds of a general investigation.

Following is a summary of the literature on detergent analysis. This is intended only as a useful overview. The information has been divided into five categories:

1) Detergent ingredients
2) Purposes of detergent analysis
3) Importance of sample preparation
4) Analytical methods
5) Sources of information

7.1. Detergent Ingredients: [238 – 243, 294 – 298]

Inorganic components: [299 – 306]
Organic components: [307 – 312]

7.2. Purposes of Detergent Analysis: [313]

Raw material analysis (triphosphate and alkylbenzenesulfonate): [314 – 316]
Production control (EO products): [317]
Product quality control (detergent analysis): [318, 319]
Market analysis (detergents): [310], [311, pp. 991–1045], [320 – 325], [326, pp. 25–43 (1977)]
Trace analysis (vinyl chloride in PVC): [327 – 329]
Analysis of formulated products: [330]
Analysis of detergent formulation: [331]
Detergent analysis: [332]

7.3. Sample Preparation: [310, pp. 21–26]

Liquids, solutions, suspensions, emulsions, and pastes: [315, p. 6], [333, H–I 4 (65)]
Detergent bars: [333, G–I 1–3 (50)], [334, Da 1–45]
Powders: [333, H–I 5], [334, Dc 1–59, Dd 1–59], [335, ISO 607 (1980)], [336], [337, DIN 53 911]

7.4. Analytical Methods

7.4.1. Qualitative Analysis: [310, pp. 27–49], [322, pp. 166–178]

General Analysis:
Solubility, pH: [333, H–III 1 (65)]
Elemental analysis: [333, H–III 5 (65)]
Analysis for Inorganic Ingredients: [338, 339]
Active oxygen: [338, p. 171], [339, pp. 492–493]
Peroxide accumulation in detergents: [340]
Active chlorine: [338, p. 159], [341, Part 2, vol. VII, p. 35]
Ammonia: [338, p. 241], [342, E 5]
Boron: [338, p. 230], [339, p. 776],
Water-insoluble inorganic builders: [243, 343 – 347]
Analysis for Organic Ingredients: [348 – 350]
Alkanolamines after hydrolysis: [310, p. 32],[312, 351]
Sequestrants (NTA, EDTA): [352, 353]
Urea: [354], [355, p. 235], [356]
Fluorescent whitening agents: [310, p. 414], [357 – 359]
Anionic surfactants: [360 – 363]
Alkylene oxide adducts (EO, PO, aromatics): [310, p. 46], [348, p. 138], [362, 364, 365]
Hydrotropic substances: [354]
Enzymes: [366]
Carboxymethyl cellulose: [367 – 371]

Amphoteric surfactants: [372]
Alkyl polyglycosides: [373, 374]
Polycarboxylates: [375]

7.4.2. Sample Preparation

Extraction, Perforation, and Sublation
Ethanol solubles: [310, p. 148], [333, H–III 4 (65)]
Soap fatty acids: [333, G–III 6 b (57) and G–III 8 (61)], [376, p. 1352]
Monosulfonate in alkanesulfonate: [315, p. 39],[333,H–IV 2 b (76)], [335, ISO 3206 (1975)]
Acetone solubles: [310, pp. 253 and 362]
Nonionic surfactants in wastewater: [377]
Nonionic surfactants in sludge: [378]
Polyglycol in ethylene oxide adducts: [379]
Distillation: [380]
Ion Exchange: [310, pp. 50–58], [381]
Nonionic surfactants, anionic surfactants, cationic surfactants, soap, fatty acids, and urea: [382 – 389]
Antimicrobicides: [390]
Phosphate types: [310, p. 467]
Carbon Methods
Sulfate: [333,H–III 8 b (76)]
Phosphate: [391]
Decomposition Methods: [392]
Residue on ignition, with hydrofluoric acid digestion: [393, p. 8]
Sulfate ash: [310, p. 335], [333, H–III 11], [394]
Borate: [310, pp. 474–476]
Borate decomposition for X-ray fluorescence analysis: [316, 395, 396]
Potassium disulfate melt for Al and Fe: [393, p. 12]
Soda–potash decomposition for water glass: [393, p. 9]
Wet Decomposition
Sulfuric acid–nitric acid, perchloric acid, nitric acid–perchloric acid: [397, pp. 188–209]
Measurement of zeolite, silicate, and phosphate: [347]
Chemical Transformations: [398]
Sodium hydrogen sulfite–acid hydrolysis: [221, p. 152], [333, H–IV 7 a], [399], [376, pp. 1474 and 1353], [400, 401]
Cleavage of ether sulfates: [310, pp. 279 and 323], [402 – 404] [405 – 407]
Esterification of fatty acids: [310, p. 572], [408 – 411]
Desulfonation (alkylbenzenesulfonates, alkanesulfonates): [310, p. 197], [403, 404, 412 – 417]

Acetylation (fatty alcohols): [418, p. 543]
Determination of industrial grade carboxymethyl celluloses: [369]
Alkaline Hydrolysis: [310, p. 150]
UV photolysis: [419]

7.4.3. Quantitative Analysis

Gravimetric Analysis
Inorganic Ingredients
Dry residue: [310, p. 434], [420]
Carbon dioxide: [341, Part 3, vol. IV aα pp. 225–229]
Silicon dioxide: [407, D 501–558], [421], [310, p. 447], [341, Part 3, vol. IV aα, p. 456]
Phosphorus pentoxide: [310, p. 452], [335, ISO/ TC 91 444F], [341, Part 3, vol. V aβ, p. 118], [422]
Sulfate: [333, H–III 8 b (76)], [423]
Magnesium: [310, p. 492], [341, Part 3, vol. II a, pp. 141–149]
Organic Ingredients: [376]
Ethanol solubles: [310, p. 148], [333, H–III 4 (65)]
Loss on drying: [310, p. 548], [420]
Unsulfated fraction (US): [310, p. 555], [335, ISO–893 (78) and ISO–894 (77)]
Nonsaponifiable fraction: [310, p. 555], [333, C–III 1 a and 1 b (77)]
Soap fatty acids: [310, p. 550], [333, G–III 6 b (57) and G–III 8 (61)], [376, p. 1352]

Nonionic surfactants: [310, p. 167], [333, H–IV 9 b]
Volumetric Analysis: [310, p. 215], [424, 425]
Acidimetry
Alkalinity: [310, pp. 334 and 472]
Sodium oxide content in soaps: [310, p. 554], [333, H–III 7 a (65)], [426]
Borate: [310, p. 474], [341, Part 3, vol. III, pp. 11–60]
Potentiometry: [427 – 429]
Phosphate: [310, pp. 455 and 480]

Borate: [310, p. 480], [430]
Chloride: [310, p. 446], [341,Part 3, vol. VII aβ, pp. 107–109]
Fluoride: [431, 432]
Nonionic surfactants in wastewater: [377]
Nonionic surfactants in detergents: [433]
Ionic surfactants: [434 – 441]
Ethylenediaminetetraacetate: [442]
Complexometry: [443, 444]
Ethylenediaminetetraacetate: [310, pp. 403 and 487], [335, ISO TC 91/490]
Nitrilotriacetate: [310, p. 403], [335, ISO TC 91/513 E]
Citrate: [445]
Calcium: [342, E 3], [446, p. 25]
Magnesium: [443, pp. 136–137], [446, p. 28], [342, E 4]
Nonionic surfactants: [447]
Conductometry: [448]
Sulfate: [449]
Two-Phase Systems: [310, pp. 215 and 218], [320, p. 427], [315, p. 16], [450, p. 221]
Anionic surfactants: [310, p. 234], [333, H–III 10], [401, 451, 452]
Cationic surfactants: [335, ISO 2871 (73)], [453]
Turbidity titration: [454]
Miscellaneous Titrimetric Analyses:
Sulfate: [310, p. 442], [315, H–III 8 a (76)], [333, H–III 8 a (76), p. 24; H–III 8 d (76)], [455, 456]
Chloride: [310, p. 445], [341, Part 3, vol. VII a, pp. 98, 99]
Water by Karl Fischer method: [310, p. 435], [333, H–III 3 a (65)], [335, ISO 4317 (77)]
Active oxygen: [310, p. 483], [335, R 607], [341, Part 3, vol. VI aα, pp. 193–207]
Active chlorine: [310, p. 487], [341, Part 3, vol. VII aβ II, pp. 74–76]
Calcium and magnesium binding capacity: [457, 458]
Gluconate: [459]
Bleach activator: [460]
Spectrophotometric Determinations: [397]
Perborate: [461, A 124]
Phosphate: [310, p. 457], [422, 462 – 464]
Enzymes: [465 – 467]
Enzymatic analysis: [468, 469]
Alkylpolyglycosides: [470]
Polycarboxylates: [471]
Trace elements: [397]
Anionic surfactants in wastewater: [423, 472 – 474]
Cationic surfactants in wastewater: [475]
Size Exclusion Chromatography:
Polycarboxylates: [476]
Flame Photometry: [477]
Sodium and potassium: [477]
Atomic Absorption Spectrometry: [478 – 481]
Traces of Cu, Fe, Zn, and Mn: [482, 483]
Traces of As: [484]
Traces of Co: [485]
Atomic Emission Spectrometry: [344, 486]
X-Ray Fluorescence: [316, 487 – 490]
X-Ray Diffraction: [491]
Phosphate phases I and II: [492]
Zeolite types: [493]
Radiometric Methods: [494 – 496]

7.4.4. Separation Methods: [497], [498, p. 21]

Capillary Zone Electrophoresis:
Anionic surfactants: [499, 500]
Polycarboxylates: [501]
Paper Chromatography: [502]
Phosphate types: [310, p. 460], [314, 503]
Alkanolamines: [504]
Alkanolamides: [310, p. 391]
Urea: [497, pp. 360 and 402], [502, p. 210], [505]
Thin-Layer Chromatography: [310, p. 45], [350, 506, 507]
NTA, EDTA: [508]
Akanolamines: [312]
Ether sulfates: [509]
Surfactant mixtures: [362, 363], [450, p. 47], [510]
Free fatty alcohols in ethylene oxide adducts: [511, 512]
Bactericides: [390, 513]
Amphoteric surfactants: [310, p. 371], [514]
Anionic surfactants: [515 – 517]
Optical brighteners: [518]
Nonionic surfactants in wastewater: [519, 520]
Column Chromatography: [310, pp. 51–63], [498, 521, 522]
Glycerides [376, Part I, p. 832], [523]
Fatty acid polyglycol esters: [524]
Petroleum sulfonates: [310, p. 61], [525]

Nonionic surfactants: [310, p. 167], [526]
Gas Chromatography: [310, p. 80], [527 – 530]
Fatty acid carbon-chain distribution: [411, 531 – 534]
Alkylbenzenesulfonate chain distribution: [535]
Fatty alcohol carbon-chain distribution: [536]
Solvents from cleansers: [537, 538]
Propellants in sprays: [539 – 541]
Dioxane in ethylene oxide adducts: [542]
Antimicrobial agents: [513]
Alkoxy content of cellulose ethers: [368]
Nitro musk compounds: [543]
Pentaacetylglucose: [544]
Fragrances by head space analysis: [545]
Supercritical Chromatography:
Nonionic surfactants: [546]
Anionic surfactants: [546]
Liquid Chromatography: [529, 547 – 549]
Poly(ethylene glycol) in nonionic surfactants: [550 – 552]
Unreacted alkylbenzenes and fatty alcohols in their sulfonation or esterification products: [553]
Bleach activators: [554, 555]
Fatty acid amides: [556]
Separation of fluorescent whitening agents: [518, 557 – 559]
Bacteriostatic substances in soaps: [560]
Nonionic surfactants: [529, 561 – 564, 566]
Anionic surfactants: [562, 563, 565, 567 – 569]
Separation of sulfonic acids: [570, 571]
Separation of lower alcohols: [572]
Separation of sulfur-containing surfactants: [573]
Phosphonates, polyphosphates: [574]
Ion Chromatography: [575 – 577]
Phosphates: [578 – 580]
Phosphonates: [578, 581, 582]
Capillary Electrophoresis: [583 – 585]

7.4.5. Structure Determination

UV Spectrometry: [310, p. 74], [586 – 589]
Alkylbenzenesulfonates: [334, Dd 3–60], [590 – 592]
Toluenesulfonate, xylenesulfonate, cumenesulfonate: [310, p. 205], [593, 594]
Structural information of surfactants: [525, 595]
Toluenesulfonamides: [596]
IR Spectrometry: [310, p. 69], [322, 597 – 604]
Petroleum sulfonates: [525]
Degree of branching in alkylbenzenesulfonates: [605, 606]
Surfactant sulfonates: [607, 608]
Soil antiredeposition agents, cellulose derivatives: [609]
NMR Spectrometry: [310, p. 86], [610, 611]
Ethylene oxide–propylene oxide adducts: [310, p. 301], [529, 612, 613]
LAS: [614, 615]
Mass Spectrometry: [310, p. 76], [616 – 619]
Fragrances (GC–MS): [620, 621]
Bactericides (TLC–MS): [310, p. 426], [626]
Nonionic surfactants (LC/FAB–MS): [622, 623]
Nonionic surfactants (MS): [624]
Nonionic surfactants (GC/CI–MS): [625]

7.4.6. Determination of Characteristic Values

Elemental Analysis: [627]
Nitrogen: [310, pp. 124 and 444], [627, p. 178], [628]
Halogen: [310, p. 128], [629, 630]
Sulfur: [310, pp. 126 and 439], [627, pp. 252 and 307], [630]
Phosphorus: [310, p.129], [397, pp. 899 and 902], [627, p. 360]
Fat Analysis Data: [376, 631]
Acid number: [310, p. 182], [333, C–V 2 (77)]
Saponification number: [310, p. 183], [333, C–V 3 (57)]
Iodine number: [310, p. 183], [333, C–V 11 a–d (53)]
Hydrogen iodide number: [333, C–V 12 (53)]
Hydroxyl number: [310, pp. 186 and 303], [333, C–V 17 a + b (53)]

7.4.7. Analysis Automation: [632]

Phosphorus pentoxide: [633 – 635]

Enzymes: [636, 637]

7.5. Sources of Information: [326, p. 403 (1981)]

Analytical Commissions: [638, 639]
Germany
GAT (Gemeinschaftsausschuß für die Analytik von Tensiden) [640]; member organizations of GAT:

a) DGF (Deutsche Gesellschaft für Fettwissenschaften) Varrentrappstraße 40–42, 60486 Frankfurt, Germany (http://www.dgfett.de).
b) TEGEWA (Verband der Textilhilfsmittel-, Lederhilfsmittel-, Gerbstoff- und Waschrohstoff-Industrie e.V., Karlstraße 21, 60329 Frankfurt, Germany, and the IKW (Industrieverband Körperpflege und Waschmittel e.V., Karlstraße 21, 60329 Frankfurt, Germany (http://www.tegewa.de).
c) NMP (Normenausschuß Materialprüfung im DIN) Burggrafenstraße 4–10, Postfach 11 04, 10787 Berlin, Germany (http://www.din.de/gremien/nas/nmp).

International Organizations
CESIO/AISE (Comité Européen d'Agents de Surface et Intermédiaires Organiques/Association Internationale de la Savonnerie, de la Détergence et des Produits d'Entretien) Working Group on Analysis
ISO (International Organization for Standardization)
IUPAC (International Union of Pure and Applied Chemistry) Commission on Oils, Fats, and Derivatives in the Applied Chemistry Division

Collections of Methods, and Sources from Which They May Be Obtained:
DGF-Einheitsmethoden.
Single pages: Deutsche Gesellschaft für Fettwissenschaften e.V., Varrentrappstraße 40–42, 60486 Frankfurt, Germany.
DIN Standards, ASTM Standards, ISO Standards.
Beuth Verlag GmbH, Burggrafenstraße 6, 10787 Berlin, Germany.
AOCS Official and Tentative Methods.
American Oil Chemists Society, 508 South Sixth Street, Champaign, Illinois 61820, USA.
ISO International Standards
ISO Central Secretariat, 1 Rue de Varembé, 1211 Geneva 20, Switzerland.

8. Test Methods for Laundry Detergents

All detergent test methods have the establishment of product quality as their goal. "Quality" is taken to include all those subjectively judgeable and objectively measurable properties of a product that play a significant role in its application.

Many possible approaches are available for measuring detergent quality, but these can be subdivided into three major groups:

Laboratory methods
Practical evaluation
Consumer tests

8.1. Laboratory Methods

Practical evaluations require careful statistical analysis of increasing numbers of test series, most of which are time-consuming and associated with high labor and material costs. For this reason, development of preliminary and screening test procedures has become necessary, as well as suitable laboratory methods for obtaining useful and relevant information, especially for development work. Such tests are designed to approximate service conditions to which the products will actually be subjected. The tests generally produce valuable clues to product quality, but their results should not be overrated, since the circumstances under which they are obtained are not totally identical to those that apply in the field. Normally these tests are conducted with laboratory equipment specially designed for small-scale tests (e.g., Launderometer [641], Linitest equipment [642], Terg-o-tometer [643], and various foam testing equipment [644, 645]).

In place of naturally soiled domestic laundry, use of artificially soiled and stained textiles and test swatches is standard in laboratory tests.

These swatches are prepared by using various standardized fabrics and knitware and various soils and stains. The swatches are often commercially available [646]. In addition, individual stains are prepared in the laboratory for specific purposes, either by hand or using special equipment. Artificially soiled fibers or fabrics must be prepared so that they respond selectively to different detergent components and different sets of washing conditions if adequate differentiation is to be assured. The amount of a certain soil on every swatch is identical and it is evenly distributed.

Criteria for testing detergents are subdivided as follows:

- Single wash cycle performance (soil and stain removal and bleaching)
- Multiple wash cycle performance, e.g., after 25 or 50 washes (soil antiredeposition properties, degree of whiteness, buildup of undesirable deposits, fiber damage, stiffness, color change, fluorescent whitening)
- Special characteristics (powder characteristics such as density, free flowability, dispensing in a washing machine, homogeneity, dusting properties, solubility, foaming, rinse behavior, and such storage characteristics as chemical and physical stability, hygroscopicity, color, odor, and tendency to form lumps)

The literature describes numerous methods for testing according to the above criteria, some of which are standardized. Standardization is a concern not only of national bodies (e.g., ANSI, the American National Standards Institute; JISC, the Japanese Industrial Standards Committee; DIN, Deutsches Institut für Normung [647]; AFNOR, Association Française de Normalisation; BSI, British Standards Institution), but also of international groups (e.g., ISO, International Organization for Standardization).

The above national organizations are all members of ISO and can, therefore, exercise influence on questions of international standardization [648]. Another particularly important organization concerned with international standardization of test methods is the CID (Comité International des Derivés Tensio-Actifs) with its subcommittee, the CIE (Commission Internationale d'Essai). This organization was disbanded in 1978, but in the meantime, its activities have been taken over and carried forward by the Working Group TMS (Test Methods for Surfactants) of the CESIO (Comité Européen d'Agents de Surface et Intermédiaires Organiques).

8.2. Practical Evaluation [649 – 653]

Practical wash evaluation is conducted with the exclusive use of commercial washing machines. In contrast to pure laboratory investigations and screening procedures, this work is carried out with laundry whose soil has been acquired naturally. The purpose of such experiments is to obtain results that can be regarded as far more realistic. Nonetheless, it is impossible to encompass the entire spectrum of conditions that might be encountered in practice. In many cases, so-called anonymous domestic laundry is tested: laundry of household origin, but not otherwise identified. Far better, however, even though more difficult to arrange, are wash-and-wear tests conducted by using specific new fabric items, e.g., terry towels, dish towels, shirts, pillow cases, etc., that have been distributed for regular use during a certain period (usually a week) in a sufficient number of representative households. After each use cycle, the samples are collected in the laboratory and washed with the detergent being evaluated. The same item must always be washed with precisely the same detergent and under identical washing conditions. The number of necessary wash-and-wear cycles depends on the nature of the questions being investigated. Nonetheless, usually ca. 15–25 wash-and-wear cycles are necessary, occasionally even more. The detergent is generally used according to the manufacturer's directions, although effects of under- and overdosing may also be taken into account. Analytical precision is improved if instrumental methods are used to evaluate test results, but such measurements are sometimes difficult both to interpret and to conduct. Thus, an alternative evaluation is based primarily on visual evaluation, with physical measurements used only in a supporting role. Recommendations for comparative testing of detergent use characteristics have been the subject of a tentative DIN standard [654], as well as an ISO standard [655].

In the interest of both manufacturers and especially consumers, comparative use tests must be conducted under the most realistic conditions.

This is particularly true of tests designed for consumer education, such as those carried out by consumer organizations; otherwise the results are of little value. Therefore, good communication and close contacts between the consumer organizations and the product manufacturers regarding test methods and criteria to be applied prior to conducting the tests is of paramount importance. Oversimplified tests, e.g., those using artificially soiled swatches only, are incapable of making the sorts of distinctions that consumers consider relevant, and are a disservice to all parties concerned. Ultimate product judgements must be based not only on wash results, but on all of the relevant factors affecting use value.

8.3. Consumer Tests

To obtain a final degree of certainty with respect to a detergent's use characteristics and acceptance, consumer tests are conducted in addition to laboratory tests as a regular control to determine consumer acceptance. In such tests, consumers subject a product to their own sets of conditions; i.e., they use their own washing machines on their own laundry and soil, and it is assumed that they introduce the detergent in the manner to which they are accustomed. The precision of such in-home use tests is of course limited, because consumers judge, they do not measure. Therefore, it is necessary that a large number of consumers (usually 200–300 households) participate if statistically meaningful conclusions are to be drawn. Experience has shown that such tests are of high value in addition to laboratory results, both for determining product weaknesses and for identifying particular strengths. By their very nature, consumer tests subject a product to a great many different sets of circumstances, and they lead to very meaningful results. Thus, consumer tests are very frequently conducted prior to marketing a new product.

9. Economic Aspects

The worldwide economic role of detergents represents a factor of considerable importance. Nonetheless, laundry detergent consumption per capita varies tremendously from country to country (Fig. 91 [216]). Changing patterns of use for laundry detergents and increased standards of living over the last 40 years reveal a remarkable growth in worldwide consumption, with the absolute quantity rising from approximately 10×10^6 t in the late 1950s to 21.5×10^6 t worldwide in 1998.

Total consumption of laundry, dishwashing, and cleaning detergents in Europe was 6.4×10^6 t in 1991. Eight years later it was 6.8×10^6 t, corresponding to a volume growth rate of nearly 1 % per year (Table 42 [656]).

Table 42. World total surfactant consumption by end use, 1992 and 1998 (10^3 t) [657]

End use	1992	1998
Household		
North America	866	1115
Western Europe	1061	1119
Asia	1208	1599
Other regions	998	1235
Total	**4133**	**5068**
Personal care		
North America	204	210
Western Europe	143	163
Asia	119	194
Other regions	86	155
Total	**552**	**722**
Industrial and institutional cleaners		
North America	263	311
Western Europe	155	192
Asia	60	75
Other regions	42	53
Total	**520**	**631**
Industrial process aids		
North America	1191	1316
Western Europe	644	676
Asia	858	1085
Other regions	355	364
Total *	**3048**	**3441**
Grand total	**8253**	**9862**

* Does not include soap.

Only a handful of large companies manufacture most of the world's laundry detergents and cleansing agents. The leaders are Procter & Gamble (Cincinnati, Ohio, USA) and Unilever (London, UK and Vlaardingen, The Netherlands). Further internationally operating major manufacturers are Colgate Palmolive (New York, USA), Henkel (Düsseldorf, Germany), Reckitt Benckiser (Windsor, UK), Kao (Tokyo, Japan), and Lion (Tokyo, Japan). In the U.S. market, four firms (Procter & Gamble, Lever, Dial, and Colgate) dominate. The remainder

is split among many smaller local or regional producers. In Europe, for example, the Association Internationale de la Savonnerie, de la Détergence et des Produits d'Entretien (AISE), a trade association made up of European national detergent industry groups, estimates that its members represent some 1200 individual firms.

9.1. Detergent Components

Production capacity and demand for the major detergent raw materials by volume are briefly reviewed in the following sections.

9.1.1. Surfactants [657]

World surfactant consumption was estimated at 9.9×10^6 t in 1998, corresponding to an increase of 20 % since 1992 (Table 42). The two largest end uses, household products and industrial processing aids, together account for for 86 % of the total. Household products alone accounted for nearly 5.1×10^6 t, or half of the world total. Household products include heavy-duty powder and liquid detergents, light-duty liquids, and fabric softeners. The regional distribution of surfactants consumption for houshold use in 1998 is shown in Figure 119.

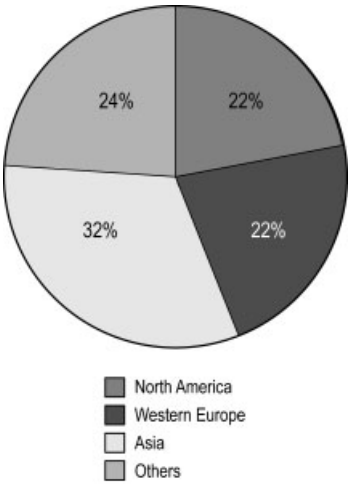

Figure 119. Surfactant demand for household use in 1998 [657]

Industrial processing aids cover five major end uses: plastics and elastomers, textiles, agricultural chemicals, leather and paper chemicals, and other miscellaneous uses. The other major categories featured include personal care products and industrial and institutional (I& I) cleaners, which together consumed about 1.4×10^6 t of surfactants in 1998. I& I products include laundry products, hard-surface cleaners, and dishwashing detergents. From data shown in Figure 120 it is obvious that the two major sectors of the total world surfactant market are the household and I& I areas. Surfactant consumption grew most rapidly in Asia in the 1990s (> 30 %), whereas Europe showed only a moderate growth (< 10 %) in that period.

Breakdown of total surfactant consumption by type is depicted in Table 43. Apart from soap, which is the world's largest surfactant by volume (worldwide consumption 8.8×10^6 t in 1999 [658]), surfactant consumption is headed by anionic surfactants, viz. linear alkylbenzenesulfonates, alcohol sulfates, and alcohol ethoxy sulfates. Second in consumption are alcohol ethoxylates and alkylphenol ethoxylates. LAS has remained the most widely used synthetic surfactants. They have weathered the environmental concerns and experienced a comeback in Europe. They continue to dominate in the Americas and Asia.

Table 43. World major surfactant consumption by type, 1992 and 1998 (10^3 t) [657]

Surfactant type	1992	1998
Linear alkylbenzenesulfonates	2385	3027
Branched alkylbenzenesulfonates	411	198
Alcohol sulfates	466	479
Alcohol ethoxysulfates	511	911
Alcohol ethoxylates	742	849
Alkylphenol ethoxylates	652	701
Quaternaries	312	434
Others *	2774	3263
Total	8253	9862

* Does not include soap.

9.1.2. Builders

Sodium triphosphate has been the most important detergent builder in powder detergents since

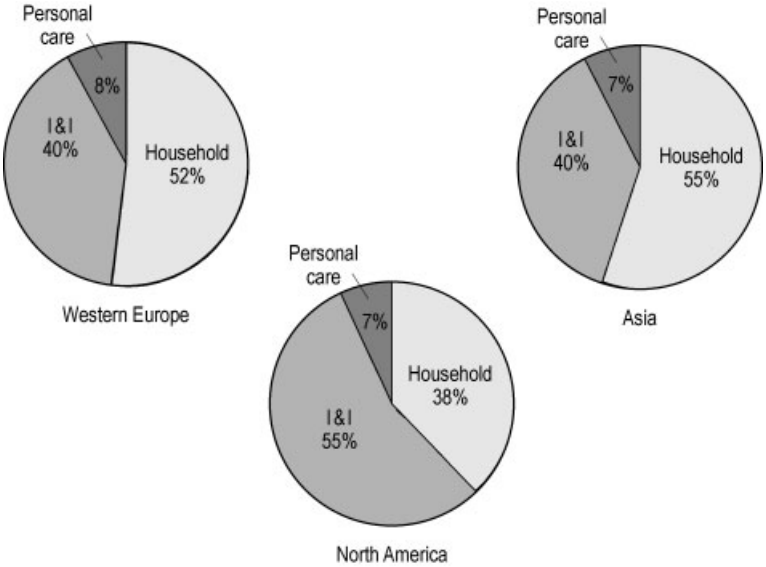

Figure 120. Total surfactant end user markets in 1998 [657]

the 1950s worldwide. Its use has become a subject of intense public discussion in many countries in the last 30 years as a result of phosphate-induced regional overfertilization of surface waters. In countries where overfertilization has been particularly severe, this has resulted in both voluntary agreements and legal restrictions, which have caused a sharp decrease in the use of sodium triphosphate for washing and cleansing purposes, mainly in the USA, Europe, and Japan. Within ten years, the use of sodium triphosphate in detergents dropped by more than 50 % in these regions [659, 660]. Worldwide production capacity has risen dramatically for the most important phosphate substitute zeolite A. Figure 121 shows the increase in production capacities for zeolite A in Western Europe in the 1980s and 1990s [169]. Additional major manufacturing capacities are located in the USA, Japan, and Taiwan. Meanwhile, production capacities for detergent-grade zeolites have largely surpassed demand.

Quantitatively, detergent builder zeolites represent the largest application field for zeolites [660]. Almost 90 % or 1 050 000 t/a of zeolites produced worldwide in 1994 were used for detergents. Of this, ca. 950 000 t/a were zeolite A and some 100 000 t/a were of the zeolite P type.

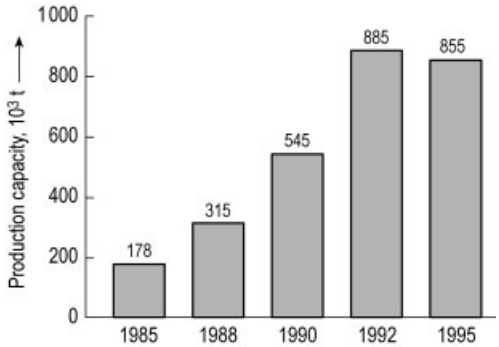

Figure 121. Production capacities of zeolite A in Europe (10^3 t) [169]

In Japan, for example, since the mid-1980s more than 95 % of all detergents produced and sold are phosphate-free, containing zeolite instead. Also in the USA meanwhile nearly all laundry detergents are nonphosphate products. A large portion of them contains zeolite A as the main builder. The situation in Europe is illustrated in Figure 99 [218]. Here, too, zeolite is the main builder for phosphate-free laundry products. The replacement of phosphate by zeolite A has gained additional momentum as a result of changes in the prices for raw materials.

The economic situation with respect to other detergent ingredients is highly variable from continent to continent. Moreover, these mate-

rials are used in smaller quantities. Hence, their economic impact is less significant and will not be dealt with here.

9.2. Laundry Detergents

In 1998, total world production of soap bars and laundry detergents was 21.5×10^6 t. Again, the tonnage varies among the global regions and individual countries (Table 44).

Worldwide, powder detergents continue to constitute the largest section of the total textile washing and cleaning products market, both in volume and value terms. The powder detergents category is prominent due to the well-established nature of the standard powder delivery form, and latterly the market performance of the tablet laundry detergents, introduced into several markets in the late 1990s and 2000. Exceptions are the USA, where liquid laundry products took over the lead in the 1990s; Indonesia, whose market is dominated by detergent pastes; and the Philippines, where the vast majority of the consumers use soap bars. India is the world's largest single market by volume for soap bars, but here, too, powders dominate and continue to grow at the expense of bars.

In terms of value, total textile washing products amounted to sales of $ 17.7×10^9 in the seven major single markets (USA, Japan, Germany, UK, France, Italy, and Spain) in 1999 [219]. These markets are best characterized by having a high degree of maturity, along with fierce price competition and high advertising expenditures. Four of these markets registered positive gross value growth over the period 1995 through 1999. This translates into a positive real value growth of the U.S. market of almost 13 %, whereas Germany, Italy, and Japan recorded real value declines of 6.5 %, 11.7 %, and 7.6 %, respectively. Due to their high levels of saturation and maturity, continuous product innovations play an important role in stimulating these major markets.

Within the powder detergents sector, conventional products claim most of the sales in France, Italy, Spain, and the UK, whilst the (super)concentrated powders are more popular in Germany, Japan, and the USA for various reasons, such as space-constraints in the home (Japan), reduced packaging costs, space-saving benefits, and consumers' greater environmental awareness (Germany), as well as consumer perceived greater convenience and product benefits vs. retail price (USA).

In the liquid detergent sector, concentrated products form the larger subsector in Germany, Japan, the UK, and the USA. Conventional liquid detergents lead in France, Italy, and Spain.

Within the category of other detergents, fine-fabric specialty detergents lead in most of the seven major markets, ahead of hand-washing detergents and detergent bars. Changing lifestyles and increased saturation of the washing-machine market have contributed to considerably reduced consumption of products of the last two subsectors.

9.3. Fabric Softeners

Europe is the major consumer of rinse-cycle fabric softeners. A number of reasons explain this, including the fact that European tap water on average is relatively hard and the fact that the vast majority of drum-type washing machines are equipped with separate dispensers for laundry aids in Europe. Apart from the USA and Japan, products in this class continue to be relatively insignificant in volume in the restof the world. Figure 122 provides information about European fabric softener consumption for the year 1999. These consumption figures represent both conventional fabric softeners, which contain 4 to 8 % active ingredients, and concentrated softeners, which usually have 12 to 30 % active cationics. After a period of decreasing consumption in the 1980s, fabric softeners enjoyed growth in both value and volume in the 1990s.

Concentrated fabric softeners have dramatically increased their market share at the expense of diluted conventional softeners. The shares of concentrated softeners in Europe in 1998 varied from some 20 % in Italy to 98 % in Germany. Use of concentrated products saves packaging materials and transportation costs and reduces the amount of waste. To further decrease waste, part of the concentrated fabric softeners are sold as refill packages in many European countries. Their share of total concentrates varies from 10 % in Belgium to some 65 % in France. Fabric softener sheets, which are very popular in the USA as laundry dryer aids, find only very limited

Table 44. World production of soap bars and powder, liquid and paste laundry detergents in 1998, 10^3 t

Country	Soap bars	Powder detergents	Liquid detergents	Paste detergents	Total laundry detergents
Austria	-	61	7	-	68
Belarus	-	-	31	-	31
Belgium	-	226	46	-	272
Bosnia-Herzegowina	-	2	-	-	2
Bulgaria	-	7	-	-	7
Croatia	-	31	2	-	33
Cyprus	-	4	-	-	4
Czechia	10	306	10	-	326
Denmark	-	29	5	-	34
Finland	-	26	-	-	26
France	-	467	111	-	578
Germany	-	578	46	-	624
Greece	11	48	5	3	64
Hungary	-	65	10	-	78
Ireland	5	41	1	-	47
Italy	20	485	190	1	695
Macedonia	1	20	1	-	23
Norway	-	6	-	-	6
Poland	31	196	10	-	237
Portugal	27	91	10	-	128
Rumania	-	42	1	5	43
Russia	70	340	6	-	421
Serbia	-	61	-	-	61
Slovakia	2	20	-	1	22
Slovenia	-	21	1	-	23
Spain	6	406	34	-	446
Sweden	-	9	4	10	13
Switzerland	-	41	-	1	51
The Netherlands	-	31	5	-	37
United Kingdom	-	337	80	-	417
Total Europe	*183*	*3997*	*616*	*21*	*4817*
Total North America	-	*1493*	*1493*	-	*2986*
Argentina	62	143	18	-	223
Brazil	550	640	10	-	1200
Mexico	241	846	41	4	1132
Others	544	584	13	10	1151
Total Latin America	*1397*	*2213*	*82*	*14*	*3706*
Congo	57	81	-	-	138
Egypt	220	350	-	1	571
Kenya	122	82	10	31	245
Nigeria	151	82	-	-	233
South Africa	63	163	9	2	237
Others	120	269	5	1	395
Total Africa	*733*	*1027*	*24*	*35*	*1819*
China	274	1508	13	107	1902
India	965	1500	9	-	2474
Indonesia	42	180	1	260	483
Iran	5	265	10	-	280
Japan	23	522	50	-	595
Philippines	244	61	-	-	305
Saudi-Arabia	12	122	-	4	138
South Korea	35	109	2	-	146
Turkey	82	230	2	-	314
Others	574	719	56	64	1413
Total Asia	*2256*	*5216*	*143*	*435*	*8050*
Total Australia	*1*	*115*	*44*	-	*160*
World Grand Total	*4570*	*14061*	*2402*	*505*	*21538*

(Source: Ciba Speciality Chemicals Ltd., Switzerland)

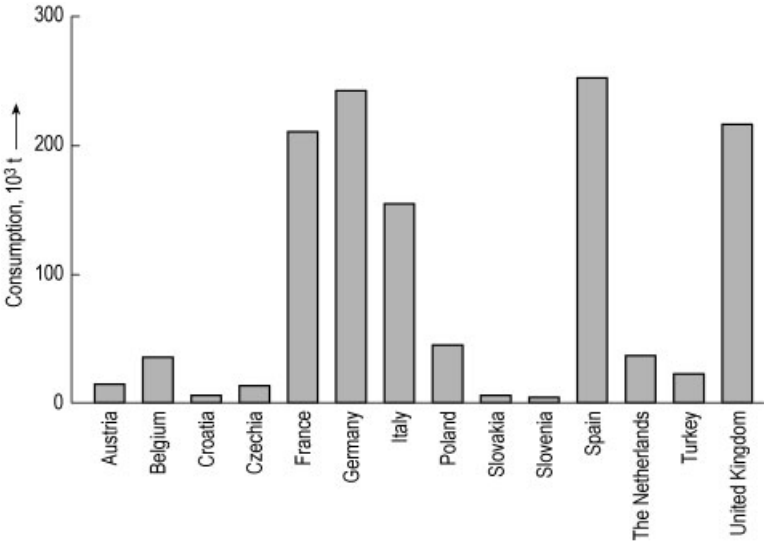

Figure 122. Comsumption of fabric softeners in Europe in 1999 (source: Leading International Market Research Institute)

use in Europe and Japan. The share of softener sheets ranges from 0 to 5 % of the fabric softener market in European countries.

Total consumption of fabric softeners in Japan was some 250 000 t in 1998, about 40 % of which were concentrated products.

9.4. Other Laundry Aids

In the seven major detergent markets, laundry aids comprise a wide number of subcategories, with spot and stain removers and laundry bleach claiming the leading positions in most key markets. Although chlorine bleaches have been used for textile washing for many years, the introduction of color-safe, non-chlorine bleaches led to a considerable consumption of this subcategory in the 1990s. Laundry bleach accounts for the largest share of sales in the USA, Spain, and Italy.

10. Ecology

10.1. Laundry, Wastewater, and the Environment

The washing process represents a complex interaction between soiled laundry, water, mechanical and thermal energy, and detergents. Laundry is repeatedly cycled through the system, and clean laundry can be regarded as the product. Wastewater is the troublesome byproduct which has the potential for causing a number of undesirable effects in sewage treatment plants and in the environment. Virtually all of the clean water brought into the process is later released to the sewage system in the form of contaminated wastewater containing additional energy (heat), soil from the laundry, lint, dyes, finishing agents, and detergents. Detergent components are released to the wastewater either in essentially unchanged form or as the products of reaction with other materials present.

Laundry wastewaters vary considerably in concentration and composition. These differences arise partly due to variations in laundry soil levels, although washing technology and the composition and amount of added detergent are also significant factors. The differences are particularly great between the wastewaters generated by household laundry and those from I& I laundries. The latter generally contain lower portions of contaminants originating from detergents, largely as a result of more economical use of typical I& I laundry products and the use of water-softening equipment. Little doubt exists that laundry wastewater must be generally regarded as a heavily contaminated medium; it should not be returned to receiving waters in untreated form.

Fortunately, the wastewater emanating from the hundreds of millions of household washing machines and from washing by hand worldwide is widely distributed, and in a pattern consistent with that of general water consumption. As a result of dilution in the public sewerage system and in sewage treatment plants, both the temperature and the relatively high pH value of wastewater are considerably decreased, and the effect of peak loads is also somewhat compensated for. Only for this reason is biological treatment of laundry wastewater in a normal sewage treatment plant feasible. If laundry wastewater were treated separately, major problems would be encountered in dealing with the load of organic pollutants introduced by household and commercial laundry operations. Dilution of laundry wastewater within a typical public sewage system normally exceeds a factor of ten.

10.2. Contribution of Laundry to the Sewage Load

Table 45 presents calculated estimates of average theoretical concentrations of the detergent and cleanser components in public wastewater. These estimates are derived from statistical data collected in Germany and depend on both product consumption and the average water consumption in households. These calculated average wastewater concentrations are "worst-case" assumptions, since the average per capita wastewater production rate (combined household and commercial) is around 200 L/d in the EU member states.

In addition, the calculated concentrations of some of the chemicals are only partly due to the use in laundry detergents since other applications (e.g., dish washing and cleansing products) may contribute to a considerable extent. Furthermore, actual values for concentrations and loads show considerable variation. Time of day is a factor, as is the day of week (with Saturdays and Mondays producing the highest values). Even seasonal variations are apparent, but all of these fluctuations are reasonable when examined in the light of known household laundry habits [662].

A significant portion of domestic water is used in the household laundering operation. According to [663], the per capita amount of laundry wastewater generated daily is around 17 L, which translates into 13 % of the domestic water consumption or about 8 % of the total hydraulic load reaching a sewage treatment plant. Not only in terms of the hydraulic load but also with respect to the chemical load, the laundry washing process is a major factor influencing municipal wastewater. As shown in Table 46, the detergents contribute about 10 % to the COD, the biodegradable organic matter (usually reported as the BOD_5), and the dry solid matter content of the municipal wastewater load, while the contribution to the nitrogen and phosphorus load is low. In addition, of course, the soil contained in soiled laundry represents another important contributor to the load of wastewaters. About half of the biodegradable contaminants load of mixed wastewater derived from laundering is due to laundry soil and stains. The amount of organic contamination originating from soil is doubled or tripled in institutional laundry wastewaters, compared to laundry wastewater from households.

10.3. Detergent Laws

Detergents represent a typical and important group of environmental chemicals, i.e., substances entering the environment after their use. The per capita consumption of detergents is nowadays at a level considerably below the figures of the 1980s (1980: 33.3 g per capita and day in Western Germany; 1997: 21.1 g per capita and day in Germany). Nevertheless, the 1999 consumption was in the range of 660 000 t/a (Table 45). These large amounts of chemicals go down the drain and enter receiving waters directly or after sewage treatment. Since the wastewater treatment situation was relatively poor in most European countries a few decades ago, it is not surprising that certain detergent ingredients, i.e., surfactants and phosphates, exhibited an apparent specific impact on the aquatic environment and became subject of specific laws and regulations.

10.3.1. Development of the European Detergent Legislation

One of the earliest pieces of detergent legislation was the German Detergent Law of 1961

Table 45. Consumption figures and average wastewater concentrations of ingredients of household detergents and cleansers (Germany, 1999)

Product ingredient	Consumption, t/a [661]	Per capita consumption, g/d[a]	Calculated average wastewater concentration, mg/L[b]
Total ingredients	688 000	23.0	177
Anionic surfactants	104 700	3.50	26.9
Nonionic surfactants	53 000	1.77	13.6
Cationic surfactants	30 400	1.02	7.8
Zeolites	139 000	4.64	35.7
Polycarboxylates	15 700	0.52	4.0
Na carbonate (soda ash)	98 800	3.30	25.4
Phosphates	23 600	0.79	6.1
Na citrate	15 600	0.52	4.0
Na perborate tetrahydrate	49 500	1.65	12.7
Na percarbonate	27 200	0.91	7.0
TAED	13 100	0.44	3.4
Phosphonates	2 900	0.10	0.7
NTA	400	0.01	0.1
Carboxymethyl cellulose	2 200	0.07	0.6
Dye transfer inhibitors	400	0.01	0.1
Silicates	22 300	0.74	5.7
Enzymes	6 000	0.20	1.5
Optical brighteners	500	0.02	0.1
Paraffins	1 200	0.04	0.3
Soil repellents	600	0.02	0.2
Perfumes	5 700	0.19	1.5
Dyes	100	0.003	0.03
Na sulfate	74 600	2.49	19.2

[a] *a* Population in Germany (1998): 82×10^6 [661].
[b] *b* Average per capita water consumption in households in Germany: 130 L/d (1998) [661].

Table 46. Contribution of laundry washing (without laundry soil) to the chemical load of municipal wastewater

Chemical load parameter	Concentration in laundry waste water [664], mg/L	Per capita load of laundry wastewater * , g/d	Per capita load of municipal wastewater [664], g/d	Contribution of laundry washing to the chemical load in wastewater, %
COD	600	10.2	120	8.5
BOD_5	350	5.95	60	10.0
Dry solid matter	450	7.65	72	10.6
Total nitrogen	7	0.12	12	1.0
Total phosphorus	2	0.03	2.4	1.0

* Derived from a water consumption of 17 L/d for laundry washing [664].

[44]. With its specific requirement for a minimum level of biodegradability, this law subjected a fundamental product parameter to legal control for the first time. The trigger for setting minimum biodegradability standards for synthetic detergents was the fact that the first generation of synthetic anionic surfactants was represented by the poorly biodegradable tetrapropylenebenzenesulfonate. The highly branched chemical structure of the anionic surfactant prevented its microbial degradation in sewage plants and rivers and caused formation of large mountains of foam in sewage treatment plants, weirs, locks and rivers in Germany, particularly during the droughtlike summers of 1959 and 1960. The law and the statute that followed (1962) [45] placed a strict requirement of at least 80 % biodegradability on all anionic surfactants. This legal move successfully displaced TPS from detergent formulations by the readily biodegradable linear alkylbenzenesulfonates. Thus, by 1965 foaming problems in German sewage treatment plants and surface waters had been eliminated.

In the late 1960s most of the European countries had limited the use of TPS in household

detergents. Based on an agreement of the Council of Europe [665] requiring all types of surfactants to be biodegradable, the EEC released two directives on the biodegradability of surfactants in 1973: The EEC Directive 73/404/EEC stipulated an average biodegradability of at least 90 % for all types of surfactants used in detergents, i.e., anionic, nonionic, cationic and amphoteric surfactants [666]. The Directive 73/405/EEC described the biodegradability test methods to be applied for anionic surfactants and specified a minimum of 80 % degradation in a single test [667].

On the national level these European Directives were translated into detergent laws by the member states, e.g., in Italy in 1974, France in 1977, The Netherlands in 1977, and the UK in 1978 [668]. In Germany the detergent law of 1975 [669] and its statutory orders of 1977 [670] and 1980 [671] regulated the "environmental compatibility of laundry detergents and cleansers". According to § 1 (1) of the German Detergent Law placement of detergents and cleansers on the market is permitted only if the absence of all avoidable deterioration in the quality of surface water can be ensured. Particular attention is paid to the potential role of surface water in the drinking water supply and to problems related to the operation of sewage treatment plants. The general requirements and authorizations for implementation apply directly to the following areas:

1) Unambiguous product labeling, including qualitative declaration of all significant components present
2) Deposition of frame formulations for all products at the Federal German Environmental Agency
3) Utilization only of biodegradable organic components, particularly surfactants and other organic compounds as specified in a statutory order
4) Provision of directions for dosage of detergents on all packages, including variations applicable to water of differing hardness, where hardness is defined in terms of four ranges
5) Obligation on the part of local water supply authorities to publish the hardness characteristics of their water
6) Limitations on permitted phosphate levels, including some cases of a total phosphate ban, provided that ecologically sound substitutes are available (see Section 10.3.4)

Products covered by the Detergent Law include: household laundry detergents; household cleansers; dishwashing agents; rinsing and other laundry aids; detergents for I& I laundries; industrial cleansers; and cleansing agents employed in the leather, tanning, paper, and textile finishing industries that might be released into wastewater.

Thus, the definition of detergent products according to the German Detergent Law is more comprehensive than the EU definition, that only covers surface active substances which are main components in detergents and designed for cleaning purposes [668].

10.3.2. Regulatory Limitations on Anionic and Nonionic Surfactants

The Directive 73/405/EEC [667] and its updated version Directive 82/243/EEC [672] specify the minimum biodegradability of anionic surfactants contained in detergents. In addition, the procedures are described to be applied for measurement of biodegradability. In close analogy the Directive 82/242/EEC [673] specifies an 80 % biodegradation pass limit for nonionic surfactants used in detergents, and the test methods to be used. A corresponding national regulation has existed in the Federal Republic of Germany already since 1977 [670] ("Tensidverordnung"). The regulation requires a minimum of 80 % biodegradability for the anionic and nonionic surfactants present in a packaged detergent. Anionic surfactants are determined as "methylene blue active substances" (MBAS), i.e., materials forming a chloroform soluble complex with the cationic dye methylene blue. Nonionic surfactants are defined as "bismuth active substances" (BiAS), i.e., materials forming an insoluble complex with the bismuth-containing Dragendorff reagent [674].

Although not all anionic and nonionic surfactants can be analytically measured by the MBAS and BiAS method [675], these parameters are used for the evaluation of the biodegradability according to the legal requirements. Essentially, the percentage of MBAS or BiAS loss describes

the disappearance of a specific analytical reaction of the parent surfactant. This loss of reactivity is due to microbial transformation of the chemical structure of the surfactant, resulting in the loss of typical surfactant properties such as surface activity and foaming power. Therefore, it is called "primary or functional biodegradation" (see Section 10.4.2.1).

10.3.3. Primary Biodegradation Test Procedures

According to the Directives 82/243/EEC and 82/242/EEC and corresponding national legislations two types of test methods are applicable for determining the (primary) biodegradability of anionic and nonionic surfactants, respectively. The first type is a discontinuous shake flask test operating with a low bacterial inoculum, that is incubated with 5 mg/L of MBAS or BiAS as the sole carbon source. In this *OECD Screening Test* [676] the loss of MBAS or BiAS is determined periodically up to 19 d. The results are compared with the degradation behavior of two control substances: the poorly degradable TPS (< 35 % MBAS decrease) and the readily biodegradable LAS (about 92 % decrease). The biodegradability determination in this test is illustrated in Figure 123. The second test type is represented by the *OECD Confirmatory Test* [676] simulating the biodegradation process occurring in an continuous activated sludge plant (Fig. 124). In this procedure, MBAS (20 mg /L) or BiAS (10 mg /L) is dissolved in synthetic sewage (containing peptone, meat extract, urea, and mineral salts) and fed continuously into a model sewage treatment plant with a hydraulic retention time of 3 h. After inoculation of the test system with the effluent of a predominantly domestic sewage treatment plant and after growth of the activated sludge, the effluent is analyzed periodically. This running-in period lasts six weeks at maximum and is followed by a 21-d evaluation period after the degradation rate has become regular. The average degree of biodegradation during the evaluation phase is calculated as a mean value of at least 14 separate results based on samples from 24-h collection periods (Fig. 125).

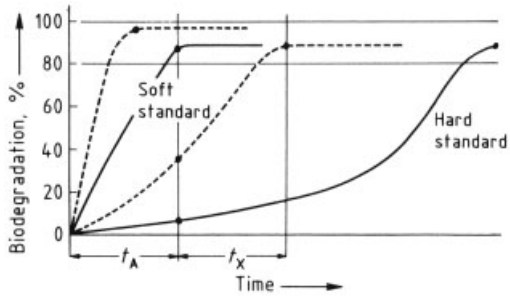

Figure 123. Biodegradability evaluation in the OECD Screening Test
t_A: maximum 14 d; $t_A + t_X$: maximum 19 d
— examples for two surfactants; one is biodegrated more easily (t_A) than the second one ($t_A + t_X$)

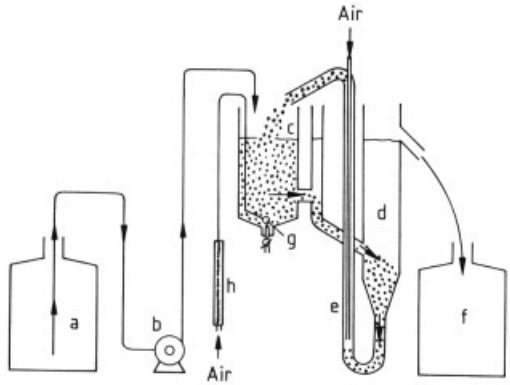

Figure 124. Experimental arrangement for the OECD Confirmatory Test as specified by the German Detergent Law
a) Sample container; b) Dosage pump; c) Activated sludge vessel (capacity 3 L); d) Settling vessel; e) Air lift; f) Collection vessel; g) Fritted disk; h) Air flow meter

The discussed test types represent a hierarchical system of the biodegradability evaluation. Usually, surfactants are evaluated on the basis of die-away screening tests such as the OECD Screening Test or the French AFNOR Test. In addition, the British Porous Pot Test, which is a continuous sewage works simulation test, may be used at this stage of data generation [672, 673]. However, if the degradation proves MBAS or BiAS loss lower than 80 %, or if doubt remains, a subsequent confirmatory test is required, and the outcome of this OECD Confirmatory Test is regarded as definitive.

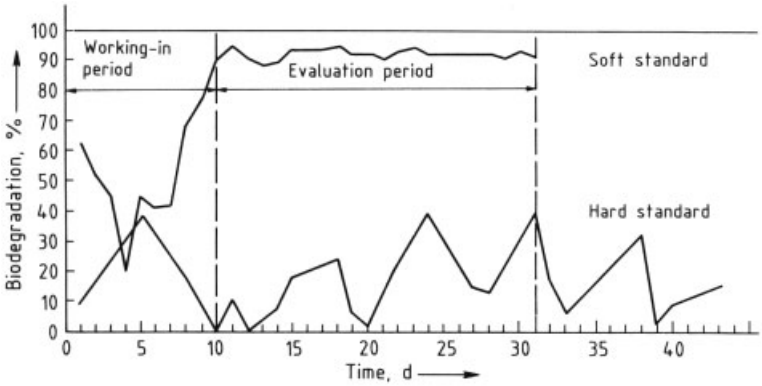

Figure 125. Determination of a rate of biodegradation according to the method specified in the OECD Confirmatory Test

10.3.4. Regulation of Maximum Phosphate Content in Detergents

The detergent regulations at the European Union level address exclusively the surfactants and their biodegradability. However, at national levels, regulations or voluntary agreements on the phosphate content of detergents exist in many countries, e.g., Austria, Germany, Italy, The Netherlands, Norway, Sweden, Switzerland, USA, Canada, and Japan.

In itself, phosphate is not harmful but is a natural and essential macronutrient for all living organisms. Normally, the phosphate concentration of surface waters is so low that it is a limiting factor for growth of algae and higher plants. Consequently, when excess phosphate is released into the aquatic environment, the resulting overfertilization leads to increased growth of algae. Due to the concomitant secondary processes (organic load of waters, oxygen depletion after the organic bio-mass is microbially degraded) the overall water quality may be considerably reduced. Detergent phosphates released with laundry wastewater are quickly converted into orthophosphate. Thus, the use of sodium triphosphate in detergents came under critical scrutiny although many other phosphate sources exist that contribute to eutrophication of surface waters.

According to a profound study on phosphates in Germany in 1975, some 60 % of the phosphate contained in municipal sewage originated from detergents and cleansers [677]. As a consequence of the partial removal of phosphate in sewage treatment plants and the input of phosphates by other sources (human excretion, food industry, agricultural fertilizers), the share of detergent phosphates in surface waters was estimated to be about 40 %. This balance showed already that the reduction of phosphates in detergents is an important but not the sole factor in solving the eutrophication problem of surface waters.

The most comprehensive approach involves chemical elimination of phosphates in the sewage treatment plant (tertiary treatment) removing the total phosphorus content of the wastewater. This approach is realized in some countries to a greater or lesser extent. Nevertheless, phosphate reduction in detergents, i.e., at its source, provides immediate relief to receiving waters and, ultimately, also to the coastal areas of the seas, which are increasingly confronted with eutrophication problems. The phosphate regulation of 1980 [671] implementing the pertinent authorization of the German Detergent Law of 1975 (see Section 10.3.1) had a considerable impact on the development of the phosphate loads in German rivers (Fig. 126). The phosphate reduction in detergents was enforced in a two-step decree starting with an overall relative decrease of 25 % of the detergents' phosphate in 1981 and around 50 % in late 1983. The second step required the availability of suitable phosphate substitutes for detergents (see Section 10.5.2). Their good performance in phosphate-reduced and -free detergents and the acceptance of nonphosphate detergents by the consumer has meanwhile led to a complete substitution of phosphate-containing detergents by phosphate-free ones in many countries, for example in Eu-

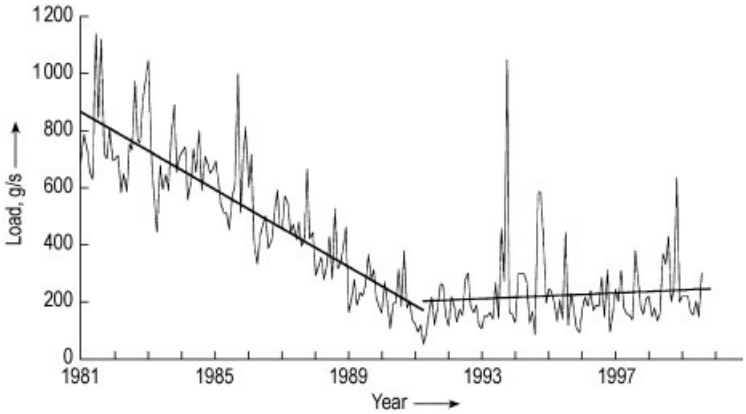

Figure 126. Variations of the orthophosphate load in the river Rhine from 1981 to 1999 (sampling site: Düsseldorf-Himmelgeist, Germany)

rope in Germany, the Scandinavian countries, Italy, Austria, The Netherlands, and Switzerland (see Fig. 99). Thus, the legal requirements on reduction or banning phosphates in detergents have been superseded in many cases by the industrial supply of efficient phosphate-free detergents and their acceptance by the consumer.

The legislative pressure towards phosphate reduction in detergents accompanied by the availability of suitable substitutes has resulted in a noticeable quality improvement of a number of surface waters today [678]. Phosphate balances of the river Rhine in 1979 and 1989 showed that the measured phosphorus load reduction of about 28 400 t/a within this period corresponded very well with the expected phosphorus reduction due to the use of phosphate-reduced or -free detergents (26 700 t/a) [679].

10.4. General Criteria for the Ecological Evaluation of Detergent Chemicals

The history and the development of the detergent legislation in Europe impressingly reflects the environmental relevance of this group of anthropogenic chemicals (xenobiotics) encountered already in times when the terms "environmental safety" or "environmental risk assessment" were not yet used. The lesson of the so-called detergent problem of the late 1950s was learned both by the legislators and the detergent industry long before general criteria were developed and applied for the ecological evaluation of chemicals. Now, such broadly applicable instruments and standards have been available for several years, e.g., the risk assessment directives for new and existing chemicals of the European Union [680, 681] and the accompanying Technical Guidance Documents [682] that provide an evaluation system applicable by authorities and industry as well. This chapter deals with the concept and the methods of the environmental compatibility assessment that form the basis for the evaluation of chemicals in general.

10.4.1. Concept (see also → Ecology and Ecotoxicology, Chap. 3)

Production, use, and disposal of chemical products has a more or less significant impact on the environmental compartments water, soil, and air. To secure that such emissions will not negatively influence the biota of these compartments it is necessary that a possible ecological risk is already recognized before the product enters the environment. The environmental safety of chemical products is essentially based on the safety of their individual ingredients. Thus, the product safety assessment must concentrate on the identification of possible risks of the raw materials. Two main issues exist that are relevant to safety or risk assessment:

1) The environmental fate of detergent ingredients and the resulting concentrations in the affected environmental compartments. The final outcome of such an exposure assessment is the predicted environmental concentration (PEC).

2) The environmental effects of a chemical towards the living organisms of the compartments concerned. The final result of such an effects assessment is the predicted no-effect concentration (PNEC).

Table 47. Elements of environmental safety assessment of chemicals

	Environmental fate	Environmental effects
Substance-inherent properties (hazard identification)	biodegradability	ecotoxicity
Environmental frame conditions (healistic worst case)	exposure scenario, e.g., – usage quantities – use pattern – sewage treatment situation	extrapolation of effect data: laboratory data → real environment (extrapolation factors)
Evaluation (hisk hharacterization)	exposure analysis: predicted environmental concentration (PEC)	effects assessment: predicted no effect concentration (PNEC)
comparison: PEC vs. PNEC (No risk: PEC < PNEC)		

The comparison of these two assessment parameters, PEC versus PNEC, forms the basis of the internationally used concept for the environmental risk assessment of chemicals: a chemical is considered safe if the PEC is not higher than the PNEC. The environmental fate and effects of a chemical and the risk assessment based thereon are governed essentially by the inherent properties of the individual chemical such as biodegradation behavior and ecotoxicological properties. The connection of these intrinsic substance properties with the concrete conditions of use of the substance — e.g., the use pattern and the usage figures of the chemical, the sewage treatment situation, the application of safety factors, etc. — provides the data necessary for the calculation/measurement of PEC and PNEC (Table 47).

10.4.2. Environmental Exposure Assessment

The chemical compounds contained in detergents are typical representatives of xenobiotics intended to go down the drain after their use and thus to enter receiving river waters directly or after passing a sewage treatment plant. Depending on their physicochemical properties (for example solubility, adsorption behavior, volatility), these substances can be distributed in environmental compartments such as the water phase, sediment, sludges, soil, etc. The knowledge of the fate and the resulting environmental concentrations are therefore a prerequisite and a decisive criterion for risk evaluation.

10.4.2.1. Biodegradation

Biodegradation is the most important mechanism for the ultimate removal of organic com-

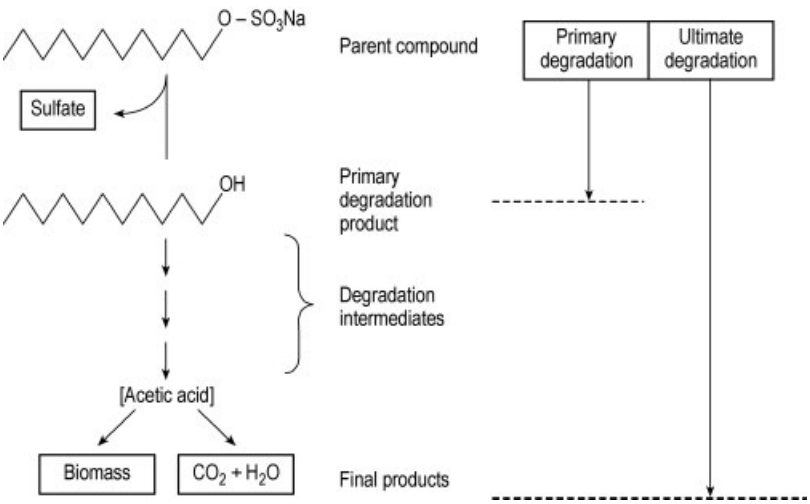

Figure 127. Terms of biodegradability evaluation exemplified by the degradation pathway of fatty alcohol sulfate

pounds from sewages, surface waters and soils. It is a stepwise process mainly effected by aerobic microorganisms, i.e., in the presence of oxygen (Fig. 127). In a first step the parent compound is transformed to a first degradation product (primary degradation), which is subsequently degraded to a second, third, etc., intermediate, i.e., to metabolites with decreasing molecular mass and structural complexity. Ultimately, the total organic structure of the starting material has been decomposed to mineralization products (carbon dioxide, water, inorganic salts) and, in parallel, partly transformed into bacterial biomass (ultimate biodegradation).

The term "primary degradation" only refers to the first steps of the degradation process. Nevertheless, the proof of a fast and quantitative removal of the parent compound is of high ecological relevance for certain chemical classes, e.g., for surfactants. As a rule, the primary degradation of surfactants significantly reduces their surface activity (including their foaming properties) and their aquatic toxicity. For that reason, the European detergent legislation requires a minimum primary biodegradability for anionic and nonionic surfactants (see Section 10.3). Nevertheless, ultimate biodegradability is the decisive criterion for the long-term fate of a chemical and its degradation intermediates in the environment, because this will guarantee the complete re-integration of all components of the substance into the natural material cycles.

10.4.2.2. Biodegradability Standard Test Methods

Primary and ultimate biodegradability of organic chemicals can be evaluated by using suitable test methods (Table 48).

The determination of the removal of the parent compound (primary degradation) requires the availability of a substance- or a substance-group-specific analytical determination method. As discussed previously (see Section 10.3), the primary biodegradation of anionic and nonionic surfactants is determined in standardized tests by measuring the removal of MBAS and BiAS, respectively. The ultimate biodegradation of chemicals can be followed in the tests by means of nonspecific analytical parameters such as carbon dioxide evolution, the BOD, or the removal of dissolved organic carbon (DOC).

Primary and ultimate biodegradability of test substances is normally evaluated by applying standardized and internationally used (OECD, ISO, EU) test procedures (Table 48). The most broadly used category of tests are the screening tests representing simple but stringent evaluation procedures. For instance, the OECD Screening Test determines the primary biodegradation of anionic and nonionic surfactants. The corresponding ultimate biodegradability screening tests are represented by the OECD tests for ready biodegradability (OECD 301 A–F) [683], e.g., the DOC die-away test (OECD 301 A), the CO_2 evolution test (OECD 301 B), or the Closed Bottle Test (OECD 301 D). If the pass level of 60 % BOD (compared to the COD) and CO_2 evolution, respectively or of 70 % COD is surpassed within the four-week test duration, it can be concluded that ultimate biodegradation of the chemical will proceed readily in the real environment [683].

It is a common feature of screening tests that a few milligrams of the test substance are diluted or suspended in a test vessel containing an aqueous mineral medium. This mixture is inoculated by adding a small amount of a bacterial suspension, in most cases originating from a municipal sewage treatment plant (30 mg/L of activated sludge or $\leq$ 10 ml/L of plant effluent). Thus, the test inoculum contains the microorganisms ubiquitously present in wastewater treatment plants and receiving river waters. The degradation process can be analytically followed by substance-specific methods as well as by summary parameters since the test substance is the only carbon source for the degrading organisms in such a screening test system.

Due to the stringency of the screening tests, a negative or poor degradation result obtained in such a system is not necessarily a proof of lack of biodegradability in the real environment. To obtain a realistic view of the biodegradation behavior under practical conditions, model test systems are used simulating the primary and ultimate biodegradation process in the biological stage of a municipal sewage treatment plant (Table 48). In contrast to screening tests, the test chemical is continuously dosed to the model plant resulting in a hydraulic retention time of 3–6 h. Another important difference to screening

Table 48. Standard test of biodegradability assessment

	Screening tests	STP-simulation tests *
Characteristics	simple discontinous flask tests test chemical is sole carbon source low bacterial inoculum stringent test conditions, no simulation test	realistic model of a municipal sewage treatment plant continuous test system/limited retention time high sludge concentration
	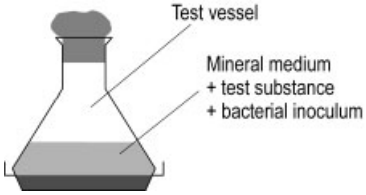	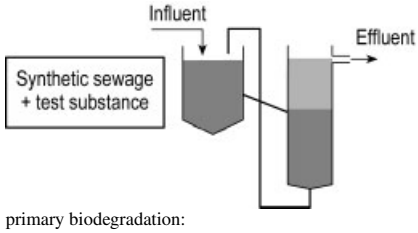
Examples	primary biodegradation: OECD Screening Test ultimate biodegradation: OECD tests for "ready biodegradability" (OECD 301 series)	primary biodegradation: OECD Confirmatory Test ultimate biodegradation: Coupled Units Test

* STP = sodium triphosphate

tests is the fact that the test material is dosed together with a large surplus of highly biodegradable substances ("synthetic wastewater") to give a competitive situation which also prevails in real domestic sewage treatment plants. The bacterial concentration in a continuous activated sludge model sewage treatment plant is relatively high (ca. 3 g/L of suspended solids). Depending on the analytical test method either the primary degradation in the simulation test (requiring substance-specific analytical methods) can be determined or the substance removal based on summary analytical parameters (DOC, COD). If a nonspecific analytical method is applied, a modified test design is necessary to obtain removal data that specifically refer to the test substance. In the *Coupled Units Test* (OECD 301 A), for example, two model plants are run in parallel [683]. One is the control, which is only fed with synthetic sewage, while the influent of the second additionally contains the test substance. The test substance-related carbon removal can be calculated from the difference of the DOC concentrations in the two plant effluents compared with the carbon content of the substance in the influent of the test unit. Monitoring data on the removal of surfactants in municipal sewage treatment plants have broadly confirmed that the elimination rates obtained in model plants are a realistic or rather a conservative description of the practical situation [684].

10.4.2.3. Supplementary Biodegradation Test Methods

Apart from these internationally used standard degradation tests additional test methods exist to extend the information about the fate of chemical substances in the environment. The results from the previously discussed screening and simulation tests are, in principle, a measure of the degradation rate, i.e., the extent of degradation within the test period. Therefore, the question remains unanswered whether the test-substance-derived material which was not biodegraded within the test period may contain recalcitrant degradation intermediates. In other words, if it is possible to show that no poorly biodegradable intermediates are formed during the degradation process, unknown negative effects from the long-term existence of those materials in the environment can also be excluded. The test for *detecting recalcitrant metabolites* (metabolite test [685]) which is experimentally based on the Coupled Units Test, represents an

instrument for answering this question. Thus, the test for recalcitrant metabolites can improve and supplement the conclusions drawn from positive results obtained in standard degradation tests.

In contrast to the Coupled Units Test, the effluents of the test and the control plants are reused each day as an influent after addition of a concentrate of the nutrients (control unit) and the test substance. Thus, a circuit is achieved which gives the degrading organisms the permanent chance to cope with the test compound and its intermediates under competitive conditions. Since the metabolite test is run over several weeks — corresponding to up to 100 cycles — even very small amounts of recalcitrant metabolites would accumulate and could be analytically detected. The proof of the absence of any recalcitrant metabolites was obtained for all important detergent surfactants by means of this method [686].

The issue of the completeness of the ultimate degradation in the environment has yet another dimension. Substances with a marked adsorption tendency such as surfactants or poorly water-soluble compounds may to a considerable extent reach environmental areas where no oxygen is present for the biodegradation processes of microorganisms. Examples are septic tanks, digesters, some river sediments, and soils. The *anaerobic biodegradability evaluation* is therefore an additional important aspect for those substances. A screening test method has been available for several years (ECETOC Test, ISO 11734 [687]) that measures the ultimate anaerobic biodegradation of a substance on the basis of the digester gas formation (carbon dioxide and methane). In addition, digester simulation tests with radiolabeled test materials have shown the anaerobic biodegradability of important surfactant groups [686].

10.4.2.4. Exposure Analysis

After identification of the environmental areas mainly exposed to the chemical(s) concerned (e.g., wastewater treatment plants, sludges, river waters and sediments, as well as soils are target areas of detergent chemicals) a preliminary exposure analysis may be conducted. The first tier of such a process is based on relatively simple but conservative standard exposure scenarios. Thus, the concentrations of detergent ingredients in raw wastewaters are calculated on the basis of the per capita usage of the product and the ingredient, respectively, and the per capita water consumption (see Table 45). The elimination of the chemical in sewage treatment plants is estimated by means of calculation models taking the physicochemical (adsorption behavior, volatility) and the biodegradability properties (deduced from degradation screening tests) of the substance into account. From the PEC of the chemical in the plant effluent it is possible to derive a PEC of the receiving water by applying a standard dilution factor of 10 for treated wastewater [682].

A more detailed and reliable environmental concentration prognosis can be obtained in the second tier of the exposure analysis when concrete data of the individual environmental situation are taken into account. These data are, for example, elimination data from sewage treatment plant simulation tests and information on the wastewater treatment (e.g., sewage volumes, hydraulic retention time, etc.) and river water situation (e.g., wastewater dilution ratio, flow rate, etc.). In addition, also the background concentration of the substance concerned must be considered, i.e., the PEC above the sewage outfall. This prognosis of the regional distribution is obtained by means of multimedia calculation models, for instance according to MacKay [688], which predict an average PEC in the environmental compartments of water, soil, and air.

The exposure analysis is — as is the whole environmental safety assessment — an iterative process. This means that the precision and reliability of the PEC values can be improved continuously by the improvement of the data and information based thereon. Therefore, it is understandable that the highest tier of the exposure analysis is represented by the measurement of real environmental concentrations, i.e., by environmental-monitoring data. If sufficiently specific and sensitive analytical methods are available and if the chemical concerned is already in use, such measurements can be conducted. Thus, environmental-monitoring data and PEC data from exposure models form a necessary mutual supplement to fill the environmental concentration data gap for most of the chemicals present in the environment. Due to the conservative character of the model cal-

culations there is consensus that the data obtained from real environmental measurements take precedence over calculated values [682].

Detergent ingredients, particularly the most broadly used surfactants (LAS, alcohol ethoxylates, alcohol ether sulfates, soap) have been thoroughly investigated by means of the multitiered approach of exposure analysis. It could be shown that the measured concentrations in raw and treated wastewaters are considerably lower than predicted by the calculation algorithm described for the tier 1 screening phase (cf. Table 49) [689, 690]. Consequently, also the predicted river water concentrations (PEC_{river}) are higher than the measured concentrations reported for several detergent surfactants [684, 691]. The conservatism of the model calculations is also connected with the fact that the biodegradation/elimination processes of chemical compounds are only considered with regard to the removal in sewage treatment plants, while an in-stream removal after entering the receiving river is not taken into account or is based on over-conservative assumptions [692]: for instance, the half-life for LAS and other readily biodegradable surfactants in rivers was determined to be in the range of 1–5 h, while the calculation model used in the EU risk assessment process assumes a value of 360 h [682].

10.4.3. Assessment of Environmental Effects

The impact of chemical substances to environmental systems can be different and may also include indirect effects, for example eutrophication of water due to a nutrient function of the substance, or remobilization of heavy metal ions. However, the most important aspect to be evaluated in the environmental risk assessment is the ecotoxicity of substances, i.e., the toxic impact to organisms living in the individual environmental compartments. Of course, it is not possible to test all these organisms in the laboratory in terms of their sensitiveness to the chemical. However, a number of internationally accepted and standardized test systems exist, which cover the different trophic levels of the aquatic food chain: starting from the algae as representatives of the plants, the plant-feeding animals, represented by the daphnia while the fish stand for higher trophic levels. Finally, the bacteria close the carbon cycle in the aquatic ecosystem due to their degradation activities (Fig. 128). In the terrestrial environment, too, the ecosystem has a similar structure, consisting of producers (plants), consumers (animals), and destruents (bacteria, fungi).

As in the case of degradation tests, there is a sensible hierarchy of ecotoxicity tests, starting with relatively simple acute toxicity tests which allow a first general evaluation; these are followed by more sensitive and more expensive chronic or subchronic tests which have a higher predictive value for the practical situation. In any case, the objective of the tests is to find out a relationship between the concentration of the test substance and the ecotoxicological effect. Therefore, the test organisms are exposed to different concentrations of the chemical under defined test conditions, and the effects are evaluated by comparison with benchmarks (without addition of the test substance).

10.4.3.1. Basic Ecotoxicity Tests (→ Chemical Products: Safety Regulations, Section 3.1.2; → Ecology and Ecotoxicology, Section 3.2)

The base set of ecotoxicological information of a chemical is obtained from acute tests on fish and daphnia and from an algal toxicity test (Table 50). Acute toxicity tests determine the lowest test concentration leading to death (fish) or immobilization (daphnia) of all tested organisms (LC_{100}/ EC_{100}). On the other hand, also the highest concentration showing no acute effect is experimentally determined (LC_0/EC_0). The acute toxicity is usually expressed by the LC_{50}/EC_{50} values, which can be calculated from the experimental LC_0/EC_0 and LC_{100}/EC_{100} data. The acute fish test is mainly conducted according to the pertinent OECD Guidelines [683] with zebra fish in a 96-h test. The acute daphnia test is also standardized [683] and has a duration of 48 h. The algal toxicity test differs from the previous tests, as the test criteria are based on the growth rate of the algal cells and the algal biomass formed within the 72- or 96-h incubation period [683]. Since this test not only covers the test organisms exposed at the beginning of the test but also the following generations of algae it is a chronic test determining the EC_0 and

Table 49. Stepwise environmental risk assessment process exemplified on a detergent-based surfactant (LAS)

Test hierarchy	Exposure analysis		Effects assessment			PEC/PNEC
	Basis	PEC	Basis	Extrapolation factor	PNEC	
Tier 1 (screening)	calculation model:		acute toxicity:			
	STP influent* : 12.3 mg/L		lowest	1000		
	Elimination in STP: 95 %		EC_{50} = 1.1 mg/L		1.1 µg/L	56
	Dilution factor: 10	62 µg/L				
Tier 2 (confirmatory)	STP simulation test:		chronic toxicity:			
	STP influent* : 12.3 mg/L		lowest	10		
	Elimination in STP: 99 %		NOEC = 0.12 mg/L		12 µg/L	1.0
	Dilution factor: 10	12 µg/L				
Tier 3 (investigative)	River monitoring:		field studies/biocenotic tests:	1–10		
	1–5 km downstream of STP release	$\leq$ 20 µg/L	NOEC = 0.25 mg/L		250/25 µg/L	0.08/0.8

* LAS-consumption: 0.9 kg/inhabitant/year
Water consumption: 200 L/inhabitant/day

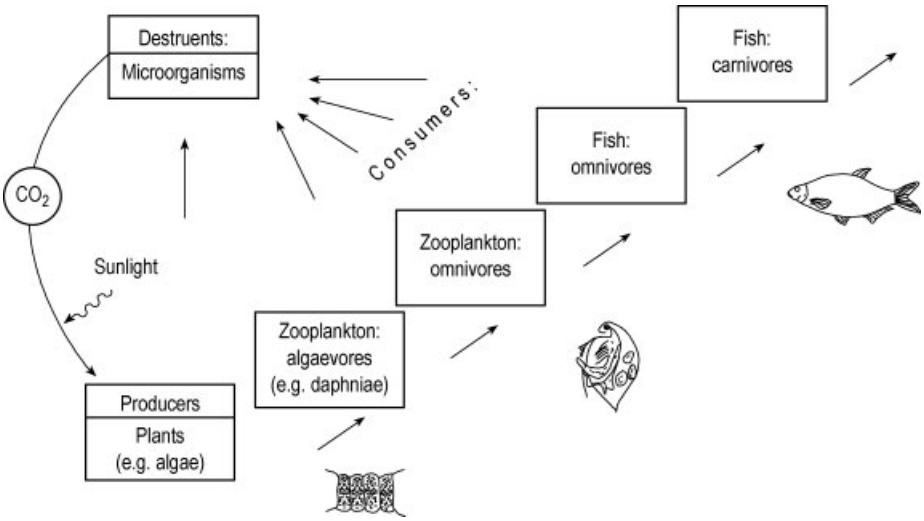

Figure 128. Scheme of the aquatic food chain

EC_{50} of algal growth (which is different from an acute algicidal effect). Despite the qualitative differences between the acute fish and daphnia tests and the chronic algae test, the LC_{50}/EC_{50} values from these three tests form the ecotoxicological criteria for the classification and labeling of substances as "dangerous for the environment" in the European Union [693]. At the same time also the environmental effects assessment for the preliminary determination of the PNEC takes reference to the lowest, i.e., most toxic LC_{50}/EC_{50} value from these tests [682]. The data from this set of three ecotoxicological characteristic values is also required when a new chemical is to be notified according to the chemical laws in EU member states [694].

Although not contained in the discussed ecotoxicological base set, the determination of the acute (e.g., bacterial respiration inhibition test [695]) or chronic (e.g., bacterial cell growth inhibition test [696]) toxicity to bacteria has a high ecological relevance for the practical situation. The chronic bacterial toxicity is measured during an 18-h incubation period using a *Pseu-*

Table 50. Ecotoxicological test methods

Test	Evaluation parameter		Relevance
Acute aquatic toxicity			
Fish	lethal effects:	LC_{50}	basic data for ecotoxicological evaluation (Safety Data Sheets)
Daphnia	ability to swim:	EC_{50}	
Algae	cell growth:	EC_{50}	EU classification "dangerous for the environment"
bacteria	respiration:	EC_0, EC_{10}, EC_{50}	initial risk assessment
Chronic/subchronic toxicity			
Fish (prolonged test/early life stage)	growth:	NOEC	
Daphnia (life cycle)	reproduction:	NOEC	aquatic risk assessment
Algae	cell growth:	EC_0	
bacteria	cell growth:	EC_0, EC_{10}	
Biocenotic toxicity			
Microcosmos test	species composition:	$NOEC_{biocoen.}$	aquatic risk assessment
Model river flow			
Terrestrial toxicity			
Earthworm (acute)	lethal effects:	LC_{50}	terrestrial risk assessment
Higher plants (chronic)	growth:	EC_0	

NOEC = No observed effect concentration
LC/EC = Lethal/Effect concentration

domonas putida shake-flask culture and determining the rate of cell multiplication. The same bacterial species or an activated sludge culture is employed in the acute test measuring the influence of the test chemical on the oxygen consumption (respiration) rate of the bacteria. The EC_0 of this test has a high predictive value for the beginning toxicity of a substance to the purification performance of a biological sewage plant [697]. Information on the basic ecotoxicological data of a chemical product as well as the environmental classification (EU classification [693], Water Hazard Class in Germany [698]) is usually provided from the pertinent safety data sheets.

10.4.3.2. Subchronic and Chronic Ecotoxicity Tests

Information on long-term ecotoxicological effects at concentration levels fairly below their acute toxicity can be gained from standardized subchronic and chronic tests with the same test organisms also used in acute tests. This type of tests aims at determining the no observed effect concentration (NOEC), i.e., the highest tested concentration that does not exhibit any effect (vs. a benchmark). The criteria for assessing effects in the subchronic fish test (test duration 3–4 weeks) are not only lethality but also changes in growth, weight, behavior, etc. In the subchronic early life stage fish test with trout (OECD 204, test duration 90 d) [683] the influence of a compound on the most sensitive life stage of the fish is tested, because the development of juvenile fish from fertilized spawn is used as criteria. The 21-d daphnia reproduction test (daphnia life cycle test, OECD 202) [683] determines the number of new daphnids produced within the test period, and thus also covers the most sensitive life processes like growth, maturation, and reproduction.

10.4.3.3. Biocenotic Ecotoxicity Tests

While the discussed test systems are single-species tests, the biota in an environmental compartment form a biocenosis with many and complex interactions. Thus, the issue must be addressed whether a single-species NOEC is suitable to allow such biocenotic implications to be assessed. Laboratory tests also exist for biocenotic effects where a self-developing biocenosis of microorganisms, algae, unicellular protozoans, and small metazoans (rotifers) represents the "test organism" [699]. The possible impact of a chemical is determined by biological analysis of the qualitative and quantitative composition of the species developing within the test period. Examples of such tests are the microcosm test and the river flow simulation test [699]. Test

results obtained in such complex systems indicate that biocenoses are not significantly more sensitive to toxic substance influences than single species used in standard chronic/subchronic tests [700].

10.4.3.4. Bioaccumulation

Bioaccumulation or bioconcentration describes the fact that a compound is enriched in the living organism to concentrations higher than that prevailing in the surrounding milieu. Experimentally, bioconcentration is determined in relatively costly tests with fish or other aquatic organisms. These organisms are exposed to a very low concentration of the test substance and the increase of the substance in the organism in the time course is analytically monitored. The bioconcentration factor (BCF) describes the ratio of the substance concentration in the organism and the surrounding medium after attaining a steady state. The BCF can also be calculated from the quotient of the substance uptake (k_1) and elimination (k_2) rate constants; the latter is determined in the depuration phase of the experiment after transfer of the exposed organisms to substance-free water. The most widely used bioaccumulation test is the flow-through bioaccumulation fish test (OECD 305) [683].

Because of the demanding analytical requirements with regard to the sensitivity and specific detection of a compound in the test organism, e.g., in a fish homogenate, in most cases the physicochemical properties of the test substance or quantitative structure–activity relationships (QSARs) are used for the bioaccumulation assessment. The descriptor mostly used for the bioaccumulation potential is the partition coefficient of a substance in a two-phase system of water and octanol (P_{ow}). The P_{ow} can be determined experimentally [683] or by calculation [701]. Substances having a BCF ≥ 100 or a log $P_{\mathrm{ow}} \geq 3$ are considered as significantly bioaccumulating and are to be classified as "dangerous for the environment" if their ecotoxicity is also high (EC/LC$_{50}$ ≤ 10 mg/L) [683]. However, it has to be borne in mind that the P_{ow} is not generally suitable for prediction of the real bioaccumulation behavior of chemicals: it is not applicable to ionic substances and surfactants [683] and does not take into account the manifold biotransformation processes in living organisms [702].

10.4.4. Process of Environmental Risk Assessment

The exposure and effects assessment implies a hierarchy of methods starting with simple standardized basic tests up to exacting simulation tests and monitoring studies. All these investigations aim at conclusively predicting the environmental concentration (PEC) and a reliable no effect concentration (PNEC). The PEC and PNEC data generation is therefore a stepwise process. Depending on the outcome of the PEC/PNEC comparison it can be decided whether or not the PEC prognoses derived from simple but stringent (screening) tests are sufficient or whether data from more realistic (simulation) tests or monitoring data is required. A similar hierarchy of the predictive value of the data is implemented in the effects assessment by the use of application (safety) factors to derive the PNEC from acute LC/EC$_{50}$ data and chronic NOEC data, respectively (see Table 49). It should be noted that the chronic toxicity test data of chemical substances already imply the increased body burden resulting from a potential bioconcentration so that no additional safety factors need to be applied for the PNEC derivation.

These principles form the basis of the European environmental risk assessment process described in Technical Guidance Documents [682]. Environmental risk assessment is an iterative process which can be stopped when the PEC/PNEC ratio is ≤ 1, which indicates the environmental safety of the chemical ("no further information/tests required"), or when all possibilities to further improve the predictive value of the data are exhausted. Risk management measures are to be taken when a PEC/PNEC ratio > 1 indicates a risk for the environment [680, 681].

According to the three-tiered process described in Section 10.4.2.4 the environmental risk of key surfactants used in detergents (LAS, alcohol ether sulfates, alcohol ethoxylates, soap) was evaluated. The LAS example (Table 49) indicates that compounds with a high environmental relevance such as detergent components often require the information even from the highest

tier to reach PEC/PNEC < 1. In the end, this risk assessment showed that the most prominent surfactants in detergents are safe for the aquatic systems in spite of their high input to the environment [689, 690].

10.5. Ecological Characterization of Main Detergent Ingredients

10.5.1. Surfactants

The surfactants are the category of detergent ingredients with the highest ecological relevance due to their high consumption figures and their potential ecotoxicological impact connected with their surface-active properties. This relevance is reflected by the fact that the two most important surfactant groups, the anionics and nonionics, are subjected to legal regulations regarding their (primary) biodegradability (see Section 10.3.2). Strictly speaking, also the second main criterion for the environmental risk assessment, the ecotoxicological behavior is included therein because the primary biodegradation of surfactants and the loss of their high aquatic toxicity are normally concomitant processes. Hence, sometimes the primacy of biodegradation is emphasized since a fast and complete (ultimate) biodegradability is the best prerequisite to exclude long-term effects in the environment.

10.5.1.1. Anionic Surfactants

The oldest and largest-volume anionic surfactant worldwide is soap. However, the superiority of synthetic anionic surfactants in terms of detergency performance considerably reduced the relevance of soap in detergents in many countries. Instead, branched-chain TPS and its successor LAS have played the most important role as detergent surfactants. The poor biodegradability of TPS became evident in many rivers and sewage plants in the late 1950s and is also recognizable from the poor results in standard primary and ultimate biodegradability tests, as shown in Table 51 (see Section 10.3.1). Unlike TPS, LAS not only fulfills the legal requirements of the European detergent legislation but is also evaluated as "readily biodegradable" in the OECD 301 tests on ultimate biodegradability (Table 51). Also in the continuous activated sludge test (OECD Confirmatory Test) a high elimination result of about 95 % MBAS removal was found, which is still lower than the > 99 % removal measured in real sewage treatment plants [684].

The success of the measures to improve the environmental compatibility of detergent surfactants can be impressively demonstrated by the results of a systematic monitoring program conducted by the Henkel company in the river Rhine in Germany since 1958 and in some of its tributaries [678, 679, 703 – 705]. Based on the MBAS analytical method, the usage of TPS in detergents is recognizable as a continuous and steep increase of the MBAS loads and concentrations (Fig. 129); the maximum concentration detected in the Rhine was 0.72 mg/L MBAS in the fall of 1964 while concentrations of even more than 1 mg/L were observed in its tributaries [703]. The replacement of TPS by LAS in 1964 marked the turning point: the MBAS values decreased continuously from then on in spite of the fact that surfactant consumption increased by 80 % from 1963 to 1975 [703].

A second reason for this positive development was the construction of biological sewage treatment plants. In highly populated areas such as Germany such technical measures are a prerequisite for a sufficiently effective reduction of the organic load present in sewage prior to entering receiving waters.

Based on the experiences from the detergent problem of the early 1960s, the development of new detergent surfactants has always been a search for improvements in terms of both their detergency performance and the reduction of the environmental impact. Besides the legally required primary biodegradability, the ultimate biodegradability behavior today is in the focus of the ecological safety assessment of surfactants used in laundry products. As shown in Table 51, there are a number of anionic surfactants which exhibit excellent ultimate biodegradability in the OECD 301 tests. This underlines that the high removal rates determined in model sewage treatment plants are mainly due to mineralization (carbon dioxide formation) and assimilation (biomass formation). Furthermore, it

Table 51. Biodegradation data of surfactants

Surfactant	Primary biodegradation			Ultimate biodegradation		
	OECD Screen. test, % removal	OECD Confirm., % removal	Ready biodegradability test, %	Coupled Units Test, % C removal	Metabolite test, % C removal	Anaerobic biodegradability
Anionics:						
TPS (C_{12})	10–50 MBAS	20–45 MBAS	301 D: < 10 BOD/COD 301 E: 37 C removal	41 ± 9 (COD)	n.d.	no
LAS (C_{10-13})	95 MBAS	93–97 MBAS	301 E: 73–84 C removal 301 D: 55–65 BOD/COD	73–82	94.9 ± 1.2	no
C_{12-14} FA + 2EO sulfate	99 MBAS	100 MBAS	301 E: 100 C removal 301 D: 100 BOD/COD	89 ± 6	102.1 ± 3.9	yes
C_{16-18} FA sulfate	99 MBAS	98–99 MBAS	301 E: 85–88 C removal 301 D: 77 BOD/COD	96	99.9 ± 1.6	yes
C_{14-18} *sec*-alkanesulfonate	96 MBAS	97–98 MBAS	301 E: 88–96 C removal 301 D: 63–95 BOD/COD	83–96	97.4 ± 0.5	no
C_{8-18} soap	n.a.	n.a.	301 D: 79–100 BOD/COD 301 B: 80–90 CO_2 evolution	n.d.	100.8 ± 1.4	yes
Nonionics:						
C_8 FA + 6 EO	99 BiAS	n.d.	301 D: 84 BOD/COD	97 ± 3	98.5 ± 2.7	yes
C_{16-18} FA + 10 EO	99 BiAS	98 BiAS	301 D: 77 BOD/COD	90 ± 16	98.8 ± 2.7	yes
C_{12-14} FA + 30 EO	99 BiAS	n.d.	301 D: 80–83 BOD/COD	94 ± 2	95.1 ± 3.1	yes
Isononylphenol + 10 EO	6–> 80 BiAS	87–97 BiAS	301 D: 5–10 BOD/COD	59 ± 22	93.6 ± 2.7	partial
C_{12-14} APG	–	99 subst.-specific	301 D: 73–88 BOD/COD 301 E: 90–93 C removal	89 ± 2	101.8 ± 2.0	yes
Cationics:						
DTDMAC	n.a.	94 DSBAS	301 D: < 5 BOD/COD	108 ± 9	n.a.	no
EQ esterquat	n.a.	93 DSBAS	301 D: 67–88 BOD/COD	89 ± 7	n.a.	yes

n.d. = not determined
n.a. = not applicable

could be shown by comprehensive investigations based on metabolic studies with radiolabeled model surfactants, or on the "test for recalcitrant metabolites", that all major anionic surfactants are biodegraded without the formation of any poorly biodegradable intermediates [706]. Surfactants exhibit a distinct adsorption behavior onto suspended solids in sewage or surface waters that cause transportation of a significant portion of these chemicals into anaerobic environmental areas (see Section 10.4.2). Thus, surfactants which have proved their ultimate biodegradability in aerobic and anaerobic environments meet a very high standard of ecological safety, as they are biodegradable under all environmental conditions. Table 51 shows that the sulfated anionic surfactants (alcohol sulfates, alcohol ethersulfates) are anaerobically degradable in contrast to the sulfonated materials (LAS,

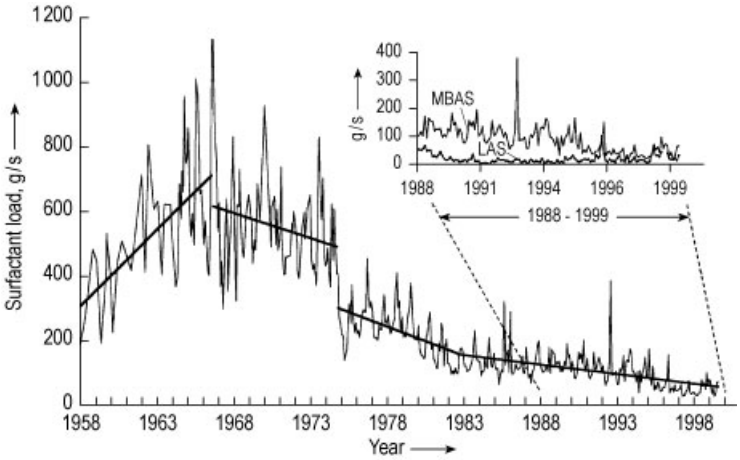

Figure 129. Changes of the anionic surfactant load (MBAS/LAS) in the river Rhine 1958–1999 (sampling site: Düsseldorf-Himmelgeist, Germany)

secondary alkane sulfonates, methyl ester sulfonates).

The second general essential criterion for the environmental risk assessment of chemicals is the effects assessment (see Section 10.4.3). The ecotoxicological behavior of the major detergent surfactants is very well investigated and documented involving acute, subacute/chronic, and biocenotic tests [689, 690]. Table 52 gives an overview of some ecotoxicological standard data of surfactants used for their evaluation according to the EU classification "dangerous for the environment" [693]. The comparison of the data reveals the marked aquatic toxicity of the surfactants with EC/LC_{50} values being predominantly in the range of 1–10 mg/L. This is a concentration range which could be achieved in surface waters if the surfactants were not biodegraded beforehand [707]. In reality, the concentrations of the most prominent detergent surfactants LAS are in the lower microgram per liter range even in small rivers receiving the effluents of biological sewage treatment plants [691, 708], while rivers receiving effluents from mechanical sewage treatment plants show LAS concentrations in the range of 500–600 µg/L [708]. It may therefore be concluded that the biological treatment of wastewaters is the decisive prerequisite for excluding an acute ecotoxicological impact of surfactants on surface waters' biota. Since an environmental safety assessment must take into account particularly long-term effects, predicted no effect concentrations derived from subacute and chronic toxicity data are more reliable than acute toxicity data. The no observed effect concentrations obtained in these tests with fish, daphnias, and algae are in the range of about 0.1–1 mg/L, i.e., one order of magnitude below the acute toxicity levels [689]; also the biocenotic NOECs determined in microcosm or mesocosm tests with anionic surfactants such as LAS and fatty alcohol sulfates [689, 700] are in the same order of magnitude. From information on the long-term ecotoxicity of key anionic surfactants a PNEC has been calculated, being 250 µg/L for LAS and 400 µg/L for alcohol ethersulfates [690]. Taking the discussed LAS concentration levels (PEC) into account it can be concluded that LAS and all other surfactants with a comparable ecological profile (readily biodegradable, ecotoxicological properties similar to LAS) will yield a ratio of PEC/PNEC $<$ 1, i.e., are safe to the aquatic environment [690].

The bioaccumulation behavior of surfactants in general was tested in several investigations that were mainly conducted using radiolabeled test materials [709]. While in some cases a bioaccumulation in fish can be definitely excluded (e.g., alcohol sulfates), in other cases only the bioconcentration of the radioactivity in the piscine body was determined without differentiation between the parent compound and the biotransformation products. A detailed investigation of the bioconcentration behavior of sev-

Table 52. Ecotoxicological data of surfactants (concentrations are given in mg/L)

Surfactant	Acute toxicity (EC/LC_{50})		Algal toxicity	Bacterial toxicity	Long-term toxicity
	fish	daphnia	(EC_{50}/EC_0)	(EC_0)	(daphnia life cycle test; NOEC)
Anionics:					
LAS	2.9	4–6	7.3/2.6	64	
C_{12-14} FA + 2 EO sulfate	7.9	79	1.8/1.0	360	0.72
C_{16-18} FA sulfate	5.2	280	30/10	35	16.5
C_{14-18} *sec*-alkanesulfonate	5	7.5	8.7 (EC_0)	390 (EC_{10})	0.6
C_{12-18} soap	46	25	10 (EC_0)	595	10
Nonionics:					
C_{12-14} FA + 9 EO	1.7	2.7	0.9/0.3	700	3.0
C_{12-14} FA + 30 EO	455	790	24/10	990	n.d.
C_{16-18} FA + 5 EO	2.7	3.9	0.3/0.1	64	n.d.
C_{16-18} FA + 14 EO	3.5	16	0.8/0.3	1000	1.0
Isononylphenol + 9/10 EO	4.6	14	2.8/1	1000	n.d.
C_{12-14} APG	3.0	7.0	12	500	1.0
Cationics:					
DTDMAC	0.6–3.0	1.1	0.06 (EC_{50})	10	0.38
EQ esterquat	3.0	78	1.8/0.3	90	2.7

n.d. = not determined

eral *n*-(*p*-sulfophenyl)alkanes, i.e., constituents of LAS, revealed that these chemical structures are not or only moderately bioaccumulative not the least due to the action of biotransformation processes [710].

10.5.1.2. Nonionic Surfactants

Alcohol Ethoxylates. Nonionic surfactants of the alcohol ethoxylate type are contained in almost all detergents and cleansers (see Section 3.1.2.1). Alcohol ethoxylates with a linear or single-branched alkyl chain are highly biodegradable. As shown in Table 51, they easily fulfill the legal requirements of a minimum primary biodegradability of 80 % BiAS removal and are "readily biodegradable" in the OECD tests on ultimate biodegradability. Alcohol ethoxylates are also bio-degradable under anaerobic conditions, thus providing all prerequisites for their fast and complete removal from the aquatic environment. The residue-free biodegradation of AE has been shown in the metabolite test [686] as well as in detailed investigations of the biodegradation pathways of these compounds [711, 712]. Consequently, also the elimination rates of alcohol ethoxylates in model plants (Table 51) and in real sewage treatment plants are very high (> 99 % [684]) and allow an environmental concentration (PEC) of a few micrograms per liter in receiving waters to be predicted. These figures are considerably lower than the BiAS concentrations determined in the river-monitoring programs of Henkel (10–100 µg/L [678, 704]). These differences are probably due to the insufficient specificity of the BiAS analytical method [713].

The ecotoxicological profile of AE has also been very well investigated. Their acute aquatic toxicity data are in the same order of magnitude as those of the anionics, although their algal toxicity seems to be more pronounced frequently (Table 52). The comprehensive set of available acute, chronic, subchronic, and biocenotic ecotoxicity data on alcohol ethoxylates was used in a recent environmental risk assessment study to derive a PNEC for alcohol ethoxylates [689]. The PEC/PNEC comparison led to the conclusion that this most prominent class of nonionic surfactants is safe to the aquatic environment [690]. Further support for this evaluation comes from comprehensive bioaccumulation studies showing that these surfactants are metabolized intensively in the piscine body [714].

Alkylphenol Ethoxylates. Another relevant group of nonionic surfactants still being used in some countries are the alkylphenol ethoxylates (see Section 3.1.2.2). In contrast to alcohol ethoxylates they show some ecological properties which have made their environmental compatibility questionable. Nonyl- and octylphenol ethoxylates have a highly branched alkyl chain which is an obstacle for microbial degradation. Thus, these surfactants are not readily biodegradable although they fulfill the legal requirement of > 80 % BiAS removal (Table 51). In winter conditions, however, even this primary biodegradation is relatively slow, so that foam problems in sewage treatment plants have been observed [715]. The major ecological reservations came from the fact that in the primary degradation process poorly biodegradable intermediates are formed which are even more toxic than the parent surfactant [716]. These intermediates (alkylphenol + 1–3 EO units) are poorly water soluble and are adsorbed strongly on sewage sludges. During the anaerobic treatment of these sludges in digesters alkylphenol is formed by splitting off the residual EO units. This metabolite shows an even increased aquatic toxicity [716, 717]. These findings led to a ban (e.g., in Switzerland in 1986) or voluntary renunciations (e.g., in Germany in 1987) of alkylphenol ethoxylates in laundry products. Since 1992 a PARCOM recommendation has existed to phase out alkylphenol ethoxylates in household laundry products by 1995 and in I& I cleaning products by the year 2000.

Alkylpolyglycosides. Developments towards more efficient and better environmentally compatible products are steadily in progress, also in the field of the nonionic surfactants. One interesting and promising example is the class of alkylpolyglycosides (see Section 3.1.2.6) These surfactants feature ready and complete ultimate biodegradability in aerobic and anaerobic environments (Table 51) [718]. Also, alkylpolyglycosides exhibit very low acute, subacute, and chronic toxicity data within the effect concentration range typical for surfactants (Table 52). A very conservative environmental risk assessment conducted on the basis of these data showed that this surfactant class is safe for the aquatic and terrestrial environment [718].

10.5.1.3. Cationic Surfactants

The cationics are the smallest of the three main surfactant groups (Table 45). They are not contained in usual detergents, but represent the active agent in fabric softeners (see Section 3.1.3). Until the beginning of the 1990s, ditallow dimethyl ammonium chloride (DTDMAC) was the most important cationic softener, with a market share of about 80 % [149] (see Section 3.1.3). This surfactant is almost quantitatively eliminated from sewage, as illustrated by the high elimination rates in model sewage treatment plants (columns 3 and 5 in Table 51). This is mainly due to adsorption onto sewage sludge, whereas its biodegradation is slow. As yet, cationic surfactants are not subjected to legal biodegradability requirements, although a group-specific analytical method exists to follow the primary biodegradation by measuring the decrease of the disulfine blue-active substance (DSBAS) [719].

Generally, degradation of cationics in usual screening tests is poor, maybe due to the fact that their bactericidal properties may prevent bacterial growth at the test concentrations required from the analytical point of view. However, even application of very low concentrations (< 1 mg/L) of ^{14}C-labeled DTDMAC did not result in substantial $^{14}CO_2$ formation under screening test conditions, and this indicates a poor biodegradability of DTDMAC even at concentrations that do not adversely affect bacterial growth. Results from activated sludge studies tend to be ambiguous due to the difficulty in distinguishing between elimination by adsorptive processes and real biodegradation. Analytical determinations of the residual amount of the parent DTDMAC present in the activated sludge of an OECD Confirmatory Test indicated that a major part of the elimination was due to (primary) biodegradation [720]. The strong adsorptivity of DTDMAC, its very slow aerobic biodegradability and its nonbiodegradability under anaerobic conditions resulted in relatively high concentrations of this surfactant in digested sludges [721]. These properties and findings in combination with the relatively high ecotoxicity of DTDMAC in standard tests (Table 52) triggered the development of alternative cationic

surfactants exhibiting more favorable ecological characteristics.

Today, the fabric softener market in industrial countries has virtually completely switched to new cationic surfactants. The major active ingredients in softeners are now esterquats (e.g., EQ, DEQ, DEEDMAC, see page 50). Esterquats are readily biodegradable under aerobic and anaerobic conditions which favors a high elimination rate in sewage treatment plants. This is attributed to genuine biodegradation (Table 51). Although the test for detecting recalcitrant metabolites is not applicable to strongly adsorbing compounds such as cationics, in-depth investigations of the degradation behavior of triethanolamine-based [149] and other esterquats [150, 722] have clearly indicated that these molecules are completely and ultimately biodegradable in the environment. The comparison of the ecotoxicological data of DTDMAC and the EQ esterquat (Table 52) points out the favorable ecological properties of the latter group of cationic surfactants. Although DTDMAC has not proved to be a risk for the aquatic environment according to the PEC/PNEC concept [723], it is even more evident that the use of esterquat cationics in fabric softeners does not have any negative impact to aquatic life.

10.5.2. Builders

As discussed in Sections 9.1.2 and 10.3.4, the complexing agent sodium triphosphate had been the main builder in almost all detergents in Europe, USA, Japan and other countries for a long time. Today, phosphate-free detergents are mainly based on builder systems consisting of the inorganic builders zeolite A and soda ash, and organic cobuilders such as polycarboxylates and/or citrate [166].

10.5.2.1. Zeolites

The major breakthrough with respect to phosphate-reduced and nonphosphate detergents was the development and introduction of inorganic water-insoluble sodium aluminum silicates of the zeolite A type. The ecological behavior of this ion exchanger had been investigated in a comprehensive research program that was started in 1973. It covered laboratory tests and field trials conducted by experts from industry, water authorities, and academia [724]. It was shown that no excessive sedimentation of the insoluble particulate matter occurred in the household sewer systems and communal sewage systems. A major part of the load of zeolite A (about 60 %) is already eliminated in the mechanical stage of sewage treatment plants (sand trap and primary settler) while total elimination is around 95 % after passing the biological stage and the secondary settler. Although the mass of the sewage sludge is increased due to zeolite, no negative effects on the purification performance of the activated sludge and on the anaerobic amenability of the sludge in the digester was found; no decreased retention of heavy metals in sewage plants has been detected in the presence of zeolite either. Investigations in the 1990s showed that zeolite A is quickly transformed into an extremely insoluble amorphous complex of basic calcium aluminum silicate phosphate under usual sewage treatment plant conditions [725].

Although only about 5 % of the zeolite A present in raw sewage will pass the sewage treatment plant, it was necessary to investigate all possible effects on the biota in receiving waters. Acute and long-term toxicity to fish and daphnia proved to be very low ($EC_0 > 250$ mg/L), while a growth-inhibiting effect on algae could be observed at concentrations > 10 mg/L in nutrient-poor culture media. This effect is, however, without any practical relevance because it is due to ion-exchange by zeolite A and the resulting depletion of essential trace elements in the test vessel [726]. In practice, zeolite A will never reach oligotrophic water bodies alone but only in conjunction with other nutrients in treated or untreated effluents. Long-term ecotoxicity data from fish (early life stage test) and daphnia (21-d reproduction test) revealed NOEC values of 87 mg/L and 37 mg/L, respectively [727].

Since no suitable analytical methods are available to determine zeolite concentrations in river water, calculations of the PEC were made by taking into account the consumption figures, elimination rates in sewage plants, and a river dilution factor of 10. Such a rough estimation leads to river water concentrations of 150–300 µg/L decreasing to 57 µg/L if a half-life of 1–2 months [728] is taken into account [725]. These

figures prove the environmental safety of zeolite A since the PEC is far below the PNEC.

10.5.2.2. Polycarboxylates

Phosphate-free detergents require cobuilders in addition to zeolite–soda ash to achieve comparable performance to STP [166] (see Section 3.2.4). By volume, the most important organic cobuilders are polycarboxylates, i.e., high-molecular mass ($M_r \leq 100\,000$) homopolymers of acrylic acid or copolymers of acrylic acid and maleic acid (Table 45). These water-soluble polyelectrolytes form sparingly soluble coagulates with excess calcium ions and adsorb on solid surfaces [729]. Based on these properties, polycarboxylates are eliminated to a high extent (> 90 %) in mechanical-biological sewage treatment plants [729, 730]. The contribution of biodegradation to this effective elimination process is minute. Polycarboxylates are poorly biodegradable except for the low molecular constituents ($M_r \leq 1000$) as was shown by investigations using ^{14}C-radiolabeled model compounds. On the other hand, polycarboxylates in real-life concentrations have proved to be without influence on the operation of sewage plants. They exhibited no adverse effects on the biological activity of sludge bacteria, the settling behavior of the sewage sludges, and heavy metal retention in laboratory tests and in a field trial [729]. Large amounts of polycarboxylates in sewage sludges do not influence the degradation process and the polycarboxylates are not redissolved in sludge digesters [729, 730].

Similarly comprehensive studies regarding fate and effects of polycarboxylates were conducted for the environmental compartments of surface waters and soil. These studies have shown that neither significant biodegradation nor heavy metal remobilization from the sediments is to be expected at relevant river concentrations (< 0.1 mg/L) [729, 730]. Due to the low ecotoxicity of polycarboxylates concluded from results of standard acute and long-term tests (acute EC_{50} > 200 mg/L, fish embryo larvae test NOEC = 40 mg/L, algal growth NOEC = 32 to > 200 mg/L, daphnia NOEC = 1.3–350 mg/L) a negative influence on the biota in surface waters can be excluded. The low no-effect concentrations within the broad range of the NOEC values observed in the chronic algae and daphnia test are connected with the precipitation of the polymer in the test system, i.e., with physical effects [730]. Bioaccumulation of polycarboxylates can be excluded because of their water solubility ($\log P_{ow} < -4.2$) and their high molecular mass.

Polycarboxylates will enter the terrestrial compartment when sewage sludges are used for agricultural purposes. The high-molecular polymers are retained in the top layer of the soil, as shown by lysimeter tests; the more mobile low-molecular fractions will hardly be present in real sludges because of their biodegradability and their less pronounced precipitation behavior [729, 730]. Thus, contamination of ground water by polycarboxylates can be ruled out. As expected, the ecotoxicity of these materials to terrestrial organisms is very low: the NOEC values in the plant growth test (> 200 mg/kg) and in toxicity tests with earthworms (1600 mg/kg) are orders of magnitude above the PEC_{soil} (< 0.1 mg/kg) [729, 730].

In spite of their proved environmental compatibility, the presently used polycarboxylates are still regarded as detergent components which are to be substituted as soon as efficient and biodegradable alternatives are available.

10.5.2.3. Citrates

Citrates are widely used in phosphate-free detergents and cleansers, especially in liquid formulations (Table 45). Since citric acid is a natural constituent and a common metabolite of most living organisms, it can be expected that its environmental behavior is very favorable. In fact, this compound is readily biodegraded in standard ultimate biodegradation screening tests and in continuous activated sludge tests [731]. Anaerobic biodegradation has also been proven [732], although only a small percentage of this water-soluble detergent ingredient will reach the anaerobic areas of the aqueous environment. Taking the citrate consumption in detergents and its excellent biodegradability into account, an influence of this weakly complexing agent on heavy metal remobilization in sludges and sediments can be excluded. The same positive conclusion is possible in regard to the ecotoxicological effects: citrate has low acute aquatic toxicity

values of $LC_{50} \geq 500$ mg/L (fish), $EC_{50} \geq 100$ mg/L (daphnia) and $EC_{50} \geq 1000$ mg/L (algae) [731, 732].

10.5.2.4. Sodium Carbonate (Soda Ash)

The inorganic compound soda ash plays an important role as a water-soluble builder and cobuilder in phosphate-containing and nonphosphate detergents. Although detergent-based consumption figures of soda ash are high (Table 45), no environmental impact is to be expected, because the ecotoxicological effects observed at high concentrations (daphnia $EC_{50} \geq 260$ mg/L, fish $LC_{50} \geq 200$ mg/L) are due to the resulting increase in the pH value of the test medium [731], which is irrelevant under real-life conditions.

10.5.2.5. Nitrilotriacetate (NTA)

In principle, NTA is an efficient builder and cobuilder in phosphate-free detergents. Its environmental compatibility was investigated in comprehensive studies [733]. Nitrilotriacetate was found to be ultimately biodegradable in aerobic conditions, although the degradation rate depends on temperature, concentration, and other ambient conditions. Heavy-metal complexes of NTA are degraded more slowly. The ecotoxicological data of NTA characterize this compound as weakly toxic to higher aquatic organisms (fish $LC_{50} \geq 100$ mg/L, daphnia $EC_{50} \geq 80$ mg/L). The NOEC of algal growth is in the range of 10 mg/L. In spite of these ecologically favorable characteristics concern arose from the strong complexing properties of NTA and, hence, the potential of heavy-metal remobilization. Due to uncertainties in regard to biodegradation rates of NTA under real-life conditions and its heavy-metal remobilization potential, it was recommended in Germany in the 1980s to use NTA only to a limited extent and to conduct further studies. The outcome of this successive program (1987–1990) is that NTA is readily biodegradable and is eliminated in sewage treatment plants to a high extent (95 %) irrespective of seasonal influences [734]. Heavy-metal removal from sewage is not impaired nor is a remobilization of heavy metals from river sediments to be anticipated below a concentration of 50 µg/L of NTA. Thus, NTA today is considerably more positively evaluated than in the past. Nevertheless, NTA is only used in minute amounts in detergents (Table 45), and it seems doubtful whether this compound will play a prominent role in the future as a detergent (co)builder.

10.5.3. Bleaching Agents (see Section 3.3)

10.5.3.1. Sodium Perborate

Sodium perborate decomposes during the washing process forming hydrogen peroxide and sodium metaborate. The water-soluble metaborate is not eliminated in wastewater treatment plants and may cause a considerable increase in the boron concentration in rivers. Although large regional variations exist due to geological conditions and degree of river pollution, the ratio of anthropogenic to naturally occuring boron in more closely investigated rivers in Germany was found to be up to 50 % in the river Rhine [703], about 70 % in the Neckar river [735], and considerably more than 100 % in the river Ruhr [736]. Nevertheless, a broad systematic study on the boron concentrations in German surface waters and drinking waters conducted in the mid–1970s [737] showed that boron concentrations in most rivers are below 0.25 mg/L and less than 0.1 mg/L in drinking waters. In small and heavily polluted rivers of Germany and other European countries higher concentrations up to 1 mg/L were found [737]. In the 1990s investigations on the boron content of drinking water in Germany showed that the concentrations had not changed significantly as compared with 1974 values [738]. Accordingly, more recent data from the river Rhine and its tributaries [678] have verified that the boron concentrations have not changed (rivers Rhine and Ruhr) or have decreased considerably (rivers Main and Neckar).

The ecotoxicological behavior of borate in water and soil has been studied comprehensively [739, 740]. The ecotoxicity of borate is relatively low. This is true for fish and invertebrates although the effective concentrations vary in a very broad range, depending on the individual studies. Fish LC_{50} values between 11 and 3400 mg/L are reported. Long-term toxicity data do

not indicate a pronounced ecotoxic action of borate either. NOEC values are clearly above a concentration of 1 mg/L [740]. However, literature data exists that indicate chronic effects towards certain aquatic organisms, e.g., fish, algae, or reed plants in the submicrogram per liter concentration range [731, 739, 740]. A critical reevaluation of these previously collected data in the light of more recent investigations showed, however, that the environmentally relevant no effect concentration in fresh waters is not lower than 1 mg/L [740]. Taking the reported river water concentrations of boron into account it is evident that boron does not constitute a problem for surface waters. However, the margin between the ecotoxicological no effect concentration and the environmental concentrations in the majority of rivers is relatively small.

With respect to the terrestrial toxicity of boron, a specific phytotoxicity towards certain agricultural plants exists, such as fruit trees, tomatoes, and vineyard stock. Therefore, it is not advisable to recycle boron-containing wastewater for irrigation purposes [739].

10.5.3.2. Sodium Percarbonate

Sodium percarbonate decomposes during the washing process forming hydrogen peroxide and sodium carbonate. Sodium carbonate is a water soluble mineral and ubiquitously present in the aqueous environment, because it is in equilibrium with carbon dioxide from the atmosphere and because of the presence of other dissolved mineral carbonates. Sodium carbonate has a very low aquatic toxicity (fish LC_{50} = 200–740 mg/L, daphnia EC_{50} = 265–565 mg/L) [731] which is attributable to the increased pH value at these high test concentrations. Thus, the use of sodium percarbonate as bleaching agent in detergents will not have any significant impact on the aquatic environment.

10.5.3.3. Tetraacetylethylenediamine (TAED)

Tetraacetylethylenediamine is the most widespread bleach activator used in laundry products (Table 45). In the washing process it is almost quantitatively perhydrolyzed to diacetylethylenediamine (DAED) and peracetate, the latter ultimately forming acetate. TAED and DAED dissolve well in aqueous alkaline solutions and have a low log P_{ow} [741]. The ready ultimate biodegradability of the two compounds has been shown in several tests [731, 741]. Accordingly, the elimination in sewage treatment plants is very high ($\geq$ 95 %) [741].

The ecotoxicity of TAED and DAED is remarkably low, since no effects were observed in the majority of tests, even at the highest concentration tested. For instance, the fish LC_{50} is well above 1000 mg/L and significant effects on daphnia and algae could not be observed at 800 and 500 mg/L, respectively [741]. Based on these biodegradation and ecotoxicological data it can be concluded that no adverse ecological effects will result from the use of TAED in laundry products.

10.5.4. Auxiliary Agents

Auxiliary agents are present in small amounts in many laundry product formulations. Nevertheless, their widespread use in these products results in considerable consumption figures (see Table 45) and, consequently, in a significant entry into the environment.

10.5.4.1. Phosphonates

Phosphonate compounds with more than one phosphonate group are suitable bleach stabilizers (see Section 3.3.4), because they are effective complexing agents and have additional chemical properties that support the washing process. Phosphonates have been well investigated with regard to their environmental behavior. They are poorly biodegradable under standard test conditions, although evidence exists that the chemically very stable C–P bond in phosphonates can be microbially attacked [742]. Nevertheless, phosphonates are not persistent in the environment, because they are subject to abiotic degradation in waters and in soil [742 – 744]. The degradation rate is strongly affected by the chemical composition of waters, e.g., the presence of metal ions like Fe^{3+}, and by sunlight. The degradation of HEDP is mainly based on photochemical processes [742,

743] while degradation of nitrogen-containing phosphonates seems to be a hydrolytic process that produces decomposition products which are, at least in part, accessible to biodegradation [742, 743]. Phosphonates are partially (about 50 %) eliminated in sewage treatment plants as shown in model experiments and field trials [742, 745]. In sewage plants having a phosphate precipitation stage the elimination of phosphonates is high (> 90 %) [745].

Phosphonates are strongly complexing agents which have the potential to affect the distribution and partitioning of metal ions in sludges, sediments, and soils. Laboratory and field studies showed no mobilization of heavy-metal ions in sewage sludges at phosphonate concentrations in the 2–10 mg/L range [742, 745]. Similar investigations on heavy metal release from river sediments made clear that the strong adsorption of phosphonates to sludges and sediments will prevent that metal ion remobilization occurs at environmentally relevant phosphonate concentrations of < 1 mg/L in wastewater and < 10 µg/L in surface waters [742].

The aquatic toxicity of phosphonates is low. The acute fish and daphnia LC_{50} values are > 100 mg/L [742]. Also, long-term NOEC data characterizes the phosphonates as unproblematic: subchronic/chronic studies on fish and daphnias revealed NOECs > 10 mg/L, i.e., orders of magnitude above the PEC levels [742]. Similarly to other complexing agents or ion exchangers, phosphonates, too, pretend to be relatively toxic to algae in standard tests. In fact, the algal NOEC was found to be in the range of 0.1–10 mg/L. However, there is clear evidence that this inhibition is due to the depletion of essential micronutrients for algal growth [742]. Ultimately, bioconcentration studies on two phosphonates (HEDP, ATMP) in fish [743, 744] proved that these chemicals do not bioaccumulate significantly.

10.5.4.2. EDTA

Ethylenediaminetetraacetate played an important role as a bleach stabilizer in the past. However, issues related to the environmental properties of this strongly chelating agent have led to a considerable decrease of its use in detergents. For instance, in Germany and some other European countries the use of EDTA in detergents ceased more than a decade ago. EDTA proved to be poorly biodegradable under environmental conditions although there is evidence that microbial degradation mechanisms exist [746]. In standard tests for ready and inherent biodegradability, this complexing agent indicated insignificant degradation rates and was not eliminated in a model sewage treatment plant (4 ± 2 % DOC removal in the Coupled Units Test) [731]. The insignificant elimination of EDTA in two-stage sewage treatment plants was confirmed in monitoring studies [747]. Thus, it is obvious that EDTA present in raw wastewater will enter receiving waters almost quantitatively. Accordingly, EDTA was found in rivers of Germany, the UK, Switzerland, and Austria in significant concentrations up to 50 µg/L [746]. At the same time similar concentrations of EDTA were detected in raw drinking water at sites where it is prepared from river water. This suggested that the processes normally employed for drinking water preparation do not eliminate low EDTA levels [747]. The influence of EDTA present in surface waters on the drinking water quality, its possible effects on remobilization of heavy-metal ions from sludges and sediments, and its deficiency of being adequately eliminated raised concerns about its environmental safety. This pushed activities towards replacing EDTA by phosphonates and NTA.

From an ecotoxicological point of view no critical aspects can be deduced. EDTA is not bioaccumulating and has a very low acute toxicity to fish and daphniae. LC/EC_{50} values are fairly above 100 mg/L [731, 746]. As expected, the algal growth NOEC in standard tests is considerably lower, i.e., in the range of 10 mg/L [731] which may be explained by the already discussed deprivation of trace metals in the test medium. On the other hand, stimulation effects of EDTA on algal growth are reported which could be explained by the photodegradation of ferric complexes to form less stable ferrous complexes with an enhanced bioavailability of growth-stimulating iron [746].

10.5.4.3. Enzymes (see Section 3.4.1)

Most detergents today contain enzymes, i.e., proteases, amylases, lipases, and cellulases. Being proteins, the ultimate biodegradation of enzymes in standard tests for ready biodegradability is high [731]. Investigations on the removal of protease in a model sewage treatment plant indeed showed a loss of proteolytic activity of more than 99 %. Of course, enzymes are also accessible to anaerobic degradation. The acute aquatic toxicity of enzymes is relatively low (fish LC_{50} = 25–350 mg/L, daphnia EC_{50} = 160 mg/L) [731]. Algal toxicity values are in the range of EC_{50} = 10–100 mg/L. The ecotoxicological effect concentrations of enzymes are better referred to enzyme activity rather than to weight.

10.5.4.4. Optical Brighteners (see Section 3.4.5)

Fluorescent whitening agents (FWA) in laundry products are mainly anionic diaminostilbene (DAS) or distyrylbiphenyl (DSBP) derivatives. They are hydrophilic water-soluble compounds having a high affinity to textile fibers and to particulate matter such as sewage sludges. A considerable portion of optical brighteners present in wastewaters is eliminated in sewage treatment plants (about 90 % in case of the most important stilbene derivative DAS 1, and about 50 % in case of DSBP, which is the major distyrylbiphenyl compound) [748]. This elimination is only due to adsorption but not to biodegradation since FWA are not readily biodegradable and seem only to undergo slow primary biological degradation [749]. In anaerobic digesters they are not biodegraded either, so that concentrations up to 1 g per kilogram of dry matter were found [749]. The portion of FWA not eliminated in wastewater treatment plants enters receiving waters and is exposed to daylight. The photodegradation of DAS 1 and DSBP was investigated in detail by experimental work and computer simulations [748]. As the kinetics of direct photodegradation is not only a function of the chemical properties of the FWA but also of the ambient environmental conditions (spectrum and intensity of daylight) the half life of DAS 1 varies from a few days to six months, whereas DSBP has a lifetime of a few hours. The photodegradation is essentially a primary degradation that forms a large number of products from DAS 1, while the degradation of DSBP follows a relatively simple path [748]. The triazinyl derivatives formed from DAS 1 during photodegradation are poorly biodegradable but weakly toxic in the acute fish test ($LC_0 > 97$ mg/L). The DSBP degradation products are predominantly accessible to biodegradation as shown in inherent biodegradability tests [748]. Recent monitoring data of the two major FWA species obtained in Germany and Switzerland revealed results that were well comparable with individual concentrations in the microgram per liter range in sewage, in the milligram per kilogram range in sludges and sediments, and in the nanogram per liter range in river waters [748].

The aquatic toxicity of FWA is low. Tests indicate fish LC_{50} values of > 320 mg/L (DAS 1) and > 70 mg/L (DSBP), respectively. The daphnia EC_{50} value for individual FWA is > 1000 mg/L. The algal toxicity is uncritical ($EC_{50} >$ 80 and > 10 mg/L, respectively) as well. The low ecotoxicity of FWA, which, in addition, do not bioaccumulate [749], is confirmed by the NOEC values determined for DSBP in the prolonged fish test (> 1 mg/L) and the three-week daphnia reproduction test (> 7.5 mg/L). The comparison of the PNEC deduced from these experimental results with the measured environmental concentrations does not indicate an environmental risk.

10.5.4.5. Carboxymethyl Cellulose

Carboxymethyl cellulose is a classical soil antiredeposition agent contained in some detergents. It is a water-soluble polymer which is only very slowly biodegraded [731, 750]. The resistance to biodegradation increases with the degree of substitution of free hydroxyl groups by carboxymethyl groups [750]. Thus, it is understandable that detergent-based CMC exhibits only poor results in ready biodegradability tests. Even in inherent biodegradability tests like the Zahn–Wellens test only a 40 % elimination value was found [731]. Also in the Coupled Units model sewage treatment plant CMC showed only a partial elimination with a 25 % DOC removal [731]. On the other hand, the polymer ex-

hibits the expected very low aquatic toxicity with EC_0 values > 1000 mg/L in the acute fish and daphnia test [731]. Based on these data CMC is not considered a detergent chemical impacting the environmental biota.

10.5.4.6. Dye Transfer Inhibitors

Dye transfer inhibitors are polymeric compounds that prevent the transfer of dyes from one laundry item to an other via the wash bath. A prominent representative of these compounds is poly(*N*-vinylpyrollidone) (PVP). Its ecological behavior is similar to that of CMC, i.e., PVP is poorly biodegradable in standard biodegradation tests and eliminated in sewage treatment plants only to a limited extent. The very low ecotoxicity of the polymer (chronic toxicity NOEC > 100 mg/L) on the other hand assures that no detrimental effects from PVP will result in the aquatic environment.

10.5.4.7. Fragrances

Fragrances are mostly present at concentrations < 1 % in almost all laundry products. Usually, they are very complex mixtures of numerous individual perfume components that differ in their chemical structure and therefore in their ecological behavior. Due to their volatility and adsorbing properties they have been considered hardly relevant to the aquatic environment. However, more recent studies have revealed that certain perfume compounds, e.g., musk xylene, can be detected in fish and river water, obviously due to a bioaccumulation process [751]. This nitro musk compound has a high log P_{ow} and is poorly biodegradable. Since this compound has also been detected in human adipose tissues and mother's milk [752], the detergent and cosmetic industry in some European countries (Germany, The Netherlands) has discontinued using this component in its products. This case exemplifies the necessity of a sufficient ecological knowledge on any detergent ingredient irrespective of whether it is contained in the product in high or very low concentrations. Presently, extensive investigation programs are under way to found a sound basis for the ecological evaluation of perfume ingredients used in detergents and cosmetics.

10.5.4.8. Foam Regulators (see Section 3.4.3)

Two key groups of chemicals used as antifoam agents are silicone fluids (polydimethylsiloxanes, PDMS) and paraffins (Table 45). Paraffins are accessible to ultimate biodegradation but the rate is strongly dependent on the chain length and the molecular mass, respectively. Polydimethylsiloxanes are not biodegradable but virtually quantitatively eliminated in sewage treatment plants due to their strong adsorptivity on sludges. PDMS that enter the terrestrial environment when sewage sludges are used for agricultural purposes are not persistent as they are abiotically decomposed in the presence of clay soils. Volatile cyclic products and linear water soluble siloxanols are formed which may undergo photolytic oxidative degradation [753]. PDMS and paraffins exhibit a low aquatic toxicity and have no bioaccumulation potential due to their polymeric properties. They will not have adverse effects on the environment.

10.5.4.9. Soil Repellents

The most important soil repellents in detergents are anionically modified polymers from terephthalic acid and poly(ethylene glycol). Due to their polymeric nature they are not readily biodegradable but, nevertheless, accessible to ultimate biodegradation under aerobic and anaerobic conditions. Based on results obtained in a static activated sludge biodegradation test system (OECD 202 B), the extent of elimination under sewage treatment plant conditions is expected to be very high. The ecotoxicity of these polymers is very low. They exhibit no acute toxic effects to fish, daphnids, and algae at the highest concentration tested (100 mg/L) which is clearly above their water solubility. Taking this data and the relatively low consumption figures (Table 45) into account it can be concluded that soil repellents pose no risk to the aquatic environment.

10.5.4.10. Dyes

Dyes are sometimes used in powder and liquid detergent products. The total usage of these complex compounds in detergents is very low (Table 45). Available ecological information on these materials from the producers' safety data sheets characterize them as poorly biodegradable and having low acute aquatic toxicity.

10.5.4.11. Sodium Sulfate

Sodium sulfate is used mainly as a filler in conventional powder detergents. The German consumption figure of this inorganic salt in detergents has decreased considerably within the last few years (1997: 100 100; 1999: 74 600 t/a [754]), mainly due to the growing market share of compact detergents. Sodium and sulfate ions are natural constituents of surface waters. Based on recent consumption figures (Table 45) and sulfate concentrations in several German rivers (30–160 mg/L), it can be calculated that the contribution of detergents to the sulfate load in rivers is less than 1 %. The aquatic toxicity of sodium sulfate is very low. The LC_{50} is in the range of $>$ 1 g/L [731]. Hence, regarding the mass balance and the ecotoxicity of this compound it is obvious that sodium sulfate from detergents does not represent a significant factor influencing the aquatic environment.

11. Toxicology

Detergents are products designed for everyday use; thus, they are found in large quantity in nearly every household. The home is quite unlike the industrial workplace: one cannot assume that every user of a household product plays strict attention to warning labels or recommended safety precautions. For this reason, any product offered on the market must insofar as possible be formulated so that it represents no significant or foreseeable health hazard, regardless of whether or not the product is properly handled. This in turn means that even at the earliest stages of raw material selection, great value should be attached to chemicals known to be harmless to humans following any realistic type of exposure.

The following types of exposure must be taken into account in assessing the risk presented by detergents and their components:

Skin contact (wash bath, residues remaining on laundered items, or exposure during manufacture)
Ingestion (accidents, especially with children, and trace residues in drinking water)
Inhalation (during production or use of powder detergents)

Extensive toxicological studies are required, and the results must be carefully weighed, taking into account the amount and stability of individual substances. Such studies should examine local effects (skin irritation, development of allergic reactions on contact, skin penetration), systemic effects (both acute and chronic), and potential hazards of a more subtle nature (mutagenicity, embryotoxicity, and carcinogenicity).

This chapter begins by examining the toxicology of the main detergent constituents, after which the properties of complete detergent formulations are explored. The approach is analogous to that followed in practice by a manufacturer in the course of product development.

11.1. Detergent Ingredients

11.1.1. Surfactants

Most of the biological properties of surfactants can be understood in terms of interactions occurring between surfactant molecules and such fundamental biological structures as membranes, proteins, and enzymes. Contact between a surfactant and a membrane leads to changes in membrane permeability, and in extreme cases, can even result in membrane solubilization. The most obvious potential consequence is interference with material transport and, ultimately, cell damage. A further danger is the fact that surfactants can significantly affect the resorption of other chemicals.

Proteins form adsorption complexes with both anionic and cationic surfactants. Complex formation is often a consequence of polar interactions between the hydrophilic residue of a surfactant and charged sites on the protein molecule, but hydrophobic interactions can also play a role. Such complex formation results in

protein denaturation, which in the case of an enzyme implies a reduction or even total loss of catalytic activity, corresponding to a change in metabolic function [755].

Nonionic surfactants are characterized by their lack of strongly polar functional groups. Thus, compounds of this class rarely cause protein denaturation. They can induce a limited amount of protein solubilization, but the concentrations required for the appearance of harmful effects are usually greater than those of ionic surfactants.

Transport through skin is an important consideration with respect to detergent ingredients, which tends to be quite low for anionic and cationic surfactants [756, 757]. Transport is greater for nonionic surfactants, but again the amount of material capable of entering an organism by this route is so low that it can be essentially disregarded as a potential hazard [758]. By contrast, both anionic and nonionic surfactants can be readily resorbed through the gastrointestinal tract following their ingestion, whereas the intestinal resorption of cationic surfactants is low [759]. Even absorbed surfactants are relatively harmless because they are rapidly metabolized, mainly by β- or ω-oxidation of alkyl chains and some cleavage of poly(alkylene glycol) ether linkages. Elimination occurs through the bile and the urine; significant accumulation within the body has never been demonstrated.

The ability of surface-active agents to emulsify lipids means that repeated or prolonged exposure to surfactant-containing solutions can cause damage to the lipid film layer that covers the skin surface. As a consequence, the barrier function of the lipids is impaired, leading to increased permeability and loss of moisture. This is evidenced by dryness, roughness, and flaking of the skin. Very prolonged exposure to concentrated surfactant solutions can lead to serious damage and even necrosis. Skin tolerance varies widely among the compounds making up each class of surfactants. Nevertheless, one can generalize that tolerance to surfactants tends to increase in the order cationic, anionic, nonionic materials.

Broadly speaking, all of the commercially relevant surfactants are well-tolerated in the concentration ranges applicable to detergent use [760, 761]. However, certain structure–activity relationships have been found. With nonionic surfactants, for example alcohol ethoxylates, skin irritation diminishes with an increasing degree of ethoxylation [762]. For anionic surfactants, the length of the alkyl chain is closely related to skin irritability. Thus, within a given homologous series, compounds containing saturated alkyl chains of 10–12 carbon atoms show the strongest effects [763].

The eye is much more sensitive than skin to damage by small amounts of surfactants. Anionic surfactant solutions with a concentration above ca. 1 % can produce minor eye irritation, although this is normally reversible [763]. Serious damage is likely only if the eye comes in direct contact with a concentrated surfactant solution and if this contact is not followed by immediate and intensive flushing with water.

Virtually all surfactants that are employed in detergents have been the subject of extensive investigation with respect to their allergenic properties. In no case has an increased risk of allergy for the consumer been demonstrated. A few isolated examples of sudden, multiple allergic reactions in connection with anionic surfactants (alkyl ether sulfates or α-olefinsulfonates) were later traced to impurities of 1,3-sultones and chlorosultones, compounds that have long been recognized as potent allergens and that can arise by bleaching of surfactants with hypochlorite at low pH [764].

The acute oral toxicity of surfactants is low; LD_{50} values normally fall in the range of several hundred to several thousand milligrams per kilogram of body weight [765]. The major detrimental effect of surfactants is damage to the mucous membranes of the gastrointestinal tract. High doses lead to vomiting and diarrhea. If a surfactant reaches the circulatory system, it can cause damage even in very low concentration as a result of interactions with erythrocyte cell membranes, ultimately resulting in destruction of the cells (hemolysis). Inhalation of dust or aerosol containing high concentrations of surfactants can interfere with pulmonary functions [766]. This interference can be partly attributed to interactions with the surface-active film of the vesicles of the lung [767].

The possibility of chronic toxicity has been a subject of intense investigation with members of all of the surfactant classes. Tests with experimental animals involving exposures for up to two years in dosage ranges of several thousand

parts per million have without exception shown complete safety [760, 761, 768]. Moreover, a number of investigations have been conducted over long periods of time with human volunteers. Substantial amounts of both anionic and nonionic surfactants were administered, but no serious side effects were discovered [765].

Neither long-term oral ingestion nor continuous skin exposure has ever suggested that surfactants within the three major groups possess carcinogenic activity [760, 769]. Similarly, surfactants can be regarded as completely devoid of both mutagenic [770] and teratogenic [760, 769] characteristics.

11.1.2. Builders

The most important substances in this category are sodium triphosphate, zeolite A, and sodium nitrilotriacetate (NTA).

Pentasodium triphosphate can be regarded as innocuous. Effects of triphosphate ingestion first become acute with dosages of several grams per kilogram of body weight, and even these manifestations are simply a result of localized damage caused by the high alkalinity of concentrated phosphate solutions [771].

Zeolite A has undergone intensive toxicological investigation and has been found to possess absolutely no acute toxicity. The compound is also tolerated well locally. No adverse effects are seen with eye tissue beyond those associated with the presence of any foreign body. Studies on chronic toxicity or carcinogenicity following oral ingestion have also shown zeolite A to be completely safe when used in detergents, even considering long-term use of the products. Inhalation of the material leads neither to silicosis nor to the development of tumors. Isolated reports of alleged carcinogenic activity have been associated with synthetic zeolites, but these have been shown to be related to the presence of natural zeolites possessing a fibrous morphology. The latter are indeed capable of inducing tumors, unlike zeolite A, which consists of cubic crystals. In summary, zeolite A appears to present no health risks [772].

NTA shows little acute toxicity with respect to both inhalation and oral ingestion. Moreover, the compound produces minimal adverse effects on the skin even after repeated exposure, and it has no sensitizing properties. NTA is readily absorbed through the gastrointestinal tract, but it is then quickly eliminated without being metabolized. Numerous investigations support the conclusion that NTA is not a mutagen. Results from animal studies have shown that repeated applications of large amounts of NTA can cause changes in the urinary tract, including tumor formation, but these changes seem to be due to alterations in the metabolism of certain divalent cations by interaction with the complexing agent NTA. This effect is not considered relevant to conditions likely to be found when NTA is used in detergents. The possibility cannot be excluded that use of NTA could lead to trace amounts in drinking water. Nevertheless, concentrations to date have never been found to exceed a few micrograms per liter, and toxicological data suggest that this presents no human health risk [773 – 778].

11.1.3. Bleach-Active Compounds

Sodium perborate and sodium percarbonate show only slight acute toxicity. Extensive contact with concentrated solutions can cause irritation to the skin and mucous membranes, but this is a consequence of alkalinity rather than a specific effect of the substances themselves. The rapid hydrolysis of the bleaching agent means that any physiological effects that are observed will be those of borates or carbonates, respectively. While it is true that these substances pass readily through the mucous membranes or through damaged skin surfaces, intact skin prevents excessive amounts from being absorbed. However, the intestinal tract is capable of admitting substantial amounts of boric acid, and the substance is considered lethal to adults at a dose of 18–20 g [779]. Studies have revealed reprotoxic effects in animals treated with borate [780]. However, with respect to the exposure situation of normal use of detergents a human health risk can be excluded.

11.1.4. Auxiliary Agents

Enzymes. The proteolytic enzymes employed in detergents are quite safe from a toxicological standpoint. Evidence of local skin ir-

ritation appears to be rare, as are systemic effects following ingestion. As with all proteins, however, human allergic reactions are possible. Indeed, exposure to industrial dust in the early days of the production of enzyme-containing detergents resulted in a number of asthmatic reactions among factory workers. This risk has been eliminated by drastic reduction in the enzyme concentration of industrial dust and by the introduction of coated enzyme granulates, prills, and extrudates.

The possibility that a consumer might inhale significant amounts of enzyme from detergent dust can be essentially excluded. Thus, the only plausible and relevant form of consumer exposure is skin contact in the course of dealing with hand-washable items. At one time a definite relationship was assumed between skin reactions and use of enzyme-containing detergents, but a large body of evidence now shows the contrary. It appears safe to say that the presence of enzymes does not contribute to skin irritation by detergents, nor does it lead to increased risk of allergenic reactions [781 – 783].

Fluorescent Whitening Agents. Much information is now available that supports the conclusion that the fluorescent whitening agents employed in detergents are safe materials. All studies on toxicity, teratogenicity, mutagenicity, and carcinogenicity are consistent in affirming that such compounds present no risk. Local tolerance of the substances is also good, and they display no tendency to cause sensitization. Moreover, long-term studies involving cutaneous application have failed to reveal any carcinogenic potential. Fluorescent whitening agents bind rather tightly to proteinaceous material; as a result, they are firmly retained by the upper keratin layer of the skin, ruling out any significant percutaneous transport. The correspondingly firm binding to textile fibers implies that risks of exposure by way of clothing can also be disregarded. In other words, all available toxicological and dermatological evidence supports the belief that low-concentration detergent use of fluorescent whitening agents presents no human health risk [784, 785].

11.2. Finished Detergents

Appropriate choice of ingredients has enabled manufacturers to assure the virtual absence of health risks associated with detergent use. Potential long-term and systemic risks, such as mutagenicity, carcinogenicity, allergy, etc., all are examined and eliminated in the course of the thorough evaluations that routinely accompany raw material considerations during product development. Nevertheless, the possibility remains that finished products could reveal harmful characteristics despite the safety of the raw materials. Acute toxicity and localized reactions are both potential areas of concern, because complex mixtures occasionally show behavior in these respects different from that predicted for the sum of the components.

In fact, the usual commercial detergents display only minimal acute toxicity. Oral ingestion leads usually to vomiting, and this may be accompanied by irritation of the mucous membranes of the gastrointestinal tract and diarrhea. Low toxicity and the fact that vomiting is virtually unpreventable after ingestion of large quantities of detergent essentially exclude any serious threat of poisoning through product misuse, even with respect to children.

Skin irritation is also practically ruled out for modern detergents, provided they are used as directed. Nevertheless, the surfactants that are present and the alkalinity of detergent solutions cause natural lipids to be removed from exposed skin, which can result in minor irritation. Most reported instances of skin irritation can be attributed to extreme levels of exposure coupled with inadequate skin care.

Similarly, no risks appear to be associated with the use of laundry and clothing that has been treated with modern detergents. Only very small amounts of the detergent components (e.g., surfactants, zeolites, fabric softeners) remain on the fabric after rinsing [786]. Of these minor remnants again only traces can be transferred to the skin, amounts well under the limits of local or systemic toxicity [786, 787].

11.3. Conclusions

The constituents of modern detergents can be regarded as toxicologically well characterized and

safe for the consumer. The occupational safety concerns that remain in the production process center around surfactants, alkaline components, and the local effects these elements can have on the skin and mucous membranes. Detergent enzymes present the potential for allergenic problems, but the risk is minimal if suitable protective measures are implemented.

From the standpoint of consumer protection, the standard commercial detergents are entirely safe. Even inadvertent misuse of the products can be expected to lead to only minor effects, which are readily reversible.

12. Textiles

A dominant characteristic of the current textile market is the large range of available products. Moreover, essentially all of the fiber types that are worked into textiles for everyday use lend themselves to washing with water — although not all fabrics and finished goods are equally able to withstand high wash temperatures, the alkaline medium, and vigorous mechanical action. Of course, articles that cannot be regarded as washable require dry cleaning; examples include wool suits and dresses, silk neckties, and various highly ornamented items made from a variety of different materials. Nevertheless, the overwhelming majority of fabrics and knitwear is washable and is washed — in some cases very frequently. Proper washing conditions vary from fabric to fabric, depending on a number of parameters.

Questions regarding washability and, in particular, choice of optimum washing conditions (e.g., selecting the proper cycle with an automatic washing machine), can no longer be resolved simply by determining what type of fibers is involved. Indeed, it has become increasingly necessary to take into account such factors as the types of dyes present and finishing operations to which a fabric may have been subjected.

Dyeing and Printing.

Dyes for Cellulosic Fibers. All cellulosic fibers (cotton, linen, viscose, etc.) resemble one another in their properties and can all be dyed with the same materials and by similar procedures. What follows is a description of the major types of dyes appropriate to these fibers.

Direct dyes are transferred onto the fibers directly from a hot dye bath. Materials that have been subjected to direct dyeing are quite capable of withstanding a gentle wash cycle at 40 °C, although subsequent treatment with special chemicals can further improve both wash- and water-fastness.

If a textile color is to withstand hot water washing (60 °C) and if a direct dye is to be used, the dye must be carefully selected and a special treatment (fixation) must be administered to increase fastness. Alternatively, dyes can be chosen that lend themselves to after-coppering. In any event, dyes must be chosen from the appropriate fastness grouping at the time of original dyeing.

If a dyed cellulosic fabric is to be subjected to particularly high demands for fastness, a dye of an entirely different type is commonly selected. Among these are vat dyes, reactive dyes, or dyes in the naphthol, phthalocyanine, or sulfur categories. Reactive dyes are capable of forming chemical bonds with cellulose molecules, which display exceptionally high degrees of water fastness.

Dyes for Wool and Silk. Only relatively few such dyes are sufficiently versatile for application to wool and silk. Wool is normally treated with acid dyes, at least in the absence of extraordinary requirements for fastness. Metal-complex dyes have been specially developed for the treatment of wool, and they possess unusually good fastness properties.

Dyes for Synthetic and Cellulose Acetate Fibers. Synthetic and cellulose acetate fibers are generally treated with disperse dyes or, in some cases, pigment dyes. The insoluble coloring agent present is subsequently fixed to the fabric with the aid of heat-setting resins.

The printing of textiles entails the use of essentially the same coloring agents that are employed in dyeing; a separate discussion of this process is therefore unnecessary. In any case, a crucial step in dying and printing of textiles is a special treatment to maximize color fastness. If the color fastness is insufficient, the dyes will severely fade upon washing. The lower the fastness, the more easily colors will fade, independent of the composition and ingredients of the detergents used. Low quality colored textiles may already fade when only water and mechanical action are applied to them.

Finishing. Consumers have come to expect fabrics emerging from washing with relatively few wrinkles. Wrinkles that remain should disappear quickly or easily be ironed off without the need for special treatment. Meeting these expectations has required manufacturers to develop special finishing methods involving the use of special resins. Appropriate water-soluble resin precondensates are applied to the raw fabric, which are induced to condense during a subsequent drying process.

With the assumption that washing is conducted under the proper conditions, resin finishing assures that the original crease angle is largely restored as the fabric dries, thereby eliminating the need for ironing. Furthermore, garments treated in this manner retain their form as they are worn. Resin-finished fabrics are frequently called permanent-press, easy-care, wash-and-wear, or no-iron fabrics.

Fabric finishes tend not to be permanent: the benefits they confer are gradually lost as a garment is subjected to repeated washing. Certain finishing characteristics can be restored by laundry aftertreatment aids, however, particularly softness, stiffness, or antistatic properties.

In some countries (e.g., the USA and Canada), flameproofing is another form of fabric finishing commonly applied to household goods. Items that have been so treated are generally identified by special labels. Strict adherence to manufacturer's washing instructions is necessary if lack of flammability is to be maintained.

A recent form of fabric finishing is the application of sun protection for special textiles such as garments for leisure, sports, and children's wear. Sun protection for fabrics is considered particularly valuable for regions with increased UV radiation due to ozone depletion in the stratosphere. Sun-protection finishings are commercially available, and are permanent and washfast [788, 789].

Washable Fabrics. Textile demand for washable fabrics in Europe (i.e., in the countries comprising the EU) for 1994 is outlined in Table 53 [790]. Three types of material can be seen to account for most of the sales: cotton (47 %), wool (9 %), and synthetic fibers (42 %). Attention is drawn to the fact that by far the largest part of synthetic fibers for domestic use is blended with natural fibers such as cotton and wool. Pure synthetic fibers are used only to a lesser extent for apparel and home furnishing textiles except upholstery and curtains. Table 53 also reveals the distribution of these fabrics over the principal categories of household washables.

In Western Europe, the proportion of white goods in a typical laundry has been quite constant since 1988. This trend is illustrated for Germany in Figure 130. These data can be extrapolated to reflect generally the situation in the European Community as a whole.

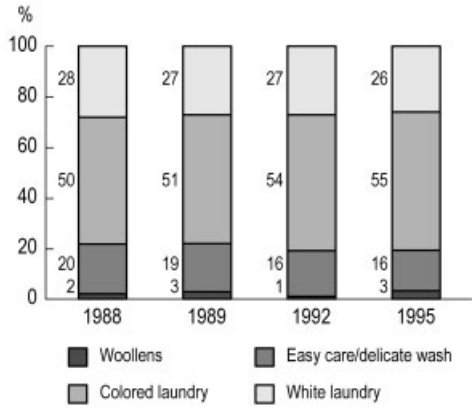

Figure 130. Proportions of wash loads in Germany (%)

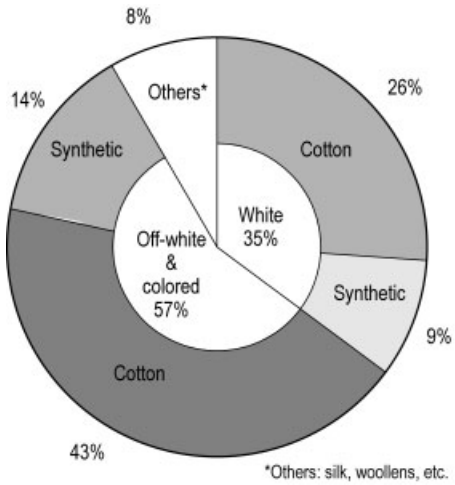

Figure 131. Distribution of white and colored washables in Japan

Table 54 shows the distribution of colored and white washables in the USA. The proportion of white fabrics in Japan is somewhat larger,

Table 53. Fiber consumption in the domestic production of the most important washable textile articles in the European Union 1994 * (1000 t)

Textiles	Total	Wool	Cotton	Cellulosics	Polyamide	Polyester	Acrylics	Others
Apparel								
Workwear	65.3		41.9	1.3	1.3	19.9	0.3	0.6
Skirts	61.8	11.6	17.5	5.5	0.8	23.4	2.1	0.9
Long trousers	70.0	8.7	40.8	1.9	1.7	15.0	1.2	0.7
Sweaters	251.0	66.5	40.9	3.3	6.1	18.1	115.2	0.9
T-shirts and sweater shirts	54.9	1.9	40.7	3.8	0.4	5.3	2.6	0.2
Blouses and shirts	81.8	0.3	43.1	9.8	0.6	26.5	0.9	0.6
Underwear	66.6	0.1	54.1	1.6	4.2	5.4	0.9	0.3
Night wear	37.8	0.5	24.0	0.8	2.6	7.8	1.8	0.3
Footwear	139.0	9.3	49.4		66.5	1.0	8.6	4.2
Brassieres and foundation garments	9.0	0.3	0.8		4.8	1.3	0.2	1.6
Home furnishing								
Towels	71.8		70.3	0.9	0.2	0.1		0.3
Bed linen	181.0		132.9	4.5	0.9	39.2		3.5
Blankets	47.7	12.9	3.1	1.9	0.7	2.0	25.1	2.0
Net curtains	67.4		7.2	2.3	0.2	51.9	5.6	0.2
Table linen	21.5		12.0	1.4	0.3	3.6	1.2	3.0
	100 %	9.1 %	47.1 %	3.2 %	7.5	18.0 %	13.5 %	1.6 %

* Excluding Denmark, Ireland, and Greece.

it accounted for 35 % of total laundry in 1994 (see Fig. 131). The washable fabrics comprise home-washed domestic laundry including clothing, towels, bed sheets, etc. [791].

Washing Conditions. Washable woven and knitted fabrics are normally washed either in the household, in a launderette, or by a professional laundry. The distribution among the three varies widely from country to country. In Germany, for example, more than 90 % of household laundry is done at home, usually in a horizontal-axis drum-type machine. In Europe generally, wherever the percentage of households with washing machines is high, the fraction of the laundry done at home is also high. A major reason for this pattern is the substantial reduction of effort brought by the technological advances incorporated in current automatic washers. The process has become even simpler with the advent of wash-and-wear fabrics. The finishing work associated with a load of laundry currently entails little more than folding and putting away the washed and dried articles (Fig. 132).

Establishing proper washing conditions — temperature, time, mechanical input, wash liquor ratio, and detergent — requires consideration of the characteristics of each of the materials that make up a given article of laundry. For this reason, washability is a characteristic that can only be measured with respect to finished goods. However, conditions imposed during washing, drying, and ironing differ, depending on whether the process is carried out at home or in a commercial laundry. Differences occur in times required for washing and drying, the stress imposed in conjunction with hot-air drying, and the pressure and temperature associated with ironing.

Throughout Europe, white and colorfast cottons are usually washed with a heavy-duty detergent and a normal machine program, i.e., at 60–95 °C with a low wash liquor ratio and a standard amount of mechanical input. Cottons that are not colorfast are usually washed at 40–60°C, frequently with color detergents. Pastel coloreds often require the use of specialty detergents or heavy-duty detergents designed for colored fabrics; their main advantage is the absence of fluorescent whiteners and bleaching agents.

Woolens and silks are regarded as especially sensitive to washing conditions. In some cases, these fabrics must simply be accepted as nonwashable and submitted to dry cleaning. The surface of a wool fiber has a unique "scaly" construction; this peculiar morphology contributes to wool's characteristic tendency toward *felting*. This term is used to describe the appearance of wool after mechanical action has caused individual fibers to become entangled with one another. Wool can be washed much more successfully if it is first given a special treatment known as an

Table 54. Distribution of colored and white fabrics in the USA

Fabric color	Percentage			
(% white)	1993	1995	1997	1999
Only white (100 %)	18	17	18	17
Mostly white (75 %)	12	13	12	12
Mixed white (50 %)	21	19	18	18
Mostly colored (25 %)	16	17	16	17
Only colored (0 %)	33	34	36	36

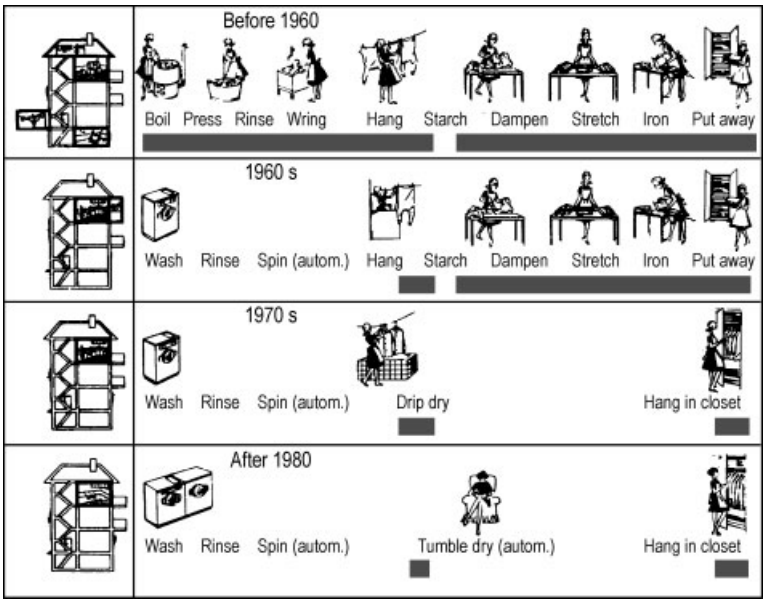

Figure 132. Effort involved in the laundry process in Germany [792]
The reduction in energy expenditure accompanying various stages of technological development through the last 50 years is symbolized by the length of the black bars under the various pictures. The more automated the process and the more nearly the fabrics show "easy-care" ("wash and wear") characteristics, the less the effort required.

antifelting finish. Woolens and silks should always be washed at low temperature (maximum 30 °C), using a high wash liquor ratio and a reduced level of mechanical input. The detergent employed should be either designed for easy-care fabrics or else a product specifically intended for wool.

Synthetic fabrics made from regenerated cellulose fibers (e.g., rayon staple or artificial wool), require a wash temperature of 60 °C, but if sensitive colors are present, the temperature should be reduced to ca. 40 °C. A high wash liquor ratio is recommended, as is a limited amount of mechanical input. Cellulose acetate is generally washed at 40 °C.

Pure synthetics are best washed using a gentle program with a high wash liquor ratio. A heavy-duty detergent is considered appropriate, as is one designed for easy-care fabrics. In some cases a specialty detergent for colored fabrics is necessary (e.g., with uniformly colored pastels, when optical brighteners are to be avoided).

White polyamide fabrics can be washed at 60 °C, colored polyamide fabrics at 30 or 40 °C.

White polyester fabrics are washed at 40 or 60 °C, whereas colored polyesters can only withstand a wash temperature of 40 °C.

Knitware made of polyacrylonitrile is treated like wool, i.e., it is washed in cold water (or at a maximum of 30 °C). White polyurethane-containing undergarments can be washed at 40–60 °C, and colored items at either 40 or 60 °C, depending on the degree of colorfastness.

Many of the easy-care fabrics on the market consist of blends of two or more fiber types brought together as the material is woven or

Table 55. Washability of textiles

Fibers	White		Colored		Pastel colored	
	Temperature, °C	Bath ratio *	Temperature, °C	Bath ratio *	Temperature, °C	Bath ratio *
Natural fibers						
Cotton	95	low	40, 60, or 95 **	low	95 or 60	low or high
Linen	95	low	40, 60, or 95 **	low	95 or 60	low or high
Wool	cold, <30	high	cold, <30	high	cold, <30	high
Silk	cold, <30	high	cold, <30	high	cold, <30	high
Chemical fibers (cellulosics)						
Rayon staple	60	high	60 or 40	high	60 or 40	high
Acetate	40	high	40	high	40	high
Chemical fibers (synthetics)						
Polyamide	60	high	30 or 40	high	30 or 40	high
Polyester	30–60	high	30 or 40	high	30 or 40	high
Polyacrylonitrile	cold, <30	high	cold, <30	high	cold, <30	high
Polyurethane	40–60	high	40–60	high	40–60	high

* Low: bath ratio 1 : 5 and normal mechanical input; high: bath ratio 1 : 20 to 1 : 30 and decreased mechanical input.
** Colorfast items.

knitted. Such blends are treated like easy-care fabrics, using a high wash liquor ratio and gentle agitation. Recommended wash temperatures vary, depending on the nature and composition of the blend. In general, the recommended laundering conditions are those that apply to the major constituent. Wool/polyacrylonitrile or wool/polyester blends are always washed at low temperature (maximum 30 °C), whereas with permanent-press cotton/polyester blends, 60 °C is appropriate.

Items that consist of two or more types of fabric or yarn (e.g., linen tablecloths with wool embroidery, cotton shirts with polyamide piping, jackets with an easy-care exterior and a standard rayon lining, etc.) require different treatment than blended fabrics. The same applies to materials containing dyes of varying degrees of fastness. Such items must always be washed, dried, and ironed under the conditions applicable to the most sensitive fabric or color present, even if this fabric constitutes only a small portion of the whole. For example, a cotton tablecloth with a small amount of wool embroidery requires the same treatment as if it were made entirely of wool. Table 55 provides a summary of the washability of various fibers.

Care Labeling. The European Community (EC), acting in the interest of consumers, has developed a uniform and comprehensive system of care labeling for textile goods. This system has required the cooperation of textile fiber manufacturers, the textile and detergent industries, and representatives of the laundry and dry-cleaning trades. The results are illustrated in Table 56 [793]. Labels of the approved type are permanently affixed to items of clothing prior to sale, providing the consumer with a straightforward guide to proper washing and handling. The symbols themselves are internationally protected trademarks, the rights to which are held by the Groupement International d'Etiquetage pour l'Entretien des Textiles (GINETEX). The German trademark rights rest with the Arbeitsgemeinschaft Pflegekennzeichen für Textilien in der Bundesrepublik Deutschland, itself a member of GINETEX. The German organization also confers the right to use these symbols. No license fees are entailed, but the symbols must be used correctly and in their entirety, as specified in the appropriate guidelines. Improper use can lead to loss of the right to employ the symbols. Care labels must be understood merely as sources of advice; they do not in any sense represent evidence of product quality. Nevertheless, any label that appears must be appropriate to the item in question and must be legible and permanently affixed. In NAFTA countries care symbols used are nearly identical to those employed in the EC. Only in a few cases do they differ somewhat.

Table 56. GINETEX internationalcare labeling symbols

Washing

Symbol	Washing process
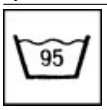	– maximum temperature 95 °C – mechanical action normal – rinsing normal – spinning normal
	– maximum temperature 95 °C – mechanical action reduced – rinsing at gradually decreasing (cool down) – spinning reduced
	– maximum temperature 70 °C – mechanical action normal – rinsing normal – spinning normal
	– maximum temperature 60 °C – mechanical action normal – rinsing normal – spinning normal
	– maximum temperature 60 °C – mechanical action reduced – rinsing at gradually decreasing temperature (cool down) – spinning reduced
	– maximum temperature 50 °C – mechanical action reduced – rinsing at gradually decreasing temperature (cool down) – spinning reduced
	– maximum temperature 40 °C – mechanical action normal – rinsing normal – spinning normal
	– maximum temperature 40 °C – mechanical action reduced – rinsing at gradually decreasing temperature (cool down) – spinning reduced

Table 56. Continued

	– maximum temperature 40 °C – mechanical action much reduced – rinsing normal – spinning normal – do not wring by hand
	– maximum temperature 30 °C – mechanical action much reduced – rinsing normal – spinning reduced
	– hand wash only – do not machine wash – maximum temperature 40 °C – handle with care
	– do not wash – be cautious when treating in wet stage

Chlorine-based bleaching (hypochlorite)

Symbol	Process
	– chlorine-based bleaching allowed – only cold and dilute solutions
	– do not use chlorine-based bleach

Ironing

Symbol	Process
	– iron at a maximum sole-plate temperature of 200 °C
	– iron at a maximum sole-plate temperature of 150 °C
	– iron at a maximum sole-plate temperature of 110 °C – steam-ironing may be risky
	– do not iron – steaming and steam treatments are not allowed

Table 56. Continued

Dry Cleaning	
Symbol	Process
	– dry-cleaning in all solvents normally used for dry-cleaning — this includes all solvents listed for the symbol P plus trichlorethylene and 1,1,1-trichlorethane
	– dry-cleaning in tetrachloroethylene, monofluorotrichloromethane and all solvents listed for the symbol F – normal cleaning procedures without restrictions
	– dry-cleaning in the solvents listed in the previous paragraph – strict limitations on the addition of water and/or mechanical action and/or temperature during cleaning and/or drying – no self-service cleaning allowed
	– dry-cleaning in trifluorotrichloroethane, with spirit (distillation temperature between 150 °C and 210 °C, flash point 38 °C to 60 °C) – normal cleaning procedures without restrictions
	– dry-cleaning in the solvents listed in the previous paragraph – strict limitations on the addition of water and/or mechanical action and/or temperature during cleaning and/or drying – no self-service cleaning allowed
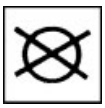	– do not dry-clean – no stain removal with solvents

Laundry Sorting. In Europe, all laundry is customarily sorted on the basis of the care symbols described above. The situation in the USA is rather different, primarily because of the use of different washing conditions (cf. Table 18). In particular, average wash temperatures in North America are significantly lower than in Europe, partly due to the absence of heating units in North American washing machines. Temperature recommendations applicable under these conditions are outlined in Table 57. Sorting of laundry in North America is usually limited to that required for producing entire loads subject to a common wash temperature, degree of agitation, and spin speed, as well as to the same laundry aid treatment (Table 58).

Table 57. Wash water temperatures in the USA

Water temperatures	Use
Hot, 130 °F (54 °C) or above	white and colorfast pastels
	heavy and greasy soils
Warm, 90–110 °F (30–45 °C)	dark, bright, or noncolorfast colors
	moderate to light soils
	knits
Cold, less than 90 °F (< 32 °C)	lightly soiled fabrics
	colors that bleed or fade *

* Heavily bleeding coloreds are to be washed separately from other laundry items.

Table 58. Cloth sorting in the USA

Color	Amount of soil	Fabric and construction
White	heavy	permanent-press and synthetics
Colorfast	normal	towels, jeans, and denims
		synthetic knits
Noncolorfast	light	delicates

In Japan, laundry is customarily sorted into three categories, defined solely by the degree of soil: light, normal, and heavy.

Fabric Labeling. Current EC guidelines specify that all textile products offered to consumers must be accompanied by a detailed description of their content of different fibers. A "textile product" in this regulation is any item whose weight is made up of $\geq 80\,\%$ textile raw material.

The guidelines enumerate a very large number of categories covering the entire range of animal, vegetable, and synthetic fibers, including the newest man-made materials. In this context, the term "synthetic" is in itself not considered sufficient and definitive; more explicit identification of material category is required (e.g., polyester, polyacrylonitrile, polyamide, polyurethane, etc.).

Fiber content is specified on the basis of a percentage of net weight, e.g.:

70 % cotton
20 % polyester
10 % silk

Constituent textile raw materials whose contribution is less than 10 % may be collected under the single heading "other fibers." Thus, a product comprised of 72 % cotton, 7 % polyester, 7 % polyamide, 7 % viscose, and 7 % acetate could be labeled as follows:

72 % cotton
28 % other fibers.

13. Washing Machines and Wash Programs (Cycles)

In contrast to the situation in the USA, the beginning of extensive mechanization of the washing process in Europe was delayed until the late 1950s. A major factor in its ultimate arrival was the development of automatic washers operating on regular home electric current, particularly those whose spin cycle was sufficiently balance-controlled so that the machine would not need to be permanently fastened to the floor.

Automation to some extent has long been present in institutional laundries. This has culminated in systems designed to operate on a production-line basis, resulting in a multifold increase in hourly output per employee and a major reduction in required man power. Nevertheless, only a small fraction of household laundry is washed commercially, primarily because the effort involved in doing laundry at home has been so dramatically reduced by the advent of the household automatic washer.

13.1. Household Washing Machines

13.1.1. Classification

Two basic types of home automatic washers exist: vertical-axis machines and horizontal-axis drum machines. In the Americas and Asia vertical-axis washers of the agitator (Fig. 133 [794]) and the related impeller (also called pulsator) types (Fig. 134 [795]) dominate. They both lack internal heating facilities and, therefore, must be connected to an external source of hot water if a heated wash liquor is desired. By contrast, this type of machine has lost virtually all of its former significance in Europe, where it has been replaced entirely by the horizontal-axis drum-type washer (Fig. 135 [796]).

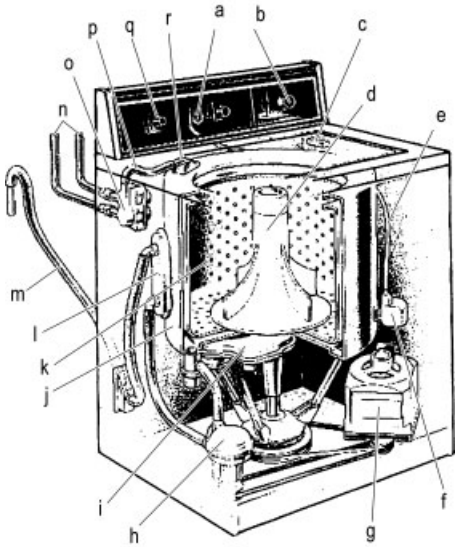

Figure 133. Agitator-type vertical-axis washer
a) Pressure switch; b) Timer, c) Lid switch; d) Agitator; e) Air-pressure tube; f) Air-pressure dome; g) Motor; h) Water pump; i) Transmission; j) Tub; k) Basket; l) Filter; m) Drain hose; n) Fill hoses; o) Water inlet mixing valve; p) Inlet hose; q) Temperature dial; r) Water inlet

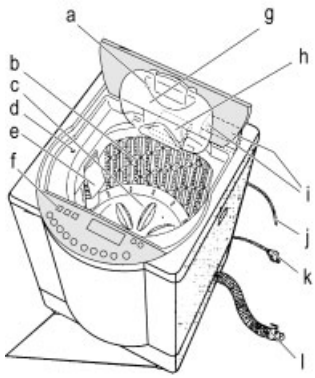

Figure 134. Impeller- (pulsator-) type washing machine
a) Water supply valve; b) Basket; c) Agitator; d) Spirit level; e) Lint filter; f) Control panel; g) Inside lid; h) Detergent and fabric softener dispenser; i) Water supply pump; j) Earth wire; k) Cord and plug; l) Drain hose; m) Leveling feet

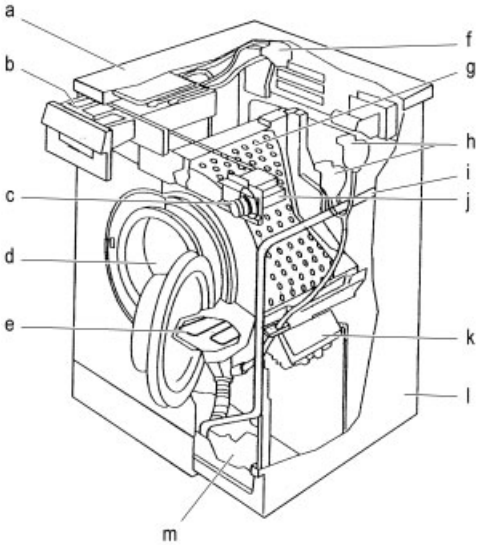

Figure 135. Horizontal-axis drum-type washer
a) Top; b) Dispenser drawer; c) Program control knob; d) Front door; e) Heating coil; f) Magnetic valves; g) Drum and tub; h) Water level control; i) Drain hose; j) Program control; k) Driving motor; l) Casing; m) Drain pump

Agitator Machines. Most American agitator-type washing machines mostly have an enameled metal laundry tub. Inside the tub is an open, perforated vertical axis basket, usually of a size suitable for holding up to 5 kg of laundry.

The laundry tub is filled with 50 to 75 L of water, depending on the wash program chosen and the type and weight of the load (medium or large load). Manufacturers recommend that water and detergent be introduced into the machine first, followed by the laundry, which should always be distributed as evenly as possible. Machines of this type normally have two water connections, one for hot water and one for cold. The desired wash temperature (hot, warm, or cold) is selected by proper setting of the controls, and the selected temperature is achieved by automatic mixing of water from the two input sources. Sodium hypochlorite or oxygen bleach may be added separately by hand during the wash cycle, although quite a few machines are equipped with automatic dispensers for the purpose.

After the desired water level has been reached, the agitator is set to function, and the laundry is jostled for ca. 8–18 min by the rhythmic motion of the agitator (Fig. 133). When the wash cycle is complete, the used wash liquor is pumped off. American machines that are set for the regular cycle then begin a series of spins and spray rinses, whereby fresh cold water is sprayed on the load during spinning to remove residual wash water. Water is again allowed to fill the tub for a brief rinse (ca. 2 min), after which it is once more drained off and a final spin drying occurs.

Impeller (Pulsator) Machines. These fully automatic washing machines are used in Japan, China, and other Asian countries. They use a rotating ribbed disk mostly mounted at the bottom of the tub as their source of mechanical input (Fig. 134). Japanese machines have a capacity of 3 to 8 kg of laundry.

Horizontal-Axis Drum Washing Machines. In automatic drum-type washers, the laundry is placed in a horizontal-axis perforated drum which rotates in a tub in alternating directions (Figs. 135 and 136). For a regular cotton wash, only the lower quarter of the drum is filled with wash liquor (bath ratio ca. 1 : 4), which means that, in contrast to agitator machines and pulsator washers, the laundry is only partly submerged. Individual laundry items are repeatedly lifted by paddles located at the inner wall of the drum, each time falling again and tossed into the wash liquor for renewed soaking and rubbing. The wash liquor is heated internally either by means of electrical heating coils located at the bottom of the tub or by a continuous flow heater. A relatively small number of automatic drum-type washing machines are equipped with a separate hot-water inlet for use in households that have access to an economical external hot-water source. In this case the bath is often tempered initially to 30–40 °C by introduction of cold water, thereby preventing proteinaceous soils and stains from being thermally denatured and set in the fabric. Subsequent heating can then take place if this is required by the particular automatic cycle selected.

Laundry is loaded either from the top through an open lid in the side of the drum (top loader) or through its open front (front loader).

Detergent is introduced through a dispenser. In a front loader, the dispenser is usually a device resembling a drawer located at the front of the machine. The dispenser for a top loader

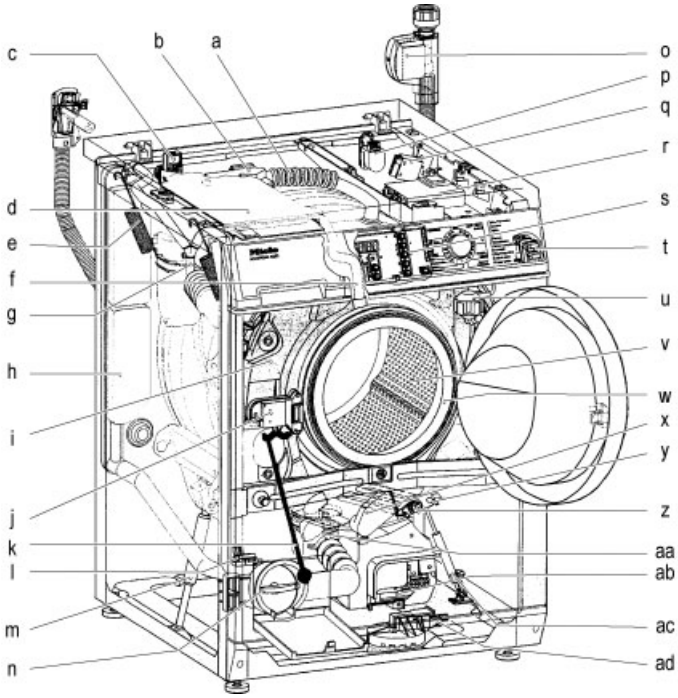

Figure 136. Front loading washing machine
a) Suds container air vent; b) Water path control system; c) Water Control System valve; d) Detergent dispenser; e) Suspension springs; f) Water intake via door glass; g) Drain pump vent/Non-return valve; h) Non-return valve; i) Counterweight; j) Door lock; k) Emergency door release; l) Drain pump; m) Hydraulic shock absorber (Only on models with loading level indication); n) Fluff filter housing; o) WaterProof System valve; p) Interference suppression filter; q) Heater relay; r) Power module; s) Control module; t) Pushbutton switch unit; u) Imbalance sensor (Only on models with loading level indication); v) Inner drum; w) Door seal; x) Air trap; y) Heating; z) NTC resistor; aa) Ball-and-seat valve; ab) Motor; ac) Load sensor (Only on models with loading level indication); ad) WaterProof System switch.

is located beneath the lid of the machine. Dispensers permit automatic sequences consisting of either one or two wash cycles (i.e., a prewash and a main wash). Additional compartments are also provided for the automatic introduction of laundry aftertreatment aids and, especially in Southern Europe, additional automatic dosage of a chlorine-containing bleach during the rinse cycles.

Fully automatic drum type washers utilize one drum for successively accomplishing three separate consecutive operations: washing, rinsing, and spin drying. The machines are equipped with special mechanical devices that compensate for load imbalances. The trend toward space-saving machines has led to the development of smaller washers and washers fitted for built-in kitchens.

Modified European-type horizontal axis drum washers have been introduced in the 1990s in the USA and Japan for reasons of improved efficiency versus current washers in terms of water and energy conservation. Both markets indicate increasing shares of these front loaders. The market shares are, however, still very low.

13.1.2. Operational Parameters

Four factors must be regulated in the washing operation: chemistry, mechanical input, wash temperature, and time.

The effect of each factor on wash performance varies, depending on the washing techniques employed. Figure 137 illustrates graphically by a Sinner's circle the rough percentage influence of the individual parameters on the overall process.

The inner circle implies that the effectiveness of the four factors is a consequence of the medium that unites them, i.e., water.

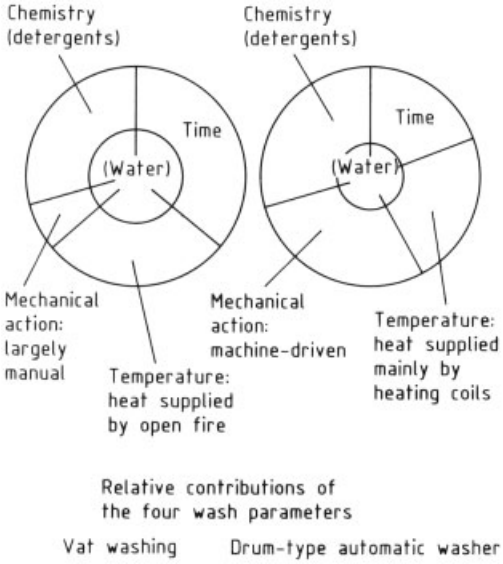

Figure 137. Circular laundry chart (Sinner's circle)

In the days when laundry was done in an open vat, the required amount of water was large, the role of mechanical action was very limited, and time was a very important factor. A washing machine provides considerably more mechanical action; thus, the wash time has been shortened.

Wash Liquor Ratio. "Wash liquor ratio" (or bath ratio) means the ratio of dry laundry (in kilograms) to the volume of wash liquor (in liters). The total amount of wash liquor required in the overall process is made up of two portions: that which is absorbed by the laundry (bound wash liquor) and that which remains in excess (free wash liquor).

Various washing processes have different wash liquor requirements. North American agitator-type washing machines typically use some 25 L of water per kilogram of laundry, i.e., a wash liquor ratio of 1 : 25, whereas Asian pulsator machines require a ratio of ca. 1 : 20 to 1 : 15. Modern horizontal axis drum-type automatic washers are characterized by a low 1 : 4 wash liquor ratio with cotton, but rather high values of up to 1 : 15 for easy-care and delicate fabrics.

Wash Liquor Level and Reversing Rhythm. A drum-type washer is programmed to rotate in a reversing fashion, i.e., first in one direction and then in the opposite. The extent of the mechanical action imparted to the laundry can be altered by changing either the rhythm of reversal or the wash liquor level (or both). A low wash liquor level (wash liquor ratio of ca. 1 : 4) and relatively rapid reversal (e.g., a reversing rhythm of 12 s of rotation/4 s of pause) causes a large mechanical effect on the laundry. In contrast, a high wash liquor level (wash liquor ratio 1 : 15) and slower reversal (e.g., reversing rhythm 4 s of rotation/12 s of pause) provides less mechanical action, hence the term "gentle cycle".

Residual Moisture. Residual moisture is defined as the amount of water remaining in the laundry after draining and spinning. Data is reported as a percentage of the mass of air-dried laundry. Separate spin drying with a vertical drum permits relatively high rates of rotation: up to ca. 2800 rpm. Corresponding residual moisture levels for a cotton laundry load are typically ca. 40–50 %. Automatic drum-type machines spin dry the load at rates of 400–1800 rpm and result in residual moisture levels for cotton of 100–45 %, respectively.

Wash Temperature. In Europe clothes are commonly washed over a wide temperature range of 30–90 °C, whereas elsewhere much lower temperatures are used (e.g., in North America up to ca. 55 °C, in Asia and South America only up to 25 °C or, in exceptional cases, 40 °C). The high wash temperatures frequently used in Europe for decades before the 1980s had been based largely on tradition and on the firmly held belief that "only clothes that have been boiled are really clean." Nonetheless, this attitude has been changing as a result of the increasing popularity of colored and permanent-press fabrics, the desire to conserve energy, and the introduction of more effective multifunctional heavy-duty detergents, designed for use at a lower temperature (cf. Section 4.1). Today, the predominant washing temperature in Europe is 40 °C. Lower temperatures have long been the rule in North America and in Japan, a fact partly explained by the hygienic function associated with use of chlorine-containing bleaches. Nevertheless, the majority of the world's population (Asia, Latin America, Africa) usually washes at tap-water temperature, i.e., cold.

Table 59. Predominant machine washing conditions

	Europe	Japan	USA
Machine type	Front loader (drum)*	Top loader (pulsator)**	Top loader (agitator)**
Washing time, min	20–60	5–15	8–18
Amount of water, L	8–15	30–60	75
Bath ratio	4 : 1	10 : 1	25 : 1
Temperature, °C	40–60	5–25	10–40
Water hardness	medium/high	very low	low/medium
Detergent dosage, g/L	5–10	0.5	1–2

* Horizontal-axis type.
** Vertical-axis type.

Washing Times. Customary wash times also differ substantially between Europe on the one hand and North America and Japan on the other. Much of the difference is attributable to differences in washing temperatures: relatively low temperatures and the absence of built-in heating facilities in North America and in Japan, as compared to Europe, where a typical washer can heat water to 90 °C. The heating process also consumes a great deal of time. Nearly 40 min are required to raise the temperature of the wash liquor in a drum-type machine to a preset value of 90 °C, not counting the time occupied by a prewash. A wash cycle that includes a prewash entails draining about half of the first wash liquor after it has reached a temperature of about 40 °C at the end of the prewash cycle, which adds ca. 15 min to the process.

The cycles provided for most drum-type automatic washers use the temperature as their time controlling variable. Thus, a high wash temperature dictates a long wash cycle whereas a low temperature requires less time. Meanwhile it has become common for manufacturers to install optional "energy-saving cycles," where a lower wash temperature is compensated for by increasing the wash time. The resulting increase in time and mechanical input accomplishes essentially the same end as a high-temperature wash; the corresponding Sinner's circle reveals that a temperature deficit has been balanced by a gain in time and mechanical action.

Washing machines in the Americas and Asia operate considerably more rapidly. They normally require only 8–18 min for the wash cycle, followed by one or two brief rinse cycles. A comparison of machine washing conditions in Europe, Japan, and the USA is given in Table 59 [797].

13.1.3. Wash Programs

Since laundry habits and washing machines vary considerably in different parts of the world, discussion of washing programs under separate headings reflects the three major traditions: those found in Japan, North America, and Europe.

13.1.3.1. Japanese Washing Machines and Washing Conditions

The washing machines found in Japanese households are of the automatic vertical axis impeller/pulsator type. A declining 25 % of the machines were still semiautomatic twin-tub type washers in 1998. They feature a single housing that contains one separate tub each for washing and spin drying, with the impeller/pulsator located in the former. A characteristic of most Japanese washing machines is the absence of a provision for temperature control, since the process is predominantly conducted at tap water temperature.

The most important variables with a Japanese washing machine are the water level, the weight of the wash load, and the extent of mechanical action. Washing time is adjusted to match the degree of soil of the laundry (ca. 5–15 min). Table 60 depicts program selections of a typical Japanese washer.

Table 60. Program selections of a typical Japanese washing machine

Program		Application
Wash Selector	Automatic program	
	Normal	For normally soiled clothes
	Gentle	For lightly soiled clothes
	Heavy	For heavily soiled clothes
	Hand Wash	For delicate fabrics with hand wash/dry mark
	Dry Cleaning	
	Blanket	For blankets, etc.
	Optional program	
	Wash	– To set preferred program (only wash, wash & rinse, only spin, only drain, etc.) – To set automatic/original wash program.
	Rinse	
	Spin/Dry	
Wash Timer		To set washing time
Water Selector	Level	To set water level. If not selected, washer adjust the level automatically
	Bathtub Water	To recycle used wash water for washing/rinsing. (Tap water is supplied to the last rinse.)
Optional Selectors*	Dosage Selector	For automatic detergent dispensing
	Basket Wash	To remove soap scum from basket using chlorine bleach. (Scum may cause malodor on clothes.)
	Spin Selector	To select spin level, depending on type of garment

* Depending on manufacturer of the washer.

13.1.3.2. North American Washing Machines and Washing Conditions

Although horizontal-axis drum-type machines have gained importance in the U.S. [798] and Canadian markets in the late 1990s, the automatic agitator washing machine still is by far the most common in North American households. Such machines perform a full set of laundry operations in the usual sequence, i.e., wash, rinse, and spin dry. North American machines resemble their Asian counterparts in that they lack heating coils. Therefore, they are designed for connection to domestic supply of both hot and cold water. Three wash temperatures are generally used: hot (55 °C/130 °F), warm (30 to 45 °C/90 to 110 °F), and cold (10 to 25 °C/50 to 80 °F). Rinse temperatures can also be selected, usually from among the two alternatives, warm and cold. One further variable is the water level, which is adjusted according to the quantity of laundry, e.g., medium and large load, and type of laundry to be washed. The mechanical input is also adjustable within limits by introducing changes in the agitator speed and the wash time. The following set of washing programs is typical (cf. Tables 18 and 61):

Regular (heavy-duty wash)
Permanent press
Knit/gentle
Woolens
Soak
Prewash

13.1.3.3. European Washing Machines and Washing Conditions

Horizontal-axis drum-type washing machines in which the water can be heated to up to 90 °C dominate the household market in Europe. To meet the demands posed by current laundry, such machines are provided with a wide variety of wash cycles. In choosing the appropriate cycle for a given load of laundry, the first decision is whether or not a prewash is desired. Options in later stages are supported by the presence of a special third compartment in the detergent dispenser, which permits addition of aftertreatment aids to the final rinse. Machines built for the Southern European market commonly offer a fourth compartment, which allows addition of chlorine bleach to the second rinse.

Selection of the proper cycle is accomplished in one of several ways. Some machines are equipped with one control knob for establishing the program and a separate infinitely variable temperature regulation control. Alternatively, both functions can be combined in one knob while other models offer a set of pushbutton controls. Separate temperature control provides the greatest degree of flexibility.

Management of any wash cycle requires a more or less complicated switching system. Very few machines (mostly those in the lowest price category) continue to be equipped with time-based switching, an approach in which each operation is allowed to continue for a precise and

Table 61. Typical wash programs in North America

Wash program	Fabric	Degree of soiling	Wash time, min	Water temperature	
				Wash	Rinse
Regular heavy	cotton, linen	heavy	12–18	warm or hot	cold
		moderate	10–12	warm	cold
		light	8–10	cold or warm	cold
Knits/gentle	synthetic	moderate	10–12	warm	cold
	delicate	light	8–10	warm	cold
Permanent press	Permanent press cotton, cotton/synthetic fiber blends	heavy	12	warm or hot	cold
		light to moderate	8–12	warm	cold
Woolens	wool	light	4		
Soak	colorfast	heavy	30	cold or warm	none
		stained	30	cold or warm	none
Prewash	colorfast	heavy	6	cold or warm	none
		stained	6	cold or warm	none

Table 62. Typical wash programs for a European drum-type automatic washer

Designation of program	Application	Prewash temperature, °C	Main wash temperature, °C	Cycle		Water level		Number of rinse cycles	Final rinse water retained in drum	Spin drying	
				Regular	Gentle	Low (12–15 L)	High (20 L)			Spin before removal	Regular
Normal ("boiling") cycle, 90°C	whites and fast coloreds (cotton, linen, and their blends)	40	85–90	+		+		3–4			+
Colored fabrics, 60 °C	nonfast colors (cotton, linen, and their blends)	40	55–60	+		+		3			+
White easy-care, 40 or 60 °C	white synthetics (including polyamides, polyesters, and their blends with cotton)	40	40 or 55–60		+	+		3	+	+	
Colored easy-care or delicate washables, 30 °C	colored synthetics, e.g., polyamides and polyesters; polyacrylonitrile, polyacrylonitrile/wool, rayon, acetate, and cupro	30	30		+	+	+	3	+		+
Wool, silk, 30 °C	washable woolens		30/40		+	+	+	2–3	+	+	+
Fabric softener	normal and colored fabrics			+		+					+

predetermined length of time. The principal disadvantage of such a system is that temperature cannot be properly established without taking into account the temperature and volume of the water being introduced and the ambient temperature, all of which can vary.

Modern washing machines are mostly equipped with integrated electronic circuitry based on microprocessors, which may be regarded as microcomputers. These circuits take full functional control of the machine. So-called fully electronic washing machines of this type lack any complex mechanical switching formerly typical of automatic washers.

Modern washers frequently also have so-called fuzzy logic control governing partial processes of the washing program. Fuzzy logics are able to automatically provide, for example, variable speed of reversion of the drum depending on the amount of foam during the wash cycle

[799], provide an additional rinse cycle depending on the quantity of residual foam after the regular final rinse, control the quantity of water intake and washing time depending on weight and type of wash load, etc. Fuzzy logic provides for increased fabric care and water and energy conservation. European washing machines have been reviewed [800].

Today, most garments sold in Europe have uniform labels specifying the use of one of four wash temperatures: 30 °C, 40 °C, 60 °C, and 95 °C (see Table 56). Table 62 gives examples of the wash cycles commonly provided to meet these needs (cf. also Table 18). Similar care labels are found in USA and Japan.

13.1.4. Energy and Water Consumption

Mechanization of the washing process has relieved much of the burden formerly resting on the person doing the wash. At the same time, however, steady increases in energy and water supply costs have made the use of a household drum-type washing machine increasingly expensive. In particular, the cost increases as the selected wash temperature is increased (Table 63), due to the fact that washers are usually poorly heat insulated.

Table 63. Typical energy consumption for various wash programs with a European drum-type automatic washer

Program	Energy consumption, kWh
90 °C Boiling wash program	1.8
60 °C Colored fabrics	1.2
60 °C Synthetics	1.2
40 °C Synthetics	0.6
30 °C Woolens	0.3

It is not surprising that efforts to reduce the costs have begun with the 90 °C cycles, since these hold the greatest potential for energy savings (e.g., by lowering the water temperature from 90 °C to 60 °C). Some washers have herefore incorporated an optional energy conservation program. Such cycles are usually designed to function at 60 °C. Examples of typical energy conservation measures are shown in Table 64. (For information on the related trend in the USA toward washing at lower temperatures, see Fig. 95.)

With respect to conserving water and energy, European washers have made considerable progress during the 1980s and 1990s. This has been made possible through automatic control of the amount of water intake depending on the quantity of the wash load, by omitting the former intermediate cool-down rinse at the end of a hot or warm wash cycle, by considerably increasing the number of revolutions per minute for spinning, applying spray-spinning and, last but not least, by reducing the number of rinse cycles. To encourage consumers to use washing machines that operate with lower consumption of electric power and water all new washing machines manufactured in Europe since 1996 have to display the Energy Label (Fig. 138) as per Official Directive of the European Union. The label indicates the efficiency of the appliance in terms of energy and water consumption, washing performance, capacity, etc. [220].

13.1.5. Construction Materials Used in Washing Machines

Materials used in the construction of washing machines must exhibit a high degree of resistance to various mechanical, heat, and chemical influences. The following materials are particularly well-suited to the purpose:

Stainless Steel. The vast majority of drums are made of stainless steel. Chromium steel continues to be widely used for the construction of tubs, heating coils, and thermostat sensors. Wash tubs have been increasingly constructed from plastic in the 1990s.

Sheet Steel. Casings and lids are made from sheet steel with a protective surface (e.g., enamel paint, baked enamel, galvanization). Baked enamel provides the most permanent form of protection.

Plastics. Certain components can be made from alkali-resistant plastics, usually polypropylene. Typical applications include detergent dispensers, lint filters, paddles, operating controls, mechanical parts, and various small components. Glass fiber reinforced polypropylene has also been employed for tub construction in drum-type machines..

Table 64. Examples of energy conservation programs for European automatic washers

Modification of program	Wash load, kg	Savings in comparison to usual 90 °C-programs, %	
		Electricity	Water
No prewash cycle	4	10	11
Reduction of the washing temperature from 90 °C to 60 °C	4	30	–

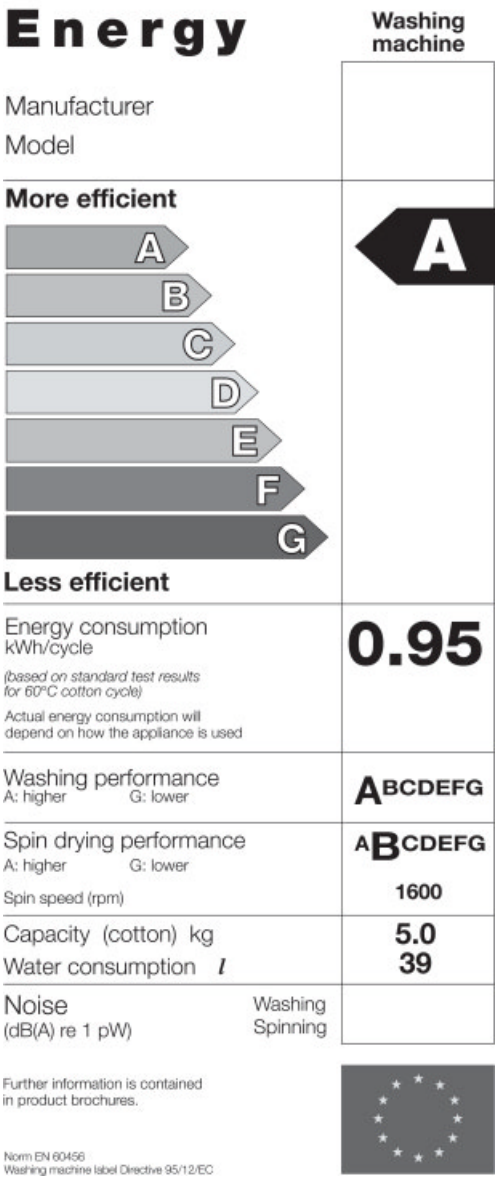

Figure 138. European Union energy label

Rubber. Molded rubber is the material of choice for components that require a high degree of elasticity. This applies particularly to gaskets, such as those around the door, as well as to various hoses, hose connectors, and shaft seals. Synthetic rubber is used exclusively for such purposes. A drawback of rubber is its tendency to deteriorate with age. Deterioration is accelerated by exposure to grease, by which the rubber gradually acquires a soft and greasy character.

13.1.6. The Market for Washing Machines

The majority of the world population cleans laundry by hand, washing either in bowls, tubs, sinks, or in river waters. Washing machines dominate only in industrialized global regions.

The percentage of households equipped with washing machines varies considerably among the countries of the EU. Percentages for 1998 are as follows [801]:

Austria	88
France	91
Germany	95
Great Britain	93
Italy	99
Sweden	63
The Netherlands	95
Europe, average	89

In 1998, 3.36×10^6 automatic washers were produced in Germany alone. Total European production was more than 16×10^6 units with a total consumption number of some 11×10^6 in 1998 [801]. A report and analysis on washing machines in Europe has been published. It contains detailed statistical figures for all countries of the European Union [802].

Worldwide, the USA represents the largest single washing machine market. Nevertheless, market penetration in the USA is lower than that in Europe and Japan; 80 % of American households owned a washing machine in 1998 [803]. In the same year, the USA manufactured some 7.0×10^6 washers [803].

More than 99 % of the households in Japan are equipped with a washing machine. The Japanese manufactured 4.4×10^6 units in 1999. Of these, 3.8×10^6 were fully automatic, whereas 0.6×10^6 units were of the twin-tub type.

Market penetration of washing machines in Brazil was 28 % of all households in 2000, in China it was 10 %, and in India 7 % [803].

13.2. Laundry Dryers

Laundry dryers represent a further stage in the development of laundry technology. Dryers are machines that utilize an influx of heat to cause evaporation of the moisture remaining in spin-dried laundry. Electricity is normally used to supply the heat, although some of the dryers in the USA and Europe are heated by gas. Several basic types of dryers exist, including those based on circulation of heated air, heat transfer from metal surfaces, and radiant heating. Most are drum-type machines, so-called tumblers. They can be categorized by the way they dispose of the resulting steam:

Air-Vented Dryers. Dryers in this category function on the same principle as a fan. Ambient air is drawn in by suction, passed through a heated zone, and then introduced into the laundry drum. As the air becomes enriched in moisture, it is vented through a lint filter either directly to the room or out of the building through an exhaust hose or duct.

Condenser Dryers. In these dryers the air needed for drying is drawn from the ambience and passed over a heating device before encountering the laundry. Moist air leaving the tumbler drum is cooled by an additional supply of incoming ambient air. This cooling takes place in the presence of a drip screen, which collects precipitated drops of water. Water from the drip screen collects in a condensate vessel which has to be emptied after each drying cycle. The working principle is depicted in Fig. 139 [796]. The most recent developments comprise condenser dryers that use heat generated by a heat pump [804].

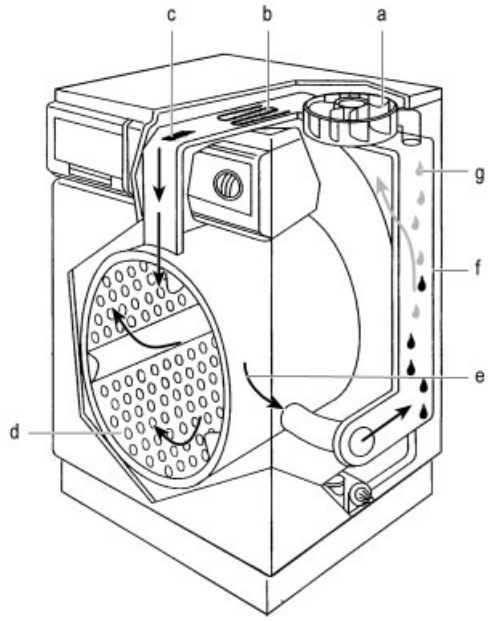

Figure 139. Condenser dryer
a) Ventilator; b) Heater; c) Drying air; d) Tumbler drum; e) Moist air exit; f) Heat echanger; g) Condensing water

Combination Washer–Dryers. The European washer–dryer is a further development of the drum-type washing machine. It is comprised of a washer combined with a water-cooled condensation dryer, and the same drum is used for washing, spinning, and drying. A drawback of these machines is the fact that the small size of the washing drum requires a full load of laundry to be divided into two batches before it can be dried. The heat necessary for drying is supplied either by radiation or by heated air. The market share of these combines in Europe has remained low. In Germany it was only some 4 % in 1999.

Market for Laundry Dryers. The market penetration for laundry dryers in the EU was 27 % on average in 1998 [801]. Percentages for individual countries vary considerably:

Austria	12
France	28
Germany	31
Great Britain	42
Italy	2
The Netherlands	44
Sweden	30

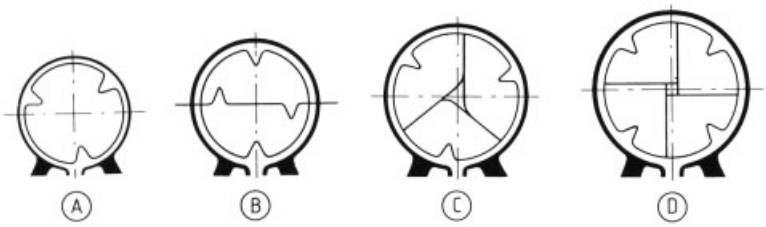

Figure 140. Four common designs for commercial washing machines
A) Open machine; internal drum diameter up to 900 mm; inner drum lacking subdivisions; B) Pullmann type machine; internal drum diameter 900–1400 mm; inner drum divided into two equivalent compartments by means of an axial partition; C) Y-type machine; internal drum diameter 1400–1700 mm; inner drum divided into three compartments by three axial partitions joined along the axis; D) Star-pattern machine, internal drum diameter > 1700 mm; inner drum divided into four compartments by four axial partitions joined along the axis

Major manufacturing countries are Great Britain (1.2×10^6 units in 1998), Germany (0.9×10^6), France (0.5×10^6) and Italy (0.3×10^6). The majority (65 %) of the dryers in Europe are air-vented dryers and the remainder (35 %) represent dryers that use condensation technology.

The situation in the USA is quite different from that in Europe. In the USA, 74 % of all households are equipped with dryers, of which some 27 % are gas-heated, the remaining 73 % being electrically heated units. U.S. production of household dryers amounted to 5.2×10^6 units in 1998 of which 23 % were gas-heated dryers [803].

In 1999, some 20 % of the Japanese households operated a laundry dryer. Approximately 45 % of the dryers have a capacity of more than 5 kg of laundry. The largest part of the remaining 55 % consists of units that have a capacity of 4–4.5 kg.

13.3. Washing Machines for Institutional Use

Washing machines designed for use by commercial laundries are generally designed to keep operating costs as low as possible. As a result, the more labor-intensive batch laundry systems are being increasingly replaced by more efficient continuous devices. In contrast to European households, where laundry water is heated electrically, commercial laundries worldwide usually employ high-pressure steam as their source of heat. Heat from the steam is led into the wash liquor either directly or indirectly, resulting in more rapid heating and shorter wash times.

13.3.1. Batch-Type Machines

There are many situations even today in which batch-type commercial washing machines are indispensable. This is particularly true where a frequent need exists to process numerous small batches of different kinds of laundry. Batch-type machines require a substantial amount of manual supervision; thus, their commercial use is expensive.

Conventional Drum-Type Devices. Such devices are basically scaled-up versions of household washing machines. Commercial front loaders often allow the processing of up to 400 kg of laundry, with special designs allowing even higher amounts. Reversal and rotation rates commonly produce *g*-factors during washing of 0.7–0.9. During spin extraction 400 *g* and higher rates are usual. The large drums require that the effect of mechanical action resulting from great fall heights and excessive weight be moderated and that loading and removing the laundry be simplified. These targets are generally accomplished by introducing into the drum a set of partitions that divide the drum into compartments. Three partitioning schemes (the Pullmann, Y, and star-pattern systems) are illustrated in Figure 140.

Machines of this type are nowadays controlled by microprocessors. Final drying is also performed in large commercial-type equipment.

Wash Extractors. Machines in this category, with a capacity of up to 360 kg, are capable of not only the wash step, but also a spin drying and "fluffing" of the laundered items. Spinning occurs after each wash and rinse operation, an expedient that has the effect of minimizing the number of required rinses. This in turn conserves water and reduces overall wash time to ca. 40 min (Fig. 141). Wash extractors are used in all parts of the world. The machine types, however, vary from manufacturer to manufacturer.

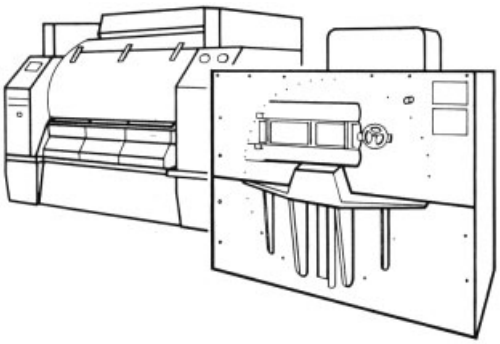

Figure 141. Schematic diagram of a wash extractor

13.3.2. Continuous Batch Washers

The development of the continuous batch washer is a consequence of the demand for a laundry system requiring even less manual intervention. Such machines operate uninterrupted and are best suited to institutions faced with large amounts of relatively uniform laundry. Soiled laundry is loaded onto either a conveyor belt or a suspended conveyor system. As it progresses through the machine, the laundry is exposed to a series of zones designed to accomplish the equivalent of wetting, prewashing, main washing, and rinsing. Most of the zones operate on the countercurrent principle: all water encountered by the laundry flows continuously and in a direction opposite to the motion of the laundry. Each zone leads directly into the next, resulting in significant savings of energy, water, and detergent. Typical water consumption for hospital linen is 8–12 L/kg, sometimes even less.

Following the actual washing process, the laundry is directly loaded into an extraction press or a spin extraction to remove the water. The damp goods are then passed to steam or gas heated batch dryers. The entire process occurs without manual intervention. Continuous batch washers are found all around the globe. But especially in Europe high amounts of goods are treated in this type of machines. Figure 142 illustrates the design of a typical continuous batch washer, in this case one of the Senking type [805].

14. References

1. K. Lindner: *Tenside-Textilhilfsmittel-Waschmittel-Waschrohstoffe*, Wissenschaftl. Verlagsges. Stuttgart, vols. I and II, 1964, vol. III, 1971.
2. A. M. Schwartz, J. W. Perry, J. Berch: *Surface active Agents*, vol**s. I and II**, Interscience Publ., New York 1958.
3. *Kirk-Othmer*, **22**, 332–432.
4. J. Falbe: *Surfactants in Consumer Products*, Springer Verlag, Heidelberg 1987.

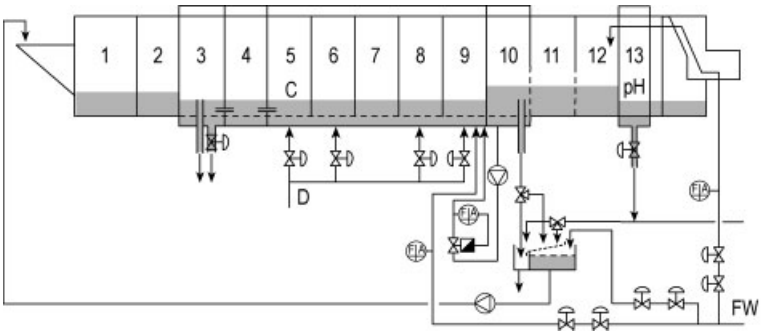

Figure 142. Continous batch washer

5. Kao Corporation: *Surfactants*, Chuo Toshado Printing Co., Tokyo 1983.
6. M. J. Rosen: *Surfactants and Interfacial Phenomena*, J. Wiley & Sons, New York–Chichester–Brisbane –Toronto 1978.
7. H. Stache: *Tensid-Taschenbuch*, 2nd ed., Hanser Verlag München–Wien 1981; 3rd ed., 1990.
8. E. H. Lucassen-Reynders: *Anionic Surfactants*, Marcel Dekker, New York 1981.
9. K. Laux et al.: „Tenside," in *Winnacker-Küchler*, 4th ed., vol. **7** (1986) pp. 84–148.
10. B. Werdelmann in: *Proceedings of the World Surfactants Congress München*, vol. **I**, Kürle Verlag, Gelnhausen 1984, p. 3.
11. H. Andree, P. Krings, H. Verbeek: „Tenside und ihre Kombinationen in zukünftigen Waschmittelformulierungen," *Seifen Oele Fette Wachse* **108** (1982) 277–282.
12. K. J. Bock, H. Stache: "Surfactants" in: The Handbook of Environmental Chemistry, 3, Part B, Springer Verlag, Berlin–Heidelberg–New York 1982, pp. 163 –199.
13. Henkel KGaA: *Fatty Alcohols. Raw Materials*, Methods, Uses, Düsseldorf 1981.
14. W. Kling: „Waschmittel 1928, 1968 und 2000," *Chem. Ind. (Düsseldorf)* **20** (1968) 394.
15. Henkel & Cie, GmbH: *Waschmittelchemie*, Hüthig Verlag, Heidelberg 1976.
16. H. Hauptmann: *Grundlagen der Wäschereichemie*, Spohr-Verlag, Frankfurt 1977, p. 451.
17. Surfactant Science Series, vol. **1–87**, Marcel Dekker, New York.
18. *Tensid Taschenbuch*, 1st, 2nd, and 3rd ed., Hanser Verlag, Munich-Vienna 1981, Leipzig 1990.
19. W. G. Cutler, R. C. Davis: *Detergency*, **Part I–III**, Marcel Dekker, New York 1973 –1981.
20. „Chemie der Waschmittel und der Chemisch-Reinigung," *Chem. Ztg.* **99** (1975) 161–211.
21. "Proceedings-World Conference on Detergents," 1978–1998, A. Cahn (ed.), AOCS Press, Champaign (IL), USA.
22. H. Strone, W. Löbrich, R. Senf, K. D. Wetzler: *Bildungsfibel*, **Part 1**, Deutscher Textilreinigungs-Verband, Bonn 1980.
23. Henkel KGaA: *Industrielles Waschen*, Rheinisch-Bergische Druckerei GmbH, Düsseldorf 1978.
24. A. Davidsohn, B. M. Milwidsky: *Synthetic Detergents*, 6th ed., G. Godwin Ltd., London, J. Wiley & Sons, New York 1978.
25. M. J. Rosen, H. A. Goldsmith: *Systematic Analysis of Surface Active Agents*, 2nd ed., J. Wiley & Sons, London 1972.
26. D. Hummel: *Analyse der Tenside*, Hanser Verlag, München 1962.
27. R. Wickbold: *Die Analytik der Tenside*, Chem. Werke Hüls AG, Marl 1976.
28. B. M. Milwidsky: *Practical Detergent Analysis*, Mac Nair-Dorland Comp., New York 1970.
29. H. König: *Neuere Methoden zur Analyse von Tensiden*, Springer Verlag, Berlin 1971.
30. G. F. Longman: *The Analysis of Detergents and Detergent Products*, J. Wiley & Sons, London 1975.
31. L. H. T. Tai (ed.): *Formulating Detergents and Personal Care Products: A Guide to Product Development*, AOCS Press, Champaign (IL) 2000.
32. M. J. Rosen, M. Dahanayake (eds.): *Industrial Utilization of Surfactants*, AOCS Press, Champaign (IL) 2000.
33. V. K. Pillai, D. O. Shah (eds.): *Dynamic Properties of Interfaces and Association Structures*, AOCS Press, Champaign (IL) 2000.
34. F. Bohmert: *Hauptsache sauber? — Vom Waschen und Reinigen im Wandel der Zeit*. Stürtz-Verlag, Henkel KGaA, Düsseldorf 1988.
35. H. Eierdanz: *Perspektiven nachwachsender Rohstoffe in der Chemie*. VCH Verlagsgesellschaft, Weinheim, Germany 1996.
36. Fonds der Chemischen Industrie: Tenside. Folienserie 14 des Fonds der Chemischen Industrie, Frankfurt 1992.
37. Grugel/Puchta/Schöberl: *Waschmittel und Wäschepflege*, Falken-Verlag, Niedernhausen 1996.
38. Industrieverband Körperpflege und Waschmittel e.V. (IKW): Wäsche und Pflege. Informationen zu Waschmitteln, Textilwäsche und Umwelt, jährliche Berichte und weitere Materialien, Frankfurt 1996.
39. G. Jakobi, A. Löhr: *Detergents and Textile Washing*, VCH, Weinheim, Germany 1987.
40. H. Krüßmann, R. Bercovici: "Chemische Aspekte beim Waschen," *Tenside, Surfactants, Deterg.* **27** (1990) H. 1, S. 8 ff.
41. P. Kuzel, T. Lieser: "Bleichsysteme," *Tenside, Surfactants, Deterg.* **27** (1990) 23.
42. F. Lang, H. Berenhold: "Weichspüler — ein Beitrag zur modernen Wäschepflege," *Seifen, Oele, Fette, Wachse* **117** (1991) 690.

43. Henkel KGaA: *Fettalkohole*, Rohstoffe, Verfahren und Verwendung, Düsseldorf 1981.
44. *Gesetz über Detergentien in Wasch- und Reinigungsmitteln*, 5. 9. 1961, BGBl. I, p. 1653.
45. *Verordnung über die Abbaubarkeit von Detergentien in Wasch- und Reinigungsmitteln*, 1. 12. 1962, BGBl. I, pp. 698–706.
46. G. Jakobi, M. J. Schwuger, *Chem. Ztg.* **99** (1975) 182–193.
47. 759. M. J. Schwuger in E. H. Lucassen-Reynders (ed.): *Anionic Surfactants*, Marcel Dekker, New York 1981.
48. 760. G. Jakobi, M. J. Schwuger, *Chem. Ztg.* **99** (1975) 182–193.
49. 761. H. D. Doerfler, *Grenzflächen- und Kolloidchemie*, VCH Weinheim, Germany 1994, p. 39.
50. 762. J. A. Finch, G. W. Smith in E. H. Lucassen-Reynders (ed.): *Anionic Surfactants*, Marcel Dekker, New York 1981.
51. 763. T. Young, *Philos. Trans. R. Soc. London* **95** (1805) 65, 82.
52. 764. E. G. Shafrin, W. A. Zisman, *J. Phys. Chem.* **64** (1960) 519.
53. 765. Y. Shen, A. Couzis, J. Koplik, C. Maldarelli, M. Silvina Tomassone, *Langmuir* **21** (2005) 12, 160.
54. 766. M. Buzzacchi, P. Schmiedel, W. von Rybinski, *Colloids Surf.* **273** (2005) 47.
55. 767. P. Berth, M. J. Schwuger, *Tenside Det.* **16** (1979) 3.
56. 768. D. Nickel, C. Nitsch, P. Kurzendörfer, W. von Rybinski, *Progr. Colloid Polym. Sci.* **89** (1992) 249–252.
57. 769. T. Engels, W. von Rybinski, P. Schmiedel, *Progr. Colloid Polym. Sci.* **111** (1998) 117.
58. 770. D. Nickel, W. von Rybinski, E.-M. Kutschmann, C. Stubenrauch, G. H. Findenegg, *Fett/Lipid* **98** (1996) 363.
59. 771. E.-M. Kutschmann, G. H. Findenegg, D. Nickel, W. von Rybinski, *Colloid Polym. Sci.* **273** (1995) 565.
60. 772. W. von Rybinski, M. J. Schwuger, *Colloids Surf.* **26** (1987) 291–304.
61. 773. M. J. Schwuger, *Ber. Bunsenges. Phys. Chem.* **83** (1979) 1193.
62. 774. M. L. Gum, M. P. Aronson, E. D. Goddard, *Colloids Surf.* **4** (1982) 87.
63. 775. B. Dobias, X. Qiu, W. von Rybinski, *Solid-Liquid Dispersions*, M. Dekker, New York 1999, p. 340.
64. 776. J. Leja, *Surface Chemistry of Froth Flotation*, Plenum Press, New York 1982, p. 296.
65. 777. D. Balzer, M. J. Schwuger, *Colloid Polym. Sci.* **257** (1979) 292.
66. 778. M. J. Schwuger, H. G. Smolka, *Colloid Polym. Sci.* **254** (1976) 1062.
67. 779. M. J. Schwuger, H. G. Smolka, *Tenside Deterg.* **16** (1979) 233–239.
68. 780. M. J. Schwuger, *J. Am. Oil Chem. Soc.* **59** (1982) 265.
69. 781. H. Upadek, B. Kottwitz, B. Schreck, *Tenside Surf. Det.* **33** (1996) 385.
70. 782. E. Smulders: *Laundry Detergents*, Wiley-VCH, Weinheim 2002, p. 30.
71. 783. H. G. Smolka, M. J. Schwuger, *Tenside Deterg.* **14** (1977) 222.
72. 784. G. Jacobi, *Angew. Makromol. Chem.* **123/124** (1984) 119–145.
73. 785. H. Upadek, P. Krings, *Seifen Oele Fette Wachse* **117** (1991) 554–558.
74. 786. P. Zini (ed.): *Polymeric Additives for High Reforming Detergents*, Technomic, Lancaster 1995, pp. 179–212.
75. 787. P. Zini, *J. Chem. Technol. Biotechnol.* **50** (1991) 351–359.
76. 788. M. J. Schwuger, M. Liphard, *Colloid Polym. Sci.* **267** (1989) 336–344.
77. 789. A. de Lucas, L. Rodriguez, P. Sanchez, J. Lobato, *Tenside Surf. Det.* **39** (2002) 94.
78. 790. W. Kling, *Kolloid Z.* **115** (1949) 37.
79. 791. M. J. Schwuger, H. G. Smolka, *Colloid Polym. Sci.* **255** (1977) 589.
80. 792. D. Nickel, H. D. Speckmann, W. von Rybinski, *Tenside Surf. Det.* **32** (1995) 470–474.
81. 793. R. Hofmann, D. Nickel, W. von Rybinski, *Tenside Surf. Det.* **31** (1994) 63–66.
82. 794. C. P. Kurzendoerfer, H. Lange, *Fette Seifen Anstrichm.* **71** (1969) 561.
83. 795. F. Jost, H. Leiter, M. J. Schwuger, *Colloid Polym. Sci.* **266** (1988) 554.
84. 796. M. J. Rosen, X. Y. Hua, *J. Am Oil Chem. Soc.* **59** (1982) 582.
85. 797. F. Schambil, M. J. Schwuger, *Colloid Polym. Sci.* **265** (1987) 1009.
86. 798. M. Kahlweit, *Tenside Surf. Det.* **30** (1993) 83.
87. 799. H. L. Benson, K. R. Cox, J. E. Zweig, *HAPPI* **22** (1985) 50.
M. Kahlweit, R. Strey in H. L. Rosano, M. Clausse (eds.): *Microemulsion Systems*, Marcel Dekker, New York–Basel, 1987, p. 12.
88. 800. C. A. Miller, K. H. Raney, *Colloids Surf. A* **74** (1993) 169.
89. 801. E. Ryjkina, H. Kuhn, H. Rehage, F. Müller, J. Peggau, *Angew. Chem. Int. Ed.* **41** (2002) 983.

90. 802. K. Siekmann, M. J. Schwuger, *Physik in unserer Zeit* **18** (1987) 178.
91. 803. E. J. W. Verwey, J. T. G. Overbeek: *Theory of the Stability of Hydrophobic Colloids*, Elsevier, Amsterdam 1948.
92. 804. H. Sonntag, K. Strenge: *Koagulation und Stabilität disperser Systeme*, VEB Deutscher Verlag der Wissenschaften, Berlin (GDR) 1970.
93. 805. H. Lange in K. H. Mittal (ed.): *Adsorption at Interfaces*, ACS Symp. Ser. no. 8 (1975) 270.
94. 806. B. Dobias, X. Qiu, W. von Rybinski, *Solid-Liquid Dispersions*, Marcel Dekker, New York, Basel 1999, pp. 159, 257.
95. 807. R. J. Hunter, A. E. Alexander, *J. Colloid Sci.* **18** (1963) 820.
96. 808. D. E. Yates, T. W. Healy, *J. Chem Soc. Faraday Trans.* 1 **76** (1980) 9.
97. 809. R. Hage, A. Lienke, *Angew. Chem.* **118** (2006) 212; *Angew. Chem. Int. Ed.* **45** (2006) 206.
98. 810. M. Spiro, W. P. Griffith, *Textile Chem. Colorist* **29** (1997) 12.
99. 811. J. Dannacher, W. Schlenker, *Textilveredelung* **25** (1990) 205.
100. 812. J. Dannacher, W. Schlenker, *Textile Chem. Colorist* **28** (1996) 24.
101. 813. M. E. Burns, G. S. Miracle, A. D. Willey in M. S. Showell (ed.): *Powdered Detergents*, Marcel Dekker, New York–Basel–Hong Kong, 1998, p. 165.
102. 814. D. M. Davies, M. E. Deary, *J. Chem. Soc., Perkin Trans. 2* **10** (1991) 1549.
103. 815. N. J. Milne, *J. Surf. Det.* **1** (1998) 253.
104. 816. A. P. James, I. S. Mackirdy, *Chem. Ind.* (1990) 641.
105. 817. "Enzymes in Household Detergents", in W. Aehle (ed.): *Enzymes in Industry, Ullmann's Encyclopedia of Industrial Chemistry*, Weinheim: Wiley-VCH, Weinheim 2004, Chap. 5.2.1. p. 155.
106. 818. E. Gormsen, E. Marcussen, T. Damhus in M. S. Showell (ed.): *Powdered Detergents*, Marcel Dekker, New York–Basel–Hong Kong 1998, p. 137.
107. 819. C. Bessler, J. Schmitt, K.-H. Maurer, R. D. Schmid, *Protein Sci.* **12** (2003) 2141.
108. 820. K.-H. Maurer, *Curr. Opin. Biotechnol.* **15** (2004) 330.
109. 821. A. J. O'Lenick, Jr., *J. Surf. Det.* **2** (1999) 553.
110. 822. E. P. Grosselink in M. S. Showell (ed.): *Powdered Detergents*, Marcel Dekker, New York–Basel–Hong Kong, 1998, p. 205.
111. 823. Textile Outlook International, Summary Report: Profile of Nano-Tex, Textiles Intelligence Limited, 2005, p. 139.
112. 824. J. Zhang, P. France, A. Radomyselskiy, S. Datta, J. Zhao, W. van Ooij, *J. Appl. Polym. Sci.* **88** (2003) 1473.
113. 825. M. G. McCord, Y. J. Hwang, Y. Qiu, L. K. Hughes, M. A. Bourham, *J. Appl. Polym. Sci.* **88** (2003) 2038.
114. 826. A. Calvimontes, V. Dutschk, B. Breitzke, P. Offermann, B. Voit, *Tenside Surf. Det.* **42** (2005) 17.
115. 827. A. Calvimontes, V. Dutschk, H. Koch, B. Voit, *Tenside Surf. Det.* **42** (2005) 210.
116. 828. T. Foerster, W. von Rybinski in B. P. Binks (ed.): *Modern Aspects of Emulsion Science, The Royal Society of Chemistry*, Cambridge, UK, 1998, pp. 418, 448.
117. 829. C. Solans, J. Garcia Dominguez, S. E. Friberg, *J. Disp. Sci. Technol.* **6** (1985) 523.
118. 830. K. Braeuer, H. Fehr, R. Puchta, *Tenside Det.* **17** (1980) 281.
119. E. Smulders, P. Krings, H. Verbeek, Proceedings of the Annual Meeting of the German Society for Fat Research (DGF) 1991.
120. D. Nickel, W. von Rybinski, E. M. Kutschmann, C. Stubenrauch, G. H. Findenegg, *Proc. 4th World Surfactant Congress* **2** (1996) 371–382.
121. H. G. Smolka, M. J. Schwuger, *Tenside Deterg.* **14** (1977) 222.
122. H. G. Smolka, M. J. Schwuger, *Colloid Polym Sci.* **256** (1978) 270.
123. M. J. Schwuger, H. G. Smolka, *Tenside Deterg.* **16** (1979) 233.
124. G. Jacobi, *Angew. Makromol. Chem.* **123/124** (1984) 119–145.
125. R. Puchta, W. Grünewälder: *Textilpflege*, Waschen und Chemischreinigen, Verlag Schiele und Schön, Berlin 1973, pp. 47–100.
126. M. J. Schwuger, *Chem. Ing. Tech.* **42** (1970) 433 –438.
127. E. Smulders, P. Krings, *China Surfactant Deterg. & Cosmet.* **1** (1993) 18–24.
128. A. M. Rymond, *Proceedings of the 4th World Surfactant Congress* **1** (1996) 21–32.
129. G. Jakobi in H. Stache (ed.): *Tensid-Taschenbuch*, 2nd ed., Hanser Verlag, München-Wien 1981, pp. 253–337.
130. M. F. Cox, T. P. Matson, J. L. Berna, A. Moreno, S. Kawakami, M. Suzuki, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 330.
131. H. Andree, P. Krings, *Chem. Ztg.* **99** (1975) 168–174.

132. J. M. Quack, M. Trautmann, *Tenside Deterg.* **22** (1985) no. 6, 281–289.
133. F. Asinger in *Chemie und Technologie der Paraffinkohlenwasserstoffe*, Akademie Verlag, Berlin 1959.
134. H. G. Hauthal, B. Makowka, M. Pieroth, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 873–882.
135. H. G. Hauthal, in H. W. Stache (ed.): *Anionic Surfactants: Organic Chemistry* Surfactant Sci. Ser. **56** (1996) 143–221.
136. H. Andree, P. Krings in Henkel & Cie GmbH (ed.): *Waschmittelchemie*, Hüthig Verlag, Heidelberg 1976, pp. 73–90.
137. F. Schambil, J. Schwuger, Proceedings of the 2nd World Surfactants Congress (CESIO), 1988, 505–524.
138. T. Satsuki, *INFORM* **3** (1992) 109–110.
139. E. Smulders, P. Krings, *Chem. Ind.* **6** (1990) 160.
140. A. J. Kaufmann, R. J. Ruebusch, Proceedings of the World Conference on Oleochemicals, Kuala Lumpur, 1990, 10.
141. R. A. Peters, Proceedings of the World Conference on Oleochemicals, Kuala Lumpur, 1990, 181.
142. H. J. Richtler, J. Knaut, 2nd World Surfactant Congress (CESIO), Paris, 1988.
143. G. Czichocki, H. Brämer, I. Ohme, *Z. Chem.* **20** (1980) 90–93.
144. G. Czichocki, A. Greiner, B. Koch, *Tenside Deterg.* **21** (1984) no. 2, 62–66.
145. E. Smulders, Proceedings of the International Conference of the European Chemical Marketing Research Association (ECMRA) (1984) 305–348.
146. K. Hill, W. von Rybinski, G. Stoll (eds.): *Alkyl Polyglycosides — Technology*, Properties, and Applications. Wiley-VCH, Weinheim, Germany 1997.
147. K. Bräuer, H. Fehr, R. Puchta, *Tenside Deterg.* **17** (1980) 281–287.
148. W. M. Linfield in E. Jungermann (ed.): *Cationic Surfactants* Surfactant Sci. Ser. **4** (1970) 49.
149. R. Puchta, P. Krings, P. Sandkühler, *Tenside Surf. Det.* **30** (1993) 186–191.
150. S. T. Giolando, R. A. Rapaport, R. J. Larson, T. W. Federle, *Chemosphere* **30** (1995) 1067–1083.
151. F. Lang, H. Berenbold, *Seifen, Oele, Fette, Wachse* **117** (1990) 690–694.
152. H. Berenbold, *Tenside, Surfactants, Deterg.* **27** (1990) 34–40.
153. G. C. Schweiker, *J. Am. Oil Chem. Soc.* **58** (1981) 58–61.
154. G. Jakobi, M. J. Schwuger in Henkel & Cie GmbH (ed.): *Waschmittelchemie*, Hüthig Verlag Heidelberg 1976, pp. 91–120.
155. Henkel, DE-OS 2 412 837, 1974.
156. M. Schwuger, E. Smulders: "Inorganic Builders" in G. Cutler, E. Kissa (eds.): *Detergency: Theory and Technology*, Surfactant Sci. Ser. **20** (1987) 371–439.
157. Henkel, DE-OS 2 412 838, 1975.
158. G. Jakobi, *Angew. Makromol. Chem.* **123/124** (1984) 119.
159. H. Krüssmann, P. Vogel, H. Hloch, H. Carlhoff, *Seifen Oele Fette Wachse* **105** (1979) 3–6.
160. M. Ettlinger, H. Ferch, *Seifen Oele Fette Wachse* **105** (1979) 131–135, 160.
161. W. E. Adam, K. Neumann, J. P. Ploumen, *Fette Seifen Anstrichm.* **81** (1979) 445–449.
162. H. Nüsslein, K. Schumann, M. J. Schwuger, *Ber. Bunsenges. Phys. Chem.* **83** (1979) 1229–1238.
163. C. P. Kurzendörfer, M. J. Schwuger, H. G. Smolka, *Tenside Deterg.* **16** (1979) 123–129.
164. P. Berth, *Tenside Deterg.* **14** (1978) 176–180.
165. H. Upadek, P. Krings, E. Smulders, CHIMICAoggi, January–February 1990, 61–67.
166. H. Upadek, P. Krings, *Seifen Oele Fette Wachse* **117** (1991) 554–558.
167. P. Zini (ed.): *Polymeric Additives for High Reforming Detergents*, Technomic, Lancaster 1995, pp. 179–212.
168. P. Zini, *J. of Chem. Technol. and Biotechnol.* **50** (1991) 351–359.
169. H. Upadek, B. Kottwitz, B. Schreck, *Tenside Surf. Det.* **33** (1996) 385–392.
170. H. P. Rieck, *Seifen Oele Fette Wachse* **122** (1996) 376–391.
171. Unilever, EP 0 384 070, 1989 (G. T. Brown, T. J. Osinga, M. J. Parkington, A. T. Steel).
172. Joseph Crosfield & Sons, EP 0 565 364, 1993 (A. Araya).
173. P. Kuzel, *Seifen Oele Fette Wachse* **105** (1979) 423–424.
174. P. Kleinschmitt, B. Bertsch-Frank, Th. Lehmann, P. Panster, Ullmann's Encyclopedia of Industrial Chemistry, Vol. A 19 (1991) 177–197.
175. V. Croud: "Oxygen Bleaches" in U. Zoller, G. Broze (eds.): *Properties*, Surfactant Sci. Ser. **82** (1999) 597–617.
176. G. Whalley, *Manuf. Chem.* April 1998, 31–33.

177. G. Reinhardt: " New Bleach Systems" Proceedings of the 4th World Conference on Detergents, AOCS Press Champaign (IL) 1999, 195–203.
178. P. Kuzel in J. Falbe (ed.): *Surfactants in Consumer Products*, Springer Verlag, Heidelberg 1987, p. 267.
179. J. A. Church: "Hypochlorite Bleach" in U. Zoller, G. Broze (eds.): *Handbook of Detergents: Properties*, Surfactant Sci. Ser. **82** (1999) 619–629.
180. M. Julémont: "Application of Hypochlorite" in U. Zoller, G. Broze (eds.): *Handbook of Detergents: Properties*, Surfactant Sci. Ser. **82** (1999) 631–637.
181. W. L. Smith, Proceedings of the 4th World Conference on Detergents: Strategies for the 21st Century, AOCS Press Champaign (IL) 1999, 137–141.
182. H. Bloching, *Sonderheft „Haushaltswaschmittel"*, Ciba-Geigy, 1975, p. 17.
183. D. Coons in J. Falbe (ed.): *Surfactants in Consumer Products*, Springer Verlag, Heidelberg 1987, p. 274.
184. O. Annen, H. Ulshöfer, *Melliand Textilber.* (2000) 756–759
185. G. Reinhardt: "New Bleach Systems", Proc. World Conf. Deterg.: Strategies 21 Century, 4th, 1999, 195–203.
186. Noury van der Lande, DE 1 162 967, 1960.
187. J. Mathews, *J. Soc. of Dyers Colour.* **115** (1999) 154–155.
188. J. Mathews, *Book. Pap.-Int. Conf. Exhib., AATC*, 1997, 462–470.
189. V. B. Croud, S. J. Tompsett, Proceedings of the 4th World Surfactants Congress (CESIO), 1996, 257–267.
190. M. A. Chadwick, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 1227–1236.
191. V. B. Croud, I. M. George, *HAPPI Household Pers. Prod. Ind.* **34** (1997) no. 1, 82–92.
192. I. M. George, V. B. Croud, *Biocides Today* (1998) 25–27.
193. Procter & Gamble, US 4 412 934, 1983.
194. A. Gilbert, *Deterg. Age* **4** (1967) July, 30–33.
195. C. Render, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 1217–1226.
196. R. Hage et al., *Nature (London)* **369** (1994) 637–639.
197. A. Crutzen, M. L. Douglass: "Detergent Enzymes: A challenge!" in U. Zoller, G. Broze (eds.): *Handbook of Detergents: Properties*, Surfactant Sci. Ser. **82**, 1999, pp. 639–690.
198. J. H. van Ee: "Historical Overview" in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency*, Surfactant Sci. Ser. **69**, 1997, pp. 1–10.
199. Novo Nordisk A/S, Product Sheet, Enzyme Business B 345e-GB, 1998.
200. J. H. Houston: "Detergent Enzymes: Market" in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency*, Surfactant Sci. Ser. **69**, 1997, pp. 11–21.
201. H. Upadek, B. Kottwitz: "Application of Amylases in Detergents" in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency*, Surfactant Sci. Ser. **69**, 1997, pp. 203–212.
202. H. Andree, W. R. Müller, R. D. Schmid, *J. Appl. Biochem.* **2** (1980) 218.
203. B. Kottwitz, H. Upadek, *Seifen Oele Fette Wachse* **120** (1994) 794–799.
204. G. K. Greminger, Jr. *Soap Cosmet. Chem. Spec.* **2** (1978) Nov., 28–31, 38.
205. G. K. Greminger, *J. Am. Oil Chem. Soc.* **55** (1978) 122–126.
206. F. Lang, *Chim. Oggi* **9** (1998) 14.
207. E. Kissa, in *Detergency: Theory and Technology*, Surfactant Sci. Ser. **20**, (1986) 333.
208. E. P. Grosselink, in *Powder Detergents* Surfactant Sci. Ser. **71**, (1998) 215.
209. Procter & Gamble, DE-OS 2 338 464, 1974.
210. Dow Corning, DE-OS 2 402 955, 1974.
211. G. Rossmy, *Fette Seifen Anstrichm.* **71** (1969) 56.
212. J. Blajely, S. Stassen, A. Bouadjaj, E. Mallen, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 1157–1168.
213. R. Höfer, F. Jost, M. J. Schwuger, R. Scharf, J. Geke, J. Kresse, H. Lingmann, R. Veitenhansel, W. Erwied: "Foams and Foam Control" in Ullmann's Encyclopedia of Industrial Chemistry, Vol. A 11 (1988) 465–490.
214. A. E. Siegrist, C. Eckhardt, J. Kaschig, E. Schmidt: "Optical Brighteners" Ullmann's Encyclopedia of Industrial Chemistry, Vol. A 18 (1991) 153–173.
215. J. Kaschig, C. Puebla, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 1147–1156.
216. E. Smulders, M. Osset Hernandez, Comun. Jorn. Com. Esp. Deterg., 28th, (1998) 13–32.
217. E. Smulders, P. Krings, H. Verbeek, *Tenside, Surfactants, Deterg.* **34** (1997) 386–392.

218. Th. Müller-Kirschbaum, P. Sandkühler, G. Wagner, *Naturwissenschaften im Unterricht Chemie* **12** (2001) no. 3, 11–15.
219. Textile Washing Products, Euromonitor International, London, Chicago, Singapore, September 2000.
220. Th. Müller-Kirschbaum, E. Smulders: "Facing future's Challenges — European Laundry Products on the Treshold of the 21st Century", in *Proc. World Conf. Deterg.: Strategies 21st Century*, 4th 1999, 93–106.
221. R. B. McConnell, *Am. Dyest. Rep.* **67** (1978) July, 31–34.
222. H. W. Bücking, K. Lötzsch, G. Täuber, *Tenside Deterg.* **16** (1979) 1–10.
223. L. Hughes, J. M. Leiby, M. L. Deviney, *Soap Cosmet. Chem. Spec.* **2** (1975) Oct., 56–62.
224. W. Bechstedt, W. A. Roland, *Reiniger Wäscher* **33** (1980) no. 6, 21–29, no. 7, 18–24.
225. E. Smulders, Th. Müller-Kirschbaum, P. Sandkühler, *Chemie-Technik* **1** (1997) 96–100.
226. H. Upadek, P. Krings, *Seifen, Oele, Fette, Wachse* **117** (1991) 554–558.
227. Henkel KGaA, EP 0 486 592 B1, 1990.
228. Henkel, EP 595 946, 1997.
229. Henkel, EP 828 818, 1999.
230. Henkel, WO 98/12299, 1998.
231. F. Schambil, M. Böcker, *Tenside Surf. Det.* **37** (2000) 1, 48–50.
232. F. Schambil, M. Böcker, G. Blasey, *Chim. Oggi*, March/April 2000 48–50.
233. G. Whalley, *Manuf. Chem.* October 1998, 41–45.
234. A. Syldath, K. H. Schmid, D. Kischel, Proceedings of the 5th World Surfactants Congress (CESIO), 2000, 1141–1146.
235. E. Smulders, P. Krings, Proceedings of the 2nd World Surfactants Congress (CESIO), 1988.
236. D. Nickel, Proceedings of the 5th World Surfactant Congress (CESIO), 2000, 1106–1111.
237. B. Jewett, *Soap & Cosmetics* **77** (2000) no. 9, 32–34.
238. H. J. Lehmann, *Chem. Unserer Zeit* **7** (1973) 82–89.
239. E. P. Frieser, *Österr. Chem. Z.* **76** (1975) no. 9, 4–8; **76** (1975) no. 10, 1–6; **76** (1975) no. 11, 5–7; **76** (1975) no. 12, 5–9.
240. K. Henning, *Chem. Lab. Betr.* **27** (1976) 46–50, 81–86.
241. G. Gawalek: *Tenside*, Akademie Verlag, Berlin 1975.
242. Henkel: *Waschmittelchemie*, Hüthig Verlag, Heidelberg 1976.
243. P. Berth, G. Jakobi, E. Schmadel, J. Schwuger, C. Krauch, *Angew. Chem.* **87** (1975) 115–123; *Angew. Chem. Int. Ed. Engl.* **14** (1975) 94.
244. R. Puchta, P. Krings, A. Wilsch-Irrgang, *Hauswirtschaft und Wissenschaft* **42** (1994) 70–75.
245. J. Waters, K. S. Lee, V. Perchard, M. Flanagan, P. Clarke, *Tenside Surf. Det.* **37** (2000) 3, 161–171.
246. W. Pietsch: *Size Enlargement by Agglomeration*, John Wiley & Sons, New York 1990.
247. D. Jung: *Ullmann*, vol. **A8**, 1985, p. 385 ff.
248. K. Masters: *Spray Drying Handbook*, Fifth Edition, Longman Scientific & Technical, New York 1991.
249. Henkel KGaA, US-Patent 4 741 803, 1985.
250. G. F. Moretti, I. Adami: Lecture presented at 6e Giornate CID, Roma 1995.
251. J. Höhmann: *Wissensbasierte Systeme zur Prozessführung verfahrenstechnischer Anlagen am Beispiel der Sprühtrocknung*, Verlag Shaker, Aachen 1993.
252. Henkel KGaA, EP 0 550 508 B1, 1991.
253. Henkel KGaA, WO 95/21010, 1995.
254. Henkel KGaA, EP 0 625 923, 1992.
255. W. Rähse: Drying with Superheated Steam in a Spray-Drying Tower, Henkel-Referate 35, Düsseldorf 1998.
256. W. H. de Groot, I. Adami, G. F. Moretti: *The Manufacture of Modern Detergent Powders*, H. de Groot Academic Publisher, Wassenaar 1995.
257. Procter & Gamble, EP 0 643 130 A1, 1993.
258. Henkel KGaA, P 38 03 966.4, 1992.
259. Henkel KGaA, P 38 12 530.7, 1988.
260. Procter & Gamble, WO 95/32276, 1994.
261. Procter & Gamble, EP 0 663 439 A1, 1994.
262. Procter & Gamble, WO 95/00630, 1995.
263. O. T. Kraph, A. Kraglund, *Chem. Eng. (London)* **367** (1981) 149–153.
264. G. F. Moretti, I. Adami: "Drying and Agglomeration Process for Traditional and Concentrated Detergent Powders", Ballestra S.p.A., Milano 1994.
265. Henkel KGaA, DE 19844522.9, 1998.
266. Henkel KGaA, DE 40 07601 A1, 1990.
267. Henkel KGaA, DE 197 10 254.9, 1997.
268. Procter & Gamble, EP 846 745, EP 846 755, EP 846 756, 1996.
269. Unilever, EP 481 793, 1990.
270. Unilever, EP 711 827, 1994.
271. A. Behler, A. Syldath, Proceedings of the 5th World Surfactants Congress, 29.5–2.6.2000, Firenze, Vol. 1, p. 382.
272. Korsch Pressen GmbH, Berlin, Germany.

273. L. Spitz, *Soaps and Detergents*, AOCS Press, San Diego 1996.
274. Henkel KGaA, DE 195 46 735, 1995.
275. Henkel KGaA, EP 319 819, 1988.
276. Henkel KGaA, DE 44 25 968 A1, 1994.
277. Vomm Inc., Rozzano/Milan, Italy.
278. Bayer AG, EP 0 163 836, 1984.
279. Glatt GmbH, Weimar, Germany.
280. G. Dieckelmann, H. J. Heinz: *The Basis of Industrial Oleochemistry*, Peter Pomp GmbH, Essen 1988.
281. K. Winnacker, L. Küchler, H. Harnisch, R. Steiner (eds.): *Chemische Technologie*, 4. Auflage, **Bd. 7**, Carl Hanser Verlag München 1986.
282. M. S. Showell (ed.): *Powdered Detergents*, vol. **71**, Surfactants Science Series, 1998.
283. Hoechst, DE-A-34 17 649, 1984.
284. Hoechst, DE 196 01 063 A1, 1995.
285. Hoechst, DE-A-41 42 711, 1991.
286. Henkel KGaA, DE 444 63 63, 1994.
287. Henkel KGaA, DE 195 33 790, 1995.
288. Unilever/Crosfield, WO 94/26662, 1993.
289. Unilever/Crosfield, EP 565 364, 1992.
290. Unilever, WO 94/28098, 1993.
291. K. Winnacker, L. Küchler, H. Harnisch, R. Steiner (eds.): *Chemische Technologie*, 4. Auflage, **Bd. 2**, Carl Hanser Verlag München 1986, p. 584 ff.
292. W. Rähse, F.-J. Carduck, Microfiltration of Fermenter Broths, *Chem.-Ing.-Tech.* **57** (1985) 747.
293. Biozym GmbH, Kundl, Austria.
294. G. Sorbe, *Seifen Oele Fette Wachse* **105** (1979) 251–253.
295. W. M. Linfield: *Anionic Surfactants*, vol. **7** (Parts I and II), Marcel Dekker, New York-Basel 1976.
296. P. Berth, M. J. Schwuger, *Tenside Deterg.* **16** (1979) 175–184.
297. P. Berth, J. Heidrich, G. Jakobi, *Tenside Deterg.* **17** (1980) 228–235.
298. B. Ziolkowsky, *Seifen Oele Fette Wachse* **12** (1981) 333–336, 339–343.
299. Th. Kunzmann, *Seifen Oele Fette Wachse* **98** (1972) 179–181.
300. M. J. Schwuger, H. G. Smolka, *Colloid Polym. Sci.* **254** (1976) 1062.
301. M. J. Schwuger, H. G. Smolka, C. P. Kurzendörfer, *Tenside Deterg.* **13** (1976) 305.
302. C. P. Kurzendörfer, M. J. Schwuger, H. G. Smolka, *Tenside Deterg.* **16** (1979) 123–129.
303. H. G. Smolka, M. J. Schwuger, *Tenside Deterg.* **14** (1977) 222.
304. H. G. Smolka, M. J. Schwuger, *Colloid Polymer. Sci.* **256** (1978) 270–277.
305. M. J. Schwuger, H. G. Smolka, *ACS Symp. Ser.* **40** (1971) 696–707.
306. M. J. Schwuger, H. G. Smolka, *Tenside Deterg.* **16** (1979) 233–239.
307. W. Kling, *Angew. Chem.* **65** (1953) 201–212.
308. W. Kling, *Parfüm. Kosmet.* **45** (1964) 1–5, 29–31.
309. W. Hagge, *CID-Kongreß 1968*, vol. **I**, pp. 1–19.
310. G. F. Longman: *Analysis of Detergents and Detergent Products*, J. Wiley & Sons, London 1975.
311. N. Schönfeld: *Grenzflächenaktive Ethylenoxid-Addukte*, Wissenschaftl. Verlagsges., Stuttgart 1976, pp. 991–1045.
312. J. Wiesner, L. Wiesnerova, *J. Chromatogr.* **114** (1975) 411–412.
313. E. Kunker, *Tenside Deterg.* **18** (1981) 301–305.
314. E. Heinerth, *Seifen Oele Fette Wachse* **5** (1964) 105–110.
315. R. Wickbold: *Die Analytik der Tenside*, Chemische Werke Hüls AG, Marl 1976.
316. G. Graffmann, W. Hörig, P. Sladek, *Tenside Deterg.* **14** (1977) 194–197.
317. M. Teupel, *CID-Kongreß 1960*, vol. **C/III**, pp. 177 –183.
318. G. Schwarz, *Seifen Oele Fette Wachse* **22** (1961) 715–717.
319. H. König, *Fresenius' Z. Anal. Chem.* **293** (1978) 295–300.
320. M. J. Rosen, H. A. Goldsmith: *Sytematic Analysis of Surface-Active Agents*, 2nd ed., J. Wiley & Sons, London 1972.
321. Schweizerische Gesellschaft f. Analytische und Angewandte Chemie: *Seifen und Waschmittel*, Verlag H. Huber, Bonn-Stuttgart 1955, 1963.
322. D. Hummel: *Analyse der Tenside*, Hanser Verlag, München 1962.
323. B. M. Milwidsky, *Practical Detergent Analyses*, Mac Nair-Dorland Comp., New York 1970.
324. British Standard 3762 : 1964, "Methods of Sampling and Testing Detergents."
325. G. F. Longman, *Talanta* **22** (1975) 621–636.
326. *Jahrbuch für den Praktiker aus der Öl-*, Fett-, Seifen- und Waschmittel-, Wachs- und sonstigen chemischtechnischen Industrie 1981, Verlag für chemische Industrie H. Ziolkowsky KG, Augsburg.
327. J. Puschmann, *Angew. Makromol. Chem.* **47** (1975) 29–41.
328. H. Specker, *Angew. Chem.* **80** (1968) 297–304; *Angew. Chem. Int. Ed. Engl.* **7** (1968) 252.

329. G. Tölg, *Naturwissenschaften* **63** (1976) 99–110.
330. D. C. Cullum, P. Platt, *Crit. Rep. Appl. Chem.* **32** (1991) (Recent Dev. Anal. Surfactants) 5–74.
331. G. T. Battaglini, Soaps Deterg. Theor. Pract. Rev., [Soaps Deterg. Technol. Today Conf.] (1996), Meeting Date 1994, 330–380.
332. B. M. Milwidsky: *Practical Detergent Analyses*, Mac Nair-Dorland Comp., New York 1970.
333. DGF-Einheitsmethoden, Wissenschaftl. Verlagsges. Stuttgart 1950–1984.
334. Official and Tentative Methods of The American Oil Chemists Society, 508 South Sixth Street, Champaign, Illinois 61 820.
335. ISO-TC 91.
336. G. Graffmann, H. Domels, W. Hörig, *Fette Seifen Anstrichm.* **77** (1975) 364–365.
337. DIN Deutsches Institut für Normung e.V., Beuth-Vertrieb GmbH, Berlin-Köln.
338. Jander-Blasius: *Lehrbuch der analytischen und präparativen anorganischen Chemie*, Hirzel Verlag, Stuttgart 1973.
339. Anorganikum, VEB Deutscher Verlag der Wissenschaften, Berlin (GDR) 1967.
340. J. Jaeger, K. Sorensen, S. P. Wolff, *J. Biochem. Biophys. Methods* **29** (1994) no. 1, 77–81.
341. Fresenius-Jander: *Handbuch der Analytischen Chemie*, Springer Verlag, Berlin 1953.
342. Deutsche Einheitsverfahren zur Wasseruntersuchung, Verlag Chemie, Weinheim 1960.
343. H. Puderbach, *J. Am. Oil Chem. Soc.* **55** (1978) 156–162.
344. I. Kawase, A. Nakae, K. Tsuji, *Anal. Chim. Acta* **131** (1981) 213–222.
345. D. Millers, *Comun. Jorn. Esp. Deterg.* **21** (1990) 99–112.
346. H. Bauer, *Seifen Oele Fette Wachse* **121** (1995) no. 3, 168–170.
347. A. Kawauchi, M. Ishida, *J. Am. Oil Chem. Soc.* **73** (1996) no. 1, 131–135.
348. F. Feigl: *Spot Tests in Organic Analysis*, 7th ed., Elsevier, Amsterdam 1966.
349. F. Cramer: *Papierchromatographie*. 5th ed., Verlag Chemie, Weinheim, Germany 1962.
350. E. Stahl: *Dünnschichtchromatographie*. 2nd ed., Springer Verlag, Berlin 1967.
351. E. Heinerth, J. Pollerberg, *Fette Seifen Anstrichm.* **61** (1959) 376–377.
352. R. Pribil, V. Vesely, *Chem. Analyst* **56** (1967) 51–53.
353. E. Heinerth, *Fette Seifen Anstrichm.* **70** (1968) 495–498.
354. J. Drewry, *Analyst (London)* **88** (1963) 225–231.
355. D. Hummel: *Kunststoff-, Lack- und Gummianalyse,* vol. I, Hanser Verlag, München 1958.
356. E. Knappe, I. Rhodewald, *Fresenius' Z. Anal. Chem.* **223** (1966) 174–181.
357. K. Figge, *Fette Seifen Anstrichm.* **70** (1968) 680–687.
358. G. Lehmann, M. Becker-Klose, *Tenside Deterg.* **13** (1976) 7–8.
359. H. Bloching, W. Holtmann, M. Otten, *Seifen Oele Fette Wachse* **105** (1979) 33–38, 82–83.
360. S. R. Epton, *Nature (London)* **160** (1947) 795–796.
361. R. Matissek, E. Hieke, W. Baltes, *Fresenius' Z. Anal. Chem.* **300** (1980) 403–406.
362. H. Brüschweiler, V. Sieber, H. Weishaupt, *Tenside Deterg.* **17** (1980) 126–129.
363. M. Däuble, *Tenside Deterg.* **18** (1981) 7–12.
364. J. M. Rosen, *Anal. Chem.* **27** (1955) 787–790.
365. K. Bürger, *Fresenius' Z. Anal. Chem.* **196** (1963) 251–259.
366. *Erfahrungsaustausch der Seifen-*, Wasch- und Reinigungsmittelindustrie, vol. **1, pp. 9–12**, 1944;
Vagda-Kalender, 5th ed., 1943, p. 87.
367. A. Hintermaier, *Fette Seifen* 1944, Sept. 1, 367–368.
368. P. Friese, *Fresenius' Z. Anal. Chem.* **303** (1980) 279–288.
369. J. F. Kennedy, E. H. M. Melo, *Cellul. Sources Exploit.* (1990) 371–376.
370. J. F. Kennedy, E. H. M. Melo, *Eur. Pol. Paint Colour J.* **182** (1992) no. 4312, 430–431, 436.
371. J. F. Kennedy, E. H. M. Melo, *Eur. Chem. Paint Colour J.* **181** (1991) no. 4289, 478, 480, 482.
372. B. Gruening, D. Kaeseborn, H. Leidreiter, *Parfuem. Kosmet.* **77** (1996) no. 4, 244–248.
373. R. Spilker, B. Menzebach, U. Schneider, I. Venn, *Tenside, Surfactants, Deterg.* **33** (1996) no. 1, 21–25.
374. H. Waldhoff, J. Scherler, M. Schmitt, World Surfactants Congr., 4th (CESIO) 1, (1996) 507–518.
375. W. Leonhardt, R. Peldszus, H. Wegert, *Seife Oele Fette Wachse* **113** (1987) no. 15, 511–513.
376. H. P. Kaufmann: *Analyse der Fette und Fettprodukte*, Springer Verlag, Berlin 1958.
377. R. Wickbold, *Tenside Deterg.* **9** (1972) 173–178.
378. H. Hellmann, *Fresenius' Z. Anal. Chem.* **300** (1980) 44–47.

379. B. Weibull: *3. Int. Kongreß für grenzflächenaktive Stoffe*, Köln 1960, vol. **3**, Universitätsdruckerei, Mainz 1961, p. 121.
380. B. Milwidsky, *Tenside Deterg.* **22** (1985) no. 3, 136–138.
381. W. Riemann, H. Walton: *Ion Exchange in Analytical Chemistry*, Pergamon Press, Oxford 1970.
382. P. Voogt, *Recl. Trav. Chim. Pays-Bas* **77** (1958) 889–901.
383. R. Wickbold, *Seifen Oele Fette Wachse* **86** (1960) 79–82.
384. N. Blumer, *Schweiz. Arch.* **29** (1963) 171–180.
385. M. E. Ginn, C. C. Church, *Anal. Chem.* **31** (1959) 551–555.
386. K. Bey, *Fette Seifen Anstrichm.* **67** (1965) 25–30.
387. P. Voogt in H. A. Boeckenoogen (ed.): *Analysis and Characterization of Oils*, Fats, and Fat Products Interscience Publ. Inc., London 1964.
388. R. Wickbold, *Tenside Deterg.* **13** (1976) 177–180.
389. R. Wickbold, *Tenside Deterg.* **13** (1976) 181–187.
390. G. Schwarz, *Fette Seifen Anstrichm.* **71** (1969) 223–226.
391. E. Heinerth, *Fette Seifen Anstrichm.* **63** (1961) 181–183.
392. R. Bock: *Aufschlußmethoden der anorg. und organischen Chemie*, Verlag Chemie, Weinheim, Germany 1972.
393. Verein Deutscher Eisenhüttenleute: *Handbuch für das Eisenhüttenlaboratorium,* vol. 1, Stahleisen-Verlag, Düsseldorf 1960.
394. *Fette Seifen Anstrichm.* **79** (1977) 203.
395. G. Staats, H. Brück, *Fresenius' Z. Anal. Chem.* **250** (1970) 289–294.
396. E. Vaeth, E. Griessmayer, *Fresenius' Z. Anal. Chem.* **303** (1980) 268–271.
397. O. G. Koch, A. Koch-Dedic: *Handbuch der Spurenanalyse*, Springer Verlag, Berlin 1974.
398. J. Drozd, *J. Chromatogr.* **113** (1975) 303–356.
399. *Fette Seifen Anstrichm.* **74** (1972) 31.
400. A. Hintermaier, *Angew. Chem.* **60** (1948) 158–159.
401. V. W. Reid, G. F. Longman, E. Heinerth, *Tenside* **4** (1967) 292–304.
402. S. Lee, N. A. Puttnam, *J. Am. Oil Chem. Soc.* **43** (1966) 690.
403. R. Denig, *Fette Seifen Anstrichm.* **76** (1974) 412–416.
404. R. Denig, *Tenside Deterg.* **10** (1973) 59–63.
405. W. Kupfer, K. Künzler, *Fresenius' Z. Anal. Chem.* **267** (1973) 166–169.
406. T. H. Liddicoet, L. H. Smithson, *J. Am. Oil Chem. Soc.* **42** (1965) 1097–1102.
407. D 501–67, ASTM Standards of Soaps and other Detergents, Part 30, 1974.
408. A. J. Sheppard, J. L. Iversen, *J. Chromatogr. Sci.* **13** (1975) 448–452.
409. H. Schlenk, J. L. Gellerman, *Anal. Chem.* **32** (1960) 1412–1414.
410. L. D. Metcalf, A. A. Schmitz, *Anal. Chem.* **33** (1961) 363–364.
411. A. Seher, H. Pardun, M. Arens, *Fette Seifen Anstrichm.* **80** (1978) 58–66.
412. J. D. Knight, R. House, *J. Am. Oil Chem. Soc.* **36** (1959) 195–200.
413. J. Pollerberg, *Fette Seifen Anstrichm.* **67** (1965) 927–929.
414. M. J. Rosen, G. C. Goldfinger, *Anal. Chem.* **28** (1956) 1979–1981.
415. J. Borecky, *Mikrochim. Acta* 1962, 1137–1145.
416. S. Nishi, *Bunseki Kagaku* **14** (1965) 917; *Chem. Abstr.* **64** (1966) 3858 f.
417. R. Wickbold, *Tenside Deterg.* **12** (1975) 25–27.
418. „Sauerstoffverbindungen III", *Houben-Weyl*, **8**
419. P. L. Buldini, J. L. Sharma, D. Ferri, *J. Chromatogr.* **654** (1993) no. 1, 129–134.
420. H. Biltz, W. Biltz: *Ausführung quantitativer Analysen*. Hirzel Verlag, Stuttgart 1965.
421. British Standard 3984: 1966"Sodium Silicates".
422. G. Graffmann, W. Schneider, L. Dinkloh, *Fresenius' Z. Anal. Chem.* **301** (1980) 364–372.
423. J. Longwell, W. D. Maniece, *Analyst (London)* **80** (1955) 167–171.
424. M. Kolthoff, V. A. Stenger: *Volumetric Analysis*, 2nd ed., vols I and II, Interscience Publ., New York 1942, 1947.
425. G. Jander, K. F. Jahr, H. Knoll: *Maßanalyse*, De Gruyter, Berlin 1966.
426. E. Heinerth, *Tenside* **4** (1967) 45–47.
427. S. Ebel, U. Parzefall: *Experimentelle Einführung in die Potentiometrie*, Verlag Chemie, Weinheim, Germany 1975.
428. H. Malmerig, *GIT Fachz. Lab.* **19** (1975) 400–404.
429. S. Ebel, A. Seuring, *Angew. Chem.* **89** (1977) 129–141; *Angew. Chem. Int. Ed. Engl.* **16** (1977) 157.
430. A. Hofer, E. Brosche, R. Heidinger, *Fresenius' Z. Anal. Chem.* **253** (1971) 117–119.
431. J. Koryta, *Anal. Chim. Acta* **61** (1972) 329–411.
432. *Orion Ion- and Gas-sensitive Electrodes*, Orion Research Inc., Cambridge 1973.

433. D. L. Jones, G. J. Moody, J. D. R. Thomas, B. J. Birch, *Analyst (London)* **106** (1981) no. 1266, 974–984.
434. M. Bos, *Anal. Chim. Acta* **135** (1982) no. 2, 249–261.
435. T. Kobayashi, M. Kataoka, T. Kambra, *Talanta* **27** (1980) no. 2, 253–256.
436. H. W. Leuchte, H. D. Kahleyss, J. Faltermann, *Fett Wiss. Technol.* **94** (1992) no. 2, 64–66.
437. M. Tehrani, M. Thomae, *Am. Lab. (Shelton, Conn.)* **23** (1991) no. 19, 8, 10.
438. G. C. Dilley, *Analyst (London)* **105** (1980) no. 1252, 713–719.
439. L. I. Makovetskaya, I. L. Maislina, S. P. Volosovich, *Seife Oele Fette Wachse* **117** (1991) no. 15, 565–571.
440. M. Gerlache, Z. Sentuerk, J. C. Vire, J. M. Kauffmann, *Anal. Chim. Acta* **349** (1997) no. 1–3, 59–65.
441. B. J. Birch, R. N. Cockcroft, *Ion-Sel. Electrode Rev.* **3** (1981) no. 1, 1–41.
442. R. Calapaj, L. Ciraolo, F. Corigliano, D. Di Pasquale, *Analyst (London)* **107** (1982) no. 1273, 403–407.
443. G. Schwarzenbach, H. Flaschka: *Die komplexometrische Titration*, Enke Verlag, Stuttgart 1965.
444. R. Pribil: *Komplexometrie*, VEB Deutscher Verlag für Grundstoffind., Leipzig 1960.
445. G. Graffmann, H. Domels, M. L. Sträter, *Fette Seifen Anstrichm.* **76** (1974) 218–220.
446. Merck AG: *Komplexometrische Bestimmungen mit Titriplex*, 3rd ed., Darmstadt 1966.
447. R. Wickbold: *VI. Int. Kongreß für grenzflächenaktive Stoffe*, Zürich 1972, vol. **I**, Section A, Hanser Verlag, München 1973, p. 373.
448. F. Öhme: *Angewandte Konduktometrie*, Hüthig Verlag, Heidelberg 1962.
449. K. Kiemstedt, W. Pfab, *Fresenius' Z. Anal. Chem.* **213** (1965) 100–107.
450. J. Cross: *Anionic Surfactants — Chemical Analysis*, vol. **8**, Marcel Dekker, New York-Basel 1977.
451. *Fette Seifen Anstrichm.* **73** (1971) 683.
452. H. R. Hoffmann, W. Böer, G. W. G. Schwarz, *Fette Seifen Anstrichm.* **78** (1976) 367–368.
453. "Benzalkonium (Chlorure de)," *Pharmacopée Française* 9th ed., Adapharus, Paris 1972.
454. R. Wickbold, *Seifen Oele Fette Wachse* **85** (1959) 415–416.
455. D. C. White, *Mikrochim. Acta* 1959, 254–259.
456. E. E. Archer, *Analyst (London)* **82** (1957) 208–209.
457. D. Miller, *Comun. Jorn. Com. Esp. Deterg.* **21** (1990) 99–112.
458. H. Bauer, *Seifen Oele Fette Wachse* **121** (1995) no. 3, 168–170.
459. R. Hwang, *J. Test. Eval.* **14** (1986) no. 2, 128–130.
460. L. Baini, G. Bolzoni, G. Carrer, E. Facetti, L. Valtora, E. Guiati, C. Pacchetti, C. Ruffo, O. Cozzoli, L. Sedea, *Riv. Ital. Sostanze Grasse* **69** (1992) no. 12, 615–617.
461. AFNOR, Doc. Française A-124 (7. 7. 70)
462. R. E. Kitson, M. G. Mellon, *Ind. Eng. Chem. Anal. Ed.* **16** (1944) 379–383.
463. H. Bernhardt, (Ausschuß Phosphate und Wasser), *Zeitschr. f. Wasser Abwasser Forsch.* **7** (1974) 143–146.
464. C. Burgi, R. Bollhalder, *Tenside, Surfactants, Deterg.* **31** (1994) no. 2, 86–89.
465. G. van Raay, *Tenside Deterg.* **7** (1970) 125–132.
466. H. Jaag: *Über proteolytische Enzyme*, deren Prüfmethoden und Einsatzmöglichkeiten in der Waschmittelindustrie, Verlag H. Lang & Cie AG, Bern 1968.
467. T. M. Rothgeb, B. D. Goodlander, P. H. Garrison, L. A. Smith, *J. Am. Oil Chem. Soc.* **65** (1988) no. 5, 806–810.
468. Boehringer, Mannheim: *Methoden der enzymatischen Lebensmittelanalytik*, 1980.
469. E. Klein, J. Freiberg, S. Same, M. A. Carroll, *J.-Assoc. Off. Anal. Chem.* **72** (1989) no. 6, 881–882.
470. N. Buschmann, S. Wodarczak, *Tenside Surfactants Deterg.* **32** (1995) no. 4, 336–339.
471. F. Hui, P. Garcia-Camacho, R. Rosset, *Analysis* **23** (1995) no. 2, 58–65.
472. L. K. Wang, *J. Am. Oil Chem. Soc.* **52** (1975) 339–344.
473. K. Toel, H. Fujii, *Anal. Chim. Acta* **90** (1977) 319–322.
474. St. Janeva, R. Borissova-Pangarova, *Talanta* **25** (1977) 279–282.
475. W. Kupfer, *Tenside Deterg.* **12** (1975) 40.
476. G. Krusche, J. Illert, H. Mandery, *Chromatographia* **31** (1991) no. 1–2, 17–20.
477. R. Hermann, C. Th. J. Alkemade: *Flammenphotometrie*, Springer Verlag, Berlin-Göttingen-Heidelberg 1960.
478. B. Welz: *Atom-Absorptions-Spektroskopie*, Verlag Chemie, Weinheim, Germany 1972.
479. H. Bernd, E. Jackwerth, *Spectrochim. Acta Part B* **30 B** (1975) 169–177.
480. G. Volland, G. Kölblin, P. Tschöpel, G. Tölg, *Fresenius' Z. Anal. Chem.* **284** (1977) 1–12.
481. I. Ebdon, A. S. Fisher, *Anal. Spectrosc. Libr. (At. Absorpt. Spectrom.)* **5** (1991) 463–514.

482. E. Jackwerth, J. Lohmar, G. Wittler, *Fresenius' Z. Anal. Chem.* **266** (1973) 1–8.
483. A. Dornemann, H. Kleist, *Fresenius' Z. Anal. Chem.* **291** (1978) 353–359.
484. M. H. Arab-Zavar, A. G. Howard, *Analyst (London)* **105** (1980) no. 1253, 744–750.
485. J. Vilaplana, F. Grimalt, C. Romaguero, J. M. Mascaro, *Contact Dermatitis* **16** (1987) no. 3, 139–141.
486. K. E. Jarvis, J. G. Williams, B. C. H. Gibson, E. Temmerman, C. De Cuyper, *Analyst (London)* **121** (1996) no. 12, 1929–1933.
487. R. O. Müller: *Spektrochemische Analyse mit Röntgenfluoreszenz*, R. Oldenbourg Verlag, München 1967.
488. E. Vaeth, E. Griessmayr, *Fresenius' Z. Anal. Chem.* **303** (1980) 268–271.
489. E. Schenbeck, Ch. Jörrens, *Fresenius' Z. Anal. Chem.* **303** (1980) 257–264.
490. J. Williams, *Analyst (London)* **111** (1986) no. 2, 175–177.
491. H. Krischner: *Einführung in die Röntgenfeinstrukturanalyse*, Vieweg & Sohn, Braunschweig 1974.
492. J. M. Mabis, O. T. Quimby, *Anal. Chem.* **25** (1953) 1814–1818.
493. D. W. Breck: *Zeolite Molecular Sieves: Structure*, Chemistry and Use, J. Wiley & Sons, New York 1974, pp. 347–378.
494. K. Lötzsch, *Seifen Oele Fette Wachse* **105** (1979) 261–267.
495. F. Y. Iskander, *Appl. Radiat. Iso.* **37** (1986) no. 5, 435–437.
496. J. Saint-Pierre, L. Zikovsky, *J. Radioanal. Chem.* **71** (1982) no. 1–2, 19–28.
497. G. Zweig, J. Sherma: *Handbook of Chromatography*, CRC-Press, Cleveland, Ohio 1972.
498. F. Korte: *Methodicum Chimicum,* vol. 1: *Analyticum,* Part 1, G. Thieme Verlag, Stuttgart 1973.
499. K. Heinig, C. Vogt, G. Werner, *J. Capillary Electrophor.* **3** (1996) no. 5, 261–270.
500. M. W. F. Nielen, *J. Chromatogr.* **712 A** (1995) no. 1, 269–284.
501. J. P. Wiley, *J. Chromatogr.* **692 A** (1995) no. 1+2, 267–274.
502. J. M. Hais, K. Macek: *Handbuch der Papierchromatographie*, VEB Fischer Verlag, Jena 1963.
503. H. Grunze, E. Thilo: *Die Papierchromatographie der kondensierten Phosphate*, Sitzungsberichte der Deutschen Akademie der Wissenschaften zu Berlin, Akademie Verlag, Berlin 1955.
504. E. Heinerth, J. Pollerberg, *Fette Seifen Anstrichm.* **61** (1959) 376–377.
505. R. E. Cline, R. M. Fink, *Anal. Chem.* **28** (1956) 47–52.
506. J. C. Touchstone, M. F. Dobbins: *Practice of the Thin Layer Chromatography*, J. Wiley & Sons, New York 1978.
507. F. Geiss: *Die Parameter der DC*, Vieweg und Sohn, Braunschweig 1972.
508. E. Heinerth, *Fette Seifen Anstrichm.* **70** (1968) 495–498.
509. A. Breyer, M. Fischl, E. Setzer, *J. Chromatogr.* **82** (1973) 37–52.
510. K. Bey, *Fette Seifen Anstrichm.* **67** (1965) 217–221.
511. U. Hezel, *Angew. Chem.* **85** (1973) 334–342; *Angew. Chem. Int. Ed. Engl.* **12** (1973) 298.
512. J. Touchstone, T. Murawec, M. Kasparow, W. Wortmann, *J. Chromatogr. Sci.* **10** (1972) 490–493.
513. D. H. Liem, *Cosmet. Toiletries* **92** (1977) 59–72.
514. H. König, *Fresenius' Z. Anal. Chem.* **251** (1970) 359–368.
515. R. Matissek, E. Hieke, U. Baltes, *Fresenius' Z. Anal. Chem.* **300** (1980) 403–406.
516. M. Köhler, B. Chalupka, *Fette Seifen Anstrichm.* **84** (1982) 208–211.
517. R. Matissek, *Tenside Deterg.* **19** (1982) 57–66.
518. J. L. Jasperse, P. H. Steiger, *J. Am. Oil Chem. Soc.* **69** (1992) no. 7, 621–625.
519. S. J. Patterson, E. C. Hunt, K. B. E. Tucker, *J. Proc. Inst. Sewage Purif.* 1966, 190.
520. Ch. Buergi, T. Otz, *Tenside, Surfactants, Deterg.* **32** (1995) no. 1, 22–24.
521. G. Hesse: *Chromatographisches Praktikum*, Akademische Verlagsges., Frankfurt 1968.
522. K. Ehlert, R. Engler, *GIT Fachz. Lab.* **23** (1979) 659–664.
523. P. Quinlin, H. J. Weiser, *J. Am. Oil Chem. Soc.* **35** (1958) 325–327.
524. R. Wickbold, *Fette Seifen Anstrichm.* **74** (1972) 578–579.
525. W. Kupfer, K. Künzler, *Fette Seifen Anstrichm.* **74** (1972) 287–291.
526. M. J. Rosen, *Anal. Chem.* **35** (1963) 2074–2077.
527. R. E. Kaiser: *Chromatographie in der Gasphase*, Bibliographisches Institut, Mannheim 1973.
528. G. Schomburg, R. Dielmann, H. Husmann, F. Weeke, *J. Chromatogr.* **122** (1976) 55–72.
529. H. J. Vonk et al.: *Int. Surfactants Congress*, Moscow 1976, vol. **1, Section A**, pp. 435–449.

530. G. Schomburg: *Gaschromatographie*, Verlag Chemie, Weinheim, Germany 1977.
531. A. Kuksis, *Fette Seifen Anstrichm.* **73** (1971) 130–138.
532. H. Hadorn, K. Zürcher, *Dtsch. Lebensm. Rundsch.* **66** (1970) 77–87.
533. A. F. Prevot, F. X. Mordret, *Rev. Fr. Corps Gras* **23** (1976) 409–423.
534. L. Julin, P. Sunila, Proc.-Scand. Symp. Lipids, 12th (1984), R. Marcuse (ed.), LIPIDFORUM, Goeteborg, 57–64.
535. W. J. Carnes, *Anal. Chem.* **36** (1964) 1197–1200.
536. E. Link, H. M. Hickman, R. A. Morrissette, *J. Am. Oil Chem. Soc.* **36** (1959) 20–23, 300–303.
537. D. G. Anderson, *J. Paint. Technol.* **40** (1968) 549 –557.
538. B. Bonney, N. F. Glennard, A. M. Humphrey, *Chem. Ind. (London)* 1973, 749–751.
539. R. Schicker, *Aerosol Rep.* **13** (1974) 149–163.
540. R. Schubert, L. Ketel, *J. Soc. Cosmet. Chem.* **23** (1972) 115–124.
541. Hoechst, DE-OS 2 161 702, 1971.
542. T. J. Birkel, C. R. Warner, T. Fario, *J. Am. Oil Chem. Soc.* **62** (1979) 931–936.
543. C. Sommer, *Dtsch. Lebensm.-Rundsch.* **89** (1993) no. 4, 108–111.
544. L. Sedea, G. Toninelli, B. Sartorel, *Tenside Deterg.* **22** (1985) no. 5, 269–271.
545. F. Etzweiler, N. Neuner-Jehle, E. Senn, *Seifen Oele Fette Wachse* **106** (1980) no. 15, 419–427.
546. P. Sandra, F. David, *J. High Resolut. Chromatogr.* **13** (1990) no. 6, 414–417.
547. H. Engelhardt: *Hochdruck-Flüssigkeits-Chromatographie — HPLC*, Springer Verlag, Berlin-Göttingen-Heidelberg-New York 1977.
548. J. J. Kirkland: *Modern Practise of Liquid Chromatography*, Wiley-Interscience, New York 1971.
549. H. M. McNair, *Int. Lab.* 1980, no. 5/6, 51–59.
550. L. P. Turner et al., *J. Am. Oil Chem. Soc.* **53** (1976) 691–694.
551. H. Henke, *Tenside Deterg.* **15** (1978) 193–195.
552. R. Murphy, A. C. Selden, M. Fisher, E. A. Fagan, V. S. Chadwick, *J. Chromatogr.* **211** (1981) 160–165.
553. J. H. van Dijk et al., *Tenside Deterg.* **12** (1975) 261–263.
554. L. Sedea, *Tenside, Surfactants, Deterg.* **26** (1989) no. 3, 211–214.
555. J. Weiss, *Tenside, Surfactants, Deterg.* **23** (1986) no. 5, 237–244.
556. J. Truchan, H. T. Rasmussen, N. Omelczenko, B. P. McPherson, *J. Liq. Chromatogr. Relat. Technol.* **19** (1996) no. 11, 1785–1792.
557. D. Kirkpatrick, *J. Chromatogr.* **121** (1976) 153–154.
558. B. P. McPherson, N. Omelczenko, *J. Am. Oil Chem. Soc.* **57** (1980) no. 11, 388–391.
559. G. Micali, P. Curro, G. Calabro, *Analyst (London)* **109** (1984) no. 2, 155–158.
560. Th. Wolf, D. Semionow, *J. Soc. Cosmet. Chem.* **24** (1973) 363–370.
561. A. Nozawa, T. Ohnuma, *J. Chromatogr.* **187** (1980) 261–263.
562. A. Marcomini, F. Filipuzzi, W. Giger, *Chemosphere* **17** (1988) no. 5, 853–863.
563. A. Marcomini, W. Giger, *Anal. Chem.* **59** (1987) no. 13, 1709–1715.
564. H. Yoshimura, T. Sugiyama, T. Nagai, *J. Am. Oil Chem. Soc.* **64** (1987) no. 4, 550–555.
565. A. Marcomini, S. Stelluto, B. Pavoni, *Int. J. Environ. Anal. Chem.* **35** (1989) no. 4, 207–218.
566. H. Waldhoff, J. Scherler, M. Schmitt, World Surfactants Congr., 4th (CESIO) (1996), Volume 1, 507–518.
567. K. Nakamura, Y. Morikawa, I. Matsumoto, *J. Am. Oil Chem. Soc.* **58** (1981) Jan., 72–77.
568. H. Ullner, I. König, C. Sander, Schwenk, *Tenside Deterg.* **17** (1980) 169–170.
569. N. Chen, Jr., V. P. Nero, *J. Chromatogr.* **549** (1991) no. 1–2, 247–256.
570. H.-U. Ehmcke, H. Kelker, K.-H. König, H. Ullner, *Fresenius' Z. Anal. Chem.* **294** (1979) 251–261.
571. D. R. Zornes et al., *Petr. Eng. J.* **18** (1978) 207–218.
572. W. Winkle, M. Köhler, *Chromatographia* **13** (1980) 357–363.
573. K. J. Irgolic, J. E. Hobill, *Spectrochim. Acta, Part B* **42 B** (1987) no. 1–2, 269–273.
574. H. Wagner, G. Metzner, R. R. Wagner, *Muench. Beitr. Abwasser-, Fidch.- Flussbiol. (Umweltrverträglichkeit Wasch-Reinigungsm.)* **44** (1990) 218–230.
575. P. L. Buldini, J. Sharma, D. Ferri, *J. Chromatogr.* **654** (1993) no. 1, 129–134.
576. J. Weiß (ed.): *Ion chromatography*, Wiley-VCH, Weinheim, Germany 1995.
577. J. Weiß, *Tenside, Surfactants, Deterg.* **23** (1986) no. 5, 237–244.
578. E. Vaeth, P. Sladek, K. Kenar, *Fresenius' Z. Anal. Chem.* **329** (1987) no. 5, 584–589.
579. W. D. MacMillan, *HRC CC, J. High Resolut. Chromatogr. Chromatogr. Commun.* **7** (1984) no. 2, 102–103.

580. P. Linares, M. D. Luque de Castro, M. Valcarcel, *J. Chromatogr.* **585** (1991) no. 2, 267–271.
581. H. Felber, K. Hegetschweiler, M. Mueller, R. Odermatt, B. Wampfler, *Chimia* **49** (1995) no. 6, 179–181.
582. T. K. Mahabir-Jagessar, S. A. M. van Stroe-Biezen, *J. Chromatogr. A* **771** (1997) no. 1+2, 1155–1161.
583. E. L. Pretswell, A. R. Morrison, J. S. Park, *Analyst (Cambridge, U.K.)* **118** (1993) no. 10, 1265–1267.
584. T. Wang, S. F. Y. Li, *J. Chromatogr. A* **723** (1996) no. 1, 197–205.
585. J. M. Jordan, R. L. Moese, R. Johnson-Watts, D. E. Burton, *J. Chromatogr. A* **671** (1994) no. 1–2, 445–451.
586. M. Pestemer: *Anleitung zum Messen von Absorptionsspektren im Ultraviolett und Sichtbaren*, Thieme Verlag, Stuttgart 1964.
587. E. J. Stearns: *The Practice of Absorption Spectrophotometry*, Wiley-Interscience, New York 1969.
588. C. N. R. Rao: *Ultraviolet and Visible Spectroscopy*, 3rd ed., Butterworths, London 1975.
589. G. Milazzo et al., *Anal. Chem.* **49** (1977) no. 6, 711–717.
590. W. J. Weber, J. C. Morris, W. Stumm, *Anal. Chem.* **34** (1962) 1844–1845.
591. R. M. Kelley, E. W. Blank, W. E. Thompson, R. Fine, *ASTM Bull.* **TP 90** (1959) 70–73.
592. M. Uchijama, *Water Res.* **11** (1977) 205–207.
593. E. Heinerth, H. G. van Raay, G. Schwarz, *Fette Seifen Anstrichm.* **62** (1960) 825–826.
594. H. Hellmann, *Tenside, Surfactants, Deterg.* **31** (1994) no. 3, 200–206.
595. V. W. Reid, T. Alston, B. W. Young, *Analyst (London)* **80** (1955) 682–689.
596. F. N. Stewart et al., *Anal. Chem.* **31** (1959) no. 11, 1806–1808.
597. H. J. Hediger: „Infrarotspektroskopie," *Methoden der Analyse in der Chemie*, vol. **11** Akademische Verlagsges., Frankfurt 1971.
598. L. J. Bellamy: *Ultrarot-Spektrum und chemische Konstitution*, D. Steinkopff Verlag, Darmstadt 1966.
599. Weissberger: *Technique of Organic Chemistry,* vol. IX: "Chemical Application of Spectroscopy," Interscience Publ. Inc., New York 1956.
600. H. König: *Neuere Methoden zur Analyse von Tensiden*, Springer Verlag, Berlin-Göttingen-Heidelberg-New York 1971.
601. E. Kunkel, *Mikrochim. Acta* 1977, 227–240.
602. H. Günzler, H. Böck: *IR-Spektroskopie*, Verlag Chemie, Weinheim, Germany 1975.
603. G. Socrates: *Infrared Characteristic Group Frequencies*, J. Wiley & Sons, New York 1980.
604. Z. Huren, D. Hajman, D. Vucelic, *J. Serb. Chem. Soc.* **58** (1993) no. 9, 721–724.
605. H. G. van Raay, M. Teupel, *Fette Seifen Anstrichm.* **75** (1973) 572–578.
606. C. D. Frazee, R. O. Crister, *J. Am. Oil Chem. Soc.* **41** (1964) 334–335.
607. H. Hellmann, *Fresenius' Z. Anal. Chem.* **294** (1979) 379–384.
608. H. Hellmann, *Fresenius' Z. Anal Chem.* **293** (1978) 359–363.
609. P. Friese, *Fresenius' Z. Anal Chem.* **305** (1981) 337–346.
610. H. Günther: *NMR-Spektroskopie–Eine Einführung*, Thieme Verlag, Stuttgart 1973.
611. H. König, *Fresenius' Z. Anal. Chem.* **251** (1970) 225–262.
612. A. Mathias, N. Mellor, *Anal. Chem.* **38** (1966) 472–477.
613. M. M. Crutchfield, R. R. Irani, J. T. Yoder, *J. Am. Oil Chem. Soc.* **41** (1964) 129–132.
614. K. H. Choe, S. J. Kwon, *Taehan Hwahakhoe Chi* **27** (1983) no. 6, 419–423.
615. K. H. Choe, S. J. Kwon, *Taehan Hwahakhoe Chi* **25** (1981) no. 1, 54–56.
616. H. Budzikiewicz: *Massenspektrometrie–Eine Einführung*, Verlag Chemie, Weinheim, Germany 1972 (Taschentext 5).
617. M. Köhler, M. Höhn, *Chromatographia* **9** (1976) 611–617.
618. H. A. Boekenoogen: *Oils, Fats and Fat Products,* vol. 2, Interscience Publ., New York 1968.
619. D. Jahr, P. Binnemann, *Fresenius' Z. Anal. Chem.* **298** (1979) 337–348.
620. W. McFadden: *Techniques of Combined Gas Chromatography/Mass Spectrometry*, J. Wiley & Sons, London 1973.
621. R. ter Heide, N. Provatoroff, P. C. Traas, P. J. Valois, *J. Agric. Food Chem.* **23** (1975) 950–957.
622. T. Takeuchi, S. Watanabe, N. Kondo, M. Goto, D. Ishii, *Chromatographia* **25** (1988) no. 6, 523–525.
623. A. L. Rockwood, T. Higuchi, *Tenside, Surfactants, Deterg.* **29** (1992) no. 1, 6–12.
624. P. Rudewicz, B. Muson, *Anal. Chem.* **58** (1986) no. 4, 674–679.
625. E. Stephanou, *Chemosphere* **13** (1984) no. 1, 43–51.

626. M. Köhler, *Chromatographia* **8** (1975) 685–689.
627. F. Ehrenberger, S. Gorbach: *Methoden der organischen Elementar- und Spurenanalyse*, Verlag Chemie, Weinheim, Germany 1973.
628. J. Monar, *Mikrochim. Acta* **2** (1965) 208–250.
629. R. Wickbold, *Angew. Chem.* **69** (1957) 530–533.
630. F. Ehrenberger, *GIT Fachz. Lab.* **21** (1977) 944.
631. R. Kaiser: *Quantitative Bestimmung Organischer Funktioneller Gruppen*, Akademische Verlagsges., Frankfurt 1966.
632. *Technicon-Symposia, Technicon Intern. Congress 1969,* vol. II, White Plains, Mediad 1970.
633. D. P. Lundgren, *Anal. Chem.* **32** (1960) 824–828.
634. *Technicon Industrial Method 3–68 W*, Technicon Corporation Tarrytown/New York 10591.
635. W. Gohla et al., *GIT Fachz. Lab.* **23** (1979) 89–96.
636. H. Pfeiffer, K. H. Lange, *Fette Seifen Anstrichm.* **75** (1973) 438–442.
637. C. Pipinato, L. Sedea, S. Marinaz, G. Toninelli, *Fette, Seifen, Anstrichm.* **82** (1980) no. 7, 283–289.
638. E. Heinerth, *Tenside Deterg.* **7** (1970) 23–24.
639. H. Grossmann, *Seifen Oele Fette Wachse* **101** (1975) 521–524.
640. *Fette Seifen Anstrichm.* **77** (1975) 80.
641. Manufacturer: Atlas Electric Devices Comp., 4114 N. Ravenswood Ave., Chicago.
642. Manufacturer: Heraeus-Original Hanau, D-6450 Hanau.
643. Manufacturer: United States Testing Co. Inc., Hoboken N.J.
644. DIN 53 902, Part 1: Bestimmung des Schäumvermögens, Lochscheibenschlagverfahren.
645. DIN 53 902, Part 2: Bestimmung des Schäumvermögens, modifiziertes Ross-Miles-Verfahren.
646. Manufacturer: wfk-Forschungsinstitut, Campus Fichtenhain 11, D-47807 Krefeld. EMPA Eidgenössische Materialprüfungs- und Versuchsansalt, Lerchenfeldstrasse 5, CH-9014 St. Gallen.
647. H. Stache in: *Tensid-Taschenbuch*, 2nd ed., Hanser Verlag, München-Wien 1981, pp. 460–461.
648. H. Stache in: *Tensid-Taschenbuch*, 2nd ed., Hanser Verlag, München-Wien 1981, pp. 462–466.
649. U. Sommer, H. Milster, *Seifen Oele Fette Wachse* **103** (1977) 295.
650. H. Harder, D. Arends, W. Pochandke, *Seifen Oele Fette Wachse* **102** (1976) 421–426.
651. H. Krüssmann, *J. Am. Oil Chem. Soc.* **55** (1978) 165.
652. H. Harder, *Seifen Oele Fette Wachse* **94** (1968) 789–794, 825.
653. H. Harder, D. Arends, W. Pochandke, *Seifen Oele Fette Wachse* **103** (1977) 180–183.
654. DIN 53 990:Waschmittel zum Waschen von Textilien, Empfehlungen für die vergleichende Prüfung von Gebrauchseigenschaften.
655. ISO TC 91–4319- 1977: Surface active agents–Detergents for washing fabrics–Guide for comparative testing of performance.
656. A. Trius, B. Brackmann, Proceedings of the 5th World Surfactants Congress (CESIO), Firenze 2000, 56–72.
657. "Higher Alcohols Market Forecast", Colin A. Houston & Associates, Inc., Pound Ridge (NY) 2000.
658. M. Elton, *Soap, Cosmet., Perfum.* **73** (2000) no. 22, 51–56.
659. Hauptausschuß Phosphate: *Phosphor*, Wege und Verbleib in der Bundesrepublik Deutschland. Verlag Chemie, Weinheim-New York 1978, pp. 29–30.
660. E. Roland, P. Kleinschmit: "Zeolites" in *Ullmann's Encyclopedia of Industrial Chemistry*, Sixth Edition, 2001, Electronic Release, Wiley-VCH, Weinheim, Germany 2001.
661. Statistisches Bundesamt 1998.
662. R. Wagner, *GWF, Gas Wasserfach: Wasser/Abwasser* **119** (1978) 235–242.
663. B. Böhnke, Gutachten zur Belastung der Gewässer durch Waschmittelinhaltstoffe aus Regenüberläufen und kommunalen Kläranlagen mit alleiniger mechanischer Reinigung. Industrieverband Körperpflege- und Waschmittel (IKW), Frankfurt, Dezember 1991.
664. B. Böhnke, Gutachten zur kostenmäßigen Beurteilung des Verhaltens von Waschmitteln in kommunalen Kläranlagen. Industrieverband Körperpflege- und Waschmittel (IKW), Frankfurt, Dezember 1990.
665. Europäisches Übereinkommen über die Beschränkung der Verwendung bestimmter Detergentien in Wasch- und Reinigungsmitteln vom 16. September 1968, BGBl 1968 I, p. 1718.
666. Council directive 73/404/EEC, of 22nd November 1973 on the approximation of the

laws of the Member States relating to detergents. OJ L347, p. 51.
667. Council directive 73/405/EEC, of 22nd November 1973 on the approximation of the laws of the Member States relating to detergents. OJ L347, p. 53.
668. J. Au, *Tenside, Deterg.* **18** (1981) 280–285.
669. Gesetz über die Umweltverträglichkeit von Wasch- und Reinigungsmitteln (Waschmittelgesetz) vom 20.08.1975, BGBl 1975 I, p. 2255.
670. Verordnung über die Abbaubarkeit anionischer und nichtionischer grenzflächenaktiver Stoffe in Wasch- und Reinigungsmitteln vom 30.01.1977 BGBl 1977 I, p. 244 (geändert durch Verordnung vom 18.06.1980, BGBl I, p. 706.)
671. Verordnung über Höchstmengen für Phosphate in Wasch- und Reinigungsmitteln (Phosphathöchstmengenverordnung — PHöchstMengV) vom 04.06.1980, BGBl 1980 I, p. 664.
672. Council Directive 82/243/EEC, of 31 March 1982 amending Directive 73/405/EEC on the approximation of the laws of the Member States relating to methods of testing the biodegradability of anionic surfactants. OJ L109, p. 18.
673. Council Directive 82/242/EEC, of 31 March 1982 amending Directive 73/405/EEC on the approximation of the laws of the Member States relating to methods of testing the biodegradability of nonionic surfactants. OJ L109, p. 1.
674. R. Wickbold, *Tenside* **9** (1972) 173–177.
675. D. Brown: "Introduction to surfactant biodegradation", in D. R. Karsa, M. R. Porter (eds.): *Biodegradability of surfactants*, Blackie Academic & Professional, London 1995, p. 1.
676. Proposed method for the determination of the biodegradability of surfactants used in synthetic detergents. OECD, Paris 1976.
677. Hauptausschuß Phosphate, Phosphor, Wege und Verbleib in der Bundesrepublik Deutschland, VCH, Weinheim, Germany 1978.
678. P. Gerike, K. Winkler, W. Jakob, *Tenside Surf. Det.* **26** (1989) 270–275.
679. P. Gerike, K. Winkler, W. Schneider, W. Jakob, J. Steber, *Tenside Surf. Det.* **28** (1991) 86–89.
680. Commission Directive 93/67/EEC on the assessment of risks to man and the environment of substances notified in accordance with Directive 67/584/EEC on dangerous substances. Off. J. Europ. Comm. L227 of Sept. 8, 1993.
681. Council Regulation (EEC) No. 793/93 of 23 March 1993 on the evaluation and control of the risks of existing substances. OJ L84. p. 1.
682. Technical Guidance Documents in support of the Commission Directive 93/67/EEC on risk assessment for new notified substances and the Commission Regulation (EC) 1488/94 on risk assessment for existing substances, ECB, Ispra April 19th, 1996.
683. OECD Guidelines for testing of chemicals, Volume 1, OECD, Paris 1993.
684. J. Waters, T. C. J. Feijtel, *Chemosphere* **30** (1995) 1939–1956.
685. P. Gerike, W. Holtmann, W. Jasiak, *Chemosphere* **13** (1984) 121–141.
686. J. Steber, *Textilveredlung* **26** (1991) 348–354.
687. R. Birch, C. Biver, R. Campagna, W. E. Gledhill, U. Pagga, J. Steber, H. Reust, W. J. Bontinck, *Chemosphere* **19** (1989) 1127–1550.
688. ECETOC, Technical report No. 61: "Environmental exposure assessment", European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), Brussels, November 1992.
689. T. Feijtel, E. J. van de Plassche, Environmental risk characterization of 4 major surfactants used in The Netherlands, RIVM report No. 679101, National Institute of public health and environmental protection, Bilthoven 1995.
690. AISE/CESIO: "Environmental risk assessment of detergent chemicals", Proceedings of the A.I.S.E./CESIO Limelette III Workshop on 28–29 November 1995. AISE Brussels 1996.
691. R. Schröder, *Tenside Surf. Det.* **32** (1995) 492–497.
692. J. Steber: "Biodegradation kinetic data for environmental risk assessment on surface waters and soil," in S. G. Hales, T. Feijtel, H. King, K. Fox, W. Verstraete (eds.): Biodegradation kinetics, Generation and Use of Data for Regulatory Decisions Making, Proceedings of the SETAC-Europe workshop 1996, SETAC Europe, Brussels, p. 81.
693. Council Directive 92/32/EEC of 30.04.1992 amending for the seventh time, Directive 67/548/EEC on the approximation of the laws, regulations and administrative provosion relating to the classification, packaging and labelling of dangerous substances. OJ L154, p. 1.
694. Verordnung über Anmeldeunterlagen und Prüfnachweise nach dem Chemikaliengesetz (ChemG Anmelde- und PrüfnachweisV), Änderung vom 31.5.1989, BGBl 1989 I, p. 1074.

695. DIN 38412, L127: "German standard methods for the examination of water, waste water and sludge; bio-assays (group L); determination of the inhibitory effect of waste water on the oxygen consumption of Pseudomonas putida", November 1992.
696. EN ISO 10712, "Water quality — Pseudomonas putida growth inhibition test (Pseudomonas cell multiplication inhibition test), December 1995.
697. W. Guhl, P. Gode, *Vom Wasser* **72** (1989) 165–173.
698. Allgemeine Verwaltungsvorschrift zum Wasserhaushaltsgesetz über die Einstufung wassergefährdender Stoffe in Wassergefährdungsklassen vom 17. Mai 1999. Bundesanzeiger, Jahrgang 51, Nr. 98a, ausgegeben am 29.05.1999.
699. W. Guhl, *Z. Angew. Zool.* **74** (1987) 385–409.
700. W. Guhl, P. Gode, *Tenside Surf. Det.* **26** (1989) 282–287.
701. W. J. Lyman, E. F. Reehl, D. H. Rosenblatt (eds.): *Handbook of chemical property estimation methods*, McGraw Hill Book Comp., New York 1982.
702. ECETOC, Technical Report No. 67: "The role of bioaccumulation in environmental risk assessment: The aquatic environment and related food webs". European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC). Brussels, March 1995.
703. W. K. Fischer, K. Winkler, *Vom Wasser* **47** (1976) 81–129.
704. P. Gerike, K. Winkler, W. Schneider, W. Jakob, *Tenside Surf. Det.* **26** (1989) 21–26.
705. R. Schröder, *SOFW-J.* **121** (1995) 420–427.
706. J. Steber, H. Berger: "Biodegradability of anionic surfactants", in D. R. Karsa, M. R. Porter (eds.): *Biodegradability of surfactants*, Blackie Academic & Professional, London 1995, p. 134.
707. P. Berth, P. Gerike, P. Gode, J. Steber, *Tenside Surf. Det.* **25** (1988) 108–115.
708. P. Schöberl, *Tenside Surf. Det.* **32** (1995) 25–35.
709. J. Tolls, P. Kloepper-Sams, D. T. H. M. Sijm, *Chemosphere* **29** (1994) 693–717.
710. J. Tolls, M. Haller, I. de Graaf, M. A. T. C. Thijssen, D. T. H. M. Sijm, *Environ. Sci. Technol.* **31** (1997) 3426–3431.
711. J. Steber, P. Wierich, *Appl. Environ. Microbiol.* **49** (1985) 530–537.
712. J. Steber, P. Wierich, *Water. Res.* **21** (1987) 661–667.
713. H. Hellmann, *Vom Wasser* **64** (1985) 29–42.
714. C. S. Newsome, D. Howes, S. J. Marshal, R. A. van Egmont, *Tenside Surf. Det.* **32** (1995) 498–503.
715. Shell Chemicals U.K. Ltd., *Chim. Oggi* **8** (1983) 19–22, 53.
716. E. Stephanou, W. Giger, *Environ. Sci. Technol.* **16** (1982) 800–805.
717. M. Tschui, P. H. Brunner, *Vom Wasser* **65** (1985) 9–19.
718. J. Steber, W. Guhl, N. Stelter, F. R. Schröder, *Tenside Surf. Det.* **32** (1995) 515–521.
719. H. A. Painter: "Biodegradability testing", in D. R. Karsa, M. R. Porter (eds.): *Biodegradability of surfactants*, Blackie Academic & Professional, London 1995, p. 65.
720. P. Gerike, *Tenside Surf. Det.* **19** (1982) 162–164.
721. H. Hellmann, *Z. Wasser Abwasser Forsch.* **22** (1989) 131–137.
722. J. Waters, H. H. Kleiser, M. J. How, M. D. Barratt, R. R. Birch, R. J. Fletcher, S. D. Haigh, S. G. Hales, S. J. Marshal, T. C. Pestell, *Tenside Surf. Det.* **28** (1991) 460–468.
723. D. J. Versteeg, T. C. J. Feijtel, C. E. Cowan, T. E. Ward, R. A. Rapaport, *Chemosphere* **24** (1992) 641–662.
724. Umweltbundesamt: Die Prüfung des Umweltverhalten von Natrium-Aluminium-Silikat Zeolith A als Phosphatersatzstoff in Wasch- und Reinigungsmitteln, Materialien 4/1979, E. Schmidt-Verlag, Berlin 1979.
725. C. P. Kurzendörfer, P. Kuhm, J. Steber: "Zeolites in the environment" in M. J. Schwuger (ed.): *Detergents in the environment*, Marcel Dekker Inc., New York 1997, p. 127.
726. P. Gode, *Z. Wasser Abwasser Forsch.* **16** (1983) 210–219.
727. A. W. Maki, K. J. Macek, *Environ. Sci. Technol.* **12** (1978) 573–580.
728. H. E. Allen, S. H. Cho, T. A. Neubecker, *Water Res.* **17** (1983) 1871–1879.
729. H.-J. Opgenorth: "Polymeric materials polycarboxylates", in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 337.
730. ECETOC, Joint Assessment of Commodity Chemicals No. 23: "Polycarboxylate polymers as used in detergents." European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) Brussels 1993.
731. P. Schöberl, L. Huber, *Tenside Det. Surf.* **25** (1988) 99–107.

732. H. L. Hoyt, H. L. Gewanter: "Citrate" in O. Hutzinger (ed.): *The Handbook of environmental chemsitry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 229.
733. H. Bernhardt: *Studie über die aquatische Umweltverträglichkeit von Nitrilotriacetat*, Verlag Hans Richarz, Sankt Augustin 1984.
734. "Aquatische Umweltverträglichkeit von Nitrilotriessigsäure (NTA)", Kernforschungszentrum Karlsruhe GmbH, Projektträger Wassertechnologie und Schlammbehandlung (PtWT) 1991.
735. G. Müller, U. Nagel, I. Purba, *Chem. Ztg.* **102** (1978) 169–178.
736. F. Dietz, *GWF, Gas Wasserfach: Wasser/Abwasser* **116** (1975) 301–308.
737. G. Graffmann, P. Kuzel, H. Nösler, G. Nonnenmacher, *Chem. Ztg.* **98** (1974) 499–504.
738. B. Wiecken, S. Wübbold-Weber, *SOFW-J.* (1994).
739. K. Raymond, L. Butterwick: "Perborate" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 287.
740. W. Guhl, *SOFW-J.* **18** (1992) 1159–1168.
741. P. A. Gilbert: "TAED-Tetraacetylethylenediamine" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 319.
742. W. E. Gledhill, T. C. J. Feijtel: "Environmental properties and safety assessment of organic phosphonates used for detergent and water treatment applications" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F, Detergents**, Springer-Verlag Berlin Heidelberg 1992, p. 261.
743. J. Steber, P. Wierich, *Chemosphere* **15** (1986) 929–945.
744. J. Steber, P. Wierich, *Chemosphere* **16** (1987) 1323–1337.
745. G. Müller, J. Steber, H. Waldhoff, *Vom Wasser* **63** (1984) 63–78.
746. K. Wolf, P. A. Gilbert: "EDTA-Ethylenediaminetetraacetic acid" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 243.
747. AIS Environmental Safety Working Group: An assessment of the implications of the use of EDTA in detergent products. AIS, Brussels, April 1987.
748. J. Kaschig, R. Hochberg, M. Zeller: "Concerning the environmental fate of fluorescent whitening agents", in: Proc. of the 4th World Surfactants Congress 3 (CESIO), 1996, 155.
749. J. B. Kramer: "Fluorescent whitening agents" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 351.
750. J. G. Bartelaan, C. G. van Ginkel, F. Balk: "Carboxymethylcellulose (CMC)" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part F**, Springer-Verlag Berlin Heidelberg 1992, p. 329.
751. G. Rimkus, M. Wolf, *Chemosphere* **30** (1995) 641–651.
752. H. F. Geyer, G. Rimkus, M. Wolf, A. Attar, C. Steinberg, A. Kettrup, *UWSF-Z. Umweltchem. Ökotox.* **6** (1994) 9–17.
753. N. J. Fendinger, R. G. Lehmann, E. M. Mihaich, "Polydimethylsiloxane" in O. Hutzinger (ed.): *The Handbook of environmental chemistry*, Vol. **3, Part H**, Springer-Verlag Berlin Heidelberg 1997, p. 181.
754. Industrieverband Körperpflege und Waschmittel, Frankfurt, 2000.
755. M. J. Schwuger, F. G. Bartnik in Ch. Gloxhuber (ed.): *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Surfactant Sci. Ser. **10**, 1981, pp. 1–49.
756. J. G. Black, D. Howes in Ch. Gloxhuber (ed.): *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Surfactant Sci. Ser. **10**, 1981, pp. 51–85.
757. R. B. Drotman in V. A. Drill, P. Lazar (eds.): *Cutaneous Toxicity*, Academic Press, New York 1977, pp. 96–109.
758. R. B. Drotman, *Toxicol. Appl. Pharmacol.* **52** (1980) 38.
759. B. Isomaa, *Food Cosmet. Toxicol.* **13** (1975) 231.
760. A. Siwak, M. Goyer, J. Perwak, P. Thayer in K. L. Mittal, E. J. Fendier (eds.): *Solution Behavior of Surfactants*, vol. **I**, Plenum Publ. Corp., New York 1982, p. 161.
761. R. A. Cutler, H. P. Drobeck in E. Jungermann (ed.): *Cationic Surfactants*, Surfactant Sci. Ser. **5**, 1970, p. 527.
762. D. L. Opdyke, C. M. Burnett, *Proc. Sci. Sect. Toilet. Goods Assoc.* **44** (1965) 3.

763. W. Kästner in Ch. Gloxhuber (ed.): *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Surfactant Sci. Ser. **10**, 1981, pp. 139–307.
764. Ch. Gloxhuber, M. Potokar, S. Braig, H. G. van Raay et al., *Fette Seifen Anstrichm.* **76** (1974) 126.
765. R. D. Swisher, *Arch. Environ. Health* **17** (1968) 232.
766. W. B. Coate, W. M. Busey, W. H. Schoenfisch, N. M. Brown et al., *Toxicol. Appl. Pharmacol.* **45** (1978) 477.
767. W. Kissler, K. Morgenroth, W. Weller, *Prog. Respi. Res.* **15** (1981) 121.
768. M. S. Potokar in Ch. Gloxhuber (ed.): *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Surfactant Sci. Ser. **10**, 1981, pp. 87–126.
769. K. Oba in Ch. Gloxhuber (ed.): *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Surfactant Sci. Ser. **10**, 1981, pp. 327–403.
770. Y. Yam, K. A. Booman, W. Broddle, L. Geiger et al., *Food Chem. Toxicol.* **22** (1984) no. 9, 761–769.
771. Ch. Gloxhuber, *Fette Seifen Anstrichm.* **74** (1972) 49.
772. Ch. Gloxhuber, M. Potokar, W. Pittermann, S. Wallat et al., *Food Chem. Toxicol.* **21** (1983) no. 2, 209–220.
773. G. A. Nixon, *Toxicol. Appl. Pharmacol.* **18** (1971) 398–406.
774. W. R. Michael, J. M. Wakim, *Toxicol. Appl. Pharmacol.* **18** (1971) no. 2, 407–416.
775. J. A. Budny, J. D. Arnold, *Toxicol. Appl. Pharmacol.* **25** (1973) no. 1, 43–53.
776. G. A. Nixon, E. V. Buehler, R. J. Nieuwenhuis, *Toxicol. Appl. Pharmacol.* **21** (1972) no. 2, 244–252.
777. National Cancer Institute, NCI Techn. Rep. Ser. no. 6, DHEW Publication no. (NHI) 77–806 (1977).
778. P. S. Thayer, C. J. Kensler, A. D. Little, *CRC Crit. Rev. Environ. Control* **3** (1973) 335–340.
779. Ch. Gloxhuber, *Med. Welt* **19** (1968) 351–357.
780. ECETOC Technical Report No. 63: "Reproductive and General Toxicology of some Inorganic Borates and Risk Assessment for Human Beings", Brussels Febr. (1995), p. 1–91.
781. Ch. Gloxhuber, J. Malaszkiewicz, M. Potokar, *Fette Seifen Anstrichm.* **73** (1971) 182–189.
782. G. J. Schmitt, *Z. Hautkr.* **49** (1974) 901.
783. P. H. Andersen, C. Bindslev-Jensen, H. Mosbech, H. Zachariae, K. E. Andersen, *Acta Derm.-Venereol* **78** (1998) 60–62.
784. F. Coulston, F. Korte (eds.): *Fluorescent Whitening Agents*, Georg Thieme Verlag, Stuttgart 1975.
785. A. W. Burg, M. W. Rohovsky, C. J. Kensler, *CRC Crit. Rev. Environ. Control* **7** (1977) 91–120.
786. W. Matthies, A. Löhr, H. Ippen, *Dermatosen Beruf Umwelt* **38** (1990) 184–189.
787. G. J. Schmitt, *Tenside Deterg.* **16** (1979) 226–228.
788. R. Teichmann, C. Vogel, G. Busch, W. Körber, *Melliand Textilber.* **80** (1999) 638–639.
789. J. Rieker, T. Guschlbauer, *Textilveredlung*, **34** (1999) no. 11/12, 4–6.
790. CIRFS (Comité International de la Rayon et des Fibres Synthétiques, Paris): *Data on Synthetic Fibers*, 1996, pp. –51.
791. S. Mizobuchi, S. Sekiguchi, Household Science Laboratories, Lion Corporation, Tokyo 2000.
792. R. Weber, *Textilveredlung* **15** (1980) 380–385.
793. ISO 3758, EN DIN 23758.
794. Brochure No. 3951059 Rev. B, Whirlpool Co., Benton Harbor MI, USA 1999.
795. Brochure 9/2000, National Matsushita Electric Industrial Co., Tokyo 2000.
796. Hauptberatungsstelle für Elektrizitätsanwendung (HEA) e.V., Bilderdienst, Frankfurt 1992.
797. A. Cahn, *HAPPI* **32** (1995) no. 6, 77–86.
798. J. Motavalli, *E/The Magazine* **8** September—October 1997 no. 8, 77–86.
799. M. Liphard, A. Griza, *Tenside Surf. Det.* **34** (1997) no. 6, 410–416.
800. H. G. Hloch, Proceedings of the 50th International Detergency Conference (wfk), Luxembourg 1999, 31–42.
801. Appliance, Nov. 1999, 61–62.
802. European Energy Network, Working Group for Efficient Appliances, Danish Energy Agency, Copenhagen 1995.
803. Appliance, Sept. 1999, 77–80.
804. AEG, Brochure of Washers and Dryers, 2000.
805. Senking Wäschereianlagen, Hildesheim, Germany, 1997.

Dextran

ANTHONY N. DE BELDER, TdB Consultancy AB, Uppsala, Sweden

1. **Introduction** 1
2. **Structure, Chemical, and Physicochemical Properties** 1
2.1. **Structure** 1
2.2. **Physicochemical Properties** 2
2.3. **Reactivity** 3
3. **Production** 3
3.1. **Biosynthesis** 3
3.2. **Production of Clinical Dextran** 3
3.3. **Future Trends** 4
4. **Quality Specifications** 4
5. **Storage** 4
6. **Uses** 4
6.1. **Clinical Products** 4
6.2. **General Uses** 5
7. **Economic Aspects** 5
8. **Toxicology** 5
9. **References** 6

1. Introduction

The term dextran [*9004-54-0*] refers to those polysaccharides which are composed primarily of $1 \rightarrow 6$ linked α-D-glucopyranose units. Many dextrans contain branches attached to C-2, C-3, or C-4. Others may also contain $1 \rightarrow 3$ linkages in the main chain.

The formation of slimes and jellies during the processing of wines and in sugar refining has long been an undesirable complication. In the middle of the 19th century, reports by PASTEUR [1] and VAN TIEGHEM [2] implicated a bacterial mechanism. The name dextran appears to have been assigned ca. 1870 by SCHEIBLER [3], who also established that dextran was a polymer of glucose. Further references to the early history of dextran are available [4].

Dextrans are synthesized from sucrose by a large number of organisms. JEANES and coworkers have characterized over 96 strains of dextran-producing bacteria [5]. These bacteria are confined to the family Lactobacillaceae and in particular to the genera *Lactobacillus*, *Leuconostoc*, and *Streptococcus*. Several members of the Streptococcus group are implicated in the development of cariogenic plaque on tooth surfaces [6], [7].

Leuconostoc and *Lactobacillus* dextrans are also undesirable contaminants in the sugar refining industry, in which they adversely affect the filterability and crystallization of sucrose [8].

Only the slightly branched dextran from *Leuconostoc mesenteroides B-512 (F)* is of commercial interest. Most of the chemical, physicochemical, and biological reports have therefore focused on this dextran. Unless otherwise stated, the discussions in this article are concerned only with B-512 (F). Figure 1 shows its structure.

2. Structure, Chemical, and Physicochemical Properties

2.1. Structure

The dextran elaborated by *Leuconostoc mesenteroides NRRL B-512 (F)* consists of an $\alpha\,(1 \rightarrow 6)$-linked glucan with side chains attached to the 3-positions of the backbone glucose units. From periodate and methylation analyses, the degree of branching is estimated to be 5 %. Upon hydrolysis, the degree of branching is found to decrease slightly with decreasing molecular mass.

The ^{1}H- and ^{13}C-NMR spectra afford compelling evidence for the main structural features of dextran [9]. The degree of branching of clinical dextrans as determined by ^{13}C-NMR is 4.0 – 6.0 %.

Many details of the fine structure are unresolved, in particular the length and distribution of the side chains. LARM and colleagues concluded that 40 % of the side chains are one unit long, 45 % are two units long, and the remaining 15 % are still longer [10]. The preponderance of single unit branches in several other dextrans, *Leuconostoc mesenteroides B-1375, B-1415*, and *B-1416*, has been established

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_449

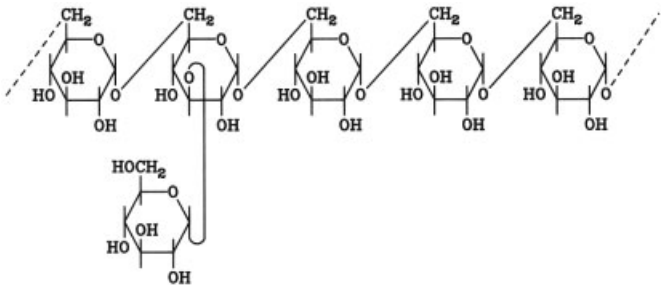

Figure 1. Structure of a fragment of the dextran B-512 (F) chain

[11]. The availability of dextranases with well characterized properties has permitted WALKER and colleagues to corroborate these results [12]. The presence of a small proportion of longer branches (up to 50 units) is implicit in the studies of the viscosity of dextran [13]. COVACEVICH and RICHARDS [14] concluded that the branches were distributed in a relatively regular manner.

2.2. Physicochemical Properties

Dextran B-512 (F) is freely soluble in water, methyl sulfoxide, formamide, ethylene glycol, glycerol, *N*-methylmorpholine-*N*-oxide, and hexamethylphosphoramide. Some dextran fractions may have adopted a certain degree of crystallinity and can only be brought into solution by strong heating.

The molecular mass of hydrolyzed natural dextran NRRL B-512 (F) is 9×10^6 to 500×10^6 [13], [15]. Measurements in a variety of solvents (e.g., 4 M aqueous sodium chloride and 6 M aqueous urea) failed to reveal any evidence of association.

The relationship between the mass-average molecular mass $\bar{M}_\mathrm{w}$ and the intrinsic viscosity is shown in Figure 2 [13], [16]. GRANATH obtained the viscosity dependence of dextran in the clinical range [16]:

$$[\eta] = 2.23 \times 10^{-3} \bar{M}_\mathrm{v}^{0.43} \ \left(\text{for } M_\mathrm{r} < 5 \times 10^5\right)$$

where $[\eta]$ = intrinsic viscosity and $\bar{M}_\mathrm{v}$ = viscosity-average molecular mass.

The colloid osmotic pressure of dextran solutions significantly affects their plasma volume expansion [17].

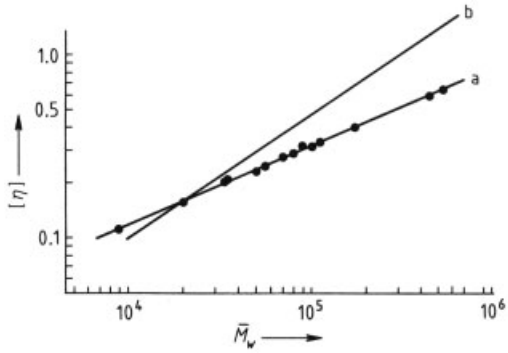

Figure 2. Log–log plot of intrinsic viscosity $[\eta]$ against mass-average molecular mass $\bar{M}_\mathrm{w}$ for dextran B-512 (a) and for a hypothetical linear dextran (b)

The $\alpha(1\rightarrow6)$ linked polysaccharides represent a class of very flexible and extended polymers. Table 1 shows the relationship between mass-average molecular mass and the radius of gyration. At high molecular masses, the molecules display increased symmetry [13], [16].

Table 1. Molecular dimensions of dextran B-512(F) [16]

Molecular mass $\bar{M}_\mathrm{w}$	Radius of gyration, nm
2×10^6	38
5×10^5	20
1×10^5	9.5
5×10^4	6.8
(Albumin)	3.5

At M_r 2000, the solution properties are best explained by a transition from a coil to a rodlike conformation. Several attempts have been made to crystallize dextran, and CHANZY et al. grew single crystals from a dextran fraction having M_r 19 900 and proposed a ribbon-like conformation [18].

2.3. Reactivity

As with other glucans, the reactivity of the hydroxyl group at position 2 toward alkylating agents is higher than at position 3 or position 4. Studies on the partial methylation of dextran revealed the following relative reactivities: $k_2 : k_3 : k_4 = 16 : 2 : 7$. Migration of the substituents may subject acylations to thermodynamic control. Thus, although partial acetylation of dextran with acetic anhydride – pyridine in formamide as solvent gave $k_2 \gg k_3 = k_4$, the reactivities were virtually identical when the reaction was carried out with acetic anhydride in aqueous alkali [19].

3. Production

3.1. Biosynthesis

The early work of Beijerinck [20] and Hehre [21] showed that dextran was elaborated by an extracellular enzyme designated dextransucrase (sucrose-1,6-α-D-glucan 6-α-glucosyltransferase, E.C. 2.4.1.5).

The dextransucrases from many *Leuconostoc* and *Streptococcal* strains have been isolated and their properties studied. The purified B-512 (F) enzyme, which binds dextran strongly, rapidly loses activity at 4 °C and even at − 15 °C. Its activity decreases by 60 % over 20 d. The enzyme can be stabilized by adding dextran (> 4 mg/mL) [22].

Dextransucrase appears to be a glycoprotein (M_r 64 000) with mannose being the primary sugar. Calcium ions appear to be essential for the activity of the enzyme.

A mechanism for the biosynthesis has been proposed whereby the enzyme serves two functions [23]. It first hydrolyzes the sucrose and binds the glucosyl moiety, and thereafter it builds up the dextran chain by an insertion mechanism (Fig. 3).

The biosynthesis can be terminated by any one of a large number of acceptors, and if so, the dextran chain is released. Examples of acceptors are maltose, isomaltose, and methyl-α-D-glucoside. Dextran itself can also function as an acceptor [24].

The formation of branches is also an acceptor reaction in which a ring hydroxyl group on a dextran molecule is inserted in the growing dextran chain.

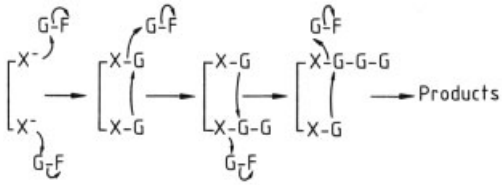

Figure 3. Two-site mechanism for the biosynthesis of dextran chains by dextransucrase
G–F represents a sucrose molecule composed of glucose (G) and fructose (F). G–G represents α (1 → 6) linked glucose residues. The binding sites X may not be identical.

3.2. Production of Clinical Dextran

In the West, most major manufacturers of dextran employ a process based on the batchwise culture of *Leuconostoc mesenteroides B-512 (F)* in the presence of sucrose. The viscous culture fluid is precipitated in an alcohol, and the native dextran then is hydrolyzed and fractionated to the desired molecular mass range. In Japan and Eastern European countries, different strains of *Leuconostoc* are used, but the dextrans produced are of similar structure. The preparation of dextran on a laboratory scale has been described in detail [25], and the industrial production of dextran has been reviewed [26–28].

To obtain vigorous growth of the organism, the culture medium must contain various nutrients in addition to sucrose. In practice, these are supplied by addition of either yeast extracts, corn steep liquors, or malt extracts with peptone or tryptone broth. The pH of the medium affects both the production of dextransucrase by the organism and the activity and stability of the enzyme; Figure 4 shows the pH changes. The influence of sucrose concentration on production of dextran has been reviewed [26]. A sucrose concentration of ca. 10 % is satisfactory; at higher concentrations the yield of dextran with $\bar{M}_w > 5000$ decreases. The cultures are maintained at a temperature of ca. 25 °C, at which the fermentation is complete after 24 – 48 h. Prolongation of the cultures for more than 48 h may lead to a decrease in molecular mass of the dextran.

The native dextran that is obtained is then subjected to partial hydrolysis to form products of appropriate molecular size, and fractionation by ethanol or methanol to give clinical fractions.

Extensive studies on the acid hydrolysis and fractionation of native B-512 (F) dextran have been carried out and conditions for maximal yields have been established [29], [30]. Following the partial hydrolysis, the clinical fraction can be precipitated with 39 – 46 % aqueous ethanol under careful temperature control (25 °C). Membrane technology now often replaces ethanol fractionation in purification operations. The product is then redissolved and spray dried.

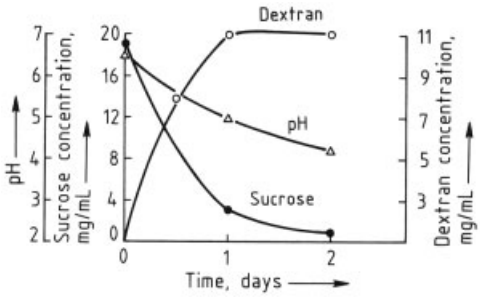

Figure 4. Changes in sucrose concentration, pH, and dextran concentration during the fermentation of *Leuconostoc mesenteroides*

3.3. Future Trends

Several alternative processes have been devised over the years:

1) Fermentation with cell-free extracts
2) Fermentation in the presence of acceptors
3) Continuous processes
4) Use of immobilized dextransucrase

About 40 years ago, a supplement (1 mg/mL) of low molecular mass dextran was found to yield a product dextran of lower molecular mass [31]. Few producers have adopted this type of process for clinical dextrans. One drawback is that it requires a production facility for low molecular mass dextran.

Promising results have been obtained with immobilized dextransucrase but only on a laboratory scale [32].

4. Quality Specifications

Monographs for quality specifications of dextran 40, 60, and 70 substance have been published in the European Pharmacopoeia and for dextran 40 and 70 in USP. The molecular size is now determined by size exclusion chromatography. Other quality parameters are acidity/alkalinity, nitrogen impurities, residual solvents, heavy metals, sulfated ash, bacterial endotoxins, and microbial contamination.

5. Storage

Dextran is stable indefinitely when stored in the absence of light, excessive heat, and moisture. Sterile solutions of clinical dextrans with pH 4 – 7 have excellent chemical stability (at least 10 years) when stored at 4 – 40 °C.

Clinical dextran solutions tend to form small amounts of flakes if kept in glass bottles and subjected to considerable temperature fluctuations. These flakes have been found to consist of dextran with the same M_r as the substance in solution. They redissolve upon heating.

6. Uses

6.1. Clinical Products

Dextran 70 (M_r ca. 70 000) is marketed in most countries as a 6 % solution in normal saline and, as such, represents the plasma volume expander of choice. It is one of two plasma volume expanders included in the WHO list of essential drugs. Clinical experience supports its use in the treatment of shock or impending shock as a result of hemorrhage, burns, or surgery. Dextran 60 has replaced dextran 70 in some parts of Europe.

Dextran 40 ($\bar{M}_w$ ca. 40 000) was introduced in 1961 following studies of the erythrocyte disaggregating and blood flow improvement properties of lower molecular mass dextrans. It is marketed as a 10 % solution in normal saline or 5 % glucose which provides rapid plasma volume expansion and promotes microcirculatory blood flow. Both products are used extensively to prevent fatal pulmonary complications of surgery, trauma, and shock. A 6 % solution of dextran 70 in a hypertonic saline solution (7.5 % NaCl) has been shown to be exceptionally effective as a plasma volume expander for primary (prehospital) resuscitation in severe trauma and hemorrhagic shock [33].

In 1982, the monovalent hapten dextran 1 (a fraction of $\bar{M}_w$ ca. 1000) was launched as prophylaxis against dextran-induced anaphylactoid

reactions, which, although rare, may cause undesirable complications in the clinic. Dextrans are commonly added to infusion and perfusion solutions, for example, in connection with transplants. A dextran sulfate of $\bar{M}_w$ 7000 and sulfur content of 16 % was tested clinically as an anticoagulant, but at the doses used, it caused toxic symptoms [34]. A dextran sulfate with M_r approx. 8000 with in vitro antiviral properties has been tested in patients with HIV infection. The results were not encouraging [35]. A diethylaminoethyl derivative of dextran ($\bar{M}_w$ ca. 500 000) has been found to be effective in reducing serum cholesterol.

Iron-dextran products prepared from low molecular mass dextrans are used widely for treating anemic piglets and humans.

In 1977, a cross-linked dextran in bead form (Debrisan) was introduced as a wound-cleansing agent for secreting and infected wounds in particular.

6.2. General Uses

Purified dextran fractions are currently used in the cosmetic and photographic industries. A comprehensive bibliography covering dextran literature and patents is available [36]. Dextran fractions find application for partitioning subcellular particles in two-phase polymer systems and as cryoprotective agents.

Sephadex gels, prepared by emulsion polymerization of dextran, have occupied a dominant place among gel filtration media for many years. Sephadex G-25 is used on an industrial scale for desalting operations during the purification of biopolymers of medical importance, e.g., insulin. Many of the ion exchangers derived from Sephadex are also used commercially for separations [37].

Dextran fractions have been used extensively for preparing conjugates with biologically active substances, e.g., drugs, enzymes, and hormones [38]. Conjugation prolongs the lifetime of the active component in vivo, increases its stability, and facilitates the targeting of the active component. Many derivatives of dextran have been described and numerous patents have appeared. However, few have attained significant commercial interest. Dextran sulfate ($\bar{M}_w$ ca. 500 000; sulfur content 18 %) and diethylaminoethyl dextran ($\bar{M}_w$ ca. 500 000) have revealed an astonishing range of biological activities [39].

Several microcarriers based on cross-linked dextran substituted with cationic substituents (e.g., Cytodex) are used commercially for culturing anchorage-dependent animal cells. As such, they have proved invaluable for the production of vaccines, viruses, and interferon.

7. Economic Aspects

Exact figures for the annual world production are not available but the size of magnitude would be somewhat in excess of 500 t.

The major producer of *dextran* is Amersham Pharmacia Biotech (Uppsala, Sweden); other producers are Meito (Japan), Polfa (Poland), Pharmacosmos (Denmark). The major producers of *clinical dextran solutions* are: Abbott (United States), Baxter-Travenol (Unites States and Europe), Polfa (Poland), Infusia (Czech Republic), Otsuka and Green Cross (Japan), B. Braun (Germany) and Pharmalink (Sweden).

8. Toxicology

The infusion of clinical dextran solutions in humans is now approved by the regulatory authorities in most countries. No adverse effects are to be expected, provided the recommended doses are observed. However, a low incidence of dextran-induced anaphylactoid reactions has been reported. The risk of this type of reaction has been virtually eliminated by the preinjection of 20 mL of a 15 % dextran 1 (M_r ca. 1000) solution [40].

Dextran B-512 (F) given orally is degraded by bacteria in the gut and the products produce a rapid increase in blood sugar and liver glycogen. Dextran administered intravenously is degraded in the liver to carbon dioxide and water.

Dextrans are not permitted as food additives in the United States and Western Europe [41], [42], but they are considered safe as components of food packing materials. Dextrans may be used as additives in pharmaceutical and cosmetic formulations provided that the necessary safety documentation is presented.

9. References

1. L. Pasteur, *Bull. Soc. Chim. Fr.* **1861,** 30 – 31.
2. P. van Tieghem, *Ann. Sci. Nat. Botan. Biol. Veg. s 2,* **7** (1878) 180 – 203.
3. G. Scheibler, *Z. Ver. Dtsch. Zucker-Ind.* **19** (1869) 472; **24** (1874) 309.
4. T. H. Evans, H. Hibbert, *Adv. Carbohydr. Chem.* **2** (1946) 209 – 219.
5. A. Jeanes, W. C. Haynes, C. A. Wilham, J. C. Rankin et al., *J. Am. Chem. Soc.* **76** (1964) 5041 – 5052.
6. R. L. Sidebotham, *Adv. Carbohydr. Chem. Biochem.* **30** (1974) 435 – 444.
7. G. J. Walker: "Int. Rev. of Biochem.," in D. J. Manners (ed.): *Biochem. of Carbohydrates II,* **vol. 16,** Univ. Park Press, Baltimore 1978, pp. 75 – 110.
8. P. A. Inkerman in M. B. Lopez, C. M. Madrazo (eds.): *Proceedings XVII Congress, International Society of Sugar Cane Technologists,* **vol. 3,** Executive Committee of ISSCT, Metro Manila 1980, pp. 2411 – 2427.
9. D. Gagnaire, M. Vignon, *Makromol. Chem.* **178** (1977) 2321 – 2333.
10. O. Larm, B. Lindberg, S. Svensson, *Carbohydr. Res.* **20** (1971) 39 – 48.
11. E. J. Bourne, D. H. Hutson, H. Weigel, *Biochem. J.* **86** (1963) 555 – 562.
12. C. Taylor, N. W. H. Cheetham, G. J. Walker, *Carbohydr. Res.* **137** (1985) 1 – 3.
13. F. R. Senti, N. N. Hellman, N. H. Ludwig, G. E. Babcock et al., *J. Polym. Sci.* **17** (1955) 527 – 546.
14. M. T. Covacevich, G. N. Richards, *Carbohydr. Res.* **54** (1977) 311 – 315.
15. L. H. Arond, H. P. Frank, *J. Phys. Chem.* **58** (1954) 953 – 958.
16. K. A. Granath, *J. Colloid Sci.* **13** (1958) 308 – 328.
17. U. F. Gruber: *Blutersatz,* Springer Verlag, Berlin 1968, p. 68.
18. C. Guizard, H. Chanzy, A. Sarko, *Macromolecules* **17** (1984) 100 – 107.
19. A. N. de Belder, B. Norrman, *Carbohydr. Res.* **8** (1968) 1 – 6.
20. M. W. K. Beijerinck, *Akad. Wetensch. Amsterdam Proc. Sect. Sc.* **12** (1910) 635.
21. E. J. Hehre, *Science (Washington, D.C., 1883)* **93** (1941) 237 – 238.
22. M. Kobayashi, K. Matsuda, *Biochem. Biophys. Acta* **614** (1980) 46 – 62.
23. J. F. Robyt, B. K. Kimble, T. F. Walseth, *Arch. Biochem. Biophys.* **165** (1974) 634 – 640.
24. J. F. Robyt, S. H. Eklund, *Bioorg. Chem.* **11** (1982) 115 – 132.
25. A. Jeanes in R. L. Whistler (ed.): *Methods in Carbohydrate Chemistry,* **vol. 5,** Academic Press, New York 1965, p. 118.
26. R. M. Alsop, *Prog. Ind. Microbiol.* **18** (1983) 1 – 44.
27. G. H. Bixler, G. H. Hines, R. M. McGhee, R. A. Shurter, *Ind. Eng. Chem.* **45** (1953) 692 – 705.
28. F. H. Foster, *Process Biochem.* **3** (1968) 15 – 19.
29. I. A. Wolff, C. L. Mehltretter, R. L. Mellies, P. R. Watson et al., *Ind. Eng. Chem.* **46** (1954) 370 – 377.
30. M. Zief, G. Brunner, J. Metzendorf, *Ind. Eng. Chem.* **48** (1956) 119 – 121.
31. H. M. Tsuchiya, N. N. Hellman, H. J. Koepsell, J. Corman et al., *J. Am. Chem. Soc.* **77** (1955) 2412 – 2419.
32. A. Lopez, P. Monsan, *Biochimie* **62** (1980) 323 – 329.
33. G. C. Kramer, G. I. Elgjo, L. F. Poli de Figueiredo, C. E. Wade, *Bailliere's Clin. Anaesthesiology,* **11** (1997) 143 – 161
34. P. Hjort, H. Stormorken, O. Gilje, *J. Clin. Lab. Invest.* **9, Suppl. 29** (1957).
35. C. Flexner et al., *Antimicrob. Agents Chemother.,* **35** (1991) 2544 – 2550.
36. A. Jeanes: *Dextran Bibliography,* Misc. Publication no. 1355, U. S. Department of Agriculture, Washington 1978.
37. J. M. Curling (ed.): *Separation of Plasma Proteins,* Pharmacia Fine Chemicals AB, Uppsala 1983.
38. L. Molteni in H. Bundgaard, A. B. Hansen, H. Kofod (eds.): *Optimization of Drug Delivery,* Munksgaard, Copenhagen 1982, pp. 285 – 301. Z. A. Rogovin, A. D. Virnik, K. P. Khomiakov, O. P. Laletina et al., *J. Macromol. Sci. Chem.* **A 6** (1972) 569 – 593.
39. A. N. de Belder in R. L. Whistler (ed.): *Industrial Gums: Polysaccharides and Their Derivatives,* 3rd ed.,Academic Press, 1993.
40. A. W. Richter, H. I. Hedin, *Immunol. Today* **3** (1982) 132 – 138.
41. M. Glicksman in M. Glicksman (ed.): *Food Hydrocolloids,* **vol. 1,** CRC Press, Boca Raton, Florida, 1982, p. 157.
42. *Fed. Regist.* **42** (1977) Nov. 18, FR 59 518-21.

Dextrose → Glucose and Glucose-Containing Syrups
Diacetone Alcohol → Acetone
Diacyl Peroxides → Peroxy Compounds, Organic

Diagnostic Reagents

For X-ray contrast materials, see → Radiopaque Media

William J. Litchfield, E. I. Du Pont de Nemours & Co., Wilmington, Delaware 19898, United States (Chaps. 1 – 3)

James C. Richards, Standard Oil Co. (Indiana), Naperville, Illinois 60566, United States (Chap. 4)

John A. Wehrly, E. I. Du Pont de Nemours & Co., Wilmington, Delaware 19898, United States (Chap. 5)

James F. Pierson-Perry, E. I. Du Pont de Nemours & Co., Wilmington, Delaware 19898, United States (Chap. 5)

Maria P. Liteplo, E. I. Du Pont de Nemours & Co., N. Billerica, Massachusetts 01862, United States (Chap. 6)

1. Introduction 4
1.1. Medical Significance 4
1.2. Economic Significance 4
1.3. Regulatory Controls 5
1.4. Classifications of Diagnostic Reagents 5
2. Reagents for General Clinical Chemistry 5
2.1. STAT and Routine Testing 5
2.2. Quantitation of Electrolytes 5
2.3. General Chemistry Tests 6
2.4. Quantitation of Serum Proteins and Enzymes 8
2.5. Therapeutic Drug Monitoring 10
2.6. Measurement of Hormones and Tumor Markers 11
3. Immunodiagnostic Reagents 11
3.1. Polyclonal and Monoclonal Antibodies 11
3.2. Radioimmunodiagnostic Reagents . 12
3.2.1. Radioimmunoassays 12
3.2.2. Radioisotopic Dilution Assays 12
3.2.3. Immunoradiometric Assays 12
3.3. Reagents for Nonisotopic Homogeneous Immunoassays 13
3.3.1. Latex Agglutination 13
3.3.2. Enzyme-Tagged Immunoassays 13
3.3.3. Fluorescence-Based Immunoassays . 14
3.3.4. Fluorescence Polarization 14
3.3.5. Liposome-Based Immunoassays . . . 14
3.4. Reagents for Nonisotopic Heterogeneous Immunoassays 15
3.4.1. Enzyme-Linked Immunosorbent Assays 15
3.4.2. Sandwich Immunoassays 16
3.4.3. Affinity Column Mediated Immunoassays 16
3.4.4. Avidin and Biotin 17
4. Microbiologic Reagents 17
4.1. Collection, Transport, Processing Devices, and Reagents 18
4.2. Growth-Dependent Microbial Reagents 19
4.3. Immunodiagnostic Reagents in Clinical Microbiology 19
4.4. Monoclonal Antibody and Nucleic Acid Assays 21
5. Coagulation Diagnostic Reagents . . 23
5.1. Introduction to Hemostasis 23
5.2. Screening Tests 24
5.3. Specific Analytes 25
5.3.1. Coagulation Factors 25
5.3.2. Fibrinolytic Pathway 26
5.3.3. Activators 27
5.3.4. Inhibitors 27
5.4. Platelets 28
5.5. Control of Anticoagulant Therapy . 28
5.5.1. Oral Anticoagulants 28
5.5.2. Heparin 28
6. Reagents for Nuclear Medicine . . . 29
6.1. Nuclear Imaging (Radioscintigraphy) 29
6.2. Detection Devices 29
6.3. Radionuclides 30
6.3.1. Selection Criteria 30
6.3.2. ^{99m}Tc 31
6.3.3. Other γ-Emitters 32
6.3.4. Positron Emitters 32
6.4. Diagnostic Radiopharmaceuticals . 32
6.4.1. Heart Imaging 32
6.4.2. Brain Imaging 33
6.4.3. Bone Imaging 33
6.4.4. Lung Imaging 34
6.4.5. Hepatobiliary Imaging 34
6.4.6. Reticuloendothelial System 35
6.4.7. Renal Imaging 35

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_455

6.4.8. Malignant and Inflammatory Regions 35
6.5. Experimental Agents 35
6.6. Quality Control 37
6.7. Dosimetry 37
7. References 38

The following is a list of specific abbreviations used in this article.

AACC American Association for Clinical Chemistry
AAP 4-aminoantipyrine
ACT activated clotting time
ACTH adrenocorticotropic hormone
ADP adenosine 5′-diphosphate
AFP α-fetoprotein
AMP adenosine 5′-monophosphate
APTT activated partial thromboplastin time
ARD antimicrobial removal device
ARIES apoenzyme reconstitution immunoassay system
AT III antithrombin III
ATCC American Tissue Culture Collection
ATP adenosine 5′-triphosphate
B-HCG human chorionic gonadotropin, B-subunit
BB CPK isoenzyme
BHI brain heart infusion
BPS bathophenanthroline sulfonate
BUN blood urea nitrogen
CAP College of American Pathologists
CAT computerized axial tomography
CE cholesterol esterase
CEA carcinoembryonic antigen
CF complement fixation
CFP cerebrospinal fluid protein
CK creatine kinase
CKMB creatine kinase, MB isoenzyme
CO cholesterol oxidase
DEA N,*N*-diethylaniline
DEAE diethylaminoethyl
DMSA dimercaptosuccinic acid
DNA deoxyribonucleic acid
DTPA diethylenetriamine-*N,N,N′,N″,N″*-pentaacetic acid
E.C. Enzyme Commission
EDTA ethylenediaminetetraacetic acid
EGTA ethylenebis(oxyethylenenitrilo)tetraacetic acid
EHDP ethane 1-hydroxy-1, 1-diphosphonate
EIA enzyme immunoassay
ELISA enzyme-linked immunosorbent assays
EMIT enzyme-multiplied immunoassay technique
FA immunofluorescent staining
FDP fibrin degradation products
FIA fluorescence immunoassay
FSH follicle-stimulating hormone
G6P glucose 6-phosphate
G6PDH glucose 6-phosphate dehydrogenase
GD α-glucosidase
GLDH glutamate dehydrogenase
GOT glutamic oxalacetic transaminase
GPT glutamic pyruvic transferase
HAD hemadsorption
HAM human albumin microspheres
HCG human chorionic gonadotropin
HDL high-density lipoprotein
HGH human growth hormone
HI hemagglutination-inhibition test
HIDA *N*-(2,6-dimethylphenylcarbamoylmethyl) iminodiacetic acid
HIPDM *N,N,N*-trimethyl-*N*-′-[2-hydroxyl-3-methyl-5-iodobenzyl]-1,3-propanediamine
HK hexokinase
HMDP hydroxymethylene diphosphonate
HRP horseradish peroxidase
HTLV human T-cell leukemia virus
IMP *N*-isopropyl-*p*-iodoamphetamine
IAH immunoadherence hemagglutination
IEM immunoelectron microscopy
IFA indirect immunofluorescent staining
IRMA immunoradiometric assay
KG α-ketoglutarate
LAO left anterior oblique
LDH lactate dehydrogenase
LFOV large field of view
MAA macroaggregates of albumin
MB CPK isoenzyme
MDH malate dehydrogenase
MDP methylene diphosphonate
MIRD Medical Internal Radiation Dose
MM one of three main iso-enzymes of creatine kinase
MTB methylthymol blue
NAD nicotinamide–adenine dinucleotide
NADH reduced NAD

NADP	nicotinamide – adenine dinucleotide phosphate
NADPH	reduced NADP
NAPH	*N*-acetylprocainamide
NCCLS	National Committee for Clinical Laboratory Standards
NI	neutralization test
OCPC	*o*-cresolphthalein
PAP	prostatic acid phosphatase
PEP	phosphoenolpyruvic acid
PEP-C	phosphoenolpyruvate carboxylase
PET	positron emission tomography
P_i	inorganic phosphate
PMA	pre-market approval application
PNB	*p*-nitrobenzenediazonium tetrafluoroborate
PT	prothrombin time
PTT	partial thromboplastin time
RBC	red blood cell
RE	reticuloendothelial
RIA	radioimmunoassay
RID	radioimmunodiffusion
RNA	ribonucleic acid
SFOV	standard field of view
SLFIA	substrate-labeled fluorescent immunoassay
SPB	supplemented peptone broth
SPECT	single photon emission computed tomography
SPS	sodium polyanetholesulfonate
STAT	medical term meaning "immediate"
TEG	thromboelastogram
TRCHII	tanned red cell hemagglutination inhibition immunoassay
TSB	trypic or trypticase soy broth
TSH	thyroid-stimulating hormone

1. Introduction

Diagnostic reagents can be defined broadly as those chemicals used for the in vitro and in vivo testing, identification, or monitoring of a disease. Although their composition depends on the disease being investigated, these reagents can be composed of a wide variety of materials, ranging from solutions of simple inorganic chemicals to freeze-dried powders or suspensions of complex biologic molecules. Diagnostic reagents are manufactured by many companies worldwide and are packaged in numerous forms and devices for either manual or automated use. These reagents are most often employed in the general professional fields of clinical chemistry, clinical biochemistry, and pathology, as well as in other areas such as medical microbiology, serology, radiology, blood banking, and hematology. Their applications extend from medical research to patient care and even to patient self-testing in the home.

This article describes diagnostic reagents and their applications in general clinical chemistry, immunodiagnostics, microbiology, coagulation, and nuclear medicine. Inorganic reagents [1] and organic solvents [2] used routinely in analytical tests, together with older techniques for blood typing and histology [3], are not described, and the reader is referred to the appropriate references (→ Blood). Flow cytometry, thin-film reagent testing devices, two-dimensional electrophoresis, and immunoelectrophoresis are also not included.

1.1. Medical Significance

Because of their effective use in medical devices, diagnostic reagents have both improved the quality of health care and expanded knowledge of disease. Prior to 1963, most diagnostic reagents were either formulated in clinical laboratories where they were used, or purchased directly from a few chemical supply houses and pharmaceutical companies. The reagents were often used in manual semiquantitative assays or in laborious quantitative methods for determining certain enzymes and metabolites. Quantitative tests generally needed to be done one at a time by using simple spectrophotometers. A few test kits were available for measurement of simple analyte concentrations. However, even these tests were semiquantitative and labor intensive.

The advent of laboratory automation began in 1963, when Technicon Corp. first introduced an *automated clinical analyzer,* based on the continuous flow of reagents, which could perform multiple tests on each sample. Other automated systems, both discrete and continuous flow, have since been manufactured and have changed the way all clinical laboratories operate. Over 60 different tests can now be performed on some automated systems. *Manual kits* are also available for most routine clinical tests, and unlike their predecessors, tests can be run conveniently

and accurately on many samples at the same time by using centrifugal analyzers or multichannel spectrophotometers.

1.2. Economic Significance

In the United States alone, ca. $\$200 \times 10^9$ was spent in 1979 for health care, and over 5.7×10^9 laboratory tests were performed [4]. About 10 % of the total health care costs was attributed to testing, and 5 – 15 % of this was for diagnostic reagents. These numbers have increased during the 1980s, although not as rapidly as during the 1970s, and growth should continue because of the introduction of new diagnostic methods and reagents. The market for diagnostic reagents based on new developments in biotechnology, such as monoclonal antibodies and DNA probes, is expected by some investors to top $\$500 \times 10^6$ by 1987 and to grow to $> \$1 \times 10^9$ during the 1990s [5]. Factors slowing the growth rate of the diagnostic testing market include new legislation in the United States aimed at reducing Medicare payments, a decline in the birth rate, shortened hospital stays, increased cost consciousness by patients and physicians, and increased costs of developing new reagents and methods.

1.3. Regulatory Controls

The introduction of new diagnostic reagents is tightly controlled in the United States under the Federal Food, Drug, & Cosmetic Act of 1938, which was amended in 1976 to include medical devices. All medical devices marketed prior to May 28, 1976, were required by law to be classified by the Food and Drug Administration (FDA) into one of three categories [6]:

1) Class I: devices requiring general controls
2) Class II: devices requiring performance standards
3) Class III: devices for which insufficient information exists for developing controls or performance standards to guarantee safety and effectiveness

In order to sell a novel device classified in Class III, a manufacturer must first submit a premarket approval application (PMA) containing sufficient data to establish its safety and effectiveness. In designing new reagents and test kits, manufacturers may also follow a number of voluntary guidelines established by the American Association for Clinical Chemistry (AACC) [7] and the National Committee for Clinical Laboratory Standards (NCCLS) [8].

Another organization that supplies guidelines, as well as control materials for checking performance of laboratory tests, is the College of American Pathologists (CAP) [9]. Many state agencies now require laboratories to use these control materials to remain accredited.

1.4. Classifications of Diagnostic Reagents

In this review, diagnostic reagents have been classified into three broad categories on the basis of their applications. They are as follows: (1) reagents used to perform general clinical chemistry tests, including the quantitation of electrolytes, serum proteins, enzymes, therapeutic drugs, hormones, and tumor markers (Chap. 2), (2) reagents used to identify microorganisms (Chap. 4), and (3) reagents used to diagnose disorders of the coagulation pathways (Chap. 5). The following two additional categories are based on reagent technology: (4) immunodiagnostic reagents, including both isotopic and nonisotopic materials (Chap. 3), and (5) in vivo reagents for nuclear medicine, including imaging and radiopharmaceuticals (Chap. 6), (→ Radiopaque Media).

Traditionally, microbiologic tests, coagulation tests, and radioisotopic tests have been performed in different laboratories from those in which general clinical chemistry methods are used. This segmentation of effort within hospitals is disappearing as more in vitro diagnostic reagents that bridge the various medical disciplines are developed.

2. Reagents for General Clinical Chemistry

2.1. STAT and Routine Testing

In general, all tests performed in clinical laboratories can be further divided into two operational categories, depending on the priority

of patient treatment. The *STAT tests* are those that typically must be performed quickly in an emergency. These tests usually include analyses of electrolytes (sodium, potassium, chloride, and hydrogen carbonate), creatinine, urea, glucose, and creatine kinase. Results from *routine tests* are not needed as quickly. Examples include methods for measuring serum proteins, hormones, most enzymes, and some therapeutic drugs. Because STAT tests have priority, these assays should be less time consuming and less labor intensive than routine methods. Reagents for STAT methods are ideally more stable and easier to use than others, so that the frequency of calibration is reduced and the time to the first result is minimized.

2.2. Quantitation of Electrolytes

Sodium, potassium, hydrogen carbonate, and chloride are the four most commonly measured ions in serum. Usually, all four are measured on a single sample from the patient because of the important relationships among the ions in terms of acid–base and cation–anion balance. *Ion-selective electrodes* are commonly used to determine concentrations of these ions. However, simple spectrophotometric assays are available for chloride and hydrogen carbonate.

Hydrogen Carbonate. All forms of carbon dioxide (dissolved CO_2, HCO_3^-, and CO_3^{2-}) can be converted to the hydrogen carbonate anion with an alkaline buffer. Hydrogen carbonate can then be accurately *measured enzymatically* [10], using phosphoenolpyruvate carboxylase (PEP-C) (E.C. 4.1.1.32) [*9013-08-5*] in the presence of excess phosphoenolpyruvate ion [*138-08-9*] (PEP) toform oxalacetate ion and inorganic phosphate (P_i). The oxalacetate ion thus formed can be reduced to malate ion by malate dehydrogenase (MDH) with the simultaneous oxidation of reduced nicotinamide–adenine dinucleotide (NADH). The change in absorbance at 340 nm caused by the disappearance of NADH is directly proportional to the total amount of hydrogen carbonate in the sample.

Sodium and Potassium. Measurements of sodium and potassium are most readily performed by using *ion-selective electrodes* that have specific ionophores, such as valinomycin, embedded in the membrane layer. Concentrations of sodium and potassium are calculated from the electrode potential, using the Nernst equation [11–13].

Chloride. The chloride ion, like sodium, is primarily found in extracellular fluids. Hypochloremia (i.e., decreased plasma or serum chloride) may be caused by prolonged vomiting, chronic diarrhea, diabetic ketoacidosis, lactic acidosis, adrenal disease, and renal failure. Hyperchloremia may be caused by hyperventilation, large doses of potassium chloride, and dehydration.

Although there are numerous methods for measuring chloride concentration, including the use of ion-selective electrodes, one easy *spectrophotometric assay* is based on the reaction of chloride with excess mercury(II) thiocyanate to form mercury(II) chloride and thiocyanate, which further reacts with iron(III) ions from iron(III) nitrate [14]. The absorbance at 452 nm caused by formation of red iron(III) thiocyanate is directly proportional to the chloride concentration in the sample.

$$2\,Cl^- + Hg(SCN)_2 \longrightarrow HgCl_2 + 2\,SCN^-$$
$$SCN^- + Fe^{3+} \longrightarrow Fe(SCN)^{2+}$$

2.3. General Chemistry Tests

Typical analyses performed in clinical chemistry laboratories include tests for the concentration of albumin, total protein, calcium, bilirubin, cerebrospinal fluid protein, glucose, cholesterol, high-density lipoprotein (HDL), iron, magnesium, phosphorus, triglycerides, uric acid, and urea. All of the methods for measuring these materials are now automated. Reagent kits are also available for performing these tests manually.

Human serum albumin is commonly measured by its reaction with *bromocresol green* [*76-60-8*] to form a colored complex that absorbs light at 600 nm [15]. Methods with greater accuracy have been developed [16], [17], such as using *bromocresol purple* [*115-40-2*] with a low pH buffer; other determinations rely on either *immunoprecipitation* or *electrophoresis*. Overall, agreement is good between these approaches.

Total protein measurements have been traditionally done by using *biuret* [*108-19-0*] [18] or *Lowry* [19] *reagents,* which form soluble blue complexes. Copper(II) sulfate reacts with peptide linkages of proteins in alkaline solution; the resulting complexes, measured at 540 nm, are proportional to the amount of protein in the sample. A commonly accepted reference method for total protein is available [20] as are numerous other methods.

Calcium [*7440-70-2*] **magnesium** [*7439-95-4*]**, and iron** [*7439-89-6*] can be measured by using *organic dyes,* namely *o*-cresolphthalein [*596-27-0*] (OCPC), methylthymol blue [*1945-77-3*] (MTB), and 4,7-diphenyl-1,10-phenanthroline (bathophenanthroline) sulfonate (BPS), respectively. *Calcium* reacts with OCPC to form a purple complex absorbing at 577 nm, the concentration of which is proportional to the calcium level in serum [21]. Magnesium ions interfere and are removed by chelation with 8-quinolinol.

Likewise, *magnesium* ions form a blue complex with MTB [22], and calcium interference can be minimized by adding the barium salt of ethylenebis(oxyethylenenitrilo)tetraacetic acid (EGTA). Both of these methods are easy to perform and can be completed in < 5 min.

Under acidic conditions (pH 4.2), the protein transferrin can be reduced with hydroxylamine. The *iron(II)* that is bound to this protein is released and forms a red complex with BPS, which absorbs at 540 nm [20, p. 680].

Phosphorus [*7723-14-0*] is usually determined by using a method in which a mixture of 4-methylaminophenolsulfate and bisulfite are used to reduce 12-molybdophosphoric acid [23], [24]. The product of this reaction absorbs at 340 nm and is proportional to the level of inorganic phosphorus in the sample.

Bilirubin [*635-65-4*] . The techniques available for measuring bilirubin include *direct absorbance* [25], *reaction with diazonium salts* [26], and *liquid chromatography.* The most commonly employed methods [27], [28] have been modified extensively by instrument manufacturers. Bilirubin reacts with *p*-nitrobenzenediazonium tetrafluoroborate (PNB) to form a red complex that absorbs at 540 nm. Surfactants, such as Tween 20, can be added to obtain measurements of total bilirubin; in the absence of surfactants the absorbance is more indicative of conjugated bilirubin (attached to glucosyl residues).

Cerebrospinal fluid protein (CFP) is often elevated if the brain or spinal cord are diseased. This material is readily measured by adding *trichloroacetic acid* [*76-03-9*] and determining the resulting turbidity of the protein precipitate at 340 nm [20, p. 423]. Methods using sulfosalicylic acid [29] are also available and generally give similar results.

Glucose [*50-99-7*] . Analysis for glucose is perhaps the most frequently performed assay in the clinical laboratory. Analyses are usually based on well-known enzymatic procedures that use either *glucose oxidase* (E.C. 1.1.3.4) [*9001-37-0*] or *hexokinase* (E.C. 2.7.1.1) [*9001-51-8*] [30]. The former approach relies on the production of hydrogen peroxide, which can react with peroxidase and a colorimetric substrate, whereas the latter depends on the coupling of the hexokinase (HK) reaction product with another enzyme, glucose 6-phosphate dehydrogenase (G6PDH) (E.C. 1.1.1.49) [*9001-40-5*]. Thephosphorylation of glucose by HK in the presence of adenosine 5′-triphosphate(ATP) produces glucose 6-phosphate (G6P). As G6P is consumed by excess G6PDH, nicotinamid – adenine dinucleotide phosphate (NADP) is reduced, producing 6-phosphogluconolactone and NADPH, which has an absorbance at 340 nm that is proportional to the initial concentration of glucose.

Cholesterol [*57-88-5*] test methods are similar to those for glucose and are primarily based on *enzymatic procedures* [31], [32]. Cholesterol esterase (CE) (E.C. 3.1.1.13) [*9026-00-0*] catalyzes the hydrolysis of cholesterol esters to produce free cholesterol, which, along with preexisting free cholesterol, is oxidized in a reaction catalyzed by cholesterol oxidase (CO) (E.C. 1.1.3.6) [*9028-76-6*] to form cholest-4-ene-3-one and hydrogen peroxide. In the presence of horseradish peroxidase (HRP) (E.C. 1.11.1.7) [*9003-99-0*], the peroxide is used

to oxidize chromogenic substrates, such as *N,N*-diethylaniline (DEA) and 4-aminoantipyrine (AAP). The absorbance caused by the colored products is directly related to the cholesterol concentration of the sample. Because **high-density lipoproteins** (HDL) contain cholesterol, levels of HDL can also be estimated accurately by first isolating the fraction of serum proteins containing them and then using the cholesterol methods. At pH 5.7, low-density and very low-density lipoproteins are selectively precipitated from serum by adding buffered phosphotungstate. The remaining clear fraction containing HDL can be decanted and assayed directly.

Triglycerides can be quantitated enzymatically in serum by first using lipase (E.C. 3.1.1.3) [*9001-62-1*] to produce glycerol and then coupling this product to glyceroldehydrogenase (E.C. 1.1.1.29) [*9028-37-9*] [33].The latter enzyme simultaneously reduces NAD to NADH, which absorbs at 340 nm.

Uric acid [*69-93-2*] , which absorbs *light* at 293 nm, may be measured by various procedures. However, one convenient method is to follow the loss of 293 nm absorbance as the uric acid is enzymatically converted to allantoin by uricase (E.C. 1.7.3.3) [*9002-12-4*] [20, p. 538]. The rate of change of absorbance is directly proportional to the concentration of uric acid in serum.

Urea [*57-13-6*] (**1**), or blood urea nitrogen (BUN), is important for diagnosing kidney disease and can be determined by coupling two enzymes [34], namely, urease (E.C. 3.5.1.5) [*9002-13-5*] and glutamate dehydrogenase(GLDH) (E.C. 1.4.1.3) [*9029-12-3*]. Urease specifically hydrolyzes urea to form ammonia and carbon dioxide; the ammonia is a substrate for GLDH, which reductively aminates α-ketoglutarate (KG) ion. In the process, NADH is oxidized to NAD with a loss of absorbance at 340 nm.

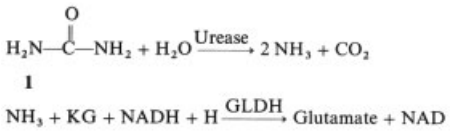

2.4. Quantitation of Serum Proteins and Enzymes

Specific serum proteins , such as the immunoglobulins IgG, IgM, IgA, IgD, and IgE, are important for immunity to infections (→ Blood). Abnormalities in plasma IgG, IgA, and IgM occur in many diseases; however, the most noted has been *multiple myeloma,* which results from the rapid uncontrolled growth of antibody-producing cells. A malignant plasma cell tumor may produce any one of the five classes of immunoglobulins, but ca. 50 % of the tumor cells produce IgG and 25 % IgA. Another 20 % produce only single polypeptide chains, usually short chains, which can be detected in the urine as Bence–Jones proteins. Rarely do tumors produce IgD or IgE. *Hypo-γ-globulinemias*, or decreases in immunoglobulins, are usually linked either to hereditary deficiencies in infants or to certain infectious diseases in adults.

Immunoglobulins can be measured turbidometrically, using immunoprecipitation techniques. In *turbidometric* approaches, polyclonal antiserum to human IgG or another immunoglobulin class is prepared from another species (usually rabbit or goat) and used as the main reagent. The pH of the reaction mixture is usually maintained at ca. 7; as the sample and main reagent are mixed, aggregates form between antibodies and human immunoglobulin antigens. The degree of aggregation, measured by the turbidity at 340 nm, is related to the concentration of immunoglobulin in the sample. Antisera for these methods must be chosen carefully to avoid cross reaction by immunoglobulins in other classes and to give reproducible standard curves.

Enzymes in serum or plasma are typically measured to determine the status of organ function because circulating concentrations are usually elevated by diseases affecting specific organs. Lipase and amylase are elevated during episodes of acute pancreatitis. Certain isoenzymes, like creatine kinase-MB, are measured to aid in the diagnosis of myocardial infarction, and different isoenzymes of lactic dehydrogenase are elevated in many disorders, including liver disease and myocardial infarction. Acid phosphatase is used as a tumor marker.

The serum enzyme concentrations of *lipase* and *amylase* usually parallel each other and are often analyzed together on the same sample. Like almost all enzyme methods, these are measured by monitoring the rate of enzymatic reaction either spectrophotometrically or turbidometrically after adding the appropriate enzyme substrate. Substrate levels are typically held well above the Michaelis constant (K_m) of the enzyme to minimize errors resulting from variations in the amount of substrate added [35].

Lipase (E.C. 3.1.1.3) [*9001-62-1*] can be measured directly by using a triolein emulsion, which loses turbidity when it is hydrolyzed to glycerides and fatty acids [36]. Amylase (E.C. 2.4.1.1) [*9035-74-9*], on the other hand, is usually measured with maltopentaose (a linear polymerof five glucose units in an α-1,4-linkage) as a substrate in a reaction mixture containing at least one other enzyme, α-glucosidase (GD), which further breaks down the products of the amylase reaction to glucose. Glucose can then be detected spectrophotometrically with the simultaneous addition of three enzymes, GD, hexokinase (HK), and glucose 6-phosphate dehydrogenase (G6PDH), as well as adenosine 5′-triphosphate (ATP) and nicotinamide–adenine dinucleotide (NAD).

$$\text{Maltopentaose} \xrightarrow{\text{Serum amylase}} \text{Maltotriose} + \text{Maltose}$$
$$\text{Maltotriose} + \text{Maltose} \xrightarrow{\text{GD}} 5\ \text{Glucose}$$
$$5\ \text{Glucose} + 5\ \text{ATP} \xrightarrow{\text{HK}} 5\ \text{G6P} + 5\ \text{ADP}$$
$$5\ \text{G6P} + 5\ \text{NAD} \xrightarrow{\text{G6PDH}} 5\ \text{6-Phosphogluconolactone} + 5\ \text{NADH (absorbs at 340 nm)}$$

Creatine kinase (E.C. 2.7.3.2) [*9001-15-4*], (CK) which catalyzes the transfer of phosphate from creatine phosphate to adenosine 5′-diphosphate(ADP), is also determined by using a method that requires complex reagents. As in the case of amylase, both hexokinase and glucose 6-phosphate dehydrogenase are employed; however, ATP becomes the intermediate reactant and an excess of G6P is added together with NAD [37]. To keep the reagents from reacting before they are needed, they are segregated into convenient solutions or packaged in tablet form. In most methods for CK, adenosine monophosphate and 1,5-di(adenosine 5-pentaphosphate) are added to inhibit adenylate kinase, and ethylenediaminetetraacetic acid (EDTA) is added to suppress calcium inhibition of creatine kinase activity. A sulfhydryl reagent, like dithioerythritol, is also needed to activate the kinase.

$$\text{Creatine phosphate} + \text{ADP} \xrightarrow{\text{CK}} \text{Creatine} + \text{ATP}$$
$$\text{Glucose} + \text{ATP} \xrightarrow{\text{HK}} \text{G6P} + \text{ADP}$$
$$\text{G6P} + \text{NAD}^+ \xrightarrow{\text{G6PDH}} \text{6-Phosphogluconolactone} + 5\ \text{NADH (absorbs at 340 nm)}$$

The method used to measure CK can also be applied to *CKMB,* the MB isoenzyme of creatine kinase. However, CKMB must first be isolated from the two other main isoenzymes of CK, i.e., MM and BB. This can be done with small ion – exchange columns containing cationic and anionic resins, e.g., Biorex 70 and DEAE-sephadex. An immunoinhibition method for CKMB, which actually measures both MB and BB, is available; this method does not require a column separation step. A two-site sandwich immunochemical method, in which the mass rather than the activity of the isoenzyme is measured, is also on the market. All of these approaches give clinically acceptable performance.

Lactate dehydrogenase (E.C. 1.1.1.27) [*9001-60-9*] catalyzes the interconversion of lactic and pyruvic acids and, like creatine kinase, has several isoenzymes that are designated from LDH_1 to LDH_5 by their typical electrophoretic patterns. Liver tissue produces a very high proportion of LDH_5, whereas heart tissue produces a high proportion of both LDH_1 and LDH_2. Total levels of lactate dehydrogenase in serum can be measured by following the rate of absorbance change at 340 nm after adding two reagents, L-lactic acid and oxidized NAD [38]. The liver form of the enzyme (LDH_5) can be measured similarly after its separation from other isoenzymes by using DEAE-sephadex [39]. Another substrate, α-ketobutyrate, works well for the LDH_1 isoenzyme, but not for other forms, and can be used together with reduced NADH to measure LDH_1 activity [40]. When measured in this fashion, the enzyme is often referred to as β-hydroxybutyrate dehydrogenase.

Acid phosphatase [*9001-77-8*] also has numerous isoenzymes; however, only prostatic acid phosphatase is clinically significant because its levels are elevated only during carcinoma of the prostate gland. Although newer methods for measuring this isoenzyme are based on radioimmunoassays and immunoenzymo-

metric assays many laboratories still use a total acid phosphatase method for speed and convenience. Total acid phosphatase can be measured under acidic conditions by adding thymolphthalein monophosphate as the substrate and monitoring the rate of hydrolysis at 600 nm [41].

Serum concentrations of *alkaline phosphatase* (E.C. 3.1.3.1) [*9001-78-9*] are usually elevated by either increased osteoblast activity, leading to the formation of new bone, or by liver disease, in which the bile ducts are obstructed. Because specific isoenzymes of alkaline phosphatase are rarely measured, the methods for determining total concentration of enzyme, using 4-nitrophenylphosphate as a chromogenic substrate, are typically straightforward. As this substrate is hydrolyzed by the enzyme at pH 10, absorbance at 405 nm increases [42]. A variety of buffers have been used with this method; the most preferred is 1-aminopropanol, which acts as a phosphate acceptor and prevents inhibition of the enzyme by phosphate.

Aspartate aminotransferase (E.C. 2.6.1.1) [*9000-97-9*], also known as glutamic oxalacetic transaminase (GOT), is found largely in the liver and heart and is responsible for catalyzing the formation of glutamate ion from α-ketoglutarate by transferring an amino group from aspartate ion. Together with other enzymes, such as creatine kinase, lactic dehydrogenase, and alanine aminotransferase, this enzyme can give useful diagnostic information. Myocardial infarction as well as diseases affecting the liver, such as infectious hepatitis, mononucleosis, cirrhosis, and obstruction of the bile duct, may elevate levels of GOT in the serum. The activity of this enzyme can be conveniently determined by coupling one enzyme product, oxalacetate, to MDH. In doing so, the rate of NAD formation, or loss of absorbance at 340 nm, is proportional to the level of GOT [43].

$$\text{L-Aspartate} + \alpha\text{-Ketoglutarate} \xrightarrow{\text{GOT}} \text{L-Glutamate} + \text{Oxalacetate}$$

$$\text{Oxalacetate} + \text{NADH} \xrightarrow{\text{MDH}} \text{Malate} + \text{NAD}$$

(absorbs at 340 nm)

Alanine aminotransferase (E.C. 2.6.1.2) [*9000-86-6*] , which is also called glutamic pyruvic transferase (GPT), and GOT are usually measured together because changes in concentration of the two enzymes frequently parallel each other. However, GPT typically does not become elevated during a myocardial infarction. Like GOT, this enzyme can be coupled to a dehydrogenase (LDH) through the intermediate pyruvate ion [44].

With most serum enzymes that are measured in clinical medicine, increased levels signify the possibility of disease. With *pseudocholinesterase* [*9001-08-5*], inactivity is usually of interest. This enzyme catalyzes the hydrolysis of choline esters in blood, and its activity can be lowered by organophosphorus poisoning, succinyldicholine hypersensitivity, and liver disease. Because the enzyme can hydrolyze butyrylthiocholine to butyric acid and thiocholine, measurements can be done spectrophotometrically by using 2,6-dichlorophenolindophenol, which reacts with thiocholine [45]. As the reaction proceeds, absorbance is lost at 600 nm.

2.5. Therapeutic Drug Monitoring

Since the mid 1960s, the rational use of therapeutic drugs to combat infection and ease suffering from a variety of diseases has increased dramatically [46]. Drugs, such as digoxin, theophylline, phenobarbital, phenytoin, and aminoglycoside antibiotics, lead the list of agents most often monitored and now account for ca. 25 % of all tests ordered from the clinical laboratory (see Table 1). Monitoring allows physicians to adjust doses, to meet the needs of individual patients, to be certain the drugs are being taken and in therapeutic amounts, and to be sure the levels are below those in which adverse reactions are seen. Many drugs have a narrow range between therapeutic and toxic effects and need to be monitored periodically to assure a safety margin. Other drugs are ineffective unless given at the proper dose; they may have a wider range between the therapeutic and toxic levels. Antibiotics, like gentamicin and tobramycin, may severely damage the kidneys if given in excess.

Methods for measuring the concentration of these drugs by radioimmunoassay (RIA), enzyme immunoassay (EIA), fluorescence immunoassay (FIA), and latex agglutination have been commercialized and are displacing the older analytical techniques [47]. For a detailed

discussion of these new techniques and reagents, see Sections 3.3 and 3.4.

Table 1. Drugs commonly measured in clinical chemistry

Major use	Drug	General therapeutic levels, mg/L
Antiarrhythmic	digoxin	0.001 – 0.002
	lidocaine	1.5 – 5.0
	procainamide	4.0 – 10
	N-acetylprocainamide	4.0 – 10
	quinidine	2.0 – 5.0
Antiasthmatic	theophylline	10 – 20
Anticonvulsant	carbamazepine	3.0 – 9.0
	ethosuxide	40 – 100
	phenobarbital	10 – 25
	primidone	5 – 10
	phenytoin	10 – 20
	valproic acid	∼50 – 100
Antibacterial	gentamicin	2 – 10
	tobramycin	4 – 10

2.6. Measurement of Hormones and Tumor Markers

The development of radioimmunoassay (RIA) technology, with high sensitivity and specificity, has led to many methods for measuring hormones and tumor markers that otherwise would have been impractical. The main reason is that these biologic molecules are present in extremely low concentration (ranging from 10^{-16} to 10^{-9} M) and, thus, are beyond the limits of chromatographic techniques.

Originally, RIA techniques were developed for determining high molecular mass peptides, using competitive binding between labeled antigen and unlabeled antigen from the sample with a predetermined concentration of polyclonal antibody. These methods were subsequently commercialized for low mass haptens, such as thyroxine and cortisol, and a newer technique known as the immunoradiometric assay (IRMA) was developed in the late 1970s for measuring the concentration of proteins. Methods based on IRMA technology give much higher selectivity for the specific analyte compared to RIA methods and employ two monoclonal antibodies, one bound to a solid support and the other isotopically labeled, usually with iodine–125. A sandwich immunoenzymometric assay uses the same approach but with an enzyme covalently bonded to the second antibody (for details, see Sections 3.2 and 3.4).

Although a number of *tumor markers* have been claimed in the research literature, only three are ordered frequently by physicians. These are (1) carcinoembryonic antigen (CEA) (2) prostatic acid phosphatase (PAP), and (3) α-fetoprotein (AFP). On the other hand, the number of *hormones* frequently ordered is considerably larger and includes human chorionic gonadotropin (HCG), thyroid-stimulating hormone (TSH), follicle-stimulating hormone (FSH), human growth hormone (HGH), prolactin, luteinizing hormone, insulin, adrenocorticotropic hormone (ACTH), parathyroid hormone, testosterone, progesterone, aldosterone, cortisol, estriol, and thyroxines (T 3 and T 4). The competitive RIA techniques can also be used to measure other low-concentration substances, such as renin, gastrin, cyclic adenosine 3,5-monophosphate (cyclic AMP), and digoxin.

3. Immunodiagnostic Reagents

Sales of reagents used in immunoassays to diagnose disease and to measure therapeutic agents have been growing exponentially since 1965. This accelerating pace can be attributed to advances in medical technology that have led to the widespread use of therapeutic drugs and to the development of new and convenient in vitro tests for measuring substances. Immunodiagnostic reagents are now used in almost all blood banks to screen samples for infectious disease; in most hospitals, clinics, and commercial laboratories to measure therapeutic drugs, hormones, and tumor markers; and in some physician's offices to monitor levels of certain drugs and to test for pregnancy. These reagents are also part of a growing market for home testing, as exemplified by over-the-counter pregnancy tests that rely on visual detection. Increasingly, methods that use these reagents are becoming both more technologically sophisticated and easier to use because of their packaging, use of nonisotopic markers, and better defined characteristics.

3.1. Polyclonal and Monoclonal Antibodies

Historically, most immunoassays have been developed by using heterogeneous mixtures of an-

tibodies obtained from the serum of animals (usually rabbits, goats, sheep, and horses) immunized with the antigen of interest. These polyclonal antibodies are produced by numerous B cells within the host animal and can differ from each other in terms of their relative concentrations, affinities for the antigen, and recognition of antigen binding sites. The polyclonal antibody populations are also different from one animal to another, which leads to problems in preserving assay performance when one lot of antiserum is depleted.

These concerns can be largely circumvented by using monoclonal antibodies, which were first prepared in 1975 [48], (→ Monoclonal Antibodies). Monoclonal antibodies used in diagnostic assays are typically produced by removing cells from the spleens of immunized mice, fusing these cells with a myeloma cell line, and growing the resulting hybridoma cells in culture. The hybridoma cells are usually screened repeatedly to select those cells producing the desired antibody; the selected cells are then injected into the peritoneal cavities of other mice to produce gram quantities of the specific antibody. Because only one cell line is used to produce the desired antibody, the immunodiagnostic reagent has inherently more uniform properties of affinity and antigen recognition than antibody produced from a combination of cell lines. In addition, monoclonal antibody producing cells can be frozen for long-term storage, reducing lot-to-lot variations. Monoclonal antibodies are now being used in vitro for both radioimmunoassays and nonisotopic immunoassays, as well as in vivo for immunoscintigraphy.

3.2. Radioimmunodiagnostic Reagents

Soon after the first radioimmunoassay (RIA) technique was developed in 1959 [49], there was an abundance of publications regarding new RIA methods for measuring hormones, metabolites, tumor markers, drugs, and a variety of other substances found in low concentration. These early methods typically relied on immunoprecipitation techniques and used one polyclonal antibody reagent per test, together with a radioisotope-labeled antigen. Many of these tests are still being performed, but they have been modified to use other separation techniques and monoclonal antibodies. Radioassays that use binding agents other than antibodies have been commercialized, and many methods now use two or more antibodies per test. The worldwide market for radioimmunodiagnostic reagents (using iodine–125, carbon–14, phosphorus–32, and tritium as labels) has grown to $\$ 1 \times 10^9$/a.

3.2.1. Radioimmunoassays

The first RIA methods were based on *competitive binding reactions* between radioisotope-labeled antigens (such as iodine–131-labeled insulin) and sample antigens for antibody binding. Generally, the resulting immune complexes were precipitated, and the amount of label determined in either the liquid or the precipitate. To obtain well-defined precipitates more quickly, this technique was modified, e.g., by adding ammonium sulfate and a second antibody, such as antirabbit immunoglobulin. Techniques were also developed that used charcoal, which quickly adsorbed any unbound low molecular mass antigens but not antigen bound to antibodies [50].

Most radioimmunoassays employ a *solid phase* that has the specific antibody attached through either covalent bonds or adsorption. The solid phase, such as a coated test tube, is exposed to the sample antigen together with a defined amount of labeled antigen. After an incubation period (ranging from minutes to hours), the liquid in the test tube is decanted and counted to determine the amount of unbound radioactivity. These tests are most often performed on batches of samples and are now being automated [51]. In automated systems, both the bound and free fractions can be measured for radioactivity, which gives an additional check on assay performance.

3.2.2. Radioisotopic Dilution Assays

Methods using this technique are similar to radioimmunoassays in design except that the antibodies are replaced with other binding agents, such as intrinsic factor for measuring vitamin B_{12} and milk-binding protein for measuring folate. These binding agents must be purified from biologic tissues and are now commercially available. Many of these tests also require a step in which the analyte is first separated from an

endogenous-binding protein by boiling the sample or by adding a denaturing agent like sodium hydroxide [52].

3.2.3. Immunoradiometric Assays

An immunoradiometric assay (IRMA) differs from a conventional RIA technique in that *two antibodies are employed;* each recognizes a different binding site on the antigen. One antibody (Ab_1) is immobilized on a solid support and the other (Ab_2) is tagged with ^{125}I.

Immobilized Ab_1 + Ag $\longrightarrow$ Immobilized Ab_1–Ag

Immobilized Ab_1–Ag + Ab_2–^{125}I $\longrightarrow$ Immobilized Ab_1–Ag–Ab_2–^{125}I

These assays do not depend on competitive reactions, but instead are performed in multiple steps:

The sample is exposed to the solid phase (immobilized Ab_1), incubated, washed, and then reacted with the second antibody–tag (Ab_2–^{125}I). After another washing step, the amount of radioactivity bound to the support is measured by using a γ-counter and is directly proportional to the concentration of antigen present in the sample.

Methods that use monoclonal antibodies are becoming more popular for measuring proteins with multiple subunits, such as ferritin, CKMB, IgE, TSH, and B-HCG, and are more specific for the antigen than methods that use only one antibody [53].

3.3. Reagents for Nonisotopic Homogeneous Immunoassays

Most therapeutic drugs and substances that are in the nanomolar concentration range or higher can be measured by using nonisotopic homogeneous immunoassays. These techniques have become popular because they give rapid results, use inexpensive reagents, and circumvent the safety concerns regarding disposal of radioisotopes. In addition, they can be readily automated because no separation step is required.

3.3.1. Latex Agglutination

The first practical homogeneous immunoassay was based on latex agglutination [54]. Micrometer-size polystyrene latex particles coated with adsorbed immunoglobulin were used and were quickly agglutinated by the rheumatoid factor. Latex agglutination assays have since been developed with both antibody- and antigen-coated particles and have been used to measure a wide variety of clinically important substances.

Although most of the early latex agglutination tests were semiquantitative, many have now been developed that can be automatically *monitored* by turbidity, nephelometry, lightcattering, and particle-counting detectors. These tests are generally as quantitative as any state-of-the-art homogeneous immunoassays. The analysis has been stabilized by covalently bonding, rather than adsorbing, antigens and antibodies to latex particles coated with reactive shells. Different polymers with high refractive indices have been used to improve the signal, and in addition, particles of defined size have been synthesized [55]. Agglutination rates are faster for particles in the submicrometer size range, which can be important when designing tests for automated analyzers.

Agglutination tests for the concentration of C-reactive protein, fibrin degradation products, human chorionic gonadotropin, and other proteins usually use particles coated with an excess of a specific antibody. To reduce interference from human rheumatoid factors, either the antibody can be first treated with pepsin to remove the rheumatoid factor binding site, or an excess of rabbit γ-globulin can be added during the assay. Antigen-coated particles are typically employed in assays for low molecular mass substances, such as the therapeutic drugs. In the absence of drug, particles agglutinate when a specific antibody is added, while in the presence of drug, the amount and rate of antibody-induced agglutination is decreased. Drugs, or their analogues that react with the antibody, can be either directly bound to the particle surface, attached through protein conjugates, such as bovine serum albumin, or linked through synthetic spacer molecules, such as polyoxyethylene–polyamine. The choice of attachment chem-

istry is often important for optimal performance and reagent stability.

3.3.2. Enzyme-Tagged Immunoassays

Prior to the advent of quantitative agglutination tests, an enzyme-based immunoassay was developed by the Syva Co. which used a simple homogeneous assay protocol which could be followed by laboratory technicians without special training. As a result, this *enzyme-multiplied immunoassay technique* (EMIT) quickly gained acceptance and is now used routinely for measuring the concentration of therapeutic drugs. The test relies on an enzyme, usually glucose 6-phosphate dehydrogenase (E.C. 1.1.1.49) [*9001-40-5*], to which the drug of interest is covalently bonded nearthe active site. This antigen– (or drug–)enzyme conjugate retains activity until bound by an antibody and, thus, can be used as a marker for the amount of antibody not bound by the sample antigen [56]. Rates of antigen–enzyme conjugate activity are conveniently measured at 340 nm by following NADH production and are inversely proportional to the amount of sample antigen present.

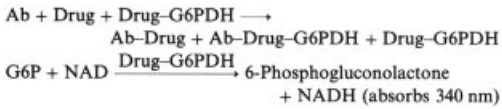

Another homogeneous enzyme-based immunoassay that is used routinely has been developed by the Ames Division of Miles Laboratories. This assay, called the *apoenzyme reconstitution immunoassay system* (ARIES), requires that the antigen be covalently attached to the prosthetic group of an enzyme, such as glucose oxidase (E.C. 1.1.3.4) [*9001-37-0*]. For measuring theophylline levels, an analogue of the drugis covalently bonded to flavin–adenine mononucleotide, and the resulting molecule is added to apoglucose oxidase to reconstitute enzymatic activity. Reconstitution of this activity is inhibited by the presence of free antitheophylline antibody, which is affected by the amount of sample theophylline present. Unlike EMIT methods, which have been adapted to many different clinical analyzers, ARIES technology is currently used only on the Ames Seralyzer.

3.3.3. Fluorescence-Based Immunoassays

As an analytical tool, fluorescence detection is generally superior to absorbance when low levels of fluorescent substances are measured. However, this does not mean that fluorescence–based immunoassays are more sensitive than those based on spectrophotometry because the affinity of the antibody for the antigen is generally the limiting parameter in homogeneous immunoassays. Nevertheless, some fluorescence-based immunoassays have been developed, the most notable of which is the Ames *substrate-labeled fluorescent immunoassay* (SLFIA), automated on the Ames/Gilford Optimate analyzer. This technique employs a fluorogenic enzyme substrate, formed by linking a galactose residue to methylumbelliferone and antigen. β-Galactosidase (E.C. 3.2.1.23) [*9031-11-2*] can hydrolyze the substrate to a fluorescent product, but this reaction cannot occur if the substrate is bound by antibody. Thus, the amount of fluorescence is proportional to the concentration of competing antigen present in the sample. Other techniques based on fluorescence quenching and time-resolved fluorescence have also been reported, but have not gained wide acceptance in the clinical field, partly because they require special instrumentation and complicated manual procedures.

3.3.4. Fluorescence Polarization

When a low molecular mass substance (like phenobarbital) is attached to a fluorophore or its derivative, changes can be measured in the polarization of fluorescence when the conjugate binds to an antibody. This occurs because the rotation of a fluorophore bound to a high molecular mass antibody is considerably slower than that of the free fluorophore and results in a more polarized emission of light. Assays based on this approach rely on competition between the substance to be determined (an analyte) and the analyte-fluorophore conjugate for antibody binding and are generally rapid and reliable. Fluorescence polarization can be used to measure most of the current therapeutic drugs; however,

this method is limited in sensitivity by the fluorescence quantum yield of the selected fluorophore and in scope by the mass of the desired analyte to be measured.

3.3.5. Liposome-Based Immunoassays

Liposomes are small bilayer membrane vesicles that can be prepared by suspending a variety of lipids in water. They can be either multilamellar, containing many internal aqueous compartments, or unilamellar, containing only one compartment per vesicle. When prepared in the presence of soluble marker molecules (e.g., enzymes, ions, fluorophores, enzyme substrates, or inhibitors), the markers are entrapped within the bilayer membranes and subsequently are released during an immunoassay.

The most commonly reported *liposome assay design* uses antigens or small haptens covalently bonded to the lipid membrane [57]. This analysis can be done by linking the antigen to phosphatidylethanolamine (this complex is later used to form liposomes), or by preforming the liposomes and attaching the antigen through a bifunctional cross-linking agent. Specific antibodies bind to antigens on the liposome surface, and the vesicles are then lysed by adding a complement (i.e., a lytic protein typically obtained from guinea pig serum). Antigen from the sample competes with the liposome-attached antigen for antibody binding.

Other immunoassay designs use liposomes that are not coated with antigens, but are instead lysed by hapten–cytolysin conjugates, such as digoxin covalently linked to bee venom melittin [58]. These conjugates retain their lytic activity until they are bound by antibody and can thus be used to monitor competition between the conjugate and analyte in the sample. Heterogeneous assays using antibody-coated tubes together with hapten-coated liposomes have also been developed.

Uses. Liposome-based assays offer potentially a higher signal over other immunoassay techniques because many copies of an enzyme or marker can be loaded within the vesicle. Furthermore, these methods can be used with a variety of detection systems, including spectrophotometers, fluorometers, ion-selective electrodes, and nuclear magnetic resonance spectrometers, depending on the nature of the marker employed. To date, liposomes have been used to measure specific antibodies, IgG, hormones, and therapeutic drugs, such as digoxin and theophylline.

The main concerns regarding these reagents are the preparation of stable liposomes that can be used over several months, the cost of a complement, and the selection of antibodies with the ability both to activate a complement and to bind the antigen with high affinity. As with other homogeneous immunoassays, the ultimate sensitivity for analyte is usually limited by the affinity of the antibodies.

3.4. Reagents for Nonisotopic Heterogeneous Immunoassays

Like most radioimmunoassays, these tests *require preattaching of either an antibody or an antigen to a solid support.* The attachment can be through covalent bonds, simple adsorption, or binding of the immunochemical to a secondary binding agent, such as protein A (from *Staphylococcus aureus*) or another antibody. An important difference between radioimmunoassays and nonisotopic assays is that the latter generally employ enzymes instead of radioisotopes as markers. The enzymes can be covalently linked to either antibodies or antigens by using glutaraldehyde, periodate oxidation, or other bifunctional coupling agents, and the choice of coupling conditions depends on the method being developed. Horseradish peroxidase (E.C. 1.11.1.7) [*9003-99-0*], alkaline phosphatase(E.C. 3.1.3.1) [*9001-78-9*], and β-galactosidase (E.C. 3.2.1.23) [*9031-11-2*] are commonly used because of their high substrate turnovernumbers and because the products of these enzymes can be measured either spectrophotometrically or fluorometrically. Some conjugates have also been prepared by using bacterial luciferase, which can be measured with a bioluminescence detector, as the enzyme.

3.4.1. Enzyme-Linked Immunosorbent Assays

These assays, generally referred to as ELISA, have gained wide acceptance since the mid

1970s for measuring high and low concentrations of high and low molecular mass antigens, primarily infectious disease agents, tumor markers, hormones, and other antibodies [59].

In the most basic case, an antibody is first adsorbed from solution onto a polymeric (often polystyrene) support, which can be premolded into a variety of convenient shapes, such as test tubes, small 6.4 mm beads, and microtiter plates (an array of 96 wells in a plastic tray). After the adsorption step, the support is washed with buffer to remove any unbound antibody, and the sample antigen is added. An enzyme-tagged antigen can be included at the same time to set up a competitive binding reaction; however, this reagent is usually added after a few minutes. The incubation time may vary from minutes to hours before the support is washed again with buffer, depending on the concentration of sample antigen. The second wash step is critical for removing any unbound enzyme-tag and is followed by adding the enzyme substrate. The amount of enzyme-tagged antigen detected on the support is, therefore, inversely related to the concentration of sample antigen.

One frequent *research use* of this technique is to screen monoclonal antibodies for further diagnostic and therapeutic applications. In this case, the antigen is usually adsorbed to the inner surface of microtiter plates, and a second antibody (such as rabbit antimouse IgG) is linked to the enzyme. After the support is washed, samples containing antibodies, which can bind to the adsorbed antigen, are added, and the bound antibodies are detected with the second antibody–enzyme conjugate. This technique gives informa-tion about which cell lines are producing monoclonal antibodies of the IgG class, as well as information about the antibody titer and affinity.

3.4.2. Sandwich Immunoassays

Many antigens, such as HCG, TSH, and ferritin, have multiple epitopic sites and can be measured by using the sandwich immunoassay approach, wherein one antibody is attached to a solid support and another is linked to an enzyme marker. The antigen, therefore, becomes sandwiched between the two antibodies, and the amount of bound antibody–enzyme conjugate is directly proportional to the concentration of antigen in the sample. These assays are more sensitive than competitive type methods because an excess of both antibodies can be used to drive the reaction with antigen (overcoming constraints usually imposed by antibody affinity). The strength of signal generated can also be enhanced simply by incubating the conjugate with the substrate for a longer time.

Tests using this approach can be performed in either of three modes, i.e., forward, simultaneous, or reverse, depending on the order of reagent addition. In the *forward mode,* a sample containing antigen is first added to the solid support to which the antibody is bound. An antibody–enzyme conjugate is then added to react with any captured antigen. After the support is thoroughly washed, an enzyme substrate is added. The wash step in this case is necessary to remove all conjugate that is not bound to the sandwiched antigen. In the *simultaneous mode,* the sample and both antibody reagents are added together. In the *reverse mode,* the sample is first incubated with the antibody–enzyme conjugate. The reverse mode often gives a stronger signal than the forward mode because the antibody–enzyme conjugate can react faster with the antigen when it is in solution than when it is bound to a solid support [60]. Supports used in commercial methods include small polystyrene beads, coated tubes, and magnetic particles.

A *modification* of this approach, which uses two monoclonal antibodies, is described as a Tandem immunoassay [53]. Using monoclonals, Hybritech has commercialized tests for TSH, HCG, ferritin, and carcinoembryonic antigen and has reported others for IgE, luteinizing hormone, and α-fetoprotein.

3.4.3. Affinity Column Mediated Immunoassays

Low molecular mass substances, like small haptens, can be measured accurately by using antibody–enzyme conjugates together with short affinity columns or solid supports containing either the hapten or its analogue [61].

The sample is first incubated with specific antibody–enzyme conjugate, and the resulting solution is then applied to an affinity column. Any specific antibody that does not re-

act with sample hapten binds to the hapten on the solid support; the amount of conjugate passing through the column is directly related to the quantity of sample hapten. Enzyme activity can be measured on either the fraction passing through the column or the bound fraction, depending on the amount of nonspecific background signal that can be tolerated and the purity of the conjugate employed. β-Galactosidase is preferred as a marker when the unbound fraction is measured because this enzyme is not normally present in human serum. Other enzymes, however, can be used when the bound conjugate fraction is determined.

3.4.4. Avidin and Biotin

Avidin, a common protein found in egg white, can bind four molecules of biotin [*58-85-5*] (hexahydro-2-oxo-1*H*-thieno[3,4-*d*]imidazole-4-pentanoic acid), with an affinity constant of 10^{15} L/mol. This tight interaction is essentially nonreversible and has been used in the design of heterogeneous immunoassays in which one immunochemical must be linked either to a solid support or to an enzyme. Many investigators report a higher signal when conjugates of avidin and enzyme are employed with biotinylated antibodies [62]; however, theoretically, no enhancement should be more than threefold above the value found with the use of direct antibody–enzyme conjugates. Biotinylated antibodies have also been linked to supports containing bound avidin, which in some circumstances can lead to slightly increased sensitivity. The incremental costs of using avidin and biotin, therefore, become important when a decision is made whether an assay with an enhanced signal or sensitivity should be developed. Bacterial streptavidin is also available and can be substituted for avidin.

4. Microbiologic Reagents

Most microorganisms, as members of the kingdom Protista [63], are single-cell entities that can only be seen individually with the aid of a light or electron microscope. Viruses are obligate intracellular parasites that do not have cell structure; they possess either DNA or RNA, usually surrounded by a protein-containing coat or capsid and can only be visualized directly by electron microscopy.

This chapter reviews selected diagnostic assays as well as reagents used to determine the quantity and identity of bacteria, fungi, protozoa, viruses, chlamydia, and rickettsia. All microorganisms cited have proven roles in human disease and infectious processes. A short section is included describing products that facilitate collection, transport, and detection of microorganisms from specimens, such as human body fluids. Reagents used in standard or typical methods for microbial detection or identification are also described, and in many cases references and commercial sources of reagents are provided.

Most standard practices in diagnostic microbiology require a culture of a suspected pathogen. The isolated pathogen is typically grown in vitro to increase the population to ca. $10^8 - 10^9$ organisms. This increase in number of organisms permits subsequent diagnostic testing. However, technological advances, such as nonradioisotopic methods of labeling antibodies and nucleic acids, have stimulated development of methods for direct microbial identification. In many instances, these newer direct detection procedures do not require in vitro cultivation of the pathogen. Therefore, diagnostic reagents for detection of microorganisms can be divided operationally into (1) growth-dependent and (2) growth-independent methods. *Growth-dependent techniques* include all methods and reagents that enhance or support growth of target organisms. The organisms are detected by macroscopic evaluation, e.g., of turbidity in broth or of colonies on agar-based growth media. *Growth-independent methods* of detection include microscopy, serology, and newer techniques, such as specific antigen and nucleic acid detection.

Genus- or species-specific antigen detection is more feasible today because of the availability of monoclonal antibodies that are capable of recognizing a specific antigenic determinant (i.e., an epitope) of the target organism. Nucleic acid probes are reagents, like monoclonal antibodies, that can be simultaneously used to detect and identify an organism. The degree of specificity achieved with monoclonal antibodies or

nucleic acid probes is a function of the selection process used to identify and purify the reagent. The major reason that direct specimen detection and identification methods and reagents are gaining popularity is that they save money and reduce time from sample collection to identification compared to traditional approaches. Consequently, the economic impact is the driving force to develop new direct diagnostic methods and reagents.

4.1. Collection, Transport, Processing Devices, and Reagents

Diagnostic tests and associated reagents for use in the clinical laboratory to detect and identify bacteria predominantly depend on bacterial growth. Special precautions, such as antisepsis of the sample site, use of selective media, and microscopy to count the specimen, need to be taken to avoid contaminating patient samples taken from normally sterile body sites. In addition, a properly collected specimen is worthless to the clinical laboratory if it arrives in a condition that reduces or eliminates the identifiability of the disease-producing agent. A comprehensive review of specimen collection and transport practices and procedures used at the Mayo Clinic, Rochester, Minnesota, United States, can be found in *Laboratory Procedures in Clinical Microbiology* [64].

Increasing emphasis has been placed on methods of collecting and transporting clinical samples to the testing laboratory. Many new commercial devices that enhance recovery efficiency, stabilize the viable microbial concentration, or process the specimen to enhance recovery have been introduced. For example, septicemia requires a diagnostic test for detection of bacteria or fungi in blood. Currently, this can be accomplished only by isolating or growing the organism from the sample. Typically, blood is aseptically collected in devices containing a suitable anticoagulant, e.g., sodium polyanetholesulfonate (SPS), for transport to the clinical microbiology laboratory.

Blood Collection Devices. The simplest device is a sterile, rubber-stoppered, SPS containing Vacutainer, manufactured by Becton Dickenson. Another blood collection, transport, and processing device is the Du Pont Isolator microbial tube. In addition to containing SPS, the Isolator allows for concentration of microorganisms by centrifugation. The concentrated bacteria or fungi are plated directly onto various agar-based growth media in petri plates, which results in the growth of isolated colonies 18 – 24 h sooner than broth culture methods [65]. Other blood transport devices are blood culture bottles and tubes containing SPS and a liquid growth medium or broth, including soybean-casein digest, e.g., Trypic or Trypticase soy broth (TSB), brain heart infusion (BHI), Brucella, Thiol, thioglycolate, Columbia, and supplemented peptone broth (SPB). United States manufacturers of blood culture media include Difco, BBL, Scott, and Roche.

A new device made by Roche called Septi-Chek has a built-in slide subculturing chamber, which has chocolate, MacConkey, and malt agar surfaces. Microorganisms growing in the broth phase are simultaneously subcultured onto the agar surfaces and form colonies. The antimicrobial removal device (ARD) by Marion Laboratories is a sterile bottle containing ion-exchange resins that remove antibiotics from the patient's blood. Following antibiotic removal, the sample is transferred to a culture system. The ARD is intended for use in those cases in which patients are known to be receiving antimicrobial therapy and septicemia is suspected.

Urine collected for microbiologic analyses, e.g., bacteriuria, requires either prompt analysis (ca. 2 h) or refrigeration because urine is an excellent growth medium, and the concentration of bacteria or fungi in urine is indicative of urinary tract infection. The BD Urine Culture Kit is a new device and approach to urine transport. Patient urine is added to a sterile tube containing glyceroboric acid in sodium formate. These reagents stabilize and maintain the microbial count from the time the sample is collected until it is analyzed.

Although not a collection or transport device, the new Chemstrip LN (Boehringer Mannheim) is a *rapid urine screen for bacteriuria* and is claimed to have a negative predictive value of 97.2 %. The Chemstrip LN is a 2-min dip-strip test specifically for leukocyte esterase and nitrate reductase. Urine culture is indicated only if the urine sample shows a positive Chemstrip LN

test. Other microbial devices and reagents include Meridian ParaPak for recovery or straining of protozoan parasites from feces, Culturette collection systems (Marion Scientific) for swab samples, and Anaerobic Specimen Collector (Becton Dickenson), which is an anaerobic swab collection system for wounds and abscesses.

4.2. Growth-Dependent Microbial Reagents

Growth-dependent methods of microbial diagnosis have, in many cases, been selected from among processes occurring for over 100 years.

The success of culture-based methods for detecting and identifying bacteria, fungi, protozoa, viruses, chlamydia, and rickettsia depends on establishing *optimal environmental and chemical growth conditions.* The growth medium or reagent must possess all necessary nutritional requirements to support microbial multiplication. Bacteria, fungi, and protozoan parasites can be cultivated in vitro by using chemically defined media. Growth of protozoan parasites often seems to be more art than science [66], because complex biologic fluids, e.g., egg yolk, gastric mucin from hog stomach, and serum, must be included to achieve optimal results. Occasionally, *fastidious* bacteria and *mycoplasma* species require additional factors, such as blood, hemoglobin, serum, coenzyme(s), or hematin.

Recommendations and formulations for cultivation of the various categories of microorganisms are found in the following publications: bacteria and fungi [64], [67–72], protozoa [66], [67], viruses [64], [67–70], and chlamydiae/rickettsiae [64], [67], [68], [70]. Manufacturers of prepared media include BBL Microbiology Systems, Difco Laboratories, Fisher Scientific Co., GIBCO, M. A. Bioproducts, Flow Laboratories, K. C. Biologicals, and Scott Laboratories. Linscott's *Directory of Immunochemical and Biological Reagents* (1984 – 1985) lists worldwide suppliers of media and reagents. All manufacturers provide the chemical formulations for commercial media, which also serve as references for media composition.

Viruses and chlamydiae are all obligate, intracellular parasites. These organisms require *host cells* for in vitro cultivation. Host cells are typically primary, secondary, or continuous [63], [69] as well as either fibroblastic or epithelial in origin. In vitro isolation and detection of viruses and chlamydiae require permissive or susceptible host cells. The pathogen must be capable of entering the cells and multiplying. In most cases, clear characteristics of infection are obvious by low-power (40 – 100 ×) phase-contrast microscopy after 1 – 2 d for Herpes simplex or polio virus, or weeks for cytomegalovirus [63], [69].

Cell cultures can be purchased or obtained as gifts from *commercial sources.* In the United States, Flow Laboratories, M. A. Bioproducts, and BBL Microbiology Systems are major suppliers of cell cultures. The American Tissue Culture Collection (ATCC) is another source for cell cultures. Primary cell cultures are routinely prepared in large virology laboratories from rabbits (kidney), guinea pigs (whole embryo), or chickens (whole embryo) [69] or can be purchased as ready-to-use monolayer cultures from commercial sources.

Growth-dependent methods for recovery of viruses and chlamydiae are costly and time consuming. However, in many cases, culture remains the method of choice based on consideration of clinical sensitivity and specificity.

4.3. Immunodiagnostic Reagents in Clinical Microbiology

The *serological determination* of bacterial infection is performed by measuring the quantity of antibody present in patient serum that specifically reacts with a selected antigen preparation. Antigen selection and preparation for use in a diagnostic assay are the most important reagent-related variables. Table 2 lists clinically significant bacteria, fungi, and protozoan parasites whose detection or identification often is performed by using serological methods. Similarly, Table 3 lists serological assays, tests, and reagents used to diagnose viral, rickettsial, and chlamydia infections. The antigenic preparation used for each organism is also listed along with literature references and, in many cases, commercial sources for the specific organism or antigen preparation. Standard serological tests used to identify and type bacteria include precipitation reactions (e.g., ring test,

quantitative precipitation, capsular swelling, gel diffusion methods, and immunoelectrophoretic methods), complement fixation, neutralization tests, cytolytic methods, agglutination assays, flow cytometric analysis, and immune electron microscopy [72], [77].

Table 2. Antigen reagents used in serological assays for selected bacteria, fungi, and protozoan parasites

Microorganism	Antigen reagent and commercial source
Group A β-hemolytic streptococci	streptolysin O; deoxyribonuclease β; hyaluronidase, Difco, GIBCO, Fisher [68], [73]
Streptococcus pneumoniae	capsule polysaccharides; Difco [73]
Neisseria meningitides	capsule polysaccharides; outer membrane proteins; lipopolysaccharides; pili [67], [70], [73]
Neisseria gonorrhoeae	phenol extracts, lipopolysaccharide; pooled cells; soluble fraction from physically disrupted cells; pili [73]
Salmonella typhi	somatic (0 antigens) flagellar antigens, H antigens; Difco
Vibrio cholerae	somatic (0 antigens) heat labile enterotoxin; Y cholerae serotype Ogarva (strain VC12) and serotype Inaba (strain VC13)
Yersinia enterocolitica	0 antigen (lipopolysaccharide)
Yersinia pseudotuberculosis	0 antigen (lipopolysaccharide)
Pasteurella multocids	soluble extract from heated cells [73]
Francisella tularensis	formalin-treated whole cells or commercial preparations; Lederle, Difco, BBL [73]
Haemophilus influenzae Type B	capsule polysaccharide, polyribose phosphate (PRP) [68], [70]
Brucella abortus	surface antigens A and M
Brucella suis	surface antigens A and M
Brucella melitensis	B. aborus strain 1119 used as antigen standard
Bordetella pertussis	whole cell preparation B., pertussis strain 460 [a]
Legionella pneumophila	*L. pneumophila*
Microorganism	Antigen reagent and commercial source
	A antigens serotypes 1 – 4 heat-killed organisms [b]
Pseudomonas pseudomallei	*P. pseudomallei* strains 8202, China 3 and 292 heat-killed, cell-free supernatant [73]
Listeria monocytogenes	*L. monocytogenes* serotype 1 a, NCTC 7973, H.P.R. Seeliger, ATCC 19 111 and serotype 4 b, F 4, J. Donker-Voef [b]
Corynebacterium diphtheriae	*diphtheriae toxoid;* "Schick test" [73]
Cloareisium tetani	tetanus toxoid "Schick test" [73]
Mycobacterium tuberculosis	*M. tuberculosis* strain DT-612 [c]
Treponema pallidum	*T. pallidum* antigens available from a variety of manufacturers [68], [70], [73], [74]
Mycoplasma pneumoniae	*M. pneumoniae*
Mycoplasma lomines	fraction from
Ureaplasma urealyticium	chloroform – methanol extraction *M. homines* and *U. urealyticum* nonlipid fractions
Aspergillus fumigatus	culture filtrate,
Aspergillus flavus	acetone precipitated;
Aspergillus niger	Meridian M.A. Bioproducts
Micropolyspora faeni	Greer Labs
Thermoactinomyces candidus	
Thermoactinomyces virlgaris (actinomycotic hyper-sensitivity pneumonitis)	
Paracoccidioides brasiliensis (paracoccidioidomycosis)	concentrated yeast filtrate antigens; [12]; CDC stock P. brasiliensis B339, B341 & B1183
Skporothrix schenckii	*S. schenckii* reagents not commercially available [73]
Giardia lamblia	[66], [67], [73]
Entamoeba listolytica (amebiasis)	[66], [67], [73]
Trichomonas vaginalis	[66], [67], [73], [75]
Trypanosoma cruzi	[67], [73], [75]
Trypanosoma gambiense (African trypanosomiasis 0)	
Leishmania donovani	
L. tropica	[73]
L. braziliensis (leishmaniasis)	Burroughs Wellcome
Plasmodium falciparum	[73]

Table 2. (Continued)

Microorganism	Antigen reagent and commercial source
P. vivax (malaria)	
Pneumocystic carinii (pneumocystosis)	[73]
Toxoplasma gondii (toxoplasmosis)	[73]
Other miscellaneous parasitic diseases: echinococcosis, schistosomiasis, fascioliasis, trichinosis, filariasis, ascariasis, Chagas' disease	[73]

[a] Bacterial Immunology Branch, Bacteriology Division, Center for Disease Control, Atlanta, Ga. 30333, United States
[b] Biological Products Division, Center for Disease Control, Atlanta, Ga. 30333, United States
[c] Mycobacteriology Branch, Center for Disease Control, Atlanta, Ga. 30333, United States

An example of one of the most commonly employed immunoassay test configurations for serodiagnosis uses antigen immobilized on a solid support, e.g., on the walls of a polymer tube, on the surface of polymer spheres, or in the wells of a plastic microtiter plate. Serum samples are added to the solid supports and allowed to react. Unreacted material is washed away and a labeled second antibody is added to all test chambers. As in many immunoassays, a blocking substance, e.g., 1 % bovine serum albumin or ovalbumin, is added to minimize nonspecific binding reactions. Unreacted material is washed away and the amount of specific human antibody bound to immobilized antigen is measured.

This type of immunoassay is designated an *indirect, heterogeneous immunoassay* because two antibody reagents are employed and reagents must be added sequentially and separated by washes to remove unreacted reagents [73], [74], [76]. Direct antigen and nucleic acid detection are discussed in Section 4.4.

4.4. Monoclonal Antibody and Nucleic Acid Assays

Monoclonal antibodies and nucleic acids (DNA probes) are highly sensitive and specific diagnostic reagents. A **monoclonal antibody reagent** consists of one antibody type that recognizes a specific antigen or epitope. Monoclonal antibodies are produced by specialized hybrid cells called *hybridomas,* which are made by fusing antibody–secreting cells with cancer cells. Hybridomas can be frozen for long-term storage, ensuring a uniform source of antibodies.

Monoclonal antibody reagents can be produced continuously in a reproducible fashion and can be highly purified and characterized. However, in spite of the obvious advantages, the ultimate role of monoclonal antibodies in microbial diagnostics remains uncertain. The reason for this uncertainty is that microorganisms are capable of altering their antigenic appearance as a result of genotypic (mutagenic) or phenotypic (environmental) changes. Such changes can prevent the highly specific monoclonal antibody from recognizing a desired target (epitope) and, thus, render the assay useless. One method of circumventing this problem is to choose monoclonal antibody reagents that react with highly conserved microbial antigens.

Regardless of strategy employed to produce a monoclonal antibody, it is highly desirable and advantageous if the selected antibody has both suitable affinity and avidity [72].

This distinction between diagnostic targets subject to genetic or environmental changes (e.g., microorganisms) and those targets that are constant (e.g., drugs and hormones), more than any other differences, accounts for the broader use of monoclonal antibodies in clinical chemistry than in diagnostic microbiology. Continued research on the application of monoclonal antibodies for microbial diagnostics will undoubtedly minimize the importance of these distinctions.

DNA probes are pieces of nucleic acid chosen because they specifically hybridize to a target nucleic acid located within an organism. The DNA probe is generally a *single-stranded nucleic acid* consisting of > 20 bases and is labeled in some way to permit detection. Common *labels* include radioisotopes, enzymes, or chemiluminescent or fluorescent molecules.

Labeled DNA probes can be reacted (i.e., hybridized) directly with a target nucleic acid (complementary and single-stranded) bound to a solid support. An example is filter hybridization in which single-stranded target nucleic acid is bound to a nylon or nitrocellulose membrane

Table 3. Serological assays [a], tests, and reagents used to diagnose viral, rickettsial, and chlamydial infections

Organisms	Tests [b]	Reagent sources and references
Herpes simplex	CF, IFA, NI, RIA, EIA	ATCC, Ortho, Boots-Celltech, Syva, Biotech Research Labs, Wellcome Diagnostics, Gull Labs, Litton Electro-Nucleonics, Dako [70], [72], [76]
Cytomegalovirus (CMV)	latex agglutination, EIA	Hynson, Westcott & Dugging, M. A. Bioproducts, Litton Bionetics, Biotech Research Labs, Wellcome Diagnostics [73], [76]
Varicella-Zoster virus (VZV)	CF, FA against membrane antigen (FAMA)	Flow, M. A. Bioproducts [70], [73]
Epstein-Barr virus (EBV)	elevated Ab levels (IFA) to EBV viral capsid (VCA), EBV membrane antigen, EBV nuclear antigen (EBNA), and EBV early antigens or heterophile antibody test	[68], [69], [73], [74]
Influenza virus Type A	HI, HAD	GIBCO, Flow, M. A. Bioproducts [69], [73], [74]
Influenza virus Type B		
Respiratory syncyteal virus (RSV)	CF, HI, NI, EIA	M. A. Bioproducts, Flow, BBL [63], [69], [70], [73], [74], [76]
Parainfluenza virus (Types 1 – 4)	CF, HI, NI, EIA	M. A. Bioproducts, Flow, BBL [63], [69], [70], [73], [74], [76]
Measles	HI, CF	Flow, M. A. Bioproducts [68], [73], [74]
Mumps	HI	Flow, M. A. Bioproducts [68], [73], [74]
Adenovirus	CF, HI, NI	M. A. Bioproducts, Flow [63], [69], [73], [74]
Rubella	HI, NI, CF, EIA latex agglutination, passive HA	Abbott Labs, Ortho Diagnostics, Flow, M. A. Bioproducts [70], [73]
Togaviridae (Eastern equine) encephalomyelitis, Western equine encephalomyelitis, yellow fever, tick-borne	HI, CF, NI; also EIA, FA, RIA	[69], [73], [74]
Bunyaviridae (Rift Valley fever, Phlebotomus fever, and LaCrosse)	Hi, CF, NI; also EIA, FA, RIA	[69], [73], [74]
Arinaviridae (lymphocytic choriomeningitis, Lassa virus, and Tacaribe, Pichinde, and Flexal viruses)	CF, IFA, FA, NI	[63], [69], [70], [73], [74]
Entero viruses (polio virus and echovirus)	isolation (culture) usually preferred to antigenic variance NI, CF (polio) HI, FA	M. A. Bioproducts [73]
Reovirus (Types 1, 2, & 3)	HI, NI, CF	M. A. Bioproducts [73]
Hepatitis viruses Hepatitis A	detection of Hepatitis A antigen in stool –	[73], [74], [76]

[72], [77], [78]. Labeled, single-stranded nucleic acid is then hybridized to complementary sequences bound to the filter, and unhybridized DNA probe is washed away. The labeled, hybridized DNA is detected and the amount that is bound is compared to positive and negative controls. DNA probe reagents can be *amplified* either biologically through a recombinant vector or synthetically by oligonucleotide syntheses. If desired, nucleic acid probes can be *synthesized* on the basis of known amino acid sequence data. Finally and most importantly, the *chemical basis* of all nucleic acid probe–nucleic acid target hybridization reactions (which defines clinical assay specificity) is complementary base-pairing (hydrogen bond formation) and stabilization of the resulting duplex via hydrophobic interactions between stacked bases.

Table 3. (Continued)

Organisms	Tests [b]	Reagent sources and references
	RIA, EIA, IEM	
Hepatitis B	HBsAg; anti-HBs detection; anti-HBc detection – RIA, EIA, latex agglutination	[73]
Hepatitis non-A, non-B	essentially none available; negative test results for Hep A and Hep B indicate presumptive positive for Hep non-A, non-B	none available
Rotavirus (infectious gastro-enteritis in children)	demonstration of rotavirus in stool using EIA, RIA, IEM or rotavirus-specific antibody in serum – EIA, RIA, CF	Abbott Labs [73]
Chlamydia trachomatis (Trachoma and venereal disease)	CF, FA, IFA	[73], [74], [76]
Rickettsiae (Rocky Mountain spotted fever, Q fever, etc.)	IFA, CF, Weil-Felix	[73]
Human tumor leukemia HTLV-III (cause of human-acquired immunodeficiency syndrome, AIDS)	IA, RIA, (HTLV-III antigens coated onto solid phase)	Abbott Labs, Litton-Bionetics, Electronucleonics, Du Pont

[a] A variety of the serological assays, e.g., neutralization test, hemadsorption, hemadsorption inhibition, and fluorescent antibody (direct and indirect) often require virus-infected cell cultures; such assays are growth-dependent methods; however, newer test methods that employ an antigen-coated solid phase, e.g., latex agglutination, EIA, RIA, do not require cell cultures and are appropriately listed as growth-independent methods and reagents.
[b] Tests: CF = complement fixation; EIA = enzyme immunoassay; HAD = hemadsorption; HI = hemagglutination–inhibition test; IAH = immuneadherence hemagglutination; IEM = immunoelectron microscopy; IFA = indirect immunofluorescent staining; FA = immunofluorescent staining; NI = neutralization test ; RIA = radioimmunoassay.
[c] Biological Reagents Branch, Scientific Resources Division, Bureau of Laboratories, Center for Disease Control, Atlanta, Ga., United States.

The pressure to *reduce health care costs* may influence the choice of technology used for microbial detection and identification. Methods that detect and identify pathogens directly from the patient specimens produce faster results than culture-based methods. Improvement in test turnaround time, i.e., time from sample collection to result, should produce improved administration of chemotherapy. Use of rapid diagnostic methods could also reduce the time of hospitalization for some patients. Consequently, in many cases, traditional culture-based, growth-dependent microbial diagnostic methods will be displaced by nonculture, growth-independent methods. The decision, however, to adopt new direct detection methods must be based on clinical utility and cost effectiveness.

5. Coagulation Diagnostic Reagents

5.1. Introduction to Hemostasis

Hemostasis is the complex mechanism by which blood is maintained in a fluid state within the circulatory system (→ Blood), [79–87]. It involves the intimate interaction of blood vessels, platelets, and soluble plasma proteins. Disorders of any of these components can lead to serious bleeding or clotting problems. Because of the complex nature of the system and the diversity of pathological consequences, many varied tests are required to evaluate the integrity of the hemostatic system.

Blood Vessels. Although some new tests are emerging, diagnosis of disorders of the blood vessels that affect hemostasis is difficult. A physical examination may be necessary for recognition of overt vascular defects. Abnormal bleed-

ing times or tourniquet tests are two nonspecific indicators of vascular disorders. As the molecular mechanism of the interaction of platelets and soluble plasma proteins with the vascular endothelia becomes better understood, more direct clinically significant tests may emerge.

Platelets. Platelets are circulating cells containing proteins and other factors that interact with the vessel wall and with soluble plasma proteins. Pathological disorders of platelets can cause bleeding disorders and may also contribute to problems such as atherosclerosis. Measurement of platelet number and function (bleeding time) are routinely performed and other specialized tests are becoming more widely used.

Plasma Factors: Screening and Specific Tests. The common screening tests for hemostasis *evaluate the overall integrity of the blood coagulation cascade* (→ Blood, Chap. 2.1.). The coagulation cascade is extremely complex, and generalized screening tests are useful to confirm normal function. Certain disorders require more specific diagnostic tests for individual factors or for inhibitors or activators that modulate the factors. The proper dissolution of blood clots through the fibrinolytic pathway is also required for normal hemostasis and several diagnostic tests are available and useful in detecting fibrinolytic disorders.

Anticoagulants. Cardiovascular disorder, often caused or potentiated by the improper regulation of the hemostatic system, is a leading cause of death. Anticoagulants are widely prescribed for *thrombotic episodes.* Proper therapeutic maintenance requires specialized tests that monitor the function of these drugs.

5.2. Screening Tests

Screening tests are widely used to evaluate the overall integrity of the coagulation pathways. In general, they are used for presurgical screening to avoid bleeding complications, for monitoring anticoagulant therapy, and for evaluating bleeding disorders. Screening tests are of limited value in detecting thrombotic disorders.

Prothrombin Time. The prothrombin time (PT) test was developed in 1935 and is *used to determine the status of the extrinsic system* (factors I, II, V, VII, and X). Patient's plasma (0.1 mL) is mixed with 0.1 mL of tissue thromboplastin and 0.1 mL of a calcium chloride solution at 37 °C, and the time to clot formation is measured. Various manufacturers produce the thromboplastin reagent, which contains extracts from animal or human brain or lung tissue (see Table 4). The PT results can be reported directly as the clotting time (along with the clotting time for a control plasma), a ratio (test : control), or a percentage (based on dilution of normal plasma with saline). Control plasma materials are commercially available and can be prepared from a normal plasma pool.

The test procedure is fairly well standardized, and results are generally reproducible within a laboratory. Unfortunately, test results vary between laboratories because each thromboplastin reagent has a different pattern of sensitivity to the various extrinsic coagulation factors. Besides its use for general screening, the PT test is used to monitor oral anticoagulant therapy.

Activated Partial Thromboplastin Time. The activated partial thromboplastin time (APTT) test *measures the integrity of the intrinsic system of coagulation.* It is sensitive to deficiencies in factors I, II, V, VIII, IX, X, XI, and XII. A mixture of 0.1 mL of patient's plasma and 0.1 mL of a phospholipid–surface activator mixture is incubated. Calcium chloride solution (0.1 mL) is added to initiate clotting, and the time required to form the initial fibrin clot is measured. The phospholipid–surface activator mixture is critical to the performance of the APTT test. Different sources of phospholipid (e.g., rabbit or bovine lung tissue or soybean extracts) and different surface activators (e.g., kaolin, silica, or ellagic acid) are used in various combinations for the APTT test. Table 4 shows the major manufacturers and their APTT reagent preparations. The APTT test results are reported as direct time in seconds. Control plasma materials are also available, or the laboratory may use an in-house normal plasma pool. The APTT test is not as well standardized as the PT test. Different reagents can yield different results (anywhere from 25 s to 45 s); sensitivity also can vary. Besides measuring the integrity of the intrinsic system, the APTT test is most commonly used to monitor heparin therapy.

Table 4. PT/APTT Reagent manufacturers

Reagent name or trade name		
PT	APTT	Manufacturer
Thromboplastin C	Actin, Actin-FS	Dade
Brain thromboplastin	Thrombofax	Ortho
Simplastin	Platelin Plus	General Diagnostics
Liquid thromboplastin	APTT Reagent	Hyland
Thromborel	Pathrombin	Behringwerke
Thromboplastin	Cephalin	Helena
Neoplastine	CK-Prest	Stago
Thrombotest	Cephotest	Nyegaard
Thromboplastin with calcium	Activated Cephaloplastin	Merieux
Lyoplastin		Mochida
Calcium thromboplastin	PTT-Reagent	Boehringer

Other Screening Tests. *Thrombin time* is determined by adding thrombin [*9002-04-4*] to the patient's plasma. Times are abnormal if fibrinogen is abnormal or deficient or if thrombin inhibitors, such as heparin or fibrin degradation products (FDP), are present.

The *thromboelastogram* (TEG) is a means to measure the dynamics of the blood clotting process. The instrument cuvette is filled with venous blood, the blood is coated with paraffin oil, and a steel cylinder is inserted. The cuvette is placed into the thromboelastograph and put into torsional motion. As the blood begins to clot, fibrin threads, which cause the steel cylinder to move, are formed. This movement is detected via an optical system and recorded. Disorders in the coagulation system cause characteristic deviations from the normal thromboelastogram curve. This TEG test provides significantly more information than the other screening tests, but is not widely used because of the need for specialized instrumentation and the complexity of interpreting the results.

Euglobulin lysis time measures the status of the fibrinolytic system. Blood is allowed to coagulate in a tube. Normally the clot does not dissolve within 24 h; however, if it does dissolve within this period, the fibrinolytic system is assumed to be activated. The increase is clinically significant only if the clot dissolves within 1 h.

5.3. Specific Analytes

Assays for specific components of the coagulation–fibrinolytic system are made to diagnose the cause of abnormalities detected in preliminary screening tests. These assays may also be done to *monitor* patient response to therapy or to *counsel* a patient with a familiar genetic hemostasis disorder.

Essentially, all disorders of coagulation components are caused by *decreased biologic activity of specific proteins.* These disorders may be a result of defective protein synthesis, altered synthesis or rates of clearance from the circulation, or presence of pathologic inhibitors. Although immunologic testing may indicate normal analyte levels, biologic activity may still be decreased. Specific assays for biologic activity are not available for all components.

5.3.1. Coagulation Factors

Qualitative identification of an abnormal coagulation factor is needed prior to determining the actual factor activity and to differentiating between whether the factor is deficient or whether a circulating acquired inhibitor is present. If the screening test (PT or APTT) result is abnormal, the test is repeated with a 1 : 1 mix of test plasma and normal human plasma. If the test result is corrected by adding normal plasma, a factor deficiency is indicated; uncorrected test results suggest a circulating inhibitor. Further retesting with 1 : 1 mixtures of test plasma and aluminum hydroxide adsorbed plasma (or aged serum) can help pinpoint the deficient plasma factor.

Fibrinogen [*9001-32-5*] (Factor I). Fibrinogen is most commonly measured by the *Clauss method.* Test plasma is diluted 1 : 10 with veronal buffer, mixed with excess thrombin, and

then the clotting time is measured. Heparin and FDPs interfere with the test. Antisera and radioimmunodiffusion (RID) plates are also available for immunologic testing of fibrinogen.

Factors II – XII. These factors are measured by using a *modified PT* (factors II, V, VII, and X) and *APTT* (factors VIII, IX, XI, and XII) tests. Test plasma is diluted 1 : 10 with barbital buffer and mixed 1 : 1 with an appropriate factor-deficient plasma, and the clotting time is measured. Because these assays are based on the PT and APTT tests, complications, such as heparin interference or clot time shortening by elevated coagulation factors, may lead to erroneous results.

Alternatively, *synthetic substrate methods* are available for analyzing all factors. Several of these assays (factors II, VIII, and X) can be readily performed with commercial kits, whereas the rest require special reagents. Synthetic substrate assays afford improved sensitivity, specificity, and reproducibility over traditional clotting assays, but generally have not been fully evaluated in many pathological states, and some discrepancies versus clotting assays have been reported.

Factor XIII. Deficiencies in factor XIII are not detected by routine screening tests and must be recognized by *clinical symptoms.* Suspected deficiencies can be screened by testing *clot solubility* in 5 M urea; normal samples are stable whereas abnormal ones form clots that dissolve within 1 h. Heparin can interfere with this test. An alternative method uses 1 % chloroacetic acid, which speeds abnormal clot dissolution but is more susceptible to interference. Antisera are available for immunologic testing.

Prekallikrein [*9055-02-1*] (Fletcher Factor). Suspected deficiencies can be screened by running the *APTT test,* using kaolin or celite as activator and allowing a 10 min incubation. If this corrects the APTT time, then prekallikrein deficiency is probable. APTT activators that contain ellagic acid are usually insensitive to this deficiency. Quantification can be done by using the clotting assay described for factor XIII with prekallikrein deficient plasma or by using a commercial synthetic substrate kit. Antisera are available for immunologic testing.

High Molecular Mass Kininogen (Fitzgerald Factor). Suspected deficiencies show prolonged APTT times that are not corrected by increased incubation time; usually the patient does not have a history of overt bleeding. Quantification is done with a *clotting assay* (as described) and Fitzgerald factor deficient plasma.

Passovoy Factor. Suspected deficiencies show prolonged APTT times that are not corrected by increased incubation time; the patient has a history of bleeding problems and normal levels of other factors. Quantification is done with a *clotting assay* (as described) and passovoy factor deficient plasma.

5.3.2. Fibrinolytic Pathway

Plasminogen [*9001-91-6*] . The most common plasminogen assay is a *synthetic substrate method.* Plasminogen, activated by addition of either streptokinase or urokinase, hydrolyzes the substrate. The results agree well with the older, more laborious caseinolytic assay. Both RID plates and antisera are available for use in immunologic testing.

FDP. Fibrin and fibrinogen degradation products are readily measured with a *latex agglutination slide assay.* Latex particles coated with antifibrinogen antibodies form visible aggregates when reacted with serum samples containing FDP.

Special sample collection tubes containing thrombin and soybean trypsin inhibitor are used to remove cross-reactive fibrinogen from the sample and prevent in vitro formation of FDP by endogenous proteases after sample collection. Heparin can cause a positive bias because it prevents complete removal of fibrinogen. Batroxobin, a thrombin-like enzyme that is insensitive to heparin, may be substituted for thrombin with samples containing heparin.

A more sensitive, but technically demanding, procedure is the *tanned red cell hemagglutination inhibition immunoassay* (TRCHII), which uses human erythrocytes coated with fibrinogen. Serum (collected as described for the latex assay) is incubated with antifibrinogen antibody and then mixed with the coated erythrocytes. Presence of FDP in the test serum de-

creases visible agglutination. Both the latex particle and TRCHII assays are semiquantitative, and results are reported in titers (equal to the mass of the substance equivalent to a liter of standard solution). Different antibody preparations have different sensitivities to individual degradation products and, therefore, can cause test performance to vary.

5.3.3. Activators

Thromboplastin. A standard assay for thromboplastin activity is not available. The PT test can be modified by using excess normal human plasma and by making thromboplastin the limiting reagent to give an assay for thromboplastin.

Urokinase [*9039-53-6*] . *Synthetic substrates* are available that react directly with urokinase.

Streptokinase [*9002-01-1*] . The *synthetic substrate assay for plasminogen can be modified* to give an assay for streptokinase. Excess purified plasminogen is added to the reaction mixture; the sample that is used contains streptokinase, which is the limiting reagent.

5.3.4. Inhibitors

Two types of inhibitors are present in the human blood system: (1) *naturally occurring plasma proteins,* which are integral to normal hemostasis, and (2) circulating *acquired inhibitors,* which react against specific coagulation factors arising from a pathological cause. Acquired inhibitors can lead to bleeding problems and are differentiated from factor deficiencies by a mixing test (see Section 5.3.1). Typically these inhibitors react with factor VIII or are of the lupus anticoagulant type. Serial testing that uses the APTT assay or a 1 : 1 mix of test plasma with normal human plasma can differentiate the type of acquired inhibitor. The APTT is longer if an inhibitor to factor VIII is present, but is not changed with lupus anticoagulant-type inhibitors.

Factor VIII Inhibitor. Test plasma containing this inhibitor is mixed with normal human plasma and incubated; residual factor VIII activity is measured. The result is compared with a control plasma incubated for the same period of time; the control is needed to correct for the natural lability of factor VIII. The corrected residual factor activity is reported in Bethesda units.

Lupus Anticoagulant. Lupus anticoagulant-type inhibitors act on thromboplastin and can be measured with diluted thromboplastin, using a *modified PT test.*

Antithrombin III (AT III). The biologic activity of AT III is commonly measured by its *heparin-catalyzed inhibition of thrombin;* a synthetic substrate method is used. Bovine rather than human thrombin is recommended to prevent interference from heparin cofactor II. A variation of this method is to substitute factor Xa for thrombin, but limited clinical data are published for this variation. Both RID plates and antisera are available for immunologic testing.

α_1-Antitrypsin [*9041-92-3*] . Biologic activity can be measured by the *inhibitory effect of α_1-antitrypsin on trypsin or elastase;* a synthetic substrate method is used. This inhibitor is most commonly tested with RID plates or other immunologic methods.

β_2-Macroglobulin. Biologic activity can be measured indirectly through a synthetic substrate method that uses *trypsin.* β-Macroglobulin is most commonly tested with RID plates or other immunologic methods.

α_2-Antiplasmin. This compound is most commonly measured by its *inhibition of rapid plasmin;* a synthetic substrate method is used. Both RID plates and antisera are available for immunologic testing.

Protein C. Preliminary assays for biologic activity, which use synthetic substrate methods, have been developed. These analyses require specialized biochemical reagents and are only used in research laboratories. Most Protein C testing is done by ELISA or other immunologic assays.

5.4. Platelets

Normal platelet number and function are required for proper hemostasis. *Platelet number* is measured either manually with a microscope or automatically with an electronic or optical counter. The normal platelet count is 250 000 – 450 000/mm ^{3}.

Platelet function can be tested with the *bleeding time test.* A small incision is made on the patient's arm, usually using a template to improve test standardization, and increased venous pressure (5.3 kPa) is maintained with a blood pressure cuff. The time required until bleeding stops is measured; normal time ranges from 2 to 6 min. Prolonged time indicates abnormal vessel contraction, low platelet count, or abnormal platelet function. Drugs, such as aspirin, or von Willebrand factor deficiency can prolong bleeding time. This test may not detect coagulation disorders, such as hemophilia, because it measures the integrity of the initial platelet clot.

Platelet function can also be measured with a *platelet aggregometer.* Platelet-rich plasma is stirred in a cuvette, and the transmittance of light through the sample is measured. When an aggregating agent is added, the platelets aggregate, causing the plasma to clear, and the increase in light transmittance is recorded. Agents, such as ADP, thrombin, epinephrine, and ristocetin, induce the platelets to aggregate or cause them to release endogenous ADP. Abnormal platelet aggregation profiles can be indicative of various disorders, such as von Willebrand's disease, idiopathic thrombocytopenic purpura, and platelet storage pool disease, or of drug presence (aspirin, heparin, etc.).

Other specialized platelet function tests are sometimes used to determine platelet retention, adhesiveness, and clot retraction. Specialized tests for platelet factor IV or glycoprotein surface receptors are available to help diagnose specific disorders. These tests are normally performed by using specific immunologic reagents.

5.5. Control of Anticoagulant Therapy

5.5.1. Oral Anticoagulants

Oral anticoagulants act by preventing a vitamin K dependent carboxylation step in the synthesis of factors II, VII, IX, and X. This does not affect the quantities of the factors produced, but does result in them having no biologic activity. Management of the anticoagulant dose regulates the balance between active and inactive factor production.

The *PT test* is preferred for monitoring oral anticoagulant therapy as it is sensitive to three of the affected factors (II, VII, and X). Variation in the manufacture of thromboplastins used in this test results in wide variation sensitivity to the various factors, with resulting changes in the therapeutic range. A proposal to have manufacturers calibrate their thromboplastins with respect to international standards may eventually lead to improved standardization.

An alternative approach is to *control therapy on the basis of a single affected factor,* using a synthetic substrate method. This may improve specificity for the effect of the drug by eliminating nonspecific interferences that occur with the PT test. Comparison studies have shown good correlation between the PT test and factors II and X for patients on long-term therapy.

5.5.2. Heparin

Global Clotting Assays. Standard heparin [*9005-49-6*] therapy is usually controlled by the APTT test. Alternative assays in use include the thrombin clotting time and whole blood clotting time tests. For patients undergoing extracorporeal circulation, with its associated high heparin levels, the activated clotting time (ACT) is assumed to be the same as the APTT. This assumption is unreliable at high dose. Clotting assays, such as APTT and PT, ultimately rely on thrombin formation and are susceptible to various nonspecific interferences. They are also insensitive to low molecular mass heparin (M_r 6 000 – 20 000).

Specific Assays. Heparin can be assayed directly because of its ability to catalyze the AT III inhibition of thrombin or factor Xa. These assays can be classified as measuring either total heparin activity (when the assay uses excess AT III) or effective heparin activity (test plasma is the sole source of AT III). Typically these assays employ synthetic substrates.Factor Xa is the preferred enzyme reagent because it allows for measuring both standard and low molecular mass

heparin preparations. Data are not yet available on the relative clinical efficacy of these heparin-specific assays versus the global clotting assays.

6. Reagents for Nuclear Medicine

6.1. Nuclear Imaging (Radioscintigraphy)

Purpose. The object of nuclear medicine procedures is to *diagnose disease and dysfunction by external imaging of the distribution of a radioactive compound in the body* [88–91]. The radionuclides used in imaging procedures are selected from those that decay by emission of γ-rays, which penetrate the body and permit accurate and noninvasive external detection with imaging devices [92].

Diagnostic radiopharmaceuticals may be simple *radioactive ions* in aqueous solution or *labeled compounds*, which are radionuclide-containing inorganic, organic, or biologic molecules. Certain radiopharmaceuticals have biologic properties that permit accumulation, and therefore visualization, of the internal organs. Within these organs, areas of disease or decreased function may be detected as regions of lower concentration of radioactivity or *cold spots.* In other cases, such as in bone disorders, abnormal areas may accumulate radioactivity to a greater degree than normal tissue and appear as *hot spots.* In other procedures, radiopharmaceuticals are used to measure or visualize a *dynamic function,* e.g., blood flow through the heart, renal excretion, or pulmonary ventilation. In these procedures, the rate of accumulation of the radiopharmaceutical and clearance of activity is diagnostic of normal or abnormal function.

Because γ-emitting nuclides can be detected with such high sensitivity, doses of radiopharmaceuticals can be kept well below levels that produce any pharmacologic or adverse effects in the patient.

Typical Procedure. Diagnostic nuclear imaging procedures are performed in hospitals or nuclear medicine clinics that are licensed for radioisotope use. In most procedures, the radiopharmaceutical is administered to the patient intravenously, although other routes, such as oral administration or inhalation, are appropriate with certain agents. After administration, the diagnostic agent is permitted to distribute within the body for a specified period of time, usually from 10 min to 3 h. This is required for a sufficient accumulation of radioactivity in the target organ or site and sufficient clearance from the blood or immediately adjacent sites to provide adequate contrast. Typically, good images can be produced if the target to nontarget ratio of activity is 5 : 1 or greater.

After the appropriate time, the patient is positioned under an imaging device, the γ-camera, which can detect the spatial distribution of the radiopharmaceutical in the body. Most commonly, a single view (or multiple views) of the area of interest is obtained to provide a two-dimensional image of the organ. With newer rotating tomographic cameras, imaging data are obtained at preset angles around the body. A three-dimensional image of the organ is then reconstructed from the imaging data.

6.2. Detection Devices

Anger Camera. The most common imaging device in nuclear medicine is the Anger scintillation camera, which provides two-dimensional planar images for a variety of diagnostic applications. The *detection system* of the basic Anger camera is shown in Figure 1.

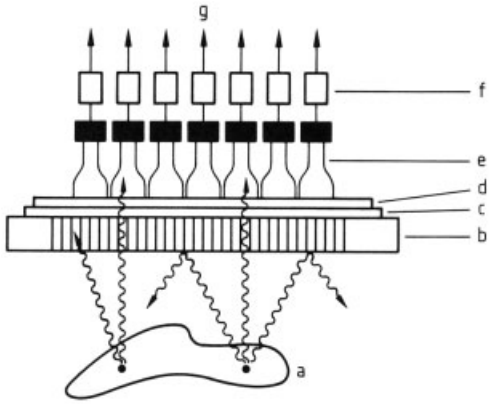

Figure 1. Schematic of the Anger scintillation camera
a) Source of γ-radiation; b) Collimator; c) NaI (Tl) Crystal; d) Light pipe; e) Photomultiplier tubes; f) Preamplifiers; g) Computing circuits.

The *collimator* (b) is a thick plate of high-density material, generally lead, which absorbs radiation. An array of channels is geometrically arranged over the surface. The role of the collimator is to limit the field of view and allow only photons emitted from a defined spatial location within the body to reach the detector crystal. Detection sensitivity and image resolution are a function of the distance of the collimator from the source of radiation, as well as a function of the collimator design.

The *thallium-activated sodium iodide crystal* (c), NaI(Tl), traps the γ-rays and converts them to small flashes of visual light, i.e., scintillations. These scintillations are subsequently converted from light to current flow by the *photomultiplier tubes* (e). In the next phase, the *computing circuit* (g), in combination with the photomultiplier array, determines the $x-y$ location of the scintillations occurring within the crystal.

A range of cameras, which offer various fields of view, is available commercially. The standard field of view (SFOV) camera can detect a field with a diameter of ca. 25 cm. The large field of view (LFOV) camera can image an area with a diameter of 35.6 – 41.9 cm, whereas a new extra large field of view camera (the jumbo camera) can image a 54.6 cm field. The particular advantage of large field cameras is that a whole body survey can be obtained rapidly with a single pass of the detector. The total camera system resolution of most state-of-the-art Anger cameras ranges from ca. 13 to 18 mm. However, if contrast is good, lesions of considerably smaller size can be detected.

SPECT Imaging. A limitation of current Anger cameras is that they provide only two-dimensional images of the three-dimensional object. Frequently underlying and overlying structures accumulate activity and obscure the image of the target organ or site. Advances in scintillation camera technology have resulted in single photon emission computed tomographic (SPECT) systems [93]. These systems use single- or multiheaded scintillation cameras, which provide multiple views at different angles of rotation around the subject. Computerized data collection then permits a reconstruction of a three-dimensional image or a slice-by-slice representation.

Positron Camera. Positron emission tomography (PET) involves the use of radiopharmaceuticals labeled with nuclides that decay by positron (positively charged electron) emission. In this decay process, a positron is emitted from the nucleus and is annihilated in a collision with a negatively charged orbital electron. In the annihilation reaction, the two particles are converted into two 8.19×10^{-14} J γ-rays, which travel away from each other at an angle of 180 °. In PET imaging, two opposed detectors simultaneously record the two high-energy γ-rays. Because only those γ-rays that strike the two opposed detectors simultaneously are recorded (coincidence counting), no collimator is necessary in a positron camera. Modern PET systems involve multidetector arrays arranged in a circle. The total system resolution of the multidetector PET camera is 7 – 10 mm. Consequently, both resolution and sensitivity of positron emission tomography is greater than that of single photon imaging.

Despite major imaging advantages of positron emission tomography, the procedure is currently limited to large medical research institutions. In addition to the considerable cost of modern PET systems ($> \$1\times10^6$), the positron emitting radionuclides of greatest interest (see Section 6.3) have exceedingly short half-lifes, e.g., ^{11}C, $t_{1/2} = 20$ min; ^{13}N, $t_{1/2} = 10$ min; ^{15}O, $t_{1/2} = 2$ min; and ^{18}F, $t_{1/2} = 110$ min. These isotopes must be produced by a cyclotron at the site of use. The establishment of a cyclotron facility and positron imaging laboratory typically requires an investment of $> \$2\times10^6$. In 1986 ca. 50 institutions worldwide had such facilities.

6.3. Radionuclides

6.3.1. Selection Criteria

The first consideration in selecting a γ-emitting isotope for use in conjunction with the Anger camera is its *decay scheme,* particularly the energy and abundance of its γ-emissions [94]. For good detection and collimation, the ideal isotope should emit γ-radiation with energy of $(1.6-3.2)\times10^{-13}$ J. Photons of lower energy are significantly attenuated, i.e., scattered within the body, and are poorly resolved by the camera; photons of energy $>6.4\times10^{-13}$ J penetrate

most collimators and flood the crystal with uncollimated radiation. The isotope should also be free of any "useless" radiation, such as particulate radiation (α or β) or γ-photons, which are too high or too low in energy, as these may degrade the image and also contribute significantly to the radiation dose received by the patient.

A further consideration is the *half-life* of the radionuclide. In most procedures currently in use, imaging is performed between 10 min and 4 h after administration of the radiopharmaceutical. Therefore, nuclides with short half-lifes, such as ^{99m}Tc and ^{123}I, are generally the most appropriate. Isotopes with ultrashort half-lifes, i.e., $t_{1/2} < 1$ h, and isotopes with intermediate half-lifes, i.e., $t_{1/2} > 4$ d, are not ideal. γ-Emitting radionuclides of greatest medical interest are shown in Table 5.

Table 5. γ-Emitting radionuclides for diagnostic imaging

Nuclide	$t_{1/2}$, h	γ-Energy, $\times 10^{-14}$ J
Ideal	~ 4	1.6 – 3.2
^{99m}Tc*	6	2.24
^{123}I	13	2.55
^{203}Pb	52	4.47
^{111}In	68	2.77
		3.96
^{97}Ru	70	3.46
^{201}Tl	73	1.1 – 1.3
^{67}Ga	78	1.49
		2.96
		4.81
		6.31

* m = Metastable.

Some nuclides are produced by nuclear reactors or particle accelerators, such as cyclotrons or linear accelerators. Other short-lived and ultrashort-lived nuclides are obtained from the decay of a parent nuclide through the use of a radioisotope generator [95]. The generator is generally a chromatographic column packed with media, such as alumina or an organic or inorganic ion-exchange resin, on which the parent isotope is adsorbed. As the parent decays, the daughter isotope concentration increases on the column and may be eluted with a suitable solvent, generally an aqueous solution. A generator, therefore, permits the delivery of a short-lived daughter nuclide by using the longer half-life of the parent.

6.3.2. ^{99m}Tc

^{99m}Tc [*14133-76-7*]is the isotope of choice because of its *excellent imaging characteristics* and its *ready availability*. The chemistry of technetium is complex but versatile, and a wide range of ^{99m}Tc radiopharmaceuticals is both known and practical.

Technetium [^{99m}Tc] is obtained from its parent isotope ^{99}Mo, which is a product of nuclear fission. Molybdenum [^{99}Mo] decays by γ-emission, with $t_{1/2} = 66.5$ h, to the isotope of greatest medical interest, the metastable ^{99m}Tc, which decays to ^{99}Tc by emitting a single γ-photon (2.24×10^{-14} J, $t_{1/2} = 6$ h).

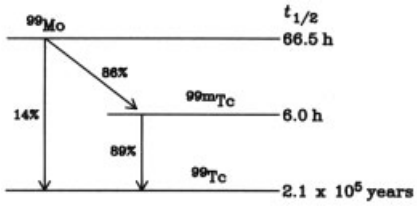

Although ^{99}Tc is a β-emitter, the quantity of this long-lived radionuclide ($t_{1/2} = 2.1 \times 10^5$ years) present in radiopharmaceutical preparations is so low that solutions are considered nonradioactive.

Most commonly, ^{99m}Tc is delivered as a ^{99}Mo – ^{99m}Tc radionuclide generator. The parent isotope is adsorbed on an alumina column and a maximum concentration of ^{99m}Tc forms every 23 h. Typically, the generator is eluted once daily with sterile physiologic saline solution, and the resulting ^{99m}Tc eluate is available for preparing other radiopharmaceutical compounds.

The ^{99m}Tc eluate contains $^{99m}Tc^{7+}$ as $^{99m}TcO_4^-$, which is used directly for some diagnostic applications; however, most radiopharmaceutical uses of this nuclide require the preparation of other ^{99m}Tc-labeled compounds. A wide variety of ^{99m}Tc radiopharmaceuticals can generally be prepared by using commercially available kits, which contain chemicals required to reduce $^{99m}Tc^{7+}$. The reduced ^{99m}Tc is subsequently complexed with an organic molecule, which produces the radiopharmaceutical with the desired avidity for an organ or a disease site. Typically, the desired radiopharmaceutical

is prepared on the day of use with a commercially available freeze-dried kit and ^{99m}Tc eluate from a ^{99}Mo – ^{99m}Tc generator. The ^{99m}Tc radiopharmaceutical forms rapidly and quantitatively, and no further purification is required prior to administration [96].

The γ-energy of ^{99m}Tc is ideally collimated and detected by the Anger camera system. Because ^{99m}Tc is a pure γ-emitter with a short half-life, doses of 3 – 30 mCi are acceptable without significantly irradiating the patient. Technetium (^{99m}Tc) is the isotope of choice as long as imaging can be completed within 24 h of the administration of the agent and the desired agent can be prepared satisfactorily.

Table 6. Ultrashort half-lifed generator-produced nuclides for diagnostic imaging

Nuclide	$t_{1/2}$, s	γ-Energy∗, $\times 10^{-14}$ J
^{191m}Ir	4.90	2.07
^{81m}Kr	13.00	3.04
^{195m}Au	30.00	3.46
^{82}Rb	75.00	8.19

∗ Of major imaging photons.

6.3.3. Other γ-Emitters

Among alternatives to ^{99m}Tc, ^{123}I [*15715-08-9*] (see Table 5) isan attractive isotope for general use. A singular advantage of ^{123}I, in addition to its ideal energy, is the availability of well-established iodination reactions. Iodine can be easily added to a wide variety of molecules without disturbing the biologic properties of these compounds. Because no generator system is possible for ^{123}I, it must be prepared frequently with a cyclotron. In addition, most low-energy cyclotrons produce ^{123}I contaminated with 3 – 5 % of the longer lived ^{124}I. This radionuclide cannot be separated from ^{123}I by conventional chemical means and, when present in higher quantities, may degrade the image (because of high-energy γ-photons) as well as increase the radiation dose to the patient.

For those applications that require imaging 24 h or longer after administration of the agent, other radionuclides with intermediate half-lifes may be considered. Most attractive among these is ^{111}In [*15750-15-9*] (see Table 5), whichis available in good abundance; ^{203}Pb [*14687-25-3*] and ^{97}Ru [*15758-35-7*]also are suitable. However, because In, Pb, and Ru ions are not present inmolecules of biologic interest, their use requires preparing novel radio-labeled compounds.

In newer applications, γ-emitting radionuclides with $t_{1/2}$ <10 min have been used for studying processes that have very rapid biokinetics, such as blood flow, cardiac-wall motion, and pulmonary ventilation. These *ultrashort-lived nuclides* (see Table 6) have the unique advantage of producing low radiation exposure to the patient, which permits administration of high millicurie doses. In addition, because the radioactivity decays within minutes, repeat studies may be performed quickly.

6.3.4. Positron Emitters

Large medical centers with positron-imaging capabilities and a nuclear reactor or cyclotron on site can use positron-emitting nuclides for medical diagnosis. *Cyclotron-produced* isotopes, such as ^{11}C [*14333-33-6*] ($t_{1/2}$ = 20.50 min), ^{13}N [*13981-22-1*] ($t_{1/2}$ = 9.95 min), ^{15}O ($t_{1/2}$ = 2.05 min), and ^{18}F [*13981-56-1*] ($t_{1/2}$ = 110.00 min), have been used to label biologic molecules to produce true physiological tracers that retain native biologic activity. Because of the short half-lifes, on-site production capability for these positron isotopes is required.

If a cyclotron is not available on site, two *generator-produced,* positron-emitting isotopes are available in radiochemical (not radiopharmaceutical) grade: ^{68}Ga [*15757-14-9*]($t_{1/2}$ = 68.00 min) and ^{82}Rb [*14391-63-0*] ($t_{1/2}$ = 1.30 min).

6.4. Diagnostic Radiopharmaceuticals

6.4.1. Heart Imaging

Diagnostic agents are available for use in studying the functions of the heart. *Thallium* [^{201}Tl^{+}], an analogue of K^{+}, is efficiently extracted by the myocardium and accumulates in the heart muscle in proportion to its concentration in the blood. When viewed from the left anterior oblique (LAO) position relative to the chest, the walls of the left ventricle, in a normal subject, appear as a “donut” (Fig. 2A). In a patient with significant coronary artery disease, blood flow may be entirely adequate at rest; however, with

exercise (or pharmacologic) stress, the affected regions may appear transiently to have inadequate blood flow, appearing as a "broken donut" (Fig. 2B). Old myocardial infarctions appear as cold spots both at rest and stress.

The wall motion and blood pumping functions of the left and right ventricle can be studied by using ^{99m}*Tc*-labeled red blood cells (RBCs). Typically, the patient receives a first dose of tin (II) pyrophosphate ($Sn_2P_2O_7$), which is taken up by the RBCs. The Sn^{2+} reduces $^{99m}TcO_4^-$, which is administered 30 min later. To study the heart at various phases of the contractile cycle, the γ-camera is controlled by an electrocardiograph. Quantitative measurements of the injection fraction may be obtained, in addition to qualitative observations of wall motion during contraction.

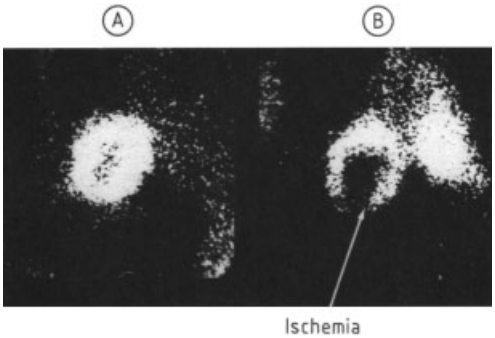

Figure 2. Heart images (^{201}Tl) of a patient with coronary artery disease (LAO views)
A) At rest; B) After exercise stress.

In some cases, ventricular function is assessed by observing the first passage of a bolus of radioactivity through the heart. Any ^{99m}Tc radiopharmaceutical (except one that is extracted by the lungs) may be used for first-pass studies. The generator-produced ultrashort-lived isotope ^{195m}Au (see Table 6) has been used for multiple-repeat first-pass studies. Because of its high γ-energy, it can be used simultaneously with ^{201}Tl or ^{99m}Tc. A disadvantage, however, is that the short half-life of the parent nuclide, ^{195m}Hg ($t_{1/2} = 40$ h), limits the shelf life of the generator to 3 d. The ^{195m}Hg – ^{195m}Au generator has recently become commercially available in Europe.

Fatty acids labeled with ^{123}*I* are available in Europe for clinical studies of aerobic metabolism in the myocardium. Research has shown that ^{11}*C-palmitic acid* can give early and specific information on developing cardiomyopathies, as well as ischemic disease.

6.4.2. Brain Imaging

Use of radiopharmaceuticals, such as $^{99m}TcO_4^-$, ^{99m}Tc-DTPA (diethylenetriamine-*N,N,N′,N″,N″*-pentaacetic acid) [*80908-06-1*], and ^{99m}Tc-glucoheptonate for brain imaging-has declined because of increased availability of *computerized transaxial tomography* (CAT scanning). Radiopharmaceuticals are still in use, however, and provide information on the integrity of the blood brain barrier. Extravascular accumulation of activity suggests a brain tumor, abscess, or other pathology.

Some experimental radiopharmaceuticals currently under development for brain imaging are discussed in Section 6.5.

6.4.3. Bone Imaging

The *agents of choice for bone imaging* are ^{99m}Tc-labeled MDP (methylene diphosphonate), EHDP (ethane 1-hydroxy-1-diphosphonate), and HMDP (hydroxymethylene diphosphonate) although ^{99m}Tc-pyrophosphate is also used. All can be *prepared from commercially available kits.*

In normal subjects, these agents are quickly absorbed by bone (but not by blood and soft tissue) so that images of the skeleton can be produced within 1 h after administration. In patients with bone disorders, areas with low concentration indicate bone destruction or necrosis. Other lesions, such as septic or aseptic inflammation, fracture healing, and tumor growth, are detected as areas of high radioactivity. This is probably because of increased bone blood flow and increased extraction efficiency in those areas of higher metabolic activity.

An image of a patient with multiple tumor metastases is shown in Figure 3. Nuclear medicine procedures provide earlier detection of metastatic lesions, where no significant bone destruction or density change has occurred, than does conventional radiology (X-ray). The image also illustrates the use of a wide field of view camera to produce a whole body survey for diagnosis of occult lesions.

6.4.4. Lung Imaging

Imaging of lung perfusion and ventilation, either separately or together, is *used to diagnose pulmonary disorders.* Suspensions of labeled macroaggregates of albumin (^{99m}Tc-MAA) or human albumin microspheres (^{99m}Tc-HAM) are injected intravenously to study the perfusion of the lung capillary bed. If the vascular system is obstructed, the concentration of activity in the involved lung segments is decreased. Although this technique is helpful in diagnosing pulmonary embolism, abnormal lung perfusion is nonspecific and may occur with other disorders.

Figure 3. Bone image (^{99m}Tc-MDP) of a patient with multiple metastases to the skeleton (seen as hot spots)

Three *radioactive noble gases* have been used to study pulmonary ventilation: ^{127}Xe [*13994-19-9*], ^{133}Xe [*14932-42-4*], and ^{81}Kr [*15678-91-8*]. Xenon [^{133}Xe] is widely used clinically for single-breath studies and also for lung volume and washout imaging. The physical characteristics of ^{127}Xe are preferable to those of ^{133}Xe, but availability is more limited. Krypton [^{81m}Kr], a generator-produced isotope with a 13-s half-life, can be used for single breath imaging studies.

Radioactive aerosols, prepared with ^{99m}Tc agents, have been used in place of the radioactive noble gases for ventilation studies. The advantages of ^{99m}Tc aerosols include the ready availability and optimum characteristics of ^{99m}Tc and the longer residence time of the agent in the lung. Because the radioactive aerosol deposits in the smaller bronchi and alveoli, multiple images in several projections may be obtained. However, with chronic obstructive pulmonary disease, deposition of the aerosol in the major airways may be irregular, making interpretation of images difficult.

Combined ventilation and perfusion studies are frequently used to differentiate the perfusion disorders of pulmonary embolism from obstructive airway disorders.

6.4.5. Hepatobiliary Imaging

After intravenous administration, hepatobiliary imaging agents are rapidly cleared from the blood by the hepatocytes, or polygonal cells, which comprise 60 % of the liver mass in humans. The radiopharmaceuticals are then concentrated in the bile, transiently stored in the gall bladder, and excreted via the bile ducts into the small intestine. If all of the organs, i.e., the liver, gall bladder, bile ducts, and intestines, are not visualized sequentially, or if transit is delayed, the study is abnormal. If no activity appears in the intestinal tract, even after a prolonged period of time, complete obstruction of the bile duct is likely. Nonvisualization of the gall bladder in a patient who has been fasting for several hours suggests acute cholecystitis. Poor blood clearance and poor liver uptake is consistent with either primary hepatocellular obstruction or biliary obstruction. In jaundiced patients, serum bilirubin competes for the anion receptor on the hepatocyte membrane, resulting in decreased liver uptake and increased renal clearance of the radiopharmaceutical.

A series of ^{99m}Tc complexes of *N*-(2,6-dimethylphenylcarbamoylmethyl)iminodiacetic acid [*59160-29-1*] (HIDA) and its derivatives have been developed.

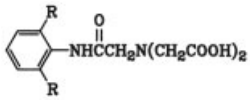

HIDA, R = $-CH_3$

Complexes of ^{99m}Tc with two HIDA derivatives, where R is ethyl or isopropyl, have been used extensively in patients in Europe and the United States. Other derivatives, e.g., the 3-bromo-2,4,6-trimethyl analogue of HIDA, have also been used experimentally.

6.4.6. Reticuloendothelial System

Radioactive colloids with particle sizes $<3\,\mu$m in diameter have been used to visualize major organs of the reticuloendothelial (RE) system, i.e., the liver, spleen, and bone marrow. After intravenous injection, the colloidal suspensions are rapidly extracted from the blood by phagocytosis. In the presence of decreased hepatic mass, decreased hepatic blood flow, or focal lesions within the liver, regional hepatic activity decreases and the activity in the spleen and bone marrow increases.

The ^{99m}*Tc-sulfur colloid* is the most widely used agent for RE imaging. A kit for the instant ^{99m}Tc labeling of a preformed colloid, is commercially available.

6.4.7. Renal Imaging

Three classes of radiopharmaceuticals have been employed for studying renal function. *2-Iodohippuric acid* [*147-58-0*], *labeled with* ^{131}*I* [*1203-90-3*], *or* ^{123}*I* [*54378-29-9*], is the only commercially available agent for estimating effective renal plasma flow. The glomerular filtration rate may be measured with radioactive metal chelates, e.g., ^{99m}Tc-DTPA.

Renal imaging is most frequently performed with ^{99m}*Tc-labeled glucoheptonate, 2,3-dimercaptosuccinic acid* [*75014-98-1*] *(DMSA), or DTPA*. All of these agents are cleared rapidly from the blood into the urine; however, with glucoheptonate and DMSA, sufficient activity is retained within the kidneys to permit visualization of the renal cortex and the collecting system for several hours.

6.4.8. Malignant and Inflammatory Regions

Gallium ($^{67}Ga^{3+}$) [*14119-09-6*] has been used to image and monitor the recurrence of soft tissue tumors, especially lymphomas, malignant melanomas, testicular and leukemic tumors, and sites of Hodgkin's disease. When injected intravenously, ^{67} Ga binds to transferrin and to other circulating Fe-binding proteins. The plasma clearance of ^{67}Ga is slow; ca. 25 % is excreted in the urine and feces in the first week after injection. Because of its high intestinal activity, which may be decreased by the use of laxatives and enemas, ^{67}Ga images thoracic and neck lesions generally better than abdominal sites. Typically, ^{67}Ga imaging is performed between 24 and 72 h after injection, depending on the need for background clearance from various sites.

^{67}Ga has been used more frequently for detecting abscesses and focal inflammatory lesions. The nuclide is known to concentrate in recent surgical incisions and in cases of acute thyroiditis, pneumonia, turberculosis, sarcoidosis, inflammatory cardiomyopathy, pyelonephritis, and osteomyelitis.

Although it continues to be useful, the *disadvantages* of ^{67}Ga include its long half-life, high-energy γ-emission, slow plasma and abdominal clearance, and the inability to differentiate tumors from inflammatory lesions.

More recently, an *uncharged lipophilic complex of* ^{111}*In with oxine* (8-hydroxyquinoline) has been used as a nonspecific label for blood cells. Populations of leukocytes, especially neutrophils, can be labeled in vitro with ^{111}In-oxine in a suspension containing only limited plasma proteins. Although separating and labeling leukocytes is time-consuming, ^{111}In-oxine is advantageous over ^{67}Ga. The labeled leukocytes concentrate in inflammatory lesions and can be optimally imaged within 24 h; they are not eliminated via the intestine. The specificity of abscess detection with ^{111}In-labeled leukocytes is >95 %.

6.5. Experimental Agents

Heart Imaging with ^{99m}Tc Agents. Extensive research in both academic and industrial laboratories has been directed to developing new radiopharmaceuticals for diagnostic imaging. In the area of diagnostic nuclear cardiology, new

^{99m}Tc agents for the study of myocardial perfusion are being sought to replace ^{201}Tl, whose isotopic properties (low γ-energy, long half-life) are nonideal. A series of ^{99m}*Tc-isonitrile complexes,* which concentrate in the myocardium in proportion to regional blood flow, have been reported [97]. The borders of the ventricular wall can be well resolved with these complexes because of the superior characteristics of ^{99m}Tc and its higher permissible dose. Therefore, these agents may be useful for simultaneous studies of both myocardial perfusion and ventricular function, which currently require two separate procedures.

Brain Perfusion Imaging. Radiopharmaceuticals have been studied for regional cerebral brain perfusion [98]. *N-Isopropyliodoamphetamine* labeled with ^{123}*I* (**2**) and a related amine, ^{123}*I-HIPDM* (*N,N, N′*-trimethyl-*N′*-[2-hydroxyl-3-methyl-5[^{123}I]-iodobenzyl]-1,3-propanediamine), (**3**) distribute in the brain in proportion to regional blood flow. These compounds may be used for single photon emission computed tomographic (SPECT) imaging of the brain because their residence time in the brain is long and their ability to differentiate gray-white matter is good.

There has also been considerable interest in developing ^{99m}Tc-labeled agents with properties analogous to ^{123}I-IMP (*N*-isopropyl-4-[^{123}I] iodoamphetamine). The ^{99m}Tc complexes of HMPAO ([^{99m}Tc]-oxo-3,6,6,9-tetramethyl-4,8-diazaundecane-2,10-dionedioxime) [*99944-78-2*] (**4**) [99]and NEP ([^{99m}Tc]-oxo-3,6-diaza-1,8-dithio-1,1,5,5,8,8-hexamethyl-3-*N*-piperidinylethyloctane) [*67-43-6*] (**5**) [100] appear promising for brain imaging studies.

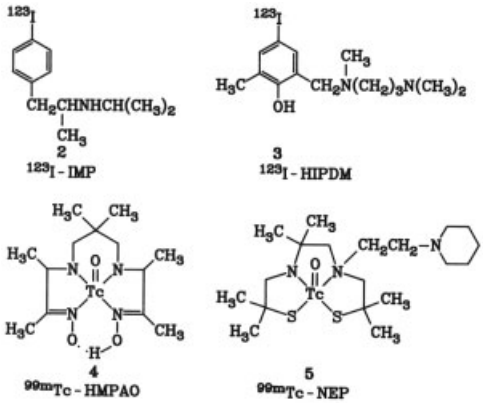

Specific Brain Receptors. Extensive research has been directed toward developing and using *positron-emitting radiopharmaceuticals* that will image specific brain receptors. An example is ^{11}C-labeled *N*-methylspiperone, which is used in vivo to image dopamine receptors [101]. These and other agents may provide means for the noninvasive quantitation of specific receptors and for the study of their function in health and disease. It has been suggested that receptor-specific radiopharmaceuticals may also be useful for the objective diagnosis of senile dementias and psychiatric disorders as well as for drug monitoring.

Brain Metabolism. ^{11}*C-2-Deoxyglucose* and ^{18}*F-2-deoxyglucose* have been used experimentally in animals and humans to measure brain metabolism in disorders, such as stroke, epilepsy, and senile dementias, as well as in normal volunteers undergoing sensory stimulation.

Monoclonal Antibodies. Advances in biotechnology led to new areas of radiopharmaceutical research. Monoclonal antibodies targeted toward melanomas, as well as toward tumors of the lung, breast, and colon, are being explored as *carriers of diagnostic and therapeutic radionuclides.* Figure 4 shows the localization of a murine antihuman breast carcinoma antibody conjugated with ^{99m}Tc-metallothionein in a mouse bearing a human breast tumor [102].

Therapeutic Agents. Therapeutic nuclides are being studied because of the anticipated greater specificity of radiopharmaceuticals based on monoclonal antibodies. The therapeutic nuclides of greatest interest (see Table 7) are β-emitters with sufficient β-energy to produce a cytotoxic effect with a radiation range of several cell diameters.

Table 7. Nuclides for therapeutic applications

Nuclide	$t_{1/2}$, d	β-Energy (max), $\times 10^{-14}$ J
^{90}Y	2.7	3.64
^{188}Re	0.71	3.40
^{32}P	14.3	2.74
^{111}Ag	7.5	0.40
^{131}I	4.15	

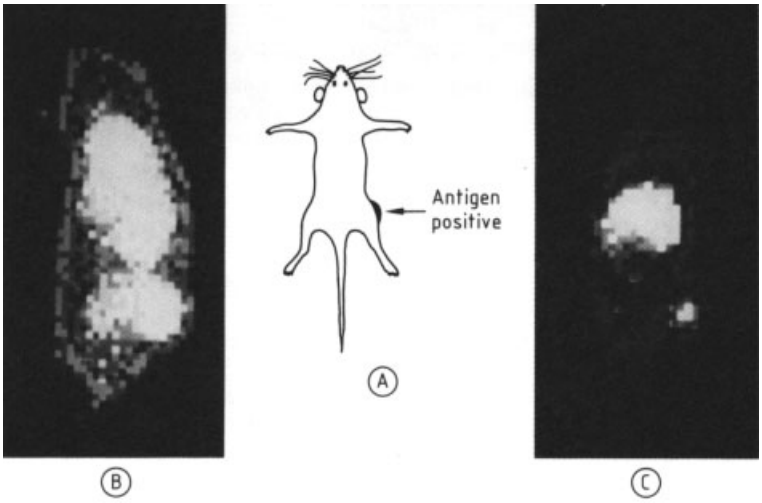

Figure 4. Detection of a subcutaneous human breast tumor in a mouse with the B6.2 antibody conjugated with ^{99m}Tc-metallothionein
A) Antigen positive; B) ^{125}I-Antibody; C) ^{99m}Tc-Antibody.

6.6. Quality Control

Radiopharmaceuticals used in diagnostic and therapeutic procedures must meet most requirements for drugs and injectables. Agents must be sterile, free of compounds that cause fever, free of foreign matter, and well labeled. Simple and rapid thin-layer and paper chromatographic procedures have been available for detecting unreacted $^{99m}TcO_4^-$ and reduced hydrolyzed $^{99m}TcO_2$ in radiopharmaceutical solutions [103]. Newer TLC and HPLC procedures have been employed to assess the radiochemical purity of radiopharmaceutical solutions, i.e., the presence of a single, defined radiochemical compound.

Biodistribution as well as toxicological tests are generally performed on random samples from new lots of radiopharmaceuticals. Although these tests are necessary, the rate of adverse reactions to radiopharmaceuticals is low because of the lower administered doses of chemicals employed in diagnostic procedures as compared to typical therapeutic doses. From 1976 to 1979, the number of adverse reactions reported to the Registry of the Society of Nuclear Medicine was ca. 1 – 6 (per 100 000), of which only 3 % were severe.

In addition, routine quality control monitoring of the scintillation camera system is required. The pulse height analyzer is adjusted to give a symmetric window about the primary photo peak corresponding to the nuclides to be imaged. Uniformity of detector response and checks for linearity and spatial distortion, as well as good spatial resolution, are performed daily prior to beginning studies on patients.

6.7. Dosimetry

The administration of isotopes to patients for diagnostic studies is limited by the *potential radiation damage* that these isotopes can cause to various organs and tissues. The radiopharmaceutical, which the target tissue absorbs, is the primary source of this radiation. In addition, the target tissue can be irradiated by other proximal organs if they contain sufficient radioactivity. The amount of the absorbed dose is determined by the concentration of activity and by the residence time of the radiopharmaceutical in the target and proximal sites. For most radionuclides used for diagnostic purposes (which are substantially free of particulate radiation), the dose to the patient is in direct proportion to the half-life of the radionuclide. Therefore, short-lived and ultrashort-lived radionuclides deliver lower doses of radiation per millicurie administered. A significant limitation is placed on the permissible radioactivity dose of intermediate

half-lifed radionuclides because of their higher absorbed radiation dose per millicurie. Standard methods for performing internal dose calculations are described in the publications of the Medical Internal Radiation Dose (MIRD) Committee of the United States Society of Nuclear Medicine [104].

7. References

1. Committee on Analytical Reagents (ed.): *Reagent Chemicals,* 4th ed., Amer. Chem. Soc., Washington, D.C., 1968.
2. D. D. Perrin, W. L. F. Armarego, D. B. Perrin: *Purification of Laboratory Chemicals,* 2nd ed., Pergamon Press, New York 1982.
3. H. C. Damon, J. W. King (eds.): *Practical Manual for Clinical Laboratory Procedures,* CRC Press, Cleveland, Ohio, 1965.
4. J. L. Bennington in J. A. Halsted, C. H. Halsted (eds.): *Role of Laboratory in Clinical Medicine,* W. B. Saunders Co., Philadelphia 1981.
5. Biotechnology & the Diagnostic Industry, report no. 113, Theta Technology Corp., Wethersfield, Conn., 1983.
6. K. J. Aziz, R. S. Kennedy, *Clin. Chem. (Winston-Salem)* **26** (1980) 539.
7. L. Edwards, *Clin. Chem. (Winston-Salem)* **29** (1983) 890.
8. Current NCCLS Publications, National Committee for Clinical Laboratory Standards, Villanova, Pa., 1984.
9. CAP Surveys, The Interlaboratory Comparison Program, College of American Pathologists, Skokie, Ill., 1985.
10. R. C. Menson, V. Narayanswamy, R. W. Bussian, T. H. Adams, *Clin. Chem. (Winston-Salem)* **20** (1974) 872.
11. G. Eisenman, D. O. Rubin, J. U. Casby, *Science* **126** (1957) 831.
12. M. S. Frant, J. W. Ross, Jr., *Science* **167** (1970) 987.
13. J. C. Sternberg, S. J. Updike, D. P. Lehane in M. Werner (ed.): *Microtechniques for the Clinical Laboratory Concepts and Applications,* J. Wiley & Sons, New York 1976, p. 129.
14. R. H. Hamilton, *Clin. Chem. (Winston-Salem)* **12** (1966) 1.
15. B. T. Doumas, W. A. Watson, H. G. Biggs, *Clin. Chim. Acta* **31** (1966) 87.
16. A. E. Pinnell, B. E. Northam, *Clin. Chem. (Winston-Salem)* **24** (1978) 80.
17. P. Carter, *Microchem. J.* **15** (1970) 531.
18. G. R. Kingsley, *J. Lab. Clin. Med.* **27** (1942) 840.
19. O. H. Lowry, N. J. Rosebrough, A. L. Farr, R. J. Randall, *J. Biol. Chem.* **193** (1951) 265.
20. R. J. Henry: *Clinical Chemistry, Principles and Technics,* Harper & Row Publ., New York 1974, p. 407.
21. B. C. Sarkar, U. S. P. Chauhan, *Anal. Biochem.* **20 A** (1967) 155.
22. H. V. Connerty, H. S. C. Lau, A. R. Briggs, *Clin. Chem. (Winston-Salem)* **17** (1971) 661.
23. C. H. Fiske, Y. Subbarrow, *J. Biol. Chem.* **66** (1925) 375.
24. P. Drewes, *Clin. Chim. Acta* **39** (1972) 81.
25. H. Hertz, R. Dybkaer, M. Lauritzen, *Scand. J. Clin. Lab. Invest.* **33** (1974) 215.
26. A. A. H. Van den Bergh, I. Snapper, *Dtsch. Arch. Klin. Med.* **110** (1913) 540.
27. H. T. Malloy, K. A. Evelyn, *J. Biol. Chem.* **119** (1937) 481.
28. L. Jendrassik, P. Grof, *Biochem. Z.* **297** (1938) 81.
29. N. W. Tietz: *Fundamentals of Clinical Chemistry,* W. B. Saunders Co., Philadelphia 1970, p. 206.
30. M. W. Slein, H. U. Bergmeyer: *Methods of Enzymatic Analysis,* Academic Press, New York 1963, p. 117.
31. T. C. Stadtman in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology,* **vol. 3,** Academic Press, New York 1957, p. 392.
32. W. Richmond, *Clin. Chem. (Winston-Salem)* **19** (1973) 1350.
33. J. H. Hagan, P. B. Hagen, *Can. J. Biochem. Physiol.* **40** (1962) 1129.
34. H. Talke, G. E. Schubert, *Klin. Wochenschr.* **41** (1965) 174.
35. H. R. Mahler, E. H. Cordis: *Biological Chemistry,* 2nd ed., Harper & Row Publ., New York 1971, p. 267.
36. P. A. Verduin, J. M. H. M. Punt, H. H. Kreutzer, *Clin. Chim. Acta* **46** (1973) 11 – 19.
37. N. W. Tietz: *Fundamentals of Clinical Chemistry,* 2nd ed., W. B. Saunders Co., Philadelphia 1976, p. 684.
38. R. J. Gay, R. B. McComb, G. N. Bowers, Jr., *Clin. Chem. (Winston-Salem)* **14** (1968) 740.
39. B. Hess, *Ann. N. Y. Acad. Sci.* **75** (1958) 292.
40. S. B. Rosalki, J. H. Wilkinson, *Nature (London)* **188** (1960) 1110.
41. A. V. Roy, M. E. Brower, J. E. Hayden, *Clin. Chem. (Winston-Salem)* **17** (1971) 1093.

42. G. N. Bowers, R. B. McComb, *Clin. Chem. (Winston-Salem)* **12** (1966) 70.
43. R. J. Henry, N. Chiamori, O. J. Golub, S. Berkman, *Am. J. Clin. Pathol.* **34** (1960) 381.
44. H. U. Bergmeyer, E. Burnt: *Methods of Enzymatic Analysis,* Academic Press, New York 1965, p. 846.
45. J. Gal, E. Roth, *Clin. Chim. Acta* **2** (1957) 316.
46. C. E. Pipenger, D. D. Schottelius, *J. Clin. Immunoassay* **7** (1984) 254.
47. D. M. Baer, W. R. Dito (eds.): *Interpretations in Therapeutic Drug Monitoring,* American Society of Clinical Pathologists, Chicago 1981.
48. G. Kohler, C. Milstein, *Nature (London)* **256** (1975) 495.
49. S. A. Berson, R. S. Yalow, *J. Clin. Invest.* **38** (1959) 1996.
50. S. Spector, *Ann. Rev. Pharmacol.* **13** (1973) 359.
51. N. L. Alpert, *Clin. Instrument Systems* **6** (1985) 1.
52. University Patents, US 4 451 571, 1984 (R. H. Allen).
53. Hybritech, US 4 376 110, 1983 (G. S. David, H. E. Greene).
54. J. M. Singer, C. M. Plotz, *Am. J. Med.* **21** (1956) 888.
55. W. J. Litchfield, A. R. Craig, W. A. Frey, C. C. Leflar et al., *Clin. Chem. (Winston-Salem)* **30** (1984) 1489.
56. K. E. Rubenstein, R. S. Schneider, E. F. Ullman, *Biochem. Biophys. Res. Commun.* **47** (1972) 846.
57. J. C. Braman, R. J. Broeze, D. W. Bowden et al., *Bio/Technology* **1** (1984) 349.
58. W. J. Litchfield, J. W. Freytag, M. Adamich, *Clin. Chem. (Winston-Salem)* **30** (1984) 1441.
59. B. K. Van Weeman, *J. Virol. Methods* **10** (1985) 371.
60. Corning, US 4 098 876, 1978 (R. N. Piasio, J. W. Ryan, J. E. Woiszwillo).
61. J. W. Freytag, J. C. Dickinson, S. Y. Tseng, *Clin. Chem.* **30** (1984) 417.
62. R. C. Kennedy, R. E. Lanford, R. D. Henkel, G. R. Dreesman, *Lab. Manage.* **23** (1985) 53.
63. B. D. Davis, R. Dulbecco, H. N. Eisen, H. S. Ginsberg: *Microbiology,* 3rd ed., Harper & Row Publ., Philadelphia 1980.
64. J. A. Washington (ed.): *Laboratory Procedures in Clinical Microbiology,* Springer-Verlag, New York 1981.
65. T. E. Kiehn, B. Wong, F. F. Edwards, D. Armstrong, *Clin. Microbiol.* **18** (1983) 300.
66. L. S. Diamond in J. B. Jensen (ed.): *In Vitro Cultivation of Protozoan Parasites,* CRC Press, Boca Raton, Fla., 1983, p. 65.
67. R. Cote (ed.): *ATCC Media Handbook,* 1st ed., American Type Culture Collection, Rockville, Md., 1984.
68. Difco Manual, 10th ed., Difco Laboratories, Detroit, Mich., 1984.
69. G. D. Hsiung: *Diagnostic Virology,* 3rd ed., Yale Univ. Press, New Haven-London 1982.
70. E. H. Lennette, A. Balows, W. J. Hausler, J. P. Truant: *Manual of Clinical Microbiology,* 3rd ed., American Society for Microbiology, Washington, D.C., 1980.
71. W. H. R. Lumsden in D. M. Weir (ed.): *Handbook of Experimental Immunology,* 3rd ed., Blackwell Scientific Publications, Oxford 1978, p. 421.
72. D. Richman, P. Cleveland, D. Redfield, M. Oxman et al., *J. Infect. Dis.* **149** (1984) 298.
73. N. R. Rose, H. Friedman: *Manual of Clinical Immunology,* 2nd. ed., American Society for Microbiology, Washington, D.C., 1980.
74. S. D. Henriksin in T. Bergen, J. R. Norris (eds.): *Methods in Microbiology,* **vol. 12,** Academic Press, London 1978, p. 1.
75. T. C. Jones in G. L. Mandell, R. G. Douglas, J. E. Bennett (eds.): *Principles and Practice of Infectious Disease,* J. Wiley & Sons, New York 1979, p. 2085.
76. G. P. Youmans, P. Y. Paterson, H. M. Somners: *The Biologic and Clinical Basis of Infectious Diseases,* 2nd ed., W. B. Saunders Co., Philadelphia 1980.
77. D. Richman, N. Schmidt, S. Plotkin, R. Yolken et al., *J. Infect. Dis.* **150** (1984) 941.
78. J. K. Kulski, M. Norval, *Arch. Virol.* **83** (1985) 3.
79. R. W. Colman, J. Hirsh, V. J. Mader, E. W. Salzman (eds.): *Hemostasis and Thrombosis,* J. B. Lippincott, Philadelphia 1982.
80. J. M. Thomson (ed.): *Blood Coagulation and Haemostasis,* Churchill Livingston, London 1980.
81. J. Fareed, H. L. Messmore, J. W. Fenton, K. M. Brinkhous (eds.): *Perspectives in Hemostasis,* Pergamon Press, New York 1981.
82. D. A. Triplett, C. S. Harms, P. Newhouse, C. Clark: *Platelet Function,* American Society of Clinical Pathologists, Chicago, Ill., 1978.
83. P. J. Gaffney, S. Balkuv-Ulvtin (eds.): *Fibrinolysis,* Academic Press, London 1978.
84. M. F. Scully, V. V. Kakkar (eds.): *Chromogenic Peptide Substrates: Chemistry and Clinical Usage,* Churchill Livingston, London 1979.

85. I. Witt (ed.): *Heparin: New Biochemical and Medical Aspects,* De Gruyter, Berlin 1983.
86. J. Hirsh, E. A. Brain: *Hemostasis and Thrombosis: A Conceptual Approach,* Churchill Livingston, Berlin 1983.
87. R. Biggs, R. G. MacFarlane: *Human Blood Coagulation, Haemostasis, and Thrombosis,* Blackwell, Oxford 1972.
88. P. J. Early, D. B. Sodee: *Principles and Practice of Nuclear Medicine,* C. V. Mosby Co., St. Louis, Mo., 1985.
89. N. P. Alazraki, F. S. Mishkin (eds.): *Fundamentals of Nuclear Medicine,* The Society of Nuclear Medicine, New York 1984.
90. L. M. Freeman (ed.): *Freeman and Johnson's Clinical Radionuclide Imaging,* 3rd ed., Grune & Stratton, New York 1984.
91. J. E. Carey, Jr., R. C. Kline, J. W. Keyes, Jr. (eds.): *CRC Manual of Nuclear Medicine Procedures,* 4th ed., CRC Press, Boca Raton, Fla., 1983.
92. R. M. Sarper: *Nuclear Medicine Instrumentation,* Charles C. Thomas, Springfield, Ill., 1984.
93. "Emission Computed Tomography: The Single Photon Approach," U.S. Dept. of Health & Human Services Publication, FDA 81-8177, U. S. Govt. Printing Office, Washington, D.C., Aug. 1981.
94. "Nuclear Medicine – Factors Influencing the Choice and Use of Radionuclides in Diagnosis and Therapy," NCRP report no. 70, National Council on Radiation Protection & Measurements, Bethesda, Md., Jun. 1982.
95. F. F. Knapp, Jr., T. A. Butler (eds.): *Radionuclide Generators – New Systems for Nuclear Medicine Applications,* ACS Symposium Series no. 291, Amer. Chem. Soc., Washington, D.C., 1984.
96. G. B. Saha: *Fundamentals of Nuclear Pharmacy,* 2nd ed., Springer, New York 1984.
97. B. L. Holman, A. G. Jones, J. Lister-James et al., *J. Nucl. Med.* **25** (1984) 1350 – 1355.
98. P. Magistretti (ed.): *Functional Radionuclide Imaging of the Brain,* Raven Press, New York 1983.
99. S. B. Chaplin, P. O. Oberle, T. J. Hoffman et al., *J. Nucl. Med.* **26** (1985) 18.
100. S. Z. Lever, H. D. Burns, T. M. Kervitsky et al., *J. Nucl. Med.* **26** (1985) 18.
101. D. F. Wong, H. N. Wagner, Jr., R. F. Dannals et al., *J. Nucl. Med.* **25** (1984) 73.
102. G. L. Tolman, R. J. Hadjian, M. M. Morelock et al., *J. Nucl. Med.* **25** (1984) 24.
103. P. J. Robbins: *Chromatography of Technetium-99m Radiopharmaceuticals,* The Society of Nuclear Medicine, New York 1984.
104. R. Loevinger, M. Berman: "NM/MIRD," Nuclear Medicine Supplement Pamphlet no. 1, The Society of Nuclear Medicine, New York 1984.

Dialkyl Peroxides → Peroxy Compounds, Organic

Dialkyl Sulfates and Alkylsulfuric Acids

Karl Weisenberger, Hoechst AG, Frankfurt, Federal Republic of Germany (Chaps. 1–9)

Dieter Mayer, Hoechst AG, Frankfurt, Federal Republic of Germany (Chap. 10)

Stanley R. Sandler, ATOFINA Chemicals, Inc., King of Prussia, Pennsylvania, United States

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. From Ethers and Sulfur Trioxide . . 3
4.2. From Olefins and Sulfuric Acid . . . 4
4.3. From Alcohols and Chlorosulfuric Acid 4
4.4. From Alcohols and Sulfuryl Chloride or Thionyl Chloride 5
4.5. From Alcohols and Sulfuric Acid . . 5
5. Quality Specifications and Analysis . 5
6. Storage, Transportation, and Handling 6
7. Environmental Protection 8
8. Uses 8
8.1. Alkylation 8
8.2. Other Uses 12
9. Economic Aspects 12
10. Toxicology and Occupational Health 12
10.1. Dimethyl Sulfate 12
10.2. Diethyl Sulfate 13
11. References 13

1. Introduction

Dialkyl sulfates **1** are neutral esters of sulfuric acid and aliphatic alcohols. The corresponding monoesters **2** (alkyl hydrogensulfates, monoalkyl sulfates) are known as alkylsulfuric acids. They are available only as the stable alkali metal salts. Cyclic sulfates **3** (alkylene sulfates) represent a special group of neutral sulfuric acid esters. They result from the reaction of one molecule of a diol with one molecule of sulfuric acid.

1

2

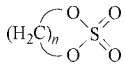
3

where R = alkyl

The most commercially important dialkyl sulfates, dimethyl sulfate [*77-78-1*] and diethyl sulfate [*64-67-5*], were first prepared in 1835 by Dumas and Peligot, who distilled a mixture of sulfuric acid and the corresponding alcohol. The general applicability of dimethyl sulfate as an alkylating agent was recognized by Ullmann and Wenner in 1900. Around this time, commercial production of dimethyl sulfate from sulfur trioxide and dimethyl ether began in Lyon, France.

The alkali metal salts of higher alkylsulfuric acids (fatty alcohol sulfates) are important surfactants and are not treated in this article (see → Laundry Detergents, → Surfactants).

2. Physical Properties

Up to dihexyl sulfate, the dialkyl sulfates are liquids at room temperature. They are completely miscible with polar organic solvents and aromatic hydrocarbons, but their solubility in water is poor. The first members of the homologous series of dialkyl sulfates are only partially miscible with saturated hydrocarbons. Boiling points are relatively high. Some physical properties of symmetrical dialkyl sulfates are listed in Table 1.

Dimethyl sulfate forms an azeotrope with water (73 wt % water, *bp* 98.6 °C).

Solubility of dimethyl ether in dimethyl sulfate at 0.1 MPa [9]:

t, °C	0	10	20	30	40	50
Solubility wt %	6.3	5.2	4.3	3.6	3.0	2.6

The lower cyclic sulfates have comparatively high melting points, which decrease with increasing molecular mass (see Table 2).

Alkylsulfuric acids, which are usually described as viscous oils, are unstable compounds.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_493

Table 1. Physical properties of dialkyl sulfates, $(RO)_2SO_2$

R	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C (160 mm)	ϱ (20 °C), g/cm^3	Ref.
Methyl	[*77-78-1*]	$C_2H_6O_4S$	126.11	−31	188.8 (160)	1.329	[1]
Ethyl	[*64-67-5*]	$C_4H_{10}O_4S$	154.18	−24.4	208 (160)	1.180	[2], [8]
Isopropyl	[*2973-10-6*]	$C_6H_{14}O_4S$	182.24	ca. −19	94 (7)	1.096	[3]
n-Propyl	[*598-05-0*]	$C_6H_{14}O_4S$	182.24		108 – 109 (11)	1.109	[4]
n-Butyl	[*625-22-9*]	$C_8H_{18}O_4S$	210.29		251.9 (160)	1.062	[2]
n-Pentyl	[*5867-98-1*]	$C_{10}H_{22}O_4S$	238.34	14	128 – 130 (3)	1.029	[5]
n-Hexyl	[*7722-57-8*]	$C_{12}H_{26}O_4S$	266.40	ca. 12	147 (3)	1.006	[6]
n-Heptyl	[*84833-74-9*]	$C_{14}H_{30}O_4S$	294.45		146.6 (1.5)	0.986	[7]

Viscosity of dimethyl sulfate [9]:

t, °C	20	30	40	50	60	70	80	90	100
η, mPa · s	1.18	11.54	1.33	1.16	1.02	0.90	0.81	0.73	0.66

Vapor pressure of dimethyl sulfate (figures in brackets are extrapolated):

t, °C	0	20	40	60	80	100	120	140	160
p, kPa	(0.01)	(0.06)	(0.25)	0.82	2.5	12.5	25.0		47.0

Table 2. Physical properties of cyclic sulfates

Sulfate	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C	ϱ, g/cm^3
Methylene sulfate *	[*56639-44-2*]	$(CH_2SO_4)_2$	220.17	155	–	
Ethylene sulfate	[*1072-53-3*]	$C_2H_4SO_4$	124.11	ca. 98	sublimes	1.735 (13 °C)
1,3-Propylene sulfate	[*1073-05-8*]	$C_3H_6SO_4$	138.14	ca. 60	ca. 140 (0.8 kPa)	1.513 (20 °C)
1,4-Butylene sulfate	[*5732-44-5*]	$C_4H_8SO_4$	152.17	ca. 37	–	
1,3-Butylene sulfate	[*4426-50-0*]	$C_4H_8SO_4$	152.17	ca. 44	ca. 150 (2 kPa)	1.490 (20 °C)

*

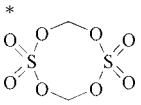

They are strong acids. Methyl hydrogensulfate [*75-93-4*], for example, is readily soluble in water; it is strongly dissociated in aqueous solution [10]. The preparation of neopentyl hydrogensulfate as a crystalline solid which melts on warming to room temperature has been described [11]. The alkali and alkaline earth metal salts of alkylsulfuric acids are stable, crystalline compounds, which often contain water of crystallization. The ammonium salts have well-defined melting points (ammonium methyl sulfate [*19803-43-1*], *mp* 135 °C; ammonium ethyl sulfate [*513-13-3*], *mp* 97 °C; ammonium *n*-propyl sulfate, *mp* 132 °C) and can be used to identify the alcohols.

3. Chemical Properties

Pure dialkyl compounds are not acidic. Traces of acid can be removed with cold hydrogen carbonate solution. Dimethyl sulfate may be stored for several days over dry anhydrous sodium carbonate, dry calcium oxide, or dry calcium carbonate, without undergoing any changes.

An equilibrium exists between monoalkyl sulfates and dialkyl sulfates:

$$2\,RHSO_4 \rightleftharpoons R_2SO_4 + H_2SO_4$$

This equilibrium can be shifted completely to the right by chemical binding of the sulfuric acid or by distilling the dialkyl sulfate.

Even at temperatures as low as 100 °C, unsymmetrical dialkyl sulfates are converted partially into the symmetrical esters:

$$2\,RR'SO_4 \rightleftharpoons R_2SO_4 + R'_2SO_4$$

Above 100 °C, acidic decomposition products such as $RHSO_4$ and SO_3 also form.

Dialkyl sulfates undergo acid- or base-catalyzed hydrolysis. In the acid-catalyzed reaction, free sulfuric acid forms rapidly. In alkaline solution, only the alkali metal salt of the alkylsulfuric acid is produced at room temperature. Higher temperatures are necessary for further hydrolysis. The alkyl groups are split off as alcohols, but small amounts of the corresponding ethers may also form.

Dimethyl sulfate is hydrolyzed slowly in cold water, but at 50 °C it takes 20 min to produce a 99.9 % methylsulfuric acid and methanol [12]. Further hydrolysis leads to sulfuric acid and methanol.

The most important property of dialkyl sulfates is their ability to transfer alkyl groups to other molecules (for examples, see Chap. 8). Reactivity decreases with increasing chain length. In unsymmetrical dialkyl sulfates, the shorter alkyl group reacts before the longer one. Above 160 °C, the lower alkyl sulfates react with metal oxides, halides, carbonates, and other inorganic compounds to form metal sulfates.

Under ultraviolet light and at high temperature, one methyl group of dimethyl sulfate can be chlorinated to yield the monochloro derivative, which is more reactive than dimethyl sulfate. In alkylation, only the chloromethyl group is transferred.

Alkylsulfuric acids are hydrolyzed in water or aqueous mineral acid to sulfuric acid and the corresponding alcohol. The alkali-metal salts serve as mild alkylating agents. Alkyl fluorosulfates, $ROSO_2F$, can be prepared readily from dialkyl sulfates and fluorosulfuric acid; they have greater alkylating potential than dialkyl sulfates [13].

4. Production

4.1. From Ethers and Sulfur Trioxide

The addition of sulfur trioxide to dimethyl ether is used industrially for the production of dimethyl sulfate.

$$(CH_3)_2O\,(g) + SO_3\,(l) \longrightarrow (CH_3)_2SO_4$$

$$\Delta H_{76\,^\circ C} = -92 \pm 7\,\text{kJ/mol}$$

The reaction is carried out in water-cooled, vertical aluminum or stainless steel tubes (diameter, 0.2 m; height, 4 m). In a continuous process at ca. 70 °C, stoichiometric quantities of liquid sulfur trioxide and gaseous dimethyl ether are fed from the top and bottom, respectively, into the dimethyl sulfate that is already present. The two reactants quickly form an acidic adduct, which slowly rearranges to the neutral ester. The product drawn off from the bottom of the tubes still contains a considerable amount of the acidic adduct. To achieve a quantitative rearrangement, the product is held at 70°C for several hours. A tower and its associated holding vessel have an output of up to 120 kg of dimethyl sulfate per hour [14], [15].

Technical-grade dimethyl sulfate contains small amounts of acidic compounds: water leads to formation of an equivalent amount of sulfuric acid; methanol introduced with the dimethyl ether results in methylsulfuric acid; and saturated hydrocarbons present in the dimethyl ether can be converted to sulfonic acids. Technical-grade dimethyl sulfate may also contain small amounts of dimethyl ether. Most of the impurities can be removed by vacuum distillation over anhydrous sodium sulfate.

This process is unsuitable for higher dialkyl sulfates because ethers with carbon atoms that are not bound to either oxygen or chlorine will undergo oxidation.

4.2. From Olefins and Sulfuric Acid

Sulfuric acid can add to the double bond of olefins:

$$2(CH_3)_2C{=}CH_2 + H_2SO_4 \longrightarrow [(CH_3)_3CO]_2SO_2$$

An excess of olefin leads almost exclusively to formation of the dialkyl sulfate. Further addition of sulfuric acid results in a decrease in the concentration of dialkyl sulfate and an increase in the concentration of alkylsulfuric acid [16].

Generally, tertiary olefins react much more readily than secondary or primary ones. According to Markovnikov's rule, the hydrogen of sulfuric acid always adds to the carbon atom with

the greater number of hydrogen atoms. Therefore, even olefins with a terminal double bond do not form *n*-alkyl sulfates.

The reaction of olefins in aqueous sulfuric acid leads not to a single product, but to a mixture of dialkyl sulfate, alkylsulfuric acid, sulfuric acid, water, alcohol, ether, and olefin. The mixture is characterized customarily by the concentration of acid used and the molar ratio of absorbed olefins. A higher olefin pressure and a higher sulfuric acid concentration favor formation of the corresponding dialkyl sulfate. Thus, the reaction between ethylene and 96 % sulfuric acid at standard pressure leads almost exclusively to ethyl hydrogensulfate, whereas anhydrous sulfuric acid at 0.35 MPa (3.5 bar) yields a mixture of 35 % diethyl sulfate and 51 % ethyl hydrogensulfate. At 1 MPa (10 bar), 62 % of diethyl sulfate and 30 % of ethyl hydrogensulfate are formed [17].

Sulfuric acid always causes some polymerization of olefins. Therefore, the sulfuric acid concentration and the reaction temperature should be kept as low as possible. Indeed, the reaction is most often carried out at or below ambient temperature.

The reaction of olefins with sulfuric acid is used industrially to extract isobutene [*115-11-7*] (2-methylpropene) from mixtures of C_4 hydrocarbons (65 % sulfuric acid) (→ Butenes); 80 % sulfuric acid is required to remove isoamylene [*563-45-1*] (3-methyl-1-butene) from mixtures of C_5 hydrocarbons [18].

4.3. From Alcohols and Chlorosulfuric Acid

The reaction between chlorosulfuric acid [*7790-94-5*] and an alcohol at low temperature yields the corresponding alkylsulfuric acid [19].

$$\mathrm{ROH + ClSO_3H \longrightarrow ROSO_3H + HCl \text{ where } R = alkyl}$$

The alkylsulfuric acid cannot be isolated by distillation because it is less volatile than other constituents and rearranges rapidly in the acidic environment. Crystalline methylsulfuric acid (*mp* − 30 to − 35 °C) is obtained if the reaction is carried out at − 40 °C and the hydrogen chloride generated is driven off with nitrogen [10].

Reaction between an alcohol and amidosulfuric acid leads directly to the ammonium salt of the corresponding alkylsulfuric acid.

Conversion of alkylsulfuric acids into dialkyl sulfates by vacuum distillation produces high yields only with methylsulfuric acid [20].

$$\mathrm{2\,RHSO_4 \rightleftharpoons R_2SO_4 + H_2SO_4}$$

The isolation of dialkyl sulfates from the equilibrium mixture progresses more smoothly if a neutral sodium salt is added, which binds the liberated sulfuric acid as sodium hydrogen sulfate [21]. The following procedure is used to prepare diethyl sulfate:

Ethylsulfuric acid (1100 kg, *d* 1.41 at 15 °C) is distilled under vacuum (0.65 kPa) in an enameled, stirred tank of 3-m^3 capacity over 1950 kg of anhydrous sodium sulfate. The condenser and recipient vessel are made of aluminum. The yield is 580 kg of diethyl sulfate (99.8 % purity), which is equivalent to 86 % of the theoretical yield.

The drawback of such processes is the production of relatively large amounts of sodium hydrogensulfate contaminated with residual dialkyl sulfate.

4.4. From Alcohols and Sulfuryl Chloride or Thionyl Chloride

If alcohols and sulfuryl chloride react in a molar ratio of 1 : 1, alkyl chlorosulfates are produced.

$$\mathrm{ROH + Cl{-}SO_2{-}Cl \longrightarrow RO{-}SO_2{-}Cl + HCl} \qquad (1)$$

Alcohols and thionyl chloride react readily to produce dialkyl sulfites [22].

$$\mathrm{2\,ROH + Cl{-}SO{-}Cl \longrightarrow RO{-}SO{-}OR + 2\,HCl} \qquad (2)$$

Dialkyl sulfites prepared according to Equation (2) react directly with the alkyl chlorosulfates to form dialkyl sulfates. Prior purification is unnecessary.

$$\mathrm{RO{-}SO_2{-}Cl + RO{-}SO{-}OR \longrightarrow RO{-}SO_2{-}OR + RO{-}SO{-}Cl} \qquad (3)$$

Reaction (3) can be carried out at 130 °C or, if zinc chloride is used as catalyst, at 90 °C; the reaction time is ca. 6 h [23]. Under these conditions, an exchange of alkyl groups takes place

(see Chap. 3); therefore, only symmetrical dialkyl sulfates can be produced.

Another process operates at 45 °C, with a reaction time of 2 h [24]. This enables a reaction between sulfites and chlorosulfates with different alkyl groups; no exchange of alkyl groups occurs. However, the alkyl chlorosulfates, especially those with long-chain alkyl groups, must be stabilized, e.g., by an alkyl chlorosulfite. Chlorosulfites are always present when the dialkyl sulfite rather than the alcohol is used in reaction (1) [25]:

$$\mathrm{RO{-}SO{-}OR + Cl{-}SO_2{-}Cl \longrightarrow RO{-}SO{-}Cl + RO{-}SO_2{-}Cl} \quad (4)$$

Despite an increased amount of valueless byproduct (RO−SO−Cl), this variant of Equation (1) is often preferred in the process operating at higher temperatures because of its smoother progress.

To produce a symmetrical dialkyl sulfate, only one sulfite is needed. If an unsymmetrical dialkyl sulfate is desired, however, the reaction of two different alcohols according to Equation (2) is necessary.

At the relatively high temperatures required in the older process, alkyl chlorosulfites decompose to sulfur dioxide and alkyl chloride. In the low-temperature process subsequent decomposition in aqueous alkali is necessary.

In the production of unsymmetrical dialkyl sulfates, yields are considerably higher if the longer-chain sulfite is used in Equation (4) and the sulfite with the smaller alkyl group in Equation (3). The overall reaction is as follows:

$$\mathrm{4\,ROH + 2\,SOCl_2 + SO_2Cl_2 \longrightarrow R_2SO_4 + 4\,HCl + 2\,RCl + 2\,SO_2} \quad (5)$$

This process is especially suitable for the production of higher, unsymmetrical dialkyl sulfates.

For further production methods, see [26], [27], [28].

4.5. From Alcohols and Sulfuric Acid

Reaction of alcohols with sulfuric acid can be used to produce alkylsulfuric acids.

$$\mathrm{ROH + H_2SO_4 \rightleftharpoons ROSO_3H + H_2O}$$

With excess sulfuric acid, the equilibrium can be shifted to the right. The alkylsulfuric acid may be isolated in the form of its alkali-metal salt, after neutralization in the cold and precipitation of the sulfate anion as barium sulfate. At higher temperature (120 – 140 °C), dialkyl ethers are produced. Use of oleum or sulfur trioxide leads to sulfonation. Thus ethionic acid [*461-42-7*], $\mathrm{HO_3S{-}CH_2CH_2{-}O{-}SO_3H}$, is produced readily from ethanol and sulfur trioxide [19].

5. Quality Specifications and Analysis

Commercial dimethyl sulfate (99.8 – 99.9 wt %) contains 0.1 – 0.2 % acid, calculated as sulfuric acid, and dissolved dimethyl ether in concentrations of a few parts per million. Other impurities are sulfuric acid and alkanesulfonic acids (see Section 4.1). Although the following specifications and analytical procedures pertain to dimethyl sulfate they are, in principle, also applicable to its nearest homologues.

Assay. (1) A sample is boiled for at least 30 min under reflux with dilute hydrochloric acid. The hydrochloric acid is then removed completely by evaporation with repeated addition of water. The sulfuric acid left behind is taken up in water and titrated [29] .

(2) A sample is stirred for 12 h at 20 °C in water to which an approximately equivalent amount of sodium hydroxide has been added. The excess hydroxide is then back-titrated. Because hydrolysis stops at the monomethyl sulfate, free sulfate ions can be determined with barium chloride in the cold solution after a small amount of hydrochloric acid has been added.

In both methods, the free acid must be determined separately and subtracted from the result.

Another method is based on the formation of methyl iodide from dimethyl sulfate and potassium iodide.

$$\mathrm{(CH_3)_2SO_4 + KI \longrightarrow CH_3SO_4^- + K^+ + CH_3I}$$

An aqueous methanol solution of a measured amount of potassium iodide is added to the sample with mild heating. Excess potassium iodide is titrated according to the Volhard method or potentiometrically with silver nitrate [29].

Determination of the Free Acid. The sample is suspended in ice – water or extracted by shaking with ice – water. The dissolved acid is then titrated with sodium hydroxide, by using methyl orange as indicator. Rapid assay at low temperature is essential to prevent significant formation of acids by hydrolysis.

The free acid can also be titrated with sodium acetate in glacial acetic acid, by using crystal violet as indicator. This accounts only for the first dissociation step of any sulfuric acid present [29].

The alkylsulfuric acid and sulfuric acid contents of dimethyl sulfate can be determined separately by potentiometric titration (glass electrode) with tributylethylammonium hydroxide [*16208-32-5*] in a water-free medium. This procedure is also suitable for diethyl sulfate [30].

Determination of Dimethyl Ether. The liquid sample is dripped by means of a syringe onto diatomaceous earth; a given amount of air is then passed over the absorbed liquid and collected in an evacuated bottle. The ether content of the air is determined in the usual manner by gas chromatography [29].

Determination of Dimethyl Sulfate Vapor in Air. Small amounts of dimethyl sulfate in air can be determined by using special test tubes in which concentrations can be read directly from color changes (Manufacturer: Drägerwerk AG, Lübeck, Germany). Reliable detection of levels as low as 0.005 ppm can be achieved.

In addition, test kits are available such as the MSA Universal Tester and the Dräger Multigas Detector and these can measure dimethyl sulfate down, to 0.005 ppm (Dräger) and 1.0 ppm (MSA) by volume.

One process involves the following steps (sample volume 500 L): (1) absorption from air onto silica gel; (2) elution with acetone; (3) reaction with nitrophenol by reaction of dimethyl sulfate or with sodium *p*-nitrophenoxide (DuPont Analytical Methods) [31]; (4) separation of the methylation product by thin-layer chromatography; (5) staining of the elution zone by reduction, diazotization, and coupling; and (6) comparison of the color intensity with a standard [32]. In one variant of this process, nitroanisole is determined by gas chromatography after step 3.

Continuous monitoring of rooms in which dimethyl sulfate is processed can be achieved with flame photometry using a sulfur detector (Messer Griesheim AG, Düsseldorf). If no other sulfur compounds are present (particularly SO_2 and CS_2), dimethyl sulfate levels from 0 to 0.5 ppm can be indicated. The last two methods are recognized by the chemical industry trade association in Germany [33].

6. Storage, Transportation, and Handling

Employees should be thoroughly informed and trained on the precautions that must be followed in the safe handling of dimethyl sulfate (DMS) and diethyl sulfate (DES). The area where DMS and DES are handled must be well ventilated and clearly marked to prevent entry by nonoperating personnel. The proper personal protective equipment must be worn to protect from any exposure.

It is important to refer to the manufacturers (for example see [31]) guidelines for the storage, transportation, and handling of DMS and DES. Since shipping regulations change over time, the user should refer to current applicable regulations such as those issued by the proper governmental authorities (for example, in the U.S. they are the DOT, EPA, OSHA, and The U.S. Food and Drug Administration) for additional information.

As a general guideline the physical and chemical properties of DMS and DES are shown in Table 3.

The presence of dimethyl sulfate can be detected by using the color reaction with 4-(4-nitrobenzyl)pyridine [*1083-48-3*] [35]. The sensitivity is about 20 mg/L. This method is also suitable for spot analysis.

First aid treatment after accidents with liquid or gaseous dimethyl sulfate involves thorough flushing of affected skin areas with 5 % ammonia solution. All clothing soiled with dimethyl sulfate should be removed immediately and decontaminated. A doctor must be consulted; symptoms may develop after a latency period of several hours (see Chap. 10).

Absorption by diatomite is the best way to clean up spilled dimethyl sulfate; 1 kg of diatomite binds 5 – 6 kg of dimethyl sulfate to form a doughlike mass.

A 5–10 % aqueous ammonia solution is commonly used to break down dimethyl sulfate. The resulting compounds are methylamine, dimethylamine, and methyl sulfate anion. Without mixing and at ambient temperature, the dimethyl sulfate tends to settle (ϱ 1.329 g/cm^3) and several days may be necessary for the reaction to be complete. However, with effective stirring, breakdown is complete in ca. 3 min. Methylamine and dimethylamine have very unpleasant, persistent odors. The odor nuisance can be avoided by using an approximately 2 M aqueous solution of hexamethylenetetramine [*100-97-0*] instead of ammonia. In this case, a nonvolatile methyl sulfate salt is formed [36]. If hexamethylenetetramine is dissolved in a mixture of two parts water and one part diethylene glycol monobutyl ether, a homogeneous phase is obtained with dimethyl sulfate, and stirring is unnecessary. Because the breakdown of dimethyl sulfate with ammonia or hexamethylenetetramine is always an exothermic reaction, the volume of breakdown agent should be five times that of dimethyl sulfate. Solid hexamethylenetetramine reacts explosively with dimethyl sulfate. Dry alkali such as soda or lime is unsuitable as a breakdown agent.

Table 3. Physical and chemical properties of dimethyl sulfate and diethyl sulfate

	Dimethyl sulfate	Diethyl sulfate
Solubility in water, g/100 g	2.8 (18 °C) (reacts)	0.7 (reacts)
Boiling point, °C	188 (decomp.)	208 (decomp.)
Density of vapor (air = 1)	4.35	5.31
Flash point, °C (Closed Cup)	83	104
Ignition temperature (DIN 51 794), °C	470	395
Temperature class (VDE 0165)	T 1	T 2
Explosive limits, vol %		
Lower	3.6	1.8
Upper	23.2	12.2
Heat of combustion at constant pressure [25 °C; CO_2, H_2O, H_2SO_4 (aqueous)], J/g	11 970	18 110
Heat of vaporization at 25 °C, J/g [10]	385	368
Heat of formation at 25 °C (liquid), kJ/mol	736	812

The higher homologues of dimethyl sulfate behave similarly, but longer reaction times are required.

If dimethyl sulfate gets into insulation, complete dismantling and destruction of the porous materials are necessary.

Large quantities of water and dimethyl sulfate must not be allowed to come into contact in an uncontrolled manner: the acid and heat produced during hydrolysis may result in a continuously increasing speed of reaction. In addition, the acid attacks iron containers, and a substantial buildup of hydrogen pressure may result.

Technical-grade dimethyl sulfate contains small amounts of dimethyl ether, a flammable gas, which can collect in the air space in containers during prolonged storage. Some dimethyl ether may also be generated during reaction in alkaline solution. In this case, it may be found in either the waste gas or the final product; a considerable amount of dimethyl ether can accumulate in the solvent.

Further information on safety precautions can be found in [31], [37–39].

7. Environmental Protection

Buildings in which large amounts of dimethyl sulfate are stored or handled must not have drainage that leads directly into the public sewer system; traps must be installed to test and, if necessary, monitor effluents. Disposal of waste liquid streams containing dialkyl sulfates must follow the regulations in force on the local, state, and national levels.

In most alkylations with dialkyl sulfates in aqueous alkaline media, the effluent contains half of the organic component of the dialkyl sulfate in the form of salts of alkylsulfuric acid. These salts are biodegradable. Alternatively, if permitted, they can be hydrolyzed completely under pressure and at higher temperature; the alcohol produced is recovered by distillation. Acid hydrolysis at low temperature is also possible, but subsequent neutralization is required.

Activated carbon may be used to absorb dimethyl sulfate vapors in waste gas. The carbon can be regenerated by treatment with dilute ammonia solution [40] or with water for several hours. Drying is not necessary.

8. Uses

8.1. Alkylation

The most important use of dialkyl sulfates is in alkylation. In this application, dialkyl sulfates must compete with alkyl halides, which are also important alkylating agents (→ Acylation and Alkylation). Halides, especially chlorides, are considerably cheaper; in addition, chlorides result in a lower salt load in the effluent. However, dimethyl and diethyl sulfate can be processed at atmospheric pressure because of their higher boiling points, and their high reactivity enables conversion to occur at lower temperatures. Dimethyl sulfate is by far the most important dialkyl sulfate because it is cheaper and more reactive than its homologues. Diethyl sulfate also has commercial significance, whereas higher homologues or unsymmetrical and cyclic dialkyl sulfates are used only in special applications.

Alkylation is usually carried out in stirred vessels. The alkylating agent is added slowly to an aqueous alkaline solution or suspension of the compound to be alkylated. Where large volumes are concerned, gradual addition of the alkali during the reaction is preferred, to maintain a certain pH. In the alkaline range, steel equipment may be used. Stainless steel or enamel is recommended for weakly acidic media or for sensitive products (e.g., quaternized dyes).

If the reaction does not proceed in an aqueous alkaline medium, treatment of the dry alkali-metal salt with dialkyl sulfate, in the presence of an inert solvent if necessary, is often successful. Occasionally, the pure substance is fed into dialkyl sulfate above 100 °C. However, increased side reactions (e.g., sulfonations) must be expected under these conditions.

Etherification. The classic use of dimethyl sulfate is to etherize an aromatic hydroxyl group in aqueous alkaline solution:

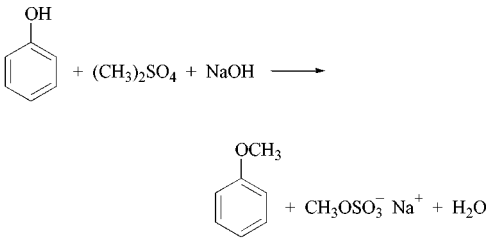

Dimethyl sulfate is added at 10 °C over 1 h, and the mixture is heated at 80 °C for 30 min; the yield is 90 % [41]. This reaction requires the presence of phenolate ions and therefore proceeds only at sufficiently high pH. The second methyl group of dimethyl sulfate may also be utilized if the mixture is refluxed for 15 h, but this is accompanied by reduced yield.

In strongly alkaline solution, the hydroxyl group reacts more rapidly than the carboxyl group:

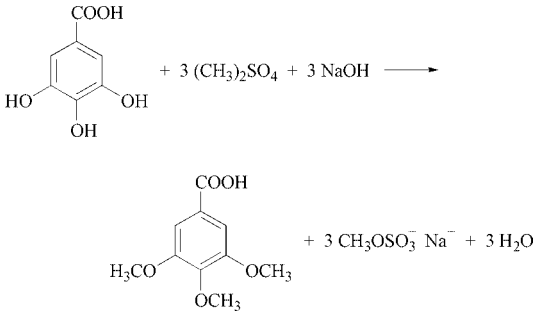

Small amounts of trimethoxybenzoic acid ester formed as byproduct are hydrolyzed by heating at 100 °C for 2 h.

Near neutral pH, however, the carboxyl group is alkylated more rapidly than the phenolic hydroxyl group. An example is the production of methyl salicylate [43]:

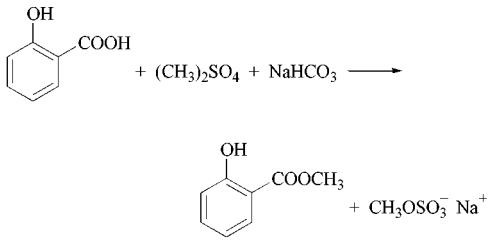

Because of steric hindrance, the two *ortho* hydroxyl groups of dimethyl dihydroxyterephthalate [*610-92-4*] can be etherized only under special conditions:

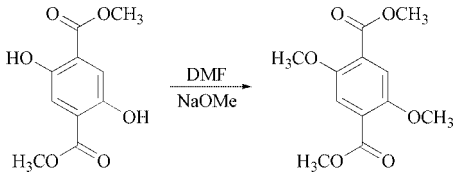

Sodium methoxide is used as the base, and the alkylation is carried out in dimethylformamide

(DMF) at 65 °C. The reaction is completed by heating rapidly to 100 °C. The product shows a strong bluish fluorescence and can be used as a brightener for polyester fibers [44].

Steric hindrance also affects alkylation of 2,4,6-trihydroxybenzoic acid [*83-30-7*]. Both hydroxyl groups in the *ortho* position remain free although esterification occurs [45]:

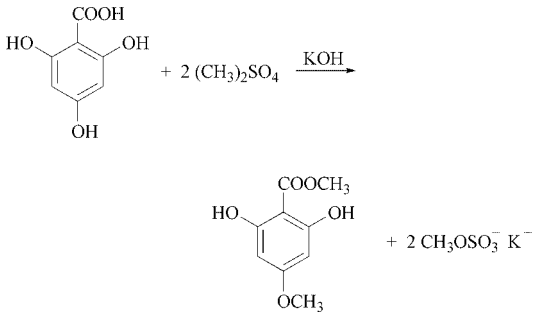

Selective methylation of one hydroxyl group is also possible. For example, hydroquinone monomethyl ether is obtained in 84 % yield if alkylation is carried out in a relatively large amount of an inert solvent (benzene), which extracts the monomethyl ether from the reaction mixture [46]:

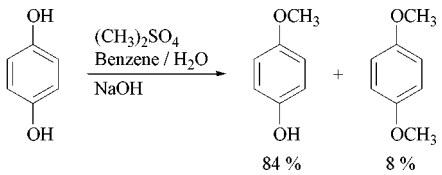

The aliphatic hydroxyl group can also be alkylated with dialkyl sulfates in the usual way if it is sufficiently activated by neighboring groups:

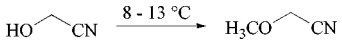

Dimethyl sulfate is partially recovered, and the low temperature reduces losses from hydrolysis [48]. In a strongly acidic medium, the non-activated aliphatic hydroxyl group is selectively methylated before the amino group, although the yield is low (30 %) [49]:

An unexpected byproduct of the methylation of phenol and other hydroxyaromatic compounds by dimethyl sulfate in the presence of tetrahydrofuran (THF) and potassium carbonate is the 4-methoxybutyl ether. In some cases, this is obtained in considerable quantities in addition to the normal methyl ether [50]:

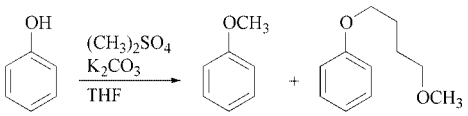

Esterification. Fatty acid ethyl esters, which can be used as plasticizers, are produced according to the following reaction [47]:

$$2\,R{-}COONa + (C_2H_5)_2SO_4 \longrightarrow 2\,R{-}COOC_2H_5 + Na_2SO_4$$

Both ethyl groups of the sulfate are utilized and therefore prolonged heating at 150 °C is necessary. The yield drops sharply outside the pH range of 7.5 – 8.5. Diethyl sulfate should be added at temperatures between 70 and 90 °C, never above 95 °C. If calcium salts are used instead of sodium salts, the second ethyl group does not react, because calcium bis(ethyl sulfate) [*926-03-4*] is unstable above 100 °C. Monomers of acrylic resins (raw materials for varnishes) and acrylic glass can also be produced by esterification with diethyl sulfate. The reaction is of commercial interest because diethyl sulfate is continually regenerated by feeding in ethylene. The process is also applicable to lower saturated fatty acids [47].

N-Alkylation. Ammonia and aliphatic or aromatic amines are alkylated readily by lower dialkyl sulfates. The reaction may progress to the tetraalkylammonium ion (quaternization). Quaternization leads to the formation of an organic cation; i.e., a positively charged, strongly hydrophilic center is created within the molecule. The corresponding anion is usually the alkyl sulfate ion (RSO_4^-) or the sulfate ion (SO_4^{2-}) produced during the reaction, although in some cases the cation can be associated with other suitable anions.

Quaternization can increase the water solubility of a drug or the affinity of a dye molecule for a negatively charged fiber. Numerous organic ammonium derivatives are used as raw materials for detergents, textile auxiliaries, and antistatic agents.

Most quaternizations can also be achieved with methyl chloride. However, methylsulfuric acid salts (methosulfates) often yield preparations with better consistency and greater solubility in water than the chlorides. The sodium methyl sulfate produced during methylation can often be tolerated in the product and does not have to be removed. For plastics additives that are put into the melt, the greater heat stability of methyl sulfates is a desirable property. In addition, many products must be free of chloride ions to prevent corrosion during subsequent processing.

Quaternization is feasible with aromatic, aliphatic, and heterocyclic compounds [51]. For example, pyridine reacts with dimethyl sulfate in the absence of a solvent to form methylpyridinium methyl sulfate [48]. Treatment of nitro-2-hydroxyquinolines with dimethyl sulfate also yields the quaternary base (**4**). If the nitro group is in the 8-position, i.e., adjacent to the nitrogen, steric hindrance inhibits *N*-alkylation, and 2-methoxyquinoline is the main product (**5**) [52].

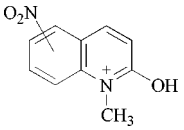

4

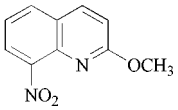

5

Alkylation of the dry potassium salt of carbazole in chlorobenzene at 100 °C gives *N*-ethylcarbazole [*86-28-2*] in 99 % yield; both alkyl groups react [53]:

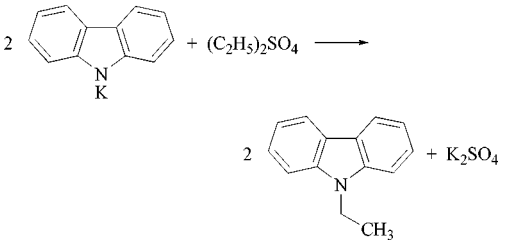

Also the alkylation of *p*-toluidine with two equivalents of dimethyl sulfate in the presence of sodium bicarbonate leads to *N,N*-dimethyl-*p*-toluidine [54].

Indirect methods must often be used to obtain the desired product. Thus, *N*-methylbutylamine [*110-68-9*] is produced from butylamine [*109-73-9*] in 49 % yield by the following reaction sequence [55]:

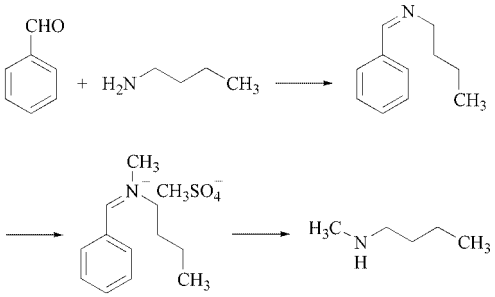

To prevent more than one methyl group from attaching to the nitrogen atom, the first step is formation of the Schiff base with benzaldehyde. After methylation, the Schiff base is then hydrolyzed. Reaction of urea with dialkyl sulfates usually gives the alkyl compound of the enol form, −C(OR)=NH. The cyclic amide caprolactam behaves similarly [56]:

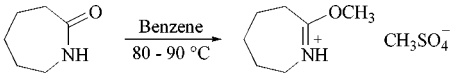

However, some urea derivatives react in the keto form, $-CO-NH_2$, which yields *N*-alkylated compounds. An example is the final step in the synthesis of a chlorinated *N*-phenylurea [57]:

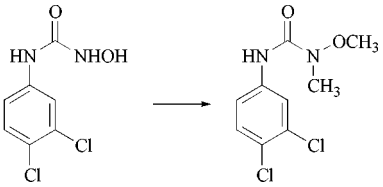

The hydroxyl group and the amino group are methylated in a single operation. The resulting *N*-dichlorophenyl-*N′*-methyl-*N′*-methoxyurea and compounds of similar structure are used in large quantities as herbicides.

Hydrazine is also methylated readily. An indirect pathway via a diacyl compound is pre-

ferred for the synthesis of symmetrical dimethylhydrazine [*540-73-8*] [58]:

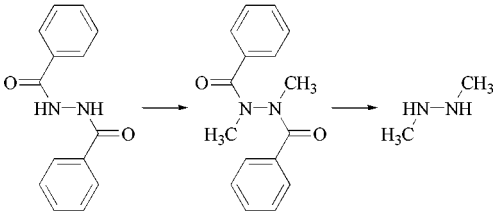

C-Alkylation. Compounds that can react as either ketones or enols (ambident nucleophiles, e.g., acetoacetic ester) may undergo *C*-alkylation (−CO−CHR−) in addition to *O*-alkylation (−C(OR)=CH−). Aromatic compounds with several hydroxyl groups in the meta position, e.g., resorcinol [*108-46-3*] and phloroglucinol [*6099-90-7*], react accordingly. In addition to ethers, small amounts of compounds in which the alkyl group is located at the aromatic nucleus are formed [59].

S-Alkylation. Phenols carrying a thiocyano group react with dimethyl sulfate in aqueous alkali to form thioethers.

The hydroxyl group remains unaltered and the cyano group is split off [60].

In addition, mercaptan salts are rapidly methylated with dimethyl sulfate. The thioethers can further react to form the water soluble sulfonium salts [61].

Alkylation of Inorganic Salts. Water-soluble metal halides can be converted to alkyl halides by reaction with dialkyl sulfates. A well-known example is the production of methyl iodide from potassium iodide [62]. Other salts can be alkylated similarly, e.g., sodium sulfide to diethyl sulfide or sodium sulfite to methanesulfonic acid [64] and sodium disulfide to dimethyl disulfide [63]. Di-*tert*-butyl peroxide, which is used as a polymerization catalyst, is produced continuously in a cooled, vertical reaction tube from hydroperoxide and a mixture of *tert*-butyl hydrogen sulfate and di-*tert*-butyl sulfate [65].

Sulfonation. At high temperature, reaction with dialkyl sulfates may lead to sulfonation. This results from the equilibrium between dialkyl sulfates, on the one hand, and sulfur trioxide and the corresponding ether, on the other (see Section 4.1).

$$R_2SO_4 \rightleftharpoons SO_3 + R_2O$$

For example, if diphenyl ether or diphenylmethylamine is heated with diethyl sulfate for several hours at 150 °C, phenylsulfonic acids are produced [66].

8.2. Other Uses

The lower dialkyl sulfates are used occasionally as a source of acidifiers in thermosetting resins, as stabilizers for various chemicals, and as selective solvents for aromatic compounds in petroleum fractions.

An interesting application in preparative chemistry is the protection of the hydroxyl group against oxidation by esterification with sulfuric acid. For example, *tert*-butyl sulfate is oxidized with nitrogen dioxide at about 15 °C, leading to the conversion of a methyl group into a carboxyl group. After hydrolysis, a sulfuric acid solution of α-hydroxyisobutanoic acid [*594-61-6*] is obtained [67].

Dimethyl sulfate can also act as a catalyst in certain reactions. For example, when it is mixed with titanium or vanadium halides it acts to catalyze the polymerization of olefins at low temperatures [68].

9. Economic Aspects

The production of dimethyl sulfate in Western Europe totals approximately 50 000 t/a [69]. There are no production figures reported for the United States. Diethyl sulfate production is estimated to be much lower. The higher dialkyl sulfates are not commercially available on a large scale, although small quantities are available for speciality synthesis.

The following companies produce dimethyl sulfate: Celanese and DMS Chemie Produktionsgesellschaft (Germany); Rhodia (France); Säurefabrik Schweizerhalle (Switzerland); DuPont, and Schweizerhall (United States); Lei Chang Yung Chemical Industry, Heng Chemical and Consul Chim (Taiwan); Hodogaya Chem., and Tsujimoto Kagaku Kogyo (Japan); and many producers in China, and

India [69]. Other countries which are listed as producing DMS are Mexico and the United Kingdom.

Diethyl sulfate is produced by Union Carbide (soon to be merged into Dow Chemical) in the United States. It is also produced by Nisso Maruzen (Japan); Dharamsi Morarji Chem and Industrial Solvents & Chemical (India).

The selling price of dimethyl sulfate in October 1999 was listed as $ 1.23/kg and for diethyl sulfate as $ 2.36/kg [70].

10. Toxicology and Occupational Health

10.1. Dimethyl Sulfate

Acute Toxicity. Inhalatory LC_{50} values in laboratory rodents (30 – 60 min) are between 30 and 75 ppm in the inhaled air [71], [72]. The 4-h LC_{50} is 45 mg/m^3 [73]. When humans are exposed to low concentrations, conjunctivitis with photophobia and irritation of the mucosal membranes of the upper respiratory tract result. Severe intoxication is characterized first by minor symptoms of local irritation of eyes, nose, throat, trachea, and bronchi; after a latency period of several hours lung edema may develop. If the lung edema is not lethal, bronchopneumonia and lung emphysema are typical secondary symptoms [74–85]. The oral LD_{50} values are between 205 and 404 mg/kg in laboratory animals [73]. In addition, dimethyl sulfate is corrosive to the eyes and skin [86].

Chronic Toxicity, Carcinogenicity, and Mutagenicity. Inhalation by mice, rats, and hamsters of 0.5 ppm over a period of 15 months caused severe inflammation in the mucosa of the nose and in the lungs as well as increased mortality [87]. Long-term inhalation of 0.5 ppm of dimethyl sulfate induced malignant tumors in the nose and the sinusoidal membranes in rats, mice, and hamsters [88], [89]. Compared to other compounds with alkylating properties, dimethyl sulfate has a rather weak methylating activity, probably because it is quickly hydrolyzed after contact with wet mucosal membranes. Therefore, it induces local tumors but no tumors remote of the site of the primary contact. Dimethyl sulfate is a directly acting mutagen [90–94]. It methylates DNA in rat liver, lung, and intestine [95]. For additional references, see [86], [96–98].

10.2. Diethyl Sulfate

In animal experiments, diethyl sulfate is a more potent alkylating agent than dimethyl sulfate.

Acute Toxicity. The acute toxicity of diethyl sulfate is lower than that of dimethyl sulfate: LC_{50} (rat, 4 h) > 250 ppm [99], LD_{50} (rat, oral) 350 – 1000 mg/kg [99]. Also the LD_{50} (rabbit, dermal) is reported to be 600 mg/kg [100]. The corrosive effect is weaker than that of dimethyl sulfate because hydrolysis occurs slower, but solutions containing 40 % burn the eyes of rabbits [106]. Lethal intoxications are characterized by lung edema occurring after a latency period of 10 – 20 h. Diethyl sulfate appears to be severely irritating or corrosive to the eyes and skin.

Chronic Toxicity and Carcinogenicity. After chronic oral administration of diethyl sulfate to rats (25 and 50 mg/kg in weekly intervals) benign tumors of the forestomach and one squamous cell carcinoma, but no systemic tumors, occurred. This probably results from rapid hydrolysis of diethyl sulfate in the stomach [101]. Subcutaneous administration also produced only local tumors at the site of injection [74], [107], [108]. In offsprings of dams who had been treated with diethyl sulfate during pregnancy, tumors of the central nervous system and multiple mammary cancer occurred [101]. Some additional references are worth consulting [102–105].

Mutagenicity. Diethyl sulfate is mutagenic in bacteriophages [109], in *E. coli* [110], [111], in *Drosophila* [112], [113], and in the dominant lethal assay in mice [114]. Cytogenetic studies in mice induced increased frequencies of aneuploidy in bone marrow but no structural chromosomal aberrations [115]. Diethyl sulfate is also positive in the point-mutation assay in Chinese hamster ovary cells [108].

11. References

1. DuPont Technical Bulletin: "Dimethyl Sulfate, Properties, Uses, Storage and Handling", no. H-32109 (3/92).
2. Am. Inst. Of Chemical Engineers (AIChE) Design Institute Physical Property Data Base (1982 – 1999).
3. A. Medwedew, *Chem. Ber.* **65** (1932) 135.
4. Levaillant, *Am. Chim., (Paris)* **11** (1936) no. 6, 534.
5. Levaillant, *C. R. Hebd. Seances Akad. Sci.* **200** (1935) 940.
6. Levaillant, *C. R. Hebd. Seances Akad. Sci.* **200** (1935) 941.
7. O. Barkerbus, *J. Am. Chem. Soc.* **56** (1934) 1206.
8. Union Carbide Material Safety Data sheet for Diethyl Sulfate, p. 7 of 15, July 23, 1998.
9. G. Seydel, *Chem.-Ing.-Tech.* **43** (1971) 140.
10. F. Klages, H. A. Jung, P. Hegenberg, *Chem. Ber.* **99** (1966) 1704.
11. S. R. Sandler, W. Karo: *Organic Functional Group Preparations,* 2nd ed., **vol. III,** Academic Press, New York 1989, pp. 135 – 136.
12. R. E. Robertson, S. E. Sugamon, *Can. J. Chem.* **44** (1966) 1728.
13. M. G. Ahmed, G. W. Alder, G. H. James, M. C. Sinnot, M. G. Whiting, *Chem. Commun.* 1968, 1533.
14. J. Avery et al., BIOS Final Report no. 986, 1946 – 1948, pp. 175, 227.
15. DE 4 121 409, 1993 (E. Eberhard et al.).
16. E. Clippinger, *Ind. Eng. Chem. Prod. Res. Develop.* **3** (1964) 3.
17. Société Anon. des Manufacturers de Glaces et Products Chimiques de Société Gobain, Chauny et Cirey, FR 1 006 211, 1952.
18. F. Asinger: *Die petrochemische Industrie,* Akademie-Verlag, Berlin 1971, Part 1, pp. 279 ff, 379 ff, 396 ff.
19. Union Chimique–Chemische Bedrijven S. A., GB 1 101 238, 1967 (A. v. Gysel, J. Colle).
20. *Houben-Weyl,* **6**/2, p. 481.
21. K. Venkataraman, *J. Sci. Ind. Res., Sect. A,* **8** (1949) 141.
22. *Houben-Weyl,* **6**/2, p. 437.
23. R. Levaillant, *C. R. Acad. Sci.* **189** (1929) 465 ; **200** (1935) 140.
24. VEB Arzneimittelwerk Dresden, DD 32 781 ; DE-AS 1 200 290, 1962 (W. Lugenheim, E. Carstens, H. Fürst).
25. *Houben-Weyl,* **6**/2, pp. 470.
26. G. A. Sokolskij, N. N. Bogdanov, V. M. Pavlov, *Zh. Obshch. Khim.* 1971, 2530; *Chem. Abstr.* **76** (1972) 112 598.
27. *Houben-Weyl,* **6**/2, pp. 475 ff.
28. *Kirk-Othmer,* 4th ed., **vol. 23,** 409 – 428.
29. Hoechst, Analytical Laboratory, personal communication.
30. W. M. Banick, E. C. Francis, *Talanta* **13** (1966) 979. D. K. Banerjee, M. J. Fuller, H. Y. Chen, *Anal. Chem.* **36** (1964) 2046.
31. DuPont Technical Bulletin no. H-32109 on "Dimethyl Sulfate, properties, Uses, Storage and handling", 1992, pp. 23 – 25.
32. J. Keller, *Fresenius Z. Anal. Chem.* **269** (1974) 206.
33. *Krebserzeugende Arbeitsstoffe, Anerkannte Analysenverfahren,* no. ZH 1 120.7, Carl Heymanns Verlag, Köln 1983.
34. H. Wendt, *Chem.-Tech.* **9** (1980) no. 1, 31 – 34.
35. J. Epstein, R. W. Rosenthal, R. J. Ess, *Anal. Chem.* **27** (1955) no. 9, 1435 – 1439.
36. F. L. Hahn, H. Walter, *Ber. Dtsch. Chem. Ges.* **54** (1921) 1532.
37. *Merkblatt für Arbeiten mit Dimethylsulfat,* Berufsgenossenschaft der Chemischen Industrie, no. M 013, Verlag Chemie, Weinheim 1982.
38. *Unfallverhütungsvorschrift der Berufsgenossenschaft der Chemischen Industrie,* **Appendix 4, no. 34,** Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg 1982.
39. *Jahrbuch "Betriebswacht",* Chap. MAK-Werte ff, Universum-Verlagsanstalt, Wiesbaden.
40. Hoechst DE-OS 2 063 071, 1970 (R. Grimm, W. Herzog, R. Lademann).
41. *Org. Synth. Coll. Vol. I,* p. 58.
42. *Org. Synth. Coll. Vol. I,* p. 537.
43. M. Pailler, P. Bergthaller, *Monatsh. Chem.* **99** (1968) 103. U. N. Hirwe et al., *J. Univ. Bombay, Sect. A,* **22,** part 5; *Chem. Abstr.* **49** (1953) 11594 h.
44. Alpine Chem. AG, FR 1 533 584 1967 (J. Hrach, W. Zeschmar).
45. Kabushi Kaisha Ricoh, DE-OS 1 917 055, 1969.
46. Universal Oil Products, US 3 274 260, 1964 (J. Levy, A. Friedmann).
47. B. K. Zeinalov, J. A. Panteeva, *Vses. Soveshch. Sin. Zhirosdmen, Poverkhùoslnoaktiv. Vesheheslvam Moyushch. Sredslvam, 3rd Shebekino* 1965, pp. 225 – 229 ; *Chem. Abstr.* **66** (1967) 30 233 h; **67** (1967) 11 140.
48. *Org. Synth. Coll. Vol. II,* p. 387.

49. C. Pfizer Co. Inc ., GB 1 070 307, 1966.
50. Hori Zenichi et al., *Chem. Pharm. Bull.* **20** (1972) 827.
51. H. Quast, E. Schmitt, *Chem. Ber.* **101** (1968) 4012.
52. H. Decker, *Ber. Dtsch. Chem. Ges.* **38** (1905) 1148.
53. Hoechst, DE-OS 2 132 961, 1971 (Th. Papenfuhs).
54. S. Hunig, *Chem. Ber.* **85** (1956) 1056.
55. *Org. Synth. Coll. Vol. V,* p. 736.
56. *Org. Synth. Coll. Vol. IV,* p. 588.
57. O. Scherer, G. Hörlein, K. Härtel, *Angew. Chem.* **75** (1963) 852.
58. *Org. Synth. Coll. Vol II,* p. 208.
59. E. Ott, E. Nauen, *Ber. Dtsch. Chem. Ges.* **55** (1922) 928.
60. Dow Chemical, US 3 655 773, 1970 (W. Reifenschneider).
61. Gulf Research & Development Co., US 3 661 771, 1972 (R. H. Havens).
62. *Org. Synth. Coll. Vol. II,* p. 404.
63. CN 1 075 140 A, 1993 (Y. Ren, C. Zhu, B. Jun); *Chem. Abstr.* **121** (1993) 136593.
64. O. I. Kajurin, SU 335 240, 1969 ; *Chem. Abstr.* **77** (1972)33 931.
65. G. Nettesheim, *Erdöl Kohle Erdgas Petrochem.* **19** (1966) 265.
66. V. N. Belov, M. Z. Finkelstein, *Khim. (USSR)* **16** (1946) 1248 ; *Chem Abstr.* **41** (1947) 3065 i.
67. Nissan Chem. Ind. Co., JP 6 726 611, 1962 ; 6 726 612, 1962.
68. US 3 389 128, 1968 (J. W. Bayer, W. C. Grinonneau).
69. "SRI 1998/1999 Directory of Chemical Producers," covering Europe and the United States.
70. *Chemical Marketing Reporter* **256** (1999) no. 14, 28.
71. N. Hein, *Med. Inaug.-Diss.,* Würzburg 1969.
72. L. Ghiringhelli, U. Colombo, A. Monteverde, *Med. Lav.* **48** (1957) 634.
73. World Health Organization (WHO): "Dimethyl Sulfate", Environmental Health Criteria no. 48, 1985.
74. S. Weber, *Naunyn-Schmiedeberg's Arch. Exp. Path. Pharmakol.* **47** (1902) 113.
75. F. D. Mohlau, *J. Ind. Hyg.* **2** (1920) 238.
76. H. Strothmann, *Klin. Wochenschr.* **8** (1929) 493.
77. J. Balazs in H. Fühner, *Sammlung von Vergiftungsfällen* **5** (1934) 47.
78. S. v. Nida, *Klin. Wochenschr.* **24/25** (1947) 633.
79. S. Tara, A. Cavigneaux, Y. Delplace, *Arch. Mal. Prof.* **15** (1954) 291
80. T. R. Littler, R. B. McConnel, *Br. J. Ind. Med.* **12** (1955) 54.
81. M. G. Duverneul, *Arch. Mal. Prof.* **18** (1957) 306.
82. W. Nebelung, *Arch. Gewerbepath. Gewerbehyg.* **15** (1957) 581.
83. W. Tilling, B. Knick, *Med. Klin.* **35** (1960) 1534.
84. H. Leithoff, J. Weinreich, *Beitr. gerichtl. Med.* **22** (1963) 196.
85. L. Roche, J. M. Robert, P. Paliard, *Can. J. Chem.* **44** (1966) 1728.
86. *Documentation of the Threshold Limit Values and Biological Exposure Indices,* 6th ed., American Conference of Governmental Industrial Hygienists (ACGIH) 1991, p. 497.
87. F. A. Schlögel, *Med. Inaug.-Diss.,* Würzburg 1972.
88. H. Druckrey, H. Kruse, R. Preussmann, S. Ivankovic, Ch. Landschütz, *Z. Krebsforsch.* **74** (1970) 241.
89. F. A. Schlögel, P. Bannasch, *Naunyn-Schmiedeberg's Arch. Exp. Path. Pharmakol.* **266** (1970) 441.
90. J. R. Connell, A. S. C. Medcalf, *Carcinogenesis* **3** (1982) 385.
91. M. Fox, J. Brennand, *Carcinogenesis* **1** (1980) 795.
92. K. Hemminki, K. Falck, H. Vainio, *Arch. Toxicol.* **46** (1980) 277.
93. H. Kubinski, G. E. Gutzke, Z. O. Kubinski, *Mutat. Res.* **89** (1981) 95.
94. R. B. Painter, R. Howard, *Mutat. Res.* **92** (1982) 427.
95. P. F. Swann, P. N. Magee, *Biochem. J.* **110** (1968) 39.
96. *Re-evaluation of some organic chemicals, hydrazine and hydrogen peroxide,* **vol. 71,** IARC 1999, p. 575.
97. *Overall evaluations of carcinogenicity: An updating of IARC Monographs,* **vol. 1 to 42,** Suppl. 7, IARC 1987, p. 200.
98. *Some aromatic amines, hydrazine and related substances, N-nitroso compounds and miscellaneous alkylating agents,* **vol. 4,** IARC 1974, p. 271.
99. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, *J. Ind. Hyg.* **31** (1949) 60.
100. *Occupational exposures to mists and vapors from strong inorganic acids; and other industrial chemicals,* **vol. 54,** International Agency for Research on Cancer (IARC) 1992, p. 213.

101. H. Druckrey, H. Kruse, R. Preussmann, S. Ivankovic, Ch. Landschütz, *Z. Krebsforsch.* **74** (1970) 241.
102. *Re-evaluation of some organic chemicals, hydrazine and hydrogen peroxide,* **vol. 71,** IARC 1999, p. 1405.
103. *Occupational exposures to mists and vapors from strong inorganic acids; and other industrial chemicals,* **vol. 54,** IARC 1992, p. 213.
104. *Overall evaluations of carcinogenicity: An updating of IARC Monographs,* **vol. 1 to 42,** Suppl. 7, IARC 1987, p. 198.
105. *Some aromatic amines, hydrazine and related substances, N-nitroso compounds and miscellaneous alkylating agents,* **vol. 4,** IARC 1974, p. 277.
106. C. P. Carpenter, H. F. Smyth, Jr., *Amer. J. Ophthal.* **29** (1946) 1363.
107. W. B. Deichmann, H. W. Gerarde: *Toxicology of Drugs and Chemicals,* Academic Press, New York 1969, p. 217.
108. D. B. Couch, N. L. Forbes, A. W. Hsie, *Mutat. Res.* **57** (1978) 217.
109. A. Loveless, *Proc. R. Soc. B* **150** (1959) 497.
110. Yu. A. Ratiner, S. K. Kanareikina, V. M. Bondarenko, I. V. Golubeva, *Zh. Mikrobiol. (Moscow)* **5** (1976) 117.
111. C. H. Clarke, D. M. Shankel, *Mutat. Res.* **53** (1978) 168.
112. M. Pelecanos, *Nature (London)* **210** (1966) 1294.
113. M. Pelecanos, *Experientia (Basel)* **27** (1971) 473.
114. A. M. Malashenko, I. K. Egorov, *Sov. Genet.* **4** (1972) no. 1, 14.
115. N. I. Surkova, A. M. Malashenko, *Sov. Genet.* **8** (1974) no. 11, 1386.

Diallyl Phthalates → Polyesters
Diamines → Amines, Aliphatic
Diaminotoluenes → Phenylenediamines and Toluenediamines
Diamond → Carbon
Diaphragm Cell → Chlorine
Diatomite → Silica

Diazo Compounds and Diazo Reactions

Klaus Forstinger, Clariant GmbH, Frankfurt/Main, Federal Republic of Germany

Hans Joachim Metz, Clariant GmbH, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Aliphatic Diazo Compounds 1
3. Aromatic Diazo Compounds 3
3.1. Diazonium Salts 3
3.1.1. Properties 3
3.1.2. Diazotization 5
3.1.3. Other Methods of Preparation 7
3.1.4. Uses 8
3.2. Diazotates 8
3.3. Diazosulfonates 8
3.4. Triazenes 9
3.5. Other Aromatic Diazo Compounds 11
3.6. Analysis 12
4. Diazo Reactions 12
4.1. Chemical Transformation with Retention of the Diazo Group 12
4.2. Reactions at the Diazo Group 13
4.3. Dediazotization (Replacement of the Diazo Group) 15
5. Environmental Protection 18
6. Safety 20
7. Toxicology 20
8. References 20

1. Introduction

A diazo compound is an aliphatic, aromatic, or heterocyclic compound in which an $-N_2$ group is attached to a carbon atom. This group can be present as such (diazo compounds in the narrowest sense), it can carry a positive charge (diazonium salts), or the nitrogen atom in the β-position can carry a substituent (diazo compounds in the general sense). Compounds containing the $RN_2^+X^-$ group are named by adding the suffix "diazonium" to the name of the parent compound RX, followed by the name of X^- (e.g., benzenediazonium chloride), whereas compounds having the N_2 group in the neutral form attached to a carbon atom are named by adding the prefix "diazo" to the name of the parent compound (e.g., diazomethane). In azo compounds, the N_2 group is substituted on both sides by carbon atoms. A overview on nomenclature is given in [6].

Diazo- and diazonium compounds are reactive molecules. They readily take part in a variety of reactions, which gives them their value to the chemical industry. The reactions, which normally take place rapidly and smoothly, lead to a number of important intermediates and end products. Their main importance, however, is as a first step in the preparation of azo pigments and azo dyes (→ Azo Dyes)

2. Aliphatic Diazo Compounds

Aliphatic diazo compounds have an open chain structure, which may be represented by the following mesomeric formulas:

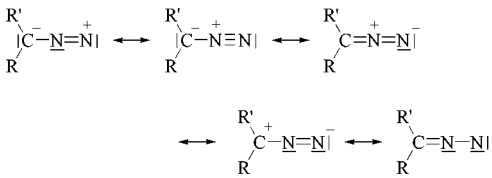

With few exceptions, aliphatic diazo compounds lack the ability for azo coupling, and as a result they are of less industrial importance. While handling diazomethane or other diazo compounds appropriate care has to be applied. Nevertheless, due to the great synthetic potential in organic chemistry, quite a number of diazo compounds and precursors are commercially available [7]. The simplest diazo compound is diazomethane [*334-88-3*], CH_2N_2, *bp* −23 °C, *fp* −145 °C, a yellow gas, which is explosive both as gas and in concentrated solution. It is prepared by reacting an alkali metal hydroxide with *N*-nitroso-*N*-methylurea [*684-93-5*] or *N*-nitroso-*N*-methyl-4-toluenesulfonamide [*80-11-5*], e.g., in diethyl ether. *N*-Methyl-*N*-nitro-*N*-nitrosoguanidine is also used in the synthesis of diazomethane (for

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_505

details, see [7] and references cited therein). Solutions of diazomethane decompose slowly, even at room temperature, liberating nitrogen.

Diazomethane has been used in the esterification of unstable carboxylic acids and generally as a methylating agent. Because of its toxicity (see Chap. 7) and explosive properties, however, its industrial use is rather limited.

Many synthetic methods for the preparation of diazo compounds and diazocarbonyl compounds are available [6–9]. A covenient route is the diazo transfer reaction [8], in which the N_2 group is transferred from an N_2 donor (a sulfonyl azide) to a suitable acceptor molecule. For example, tosyl azide TsN_3 reacts with *tert*-butyl acetoacetate to give the corresponding*tert*-butyl α-diazoacetoacetate.

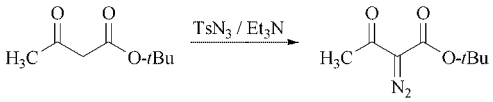

In the Arndt–Eistert synthesis, diazomethane is used to convert an acid chloride into a carboxylic acid with an additional carbon atom ($RCOCl \rightarrow RCH_2COOH$). The reaction proceeds via Wolff rearangement of the initially formed diazoketone on treatment with water and silver ion catalysis or photolysis. With alcohols $R'OH$ or ammonia instead of water the esters RCH_2COOR' or the amide are obtained, respectively [6].

Ring contraction occurs during Wolff rearrangement of cyclic 2-diazoketones, whereas ring expansion of a cyclic ketone to the next higher homologue can be achieved by reaction with diazomethane (e.g., cyclohexanone to cycloheptanone). A prominent example for industrial application of the Wolff rearrangement is the diazonaphthoquinone(DNQ)/novolak system used in photoresists [14–17]. In the early 1940s, Süs observed inhibition of novolak dissolution by a diazonaphthoquinone sulfonate. However, on exposure of these coatings to sunlight, they dissolved faster than the pure novolak. Süs realized the potential of his observation and the first Kalle positive-tone printing plate consisting of a DNQ/novolak system on aluminum was introduced in 1950 (Ozatec). The novolak/DNQ systems have been optimized to maximize this effect. In modern resists, dissolution rate between exposed and unexposed resist regions differ by over three orders of magnitude. DNQ loading is ca. 5 – 10 %. Diazonaphthoquinone (DNQ) is photolytically converted into a 3-carboxyindene, which unlike the diazonaphthoquinone is more soluble in aqueous bases.

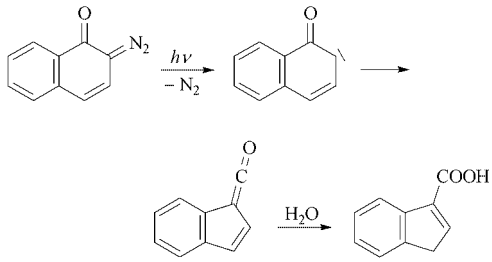

Substituted DNQs used in photoresists include 1,2-diazonaphthoquinone-4- or 5-sulfonate esters [16], [17] and 7-hydroxy-1,2-naphthoquinone-2-diazide-4-sulfonic acid. Other aliphatic diazo reagents used in organic synthesis are diphenyldiazomethane [*883-40-9*] and ethyl diazoacetate [*623-73-4*].

Alkyldiazoacetates react with a hydrazone aldehyde in the presence of Lewis acids to form hydrazone β-ketoesters, which are precursors for pyridazinones [18]. Certain pyridazinones are plant-growth regulators.

A great number of diazo- and diazo carbonyl compounds are accessible. Synthesis and versatility in organic transformations are covered by comprehensive reviews [6–9].

3. Aromatic Diazo Compounds

The most important aromatic diazonium compounds are the diazonium salts (Section 4.1). Other diazo compounds are:

- Aryldiazotates
- Aryldiazoethers
- Aryldiazoesters
- Aryldiazosulfides
- Aryldiazosulfones
- Aryldiazosulfonates
- Aryldiazophosphorous esters
- Aryldiazoamino compounds

Since the above-mentioned compounds show typical coupling reactions they can be considered as masked diazonium compounds.

3.1. Diazonium Salts

The diazonium cation is a mesomeric system that can be represented by the following mesomeric structures:

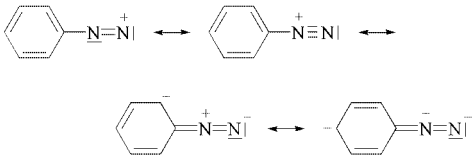

Conjugation with the π electrons of the aromatic system renders aryldiazonium ions much more stable than alkyldiazonium ions. Aromatic diazonium salts can be isolated as moderately stable compounds.

3.1.1. Properties

Diazonium salts are known with many inorganic and organic anions. Salts of strong acids are completely dissociated in aqueous solution, and such solutions are neutral, because aryldiazonium ions show little tendency to hydrolyze. A large number of salts with complex anions are known [1], [2], [19]. Salts of industrial relevance contain anions such as chloride, hydrogensulfate, tetrafluoroborate, tetrachlorozincate, and naphthalene-1,5-disulfonate, the solubilities of which decrease in that order. Diazonium compounds formed from aromatic sulfonic or carboxylic acids, being zwitterions, are sparingly soluble unless additional solubilizing groups are present.

Many diazonium salts are explosive when dry, some even when moist, an extreme example being diazotized aminotetrazole, which can explode even in aqueous suspension. Explosion can be initiated by heat, friction, or impact. Extremely unstable salts occasionally explode without any apparent external cause. Nitrates, perchlorates, and iodides must be handled with great care. The risk of explosion decreases in the order iodide, bromide, and chloride; and in the order tetrachlorozincate, hydrogensulfate, tetrafluoroborate, and naphthalene-1,5-disulfonate; the latter group of anions is preferred for safety reasons. Nitro groups on the aromatic nucleus increase the risk of explosion, while alkylamino and especially arylamino groups reduce it. The pressure increase in the exothermic decomposition was investigated [20], and the explosion properties were studied calorimetrically [21].

Gradual decomposition of diazonium salts occurs at or just above room temperature, and nitrogen is cleaved from the molecular ion [22]. For many industrial diazonium salts, the extent of decomposition is less than 1 % after one month at 40 °C.

Diazonium salt solutions are also not stable indefinitely, especially at elevated temperatures. Decomposition in dilute mineral acid is a first-order reaction. The rate of decomposition, which is largely independent of the anion and almost entirely determined by the cation, increases about fourfold for a 10 °C rise in temperature [23]. Solutions of diazonium naphthalenesulfonates are exceptional because of π-complex formation, which stabilizes the cations [4]. Figure 1 shows the rate of decomposition of some diazonium chloride solutions at various temperatures.

For mineral acid concentrations between 0.1 and 10 wt %, the rate of decomposition is practically constant. Solutions in more highly concentrated sulfuric acid are usually more stable: the rate of decomposition of 2,4-dinitrophenyldiazonium ion in 85 wt % sulfuric acid is ca. 60 times less than in 63 wt % sulfuric acid. The system H_2O/H_2SO_4 is very complex [24] and the effect can be understood by the absence of reactive particles and the decreasing activity of water and sulfate ions. Many diazonium salts of heterocyclic amines, e.g., of 2-aminobenzothiazole, are stable only in concentrated sulfuric acid; dilution with water leads to rapid decomposition. Increasing pH also accelerates decomposition. Although a pH > 3 may be suitable for azo coupling, the rate of decomposition can exceed that in mineral acids by several orders of magnitude. The outcome of any such experiment is controlled kinetically, for instance, by very rapid coupling. The maximum rate of decomposition occurs at pH values at which concentrations of diazonium ion and diazotate ion are equal [25].

3.1.2. Diazotization

Although continuous diazotization processes are known [26–31], crystalline diazonium salts are

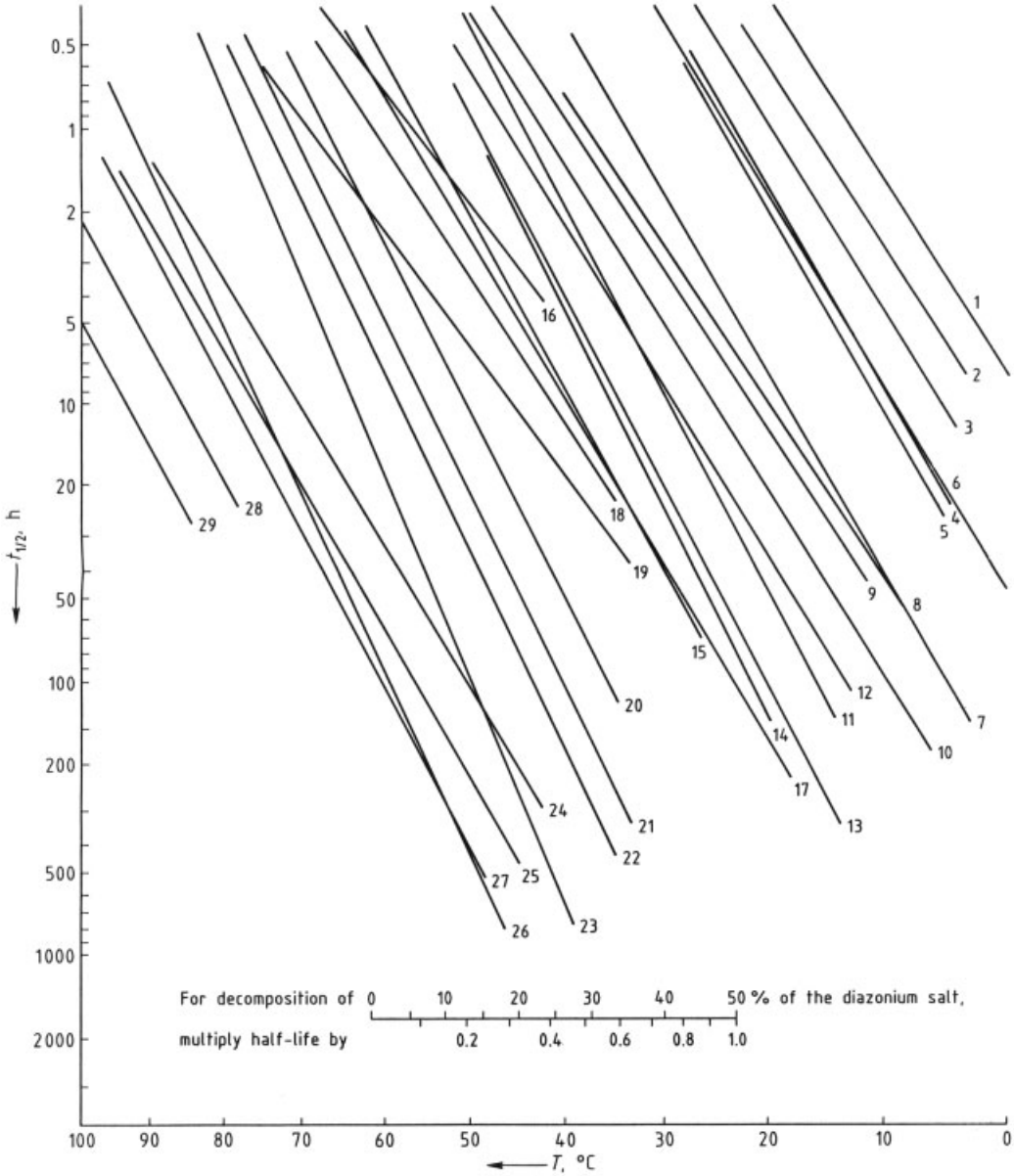

Figure 1. Decomposition half-life of mineral acid solutions of diazonium salts as functions of temperature (partly extrapolated) Diazonium chloride solutions (ca. 0.1 mol/L, pH ca. 1.5) prepared from: 1) 5-Amino-1,4-dimethylbenzene; 2) 3-Aminophenol; 3) 3-Ethylaniline; 4) 2-Aminotoluene; 5) 3-Aminotoluene; 6) 3-Aminoanisole; 7) Aniline; 8) 4-Amino-1,3-dimethylbenzene; 9) 4-Chloro-2-aminotoluene; 10) 2-Aminonaphthalene; 11) 3-Aminobenzenesulfonic acid; 12) 6-Chloro-2-aminotoluene; 13) 4-Aminotoluene; 14) 2-Aminobenzenesulfonic acid; 15) 4-Aminobenzenesulfonic acid; 16) 4-Aminoazobenzene; 17) 3-Chloroaniline; 18) 5-Nitro-2-aminotoluene; 19) 4-Nitro-2-aminotoluene; 20) 2-Aminophenol; 21) 4-Nitroaniline; 22) 4-Chloroaniline; 23) 3-Nitroaniline; 24) 4-Aminophenol; 25) 2-Nitroaniline; 26) 4-Aminoanisole; 27) 2-Chloroaniline; 28) 1,4-Diaminobenzene (only one amine group diazotized); 29) 2,4-Dichloroaniline

still made by the batch method without exception. Apart from the well-known problems associated with continuous operation of a multiphase process, the low production volumes could well be the reason for batch operation. For the details of specific industrial diazotization procedures, see [32–34]. Whenever suitable, *direct diazotization* in dilute hydrochloric acid with sodium nitrite is the process of choice. The amine should either be completely dissolved in the liquid phase or prepared as fine, lump-free suspension of the free amine or the amine hydrochloride. Amines that are liquid at room temperature or that melt below ca. 70 °C are slowly added as liquids with vigorous agitation to dilute hydrochloric acid at temperatures about equal to or higher than the melting point of the amine. For each mole of amine, 2.3 – 3.0 mol of hydrochloric acid and 180 – 500 mL of water are required, depending on the molecular mass of the amine. To prevent lump formation, only the amount of hydrochloric acid necessary to form the hydrochloride (ca. 1.2 mol) is used; the remainder required for diazotization (1.1 – 1.8 mol) is added at a later stage.

Amines melting above 80 °C are slurried in water at ambient temperature to prepare a fine suspension before mineral acid is added. Hard agglomerates are broken down by stirring or with high-speed rotor – stator homogenizers or auxiliary disk mills. The hydrochloric acid is then added slowly to form the hydrochloride.

Diazotization is usually carried out between 0 and 15 °C with external cooling to avoid side reactions and decomposition. When the starting temperature has been established, a concentrated solution of sodium nitrite is added as quickly as possible without the evolution of nitrous gases (NO_x). The rate of nitrite addition is usually adjusted to match the rate of consumption, monitored by spot testing with 3,6-diamino-9*H*-thioxanthene-10,10-dioxide [*10215-25-5*] ("sulfone reagent") or redox measurement. The heat of reaction, ca. 100 kJ/mol, must be removed continuously, which frequently imposes a limit on the speed of the process.

Nitrite depletion of the reaction mixture must be avoided to suppress side products. Certain amines require quick addition of all the nitrite, in practice within 1 – 2 min, 4-nitroaniline being a well-known example. In such cases, the heat of reaction cannot be removed fast enough by external cooling; hence, either the starting temperature must be very low (−10 to −15 °C) and/or ice must be added for internal cooling. In the diazotization of aminoazo compounds or aminonaphtholsulfonic acids and in some continuous diazotization processes, higher reaction temperatures up to 30 – 50 °C are used.

Amines that form sparingly soluble diazonium chlorides must be diazotized in a more diluted state; otherwise a clear solution is not obtained. For many diazoazobenzenes, concentrations as low as 0.03 mol/L are used, compared with ca. 1 mol/L for diazonium compounds of chlorobenzenes or chlorotoluenes.

Amines with a tendency to form triazenes require more acid, (e.g., 2,5-dichloroaniline requires 4 – 5 mol HCl), whereas the easily oxidizable aminonaphthols are diazotized satisfactorily only with the stoichiometric amount of acid.

Aromatic aminosulfonic acids are usually only slightly soluble in water (zwitterions), as are their diazonium salts. Sulfanilic acid may be diazotized by a standard procedure. For larger molecules, *indirect diazotization* is a better approach: the aminosulfonic acid is dissolved by the addition of an equivalent amount of sodium hydroxide solution to give the sodium salt. Sodium nitrite solution is then added, and the mixture poured into dilute ice-cooled hydrochloric acid to effect diazotation.

Diazotation of *weakly basic aromatic amines* (dinitroanilines, dinitrocyanoanilines, and heterocycles such as aminonitrothiazoles), which has gained increasing industrial importance due to the development of blue to red disperse dyes based on azo chemistry, calls for special techniques. The poor nucleophilicity of these amines requires strong electrophiles such as the nitrosonium cation, which is formed quantitatively from nitrous acid in highly acidic media like sulfuric or perchloric acid above 65 wt % acid. The highly electrophilic diazonium ions formed are stable only at very high acidities of the solvent; typically > 85 wt % of sulfuric acid is required to avoid decomposition. Reaction kinetics can be complex, and unusual dependences of rate constants on overall concentration of amine have been observed [10, B1 – B52]. Quenching the reaction mixture with ice in many cases produces a diazonium salt solution which can be further

processed in the same way as a product made by the hydrochloric acid – nitrite route.

To remove undiazotized material and byproducts *clarification* of diazonium salt solutions may be necessary. The use of a filtration aid such as diatomite is recommended; activated carbon is used to obtain the best results. The solution then usually goes directly into the azo coupling reactor. Diazonium salt is isolated only for special products, such as pure materials for reprography or the textile industry, where inert salts in the product are not tolerated.

Crystalline diazonium salts can be precipitated from solution by the addition of a solution of zinc chloride, arylsulfonate, sodium chloride, etc. Solid precipitants should be introduced gradually. In a few cases, the diazonium salt solution is added to the precipitant. Temperature, time, and stirring intensity are chosen to give crystals as large and compact as possible, to facilitate filtration. The adhering mother liquor can be removed largely by a basket centrifuge, giving residual moisture contents of 5 – 30 %, depending on the product. Some materials not suitable for centrifugation, e.g., those with platelet or needlelike crystals, may be dewatered in a hydraulic press. Continuous centrifuges enable filtration and dewatering to be done in a single operation, but care must be taken to avoid heating the diazonium salt excessively, with the consequent danger of fire when the diazonium salt is removed by the scraper blade. The yield depends on the choice of amine and the method of isolation. Tetrachlorozincates give the best results, with yields as high as 96 %.

Reaction vessels can have a volume up to and sometimes exceeding 20 m^3. Preferably, they should be glass-lined and have a cooling jacket for water or brine. With rubber- or brick-lined vessels, the only cooling method possible is addition of ice. Pipework can be made from rubber- or glass-lined steel, polypropylene, or glass-fiber-reinforced polyester resin. The filtration equipment (leaf filters, pressure filters, and filter presses) usually is made of rubber-lined steel.

Commercially important diazonium salts used to be the chlorides; the hydrogensulfates; the tetrachlorozincates, $ZnCl_4^{2-}$; the tetrafluoroborates BF_4^-; and the arylsulfonates, $ArSO_3^-$. Progress in textile dyeing consumed much of the market for these products.

Benzenediazonium chloride and its simple substituted derivatives cannot efficiently be crystallized from solution because of their high solubilities. Higher molecular mass is required to give satisfactory yields, e.g., salts with two or more benzene rings per diazonium group. The ease of decomposition of crystalline diazonium chlorides makes their isolation risky and unfavorable. Therefore, only the most stable diazonium cations are produced industrially as chlorides. Diphenylamine-4-diazonium chloride [*101-56-4*] and 4′-methoxydiphenylamine-4-diazonium chloride [*101-69-9*] had special importance because of their use as diazo components of blue dyes. The amino group in the *para* position has a stabilizing effect, giving the diazonium chlorides good storage properties which surpass those of most other industrial diazonium salts.

Diazonium hydrogensulfates are more sparingly soluble than the corresponding chlorides and can be precipitated from the solution obtained in a diazotization reaction by adding sulfuric acid or sodium hydrogensulfate (in practice, 78 % sulfuric acid plus sodium sulfate). The second method of isolation has proved to be very successful in the production of azobenzene diazonium salts.

Diazonium tetrachlorozincates display extremely good crystallization properties. They have high solubility in water or dilute mineral acids but not in concentrated sodium chloride solution, which allows their isolation in good yield. Their large, hard, compact crystals are suitable for further processing and are usually very stable, so that these products are the most common solid diazonium salts.

The first step is the clarification of the diazonium salt solution by filtration. Over a period of ca. 1 h, zinc chloride (0.6 – 0.7 mol per mole of diazonium salt) is added in the form of a ca. 65 % technical-grade solution. For the less-soluble tetrachlorozincates, a dilute solution is used to avoid formation of an amorphous precipitate. Addition of 10 – 20 vol % of sodium chloride completes the precipitation.

Tetrachlorozincates may also be produced from the solutions in concentrated sulfuric acid that are obtained by adding nitrosyl diazotization products to ice. Zinc chloride solution

and sodium chloride are added, precipitating the tetrachlorozincate. However, usually the diazonium hydrogensulfate is precipitated, filtered off, dissolved in water, and then precipitated as the tetrachlorozincate.

Diazonium tetrafluoroborates may be obtained from diazonium salt solutions in hydrochloric acid by addition of tetrafluoroboric acid or its salts, usually the sodium salt. For good yields, an excess of 30 – 50 % is required. Solubilities are similar to those of the tetrachlorozincates, but crystallization properties and yields are inferior. They are used in industry because they are less explosive than other salts.

Diazonium arylsulfonates are obtained by addition of arylsulfonic acids or their alkali metal salts to diazonium salt solutions. If necessary, precipitation can be completed by addition of sodium chloride. Their solubility, for a given cation, depends strongly on the type of arylsulfonic acid.

Drying of diazonium salts must be carried out gently because of their instability. Usually drying cabinets with trays are ventilated with air at ca. 45 – 70 °C and provided with emergency water sprays that come into operation at a predetermined temperature. Vacuum drying is avoided because the gas that is evolved when the material decomposes can destroy equipment. For safety reasons, no drying processes are used that subject the product to mechanical stress.

Only a few diazonium salts are stable enough to be dried in the form in which they are produced. Many are first blended with inert salts, such as sodium sulfate or sodium chloride, to reduce the danger of explosion. Any adhering acid is neutralized by mildly alkaline materials such as borax or magnesium oxide.

The *final processing* of a diazonium salt includes grinding and other processing if the diazonium salt is to be used as a dye intermediate, i.e., incorporation of additives, adjustment of product content, and removal of fines.

Additives suppress undesirable properties in the product and widen the range of uses. The prime concern is to reduce the risk of explosion, e.g., by addition of inert materials such as sodium chloride or sodium sulfate. Another approach is to add compounds with a large heat of fusion or vaporization. In practice, only hydrates, usually aluminum sulfate hydrates, are used [35]. These methods desensitize the products, rendering them completely safe even under severe stress. Some metal salts, such as chromium acetate or aluminum sulfate, or boric acid [36], [37] can be added as buffering agents. Metaboric acid or boric anhydride improves the stability and avoids discolorization or caking [38].

3.1.3. Other Methods of Preparation

Other methods of producing diazonium salts have some importance, e.g., the introduction, replacement, and modification of substituent groups in diazonium salts that are already formed (Section 4.1), and the reaction of nitrous acid with arylamine-*N*-sulfonic acids. Weakly basic amines may be diazotized more easily in this way than in the unsubstituted form, probably because of the solubility imparted by the *N*-sulfonic acid group.

$$\mathrm{ArNHSO_3H + HNO_2 \longrightarrow ArN{\equiv}N^+ + HSO_4^- + H_2O}$$

3.1.4. Uses

Solid diazonium salts with blended additives are used mainly for the manufacture of developing dyes for textile dyeing and printing (→ Azo Dyes, Chap. 3., with tables of trade names, chemical constitutions, and Colour Indexes in → Azo Dyes). The industrial manufacture of stabilized diazonium salts according to the processes described above is declining due to the continuing displacement of developing dyes by reactive dyes. However, diazonium salts prepared as intermediates for azo pigments continue to play an important role in the color industry. Another application used to be diazotype photocopying processes (→ Imaging Technology, Chap. 3.2.1.). Many other uses of diazonium salts are discussed in the following sections.

3.2. Diazotates

Diazonium ions are capable of accepting a hydroxyl ion to form a diazohydroxide which is deprotonated to yield a diazotate anion. [4], [6].

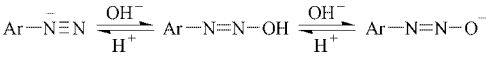

The reaction takes place in weakly alkaline media if strongly electron withdrawing substituents are present which enhance the Lewis acidity of the diazonium ions. For less acidic ions strongly alkaline pH is required. *Z*-Diazotates, formed under kinetic control are converted to the more stable *E*-diazotates in strongly alkaline solution. [39] Experimental evidence points to a direct rearrangement rather than a dissociation/addition mechanism [4].

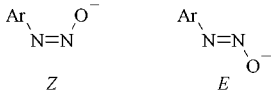

The rate of conversion again depends on the substitution pattern and is fastest for more acidic compounds [40]. Only *E*-diazotates are used in industry. At acidic pH diazotates form diazonium ions in the reverse of their formation. This reaction is the reason for their former use in the textile printing industry, where they served as a latent form of the diazonium ion ("Rapid Fast Dyes"). Sodium and potassium diazotates are crystalline substances that store well. They decompose on heating, though not violently. The industrial use of this class of compounds has been greatly diminished by reactive dye textile printing. For an overview, see [12, 10/3], [34, C1].

3.3. Diazosulfonates

Diazosulfonates have been known since 1869, but due to the lack of appropriate analytical methods their structure was a matter of discussion for years. Diazosulfonates exist in two isomeric forms. The stable *E* (*anti* or *trans*) isomer and the labile *Z* (*syn* or *cis*) isomer.

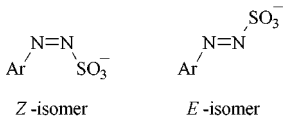

Formation, isomerization, kinetic investigations and photo-isomerization were studied extensively in the 1950s and 60s [41–43]. In the S-coupling reaction of an aromatic diazonium salt with SO_3^{2-}, primarily the *Z* isomer is formed, which rearranges to the *E* isomer. The labile *Z* isomer is in equilibrium with the starting diazonium compound.

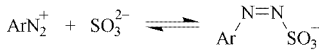

Thermal or photochemical decomposition proceeds via an ionic or radical pathway. The presence of diazonium ions can be detected by coupling reactions with, e.g., β-naphthol. ESR studies proved the existence of radicals on photolysis of diazosulfonates [44]. Industrial applications of diazosulfonates are mainly based on their conversion to reactive diazonium ions and subsequent coupling reactions [45].

Properties. The sodium diazosulfonates are yellow to orange solids and crystallize well. The *Z* isomers have a tendency to explode on isolation as dry material [44], [45]. The *E* isomers usually exhibit considerable stability and can be stored for years without decomposition. For purification they can be recrystallized from water, although prolonged heating should be avoided. Unsubstituted *E*-diazosulfonates and those with electron-withdrawing substituents are usually stable towards dissociation. Only diazosulfonates with electron-donating substituents, e.g., those derived from 4-aminodiphenylamine, generate diazonium ions and form azo dyes in presence of a coupling component. Reducing agents like dithionite and excess of sulfite convert diazosulfonates at least to some extent (depends on excess and quantity of reducing agent used) into phenylhydrazine-*N*-sulfonic acids, which are intermediates in the synthesis of phenylhydrazines. Reduction of isolated diazosulfonates can also be achieved by catalytic hydrogenation [46].

Preparation. Diazosulfonates are prepared by the reaction of an aromatic diazonium salt (prepared in situ) with alkali metal sulfite solution. Usually the diazonium salt solution is added to a solution of sodium sulfite at below 10 °C. Since only the sulfite ion but not the hydrogensulfite ion is suitable for S-coupling, a neutral to weakly alkaline medium is required, that is, the pH should be kept below 8. This can be achieved by the parallel addition of sodium carbonate or sodium hydrogencarbonate.

Immediately after mixing the two solutions, the color turns yellow, indicating product formation. It is recommended to continue stirring after complete addition at slighty elevated temperature (30 – 40°C) for at least one hour, to ensure complete isomerization to the *E* isomer.

More alkaline conditions favor decomposition; in addition, diazotates are formed in a side reaction. To avoid partial reduction to the arylhydrazine-*N*,*N*′-disulfonic acid, only a slight excess (ca. 10 %) of alkali metal sulfite has to be applied. Due to the good solubilities of the diazosulfonates, it is recommended to use the highest possible concentration of diazonium salt solution to achieve acceptable yields. Since potasium diazosulfonates are less soluble, it is sometimes advantageous to use them instead of sodium sulfite. Yields can also be increased by the addition of sodium chloride during precipitation.

The above-mentioned reaction finds broad application, and a large number of diazosulfonates have been synthesized [44], [45].

Uses. Sodium 4-dimethylaminobenzene-*E*-diazosulfonate [*140-56-7*] has fungicidal properties and is used as a soil disinfectant (Dexon) to combat damage to seedlings and roots. Diphenylamine-4-*E*-diazosulfonate and its 4′-methoxy derivative are used in textile printing for producing blue and black shades. They are marketed as mixtures with the coupling components, usually 2-hydroxynaphthalene-3-carboxylic acid anilide. The azo dyes are formed by steaming the prints. Textile prints of *E*-diazosulfonates and Naphtol AS components can also be developed by irradiation. Other applications have been found in reprography [47]. Applications of polymer-bound diazosulfonates as photoresists in offset printing have been described [44]. Diazosulfonates can be considered as thermally stable sources of diazonium salts and and thus could have applications in analytical quick assays [44].

3.4. Triazenes

Triazenes are formed by N-coupling of diazonium ions. Their chemistry forms a broad subject that is discussed in detail in [48].

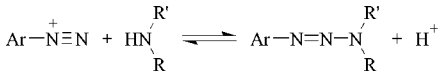

In most cases in solution, a true pH-dependent equilibrium is observed which can be used to regenerate the diazonium ions. For R′ = H tautomerization is possible and results in mixed diazonium ions for the case R = Ar′.

Properties. The stability of triazenes covers a wide range; most compounds will rearrange in weakly acidic media to form aminoazo compounds, while some compounds like 2,2′,4,4′,5,5′-hexachlorodiphenyltriazene are almost inert to aqueous mineral acid [49].

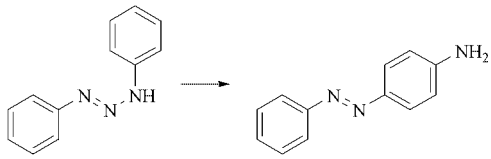

In alkaline media triazenes are stable. Protic compounds can be deprotonated to form intensely colored salts.

Preparation. The only method of importance in triazene manufacture is N-coupling in buffered solutions or suspensions of the amine. Low reaction temperatures (0 – 10 °C) are required to avoid decomposition. Products usually precipitate from the reaction solution; in some cases salting out or spray drying is used to isolate the more soluble products. The reaction of diazonium salts with secondary aliphatic amines takes place smoothly and usually with an excellent yield; industrially relevant triazenes are produced from amines with solubilizing groups such as sarcosine and methyltaurine.

CH_3NHCH_2COOH	$CH_3NHCH_2CH_2SO_3H$
Sarcosine	Methyltaurine
[*107-97-1*]	[*107-86-6*]

A good example is the preparation of sodium 3-methyl-1-(5-chloro-2-methoxyphenyl)-3-(carboxylmethyl)triazene. 4-Chloro-2-amino-1-methoxybenzene hydrochloride is mixed with water, and 32 % hydrochloric acid is added. Ice is added to bring the temperature to 0 °C, and nitrite is then added as a 30 % solution, maintaining the temperature at 0 °C by further

ice addition. After the diazotization, a small amount of diatomite is added, and the liquid is filtered with a clarifiying press. A mixture is prepared of sarcosine, sodium carbonate, sodium chloride, water, and ice. The diazotized solution is added over ca. 30 min, keeping the temperature at 0 °C by ice addition. The precipitated triazene is filtered and dried at 80 °C.

When diazonium salts are treated with primary or secondary aromatic amines, triazene formation competes with azo coupling. The dominant pathway depends on the nucleophilicity of the amine. If the nucleophilicity is not increased by electron-releasing substituents or fused rings, triazene is formed preferentially. Triazene formation is also favored by electron-withdrawing substituents and when the position *para* to the amine group is occupied. In cases where azoamino and triazene formation compete, high pH of the reaction medium favors the triazene pathway, because cleavage and rearrangement of the triazene chain is catalyzed by acid. Secondary aromatic amines have a greater tendency to form triazenes than the corresponding primary amines. Water-soluble triazenes are formed without difficulty from **1** and especially from **2** and **3**.

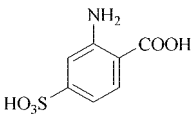

1

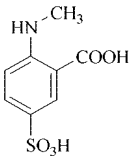

2

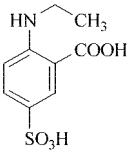

3

The sodium derivative of 3-methyl-1-(4-nitro-2-methoxyphenyl)-3-(4-sulfo-2-carboxyphenyl)triazene (**4**) is prepared by first diazotizing 5-nitro-2-amino-1-methoxybenzene at 10 – 15 °C in dilute sulfuric acid and adding this to a solution of 2-methylamino-5-sulfobenzoic acid that has been neutralized with sodium hydroxide solution and mixed with calcium carbonate. The pH is adjusted to ca. 12 with sodium hydroxide solution, thus precipitating calcium sulfate, which can be filtered off. The filtrate is then spray dried [50, ; p. 25.]

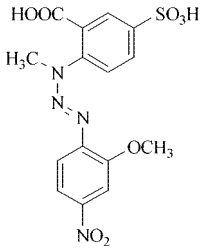

4

Symmetrical 1,3-diaryltriazenes that are to be rearranged to *p*-aminoazo compounds, can be prepared by reacting the amine with half an equivalent of nitrous acid in a moderately acidic medium. The reaction is usually carried out with an excess of amine, which acts as a reaction medium and speeds up the rearrangement. [11].

Uses. Like diazotates, triazenes were also used as diazonium precursors for developing dyes ("Rapidogen Dyes"), but again this application has declined due to the emergence of more efficient dyes and procedures. Some triazenes are active against bacterial and especially against protozoal infections. Berenil [*536-71-0*] was developed to treat diseases caused by trypanosoma and babesias.

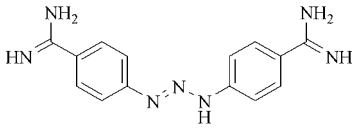

Berenil

1-Aryl-3,3′-dimethyltriazenes can act as prodrugs for 1-aryl-3-methyltriazenes, which are efficient in vivo antitumor agents. One of these compounds, Dacarbazine, is used in the treatment of melanoma and soft tissue sarcoma [51].

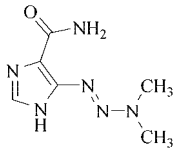

Dacarbazine

3.5. Other Aromatic Diazo Compounds

Diazotized *o*- and *p*-aminophenols (**5**) exist only in mineral acid solution. The diazonium group renders the hydroxyl group much more acidic, increasing the ionization constant by several orders of magnitude, e.g., from 10^{-10} for phenol to 10^{-3} for **5**. The mesomeric system formed is known as quinone diazide (**6**).

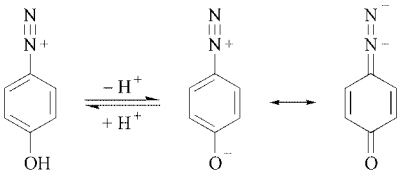

5 **6a** **6** **6b**

Quinone diazides are also known as diazophenols in the literature, yet they differ distinctly in their chemical properties from diazonium salts. Yellow in color, the compounds are sparingly soluble in water and quite explosive, especially when nitro groups are present in the molecule. Picramic acid [*96-91-3*] may be diazotized to give dinitrobenzoquinone diazide [*7008-81-3*] (**7**), which is used as a detonator.

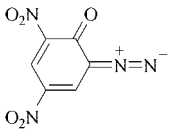

7

Quinone diazides can be employed in the the Sandmeyer reaction, and they form azo dyes. However, their reactivity in coupling is considerably less than that of the corresponding substituted diazonium salts. 2,2′-Dihydroxyazo dyes are formed by coupling with substituted phenols and are capable of forming metal-complex dyes (→ Azo Dyes, Chap. 3.2.1.4. and → Azo Dyes, Chap. 4.1.8.). By the action of light, naphto-*o*-quinone diazides are converted to indene-carboxylic acids, with liberation of nitrogen and ring contraction [11]; thus, they can be used in reprography.

Quinone diazides in the benzene series are formed through diazotization of aminophenols. In some cases, buffering with sodium acetate is necessary. The products are used in the form of a suspension to avoid the danger of explosion.

When *o*-aminonaphthols are treated with nitrite in a mineral acid medium, they are oxidized to naphthoquinones. Diazotization can be achieved in weakly acidic solution in the presence of a small amount of a metal salt, preferably of copper(II) [12, 10/3], [52], [85]; under these conditions the reaction proceeds smoothly. The naphthoquinone diazides formed are stable compounds; for example, they can be crystallized from boiling water, be nitrated, or be chlorinated (see page 13). Moist products can be stored and marketed safely. For details on the production of 4-diazo-3,4-dihydro-3-oxo-naphthalene-1-sulfonic acid [*4857-47-0*], see [52]. Naphthoquinone diazides are important starting materials in the manufacture of metal-complex dyes and are used in reprography (especially films for offset litho printing and IC manufacture (→ Imaging Technology).

Aryldiazo ethers, esters, sulfides, thioethers, sulfones, cyanides, and phosphoric esters have very few practical applications. One or two of these have been suggested as polymerization catalysts and modifiers. The acaricide Milbex [*2274-74-0*] is a tetrachloroaryldiazo thioether.

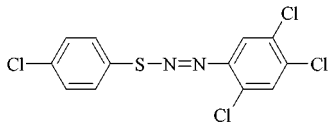

Milbex

3.6. Analysis

The best way to detect aromatic diazo compounds is by the azo coupling reaction. The substance under investigation is treated with (1) a phenolic azo component, such as resorcinol, β-naphthol derivatives or H-acid dissolved in sodium hydroxide or carbonate solution, or (2) an aromatic amine capable of azo coupling dissolved in dilute acetic acid. The most convenient method is to use a spot test [53].

E-Diazotates and triazenes are converted to diazonium salts by acid treatment, which is not successful in all cases since some triazenes are very stable even under these conditions. Non-coupling *E*-diazosulfonates may be converted to diazonium salts by treatment with oxidizing agents. To identify the aromatic part of the

molecule, either direct spectroscopy (IR, NMR, MS) is used or the diazonium salts are converted to chlorinated aromatic compounds by the Sandmeyer reaction, which are then identified by standard methods. The universal chemical method for quantitative analysis of diazonium salts is controlled decomposition by treatment with acid and copper(I) ions and measurement of the nitrogen evolved [54], [55]. In industry, diazonium salts and their solutions are often analyzed by titration. The diazonium salt solution is gradually added to a known amount of coupling component, e.g., R salt (disodium 3-hydroxy-2,7-naphthalenesulfonate [*135-51-3*]), β-naphthol, or 2,4-diaminotoluene. The end point is detected by the disappearance of the coupling component or the appearance of the diazonium salt, as detected by spot test methods. Unreactive or easily decomposed diazonium compounds may be treated with an excess of the coupling component, followed by back-titration. Quantitative analysis can also be accomplished by spectophotometry provided internal or external standardization is possible. The products must be fairly pure so that secondary or decomposition products do not interfere with the absorption maximum of the diazo compound.

4. Diazo Reactions

From a practical point of view, the most important reactions of the diazo compounds can be placed in three groups: (1) reactions in which the diazo group is retained but substituents are introduced or replaced, (2) reactions in which the diazo group is modified, and (3) reactions in which the diazo group is replaced.

4.1. Chemical Transformation with Retention of the Diazo Group

Substituents can be introduced or replaced without changing the diazo group.

Electrophilic Substitution. The diazonium group is the strongest electron-withdrawing substituent known. Therefore, electrophilic substitutions on the same nucleus are very difficult to achieve and largely of theoretical interest only. Under special circumstances, halogen atoms or nitro or sulfonic acid groups may be introduced into the aromatic ring. An example is the production of 6-nitro-1-diazo-2-naphthol-4-sulfonic acid [*5366-84-7*] [56].

Nucleophilic Substitution. Owing to the strongly electron-withdrawing character of the diazonium group, nucleophilic replacement of *ortho* and *para* substituents is possible. Ease of replacement increases in the order SO_3H, OCH_3, Cl, NO_2, and $-N{\equiv}N^+$. In the benzene-1,4-bis(diazonium) ion, one of the diazonium groups can be easily replaced by a halogen or hydroxyl group. Replacement of a nitro group can take place only in the presence of an additional electron-withdrawing substituent; therefore, solutions of *o*- and *p*-nitrodiazonium ions, which are heavily used in industry, are stable under typical handling conditions.

On the other hand, the *o*-nitro group in diazotized 2,4-dinitroaniline can be replaced by nucleophiles. This reaction muddies the color of the pigment derived from 2,4-dinitroaniline and β-naphthol during production. Care has to be taken to avoid these side reactions when diazonium salts with strong electron-withdrawing substituents are prepared or converted.

The replacement of chlorine by a hydroxyl group is used in the manufacture of 4-hydroxy-3,5-bis(diazo)benzenesulfonic acid, a component for complex-forming azo dyes.

4.2. Reactions at the Diazo Group

Azo Coupling. Diazonium salts combine with many electron-rich aromatic compounds and compounds with active CH or NH groups to form azo compounds, which may tautomerize subsequently. Coupling with phenols, aromatic amines, and acetoacetarylides is most important for the dyes and pigments industry; for a detailed discussion see → Azo Dyes.

Formazan Formation. Diazonium ions react with functional groups carrying active hydrogen as mentioned above. Starting from an aldehyde hydrazone, the product is a formazan:

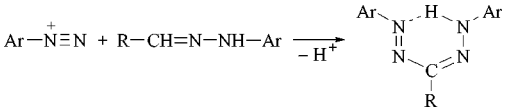

Compounds with a doubly activated methylene group, e.g., acetoacetyl compounds, first form an arylhydrazone, which can react in an alkaline medium with another diazonium ion to give the formazan. In this reaction one of the two activating radicals is split off (Japp – Klingemann reaction)

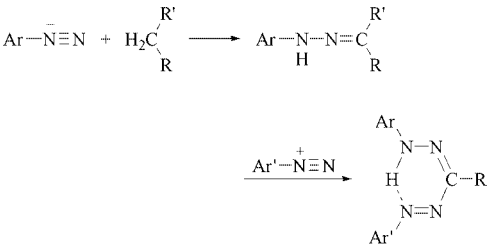

R, R' = CHO, $COCH_3$, CN, $COOCH_3$, $CONH_2$, COOH

By means of kinetic control, unsymmetrical formazans can be produced by using two different diazonium salts, provided their rates of reaction differ.

Formazans are intensely colored compounds. They show amphiphilic properties and hence behave as indicators with changes in pH. For this reason and because of their poor lightfastness, simple formazans are rarely used for textile dyeing. However, they are used in photography, e.g., for yellow filters and color developers.

Formazans form heavy-metal complexes, the industrially relevant ones being those of Co, Cr, Cu, and Ni. These complex dyes have considerably improved stability, owing to the presence of additional functional groups (OH, COOH) in the *ortho* positions of the aryl radicals attached to the nitrogen atom, which act as ligands for the metal ion. Excellent lightfastness and low sensitivity to acids and bases make some of these complexes valuable textile dyes (→ Metal-Complex Dyes).

Cyclization Reactions. Reactive *ortho* substituents react with diazo groups to give heterocyclic products. [57], [58]. This pathway is exploited to convert diazotized *o*-toluidines to indazoles. Although the parent substance *o*-toluidine and derivatives with an electron-releasing substituent undergo this reaction only to a minor extent, the mononitro derivatives in acetic acid solution or the dinitro derivatives in sulfuric acid solution give high yields of indazole. A frequent side product is a 3-arylazoindazole. *o*-Aminobenzamides give benzotriazines under diazotizing conditions e.g. Azinphos-methyl [*86-50-0*], a component of the insecticide Gusathion.

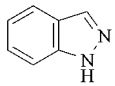

Indazole

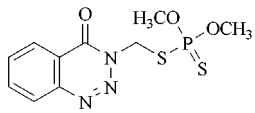

Azinphos-methyl

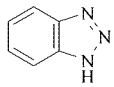

Benzotriazole

o-Diamines also do not yield diazonium salts; instead, they are converted to cyclic triazoles when treated with nitrous acid. The preparation of benzotriazole and its simple derivatives is best carried out in dilute acetic acid [59]. Nitrite is added at once, and no cooling is required. Unlike the open-chain triazenes, triazoles appear to be stable toward mineral acid. Triazoles can form salts with acids or bases. Benzotriazoles and tolutriazoles are corrosion inhibitors.

Reduction. Diazonium salts and covalent diazo compounds such as diazotates, diazosulfonates, diazosulfones, and triazenes can be reduced to arylhydrazines. Most common reducing agents are effective. Electrochemical reduction has also been described [60], [61]. Tin(II) chloride reduces diazonium salts under mild conditions (low temperature) in highly acidic media and provides a convenient method for aromatic hydrazines. Due to the large excess of tin(II) chloride which finally enters the wastewater, this method is limited to laboratory scale. In industry, sodium sulfite and sodium hydrogensulfite are used almost exclusively. When the reaction proceeds sluggishly, it can be brought to completion by the addition of zinc dust. The sulfite reduction of aryldiazonium salts to arylhydrazines takes place stepwise via aryldiazosulfonates (**8**), arylhydrazine-α, β-disulfonic acids (**9**), and arylhydrazine-β-sulfonic acids (**10**).

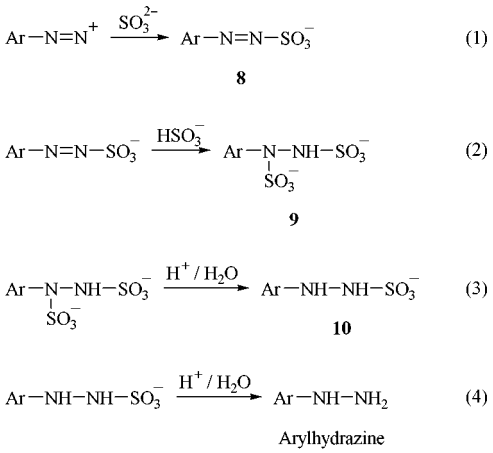

Reactions (1) and (2) take place in a weakly acidic to weakly alkaline medium, reaction (3) in weak to moderately strong acid, and reaction (4) in the mineral acid range. Commercial 40 % sodium hydrogensulfite solution is converted to sulfite by addition of sodium carbonate or sodium hydroxide solution. Each diazonium cation requires two sulfite equivalents for reduction. Usually 10 – 20 % excess of sulfite is applied to achieve optimum conversion and yields. In the preparation of unsubstituted phenylhydrazine, 102 % of the theoretical amount suffices [26].

The diazonium solution or suspension is added to the sulfite solution. If no electron-withdrawing substituents are present, the mixture is heated at 60 – 80 °C to give **10**. During reduction, the pH should be held between 5 and 7 by means of inorganic bases. Colored impurities can be reduced with a little zinc dust. At this stage, if necessary, a charcoal filtration may be carried out to remove tarry or insoluble materials.

Hydrolysis of the arylhydrazine-α, β-disulfonic acid to arylhydrazine-β-sulfonic acid already takes place under mild acidic conditions. To minimize salt generation it is also possible to reduce the diazosulfonate in good yields catalytically with hydrogen in the presence of Ni or noble metal catalysts to arylhydrazine-β-sulfonic acid.

The second sulfonic acid group is much more firmly bonded than the first, and the conversion to arylhydrazine requires warming under strongly acidic conditions, usually accomplished by adding concentrated hydrochloric acid or sulfuric acid to the arylhydrazine-α, β-disulfonic acid or arylhydrazine-β-sulfonic acid and heating at 6 – 100 °C. In some cases solvents were found to be advantageous in hydrolysis [62].

Arylhydrazines with sulfonic acid groups in the aromatic nucleus precipitate as sparingly soluble zwitterions. If there are no sulfonic acid groups present, the hydrazines are isolated as sulfates or hydrochlorides. Free hydrazines can be precipitated by neutralizing the acidic solution solution with alkali. If solid, they are filtered off; if liquid, they are dissolved in solvents and isolated by fractional vacuum distillation.

Phenylhydrazine hydrochloride can also be manufactured by a continuous process [33], but most arylhydrazines are made by batch processing. Arylhydrazines are starting materials for pyrazolones, which serve as dye precursors. They react with dioxo compounds to give azo dyes (→ Azo Dyes, Chap. 2.4.4.). In addition, phenylhydrazines are versatile building blocks for heterocycles such as indoles, triazolinones, pyrazoles and pyridazinones, which are valuable intermediates in the pharmaceutical and agrochemical industries [18], [63–65].

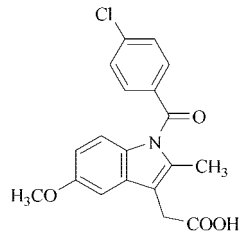

Indomethacin

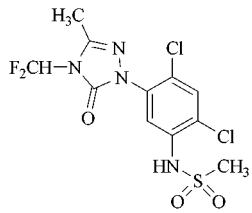

Sulfentrazone

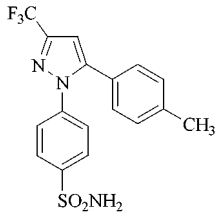

Celecoxib

4.3. Dediazotization (Replacement of the Diazo Group)

The term dediazotization describes reactions of diazo- and diazonium compounds in which the N_2 molecule is a reaction product. In the following, dediazotization of arenediazonium compounds is discussed. The diazo group can be replaced by a hydrogen, hydroxyl, halogen, cyanide group, etc.

Hydrodediazotization. The replacement of the diazo group by hydrogen is not only used on a laboratory scale but finds industrial application for the manufacture of compounds that are not accessible by direct electrophilic aromatic substitution reactions. For example, 1,3,5-trichlorobenzene can be prepared by chlorination of aniline, diazotization and hydrodediazotization. Treatment of diazonium salts in boiling ethanol leads to the desired product, but quite often ethoxydediazotization occurs. Excellent results can be obtained by using hypophosporous acid as the reducing agent. For some special purposes salts, esters, and amides of formic acid or even formic acid itself are used [66]. For example, 3-hydroxynaphthalene-1-sulfonic acid [*6357-85-3*], a coupling component for dyes, can be manufactured in excellent yields.

Hydroxydediazotization. Thermal decomposition of diazonium ions in aqueous acid gives transient aryl ions, which react with water to give phenols:

$$Ar{-}N^+ \equiv N \longrightarrow Ar^+ + N_2$$
$$Ar^+ + HOH \longrightarrow ArOH + H^+$$

In the presence of other nucleophiles Z^-, side reactions can occur:

$$Ar^+ + Z^- \longrightarrow Ar{-}Z$$

Therefore, diazotization and dediazotization are best carried out in sulfuric acid solution. The simplest method is to heat the diazonium salt solution to 80 – 100 °C, but the products obtained are usually not very pure. Higher yields and purities can be achieved by adding the diazonium salt solution to hot (boiling), dilute sulfuric acid. To avoid diazo coupling of the phenol with unconverted diazonium salt, the acidity of the reaction mixture should be sufficiently high. At sulfuric acid concentrations of 45 – 60 %, reaction temperatures can be as high as 118 – 142 °C. The sulfuric acid can be partly replaced by sodium sulfate. Sometimes Cu salts are used in the decomposition reaction [67].

During hydroxydediazotization, side products, sometimes dark in color and insoluble, are often formed in considerable amounts, e.g., diphenyl sulfates, diphenyl ethers, hydrocarbons, nitrosophenols, indazenes, dioxazines, hydrazobenzenes, and arylsulfonic acids. This formation of diazo resin, which visibly builds up during the course of the reaction, can be minimized by careful diazotization and avoidance of excess nitrite. Continuous removal of the phenol, e.g., by steam distillation, is favorable. Conducting the reaction in a continuous manner can also improve yield and quality. Yields are usually in the range of 50 – 75 % but can exceed 90 % in some cases if phenols are continously removed from the reaction mixture [6, ; pp. 6, 225], [67]. Improved yields can sometimes be achieved by dissolving the formed phenol in an organic solvent such as xylene.

The preparation of hydroxyanthraquinones from aminoanthraquinones can be accomplished in concentrated or fuming sulfuric acid. Diazonium salts that do not react in sulfuric acid even at elevated temperatures can be converted in phosphoric acid at ca. 200 °C.

The industrial importance of hydroxydediazotization lies not only in the transformation of an amino group into a hydroxyl group but also provides a method to manufacture chloro- or nitro-substituted phenols and cresols at the *meta* position, e.g., the preparation of 2,3-dichlorophenol [*576-24-9*] [68] or 4-methyl-3-nitrophenol [*2042-14-0*].

Chloro- and Bromodediazotization. Treatment of aromatic diazonium salts in aqueous solution with chloride or bromide ions in presence of the corresponding Cu halide leads to aryl chlorides and bromides. The reaction

has been known since 1884 and is called the Sandmeyer reaction; the following mechanism is proposed:

$$2\,ArN_2^+ + CuCl_2^- \longrightarrow Ar^\bullet + N_2 + CuCl_2$$

$$Ar^\bullet + CuCl_2 \longrightarrow ArCl + CuCl$$

The first step is the reduction of the diazonium group by electron transfer from the CuX_2^- and formation of an aryl radical. Finally the radical abstracts a halogen from the copper(I) chloride to give the arylhalogenide and CuX. Azo compounds and diphenyls can be formed as side products in Sandmeyer reactions.

$$2\,ArN_2^+ + CuCl \longrightarrow ArN{=}NAr + N_2 + 2\,Cu^{2+} + 2\,Cl^-$$

$$2\,ArN_2^+ + + CuCl \longrightarrow Ar{-}Ar + 2\,N_2 + 2\,Cu^{2+} + 2\,Cl^-$$

These can be minimized by increasing the concentration of halide in the solution. Enough hydrochloric acid should be added that its concentration in the reaction mixture does not fall below 5 – 10 wt % during addition of the diazonium salt solution. Additionally, the concentration of the diazonium salt should be kept as high as possible.

In industry, a molar ratio of copper(I) halide to diazonium salt from 1/10 to 1/5 is used, although in some cases considerably more catalyst is required. For example, one molar equivalent is used in the manufacture of 2-chloro-5-nitrophenol [*619-10-3*] from 5-nitro-2-aminophenol. Copper(I) chloride is sparingly soluble in water and is usually used in the form of the more soluble chloro complex. The Sandmeyer reaction can be conducted in glass-lined, rubber-coated, or brick-lined (acid resistant) vessels. The filtration equipment and pipework are made of rubber-lined steel, and the pipework is often made of suitable plastic. Reaction temperatures are usually in the range of 10 – 30 °C, although in a few instances, they are higher, e.g,. 75 °C for 4-chlorotoluene-2-sulfonic acid. The reaction product is isolated by steam distillation or filtration. The copper is precipitated from the mother liquor and recycled.

3-Chlorotoluene [*108-41-8*] is manufactured on an industrial scale by a Sandmeyer reaction [70]. The Sandmeyer reaction may also be carried out with copper(I) bromide and hydrobromic acid [71].

Cyanodediazotization. Replacement of the diazonium group by cyanide in presence of Cu catalysts to form aromatic nitriles (cyano Sandmeyer reaction) is similar to the reaction with Cl^- and Br^-. The reaction is usually conducted in a neutral medium to avoid liberation of HCN.

$$ArN_2^+ + CN^- \longrightarrow ArCN + N_2$$

Because copper(I) cyanide is insoluble, sodium copper(I) cyanide, $NaCu(CN)_2$ is used as catalyst. The amount of catalyst required is larger compared with chloro-dediazotization, usually 25 – 30 mol %. Reactions can be performed at ambient temperature. The large excess of cyanide (180 – 300 % of theory) may result in unwanted diazocyanides. To minimize this side reaction, only part of the alkali metal cyanide is added with the Cu catalyst; then the remaining quantity is added in portions along with the diazonium salt solution, e.g., in the production of 5-chloro-4-nitro-2-methylbenzonitrile [72].

Iodo- and Fluorodediazotization. Unlike aryl chlorides and bromides, which are accessible from the Sandmeyer reaction, aryl fluorides cannot be achieved in this way. In case of the iodo derivatives, no copper(I) catalyst is required since the redox potential of iodine is sufficient for electron transfer to the diazonium salt. It has been proposed that initially aryl radicals as well as iodine and triiodide are formed followed by coupled chain reactions [6, ; p. 236].

$$ArN_2^+ + I^- \longrightarrow ArI + N_2$$

For aryl iodides the reaction is carried out in mineral acid, usually hydrochloric or sulfuric acid, and proceeds much faster than with the other halides. A concentrated aqueous solution of alkali metal iodide is added to a diazonium salt solution. Inverse addition is also possible. For each mole of diazonium salt, 1.1 – 2.0 equivalents of iodide is used. Elevated reaction temperatures to increase rate of conversion are not required. Volatile aryl iodides may be recovered by steam distillation. Less volatile products are isolated by filtration or extracted by solvents such as toluene. The plant growth regulator 2-iodo-3-nitrobenzoic acid [*5398-69-6*] is manufactured on commercial scale by iododediazotization [69].

If solutions of diazonium fluorides in anhydrous hydrofluoric acid are warmed, fluoroaromatics are obtained. The diazonium fluorides are prepared by dissolving the amine in 2 – 3 equivalents of hydrofluoric acid and diazotized by adding dry sodium nitrite at ca. 0 °C. When the reaction mixture is heated to reflux (ca. 30 °C), nitrogen is released. The reaction takes place smoothly over several hours.

Another important route to fluoroaromatics is the thermal decomposition of neat diazonium fluoroborates (Balz – Schiemann reaction).

$$(ArN{\equiv}N)^{+}BF_4^{-} \longrightarrow ArF + N_2 + BF_3$$

Diazoniumtetrafluoroborates are precipitated by addition of $NaBF_4$ to a cold solution of diazonium salt prepared as usual with sodium nitrite in hydrochloric acid. The BF_4^- salts are comparatively stable and can be isolated in high yields.

Mercapto- and Chlorosulfodediazotization. Thiophenols can not be synthesized by a similar procedure to phenols since the thiophenol catalyzes homolytic dediazotization. However they can be synthesized by treating sodium disulfide with diazonium salt to form a diaryl disulfide, which can be reduced to the thiophenols by means of Zn – acetic acid or Fe – H_2SO_4 and other reducing agents. Another route to thiophenols is the reaction of a diazonium salt solution with potassium ethylxanthogenate [6, ; p. 227], [73] and subsequent alkaline hydrolysis of the intermediate formed. The dediazotization is carried out at elevated temperature to immediately decompose the unstable intermediates (appropriate care has to be taken to avoid accumulation). On acidification, thiophenols are obtained in reasonable yields (63 – 75 %).

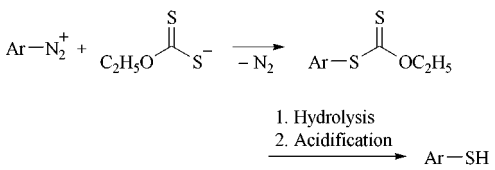

Aromatic sulfochlorides are of industrial importance as precursors in the manufacture of sulfonamides, alkyl aryl sulfones, etc. They can be prepared from diazonium salts and sulfur dioxide. Copper(I) chloride is used as a catalyst.

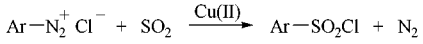

Catalyst: Cu(F)

The reaction is carried out under reaction conditions that exclude formation of diazosulfonates (either strong hydrochloric acid or in the absence of water). Sulfur dioxide is present in large excess. Electron-withdrawing substituents in the aromatic nucleus favor the reaction and improve the yield. The sulfochlorides can be reduced to thiophenols by means of zinc – sulfuric acid. Methyl-2-*tert*-butyl-5-thiophenol is thus synthesized and can be used as stabilizer for vinyl chloride polymers, as a chain-transfer agent, and as a peptizer [74].

Nitrodediazotization can be applied if a nitro group is to be introduced into an aromatic nucleus which due to the substitution pattern cannot be easily synthesized by direct nitration. The reaction is performed with sodium nitrite in neutral or weakly alkaline medium in the presence of a Cu catalyst . Generous proportions of nitrite and catalyst are necessary: 5 mol of sodium nitrite and equimolar quantities of catalyst per mole of diazo compound. Catalysts used include copper powder, CuO, $CuSO_4$, $CuCO_3$, and $Cu_3(SO_3)_2 \cdot 2\,H_2O$, the latter being especially efficient. If electron-withdrawing substituents are present in the substrate, no catalyst is required.

Acyldediazotization. Diazonium salts are transformed with aldoximes into aryl oximes which are hydrolyzed to ketones. A copper sulfate/sodium sulfite catalyst is essential. In most cases, yields are moderate.

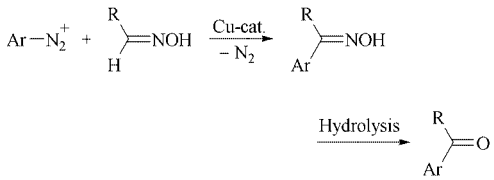

Heck Vinylation and Cross-Coupling Reactions. Palladium-catalyzed vinylation (Heck reaction) and cross-coupling reactions are versatile synthetic methods for C–C coupling [75–77].

Apart from aryl and vinyl halides, which represent the classical electrophiles in the Heck reaction, its synthetic scope has been broadened by succesful application of other substrates, e.g., benzoic anhydrides, aroyl chlorides, and in particular arenediazonium salts [78].

Diazonium tetrafluoroborates react with various olefins in alcohols in the presence of $Pd-CaCO_3$, or $Pd(OAc)_2$ with additional $CaCO_3$ as base under mild conditions to give substituted styrene derivatives in good to excellent yields (61 – 93 %).

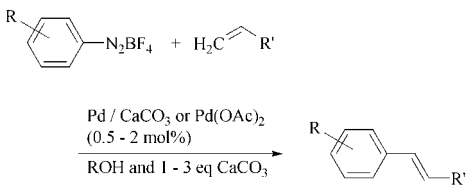

This method furnishes a convenient route to cinnamic acid derivatives, which are of industrial importance as UV absorbers, as antioxidants, and as key intermediates for pharmaceuticals, agrochemicals, and fragrances [79]. Tetraaryls and extended conjugated aromatic systems have been synthesized by twofold Heck reaction of bis-diazonium salts with acrylates and styrene or twofold cross-coupling with phenylboronic acid [81]. Fluorescent whitening agents have been synthesized by Heck vinylation of e.g. diazotized 2-aminobenzenesulfonic acid with 4,4′-divinylbiphenyl [80].

Other Reactions. Due to their high reactivity, diazo and diazonium compounds are capable of reacting in manifold ways and with various other substrates to yield, e.g., biphenyls via Gomberg – Bachmann reaction, aryl azides, aryl isocyanates, arylarsonic acids (Bart reaction), or carbonylation products. Details of these and other reactions are given in [6–10].

5. Environmental Protection

In general, no problems arise with waste gases from the manufacture of diazo compounds or from industrial diazo reactions. Larger quantities of waste gases occur during manufacture of certain azobenzenediazonium salts and Sandmeyer nitrile synthesis due to the excess of nitrite used or contamination of liberated nitrogen with hydrogen cyanide. These substances can be removed by alkali washing or by treatment with oxidizing reagents.

The wastewater contains dissolved reaction and decomposition products together with quite a large amount of mineral acid or caustic alkali. If the diazonium salts are isolated as tetrachlorozincates, the wastewater also contains considerable amounts of zinc ions. If all wastewaters are led into one collecting tank, partial neutralization takes place, which can be completed by the addition of a supension of calcium hydroxide in water, or powdered limestone. Many diazonium salts rapidly decompose under these conditions, usually into insoluble products that precipitate along with zinc hydroxide or carbonate. Residual products can be converted to insoluble azo dyes by adding wastewater containing naphtholate or other coupling components. The solids are concentrated by sedimentation, separated by filtration, and then incinerated or dumped on a certified landfill. The wastewater is discharged, sometimes after filtration through a sand filter, to a biological treatment facility to reduce the organic content still further.

Wastewater from the Sandmeyer reaction has to be treated to remove all traces of copper, preferably by adding sodium sulfide solution or by reductive treatment with, e.g., sodium borohydride. The precipitated copper sulfide or copper is isolated and recycled. The mother liquor from the cyanide replacement reaction contains cyanide. It is usually oxidized with chlorine, hypochlorite, or hydrogen peroxide to form cyanate. It may be also hydrolyzed to sodium formate and ammonia at a temperature of 200 °C [82].

Other solid wastes include filter residues from clarifications during diazotization processes, which mainly consist of activated carbon and filter aids such as diatomite. If the diazotization reactions are well controlled, only a small amount of undiazotized bases and insoluble decomposition products is present. These solid wastes can be incinerated like other organic waste products.

Table 1. Acute toxicity of some diazonium salts [87]

Diazonium salt *	Toxicity	
	LD_{50} (rat, oral), mg/kg	LC_0 (fish **), mg/L
2-Chlorobenzenediazonium tetrachlorozincate [*14263-92-4*]	1900	10 – 100
3-Chlorobenzenediazonium tetrafluoroborate [*456-39-3*]	1000	<10
5-Chloro-2-methoxybenzenediazonium tetrachlorozincate [*68025-25-2*]	1500	100 – 500
4-Chloro-2-nitrobenzenediazonium tetrachlorozincate [*14263-89-9*]	1600	10 – 100
4-Chloro-2-trifluormethylbenzenediazonium 1,5-naphthalenedisulfonate [*85222-98-6*]	3400	<10
2,5-Dichlorobenzenediazonium tetrachlorozincate [*14239-23-7*]	2200	<10
2,5-Dimethoxy-4-[(4-nitrophenylazo)]benzenediazonium tetrachlorozincate [*64071-86-9*]	3700	100 – 500
2-Methoxy-4-nitrobenzenediazonium 1,5-naphthalenedisulfonate [*49735-71-9*]	2500	
2-Methoxy-5-nitrobenzenediazonium tetrachlorozincate [*61919-18-4*]	1900	100 – 500
4-Methoxy-2-nitrobenzenediazonium tetrachlorozincate [*14239-24-8*]	2100	10 – 100
4-[(4-Methoxyphenyl)amino]benzenediazonium chloride [*101-69-9*]	200	
2-Methyl-4-[(2-methylphenyl)azo]benzenediazonium hydrogen sulfate [*101-89-3*]	2000	
2-Methyl-4-nitrobenzenediazonium tetrafluoborate [*455-90-3*]	1200	
2-Methyl-5-nitrobenzenediazonium 1,5-naphthalenedisulfonate [*49735-69-5*]	1800	10 – 100
2-Nitrobenzenediazonium tetrafluoroborate [*365-33-3*]	2000	
4-(Phenylamino)benzenediazonium chloride [*101-56-4*]	400	<10
3,3′-Dimethoxy-[1,1′-biphenyl]-4,4′-bis-(diazonium) tetrachlorozincate [*14263-94-6*]	1600	<10
3,3′-Dichlorobenzidinetetrazonium salt	>2000	70.7
2,4,5-Trichloranilinediazonium chloride	200 – 2000	11.9
2-Methoxy-4-nitroanilinediazonium chloride	200 – 2000	67.1
4-Chloro-2-nitroanilinediazonium chloride	200 – 2000	31.5

* As a commercial azoic diazo component: the amount of diazonium salt in a commercial diazo component is ca. 25 – 50 wt %, the rest being inorganic salts such as sodium sulfate and aluminum sulfate;
** *Lanciscus idus.*

6. Safety

Many diazo compounds are explosive, some even having the brisance of detonators. Therefore, a diazo product whith unknown explosive properties should be prepared in small quantities first and subjected to ignition and impact tests, as well as differential thermal analysis [83]. Before full-scale or pilot-plant production is initiated, an industrial safety investigation should be performed, e.g., following the guidelines of the Bundesamt für Materialprüfung (German Federal Office for Materials Testing) [84]. Commercial products must fulfill the criteria for non-explosive substances; otherwise, they can be marketed only under restrictive conditions. The same restrictions should be applied to intermediates.

In addition, flammability must not be overlooked. Although commercial products are usually self-extinguishing because of the presence of extenders, many intermediates, especially those containing nitro groups, can deflagrate.

Aqueous solutions of diazonium salts can decompose, but do not explode. They are protected against dangerous increase in temperature by the high specific heat and heat of vaporization of water. However, diazonium salt solutions in concentrated sulfuric acid, as formed in nitrosyl-diazotization, are a special case. Because of the low specific heat and the high boiling point of sulfuric acid, the heat of decomposition can raise the temperature sufficiently to trigger explosion [85].

7. Toxicology

Diazomethane is the most thoroughly investigated diazo compound with regard to toxicology. Acute poisoning follows a course similar to that of phosgene poisoning: a latent period is followed by severe difficulty in breathing and, in severe cases, by pulmonary edema. Animal trials have revealed carcinogenic properties; therefore, diazomethane is included under carcinogens Section 2 (former Section III A2) of the Technische Regeln für gefährliche Arbeitsstoffe; an MAK value is not quoted. For manufacture and use of diazomethane special rules must be observed [86]. For another aliphatic diazo compound, ethyl diazoacetate [*623-73-4*], the LD_{50} value for acute oral toxicity (rat) is 400 mg/kg.

Table 1 provides a survey of the acute toxicity of some aromatic diazonium salts used as azoic diazo components.

The acute oral toxicity of aromatic diazo compounds is somewhat greater: the LD_{50} values are 430 mg/kg for 3,3-dimethyl-1-phenyltriazene [*7227-91-0*], 224 mg/kg for 1,3-diphenyltriazene [*136-35-6*], and 63 mg/kg for 4-dimethylaminobenzene-*E*-diazosulfonate [*140-56-7*] [88].

8. References

General References

1. S. Patai: *The Chemistry of Diazonium and Diazo Groups,* J. Wiley & Sons, Chichester, England, 1978.
2. K. Holzach: *Die aromatischen Diazoverbindungen,* Enke Verlag, Stuttgart 1947.
3. K. H. Saunders: *The Aromatic Diazo Compounds,* 3rd ed., E. Arnold, London 1985.
4. H. Zollinger: *Chemie der Azofarbstoffe,* Birkhäuser, Basel-Stuttgart 1958.
5. H. Zollinger: *Azo and Diazo Chemistry,* Interscience Publ., London-New York 1961.
6. H. Zollinger: *Diazo Chemistry I and II,* VCH, Weinheim, 1994.
7. M. P. Doyle, M. A. McKervey, T. Ye: *Modern Catalytic Methods for Organic Synthesis with Diazo Compounds,* Wiley , New York, 1998.
8. M. Regitz, G. Maas: *Diazo Compounds,* Academic Press, New York, 1986.
9. K.-P. Zeller, *Angew. Chem. Int. Ed. Engl.* **14** (1974) 32 – 43.
10. P. Rys: "Diazotation of Weakly Basic Aromatic Amines", in *Physico-Chemical Principles of Color Chemistry* , Blackie 1996.
11. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin-Göttingen-Heidelberg 1964, pp. 177 – 208.
12. *Houben-Weyl,* **10/2,** 169 – 692; **10/3,** 1 – 212, 545 – 626, 631 – 743; **10/4,** 473 – 893.
13. I. G. Laing: "Arenediazonium Salts and Diazo Compounds," in *Rodd's Chemistry of Carbon Compounds,* vol. III, Section C, Elsevier, Amsterdam-London-New York 1973, pp. 11 – 88.

Specific References

14. "Novolak-Diazoquinone Resists: The Imaging System of the Computer Chip", *Angew. Chem. Int. Ed. Engl.* **35** (1996) 2428 – 2440.
15. R. Dammel: *Diazonaphthoquinone-Based Resists,* SPIE Optical Engineering Press, Washington, 1993.
16. Clariant, EP 0651893 B1, 1999 (E. C. Jacovich).
17. Hoechst, EP 0413201 B1, 1994(S. Scheler, G. Buhr).
18. Monsanto, WO 98/08807, 1997 (A. S. Shah, J Clark, Y. Ma, J. Peterson, L. Patelis).
19. A. N. Nesmejanow, K. A. Konzeschkow, W. A. Klimowa, *Ber. Dtsch. Chem. Ges.* **68** (1935) 1877 – 1883.
20. O. Klais, T. Grewer, *Inst. Chem. Eng. Symp. Ser.* **82** (1983) C24 – C34.
21. P. D. Storey, *Inst. Chem. Eng. Symp. Ser.* **68** (1981) 3/P/1 – 3/P/9.
22. J. Ribka, *Angew. Chem.* **70** (1958) 241 – 244.
23. D. F. DeTar, A. Ballantine, *J. Am. Chem. Soc.* **78** (1956) 3916 – 3920.
24. W. F. Giauque et al., *J. Am. Chem. Soc.* **82** (1960) 62.
25. H. Zollinger, *Angew. Chem.* **90** (1978) 157; *Angew. Chem. Int. Ed. Engl.* **17** (1978) 147.
26. H. Hupfer, *Angew. Chem.* **70** (1958) 244 – 246.
27. H. Kindler, D. Schuler, *Chem. Ing. Tech.* **37** (1965) 402 – 405.
28. Bayer, DE-OS 2 741 925, 1977 (K. Breig, G. Demel, N. Hamm).
29. Hoechst, DE-OS 2 825 655, 1978 (H. Behringer, K. Karrenbauer).
30. ICI, EP 3 656, 1979 (H. Alterton, I. Hodgkinson).
31. Ciba, US 4 246 171, 1981 (A. Hamilton, C. Nelson).
32. Fiat Final Report 1023, p. 6.
33. Hoechst, DE 1 543 623, 1966 (H. Hertel).
34. BIOS DOCS 1156/1121.
35. Hoechst, DE 2 521 650, 1975 (H. Hertel, R. Kostka, H. Milz, R. Ullrich).
36. Hoechst, DE 2 433 232, 1974 (H. Hertel).
37. Hoechst, DE 2 527 264, 1975 (H. Hertel).
38. Showa Chemical Co., JP-Kokai 82 139 146, 1981.
39. E. S. Lewis, H. Suhr, *Chem. Ber.* **91** (1958) 2350 – 2358.
40. V. Sterba, Scientific Papers of the University of Pardubice, Ser. A - Fac. Of Chem. Techn. 1 (1995).
41. R. Dijkstra, J. de Jonge, *Recl. Trav. Chim. Pays-Bas* **77** (1958) 538.
42. E. S. Lewis, H. Suhr, *Chem. Ber.* **92** (1959) 3031.
43. L. K. H. van Beek, J. Helferich, *Recl. Trav. Chim. Pays-Bas* **87** (1968) 997.

44. O. Nuyken, B. Voit, *Macromol. Chem. Phys.* **198** (1997) 2337 – 2372.
45. *Houben-Weyl,* **10/3,** 570 ff.
46. Hoechst AG, EP 0 073 000, 1981 (R. Berthold).
47. Ricoh, JP 56089736 A2 1981.
48. K. H. Saunders, R. L. M. Allen: *Aromatic Diazo Compounds* , 3rd ed., Arnold, London 1985.
49. V. O. Lukashevich, E. S. Lisitsyna, *Dokl. Akad. Nauk SSSR* **160** (1965) 129 – 132.
50. BIOS Final Report 988.
51. H. Breithaupt, A. Dammann, K. Aigner, *Cancer Chemoter. Pharmacol.* **9** (1982) 103.
52. BIOS Final Report 986, p. 146.
53. R. F. Muraca, *Treatise Anal. Chem. 1959 Part II* **15** (1976) 251 – 384.
54. *Houben-Weyl,* **2,** 700.
55. E. Bitterlin, *Z. Anal. Chem.* **253** (1971) 120 – 122.
56. BIOS Final Report 1152, p. 91.
57. R. C. Elderfield (ed.): *Heterocyclic Compounds,* **vols. 5 – 8,** J. Wiley & Sons, New York-London 1957 –1967.
58. A. Weissberger (ed.): *The Chemistry of Heterocyclic Compounds,* **vol. 4, 22, and 27,** Interscience, New York-London-Sidney1950 et seq.
59. Hoechst, DE 2 351 595, 1973 (K. Gengnagel, T. Papenfuhs).
60. C. Hamet-Elfakir, C. Caullet, *Bull. Soc. Chim. Fr.* **5** (1986) 687 – 692.
61. P. Rütschi, G. Trümpler, *Helv. Chim. Acta* **36** (1953) 1649 – 1658.
62. Clariant, DE 19 822 316 A1, 1999 (H. Volk, M. Koch, T. Schach).
63. FMC, GB 2 230 261, 1990 (G. Theodoridis).
64. Searle & Co., US 5 466 823, 1995 (M Granetto, J. Miyashiro, J. Talley, P. Collins, K. Khanna Ish, J. Malecha, R. Rogers, S. Bertenshaw, J. Carter, T. Penning, D. Rogier Jr.).
65. A. Kleemann, J. Engel: *Pharmaceutische Wirkstoffe,* Thieme, Stuttgart, 1982, p. 488 – 489.
66. Cassella, DE 901 175, 1943 (W. Zerweck, M. Schubart, F. Fleischhauer).
67. Olin, DE 2 426 994 A1 (D. F. Gavin, J. H. Tobin).
68. BIOS Final Report 1153, p. 20.
69. *Houben-Weyl,* **5/4,** pp. 639 – 647.
70. BIOS Final Report 1145, p. 19.
71. *Houben-Weyl,* **5/4,** 438 – 447,
72. BIOS Final Report 1149, p. 32.
73. L. D. Basnagoudar, *Can. J. Chem.* **45** (1967) 1807.
74. Hoechst France, EP 0 368 697, 1990 (Y. Christidis).
75. B. Cornils, W. A. Herrmann, *Applied Homogenous Catalysis with Organometallic Compounds,* VCH, Weinheim, 1996.
76. R. F. Heck: *Palladium Reagents in Organic Synthesis,* Academic Press, London, 1985.
77. S. Bräse, A. de Meijere *in Metal-Catalyzed Cross-Coupling Reactions,* Wiley-VCH, Weinheim, 1998.
78. H. Brunner, N. Le Cousturier de Courcy, J. P. Genet, *Tetrahedron Lett.* **40** (1999) 4815 – 4818.
79. *Kirk-Othmer*, 3rd ed., **7**, p. 153.
80. Ciba Specialty Chemicals Holding Inc., EP 0 849 244 A1, 1998 (V. Eliu, W. Kanert, P. Baumeister, J Völkel).
81. S. Sengupta, S. K. Sadhukhan, S. Bhttacharyya, J. Guha, *J. Chem. Soc. Perkin Trans. 1* **1998** , 407 – 410.
82. J. Hoerth, U. Schindewolf, W. Zbinden, *Chem. Ing. Tech.* **45** (1973) 641 – 646.
83. J. Luetolf, *Staub Reinhalt. Luft* **31** (1971) 93 – 97.
84. Bundesgesetzblatt, part 1, no. 85, pp. 1357 – 1404 (1969).
85. P. Bersier, L. Valpiana, H. Zubler, *Chem. Ing. Tech.* **43** (1971) 1311.
86. *Gefahrstoffverordnung,* Deutscher Bundes-Verlag, Bonn 1986.
87. Hoechst, *Sicherheitsdatenblätter der Echtfärbesalze* (1978 – 1986).
88. E. J. Fairchild (ed.): *Registry of Toxic Effects of Chemical Substances,* U.S. Dept. of Health (1977).

Dibenzofuran → Furan and Derivatives
Dibenzothiophene → Thiophene
1,2-Dibromoethane → Bromine Compounds

Dicarboxylic Acids, Aliphatic

Adipic Acid, Citric Acid, Ethylenediaminetetraacetic Acid and Related Chelating Agents, Maleic and Fumaric Acids, Malonic Acid and Derivatives, Nitrilotriacetic Acid, Oxalic Acid and Tartaric Acid are separate keywords.

Boy Cornils, Hoechst AG, Frankfurt, Federal Republic of Germany

Peter Lappe, Ruhrchemie AG, Oberhausen, Federal Republic of Germany

1. Introduction 1
2. Saturated Dicarboxylic Acids . . . 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.3. Production 4
2.3.1. Degradative Methods 5
2.3.2. Processes Maintaining the Carbon Structure 5
2.3.3. Syntheses from Smaller Units 7
2.4. Individual Saturated Dicarboxylic Acids 8
2.4.1. Succinic Acid 8
2.4.2. Glutaric Acid 9
2.4.3. Dimethylglutaric Acids 9
2.4.4. Trimethyladipic Acid 9
2.4.5. Pimelic Acid 9
2.4.6. Suberic Acid 10
2.4.7. Azelaic Acid 10
2.4.8. Sebacic Acid 10
2.4.9. 1,12-Dodecanedioic Acid 11
2.4.10. 1,13-Tridecanedioic Acid (Brassylic Acid) 11
2.4.11. C_{19} Dicarboxylic Acids 11
3. Unsaturated Dicarboxylic Acids . 12
3.1. Physical Properties 12
3.2. Chemical Properties 12
3.3. Production 12
3.4. Individual Unsaturated Dicarboxylic Acids 14
3.4.1. Itaconic Acid 14
3.4.2. Dimer Acids 14
4. Quality Specifications and Analysis 15
5. Storage, Transportation, and Handling 15
6. References 15

1. Introduction

Aliphatic ω,ω'-dicarboxylic acids (or diacids) can be described by the following general formula:

$$HOOC-(CH_2)_n-COOH$$

According to IUPAC nomenclature, dicarboxylic acids are named by adding the suffix dioic acid to the name of the hydrocarbon with the same number of carbon atoms, e.g., nonanedioic acid for $n=7$. The older literature often uses another system based on the hydrocarbon for the $(CH_2)_n$ carbon segment and the suffix dicarboxylic acid, e.g., heptanedicarboxylic acid for $n=7$. However, trivial names are commonly used for the saturated linear aliphatic dicarboxylic acids from $n=0$ (oxalic acid) to $n=8$ (sebacic acid) and for the simple unsaturated aliphatic dicarboxylic acids; these names are generally derived from the natural substance in which the acid occurs or from which it was first isolated.

Aliphatic dicarboxylic acids are found in nature both as free acids and as salts. For example, malonic acid is present in small amounts in sugar beet and in the green parts of the wheat plant; oxalic acid occurs in many plants and in some minerals as the calcium salt. However, natural sources are no longer used to recover these acids.

The main industrial process employed for manufacturing dicarboxylic acids is the ring-opening oxidation of cyclic compounds.

Adipic acid is the most important dicarboxylic acid. Oxalic, malonic, suberic, azelaic, sebacic, and 1,12-dodecanedioic acids, as well as maleic and fumaric acids, are also manufactured on an industrial scale.

Dicarboxylic acids are important feedstocks in the manufacture of polyamides or of di- and polyesters. Esters produced by the reaction of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_523

dicarboxylic acids with monofunctional alcohols serve as plasticizers or lubricants. In addition, dicarboxylic acids are used as intermediates in many organic syntheses.

2. Saturated Dicarboxylic Acids

The most important saturated aliphatic dicarboxylic acids are treated under separate keywords (see → Adipic Acid, → Malonic Acid and Derivatives, → Oxalic Acid).

2.1. Physical Properties

Dicarboxylic acids are colorless, odorless crystalline substances at room temperature. Table 1 lists the major physical properties of some saturated aliphatic dicarboxylic acids.

The lower dicarboxylic acids are stronger acids than the corresponding monocarboxylic ones. The first dissociation constant is considerably greater than the second. Density and dissociation constants decrease steadily with increasing chain length. By contrast, melting point and water solubility alternate: Dicarboxylic acids with an even number of carbon atoms have higher melting points than the next higher odd-numbered dicarboxylic acid. In the $n = 0-8$ range, dicarboxylic acids with an even number of carbon atoms are slightly soluble in water, while the next higher homologues with an odd number of carbon atoms are more readily soluble. As chain length increases, the influence of the hydrophilic carboxyl groups diminishes; from $n = 5$ (pimelic acid) onward, solubility in water decreases rapidly. The alternating solubility of dicarboxylic acids can be exploited to separate acid mixtures [1], [2]. Most dicarboxylic acids dissolve easily in lower alcohols; at room temperature, the lower dicarboxylic acids are practically insoluble in benzene and other aromatic solvents.

2.2. Chemical Properties

The chemical behavior of dicarboxylic acids is determined principally by the two carboxyl groups. The neighboring methylene groups are activated generally to only a minor degree; malonic acid derivatives (esters and nitriles) are an exception (→ Malonic Acid and Derivatives).

Thermal decomposition of dicarboxylic acids gives different products depending on the chain length. Acids with an even number of carbon atoms require higher decarboxylation temperatures than the next higher odd-numbered homologues; lower dicarboxylic acids decompose more easily than higher ones. To avoid undesired decomposition reactions, aliphatic dicarboxylic acids should only be distilled in vacuum. When heated above 190 °C, oxalic acid decomposes to carbon monoxide, carbon dioxide, and water. Malonic acid is decarboxylated to acetic acid at temperatures above 150 °C:

$$HOOC-CH_2-COOH \longrightarrow CH_3COOH + CO_2$$

When malonic acid is heated in the presence of P_2O_5 at ca. 150 °C, small amounts of carbon suboxide (C_3O_2) are also formed. Succinic and glutaric acids are converted into cyclic anhydrides on heating:

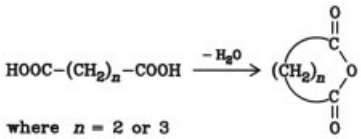

When the ammonium salt of succinic acid is distilled rapidly, succinimide is formed, with the release of water and ammonia.

Higher dicarboxylic acids from $n = 4$ (adipic acid) to $n = 6$ (suberic acid) split off carbon dioxide and water to form cyclic ketones:

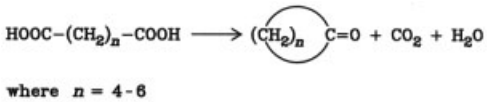

The decomposition of still higher dicarboxylic acids leads to complex mixtures. With the exception of oxalic acid, dicarboxylic acids are resistant to oxidation. Oxalic acid is used as a reducing agent for both commercial and analytical purposes. Dicarboxylic acids react with dialcohols to form polyesters and with diamines to form polyamides. They also serve as starting materials for the production of the corresponding diamines. Reaction with monoalcohols yields esters. All of these reactions are commercially important. Several reactions with malonic and glutaric acids are of interest in organic syntheses: the Knoevenagel condensation,

Table 1. Physical properties of saturated dicarboxylic acids

IUPAC name	Common name	CAS registry number	Formula	M_r	*mp*, °C	*bp* at 13.3 kPa, °C	ϱ at 25 °C, g/cm^3	Solubility in H_2O at 20 °C, wt %	Decarboxylation temperature, °C	Ionization constants K_1	K_2
Ethanedioic acid	oxalic acid	[*144-62-7*]	HOOC–COOH	90.03	189.5	(sublimes)	1.653	8.0	166–180	5.36×10^{-2}	5.42×10^{-5}
Propanedioic acid	malonic acid	[*141-82-2*]	HOOC–CH_2–COOH	104.06	135		1.619 (16 °C)	73.5	140–160	1.42×10^{-3}	2.01×10^{-6}
Butanedioic acid	succinic acid	[*110-15-6*]	HOOC–$(CH_2)_2$–COOH	118.08	188	235 [a]	1.572	5.8	290–310	6.21×10^{-5}	2.31×10^{-6}
Pentanedioic acid	glutaric acid	[*110-94-1*]	HOOC–$(CH_2)_3$–COOH	132.11	99	200 (2.7 kPa)	1.424	63.9	280–290	4.58×10^{-5}	3.89×10^{-6}
2,2-Dimethylpentanedioic acid	2,2-dimethylglutaric acid	[*681-57-2*]	**1** [d]	160.17	85			[b]		5.25×10^{-5}	3.8×10^{-6}
Hexanedioic acid	adipic acid	[*124-04-9*]	HOOC–$(CH_2)_4$–COOH	146.14	153	265	1.360	1.6	300–320	3.85×10^{-5}	3.89×10^{-6}
2,4,4-Trimethylhexanedioic acid	2,4,4-trimethyladipic acid	[*3937-59-5*]	**2**	188.22	68			1.075 [c]			
Heptanedioic acid	pimelic acid	[*111-16-0*]	HOOC–$(CH_2)_5$–COOH	160.17	106	272	1.329 (15 °C)	5.0	290–310	3.19×10^{-5}	3.74×10^{-6}
Octanedioic acid	suberic acid	[*505-48-6*]	HOOC–$(CH_2)_6$–COOH	174.19	144	279	1.266	0.16	340–360	3.05×10^{-5}	3.85×10^{-6}
Nonanedioic acid	azelaic acid	[*123-99-9*]	HOOC–$(CH_2)_7$–COOH	188.22	108	287	1.225	0.24	320–340	2.88×10^{-5}	3.86×10^{-6}
Decanedioic acid	sebacic acid	[*111-20-6*]	HOOC–$(CH_2)_8$–COOH	202.25	134.5	295	1.207	0.10	350–370	3.1×10^{-5}	3.6×10^{-6}
Undecanedioic acid		[*1852-04-6*]	HOOC–$(CH_2)_9$–COOH	216.27	110			0.014			
Dodecanedioic acid		[*693-23-2*]	HOOC–$(CH_2)_{10}$–COOH	230.30	131	254 (2.0 kPa)		0.004			
Tridecanedioic acid	brassylic acid	[*505-52-2*]	HOOC–$(CH_2)_{11}$–COOH	244.33	114		1.150 (18 °C)	0.0025			
Tetradecanedioic acid		[*821-38-5*]	HOOC–$(CH_2)_{12}$–COOH	258.35	129						

[a] Forms the anhydride (101.3 kPa).
[b] Readily soluble in water.
[c] Mixture of 40 % 2,2,4- and 60 % 2,4,4-trimethyladipic acid.
[d] **1** = HOOC–$C(CH_3)_2$–$(CH_2)_2$–COOH **2** = HOOC–$CH(CH_3)$–CH_2–$C(CH_3)_2$–CH_2–COOH

Michael addition, and malonic ester synthesis (→ Malonic Acid and Derivatives) [3], [4].

Succinic acid ester reacts with aldehydes or ketones in the presence of sodium ethoxide or potassium *tert*-butoxide to form alkylidenesuccinic acid monoesters (Stobbe condensation). These can subsequently be converted into monocarboxylic acids by hydrolysis, decarboxylation, and hydrogenation [5]:

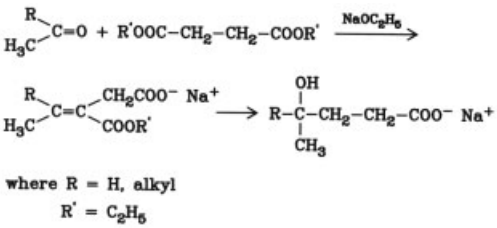

Cyclic ketones are obtained from C_6–C_8 dicarboxylic acid esters and sodium methoxide (Dieckmann reaction) [6]. Esters of adipic, pimelic, and suberic acids can be converted in good yields; esters of higher dicarboxylic acids cannot be cyclized by this method.

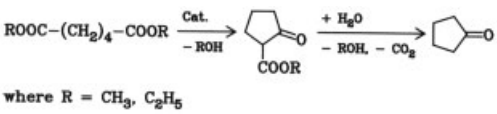

Acyloin condensation with metallic sodium gives cyclic acyloins; this method is particularly suitable for synthesis of large rings [7]:

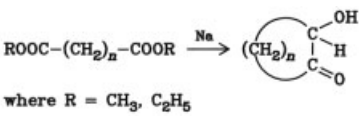

Detailed summaries of reactions with dicarboxylic acids can be found in [8].

2.3. Production

A number of straight-chain aliphatic dicarboxylic acids and their derivatives occur in nature. However, isolation from natural substances has no commercial significance. Although many syntheses for the production of aliphatic dicarboxylic acids are known, only a few have found industrial application. This is due partly to the shortage of raw materials.

The most important processes for the manufacture of saturated aliphatic dicarboxylic acids are the following:

1) Oxidative cleavage of cyclic compounds (e.g., adipic acid from cyclohexane, 1,12-dodecanedioic acid from 1,5,9-cyclododecatriene)
2) Oxidative cleavage of unsaturated monocarboxylic acids (e.g., azelaic acid from oleic acid)
3) Alkaline cleavage of substituted monocarboxylic acids (e.g., sebacic acid from ricinoleic acid)
4) Hydrogenation of unsaturated dicarboxylic acids (e.g., succinic acid from maleic acid)
5) Oxidation of ω,ω'-diols (e.g., pimelic acid from 1,7-heptanediol)
6) Carbonylation reactions (e.g., suberic acid from 1,6-hexanediol)

Some special syntheses are also of interest. The following sections treat the most important manufacturing processes, which can be subdivided into degradative methods, processes in which the carbon structure is maintained, and synthetic methods starting from smaller units.

2.3.1. Degradative Methods

Ozonolysis of Oleic Acid. Ozonolysis of oleic acid [*112-80-1*] followed by oxidative cleavage gives pelargonic acid [*112-05-0*] and azelaic acid [9]:

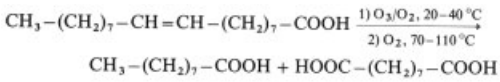

Figure 1 shows a commercial process for the production of azelaic acid from oleic acid.

Oleic acid is cleaved by ozonolysis (O_3 concentration in the air: 1.0 vol %) at 20 – 40 °C in pelargonic acid and water. The alkene residence time is about 10 min. The ozonide is then cleaved with oxygen at 70 – 110 °C. Pelargonic and azelaic acids are separated from higher boiling compounds by subsequent distillation. Azelaic acid is subjected to extraction to remove monocarboxylic acids; distillation of the extractant finally yields pure acid.

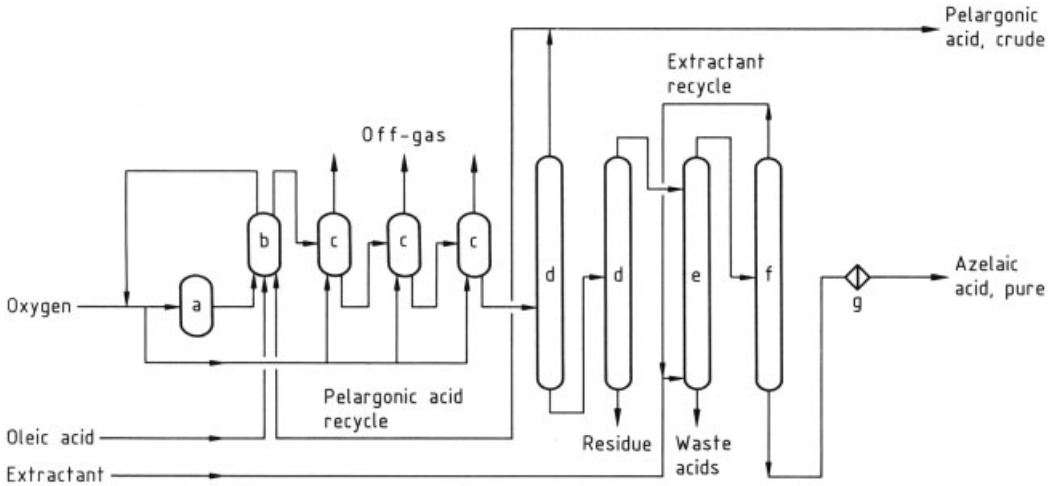

Figure 1. Manufacture of azelaic acid by ozonolysis of oleic acid
a) Ozone generator; b) Ozone absorber; c) Reactor; d) Distillation column; e) Extraction column; f) Distillation of the extractant; g) Flaking

Cleavage of Ricinoleic Acid. Alkaline cleavage of ricinoleic acid [*141-22-0*] (12-hydroxy-9-octadecenoic acid) under pressure and at high temperature leads to the formation of sodium sebacate and 2-octanol [10]:

$$CH_3-(CH_2)_5-CH(OH)-CH_2-CH=CH-(CH_2)_7-COOH \xrightarrow{NaOH} NaOOC-(CH_2)_8-COONa + CH_3-(CH_2)_5-CH(OH)-CH_3$$

In industry, castor oil [*8001-79-4*], which contains about 87 % ricinoleic acid, is normally used instead of the pure acid.

Oxidation with N_2O_4. Oxidative degradation of monocarboxylic acids generally produces dicarboxylic acid mixtures; the composition of the reaction products shifts toward the higher dicarboxylic acids as the chain length of the monocarboxylic acids increases. Oxidation of stearic acid [*57-11-4*] with N_2O_4 yields a mixture consisting mainly of sebacic and caprylic [*124-07-2*] acids. In the same way, palmitic acid [*57-10-3*] can be oxidized with nitric acid – N_2O_4 to form suberic acid [11].

Commercial production of adipic acid from cyclohexanol – cyclohexanone yields two major byproducts, succinic and glutaric acids, which can be separated easily (→ Adipic Acid, Chap. 5.).

Oxidation of Hydrocarbons. Oxidative degradation of hydrocarbons is also a common manufacturing process. The best-known example is the oxidation of benzene to maleic acid by means of vanadium pentoxide catalysts (→ Maleic and Fumaric Acids).

2.3.2. Processes Maintaining the Carbon Structure

Cleavage of Cyclic Compounds. Many processes for the manufacture of dicarboxylic acids by oxidative cleavage of cyclic compounds are commercially significant; however, the oxidation of cyclohexane via cyclohexanol – cyclohexanone is the most important (→ Adipic Acid). Similar processes are employed to convert cyclopentanol – cyclopentanone to glutaric acid, cycloheptanone to pimelic acid, and cyclododecanol – cyclododecanone to 1,12-dodecanedioic acid [12]. Figure 2 shows a process for the manufacture of 1,12-dodecanedioic acid from cyclododecanol – cyclododecanone [13].

The oxidation is carried out in a stirred reactor (b) fed continuously with nitric acid and cyclododecanol – cyclododecanone; ammonium vanadate is used as a catalyst. The nitric oxides formed during oxidation are recycled via the

condenser (d) and cooler (e). The reaction slurry passes to the postreaction stage and then to the crystallizer (g), in which most of the acid crystallizes. The solids are filtered off and the mother liquor is recycled to the reactor. In industrial-scale processes, selectivities toward the acid of about 90 % are achieved.

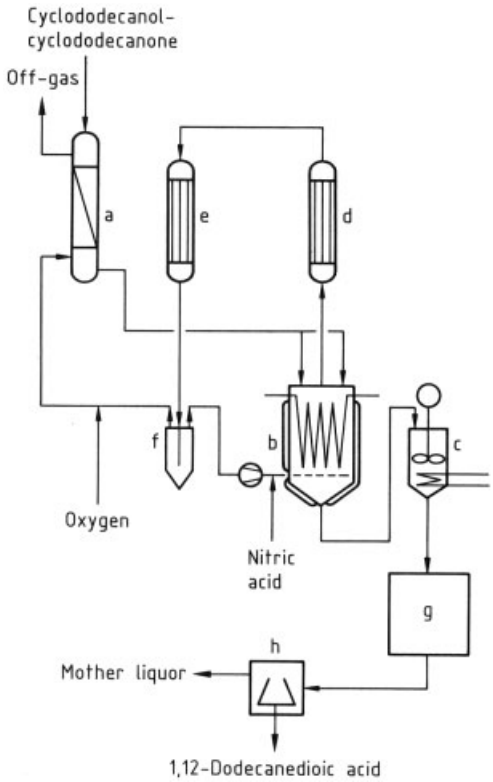

Figure 2. Manufacture of 1,12-dodecanedioic acid from cyclododecanol – cyclododecanone
a) Scrubber; b) Reactor; c) Downstream reactor; d) Condenser; e) Cooler; f) Separator; g) Crystallizer; h) Filter

Cyclic ethers can also be used as starting materials. Thus, pimelic acid is obtained from potassium tetrahydrofurylpropionate [14] or from hydroxycyclohexanoic acid [15]. Hydrolysis of dihydropyran produces 5-hydroxypentanal, which is converted to glutaric acid by subsequent oxidation with nitric acid [16].

Ozonolysis of Cyclic Olefins. Cyclic olefins can be converted to dicarboxylic acids by ozonolysis and subsequent oxidative cleavage. For example, 1,12-dodecanedioic acid can be obtained by ozonolysis of cyclododecene (see Section 2.4.9) [17].

Oxidation of Bifunctional Compounds. Dicarboxylic acids can be produced by oxidation of bifunctional compounds with HNO_3 in the presence of ammonium vanadate, with N_2O_4, or with oxygen in the presence of palladium on carbon. Diols are preferred as bifunctional starting materials. Well-known examples of this process are the syntheses of pimelic acid from 1,7-heptanediol and of succinic acid from 1,4-butanediol [18].

Nitrile Hydrolysis. Saponification of dinitriles also yields dicarboxylic acids. Thus, glutaric acid can be produced from glutarodinitrile, which is obtained by the reaction of 1,3-dihalopropane with sodium cyanide. Saponification of the nitrile group can take place concurrently with oxidation of a carbonyl group; e.g., 4-cyano-2,2-dimethylbutanal, obtained by the addition of isobutanal to acrylonitrile, gives 2,2-dimethylglutaric acid [19].

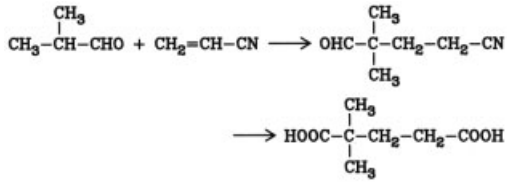

Hydrogenation. Hydrogenation of unsaturated dicarboxylic acids or their anhydrides produces good yields of the corresponding saturated compounds. Succinic acid is obtained by this method from maleic acid or maleic anhydride [20].

Fermentation. Numerous alkane-based fermentation processes have been described for the manufacture of dicarboxylic acids [21]. However, these biotechnical processes have not yet become standard commercial practice.

2.3.3. Syntheses from Smaller Units

The main addition reactions leading to dicarboxylic acids are variants of carbonylation. Diolefins, dialcohols, and unsaturated monocarboxylic acids are used as starting materials

[22]. Reppe carbonylation of 1,6-hexanediol produces suberic acid; C_{19} dicarboxylic acids are obtained from oleic acid [23].

The dimerization of monomethyl adipate to sebacic acid is an electrochemical process which has achieved commercial significance. Figure 3 illustrates this process [24].
The reaction takes place in three stages:
Esterification:

$$\mathrm{HOOC{-}(CH_2)_4{-}COOH + CH_3OH \longrightarrow HOOC{-}(CH_2)_4{-}COOCH_3 + H_2O}$$

Electrolysis:

$$\mathrm{{}^{-}OOC{-}(CH_2)_4{-}COOCH_3 - e^- \longrightarrow 1/2\,CH_3OOC{-}(CH_2)_8{-}COOCH_3 + CO_2}$$

$$\mathrm{H^+ + e^- \longrightarrow 1/2\,H_2}$$

Hydrolysis:

$$\mathrm{CH_3OOC{-}(CH_2)_8{-}COOCH_3 + 2\,H_2O \longrightarrow HOOC{-}(CH_2)_8{-}COOH + 2\,CH_3OH}$$

In the first stage, adipic acid reacts with methanol at 80 °C to form monomethyl adipate. Ion exchangers containing sulfonic acid groups are used as catalysts and also prevent the formation of byproducts such as cyclopentanone. The monomethyl adipate is separated by distillation.

In the second stage, the potassium salt of monomethyl adipate is dimerized electrolytically either continuously or batchwise. The electrolyzer is equipped with bipolar electrodes. Electrolysis takes place at 50 – 60 °C. Aqueous methanol is used as solvent, the H_2O concentration being between 0.15 and 0.30 %. The resulting dimethyl sebacate solution is distilled, and unreacted potassium methyl adipate is returned to the electrolysis process. Dimethyl sebacate is purified by distillation.

In the third stage, dimethyl sebacate is hydrolyzed at 160 – 180 °C and a pressure of about 0.9 MPa (9 bar). Methanol is removed, and the crude sebacic acid is treated with activated carbon and then dried.

The Wurtz synthesis can also be used [25]:

$$\mathrm{2\,Cl{-}CH_2{-}COOR + 2\,Na \longrightarrow ROOC{-}(CH_2)_2{-}COOR + 2\,NaCl \xrightarrow{H_2O} HOOC{-}(CH_2)_2{-}COOH}$$

The reactivity of the methylene group in malonic ester is exploited in many dicarboxylic acid syntheses (malonic ester synthesis; → Malonic Acid and Derivatives).

The Stetter dicarboxylic acid synthesis is another important process [26]:

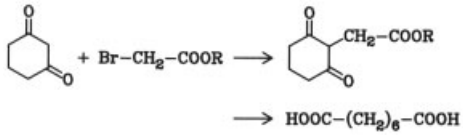

Long-chain dicarboxylic acids can be prepared in the following manner [27]:

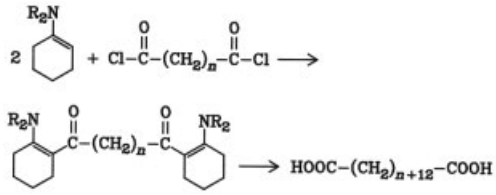

Many processes used to manufacture monocarboxylic acids are also suitable for the synthesis of dicarboxylic acids. This field has been reviewed extensively [28] (→ Carboxylic Acids, Aliphatic).

2.4. Individual Saturated Dicarboxylic Acids

Dicarboxylic acids are used mainly as intermediates in the manufacture of esters and polyamides. Esters derived from monofunctional alcohols serve as plasticizers or lubricants. Polyesters are obtained by reaction with dialcohols. In addition, dicarboxylic acids are employed in the manufacture of hydraulic fluids, agricultural chemicals, pharmaceuticals, dyes, complexing agents for heavy-metal salts, and lubricant additives (as metal salts).

2.4.1. Succinic Acid

Succinic acid is found in amber, in numerous plants (e.g., algae, lichens, rhubarb, and tomatoes), and in many lignites.

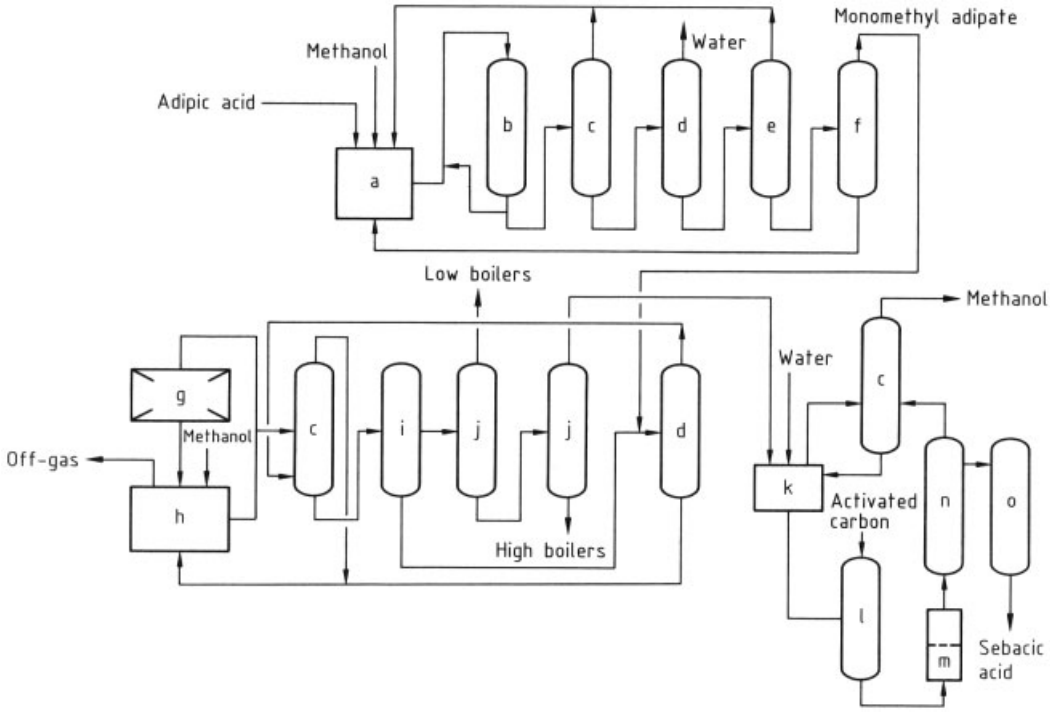

Figure 3. Manufacture of sebacic acid by electrochemical dimerization of monomethyl adipate
a) Mixing tank; b) Reaction column; c) Methanol stripper; d) Water stripper; e) Dimethyl adipate stripper; f) Adipic acid cutting column; g) Electrolyzer; h) Electrolyte tank; i) Decanter; J) Distillation column; k) Reactor; l) Mixing tank; m) Filter; n) Dehydrator; o) Prill tower

Production. A large number of syntheses are used to manufacture succinic acid. Hydrogenation of maleic acid, maleic anhydride, or fumaric acid produces good yields of succinic acid; the standard catalysts are Raney nickel [20], Cu, NiO, or CuZnCr [29], $Pd-Al_2O_3$ [30], $Pd-CaCO_3$ [31], or Ni – diatomite [32]. 1,4-Butanediol can be oxidized to succinic acid in several ways: (1) with O_2 in an aqueous solution of an alkaline-earth hydroxide at 90 – 110 °C in the presence of Pd – C; (2) by ozonolysis in aqueous acetic acid; or (3) by reaction with N_2O_4 at low temperature [18]. Succinic acid or its esters are also obtained by Reppe carbonylation of ethylene glycol, catalyzed with $RhCl_3$ – pentachlorothiophenol [33]; Pd-catalyzed methoxycarbonylation of ethylene [34]; and carbonylation of acetylene, acrylic acid, dioxane, or β-propiolactone [35], [36] (→ Carbonylation).

Acid mixtures containing succinic acid are obtained in various oxidation processes. Examples include the manufacture of adipic acid [2], [37–39] (→ Adipic Acid, Chap. 4.1.); the oxidation of enanthic acid [40]; and the ozonolysis of palmitic acid [41].

Succinic acid can also be obtained by phase-transfer-catalyzed reaction of 2-haloacetates [42], electrolytic dimerization of bromoacetic acid or ester [43], oxidation of 3-cyanopropanal [44], and fermentation of *n*-alkanes [45].

Uses. Succinic acid is used as a starting material in the manufacture of alkyd resins, dyes, pharmaceuticals, and pesticides. Reaction with glycols gives polyesters; esters formed by reaction with monoalcohols are important plasticizers and lubricants.

2.4.2. Glutaric Acid

Glutaric acid occurs in washings from fleece and, together with malonic acid, in the juice of unripened sugar beet.

Production. Glutaric acid is obtained from cyclopentane by oxidation with oxygen and cobalt (III) catalysts [46], [47] or by ozonolysis [48]; and from cyclopentanol – cyclopentanone by oxidation with oxygen and $Co(CH_3CO_2)_2$, with potassium peroxide in benzene, or with N_2O_4 or nitric acid [12], [49–51]. Like succinic acid, glutaric acid is formed as a byproduct during oxidation of cyclohexanol – cyclohexanone (→ Adipic Acid).

Other production methods include reaction of malonic ester with acrylic acid ester [52], [53], carbonylation of γ-butyrolactone [22], oxidation of 1,5-pentanediol with N_2O_4 [18], and oxidative cleavage of γ-caprolactone [54].

Uses. The applications of glutaric acid, e.g., as an intermediate, are limited. Its use as a starting material in the manufacture of maleic acid has no commercial importance.

2.4.3. Dimethylglutaric Acids

2,2-Dimethylglutaric acid is manufactured from dimethyl-γ-butyrolactone by carbonylation using $HF-SbF_5$ as a catalyst or by reaction with formic acid in stronger acids such as $H_2SO_4-SO_3$ [55], [56]. 4-Cyano-2,2-dimethylbutanal, which is obtained by addition of isobutanal to acrylonitrile, can be converted to the acid by oxidation of the formyl group and subsequent hydrolysis of the nitrile group [19], [57].

2,2-Dimethylglutaric acid is used in the manufacture of diglycidyl esters (for coating materials) [58], pyrethroids (for insecticides and acaricides) [59], and antibiotics [60].

3,3-Dimethylglutaric acid [*4839-46-7*] is manufactured from isophorone by oxidation with H_2O_2 in the presence of concentrated sulfuric acid or by ozonolysis in methanolic solution and subsequent oxidation with H_2O_2 [61]. This acid is used in the manufacture of pesticides and lubricating oil additives.

2.4.4. Trimethyladipic Acid

Commercial trimethyladipic acid is a mixture of ca. 40 % 2,2,4-trimethyladipic acid and 60 % 2,4,4-trimethyladipic acid.

Production. Trimethyladipic acid is manufactured by oxidative cleavage of 3,3,5-trimethylcyclohexanol [*116-02-9*] (produced from acetone) with 65 % nitric acid at 50 °C [62]. To separate the short-chain dicarboxylic acids, the mixture is heated to 180 – 250 °C and the cyclic anhydrides formed are distilled off [63].

Uses. Trimethyladipic acid is used in the production of synthetic lubricating oils [64], polyesters [65], and polyamides [65], and in the modification of terephthalic acid esters [66].

2.4.5. Pimelic Acid

Pimelic acid is an oxidation product of fats.

Production. Pimelic acid can be manufactured with good selectivity by oxidation of cycloheptanone [*502-42-1*] (suberone) with N_2O_4 [12]. It is also obtained in a mixture with other dicarboxylic acids by oxidative cleavage of palmitic acid [41]. Other manufacturing processes include oxidation of 1,7-heptanediol [67], carbonylation of ε-caprolactone [22], and acid cleavage of tetrahydrosalicylic acid with potassium hydroxide at 300 °C under pressure [15].

Uses. Pimelic acid has slight significance as a starting material in the manufacture of polyesters and polyamides.

2.4.6. Suberic Acid

Suberic acid is formed from the action of nitric acid on cork.

Production. Suberic acid is manufactured by oxidation of cyclooctene with ozone – oxygen [48], [68] or with ozone – H_2O_2 [69]. The acid is formed together with other dicarboxylic acids during ozonolysis of palmitic acid [41] as well as during cleavage of ricinoleic acid with nitric acid [70]. Other manufacturing processes include oxidation of cyclooctanol – cyclooctanone with N_2O_4 or HNO_3 [12], [71], carbonylation of 1,6-hexanediol [72], and oxidative cleavage of 2-(cyclohexanon-2-yl)acetic acid ethyl ester [26].

Uses. Suberic acid has been used in the manufacture of mono- and diesters as well as polyamides. Nylon 6,8 is obtained by reaction of suberic acid with hexamethylenediamine,

and nylon 8,8 by reaction with octamethylenediamine. Polyamides of suberic acid with diamines such as 1,3-bis(aminomethyl)benzene, 1,4(bisaminomethyl)cyclohexane, and bis(4-aminocyclohexyl)methane are also of commercial interest. Esters of suberic acid with mono- and bifunctional alcohols are used as lubricants.

2.4.7. Azelaic Acid

Azelaic acid occurs in many natural substances containing long-chain fatty acids.

Production. Azelaic acid is obtained by oxidative cleavage of oleic acid with oxidants such as RuO_4 [73]; $Cl_2 - RuO_2$ or $Cl_2 - RuCl_2$ [74], [75]; $KMnO_4$ [76]; $NaOCl - RuO_4 - OsO_4$ [77]; and HNO_3 [78]. The industrially most important process is the ozonolysis of oleic acid (see Section 2.3.1) [9], [79], [80].

Other means of synthesizing azelaic acid include carbonylation of 1,5-cyclooctadiene [81], oxidation of 1,9-nonanedial with oxygen [82], oxidative cleavage of 2-cyanoethylcyclohexanone [83], and fermentation of pelargonic acid [84]. A mixture of azelaic and other dicarboxylic acids is obtained during ozonolysis of palmitic acid [41].

Uses. Monoesters of azelaic acid with 2-ethylhexanol are used as plasticizers. Mono- and diesters with other alcohols act as hydraulic fluids and lubricating oils; their metal salts are recommended as lubricating oil additives. Reaction with hexamethylenediamine leads to nylon 6,9, which is used as extruded film for food packaging, as a coating for wire, and in the electronics and automobile industries. Unsaturated polyesters are employed as resins, laminates, and adhesives.

2.4.8. Sebacic Acid

Production. The most important processes for manufacturing sebacic acid are alkaline cleavage of ricinoleic acid (see Section 2.3.1) [85–88] and electrolytic dimerization of monomethyl adipate (see Section 2.3.3) [89–91]. 2-Octanol is formed as a byproduct during ricinoleic acid cleavage. Other methods used to manufacture sebacic acid are oxidation of stearic acid by N_2O_4 [11], oxidation of 1,10-decanediol [18], and various fermentation processes [21], [92]. A mixture of sebacic acid and other dicarboxylic acids is formed during ozonolysis of palmitic acid [41]. Processes for purifying sebacic acid are described in [93–95].

The C_{10} dicarboxylic acid mixture obtained by dimerization of butadiene and subsequent reaction with CO_2 is called isosebacic acid; it consists of ca. 75 % 2-ethylsuberic acid, 15 % diethyladipic acid, and 10 % sebacic acid. Because of the varying composition of different production batches, this mixture has not been able to gain a foothold in the market.

Uses. The polyamide nylon 6,10 obtained by reaction of sebacic acid with hexamethylenediamine no longer has great industrial significance. The sebacates of various oxo and straight-chain alcohols are important plasticizers. Their main characteristics are high migration resistance and good low-temperature resistance. The esters are also used as components of lubricating oils and as diluents; because of their low toxicity they are important components of packaging films. Sebacic acid-based alkyd resins are characterized by marked flexibility.

2.4.9. 1,12-Dodecanedioic Acid

Over the past few years, 1,12-dodecanedioic acid has achieved industrial importance.

Production. The starting compound for industrial-scale production of 1,12-dodecanedioic acid is 1,5,9-cyclododecatriene (CDT), which is obtained by trimerization of butadiene (→ Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene). Cyclododecatriene can react to form the acid by two different processes. In a three-stage reaction sequence, 1,5,9-cyclododecatriene is first hydrogenated to cyclododecane on nickel catalysts; cyclododecane is then oxidized with oxygen or air to a cyclododecanol – cyclododecanone mixture; and this mixture is finally oxidized with nitric acid to 1,12-dodecanedioic acid (→ Cyclododecanol, Cyclododecanone, and Laurolactam) [96–100].

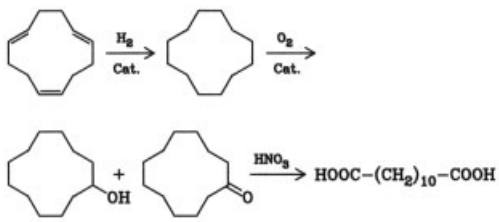

The second route consists of partial hydrogenation of 1,5,9-cyclododecatriene to cyclododecene and subsequent oxidative ozonolytic cleavage to the acid [101–103]; ozonolysis of cyclododecanol has also been described [17].

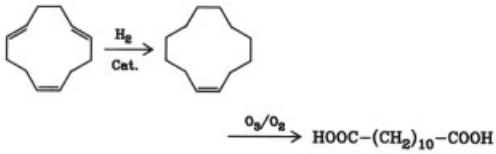

Other manufacturing processes such as the fermentation of *n*-dodecane [21], [104], [105] and the oxidation of analogous mono- and diformyl compounds [106] have no industrial importance. Processes for purifying the acid have been described [107–110].

Uses. 1,12-Dodecanedioic acid is used mainly in manufacturing polyamides and polyesters. Reaction with hexamethylenediamine gives nylon 6,12; reaction with *trans,trans*-bis-(4-aminocyclohexyl)methane yields the polyamide known as Qiana. 1,12-Dodecanedioic acid is also used for the manufacture of lubricating oils and plasticizers.

2.4.10. 1,13-Tridecanedioic Acid (Brassylic Acid)

Production. The most important raw material for the production of brassylic acid is erucic acid [*112-86-7*] (*cis*-13-docosenoic acid), which occurs in large quantities in the seed oil of rape, mustard, wallflower, and cress. Yields of 82 – 92 % are obtained by ozonolysis of erucic acid in acetic acid and subsequent oxidation with oxygen [111], [112]. The byproduct, pelargonic acid, can be separated easily. Oxidative cleavage of erucic acid by reaction with nitric acid is also possible [78].

$$CH_3-(CH_2)_7-CH{=}CH-(CH_2)_{11}-COOH \longrightarrow$$
$$CH_3-(CH_2)_7-COOH + HOOC-(CH_2)_{11}-COOH$$

Fermentation of *n*-tridecane has been investigated over the past few years, especially in Japan [113–117]; purification of the resulting acid is described in [118–121].

Uses. Brassylic acid is used in the manufacture of polyamides (nylon 13,13) and esters which are employed as low-temperature plasticizers for poly(vinyl chloride) (PVC) and as lubricant components. It is also a starting material for synthetic musk.

2.4.11. C_{19} Dicarboxylic Acids

Production. The composition of C_{19} dicarboxylic acid mixtures depends on the manufacturing process. Three processes, based on oleic acid or oleic acid esters, are used industrially: (1) Reppe carbonylation catalyzed by $Ni(CO)_4$ or metal complexes such as $PdCl_2$ – triphenylphosphine; (2) Koch reaction in concentrated sulfuric acid at 10 – 20 °C [23], [122], [123] or with HF catalysis at 30 °C [124]; and (3) hydroformylation. Hydroformylation gives a mixture of isomeric formylstearates, which are subsequently oxidized with air or oxygen. Oxidation takes place in an aqueous emulsion at 20 – 25 °C in the presence of calcium acetate or manganese naphthenate catalyst [125], [126]. Potassium permanganate [127–129] or potassium dichromate [128], [129] can also be used as the oxidizing agent.

Uses. Esters of C_{19} dicarboxylic acids are used as plasticizers for PVC. The esterification rate of the terminal carboxyl group is considerably higher than that of the central group. This allows selective synthesis of mixed esters.

The C_{19} dicarboxylic acids are also used as starting materials in the production of polyamides, epoxy resins, unsaturated polyester resins, lubricants, and adhesives.

3. Unsaturated Dicarboxylic Acids

The most important derivative of unsaturated dicarboxylic acids is maleic anhydride (→ Maleic and Fumaric Acids). Unsaturated C_{36} dicarboxylic acids containing cyclic structures, which are known as dimer acids, also have some industrial significance.

3.1. Physical Properties

Table 2 lists the most important physical properties of some unsaturated aliphatic dicarboxylic

acids. The lower members of the series are colorless, crystalline substances at room temperature; the dimer acids, which are commercially available as isomeric mixtures, are viscous liquids at 25 °C. Melting point, solubility in water, and dissociation constants of the lower homologues are influenced by configuration. For example, the melting points of fumaric and mesaconic acid (trans) are considerably higher than those of the cis isomers maleic and citraconic acid, respectively; the trans isomers are also much less water soluble and less acidic. An indication of the higher stability of the trans form is the higher heat of combustion of maleic acid compared with fumaric acid.

3.2. Chemical Properties

The chemical behavior of unsaturated aliphatic dicarboxylic acids is determined primarily by the two carboxyl groups (see Section 2.2) and the olefinic double bond. Reactions of the carboxyl groups can also be influenced by the olefinic bond. For example, when maleic or citraconic acid is heated above 100 °C, water is split off, and maleic or citraconic anhydride is obtained (→ Maleic and Fumaric Acids). Fumaric acid, however, forms no anhydride; above 230 °C, decomposition occurs and maleic anhydride, water, and an appreciable amount of residue are formed.

Addition of halogen to the carbon – carbon double bond yields dihalodicarboxylic acids; reaction with ozone gives formylcarboxylic acids; and hydroxydicarboxylic acids are formed by addition of water. Catalytic hydrogenation leads to saturated dicarboxylic acids; the cis isomers generally react much more quickly than the trans isomers.

The Diels – Alder reaction of maleic anhydride with conjugated dienes is used both industrially and in preparative organic chemistry [130]; for example, tetrahydrophthalic anhydride is formed by reaction with butadiene:

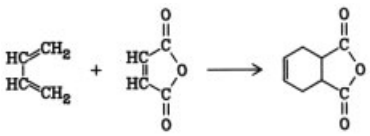

Comprehensive information on reactions with unsaturated dicarboxylic acids can be found in [131].

3.3. Production

Only a few processes are used industrially for the production of unsaturated dicarboxylic acids:

1) Oxidation of hydrocarbons (maleic anhydride from benzene and C_4 hydrocarbons such as *n*-butane and *n*-butene; see → Maleic and Fumaric Acids)
2) Diels – Alder reaction of unsaturated acids (dimer acids from oleic or linoleic acid; see Section 3.4.2)
3) Fermentation (itaconic acid)

3.4. Individual Unsaturated Dicarboxylic Acids

Unsaturated dicarboxylic acids are used mainly to manufacture unsaturated polyester resins, copolymers, or polyamides, and as intermediates in the synthesis of herbicides, insecticides, fungicides, surfactants, lubricants, and plasticizers.

Maleic anhydride is also employed as a starting material for the manufacture of DL-tartaric acid, DL-malic acid, glyoxylic acid, and tetrahydrophthalic anhydride, as well as γ-butyrolactone, 1,4-butanediol, and tetrahydrofuran.

3.4.1. Itaconic Acid

Itaconic acid (2-methylenebutanedioic acid) is soluble in water; moderately soluble in chloroform, benzene, and ligroin; and slightly soluble in ether. When distilled at normal pressure, itaconic acid or itaconic anhydride yields citraconic anhydride.

Production. Itaconic acid is produced by fermentation [132–138]. A mixture of itaconic acid, citraconic acid, and citraconic anhydride is obtained by reaction of succinic anhydride with formaldehyde at 200 – 500 °C in the presence of alkali or alkaline-earth hydroxides [139]; $SiO_2 - Al_2O_3$ or $SiO_2 - MgO$ can also be used

Table 2. Physical properties of unsaturated dicarboxylic acids

IUPAC name	Common name	CAS registry number	Formula	M_r	*mp*, °C	ϱ at 20 °C, g/cm³	Solubility in H_2O, wt %	Ionization constants K_1	K_2
cis-Butenedioic acid	maleic acid	[*110-16-7*]	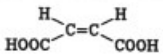	116.07	143.5	1.590	78.9 (25 °C)	1.14×10^{-2}	5.95×10^{-7}
trans-Butenedioic acid	fumaric acid	[*110-17-8*]	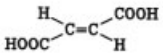	116.07	296–298	1.635 (25 °C)	0.428 (15.5 °C)	9.57×10^{-4}	4.13×10^{-5}
2-Methyl-*cis*-butenedioic acid	citraconic acid	[*498-23-7*]	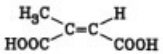	130.10	93	1.617		1.10×10^{-3}	1.0×10^{-6}
2-Methyl-*trans*-butenedioic acid	mesaconic acid	[*498-24-8*]	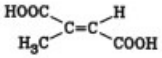	130.10	204.5	1.466	2.7 (18 °C)	4.78×10^{-4}	1.86×10^{-5}
2-Methylenebutanedioic acid	itaconic acid	[*97-65-4*]	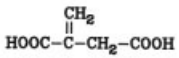	130.10	175	1.632	8.3	1.5×10^{-4}	2.2×10^{-6}
2-Methylenepentanedioic acid	2-methylene-glutaric acid	[*3621-79-2*]	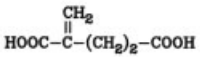	144.12	132–133				
2,4-Hexadienedioic acid	muconic acid (cis–cis)	[*3588-17-8*]	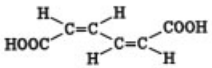	142.11	194–195				

as catalysts [140]. Other methods involve carbonylation of propargyl chloride with metal carbonyl catalysts [141] and thermal decomposition of citric acid.

Uses. Itaconic acid can be used as a comonomer and in the separation of triorganophosphine mixtures [142].

3.4.2. Dimer Acids

The only higher dicarboxylic acids of commercial importance are unsaturated cyclic C_{36} dicarboxylic acids known as dimer acids. The main difference among various standard proprietary products is the content of trimer compounds; the ratio of dimer to trimer acids can vary from 36 : 1 to about 0.7 : 1 [143]. The composition of dimer acids depends to a large degree on the feed materials and the manufacturing process.

Production. Dimer acids are produced by intermolecular condensation of unsaturated C_{18} carboxylic acids or their esters. Tall oil fatty acids are the main feed materials, but oleic and linoleic acids can also be used. The reaction is conducted preferably in the presence of special aluminum silicates (montmorillonites) at 190 – 240 °C; thermal dimerization at 270 – 290 °C is also possible [144–147]. Numerous catalyst modifications such as addition of alkali, amines, or sulfonic acid halides are described in the literature. A summary of the manufacturing processes for dimer acids can be found in [151–153]. Diels – Alder and free-radical reactions have been suggested as mechanisms for the thermal dimerization [148]; when aluminum silicates are used, ionic intermediates may be formed. Commercial processes yield mixtures of dimer acids, higher polycarboxylic acids, and various isomeric monomer acids, which are separated by distillation using film evaporators. After removal of the monomer acid fraction and the polycarboxylic acids, the remaining acids can be separated by further distillation into dimer and trimer acids. Figure 4 shows some dimer acids found in commercial mixtures. Investigations of the structures of dimer acids can be found in [149], [150].

Uses. The dimer acids produced on an industrial scale are used in the manufacture of polyamides, polyesters, epoxy resins, lubricants, plasticizers, and pesticides [154–157].

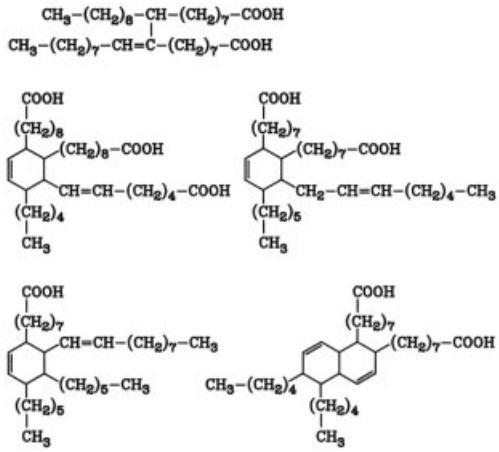

Figure 4. Dimer acids found in commercial mixtures

4. Quality Specifications and Analysis

Quality control of dicarboxylic acids covers the determination of content, melting point, color, traces of heavy metals, and solubility in water or other solvents. High purity is generally demanded of dicarboxylic acids.

The content of dicarboxylic acids is usually determined by acidimetric titration. Specific regulations exist for properties; for regulations concerning dimer acids, see [158, 159]. Specifications may include condition, color, content, and ash. Melting point, density, refractive index, water content, steam pressure, specific heat, dissociation constants, and solubility in water and other solvents are also often determined.

The most important qualitative or quantitative analytical method used industrially is gas chromatography. Dicarboxylic acids are first converted into their esters (preferably methyl esters) because free acids generally undergo undesirable secondary reactions during chromatography. Both packed columns and capillary columns with stationary liquids of different polarities are used. Calcined kieselguhr is the most common carrier material. Other chromatographic methods such as HPLC, paper and thin-layer chromatography, and gel chromatography are also widely employed.

Alkalimetric titration is commonly used to monitor the different production steps and to

identify pure dicarboxylic acids or their mixtures. In the absence of other reducing agents, oxalic acid is determined by titration with potassium permanganate. Crystalline derivatives such as phenacyl esters and amides are particularly suitable for chemical determination of dicarboxylic acids.

With infrared spectroscopy, dicarboxylic acids can be detected by the intense carbonyl stretching frequency in the range of $1650-1740\,cm^{-1}$. In the ^{1}H-NMR spectrum, the hydroxyl proton signals can be found at $\delta = 10-13$ ppm; the signals for the methylene groups of malonic and succinic acids are around 3.4 and 2.6 ppm, respectively. In the ^{13}C-NMR spectrum, the absorption range of the carboxyl carbon atom is around $\delta = 160-180$ ppm.

For further details on analysis, see [158], [160].

5. Storage, Transportation, and Handling

At room temperature, straight-chain dicarboxylic acids are solid compounds that are delivered and stored as crystals or – particularly if a melt or distillation is used to recover the pure substance – as flakes.

Dicarboxylic acids are stored and transported in drums made of plastic-coated steel, stainless steel, or aluminum. Polyethylene-lined paper sacks are also used. These acids are hygroscopic and should be stored in cool, dry rooms to avoid clumping. Shipping regulations and hazard classification for dicarboxylic acids depend on the specific properties of the compounds such as flash point, decomposition temperature, water solubility, toxicity, and ignition temperature.

Aliphatic dicarboxylic acids are local irritants, especially to the mucous membranes; this effect decreases with increasing chain length. Oxalic acid is absorbed readily by the outer layers of the skin and can upset the body's calcium balance through the formation of calcium oxalate. Therefore, gloves and safety glasses must be worn when dicarboxylic acids are handled. To avoid dust that can damage health, dicarboxylic acids are normally supplied in the form of flakes or laminated moldings.

Waste gases from dicarboxylic acid production facilities are generally drawn off at a central point and fed into a combustion chamber. The wastewater is subjected to chemical and biological treatment.

6. References

1. BP Chemicals Ltd., JP-Kokai 77 19 618, 1977.
2. Monsanto Co., US 4 254 283, 1981 (G. H. Mock).
3. G. Jones, *Org. React. (N.Y.)* **15** (1967) 204.
4. H. R. Kaenel, M. Brossi, *SLZ Schweiz. Lab. Z.* **41** (1984) 197 – 9. H. O. House, *Modern Synthetic Reactions,* W. A. Benjamin Inc., New York – Amsterdam 1965.
5. R. N. Hurd, D. H. Shah, *J. Org. Chem.* **38** (1967) 607. W. S. Johnson, G. H. Daub, *Org. React. (N.Y.)* **6** (1951) 1.
6. C. W. Schimelpfenig, V. T. Lin, J. F. Waller, *J. Org. Chem.* **28** (1963) 805.
7. J. J. Bloomfield, D. C. Owsley, J. M. Nelke, *Org. React. (N.Y.)* **23** (1976) Chap. 2, 259.
8. V. Matthews: "Aliphatic Dicarboxylic Acids and related Compounds," Chap. 17, in: Rodd's Chemistry of Carbon Compounds, Elsevier, Amsterdam 1973. A. Cox: "Dicarboxylic and Polycarboxylic Acids," in: Comprehensive Organic Chemistry, Pergamon Press, Oxford 1979. J. Falbe: "Carbonsäuren," in: Methodicum Chimicum, vol. 5, Thieme Verlag, Stuttgart 1975. R. W. Johnson, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 241.
9. Emery Industries, US 2 813 113, 1957 (C. G. Goebel). N. A. Bogdanova, M. L. Kolesov, *Zh. Khim. Promsti* **4** (1978) no. 49; *Chem. Abstr.* **88** (1978) 191 525. A. Heins, M. Witthaus, *Henkel-Referate* **20** (1984) 42.
10. D. D. Nanavati, *J. Sci. Ind. Res.* **35**(1976) 163 – 8. A. P. Kudchadker et al., *Encycl. Chem. Process. Des.* **6**(1978) 401 – 20. Intreprinderea de Sapun, RO 60 094, 1976 (C. Razvan, D. Lungu, T. Popescu, M. Petrescu).
11. General Anilin and Film Corp., US 2 821 534, 1954 (W. N. Alexander).
12. Chem. Werke Hüls, EP 48 476, 1982 (H. Heumann, W. Hilt, H. Liebing, M. Schweppe). J. G. D. Schulz, A. Onopchenko, *J. Org. Chem.* **45** (1980) 3716 – 19. Kokai Boshi Chosa Kenkyusho K. K., DE-OS 2 638 046, 1977 (W. Ando, I. Nakaoka).
13. Chem. Werke Hüls, DE-OS 2 001 182, 1970 (H. Röhl, W. Eversmann, P. Hegenberg, E. Hellemanns). Chem. Werke Hüls, DE-OS 1 919 228, 1969 (H. Röhl, W.

Eversmann, P. Hegenberg, E. Hellemanns). Chem. Werke Hüls, DE-OS 2 217 003, 1972 (E. Hellemanns, H. Röhl, P. Hegenberg, W. Eversmann). Du Pont, DE-OS 1 912 569, 1969 (J. O. White, D. D. Davis).
14. F. Runge, R. Hueter, H. D. Wulf, *Chem. Ber.* **87** (1954) 1430.
15. J. H. Pistor, H. Plieninger, *Justus Liebigs Ann. Chem.* **562** (1939) 239.
16. J. English, J. E. Dayan, *Org. Synth., Coll. Vol.* **IV** (1963) 499.
17. Agency of Ind. Sciences and Technology, JP-Kokai 82 64 637, 1982. Chem. Werke Hüls, DE-OS 2 942 279, 1981 (K. D. Dohm, P. Hofmann). Dainippon Ink and Chemicals Inc., JP-Kokai 82 32 245, 1982.
18. Mitsui Toatsu Chemicals Inc., JP-Kokai 77 151 117, 1977 (M. Kawamata, S. Fujikake, K. Tanabe). S. Miyazaki, Y. Suhara, *J. Am. Oil Chem. Soc.* **55** (1978) 536. Kogai Boshi Chosa Kenkyusho K. K., DE-OS 2 826 065, 1979 (W. Ando, I. Nakaoka).
19. BASF, DE-OS 1 618 177, 1975 (F. Merger). Dynamit Nobel, FR 2 008 138, 1970 (H. Aus der Fuenten, H. Richtzenhain). Dynamit Nobel, BE 742 378, 1970 (H. Nestler, H. Richtzenhain).
20. Intreprinderea Chimica Dudesti, RO 79 020, 1982 (S. D. Paucescu, L. Toth, G. Munteanu, M. Badea, S. Kurti). H. Mueller, U. Resch, DD 146 454, 1981.
21. Asahi Denka Kogyo K. K., JP 793 950, 1979 (K. Yamada, T. Hattori, Y. Shirakawa). Bio Research Center Co. Ltd., DE-OS 2 853 847, 1979 (A. Taoka, S. Uchida). Dainippon Ink and Chemical Inc., JP-Kokai 82 129 694, 1982. Mitsui Petrochemical Industries Inc., JP-Kokai 82 206 394, 1982.
22. J. Falbe (ed.): *Carbon Monoxide in Organic Synthesis,* Springer Verlag, Berlin – Heidelberg – New York 1970. J. Falbe (ed.): *New Syntheses with Carbon Monoxide,* Springer Verlag, Berlin – Heidelberg – New York 1980. Noguchi Research Foundation, JP-Kokai 79 92 913, 1979 (Y. Sado, K. Tajima). Shell Intern. Res. Maats. B. V., EP 42 633, 1981 (E. Drent).
23. N. E. Lawson, T. T. Cheng, F. B. Slezak, *J. Am. Oil Chem. Soc.* **54** (1977) 215. E. N. Frankel, E. H. Pryde, *J. Am. Oil Chem. Soc.* **54** (1977) 873A.
24. M. Seko, A. Yomiyama, T. Isoya, *Hydrocarbon Process.* 1979, Dec., 117.
25. J. W. Conolly, G. Urry, *J. Org. Chem.* **29** (1964) 619. *Houben-Weyl,* **V/1a,** Part 1, 480.
26. H. Stetter, W. Dierichs, *Chem. Ber.* **86** (1953) 693. H. Stetter, R. Engl, H. Rauhut, *Chem. Ber.* **91** (1958) 2882. H. Stetter, R. Engl. H. Rauhut, *Chem. Ber.* **92** (1959) 1184.
27. S. Hünig, H. Hoch, *Chem. Ber.* **105** (1972) 2197, 2216. S. Hünig, *Fortschr. Chem. Forsch.* **14** (1970) 235.
28. *Winnacker-Küchler,* 4th ed., **vol. 6,** Chap. 8.6. E. C. Leonard in E. H. Pryde (ed.): *Fatty Acids,* Chap. 25, The American Oil Chemists' Society, Champaign, Ill., 1979. V. Matthews in: *Rodd's Chemistry of Carbon Compounds,* **vol. I,** Chap. 17, Elsevier, Amsterdam 1973. A. Cox in: *Comprehensive Organic Chemistry,* **vol. 2,** Chap. 9.2, Pergamon Press, Oxford 1979. J. Falbe (ed.): "C–O-Verbindungen" in: *Methodicum Chimicum,* **vol. 5,** Chap. 8, Thieme Verlag, Stuttgart – New York 1975.
29. M. Polievka, V. Macho, L. Uhlar, J. Kordik, CS 195 860, 1982.
30. P. Ruiz et al., *ACS Symp. Ser.* **237** (1984) 15 – 36; *Chem. Abstr.* **100** (1984) 23 918.
31. H. Müller, U. Resch, DD 146 454, 1981.
32. S. Loktev et al., *Neftepererab. Neftekhim. (Moscow)* **2** (1978) 34; *Chem. Abstr.* **88** (1978) 169 557.
33. Shell Int. Res. Maats. B. V., EP 42 633, 1981 (E. Drent).
34. J. Stille, R. Divakatumi, *J. Org. Chem.* **44** (1979) 3474 – 82.
35. Lonza, DE 1 133 359, 1956 (G. Natta, P. Pino). Lonza, US 2 851 456, 1957 (E. S. Rothman, M. E. Wall).
36. BASF, US 2 604 490, 1951 (W. Reppe).
37. Gulf Research and Development Co., US 4 263 453, 1981 (J. G. D. Schulz, A. Onopchenko).
38. BP Chemicals Ltd., GB 7 533 012, 1975.
39. ICI Ltd., EP 2598, 1979 (B. Baker).
40. B. L. Moldavskii, R. J. Rudakova, *Zh. Prikl. Khim. (Leningrad)* **33** (1960) 417.
41. Agency of Ind. Sciences and Technology, JP-Kokai 78 46 914, 1978 (S. Miyazaki).
42. V. N. Gogte et al., *Tetrahedron Lett.* **24** (1983) 4131 – 4.
43. M. T. Ismail et al., *Bull. Fac. Sci. Assiut. Univ.* **11** (1982) 121 – 6.
44. I. B. Blanshtein, V. Bunin, E. I. Leenson; SU 70 19 94, 1979.
45. J. Tian et al., *Wei Sheng Wu Hsueh Pao* **21** (1981) 229 – 33; *Chem. Abstr.* **95** (1981) 113 347.
46. All-Union Scientific Res. and Dev. Inst., SU 937 444, 1982 (R. Nesterova et al.).

47. Gulf Research and Development Co., US 4 158 739, 1979 (J. G. D. Schulz, A. Onopchenko).
48. R. M. Habib et al., *J. Org. Chem.* **49** (1984) 2780 – 4.
49. Z. G. Petrova et al., *Deposited Doc.* 1981, 1336 – 81; *Chem. Abstr.* **96** (1982) 217 201.
50. M. Lissell, E. V. Dehmlov, *Tetrahedron Lett.* **39** (1978) 3689.
51. K. Isogai, K. Awano, *Yuki Gosei Kagaku Kyokaishi* **35** (1977) 280 – 4; *Chem. Abstr.* **87** (1977) 52 699.
52. All-Union Scientific-Research Inst., SU 891 630, 1981 (L. V. Aleksanyan et al.).
53. Erevan Plant of Chem. Reagents, SU 535 283, 1976 (K. G. Akopyan et al.).
54. Mitsubishi Chemical Ind. Co. Ltd., JP-Kokai 78 12 807, 1978 (T. Maki, T. Ochiai, T. Yamaura).
55. Y. Takahashi et al., *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **28** (1983) 392 – 6.
56. N. Yomeda et al., *Chem. Lett.* **6** (1981) 767 – 8.
57. Dynamit Nobel, FR 2 017 352, 1970 (H. Richtzenhain, H. Aus der Fuenten).
58. Ciba S. A., BE 739 526, 1968.
59. Bayer, DE 2 923 775, 1979 (F. Maurer, U. Priesnitz, H. J. Riebel).
60. Merck & Co. Inc., US 4 262 009, 1979 (B. G. Christensen, D. H. Shin).
61. Lonza, EP 64 633, 1981 (P. Lehky, V. Franzen). Lonza, EP 64 634, 1981 (P. Lehky, P. Hardt). Lonza, EP 65 706, 1981 (P. Lehky).
62. Veba-Chemie AG, DE 1 418 074, 1976 (E. Rindtorff, K. Schmitt, H. Heumann).
63. Chem. Werke Hüls, DE-OS 3 200 065, 1983 (W. Huebel, J. Reiffer).
64. F. Debuan, P. Häussle, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) 511.
65. Vianova Kunstharz AG, DE 2 638 464, 1975 (H. Lackner, J. Manger, P. Thaller). Chem. Werke Hüls, EP 25 828, 1979 (S. Maman).
66. Ciba-Geigy, DE 2 651 650, 1975 (L. Buxbaum, J. Habermeier).
67. W. Langenbeck, M. Richter, *Chem. Ber.* **89** (1956) 202.
68. V. Odinokov et al., *Zh. Org. Khim.* **14** (1978) 54 – 9; *Chem. Abstr.* **88** (1978) 151 987.
69. Dainippon Ink and Chemicals Inc., JP-Kokai 82 142 940, 1982.
70. P. E. Verkade, *Rec. Trav. Chim. Pays Bas* **46** (1927) 137.
71. T. Antonova et al., *Ref. Zh. Khim.* 1981, Abstr. No. 10N44; *Chem. Abstr.* **95** (1981) 61 415.
72. W. Reppe, *Justus Liebigs Ann. Chem.* **582** (1953) 1 – 161.
73. Y. Nakauo, T. A. Foglia, *J. Am. Oil Chem. Soc.* **59** (1982) 163 – 6.
74. Nippon Oils and Fats Co., JP-Kokai 81 169 640, 1981.
75. Nippon Oils and Fats Co., JP-Kokai 82 4940, 1982.
76. N. Garti, E. Auni, *J. Am. Oil Chem. Soc.* **58** (1981) 840-1.
77. T. A. Foglia et al., *J. Am. Oil Chem. Soc.* **54** (1977) 870A – 872A.
78. V. W. Advani et al., *J. Oil Technol. Assoc. India* **8** (1976) 27 – 30.
79. Welsbach Corp., US 2 865 937, 1958 (A. Maggiolo).
80. Welsbach Corp., US 2 897 231, 1959 (S. J. Niegowski, A. Maggiolo).
81. Inst. of Heteroorganic Compounds, Acad. of Sc., SU 1 092 150, 1984 (L. I. Zakharin, V. V. Guseva).
82. Kuraray Co., JP-Kokai 83 140 038, 1983.
83. Brichima, DE-OS 3 027 111, 1981 (F. Minisci, P. Maggioni, A. Citterio).
84. Mitsubishi Petrochemical Co., JP-Kokai 82 79 889, 1982.
85. A. P. Kudehandker et al., *Encycl. Chem. Process. Des.* **6** (1978) 401 – 20.
86. Intreprinderea de Sapun, RO 60 094, 1976 (C. Razvan, D. Lungu, T. Popescu, M. Petrescu).
87. D. D. Nanavati, *J. Sci. Ind. Res.* **35** (1976) 163 – 8.
88. M. J. Diamond et al., *J. Am. Oil Chem. Soc.* **44** (1967) 656.
89. Asahi Chemical Ind. Co., US 4 237 317, 1980 (K. Yamataka, Y. Matsuoka, T. Isoya).
90. Asahi Chemical Ind. Co., DE-OS 3 019 537, 1980 (K. Yamataka, T. Isoya, C. Kawamura).
91. Asahi Chemical Ind. Co., DE-OS 2 830 144, 1979 (K. Yamataka, Y. Matsuoka, T. Isoya).
92. Daicel Ltd., JP 78 25 032, 1978 (K. Yamada, A. Nishihara, Y. Shirakawa, T. Nakazawa).
93. S. I. Popovick et al., *Plast. Massy* **9** (1983) 62; *Chem. Abstr.* **99** (1983) 176 712.
94. Asahi Chemical Ind. Co. Ltd., JP-Kokai 79 46 710, 1979 (K. Yamataka, T. Isoya, Y. Matsuoka).
95. Asahi Chemical Ind. Co., JP-Kokai 78 82 717, 1978 (J. Nishikido, N. Tamura).
96. I. S. Chevilenko et al., *Zh. Prikl. Khim. (Leningrad)* **55** (1982) 2742 – 6.
97. I. S. Chevilenko et al., *Khim. Promst. (Moscow)* **2** (1980) 123.
98. Asahi Denka Kogyo K. K., JP 78 15 051, 1978 (H. Yamamoto, I. Hisano, T. Okamota).
99. I. S. Chevilenko et al., *Khim. Promst. (Moscow)* **1** (1977) 14 – 16.

100. BASF, GB 1 092 603, 1967 (G. Riegelbauer, A. Wegerich, A. Kuerzinger, E. Haarer). Chem. Werke Hüls, DE 1 643 854, 1968 (E. Vangermain).
101. Dainippon Ink and Chemicals Inc., JP-Kokai 82 32 245, 1982.
102. Chem. Werke Hüls, DE-OS 2 942 279, 1981 (K. D. Dohm, P. Hofmann).
103. Y. N. Yurev et al., SU 322 984, 1981.
104. Bio Research Center Co., JP-Kokai 81 154 993, 1981.
105. Bio Research Center Co., DE-OS 2 909 420, 1979 (A. Watanabe, A. Taoka, S. Uchida).
106. SNIA Viscosa, DE-OS 2 945 004, 1980 (F. Siclari, L. Canavesi, P. P. Rossi).
107. Ube Industries, JP-Kokai 80 104 226, 1980.
108. Du Pont, US 4 149 013, 1979 (D. A. Klein).
109. Asahi Chem. Ind. Co., JP-Kokai 78 82 718, 1978 (J. Nishikido, N. Tamura).
110. Toa Gosei Chemical Industry Co., JP-Kokai 76 108 015, 1976 (Y. Yasuda, T. Matsubara).
111. J. J. Jaskierski et al., *Zesz. Probl. Postepow. Nauk Roln.* **211** (1981) 159 – 165, 311 – 319.
112. K. D. Carlson et al., *Ind. Eng. Chem. Prod. Res. Dev.* **16** (1977) 95 – 101.
113. Mitsui Petrochemical Ind. Ltd., JP-Kokai 83 121 797, 1983.
114. Mitsui Petrochemical Ind. Ltd., JP-Kokai 83 60 994, 1983.
115. Mitsubishi Petrochemical Ind. Ltd., JP-Kokai 82 102 191, 1982.
116. Mitsubishi Petrochemical Co. Ltd., JP-Kokai 81 11 796, 1981.
117. Nippon Mining Co. Ltd., JP-Kokai 81 11 796, 1981.
118. Mitsui Petrochemical Ind., JP-Kokai 81 15 693, 1981.
119. Mitsui Petrochemical Ind., JP-Kokai 81 15 694, 1981.
120. Mitsui Petrochemical Ind., JP-Kokai 81 15 695, 1981.
121. Bio Research Center Co., DE-OS 2 951 177, 1980 (A. Watanabe, S. Uchida, A. Toaka).
122. United States of America, Secretary of Agriculture, US 3 270 035,1966 (E. T. Roe, D. Swern).
123. Matsubara et al., *Kogyo Kagaku Zasshi* **71** (1968) 1179.
124. Armour and Company, US 3 481 977, 1969 (E. J. Miller, A. Mais, D. Say).
125. J. P. Friedrich, *J. Am. Oil Chem. Soc.* **53** (1976) 125.
126. E. H. Pryde, *J. Am. Oil Chem. Soc.* **61** (1984) 419.
127. E. N. Frankel, *J. Am. Oil Chem. Soc.* **48** (1971) 248.
128. A. W. Schwab et al., *J. Am. Oil Chem. Soc.* **49** (1972) 75.
129. E. J. Dufek et al., *J. Am. Oil Chem. Soc.* **49** (1972) 302.
130. H. M. R. Hoffmann, *Angew. Chem.* **81** (1969) 597; *Angew. Chem. Int. Ed. Engl.* **8** (1969) 556.
131. A. Cox in: *Comprehensive Organic Chemistry,* **vol. 2,** Chap. 9.2, Pergamon Press, Oxford 1979. V. Matthews in: *Rodd's Chemistry of Carbon Compounds,* **vol. 1,** Chap. 17,Elsevier, Amsterdam 1973. E. C. Leonard in: *Fatty Acids,* Chap. 25, The American Oil Chemists Society, Champaign, Ill., 1979.
132. Iwata Kagaku Kogyo K. K., JP-Kokai 84 63 190, 1984.
133. Shizuoka Prefecture, Banda Kagaku Kogyo K. K., JP-Kokai 81 137 893, 1981.
134. N. Nakagawa et al., *J. Ferment. Technol.* **62** (1984) 201 – 203.
135. Mitsubishi Chemical Industries Co. Ltd., JP-Kokai 80 34 017, 1980.
136. S. Ikeda, *Hakko Kogaku Kaishi* **60** (1982) 208 – 210.
137. L. B. Lockwood, *Microb. Technol.,* 2nd ed., **vol. 1,** Academic Press, New York 1979, pp. 355 – 387.
138. L. M. Miall, *Econ. Microbiol.* **2** (1978) 47 – 119.
139. Nissan Chemical Industries Ltd., JP-Kokai 78 15 316, 1978 (A. Murata, T. Ishii).
140. Denki Kagaku Kogyo K. K., JP-Kokai 74 101 326, 74 101 327, 1973 (T. Shimizui, C. Fujii).
141. J. Tsuji, T. Nogi, *Tetrahedron Lett.* 1966, 1801 – 04.
142. Union Carbide Corp., EP 19 296, 1980 (D. R. Bryant, R. A. Galley).
143. E. C. Leonard in: *Fatty Acids,* Chap. 25, The American Oil Chemists Society, Champaign, Ill., 1979,p. 521.
144. Yaroslavl Polyt. Inst., SU 1 057 515, 1983 (V. V. Solovev, B. N. Bychkov, G. N. Koshel, L. A. Rodivilova, R. D. Zhilina, E. A. Bondareva).
145. Toei Chemical K. K., JP-Kokai 78 23 306, 1978 (N. Fujihana, M. Terauchi).
146. Emery Industries Inc., CA 1 019 343, 1977 (R. J. Sturwold, H. J. Sharkey).
147. Agency of Industrial Sciences and Technology, JP-Kokai 76 88 910, 1976 (O. Suzuki, K. Tanabe, T. Hashimoto).

148. R. W. Johnson in: *Fatty Acids,* Chap. 17,The American Oil Chemists Society, Champaign, Ill., 1979,
149. D. H. Wheeler, A. Milum, F. Linn, *J. Am. Oil Chem. Soc.* **47** (1970) 242 – 244.
150. D. H. Wheeler, J. M. White, *J. Am. Oil Chem. Soc.* **44** (1967) 298 – 302.
151. E. H. Pryde (ed.): *Fatty Acids,* Chaps. 17, 25, The American Oil Chemists Society, Champaign, Ill., 1979.
152. E. C. Leonhard (ed.): *The Dimer Acids,* Humbko Sheffield Chemicals, 1975.
153. E. C. Leonhard, *J. Am. Oil Chem. Soc.* **56** (1979) 782A – 785A.
154. Sankyo Co. Ltd., JP-Kokai 76 41 441, 1976 (H. Takehara).
155. W. E. A. De Mear, R. J. Bellamy, *Polym. Paint Colour J.* **164** (1974) 233.
156. *Aust. OCCA Proc. News* **9** (1972) 5 – 14.
157. G. C. Guainazzi, *Boll. Assoc. Ital. Tec. Ind. Vernici Affini* **61** (1978) 3 – 12.
158. J. P. Nelson, A. J. Milun, *J. Am. Oil Chem. Soc.* **52** (1975) 81 – 83.
159. A. Zeman, H. Sharmann, *Fette, Seifen, Anstrichm.* **71** (1969) 957 – 960.
160. E. Pretsch, T. Clerc, J. Seibl, W. Simon: *Tabellen zur Strukturaufklärung organischer Verbindungen,* Springer Verlag, Heidelberg 1976. D. H. Williams, J. Fleming: *Spektroskopische Methoden in der organischen Chemie,* Thieme Verlag, Stuttgart 1971. H. Roth: "Analytik der Carboxylgruppen," in: Houben-Weyl, 2, 487. F. Korte (ed.): "Analytik," in: Methodicum Chimicum, vol. 1, Thieme Verlag, Stuttgart 1973.

Dichloroacetic Acid → Chloroacetic Acids
Dichloroethanes → Chlorinated Hydrocarbons
Dichloroethylene → Chlorinated Hydrocarbons
Dichloromethane → Chlorinated Hydrocarbons
Dicyandiamide → Cyanamides
Dicyanogen → Cyano Compounds, Inorganic
Diesel Fuel → Automotive Fuels
Diethanolamine → Ethanolamines and Propanolamines
Diethyl Ether → Ethers, Aliphatic
Diethyl Sulfate → Dialkyl Sulfates and Alkylsulfuric Acids
Diffusion Separation Methods → Isotopes, Natural
Diffusion Separation Methods → Separation Processes, Introduction
Digital Displays → Display Technology
Digitonin → Saponins
Diglyme → Ethylene Glycol
Diketene → Ketenes
Dimer Acid → Dicarboxylic Acids, Aliphatic
Dimethylamine → Methylamines

Dimethyl Ether

Manfred Müller, DEA Mineraloel AG, Wesseling, Federal Republic of Germany

Ute Hübsch, DEA Mineraloel AG, Wesseling, Federal Republic of Germany

1. Introduction 1
2. Properties 1
3. Production 2
4. Uses 3
5. Toxicology 4
6. References 4

1. Introduction

Dimethyl ether [*115-10-6*], C_2H_6O, is the simplest aliphatic ether.

CH_3–O–CH_3

For a long time it was obtained on an industrial scale as a byproduct in the high-pressure production of methanol. Recently the high-pressure methanol plants have been almost completely replaced by low-pressure plants; this has led to the erection of special plants for the synthesis of dimethyl ether. The production of dimethyl ether in Western Europe is approx. 50 000 t per year. Dimethyl ether is industrially important as the starting material in the production of the methylating agent dimethyl sulfate and is used as an aerosol propellant.

2. Properties

Dimethyl ether (DME, methoxymethane), M_r 46.07, is a colorless, almost odorless gas at room temperature and atmospheric pressure and has the following physical properties:

M_r	46.07
bp at 0.1 MPa	− 24.8 °C
mp	− 141 °C
Critical pressure	5.28 MPa (52.84 bar) [1]
Critical temperature	400.29 K [1] (127.1 °C)
Critical density	269.9 kg/m^3 [1]
Heat of combustion (gas)	31.75 MJ/kg
Heat of formation	−183 kJ/mol
Specific heat capacity (at −24 °C)	2.26 kJ kg^{-1} K^{-1}
Heat of vaporization (at −20 °C)	410.2 kJ/kg
Autoignition temperature	235 °C [2]
Explosive limits	3.0 – 17 vol % in air
Flash point	− 41 °C
Relative density (gaseous, air = 1)	1.59
Density	
(at 20 °C)	668.3 kg/m^3 [2]
(at 50 °C)	615.0 kg/m^3 [2]
Solubility in water	
(at 20 °C, 1 bar)	5.7 wt % [2]
(at 20 °C, 4.8 bar)	36 wt % [2]
Surface tension (liquid)	0.0125 N/m [2]
Viscosity	
gaseous	0.0091 mPa · s [2]
liquid	0.11 mPa · s

Density of liquid dimethyl ether [2]:

t, °C	10	20	30	40	50	60	70	80
ϱ, kg/m^3	682	666	649	631	612	593	573	552

Vapor pressure [2]:

t, °C	−20	−10	0	10	20	50
p, MPa						
measured	0.124	0.184	0.266	0.375	0.512	1.149
calculated	0.114	0.185	0.267	0.374	0.512	1.152

Dimethyl ether is miscible with most polar and nonpolar organic solvents. It is also partly miscible with water (76 g in 1 L of water at 18 °C). In addition, numerous polar and nonpolar substances readily dissolve in dimethyl ether.

Chemical Properties. Unlike most other aliphatic ethers, dimethyl ether is not susceptible to autoxidation. This is of considerable importance for industrial applications. Numerous studies have confirmed that dimethyl ether is stable to atmospheric oxygen and does not form peroxides [3].

Dimethyl ether is converted to dimethyl sulfate [*77-78-1*] by sulfur trioxide (→ Dialkyl Sulfates and Alkylsulfuric Acids).

$SO_3 + CH_3$–O–$CH_3 \longrightarrow (CH_3)_2SO_4$

Addition of boron trifluoride to dimethyl ether results in the formation of $BF_3 \cdot CH_3OCH_3$ [*353-42-4*], *mp* −14 °C, *bp* 127 °C, a distillable liquid which fumes when exposed to moist air.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_541

This addition compound is easier to handle than gaseous boron trifluoride.

In the presence of CoI_2, dimethyl ether reacts with carbon monoxide and water to form acetic acid [*64-19-7*] (→ Acetic Acid, Chap. 4.1.) [4].

$$CH_3\text{–}O\text{–}CH_3 + H_2O + 2\,CO \longrightarrow 2\,CH_3COOH$$

The reaction of dimethyl ether with hydrogen sulfide in the presence of a catalyst, e.g., tungsten sulfide (WS_2), gives dimethyl sulfide [*75-18-3*] [5].

$$CH_3\text{–}O\text{–}CH_3 + H_2S \longrightarrow CH_3\text{–}S\text{–}CH_3 + H_2O$$

The reaction of dimethyl ether to give unsaturated and saturated hydrocarbons proceeds in the presence of zeolitic catalysts [6].

$$CH_3\text{–}O\text{–}CH_3 \longrightarrow H_2C{=}CH_2 + H_2O$$

The catalytic conversion of dimethyl ether to formaldehyde has also been observed [7].

$$CH_3\text{–}O\text{–}CH_3 + O_2 \longrightarrow 2\,CH_2O + H_2O$$

3. Production

Until about 1975 dimethyl ether was obtained as a byproduct in the high-pressure production of methanol [8]. In this process up to 3 – 5 wt % dimethyl ether is formed. Dimethyl ether can be recovered in pure form by distillation of crude methanol. The development of the low-pressure methanol synthesis, particularly by Lurgi [9] and ICI [10], has resulted in the almost complete replacement of all high-pressure plants by 1980. The low-pressure processes, which require less severe conditions, produce only very small amounts of dimethyl ether (→ Methanol). As a result, special catalytic processes have been developed for the production of dimethyl ether.

The preparation of dimethyl ether from methanol in the presence of acidic catalysts on a laboratory scale has been known for many years [11]. Numerous methods have been discussed in the patent literature. For instance, aliphatic ethers can be prepared by heating alcohols in the presence of zinc chloride [12]. Other suitable catalysts are iron chloride, copper sulfate, copper chloride, manganese chloride, aluminum chloride, aluminum sulfate, chromium sulfate, alums, thorium compounds, aluminum oxide, titanium oxide, barium oxide, silica gel, and aluminum phosphate [13]. Aluminum oxide and aluminum silicate, with or without doping, are the most important catalysts for industrial use [14]. Figure 1 illustrates the process developed by Union Rheinische Braunkohlen Kraftstoff AG (now DEA Mineraloel AG, Wesseling) for the production of high-purity dimethyl ether (purity > 99.99). This quality is virtually odorless.

The catalytic dehydration of pure, gaseous methanol is carried out in a pipe reactor. The product is cooled in two stages and subsequently distilled to yield pure dimethyl ether. Small amounts of dimethyl ether are recovered from the off-gas in a scrubber and recycled to the reactor. Unreacted methanol is separated from water in a second column and also recycled.

Very pure dimethyl ether which is suitable as an aerosol propellant is obtained by special rectification processes.

Direct synthesis of dimethyl ether from synthesis gas ($CO + H_2$) has also been described [15].

Safety and Environmental Aspects. Dimethyl ether is a flammable gas. Water, foam, carbon dioxide, and dry chemical powders can be employed to control dimethyl ether fires. Fire extinguishers suitable for class C (Europe) or class B (United States) can be used. Endangered vessels must be cooled.

Although dimethyl ether is soluble in water and biologically easily degradable, large amounts of this compound should not enter the wastewater system because evaporation can cause the formation of explosive mixtures over the surface of the water.

Available data do not indicate a danger for the environment [2].

Quality Specifications and Analysis. Dimethyl ether is available in two commercial qualities:

- Technical grade contains up to 0.05 % methanol and contaminations with a strong odor.
- High-purity dimethyl ether, practically free of sulfur-containing and other substances with an unpleasant odor, can be used in

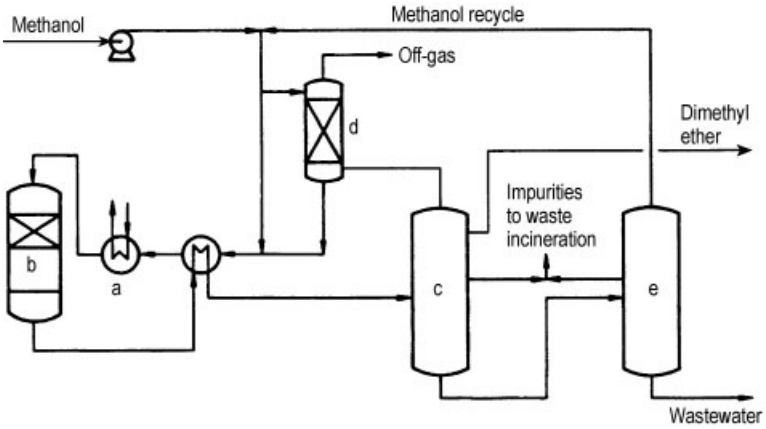

Figure 1. Dimethyl ether production by dehydration of methanol
a) Vaporizer; b) Reactor; c) Dimethyl ether column; d) Scrubber; e) Methanol column

the aerosol industry; the methanol content should not exceed 10 mg/kg for this quality.

The purity of dimethyl ether is determined by GC analysis. Oil and ash content are measured by special methods of evaporation and combustion.

Storage and Transportation. Dimethyl ether is usually stored as a liquid under pressure. It is transported in railway pressure tanks, tank trucks, and other pressure containers. Overseas transport is carried out in ISO tankers (International Shipping Organization).

Transportation is subject to the following regulations: GGVS/ADR GGVE/RID class 2, no. 2 F; ADN/ADNR class 2, no. 2 F; UN no. 1033; IMDG Code (Amendment 20-82), class 2.1, p. 2052.

Suppliers are DEA Mineraloel AG in Germany, Akzo Nobel Chemicals B.V. in the Netherlands, and DuPont & Nemours in Switzerland.

4. Uses

Up to the 1980s, the main industrial use for dimethyl ether was its conversion to dimethyl sulfate by treatment with sulfur trioxide. Dimethyl sulfate is used as a methylating agent. Of the 50 000 t of dimethyl ether produced in Western Europe in 1998, about 15 000 t was used in the production of dimethyl sulfate. The remaining 35 000 t of DME were used in the aerosol industry as a propellant. Owing to its high solubility coefficient, it also acts as a solvent in aerosol formulations. This property is of particular value in aerosol formulations which contain substances that are difficult to dissolve.

The reaction of dimethyl ether with carbon monoxide and water can be used in large-scale production of acetic acid instead of the methanol – carbon monoxide reaction. Future industrial uses of dimethyl ether could include the production of olefins, especially ethylene, propene, and butenes in the presence of zeolitic catalysts.

DME is also considered an alternative fuel for diesel engines (compression ignition) [16]:

- It has a low autoignition temperature
- It is an oxygenated fuel, and as a gas it helps establishing a good air – fuel mixture and consequently prevents soot formation
- It enables NO_x reduction.

5. Toxicology

Pure dimethyl ether is nontoxic. Inhalation experiments conducted on rats using concentrations of up to 20 000 ppm (2 vol % in air) over a period of 8 months did not lead to any deaths [3]. Contact with dimethyl ether is not irritating to the skin.

6. References

1. R. Graviliuc: *Dimethylether–Begründung für zusätzliche thermodynamische Forschung,* Institut für Technische Thermodynamik, Universität Karlsruhe 1998.
2. *DME Handbook,* DEA Mineraloel AG, 1998.
3. L. J. M. Bohnenn, *Manuf. Chem. Aerosol News* 1978, August, 39.
4. *Hydrocarbon Process.* **50** (1971) no. 11, 115.
5. Union Rheinische Braunkohlen Kraftstoff AG, Wesseling, DE 1 016 261, 1956 (F. Hübenett, N. Schnack).
6. F. X. Cormerais, G. Perot, M. Guisnet, *Zeolites* **1** (1981) October, 141. W. Lee, J. Mazink, V. W. Weckman Jr., S. Yurchak, *Biomass Wastes* **4** (1980) 721. C. D. Chang, C. T.-W. Chu, *J. Catal.* **74** (1982) 203. J. Haggin, *Chem. Eng. News* **63** (1985) March 25, 39.
7. Shell Oil Company, US 4 439 624, 1982 (R. M. Lewis, R. V. Ryan, L. H. Slaugh); US 4 442 307, 1982 (R. M. Lewis, R. C. Ryan, L. H. Slaugh).
8. *Winnacker-Küchler,* 4th ed., **vol. 5,** Carl Hanser Verlag, München–Wien 1981, p. 512.
9. E. Supp, *Chem. Technol.* **3** (1973) 430.
10. P. L. Rogerson, *Chem. Eng. (N.Y.)* **80** (1973) no. 19, 112.
11. *Houben-Weyl,* **vol. VI/3,** part 3, Georg Thieme Verlag, Stuttgart 1965, pp. 17, 18.
12. Dr. Alexander Wacker, Gesellschaft für elektrochemische Industrie GmbH, DE 680 328, 1934 (P. Halbig, O. Moldenhauer).
13. R. L. Brown, W. W. Odells, US 1 873 537, 1927. N. V. de Bataafsche Petroleum Maatschappij, FR 701 335, 1930;GB 332 756, 1929;GB 350 010, 1931;GB 403 402, 1932.
14. Mobil Oil Corp., DE 2 818 831, 1978 (F. G. Dwyer, A. B. Schwartz); DE-OS 3 201 155, 1982 (W. K. Bell, C. Chang). Du Pont, EP-A 99 676, 1983 (D. L. Brake). Mitsubishi Chemical Industries, EP-A 124 078, 1984 (N. Murai, K. Nakamichi, M. Otake, T. Ushikubo).
15. Snamprogetti, SpA., DE 2 362 944, 1973 (G. Giorgio); DE 2 757 788, 1977 (G. Manara, B. Notari, V. Fattore);DE 3 220 547, 1982 (G. Manara).
16. R. P. Verbeek, A. van Doorn, Global Assessment of Dimethyl Ether as an Automotive Fuel, 2nd ed., TNO-report 96.O.VM.029.1/RV, July 1996.

Dimethylformamide → Formamides
Dimethyl Sulfate → Diakyl Sulfates and Alkylsulfuric Acids
Dimethyl Sulfoxide → Sulfones and Sulfoxides
Dimethylterephthalate → Terephthalic Acid and Dimethylterephthalates
Diols → Alcohols, Polyhydric
Diols → Ethylene Glycol
Diosgenin → Saponins

Dioxane

KENNETH S. SURPRENANT, Dow Chemical Co., Midland, Michigan 48640, United States

1. **Physical Properties** 1
2. **Chemical Properties** 1
3. **Production** 3
4. **Environmental Protection** 3
5. **Quality Specifications** 3
6. **Storage and Transportation** 4
7. **Legal Aspects** 4
8. **Uses** 4
9. **Economic Aspects** 4
10. **Toxicology and OccupationalHealth** . 5
11. **References** 6

1. Physical Properties

The most important *physical and flammable properties* of dioxane are as follows:

bp	101.3 °C
fp	11.8 °C
d_{20}^{20}	1.0356
Latent heat of evaporation	413 kJ/kg
Heat of fusion	141 kJ/kg
Specific heat (20 °C)	1.76 kJ kg^{-1} K^{-1}
n^{20}	1.4224
Vapor pressure (20 °C)	4.13 kPa (41.3 mbar)
Dielectric constant (20 °C)	2.23
Surface tension (20 °C)	36.9 mN/m
t_{crit}	312 °C
p_{crit}	5.14 MPa
Flash point (closed cup)	11 °C
Flammable limits (in air)	
Lower limit	2 vol %
Upper limit	22 vol %
Heat of combustion	27 600 kJ/kg
Autoignition temperature	180 °C

1,4-Dioxane [*123-91-1*], $C_4H_8O_2$, M_r 88.11, is an extraordinary solvent, capable of solubilizing most organic compounds, water in all proportions, and many inorganic compounds.

Dioxane is a cyclic diether forming a six-membered ring with the following structure:

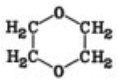

Synonyms for dioxane are *p*-dioxane, diethylene oxide, diethylene dioxide, diethylene ether, 1,4-dioxacyclohexane, dioxyethylene ether, and dioxan.

At ambient temperature, dioxane is a clear liquid with an ether-like odor. Its relative stability is similar to other aliphatic ethers, and it forms peroxides on exposure to air as they do. In spite of its exceptional solvent properties, its largest use is in stabilizing 1,1,1-trichloroethane and keep it from reacting with aluminum (see Chap. 8).

Reportedly, 1,4-dioxane was prepared by A. V. LOURENÇO in 1863 by reacting ethylene glycol and 1,2-dibromoethane [1]. However, commercial production did not occur until 1929.

Infinite *solubility* in dioxane can be expected from lower molecular mass aliphatic and aromatic liquid hydrocarbons, ethers, alcohols, ketones, and chlorinated hydrocarbons. Dioxane has solubility for animal and vegetable oils, paraffin oils, synthetic and natural resins, and some inorganic compounds and elements, such as iron chlorides, mercuric chloride, hydrochloric, sulfuric, and phosphoric acids, bromine, chlorine, and iodine.

Dioxane forms *azeotropic mixtures* with many compounds. Some common binary systems are given in Table 1.

2. Chemical Properties

As is characteristic of ethers, dioxane forms *peroxides* by air oxidation. These peroxides are often highly explosive and unstable. The peroxide of dioxane is 2-hydroperoxide 1,4-dioxane [*4722-59-2*] (**1**) [2], which decomposes to dioxanone [*3041-16-5*] (**2**) and water, and subsequently is hydrolyzed to form 2-(2-hydroxyethoxy)acetic acid [*13382-47-3*] (**3**):

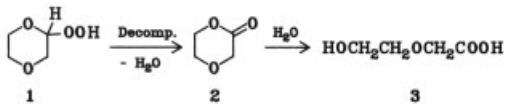

The reaction rates favor formation of **3** and limit the peroxide concentration. However, di-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_545

Table 1. Dioxane azeotropic data

Component	Component *bp*, °C	Dioxane, wt %	Azeotropic *bp*, °C
Formic acid	100.7	57	113.35
Nitromethane	100.8	43.5	100.55
Acetic acid	117.9	20.5 – 23	119.4
Ethanol	78.5	9.3	78.1
3-Iodo-1-propene	103	44	98.5
1-Propanol	97.4	45	95.3
tert-Amyl alcohol	102.5 *	20	100.65
Cyclohexane	80.7	24.6	79.5
Heptane	98.4	44	91.85
Water	100	82	87.8

* At 8 kPa.

oxane should be treated with the same precautions as other ethers to prevent peroxide formation. Inert gas padding (e.g., nitrogen atmosphere) may be used. Peroxide-containing ethers should not be evaporated or distilled to dryness. Peroxides can be destroyed by reducing agents, such as iron (II) or tin (II) chlorides or sodium bisulfite. Activated carbon, activated basic alumina, and the hydroxyl form of anion-exchange resins also remove peroxides [3].

Dioxane solubilizes and complexes with Cl_2, Br_2, I_2, and the halogen acids. These *complexes* can moderate halogenation or halogen acid reactions. For example, dioxane dibromide complex can brominate phenols, aromatic aldehydes or amines, and ketones. Conversely, chlorination can produce chlorinated dioxanes containing from one to eight chlorine atoms, with isomeric yields controlled by temperature, reaction conditions, and catalysts employed. Dioxanes containing fewer chlorine atoms can be used to prepare dioxane addition compounds, e.g., addition of ethoxy or ethyl groups using $NaOCH_2CH_3$ or CH_3CH_2MgBr, respectively.

2,3-Dichlorodioxane is particularly reactive. It condenses with aliphatic and aromatic acids, alcohols, and glycols to give diesters and ethers [4–6]. With ethylene glycol, the dicyclic isomers of 1,4,5,8-tetraoxadecahydronaphthalene are formed [7]. 2-Chloro-1,4-dioxene, made by dehydrohalogenation of 2,3-dichlorodioxane, can be polymerized or hydrolyzed to give dioxanone [8].

Although the *dioxane ring* is quite stable, it can be cleaved by strong acids at high temperature and pressure. Under such conditions, dioxane forms ethylene glycol diacetate and diethylene glycol diacetate when reacted with acetic acid anhydride [9]. In addition, benzoyl chloride and acetyl chloride catalyzed by $TiCl_4$ react with dioxane to form the 2-chloroethyl esters of benzoic and acetic acids [10], respectively.

Benzoyl peroxide catalyzes the reaction between α-olefins and dioxane to form *polymeric condensation products* [11]. The reaction requires a temperature of 80 – 140 °C and a reaction time of 5 – 10 h. The polymers are useful as lubricants. Similarly, tetrafluoroethylene reacted with dioxane at 110 °C with peroxide catalysis under pressure for 8 h forms analogous polymers of the following general structure [12]:

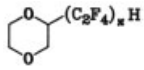

Numerous alternative catalysts are reported to be effective, including organic and inorganic peroxides, bases (e.g., borax, disodium phosphate, and hydrazine), zinc chloride, and phosphoric acid. Chlorofluoroethylenes containing at least two fluorine atoms and only one hydrogen atom yield similar products. The polymers formed can contain values of x from 1 to 25. A second polymer chain at another site on the dioxane ring can be formed by repeating the reaction with the first polymer product. The lower molecular mass products are volatile liquids of potential use as solvents. Higher molecular mass polymers are waxy solids and are regarded as possible lubricants.

Many *stable complexes* are reported in addition to those of the halogens mentioned previously. Some of the agents that give dioxane complexes are the following:

$AlBr_3$	LiBr	HCl
$AlCl_3$	LiCl	H_2SO_4
$AuCl_3$	LiI	H_3PO_4
BBr_3	$MnCl_2$	SO_2
BCl_3	$NiCl_2$	SO_3
BF_3	PF_5	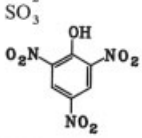
$CuCl_2$	$PtCl_4$	
$CoCl_2$	$SbCl_5$	
$FeCl_3$	$ZnCl_2$	
$HgCl_2$	ZnF_2	

These complexes can be employed as catalysts, as moderating agents in carrying out reactions, such as bromination or sulfonation, or as anhydrous systems for acid reactions.

3. Production

1,4-Dioxane is manufactured commercially by *dehydration and ring closure of diethylene glycol.* Concentrated sulfuric acid (ca. 5 %) is used as a catalyst, although phosphoric acid, *p*-toluenesulfonic acid, and strongly acidic ion-exchange resins are recognized alternatives. Operating conditions vary; temperatures range from 130 to 200 °C and pressures range from a slight pressure to a partial vacuum (25–110 kPa). A favorable temperature is reported to be 160 °C. The process is continuous, with dioxane vaporized from the reaction vessel. The vapors are passed through an acid trap and two distillation columns to remove water and to purify the product. Yields of ca. 90 % are possible. 2-Methyl-1,3-dioxolane, acetaldehyde, crotonaldehyde, and polyglycol are unwanted byproducts.

Dioxane can also be prepared by dehydrohalogenation of 2-chloro-2′-hydroxydiethyl ether, by reacting ethylene glycol with 1,2-dibromoethane, and by dimerizing ethylene oxide either over $NaHSO_4$ [1], SiF_4, or BF_3 [13], or at elevated temperature with an acidic cation-exchange resin [14].

4. Environmental Protection

Dioxane, emitted into the atmosphere, is decomposed rapidly by hydroxyl radical oxidation, as are other hydrocarbons. The half-life is estimated to be 3.9 d [15]. The intermediates are probably short-lived hydrocarbon radicals; the ultimate products are carbon dioxide and water.

Dioxane is highly resistant to biotransformation, although it may possibly be oxidized by acclimated sewage microorganisms. Carbon adsorption purification of water containing dioxane has limited practicality because of the high (ca. 100-fold) ratio of carbon to adsorbed dioxane required for complete removal. Aeration of water containing dioxane is inefficient because of dioxane's extreme solubility and low vapor pressure. Dioxane can be decomposed by H_2O_2 by using $FeSO_4$ to catalyze the reaction at 25 °C. Oxidation of dioxane by chlorination is effective at 75 °C, but not at 25 °C [16].

5. Quality Specifications

The American Chemical Society (ACS) has published specifications for 1,4-dioxane [17]. A summary of the ACS requirements is as follows:

Color (APHA), max.	20
Peroxide (as H_2O_2), max.	0.005 %
fp, min.	11.0 °C
Nonvolatile residue, max.	0.005 %
Neutrality	to pass test
A maximum of 1.0 mL of 0.01 M H_3PO_4 to neutralize a 50-mL sample.	
Carbonyl (as HCHO), max.	0.01 %
Water, max.	0.05 %

Industrial grades are sold under producer specifications that are more or less demanding than for ACS grade. These specifications often include boiling and relative density ranges and a maximum acidity. Industrial specifications are accompanied by analytical methods or referenced methods.

Higher purity dioxane may be required for research or special industrial processes. Techniques have been described in the literature to remove traces of peroxides, acids, stabilizers, and aldehydes [3].

Samples should be handled to prevent air contact, and nonvolatile residue tests should not be performed on material that contains peroxides.

6. Storage and Transportation

The most important consideration in shipping and storing dioxane is to prevent peroxide formation. This is achieved by shipping and storing with a nitrogen pad. Failure to maintain an inert

atmosphere causes the material to exceed acid, peroxide, and possibly other specifications.

Mild steel drums and tanks are adequate for shipping and storage. To limit the fire hazard, all containers should be grounded to prevent static electrical charges from accumulating during transfer operations. All areas of storage must be equipped with explosion-proof equipment and be well ventilated. The material must be maintained above its freezing point in cases where temperatures may be $< 12\,^{\circ}C$ by use of insulation or heat.

The United Nations' recommendations include the shipping name "Dioxane," number UN 1165, and labeling using the warning as a flammable liquid hazard class 3. The International Maritime identifies dioxane in class 3.2.

7. Legal Aspects

The U.S. National Toxicology Program (NTP) and the International Agency for Research on Cancer (IARC) have listed dioxane as a possible carcinogen on the basis of extreme high-dose lifetime animal studies discussed in Chapter 10. These listings may require special labeling or restrictions; national and local regulations should be consulted [18].

8. Uses

The largest single use of 1,4-dioxane is the *stabilization of 1,1,1-trichloroethane* against chemical attack by aluminum [19], [20]. The aluminum oxide film prevents the metal from reacting with 1,1,1-trichloroethane until the film is disturbed. When the bare metal contacts 1,1,1-trichloroethane, chlorine is extracted from the solvent to form $AlCl_3$, which causes dehydrohalogenation of 1,1,1-trichloroethane to produce HCl and vinylidene chloride ($CH_2 = CCl_2$). Dioxane is the foremost inhibitor of these reactions. An insoluble dioxane–aluminum chloride complex is formed, which deactivates the catalyst and seals any openings in the Al_2O_3 film.

Dioxane is used as a *solvent* in the formulation of inks, coatings, and adhesives, and as solvent for extracting animal and vegetable oils. As a *chemical intermediate,* dioxane reaction products are useful as insecticides, herbicides, plasticizers, and monomers. The oxonium complexes of dioxane with salts, mineral acids, halogens, and sulfur trioxide are used as *catalysts* and as *reagents* for anhydrous acid reactions, brominations, and sulfonations. In the laboratory, dioxane is useful as a *cryoscopic solvent* for molecular mass determinations and as a *stable reaction medium* for diverse reactions. Historically, dioxane is reported to be valuable in *polymer manufacture* and as a *solvent* for natural and synthetic resins, including cellulose derivatives, polyvinyl acetal resins, acrylonitrile – methyl vinyl pyridine copolymers, and others.

9. Economic Aspects

Recognized *manufacturers* of dioxane are

BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
CPS Chemical Co., Old Bridge, New Jersey, United States
Dow Chemical Co., Terneuzen, The Netherlands
Grant Chemical Co., Baton Rouge, Louisiana, United States
Osaka Organic Chemical Industries Co., Japan
Sanraku-Ocean Co., Japan
Toho Chemical Industry Co., Japan
Ugine-Kuhlmann, Frankfurt, Federal Republic of Germany
Union Carbide, South Charleston, West Virginia, United States

Capacity of known producers was estimated to be 11 000 – 14 000 t in 1985.

10. Toxicology and OccupationalHealth

The *acute toxicity* of dioxane is relatively low. The oral LD_{50} (rats) is 5170 mg/kg; by skin absorption, the LD_{50} (rabbits) is 7600 mg/kg. The LC_{50} (rats) with a 4-h exposure is 14 260 ppm (51 880 mg/m^3) [21]. *Acceptable occupational standards* are as follows:

American Conference of Governmental Industrial Hygienists (ACGIH), TLV–TWA	25 ppm (90 mg/m^3) [22] "skin" notation
U.S. Occupational Safety and Health Administration (OSHA)	100 ppm (360 mg/m^3)
Deutsche Forschungsgemeinschaft (MAK)	>50 ppm (180 mg/m^3) [23] IIIB notation: suspected of carcinogenic potential
Sweden (1975)	25 ppm (90 mg/m^3)
USSR (1980)	3 ppm (10 mg/m^3)

Processes that employ dioxane should be designed to protect workers against skin and eye exposure, as well as excessive vapor inhalation. The odor warning properties of dioxane are not adequate to afford protection. Exposure to high concentrations may cause respiratory irritation and damage to the kidney and liver.

Early, chronic *animal testing* (rodents) at extremely high doses (7 – 18 000 mg/kg in water) resulted in nasal and liver tumors and in liver and kidney damage. A follow-up study with exposures at 10 000, 1000, and 100 mg/kg in water confirmed the earlier results at high dose, but the lower dose testing showed definite indications of a threshold [24]. Animals exposed to 1000 mg/kg in water suffered varying degrees of kidney and liver injury but no treatment-related tumors. The animals dosed with 100 mg/kg (in water) experienced no adverse health effects.

A number of tests have been employed to determine whether or not dioxane caused these nasal and liver *tumors* by direct interaction with deoxyribonucleic acid (DNA) or by some other mechanism. Included were mutagenicity tests on microorganisms to observe if genetic changes occurred, e.g., the Ames bacterial mutagenicity test and a test to detect DNA repair in liver cells [25], [26]. Further tests using relatively high doses of ^{14}C-radiolabeled dioxane were conducted; the liver DNA was isolated and measured for trace quantities of radioactivity from DNA–dioxane reaction products [26]. Strong positive responses in these tests would have provided evidence of possible carcinogenicity. However, no mutagenicity or DNA reaction products were discovered [25], [26]. Therefore, dioxane may cause tumors in rodents via an alternate mechanism, possibly by recurrent tissue damage [26].

Pharmacokinetic and metabolic studies using ^{14}C-radioactive dioxane demonstrated that low doses of dioxane are eliminated in the urine as 2-hydroxyethoxyacetic acid (HEAA) [27–30]. Conversion of dioxane to HEAA is nearly complete, as indicated by the observed 118 : 1 ratio of HEAA to dioxane; the reaction is rapid, with a dioxane half-life of ca. 1 h. However, the process has a saturation point, and when this is exceeded by a high dose, elimination requires much more time.

The threshold effect is further supported by a 2-year inhalation study conducted on rats at a concentration of 111 ppm [31]. Unlike the high dose drinking water studies, no adverse health effects were observed, including no increase in tumor incidence.

Human exposures, estimated to have been at 470 ppm, have resulted in deaths. Three epidemiologic studies have been made of plant workers exposed to concentrations of up to 51 ppm. Each of these studies found no evidence of any adverse health effects.

These data were reviewed by the ACGIH with the following conclusion: "... In view of the findings of tumors in both liver and lung at or near the 10 000 ppm dietary level, and the lack of such findings at inhalation exposure levels slightly above 100 ppm for two years, dioxane has been classed by the TLV Committee as an animal tumorigen of such low potential as to be of no practical significance as an occupational carcinogen at or around the former TLV of 50 ppm." The carcinogenicity of dioxane at high-dose exposures appears to be the result of repeated tissue injury. Therefore, controlling exposures to levels below those that would cause organ injury should prevent any significant risk of cancer.

11. References

1. M. Sainsbury in S. Coffey (ed.): *Rodd's Chemistry of Carbon Compounds,* 2nd ed., **vol. 4,** Elsevier Scientific Publ., Amsterdam 1978, part H, p. 382.
2. W. Stumpf: *Chemie und Anwendungen des 1,4-Dioxans,* Verlag Chemie, Weinheim 1956, p. 33.
3. J. F. Coetzee, T. H. Chang, *J. Pure Appl. Chem.* **57** (1985) no. 4, 633 – 638.
4. Dow Chemical, US 2 195 386, 1940 (E. C. Britton, H. R. Slagh).

5. Dow Chemical, US 2 843 602, 1958 (N. B. Lorette).
6. Dow Chemical, US 2 887 491, 1959 (A. E. Gurgiolo).
7. J. Boeseken et al., *Recl. Trav. Chim. Pays-Bas* **50** (1931) 909.
8. Olin Mathieson, US 2 756 240, 1956 (M. J. Astle, W. C. Gergel).
9. M. MacLeod, *J. Chem. Soc.* **50** (1928) part 2, 3092.
10. L. M. Smorgonski, Y. L. Goldfarb, *J. Gen. Chem. USSR* **8** (1938) 1516. *Chem. Abstr.* **33** (1939) 4593.
11. Socony Mobil Oil, US 2 743 281, 1956 (C. F. Feasley).
12. Du Pont, US 2 433 844, 1948 (W. E. Hanford).
13. H. Schecker, W. Kochler, B. Sander, US 3 825 568, 1974.
14. Mathieson, CA 540 278, 1957 (M. J. Astle, J. A. Zaslowsky).
15. L. T. Cupitt, EPA report no. 600/S3-80-084, 1980.
16. G. M. Klecka, S. J. Gonsior, *J. Hazard. Mater.* **13** (1968) 161 – 168.
17. W. E. Schmidt: *Reagent Chemicals,* 5th ed., Amer. Chem. Soc., Washington, D.C., 1974, pp. 237 – 239.
18. CFR 29, U.S. Hazard Communication Standard, 1910.1200, Washington, D.C., 1983.
19. W. L. Archer, *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1979) no. 2, 131 – 135.
20. Dow Chemical, US 2 811 252, 1957 (H. J. Batchel).
21. V. K. Rowe, M. A. Wolf in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 C,** Wiley-Interscience, New York 1981, pp. 3947 – 3956.
22. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* ACGIH, Cincinnati, Ohio, 1986 – 1987.
23. DFG (ed.): *Maximum Concentrations at the Workplace (MAK),* VCH Verlagsgesellschaft, Weinheim 1986.
24. R. J. Kociba, S. B. McCollister, C. Park, T. R. Torkelson, *Toxicol. Appl. Pharmacol.* **30** (1974) 275 – 286.
25. S. Haworth, T. Lawlor, K. Mortelmans, W. Speck et al., *Environ. Mutagenesis* **5** (1983) Suppl. 1, 3 – 142.
26. W. T. Stott, J. F. Quast, P. G. Watanabe, *Toxicol. Appl. Pharmacol.* **60** (1981) 287 – 300.
27. J. D. Young, W. H. Braun, L. W. Rampy, M. B. Chenoweth et al., *J. Toxicol. Environ. Health* **3** (1977) 507 – 520.
28. J. D. Young, W. H. Braun, P. J. Gehring, *J. Toxicol. Environ. Health* **4** (1978) 709 – 726.
29. W. H. Braun, J. D. Young, *Toxicol. Appl. Pharmacol.* **39** (1977) 33 – 38.
30. F. K. Dietz, W. T. Stott, J. C. Ramsey, *Drug Metab. Rev.* **13** (1982) no. 6, 963 – 981.
31. T. R. Torkelson, B. K. J. Leong, R. J. Kociba, W. A. Richter et al., *Toxicol. Appl. Pharmacol.* **30** (1974) 287 – 298.

Diphenyl → Hydrocarbons
Diphenylamine → Aniline
Diphenylmethane → Hydrocarbons
Direct Dyes → Dyes, General Survey

Disinfectants

HANS-P. HARKE, Schülke & Mayr, Hamburg/Norderstedt, Federal Republic of Germany

1. Introduction 1
2. Definitions 2
3. Application 3
4. Efficacy Testing Methods, Requirements for Efficacy 3
5. Formulations 4
6. Active Ingredients 5
6.1. Alcohols 5
6.2. Aldehydes 6
6.3. Peroxy and Peroxo Acids 7
6.4. Phenols 8
6.5. Cation-Active Compounds 10
6.6. Amphoteric Compounds 12
6.7. Alkylamines 12
6.8. Inorganic Compounds 13
7. Toxicology 14
8. References 15

1. Introduction

Systematic pioneering studies in the 19th century, especially those of LOUIS PASTEUR (1822 – 1895) and ROBERT KOCH (1843 – 1910), showed that diseases are caused by microorganisms. Simultaneously, microorganisms were found to cause spoilage of food, cosmetics, drugs, and other materials. Preventing microorganisms from causing disease or destroying products led to the concept of disinfection.

Even prior to Koch's studies, IGNAZ PHILIP SEMMELWEIS, working in Vienna, Austria in 1847, dramatically reduced maternal mortality in his hospital when he ordered physicians to wash their hands with chlorinated lime prior to assisting in birth. Although misunderstood by many of his colleagues, SEMMELWEIS was successful in introducing hand disinfection without realizing the importance of bacteria [1], [2].

JOSEPH L. LISTER, who was influenced by Pasteur's studies, was equally successful in England in 1867, when he introduced the use of phenol solutions for treating skin wounds, thereby preventing severe infection [3].

Thereafter, successes followed rapidly. In 1888, PAUL F. FÜRBRINGER introduced the use of alcohol for disinfecting hands. The first proprietary disinfectant was Lysol (Schülke & Mayr), which appeared on the market in 1892 and proved its efficacy during the 1893 cholera epidemic in Hamburg, Germany. The disinfective properties of formaldehyde were recognized ca. 1900. This was followed in 1912 by the introduction of less toxic phenolic compounds, which superseded the use of cresols and tar oils for more delicate disinfection procedures.

Consequently, disinfectants and antiseptics rank high and they are becoming more and more important as protection against infections in the medical field and for production hygiene, particularly in food processing.

The discovery and development of antibiotics in the second half of this century seemed to have banned the threat posed by infectious diseases so that the significance of disinfectants rapidly declined for physicans and in public opinion. This development went the wrong way and has changed meanwhile.

1) All over the world nosocomical infections (i.e., in hospital acquired infectious diseases) play an increasingly important role not only regarding mortality and morbidity of patients but also in view of economy. The German Hospital Society reported 638 000 – 706 000 cases of nonsocomial infections in 1987 for Germany on average. These patients had to stay in hospitals 10 days longer than normal [4]. In the United States additional costs of US $ 2100 per nosocomical infection have been calculated [5]. In Germany it is estimated that nowadays of all patients being hospitalized for different reasons 1×10^6 suffer from nosocomical infections in addition to their original disease. The most important prophylactic measure is the compliance with hygienic and disinfection plans.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_551

2) Many well known pathogens which were supposed to have been overcome have developed a resistance against antibiotics. Infectious diseases caused by multiresistant microorganisms are difficult to treat. Their transmission can be restricted or inhibited by specific disinfection measures [6].
3) New infective agents are appearing or identified for the first time, e.g., *Helicobacter pyloris* causing gastritis or gastric carcinoma, Hepatitis-C-virus against which vaccination is still not possible, or *Escherichia coli* type EHEC (*E. coli* 0157:H7), causing the hemolytic uremic syndrom through food poisoning [7]. A careful hygiene and disinfection help to significantly reduce the danger emanating from such pathogens.

2. Definitions

The goal of *disinfection procedures* is to prevent transmission or spread of organisms that cause disease or spoilage. The following definition is frequently quoted in international publications: "Disinfection is the specific elimination of undesirable microorganisms for the purpose of preventing their transmission by interfering with their structure or metabolism irrespective of their functional state" [8]. This rather broad statement has been replaced by more manageable definitions, all of which are basically similar but which address only the medical aspect of this term. Disinfection has been defined as:

1) killing or inactivating all disease-causing organisms by chemical agents or physical procedures (Roche Lexikon, 1984) [9]
2) removing or destroying pathogenic microorganisms (Oxford Companions to Medicine, 1986) [10]
3) killing or irreversibly inactivating all causative agents of transmissible diseases [11].

In some circumstances, notably in the United States, the term "disinfectant" is customarily used only for active agents applied to inanimate objects. According to the EPA definition: "Disinfectants destroy or irreversibly inactivate infectious or other undesirable bacteria, pathogenic fungi, or viruses on surfaces or inanimate objects." Agents for application to living tissue are referred to as "antiseptics." In this article, "disinfectant" refers to active agents for both living tissue and inanimate objects.

Common to all definitions is the fact that disinfection is directed only against organisms that are considered harmful. The irreversible inactivation of the causative agent is considered to be essential by most definitions, except the British one, which also permits removal. Reber [8] is the only one to extend the definition of disinfection to include product protection.

Distinction should be made among disinfection, sterilization, and preservation. *Sterilization* refers to the complete destruction or irreversible inactivation of all microorganisms. *Preservation* is aimed at inhibiting multiplication of microorganisms in materials that are likely to spoil, with activity extending over a relatively long period.

Applying disinfectants means a careful selection of the range of microorgansims that will be inactivated by the corresponding preparation. The contained active agents as well as use concentrations and contact times are decisive.

In general it can be stated that disinfectants are formulated to inactivate bacteria and fungi. Efficacy against viruses, in particular against uncovered viruses, can be limited. Not all preparations inactivate bacterial spores, high concentrations of active agents are absolutely necessary.

3. Application

The aim of a disinfection measure will only be achieved if an appropriate disinfectant for the intended purpose is chosen observing the recommended use concentration and time and provided all parts of the body or the object to be treated come into contact with the preparation or its solution.

Typical application fields for disinfectants are in hospitals and other medical institutions in veterinary medicine; in many industrial areas for maintenance of a good production hygiene (e.g., in the pharmaceutical, food processing and cosmetics industry, in catering establishments, etc.); and in public facilities (swimming pools, prisons, barracks etc.).

Disinfection is important in agriculture for disinfection of barns used for intensive animal

husbandry and disinfection of working surfaces and equipment used in greenhouses (to maintain a monoculture free of outside germs).

In public services, sanitation of drinking water and swimming pools must be considered.

In private households disinfection measures may be advisable for the care of seriously ill patients or for food processing.

Within the mentioned application fields specific disinfection procedures with different types of preparations are necessary for different purposes. What is meant by "different purposes" may be demonstrated in the field of medical applications.

1) Skin and mucous membrane disinfection of patients, e.g., before surgical treatment, injections, punctures, catheterization of the bladder, or for antiseptic wound treatment. With these measures a transmission of microorganisms in primary sterile areas will be avoided.
2) Hygienic hand disinfection to inactive microorganisms on the hands of medical staff to avoid the transmission of microorganisms from one patient to the other via the hands or from contaminated surfaces and equipment to patients via the hands.
3) Surgical hand disinfection is executed before surgical treatment by the operating team before gloves are put on. The endogenic mixed flora, too, must be inactivated in order to prevent microorganisms from entering the opened body during operation in case the gloves are defect.
4) Instrument disinfection by immersion of used instruments into disinfection solutions with the aim to protect the staff against infections before further handling (e.g., cleaning, sterilization, etc.). Patients are also protected if thermolabile material is handled (e.g., endoscopes).
5) Thermochemical instrument disinfection in (washing) machines at 50 – 60 °C instead of immersion into disinfection solutions.
6) Surface disinfection by rubbing and wiping the surfaces of treatment tables in consulting rooms, the surfaces of equipment for treatment and diagnosis, etc., or floors.
7) Laundry disinfection, disinfection of excrements and other special applications.

4. Efficacy Testing Methods, Requirements for Efficacy

Up to now many countries have their own test methods to evaluate a disinfectant. These methods may be found compiled in national guidelines for normalization or in guidelines published by scientific associations. An example are the test methods of the DGHM (German Society for Hygiene and Microbiology) [12]

Suspension tests and germ carrier tests are most important for testing disinfectants.

In *suspension tests* a suspension of the microorganisms is mixed with the disinfectant in a given test concentration. After the desired exposure time the effect of the disinfection is interrupted by suitable methods (dilution or chemical neutralization), surviving germs are counted and deducted from the amount of germs that remained without any influence of the disinfectant. From both figures the logarithmic germ count reduction is calculated. The suspension test also allows to determine the influence of disturbing substances such as blood, or albumin on the efficacy.

Germ carrier tests allow to simulate practice-relevance conditions. Microorganisms are brought onto germ carriers made of glass, metal, plastic, textiles, etc. After drying, the carriers are placed into the disinfection solution or they are wetted with the solution. The further procedure is similar to the suspension test.

For hand, skin or mucous membrane disinfectants tests on the corresponding epithelia are carried out. These tests have to be organized like chemical trials.

Since 1990 a European group of experts associated to CEN (European Committee for Standardization) in Brussels has been establishing European standards for testing disinfectants and antiseptics and an obligatory profile for tests and requirements. After tests in phase 1 where it is evaluated whether a formulation has disinfecting properties at all, the suitability for a particular application field is investigated in phase 2.

Corresponding standards have already been released for the application field hand disinfection in the medical area. Standing for all other test methods, they are introduced as follows: In a suspension test EN 12054 the bactericidal effect

is investigated against a range of four bacterial strains in short exposure times. These bacteria represent all other bacteria. If the test requirements are met, the efficacy on hands is investigated in a test under practical conditions. In case the formulation is intended for hygienic hand disinfection, the hands of volunteers are heavily contaminated with *E. coli* and afterwards treated with the disinfectant or with 2-propanol (60 vol %) in a reference procedure (EN 1500). The obtained germ count reductions are compared. The test preparation must be at least as effective as 2-propanol in the reference procedure.

The efficacy of a formulation for surgical hand disinfection is tested on the normal hand, i.e., without special contamination and is compared with a reference procedure using 1-propanol (60 vol %) EN 12791.

For the application field hygienic hand disinfection two different methods have been accepted throughout Europe. In Germany, Austria, and in parts of Switzerland alcohol preparations are rubbed into the hands, in other countries aqueous washing preparations with a bactericidal effect are used. The aforementioned EN 1500 refers to such rubb-in preparations. EN 1499 is effective for washing preparations. In the reference procedure (EN 1499) potassium soap is used. The test preparation must have a better effect than the soap in a test under practice-relevance conditions on the hand. If necessary, efficacy tests against mycobacteria, fungi, or viruses must also be added. Corresponding standards are being worked out.

For all other application fields for disinfectants and antiseptics test methods and requirement profiles have also been issued as ENs or are being developed.

All valid standards can be obtained at DIN (in German language), BSI (in English language) or at AFNOR (in French language).

5. Formulations

Disinfectant formulations mostly contain a combination of active ingredients to satisfy user demands, and to fulfill the appropriate test procedures. In addition, additives must be incorporated to achieve good skin tolerance, materials compatibility, cleaning power, and other goals (see Table 1).

The specific spectrum of effectiveness of the active ingredients is discussed in Chapter 6. However, because closely related chemical substances commonly exhibit marked differences in potency, referring globally to aldehydes, phenols, etc. as disinfectants is inappropriate.

Currently, the preferred active ingredients are from the following classes of chemicals: aliphatic and aromatic alcohols, aldehydes, peroxy acids, phenols, cation-active compounds, amphoteric surfactants, alkylamines and inorganic oxidizing agents.

The benefits and risks associated with commercial products can only be assessed on the basis of test results for the complete preparation. Definitive conclusions are impossible if they are based solely on the properties of individual ingredients.

Examples of some preparations used for different applications are given in Table 2. Some preparations are used without further dilution, whereas others must be diluted prior to use. Many products contain active ingredients from several different chemical classes.

6. Active Ingredients

6.1. Alcohols

Ethanol [*64-17-5*] and the two propanols (1-propanol [*71-23-8*] and 2-propanol [*67-63-0*]) are essentially the only aliphatic alcohols used, even though they are effective only at high concentration. Their advantage lies in their water miscibility, rapid evaporation, and favorable pharmacologic and toxicologic properties [13]. Higher alcohols are more effective as antimicrobial agents, but their low volatility, unpleasant odor, and danger of resorption exclude their use.

The concentration of primary alcohol effective within 2 min against gram-negative and gram-positive bacteria, when tested in a suspension test, is as follows [14]:

Chain length	C_1	C_2	C_3	C_4	C_6	C_{12}
Concentration, %	60	50	20	10	5	15

Maximum effectiveness is reached with alcohols having six to eight carbon atoms. 2-Ethylhexanol [*104-76-7*] appears to be the most

Table 1. Common disinfectant additives

Additive	Disinfectant use [a]			
	Hand	Skin	Instrument	Surface
Germicides	+	+	+	+
Germistatics	+	+	−	(+)
Surfactants [b]	−	−	+	+
Wetting agents	+	+	−	−
Anticorrosives	−	−	+	+
Indicator dyes	−	+/−	−	(+)
Fragrances [c]	+	−	(+)	(+)
Solvents	+	+	+	+

[a] Key: + Useful; −Unnecessary; (+) Occasionally useful.
[b] Cleansing agents.
[c] Fragrances.

Table 2. Typical disinfectants for various applications

Application and trade names	Chemical class	Producer	Application form
Hand disinfection			
Desderman N	alcohol – phenol	Schülke & Mayr	undiluted
Sterillium	alcohol – cationic	Bode Chemie	undiluted
Skin disinfection			
Betaisodona solution	PVP ∗ – iodine	Mundipharma	undiluted
Braunol 2000	PVP ∗ – iodine	Braun Melsungen	undiluted
Cutasept	alcohol – cationic	Bode Chemie	undiluted
Kodan Tinktur forte	alcohol – phenol	Schülke & Mayr	undiluted
Spitaderm	alcohol – cationic	Henkel	undiluted
Mucous membrane antiseptics			
Betaisodona (range)	PVP ∗ – iodine	Mundipharma	undiluted
Octenisept	Octenidine – aromatic alcohol	Schülke & Mayr	undiluted
Medical instrument disinfection			
Alhydex compact	aldehyde	Johnson & Johnson	undiluted
Desoform	aldehyde	Lysoform	diluted
Gigasept FF	aldehyde	Schülke & Mayr	diluted
Kohrsolin ID	aldehyde	Bode Chemie	diluted
Korsolex FF	aldehyde – amine	Bode Chemie	diluted
Lysetol AF	aromatic alcohol – cationic – alkylamine	Schülke & Mayr	diluted
Sekusept plus	alkylamine – aminoacid	Henkel	diluted
Surface disinfection			
Buraton 10 F	aldehyde	Schülke & Mayr	diluted
Dettol	phenol	Reckitt & Colman	diluted
Incidin perfekt	aldehyde – cationic	Henkel	diluted
Kohrsolin	aldehyde	Bode Chemie	diluted
Lysoformin	aldehyde	Lysoform	diluted
Perform	peroxy acid	Schülke & Mayr	diluted
Terralin	aromatic alcohol – cationic	Schülke & Mayr	diluted

∗ PVP = Polyvinylpyrollidone

effective. When combined with solvents and emulsifiers, a 0.2 % solution of 2-ethylhexanol kills gram-negative organisms within 15 min [15]. Despite its odor, these favorable properties have led to limited use of 2-ethylhexanol in combination with other agents in concentrated surface disinfectants that are diluted with water before use.

The most important application of short-chain alcohols is for disinfecting skin and hands. The alcohols are mostly combined with substances having germistatic activity. These alcohols are not suitable for disinfection of mucous membranes. Another application is for surfaces that must be disinfected rapidly, e.g., examination tables and working surfaces.

The maximum potency of ethanol is as a 70 % aqueous solution. For example, in a germ carrier test using gauze as the substrate, *Staphylococcus aureus* is killed by a 50 or 80 % solution within 10 min, but is killed within 5 min by a 60 – 75 % solution [16]. The alcohols are ineffective against spores of bacteria. Ethanol in concentrations below 75 – 80 %, 2-propanol, and 1-propanol are ineffective against unenveloped viruses like the polyvirus as a representative for

enteroviruses. However, lipophilic viruses like the herpesvirus are rapidly and effectively inactivated by all three alcohols [17], [19]. The effectiveness of an ethanol-based hand disinfectant against polio virus within 10 – 60 s has been proven in a clinical trial. 2-propanol (60 vol %) in comparison was absolutely ineffective even after 60 s [18].

Among aromatic alcohols, benzyl alcohol [*100-51-6*], $C_6H_5CH_2OH$, phenethyl alcohol [*60-12-8*], $C_6H_5CH_2CH_2OH$, and phenoxyethanol [*122-99-6*], $C_6H_5OCH_2CH_2OH$, and the mixture of isometric phenoxypropanols (1-phenoxypropanol-2 [*770-35-4*] and 2-phenoxypropanol-1 [*4169-04-4*] find limited application, generally in combination with other active agents. These compounds have germicidal as well as germistatic activity (Table 3). They are more toxic than short-chain aliphatic alcohols [20]. An overview of the properties of alcohols as disinfectants has been published [21].

6.2. Aldehydes

Certain aldehydes have proved to be excellent agents for disinfecting instruments and surfaces. However, aldehydes are not used as skin disinfectants mainly because their local tolerance is unacceptable. Within the large group of aldehydes only very few have microbiocidal properties. For example, succinaldehyde and glutaraldehyde are extremely effective whereas adipaldehyde and malonaldehyde are ineffective. On the other hand the few active aldehydes are effective against a broad spectrum of infective agents provided the contact times and concentrations are chosen correctly.

The most commonly used aldehydes are formaldehyde [*50-00-0*], HCHO; succinaldehyde [*638-37-9*], $HCOCH_2CH_2CHO$; glutaraldehyde [*111-30-8*], $HCO(CH_2)_3CHO$; 2-ethylhexanal [*123-05-7*],

$CH_3(CH_2)_3CH(CH_2CH_3)CHO$; and glyoxal [*107-22-2*], HCOCHO. As with alcohols, the most effective monoaldehydes have seven or eight carbon atoms, but their pungent odor limits their use. In this connection formaldehyde must be regarded differently. This might be due to its pronounced tendency to form oligomeres on the one hand and hydrates on the other hand. Thus, marked differences in efficacy are noticeable when the results of surface and suspension tests are compared [15]. In the suspension test, a 1 % solution of formaldehyde is effective in less than 30 min against a wide range of fungi and bacteria and, thus, is considerably less effective than, e.g., 2-ethylhexanal, a 0.1 % solution of which kills all microorganisms within 2.5 min. On the other hand, a 0.3 % solution of formaldehyde kills *Staphylococcus aureus* on surfaces within 4 h, whereas an equally strong solution of 2-ethylhexanal is inadequate even after contact for 6 h.

These results clearly demonstrate the advisability of combining active ingredients to formulate effective and acceptable preparations. This is especially important because the effects of combination preparations are frequently not only additive, but also synergistic. For example, a product (Buraton 10 F), which at its working strength contains 0.045 % formaldehyde, 0.01 % glutaraldehyde, 0.0025 % 2-ethylhexanal, and 0.07 % glyoxal, meets the DGHM specification for a surface disinfectant because it is effective within 1 h both in the suspension test and the surface test. Often aldehydes are combined with cationic compounds in commercial preparations. So the concentration of aldehydes can be lowered and the acceptability is increased (reduction of smell).

Among the dialdehydes, glyoxal was ineffective against the usual spectrum of microorganisms in the suspension test (30 min exposure, 1 % solution). On the other hand, a 0.2 % solution of succinaldehyde or glutaraldehyde was effective within 15 and 2.5 min, respectively (if the effect on *Candida albicans*, which required a longer contact time, is excluded). The same concentration was also effective within 1 h against *Staphylococcus aureus* on surfaces; for glyoxal this effect required only a 0.1 % solution. Therefore, glyoxal is useful as a surface disinfectant only when combined with other active ingredients.

Aldehydes are effective virucidal agents against lipophilic viruses and enteroviruses, such as the Coxsackie viruses or polioviruses. They are to some extend less effective against papovaviruses such as the SV 40 virus [22], [23].

Succinaldehyde and glutaraldehyde have a low tendency to be adsorbed onto plastic materials. This property, combined with their gen-

Table 3. Effective concentration and contact time to kill germs with aromatic alcohols [15]

Aromatic alcohol	Concentration, %	Contact time, min			
		Escherichia coli	*Pseudomonas aeruginosa*	*Proteus mirabilis*	*Staphylococcus aureus*
Benzyl alcohol	1	>30	>30	>30	>30
Phenethyl alcohol	1.25	2.5	2.5	2.5	>30
	2.5	2.5	2.5	2.5	5
Phenoxyethanol	1.25	15	2.5	2.5	>30
	2.5	2.5	2.5	2.5	>30

erally broad activity and excellent materials compatibility, has led to their use as agents capable of disinfecting (contact period 15 min) or almost sterilizing (contact period 10 h) delicate thermolabile instruments, e.g., endoscopes. The best known products are Gigasept/Gigasept FF (succinaldehyde – dimethoxytetrahydrofuran) and Alhydex – Cidex (glutaraldehyde). In U. S. journals such preparations are defined as "sterilizer". The treatment of an instrument with the preparation (immersion, rinsing, drying) does not guarantee sterility of the instrument. Therefore, this term is not used in Europe, and the term "disinfection" is preferred. A survey on the properties of glutaraldehyde is given in [24].

6.3. Peroxy and Peroxo Acids

Peroxycarboxylic acids and inorganic peroxo acids contain active oxygen and are extremely effective disinfectants. Because these compounds also act as strong oxidants, formulating stable, liquid preparations that contain these active ingredients combined with additives that have cleaning, skin protective, anticorrosive, or other properties is difficult.

Peroxycarboxylic acids may be prepared in solution from reactive carboxylic acid derivatives (e.g., acid chlorides, acid anhydrides, and possibly esters or amides) and excess H_2O_2. The direct reaction of carboxylic acids with H_2O_2 requires the presence of a strong mineral acid such as sulfuric acid. Reviews on peroxycarboxylic acids as disinfectants have been published [25], [26].

Solid reactive acid derivatives can be mixed with solid hydrogen peroxide sources (e.g., peroxycarbonates, peroxyborates, or similar substances). In this case, the peroxycarboxylic acid is not formed until water is added. Other additives may be included in such a formulation, because the final solution must remain stable for only a few hours. Such solutions also exhibit a broad spectrum of antimicrobial activity [27], [28].

Aliphatic Peroxycarboxylic Acids.

Peracetic acid [*79-21-0*], CH_3CO_3H, was described as a germicidal substance as early as 1902 [29].

In spite of its effectiveness at low concentration against all types of microorganisms (Table 4), peracetic acid has not found wider acceptance because of its corrosive properties, penetrating odor, the danger of explosive reactions, and its limited stability. There are also some discussions about possible toxicity of peracetic acid [30], [31]. Peracetic acid is used for disinfection of artificial kidneys, linen, for treating ion exchange resins or for use in food industry.

Monoperoxyglutaric acid [*3851-97-6*], $HO_2C(CH_2)_3CO_3H$, exhibits a broad spectrum of antimicrobial activity. The compound is odorless, essentially nonvolatile, and quite stable in solution [35].

Monoperoxysuccinic acid [*3504-13-0*], $HO_2CCH_2CH_2CO_3H$, is generated only during the preparation of the aqueous application solution as described previously. The compound is relatively unstable. It inactivates viruses not only by capsid alteration, but also by RNA fragmentation [36].

Peroxyformic acid and peroxypropionic acid resemble peracetic acid in their effectiveness, but both are difficult to apply and are used only occasionally.

Aromatic Peroxycarboxylic Acids. These compounds are frequently much more stable than the aliphatic acids. However, their low water solubility and limited stability in alcohols have prevented their practical application. The antimicrobial properties of some members of this class, e.g., peroxybenzoic acid, peroxyanisic

Table 4. Contact time to kill germs with four dilutions of peracetic acid [33], [34]

Microorganism	Contact time, min			
	0.5 %	0.05 %	0.005 %	0.0005 %
Bacillus subtilis (spores)	1	>30	>30	>30
Escherichia coli	1	1	1	10
Staphylococcus aureus	1	1	5	10
Trichophyton mentagrophytes	5	>30	>30	>30

acid, and chloroxyperbenzoic acid, have been described [37].

Monoperoxyphthalic acid is commercially available as the hydrated magnesium salt [38]. The compound is used in a surface disinfectant which is a white granulate.

Inorganic Peroxo Acids. Peroxomonosulfuric acid [*7722-86-3*], HO_3SOOH, and its alkali-metal salts have been found to be as effective as organic peroxy acids . These compounds also have good overall effectiveness against viruses [39]. Viruses including papovaviruses which are resistant to other disinfectants are deactivated. The absence of toxic vapors that might endanger personnel or patients is noteworthy. However, peroxomonosulfuric acid can oxidize chloride to chlorine. So, if the solutions are prepared with tap water which contains relatively high amounts of chloride, the chlorine smell may influence the user in favor of the compound.

6.4. Phenols

Phenol [*108-95-2*], C_6H_5OH, the simplest member of this class, was introduced as an antiseptic in 1867 by Joseph L. Lister. Systematic studies were undertaken around the turn of the century to determine the effect of substitution on disinfectant properties [40], [41]. Alkyl substituents were found to increase efficacy, with those having 5 – 7 carbon atoms being the most effective. Substitution by chlorine or bromine exerted an even greater effect. These relationships have been discussed in detail [42].

In general, phenols are effective germicidal and germistatic agents against bacteria, fungi, and lipophilic viruses, but are not effective against spores and hydrophilic viruses. They are markedly more effective in a weakly acidic rather than in an alkaline environment. However, their water solubility under acidic conditions is low. Surfactants may impair the efficacy of phenols.

Phenols are frequently incorporated into commercial soap concentrates; the diluted solution is used to disinfect surfaces and instruments, and occasionally skin and hands. Phenol derivatives are also incorporated as the active germistatic agents in alcohol-based products.

Increased discussions on pharmacological, toxicological, and ecological properties of phenols have reduced use to relatively few, preferably nonhalogenated compounds. This is especially true for Middle European and Scandinavian countries, but not for English speaking countries. Polychlorinated phenols, such as pentachlorophenol and even hexachlorophene, are no longer used.

The efficacy and pharmacologic and chemical properties of the phenols have been described in detail [43].

Nonhalogenated Phenols. *2-Phenylphenol* [*90-43-7*], $C_6H_5C_6H_4OH$, M_r 170.21, (2-hydroxybiphenyl, 2-hydroxydiphenyl, *ortho*-phenylphenol) *mp* 55.5 – 57.5 °C, *bp* 280 – 284 °C, vapor pressure 70 Pa at 100 °C, solubility in water 0.07 g in 100 g, forms colorless to pale yellow crystalline flakes.

This substance is more effective against gram-negative bacteria and fungi than against gram-positive cocci (Table 5) [44]. No effect was observed on enteroviruses when exposed for 10 min at concentrations up to 12 % [45]. However, adenoviruses and vacciniaviruses were rapidly inactivated by 0.12 % solutions. A combination of 2 % 2-phenylphenol, 18 % 1-

Table 5. Germicidal concentration after 30 min contact with 2-phenylphenol

Microorganism *	Germicidal concentration, %		
	pH 4.5	pH 7	pH 8.5
Gram-negative bacteria	0.005 – 0.01		0.0125 – 0.025
Staphylococcus aureus	0.1		0.25
Candida albicans	0.01 – 0.0125	0.025	

* Initial concentration 10^9/mL.

propanol – 2-propanol, and soap was effective against hepatitis B viruses [46].

Based on its well-researched and generally favorable toxicologic properties, 2-phenylphenol is currently the most widely used phenol. It is approved in many countries for treating citrus fruit. The WHO has set the acceptable daily intake to 0.2 mg/kg and limited maximum daily intake to 1 mg/kg.

Thymol [*89-83-8*], 2-[$(CH_3)_2CH$]C_6H_3-5-(CH_3) OH, M_r 150.22, 2-isopropyl-5-methylphenol, *mp* 51.5 °C, *bp* 232 °C, vapor pressure 12.7 Pa at 40 °C, solubility in water 0.098 g in 100 g at 25 °C, forms colorless crystals.

Thymol is a component of many natural essential oils and, probably for this reason, has been incorporated as a component of antimicrobial mouth and throat antiseptics. Solutions containing 0.03 – 0.05 % are effective against fungi [47].

Eugenol [*97-53-0*], 4-$(H_2C{=}CHCH_2)$ C_6H_3-2-(OCH_3) OH, M_r 164.20, 2-methoxy-4-(2-propenyl)phenol, *mp* – 9 °C, *bp* 225 °C, is a clear liquid.

Eugenol, also a natural substance, is quite effective against bacteria and fungi. Its tendency to form complexes with metal ions, its sensitivity to oxidation, and its intensive smell of cloves limit general use of this compound.

4-tert-Amylphenol [*80-46-6*],

$C_2H_5C(CH_3)_2C_6H_4OH$, M_r164.25, 4-*tert*-phenylphenol, 4-(2-methylbutyl)phenol *mp* 88 – 89 °C, *bp* 225 °C, is extremely effective against gram-positive bacteria and for this reason is still used in the United States in surface disinfectants.

Although they are quite effective, cresols and xylenols are no longer extensively used because of their toxic properties and strong odor. They were once used extensively in cresol – soap solutions as described in various pharmacopoeias.

Halogenated Phenols.

4-Chloro-3-methylphenol [*59-50-7*], $ClC_6H_3(CH_3)OH$, M_r 142.59, 4-chloro-*m*-cresol, *mp* 55.5 °C, *bp* 235 °C, vapor pressure 700 Pa at 100 °C, solubility in water 0.4 g in 100 g, is a colorless powder.

4-Chloro-3-methylphenol is currently an important chlorophenol; it is used as a commercial preservative as well as an active agent in disinfectants. In suspension tests, chlorocresol was found to be effective within 30 min against dermatophytes at concentrations of 0.005 – 0.01 % (pH 4.5) and at 0.0125 – 0.025 % against *Candida albicans*.

2-Benzyl-4-chlorophenol [*120-32-1*], $C_6H_5CH_2C_6H_3(Cl)OH$, M_r 218.69, Clorophene, *mp* 48.5 °C, *bp* 160 – 162 °C at 466 Pa, vapor pressure 13 Pa at 100 °C, solubility in water 0.007 g in 100 g, occurs as colorless crystals.

Chlorophenol is extremely effective. In the suspension test, a concentration of 0.025 % is adequate to kill all types of microorganisms within 30 min [44].

4-Chloro-3,5-dimethylphenol [*88-04-0*], $(CH_3)_2ClC_6H_2OH$, M_r 156.61, chloroxylenol, *mp* 115.5 °C, *bp* 246 °C, solubility in water 0.05 g in 100 g, occurs as colorless prisms or needles.

Chloroxylenol is used in combination with other phenols. It is more effective against gram-positive bacteria than against gram-negative bacteria [48].

Other halogenated phenols used as disinfectants include chlorothymol, 2-chloro-6-methyl-4-benzylphenol, 4-bromo-2,6-dimethylphenol, 2,3,4,5-tetrabromo-6-methylphenol, and 2,4,4′-trichloro-2′-hydroxydiphenyl ether (Triclosan).

Triclosan is used as active agent in germicidal soap.

6.5. Cation-Active Compounds

Cation-active compounds were introduced as antimicrobial agents in 1935 by GERHARD D. DOMAGK. This class of disinfectants includes compounds of widely differing structures, such as quaternary ammonium compounds and guanidinium and pyridinium derivatives. The first generation of cation-active compounds showed effectiveness against gram-positive bacteria at low concentration, low activity against gram-negative bacteria, and ineffectiveness against mycobacteria. Continuous development and testing of new structures for activity have produced the relatively well-balanced compounds currently available. However, even such substances are inactive against mycobacteria.

Benzalkonium chloride [*68391-01-5*], $[C_6H_5CH_2N(CH_3)_2C_nH_{2n+1}]^+Cl^-$, is a mixture of alkyldimethylbenzylammonium chlorides. The alkyl chains have a length of 8 – 18 carbon atoms. According to the European Pharmacopeia 1997, the mean molecular mass should be ca. 354 [49].

The effect of alkyl chain length of benzalkonium chloride on gram-negative and gram-positive bacteria has been tested in a suspension test [50]. The effective concentration against *Staphylococcus aureus* and *Pseudomonas aeruginosa* is shown in Table 6. This test indicates that gram-negative bacteria are less susceptible. The most effective benzalkonium chlorides are compounds in which the alkyl residues have 12 or 14 carbon atoms.

Benzalkonium chloride is used as an active ingredient in surface disinfectants. Its lack of activity against mycobacteria would disallow its use as an instrument disinfectant.

Benzalkonium chloride has low oral toxicity [51]. Because it is effective in low concentration, this agent is frequently chosen as an ingredient for disinfectants to be used in the vicinity of food. However, benzalkonium chloride is inactivated, as are other cation-active compounds, by many anions, including anionic detergents.

Table 6. Concentration of benzalkonium chloride effective within 10 min against *Staphylococcus aureus* and *Pseudomonas aeruginosa*

Chain length	Concentration, ppm	
	Pseudomonas aeruginosa	*Staphylococcus aureus*
C_8	6000	3000
C_{10}	1200	450
C_{12}	120	45
C_{14}	40	15
C_{16}	200	30
C_{18}	1000	450

Didecyldimethylammonium chloride [*2390-68-3*], is an interesting alternative to benzalkonium chloride. Toxicological and ecotoxicological properties have been thoroughly investigated (Lonza AG) with favorable results.

Poly(hexamethylenebiguanide) hydrochloride [*32289-58-0*] is also used as a surface disinfectant and is alleged to be suitable for skin disinfection. This compound has a slow effect and does not meet the practical requirements for prophylactic antiseptics in this respect. Although it is somewhat less effective than benzalkonium chloride (Table 7), it is sometimes used instead of benzalkonium because it is less foam producing under use conditions.

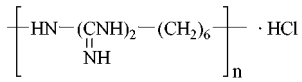

Poly(hexamethylenebiguanide) hydrochloride

Cetylpyridinium chloride [*123-03-5*] is used for treatment of mucous membranes and as throat medicine. Furthermore the substance is applied as preservative due to its excellent bacteria inhibiting property. In this respect it is of some interest that gram-negative bacteria are less sensitive by a factor of 4 than gram-positive bacteria [53]. The substance does not inactivate polioviruses or adenoviruses [54].

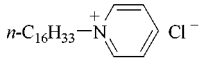

Cetylpyridinium chloride

Chlorhexidine [*55-56-1*], a bisbiguanide, is a strong base that is only slightly soluble in water (solubility 0.008 % at 20 °C). It is used worldwide, especially in skin and hand disinfectants, generally in the form of the digluconate, as a bacteriostatic. A group of experts at the German BGA summarized the available results and cer-

Table 7. Required exposure time as determined in suspension test [52]

Microorganism	Exposure time, min			
	A *	A * plus 10 % serum	B **	B ** plus 10 % serum
Staphylococcus aureus	5	5	15	15
Pseudomonas aeruginosa	30	60	15	15
Proteus vulgaris	15	15	60	60
Candida albicans	5	15	15	30

* 0.05 % Benzalkonium chloride.
** 0.05 % Poly(hexamethylenebiguanide) hydrochloride.

tified that chlorhexidine has above all an excellent bacteriostatic efficacy against gram-positive bacteria. The minimum inhibiting concentration is ≥ 1 µg/L. Significantly higher concentrations are given for gram-positive bacteria and fungi (10 to > 73 µg/mL). In the presence of blood or protein the efficacy is reduced by a factor of 100 to 1000 [55].

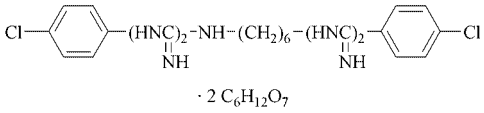

Chlorhexidine digluconate

4 % solution of chlorhexidine digluconate is frequently used in hand wash lotions. Application of such a lotion leads to ca. 90 % germ reduction, which can be increased to 98 % after 6 applications [56]. However, it does not meet specifications established nowadays in Europe for a hand disinfectant. Under the test conditions of the European Standard EN 1499 for the efficacy testing of hygienic wash preparations it turned out that a logarithmic reduction of 3.1 is reached with a 4 % chlorhexidine solution which is customary in trade. There is no significant difference in the efficacy between this solution and soap [57]. In a multicenter study performed in six European countries which included 28 hospitals it was observed that the rate of wound infection had not been influenced by total body washings with a chlorhexidine-containing detergent the night before operation. In this study more than 2800 patients were included [58]. This result may be regarded in connection with investigations showing that a 0.5 % chlorhexidine solution reduces the number of organisms by only ca. 60 % [59], [60].

Chlorhexidine has bacteriostatic effects at low concentration and, like other cation-active compounds, remains on the skin. These properties suggest a combination of chlorhexidine with alcohol. Preparations containing both agents act rapidly, and the presence of chlorhexidine assures a residual effect.

Chlorhexidine is ineffective against polioviruses and adenoviruses. The effectiveness of a 0.1 % solution against herpes viruses has not yet been established unequivocally [54].

Octenidine [*71251-02-0*], a new cation-active agent, is a bispyridine that was used for the first time in an approved disinfectant in 1987 [61], [62]. It may be even more suitable as a remanent active ingredient than chlorhexidine combined with alcohols because it is more effective and better balanced (Table 8). Residual octenidine, after skin disinfection with an alcohol-based, octenidine-containing preparation, was capable of inactivating micrococci introduced later [63].

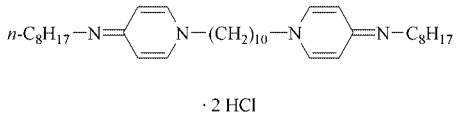

Octenidine dihydrochloride

Octenidine has been applied as active agent in a mucous membrane antiseptic since 1990 [64], [65]. Today it is one of the preparations prevalently used in many European countries for the treatment of the urogenital tract, buccal cavity [66], and wounds. Octenidine is absorbed neither through the skin nor through mucous membranes nor via wounds and does not even pass the placental barrier [67].

Other cation-active compounds that are used include polymeric quaternary ammonium salts (Buckman Laboratories or Onyx Chemical Co.), and guanidine [68].

6.6. Amphoteric Compounds

Amphoteric compounds, which are derived from glycine by replacing one amino hydrogen atom with an alkylamine group, exhibit antimicrobial activity. Solutions containing such agents are commercially available under the trade name Tego (Goldschmidt, Essen, Federal Republic of Germany) and have the following basic formula:

$RNH(CH_2)_2NH(CH_2)_2NHCH_2COOH$

Solutions of 1 % are effective within a short period against gram-positive bacteria and to a certain extent against gram-negative bacteria as well [69]. Lipid viruses are also inactivated but hydrophilic viruses, such as the poliovirus, are not [70].

The efficacy of these amphoteric compounds is less affected by a high protein burden than other agents generally are. Therefore, such compounds are well suitable for use in the food industry to disinfect working surfaces and equipment. These compounds have surface-active properties similar to those of quaternary ammonium compounds. They have low oral toxicity [71], and toxicologically significant amounts are apparently not absorbed through the skin.

Table 8. Comparison between octenidine and chlorhexidine determined by the suspension test *

	Effective concentration, %	
	Octenidine dihydrochloride	Chlorhexidine digluconate
Staphylococcus aureus	0.025	>0.2
Escherichia coli	0.025	0.1
Proteus mirabilis	0.025	>0.2
Pseudomonas aeruginosa	0.025	>0.2
Candida albicans	0.01	0.025

* Exposure time 5 min.

6.7. Alkylamines

It has already been known for quite some time that selected alkylamines have antimicrobial properties. Their activity against sulfate-reducing bacteria made them of particular interest for the oil-producing industry. Cocospropylenediamine [*61791-63-7*], e.g., is applied in this area. The importance of these compounds in other fields of disinfection increased after a successful strengthening of their antimicrobial properties by further derivatization and/or improvement of their compatibility [72].

N,N-bis-(3-aminopropyl)laurylamine [*2372-82-9*] has a surprisingly high effect against mycobacteria. This compound can be used instead of aldehydes for formulating tuberculocidal preparations [73]. 1.5 % solutions are capable to kill a great deal of mycobacteria in the presence of 0.5 % bovine albumine.

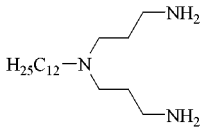

N,N-bis-(3-aminopropyl)laurylamine

Glukoprotamine [*2372-82-9*], a reaction product of cocospropylene-1,3-diamine and L-glutamine acid is present as a mixture of the amide and the corresponding ammonium salts [74]. Glukoprotamine has a well balanced microbiological effect and also kills mycobacteria. An 0.5 % solution is effective in a germ carrier test at pieces of cotton within 60 min against *Mycobacterium terrae*. Enteroviruses are not inactivated.

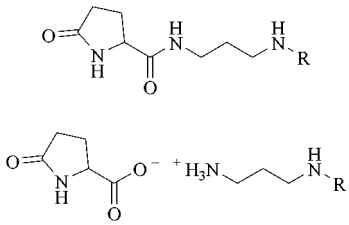

Glukoprotamine

The effectiveness of both compounds remains for cocospropylenediamine guanidinium diacetate [*67-63-0*], whereas the mycobactericidal effect is restricted.

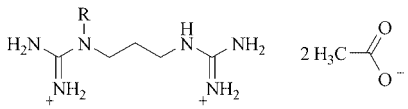

Cocospropylenediamine guanidinium diacetate

6.8. Inorganic Compounds

Chlorine. Chlorine [*7782-50-5*] is used in large quantities in some countries and has long been known as a disinfectant. The term chlorine is frequently used rather imprecisely. In the present context, the term refers generally to an aqueous solution of chlorine, hypochlorite, or hypochlorous acid. Occasionally, even chlorine-releasing compounds, such as *N*-chlorosulfonamides (e.g., Chloramin T) or *N*-chloroisocyanuric acid, and their salts are included in this group.

Aqueous solutions of chlorine and hypochlorite are extremely effective against a broad spectrum of microorganisms. Frequently, a concentration of < 1 ppm of available chlorine is sufficient to kill bacteria. Viruses are also inactivated under similar conditions [75]. Higher concentration is required to kill spores and mycobacteria.

Because chlorine is so chemically reactive, contaminants, especially organic substances, rapidly impair its effectiveness, which is one of the reasons that chlorine cannot be used universally for disinfection [76].

Other conditions may also significantly alter chlorine effectiveness. One example is pH. Spores of *Bacillus macerans* were reduced by 99 % in 8.5 min when 15 ppm chlorine was used at pH 6, but 42 min were required at pH 8 [77].

Chlorine is used mainly to disinfect drinking water, swimming pools, and food industry installations. Chlorine for household use or for wastewater treatment is questionable for ecological reasons, as the reaction between chlorine and organic material leads to the formation of undesirable organic chlorine compounds.

Iodine. Like chlorine, iodine [*7553-56-2*] has a broad spectrum of effectiveness. It was used mainly to disinfect wounds, skin, and mucous membranes. However, a number of adverse reactions made its use questionable and led to changes in its mode of application. Adverse reactions include local effects, primarily skin irritation, pain when applied to wounds, and the possibility of sensitization. In addition, systemic reactions, notably those affecting the thyroid gland, are known.

Since the mid-1950 PVP – iodine [*25655-41-8*], (providone – iodine), an iodine – polyvinylpyrrolidone complex with 10 % available iodine is being used instead of the old iodine tincture.

This change in the galenic of the iodine preparation had the effect that some of the abovementioned disadvantages of the iodine tincture were significantly reduced, including sensitization, primary skin irritation, or pain.

Around 1980 it was observed that the PVP – iodine use solutions did not have the same efficacy as diluted test solutions [79], [78]. Very soon it was discovered that this phenomenon was due to the fact that aqueous solutions of 10 % PVP – iodine contain only a few ppm of free iodine [80]. The contents of free iodine increases when diluted with water and in consequence efficacy increases, too. However, at the same time the effectiveness decreases when the PVP – iodine solution is loaded with blood, protein, etc. to an inadmissible point because the contents of free available iodine decreases.

By optimizing the galenics it was possible to increase the level of free iodine even in 10 % PVP–iodine solutions to such a point that even undiluted the desired effect was obtained [81].

The use of PVP – iodine is counterindicated in pregnant women and in newborns because the thyriod gland of a fetus or a newborn is not fully functional.

Like chlorine, iodine kills vegetative bacteria rapidly and effectively in the ppm range [82]. Organic materials severely impair its activity. Spores are inactivated only at high concentration of free iodine and longer contact times are used. The use of PVP – iodine on skin for this purpose is questionable. In addition, PVP – iodine is ineffective against enteroviruses, but is active against lipophilic viruses [83], [84].

Acids and bases. Strong alkali-metal and alkaline-earth hydroxides are occasionally employed for disinfecting excretions or as disinfectants in animal husbandry. The concentration must be chosen carefully so that the hydrolytic reactions of microorganisms as well as of organic contaminants proceed at ambient temperature. The same applies to nonoxidizing inorganic acids like sulfurous acid or SO_2.

Metals. Metals, especially heavy metals, have been used for years because of the so-called oligodynamic effect; that is, they are toxic to

bacteria at low concentration. The metals react with thiol (SH) or amine (NH) groups of enzymes or proteins, a mode of action somewhat different from that of the other active agents and one to which microorganisms may develop resistance. Such resistance may be transmitted by plasmids [85–87].

Only mercury and silver compounds are used currently, and then just for specific applications. For example, silver sulfadiazine [*22199-08-2*] is used as a prophylactic antiseptic ointment for extensive burns. An equilibrium dispersion of colloidal silver with dissolved silver ions can be used to purify drinking water at sea [88].

Phenylmercuric borate [*8017-88-7*] or acetate [*62-38-4*] were used for disinfecting mucous menbranes at an effective concentration of 0.07 % in aqueous solutions [89]. Due to toxicological and ecotoxicological reasons phenylmercury salts are no longer applied nowadays. Nevertheless, surgeons use mercurochrom [*129-16-8*] even today and despite toxicological objections. The benefit, however, may be more an adstringent effect than an antimicrobial one.

7. Toxicology

When the toxicology of disinfectants is evaluated, their widely varying methods of application must be considered. Important factors include the final use concentration, whether direct body contact is intended, and the volatility of the substance with the accompanying danger of inhalation. Because coverage of the many possible conditions would be difficult, the present discussion is restricted to a few general comments, and references to toxicologic studies of the active agents.

Short-chain alcohols are considered safe, provided they are properly applied to skin or small areas. They are not absorbed by the skin. *Phenols* used in disinfectants have commonly been well investigated with regard to their pharmacological toxicological properties. For 2-phenylphenol [*90-43-7*] cancerogenicity and teratogenicity tests are available [90]. Even in the applied high dosages neither cancerogenic nor teratogenic effects were observed. These dosages are significantly and by more than a factor 1000 higher than those amount of substances being applied for hand disinfection (Desderman N). Noteworthy is the fact that phenols are readily absorbed through the skin.

Aldehydes may cause local irritation and may elicit sensitization. A formaldehyde concentration as low as 1 ppm in air irritates the upper respiratory tract and the mucous membrane of the eye. This may be an advantageous warning that the facility needs to be ventilated. When used correctly as a surface disinfectant for washing and scrubbing floors, the maximum concentration attained for a short time is just above 0.1 ppm [91]. Concentrations of formaldehyde that cause mucous membrane carcinomas in animals are not likely be reached when appropriate disinfectants are used properly [92], [93].

Modern preparations containing *active oxygen* are characterized by low volatility. They are a genuine improvement over aldehyde-based preparations, especially considering their broad range of activity. Their acute oral toxicity is low and they cause only mild local irritation.

Cation-active and amphoteric compounds are not readily absorbed through the skin or the mucous membrane. These compounds cause local irritation and are extremely poisonous when administered parenterally [51]. Specific toxic effects have been observed occasionally as, for example, effects of chlorhexidine on hearing and balance [94].

With *iodine*, the systemic effect and the effect on thyroid function are important. With *metals*, central nervous system involvement may be a consideration.

In summary, the toxic properties of the products must be made known by the manufacturers, and proper precautions must be taken by the consumers. If directions for use are followed, most preparations can be handled without presenting any danger.

8. References

1. S. W. B Newsom, *J. Hospital Infect.* **23** (1993),175 – 187.
2. M. L. Rotter, *Hyg. Med.* **22** (1997) 332 – 339.
3. N. V. Nussbaum, *Ärztliches Intelligenzblatt* **5** (1875)43 – 47.
4. H. Rüden, F. Daschner, M. Schumacher in Bundesministerium für Gesundheit (ed.): *Nosokominale Infectionen in Deutschland,*

vol. 56, Nomos Verlag, Baden-Baden 1995,pp. 13 – 14.
5. W. J. Martone, W. R. Jarvis, D. H. Culver, R. W. Haley: "Incidence and nature of endemic and epidemic nosocomical infections," in J. V. Benett, P. S. Brachman (eds.) *Hospital Infection,* Little Brown and Co., Boston 1992, pp. 577 – 596.
6. M. Kresken, A. Jansen, B. Wiedemann, *J. Antimicrob. Chemother.* **25** (1990) 1022 – 1024.
7. D. Bitter-Suermann, M. Exner, K.-O. Gundermann, H.-P. Harke: *The Threat Posed by Infectious Diseases,"* Rudolph Schülke Stiftung, mhp-Verlag, Wiesbaden 1996.
8. H. Reber, *Zentralbl. Bakteriol. Mikrobiol. Hyg., Abt. 1 Orig. B* **157** (1973) 421 – 438.
9. Hoffmann-La Roche (eds.): *Roche Lexikon Medizin,* Urban & Schwarzenberg, München-Wien-Baltimore 1984, p. 340.
10. J. Walton, P. B. Beeson, R. B. Scott (eds.): *The Oxford Companion to Medicine,* Oxford Univ. Press, Oxford-New York 1986, p. 314.
11. J. Borneff: *Hygiene,* 4th ed., Georg Thieme Verlag, Stuttgart-New York 1982, p. 416.
12. Desinfektionsmittelkommission der DGHM (ed): *Prüfung und Bewertung chemischer Desinfektionsverfahren,* Status 12., July 1991, mhp-Verlag, Wiesbaden 1991.
13. H.-P. Harke, *Hyg. Med.* **3** (1978) 76 – 80.
14. H. Eggensperger, *Hosp.-Hyg., Gesundheitswes. Desinfekt.* **67** (1975) 321 – 329.
15. H. Eggensperger, *Hosp.-Hyg., Gesundheitswes. Desinfekt.* **68** (1976) 39 – 55.
16. C. Harrington, H. Walker, *Boston Med. Surg. J.* **148** (1903) 548 – 552.
17. H. J. Eggers, *Zbl. Bakt.* **273** (1990) 36 – 51.
18. A. Böse-Eggert, J. Steinmann, Evaluation of Desderman N against Polivirus Type 1 (Sabin) using Fingerpads of Adult Volunteers, Report Bremen 1997.
19. W. Schürmann, H. J. Eggers, *Antiviral Res.* **3** (1983) 25 – 41.
20. H.-P. Harke, *Hyg. Med.* **4** (1979) 401 – 403.
21. E. L. Larson, H. E. Morton in S. S. Block (ed.): *Disinfection, Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia 1991, pp. 191 – 203.
22. O. Thraenhart, E. Kuwert, *Zentralbl. Bakteriol. Mikrobiol. Hyg., Abt. 1 Orig. B* **161** (1975) 209 – 232.
23. O. Thraenhart, E. Kuwert, *Zentralbl. Bakteriol. Mikrobiol. Hyg., Abt. 1 Orig. B* **164** (1977) 22 – 44.
24. E. M. Scott, S. P. Gorman in S. S. Block (ed.): *Disinfection, Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia 1991, pp. 596 – 614.
25. B. Tichacek, V. Merka, A. Kramer in W. Weuffen, G. Berencsi, D. Gröschel, B. Kemter et al. (eds.): *Handbuch der Antiseptik,* **vol. 2,** part 2, G. Fischer Verlag, Stuttgart-New York 1983, pp. 160 – 176.
26. A. Kramer, W. Weuffen, V. Merka, B. Tichacek in [25], pp. 177 – 187.
27. H. Eggensperger, *Dtsch. Apoth.-Ztg.* **118** (1978) 1073 – 1076.
28. H. Eggensperger, *Dtsch. Apoth.-Ztg.* **122** (1982) 2599 – 2602.
29. P. C. Freer, F. G. Novy, *Am. Chem. J.* **27** (1902) 161.
30. A. Kramer et al., *Hyg. Med.* **16** (1991) 279 – 286.
31. *Bundesgesundheitsblatt* 1997, 72 – 74.
32. W. Weuffen, A. Kramer, V. Adrien, *Hyg. Med.* (1987) in press.
33. V. Merka, F. Sita, V. Zikes, *J. Hyg. Epidemiol., Microbiol., Immunol. (Prague)* **9** (1965) 220 – 226.
34. V. Merka, J. Dvorak, *J. Hyg. Epidemiol., Microbiol., Immunol. (Prague)* **12** (1968) 115 – 121.
35. Schülke & Mayr, DE 2 654 164, 1976 (H. Eggensperger, W. Beilfuss).
36. J. Sporkenbach-Höffler, K. J. Wiegers, R. Dernick, *Zentralbl. Bakteriol. Mikrobiol. Hyg., Abt. 1 Orig. B* **177** (1983) 469 – 481.
37. Schülke & Mayr, DE-OS 2 653 738, 1976 (H. Eggensperger, W. Beilfuss, W. Zerling).
38. Interox Chemicals Ltd., Technisches Merkblatt,Cheshire, UK.
39. Schülke & Mayr, DE 3 046 769, 1980 (J. Sporkenbach, H. Eggensperger, L. Bücklers, H. H. Ehlers et al.).
40. H. Bechthold, P. Ehrlich, *Z. Physiol. Chem.* **47** (1906) 173 – 199.
41. K. Laubenheimer: *Phenol und seine Derivate als Desinfektionsmittel,* Urban & Schwarzenberg, Berlin 1909.
42. D. O. O'Connor, J. R. Rubino in S. S. Block (ed.): *Disinfection, Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia 1991, pp. 204 – 224.
43. W. Beilfuss, L. Bücklers, U. Eigener, H.-P. Harke et al., [25], part 2.
44. W. Beilfuss, H. Nolte, Personal communication, 1987.
45. M. Klein, A. Deforest, *Soap Chem. Spec.* **39** (1963) 70 – 97.

46. H.-P. Harke, *GIT Labor-Med.* **5** (1982) 35 – 36.
47. K. H. Wallhäusser: *Sterilisation Desinfektion Konservierung,* G. Thieme Verlag, Stuttgart 1978, p. 348.
48. W. B. Hugo, A. D. Russel: *Pharmaceutical Microbiology,* Blackwell Scientific Publications, Oxford 1977, pp. 163 – 169.
49. *European Pharmacopeia,* 3rd. ed., Council of Europe, Strasbourg; official german edition: Deutscher Apotheker Verlag, Stuttgart and Govi Verlag, Eschborn 1997.
50. R. A. Cutler, E. B. Cimijotti, T. J. Okolowich, W. F. Wetterau: *Chemical Specialties Manufacturers Association (CSMA) Proceedings of the 53rd Annual Meeting,* 1966, pp. 102 – 113.
51. R. A. Cutler, H. P. Drobeck in E. Jungermann (ed.): *Cationic Surfactants,* Marcel Dekker, New York 1970.
52. W. Beilfuss, Personal communication, 1986.
53. K.-H. Wallhäusser: *Praxis der Sterilisation Desinfektion Konservierung,* G. Thieme Verlag, Stuttgart–New York 1995,p. 597.
54. B. Eckhoff, I. Grell-Büchtmann, J. Steinmann, *Hyg. Med.* **11** (1986) 328 – 330.
55. *Bundesanzeiger no. 159,* 28.08.1994, p. 9122.
56. K. Askgaard, *Ugeskr. Laeg.* **137** (1975) 2515 – 2518.
57. M. L. Rotter: "Hand Washing and Hand Disinfection", in: C. G. Mayhall (ed.): *Hospital Epidemiology and Infection Control,* Williams & Wilkins, Baltimore 1996, pp. 1052 – 1068.
58. M. L. Rotter et al., *J. Hosp. Infect* **11** (1988)310 – 320.
59. E. J. L. Lowbury, H. A. Lilly, *Br. J. Surg.* **61** (1974) 19.
60. G. A. J. Ayliffe, J. R. Babb, K. Bridges, *J. Hyg.* **75** (1975) 259 – 274.
61. D. M. Bailey, C. G. DeGrazia, S. J. Hoff, P. L. Schulenberg et. al., *J. Med. Chem.* **27** (1984) 1457 – 1464.
62. H.-P. Harke, *Zbl. Hyg.* **188** (1989) 188 – 193.
63. U. Eigener, U. Behrens, *Hyg. Med.* **10** (1985) 475 – 478.
64. G. Wewalka et al., *Hyg. Med.* **16** (1991) 335 – 345.
65. V. Hanf, M. Dominguez, P. Heeg, H. R. Tinneberg, *Fertilität* **7** (1991)140 – 143.
66. A. Kramer et al., *Zent.bl. Hyg. Umweltmed.* **200** (1998)443 – 456.
67. E. R. Weissenbacher et al., *Int. J. Experimt. Clinical Res.* **10** (1997) Supl. 1.
68. K. H. Wallhäusser: *Praxis der Sterilisation Desinfektion Konservierung,* 5th ed., Georg Thieme Verlag, Stuttgart–New York 1995, pp. 598–604.
69. S. S. Block in S. S. Block (ed.): *Disinfection Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia–London 1991 pp. 263 – 273.
70. P. G. Micheletti, R. Ponti, C. Cantoni, *Arch. Lebensmittelhyg.* **29** (1978) 94 – 96.
71. H. Edelmeyer, A. Laqua, M. Wiemann, *Arch. Lebensmittelhyg.* **29** (1978) 62 – 65.
72. H. W. Rossmoore in S. S. Block (ed.): *Disinfection Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia–London 1991, pp. 290 – 321.
73. Lonza AG, EP 0 343 605 B 1, 1992 (S. Güller, J. Fritschi, F. Lichtenberg).
74. K. Disch, "Glucoprotamin – ein neuer antimikrobieller Wirkstoff", *Hyg. Med.* **17** (1992) 529 – 534.
75. N. A. Clarke, S. L. Chang, *J. Am. Water Works Assoc.* **51** (1959) 1299 – 1317.
76. J. Peters, *Bundesgesundheitsblatt* (1987) in press.
77. W. A. Mercer, I. I. Sommer, *Adv. Food Res.* **7** (1957) 129 – 160.
78. H.-P. Werner, *Hyg. Med.* **7** (1982) 205 – 212.
79. K. O. Gundermann, H.-P. Harke, D. Horn, G. Koppensteiner et al., *Hyg. Med.* **8** (1983) 175 – 178.
80. D. Horn, W. Ditter in G. Hierholzer, G. Görz (eds.): *PVP Jod in der operativen Medizin,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1984, pp. 7 – 19.
81. E. Pinter, H. Rackur, R. Schubert, *Pharm. Ind.* **46** (1984) 640 – 645.
82. W. Gotthardi in S. S. Block (ed.): *Disinfection Sterilization and Preservation,* 4th ed., Lea & Febiger, Philadelphia–London 1991 152 – 166.
83. J. Sporkenbach, *Hyg. Med.* **5** (1980) 357 – 362.
84. C. H. Carter, *Proc. Soc. Exp. Biol. Med.* **165** (1980) 380.
85. E. J. L. Lowbury, J. R. Babb, K. Bridges, D. M. Jackson, *Br. Med. J.* **1** (1976) 493 – 496.
86. R. J. Pinney, *J. Pharm. Pharmacol.* **30** (1978) 228 –232.
87. A. O. Summer, S. Silver, *J. Bacteriol.* **112** (1972) 1228 – 1236.
88. G. E. Pierce, J. J. Brimm, E. E. Reiber, J. H. Litchfield, *Dev. Ind. Microbiol.* **20** (1979) 455 – 461.
89. A. Cremieux, *Ann. Pharm. Fr.* **37** (1979) 165 – 174.

90. National Institute of Health: *Toxicology and Carcinogenesis Studies of ORTHO-Phenylphenol,* NIH Publication 86-2557, 1986.
91. U. Knecht, H.-J. Woitowitz, *Oeff. Gesundheitswes.* **41** (1979) 715 – 723.
92. W. D. Kerns, K. L. Pavkov, D. J. Donofrio, E. J. Gralla et al., *Cancer Res.* **43** (1983) 4382.
93. J. K. McLaughlin, *Int. Arch. Occup. Envion Health* **66** (1994) 295 – 301.
94. T. Morizono, B. M. Johnstone, E. Hadjar, *J. Oto-Laryngol. Soc. Aust.* **3** (1983) 550.

Disperse Dyes

Rainer Hamprecht, DyStar GmbH & Co. Deutschland KG, Leverkusen, Federal Republic of Germany (Chaps. 1 – 5, 7, 8)

Aloys Westerkamp, DyStar GmbH & Co. Deutschland KG, Leverkusen, Federal Republic of Germany (Chap. 6)

1. Introduction 1
2. Anthraquinone Dyes 2
3. Azo Dyes 3
3.1. Monoazo Dyes 3
3.1.1. Aromatic Amines as Coupling Components 3
3.1.2. Aromatic Hydroxy Compounds as Coupling Components 7
3.1.3. Heterocyclic Compounds as Coupling Components 7
3.1.4. Production of Monoazo Dyes 8
3.2. Disazo Dyes 9
4. Other Chromophores 10
5. Production of Disperse Dyes 11
6. Toxicology and Ecotoxicology 12
7. Trade Names 13
8. Economic Aspects 13
9. References 13

1. Introduction

Disperse dyes are colorants with low water solubility that, in their disperse colloidal form, are suitable for dyeing and printing hydrophobic fibers and fabrics. Forerunners of the disperse dyes were the ionamine dyes of British Dyestuffs Corp.; these were *N*-methanesulfonic acids of aminoazo or aminoanthraquinone dyes that released the *N*-methanesulfonic acid group in the dyeing process and, thereby, precipitated as disperse dyes on the acetate fibers. The understanding of this mechanism in 1923 initiated the development of genuine disperse dyes. British Celanese and British Dyestuffs Corp. were the first companies to introduce these dyes into the market, under the names SRA dyes and Dispersol dyes, respectively, for coloring acetate fibers. The SRA dyes were dispersed with sulforicinoleic acid [*8002-33-3*], and the Dispersol dyes, with soap or Turkey-red oil [1].

From 1924 to 1930, products of other companies appeared on the market, initially as pastes; later, when the materials could be dried successfully without interfering with their dispersibility, they were also marketed as powders. Since 1950, the production of disperse dyes has increased sharply, closely following the growth in worldwide production of synthetic fibers, especially polyester [poly(ethylene terephthalate)] fibers, production of which has grown steadily from ca. 2×10^6 t/a in 1970 to ca. 16×10^6 t/a in 1999 [2].

Furthermore, new dyeing processes necessitated the development of special disperse dyes. For instance, dyes characterized by special ease of sublimation are preferred for transfer printing [3]. The demand for new fastness properties such as thermomigration fastness and automotive light fastness [4–7] also led to new dyes, as has the ongoing pressure on market prices.

Models for the dyeing of polyester fibers with disperse dyes have been developed [8]. When the dye is applied from aqueous medium, it is adsorbed from the molecularly dispersed aqueous solution onto the fiber surface and then diffuses into the interior of the fiber. The following parameters determine the rate of dyeing and, to some extent, the leveling properties: (1) the dissolution rate during the transition from the dispersed crystalline state of the dye into the molecularly dispersed phase, and (2) the diffusion rate at the fiber surface and, especially, in the interior of the fiber. The rates of both processes vary with temperature.

Differences in geometry and polarity of the dye molecules can lead to wide variations in these finishing or dye-specific properties and can

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_565

have a marked effect on the absorption characteristics of all dyes, irrespective of whether single-component or combination dyeing processes are used. For instance, uneven dyeing may occur when an unequal distribution of particle size results in insufficient dispersion stability and, thus, crystal growth and precipitation at the substrate surface.

Dye Classes. Industrially applied disperse dyes are based on numerous chromophore systems. Approximately 60 % of all products are azo dyes and ca. 25 % are anthraquinone dyes, with the remainder distributed among quinophthalone, methine, naphthalimide, naphthoquinone, and nitro dyes [10]. Azo dyes are employed to create almost the entire range of shades; anthraquinone derivatives are used for red, violet, blue, and turquoise. The remaining dye classes are used mainly to produce yellow shades. In this article, the disperse dyes are grouped according to their chemical constitution rather than their color properties. Production methods are discussed only briefly; details are given in the specific dye-related articles, e.g., → Anthraquinone Dyes and Intermediates, → Azo Dyes.

Of the huge range of disperse dyes published in journals and patents, only those are mentioned which have attained some industrial importance. Fastness properties and often also the structures are disclosed in the Colour Index [11].

2. Anthraquinone Dyes

The brilliant red, blue, and turquoise anthraquinone dyes have great industrial significance. The most important red shades are produced by alkyl or aryl ethers of 4-amino-1,3-dihydroxyanthraquinone [*81-51-6*] (**1**).

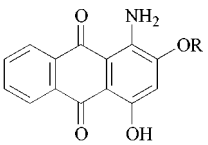

1
R = alkyl, aryl

Blue and turquoise dyes also play a dominant role. The most important blue dyes come from ring- or N-substituted derivatives of the two isomers of aminodihydroxyanthraquinone (**2** and **3**).

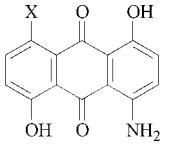

2
X = NH_2, NO_2

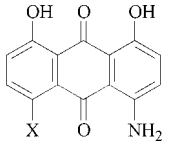

3

Turquoise shades are obtained from derivatives of 1,4-diaminoanthraquinone-2,3-dicarboximide [*128-81-4*] (**4**).

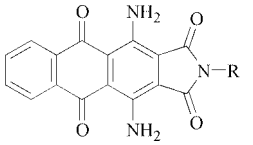

4
R = alkyl

In comparison to azo dyes of similar shades, the anthraquinone derivatives are frequently characterized by greater clarity and better stability against hydrolysis and reduction. The anthraquinone disperse dyes are described in detail under → Anthraquinone Dyes and Intermediates.

A serious handicap of anthraquinone dyes is their poor dyeing strength, which is indicated by a low molar extinction coefficient ε. The ε value usually has a range of 12 000 – 34 000 L mol^{-1} cm^{-1} for anthraquinone dyes, depending on the shade, whereas azo dyes of similar shade attain ε values of 30 000 – 84 000 L mol^{-1} cm^{-1}. This is the reason why disperse anthraquinone dyes have been increasingly displaced by azo dyes. Reviews of this ongoing process are given in [12], [13]. Research on new dyestuffs has therefore been concentrated on nonanthraquinoid dyestuffs. The only exception to this is the brilliant blue dye **5** [14].

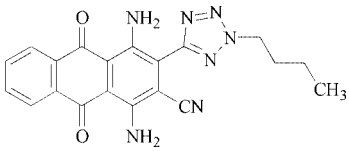

5
[*93686-63-6*] [14]

3. Azo Dyes

Azo dyes represent the largest group of disperse dyes for two reasons: (1) the ease with which an extraordinary number of molecular combinations can be generated by varying the diazo and the coupling components and (2) the relatively simple process by which the dyes can be produced. With this class of dyes, manufacturers can respond much more easily to customers' and end users' requests for special shades and fastness characteristics.

In the following survey, the most important disperse azo dyes are divided into mono- and disazo types; then each of these classes is subdivided according to the diazo and the coupling components. The diazo component is further subdivided where appropriate into aromatic and heteroaromatic amines.

3.1. Monoazo Dyes

About 50 % of all disperse dyes are monoazo dyes, which thus represent the largest single group [10]. Relatively simple syntheses enable a range of shades from greenish yellow to cyan to be produced with this chromophore system.

3.1.1. Aromatic Amines as Coupling Components

Carbocyclic Aromatic Amines as Diazo Components. Dyes of structure **6**, in which aminobenzenes are the coupling components, predominate.

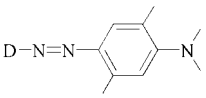

6
D denotes an aromatic or heterocyclic group of the diazo component

Of all the disperse azo dyes, this class has the greatest economic importance. Commercial products are most often represented by structure **7**, in which 4-nitroaniline [*100-01-6*] and its substituted derivatives constitute the diazo component.

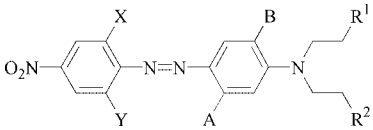

7
X, Y = H, Cl, Br, CN, NO_2, CH_3SO_2
A = H, CH_3, Cl, NHCOR
B = H, CH_3O, C_2H_5O, Cl
R^1, R^2 = H, CN, OH, OCOR, COOR, aryl (R = alkyl)

The most important diazo components are 4-nitroaniline [*100-01-6*], 2-chloro-4-nitroaniline [*121-87-9*], 2-cyano-4-nitroaniline [*17420-30-3*], 2,4-dinitroaniline [*97-02-9*], 2,6-dichloro-4-nitroaniline [*99-30-9*], 2-bromo- [*1817-73-8*] and 2-chloro-4,6-dinitroaniline [*3531-19-9*], and 2-bromo-6-cyano-4-nitroaniline [*17601-94-4*]. The coupling components are derived from aniline, 3-aminotoluene [*108-44-1*], 3-chloroaniline [*108-42-9*], 3-aminoacetanilide [*102-28-3*], and 3-amino-4-alkoxyacetanilide by *N*-alkylation.

This type of dye quickly became important for dyeing acetate fibers. It has been adapted to the requirements of polyester fibers and of different dyeing processes, mainly by varying the substituents R^1 and R^2 [15–18].

For example, the β-hydroxyalkyl groups (R^1 and/or R^2 = OH), typical for many acetate dyes, were replaced by less hydrophilic groups such as CN, OCOR, and COOR, which enhanced the affinity for polyester fibers and, in many instances, the fastness to light and sublimation [10], [19].

The shade of dyes with formula **6** depends largely on the substituents X, Y, A, B, R^1, and R^2. If the coupling component is held constant and X = H, variation of Y results in a bathochromic shift, which increases in the following order: H, Cl, NO_2, CH_3SO_2, CN.

If X and Y $\neq$ H, the situation is more complicated because of possible steric effects on the azobenzene system [20]. For example, if X = Y = Cl, steric hindrance prevents a planar alignment of the azobenzene molecule, which leads to a hypsochromic shift with simultaneous loss of clarity. This effect is exploited in the industry to produce brown shades. Cyano groups exert no steric effects, and the strongest bathochromic shifts can be obtained by using X = Y = CN; X = CN, Y = NO_2; and X = CN, Y = CH_3SO_2. When the diazo component is held

constant, a bathochromic shift occurs which increases in the order A = Cl, H, CH_3, NHCOR. The most extreme bathochromic effect is obtained when A = NHCOR and B = OR. In combination with 2-bromo- [*1817-73-8*] or 2-chloro-4,6-dinitroaniline [*3531-19-9*] as the diazo component, commercially important navy blue dyes are obtained; of all the monoazo dyes, these are produced in the largest quantity.

The significant influence of R^1 and R^2 on color is quite surprising because these substituents are not components of the chromophore system. Ester and, especially, cyano groups in these positions cause a distinct hypsochromic shift [10].

A few examples of the many dyes from this series that are used to dye polyesters are listed below:

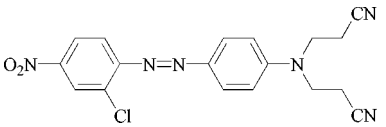

Orange [*4058-30-4*] [21]

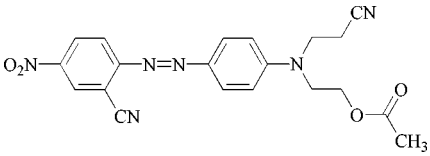

Red, C.I. 11114, [*12223-39-1*] [22]

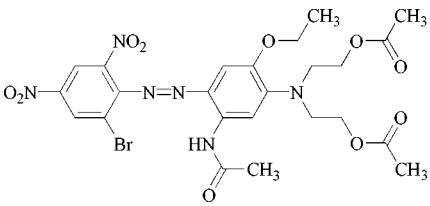

Navy blue, C.I. 11345, [*3956-55-6*] [23]

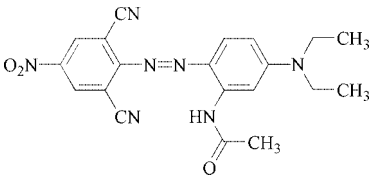

Disperse Blue 165 [*41642-51-7*] [24]

Representatives of this class are also suitable for dyeing cellulose triacetate [*9012-09-3*]. However, sufficient lightfastness is not obtained on synthetic polyamides [10].

Other industrially important 2,6-dicyanoazo dyes are obtained by using diethyl-*m*-toluidine and 3-diethylaminophenylmethansulfonamide [*52603-47-1*] as coupling components (see **8** and **9**, respectively).

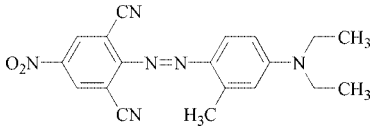

8
Disperse Blue 366 [*84870-65-5*] [25]

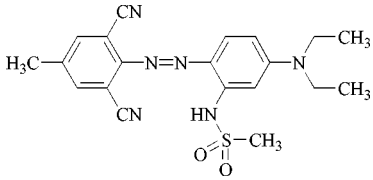

9
Red [*68385-96-6*] [26]

By mixing the reddish blue Disperse Blue 366 with Disperse Blue 165 [27] or heteroarylazo dyes [28], technically equivalent replacement products [13] for the anthraquinoid Disperse Blue 56 were obtained. The red dye **9** has been similarly used as a substitute [12] for the anthraquinoid Disperse Red 60.

Heteroaromatic Amines as Diazo Components. Developments in this field started with the finding that 2-amino-5-nitrothiazole [*121-66-4*] [29] gave bright blue shades of reasonable fastness in azo dyes for cellulose acetate (e.g., **10**, [30]). Because of their good dischargeability but limited fastness properties, these dyes were also used on polyester. The importance of this class of dyes has diminished since some of them have been found to have sensitizing properties (see Chap. 6).

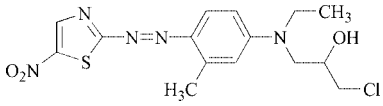

10
Blue [*70865-21-3*] [31]

2-Amino-6-nitrobenzthiazole [*6285-57-0*] and 2-amino-5,6- or 6,7-dichlorobenzthiazole [*24072-75-1*], [*25150–27-0*] are widely employed to give scarlet to ruby shades because of their facile accessibility by ring closure of 4-nitrophenylthiourea [32] or 3,4-dichlorophenylthiourea, respectively. In the latter case the resulting mixture of isomers is directly utilized for diazotization and transformed into the mixture of dye isomers **12**.

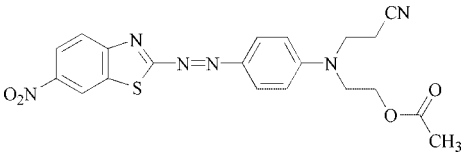

11
[*68133-69-7*] [33]

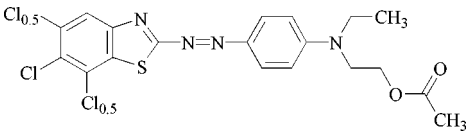

12
[*117541-97-6*], [*117541-98-7*] [34–36]

The red dye **13** is an example for the use of the methanesulfonyl group as an acceptor in azo dyes.

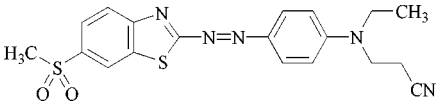

13
Red [*63467-01-6*] [37]

2-Amino-5-ethylsulfanyl-1,3,4-thiadiazole [*25660-70-2*], accessible by alkylation of bithiourea with diethyl sulfate and ring closure [38], is suitable for the manufacture of bright reds, e.g., **14**.

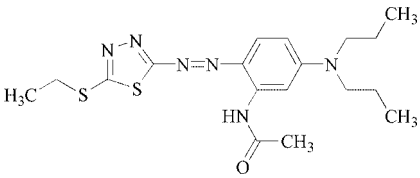

14
Red [*63134-15-6*] [39]

2-Amino-4,5-dicyanoimidazole [*40953-34-2*] can be made by reaction of diaminomaleonitrile with cyanogen chloride. By coupling with aniline couplers and alkylating the resultant azo dye precursor bright reds such as **15** are obtained.

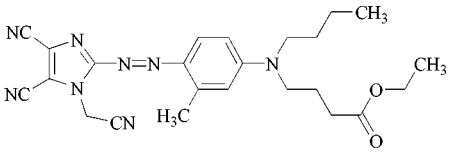

15
Red [*86772-44-3*] [40]

3-Amino-5-nitro-2,1-benzisothiazole is manufactured by thiolysis of 2-amino-5-nitrobenzonitrile and oxidative ring closure. With aniline coupling components, shades from navy to greenish blue are obtainable (e.g., **16** and **17**).

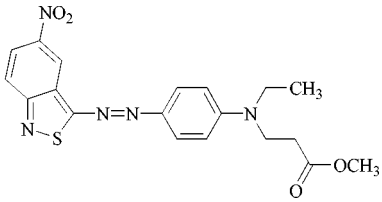

16
Disperse Blue 148 [*52239-04-0*] [41]

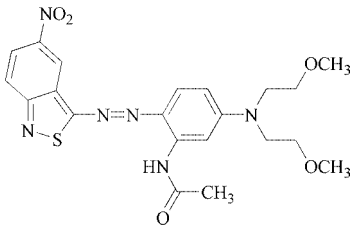

17
Blue [*105076-77-5*] [42]

Since the first appearance of thienylazo dyes like Disperse Green 9 in the early 1970s, this field has been intensely exploited. The interest originates from the fact that thienylazo dyes are nearly 100 nm more bathochromic than their carbocyclic counterparts (cf. Disperse Green 9 and **18**) and exceed these in the molar extinction coefficient and hence also in dyeing strength.

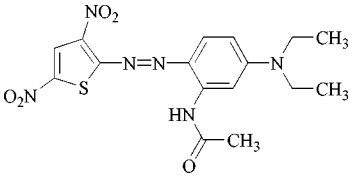

Disperse Green 9 [*58979-46-7*] [43]
$\lambda = 644$ nm $\varepsilon = 87\,400$ L mol^{-1} cm^{-1} [12]

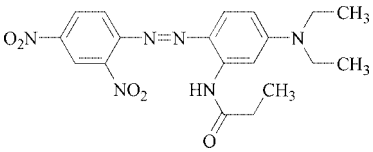

18
$\lambda = 545$ nm $\varepsilon = 46\,000$ L mol^{-1} cm^{-1} [44]

Furthermore, the Gewald reaction [45] opened a simple and economic access from aliphatic reactants to aminothiophenes that were formerly only accessible from multistep reactions [46] (see, e.g., the dye **19**. For example, 2-amino-3-carbethoxy-5-nitrothiophene, a precursor for the dye **20**, can be prepared by reaction of 1,4-dithiane with ethyl cyanoacetate and subsequent nitration.

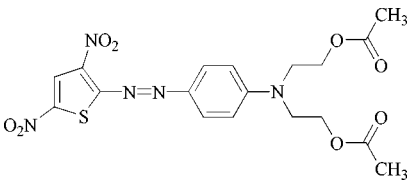

19
Blue [*42783-06-2*] [47]

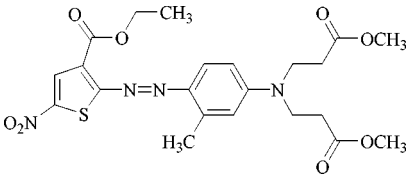

20
Blue [*210758-04-6*] [48]

Whereas the blue dyes **19** and **20** are claimed to have good thermomigration fastness, **21**, **22**, and **23** have superior pH stability and are proposed as substitutes for bright anthraquinone blues.

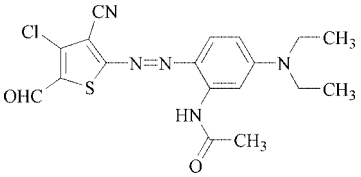

21
Blue [*104366-25-8*] [49]

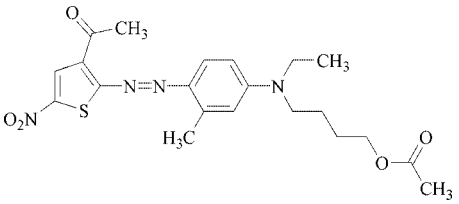

22
Blue [*122063-39-2*] [50]

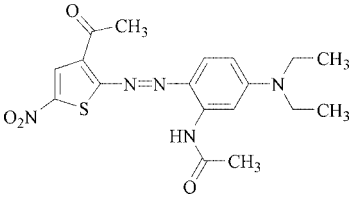

23
Blue [*167940-11-6*]

By combination of the thiophene diazo moiety with a diaminopyridine coupler, the hue is shifted to a bright red with good lightfastness (**24**).

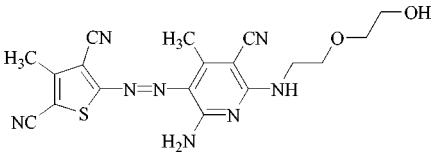

24
Red [*107815-88-3*] [51]

The chemistry and physics of azo dyes based on heterocyclic amines has been surveyed in many reviews [52–54]

Aminonaphthalenes as Coupling Components. In comparison to aminobenzenes, aminonaphthalenes have played only a minor role as coupling components for disperse dyes and are encountered in only a few industrial appli-

cations, e.g., **25** for coloring of cellulose acetate fibers.

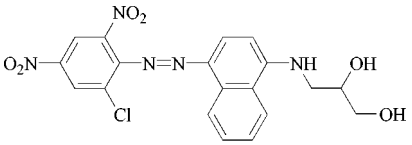

25
Blue [*69828-87-1*] [55]

3.1.2. Aromatic Hydroxy Compounds as Coupling Components

Dyes with aromatic hydroxy compounds as the coupling components, of which Cellitone Fast Yellow G is a representative example, have been used many years to dye acetate fibers. Its use has been questioned for ecological reasons (see Chap. 6). For application to polyesters, sublimation fastness has been enhanced by increasing the size of the molecule.

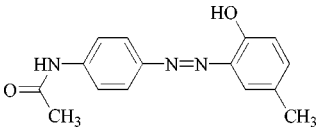

Disperse Yellow 3, C.I. 11855 [*2832-40-8*]

When 2-hydroxynaphthalene [*135-19-3*] and its derivatives are used as coupling components, clear orange (e.g., **26**) and red shades result. Their commercial importance is low.

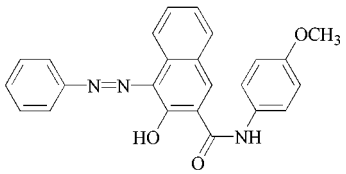

26
Orange [*17947-32-9*] [56]

The chemistry of 4-hydroxynaphthalimides as coupling components has mainly been investigated in Japan. Brilliant reds with good fastness properties are obtainable (e.g., **27**).

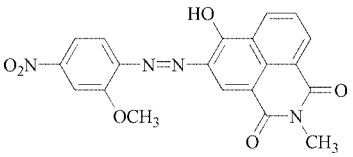

27
Red [*42357-98-2*] [57]

3.1.3. Heterocyclic Compounds as Coupling Components

The importance of the formerly widespread pyrazolone, aminopyrazole, and 4-hydroxyquinolone based yellow azo dyes has greatly diminished with the advent of the tinctorially superior and therefore more economical pyridone azo dyes. Only a few examples have survived in special applications such as for dyeing of acetate fibers. Examples are Disperse Orange 56 and Disperse Yellow 5.

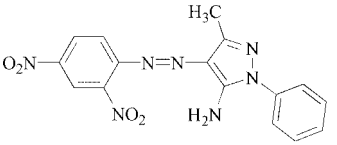

Disperse Orange 56 [*67162-11-2*] [58]

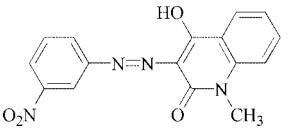

Disperse Yellow 5 [*6439-53-8*]

By reaction of ethyl acetoacetate with cyanoacetamides, pyridones are easily accessible. By selection of suitable substituents in the diazo components, shade and fastness properties and build up can be steered (e.g., **28**, **29**, **30**, **31**, and Disperse Yellow 211).

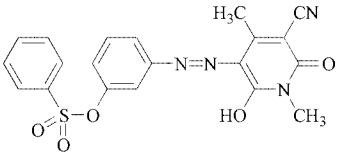

28
Greenish Yellow [*59312-61-7*] [59]

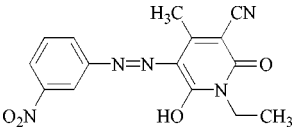

29
Greenish Yellow [*37781-00-3*] [60]

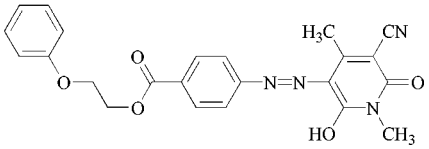

30
Greenish Yellow [*88938-37-8*] [61]

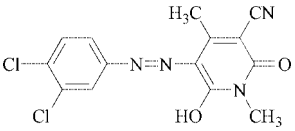

31
Greenish Yellow [*83249-52-9*] [62]

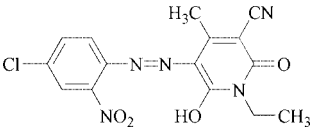

Disperse Yellow 211 [63]

Chlorination of 2,6-dihydroxy-3-cyano-4-methylpyridine with phosphorus oxychloride and successive reaction with amines yields 2,6 diaminopyridines. With these coupling components brilliant orange and red dyes with excellent lightfastness are obtained (e.g., **32**, [4]).

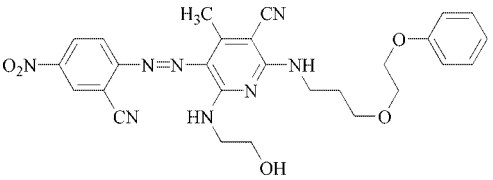

32
Red [*63833-78-3*] [64]

3.1.4. Production of Monoazo Dyes

The production of monoazo disperse dyes by diazotization and coupling is described in detail under → Azo Dyes. Additional chemical treatment subsequent to synthesis of the azo compounds is performed only rarely.

Frequently, weakly basic amines, such as negatively substituted 4-nitroanilines or aminoheterocycles, must be diazotized. These reactions require concentrated sulfuric acid, phosphoric acid (85 wt %), glacial acetic acid, propionic acid, or mixtures of these acids as the reaction medium, with nitrosylsulfuric acid as the reagent. On an industrial scale, this reaction is conducted preferably in enameled, stirred vessels. Whenever the stability of the amine or the diazonium salt permits, diazotization is carried out in concentrated sulfuric acid at slightly elevated temperature (10 – 40 °C). With regard to safety, factors such as temperature and concentration must be controlled carefully to avoid explosion [65].

The sulfuric acid diazotization mixture is generally added directly to the ice-cold solution or suspension of the coupling component. If the diazonium salt is stable in dilute sulfuric acid, the diazotization mixture may first be quenched with ice, possibly clarified, and only then combined with the coupling component.

Strongly negatively substituted 4-nitroanilines, such as 2,4,6-trinitroaniline [*489-98-5*], 2,4-dinitro-6-cyanoaniline [*22603-53-8*], 2-methanesulfonyl-4-nitro-6-cyanoaniline, and 2,6-dicyano-4-nitroaniline [*20033-48-1*], are difficult to diazotize, and their diazonium salts tend to decompose extensively during the coupling process [15]. This results in diminished yields and loss of clarity of the dyes.

These difficulties may be avoided if 2-halogenoazo dyes, preferably 2-bromoazo dyes, are generated first, and the halogen is then exchanged for the NO_2 or CN group. These copper-catalyzed reactions proceed especially smoothly in dipolar aprotic solvents such as ethylene glycol monoalkyl ethers, dimethylformamide, dimethyl sulfoxide, or *N*-methylpyrrolidone, and they occur under considerably milder conditions than those in the starting amine [66–68]. In this process, the bromo – cyano exchange yields clear blue polyester dyes (see formula **33**) of a shade which, formerly, could only be obtained by using anthraquinones.

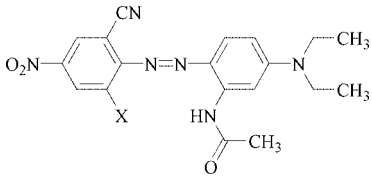

33
X = CN, NO_2

By modification of the substituent in the 4-position of the diazo component, this process yields further 2,6-dicyanoazo dyes, for example, 4-methyl- or 4-bromo-2,6-dicyano derivatives, where the basic diazo component is otherwise difficultly accessible. In this way clear red shades on polyester can be obtained.

3.2. Disazo Dyes

About 10 % of all disperse dyes are disazo compounds [10]. Even the simplest hydroxy disazo dyes, such as 4-aminoazobenzene coupled to phenol and 4-aminoazobenzene coupled to *o*-cresol, have good affinity for polyester fibers and yield lightfast reddish yellow hues. However, these shades are frequently less bright than those obtained with monoazo dyes.

The introduction of an alkoxy group into the central benzene ring causes a distinct shift toward orange. A similar bathochromic effect is obtained by replacing this benzene with naphthalene.

Substitution of the first benzene nucleus by electron acceptors also causes bathochromic shifts. Disazo dyes of this type such as **34** are frequently incorporated as components of black mixtures.

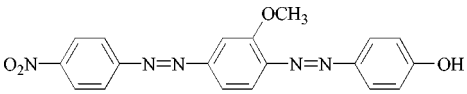

34
Orange, C.I. 26077 [*19800-42-1*] [69]

The simplest route to disazo dyes is by using *p*-aminoazobenzene [*60-9-3*]. As *p*-aminoazobenzene has been classified in the EU as carcinogenic, the use of derived disazo dyes like Disperse Red 151, Disperse Yellow 7, Disperse Yellow 23, and Disperse Yellow 56 has diminished.

Many disazo dyes with derivatives of aniline as the terminal component have been described in the literature, but only a few have become industrial products. Replacement of the central benzene by heterocycles such as thiazole or thiophene causes marked bathochromic shifts.

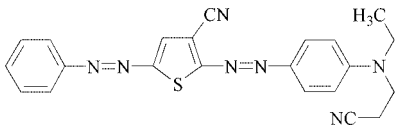

35
Blue, [*84425-43-4*] [70]

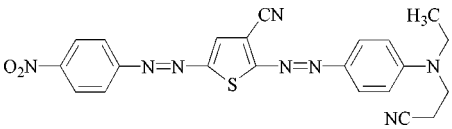

36
Blue [*87606-56-2*]

4. Other Chromophores

Quinophthalones. 3′-Hydroxyquinophthalone [*7576-65-0*] is employed extensively in a number of product lines used to dye synthetic fibers in greenish yellow hues with good lightfastness and generally sufficient sublimation-fastness. Suitable substitution in the phthalic acid residue or the quinoline nucleus may improve thermosetting fastness [71]. For synthetic methods, see → Methine Dyes and Pigments.

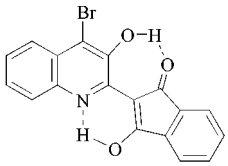

Disperse Yellow 64 [72]

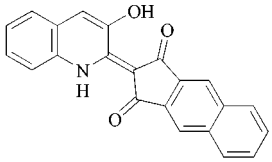

37
Yellow [*75216-43-2*] [73]

Methine Dyes. Condensation products of 4-dialkylaminobenzaldehydes with cyanoacetic esters have long been used to dye acetate fibers. Brilliant greenish yellow dyes with excellent lightfastness are obtained on polyester fibers with the corresponding condensation products of malonodinitrile. The sublimation fastness of this dye type can be improved by introducing suitable substituents into the alkyl residue of the amino group or by doubling the molecular size (**38**) [16], [74].

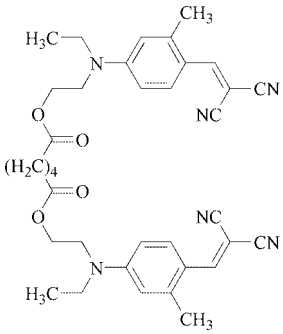

38
Yellow [*25857-05-0*] [75]

Brilliant blue dyes such as **39** are produced from aminobenzaldehydes and benzothiophene derivatives.

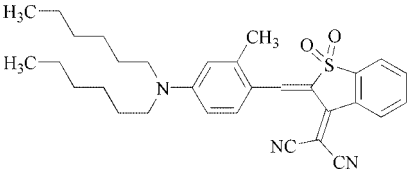

39
Blue [*74239-96-6*] [76]

By condensation of 2-hydroxy-4,4′-diethylaminobenzaldehydes with hetarylacetonitriles, 3-heteroaryl-7-diethylaminocoumarins such as **40** and **41** are obtained, which dye polyester fibers fluorescent yellow with a strong green tint.

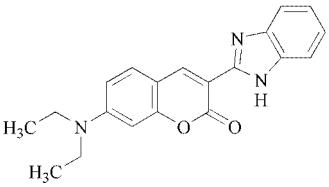

40
Yellow [*27425-55-4*] [77]

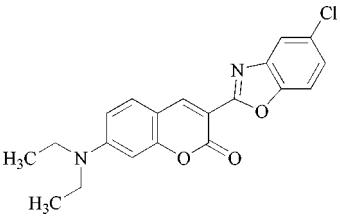

41
Brilliant Yellow [*35773-43-4*] [78]

Further condensation with malonodinitrile results in fluorescent red dyes such as **42**.

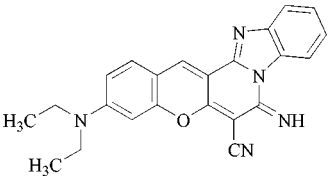

42
Red [*53272-39-1*] [79]

Oxidative cyanation of yellow 3-heteroaryl-4*H*-7-diethylaminocoumarin dyes such as **40** and **41** introduces a 4-cyano group and yields fluorescent red dyes [9]. For example, treatment with NaCN followed by Br_2 gives **43**.

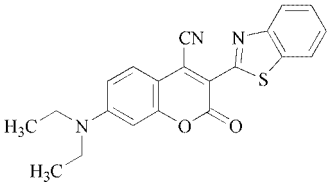

43

By elongation of the methine chromophore even brilliant blues such as **44** can be produced. Because of limited fastness properties, this interesting new dye could not yet pose a threat to anthraquinone dyes.

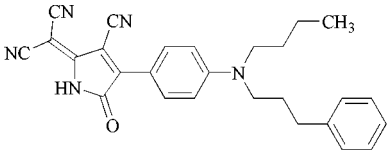

44
Brilliant blue [*108948-36-3*] [80]

Condensation of mandelic acids and hydroquinone, followed by oxidation, results in yellow to red polyester dyes [81]. This new chromophore shows excellent thermomigration fastness and has been frequently varied to cover all red shades (e.g., **45**) and to improve build up.

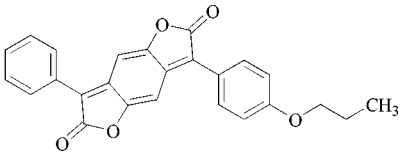

45
Red [*79694-17-0*] [81]

Naphthalimide Dyes. Derivatives of 4-aminonaphthalimide were used initially to dye acetate fibers. For polyesters, condensation products of 1,8-naphthalenedicarboxylic acid [*518-05-8*] or 1,4,5,8-naphthalenetetracarboxylic acid [*128-97-2*] with 1,2-diaminobenzenes are used.

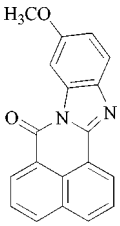

49
Yellow [*15220-29-8*]

Indigo Dyes. Whereas indigo itself is not suitable for exhaustion dyeing of polyester, thioindigo gives fast brilliant red shades in pale to medium depth [82].

Nitro Dyes. 2-Nitrodiphenylamines are readily obtained by condensation of derivatives of 2-nitrochlorobenzene [*88-73-3*] with suitable aromatic amines. Because of their accessibility and good lightfastness, these dyes became very important for dyeing cellulose acetate and, more recently, have gained a solid position as disperse dyes for polyester fibers. This is especially true for the reaction product of 1 mol of 3-nitro-4-chlorobenzenesulfonyl chloride [*97-08-5*] and 2 mol of aniline. An exhaustive review of the constitution and color of nitro dyes is given by Merian [83]. The yellow nitroacridones may also be classified in this group.

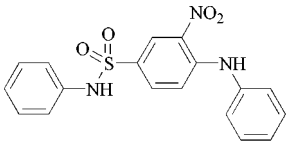

50
Yellow, C.I. 10338 [*5124-25-4*]

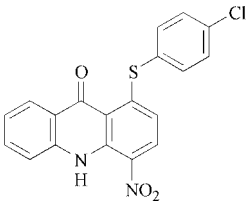

51
Yellow [*19309-55-8*] [84]

5. Production of Disperse Dyes

Chemical synthesis produces dyes of varying particle size. When the dyes are applied in this form, uneven and spotty dyeing results, and the dyeing process may be slow and frequently accompanied by incomplete absorption [85]. To assure high yield, good reproducibility, and faultless dyeing and printing in commercial use, especially when densely woven fabric or wound material is involved, the dye must be applied as a fine dispersion that is stable under the process conditions.

Therefore, the first requirement is that the dye itself be present in a stable crystal modification that does not change during application. If such a modification cannot be obtained by synthesis, attempts must be made to generate it by thermal posttreatment, for instance, by heating the dye in an aqueous suspension in the presence of nonionic detergents.

The eventual dispersion process consists of grinding the moist dye, generally in a ball, bead, or sand mill, in the presence of dispersing agents. Since the dyeing rate depends on particle size, the aim is to reduce the particle size to $< 1\ \mu m$. A narrow particle-size distribution is important to minimize recrystallization during storage and application.

The dispersion agents used are primarily ligninsulfonates (sulfite cellulose liquors); condensation products of naphthalene, sulfuric acid, and formaldehyde; condensation products of *m*- and *o*-cresol, formaldehyde, and 2-hydroxynaphthalene-6-sulfonic acid [*93-01-6*]; or mixtures of these products [85]. Literature on dispersing agents is relatively scarce [86–88].

Difficulties in the production of disperse dyes may be encountered, for example, in grinding or drying dyes with low melting points. Therefore, only dyes that have a melting point $> 120\,°C$, and preferably $> 140\,°C$, should be used. Disperse dyes are sold either as aqueous pastes obtained by grinding or as powders obtained by gentle drying of the paste. Powders are generally produced by spray drying. Important requirements for industrial products are the storage stability of the paste and the ease with which the powder can be redispersed in water. The actual dye content of commercial products is generally 20 – 40 % for pastes and 20 – 50 % for powders.

6. Toxicology and Ecotoxicology

Ecotoxicology. Due to their poor water solubility (< 1 mg/L), disperse dyes have low acute ecological impact. Especially the acute toxicity to aquatic life is generally low. Nevertheless, according to the European chemicals legislation disperse dyes are classified as substances, that "may cause long-term adverse effects in the aquatic environment" [89], [90], because they are not easily biodegradable and suspected as

being potentially bioaccumulative due to their hydrophobicity.

In practice, unfixed disperse dyes discharged from spent dye baths to wastewater treatment plants are easily eliminated by coprecipitation with the sewage sludge and may be anaerobically degraded in the digestion process.

Most commercial disperse dyes are formulated with naphthalene sulfonate or lignosulfonate dispersing agents. These dispersants are completely discharged with the spent dyebath. They have poor biodegradability and will pass the treatment plants to a notable extent. Although not harmful to aquatic life, these dispersants contribute to the persistent COD load of the wastewater. Recently, disperse dyes containing highly effective biodegradable dispersants were brought onto the market (Dianix ECO types [91]).

Toxicology. Disperse dyes generally have low mammalian toxicity as a result of their low solubility and bioavailability. Like any organic chemical, disperse dyes, mostly as powder formulations, must be handled according to good industrial working practice [92].

Some specific azo-type disperse dyes, which contain theoretically azo-cleavable carcinogenic amines, are banned in some countries, e.g. Germany, for the dyeing of consumer goods such as garments [93]. Examples of this class which seem to be still on the market presently are C.I. Disperse Orange 60, C.I. Disperse Red 221 and C.I. Disperse Yellow 218 [94].

Some disperse dyes occasionally have caused clinically documented contact allergies in consumers wearing polyamide hosiery, leggings, and similar disperse-dyed articles. On polyamide, cellulose acetate, and elastane fibers, disperse dyes exhibit poor wet fastness properties and hence can migrate onto the skin of the consumer and potentially elicit allergenic reactions in sensitized persons [95].

Nine generic disperse dyes with epidemiologically proven sensitizing properties are intended for a total ban in the dyeing of garments by the EU: C.I. Disperse Blue 1, Disperse Blue 3, Disperse Blue 35, Disperse Blue 106, Disperse Blue 124, Disperse Yellow 3, Disperse Orange 3, Disperse Orange 37/76, and Disperse Red 1 [96]. These species are already banned by the most important private ecological standard, namely, Oeko-Tex Standard 100 [97], and company related eco standards.

7. Trade Names

Disperse dyes are used to dye cellulose 2,5-acetate, cellulose triacetate, synthetic polyamides, and to a lesser degree, polyacrylonitrile and polypropylene. Their major application is clearly for dyeing polyesters.

Since only a few individual dyes have satisfactory properties for all synthetic fibers, many dye producers are offering separate lines for individual fibers as well as for various application methods. The following list includes the best-known disperse dye lines for dyeing polyester fibers:

Bucron (Blackmann-Uhler)
Dianix (DyStar)
Dispersol (BASF)
Foron (Clariant)
Kayalon Polyester (Nippon Kayaku)
Miketon Polyester (Mitsui Toatsu)
Palanil (BASF)
Serilen (Yorkshire)
Sumikaron (Sumitomo)
Terasil (Ciba Specialities)

As a result of closing of production, merging of producers, or renaming of disperse dye ranges the list of global players has greatly diminished, whereas local producers especially in the Far East, have grown. Parallel to this development, concentration on major products and streamlining of ranges has reduced the number of individual colorants.

8. Economic Aspects

From 1975 to 1998 the share of polyester fibers in world man-made fiber production has nearly doubled from ca. 30 to almost 60 %. Production of disperse dyes has accompanied this tremendous growth. In 1997 disperse dyes accounted for ca. 18 % of the total world market for textile dyes of 800 000 t/a and ca. 20 % of the turnover of € 6×10^9.

9. References

1. R. K. Fourness, *J. Soc. Dyers Colourists* **72** (1956) 513.
2. Acordis, The Man-made Fibre Year 1998.
3. N. L. Moore, *J. Soc. Dyers Colourists* **90** (1974) 318.
4. T. Hihara, *Textile Art* **3** (1997) 23.
5. T. Lampe, P. Neumayer, W. Roggenbach, *Melliand Textilberichte* **1992**, 402 – 406.
6. K. H. Ulbrich, *Melliand Textilberichte* **1992**, 406 – 410.
7. P. Vonhöne, J. Stuck, *Melliand Textilberichte* **1922**, 411 – 416.
8. W. McDowell, *Melliand Textilber.* **61** (1980) 946. E. Schollmeyer, G. Heidemann, A. Bossmann, S. Dugal, C. Heinrichs, *Textilveredlung* **20** (1985) 190.
9. Ciba-Geigy, US 4 547 579, 1985 (P. Möckli).
10. C. Müller, *Chimia (Aarau)* Suppl. 1968, 69.
11. *Colour Index International,* 3rd ed., The Society of Dyers and Colourists, American Association of Textile Chemists and Colorists, Bradford/England, 1992.
12. O. Annen, R. Egli, R. Hasler, B. Henzi, H. Jakob, P. Matzinger, *Rev. Progr. Coloration* **17** (1987) 72.
13. A. T. Leaver, B. Glover, P. W. Leadbetter, *Textile Chemist and Colorist* **24** (1992) 18
14. Ciba-Geigy, EP 79 862, 1981 (J. M. Adam).
15. C. V. Stead, *Rev. Prog. Color. Relat. Top.* **1** (1970) 26.
16. J. F. Dawson, *Rev. Prog. Color. Relat. Top.* **3** (1972) 18.
17. *Ullmann,* 3rd ed., Suppl., p. 222.
18. J. F. Dawson, *Rev. Prog. Coloration* **9** (1978) 25.
19. *Ullmann,* 3rd ed., **4,** 128.
20. A. T. Peters, *J. Soc. Dyers Colourists* **101** (1985) 361.
21. DuPont, US 2 782 187, 1954 (M. F. Sartori).
22. Sandoz, DE-AS 1 065 112, 1955 (E. Merian).
23. Sandoz, DE-AS 1 069 313, 1956 (E. Merian).
24. Bayer, DE-OS 1 544 563, 1966 (A. Gottschlich, K. Leverenz).
25. Synthetic Dye Technology Research Association, JP 03 006 275, 1991 (M. Kurosawa et al.).
26. Bayer, DE-OS 2 711 130, 1977 (R. Hamprecht).
27. Cassella, EP 277 529, 1987 (U. Bühler, K. Hofmann, M. Hähnke).
28. Bayer, EP 347 685, 1988 (M. Hoppe, D. Wiegner, K.-P. Sagner, H. D. Jordan, H. Brandt).
29. Monsanto, US 2 617 809, 1950 (H. L. Hubbard, G. W. Steahly).
30. Eastmann Kodak, US 2 683 709, 1951 (J. B. Dickey, E. B. Towne).
31. ICI, GB 840 903, 1957 (J. R. Atkinson, G. Booth, E. L. Johnson).
32. A. T. Peters, S. S. Yang, *Dyes Pigments* **28** (1995) 151.
33. Saul and Co, US 2 891 942, 1959.
34. A. T. Peters, E. Tsatsaroni, M. Xisai, *Dyes Pigments* **20** (1992) 41.
35. Mitsubishi, JP 631 355 79, 1988 (K. Himeno, J. Yoshihara).
36. A. T. Peters, N. M. A. Gbadamosi, *Dyes Pigments* **18** (1992) 115.
37. Eastman, DE-AS 1 065 542, 1952 (J. M. Straley, D. J. Wallace).
38. Eastmann Kodak, US 4 374 993, 1981.
39. Eastman, US 3 657 215, 1969 (M. A. Weaver, D. J. Wallace).
40. R. W. Sabnis, D. W. Rangnekar, N. D. Sonawane, *J. Heterocycl. Chem.* **36** (1999) 333. Mitsubishi, JP 5 829 859, 1981.
41. BASF, DE-AS 1 544 375, 1965 (M. Seefelder, H. G. Wippel, H. Armbrust).
42. Bayer, EP 167 913, 1984 (W. Kruckenberg, K. Leverenz, H.-G. Otten).
43. ICI, DE-OS 2 304 218, 1972 (D. B. Baird, A. T. Costello, B. R. Fishwick, R. D. McClelland, P. Smith).
44. A. T. Peters, *Dyes Pigments* **22** (1993) 223.
45. K. Gewald, *Chimica* **34** (1980) 101.
46. J. B. Dickey, E. B. Towne, M. S. Bloom, W. H. Moore, B. H. Schmith, Jr., D. G. Hedberg, *J. Soc. Dyers Colour.* **74** (1958) 123.
47. ICI, GB 1 394 365, 1972 (B. D. Baird, A. T. Costello, R. B. Fishwick, R. McClelland).
48. ICI, DE-OS 2 612 792, 1976 (D. B. Baird, A. T. Costello, B. R. Ribbons R. D. McClelland, P. Smith).
49. Clariant, DE 3 529 831, 1984 (R. Egli, B. Henzi).
50. ICI, EP 297 710, 1988 (D. Brierley, R. W. Kenyon, D. R. A. Ridyard).
51. BASF, EP 214 445, 1986 (J. Dehnert, G. Lamm, H. Loeffler).
52. E. Hahn, *J. Heterocycl. Chem.* **24** (1987) Suppl. Vol. 9, 9 – 13.
53. M. A. Weaver, L. Shuttleworth, *Dyes Pigments* **3** (1982) 81.
54. A. D. Towns, *Dyes Pigments* **42** (1999) 3.
55. Yorkshire, GB 985 254, 1963. (E. Stead)
56. Hoechst, DE-OS 1 810 305, 1968.
57. Mitsubishi, GB 1 384 457, 1972 (S. Imahori, M. Kaneko, Y. Kato).

58. Geigy, BE 594 867, 1959 (W. Bossard, F. Farre).
59. Casella, DE-OS 1 813 385, 1968 (H. von Brachel, E. Heinrich, O. Gräwinger, K. Hintermeier, H. Kindler).
60. ICI, DE-AS 1 932 806, 1968 (A. H. Berrie, N. Hughes).
61. Mitsubishi, JP 58 136 656, 1982
62. Bayer, EP 74 562, 1981 (H. Brandt, K. Leverenz, H.-G. Otten).
63. ICI, DE-OS 1 932 806, 1970 (H. H. Berrie, N. Hughes).
64. BASF, DE-OS 2 525 505, 1975 (N. Zimmermann, E. Daubach).
65. P. Bersier, L. Valpina, H. Zubler, *Chem. Ing. Tech.* **43** (1971) 1311.
66. Bayer, DE-OS 1 544 563, 1966.
67. ICI, DE-OS 1 807 642, 1967.
68. Sandoz, DE-OS 1 966 124, 1968.
69. Koppers, BE 604 935, 1960 (D. A. Zanella, C. R. Conway).
70. Mitsubishi, DE-OS 3 151 114, 1980 (S. Imahor, K. Himeno, S. Maeda).
71. B. K. Manukian, A. Mangini, *Chimia (Aarau)* **24** (1970) 328.
72. BASF, GB 1 036 389, 1965 (A. Schöllig).
73. Mitsubishi, JP 7 308 319, 1973.
74. J. M. Strayley in: *The Chemistry of Synthetic Dyes,* **vol. III,** Academic Press, New York 1970, p. 385.
75. Ciba, GB 1 201 925, 1966 (R. Peter, H.-J. Anliker).
76. Sandoz, DE-OS 2 929 001, 1978 (W. Baumann).
77. Geigy, GB 1 085 456, 1964 (J. Voltz, H. Heusermann).
78. Bayer, DE-OS 2 065 076, 1972 (H. Harnisch).
79. BASF, DE-OS 2 253 538, 1972 (W. Mach, D. Augart, H. Scheuermann).
80. Nippon Kayaku, DE-OS 3 716 840, 1987 (H. Matusmoto, H. Imai, S. Tada).
81. ICI, EP 33 583, 1980 (C. W. Greenhalgh, N. Hall, D. F. Newton).
82. Hoechst, EP 26 463, 1981.
83. E. Merian, *Angew. Chem.* **72** (1960) 766.
84. Geigy, GB 1 127 721, 1966 (K. Burdeska, H. Bosshard, A. Pugin).
85. J. F. Dawson, *Rev. Prog. Coloration* **14** (1984) 90.
86. A. Murray, K. Mortimer, *J. Soc. Dyers Colourists* **87** (1971), no. 6.
87. N. J. Sardesai, *Colourage* **36** (1989), Moscow Seminar Special Issue.
88. "Dispersing Agents, Review of Progress in Colouration and Related Topics", *Soc. Dyers Colourists* **23** (1993) 48.
89. EU Directive 93/21/EEC.
90. EU Common Position Paper No 54/98.
91. J. Gruetze, "Dianix ECO Dyes – a New Approach to Dispersing Dyestuffs" 25th, Aachen Textile Conference, Nov. 25/26, 1998.
92. ETAD, "Safe Handling of Dyes – a Guide for the Protection of Workers Handling Dyes", 1995.
93. ETAD, "German Ban of Use of Certain Azo Compounds in Some Consumer Goods", Oct. 1998.
94. DyStar Textilfarben GmbH, "German Legislation on Azo Dyes", 1998.
95. T. Platzek, "How Serious is the Actual Health Hazard Caused by Textiles?" *Melliand Textilberichte* **77** (1996) no. 11, E165 – E167.
96. ETAD, "Position Paper on Skin Sensitizing Dyes" Oct. 1998.
97. DyStar Textilfarben GmbH, "The Oeko-Tex Standard 100", 1998.

Disperse Systems and Dispersants

Rudolf Heusch, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 1 – 5)

Karl Reizlein, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 3)

1. Terminology 1
2. Production of Disperse Systems . . 3
3. Stabilization of Disperse Systems . 3
3.1. Attraction Forces 4
3.2. Repulsion Forces 4
3.2.1. Electrostatic Interaction 4
3.2.2. Steric Stabilization 5
3.3. DLVO Theory 8
4. Destruction of Disperse Systems . 9
4.1. Forces and Mechanisms 9
4.2. Practical Methods 10
5. Dispersants 10
5.1. Dispersant Classes 11
5.1.1. Ligninsulfonic Acids 11
5.1.2. Formaldehyde Condensation Products 12
5.1.3. Protein Condensation Products . . . 14
5.1.4. Polymers 14
5.1.5. Surfactants 15
5.1.6. Polyphosphates and Derivatives . . . 15
5.1.7. Other Dispersants 15
5.2. Properties and Mode of Action . . 15
5.3. Applications 17
5.3.1. Conditioning of Dyes or Pigments . 18
5.3.2. Disperse Dyes and Dispersants for the Textile Industry 18
5.3.3. Production of Polymer Dispersions . 19
5.3.4. Application of Polymer Dispersions 20
5.3.5. Adhesive Dispersions 22
5.3.6. Detergents 22
5.3.7. Photographic Industry 22
5.3.8. Agriculture 23
5.3.9. Cosmetic and Pharmaceutical Industry 23
5.3.10. Cleaning and Polishing Agents . . . 24
5.4. Testing 24
6. References 25

1. Terminology

If two or more phases that are insoluble or only slightly soluble are finely distributed in one another, the term *disperse system* or, more simply, *dispersion* is used. The continuous, or external phase is termed the *dispersion phase* or *dispersion medium,* while the substance (finely) distributed in it is termed the *internal phase* or *dispersed phase.* The preparation of a dispersion is known as *dispersing,* and the products that are able to promote or stabilize the formation of a dispersion are termed *dispersants* [1–3].

Table 1 lists disperse systems on the basis of the state of aggregation of the external and internal phases [4], [5]. Unlike true solutions, the disperse systems are termed *sols,* and this term is qualified on the basis of the dispersion medium used. If the latter is a gas, the sol is described as an *aerosol;* if liquids are used, the sol is described as a *lyosol;* and if a solid is used, the sol is described as a *solid sol.* In describing the solid sols, the establishment of concepts is difficult since there are no particles in the sense of kinetic entities and often not even any clearly defined interfaces.

Literature often fails to make a precise distinction between *dispersions* and *suspensions.* In describing disperse systems, the term *dispersion* should be used as the generic term, with solid/liquid systems then being described as *suspensions.*

Several peculiar terms persist in general use. The term "photographic emulsions" is used even though solid silver halide particles are finely distributed in the gelatine, so that the term "suspensions" would be more correct. The term "dispersant" is applied to products or product mixtures that are used to disperse solids finely in a liquid phase, whereas on the basis of Table 1 the term "suspension agent" is more appropriate. Paints in which binders and pigment particles are stably dispersed in an aqueous phase are described in English use as "emulsion paints."

The mean particle diameter determines the degree of dispersion (Table 2) [5]. The particle size is decisive in determining the behavior

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_577

Table 1. Disperse systems

Disperse systems (sols)	Dispersion medium (continuous phase)	Dispersed substance (dispersed phase)	Description
Aerosols	gas	gas	
	gas	liquid	mist
	gas	solid	smoke
Lyosols	liquid	gas	foam
	liquid	liquid	emulsion
	liquid	solid	suspension (dispersion)
Solid sols	solid	gas	solid foam
	solid	liquid	solid emulsion
	solid	solid	solid suspension

of a disperse system under the action of gravity. If the particles are colloidal, they are kept in suspension by Brownian movement. However, this is not the case in disperse systems containing coarse particles, which must be stabilized by adding stabilizers.

Table 2. Degree of dispersion [2]

Dispersion	Mean particle diameter, m
Coarsely disperse	$> 10^{-6}$
Colloidal	10^{-6} to 10^{-9}
Highly disperse	$< 10^{-9}$

In most disperse systems, the particle size is variable and Figure 1 shows the particle size distribution F. If the particle size distribution curve is very narrow (A), the system is said to be monodisperse [6], but if the distribution curve is sufficiently wide, the system is termed polydisperse (B). If the concentration of monodisperse spherical particles is sufficiently high, the configuration of the internal phase is one of closely packed spheres. Since the areas of contact between spheres of equal size are greater than those of spheres of different size, the viscosity of monodisperse systems is higher than that of polydisperse ones [7]. In the case of polydisperse systems, achievement of a higher volume fraction of particles is also possible. For an account of the rheology of concentrated suspensions, see [8].

The goal of particle size depends on the particular application. For the manufacture of ionexchanger beads, a mean bead diameter of ca. 1 mm is desirable, while for microemulsions the particle diameter is about 10^{-8} m [9].

The energy required to produce a dispersion increases with increasing division of the particles because their surface area dS increases. The energy is proportional to the work dW which has to be performed. The proportionality factor is the interfacial tension γ between the phases to be dispersed:

$$\mathrm{d}W = \gamma \mathrm{d}S \tag{1}$$

As γ decreases, so does the energy that must be expended to produce a dispersion. For this purpose chemicals – the dispersants that lower the interfacial tension of the phases to be dispersed – are added to a disperse system.

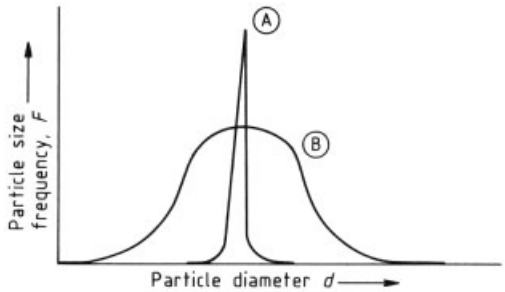

Figure 1. Particle size frequency F as a function of the particle diameter d
A) Monodisperse emulsion B) Polydisperse emulsion

Once produced, particles must be protected against immediately coalescing again. The added dispersing agents, such as emulsifiers and ionic and nonionic surfactants, as well as those dispersants that consist of macromolecules and polyelectrolytes, not only promote dispersion during the preparation of a disperse system, but also make a decisive contribution to stabilizing the particles produced [10]. They create barriers that counteract the natural tendency of individual particles to coalesce again. Thickening agents that increase the viscosity of disperse systems are often added to maintain these barriers.

2. Production of Disperse Systems

The economic advantage, which is associated with the dilution of an active substance in a dispersion, as well as the safety of using aqueous systems rather than the pure active substance of its solution in an organic solvent, has provided a considerable stimulus for the production and application of disperse systems. In principle, there are two methods of producing disperse systems [3]: (1) *condensation,* in which the disperse phase is built-up from atomic or molecular units, and (2) *fine division,* in which coarse particles are reduced in size.

There are many examples of condensation, such as the production of metal sols [11], of pigments with a precisely defined crystal size and shape or the production of latices by emulsion polymerization (→ Emulsions).

In the case of fine division, the energy required to produce a dispersion may be introduced into the system mechanically, chemically, electrochemically, or electrically.

Gases, liquids, or solutions that are not overly concentrated may, e.g., be atomized or sprayed. Liquid phases may also be nebulized [12], [13]. A number of emulsifying machines have been developed for incorporating liquids into aqueous phases; see also → Emulsions. Solids generally must be comminuted mechanically before being introduced into another phase [14], [15], and suitable machines such as mills [16] or ultrasonic generators [17] are used. Finely divided solids are then incorporated in the external phase by stirring or kneading.

A number of studies have been made of the dispersion of particles in shearing fields [16] and in turbulent flows, without any standard theory of mechanical dispersion emerging. The difficulty of arriving at results of fundamental importance is demonstrated by the work of Koglin, Pawlowski, and Schnöring [18].

Despite the existence of machines to produce disperse systems, natural or synthetic dispersants are used which reduce, or almost completely replace, the required mechanical energy, → Emulsions.

3. Stabilization of Disperse Systems

[5], [9], [10], [19–23]

Particles that are dispersed in a dispersion medium are attracted by van der Waals forces. On the other hand, these particles are repelled by electrostatic forces, steric interactions, and Born and solvation interactions. The latter two forces only have a short range. If the forces of attraction dominate, the particles coalesce. If the forces of repulsion dominate, the particles can move freely and a more or less stable system results.

3.1. Attraction Forces

Calculation of van der Waals forces is only possible by using quantum mechanics. The interaction between particles having different geometries in vacuo or in a specified medium has been calculated, using the polarizability of the molecules or the dielectric constants of the phases as starting points [24].

Thus, for example, the following relationship for the energy of interaction V_a has been cited for the special case of similar spherical particles in vacuo [25]:

$$V_a = -(A/12)(a/d) \tag{2}$$

where a is the particle radius, d is the distance between two particle surfaces, which must be distinctly less than a ($d \ll a$), and A is a material constant which is known as the Hamaker constant. The force is obtained by differentiating V_a with respect to d. If the particles are contained in a liquid, a corrected Hamaker constant A_k is approximately valid:

$$A_k = \left(\sqrt{A_{11}} - \sqrt{A_{22}}\right)^2 \tag{3}$$

where A_{11} and A_{22} are the Hamaker constants of the dispersed substance and the dispersion medium, respectively. A_k becomes small if the difference in Eq. (3) decreases; i.e., the attractive forces between dispersed particles can be reduced in two ways: They approach zero (1) if the substance to be dispersed and the dispersion medium have similar Hamaker constants or (2) if the auxiliary agent that is used has a Hamaker constant equal to that of the dispersion medium.

When the particles are enveloped in a solvating layer, the enveloped particles more or less acquire the Hamaker constant of the external phase [26–28]. In this connection hydrate formation of dispersants in aqueous systems is significant [29–31]. Some Hamaker constants are listed in [32]. The Hamaker constants vary with distance since the coupling between the electron oscillators decreases continuously as the distance between them increases [24].

3.2. Repulsion Forces

3.2.1. Electrostatic Interaction

The electrostatic charging of particles dispersed in liquids may occur as a result of the following processes:

1) Rubbing of the particles against each other or against walls of the equipment
2) Dissociation of ionic groups at the surface of the particles
3) Adsorption of ions or oriented dipoles from the dispersion medium on the dispersed particles
4) Adsorption of ionic surfactants on hydrophobic particles
5) Ion exchange between ions of the internal and external phases
6) Difference in the solubility of oppositely charged ions of the disperse phase (isomorphous substitution, e.g., in the case of inorganic auxiliary agents by replacement of Mg^{2+} by Al^{3+}).

In weakly polar dispersion media, only the first three mechanisms are important.

An electrical double layer then tends to form around the particles because they attempt to capture oppositely charged ions from the solution to reestablish electrical neutrality. Brownian molecular motion, on the other hand, tends to distribute the oppositely charged ions uniformly in the external phase. The two opposite effects result in the formation of a diffuse double layer whose characteristic thickness l_D can be determined from the Debye–Hückel theory [21], [33] and by the Poisson–Boltzmann equation. This indicates that the extent of the layer must increase with the absolute temperature T and the dielectric constant ε of the dispersion medium. It decreases as the electrolyte concentration n increases since the particles shield each other over paths of decreasing length, and it decreases as the charge q of the electrolyte's ions increases. The results is that

$$l_D = \sqrt{\frac{\varepsilon k T}{2e^2 n z^2}} \tag{4}$$

where k is the Boltzmann constant, e is the elementary charge, and z is the valence of a symmetrical electrolyte. At large distances from the particle surface, the effective charge approaches zero. This obviously means that particles that are widely separated have no effect on each other. Repulsive forces only come into operation if the particles are so near that their ion clouds are in contact.

The electrostatic interaction of dispersed particles can be assessed quantitatively if the potential gradient in the diffuse electrical double layer is known. For the energy of interaction V_w of spherical particles of equal size having radii a, which should be large compared with $1/x$, and having low, equal interfacial potentials ψ_0, the following approximation is valid [21]:

$$V_w = (1/2)\,\varepsilon a Y_0^2 \cdot \ln\left[1 - \exp\left(-x \cdot d\right)\right] \tag{5}$$

where d is the shortest distance between the particles. It is impossible to determine ψ_0 experimentally since moving particles entrain films of liquid, but the zeta potential ζ is measured in the shear surface of the diffuse double layer of the particles; this is not identical to the interfacial potential, but both quantities are virtually identical in weakly polar phases.

For methods of measuring ζ, see also [9, p. 216], [32].

3.2.2. Steric Stabilization

Steric stabilization of colloidal dispersions by means of natural polymers has been in use for ca. 5000 years, when instant inks for papyrus were prepared in Egypt by mixing lamp-black with a solution of a natural steric stabilizer (e.g., casein from milk, egg albumin, or gum arabic). The suspension was then poured into mold and dried. When required for use, it was simply dipped in water and the sterically stabilized soot particles redispersed spontaneously.

Steric stabilization is used in many industrial products, such as paints, glues, inks, pharmaceutical and foodstuff emulsions, latices, or dispersions of magnetic pigments for manufacturing magnetic tapes.

Steric stabilization has several advantages over other methods of stabilization: (1) it is relatively insensitive to high electrolyte concentrations; (2) the dispersions can be redispersed; (3) dispersions with a high solids content have relatively low viscosities; and (4) steric stabilization is equally effective in aqueous and nonaqueous systems. Sterically stabilized dispersions also exhibit good stability with respect to freezing and thawing.

In contrast to electrostatic repulsion, steric repulsion does not have a long range, but repulsion increases asymptotically at small distances. This latter situation is also described as hard interaction, as distinct from (soft) electrostatic interaction. Since steric repulsion is stronger than van der Waals interaction, sterically stabilized dispersions such as, e.g., latices in dispersion media that are good solvents for the stabilizer, can be thermodynamically stable; therefore, they do not fit into the lyophobic–lyophilic colloids scheme.

Stabilizers. Steric stabilization is caused by macromolecules or by surfactant molecules which are (1) absorbed, (2) grafted on, or (3) form part of the particle.

The best stabilization for colloidal dispersions is produced by macromolecules that form part of the particle, are anchored to it by a chemical bond, or are grafted onto it. Modifications of this type have been carried out with clays, bentonites, and SiO_2 particles [34]. With adsorbed polymer stabilizers, a distinction must be made between the homopolymers and block or graft copolymers.

For adsorption to occur, the interaction between the polymer and the particle must be strong (hydrogen bond formation, dipole forces, or dispersion forces) and that between the polymer and the solvent weak; otherwise, the shear stress produced in a collision would cause lateral movement or even desorption of the stabilizer. On the other hand, for steric stabilization the interaction between the polymer and the solvent must be good. Therefore, steric stabilization is not easily achieved with homopolymers. However, the problem can be solved by using block and graft copolymers in which insoluble anchoring segments are combined with soluble stabilizing units. Simple examples of effective steric stabilizers for aqueous systems are poly(vinyl alcohol – vinylacetate) graft polymers or poly-(styrene – oxyethylene) block polymers and, for nonaqueous systems, poly(acrylonitrile –styrene) block polymers and poly(vinyl chloride – isobutene) graft polymers.

Most polymers are present in solution as a random coil, which is adsorbed at the interface, where in some cases it remains unchanged (Fig. 2 A), but in others it spreads out (Fig. 2 B). If the coils spread out tails (a), loops (b), and trains (c) may be produced at the interface.

Figure 2. Models representing the adsorption of tangled molecules on solid surfaces
A) Adsorption as a tangled molecule B) Adsorption as a spread-out macromoleculea) Tail; b) Loop; c) Train

Depending on the concentration and structure of the polymer and the interactions of the systems polymer – substrate, polymer – solvent, polymer – polymer, and solvent – substrate, very different configurations may be produced and the layer thicknesses may vary considerably, Graft polymers with good anchoring structures may attach themselves so that the main chain forms the train and the grafted side chains form the tails (Fig. 3).

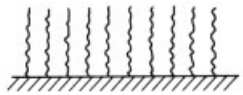

Figure 3. Configuration of graft polymers at a solid surface

With block polymers, it is believed that almost any relationship between loops and tails may be established, depending on the arrangement of the anchoring blocks.

The thickness of the polymer layer (10 – 100 nm) can be determined from the hydrodynamic radius of the particle by viscosity measurements, electrophoresis, and dynamic light scattering [35] or by means of neutron scattering experiments [36]. The interactions of

polymer layers on mica have been measured [37].

Stabilization Mechanisms [22]. The principles of steric stabilization are now well understood, an essential distinction being made between entropy stabilization and osmotic stabilization.

Entropy Stabilization (volume restriction). When two particles approach each other, overlapping takes place between the adsorption layers consisting of surfactants or polymers, which restricts the number of possible configurations (Fig. 4).

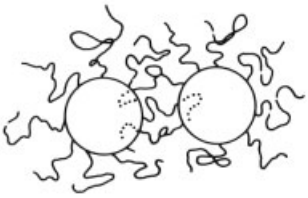

Figure 4. Volume restriction

MACKOR first calculated the loss of configurational entropy for a small rigid rod attached to a platelike particle, otherwise free to move, as a result of the approach of a second platelet [38]. This theory was later extended to spherical particles and to neighboring molecules, that obstruct each other [39]. Other theories take account of the flexibility of the chains [22]. However, because entropy stabilization could not explain the coagulation of a dispersion with rising temperature, a second stabilization mechanism had to be looked for.

Osmotic (Enthalpic) Stabilization. When two particles approach each other, solvent is displaced from the region of macromolecular overlap (Fig. 5). As a result, the concentration of the macromolecules in the region of overlap increases and osmotic forces are produced, pulling the solvent back into this region and, thus, producing a separation of the particles [40]. The osmotic stabilization is greatly influenced by temperature and solvent. This explains the dependence of steric stabilization on the solvent and the flocculation of a sterically stabilized dispersion at the Θ-point, the temperature at which interaction between macromolecules and solvent is just as high as that between the segments of the macromolecules.

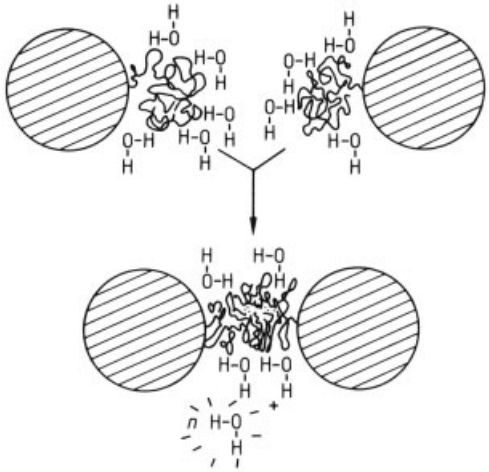

Figure 5. Schematic illustration of osmotic stabilization

Temperature Dependence of Steric Stabilization. Depending on whether stabilization is entropic or enthalpic, destabilization may occur on cooling or heating (Table 3). Stabilization in aqueous media is predominantly enthalpic, and in nonaqueous media it is predominantly entropic [41]. The effect of temperature is more pronounced for steric stabilization than for electrostatic stabilization, and a sharp transition occurs from long-term stability to flocculation, which is generally reversible, at the critical flocculation temperature (CFT). This temperature may be affected by the presence of an electrolyte.

Further Polymer Effects. Steric stabilization becomes more effective as the molecular mass increases. Complete coverage of the particle surface is required. If this latter condition is not fulfilled, the dispersion is destabilized. In addition, free polymers can also affect the stability of dispersions.

Bridging Flocculation. If the dispersed particles are only partially covered with polymers and the macromolecules of the latter have a sufficiently high molecular mass, they may anchor themselves to two particles at the same time, thus forming polymer bridges and causing flocculation to occur. This phenomenon is termed bridging flocculation [43], [44] and is of great importance in clarifying wastewaters, processing minerals, precipitating sols, and dewatering sludges. These applications require a few mg/kg of a high molecular mass polymer (molecular

Table 3. Thermal stability of sterically stabilized emulsions [42]

Emulsion type	Stabilizing moiety [a]	Anchor polymer [a]	Disperse phase	Dispersion medium	Flocculation	Stability type	CFT, [b] K
Oil-in-water	PEO	PVAc	toluene	0.09 M $MgSO_4$	heating	enthalpic	318 ± 2
Oil-in-water	PVA	PVAc	toluene	2.0 M NaCl	heating	enthalpic	297 ± 3
Oil-in-water	PAM	PVAc	toluene	2.1 M $(NH_4)_2SO_4$	cooling	entropic	297 ± 2
Oil-in-water	PAA	PVAc	toluene	0.2 M HCl	cooling	entropic	298 ± 2
Water-in-oil	PLM	PAA	water	1-pentanol	cooling	entropic	301 ± 3
Water-in-oil	PS	PEO	water	cyclohexane	cooling	entropic	308 ± 3

[a] PEO = poly(ethylene oxide); PVA = poly(vinyl alcohol); PAM = polyacrylamide; PAA = poly(acrylic acid); PLM = poly(lauryl methacrylate); PS = polystyrene; PVAc = poly(vinyl acetate);
[b] CFT = critical flocculation temperature

mass at least 10^6) such as polyacrylamide. The polymers may be nonionic, anionic, or cationic. It should be noted that ionic polymers of even relatively low molecular mass (ca. 10^4) may initiate flocculation, but in this case the surface charge density is reduced by charge neutralization.

Sensitization. The way polymer solutions are mixed with dispersions determines the result decisively because of kinetic irreversibility of the adsorption. Dispersions, in which particles are unevenly covered with polymers, will become more sensitive to the effect of electrolytes than those containing particles with no polymer layer on them; see, for example, [43], [45].

Depletion Flocculation and Depletion Stabilization. The question of depletion flocculation and depletion stabilization with high polymer concentration and high solids content has been under discussion for several years [46]. Both effects are initiated by freely dissolved nonionic polymers.

Figure 6 shows a plot of free energy versus distance for spheres in aqueous solutions of poly-(ethylene oxide) for different polymer concentrations. If the maximum is higher than 15 – 20 kT, *depletion stabilization* occurs (a). If the maximum is between 5 kT and 15 kT, *depletion flocculation* results (b), and if the maximum is smaller than 5 kT, no effect can be observed (c).

Thus, at relatively low concentrations flocculation occurs (even in sterically stabilized dispersions), whereas a high concentration results in stabilization.

When two particles approach each other, the concentration of free polymer between them is decreased. In a good solvent this is energetically unfavorable, and the energy of the system increases. If the particles come still closer together, the polymer is almost completely displaced from the volume between them, which consequently is filled with almost pure solvent. At this point a closer approach is energetically favorable since it results in mixing of almost pure solvent with polymer solution, and the result is depletion flocculation. If it is not possible to displace the polymer from the volume between the particles, depletion stabilization occurs.

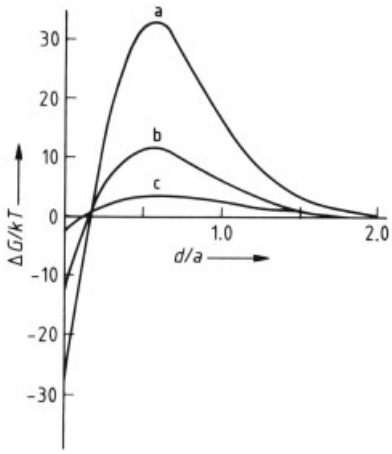

Figure 6. Energy diagram for two spheres (radius 77.5 nm) for aqueous solutions of poly(ethylene oxide) (M_r 20 000) at polymer volume fractions of a) 0.07; b) 0.04; and c) 0.02

Depletion stabilization differs from steric stabilization in two respects: (1) the stabilizing polymer can move freely in the solution and is not adsorbed on the particle and (2) stabilization due to free polymers is kinetic (and thermodynamically metastable) whereas stabilization due to adsorbed polymers is thermodynamic.

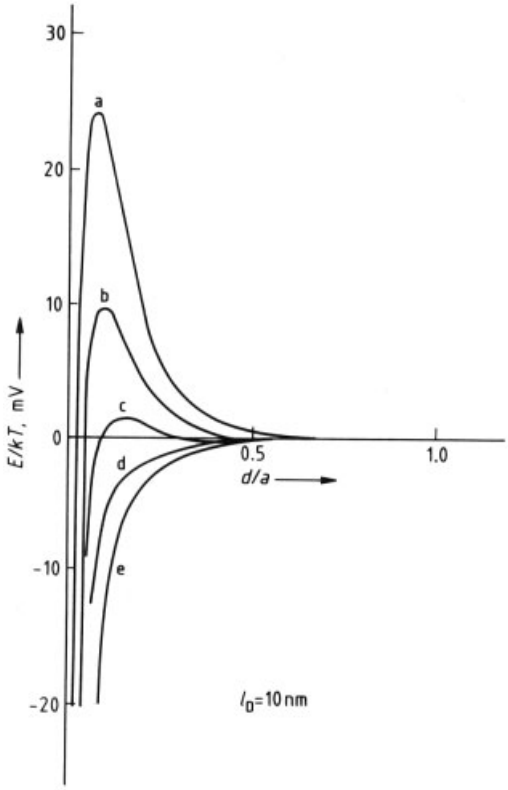

Figure 7. Variation in potential between two particles according to the DLVO theory [19], [21], [50] at various electrostatic potentials
a) ψ_0 32 mV; b) ψ_0 25.6 mV; c) ψ_0 19.2 mV;d) ψ_0 12.8 mV; e) ψ_0 0 mV

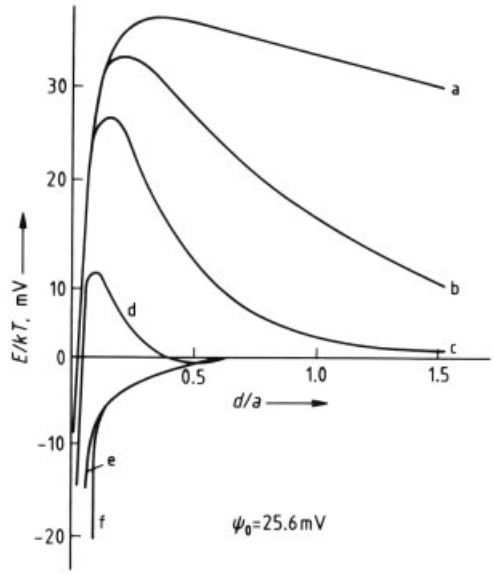

Figure 8. Variation in potential between two particles at varying Debye length l_D [19], [21], [50]
a) $l_D \to \infty$; b) l_D 100 nm; c) l_D 33.3 nm; d) l_D 10 nm; e) l_D 3.3 nm; f) l_D 1 nm

3.3. DLVO Theory

Precise knowledge of the potential as a function of the distance between two particles in a dispersion medium would be an excellent aid in assessing a dispersion. Initially, the variation in potential was determined for model dispersions of spherical particles with a narrow size distribution [5], [36], [47] and for dispersions of platelike particles [48] by measuring the osmotic pressure [5], [48]. Dynamic light scattering [47] and neutron scattering experiments [48], [36] were also carried out. For the purpose of calculation, the teams of Verwey and Overbeek [21] and Derjaguin and Landau [49] (DLVO) independently superimposed the interaction of the energy of attraction E_{attr} and repulsion E_{rep} on each other. The resultant energy of interaction E between two highly charged, spherical, dispersed particles is then given by

$$E = E_{\mathrm{rep}} + E_{\mathrm{attr}} \tag{7}$$

Figure 7 shows the resultant interaction potential E versus the reduced distance d/a of the particles. The particle radius a is 100 nm and the Hamaker constant A has a value of 10^{-19} J. The electrostatic potential ψ_0 varies. For $\psi_0 = 12.8$ mV, a decreasing potential results in an unimpeded coalescence, whereas doubling ψ_0 to 25.6 mV ensures complete stability.

If two particles approach each other at an electrostatic potential of $\psi_0 = 19.2$ mV, the approach is initially favored by a weak minimum. Agglomeration results if the thermal energy of the particles is insufficient to overcome this minimum. The stability of the agglomerates depends

on the depth of the minimum. If the droplets overcome the minimum as a result of their thermal energy, they are again able to move freely in the external phase. Coalescence is initially prevented during further approach because work must be performed in order to overcome the rise in the potential energy.

To prevent the minimum which results in agglomeration, the electrostatic repulsion potential must be increased.

In Figure 8 the electrostatic shielding length, l_D (Debye length) is varied at a constant electrostatic potential of 25.6 mV and a Hamaker constant A of 10^{-19} J. If the Debye length is increased > 10 nm, the stability of a dispersion constantly increases. However, if the Debye length is reduced, the electrostatic repulsion potential becomes virtually insignificant, just like case (e) in Figure 8 [19], [21], [50].

In its original form, DLVO theory applies to highly charged boundary layers with no account taken for steric factors. Expanded model calculations are now available for weakly charged and uncharged boundary layers [51]. To assess the stabilizing action of polymeric substances, the Flory–Huggins parameter was included in DLVO theory. Usable results are claimed if the polymers that are used have no appreciable interfacial properties [52].

4. Destruction of Disperse Systems

4.1. Forces and Mechanisms

To destroy a disperse system, the forces that stabilize a dispersion must be weakened or eliminated while the attraction forces and bonding mechanisms operating between the dispersed particles must be strengthened.

As a result of Brownian molecular motion and van der Waals forces, the dispersed particles approach each other so closely that a combination may take place (cf. Section 3.1). Figure 9 shows the bonding mechanisms that may operate in such a combination. Coalescence of the particles may take place with (A) or without (B) the formation of a material bridge.

When a *material bridge* is formed between two particles, intergrowth may be caused by crystallization effects, but the particles may also combine by fusion or sintering. This produces aggregates that can no longer be dispersed. *Liquid bridges* (b) also enable particles to combine. Substances may crystallize inside the liquid bridge, which causes the particles to fuse together, and if this occurs, redispersion is no longer possible. Liquid bridges may be formed by highly viscous liquids or by capillary forces; it depends on the forces inside the bridge whether indestructible aggregates are formed or whether agglomerates that can be redispersed by supplying sufficiently high mechanical energy are formed.

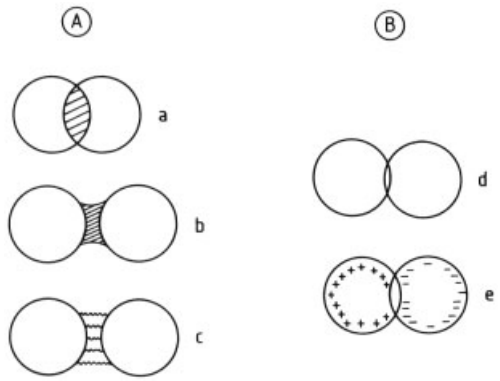

Figure 9. Bonding mechanisms between particles with the formation of a material bridge (A) and without the formation of a material bridge (B)
a) Intergrowth; b) Liquid bridges; c) Steric bridges; d) Coalescence; e) Electrical forces

Steric bridges (c) are formed when polymers are used and when the interfaces are incompletely covered (bridging flocculation,see page 7); these bridges generally result in the formation of aggregates that flocculate.

Without the formation of a material bridge, two particles coalesce if the liquid layer between them is ruptured when they approach each other, and coalescence (d) takes place if van der Waals attraction forces considerably exceed the repulsion forces. Coalescence leads to aggregation, and redispersing is then no longer possible. Oppositely charged particles also combine without a material bridge being formed (e), and the addition of oppositely charged substances is a method used to destroy a disperse system. Thus, a dispersion that has been stabilized, for example, by an anionic surfactant is broken by adding

a cationic surfactant. As a rule, a sparingly soluble complex is produced which precipitates from the solution.

4.2. Practical Methods

To destroy a dispersion, the mechanisms that counteract agglomeration or aggregation must be eliminated. In addition, the molecules absorbed on the particle surfaces must be removed. Following is a list of the possible methods to destroy a dispersion:

1) *High temperature* increases the thermal energy, which overcomes the repulsive forces. Elevated temperature may also change the adsorption conditions, producing, for example, melting of the dispersant, increased crystal growth, or a change in the crystal form.
2) *Electrolyte addition* reduces the electrical double layer and the electrostatic repulsion potential. Electrolytes may also have a dehydrating action.
3) *Electric fields* result in the particles being discharged.
4) *Desorption of dispersant particles* can be achieved by precipitating them or by displacing or detaching them from the particle surface. The dispersants can be precipitated by using flocculating agents or by decreasing the solubility of the dispersant either by adding a salt or by changing the pH. The addition of a less effective dispersant may displace the more effective one. The adsorption of the dispersant by competing adsorption to another substance also leads to dispersant detachment.
5) *Mechanical loads,* particularly a *shearing load,* may destroy a dispersion.
6) *Freezing* may destroy the dispersion.

5. Dispersants

Definitions. *Dispersants* are products or mixtures of products that can promote the formation of a dispersion or stabilize a dispersion. The term *dispersion* is applied to a system of several phases in which one is continuous and at least one other is finely distributed [53]. The preparation of a dispersion is called *dispersal.*

A dispersant generally does not change the solubility of the substance to be dispersed. In contrast, *hydrotropic substances* increase the solubility of a substance that is sparingly soluble in water; the more general term is a *lyotropic substance* [54]. Whenever the additive is a surfactant, the term *solubilization* is used [55]. *Sequestration* occurs when sparingly soluble salts are solubilized by chelating their cations [54]. In practice, no sharp division exists between dispersants (no solubilization) and lyotropic substances (significant solubilization). In general usage, the term *dispersant* is used for a product that can disperse sparingly soluble solids in liquid phases and keep them finely dispersed [56]. In the strict sense, it is more precise to use the term *suspending agent* whenever a solid is finely dispersed in a liquid (→ Emulsions).

The term *flocculation protection* describes a dispersant used to prevent a solution from turning into a colloidal dispersion or becoming turbid [54], [58]. The re-formation of a stable dispersion from flocks or aggregates is called *peptization* or *deflocculation* [54].

In a dispersion of solid particles, such primary particles as agglomerates and aggregates may be encountered. *Agglomerates* are combinations of coalesced particles that can be redispersed. In contrast, *aggregates* can no longer be dispersed because the particles have, for example, undergone intergrowth [57].

5.1. Dispersant Classes

Dispersants are generally polyelectrolytes that are readily soluble in water. The most common representatives are alkali-metal polycarbonates, polysulfonates, or polyphosphates, usually sodium salts. The use of ligninsulfonates and of condensation products of aromatic sulfonic acid with formaldehyde is very widespread.

At present, polyphosphates are still used in the detergent industry, e.g., for colloidally dispersing dirt detached from the fibers and for sequestering calcium and magnesium ions in the detergent liquor. In low-phosphate detergents, polyphosphates are completely or partially replaced by zeolites (→ Laundry Detergents) [59]. Polycarbonates have also been proposed for this purpose [60]. The structural details of common

dispersants have not yet been fully elucidated; therefore, only schematic structures are given in the following for the various dispersants. In the case of formaldehyde reaction products, the degree of condensation is not known since. When two to ten aromatic nuclei are linked together, there is no reliable method known for determining the molecular mass.

5.1.1. Ligninsulfonic Acids

ligninsulfonic acids [*8062-15-5*] are made from native lignin [*9005-53-2*]. Two different starting materials are available: (1) lignin made from sulfite pulping and (2) lignin made from alkaline (kraft) pulping of wood (→ Lignin). In the first case, the sulfite waste liquor is purified and marketed as a liquid or a solid. Lignin from kraft pulping is first purified and then sulfonated.

Industrial ligninsulfonates are globular colloids with molecular masses between 2000 and 100 000 whose structures have not yet been fully elucidated. Their basic structural unit is phenylpropane. The structural formula shows five structural units linked with the parent unit in various ways (Fig. 10) [61]. The degree of sulfonation varies between 0.3 and 1.0 sulfite groups per phenylpropane unit [57].

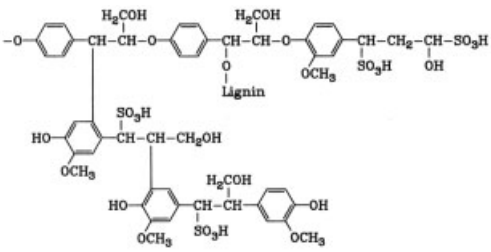

Figure 10. Partial structure of ligninsulfonate

The suitability of ligninsulfonates for various industrial purposes depends on the purity, degree of sulfonation, number of hydroxyl groups, and molecular mass distribution of the sulfonates. In addition to highly dispersive action during milling and ease of redispersion, ligninsulfonates must be easily soluble and the dispersions stable at higher temperature. These properties are achieved by suitable purification and chemical treatment of the crude product (→ Lignin).

Ultrafiltration of the sulfite waste liquor can be used to separate the ligninsulfonates into high molecular mass and low molecular mass fractions. Products with a low degree of sulfonation and an increased hydroxyl and carboxyl content are produced by *alkaline and oxidative treatment* [62]. Dispersants for pigments and dyes can also be prepared by *thermal degradation* of ligninsulfonic acid [63].

Ligninsulfonates with a relatively high degree of sulfonation are very effective milling additives. On the other hand, products with fewer sulfonic acid groups are more thermally stable; these products adhere more strongly to the hydrophobic surfaces of the dye particles than highly sulfonated products and are not as readily separated by Brownian thermal movement [64]. Therefore, thermally stable dispersions are obtained by using dispersants made from kraft lignin with few sulfonic acid groups [65] (→ Lignin). Leveling agents for dyes have long been prepared by reacting the ligninsulfonates with various monofunctional (aniline or chloroalkanes) or bifunctional (ethylene oxide or aliphatic diamines) compounds [66]. Similar reaction products have been proposed as dispersants for paint manufacture and dyeing [67]. For example, Schutzkolloid 63-Bayer (protective colloid 63) is made by reacting ligninsulfonic acid with aniline.

Disadvantages of ligninsulfonates are their reductive effect on sensitive azo dyes in high-temperature dyeing (130 °C) and their tendency to soil fibers, which with polyamide and wool may lead to a yellow discoloration of the fibers when exposed to light. The foaming ability of these products may also occasionally cause problems. Improved ligninsulfonates have been developed [68].

Commercial product names for ligninsulfonates and their derivatives are as follows [69]: Dekol N, CN (BASF); Lamepon N (Grünau); Lignosol (Lignosol, Canada); Lyocol A powder (Sandoz); Reax brands (Westvaco); Sulfite waste liquor, Waldhof liquor, Zewa powder (Zellstoff Waldhof); Schutzkolloid 63 (Bayer); Ufoxanes, Ultrazines (Messrs. Borregards, Norway); Unisol N powder (PUC-FMC); Irgasol AT (Ciba-Geigy).

5.1.2. Formaldehyde Condensation Products

Table 4 summarizes important dispersants produced by condensation of aromatic compounds with formaldehyde. The following describes a typical production of a dispersant of this category. With suitable variations it can be applied to a number of products formed by condensation with formaldehyde [*50-00-0*]. [70], [71], [72], [73], [74], [75], [76], [77], [78], [79]

Naphthalene (200 parts) is reacted with 200 parts of 98 % sulfuric acid for 5 h at 160 – 165 °C in an iron tank. After the solution has been cooled to 80 °C, the mixture is diluted with 90 parts of water in a lead-lined stirring vessel. Then 107 parts of a 30 % formaldehyde solution are added and the mixture is stirred for 20 h at ca. 80 °C. Toward the end of this period, the mixture is gradually heated to 100 °C. Neutralization is carried out at 20 – 25 °C with ca. 165 – 180 parts of a 25 % ammonia solution. The neutralization product is then filtered and, if necessary, dried in a vacuum drum drier [80].

Numerous variations of this synthesis are possible, and a wide range of aromatic compounds and their derivatives can react with aldehydes, ketones, and compounds that eliminate aldehyde groups. The following three typical examples supplement those listed in Table 4:

1) Dispersants produced by condensation of aromatic sulfonic acids and benzyl chloride [*100-44-7*] or benzoin [*119-53-9*] [81]
2) Dispersants produced by condensation of various alkylarylsulfonic acids with a halogen arylsulfonic acid [82]
3) Dispersants produced by condensation of sulfonated phenols, 2-naphthols, etc. with formaldehyde and various nitrogen compounds [83]

Finally, special processes have been developed for improving the quality of dispersants or for adapting them to special applications [57]. Thus, the following specific products can be prepared:

Products with low sulfate content [84]
Low-foam, alkali-resistant sulfomethylation products obtained from mono- and polynuclear phenols [85]
Thermostable condensation products [86]
Low-foam condensation products obtained from diphenyl derivatives [73]
Ethoxylated and sulfated condensation products obtained from alkylphenols and formaldehyde [87]
Thermostable sulfomethylation products [88]
Phenol – formaldehyde condensation products containing sulfonic groups and having improved dispersive action [89]
Products that have no reducing action on sensitive dyes
Almost colorless sulfomethylation products that are also suitable for optical brighteners [85], [90]
Dispersants that do not stain fibers [87], [91]
Postcondensed products with improved dispersive properties [92]
Phenol – formaldehyde – hydrogen sulfite condensation products that contain nitrogen and are low in alkali concentration

Commercial polysulfonate products are listed as follows [69].
Condensed sulfonates: Avolan IS (Bayer); CHT dispersants SMS (CHT); Dispersant EB (PUC-FMC); Dupramin DS (AKZO); Dyapanol (Yorkshire); Irgasol DAM (Ciba-Geigy); Lyocol O, Lyocol RD, Lyogen PN (Sandoz); Matexil DA-AC (ICI); Osimol DPA (Grünau); Setamol WS (BASF); Solegal A, P (Hoechst).
Liquids: Tamol NH, NNO, NOP (BASF); Tanede DPA (Tanatex); Texamol WS (ACNA); Vultamol (BASF); Wotamol WS (VEB Wolfen); Baykanol PQ, Baykanol LS, Dispersant SS, Dispersant 523 (Bayer); Melmeny (SKW Trostberg), Lissatan AC (ICI); Demol N (Bayer, Japan).

These important dispersants find applications in the following fields [93]:

1) Milling additives and dispersants for pigments and dyes
2) Dispersants and plasticizers in the building industry
3) Dispersants for polishing and abrasive powders
4) Stabilizing agents for resin glue emulsions, e.g., Tamol NOP (BASF)
5) Auxiliary agents in the emulsion polymerization of synthetic rubber
6) Dispersants for fillers, pigments, and vulcanizing agents in natural and synthetic rubber latices

Table 4. Aromatic formaldehyde condensation products

Production	Structural unit	References
Naphthalene [*91-20-3*] + H_2SO_4 + formaldehyde	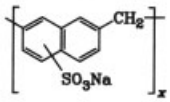	[70]
Naphthalene [*91-20-3*] + cresol [*1319-77-3*] + H_2SO_4 + formaldehyde	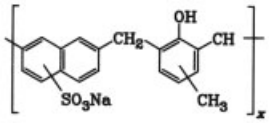	[71], [72]
Diphenyl ether [*101-84-8*] + H_2SO_4 + formaldehyde	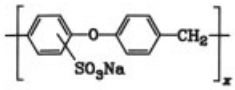	[73]
Toluene [*108-88-3*] + H_2SO_4 + formaldehyde	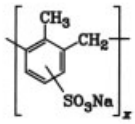	[74]
Isopropylbenzene [*98-82-8*] + H_2SO_4 + formaldehyde	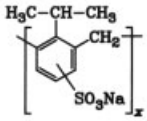	[75]
Cresol [*1319-77-3*] + H_2SO_4 + formaldehyde	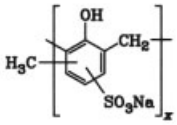	[76]
Phenol [*108-95-2*] + formaldehyde + sodium sulfite	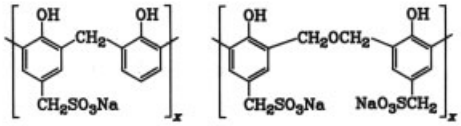	
Cyclohexanone [*108-94-1*] + formaldehyde + sodium sulfite	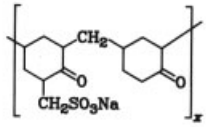	[77]
Maleic anhydride [*108-31-6*] + formaldehyde + sodium sulfite (melamin)		
Phenol [*108-95-2*] + H_2SO_4 + formaldehyde	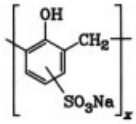	[78]
Aliphatic ketones + formaldehyde + sodium hydrogen sulfite		[79]

7) Boiling-temperature degreasing agents to which Tamol N brands (BASF) have been added to improve the stability of the solution and increase its dirt-suspending power

5.1.3. Protein Condensation Products [69, p. 523]

Protein-derived dispersants are obtained, for example, by condensation of natural protein hydrolyzates from collagen with fatty acids or abietic acid; the products are marketed as potassium or triethanolamine salts [94]. Examples of commercial products are as follows:

$C_{11}H_{23}CONHCHRCO-(NHCHRCO)_n-NHCHRCOOK$
Lamepon S (Grünau)
$C_{11}H_{19}CO-(NHCHRCO)_n-NHCHRCOOK$
Lamepon UD (Grünau)
$C_{19}H_{29}CONHCHRCO-(NHCHRCO)_2-NHCHRCOOH \cdot N(CH_2CH_2OH)_3$
Lamepon PA-TR (Grünau)

Slightly degraded polypeptides are as follows: Lamasan CL (Sandoz); Ofnapon AS, Ofnapon ASN (Hoechst); Protepon (Protex).

5.1.4. Polymers [69]

Water-soluble polymers are used both in the production of dispersions and in the preparation of dye dispersions and in the preparation of polymer dispersions. Examples of suitable compounds are

poly(vinyl alcohol) [*9002-89-5*]
poly(vinyl pyrrolidone) [*9003-39-8*]
poly(vinyl sulfates)
polyacrylates
polyacrylamide [*9003-05-8*] [95]
maleic acid–styrene copolymers

Polymeric compounds often have a synergistic effect in disperse systems. Even small additions may decisively improve the quality of dye dispersions [96]. The polymers probably form complexes with the dispersants, and these complexes have a solubilizing effect on certain dyes. Such complexes have been found in mixtures of poly(vinyl pyrrolidone) and sodium alkylsulfates [97]. Alternatively, the quality-improving effect of polymers on dye dispersions may be caused by complex formation between the polymer and the dye, because poly(vinyl pyrrolidone) and alkyl or alkylphenol polyglycol ether have been found to form complexes with suitable dyes.

A further field of application of polymeric dispersants is the plastic industry. Many polymers are produced as dispersions (latices), and these have found varied and widespread fields of application, often by adding further emulsifiers or dispersants [112].

A distinction is made between self-dispersable polymers such as, for example, polyurethanes, which have ionic and nonionic water-soluble groups, and those polymers in which dispersability must be achieved by adding dispersants [9].

5.1.5. Surfactants

Surfactants [98] are not usually considered as dispersants although they often have dispersive properties. Their use is of decisive importance, for example, in the conditioning of pigments that are sparingly soluble in water or in the further processing of such pigments in polymer dispersions. The wetting properties of the surfactants are particularly important when pigments are mixed with polymer dispersions. In this case, the surfactants must be compatible with the dispersions and cationic surfactants are usually ruled out, because they form insoluble complexes with the anionic dispersants which than destroy the dispersions.

The following anionic or nonionic surfactant classes are particularly suitable: alkylpolyethoxylates, e.g., Ekalin F (Sandoz); alkylphenol polyethoxylates, e.g., Alkol SMK (Sandoz); alkylamine polyethoxylates; carboxymethylated alkylphenol polyethoxylates, e.g., Lyocol AC (Sandoz); polyalkylene oxide compounds and their sulfuric acid or phosphonic acid esters; alkylsulfonates or alkylsulfates; alkylarylsulfonates; sulfated fats or oils, e.g., Lyogen P (Sandoz); salts of fatty or resin acids; reaction products of diamines or triamines with ethylene oxide, e.g., Sandogen NH (Sandoz) [99]; weakly anionic, amphoteric fatty acid amides containing sulfonic groups, e.g., Tergolix A (Sandoz).

In addition to the products already mentioned, mixtures of nonionogenic and anionic dispersants that are supplemented with noninterfering foam-suppressing agents are also commercially available. Examples are Lyocol WPN (weakly anionic) and Lyocol ACN (anionic), both made by Sandoz.

5.1.6. Polyphosphates and Derivatives

Polyphosphates, particularly pentasodium triphosphate [*7758-29-4*], $Na_5P_3O_{10}$, M_r 367.93, have been long used in detergents for dispersing dirt detached from the fibers and sparingly soluble calcium and magnesium soaps in the wash liquor. The use of these dispersants prevents graying of the laundry and improves the degree of whiteness. However, because of the eutrophication problem, detergent manufacturers have begun to replace the polyphosphates with other complexing agents (→ Laundry Detergents).

5.1.7. Other Dispersants

Some dispersants are difficult to classify. Examples are disproportionated abietic acid (Dresinat; Hercules) [70, p. 408] and starch derivatives [80, p. 2769].

5.2. Properties and Mode of Action

Surface Tension. All dispersants that are readily soluble in water do not lower the surface tension γ of water as much as surfactants. Table 5 shows the surface tension of solutions containing 5 g/L of dispersant. As a comparison, the surface tension of a surfactant (dodecyl alcohol heptaglycol ether, DA-7) [55] is quoted for the same concentration and temperature.

Time-dependent measurement shows that the surface tension of dispersants varies with time in a characteristic way (Fig. 11) [101]. This phenomenon is explained by the presence of several compounds, e.g., in Dispersant SS; these compounds diffuse to the surface of their solution at different rates and will be adsorbed there one after another.

Table 5. Surface tension of dispersant solutions *

Dispersant	Surface tension, mN/m
Waldhof liquor	52.7
Schutzkolloid 63	64.7
Baykanol PQ	71.5
Baykanol SL	54.6
Dispersant 523	48.0
Dispersant SS	61.6
Avolan IS	68.3
Melment	71.3
Demol N	71.9
DA-7	31.6

* Surface tension measured by the ring method [100] after leaving the solution for 1 h at 20 °C

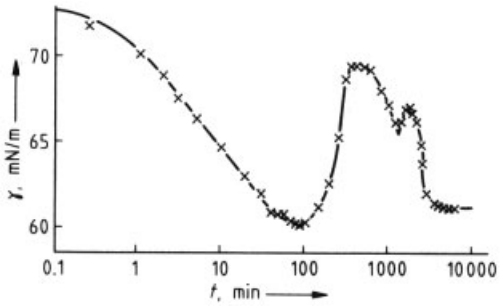

Figure 11. Surface tension of a solution of an anionic dispersant * (1 g/L) as a function of time at 20 °C
* Dispersant SS, made by condensation of cresol, naphthalenesulfonic acid, and formaldehyde

Dispersive Action. The main effect of dispersants is their ability to disperse finely ground solids in aqueous media and to stabilize the dispersion. Although certain models of dispersive action have been developed, these models fail to satisfactorily explain the action of a dispersant [98], [69]. As a phase is further dispersed within another phase, the interface area between the two phases increases. Therefore, the properties of the interface determine the behavior of a disperse system [102], and the action of the dispersant molecules at the solid particle interface acquires special significance. Figure 12 summarizes diagrammatically the possibilities encountered [103]. For monomeric and polymeric molecules, the attractive and repulsive forces and their interaction are treated in more detail in Chapter 3.

Dispersants have several hydrophilic groups along their chain and a comblike structure. In contrast, surfactants have one polar hydrophilic group and a long hydrophobic chain. Therefore

surfactants are alignes with their long axes perpendicular to the particle surface, whereas the absorbed polyelectrolyte molecules are parallel to the particle surface with the polar groups directed outward [104]. The particles of the substance to be dispersed are generally negatively charged; therefore, they push the negatively charged groups of the polyelectrolytes away toward the solution. The recombination of particles by agglomeration or aggregation is counteracted by the following mechanisms:

1) *Electrical repulsion* of the negative groups directed toward the solution. The stability of dispersions depends very strongly on electrolyte additives and the thickness of the electric double layer decreases as the ion concentration increases. Therefore, the stability of the dispersion decreases both as the electrolyte concentration increases and as the valence of the counterions increases (Shultze–Hardy rule). According to calculations, monodisperse systems are more stable than dispersions of compounds with varying radii and surface potentials [69], [105].
2) *Steric action* of the hydrated ionic groups. Dispersants are readily soluble in water, which suggests that their numerous ionic groups have extensive hydrate envelopes. These hydrated groups act as barriers when the particles approach distances at which the attraction due to van der Waals–London forces could become effective.
3) *Structural change of the aqueous phase* by the saltlike nature of dispersants. Figure 13 shows that small additions of a dispersant may raise the cloud point [106] of a polyglycol ether solution significantly. Raising the cloud point is approximately equivalent to increasing the rigidity of the water structure, and this counteracts agglomeration or aggregation.

The action of a solubilizing substance differs from that of a dispersant. Solubilizing surfactants form micellae that can include the sparingly soluble compound in their interior hydrophobic parts [55]. Surfactants have limited solubility in water, whereas dispersants always have a good, often very high, solubility in water.

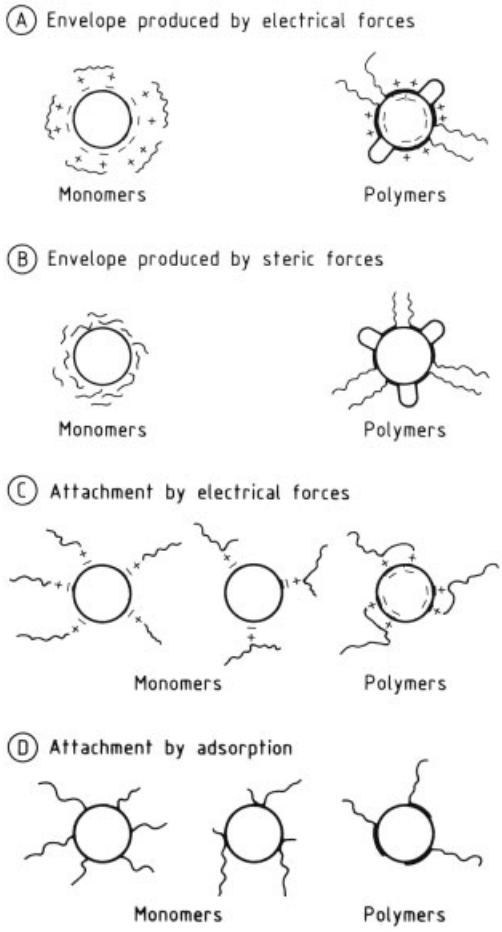

Figure 12. Stabilization of a dispersion

Destruction of Dispersions. The mechanisms and methods that lead to the destruction of a dispersion are treated in Chapter 4.

5.3. Applications

Since the fine dispersion of sparingly soluble solids in aqueous media is an inexpensive process, which presents few hazards, the application of dispersants is varied and widespread. Large quantities of dispersants are used in the following fields:

Conditioning of dyes or pigments
Disperse dyes and dispersants for the textile industry
Production of polymer dispersions

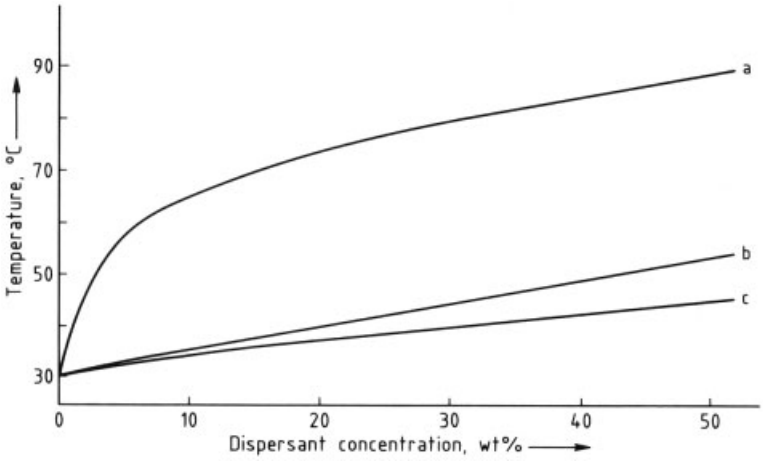

Figure 13. Cloud point of a polyglycol ether solution (NP-8.5: made by reacting 4-isononylphenol with 8.5 mol of ethylene oxide) as a function of dispersant concentration
a) Baykanol PQ; b) Sulfite liquor 28.3 (based on 100 %); c) DM 2883 [77]

Application of polymer dispersions
Adhesive dispersions emulsions/dispersions
Detergents
Photographic industry
Agriculture
Cosmetic and pharmaceutical industries
Cleaning and polishing products

5.3.1. Conditioning of Dyes or Pigments
[57], [69], [71]

Sparingly soluble dyes or pigments are used on a large scale in dispersed form. Such dispersions, actually suspensions, of colorants are industrially important, especially as pigments in binders and as emulsion paints in water or other media. When applied, pigments continue to form a colloidal two-phase system (pigment in binder), whereas disperse dyes, when used in the suspended form, are converted into a molecularly dispersed state on or in the fiber (→ Disperse Dyes).

Disperse dyes are used chiefly for textile dyeing, but pigments find application in many different industry branches such as car paints, industrial paints, indoor and outdoor paints, printing inks in the graphics industry, coloration of plastics, and textile printing.

The synthesis of sparingly soluble dyes produces coarse crystals (filter cakes), the particle size of which must be reduced in the presence of dispersants before or after drying. Various devices are used for grinding the filter cake, depending on the rheological properties [105]. Normally beater additives (e.g., Tamol N brands; Bayer) are used, which are capable of altering the rheological properties of the filter cake.

During grinding the dispersants penetrate the submicroscopic cracks and crevices in the dye particles, assist in separating the primary particles from each other, and protect the particles against agglomeration. However, the dispersants also ensure optimum viscosity for the grinding process and the stability of the dispersion. If the aqueous dispersions are then dried to form powders, further dispersants are added. In this way, the high degree of dispersion required in the final product can be achieved again when the dye is introduced into the dye bath. However, additives are also included in liquid preparations of pigment dispersions to prevent rapid settling and drying [107].

Dyes with hydrophilic groups are easier to disperse than hydrophobic dyes. The stability of dye dispersions also depends on the tendency of the dyes to crystallize, which is influenced by their purity and solubility. Particle size distribution is also a factor, since monodisperse systems are more stable than polydisperse systems [106]. Finally, the stability depends on the dispersant used.

Consi deration must already be given to all these factors during the manufacture of the dye dispersions. Therefore, the choice of dispersant depends both on manufacturing requirements and on the subsequent application. The dispersants used for dyes and pigments are polyelectrolytes such as sulfite waste liquors and their derivatives or sulfonates of aromatic, cycloaromatic, or aliphatic condensation products. Anionic or nonionic surfactants are also used. In addition to their dispersive properties, surfactants enhance the wettability of dye powders [107].

5.3.2. Disperse Dyes and Dispersants for the Textile Industry

After their conditioning vat dyes or disperse dyes are finely divided, generally aqueous, dispersions, pastes, or powders. All three forms must be readily dispersable in the dye bath and the dispersions that are obtained must be sufficiently stable throughout the entire dyeing process. Moreover, the dispersant must not adversely affect the dyeing behavior of the dyes or the results of dyeing. To disperse the sparingly soluble dyes in the dye bath, concentrated emulsions of liquid carriers or leveling agents are usually employed, and the composition of these is described elsewhere (→ Emulsions).

In addition, paraffin, wax, polyethylene, or silicone dispersions are used in the textile industry as waterproofing agents [108], [109]. Inorganic hydrated oxides and aluminum salts and heavy-metal salts of fatty acids such as zirconium stearate are usually added [108].

To disperse crystalline paraffin, ethoxylated sorbitan fatty acid esters, usually mixtures having various HLB values may be used (HLB = hydrophilic – lipophilic balance; for explanation, see → Emulsions). Mixtures of ethoxylated sorbitanyl oleate (Tween 80; Atlas Chemie) and ethoxylated oleyl alcohol have also been used [110] as well as mixtures of ethoxylated oleyl alcohol with various HLB values [111]. Protein – fatty acid condensates are used as codispersants [94].

Waxes may be dispersed with polyethoxylated oleyl alcohol, and polyethylene may be dispersed with ethoxylated fatty alcohols having an HLB value of 10 – 12 or mixtures of lauryl polyglycol ether (HLB = 9) with partially ethoxylated sorbitan trioleate (Span 85; Atlas Chemie). The silicones used are generally liquids; their emulsification is treated elsewhere (→ Emulsions).

5.3.3. Production of Polymer Dispersions

[56], [112], [70, p. 403]

The manufacture of polymer dispersions has been improved to such an extent that new fields of application have been opened up. Thus, for example, the increased durability of outdoor paints enabled colored paints to be used.

A distinction is made between self-dispersable polymers and those which become finely distributed by adding dispersants. Polyurethanes with ionic or nonionic water-soluble groups can be dispersed without further additives. Common polymers are finely dispersed by adding dispersants and additional surfactants, particularly wetting agents.

The manufacture of polymer dispersions is generally carried out by emulsion polymerization [112, part 1]. In this process, the sparingly soluble monomer is emulsified in water. The actual polymerization is initiated by water-soluble radical initiators and then proceeds in the aqueous phase in the polymer particles enveloped by the emulsifier. The emulsified droplets are only responsible for maintaining the supply of the monomer. The emulsion disintegrates during the polymerization, and a dispersion (suspension) of the polymer is produced. The final product is a "synthetic latex" containing particles that are 10 – 100 times smaller than the emulsified monomer droplets present at the beginning of the reaction. The size of the latex particles varies between 50 nm and 3 μm and can be controlled by the choice and quantity of dispersant and by the way in which it is added.

The advantages of emulsion polymerization over block polymerization or over the use of organic solvents in production and application are as follows:

1) Easy removal of the heat of reaction
2) Achievement of higher molecular masses, which results in better product characteristics such as film formation
3) Lower viscosity of the mixes

4) More convenient workup due to the elimination of solvent recovery and avoiding the hazards associated with the handling of organic solvents

Since the polymer latices that are formed are often used directly as coating compounds, adhesives, or for coating and impregnating paper or textiles, they must be, for example, resistant to storage, frost, and electrolytes. Their film-forming properties must not be adversely affected by the addition of pigments, plasticizers, or thickening agents. The adhesion of the dispersant molecules to the polymer particles contributes to the stability of the dispersion, which depends on the dispersant's degree of hydration and on the polymer type [113].

Cationic products are seldom used, since they may interfere with the common initiator systems or precipitate during subsequent use when anionic surfactants are added.

Although condensation products are also used, they are less widespread than anionic, nonionic, and especially hydrophilic surfactants because of their dark color. Most commonly used are anionic alkyl or arylalkylsulfonates, alkylsulfates or salts of natural fatty acids, and resin acids. These produce a relatively large decrease in surface tension and in the solid – liquid interface tension and an increase in the electrical charge of the negatively charged monomer droplets. Both factors result in finely divided, stable polymer dispersions even when small amounts are added [112, vol. I, p. 64]. Nonionic surfactants such as alkyl or alkylphenyl polyglycol ethers are preferred because their critical micelle concentration is markedly lower than that of anionic surfactants. If mixtures of nonionic and anionic surfactants are used, the low critical micelle concentration can be suitably combined with the negative electrical charges. Therefore, the use of such mixtures is common practice in the plastics industry.

Salts of natural fatty acids or resin acids are preferred for the *emulsion polymerization of butadiene* [*106-99-0*] or for the copolymerization of butadiene with styrene or acrylonitrile, but anionic sulfonates or sulfates are also used.

Abietic acid derivatives, which have become known as Dresinat (Hercules), are commonly used in the manufacture of *polychloroprene* [*9010-98-4*]. The use of alkylphenol polyglycol ethers has also been described [114].

If natural *rubber* is being processed, a protein – fatty acid condensate is added to abietic acid as a codispersant [94].

In the polycondensation of certain *polyurethanes,* small quantities of condensed dispersants or readily water-soluble chlorides are added to abietic acid derivatives.

The polymerization of *cis*-1,4-isoprene in solution or the manufacture of *butyl rubber* [*9010-85-9*] has been carried out in the presence of potassium soaps, alkylbenzene sulfonates, cetylpyridinium chloride, or ethoxylated alcohols or alkylphenols [113].

For the *polymerization of styrene* [*100-42-5*], alkylbenzene sulfonates, sulfated alcohols, often together with nonionic products such as polyglycol ethers and polyglycol esters, are used.

In the emulsion polymerization of *vinyl monomers* such as vinyl chloride [*75-01-4*] or vinyl acetate [*108-05-4*], soaps, alkylsulfonates, and sulfosuccinic acid esters are chiefly employed. Vinyl acetate can also be polymerized in the presence of protein – fatty acid condensates [94]. Stabilization with poly(vinyl alcohol) is common in the dispersion of poly(vinyl acetate).

Sulfated alcohols, condensed amines, and recently perfluoro compounds are used for the emulsion polymerization of *ethylene* [*74-85-1*].

For the emulsion polymerization of *ABS polymers* [*9003-56-9*] (copolymer of acrylonitrile, butadiene, and styrene), hydrolyzed, disproportionated colophonium (Dresinat; Hercules), and sodium oleate and sodium alkylbenzenesulfonate are suitable [115], [116].

Water-soluble monomers such as acrylonitrile [*107-13-1*] are polymerized in water-in-oil emulsions. Soaps, sorbitan fatty acid esters, and poly(vinyl pyridinium) salts are used as emulsifiers. The mechanism of this polymerization differs from the described mechanism in that no new particles are formed during the polymerization [117].

To carry out the polycondensation of polyamides or polyesters in emulsion, a mixture of condensed sulfates with suitable polyglycol ethers may be used. Finally, the polycondensation of phenol – formaldehyde resins in emulsion has been attempted, using hydrolyzed proteins and zinc soaps.

5.3.4. Application of Polymer Dispersions

The processing of polymer dispersions involves the addition of further dispersants or dispersion aids, depending on the dispersion's field of application [9, p. 389]. The main fields of application of such polymer dispersions are paints and lacquers, building chemicals, auxiliary materials for the manufacture of paper and paper coatings, textile coating, and colorants for plastics.

Paints and Lacquers. Painting materials contain essentially two components, pigments and binders. A distinction is made between aqueous and nonaqueous systems.

Nonaqueous Systems. In nonaqueous systems, the natural or synthetic resins or oils are dissolved in an organic solvent such as benzene, xylene, or a suitable ester. Pigments are generally finely distributed in these liquids by using dispersing machines (grinding units) or by adding dispersants, antisettling agents, or leveling agents. The dispersants that are used are substances such as disproportionated abietic acid, sparingly soluble naphthenates, or sulfated fats and oils. Common antisettling agents are, for example, thickeners such as fatty acid ethanolamides. Despite their high viscosity, lacquers have excellent brushability. Finally, silicon oils or oligomeric acrylates are used as leveling agents. Lacquers which provide shiny coatings are generally nonaqueous systems.

Aqueous Systems. In aqueous systems, the natural resins or oils must first be emulsified. As a rule, polymer dispersions produced by emulsion polymerization or emulsion polycondensation must first be stabilized before pigments can be added. Liquid – liquid or solid – liquid systems are encountered, but systems exist in which solid and liquid constitutents are dispersed alongside each other. The dispersants used are polyelectrolytes such as ligninsulfonate or condensed sulfonates. Alkylbenzene sulfonates and disproportionated abietic acid are also used. In addition, nonionic polyglycol ethers of alcohols, alkylphenols, condensed phenols, abietic acid, fatty acids, and fatty acid esters are added. To avoid the disturbing effects of polar additives on polymer dispersions with polar structures, nonionic surfactants are often preferred.

The dispersions of natural resins or oils or the synthetic latices are then supplemented with pigments, again by mechanical dispersion. Pigment dispersants, so-called milling agents, are added; condensed sulfonates such as brands of Vultamol (BASF) are suitable [118]. Salts of polycarboxylic and polyacrylic acids or polyphosphoric acid with a medium chain length, cellulose derivatives, or alkylphenyl polyglycol ether, as well as copolymers of ethylene and propylene oxide, are also used. In the final emulsion paints, the binder and pigment particles are stably dispersed alongside each other in the aqueous phase.

The dispersants in such paints, apart from having stabilizing properties, are intended to improve the adhesion of the polymer and pigment particles to the foundation and to promote film formation. The color intensity and the spreading power of the pigments are generally improved by interface-active additives [119, p. 192].

When polymer dispersions are used on plaster, concrete, masonry, cement, or wood for painting facades, walls, or ceilings, or as roofing felt paints, the dispersants added must primarily have a beneficial effect on the rheological properties of the polymer films produced and keep them stable for a long period. In addition, they must contribute to the shelf life and the electrolyte stability. Since the action of each individual dispersant is specific to certain polymer dispersions, there are no general rules governing their use [112, vol. II, p. 48]. Addition of the emulsifier used in emulsion polymerization is recommended, provided that its nature is known or can be determined. In addition to natural or synthetic dispersants, auxiliary agents such as protective colloids or thickeners are often used.

Special mention should be made of polymer dispersions for metal painting such as corrosion protection painting, roof painting, and to an increasing extent, for priming vehicle bodies. In these paints, the choice of suitable dispersants is critical for the formation of smooth films with good adhesion on such nonabsorbent substrates [112, Vol. II, p. 65]. Aqueous polymer dispersions have not acquired any importance as baking enamels thus far.

Building Chemicals. Polymer dispersions are widely used in building materials. The dispersions, which are chiefly manufactured by

emulsion polymerization, are used in plasters and coatings [112, vol. II, p. 32] and as additives in cement mortars and filling compounds [112, vol. II, p. 80], usually with emulsifiers or dispersants also added.

To enable the use of polymer dispersions as plaster in the building industry, as little water as possible should be present to achieve a high flow resistance, which is beneficial in handling the material, and to avoid the risk of shrinkage cracks forming. Dry fillers and pigments are usually worked directly into the polymer dispersion. To prevent coagulation, dispersants such as ligninsulfonate or condensed aromatic sulfonates are added beforehand.

If polymer dispersions are used as filling compounds primarily for producing jointless built-up floor coverings, filler compatibility is an important application factor. Interface properties such as wettability are, therefore, especially important and can be controlled by adding surfactants. If interface-active copolymers are used, the choice of the hydrophilic – hydrophobic polymer ratio is decisive.

Addition of surfactants is essential to produce jointless floor coverings, for example, with an anhydrite base by using the Anhyplan process (BAVG, Bayer Anhydrit Vertriebsgesellschaft), and for the mechanical application of ceiling or wall plasters or facade stuccos. Rapid mechanical application of the material, for example, by means of a pump, must be possible. However, after the plaster has adhered, it must set relatively rapidly. Alkanesulfonates, alkylbenzene sulfonates, and polyglycol ethers of alcohols or alkylphenols are used for this purpose.

Paper Auxiliaries. When pulp or paper coating pulp are manufactured, polyphosphates, complex phosphates, or polyacrylate salts are suitable for the dispersal of white pigments such as china clay or titanium dioxide in water. On the other hand, in colored pigment dispersions, so-called fine pastes, which are used to stain high-quality paper such as laminate paper, alkylnaphthalene sulfonates are generally used. Additions of alkylsulfonates have proved to be successful when plastic dispersions based, for example, on butadiene – styrene or acrylate polymers are used as binders for pigmented paper plasters.

Additions of alkylsulfonates are suitable if polymer plastic dispersions based, for example, on poly(vinylidene chloride), polyurethane, or silicon oil are used to achieve effects on the paper surface such as water repulsion, greaseproofness, or separating action towards adhesive substances.

Textile Coating. Dispersions are used as flaking or coating agents or as binding agents in the textile industry. Fields of application are, for example, the manufacture of artificial leather or carpet backing [120]. Virtually all the polymers are used, especially poly(vinyl acetate), polyurethanes, or polyvinyl – polyalkylene copolymers.

Coloration of Plastics. For coloring plastics, the pigments are added to the polymer dispersions. The dispersants used in the polymer dispersions must be compatible with the pigment dispersants.

5.3.5. Adhesive Dispersions

Dispersions of numerous polymers are used as emulsion-based adhesives. Their oil-in-water emulsions (dispersions) or mixtures of finely divided substances and liquids are also used in various industrial applications.

For bonding wooden furniture parts together, poly(vinyl acetate) or polyamide dispersions are generally used. The packaging industry prefers dispersions of poly(vinyl acetate) or polyacrylate or their copolymers and polychloroprene or polyurethane emulsions as paper adhesives. In the building trade, dispersions of poly-(vinyl acetate), polyacrylate, or polychloroprene and dispersed copolymers of styrene and acrylate are used, e.g., for bonding tiles. The leather industry makes use of emulsion/dispersion-based adhesives that contain polyurethanes, polychloroprene, or vinyl copolymers. Dispersions of polyvinyl – polyalkylene copolymers are required by the textile industry for laminating.

The dispersants used generally lower the interface tension between paper, wood, or other materials and the adhesive used. Since organic solvents themselves have a low surface tension, surfactants are normally only added to aqueous glues to reduce the high surface tension of water.

The governing rule for rubber adhesives is that anionic surfactants are used for anionic

latices, and cationic surfactants are used for cationic latices. To prevent errors, nonionic surfactants are generally used. Alkylsulfonates, alkyl ether sulfates, or ethylene oxide – propylene oxide adducts have been suggested for vinyl – resin adhesives [121].

5.3.6. Detergents

Polyphosphates are used in the detergent industry, but they are being increasingly replaced by other additives. In Switzerland, for example, the use of detergents was restricted to those free of phosphates from 1986. Substitutes are, for example, zeolites and nitrilotriacetic acid [120]. The addition of optical brighteners is also rejected because they are difficult to degrade in lakes, rivers, and other bodies of water.

5.3.7. Photographic Industry

The coating solutions and dispersions manufactured in the photographic industry contain microcrystals of silver halides dispersed in gelatin (Fig. 14).

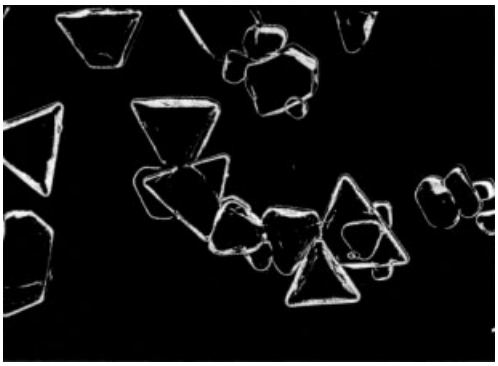

Figure 14. Silver halide microcrystals in a photographic "emulsion" [122]

The wetting agents, antistatic agents, and lubricants added must not alter the original properties of the silver halide microcrystals, but on the contrary, they should result in uniform and faultless film formation on various backings such as paper or photographic film. Therefore, careful choice and precise measurement of the quantity are indispensable.

The oldest and most well-known photographic *wetting agent* is the natural product saponin, which despite variations in the quality of the starting material (quillaia bark) has not yetbeen completely displaced because of its good wetting properties and photographic neutrality. Synthetic anionic surfactants contain sulfate, sulfonate, or carboxylic groups as hydrophilic groups (e.g., fatty acid sulfates, alkylarylsulfonates, sulfosuccinic acid derivatives). On the other hand, nonionic wetting agents contain alkyl or alkylaryl groups (often with several alkyl chains containing a total of 8 – 18 carbon atoms) as hydrophobic groups and ester – ether – amide or polyethylene – glycol groups as the hydrophilic part (e.g., polyhydric alcohols, glycerol, sorbitol, and sugar). Perfluorinated surfactants such as Bayer FT 248 have proved to be successful for use with modern multilayer pourers and backing speeds of ≥ 100 m/min.

Antistatics are used in photographic manufacturing to prevent the disadvantages associated with electrical charging, such as inadvertent exposure or attraction of dust. Readily hydratable ionic surfactants such as sulfated alcohols or sulfosuccinic acid esters are used, as are betaines, amine oxides, or short-chain alkyl polyglycol ethers.

Lubricants ensure satisfactory film running in cameras and projectors. They consist of dispersions, for example, of stearic acid, wax, polyethylene, or silicon oil, which are added as oil-in-water emulsions to the uppermost layer of the material. Short-chain alkyl polyglycol ethers are used. Since a number of antistatic agents and lubricants cause faults in the uniformity of the pouring, anionic wetting agents must be added.

In addition, color components that couple to the dyes during photographic development must also be dispersed because they are generally insoluble in gelatin. Alkylsulfonates, condensed sulfonates, sulfated fatty acid amides, or alkylnaphthalene sulfonates are used. Additions of saponin and nonionic surfactants have been described [123].

5.3.8. Agriculture

Dispersants are used to prevent caking in the manufacture of fertilizers. They also ensure uniform distribution of mineral fertilizers in the soil. For these purposes, solid fertilizers are

sprayed with dilute solutions of ethoxylated alcohols, alkylbenzenesulfonates, or occasionally solutions of cationic surfactants.

In plant protection and in pest control, dispersable powders and concentrated suspensions are used [70, p. 394; 119, p. 182]. *Wettable powders* contain the active ingredient, fillers such as kaolin or calcite, dispersants such as ligninsulfonate, sulfonates of condensed aromatic compounds or polyglycol ethers of condensed phenols, and wetting agents such as alkylnaphthalene sulfonate. The powders are manufactured by dry grinding the thoroughly mixed constituents to a particle size of 5 – 40 µm. Grinding additives such as kieselgur or synthetic silica prevent sticking and clotting. Concentrated *suspensions* usually contain 30 – 50 % solids. Mixtures of dispersants and emulsifiers have proved to be suitable for such suspensions, which must remain stable for over 2 years under a wide variety of storage conditions. The suspensions are generally manufactured by wet grinding to a particle size of < 5 µm.

5.3.9. Cosmetic and Pharmaceutical Industry

To manufacture ointment bases of the water-in-oil type, sorbitan fatty acid esters are used, whereas the preparation of an ointment base of the oil-in-water type that can be washed off involves ethoxylated stearic acid or ethoxylated sorbitan fatty acid esters, often mixed with nonethoxylated esters.

Stearic acid and its sodium salt can be dispersed by using mixtures of sorbitan fatty acid esters and their ethoxylated products, as well as with ethoxylated alcohols or fatty acids [110].

Partial fatty acid esters or their ethoxylated products are common in the dispersion of fats. Polyglycol ethers of fatty alcohols and fatty acids and of lanolin or propylene glycol are also used. In addition, mixtures of anionic and nonionogenic emulsifiers are used, as are mixtures of purely nonionic interface-active substances with various HLB values. Lanolin is usually emulsified with fatty alcohol polyglycol ethers. Sulfated esters such as "turkey-red oil" are added to dental care products and, in addition to their dispersive action, are intended to prevent tartar formation.

5.3.10. Cleaning and Polishing Agents

The dispersal of polishing and scouring powders is usually promoted by adding polycyclic aromatic sulfonates; Tamol N brands are suitable for this purpose [93].

5.4. Testing

Since dispersants are used in various industries, the fact that there are no generally applicable test methods is not surprising. The published test methods are usually directed at a special field of application. The following methods exist.

Zinc Oxide Test. The zinc oxide test gives an initial indication of whether inorganic materials are dispersed; however, this test is either not applicable or applicable only with restriction to dye dispersions.

After the dispersant to be tested (3 g) has been shaken with 10 g of zinc oxide in a measuring cylinder containing 200 mL of water, the settling of the zinc oxide is observed and assessed. The settling can be recorded photometrically.

Flocculation Preventing Power [124]. The determination of the flocculation preventing power is based on precipitation of an insoluble compound by a chemical reaction and subsequent determination of the concentration of a dispersant that just prevents the flocculation of the insoluble compound. The standard DIN 53 903 specifies the determination of the ability to disperse calcium soap. The result is given as a percentage of pure sodium oleate, and the flocculation preventing power is inversely proportional to the quantity of dispersant necessary to disperse lime soap [125].

The standard DIN 53 908 specifies the determination of dispersive power with respect to dyes. This involves the precipitation of an insoluble dye from a solution of a leuco ester of a vat dye in the presence of a defined quantity of the surfactants to be tested. Conclusions are drawn about the dispersive action of the material from the quantity of dispersant necessary to achieve a stable dispersion. The stability of the dispersion is assessed from the spreading pattern that results when a defined quantity of the dye dispersion is applied to a dry filter paper [58].

Filtration Methods. Filtration methods also yield information on the stability of dispersions in dyeing processes.

Thus, for example, 250 mg of a commercial dye (C.I. Disperse Red 54) is dispersed in 500 mL of water by adding the dispersant and subsequently treated for 30 min at 125 °C in a laboratory dyeing apparatus. The dispersions are then cooled to 80 °C, filtered by suction through a filter paper (e.g., Schleicher & Schüll No. 1450 CV) and placed on a vacuum filter of 90 mm diameter [69, p. 517]. As less dye remains on the filter, the dispersive power is better.

Spot Test. The filter paper spot test, the so-called *flow test,* can be used as a rapid test for the finenese of a pigment dispersion and, consequently, for the effectiveness of a dispersant. A small amount of the pigment dispersion is dropped onto a piece of filter paper, and the pattern that is produced as the drop spreads is assessed. If the dispersion is fine, the drop spreads uniformly, and as the edge is sharper, the finenese of the dispersion is poorer [69, p. 517].

Practical Test Methods. Standard dyeings in laboratory dyeing apparatuses can be used to determine the effect of dispersants on levelness and rubbing fastness in the printing and thermosol dyeing of fabrics [107]. In addition to these practical test methods, physical methods are being used increasingly to determine the particle size and distribution.

Physical Test Methods. The preparation of electron micrographs [126–128] that can be counted yields the particle size distribution of a dispersion [53], [93], [94], as does sedimentation analysis under gravity and in a centrifugal field [107], [118], [127], [129]. Gravity covers particles of up to 1 µm diameter and the particle size of the disk centrifuge ranges between 0.1 and 10 µm [130]. Counts made with the Coulter counter [130] or optical methods, such as light scattering measurements or laser correlation spectroscopy, are suitable for determining the average particle size or the particle size distribution.

6. References

1. *Vokabularium von Tensiden (Tenside vocabulary),* 2nd ed., C.I.D.-Commission de Terminologie, Paris 1972.
2. DIN 53 900, July 1972 issue.
3. B. Koglin, *Chem. Ing. Tech.* **46** (1974) 720.
4. J. Stauff: *Kolloidchemie (Colloid Chemistry)* , Springer Verlag, Berlin 1960.
5. R. Menold, *Phys. Bl.* **33** (1978) 203.
6. R. E. Wachtel, V. K. La Mer, *J. Colloid Sci.* **17** (1962) 531.
7. E. G. Richardson, *J. Colloid Sci.* **8** (1953) 367.
8. I. W. Goodwin: *Colloidal Dispersions,* Special Publication no. 43 of the Royal Society of Chemistry, London 1983, p. 165.
9. R. Heusch in: *Tensidtaschenbuch* (Tenside Handbook), 2nd ed., Hanser Verlag, München–Wien 1981; p. 183.
10. R. Heusch, *Chem. Ing. Tech.* **55** (1983) 608.
11. R. Zsigmondy: *Kolloidchemie (Colloid chemistry),* **vol. II,** Verlag Otto Spamer, Leipzig 1927.
12. K. Kuchta, *Chem. Ind. (Düsseldorf)* **28** (1976) 285, 110.
13. W. Wiedmann, H. Blenke, *Chem. Anlagen + Verfahren* 1976, no. 4, 82 – 90, 110.
14. H. Ries, *Aufbereit. Tech.* **20** (1979) no. 1, 1 – 25, no. 2, 78 – 98.
15. B. Koglin, *Maschinenmarkt/MM-Industriejournal* **86** (1980) 346.
16. A. Anbarci, *Int. Z. Lebensm. Technol. Verf. Tech.* **37** (1986) 336.
17. O. K. Kubelik, *Erdöl Kohle Erdgas Petrochem.* **23** (1970) 381.
18. B. Koglin, J. Pawlowski, H. Schnöring, *Chem. Ing. Tech.* **53** (1981) 641.
19. *Winnacker-Küchler,* 3rd ed., **7,** 6.
20. H. Sonntag, K. Strenge: *Koagulation und Stabilität disperser Systeme (Coagulation and stability of disperse systems)*, VEB Deutscher Verlag der Wissenschaften, Berlin 1970.
21. E. J. W. Verwey, J. Th. G. Ovberbeek: *Theory of Stability of Lyophobic Colloids,* Elsevier, Amsterdam 1948.
22. D. H. Napper: *Polymeric Stabilisation of Colloidal Dispersions,* Academic Press, London 1983.
23. P. C. Hiemenz: *Principles of Colloid and Surface Chemistry,* Marcel Dekker, New York–Basel 1977.
24. I. N. Israelachvili, D. Tabor, *Prog. Surf. Membr. Sci.* **7** (1973) 1.
25. C. H. Hamaker, *Physica (Amsterdam)* **4** (1937) 1058.

26. M. J. Vold, *J. Colloid Sci.* **16** (1961) 12.
27. D. J. M. Osmond, B. Vincent, F. A. Weix, *J. Colloid Interface Sci.* **42** (1973) 262.
28. B. Vincent, *J. Colloid Interface Sci.* **42** (1973) 270.
29. R. Heusch, *Tenside Deterg.* **21** (1984) 173, 298; **20** (1983) 1.
30. R. Heusch, *Ber. Bunsen Ges. Phys. Chem.* **88** (1984) 1093; **83** (1979) 834; **82** (1978) 970.
31. G. Boehmke, R. Heusch, *Fette Seifen Anstrichm.* **62** (1960) 87.
32. I. Vissar, *Adv. Colloid Interface Sci.* **3** (1972) 331.
33. P. Debye, E. Hückel, *Phys. Z.* **24** (1923) 185.
34. H. de Hek, A. Vrij, *J. Colloid Interface Sci.* **79** (1981) 289.
35. B. Vincent, *Adv. Colloid Interface Sci.* **4** (1974) 193.
36. R. H. Ottewill in: J. W. Goodwin (ed.) *Colloidal Dispersions,* Special Publication no. 43 of the Royal Society of Chemistry, London 1983, p. 143.
37. J. Klein, *Makromol. Chem. Makromol. Symp.* **1** (1986) 125.
38. E. L. Mackor, *J. Colloid Sci.* **6** (1951) 492.
39. P. Bagchi, R. D. Vold, *J. Colloid Interface Sci.* **33** (1970) 405.
40. J. B. Smitham, R. Evans, D. H. Napper, *J. Chem. Soc. Faraday Trans. 1* **71** (1975) 285.
41. R. H. Ottewill, *Annu. Rep. Prog. Chem. Sect. A: Gen. Phys. Inorg. Chem.* **66** (1969) 212.
42. G. C. March, D. H. Napper, *J. Colloid Interface Sci.* **61** (1977) 383.
43. G. J. Fleer, J. Lyklema, *J. Colloid Interface Sci.* **46** (1974) 1.
44. A. T. Clark, M. Lal in : Th. F. Tadros (ed.): *The Effect of Polymers on Dispersion Properties,* Academic Press, London 1982.
45. *G. J. Fleer, J. Lyklema, J. Colloid Interface Sci.* **55** (1976) 228.
46. R. E. Feigin, D. H. Napper, *J. Colloid Interface Sci.* **75** (1980) 525.
47. W. Härtl, H. Versmold, U. Wittig, *Ber. Bunsen Ges. Phys. Chem.* **88** (1984) 1063.
48. R. H. Ottewill, *Ber. Bunsen Ges. Phys. Chem.* **88** (1984) 1063.
49. B. V. Derjaguin, L. D. Landau, *Acta Physicochim. URSS* **14** (1941) 633.
50. J. Th. G. Overbeek in H. R. Kruyt (ed.): *Colloid Science,* vol. I, Elsevier, Amsterdam 1952.
51. J. Fleer: "Adsorption aus Lösungen" (Adsorption from solutions), *Seminar of the Colloid Society,* Ruhr University, Bochum 1980, p. 43.
52. J. Lyklema, *Adv. Colloid Interface Sci.* **2** (1968) 65.
53. DIN 53 900.
54. *Vokabularium von Tensiden,* 2nd ed., Comité Intern. Des Derivés Tensio-Actifs (C.I.D.) Commission De Terminologie, Paris 1972.
55. R. Heusch, *Prog. Colloid Polym. Sci.* **69** (1984) 29.
56. G. D. Parfitt: *Dispersions of Powders in Liquids,* Applied Science Publ., London 1973.
57. S. Heimann, *Melliand Textilber.* **12** (1982) 885.
58. DIN 53 908, August 1982.
59. F. Jost, H. Andree, M. J. Schwuger, *Colloid Polym. Sci.* **264** (1986) 56.
60. A. Hettche, W. Trieselt, P. Diessel, *Tenside Deterg.* **23** (1986) 12.
61. K. V. Sarkanen, C. H. Ludwig: *Lignins,* J. Wiley, New York 1971.
62. Marathon Corp., US 2 576 418, 1951, US 2 505 304, 1947. Dryden Chem., US 293 898, 1972.
63. Zellstoff Waldhoff, DE-OS 1 769 903, 1968.
64. "*Ultrazine*" and "*Ufoxane*", Company Publications of Messers Borregaards, Postbok 162, N-1701 Sarpsborg, Norway 1978.
65. G. Prazak, *Am. Dyest. Rep.* **59** (1970) 44.
66. I.G. Farben, DE 752 830, 1943; Westvaco, US 3 769 272, 1973; US 3 672 817, 1970.
67. Westvaco, US 3 672 817, 1970; US 3 769 277, (1972); US 3 841 887, 1972; US 3 865 503, 1972; US 3 956 261, 1975.
68. S. Y. Lyn, *Text. Chem. Color.* **13** (1981) 261.
69. A. Chwala, V. Anger in: *Handbuch der Textilhilfsmittel,* Verlag Chemie 1977, pp. 502 ff, 523.
70. R. Heusch in: *Tensid Taschenbuch,* 2nd ed., Hanser Verlag, München 1981.
71. FIAT-Report 1013 (1946).
72. Kao Soap, JA 977 273, 1975.
73. Ciba-Geigy, BE 806 793, 1972.
74. Hoechst, DE-OS 2 442 514, 1974.
75. A. M. Schwarz, J. W. Perry, J. Berch in: *Surface Active Agents and Detergents,* vol. I – II., Intersci. Publ., New York 1949 and 1958.
76. Casella, DE-OS 1 165 753, 1971.
77. Bayer, DE 2 542 090, 1975.
78. Westvaco, US 236 786, 1973.
79. Ciba-Geigy, DE-AS 2 657 453, 1975.
80. K. Linder: *Tenside–Textilhilfsmittel–Waschrohstoffe,* **vol. I – III,** Wissenschaftliche Verlagsanstalt mbH, Stuttgart 1964 and 1971, respectively.

81. Ciba-Geiby, GB 1 101 671, 1964; DE-OS 2 353 691, 1972. ICI, DE-OS 1 902 110, 1969.
82. Hoechst, DE-AS 1 163 284, 1960.
83. BASF, DE-AS 1 241 985, 1962; DE-AS 1 246 752, 1962; DE-AS 1 247 329, 1964.
84. BASF, DE-OS 1 960 812, 1969.
85. Hoechst, DE-OS 2 032 926, 1970.
86. Ugine Kuhlmann, FR 2 122 710, 1971.
87. Hoechst, NL 7 302 331, 1972.
88. Westvaco, DE-OS 2 314 326, 1972; US 3 870 681, 1973.
89. BASF, DE-OS 2 301 638, 1973.
90. BASF, DE-AS 2 401 818, 1974.
91. Mitsui Toatsu Chem., DE-OS 2 439 878, 1974.
92. BASF, DE-AS 2 535 447, 1975.
93. BASF, Ti/P 2507 d, 2684 d, 2685 d (technical informations).
94. Chemische Fabrik Grünau GmbH, Illertissen, F. R. of Germany, company publication.
95. ICI, GB 717 840, 1951. BASF, DE 1 180 343, 1962.
96. I. Wolf, M. Bauer, *Melliand Textilber.* **49** (1968) 464. F. Wolf, U. Koch, W. Homeister, *Melliand Textilber.* **52** (1971) 456.
97. H. Lange, *Kolloid Z. Z. Polym.* **234** (1971) 101.
98. M. J. Rosen in: *Surfactants and Interfacial Phenomena,* J. Wiley & Sons, New York 1978.
99. Kemisk, Vaerk, Køge, NL-A 6 611 672, 1964.
100. R. Heusch, *Fette Seifen Anstrichm.* **72** (1972) 969.
101. R. Heusch, *Chem. Phys. Chem. Anwendungstech. Grenzflächenakt. Stoffe Ber. Int. Kongr. 6th 1972* **1** (1973) 476.
102. H. Nassenstein, *Phys. Bl.* **32** (1976) 249.
103. R. Heusch, *Chem. Ing. Tech.* **55** (1983) 608.
104. E. Valko, *Rev. Prog. Color Relat. Top.* **3** (1972) 50.
105. W. D. Cooper, *Kolloid Z. Z Poly.* **250** (1972) 38.
106. DIN 53 917,Jan. 1981.
107. H. Leube, H. Uhrig, *Textilveredlung* **9** (1974) 97.
108. US (Defensive Publication) T 912 002, 1. May 1972.
109. W. Dietsche, *Fette Seifen Anstrichm.* **75** (1973) 415.
110. Das Atlas HLB-System, company publication of Atlas Chem. GmbH, Goldschmidtstraße 100, D-4300 Essen 1, Federal Republic of Germany.
111. E. Hantz et al., *J. Phys. Chem.* **89** (1985) 5832.
112. F. Hölscher, H. Reinhard: *Dispersionen synthetischer Hochpolymerer,* Springer Verlag, Berlin–Heidelberg–New York 1969.
113. V. T. Samorodov, L. V. Kosmodem'jankij, E. A. Lazuhrin, *Kolloidn. Zh.* **40** (1978) 1206.
114. H. Schuller, *Chem. Ing. Tech.* **45** (1973) 1123.
115. R. M. Panitsch, N. N. Iwanova, S. S. Wojuzkij, *Kolloid Z. Z. Polym.* **232** (1969) 669.
116. Chem. Systems Inc., 747 third Avenue, New York N.Y. 10017, Report Nr. 74-7 (1975); p. 89.
117. J. W. Vanderhoff, E. B. Bradfort, H. L. Terkowski, J. B. Schäffer, R. M. Wiley, *Adv. Chem. Ser.* **34** (1962) 32.
118. E. Afherton, C. A. Cooper, *Teintex* **30** (1965) 453.
119. J. Falbe, M. Hesserodt: *Katalysatoren, Tenside Mineralöladditive,* Thieme Verlag, Stuttgart 1978.
120. F. F. Hartmann, *Chemiefasern/Textilind.* **24** (1974) 207.
121. H. Hadert, *Seifen Öle Fette Wachse* **93** (1967) 139.
122. *Bayer-Ztg. "direkt",* 1986, no. 4.
123. M. Oguchi, Y. Kuwabara, *Nippon Shashin Gakkaishi* **26** (1963) no. 2, 73.
124. R. Wirth, *Dtsch., Textiltech.* **11** (1961) 89.
125. DIN 53 902,Oct. 1955.
126. R. Kath, D. Mildebrand, *Z. Gesamte Textilind.* **89** (1967) 329, 405, 473.
127. J. Beresford, *J. Oil Colour Chem. Assoc.* **50** (1967) 594.
128. W. Biedemann, *J. Soc. Dyers Colour.* **87** (1971) 105; **88** (1972) 329.
129. M. R. Fox, *Teintex* **30** (1965) 455.
130. A. F. Bogenschutz, U. George, J. Jentsch, *Fachber. Oberflächentech.* **10** (1972) 1.

Display Technology

DIETMAR THEIS, Siemens AG, Corporate Technology, München, Federal Republic of Germany

1. Introduction 2
2. Display Categories 2
3. Cathode Ray Tubes (CRTs) 4
3.1. Color CRTs 5
3.2. Electron Optics 6
3.3. Phosphors 6
3.4. Trends 7
4. Flat Panels and Matrix Addressing 8
4.1. Light-Generating (Active) Displays 8
4.1.1. Gas Discharge 9
4.1.1.1. Alternating-Current Plasma Displays 9
4.1.1.2. Plasma Display Addressing 10
4.1.1.3. Remaining Problems for Plasma Displays 12
4.1.2. Cathodoluminescence 12
4.1.2.1. Vacuum Fluorescence Displays (VFDs) 12
4.1.2.2. Field-Emission Displays (FEDs) . . 13
4.1.3. Electroluminescence 14
4.1.3.1. Visible Light Emitting Diodes (LEDs) 14
4.1.3.2. Organic and Polymer Light Emitting Diodes 16
4.1.3.3. Electroluminescence in Inorganic Materials 18
4.2. Light Modulating (Passive) Displays 19
4.2.1. Liquid Crystal Displays 19
4.2.1.1. Fundamental Operation 20
4.2.1.2. Simple Matrix-Addressed LCDs . . 20
4.2.1.3. Novel Addressing Schemes for Direct-Addressed STN LCDs 22
4.2.2. Reflective LCDs for Low-Power Systems 22
4.2.2.1. Reflective LCDs with Polarizers . . 22
4.2.2.2. Polarizer-Free Reflective LCDs . . . 22
4.2.3. Bistable Liquid Crystal Devices . . 23
4.2.4. Active Matrix-Addressed Liquid Crystal Displays (AMLCDs) 25
4.2.4.1. Fundamental Properties 25
4.2.4.2. Manufacturing Issues 25
4.2.4.3. Viewing Angle Improvements . . . 26
4.2.4.4. System Performance 28
4.2.4.5. Plasma-Addressed Liquid Crystal Displays 28
4.2.5. Microparticle-Based Displays 29
4.2.5.1. Electrophoretic Displays 29
4.2.5.2. Rotating Ball Displays (Gyricon) . 30
5. Projection Displays 30
5.1. Projection CRT 31
5.2. Liquid Crystal Projection 32
5.2.1. Transmissive Projection LCDs . . . 32
5.2.2. Reflective Projection LCDs 32
5.3. Micromechanical Light Valve Projection 33
6. References 33

1. Introduction

The "information age" and its enabling technologies will greatly influence our future daily life in the next decades. The electronic provision of information is the backbone of the new age. The human eye is the most versatile sensing organ, easily being able to handle huge data streams either in sequence (e.g., written text) or in parallel (e.g., pictures). The transformation of electronically generated information into specially configured visible signals is carried out by a display device.

An electronic display device consists of several components (Fig. 1). First, there are the electro-optic display elements, which modify their optical appearance by the action of an electrical signal. The electrical information is usually processed by some logic and amplified before it activates the display elements. Sometimes these elements must be backed up by external power sources. In many cases, ancillary elements are added to make detection of the displayed information easier.

The display device is the visual interface between user and machine. Two requirements are placed on this interface: (1) optimum correspondence of the signals generated from the electro-optic display element to the physiological sensory requirements of the human eye and (2) their

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a08_603

adaptation to the requirements of the subsequent neural signal processing for easy interpretation and minimum reading error. Accordingly, display development is of an interdisciplinary nature, including physiology, engineering, optics, materials science, and computer science. The emphasis in this article is on technology, i.e., the engineering and materials science aspects of electronic display devices.

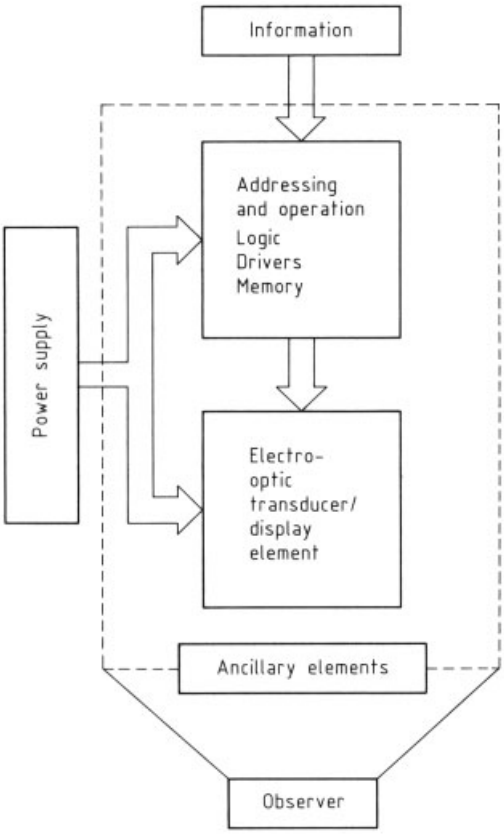

Figure 1. Structure of an electronic display device. The area enclosed by the dashed lines is the display unit.

2. Display Categories

A multitude of very different displays is already available on the market. For structuring this complexity at least three criteria have been put forward to categorize display devices.

The first is a classification according to the *information content* of the display device, which implicitly also largely determines the application area of the display. The information content is related to the number of picture elements (pixels) available. A technically meaningful upper limit for the number of pixels per inch (ppi) or centimeter (ppc) is given by the minimum angle of resolution of the eye. The average naked eye is not able to resolve pixels less than about one minute of arc apart. For a typical computer-screen viewing distance of about 50 cm this means that the meaningful lower limit for the pixel pitch is about 0.15 mm or 170 ppi (67 ppc). (Note that for present automatic bonding processes for connecting external circuitry to the display rows and columns, 130 ppi is the upper limit for production scale.) Figure 2 plots the total number of pixels versus panel size, and includes the areas of different applications, from cellular phone displays via handheld computers and portable TV, e-books, web pads, notebooks and desktop monitors, workstations, large TVs, to large-area professional applications with more than 30 diagonal inches. Pixel formats of panels are often referred to in terms of standardized sets such as VGA (640×480), XGA (1024×768), and UXGA (1280×1600). Note that for colored displays "pixel" always means a entity of the three (red, green, and blue) subpixels; hence, the number of individual subpixels is three times that of simple black and white dislays. Once the pixel format is given, the panel size is determined by the resolution, i.e., the pixel pitch or the number of pixels per inch (or centimeter). Figure 2 includes plots for three high to ultrahigh resolutions.

The second classification is based on the *method of addressing*. In low-information-density displays, e.g., numerical seven-segment displays, each picture element can be addressed individually. In displays with medium-to-high information contents, two addressing techniques allow many picture elements to be addressed in a short time: (1) scanned displays and (2) matrix displays. In the first, the viewing screen is scanned by an electron beam or a light beam that is modulated to produce the desired pattern. The most prominent scanned display is the cathode ray tube. In the second, the display screen is made up of a two-dimensional matrix of pixels, and the pattern is obtained by activating each individual element electrically. In most cases, activation is carried out line by line: all pixels in one line are addressed and activated by the

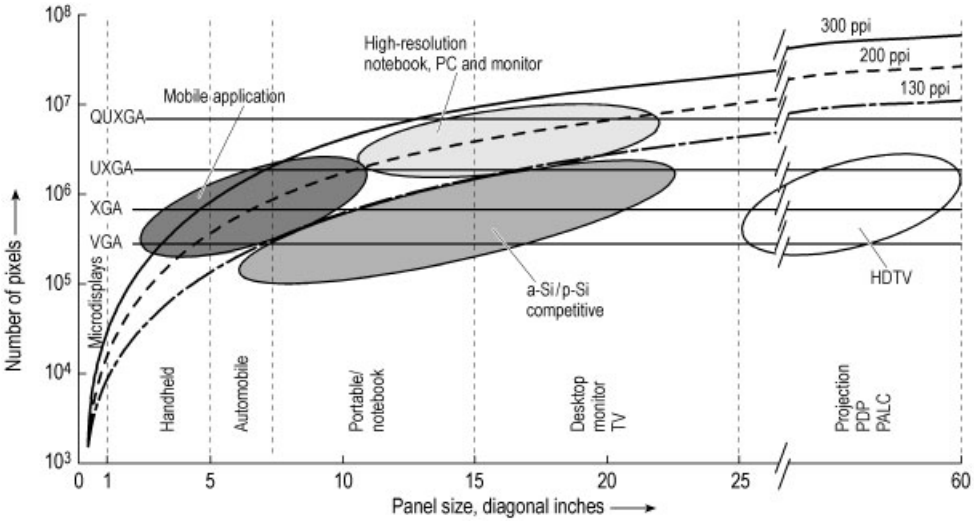

Figure 2. Different application fields require matched panel sizes. For a given standard pixel format, the size of the panel is determined by the resolution required.

signal on their respective intersecting column. One line after another is activated to build up the image within the frame time. Liquid crystal displays and the vast majority of all other flat panel displays are matrix displays.

The third classification is based on the *method and technology of presenting the information*: (1) off-screen, (2) projection, and (3) direct-view display devices (Fig. 3). *Off-screen* devices include those in which the image does not appear on a screen, but rather is perceived to exist in space as a real or a virtual image. Examples are holographic displays, "head-up" displays, and displays which generate the image directly on the retina of the eye. *Projection displays* include all those that use a viewing screen separate from the optical image source. Currently, projection displays are used primarily when a group of people or a larger audience must be informed or entertained. In *direct-view displays*, the image is generated at, or very close to, the viewing surface of the device. Direct-view devices are the most common display devices.

A prominent direct-view display device is the *cathode ray tube* (CRT), which was the first electronic device to produce images. A great variety of CRTs was invented in the last 100 years, but the basic mechanism of operation is the same in all devices. The quality of images produced by CRTs has been accepted widely as a standard. However, CRTs are large, deep, heavy, and require high voltage.

The desire for portable electronic display devices with the same visual performance as the CRT but without its drawbacks has fuelled a multibillion-dollar industry in the last two to three decades. In 2000 roughly half of the worldwide display sales of about $\$ 52 \times 10^9$ were for flat panel displays, of which liquid crystal displays (LCDs) accounted for close to 90 %. Of growing importance are a number of other new technologies which also provide *flat panel displays*. The technological development in this field is still very dynamic. For 2005 a total revenue of roughly $\$ 100 \times 10^9$ is expected for all types of display modules[11]. Since the compound annual growth rate for flat panel displays is expected to be about four times larger than the growth rate for CRTs, the market share of flat panel products in 2005 is expected to be ca. 63 % of the total market.

Many physical and chemical effects have been explored for their utility in flat panel information presentation. Figure 3 classifies them as emissive or nonemissive technologies. The strong dominance of LCD technologies is expected to persevere until at least 2010, although emissive technologies such as plasma and organic light emitting diodes (LEDs) are expected to gain important market shares.

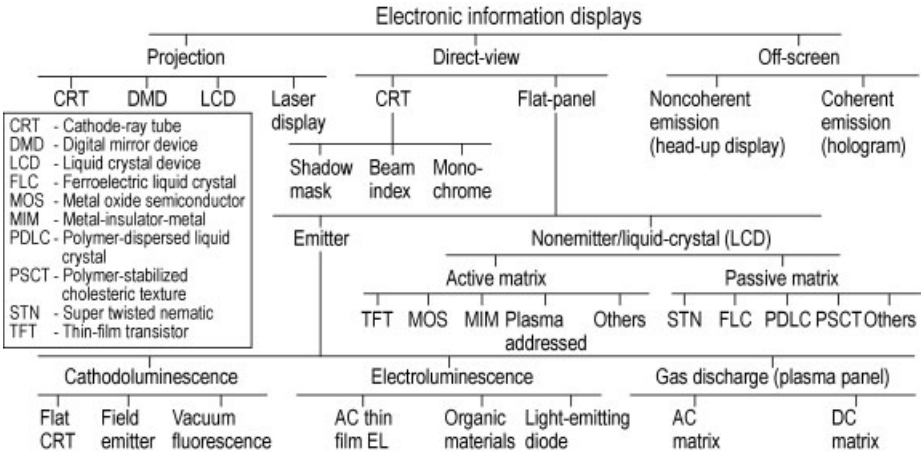

Figure 3. Categorization of electronic information displays according to the way of viewing the display and the required enabling technologies

3. Cathode Ray Tubes (CRTs)

Cathode ray tubes (CRTs) are the oldest display devices. In 1896, Ferdinand Braun invented a device that already contained all the elementary features of modern CRTs. A monochrome CRT basically consists of an evacuated bulb, an electron gun, a beam-focusing and -deflection system, and a phosphor screen laid on the front face of a sealed bulb. The pressure inside the bulb is usually less than 10^{-5} Pa. In most cases, the bulb is made of glass. The electron gun provides a high-density electron beam, which can be focused and deflected to address any desired point on the phosphor screen. The faceplate carries a thin layer of phosphor particles which emit light under electron bombardment. Usually this layer is covered by a thin layer of aluminum to establish the electrical potential at the screen and to reduce charging effects. The faceplate may be treated specially to improve display contrast, e.g., with an antireflective coating.

3.1. Color CRTs

Color is incorporated into the CRT by providing separate phosphor dots or stripes to generate the three primary colors – red, green, and blue – on the screen (Fig. 4). Proper addressing of these dots is achieved with the aid of a metal mask containing aperture holes or slots. The thickness of the masks varies between 25 and 250 μm, and they are made from low-carbon steel or, in high-end applications, from iron – nickel alloy. Depending on the mass fabrication technology the mask material is either drawn with a pressing tool beyond the yield point to obtain plastic deformation for acquiring a prescribed curved contour, or the mask is welded to a frame that results in tension in the mask material. The latter technology requires strong and therefore heavy frames [13]. The shadow mask in its frame is positioned close to the screen. Three electron beams emanating from electron guns positioned close to the longitudinal axis of the tube in delta or "in-line" configuration. These beams are individually modulated for the three primary colors.

If one gun is regarded as a point source of light, a shadow of the aperture mask falls on the inside of the faceplate. At the same time illuminated dots of the specific color appear on the screen. Arranging the three guns in a symmetrical triangle around the longitudinal axis produces a set of three illuminated dots on the faceplate, one for each primary color. The positions of the three phosphor dots on the screen must fit the shadow mask. This distribution can actually be fabricated in a self-aligned technique that uses the shadow mask during the photolithographic steps for phosphor deposition.

Shadow masks were introduced into large-scale production in 1954 (see, e.g., [15]). Substantial improvements followed and resulted in a strong increase in CRT screen brightness.

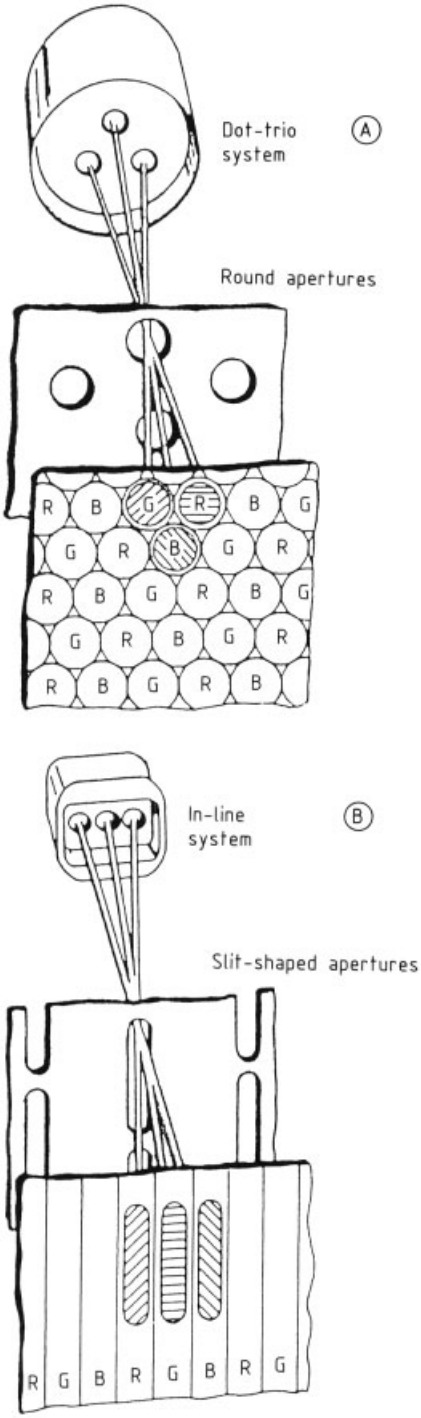

Figure 4. Triangular (delta) dot system (A) compared with the in-line system (B), showing the beam paths and the screen structures (after [14])
R = red; G = green; B = blue
Courtesy of the Society for Information Display

Figure 4 B shows a shadow-mask CRT with an in-line configuration of the three electron guns, slot apertures in the shadow mask, and continuous vertical phosphor lines on the screen. This technology allows for larger deflection angles, improved beam convergence, and better resolution, and has been further refined by using strip masks.

To obtain a perfect picture, severe requirements must be imposed on the tube design. Focal and color purity must be maintained, as well as mechanical tolerances, which must be kept low. When the tube is operated, ca. 75 – 80 % of the electron beam is intercepted by the dotted shadow mask. This substantially reduces the brightness and causes heating. In a typical receiver, ca. 35 W is absorbed by the mask and converted to heat. The resulting expansion of the mask must be counteracted by sophisticated bimetal support elements that keep the expanding mask aligned to the electron beam.

3.2. Electron Optics

Figure 5 shows three important processes taking place in the neck of the CRT: (1) beam formation, (2) focus, and (3) deflection.

Electrons escape from the heated cathode surface and are accelerated by the positive potential on grid 2. Application of an appropriate negative potential to grid 1 controls the electron beam intensity. Cathode design is important because CRT lifetime is governed mainly by *cathode life*.

Two cathode configurations are depicted. Figure 5 A shows the conventional oxide cathode, while Figure 5 B shows the dispenser cathode. The *oxide cathode* consists of a barium – strontium oxide layer 10 – 20 µm thick applied to the capped end of a nickel cylinder. The heavy metal oxide reduces the work function of the surface and increases electron emission. In practice, a 10 000-h cathode lifetime is achieved. In the *dispenser cathode*, the barium compound diffuses through a porous tungsten structure to the surface and ensures that the barium content at the surface is replenished, allowing for a much longer lifetime.

As shown in Figure 5, the electron beam can be focused electrostatically or magnetically.

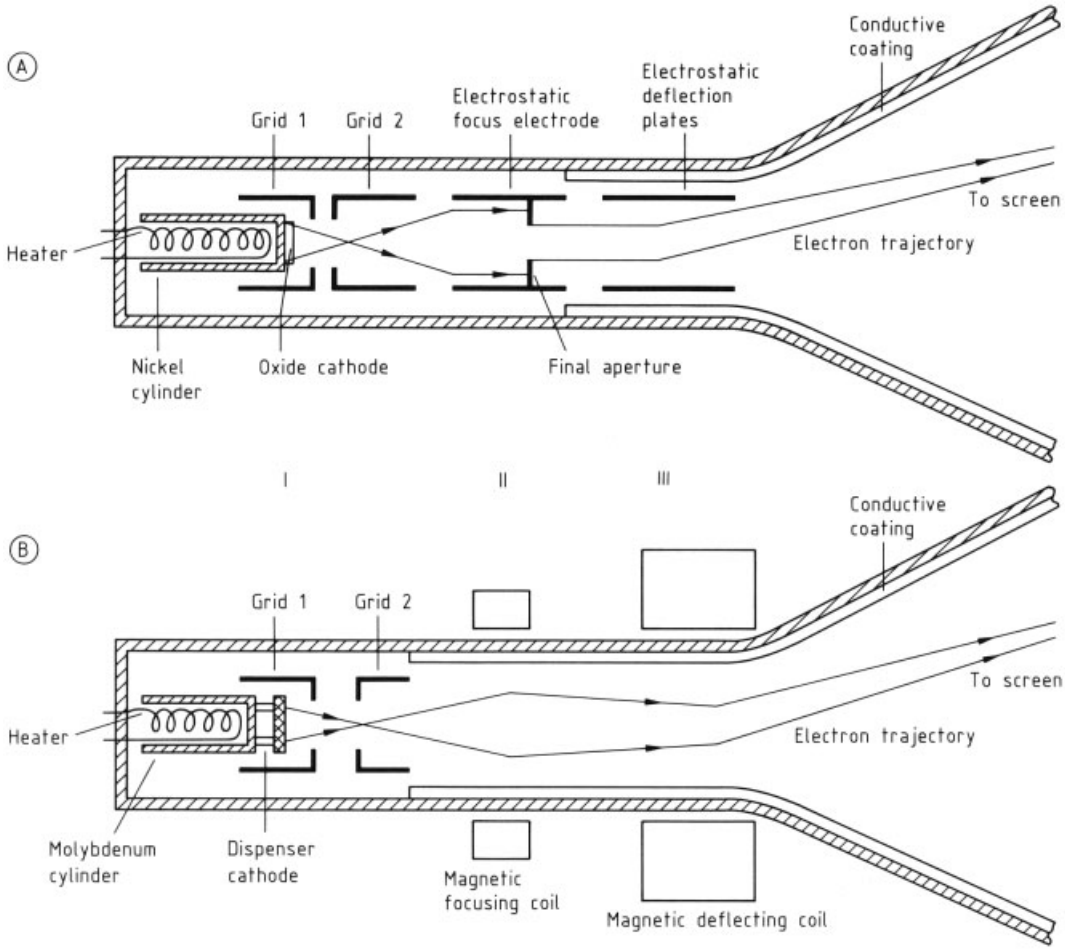

Figure 5. Schematic presentation of the beam forming (I), focusing (II), and deflection (III) regions in the CRT neck
A) Normal oxide cathode with electrostatic focusing and deflection; B) Dispenser cathode with magnetic focusing and deflection systems

Magnetic focusing provides the highest resolution, because magnetic lenses have fewer aberrations than electrostatic lenses. Likewise, deflection can be electrical or magnetic. Electrostatic deflection requires two sets of metal plates (for x and y deflection) set at right angles to each other. Magnetic deflection employs two pairs of coils, the coils in each pair being located on opposite sides of the neck of the tube to provide orthogonal magnetic fields. Electrostatic deflection provides the highest deflection speed but only moderate resolution. Magnetic deflection provides lower deflection speed, but gives satisfactory resolution, especially when combined with electrostatic focus. This combination of magnetic deflection and electrostatic focusing is used widely in television and computer terminal tubes. Magnetically deflected and focused tubes provide the highest resolution.

3.3. Phosphors

The production of visual images from the electronic signals occurs at the phosphor-covered faceplate (→ Luminescent Materials). Table 1 lists the three RGB phosphors mainly used in CRTs with their composition, peak wavelength, decay time, and luminous efficiency.

The *luminous efficiency* of many phosphors (ratio of luminous flux to incident energy) in-

Table 1. Selected phosphors for CRTs

Composition	Peak wavelength, nm	Decay time to 10 % level, ms	Luminous efficiency, lm/W	CIE color coordinates *x,y*
ZnS : Ag	440	25	35 – 60	0.15, 0.06
ZnS – CdS : Ag	530	60	40 – 65	0.32, 0.60
Y_2O_2S:Eu	627	900	30 – 50	0.66, 0.35

creases with the beam energy, tending to saturate only between 15 and 30 kV. For maximum efficiency, CRTs are operated at ca. 20 kV. Luminous efficiency is also affected by current density, usually decaying strongly above 30 mA/cm^2. Under normal operating conditions, luminous efficiencies of phosphors range between 20 and 70 lm/W. The power required by the filament reduces these figures by ca. 30 %. The luminous efficiency of the system takes into account all power necessary to drive the display: on average, 15 lm/W for monochrome CRTs and 6 lm/W for color television.

The luminance of all phosphors decreases under continued electron bombardment. The *decay of efficiency* with the number of coulombs deposited is different for each phosphor. After cathode aging, phosphor aging is the most important factor affecting the CRT lifetime.

3.4. Trends

CRT technology is very mature, and manufacturing costs for commodity parts are relatively low. Typically, materials costs constitute about 65 % of the overall cost in CRT production. The price reduction for color CRTs has been comparatively moderate in the last decades: 30 % for TV tubes and 50 % for color monitor tubes between 1990 and 2000 [17].

In 1997 real-flat TV tubes were first introduced on the market. In theses CRTs the outside surface of the faceplate is practically flat, whereas the inner surface is curved to accommodate the curved mask. Some effort has still to be made to overcome transmission losses associated with the fact that in these faceplates the glass at the corners is much thicker than in the center of the faceplate [18], [19].

4. Flat Panels and Matrix Addressing

The term *flat panel display* essentially covers all the technologies in Figure 3 for non-CRT direct-view display. Flatness in this context usually means that the ratio of device depth to screen diagonal length is less than 1:10, including driving circuitry boards.

Nearly all flat panel designs rely on a matrix arrangement of the picture elements. This has important consequences common to all types of flat panels. Normally, each picture element has one electrode connected to all other picture elements in a row, and the second electrode is connected to all other picture elements in the orthogonal column. When a picture element is selected, its row and column leads are energized to generate a potential across the desired picture element. Analysis of such a network with Kirchhoff's voltage and current laws shows that the minimum potential across all the Off picture elements is one-third of that across the On picture elements because there are three sets of parallel branches between the *x* and *y* terminals of the selected On picture element [20].

The 1/3 voltage is the minimum inherent and unavoidable cross-coupling in matrix addressing. This cross-coupling potential partially excites the display medium. If the luminance – voltage relationship for the picture element is linear, then an image background is seen, impairing the contrast. Therefore, strongly nonlinear luminance – voltage dependence is mandatory for matrix addressing.

Another problem arises from addressing two picture elements of different columns and rows simultaneously. Two additional picture elements light up, and together with the addressed picture elements they formcorners of the rectangle given by the intersection of the two rows and columns.

The addressing of unwanted picture elements is avoided by a *line-at-a-time addressing* technique, in which only one line lead is addressed at any instant. All column leads are addressed

in parallel. With this scheme the Off picture element is excited by 1/3 voltage cross-coupling each time a line is addressed. The On picture element is excited to full On only when its line is addressed and is excited to 1/3 voltage cross-coupling every time the other rows are addressed.

In line-at-a-time addressing, the time available to activate each line is the frame time divided by the number of display lines. The frame time is the total time available to address the entire display array, usually ca. 16.6 ms (60 Hz) for flicker-free images. In refreshed displays, the activation of the picture element and the persistence of the electro-optic effect make up the actual time that a display effect is visible. If a very large number of lines is to be scanned during frame time, the entire performance is degraded because insufficient time is left for proper activation.

In display configurations in which the display material exhibits inherent memory, these problems do not occur. The images formed by the picture elements can be sustained without refreshing until another image is desired. Memory can also be introduced artificially by a proper circuit for each picture element (extrinsic or active matrix addressing; see Section 4.2.4).

4.1. Light-Generating (Active) Displays

Flat panel light-generating displays make information visible by emission of light as a result of picture element activation. Important is the *light output conversion efficiency*, i.e., the ratio of photon output power, $P_{\mathrm{out}} = h\nu \cdot p/t$, to electrical input power, $P_{\mathrm{in}} = e\,V \cdot n/t$ ($h\nu$ is the energy per photon, p/t is the number of photons generated per unit time t, n/t is the number of inflowing electrons per unit time, and eV is the average electron energy before photon emission).

The ratio

$$\eta = P_{\mathrm{out}}/P_{\mathrm{in}} = (h\nu/eV)\cdot p/n = \eta_{\mathrm{E}} \cdot \eta_{\mathrm{n}}$$

can be factored into the photon – electron energy ratio η_{E} and the photon – electron conversion efficiency ratio η_{n}. The former is close to unity in most light-emitting displays. Even in high electron energy devices like the CRT, down conversion of electron energy occurs in the phosphor through creation of secondary electrons with energies comparable to the photon energy before the onset of light emission. However, η_{n} is one or two orders of magnitude below unity, which is why active displays always require more power than ambient-light-modulating displays.

4.1.1. Gas Discharge

Gas discharge is the fundamental phenomenon for all types of plasma displays, and in this sense these displays are similar in operation to the familiar neon lamp. The displays consist of two plates of glass enclosing a gas-filled volume. The gas mixture usually consists of chemically unreactive noble gases like neon plus 3 – 8 % xenon with traces of argon or helium at a pressure of about 7 kPa. In matrix displays both glass plates carry electrode strips oriented orthogonally to each other. When the electrodes are activated a discharge is created and the xenon atoms – excited by energetic electrons – begin to radiate ultraviolet light which can be transformed into visible light by appropriate photoluminescent phosphors.

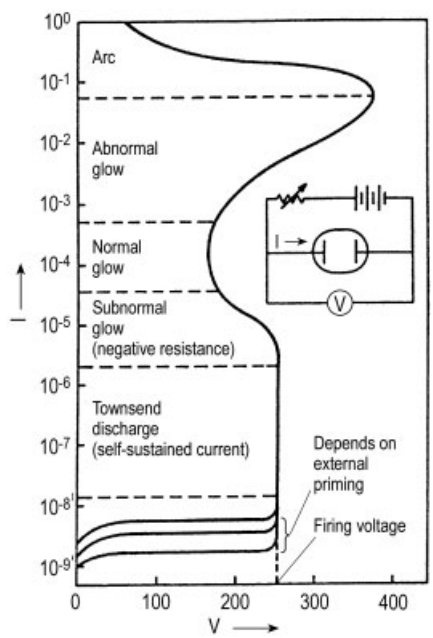

Figure 6. Schematic current – voltage ($I - V$) characteristic of a gas discharge

Figure 6 schematically shows the logarithmic current – voltage ($I - V$) characteristic of a typical gas discharge. In such a reaction numerous different ionization and excitation processes occur. The most important one is the ionization of neutral atoms by electrons, resulting in a positively charged ion and an additional free electron. The ions flow to the cathode and induce secondary electron emission there, which in turn leads to new avalanches and the growth of the discharge current. Depending on the voltage across the gas the feedback loop of the process described can be greater than unity, and without any external resistance the current would grow unrestricted. This strong nonlinearity is one of the major advantages of a gas discharge, allowing practically unlimited multiplexing without any additional nonlinear devices.

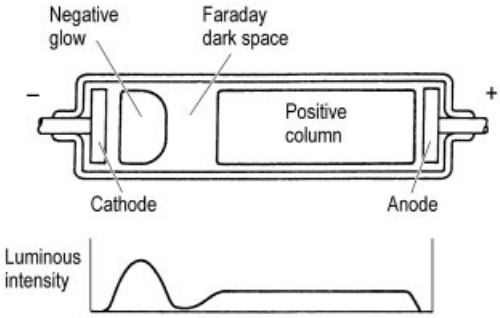

Figure 7. Characteristic luminous regions of a gas discharge

The principal luminous regions of a gas discharge are shown in Figure 7. The positive column discharge is used in fluorescent lamps for room lighting and in neon signs for commercial advertising. In plasma displays, however, the reduction of the electrode separation shortens the available portion of the positive column, while the negative glow length is maintained. In commercially available a.c.-addressed plasma displays the negative glow dominates in light generation over the positive column. Unfortunately, the negative glow is a much less efficient light generator than the positive column. Designs that make better use of the positive column would help to increase the luminous efficiency [21].

4.1.1.1. Alternating-Current Plasma Displays

Today the technically and commercially most successful plasma displays are color a.c. plasma panels. Their fundamental structure is shown in Figure 8 A together with a cross-sectional view in Figure 8 B. The devices consist of glass back and front plates.

The back plate carries the silver x-axis address electrodes; a 20 µm-thick layer of dielectric glass; screen-printed, sandblasted, or photocast barrier ribs of about 150 µm in height between the address electrodes, which suppress optical cross-talk between neighboring discharge cells; and phosphors that emit red, green, or blue light on UV excitation. In color plasma displays the xenon component of the gas mixture emits UV light in the 147 nm range. In a line-at-a-time scanned matrix display the overall brightness decreases with increasing number of scanned lines, and therefore the inherent memory function is extremely valuable. RGB light emission is achieved by the photoluminescence of phosphors excited by the ultraviolet light. Typical phosphors for plasma displays are $(Y,Gd)BO_3:Tb^{3+}$ for red, $ZnSiO_4:Mn^{2+}$ for green, and $BaMgAl_{10}O_{17}:Eu^{2+}$ for blue emission.

The front glass plate is equipped with two transparent parallel electrodes, the sustain and scan electrodes, which are covered by a thick-film dielectric The inner surface of this dielectric layer, which is in contact with the noble gas, is given a special 0.5 µm-thick coating of MgO. This coating ensures longer display life by preventing ions from bombarding the dielectrics. It also has a low threshold for secondary electron emission, thereby allowing for lower operating voltages.

The insulating dielectric layers have two basic functions. First, they limit the current (and thus prevent arcing) by coupling the electrodes capacitively to the gas discharge, and second they help to achieve memory function by forming a capacitor in series with each pixel.

The establishment of a discharge during one cycle induces charge on the capacitance formed by the dielectrics with a field direction opposite to the polarity of the applied potential. These wall charges add to the applied voltage when the voltage polarity is reversed in the next cy-

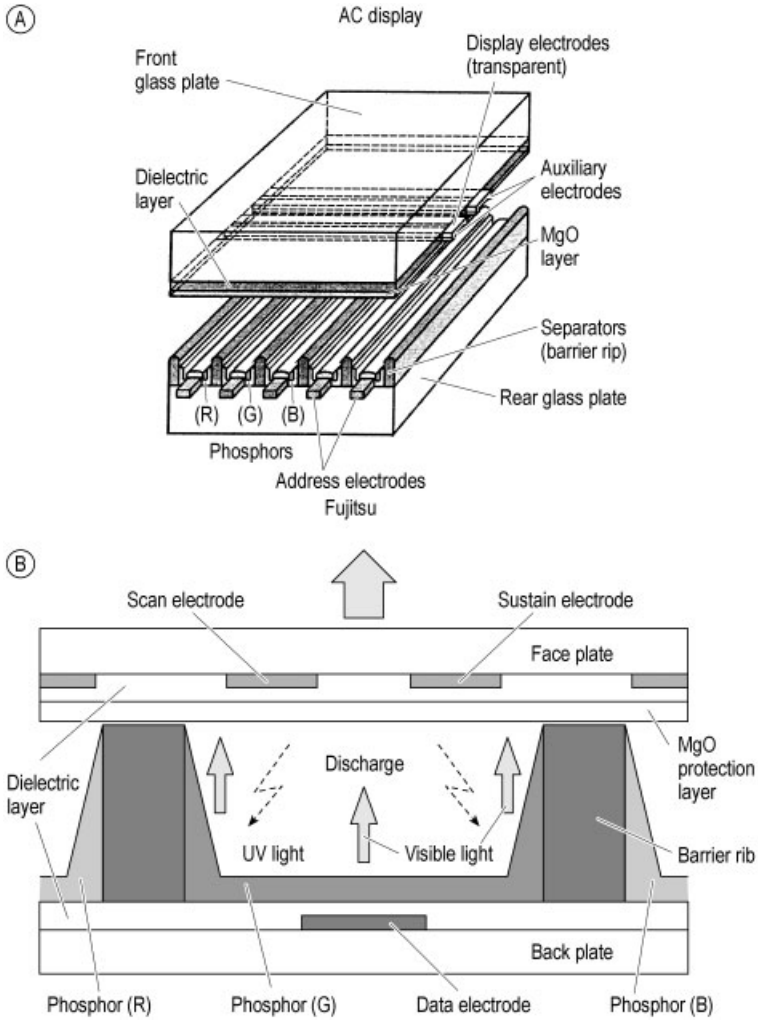

Figure 8. A) Fundamental structure of an a.c.-addressed plasma display (Fujitsu); B) Cross section of a (green) subpixel discharge cell in such displays

cle and thus allow breakdown at a lower applied potential. The discharge current and the resulting light occur in short pulses. At typical operating frequencies of 50 or 60 kHz, the cumulative effect of many pulses results in typical luminance values between 300 and 600 cd/m^2. The stored charge permits display memory. The entire matrix array is supplied with a repetitive pulse voltage, the *sustain voltage*, below that required to start the discharge. An additional pulse across a selected picture element in the matrix starts the discharge at this picture element, and the discharge sequence is sustained by repeated pulses with the aid of the charge stored in the series capacitance. The local discharge is extinguished when the stored charge is reduced by one smaller *erase pulse*, which makes the field across the capacitance too low for firing.

Figure 9 Shows the voltage waveforms for addressing plasma display pixels as described above.

4.1.1.2. Plasma Display Addressing

The fundamental method for addressing practically all presently available a.c. plasma displays is called address display separation (ADS). With

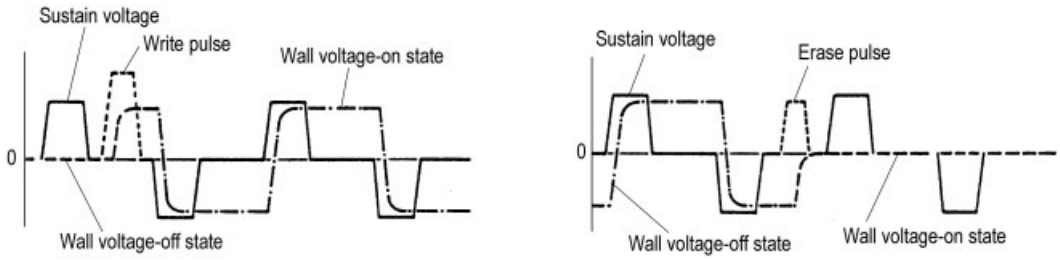

Figure 9. Voltage waveforms for addressing plasma display pixels

this method an a.c. sustain voltage is applied to all the cells of a line. To write the cells an additional pulse is applied to all cells during the address period. Those cells which have to stay Off receive a corresponding column pulse that cancels the effect of the write pulse. The obtained distribution is maintained for the required sustain period by the sustain signal until a special erasing signal quenches the discharge for all pixels. This signal is often called "setup" or "reset". This signal has two important purposes. First, it causes some minimal level of discharge activity that primes the next gas discharge. Priming is important in plasma displays since it provides a controlled low level of ionization necessary to achieve reliable starting conditions for the discharge. The other important function of the setup pulse is to place the Off-state wall voltage to a precise and reproducible level that will allow the "write" operation to occur reliably [21]. After erasure, all cells of the line are in Off state waiting for a new writing process.

To achieve gray scale with binary plasma displays, the setup – address – sustain sequence is used in a multiple-subfield arrangement, as shown in Figure 10. In this example the image frame time is divided into eight subframes, each weighted with a binary power multiple of the smallest time. Each of them has a setup and address period of equal length. In each subframe the entire screen image is addressed, selectively and sequentially one line at a time from top to bottom. This writes an image into the plasma display's inherent memory. Then a sustaining cycle is executed which produces light from those pixels which were written in the On state. The eight bits of gray-scale code for each pixel determines which subframes must be turned on for that pixel during the image frame time. This scheme works well for still images, but has to be modified for motion pictures due to visible artefacts.

4.1.1.3. Remaining Problems for Plasma Displays

There are essentially three major issues still to be solved. One is *phosphor degradation*. The materials used are basically those developed by the flourescent lamp industry. Of these, the blue-emitting material has the shortest life, while the red material is very stable. Most manufacturers quote a lifetime of plasma displays to half luminance to be between 10 000 and 30 000 h. The mechanism of phosphor degradation is not well understood, but a doubling of the lifetime is desirable.

The current *manufacturing cost* of plasma panels excludes its use as a mass consumer product. A large proportion of the manufacturing cost is determined by material cost [22], with an important contribution from electronic drivers. However costs can be cut by using cheaper glass, by further improvement in the driver area by using silicon-on-insulator technologies, and advanced functional integration, which may lead to the production of a complete 42-inch RGB module for about $ 1000.

Probably the greatest challenge for plasma display engineers is to *increase the efficiency* of the devices. Most plasma displays today operate at a luminous efficiency of about 1 lm/W. There is considerable effort in academia and industry to find ways to improve the efficiency. Flourescent lamps show efficiencies around 80 lm/W, and the target for plasma displays has been set at 5 lm/W for the near future. To achieve this goal improvements in the discharge efficiency,

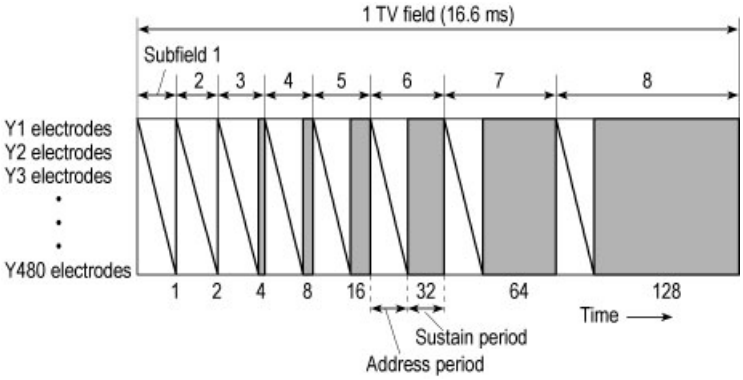

Figure 10. Address display separation (ADS) technique for gray scale

better optical coupling, and higher phosphor efficiencies have to be attained.

4.1.2. Cathodoluminescence

In the last fifty years several suggestions have been made for making *flat CRTs* with ingenious deflection plates, multiple electron beams, decoupling of electron beam selection and high-voltage acceleration or the provision of large-area screen-sized cathodes, whereby the electrons could be extracted through the aperture holes of a matrix-addressed grid separating cathode and screen. None of these many attempts made great inroads in computing or entertainment applications. The only cathodolominescent flat panel devices with significant sales are vacuum flourescence displays. They are briefly discussed below together with field-emission display technology, which some companies believe to be the ultimate solution for the flat CRT problem.

4.1.2.1. Vacuum Fluorescence Displays (VFDs)

Vacuum fluorescence displays (VFDs) based on low-voltage cathodoluminescence were developed in Japan in the 1960s and are still being manufactured, mainly in that country. The technology is well suited for low-information devices, like bar graphs and numerics, but is also capable of alphanumerics and low-resolution video display.

The simple numeric device consists of an oxide-coated heated emitter filament and seven anodes in a figure-of-eight configuration mounted in a flat evacuated glass tube (Fig. 11). The anodes are covered with the low-voltage phosphor ZnO:Zn, which gives a bluish green luminance when bombarded with ca. 20-eV electrons. The efficiency is 5 – 10 lm/W. The thin filament is located in front of the anodes. The anode construction is shown in Figure 11 C. The filament is kept below the temperature that leads to visible thermal emission and therefore is barely visible. By raising the potential at the selected anode segments to + 15 V relative to the cathode, any required numeral can be displayed. This is improved by introducing a grid between anodes and cathode, allowing several digits to be multiplexed. In this approach, the corresponding anode segments are interconnected and the digit position is selected by the grid potential. Only the digit that receives both anode and cathode grid signals produces light.

These displays are normally viewed from the back, i.e., through the grid. Transparent anode connectors and thin phosphor layers enable the construction of VFDs that are viewed from the front, thus allowing for a larger viewing angle.

The grid-controlled flat multidigit packages had a significant impact on the market for digital clocks, hand-held calculators, and numeric indication in audio and video equipment. Power dissipation is quite low. A nine-digit tube giving 600 cd/m^2 requires a total of ca. 40 mW.

The incorporation of colors other than blue-green requires new low-resistivity phosphors. A number of phosphor mixtures have been tried,

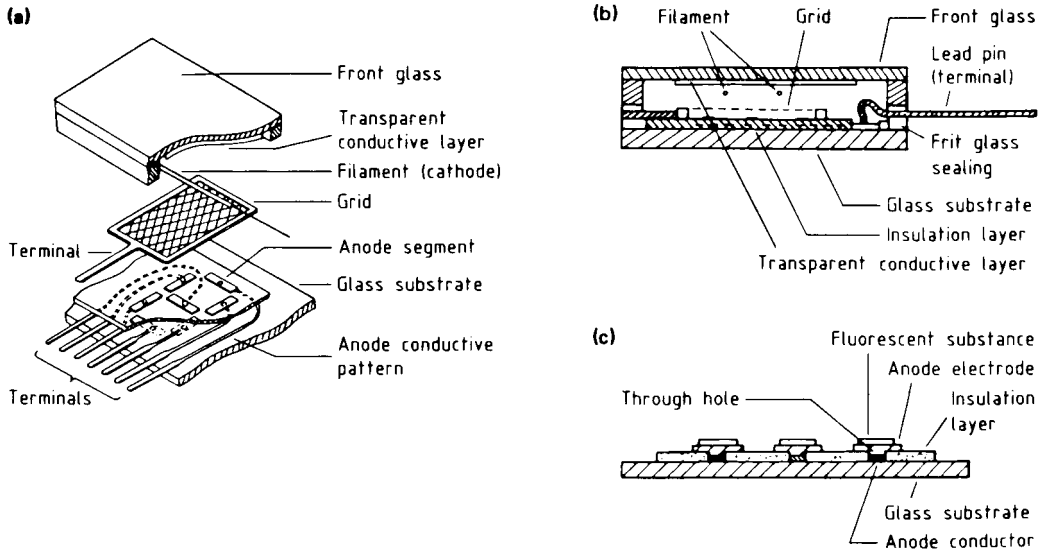

Figure 11. Vacuum fluorescence displays: (a) exploded view of a seven segment numeric display (b) cross section of the same device (c) details of the anode structure

but none matches the efficiency of ZnO:Zn. Nevertheless, VFDs with green-blue, yellow, and orange emission are available.

4.1.2.2. Field-Emission Displays (FEDs)

The operating principle of these vacuum electron devices is comparable to vacuum fluorescent displays with the important difference that the electrons are not emitted by thermal activation from hot filaments but rather by field emission from a cold cathode. Electrons escaping from the cathode by quantum-mechanical tunnelling are accelerated towards a phosphor screen, where their energy is partly transformed into light (Figure 12). Many FEDs use microscopic tips for which the field is determined by the microscopic tip geometry and the voltage difference between the cathode and closely positioned gates. The statistical probability that electrons in the metal tip will tunnel through the metal – vacuum barrier is given by the Fowler – Nordheim tunneling equation for field emission. This equation relates the emission current density with the applied electrical field, the work function of the emission surface and its geometry.

Field enhancement induced by tips is the most widely investigated method to obtain useful currents at moderate macroscopic fields. One way of fabricating microtips is to make use of the wet-etching properties of *n*-type silicon. [23] Meyer developed a fabrication process for molybdenum microtip cathodes on glass. The process also provides a gate electrode with a 1.4 μm self-aligned hole around each cathode tip. This original design was refined through the addition of a ballast resistive layer to average out random fluctuations of current yield between tips. The negative electron affinity on diamond-like carbon surfaces has fuelled research in diamond emitters as a low-cost manufacturing alternative. A more recent version of a field emitter was introduced by Canon and is based on the creation of nanometer-wide cracks in thin-film palladium oxide containing nanocrystalline particles. When a potential is applied across the small gap, electrons tunnel across, and some of them are inelastically scattered at the positive edges of the gap, from where they are attracted to the positive anode. An overview on field emitters is given in [24].

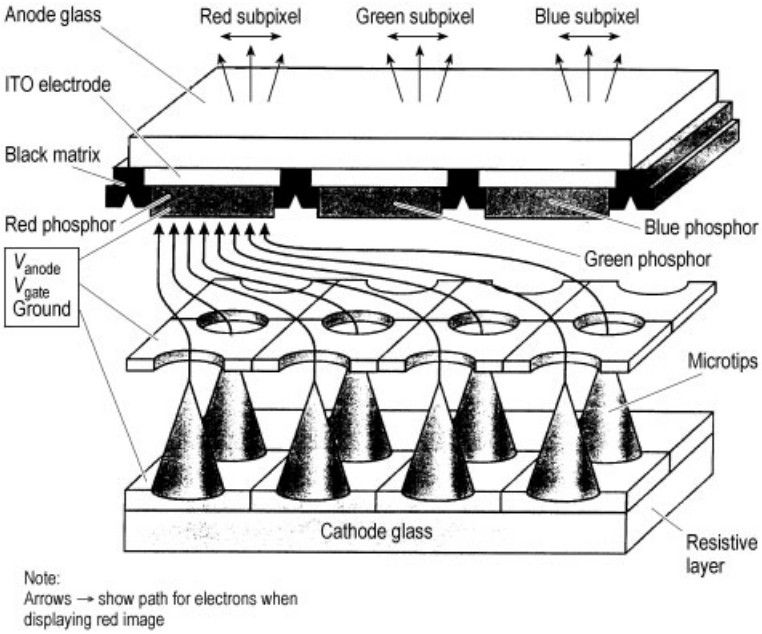

Figure 12. Basic structure of a field emission display

4.1.3. Electroluminescence

The term *luminescence* generally describes the generation of light not caused solely by a heated black body radiator. *Electroluminescence* denotes light emission phenomena related to the action of electric fields in crystalline or organic solids. Due to the fundamentally different physical mechanisms, a distinction is made between light generation due to low or high electric fields. The former case refers to the emission that occurs after the injection of minority carriers across the forward-biased *p-n* junction and the subsequent radiative recombination of electrons and holes in single-crystal semiconductor diodes or equivalent structures made of organic material. The latter is related to the emission of light in polycrystalline material exposed to high electrical fields, where generation of hot electrons and impact excitation of activators (dopants) become possible.

The term *light-emitting diode* (LED) originally was used for displays made of inorganic single-crystal semiconductors. With the recent advent of recombination radiation from organic materials, the family of light emitting diodes was enlarged to include (small molecule) organic material based OLEDs and similar devices based on organic polymers material, often called PLEDs.

The expression *electroluminescent display* (ELD, sometimes also called high-field electroluminescent display or thin film electroluminescent (TFEL) display, is generally reserved for displays made of thin film polycrystalline phosphors or of powder phosphor material in a binder.

4.1.3.1. Visible Light Emitting Diodes (LEDs)

In forward biased *p-n* junction LEDs the emission resulting from radiative recombination of electrons and holes has a wavelength corresponding to the band gap of the material used.

The efficiency of conversion of the electrical energy input to the LED into the desired light emission depends on three factors:

1) The excitation of excess minority carriers. At a *p-n* junction passing $10\,\mathrm{A/cm^2}$ minority carriers are injected at a rate of about 10^{24} carriers/cm^3 within a narrow region of $1-2\,\mu\mathrm{m}$. Assuming a lifetime of 10^{-9} s this generation yields a net surplus of 10^{15} carriers/cm^3, much above the typical equilibrium minority carrier concentration.

2) Radiative recombination of these carriers to the ground state can occur via many pathways. Some are radiative, others are not, depending on material properties such as band structure, defects, and dopant nature and concentration. Radiative transitions can occur from the conduction band to the valence band (with the emitted photon energy corresponding to the gap between the two) or via intermediate impurity levels.
3) Extraction of the generated light from the volume of the material to the surroundings. Since the materials in which the recombination radiation is generated usually have high indices of refraction much of the light cannot leave the volume of the material due to total reflection and eventual internal absorption. One way of improving the situation is to produce a micro-roughness on the semiconductor surfaces.

Visible LEDs have been on the market since the 1960s. However, only in the last few years have high-performance LEDs become available in all colors, allowing the fabrication of full-color flat panel displays viewable in direct sunlight. Before 1985, when LED performance was generally in the range of less than 2 lm/W, LEDs were mostly limited to indicator-type applications. Since 1985 the perfomance has increased to as high as 50 lm/W for the highest performance type of red-emitting AlInGaP LEDs. With the development of GaInN devices, the range of colors has expanded to include violet, blue, and pure green, so that high-performance LEDs are available throughout the color spectrum. With an efficiency that exceeds many types of conventional lighting sources, LEDs can now be used in a variety of new application areas such as vehicle lighting, traffic signals, road signs, and large-area displays for advertisment. Red-emitting high-performance devices based on AlGaAs technology were first introduced to the market around 1983. In 1991 red, orange, yellow, and green LEDS based on AlInGaP appeared [25], [26]. In 1993 this was followed by the succesful commercialization of a new materials system for the green and blue part of the visible spectrum: the InGaN system [27].

The external quantum efficiency versus emission wavelength for LEDs is shown in Figure 13

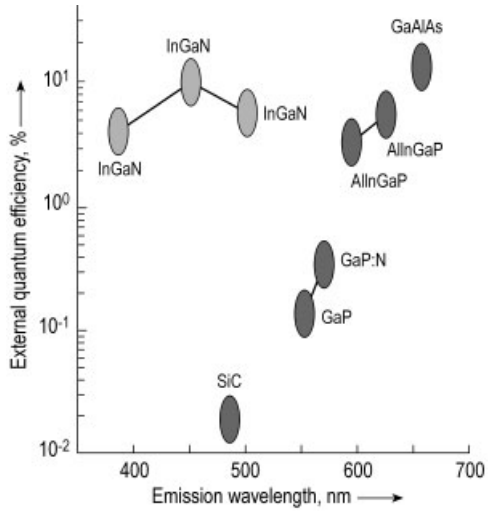

Figure 13. External quantum efficiency of commercial light emitting diodes plotted versus their peak emission wavelength

The AlInGaP system shows the highest performance in the red and amber region, but the efficiency degrades sharply towards the green region. Nevertheless, green AlInGaP LEDs can be produced with an efficiency still doubling that of the conventional GaP:N green devices. The wavelength, however, being 570 nm or greater is perceived as a yellowish green and therefore is not suitable for many full-color applications. The efficiency degradation towards green is due to the fundamental transition in the band structure of the semiconductor alloy, which changes from direct-type to indirect type with much lower light-generating efficiency. The direct to indirect change is also the reason why AlGaAs LEDs work well in the infrared and red but not in the green part of the spectrum. The InGaN system is a direct-type semiconductor all the way from the red to the ultraviolet. The best devices have been made in the blue and green regions. In the yellow there is a sharp fall off due to problems with materials growth and a bad match in atomic spacing between the different layers of the device structure.

Table 2 shows a selection of commercially available LEDs and their characteristics (data taken from [69]), Figure 14 illustrates their chromaticity in the C.I.E. diagram.

Table 2. Comparison of red, geen, and blue LEDs

Material	Colour	Chromaticity		Peak wavelength,	Half-width,	External quantum efficiency,	Luminous efficiency
		x	y	nm	nm	%	lm/mW
InGaN	ultraviolet			370	11	3	-
InGaN	blue	0.13	0.08	465	30	11	7
InGaN	blue-green	0.18	0.40	495	35	10	18
InGaN	green	0.10	0.55	505	35	10	25
InGaN	green	0.17	0.70	520	40	8	25
GaP:N	yellow-green	0.45	0.55	565	30	0.4	2.4
AlInGaP	amber	0.57	0.43	590	15	5	20
AlInGaP	red	0.70	0.30	635	18	6	20
GaAlAs	red	0.72	0.28	655	25	15	7
InGa/YAG	white: 6500 K	0.31	0.32	460/555	30/150	9	15
W filament∗	white: 2856 K	0.45	0.41				12
Phosphor∗∗	white: 6500 K	0.31	0.33				68

∗ Incandescent. ∗∗ Fluorescent.

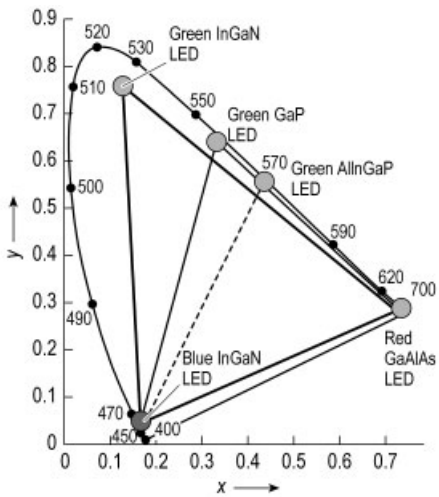

Figure 14. Chromaticity diagram for red, green, and blue semiconductor LEDs

With the availability of efficient blue and even UV LEDs the application of LEDs and phosphors for illumination purposes has gained momentum. Three methods for obtaining white light from LEDs are being followed. The first is to combine three individual blue, green, and red LEDs and the second is a binary complementary system of blue and yellow or blue-green and amber emissions, based on InGaN, GaP, and AlInGaP. The third method involves blue- or UV-emitting $In_xGa_{1-x}N$ and corresponding RGB phosphors or blue-emitting $In_xGa_{1-x}N$ LEDs and YAG:Ce^{3+} phosphor.

4.1.3.2. Organic and Polymer Light Emitting Diodes

There is a class of conjugated organic compounds and polymers with alternating double (C=C) and single (C–C) bonds which show many of the characteristics of semiconductors. They have energy gaps of about the same magnitude, they can be doped to conduct electrons or holes, and their conductivity is rather low. Valence and conduction bands in inorganic semiconductors are determined by their tetrahedral bonding structures formed by sp^3 hybridization. In organic LEDs the analogue to the valence band is the highest occupied molecular orbital (HOMO). The conduction band corresponds to the lowest unoccupied molecular orbital (LUMO). Whereas in semiconductor LEDs the *p-n* junction is usually formed by *p*- and *n*-type doped regions of the same material; in the organic case the *p* and *n* materials are different. In device operation electrons are injected from the metal cathode and holes from the indium tin oxide (ITO) anode into the electron- (ETL) or hole-transporting (HTL) organic layers, respectively. The carriers move through the organic films by field-assisted tunneling from one molecule to the next (hopping). On their way the carriers "meet" at a single molecule and form an exciton. These excitons recombine across the energy difference between the LUMO and the HOMO. For doping devices the exciton is transferred from the host molecule to a fluorescent guest molecule. Due to quantum-mechanical reasons the electron – hole capture

is expected to produce excitons with spin wavefunctions in the triplet and singlet configurations in the ratio 3:1. Only the singlet state is capable of efficient radiation. The other three long-lived triplet states usually recombine nonradiatively. This limitation and the trapping of light limits the maximum possible external quantum efficiency to 8 % or below unless new mechanisms are found to circumvent these conditions. One way of further increasing the quantum efficiency is the use of a phosphorescent dye, which permits triplets to be used to generate photons. Actual devices already run as high as 6.6 % external quantum efficiency for the red color [28].

Following pioneering work [29], [30], research efforts have been directed at developing efficient organic EL devices based on thin films, driven typically below 10 V. Substantial improvements in efficiency, color, and stability of the devices have been achieved.

Organic EL devices generally are bilayer devices, consisting of a hole-transport layer (HTL) and an electron-transport layer (ETL) sandwiched between current injecting electrodes. Potential barriers at the interface of the two layers localize electrons and holes and thus enhance the probability of their recombination in this area. Also, the channel for radiative recombinations can be further optimized by introducing fluorescent centers near this interface. The two-layer structure is thus extended to include a thin emitting layer (EML) between HTL and ETL. The three-layer structure is particularly useful in tuning the EL colors [31].

The current in the organic layers in the EL device is principally space-charge or injection limited. Space-charge limitation is avoided by using organic films only a few hundred nanometers thick. As ETL organic materials often a complex of 8-hydroxyquinoline with a tris-ligand configuration around the central Al ion (called Alq) is used, while the EML contains a small amount of flourescent molecules, and the HTL is composed of an aromatic amine material. Another successful approach is the use of light-emitting polymer material for the fabrication of flexible full-color displays. Low-barrier contacts are designed by matching the organic layers with electrode materials with appropriate work functions like Mg:Ag or Al:Li as cathode material and ITO for the anode. Figure 15 shows the schematic structures of OLED and PLED devices.

OLEDs and PLEDs are excellent candidates for display application. They have fast response, high brightness, high efficiency, wide viewing angle, and easy gray-scale control. A major problem in manufacturing is the incompatibility of the materials with conventional photolithography for patterning. As a consequence shadow masks are used. The device constituents including the cathodes are very sensitive to moisture and oxygen, so extremely good sealing is required, including the incorporation of some dessicant inside the seal.

4.1.3.3. Electroluminescence in Inorganic Materials

Electroluminescent powder-based devices consist of suitably doped ZnS powder suspended in a transparent dielectric binder and sandwiched between two electrodes, one of which is transparent. Activated by the application of about 150 V a.c., these devices are only employed for low-level flat lighting applications. Their d.c.-driven powder counterparts consist of small grains of doped ZnS with an initial onionskin coating of CuS_2 on each particle, which undergo a forming process after initial activation, which results in a high-field zone in the device where light is generated. Although moderately multiplexible these products have disappeared from the market as information displays.

The only inorganic electroluminescent device on the market is the a.c. thin-film (ACFEL or TFEL) device. It consists of a transparent (glass) substrate with indium tin oxide electrodes, onto which a sandwich of a few hundred nanometers of insulator, 500 nm of phosphor, and again a few hundred nanometers of insulator thin films is deposited. On top of this stack a metallic or transparent electrode is applied.

Conceptually of striking simplicity electroluminescence (EL) in thin films of inorganic semiconductors has nevertheless never gained a significant share of the display market. The main reason is the still poor performance as full-color displays. Although encouraging progress has been made in recent years in improving color phosphor materials with blue and broad-band emissions for ELDs, it is not yet sufficient for a large-scale breakthrough on the market. For

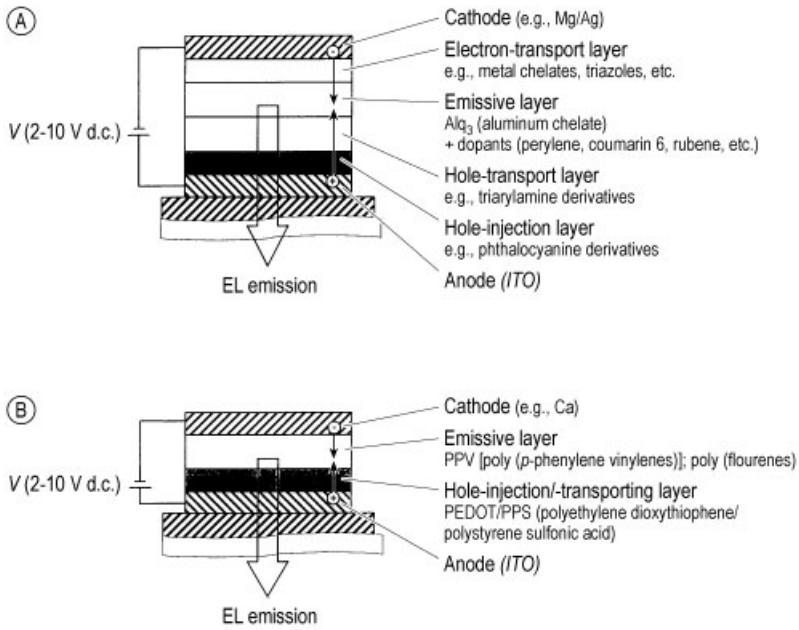

Figure 15. Typical layer structure and materials for A) OLEDs and B) PLEDs

a comprehensive book on inorganic high-field electroluminescent devices, see [32].

The classical EL phosphor is the yellow-emitting ZnS:Mn. This large-band-gap semiconductor is sandwiched between two insulating thin films and operated with a.c. voltage. The power dissipated in generating light is 8 mW/cm^2. Accordingly the power required to generate light from a 10-inch VGA display is just over 1 W, and the luminous efficiency is 6 lm/W.

Two parallel roads have been taken to achieve color EL:

1) The development of an efficient white (broad-band) phosphor that can be filtered to yield red, green, and blue emission (color by white)
2) The development of efficient red, green, and blue primary-color EL phosphors

Efficient red emission (1 lm/W) can be acieved with a combination of the standard yellow ZnS:Mn phosphor and a red filter [33]. The green ZnS:Tb phosphor also has an efficiency near 1 lm/W [34]. In the blue part of the spectrum for some time the efficient SrS:Ce phosphor was available. Actually this phosphor has a blue-green chromaticity and requires filtering to achieve a true blue color. Barrow et al. [35] achieved an important breakthrough by showing that a group of ternary sulfides (thiogallates: MGa_2S_4, M = Ca, Sr, or Ba) activated with Ce are efficient blue-emitting phosphors.

White EL phosphors were first based on the broad emission of the system SrS:Ce,Eu, but the achieved brightness did not satisfy the needs of high-information displays. White emission from stacked layers of ZnS:Mn and SrS:Ce phosphors came close to the luminance level required for filtered-color ELDs [36]. A deep blue emission (and better stability) was claimed for the phosphor SrS:Cu, which in combination with ZnS:Mn emits a true white color [37]. Alternatively to the above thin-film layer structures a structure was proposed with a 30-mm-thick film dielectric layer on a ceramic substrate, which allows higher processing temperatures than glass. On top of this dielectric layer patterned SrS:Ce and (Zn,Mg)S:Mn phosphors are produced, followed by a thin-film dielectric layer, a transparent electrode, and filters which are aligned to the phosphors, thus allowing for better saturation of the blue subpixels [70].

4.2. Light Modulating (Passive) Displays

Passive, nonemissive displays present information by changing the ambient light pattern at the observer. Light radiation can be modulated by changing the amplitude of the electromagnetic wave (absorption), the direction of wave propagation (scattering or refraction), or the polarization of the wave. Detection of polarization changes requires the use of polarizers.

Nonemissive displays are often viewed in reflection. Light is modulated on the way to the reflecting nontransparent mirror and/or on the way back from the mirror to the observer. These reflective passive displays have two distinct features: (1) they require much less power and (2) they do not lose contrast (wash out) in high ambient light. Problems arise, however, if the display is to be read in medium- to low-light conditions and particularly if color is required. Then the displays have to be transmissive, and are supported by a backlight. Thus they gain in visual performance but largely lose the advantage of very low power consumption. To accommodate the different trade-offs transflective passive displays are becoming popular. These displays make use of reflected light under bright ambient conditions and can activate a backlight when the surroundings are not bright enough.

4.2.1. Liquid Crystal Displays

Liquid crystal displays (LCDs) are currently the most important flat panel displays on the market. Liquid crystal displays can be used in a transmissive or a reflective mode, and for some applications in both. The liquid crystal state (mesomorphic state) is an intermediate state between solid and liquid (→ Liquid Crystals). Liquid crystals occur in several mesophases which exist between the isotropic liquid phase and the highly ordered crystalline structure. Technically, the most important mesophases are the nematic (stringlike) and the smectic (soapy) mesophases. Nematic liquid crystal molecules consist of a flexible component such as alkyl groups and a rigid component such as biphenyl group and a cyano group. The cyano group serves to produce positive dielectric anisotropy (the difference between the dielectric constant in the direction of the long axis of the liquid crystal molecule and the dielectric constant perpendicular to the long axis). Molecular design efforts in recent years have involved replacing or modifying the three parts of a liquid crystal to synthesise new liquid crystals. This is illustrated in Figure 16, which shows the main results of chemical modifications on the device performance [38]. Properties that cannot be produced by molecular design are achieved by mixing various liquid-crystal materials, e.g., to increase the practical temperature range.

4.2.1.1. Fundamental Operation

LCDs with the twisted nematic operating mode are one of the earliest technical realizations and still the dominant technology. To build such a device, an orientation layer (polyimide) is applied to the surface of two glasses, equipped with transparent electrodes, and the layer is rubbed. Spacers of about 5 – 8 μm are sprayed onto the surfaces, and a sealing material is applied to the edges of the display The display is assembled by gluing together the two substrates with the rubbing directions orthogonal to each other. When nematic liquid crystals are injected into the gap between the two substrate glasses the director pattern (the unit vector parallel to the average local orientation of the liquid-crystal molecules) changes gradually from the parallel orientation in the immediate vicinity of the top substrate to a 90°-twisted orientation at the bottom substrate. The alignment of the molecules is homogenous with a slight tilt to ensure that all molecules turn in the same direction above threshold. To the outside face of the substrates polarizer films are laminated, so that their polarization axis are either at 90° or parallel to each other.

When light enters the panel it is first linearly polarized and enters the liquid crystal containing gap in a direction parallel to the director of the molecules. The light passing through the first molecular layer rotates along the molecules and enters the second molecular layer with the net effect being that the angle of the linearly polarized light is twisted by an angle equal to that by which the molecular layer is twisted. This occurs repeatedly from layer to layer until at the bottom the linearly polarized light has rotated an amount equal to the twist of the molecules.

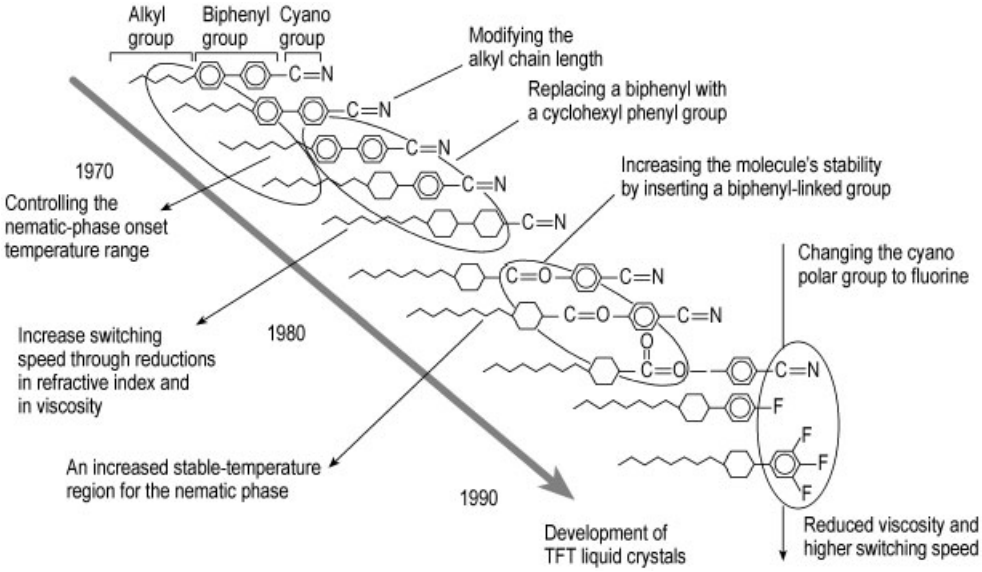

Figure 16. Molecular design for liquid-crystal molecules

Having passed the liquid crystal layer the light crosses the second polarizer film, and depending on the arrangement of the film polarization axis it is totally transmitted (normally white mode) or totally absorbed (normally black mode), as is schematically illustrated in Figure 17.

When an electric field is applied to the liquid crystal layer the anisotropy of its dielectric constant causes the long axis of the molecules to turn parallel to the direction of the electric field, and they line up perpendicular to the substrates. If linearly polarized light now enters the liquid-crystal layer it does not undergo any change. Accordingly, the action of an electric field will cause the normally white pixel to turn to dark, since the polarized light is absorbed in the second polarizer film. Vice versa, in the normally dark mode the electric field switches the pixel on.

4.2.1.2. Simple Matrix-Addressed LCDs

The simple *matrix-addressed LCD* consists of two plates of glass with orthogonal rows and columns of electrodes enclosing the liquid crystal material. Scan pulses are applied sequentially to the rows, and data pulses correlated with scan pulses are fed into the columns. A trade-off exists between the number of elements to be multiplexed and both the display contrast and the response rate.

To turn on an LCD picture element, the sum of voltages on the row and column must exceed the turn-on voltage V_t. If the other elements on the same row or column are to remain off, the potential across these elements must stay below V_t. However, liquid crystals are sluggish and exhibit a time-integrated root-mean-square (rms) voltage response rather than an instantaneous response. The Off elements see the integral of all the column potentials, and as the number of rows N increases, they tend to turn on. The permissible root-mean-square On potentials can just barely exceed V_t, while the Off voltages can be only marginally below it [39]. Since the driving potential should be at least 2 – 3 times V_t for good contrast and viewing angle, multiplexing seriously degrades performance. As the electro-optic characteristics are made sharper, a greater number of lines can be addressed.

In LCD matrix displays the driving circuitry sequentially selects the address lines, while the data information is sent simultaneously to all pixels in the selected line. Twisted nematic liquid crystals have relatively slow turn-on times, and they respond to the time average voltage (rms) instead of the instantaneous voltage. A se-

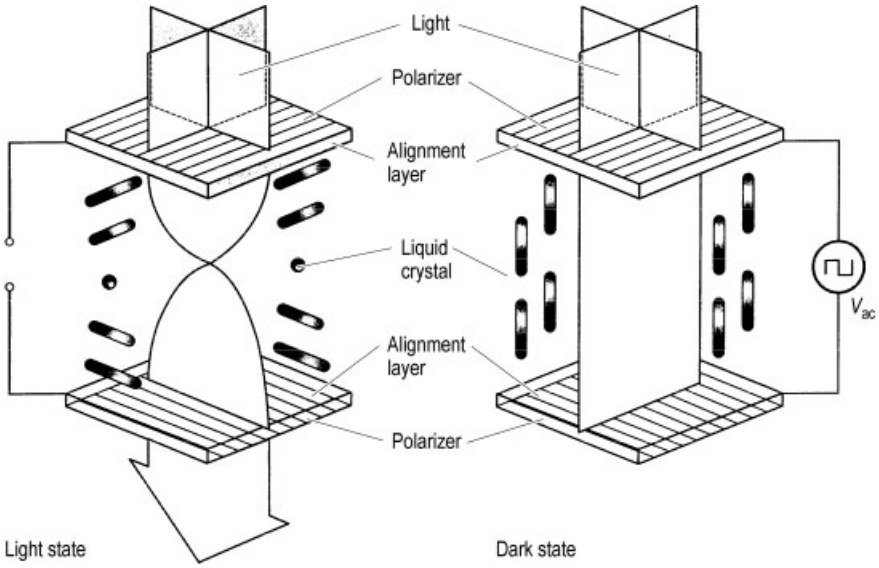

Figure 17. Schematic represemtation of light and dark state in twisted nematic LCDs

lected pixel, which spends most of its time with its driving voltage below threshold voltage, has only a slightly higher rms voltage than an unselected pixel. All contrast must be produced with only this small difference in voltage. This is why a highly nonlinear response is required, as is provided by supertwist (STN) LCDs. Adding some chiral nematic liquid crystals and chiral compounds to nematic liquid crystals causes the nematic liquid crystal to assume a helical structure. A critical factor is the angle by which the helix rotates. The original smooth threshold characteristics can be significantly improved by setting the twist angle to 240 – 270°. This operating mode is referred to as the STN mode.

A drawback of STN is the inherent comparatively long response time (100 – 300 ms) and the chromaticity of the cell. To eliminate the chromaticity double supertwisted nematic cells (D-STN) have been developed. Here a second LC panel containing a helix of the opposite orientation is placed in front of the first device. This technique later gave way to film-compensated STN (F-STN), in which the undesired coloration was compensated for by a phase-difference film.

4.2.1.3. Novel Addressing Schemes for Direct-Addressed STN LCDs

The inherent response time of the display must be many times longer than the refresh time to satisfy the rms response condition. This generally rules out displays with response times short enough to show moving images at video rates. Indeed, if the display uses a slow-responding LC mixture, the difference between Off and On states is never large enough to generate good contrast. If the display uses a fast-responding LC, then the contrast of an individual row begins to fade before the next refresh cycle begins, again deteriorating contrast. This phenomenon is called frame response; instead of responding to the average voltage applied over several frames the LCD reacts to the relatively high voltage row pulse within a frame.

The problem can be avoided by different approaches (Active Addressing [40] and Multiple-Line Selection [41]). Instead of sending the high-voltage pulses sequentially down the rows, a set of predetermined signals is applied to multiple rows simultaneously and a unique calculated analogue voltage is applied to the columns, depending on the state of the pixels in that column. The key is to determine which rows need which signal and to calculate the proper column voltage within the time frame available. An LCD cell is

needed that is capable of switching at video rates (i.e. by decreasing the cell thickness and using fast material).

4.2.2. Reflective LCDs for Low-Power Systems

Reflective color LCD without back light would be the enabling technology for multifunctional portable information terminals with which one could receive and send any information any time at any place. Mechanisms such as phase retardation, rotation of polarization, absorption, light scattering, and Bragg reflection have been used to modulate reflected ambient light. Such displays do not need heavy batteries and represent a promising approach to ″electronic paper″ today. (Further technologies for this application are described in Section 4.2.5).

4.2.2.1. Reflective LCDs with Polarizers

Making use of phase retardation and polarization rotation requires the use of two or at least one polarizer for the LCD. In a simple 90°-twisted nematic cell with two crossed polarizers, the second polarizer is placed between the back substrate and the reflector. It reduces the brightness by about 30 % and causes parallax (double images), limiting the device resolution. If the second polarizer is removed the parallax is eliminated. The light however still has to traverse the polarizer twice, and this limits the throughput to 38 %. A single-polarizer LCD exhibits a better contrast ratio than a nonpolarizer LCD, although its reflectivity is lower. A high contrast ratio is essential for color displays. The single-polarizer LCD offers a compromise between display brightness and contrast ratio, and actually represents the mainstream approach. A comprehensive description of the basic LC modes and design concepts is been given in [42].

4.2.2.2. Polarizer-Free Reflective LCDs

An LCD that does not require a polarizer shows good reflective properties and a wide viewing angle, but the contrast is often limited to values of around 10:1. Three different types of polarizer-free LCDs have been developed: The chiral nematic (or cholesteric) LCD, the polymer-dispersed liquid crystal display (PDLC), and the guest – host (GH) display.

In *cholesteric liquid crystals* the chiral components cause the molecules to spontaneously align on the surface of a helix. These helices are stabilized by the anchoring forces of the orientation layer. If in the voltage-off state the product of the pitch length p of their helical structure and the average refractive index n is equal to the wavelength of the light l (i.e., $\lambda = np$), the Bragg condition of constructive interference is fulfilled, and the material has the property of selectively reflecting light of a specific wavelength. In helical structures the incident white light is decomposed into right and left circularly polarized components, one of which is reflected and the other transmitted. The transmitted light is absorbed by a black layer on the rear inner substrate surface. In the voltage-on state the periodic helical structures are changed to focal-conic and hence the Bragg reflection is interrupted. The incoming light is absorbed by the black layer and the entire pixel appears dark. Cholesteric liquid crystals can form one of three major textures depending on boundary conditions and applied fields [43]: planar, focal-conic, or homeotropic (Fig. 18). With increasing field strength the planar structure first transforms into a focal-conic structure, then into a homeotropic aligned nematic state, which is highly transparent. The planar texture is highly reflective at a preselected wavelength and bandwidth. By introduction of a polymer network or suitable surface treatment of the cell substrates, a domain structure is introduced. This broadens the viewing angle, stabilizes the focal-conic texture to create bistable switching, and introduces a gray-scale regime. The ability of the display to maintain pixels both in the On and Off states once they are written (without refresh) allows for unlimited resolution on a low-cost passive matrix without sacrificing contrast. Full-page-size prototypes [44] have been demonstrated.

In the first engineering samples a total reflection of around 40 % was observed, substantially greater than that of an ordinary reflective twisted nematic LCD. The achievable contrast is determined by the transparency of the focal-conic structure, which can exhibit slight scattering. Different driving schemes are still under

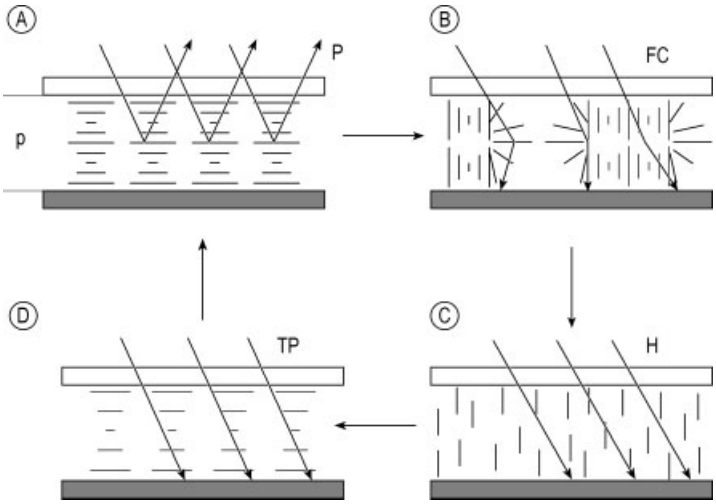

Figure 18. Schematic textures in cholesteric liquid crystal material
a) Light-reflecting planar texture; b) Nearly transparent focal-conic texture; c) Transparent homeotropic texture; d) Transient planar structure

development. Real-time television has already been viewed on a cholesteric display cell at Kent State University [45].

PDLC operation is based on light-scattering induced by the refractive index mismatch between micron-sized liquid crystal droplets and the surrounding polymer matrix [72]. In the normally black mode, the LC droplets scatter light due to index mismatch at zero field and transmit light above a threshold field. (The reverse mode is also possible.) In the transparent mode, due to the application of the electric field, the LC droplets are aligned nearly perpendicular to the substrates, with the refractive index of the extraordinary ray in the liquid crystals matching that of the surrounding polymer matrix. The light passes through the PDLC layer and is reflected from the rear reflector. PDLC devices are much brighter than ordinary reflective twisted nematic cells, but their contrast ratio is lower. A new development is using PDLCs to form switchable reflective holograms. Such gratings are formed by curing a prepolymer in the interference fringes of two coherent counterpropagating laser beams. The periodic variation in intensity results in localized regions of higher and lower polymerization rates. The resulting films have periodic LC domains separated by dense polymer, and consequently a periodic modulation of the refractive index. In the field-off state, diffraction of light due to the presence of the holographic grating is observed. This can be turned on and off by the application of an electric field. Commercial application is however still a few years away [46].

Guest – host displays employ the absorption anisotropy of dichroic dye molecules dissolved in amounts of 1 – 5 % in a liquid-crystal host. The host material is highly transparent for the visible spectrum, whereas the dye dopants absorb one polarization of the incoming light and transmit the other. The use of polarizers has been avoided by, for example, a quarter-wave film behind the rear substrate and the reflector (Cole – Kashnow) or the application of a chiral dopant to form twisted guest – host layers that absorb both polarizations of the incoming light while traversing the doped layer twice (White – Taylor).

4.2.3. Bistable Liquid Crystal Devices

An LCD is called bistable if it assumes two different configurations at zero electric field. Which configuration is displayed depends on the addressing history of the device. The respective two stable states represent minima in the total energy stored. Bistable devices are able to maintain an image memory at zero field, like a printed page of a magazine. Clearly they consume substantially less energy than those in which the

image must be refreshed periodically. The most prominent representative of bistable LCDs is the ferroelectric liquid crystal display.

Optically active (chiral) smectic liquid crystals have a permanent electric dipole moment in the absence of an electric field, and for this reason they are also called ferroelectric liquid crystals (FLCs). The intrinsic polarization of these molecules, which is perpendicular to their long molecular axis (director), has the consequence that when subjected to an external field the molecules respond to the direction as well as the magnitude of the field. The helical structure would average out the direction of the spontaneous dipole, thus suppressing the ferroelectric behavior. Rubbed polymer surface alignment layers, similar to those in supertwist LCDs, remove the helical structure. The spacing between the two glass substrates should be about 2 μm. This spacing is also required for the ferroelectric liquid crystal cell to act as a half-wave plate. The operation of a FLC display depends on the switching of the FLCs dipole, whose orientation depends on the polarity of the addressing pulse.

The FLC molecule cannot simply rotate about its own axis; instead it is constrained to rotate about an imaginary cone (Fig. 19). If the cone half-angle Θ, which is equal to the tilt angle of the molecules in the smectic layer, is approximately equal to 22.5°, then the change in the orientation of the FLC molecules 2Θ is 45°. This is the angular change that produces the optimum values of transmission and contrast ratio [47].

The two main attributes of FLC displays are very fast switching, typically 100 ms per row, and bistability. This property can be made use of to produce a nonvolatile display in which the image is maintained at zero potential. The molecules switch in the plane of the display rather than perpendicular to it (as in supertwist and active matrix LCDs). This leads to a very good viewing angle. The main obstacles in the commercialization of ferroelectric LCDs have been production issues (very small distance between substrate plates), the absence of gray scale, and susceptibility to mechanical shock, although it is said that there are solutions to these problems. In addition to ferroelectric LCDs there are also antiferroelectric LCDs. Yamada et al. [48] have developed a display with gray scale and have demonstrated high-quality video images on a 125×80 mm^2 display.

4.2.4. Active Matrix-Addressed Liquid Crystal Displays (AMLCDs)

4.2.4.1. Fundamental Properties

Active matrix addressing leaves the fundamental twisted nematic operation (Section 4.2.1.1) unchanged and uses pixel-external nonlinear devices to enhance the contrast of large multiplexed displays. The devices do not respond to the rms time-average voltages but rather to the instantaneous peak voltage. Two-terminal (metal – insulator – metal structures) or three-terminal (thin-film transistors) devices act as switches, electrically isolating a pixel until its input voltage exceeds a certain specified threshold. Pixels in active matrix devices are driven at duty cycles close to 100 %. The voltage at the pixel remains practically constant during each frame, as if the pixel were directly driven. The pixel acts as a capacitor, storing the charge and maintaining the voltage during the frame time. Thin-film transistors (TFTs) formed by sophisticated deposition and patterning technologies are now the most frequently used switches in commercially available AMLCDs. A schematic TFT matrix, the equivalent circuitry, the addressing pulse sequence, and the physical layout of an AMLCD pixel is shown in Figure 20 The semiconductor material used for the TFT in most cases is amorphous silicon because it can be fabricated at relatively low temperatures and generates high image quality. The use of poly-silicon material would allow the integration of driver circuits on the same glass plate due to the much higher mobility achievable. The high temperatures necessary to process poly-silicon are not compatible with normal glass substrates. Solid-phase crystallization and excimer laser annealing for large-grain Si crystallisation and special furnace or laser annealing for doping activation are key processes for the advancement of poly-silicon AMLCD technology. For a recent review of the basics of AMLCDs, see [49].

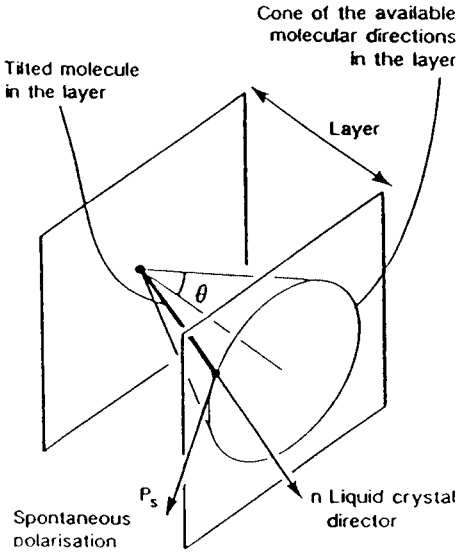

Figure 19. The FLC molecule is constrained to rotate about an imaginary axis

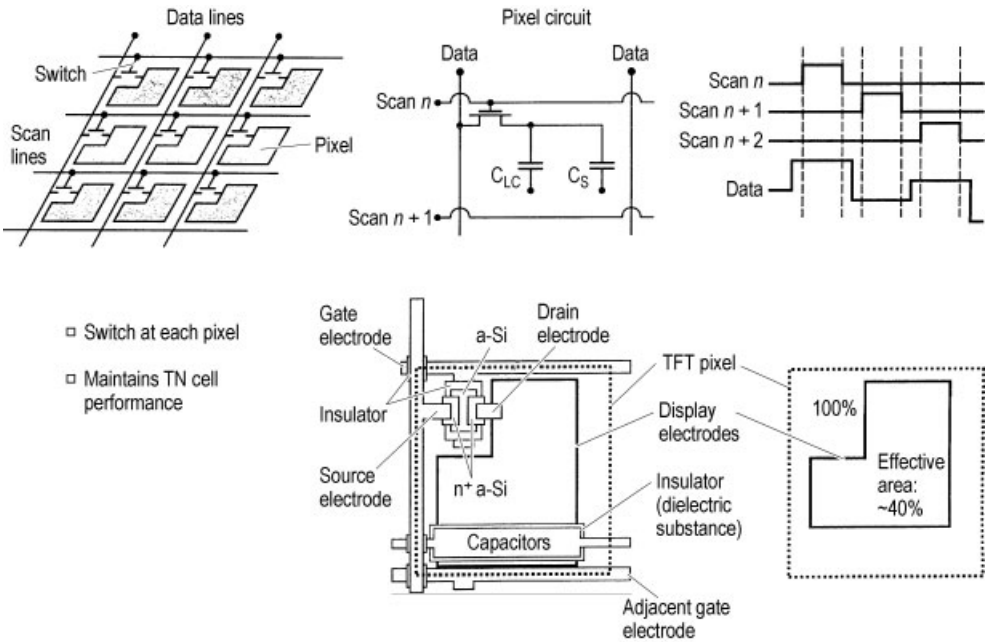

Figure 20. TFT matrix, equivalent circuit, addressing pulses, and physical layout of an AMLCD pixel

4.2.4.2. Manufacturing Issues

AMLCD manufacturers are now working on the fourth generation with significant improvement in substrate size (730×920 mm^2, accomodating, e.g., six 17-inch faceplates). The coming wave of AMLCDs is intended to at least partially replace CRT monitors. This means that competitive size, cost, and performance of the flat displays will be of paramount importance. Manufacturers will ultimately be targeting mother glass sizes of one meter diagonal. They will simplify their TFT processes by lowering the number of photo masks to four and simplify LC-specific fabrication steps such as the preparation of the orientation layer, LC filling methods, and lamination of polarizers and color filters. They will have to develop means to lower the resistance of bus lines. At the same time they are striving to widen the viewing angle and reduce power consumption of the panels. Figure 21 shows a cross section of a presently available color AMLCD module and indicates how the driving circuitry is connected to the glass substrate carrying the TFTs. Mostly anisotropic conducting films are used to connect the ITO pads on the glass to the driver LSIs. A method of growing importance is the use of chip-on-glass architectures.

Many manufacturers are already envisioning the evolution from the near-future implementation of TFT drivers in poly-Si technology to the ultimate system on glass with the signal processing and driving units fabricated all in poly-Si thin-film technology.

4.2.4.3. Viewing Angle Improvements

The viewing angle is a critical problem for large panel liquid crystal displays. Several technologies have been proposed to improve the originally rather narrow and nonuniform viewing characteristic of single-domain twisted nematic, homogenous (LC molecules oriented nearly – depending on pretilt angle – parallel to the substrate), homeotropic (LC molecules oriented perpendicular to the substrate, vertical alignment) cells. Methods of improving the viewing angle can be external, internal, or a combination of both. In external methods the LC cell is provided with a *phase-compensation film* and/or diffusive optics laminated outside the display. In the internal method the LCD structure is modified internally. Whereas external methods are easy to manufacture and accordingly cheap, they do not improve the viewing angle to more than $\pm 60°$. State-of-the-art approaches generally involve both internal modifications and the usage of external compensation films.

One major internal technology is the formation of *multi-domain cells* with reversed LC orientations in each domain. Each pixel is divided, e.g., into four subpixel domains. The generation of these four domains can be done in a variety of ways: reverse rubbing [50], evaporation, photo-alignment [51], or domain control electrodes [52].

A special modification in the case of vertical alignment (VA) cells is the protrusions method. One group has obtained a *multi-domain VA (MVA) cell* by implementing roof-shaped protrusions on both the TFT and color filter substrates. The domains are automatically controlled by the slopes of the protrusions [53].

Another wide-view method is the *axially symmetric-aligned microcell (ASM)*. Here the LC alignment has a spiral distribution and is symmetric with respect to the cell normal. The ASM mode has been applied to a 42-inch plasma-addressed liquid crystal display (see Section 4.2.4.5) [48].

In common twisted nematic cells the applied electric field is in the longitudinal direction, and the tilted molecule directors cause a different phase retardation viewed from either the right or the left direction. This is the origin of the narrow and asymmetric viewing angle. If the electric field is transverse the LC molecule directors are rotated in the plane. This *in-plane switching* was first devised by Kiefer et al. [54] and implemented by Kondo et al. [55].

Standard AMLCDs have five electrodes, three for gate, source and drain of the transistor, one for the pixel electrode, and one common electrode. In most cases the common electrode is formed on the countersubstrate. As shown in Figure 22, in-plane switching uses an interdigital electrode structure. The five electrodes are arranged in two layers on the TFT substrate side. The lower layer carries the gate and common electrodes, the upper one the source and drain electrodes which are co-used as pixel electrodes. This approach avoids the transparent in-

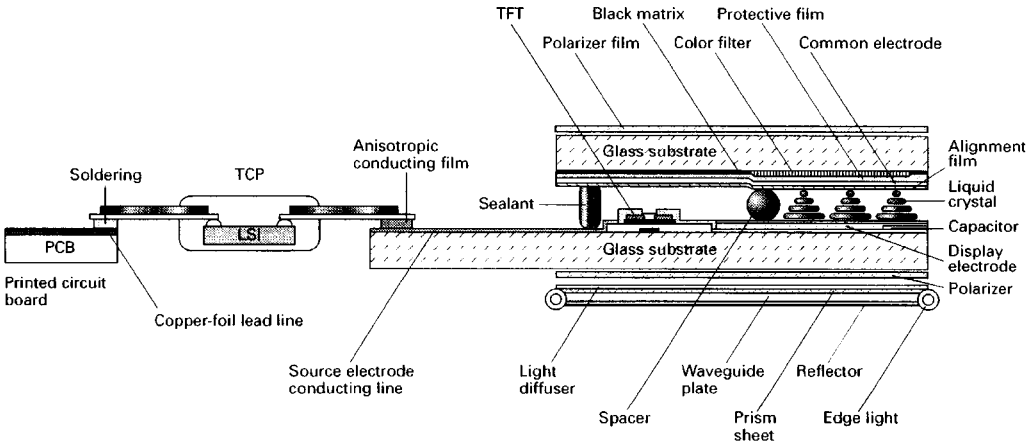

Figure 21. Cross section of a color AMLCD and connection to the addressing circuitry (after [38])

dium tin oxide electrode on the countersubstrate. The electric field is no longer vertically but horizontally oriented, as indicated by the arrows in Figure 22. Under the influence of this field the optical axis of the nematic LC molecules deviates from the polarization axis and the transmittance of incident light through crossed polarizers is gradually increased. Horizontal and vertical viewing angles of 70° have been achieved at a contrast ratio of 10.

Another implementation of a wide viewing angle cell is the *film-compensated p-cell*. Basically p-cells resemble homogenous cells. The difference is that in homogenous cells the directions of the pretilt angle at the upper and lower substrate are parallel, whereas in p-cells the directions of the pretilt angles of the upper and lower substrate are antiparallel. These opposite pretilt angles result in a rather wide and symmetric viewing angle. As the applied voltage increases the LC directors go through splay and then bend formation. With proper phase compensation (e.g., biaxial retardation film) the *optically compensated bend (OCB) cell* exhibits a wide and nearly symmetric viewing angle with fast response times [56].

4.2.4.4. System Performance

The overall transmissivity of AMLCDs is determined by the transmissivities of the glass (95 %), the polarizers (35 %), the color filters, (25 %), and the cell aperture (that portion of the cell which is transparent to light), which depending on design can vary between 40 and 65 %. As a result the overall transmissivity of the cell is not more than 3.3 – 5.0 %. A major research goal for AMLCDs is to improve transmissivity to reduce power consumption. Two other targets for further power reduction are improvement of the efficiency of backlights and reduction of the power consumption of the drive circuit unit. Backlights consume about 70 % of the total power necessary to operate a transmissive AMLCD module. Therefore, for portable applications much research effort still goes into the improvement of reflective color AMLCDs. Up to now the reflectance of such devices reaches only about 40 % as compared to newspapers or magazines, which achieve 60 % and more [71].

Improving legibility, particularly under conditions of low ambient light, requires some form of either front- or backlight illumination. Transflective AMLCDs are characterized by a reflective layer behind the LC layer (as seen from the viewer) that is structured to allow the additional backlight illumination pass through the reflective layer, e.g., via holes. They retain good picture quality under all ambient-light conditions at reasonable energy consumption.

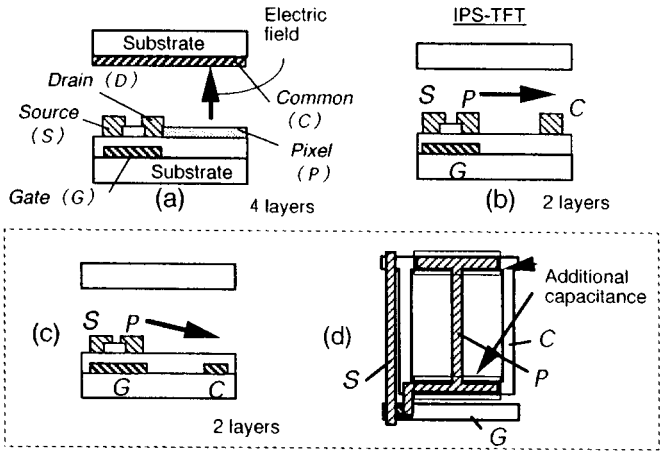

Figure 22. Cross section of a conventional AMLCD (a) and of an in-plane switching AMLCD (IPS TFT) (b). In the IPS version (c) the common electrode is located in the gate electrode layer. d) Planar arrangement of c) (after [55])

4.2.4.5. Plasma-Addressed Liquid Crystal Displays

Despite their immense success in the portable display area and the anticipated large market share they will conquer in the monitor market in the near future it is not predicted that AMLCDs will be available as meter-sized panels for consumer HDTV in the next few years. The scaling difficulties of photolithography and metallization to sizes larger than 40 inches can elegantly be bypassed by a new approach called plasma addressed liquid crystal (PALC) [57], [58]. PALC uses the electrical switching properties of an ionized gas to actively address the liquid crystal pixels. In the integrated LCD/plasma structure of the PALC, each scan line is defined by a plasma channel. The cathodes in the channels are sequentially scanned by applying a plasma discharge voltage that produces a full plasma within several microseconds. The analogue video signals are fed to each column electrode. Except for the one row in which the gas is ionized, the analogue voltage has no effect on the light transmissivity of the liquid crystal since there is no electrical connection or conductor on the bottom side of the liquid crystal. Along the line above the ionized channel, however, the pixel capacitance can charge to the value determined by the data voltage. In effect, the plasma completes the electrical circuit between the data electrodes and the grounded channel electrode. Kakizaki et al. [59] have presented a 25-inch PALC display for wall-hanging TV with 1.8 lm/W light efficiency, 250 cd/m^2, and high contrast, even in bright ambient light.

4.2.5. Microparticle-Based Displays

The common basis of all microparticle-based display techniques discussed here is the manipulation of small objects such as pigments or rotating microscopic balls by means of electrical forces. The mechanical manipulation of these objects allows for a spatial modulation of the optical properties of display cells, such as their reflectance or transmission behavior for light. In the following, the most promising emerging techniques in this field are reviewed and several alternative approaches are discussed briefly.

4.2.5.1. Electrophoretic Displays

The electro-optical effect used in these types of display cells results from the movement of charged microparticles in a colloidal suspension.

The migration of small charged particles through a liquid induced by an electrical field is known as electrophoresis. Since the pigments are chosen to have reflectance properties different from those of the liquid, the overall optical appearance varies depending on the distribution of the pigment particles in the volume of the display cell. Research activities on these electrophoretic displays (EPDs) were already reported in the early 1970s [60]. This nonemissive display technology offers a very wide angular viewing range and high contrast ratio over a wide range of illumination. After a slowdown in the 1980s recently considerable effort has been spent on the development of prototypes with properties comparable to printed paper. Hence the popular terms "electronic ink" or "electronic paper".

As shown in Figure 23, the EPD suspension is commonly composed of a darkly dyed liquid with low reflectivity, whereas the microparticles are of a light color. For the microparticles, titanium dioxide particles with sizes ranging from the submicrometer regime to several micrometers are used. When a thin layer of this suspension is sandwiched between two parallel electrodes, of which at least that directed towards the observer is transparent, the particles either move to the top or to the bottom electrode, depending on the polarity of the applied field. The observer either sees the opaque dark dye (off state) or the reflective white particles (on state). Since this effect is only controlled by an electric field perpendicular to the electrodes, the addressing of many pixel elements can be realized in parallel by means of appropriate electrode structure. One of the most important advantages of EPDs with respect to energy efficiency is the bistability of the display cells, i.e., the cell remains in its actual state (on or off) after the electric field has been removed. This is achieved by matching the densities of the suspension and the microparticles and thus establishing gravitational stability. Mechanical shocks or thermal influences will disturb this stability. In general, the resulting image contrast is determined by the dye concentration in the suspension, by the thickness of the layer, and by the effective electrical field which is applied to the electrophoretic layer.

When the particles are bunched together or even compressed, the danger of agglomeration and clustering occurs at high electric fields. This has a negative effect on lifetime and quality of the display cell. Therefore, the additional introduction of repulsive forces is necessary to prevent the individual microparticles from sticking together. Colloidal suspensions inherently tend to be unstable. The large surface free energy is reduced as soon as the particles bunch together, which in turn enlarges the attractive forces. Establishing a colloidal stability of the suspension is the most crucial aspect in EPDs and can be achieved by electrostatic or by steric repulsion. While electrostatic repulsion is based on a precise charging of the pigments, steric repulsion is accomplished by covering their surfaces with special polymer molecules. Finally, dye, solvent, and the microparticles must be prevented from interacting chemically.

Several proposals for modifications of cell design and electrode arrangements have been made, such as etched cavities and walls for each display cell for a compartmentation of the particles. In addition to these architecture-related approaches, a materials-related approach in which the electrophoretic suspension is confined in microcapsules (microencapsulation) has been very successful [61]. The microcapsules, with typical diameters between 30 and 300 μm, result from in situ polycondensation of urea and formaldehyde. The capsules have mechanically strong, optically transparent walls with a relatively high dielectric constant. This technology is compatible with conventional techniques like screen and ink-jet printing. The schematic structure of the microencapsulated EPD is shown in Figure 24

4.2.5.2. Rotating Ball Displays (Gyricon)

The basic operating principle of a Gyricon display is depicted in Figure 25. A clear and transparent elastomer sheet is sandwiched between two opposite electrodes. Small balls or beads are randomly dispersed in this substrate sheet, each embedded into a liquid-filled cavity. The balls are composed of two different materials resulting in a structure of two hemispheres with contrasting color and opposite electrical charge. Thus, each ball has a macroscopic electrical dipole character and is free to rotate within its cavity when an electrical field is applied. The observer thus sees either the dark or the light side of the balls, which finally produces an image. For ball sizes of 10 μm and a thickness of the sub-

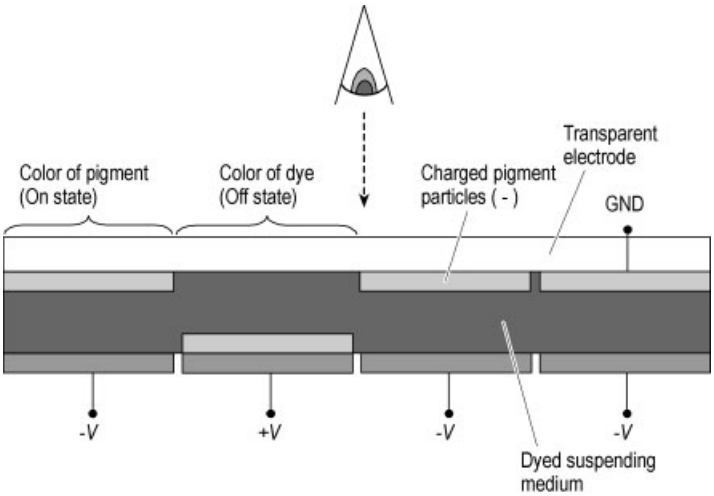

Figure 23. Schematic structure of an electrophoretic display. The white, negatively charged microparticles are suspended in liquid dyed dark blue and can modulate the reflectivity of the display surface under the action of an electric field.

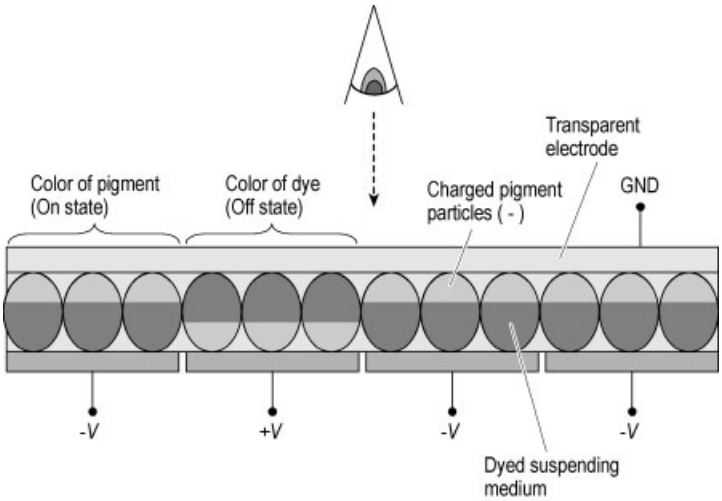

Figure 24. The microencapsulation of particles in the electrophoretic cell successfully solved the degradation problem due to particle agglomeration

strate sheet of 60 µm, minimum programming voltages of between 50 and 150 V have been reported [62].

5. Projection Displays

In a projection display the image is not viewed directly but enlarged through an optical system and projected on a screen. The two major configurations of projection displays are *front projection* (viewer and projector on the same side of the reflective screen) and *rear projection* (viewer and projector on opposite sides of the transmissive screen). According to the mechanism used to generate the image three categories are distinguished: 1) CRT projection (image formed on the CRT face-plate), 2) light valve projection (light valve on which the image is produced modulates the light of a light source), 3) laser projection (image generated on screen by laser beam deflection).

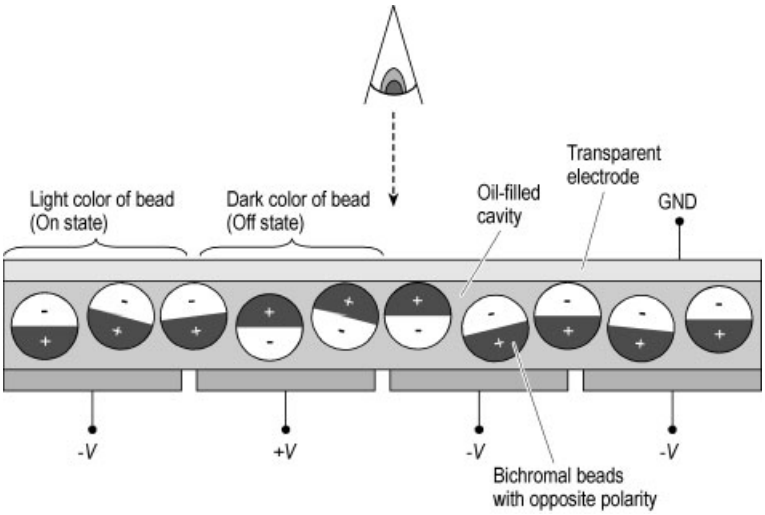

Figure 25. Principle of a Gyricon rotating ball display. Small hemispheric bichromal balls are dispersed in a thin elastomer sheet. Each ball is placed in an oil filled cavity and freely rotates to adjust its polarity to an applied field.

5.1. Projection CRT

In the standard configuration still widely used today the output of three CRTs with red, green, and blue phosphors are collected by three optical lenses and focused onto a screen. The lenses are often coupled to the CRT faceplates by index-matched fluids to enhance throughput and to provide cooling for the monochrome, shadow-maskless CRTs, which are driven at high current. Lenses are often lenticular projection lenses with contrast-enhancing black stripes. The optics of the lenses and the screen are designed to converge the three color images and direct the image to the viewer in spite of the off axis projection. Limiting factors of this mature technology include the resolution/brightness trade-off inherent in CRT technology. As the electron beam current is increased space charge repulsion within the beam causes beam "blooming", and light scattering in the CRT phosphor itself further lowers the spatial resolution.

Taking the minimum electron beam spot size to be 0.23 mm, a typical 7-inch CRT (with 130 to 140 mm usable diagonal) is capable of only 450 – 490 distinct spots in a horizontal scan line.

The flexibility provided by the multisync raster scan of a CRT display is a significant advantage when displaying signals with different formats and resolutions. Matrix displays have fixed format and resolution and consequently must incorporate image scaling and format conversions.

5.2. Liquid Crystal Projection

The development of small, high-resolution liquid crystal light valves has led to the successful introduction of LC projection displays for use in conference rooms and portable applications. Amorphous, polycrystalline, and single-crystal silicon active matrix arrays have been used to control the LC layer. As higher resolutions are demanded the trend is to polycrystalline single-crystal silicon address circuitry. The general idea of these projectors is to illuminate the light valve by an intense arc lamp and to magnify and project the transmitted or reflected image onto a viewing screen. Since single-crystal silicon is opaque at visible wavelengths, liquid crystal on silicon (LCOS) devices can only operate in reflection. (An exception to this is the silicon on insulator transfer technology for microdisplays [63]).

5.2.1. Transmissive Projection LCDs

The current market for front projection displays for business applications is dominated by polysilicon light valve technology. This technology

is similar to that described in Section 4.2.4. The ability to integrate driver circuits in the poly silicon material has been important in enabling the technology to reach rather high resolution in an economical package. The AMLCD light valves are typically one inch in size and have pixel sizes ranging from 20 to 40 µm. Considerable effort has been spent on improving the optical throughput (aperture and microlens arrays) and the optical shielding of the active circuits (elimination of photo-excitation) in the device.

The common optical layout for AMLCD projectors provides for the collection of light from a small arc lamp and its division into RGB components. Each color is then separately modulated in transmission by an AMLCD light valve. The resulting three images are recombined into a full color image by a dichroic color-combining x-cube. Finally the image is magnified and projected to the screen. Alternatively, tilted dichroic mirrors can be used to direct the three color components of an illumination system to a single AMLCD light valve containing RGB subpixels.

In addition to the optical components like dichroic mirrors and x-cubes, recently very significant improvements have been achieved in the arc lamp and illumination optics. Small arcs for the efficient collection and focusing ofthe light on the small light valve and long lamp life were the main development goals. Since AMLCDs use polarized light, 50 % is immediately lost from the unpolarized lamp light. In principal this "wrong"-polarization component can be recovered by smart polarization conversion systems, which can improve the system throughput by as much as 80 % [64].

5.2.2. Reflective Projection LCDs

Liquid crystal on silicon (LCOS) light valves have been developed for applications ranging from highly mobile microdisplays to high-brightness, high-resolution projection displays. These devices employ an active matrix array manufactured with standard CMOS processing on single-crystal silicon. The array transistors supply the desired voltage to reflective aluminum mirrors rather than to the ITO transmissive electrodes as in poly-Si AMLCDs. Since the layout of the reflective mirror is made to cover row and column lines and the circuit itself it is evident that this design has an aperture ratio advantage over transmissive designs. In fact aperture ratios of about 90 % are routinely achieved. The CMOS transistor array fabrication is followed by a few levels of additional processing which provide the light-blocking layer, the reflective mirrors, and their contacts to the transistor array. The use of chemical mechanical polishing, a technique developed for multilevel wiring on VLSI circuit chips, provided a major advance in the performance of LCOS devices because of the requirement that the mirrors be highly specularly reflecting. One great advantage of LCOS is its scalability to small pixels and hence to small and inexpensive devices. Further advantages of the technology are the good thermal conductivity of silicon, which helps to avoid problems associated with heating, and the ability to inegrate function by using high-quality CMOS circuitry on the array chip.

A variety of LC modes, including twisted nematic, vertically aligned homeotropic, polymer-dispersed liquid crystal, and ferroelectric, have been successfully implemented. The optical architecture must effectively separate the incident illumination from the reflected image. This is usually done by three polarizing beam splitters in the RGB channels, which polarize and analyze the light and separate the incident light from the reflected image.

5.3. Micromechanical Light Valve Projection

Three separate light valve technologies based on arrays of moving micromechanical devices, which either reflect or diffract light, have been developed. Switching is fast enough in each of these technologies to allow for color sequential operation. The optical throughput is larger than in polarization-sensitive LCD devices, since these devices do not require polarized light.

The activated mirror array consists of an array of piezoelectrically activated thin-film micromirrors [65]. These mirrors are driven by an active matrix array fabricated under the mirrors by CMOS technology. The angular deflection of the mirrors is a function of the supplied voltage, giving an analog light intensity versus voltage response behaviour.

The grating light valve consists of a large number of individually addressed small diffraction gratings [66]. Depending on the voltage applied the grating acts on the illuminating light either as a reflecting mirror or as a diffracting element with an analog diffraction response behavior. A Schlieren optical system allows the first-order diffracted light to enter the projection lens while rejecting the reflected light. The linear array of diffraction gratings is illuminated by a laser source.

Whereas the above-mentioned two technologies have not yet entered the market, digital micromirror devices (DMD) and the digital light processing (DLP) [67], [68] have had significant commercial impact. The DMD consists of an array of many individual reflecting $17 \times 17\,\mu m^2$ micromirrors, each which can be electrostatically deflected by address voltages. The mirrors are binary. In one state they reflect incident light into the projection lens, and in the other do not. During operation light from an arc lamp is chopped by a color wheel into its RGB components which sequentially illuminate the single DMD device. The device is driven rapidly with the appropriate RGB image data. The three monochrome images are reflected into the lens and temporally averaged by the human visual system to form a full color image. Gray scale is achieved by binary pulse width modulation.

6. References

Selected comprehensive conference reports: *Society for Information Display (SID) International Symposium*, *Digest of Technical Papers; SID Seminar Lecture Notes*; *SID Short Courses*; *SID Application Seminars*; *Conference Record International Display Research Conference/Eurodisplay/Asia Display*; *International Display Workshops*.

General References

1. J. A. Castellano, *Handbook of Display Technology,* Academic Press, Inc. San Diego, 1992.
2. V. G. Chigrinov, *Liquid Crystal Devices,* Artech House, Boston, London, 1999.
3. P. S. Drzaic, *Liquid Crystal Dispersions,* World Scientific Publishing Co., Singapore, 1995.
4. P. A. Keller, *Electronic Display Measurement,* John Wiley & Sons, Inc., 1997.
5. E. Lueder, *Liquid Crystal Displays,* John Wiley & Sons, Inc., 2001.
6. L. W. MacDonald, A. C. Lowe (eds.): *Display Systems,* John Wiley & Sons, Inc., 1997.
7. T. J. Nelson, J. R. Wullert II, *Electronic Information Display Technologies,* World Scientific Publishing Co., Singapore, 1997.
8. N. Horie, T. Sugawara, T. Oyama S. Sherr, *Electronic Displays,* John Wiley & Sons, Inc., 1993.
9. S. Sherr, *Applications for Electronic Displays,* John Wiley & Sons, Inc., 1998.
10. L. E. Tannas, Jr., *Flat Panel Displays and CRTs,* Van Nostrand Reinhold, New York, 1985.
11. R. Young, (2001) SDI Seminar Lecture Notes, M-1/1–125.
12. S.-T. Wu, D.-K. Yang, *Reflective Liquid Crystal Displays,* John Wiley & Sons, Inc., 2001

Specific References

13. D. den Engelsen (2000), SID Seminar Lecture Notes, M-7/1-23.
14. A. M. Morell (1981), SID '81 Digest, 3 – 9.
15. H. R. Seelen et al., *RCA Rev.* **16** (1955) 122 – 139.
16. F. C. Gehring, T. van Zutphen (1999), SID '99 Digest, 364 – 367.
17. D. den Engelsen (2001), SID '01 Digest, 1310 – 1313.
18. H. Horie, T. Sugawara, T. Oyama (2001), SID '01 Digest, 1319 – 1321.
19. Y.-M. Teng, H.-S. Tong, C.-M. Hu (2001), SID '01 Digest, 1322 – 325.
20. L. E. Tannas Jr. (ed.): *Flat Panel Displays and CRTs,* Van Nostrand Reinhold, New York, 1985, pp. 91 – 134.
21. L. F. Weber (2001), SID Seminar Lecture Notes, F-3/1-54.
22. H. Tolner (2001), SID Seminar Lecture Notes, F-4/1-21.
23. R. Meyer (1991) Tech. Digest 4th Intl. Vac. Microelectronics Conf., Nagakama, 6 – 9.
24. D. Temple, *Mat. Science a. Engineering* **R24** (1999) 185 – 189.
25. C. P. Kuo, R. M. Fletcher, T. D. Osentowski, M. C. Lardizable, M. G. Craford, *Appl. Phys. Lett.* **57** (1990) 2937 – 2939.
26. H. Sugawara, M. Ishikawa, G. Hatakoshi, *Appl. Phys. Lett.* **58** (1991) 1010 – 1012.
27. S. Nakamura, M. Senoh, N. Iwasa, S. Nagahama, T. Yamaka, T. Mukai, *J Appl. Phys.* **34** (1995) L1332 – L1335.

28. E. W. Forsythe, A. J. Campbell (2001), SID Seminar Lecture Notes, M-10/1-58.
29. C. W. Tang, S. A. Van Slyke, *Appl. Phys. Lett.* **151** (1987) 913 – 915.
30. J. H. Burroughes, D. D. C. Bradley, A. R. Brown, R. N. Marks, K. MacKay, R. H. Fiend, P. C. Burn, A. B. Holmes, *Nature* **347** (1990) 539 – 543.
31. C. W. Tang (1996), SID '96 Digest, 181 – 184.
32. Y. A. Ono (1995), *Electroluminescent Displays*, World Scientific Publishing, Singapore.
33. R. T. Tuenge, J. Kane (1991), SID '91 Digest, 279 – 281.
34. G. Härkönen, K. Härkönen, R. Trönquist, (1990) SID 90 Digert, 232 – 5.
35. W. A. Barrow, R. C. Coovert, E. Dickey, C. N. King, C. Laakso, S. S. Sun, R. Tuenge, R. Wentross, (1993) SID 93 Digert, 761 – 764.
36. K. O. Velthaus, U. Troppenz, B. Hüttl, R. Herrmann, R. H. Mauch (1994), Intl. Display Research Conf., 346 – 348.
37. E. Soininen, G. Härkönen, M. Lahonen, R. Törnqvist (1999), Eurodisplay '99, 147 – 150.
38. T. Yamazaki, H. Kawakami, H. Hori (1996) in SEMI Standard FPD Technology Group (ed.) Color TFT Liquid Crystal Displays, Mountain View,CA: Semiconductor Equipment and Materials International.
39. P. Alt, P. Pleshko, *IEEE Trans. Electron Devices* **21** (1974) 146 – 155.
40. T. J. Scheffer, B. Clifton (1992), SID '92 Digest, 228 – 231.
41. S. Ihara, Y. Sugimoto, Y. Nakagawa, T. Kuweta, H. Koh, H. Hasebe, T. N. Ruckmongathan (1992), SID '92 Digest, 232 – 235.
42. S.-T. Wu, D.-K. Yang, *Reflective Liquid Crystal Displays,* John Wiley & Sons, Inc., 2001.
43. K. Yang, L. C. Chien, J. W. Doane (1991) Intl. Display Research Conf., 49 52.
44. M. Pfeiffer, D. K. Yang, J. W. Doane (1995), SID '95 Digest, 706 – 709.
45. J. W. Doane, M. E. Stefanov, *Information Display* **12** (1996) no. 12, 18 – 21.
46. T. J. Bunning, L. V. Natarajan, V. P. Tondiglia, R. L. Sutherland (2001), Proc. Asia Display, IDW '01, 37 – 40.
47. A. Mosley in L. W. MacDonald, A. C. Lowe (eds.): *Display Systems,* John Wiley & Sons, Inc., 1997, 173 – 189.
48. Y. Yamada, N. Yamamoto, K. Nakamura, N. Koshobu (1995) SID '95 Digest, 789 – 792.
49. F. Okumura (2001), SID Seminar Lecture Notes, M-3/1-37.
50. J. Chen, P. J. Bos, D. R. Bryant, D. L. Johnson, S. H. Jamal, J. R. Kelly (1995), SID '95 Digest, 865 – 868.
51. M. S. Nam, J. W. Wu, Y. J. Choi, K. H. Yoon, J. H. Jung, J. Y. Kim, K. J. Kim, J. H. Kim, S. B. Kwon (1997), SID '97 Digest, 933 – 936.
52. A. Lien, R. A. John (1993), Eurodisplay '93, 21 – 24.
53. A. Takeda, S. Kataoka, T. Sasaki, H. Chida, H. Tsuda, K. Ohmaro, Y. Koike, T. Sasabayashi, K. Okamoto (1998), SID '98 Digest, 1077 – 1080.
54. R. Kiefer, B. Weber, F. Windscheid, G. Baur, *J. Displ.* **92** (1992) 547 – 550.
55. K. Kondo, K. Kinugawa, N. Konishi, H. Kawakami (1996), SID '96 Digest, 82 – 84.
56. C. L. Kuo, T. Miyashita, M. Suzuki, T. Uchida (1994), SID '94 Digest, 927 – 930.
57. T. Buzak (1990), SID '90 Digest, 420 – 423.
58. T. Buzak (1999), Proc. Eurodisplay '99, 19 – 25.
59. T. Kakizaki, S. Tanamachi, M. Hayashi (1996), SID '96 Digest, 232 – 235.
60. A. L. Dalisa (1980), in Topics in Applied Physics **40**, Pankove J I (ed.), 213 – 232.
61. B. Comiskey, J. D. Albert, J. Jacobson (1997), SID '97 Digest, 75 – 76.
62. M. E. Howard, E. A. Richley, R. Sprague, N. K. Sheridon (1998), SID '98 Digest, 1010 – 1013.
63. J. P. Salerno, D. P. Vu, B. D. Dingle, M. W. Batty, A. C. Ipri, R. G. Steward, D. L. Jose, M. L. Tilton (1992), SID '92 Digest, 63 – 66.
64. Y. Itoh, J.-I. Nakamura, K. Yoneno, H. Kamakura, N. Okamoto (1997), SID '97 Digest, 993 – 996.
65. S.-G. Kim, K.-H. Hwang, J. Hwang, M.-K. Koo, G.-W. Lee (1998) Proc. Asia Display '98, 329 – 334.
66. D. T. Amm, R. W. Corrigan (1998), SID '98 Digest, 29 – 32.
67. J. B. Sampsell (1994), SID '94 Digest, 669 – 672.
68. L. J. Hornbeck, *Proc. SPIE, Proj. Displays V* **3634** (1999) 158 – 170.
69. A. Vecht (2001), SID Seminar Lecture Notes, M-9/1-63.
70. D. Seale, X. Wu (1999), Proc IDW'99, 861 – 864.
71. Y. Ishii, M. Hijikigawa (2001), Proc. Asia Display, IDW '01, 457 – 460.
72. J. L. Fergason (1985), SID '85 Digest, 68 – 69.

Dithiocarbamic Acid and Derivatives

RÜDIGER SCHUBART, Bayer AG, Leverkusen, Federal Republic of Germany

1.	**Introduction**	1
2.	**Physical and Chemical Properties**	2
3.	**Manufacture of Dithiocarbamate Salts**	2
4.	**Transformation Products**	3
4.1.	**Thiuram Sulfides**	3
4.1.1.	Thiuram Disulfides	3
4.1.2.	Thiuram Monosulfides	6
4.1.3.	Thiuram Trisulfides and Tetrasulfides	6
4.2.	**Thiocarbamoylsulfenamides**	7
4.3.	**Thiocarbamoyl Chlorides**	8
4.4.	**Dithiocarbamic Acid Esters**	8
4.5.	**Heterocyclic Compounds**	11
4.6.	**Other Derivatives**	13
4.6.1.	Dithiolanylium Salts	13
4.6.2.	Alkylthioformimidic Chlorides	13
4.6.3.	Isothiocyanates, Isocyanates, and Thioureas	13
4.7.	**Photolysis of Dithiocarbamates**	14
5.	**Uses**	14
5.1.	**Vulcanization Accelerators**	14
5.2.	**Pesticides**	14
5.3.	**Medical Applications**	14
5.4.	**Radioprotective Agents**	20
5.5.	**Imaging Technology**	20
5.6.	**Other Uses**	20
6.	**Toxicology and Occupational Health**	21
7.	**References**	21

1. Introduction

Dithiocarbamic acids are monoamides of dithiocarbonic acid. As a rule, one or both of the hydrogen atoms on the nitrogen atom are replaced by alkyl, aralkyl, cycloalkyl, or aryl radicals; the amide nitrogen atom may also belong to a ring:

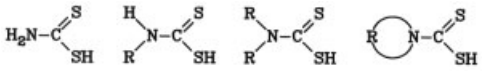

Acyl- and thioacyldithiocarbamic acids have also been described:

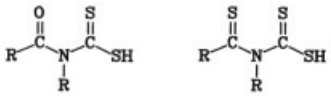

Industrially important derivatives of dithiocarbamic acids are ammonium and metal salts, esters, thiuram sulfides, thiocarbamoylsulfenamides, and a number of other products. They are used mainly as vulcanization accelerators in the rubber industry and as fungicides in agriculture.

2. Physical and Chemical Properties

Free dithiocarbamic acid is unstable in the solid state (*mp* 35.7 °C) and in solution. Thermal decomposition results in carbon disulfide, hydrogen sulfide, and ammonium thiocyanate [11].

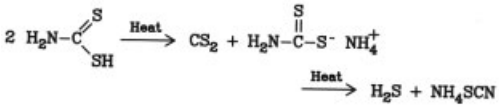

Free *N*-substituted dithiocarbamic acids are also unstable. However, diphenyldithiocarbamic acid [12] is stable, as are several acyldithiocarbamic acids, such as formyl- [13], 2-ethoxybenzoyl- [14], and furoyldithiocarbamic acid [14], as well as 2-pyrrolidone-1-dithiocarbamic acid [15] (Table 1); these acids can be produced from the alkali salts by acidification with HCl gas in solvents such as ether. *N*-Thiobenzoyldithiocarbamic acid cannot be obtained by this method [16].

Salts of dithiocarbamic acids are more stable than the free acids [17]. The physical properties of some alkali-metal, heavy-metal, and ammonium dithiocarbamates are listed in Tables 2, 3, 4. Salts of unsubstituted and *N*-monosubstituted

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_001

Table 1. Physical properties of dithiocarbamic acids

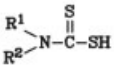

R^1	R^2	CAS registry number	Physical form	Molecular formula	M_r	*mp*, °C
H–	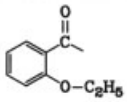	[*79388-79-7*]	yellow crystals	$C_{10}H_{11}NO_2S_2$	241.3	122–123
H–		[*79388-80-0*]	orange crystals	$C_6H_5NO_2S_2$	187.2	77–79
C_6H_5–	C_6H_5–	[*7283-79-6*]		$C_{13}H_{11}NS_2$	245.36	147
H–		[*100647-03-8*]	bright yellow	$C_2H_3NOS_2$	121.18	66–67
		[*71408-15-6*]	yellow crystals	$C_5H_7NOS_2$	161.25	101–102

dithiocarbamic acids decompose on heating and are sensitive to oxidation (see Section 4.6.3). Derivatives of the tautomeric imide form of N-monosubstituted dithiocarbamic acids are also known [14].

[18] , [19], [20], [21], [22]

Salts and derivatives of *N,N*-disubstituted dithiocarbamic acids are very stable. These salts (especially the heavy-metal salts) are strongly colored; they are often volatile and soluble in organic solvents. For example, the copper, nickel, and cobalt salts of *N,N*-diisobutyldithiocarbamic acid can be distilled at low pressure. These compounds are apparently internal complexes.

Reactions of dithiocarbamate salts to form esters, thiuram sulfides, and other derivatives are discussed in detail in Chapter 4. All dithiocarbamic acids and their derivatives are cleaved by mineral acids to carbon disulfide and ammonium salts of the mineral acid. This reaction can be used to determine the structure of unknown derivatives [23].

3. Manufacture of Dithiocarbamate Salts

The parent compound in this series, ammonium dithiocarbamate [*513-74-6*], is prepared most easily from ammonia and carbon disulfide in alcohol or an ester solvent [24]. *N*-Alkyldithiocarbamic acids can be obtained as ammonium salts from two moles of a primary amine and one mole of carbon disulfide:

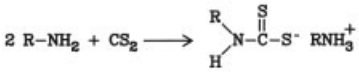

If the reaction is carried out in the presence of an alkali or alkaline-earth hydroxide, the corresponding metal salts are formed. To prepare *N*-aryldithiocarbamates a strong base is required; otherwise *N,N*-diarylthioureas are formed. Heavy-metal dithiocarbamates are obtained from ammonium or alkali dithiocarbamates and the corresponding heavy-metal salts in aqueous solution, or directly by reaction of an aromatic amine and carbon disulfide in the presence of finely divided metal oxide [25]. The deeply colored salts of formyldithiocarbamic

Table 2. Physical properties of ammonium salts of dithiocarbamic acids

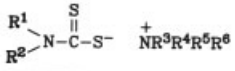

R^1	R^2	R^3	R^4	R^5	R^6	CAS registry number	Physical form	Molecular formula	M_r	*d*	*mp*, °C	Ref.
CH_3–	CH_3–	CH_3–	CH_3–	H–	H–	[*598-64-1*]	colorless crystals	$C_5H_{14}N_2S_2$	166.3		115–120	
C_2H_5–	C_2H_5–	C_2H_5–	C_2H_5–	H–	H–	[*1518-58-7*]	colorless crystals	$C_9H_{22}N_2S_2$	222.4		88	
C_2H_5–	C_6H_{11}–	C_2H_5–	C_6H_{11}–	H–	H–	[*13167-44-7*]	yellowish powder	$C_{17}H_{34}N_2S_2$	360.6	1.08	92	
				H–	H–	[*98-77-1*]	yellowish powder	$C_{11}H_{22}N_2S_2$	246.45	1.15	179	
H–	HCO–	H–	H–	H–	H–	[*100647-07-2*]		$C_2H_6N_2OS_2$	138.22	1.48	114 (decomp.)	a
H–	HCO–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	[*100647-09-4*]	yellow	$C_{18}H_{38}N_2OS_2$	362.63	1.06	79–81	a
CH_3–	HCO–	H–	H–	H–	H–	[*102564-66-9*]		$C_3H_8N_2OS_2$	152.24		87–88	b
H–	HCS–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	[*97960-84-4*]	brilliant red	$C_{18}H_{38}N_2S_3$	378.69	1.06	90.5	c
CH_3–	HCS–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	*n*-C_4H_9–	[*101901-15-9*]	blackish red	$C_{19}H_{40}N_2S_3$	392.87	1.09	89–91 (decomp.)	d
H–	HCS–	H–	H–	H–	H–	[*97960-89-9*]	orange	$C_2H_6N_2S_3$	122.22		73–74 (decomp.)	e
C_6H_5–	C_6H_5–	H–	H–	H–	HO–		yellow	$C_{13}H_{14}N_2OS_2$	278.39		65 (decomp.)	e

Table 3. Metal salts of ethylenebis(dithiocarbamic acids)

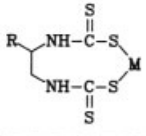

R	M	CAS registry number	Physical form	*mp*, °C
H–	2 Na	[*142-59-6*]	white crystals	82
H–	Zn	[*12122-67-7*]	light-colored powder, polymer	decomp. on heating
H–	Mn	[*12427-38-2*]	yellow powder, polymer	decomp. on heating
CH_3–	Zn	[*12071-83-9*]	pale yellow powder, polymer	decomp. above 160 °C

Table 4. Physical properties of metal salts of dithiocarbamic acids

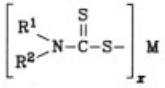

R^1	R^2	M	x	CAS registry number	Form	Molecular formula	M_r	d	mp, °C
C_2H_5-	C_2H_5-	Na	1	[*148-18-5*]	yellow crystals	$C_5H_{10}NNaS_2$	171.27		97–98
CH_3-	H–	Na	1	[*137-42-8*]	colorless crystals	$C_2H_4NNaS_2$	129.19		decomp.
CH_3-	CH_3-	Zn	2	[*137-30-4*]	white powder	$C_6H_{12}N_2S_4Zn$	305.82	1.65	248–250
CH_3-	CH_3-	Fe	3	[*14484-64-1*]	black powder	$C_9H_{18}FeN_3S_6$	416.51		180 (decomp.)
C_2H_5-	C_2H_5-	Zn	2	[*14324-55-1*]	white powder	$C_{10}H_{20}N_2S_4Zn$	361.93	1.47	178–180
n-C_3H_7-	n-C_3H_7-	Zn	2	[*15694-56-1*]	white powder	$C_{14}H_{28}N_2S_4Zn$	418.03	1.35	107–108
n-C_4H_9-	n-C_4H_9-	Zn	2	[*136-23-2*]	white powder	$C_{18}H_{36}N_2S_4Zn$	474.15	1.24	105–108
n-$C_5H_{11}-$	n-$C_5H_{11}-$	Zn	2	[*15337-18-5*]	yellowish powder	$C_{22}H_{44}N_2S_4Zn$	530.25	1.14	65.5
		Zn	2	[*13878-54-1*]	white powder	$C_{12}H_{20}N_2S_4Zn$	385.95	1.55	222–225
$C_6H_5CH_2-$	$C_6H_5CH_2-$	Zn	2	[*14726-36-4*]	white powder	$C_{30}H_{28}N_2S_4Zn$	610.21	1.41	165–175
C_6H_5-	C_2H_5-	Zn	2	[*14634-93-6*]	light-colored	$C_{18}H_{20}N_2S_4Zn$	458.02	1.46	205
CH_3-	CH_3-	Pb	2	[*19010-66-3*]	pale gray powder	$C_6H_{12}N_2PbS_4$	447.65	2.38	320–323
C_2H_5-	C_2H_5-	Pb	2	[*17549-30-3*]	pale gray powder	$C_{10}H_{20}N_2PbS_4$	503.76	1.87	206–207
n-C_3H_7-	n-C_3H_7-	Pb	2	[*70995-63-0*]	grayish brown powder	$C_{14}H_{28}N_2PbS_4$	559.87	1.65	97–98
n-C_4H_9-	n-C_4H_9-	Pb	2	[*27803-77-6*]	light-colored powder	$C_{18}H_{36}N_2PbS_4$	615.98	1.42	72–73
n-$C_5H_{11}-$	n-$C_5H_{11}-$	Pb	2	[*36501-84-5*]	pale yellow liquid	$C_{22}H_{44}N_2PbS_4$	672.78	1.40	–
		Pb	2	[*41556-46-1*]	pale gray powder	$C_{12}H_{20}N_2PbS_4$	527.78	2.79	249–250
		Cd	2	[*14949-59-8*]	yellow powder	$C_{12}H_{20}CdN_2S_4$	432.98	1.82	240–245
n-C_4H_9-	n-C_4H_9-	Ni	2	[*13927-77-0*]	green powder	$C_{18}H_{36}N_2NiS_4$	467.48	1.29	~88
CH_3-	CH_3-	Cu	2	[*137-29-1*], [*23726-35-4*]	pale brown powder	$C_6H_{12}CuN_2S_4$	303.98	1.7	360
C_2H_5-	C_2H_5-	Cu	2	[*13681-87-3*]	dark brown powder	$C_{10}H_{20}CuN_2S_4$	360.09	1.49	196–197
n-C_3H_7-	n-C_3H_7-	Cu	2	[*14354-08-6*], [*2102-05-8*]	black powder	$C_{14}H_{28}CuN_2S_4$	416.20	1.24	100–101
n-C_4H_9-	n-C_4H_9-	Cu	2	[*52691-95-9*], [*13927-71-4*]	black powder	$C_{18}H_{36}CuN_2S_4$	472.31	1.26	77–78
n-$C_5H_{11}-$	n-$C_5H_{11}-$	Cu	2	[*36190-66-6*]	black powder	$C_{22}H_{44}CuN_2S_4$	528.41	1.15	50
CH_3-	CH_3-	Se	4	[*21559-13-7*], [*144-34-3*]	yellow powder	$C_{12}H_{24}N_4S_8Se$	559.84	1.57	146
C_2H_5-	C_2H_5-	Se	4	[*105618-13-1*], [*21559-14-8*]	dark yellow powder	$C_{20}H_{40}N_4S_8Se$	672.06	1.32	62–86
n-C_3H_7-	n-C_3H_7-	Se	4		orange powder	$C_{28}H_{56}N_4S_8Se$	784.28	1.23	45
n-C_4H_9-	n-C_4H_9-	Se	4		dark red liquid	$C_{36}H_{72}N_4S_8Se$	896.49	1.11	–
n-$C_5H_{11}-$	n-$C_5H_{11}-$	Se	4		dark red liquid	$C_{44}H_{88}N_4S_8Se$	1008.71	1.03	–
C_2H_5-	C_2H_5-	Bi	3	[*20673-31-8*], [*96126-13-5*]	yellow powder	$C_{15}H_{30}BiN_3S_6$	653.83	2.06	230 (decomp.)
H–	HCO–	K	1	[*97288-11-4*]	yellow	$C_2H_2KNOS_2$	159.27	1.83	186 (decomp.)
CH_3-	HCO–	Na	1	[*102118-28-5*]	yellow	$C_3H_4NNaOS_2$	157.2		161–162
CH_3-	HCS–	K	1	[*102137-68-9*]	pale red	$C_3H_4KNS_3$	189.36	1.57	94–95
H–	HCS–	K	1	[*97960-86-6*]	orange–yellow	$C_2H_2KNS_3$	175.34		142

acid are produced from potassium formyldithiocarbamate [18].

Amine and ammonium salts of *N,N*-disubstituted dithiocarbamic acids are produced from carbon disulfide by reaction with two equivalents of amine or one each of amine and ammonia. An example is the piperidinium salt of pentamethylenedithiocarbamic acid, which is used as a vulcanization accelerator.

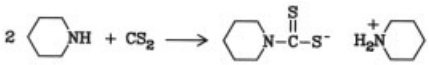

Guanidine and arylguanidine react analogously [26]; hydroxylamine derivatives have also been described [27].

Alkali and heavy-metal salts of *N,N*-disubstituted dithiocarbamic acids are produced like the monosubstituted salts [28], [29].

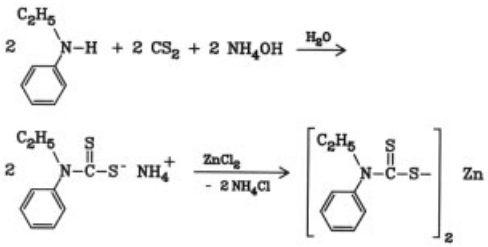

Diamines react with two moles of carbon disulfide to yield bis(dithiocarbamates). Zinc, iron, and copper ethylenebis(dithiocarbamates) are used as fungicides and nematocides [30]. With one mole of carbon disulfide, diamines form inner salts or mixtures of inner salts and amine salts of bis(dithiocarbamic acids). Such mixtures are produced by piperazines [31].

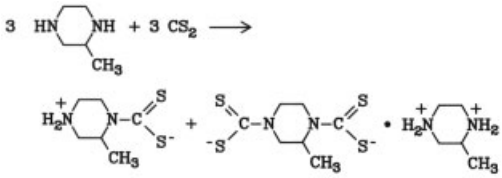

Hydrazine reacts with one mole of carbon disulfide in the presence of alkali hydroxide to form alkali salts of aminodithiocarbamic acid (dithiocarbazic acid) [32]:

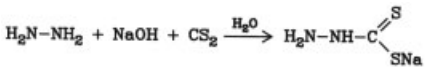

With two moles of carbon disulfide, salts derived from the tautomeric imide form are obtained [33]:

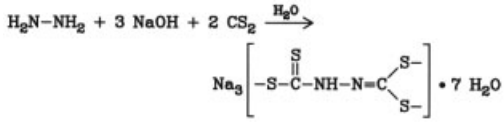

Metal salts, particularly zinc dithiocarbamates, are used as vulcanization accelerators, oil additives, bactericides, and fungicides.

4. Transformation Products

4.1. Thiuram Sulfides

4.1.1. Thiuram Disulfides

The oxidation of salts of dithiocarbamic acids usually gives thiuram disulfides (bis(aminothiocarbonyl)disulfanes) [34–36]. Table 5 lists physical properties of some thiuram sulfides. Only the tetraalkyl derivatives are stable. Oxidizing agents include iodine, bromine, hydrogen peroxide, and potassium peroxodisulfate [37]; ammonium peroxodisulfate [38]; sodium tetrathionate [39]; and chlorine or bromine in the presence of buffers [40]. The colorless compound tetramethylthiuram disulfide, for example, is obtained from sodium dimethylthiocarbamate and hydrogen peroxide [41]:

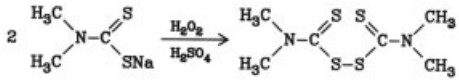

Other methods use hypochlorites [42], sodium nitrite [43], oxygen in the presence of catalysts such as copper or manganese salts [44] (see also Section 4.6.3), or electrolytic oxidation [45]. Tetraalkylthiuram disulfides are powerful vulcanization accelerators [46].

N-Acyl derivatives are also known. For example, oxidation of *N*-methyl-*N*-formyldithiocarbamate with iodine in ether yields the corresponding thiuram [47]:

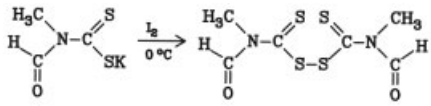

N-Methyl-*N*-thioformyldithiocarbamate gives a cyclic, pale yellow thiuram (*mp* 108.5 °C) [48]:

Table 5. Physical properties of thiuram sulfides

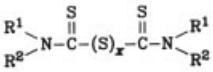

R^1	R^2	CAS registry number	Physical form	Molecular formula	M_r	d	mp, °C
Monosulfides ($x = 1$)							
CH_3-	CH_3-	[*97-74-5*]	yellow powder	$C_6H_{12}N_2S_3$	208.38	1.39	107
C_2H_5-	C_2H_5-	[*95-05-6*]	brown liquid	$C_{10}H_{20}N_2S_3$	264.48	1.13	28.5–29.8
n-C_3H_7-	n-C_3H_7-	[*2556-41-4*]	brown liquid	$C_{14}H_{28}N_2S_3$	320.59	1.04	
n-C_4H_9-	n-C_4H_9-	[*93-73-2*]	dark red liquid	$C_{18}H_{36}N_2S_3$	376.70	0.99	
n-$C_5H_{11}-$	n-$C_5H_{11}-$	[*6405-88-5*]	dark red liquid	$C_{22}H_{44}N_2S_3$	432.81	0.96	
$H-$	C_2H_4-	[*3082-38-0*]	monomer	$C_4H_6N_2S_3$	178.31		125–126
$H-$	C_2H_4-	[*27321-95-5*]	polymer	$(C_4H_6N_2S_3)_n$	$(178.31)_n$		145–147
Disulfides ($x = 2$)							
CH_3-	CH_3-	[*137-26-8*]	white powder	$C_6H_{12}N_2S_4$	240.44	1.29	146–148
C_2H_5-	C_2H_5-	[*97-77-8*]	almost colorless powder	$C_{10}H_{20}N_2S_4$	296.55	1.17	65–70
n-C_3H_7-	n-C_3H_7-	[*2556-42-5*]	pale yellowish powder	$C_{14}H_{28}N_2S_4$	352.66	1.09	49.5–50.5
n-C_4H_9-	n-C_4H_9-	[*1634-02-2*]	dark red liquid	$C_{18}H_{36}N_2S_4$	408.77	1.03	
n-$C_5H_{11}-$	n-$C_5H_{11}-$	[*5721-31-3*]	dark red liquid	$C_{22}H_{44}N_2S_4$	464.87	0.99	
CH_3-	C_6H_5-	[*10591-84-1*]	white powder	$C_{16}H_{16}N_2S_4$	364.58	1.33	206
CH_3-	$HCO-$	[*102127-57-1*]	pale yellow	$C_6H_8N_2O_2S_4$	268.40	1.088	69–70
$H-$	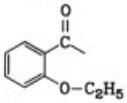	[*79388-81-1*]	yellow crystals	$C_{20}H_{20}N_2O_4S_4$	480.6		137–138
Tetrasulfides ($x = 4$)							
CH_3-	CH_3-	[*97-91-6*]	pale yellowish powder	$C_6H_{12}N_2S_6$	304.57	1.52	~90
CH_3-	C_6H_5-	[*103597-02-0*]	pale yellowish powder	$C_{16}H_{16}N_2S_6$	428.72		117
C_2H_5-	C_6H_5-		pale yellowish powder	$C_{18}H_{20}N_2S_6$	456.77		142
		[*120-54-7*]	white powder	$C_{12}H_{20}N_2S_6$	384.70	1.41	117
		[*103597-04-2*]	white powder	$C_{10}H_{16}N_2O_2S_6$	388.65		111.5–114.5

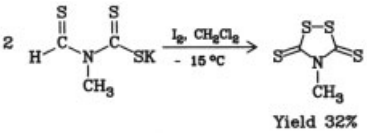

N-Thioformyldithiocarbamate, however, yields 1,2,4-dithiazole-3-thione (*mp* 73–75 °C) [47]:

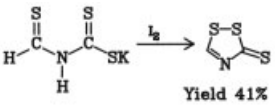

Oxidation of *N*-(2-ethoxybenzoyl)dithiocarbamate gives the corresponding disulfide (*mp* 138 °C) which, when heated in the presence of pyridine, yields 2-ethoxybenzoyl disulfide (*mp* 137–139 °C) [14].

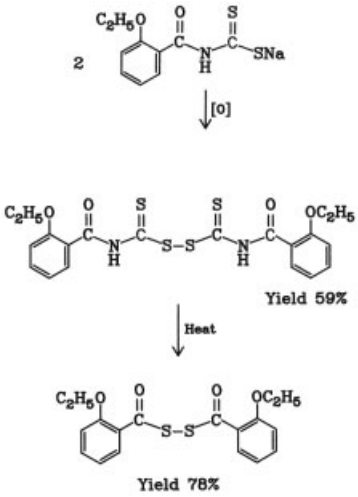

4.1.2. Thiuram Monosulfides

Thiuram disulfides may lose one sulfur atom to form thiuram monosulfides (bis(aminothiocarbonyl)sulfanes), which are also used as vulcanization accelerators [46].

Tetramethylthiuram disulfide gives the yellow monosulfide on heating with triphenylphosphine [49] or potassium cyanide [50] (Fig. 1, Reaction 1). The monosulfide can also be obtained directly from sodium dimethyldithiocarbamate by reaction with phosgene [51], [52] or cyanogen chloride [51], [53] (Reaction 2). Another method uses the ammonium dithiocarbamate and thiocarbamoyl chloride [54] (Reaction 3).

4.1.3. Thiuram Trisulfides and Tetrasulfides

Sulfur dichloride reacts with alkali dithiocarbamates to form thiuram trisulfides (bis(aminothiocarbonyl)trisulfanes). Disulfur dichloride gives the corresponding tetrasulfides, e.g., the tetrasulfide of pentamethylenedithiocarbamic acid, an important vulcanization accelerator [55]:

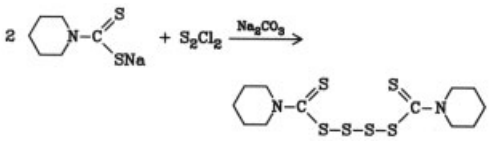

4.2. Thiocarbamoylsulfenamides

Thiocarbamoylsulfenamides are obtained from *N,N*-disubstituted dithiocarbamates by mild oxidation with iodine in the presence of primary or secondary amines [56] or by electrolytic oxidation [57]:

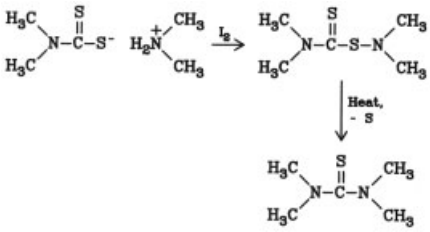

These compounds are used as vulcanization accelerators, fungicides [58], and bactericides [59]. Thermolysis removes one atom of sulfur to give thioureas. The morpholine derivative is manufactured by oxidation with sodium hypochlorite [60–69].

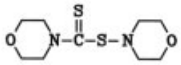

Substituted thiocarbamoylsulfenamides are obtained by reaction with isocyanates [70]:

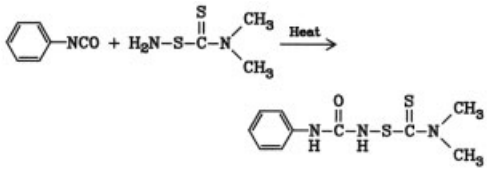

Polymeric thiocarbamoylsulfenamides have also been reported [71].

4.3. Thiocarbamoyl Chlorides

Thiuram disulfides react with chlorine in organic solvents to give disubstituted thiocarbamoyl chlorides, which are useful intermediates [72], [73].

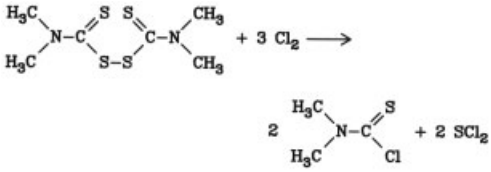

The methyl groups can be replaced by ethyl, propyl, butyl, or phenyl groups; the nitrogen can also be part of a piperidine ring [74]. Alternatively, thiocarbamoyl chlorides can be obtained from amines and thiophosgene [75] or by chlorination of thioformamides [76]. Further chlorination leads to iminium salts [77]:

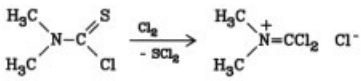

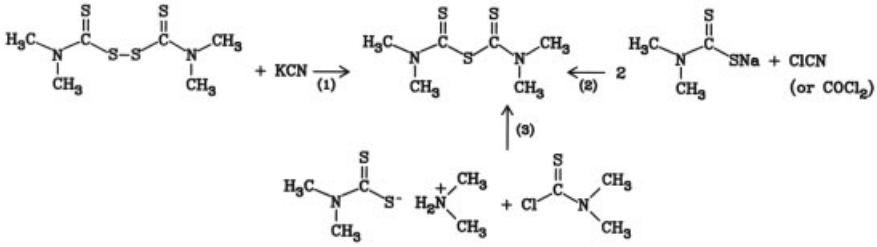

Figure 1. Syntheses of tetramethylthiuram monosulfide

4.4. Dithiocarbamic Acid Esters

Dithiocarbamic acid esters (dithiourethanes) are used in agriculture as nematocides, fungicides, and herbicides. Physical properties of some dithiocarbamates are given in Table 6. The most important methods of preparation are (1) alkylation of alkali dithiocarbamates with halogen compounds or other alkylating agents and (2) insertion reactions of carbon disulfide with tertiary amines.

Suitable halogen compounds include alkyl chlorides, benzyl chloride [78], phenacyl chloride [79], acyloxymethyl chloride [80], and allyl chlorides [81]. For example, the reaction of sodium *N*,*N*-dimethyldithiocarbamate with 1,3-dichloro-2-butene gives dimethyldithiocarbamic acid 3-chloro-2-butenyl ester, which is used as a herbicide:

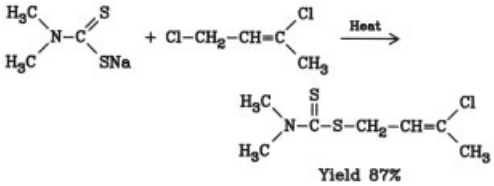

Other alkylating agents are (1) halogenated aldehydes, ketones, and chlorohydrin [78]; (2) dihalogen compounds such as benzal chloride [82] and 1,3-dichloropropane [83]; (3) trihalogen compounds such as cyanuric chloride [84] and chloroform [85]; and (4) polychlorinated compounds [86]. The reaction of an alkali dithiocarbamate with chloroform, for example, yields an amino diester and carbon disulfide:

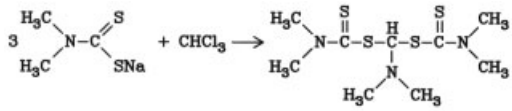

High yields are also obtained from alkali acyldithiocarbamates and alkyl halides [48], [87], [88]:

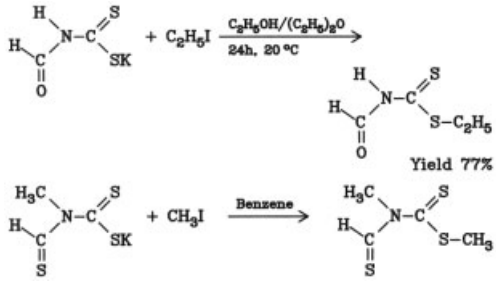

Arylation is possible with strongly activated halobenzenes such as 2,4-dinitrochlorobenzene [82] or with diazonium salts [89]. Phenyl esters are also produced by reaction of phenyldithioformic acid chloride with amines [90]. *o*-Phenylene diesters are obtained from *o*-nitrophenyl monoesters by substitution of the nitro group [91].

Sulfenyl halides [92] and thiosulfates [93] react similarly with alkali dithiocarbamates to form thiocarbamoylalkyldisulfanes:

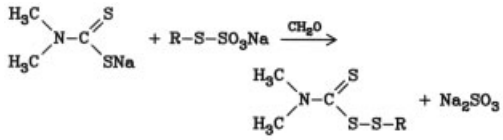

Dithiocarbamic acid esters are also obtained by reaction of thiocarbamoyl chlorides with thiophosphates [94] or by addition of thiols to isothiocyanates [95], [96].

Table 6. Physical properties of dithiocarbamic acid esters

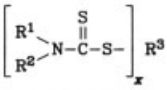

R^1	R^2	R^3	x	CAS registry number	Physical form	Molecular formula	M_r	d	mp, °C
CH_3–	CH_3–	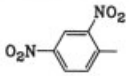	1	[*89-37-2*]	yellow powder	$C_9H_9N_3O_4S_2$	287.33	1.75	140–145
		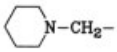	1	[*10254-56-5*]	brown crystals	$C_{12}H_{22}N_2S_2$	258.46	1.11	59
CH_3–	CH_3–		2	[*49773-60-6*]	white powder	$C_{13}H_{18}N_2S_4$	330.57		180–182
CH_3–	CH_3–	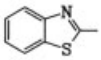	1	[*3432-25-5*]	yellowish powder	$C_{10}H_{10}N_2S_3$	254.4		106–112
C_2H_5–	C_2H_5–	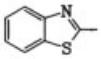	1	[*95-30-7*]	yellowish powder	$C_{12}H_{14}N_2S_3$	282.46	1.27	70–73
C_2H_5–	C_2H_5–	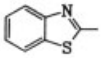	1		yellowish powder	$C_{16}H_{14}N_2S_3$	330.50		105–108
		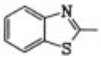	1	[*86443-84-7*]	yellowish powder	$C_{13}H_{14}N_2S_3$	294.47		133–137
CH_3–	HC(=O)–	C_2H_5–	1	[*102127-59-3*]	pale yellow	$C_5H_9NOS_2$	163.26	1.179	73.5
H–	HC(=O)–	C_2H_5–	1	[*102933-65-3*]	white–yellow	$C_4H_7NOS_2$	149.24	1.292	77–81
H–	HC(=S)–	C_2H_5–	1	[*102118-26-3*]	orange	$C_4H_7NS_3$	165.30	1.33	67.5
H–	H_2N–	CH_3–	1	[*5397-03-5*]	yellowish	$C_2H_6N_2S_2$	122.22		

* Boiling point at 0.01 mm Hg (1.3 Pa).

Vinyl esters can be prepared from phenolates [97]:

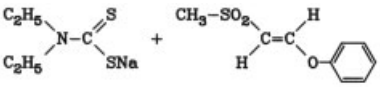

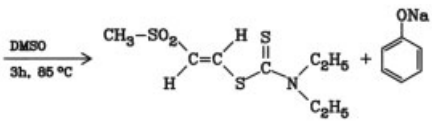

Alkylation of ammonium dithiocarbamates with epoxides is also possible [98]:

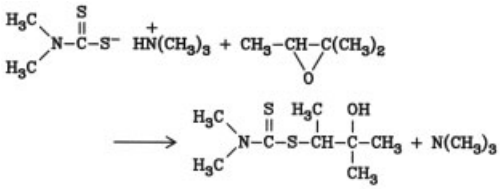

Styrene oxide and sodium diethyldithiocarbamate react similarly [99]. At low temperature, epichlorohydrin reacts like an epoxide to form the dithiocarbamic acid ester; when heated in the presence of a tertiary amine, this ester rearranges to the thiocarbamate [100]:

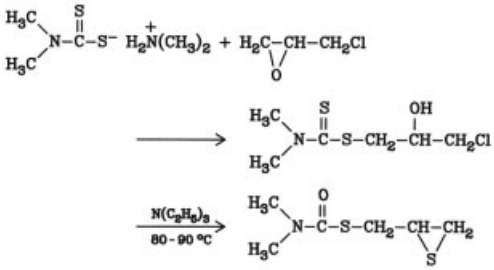

Whereas primary or secondary amines react with carbon disulfide to form mono- or disubstituted salts of dithiocarbamic acids (see Chap. 3), some tertiary amines such as aziridines form cyclic esters. For example, *N*-phenylthiazolidine-2-thione is obtained from *N*-phenylaziridine [101]:

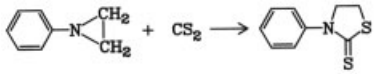

β-Amino ketones [102] and benzylamines [103] also react with insertion of carbon disulfide; for example, 2,6-di-*tert*-butyl-4-dialkylaminomethylphenols form esters that are used as stabilizers and intermediates [104]:

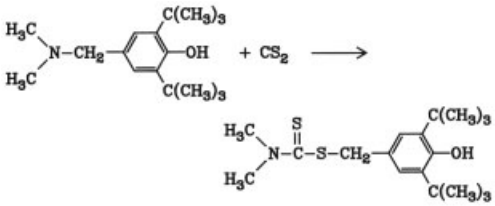

Silylamines also give insertion products, but no insertion is observed with silylaminophosphine [105]:

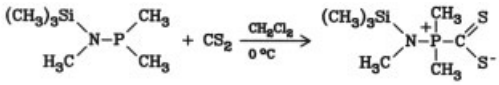

Methylenediamines, such as bis(piperidino)-methane, and carbon disulfide form aminomethyl dithiocarbamates [106]:

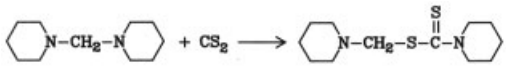

Piperidinomethyl carbamates can also be produced from methylolpiperidine and salts of dithiocarbamic acids [107]. Methylene bis(dithiocarbamates) are obtained from formaldehyde and dithiocarbamic acids in acidic solution [108]:

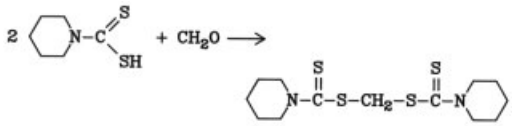

Bis(dithiocarbamates) are also formed by cleavage of ethylene trithiocarbonate with amines [4], [109]. Benzothiazolyl dithiocarbamates, which are used as accelerators for rubber, can be obtained in high yield from benzothiazole disulfide and thiuram disulfide in the presence of sodium cyanide [110]:

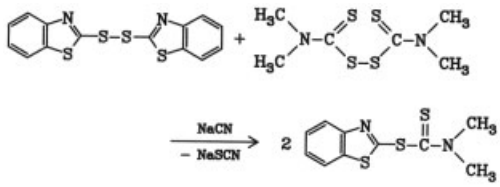

Dithiocarbamates are also prepared by addition of dithiocarbamic acids to unsaturated compounds such as vinylsulfonates [111], vinyl ether [112], acrylonitrile [113], vinylpyridine [114], mesityl oxide [115], acetylenedicarboxylate [116], acetylene ketones [117], and even acetylene [118]:

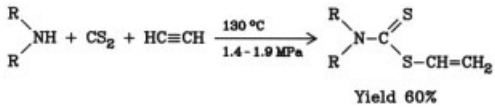

Unsaturated dithiocarbamates have been used in the stereoselective synthesis of polyenes [119]. The chlorination of monosubstituted dithiocarbamates leads to isocyanide dichlorides and removal of the sulfur [120], [121]:

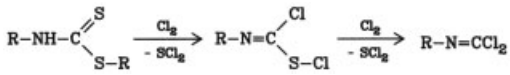

4.5. Heterocyclic Compounds

Five-membered Heterocycles. Dithiocarbamic acid and its salts react with

monochloroacetic acid to yield first the dithiocarbamic acid ester and then rhodanine [122], [123]:

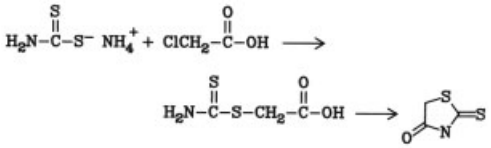

Rhodanine reacts with aldehydes [122], [124]; the products are stimulants [125]. Thiorhodanine (thiazolidinedithione) [126], *N*-aminorhodanine [127], and other substituted rhodanines [128], [129] have also been prepared. Certain rhodanines act as cross-linking agents [130].

Ammonium dithiocarbamate and α-halogenated ketones or aldehydes give 2-mercaptothiazoles; *N*-alkyldithiocarbamates form *N*-alkylthiazoline-2-thiones [129], [131]:

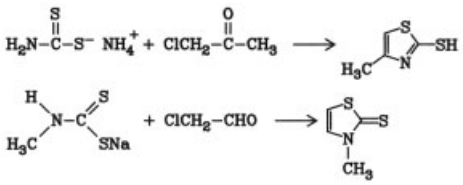

The reaction of dithiocarbamic acid methyl ester and chloroacetaldehyde leads to 2-methylthiothiazole, which is easily converted to thiazole with lithium in liquid ammonia [132]. Further thiazoline-2-thione syntheses have been described [133], [134].

Thiazolidine-2-thione is prepared by treating methyl thioxanthate with bromoethylamine [135]; substituted thiazolidine-2-thiones are obtained by heating aminoethanols with carbon disulfide [136]:

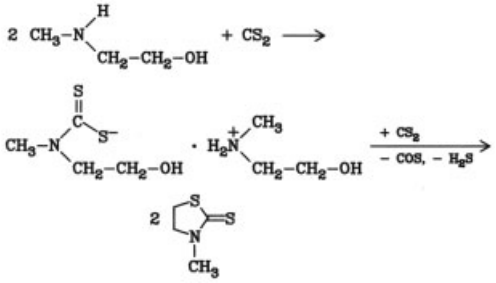

N-Methylthiazolidine-2-thione [137] is a good accelerator for polychloroprene.

Bis(carboxyaminoethyl)disulfane and carbon disulfide yield carboxythiazolidine-2-thione [138]:

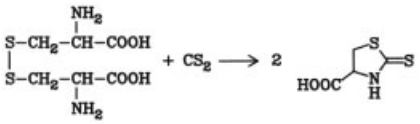

The reaction between acylaminodithiocarboxylic acid esters and aziridine leads to 2-acyliminothiazolidines [139]:

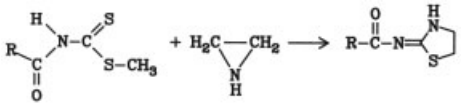

Thiazolotriazinium salts are prepared from dithiocarbazic acid methyl ester [140]. Aroylcarbimidodithioates easily cyclize with hydrazine to triazoles [14]:

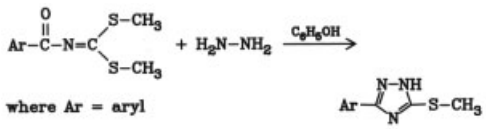

The cyclization of *N*-methylquinoxalinium iodide with dithiocarbamates gives tetrahydrothiazoloquinoxalines [141]:

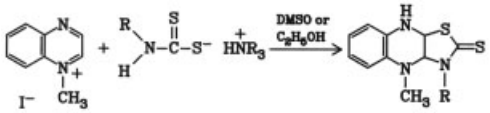

Six-membered Heterocycles. The reaction of *N*-methyldithiocarbamic acid with propiolic acid after several steps leads to 3-methyl-3,4-dihydro-4-oxo-2-thioxo-2*H*-1,3-thiazine [142]:

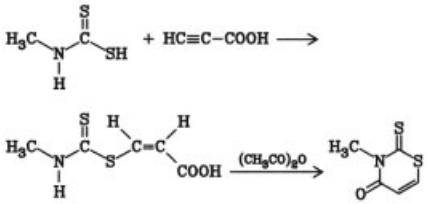

Heating this compound with primary amines gives 2-thiouracyls. 3*H*-1,3-Thiazine-2,6-

dithiones have also been reported [143]. Saturated six-membered rings are synthesized by reaction of dithiocarbamic acids with acrylonitrile [144] or propiolactone [145]. Acrylamides and carbon disulfide give 2-thioxoperhydro-1,3-thiazin-4-ones [146] and other derivatives [147]:

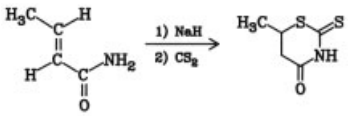

Oxidation removes the thione sulfur [148]:

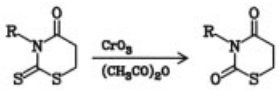

3-Amino-2-thioxotetrahydro-1,3-thiazin-4-ones are obtained from hydrazine or its derivatives [149]:

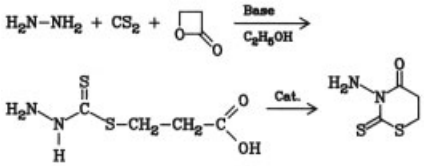

1,3-Thiazines can also be prepared from thioureas [150]:

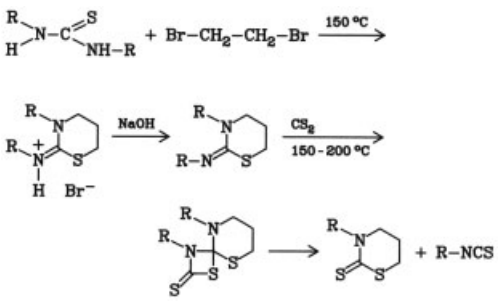

Reactions at the nitrogen atom of unsubstituted 1,3-thiazine-2-thione have been reported [151]. Mercaptothiazines are produced from ammonium dithiocarbamate and α,β-unsaturated ketones; the reaction of mesityl oxide with dithiocarbamic acid gives 2-mercapto-4,4,6-trimethyl-4*H*-1,3-thiazine [152]:

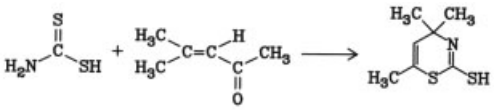

Thiadiazinethiones are produced in high yield by condensation of *N*-monosubstituted ammonium dithiocarbamates with aldehydes. For example, the reaction of methylammonium *N*-methyldithiocarbamate with formaldehyde in water gives 3,5-dimethyltetrahydro-2*H*-1,3,5-thiadiazine-2-thione, an accelerator for polychloroprene [176], in 90 % yield [153]:

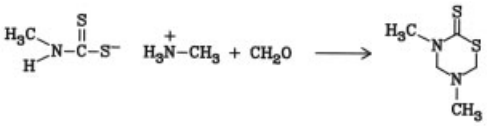

The corresponding *N*-benzyl derivative, known as sulbentine, is an antimycotic agent [154]. The reaction of methylamine and aminoacetic acid with carbon disulfide and formaldehyde yields 3-methyl-5-carboxymethyltetrahydro-2*H*-1,3,5-thiadiazine-2-thione, a nematocide [155].

4.6. Other Derivatives

4.6.1. Dithiolanylium Salts

2-Dialkylamino-1,3-dithiolanylium salts are prepared by reacting sodium *N*,*N*-dimethyldithiocarbamate with 1,2-dichloroethane in tetrahydrofuran in the presence of sodium perchlorate [156]:

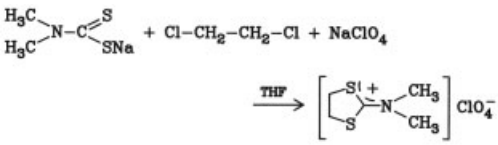

These compounds react with many nucleophiles [156], [157] (see Fig. 2). Unsaturated dithiolium salts can be used to prepare tetrathiafulvenes [158] and 1,2-dimercaptoalkenes [159].

4.6.2. Alkylthioformimidic Chlorides

Alkylthioformimidic chlorides are obtained from *N*,*N*-dialkyldithiocarbamic acid esters and phosgene [160]:

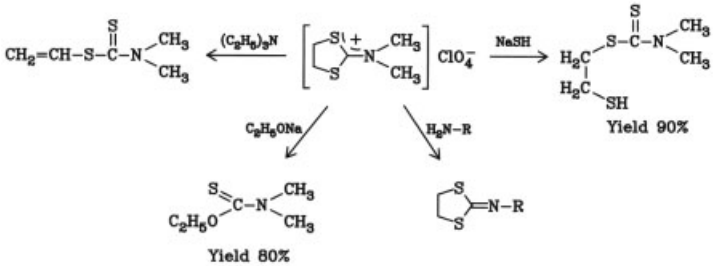

Figure 2. Reactions of 2-dialkylamino-1,3-dithiolanylium salts with nucleophiles

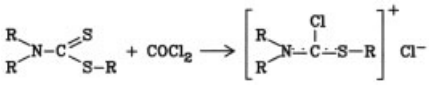

These salts are very reactive; for example, methylthioformimidic chloride and pyrrole give a 2-pyrrolyl derivative of methylthiocarboximidic chloride [161].

4.6.3. Isothiocyanates, Isocyanates, and Thioureas

Isothiocyanates (mustard oils) can be produced from monosubstituted salts or esters of dithiocarbamic acids by oxidation or heat (see also Section 4.1.1 and → Thiocyanates and Isothiocyanates, Organic) [162], [163]:

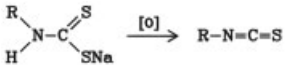

N-Acyl and *N*-thioacyl derivatives react differently (see Section 4.1.1).

Isocyanates are produced from monosubstituted dithiocarbamates by reaction with phosgene or chloroformates (→ Isocyanates, Organic) [164]:

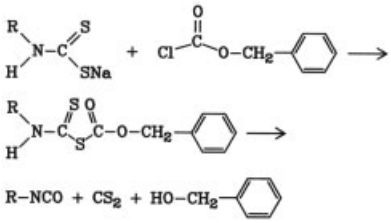

Thioureas can be obtained from isothiocyanates and primary or secondary amines, by pyrolysis of thiuram disulfides [165], [166], or from dithiocarbamate salts [167] (→ Thiourea):

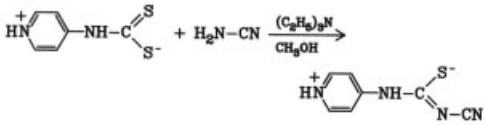

The synthesis of ethylenethiourea, an accelerator for polychloroprene, also proceeds via a dithiocarbamate salt:

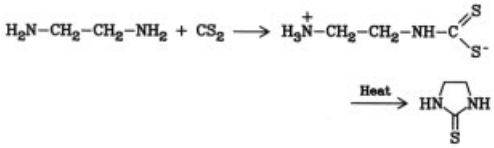

N-Substituted ethylenethiourea derivatives have also been described [168].

4.7. Photolysis of Dithiocarbamates

Irradiation of acyl dithiocarbamates gives acyl and dithiocarbamic acid radicals [169]:

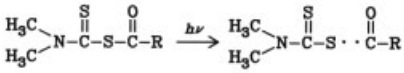

These radicals can initiate polymerizations of olefins. The living radical photopolymerization of styrene with the aid of benzyl *N*,*N*-dialkyldithiocarbamates has also been described [170]:

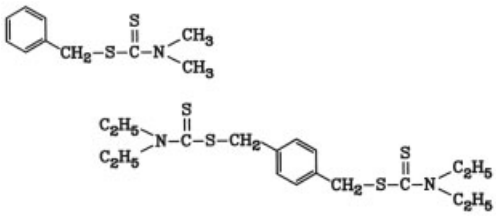

Other dithiocarbamates have been proposed as photoinitiators for methyl acrylate [171].

5. Uses

5.1. Vulcanization Accelerators

Thiuram sulfides and salts, especially zinc salts, of *N,N*-dialkyldithiocarbamic acids are vulcanization accelerators for natural and synthetic rubber [46], [172], [173]; some examples of commercial products are given in Tables 7 and 8. Water-soluble derivatives are used in low-temperature processing of latex. Thiuram disulfides and polysulfides give heat-resistant rubber products. Sodium and potassium dimethyldithiocarbamates are modifiers in emulsion polymerization [174]. 4-[(Morpholinothiocarbonyl)thio]morpholine is a vulcanization accelerator which gives good resistance to reversion and low compression set [175]. Thiadiazines are also utilized as accelerators [176].

Dithiocarbamic acids are intermediates in the production of thioureas, particularly ethylene-, *N,N′*-diphenyl-, and tetramethylthiourea, which are used as accelerators for polychloroprene. Dust production has been reduced by formulation as polymeric granules. Under certain circumstances, thioureas can be injurious to health (see Chap. 6) [177]; 3-methylthiazolidine-2-thione is a polychloroprene accelerator that does not have this disadvantage [137]. Other nonhazardous vulcanization accelerators have been studied [178].

5.2. Pesticides

Salts and derivatives of various dithiocarbamic acids are used as fungicides, nematocides, or bactericides (Table 9). These compounds liberate mustard oils or dialkyldithiocarbamic acids [179], [180]. Dithiocarbamate pesticides have been the subject of intensive metabolic [181], [182] and environmental [183–186] investigations. Analytical questions are discussed in [187].

5.3. Medical Applications

Tetraethylthiuram disulfide, known as disulfiram, is a withdrawal agent used in the treatment of alcoholism (Table 10). Nontoxic penicillins containing a dithiocarbamic acid structure are effective against penicillin-resistant organisms [188]. Dithiocarbamic acid derivatives are also used as antimycotics [154], [189] and nickel-poisoning antidotes [190] (see Table 10).

A broad spectrum of activity is shown by 2,6-bis-(thiocarbamoylthiomethyl)pyridine [191]. Dithiocarbamic acids affect the biosynthesis of catecholamines [192] and hepatosine [193], cause immune responses when used with levamisole [194], and reduce cholesterol production in rat liver [195]. Tetraethylthiuram monosulfide is used in conjunction with salicylic acid in the treatment of leishmaniasis [196]. Fungal infections of the skin can be treated with zinc ethylenebis(dithiocarbamate) [197]. Dithiocarbamates are anticaries agents [198]. 3-Aminotetrahydro-1,3-thiazine-2,4-diones [199], produced from cyclic dithiocarbamic acids, are skin-protecting agents. Certain tetrahydro-β-carbolines containing a dithiocarbamic acid structure are effective in liver therapy [200].

5.4. Radioprotective Agents

Dithiocarbamic acid esters and other derivatives provide protection against γ-radiation [201–206]. The metabolism of some of these agents in mice has been studied [207]. Polymers, e.g., poly(vinyl chloride), that contain dithiocarbamic acid substituents are more radiation resistant than unsubstituted polymers [208].

5.5. Imaging Technology

Dithiocarbamic acids and their derivatives have many applications in photographic and recording materials [209–211].

Heat-sensitive copying materials contain, for example, tetramethylthiuram disulfide and zinc dithiocarbamates [212]. Light-sensitive photographic materials for high-contrast negatives and direct-positive photographic materials employ certain dithiocarbamates [213], [214].

Table 7. Metal dithiocarbamates used as vulcanization accelerators

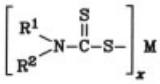

R^1	R^2	M	x	CAS registry number	Trade name	Manufacturer
CH_3-	CH_3-	Na	1	[*128-04-1*]	ACCEL SDD KUMAC SDD	Kawaguchi KUMHO (Korea)
CH_3-	CH_3-	Zn	2	[*137-30-4*]	Vulkacit L Vulcafor ZDMC Methazate ACCEL PZ	Bayer Vulnax Uniroyal Kawaguchi
CH_3-	CH_3-	Pb	2	[*19010-66-3*]	Methyl Ledate	R. T. Vanderbilt
CH_3-	CH_3-	Cu	2	[*137-29-1*]	Ekaland CDMC Methyl Cumate	Atochem R. T. Vanderbilt
CH_3-	CH_3-	Fe	3	[*67055-36-1*]	Royalac 133	Uniroyal
CH_3-	CH_3-	Bi	3	[*21260-46-8*]	Robac BJDD Bismate	Robinson Brothers R. T. Vanderbilt
CH_3-	CH_3-	Se	4	[*19632-73-6*]	Methyl Selenac	R. T. Vanderbilt
C_2H_5-	C_2H_5-	Na	1	[*148-18-5*]	Oricel ESL Eveite L Vulcafor SDC	Oriental ACNA (Italy) Vulnax
C_2H_5-	C_2H_5-	Zn	2	[*14324-55-1*]	ACCEL EZ Eveite Z Robac ZDC Vulkacit LDA Ethasan Ethazate Ethyl Zimate Perkacit ZDEC	Kawaguchi ACNA Robinson Brothers Bayer Monsanto Uniroyal R. T. Vanderbilt Akzo
C_2H_5-	C_2H_5-	Se	4	[*5456-28-0*]	Ethyl Selenac	R. T. Vanderbilt
C_2H_5-	C_2H_5-	Te	4	[*20941-65-5*]	Ethyl Tellurac Ekaland TE DEC Ethyl Telluram	R. T. Vanderbilt Atochem Prochimie
C_2H_5-	C_2H_5-	Cd	2	[*14239-68-0*]	Ethyl Cadmate	R. T. Vanderbilt
C_4H_9-	C_4H_9-	Zn	2	[*136-23-2*]	Butasan Vulcafor ZDBC Butazate Butyl Ziram Butyl Zimate Vulkacit LDB	Monsanto Vulnax Uniroyal Prochimie R. T. Vanderbilt Bayer
C_4H_9-	C_4H_9-	Na	1	[*136-30-1*]	Oricel TP	Oriental
C_4H_9-	C_4H_9-	Ni	2	[*13927-77-0*]	NDBC Antage NBC Ekaland NBC Antigene NBC NJBUD	Prochimie Kawaguchi Atochem Sumitomo Akron
C_4H_9-	C_4H_9-	Cu	2	[*52691-95-9*]	Ekaland CDBC	Atochem
$C_5H_{11}-$	$C_5H_{11}-$	Pb	2	[*36501-84-5*]	Amyl Ledate	R. T. Vanderbilt
$C_5H_{11}-$	$C_5H_{11}-$	Zn	2	[*15337-18-5*]	Amyl Zimate	R. T. Vanderbilt
		Cd	2	[*14949-59-8*]	Robac CPD	Robinson Brothers
		Pb	2	[*81195-36-0*]	Robac LPD	Robinson Brothers

Table 7. Continued

	Zn	2	[*13878-54-1*]	Vulkacit ZP Nocceler ZP	Bayer Ouchi
C_2H_5- C_6H_5-	Zn	2	[*14634-93-6*]	Hermat FEDK Vulkacit P extra N Eveite P ACCEL PX	Chemapol Bayer ACNA Kawaguchi
	Zn	2	[*14726-36-4*]	Robac ZBED Arazate Perkacit ZBEC	Robinson Brothers Uniroyal Akzo

Table 8. Other dithiocarbamate derivatives used as vulcanization accelerators

Compound	CAS registry number	Trade name	Manufacturer
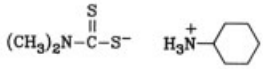	[*78600-26-7*]	AZ-100	Monsanto
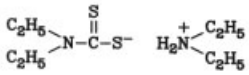	[*1518-58-7*]	Vulcafor DDCN	Vulnax
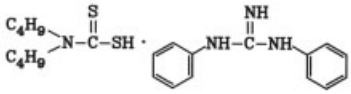	[*61792-03-8*]	Ultex	C. P. Hall
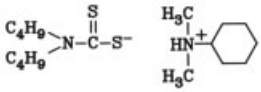	[*149-82-6*]	RZ-100	Monsanto
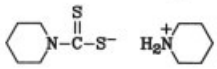	[*98-77-1*]	Nocceler PPD Vulkacit P Robac PPD	Ouchi (Japan) Bayer Robinson Brothers
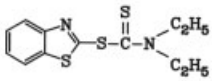	[*95-30-7*]	Ethylac	Pennwalt
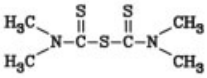	[*97-74-5*]	Eveite MST Vulkacit Thiuram MS Vulkafor TMTM Mono Thiurad Monex Unads ACCEL TS Sanceller TS Soxinol TS Nocceler TS	ACNA Bayer Vulnax Monsanto Uniroyal R. T. Vanderbilt Kawaguchi Sanshin Sumitomo Ouchi (Japan)
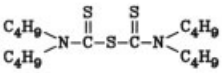	[*93-73-2*]	Pentex	Uniroyal

Table 8. Continued

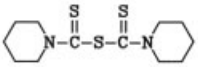	[*725-32-6*]	Robac PTM	Robinson Brothers
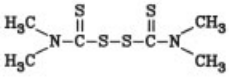	[*137-26-8*]	ACCEL TMT Eveite 4 MT Perkacit TMTD Vulkacit Thiuram Vulcafor TMTD Thiurad Tuex Robac TMT Accelerator TMTD Methyl-Tuads Super Accélerator 501 Sanceller TT	Kawaguchi ACNA Akzo Bayer Vulnax Monsanto Premier (Taiwan) Robinson Brothers Bann (Brazil) R. T. Vanderbilt Rhône-Poulenc Sanshin
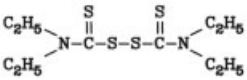	[*97-77-8*]	Eveite T Thiuram E Ethylthiurad Ethyl Tuex Ethyl Tuads Etiurac	ACNA Du Pont Monsanto Uniroyal R. T. Vanderbilt Pennwalt
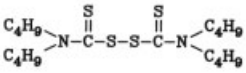	[*1634-02-2*]	Butyl Tuads Oricel TBT Robac TBTU	R. T. Vanderbilt Oriental Robinson Brothers
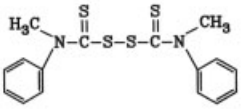	[*10591-84-1*]	Vulkacit J	Bayer
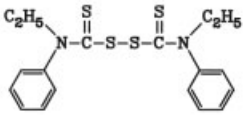	[*41365-24-6*]	PETD	Bozzetto
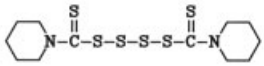	[*120-54-7*]	Tetrone A DPTT Ekaland TSPM	Du Pont Akron Atochem
	[*1908-87-8*]	Vulkacit CRV	Bayer
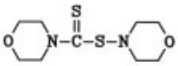	[*13752-51-7*]	Cure-rite 18	B. F. Goodrich

Table 9. Dithiocarbamic acid derivatives used in agriculture

Compound	CAS registry number	Application	Trade name	Manufacturer
$CH_3-NH-C(=S)-SNa \cdot 2\ H_2O$	[*137-42-8*]	soil fungicide herbicide	Vapam	Stauffer Chemical
$(H_3C)_2N-C(=S)-SNa \cdot 2\ H_2O$	[*72140-17-1*]	herbicide	Na-DMDT, Diram	Stauffer Chemical
$[(H_3C)_2N-C(=S)-S-]_2Zn$	[*137-30-4*]	protective fungicide	Zerlate, Milban Fuklasin Pomarsol-Z Zinc carbamate	Du Pont Schering Bayer Bayer
$[(H_3C)_2N-C(=S)-S-]_3Fe$	[*14484-64-1*]	fungicide	Fermate	Du Pont
$[(H_3C)_2N-C(=S)-S-]_2Ni$	[*15521-65-0*]	rice bactericide	Sankel	Sankyo
$(H_3C)_2N-C(=S)-S-S-C(=S)-N(CH_3)_2$	[*137-26-8*]	fungicide	Arasan, Tersan Nomersan Pomarsol	Du Pont Plant Protection Bayer
$(H_3C)_2N-C(=O)-S-S-C(=S)-N(CH_3)_2$	[*1115-06-6*]	fungicide	Niagara 9130	R. T. Vanderbilt
$NH(C(=S)-SNa)-CH_2-CH_2-NH(C(=S)-SNa)$	[*142-59-6*]	soil fungicide	Dithane D-14 Parzate, DSE Amobam	Rohm and Haas Du Pont Roberts
$NH(C(=S)-S)-CH_2-CH_2-NH(C(=S)-S)Zn$	[*12122-67-7*]	leaf fungicide soil fungicide	Dithane Z-78 Parzate, Zineb Lonacol	Rohm and Haas Du Pont Bayer
$NH(C(=S)-S)-CH_2-CH_2-NH(C(=S)-S)Mn$	[*12427-38-2*]	leaf fungicide soil fungicide seed disinfectant	Dithane M-22 Manzate, Maneb Dithane M-22 Maneb Wettable Powder	Du Pont Du Pont Rohm and Haas Bayer
$NH(C(=S)-S)-CH_2-CH_2-NH(C(=S)-S)(Mn + Zn)$ 20% 2.5%	[*8018-01-7*]	leaf fungicide fungicide	Dithane M-45 (mancozeb)	Rohm and Haas
$[-Zn-S-C(=S)-NH-CH_2-CH(CH_3)-NH-C(=S)-S-]_n$	[*12071-83-9*]	fungicide	Antracol	Bayer

Table 9. Continued

$\left[-\overset{S}{\overset{\|}{C}}-NH-CH_2-CH_2-NH-\overset{S}{\overset{\|}{C}}-S-\right]_n$	[*58855-93-9*]	fungicide	Thioneb	Du Pont
$\left[-S-\overset{S}{\overset{\|}{C}}-NH-CH_2-CH_2-NH-\overset{S}{\overset{\|}{C}}-S-\right]_n$	[*9006-42-2*]	fungicide	Polyram	BASF
$(C_2H_5)_2N-\overset{S}{\overset{\|}{C}}-S-CH_2-CCl=CH_2$	[*95-06-7*]	soil herbicide	Vegadex	Monsanto Chemical
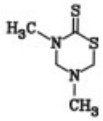	[*533-74-4*]	nematocide fungicide	Mylone, Dazomet N-521 Basamid Drozine	Union Carbide Stauffer Chemical BASF
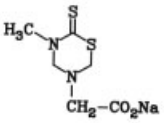	[*3655-88-7*]	nematocide soil fungicide	Terracur	Bayer
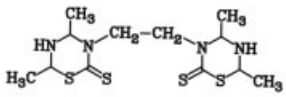	[*3773-49-7*]	fungicide	Du Pont Fungicide 328 Milneb Banlate	Du Pont

Table 10. Medical applications of dithiocarbamic acid derivatives

Compound	CAS registry number	Application	Trade name	Manufacturer
$(C_2H_5)_2N-\overset{S}{\overset{\|}{C}}-S-S-\overset{S}{\overset{\|}{C}}-N(C_2H_5)_2$	[*97-77-8*]	alcoholism withdrawal agent	Antabuse (disulfiram)	Tosse
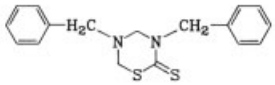	[*350-12-9*]	antimycotic	Fungiplex (sulbentine)	Hermal
$(C_2H_5)_2N-\overset{S}{\overset{\|}{C}}-SNa$	[*148-18-5*]	nickel-poisoning antidote	Dithiocarb	
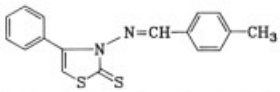	[*15387-18-5*]	antimycotic	Polyodin (fezatione)	Takeda

Dithiocarbamates are used in color photography [215] and in electrophotographic materials [216]. Photopolymerizable printing compositions contain thiuram sulfides [217], and latent image developers contain zinc dithiocarbamates [218]. Developer compositions for electrolyte photography may contain small amounts of dimethyldithiocarbamates [219]. Dithiocarbamates are likewise used in the bleaching of photographic materials [220] and in photothermographic materials [221]. Thermodeveloping photographic materials contain metal derivatives of dithiocarbamic acid [222], [223].

Dialkyldithiocarbamates [224] or rhodanines [225] are employed in photosensitive resin compositions, and certain rhodanines are used in sensitizing dyes [226]. Photosensitive materials for high resolution may contain cyclic dithiocarbamic acid esters [227]. Solutions containing dithiocarbamic acid or its salts are used in the production of photographic relief images [228], [229]. Dialkyldithiocarbamic acid derivatives are used in the bleaching of silver halide-containing photographic materials [230]. Small amounts of dithiocarbamic acid derivatives increase the brilliance of photographic silver images [231]. Ammonium dithiocarbamate is a heat and moisture stabilizer for photographic emulsions [232]. Combinations of sulfinic acids and ammonium salts of dithiocarbamic acid, as well as thiuram disulfides, increase the sensitivity of photographic emulsions [233–238]. Certain dithiocarbamic acid derivatives increase the sensitivity of diazole compounds in diazotype printing [239]. Dialkyldithiocarbamic acid derivatives stabilize silver halide emulsions [240]. Gel formation in photopolymerization is prevented by dithiocarbamates [241], which produce positive prints with small amounts of silver halides [242]. Color fading in silver halide images can be prevented by treatment with a dithiocarbamate salt solution [243]. Color images are produced in direct-positive processes with the aid of dithiocarbamic acid derivatives [244]. Silver ions and a solution containing sodium *N*-methyldithiocarbamate [245] participate in the photocuring of high molecular mass substances containing amino or amido groups. Dithiocarbamates are used in the recovery of silver from photographic processing solutions [246].

5.6. Other Uses

Curing Agents. Heavy-metal salts of dibutyldithiocarbamic acid in epoxy resins improve adhesion properties and moisture resistance, and inhibit rust formation [247]. Treatment of halogen-containing elastomers with a dimercaptothiadiazole – dithiocarbamate mixture improves flow behavior and cure rates [248].

Separation of Heavy Metals. Dithiocarbamates precipitate heavy metals such as cadmium, nickel, and zinc from aqueous solution [249], [250]. Polydithiocarbamate resins remove heavy-metal ions from wastewater [251]. Dithiocarbamic acid derivatives are employed as flotation agents for copper, zinc, nickel, lead, and iron ores [252].

Complexation. Dithiocarbamic acid derivatives stabilize transition metals in high states of oxidation [253]. The resulting metal complexes have been the subject of electrochemical [254], crystallographic [255], thermochemical [256], and structural [257] studies.

Stabilizers for Polymers and Oils. Zinc dithiocarbamates improve the thermal stability of polymers, especially polyurethane elastomers [258]; combinations with sterically hindered phenols show synergistic stabilizing effects [259], [260]. Zinc dialkyldithiocarbamates are antioxidants [261]. Nickel dithiocarbamate inhibits the photodegradation of polyethylene [259], [262]. Dithiocarbamates improve the ozone resistance of printing plates [263] and the stability of photosensitive resins [264]. Protective effects are found in the radiofrequency range [265]. Certain metal derivatives serve as oil [266] and lubricant additives [267]. Aqueous synthetic resin dispersions containing tetramethylthiuram disulfide as a fungicide have improved storage stability [268]. Organopolysiloxane-based sealants used in the construction industry are protected from fungus attack by tetraalkylthiuram disulfide [269]. Biocides consisting of copper salts of alkylenebis(dithiocarbamates) and triphenyltin hydroxide considerably extend the life of underwater paints [270].

Analysis. Dithiocarbamic acids are useful in inorganic [271], [272] and other analyses [273], especially to determine trace amounts of heavy metals in organic and biological materials [186],

[274–276]; they can also be used as chelating agents in HPLC [277].

6. Toxicology and Occupational Health

The toxicity of dithiocarbamic acid derivatives [278–280] and oxidation products of the dithiocarbamates [281] has been investigated extensively. The biological activity of some dithiocarbamic acid derivatives is treated in [282]. The embryotoxicity of dithiocarbamates is being studied, with particular attention to the derivatives used in the rubber industry [283] and in pesticides [284]. The teratogenicity potential of such compounds [285] and their metabolites [286] has also been investigated. Mutagenicity studies of dithiocarbamic acid pesticides [287–289] and of products used in the rubber industry [290–292] have been reported. Cytological and genetic effects of ethylenebis(dithiocarbamates) used as fungicides have been investigated with the aid of *Allium cepa* [293]. Genotoxic effects have been observed for both synthetic elastomer additives (especially thiuram sulfides and dithiocarbamates) and pesticides [294].

Formation of Nitrosamines. During the vulcanization of rubber, dithiocarbamic acid derivatives used as accelerators form trace amounts of amines [295]. These amines react with nitrogen oxides from the air or with nitrosating ingredients of rubber compounds to form highly toxic nitrosamines [296]. The carcinogenic potential of these nitrosamines is being studied.

The TLV-TWAs and MAK values are 5 mg/m^3 for thiram (tetramethylthiuram disulfide) and 2 mg/m^3 for disulfiram (tetraethylthiuram disulfide).

7. References

General References

1. *Houben-Weyl,* **9,** 824; E 4, 458.
2. G. D. Thorn, R. A. Ludwig: *The Dithiocarbamates and Related Compounds,* Elsevier, Amsterdam-New York 1962.
3. G. Scheuerer, *Fortschr. Chem. Forsch.* **9** (1967) 254.
4. W. Walter, K.-D. Bode, *Angew. Chem.* **79** (1967) 285; *Angew. Chem. Int. Ed. Engl.* **6** (1967) 281.
5. L. A. Summers, *Rev. Pure Appl. Chem.* **18** (1968) 1.
6. G. Gattow, W. Behrendt: "Carbon Sulfides and Their Inorganic and Complex Chemistry" in A. Senning (ed.): *Topics in Sulfur Chemistry,* vol. 2, Thieme-Verlag, Stuttgart 1977.
7. E. E. Reid: *Organic Chemistry of Bivalent Sulfur,* **vol. 4,** Chemical Publishing Co., New York 1962, p. 131 ff.
8. F. Duus: "Thionocarbamic and Dithiocarbamic Acids and their Derivatives," in D. Barton, W. D. Ollis (eds.): *Comprehensive Organic Chemistry,* vol. 3, Pergamon Press, Oxford 1979, p. 469.
9. R. Zahradnik, *Chem. Tech. (Leipzig)* **10** (1958) 546.
10. D. Coucouvanis, *Prog. Inorg. Chem.* **11** (1970) 233; **26** (1979) 301.

Specific References

11. G. Gattow, V. Hahnkamm, *Z. Anorg. Allg. Chem.* **364** (1969) 161. G. Gattow, V. Hahnkamm, *Angew. Chem.* **78** (1966) 334; *Angew. Chem. Int. Ed. Engl.* **5** (1966) 316. J.-L. Fourquet, *Bull. Soc. Chim. Fr.* 1969, no. 9, 3001. D. De Filippo, P. Deplano, F. Devillanova, E. F. Trogu, G. Verani, *J. Org. Chem.* **38** (1973) 560. S. J. Joris, K. J. Aspila, Ch. L. Chakrabarti, *J. Phys. Chem.* **74** (1970) 860.
12. D. Graig, A. E. Juve, W. L. Davidson, W. L. Semon, D. C. Hay, *J. Polym. Sci.* **8** (1952) 321.
13. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **524** (1985) 111.
14. M. Sato, N. Fukada, M. Kurauchi, *Synthesis* 1981, 554.
15. T. Takeshima, M. Ikeda, M. Yokoyama, N. Fukada, *J. Chem. Soc. Perkin Trans. I* 1979, 692.
16. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **525** (1985) 112.
17. V. Hahnkamm, G. Kiel, G. Gattow, *Z. Anorg. Allg. Chem.* **368** (1969) 127. F. Christiani, F. A. Devillanova, G. Verani, *Int. Symp. Interact. Mol. Ions [Proc.] 3rd,* **vol. 1,** 103; *Chem. Abstr.* **88** (1978) 104 189.
18. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **522** (1985) 145.
19. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **526** (1985) 122.
20. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **524** (1985) 122.
21. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **528** (1985) 157.

22. C. W. Voigt, G. Gattow, *Z. Anorg. Allg. Chem.* **437** (1977) 226.
23. T. Callan, M. Strafford, *J. Soc. Chem. Ind. Trans.* **43** (1924) 7.
24. B. F. Goodrich, US 2 117 619, 1937; US 2 123 370, 1937; US 2 123 373, 1937. Amer. Cyanamid, US 2 235 747, 1939.
25. Sharples Chem., US 2 492 314, 1945.
26. G. Gattow, W. Eul, *Z. Anorg. Allg. Chem.* **483** (1981) 103.
27. C. W. Voigt, G. Gattow, *Z. Anorg. Allg. Chem.* **437** (1977) 233.
28. Sharples Chem., US 2 443 160, 1944.
29. R. T. Vanderbilt Co., US 2 347 128, 1943. J. Stary et al., *Talanta* **15** (1968) 505. N. K. Wilson, L. Fishbein, *J. Agric. Food Chem.* **20** (1972) 847. O. Foss, *Acta Chem. Scand.* **5** (1951) 115.
30. Rohm & Haas Co., US 2 317 765, 1941.
31. H. Nishimura, T. Kinugasa, *Chem. Pharm. Bull.* **17** (1969) 94. J. Dunderdale, T. J. Watkins, *Chem. Ind. (London)* 1956, 174.
32. G. Gattow, S. Lotz, *Z. Anorg. Allg. Chem.* **531** (1985) 101.
33. G. Gattow, S. Lotz, *Z. Anorg. Allg. Chem.* **531** (1985) 97.
34. G. D. Thorn, R. A. Ludwig: *The Dithiocarbamates and Related Compounds,* Elsevier, Amsterdam 1962.
35. G. Gattow, W. Behrendt, "Carbon Sulfides and Their Inorganic and Complex Chemistry," in A. Senning (ed.), *Topics in Sulfur Chemistry,* Thieme-Verlag, Stuttgart 1977.
36. Du Pont, US 3 276 950, 1965 (A. W. Engelhard); *Chem. Abstr.* **66** (1967) 1871.
37. J. v. Braun, W. Kaiser, *Ber. Dtsch. Chem. Ges.* **56** (1923) 550. Naugatuck Chem. Co., US 1 782 111, 1925.
38. Wingfoot Corp., US 2 014 353, 1931.
39. Silesia, DE 444 014, 1925.
40. Monsanto Chem. Co., US 2 375 083, 1943. J. v. Braun, *Ber. Dtsch. Chem. Ges.* **35** (1902) 819.
41. *Fiat* **1018** (1947) 52, 58. Goodyear, DE-OS 2 349 313, 1973 (J. J. Tazuma, B. A. Bergorni). UCB, DE-OS 2 527 898, 1975 (J. M. G. Lietard, G. Matthijis).
42. Roessler & Hasslacher Chem. Co., US 1 796 977, 1928. Monsanto Chem. Co., GB 555 874, 1942.
43. Naugatuck Chem. Co., US 1 782 111, 1925 (H. S. Adams, L. Meuser). SU 56 086, 1939. Sharples Chem. Inc., US 2 325 194, 1941.
44. Akzo, DE 3 105 587, 1981 (L. Eisenhut, H. G. Zengel, M. Bergfeld); BE 892 143, 1981; *Chem. Abstr.* **97** (1982) 144 394. Akzo, DE 3 105 622, 1981 (L. Eisenhuth, H. G. Zengel, M. Bergfeld); BE 892 144, 1981; *Chem. Abstr.* **97** (1982) 144 395. C. J. Swan, D. L. Trimm, *J. Appl. Chem.* **18** (1968) 340.
45. S. Torii, H. Tanaka, K. Mishima, *Bull. Chem. Soc. Jpn.* **51** (1978) 1575; *Chem. Abstr.* **89** (1978) 67 437.
46. *Bayer Manual for the Rubber Industry,* Bayer, Leverkusen, Rubber Division, Technical Service Section, 1971.
47. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **527** (1985) 125.
48. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **528** (1985) 168.
49. A. Schönberg, *Ber. Dtsch. Chem. Ges.* **68** (1935) 163.
50. J. v. Braun, F. Stechele, *Ber. Dtsch. Chem. Ges.* **36** (1903) 2280. Naugatuck Chem. Co., US 1 682 920, 1926.
51. I.G. Farben, DE 519 445, 1930.
52. Du Pont, US 2 048 043, 1931.
53. Du Pont, US 1 788 632, 1928.
54. A. Cambron, *Can. J. Res.* **2** (1930) 341.
55. Roessler & Hasslacher Chem. Co., US 1 681 717, 1926. Du Pont US 2 414 014, 1943.
56. E. L. Carr, G. E. P. Smith, G. Alliger, *J. Org. Chem.* **14** (1949) 921. R. A. Donia, J. A. Shotton, L. O. Beutz, G. E. P. Smith Jr., *J. Org. Chem.* **14** (1949) 946. G. E. P. Smith, G. Alliger, E. L. Carr, K. C. Young, *J. Org. Chem.* **14** (1949) 935. Monsanto Chem. Co., US 3 732 222, 1970; *Chem. Abstr.* **79** (1973) 54 669.
57. S. Torii, H. Tanaka, M. Ukita, *Jpn. Kokai Tokkyo Koho* **79** (1979) 115, 323; *Chem. Abstr.* **92** (1980) 84 927.
58. Denki Kagaku Kogyo K., Kanesha Co. Ltd., JP 75/7134, 1975; *Chem. Abstr.* **83** (1975) 142 999.
59. Vanderbilt R. T. Co. Inc., DE-OS 2 446 555, 1975 (K. S. Karsten); *Chem. Abstr.* **82** (1975) 173 000.
60. R. N. Datta, M. M. Das, D. K. Basu, A. K. Chaudhuri, *Rubber Chem. Technol.* **57** (1984) 879.
61. W. Hofmann, *Plast. Rubber Process Appl.* **5** (1985) 209.
62. B. Adhikari et al., *Rubber Chem. Technol.* **56** (1983) 327.
63. B. F. Goodrich Co., DE-OS 2 325 027, 1973 (R. D. Taylor); (1974) *Chem. Abstr.* **80** 109 604.
64. B. F. Goodrich Co., DE-OS 2 324 934, 1973 (R. D. Taylor); *Chem. Abstr.* **81** (1974) 26 970.

65. B. F. Goodrich Co., DE-OS 2 324 933, 1973 (R. D. Taylor); *Chem. Abstr.* **80** (1974) 146 841.
66. Amer. Cyanamid, DE-OS 2 827 933, 1979 (Franz, Curtis Allen); *Chem. Abstr.* **90** (1979) 122 845.
67. B. F. Goodrich Co., US 3 985 743, 1972, 1976 (R. D. Taylor); *Chem. Abstr.* **86** (1977) 30 837.
68. Th. Kempermann, *Kautsch. Gummi Kunstst.* **20** (1967) 126. R. D. Taylor, "Thiocarbamyl-Sulfenamide als Vulkanisationsbeschleuniger," *Rubber Chem. Technol.* **47** (1974) 906.
69. B. F. Boodrich Co., DE-OS 2 324 981, 1973 (R. N. Taylor); *Chem. Abstr.* **81** (1974) 26 969.
70. Uniroyal Inc., EP 154 437, 1984 (J. K. Stieber); *Chem. Abstr.* **104** (1986) 111 175. G. E. P. Smith, G. Alliger, E. L. Corr, K. C. Young, *J. Org. Chem.* **14** (1949) 935.
71. A. K. Khamrai, B. Adhikari, M. M. Maiti, S. Maiti, *Angew. Makromol. Chem.* **143** (1986) 39.
72. Sharples Chem. Inc., US 2 466 276, 1946; FR 960 006, 1948.
73. O. C. Billeter, H. Rivier, *Ber. Dtsch. Chem. Ges.* **37** (1904) 4319. G. M. Dyson, H. J. George, *J. Chem. Soc.* **125** (1924) 1702.
74. M. S. Newmann, F. W. Hetzel, *J. Org. Chem.* **34** (1969) 3604. H. G. Viehe, Z. Janousek, *Angew. Chem.* **83** (1971) 614; *Angew. Chem. Int. Ed. Engl.* **10** (1971) 573. R. H. Goshorn, W. W. Levis Jr., E. Jaul, E. J. Ritter, T. L. Cairns, H. E. Cupery: *Organic Syntheses,* Coll. vol. IV, Wiley, New York 1963, pp. 307, 310.
75. *Houben-Weyl,* **E 4,** 416.
76. S. Scheithauer, R. Mayer: "Thio- and Dithiocarboxylic Acids and Their Derivatives," in A. Senning (ed.): Topics in Sulfur Chemistry, vol. 4, Thieme, Stuttgart 1979, p. 293. *Houben Weyl,* **E 4,** 416. W. Walter, R. F. Becker, *Justus Liebigs Ann. Chem.* **755** (1972) 145.
77. N. N. Yarovenko, A. S. Vasil'eva, *Zh. Obshch. Khim.* **29** (1959) 3786; *J. Gen. Chem. USSR* (*Engl. Transl.*) **29** (1959) 3747. L. M. Yagupol'skii, M. I. Dronkina, *Zh. Obshch. Khim.* **36** (1966) 1309; *J. Gen. Chem. USSR* (*Engl. Transl.*) **36** (1966) 1323. Bayer, DE-OS 3 044 216, 1982 (B. Baasner, G. M. Petruck, H. Hagemann, E. Klauke); *Chem. Abstr.* **97** (1982) 181 932.
78. G. Nochmias, *Ann. Chim.* **12** (1952) 584.
79. A. Bhai, A. Das, S. Medheker, K. S. Boparai, *J. Indian Chem. Soc.* **58** (1981) 295.
80. A. M. Kuliev, M. A. Kulieva, T. N. Kulibekova, A. K. Ibazade, *Neftekhimiya* **25** (1985) 679.
81. M. W. Harman, J. J. D'Amico, *J. Am. Chem. Soc.* **75** (1953) 4081.
82. Naugatuck Chem. Co., US 1 726 647, 1928.
83. J. v. Braun, *Ber. Dtsch. Chem. Ges.* **42** (1909) 4568.
84. I.G. Farben, DE 575 372, 1931; Monsanto Chem. Co., US 2 695 901, 1953.
85. Bayer, DE-AS 1 245 358, 1964.
86. A. I. Griogorieva, I. F. Titova, V. N. Konygin, SU 1 065 396, 1982; *Chem. Abstr.* **101** (1984) 38 146.
87. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **524** (1985) 117.
88. R. Gerner, G. Gattow, *Z. Anorg. Allg. Chem.* **527** (1985) 130.
89. A. M. Clifford, J. G. Lichty, *J. Am. Chem. Soc.* **54** (1932) 1163.
90. A. Rieche, G. Hilgetag, D. Martin, I. Kreyzi, *Arch. Pharm. (Weinheim Ger.)* **296** (1963) 310. Y. Iwakura, A. Nabeya, F. Nishigushi, K. H. Okkawa, *J. Org. Chem.* **31** (1966) 3352. J. v. Braun, *Ber. Dtsch. Chem. Ges.* **35** (1902) 3377. A. Kaji, *Bull. Chem. Soc. Jpn.* **34** (1961) 1147.
91. Bayer, DE 1 768 874, 1968.
92. Phillips Petroleum Co., US 2 690 440, 1951.
93. A. A. Watson, *J. Chem. Soc.* **1964,** 2100. Phillips Petroleum Co., US 2 862 850, 1954.
94. M. G. Zimin, M. M. Afanasev, A. V. Mironov, A. N. Pudovik, *Zh. Obshch. Khim.* **51** (1981) 470.
95. M. Roshdestwenski, *Zh. Russ. Fiz. Khim. Ova. Chast. Khim.* **41** (1909) 1438; *Chem. Zentralbl.* 1910, 910.
96. Olin Mathieson Chem. Corp., US 2 940 978, 1958.
97. T. I. Bychkova, M. A. Vasil'eva, L. B. Krivdin, A. V. Kalabina, *Zh. Org. Khim.* **20** (1984) no. 10, 2114.
98. Amer. Cyanamid, US 3 407 222, 1965; *Chem. Abstr.* **70** (1969) 37 247. T. Nakai, M. Okawara, *Bull. Chem. Soc. Jpn.* **41** (1968) 707.
99. U. Stanior, M. Weißler, *Arch. Pharm. (Weinheim Ger.)* **317** (1984) 1042.
100. Gulf Research and Development Co., US 3 676 479, 1969; *Chem. Abstr.* **77** (1972) 100 872.
101. A. P. Sineokov, F. N. Gladsheva, V. S. Etlis, *Chem. Heterocycl. Comp. (Engl. Transl.)* 1970, 562. H. Stamm, *Pharm. Zentralhalle* **107** (1968) 440. T. A. Foglia, L. M. Gregory, G. Maerker, S. F. Osman, *J. Org. Chem.* **36** (1971) 1068.

102. N. Kreutzkamp, H. Y. Oei, H. Peschel, *Arch. Pharm. (Weinheim Ger.)* **304** (1971) 649.
103. A. O. Fitton et al., *J. Chem. Soc. C* **1969,** 230;**1968,** 996; **1971,** 1245. A. O. Fitton, J. Hill, M. Clutob, A. Thompon, *J. Chem. Soc. Perkin Trans. I* **1972,** 2658. N. Kreutzkamp, H. Y. Oei, H. Peschel, *Arch. Pharm.* (*Weinheim Ger.*) **304** (1971) 648. ICI, FR 1 559 120, 1968; *Chem. Abstr.* **71** (1969) 113 737. Uniroyal Inc., DE-OS 1 815 221, 1968; *Chem. Abstr.* **72** (1970) 3218.
104. Uniroyal F. X. O'Shea, US 3 330 804, 1963; NL 64/8883, 1963; *Chem. Abstr.* **63** (1965) 17 974 E
105. D. M. Morton, R. H. Neilson, *Phosphorus Sulfur* **25** (1985) 315.
106. Rubber Service Laboratories Co., US 1 586 121, 1925.
107. R. A. Donia, J. A. Shotton, L. O. Beutz, G. E. P. Smith jr., *J. Org. Chem.* **14** (1949) 952.
108. Wingfoot Corp., US 2 238 331, 1936.
109. T. P. Johnston, C. R. Stringfellow jr., A. Gallagher, *J. Org. Chem.* **27** (1962) 4068. R. Dalaby et al., *C. R. Hebd. Séances Acad. Sci.* **232** (1951) 1676. Monsanto Chem. Co., US 3 372 150, 1966; *Chem. Abstr.* **67** (1967) 3088. US 3 361 752, 1964;BE 6 704 434, 1964; *Chem. Abstr.* **65** (1966) 15 383 E.
110. Wingfoot Corp., GB 675 450, 1950; US 2 597 988, 1950. Monsanto Chem. Co., US 3 726 866, 1970; *Chem. Abstr.* **79** (1973) 20 064.
111. N. K. Blizuyuk et al., SU 175 054, 1966.
112. Bayer, DE-AS 1 178 417, 1964; DE-AS 1 178 418, 1964. G. Buchmann, O. Wolniak, *Pharmazie* **21** (1966) 650. E. G. Novikov, J. N. Tugarinova, *Chem. Heterocycl. Comp.* 1968, 207.
113. R. Delaby, R. Daumiens, R. Seyden-Penne, *C. R. Hebd. Séances Acad. Sci.* **238** (1954) 121.
114. E. Profft, R. Schmuck, *Arch. Pharm. (Weinheim Ger.)* **296** (1963) 209.
115. J. L. Garraway, *J. Chem. Soc.* 1964, 4004, 4008.
116. C. S. Angadiyavar, M. N. Gudi, M. V. George, *Indian J. Chem. Soc.* **10** (1972) 888.
117. V. N. Elokhina, A. S. Nakhmanovich, A. E. Aleksandrova, B. J. Vishnevskii, J. D. Kalikhman, *Khim. Farm. Zh.* **20** (1986) 1061.
118. J. C. Sauer, *J. Org. Chem.* **24** (1959) 1592.
119. T. Hayashi, T. Oishi, *Chem. Lett.* 1985, 413.
120. *Houben-Weyl,* **E 11,** 82; **E 4,** 551.
121. T. Olijusma, I. B. F. N. Engberts, *Synth. Commun.* **3** (1973) 1. H. C. Hansen, A. Senning, *Org. Prep. Proced. Int.* **17** (1985) 275. *Houben Weyl,* **E 4,** 524.
122. W. J. Stephen, A. Townshend, *J. Chem. Soc.* 1965, 5127.
123. E. Cherbuliez, J. Marszalek, J. Rabinowitz, *Helv. Chim. Acta.* **47** (1964) 1666.
124. L. Musial, J. Staniec, *Rocz. Chem.* **44** (1970) 1801.
125. L. Y. Ladnaya, E. M. Protsenko, *Pharm. Chem. J. (Engl. Transl.)* 1968, 257.
126. S. O. Abdallah, H. H. Hammouda, *J. Heterocycl. Chem.* **22** (1985) 497.
127. E. K. Mikitenko, N. N. Romanov, *Khim. Geterotsikl. Soedin.* 1981, 141.
128. J. Kinugawa, H. Nagase, *Yakugaku Zasshi* **86** (1966) 95, 101.
129. A. Miolati, *Gazz. Chim. Ital.* **23** (1894) 578. H. Erlenmeyer, M. Simon, *Helv. Chim. Acta* **25** (1942) 362. M. O. Kolosowa, V. I. Stavroskaya, *J. Gen. Chem. USSR (Engl. Transl.)* **33** (1963) 2706.
130. Agency of Ind. Sci. Tech., JP 60 096 604-A, 1983; *Chem. Abstr.* **103** (1985) 179 168.
131. C. M. Roussel, R. Gallo, M. Chanon, J. Metzger, *Bull. Chim. Fr.* 1971, 1902. J. J. D'Amico, T. W. Bartman, *J. Org. Chem.* **25** (1960) 1336.
132. L. Brandsma, R. L. P. De Jong, H. D. Verkruijsse, *Synthesis* 1985, 948.
133. T. Chiba, H. Sato, T. Kato, *Heterocycles* **21** (1984) 613.
134. D. H. R. Barton, D. Bridon, S. Z. J. Zard, *J. Chem. Soc., Chem. Commun.* 1985, 1066.
135. G. J. Pustoshkin, F. Yu. Rachinski, *J. Org. Chem. USSR (Engl. Transl.)* 1966, 1254.
136. Amer. Cyanamid, US 3 215 704, 1963.
137. Amer. Cyanamid, US 3 215 703, 1962 (F. A. V. Sullivan, A. C. Lindaw); *Chem. Abstr.* **64** (1966) 20 910. Bayer, DE 2 701 215, 1977 (R. Schubart, U. Eholzer); *Chem. Abstr.* **89** (1978) 179 990. Bayer, EP 17 039, 1980 (R. Schubart, U. Eholzer); DE 2 911 662, 1979; *Chem. Abstr.* **94** (1981) 65 669. Bayer, EP 16 420, 1980 (D. Hüllstrung, J. Trimbach); DE 2 911 661, 1979; *Chem. Abstr.* **94** (1981) 84107.
138. J. Kopecký, J. Smejkal, *Bull. Soc. Chim. Belg.* **93** (1984) 231.
139. Luther-Universität Halle, DD 200 958-A, 1984; *Chem. Abstr.* **104** (1986) 68 845.
140. Yu. P. Kovtun, N. N. Romanov, *Khim. Geterotsikl. Soedin.* 1985, 498.
141. V. N. Charushin, V. G. Baklykov, O. N. Chupakhin, G. M. Petrova, E. O. Sidorov, *Khim. Geterotsikl, Soedin.* 1984, 680.

142. R. N. Warrener, E. N. Cain, *Chem. Ind. (London)* **48** (1964) 1989. E. N. Cain, R. N. Warrener, *Aust. J. Chem.* **23** (1970) 51.
143. W. Schroth, A. Hildebrandt, U. Becker, S. Freitag, M. Akram, R. Spitzner, *Z. Chem.* **25** (1985) 20.
144. N. M. Turkevic, B. S. Zimenkovskii, *Khim. Geterotsikl. Soedin.* 1967, 845.
145. T. L. Gresham, J. E. Jansen, E. W. Shawer, *J. Amer. Chem. Soc.* **70** (1948) 1001. E. Cherbuliez, A. Buchs, J. Marszalek, J. Rabinowitz, *Helv. Chim. Acta* **48** (1965) 1414.
146. T. Takeshima, N. Fukada, E. Ohki, M. Muroaka, *J. Chem. Research* 1979, 212.
147. W. Hanefeld, G. Glaeske, *Liebigs Ann. Chem.* 1981, 1388.
148. Wella AG, DE 3 304 871-A, 1983; *Chem. Abstr.* **101** (1984) 211 158.
149. H. Hanefeld, G. Glaeske, H. J. Staude, *Arch. Pharm. (Weinheim Ger.)* **315** (1982) 103.
150. W. Hanefeld, E. Bercin, *Arch. Pharm. (Weinheim Ger.)* **314** (1981) 413.
151. W. Hanefeld, E. Bercin, *Arch. Pharm.* (*Weinheim Ger.*) **318** (1985) 848. Wella AG, DE 3 403 147, 1984; *Chem. Abstr.* **104** (1986) 34 095.
152. B. F. Goodrich Co., US 2 440 095, 1944 (J. E. Jansen). P. L. Ovechkin, L. A. Ignatova, B. V. Unkovskii, *Chem. Heterocycl. Comp. (Engl. Transl.)* 1971, 882.
153. K. Bodendorf, *J. Prakt. Chem.* **126** (1930) 233. P. Kristian, J. Bernat, *Tetrahedron Lett.* **1968,** 679. Union Carbide Corp., US 2 838 389, 1952. Henkel, DE-AS 1 284 043, 1967; *Chem. Abstr.* **70** (1969) 87 861. A. Rieche, D. Martin, W. Schade, *Arch. Pharm.* (*Weinheim Ger.*) **296** (1963) 770. Geigy, FR 6521 M 1967; *Chem. Abstr.* **74** (1971) 88 072.
154. A. Rieche, G. Hilgetag, A. Martini, O. Nejedly, J. Schlegel, *Arch. Pharm. (Weinheim Ger.)* **293** (1960) 957.
155. Bayer, DE-AS 1 120 801, 1957.
156. T. Nakai, Y. Ueno, M. Okawara, *Tetrahedron Lett.* 1967, 3831, 3835.
157. T. Nakai, M. Okawara, *Bull. Chem. Soc. Jpn.* **43** (1970) 1864; **43** (1970) 3882, 3528. K. Hartke, E. Schmidt, M. Castillo, J. Bartulin, *Chem. Ber.* **99** (1966) 3268.
158. C. Polycarpe, E. Torreilles, L. Giral, A. Babeau, N. H. Tinh, H. Gasparoux, *J. Heterocycl. Chem.* **21** (1984) 1741.
159. D. J. Rowe, C. D. Garner, J. A. Joule, *J. Chem. Soc. Perkin Trans. I,* 1985, 1907.
160. BASF, BE 660 941, 1965; *Chem. Abstr.* **64** (1966) 3364 H. H. Eilingsfeld, L. Möbius, *Chem. Ber.* **98** (1965) 1293.
161. R. L. N. Harris, *Tetrahedron Lett.* 1970, 5217.
162. *Houben Weyl,* **E 4,** 853. Story Chem. Corp., US 3 923 852, 1974; *Chem. Abstr.* **84** (1974) 58 620.
163. G. Btotny, *Liebigs Ann. Chem.* 1982, no. 10, 1927–1932.
164. A. Danopoulos, M. Avouri, S. Paraskewas, *Synthesis* 1985, 682.
165. J. v. Braun, K. Weissbach, *Ber. Dtsch. Chem. Ges.* **63** (1930) 2846.
166. Berezniki Polymer, SU 1 154 275-A, 1983; *Chem. Abstr.* **103** (1985) 70 954.
167. E. T. Hansen, J. J. Petersen, *Synth. Commun.* **14** (1984) 1275.
168. Farma Khim, Sofia, BE 868 745, 1978 (A. Georgiew, Kh. Dasholov, V. Mikhailov, K. Konstantinova); *Chem. Abstr.* **90** (1979) 168 327.
169. Research Institute for Medicine and Chemistry, CH 467 225, 1963.
170. T. Otsu, A. Kuryama, *Polym. J.* **17** (1985) 97. T. Otsu et al., *Polym. Bull.* **7** (1982) no. 1, 45. T. Otsu et al., *Macromolecules* **19** (1986) 287.
171. Wako Pure Chemical Industries, Ltd., JP 55/3344, 1980; *Chem. Abstr.* **94** (1981) 104 125.
172. *Ullmann,* 4th ed., **13,** 637. C. F. Martin, T. O. Martin Jr., US 4 339 506, 1981; *Chem. Abstr.* **97** (1982) 183 807. B. Banerjee, "Thiuram-Vulkanisation von Naturkautschuk in Gegenwart von Aminen," *Kautsch. Gummi. Kunstst.* **37** (1984) 21. Osaka Soda KK, JP 59 142 239, 1983; *Chem. Abstr.* **102** (1985) 114 936. Japan Synthetic Rubber, JP 60 047 040, 1983; *Chem. Abstr.* **103** (1985) 7627. Du Pont, JP 56 122 812, 1980; *Chem. Abstr.* **96** (1982) 70 243. P. K. Bandyopadhyay, S. Banerjee, *Kautsch. Gummi, Kunstst.* **32** (1979) 961. Bando Chem. Ind., JP 59 124 942, 1982; *Chem. Abstr.* **101** (1984) 231 782. Japan Synthetic Rubber, JP 60 047 040, 1983; *Chem. Abstr.* **103** (1985) 7627. Asahi Chemical, JP 59 197 449, 1983; *Chem. Abstr.* **102** (1985) 133 696. Osaka Soda KK, JP 59 142 239, 1983; *Chem. Abstr.* **102** (1985) 114 936. Uniroyal Inc., EP 93 500, 1983 (A. E. Crepeau); *Chem. Abstr.* **100** (1984) 35 959. W. R. Grace & Co., US 4 348 515, 1980, 1982 (C. R. Morgan); *Chem. Abstr.* **97** (1982) 199 101.
173. J. Kelm, D. Gross, "Untersuchung über Zersetzungsprodukte von Vulkanisationsbeschleunigern," *Rubber Chem. Technol.* **58** (1985) 37.

174. Japan Synthetic Rubber Co. Ltd., GB 2 092 604, 1982 (Y. Nakanishi); NL 82/341, 1981; *Chem. Abstr.* **97** (1982) 183 428.
175. M. M. Das, D. K. Basu, A. K. Chaudhuri, *Kautsch. Gummi, Kunstst.* **36** (1983) 569. R. D. Taylor, *Rubber Chem. Technol.* **47** (1974) 900. W. Hofmann, *Kautsch. Gummi, Kunstst.* **36** (1983) 1044. N. G. Chiew Sum, *J. Rubber Res. Inst. Malays.* **29** (1981) 146.
176. Bayer, DE-OS 2 227 338, 1972 (R. Schubart, U. Eholzer, E. Roos, Th. Kempermann); *Chem. Abstr.* **80** (1974) 109 596.
177. D. M. Smith, *Br. J. Ind. Med.* **41** (1984) 362. J. A. Ruddick, W. H. Newsome, L. Nash, *Teratology* **13** (1976) 263. Du Pont, *Material Safety Data Sheet* (NA-22 F). Bayer, *Sicherheitsdaten-Blatt zu Vulkacit NPV/C.* Du Pont, NA-22 Handling Precautions and Toxicity. M. Bruze, S. Fregert, *Contact Dermatitis* **9** (1983) 208.
178. G. Eisenbrand, R. Preussmann, B. Spiegelhalder, DE 3 243 141, 1982; *Chem. Abstr.* **101** (1984) 132 259. Sumitomo, EP 184 179, 1985 (A. Okamoto, T. Yamaguchi, H. Okamura, E. Okino).
179. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* Springer Verlag, Berlin, Heidelberg, New York, vol. 2, p. 4 (1970); vol. 4, p. 120 (1977); vol. 6, p. 390 (1981).
180. K.-H. Büchel: *Pflanzenschutz- und Schädlingsbekämpfungsmittel,* Thieme, Stuttgart 1977. N. M. Golyshin, *Zashch. Rast. (Moscow)* **8** (1979) 31; *Chem. Abstr.* **91** (1979) 205 502. K. Yoneyama, *Nippon Noyaku Gakkaishi* **6** (1981) 452; *Chem. Abstr.* **96** (1982) 157 159. O. A. Korotkova, *Khim. Sel'sk Khoz.* **12** (1974) 869; *Chem. Abstr.* **82** (1975) 81 240. T. Ragemath, *Pesticides* **12** (1978) XIX; *Chem. Abstr.* **89** (1978) 210 231.
181. I. Nitsche, J. Siemrova, K. Ballschmiter, F. Selenka, *Environ. Qual. Saf. Suppl.* **3** (1975) (Pesticides) 292.
182. R. Engst, W. Schnaak: *Pesticides,* Thieme Verlag, Stuttgart 1975, p. 62. W. R. Lyman, R. J. Lacoste: *Pesticides,* Thieme Verlag, Stuttgart 1975, p. 67. A. Kaars Sijpesteijn, J. W. Vonk: *Pesticides,* Thieme Verlag, Stuttgart 1975, p. 57.
183. L. Fishbein, *J. Toxicol. Environ. Health* **1** (1975) 713. N. Shindo, *Shokubutsu Bocki,* **30** (1976) 107; *Chem. Abstr.* **85** (1976) 187 557.
184. R. Engst, W. Schnaak, *Environ. Qual. Saf. Suppl.* **3** (1975) (Pesticides) 62. M. Sh. Vekshtein, M. A. Klisenko, *Probl. Anal. Khim.* **2** (1972) 21; *Chem. Abstr.* **79** (1973) 14 277. H. M. Dekhuijzen, J. W. Vonk, A. Kaars Sijpesteijn, *Pestic. Terminal Residues Invited Pap. Int. Symp.* 1971, 233. *Pure Appl. Chem.* **49** (1977) 675; *Chem. Abstr.* **87** (1977) 195 207.
185. Food and Agriculture Organization, *WHO Pestic. Residues Ser.* **4** (1975) 545; *Chem. Abstr.* **84** (1976) 103 877.
186. A. R. C. Hill, J. W. Edmunds, *Anal. Proc. (London)* **19** (1982) 433.
187. G. Pfeifer, *Mag. Kem. Lapja* **8** (1983) 333.
188. ICI, GB 916 204, 1960; GB 916 205, 1960.
189. E. Mutschler: *Arzneimittelwirkungen,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1981, p. 624.
190. S. Ebel: *Synthetische Arzneimittel,* Verlag Chemie, Weinheim 1979, p. 571.
191. Banyu Pharmaceutical Co., Ltd., NL 7 207 077, 1971;DE 2 225 482, 1971; *Chem. Abstr.* **78** (1973) 58 249.
192. M. V. Korablev, N. M. Kurbat, *Farmakol .*, Toksikol. (Moscow) 40 (1977) 230; *Chem. Abstr.* **86** (1977) 182 718.
193. Newport Pharmaceuticals International, Inc., EP 58 857, 1982 (G. E. Renoux, M. J. Renoux); *Chem. Abstr.* **97** (1982) 222 930.
194. G. Renoux, *Trends Pharmacol. Sci.* **2** (1981) 248.
195. Kumiai Chemical Industry Co., Ltd., JP 57/46 916, 1982; *Chem. Abstr.* **97** (1982) 28 602.
196. M. Gh. Alwafaie, GB 2 038 626, 1980; *Chem. Abstr.* **94** (1981) 90 344.
197. M. Morioka, JP 54/12 674, 1974; *Chem. Abstr.* **92** (1980) 135 423.
198. Colgate-Palmolive Co., US 4 007 281, 1966, 1977 (R. A. Baumann); *Chem. Abstr.* **86** (1977) 177 317. US 3 956 479, 1966, 1976; *Chem. Abstr.* **85** (1976) 83 229.
199. Wella AG, DE 3 309 400, 1983; *Chem. Abstr.* **102** (1985) 62 252.
200. Tanabe Seiyaku Co. Ltd., EP 165 017, 1985 (I. Iijiima, Y. Seiga, T. Miyagishima, Y. Matsuka, M. Matsumato).
201. W. O. Foye, D. H. Kay, P. R. Amin, *J. Pharm. Sci.* **57** (1963) 1793. C. van der Meer, D. W. Bekkum, *Int. J. Radiat. Biol.* **1** (1959) 5. J. R. Piper et al., *J. Med. Chem.* **9** (1966) 911.
202. M. J. Allalunis-Turner, J. D. Chapman, *Int. J. Radiat. Oncol. Biol. Phys.* **10** (1984) 1569.
203. J. H. Barnes, M. Fatone, G. F. Esslemont, L. Andrien, E, Bargy, *Eur. J. Med. Chem. Chim. Ther.* **10** (1975) 619.

204. W. O. Foye, D. H. Kay, *J. Pharm. Sci.* **57**(1968) 345.
205. J. R. Piper, C. R. Springfield jr., Th. P. Johnston, *J. Med. Chem.* **9** (1966) 911.
206. J. Barnes, G. Esslemont, P. Holt, *Makromol. Chem.* **176** (1975) 275.
207. W. O. Foye, R. S. F. Chu, K. A. Shah, W. H. Parsons, *J. Pharm. Sci.* **60** (1971) 1839.
208. T. Nakagawa, Y. Fujiwara, *J. Appl. Polym. Sci.* **20** (1976) 753.
209. Canon KK, JP 60 195 731-A, 1984.
210. TDK Corp., JP 60 201 988-A, 1984; *Chem. Abstr.* **85** (1976) 126 945.
211. Asahi Chemical Industry Co., Ltd., JP 59/135 187, 1984; *Chem. Abstr.* **102** (1985) 176 585. Mitsubishi Paper Mills, Ltd., JP 59/41 296, 1984; *Chem. Abstr.* **100** (1984) 219 119. Honshu Paper Co., Ltd., JP 56/148 587, 1981; *Chem. Abstr.* **97** (1982) 14 828. Mitsubishi Electric. Corp., JP 55/39340, 1980 (K. Enmanji, K. Takahashi, T. Kitagawa); *Chem. Abstr.* **93** (1980) 85 201.Canon KK, JP 51/114 124, 1976 (T. Ohta, S. Togano); *Chem. Abstr.* **89** (1978) 51 405.
212. H. W. Reinhard, O. Riessner, H. J. Smolenski, W. Walther, DD 123 282, 1976; *Chem. Abstr.* **87** (1977) 60 796. Orchard Paper Co., BE 631 000, 1962.
213. H. Mifune, S. Takada, Y. Akimura, S. Hirano,DE-OS 2 941 428, 1980; *Chem. Abstr.* **94** (1981) 9957. S. Hirano, Y. Takagi, DE-OS 3 203 661, 1982; *Chem. Abstr.* **99** (1983) 46 004.
214. Fuji Photo Film Co., Ltd., JP 55/151 635, 1980; *Chem. Abstr.* **94** (1981) 200 769.
215. Fuji Photo Film Co., Ltd., JP 58/95 347, 1983; *Chem. Abstr.* **101** (1984) 201 327.
216. Mita Industrial Co., Ltd., EP 98 089, 1984 (K. Nakatani, N. Miyakawa, T. Higashiguchi, T. Edo); *Chem. Abstr.* **100** (1984) 129 871.
217. Fuji Photo Film KK, DE 3 534 527, 1984; *Chem. Abstr.* **105** (1986) 105 790.
218. Canon KK, JP 60 196 775, 1984; *Chem. Abstr.* **104** (1985) 99 511.
219. Y. Ohba, JP 50/63 938, 1975; *Chem. Abstr.* **83** (1975) 170 880.
220. Fuji Photo Film Co., Ltd., JP 55/26 506, 1980 (M. Yamada, I. Ito, K. Watase); *Chem. Abstr.* **93** (1980) 85 147. Konishiroku Photo Industry Co., Ltd., JP 52/20 832, 1977 (K. Sakamoto, I. Fushiki, S. Koposhi, E. Sakamoto); *Chem. Abstr.* **87** (1977) 175 640.
221. Canon KK, JP 50/99 720, 1975 (T. Minegishi); *Chem. Abstr.* **84** (1976) 114 193.
222. Canon KK, DE-OS 2 537 935, 1976 (N. Hasegawa, H. Kobayashi, J. Endo, K. Kinjo); *Chem. Abstr.* **85** (1976) 184 825.
223. Canon KK, US 4 245 033, 1974 (T. Eida, I. Endo); *Chem. Abstr.* **94** (1981) 183 422.
224. Japan Synthetic Rubber Co., Ltd., JP N 54/31 727, 1977 (Y. Hosaka, Y. Haruta, M. Kurokawa, K. Harada); *Chem. Abstr.* **91** (1979) 166 399. Pittsburgh Plate Glass Co., US 2 673 151, 1954 (H. L. Gerhart).
225. Agency Ind. Sci. Tech., JP 60 076 740, 1983; *Chem. Abstr.* **103** (1985) 96 392. Agency Ind. Sci. Tech., JP 60 076 735, 1983; *Chem. Abstr.* **103** (1985) 79 496.
226. Konishiroku Photo Industry Co., Ltd., DE 2 734 335, 1977 (M. Fujiwhara, S. Matsuo, T. Masukawa, Y. Kaneko, M. Kawasaki); *Chem. Abstr.* **88** (1978) 161 442.
227. Fuji Photo Film KK, JP 55 090 943, 1979; *Chem. Abstr.* **94** (1981) 74 729.
228. Fuji Photo Film Co., Ltd., JP 50/40 645, 1973; *Chem. Abstr.* **83** (1975) 211 256.
229. Horizons, US 4 018 604, 1976, 1977 (P. L. Bachmann); *Chem. Abstr.* **86** (1977) 198 032. Asahi Chemical Industry Co., Ltd., JP 51/58 327, 1976 (T. Shiga, I. Kikuchi, M. Yoshino); *Chem. Abstr.* **85** (1976) 200 565.
230. Fuji Photo Film KK, JP N 55 026 506, 1978; *Chem. Abstr.* **93** (1980) 85 147. Konishiroku, JP N 5 220 832, 1975; *Chem. Abstr.* **87** (1977) 175 640.
231. Agfa, US 2 124 159, 1934, 1938 (E. Weyde).
232. Kodak, US 2 453 346, 1945, 1948 (H. D. Russel).
233. Eastman Kodak Comp., US 3 144 336 1961, 1964 (A. H. Herz).
234. Du Pont, US 3 505 069, 1966, 1970 (J. H. Bigelow); *Chem. Abstr.* **73** (1970) 40 458.
235. Konishiroku Photo, JP 22 943/69, 1966.
236. Konishiroku Photo, JP 22 944/69, 1966.
237. Du Pont, US 3 597 210, 1968, 1971 (J. H. Bigelow); *Chem. Abstr.* **75** (1971) 103 662.
238. Fuji Photo KK, DE 3 534 527, 1984; *Chem. Abstr.* **105** (1986) 105 790.
239. Canon, DE-OS 2 117 044, 1971 (E. Inoue, T. Yamase); *Chem. Abstr.* **76** (1972) 8 953.
240. Agfa-Gevaert, DE-OS 2 043 205, 1970 (W. Himmelmann, A. v. König, F. Moll, W. Saleck); *Chem. Abstr.* **77** (1972) 54 873. Agfa-Gevaert, DE-OS 2 044 622, 1970 (W. Himmelmann, A. v. König, F. Moll, D. Rücker, W. Saleck); *Chem. Abstr.* **77** (1972) 95 349.
241. Continental Can., DE-OS 2 505 448, 1975 (G. Pasternack, T. P. Johndal); *Chem. Abstr.* **84** (1976) 46 258.

242. Du Pont, DE-OS 2 651 941, 1976 (R. K. Blake); *Chem. Abstr.* **87** (1977) 109 399.
243. Oriental Photo, JP 7 729 611, 1972.
244. Canon, US 3 718 467, 1971, 1973 (E. Inoue, T. Yamase); *Chem. Abstr.* **78** (1973) 167 050.
245. Fuji, JP 5 040 645, 1973; *Chem. Abstr.* **83** (1975) 211 256.
246. Oriental Photo Industrial Co., Ltd., JP 51/111 756, 1976 (T. Koura); *Chem. Abstr.* **87** (1977) 58 112.
247. Toyota Cent. Res. & Dev., JP 60/110 721-A, 1983; *Chem. Abstr.* **104** (1986) 20 389.
248. Federal Mogul Corp., US 4 524 185, 1983 (R. F. Hinderer); *Chem. Abstr.* **103** (1985) 124 799.
249. Unie van Kunstmestfab. BV, EP 116 988, 1983; *Chem. Abstr.* **101** (1984) 173 936.
250. Y. Kato, O. Katsuki, *Kogyo Yosui* **290** (1982) 37; *Chem. Abstr.* **98** (1983) 95 142.
251. Olin. Corp., US 4 518 760, 1984 (M. R. Smith, H. B. Cochran); *Chem. Abstr.* **103** (1985) 161 294.
252. Phillips Petroleum Co., US 4 554 108, 1983; *Chem. Abstr.* **104** (1986) 190 296. Phillips Petroleum Co., US 4554-068-A, 1983; *Chem. Abstr.* **104** (1986) 190 296. Bresson CR, US 4514-293-A, 1984; *Chem. Abstr.* **103** (1985) 40 425.
253. J. J. Steggerda, J. A. Cras, J. Willemse, *Recl. Trav. Chim. Pays Bas* **100** (1981) 41.
254. A. M. Bond, R. L. Martin, *Coord. Chem. Rev.* **54** (1984) 23. V. F. Toropova, G. K. Budnikov, N. A. Ulakhovich, *Elektrodyne Protsessy Adsorbtsiya* **2** (1978) 3; *Chem. Abstr.* **91** (1979) 183 779. R. L. Martin, *Coord. Chem. – Invited Lect. Int. Conf. 20th* 1979, 255; *Chem. Abstr.* **94** (1981) 75 848.
255. J. Garaj, *Proc. Conf. Coord. Chem. 9th* 1983, 83; *Chem. Abstr.* **99** (1983) 61 800.
256. J. O. Hill, R. J. Magee, *Rev. Inorg. Chem.* **3** (1981) 141.
257. J. Garaj, V. Kettmann, *Int. Semin. Cryst. Chem. Coord. Organomet. Compd. [Proc.], 3rd* 1977, 166; *Chem. Abstr.* **93** (1980) 36 047.
258. Mitsui-Nisso Urethane KK, *Chem. Abstr.* **96** (1982) 201 081. Mitsui-Nisso Urethane KK, JP 56 135 545, 1980; *Chem. Abstr.* **96** (1982) 53 594. Mitsui-Nisso Urethane KK, JP 57 159 846, 1981; *Chem. Abstr.* **98** (1983) 127 531.
259. ICI, FR 1 559 120, 1968; *Chem. Abstr.* **71** (1969) 113 737. A. O. Fitton, A. Rigby, R. J. Hurlock, *J. Chem. Soc. C* 1969, 230. Uniroyal Inc., DE-OS 1 815 221, 1968; *Chem. Abstr.* **72** (1970) 3218.
260. E. G. Kolawole, *J. Appl. Polym. Sci.* **27** (1982) 3437.
261. A. Al-Malaika et al., *J. Appl. Polym. Sci.* **30** (1985) 789.
262. G. Scott, D. Gilead , DE-OS 2 839 867, 1979; *Chem. Abstr.* **90** (1979) 205 265.
263. Du Pont, US 4 452 879, 1983 (M. G. Fickes, P. F. Warfield); *Chem. Abstr.* **101** (1984) 63 698.
264. Japan Synthetic Rubber Co., Ltd., JP 54/31 727, 1977 (Y. Hosaka, Y. Haruta, M. Kurokawa, K. Harada); *Chem. Abstr.* **91** (1979) 166 399.
265. New England Nuclear Corp., US 4 411 881, 1982 (N. R. Tzodikov); *Chem. Abstr.* **100** (1984) 39 659.
266. A. T. Polishuk, H. H. Farmer, *NLGI Spokesman* **43** (1979) 200; *Chem. Abstr.* **92** (1980) 44 111. Idemitsu Kosan KK, JP 60 084 394, 1983; *Chem. Abstr.* **103** (1985) 180 736. Nitto Chem. Ind. KK, JP 61 028 694, 1984; *Chem. Abstr.* **105** (1986) 100 209. Lengd. Petro Oil WKS, SU 1 065 396, 1982; *Chem. Abstr.* **101** (1984) 38 146. General Sekiyu KK, JP 59/199 796-A, 1983; *Chem. Abstr.* **102** (1985) 169 528. Pennwalt Corp., EP 128 262-A, 1983; *Chem. Abstr.* **102** (1985) 116 430.
267. T. Sakurai et al., *Bull. Jpn. Pet. Inst.* **13** (1971) 243.
268. VEB Chemiekombinat Bitterfeld, DD 214 855, 1983; *Chem. Abstr.* **103** (85) 106 382.
269. Rhône-Poulenc Specialités Chimiques, EP 141 685, 1984 (G. Fabre, P. Perrin); FR 2 550 541, 1983; *Chem. Abstr.* **103** (1985) 55 277.
270. Denki Kagaku Kogyo KK, JP 59/053 560, 1982; *Chem. Abstr.* **101** (1984) 112 538.
271. K. Gleu, R. Schwab, *Angew. Chem.* **62** (1950) 320. E. Eckert, *Z. Anal. Chem.* **155** (1957) 23. H. Bode, *Z. Anal. Chem.* **143** (1954) 182; **144** (1955) 90, 166. H. Bode, *Z. Anorg. Allg. Chem.* **289** (1957) 207.
272. B. Ch. Verma, S. Kumar, *Rev. Anal. Chem.* **4** (1978) 61.
273. R. J. Magee, *Rev. Anal. Chem.* **1** (1973) 335. R. J. Magee, J. O. Hill, *Rev. Anal. Chem.* **8** (1985) 5.
274. W. D. McFarlane, *Biochem. J.* **26** (1932) 1022. B. Eisler, K. G. Rosdahl, R. Theorell, *Biochem. Z.* **285** (1936) 76. M. Picotti, G. Baldari, *Mikrochim. Acta* **30** (1947) 77.
275. R. M. Barnes, *Biol. Trace Chem. Res.* **61** (1984) 93.
276. R. Neeb, *Pure Appl. Chem.* **54** (1982) 847.

277. Ch. Fu, B. Zuo, *Fen Hsi Hua Hsueh* **9** (1981) 737; *Chem. Abstr.* **97** (1982) 84 123.
278. M. V. Korablev, *Farmakol. Toksikol. (Moscow)* **32** (1969) 356; *Chem. Abstr.* **71** (1969) 68 928.
279. A. Rolandi, E. De Marinis, M. De Caterina, *Mutat. Res.* **135** (1984) 193.
280. O. P. Chepinoga, *Tr. S'ezda Gig. Ukr. SSR. 8th 1970* 1971, 209; *Chem. Abstr.* **77** (1972) 70 953.
281. I. Schuphan, Y. Segall, J. D. Rosen, J. E. Casida, *ACS Symp. Ser.* **158 (Sulfur Pestic. Action Metab.), 65** (1981); *Chem. Abstr.* **95** (1981) 126 730.
282. A. P. Shitskova, R. A. Ryanzanova; *Vsestoronnii Anal. Okruzh. Prir. Sredy, 3rd* **1978,** 96; *Chem. Abstr.* **93** (1980) 39 117.
283. A. Korhonen, K. Hemminki, H. Vainio, *Teratog., Carcinog., Mutagen.* **3** (1983) 163. A. Korhonen, K. Hemminki, H. Vainio, *Scand. J. Work Environ. Health* **9** (1983) 115. A. Korhonen, K. Hemminki, H. Vainio, *Scand. J. Work Environ. Health* **8** (1982) 63. L. Ivanova-Chemishanska, T. Petrova-Vergieva, E. Mirkova, *Eksp. Med. Morfol.* **14** (1975) 29; *Chem. Abstr.* **83** (1975) 158 773. F. B. Kuzan, K. V. Prahlad, *Poult. Sci.* **54** (1975) 1054; *Chem. Abstr.* **84** (1976) 70 028.
284. M. Dirimanov, A. Mateeva, *Nauchni Tr. Vissh Selskostop. Inst. Vasil Koralov, Plovdiv* **24** (1979) 125; *Chem. Abstr.* **94** (1981) 1035. H. V. Ghate, L. Mulherkar, *Indian J. Exp. Biol.* **18** (1980) 1040; *Chem. Abstr.* **93** (1980) 232 143. H. V. Ghate, *Toxicol. Lett.* **19** (1983) 253.
285. L. Vicari, G. De Dominicis, M. Vito, C. Placida, E. De Marinis, *Boll. Soc. Ital. Biol. Sper.* **61** (1985) 271. H. V. Ghate, J. K. Pal, *Curr. Sci.* **53** (1984) 662. R. D. Short, J. L. Ninor, T. M. Unger, Report EPA-600/1-80-017, Order no. PB80-181 175, 28 pp., *Chem. Abstr.* **93** (1980) 180 594. E. Arias, T. Zavanella, *Bull. Environ. Contam. Toxicol.* **22** (1979) 297. N. Chernoff, R. J. Kavlock, E. H. Rogers, D. B. Carver, S. Murray, *J. Toxicol. Environ. Health* **5** (1979) 821. K. S. Larsson, Cl. Arnander E. Cekanova, M. Kjellberg, *Teratology* **14** (1976) 171. M. A. Pilinskaya, A. I. Kurinnyi, T. I. Kondratenko, *Mol. Mekh. Genet. Protsessov* **295** (1976); *Chem. Abstr.* **87** (1977) 34 142. T. Petrova-Vergieva, *Khig. Zdraveopaz.* **19** (1976) 435; *Chem. Abstr.* **87** (1977) 16 806.
286. D. J. Clegg, K. Khera, *Pestic. Environ. Contin. Controversy Pap. Inter Am. Conf. Toxicol. Occup. Med., 8th* 1973, 276; *Chem. Abstr.* **84** (1976) 145 460.
287. A. Rannug, U. Rannug, *Chem. Biol. Interact.* **49** (1984) 329. V. A. Kiryushiu, *Gig. Sanit.* 1975, no. 9,43; *Chem. Abstr.* **83** (1975) 202 555. Y. Shirasu, M. Moriya, H. Tezuka, S. Teramoto, T. Ohta, T. Inoue, *Environ. Sci. Res.* **31** (1984) 617. M. Moriya, T. Ohta, K. Watanabe, T. Miyazawa, K. Kato, Y. Shirasu, *Mutat. Res.* **116** (1983) 185. L. Fishbein, *Environ. Mutagens Carcinog. Proc. Int. Conf. 3rd* 1981, 371; *Chem. Abstr.* **98** (1983) 12 428. Y. Shirasu, M. Moriya, K. Kato, A. Furuhashi, T. Kada, *Mutat. Res.* **40** (1976) 19. M. E. Gerzoni, L. Del Cupolo, I. Ponti, *Riv. Sci. Tecnol. Alimenti Nutr. Um.* **6** (1976) 161. H. C. Sikka, P. Florczyk, *J. Agric. Food Chem.* **26** (1978) 146. F. De Lorenzo, N. Staiano, L. Silengo, R. Cortese, *Cancer Res.* **38** (1978) 13. V. A. Kiryushin, *Gig. Aspekty Okruzh. Zdorov'ya Naseleniya* **196;** *Chem. Abstr.* **89** (1978) 85537. M. I. Kulik, *Genet. Posledstviya Zagryaz. Okruzh. Sredy* **177;** *Chem. Abstr.* **88** (1978) 131 630. M. Moriya, T. Ohta, K. Watanabe, Y. Watanabe, F. Sugiyama, T. Miyazawa, Y. Shirasu, *Cancer Lett. (Shannon Irel.)* **7** (1979) 325. V. Vasudev, N. B. Krishnamurthy, *Mutat. Res.* **77** (1980) 189; *Chem. Abstr.* **92** (1980) 122 984. R. A. Ryazanova, T. V. Gafurova, *Gig. Sanit.* 1980, no. 1,80; *Chem. Abstr.* **92** (1980) 105 449. M. Tsoneva, V. Georgieva, Chan Chi Lien, *Genet. Sel.* **12** (1979) 118; *Chem. Abstr.* **93** (1980) 144 099. J. D. Rosen, I. Schuphan, Y. Segall, J. E. Casida, *J. Agric. Food Chem.* **28** (1980) 880. Chii-hing Jeang, Gwo-Chen Li, *K'o Hsueh Fa Chan Yueh K'an* **8** (1980) 551; *Chem. Abstr.* **93** (1980) 180 338. Y. Kato, M. Tanaka, K. Umezawa, S. Takayama, *Toxicol. Lett.* **7** (1980) 125. F. Decloitre, G. Hamon, *Mutat. Res.* **79** (1980) 185. K. Hemminki, K. Falck, H. Vainio, *Arch. Toxicol.* **46** (1980) 277. P. Principe, E. Dogliotti, M. Bignami, R. Crebelli, E. Falcone, M. Fabrini, G. Conti, P. Comba, *J. Sci. Food Agric.* **32** (1981) 826. M. Zdzienicka, M. Zielenska, M. Trojanowska, T. Szymczyk, *Bromatol. Chem. Toksykol.* **15** (1982) 83; *Chem. Abstr.* **98** (1983) 29 291.
288. C. J. van Leeuwen, J. L. Maas Diepeveen, G. Niebeek, W. H. A. Vergouw, P. S. Griffioen, M. W. Luijken, *Aquat. Toxicol.* **7** (1985) 145. L. Fishbein, *J. Toxicol. Environ. Health* **1** (1976) 713.

289. C. J. van Leeuwen, F. Moberts, G. Niebeek, *Aquatic Toxicology* **7** (1985) 165. *Carbamates, Thiocarbamates and Carbazides, IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man,* **vol. 12,** Lyon 1976.
290. A. Hedenstedt, U. Rannug, C. Ramel, C. A. Wachtmeister, *Mutat. Res.* **68** (1979) 313. X. You, Y. Zhou, Y. Hu, *Huan Ching K'o Hsueh* **3** (1982) 39; *Chem. Abstr.* **98** (1983) 84 705. M. Donner, K. Husgafvel-Pursiainen, D. Jenssen, A. Rannug, *Scand. J. Work Environ. Health* **9** (1983) (Suppl. 2), 27.
291. A. Hedenstedt, *SGF Publ.* **57** (1981) 39; *Chem. Abstr.* **97** (1982) 105 398.
292. M. Donner, *SGF Publ.* **57** (1981) 47; *Chem. Abstr.* **97** (1982) 86 819.
293. S. K. Mann, *Environ. Exp. Bot.* **17** (1977) 7.
294. A. Rannug, U. Rannug, C. Ramel, *Prog. Clin. Biol. Res.* **141** (1984) 407. S. De Flora, P. Zanacchi, A. Camoirano, C. Bennicelli, G. S. Badolati, *Mutat. Res.* **133**(1984) 161. M. D. Waters et al., *Basic Life Sci.* **21** (1982) 275.
295. J. Kelm et al., *Kautsch. Gummi Kunstst.* **36** (1983) 274.
296. H. Druckrey, R. Preussmann, S. Ivankovic, D. Schmähl, *Z. Krebsforsch.* **69** (1967) 103. B. L. Pool, B. Bertram, M. Wiessler, *Carcinogenesis (London)* **3** (1982) 563.

Dithionites → Sulfites, Thiosulfates, and Dithionites

Diuretics

Max Hropot, Hoechst AG, Hoechst Marion Roussel, Frankfurt/Main, Federal Republic of Germany

Hans-Jochen Lang, Hoechst AG, Hoechst Marion Roussel, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Carbonic Anhydrase Blockers . . . 2
3. Saluretics 3
3.1. Thiazide-like Diuretics 3
3.1.1. Benzothiadiazine Derivatives 3
3.1.2. Other Saluretics with Thiazide-like Activity 4
3.2. Loop Saluretics 4
3.2.1. Furosemide 4
3.2.2. Other Loop Diuretics 5
4. Potassium-Retaining Diuretics . . . 7
4.1. Aldosterone Antagonists 7
4.2. Other Potassium-Retaining Diuretics 7
5. Aquaretics 8
6. Adenosine A1 Receptor Antagonists 9
7. References 9

1. Introduction

Diuretics are generally defined as substances that cause an increased excretion of urine. The name saluretic or salidiuretic refers to the stimulation of salt excretion caused by these substances. Since urine always contains inorganic salts and since the total concentration of ions (predominantly sodium, potassium, chloride, and hydrogencarbonate ions) is only slightly influenced by drugs, the terms diuretic and saluretic are often used synonymously.

In the recent literature, however, it has become customary to define saluretics as substances that directly stimulate the excretion of salts by inhibiting electrolyte reabsorption in the kidney tubules, sodium being the main cation and chloride being the main anion excreted. In this article, the term saluretic is used in this sense.

Physiological Basis of Diuretic Function. The smallest functional unit of the kidney is the nephron, which can be divided into four different sections: the glomerulus (a capillary filter), the proximal tubule, the loop of Henle, and the distal tubule, which terminates in the collecting duct. Blood plasma is filtered through the glomerular capillaries (approximately 180 L of plasma per day in an adult) and enters the proximal tubule as an ultrafiltrate. In the proximal tubule 50 – 60 % of the filtrate is reabsorbed. Apart from Na^+, Cl^-, HCO_3^-, and K^+ ions, many other substances, such as glucose and amino acids, are reabsorbed. Hydrogencarbonate reabsorption in this section of the nephron is inhibited primarily by carbonic anhydrase blockers.

Approximately 30 % of the glomerular filtrate is reabsorbed in the following loop of Henle, which is divided into an ascending and a descending limb. The thick ascending limb, in which electrolytes are transported through a water impermeable membrane, is especially important for diuretic activity. An active $Na^+ 2\,Cl^- K^+$ cotransport takes place, which is inhibited by loop diuretics.

Only ca. 10 % of the glomerular filtrate is reabsorbed in the distal tubule. Sodium and chloride ions are actively reabsorbed against a concentration gradient. Diuretics having thiazide-like activity inhibit ion transport in the early part of the distal tubule. In the late part of the distal tubule and in the collecting ducts, where final adjustments in the composition of urine occur, NaCl reabsorption is accompanied by secretion of potassium and hydrogen ions. This formal sodium – potassium exchange is prevented by potassium-retaining diuretics or aldosterone antagonists.

Potassium secretion in the late part of the distal tubules and in the collecting ducts is stimulated by an accelerated tubular flow rate and an increased delivery of sodium. Consequently, diuretics that act in an earlier part of the nephron cause increased potassium secretion.

Diuretics produce a loss of electrolytes and fluids; this induces compensation

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_029

phenomena, such as activation of the renin – angiotensin – aldosterone system. A long-term diuretic therapy, therefore, can result in a marked decline in the effect produced.

Fields of Application. Diuretics are administered when an increased excretion of extracellular fluids is desired. They are used to treat acute forms of edema, like cerebral and pulmonary edema, and to control chronic cardiac, hepatic, or renal edema or edema accompanying a toxic – allergic condition.

The discovery of the antihypertensive properties of saluretics in the early 1960s considerably broadened the scope of application of these drugs. Saluretics are especially important in the treatment of essential hypertension, which makes up approximately 80 % of all cases of hypertension. They are firmly established as being safe and reliable. Indeed, saluretics and β-adrenoceptor antagonists are the basic drugs used to treat high blood pressure and are used either alone or in combination with other antihypertensive medication, depending on the seriousness of the condition (→ Antihypertensives).

Historical Development. Until 1950, apart from plant extracts and the so-called osmotic diuretics, doctors had only two classes of substances at their disposal for the treatment of edema: the *xanthine derivatives* and related nitrogen heterocyclic compounds, primarily pyrimidine and triazine derivatives, and an extensive number of *organic mercurial compounds.* Typical examples of these groups are theophylline [*58-55-9*]

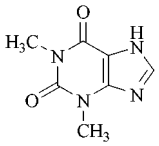

and mersalyl [*492-18-2*].

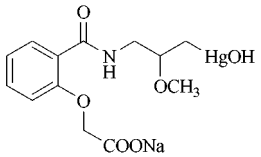

Mercurial compounds have been almost completely replaced by the sulfonamide saluretics, which are considerably better tolerated. As diuretics the xanthine drugs still on the market are not very potent; they are employed mainly because of their cardiotonic and bronchodilator activity (→ Antiasthmatic Agents, Chap. 2.2.).

After years of stagnation, diuretic research has received new impetus. Drugs like renal vasodilators (especially dopamine agonists like ibopamine [*66195-31-1*]; see also → Cardiac Glycosides and Synthetic Cardiotonic Drugs, Chap. 3.1.) [3], vasopressin antagonists [4], and the atrial natriuretic factor (ANF) and its derivatives [5] are currently being developed.

The most important diuretics currently used can be classified into three main groups, according to their site of action and mechanism of action: (1) carbonic anhydrase blockers, (2) saluretics, and (3) potassium-retaining diuretics.

2. Carbonic Anhydrase Blockers

The enzyme carbonic anhydrase catalyzes the hydration of carbon dioxide in tubular cells. Carbonic anhydrase inhibition results in a shortage of hydrogen ions and an increased excretion of sodium and hydrogencarbonate ions. This leads to alkalinization of urine, a fall in blood pH (hyperchloremic acidosis), and an increase in potassium excretion. However, the use of carbonic anhydrase inhibitors has declined. They are used mainly in the treatment of glaucoma and to some extent in epilepsy (→ Antiepileptics, Chap. 2.8.). As a result of the acidification effect, they are also employed as adjuvants in the treatment of alkalosis.

Acetazolamide [*59-66-5*], 2-acetylamino-1,3,4-thiadiazole-5-sulfonamide, $C_4H_6N_4O_3S_2$, M_r 222.2, the most important commercial preparation, was developed by Roblin and Clapp in 1950 [6]. It crystallizes from water in the form of colorless prisms, *mp* 268 – 270 °C (decomp.); it is sparingly soluble in cold water and readily soluble in 1 N sodium carbonate solution.

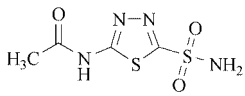

Acetazolamide is preferably administered orally in daily doses of up to 1.0 g. The diuretic

effect starts ca. 2 h after administration. After several days of treatment, however, the diuretic effect declines steadily because of increasing acidosis. The drug should then be discontinued for ca. 4 d, but it regains full activity as soon as normalization of plasma pH occurs.

The high potassium excretion is the most serious side effect caused by acetazolamide. Indeed, it is imperative that patients subjected to long-term medication be given potassium supplements (fruit juices, bananas, KCl preparations) to prevent potassium deficiency symptoms like paresthesia, muscle cramps, and drowsiness.

Synthesis. Oxidative chlorination of 2-acetylamino-5-mercapto-1,3,4-thiadiazole or its *S*-benzyl or *S*-acyl derivatives to the corresponding sulfonyl chlorides followed by reaction with ammonia [7] has proved to be an optimal method for the synthesis of acetazolamide.

Trade Names. Diamox (Cyanamid-Novalis), Glaupax (Dispersa).

3. Saluretics

Saluretics are currently the most important drugs used in the treatment of edema. They have also become increasingly important in the treatment of hypertension (→ Antihypertensives). Their structure differs from that of carbonic anhydrase blockers in that the heterocyclic ring substituted with the sulfonamide group is replaced by a phenyl moiety. They have a very weak inhibitory effect on carbonic anhydrase and, consequently, only a small influence on the excretion of HCO_3^- in the urine. Sodium chloride excretion is predominantly enhanced. There are marked differences within the group with respect to potency, efficacy, duration of action, and potassium excretion.

3.1. Thiazide-like Diuretics

Thiazide-like saluretics inhibit sodium reabsorption only in a very limited area of the nephron, namely the cortical segment of the loop of Henle and the early part of the distal tubule. In addition, a site of action in the proximal tubule has been found. The efficacy of these drugs is not very high; the duration of action is prolonged. They are especially suitable in the treatment of hypertension, either alone or in combination with other drugs.

3.1.1. Benzothiadiazine Derivatives

Chlorothiazide. This preparation, developed by Novello and Sprague [8], was the very first saluretic on the market. Chlorothiazide [*58-94-6*], 6-chloro-7-sulfamoyl-2*H*-1,2,4-benzothiadiazine 1,1-dioxide, $C_7H_6ClN_3O_4S_2$, M_r 295.7, decomposes at 345 °C. Solubility in water is ca. 0.04 % at pH 4 and 0.15 % at pH 7. The compound is readily soluble in aqueous alkalies, in which it is slowly hydrolyzed with ring opening.

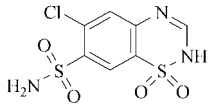

Chlorothiazide possesses one tenth of the carbonic anhydrase inhibitory effect shown by acetazolamide. In oral doses of 1 – 2 g, the drug evokes a copious salidiuresis, which reaches a maximum after 3 – 4 h. Even after long-term administration, it remains fully effective.

Synthesis. Chlorothiazide is synthesized starting from 3-chloroaniline. This compound is converted to the corresponding 4,6-disulfonamide in two steps by reaction with chlorosulfuric acid and ammonia; ring closure is effected with formic acid [9].

Trade Names. Chlotride, Saluric, Diuril, Diupres (all Merck Sharp & Dohme).

Hydrochlorothiazide. Chlorothiazide was on the market only a year when de Stevens [10] in Switzerland developed hydrochlorothiazide. This new drug proved to be of considerable therapeutic importance.

Hydrochlorothiazide [*58-93-5*], 6-chloro-3,4-dihydro-7-sulfamoyl-2*H*-1,2,4-benzothiadiazine 1,1-dioxide, $C_7H_8ClN_3O_4S_2$, M_r 297.7, crystallizes from aqueous ethanol as colorless prisms, *mp* 274 – 275 °C. It is soluble in 1 N sodium carbonate solution, in which it is gradually hydrolyzed, even at room temperature, to give the disulfonamide and formaldehyde.

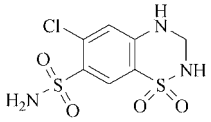

Hydrochlorothiazide is 15 – 20 times more potent than chlorothiazide and exerts an even smaller inhibitory effect on carbonic anhydrase [11]. It is much better tolerated than chlorothiazide. Combination preparations with potassium-retaining diuretics (triamterene and amiloride, see Section 4.2) are also on the market.

Synthesis. There are several methods available for the production of hydrochlorothiazide. Chlorothiazide can be reduced electrolytically or with sodium borohydride [12] or catalytically [13]. On the other hand, 3-chloro-4,6-disulfamoylaniline can be cyclized with formaldehyde [14] to give hydrochlorothiazide.

Trade Names. Di-Chlotride, Hydro-Diuril, Hydro-Saluric (all Merck Sharp & Dohme); Esidrix (Ciba); Oretic (Abbott); Dytide H (combination with triamterene, Röhm Pharma); Moduretic (combination with amiloride, Merck Sharp & Dohme).

Hydrochlorothiazide Analogues. Substituents in the 3-position or in the 2- and 3-positions can substantially increase potency. However, replacement of the chlorine atom by a trifluoromethyl group has almost no influence on potency. Some well-known analogues are listed in Table 1. Despite the large number of newer thiazine dioxide derivatives, hydrochlorothiazide remains the most important thiazide-type drug on the market. In their biological action, these chemical analogues differ from hydrochlorothiazide mainly by their lower threshold dose and, consequently, the therapeutic dose (potency); the maximum saluretic effect (efficacy), however, is not larger. The most potent compound of this series is cyclopenthiazide, which was put on the market in 1961. Its threshold dose in humans is 0.125 mg. The maximum response is obtained in therapeutic doses as low as 1.0 mg [22].

3.1.2. Other Saluretics with Thiazide-like Activity

Table 2 lists some other saluretics with thiazide-like activity. With the exception of tienilic acid (ticrynafen) all these compounds are sulfonamide derivatives. Tienilic acid differs from the other saluretics in that it causes an increased excretion of uric acid (uricosuric activity). However, this drug has been removed from the market in most countries because it is believed to be connected with hepatotoxic effects [32].

3.2. Loop Saluretics

Loop diuretics, also known as high-ceiling diuretics, act by inhibiting reabsorption of sodium chloride in a more extensive area of the nephron than the thiazide-like saluretics. They primarily act in the medullary segment of the loop of Henle and to a lesser extent in the proximal tubule. Their efficacy is much higher than that of the thiazide-like saluretics. A further characteristic is the shorter duration of action.

3.2.1. Furosemide

Furosemide [*54-31-9*], frusemide, 4-chloro-*N*-(2-furylmethyl)-5-sulfamoyl anthranilic acid, $C_{12}H_{11}ClN_2O_5S$, M_r 330.8, was the first high-ceiling diuretic on the market and is still the most important product in this group. It is sparingly soluble in water and crystallizes from ethanol or nitromethane as colorless prisms, which decompose at 208 °C. The sodium salt can be recrystallized from water; the solubility of the salt in water at room temperature is 6 %.

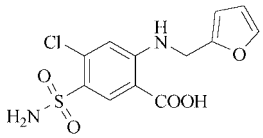

Furosemide was discovered in 1959 [33] and was put on the market in the Federal Republic of Germany in March 1964. A free carboxyl group gives it a strong hydrophilic character, which causes rapid onset of diuresis (30 – 60 min after oral administration, a few minutes after intravenous injection) and rapid elimination. The efficacy of this drug is extremely high. With respect to the maximum natriuresis produced, furosemide is 8 – 10 times more effective than

Table 1. Hydrochlorothiazide analogues

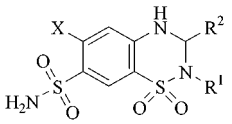

X	R^1	R^2	*Generic name* Trade name	CAS registry number	Therapeutic dose, mg
$-CF_3$	H	H	*hydroflumethiazide* [15] Naclex (Glaxo)	[*135-09-1*]	25 – 150
Cl	H	isobutyl	*buthiazide, thiabutazide* Saltucin (Boehringer Mannheim)	[*2043-38-1*]	20 – 100
$-CF_3$	H	benzyl	*bendroflumethiazide* [15] Naturetin (Squibb)	[*73-48-3*]	5 – 20
Cl	H	$-CHCl_2$	*trichlormethiazide* [17] Esmarin (E. Merck)	[*133-67-5*]	5 – 20
Cl	$-CH_3$	$-CH_2Cl$	*methylclothiazide* [18] Enduron (Abbot)	[*135-07-9*]	2 – 10
Cl	$-CH_3$	$-CH_2SCH_2CF_3$	*polythiazide* [19] Renese, Drenusil (Pfizer)	[*346-18-9*]	1 – 5
Cl	H		*cyclothiazide* [20] Anhydron (Lilly)	[*2259-96-3*]	1 – 2
Cl	H		*cyclopenthiazide* [21] Navidrex (Ciba)	[*742-20-1*]	0.5 – 1

the thiazides [34]. Hence, the therapeutic efficacy of furosemide is extremely high, especially as the drug shows practically no toxic side effects.

Synthesis. Furosemide is produced by treatment of 2,4-dichloro-5-sulfamoylbenzoic acid with excess furfurylamine in the absence of solvent at 120 – 130 °C [33], [35].

Trade Names. Lasix, Lasilix, Seguril (Hoechst).

3.2.2. Other Loop Diuretics

As shown in Table 3, loop diuretics are derived from completely different classes of chemical substances. The two sulfonamide derivatives, *bumetanide* [36] and *piretanide* [37], have a pharmacological activity profile very similar to that of furosemide, but are much more potent. Piretanide has proved to be very useful in the treatment of hypertension. *Ethacrynic acid* [38] is 5 times more effective than the thiazides; however, it has not attained particular importance, presumably because of side effects. The pyrazolone derivative *muzolimine* [39] combines the saluretic pattern of loop diuretics with a prolonged duration of action, whereas *etozolin* [40] shows reduced side effects on glucose metabolism. In 1987, shortly after the introduction of muzolimine onto the market in Germany, it was withdrawn for toxicological (polyneuropathy) reasons.

4. Potassium-Retaining Diuretics

4.1. Aldosterone Antagonists

Spironolactone [*52-01-7*] [41] and potassium canrenoate [42] [Aldactone (Searle, Boehringer Mannheim), Osyrol (Hoechst), Soldactone (Searle)] are used as aldosterone antagonists. Aldosterone stimulates sodium and water reabsorption in the late part of the distal tubule and in the collecting ducts.

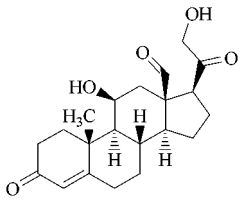

Aldosterone

Table 2. Other saluretics with thiazide-like activity

Class of substance Formula	R^1	R^2	*Generic name* Trade name	CAS registry number
Benzenedisulfamides				
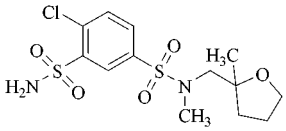			*mefruside* [23] Baycaron (Bayer)	[*7195-27-9*]
Quinazolines				
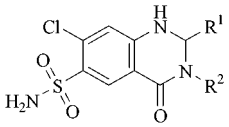	$-C_2H_5$	H	*quinethazone* [24] Aquamox	[*73-49-4*]
	$-CH_3$		*metolazone* [25] Zaroxolyn (Pennwalt, Searle)	[*17560-51-9*]
Benzamides and benzhydrazines				
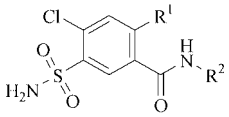	$-OH$	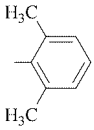	*xipamide* [26] Aquaphor (Beiersdorf)	[*17560-51-9*]
	H	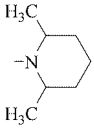	*clopamide* [27] Brinaldix (Sandoz)	[*636-54-4*]
	H	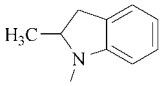	*indapamide* [28] Natrilix (Servier, Itherapia)	[*26807-65-8*]
	H	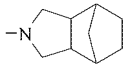	*tripamide* [26] Normonal (Eisai)	[*7308-48-2*]
Others				
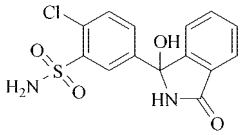			*chlorthalidone* [30] Hygroton (Ciba-Geigy)	[*77-36-1*]
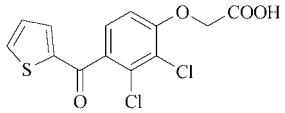			*ticrynafen* [31] Diflurex (Crephar)	[*40180-04-9*]

Aldosterone antagonists, on the other hand, enhance sodium excretion and diminish potassium secretion. After administration of these compounds only about 2 % of the filtered sodium is excreted. Aldosterone antagonists act only in the presence of aldosterone. The response to these drugs is slow in onset. The complete saluretic activity will be reached only after several days of treatment. Spironolactone (for details on synthesis, see [42]) is insoluble in water, but potassium canrenoate is water-soluble and hence can be given by parenteral administration. Both spironolactone and canrenoate are rapidly converted to the active metabolite canrenone, which has a half-life of ca. 20 h.

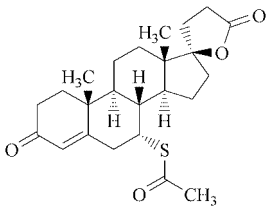

Sipronolactone

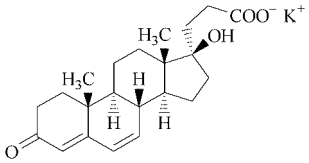

Potassium canrenoate

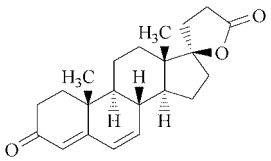

Canrenone

Aldosterone antagonists are always indicated in the treatment of edema accompanied by hyperaldosteronism. The most dangerous side effect is hyperkalemia.

4.2. Other Potassium-Retaining Diuretics

Triamterene and amiloride also decrease the excretion of potassium. However, their mechanism of action differs from that of spironolactone. They do not inhibit aldosterone; modern electrophysiological methods indicate that they block the luminal sodium binding site at the sodium – potassium exchange mechanism in the late part of the distal tubule (see also → Antihypertensives, Chap. 2.2.). Triamterene and amiloride are used mainly in combination with hydrochlorothiazide and other saluretics (see Section 3.1.1) in order to prevent diuretic induced hypokalemia.

Table 3. Loop diuretics

Formula	*Generic name* Trade name	CAS registry number
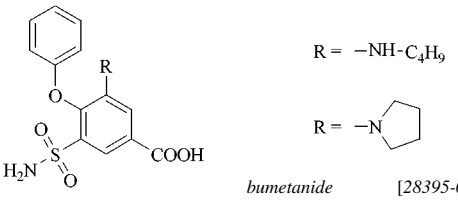	*bumetanide* Burinex (Leo)	[*28395-03-1*]
	piretanide Arelix (Hoechst)	[*55837-27-9*]
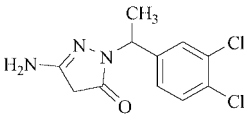	*muzolimine* Edrul (Bayer, Zyma)	[*55294-15-0*]
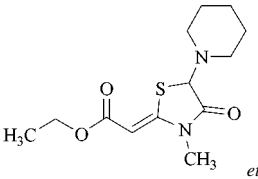	*etozolin* Elkapin (Gödecke)	[*73-09-6*]
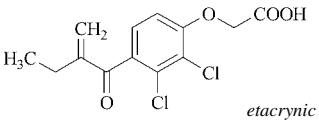	*etacrynic acid* Edecrin, Hydromedin (Merck Sharp & Dohme)	[*58-54-8*]

Triamterene [*396-01-0*], $C_{12}H_{11}N_7$, M_r 253.3, *mp* 316 °C, is obtained by condensation of 2,4,6-triamino-5-nitrosopyrimidine with benzyl cyanide [43].

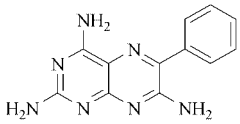

Trade Names. Dyrenium, Triamterene (Smith, Kline & French); Teriam (Roussel); Jatropur (Röhm Pharma); Dytide H (combination with hydrochlorothiazide, Röhm Pharma).

Amiloride [*2609-46-3*], $C_6H_9Cl_2N_7O$, M_r 266.2, *mp* 285 – 288 °C (decomp.), is synthesized as follows: Methyl 3-aminopyrazine-2-carboxylate is chlorinated with sulfuryl chloride to methyl 3-amino-5,6-dichloropyrazine-2-carboxylate; treatment with ammonia gives methyl 3,5-diamino-6-chloropyrazine-2-carboxylate, which is converted to amiloride by reaction with guanidine [44].

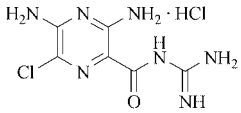

Trade Names. Colectril, Midamor; Moduretic (combination with hydrochlorothiazide); all Merck Sharp & Dohme.

5. Aquaretics

Aquaretics are agents which selectively increase the renal excretion of osmotically free water without major changes in electrolyte homeostasis. They dilute urine to an osmolality distinctly below plasma osmolality values and produce a hypotonic urine. Since vasopressin is the primary hormonal regulator of total body water, the discovery of diuretic vasopressin analogues has initiated interest in vasopressin receptor antagonists as potential specific water diuretics [45]. On the basis of the mechanism of action and chemical structure, one can differentiate three main types of aquaretics:

1) Vasopressin-analogous V2 receptor antagonists
2) Nonpeptide V2 receptor antagonists such as OPC-31260
3) Nonpeptide κ-opioid receptor antagonists

The benzazepine derivative OPC-31260, 5-dimethylamino-1-{4-(2-methylbenzoylamino)benzoyl}-2,3,4,5-tetrahydro-1*H*-benzazepine, is a selective nonpeptide V2 receptor antagonist [46]. OPC-31260 inhibits arginine vasopressin binding to V1 and V2 receptors in a competitive manner, whereby it is 100 times more selective for V2 than for V1 receptors. Furthermore, OPC-31260 does not have partial agonistic activity, and after oral administration to rats, causes a dose-dependent diuretic effect, significantly lowering the osmolality of the urine. OPC-31260 has entered phase II studies.

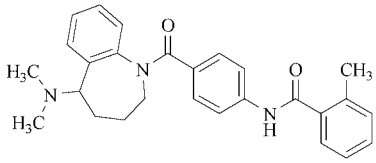

6. Adenosine A1 Receptor Antagonists

Adenosine A1 receptor antagonists are a relatively new class of compounds which may have therapeutic potential in treatment of edematous states of different origin and in acute renal failure (ARF). Adenosine, a degradation product of ATP, plays a significant role in normal and ischemic kidney function [47]. Adenosine is a vasodilator in most vascular beds but acts as a vasoconstrictor in the kidney. In the kidney, adenosine A1 receptors are expressed in the renal vasculature, juxtaglomerular cells, in proximal tubules, the thick ascending limb of Henle's loop and in the papillary collecting duct. Adenosine vasoconstricts preglomerular vessels in the kidney, while vasodilating efferent arterioles. Thus adenosine has been shown to decrease glomerular filtration rate (GFR) and renal blood flow (RBF) by means of A1-receptor-mediated renal afferent arteriolar vasoconstriction and A2-receptor-mediated efferent arteriolar vasodilation. Furthermore, adenosine modulates tubular handling of electrolytes as well as tubuloglomerular feedback (TGF) and inhibits

renin release. Adenosine also seems to be involved in ischemia-induced ARF, renovascular hypertension, and in various nephrotoxic drug reactions.

Theophylline, a nonselective adenosine receptor antagonist, exerted therapeutic effects in various animal models and in contrast medium-induced renal insufficiency in humans. Recently developed selective adenosine A1 receptor antagonists CVT-124 (1,3-dipropyl-8-[2-(5,6-epoxy)norbornyl]xanthine, KW-3902 [8-(noradamantan-3-yl)-1,3-dipropylxanthine] and FK-453 {(+)-R-1-[(E)-3-(2-phenylpyrazolo[1,5-a]pyridin-3-yl) acrylol]-2-piperidine} have been described in preclinical studies as potent proximally acting diuretic agents with therapeutic efficacy in ischemia-induced ARF and in renal insufficiency provoked by nephrotoxic drugs such as cisplatin, gentamicin, and cephaloridine [48]. Clinical development status of these agents is Phase I/II. Currently, experimental and clinical data strongly support the hypothesis that blockage of the adenosine A1 receptors may be a new therapeutic principle in renal insufficiency.

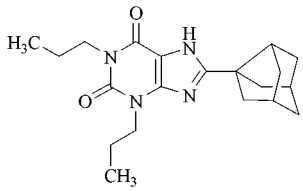

KW-3902

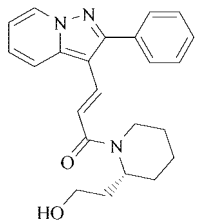

FK-453

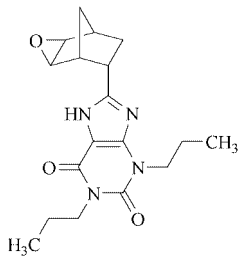

CVT-124

7. References

General References

1. E. J. Cragoe, Jr. (ed.): *Diuretics,* J. Wiley & Sons, New York 1983.
2. J. B. Puschett, A. Greenberg (eds.): *Diuretics,* Elsevier, Amsterdam 1984.

Specific References

3. R. L. Smith in E. J. Cragoe, Jr. (ed.): *Diuretics,* J. Wiley & Sons, New York 1983. p. 555.
4. R. W. Schrier, J. K. Kim in A. Greenberg (eds.): *Diuretics,* Elsevier, Amsterdam 1984. p. 56.
5. R. Palluk, W. Gaida, W. Hoefke, *Life Science* **36** (1985) 1415.
6. R. O. Roblin, J. W. Clapp, *J. Am. Chem. Soc.* **72** (1950) 4890.
7. American Cyanamid, US 2 554 816, 1950 (J. W. Clapp, R. O. Roblin); US 2 744 907, 1954 (R. W. Young).
8. F. C. Novello, J. M. Sprague, *J. Am. Chem. Soc.* **79** (1957) 2028.
9. Merck & Co., BE 557 149, 1957 (F. C. Novello); BE 568 621, 1958 (E. F. Schoenewaldt, A. E. Erickson, J. M. Chemerda).
10. G. de Stevens, *Experientia* **14** (1958) 463.
11. R. Ritscherlich, *Experientia* **14** (1958) 458; *Schweiz. Med. Wochenschr.* **89** (1959) 353 – 361.
12. Ciba, CH 337 205, 1958 (G. de Stevens).
13. Simes, BE 580 881, 1959.
14. Ciba, CH 337 203, 1958 (G. de Stevens).
15. C. T. Holdrege, R. B. Babel, L. C. Cheney, *J. Am. Chem. Soc.* **81** (1959) 4807.
16. L. H. Werner, A. Halamandaris, S. Ricca, Jr., L. Dorfmann, G. de Stevens, *J. Am. Chem. Soc.* **82** (1960) 1161.
17. G. de Stevens, L. H. Werner, W. E. Barrett, J. J. Chart, A. H. Renzi, *Experientia* **16** (1960) 113.
18. W. J. Close, L. R. Swett, L. E. Brady, J. H. Short, M. Vernstein, *J. Am. Chem. Soc.* **82** (1960) 1132.
19. Ciba, DE 1 093 370, 1959 (R. B. Margerison, A. C. Shabica, J. B. Ziegler).
20. Thomae, DE 1 125 938, 1960 (E. Müller, K. Hassbacher).
21. C. W. Whitehead, J. J. Traverso, H. P. Sullivan, F. J. Marshall, *J. Org. Chem.* **26** (1961) 2814.
22. B. Truninger, B. Siegenthaler, *Schweiz. Med. Wochenschr.* **91** (1961) 87.
23. H. Horstmann, H. Wollweber, K. Meng, *Arzn. Forsch.* **17** (1967) 653.

24. E. Cohen, B. Klarberg, J. R. Vaughan, Jr., *J. Am. Chem. Soc.* **81** (1959) 5508; **82** (1960) 2731.
25. B. V. Shetty et al., *J. Med. Chem.* **13** (1970) 886.
26. Beiersdorf, US 3 567 777, 1966 (W. Liebenow).
27. E. Jucker, A. Lindenmann, *Helv. Chim. Acta* **45** (1962) 2316. E. Jucker, A. Lindenmann, E. Schenker, E. Flückinger, M. Taeschler, *Arzn. Forsch.* **13** (1963) 269.
28. Science Union, US 3 565 911, 1969 (L. Beregi, P. Hugon, R. Malmaison, M. Laubie).
29. Eisai, JP 73/05 585, 1970 (S. Hamano, T. Nakamura, S. Kuriyama, M. Yamanaka).
30. W. Graf, E. Girod, E. Schmid, W. G. Stoll, *Helv. Chim. Acta* **42** (1959) 1085. Geigy, DE 1 244 188, 1958 (W. Graf, E. Schmid, W. Stoll).
31. G. Thuillier, J. Laforest, B. Cariou, P. Bessin, J. Bonnet, J. Thuillier, *Eur. J. Med. Chem.* **9** (1974) 625.
32. W. R. Barcley, *J. Am. Med. Assoc.* **243** (1980) 771. D. G. Beevers, J. M. Walker, *Lancet* 1980, 1417.
33. Hoechst, DE 1 122 541, 1959 (K. Sturm, K. Siedel, R. Weyer).
34. L. E. Early, *N. Engl. J. Med.* **276** (1967) 996, 1023.
35. K. Sturm, W. Siedel, R. Weyer, H. Ruschig, *Chem. Ber.* **99** (1966) 328.
36. P. W. Feit, *J. Med. Chem.* **14** (1971) 432. Loewens, GB 1 249 490, 1968.
37. W. Merkel, D. Bormann, D. Mania, R. Muschaweck, M. Hropot, *Eur. J. Med. Chem.* **11** (1976) 399.
38. E. M. Schultz, E. J. Cragoe, Jr., J. B. Bicking, W. A. Bolhofer, J. M. Sprague, *J. Med. Pharm. Chem.* **5** (1962) 660.
39. Bayer, US 4 018 890, 1973 (E. Möller, K. Meng, E. Wehinger, H. Horstmann).
40. G. Satzinger et al. in E. J. Cragoe, Jr. (ed.): *Diuretic Agents,* ACS Symposium Series 83, American Chemical Society, Washington, D.C., 1978, p. 155.
41. R. M. Dodson, R. C. Tweit, *J. Am. Chem. Soc.* **81** (1959) 1224.
42. J. A. Cella, R. C. Tweit, *J. Org. Chem.* **24** (1959) 1109.
43. R. G. W. Spicket, G. M. Timmis, *J. Chem. Soc.* **1954,** 2887. SK&F, US 3 081 230, 1960 (J. Weinstock, W. D. Wiebelhaus).
44. E. J. Cragoe, Jr., O. W. Woltersdorf, J. B. Bicking, S. F. Kwong, J. H. Jones, *J. Med. Chem.* **10** (1967) 66.
45. M. Manning, B. Lammek, A. M. Kolodziejczyk, J. Seto, W. H. Sawyer, *J. Med. Chem.* **24** (1981) 701.
46. Y. Yamamura, H. Ogawa, H. Yamashita, T. Chihara, H. Miyamoto, S. Nakamura, T. Onogawa, T. Yamashita, T. Hosokawa, T. Mori, M. Tominaga, Y. Yabuuchi, *Br. J. Pharmacol.* **105** (1992) 787.
47. H. Oßwald, H. J. Schmitz, R. Kemper, *Pflüger's Arch.* **371** (1977) 45.
48. C. E. Müller, *Exp. Opin. Ther. Patents* **7** (1997) 419.

DNA → Nucleic Acids
Dodecanol → Fatty Alcohols

Drawing and Writing Materials

Michael Colditz, A. W. Faber-Castell GmbH & Co., Nürnberg, Federal Republic of Germany (Chaps. 1 and 2)

Ernst Kunkel, Pelikan Aktiengesellschaft, Hannover, Federal Republic of Germany (Chaps. 3 and 4)

Karl-Heinz Bohne, Pelikan Aktiengesellschaft, Hannover, Federal Republic of Germany (Chaps. 3 and 4)

1. Introduction 1
2. Pencils 3
2.1. Chemical Composition 3
2.2. Writing Properties 3
2.3. Production 4
3. Ink 6
3.1. Drawing Ink 6
3.2. Indelible Ink 6
3.3. Ink for Fountain Pens 6
3.3.1. Blue-Black Ink 6
3.3.2. Colored Ink 7
3.4. Ballpoint Pens and Inks 7
3.5. Ballpoint Ink Pens 9
3.6. Fiber-Tip Pens 10
4. Charcoal, Chalk, and Crayons . . . 11
4.1. Charcoal 11
4.2. Pastel Chalk 11
4.3. Blackboard Chalk 11
4.4. Crayons 12
5. Economic Aspects 12
6. Toxicology 13
7. References 13

1. Introduction

Ink, especially ballpoint ink, and pencil leads are the main drawing and writing media today. Charcoal, chalk, and crayons, although less important, are widely used.

History. Cave paintings of the ice age are considered to be the oldest documentation of human ingenuity, representing the first use of drawing materials. A piece of flint or a stone ax, the first writing tool, was used to scratch pictures on the walls of caves, pictures that were then painted with colored earth. Stone remained the most important writing surface for a long time. Even after the appearance of hieroglyphics next to the painted pictures and signs, stone and slate pencils continued to be the most common and long-lasting writing utensils.

Later, the flint and ax were supplemented by the chisel, for rough writing, and the gouge, for fine writing. In fact, until the beginning of the twentieth century, children learned to write with a slate and a slate pencil. However, clay tablets are easier to write on and also easier to transport. The writing was scratched into the soft, plastic surface, and the clay was then fired. Similarly, wax tablets, employed by Mediterranean traders, are light and easy to write on [5], [6]. The slate pencil was especially suitable for inscribing wax plates.

Even in early times, in addition to stone, metal, clay, and wax, thin rolls were available for writing. The Aztecs used fig leaves, and as far back as 3000 B.C., Egyptians pioneered the use of papyrus, which they prepared from the pith of the papyrus plant. Papyrus was employed until the ninth century A.D. [6], and a reed was used to write on papyrus. India ink is the oldest writing and drawing medium. As long ago as 2600 B.C., India ink sticks were prepared in China from lampblack and animal glue, and used for both writing and drawing. When rubbed with water, the sticks gave an aqueous dispersion of carbon – the basis for preparing India ink even today. The glue served as the protective colloid.

Parchment was discovered in Asia Minor around 200 B.C. Depilated, untanned goat- or calfskin provided the starting material. Parchment was used until the Middle Ages. Precious manuscripts still bear witness to the fine art of medieval manuscript writing [5], [7].

Tsai Lun discovered paper in 105 A.D. in China. However, this knowledge did not reach Europe until the eleventh century, at which time

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_037

the use of India ink and other inks for writing spread rapidly.

Pencils. In ancient times, CATULLUS and PLINY mentioned the use of graphite (Greek: *graphein*, to write) or lead disks to draw lines. The so-called silver pencil, an alloy of tin and lead, was used as a drawing instrument from the twelfth century on.

In the sixteenth century, CONRAD GESNER described graphite stick-shaped leads, which originated solely in England, and which were later protected by cartridges from breakage and abrasion. Deposits of very pure sheet graphite (later a crown monopoly) were found in Borrowdale, England, in 1564. Over 200 years later, in 1782, the chemist C. W. SCHEELE proved that graphite is hexagonal carbon, not lead or molybdenite. However, the term lead pencil was by that time firmly entrenched in the language. The manual production of wood-encased graphite leads began in Nürnberg, Germany, around 1662. In 1794, NICOLAS JACQUES CONTÉ, in France, ground graphite, mixed it with clay as a binding agent, and fired the mixture at ca. 1000 °C to obtain preformed sticks of ceramic leads. These showed three substantial improvements: (1) higher strength, preventing breakage during writing, (2) variable hardness achieved by changing the ratio of graphite to clay, and (3) more efficient utilization of graphite.

Pens. The Chinese used a fine hair brush to write on paper. For centuries, bird feathers, especially goose quills, were the instrument of choice for writing with ink. Pen nibs were made from copper, iron, and brass as early as the sixteenth century. In 1798, ALOIS SENEFELDER, the inventor of lithography, produced the first nib made from hardened steel [8], [9].

The Leipzig engineer SCHELLER is said to have constructed a fountain pen in 1781, and L. E. WATERMAN invented a fountain pen with a vulcanized rubber reservoir in 1883. Since 1887, generations of German students have learned writing and arithmetic while using Soennecken pens. In 1889, GUSTAV PICKHARDT patented a fountain pen consisting of a glass tube drawn out to a point at one end, a holder for the nib, and a wick made from a capillary fabric that transported the ink [10].

Ink. Progress in writing was closely connected with the development of India ink and other inks. The first ink was made by rubbing soot or a colored pigment such as cinnabar with water-soluble binding agents, e.g., glue or gum arabic. Pieces of solid India ink were rubbed against a stone with water to form a writing liquid, which could then be applied with a brush [9].

A few inks in use today were employed before the birth of Christ. For example, the bones of cattle were boiled to give a brew that, when mixed with iron vitriol and thickened, darkened on exposure to light and could be used as ink.

The Middle Ages brought forth recipes for iron – gallic inks that contained aqueous extracts of nut gall, iron vitriol, and gum arabic [11]. The first iron – gallic inks contained the actual ink dye – iron oxides of tannic and gallic acids – suspended in the liquid. The binding agent, gum arabic, was responsible for keeping the darkly colored iron salts suspended in the liquid and binding them to the surface of the paper.

AUGUST LEONHARDI's patent, in 1856, describing the production of so-called alizarin ink, was a breakthrough in ink chemistry. The innovation was the addition of a solution of iron(II) sulfate and indigo to the aqueous extract of Aleppo gall to produce a clear, filterable liquid containing dissolved iron(II) gallate. Conversion to the iron(III) salt by atmospheric oxygen gave rise to the actual ink dye only on the paper itself (this is also the reason for the sludge present in inkwells until the end of the nineteenth and beginning of the twentieth centuries). Another dye had to be present to make the ink visible immediately, and an extract of madder root, which contains alizarin, was employed for this purpose [12].

A formulation for indelible ink that met official German regulations in 1888 consisted of

23.4 g tannic acid
7.7 g gallic acid
30.0 g iron(II) sulfate
10.0 g gum arabic
10.0 g hydrochloric acid (25 wt %)
1.0 g phenol

diluted to 1 L with water. As early as 1890, the Leonardi company in Dresden produced 0.5×10^6 kg of ink annually [12].

Ballpoint Pens. Ballpoint pens were mass-produced in the United States in the 1940s and

introduced to Europe during World War II. However, pens having a ball instead of a nib to transfer ink or colored paste had been patented earlier [13]. Jose Biro, who has ca. 700 patents to his name, is largely responsible for the further development of ballpoint pens.

Chalk and Crayons. The simplest and one of the oldest writing materials is charcoal. Natural red iron oxide together with clay minerals was used as ruddle or bole chalk (fifteenth and sixteenth centuries), a forerunner of the pastels of the eighteenth century. Lumps of naturally occurring chalk were formed into sticks and used as white drawing chalk. Gray slate completed the palette of naturally occurring materials.

Wax is one of the oldest drawing materials. In ancient Roman wall and panel paintings, a wax drawing technique – encaustic, which employs melted wax – was used. These encaustic waxes were presumably formed into sticks, which were the forerunners of wax crayons.

2. Pencils

Today, a pencil refers to a dry writing instrument, usually 178 mm long, made from wood, metal, plastic, etc., with a sharpened, homogeneous graphite lead at its center.

2.1. Chemical Composition

A lead consists of several components.

Dye	graphite [*7782-42-5*], organic and inorganic pigments, triphenylmethane dyes
Binding agent	aluminum silicates (clays), plastics, methylcellulose [*9004-67-5*], lignin
Additive	fats, wax mixtures
Filler	kaolin [*1332-58-7*], talc [*14807-96-6*]

The binding agents and fillers function as the supporting framework of the lead in which the graphite and pigments are deposited. Both the nature of the binding agents and the ratio of the components determine the structure and hardness of the lead; the pigment determines the intensity of color. The pore volume of the lead is filled by dipping it in a hot fat – wax mixture, which decreases friction during writing. The choice of hard or soft fats – waxes allows some control over the degree of hardness. In addition, the adhesiveness, breaking strength, and color intensity of the lead can be increased in this way. The clay content varies from 15 wt % in the softest lead to ca. 60 wt % in the hardest. A mixture of about 2 parts of graphite and 1 part of clay gives the usual HB or 2 1/2 degrees of hardness.

2.2. Writing Properties

The quality of a lead is determined by its hardness, breaking strength, sliding ability, covering power, color intensity or blackness, brightness, light fastness, and erasability.

The hardness is a measure of the abrasion resistance of the lead. The Castell degrees of hardness have existed since 1910. This scale is abbreviated by using the first letters of the traditional English markings: B = black/soft, HB = medium, F = firm, and H = hard. Today, the scale encompasses 17 grades of hardness: 8 B – B, HB, F, H – 7 H. A same scale of hardness is being compiled for the ISO standards. The hardness scale in use in the United States employs only numbers, but the most important grades in the two scales have a one-to-one relationship.

3 B	2 B	B	HB	F	H	2 H	3 H	4 H
1	1 1/2	2	2 1/2	3	3 1/2	4	4 1/2	5

The breaking strength and sliding ability, opposing factors, are largely dependent on the proportion of binder to graphite.

On magnification, a pencil line is seen to be not an exact line but rather a collection of lead particles scraped off by the rough surface of the paper. The covering power and color intensity of a pencil line are determined by the adhesiveness, the rubbing off of the lead, and the intensity of the dyes.

2.3. Production

Graphite Leads. In the classical procedure, the starting materials – ground natural graphite and very fine clay – are first freed from sand,

oversized material, and impurities. The microcrystalline structure of graphite must be retained during grinding. The moistened mixture (15 wt % moisture) is homogenized in kneading machines and rolling mills, and subsequently extruded, cut, and dried. The leads are then fired in furnaces, similar to those used in ceramics, at 950 – 1200 °C. Although the graphite is not structurally changed, the clay sinters, releasing water to form a hard, firm structure. After firing, the lead is dipped into a hot fat – wax mixture and then either enclosed in a protective covering, usually wood, or used loose in mechanical pencils.

Polymerized Graphite Leads. Polymer leads contain, for example, a synthetic thermoplastic [14] or lignin [15] rather than clay. As in the classical procedure, graphite is mixed with the binder during grinding. After homogenization, the mixture is extruded into leads. The plastic pyrolyzes in the subsequent firing to produce a carbon skeleton, which assumes the function of the clay framework but far surpasses the breaking strength and elasticity of conventional leads. Polymer leads are used to prepare fine refills with diameters of 0.3, 0.5, 0.7, and 0.9 mm. The leads are impregnated with a fat – wax mixture, and the fat is then removed from the lead surface. Thus, they can be used as loose leads in mechanical pencils. The principal advantage of these is that they no longer have to be sharpened. These intensely colored leads, which possess superior breaking strength and sliding ability, can still be erased completely.

Colored Leads. Both the composition and the production of colored leads differ greatly from those of graphite leads. Because the pigments are not resistant to high firing temperatures, clay or carbon cannot be used as binding agents. The binding agent must be a substance that is capable of swelling in water, such as methylcellulose, a cellulose derivative; a synthetic thermoplastic; or a substance, such as cellulose nitrate or ethylcellulose, that dissolves in an organic solvent and hardens as a result of the evaporation of the solvent at moderate heat. The pigments are homogenized with the binders, fillers (e.g., kaolin, powdered talc, chalk), additives, and fungicides by kneading and rolling. The mixture is subsequently passed through an extruder and dried below 100 °C. The extruded and dried leads are dipped into a fat – wax bath. Fat improves the consistency, sliding ability, and color intensity; the lead is wipe-resistant and nonfading, and can have either bright or pastel color (e.g., Polychromos). The resistance to light depends primarily on the quality of the pigment. Binding agents that are insoluble in water are not attacked by atmospheric humidity.

In addition to these oily, chalky leads, there are hard emulsion leads that contain fat and emulsifiers in the initial mixture and therefore do not have to be treated afterward. These leads may be blended like watercolors and can be used with water for painting (traces remain).

Indelible Leads. Indelible pencils (copying pencils), which allow duplication in the copying press, are produced much like colored pencils. The dyes, however, are water soluble and usually organic, e.g., triphenylmethane dyes. Indelible pencils and colored indelible pencils are not erasable. The ink dye penetrates the fibrous tissue of the paper through the moisture on the surface. Black indelible pencils contain up to 60 % graphite and methyl violet; consequently, they write black and copy violet. The colored leads do not contain graphite but rather water-soluble ink dyes. Indelible pencils have been almost entirely replaced by indelible ballpoint pens.

Wooden Pencils. California cedar (*Libocedrus decurrens*) is the preferred wood for pencils. In addition, alder, linden, fir, poplar, and some tropical woods (e.g., jelutong) are suitable. The wood is cut into boards, called *slats*, half the thickness of a pencil and then impregnated to improve the cutting quality and standardize the shade. The slats, most commonly 184 × 70 × 5 mm, must be knotless, straight, free from tension, dry, and soft enough to be sharpened without splintering. Five to ten half-grooves are cut into the slats on an automatic pencil assembly belt. After insertion of the lead and gluing of two slats to enclose the lead, the glued slats are dried under pressure. They are then cut into hexagon, oval, or round pencils and planed to size. Subsequently, the raw pencils are sent to the polishing room, where up to nine layers of varnish are applied. Labeling is carried out by foil-stamping, direct printing, or offset printing. Despite intensive automation, more than 30 steps are neces-

sary to convert lead and wooden slat into a pencil (Fig. 1)

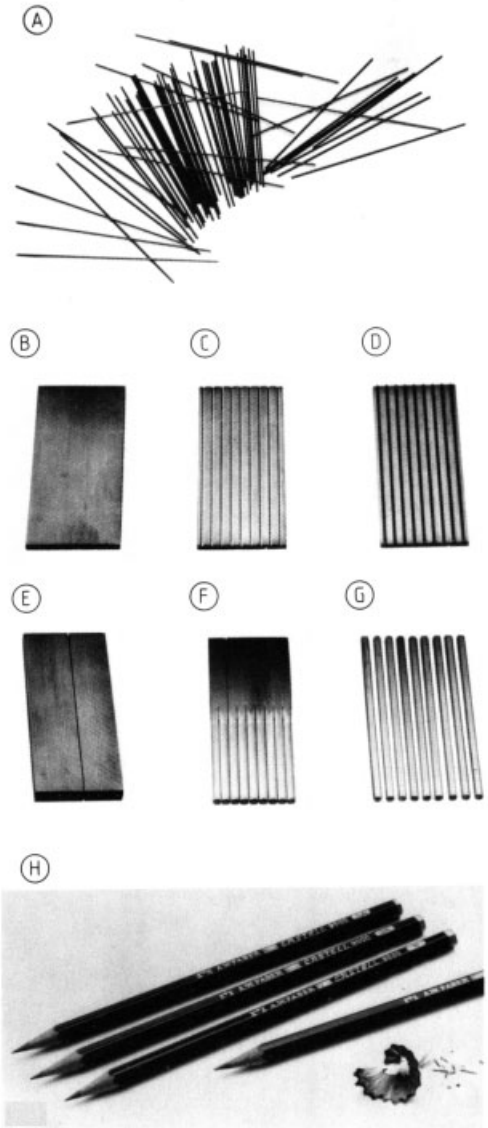

Figure 1. Pencil production
A) leads; B) seasoned, impregnated cedar slat; C) grooved slat; D) grooved slat with leads inserted; E) lead sandwich; F) final shaping; G) unfinished pencils; H) pencils

Pencils produced by plastic coextrusion could not match the high quality of wooden pencils, and were not accepted by the consumer.

3. Ink

The inks described in this chapter are writing and drawing liquids having a variety of compositions and intended for a variety of uses. There are two basic types of inks, one that contains fine carbon or a colored pigment suspended in water with the aid of a binder (e.g., India ink), which has tended to be used in drawing, and one that contains water-soluble dyes (e.g., fountain pen ink), which has tended to be used for writing. Photographs courtesy of A.W. Faber-Castell GmbH & Co.

3.1. Drawing Ink

Liquid drawing inks, for example, *India ink,* have been produced and sold commercially since the turn of the century. These inks consist of a suspension of carbon black in a colloidal solution of shellac soap, which is stabilized with glue, the protective colloid, and provided with a suitable preservative. Poly(acrylic acid ester) can be used in place of the shellac soap. Carbon black [*1333-86-4*] is the best form of carbon (→ Carbon, Chap. 6.). The shellac is converted into a soap by means of a weak base – borax, ammonia, morpholine, or ammonium hydrogen carbonate.

The best protective colloids are the nongelling animal glues; however, poly(vinyl alcohol), methylcellulose, hydroxyethylcellulose, carboxymethylcellulose, other high molecular mass, water-insoluble polymers, and surfactants are also used. Formerly, the most common preservative was phenol, but this is being increasingly replaced by less toxic compounds.

The carbon black is suspended in a concentrated solution of the other ingredients in roller mills, ball mills, or homogenizers. The ink is allowed to stand several weeks in wood or plastic containers, or it is centrifuged, before being put into bottles. Quality standards for drawing inks exist in the United States, Great Britain (BS 2490), Canada, Israel, etc., and usually concern composition and stability with respect to light and water.

When the ink dries, the hardened shellac makes a water-insoluble, nonfading line. Generally speaking, India ink should have excel-

lent covering power and produce lines of uniform width and a sharp edge. Ink used in automatic drawing devices should flow quickly and smoothly.

The high concentration of pigment and binding agents tends to cause ink to dry on the point of the pen. This tendency can be reduced by the use of moisturizers, but then the ink also drys more slowly after application.

Colored inks may be either opaque or transparent. Opaque inks contain organic or inorganic pigments, whereas transparent inks contain ordinary water-soluble dyes of the sort used in writing inks. Otherwise, colored inks are made much like India ink and contain similar substances. Exceptions are the inks used on transparent plastic films, such as those employed by cartographers. Here the writing surface is hydrophobic and nonabsorbing. These inks are made with solvents that etch the surface of the plastic film; the etching promotes adhesion. In addition, these inks must be resistant to water, gasoline, and alcohol.

3.2. Indelible Ink

Indelible inks are iron – gallic inks containing specified amounts of tanning agents and iron sulfate. They are used primarily to sign notarized contracts and important documents. Today, the quantity of indelible inks purchased is relatively small.

3.3. Ink for Fountain Pens

3.3.1. Blue-Black Ink

Ink for fountain pens should flow well; it contains less iron and tannic acid than indelible ink but more water-soluble dyes. Originally, these inks had a bright blue color, on paper, which darkened with time when exposed to light. They are often called blue-black inks. An example of their composition is

2.0 g tannic acid
4.0 g gallic acid
7.5 g other tanning agents
5.7 g iron(II) sulfate
1.2 g sulfuric acid (94 – 95 wt %)
7.0 g ink blue R
1.5 g amino blue A
3.0 g gum arabic
1.5 g preservatives

diluted to 1 kg with water.

A sketch depicting the construction of a modern cartridge fountain pen is shown in Figure 2. A photograph of a modern fountain pen is shown as Figure 3 A, along with other types of pens that use ink.

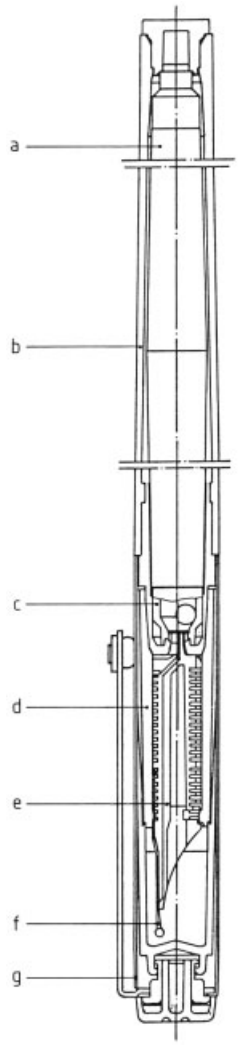

Figure 2. Cartridge pen:
a) spare cartridge; b) shaft; c) cartridge with ball stopper; d) front; e) ink supplier; f) nib; g) cap with clip

Figure 3. Pens:
A) modern fountain pen; B) ballpoint pen; C) ballpoint ink pen; D) fiber-tip pen
photographs courtesy of Pelikan Aktiengesellschaft

3.3.2. Colored Ink

The inks used most often today for fountain pens are colored inks. These write better than iron – gallic inks. The addition of a moisturizer (glycerin, glycol, or sorbitol) keeps the ink liquid in the pen and at the nib, even if the pen has been left open. A triacylmethane dye is the basis for the most common blue ink.

Ink blue (Acid Blue 93, C.I. 42780), patented by A. W. Hofmann in 1858, and its modifications are the most frequently used ink dyes. Ink containing this dye has the advantage of being completely removable (i.e., erasable) with sulfite solution, which was largely responsible for the increased use of fountain pens, especially among students. Young consumers today increasingly employ inks having a color other than blue.

An example of the composition of a blue ink is

17.0 g moisturizer
2.5 g sulfuric acid (94 – 95 wt %)
5.0 g ink blue
0.1 g surfactive agents
0.5 g preservatives

diluted to 1 kg with water.

These inks are produced easily. The moisturizer, usually glycerol or sorbitol, is diluted with water; sulfuric acid added; and the dye stirred into the mixture. The resulting liquid is heated to 80 °C and stirred until the dye is completely dissolved. After cooling, the surfactive agents and preservatives are added.

If preservatives are not used, molds tend to grow in both iron – gallic and colored inks; these can clogs fountain pens. Formerly, most inks were preserved with phenol; however, today only preservatives that comply with food laws are used.

3.4. Ballpoint Pens and Inks

The functional core of a ballpoint pen (Fig. 3 B) is the refill which, in the final analysis, is res-

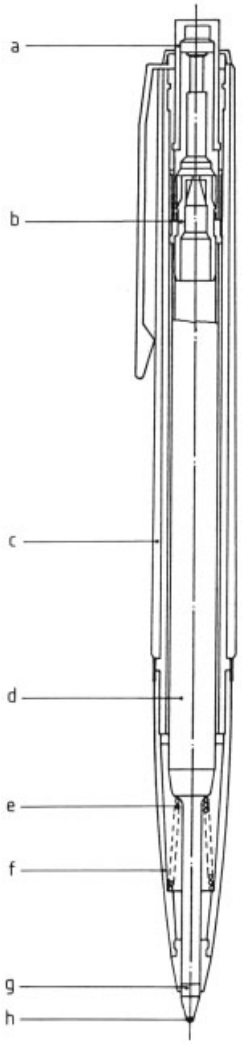

Figure 4. Ballpoint pen:
a) push button; b) switching mechanism; c) top part with clip; d) refill; e) compression spring; f) front; g) point; h) ball

ponsible for the quality of the pen. In the Federal Republic of Germany, a quality refill must satisfy the requirements stipulated in DIN 16554. The main components of a refill are the ball, the ball bed, the cartridge that contains the paste, and the plug responsible for leakproofing.

The materials used to make a ballpoint refill are usually tungsten carbide or sapphire for the ball and brass, stainless steel, plastic, or a copper – nickel alloy for the bed. The cartridge is generally made from plastic or brass. The refill is made leakproof by means of a plastic foam plug that is permeable to air. The so-called large-capacity refills have a plug made from soft, pasty hydrocarbons or silicone grease that floats on the surface of the colored paste and moves along with it. Ballpoints, like other pens, depend on capillarity to function (see Fig. 4). The rolling ball transfers a film of color from its surface to the paper; the ball itself is constantly wetted by the paste in the cartridge. A uniform wetting of the surface of the ball is made possible by a narrow gap between the ball and the bed, and by distributing chanels in the bed. The surface tension and viscosity of the writing paste must be suitable for the pen. Many pastes are characterized by a viscosity of 10 – 20 Pa s and a surface tension of ca. 40 mN/m at 20 °C.

A dye concentration up to 40 % ensures intense color. Dyes digested in oleic acid or acid resins, or metal – complex dyes soluble in organic solvents (e.g., phthalocyanine dyes) are suitable. Metal – complex dyes make the paste resistant to light and chemicals (as required by DIN 16554). The flowability of ballpoint pastes is adjusted by adding resins or polyvinylpyrrolidone. Low-volatility solvents, such as benzyl alcohol, glycols, glycol ether, or poly(ethylene glycol) are used. An example of the paste composition by weight is

30 % dye base
10 % oleic acid
5 % pigments
5 % softener
30 % phenylglycol
2 % benzyl alcohol
10 % synthetic resin
0.2 % polyvinylpyrrolidone
7.8 % 1,2-propanediol

Ballpoint pens write only when the refill is tilted downward. If the pen is held horizontally or if the tip is held upward, the force of gravity overpowers the capillarity of the system, and the pen does not function. This problem can be solved with so-called *pressurized refills*, which were described in U.S. patents many years ago [16].

Fluorinated hydrocarbons are the predominant propellants; however, air or carbon dioxide may be employed. The ballpoint paste must be extremely viscous, ca. 100 – 150 Pa s, to prevent it from flowing out of the space bet-

ween ball and bed under normal conditions. The compositions of the paste is such that it possesses thixotropic, cohesive, and adhesive properties. Apart from the dyes used in normal ballpoints, special acrylics and highly viscous polyvinylpyrrolidone give the paste the desired properties.

As a rule, ballpoint pastes should be indelible (stipulated in DIN 16554). However, erasable ballpoint refills are available. The composition of one such paste is described in [17]: rubber or polyisoprene with a molecular mass as high as 100 000 – 900 000 is dissolved in a mixture of a low-boiling and a high-boiling solvent, and ca. 20 % pigment is finely dispersed in the resulting solution. Lubricants, resins, dispersing agents, etc., can also be added to improve writing properties and adjust erasability.

3.5. Ballpoint Ink Pens

The ballpoint ink pen (Fig. 3 C) combines the advantages of the ballpoint pen and the fiber-tip pen; it ensures continuous, uniform writing. The ballpoint ink pen was first produced in Japan and has become as popular as ballpoint pens were earlier. The ballpoint ink pen has a stainless steel, cemented carbide, or ceramic ball, which rests in a plastic or stainless steel bed. The ink is free flowing. The pen writes more easily than a ballpoint pen and produces intense colors.

The inks are watery and, therefore, the pen must be equipped with a so-called tampon, a highly absorptive reservoir. The reservoir is made from bundles, about 10 cm long, of staple fibers of cellulose acetate or polyester. The tampon is filled with ink. The surface tension of the ink and capillarity enable the ink to flow out of the reservoir through a wick to the ball and onto the paper. A typical composition by weight consists of a mixture of

20 – 30 % polydiols
2 – 8 % dyes
0.5 % surfactive agents
0.5 % anticorrosive agents
0.2 % preservatives

diluted to 100 % with water.

Although the polydiols prevent the ink from drying on the ball or in the pen, they function primarily as lubricants for the ball and reduce rolling friction. The surfactive agents regulate the surface tension of the ink and adjust it to the reservoir, wick, ball bed, and ball.

3.6. Fiber-Tip Pens

The *felt-tip pen*, used predominantly for large-sized lettering and marking of packages, was the forerunner of the fiber-tip pen [18]. The felt-tip pen consists of a metal or plastic shell and a tip of cut pressed resin-soaked felt. It is filled with an ink consisting of a dye in a solvent that evaporates quickly. The felt-tip pen was developed in Japan.

Fiber-tip pens (Fig. 3 D) also originated in Japan and simplified the Oriental art of calligraphy. The fiber-tip pen is available in diverse variations and can be used for painting, marking, or fine writing.

All these pens have the same basic features: the ink supply is accommodated in a reservoir of synthetic fibers and conveyed to the writing point by a wick, which simultaneously functions as the writing point (see Fig. 5). This writing point is made from nylon, acrylonitrile, or polyester fibers or from carefully bundled, extruded fibers of a nonabrasive plastic like polyoxymethylene. The latter is especially suitable for writing and noted for uniform fine strokes.

Fiber-tip pens are assembled from the individual part by automated machines. The tip holder is surrounded with poly(vinyl chloride) or polystyrene to make the front of the pen, and this is fused to an extruded polypropylene tube. The fiber tip is placed and fixed in the front, the reservoir going into the open tube. The reservoir is filled with ink, and the tube is sealed with a plastic plug. The pressure inside the pen is kept the same as that outside by the presence of a vent, so that the ink always flows smoothly [19].

The fiber-tip pen is widely used today for both writing and drawing. As a result of its manifold applications, many types of ink have been developed. Inks having an aqueous base are used mainly for writing and drawing on paper. Inks containing volatile organic solvents and synthetic resin adhesives are used to write on or mark nonabsorbing backgrounds like glass, metal, or plastic foil. An example of the com-

position by weight of an aqueous fiber-tip ink is

20 – 30 % glycols
1 – 10 % dyes
0.2 % preservatives
0.1 % surfactive agents
2 % binding agents

diluted to 100 % with water. A typical solvent-containing ink for marking contains

2 – 15 % dyes
10 % butanol
15 % methoxypropanol
10 % binding agents

diluted to 100 % with ethanol.

Fiber-tip ink has low viscosity – if possible <4 mPa s at 25 °C – and low surface tension to guarantee good ink flow and quick drying. The aqueous inks contain 10 – 30 wt % moisturizer, such as glycols or polyglycols, to prevent the ink from drying if the pen is left open. The flowability is improved by surfactive substances. Preservatives prevent the growth of mold. Generally, acidic or basic dyes are employed, and so-called pigment preparations are used to produce nonfading inks.

Fiber-tip pens for children contain food-coloring dyes; other fiber tips contain dyes easily removed with sodium sulfite, e.g., ink blue R (Hoechst AG) or basacryl dyes (BASF). The so-called text markers contain fluorescent dyes, e.g., uranin (disodium derivative of fluorescein), pyranin (disodium derivative of 3-oxypyrene-5,8,10-trisulfonic acid), or rhodamine in combination with phthalic acid resins.

Quick-drying inks contain mainly ethanol as a solvent, in addition to butanol and other high-boiling alcohols and glycol esters. In some countries, toluene and xylene are also employed, although inclusion of aromatic solvents in goods that may be used by children is prohibited in the Federal Republic of Germany. Suitable solvents in this case are basic dyes and color bases. Natural and synthetic resins serve as film formers and improve adhesive strength.

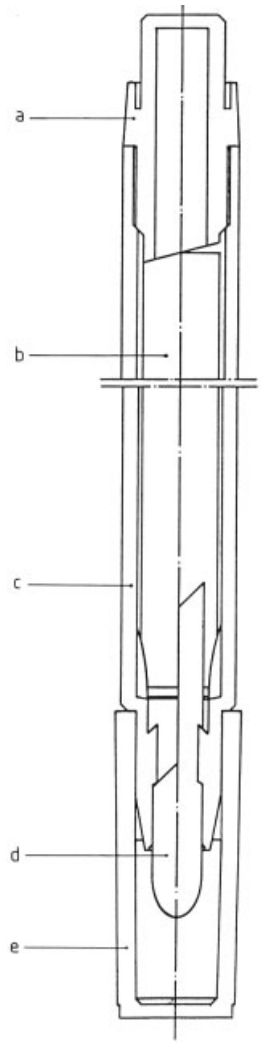

Figure 5. Fiber-tip pen:
a) plug; b) ink reservoir; c) shaft with socket; d) writing wick; e) cap

4. Charcoal, Chalk, and Crayons

4.1. Charcoal

All types of wood form charcoal when heated in the absence of air. The best charcoal sticks for drawing are those from linden or willow.

4.2. Pastel Chalk

A mixture of nonfading organic – inorganic pigments and mineral fillers like chalk, gypsum,

powdered talc, or kaolin is added to a solution of binding agent like tragacanth gum derivatives, cellulose derivatives, alginates, gum arabic, or poly(vinyl alcohols) in a heatable – coolable mixing and kneading aggregate, and is intensively mixed. Clay or bentonite improves the plasticity of the mass and helps produce an easily pressable consistency. The amount of water added must be adjusted to the pressability of the chalk mass and can be regulated, if necessary, by evaporation of the excess. Typically, if water is ignored, the mixture consists by weight of

60 – 70 % mineral fillers
10 – 15 % organic – inorganic pigments
15 – 20 % binder solution (5 – 10 %)
3 – 5 % bentonite

The cold mass prepared in this way is pressed through the nozzle of a piston or worm press to form strands; these are marked with notches that determine the length and are then air dried. After drying, the notched strands are broken to give pieces of chalk. The amount of binding agent must be determined so that the dry sticks have sufficient tensile strength and optimal writing ability.

4.3. Blackboard Chalk

Casting Process. Anhydrous gypsum is mixed with water in a stirred mixer to give a thick lump-free paste. The paste is then passed once or twice through a 0.4-mm sieve and poured into stick-shaped forms.

After the gypsum – water mixture has set, the forms are opened and the sticks are removed. The sticks are either air dried or dried in an oven at $\leq 50\,^{\circ}C$. The mixing ratio of water to gypsum regulates the hardness and writing ability of the chalk. A water-to-gypsum ratio of 5 : 6 gives chalk that writes well and has good tensile strength. Powdered talc, kaolin, or chalk can be added to improve writing ability. Trace amounts of cellulose or gum arabic solutions bind dust. An assortment of colored chalks can be obtained by adding pigments to the basic mixture.

Pressing Process. The choice of the basic raw materials is the same as that described for pastel chalks. Only the amount of binding agent has to be adjusted to increase the hardness of the chalk.

Wetting agents are also added. These additives improve the washability of the blackboard. The pressing into strands and the preparation of individual pieces of chalk are identical with those of pastel chalks. Such a pressed chalk might have the following composition by weight:

ca. 67.5 % elutriated natural chalk
5 % powdered talc
5 % bentonite
10 % binder solution consisting of 20 wt % carboxy-
methylcellulose in water
2 % wetting agent
0.5 % preservatives
10 % water

Chalks have been produced that are efficiently pressed, can be processed safely up to the finished product, and have a high tensile strength [20].

4.4. Crayons

Crayons are produced to meet varying needs, from painting and drawing to signing. The production of crayons is carried out by either casting or pressing.

Casting. A typical composition by weight for crayon casting is

ca. 40 % C_{16}–C_{18} fatty acids
15 % microcrystalline wax
25 % ester wax
15 % mineral fillers
5 % pigments.

A typical composition for crayons that can be water painted is

ca. 15 % C_{16}–C_{18} fatty acids
30 % fatty alcohol, polyglycol ether or ester
5 % microcrystalline wax
25 % ester wax
20 % mineral filler
5 % pigment

The cast sticks are produced by melting the components and dispersing the fillers and pigments in the wax melt. A double-walled heatable container is used. The rate of rotation of the stirrer (600 – 1000 rpm) is adjusted to give a fine distribution of solid components. The homogeneous liquid crayon melt is then poured into casting forms that can be heated or cooled.

The casting temperature depends on the melting point of the wax mixture, being higher than the melting point to ensure sufficient heat capacity during the filling of the preheated casting forms. The temperatures of the wax mixture and the casting forms must be almost perfectly adjusted to produce crayons having superior surface quality and good tensile strength.

Various types of casting machines are available today, ranging from highly efficient, computerized, rotating casting machines to simple, manually operated, casting tables.

Pressing/Extruding. A typical crayon mixture for pressing has the following composition by weight:

ca. 30 % C_{16}–C_{18} fatty acids
10 % microcrystalline wax
15 % ester wax
10 % stearates
30 % kaolin
5 % pigments

The type of wax depends on the application: writing on glass, metal, or wood; painting on cloth; or other artistic purposes.

The wax mass is prepared in a kneader that can be heated or cooled. The components are homogenized, and the consistency is adjusted to guarantee a fine distribution. The cooled wax mass is pressed into strands by the heated nozzle of a piston press or extruder. The strands are cut to length and sharpened.

5. Economic Aspects

Pencils. In 1985, the number of pencil factories in the world was around 90, with a total production of ca. 4×10^9 pencils. For the last 20 years, the number of pencils produced annually in the Federal Republic of Germany has been ca. 500×10^6 ($\pm$ 15 %); more than half of these are exported. The ratio of black pencils to colored pencils has been ca. 3 : 2. The most important Western producers are the United States (ca. 1.5×10^9 pencils), Germany, Brazil, Mexico, the United Kingdom, and Japan.

Pens and Inks. Expensive, high-quality pens are made by a small number of producers, whereas inexpensive writing materials, especially fiber-tip pens, are made by a large number of producers worldwide. Reliable production information, either by quantity or by value, is not available.

6. Toxicology

In general, the substances present in ink, crayons, chalks, pencils, and other writing materials are not toxic, and all are subject to strict regulation. The end products must meet a series of legal standards that depend on the consumer group. For example, products used by children must be nontoxic in every respect. In industrialized countries, regulations cover the chemicals, the consumer goods, labeling, and other restrictions. For example, in Europe heavy-metal content must comply with EN 71, Part 3, which prohibits the use of soluble compounds of arsenic, antimony, barium, cadmium, chromium, lead, and mercury.

The poison centers in the Federal Republic of Germany and in West Berlin can provide information in cases of acute poisoning, which occur occasionally when children consume large quantities of ink or coloring matter. Any doctor in the Federal Republic of Germany has a list of these centers. Other industrialized countries have such poison centers.

7. References

General References

1. R. Endler: "Die Bearbeitung des Holzes und Fertigstellung des Bleistiftes," *Chem.-Ztg.* **65** (1941) 284.
2. R. Gööck: *Die großen Erfindungen, Schrift-Druck-Musik,* Edition Sigloch, Künzelsau-Stäfa-Salzburg 1984, pp. 158 – 163.
3. E. Sattmann: *Vom Faustkeil zum Bleistift,* P. Basten, Aachen 1952.

4. R. Zückert: "Die Bleistiftfabrikation," *Chem.-Ztg.* **65** (1941) 281.

Specific References

5. Wilhelm Eule: *Mit Stift und Feder,* Fachbuchverlag, Leipzig 1955.
6. Gottfried Schweizer: "Papier-historischer Kulturträger und modernes Industrieprodukt," *Archivalische Zeitschrift,* **70** (1974) 67 – 69.
7. Hans Kotte: *Welches Papier ist das?* **vol. 2,** DPW Verlag, Heusenstamm 1982, p. 52 P.
8. Erhard Sattmann: *Vom Faustkeil zum Bleistift,* Verlag Peter Basten, Aachen.
9. Rudolf Blanckertz: "Schreibwerkzeug und Schriftform," *Deutsches Museum, Abhandlungen und Berichte,* **5** (1933) no. 4.
10. Gustav Pickhardt, DE 52 640, 1889.
11. Horst Roselieb: "Die Chemie alter und neuer Tinten," *Archivalische Zeitschrift,* **70** (1974) 74 – 78.
12. Friedrich Wilhelm Hinrichsen: *Die Untersuchung von Eisengallustinten, Die Chemische Analyse,* Enke Verlag, Stuttgart 1909.
13. Michel Baum, DE 239 050, 1910. Wenzel Klimes, DE 670 809, 1935.
14. Dainihonbungu, DE-AS 1 519 102, 1964 (Tsukahara, Yasushi).
15. Pilot, JA-AS Sho-51-26849, 1976 (M. Shinomoto).
16. Scripto, US 3 099 252, 1973 (E. P. Cofield, Jr.). Fisher, US 3 425 779, 1967 (P. C. Fischer, H. Schub, R. A. Kennedy).
17. Gilette, US 4 097 290, 1976 (F. A. Muller, H. Peper, Jr.).
18. Marlock, DE 831 961, 1950. Carters Ink Co., DE 1 267 137, 1962.
19. Tokyo Boshi, DT 2 456 905, 1973; DT 2 621 544, 1975.
20. J. S. Staedtler, DE 2 645 812 C2, 1978; DE 3 120 241 C2, 1985.

Driers → Metallic Soaps
Drug Delivery Systems → Pharmaceutical Dosage Forms

Dry Cleaning

David C. Tirsell, Dow Chemical Co., Midland, Michigan 48640, United States

1. Introduction 1
2. Solvent 1
3. Process Description 4
4. Aftertreatment 5
5. References 5

1. Introduction

Dry cleaning is a process for removing soils and stains from fabrics and garments, which uses special equipment and a nonaqueous organic solvent with detergent added. Dry cleaning is popular for *cleaning natural fibers,* such as cotton and wool; it provides good cleaning without the garment shrinkage experienced with aqueous cleaning. Dry cleaning is also popular because of its *convenience,* providing garment care with equipment and skills beyond the reach of the typical consumer.

The dry-cleaning industry began in France in 1845 [1], and although it may be mature in age, it is dynamic in terms of adapting to a changing array of fabric, solvent, machinery, and regulatory issues that confront it.

Type of Operation. The industry has evolved into several forms in providing a cleaning service. The smallest operations are often the *coin-operated* dry cleaners or "coin-ops." The dry cleaning is provided as a sideline to a coin-operated laundry service, which is the principal business on the premises. The machines employed are operated by the customer or by an attendant and are usually for small-batch cleaning of 3 – 5 kg at a time; additional pressing or finishing services are not provided. Coin-ops are a significant, though declining, portion of the solvent-cleaning industry in terms of the number of establishments in North America [2], and are less prevalent in other countries.

The next largest industry segment in terms of garments cleaned per plant is *retail* or *commercial* operations. Garments are cleaned and finished by trained personnel, with aqueous cleaning a minor portion of the business. These plants may have several cleaning machines, cleaning 5 – 30 kg at a time, perhaps with more than one solvent available to give flexibility in cleaning various fabrics.

The largest volume industry segment is *industrial* cleaning. Industrial cleaning plants specialize in cleaning uniforms, protective garments, gloves, towels, or linen for large-volume institutional or industrial customers. Aqueous cleaning is often a significant portion of these businesses, with the proportion swinging between dry and aqueous cleaning with a change in relative economics or garment type chosen by the customer. Machines may have a range of 25 – 180 kg of garment capacity, again with more than one solvent usually available on-site.

2. Solvent

The characteristics of the cleaning process in whatever type of plant that is involved center on the solvent employed. Solvent *economics* is important; North American data show that the total supply cost for a retail plant, including solvent, other chemicals, hangers, and packaging materials, ranges from 9 to 11.5 % of all costs [3]. However, other factors beyond economics need to be considered, e.g., processing features, garment compatibility, and health-related issues, before an optimum solvent can be properly chosen. Because of all of these factors, four major dry-cleaning solvents are used:

1) petroleum blends, such as Stoddard solvent,
2) perchloroethylene,
3) fluorocarbon 113 (1,1,2-trichloro-1,2,2-trifluoroethane), and
4) 1,1,1-trichloroethane.

Properties of interest for a dry-cleaning solvent start with a *high affinity for greases and oils,*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_049

the major function of the solvent in the fabric-cleaning process, balanced with a *low affinity for fabric dyes.* A good dry-cleaning solvent must also be *noncorrosive* to the common metals used in dry-cleaning equipment as well as be *chemically* and *thermally stable* under the variety of conditions experienced in the process. The solvent should be *compatible* with the wide variety of detergents and spotting chemicals in common use. The solvent needs to be sufficiently *volatile* to leave the garment essentially free of odors after drying, evaporating at a sufficiently low temperature to protect the garment fabric from damage and to allow separation from contaminants and detergents by distillation.

The last major property of interest for a dry-cleaning solvent is its relative *safety* in terms of human health and toxicity. Regulations vary throughout the world, but clearly the trend is toward a reduction in solvent use and worker exposure to all solvents. This concern for exposure, coupled with economic considerations, has encouraged the use of improved ventilation and maintenance practices, as well as recovery devices, such as refrigerated condensers or activated-carbon beds. In terms of allowable worker-exposure levels, the TLVs and MAKs of the four major solvents are given in Table 1 [4].

Solvent Selection. The balance of all of the pertinent solvent properties has resulted in a niche in the industry for each solvent. *Petroleum* has moved from a dominant position in the market in the 1930s to a stable 10 – 12 % share of the cleaning market in North America and a larger though still minority market share in Europe and Japan. Leather and suede cleaning is a major use, with the relatively gentle solvency of petroleum blends making petroleum popular for a range of textile garments as well. Although these blends are still widely used in many areas, insurance, zoning, and safety considerations often favor a nonflammable solvent substitute.

Perchloroethylene dominates the world market for dry-cleaning solvents as the most successful nonflammable replacement for petroleum, and has over 80 % of the North American dry-cleaning market. Its more aggressive solvency and lower boiling point have increased its popularity for faster and easier cleaning.

Fluorocarbon 113 was first introduced to dry cleaning in the 1960s and developed a small market on the basis of its extreme gentleness toward garments and drapes, allowing the cleaning without damage of items with more solvent-sensitive adhesives or ornamentation. The lower drying temperature reduces wrinkling and speeds finishing of a garment.

The high solvency of *1,1,1-trichloroethane,* coupled with less graying and a lower energy use, encourages a growing market for that solvent in Japan, Europe, and North America. Some of the advances in equipment, refrigeration, and recovery design developed for perchloroethylene to decrease solvent consumption and control solvent losses with the more volatile fluorocarbon solvents have contributed to machinery designed for 1,1,1-trichloroethane use.

Physical Properties. Table 2 compares the important physical properties of the four major dry-cleaning solvents. The values for petroleum solvents are expressed as ranges covering the most common blends now in use in the industry.

The importance of each property in determining the optimum solvent for a particular dry-cleaning process is as follows:

Kauri–Butanol (K–B) Value. The higher the K–B value, the greater the amount of grease and oil that will dissolve in the solvent; in addition, cleaning can be done faster with a higher K–B value solvent than with a lower K–B value solvent. Although dye stability is also affected by moisture levels and temperatures for cleaning and drying, a lower K–B value solvent is generally safer for loosely fixed dyes or fabrics that are more sensitive to solvent attack.

Surface Tension. Although modified by the addition of detergents, a lower solvent surface tension value enables the solvent to penetrate the fabric easier, and soils can be removed quicker. Once the solvent has wetted the fabric, it must have adequate solvency for oils and soils for this property to be of use.

Dielectric Constant. This property is a measure of the affinity of the solvent for polar or ionic substances, such as some soils and stains. A higher value indicates quicker removal of such stains and reduced static electricity buildup on the fabric surface. The dielectric constant also affects "graying" of fabrics, i.e., the redeposition of soils onto fabrics after they have been

Table 1. Occupational exposure limits to dry-cleaning solvents [4]

Solvent	TLV				MAK	
	TWA		STEL ∗			
	ppm	mg/m^3	ppm	mg/m^3	ppm	mg/m^3
Stoddard	100	525	(200)	(1050)		
Perchloroethylene	50	335	200	1340	50	345
1,1,2-Trichloro-1,2,2-trifluoroethane	1000	7600	1250	9500	1000	7600
1,1,1-Trichloroethane	350	1900	450	2450	200	1080

∗ The ACGIH intends to delete the values in parentheses in 1986 – 1987.

Table 2. Physical properties of dry-cleaning solvents [a] [5–10]

Property	Petroleum	Perchloro-ethylene	Fluorocarbon 113	1,1,1-Tri-chloroethane
CAS registry number	[*8052-41-3*] [b]	[*127-18-4*]	[*76-13-1*]	[*71-55-6*]
Molecular formula	hydrocarbon blends	C_2Cl_4	$C_2Cl_3F_3$	$C_2H_3Cl_3$
K–B [c] value	27 – 45	90	31	124
Surface tension (at 20 °C), N/m	0.027 – 0.050	0.0323	0.0196	0.0256
Dielectric constant	2.00 – 3.00	2.36	2.46	7.53
Relative density	0.75 – 0.85	1.620	1.577	1.315
bp, °C	150 – 210	121.0	47.6	74.0
Evaporation rate (relative to CCl_4 = 100)	6 – 9	39	170	91
Latent heat of vaporization (at *bp*), J/g	284.3	209.8	147.0	234.9
Flash point (TCC [d]), °C	38 – 60	none	none	none
Solubility of water, wt %	0.01	0.01	0.009	0.05

[a] All values at 25 °C unless noted.
[b] For Stoddard solvent.
[c] Kauri–Butanol.
[d] Tagliabue closed cup.

initially removed in cleaning. With solvents having a higher dielectric constant, the more polar soils stay in solution rather than redepositing on the fabric surface.

Relative Density. A solvent with a higher relative density adds more mass to a garment and increases its soil-removing mechanical action, because a garment absorbs solvent in cleaning and retains it through extraction. A lower relative density solvent reduces fabric damage from this mechanical action and is desired for cleaning all but the most heavily soiled garments.

Boiling Point. The temperature needed to dry a garment is directly related to the boiling point of the solvent. A lower boiling point reduces the solubility of fabric dyes, the shrinkage of heat-sensitive fabrics, and the energy needed to distill the solvent from detergents and oils. Distillation is facilitated by large differences between the boiling point of the solvent and that of the various detergents, chemicals, and contaminants encountered.

Evaporation Rate. A higher evaporation rate shortens the time needed for garment drying, but also increases the need for greater care in machine design and solvent storage practices to avoid greater loss of solvent to the air.

Latent Heat of Vaporization. A higher value indicates a greater amount of energy needed to convert the liquid solvent to vapor in distillation or drying. Condensation of the solvent vapor is exothermic, a factor in designing adequate cooling systems.

Flash Point. For all practical purposes, three of the four solvents are nonflammable, with the flash point of petroleum dependent on the particular blend chosen. The flammability of petroleum solvents requires care in equipment design, plant lighting, and wiring.

Solubility of Water. A lower solubility of water in a solvent indicates a quicker separation of excess moisture from the solvent in the water separator. At the same time, some solubility is desired to help the detergent remove water-soluble soils.

3. Process Description

The flow of solvent within a dry-cleaning plant is shown in Figure 1.

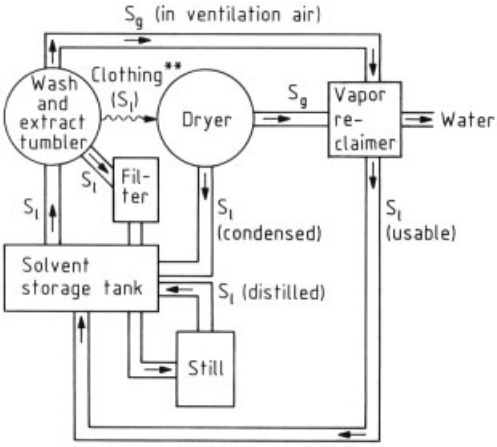

Figure 1. Flow diagram of a typical dry-cleaning plant ∗ [10]
∗ S_l = solvent in liquid state; S_g = solvent in gaseous state.
∗∗ Clothing wet with liquid solvent.

Cleaning. The first step in the dry-cleaning process is a 3 – 15 min *soaking* and *tumbling* of the garments in the solvent in a machine much like a front-loading washing machine (see Fig. 2). A *detergent* is usually added to the solvent at base-tank concentrations of 0.50 – 1.25 vol % to aid in removing water-soluble stains (→ Laundry Detergents). An alternative procedure is to inject the detergent at a lower concentration into the wash cycle with each cleaning load, often followed by distillation of the solvent. Detergents enable the system to carry moisture above the solvent's usual water saturation point, preferably maintaining a 75 – 85 % solvent relative humidity. Most of the detergents used are anionic, although a few cationic and nonionic *surfactants* are available (→ Surfactants). All are blended by the manufacturer with inert fillers added to facilitate volume measurement. In addition, various *prespotting chemicals* may be applied to specific stains before the garment is placed into the tumbler. Solvent-soluble oils and greases are removed by the action of the solvent on the garment fibers, and insoluble soils are removed by the mechanical tumbling action.

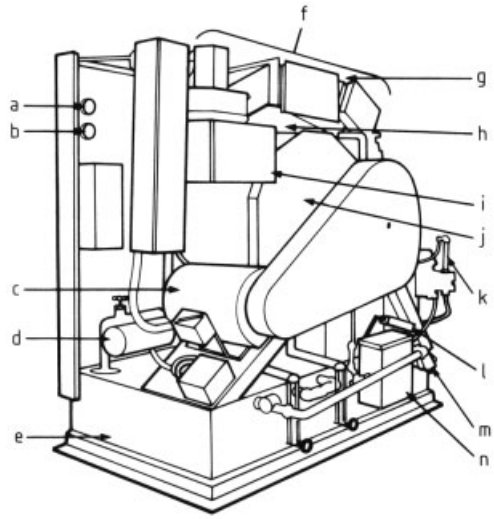

Figure 2. Dry-cleaning machine ∗, one-bath, dry-to-dry [10]
a) Pressure gauge; b) Temperature gauge; c) Motor; d) Pressure pump with in-line strainer; e) Base tank; f) Stainless steel recovery housing; g) Steam valves, variable temperature control; h) 100 % Solvent flow-through basket; i) Lint trap; j) Stainless steel tumbler; k) Variable solvent level; l) Flow valves; m) Solvent-resistant paint; n) Water separator
∗ Rear view.

Drying. The *extraction* step removes most of the solvent from the garments before they are tumbled with heated air. The temperature of the air and the length of the drying cycle are determined by the solvent used and by the garment fabric material and construction. Figure 3 shows the dynamics of the drying cycle. *Water separators* on the dryer and on the still help to control solvent moisture levels by slowing the solvent flow returning to the base tank enough to allow excess water to separate.

Drying has been done traditionally in a piece of equipment separate from the washer, although the portion of the industry using nonflammable solvents is moving to the use of a single, dry-to-dry unit for both washing and drying (Fig. 2).

This unit saves space and solvent and decreases worker contact with solvent by eliminating the transfer of solvent-laden garments.

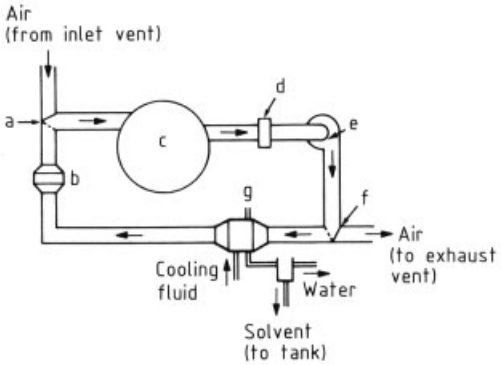

Figure 3. Reclaiming dryer system [10]
a) Air inlet valve; b) Heater; c) Dryer; d) Lint screens; e) Fan assembly; f) Air exhaust valve; g) Cooling coil

4. Aftertreatment

The solvent is continuously *filtered* during the washing cycle to remove insoluble soils, such as lint, dust, and dirt. Most processes use cartridge filters, containing a large surface area of paper for lint and soil removal and a central core of activated carbon for removing odors and dyes from the solvent. Even greater dye removal capacity is available with cartridges composed entirely of activated carbon. Cartridge filters that also contain activated clay to help remove grease and oil from the solvent are also used. Other systems are equipped with tubular filters that employ diatomaceous earth and activated carbon as the filter media, which was the standard of the industry before the more convenient cartridge filtration systems were introduced.

The solvent can be further cleansed by either continuous or batch *distillation.* Nonflammable solvents can be processed in an atmospheric still, whereas the higher boiling petroleum blends are distilled at reduced pressure. The distilled solvent is returned to the machine-base tank or other storage tank for reuse. The still bottoms are soils, greases and oils, detergent, and spotting chemicals, which are discarded appropriately.

5. References

1. E. Garfield (ed.): "Drycleaning, the Process and Its History: from Starch to Finish," *Current Contents* , no. 22, Institute for Scientific Information Press (ISI), Philadelphia 1985, part 1, pp. 3 – 12.
2. Reference 1, "Drycleaning, Commercial History and Social Impacts: Ironing Out Some Pressing Problems," no. 23, part 2, pp. 3 – 10.
3. Neighborhood Cleaners Associations (NCA), Cost Chart, NCA Bulletin, New York, Apr., 1985.
4. ACGIH (eds.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* Cincinatti, Ohio, 1986 –1987, pp. 23, 26, 29, 32. DFG (ed.): *Maximum Concentrations at the Workplace (MAK),* VCH Verlagsgesellschaft, Weinheim 1986.
5. W. L. Archer, L. S. Violette, *Ind. Eng. Chem.* **16** (1977, Dec) 319 – 325.
6. Chemcentral Corp., Physical Properties of Common Organic Chemicals & Solvents, Chicago, Dec., 1982.
7. Du Pont, Freon Solvent Formulation Data, Bulletin FST-5, Wilmington, Del., 1984.
8. International Fabricare Institute (IFI), Characteristics of Drycleaning Solvents, Bulletin T-536, Silver Spring, Md., 1980.
9. I. Mellan: *Industrial Solvents Handbook,* Noyes Data Corp., Park Ridge, N.J., 1970, pp. 8, 15, 21, 48.
10. Dow Chemical, *Dry Cleaning: a Basic Handbook,* Midland, Mich., 1984, pp. 9, 10, 14.

Dry Ice → Carbon Dioxide
Drying Agents → Drying of Solid Materials

Drying Oils and Related Products

Ulrich Poth, BASF Coatings AG, Münster, Federal Republic of Germany

1. Definitions . . . 1
2. History . . . 2
3. Chemical Compositions . . . 3
4. Sources and Types . . . 4
4.1. Linseed Oil . . . 4
4.2. Perilla Oil . . . 4
4.3. Tung Oil . . . 4
4.4. Oiticica Oil . . . 4
4.5. Fish Oils . . . 4
4.6. Safflower Oil . . . 4
4.7. Sunflower Oil . . . 6
4.8. Soybean Oil . . . 6
4.9. Cottonseed Oil . . . 6
4.10. Dehydrated Castor Oil . . . 6
4.11. Tall Oil . . . 6
5. Production . . . 6
6. Testing and Specifications . . . 7
7. Film Formation . . . 7
7.1. Properties of Drying Oils for Film Formation . . . 7
7.2. Cross-Linking by Reaction with Oxygen . . . 8
7.3. Influence of Catalysts on Cross-Linking . . . 9
7.4. Yellowing . . . 10
8. Modifications . . . 10
8.1. Boiled Oils . . . 11
8.2. Stand Oils . . . 11
8.3. Blown Oils . . . 11
8.4. Factice, Sulfurized Oils . . . 12
8.5. Isomerized Oils . . . 12
8.6. Copolymers . . . 12
8.7. Maleinized Oils . . . 12
8.8. Oil Boiling with Rosin, Copal, and Amber . . . 13
8.9. Oil Boiling with Phenolic Resins . . . 13
8.10. Modifications by Transesterification 13
8.11. Modifications Based on Fatty Acids 14
9. Application, Importance, and Future Aspects . . . 14
10. References . . . 14

1. Definitions

Drying oils are fatty oils (→ Fats and Fatty Oils) from different natural sources, which are able to form coating films by reaction with atmospheric oxygen. The reaction may take place with the original, unmodified oils or after modification. Oils which are able to form films on their own, without being modified, are called drying oils in the narrow sense. Oils which are able to form films if they are modified or mixed with other film-forming components are called *semidrying oils*. *Nondrying oils* are also used to produce binders for coatings formulations, but these oils and there modifications are not able to form films by reaction with atmospheric oxygen.

The reaction with atmospheric oxygen (autoxidation) takes place with unsaturated fatty acids containing more than two double bonds in the molecule. Triglycerides containing a sufficient amount of these fatty acids are drying oils in the narrow sense. Triglycerides with a significant content of fatty acids with two double bonds are semidrying oils. Oils containing predominantly fatty acids with only one double bond are not able to dry.

Drying oils in the narrow sense are linseed oil, perilla oil, tung oil, oiticica oil, and different fish oils. Semidrying oils are cottonseed oil, soybean oil, sunflower oil, safflower oil, hemp oil, poppy oil, walnut oil, sesame oil, maize oil, grape seed oils and the oils of Euphorbia, Coriander, Brassica, and Limanthes species. Peanut oil, castor oil, palm tree oil, olive oil, and almond oil are nondrying oils.

Tall oil is not a triglyceride, but a mixture of mainly unsaturated fatty acids and rosin. Tall oil is formed as a byproduct in kraft (sulfate) pulping (→ Tall Oil). Refined tall oil can be used in the same way as fatty acids from soybean oil in the coating industry.

Dehydrated castor oil is a drying oil. Dehydrated castor oil is prepared from castor oil by dehydration with sulfonic acids or at high temperatures. Dehydrated castor oil contains a rather high amount of 9,11-linoleic acid.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_055

2. History

It was in the Stone Age that humans began to grow plants and the early advanced civilizations developed an agricultural system. Important was the cultivation of flax to produce linen, but linseed was also used for production of linseed oil. First documents which give information on the growing of flax are dated to the third dynasty in the Egyptian "Old Empire" (Pharaoh Snofru, approx. 2650 B.C.). These documents only describe the use of linseed oil as food and for lamps, there is no indication that they were used for painting. But it can be assumed that oils were used as binders for pigments in addition to inorganic binders, albumin, waxes, glues, and blood.

First paintings were artistic paintings for religious objects or for objects representing those in power. In the classic period of ancient Greece paints were also used to protect objects. Wooden ships were painted with varnishes prepared from mixtures of waxes, rosins, and oils (semidrying oils).

The first written document how to prepare oil varnishes dates back to the 11th century. The monk RODGERUS OF HELMARSHAUSEN named THEOPHILUS PRESBYTER described in his "Schedula diversarum artium" [1] how to prepare a varnish by melting amber (lat. glassa) together with linseed oil. In the Middle Ages drying oils were combined with metal oxides (white lead, red lead or minium, brown stone) not only to prepare different colors, but also to improve drying (early 13th century, HERAKLIUS, "De coloribus et artibus Romanorum" [2]). Furthermore, it was found by empirical experiments that the drying properties of oil varnishes could be improved if the oils were exposed to sunlight (stand oil), or cooked (boiled oil). Use of other driers (metal vitriols) in oil varnishes was introduced. The knowledge to prepare advanced oil varnishes was kept as an important secret [3].

In 1298 MARCO POLO reported the use of tung oil as raw material in paints for the ships in the Chinese Empire. But it was not until the late 19th century that tung oil was shipped to Europe for production of advanced oil varnishes.

In the beginning of the 15th century the use of turpentine oil (prepared by the new technique of distillation, developed in the Oriental Empires) was introduced into the preparation of oil varnishes (VAN EYCK brothers in Brugge and Gent). This invention formed the basis for the works of the most famous painters in the Netherlands at that time.

In the 16th century the first Chinese and Japanese painted articles were imported to Europe and the Europeans started to copy those techniques by using oil-based coatings. The demand for painted articles became larger and the first coating manufactories were founded in France and later in Germany. In the 18th century the method to produce sheet metal was developed in England. The coatings for the articles from sheet metal consisted of drying oil varnishes modified with furnace soot and asphalt. At the end of the 18th century the coating technique was also introduced in France and Germany. The development of traffic gave a new impetus for founding and extension of paint industries, because of an increased demand for coatings for horse coaches and ships. In the 19th century the industrial production of coatings and paints started in England, the Netherlands, and in Germany. Raw materials for the oil based coatings were linseed oil, tung oil, amber, copals, and different types of rosin. The English varnish processing plants were famous.

When the natural raw materials became rare (mainly amber and copals), chemists started to look for synthetic replacements. The first synthetic resins which were combined with oils were the phenolic resins. Later so-called oil-reactive phenolic resins were prepared. The alkyd resins formed the most important group of synthetic resins which combined the air drying of oils or oil-derived fatty acids with the physical drying of a polyester molecule. Alkyds became the most important group of resins for paints and coating from the 1930s to the 1980s and nearly totally replaced the older varnishes based on oils combined with amber, copals, and rosins. Currently alkyds are also being replaced by other binder systems.

3. Chemical Compositions [5]

Fatty Acids with Isolated Double Bonds. Important contents of drying oils are the fatty acids with isolated double bonds, i.e., linoleic, linolenic, arachidonic, and clupanodonics acid.

Linoleic acid [*60-33-3*], all-*cis*-9,12-octadecadienoic acid, M_r 280.45, *mp* $-5.0\,°C$, *bp*

(1333.2 Pa) 224 °C, iodine value 181 %, acid number 200. Various oils, which are used for the production of air-drying coatings, contain a significant amount of linoleic acid: safflower oil, sunflower oil, cottonseed oil, and soybean oil. Linoleic acid is also the main component of tall oil and a constituent of other semidrying oils which are used for specialties or artist paints.

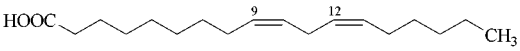

9, 12-Linoleic acid

Linolenic acid [*463-40-1*], all-*cis*-9,12,15-octadecatrienoic acid, M_r 278.44, *mp* − 5.0 °C, *bp* (1333.2 Pa) 225 °C, iodine value 274 %, acid number 202. Linolenic acid is the main constituent of linseed oil and perilla oil.

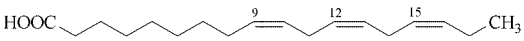

9, 12, 15-Linolenic acid

Arachidonic acid [*7771-44-0*], all-*cis*-5,8,11,14-eicosatetraenoic acid, M_r 304.47, *mp* − 49.5 °C, iodine value 334 %, acid number 184, and *clupanodonic acid* [*2548-85-8*], all-*cis*-4,8,12,15,19-dodecosapentaenoic acid, M_r 330.52, *mp* − 78.0 °C, *bp* (666 Pa) 236 °C, iodine value 384 %, acid number 170. These fatty acids are components of fish oils, which have nearly the same properties (for coatings and varnishes) as linseed oil.

Fatty Acids with Conjugated Double Bonds. Also fatty acids with conjugated double bonds are important constituents of industrially used drying oils, i.e., α-eleostearic acid, α-licanic acid, and conjugated linoleic acid.

α-Eleostearic acid [*4337-71-7*], 9-*cis*,11-*trans*,13-*trans*-octadecatrienoic acid, M_r 278.44, *mp* + 49 °C, *bp* (1560 Pa) 235 °C, iodine value 274 %, acid number 202, is a component of tung oil.

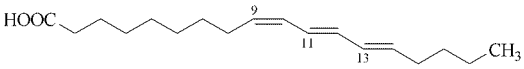

α-Eleostearic acid

α-Licanic acid, 9-*cis*,11-*trans*,13-*trans*, *trans*-4-ketooctadecatrienoic acid, M_r 292.40, *mp* + 75 °C, iodine value 260 %, acid number 202, is a constituent of oiticica oil.

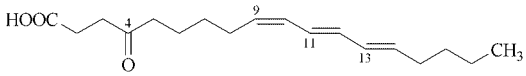

α-Licanic acid

Conjugated linoleic acid, all-*trans*-9,11-octadecadienoic acid, M_r 280.45, iodine value 260 %, acid number 202, is prepared by isomerization of 9,11-linoleic acid (see Sections 8.5) or by dehydration of ricinolic acid form castor oil (see Section 4.10).

There are other high unsaturated fatty acids in Boleco oil, Isano oil, in the oil of Marigold and in the oils of Balsaminaceae species but until now they did not find any industrial use.

4. Sources and Types [6], [7]

The composition of important drying oils is shown in Table 1. Properties of commercial drying oils are listed in Table 2.

4.1. Linseed Oil [8], [9]

Linseed oil is a drying oil. It is the oil of the seed of flax (*Linum usitatissimum*) which has been cultivated by humans since the Stone Age. Flax is mainly grown to produce fibers (linen) and is cultivated in the Baltic states, Canada, USA, Argentina, India, and since the nineties also (again) in middle Europe. The seed consists of 32 – 43 % oil, the main fatty acids are linoleic acid and linolenic acid. Linseed oil is mainly used as a raw material for paints (varnishes, stand oils, alkyds, combinations with phenolic resins, maleinized oils, urethane oils, styrolized oils) and glaziers putty and to produce linoleum, normally not for foods. Films from linseed oil are stable to water and other agents but they are not sufficiently light-stable and have a tendency for yellowing.

4.2. Perilla Oil

Perilla oil is prepared from the seeds of *Perilla fructescens* (Labiatae species), grown in India

Table 1. Composition of drying oils (figures are given in %)

	number of C-atoms : number of double bonds													
Name of oil	14 : 0	16 : 0	18 : 0	20 : 0	22 : 0	24 : 0	16 : 1	18 : 1	18 : 2	18 : 3	20 : 1	20 : 4.5	22 : 1	22 : 5.6
Cottonseed oil	0 – 2	17 – 29	1 – 4	0 – 1			0 – 2	13 – 44	33 – 58	0 – 3				
Soybean oil	traces	7 – 11	2 – 6	0 – 3			traces	15 – 33	43 – 56	5 – 11				
Sunflower oil	traces	3 – 7	1 – 5	0 – 1	0 – 1	traces		14 – 33	44 – 68	traces				
Safflower oil	traces	3 – 6	1 – 4					13 – 21	73 – 79	traces				
Linseed oil		4 – 6	2 – 5	0 – 1				12 – 34	17 – 24	35 – 60				
Herring oil	3 – 8	8 – 13	1 – 3				6 – 9	17 – 22	1 – 4	0 – 1	9 – 15	6 – 9	1 – 16	6 – 18
Menhaden oil	7 – 8	17 – 19	3 – 4				7 – 10	13 – 16	0 – 1	0 – 1	1 – 2	13 – 19[a]	0 – 2	11 – 17
Sardine oil	6 – 8	16 – 19	2 – 4				8 – 12	10 – 16	1 – 3	0 – 2	2 – 8	13 – 29[a]	1 – 6	6 – 18
Tung oil		3 – 4	1 – 3					4 – 9	8 – 10	77 – 86[b]				
Oiticica oil		6 – 7	4 – 5					8 – 13	11 – 25	73 – 83[c]				

[a] and C 18 : 4. [b] α-Eleostearic acid. [c] α-Licanic acid.

Table 2. Properties of drying oils (commercial, refined)

Name of oil	Color Gardner (max.)	Acid number (max.)	Iodine value (Wijs)	Saponification number	Unsaponifiables (max.), %	Viscosity at 20 °C, mPa · s	Titer[a], °C	Refractive index n_{20}^{D}	Density at 20 °C, g/cm^3	Specifications
Cottonseed oil [*8001-29-4*]	4	1	110 – 120	188 – 196	1.5	approx. 60	30 – 37		0.918	
Soybean oil [*8001-22-7*]	2	0.4	120 – 141	188 – 195	1.5	approx. 60	20 – 24	1.472 – 1.477	0.917 – 0.925	DIN 55938 ASTM D 1462 BS 5170
Sunflower oil [*8001-21-6*]	2	1	125 – 135	188 – 196	1.5	approx. 60	16 – 24	1.475	0.920	
Safflower oil [*8001-23-8*]	2	1	140 – 150	188 – 196	1.5	approx. 60	16 – 17	1.475 – 1.478	0.923 – 0.926	DIN 55942
Linseed oil [*8001-26-1*]	4	0.6	160 – 200	188 – 195	1.5	approx. 60	19 – 21	1.4785 – 1.4825	0.926 – 0.933	DIN 55933 DIN 55934 ISO 150 BS242 BS 243
Fish oil [*8016-13-5*]	6	4	190 – 205	188 – 194	1.5	approx. 50	25 – 31	1.482	0.930	
Tung oil [*8001-20-5*]	8	5	160 – 180	189 – 195	1.0	200 – 400	36 – 37	1.511 – 1.522	0.930 – 0.940	DIN 55936 ASTM D 12 ISO 277 BS 391
Oiticia oil [*8016-35-1*]	11	10	145 – 155	186 – 193	1.5		44 – 47	1.510	0.977 – 0.988	ASTM D 601

[a] Temperature range where the oil becomes turbid during cooling.

and Indochina. The seed consists of 30 – 51 % oil. The composition of Perilla oil is similar to that of linseed oil. The oil is of local importance.

4.3. Tung Oil [10–16]

Tung oil is isolated from the nut-like seeds of the tung tree (*Aleurites fordii*, Euphorbiaceae species) in China and Japan. Tung trees are also cultivated in USA, Malawi, and Argentina. The seed consists of 25 – 45 % oil. 80 % of the oil consists of the triple unsaturated, conjugated α-eleostearic acid. Tung oil is easy to polymerize and is still being used for coatings. Tung oil is used for preparation of stand oils, for alkyds, and in combinations with phenolic resins. Tung oil is often mixed with linseed oil. Oxygen uptake during drying is lower than that of linseed oil. Films from tung oil are more elastic than those from linseed oil and more stable to yellowing. Products based on tung oil and its modifications are used for water-resistant paints (marine varnishes) and corrosion protection. A special use is the preparation of wrinkle finishes. These finishes dry very fast in the presence of driers and form regularly wrinkled surfaces (paints for machine facings).

4.4. Oiticica Oil [17]

Oiticica oil is isolated from the seeds of *Licania rigida* (Chrysobalanaceae species) in Brazil. The seed consists of 50 – 60 % oil, which consists mainly of licanic acid. Like α-eleostatic acid licanic acid contains three conjugated double bonds and has a keto group in addition. That is why oiticica oil has nearly the same properties as tung oil. It is of local interest.

4.5. Fish Oils [18–21]

The various fish oils originate from different organs of fish and differ in their composition. Oils from herring, menhaden, and sardine are of industrial interest. All of these oils contain fatty acids with 18 to 22 carbon atoms and with up to fife isolated double bonds. The drying properties of fish oils are comparable with those of linseed oil. Fish oils have a specific odor, the same holds true for products from fish oil. Fish oils are used for varnishes and alkyds for anticorrosive coatings and primers.

4.6. Safflower Oil [22], [23]

Safflower oil is a semidrying oil. It is obtained from the seeds of Safflower (*Carthamus tinctorius*, Compositae species), which is grown in Egypt, Turkey, and North America. The pure seed consists of 46 – 50 % oil. The oil contains approx. 70 % linoleic acid. Safflower oil is for high-quality foods and in the coating industry mainly for nonyellowing alkyds.

4.7. Sunflower Oil [24], [25]

Sunflower oil is also a semidrying oil. It is isolated from seeds of sunflower (*Helianthus annuus*), which is grown in Eastern Europe, China, North America, Middle America, Australia, and Eastern Africa. The seeds consist of 42 – 63 % oil. The oil contains approx. 60 % linoleic acid. In the coating industry sunflower oil is applied in high quality alkyds (nonyellowing) and in artists' paints.

4.8. Soybean Oil [26], [27]

Soybean oil is a semidrying oil. Soybeans are the fruits of soy (*Glycine max*, Leguminosae species) which is mainly grown in China, Argentina, Brazil, and USA. The beans consist to 18 – 20 % of oil. The oil contains approx. 50 % linoleic acid. Soybean oil is mainly used for foods, the paint industry uses only a small part of the large quantities of soybean oil produced. Fatty acids from soybean oil are used for alkyds. Also stand oil are prepared from soybean oil. Soybean oil is much slower drying than linseed oil. But the films are more durable and stable to yellowing.

4.9. Cottonseed Oil

Cottonseed oil is also a semidrying oil. It is obtained from the seeds of the cotton plant (*Gossypium arboreum, G. hirsutum, G. barbadeuse*, and other species), which are located on the cotton fibers. Cotton is mainly cultivated in Egypt, China, Japan, North, and Middle America. The seed kernels consists of 14 – 25 % oil. The oil contains approx. 40 % linoleic acid. Cottonseed oil is used for foods and a raw material for fatty acids for alkyds (nonyellowing).

4.10. Dehydrated Castor Oil [28–30]

Castor oil consists of a mixture of fatty acids which contain up to 90 wt % ricinoleic acid (12-hydroxy-9-octadecanoenic acid). Castor oil is isolated from the bean-like seeds of the castor plant (*Ricinus communis*), which is cultivated in Brazil, India, the Philippines, South Africa, and Southeast Europe. The seed consists of 50 – 60 wt % oil. Castor oil is a nondrying oil. In the paint industry castor oil is used for polyurethane coatings, alkyds, and for the preparation of dehydrated castor oil and its fatty acids.

Dehydrated castor oil is prepared by dehydration of castor oil with sulfonic acids and removing the crystalline hydrates. Dehydrated castor oil consists mainly of conjugated 9,11-linoleic acid and some 9,12-linoleic acid. During dehydration polymerization may occur as side reaction. Dehydrated castor oils are available with different viscosities. Dehydrated castor oil is a drying oil. The film forming properties are similar to those of tung oil, but the drying velocity of dehydrated castor oil is lower. The films are stable against yellowing, are elastic and water resistant. Dehydrated castor oil is used for oil varnishes – often combined with other oils – for pigment pastes and for printing inks.

4.11. Tall Oil (→ Tall Oil) [31–37]

Tall oil is a byproduct of the kraft (sulfate) pulping process. Crude tall oil consists of a mixture of fatty acids (45 – 55 wt %), rosin (30 – 40 wt %), and unsaponifiable components (10 – 13 wt %). The fatty acids can be separated and used as components for alkyds. Tall oil has a similar fatty acids composition as soybean oil, but normally the content of oleic acid is higher. Furthermore, tail oil fatty acids contain unusual isomers, e.g., octadecadienoic acids with double bonds in the 5-,9-, and 5-,12-positions.

5. Production

The industrial production of oils from vegetable materials (oilseeds) and animal materials (selected fish organs) by pressing (cold and hot) and extraction is described in detail in → Fats and Fatty Oils.

Refining [6], [38] (→ Fats and Fatty Oils, Chap. 6.). Some of the trace components contained in fatty oils are accepted for use of oils in foods. But when used in coatings the trace components may negatively affect the film forming properties. The impurities extracted together with fatty oils include gums, phosphatides, carbohydrates, proteins, colorants, and waxes.

The classic refining procedure consists of the following sequential, stepwise treatments of the oil: degumming, bleaching, deacidizing, and, if necessary, deodoration and dewaxing.

Degumming is performed by adding approximately 0.1 wt % of phosphoric acid or the equivalent amount of citric acid and subsequent agitation of the oil for approx. 30 min at 80 – 90 °C. The acids coagulate the gums and the coagulate is removed by filtration or by centrifugation. The process can be supported by a treatment with fuller's earth.

Bleaching removes the vegetable colors, such as chlorophyll and carotene, autoxidation products, soaps, traces of heavy metals, residues of phosphorus-containing materials, and odorants. In bleaching solid adsorbents, such as activated fuller's earth and in special cases activated carbon, are added. Bleaching is carried out in closed, agitated tanks under partial vacuum at 80 – 90 °C in the presence of 1 – 5 wt % fuller's earth for up to 30 min. Afterwards, the oil is filtered over filter presses.

Deacidizing. Crude oils may contain 1 – 3 wt % of free fatty acids. This content can be lowered to below 0.1 wt % by deacidizing. In the older process equivalent amounts of caustic soda (sodium hydroxide) are added which convert the free fatty acids to sodium soaps at 90 °C.

The process is carried out under slow stirring to avoid emulsifying. The sodium soaps are separated in form of an aqueous, concentrated soap solution (soap stock). The soap stock is separated and removed and the oil is dried under vacuum. The process is problematic regarding the residual soap contents in the oil and the oil losses by saponification.

In the newer process the fatty oils are deacidized by distillation (physical deacidizing). The oil, which is free from other impurities, which may form decomposition products during the process, is vacuum steam-distilled at 210 – 270 °C and the fatty acids are distilled off. The final acid content is 0.01 wt %. This process includes also a deodorization step. The process avoids contamination of the oil with soaps and the loss of neutral oil is lower than in the older process.

Deodorization is essentially a vacuum steam-distillation, by which odorous volatile compounds, mainly aldehydes and ketones, are driven from the nonvolatile oils. Other volatiles, such as free acids, sterols and tocopherol, are also removed. Deodorization is carried out with steam under reduced pressure at 1.0 – 2.5 kPa and 180 – 220 °C for 2 – 6 h.

Dewaxing. Linseeds and sunflower seeds are covered with a thin layer of wax. Particles of this wax were dissolved during oil pressing and extraction. Wax dissolved in the oil crystallizes during storage, especially during cooling and the oil becomes turbid. The wax contents would result in film defects in the coatings. Dewaxing consists of cooling the oil to 4 – 10 °C and holding it at this temperature until the crystallization process is finished. Then the wax is removed by filtration.

6. Testing and Specifications [39]

The properties of drying oils are tested in specific tests. The methods and target values are defined in national (DIN, DGF, ASTM, BS) and international (ISO) standards and specifications. Oils are characterized by color, acid number, saponification value, amount of unsaponifiable contents, iodine number (amount and activity of double bounds), viscosity, refractive index, and specific density (→ Fats and Fatty Oils, Chap. 11.3., see Table 2). The fatty acid composition is determined by gas and liquid chromatography [40].

7. Film Formation

7.1. Properties of Drying Oils for Film Formation

Besides their ability to dry by reaction with atmospheric oxygen, drying oils have other specific properties. Due to the long aliphatic chains of the fatty acids and their double bond content, the coatings based on oils and related products have a low viscosity and excellent wetting properties on different types of substrates (metals, wood) and pigments. That is why drying oils can form the base for "high-solids" paints, corrosion protection paints, pigment pastes, fillers, and artist's paints (→ Artists Colors). In addition, drying oils have an excellent solubility in a broad variety of solvents, including simple aliphatic hydrocarbons, which are low cost solvents, odorless, and physiologically safe. Also the oils themselves are no hazardous products. Their molecular structure imparts excellent flow and leveling properties during film formation. When cross-linking is finished the films still contain mainly aliphatic structures, which are hydrophobic and rather stable to water and some water-soluble agents. A disadvantage is their sensitivity to basic agents, which will saponify the ester groups. That is why oils and related products are not the optimum bases of waterborne coatings. But in principle the hydrophobicity of the fatty acids is an advantage also for waterborne coatings, if the drying oils are incorporated well and protected from saponification (e.g., in aqueous polyurethane dispersion based on alkyds with unsaturated fatty acids). Another disadvantage is that drying requires much time and cannot be accelerated sufficiently by higher temperatures and catalysis.

7.2. Cross-Linking by Reaction with Oxygen

The mechanisms of film formation of drying oils have been studied in investigations on oil molecules and appropriate model substances

[41]. The process occurs in different overlapping steps and the total reaction sequence and the molecular structure of the film forming polymers are complex [42], [43]. Different reaction mechanisms are followed, depending on the type of double bonds and the use of catalysts.

Oils with Nonconjugated Double Bonds. The initial autoxidation step in the case of oils with nonconjugated double bonds is the formation of hydroperoxides in allyl position to the double bonds [44–48]. Since the 11-position of linoleic and linolenic acids and – in the case of linolenic acid also the 14-position – are activated by two double bonds, the reactivity of these positions is the highest. First a radical is formed at these positions by dehydrogenation, then oxygen is added and a hydroperoxide is formed by recombination with the hydrogen radical. The reactivity of drying oils is based on the mesomeric stabilization of the radical intermediate: the unpaired electron is delocalized over several carbon atoms along the chain, and less energy is required to eliminate the hydrogen radical [49].

Additionally isomerization may take place, analytical investigation on esters of linoleic acid, for example, showed higher amounts of conjugated unsaturated 9- and 13-hydroperoxides besides hydroperoxides in 8-, 10-, 12- and 14-position [50–53].

Since the next steps of film forming lead to different higher molecular products, the analysis of these products is much more complex. Decomposition products of the film in different states, byproducts of the film-forming reaction, were investigated by mass spectroscopy [54–56]. Time-laps infrared spectroscopy was used to follow the process [57]. The analytical result led to the interpretation that the next important step is the formation of peroxy radicals by decomposition of the hydroperoxide (bimolecular disproportionation at higher concentration). Then a free-radical chain process is started by formation of oxygen bridges between fatty acid molecules and by activation of C-radicals and formation of C–C bridges. The reaction is terminated by recombination of different radicals.

The reaction is strongly influenced by temperature, light, and traces of substances which may act as catalysts or inhibitor. For example, at lower temperatures more C–O–C bonds are produced than at higher temperatures, where formation of C–C bridges is favored [58].

Oils with conjugated double bonds react differently from those with isolated double bonds. Tung oil for example absorbs less oxygen during drying, 12 wt % instead of 16 wt % in case of linseed oil [49], [59–61]. Analysis shows that the starting reaction is not the formation of hydroperoxides, but of cyclic peroxides formed by direct oxygen attack on the conjugated bond system [62]. Reaction of the peroxides with allylic methylene groups or dissociation leads to radicals. The radicals initiate a radical chain reaction, yielding C–O–C bonds and C–C bridges [63]. Due to the easy 1,4-addition of radicals to conjugated double bonds, formation of C–C bridges is preferred. The reaction is terminated by recombination of radicals or disproportionation.

7.3. Influence of Catalysts on Cross-Linking

All oxidative drying systems are formulated by adding small amounts of catalysts for drying, the so-called driers. Driers (siccatives) are oil-soluble metal salts of organic acids. The soaps with best solubility are obtained from 2-ethylhexanoic acid and naphthenic acids, in earlier times rosin acids and linoleic acid were in use [64], [65].

The soaps are produced from metal oxides or hydroxides and the acids, either by fusion or by precipitation [66]. The driers are usually dissolved in aliphatic or aromatic hydrocarbons, water-emulsifiable versions are available [67], [68].

The driers are classified in two groups, distinguished by the mode of action: active driers and auxiliary driers.

Active driers consist of metals occurring in several oxidation states and being capable to undergo redox reactions. The degree of activity may be also influenced by the molecular compatibility of the metal ion with the double bond system. Examples of metals in active driers, in order of decreasing activity, are:

Co > Fe > Mn > Ce > Pb > Zr

The metals are added to drying oils and related products (drying alkyds) at a concentration of 0.005 – 0.2 wt %. The driers speed up two reaction steps of the drying process, the activation of methylene bonds to form hydroperoxides and, more important, the redox reaction in which radicals are formed from the hydroperoxides [69–75]. If the hydroperoxides are transformed faster to radicals, the radical chain polymerization takes place more efficiently. The films are more homogeneous and become stable to dissolution in a shorter time. The metal ions are reoxidized during the process. A simplified illustration of the reaction of a drying oil with oxygen in the presence of cobalt driers is shown in Figure 1.

Suitable combinations of metals allow synergistic improvement of the total drying effect. Thus, in the past most siccatives consisted of combinations of cobalt and lead. Cobalt promotes the formation of hydroperoxides rather fast on the film surface. This is connected with the risk that the surface is cross-linked more quickly than the other coating layers. Thus, oxygen cannot diffuse into the film and wrinkles can be formed. Lead promotes hydroperoxide decomposition and accelerates polymerization and ensures that the film is formed evenly over all coating layers. Also zirconium ions form complex compounds with hydroperoxides that decompose slowly [76], zirconium has the same effect as lead. Since lead is defined as a hazardous component, most of the new drier compositions consist of mixtures of cobalt and zirconium. But also cobalt is in discussion as possible health risk. Because no real alternative to cobalt exist, the newer formulation contain smaller amounts of cobalt driers.

Auxiliary driers are soaps of metals which form only one oxidation degree but which activate the primary active driers. They do not catalyze autoxidation by themselves. Auxiliary driers are added in higher concentration than the primary driers, up to 0.6 wt % of metal. Examples of metals for auxiliary driers are calcium, barium, and zinc. The effect is explained by the basic reaction of these driers, which mainly compensate the effect of acid groups present the oil, and thus prevent precipitation of the active drier metals [77]. Auxiliary driers can also be absorbed on surfaces, e.g., from pigments instead of the active driers and the active driers are not involved in the process [78]. Barium is more toxic than lead and should no longer be used.

If the total content of metals in a film is rather high, they will be active indefinitely, causing oxidative breakdown of the film. Therefore, the amount of driers used must be carefully balanced to achieve a high degree of cross-linking but to avoid an early film decomposition.

During storage of wet paints reaction with oxygen must be avoided. The addition of oximes (methyl ethyl ketoxime) stabilizes against skin formation by reaction with oxygen in the container [79]. But now also methyl ethyl ketoxime is defined as a toxic substance and has to be substituted.

7.4. Yellowing

Films of highly unsaturated oils and related products with isolated double bonds tend to yellow under specific conditions. Films of linseed oil and related products yellow in the dark, but the yellowing will largely disappear if the film is exposed to light. Irreversible yellowing takes place if films from linseed oil and related products are aged or baked at higher temperatures. The yellowing is caused by formation of colored polyenes or quinone structures, starting from conjugated unsaturated hydroperoxides via ketones as intermediates [80–82]. The formed 1,4-diketones may also react with ammonia (farmhouses, stables, and rest rooms) to yield substituted pyrroles that can form colored products by oxidation [83], [84]. For white and other light colors linseed oil should be replaced by less yellowing products from semidrying oils, or better, alkyds based on these oils.

But also the driers (manganese, cobalt) can form complexes with oximes, which may cause discoloration in films, mainly by aging [85].

8. Modifications

The rather effective cross-linking of drying oils by reaction with atmospheric oxygen suffers from the disadvantages of a low viscosity of the unsaturated triglycerides, and the fact that the drying process has to proceed to a significant

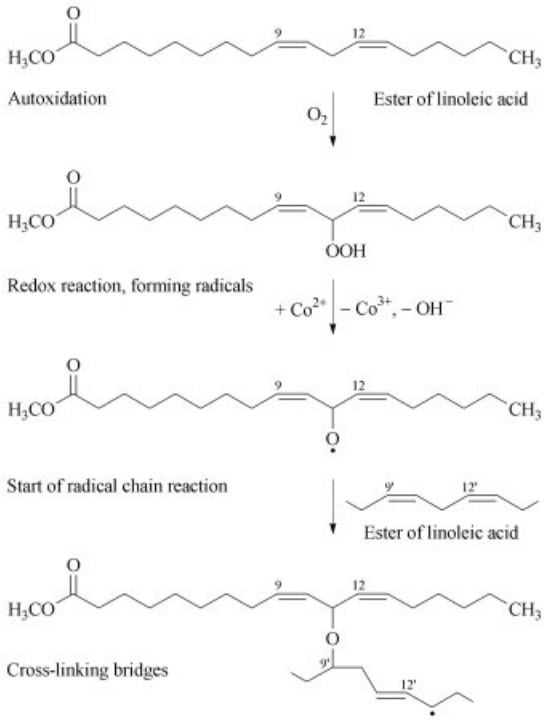

Figure 1. Simplified reaction of a drying oil with oxygen in the presence of a cobalt drier

extent until a rather hard solvent- and chemicals-resistant coating film is formed. This process requires a lot of time. Early in history the painters tried to improve the drying properties of oils. One way was the combination of the chemical drying of oils with products which form films by physical drying (waxes, rosins, amber). When the compatibility or solubility constituted a problem, the components were heated to form stable mixtures. Furthermore, it was found that addition of particular pigments to the oils improved the drying velocity of the coating films. If oils were stored in glass bottles in the sun light or heated in the absence of oxygen the oil became more viscous. The higher viscosity of these oils results also in a reduction of drying time.

When rosins, copals, and amber became rare or were too expensive, various synthetic resins were developed as combination partners for oils. In the beginning mainly phenolic resins were used. Phenolic resins were later replaced by alkyds, which combine the chemical drying of double bonds of fatty acids with physical drying of a phthalic polyester.

8.1. Boiled Oils

Boiled oils are combinations of drying oils and stand oils with siccatives. To accomplish a homogeneous distribution of the driers in the oil, the oxides or salts of the driers are incorporated by heating the mixture to higher temperatures ($\geq 100\,°C$, boiling). The metal oxides or salts form salts with part of the fatty acids by saponification. The main bases for boiled oils is linseed oil. The driers may consist of 0.1 – 0.3 wt % cobalt or about 0.4 wt % lead, sometimes combined with 0.1 wt % of manganese.

Boiled oils are used for varnishes, mainly lithographic varnishes, and printing inks (pigment pastes).

Currently it is easier to produce varnishes in a cold process, i.e., by mixing oils and stand oils (or their mixtures with other resins) with solutions of metal driers.

8.2. Stand Oils

Stand oils are drying oils with increased viscosity. The viscosity increase results from heating of the oils in the absence of oxygen. During heating an intermolecular polymerization process occurs. The main reaction is a Diels – Alder reaction between unsaturated fatty acids, acting as partly as dienes and partly as dienophiles [86], [87]. The step that controls the reaction rate of oils with *nonconjugated double bonds* (linseed oil) is the isomerization into conjugated double bonds with concurrent trans rearrangement [88–90]. Therefore, oils with conjugated double bonds react very fast. Stand oils consist of esters of polymeric fatty acids, containing a significant amount of cyclic polymeric fatty acids [91]. The reaction may be catalyzed by the addition of small amounts of anthraquinone [92].

The process used for manufacturing of stand oils must be adjusted to the reactivity of the drying oils. Stand oils of linseed oil and soybean oil are produced in agitated stainless steel reactors at 280 – 300 °C in an inert gas stream (carbon dioxide, nitrogen). The reactor may be heated by heat-transfer oils, superheated steam, electrical induction, or resistance heating. The reactor should be equipped with water cooling facilities (half-pipes outside the reactor wall or internal cooling coils). The process is controlled by temperature and time, the progress of reaction is determined by measuring the viscosity and the acid number of samples. A linseed stand oil needs 40 – 45 h at 280 °C to attain viscosities of about 10 – 15 Pa · s. Viscosity increase of soybean oil is much slower. Decomposition products of the production of stand oils are removed by distillation, under vacuum if necessary.

Tung oil is heated in small batches (20 – 50 kg) rapidly to 260 °C, held for 7 – 12 min at that temperature (exothermic reaction takes place) and then flash-cooled with minimum the same amount of cold tung stand oil.

Stand oils are used in varnishes, oil paints, enamel varnishes, and fillers. Stand oils dry more homogeneously than their starting oils and give more elastic and water-resistant coating films. Although stand oils dry more slowly than drying alkyds, stand oils are used in combination with alkyds to raise the solids of paints to reduce the amount of solvent to evaporize. Linseed stand oil has a lower tendency for yellowing than linseed oil. Tung stand oil and soybean stand oil are even more stable to yellowing than linseed oil. Tung stand oil is superior to linseed stand oil in forming hard and water-resistant films, but the films are more brittle. Thus, linseed stand oil is often combined with tung stand oil in varnishes.

8.3. Blown Oils

For production of blown oils, drying oils are blown with air at 90 – 120 °C. During this process, hydroperoxides are formed and decompose into radicals which lead to increase in molecular mass by a radical chain transfer reaction. In addition, keto groups and hydroxy groups are formed. Blown oils have viscosities up to 3 Pa · s and are much more polar than stand oils. The drying properties are reduced.

Blown oils from linseed oil and soybean oil are used as plasticizers for polymers which dry physically, for printing inks, for putties, and sealings. Blown linseed oil is used for core binders for casting.

8.4. Factice, Sulfurized Oils

Factice is a mixture of drying oils with sulfur or disulfur dichloride. Like oxygen, sulfur is able to react with the double bond system in a radical chain reaction. The polymerization occurs very effectively over the whole film layer. Varnishes containing factice oils can be applied in a wet-on-wet process.

8.5. Isomerized Oils

During isomerization isolated double bonds are transferred into conjugated double bonds and cis double bonds are converted into trans double bonds. The process is carried out with metal catalysts at 100 – 140 °C. Linseed oil and safflower oil are isomerized in the presence of a nickel catalyst, finely distributed on diatomaceous earth or active carbon [93–95]. Isomerized oils take up less oxygen during drying than the corresponding starting oils and form elastic films with higher water resistance.

The alkali isomerization process is carried out with the methyl ester of fatty acids to afford fatty acids with conjugated double bonds. They are used in alkyds, mainly for stoving enamels [96–98].

8.6. Copolymers

Copolymers of drying oils are produced by heating unsaturated oils with unsaturated monomers [99], [100], preferably styrene. Additionally, vinyltoluene [101], α-methylstyrene, dicyclopentadiene [102], pentadiene, and indene [103] are used as monomers, also methylmethacrylate is described. Copolymerization is carried out at higher temperatures, catalysts are peroxides, e.g., di-*tert* butyl peroxide [104]. The process may be run in bulk or in solution. A feed process for monomer dosing is preferred for most exothermic reactions to guarantee a reproducible process. Unreacted monomers are distilled off after the process. The content of styrene in styrenated oils is 20 to 45 wt %. During the polymerization of nonconjugated oils and styrene, grafting of styrene molecules to chains onto oil molecules occurs. Starting from the active methylene groups of the oil, polysterene bridges are formed [105], [106]. It is important to avoid the formation of styrene homopolymer, because it is not compatible with the oil. *Conjugated oils* form Diels – Alder adducts connected by short bridges of polystyrene, the viscosity increases more rapidly than the viscosity resulting from the grafting nonconjugated oils [107], [108]. The compatibility of styrenated oils with other polymers is limited. Using vinyltoluene, α-methylstyrene, or dicyclopentadiene gives copolymers with a much better compatibility and solubility. Cyclopentadiene and indene form only Diels – Alder adducts both with conjugated and nonconjugated oils.

Copolymers of oils have improved drying properties, the resulting films are harder and more resistant to water and chemical agents than the films from drying oils and alkyds.

8.7. Maleinized Oils [109–112]

Maleic anhydride is a dienophile and easily reacts with conjugated double bonds in a Diels – Alder reaction to form a substituted cyclohexane dicarboxylic anhydride. The process is carried out at 80 – 120 °C [113], [114]. Nonconjugated unsaturated oils react with maleic anhydride in a ene reaction via the active methylene groups to form substituted succinic anhydride derivatives [115], [116]. But also Diels – Alder reactions may take place if the isolated double bonds are converted into conjugated during the process. The reaction is carried out at 180 – 240 °C.

Maleinized oils with 5 – 8 wt % of maleic anhydride (less than 1 mol per triglyceride) are esterified with glycerol and other polyols to form oil-modified polyesters. Their properties are similar to those of alkyds, and oil-modified polyesters dry faster than the original oils and form harder and more water-resistant films.

Maleinized oils with 16 – 24 wt % maleic anhydride are reacted with water or alcohols to yield maleinized oils with high acid numbers (100 – 200). After neutralization of the acid groups with ammonia or – more preferably – amines, the maleinized oil becomes water-thinnable [117]. The product is combined with phenolic or amino resins [118–120]. These water-thinnable combinations are mainly used as anticorrosive primers for electrodeposition. In electrodeposition the part to be coated is immersed in a tank and direct current is applied. In the process the binder systems (together with pigments and additives) are coagulated on the anode, which is the surface of the substrate to be coated [121–123]. The coagulated film can be rinsed with water and is then stoved at 175 – 180 °C for 20 – 30 min. The process ensures that film formation takes place on all parts which are immersed in the tank. Starting in the 1960s, such systems gained importance as corrosion protective primers in the automotive industry.

In the late 1970s the systems were mainly substituted by a resin system for a cathodic deposition process. But the maleinic oils are still in use for industrial primers.

8.8. Oil Boiling with Rosin, Copal, and Amber

In the past oils – and also drying oils – were used as solvent for different natural resins. The

target was to convert the natural resins into a suitable form for application and to combine the film properties of the natural resins with the drying properties of the oil. The higher the melting point of the natural resins, the better the film properties. Copals were better than rosins, and amber was better than copals. To improve the properties of rosins, rosins were transferred into rosin salts with calcium and zinc, into rosin esters with glycerol and pentaerythrol, and maleinized and esterified (→ Resins, Natural, Chap. 5.5.). It was not easy to melt the high molecular natural resins (copals and amber) and starting in England special know-how was gathered for this process.

The combinations of oils, different natural resins and related products and siccatives formed the basis for the oil varnishes, the first industrially used coatings.

8.9. Oil Boiling with Phenolic Resins

When the natural resins became rare the investigation to prepare synthetic resins began. The first synthetic resins used for industrial coatings were the phenolic resins (→ Phenolic Resins). But the normal phenolic resins (resols) are not compatible with oil. The first step to overcome the problem was the combination of rosin-modified phenolic resins with drying oils [124]. The second step was the preparation of so-called alkylphenolic resins (from alkyl-substituted phenols) [125], [126]. These alkylphenolic resins are able to react with oils, preferable tung oil, to yield adducts between the double bond system and the methylol and OH group of the phenolic resin [127]. The process is carried out at 240 °C, using a mixture of 2–4 parts tung oil and 1 part alkylphenolic resin. The reaction products have good drying properties and the films are highly resistant to water and chemicals. The product were used for industrial coatings and especially for yacht varnishes. For this application they were much more suited than alkyd varnishes. These combinations are often replaced by polyurethane coatings, but yacht varnishes are still available.

8.10. Modifications by Transesterification

If drying oils are heated with an excess of polyols, glycerol, pentaerythrol, or dipentaerythrol at 230 to 260 °C, mono- and diglycerides [128–131] are obtained. The alcoholysis (transesterification) is an equilibrium process. Salts of alkali or alkaline-earth metals serve as catalysts. To avoid the formation of incompatibilities of the catalyst in the reaction mixture or the final product, the use of lithium salts is preferred. The reaction is controlled by the added quantity of polyols, the reactivity of the OH groups, the temperature, and the time. The progress of the reaction is determined by measuring the solubility of the reaction mixture in methanol or ethanol. Transesterification products are starting materials for production of alkyds, urethane oils, and urethane alkyds.

Alkyds (→ Alkyd Resins) are prepared by reaction of the transesterification products and phthalic anhydride. The process was developed in 1927 and was the first important process for production of air-drying alkyds.

Urethane oils [132], [133] are prepared from the transesterification products and toluylene diisocyanate. Products obtained by reaction of the transesterification products with mixtures of phthalic anhydride (first esterified) and toluylene diisocyanate are called urethane alkyds. Urethane oils and urethane alkyds dry faster than alkyds to give harder and more resistant films, but the films are not weather-resistant. Urethane oils and urethane alkyds are used for wood coatings, preferably for parquet varnishes.

8.11. Modifications Based on Fatty Acids

Fatty acids are produced from drying oils by saponification. The fatty acids are the basis for the production of alkyds (→ Alkyd Resins) and epoxy esters. In contrast to oil varnishes and paints, alkyds and epoxy esters based on unsaturated fatty acids are still important coating materials.

9. Application, Importance, and Future Aspects

Oil varnishes and oil paints have been largely replaced by more modern systems, mainly alkyds, but there are still some oil paint products on the market. For people, who want to use natural raw materials, oil-based paints are still of interest for furniture coatings (mainly for "do-it-yourself"-applications), but there is nearly no industrial use. Products derived from drying oils, i.e., air-dryings alkyds and epoxy esters, and maleinized oils, still play a more important role in the paint industry. In Germany the market volume of solventborne coating materials is estimated to be approximately 300 000 t in 1999. Oil paints and oil varnishes have a share of only 0.6 %, but this is approximately 12 % more than in 1998. Coatings based on air-drying alkyds made up nearly 20 % of this volume, but the amount is decreasing in comparison to 1998 [134].

Advantages of oil-based or fatty acid-based coating materials are: excellent wetting of pigments and surfaces, easy application, good flow and leveling, low health risks, environmental compatibility, and optimum hydrophobicity of films. However, they suffer from two important disadvantages: (1) Oil- or fatty acid-based coating materials need too much time for film formation by air drying to be used in advanced industrial coatings application processes (e.g., automotive coating application) and (2) these coating materials are not easy to transfer into water-based systems – to avoid organic solvents for ecological and economical reasons – due to the sensitivity to react by saponification. On the other hand, since the 1980s there has been growing interest to increase the use of natural renewable raw materials. Drying oils and related products meet this demand. Several institutes started to investigate the possibility to decrease the agricultural food production in Europe and to replace growing of plants for food partly by growing of plants which produce raw materials for industrial use [135–140]. There are new proposals to grow plants which contain drying or semidrying oils or plants which contain fatty acids plus other chemically reactive groups.

10. References

General References

1. Albert Ilg (ed.): *Presbyter, Theophilus: Rodgerus von Helmarshausen: Schedula diversarium artium,* Verlag Wilh. Braumüller, Wien 1874.
2. Albert Ilg (ed.): *Heraklius: De coloribus et artibus Romanorum,* Verlag Wilh. Braumüller, Wien 1873.
3. K. Herberts: *Die Maltechniken,* Econ-Verlag, Düsseldorf 1957.
4. L. Stryer: *Biochemie,* Spectrum Verlag, Heidelberg 1990.
5. Unichema: Fatty Acid Data Book. 3rd. edition 1992.
6. E. Gulinsky: *Pflanzliche und tierische Fette u. Öle,* C. R. Vincentz Verlag, Hannover 1963.
7. H. Kittel: *Lehrbuch der Lacke und Beschichtungen.* Verlag W. A. Colomb, Stuttgart-Berlin 1971.

Specific References

8. W. Schuster, R. Marquard, *Fette, Seifen, Anstrichm.* **76** (1974) 207.
9. J. D. v. Mikusch, *Farbe + Lack* **58** (1952) 303 – 306.
10. J. D. v. Mikusch, *Farben, Lacke, Anstrichst.* **4** (1950) 149.
11. J. Baltes, *Fette Seifen* **52** (1950) 462.
12. J. Baltes, F. Weghorst, O. Wechmann, *Fette, Seifen, Anstrichm.* **63** (1961) 413.
13. A. E. Rheineck, D. D. Zimmerman, *Farbe + Lack* **70** (1964) 641.
14. C. I. Atherton, A. F. Kertess, *Fette, Seifen, Anstrichm.* **68** (1966) 279.
15. E. Fonrobert: *Das Holzöl,* Berliner Union-Verlag, Stuttgart 1951.
16. T. P. Hilditch, A. Mendelewitz, *J. Sci. Food Agric.* **2** (1951) 548.
17. M. Hassel, W. Lawrence, *Paint Ind.* **74** (1959) no. 7, 10, and no. 8, 15.
18. J. D. Hetchler, *J. Am. Oil Chem. Soc.* **31** (1954) 503.
19. R. J. De Sesa, *Farbe + Lack* **69** (1963) 752.
20. H. W. Chatfield, *Paint Manuf.* **30** (1960) 45.
21. H. A. Bhatt, P. V. Tagdiwala, *Paintindia* **29** (1970) no. 4, 19; no. 7, 19.
22. G. B. Kromer, *Oleagineux* **18** (1963) 21.
23. A. E. Rheineck, L. O. Cummings, *J. Am. Oil Chem. Soc.* **43** (1966) 409.
24. W. Schuster, *Fette, Seifen, Anstrichm.* **74** (1972) 150.
25. A. Müller, *Farbe + Lack* **57** (1951) 240.
26. F. G. Sietz, *Fette, Seifen, Anstrichm.* **67** (1965) 411.

27. N. Minoru et al., *Farbe + Lack* **71** (1965) 839.
28. G. W. Priest, J. D. v. Mikusch, *Ind. Eng. Chem.* **32** (1940) 1314.
29. N. A. Ghanem, Z. H. Abd El-Latif, *J. Paint Technol.* **39** (1967) 144.
30. C. I. Atherton, A. F. Kertess, *J. Oil Colour Chem. Assoc.* **49** (1966) 340.
31. W. Sandermann: *Naturharze, Terpentinöl, Tallöl,* Springer Verlag, Berlin 1960.
32. W. Asche, *Farbe + Lack* **68** (1962) 448 – 453, 518 – 525.
33. H. W. Chatfield, *Paint Oil Colour J.* **141** (1962) 1262.
34. G. Eick, *Farbe + Lack* **69** (1963) 616.
35. W. M. Kraft et al., *J. Am. Oil Chem. Soc.* **42** (1965) 96.
36. K. B. Gilkes, T. Hunt, *J. Oil Colour Chem. Assoc.* **51** (1968) 389.
37. J. T. Geoghegan, H. G. Arlt, C. O. Myatt, *J. Paint Technol.* **40** (1968) 209.
38. R. V. V. Nicholls, W. H. Hoffman, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **24** (1952) no. 4, 245 – 260.
39. H. P. Kaufmann: *Analyse der Fette und Fettprodukte,* Springer Verlag, Berlin 1958.
40. DIN 53 782 and DIN 55 957.
41. G. H. Hutchinson, *J. Oil Colour Chem. Assoc.* **56** (1973) 44.
42. P. O. Powers, *Ind. Eng. Chem.* **41** (1949) 304.
43. J. R. Chipault, E. C. Nickell, W. O. Lundberg, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **23** (1951) no. 11, 740.
44. L. Dulog, *Congr. FATIPEC 8th, 1966,* 17 – 25.
45. H. R. Rawls, P. J. Van Santen, *J. Am. Oil Chem. Soc.* **47** (1970) 121.
46. R. Criegee, *Justus Liebigs Ann. Chem.* **522** (1936) 84.
47. A. Rieche, *Angew. Chem.* **50** (1937) 520.
48. E. H. Farmer, *Trans. Faraday. Soc.,* **38** (1942) 340.
49. H. P. Kaufmann, *Fette, Seifen,* Anstrichm. 59 (1957) 153.
50. F. Hasbeck, W. Grosch, *Lipids* **18** (1983) 706.
51. E. N. Frankel, C. D. Evans, D. G. McConnell, E. Selke, et al., *J. Org. Chem.* **26** (1961), 4663.
52. F. Haslbeck, W. Grosch, J. Firl, *Biochim. Biophys. Acta* **750** (1983) 185.
53. F. Haslbeck, W. Grosch, *Fette, Seifen, Anstrichm.* **86** (1984) 408.
54. A. M. Gaddis, R. Ellis, G. T. Currie, *J. Am. Oil Chem. Soc.* **38** (1961) 371.
55. J. H. Skellon, *J. Oil Colour Chem. Assoc.* **46** (1963) 1001.
56. R. Clement, *Congr. FATIPEC 14th, 1978,* 177 – 182.
57. J. H. Hartshorn, *J. Coat. Technol.* **54** (1982) no. 687, 53 – 61.
58. P. S. Hess, G. A. O'Hare, *Ind. Eng. Chem.* **42** (1950) 1424.
59. M. Dyck, *Farbe + Lack* **67** (1961) 442.
60. H. Wexler, *Chem. Rev.* **64** (1964) 591.
61. W. Treibs, *Fette Seifen* **54** (1952) 3.
62. R. N. Faulkner, *J. Appl. Chem.* **8** (1958), 448.
63. L. A. O'Neill, *Chem. Ind. (London)* 1954, 384.
64. W. J. Stewart, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **26** (1954) no. 6, 413 – 425.
65. DIN 55 901.
66. J. Skalsky, *Prog. Org. Coat.* **4** (1976) 137.
67. R. Hurley, F. Buono, *J. Coat. Technol.* **54** (1982) no. 694, 55 – 61.
68. R. G. Middlemiss, *J. Water Borne Coat.* **8** (1985) no. 11, 3 – 9.
69. C. E. H. Bawn, *Discuss. Faraday Soc.* **14** (1953) 181.
70. E. R. Mueller, *Ind. Eng. Chem.* **46** (1954) 562.
71. N. Uri, *Chem. Ind. (London)* 1956, 515 – 517.
72. C. E. H. Bawn, J. A. Sharp, *J. Chem. Soc.* 1957, 1854.
73. F. W. Heaton, N. Uri, *J. Lipid Res.* **2** (1961) 152.
74. M. Allard, *Double Liaison* **18** (1971) no. 194, 503 – 523.
75. N. S. Baer, N. Indictor, *J. Coat. Technol.* **48** (1976) no. 623, 58 – 62.
76. M. Giesen, *Dtsch. Farben Z.* **20** (1966) 236.
77. E. Krejcar, K. Hájek, O. Kolář, *Farbe + Lack* **74** (1968) 115.
78. G. Rieck: *Congr. FATIPEC 10th, 1970,* 239.
79. M. Giesen, *Congr. FATIPEC 7th, 1964,* 349.
80. F. Franks, B. Roberts, *J. Appl. Chem.* **13** (1963) 302.
81. A. C. Elm, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **29** (1957) no. 4, 351 – 385.
82. O. S. Privett, M. L. Blank, J. B. Covell, W. O. Lundberg, *J. Am. Oil Chem. Assoc.* **38** (1961) 22.
83. L. A. O'Neill, S. M. Rybicka, T. Robey, *Chem. Ind. (London)* 1962, 1796 – 1797.
84. L. A. O'Neill, *Paint Technol.* **27** (1963) 44.
85. G. Link, *Farbe + Lack* **92** (1986) 279.
86. C. Boelhouwer, J. Th. Knegtel, M. Tels, *Fette, Seifen, Anstrichm.* **69** (1967) 432.
87. D. H. Wheeler, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **23** (1951) no. 11, 661 – 668.
88. R. F. Paschke, D. H. Wheeler, *J. Am. Oil Chem. Soc.* **26** (1949) 278.
89. R. F. Paschke, D. H. Wheeler, *J. Am. Oil Chem. Soc.* **31** (1954) 208.
90. A. L. Clingman, D. E. A. Rivett, D. A. Sutton, *J. Chem. Soc.* 1954, 1088 – 1090.

91. J. A. MacDonald, *J. Am. Oil Chem. Soc.* **33** (1956) 394.
92. G. H. Hutchinson, *J. Oil. Colour Chem. Assoc.* **41** (1958) 474.
93. S. B. Radlove, H. M. Teeter, W. H. Bond, J. C. Cowan, J. P. Kass, *Ind. Eng. Chem.* **38** (1946) 997.
94. J. D. von Mikusch, *Farbe + Lack* **57** (1951) 341 – 345.
95. J. D. von Mikusch, *Farbe + Lack* **57** (1951) 393 – 399.
96. J. D. von Mikusch, *Farben, Lacke, Anstrichst.* **4** (1950) 149.
97. P. L. Nichols, S. F. Herb, R. W. Riemenschneider, *J. Am. Chem. Soc.* **73** (1951) 247.
98. J. Baltes, F. Weghorst, O. Wechmann, *Fette, Seifen, Anstrichm.* **63** (1961) 413.
99. BE 13 378, 1900 (A. Kronstein).
100. J. Scheiber, *Farbe + Lack* **63** (1957) 443.
101. S. R. Vranish, R. M. Jackman, R. F. Helmreich, W. A. Henson, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **29** (1957) no. 1, 56 – 74.
102. H. P. Kaufmann, H. Gruber, *Fette, Seifen, Anstrichm.* **62** (1960) 607.
103. H. P. Kaufmann, H. Brüning, *Fette, Seifen, Anstrichm.* **62** (1960) 1146.
104. F. R. Mayo, C. W. Gould, *J. Am. Oil Chem. Soc.* **41** (1964) 25.
105. K. Hamann, O. Mauz, *Fette, Seifen, Anstrichm.* **58** (1956) 528.
106. J. B. Crofts, *J. Appl. Chem.* **5** (1955) 88.
107. E. F. Redknap, *J. Oil Colour Chem. Assoc.* **43** (1960) 260.
108. M. Dyck, *Farbe + Lack* **67** (1961) 148.
109. US 2 033 131, 1931 (Ellis-Foster).
110. US 2 033 132, 1931 (Ellis-Foster).
111. US 2 188 882, 1934 (E. T. Clocker).
112. Springer & Möller, DE 635 926, 1935.
113. W. G. Bickford, E. F. DuPre, C. H. Mack, R. T. O'Connor, *J. Am. Oil Chem. Soc.* **30** (1953) 376.
114. J. Sauer, R. Sustmann, *Angew. Chem.* **92** (1980) 779.
115. C. P. A. Kappelmeier, J. H. van der Neut, *Kunststoffe* **40** (1950) 81.
116. A. E. Rheineck, T. H. Khoe, *Fette, Seifen, Anstrichm.* **71** (1969) 644.
117. A. E. Rheineck, R. A. Heskin, *J. Paint Technol.* **40** (1968) 450.
118. Glasurit Werke, DE-AS 15 19 167, 1965 (G. Tröger, H. Noak).
119. Am. Cyanamid, DE-OS 16 69 593, 1967 (J. N. Koral).
120. PPG, DE-OS 1938 191, 1969 (E. J. Kapalko, W. H. English).
121. Ford Motor Co., DE-AS 15 46 944, 1962 (A. B. Gilchrist).
122. ICI, DE-OS 15 46 995, 1963 (N. McPherson, J. P. Burden).
123. Herberts, DE-AS 2 737 174, 1977 (H.-P. Patzschke, D. Saatweber).
124. G. Dantlo, *Peintures, Pigments, Vernis* **23** (1947) 171.
125. Koller & Co., EP 334 527, 1929; FP 676 456, 1929 (H. Hönel, Beck).
126. V. H. Turkington, I. Allen, *Ind. Eng. Chem.* **33** (1941) 966.
127. H. Hultzsch, *Kunstst.* **37** (1947) 43.
128. K. Hajek, J. Stanek, J. Hires, *Plaste Kautsch.* **15** (1968) 679.
129. R. Schöllner, L. Läbisch, *Fette Seifen Anstrichm.* **69** (1967) 426.
130. N. A. Ghanem, F. F. Abd. El-Mohsen, *J. Oil Colour Chem. Assoc.* **50** (1967) 441.
131. F. Mort, *J. Oil Colour Chem. Assoc.* **47** (1964) 919.
132. A. C. Jolly, *J. Oil Colour Chem. Assoc.* **47** (1964) 919.
133. K. Hajek, J. Kitzler, J. Stanek, E. Krejcar, *Farbe + Lack* **77** (1971) 422.
134. *Statistics from Verband der Lackindustrie e.V.,* Frankfurt, Germany 1999.
135. J. T. P. Derksen, B. G. Muuse, F. P. Cuperus in D. J. Murphy (ed.): *Designer Oil Crops, Breeding, Processing and Biotechnology,* VCH, Weinheim 1993.
136. J. T. P. Derksen, F. P. Cuperus, *Industrial Crops and Products* **3** (1995) 57.
137. B. G. Muuse, F. P. Cuperus, J. T. P. Derksen, *Industrial Crops and Products* **3** (1992) 57.
138. H. G. Hauthal, "Bunsentage 1992, Nachwachsende Rohstoffe," *Nachr. Chem. Techn. Lab.* **40** (1992) no. 9, 996.
139. E. M. S. Hammersveld, F. P. Cuperus, J. T. P. Derksen: Modified and Novel Veg. Oils in a New Gen. of Emulsion Paints, Agrotechnical Research Institute, Wageningen, Netherlands 1994.
140. Project: VOICI (Vegetable Oils for Innovation in Chem. Ind.) in ECLAIR (Europ. Collaborative Linkage of Agriculture and Industry through Research), Brüsselsss 1993.

Dyes, General Survey

Gerald Booth, Imperial Chemical Industries PLC, Manchester, United Kingdom (Chap. 1)

Heinrich Zollinger, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland (Chap. 2)

Keith McLaren, Consultant, Upper Woolhampton, Berkshire, United Kingdom (Chap. 3)

William Gibbard Sharples, Imperial Chemical Industries PLC, Manchester, United Kingdom (Chaps. 4 – 6)

Alan Westwell, Imperial Chemical Industries PLC, Manchester, United Kingdom (Chaps. 4 – 6)

1. Survey 2
1.1. History 2
1.2. Classification and Definitions 3
1.3. Manufacture 5
1.3.1. Unit Operations 5
1.3.2. Formulation and Standardization . . . 6
1.4. Applications 6
1.4.1. Wool 6
1.4.2. Cellulosics 7
1.4.3. Polyester 8
1.4.4. Polyamide 9
1.4.5. Acrylics 9
1.4.6. Pigments 9
1.5. Economic Aspects 9
1.6. Environment, Ecology, and Toxicology 10
2. Color and Structure 11
2.1. Basic Principles of Color 11
2.2. Empirical Correlations 14
2.3. Quantum Chemical Methods 15
2.4. Fluorescence and Phosphorescence 20
2.5. Quantitative Treatment of Light Absorption 22
2.6. The Influence of Substituents on the Spectra of Aromatic Compounds . . 26
3. Color Measurement and Colorant Formulation 28
3.1. Measurement of Reflectance Factors 28
3.2. Calculation of Tristimulus Values . 29
3.3. Chromaticity Diagram 31
3.4. Conversion of *XYZ* Values into *L*a*b Coordinates** 32
3.5. Calculation of Color Differences . . 34
3.6. Colorant Formulation 35
4. Chemical Analysis and Tests 36
4.1. Investigation of Individual Dyes . . 36
4.1.1. Testing for Homogeneity and Separation of Mixtures 36
4.1.2. Classification 38
4.1.3. Identification 39
4.1.4. Quantitative Determination 40
4.1.5. Determination of Structure 41
4.2. Investigation of Dyes on the Fiber . 43
4.2.1. General Tests 43
4.2.2. Systematic Analysis 44
4.2.3. Special Tests 44
5. Chromatography 48
5.1. Column Chromatography 48
5.2. Paper Chromatography 48
5.3. Thin-Layer Chromatography 49
5.4. High-Performance Liquid Chromatography (HPLC) 50
6. Spectroscopic Measurements 51
6.1. Nuclear Magnetic Resonance (NMR) 51
6.2. Mass Spectrometry (MS) 53
6.3. UV-Visible Spectrophotometry . . . 53
7. References 53

1. Survey

1.1. History [10]

In 1834 Runge isolated aniline from coal tar and observed the formation of aniline black on oxidation. Picric acid was used as a silk dye in 1849. The first synthetic dye of technical significance, called Mauveine after its mauve color, was obtained in 1856 by Perkin in an attempt to synthesize quinine. With his father and brother, he founded the first factory to manufacture synthetic dyes in Greenford Green, near London. This was the start of the dye and pigment industry. The year 1856 also saw the synthesis of fuchsine. The manufacture of fuchsine, later called magenta, began in 1859.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_073

In 1862 GRIESS created *azo dye* chemistry with the discovery of diazo compounds [11]. Among the earliest *azine dyes* were induline in 1863 and nigrosine in 1867. The first synthesis of an *anthraquinone dye* was that of alizarin in 1868. The first *sulfur dye* was obtained in 1873, and indigo was synthesized in 1878.

Progress became very rapid with the advent of structural organic chemistry; many thousands of dyes were synthesized within a few decades. Many of the large chemical plants in operation in Germany and Switzerland today were built as dye factories in the second half of the 19th century. The manufacture of dyes has spread from Europe throughout the world; Western Europe still accounts for ca. 40 % of the worldwide total.

Azo Dyes. Bismarck Brown was introduced in 1863, and chrysoidin was introduced in 1875. Azoic dyeing became possible in 1880 with the formation of insoluble azo dyes on the fiber. In 1884 the use of benzidine-based disazo dyes such as Congo Red made possible the dyeing of cotton without a mordant.

Post-treatment of dihydroxy azo dyes with chromium salts was introduced in 1893, and the use of arylamides of 2-hydroxy-3-naphthoic acid (Naphtol AS) as coupling components was first seen in 1911.

Simple 1 : 1 preformed chromium and copper complexes were first described in 1912 leading to 2 : 1 chromium and cobalt complexes specifically designed for wool in 1949. Disperse dyes for acetate fiber became available in 1921, and triazine polyazo dyes were introduced in 1922. Reactive dyes for wool and cellulose were introduced in 1952 and 1956, respectively.

Anthraquinone Dyes. The early history is reviewed in [12]. Alizarin (1,2-dihydroxyanthraquinone) was first synthesized in 1868. Mercury-catalyzed sulfonation of anthraquinone was discovered in 1901. Indanthrene Blue, the first anthraquinone vat dye, was produced in 1901. Benzanthrone, as well as vat dyes prepared from it, followed in 1904. The phthaloyl acridones, vat dyes of high lightfastness were discovered ca. 1910, and Caledon Jade Green (dimethoxy violanthrone) was an important discovery in 1920. Sulfuric esters of leuco vat dyes were introduced in 1924.

Sulfur Dyes. Cachou de Laval, the first useful sulfur dye, was discovered in 1873. Improvements followed in 1893, and in 1897, Immedial Black V Extra became the first major commercial product.

Indigo and Indigoid Dyes. After the historic synthesis of indigo by VON BAEYER in 1878, more economical methods were devised, and production began in 1897.

Sodium amide was first used in conversion of phenylglycine to indigo in 1901; this process is still employed. Thioindigo was first synthesized in 1906, and stable water-soluble sulfuric esters of leuco indigoid dyes were introduced in 1921.

Triphenylmethane and Xanthene Dyes. Phthaleins such as phenolphthalein and fluorescein were discovered by VON BAEYER in 1871. Malachite Green and Fuchsine S were introduced in 1877, followed in 1883 by Crystal Violet. The important xanthene dye Rhodamine B was discovered in 1877.

Phthalocyanines [13]: Copper phthalocyanine was discovered in 1927 [14], and its structure elucidated in 1933. In 1935 it came on the market as Monastral Fast Blue B (ICI) and as Heliogen Blue B (I. G.). In 1951 the discovery was made that metal phthalocyanines can be produced on the fiber [15].

1.2. Classification and Definitions

Dyes and pigments can be classified according to chemical structure or the mode of application. The most important commercial products are the azo, anthraquinone, sulfur, indigoid, triphenylmethane, and phthalocyanine dyes. The principal chromophores are as follows (Dye classes marked with an asterisk and those referred to by an arrow are separate keywords in this Encyclopedia):

Acridine (→ Triarylmethane and Diarylmethane Dyes)
Anthraquinone *
Azine (including oxazine and thiazine) *
Azo *
Benzoquinone and naphthoquinone *

Formazan (→ Metal Complex Dyes)
Indigoid ∗
Metal complex ∗
Methine (including azomethine) ∗
Naphthalamide ∗
Nitro ∗
Nitroso ∗
Phthalocyanine ∗
Quinacridone (→ Pigments, Organic)
Quinophthalone (→ Methine Dyes and Pigments)
Stilbene (→ Optical Brighteners ; → Azo Dyes)
Sulfur ∗
Tri- and diarylmethane ∗
Xanthene (→ Triarylmethane Dyes and Diarylmethane Dyes)

Application. The same chromophores are found in both dyes and pigments. The former are applied mainly to textiles, but also to paper, leather, food and other products. Pigments are used in paints, inks, and polymers. Other specialized applications are covered in separate articles (→ Artist's Colors ; → Ceramic Colorants ; → Laser Dyes ; → Photography).

The Colour Index [16] is based on the classification summarized in Table 1. Principal areas of application are emphasized. It can be seen, for example, that for dyeing cotton one can choose among five classes of dyes. Every dye and pigment manufacturing company has its own trademark to identify each product class. The Colour Index (C.I.) lists all these (some 8000 products) and, when chemical constitutions are available, the equivalent products under the same number [16]. The C. I. names and numbers are indexed in this Encyclopedia.

It is worth noting that by 1980 the world production of the main synthetic fibers (polyester, polyamide, acrylic, viscose, and acetate) had reached equality with the world production of natural fibers (cotton and wool). It follows that the consumption of textile dyes is about equally divided between natural and synthetic fibers.

1.3. Manufacture

This survey considers unit operations of general significance, but excludes unit processes such as nitration, reduction, and sulfonation, important enough to be treated separately (→ Nitro Compounds, Aromatic; → Reduction ; → Benzenesulfonic Acids Their and Derivatives). The manufacture of the chemical, the so-called color paste, and that of the finished product are separate operations, although the two are often fully integrated. Finishing (or conditioning) changes the physical form, most simply by drying and grinding; this technology is as important as the chemical synthesis, especially in the manufacture of pigments and water-insoluble dyes. These two classes should be distinguished from water-soluble dyes at the finishing stage. Earlier processing stages may also have a significant influence on the final physical form.

Although unit operations are fully described in the B-series, the dye and pigment industry is rarely mentioned, because it has very few requirements peculiar to itself.

1.3.1. Unit Operations

Reactants and the reaction medium (water or solvent) are delivered to the reactor using standard materials handling procedures. The dye industry usually employs a batch process; agitation and mixing are critical.

Crystallization and Filtration. Crystallization of the final product requires careful control to obtain the most desirable physical form; the filtration equipment is chosen accordingly, with a view to both speed and maximum liquid removal. The suspension to be filtered usually contains 1 – 10 % solids with a concentration to an average of 50 % after filtration.

Crystallization is effected by one or more of the following techniques:

1) Variation of reaction conditions
2) Cooling a solution below the crystallization temperature
3) Solvent evaporation
4) Dilution, usually by addition to water
5) Adjustment of pH
6) Salting-out by addition of an inorganic cation, usually Na in the form of salt

Table 1. Dye classification

Class	Properties	Chromophores (main)	Substrate suitability					
			Wool	Cotton, Viscose	Polyester	Polyamide	Acrylic	Others
Acid	water-soluble, anionic	azo (incl. premetallized chromium complexes), anthraquinone	X			X		leather silk
Azoic	water-insoluble azo dye formed on fiber	azo		X				
Basic	water-soluble, cationic	triphenylmethane, methine, modified azo, anthraquinone					X	acid-modified synthetics paper
Direct	water-soluble, anionic	azo (incl. copper complexes), stilbene, phthalocyanine		X				paper
Disperse	water-insoluble, nonionic	azo, anthraquinone, nitro			X	X		acetate
Mordant	water-soluble, anionic, chromium complex formed on fiber	azo	X					
Reactive	water-soluble, anionic, forms covalent bond with substrate	azo (incl. premetallized), anthraquinone, phthalocyanine	X	X				
Sulfur	temporarily solubilized with alkali sulfide	sulfur		X				
Vat	temporarily solubilized as leuco ester with alkaline sodium hydrosulfite	anthraquinone indigoid		X				
Pigment	water-insoluble, nonionic	azo anthraquinone phthalocyanine quinacridone						polymers, paints, inks, etc.

The choice of filter is based on the physical form of the product and the availability of equipment. The physical form may need to be adapted to an existing filter, or the best combination of physical form with efficient filtration may need to be developed. A complication may be introduced by the need to wash the product.

Filters are available in six types: nutsch filters (vacuum filters), pressure filters, filter presses, diaphragm presses, continuous vacuum filters, and centrifuges. The filtration product is usually termed the color paste.

Drying. Drying is usually necessary unless the color paste is to be converted to a liquid product. The water content of these pastes varies from 20 to 80 %; drying costs vary accordingly.

The choice of dryer is governed by the following:

1) *Scale,* i.e., quantity of product and of water to be removed
2) *Physical form* and characteristics of feedstock
3) *Hazards* or other decomposition problems caused by drying temperatures of 50 – 180 °C
4) *Properties* of the product
5) *Physical form requirements* for subsequent processing or marketing
6) *Capital and operating costs*

The following dryer types are used: ovens, agitated vacuum dryers, filter dryers, spray dryers, drum dryers, through-circulation dryers, and fluidized-bed dryers. Ovens produce dry lumps, spray dryers produce grains. Other types of dryers produce a variety of aggregated forms.

Milling and Grinding. Particle size is reduced by grinding or milling. Dry grinding is usually carried out in impact mills equipped to control dust. Continuous drying methods have been developed to eliminate the final product comminution stage.

The ball mill, consisting of a rotating or vibrating cylinder containing hard balls, has always been limited in use for dry grinding because of product removal problems. It has been superseded by sand and bead mills. Wet milling is increasingly important for pigments and disperse dyes. Many designs consist of vertical or horizontal cylinders equipped with high-speed agitators and continuous feed and discharge. Sand or fine glass beads remain in the cylinder and effect comminution as the suspension of dye or pigment flows through. The objective is to achieve uniform particle-size distribution with a minimum expenditure of energy. A multishear mixer in the feed vessel sometimes provides a premilling stage.

1.3.2. Formulation and Standardization

The tendency has been to integrate standardization with color manufacture. Drying is frequently integrated with color production; modern continuous dryers afford a wet-standardized product ready for sale. This is made possible by the use of rapid instrumental color matching against standard products. The finishing stage includes the incorporation of diluents and dispersing, wetting, and dust-control agents. The diluents may be common salt or Glauber's salt, but if electrolytes are undesirable, urea or dextrine may be used. Dyes isolated as free acids may be blended with alkali (sodium carbonate, sodium hydrogen carbonate, or trisodium phosphate) to form the sodium salt when the mixture is dissolved. Reactive dyes are buffered to prevent hydrolysis during storage.

Table 2. Physical form specifications

Form	Specification	
	as sold	as used
Solids		
Grains	dust-free	no precipitation
Powders	good solubility or dispersibility	no adverse additives effects
	good wetting	foam-free
Liquids		
Ionic	stable to storage	stable throughout dyeing
	pourable	
Nonionic	speck-free	

To meet market needs, it is often necessary to produce a given dye in more than one strength or brand. For hygienic reasons, an increasing number of water-soluble dyes are being produced in

aggregated or granular rather than powder form by the use of modified drying techniques.

These procedures apply equally to liquid products, such as solutions of basic dyes or dispersions of water-soluble dyes. Water, or occasionally an organic solvent, is the diluent, but additives are usually required to control pH, viscosity, stability, freezing, and evaporation.

Table 2 shows physical specifications for textile dyes.

1.4. Applications

In dyeing processes, textiles, paper, or leather are usually heated in a solution or suspension of dye. In printing, a dye or pigment paste is applied to the surface of the substrate and then heated. Increasingly sophisticated equipment is employed toward energy saving and cost reduction [17].

1.4.1. Wool

Wool consists of a protein called keratin, which is built from 18 different amino acids of the general formula $H_2N-CHR-COOH$. The chains have the general form

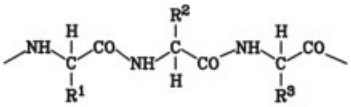

and are united at intervals by bridges derived from cystine.

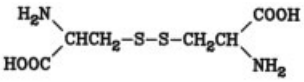

Some of the side chains end in amino groups and others in carboxyl groups; internal salt formation binds the molecules together by electrovalent linkages. The average molecular mass of keratin is ca. 60 000.

Wool has the most complex physical structure of all textile fibers. A central cortex composed of a bundle of fibrils is surrounded by a scaly covering known as the cuticle. Coarse wools may also contain a medulla, or central canal. Affinity for dyes is not uniform throughout the fiber, the cuticle being hydrophobic and resistant to penetration; however, it does not usually extend to the fiber tips. It is subject to damage during processing, where alkaline conditions are especially destructive, causing rupture of the cystine links. The fiber is more resistant to acid, but prolonged boiling causes hydrolysis of peptide linkages. Therefore, uniform dyeing of wool is difficult, especially with a mixture of dyes. By virtue of its carboxylic and amino groups, wool has affinity for basic and acid dyes but the former are rarely used. Almost all acid dyes contain the sulfonic acid group which causes adsorption by ionic attraction to the basic groups in wool. The resulting link is easily broken, however, and cannot fully explain the washfastness shown by many acid dyes. In these cases attachment to the fiber appears to depend mainly on nonpolar van der Waals forces, the strength of which is proportional to the area of contact. Acid dyes are of two general types, the leveling type that diffuses readily through the fiber and the milling type that does not and is, thus, faster to wet treatment. Leveling dyes are applied in the presence of sulfuric acid (pH 2.5 – 3.0) and Glauber's salt, but milling dyes, having greater affinity, must be applied under mildly acidic conditions (acetic acid, pH 4.5 – 5.5); in some cases, ammonium sulfate is used (pH 6.0 – 8.5). In more strongly acidic conditions, neutral-dyeing acid dyes give uneven shades.

The washfastness of wool dyed with an anthraquinone dye containing an *o*-dihydroxy group can be improved by posttreatment with a chromium salt such as sodium or potassium dichromate. Early in the 20th century, it was found that *o,o′*-dihydroxyazo dyes give similar results. Such dyeings are known as Chrome colors; their exceptional fastness is due to the formation of chromium complexes within the fiber.

Afterchroming modifies shades and creates difficulties in color matching. This disadvantage has been overcome by the use of preformed metal complexes (→ Metal Complex Dyes). The early 1 : 1 chromium complexes were applied by prolonged boiling in a strongly acidic bath which damaged the fiber. The later 1 : 2 chromium complexes required less severe conditions. Unsymmetrical 1 : 2 metal complexes meet still more varied requirements.

Reactive wool dyes provide bright fast shades when applied at carefully controlled pH. Leveling is controlled by temperature and additives.

Wool may be dyed in several forms:

1) *Loose wool* fleece has been merely scoured free of natural grease and minerals. *Slubbing* consists of thick, loose untwisted ropes of parallel wool fibers. Such wools are usually dyed with the faster acid and chrome dyes. The fibers are more easily penetrated at this stage than in yarn or piece form.
2) Yarn, i.e., in thread form, is standard for carpets or knitting yarn. Acid dyes are most commonly employed, with chrome dyes and 1 : 2 metal complexes used less often.
3) *Piece goods,* i.e., woven or knitted fabric, provide the cheapest method and is used whenever possible. It is, of course, limited to solid shades, as in dress fabrics. Dyes with good leveling and fastness properties are required.

Various finishing treatments may be employed after dyeing. *Crabbing* gives worsteds a more or less permanent set, and *potting* confers a characteristicly smooth surface on such materials for dress suits; both involve treatment with hot water. A more drastic process known as *milling* involves prolonged treatment between rollers with soap and sodium carbonate solutions; this causes the wool to "felt," i.e., become more compact. Dyes must be fast to these processes without loss of color or staining of white or lightly colored threads in the fabric.

1.4.2. Cellulosics

Cotton is almost pure cellulose [*9004-34-6*], a linear polymer of β-D-glucose units (→ Cellulose). The empirical formula is $(C_6H_{10}O_5)_n$ with *n* between 600 and 1000 and molecular mass upward of 100 000.

Viscose rayon is regenerated wood cellulose. Although similar chemically, cotton and viscose differ somewhat in structure and, therefore, in dyeing properties. Both the chemistry (especially the presence of the hydroxyl groups) and the physics of the fiber are important in dyeing.

Direct dyes are substantive to cellulose without a mordant; historically they replaced naturally occurring mordant dyes. Direct dyes are the most important single class, and are simple to use. Optimum conditions vary, but dyeing is usually carried out at or near the boiling point in the presence of common salt or Glauber's salt; the ratio of liquor to goods is fairly high (typically 20 : 1 in fabric-dyeing machinery). The lightfastness of metallizable dyes can be increased by posttreatment with copper salts, and washfastness can often be improved by resin treatment. Shades are greatly affected by metal salts but less so by resins imparting crease resistance.

Reactive dyes offer brighter shades of higher wetfastness than direct dyes. Batchwise application resembles dyeing with direct dyes ("salting" onto the fiber from hot aqueous solution) but is followed by alkaline treatment to fix the dye on the fiber by a chemical reaction with the hydroxyl groups of the cellulose (→ Reactive Dyes).

Continuous dyeing and printing are attractive for reactive dyes, permitting the alkali to be applied simultaneously in a single-stage process. All of the important chromophores are available as combinations with a variety of reactive groups [18].

Vat dyes are more expensive than direct or reactive dyes, but are used when the highest lightfastness is required, for example, for home-furnishing fabrics. They are applied from an aqueous solution of the leuco dye (vatting stage) obtained by alkaline reduction with sodium hydrosulfite (dithionite). Oxidation reforms the insoluble dye within the substrate. Subsequent treatment in a boiling detergent solution gives maximal fastness.

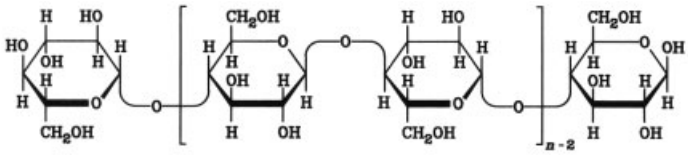

Sulfur dyes are applied similarly to vat dyes, with the dye being dissolved by sodium sulfide reduction; sodium carbonate may also be present. Dichromate is commonly used as an oxidant. Dull brown to black shades are obtained cheaply by this process but fastness properties are modest. The use of sulfide may create environmental problems, however.

Azoic colors are insoluble azo dyes produced from the intermediates directly on the fiber. They were once important but are now in marked decline. Their excellent properties are offset by their limited shade range and the complexity of the application procedure compared with, for example, those of reactive dyes. In the simplest form of the process, a coupling component with affinity for the fiber is applied from an alkaline solution to cotton or other cellulosic material. After removal of excess liquor, the color is developed on the fiber by immersion in a bath containing a diazonium compound. Pigment particles formed within the fiber are aggregated by hot soaping, which enhances resistance to wet treatment.

1.4.3. Polyester

Poly(ethylene terephthalate) [*25038-59-9*], M_r 15 000 – 20 000, is highly hydrophobic and

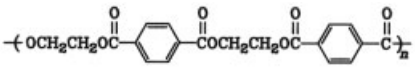

lacks reactive groups (→ Fibers, 4. Synthetic Organic). The tightly packed molecules do not easily accommodate dye molecules. Disperse dyes applied at the boiling point under conditions suitable for acetate or polyamide give only pale shades; temperatures of 120 – 130 °C in pressure equipment are usually required (→ Disperse Dyes).

A simpler process employed at atmospheric pressure makes use of an agent called a *carrier.* This is an emulsion of an aromatic compound such as methylnaphthalene, diphenyl, or trichlorobenzene. Carriers are really plasticizers that increase the segmental mobility of the polyester chain and, thus, facilitate entry of dye molecules. This method creates environmental problems, and removal of the carrier after dyeing is troublesome, but it is still widely used for polyester-wool fabric.

The *Thermosol process* is used for dyeing polyester in cellulosic blends. The scoured fabric is padded with a disperse dye and a migration inhibitor. The fabric is then dried and heated to 175 – 200 °C for ca. 1 min, during which the applied dye is absorbed by the polyester component. The cellulosic component can be dyed simultaneously or afterwards.

Slow diffusion of dyes below 100 °C usually creates good wetfastness properties. Lightfastness can be improved by structural modification of the dye molecule. Some small nonionic molecules are subject to sublimation when treated with a hot iron. This property can be exploited in transfer printing, a process of minor importance.

1.4.4. Polyamide

The commercial products include Nylon 6,6 [*9011-55-6*] (adipic acid copolymer with hexamethylene diamine, average M_r 12 000 – 20 000), and Nylon 6 [*25038-54-4*] (polymerized caprolactam).

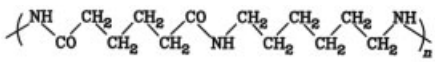

Nylon 6,6

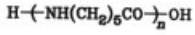

Nylon 6

Nylon 6,6 and Nylon 6 can be dyed by certain disperse dyes that show good leveling properties and are more effective than acid dyes in concealing irregularities in the filament. For wetfastness, however, acid milling or premetallized dyes are required; these dyes withstand wet treatments even better on polyamide, with its hydrophobic character, than on wool. Lightfastness is usually slightly lower on polyamide than on wool. The leveling properties of acid milling and premetallized dyes are poor on polyamide, especially Nylon 6, which contains a higher proportion of free amino groups and has greater affinity for acid dyes.

1.4.5. Acrylics

Unmodified polyacrylonitrile [*25014-41-9*] is so difficult to dye that copolymers are used to introduce acidic or basic sites. Thus, 4-styrenesulfonic acid [*98-70-4*] copolymer gives a fiber that accepts basic dyes.

The classical basic dyes, e.g., triphenylmethanes, give strong shades that are usually poor in lightfastness and modified basic dyes have been developed especially for polyacrylonitrile. These are applied from aqueous solution with careful temperature control up to the boiling point to achieve level dyeing. Fastness properties are excellent.

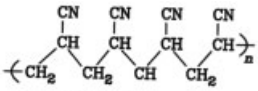

Unmodified polyacrylonitrile

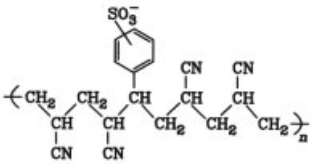

Acid-modified polyacrylonitrile

1.4.6. Pigments

A pigment is an insoluble coloring matter employed in a medium such as a paint or polymer [19]. Physical properties such as crystal form, tinctorial strength, transparency, and rheology are much more important than with water-soluble dyes (inorganic and organic pigments are treated in separate articles; → Pigments, Inorganic and → Pigments, Organic).

Toners are precipitated complexes of water-soluble dyes, e.g., calcium or barium salts of sulfonated azo dyes or phosphomolybdate salts of basic dyes. Lakes are similar to toners, but are precipitated in the presence of a substrate such as alumina to give a more dilute product. Extended (or reduced) pigments are toners or lakes diluted further with alumina or calcium carbonate.

As with textile dyes, the azo chromophore predominates, supplemented by phthalocyanine blues and greens and quinacridone reds where the highest fastness is required. Vat dyes are also used in pigment form.

1.5. Economic Aspects

Worldwide production of dyes and pigments is estimated to be ca. 750 000 t/a [20]. Very few products have a worldwide demand of more than 1000 t; 100 – 200 t/a is a more typical figure. These requirements are met by a large number of competing suppliers.

The worldwide growth since the mid 1930, shown in Table 3, reflects the decline of the Western European share from 95 % at the beginning of the century to ca. 40 %. About half the Western European production is based in the Federal Republic of Germany; Switzerland and the United Kingdom are the other principal exporters. The major Western European manufacturers have subsidiaries in the United States and elsewhere.

Figures after 1974 are available for United States [21] and Japan [22] but not for the OECD. However, United Nations statistics show a slump in 1975 followed by a leveling of world production from 1974 (Table 3) through to the 1980s [23]. The real value of production (ca. $ 5×10^9 by 1980) shows slight expansion.

Tonnage figures can be misleading since they do not take account of product strength. There is little doubt that the trend over the last ten years has been for increasing strengths and that this accounts for the apparent lack of tonnage growth.

On a worldwide basis, the consumption of dyes and pigments is ca. 200 g per person per annum; the figures for developed countries are much higher, e.g., 650 g/a for the Federal Republic of Germany. This covers everything from clothing to cars: ca. 60 % is allotted to textiles, 25 % to pigment applications, and the remainder to paper, leather, and minor applications.

A production cost of 6 $/kg might be accounted for as follows:

Raw materials	$ 1.4
Energy	$ 0.4
Manufacturing	$ 1.6
Administration, research and development	$ 1.6
Servicing fixed and working capital	$ 1.0

Table 3. World dye and pigment production, 10^3 t∗

Year	Western Europe	United States	USSR, Eastern Europe, China	Japan	Others	Total
1938	110	37	35		28	210
1958	112	80	127		27	346
1974	300	138	200	68	44	750

∗ 1 t of dye colors 41 700 men's suits; 1 t of pigment colors 16 600 VW automobiles [24]

Raw materials cost is of course largely outside the manufacturer's control. This cost increases if late-stage intermediates are used, but with a compensating reduction in manufacturing and other costs.

1.6. Environment, Ecology, and Toxicology

Environment, ecology, and toxicology are of increasing importance for manufacturers and users of dyes. Environmental considerations are of major importance in choosing new plant sites and are sometimes economically restrictive to existing operations.

Apart from the legal restrictions, which vary from country to country, on the discharge of acid, alkali, and color, there are increasingly stringent limitations in most developed countries on toxic metals (Hg, Cd, Cu, Ni, Pb, Zn, and Cr), lipophilic aromatic amines, and polychlorobiphenyl compounds (PCBs). Effluents are monitored for pH, biological oxygen demand (BOD), chemical oxygen demand (COD), and total organic carbon (TOC). Dyes are not readily biodegradable and make little contribution to BOD. Some form of adsorption or chemical treatment is often employed to minimize visible color for aesthetic reasons.

Accurate data on colorants discharged to the environment are not available, but estimates suggest a 1 – 2 % loss in production and a 1 – 10 % loss in use [25]. Factories may require some form of adsorption or sewage treatment.

Restrictions on effluents are increasingly important in process development for new products and new plants. The trend is toward process recycling and the incorporation of a detoxifying or neutralizing stage.

The Ecological and Toxicological Association of the Dyestuffs Manufacturing Industry (ETAD) was formed in 1974. The 28 member companies represent > 80 % of worldwide dye manufacture (excluding Comecon countries and China); the two main objectives of the association are as follows:

1) Coordinate and unify the efforts of manufacturers of synthetic dyes and pigments to minimize any possible damage to the environment that arises from the use and application of these products and to provide the best possible practical protection for users and consumers
2) Aid and cooperate with government departments and agencies and any other public institution concerned with the ecological and toxicological impact of these products [26]

The principal legislative acts dealing with ecological and toxicological requirements are the 6th Amendment on dangerous substances in the EEC and the Toxic Substances Control Act (TSCA) in the United States. Both laws distinguish between existing chemicals and "new chemicals" and require the preparation of an inventory and the notification of new chemicals [27].

2. Color and Structure

The correlation between structure and color has always been important in the understanding of organic compounds [28], [29].

The development of quantum theory led to quantitative treatment of these correlations but not to a direct relationship between color and constitution. Instead, the theory linked energy differences between ground and excited states with spectra absorption maxima, as shown in Section 2.3. These absorption spectra are determined in solution, and consequently such correlations, are applicable mainly to dyes, with

only limited value for pigments, which are soluble in very few solvents. The energy acquired by photon absorption is subsequently released, usually as heat, i.e., molecular vibrations. In the present context, however, the release of energy as visible-light emission is more important since it influences the color of fluorescent colorants. In Section 2.4 the two possible emission mechanisms, fluorescence and phosphorescence, are discussed which are important for commercial products (→ Optical Brighteners; → Fluorescent Dyes). In Section 2.5, two examples are given of the correlation between structure and light absorption that demonstrate the success of modern molecular orbital methods. These methods have shed light on the blue color of indigo and the influence of structure on the light absorption of the largest class of industrial dyes (→ Azo Dyes). The effect of substituents on the absorption spectra of aromatic systems is outlined in Section 2.6.

Reflection spectra in the solid state are not important industrially or in theoretical investigations of dyes. Such spectra are, however, extremely important in the application of pigments. Crystal modifications of the same chemical compound may be very different in color. Little, however, is known about correlations between chemical structure and reflection spectra of pigments (see Chap. 3; → Pigments, Inorganic and → Pigments, Organic).

2.1. Basic Principles of Color

Color vision is based on physical, chemical, physiological, and psychological processes. The part of the electromagnetic spectrum visible to humans is the range of wavelengths between 400 and 700 nm. Light can be absorbed completely, partially or not at all by gases, liquids, or solids. The unabsorbed light can be reflected at the surface of liquids or solids or can be transmitted through gases, liquids, or transparent or translucent solids. Light can reach the human retina by direct emission from a source and by reflectance and transmission. Wavelengths between 400 and 700 nm initiate a photochemical reaction followed by light-independent reactions in the light-sensitive pigments of the retina. This process results in visual perception by a transfer of information from the eye to the brain.

This Section treats simple optical concepts as the basis for the correlations between chemical structure and light absorption. Colorimetry has become important in the dyeing of textiles, plastics, and paper. Colorimetry is based on the neurobiological processes in the eye and the brain and is discussed in Chap. 3.

If visible electromagnetic rays, with energy distribution corresponding to sunlight, are subjected to complete and diffuse reflection by a solid, that solid appears *white* to the human eye. On the other hand, if the solid absorbs all the light, it appears *black*. If it absorbs a constant fraction of light in the entire range between 400 and 700 nm, it appears *grey.* White, grey, and black are called *achromatic* colors and are characterized by a constant absorption between 400 and 700 nm (Fig. 1). In contrast, solids having *chromatic colors* show one or more bands, i.e., absorption maxima and minima, in the visible spectrum. If an absorption band is located in the part of the visible spectrum with short wavelengths (400 – 430 nm), that part of the incident light is absorbed. The rest of the light is reflected and the solid appears yellow. Similarly, absorption bands at 430 – 480, 480 – 550, 550 – 600, and 600 – 700 nm yield *orange, red, violet, and blue* colors, respectively (Fig. 1). *Green solids* have absorption maxima at 400 – 450 and 580 – 700 nm. Simple inspection does not distinguish between a green color originating from one dye with two absorption maxima and a mixture of a blue and yellow dye.

The use of two or more different dyes to achieve a certain hue always leads to duller shades than those given by a single dye of the required color. Such a combination is called a *subtractive* mixture. In the ultimate case, black is obtained by "subtraction" of all parts of the visible spectrum. The color of a solution can be explained similarly in terms of light absorbed and transmitted (rather than reflected) by the constituents.

In contrast to these subtractive processes, when light from different sources combines so as to match the relative intensities of sunlight, an impression of colorless (white) results. Such a combination is called *additive.*

For the perception of color by the eye, both the position and the shape of the absorption band are important. Narrower and sharper bands give purer and more brilliant hues. The spec-

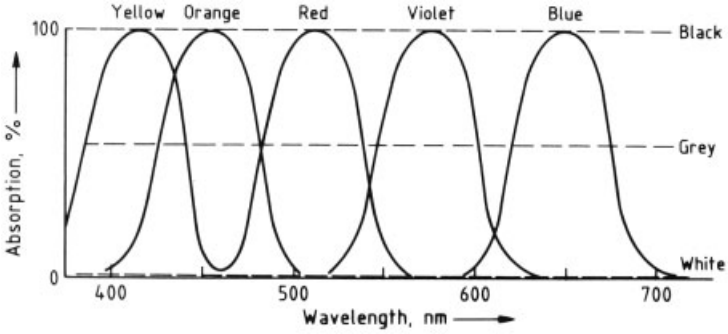

Figure 1. Schematic representation of the light absorption of colored solids
– – – – achromatic colors
—— chromatic colors

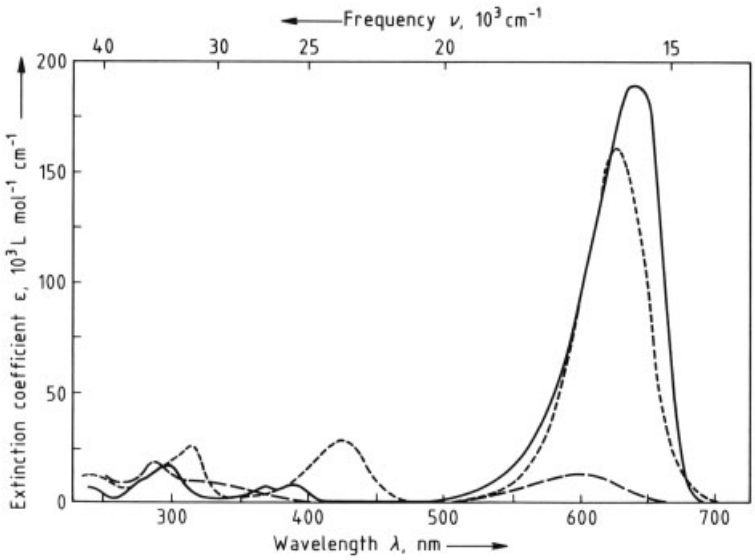

Figure 2. Absorption spectra of Indigo in $CHCl_3$, (– – –), Michlers Hydrol Blue in water (– – – – –), and Malachite Green in water (———). From [2], Figs. X.7 and XI.5.

tra of two blue dyes and one green dye are shown in Figure 2. Indigo (C. I. Vat Blue 1) [*482-89-3*], is the most important of all dyes from a historical standpoint (Chap. 1). It is applied mainly to cotton denim (blue jeans), producing a rather dull, slightly reddish blue (navy blue). The dullness is due to the broad band at 580 – 600 nm (half-width ca. 100 nm) and a very weak second band in the 400 – 520 nm region (not visible in Fig. 2). The spectrum of Michler's Hydrol Blue [*14844-71-4*] is characterized by a single narrow band in the visible spectrum (λ_{max} = 647 nm; half-width ca. 50 nm). Therefore, its color is brilliant. The third spectrum, that of Malachite Green (C. I. Basic Green 4) [*569-64-2*], is characterized by two bands in the visible spectrum. In contrast to indigo, they are narrow (λ_{max} = 621 and 427 nm; half-widths 60 and 40 nm, respectively). This combination is seen as green. The absorption of indigo is weaker than that of the two other dyes (low extinction coefficient ε).

Contrary to general belief, in many cases it is not possible to predict the colors of solutions or solids from their spectra. The eye is not equally sensitive to all parts of the spectrum; as a result, color perception is governed not only by the position and the area, but also by the *shape* of the absorption band. Furthermore, the sensitivity of the eye is not the same for day or photopic vision (using cones in the retina as photoreceptors) and for night or scotopic vision (using the more sensitive retina rods). Day vision is most sensitive at 555 nm.

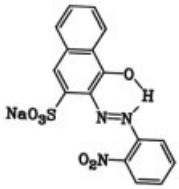

1

The perceived color of the azo dye **1** is illustrative. The absorption spectrum in weakly acidic solution shows a sharp maximum at 500 nm ($\varepsilon = 25\,000$ L mol^{-1} cm^{-1}). In alkaline solution, dissociation of the hydroxyl group gives the naphtholate ion with a broad absorption maximum at 478 nm ($\varepsilon = 14\,000$ L mol^{-1} cm^{-1}); this band extends to 700 nm (Fig. 3).

These absorption maxima would suggest an orange-red color in acid solution and a yellowish-orange hue in alkaline solution. The observed color change, however, is from orange to dull wine-red.

The utility of a colored compound depends not only on the position of the absorption band, but also on its intensity, which is given by the Lambert – Beer law:

$$\log \frac{I_0}{I} = E = \varepsilon d c \qquad (1)$$

where

I = intensity
E = extinction
d = cell length, cm
c = concentration, mol/L
ε = extinction coefficient, L mol^{-1} cm^{-1}

The ε-values of industrial dyes range from less than 10^4 to more than 10^5 L mol^{-1} cm^{-1} (Fig. 2). Dyes with high molar extinction coefficients (e.g., triarylmethine dyes $\varepsilon \geq 100\,000$) are economically efficient. Conversely, anthraquinone dyes are expensive to use because manufacturing costs are high and extinction coefficients are low. The ε-values of anionic and disperse anthraquinone dyes (→ Anthraquinone Dyes and Intermediates) range from 8000 to 18 000. Vat dyes have still lower extinction coefficients; Indanthrone [*81-77-6*] has $\varepsilon = 2600$ in dimethylformamide for $\lambda_{max} = 610$ nm [30]. As shown in Figure 2, indigo also has a low extinction coefficient ($\varepsilon = 12\,000$ in chloroform).

In industrial research, many colored compounds are prepared by varying the substitution on the same basic structure. The effects of such changes are seen schematically in Figure 4. Shifts of the absorption maximum to longer and shorter wavelengths are called *bathochromic* and *hypsochromic* shifts, respectively. Increases and decreases in extinction coefficients (absorption) are termed *hyperchromic* and *hypochromic* shifts, respectively.

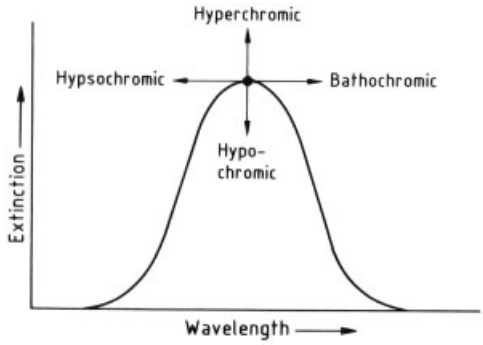

Figure 4. Designations of wavelength shifts and changes of extinction of absorption bands

2.2. Empirical Correlations

Correlations between chemical structure and color began to be studied in the early days of dye chemistry. Graebe and Liebermann recognized that all dyes contain conjugated double bonds [31]. The first comprehensive theory was developed by Witt [32] and has been the basis of most subsequent investigations. Witt postulated that color is due to the presence of groups called *chromophores,* which must be linked to the system of conjugated double bonds. Potential chromophores create color if

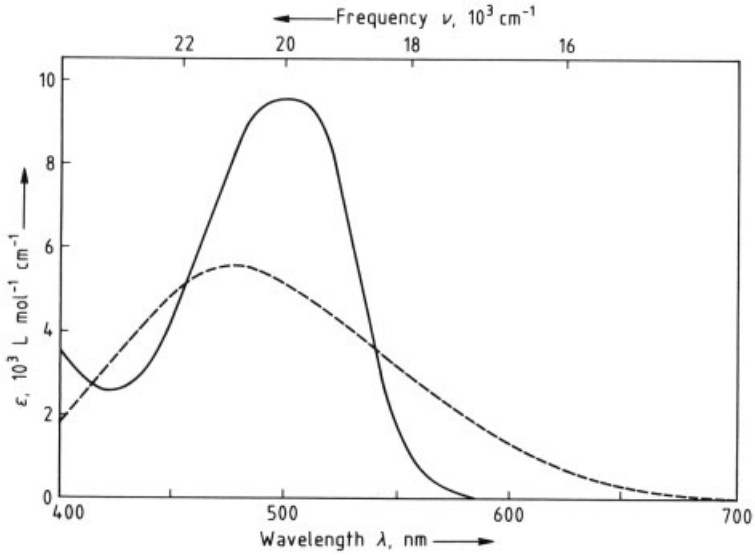

Figure 3. Visible spectra of dye **1** in acidic (——) and basic solution (– – –).

auxochromic groups are present. Some chromophores have been classed as antiauxochromes [33–35]. With the later development of electron theory, it became apparent that auxochromes are electron donors ($-NH_2$, $-OH$, $-O^-$, etc.), antiauxochromes are electron acceptors, and chromophores are linear or cyclic systems of conjugated double bonds. An assembly of these entities is sometimes called a *chromogen.*

König studied the influence of the length of the conjugated chains in polymethine dyes (**2**, n = 0, 1, 2, . . .) on the wavelength of the most bathochromic absorption band [36].

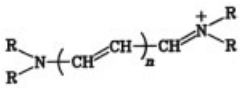

2

Dyes of type **2** were seen to support the so-called oscillation hypothesis of color that was proposed by von Baeyer [37]; this was based on the Kekulé benzene structure, which has been generalized by Robinson et al. [38] and Ingold and Ingold [39]. Empirical chemical theories of color have been reviewed [28], [29], [40–42].

The development of quantum theory in the 1920s provided a new understanding of chemical bonds and electronic spectra [43], [44].

2.3. Quantum Chemical Methods

As mentioned earlier, color is created by the absorption of electromagnetic radiation of wavelength between 400 and 700 nm. The energy of the light absorbed by the molecule promotes an electron to a higher energy level. This energy difference corresponds to 170 – 300 kJ/mol for visible light and 300 – 12 000 kJ/mol for ultraviolet light; the same absorption mechanism produces electronic spectra in both cases. Light of longer wavelength (e.g., infrared, of 7 – 500 μm) is, on the other hand, not sufficiently energetic to excite electrons, but only causes molecular vibrations.

A real understanding of electronic spectra requires knowledge of the energy of the ground states and excited electronic states of the absorbing molecules and ions. According to the Einstein – Bohr frequency condition (Eq. 2), the energy differences ΔE between these states are proportional to the frequency ν or wavenumber and inversely proportional to the wavelength λ of the absorbed light.

$$\Delta E = h\nu = \frac{hc}{\lambda} \tag{2}$$

With h being Planck's constant and c the speed of light.

In the three-dimensional wave mechanical model of electron distribution the electron orbits

in the Bohr model are replaced by orbitals, i.e., distribution functions of single electrons. The Schrödinger equation (Eq. 3) describes the motion of an electron of mass m and potential energy U as a function of the Cartesian coordinates x, y, and z, the total energy E, and the eigenfunction ψ; ψ^2 is a measurable quantity related to the electron density.

$$\frac{\delta^2\psi}{\delta x^2}+\frac{\delta^2\psi}{\delta y^2}+\frac{\delta^2\psi}{\delta z^2}+\frac{8\pi^2 m}{h^2}\left[E-U\left(x,y,z\right)\right]\psi=0 \quad (3)$$

A closed solution of the Schrödinger equation is possible only for extremely simple systems. Multielectron systems present in diatomic molecules are reduced to one-electron systems by approximation methods.

The *valence bond (VB) method* was first applied to the hydrogen molecule. The bonding effect of its electron pair is ascribed to an exchange interaction of one unpaired electron in an atomic orbital of one of the hydrogen atoms with the unpaired electron of the other atom. Thus, the two electrons are indistinguishable and can exchange positions. The wave mechanical computation shows that this process corresponds to a stabilization of H_2 relative to 2 H Extreme distributions of electrons in VB representations are called limiting structures (for the related resonance theory, see Section 2.5).

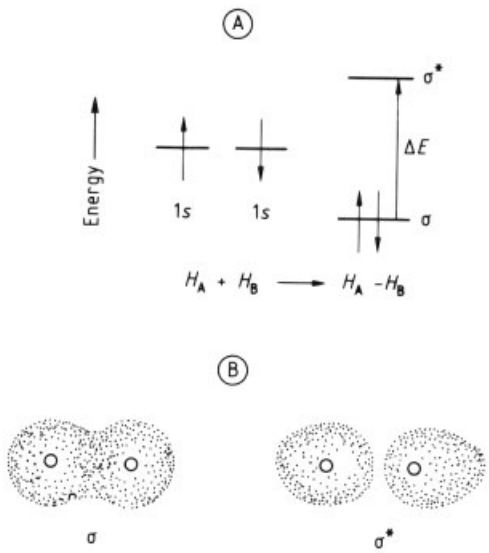

Figure 5. Formation of H_2 by the molecular orbital (MO) approximation; energy diagram (A) and form of the σ and $\sigma*$ MOs of the hydrogen molecule (electron densities, B)

In the *molecular orbital* (*MO*) *method*, the molecular wave functions (MOs) are formed by a linear combination of the atomic wave functions (AOs) of the atoms in the molecule (LCAO-MO, linear combination of atomic orbitals). Each molecular orbital formed can accommodate two electrons with opposite spin (Pauli principle). In the case of H_2, the eigenfunctions (wave functions) of the two atomic orbitals are added with the same sign to produce the bonding (or occupied) MO and with opposite sign to obtain the nonbonding (unoccupied) MO (see Fig. 5). In the ground state, both electrons are in the bonding MO. Electromagnetic energy (light) is absorbed by this molecule if the energy of the light is equal to the energy difference ΔE between the occupied and the unoccupied molecular orbitals. In the case of H_2, the two 1s molecular orbitals that are involved are denoted by σ and $\sigma*$; the H – H bond of the H_2 molecule in the ground state is a σ bond.

The energy levels of the bonding orbitals decrease as the atomic orbital overlap increases. This overlap is, therefore, the cause of bond formation in the MO model.

The absorption of light by H_2 occurs in the far ultraviolet ($\lambda = 160$ nm). The high absorption energy for this $\sigma \rightarrow \sigma*$ transition is also in agreement with the fact that in more complex molecules excitation of σ electrons requires high-energy photons.

In this excited state, since the electrons no longer occupy the same molecular orbital, the restriction on their spins that is imposed by the Pauli principle no longer applies; there are, therefore, two excited states, the singlet and the triplet state, with anti-parallel (↓↑) or parallel (↑↑) spins of the two electrons, respectively. The triplet state, which is important with respect to phosphorescence and photochemical reactivity, has a lower energy than the related singlet state. In compounds that contain elements of the second and higher rows of the periodic table, p atomic orbitals must be considered in addition to s orbitals for σ bond formation. These s and p orbitals can be combined to form hybrid orbitals; the method of combination determines the geometry of the molecule. The most characteristic types of s-p hybridization are sp^3, i.e., the combination of one s and three p orbitals (tetrahedral, specific for aliphatic hydrocarbons), sp^2 (trigonal, typical for olefins and aromatic com-

pounds), and *sp* (digonal, acetylenes). In sp^2 hybridization, only two *p* electrons (p_x and p_y) are used together with the *s* electron. The third *p* orbital (p_z) does not take part in hybridization. It is used to form the π orbital. Similarly, with *sp* hybrids, the p_y and the p_z orbitals are used for the formation of two π orbitals.

In ethylene there are only two π electrons; their combination in an MO is formally analogous to the treatment of the two *s* electrons in H_2 discussed previously (Fig. 5 A). A division is obtained between an MO occupied (in the ground state) by both electrons (π MO) and an unoccupied higher energy MO ($\pi *$).

Electrons in these π MOs are less strongly bonded and can be more easily excited into $\pi*$ orbitals by absorption of light of longer wavelength, i.e., lower energy. This explains the rule of Graebe and Liebermann (Section 2.2) that unsaturated compounds with double bonds absorb light at longer wavelengths than saturated compounds, especially if the double bonds are conjugated [31].

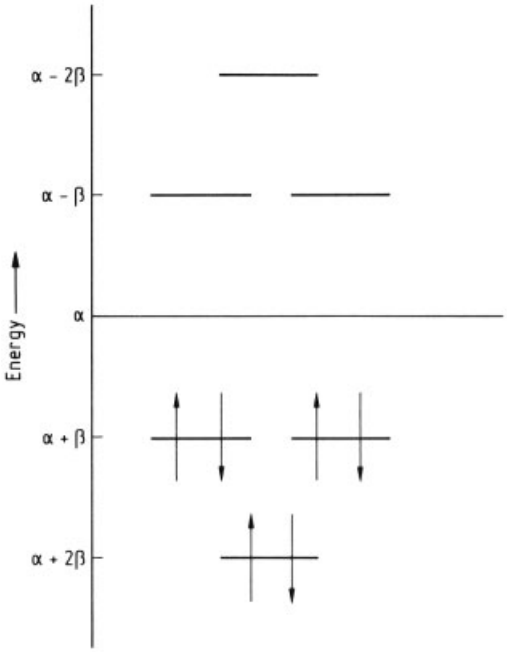

Figure 6. The π orbitals of benzene

The properties of *benzene* do not correspond to those of the hypothetical cyclohexatriene. The most bathochromic absorption band of benzene is at 250 nm. The electron distribution follows from molecular orbital considerations. As with ethylene, π orbitals overlap, but with the hexagonal symmetry not only do three pairs of adjacent π orbitals overlap, but also all six π orbitals, resulting in delocalization of the six π electrons over all six C–C bonds. The MO calculations by Hückel in 1931 show that the second and third bonding orbitals possess the same energy (*degenerate MOs*). The two lower antibonding orbitals are also degenerate. The energy levels of these orbitals are shown in Figure 6, as calculated by the simplest MO approximation method (Hückel MO or HMO) [45].

Here the so-called Coulomb integral α serves as a reference energy level for the sp^2 hybridized carbon atom, corresponding approximately to the ionization potential of an electron in a $2p$ atomic orbital. The resonance integral β is a measure of the interaction between two orbitals; it is approximately proportional to the overlap integral *S* and with aromatic hydrocarbons amounts to 74–83 kJ/mol. The energy 2β is termed the mesomeric or delocalization energy of benzene, and is responsible for its stability.

Hückel demonstrated that this particularly strong stabilization occurs only in homo- and heterocyclic compounds that have $(4n+2)\pi$ electrons in the perimetric ring system (Hückel rule). As the investigation of annulene systems showed, it is not fundamentally important whether they are mono- or polycyclic systems when $n = 1, 2, 3, \ldots$ (etc.). The cycloheptatrienylium cation, which Hückel predicted to be stable (tropylium ion, C_7H_7), was first prepared by Doering in 1956, confirming the usefulness of quantum theory in organic chemistry.

Heterocyclic compounds also possess benzenoid characteristics if they have $(4n + 2)$ overlapping π orbitals. This is true for pyridine and pyrimidine. Thiophene and pyrrole are also aromatic because the lone pairs at S and N, respectively, are in p_z orbitals and overlap with the π electrons of the carbon–carbon double bonds.

When an aromatic compound carries a group with π orbitals e.g., NO_2 or COOH on an aromatic carbon, both π electron systems overlap, leading to new bonding and antibonding electronic states. This also occurs with atoms that have lone pairs of electrons (e.g., OH or NR_2) in *p* orbitals which, for geometrical reasons, can also overlap with the aromatic π orbitals. This,

however, is not the case with the lone pair on nitrogen in pyridine. Both electrons occupy an $sp_x{}^2$ hybrid atomic orbital located in the plane of the six-membered ring perpendicular to the *p* orbitals of the aromatic π system. Therefore, interaction between this lone pair and the π system is minimal.

A molecular electronic spectrum exhibits a number of absorption bands of widely differing intensities which are partly superimposed. In principle, *one* particular electron transition corresponds to each band. However, the origins of these bands are often not clear since simultaneous changes in the vibrational states are superimposed on the electron transitions.

The essential types of energy levels are shown in Figure 7 in the simplest form for two types of compounds: a compound A with a pair each of *p,* π and *n* electrons, and a compound B with one pair of *p* and *n* each but with two pairs of conjugated π electrons.

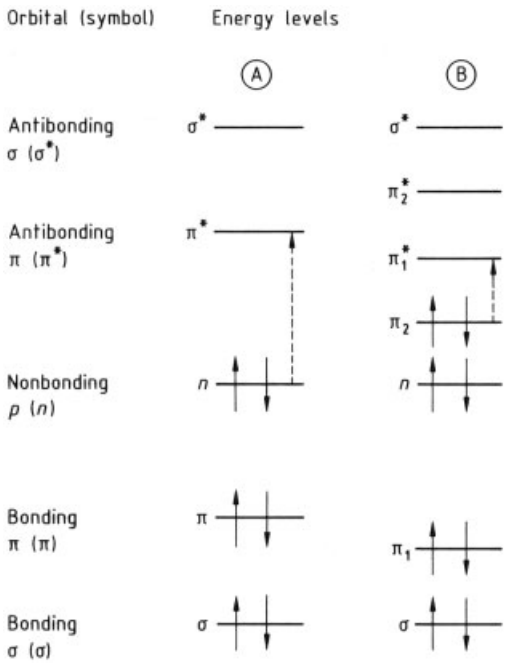

Figure 7. Designation of molecular orbitals and representation of energy levelsA: for a molecule with one *p,* one π, and one *n* electron pair
B: for a molecule with one *p* and two conjugated π electron pairs (number of *s* electrons not specified)

In both cases only one occupied σ and one unoccupied $\sigma*$ level are shown although all real compounds have more *s* electrons. These are, however, not important for the present discussion.

For compound A the transition of an *n* electron from the nonbonding orbital (*n*) to the antibonding $\pi*$ orbital requires the least energy. The $n \rightarrow \pi*$ transition, therefore, gives rise to an absorption band at a longer wavelength than the $\pi \rightarrow \pi*$ transition. In the case of compound B, the π_1 and π_2 levels are split (nondegenerate). Therefore, for the given example the band at the longest wavelength corresponds to a $\pi \rightarrow \pi*$ transition. Not all transitions, however, go to antibonding levels: in arylmethine dyes transitions to nonbonding excited states are important.

As shown above for benzene, the HMO method takes into account only the π orbitals; σ orbitals, their interaction with π orbitals, and interactions between two or more π electrons are neglected. Therefore, the HMO method can be applied to simple aromatic compounds, e.g., disubstituted benzenes, only in a very restricted way by a suitable choice of empirical parameters. In the so-called MIM (molecules-in-molecules) version of the HMO method, the only changes calculated in relatively complex molecules (which relate to the known spectra of subsystems) are those that occur when these subsystems are coupled to the whole molecule.

Until 1937, quantum mechanical calculations sought to determine the energies of ground states of molecules, particularly for compounds such as benzene. However, in that year, recognition of the importance of the excited states of a molecule for light absorption allowed the calculation of the absorption bands of a number of unsaturated hydrocarbons by early MO methods [46]. Shortly afterward PAULING applied the VB method [47]. Whereas the chemical theories discussed in Section 2.2. were concerned mainly with dyes of industrial importance, quantum mechanical methods started with very simple systems.

The free electron model (FE-MO theory) was particularly important for interpreting the electronic spectra of complex dyes [48–53]. Simple calculations based on this model show good agreement with experimental spectra, particularly for polymethine dyes [54–57].

With this model the wave functions of the π electrons are treated as waves in a box corresponding to the dye molecule or ion; in the simplest case, the box is one-dimensional and

as long as, for example, an all-*trans* polymethine chain. Stationary states are then characterized by standing waves; the length of the box (L) in which the electron moves is equivalent to an integral multiple ($n = 1, 2, 3, \ldots$) of the half-wavelength λ ($L = n\lambda/2$). The energy levels ε_n corresponding to these waves are occupied from $n = 1$ upward with 2 π electrons each and correspond to molecular orbitals. The energy difference ΔE between the highest occupied orbital E_{HOMO} (HOMO = highest occupied MO [58], [59]) and the next orbital above, i.e., the lowest unoccupied orbital E_{LUMO}, yields the $\pi \rightarrow \pi *$ transition energy with the aid of the physical constants h and m (mass of the electron) and the length (L) of the system of conjugated double bonds of the polymethine dye (Eq. 4). With the Einstein – Bohr frequency condition (Eq. 2), ΔE yields the frequency of the absorption band; the acronyms HOMO and LUMO were proposed by Fukui in the frontier orbital method [31], [32].

$$\Delta E = E_{\mathrm{HOMO}} - E_{\mathrm{LUMO}} = \frac{h^2}{8mL^2}\left(n^2_{\mathrm{HOMO}} - n^2_{\mathrm{LUMO}}\right) \tag{4}$$

In Fig. 8 A, the energy levels $\varepsilon_n = 1$, 2, 3, 4, and 5 of Equation (4) are given with the respective waves for the trimethine cation $R^2NCH{=}CHCH{=}N^+R_2$. They are compared with the energy levels of the molecular orbitals produced by the linear combination of atomic orbitals in the HMO method (Fig. 8 B). Figure 8 C shows the corresponding view as seen from above. The size of the circles is proportional to the LCAO-MO coefficients of the atomic orbitals. The HMO and FE methods are compared in Figs. 8 B and 8 C. For more complex π electron systems, the simple FE-MO model does not correspond well to experimental spectra; such systems can be handled by two- and three-dimensional FE methods. The FE-MO model is clearest for didactic purposes.

The energy of each MO is determined not simply by a single-electron configuration, as assumed in HMO and FE-MO methods, but by the interaction of several such configurations. Electron – electron interaction and configuration interaction (CI) are taken into account for π-electrons in the Pariser – Parr – Pople (PPP) method [60], [61]. This method has given successful calculations of the spectra of colorants including those with secondary absorption bands at shorter wavelengths.

The $\pi \rightarrow \pi *$ transitions are also influenced by σ electron distributions and may mix with certain $\sigma \rightarrow \sigma *$ transitions. This is taken into account in the semiempirical all-valence electron methods, the so-called CNDO, MINDO, INDO, and related approximations (CNDO, MINDO, INDO = complete, modified intermediate, and intermediate neglect of differential overlap, respectively). They are useful for small π electron systems, but the improvement over PPP for larger colorants is not significant.

The PPP method reproduces the main features of various classes of chromophores, and the calculated absorption wavenumbers differ by not more than 3000 cm^{-1} from the observed absorption maximum ([29], p. 7). This is ca. 20 nm in the 250-nm region, but more than 100 nm in the visible 600-nm region. Closer agreement can be obtained within narrow structural changes by optimizing the parameters used for the calculations.

The PPP method has been applied by Griffiths to practically all classes of dyes [28], [40, 41], [62, 63].

In principle, the intensity of absorption is also obtainable by MO methods. The extinction coefficient is a function of the shape of the orbitals and of the degree of overlap, i.e., of the so-called transition dipole moment M, which has the character of a vector. In general, this has different magnitudes along the three Cartesian coordinates (x, y, and z in Eq. 3). This explains the dichroism of dye crystals in polarized light (see Section 2.6) or of dyes dissolved in nematic phases of liquid crystals or stretched polymer sheets [64]. The extinction coefficients depend on the plane of polarization.

The transition moment contains the product of the two overlapping orbitals, the so-called overlap densities $\Phi_m \Phi_n$. This product is a measure of the localization of the two orbitals in the same part of the molecule and, therefore, describes the probability of an electronic transition from one orbital to another. The molar extinction coefficient of the absorption maximum ($\varepsilon_{\max}$) is related to the so-called oscillator strength (f) via the square of the transition moment (M^2) and the wavenumber ($\tilde{\nu}$) of that transition (Eq. 5); $\int e_{\max} \mathrm{d}\tilde{\nu}$ corresponds to the area of the band, and $\Delta\tilde{\nu}_{1/2}$ is the width of the band at half-height:

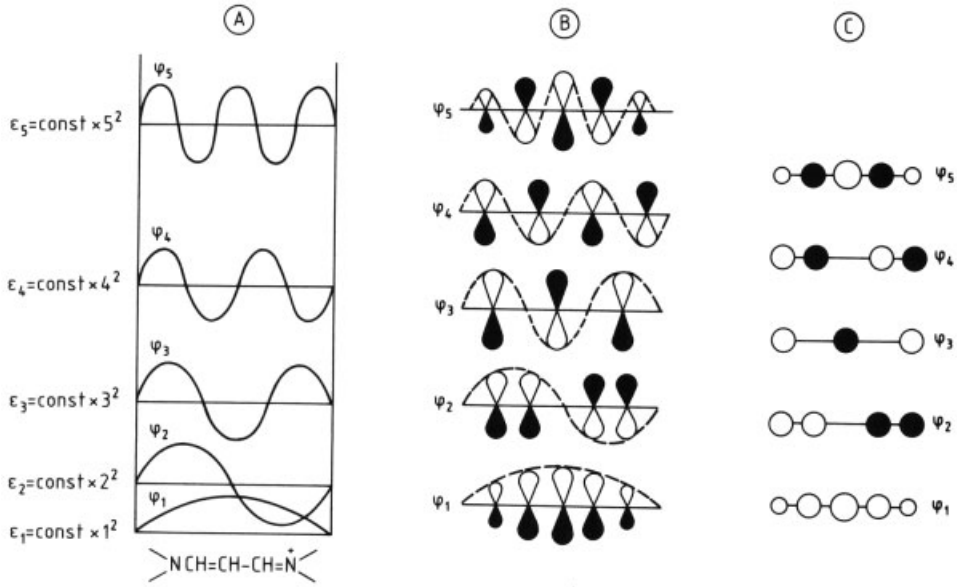

Figure 8. The π MOs of the trimethine cation $R_2NCH{=}CHCH{=}N^+R_2$ in the free electron (FE) model (A) and in the HMO model (B) and (C)
Reproduced with permission of M. Klessinger [56]

$$f = 4.703 \times 10^{29} \tilde{\nu} M^2 = 4.32 \times 10^9 \int \varepsilon_{\max} \mathrm{d}\tilde{\nu}$$
$$= 4.32 \times 10^9 \varepsilon_{\max} \Delta\tilde{\nu}_{1/2} \quad (5)$$

In this case the band is assumed to be symmetrical (Gaussian curve). However, this is rarely the case with electronic absorption bands. Unsymmetrical absorption bands result from the superposition of vibrational processes upon electronic transitions in the ground and excited states, as well as from interactions with neighboring molecules (solvatochromism).

2.4. Fluorescence and Phosphorescence

An excited molecule or ion can release energy in several ways:

1) Radiationless transition as internal conversion or intersystem crossing (heat formation)
2) Emission of radiation (fluorescence and phosphorescence)
3) Photochemical reactions

Processes of types 1 and 2 are represented in the Jablonski diagram (Fig. 9), a scheme of the essential energy levels. The lowest vibrational energy levels of the electronic states are indicated by thick horizontal lines; other horizontal lines represent higher vibrational levels. A molecule (or ion) in the ground state S_0 is brought to an excited singlet electronic state, S_1 or S_2 (or higher states S_j) by absorption of light. Radiationless transitions are shown by wavy lines. Vibrational deactivation (indicated by vertical wavy lines) leads to the lowest vibrational level of the respective excited singlet electronic state or to lower singlet states. Intersystem crossing (indicated by horizontal wavy lines) leads to triplet states. Radiation from the lowest vibrational level of the excited state S_1 to any of the vibrational levels of the ground state is called *fluorescence.* Radiation from the lowest triplet state T_1 to the ground state is called *phosphorescence.*

Photochemical reactions (e.g., rearrangement, dissociation, dimerization, photoaddition, or reaction with neighboring particles) arise most often from triplet states of lifetimes between 100 ns and 10 s. The lifetime of excited singlet states is normally too short for chemical reactions (1 – 100 ns).

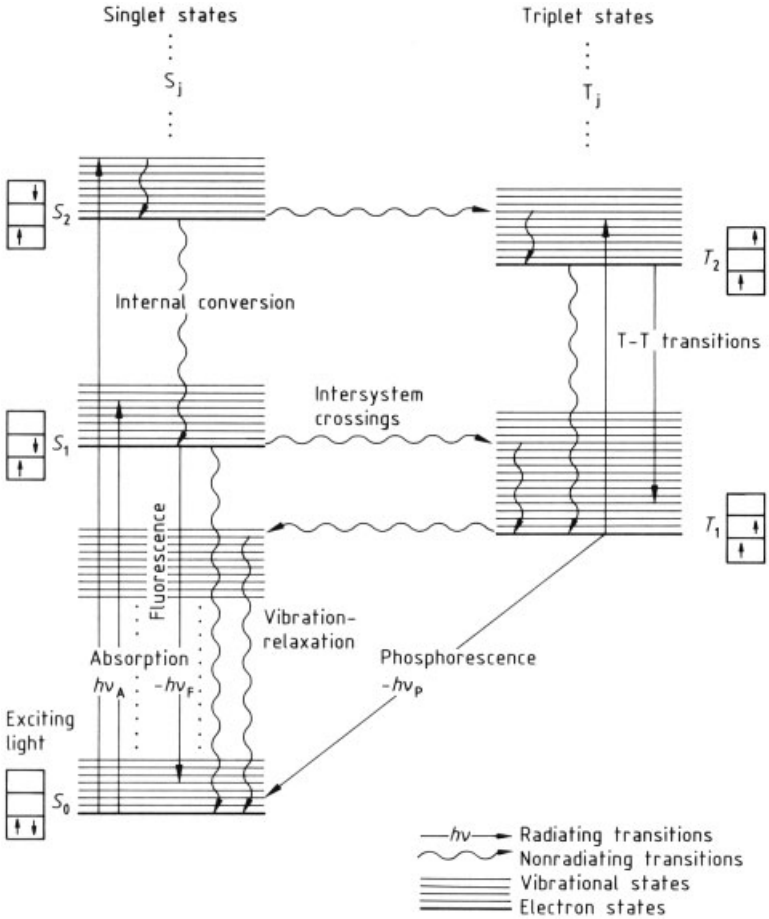

Figure 9. Jablonski scheme
Reproduced from Ciba-Rundschau 1973, no. 1, with permission

With the exception of phosphorescence, all radiation processes are very rapid, with first-order rate constants of the following orders of magnitude:

Light absorption	$10^{15}\ s^{-1}$
Vibrational relaxation	$\geq 10^{12}\ s^{-1}$
Fluorescence	$10^{6} - 10^{9}\ s^{-1}$
Intersystem crossing	$10^{-1} - 10^{5}\ s^{-1}$
Phosphorescence	$10^{-2} - 10^{4}\ s^{-1}$

As seen from Fig. 9, absorption normally occurs from the lowest vibrational level of the electronic ground state and leads to one of the higher vibrational levels of an electronic excited state. Fluorescence arises in most cases from the lowest vibrational level of the first excited state (S_1) and ends in one of the higher vibrational levels of the ground state. Therefore, the fluorescence spectrum, in ideal cases, is a mirror image of the absorption spectrum, but shifted to higher wavelengths (Stokes shift).

For phosphorescence to occur, molecules primarily excited to the S_1 singlet state must revert to the lowest level of their triplet manifold (T_1) by radiationless intersystem crossing. From T_1 the molecules can slowly emit radiation (phosphorescence) or they can return to the S_0 level without emission by intersystem crossing. Phosphorescence spectra are also usually shifted to longer wavelengths much more than fluores-

cence, because the energy of the triplet level T_1 is usually below that of the singlet S_1.

Most fluorescent organic compounds are characterized by a rigidity of structure which prevents the energy of the excited states from being lost by torsional vibration. The classic example is fluorescein [*2321-07-5*] (C.I. Acid Yellow 73, **4**), which is highly fluorescent. In contrast, phenolphthalein [*77-09-8*] (**3**) shows no fluorescence because its excited state loses energy by internal conversion (vibrations of the benzene rings), a process that is prevented in fluorescein by the ether link between two of the aromatic rings.

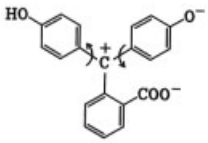

3

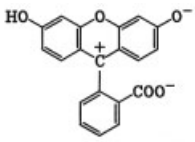

4

An interesting example of wavelength shifts is shown by the three vinyl polymethine dyes (**5**) ($m = 1$, 2, and 3). They show a bathochromic shift of the absorption (abs.) maximum for $m = 1$, 2, and 3 from 500 to 526 and 544 nm, respectively. The shifts in the fluorescence (fl.) maxima, however, are much larger. The difference λ_{max} (fl.) $-$ λ_{max} (abs.) for $m = 1$, 2, and 3 is 115, 175, and 245 nm, respectively [65]. This effect is probably due to larger energy differences between the excited singlet and triplet states in that sequence.

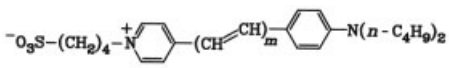

5

The fluorescence efficiency of a compound is related to the quantum yield, i.e., the fraction of the absorbed light energy that is emitted in fluorescence.

Although dyes and pigments that fluoresce in the ultraviolet and in the infrared ranges are known, most fluorescent colorants absorb and fluoresce in the visible spectrum (→ Fluorescent Dyes). Dyeings with such dyes appear brighter. They are used for fibers, plastics, as active media for dye lasers (→ Laser Dyes), and for biological staining (→ Stains, Microscopic). Optical brighteners are compounds whose most bathochromic absorption band is in the near-ultraviolet range and whose fluorescence is visible. They are used for increasing the whiteness of textiles, plastics and soaps (→ Optical Brighteners).

2.5. Quantitative Treatment of Light Absorption

For azobenzene and its derivatives, the assignment of the absorption band of longest wavelength (most bathochromic) and all other visible and near-ultraviolet bands (down to 210 nm) shows the importance of the various electronic transitions in simple compounds. *trans*-[*17082-12-1*] and *cis*-[*1080-16-6*] azobenzene (**6** and **7**) have in addition to σ electrons two times six π electrons in the two aromatic rings, two π electrons in the azo group, and two electron lone pairs on each nitrogen atom. Various $n \rightarrow \pi*$ and $\pi \rightarrow \pi*$ transitions can be expected.

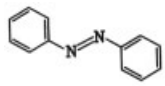

6

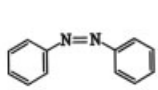

7

The absorption spectra of *cis*- and *trans*-azobenzene have been successfully interpreted by MO methods [29, pp. 46 ff]. Figure 10 shows the molecular orbital energy diagram of azobenzene with the three highest occupied and the three lowest unoccupied π orbitals and the nonbonding orbitals of the nitrogen lone pairs [66]. The last are degenerate in two orbitals (n_s and n_a), which are present in 4-aminoazobenzene derivatives.

The weak band at the longest wavelength in the visible spectrum of *trans*-azobenzene ($\lambda = 450$ nm, $\varepsilon = 463$; see Fig. 11) and *cis*-azobenzene ($\lambda = 430$ nm, $\varepsilon \approx 1500$) is due to the $n_s \rightarrow \pi_1*$ transition. The strong second bands (*trans* 330 nm, $\varepsilon = 17\,000$; *cis* 280 nm, $\varepsilon \approx 5100$, shoulder) are due to $\pi_1 \rightarrow \pi_1*$ transitions. Since *cis*-azobenzene is nonplanar, n and π-orbitals mix. This is the reason for the higher intensity of the $n_s \rightarrow \pi_1*$ band and for the weaker $\pi_1 \rightarrow \pi_1*$ absorption. The highest energy processes in the region 230 – 240 nm are due to the $\Phi \rightarrow \Phi*$ transitions; i.e., they are regarded as being localized

in the benzene rings. This is probably the reason that this part of the benzylideneaniline spectrum is similar to azobenzene and stilbene (Fig. 11). Of course, stilbene has no $n \rightarrow \pi *$ transition.

In spite of its orange color, azobenzene is not used as a dye because its molar extinction coefficient in the visible region is too small for practical purposes ($\varepsilon = 450$). If an electron-releasing group with lone pairs (e.g., dimethylamino) is introduced in the 4-position, a significant hyperchromic change results. The extinction coefficient of the longest wavelength band ($\lambda_{max} = 408$ nm) is then 27 500 (in 95 % ethanol) as in Dimethyl Yellow, C.I. Solvent Yellow 2 [*60-11-7*], a dye in the industrial sense. An additional electron acceptor group, e.g., NO_2, in the 4′-position (compound **8**) produces a definite bathochromic shift ($\lambda = 478$ nm, $\varepsilon = 33\,100$ in 95 % ethanol). However, these intense visible bands correspond not to the $n_s \rightarrow \pi_1*$ transition, but to the $\pi_1 \rightarrow \pi_1*$ transition of azobenzene. This is shifted from the ultraviolet to the visible spectrum by the influence of the substituents. Negligible $n_s \rightarrow \pi_1*$ and, presumably, weaker transitions are superimposed.

Azobenzene, 4-dimethylaminoazobenzene, and 4-dimethylamino-4′-nitroazobenzene [*2491-74-9*] (**8**) illustrate the rules of Witt [5] and Dilthey and Wizinger [6] (see Section 2.2): The absorption shifts are created by adding an auxochrome (NR_2) and an antiauxochrome (NO_2) to a chromophore.

In resonance theory, which is a qualitative version of the valence bond theory (VB, see Section 2.3), it is assumed that increasing numbers of resonance structures correspond to easier excitability of electrons, i.e., longwave absorption. Thus, it is possible to write an increasing number of polar resonance structures for azobenzene, Dimethyl Yellow, and 4-dimethylamino-4′-nitroazobenzene. Apart from the fact that the choice of resonance structures is largely arbitrary, quantum chemical treatment shows that the assumption of equal distribution among resonance structures in the ground and excited states is, in many cases, incorrect. 4-Dimethylamino-4′-nitroazobenzene (**8**) is a typical example, which can be represented as a hybrid of the three structures **8a**, **8b**, and **8c**, as well as many other structures.

The PPP method gives in the ground state a partially positive charge (+0.327) for the amino group and partially negative charges for the azo and nitro groups (−0.206 and −0.131, respectively). This corresponds qualitatively to the resonance structures, but indicates nothing about charge distributions in the excited states. According to PPP calculations the charge densities in the first excited state at the three sites mentioned are drastically different from those in the ground state, e.g. the charge density at the nitro group is not negative, but positive [63]. The electronic excitation could, therefore, be regarded as being associated with the transfer of charge from one site of the molecule to another. Figure 11 shows the absorption spectra of *trans*-azobenzene (**6**) and two iso-π-electronic compounds, *trans*-stilbene [*103-30-0*] (**9**) and *trans*-benzylideneaniline [*538-51-2*] (**10**).

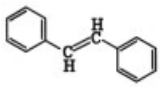

9

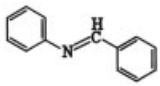

10

The similarity of the azobenzene and stilbene spectra (apart from the weak visible band of azobenzene) indicates similar electron transitions. Surprisingly, however, the spectrum of benzylideneaniline, which may be regarded as a

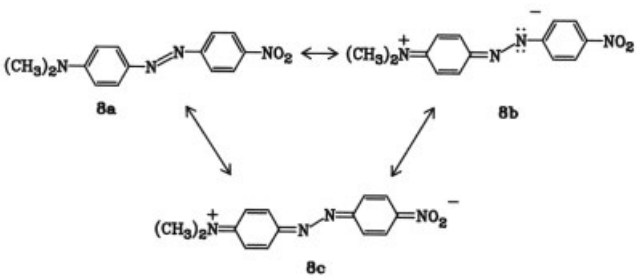

8a **8b** **8c**

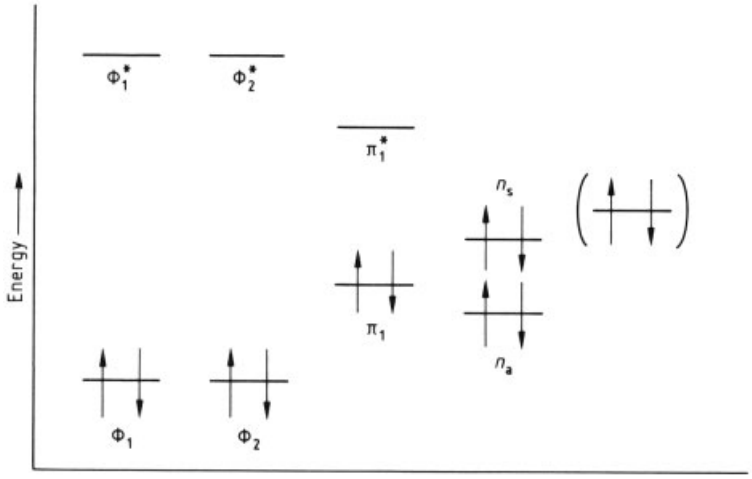

Figure 10. The molecular orbital diagram for azobenzene, including the three highest occupied, three lowest unoccupied, and the nonbonding orbitals of the nitrogen lone pairs

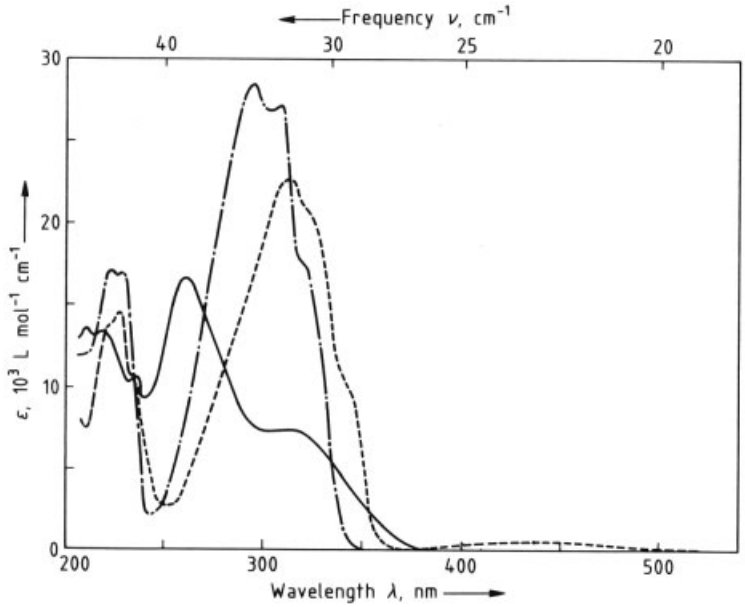

Figure 11. Spectra of *trans*-azobenzene (– – --, **6**), *trans*-stilbene (—- - —-, **9**), and *trans*-benzylideneaniline (———--, **10**) (in cyclohexane) [67]

hybrid of the two other compounds, differs considerably from those of azobenzene and stilbene. This is due to twisting around the C – N single bond by 55 °, thus dividing the conjugated system into two weakly interacting π systems. Ring closure of the methine group in **8** to the aniline ring forms the indolenine derivative **11**, which is also a planar molecule with a spectrum very similar to that of *trans*-azobenzene and *trans*-stilbene; the most intensive absorption bands in cyclohexane are $\lambda_{max} = 310$ nm, $\varepsilon = 17\,000$, and $\lambda_{max} = 235$ nm, $\varepsilon = 16\,000$ [67].

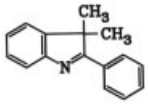

A knowledge of changes in charge densities between the ground and excited states helps to explain *solvatochromism,* the effect of a solvent on the color of a dye. Positive and negative solvatochromism should be differentiated, i.e., bathochromic and hypsochromic shifts with increasing solvent polarity. As the interaction of solvent with solute depends on the charges and the dipole moments of the various electron states of the solute molecules, a large change in electron densities causes strong solvatochromism. Experimentally, a bathochromic shift of 38 nm is found for **8** when the solvent is changed from cyclohexane to nitromethane. The spectra of azobenzene and Dimethyl Yellow are less sensitive to solvent changes. If a bathochromic shift is found with increasing polarity of the solvent, the excited state is more polar (and thus stabilized in a more polar solvent) than the ground state. Hypsochromic shifts arise when the ground state is more polar (and more strongly stabilized in a more polar solvent) than the excited state.

Association of two or more dye molecules, or ions, or aggregation of dye particles with other molecules or ions (ion pairs, aggregates, micelles) also changes the electronic spectra in most cases [68], [69].

Indigo [*482-89-3*] (C.I. Vat Blue 1, **12a**) is the most famous natural dye and was the most widely used dye in the world before the appearance of synthetic dyes. The elucidation of its structure by VON BAEYER and its first successful commercial synthesis by HEUMANN were the greatest achievements in color chemistry of the 19th century (see Chap. 1).

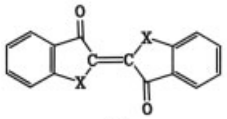

a) X = NH
b) X = S
c) X = Se

12

Qualitative and empirical theories were unable to explain either the blue color of this relatively simple molecule or the red color of its S-analogue, thioindigo [*522-75-8*] (**12b**). Consequently, the von Baeyer structure was repeatedly questioned, but was finally verified by X-ray analysis to be **12** [70], [71]. In the solid state, neighboring molecules are strongly hydrogen-bonded, which explains the high melting point (390 – 392 °C) and low solubility.

The spectrum of indigo has been investigated by HMO [72] and PPP methods [73–75]. The basic chromophore of indigo, thio-, seleno- and oxindigo is the partial structure **13**, consisting of a C=C double bond substituted by two donor groups X and two acceptor groups C=Y. This system is called a cross-conjugated or H-shaped chromophore (**13**).

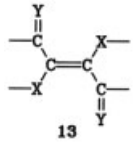

X = NH, S, Se, O
Y = O, (NR)

13

The benzene rings in indigo play only a secondary role, as shown by tetramethylpyrrolindigo (**14a**). In 1963 indigoid dyes and tetramethylpyrrolindigo were the only known examples of H-shaped chromophores.

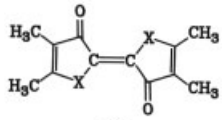

a) X = NH
b) X = S

14

Subsequent synthetic work produced compounds (**15**), (**16**), and **17**, which contain the n and π electrons of the H-shaped chromophore (**13**), but without the additional π electrons found in indigo, thioindigo, and pyrrolindigo [76–79]. The observed spectra (Table 4) and the similar shapes of the absorption bands corroborate the MO predictions. Additional support is afforded by **18** and **19**, which contain an analogous linearly conjugated system, but without cross-conjugation between the donor S atom and the carbonyl acceptor.

The color of indigo (**12a**) depends strictly on conditions: it ranges from red in the vapor phase (λ_{max} = 540 nm) to violet in a nonpolar solvent, and to blue in polar solvents and in the solid state. Amorphous and crystalline indigo differ significantly in their maxima (650 and 675 nm,

Table 4. The longest wavelength absorption bands (λ_{max}, nm, and in parentheses log ε) of indigo and thio- and selenoindigo (**12.**), cross-conjugated chromophores **15**–**17**, and not cross-conjugated chromophores **18** and **19**

Substituent	Compound					
X	**12**	**15**	**16**	**17**	**18**	**19**
NH	588 (4.30) ∗		487 (4.00)			
S	546 (4.26)	467 (3.66)	458 (4.34)	358 (3.76)	300 (4.12)	285
Se	564 (4.06)		482 (4.04)			

∗ In CCl_4

respectively). These differences probably reflect the influence of inter- and intramolecular hydrogen bonding.

The hypsochromic shift in the cross-conjugated thio series **12b**, **15b**, **16b**, and **17** is mainly due to decreasing planarity of the five- and six-membered rings. The linearly conjugated compounds **18** and **19** no longer show similarity to indigoid dyes.

After dividing **13** into two identical subsystems (by formal splitting of the C=C double bond), perturbation theory (PMO) explains the hypsochromic shift between X=NH and X=S, which is due to −S− being a poorer electron donor than −NH− [75]. The PPP method includes electron interaction, and although it is important for a quantitative description of the indigo spectrum, it does not change the general HMO and PMO results [29, pp. 115 ff], [75].

It is interesting to compare the bathochromic effect of cross-conjugated systems of type **13** (X=NH, Y=O) with those of polymethine dyes **20**and **21**, in which similar donors and acceptors are found at the ends of a linearly conjugated C=C chain (→ Methine Dyes and Pigments). In the case of the merocyanines (**20**), donor and acceptor groups are directly comparable with the respective groups of indigo (**12a**). An absorption band in the range of cross-conjugated systems containing only one C=C double bond is obtained only with a dodecahexaene system (**20b**), i.e., with six conjugated double bonds. Similarly with symmetrical cyanines (**21**), which show the strongest bathochromic shift of all polymethine dyes, at least three conjugated C=C double bonds are necessary to obtain absorption in the indigoid range.

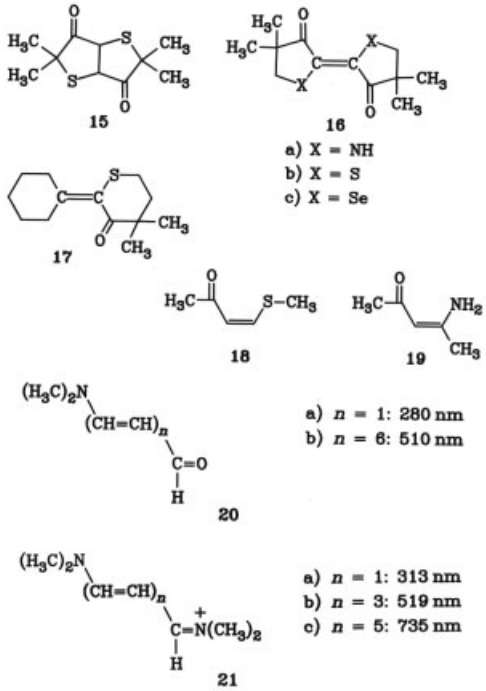

It would seem obvious that a combination of linearly conjugated double bond systems with cross-conjugated compounds should result in further bathochromic shifts. However, this expectation is not fulfilled by the series of compounds **22** and **23** (Table 5) [29, p. 124].

The examples in Table 5 show the difficulty of developing simple qualitative correlations between structure and electronic spectra of π systems.

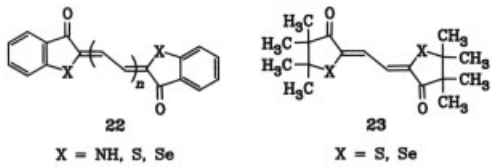

Table 5. The longest wavelength absorption maxima (λ_{max}, nm, and in parentheses log ε) of vinyl indigo blue dyes , p. 124[29]

	Substituent			
Compound	X	$n = 0$	$n = 1$	$n = 2$
22	NH	615 (4.32)	573 (4.35)	
	S	554 (4.17)	534 (4.36)	532 (4.56)
	Se	573 (4.10)	545 (4.25)	
23	S	450 (4.13)	446 (4.50)	454 (4.67)
	Se	476 (4.17)	463	

These two examples of aminoazobenzene and indigo derivatives demonstrate that three structural characteristics are important for long-wavelength electronic spectra:

1) Linear and cyclic systems of conjugated double bonds
2) Electron donors
3) Electron acceptors

Maximum overlap of the π orbitals of these systems is essential and can be achieved only in coplanar molecules and ions.

2.6. The Influence of Substituents on the Spectra of Aromatic Compounds

The discussion in the preceding Section 2.5 demonstrates that substituents significantly influence the spectra of azobenzene and indigoid dyes.

Simple benzene derivatives and the industrially important quinacridone and anthraquinone dyes show that qualitative rules based on chemical classification of substituents are applicable only within a narrow range; however, their spectra can be interpreted by quantum-mechanical methods.

The spectra of 2-, 3-, and 4-substituted nitrophenols, cyanophenols, and nitro-*N*-dimethylanilines, i.e., benzene derivatives with donor and acceptor substituents, cannot be understood from simple considerations of mesomeric structures. The *o*- and *m*-, but not the *o*- and *p*-compounds absorb very similarly over the whole range above 220 nm, whereas the absorption bands of the *p*-isomers are hypsochromically displaced and different in type. This problem was solved by the molecules-in-molecules (MIM-MO) [80] and PPP methods [81].

The introduction of a further donor (D) into a benzene derivative that already carries an electron acceptor (A) and an electron donor causes the following shifts relative to the 1,4-donor – acceptor compound:

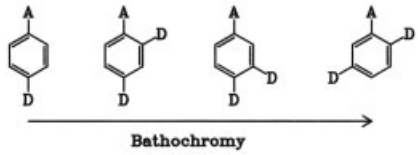

This is called the rule of the distribution of auxochromes, which has been known since 1911. It is corroborated by many examples of industrially important dyes of the azo, anthraquinone, and other classes. The customary concept of mesomeric structures is not able to explain the energy in the excited state, but an MIM investigation gives the correct answer [80].

Examples of the distribution of two donors and acceptors each on benzene are provided by the pigment quinacridone (**24**) and its isomer **25**. The former is reddish violet ($\lambda_{max} = 523$ nm), whereas **25** is yellow ($\lambda_{max} = 436$ nm).

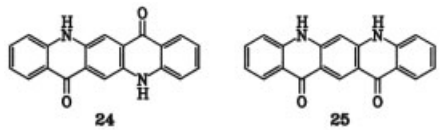

For the assignment of bands to specific transitions, spectra of oriented molecules (dye crystals or dyes adsorbed on stretched polymer films) in polarized light are useful. Dichroism, i.e., different extinctions parallel and perpendicular to the direction of light polarization, is frequently observed. This method settled a long controversy over the spectrum of anthraquinone [82]. The three bands at 320, 270, and 250 nm arise from

three $\pi \rightarrow \pi *$ transitions. The first and third bands are polarized along the axis of the three six-membered rings; the second transition arises from an excitation perpendicular to this axis.

The dichroic character of many dyes, particularly anthraquinone derivatives, has become important for liquid-crystal displays of small computers or watches. Dyes are adsorbed uniformly onto the liquid crystals, which become ordered in an electric field; the effect of dichroism can be used to produce color contrasts.

In summary, it is possible in principle to explain electronic absorption bands quantitatively. It has been stated that "the routine type and reliable calculation of wavelengths and extinctions of any organic compound is possible today" [29] and that "the search for new chromogens need not now depend on serendipity or massive synthetic attack. A pen, paper, an inventive mind and a modest computer can achieve much more, with far greater economy" [63].

At present these statements can be only partly supported. It is insufficient if the extinction maxima, calculated by PPP, are correct within only ca. 100 nm [29, p. 7]. The calculation of oscillator strengths and extinctions is even more crude, and the theories indicate almost nothing about the *band shape* which may be decisive for color (Fig. 3). Future progress is to be expected, however, with the help of faster computers and more precise approximation methods.

Fukui, in his Nobel lecture, emphasized the mutual correlations between theoretical and experimental reaction mechanisms [59]; similar considerations apply to light absorption.

Calculations of dye spectra are much more successful if a generalized set of PPP parameters, limited in number but applicable to all known classes of dyes, is used [62]. The so-called valence-state ionization potential (VSIP) of a bond is the parameter that exerts the largest influence in wavelength calculations. However, although these parameters are related to only one bond, they are very sensitive to the overall structure of the dye [83]. They are therefore not yet the final solution.

The biggest advantage of a quantum-chemical treatment lies perhaps in the comparison of energies of adjacent occupied and unoccupied molecular orbitals, as shown in the free-electron model. A second advantage is the possibility of elucidating the basic structural features of *particular* known colored compounds, as shown for indigo.

3. Color Measurement and Colorant Formulation

The color of a material is governed by the composition of the light reflected by that material into the observer's eye. Dyes modify the color of the substrate because they absorb light in the visible region of the spectrum. Nonwhite pigments also possess this ability, but in addition, all pigments scatter light; however, with dyed materials, the light is scattered by the substrate. The combined effect of absorption and scattering is measured by determining the reflectance factor at each wavelength in the visible region of the spectrum, using a reflectance spectrophotometer. Although reflectance factors completely quantify the color characteristics of the material, these factors do not constitute color measurement because perceived color also depends on the incident light, i.e., the spectral power distribution. In addition, two materials viewed under the same light may appear to have the same color, yet their reflectance curves may be different, a phenomenon known as metamerism.

Color measurement requires the mathematical reduction of the two dissimilar reflectance curves to give the same numerical definition of color; such a method was introduced by the CIE (Commission Internationale de l'Éclairage) in 1931. With minor modifications this is the only method used today. It defines a color by three numbers known as *tristimulus values, X, Y*, and *Z*. These numbers quantify the amounts of three imaginary lights which, if shone onto a pure white surface, would match the color of the material illuminated by a specified source. This method, using real lights, was first conceived by Maxwell in the 19th century.

Tristimulus values are rarely used to specify a color, but are widely used to predict how much of which colorants (dyes or pigments) would be required to produce a color (colorant formulation) and to quantify the perceived difference between materials of similar color (quality control).

3.1. Measurement of Reflectance Factors

The reflectance factor of a given material at wavelength λ is the ratio of amount of light of wavelength λ reflected by the material to that reflected by a perfect reflecting diffuser illuminated in the same manner. The instrument used for this determination is a reflectance spectrophotometer; such instruments plot a reflectance curve or display its digital equivalent.

The first reflectance spectrophotometer sequentially irradiated the specimen and a block that was coated with magnesium oxide (which was the CIE standard of reflectance until 1969) with monochromatic light, the reflection from both surfaces being compared by a photocell [84]. The reflectance factor at each wavelength is expressed as a percentage (0 – 100 scale) although calculations are made on reflectance factors expressed on a 0 – 1 scale.

However, such instruments are not suitable for fluorescent specimens since they are based on the assumption that the incident and reflected light have the same wavelength. If the specimen is fluorescent, the photodetector gives spurious results over a significant region of the spectrum.

When fluorescent brightening agents became widely used in the 1950 s, spectrophotometers needed to be designed that irradiated the specimen with polychromatic light whose spectral power distribution (SPD) resembled that of daylight. The light that comes from the specimen consists of both reflected and emitted light and is passed to the photodetector through a series of interference filters. The SPD of daylight was established by the CIE in 1963 as Illuminant D65, the 65 indicating that its color temperature was ca. 6500 K. The SPD covered the range of 300 – 830 nm, the inclusion of the ultraviolet region being essential when brightening agents were measured. Most modern spectrophotometers attempt to simulate D65, but this has proved to be difficult [85, p. 122].

Figure 12 illustrates a modern spectrophotometer. A pulse of light of 20 – 30 μs duration from a xenon arc passes through a filter, which renders the SPD closer to that of D65, through an integrating sphere to the sample being measured. The light coming from the sample passes to a monochromator, which is a reflecting diffraction grating, and to an array of 17 photodiodes spaced every 20 nm from 380 to 700 nm. A Fresnel prism is then moved over the exit port, and a second pulse of light enters the sphere. The light then falling on the monochromator comes from the sphere wall which acts as a transfer standard. This is known as dual – beam operation; it automatically compensates for the fact that a colored specimen changes the color of the irradiating light. The signals from the 17 photodiodes are electronically processed in parallel, the final output being the reflectance factors at every 20 nm from 380 to 700 nm. Every few hours the instrument is calibrated by measuring the reflectance of a white tile and a black tile, whose reflectances relative to the perfect reflecting diffuser have been determined against standard tiles.

The wall of the sphere contains a port, which permits the inclusion or exclusion of the specular component of the reflected light. This is essential because in the case of nonmetallic specimens, specularly reflected light has not the SPD of diffusely reflected light, but that of the irradiating light. These two viewing geometries are known as total/normal if the port is closed with the material used for the sphere wall and diffuse/normal if it is open and leads to a light trap. Other spectrophotometers irradiate the specimen with a beam of light normal to the surface and view it at 45°, known as 0/45 or vice versa, 45/0. These are primarily intended for the measurement of moving specimens since they minimize the effects of any variations in the distance from the port. Measurements on glossy specimens with 0/45 or 45/0 geometries (which give identical results because of the Helmholtz reciprocity principle) usually lie between those obtained with total/normal and diffuse/normal viewing geometries. Total/normal geometry is preferred for colorant formulation, and diffuse/normal geometry is preferred for evaluating color differences. All of these viewing geometries are defined by the CIE [86].

Most reflectance spectrophotometers also permit insertion of an ultraviolet absorbing filter into the incident beam (not shown in Fig. 12). This enables fluorescent specimens to be recognized, as the reflectance curves with and without the ultraviolet component are different. It also permits the effectiveness of fluorescent brightening agents to be measured.

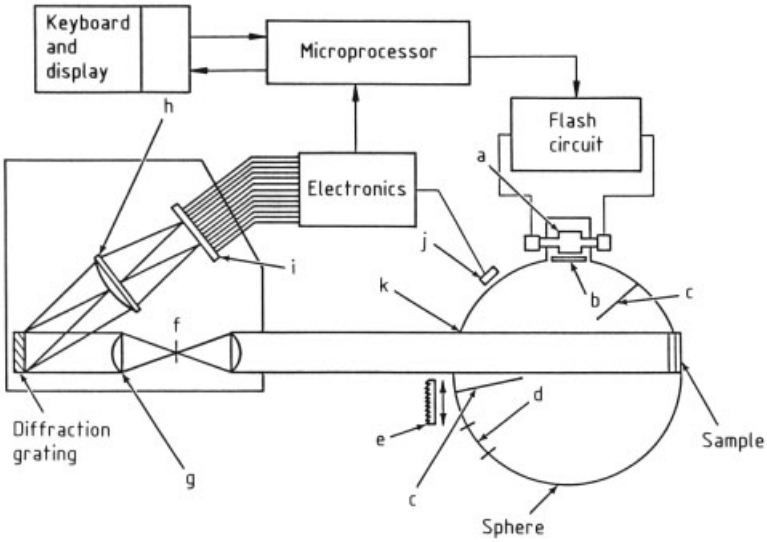

Figure 12. Reflectance spectrophotometer
a) Xenon arc; b) D65 filter; c) Baffle; d) Specular port; e) Fresnel prism; f) Entrance slit; g) Collimating lens; h) Integrating lens; i) Photodiodes; j) Reference detector; k) Viewing port

3.2. Calculation of Tristimulus Values

A reflectance curve does not quantify the relative amounts of light at each wavelength reflected by a specimen unless it is illuminated by an equal energy source, a theoretical possibility. The next stage in color measurement is, therefore, to determine those relative amounts when the specimen is illuminated by a real source. In 1931 the CIE defined three such sources, Illuminant A, representing the light from a tungsten filament lamp, and Illuminants B and C, representing daylight at a color temperature of ca. 4 900 and 6 800 K respectively. Illuminant A is still valid, Illuminant B rapidly fell into disuse, and Illuminant C was supplemented in 1963 by Illuminant D65, which (unlike C) has the correct amount of ultraviolet radiation.

Formulas and tables for calculating the SPD of daylight at any color temperature between 4000 and 25 000 K were also published, and any other SPD can be used in the calculation, e.g., that of a fluorescent tube, provided that the specimen is not fluorescent. If it is, the only valid calculation is that using the SPD of the light irradiating the specimen during measurement. Therefore, the next stage in color measurement is to multiply the reflectance factor (on a 0 – 1 scale) at each wavelength by the relative amount of light of this wavelength in the illuminant:

$$R(\lambda) \times S(\lambda)$$

Each of these products is then multiplied by the color-matching functions of the standard observer.

The Standard Observer. Most colors can be matched perfectly by mixing red, green, and blue lights. If these are defined, e.g., by wavelength and power, a given observer always requires the same amounts of red, green, and blue lights to match a given color irrespective of its spectral composition. Therefore, these amounts quantify perceived color and are known as tristimulus values *R, G,* and *B*. MAXWELL deduced that if the amounts of *R, G,* and *B* required to match monochromatic light throughout the visible spectrum were determined, these values would enable the tristimulus value of any stimulus to be calculated if the SPD of the stimulus were known. These amounts are known as color-matching functions, and those obtained by GUILD [87] and WRIGHT [88] were adopted by the CIE in 1931 to define the standard observer.

Imaginary Primaries. HELMHOLTZ and MAXWELL had observed that it was impossi-

ble to select red, green, and blue lights that, by additive mixing, could match all monochromatic lights:

$$Q \not\equiv R+G+B$$

where $\equiv$ indicates perfect color matching. However, MAXWELL discovered that a color-matching equation could be set up with any monochromatic light, provided that one of the additive primaries *R, G,* or *B* could be added to *Q*; i.e., for any wavelength $Q(\lambda)$, one of the following color-matching equations could be set up:

$$Q(\lambda)+R \equiv B+G$$
$$Q(\lambda)+G \equiv R+B$$
$$Q(\lambda)+B \equiv R+G$$

Since any monochromatic stimulus could become part of a color matching equation of this type, this also applies to any polychromatic stimulus. The laws of additive color mixing state that color-matching equations could be treated algebraically, and therefore, every light could be specified in terms of *R*, *G*, and *B,* provided that one could be negative, for example:

$$Q(\lambda) \equiv B+G-R$$

In 1931 the CIE decided that it would be unwise to establish a system of color specification employing negative amounts of one of the primaries, but that these could be eliminated by converting *R, G,* and *B* into imaginary primaries *X, Y,* and *Z.* The CIE, therefore, converted the means of the Guild – Wright data involving 17 observers into the amounts, all positive or zero, of three imaginary primaries required to match monochromatic light at every 5 nm from 400 to 700 nm. These amounts are known as the color-matching functions of the CIE 1931 standard colorimetric observer, represented by $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ [85, pp. 104 – 108].

The color-matching experiments of GUILD and WRIGHT involved a split circular field subtending an angle of 2° at the eye, approximately that of a thumbnail held at arm's length. Since industrial color matching employs larger specimens, the CIE established the CIE 1964 supplementary standard colorimetric observer using a split 10° field. These color-matching functions are designated $\bar{x}_{10}(\lambda)$, $\bar{y}_{10}(\lambda)$, and $\bar{z}_{10}(\lambda)$ and are free of the defects subsequently discovered in the 2° data [89]. Both sets have now been interpolated and extrapolated to cover the range of 360 – 830 nm at 1-nm intervals, to about eight significant figures. Unless assessments of color are made on small areas, e.g., color photographs, the color-matching functions of the 10° observer should always be used.

Tristimulus values are calculated by multiplying the product of the reflectance factor $R(\lambda)$ and the spectral power distribution of the illuminant $S(\lambda)$ by each of the three color-matching functions, ideally at every 5 nm from 380 to 780 nm; however, many modern spectrophotometers use values at every 20 nm from 400 to 700 nm. The individual values involving the three functions are then summed and multiplied by a normalizing factor *k* to give the *XYZ* tristimulus values:

$$X = k \sum_{380}^{780} R(\lambda)\, S(\lambda)\, \bar{x}(\lambda)$$
$$Y = k \sum_{380}^{780} R(\lambda)\, S(\lambda)\, \bar{y}(\lambda)$$
$$Z = k \sum_{380}^{780} R(\lambda)\, S(\lambda)\, \bar{z}(\lambda)$$

where

$$k = 100 \left[\sum_{380}^{780} S(\lambda)\, \bar{y}(\lambda) \right]^{-1}$$

The values X_{10}, Y_{10}, and Z_{10} are calculated similarly from the color-matching functions of the 10° observer.

The constant *k* results in the perfect reflecting diffuser (the CIE standard of reflectance since 1969) having a *Y* value of 100.00 under every illuminant. Every modern spectrophotometer has a built-in microprocessor that performs these calculations, and many display tristimulus values for several illuminants by using both the 2° and 10° color-matching functions.

Tristimulus Colorimeters. The first spectrophotometers were expensive and slow and required highly skilled operators. The need for a simpler instrument was met by the tristimulus colorimeter, designed to respond to spectral distributions of light like the human eye. This

was achieved by making the photodetector response spectrally equivalent to the CIE standard observer color-matching functions. For this purpose three or four filters are used [90].

The first commercial colorimeter was developed in 1952 at the National Bureau of Standards (NBS) in Washington, D.C. [90]. Although its accuracy left much to be desired, it provided rapid measurements at low cost. Spectrophotometers are just as fast and easy to use and are able to deal with metamerism, whereas the colorimeters can only duplicate a single observer – illuminant combination.

3.3. Chromaticity Diagram

There are three independent variables of perceived color; in the case of surface colors, they are usually termed lightness, hue, and saturation. Only one of these, lightness, is directly related to a single tristimulus value, the *Y* value; interpretation of the other two variables is facilitated by determining the ratio of each tristimulus value to their sum. These ratios are termed chromaticity coordinates, given by

$$x = \frac{X}{X+Y+Z}$$
$$y = \frac{Y}{X+Y+Z}$$
$$z = \frac{Z}{X+Y+Z}$$

As $x+y+z=1$, only two chromaticity coordinates are required to specify the saturation and hue of any color. Any two could be used, but *x* and *y* are always chosen. The spectrum locus and the purple line representing all the colors obtained by mixing the ends of the spectrum are shown in Figure 13. This continuous line is the boundary of all real colors; the *X, Y,* and *Z* lights, being imaginary, lie outside at the following positions:

$$X \text{ at } x=1, y=0$$
$$Y \text{ at } x=0, y=1$$
$$Z \text{ at } x=0, y=0$$

The position of the two most important illuminants, D65 (daylight) and A (tungsten light) are shown, as well as the positions of two surface colors *P* and *Q*. The hue of *P* can be quantified by drawing a straight line through the positions of D65 and *P* and extrapolating it until it cuts the spectrum locus at λ_d (*P*). The wavelength corresponding to this point is the dominant wavelength of *P* and this is defined by the CIE as the "wavelength of the monochromatic light stimulus that, when additively mixed in suitable proportions with the specified achromatic light stimulus, matches the color stimulus considered."

In the case of surface color *Q,* the line must be extrapolated in the opposite direction to cut the spectrum locus at λ_c (*Q*). The wavelength corresponding to this point is the complementary wavelength of *Q,* which means that when light of this wavelength is added in the correct proportions to light whose chromaticity coordinates are the same as *Q,* the mixture matches the specified achromatic stimulus, usually D65.

Saturation can be quantified by determining the position of the specimen on the line drawn from D65 through the position of the specimen to the boundary, i.e., the spectrum locus in the case of *P* and the purple line in the case of *Q*. A specimen that lies halfway along this line has an excitation purity p_e of 0.5. Saturation can also be quantified by multiplying the excitation purity by y_d/y, where y_d is the *y* chromaticity coordinate of the monochromatic light stimulus and *y* is that of the specimen. This measure of saturation is termed colorimetric purity.

None of these measures of hue and saturation, however, became widely used, which is fortunate because they, together with the *Y* value as a measure of lightness, are unreliable. Specimens that have the same perceived hue but with variation in purity rarely have the same dominant or complementary wavelength. Neither of the two measures of purity is a reliable measure of the perceived variable known as saturation. The value of a grey, which is perceptually midway between black ($Y \approx$ ca. 0) and white ($Y \approx 100$) has a *Y* value of 20 instead of ca. 50. This shows that the *XYZ* space is nonuniform, a fact that was recognized in the mid-1930 s, and more effort has been expended in the search for a uniform color space since that time than that required to develop the CIE system. The principal incentive for this effort, which is still being pursued, is that the distance between the posi-

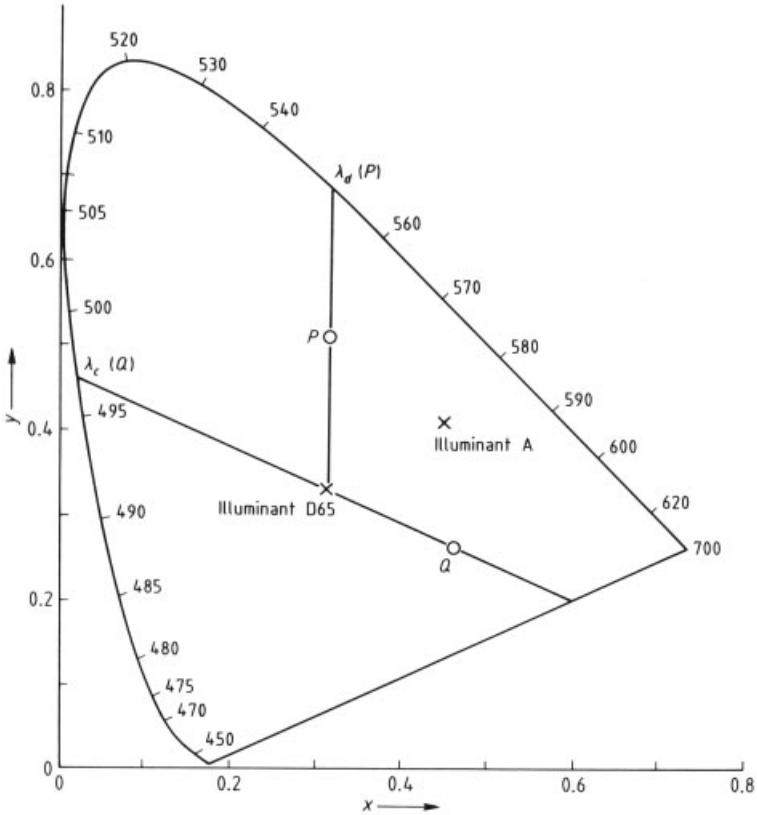

Figure 13. The CIE 1931 *xy* chromaticity diagram
For explanation of abbreviations, see text

tions of two specimens in such a space would be proportional to their perceived color difference. Such distances are given the symbol ΔE (E is derived from the German *Empfindung* for sensation). Formulas for calculating ΔE are known as color-difference equations, of which more than 20 have been devised since 1935. With most of these, the *XYZ* tristimulus values of standard and batch are first converted into the coordinates of a more uniform color space. The most uniform color spaces devised so far are the CIE 1976 ($L*u*v*$) color space known as CIELUV and the CIE 1976 ($L*a*b*$) color space known as CIELAB.

The CIELUV is primarily intended for additive mixtures of lights; the associated $u'v'$ diagram, a linear transformation of the *xy* chromaticity diagram, possesses the essential attribute that mixtures of two lights are plotted along the straight line joining their positions in the $u'v'$ diagram. The CIELAB, on the other hand, is unsuitable for lights, as the associated $a*b*$ diagram is a nonlinear transformation of the *xy* chromaticity diagram. This is, however, no disadvantage where surface colors are concerned, and CIELAB is used worldwide for this purpose. Consequently, only CIELAB is described in detail here; the equations for calculating CIELUV coordinates are given by CIE [86].

3.4. Conversion of *XYZ* Values into *L*a*b** Coordinates

When the CIE defined the imaginary primaries *X, Y,* and *Z* from the real primaries *R, G,* and *B,* which were monochromatic lights at 700, 546.1, and 435.8 nm, respectively, *X* and *Z* were chosen

to have 0 brightness. This meant that the brightness of any light or the lightness of any colored object would be quantified by the Y tristimulus value alone. The Y values of a series of grey specimens equally spaced perceptually from black to white, however, lie not on a straight line, but on a curve, the grey midway between black and white having a Y value of ca. 20 instead of ca. 50.

The first step in converting XYZ space into a uniform color space is, therefore, to convert the Y values into other values that are linear with respect to perceived lightness. In the CIELAB (and CIELUV) uniform color spaces, these values are designated $L*$ and are given by

$$L^* = 116\,(Y/Y_n)^{1/3} - 16$$
$$\text{if } Y/Y_n > 0.008856 \text{ and}$$
$$L^* = 903.3\,(Y/Y_n)$$
$$\text{if } Y/Y_n \leq 0.008856$$

where Y_n is the Y value of the perfect reflecting diffuser, which is always 100.000.

For a perfect black $L* = 0.00$, for the perfect reflecting diffuser $L* = 100.00$, a grey midway between black and white has an $L*$ value of ca. 50. Similar functions are applied to X and Z, and the CIELAB $a*b*$ coordinates are given by

$$a^* = 500\,[f\,(X/X_n) - f\,(Y/Y_n)]$$
$$b^* = 200\,[f\,(Y/Y_n) - \mathrm{f}\,(Z/Z_n)]$$

where

$$f\,(X/X_n) = (X/X_n)^{1/3} \text{ if } X/X_n > 0.008856 \text{ but}$$
$$f\,(X/X_n) = 7.787\,(X/X_n) + 16/116$$
$$\text{if } X/X_n \leq 0.008856$$
$$f\,(Y/Y_n) = (Y/Y_n)^{1/3} \text{ if } Y/Y_n > 0.008856 \text{ but}$$
$$f\,(Y/Y_n) = 7.787\,(Y/Y_n) + 16/116$$
$$\text{if } Y/Y_n \leq 0.008856$$
$$f\,(Z/Z_n) = (Z/Z_n)^{1/3} \text{ if } Z/Z_n > 0.008856 \text{ but}$$
$$f\,(Z/Z_n) = 7.787\,(Z/Z_n) + 16/116$$
$$\text{if } Z/Z_n \leq 0.008856$$

The values X_n, Y_n, and Z_n are the tristimulus values of the perfect reflecting diffuser (which are the same as those of the illuminant), the appropriate values being

Illuminant D65/10°	$X_n = 94.8$	$Y_n = 100.0$	$Z_n = 107.3$
Illuminant D65/2°	$X_n = 95.0$	$Y_n = 100.0$	$Z_n = 108.9$
Illuminant C/10°	$X_n = 97.3$	$Y_n = 100.0$	$Z_n = 116.1$
Illuminant C/2°	$X_n = 98.1$	$Y_n = 100.0$	$Z_n = 118.2$

Color differences in CIELAB unis are then given by

$$\Delta E = \left[(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2\right]^{0.5}$$

The CIELAB color space, one of the most uniform devised thus far, has another valuable characteristic: all colors are arranged around a central vertical axis, the $L*$ axis. This is the inherently logical visual color-order system that has been discovered, first in the 17th century, and then in this century by Ostwald and Munsell. If the Cartesian coordinates $L*$, $a*$, and $b*$ are converted into cylindrical coordinates, these quantify the three variables of a perceived color : lightness, saturation, and hue.

Saturation is the variable by which a surface color differs from the grey of the same lightness and is quantified by the distance from the achromatic axis. This is given by

$$\left[(a^*)^2 + (b^*)^2\right]^{0.5}$$

and is termed chroma (symbol $C*$), which is the term coined by Munsell. The term saturation should not be used for this variable because it has been defined by the CIE in such a way that it cannot be quantified by CIELAB coordinates and is of little value for object colors. Hue is quantified by the angle that the chroma radius, passing through the position of the color, makes with the positive $a*$ axis; it is given by

$$h = \arctan\,(b^*/a^*)$$

expressed on a scale from 0 to 360°. The relationship between these CIELAB coordinates is shown in Figure 14.

The CIELAB $L*$, $C*$, and h coordinates provide a numerical identification of any color that, unlike XYZ or $xy\,Y$ coordinates, could be understood by artists and designers. Therefore, it has been recommended that all collections of physical specimens, so-called color atlases, should include these coordinates [91].

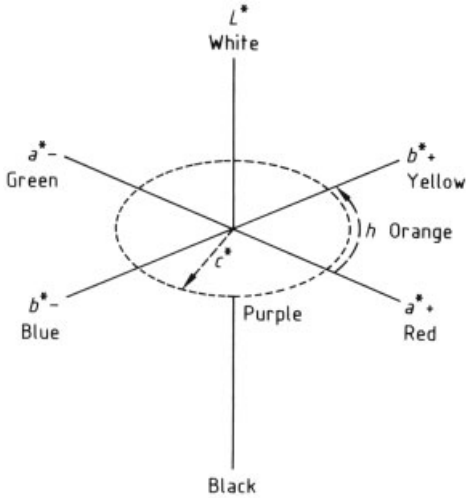

Figure 14. CIELAB color space

3.5. Calculation of Color Differences

The $L*$, $a*$, and $b*$ values of any color can be regarded as coordinates in a three-dimensional Euclidean space; two colors that are not a perfect match to the 2° or 10° observer under a specified illuminant are not located at the same point in $L*a*b*$ space, and as the match is worse the greater is the separation. This distance is easily calculated by applying the Pythagorean theorem in three dimensions:

$$\Delta E = \left[(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2\right]^{0.5}$$

where

$$\Delta L^* = L^*_{\text{batch}} - L^*_{\text{standard}}$$
$$\Delta a^* = a^*_{\text{batch}} - a^*_{\text{standard}}$$
$$\Delta b^* = b^*_{\text{batch}} - b^*_{\text{standard}}$$

and ΔE is the color difference in CIELAB units. The value $L*$ quantifies the lightness difference: a batch is lighter than standard if $\Delta L*$ is positive and darker if it is negative. Differences in the other two variables of perceived color, chroma and hue, are calculated as follows:

$$C^* = \left[(a^*)^2 + (b^*)^2\right]^{0.5}$$
$$\Delta C^* = C^*_{\text{batch}} - C^*_{\text{standard}}$$

A batch is stronger than standard if $\Delta C*$ is positive and weaker if it is negative. The hue difference, Δh, is given by

$$\Delta h = \left[(\Delta E)^2 - (\Delta L^*)^2 - (\Delta C^*)^2\right]^{0.5}$$

This can be qualified by the terms redder, yellower, greener, or bluer by reference to the $a*b*$ diagram [85, p. 156]. Although CIELAB was found to be the most reliable color difference formula devised up to its inception in 1973, it was not sufficiently reliable to permit decisions on color matching to be made with reference to a value of ΔE that depended only on the closeness of match required. Although an average of 17 % of wrong decisions were made by professional colorists. [92], instrumental decisions based on the optimum value of ΔE were higher. To overcome this defect, visual pass/fail decisions made by professional colorists on many batches dyed to match standards of hundreds of different colors were obtained. This permitted specifications to be made around any point in CIELAB color space of the ellipsoid in such a way that every batch lying on the ellipsoid surface would have the same acceptability rating. The ellipsoids were defined by three semiaxes oriented in the orthogonal directions corresponding to changes in lightness, chroma, and hue, the overall scale being such that batches lying on the ellipsoid surface would exhibit a color difference from the standard of 1 JPC79 unit. Color differences in these units are given by the position of a batch in relation to the point on the ellipsoid surface where the line from the standard location passing through the batch location cuts this surface [93].

This formula was subsequently improved by the Colour Measurement Committee of the Society of Dyers and Colourists [94] and is known as CMC ($l:c$). When $l = c = 1$, the formula quantifies the perceptibility of any small color difference, a value of 1 being about 3 times larger than a barely perceptible difference. Equally perceptible color differences are however, not always equally acceptable as industrial color matches; in particular, differences in lightness are more likely to be acceptable than equally perceptible differences in hue. To allow for this, l and c are tolerances for lightness and chroma differences relative to hue differences, and $l = 2$, $c = 1$ have been found to be the best values for textiles.

Color differences in CMC ($l:c$) units are obtained from the $L*$, $C*$, and h values of standard and $\Delta L*$, $\Delta C*$, and ΔH by using the formula

$$S_L = \frac{0.040975L^*}{1 + 0.01765L^*}$$

unless $L^* < 16$ when $S_L = 0.511$

$$S_c = \frac{0.0638C^*}{1 + 0.0131C^*} + 0.638$$

$$S_h = (fT + 1 - f)\, S_C$$

where $f = \left[\frac{(C^*)^4}{(C^*)^4 + 1900}\right]^{0.5}$

and $T = 0.36 + 0.4\cos(h + 35)$

unless h is between 164 and 34° when

$$T = 0.56 + 0.2\cos(h + 168)$$

Color differences in CMC ($l:c$) units are then given by

$$\Delta E = \left[\left(\frac{\Delta L^*}{S_L}\right)^2 + \left(\frac{\Delta C^*}{S_C}\right)^2 + \left(\frac{\Delta H}{S_H}\right)^2\right]^{0.5}$$

In addition, two formulas M & S83A and Datacolor, have been developed by optimizing the constants against masses of visual data. The former is widely used throughout the world, but since neither has been published and the use of each is confined to the instruments of a single manufacturer, they are not considered here further.

3.6. Colorant Formulation

When increasing amounts of a dye are applied to a substrate, the shape of the reflectance curve changes in a systematic but complex manner, as shown in Figure 15 for red, yellow, and blue dyes. Similarly shaped curves are also obtained when paints are compounded with increasing ratios of chromatic to white pigment. At every wavelength a simple function of reflectance varies approximately linearly with the concentration of chromatic colorant. This function is additive when more than one chromatic colorant is present. It quantifies the coefficient of absorption K divided by the coefficient of scatter S [95], [96]:

$$K/S = (1 - r)^2 / 2r$$

If the *K/S* value for a known concentration of each dye is known, the reflectance curve of any colorant mixture can be predicted from the concentration of each component, but this is of no practical value. The converse procedure, however, is colorant formulation: the reflectance curve of the target shade is measured, the *K/S* values at every 20 nm from 400 to 700 nm are calculated, and the amounts of each colorant are varied until the mixture yields these values. This can only be done effectively by computer. The first was an analogue computer introduced in 1958 in which the calculations were facilitated because all of the scatter was caused by the substrate and because a single constant for *K/S* could be used [97]. This method was successful only for pastel shade paints, because in paints of higher chroma the scattering caused by the chromatic pigment is significant. This necessitated the use of separate constants for K and S, and this could only be carried out on a digital computer.

The first program for a digital computer was devised in 1961 [98]. For textile dyeing the single-constant method was used and in 1964, a two-constant program was written for the same computer for paint formulation [99].

In addition to formulating a mixture with the same reflectance curve as the target shade, a color-matching computer also predicts mixtures yielding the same tristimulus values under D65. Such matches are, of course, metameric, but this is not serious if the degree of mismatch under any conventional illuminant is acceptable. This is indicated by the computer calculating metameric indexes, which are the color differences under every relevant illuminant.

A modern match-prediction system prints out every combination with a maximum metameric index below a preselected value together with the colorant cost, enabling the colorist to make the best choice. The mathematics of computer colorant formulation is extremely complex [85, pp. 172 – 186], [100–102].

4. Chemical Analysis and Tests

Industrial analysis of dyes aims at (1) homogeneity testing, (2) separation of mixtures, (3)

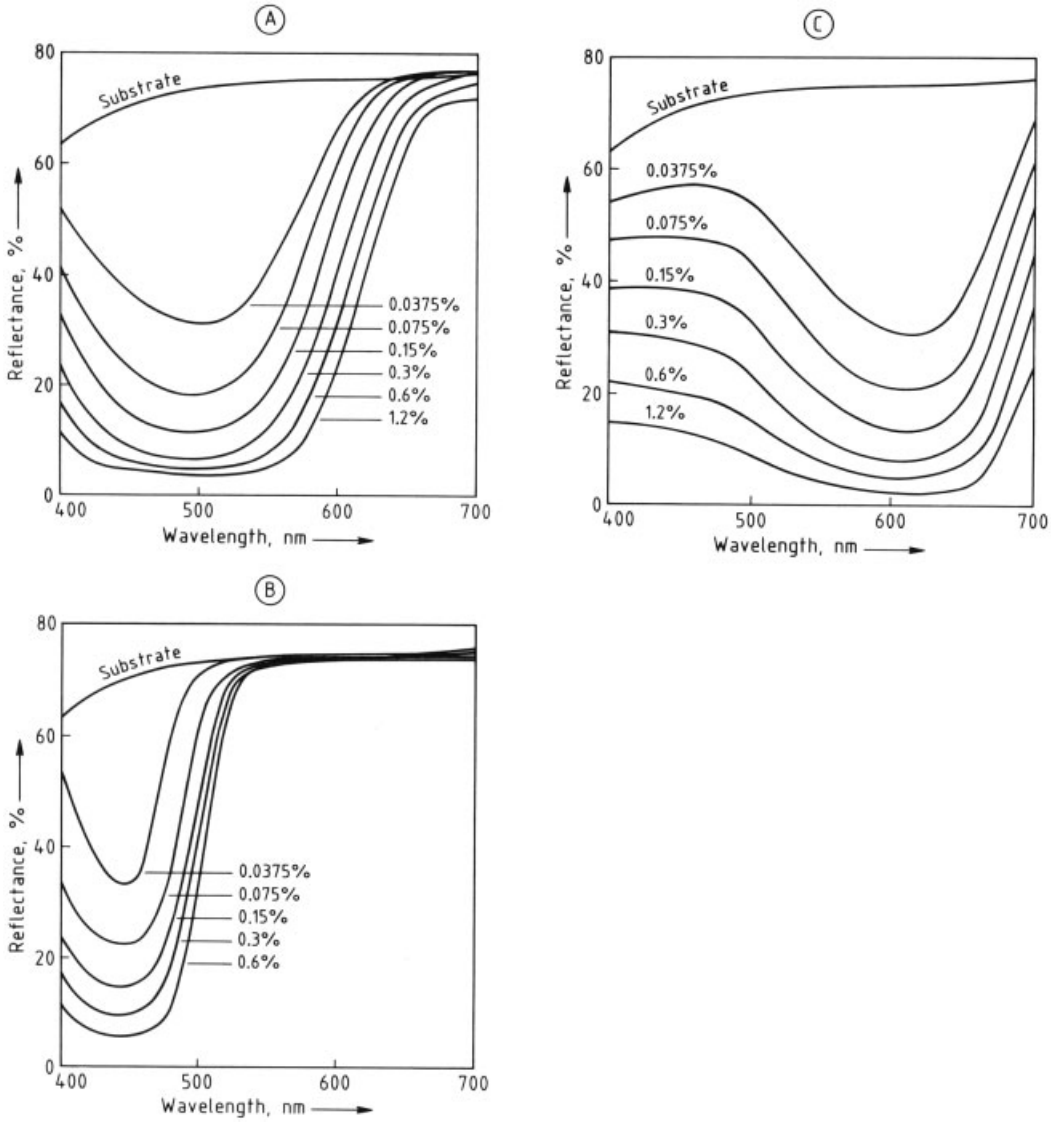

Figure 15. Reflectance curves of cotton dyes with increasing amounts of a red dye (A), a yellow dye (B), and a blue dye (C)

classification, (4) identification, and (5) quantitative determination [103].

Commercial dyes number several thousands and there is no universally valid scheme of analysis. A conclusion may be drawn about the type of dye if additives such as sodium sulfide, sodium dithionite (hydrosulfite) or substrates (alumina, barium sulfate, titanium dioxide, etc.) are found. In the dyeing of cellulose acetate, polyester, and polyamide fibers and plastics, as well as for spin dyeing, pigment printing, and other purposes, such additives as emulsifiers (→ Emulsions), dispersing agents, plasticizers, and others may be present. Water-soluble liquid preparations may contain glycols, glycol ethers, formic acid, acetic acid, or xylene, whereas liquid grades of oil-soluble dyes often contain mixtures of high-boiling aromatic compounds. On the fiber, dyes are sometimes associated with chromates, finishing agents, textile softeners, mordants, and aminoplasts, all of which must be separated for analysis.

4.1. Investigation of Individual Dyes

4.1.1. Testing for Homogeneity and Separation of Mixtures

The quickest and simplest test for dye homogeneity is the *blowing test,* in which a small sample of powdered dye on the tip of a spatula is blown against a moist filter paper, which may then be heated. In case of a dye mixture prepared by dry mixing, the particles on the paper dissolve with their characteristic color and the different dyes can be distinguished. The paper can also be moistened with ethanol, acetone, dimethylformamide, hydrochloric acid, or sodium hydroxide. Moistening first with acid and then with base often produces a color change. The dye can also be blown onto concentrated sulfuric acid in a porcelain dish to give characteristic colors. The blowing method fails when the dye mixture has been made by wet blending (of dye solutions) and with paste-type and liquid dyes.

Dye mixtures can often be separated by *fractional dissolution.* Thus, disperse dyes can be dissolved with acetone or ethyl acetate out of mixtures of acid dyes or substantive and vat dyes. Acid dyes can be separated from substantive dyes on the basis of their different solubility in ethanol, acetone, acetic acid, and other solvents. With acid solvents, the substantive or direct dye is sometimes left behind as water-insoluble free acid or salt. Basic dyes on the other hand, can be extracted from aqueous solutions by ether.

Water-soluble dyes can be separated by *solvent partition* [104]. Water is used as one phase; amyl alcohol, possibly mixed with a second solvent, is the other. In this way, dyes can be divided into eleven groups.

Other methods include *reduction and oxidation.* For example, in a mixture of water-soluble Quinoline Yellow and an azo dye, the azo dye can be reduced with ammonia and zinc dust, whereas the Quinoline Yellow remains unaltered. In a mixture of Copper Phthalocyanine Blue and Green, the blue component is oxidized by boiling nitric acid but the green remains unchanged.

Often it is possible to separate a mixture by *fractional dyeing,* first for a short time at low temperature with a small amount of material and then with fresh material at higher temperature. Repeated dyeing at various pH values may also achieve the objective. In the case of mixtures for dyeing fiber blends, fractional dyeing is often possible on the individual types of fibers.

Chromatographic methods, such as column chromatography, paper chromatography, thin-layer chromatography (in its various forms), and high-performance liquid chromatography, particularly in the reverse-phase mode, have proved excellent both in testing for dye homogeneity and in analytical separation and determination of mixtures (see Chap. 5). These methods can also be used on a preparative scale for the isolation of components of mixtures.

Paper electrophoresis is useful for indicating the electrical charge on a dye molecule or complex, particularly when the charge on the cation (Co or Cr) is modified by one or more sulfonic acid groups in the aromatic rings [106].

Dyes can be separated from foods by shaking them with polyamide powder in acid solution; the water-soluble dyes are adsorbed, but the water-soluble foods are not and are washed off. Methanol – ammonia serves for desorption of the dyes.

4.1.2. Classification

For a preliminary classification, behavior in water, dilute sodium hydroxide, and alkaline sodium dithionite is investigated. Table 6 lists the conditions of the dyeing test.

Water-soluble Dyes. Most water-soluble dyes dissolve immediately without color change. They can be classified by a few tests of their behavior toward wool, cotton, tannin-treated cotton, and cellulose acetate (Table 6).

Basic dyes give positive results on tannin-treated cotton and cellulose acetate, but not on wool (with some exceptions) or cotton.

Sodium hydroxide usually changes or decolorizes solutions of basic dyes. The dye can then be extracted with ether and back-extracted into dilute acetic acid. Basic chrome dyes are not absorbed in ether. Many basic dyes are not resistant to alkali. Some basic dyes, such as Victoria Blue grades, dye wool. A solution of 5 % tannin and 5 % sodium acetate precipitates basic dyes, and also dyes that contain sulfonic acid groups.

Table 6. Dyeing tests

Dye solution	Material (50 – 100 mg)	Procedure
5 cm^3 dye solution ∗ + 1 cm^3 15 % acetic acid	Wool and tannin-treated cotton	Boiling for a few minutes
5 cm^3 dye solution ∗ + 1 cm^3 5 % Na_2SO_4 solution	Cotton	Boiling for a few minutes, followed by a brief rinse with warm water
5 cm^3 dye solution ∗ + 1 cm^3 2.5 % soap solution	Cellulose acetate	Dyeing at 70 °C

∗ Containing 1 – 5 mg dye per 5 cm^3

Basic dyes are also precipitated with perchloric acid.

Acid dyes, *mordant dyes*, *acid metal-complex dyes*, and *leuco esters of vat dyes* dye wool, but not cotton or cellulose acetate; tannin-treated cotton is slightly dyed.

The acid metal-complex dyes are easily identified by testing an ashed sample for Cr, Co, Fe, Ni, or Cu. The leuco esters of vat dyes are detected by oxidation with sodium nitrite in sulfuric acid to the vat dyes, which can again be vatted in alkaline solution with hydrosulfite. If the tests for these two groups are negative, the wool is dyed from acid solution, followed by chroming. If there is a change of hue compared with the acid dyeing and if the washfastness improves, a chrome-developed or mordant dye is indicated; otherwise, it is an acid dye. The washing test is essential, because with many acid dyes, the hue is altered by oxidation, but washfastness is not improved.

Substantive or *direct dyes*, *diazotized dyes*, *developed dyes*, *aftertreated dyes,* and *reactive dyes* give positive results on cotton. Wool may be dyed partially; tannin-treated cotton may be dyed slightly or merely soiled; cellulose acetate is soiled.

To distinguish between an ordinary direct dye and a diazotized, developed, or coppered dye, substantive dyeing is followed by the appropriate posttreatment, and the untreated dyeing is compared with the posttreated dyeings with respect to fastness properties. If the posttreatment has improved fastness, an ordinary direct dye is probably not present.

Disperse dyes give positive results on cellulose acetate, but not on wool or cotton.

Disperse dyes can usually be extracted with ethyl acetate and are often soluble in acetone. If wool is stained in addition to cellulose acetate, dye mixtures for wool – polyester blends may be present, which in addition to disperse dyes usually contain 1 : 2 metal-complex dyes without sulfonic acid groups.

Sulfur and *vat dyes* dissolve slowly in water, often accompanied by a color change. The dyes giving positive results on wool contain reducing agents and are partly in the leuco form, in which they cannot be investigated. They must first be oxidized by blowing air into the vat solution. Alternatively they are separated from the reducing agent by dyeing and then investigated on the fiber. For this purpose, wool and cotton are dyed in the same vat solution at 60 °C.

For further differentiation, the cotton dyeing is tested for sulfur dyes with tin(II) chloride, hydrochloric acid, and lead acetate paper. Sulfur dyes and sulfur-vat dyes of the Hydron Blue type can be distinguished from vat dyes by treatment with 10 % sodium sulfide solution at 95 °C. This solution does not reduce vat dyes.

Dyes Insoluble in Water. The *alkali-soluble mordant dyes* give sparingly soluble color lakes in aqueous dispersion with solutions of metal salts when heated to ca. 70 °C, with simultaneous deepening of hue. Tests are made with chromium (III) sulfate, aluminum sulfate, and copper sulfate, or with iron(III) chloride. For confirmation, pre-chromed wool is dyed in an acetate bath, followed by fastness tests to washing and milling.

Color lakes of acid dyes leave metals, such as Ba, Sr, Ca, and Mn, in the ash. The dye is dissolved in sodium hydroxide solution and investigated in the same way as water-soluble dyes.

Alcohol-soluble dyes are usually anionic and often are salts of cyclohexylamine, dicyclohexylamine, 2-ethylhexylpropylamine, or di-o-tolylguanidine.

Color lakes from basic dyes are boiled with 10 % sodium hydroxide. The liberated dye bases are extracted into ether and back-extracted into dilute acetic acid as colored salts.

Dyes dissolving in sodium hydroxide and dithionite with a change in color, can be divided according to solubility: the *organic pigments* (azo or nitro dyes) are sparingly soluble or insoluble in ethanol, oils, and waxes; in this they differ from the *oil- and fat-soluble dyes* (Sudan dyes and Ceres dyes). The *alcohol- and ester-soluble dyes* in this group (Zapon-fast, Neozapon, Irisol-fast) are sometimes metal complexes, giving ash that contains such metals as Ni, Co, Fe, Mn, Cr, or Al.

Water-soluble vattable dyes, used for wool and cotton, are subdivided into *sulfur dyes* and *sulfur vat dyes* (e.g., Hydron Blue). These give a colored vat and can also be converted to the leuco form with sodium sulfide. Sulfur dyes are used with common salt, especially in the case of deep hues, and the Hydron Blue group is used with hydrosulfite.

If cotton dyed with dyes in this group, or the dye itself, is heated first with dilute hydrochloric acid and then with a solution of tin(II) chloride, it is possible to detect hydrogen sulfide with lead acetate paper.

Anthraquinone vat dyes in most cases give deep-colored vats, often differing considerably from the color of the dye.

Like the sulfur dyes, leuco *indigoid dyes* are only faintly colored; they do not vat with sodium sulfide, but require sodium hydrosulfite and sodium hydroxide.

4.1.3. Identification

A homogeneous dye can be identified by comparison with a known dye [106]. Identification requires comprehensive tables of reactions of dyes, dyeings, and prints [107]. Final identification requires a comparison with an authentic specimen [108–117].

An identification scheme for organic pigments and color lakes, mainly of the azo series, has been based on the crystal forms that precipitate from concentrated sulfuric acid [118].

4.1.4. Quantitative Determination

Organic dyes can be determined by a dyeing or comparison by chemical or spectroscopic methods. With dye mixtures, dyeing tests are preferred to chemical-analytical methods.

Dyeing Strength. In the determination of dyeing strength, the same conditions and the same material are used as for the standard to obtain two dyeings as nearly equal in depth as possible. If several types of fibers are to be dyed with the same dye, the color strength on each fiber must be determined. For dyes with low affinity (disperse dyes and leuco esters of vat dyes), the strength is determined at various depths of color, if a different dye is used for comparison; the result is affected by the affinity of the dye.

Elemental Analysis. Chemical methods for quantitative determination of a dye can be used if its molecular mass is known. Elemental analysis can be employed when the dye contains an element that is absent in other components of the commercial dye.

Reduction. Reduction methods can be used with most dyes except the quinophthalones, oxidation dyes, and thiazol dyes. Reduction is usually effected with titanium(III) chloride [*7705-07-9*] or titanium(III) sulfate [*10343-61-0*] [119–121]. Many azo dyes, especially if acid soluble, can be titrated with titanium(III) chloride directly; sodium tartrate, oxalate, or fluoride is often added [122]. Sometimes excess reducing agent is used and back-titrated with iron(III) salt [123]. If the end point is obscured by color, the solution can be lightened by adding ethanol or acetic acid; potentiometric titration is also used [124]. The following reductions take place:

$$R{-}N{=}NR' + 4\,Ti^{3+} + 4\,H^{+} \longrightarrow RNH_2 + R'NH_2 + 4\,Ti^{4+}$$

$$RNO_2 + 6\,Ti^{3+} + 6\,H^{+} \longrightarrow RNH_2 + 6\,Ti^{4+} + 2\,H_2O$$

Triarylmethane, azine, oxazine, and thiazine dyes are determined similarly, consuming 2 equivalents of titanium(III) for 1 mol of dye. Indigo is sulfonated before the determination.

Mixtures of dyes can be titrated successively with $TiCl_3$ in 30 % acetic acid [125]. To avoid nuclear chlorination, $Ti_2(SO_3)_4$ has

Table 7. Behavior of dyes on cellulose fibers

Dye class	Boiling with ∗ (a)	(b)	(c)	(d)	(e)	Comments
Direct dyes without treatment	+	–	–	+	(+)	wool. Tests should be performed to check if the color of the solution of both fibers is identical in concentrated H_2SO_4. If not, a mixture of acid dyes may be present.
Reactive dyes	(+)	–	–	(+)	(+)	Only small amounts of unreacted or hydrolyzed dye extracted.
Diazo dyes	–	–	(+)	–	+	
Direct dyes treated with Cu, Cr, or CH_2O	(+)	–	–	(+)	+	Test dyeing for CH_2O or ash for presence of Cu and Cr.
Insoluble azo dyes	–	(+)	+	–	–	All are destroyed by sodium hypochloride precipitated on the fiber rite and pyridine; yellow dyes are more resistant. Red dyes are resistant against sodium hypochlorite and hydrochloric acid.
Vat dyes	–	(+)	(+)	–	–	Many vat dyes are insoluble in all five solvents. Transform into vat and reoxidize on fiber.
Sulfur dyes	–	–	–	–	–	
Spin-dyeing pigments	–	–	(+)	–	–	Easily recognized on microscopic examination by pigmentation of fiber
Oxidation base dyes (aniline black)	–	–	–	–	–	
Mordant dyes	–	+	+	–	–	Give typical color changes with sodium hydroxide; some fluoresce. Basic dyes do not dissolve in ether.
Basic dyes	–	+	+	–	–	
Acid dyes	+	–	–	(+)	(+)	Used only as printing dyes on viscose rayon. Extracts dye wool (see Table 6).

∗ Abbreviations: a) water; b) alcohol; c) glacial acetic acid; d) 20 % aqueous ammonia; e) aqueous ammonia – alcohol solution; + dye is extracted; (+) dye is extracted in some instances or is partially extracted; – dye is not extracted.

Table 8. Behavior of dyes on wool fibers

Dye class	Boiling with ∗ (a)	(b)	(c)	(d)	Comments
Acid dyes ∗∗	(+)	–	–	+	
Chrome and mordant dyes	–	–	(+)	(+)	Ash of dyeings contains metal.
Acid complexed dyes	–	–	–	(+)	Ash of dyeings contains metal, which is found in redyeings with the stripped dye.
Direct dyes	(+)	–	–	(+)	
Basic dyes	(+)	+	+	(+)	
Vat dyes for wool	–	–	(+)	–	Strip with ammonia or with aqueous piperidine and dithionite.

∗ Abbreviations: (a) water; (b) alcohol; (c) glacial acetic acid; (d) 20 % aqueous ammonia; + dye is extracted; (+) dye is extracted in some instances or partially extracted; – dye is not extracted.
∗∗ Some are soluble in all solvents.

been recommended [126]. Vanadium(II) sulfate [*36907-42-3*] is also suitable for the determination of azo dyes [127]. Tin(II) chloride and sodium dithionite [*7775-14-6*] are now less important for this purpose and for the cleavage of azo dyes [128], [129].

Oxidation. Azo dyes (except insoluble ones without hydroxyl groups) and water-soluble triphenylmethane and anthraquinone dyes can be determined quantitatively by oxidation with chromic acid [*1333-82-0*]. The nitrogen present is liberated as the element and determined volumetrically [130], [131]. Various triphenylmethane dyes can be determined oxidimetrically with cerium(IV) sulfate [*17106-39-7*] [132],

[133]. This method is also suitable for the determination of benzidine derivatives [134].

Precipitation. Precipitation methods are also employed for quantitative determination. Anionic dyes are precipitated by cation-active compounds, in particular cetyltrimethylammonium bromide [135], [136]. A standardized solution of cetyltrimethylammonium bromide [*57-09-0*] in chloroform is used for titrimetric determination. Certain basic dyes can be determined with such anion-active compounds as alkyl sulfonates and sulfates. Basic dyes can also be precipitated from acid solution with phosphotungstic [*12067-99-1*] or silicotungstic acid [*12027-38-2*] [137].

4.1.5. Determination of Structure

In determining the structure of a dye, it is first necessary to identify the chemical class to which the dye belongs. Depending on its solubility, a dye can be dissolved in water, alcohol, or dimethylformamide for reduction, and the solubility can be increased by adding sodium hydroxide [107].

The behavior of *water-soluble dyes* toward reducing agents is determined by heating the dye in an aqueous solution with sodium hydroxide and dithionite.

Hue is not changed: quinophthalones, acridines, thiazoles, auramine, and azo dyes.

Hue changes and the dye is reconstituted on oxidation or another colored product is formed: anthraquinones, phthalocyanines, and azo dyes.

Dye fades and original shade returns upon exposure to oxidizing agents (air or persulfate): azines, oxazines, dioxazines, thiazines, and triphenylmethane dyes.

Dye fades and original hue does not return upon reoxidation: azo dyes, nitro dyes, and nitroso dyes.

The behavior of *water-insoluble dyes* toward reducing agents is determined by heating the dye in aqueous suspension with sodium hydroxide and dithionite:

Hue does not change: quinophthalones, dioxazine pigments, azo dyes, and phthalocyanines.

Hue changes and hue is reconstituted on re-oxidation: anthraquinones, indigoid dyes, benzo- and naphthoquinones, sulfur dyes, and phthalocyanines.

Dye is destroyed and not reconstituted upon re-oxidation: azo dyes, nitro dyes, and nitroso dyes.

Pyrazolone azo dyes can be recognized by a specific reduction technique. All dyes in this class exhibit red-violet to blue-violet hues during re-oxidation in air, when the reaction solution containing ammonia and zinc dust, with added 1-naphthol, is poured onto filter paper [107].

The color of the solution in sulfuric acid may indicate the chemical nature of the dye:

Nitro dyes	yellow to orange
Xanthene dyes	yellow to red
Triphenylmethane dyes	
basic, acid	yellow, orange, reddish-brown
mordant	red
Oxazine dyes	orange-brown, bordeaux, olive green
Azine dyes	green
Thiazine dyes	green
Dioxazine dyes	blue, green, almost black
Azo dyes	yellow to green
Anthraquinone dyes	yellow to green

It is very difficult to determine the structure of dyes, and there are no methods of general applicability. Elemental analysis can provide valuable information, but this requires a pure diluent-free specimen. The infrared spectrum can yield information on structure, but is probably used most often for comparison with known dyes. Proton and ^{13}C NMR give useful information on the aromatic-to-aliphatic ratio and the aromatic substitution. Field desorption and fast atom bombardment mass spectrometry provide data on molecular mass and structure.

Separation of Water-Soluble Additives. Before analysis, water-soluble dyes must be separated from water-soluble additives and the following methods can be used for this purpose.

1) If the dye is insoluble in acid, as with most substantive dyes, the dye can be precipitated from aqueous solution with hydrochloric acid and washed. After drying, the sodium salt can be obtained and purified by crystallization. Many polysulfonic acids do not separate from acid solution as free acids, but as acid salts. If only the chlorine content is to be

determined, water-soluble dyes can be salted out with sodium sulfate; sodium chloride is used to precipitate dyes for sulfur determination.

2) Pure sodium salts can be obtained by precipitation from aqueous solution with 2-tolylguanidine hydrochloride and by reconversion to the sodium salt by reaction with the calculated quantity of sodium methoxide in methanol solution.
3) If the dye can be precipitated from aqueous solution with sodium acetate, the precipitate is dried and the excess sodium acetate is extracted with absolute alcohol.
4) In the case of basic dyes, suspension agents (sodium sulfate, sodium chloride, dextrin) can be removed by extraction of the dye into ethanol, or it can be precipitated as the perchlorate.
5) Dispersing agents can be removed from disperse dyes by passage through a membrane filter or by dialysis.

Anthraquinone Dyes. For simple anthraquinone dyes (acid, basic, disperse, and mordant dyes) some methods permit systematic degradation (→ Anthraquinone Dyes and Intermediates). However, anthraquinone vat dyes require special techniques.

Azo Dyes. Azo dyes are usually cleaved into the individual starting amines by reduction, less often by oxidation. Tin(II) chloride [*7772-99-8*] was employed as early as 1888 [138–140]. Other suitable reductants are sodium hydrosulfite [*7775-14-6*], $Na_2S_2O_4$ [141] and titanium(III) chloride [*7705-07-9*]. Cleavage of azo dyes by catalytic hydrogenation appears to have found little application thus far for determination of structure [142], [143].

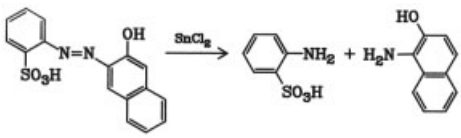

Polyazo dyes are cleaved similarly. In some cases, especially in the absence of free hydroxyl and amino groups, reduction stops at the hydrazo stage. At that time either the dye is reformed by oxidation or a benzidine or semidine rearrangement takes place in this acid environment. The goal is often achieved by reduction with titanium(III) compounds. Because many dyes contain acylamino, alkyl, or arylsulfonic or urea groups, it is sometimes advisable to hydrolyze before cleavage. With complicated polyazo dyes, cleavage may not yield information on the arrangement of the original structure. Synthesis may be required for a final answer.

Cleavage with Nitric Acid. If nitro groups are present, these are also reduced during reductive cleavage of azo dyes, or if they are in an ortho position to the azo linkage, they may form triazole rings. In this case cleavage with fuming nitric acid [*7697-37-2*] may be preferred for monoazo dyes.

$$R^1-N=N-R^2 \xrightarrow{HNO_3} [R^1-N=N]^+ + R^2-NO_2$$

A negatively substituted nucleus R^1 is obtained as a diazo compound, and a positively substituted nucleus R^2 is obtained as a nitro compound; further nitration of both nuclei is possible [114].

4.2. Investigation of Dyes on the Fiber

The dye class of a dyed article or a print is of great importance for the assessment of fastness, adjustments, processing, and finishing. The fiber analysis of a dyed article permits the exclusion of some dye classes. Identification of individual dyes requires expertise and the availability of an extensive collection of dyes for comparison. Identification of the class is more feasible than identification of an individual dye.

Tables have been available for a long time that give systematic analytical procedures for determining the chemical and dye class of dyes on fibers [144–146]. American methods have been published by the AATCC (American Association of Textile Chemists and Colorists) [147], [148]. Simple schemes are available for rapid determination of dyes on the fiber [117], [149–151]. A systematic scheme based on German experience has been published [152].

4.2.1. General Tests

Cellulose fibers burn evenly to produce a characteristic odor and a soft ash. Wool and silk can also

be recognized by burning (odor of burned horn); these fibers are soluble in boiling 5 % sodium hydroxide. Synthetic fibers can be differentiated by their solubility in solvents

If the warp and weft are different, the threads should be separated mechanically. Printed samples should be cut into areas with the same hue and treated separately. Dyed articles can be submitted to an ash test for metals (chromium, cobalt, copper, nickel, iron, tin, and aluminum) [145].

Treatment with solvents provides important information. The dye may be dissolved from the fiber completely, partly, or not at all (Tables 7 and 8). This method is especially suitable for cellulose and protein fibers [153].

When treated with sodium dithionite and aqueous sodium hydroxide, pyridine, or ammonia, many dyes are removed from the fiber or decomposed. Another reducing agent often used on the fiber is tin(II) chloride. Suitable oxidizing agents are sodium hypochlorite, permanganic acid (potassium permanganate + sulfuric acid), hydrochloric acid, and nitric acid [154].

Fluorescence can sometimes be employed for identification purposes [155]. Some dyes fluoresce on the fiber (e.g., Brilliant Sulfoflavine FF, Auramine O and G, Rhodamine B, G and 6G, Eosin, Sulforhodamine B and G, Fast Acid Pink B, Erio Fast Brilliant Flavine BBL, and Acid Red XG and XB); others fluoresce only in certain solvents or in vat form (Table 9).

Table 9. Fluorescence of dyes

Dye	Solvent	Fluorescence
Yellow dyes of the acridine class	98 % H_2SO_4	green
Indanthren Brilliant Orange GR	vat	red
Indanthren Red GG	98 % H_2SO_4	red
	pyridine	greenish yellow
Indanthren Red 5 GK	ether layer above vat	orange
Sulforhodamine BG ex.	98 % H_2SO_4	green
	pyridine	orange red
Indanthren Brilliand Pink R	benzene	yellow
Indanthren Brilliant Violet RK	pyridine	cyan
Indanthren Brilliant Violet BBK	pyridine	cyan
Indanthren Brilliant Violet 2 R	pyridine	orange red
Indanthren Deep Blue BO	benzene	reddish blue
Indanthren Deep Blue BOA	pyridine	red
Indanthren Marine Blue G	pyridine	reddish violet

4.2.2. Systematic Analysis

Detection of dyes on different fibers is covered in Tables 10, 11, 12; for details, see [152]. The dyed article is first boiled with distilled water to remove surface dyes for direct investigation. This is followed by repeated extraction with ammonia, glacial acetic acid, ether, or other solvents.

4.2.3. Special Tests

Certain methods are specifically applicable to naphthol and reactive dyes.

Naphtol AS Dyes. Combinations with yellow naphthols (Naphtol AS−G, −L3G, −L4G) remain almost unchanged by treatment with sodium hydroxide solution and dithionite, but are attacked by sodium hypochlorite and hydrochloric acid. Reductive cleavage of yellow naphthols is slow and incomplete.

Dyes from red naphthols are very resistant to sodium hypochlorite and hydrochloric acid, but they are bleached by sodium hypochlorite and pyridine. These combinations also resist sodium hydroxide and dithionite in the cold.

Dyes from the Fast Blue bases, Fast Violet bases, and Variamine Blue salts with arylamides of 2-hydroxy-3-naphthoic acid are bleached by sodium hypochlorite and pyridine and are converted to brown hues by sodium hypochlorite and hydrochloric acid. They are, however, resistant to sodium hydroxide and dithionite in the cold.

Color Reactions of Naphtol AS Combinations. With dyes from the same base, the color reactions differ only if the hydroxycarboxylic acid or acetocarboxylic acid forming the basis of the naphthol is different. Thus all red naphthols with the same base give similar color reactions. On the other hand, color reactions differ for combinations from the same base and arylamides of different hydroxycarboxylic acids. These can be distinguished from other naphthols by the yellow-green fluorescence, which is exhibited by solutions of red naphthols after cleavage with dithionite. Thus two groups of combinations are obtained: (1) from red naphthols and (2) from Naphtols AS−LB, AS−SG, AS−SR, AS−GR, AS−BT, AS−S, AS−CB, and yellow naphthols.

Table 10. Identification of dyes on cellulose * [152]

Procedure A: Pyridine is colored: sulfur, oxidation-base, mordant, acid, direct, and basic dyes; followed by test for the presence of sulfur

- Sulfur test positive: sulfur dyes (confirmation: reduction and reoxidation; blue dyeings turn violet when spotted with cold conc. nitric acid)
- Sulfur test negative: oxidation-base, acid, mordant, direct, and basic dyes; followed by testing for black dyeings with alkaline dithionite
  - Black colors turn brown and return to black on exposure to air: oxidation-base dyes; follow-up treatment with cold conc. sulfuric acid and dilution with water
    - Solution turns green: Aniline Black Test for Cr positive; turns black in potassium ferrocyanide test: positive for Fe
    - Color of solution unchanged: Diphenyl Black
  - If no black color develops or if reaction is negative: acid, mordant, direct, or basic dyes; followed by stripping with glacial acetic acid and dyeing of acrylics
    - Acrylics are dyed: basic dyes (confirmation: a) tannin/antimony potassium tartrate mordant, b) Katanol (Cassella) mordant)
    - Acrylics are not dyed: acid, mordant, and direct dyes; subsequent stripping with conc. ammonia or glacial acetic acid and dyeing reaction
      - Wool is colored more deeply: acid and mordant dyes; followed by tests for Cr, Fe, and Al
        - positive: mordant dyes
        - negative: acid dyes
      - Cotton is colored more deeply: direct dyes; followed by tests for Cu, formaldehyde, cationic products
        - positive: direct dyes treated with Cu or Cr salts, CH_2O, or cationic textile additives
        - negative: direct, diazo, or coupling dyes; differentiation should be attempted by comparison tests for fastness

Procedure B: Pyridine is not colored: sulfur, reactive, phthalocyanine, vat, and naphthol dyes; followed by test for sulfur

- Sulfur test positive: sulfur dyes (confirmation: reduction and reoxidation)
- Sulfur test negative: reactive, phthalocyanine, vat, and naphthol dyes; subsequent treatment: boiling with dimethylformamide
  - Test solution not colored: reactive dyes
  - Test solution colored: phthalocyanine, vat, naphthol dyes; subsequent treatment: boiling for 1 min with glacial acetic acid
    - Test solution not colored: phthalocyanine dyes (confirmation: treatment with cold conc. HNO_3 gives violet color; positive for Cu or Ni)
    - Test solution colored: vat and naphthol dyes; subsequent treatment: reduction and reoxidation
      - positive: vat dyes, differentiation between indigo and other blue vat dyes with HNO_3 and $SnCl_2$
      - negative: naphthol dyes (confirmation: fluorescence test; follow instructions for green naphthols)

* The microscopic examination for pigments, spin dyeings, and delustering is followed by treatment with cold pyridine.

Table 11. Paraffin tests for the identification of dyes on animal fibers [152].

- Paraffin colored: vat and naphthol dyes; reduction reoxidation
  - Hue changes and reappears: vat dyes
  - Test solution turns yellow; the test pattern does not change; if it does change, the original hue does not reappear on reoxidation: naphthol dyes
- Paraffin not colored: basic dyes, 1:1 metal complex, 1:2 metal complex, chrome, acid, direct, and fiber-reactive dyes; test for metals:
  - Positive for chromium and cobalt: 1:1 or 1:2 metal-complex or chrome dyes; extract dye with ammonia; acidify solution with hydrochloric acid and extract with ether:
    - Ether is strongly colored: 1:2 metal-complex dyes without sulfo groups
    - Ether is colored only weakly or not at all: 1:1 or 1:2 metal complex dyes with sulfo groups, chrome dyes
  - Negative for chromium and cobalt: basic, acid, direct, and reactive dyes; extraction with cold glacial acetic acid:
    - Glacial acetic acid is colored: basic dyes
    - Glacial acetic acid is not colored: acid, direct, and reactive dyes; extraction with cold pyridine:
      - Pyridine solution not colored: reactive dyes
      - Pyridine solution colored: acid and direct dyes; dyeing test:
        - Wool fiber dyed more strongly: acid dyes
        - Cotton fiber dyed more strongly: direct dyes

Table 12. Identification of dyes on synthetic fibers [152]

- Extract samples A* or B** with ethers
  - Ether layer is colored:
    - disperse, developed disperse, procinyl, naphthol, and vat dyes; fluorescence test on ether layer:
      - Positive: naphthol and developed disperse dyes: boiling in 50% aqueous pyridine
        - Dye precipitates and solution is strongly colored: developed disperse dye
        - Dye does not precipitate or solution is only faintly colored: naphthol dye
      - Negative: on PA: disperse or procinyl dyes on PE: disperse dyes boiling in glacial acetic acid, addition of ether and water:
        - Precipitate white: disperse dye
        - Precipitate colored: procinyl dye
    - disperse, metal complex, and chrome dyes; test for metals:
      - Negative: disperse dyes
      - Positive: metal-complex and chrome dyes; treatment with pyridine and ether:
        - Ether is colored: 1:2 metal-complex dyes without sulfo groups and chrome dyes
        - Ether is not colored: metal-complex dyes with sulfo groups
  - Ether layer is not colored: metal-complex, chrome, basic, acid, and vat dyes; test for metals:
    - Negative: Basic and acid dyes
      - on PA: acid dye
      - on PAN: boiling in glacial acetic acid, dyeing with acrylics
        - Acrylics are colored: basic dye
        - Acrylics are not colored: acid dye

* From polyester (PE) and polyamide (PA) fibers: caprolactam melt = sample A; vat dye test: residue of ether extraction after caprolactam melt is treated with alkaline dithionite, followed by reoxidation; changes in hue indicate presence of vat dyes.

** From polyacrylonitrile (PAN) fibers: stripping with glacial acetic acid = sample B.

With group 1 it is best to determine the base first. Combinations of red naphthols with a number of bases (Garnet GC, GBC, Corinth V, Brown, Blue and Violet) give characteristic color reactions.

With group 2, some conclusions can be drawn from the hue; investigation is simpler because only one or two arylamides of the relevant hydroxycarboxylic acids are on the market.

Naphtol AS combinations can be identified on the fiber by IR spectral comparison. The dye is removed from the fiber with dimethylformamide for the purpose. An NMR spectrum, if obtainable, helps to distinguish between closely related structures.

Reactive Dyes [156], [157]. The dyed sample is boiled with a nonionic surfactant, and tests are then conducted to establish the presence of reactive dyes.

The sample is boiled successively, with intermediate cold water rinses, with glacial acetic acid – ethanol (1 : 1), 1 % aqueous ammonia, dimethylformamide – water (1 : 1), and dimethylformamide. If no bleeding occurs during these treatments, a reactive dye is present. For identification of reactive dyes of the Remazol type (with vinyl sulfone as the reactive group), the sample can be boiled with a white wool thread in a solution containing 1 mL of concentrated sulfuric acid and 2 g of sodium sulfate per liter. Only Remazol dyes do not bleed in this test.

For further identification, the ash can be tested for Cr, Co, Cu, and Ni. For differentiation of blue reactive dyes from the anthraquinone and phthalocyanine series, the sample is treated at room temperature with a solution of 20 g of hydrosulfite and 2.5 g of sodium hydroxide per liter. Phthalocyanine derivatives exhibit blue, and anthraquinone derivatives exhibit brownish blue [158].

Other Special Tests. A simple method has been reported for the investigation of dyes on polyacrylonitrile [159]. Phthalocyanine derivatives can be identified on cellulose and distinguished from other types of dyes [160].

5. Chromatography

Theoretical principles and methods of chromatography are treated elsewhere in this Encyclopedia. In the following, only its special application to dyes is described. Chromatography of dyes is suitable for

1) Testing for purity and homogeneity
2) Separation of mixtures
3) Identification of individual dyes
4) Identification of unreacted intermediates and toxic impurities

The investigation of dyes employs column chromatography (CC), paper chromatography (PC), thin-layer chromatography (TLC), high-performance thin-layer chromatography (HPTLC), and high-performance liquid chromatography (HPLC). *Column chromatography* is suitable for separations on a preparative scale and is often better adapted than paper or thin-layer chromatography for the separation of trace substances. Trace substances in dilute solution can be concentrated by column chromatography and then eluted with a suitable solvent. *Paper chromatography and thin-layer chromatography* are simpler and more reproducible than column chromatography and produce chromatograms that can be stored.

5.1. Column Chromatography [161–165]

Only a few substances are suitable adsorbents for separating dye mixtures. Alumina [*1344-28-1*] and bleaching earths are used with organic media; mixtures of alumina with talc [*14807-96-6*] or with calcium salts are used with aqueous solutions [154, p. 56], [166–168].

Polyamide powder has been used with food and other dyes [169–174]. Gel permeation chromatography (GPC) [175], [176], and chromatography on silica gel [*7631-86-9*] [177], and anionic exchange resins [178] have also been used for food dyes.

Fluorescein and its derivatives [179] and oil-soluble dyes [180] can be separated on silica gel, and aminoanthraquinones can be separated on alumina [181], [182]. Vat dyes can be separated by high-temperature chromatography on alumina [183], and thioindigo derivatives can be separated on silica gel [184].

An alternative form of column chromatography employs polypropylene columns packed with silica-based sorbents [185].

5.2. Paper Chromatography [154, p. 52], [186–188]

Capillary analysis, a precursor of paper chromatography, has long been practiced in color laboratories. A filter paper with good absorption properties is dipped into a hot dye solution. The individual dyes rise at different rates; the components with the fastest rate of migration can be detected at the upper edge of the dye front after a short time (frontal analysis). The dye solution is supplied continuously, and zone separation can be achieved. In another procedure, water is dropped on a colored spot on a filter paper, resulting in separation in concentric zones [189].

Mobile Phases. Suitable mobile phases are critical for dye separation by paper chromatography (Table 13); these phases are sometimes chemically or physically unstable and change with time. Good reproducibility can only be achieved with mobile phases of about equal age. Accordingly, all mobile phases should be renewed regularly.

After separation of the dye from any substrate that is present [169–172], [203], [204], 0.25 – 2 % dye solutions are used for paper chromatography. Water is the preferred solvent; organic solvents are used only if water solubility is inadequate. The ascending method is preferred for determination of the most suitable mobile phases [186], [205]. The rate of migration is measured by the R_f value [186], which is only approximately constant but gives a guide to the order of separation [206–208]. It is often possible to improve separation by altering the pH [115], [119].

Paper chromatograms are evaluated in reflected and transmitted daylight and in UV light. Fluorescent substances are often present in commercial dyes as residues of intermediates.

It is never safe to conclude that a dye is homogeneous because its chromatogram shows no separation. The characterization of dyes on the chromatogram is supported by color tests [115], [119]. Paper chromatography can also be used in quantitative analysis.

The sensitivity of the method depends on the sharpness of the zone and the shade of the color spot. The limit of detection is between 0.5 and 1 % with yellow dyes and between 0.05 and 0.1 % with blue dyes. Fluorescent compounds can be detected at 1 part in 100 000 parts of the principal dye. In favorable cases, it is possible to detect as little as 1 ng (10^{-9} g).

5.3. Thin-Layer Chromatography [209–211]

For many years, dyes have been separated by adsorption on layers of alumina between two thick glass plates, the top one of which has a hole in the center [212]. This method, called disk chromatography, is a predecessor of thin-layer chromatography [213], [214].

In thin-layer chromatography (TLC) separation is accomplished on adsorbents spread on glass, aluminum, or plastic plates in a layer ca. 100 – 250 μm thick. Commercially prepared plates have the advantage of uniform thickness and they can be cut to the desired size. The separation is effected in glass chambers filled with mobile phase to a depth of ca. 1 cm (Table 14). The thin-layer plate is placed vertically in the chamber, and within 1 h the mobile phase usually rises far enough for separation. The general procedure resembles paper chromatography, but smaller specimens usually serve well.

To avoid interference during chromatography, the applied spots must first be dried in air. For separation of larger quantities, the solution is applied in a streak over the full width of the plate [209]. Plates with layer thicknesses up to 5 – 7 mm separate milligram quantities, which are useful for comparing IR spectra [210] or other spectroscopic techniques.

Silica gel [*7631-86-9*] is the substrate most widely used for dyes [215], [230], [233] – 235]. Gypsum [*13397-24-5*] is often added to improve adhesion; the product is designated Silica Gel G. Silica-based substrates have been extended to high-performance thin-layer chromatography (HPTLC) [237], [238], which provides higher resolution and speed. The coatings resemble packings used in high-performance liquid chromatography (HPLC).

Table 13. Mobile phases for paper chromatography of dyes

Dye class	Mobile phases	Reference
Acid dyes	ethyl acetate – pyridine – water (70 + 25 + 20), (35 + 25 + 20)	[190]
	trisodium citrate – 25 % aqueous ammonia – water (2 g + 20 + 80)	[191]
	1-butanol – acetic acid – water (50 + 10 + 20)	[190]
	ethyl acetate – acetic acid – water (80 + 20 + 20)	[190]
	1-butanol – ethanol – water – ammonia (100 + 20 + 44 + 1) or (20 + 10 + 10 + 10)	[192], [193]
	1-butanol – formic acid – water (100 + 24 + 36)	[190]
Direct dyes	isoamyl alcohol – pyridine – 25 % aqueous ammonia (10 + 13 + 10)	[194]
	1-butanol – pyridine – water (20 + 30 + 30)	[195]
	benzyl alcohol – dimethylformamide – water (30 + 20 + 20)	[196]
Reactive dyes	1-butanol – acetic acid – water (80 + 10 + 10) or (20 + 10 + 50)	[197], [198]
	disodium hydrogen phosphate – ammonia – water (2 g + 20 + 80)	[199]
Basic dyes	ethanol – ammonia – water (20 + 20 + 10);	[200]
	chromatography paper impregnated with lauryl alcohol	
	tetrahydrofuran – ammonia – water (10 + 4 + 7); separation using acetyl paper	[107, p. 162]
Disperse dyes	ethyl acetate – tetrahydrofuran – water (1 + 8 + 8); separation using acetyl paper	[201]
	tetrahydrofuran – ammonia – water (10 + 4 + 7); separation using acetyl paper	[201]
	chloroform – acetic acid – water (10 + 10 + 10); separation using paper impregnated with silicon	[196]
Vat dyes	tetraethylenepentamine – water – hydrosulfite (10 + 100 + 10)	[202]

Table 14. Mobile phases for thin-layer chromatography of dyes

Dye class	Mobile phases	Adsorbent	Reference
Acid dyes with sulfo groups	1-butanol – ethanol – water – acetic acid (60 + 10 + 20 + 0.5)	silica gel G ∗	[215]
	butyl acetate – pyridine – water (30 + 45 + 25)	silica gel G	[216]
	1-butanol – acetic acid – water (20 + 10 + 50) upper phase	silica gel G	[217]
Acid dyes with carboxyl groups	1-butanol – ethanol – 25 % aqueous ammonia – pyridine (40 + 10 + 30 + 20)	silica gel G	[218]
	toluene – acetic acid (65 + 35)	silica gel G	[219]
	ethyl acetate – pyridine – water (60 + 30 + 10)	silica gel G	[215]
Direct dyes	1-propanol – 25 % aqueous ammonia (60 + 30)	silica gel G	[220]
	1-amyl alcohol – pyridine – 25 % aqueous ammonia (10 + 10 + 10)	silica gel G	[220]
	1-butanol – acetone – water (50 + 50 + 30)	silica gel G	[221]
Reactive dyes	1-propanol – ethyl acetate – water (60 + 10 + 30)	silica gel G	[222]
	butyl acetate – pyridine – water (20 + 20 + 10)	silica gel G	[223]
Metal complexes 1 + 2, without sulfo groups	methanol – 25 % aqueous ammonia (95 + 5)	micropolyamide F 1700	[224]
	methanol – monoethanolamine (95 + 5)	micropolyamide F 1700	[201]
	toluene – acetic acid (70 + 30)	silica gel G	[224]
Basic dyes	chloroform – 2-butanone – formic acid (6 + 8 + 1)	silica gel G	[225]
	1-butanol – acetic acid – water (50 + 10 + 20)	silica gel G	[226]
	1-propanol – formic acid (80 + 20)	silica gel G	[227]
	carbon tetrachloride – methanol (4 + 1)	polyamide	[228]
Disperse dyes	toluene – acetone (20 + 1)	silica gel	[229]
	toluene – acetic acid (90 + 10)	silica gel G	[201]
	chloroform – methanol (95 + 5)	silica gel G	[226]
	toluene – ethanol – 25 % aqueous ammonia (850 + 150 + 15)	silica gel G	[201]
Solvent dyes	benzene	silica gel G	[214]
	toluene – acetic acid (90 + 10)	silica gel G	[201]
Organic pigments without sulfo groups	chloroform – xylene (30 + 10)	silica gel G	[230]
	dichloroethane	silica gel G	[231]
	benzene	silica gel G	[232]
	hexane – benzene – pyridine (5 + 20 + 15)	silica gel G	[226]

∗ Silica gel G: gypsum added for better adhesion

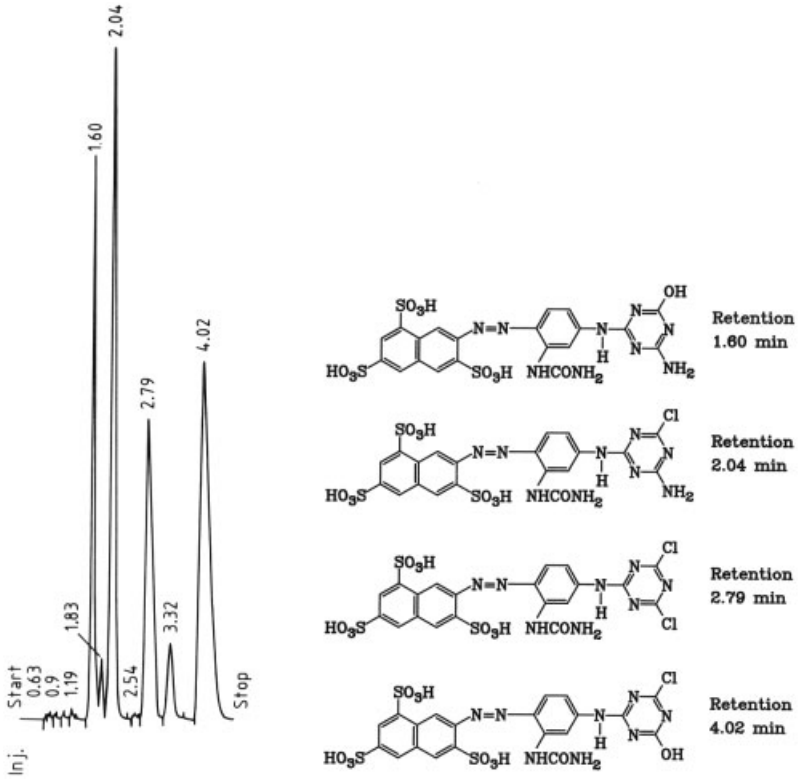

Figure 16. Separation of reactive dyes and their derivatives by HPLC
Packing material Hypersil ODS; Particle size 5 μm; Column 100×5 mm; Mobile phase 62.5 % acetonitrile in water; Flow rate 2 mL/min; Detector UV/visible at 420 nm; Temperature 40 °C; Water phase contains 0.5 g of cetyltrimethylammonium bromide per 100 mL as ion pairing agent.

The mobile phase can be selected by carrying out preliminary experiments using a series of solvent mixtures of increasing polarity (elutropic strength) [214]. Ideally, distinct, well separated zones should be obtained.

Selection can be simplified by reference to various publications [215], [218], [221], [222], [239–241]. Some mobile phases are listed in Table 14.

5.4. High-Performance Liquid Chromatography (HPLC)

High-pressure liquid chromatography, later called high-performance liquid chromatography, was developed in the early 1970s and was soon applied to dyes, where its value was greatly increased by the introduction of reverse-phase systems [242–244]. Developments utilizing stationary phases of particle sizes $< 5\ \mu$m have resulted in higher separation capability and speed [237], [238].

Mobile Phase. As with TLC, efficient separation depends on the appropriate selection of the mobile and stationary phases, but with the added facility of programmable mobile phase composition.

Reverse-phase HPLC, which uses surface modified supports and is the preferred method for polar compounds, utilizes binary (or sometimes ternary) eluents containing water, acetonitrile, tetrahydrofuran, or methanol.

Example of Dye Separation by HPLC. Figure 16 shows the chromatographic separation of a series of reactive dyes of similar structure. The area under each peak is proportional to the quantity of the component. The retention

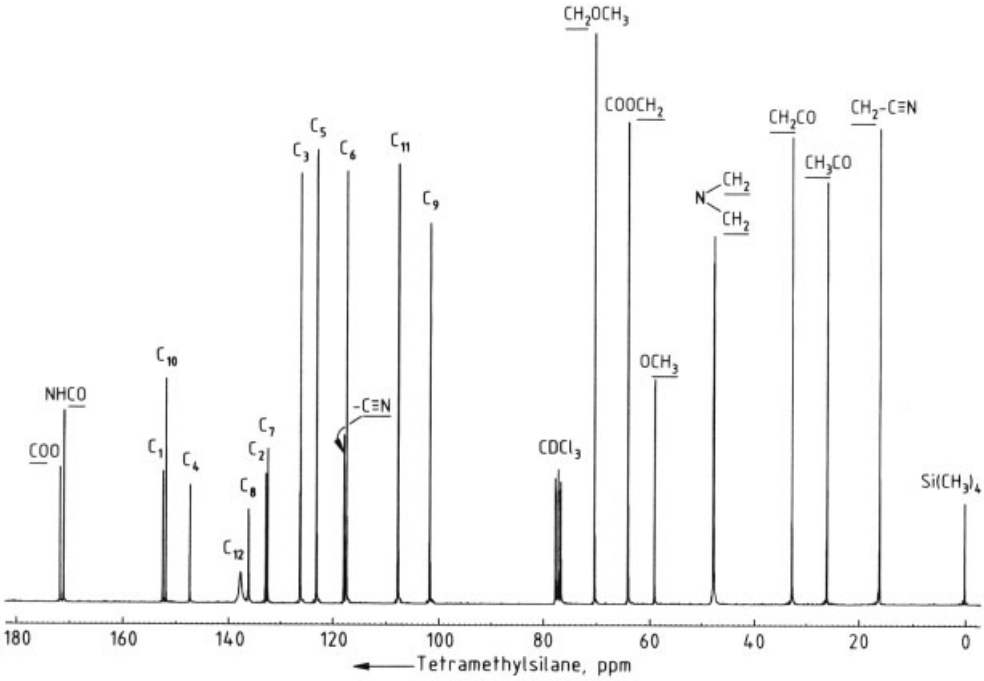

Figure 17. ^{13}C NMR spectrum of azo dye **26**, broadband proton-decoupled

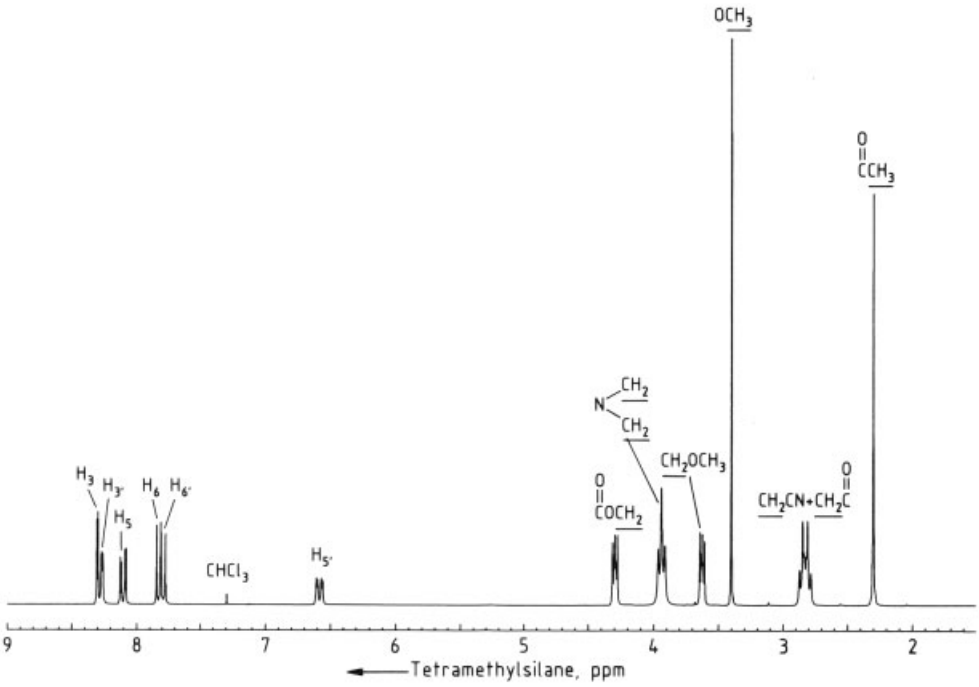

Figure 18. ^{1}H NMR spectrum of azo dye **26**

time is characteristic of the component in this reverse-phase HPLC analysis.

Allied techniques such as ion chromatography have also been of value in the detection and measurement of ionic diluents or impurities and in identifying the cations present. For the separation of dyes from impurities or additives, HPLC has been extended to preparative liquid chromatography with column diameters of 2.5 cm.

6. Spectroscopic Measurements

6.1. Nuclear Magnetic Resonance (NMR)

This technique is perhaps the single most useful means of structure determination. Modern instruments operating at a proton frequency of 90 – 400 MHz provide resolution capable of distinguishing between closely related compounds. Mass spectrometry can provide molecular mass and elemental composition.

Fourier transform NMR with ^{13}C and ^{19}F probes is valuable for investigating highly condensed aromatic or proton deficient molecules [245–247].

Example of Structure Elucidation. Disperse dye **26** was examined both by ^{1}H NMR and ^{13}C NMR. The spectra are shown in Figures 17 and 18. They are fully consistent with the structure of **26**; in fact, the ^{13}C NMR spectrum shows the peaks for each single carbon atom.

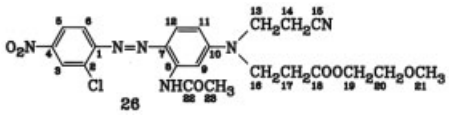

The identification of the dye can be confirmed by its synthesis or by an accurate measurement of the molecular mass by high-resolution mass spectrometry. Colors and intermediates containing other suitable nuclei such as N, P, or F can also be identified using appropriate probes.

6.2. Mass Spectrometry (MS)

In the past, mass spectrometry has been limited by the involatile nature of most dyes, but fast atom bombardment (FAB) ionization has produced high-quality mass spectra on a wide range of compounds previously regarded as intractable. Both positive and negative ions are produced, but the latter are preferred for structure elucidation because they yield more abundant fragment ions and fewer interfering ions at low mass [248], [249]. HPLC can be coupled with MS for identification of mixtures.

6.3. UV-Visible Spectrophotometry

This technique has always been of great value to manufacturers and users of dyes. Scanning instruments increased the scope for mixtures, but the real impact has been in the use of differential spectrometry, in which a microprocessor incorporated into the instrument can produce first, second, or even higher-order derivative spectra. Mixtures of colors can be measured and information provided on their exhaustion profiles (see also Chap. 3).

7. References

General References

1. B.I.O.S., F.I.A.T. reports: *German Dyestuffs and Dyestuffs Intermediates, including Manufacturing Processes, Plant Design, and Research Data,* 1945 – 1948.
2. H. E. Fierz-David, L. Blangey: *Fundamental Processes of Dye Chemistry,* Interscience, New York 1949.
3. K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* **vol. 1 – 8,** Academic Press, New York 1952 – 1974.
4. T. Vickerstaff: *The Physical Chemistry of Dyeing,* Oliver & Boyd, London 1954.
5. H. A. Lubs (ed.): *The Chemistry of Synthetic Dyes and Pigments,* Reinhold Publ. Co., New York 1955.
6. E. N. Abrahart: *Dyes and their Intermediates,* Arnold, London 1977.
7. P. F. Gordon, P. Gregory: *Organic Chemistry in Colour,* Springer, Berlin 1983.
8. H. Zollinger: *Color Chemistry, Syntheses, Properties and Applications of Organic Dyes and Pigments,* VCH Verlagsgesellschaft, Weinheim 1987.
9. Specialist journals for dyes and pigments: *Am. Dyest. Rep. DEFAZET Dtsch. Farben Z. Dyes Pigm. Farbe und Lack Int. Dyer Text.*

Printer Bleacher Finish. J. Oil Colour Chem. Assoc. J. Soc. Dyers Colour. Rev. Prog. Color. Relat. Top. (SDC) Text. Chem. Color.

Specific References

10. *Ullmann,* 3rd ed., **7,** 144 – 148.
11. W. H. Cliffe, *Chem. Ind. London* 1958, 1248 – 1255.
12. J. Houben. *Das Anthracen and die Anthrachinone,* Thieme Verlag, Leipzig 1929.
13. F. H. Moser, A. L. Thomas: *Phthalocyanine Compounds,* Reinhold Publ. Co., New York 1963.
14. G. Booth, *Chimia* **19** (1956) 201 – 207.
15. H. Vollman in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. 5, Academic Press, New York 1971.
16. Color Index 3rd ed. Society of Dyers and Colorists, Bradford 1971 (and supplements).
17. ITMA Catalogue, *1983-Esposizione Internazionale del Macchinario,* Tessile (Milan).
18. W. F. Beech: *Fibre Reactive Dyes,* Logos Press, London 1970.
19. Ed. T. C. Patton: *Pigment Handbook,* Wiley-Interscience, New York 1973
20. D. A. O'Sullivan, *Chem. Eng. News* **57** (1979) no. 9, 16.
21. U.S. International Trade Commission, *Production & Sales of Dyes & Pigments,* Annual Statistics.
22. *Jp. Chem. Annu.,* published annually.
23. United Nations (ed.): *Yearbook of Industrial Statistics,* **vol. II Commodity Production Data,** New York 1970 – 1979.
24. C. Boresch, *DEFAZET Dtsch. Farben Z.* **33** (1979) 207.
25. E. A. Clarke, R. Anliker in O. Hutzinger (ed.): *The Handbook of Environmental Chemistry,* vol. 3A, Springer, Berlin 1980.
26. ETAD Annual Report 1984: *Issued by the Ecological and Toxicological Association of the Dyestuffs Manufacturing Industry,* Basel 1985.
27. R. Anliker, E. A. Clarke, *J. Soc. Dyers Colour.* **98** (1982) 42.
28. J. Griffiths: *Colour and Constitution of Organic Molecules,* Academic Press, London 1976.
29. J. Fabian, H. Hartmann: *Light Absorption of Organic Colorants,* Springer, Berlin 1980.
30. J. Hu, P. Skrabal, H Zollinger, *Dyes Pigm.* **8** (1987) 189.
31. C. Graebe, K. Liebermann, *Ber. Dtsch. Chem. Ges.* **1** (1868) 106.
32. O. N. Witt, *Ber. Dtsch. Chem. Ges.* **9** (1876) 522.
33. W. Dilthey, R. Wizinger, *J. Prakt. Chem.* **118** (1927) 321.
34. R. Wizinger: *Organische Farbstoffe,* F. Dümmlers Verlag, Bonn 1933.
35. H. Kauffmann: *Die Auxochrome, Sammlung chem. und chem.-techn. Vorträge,* vol. 12, Stuttgart 1907.
36. W. König, *J. Prakt. Chem.* **112** (1926) 1.
37. A. von Baeyer, *Justus Liebigs Ann. Chem.* **354** (1907) 152.
38. R. Robinson, J. Allan, A. E. Oxford, J. C. Smith, *J. Chem. Soc.* **128** (1926) 401.
39. C. K. Ingold, E. H. Ingold, *J. Chem. Soc.* **128** (1926) 1310.
40. J. Griffiths, *Rev. Prog. Color. Relat. Top.* **14** (1984) 21.
41. J. Griffiths in J. Griffiths (ed.): *Developments in the Chemistry and Technology of Organic Dyes,* Blackwell Scient. Publ., Oxford 1984, p. 1.
42. S. Dähne, *Science (Washington, D.C.)* **199** (1978) 1163.
43. G. N. Lewis, M. Calvin, *Chem. Rev.* **25** (1939) 273.
44. A. Maccoll, *Quart. Rev. Chem. Soc.* **1** (1947) 16.
45. E. Heilbronner, H. Bock: *The HMO Model and its Application,* vol. 3, Verlag Chemie, Weinheim 1976.
46. A. L. Sklar, *J. Chem. Phys.* **5** (1937) 669.
47. L. Pauling, *Proc. Natl. Acad. Sci. (USA)* **25** (1939) 577.
48. N. S. Bayliss, *J. Chem. Phys.* **16** (1948) 287.
49. W. T. Simpson, *J. Chem. Phys.* **16** (1948) 1124.
50. J. R. Platt, *J. Chem. Phys.* **17** (1949) 484.
51. H. Kuhn, *J. Chem. Phys.* **16** (1948) 840; *Helv. Chim. Acta* **31** (1948) 1441.
52. H. Kuhn, *Helv. Chim. Acta* **32** (1949) 2247.
53. H. Kuhn, *Helv. Chim. Acta* **34** (1951) 2371.
54. H. Kuhn, *Fortschr. Chem. Org. Naturst.* **16** (1958) 170.
55. H. Kuhn, *Fortschr. Chem. Org. Naturst.* **17** (1959) 404.
56. M. Klessinger, *Chem. Unserer Zeit* **12** (1978) 1.
57. H.-D. Försterling, H. Kuhn: *Moleküle und Molekülanhäufungen, eine Einführung in die physikalische Chemie,* Springer Verlag, Berlin 1983, pp. 75 – 102.
58. I. Fleming: *Frontier Orbitals and Organic Chemical Reactions,* Wiley, Chichester 1978.

59. K. Fukui, *Angew. Chem.* **94** (1982) 852; *Angew. Chem. Int. Ed. Engl.* **21** (1982) 801; also published in *Science* **218,** 747.
60. R. Pariser, R. G. Parr, *J. Chem. Phys.* **21** (1953) 466, 767.
61. J. A. Pople, *Trans. Faraday Soc.* **49** (1959) 1375.
62. J. Griffiths, *Rev. Prog. Color. Relat. Top.* **11** (1981) 37.
63. J. Griffiths, *Dyes Pigm.* **3** (1982) 211.
64. J. Michl, E. W. Thulstrup: *Spectroscopy with Polarized Light,* VCH Publishers, Weinheim – New York 1987.
65. A. Grinvald, R. Hildesheim, I. C. Farber, L. Anglister, *Biophys. J.* **39** (1982) 301.
66. J. Griffiths, *Chem. Soc. Rev.* **1** (1972) 481.
67. E. Haselbach, E. Heilbronner, *Helv. Chim. Acta* **51** (1968) 16.
68. B. C. Burdett in E. Wyn-Jones, J. Gormally (eds.): "Aggregation Processes in solution," *Stud. Phys. Theor. Chem.* **26** (1983) 241.
69. V. Vitagliano in E. Wyn-Jones, J. Gormally, (eds.): "Aggregation Processes in solution," *Stud. Phys. Theor. Chem.* **26** (1983) 271.
70. T. Posner, *Ber. Dt. Chem. Ges.* **59** (1926) 1799.
71. H. von Eller, *Bull. Soc. Chim. Fr.* 1955, 1426, 1429, 1433, 1438, 1444.
72. M. Klessinger, W. Lüttke, *Tetrahedron* **19** (1963) Suppl. 2, 315.
73. M. Klessinger, *Tetrahedron* **22** (1966) 3355.
74. E. Wille, W. Lüttke, *Chem. Ber.* **106** (1973) 3240.
75. M. Klessinger, *Dyes Pigm.* **3** (1982) 235.
76. H. Hermann, W. Lüttke, *Chem. Ber.* **101** (1968) 1708, 1715.
77. E. Wille, W. Lüttke, *Angew. Chem.* **83** (1971) 853; *Angew. Chem. Int. Ed. Engl.* **10** (1971) 803.
78. L. Fitjer, W. Lüttke, *Chem. Ber.* **105** (1972) 919.
79. U. Luhmann, F. G. Wentz, B. Knierem, W. Lüttke, *Chem. Ber.* **111** (1978) 3233.
80. R. Grinter, E. Heilbronner, *Helv. Chim. Acta* **45** (1962) 2496.
81. K. P. Krishnan Namboodiri, S. Viswanathan, R. Ganesan, V. C. Jyothi Bhasu, *J. Comput. Chem.* **2** (1981) 392.
82. A. Kuboyama, S. Matsuzaki, H. Takagi, H. Arano, *Bull. Chem. Soc. Jpn.* **47** (1974) 1604.
83. J. Griffiths, K. Pender, *Dyes Pigm.* **2** (1981) 37.
84. A. C. Hardy, *J. Opt. Soc. Am.* **18** (1929) 96 – 117.
85. K. McLaren: *The Colour Science of Dyes and Pigments,* 2nd ed., Hilger, Bristol – Boston 1986.
86. CIE: *Colorimetry,* 2nd ed., CIE, Vienna, 1986.
87. J. Guild, *Philos. Trans. R. Soc. London Ser. A* **230** (1931) 149 – 187.
88. W. D. Wright, *Trans. Opt. Soc. London* **30** (1928 – 29) 141 – 164.
89. D. B. Judd, G. Wyszeki: *Color in Business, Science and Industry,* 3rd. ed., Wiley Interscience, New York 1975, p. 134.
90. R. S. Hunter, *J. Opt. Soc. Am.* **32** (1942) 509 – 538.
91. K. McLaren, *Col. Res. App.* **10** (1985) 124 – 125.
92. K. McLaren, *J. Soc. Dyers Colour.* **92** (1976) 317 – 326.
93. R. McDonald, J. Soc. Dyers Colour. **96** (1980) 372 – 376; 418 – 433; 486 – 497.
94. F. J. J. Clarke, R. McDonald, B. Rigg, *J. Soc. Dyers. Colour.* **100** (1984) 117 – 148;282.
95. P. Kubelka, F. Munk, *Zeit. Tech. Phys.* **12** (1931) 593 – 609.
96. A. Schuster, *Astrophys. J.* **21** (1905) 1 – 14.
97. H. R. Davidson, H. Hemmendinger, J. L. R. Landry, *J. Soc. Dyers Colour.* **79** (1963) 577 – 589.
98. J. V. Alderson, E. Atherton, A. N. Derbyshire, *J. Soc. Dyers Colour.* **77** (1961) 657 – 669.
99. D. R. Duncan, *Paint Research Station Bull. (UK)* **13** (1964) 132 – 150.
100. A. Brockes in *Tag. Ber. Int. Farbtagung Luzern,* Musterschmidt, Göttingen 1965, pp. 811 – 822.
101. L. Gall in *Colour 73,* Hilger, London 1973, pp. 153 – 178.
102. R. Kuehni: *Computer Colorant Formulation,* Heath, Lexington 1975. E. Allen in *Optical Radiation Measurements,* **vol. 2,** Academic Press, New York 1980, pp. 290 – 336. R. P. Best in *Golden Jubilee of Colour in the CIE,* Soc. Dye. Col. Bradford 1981, pp. 139 – 155.
103. K. Ventakamaran (ed.): *"The Analytical Chemistry of Synthetic Dyes,"* Wiley, New York 1977.
104. W. E. Mathewson, *Am. Dyest. Rep.* **22** (1933) 721 – 726, 745 – 746.
105. J. Šrámek: *"Applications of Paper Electrophoresis in Dyeing and Finishing,"* Research Institute of Textile Finishing, Dvur Králové n.L 1962. H. Tajiri, *J. Soc. Dyers Colour.* **74** (1958) 860.
106. BIOS Final Report no. 763.
107. H. Schweppe: "Qualitative Analyse von organischen Pigmenten," *FATIPEC Congr. 6th,* Verlag Chemie, Weinheim 1962, pp. 162 – 68.

108. H. C. Borghetty, K. J. Broden, *Am. Dyest. Rep.* **26** (1937) 589 – 599.
109. H. B. Bradley, D. A. Derrett-Smith, *J. Soc. Dyers Colour.* **56** (1940) 97 – 121.
110. M. Vajs, *Melliand Textilber.* **8** (1927) 611 – 619.
111. W. Bradley, *J. Soc. Dyers Colour.* **57** (1941) 9.
112. E. Herzog: *Reaktionstabelle der Küpenfarbstoffe und Indigosole,* Verlag B. Wepf & Cie., Basel 1933.
113. R. Bude, *Melliand Textilber.* **5** (1924) 602 – 606.
114. F. M. Rowe, C. Levin, *J. Soc. Dyers Colour.* **40** (1924) 218 – 231.
115. W. Bradley, *J. Soc. Dyers Colour.* **56** (1940) 296.
116. D. Haigh, *J. Soc. Dyers Colour.* **79** (1963) 242 – 256; **80** (1964) 479 – 486.
117. D. A. Derrett-Smith, J. Gray: *The Identification of Vat Dyers on Cellulosic Materials,* Pergamon Press, Oxford 1967.
118. J. J. Matiello: *Protective and Decorative Coatings,* 4th ed., J. Wiley & Sons, New York 1947, 2 vols., p. 126.
119. E. Knecht, E. Hibbert: *New Reduction Methods in Volumetric Analysis,* 2nd ed., Longmans, Green & Co., London 1925.
120. W. S. Calcott, F. L. English, *Ind. Eng. Chem.* **15** (1923) 1042 – 1044.
121. A. Stocker: *Zur Kenntnis der Analyse der Azofarbstoffe,* Dissertation, E.T.H. Zürich 1952.
122. J. Piccard, *Ber. Dtsch. Chem. Ges.* **42** (1909) 4331.
123. H. Salvaterra, *Chem. Ztg.* **38** (1914) 90.
124. J. M. Kolthoff, C. Robinson, *Recl. Trav. Chim. Pays Bas* **45** (1926) 169 – 176.
125. H. Holmes, *Am. Dyest. Rep.* **14** (1925) 415.
126. T. Callan, J. A. R. Henderson, *J. Soc. Chem. Ind. London* **41** (1922) 157 T – 161 T.
127. M. Matrka, Z. Sagner, *Chem. Prum.* **10** (1960) 474 – 475.
128. E. Grandmougin, E. Havas, *Chem. Ztg.* **36** (1912) 1167 – 1169.
129. W. Siegmund, *Monatsh. Chem.* **33** (1912) 1431 – 45.
130. F. M. Arshid et al., *J. Soc. Dyers Colour.* **69** (1953) 11 – 18.
131. M. Suzuki et al., *Tokyo Yakka Daigaku Kenkyu Nenpo* **11** (1961) 120 – 123.
132. M. Matrka, V. Hanousek, *Chem. Listy* **52** (1958) 755 – 757.
133. N. V. Rao, *Fresenius Z. Anal. Chem.* **188** (1962) 89.
134. M. Matrka et al., *Collect. Czech. Chem. Commun.* **35** (1970) 701 – 706.
135. E. Bene, I. Geczy, *Magy. Textiltech.* 1954, 370.
136. D. Sohar, P. Sohar, *Magy. Kem. Foly.* **69** (1963) 402 – 406.
137. Kozlov, *Zh. Prikl. Khim (Leningrad)* **9** (1936) 558; Payne, *J. Soc. Dyers Colour.* **50** (1934) 320.
138. P. Jacobson, F. Hönigsberger, *Ber. Dtsch. Chem. Ges.* **36** (1903) 4069, 4098.
139. A. Brunner: *Analyse der Azofarbstoffe,* Springer, Berlin 1929.
140. E. J. Cross, P. W. Carlene, *J. Soc. Dyers Colour.* **59** (1943) 191 – 192.
141. E. Grandmougin, *Ber. Dtsch. Chem. Ges.* **39** (1906) 3561, 3929; *J. Prakt Chem.* **76** (1907) 124.
142. K. W. Rosenmund, E. Pfannkuch, *Ber. Dtsch. Chem. Ges.* **56** (1923) 2258.
143. L. H. Andrews, A. Lowry, *J. Am. Chem. Soc.* **56** (1934) 1411.
144. A. G. Green: *The Analysis of Dyestuffs,* Griffin, London 1920. "Bestimmung der Farbstoffe in Substanz und auf der Faser nach Tabellen von A. Green" in Berl-Lunge: Chemisch-technische Untersuchungsmethoden, 8th ed., Springer, Berlin 1934, vol. 5, pp. 1356 – 1393.
145. E. Clayton: *Identification of Dyes on Textile Fibres,* 2nd ed., Society of Dyers and Colourists, Bradford 1963.
146. A. Bode, *Melliand Textilber.* **38** (1957) 289 – 296.
147. AATCC, New York Section, *Am. Dyest. Rep.* **47** (1958) 9 – 17; German Translation: *Z. Gesamte Textilind.* **60** (1958) 939 – 945, 1062 – 1064.
148. AATCC Forschungsausschuß RA 72, *Am. Dyest. Rep.* **57** (1968) 817 – 825; German Translation: L. Ostermeier, *Melliand Textilber.* **50** (1969) 961 – 969.
149. C. H. Giles et al., *J. Soc. Dyers Colour.* **78** (1962) 125 – 127.
150. R. V. R. Subramanian, K. S. Taraporewala, *J. Soc. Dyers Colour.* **88** (1972) 394 – 397.
151. W. Garner: *Textile Laboratory Manual,* 3rd. ed., American Elsevier, New York 1967.
152. G. Dierkes, H. Brocher, *Melliand Textilber.* **44** (1963) 387 – 391.
153. A. Seucker, BASF, personal communications.
154. A. Schaeffer: *Handbuch der Färberei,* Konradin-Verlag R. Kohlhammer, Stuttgart 1950, vol. IV, pp. 76 – 134.
155. J. Barker, F. A. Walker, *J. Oil Colour Chem. Assoc.* **23** (1940) 19 – 28.

156. A. Bode, *Melliand Textilber.* **40** (1959) 1304 – 06.
157. F. Jordinson, R. Lockwood, *J. Soc. Dyers Colour.* **78** (1962) 122; **84** (1968) 205; **84** (1970) 524; **88** (1972) 117; **90** (1974) 55.
158. O. Thumm, J. Benz, *Am. Dyest Rep.* **55** (1966) 15.
159. H. Wagner, J. Pflug, *Farben Rev.* no. 5, 1963, 62.
160. H. Wagner, J. Pflug, *Melliand Textilber.* **44** (1963) 281 – 285.
161. L. Zechmeister, L. Cholnoky: *Principles and Practice of Chromatography,* 2nd ed., Chapman & Hall, London 1941.
162. E. and M. Lederer: *Chromatography,* 2nd ed., Elsevier Publ., New York 1957.
163. H. Hrapia: *Einführung in die Chromatographie,* Wissenschaftliche Taschenbücher, Akademie-Verlag, Berlin 1965, vol. 30.
164. H. H. Strain: *Chromatographic Adsorption Analysis* Interscience Publ., New York 1954.
165. G. Hesse: *Chromatographisches Praktikum,* Akadem. Verlagsgesellschaft, Frankfurt/Main 1968.
166. P. Ruggli, P. Jensen, *Helv. Chim. Acta* **18** (1935) 624 – 643.
167. P. Ruggli, P. Jensen, *Helv. Chim. Acta* **19** (1936) 64.
168. R. H. Hanson, C. W. Gould, *Anal. Chem.* **23** (1951) 67 073.
169. G. Lehmann et al., *Z. Lebensm. Unters. Forsch.* **143** (1970) 191 – 195, 348 – 350, 418 – 420; **144** (1970) 32 – 34, 104 – 109.
170. G. Lehmann, P. Collet, H. G. Hahn, M. R. F. Ashworth, *J. Assoc. Off. Anal. Chem.* **53** (1970) 1182 – 1189.
171. G. Lehmann, P. Collet, *Arch. Pharm. (Weinheim, Ger.)* **303** (1970) 855 – 860.
172. G. Lehmann, U. Recktenwald, *Z. Lebensm. Unters. Forsch.* **146** (1971) 147 – 149.
173. R. A. Gilhooley, R. A. Hoodles, K. G. Pitman, J. Thomson, *J. Chromatogr.* **72** (1972) 325 – 331.
174. H. Schweppe: "Säulenchromatographische Trennung synthetischer Farbstoffe an Polyamidpulver," (Suppl. Chimia 1970), *5. Internat. Tagung über Trennmethoden: Säulenschromatographie,* 7. – 10.Oct. 1969, Lausanne pp. 56 – 59.
175. J. R. Parrish, *J. Chromatogr.* **33** (1968) 542 – 543.
176. A. L. Baetz, H. Diehl, *J. Chromatogr.* **34** (1968) 534.
177. Y. Yanuka, Y. Shalon, E. Weissenberg, J. Nir-Grosfeld, *Analyst (London)* **88** (1963) 872 – 876.
178. C. Graichen, J. C. Molitor, *J. Assoc. Agric. Chem.* **46** (1963) 1022 – 1029.
179. J. Santora, P. Raban, *Cesk. Farm.* **18** (1969) 522.
180. F. B. Jones, *J. Assoc. Off. Anal. Chem.* **49** (1966) 674.
181. N. R. Rao, K. H. Shah, K. Venkataraman, *Proc. Indian Acad. Sci. Sect. A* **34** (1951) 355 – 367.
182. D. Tomczyk, *Chem. Anal. (Warsaw)* **18** (1973) 365.
183. M. K. Unni, K. Venkataraman, *J. Sci. Ind. Res. Sect. B* **19** (1960), no. 9, 355 – 359.
184. M. Wyman, W. R. Brode, *J. Am. Chem. Soc.* **73** (1951) 1487 – 1493.
185. J. L. Love, *New Zealand J. Sci.* **27** (1984) no. 1, 113 – 116.
186. F. Cramer: *Papierchromatographie,* 5th ed., Verlag Chemie, Weinheim 1962.
187. K. Venkataraman: *Chemistry of Synthetic Dyes,* Academic Press Inc., New York 1952, vol. II, p. 1326.
188. I. M. Hais, K. Macek: *Handbuch der Papierchromatographie,* VEB G. Fischer Verlag, Jena 1958, vol. I: Grundlagen und Technik, p. 677; vol. II: Bibliographie, p. 473.
189. D. Paterson, *J. Soc. Dyers Colour.* **9** (1893) 110.
190. *Ullmann,* 3rd ed., **7,** 212.
191. H. Thaler, G. Sommer, *Z. Lebensm. Unters. Forsch.* **97** (1953) 345 – 365.
192. C. Graichen, *J. Assoc. Off. Agric. Chem.* **34** (1951) 795.
193. J. C. Brown, *J. Soc. Dyers Colour.* **85** (1969) 137 – 146.
194. J. Sramek, *J. Chromatogr.* **15** (1964) 57 – 64.
195. E. G. Kiel, G. H. A. Kuypers, *Tex* **22** (1963) 779 – 781.
196. J. C. Brown, *J. Soc. Dyers Colour.* **76** (1960) 536 – 542.
197. L. A. Kovzin, A. A. Charcharov, *Zavod. Lab.* **29** (1963) 542 – 543; *Fresenius Z. Anal. Chem.* **206** (1964) 159.
198. J. Perkavec, M. Perpar, *Kem. Ind.* **13** (1964) 404.
199. J. Sramek, *Textil* **17** (1962) 390 – 391.
200. J. Gasparic, M. Matrka, *Collect. Czech. Chem. Commun.* **24** (1959) 1943 – 1949.
201. H. Schweppe, unpublished results.
202. E. Klingsberg, *J. Soc. Dyers Colour.* **70** (1954) 563 – 567.
203. A. Fouassin, *Rev. Ferment. Ind. Aliment.* **7** (1952) 195 – 219.

204. H. Thaler, G. Sommer, *Z. Lebensm. Unters. Forsch.* **97** (1953) 441 – 445.
205. M. L. Puttemans et al., *J. Assoc. Off. Anal. Chem.* **65** (1982) no. 3, 730 – 744.
206. W. Ruf, *Z. Lebensm. Unters. Forsch.* **94,** (1952) 190.
207. F. Schwerdtfeger, *Chem. Tech. (Leipzig)* **6,** (1954) 192 – 196.
208. G. Zimmermann, *Fresenius Z. Anal. Chem.* **138** (1953) 321 – 332.
209. E. Stahl: *Dünnschicht-Chromatographie, ein Laboratoriumshandbuch,* Springer, Berlin – Göttingen – Heidelberg, 1st ed. 1962, 2nd ed. 1967.
210. K. Randerath: *Dünnschicht-Chromatographie,* Verlag Chemie, Weinheim, 1st ed. 1962, 2nd ed. 1965.
211. J. M. Bobbit: *Thin-Layer Chromatography,* Reinhold Publ. Co., New York 1963.
212. T. I. Williams: *Introduction to Chromatography,* Blackie & Son, Glasgow 1947.
213. J. G. Kirchner, J. M. Miller, G. J. Keller, *Anal. Chem.* **23** (1951) 420, 428.
214. E. Stahl, *Chem. Ztg.* **82** (1958) 323.
215. G. H. Rettie, C. G. Haynes, *J. Soc. Dyers Colour.* **80** (1964) 629.
216. L. Meckel, H. Milster, U. Krause, *Text. Prax.* **16** (1961) 1032.
217. S. Logar, J. Perkavec, M. Perpar, *Mikrochim. Acta* 1964, 712.
218. J. C. Brown, *J. Soc. Dyers Colour.* **80** (1964) 185.
219. M. B. Naff, A. S. Naff, *J. Chem. Educ.* **40** (1963) 534.
220. J. Gasparic, A. Cee, *J. Chromatogr.* **14** (1964) 484.
221. C. D. Sweeny, *Am. Dyest. Rep.* **61** (1972) 70 – 78.
222. J. Perkavec, M. Perpar, *Fresenius Z. Anal Chem.* **206** (1964) 356.
223. A. Cee, J. Gasparic, *Collect. Czech. Chem. Commum.* **33** (1968) 1091.
224. H. Pfitzner, H. Schweppe, *Fresenius Z. Anal. Chem.* **268** (1974) 337.
225. A. M. Arsov, B. K. Mesrob, A. B. Gateva, *J. Chromatogr.* **81** (1973) 181.
226. H. Schweppe in *Dünnschicht-Chromatographie, ein Laboratoriumshandbuch,* Springer, Berlin – Göttingen – Heidelberg, 2nd. ed. 1967, pp. 583–601.
227. A. Stier, W. Specht, *Naturwissenschaften* **50** (1963) 549.
228. R. Takeshita, N. Itoh, Y. Sakagami, *J. Chromatogr.* **57** (1971) 437.
229. J. C. Brown, *J. Soc. Dyers Colour.* **85** (1969) 137.
230. A. Mc Clure, J. Thomson, J. Tannahill, *J. Oil Colour Chem. Assoc.* **51** (1968) 580.
231. S. Fujii, M. Kamikura, *Shokuhin Eiseigaku Zasshi* **4** (1963) 135.
232. J. Gasparic, *J.Chromatogr.* **66** (1972) 179.
233. F. Schlegelmilch, W. Kuss, *Defazet Aktuell* **27** (1973) 484.
234. N. A. Fuller, *Forensic. Sci. Int.* **27** (1985) no. 3, 189 – 204.
235. J. Barek et. al., *Microchem. J.* **29** (1984) no. 3, 311 – 317.
236. J. Barek et al., *Microchem. J.* **28** (1983) no. 4, 459 – 463.
237. Proceedings 2nd Int. Symposium on Instrum. HPTLC, Interlaken, Switzerland, May 2 – 6, 1982.
238. T. A. Perenich et al., *Natl. Tech. Conf.* AATCC 1982, 177 – 184.
239. A. Cee, J. Gasparic, *Collect. Czech. Chem. Commun.* **33** (1968) 1091.
240. K. C. Walker, M. Beroza, *J. Assoc. Off. Agric. Chem.* **46** (1963) 250.
241. A. M. Arsov, B. K. Mesrob, A. B. Gateva, *J. Chromatogr.* **81** (1973) 181.
242. F. E. Lancaster, J. F. Lawrence *J. Assoc. Off. Anal. Chem.* **66** (1983) no. 6, 1424 – 1428.
243. C. Prandi, T. Venturini, *J. Chromatogr. Sci.* **19** (1981) no. 6, 308 – 314.
244. J. Park, J. Short, *Rev. Prog. Color. Relat. Top.* **12** (1982) 1 – 9.
245. A. Lycka et al., *Org. Magn. Res.* **15** (1981) no. 4, 390 – 391.
246. A. Lycka, *Collect. Czech Chem. Commun.* **48** (1983) 3104 – 3111.
247. A. M. Chippendale, *J. Chem. Soc. Perkin Trans. II* 1983, 1357 – 1359.
248. P. J. Gale et al., *Anal. Chem.* **58** (1986) 1070 – 1076.
249. C. Creaser, *Org. Mass Spectrom.* **19** (1984) 311 – 315.

Dyes, Reactive → Reactive Dyes
Dysprosium → Rare-Earth Elements
Earthenware → Whitewares

Electrochemical and Chemical Deposition

LAWRENCE J. DURNEY, Durney Associates, Inc., North Caldwell, New Jersey 07006, United States

1. Introduction . . . 2
2. Glossary of Terms . . . 3
3. Protection Against Corrosion . . . 10
4. The Plating Process . . . 11
4.1. Design for Plating . . . 12
4.2. Control . . . 12
4.3. Impurities . . . 13
4.4. Inspection . . . 14
5. Substrate Preparation . . . 14
5.1. The Nature of the Adhesive Bond 14
5.2. The Part Preparation Cycle . . . 15
5.3. Preparation of Specific Substrates 16
5.3.1. Steel . . . 16
5.3.2. Cast and Malleable Iron . . . 17
5.3.3. Copper and Copper Alloys . . . 17
5.3.4. Zinc . . . 17
5.3.5. Aluminum . . . 18
5.3.6. Magnesium . . . 19
5.3.7. Stainless Steel . . . 20
5.3.8. Nickel . . . 21
5.3.9. Nonmetallic Substrates . . . 21
6. Equipment . . . 22
6.1. Tanks . . . 22
6.2. Power Supply . . . 23
6.3. Heating Equipment . . . 23
6.4. Ventilation . . . 23
6.5. Filters . . . 24
6.6. Agitation . . . 24
6.7. Water Supply . . . 24
6.8. Racks . . . 24
6.9. Waste Disposal . . . 25
6.10. Site Location . . . 26
6.11. Layout . . . 26
7. Stripping of Deposits . . . 26
8. Specific Deposits and Plating Solutions . . . 27
8.1. Aluminum Deposits . . . 27
8.2. Antimony and Arsenic Deposits . . 27
8.3. Cadmium Deposits . . . 27
8.4. Chromium Deposits . . . 28
8.5. Cobalt Deposits . . . 29
8.6. Copper Deposits . . . 29
8.7. Gallium Deposits . . . 31
8.8. Germanium Deposits . . . 31
8.9. Gold Deposits . . . 31
8.10. Indium Deposits . . . 33
8.11. Iron Deposits . . . 33
8.12. Lead Deposits . . . 34
8.13. Manganese Deposits . . . 34
8.14. Nickel Deposits . . . 34
8.15. Niobium Deposits . . . 37
8.16. Rhodium Deposits . . . 37
8.17. Silver Deposits . . . 37
8.18. Tin Deposits . . . 38
8.19. Zinc Deposits . . . 40
9. Specific Alloy Deposits . . . 43
9.1. Chromium Alloys . . . 43
9.2. Copper Alloys – Brass Plating . . . 43
9.3. Gold Alloys . . . 45
9.4. Lead Alloys . . . 45
9.5. Nickel Alloys . . . 45
9.6. Tin Alloys . . . 46
9.7. Zinc Alloys and Other Alloys . . . 47
10. Nonelectrolytic Deposits . . . 47
10.1. Immersion Deposits . . . 47
10.2. Noncatalytic Chemically Reduced Coatings . . . 48
10.3. Catalytic Electroless Plating . . . 48
11. Chromating . . . 52
12. Oxide Coatings . . . 52
12.1. Immersion . . . 52
12.2. Electrolytic Oxidation . . . 53
12.2.1. Aluminum . . . 53
12.2.2. Magnesium . . . 55
12.2.3. Titanium . . . 56
12.3. Colors of Oxide Layers . . . 56
13. Testing of Electrodeposited Coatings . . . 56
13.1. Quality Control . . . 56
13.2. Thickness Testing . . . 57
13.2.1. Destructive Testing . . . 57
13.2.2. Nondestructive Testing . . . 58
13.3. Adhesion Tests . . . 59
13.4. Corrosion Testing . . . 59
13.5. Other Tests . . . 60
14. References . . . 60

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_125

1. Introduction

Throughout history, mankind has endeavored to change the appearance of base metals by applying metallic coatings of various types, either for decoration, or to enhance their value. There seems to be evidence that the Egyptians used a crude galvanic process for gilding and that some of the gold bars extorted by the Spanish from the Incas as ransom were actually lead gilded by the amalgam process; the application of gold leaf to various substrates has been an accepted method of decoration from the Middle Ages to the present.

Since early times, too, the ravages of corrosion have been evident, and efforts were made to alleviate this destructive process by the application of hot-dip coatings or cladding of metals by impacting or heating the parts with powdered metal as in Sheradizing (→ Metals, Surface Treatment).

For the most part, although electroplating became common practice as soon as practical direct current generators were developed, it remained more art than science until true scientific research was undertaken in the mid 1930s. The impact of World War II produced a tremendous surge in this research, which has continued to the present day.

As a result, great progress has been made in understanding the fundamentals underlying the plating process. This has brought with it the development of improved solutions with controlled plating characteristics so that the physical properties of the deposits may be controlled; new types of solutions and plated alloys; plated composites; and many nonaqueous and nonelectrolytic methods of depositing desired coatings.

An important consequence of the new mastery of plating theory and technique has been the general realization that the plated coatings are not just a cosmetic treatment applied as the final step in manufacturing, but an engineering tool that can be used to improve the performance of the product, reduce cost, or enable the product to resist particularly severe environments. The selection of the proper coating has therefore become part of the engineering design process.

The plating process is an electrochemical process that passes an electrical current through a solution containing metallic ions. These ions are reduced and deposited on the cathode as a metallic coating. If the anode is made of the metal being deposited, it dissolves to replenish the metallic ions in solution. If it is an insoluble anode, periodic additions of metal salts must be made to the solution to maintain the metal content.

The process is controlled in a general way by the usual electrical and electrochemical equations: Ohm's Law, Faraday's Law, and the equilibrium equation for the standard single electrode potential. The flow of current in the tank responds to the requirements of Ohm's Law in that the flow of current is directly proportional to the applied voltage, and inversely proportional to the resistance of the system. However, this is not a linear relationship. The system is a dynamic one, and the internal resistance in the tank is affected by many factors.

The quantity of metal deposited follows the principle of Faradays's Law in that the mass of metal deposited is directly proportional to the current flow, the time and the relative atomic mass of the metal, and inversely proportional to the oxidation state of the metal in solution. Here too, the relationship is not linear. It may be affected by competing reactions whose share of the current flow may depend on the voltage and the rate of current flow.

The single electrode potentials determine the minimum voltage required for plating, and the distribution of the available current among the various competing reactions.

2. Glossary of Terms

A large spread exists in the educational levels of even reasonably well-trained electroplating professionals. There is therefore considerable variation in the definition of the terms used. These definitions are frequently at variance with the definition used by electrochemists. The definitions that follow are those that would be accepted by the majority of knowledgeable, practicing electroplaters.

Rack Plating. The plating process in which the parts to be plated are fastened to fixtures that are carried through the plating process either manually or by some automatic mechanism. The parts are held to the fixtures by pressure contacts

that also serve to carry the required current to the parts.

Barrel Plating. The plating process in which the parts are loaded into a plastic perforated container, generally hexagonal in shape, which is then carried through the process either manually or by some automatic mechanism (Fig. 1). While in each of the solutions, the barrel is rotated so that the parts slowly tumble in the barrel and are all exposed to the plating action. Current is supplied through various types of contacts into the mass and transferred from piece to piece by the contact made during rotation. This process is especially useful for small screw machine parts and stampings.

Brush Plating. This is also sometimes called *spot plating*. Not all plating has to be carried out by total immerson. Techniques have been developed for selective plating using portable equipment and anodes surrounded by absorbent material that do not require immersion of the part. The anode is saturated with solution and rubbed over the surface of the area to be plated. Because of the movement and the slight abrasive action of the absorbent material, very high localized currents may be used. Special plating solutions are required. The entire package of equipment and solutions, as well as instruction in operation, are supplied as proprietary units. The method is not generally recommended for production but is used for repair of worn or defective parts, e.g., in situ repair of worn printing rolls or ship propeller shafts, and the plating of parts too large for immersion techniques or objects that must be plated in place; e.g., several large statues in Washington, D.C., were refurbished by gold brush plating.

Pulse Plating. The current used for plating does not always have to be applied as a continuous flow. In pulse plating, the current is applied in short bursts of high intensity followed by a period in which no current is applied. The cycles are described in terms of the ratio of On time to Off time, the *duty cycle,* which is usually expressed as a percentage, and the *frequency,* which determines the time length of each cycle. By varying the duty cycle and the frequency, desirable alterations to the physical characteristics of the deposits may often be obtained. Although the current is on for only a percentage of the total time, the current densities during the On time are sufficiently high that the average current density is usually close to or even higher than the current density recommended for regular direct current. Not all metals can be deposited using pulse plating since some metals do not tolerate variations in the applied current (*ripple*) above 5 to 10 %.

Immersion Deposits (see Section 10.1). Deposits that form on a surface by a simple replacement reaction. If a metal is immersed in a solution that contains ions of a less reactive metal, the metal immersed dissolves, in the process acting as a reducing agent for the metal ions in solution, which deposit on the surface of the immersed metal, e.g.,

$$Cu^{2+} + Fe \longrightarrow Cu^{0} + Fe^{2+}$$

The process continues until a completely continuous coating has been formed on the substrate, after which no further action occurs. The immersion coating is usually of poor structure and lacks adhesion to the substrate.

Conditions can be adjusted in some solutions so that coatings with desirable characteristics form, and such solutions are used for special purposes, e.g., a copper sulfate – sulfuric acid solution for producing a *copper liquor finish* on steel wire to act as a lubricant in drawing (see page 47). Many plating solutions form such coatings on some metals that are to be plated, with the result that the subsequently plated layer has no adhesion. These metals must be preplated with a thin layer of a nonreactive metal. The solutions used to apply these preliminary coatings are generally referred to as strike solutions.

Strike Solution. A strike solution is one that has been designed to plate thin deposits at relatively high current densities and without forming immersion deposits. The solution is used only to produce the initial thin layer of plating to completely cover the substrate. The metal used may be the same as the subsequent plated deposit, or may be different than, but compatible with, the subsequent plated layer. Typical examples: a cyanide copper strike on zinc-base die castings or on steel before bright acid copper, a silver strike on copper before silver plate, a cadmium or tin strike on cast iron before cyanide zinc plating.

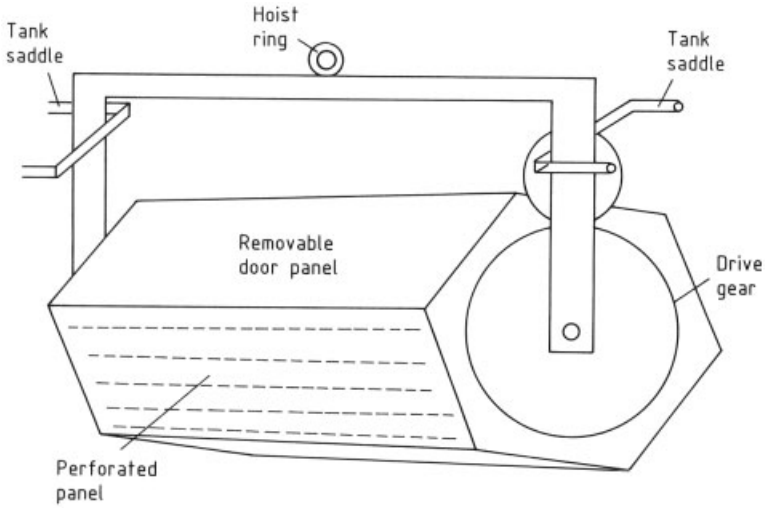

Figure 1. Schematic drawing of a plating barrel

Current Density (C.D.). The rate of current flow at any point on the part or parts being plated. This is usually expressed as amperes per square decimeter (A/dm^2) or amperes per square foot (A/ft^2). The conversion factor A/dm^2 to A/ft^2 is 9.29. Because current density is a relatively inexact measurement, however, it is common practice to use a conversion factor of 10.

Since measurement of the current density on small areas is difficult or impossible, current density is usually given as an average for the area being plated. There are now available meters that can be placed in the solution near the parts being plated to determine the current density over a relatively small area. While these meters do not measure the current density at a point, they provide more exact figures than those obtainable by dividing the total current by the total area being plated. If the parts are sufficiently large, current in the high and low current density areas can be determined.

Current Distribution. Two terms are used in the discussion of current distribution: *primary current distribution* and *secondary current distribution,* and some confusion exists as to the exact meaning of each term.

The original research into current distribution used nonconductive models of parts that were charged with static electricity. The potential at various points on the surface was then determined and related to the current that would flow if this potential existed in a plating cell. The resulting distribution of current was considered to be the primary current distribution. The secondary current distribution was then considered to be the current distribution that occurred during the actual plating operation. This included the effects of tank configuration, including rack placement, tank walls, anode placement, and the effects of polarization at both the anode and the cathode. Some electroplaters still use these definitions.

As instrumentation improved, and computers became more sophisticated, it became possible to model complete tank configurations and estimate the potentials within the tank. Brightener systems also became more complex and had greater influence on current flow. Researchers began to define primary current distribution as that which would exist in the absence of current flow as a result of the physical arrangement in the plating cell, and the secondary current distribution as that which would exist during actual plating including the effects of polarization and various brightener influences.

Most of the available information is empirical and unquantified. Further, the dynamic changes between primary and secondary current distribution vary with the brightener system, the solution composition, and operating conditions. However, an experienced plater can markedly improve the current distribution on even extremely

complicated pieces by the judicious use of supplementary anodes, shields, and rack positioning.

Deposition Potential. This is another case where the average plater is at odds with the electrochemist. For the average plater, the deposition potential is defined as the electrical potential that is just sufficient to initiate electrodeposition of the metal in question. This potential is in part a function of the solution composition, particularly the ionic concentration of the metal at the interface, and therefore is a dynamic quantity changing in response to changes in temperature, plating rate, agitation, and degree of polarization. The term overvoltage is frequently used interchangeably for deposition potential.

Hydrogen Overvoltage. Generally this term is used for the deposition potential of hydrogen at the cathode. It varies with the substrate making initial deposition on some metals difficult from some solutions. For example, the hydrogen overvoltage on graphite is extremely low. In a cyanide zinc plating bath, the deposition potential of zinc is ca. one volt. Cast iron has graphite exposed at many points on the surface. It is therefore almost impossible to plate cast iron parts in a cyanide zinc solution since hydrogen deposits preferentially. If, however, the cast iron is given a preliminary strike of cadmium or copper to cover the graphite, zinc plating proceeds normally. Deposition of hydrogen with the metal being plated affects efficiency and may produce hydrogen embrittlement of the substrate.

Hydrogen Embrittlement. Hydrogen that is codeposited with the metal being plated or that is formed during acid pickling or cleaning procedures may be absorbed by the substrate. When this occurs, the hydrogen may eventually collect at crystal interfaces at points of stress, develop extremely high pressures, and cause brittle failure of the material. A number of different mechanisms have been postulated for this occurrence, but no one mechanism has been firmly established as the absolute cause. Indeed, it is possible that under varying conditions, several different mechanisms may be operating concurrently. However, the problem is particularly severe with high strength steels. Research work to date indicates that, before plating, susceptible substrates should be stress relieved at temperatures close to the last heat-treating temperature for periods up to 24 h, the plating cycle should be designed to offer the minimum possible exposure to hydrogen, and the parts should be subjected to a postplating baking operation at a temperature approximately 28 °C below the temperature of the final heat treating step for periods of up to 24 h.

Polarization Effects. Polarization is the change that takes place in the system as a result of the current flow. Films will generally form at both the anode and cathode surfaces. These polarization films generally have a much higher resistance than the body of the solution and vary in thickness with the current density. The cathode polarization film therefore affects current distribution and efficiency. Chemical changes within the film may change the physical characteristics of the deposit. As a result of secondary reactions, the pH of the polarization films may be grossly changed, producing additional effects. The thickness of the polarization films depends on current density, temperature, and agitation.

At the cathode, depletion of the metal ions being plated increases the absolute value of the deposition potential of the metal. If this potential rises above the hydrogen overvoltage for the process, hydrogen codeposits, with a loss of plating efficiency and an increase in the pH of the film. If the pH change is sufficient, metallic hydroxides may codeposit with the plated metal, disrupting the crystal structure and producing the coarse, granular deposits generally referred to as *burning*.

Electrophoretic attraction of selected colloids into the film may increase film viscosity by a large factor. Advantage is taken of this phenomenon to alter the deposited film and produce leveling and brightening actions on the deposit.

At the anode, reactions may occur as a result of the increase of concentration of the metal ions in the polarization film. Secondary reactions may produce oxidizing materials that destroy brightener components and/or form highly resistant oxides on the anode surface resulting in *passivation* of the anodes with a loss of current carrying capability. Advantage is taken of this in alkaline tin plating to ensure that the tin

dissolves in stannic form (Sn^{IV}) rather than as stannous tin (Sn^{II}).

Efficiency. Efficiency is the ratio of the actual amount of metal deposited on the substrate to the theoretical amount calculated by Faraday's Law. It is usually expressed as a percentage. The efficiency is influenced by the difference between the dynamic deposition potential of the metal being deposited and the potentials of the secondary reactions, particularly the deposition potential for hydrogen. Efficiency is therefore affected by polarization and hence by all the factors that also affect polarization.

Covering Power. Covering power is the ability of a solution to deposit metal into recesses or holes where the current density is low. It is basically controlled by the deposition potential for the metal being plated under the operating conditions of the solution being used. Numerous ingenious ways of measuring covering power have been developed. However, they are all empirical and not easily quantified.

Throwing Power. Throwing power is a measure of the ability of the solution to plate a greater thickness of metal into the low current density areas than would be predicted from the primary current distribution. If the ratio R = actual plating thickness/predicted thickness is greater than one, the throwing power is said to be positive; if smaller, negative. Solutions that are highly complexed with considerable polarization generally have positive throwing power. In fact, certain cyanide zinc solutions plate with ratios as high as 1.25. Uncomplexed solutions with moderate polarization have little or no throwing power (R ca. 1.0), while a few solutions with deposition reactions consisting of several dependent steps, e.g., chromium plating from hexavalent solutions, have negative throwing power.

An approximation of the throwing power of a solution may be made by plating a panel in a plating test cell such as the Hull Cell, calculating the deposit thickness expected on the basis of the primary current distribution at a number of points along the panel and measuring the actual thickness obtained at these same points. If the low current density areas show $R > 1$, throwing power is positive. It is also possible to obtain absolute values for the throwing power of a solution using the Haring test cell. However, both test procedures provide results valid only for the conditions under which the test was run, and any change in these conditions may also change the results. It is difficult therefore to relate the results of such tests to the results obtained with production parts.

Investigations into the mechanisms of brightening and leveling have revealed that the throwing power mechanism exists on two dimensional levels. Furthermore, these two levels appear to have no relationship to each other. Two different terms are therefore used, macro-throwing power and micro-throwing power.

Macro-throwing power is used to describe the ability of a solution to alter the ratios on a gross scale. Macro-throwing power measures the ability of a solution to plate into very low current density areas of a part, i.e., into holes, recesses, corners, etc.

Micro-throwing power is used to describe the ability of a solution to alter the ratios on a micro scale. Micro-throwing power measures the ability of a solution to plate into fine defects on the surface, such as scratches, polishing lines, pores, pits, etc. Micro-throwing power is an essential requirement for leveling or smoothing properties in a solution.

A solution may have good macro-throwing power and brightness without having micro-throwing power or leveling capability. Conversely, it is also possible for a solution to have excellent micro-throwing power and leveling, as well as brightness, without having good macro-throwing power.

Auxiliary Electrodes. Supplemental electrodes that are used to bring current to normally low current density areas (Fig. 2). Also referred to as "inside anodes," "internal anodes," or "conforming anodes." The electrode material may be the same as the metal being plated or may be an inert material, such as lead, stainless steel, titanium, or platinized titanium, depending on the nature of the plating solution. Such electrodes are mandatory for hard chrome plating internal diameters and complex dies. The internal anode is mounted on, but insulated from, the rack holding the parts and is connected to the anode bar, which provides the current. In some cases, a small rheostat may be connected in series between the internal anode and the anode bar to

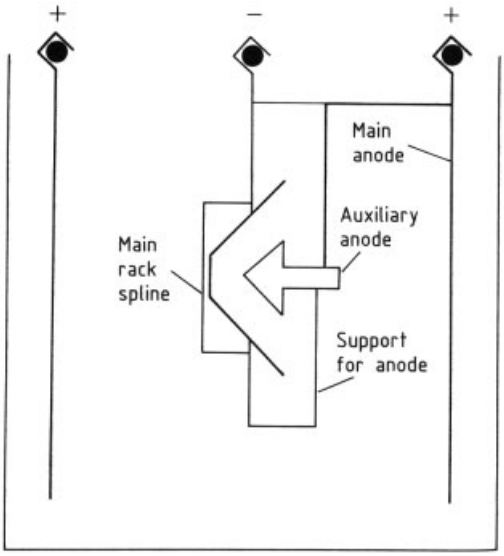

Figure 2. Auxiliary anode

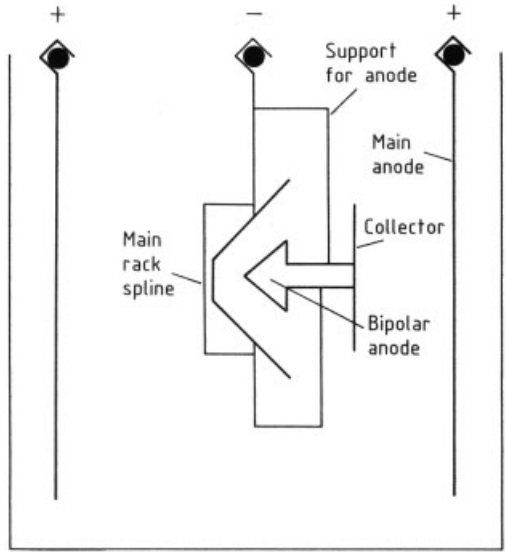

Figure 3. Bipolar anode
The bipolar anode and anode bar have no direct contact.

permit regulation of the current drawn by the internal anode.

Bipolar Electrode. A passive electrode placed between the parts to be plated (cathode) and the anode (Fig. 3). The end closest to the anode, also called the collector, becomes negative and is plated, collecting current that flows through the electrode to the end nearest the cathode, making it an anode that then functions the same as an auxiliary anode. The essential requirement for the functioning of this type of electrode is that the distance from the anode to the collector plate plus the distance from the bipolar anode to the area to be plated is less than the distance from the anode to the parts. Bipolar anodes have the advantage over auxiliary anodes that no direct contact with the anode bar needs be made. They have the disadvantage that the metal plated on the collector is wasted unless it can be stripped off and used as anode material.

Bipolar electrodes may also form accidentally and cause undesirable effects. There are three common cases:

1) In unlined conductive tanks, the distance from the anodes to the tank wall plus the distance from the tank bottom to the bottom parts on a rack may be less than the distance from the anodes to the work. The bipolar electrode that then forms may supply excessive current to the lower parts on the rack (Fig. 4). The cure is to insulate the tank, or insert a *shield* between the anodes and the tank wall or bottom.
2) In a lined tank, parts that fall into the tank and are allowed to remain there may form a bipolar electrode (Fig. 5). Additionally, the portion of the part that becomes the anode may dissolve, contaminating the solution. This is a frequent problem when nickel plating zinc-base die castings, resulting in severe contamination of the solution with copper and zinc. The cure is to regularly remove any parts that fall off the racks.
3) On automatic plating machines, a rack being lowered into the tank may temporarily become a bipolar electrode until it makes contact with the cathode bar. Hazing, clouding, or even peeling of the area that acts as the collector during this period may result. If this condition is suspected, provisions should be made for *live entry* of the rack into the solution, i.e., a connection with the cathode bar should be made before entry.

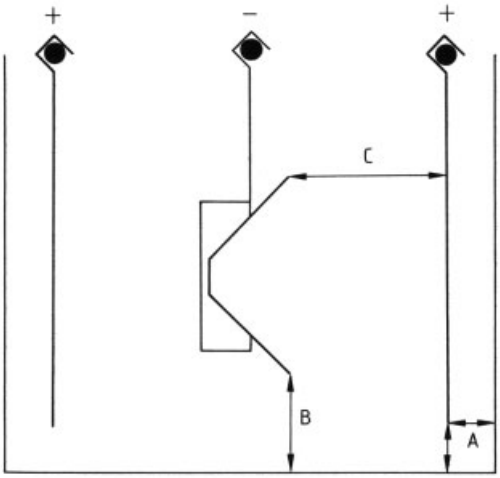

Figure 4. Bipolar short in an uninsulated tank
A bipolar short occurs when $C > A + B$.

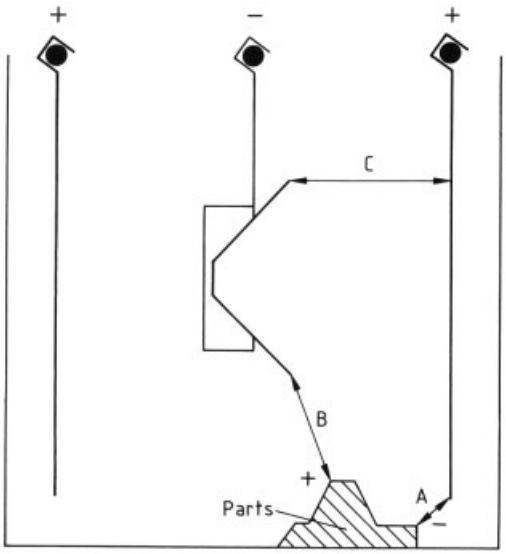

Figure 5. Bipolar short in an insulated tank
A bipolar short forms if $C > A + B$: the surface of fallen parts adjacent to the anode becomes cathodic (−), while the surface nearest the parts being plated becomes anodic (+).

Shields. Nonconductive materials placed between the anode and the parts to be plated in such a way and of such shape that they restrict the area available for current passage and therefore alter the current distribution and control the buildup of plating on high current-density areas. Shields may also be used between anodes and the walls of unlined tanks to prevent the formation of bipolar electrodes through the tank walls. Wire-reinforced glass is frequently used for this purpose (see Fig. 6)

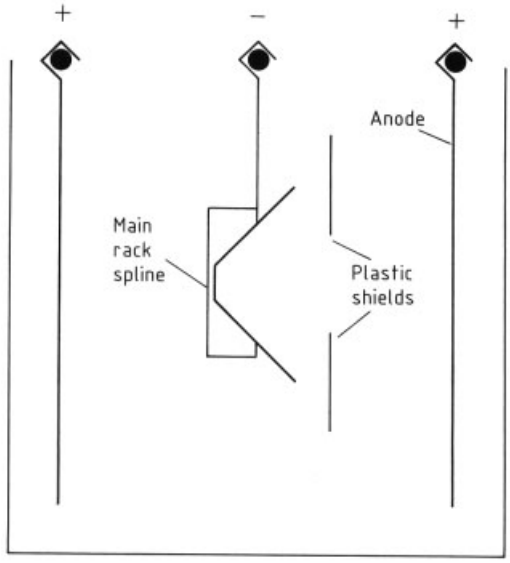

Figure 6. Use of shields to protect high current-density area

Composites. Composites are electrodeposited coatings that consist of a metallic electrodeposited matrix that surrounds and incorporates into its structure a nonmetallic material that has been suspended in the solution. The physical characteristics of the final coating are determined by the nature of the incorporated material and the ratio of incorporated material to metal. Materials that can be incorporated include fine beads of plastic, such as Teflon for added lubricity; diamonds and silicon carbide to form abrasive coatings or non-slip surfaces; microscopic particles of aluminum oxide in nickel to alter the electropotential under corrosive conditions and provide added protection in the multilayer nickel system for protecting steel and zinc-base die castings; and many other combinations.

Electroforming. Electroforming is the process of producing products by plating, such as compact recording discs, radar waveguides, pen parts, aircraft parts, complex bellows, and jewelry. Extremely complex configurations can be economically produced. A mandrel conforming

to the desired shape is prepared and plated with the metal or alloy required to the necessary thickness. If the shape allows, the part may be removed and the mandrel reused. More complex shapes use expendable mandrels made of aluminum, wax, or a low-melting bismuth – lead alloy. The mandrel is then melted or dissolved out of the finished part. Plating thickness may be as great as 6 mm (0.250″).

Alloy Plating (see Chap. 9). When two or more metals are present in the electrolyte, all of which are capable of being deposited from solution, they codeposit to form an alloy, provided the deposition potentials are sufficiently close together. The composition of the deposited alloy depends on the differences between the deposition potentials and the relative concentrations of the metals in solution. Since the polarization film affects the deposition potentials, and since the degree of polarization is affected by the current density, the composition of the deposited alloy may vary with current density. The extent of this variation depends on the diffusion characteristics of the metal ions, the amount of agitation, and the difference in the deposition potentials. Typical alloy coatings include brass, cyanide being used to control the deposition potentials of zinc and copper; bronze, cyanide being used to control the deposition potential of copper and sodium hydroxide to control the deposition potential of tin; nickel – iron, where only minimal control of the iron is needed; and many others. Ternary alloys are also possible, although in general the problems of solution and operational control restrict these to special applications.

Many alloy coatings are not necessarily recognized as alloys. The *hardeners* often used to modify the characteristics of gold deposits for various electronic uses are actually materials that are codeposited to form alloys. Similarly, metallic brighteners, such as cadmium in nickel solutions, codeposit to form an alloy that, although extremely low in cadmium, has a crystal structure that produces bright deposits.

3. Protection Against Corrosion

Metals may corrode in several different ways. Gases in the atmosphere may react with the surface of the metal to form a compound such as an oxide, a sulfide, or a carbonate. If the compounds formed by these reactions are water soluble, subsequent exposure to rain or other liquids may wash the salts away, allowing the corrosion to continue. If the compounds are insoluble and essentially nonporous, they may form a protective coating to hinder further corrosion.

Acid or alkaline materials may in the presence of moisture directly react with the surface, releasing hydrogen and forming salts with the substrate. Once again, depending on the solubility of the salts formed, the result may be additional corrosion or some protection. Acid rain has contributed greatly to the corrosion of metallic structures. The widespread use of salt to control ice and snow on roads and highways has resulted in severe corrosion of the reinforcing rod in the concrete and destruction of the supporting structures of many bridges and elevated highways.

Sooner or later, however, almost all corrosion becomes galvanic in nature. This may be because of the contact between dissimilar metals, stress within the metal, segregation of alloying components within the metal, or differential availability of oxygen as a result of dirt collection or salt formation on the surface. The only requirements for the progression of galvanic corrosion are the existence of a potential difference between points and the presence of enough moisture to allow the corrosion current to flow. This form of corrosion can be particularly damaging because it may be highly localized and penetrate the substrate in the form of deep pits.

Protection against corrosion may be obtained by several procedures. If the area to be protected can be covered by a layer of nonreactive material that is impervious to moisture and that completely encapsulates the surface to be protected, corrosion cannot occur.

Plated metals may protect surfaces in one of two ways. If the plated metal is less noble than the metal to be protected, any corrosive condition causes the plated coating to dissolve in preference to the substrate. The more active metal may even be able to protect areas that are no longer covered. Zinc and cadmium protect steel and brass by such a mechanism. Because of this characteristic, porosity in such a coating, while it reduces the life of the coating, does not produce catastrophic corrosion in the base metal.

A metal that is more noble than the substrate can only protect the substrate if the coating is completely encapsulating and pore-free. Such is the case for copper and nickel on zinc-base die castings or steel. Any porosity in the coating results in severe pit corrosion of the substrate. Furthermore, this corrosion also spreads laterally causing additional loss of the plated coating and additional corrosion.

To effectively prevent corrosion, multilayered nickel and chromium coatings are applied to automobile hardware. The several layers are deposited using different additives so that their electrochemical activities are different. Such corrosion as does occur is restricted to the plated layers, thus allowing continued protection of the substrate.

4. The Plating Process

The simplest concept of the plating process is as an electrochemical process in which metallic ions in a supporting solution are reduced to the metallic state at the cathode, forming the plated film. At the anode, metal may be oxidized from the metallic state to the ionic state, thereby replenishing the metallic ion content of the solution. The ions, released at the anode, migrate through the solution to replenish the cathode film. The electrons released by the oxidation of the metal at the anode, travel through the external circuit to the cathode. A driving potential (emf or applied voltage) to force the reaction to proceed must be provided by an external source. This source could be a battery, a generator, or a d.c. power supply (commonly referred to as rectifier). In modern practice, rectifiers are used almost exclusively.

The actual process is considerably more complicated. The dissolved metal ions are frequently held not in simple ionic form but as complexes of various types, often carrying a negative charge, so the migration tendency is not toward the cathode where they are needed, but back toward the anode increasing the concentration polarization. Some form of supplemental solution movement is needed to ensure adequate migration rates to replenish the cathodic film.

Immediately upon the application of current, a capacitive film, generally referred to as the zeta layer, is formed closely adjacent to the cathode. The incoming metallic ions are discharged to the elemental state at the surface of this layer. The atoms of metal then diffuse through this layer to the substrate surface, where they seek to establish themselves as part of the growing crystal lattice. The incoming atoms enter into the lattice in such a way as to reduce the overall free energy of the surface.

The *initial layer* on the substrate forms as submicroscopic crystallites, which grow by accretion of the incoming atoms both horizontally and vertically until a continuous coating is formed. The number and distribution of the crystallites in the initial nucleation of the surface is strongly affected by the surface preparation and in turn affects the crystal structure of the subsequent coating. Addition of selected organic materials to the electrolyte can modify the growth habit of the crystallites, restricting growth to the horizontal plane in preference to vertical growth, speeding the completion of the initial covering layer, and refining the crystal structure of the deposit. Incorporation of organic and inorganic additives in the electrolyte can be used to block selected sites on this growing lattice and force the incoming atoms into a desired growth pattern, thus producing leveling, brightness, stress relief, or other desired characteristics in the deposit.

The beneficial effect of these additives is such that in modern plating practice unmodified solutions are seldom used. Historically, the initial discovery of additive materials was an empirical process and developed as a result of the experimentation of practical platers whose scientific training was minimal. Subsequently, supply houses specializing in meeting the needs of platers recognized the need for improved solutions and mounted a determined scientific effort to develop solutions that would meet the increasingly stringent requirements of the plating industry. The result was the development of a complex and extensive series of brighteners or addition agents that could produce deposits that were bright, smooth, hard, soft, especially corrosion resistant, or with some other desired characteristic. From the original single or double type additive, e.g., the use of cadmium chloride and yellow dextrin for bright barrel plating of nickel, these additive systems have developed into extremely complex mixtures with as many as a dozen different materials in vari-

ous concentrations, some of which are extremely low. In many cases, the effective ingredient is a specially manufactured organic chemical that has been reacted under carefully controlled conditions to produce the active agent. Many of these are patented and have been developed by supply houses, which sell them to the trade as proprietary mixtures. While the composition of many of these additives can be inferred from the patents or other published literature, the difficulty of properly preparing the mixtures or some of their ingredients has militated against preparation of the materials by individual platers' homebrew, and therefore proprietaries are used almost exclusively.

With continued electrolysis, a polarization film usually forms. Since this film is thicker in the high current density areas, the added resistance of the film in those areas forces the current flow into areas of normally lower current density, producing an improvement in the throwing power of the solution. If the formation of this film increases the deposition potential of the metal sufficiently, hydrogen is codeposited with the metal, the efficiency falls off, and the throwing power improves still further.

4.1. Design for Plating

Because of the complex nature of the plating process and the way plated coatings can be used to restrict corrosion or impart other desirable properties, there are substantial benefits to be gained from having a plating expert as part of the design team. There are several aspects belonging to the design process:

1) Selection of the proper coating. This involves an evaluation of the corrosive environment to which the part will be exposed, the purpose to be served by the coating, and the nature of the substrate material. The expected life of the part, and the effects in danger and cost of a failure of the part, must be considered. The placement of the plating operation in the sequence of manufacturing may be important.
2) Part shape. Sharp corners, deep recesses, blind holes, and a number of other dimensional conditions that affect the current distribution may make plating difficult or limit the type of solution that can be used. If parts are to be rack plated, the design must make provision for easy and secure racking. If parts are to be barrel plated, design features that eliminate nesting or sticking should be incorporated.
3) Dimensional limitations. The thickness of the plated deposit must be added to the original dimensions of the part. Because of the effects of current distributions and throwing power, these dimensional changes are affected by the location of the critical area on the part. An experienced metal finisher can often accurately estimate what these effects may be and suggest either dimensional allowances be made or design effective racking and shielding procedures to hold the changes to controlled limits. This can be illustrated by the approximate effects of a deposit on the dimensions of a threaded fastener or threaded hole.
   The approximate changes to be expected are: major diameter, 2 – 2.5 times the plate thickness; root diameter, 1 – 1.5 times the plate thickness; pitch diameter, ca. 4 times the plate thickness. For an external thread the dimensions are increased by this amount; for an internal thread there is an equivalent decrease. It is entirely possible, therefore, that a fastener that mated perfectly with a threaded hole before plating will be out of tolerance after plating. Projecting threaded fasteners may receive even greater thickness of plating and have to be shielded.
4) Production sequence. The ideal place in the manufacturing sequence for the finishing operation may need consideration; e.g., it may be more economical to plate certain parts before assembly rather than after.

All of these factors must be considered and optimized to produce the desired quality parts at the lowest possible cost.

4.2. Control

Although there are numerous variables in the process, routine operation is attained with surprisingly little control. Routine chemical analysis of the composition of the solutions and careful control of operating temperatures for all solutions often suffices. For more critical opera-

tions, control of the concentrations of the addition agents may be required. For specification work, routine thickness checks on finished parts on either a random or statistically controlled basis may be needed. Statistical control of thickness is especially important in the barrel plating of gold in order to hold costs to a minimum. For especially critical work, such as aircraft maintenance and repair, the space program, or nuclear work, a much more sophisticated control procedure may be required.

4.3. Impurities

Impurities that can affect the character of the deposit include metallic ions, organic materials that can affect the polarization film, and organic materials that can adsorb onto the parts to interfere with proper adhesion.

Metallic ions affect the bath operation in two ways: they may deposit by immersion on the anode materials when the current is not flowing. In general, if a metal codeposits with the plated metal and is insoluble in the matrix so that no alloy is formed, it beneficially acts as a brightening agent. If, however, an alloy is formed, it acts as an impurity and produces undesirable characteristics in the deposit, including roughness and/or loss of brightness. The effect therefore may depend on the concentration of the extraneous ions and the current density, i.e., the effect is a function of the mole percentage of the foreign metal in the deposit. An extraneous metal ion may act as a brightener at one concentration level and current density, and as an impurity under some other set of conditions. Such is the case, for example, with cadmium or zinc in a nickel solution. Cadmium at levels up to approximately 4 mg/L acts as a brightener for current densities of ca. 0.5 – 3.0 A/dm^2. Above that level, however, it acts as an impurity and produces dark deposits in the low current-density areas. For zinc, the allowable limit is ca. 6 – 8 mg/L and the upper limit ca. 10 mg/L. Even when functioning as a brightening agent, such codeposited metals may affect other characteristics of the deposit, such as brittleness, hardness, stress, etc.

Brighteners at lower concentrations	Cd (ca. 4 mg/L) or Zn (ca. 8 mg/L) in nickel solutions	Impurities at higher concentrations

When an extraneous metal deposits on the anode during idle periods, it may subsequently deplate from the anode as solid particles that remain suspended in the bath, migrate to the cathode, and deposit as discrete particles, which cause roughness. Such is the case, for example, with lead used as a brightener in bronze plating solutions, particularly at high levels.

Metallic contamination may often be removed by electrolytic purification. Cathodes of sheet metal or scrap parts are plated at low current densities, usually ca. 0.5 A/dm^2. At properly selected current densities, the contaminating metal is preferentially removed from the solution. The process is often referred to as *dummying.*

Organic Materials. Many organic materials can interfere with the polarization film causing loss of brightness, streaking, and hazing of the deposit. Often breakdown products of organic systems produced as a result of oxidation at the anode, incomplete reduction at the cathode, or polymerization can act in this way. Typical is the deterioration of some of the aldehydes often used as brighteners in cyanide zinc plating solutions. At high sodium hydroxide levels and temperatures above 50 °C they may undergo the Cannizzaro reaction, polymerizing to long-chain materials. There is not only a selective loss of brightener ingredients, leading to an imbalance in brightener concentrations, but the long-chain materials interfere with the proper development of the polarization film, producing dullness and sometimes streaking or darkening of the deposit.

Organic materials that can adsorb and interfere with adhesion include oil dragged into the solution by improperly cleaned parts, drippage from overly lubricated equipment or introduced by improperly filtered agitation air; buffing compound from incompletely cleaned parts or airborne from nearby buffing operations; wetting agents and cleaner components from incompletely rinsed parts; and many other similar materials. These can cause pitting, roughness, hazing, and loss of adhesion. Silicones are especially damaging.

4.4. Inspection

While there is an increasing tendency for purchasers of plated coatings to specify the characteristics in some detail, especially minimum thicknesses, there are many characteristics of the plated coating that cannot be determined by physical testing. In particular, appearance and the presence of defects such as blisters, pitting, roughness, etc., are best determined visually. The rack plating process especially lends itself to rapid visual inspection since the parts can often be inspected as they are removed from the racks at the end of the process. Even with barrel plated parts a "grab" sample, taken as the parts are being emptied from the barrel, represents a random sample that can be quickly inspected to determine the general quality of the barrel load.

The areas where visual inspection is carried out must be properly lighted. A diffuse white light is most effective, and since many defects can be obscured by reflected light, the ideal inspection station has an inspection box lined with dull white surfaces and with the diffused inspection light coming from overhead. Determining coverage of chrome over nickel, stains on brass and copper, and the inspection of silver-plated items are especially demanding in this regard.

Inspection is a service that enables the production group to operate more efficiently. Fast transfer of information between the inspectors and the production staff is essential. Ideally, the inspectors should be located so that they can detect deteriorating quality early enough, and report it to the production staff promptly enough, that the process can be corrected before any appreciable number of rejects are produced. This implies as well that the inspectors should have thorough training, not only in the quality that is or is not acceptable, but also in the finer distinctions that signal that the process is becoming quality defective.

Where precious metals are barrel plated, statistical control of the plating thickness can assure meeting the specification without applying excessive thicknesses, thus becoming a potent economy tool. However, to be effective the testing facility must be located close to the plating facility so that a random sample can be taken from the barrel and tested promptly before the barrel is removed from the plating tank. In this way, no load ever has to be reloaded and reprocessed because of inadequate thickness.

5. Substrate Preparation

5.1. The Nature of the Adhesive Bond

Adhesion is a micro phenomenon, but since all tests for adhesion are essentially macro tests, the result can present at best an average of the degree of adhesion on the surface. In one sense, adhesion is an absolute: it either exists, or it does not. But since adhesion occurs at discrete points on the surface and adjacent points may have it or lack it, the test procedures actually measure the adhesion that results from the percentage of points properly activated.

There are three different forms of adhesive bonding for plated coatings. If the surface is suitably cleaned and activated, the incoming metallic atoms can crystallize as part of the lattice of the substrate and *chemical bonding* actually takes place. Because the incoming atoms in general have different radii than the substrate, and prefer a lattice of somewhat different dimensions, the boundary layer is somewhat distorted and therefore stressed. However, the adhesion of the two metals will be greater than the cohesion of the weaker metal: if the bond is later ruptured by an exterior force, the rupture occurs within the crystal lattice of the weaker of the two metals.

If the surface is prepared to a lesser degree, the incoming atoms attach themselves to the surface by an *undirected homopolar bonding,* sometimes referred to as London forces and somewhat similar to the cohesive bonding that occurs between the molecules of a liquid. No interchange of electrons take places, although the contact is sufficiently intimate that diffusion of one metal into and through the other may occur. The bond strength of this type of bond is somewhat less then the strength of the first type, but still is such that in the event of rupture the failure usually occurs within the weaker metal.

The third type of bonding is *mechanical bonding,* which occurs as a result of the incoming metal atoms depositing within the interstices of a roughened surface, forming an interlocking bond between the substrate and the deposited metal. Such a bond is obtained, for example,

on a sandblasted or abraded surface and, unless supplemented by one of the other types of bonding, is the weakest bond. It may, however, be adequate for many practical applications. Further, since such roughening produces a considerable increase in the surface area for bonding, there is always an apparent improvement in bond strength when such roughening is used.

For any of these bonds to form, the substrate must be properly wetted by the solution used for deposition so that the surface is in intimate contact with the source of metal atoms.

5.2. The Part Preparation Cycle

Platers often discuss the concept that parts must be *chemically clean* to accept an adherent deposit involving one of the three bonding mechanisms. Yet the environment in which we exist precludes such cleanliness. A truly chemically clean surface can be obtained only in a vacuum. There is, however, an acceptable level of cleanliness even under normal conditions. The preplating or part preparation cycle has as its objective the development of a surface with this acceptable level of cleanliness and activation.

The preparation procedure should not be viewed as a series of discrete steps. Each step of the process is dependent on what went before and in turn affects each step that follows. There are four steps to the preparation cycle:

1) *Gross cleaning,* the removal of the major portion of the surface soils. Gross cleaning may be accomplished by washing with a solvent; degreasing using chlorinated hydrocarbons; cleaning using emulsifiable solvents, either by soaking followed by rinsing, or by using the emulsified solvent as the cleaner; detergent cleaning; or alkaline cleaning. The aqueous materials may be used either by soaking in a tank, with or without agitation; washing in an agitated power washer; or by power spray washing.
2) *Fine cleaning,* the removal of the residues from the gross cleaning step along with adherent or electrostatically attracted solid particles, both conductive and nonconductive. Some fine particles of metal may be sufficiently attracted to the surface to adhere after this step and must be removed in a later step. This fine cleaning is usually carried out electrolytically in an alkaline cleaner. If the part is made the cathode so that hydrogen is evolved, it is referred to as *direct cleaning*; if the anode, as *reverse cleaning.* For many years, it was thought that the scrubbing action of the released gas was the main cleaning action, and direct cleaning was preferred over reverse cleaning since the volume of hydrogen released is double that of oxygen. Aside from the fact that the release of such copious amounts of hydrogen could cause serious hydrogen embrittlement, the electrolytic cleaning process is now recognized as being more complicated, involving changes in the surface oxides and the neutralization of charged colloidal particles prior to releasing them from the surface. Reverse cleaning is therefore generally used, except for lead and silver substrates.
3) *Oxide removal,* the removal of all surface oxides, both natural and those developed in the alkaline cleaning steps. The top few atomic layers of the surface may be severely distorted and stressed by the forming, machining, polishing, and buffing operations used to develop the desired surface. Adhesion of the subsequent coatings is better if these layers are removed. By proper selection of the cleaner in the electrocleaning step, these distressed layers can often be converted to oxides, which can be removed in this step. Additionally, there are many powerful detergent alkalies, such as alkaline silicates, which may adhere to the surface that has been cleaned and which must be removed. The oxide removal is usually accomplished in an acid solution, although alkaline descalers using current and occasionally periodic reversal of the current can sometimes be effective. Hydrofluoric acid, or a soluble fluoride salt, is often added to the acid mixture to remove the silicates remaining from the cleaning step or particular alloying elements that may be resistant to other acids. Prepared powdered acid mixtures containing fluorides, accelerators, inhibitors, wetting agents, deflocculants to remove fine particles, and other ingredients are an excellent way to deoxidize many alloys. In some cases, however, they may not have sufficient free acidity to remove spot welding scale and similar oxides and may

have to be supplemented with mineral acids. In other cases, mineral acids alone are effective.

4) *pH adjustment,* prior to entering the plating solution, the parts carry on their surface a film of rinse water that partially mimics the pH of the last solution used. If this pH is sufficiently different from the pH of the plating solution, some of the metal ions in the plating solution may be precipitated on the surface of the part before any plating takes place. This can cause clouding, hazing, and streaking of the plated deposit, or in extreme cases loss of adhesion. Therefore a final dip in a dilute compatible solution of similar pH to the plating solution is recommended. Typical examples are 5 wt % sulfuric acid before nickel plating or a solution of 15 g of sodium cyanide plus 7.5 g of sodium hydroxide per liter before cyanide copper plating.

Each of these steps except pH adjustment must be followed by effective *rinsing.* The rinsing between operations is as important as the operation itself. Inadequate rinsing can result in cross contamination of solutions and the production of rejects.

Drag-out is the amount of liquid carried out of a processing tank by work being transferred. *Drag-in* is the amount of liquid carried into a processing tank by work being transferred from a previous tank. Drag-in is usually unimportant unless (1) the previous solution contains contaminants harmful to the receptor tank or (2) the volume exceeds the drag-out volume and the evaporation from the receptor tank so that receptor tank volume increases. Drag-out must be reduced to reduce material loss, reduce the required rinsing volume, and reduce the load on the waste disposal system. KUSHNER has produced an excellent treatment of the entire subject [6].

Although four steps are outlined, the pretreatment cycle does not necessarily consist of four tanks and the associated rinses. The nature of the soils on the surface may require several different treatments to accomplish each stage of the cycle.

Considerable research has gone into the development of methods for evaluating the degree of cleaning that is obtained. There is an extensive body of literature on the subject, describing the use of fluorescent tracers, radioactive tracers, the atomizer test, the water-break test, and many others [1, bibliography of Chap. 3 F]. The most widely used is the *water-break test.* A surface that is to be plated in an aqueous solution must be completely wetted by a film of water. If parts that are to be plated are examined prior to entry into the plating solution, any breaks in the film of water on the surface indicates an area that is hydrophobic, rather than hydrophilic, and that is therefore improperly prepared. Because some alkalies and wetting agents in the cleaning solutions can sometimes bridge over hydrophobic areas and give a false positive result, the observation is generally carried out after the acid treatment of the parts, or for purposes of investigating trouble, by immersing the part in a weak acid before making the observation.

5.3. Preparation of Specific Substrates

Different substrate metals, because of their reactive characteristics, require somewhat different methods and solutions to carry out the four steps of the preplating cycle.

The concentrations of acids and aqueous ammonia are frequently given as percent by volume (vol %). This is the volume of concentrated solution to be mixed with water to form 100 parts of solution. For example, 20 vol % hydrochloric acid would be 2 L of concentrated hydrochloric acid (ca. 38 wt %, 22° Bé) mixed with enough water to make 10 L of solution (2 gal of acid mixed with water to make 10 gal). The concentrated solutions from which the diluted solutions are made include hydrochloric acid, nitric acid (69 wt %, 42° Bé), phosphoric acid (ca. 85 wt %), sulfuric acid (98 – 100 wt %, $d = 1.84$), and aqueous ammonia (28 – 29 wt % NH_3, 26° Bé). In addition, some solutions are mentioned which are prepared by mixing volume parts of concentrated acid and water.

5.3.1. Steel

The wide variation in steel compositions requires an equivalent variation in the preplating cycles. In general, the soak cleaners are high in alkalinity and have few restrictions on the type of surfactants used. Electrocleaners are also high in alkalinity. For best results, they should contain an appreciable percentage of soluble silicate

to act as an inhibitor and prevent surface attack. The parts are usually made anodic during cleaning, and high current densities are used. Current densities during cleaning of 7.5 – 12.5 A/dm^2 (75 – 125 A/ft^2) are not unusual. The exception is for nitrided, carburized, or very high carbon steel alloys, which may sometimes develop a carbon smut under these conditions. Lower current densities may be required for these sensitive alloys.

Oxide removal can be accomplished in hydrochloric or sulfuric acid, often with fluoride added. Proprietary powdered acids are also available. The acid should be inhibited to prevent excessive attack and the development of a carbon smut. Transfer and rinsing times must be restricted or a *rust bloom* may develop as a result of the action of the residual acid on the highly active surface that results from acid treatment. Many cycles call for treatment in an alkaline descaler or a mixture of sodium cyanide and sodium hydroxide (30 g/L of each) to remove any trace of this rust.

5.3.2. Cast and Malleable Iron

The high concentrations of carbon and silicon in the surface of cast and malleable iron alloys make them especially prone to the formation of smut. Current densities in electrocleaning should be held to 10 A/dm^2 (100 A/ft^2) maximum. Acid treatment times should be as short as possible, and the acid should always contain a fluoride salt to assist in the removal of silicon from the surface. Because the surface carbon has a very low hydrogen overvoltage, complete coverage in some solutions, e.g., cyanide zinc, may be slow or impossible to achieve. A copper, cadmium, or tin strike is often used to alter this surface characteristic and permit more rapid coverage.

5.3.3. Copper and Copper Alloys

Copper generally may be handled in much the same manner as steel with the exception that the acid used for oxide removal is more dilute. Brass and the other copper alloys are more susceptible to attack than copper itself, particularly in the alkaline cleaning step. Cleaners with lower pH are generally employed and may be fortified with complex phosphates or other complexing agents to prevent excessive oxidation. Current densities in electrocleaning range from 4 to 8 A/dm^2 (40 – 80 A/ft^2), and times should be restricted. If surface conditions demand cleaning times > 2 min, it is common practice to use two stage electrocleaning with the initial cleaning cathodic and the final 30 – 60 s anodic to remove any contaminants that may have been plated out on the surface during the cathodic cleaning. Oxide removal may be accomplished in 20 –30 vol % hydrochloric acid or 5 – 15 vol % sulfuric acid. The use of other acids, such as acetic, fluoboric (fluoroboric), or glycolic, may occasionally give better results. Specially formulated powdered acids are also effective.

5.3.4. Zinc

Zinc is an amphoteric metal that is highly reactive. It is therefore susceptible to attack by strong alkalies and even relatively weak acids, and easily acquires immersion deposits. The soak cleaners used to remove oils and buffing compounds should be detergent type, or mild alkalies buffered to a pH range below 10.5. The use of soaps in these cleaners should be avoided because zinc readily forms metallic soaps, which are most difficult to remove in subsequent steps. Rosin-based soaps, despite their effectiveness in removing buffing compounds, are especially prone to forming such metallic soaps. Their use should therefore be avoided if at all possible, or special precautions taken in the succeeding steps to ensure their removal. Moderately strong electrocleaners may be used provided they contain an adequate percentage of silicate to provide inhibition. The acid for the deoxidizing step must be a low-concentration mineral acid, i.e., 0.5 – 2 vol % sulfuric acid, 5 – 10 vol % fluoboric acid, or a proprietary powdered acid that is buffered and contains fluoride. Contamination of any of the preplating solutions with copper can result in the deposition of an invisible, or barely visible, immersion copper deposit, which may cause blistering after plating or upon heating. A copper or brass strike should be used before applying any plated coating other than zinc or cadmium.

5.3.5. Aluminum

Aluminum has many characteristics that are similar to zinc, but is even more reactive. Because of the wide variation in aluminum alloys, it cannot be considered as a single metal, but rather as a group of four or five related metals. The cycles required for proper preparation are therefore varied: the copper alloys are the simplest to process, while those containing magnesium and silicon are the most difficult.

Aluminum is rarely electrocleaned except in European practice, but etched finishes are common and chemical polishing is often used along with or instead of the oxide removal step. The precleaners should be neutral detergent types, buffered low pH type alkalies, or moderate pH alkalies inhibited with silicates. While many so-called nonetch, nonsilicated cleaners are offered, these are not actually nonetch materials, but rather carefully buffered cleaners with very low etch rates. This type of cleaner may be preferred ahead of operations such as chromating, phosphating, and electropolishing. If silicate inhibitors are used, the oxide removal step should contain a fluoride salt. The acid used for oxide removal must contain an oxidizing agent such as nitric acid or an iron(III) salt. Chromate salts may be used in cycles that are not to be followed by plating, but are considered risky in a plating cycle because of the dangers of dragover of chromium(VI) into the plating slutions and the formation of a stable chromate film on the aluminum.

The following are typical formulas for oxide removal acids:

1) Nitric acid (42° Bé) 500 mL
   Water 500 mL
2) Nitric acid 500 mL
   Water 500 mL
   Ammonium bifluoride 120 g
   (NH_4HF_2)
3) Nitric acid 500 mL
   Sulfuric acid (98 wt %) 100 mL
   Water 400 mL
   Ammonium bifluoride 60 – 120 g

Chemical polishing solutions are based on mixtures of nitric and phosphoric acid, sometimes with added acetic acid, and occasionally with accelerators such as copper(II). They are operated at temperatures close to 100 °C with agitation to avoid gas streaking. Control of temperature and water content are critical to successful operation. Two typical formulas are:

Formula 1	
Nitric acid	2.8 – 3.2 wt %
Water	17 – 23 wt %
Phosphoric acid	64 – 70 wt %
Aluminum phosphate (formed by reaction)	$\leq$ 12 wt %
Copper	0.01 – 0.02 wt %
Formula 2	
Phosphoric acid	70 wt %
Nitric acid	2 wt %
Glacial acetic acid	15 wt %
Water	13 wt %

Many other formulas are listed in the literature, including systems based on phosphoric – sulfuric – chromic acids, phosphoric – sulfuric – nitric acids, nitric – fluoride mixtures, and alkaline systems.

The final step in the preparation of aluminum for plating provides a good example of the use of a carefully prepared solution to deposit an adherent and structurally sound deposit of metal by immersion. Zinc, zinc – copper (Alumon D), zinc – copper – nickel (Alumon N), or zinc – copper – nickel – iron (Bondal), or tin may be deposited from zincate or stannate solutions. All solutions contain high amounts of sodium hydroxide and may be either concentrated (500 – 600 g/L) or dilute (120 – 300 g/L). While formulas for these solutions can be found in the literature, the most effective procedures are patented and are available as proprietary mixtures. In any event, the process is sufficiently complex, and control of raw material quality sufficiently important, that it is generally more economical to purchase prepared and quality-controlled mixtures.

Following the application of the metal film, the aluminum is usually given a cyanide copper strike before applying electrodeposits other than those from cyanide solutions. The alloy zincates permit the use of a pyrophosphate copper strike and, depending on the alloy deposited, the direct application of deposits from acid solutions within a limited pH range. The tin film calls for a special bronze strike from a cyanide solution.

Aluminum may also be plated by first anodizing it in phosphoric acid under carefully controlled conditions. Use of this procedure is limited to special applications.

5.3.6. Magnesium

Magnesium [*7439-95-4*] is very similar to aluminum in that it is highly reactive. It forms surface oxide even more readily than aluminum. It therefore requires special procedures in order to obtain adherent coatings. These procedures are quite specific and do not apply to any other metals.

There are two approaches: applying an immersion zinc coating followed by a cyanide copper strike, similar to the procedure used on aluminum; and direct plating with a special autocatalytic nickel deposit. In either case, the desired decorative or engineering coatings are applied over the initial layer.

After cleaning the parts by alkaline soak and electrocleaning as necessary, the magnesium is pickled in one of the two following solutions:

Ferric nitrate pickle	
Chromic acid (CrO_3)	180 g/L
Ferric nitrate ($Fe(NO_3)_3 \cdot 9\,H_2O$)	40 g/L
Potassium fluoride (KF)	3.5 g/L
Temperature	16 – 38 °C
Time	15 – 180 s

At 38 °C this pickle removes metal at the rate of 3 µm/surface/min, i.e., 3 µm of metal per minute is removed from each exposed surface. Where no dimensional change can be tolerated, the use of the chromic acid pickle is recommended.

Chromic acid pickle	
Chromic acid (CrO_3)	180 g/L
Temperature	16 – 93 °C
Time	2 – 10 min

The use of chromic acid pickles may leave chromate films on the surface that reduce the adhesion of the subsequently deposited coating. After rinsing, the parts are activated in the following solution:

Phosphoric acid (H_3PO_4)	20 vol %
Ammonium bifluoride (NH_4HF_2)	105 g/L
Temperature	16 to 38 °C
Time	15 – 120 s

After rinsing, the zinc coating is applied from a pyrophosphate solution:

Zinc sulfate ($ZnSO_4 \cdot H_2O$)	30 g/L
Tetrasodium pyrophosphate ($Na_4P_2O_7$)	120 g/L
Sodium fluoride (NaF)	5 g/L
or lithium fluoride (LiF)	2 g/L
Sodium carbonate (Na_2CO_3)	5 g/L

Because of the low solubility of the tetrasodium pyrophosphate, generally this solution is prepared and used at elevated temperature (70 – 85 °C). The parts are immersed for 3 – 10 min in the 79 – 85 °C solution and agitated mildly. Glass or fiberglass equipment of any kind cannot be used. Careful control is essential for best results. The pH is maintained between 10.2 and 10.4 (glass electrode, 25 °C).

For the magnesium alloys ZK-60, HM-21, HM-31, or HK-31, a double zinc immersion is required. To accomplish this, the pickling, activating, and zincating steps are repeated. Separate solutions for pickling and zincating should be used for the double zinc immersion technique.

A limited number of proprietary zincating solutions similar to the one listed are available.

The parts are then given a copper strike in either of the following:

Bath 1	
Copper cyanide (CuCN)	41 g/L
Potassium cyanide (KCN)	68 g/L
Free potassium cyanide	7.5 g/L
Potassium fluoride	30 g/L
pH	9.6 – 10.4
Temperature	54 – 60 °C
Bath 2	
Copper cyanide (CuCN)	41 g/L
Sodium cyanide (NaCN)	52.5 g/L
Free sodium cyanide	7.5 g/L
Rochelle salt ($KNaC_4H_4O_6 \cdot H_2O$)	45 g/L
pH	9.6 – 10.4
Temperature	54 – 60 °C

The parts are plated ca. 6 min. With either bath, electrical contact should be made quickly, with initial current 5 – 10 A/dm^2, then with a current of 1 – 2.5 A/dm^2. Cathode rod agitation is suggested.

To utilize the direct autocatalytic nickel method, the parts are processed through the pickling step. They are then given a chemical etch in one of the following:

Etch 1, for alloys containing aluminum	
Chromic acid (CrO_3)	120 g/L
Nitric acid, 11 vol %	to make 1 L
Etch 2, for other magnesium alloys	
Chromic acid	60 g/L
Nitric acid, 9 vol %	to make 1 L

Stainless steel tanks or tanks lined with glass, ceramic, vinyl, polyethylene, or other suitably resistant plastic material are suitable. The parts are immersed 20 – 60 s in the room-temperature solution. The parts are rinsed thoroughly and placed immediately in a hydrofluoric acid dip.

For *thorium-containing alloys* (and other alloys where dimensional change is not critical) the parts are immersed 30 – 60 s in a phosphoric acid pickle at 21 – 32 °C. Metal loss is ca. 13 µm/surface. If thorium alloys have critical dimensions, chromic acid pickle is used and followed with an immersion in a strongly alkaline cleaner to remove possible chromate films.

Magnesium parts are treated in 5.5 vol % hydrofluoric acid for 10 min at room temperature. For treating the alloy AZ61A or AZ80A 20 vol % hydrofluoric acid is used. The parts are transferred to the nickel bath immmediately after rinsing. Ammonium bifluoride, 60 – 90 g/L (8 – 12 oz/gal), can often be substituted for hydrofluoric acid, which eliminates the hazard of handling concentrated hydrofluoric acid.

Autocatalytic nickel plating is carried out in the following bath:

Basic nickel carbonate	
($2\,NiCO_3 \cdot 3\,Ni(OH)_2 \cdot 4\,H_2O$)	10 g/L
Hydrofluoric acid, 0.6 vol %	to make 1 L
Citric acid ($C_6H_8O_7$)	5.2 g/L
Ammonium bifluoride (NH_4HF_2)	10 g/L
Sodium hypophosphite ($NaH_2PO_2 \cdot H_2O$)	20 g/L

Ammonium hydroxide (commercial) is diluted 4 parts to 100 parts and is used to raise the pH. Hydrofluoric acid or ammonium bifluoride is used to reduce pH. The preferred pH is 5.5 – 6.3. In this range, the deposition rate is 20 – 25 µm/h. The presence of sulfate ion is harmful to the bath; therefore, sulfuric acid should not be used for pH adjustment. Mild mechanical agitation is used.

The solution is somewhat unstable. The addition of 1 – 2 mL of triethanolamine per liter stabilizes the bath, but reduces the deposition rate by ca. 30 – 40 %.

The parts are immersed in the nickel solution held at 77 – 82 °C with a pH range 4.5 – 6.8 (glass electrode, 25 °C) The bath sould be filtered either periodically or continuously. The use of a plastic, magnetically coupled pump eliminates problems with shaft seals.

If heavy or specialized electroless nickel deposits are to be applied, the parts can be transferred to the specialized bath after a thickness of 5 µm has accumulated.

The parts are rinsed in cold water. For maximum corrosion resistance when the nickel plate is the final deposit, immerse for 10 min in chromic acid (CrO_3, 2.5 g/L) or sodium dichromate ($Na_2Cr_2O_7 \cdot 2\,H_2O$, 120 g/L). Both solutions are operated at 88 – 100 °C. The parts are thoroughly rinsed in cold water and dried. To improve adhesion, they are baked at 200 °C for 1 h. The deposit should remain adherent and free of blisters.

5.3.7. Stainless Steel

Stainless steel is passive because of the natural oxide that rapidly forms on the surface. Various methods of activating the surface to permit application of adherent electrodeposits have been proposed. The use of cathodic treatment in an alkaline descaler containing sodium cyanide is sometimes successful, as is an acid copper strike. The most successful procedure, however, remains the use of the *Woods nickel strike.* This consists of 45 g/L (6 oz/gal) of nickel as nickel chloride in 10 – 12 vol % hydrochloric acid. It is used at room temperature at current densities of 2 – 10 A/dm^2 (20 – 100 A/ft^2) for 30 – 60 s as required to obtain complete coverage. If the subsequent deposit is to be nickel, the parts can be transferred to the plating tank without rinsing. For other deposits, rinsing should be thorough but rapid so that the nickel plate does not passivate (see Section 5.3.8).

Since the anode efficiency is much higher than the cathode efficiency, the nickel concentration tends to increase. The nickel concentration should not be allowed to exceed 75 g/L (10 oz/gal), or else the effectiveness of the strike may be lost. To reduce the nickel concentration, solution is removed and replaced with fresh hydrochloric acid.

The removed solution may be used to adjust the nickel or the chloride content of a standard nickel bath. If the addition is so large as to seriously reduce the pH of the solution, some of the acidity may be neutralized with nickel carbonate before the addition is made.

5.3.8. Nickel

Nickel rapidly forms a thin oxide and passivates. This oxide formation is sufficiently rapid that it can occur during the rinsing cycle. Such passive nickel does not cover properly in chrome plating solutions. Other deposits, particularly gold, do not adhere properly and may peel. Parts that are stored after nickel plating are almost certain to be passive. The surface must therefore be activated before electrodeposition. The severity of the activation depends on the degree of oxidation that has taken place. Wide variations in the degree and speed of passivation during rinsing can occur as a result of variations in the rinsing water, specifically, its chlorine content, degree of aeration, dwell time, temperature, etc. As a result, an activation treatment that is completely effective in one plant may be totally inadequate in another apparently similar plant. The following activation steps may be used alone or in combination:

1) Cathodic treatment in a high current-density alkaline electrocleaner
2) Treatment in strong hydrochloric acid (50 vol % or stronger)
3) Cathodic treatment in dilute sulfuric acid, with or without added fluoride salt, or a properly formulated powdered acid
4) Treatment in 5 – 10 vol % sulfuric acid with an addition of 3 – 5 g/L (0.4 – 0.7 oz/gal) potassium iodide (KI)
   In this case, activation depends on the development of iodine. In a freshly made solution, there is usually enough oxygen dissolved in the water to develop the required initial amount. However, sometimes a trace of chromic acid (CrO_3) or another oxidizer is added to initiate action.
5) Treatment in one of the numerous proprietary activators available from suppliers
6) Use of the Woods nickel strike (see Section 5.3.7), the most certain method

In all cases following activation, careful attention must be paid to rinsing procedures so that the nickel does not repassivate during the rinsing cycle.

5.3.9. Nonmetallic Substrates

The most common nonmetallic substrate plated is plastic, in particular ABS type molding material (acrylonitrile-butadiene-styrene copolymers, → Rubbers, Synthetic). This plastic may be prepared for plating with the following cycle:

1) Condition the surface with a strongly oxidizing "etch" or "conditioner" to deglaze the surface and render it hydrophilic.
2) Sensitize and activate the surface by successive immersion in tin(II) chloride in hydrochloric acid, a rinse, palladium chloride in hydrochloric acid, and a rinse or by successive immersion in a palladium/tin(IV) chloride colloidal sol, a rinse, an acid accelerator, and a rinse. Either system produces on the surface a thin film of palladium, which acts as a catalyst for a subsequent electroless deposit.
3) Apply a thin film of electroless copper or electroless nickel.
4) Plate with whatever deposit is required.

Other plastics, including SAN (styrene-acrylonitrile copolymer), DAP (diallyl phthalate), polycarbonate, polypropylene, polyethylene, and polyimides, may also be plated. The pretreatment cycle varies only in the conditioning treatments used in the first step.

For electroforming, wax or silicone rubber mandrels may sometimes be used. These can be made hydrophilic by treating with a mixture of chromic and sulfuric acid and silvered with a silver reducing solution, or treated as above starting at Step 2. Glass and nonporous ceramics may be handled similarly. Wax may sometimes be plated by brushing the surface with finely powdered graphite or a metal powder to impregnate the surface with a continuous conductive coating and then plating directly. The graphite or metal powder must be completely free of oil or soaps.

Porous materials such as leather (baby shoes), natural objects (starfish, shells, wood, etc.), and some ceramics must first be sealed with several coats of a suitable polymer such as shellac,

polyurethane, or epoxy varnish. They can then be treated as outlined for wax. They can also be brushed with graphite or powdered metal as for wax, provided the conductive material is applied before the final layer of sealer is completely dry.

6. Equipment

6.1. Tanks

Tanks for containing the various solutions must be made out of suitably corrosion resistant materials. *Alkaline cleaner tanks* are generally made from steel and contain bottom drains to permit easy drainage and replacement. Because of problems with bipolar shorts in these tanks, it is becoming increasingly common to place insulating layers between the anodes and the tank wall and on the bottom of the tank. For the same reason, lined tanks are sometimes used. The tank should also be equipped with a dam overflow to permit removal of oil and grease by overflowing the surface of the tank. Use of an exterior grease trap and recycling system to extract these oils and greases from the cleaner offers the benefit of extended cleaner life and reduced levels of hexane-soluble materials in the waste stream, of special importance if the alkaline waste stream is to be discharged to a sewer system. Provision must be made to heat the tank to the proper operating temperature with heating coils or electrical heaters, or, if a recirculating system is used, an external heat exchanger.

Electrocleaner tanks must be provided with anode and cathode rods and suitable electrodes. It is a mistake to use the tank itself as the electrode. Poor current distribution results. For most electrocleaners, steel electrodes are adequate, and if these are perforated, better current distribution is obtained.

Rinse tanks may be made of mild steel following alkaline solutions, although the use of stainless steel or plastic is becoming increasingly common. After acid solutions, the tanks must be of suitably resistant material. A single rinse following a treatment tank is rarely adequate. Double and triple rinsing is more effective. If the tanks are interconnected to flow from one to the other in the opposite direction to the direction of work travel (countercurrent rinsing), the required water flow will be tremendously reduced, with corresponding decreases in waste treatment costs, water costs, and sewer charges. Formulas for calculating the required number of rinses and the economics of using parallel feed or series feed (counterflow) are available. The savings in rinse water by using series feed are ca. 98.6 % for a two tank system and ca. 99.7 % for a three tank system. The rinse tanks should be provided with bottom drains, dam overflows, and provision for agitation. Low pressure air is a convenient and effective method of vigorously agitating the rinse water.

Tanks for the acid solutions must be made of suitably resistant material: plastic-lined steel, stainless steel, formed reinforced plastic, or similar materials. These tanks usually have no drains or overflows.

Tanks for plating solutions may be constructed of plain steel for many alkaline and cyanide solutions. For other solutions they must be constructed of suitably resistant materials or lined with a suitable plastic material. Linings should be certified as acceptable by the supplier of the brightener system, or tested before installation. Many such lining materials contain plasticizers or residual curing materials that may leach into the plating solution, causing organic contamination. Even for approved materials it is good practice to leach a new liner with a dilute solution of approximately the same pH as the plating solution and at the operating temperature of the plating bath for 24 – 48 h before use.

The tank must be equipped with anode bars and supports and a cathode or work bar unless intended for use on an automatic or semiautomatic machine that provides contact for the racks as part of the operating mechanism. Provision must also be made for heating and filtering the solution as specified in the operating instructions. Plating tanks do not usually have bottom drains because of the danger of loss of solution should the drain valve or cap fail.

The plating tanks are generally constructed of heavier gauge materials than other tanks in the line because of the higher specific gravity of the plating solution and the weight of the anodes which must be supported by the anode bars. Tanks over ca. 125 cm (4 ft) in length and over 90 cm (3 ft) deep should be reinforced at the vertical midpoint and at the tank top to prevent bowing of the tank from hydrostatic pressure.

6.2. Power Supply

The d.c. power supplies consist of a controller for the incoming voltage, a step-down transformer, and a rectifying semiconductor array. The total package is commonly referred to as a *rectifier.*

All electrified tanks should be provided with rectifiers of suitable capacity of voltage and amperage. These rectifiers should generally be sized to allow ca. 50 % expansion over the estimated amperage requirement as a safety factor. If at all possible, the rectifiers should be equipped with remote controls so that they can be located in a separate room with a clean, noncorrosive atmosphere. The controls should be located in a convenient area near the production tanks. Alternatively, the rectifiers can be located near the tanks, and clean air ducted to the ventilation intakes. Totally enclosed water-cooled rectifiers are also available. Most three-phase rectifiers have a ripple factor of ca. 5 %, which is satisfactory for most plating solutions. Where a lower ripple factor is required, as with some gold and chrome plating, it may be necessary to filter the output. Tap switch or variable transformer controls are completely satisfactory for most operations where operators are present to make necessary adjustments. At added expense, controls can provide operation at constant voltage, constant current, ramp start, or other desired settings.

6.3. Heating Equipment

Many of the solutions used require heating to a specified operating temperature. The final rinse is usually heated to facilitate drying. Heating may be done with steam, electrical heaters, or in some cases by direct-fired gas heating. For steam heating, coils or flat welded plates, Platecoils, may be used. The coil should be sized to provide a maximum heat-up time of four hours. In calculating the heat input, required consideration must be given to the total volume of the tank, the tonnage of work processed, and the specific heat of the substrate being processed. With the exception of highly concentrated solutions, such as some chemical polishing and electropolishing solutions, sufficient accuracy is obtained by assuming that the specific heat of the solution is the same as water. Automatic temperature controllers are highly desirable. If heaters are located in the tank, they are preferably located along one of the tank walls, well out of the way of anodes and racks. If the heater is located ca. 5 – 10 cm away from the tank wall, 15 – 30 cm above the bottom of the tank, and an equal distance below the surface of the solution, the natural convective chimney produced circulates cold solution from the bottom of the tank up past the heater, producing a rolling action of the solution and a much more even temperature distribution within the tank. For cleaners and similar alkaline solutions, mild steel coils or heating plates are adequate. For corrosive solutions, stainless steel, titanium, lead, or graphite may be used. Also available are plastic heating coils of various types, generally made of Teflon. Corrosion-resistant electrical heaters are also available, including quartz and steel clad with titanium or lead.

6.4. Ventilation

Local health regulations require that many solutions be ventilated, and in many cases specify the air velocity across the surface of the solution to the ventilation slot. Other local restrictions require that certain contaminants be scrubbed from the air stream before it can be released to the atmosphere. Still other regulations may require that certain types of fumes, e.g., cyanide and acid, be removed by separate systems. These regulations determine the type of exhaust and/or scrubbing systems.

Aside from these requirements, good practice requires that for operator comfort, prevention of structural corrosion, and protection of incoming or finished work, certain other tanks should also be ventilated. Any tank that releases steam or even mildly acid fumes should be ventilated.

Under normal conditions, the ventilation system is shut down when production ends. However, it will be some time before all the heated tanks cool sufficiently to stop releasing steam and other vapors into the now more or less sealed atmosphere. Extreme corrosion of structures and stored parts can result. An auxiliary system, e.g., a relatively small exhaust fan located near or in the roof is most valuable in removing these

fumes and preventing damage. This fan should either be left on continuously or interlocked with the main exhaust system so that it comes on when the main system is shut down.

6.5. Filters

Optimum results are obtained from solutions that are kept scrupulously clean. Filtration is therefore recommended for almost all plating baths. The filter should be sized to provide a flow rate that completely filters (turns over) the bath in ca. 30 min. This assures that even when the rate slows because of cake buildup an adequate filtration rate is maintained. The use of the finest-pore filtering media that provides acceptable periods between filter cleaning is suggested. This can be supplemented by the use of a filter aid such as diatomaceous earth, glass or polyester fibers, etc. As with tank linings, these filter aids should be approved, or tested before use, since they can sometimes introduce harmful materials into the plating solution.

6.6. Agitation

Many forms of mechanical agitation have been developed, particularly for plating tanks, including oscillation of the cathode bar by some type of offset cam. Modern bright solutions, however, require much more vigorous agitation than can be obtained in this way, and air is used for this purpose. Sometimes solutions require air agitation that lifts the solution as much as 15 cm above the normal quiescent level of the tank. Because of the possibility of introducing contamination from entrapped oil or other materials, high pressure air should not be used for this purpose. The air should be supplied by a low-pressure blower located where it picks up clean air, and with the intake carefully filtered. Plastic is the preferred material for distribution; the plastic distribution pipes should be located a slight distance off the bottom of the tank so that the rising solution currents produced do not pick up debris from the bottom of the tank and carry it up onto the work, causing roughness. The distribution pipes should be weighted or otherwise held down so that they do not float and tangle with the racks in the tank.

6.7. Water Supply

The water supply must be adequate and must be readily available where needed. Permanent piping through the plating room should have valved outlets at frequent intervals. Tanks that must be dumped and remade at regular intervals, i.e., cleaners and similar solutions, should have high flow filling systems so that excessive time is not lost waiting for the tank to be filled.

If the water is high in iron, chlorides, or hardness, consideration should be given to pretreatment. For makeup to plating solutions, deionization should be considered. Softening may be helpful for at least the final rinses and the rinses ahead of the plating solutions. Untreated water can be used for noncritical rinses.

6.8. Racks

Plating racks must be designed to fit the equipment being used. The working area of the rack should have a vertical length such that, when the rack is fully loaded with parts and immersed in the treatment tanks, the uppermost piece on the rack is 10 – 15 cm below the surface of the solution. The bottom of the lowest piece on the rack should be at least 15 – 30 cm above the bottom of the tank. If air agitation is used, this clearance should be between the bottom of the rack and the distribution system for the air agitation. The width of the work area should allow adequate clearance from the tank walls. On automatic equipment, there is always some movement of the racks during transfer, and the clearance must allow for this motion, or the racks must be attached to the transfer mechanism sufficiently firmly to minimize the motion. On return-type automatic machines, this dimension is particularly critical, since as the racks turn at the end of the machine, they must not catch on the tank anodes. Therefore, the turning radius of the machine becomes a factor in determining the width of the work area.

The main splines of the rack must have sufficient strength to support the work load, and be of sufficient cross section to carry the required current. Although the current carrying capacity of copper is normally given as 160 A/cm^2

(1000 A/in^2), for copper racks coated with an insulating and protective plastisol coating a safer estimate is 120 A/cm^2 (770 A/in^2). To determine the current carrying capacity of other metals, the factors are 0.6 for aluminum, 0.25 for brass, 0.12 for steel, 0.18 for phosphor bronze, 0.023 for stainless steel, and 0.03L for titanium. Because of the relationship of current carrying capacity and weight, it is common practice to make the main portion of the rack out of copper. Construction must be very firm and assure good contact between rack members, so crosspieces are usually riveted and either brazed or soldered with a high-temperature solder.

Rack tips may be designed to provide positive spring contact for holding the parts. If the parts are sufficiently heavy and have a suitable hole for hanging, the rack tips may be formed as a simple loop over which the part can be hung. Since there is considerable movement of the parts as the rack is immersed in the solution, there should be sufficient reverse curve to these loops to prevent the parts from floating off the rack as a result of this movement. In some cases, e.g., silver plating forks and spoons, the parts may be inserted in a loop or slot, the dimensions of which are such that the part is held on the rack by some projection or increase in diameter, width, etc. For this system to function properly, the center of gravity of the part must be below the supporting contact on the rack. To obtain sufficient strength and springiness to properly hold the parts, rack tips are generally made of phosphor bronze or, for added resistance to processing and metal stripping solutions, stainless steel or titanium.

To restrict current loss and wasted metal, racks are usually given a heavy coating of an insulating plastisol with only the tips left exposed. Proper maintenance, handling, and storage of racks is critical since any damage to this coating can result in unwanted plating, corrosion of the main structure of the rack, and cross-contamination of solutions due to entrapment of solutions by the damaged areas.

If the rack tips are allowed to build up with plated metal, defective parts may result from shadowing or thieving of current by the rack tip, additional metal is wasted plating the enlarged tip, and the danger of cross-contamination by solution entrapment increases because the plated metal usually is coarse and porous in structure. Additionally, the plated layer may reduce the flexibility of the tip, causing breakage and loss of productive capacity. Therefore, racks should be stripped on a regular basis to maintain the tips in a clean condition. Damaged or broken tips should be replaced promptly to avoid the loss of productive capacity.

6.9. Waste Disposal

The consideration of waste disposal is a must in designing a modern plating plant. While there may be a few places in the world where unrestricted discharge of plating wastes to streams or sewage systems is still allowed, this situation will not continue. While the design of the system depends on the production capabilities of the plant, on whether the effluent is to be discharged to a sewage system or a stream, and on the local restrictions, the basic procedure for designing the layout is similar. Furthermore, the design of the system and layout should be part of the original design of the facility. Retrofitting an existing plant with a suitable system is almost always more expensive than including it in the original design.

Waste disposal is inextricably interlocked with water supply and demand. If discharge is to a public treatment works, charges are often based on volume of effluent. Even in the absence of such charges, the cost of treatment for a given total quantity of contaminant is higher for low concentrations, since a greater volume must be treated. Water conservation and the determination of the methods of obtaining the lowest possible volume of effluent is therefore an essential part of the design process. The following steps constitute a broad outline of the design process:

1) Determine and become familiar with local restrictions, including planned changes in the requirements and the timetables for implementation.
2) Review the processes planned for installation.
   Determine the expected dragout from each process. Based on this information and the available space, determine the possibility of using reclaim tanks, the number of counterflow rinse tanks that may be used, and the possibility of counterflowing rinses for multiple use, e.g., use of the last counterflowed

rinse from an acid tank as feed for a first counterflowed rinse for a cleaner. From this information determine the lowest flow rates that can be used and calculate the total effluent flow. Determine which effluents must be treated separately, and those that can be safely mixed and treated in combination.

3) Proceed with the preliminary layout of the plating line.
4) Overlay on this preliminary layout the effluent collection troughs or piping, the location of preliminary treatment steps, and any collection sumps or other equipment that must be located in the plating area, then refine these layouts until acceptable.

6.10. Site Location

The site must be chosen on the basis of:

1) Convenience to other manufacturing operations.
2) Availability of services including power, heating facilities, sewer connections or waste treatment facilities, water supply, and ventilation.
3) Headroom to allow for hoists or automatic equipment and overhead services such as ventilation ducts.
4) Subsurface conditions to permit proper construction of foundations and sumps for waste collection.
5) Location of storage areas.

Since many local building codes require that exhaust stacks must extend a specified distance higher than any building within a specified radius, meeting these restrictions may pose special problems.

6.11. Layout

The layout of the plating facility is of particular importance if operation is to be efficient.

There are two basic layouts: U-shaped, in which the work enters and leaves the plating facility at the same point, and flow-through, in which the work enters at one end of the plating line and leaves at the opposite end. For manually operated lines, the U-shape is generally more efficient provided access to the storage areas is adequate. For automatic or hoist-assisted operation, the choice is in part dependent on the type of equipment. In either case, a complete work flow analysis should be carried out to determine that movement of work and/or operators within the facility does not result in areas of congestion or unnecessary transfer of work from operation to operation.

7. Stripping of Deposits

Plated deposits must be stripped for two reasons: (1) to repair and replate rejects or (2) to maintain the rack contacts in optimum conditions. The cost of stripping and replating a part is roughly four times the original cost of plating plus the cost of the stripping agent used. From the viewpoint of economy, therefore, parts should not be stripped unless one of the following criteria can be met: (1) the cost of stripping and replating is less than the value of the parts, (2) the parts are from a limited production run and the cost of setting up the production equipment to run additional parts is greater than the value of the parts, or (3) there is a critical need for the parts and additional parts cannot be obtained in time to satisfy the need. In most other cases, the parts should be scrapped.

Similar reasoning can be applied to the frequency of stripping for rack contacts. Tips generally do not need to be stripped after each pass through the process. (An exception might be hard chromium plating where very heavy thicknesses are applied.) There is an optimum frequency of stripping based on the rate of buildup of plating on the tips, the thickness at which problems occur, and the cost of stripping. If the parts are being used on full automatic equipment and the stripping process can be incorporated in the process cycle, stripping on each pass is both automatic and economical.

The solutions used for stripping depend on the nature of the plated deposit, the substrate, and whether stripping is done chemically or electrolytically. In general, electrolytic stripping is less expensive, but may require considerable investment in equipment, greater attention, and more time. In evaluating any stripper or stripping method, the ease of waste disposal and/or

recovery of the stripped metals should also be considered.

Immersion in *nitric acid* strips most metals from substrates resistant to the action of nitric acid, such as stainless steel, titanium, and aluminum. Nitric acid (42 °Bé) diluted with water (1 part water and 1 or 3 parts acid) is commonly used, particularly for stripping racks. Proprietary additives are available that increase the rate of stripping by 35 – 50 times.

Organic oxidizing agents such as sodium metanitrobenzoate or sodium metanitrobenzenesulfonate, which are called *nitroaromatics,* may be used with various solubilizing agents such as sodium cyanide or sulfuric acid to strip many metals from various substrates.

There are numerous electrolytic strippers, which often are specific in their action, either as to the metal stripped or the substrates not attacked. Some of the more effective are proprietary. There are also immersion strippers that are specific to particular metals and substrates.

In the past, stripping has not been considered an essential part of the plating operation. Stripping facilities therefore have often been makeshift, poorly designed, and poorly located. It should be recognized that stripping, either for maintenance of racks and other equipment or for reclaiming of rejects, is an essential part of the operation. The design, location, and installation should be given the same engineering attention as in planning a new plating line.

The methods for stripping specific deposits are given at the ends of the chapters describing the specific deposits and plating solutions.

8. Specific Deposits and Plating Solutions

8.1. Aluminum Deposits

Aluminum [*7429-90-5*] cannot be deposited from aqueous solutions because the deposition potential for aluminum is higher than the deposition potential for hydrogen. It can, however, be deposited from solutions of aluminum salts in organic solvents. Special equipment must be used because of the flammability of the solvents and the need to completely exclude even trace amounts of water from the system. While the process has had some degree of success, and while some equipment has been developed to permit continuous operation, it is still experimental. As a result, its applications are rather limited.

8.2. Antimony and Arsenic Deposits

Numerous baths for both antimony [*7440-36-0*] and arsenic [*7440-38-2*] have been listed in the literature. None is completely satisfactory and in the case of antimony the metal may sometimes be deposited in the explosive form. Modern usage generally involves cyanide-based baths containing arsenic or mixtures of arsenic, antimony, and other modifying metals such as copper to produce *antique finishes,* particularly on brass and copper. The most effective of these solutions are marketed as proprietaries.

An arsenic bath based on cyanide and caustic soda was formerly used to produce an antique pewter finish. Toxicity and waste disposal considerations have resulted in the bath being discontinued in most cases, and most handbooks no longer list the formula for the bath. The same considerations have forced the withdrawal or reformulation of many of the proprietary solutions as well.

8.3. Cadmium Deposits

Cadmium [*7440-43-9*] has been used to protect steel from corrosion because of its appearance, apparent greater resistance to marine atmospheres, the fact that the volume of the corrosion salts is less than that of zinc, and in some cases because its low coefficient of friction allows tightening fasteners with torquing equipment without overstressing the fastener. Cadmium is, however, highly toxic and has therefore come under close scrutiny by environmental protection agencies. As a result, its use has declined considerably, and replacements are used except for certain military applications. The most common bath is the cyanide bath:

Cadmium oxide	30 g/L (4.0 oz/gal)
Sodium cyanide	100 g/L (13.1 oz/gal)
or	
Cadmium cyanide	40 g/L (5.1 oz/gal)
Sodium cyanide	75 g/L (10.1 oz/gal)
Sodium hydroxide	20 g/L (2.5 oz/gal)
Control	
Cadmium metal	26.5 g/L (3.5 oz/gal)
Total cyanide (as NaCN)	100 g/L (13.1 oz/gal)
Ratio total NaCN : Cd	3.75 : 1
Temperature	20 – 35 °C
Cathode C.D.	0.5 – 5.0 A/dm^2 (5 – 50 A/ft^2)
Anode : cathode ratio	2 : 1
Efficiency	90 – 95 %

Despite its relatively high efficiency, cadmium plating has been proven to be the source of hydrogen embrittlement in susceptible steels. To overcome this, solutions containing sodium nitrate have been used with mixed success. However, baking before and after plating remains the best and safest procedure for eliminating failures due to hydrogen.

Cadmium cyanide solutions are deleteriously affected by both metallic and organic contamination. Copper and zinc not only affect appearance, but also reduce corrosion resistance. Tin produces a haze at a current density of ca. 1.5 A/dm^2. Metallic contamination may be removed by dummying at ca. 0.5 A/dm^2 for zinc, copper, and lead and at 1.5 A/dm^2 for tin. Temporary correction may be obtained by adding sodium or potassium polysulfide. Treatment with cadmium dust (toxic, avoid inhalation) followed by filtration also removes metallic contamination. Organic contamination may be removed by treatment with activated carbon. For some military applications, in an effort to reduce the possibility of embrittlement to even safer levels, an acid fluoborate solution is sometimes used:

Cadmium fluoborate	240 g/L (32.2 oz/gal)
Fluoboric acid	to proper pH
Boric acid	22.5 g/L (3.0 oz/gal)
Ammonium fluoborate	60.0 g/L (8.0 oz/gal)
Control	
Cadmium	95.0 g/L (12.6 oz/gal)
pH, colorimetric	3.0 – 3.5
Temperature	10 – 38 °C
Cathode C.D.	3 – 6 A/dm^2 (30 – 60 A/ft^2)
Anode : cathode ratio	2 : 1

Licorice (1.0 g/L, 0.134 oz/gal) may be used as a grain refiner, or proprietary brighteners may be added.

Cadmium may be stripped by immersion using nitric acid, a nitroaromatic with cyanide, 50 vol % hydrochloric acid (inhibited), or ammonium nitrate at a concentration of 120 g/L (16 oz/gal). Reverse current electrolytic strippers may also be used.

8.4. Chromium Deposits

There are two main uses for the deposits of chromium [*7440-47-3*]: *decorative chromium,* which is normally restricted to very thin deposits, and *hard chromium,* which is applied at much greater thicknesses. The essential differences between the two deposits are the thicknesses and the conditions of operation. Hard chromium is most often applied at a slightly higher temperature than decorative layers and is not as bright. Many shops alternately use the same solutions for both purposes merely by adjusting the temperature and the current densities.

Chromium is mainly deposited from baths that contain chromic acid (chromic anhydride, $Cr^{VI}O_3$) and a catalyst. The original catalyst was sulfate ion (SO_4^{2-}) provided by sulfuric acid and was added to the bath in amounts that would produce a ratio of 75 : 1 to 125 : 1 (CrO_3/SO_4^{2-}). There are two general classes of solution, those that operate at ca. 250 g/L (33 oz/gal) and those operating in the range of 375 g/L (50 oz/gal). Other anions can also function as catalysts, particularly fluoride. Many proprietary solutions have been developed based on mixtures of fluoride salts and sulfuric acid. If the fluoride salt has a limited, temperature-dependent solubility, an excess may be added to produce a self-regulating solution. By properly choosing the catalyst mixture and the operating conditions, the porosity and the degree of cracking of the chromium deposit may be controlled. For automotive use, two successive layers of differing deposit structure are sometimes used to prevent concentration of corrosive effects, thereby improving corrosion resistance. Where hard chromium is to be used as a lubricated bearing surface, the crack pattern may be specified.

The bath is operated at a temperature of 32 – 43 °C for plating decorative deposits, and 37 – 65 °C for hard chromium. Cathode cur-

rent density is 10 – 20 A/dm^2 (100 – 200 A/ft^2) for decorative work and 15 – 35 A/dm^2 (150 – 350 A/ft^2) for hard chromium. Some trivalent chromium must always be present, and optimum results are obtained when the Cr(III) concentration is approximately 5 % of the Cr(VI) concentration.

The process is sensitive to chromic acid concentration, catalyst ratio, temperature, and current density. In the past, it was reported that metallic impurities could be tolerated to a level of 20 g/L calculated as the oxides of the contaminants before the bath became unstable and difficult to control. The trivalent chromium must be included in this calculation. More recent reports have indicated that instability occurs at levels as low as 1.5 g/L of contaminant. This is particularly true of baths using mixed catalyst systems that include fluoride. Many anions are deleterious, but chloride ion is particularly harmful. Fortunately, it can be readily removed by dummying at high current density, the Cl^- being oxidized to Cl_2 at the anode and evolved from the solution.

For many years, efforts have been made to develop baths based on Cr(III) to reduce disposal problems and improve the efficiency of the bath. While several such baths were developed, the color of the deposit was somewhat dark and did not favorably compare with the blue-white color of the deposits obtained from Cr(VI) baths. Public acceptance was therefore limited despite the other advantages of the baths. A newer bath produces deposits that closely match the desired color. It has high efficiency, can be replated without stripping, and has much better throwing power than the Cr(VI) baths. It does require anode baskets that surround the anodes with a permeable membrane that prevents the Cr(VI) formed at the anode from migrating into the solution, where it acts as a poison. This bath is available only as a proprietary solution (Envirochrome).

Chromium may be stripped by immersion in 22° Bé hydrochloric acid diluted 1 : 1 with water, in warm 50 vol % sulfuric acid, or a heated solution of sodium hydrogen sulfate. It may also be stripped electrolytically in a solution of sodium hydroxide, or a strongly alkaline electrocleaner. Properly formulated alkaline electrocleaners are to be preferred for heavy deposits since they give somewhat higher stripping rates and greater protection against etching of the substrate due to localized reduction of pH.

8.5. Cobalt Deposits

Methods of electrodepositing cobalt [*7440-48-4*] have been known for over 130 years. Because of the lower cost of nickel, cobalt is not often used except as an alloying agent with nickel, especially in times of nickel shortages. The interest in such solutions has remained high, particularly in the Eastern Bloc countries.

The solutions used for depositing cobalt are essentially similar to those used for nickel. Brightening systems are also similar. Electrodeposited cobalt is somewhat more highly stressed than nickel, but this can be relieved by the proper choice of organic stress-relieving compounds. This choice is more limited in range than for nickel. The variation in cobalt ion concentration is much greater than is the case with nickel.

Chloride bath	
Cobalt chloride ($CoCl_2 \cdot 6\,H_2O$)	90 – 105 g/L
Boric acid	60 g/L
pH	2.5 – 3.5
Temperature	49 – 54 °C
Cathode C.D.	3.0 – 4.0 A/dm^2
Sulfate bath	
Cobalt sulfate ($CoSO_4 \cdot 7\,H_2O$)	330 – 565 g/L
Boric acid	30 – 45 g/L
Cobalt chloride (optional)	45 g/L
Sodium or potassium chloride (optional)	17 – 25 g/L
pH	3.0 – 5.0
Cathode C.D.	2.15 – 5.0 A/dm^2

Cobalt may be stripped by immersion using a nitroaromatic (sodium metanitrobenzenesulfonate) and cyanide, a nitroaromatic and dilute sulfuric acid (copper and copper alloys only), or nitric acid. Electrolytically it may be stripped in strong sulfuric acid, a solution of sodium nitrate at a concentration of 360 g/L (48 oz/gal), or proprietary materials.

8.6. Copper Deposits

Copper [*7440-50-8*] is widely used as an undercoating for various other plated coatings on both metals and plastics, on printed circuit boards for through-hole connection, as a stop off in selective carburizing of steel, in electroforming, for

preparing electrotypes, and as a decorative coating, either in its natural state or chemically colored and selectively buffed to create a wide range of decorator finishes. There are numerous types of copper plating solutions available: cyanide – regular, bright, and high speed; pyrophosphate; acid – sulfuric acid regular and bright; and fluoboric acid.

The *copper cyanide baths* have excellent macro-throwing power and are also often used as strike baths before applying other deposits that might produce immersion coatings. Typical cyanide copper plating formulations are shown in Table 1. The Rochelle salt bath produces a finer, denser crystal structure, which is semi-lustrous and can be partially brightened with proprietary additions. Air agitation may be used.

Generally the high-efficiency sodium and potassium baths require a preliminary strike in the standard or Rochelle salt bath. In the two high-efficiency baths, the use of specific brightener systems may require somewhat different concentrations and/or operating conditions. The deposits from the high-efficiency baths may be made full bright with proprietary additions.

Copper cyanide solutions are among the very few solutions in which the cyanide concentration is controlled as "free" cyanide, i.e., the cyanide in excess of that required to completely complex the copper. This is a highly variable quantity. The method of analysis also affects the results. The concentrations given in Table 1 should be accepted as preliminary values to be adjusted as experience dictates. Once the proper value for the operational conditions and the analytical method has been established, no problems should be encountered until operating conditions or analytical method is altered.

Lead, zinc, cadmium, and arsenic contamination may cause poor deposits, but can be removed by dummying. Chromium(VI) may cause blotchy deposits or peeling, or even stop plating completely. Chromium(VI) should be reduced with sodium hydrosulfite or a similar reducing agent, and the $Cr(OH)_3$ removed by filtration. Organic contamination may cause hazing or pitting, but can be removed by treatment with activated carbon.

High carbonates, >90 g/L (12 oz/gal), can cause porous and/or rough deposits. Carbonates may be removed from sodium baths by selective crystallization at temperatures around 0 °C and from all baths by precipitation with calcium or barium salts (substituting $Ca(OH)_2$ or $Ba(OH)_2$ for NaOH or KOH additions, using $Ba(CN)_2$ in place of NaCN or KCN) followed by filtration, or by the use of special ion-exchange resins.

Copper pyrophosphate bath is used for printed circuits or for bright plating. In a modified form the copper pyrophosphate bath may be used as a copper strike where cyanide cannot be tolerated. The unmodified bath requires a preliminary strike.

Copper pyrophosphate	345 g/L (46 oz/gal)
Potassium hydroxide	18 g/L (2.4 oz/gal)
Ammonia (29 wt %)	1 mL/L (1.24 fl oz/gal)
Analysis	
Copper	22.5 g/L (3.0 oz/gal)
Pyrophosphate	157.5 g/L (21.0 oz/gal)
Ammonia	2.0 g/L (0.25 oz/gal)
Ratio P_2O_5/Cu	7 : 1
Cathode C.D.	1 – 7.5 A/dm^2 (10 – 75 A/ft^2)
Agitation	mechanical or air
Anode : cathode ratio	1 : 1 – 2 : 1

Acid copper plating is used for producing heavy deposits for electroforming, for backup plating of electrotypes, and for applying heavy, bright deposits to plastics, zinc-base die castings, some steel automotive parts, and similar applications. Acid copper has poor macro-throwing power, but when run with the proper brighteners has outstanding micro-throwing power. This gives the bath excellent leveling and defect hiding characteristics. Additionally, although the copper is deposited from the divalent (Cu^{2+}) rather than the monovalent (Cu^{+}) state, and therefore requires twice the number of electrons that cyanide copper does, the low operating voltage results in watt-hour savings, reducing power costs.

The standard formula contains:

Copper sulfate ($CuSO_4 \cdot 5H_2O$)	195 – 250 g/L (26 – 33 oz/gal)
Sulfuric acid	30 – 75 g/L (4 – 10 oz/gal)
Control	
Copper	40 – 50 g/L (5.3 – 6.6 oz/gal)
Sulfuric acid	30 – 75 g/L (4 – 10 oz/gal)
Temperature	21 – 49 °C
Cathode C.D.	2 – 10 A/dm^2 (20 – 100 A/ft^2)

A copper strike must be used for steel and zincbase die castings. It is occasionally required for brass as well. For operation at high current densities, moderate to vigorous agitation is required. Some of the proprietary brightener sys-

Table 1. Typical cyanide copper plating formulations

Component or condition	Baths			
	Standard bath, used as a strike or for thin deposits, 1.25 – 2.5 µm	Rochelle salt bath, used for strikes or thin deposits, 1.25 – 7.5 µm	High-efficiency sodium bath, used for bright or heavy deposits, 0.75 – 50 µm	High-efficiency potassium bath ∗, used for bright or heavy deposits, 0.75 – 50 µm
Copper cyanide (CuCN)	22.5 g/L (3.0 oz/gal)	26.0 g/L (3.5 oz/gal)	120.0 g/L (16.0 oz/gal)	60.0 g/L (8.0 oz/gal)
Cyanide, sodium or potassium ∗ (MCN)	34.0 g/L (4.5 oz/gal)	35.0 g/L (4.6 oz/gal)	135.0 g/L (18.0 oz/gal)	94.0 g/L (12.5 oz/gal)
Carbonate, sodium or potassium ∗ (M_2CO_3)	15.0 g/L (2.0 oz/gal)	30.0 g/L (4.0 oz/gal)	15.0 g/L (2.0 oz/gal)	15.0 g/L (2.0 oz/gal)
Hydroxide, sodium or potassium ∗ (MOH)	to adjust pH	to pH 10.2 – 10.8	30.0 g/L (4.0 oz/gal) pH > 13	42.0 g/L (5.6 oz/gal) pH > 13
Rochelle salt NaK ($C_4H_4O_6 \cdot 4H_2O$)		45.0 g/L (6.0 oz/gal)		
Analysis				
Copper	16.0 g/L (2.1 oz/gal)	19.0 g/L (2.5 oz/gal)	85.0 g/L (11.3 oz/gal)	42.0 g/L (5.6 oz/gal)
Free, cyanide ∗∗	7.5 g/L (1.0 oz/gal)	6.0 g/L (0.8 oz/gal)	4.0 g/L (0.5 oz/gal)	7.5 g/L (1.0 oz/gal)
Temperature	32 – 43 °C	54 – 71 °C	77 – 82 °C	77 – 82 °C
Cathode current density	1 – 1.5 A/dm^2 (10 – 15 A/ft^2)	2.0 – 4.0 A/dm^2 (20 – 40 A/ft^2)	3 – 6 A/dm^2 (30 – 60 A/ft^2)	3 – 6 A/dm^2 (30 – 60 A/ft^2)
Voltage	6	6	6	6
Anode : cathode ratio	2 : 1	2 : 1	2 : 1	2 : 1
Efficiency	30 %	50 %	100 %	100 %

∗ The more expensive potassium cyanide, carbonate, and hydroxide are used only in the high-efficiency potassium bath.
∗∗ Free cyanide is the cyanide in excess of that required to completely complex the copper.

tems use very vigorous air agitation. Almost all of the bright systems require extremely careful control of the chloride ion concentration. Limits are generally in the range of 60 to 100 ppm, although some systems permit concentrations as high as 180 ppm. In some geographic areas, use of deionized water for bath makeup and replenishment is required.

Copper fluoborate baths are also occasionally used for electroforming and printed circuit manufacture. Copper fluoborate is available as a liquid concentrate. This concentrate is diluted with water to give a copper concentration of 60 – 120 g/L (8 – 16 oz/gal). Fluoboric acid is then added as required to maintain the pH at 0.6 – 1.7. The allowable current densities are somewhat higher than for the sulfuric acid bath.

Copper may be stripped by immersion in nitric acid, a nitroaromatic with sulfuric acid, a nitroaromatic with cyanide, or strong chromic acid solutions containing some sulfate ion. Electrolytically copper may be stripped in sodium cyanide – sodium hydroxide solutions, sodium nitrate solution at 360 g/L (48 oz/gal), and proprietary mixtures.

8.7. Gallium Deposits

There does not appear to be any industrial application of plating gallium [*7440-55-3*]. It has been reported that gallium may be deposited from alkaline solutions prepared by dissolving gallium hydroxide in an excess of sodium hydroxide.

8.8. Germanium Deposits

Germanium [*7440-56-4*] cannot be effectively deposited from aqueous solutions. Nonaqueous solutions and fused salt baths have been reported.

8.9. Gold Deposits

Gold [*7440-57-5*] is deposited for decorative purposes, as a contact material in electronics, and in a few cases as a lining for specialized

Table 2. Gold deposition: standard cyanide baths

Component or condition	Finish	
	Matte	Bright
Gold as the cyanide $KAu(CN)_2 \cdot 2H_2O$	2 – 12 g/L (0.25 – 1.5 oz t/gal)	4 – 16 g/L (0.5 – 2.0 oz t/gal)
Potassium cyanide	15 – 45 g/L (2 – 6 oz/gal)	15 – 90 g/L (2 – 12 oz/gal)
Potassium carbonate	0 – 45 g/L (0 – 6 oz/gal)	0 – 30 g/L (0 – 4 oz/gal)
Potassium phosphate	0 – 6 g/L (0 – 6 oz/gal)	0 – 45 g/L (0 – 6 oz/gal)
Potassium hydroxide	10 – 30 g/L (1.3 – 4 oz/gal)	10 – 30 g/L (1.3 – 4 oz/gal)
Brighteners		0.1 – 10 g/L (0.01 – 1.0 oz/gal)
Temperature	50 – 70 °C	15 – 25 °C
pH	11 – 13	
Cathode current density	0.1 – 0.5 A/dm^2 (1 – 5 A/ft^2)	
Agitation	moderate	rapid
Anodes	platinum, stainless steel, gold	platinum, stainless steel, gold

chemical reactors. Most decorative deposits are plated from the standard cyanide solution (Table 2), and the baths are often modified with alloying elements to produce various color effects. Deposits for electronics are generally plated from the neutral or acid baths, and the physical characteristics of the deposits are carefully controlled. Most of these solutions are proprietary and are supplied as prepared concentrates.

The standard cyanide solutions have good throwing power and retard the deposition of base metals. They therefore tend to hold their color well. For producing colored deposits for jewelry the following additions may be made:

14-Karat gold – copper or nickel up to 25 % of total metal.

Green gold – silver ca. 5 % of total metal with small amounts of arsenic or lead to intensify the green color or a ternary alloy of cadmium –copper – gold

Red gold – copper as required to obtain the desired color

White gold – nickel up to twice the gold content

The neutral and acid gold baths (Table 3) operate in pH ranges where cyanide solutions would normally release poisonous hydrogen cyanide gas. Because of the stability of the gold – cyanide complexes, however, this does not occur. These baths are almost all proprietary baths. The gold is supplied as a prepared salt mixture. The chelates used depend on the brightener system and the supplier of the bath, and are either patented or held as trade secrets. The acid golds produce the purest deposits.

Gold baths are especially sensitive to metallic and organic contamination. In many cases, economical purification methods do not exist. Contaminated solutions must be discarded and reclaimed. Strike solutions are used, therefore, not only to improve adhesion, but also to act as traps for contaminants that might be dragged through from the preplating cycle. All tank-mounted equipment should be either corrosion resistant or coated with corrosion-resistant materials to ensure that no corrosion salts from this equipment enters the tank. If copper anode and cathode bars are used, they are sometimes plated with a heavy coating of nickel or tin – nickel to prevent copper corrosion. Alternatively, stainless steel or titanium bars are sometimes used, but their lower conductivity requires a proportionate increase in size.

Since many of the gold baths are operated with insoluble anodes and gold concentration is replenished with gold salts, there is a buildup of salts, which eventually may interfere with operation. The supplier of the gold concentrates has often worked out procedures for removing these excess salts, and the supplier's recommendations should be followed closely.

Since many metals, silver in particular, diffuse through gold, a barrier layer of nickel is often used. Unless suitable care is exercised to prevent the nickel layer from passivating, the gold layer may fail to adhere. Prudence dictates that an activating step be used between the application of the nickel layer and the deposition of the gold. This may be an immersion in a strong HCl dip followed by rinsing; cathodic treatment in an alkaline cleaner followed by acid neutralization and rinsing; or the use of one of the many proprietary immersion activators available. Since the treatment required depends on the nickel solution used, the rinsing time, and the quality of

the rinse water, no one recommendation can satisfy all conditions. Some experimentation is required to determine the optimum procedure for each particular installation.

Gold may be stripped in several different ways, but the most effective in reclaiming defective parts are proprietary cyanide-based strippers, which do not attack nickel and copper substrates. Racks may be stripped in accelerated nitric acid solutions, which in most cases remove the gold as solid flakes by undercutting. The gold flakes may then be reclaimed from the stripping solution by skimming the floating particles or filtering the stripper. Some of the accelerators dissolve some of the gold, and in these instances the stripping solution should eventually be sent out for reclamation.

8.10. Indium Deposits

Indium [*7440-74-6*] is a soft, malleable, silvery-white metal with a low melting point (156 °C). It readily forms diffusion coatings with copper, cadmium, zinc, silver, lead, tin, and gold. The main application is as a coating on bearings to impart corrosion resistance. When used for bearing surfaces, the indium is often diffused into the substrate by heat treatment after plating. In electronic applications, indium is used as a low-pressure bonding material to join copper parts, and for low-temperature soldering as well as for corrosion proofing. Many codeposited alloys exist and are used for various specialized purposes.

The various baths differ slightly in deposit properties and operating characteristics. They may be used interchangeably, depending on the operator's preference.

Sulfate bath	
Indium sulfate ($In_2(SO_4)_3$)	45 g/L (6 oz/gal)
Sodium sulfate	9.75 g/L (1.3 oz/gal)
Control	
Indium	20.0 g/L (2.7 oz/gal)
pH, colorimetric	2.0 – 2.7
Temperature	21 – 60 °C
Cathode C.D.	0.5 – 5.0 A/dm^2 (5 – 50 A/ft^2)
Anodes	indium and platinum
Chloride bath	
Indium chloride ($InCl_3$)	58.0 g/L (7.7 oz/gal)
Potassium cyanide	158.0 g/L (21.0 oz/gal
Potassium hydroxide	40.0 g/L (5.3 oz/gal)
Dextrose	30.0 g/L (4.0 oz/gal)
Control	
Indium	30.0 g/L (4.0 oz/gal)
Total cyanide	140.0 g/L (18.7 oz/gal)
pH, colorimetric	> 13.0
Temperature	21 °C
Cathode C.D	1.5 – 3.0 A/dm^2 (15 – 30 A/ft^2)
Anodes	platinum, steel, stainless steel, graphite
Fluoborate bath	
Indium fluoborate	109.0 g/L (14.5 oz/gal)
Ammonium fluoborate	50.0 g/L (6.7 oz/gal)
Fluoboric acid	15.0 g/L (2.0 oz/gal)
Boric acid	26.0 g/L (3.4 oz/gal)
Control	
Indium	33.0 g/L (4.4 oz/gal)
pH, colorimetric	1.5 – 2.0
Temperature	21 °C
Cathode C.D.	0.5 – 2.5 A/dm^2 (5 – 25 A/ft^2)
Anodes	indium and platinum or graphite

To prevent excessive increase of indium concentration, some of the anode area should be insoluble.

8.11. Iron Deposits

The plating of iron [*7439-89-6*] is mainly used for salvaging worn parts and for specialized electroforming. It has the advantage that after deposition the composition can be altered by carburizing or nitriding to match the combination of ductility and surface hardness obtainable with steel. Some small-scale applications, e.g., the coating of soldering iron tips, also exist. While the properties and low cost of iron are attractive, the corrosiveness of the solutions increases the capital cost of the equipment so that other materials often become economically more attractive.

Even when operated at high temperature, the chloride bath produces the most highly stressed deposits. It is preferred for repair of worn bearing surfaces because of the stress-induced hardness and better wear resistance. The fluoborate bath produces the least-stressed deposits and therefore is often preferred for electroforming. The sulfate – chloride bath is a general purpose bath.

Sulfate – chloride bath	
Iron(II) sulfate ($FeSO_4 \cdot 7H_2O$)	248 g/L (33 oz/gal)
Iron(II) chloride ($FeCl_2 \cdot 4H_2O$)	30 g/L (4 oz/gal)
Ammonium chloride	22.5 g/L (3 oz/gal)
Control	
Iron	58.0 g/L (7.7 oz/gal)
pH, colorimetric	4.5 – 6.0
Temperature	38 °C
Cathode C.D.	5.0 – 10.0 A/dm^2 (50 – 100 A/ft^2)
Agitation	cathode rod and solution
Efficiency	95 – 100 %
Anodes	low carbon steel, bagged
Chloride bath	
Iron(II) chloride	300 g/L (40 oz/gal)
Calcium chloride ($CaCl_2$)	338 g/L (45 oz/gal)
Control	
Iron	84.0 g/L (11.2 oz/gal)
pH, colorimetric	1.0 – 1.5
Temperature	90 °C
Cathode C.D.	6.0 A/dm^2 (60 A/ft^2)
Other conditions are the same as for the sulfate – chloride bath.	
Fluoborate bath	
Iron(II) fluoborate	225.0 g/L (30.0 oz/gal)
Sodium chloride	10.0 g/L (1.3 oz/gal)
Boric acid	22.5 g/L (3.0 oz/gal)
Control	
Iron	55.5 g/L (7.4 oz/gal)
pH, colorimetric	3.0 – 3.7
Temperature	55 – 90 °C
Cathode C.D.	2.0 – 9.0 A/dm^2 (20 – 90 A/ft^2)
Other conditions are the same as for the sulfate – chloride bath.	

The presence of Fe(III) causes brittleness, pitting, and low efficiency. Except in the chloride bath, where high temperatures improve the appearance and ductility of the deposit, excessive temperatures should be avoided since they promote the oxidation of the Fe(II) to Fe(III). The development of Fe(III) may be inhibited by maintaining the solution in contact with iron chips, steel wool, or other forms of metallic iron. Iron(III) may be removed by passing the solution over such a source of metallic iron, by low-current-density dummying, or by raising the pH of the solution to precipitate the Fe(III) as $Fe(OH)_3$ and filtering. Iron baths are similar to nickel baths in that organic contamination can cause stress, pitting, and brittleness. Activated carbon treatment can be used to remove this contamination; however, hydrogen peroxide or permanganate treatment should not be used.

8.12. Lead Deposits

Lead deposits are used for their corrosion resistance and as bearing material. When used on bearing surfaces, lead [*7439-92-1*] is frequently deposited as an alloy (see Section 9.4) or subsequently coated with another material and subjected to heat treatment to produce a diffused surface.

The literature includes references to the deposition of lead from a wide variety of solutions, including nitrate, citrate – chloride, acetate, fluosilicate, fluoborate, perchlorate, oxalate, and many other acid solutions, as well as alkaline solutions. The fluoborate bath is most commonly used.

Lead fluoborate ($Pb(BF_4)_2$)	200 – 470 g/L (27 – 67.5 oz/gal)
Fluoboric acid	20 – 45 g/L (2.6 – 6.0 oz/gal)
Boric acid	20 – 45 g/L (2.6 – 6.0 oz/gal)
Control	
Lead	113 – 255 g/L (15 – 34 oz/gal)
pH, colorimetric	< 1.5
Temperature	25 – 70 °C
Cathode C.D.	0.5 – 60 A/dm^2 (5 – 600 A/ft^2)
Efficiency	100 %
Anode : cathode ratio	2 : 1 or 1 : 1

To improve the quality of the deposit, gelatin or glue may be used at concentrations of 0.15 – 0.38 g/L (0.02 – 0.05 oz/gal), or hydroquinone at 10 g/L (1.3 oz/gal). If the concentration of the colloidal additives is not in the proper range, the deposit may be treed, crystalline, or powdery. Organic contamination can be prevented by activated carbon treatment on a continuous or periodic basis.

8.13. Manganese Deposits

There appears to be no commercial applications for electrodeposited manganese [*7439-96-5*]. Electrowinning of the metal from low-grade ores by electrolysis has been studied, and chloride and sulfate baths for this purpose have been compared. Numerous alloy baths are cited in the literature, but none has reached commercial significance. Deposition of a manganese – nickel alloy has been reported.

8.14. Nickel Deposits

Nickel [*7440-02-0*] is one of the most important, most widely used, and most thoroughly studied of all the electrodeposited metals. Its history goes back over 140 years. It is used as a

Table 3. Gold deposition: the neutral and acid gold baths

Component or condition	Neutral gold baths		Acid gold baths	
	Matte finish	Bright finish	Matte finish	Bright finish
Gold as cyanide *	4 – 16 g/L (0.5 – 2.0 ozt/gal)	2 – 10 g/L (0.25 – 1.0 ozt/gal)	2 – 16 g/L (0.25 – 2.0 ozt/gal)	4 – 12 g/L (0.5 – 1.5 ozt/gal)
Potassium phosphate	0 – 90 g/L (0 – 12 oz/gal)	0 – 90 g/L (0 – 12 oz/gal)	0 – 100 g/L (0 – 13.5 oz/gal)	0 – 100 g/L (0 – 13.5 oz/gal)
Chelates	15 – 90 g/L (2 – 12 oz/gal)	50 – 150 g/L (6.7 – 20 oz/gal)	10 – 200 g/L (1.3 – 27 oz/gal)	10 – 150 g/L (1.3 – 20 oz/gal)
Brighteners		0.1 – 30 g/L (0.01 – 4.0 oz/gal)		0.1 – 20 g/L (0.01 – 2.5 oz/gal)
Secondary brighteners			0 – 10 g/L (0 – 1.3 oz/gal)	0.1 – 10 g/L (0.01 – 1.3 oz/gal)
Temperature	25 – 70 °C	35 – 60 °C	40 – 70 °C	20 – 50 °C
pH	6 – 8	6 – 8	3 – 5	
Cathode current density	0.2 – 1.0 A/dm^2 (2 – 10 A/ft^2)	0.5 – 2.0 A/dm^2 (5 – 20 A/ft^2)	0.1 – 0.5 A/dm^2 (1 – 5 A/ft^2)	
Agitation	moderate	rapid	moderate	rapid
Anodes	platinum, stainless steel, platinized titanium	platinum, stainless steel, platinized titanium	carbon, platinum, platinized titanium	carbon, platinum, platinized titanium

* Gold as the cyanide $KAu(CN)_2 \cdot 2H_2O$.

corrosion-resistant undercoating for chromium and gold, as a decorative and corrosion-resistant finish alone, and for numerous engineering purposes. Modern baths are almost all proprietary and deposit the metal in varied forms: bright; semi-bright but with excellent leveling characteristics; and with varied electrochemical activity so that multilayered nickel deposits provide superior corrosion resistance. Baths also exist which deposit extremely hard nickel that can be used as a machineable substitute for hard chromium. Baths that deposit nickel with low internal stress are used for electroforming.

Despite its long history, however, truly high-speed bright plating solutions were not developed until the early 1940s when the brightening effect of saccharin was discovered. Now numerous bright and/or leveling baths are available.

Most of the modern bright baths are based on the Watts type bath:

Watts type bath	
Nickel sulfate ($NiSO_4 \cdot 6H_2O$)	300 g/L (40 oz/gal)
Nickel chloride ($NiCl_2 \cdot 6H_2O$)	45 g/L (6 oz/gal)
Boric acid	30 – 37.5 g/L (4 – 5 oz/gal)
Control	
Nickel	77.0 g/L (10.3 oz/gal)
pH	2.0 – 5.2
Temperature	32 – 71 °C
Cathode C.D.	1 – 6 A/dm^2 (10 – 60 A/ft^2)
Efficiency	95 – 100

The wide range of pH listed is due to the variety of brightener systems available. For any specific system, the preferred operating pH usually lies in a range ca. 0.5 units broad. The brightener system almost always contains a wetting agent to reduce the surface tension of the solution and prevent pitting due to the adherence of gas bubbles to the surface of the parts. Sodium lauryl sulfate is frequently used, but for systems using air agitation nonfoaming types are required. The choice of wetter is extremely important since it can interact with the brightener components and, if incompatible, cause hazing and/or streaking.

Anodes generally are chips, slugs, or electroformed buttons of specially modified nickel contained in titanium anode baskets. The particular type of anode material is often dictated by the brightener system. Anodes should be bagged to prevent roughness.

Agitation varies from the very slight mechanical action obtained by the movement of the parts through the solution by automatic conveying equipment to relatively violent air agitation. For optimum results, continuous filtration is preferred. Since nickel baths are susceptible to organic contamination, either from brightener breakdown or externally introduced, maintenance of a layer of activated carbon on the filter pads is a worthwhile precaution. Even with this treatment, periodic thorough treatment with activated carbon and even hydrogen peroxide or potassium permanganate to completely remove all organics is sometimes required.

In some cases, the all-chloride bath is used to produce bright deposits. It is more often used to deposit a dull, hard nickel, which is used as a replacement for hard chromium in the repair of worn shafts and similar equipment.

All-chloride bath	
Nickel chloride	240 g/L (32 oz/gal)
Boric acid	30 g/L (4 oz/gal)
Control	
Nickel	75 g/L (10 oz/gal)
pH, electrometric	0.9 – 1.1

A wetting agent is used to reduce the surface tension to 35 – 45 mN/m and prevent gas pitting. For some engineering applications where organic additives must be avoided, 5 – 10 ppm of hydrogen peroxide achieves a similar effect. Acceptable cathode current densities are somewhat higher than in the Watts bath. Other conditions are essentially the same.

For electroforming and for building-up worn parts where hardness is not required, the fluoborate or sulfamate bath may be used.

Fluoborate bath	
Nickel fluoborate	320 g/L (29.3 oz/gal)
Boric acid	30 g/L (4.0 oz/gal)
pH, electrometric	3.0 – 4.5
Control	
Nickel	55 g/L (7.3 oz/gal)
Temperature	32 – 71 °C
Cathode C.D.	5 – 10 A/dm^2 (50 – 100 A/ft^2)
Sulfamate bath	
Nickel sulfamate	450 g/L (60 oz/gal)
Boric acid	30 g/L (4 oz/gal)
Control	
Nickel	105 g/L (14 oz/gal)
pH, electrometric	3 – 5
Temperature	49 °C
Cathode C.D	5 – 30 A/dm^2 (50 – 300 A/ft^2)

Other conditions are essentially the same as for the Watts bath. Non-pitter and proprietary stress-controlling agents may be used as required.

Nickel may be stripped in nitric acid if the substrate allows. Organic nitroaromatic strippers with cyanide or sulfuric acid may be used, depending on the substrate. Numerous electrolytic strippers are also available for various substrates.

Black Nickel Deposits. At one time, black nickel, which is a complex alloy of nickel, nickel sulfide, and zinc, was widely used for producing oxidized and relieved finishes on builder's hardware and other items. Since it is a porous, somewhat soft deposit, and has little or no intrinsic corrosion resistance, it is usually applied over a film of nickel, and it must always be protected by some sort of organic film. Its porosity enables it to hold lubricating oil, and it was applied to typewriter parts for this purpose. Development of an immersion process that could be applied over zinc on typewriter parts, and changes in decorator styles, eliminated these major uses, and the plating solution fell into disuse. The current trend toward nonreflective surfaces on automobiles, and possible use as a solar absorber, has rekindled interest in black nickel. For nonreflective surfaces, the underlying nickel must be dull or lustrous rather than bright, or the nonreflective characteristics are difficult to attain.

Baths fall into two general classes, one more concentrated than the other (Table 4). Control of the current density is probably the most important consideration in obtaining good, dull black deposits. Excess current produces gray or even pure nickel deposits. Some operators feel that small amounts of copper in the solution help to darken the deposits even more. Proper activation of the base nickel layer is necessary to ensure adhesion.

Table 4. Black nickel deposition

Component or condition	Dilute bath	Concentrated bath
Nickel ammonium sulfate ($Ni(NH_4)_2(SO_4)_2 \cdot 6H_2O$)	60 g/L (8 oz/gal)	
Nickel sulfate ($NiSO_4 \cdot 6H_2O$)		75 g/L (10 oz/gal)
Zinc sulfate ($ZnSO_4 \cdot 6H_2O$)	7.5 g/L (1 oz/gal)	30 g/L (4 oz/gal)
Sodium thiocyanate (NaCNS)	15 g/L (2 oz/gal)	15 g/L (2 oz/gal)
Control		
Nickel	9 g/L (1.2 oz/gal)	16.5 g/L (2.2 oz/gal)
pH, colorimetric	5.5 – 6.0	5.0 – 6.2
Temperature	24 °C	21 °C
Cathode current density	0.2 A/dm^2 (2A/ft^2)	0.08 – 0.2 A/dm^2 (0.75 – 2.25 A/ft^2)
Voltage	1 – 2 V	1 – 2 V

Table 5. Rhodium deposition

Component or condition	Sulfuric acid or phosphoric acid bath	Concentrated bath, for heavier deposits
Rhodium *	2 g/L (5 dwt/gal)	10 – 20 g/L (1.3 – 2.7 oz/gal)
Sulfuric acid (c.p.)	2 vol %	2.5 vol %
or		
Phosphoric acid (85 %)	2 vol %	
Temperature	40 – 45 °C	50 °C
Cathode current density	1 – 10 A/dm^2 (10 – 100 A/ft^2)	1 – 2 A/dm^2 (10 – 20 A/ft^2)
Voltage	6	6
Agitation	as required	vigorous
Cathode efficiency	80 %	85 %
Anode : cathode ratio	1 : 1	1 : 1

* The rhodium is added as a prepared concentrate.

8.15. Niobium Deposits

There is no commercial method for depositing niobium [*7440-03-1*].

8.16. Rhodium Deposits

Rhodium [*7440-16-6*] is used for thin, bright decorative and tarnish-resistant coatings (0.025 –0.15 µm). Heavier coatings (1.5 – 25 µm) provide high reflectivity, corrosion resistance, and because of its hardness, scratch resistance. Rhodium is used on jewelry, silverware, reflectors, and on electrical contacts. Except when applied over gold or the platinum group metals, an undercoating of nickel is recommended. Rhodium may be plated from either a sulfuric acid or a phosphoric acid bath (Table 5). Platinum anodes should be used. Patented additives are available to eliminate stress in and cracking of deposits. For heavier deposits, a more concentrated solution is used (Table 5).

Rhodium is resistant to the action of most stripping solutions. Light deposits may be removed by undercutting through the cracks in the deposit.

8.17. Silver Deposits

Silver [*7440-22-4*] is used for both decorative and engineering purposes. Decorative uses include silverware, jewelry, lighting fixtures, and other items. Engineering uses include bearings, electronic uses, and special applications such as the plated O-rings used to seal reactors. Many baths have been suggested including nitrate, iodide, thiourea, thiocyanate, sulfamate, and thiosulfate. Only the cyanide bath (Table 6) has received broad industrial acceptance. The main improvements have been in developing brightener systems to give full bright plate directly from the bath to eliminate buffing and in changing the metal and cyanide content to permit high-speed plating for engineering purposes, where thicknesses may be as high as 1500 µm.

Specific brightener combinations may require changes in the concentrations or conditions. With proper brightener additions, conventional and nitrate cyanide baths may be used for deposits up to 2.5 – 33 µm. Filtration is desirable, and mechanical agitation should be used. For the heavy deposit bath, agitation should be vigorous using a combination of mechanical agitation and solution agitation.

Contamination with copper may cause a pink color in the deposit. There is no economic way to remove this contamination. The bath should either be discarded and the silver content reclaimed, or the metal content depleted using stainless steel anodes and then used as a preliminary strike solution.

Table 6. Silver deposition

Component or condition	Baths		
	Conventional bath	Nitrate – cyanide bath	Heavy deposit bath
Silver cyanide	36 g/L (4.8 oz/gal)	22 – 45 g/L (3 – 6 oz/gal)	105 g/L (14.0 oz/gal)
Potassium cyanide	60 g/L (8.0 oz/gal)		113 g/L (15 oz/gal)
Sodium cyanide		33 – 45 g/L (4.4 – 6 oz/gal)	
Potassium carbonate	45 g/L (6.0 oz/gal)		15 – 115 g/L (2 – 15 oz/gal)
Sodium carbonate		22 – 45 g/L (3 – 6 oz/gal)	
Potassium nitrate		45 – 120 g/L (6 – 20 oz/gal)	
Potassium hydroxide			30 g/L (4 oz/gal)
Control			
Silver	28.7 g/L (3.5 ozt/gal)	16 – 37 g/L (2 – 4.5 oz t/gal)	84 g/L (10.2 oz t/gal)
Free potassium cyanide	43 g/L (5.7 oz/gal)	19 – 23 g/L ∗ (2.5 – 3.1 oz/gal) ∗	62 g/L (8.2 oz/gal)
Temperature	24 – 32 °C	27 – 52 °C	43 – 55 °C
Cathode current density	0.5 – 1.5 A/dm^2 (5 – 15 A/ft^2)	5 – 15 A/dm^2 (5 – 150 A/ft^2)	6 – 15 A/dm^2 (60 – 150 A/ft^2)
Efficiency	100 %	100 %	
Anodes	silver	silver	
Anode : cathode ratio	1 : 1	1 : 1	2 : 1 – 4 : 1

∗ Cyanide as potassium cyanide.

Table 7. Silver strikes

Component or condition	Copper – silver strike	Silver strike	Engineering strikes	
			Second strike	Third strike
Silver cyanide	1.7 g/L (0.23 oz t/gal)	7.0 g/L (0.9 oz/gal)	1.0 g/L (0.13 oz/gal)	5.0 g/L (0.67 oz/gal)
Potassium cyanide	75 g/L (10 oz/gal)	75 g/L (10.0 oz/gal)	45 g/L (5.0 oz/gal)	15 g/L (2.0 oz/gal)
Copper cyanide	15 g/L (2.0 oz/gal)			
Potassium carbonate			30 g/L (4.0 oz/gal)	30 g/L (4.0 oz/gal)
Control				
Silver	1.6 g/L (0.2 oz/gal)	6.0 g/L (0.7 oz t/gal)	0.8 g/L (0.1 oz t/gal)	4.0 g/L (0.5 oz/gal)
Copper	10.5 g/L (1.4 oz/gal)			
Free potassium cyanide	53 g/L (7.0 oz/gal)	71 g/L (9.5 oz/gal)	44 g/L (5.9 oz/gal)	13 g/L (1.7 oz/gal)
Temperature	24 °C	24 °C	24 °C	24 °C
Cathode current density	3 A/dm^2 (30 A/ft^2)	3 A/dm^2 (30 A/ft^2)	3 A/dm^2 (30 A/ft^2)	3 A/dm^2 (30 A/ft^2)
Time	20 s	20 s	20 s	20 s
Anodes	steel	steel	steel	steel

Because silver is a noble metal, the initial deposit must be applied from a strike solution. In many cases, a double strike is used: the first from a mixed copper – silver strike solution and the second from a standard silver strike solution (Table 7).

The copper – silver strike is used as the initial strike on steel. A nickel strike such as the Woods nickel strike (see 5.3.7) may be substituted.

The silver strike is used as the second strike on steel and as the initial strike on other metals. For engineering purposes, three strikes are recommended. The *first strike* is from a Woods nickel strike solution. The *second strike* is plated over the nickel and the *third strike* then applied (Table 7).

Silver may be stripped in nitroaromatic – cyanide strippers where applicable, many of which are now available with inhibitors to protect nickel, and sometimes copper underlayers. It may also be stripped in accelerated nitric acid strippers, which contain 25 vol % nitric acid. Electrolytically, sodium cyanide – sodium hydroxide solutions may be used.

8.18. Tin Deposits

Tin [*7440-31-5*] is soft, ductile, solderable, and resistant to corrosion and tarnish. It is used extensively in electrical and electronic work. Because it is nontoxic, it is widely used to protect items that may come into contact with food. In the presence of oxygen, tin is more noble than steel, and does not offer sacrificial protection. Thin coatings, therefore, being slightly porous, often generate severe pin-hole corrosion in the

steel substrate. However, this condition is easily reversed in the absence of oxygen, and the tin will protect the steel even if slightly porous. Therefore, thin coatings of tin are widely used on steel to be formed into "tin" cans for food. When applied in a thickness adequate to eliminate porosity, or when the coatings are melted (reflowed) after deposition to produce a bright coating and to eliminate porosity, tin offers excellent corrosion protection.

Tin may be deposited from alkaline stannate(IV) solutions or from acid tin(II) solutions. It is widely applied by continuous plating techniques to strip and wire. Baths for this purpose are usually operated under conditions that produce a matte deposit, and the coating is subsequently reflowed.

The sodium and potassium *stannate baths* (Table 8) operate similarly. The potassium bath permits higher operating temperatures, higher current densities, and more leeway in formulation. The potassium salts are more expensive, however. Where there are economies of operation to be gained by higher plating speeds, the added cost of the potassium salts may be justified. For other operations, the sodium bath may be equally satisfactory.

The sodium bath may be used for almost all base metals for applications requiring solderability and protection from corrosion. It may also be used as a strike on cast and malleable iron and steels that are difficult to cover properly. It has excellent throwing power.

Stannous ion, Sn^{2+}, is harmful to the characteristics of the deposit. It is desirable to periodically treat the bath with small amounts of hydrogen peroxide, or to maintain a slight excess of hydrogen peroxide in solution.

Proper treatment of the anodes is essential to efficient operation. The anodes must be filmed with a polarized layer of greenish white tin oxide to assure that the tin dissolves in the stannic form, Sn^{IV}. The anodes should preferably be removed during idle periods and returned to the bath a few at a time on start-up, allowing time for the anodes to film properly before more are added. Alternatively, the bath should be started with a full load of work at very high currents until the anodes film properly. Dummies may be used if necessary to protect the work from excess current during this operation. The bath should never be operated with a high anode-to-cathode ratio since under these conditions the film is lost and tin(II) dissolves into the bath.

The concentrated sodium and the potassium baths are suggested for continuous plating of wire and strip. Anodes of pure tin normally allow anode current densities of 3 – 5 A/dm^2 (30 – 50 A/ft^2). There are also more concentrated potassium solutions, which allow the use of cathode current densities up to 40 A/dm^2 (400 A/ft^2). When these solutions are used, an anode of tin alloyed with 1 % aluminum, which may be operated at 10 A/dm^2 (100 A/ft^2), makes the required anode : cathode ratio more easily attainable.

The potassium baths may also be operated using insoluble steel anodes, thus eliminating the anode problem. A specially prepared tin oxide sol is available which permits control of tin content without generating hydroxide.

For years sodium acetate was thought to be required as a buffering agent, particularly in the sodium bath. However, this is not the case. In any event, acetic acid is the acid of choice for controlling the hydroxide content of the bath.

Tin may be deposited from numerous *acid baths,* but only three are in general use. For continuous tinning of wire and strip, the du Pont Halogen process, stannous fluoborate, and stannous sulfate are the major baths in use.

Halogen bath	
Tin(II) chloride ($SnCl_2 \cdot 2\,H_2O$)	75 g/L (10 oz/gal)
Sodium fluoride	25 g/L (3.35 oz/gal)
Potassium bifluoride	50 g/L 6.7 oz/gal)
Sodium chloride	36 g/L (4.8 oz/gal)
pH	2.7
Addition agents	1 – 2 g/L (0.13 – 0.26 oz/gal)
Temperature	65 °C
Cathode C.D.	45 A/dm^2 (450 A/ft^2)
Fluoborate bath	
Tin(II) fluoborate	200 g/L (26.8 oz/gal)
Fluoboric acid	100 – 200 g/L (13 – 26 oz/gal)
Gelatin	6 g/L (0.8 oz/gal)
β-Naphthol	1 g/L (0.134 oz/gal)
Temperature	20 – 40 °C
Cathode C.D.	2.5 – 42.5 A/dm^2 (25 – 450 A/ft^2)
Stannous sulfate bath	
Tin(II) sulfate	60 – 100 g/L (8 – 13.5 oz/gal)
Sulfuric acid	40 – 70 g/L (5.36 – 9.38 oz/gal)

For strip plating, cresolsulfonic acid 30 – 60 g/L (4 – 6 oz/gal), gelatin 2 g/L (0.26 oz/gal), and β-naphthol 1 g/L (0.134 oz/gal) or

Table 8. Sodium and potassium stannate baths

Component or condition	Sodium bath	Concentrated sodium bath	Potassium bath
Sodium stannate ($Na_2SnO_3 \cdot 3\,H_2O$)	90 g/L (12 oz/gal)	140 g/L (19 oz/gal)	
Potassium stannate			80 g/L (10.6 oz/gal)
Sodium hydroxide	7.5 g/L (1 oz/gal)	15 g/L (2 oz/gal)	
Potassium hydroxide			30 g/L (4.0 oz/gal)
Sodium acetate	15 g/L (2 oz/gal)		
Control			
Tin	38 g/L (5.0 oz/gal)	60 g/L (8 oz/gal)	29 g/L (3.8 oz/gal)
Temperature	60 – 80 °C	93 – 99 °C	85 °C
Cathode current density	1 – 2.5 A/dm^2 (10 – 25 A/ft^2)	4.5 – 6.5 A/dm^2 (45 – 65 A/ft^2)	4.0 A/dm^2 (40 A/ft^2)
Efficiency	60 – 90 %	90 %	80 – 100 %
Anode : cathode ratio	1 : 1	1 : 1	1 : 1

phenolsulfonic acid 40 g/L (5.36 oz/gal), β-naphthol 1 g/L(0.134 oz/gal), and dihydroxydiphenylsulfone 5 g/L (0.67 oz/gal) are also added. Proprietary additives may be used.

For standard rack or barrel plating, proprietary additives, many based on the work of SCHLOTTER and other researchers in Germany and France are used. These brighteners generally require that the chloride content of the bath be held to a maximum of 200 ppm, but when properly operated produce brilliant deposits.

The baths are operated at 15–30 °C. Current densities range from 1 – 4 A/dm^2 (10 – 40 A/ft^2) for mildly agitated baths to 10 A/dm^2 (100 A/ft^2) for highly agitated strip lines.

Tin may be stripped in both acid and alkaline solutions. The stripping can be accomplished by immersion in a nitroaromatic and sodium hydroxide; in an accelerated 25 vol % nitric acid; in a mixture of iron(III) sulfate and sulfuric acid, 100 g/L of each; or in a 120 g/L (16 oz/gal) solution of a mixture of 75 wt % sodium hydroxide, 13 wt % sodium nitrate, and 12 wt % sodium nitrite. Electrolytically, tin can be stripped in dilute sodium hydroxide or in proprietary acid strippers.

8.19. Zinc Deposits

Zinc [*7440-66-6*] has for many years been deposited from the *cyanide bath* (Table 9).

The ratio of total cyanide : zinc should be maintained at ca. 2.75 – 3.25. The higher ratio is used where the most positive throwing power is needed. Higher temperatures cause excessive brightener consumption and eventually lead to organic contamination from brightener breakdown products. Anodes may be a combination of zinc and steel. If all zinc anodes are used, the metal content of the solution tends to increase. The voltage requirement depends on the tank configuration and the plating speed. Mechanical agitation of the cathode is recommended. For best results, the bath should be filtered regularly, if not continuously. Since zinc is reactive and fingerprints easily, a chromate post-treatment is recommended.

Poor throwing power can result from solution imbalance, particularly from a combination of high metal concentration with high caustic or high cyanide. Dark deposits may result from contamination, particularly from copper, cadmium, or lead. Even low levels of copper contamination can cause problems with subsequent chromate treatments and/or loss of corrosion resistance. Hazing of the deposits can result from solution imbalance, organic contamination, high caustic, improper rinsing, or lack of brightener. Polarized anodes may result from low cyanide, low caustic, or insufficient anode area.

Metallic contamination may be removed by low current-density plating (dummying), treatment with sodium polysulfide, or treatment with zinc dust followed by filtering. Organic contamination may be removed by activated carbon treatment.

The standard cyanide bath is a particularly tolerant bath and can be operated under adverse conditions – poor cleaning, oil drag-in, poor housekeeping – and still produce reasonably acceptable visual results. In order to reduce the problems of wastewater treatment and disposal, *low-cyanide zinc baths* (Table 9) have also been developed.

Table 9. Zinc deposition: cyanide baths

Component or condition	Cyanide bath		Low cyanide bath
	Zinc oxide formulation	Zinc cyanide formulation	
Zinc oxide (ZnO)	45 g/L (5.8 oz/gal)		10 – 22 g/L (1.3 – 3.0 oz/gal)
Zinc cyanide ($Zn(CN)_2$)		40 g/L (5.5 oz/gal)	
Sodium cyanide (NaCN)	100 g/L (13.6 oz/gal)	50 g/L (6.6 oz/gal)	2.2 – 19 g/L (0.3 – 2.5 oz/gal)
Sodium hydroxide (NaOH)	50 g/L (7.0 oz/gal)	95 g/L (12.7 oz/gal)	7.5 – 112 g/L (10 – 16 oz/gal)
Analysis or control			
Zinc, as ZnO	45 g/L (4.7 oz/gal)		7.5 – 15 g/L (1 – 2 oz/gal)
Total cyanide	102 g/L (13.6 oz/gal)		2.2 – 19 g/L (0.3 – 2.5 oz/gal)
Total NaOH	95 g/L (12.7 oz/gal)		75 – 112 g/L (10 – 15 oz/gal)
Temperature	20 – 30 °C		20 – 35 °C
pH, colorimetric	> 13.0		>13.0
Cathode current density	1 – 6 A/dm^2 (10 – 60 A/ft^2)		1 – 3.5 A/dm^2 (10 – 35 A/ft^2)
Cathode efficiency	65 – 86 %		65 – 80 %
Anode : cathode ratio	2 : 1		3 : 1
Voltage	1 – 6 V		

The operating conditions are similar to those for the standard cyanide bath, with the exception that this bath is not as tolerant. It is more sensitive to both metallic and organic impurities, as well as shortcomings in surface preparation.

The original brightener systems for zinc cyanide solutions were based on the use of aldehydes with small amounts of metallic ions as brighteners for the low current-density areas. Since aldehydes polymerize in the presence of caustic soda, these brighteners tended to decompose even when the bath was idle, resulting in an excess of metallic components and brightener system imbalance. Operation above the recommended temperature also accelerated brightener decomposition. The aldehyde components of the brightener system were often referred to as temporary brighteners and were added separately to better maintain balance.

Modern systems have replaced the aldehydes with more stable compounds to avoid this problem, but are still sensitive to operation above certain temperature limits.

In a further effort to reduce waste problems, *alkaline noncyanide baths* were developed. These consist essentially of sodium zincate formed from zinc oxide and sodium hydroxide. Complexing agents of various types may also be added. The baths are most often sold as proprietary liquid concentrates. Operating conditions are slightly different for barrel and rack solutions (Table 10).

Table 10. Zinc deposition: alkaline noncyanide baths

Component or condition	Barrel plating	Rack plating
Zinc	8 – 15 g/L (1 – 2 oz/gal)	6 – 12 g/L (0.8 – 1.6 oz/gal)
Sodium hydroxide	80 – 150 g/L (10 – 20 oz/gal)	80 – 120 g/L (8 – 16 oz/gal)
Temperature	18 – 35 °C	18 – 35 °C
Cathode current density	0.5 – 1.2 A/dm^2 (5 – 12 A/ft^2)	1 – 2.5 A/dm^2 (10 – 25 A/ft^2)

These baths are even more sensitive to improper operating conditions than the low cyanide baths.

In recent years, research has been concentrated on the *acid zinc baths.* Major developments in acid chloride zinc baths have produced baths that are not only brilliant but also have leveling action that hides many surface defects. The original work utilized ammonium chloride as an additive to increase conductivity, but these baths were somewhat difficult to control and have for the most part been replaced by baths using potassium chloride (Table 11). High purity anodes are required. Because the bath is highly corrosive, all equipment should be suitably protected.

For plating wire, strip, and other mill products, the acid sulfate bath (Table 11) is often used. Glucose (120 g/L, 16 oz/gal) or licorice (1.0 g/L, 0.33 oz/gal) may be used to refine the grain structure. Steel companies have been reported to have developed proprietary baths that allow current densities up to 10 – 60 A/dm^2

Table 11. Zinc deposition: acid zinc baths

Component or condition	Acid chloride zinc baths				Acid sulfate bath
	With ammonium chloride		With potassium chloride		
	Barrel plating	Rack plating	Barrel plating	Rack plating	
Zinc chloride	34 – 64 g/L (4.5 – 8.5 oz/gal)	74 – 95 g/L (10 – 12.5 oz/gal)	50 – 85 g/L (6.7 – 11.3 oz/gal)	75 – 85 g/L (10.0 – 11.3 oz/gal)	
Zinc sulfate ($ZnSO_4 \cdot 7\,H_2O$)					240 – 360 g/L (32 – 48 oz/gal)
Ammonium chloride	112 – 225 g/L (15 – 30 oz/gal)	90 – 120 g/L (12 – 16 oz/gal)			15 – 30 g/L (2 – 4 oz/gal)
Sodium chloride		45 – 67.5 g/L (6 – 9 oz/gal)			
Potassium chloride			180 – 270 g/L (24 – 36 oz/gal)	202 – 270 g/L (27 – 36 oz/gal)	
Boric acid or	19 – 26 g/L (2.5 – 3.5 oz/gal)	19 – 26 g/L (2.5 – 3.5 oz/gal)	22.5 – 34 g/L (3.0 – 4.5 oz/gal)	22.5 – 34 g/L (3.0 – 4.5 oz/gal)	
Ammonium acetate			7.5 – 12 g/L (1.0 – 1.6 oz/gal)	7.5 – 12 g/L (1.0 – 1.6 oz/gal)	
Aluminum sulfate ($Al_2(SO_4)_3 \cdot 18\,H_2O$)					30 g/L (4 oz/gal) *
Control					
Zinc	16.5 – 30 g/L (2.2 – 4.0 oz/gal)	36 – 45 g/L (4.8 – 6 oz/gal)	24 – 40 g/L (3.2 – 5.4 oz/gal)	35 – 40 g/L (4.8 – 5.4 oz/gal)	55 – 83 g/L (7.3 – 11 oz/gal)
Chloride	93 – 183 g/L (12.4 – 24 oz/gal)	126 – 170 g/L (16.8 – 22.6 oz/gal)	112 – 175 g/L (15 – 23 oz/gal)	135 – 175 g/L (18.0 – 23.0 oz/gal)	
Buffer **	original concentration		original concentration		
pH, colorimetric	4.4 – 5.6				3.5 – 4.2
Temperature	15 – 45 °C				
Cathode current density	0.2 – 6.0 A/dm^2 (5 – 30 A/ft^2)				1 – 3 A/dm^2 (10 – 30 A/ft^2)
Cathode efficiency	close to 100 %				close to 100 %
Anode : cathode ratio	1 : 1				1 : 1

* Sodium acetate may be used in place of the aluminum sulfate at 50 % of this concentration.
** The buffer is boric acid or ammonium acetate.

(100 – 600 A/ft^2) and operation temperatures up to 37 – 65 °C.

9. Specific Alloy Deposits

Solutions for depositing alloy coatings have been extensively investigated, and the deposition of numerous such alloys has been reported in the literature. Only a limited number of these have attained any industrial importance, however.

9.1. Chromium Alloys

There are no known commercial chromium alloy coatings. Chromium – nickel alloys can be deposited from a bath containing chromic acid and nickel sulfate, but this work is still in the research stage. The efficiency of the bath is extremely low, and commercial development does not seem likely.

9.2. Copper Alloys – Brass Plating

Since copper can be plated from both the monovalent and the divalent state, and since the deposition potential can be readily adjusted by the use of complexing agents, numerous copper alloys have been proposed and investigated. These include brass, which is probably the most widely used, an imitation bronze based on cadmium, a true bronze containing tin, and ternary alloys of copper – tin – zinc. They are used for both decorative and engineering purposes.

Brass plating was invented in 1841, but did not come into widespread use until the 1880s. It is used for decorative purposes, where the emphasis is on the color of the deposit, and as a substrate for bonding rubber to steel, where the emphasis is on the color of the deposit composition and crystal structure. For a number of years, there was interest in the use of a white brass, instead of copper, as the initial layer on zinc-base die castings that were to be nickel and chromium plated. Brass solutions are controlled using free cyanide as an analytical criterion.

The color of brass deposits is not completely dependent on composition. Furthermore, the composition of the deposit varies with current density, total metal concentration, free cyanide concentration, the copper : zinc ratio in the solution, temperature, and other operating conditions. The color of the deposits also depends on these conditions. Operation of the original plating solutions to obtain desired results was often more of an art than a science and required extensive experience to produce consistent results. Modern proprietary baths have eliminated many of these problems by use of additives that simplify solution control.

Most brass baths are only 60 – 70 % efficient, but proper control of the operating conditions in specially formulated baths can yield efficiencies up to 90 %. These are the so-called high speed (Hi-speed) baths. Brass anodes are also easily polarized, so anode efficiency is often poor and additions of metal salts to the baths are frequent. For most baths, the anode : cathode ratio should be maintained as high as possible so that anode current densities are low. Additions of Rochelle salt is sometimes helpful, and proprietary baths use a variety of complexing agents, often in combination, to obtain good anode corrosion and efficiency (Table 12).

In all of the brass baths, carbonates in excess of 90 g/L (12 oz/gal) are harmful, and carbonates form particularly rapidly in high-speed operations such as strip plating. Constant removal with a special ion-exchange resin is recommended for control under this conditions. For other operations, periodic removal by freezing or chemical precipitation usually suffices.

A bronze color may be produced by depositing a copper-rich alloy (Table 12, bronze bath). An excellent bronze color may also be produced using a cadmium bronze. A standard copper bath (see Table 1) is modified by adding cadmium cyanide or cadmium oxide dissolved in sodium cyanide until the desired color is obtained. The bath is operated with copper anodes, and the cadmium replenished as a liquid concentrate as required to maintain the color.

A noncyanide brass plating solution has been developed and is being sold as a proprietary solution. It has not made great progress in the market, nor are unbiased reports on its performance available.

Table 12. Brass deposition

Component or condition	Bath 1	Bath 2	Bath 3	Hi-speed bath	Bronze bath	White brass bath
Copper cyanide	27 g/L (3.6 oz/gal)	52.0 g/L (7.0 oz/gal)	42.7 g/L (5.7 oz/gal)	60.0 g/L (8.0 oz/gal)	53.5 g/L (7.2 oz/gal)	15.7 – 20.2 g/L (2.1 – 2.7 oz/gal)
Zinc cyanide	9.0 g/L (1.2 oz/gal)	30.0 g/L (4.0 oz/gal)	18.8 g/L (1.8 oz/gal)	3.75 g/L (05 oz/gal)	3.8 g/L (0.5 oz/gal)	34.4 – 39.7 g/L (4.6 – 5.3 oz/gal)
Sodium cyanide	56.0 g/L (7.5 oz/gal)	90.0 g/L (12.0 oz/gal)	67.7 g/L (9.0 oz/gal)	99.0 g/L (13.2 oz/gal)	62.2 g/L (8.3 oz/gal)	47.7 – 53.6 g/L (6.4 – 7.2 oz/gal)
Sodium carbonate	35.5 g/L (4.5 oz/gal)	30.0 g/L (4.0 oz/gal)	15.0 g/L (2.0 oz/gal)	15.0 g/L (2.0 oz/gal)	30.0 g/L (4.0 oz/gal)	37.5 g/L (5.0 oz/gal)
Sodium hydrogen carbonate		15.0 g/L (2.0 oz/gal)				
Rochelle salt ($NaK(C_4H_4O) \cdot 4H_2O$)		45.0 g/L (6.0 oz/gal)		45.0 g/L (6.0 oz/gal)	45.0 g/L (6.0 oz/gal)	
Sodium hydroxide				60.0 g/L (8.0 oz/gal)		30.0 – 37.5 g/L (4.0 – 5.0 oz/gal)
Ammonium hydroxide	as needed [a]	0.5 – 1.5 vol %	12.5 – 25.0 g/L (1.5 – 3.0 oz/gal)		1.0 – 5.0 g/L (0.1 – 0.7 oz/gal)	
Control						
Copper	20.0 g/L (2.5 oz/gal)	37.5 g/L (5.0 oz/gal)	30.0 g/L (4.0 oz/gal)	42.5 g/L (5.7 oz/gal)	38.0 g/L (5.1 oz/gal)	11.0 – 14.0 g/L (1.5 – 2.0 oz/gal)
Zinc	5.0 g/L (0.65 oz/gal)	15.0 g/L (2.0 oz/gal)	7.7 g/L (1.0 oz/gal)	2.25 g/L (0.3 oz/gal)	2.1 g/L (0.28 oz/gal)	29.0 – 33.5 g/L (4.0 – 5.0 oz/gal)
Cu : Zn ratio	4.0 : 1	2.5 : 1	4.0 : 1	19 : 1	18 : 1	0.38 – 0.44 : 1
Free cyanide	17.5 g/L (2.3 oz/gal)	7.5 g/L (1.0 oz/gal)	7.5 g/L (1.0 oz/gal)	36.0 g/L (4.8 oz/gal)	4.5 g/L (0.6 oz/gal)	4.7 – 6.4 g/L (0.6 – 0.85 oz/gal)
pH	10.3 – 10.7	10.3 – 10.7	9.8 – 10.3	[b]	10.3	[b]
Temperature	28 – 50 °C	43 – 60 °C	35 – 44 °C	60 °C	38 – 60 °C	21 – 29 °C
Cathode current density	0.5 A/dm^2 (5.0 A/ft^2)	0.5 – 3.5 A/dm^2 (5 – 35 A/ft^2)	0.5 – 1.6 A/dm^2 (5 – 16 A/ft^2)	up to 10 A/dm^2 (up to 100 A/ft^2) [c]	0.5 – 3.2 A/dm^2 (5 – 32 A/ft^2)	4.3 A/dm^2 (45 A/ft^2)
Cu : Zn anode ratio	80 : 20	70 : 30	80 : 20		95 : 5	35 – 65

[a] Addition of ammonium hydroxide is used to control the pH.
[b] pH not controlled.
[c] The maximum current density depends on the degree of agitation.

9.3. Gold Alloys

The use of antimony and selenium as hardeners for gold deposits is mentioned in page 10, and the use of copper, nickel, and silver as color modifiers for decorative gold deposits is described in 8.9.

9.4. Lead Alloys

The use of antimony codeposited with lead to strengthen the plated alloy has been reported. Lead – silver alloys and lead – cadmium alloys have also been reported. A ternary alloy of lead – tin – antimony has been suggested. However, the main industrial alloys continue to be the lead – tin alloys, which are used for improved solderability, particularly in electronic work (usually 50 – 70 % tin), for bearing surfaces (6 – 20 % tin, sometimes with small amounts of copper or antimony), and occasionally as a lubricant for some machining operations. The deposit varies with the ratio of the metals in the solution, the total metal content, and the current density. (An increase in current density increases the tin in the deposit.) Additives are required to improve the character of the deposit. The compositon of the anode should be the same as the deposit. A typical bath to produce lead : tin deposits is as follows:

Lead fluoborate	
Tin(II) fluoborate	see Table 13
Boric acid	16.5 g/l (2.2 oz/gal)
Fluoboric acid	30 g/L (4 oz/gal)
Glue	0.4 g/L (0.05 oz/gal)
Control	see Table 13
Boric acid	22.5 g/L (3.0 oz/gal)
Fluoboric acid	30 g/L (4.0 oz/gal)
pH, colorimetric	>0.5
Temperature	21 – 38 °C
Cathode C.D.	2.5 A/dm^2 (25 A/ft^2)
Efficiency	100 %
Anode : cathode ratio	2 : 1

A low concentration of glue causes a granular deposit, loss of throwing power, and decreased tin in the deposit. Proprietary additives may be used in place of the glue. Typical of acid baths, these baths do not have good throwing power and would normally be unsuited for plating through-hole circuit boards. However, increasing the free fluoboric acid and using peptone to replace the glue makes the bath suitable, and proprietary additive systems produce even better results.

9.5. Nickel Alloys

Numerous nickel alloys may be deposited. Cobalt easily codeposits with nickel and has been used both as a modifying alloying element and as a substitute for nickel during times of nickel shortages. Deposition of nickel alloys with tin, copper, cobalt, zinc, and iron from pyrophosphate solutions has been reported, but commercial applications have not developed. Nickel – zinc has been suggested as a coating for sheet steel. This same alloy has been suggested as a replacement for cadmium for corrosion protection of fasteners. A rhodium – nickel alloy bath has been developed. A ternary alloy of cobalt – nickel – copper from a pyrophosphate bath has been reported. The three baths in commercial use are nickel – iron, nickel – tin, and nickel – palladium.

The *nickel – iron bath* was widely promoted as an effective and economical substitute for nickel plating, and numerous baths were installed. Despite many advantages, difficulties with solution control and difficulty in stripping rejects caused an appreciable return to standard nickel baths. Many nickel – iron baths still remain in operation, however.

In addition to the lower cost of iron, and the accompanying economies, nickel – iron deposits offer corrosion protection equivalent to nickel in many applications, and exhibit less tendency to passivate. Coverage in chromium plating solutions is therefore better, and replating of rejects requires less stringent activation procedures. There is also less danger of peeling due to laminated plate should the current be interrupted in the plating cycle. Brightener systems are available to produce excellent leveling and brilliant deposits.

Typical nickel – iron bath	
Nickel sulfate	70 – 75 g/L (9 – 10 oz/gal)
Nickel chloride	70 – 75 g/L (9 – 10 oz/gal)
Iron(II) sulfate	10 – 20 g/L (1.3 – 2.6 oz/gal)
Boric acid	30 – 40 g/L (4.0 – 5.5 oz/gal)
Stabilizer	20 – 30 g/L (2.6 – 4.0 oz/gal)
pH	2.5 – 3.2
Temperature	60 – 65 °C

The stabilizer is required to maintain the iron in its 2+ state; one suggested material is trisodium citrate.

Table 13. Composition of baths to deposit specific lead – tin alloys: lead and tin components

Component	6 % tin	10 % tin	40 % tin	60 % tin
Lead fluoborate	345 g/L (46 oz/gal)	330.0 g/L (44 oz/gal)	217.5 g/L (29 oz/gal)	146.0 g/L (19.5 oz/gal)
Tin(II) fluoborate	30 g/L (4 oz/gal)	50.0 g/L (6.7 oz/gal)	202.5 g/L (27 oz/gal)	300.0 g/L (40 oz/gal)
Control				
Lead	95 g/L (12.6 oz/gal)	91.0 g/L (12.1 oz/gal)	60.0 g/L (8.0 oz/gal)	40.0 g/L (5.4 oz/gal)
Tin	6.0 g/L (0.8 oz/gal)	10.0 g/L (1.3 oz/gal)	40.0 g/L (5.4 oz/gal)	60.0 g/L (8.0 oz/gal)
Pb : Sn ratio	16 : 1	9.3 : 1	1.5 : 1	0.68 : 1

The brightener systems are in many ways similar to the standard systems used in bright nickel solutions. Specific brightener systems may require considerable deviation from the concentrations and conditions listed above. Control and purification procedures are similar to those used for nickel with the exception that hydrogen peroxide or permanganate treatment cannot be used for the removal of organic contamination.

Nickel – iron deposits may be stripped in nitric acid or in specially developed strippers. Problems in stripping high-current-density deposits sometimes occur.

Deposits of *tin – nickel alloys* were obtained in the early 1930s, but the current baths are based on developments from the 1950s. The deposit is almost inert to many reagents and possesses excellent resistance to atmospheric corrosion. It has excellent throwing power and can be soldered. The recommended bath contains:

Tin(II) chloride ($SnCl_2$)	49 g/L (6.5 oz/gal)
Nickel chloride ($NiCl_2 \cdot 6H_2O$)	300 g/L (40 oz/gal)
Ammonium bifluoride	56 g/L (7.5 oz/gal)
Ammonium hydroxide	to adjust pH to 2.0 – 2.5
Temperature	$68 \pm 3\,^{\circ}C$
Cathode C.D.	1 – 3 A/dm^2 (10 – 30 A/ft^2)

Hydrated tin(II) chloride may be used with the appropriate adjustment in concentration. The total fluorine content (initially 37.5 g/L or 5.0 oz/gal) should always be at least as high as the total tin (initially 30.0 g/L or 4 oz/gal). When the alloy is to be applied to steel, an undercoating of copper or bronze greatly improves the corrosion resistance [10], [11]. The main drawback to the coating is the extreme difficulty of stripping it without damaging the substrate.

Nickel – palladium alloys are being seriously considered as substitutes for gold in electronic applications. Varying the metal concentrations in solution yields alloys containing 80 – 90 % palladium. A typical bath contains:

$Pd(NH_3)_2Cl_2$	40 g/L (5.4 oz/gal)
Ammonia	9 vol %
Ammonium sulfate	50 g/L (6.7 oz/gal)
Nickel sulfate	45 g/L (6.0 oz/gal)
Additives	as needed
pH	8.5
Temperature	$30\,^{\circ}C$
Cathode C.D.	1 A/dm^2 (10 A/ft^2)

This bath will produce an alloy 80 wt % Pd.

Composition and concentration control of the additives may be critical to deposit properties particularly for high-speed plating.

9.6. Tin Alloys

Tin alloys with lead, copper, and nickel are discussed under lead, copper, and nickel alloys.

Tin and cadmium can be deposited in all proportions from sodium hydroxide – sodium cyanide baths. Also reported have been baths based on fluoborate, sulfate, and several other anions. The alloy has been used on threaded fasteners, and for some time the corrosion resistance of the alloy was felt to be superior to cadmium in marine environments, but this has been questioned.

A tin – zinc alloy can be plated from a similar solution and was proposed as a substitute for cadmium. It has excellent solderability. The corrosion resistance appeared variable until it was determined that the application of a chromate film by immersion gave the variable results, and an electrolytic chromate film was developed that eliminated this problem. The zinc is sufficiently segregated that the alloy exhibits the potential of zinc and therefore provides the same protection. However, when corrosion does occur, the voluminous corrosion salts associated with the corrosion of zinc do not develop.

A low antimony alloy (as low as 0.2 % Sb) is immune to tin pest, and antimony is therefore included in many baths as a low-level additive. Bismuth behaves similarly, and a process based on tin – bismuth alloy anodes has been commercialized. In view of the usefulness of tin – nickel, tin – cobalt baths were also investigated, but no commercial application has developed. Ternary alloys of tin – copper – zinc have been developed, and at least one bath was promoted commercially. However, it fell into disuse because of difficulties with control. More recently, two proprietary baths with simplified control procedures have been promoted, but applications and use appear to be limited. They were in part intended as replacements for chromium plating, and the development of a successful chromium(III) bath has contributed to the lack of acceptance. Patents exist on baths for plating tin – gold, tin – indium, tin – iron, and tin – silver, but commercial applications have been limited.

9.7. Zinc Alloys and Other Alloys

The zinc alloys with copper, nickel and tin are discussed under copper, nickel, and tin alloys. There is no other zinc alloy of commercial importance.

Numerous other alloys are mentioned in the literature, including gold – silver, gold – tin, silver – cadmium, and many others. Ternary alloys are also frequently mentioned. None of these, however, has experienced any major commercial development.

10. Nonelectrolytic Deposits

Deposits applied by nonelectrolytic methods can be classified into three groups: immersion or displacement deposits, noncatalytic chemically reduced deposits, and catalytic chemically reduced deposits (catalytic electroless plating).

10.1. Immersion Deposits

Immersion deposits form whenever the substrate is more active than the ions in solution. Although these deposits are often nonadherent and of poor physical quality, careful attention to solution composition and operating conditions can produce deposits that are acceptable for certain purposes. The zincate and stannate solutions used in plating aluminum (see 5.3.5) are examples of special cases producing acceptable deposits.

Liquor Finish. A copper – tin alloy deposited on steel wire from a solution of copper sulfate, tin sulfate, and sulfuric acid – liquor finish – is used on wire as a final finish or as a drawing lubricant. Proprietary solutions with modifying agents to improve the deposit are often used.

Copper. Copper from a dilute copper sulfate – sulfuric acid solution is used as a drawing lubricant on steel wire and occasionally on steel sheets that are to be deep drawn. Zinc, and occasionally magnesium phosphate, coatings have to a large extent replaced the use of copper.

Gold. Immersion gold is now rarely used. However, copper and its alloys may be coated in a solution of:

Potassium pyrophosphate	80 g/L (10.7 oz/gal)
Potassium cyanide	0.025 g/L (0.0033 oz/gal)
Gold as gold chloride	2.0 g/L (0.27 oz/gal)
Temperature	65 – 87 °C.

Tin. Tin is probably the most widely used immersion deposit. Pins, screws, buttons, and eyelets are typical. Tin immersion deposits are also used on the inside of brass lipstick cases to prevent discoloration of the lipstick by contact with the copper alloy. Numerous solutions exist. For example:

Solution 1	
Potassium hydrogen tartrate	saturated solution
Tin(II) chloride ($SnCl_2 \cdot H_2O$)	0.75 – 2.5 g/L (0.1 – 0.33 oz/gal)
Solution 2	
Tin(II) chloride ($SnCl_2 \cdot H_2O$)	15 g/L (2 oz/gal)
Sodium hydroxide (NaOH)	19 g/L (2.5 oz/gal)
Sodium cyanide (NaCN)	7.5 g/L (1.0 oz/gal)

Both solutions are operated boiling.

There are also proprietary solutions available with longer life, improved characteristics, and easier waste-disposal procedures.

10.2. Noncatalytic Chemically Reduced Coatings

Noncatalytic chemically reduced coatings have been in use for many years. Their two main drawbacks are that only relatively thin coatings can be deposited and that all surfaces, including the container, receive a coating. While these solutions are sometimes used by immersion, they are now more often applied by spraying the solutions from a dual spray gun system.

Silver. Silver was undoubtedly the first metal to be applied in this manner and was used to produce mirrors and coat nonconductors. One widespread use is for rendering phonographic record masters (lacquer or wax) conductive before electroforming the stampers used in the production of copies.

The original Brashear processes (sugar, Rochelle salt, silver nitrate, ammonia, and formaldehyde) have been abandoned because of the danger of explosive silver fulminate. Modern systems use an excess of ammonia, which avoids this problem. Furthermore, these solutions may be safely stored. The Brashear solutions had to be used immediately or discarded.

In the spray process, most of the silver can be made to deposit on the work by properly directing the spray and adjusting the spray pattern to avoid overspray. This is a much more efficient use of silver than in the immersion process. A typical two-part solution consists of:

Solution number 1	
Silver nitrate ($AgNO_3$)	10 g/L (1.3 oz/gal)
Aqueous ammonia (28 wt %)	4.5 g/L (0.59 fl oz/gal)
Solution number 2	
Hydrazine sulfate	19 g/L (2.56 oz/gal)
Sodium hydroxide	5 g/L (0.64 oz/gal)

An equivalent amount of potassium hydroxide may be substituted for the NaOH. Hydrazine nitrate or hydrate may be substituted in equivalent amounts for the hydrazine sulfate.

The solutions are sprayed in equal volumes from a dual spray gun. The speed of the reaction may be slowed by increasing the NH_3 content or decreasing the NaOH content.

Copper. While not as widely used, copper may be applied in a similar way.

Solution number 1	
Copper sulfate ($CuSO_4 \cdot H_2O$)	35 g/L (4.7 oz/gal)
Sulfuric acid (H_2SO_4)	50 g/L (6.7 oz/gal)
Solution number 2	
Sodium hypophosphite (NaH_2PO_2)	350 g/L (47 oz/gal)

The solutions are sprayed at a ratio of 2 parts of number 1 to 1 part of number 2 at 80 °C.

Gold. Spray solutions for gold have been patented and found some use in the early stages of the U.S. space program. The solutions are delicately balanced, relatively inefficient (ca. 50 % of the gold is wasted), and correspondingly expensive to use. They have not found general application.

10.3. Catalytic Electroless Plating

These systems stem from the work of Brenner and Riddell. The basic reaction requires the presence of a catalytic metal that allows reaction to proceed. This requires that the metal being deposited be itself a catalyst. The advantages of the systems are:

1) Provided there is adequate agitation to maintain fresh undepleted solution in contact with all surfaces at all times, the deposit is uniform over the entire surface even if the shape is highly complex.
2) Deposits are usually less porous than the same metal deposited electrolytically.
3) Racking or fixturing is greatly simplified.
4) Nonconductors may be coated.
5) Deposits often have special chemical or physical properties since the deposit is an alloy of the metal and a compound formed from the reducing agent. For example, with hypophosphite the deposit is an alloy of the metal and the metal phosphide, and with boron reducing agents the metal and the metal boride deposit.

Arsenic. Hypophosphite baths have been developed with rates of 2.5 µm/h or less and recommended thicknesses of 0.5 µm or less. No commercial application has developed.

Chromium. No truly autocatalytic bath has been verified.

Cobalt. Commercial processes are not generally available, but numerous proprietary methods of producing magnetic memory films have been developed. Published data indicates that, unlike nickel, cobalt may only be deposited from alkaline solutions.

A typical formula used mainly for memory disc production consists of:

Cobalt chloride ($CoCl_2 \cdot 10\,H_2O$)	30 g/L (4 oz/gal)
Sodium hypophosphite (NaH_2PO_2)	20 g/L (2.7 oz/gal)
Ammonium chloride (NH_4Cl)	50 g/L (6.7 oz/gal)
Sodium citrate ($Na_3C_6H_5O_7$)	35 g/L (4.7 oz/gal)
pH	9 – 10
Temperature	95 – 100 °C
Deposition rate	15 µm/h

Deposits are dull. Improved physical properties without improvement in appearance may be obtained by decreasing the hypophosphite to 10 g/L (1.3 oz/gal) and increasing the citrate to 100 g/L (13.4 oz/gal), but the rate of deposition falls to 5 µm/h. The efficiency of the bath appears to be about double the efficiency of nickel baths. Many of the additives developed by GUTZHEIT for electroless nickel are also effective in the cobalt baths.

Copper. Commercial applications date from the late 1950s. Major applications have been for through-hole printed circuit boards, the additive process for printed circuit boards, and for plating molded plastic parts.

The formulations are based in a general way on the Fehling test for aldehydes. There is a wide variation in formulas. A typical one would be:

Copper sulfate ($CuSO_4 \cdot 5\,H_2O$)	30 g/L (4.0 oz/gal)
Rochelle salt ($NaK(C_4H_4O_6) \cdot 4\,H_2O$)	135 g/L (18 oz/gal)
Formaldehyde (37 wt %)	150 g/L (20 oz/gal)
Sodium hydroxide (NaOH)	40 g/L (5.4 oz/gal)
Temperature	20 °C

Rate of deposit is up to 25 µm/h. Many different complexing agents may be used. A combination of EDTA and Rochelle salt appears to have a beneficial effect on the rate of deposition. Similarly, many different reducing agents are effective; formaldehyde and paraformaldehyde are preferred. The pH must be constantly adjusted since the progress of the reaction reduces the pH. The bath may be stabilized with very low levels of sulfur compounds, but, as also happens with nickel, slightly higher amounts slow or stop deposition. Cyanide, polysiloxanes, or substituted silanes may be used to improve the ductility of the deposit and improve its appearance.

Gold. There are several patented gold baths. It is however questionable whether they are truly autocatalytic. Useful solution life also appears limited. One bath, verifiably autocatalytic, was developed in the late 1970s. It is covered by patents and available only as a proprietary material.

Iron. Uses for electroless iron plating seem to be limited. Few verifiably effective autocatalytic baths exist. One bath developed by RUSCIOR and CROILA contains:

Ammonium iron(II) sulfate	30 g/L (4 oz/gal)
Sodium citrate	45 g/L (6 oz/gal)
Sodium hypophosphite	30 g/L (4 oz/gal)
pH (adjusted with NaOH)	>9.5
Temperature	60 – 90 °C

The rate is generally low, one figure cited being 2 µm/h. The coating is ferromagnetic.

Nickel. Nickel is by far the most widely used autocatalytic deposit. It has received the widest attention and, because of its commercial importance, has been most thoroughly researched. Much of this research has been carried out by plating solution suppliers, who have developed many highly specialized baths. One survey, conducted in 1984, showed that there were over 44 different proprietary electroless nickel solutions available from suppliers in the United States alone.

Nickel – Phosphorus Alloys. The present baths are generally based on the acid baths developed by BRENNER and RIDDELL.

The reactions are extremely complex, and a number of theories have been advanced to explain the complete mechanism. The most widely accepted theory is the hydride theory. This proposes that the interaction between the hypophosphite and water leads to the formation of a hydride ion, H^-. This hydride ion acts as the reductant for the metal and at the same time produces hydrogen, which is liberated as a gas: in acid solution,

$$H_2PO_2^- + H_2O \longrightarrow HPO_3^{2-} + 2H^+ + H^-$$

in alkaline solution,

$$H_2PO_2^- + 2OH^- \longrightarrow HPO_3^{2-} + H_2O + H^-$$

then,

$$Ni^{2+} + 2H^- \longrightarrow Ni^0 + 2H$$

Because H^- is unstable, only about one-half of the hypophosphite actually reduces nickel. Formulations consist of a source of nickel ions, sodium hypophosphite, buffers, complexing agents, accelerators, stabilizers, and wetting agents. The normal source of nickel ions is either nickel chloride or nickel sulfate. Since the baths are most generally supplied as prepared concentrates, these choices are dictated by the need for high solubility, high purity, ready availability, and moderate cost.

The concentration of hypophosphite in the bath can be critical. Low concentrations yield low rates, which increase as the hypophosphite concentration is increased to an optimum level. As the concentration exceeds the optimum level, the rate again falls off until a concentration is reached at which spontaneous decomposition of the bath occurs.

The materials used as buffers are closely linked with the complexing agents used. One or more ingredients frequently serve both functions. The main purpose of the buffers is to prevent a rapid increase in the acidity of the solution with an accompanying loss of deposition rate and changes in the characteristics of the deposits. Even with the use of appreciable concentrations of strong buffering agents, frequent additions of sodium hydroxide are needed to maintain the bath close to the optimum pH.

The complexing agents used are almost all organic acids, usually dicarboxylic acids. Their function is to reduce the concentration of free nickel ions. This helps stabilize the solution and prevent the precipitation of nickel phosphite. As the concentration of any one complexing agent is increased, the rate of deposition decreases. Therefore, combinations are often used to provide adequate complexing while maintaining rate levels. The exception to this rule is lactic acid, which in some formulations acts as an accelerator as well as serving the function of buffer and complexing agent.

Accelerators are used to overcome the rate depressing effects of complexing agents. Fluorides or sulfur-bearing compounds, such as thiourea, may be used. Concentrations are generally in the range 2 – 20 ppm.

A true electroless bath, in order to operate, must show some degree of instability. Given the right conditions, this instability may result in spontaneous decomposition, a violent reaction in which the rate increases exponentially and uncontrollably so that in a very short time the entire nickel content of the solution is reduced to a fine metallic powder. Even in the absence of complete decomposition, localized problems may occur as a result of localized overheating, local excesses of hypophosphite during additions, contamination with solids accidentally introduced, or the introduction of catalytic trace elements, such as palladium, from pretreatment processes.

Some of these problems can be prevented by proper circulation and filtration to keep the bath chemically and thermally homogeneous and free from particles.

A second preventive is the use of stabilizers. Ions such as Pb^{2+}, Cd^{2+}, Zn^{2+}, SCN^-, and CN^- have been used at concentrations of 1 – 10 ppm. Numerous other materials serving the same purpose have been found and in some cases patented. Compounds of selenium, tellurium, and a number of organic sulfur compounds, such as mercaptans, thiourea, xanthates, and sulfonates, have proved effective. Some of these stabilizers have been found to have secondary effects on the characteristics of the deposits and may improve surface luster.

Wetting agents are added to reduce the surface tension, thus bringing about the more rapid release of the hydrogen developed from the surface. This helps to eliminate pitting and gas streaking. Whatever type formulation is used, byproducts from the reactions and contaminants from pretreatment cycles and other sources eventually build up in the solution, making it necessary to discard and replace it. GUTZHEIT and others have devoted a great deal of research to methods of reclaiming solutions and establishing equilibrium conditions to permit unlimited operation, however, without complete success. The lives of the solution have been extended, but beyond certain limits they no longer function and cannot be reclaimed.

Rate is determined by the formulation, temperature, pH, and other factors such as agitation. An increase in pH also increases the rate up to the point where the hypophosphite spontaneously oxidizes and releases hydrogen, or the bath becomes so unstable as to decompose spontaneously.

The phosphorus content of the deposit varies with the operating conditions and the formulation between 2 and 14 wt % phosphorus, with most formulations yielding deposits in the 8 – 13 wt % range. Metallic ions other than nickel, introduced as additives or as contaminants, may codedeposit and alter the characteristics of the deposit. Antimony, bismuth, chromium, lead, mercury, and selenium may reduce the corrosion resistance of the deposits.

The deposits have a microcrystalline structure, usually laminar in nature, which appears to be Ni_3P in a matrix of nickel. The size of the Ni_3P crystals has been reported to range from 15 to 100 Å with an average size of 20 – 60 Å.

Heat treatment causes the nickel phosphide to segregate, increasing the hardness of the coating, developing greater wear resistance, and reducing the ductility. If the temperature exceeds 425 °C or is held close to that value for a considerable time, the deposit softens again and the ductility returns to the original level.

Although the corrosion resistance of the deposit varies from one formulation to another, it is almost always superior to that of electroplated nickel for equivalent thickness because above certain thicknesses electroless nickel is less porous than plated nickel. Great care must be taken with the pretreatment steps because of the effect of the pretreatment on porosity.

The following physical properties are typical for the deposits:

Phosphorous content	8 – 12 wt %
Hardness, as plated	550 – 650 VPN (100 g load)
heat treatable to	950 – 1050 VPN (100 g load)
Elongation before fracture	
as plated	0.5 – 2.0 %
heat treated	$\leq$6 %
Tensile strength	$>$700 MPa
Melting point	880 – 890 °C

The wear resistance of electroless nickel alloys is good in lubricated wear as long as the temperature remains low. Deposits may also be used to prevent galling between wearing surfaces provided only one surface is coated with electroless nickel.

Three typical formulations are shown:

Glycolic acid bath	
Nickel chloride ($NiCl_2 \cdot 6H_2O$)	30 g/L (4.0 oz/gal)
Sodium hypophosphite (NaH_2PO_2)	10 g/L (1.3 oz/gal)
Glycolic acid ($CH_2OHCOOH$)	35 g/L (4.7 oz/gal)
pH, adjusted with NaOH	4 – 6
Temperature	90 °C
Deposition rate	15 µm/h
Citric acid bath	
Nickel sulfate ($NiSO_4 \cdot 6H_2O$)	80 g/L (10.7 oz/gal)
Sodium hypophosphite	25 g/L (3.3 oz/gal)
Citric acid	95 g/L (12.7 oz/gal)
Ammonium chloride	50 g/L (6.7 oz/gal)
pH, adjusted with NaOH	4.8 – 5.8
Temperature	93 °C
Deposition rate	15 – 20 µm/h
Lactic – succinic acid bath	
Nickel sulfate	30.0 g/L (2.7 oz/gal)
Sodium hypophosphite	25.0 g/L (3.3 oz/gal)
Lactic acid	34.0 g/L (4.5 oz/gal)
Succinic acid	7.0 g/L (0.9 oz/gal)
Sodium fluoride	2.5 g/L (0.33 oz/gal)
Lead (Pb^{2+})	0.7 – 3.0 ppm
pH, adjusted with NaOH	4.5 – 4.7
Temperature	95 °C
Deposition rate	32 – 35 µm/h

Nickel – Boron Alloys. When a boron-containing reducing agent is used, a nickel – boron alloy is produced. These exhibit a higher as-plated hardness and have a melting point 200 – 400 °C higher than the phosphorus alloys. These nickel – boron alloys have not received the same level of industrial acceptance as the nickel –phosphorus alloys, but because of their different properties are used for special applications.

The formulations are very similar to the formulations using hypophosphite, with the exception of using a boron compound in place of the phosphorus compound. The original formulations were based on sodium borohydride. The baths had to be maintained at a pH $>$10.0, preferably 12 – 14 and required careful control to prevent solution decomposition.

Modern versions more often use dimethylaminoborane (DMAB). This reducing agent allows operation over a wider pH range, and the solutions have almost unlimited life since the breakdown products of the reducing agent have little or no effect on the operation of the solution. Deposition rates of 6 – 9 µm/h can be maintained. These rates are essentially independent of the nickel concentration and are mainly de-

pendent on the concentration of reducing agent. Two typical formulations are listed below:

Nickel sulfate	50 g/L (6.7 oz/gal)
DMAB	3 g/L (0.4 oz/gal)
Sodium pyrophosphate	100 g/L (13.3 oz/gal)
pH	10
Temperature	25 °C
Nickel chloride	24 – 48 g/L (3.2 – 6.4 oz/gal)
DMAB	3 – 4.8 g/L (0.4 – 0.64 oz/gal)
Ethylenediamine 50 %	18 – 37 g/L (5 – 10 oz/gal)
Potassium acetate	18 – 37 g/L (5 – 10 oz/gal)
pH	5.5
Temperature	70 °C

Hydrazine borane is also used in some formulations.

11. Chromating

Chromium, vanadium, molybdenum, and tungsten, when they are in their high valence states and soluble, have an unusual ability either to prevent corrosion or to limit the salt growth that occurs following corrosive action. Chromium(VI) compounds are the most widely used compounds of the group, and a whole series of chromate conversion coatings have been developed for various metals. Specifically, copper, brass, cadmium, zinc, and occasionally silver and gold are protected in this way. In general terms, the solutions consist of a soluble dichromate salt, such as sodium or potassium dichromate, and a catalyst similar to those used in chrome plating, such as SO_4^{2-}, F^-, or the other effective catalysts including complex fluorides or boric acid. Other anions may be added to alter the character of the film formed. The film is a colloidal gel that contains soluble chromium(VI), as well as salts of the base metal. The corrosion protection characteristics are dependent on the soluble Cr(VI) leaching out to protect the surface. If the parts are dried at a temperature close to or above 71 °C, the gel will dehydrate and the Cr(VI) reverts to an insoluble salt, and the protective mechanism is lost. Additionally, the gel will shrink as a result of the dehydration and exfoliate from the surface, becoming powdery and easily removed by abrasion. The original chromating solutions were relatively concentrated (120 – 240 g/L $Na_2Cr_2O_7$, 16 – 32 oz/gal) with catalyst in the ratio of 1 : 100). On zinc and cadmium, these produced iridescent yellow coatings that gave up to 72-h salt-spray resistance to the development of white corrosion salts. On copper, they provided resistance to the development of green salts of up to 24 h of salt spray, or up to 1000 h of 100 % humidity. Modern chromating solutions are almost all proprietary and operate at much lower concentrations. They fall into five categories:

1) Clear to bluish single dip chromates. Salt spray resistance ca. 24 h.
2) Clear leached type chromates. These are iridescent yellow chromates that are subsequently leached in a sodium hydroxide – sodium carbonate solution to provide a clear coating. Salt spray resistance up to 48 h.
3) Iridescent yellow chromates, single dip. Salt spray resistance up to 72 h.
4) Olive drab chromates, single dip. Salt spray resistance up to 144 h.
5) Black chromates. These may be single dip chromates in which the black color is provided by silver salts, which deposit microscopic amounts of silver in the gel. These chromates generally have less corrosion resistance than an uncolored chromate since the silver particles provide active electrochemical cells for the initiation of corrosion. Salt spray resistance 24 – 48 h.

There are also double dip systems in which an iridescent or olive drab coating is applied and subsequently dyed black using an organic dye.

Immersion times are from 10 to 30 s. Temperature and agitation affect the formation of the coating. Transfer time to the rinsing step can be important, and on automatic equipment where transfer times may run 30 – 45 s, special formulations may be necessary.

12. Oxide Coatings

12.1. Immersion

Immersion type oxide coatings are used mainly for appearance, as a paint base, or for their oil-retention characteristics.

Steel. Steel may be blackened in a high-temperature bath containing 480 – 840 g/L

(64 – 112 oz/gal) of a mixture of ca. 75 wt % sodium hydroxide and 25 wt % sodium nitrate. In this range, the solution boils under atmospheric pressure at 121 – 149 °C. The solution is controlled by monitoring the boiling point rather than by chemical analysis. Salts or water are added as required to maintain the boiling point and the solution is operated at the boil. Instead of 25 wt % sodium nitrate, 12 wt % sodium nitrate and 13 wt % sodium nitrite may be used.

The lower boiling points are obtained by including various sulfur-containing additives. The resulting oxide coating is somewhat less corrosion resistant as a result of the incorporation of iron sulfide in the film. Some of the sulfide-modified materials may also be used to blacken stainless steel. The resulting oxide, however, has relatively low corrosion resistance. The oxide coating is quite hard and therefore does not lend itself to use where decorative relieving is planned. There are proprietary immersion deposits available that are better for this purpose.

Following thorough rinsing, the oxide film should be coated with an oil, wax, or lacquer.

There are numerous gun blueing and gun browning formulas available in various craft handbooks. These are generally lengthy, labor-intensive procedures and are not used for production purposes.

Stainless steel can also be blackened in molten sodium/potassium dichromate at ca. 370 °C provided the parts are not adversely affected by the high temperature. The resulting oxide has good corrosion resistance, and the procedure is covered by several military specifications.

Copper. Copper may be blackened in a bath containing 120 g/L (16 oz/gal) of a mixture of 75 wt % sodium hydroxide and 25 wt % sodium chlorite at 93 – 100 °C. The black copper(II) oxide produced is acicular in form and presents a fuzzy appearance. On bright finished parts, this fuzzy nap disappears when lacquer is applied, restoring the lustrous appearance, or it may be laid down (not removed) by gentle buffing or tumbling in sawdust or corncob grit.

Brass may sometimes be blackened in the same solution but, depending on zinc content and metallurgical history, may not respond. Activation may be carried out by immersion in a bath containing 120 g/L of a mixture of 85 wt % sodium hydroxide and 15 wt % sodium chlorite operated at 93 – 100 °C for 5 – 10 min to dezincify the surface, after which the blackening bath will usually function normally.

Aluminum. Aluminum may be given an oxide coating by immersion by a number of different processes including the Alzac and Juratka processes. However, electrolytic processes are generally preferred.

12.2. Electrolytic Oxidation

Many different metals may be electrolytically oxidized or anodized, but in industrial practice the procedure is usually limited to aluminum, magnesium, and titanium.

12.2.1. Aluminum

Aluminum may be anodized in hundreds of different baths. Industrially, sulfuric, chromic, and oxalic acid, and mixtures of these, sometimes with added salts, are used. Sulfuric acid is probably the most widely used.

A wide variety of current forms and wave shapes may be used for the process, depending on the solution selected and the purpose of the anodic film. Direct current (d. c.), alternating current (a. c.), and d. c. with superimposed a. c. are all used in various processes.

Oxide coatings produced from *oxalic acid solutions* are hard, relatively nonporous, and thin, and have the highest dielectric strength. Such coatings have been used as insulation on high-voltage transmission wires. The process also requires the highest operating voltage, up to 65 V in some processes. Coloring the coating by subsequent dye absorption can be difficult, but some modified oxalic acid solutions have been developed which produce very attractive, startingly clear results. Others produce an opaque effect that is especially attractive in pastel shades.

Coatings produced from *chromic acid* solutions are somewhat dark in color, often have a yellow tinge, are intermediate in hardness and thickness between oxalic and sulfuric acid processes, and require operating voltages up to 40 V. The coatings may be colored by dyeing, but pastel colors are not generally obtainable. Chromic

acid anodizing is preferred by the military for anodizing assemblies with lap seams or other areas that may trap solution since any solution so trapped not only does not corrode the base metal but also provides chromium(VI) to act as a corrosion preventative. Chromic acid coatings, when sealed in a sodium dichromate solution, provide the highest corrosion resistance of any of the anodic coatings.

Sulfuric acid processes can produce the thickest, softest, most porous coatings. Because of this, they are the preferred type of coating where the coating is to be subsequently colored by dye absorption.

Hard anodizing is a special case of sulfuric acid anodizing in which the solution is maintained at a very low temperature. The coating is extremely hard and wear resistant. If maximum hardness is wanted, the coating is left in the unsealed condition since the sealing process reduces hardness.

As formed, the coatings are porous and absorptive. Advantage of this may be taken to color the coatings by absorbing dyes or forming pigments within the coating. Architectural colors (mostly browns and bronzes) may be developed as part of the anodizing process by using special alloys and solutions. The color is a function of the alloy, the anodizing solution, and the conditions of treatment.

Organic dyes, specially selected for color fastness, are available in almost all shades: pastels, imitation brass and bronze, gold, and deep shades including black. The depth of color may be controlled by the thickness and porosity of the coating (determined by the application times and temperatures), the dye concentration, pH of the dye bath, and dyeing time. For maximum color uniformity and light fastness, dyeing times should not be short; rather, the coating should be dyed to saturation and the depth of color controlled by adjusting the dye concentration. Normal conditions are 3 – 10 g/L (0.4 – 1.3 oz/ gal) of dye, pH 4.5 – 6.0 (electrometric), and a time of 5 min. For black dyes, the concentration may be 10 – 20 g/L (1.3 – 2.7 oz/gal) or more.

Light-fast *inorganic pigments* may be developed within the coating by a two-step dyeing process. The parts are successively immersed in salt solutions, which precipitate the pigment. Some typical combinations are the following:

Red brown – copper sulfate followed by potassium ferrocyanide.
Brown – silver nitrate followed by potassium dichromate or chromate.
Dark brown – lead acetate followed by ammonium sulfide.
Yellow – lead acetate followed by potassium dichromate or chromate.
White – lead acetate followed by sodium sulfate.
Blue – iron(III) sulfate or chloride followed by potassium ferrocyanide.
Black – cobalt acetate followed by sodium sulfide.

Other possible combinations can be drawn from a knowledge of inorganic pigment chemistry.

A development is the electrolytic deposition of tin and tin salts within the film to produce colors.

Whether the film is colored or not, the final step in the process is *sealing*, which closes the pores of the coating and prevents undesirable absorption.

Clear coatings may be sealed in distilled or deionized water. Coatings colored with organic dyes benefit from being sealed in a solution of nickel acetate, cobalt acetate, or a mixture of the two. The nickel or cobalt insolubilizes the dye, improves light fastness, and seals the coating.

Added corrosion resistance at the expense of developing some yellow color is obtained by sealing in sodium dichromate solution. Almost all chromic acid coatings, since they are intended for corrosion resistance, are so sealed.

A common problem in sealing is the development of *sealing bloom*, a powdery film that develops on the surface of the coating. It is due to the growth of aluminum salts out of the pores in the coating. It may be prevented by sealing to 75 – 80 % completion, allowing the final sealing to occur in air, or by incorporating a suitable bloom prevention material in the sealing solution. These are usually proprietary and are based on mild complexing agents for aluminum such as sodium gluconate, sodium gluco-heptonate, or one of the other compounds that are also useful for preventing the development of tank scaling in caustic etching solutions for aluminum.

12.2.2. Magnesium

Magnesium anodizing processes have been available for many years. They were, however, supplanted for many years by immersion chemical treatments, which gave equal or superior corrosion preventative and paint bonding characteristics. The chemical treatments still account for a high percentage of the coatings applied, but anodic treatments have been developed which produce relatively thick, dense, adherent, abrasion-resistant coatings. Two general types are in use: those based on modified acid fluorides and caustic types.

The *modified acid fluoride process* is covered by a U.S. military specification (Mil-M-45202) and is applied in three different forms:

1) Clear. A very thin 40 V coating suitable for use as a base for clear lacquers or paints and similar in appearance to clear anodizing on aluminum. Anodizing time is 1 – 2 min.
2) Low-voltage thin coating. Final voltage is 60 – 75 V to give the best combination of protective value and paint-base properties. Anodizing time is 2.5 – 5 min, depending on the alloy and the current density. The coating is light gray to pale green in color; thickness is 5 – 10 µm.
3) Full coating. Final voltage is 75 – 95 V. This type has the best combination of abrasion resistance, protective value, and paint-base characteristics. Coating is medium green in color; thickness is 23 – 30 µm.

Two types of solution are in use, and either direct current or alternating current may be used:

Ammonium hydrogen fluoride	240 g/L (32 oz/gal)
Sodium dichromate	100 g/L (13.3 oz/gal)
Phosphoric acid	9 % vol

For use with direct current the fluoride salt should be increased to 360 g/L (48 oz/gal).

Ammonium hydrogen fluoride	200 g/L (27 oz/gal)
Ammonium hydrogen phosphate	100 g/L (13.3 oz/gal)
Sodium dichromate	75 g/L (10.0 oz/gal)

For use with direct current, the fluoride salt should be increased to 270 g/L (36 oz/gal).

The bath must be heated to 82 °C to ensure complete solution of all ingredients. If allowed to cool during idle periods, the bath must be reheated to this temperature and thoroughly agitated before use. The bath does not operate below 60 °C, but may be operated up to the boil.

Best results are obtained if the process is carried out at a relatively constant current. Anodizing is therefore started at a low voltage, and the voltage continuously raised to maintain the desired current density until the desired coating is obtained. Between 2 and 50 A min dm^{-2} (20 – 500 A min ft^{-2}) is required, depending on the coating mass. The terminal voltage varies with the alloy and the current density, but the ampere minute requirement for any particular coating is a constant for any one alloy. After anodizing, parts are rinsed in cold and hot water and dried. If parts are to be left unpainted, or painted only in certain areas, they may be sealed in a water glass solution (7 wt % silicate) for 15 min at 93 – 100 °C.

The H.A.E. anodize treatment is typical of the *modified caustic treatments* and is covered by the same military specification as the acid treatment. It is applied with alternating current; direct current is not satisfactory. Again, three types of coating may be applied from the same bath with slightly different operating conditions:

1) Low voltage. A softer coating, smooth and olive-drab in color. Treatment time is 15 – 20 min at ca. 9 V, a current density of 4 A/dm^2 (40 A/ft^2) and a bath temperature of 60 – 65 °C.
2) High voltage (light coating). Similar to the low voltage coating, but tan in color. The bath is operated at 21 – 26 °C and cooling of the solution is generally required. The terminal voltage is 60 V at a current density of 1.8 – 2.0 A/dm^2 (18 – 20 A/ft^2), and processing time is ca. 8 min.
3) High voltage (hard coating). Brown in color and somewhat rougher than the low-voltage coatings. Terminal voltage is 85 V, current density is 1.8 – 2.0 A/dm^2 (18 – 20 A/ft^2), treatment time is 60 – 75 min, and temperature is 21 – 26 °C. Cooling is required.

The solution has the following composition:

Potassium hydroxide	165 g/L (22 oz/gal)
Aluminum hydroxide	34 g/L (4.5 oz/gal)
Potassium fluoride	34 g/L (4.5 oz/gal)
Trisodium phosphate	34 g/L (4.5 oz/gal)
Potassium permanganate	19 g/L (2.5 oz/gal)

After anodizing, the parts are rinsed thoroughly in cold water and then treated for 1 – 2 min in a solution containing 100 g/L (13.3 oz/gal) ammonium hydrogen fluoride and 20 g/L (2.7 oz/gal) sodium dichromate. The parts are then rinsed and dried. This posttreatment is necessary to neutralize the residual alkalinity in the coating for good paint adhesion and also to improve the protective value of the coating, particularly in unpainted areas.

12.2.3. Titanium

Titanium may be anodized in almost any conductive solution. A sodium hydroxide process has been developed in Britain; chromic acid-fluoride processes have been used; and a proprietary neutral salt solution is used to coat fasteners for aircraft. High voltages (90 – 110 V) are required to produce heavy coatings although some processes use as low as 10 V. As the coating builds in thickness, a whole range of interference colors develop, which have been used by artists working in titanium. Industrially, the coating is used to provide better adhesion for solid-film lubricants and paints and to prevent spalling of the surface.

12.3. Colors of Oxide Layers

Oxide layers, if they are thick enough, exhibit the natural color of the oxide. However if they are transparent and thin, light may be reflected and re-reflected within the coating to produce colored interference patterns, the colors varying with the thickness of the coating and the wavelength of the incident light – blues, greens, reds, and yellows – and often changing with the angle of incidence of the light and the angle at which the surface is viewed.

Two English artisans have done an extensive investigation of ways to produce various colored and textured finishes on various metals. They have collected hundreds of such formulae in a volume illustrated with color plates of the finishes described [7].

13. Testing of Electrodeposited Coatings

13.1. Quality Control

The production of quality metal finishes depends on the maintenance of a strict, comprehensive quality control program. This program should have as a minimum three phases: (1) solution control, (2) preplate inspection, and (3) postplate inspection.

Solution Control. Solutions must be analyzed on a regular basis and maintained by proper additions. Generallly, ± 10 % of the control concentration is adequate. However, some solutions require a narrower control limit, e.g., the chromic acid : catalyst ratio in chrome plating. Additions should be made during idle periods if possible. Otherwise, the salts should be dissolved separately and pumped or filtered into the bath. In this way, any solid particulate matter that has not been filtered out of the bath and has settled to the bottom is not disturbed by the addition process to cause rough deposits.

Proper solution control also requires periodic monitoring of brightener content and balance, and testing for the presence of impurities with proper corrective steps being taken as indicated by the test results. Maintenance of a uniform brightener level is particularly important. Consideration should be given to the use of brightener pumps – either continuous feed or connected to an ampere hour meter. If a continuous feed pump is used, it should be interlocked with some other piece of equipment that is on only when work is being processed, e.g., the rectifier control or the run switch of an automatic machine. Such pumps not only deliver more uniform plated deposits, but save considerable amounts of brightener.

Proper recording of analytical results is important. Trends that could eventually lead to solution problems and defective work cannot be detected from a single analysis, but will become evident when a number of analyses are seen at the same time. Graphing of analytical results

may be even more helpful in detecting these trends.

Similarly, records of rejection rates (quality defective parts) are important to serve as a baseline for troubleshooting procedures in the event that problems are encountered. Additionally, trends in these rejection rates can be seen and often point to a potential problem before it becomes critical.

Filtration should be either continuous or periodic, as required by the solution. An effective monitoring system for the filters should be implemented to ensure that they are cleaned on a regular basis so that flow rate and filtrate clarity is maintained.

A maintenance program for rectifiers and instrumentation also reduces costs. This should include cleaning and check of output for rectifiers, checking calibration of temperature controls, cleaning bus bars and connections, checking anodes or anode baskets for proper contact, condition of anode bags, etc.

Preplate Inspection. Conditions that cause defective work often exist before the parts are plated. Such conditions as poor buffing or polishing, excessive edge burrs, cracking or crack initiation in formed parts, excessive metal grain size, residual magnetism, and many other conditions can cause problems in plating or be the cause of parts being rejected after plating. These conditions should be detected and corrected before parts are processed.

Postplate Inspection. The minimum level of postplating inspection is a visual examination of some or all the parts for obvious visual defects, i.e., lack of brightness, pitting, peeling, blistering, etc. For work being plated to a specification, other tests must be added, e.g., thickness and/or thickness distribution, salt-spray resistance, abrasion resistance, or other specified characteristics.

In most of these tests, the specified property is not measured directly. Rather, some related characteristic is measured, and the results of this measurement are related to the desired property by comparison with a standard. Calibration of the measuring equipment is critical and should be repeated at intervals such that the calibration remains accurate.

Some tests, such as salt-spray and humidity test, require days to complete. Where specifications involve such tests, failures can be avoided by setting and observing other quicker inspection standards that are affected by corrosion resistance of the parts, e.g., zinc that is to be salt-sprayed can be inspected for thickness and color of the chromate coating. If proper standards for these supplementary tests are maintained, salt-spray failures are unlikely.

Some finishes, e.g., brass plating or anodized and dyed aluminum, require that a standard color be matched. Standards should be approved before production is started, and working standards provided to the inspectors. These working standards should represent the maximum allowable deviation from the original standard, thus setting limits for acceptance of work being inspected. Since these working standards may change with time as a result of use and exposure to light, they should be checked and replaced at intervals as required to maintain consistent output. The lighting of the inspection station (see Section 4.4) should be the same as that used to establish the master standards.

13.2. Thickness Testing

Deposit thickness may be tested by destructive and nondestructive methods.

13.2.1. Destructive Testing

Destructive tests include: (1) strip and weigh, (2) the dropping test, (3) the jet method, (4) the electrochemical stripping test, and (5) microscopic tests.

Strip and Weigh. This is most often used for precious metal deposits on small parts. A number of small parts are weighed, stripped in a stripping solution that does not attack the base metal, and weighed again. The area of the pieces is calculated, and the weight loss per unit area converted to thickness. This method provides an accurate measure of the average thickness of deposit.

A variation of the method does not involve weighing the pieces. The parts are stripped in the smallest workable volume of stripper, and

the stripping solution analyzed for the concentration of stripped metal. This concentration is converted to total weight of stripped metal and, using the area of the pieces, converted to an average thickness. This variation has the advantage that attack on the substrate is unimportant.

Dropping and Jet Tests. In the dropping test, a stripping solution is allowed to drop on the surface to be tested, and the number of drops required to expose the substrate is converted to thickness. In the jet method, a jet of controlled size is allowed to impinge on the surface, and the time required to expose the substrate is converted to thickness.

Both methods can be used to measure discrete areas on a surface. However, the area to be tested must have sufficient drainage that the stripping solution flows freely away from the test area.

Electrochemical Stripping Test. Various instrument manufacturers make electrochemical stripping equipment.

A small area of the surface is isolated by a rubber cup that also contains an electrode. An electrical contact is clipped to another area of the part. The rubber cup is filled with a special electrolytic stripping solution. At the beginning of the test, an electrical timer is set to zero, and a controlled current is applied to the system with the part being made anodic. When the metal being measured has been completely stripped from the test surface, the substrate is polarized by the stripping solution. The increase in voltage resulting from this polarization activates a relay that ends the test and stops the timer.

The stripping solution and the applied current are so selected that the thickness can be read directly from the timer. The method has the advantage that by properly selecting and changing the stripping solutions, successive layers of different deposits may be measured on the same spot. Charting the voltages needed to maintain the standard current may allow different layers of the same metal, e.g., nickel layers from different solutions, to be measured.

For most production purposes, the calibration given with the stripping solutions is of sufficient accuracy. For optimum accuracy, however, the instrument should be calibrated against standards plated from the same solution.

Microscopic Tests. The parts are mounted, polished, etched, and examined with a microscope. Deposit thickness is measured directly with a vernier in the eyepiece and a stage micrometer. The ASTM in its standard for such measurements, B-487, limits the use of the method to deposits at least 8 μm in thickness. This is one of the few direct measurements. However, it is subject to errors in polishing, etching, and observation, hence the restriction. Even with sophisticated equipment, the error apparently cannot be reduced to less than 0.15 – 0.25 μm.

13.2.2. Nondestructive Testing

Nondestructive thickness tests include: (1) magnetic force methods, (2) magnetic flux methods, (3) eddy current methods, (4) X-ray methods, and (5) beta backscatter methods.

Magnetic Force Methods. A magnetic method is only effective with nonmagnetic coatings on magnetic substrates or with magnetic coatings on nonmagnetic substrates. A magnet is brought into contact with the surface and the force required to release it is measured. This force must be compared against carefully prepared standards. Accuracy is affected by variations in the roughness of the surface, the substrate alloy, and for magnetic coatings the conditions under which the coating was deposited.

Magnetic Flux Measurement. These instruments measure the reluctance of a magnetic flux path passing through the coating and the substrate. They are used for the same coating – substrate combinations as the magnetic attraction methods, but have a greater range of thickness capability. Lower limit thickness is ca. 5 μm and upper limit about 1 mm. Calibration is important.

Eddy Current Methods. The probe generates a high-frequency electric field in a coil mounted in an open ferritic pot core. When the probe is in contact with a metallic substrate, eddy currents are induced in the metal. The magnitude of these currents depends on the distance

between the probe and the surface, which is essentially the thickness of the nonconductive or plated coating. These currents generate a magnetic field, which in turn interacts with the probe and produces current changes that are measured. These measurements can then be related to the coating thickness through the use of calibrated thickness standards.

Measurements are affected by the frequency used in the measuring probe, the conductivity of the coating and the substrate, the magnetic permeability of the coating and the substrate, and the roughness of the surface. Most manufacturers of these instruments do not recommend their use for conductive coatings on metal, but one manufacturer claims that if the differences in the conductivity of the coating and the substrate are sufficiently great, eddy current devices can be used when properly calibrated.

There are enough common characteristics between magnetic reluctance measuring instruments and eddy current instruments that at least one manufacturer builds a dual purpose instrument that can be switched from one mode to the other.

X-Ray Methods. Two X-ray methods are in use. One measures the X-ray fluorescence emanating from the coating, the other the X-ray fluorescence of the substrate. The fluorescence is produced by irradiation of the target area by a beam of X-rays. One of the frequencies of the expected stimulated radiation is selected for measurement and the result converted to thickness. The atomic number of the deposited metal affects the thickness range that may be measured; the higher the atomic number, the higher the upper limit for measuring thickness. The lower limit is less than 0.1 μm; hence, the method is often used for measuring gold. The upper limits are in the range of 1 – 100 μm, depending on the coating system. The accuracy can be ± 10 % or better.

Beta Backscatter Methods. These are similar to X-ray methods, but use a source of beta rays instead of X-rays.

13.3. Adhesion Tests

Adhesion is a micro phenomenon (see Section 5.1). It occurs between atoms or molecules. But by their very nature, adhesion tests are macro phenomena and therefore can only test the average adhesion across a surface. Further, there can be tremendously wide variations in the effectiveness of the part preparation cycle across the surface of the piece so that an area of excellent adhesion may be immediately adjacent to an area of zero adhesion. Adhesion tests must be combined with careful visual examination.

All adhesion tests are destructive tests and involve bending, gouging, grinding, or other methods of forcefully removing the plated layer. The results and the interpretation of the results are subjective, not absolute, and when used to evaluate production, are used as go/no go trouble shooting tests rather than for routine quality control. While there are a few tests that provide measured pull data, e.g., the Jacquet test, they involve the use of specialized and relatively expensive equipment and, with the exception of testing plated plastics using a modification of the Jacquet test, are rarely used except as research tools.

13.4. Corrosion Testing

No accelerated corrosion test can truly duplicate field corrosion conditions. As a result, correlation between performance in corrosion tests and performance in the field is almost never available. While a number of accelerated corrosion tests have been developed and are used as part of the performance requirement in specifications for plated coatings intended for outdoor exposure, confidence in the results appears to be lacking. However, as a general rule, parts that fail the corrosion test will fail in service, yet parts that pass the corrosion test may also fail in service. When failure occurs, it can often be attributed to insufficient plating thickness. Therefore, the corrosion performance test is usually supplemented by a minimum thickness requirement as part of the specification. Tests include: (1) the salt spray test, (2) the copper-accelerated acetic acid salt spray test, (3) the Corrodkote test, (4) the sulfur dioxide test, (5) the STEP test, and (6) the electrolytic corrosion test.

Salt Spray Test (ASTM Method B-117). Parts are subjected to a fog of 5 wt % sodium chloride solution under fixed conditions for specified periods of time.

Copper-accelerated Acetic Acid Salt Spray Test (Cass test, ASTM Method B-368). The 5 wt % salt solution is accelerated with 1 g of copper(II) chloride for each 3.8 L of salt solution, and the pH adjusted to 3.2 with acetic acid. The parts are then subjected to a fog of this solution under fixed conditions for specified periods of time.

Corrodkote Test. A paste containing copper(II) nitrate, iron(III) chloride, ammonium chloride, and kaolin is painted on the surface. After partial drying, the test piece is exposed in a humidity cabinet maintained at $100 + 2 - 3\,^{\circ}C$ and a relative humidity between 80 and 90 %. After cleaning the corroded surface, the corrosion sites are redeveloped in a salt spray or humidity cabinet. This test is specified as an alternate to the Cass test for automotive and other applications where severely corrosive environments are expected.

Sulfur Dioxide Test. Parts are exposed in a humidity cabinet maintained at 100 % humidity with SO_2 added to the atmosphere. The cabinet is heated to $40 \pm 3\,^{\circ}C$ for 8 h and cooled for 16 h in repeating cycles.

Simultaneous Thickness and Electrochemical Potential (STEP) Test. This test is used for the evaluation of multilayered nickel – chromium decorative coatings. It is not a corrosion test. The test consists of electrolytically dissolving the various layers and monitoring the potential changes as each layer is removed. These changes in potential can be interpreted in terms of coating performance and suggest corrective changes in the operating conditions of the plating solutions if necessary. The thickness of the nickel layers is also measured.

Electrolytic Corrosion (EC) Test. This is a rapid method for evaluating copper – nickel – chromium coatings on zinc-base or steel substrates. It is a potentiostatically controlled method of corroding the metal layers in a relationship that corresponds to outdoor exposure. Substrate attack may be observed by adding indicators to the electrolyte. An alternate method is to run the test for a specified period of time and assess the degree of corrosion by measuring the corrosion sites with an interference microscope. It is said to give the equivalence of one-year outdoor exposure in a few minutes.

13.5. Other Tests

Various other tests may be used for other requirements, e.g., abrasion tests for hard anodized coatings. The electronic industry in particular has a number of specialized tests designed to measure the durability and corrosion resistance of plated coatings used for contacts and other purposes. Solderability is a major concern in electronic manufacture, so there are a number of tests to determine ease of solderability, integrity of soldered connections, and the ability of a coating to maintain its solderability characteristics during the storage prior to assembly.

14. References

General References

The first three references contain extensive bibliographies.

Journals: The American Electroplaters and Surface Finishers Society (AESF) in the United States publishes *Plating and Surface Finishing,* and The Institute of Metal Finishing in England publishes the *Transactions of the Institute of Metal Finishing*. The annual index of these journals reveals finishing trends and sources of the latest information. Two other journals are *Metal Finishing* and *Products Finishing*. *Metal Finishing* publishes annually, usually in the February issue, a survey of technical developments during the previous year with an extensive bibliography (over 800 references in 1984).

Other sources: The AESF has an arrangement with a university to provide a data base service. A source often neglected is the sales literature and operating bulletins of supply companies, which, while obviously aimed at producing sales, often contains useful information in a condensed form.

1. L. J. Durney, ed.: *Electroplating Engineering Handbook,* 4th ed., Van Nostrand Reinhold, New York 1984.

2. F. A. Lowenheim, ed.: *Modern Electroplating,* John Wiley & Sons, New York 1974.
3. S. Wernick, R. Pinner: *The Surface Treatment and Finishing of Aluminum and its Alloys,* 4th ed., Robert Draper, London, England, 1972.
4. *The Canning Handbook on Electroplating,* 22nd ed., W. Canning, Birmingham, England, 1978.
5. *The Metal Finishing Guidebook,* Metals and Plastics Publications, Hackensack, N.J., 1986.
6. Joseph B. Kushner: *Water and Waste Control for the Plating Shop,* Gardner Publications, Cincinnatti, Ohio, 1976.
7. R. Hughes, M. Rowe: *The Coloring, Bronzing, and Patination of Metals,* Van Nostrand Reinhold, New York 1983.
8. H. W. Dettner, J. Elze: *Handbuch der Galvanotechnik,* 3 vols., Hanser, München 1963 – 1969.
9. *Praktische Galvanotechnik,*4th ed., E. G. Leuze Verlag, Saulgau 1984.

Specific References

10. ISO Specification 2179 (1972).
11. British Standard BS 3597 (1963).

Electrochemical Machining → Electrochemistry

Electrochemistry

HARTMUT WENDT, Technische Hochschule Darmstadt, Darmstadt, Germany (Chaps. 1, 2, 3.1, 3.2, 4.2 – 4.4, and 4.8)

HELMUT VOGT, Technische Fachhochschule Berlin, Berlin, Germany (Chaps. 3.3, 3.4, and 4.1)

GERHARD KREYSA, Dechema, Frankfurt/Main, Germany (Chaps. 3.5, 3.6 and 4.6)

DIETER M. KOLB, Universität Ulm, Abteilung für Elektrochemie, Ulm, Germany (Chap. 3.7)

GERALD E. ENGELMANN, Universität Ulm, Abteilung für Elektrochemie, Ulm, Germany (Chap. 3.7)

JÖRG C. ZIEGLER, Universität Ulm, Abteilung für Elektrochemie, Ulm, Germany (Chap. 3.7)

HUBERT GOLDACKER, Forschungszentrum Karlsruhe, Karlsruhe, Germany (Chap. 4.5)

KLAUS JÜTTNER, Dechema, Frankfurt/Main, Germany (Chaps. 4.6 and 4.7)

ULRICH GALLA, Forschungszentrum Karlsruhe, ITC-CPV, Karlsruhe, Germany (Chaps. 4.5 and 4.6.3)

HELMUT SCHMIEDER, Forschungszentrum Karlsruhe, ITC-CPV, Karlsruhe, Germany (Chaps. 4.5 and 4.6.3)

EBERHARD STECKHAN, Universität Bonn, Kekulé-Institut für Organische Chemie, Bonn, Germany (Chap. 5)

1. Introduction . . . 3
2. Basic Laws and Equations . . . 3
2.1. Electrode Kinetics . . . 4
2.2. Technical Electrocatalysis . . . 4
2.2.1. Principles of Electrocatalysis . . . 4
2.2.2. Technical Electrodes . . . 7
2.2.2.1. Electrocatalytically Activated Dimensionally Stable Chlorine Evolving Electrodes . . . 7
2.2.2.2. Oxygen Evolving Anodes . . . 8
2.2.2.3. Electrocatalysis of Cathodic Hydrogen Evolution . . . 10
2.2.2.4. Electrocatalysis of Cathodic Oxygen Reduction and Anodic Hydrogen Oxidation in Fuel Cells . . . 12
2.3. Ionic Conduction in Electrolytes . . . 16
2.4. Fluid Dynamics and Mass Transfer . . . 17
3. Electrochemical Cells . . . 18
3.1. Optimized Current Density . . . 19
3.2. Primary, Secondary, and Tertiary Current Distributions . . . 20
3.3. Gas-Evolving Electrodes . . . 22
3.4. Cells with Two-Dimensional Electrodes . . . 26
3.5. Special Cell Designs . . . 29
3.6. Cells with Three-Dimensional Electrodes . . . 34
3.7. Electrochemical Nanostructuring with a Scanning Tunneling Microscope . . . 39
4. Inorganic Electrochemical Processes . . . 41
4.1. Chlor-Alkali Electrolysis . . . 41
4.1.1. Molten Salt Electrolysis . . . 42
4.1.2. Production of Chlorine . . . 42
4.1.3. Electrosynthesis of Hypochlorite . . . 44
4.1.4. Electrosynthesis of Chlorate . . . 44
4.1.5. Electrosynthesis of Perchlorate . . . 45
4.2. Water Electrolysis for Hydrogen Production . . . 45
4.3. Anodic Generation of Peroxodisulfuric Acid and Peroxodisulfates . . . 49
4.4. Electrowinning and Electrorefining of Metals . . . 49
4.4.1. Aqueous Electrolytes . . . 49
4.4.2. Melts . . . 53
4.5. Electrochemical Processes in Nuclear Fuel Reprocessing . . . 55
4.5.1. Electroredox Separation Processes . . . 55
4.5.2. Electroreduction Processes . . . 57
4.5.3. Electrooxidation . . . 59
4.5.4. Corrosion . . . 60
4.6. Electrochemical Water and Effluent Treatment . . . 60
4.6.1. Cathodic Treatment . . . 60
4.6.1.1. Optimization of Cell Design . . . 61
4.6.1.2. Electrochemical Reactors and Their Applications . . . 62
4.6.1.3. Operation Data of Electrochemical Cells . . . 65
4.6.2. Electrodialysis . . . 66
4.6.3. Anodic Treatment . . . 67
4.6.3.1. Direct Oxidation at the Anode . . . 67
4.6.3.2. Indirect Oxidation . . . 68

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_183

4.7. Electrochemical Gas Purification 70
4.7.1. General Aspects 70
4.7.2. New Process Developments 71
4.8. Electrochemical Shaping 73
5. Organic Electrochemistry 76
5.1. Introduction 76
5.2. Perspectives of Electroorganic Reactions in Industry 77
5.3. Syntheses 79
5.3.1. Direct vs. Indirect or Mediated Electrochemical Reactions 79
5.3.2. Technically Important Processes Ordered by Reaction Classes 81
5.3.2.1. Electrochemical Substitution 81
5.3.2.2. Electrochemical Addition 85
5.3.2.3. Electrochemical Transformations of Functional Groups 88
5.3.2.4. Electrochemical Cleavage and Elimination 90
5.3.2.5. Electrochemical Coupling 91
5.3.2.6. Bioelectrosynthesis 93
5.3.3. Organic Electrochemistry for a Cleaner Environment 93
5.4. Experimental Factors and Techniques 93
5.4.1. Solvent – Electrolyte Systems 94
5.4.2. Electrodes 95
5.4.2.1. Cathodes 96
5.4.2.2. Anodes 96
5.4.2.3. Other Factors Influenced by the Electrode Material 97
5.4.3. Electrochemical Cells 98
5.4.3.1. General Considerations 98
5.4.3.2. Cell Design 99
5.5. Economic Aspects 103
6. References 105

Symbols

A total electrode area
a_e specific electrode area
b Tafel slope
d electrode – electrode distance (interelectrode distance); characteristic length
D diffusion coefficient
d' electrode thickness
d_b bubble diameter
d_p particle diameter
E^0 standard electrode potential
h height of electrode, bed height, heat-transfer coefficient
j current density
$j*$ "current concentration"
j_b bed current density
j_l limiting current density
j_0 exchange current density
k, k' electrochemical rate constants
k_m mass-transfer coefficient
k_1 macroconvective mass-transfer coefficient
k_2 bubble-induced mass-transfer coefficient
L electrode length
S cross sectional area
STY space – time yield
t^+, t^- transport numbers
$U_{0,c}$ equilibrium potential difference of the cell reaction
U_{cell} cell voltage
ΔU_Ω ohmic voltage drop
u linear flow velocity
$\dot{V}$ volume flow rates
v void fraction; bed voidage
x distance from electrode; distance along electrode; bed depth
y flow length
z number of electrons
Gr Grasshof number
Nu Nusselt number
Pr Prandtl number
Re Reynolds number
Sc Schmidt number
Sh Sherwood number
Wa Wagner number
α^+, α^-, α charge-transfer coefficients
δ thickness of Nernst diffusion layer
η overpotential; dynamic viscosity
κ specific conductivity
κ/κ_L effective conductivity of a dispersion relative to the conductivity of bubble-free liquid
λ equivalent conductivity
θ fraction of electrode surface shielded by adhering bubbles
ϕ potential
ϕ^e current efficiency
Φ Thiele modulus
ν kinematic viscosity
ν_i stoichiometric number
ϱ density; resistivity
ω angular velocity

Subscripts

a	anodic
ad	adsorbed
c	cathodic
conc	concentration
ct	charge transfer
G	gas
L	liquid
m	mean, mass
ox	oxidized species
p	particle
r	rising bubble
red	reduced species
rs	rising swarm of bubbles
s	solution, electrolyte

1. Introduction

Electrochemistry of solute systems deals with the exchange of charge between an electrode and electron donors or acceptors present in the solvent and with ionic charge transfer through the solvent between two electrodes. Both phenomena, charge transfer and ionic conduction, are treated more extensively in the B Series under Electrochemical Reactors. The aim of this article is to describe the interaction of the fundamental phenomena in an electrochemical process, leading to the relatively complex macroscopic response of a process to changes in the process variables.

For example, current yields and energy consumption for the production of chlorate are determined by a multitude of factors that do not interest the theoretical electrochemist but that the electrochemical engineer must skillfully master in order to build a commercially attractive process. This process is determined simultaneously by (1) the electrode kinetics of anodic chlorine and cathodic hydrogen evolution, (2) the chosen current density and process temperature, (3) the conductivity of the electrolyte, i.e., the brine concentration, (4) the velocity of the electrolyte convection and the residence time of the electrolyte within the cell and any additional reaction vessel, and (5) the cell geometry, i.e., electrode spacing and detailed structure of the electrodes forming the electrolyzer walls.

Cell geometry and electrode structure define the cross section of the cell and affect the fluid dynamics of the streaming electrolyte. They determine the mass transfer of evolved chlorine from the anode and of hydroxyl anions from the cathode into the bulk of the electrolyte, and hence the macrokinetics of hypochlorite and chlorate formation.

2. Basic Laws and Equations

The cell voltage of any operating electrolyzer is composed of four terms: (1) equilibrium potential difference of the cell reaction $U_{0,\mathrm{c}}$, (2) anodic overpotential η_a, (3) cathodic overpotential η_c, and (4) ohmic voltage drop ΔU_Ω between cathode and anode.

$$U_\mathrm{cell} = U_{0,\mathrm{c}} + \eta_\mathrm{a} - \eta_\mathrm{c} + \Delta U_\Omega \tag{1}$$

The *equilibrium potential* of the cell reaction, $U_{0,\mathrm{c}}$, is determined by the Gibbs free enthalpy of the cell reaction and can be calculated fairly accurately by adding the equilibrium electrode potentials of the two electrodes, which are the standard potentials $E^0{}_i$ for the electrode reactions corrected for the concentrations or partial pressures of the potential-determining species according to *Nernst's equation.*

$$U_{0,\mathrm{c}} = \Delta G/zF = (E_\mathrm{a} - E_\mathrm{c})_\mathrm{eq} = \left[E_\mathrm{a}^0 + \frac{RT}{zF}\ln\left(c_\mathrm{ox}/c_\mathrm{red}\right)\right]_\mathrm{a} - \left[E_\mathrm{c}^0 + \frac{RT}{zF}\ln\left(c_\mathrm{ox}/c_\mathrm{red}\right)\right]_\mathrm{c}$$

The *anodic* and *cathodic overpotentials* can be envisaged to be composed of two terms: (1) concentration polarization η_conc as induced by concentration differences between the electrolyte at the electrode surface and the bulk of the electrolyte due to mass-transfer hindrance and (2) additional overpotential η_ct required to drive the charge transfer with the desired current density. Concentration polarization overpotentials can be calculated by the modified Nernst equation:

$$\eta_\mathrm{conc} = \frac{RT}{zF}\ln\left(c_\mathrm{ox}/c_\mathrm{red}\right)^{x=0} / \left(c_\mathrm{ox}/c_\mathrm{red}\right)^{x=\infty}$$

where x is the distance from the electrode. Here, $x = 0$ refers to the electrode surface, and $x = \infty$ refers to the bulk of the electrolyte.

2.1. Electrode Kinetics

Overpotential due to charge transfer η_{ct} may be obtained from the *Butler – Volmer equation*, which correlates current density and charge-transfer overpotentials:

$$j = j_0 \left[\exp\left(\eta_{ct} F \alpha^+ / RT\right) - \exp\left(-\eta_{ct} F \alpha^- / RT\right)\right] \quad (2)$$

The anodic and cathodic charge-transfer coefficients α^+ and α^- for simple electrode reactions, in which only one electron is exchanged between the electrode and the dissolved redox system, are close to 0.5, but may vary widely for reactions composed of a series of charge-transfer and chemical steps. For small overpotentials, the Butler – Volmer equation approaches

$$j = j_0 \left(\eta_{ct} F / RT\right) \left(\alpha^+ + \alpha^-\right) \quad \eta_{ct} \ll \left(RT / F \alpha^+\right)$$

For high overpotentials, one of the two terms usually outweighs the other, and the correlation is given by a single term describing either the anodic or the cathodic current density,

$$j = j_0 \exp\left(\eta_{ct} F \alpha^+ / RT\right) \quad 1 \gg RT / \eta_{ct} F \alpha^+ \quad (3a)$$

or

$$j = j_0 \exp\left(-\eta_{ct} F \alpha^- / RT\right) \quad -1 \gg RT / \eta_{ct} F \alpha^- \quad (3b)$$

Under these conditions the equation can be transformed into the familiar *Tafel equation:*

$$\eta = a + b \ln |j| \quad (4)$$

with a equaling ($RT/F\ \alpha$) ln j_0 and b equaling $RT/F\ \alpha$, where α is either α^+ or α^-. The overpotential for a given anodic or cathodic current density j is therefore determined simultaneously by the exchange current density j_0 and the "Tafel slope" b, and thus by the charge-transfer coefficient α.

2.2. Technical Electrocatalysis

2.2.1. Principles of Electrocatalysis

Simple charge-transfer reactions, like the one-electron oxidation or reduction of a dissolved metal ion to a solvated ion of higher or lower oxidation state, most commonly exhibit charge-transfer coefficients of ca. 0.5 and often possess activation enthalpies which are much lower than those of most chemical reactions so that the exchange current densities j_0 range from 10^{-3} up to 10 A/cm^2. Whenever the electrochemical reaction is more involved, i.e., when it is composed of at least two charge-transfer steps with at least one intermediate chemical reaction, then the reaction is likely to be kinetically hindered to a significant degree. This results in very low exchange current densities. Simple charge transfer, e.g., to dissolved partners of a redox system is only moderately sensitive to catalytic acceleration. Slow intermediate chemical reactions in multielectron electrode reactions however, that are usually hindered due to chemical kinetics, can often be accelerated very strongly by means of heterogeneous catalysis at the electrode surface.

Electrocatalysis and Formal Electrode Kinetics. Electrocatalytic action of the electrode material may have two different consequences, both leading to reduced overpotential at higher current densities:

1) The mechanism of the overall electrode reaction is kept unchanged but it becomes accelerated which is manifested in an increased exchange current density j_0 (Fig. 1B)
2) The electrocatalyst accelerates the overall rection by opening a new rection pathway. This is usually manifested by a change of the charge transfer coefficient, α, which is usually decreased, leading to reduced overpotentials at higher current densities although the exchange current densities are lower than for the uncatalyzed case (Fig. 1C) in this case

Chemical Catalysis and Electrocatalysis. It is the chemistry in electrochemical reaction mechanism, which is the most hampered, needing catalytic acceleration. Therefore electrochemical catalysis does not differ very much fundamentally and mechanistically from chemical catalysis and, apart from the fact that charge transfer rates and electrosorption equilibria depend exponentially on electrode potential. In many cases electrocatalysis and catalysis of electrochemical and chemical oxidation or reduction processes follow very similar if not the same pathways.

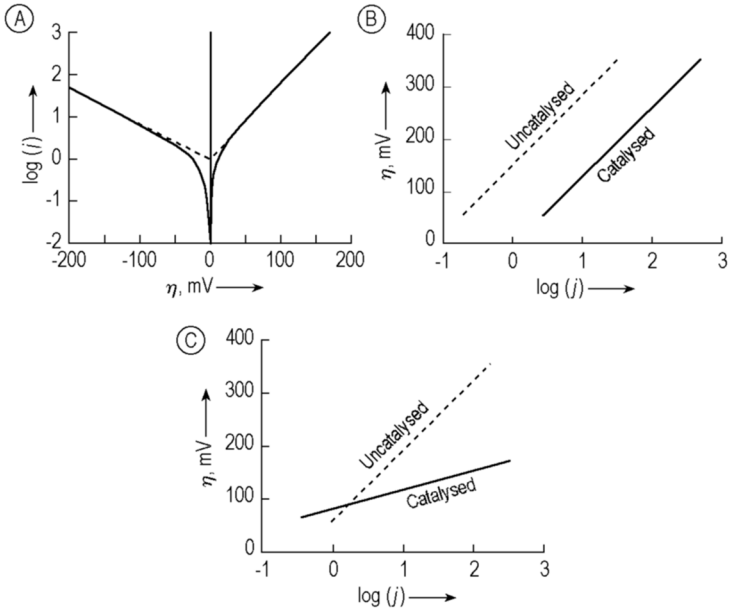

Figure 1. A) Determination of log j_0 by extrapolating the linear parts of the cathodic and anodic branches of the semilogarithmic current voltage plot. Electrocatalysis is manifested by increase of the exchange current density,j_0 or (B) by a decrease of the slope of the voltage – log current correlation (C)

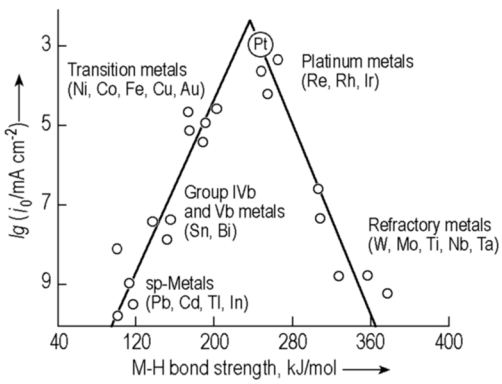

Figure 2. H_2 Volcano curve: logarithm of exchange current density of H_2 reaction vs. enthalpy of hydroge adsorption at different metals
○ Acidic

Electrochemical Hydrogen Oxidation and Generation. Since electrochemical hydrogen oxidation and generation is coupled to the chemical splitting of the H_2 molecule or its formation from adsorbed hydrogen atoms, electrocatalysts for cathodic hydrogen evolution or its oxidation and catalysts for chemical hydrogenation are essentially the same: platinum and the transition metals of group 10 (VIII) of the periodic table. For catalysis and electrocatalysis the same correlation of catalytic activity interms of log j_0 and hydrogen adsorption enthalpies at the respective metal catalyst prevails, the volcano curve (Fig. 2) [9]. The catalytic activity of those metals is optimum which adsorb hydrogen only to a moderate extent and can thus assist H_2 splitting without hindering its desorption.

$$H_2O + e^- \rightleftharpoons H_{ad} + OH^- \quad (5a)$$

$$H_{ad} + H_2O + e^- \rightleftharpoons H_2 + OH^- \quad (5b)$$

The reaction sequence Volmer reaction (5 a) and Heyrovsky reaction (5 b) is not the only one possible. Desorption of the adsorbed hydrogen may also proceed according to the Tafel reaction (5 c) by purely chemical desorptive dimerization of two adsorbed atoms for which, however, the same fundamental considerations are valid

$$2\,H_{ad} \rightarrow 2\,H_2 \quad (5c)$$

Oxygen electrochemistry, i.e., anodic oxygen evolution and cathodic reduction of dioxygen as well as anodic oxidation of organics are catalyzed by the same metals or metal oxides. The latter are also used in chemical catalysis for catalytic oxidation of organic compounds. These

catalysts/electrocatalysts undergo relatively easily redox conversion with oxygen or are oxidized anodically to a higher valency state at potentials which are close to the oxygen equilibrium potential, in this way facilitating the release of oxygen molecules or the transfer of oxygen from oxidized surface groups of the catalyst to organic substrates Eqs. (6 a) – (6 d). The relation overpotential vs. reaction enthalpy for different transition metal oxides is shown in Figure 3. Going from lower to higher oxides a type of volcano curve of anodic O_2-evolution, is obtained [10].

$$Me + H_2O \leftrightharpoons MeOH + H^+ + e^- \quad (6a)$$

$$MeOH \leftrightharpoons MeOH + H^+ + e^- \quad (6b)$$

$$2\,MeO \leftrightharpoons MeO + O_2 \quad (6c)$$

$$6d2\,MeO + RH \rightarrow 2\,Me + ROH \quad (6d)$$

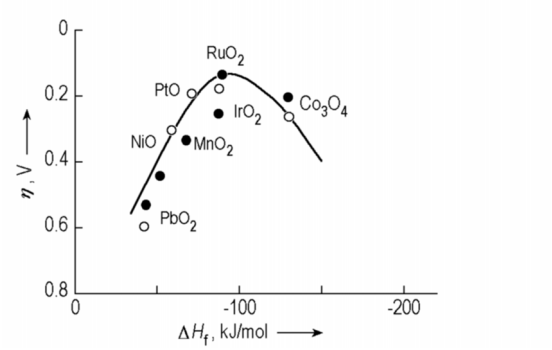

Figure 3. O_2 Volcano curve: overpotential at 0.1 mA/cm^2 vs. formation enthalpy of formation of higher from lower valent metal oxide

Utilization of Porous Electrocatalyst Particles In chemical heterogeneous catalysis of gas-phase reactions highly porous catalysts with particles of millimeter to centimeter size are used. In practical electrocatalysis, in particular in fuel cells, highly porous catalyst layers are employed as well. Accounting for the low diffusion coefficients in liquid electrolytes, 10^{-5} cm^2/s vs. 1 cm^2/s in gases, catalyst slabs which are completely flooded by the respective electrolyte have to be much smaller, i.e., of micrometer instead of centimeters size [11]; otherwise the catalyst would be utilitzed only to a very low degree.

A question of concern common to heterogeneous catalysts used in chemical catalysis and porous electrocatalytic coating is the problem of effective catalyst utilization. If the thickness of the porous catalytic coating or the diameter of the porous particle or slab does not properly match the mean diffusion length of consumed or produced species then the inner surface of the catalyst is no longer supplied sufficiently with substrate or becomes electrochemically inactive due to product accumulation and concentration polarization in the pores. In chemical engineering, the Thiele-modulus Φ_{Th} is used as a characteristic number defining the degree of utilization of a porous catalyst particle (Eq. 7 a).

$$\Phi_{\mathrm{Th}} = L\left(\frac{k}{D}\right)^{1/2} \quad (7a)$$

Where L is the pore lenght, coating thickness, or radius of the porous particle; k the volume-related rate coefficient of first order reaction; and D the effective diffusion coefficient of the reactant in the porous matrix.

For $\Phi_{\mathrm{p}} > 3$ the catalyst utilization η_{c} decreases according to a limiting $(\Phi_{\mathrm{Th}})^{-1}$ law. In a similar way the elctrochemical Thiele modulus $\Phi_{\mathrm{p,el}}$ (Eq. 7 b) is a measure for the utilization of porous electrodes in a particular fuel cell electrode of thickness L or nanoporous pellet of diameter L, which constitute a fuel cell electrode.

$$\Phi_{\mathrm{p,el}} = L\left[j\,(\eta)\,a/D\nu_{\mathrm{e}}Fc\right]^{1/2} \quad (7b)$$

(The corresponding relationship for other porous, gas evolving electrodes is mathematically more complex).

Where $j(\eta)$ denotes current density at overpotential η; a the inner electrode surface per unit volume of porous electrocatalyst; c the concentration of reactant in the electrolyte bulk; and ν_{e} the stoichiometric number of electrons for the given electrode reaction.

For a given overpotential η the the quantity $j(\eta)$ increases approximately linearly with the exchange current density j_0. For hydrogen evolving highly porous Raney nickel coatings (where j_0 equals roughly 10^{-4} A/cm^2), an overpotential of 100 mV, and a pore length $L = 1$ µm, $\Phi_{\mathrm{el,p}}$ approaches or exceeds 1 and therefore deeper pores (or thicker coatings) are no longer fully utilized. With platinum metals, which are much better catalysts than nickel ($j_0 > 10^{-2}$ A/cm^2), a dense but very thin ($d < 1$ µm) coating can be applied instead of a porous coating. The natural roughness of smooth coatings (1 µm) already matches the effective pore depht which can be fully utilized.

Functioning, Longevity, and Application of Electrocatalyst Coatings of Gas Evolving Electrodes. Electrocatalysts, composed of highly active precious metals are usually applied to the metal support (constituting the electrode proper) in the form of thin coatings, amounting to catalyst loading of milligrams per square centimeter or less. Longterm stability and firm adhesion to the support of such coatings is as important as their catalytic activity. Longevity of electrocatalyst coatings might be impaired by a number of different effects. Most important are (1) poisoning, (2) corrosion, (3) erosion, and (4) loss of internal surface by sintering and Ostwald ripening. *Poisoning* can only be reliably mitigated by using carefully purified electrolytes, raw materials, and (in fuel cells) anode and cathode gases. *Corrosion* is often a serious problem, in nonsteady condition, for instance off-times. Here redox reactions of the catalyst surface arising from floating potential render the catalyst material thermodynamically unstable or more soluble in the electrolyte than under operating potential. For instance trivalent cobalt contained in oxides (e.g., in spinels such as $NiCo_2O_4$ or Co_3O_4) is insoluble in aqueous, caustic solution, but becomes unstable at a more cathodic (i.e., more negative) potential. Cobalt(III) reacts with water under relaese of O_2 and is reduced to Co(II), which is soluble in strongly alkaline solution due to formation of $Co(OH)_4^{2-}$. *Erosion* by itself and coupled to corrosion may become a serious problem whenever electrochemical gas evolving is the main or a side reaction. Erosion is a problem if the electrode is exposed to a rapidly flowing electrolyte – in particular if the electrolyte is contaminated by dispersed solid matter. Erosion can be counteracted by modifying the structure and composition of the coating increasing its strenght and adherance and by influencing gas bubble formation in a way which removes the physical strain on the electrode due to bubble overpressure and violent two-phase flow. *Loss of internal surface* of highly porous coatings – for instance Raney nickel coatings – by slow recristallization can sometimes be prevented by internal or external dispersion hardening, i.e. by interdispersion of a second electrocatalytically inactive, oxidic highly dispersed powder which hinders creep or surface diffusion of the metallic catalyst material which would lead to agglomeration and surface loss.

The surfaces of oxidic redox catalysts which typically accelerate anodic oxygen or chlorine evolution or mediate anodic oxidation of organic substances, act catalytically as a whole not depending on active sites. They catalyze the respective reactions with every surface metal atom exposed to the electrolyte. For this type of electrocatalysts physically reinforcing the coating and strenghtening their adherance to the support is imperative. This is achieved by application of appropriate interlayers (e.g., SnO_x or TiO_y between RuO_2 and Ti-electrodes) or mixing the catalyst with inert valve metal oxides.

2.2.2. Technical Electrodes

2.2.2.1. Electrocatalytically Activated Dimensionally Stable Chlorine Evolving Electrodes

Electrocatalysis and Selectivity of Anodic Chlorine Evolution at RuO_2 Anodes. Figure 4 compares the current – voltage curves of anodic chlorine evolution measured at graphite and RuO_2 anodes in brine solution of a concentration of 300 g NaCl/dm^3, which is typical for chloralkali electrolysis. The overpotential at the RuO_2 anode is much smaller and mainly due to concentration polarization. Details of the electrocatalytic reaction are given in [11, 12, 13, 15]. The mechanism is likely to be similar to that one described [16], (Eqs. 6 a – 6 c) for anodic oxygen evolution. The same type of mechanism is also operative with anodic O_2 evolution at RuO_2 anodes. At pH 2 (the lowest pH value maintained in chlor alkali electroysis anolytes) the equilibrium potential of the oxygen anode is more cathodic than that of the chlorine electrode (+ 0.96 V vs. normal hydrogen electrode compared to + 1.26 V). Therefore, it is of ultimate importance that RuO_2 catalyzes chlorine evolution to much higher extent than oxygen evolution. However, at higher anolyte pH the oxygen potential becomes still more cathodic (60 mV/pH-unit). Thus, catalyst formulations are changed to increase the catalytic activity for the Cl_2 reaction over that of the O_2 reaction when dilute brine (30 g/dm^3 for NaOCl production) or seawater (Cl_2 production on board ships for desinfection) are electrolyzed. An increase of the Cl_2 selectivity is brought about by

adding small amounts of Pt and IrO_2 to RuO_2 or using anode coatings based on MnO_2.

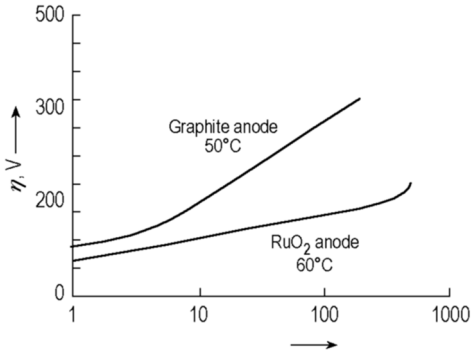

Figure 4. Semilogarithmic current density overpotential correlation of anodic chlorine evolution on graphite and RuO_2 coated titanium anodes

2.2.2.2. Oxygen Evolving Anodes

Technical Processes with Anodic Oxygen Evolution. Oxygen evolution is the counter electrode reaction in hydrogen production from alkaline solution and in cathodic electrowinning of metals from acid aqueous solution. Examples for the latter are electrowinning of zinc, copper, tin, or lead. Water electrolysis usually is performed in concentrated alkaline solutions (30 wt % KOH), whereas the electrolytes used in metal electrowinning are moderately to strongly acidic (ca.1 M H_2SO_4). Therefore, the electrode support materials and electrocatalysts are different for these two different types of electrolysis processes. In alkaline electrolysis stainless steel or nickel is used as supporting electrode material for advanced water electrolizers, where materials of improved corrosion resistance are required. For oxygen evolution from acidic electrolytes the traditional anode material – lead, passivated by in-situ formed PbO_2 – is becoming superseeded by catalyst-coated titanium anodes.

Electrocatalysis of Oxygen Evolution in Advanced Alkaline Water Electrolysis.

Coatings Containing Cobalt and Iron Oxides. In different transition metal oxides and mixed oxides (Co_3O_4, Ni_2CoO_4, etc.), which possess a relatively low solubility in alkaline solution, the redox potentials of Co(II)/Co(III)/Co(IV) oxides are relatively close to the equilibrium potential of oxygen. Because of this unique characteristic, cobalt oxides are able to catalyze anodic oxygen evolution very effectively according to Krasil'chshikov's mechanism.

Figure 5 compares the current voltage curves of anodic oxygen evolution at three different temperatures in 30 wt % KOH using RuO_2 and three different coatings containig cobalt oxide on nickel.

Although RuO_2 has the highest catalytic activity exhibiting the lowest anodic overpotential (see curves D), it is not stable in caustic solution and dissolves due to anodic formation of RuO_3 within several days. All mixed oxides containing Co show anodic overpotential close to that of the spinel phase Co_3O_4. This magnetitic oxide, however, is not stable at potentials below ca. + 0.5 V vs. the reference hydrogen electrode (RHE)

$$Co_3O_4 + 6\,OH^- \longrightarrow 3\,Co(OH)_4^{2-} + 1/2\,O_2$$

forming soluble, complexed cobalt(II).

As the dissolved cobalt is deposited at the cathode, it is completely transferred from the anode to the cathode after repeatedly cycling the electrolyzer. Cobalt oxide-coated electrodes have to be protected against dissolution in off-times by applying a protective polarization.

Figure 6 compares the current – voltage curves measured using mixed iron oxide – cobalt oxide coatings with those measured on pure Co_3O_4 and Fe_2O_3 coatings. Iron(III) oxide is a sizably better catalyst than cobalt spinel. Fe_2O_3 has, however, a relatively high solubility (of several 10^{-6} M) in aqueous concentrated KOH solution so that because of dissolution of the anode coating and cathodic iron deposition, the catalyst is readily transferred from the anode to the cathode. This renders Fe_2O_3 inherently unstable in alkaline water electrolizers. Coatings which contain both iron and cobalt are much more stable and show catalytic activities even better than that of Fe_2O_3. The longevity of this type of coating is improved to a tolerable level of some thousand hours by decreased cobalt and iron solubility.

Catalyst-Coated Titanium Electrodes for Oxygen Evolution from Acid Solutions. Catalyst-coated titanium anodes can be advantageously used for oxygen evolution from acid solutions. The anode support is passivated by a

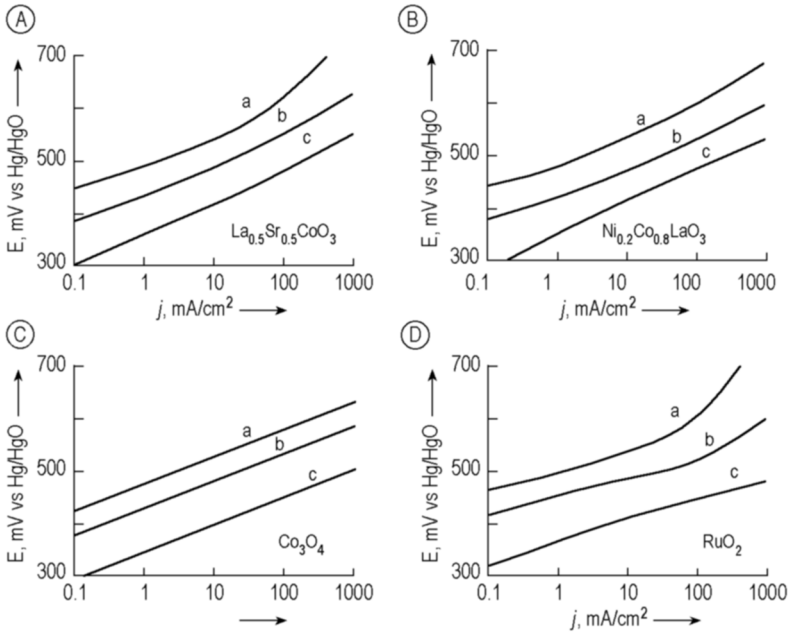

Figure 5. Semilogarithmic current potential correlation for anodic oxygen evolution from 30 wt % caustic potash at three different temperatures measured at nickel anodes with different metal oxide coatings
a) 50 °C; b) 90 °C; c) 130 °C

protective layer of TiO_2 and a suitable electrocatalyst coating reduces the oxygen overpotential substantially.

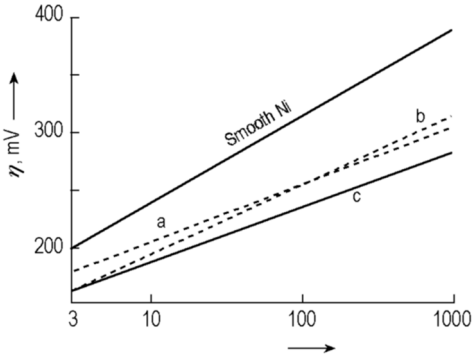

Figure 6. Semilogarithmic current voltage correlations of anodic oxigen evolution at 90 °C from 30 wt % KOH at smooth nickel anodes coated with: (a) Co_3O_4, (b) Fe_2O_3 and (c) a mixed oxide containing iron and cobalt together in roughly 1:1 molar ratio

In solutions containing approximately 20 g H_2SO_4 per liter at temperatures up to 40 °C titanium anodizes rapidly, forming a protective TiO_2 layer with time-independent thickness of fractions of micrometers. Between 50 and 70 °C continuous film growth is observed and titanium begins to dissolve actively above 70 °C. Consequently, metal-winning and metal-plating electrolyses with anodic O_2 evolution are operated at relatively low current densities allowing for relatively low process temperatures (< 40 °C) because heating by electricity passage is kept low. RuO_2 is an excellent electrocatalyst in acid solutions for anodic oxygen evolution. It is, however, oxidized and carried away in the form of relatively volatile RuO_4. Therefore, RuO_2 needs to be stabilized. Pt – Ir oxides with added TiO_2 and Ta_2O_5 – Nb_2O_5 as mechanical stabilizers are today used for anodic oxygen evolution in acid electrolytes. This type of anodes is still restricted to relative pure electrolytes, e.g., in anticorrosive tin or zinc steel-sheet plating in the automobile industry. Catalytic coatings composed of IrO_2 and Ta_2O_5 show a life time optimum at a composition of 70 wt % IrO_2/ 30 wt % Ta_2O_5. Exchanging tantalum for titanium as base metal significantly improves the stability of the combination support – coating. Such electrodes reduce the oxygen evolution

overpotential by approximately 0.5 V allowing to save more than 30 % of the energy costs which would be consumed by using conventional PbO_2 (Ag) anodes.

Water electrolysis, with hydrated cation exchange membranes (Nafion) (→ Hydrogen, Chap. 4.2.3.) cannot dispense with noble metal oxides. IrO_2 and mixtures of RuO_2/IrO_2 have succesfully been applied.

2.2.2.3. Electrocatalysis of Cathodic Hydrogen Evolution

Cathodic hydrogen evolution is the reaction of anodic chlorine evolution in chlor alkali elctrolysis. Hydrogen evolution is expected to become superseeded by the energy saving cathodic oxygen reduction only after a decennium of further development and only if energy prices will continue to increase. The chlor alkali electrolysis process is by far the most important source of electrolytically generated hydrogen, because hydrogen from water electrolysis usually cannot compete cost-effectively with hydrogen produced by steam reforming. Cathodic hydrogen evolution is also of some significance as counter electrode reaction in anodic organoelectrosyntheses.

Electrocatalyst Coating for Hydrogen Evolution from Alkaline Solution. Nowadays the following catalyst coatings are relevant for hydrogen evolution from alkaline solutions:

1) Nickel sulfide (NiS_x)
2) Raney nickel
3) Nickel alloys containing molybdenum
4) Oxidic coatings containing platinum metal oxides in particular RuO_2
5) Doped nickel oxide

Due to their relatively low costs the most frequently applied electrocatalysts are nickel sulfide and Raney nickel for alkaline conditions.

Nickel Sulfide Coatings. Two different methods are used for preparation of nickel sulfide coatings: cathodic deposition of NiS_x and chemical sulfidization of the nickel electrodes. Galvanic deposition is based on cathodic codeposition of nickel and sulfur from baths containing sulfur donors such as thiocyanate ion, thiourea, or thioglycol according to Eq. (8 a)

$$Ni^{2+} + x\,RS + 2e^- \longrightarrow NiS_x + x\,R \tag{8a}$$

The deposit is a mixture of different nickel sulfides together with amorphous nickel. The catalytic activity of the different NiS_x-phases increases with their sulfur content. Sulfidizing nickel in an atmosphere containing dihydrogen sulfide also creates smooth and close nickel sulfide layers mainly composed of NiS.

Nickel sulfide coatings are thermodynamically unstable at the hydrogen equilibrium potential. In aqueous alkaline solutions at hydrogen evolving cathodes all NiS_x-phases are reduced eventually according to Eq. (8 b) to highly disperse nickel which has properties in common with Raney nickel.

$$NiS + 2e^- \longrightarrow Ni + S^{2-} \tag{8b}$$

In so far the initial catalytic activity of the NiS_x coating is of little relevance and the real catalyst is highly dispersed nickel. Nickel sulfide coatings can only be used in nonpressurized electrolyzers working at relatively moderate temperatures (80 °C) because the cathodically released sulfide ions are aniodically oxidized to sulfate anions, which cause stress corrosion cracking with the steel parts of the electrolyzer.

Raney Nickel Coatings.

Precursor Alloys and Fabrication of Coated Cathodes. Raney nickel was introduced more than 30 years ago as an anodic electrocatalyst for alkaline fuel cells by Winsel and Justi [17]. Since more than 20 years it is used as catalytic coating for hydrogen evolving cathodes in chlor alkali electrolysis and alkaline water electrolysis. The operation conditions of the cathode are almost the same in both technologies as the catholyte contains 20 to 30 wt % NaOH or KOH and the process temperature is between 70 and 90 °C.

Although at least four different technologies for production of Raney nickel coated cathodes (cold rolling, flame spraying, Zn and Al-melt dipping, cathodic deposition of Ni/Zn-precursor alloys) have been described, only cold rolling and cathodic deposition of precursor alloys are used for commercial production.

Utilization of the Catalyst in Raney Nickel Coatings. Compared to platinum metals (e.g., Pt or Ru) which exhibit exchange current densities of the hydrogen evolution reaction in caustic electrolytes at 30 °C of the order of 10^{-3} to 10^{-2} A/cm^2, nickel is only a moderately active electrocatalyst with j_0 ranging from 5 to 7×10^{-6} A/cm^2 at ambient temperature. Therefore, considering a ratio of the exchange current densities of Pt and Ni of 10^2 to 10^3 a naive estimate would predict comparable activities of smooth Pt-metal cathodes and Raney nickel cathodes, provided the catalytically active inner surface of the Raney nickel coating would amount to 100 to 1000 cm^2 per square centimeter electrode surface. In practice, however, the internal surface of nanoporous Raney nickel catalyst coatings is usually utilized to no more than some percent. This low degree of catalyst utilization is due to slow diffusive mass transport of electrogenerated dissolved hydrogen out of the nanopores of the catalyst. The highest hydrogen concentration level in the pores is the concentration determined by the applied cathodic overpotential η according to Nernst's law (Eq. 9)

$$c_{\lim} = c_0 \exp\left\{\frac{-F\eta}{2RT}\right\} \tag{9}$$

with c_0 equaling the H_2 saturation at 0.1 MPa H_2 pressure.

Since the concentration of dissolved hydrogen in the pore increases with increasing pore depth the effective local cathodic overpotential diminishes due to concentration polarization according to Eq. 10, the overpotential η_x being always negative at H_2 evolving cathodes.

$$\eta_x = \eta_0 + \frac{RT}{2F} \ln \frac{c_x}{c_0} \tag{10}$$

The current density decreases to nil as the hydrogen concentration in the pore approaches its limiting value.

Figure 7 depicts the calculated current density, overpotential, and hydrogen concentration distribution in an assumed ideal (straight cylindrial) Raney nickel pore of a pore diameter of 2 nm and assumed virtual infinite depth calculated with an exchange current density of 10^{-4} A/cm^2, a value which is assumed to be valid for nickel at 80 to 90 °C. As can be derived from the current density distribution, the contribution of the pore walls to cathodic hydrogen evolution becomes negligible beyond 10 µm pore depth (Fig. 7D), so that a perfect, undivided Raney nickel coating of 100 µm thickness, can be utilized to only 7 to 8 %. This is the reason, why the fissures and cracks (tertiary structure of the catalyst) formed in Raney nickel coatings during the leaching process are important for improving its utilization. The utilization of gas evolving nanoporous catalysts increases with current density.

Performance and Aging of Raney Nickel Coatings. Raney nickel coatings obtained by relatively slow reduction of nickel sulfide during cell operation, which lasts for several hundred hours, shows almost no deterioration over more than 3000 h. In contrast, Raney nickel coatings formed by relatively fast caustic leaching of Ni – Zn precursor coatings deteriorate sizeable during long-term operation. This difference is very likely due to the more open and perturbed structure of the quickly formed catalyst, which causes restructuring and recrystallization accompanied by surface losses with time. Recrystallization may be retarded by dispersing TiO_2 or other metal oxides throughout the porous structure. This measure can only be used for cold-rolled or plasma-sprayed, not for cathodically deposited precursor coatings.

Coatings of Platinum Metal Oxides. Dimensionally stable RuO_2 anodes have been succesfully used also for cathodic hydrogen evolution. These electrodes are characterized by small Tafel slopes, low overpotential, and in particular insensitivity to iron poisoning. Ruthenium oxide is thermodynamically unstable and is reduced to ruthenium metal at hydrogen equilibrium potential. The metal becomes the catalyst proper although the underlying metal oxide layer keeps unchanged. It is the insensitivity of ruthenium metal to iron poisoning and the metallic reservoir stored in the bulk of ruthenium oxide which makes DSAs uniquely effective hydrogen evolving cathodes.

Active Coatings of Flame-Sprayed, Doped Nickel Oxide. Asahi Glass reports on the development and commercial use of electrodes covered by catalytically active nickel oxide coatings which are said to be fabricated by flame spraying. No detailed information of the composition, morphology and phase-content of the

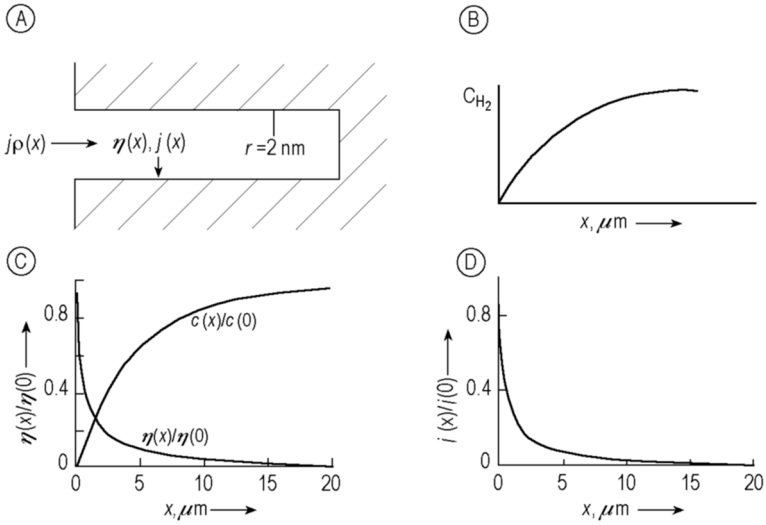

Figure 7. A) Schematic of straight nanopore; B) Concentration increase of dissolved hydrogen in the depth of a cathodically active nanopore; C) H_2 concentration and effective overpotential; D) Current density distribution in the nanopore

coating are communicated. The electrode coating is said to deteriorate slowly by progressive reduction to metallic nickel. Life times of three years in chlor alkali electrolysis membrane cells are reported by the manufacturer.

Platinum and Platinum Metal Cathodes in Membrane Water Electrolyzers. In acidic electrolytes only noble metals, and in particular the platinum metals, are resitent under unpolarized conditions and can be used as electrode materials for hydrogen evolution. Platinum in particular is a unique electrocatalyst for H_2 evolution in strongly acidic environment as prevailing in perfluorinated sulfonated polymer membranes used as solid polymer electrolyte in membrane electrolyzers (see Section 4.2).

2.2.2.4. Electrocatalysis of Cathodic Oxygen Reduction and Anodic Hydrogen Oxidation in Fuel Cells (→ Fuel Cells)

Anodic hydrogen oxidation and cathodic oxygen reduction is kinetically hampered at low temperature, so that anodic hydrogen oxidation in alkaline fuel cells (AFCs), proton exchange membrane fuel cells (PEMFCs) and phosphoric acid fuel cells (PAFCs) demands catalysts of highest activity, that means platinum metals and platinum in particular. Working temperatures of molten carbonate fuel cells (MCFCs) and solid oxide fuel cells (SOFCs), however, are high enough that the cheaper nickel metal – a much poorer catalyst than Pt – can be used in the form of relatively coarse, sintered electrodes. Platinum is indispensible as catalyst, also for cathodic oxygen reduction in low temperature fuel cells, whereas the cathodic electrocatalysts in MCFCs and SOFCs are lithiated nickel oxide and lanthanum – manganese perovskite, respectively – which are relatively cheap *p*-type semiconductors.

Structural Design of Gas Diffusion Electrodes in Low Temperature Fuel Cells. Typically, the electrodes of low temperature fuel cells contain a bimodal pore structure, comprising nanopores in the catalyst particles and micropores between the particles. They are usually constituted of PTFE bonded nanoporous catalyst particles, usually Pt doped active carbon agglomerates (Fig. 8). The nanopores are flooded by the electrolyte, the micropores are free of electrolyte and conduct the working gases to the surface of the nanoporous catalyst particles, that are covered by a thin electrolyte film. The electrolyte in the flooded catalyst particles together with the electrolyte film on their surface form a coherent electrolyte phase. The electrolyte es-

tablishes a good connection between the electrode, the electrolyte filled interelectrodic gap, and the counter electrode on the one hand and the gas-filled pores connect the interior of the electrode to the gas lumen on the other hand.

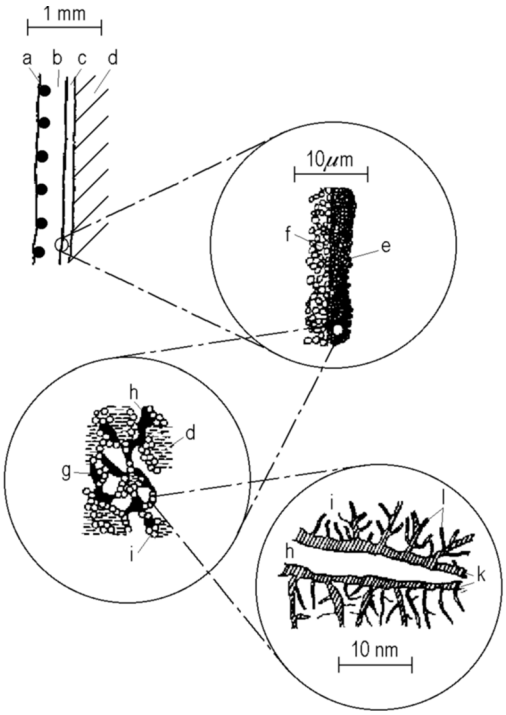

Figure 8. Morphological features of a gas diffusion electrode with PTFE bonded active carbon particles (Pt doped) for anodic H_2 oxidation and cathodic O_2 reduction a) Support; b) Gas diffusion layer; c) Hydrophobic working layer; d) Electrolyte; e) Magnified working layer; f) Adjacent non-flooded gas diffusion layer; g) Nonporous active carbon particles coated with platinum catalyst; h) Gas channels; i) PTFE particles; l) Flooded nanopores in active carbon matrix and electrolyte film on such particles; k) Electrolyte film on nanoporous particles and walls of coarse particles

Kinetic Aspects of Cathodic Oxygen Reduction at Low Temperature (T < 250 °C).

Stepwise Cathodic Reduction of Oxygen. Anodic oxygen evolution and the cathodic oxygen reduction are expressedly hindered at low temperatures so that significant overpotentials must be applied to perform the respective reaction with measurable velocities and current densities. The consequence is that microscopic reversibility is not established and oxygen reduction and anodic oxygen evolution proceed to different reaction routes.

It is almost impossible to establish the equilibrium potential of the oxygen electrodes in aqueous electrolytes (+ 1.23 V vs. reference hydrogen electrode, RHE). Usually at zero current density a potential of ca. 0.95 V vs. RHE is obtained which is established by mixed reactions, with the main reactions being the reduction of oxygen to hydroperoxide.

In *acidic media* protonation of oxygen followed by one-electron reduction is the rate limliting CE step for oxygen reduction.

C	$O_2 + H^+ \leftrightharpoons O_2H^+$	(11 a)
E	$O_2H^+ + e^- \rightarrow O_2H$	(11 b)

Where C denotes a chemical and E an electrochemical step with charge transfer.

Electrocatalyzed oxygen reduction proceeds always in the adsorbed state.

Almost in any case of electrocatalyzed oxygen reduction hydrogen peroxide is observed as a byproduct if it is not the main product – as for instance in the case of oxygen reduction in alkaline solution on graphite cathodes.

Here the two-electron ECE sequence

$(O_2H^+)_{ad} + e^- \rightarrow (O_2H)_{ad}$	Electron transfer	(12 a)
$(O_2H)_{ad} + H^+ \rightarrow (H_2O_2^+)_{ad}$	Chemical reaction	(12 b)
$(H_2O_2^+)_{ad} + e^- \rightarrow (H_2O_2)_{ad}$	Electron transfer	(12 c)
$(H_2O_2)_{ad} \rightarrow (H_2O_2)_{dissolved}$	Desorption	(12 d)

is so fast, that it cannot be resolved by usual electrochemicals methods available for electrode kinetic investigation, which have a time resolution of not better than 1 ms. The four-electron reduction of oxygen may be due to fast catalytic disproportionation of hydrogen peroxide, Eq. (13), and subsequent two-electron reduction of oxygen. This chemical diproportionation reaction is catalyzed by noble metals including silver and in particular platinum metals (stable in alkaline and acidic solution), and by a number of transition metal oxides, e.g., Fe_3O_4, MnO_2, Co_3O_4, and perovskites and magnetites containing NiO, FeO, MnO_x which are, however, only stable in alkaline solution

$$2\,H_2O_2 \rightarrow 2\,H_2O + O_2 \qquad (13)$$

Although $2e^-$ reduction with fast, chemically catalyzed H_2O_2-disproportionation seems to be the mechanism of highest importance, hydrogen peroxide can also be electrocatalytically reduced to water according to Eq. (14)

$$H_2O_2 + 2e^- \rightarrow 2\,OH^- \quad (14)$$

in a second, fast ECEC sequence, similar to (12 a) – (12 d) but accompanied by splitting of the O–O bond.

Metal Catalysts for Cathodic Oxygen Reduction. Electrocatalysis of anodic evolution and of cathodic reduction of oxygen proceeds in most cases of practical importance not on metallic, but on oxidized surfaces. Platinum like all other "noble" metals becomes already oxidized at potentials which are 0.2 mV more cathodic than the oxygen potential which is + 1.2 V vs. RHE. Other platinum metals, Rh or Ru and silver are even less noble. These oxide layers are the electrocatalysts proper for oxygen reduction. Most of the platinum metals, e.g., Pt, Rh, Ru, and Ir are corrosion resistant and do not dissolve in acidic solution (Pt not even in 100 % phosphoric acid at 200 °C). Silver is also a good electrocatalyst, which may, however, only be used in alkaline solution. Even there, silver cathodes and silver containing catalysts have to be protected from oxidation of the metal by keeping them permanently (in particular in off-time of the fuel cell) at a cathodic polarization of at least − 200 mV vs. O_2-equilibrium potential.

Catalysts for Anodic Hydrogen Oxidation. The mechanism of anodic hydrogen oxidation is much simpler than that of oxygen reduction. Reaction pathways would be revers of the Volmer – Heyrovsky or the Volmer – Tafel mechanism, i.e., with anodic adsorption (15 a)

$$H_2 \rightarrow H_{ad} + H^+ + e^- \quad (15a)$$

or dissociative chemisorption (15 b)

$$H_2 \leftrightharpoons 2\,H_{ad} \quad (15b)$$

as the initial steps.

Since the electrocatalysts must be stable under open-circuit conditions, the material of choice for alkaline fuel cells are Raney nickel [17], or carbon-supported Pt, Pd, and other platinum metals, and carbon-supported platinum, palladium, or alloys of platinum metals for fuel cells using acid electrolytes. Also tungsten carbide (WC) and platinized high-surface WC have been used succesfully in acid solution.

Properties, Preparation, and Improvement of Electrocatalysts in Gas Diffusion Electrodes.

Nanodisperse Platinum on Soot and Active Carbon. Because of their high internal surface, amounting typically to 100 m^2/g, soot and other active carbons are the most important catalyst supports for fuel cell electrodes. Platinum supported by active carbon is the fuel cell catalyst of choice for the cathode as well as for the anode.

Three different ways exist to dope the carbonaceous support material with platinum or other catalytically active noble metals.

1) Flooding of the active carbon by aqueous or organic solvents containing Pt salts with the help of wetting agents, drying of the flooded material and reduction of the deposited salts at low temperatures by hydrogen
2) Creating acidic groups (carboxylic acid and phenol groups) on the active surface by chemical oxidation with, e.g., chromic acid, fixation of cationic complexes of the catalyst (e.g., amino complexes of Pt(II) or Pt(IV)) at these groups by ion exchange, followed by chemical or electrochemical reduction of these complexes
3) Preparation of stabilized highly dispersed colloidal platinum in aqueous or nonaqueous solution by chemical reduction followed by adsorptive precipitation of the dispersed Pt particles on the surface of active carbon particles.

Ultimate goal of the catalyst preparation is to obtain very small platinum particles in order to increase the surface to volume ratio. Platinization by ion exchange and impregnation with colloidal platinum yield the best results in this respect [18].

Platinum Alloy Catalysts. By reaction of dispersed Pt on active carbon with metals of group 4 (IV B) and 5 (V B) dispersed alloys are formed which are very active electrocatalysts. Highly dispersed Pt – Cr or Pt – V alloys are formed by treating carbon impregnated with platinum and chromium or vanadium salts at 900 °C in an argon atmosphere.

Even after 9000 h of operation at 200 °C in 103 % phosphoric acid (H_3PO_4 polymer + P_2O_5) between 30 and 80 % of the nonnoble metal still can be found undissolved in the catalyst. This type of stabilized alloy catalyst is

used today in commercial fuel cell electrodes of phosphoric acid cells and contributes significantly to the technical success of this technology.

Due to catalyst aging phosphoric acid cells lose 18 % of their initial activity within 40 000 h. Two *aging mechanisms* for platinum-activated O_2 cathodes are discussed:

1) Anodic dissolution of Pt in the form of Pt(II) in off-times and diffusion of the dissolved species to the anode where Pt is deposited as metal or followed by reprecipitation of dissolved Pt in the cathode, which occurs if the cathode potential becomes more negative, results in coarsening of platinum particles.
2) Agglomeration of particles by stochastic movement on the soot particle of the supporting active carbon.

Anodic oxidation followed by dissolution of Pt and either reprecipitation or transfer to the cathode is the most important cause for deterioration. This aging mechanism is potential-dependent and accelerates as the cathode potential increases, for instance under partial load or in off-time, when the cathode potential drifts towards the oxygen equilibrium potential. Therefore, it is important that, whenever the cell is switched off, the oxygen in the cathode lumen is rapidly exchanged by inert nitrogen, and that the cell voltage under operation does not surmount 0.8 V.

Agglomeration of Pt crystallites due to Brownian motion can also be observed but is less effective. The interaction between the Pt particles and the supporting soot in presence of the electrolyte, phosphoric acid, is weak enough to allow for relatively free movement of the Pt particles. Alloy catalysts composed of platinum and non-noble metals show a reduced tendency to dissolution and agglomeration, thus their deterioration and activity loss is much slower than that of the pure platinum catalyst.

Silver Cathodes for Oxygen Reduction in Alkaline Electrolytes. Silver is one of the best electrocatalysts for cathodic oxygen reduction in alkaline solution. It is, however, readily oxidized to Ag_2O already at potentials of -200 mV vs. the oxygen equilibrium potential. In order to avoid anodic oxidation of silver followed by dissolution of the metal it is therefore essential, that in off-time the silver electrode is protected against anodic oxidation be cathodic polarization. Silver has been used successfully in alkaline fuel cells.

The silver is used in three different forms, namely as

1) Silver-doped carbon catalyst
2) Raney silver
3) PTFE-particles coated by a thin silver film in the form of Silflon cathodes

Whereas the literature on silver-doped carbon is relatively scarce more is known about Raney silver.

Raney silver – as is typical for Raney metals made from metals of relatively low melting point ($mp = 1000\,°C$) – is a much coarser material than Raney nickel ($mp = 1650\,°C$). It exhibits a specific surface of $\leq 1\,m^2/g$. The same holds for Silflon cathodes, which with silver loadings of $0.1\,g/cm^2$ of electrode area exhibit an active catalyst surface of no more than $\leq 0.24\,m^2/cm^2$ ($2.4\,m^2/g$). These Silflon electrodes are produced by chemically depositing silver on PTFE-particles of 0.5 to 1 µm size and working these silver-covered tiny particles into electrodes by calendaring or pressing. Figure 9 compares the current voltage curve of typical gas diffusion cathodes in an alkaline electrolyte. Evidently the Silflon electrode despite its relatively low active surface is superior even to the PTFE bonded Pt-activated active carbon electrode.

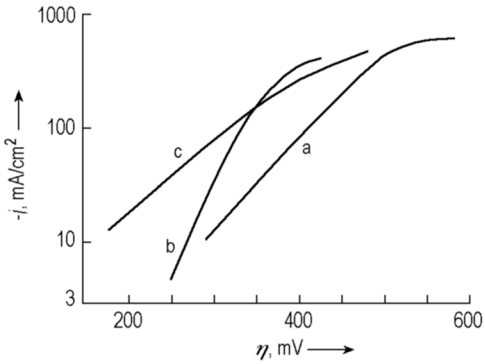

Figure 9. O_2 reduction current voltage curve in alkaline electrolyte at different cathodes
a) PTFE bonded active carbon, b) Pt-catalyzed active carbon, c) Silflon cathode

Particle Size and Utilization of Raney Nickel in Hydrogen Anodes. The particle sizes of Pt-activated soot (0.1 µm) and Raney nickel electrodes (several micrometers) differ by more than

an order of magnitude. It is the lower intrinsic catalytic activity of nickel compared to platinum in the hydrogen oxidation kinetics (j_0 (Ni) 10^{-6} A/cm^2 compared to j_0 (Pt) 10^{-3} A/cm^2) which permits to choose larger particles for gas diffusion anodes from Raney nickel than from Pt on active carbon.

Figure 10 depicts schematically the concentration profile of dissolved hydrogen in a porous catalyst particle brought about by competition of diffusive mass transport of hydrogen into the porous catalyst particles and the consumption of hydrogen by anodic oxidation at the inner surface of the catalyst.

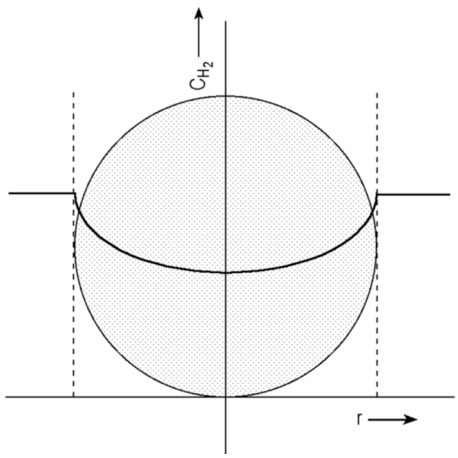

Figure 10. Concentration distribution of dissolved hydrogen in a nanoporous anodic electrocatalyst particle. Depletion of H_2 in the particle results in decreased catalyst utilization

The utilization of the porous catalyst particle is given by the ratio of the actual current produced by anodic H_2 oxidation in the particle to the hypothetical current which would be produced in a spherical particle without mass transfer limitations

$$\eta = \text{utilization} = I/I_{\max} = 3\tanh\Phi_{\text{Th,el}}/\Phi_{\text{Th,el}} \quad (16)$$

According to the different exchange current densities j_0 for hydrogen oxidation and hydrogen evolution on Ni and Pt, the catalytic activity of platinum is by a factor of several hundred to thousand higher than that of nickel. Therefore, if the utilization of Raney nickel particles below 10 µm approaches 100 %, Pt-activated active carbon particles have to be smaller by a factor of 10 (square root of 100) to 30 (square root of 1000) than Raney nickel particles to achieve full utilization.

2.3. Ionic Conduction in Electrolytes

The ohmic potential drop ΔU_Ω (Eq. 17) is determined by the current density j, the electrode – electrode distance d, and the specific conductivity κ of the electrolyte solution:

$$\Delta U_\Omega = j d \kappa^{-1} \quad (17)$$

The specific conductivity κ of dilute electrolytes is the sum of the products of the equivalent conductivities λ_i, the charges n_i, and the concentrations c_i of the various ions that constitute the electrolyte solution.

$$\kappa = \sum_{i=1}^{m} c_i \lambda_i n_i$$

For electrolyte concentrations above ca. 10 mmol/dm^3, the simple linear relationship no longer holds because the λ_i become functions of concentration, typically diminishing with concentration. In general, electrical conductivities of aqueous electrolyte solutions increase with temperature because the viscosity, which determines ionic mobilities (Walde's rule, $\lambda_i \cdot \eta_i \approx$ constant),

decreases as temperature increases. But this rule fails dramatically for concentrated bases at temperatures exceeding 150 °C. This is demonstrated in Figure 11 for aqueous solutions of caustic potash of relatively high concentrations [19]. All curves for all concentrations show a maximum in specific conductivity at a temperature of ca. 150 °C although the viscosities of these solutions keep decreasing above this temperature. The reason for this phenomenon is likely a drastic change in the water structure. The progressive destruction of water clusters begins at ca. 150 °C.

The physical properties and conductivities of concentrated aqueous electrolyte solutions and of salt melts are quite comparable, as demonstrated in Table 1, so that the electrochemical engineer needs not take into account significantly altered electrolyte properties in switching from water to ionic melts.

Table 1. Specific and equivalent conductivities of sodium chloride in aqueous solution and in NaCl melts

Electrolyte	Temperature, °C	Conductivity, specific, $\Omega^{-1}\,cm^{-1}$	Conductivity, equivalent, $\Omega^{-1}\,cm^2\,mol^{-1}$
5.09 mol/dm^3			
NaCl (aq.)	18	2.13	42
NaCl (l)	802	3.58	117

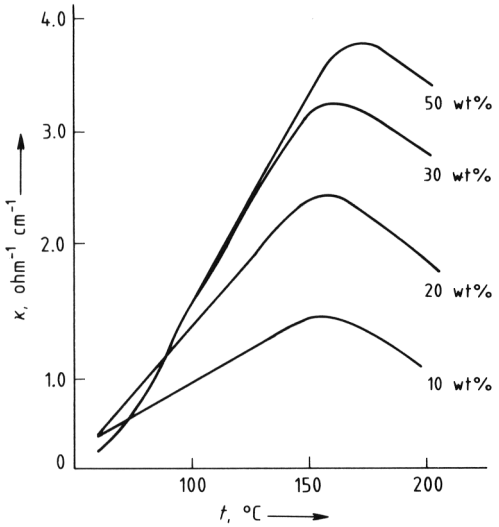

Figure 11. The specific conductivity of aqueous potassium hydroxide solutions as functions of temperature [19]

2.4. Fluid Dynamics and Mass Transfer

In many liquids with viscosities comparable to that of water, most nonpolymeric species have diffusion coefficients of the order of $10^{-5}\,cm^2\,s^{-1}$. Therefore, convective diffusion is a relatively inefficient mass-transfer process in electrolytes, even under conditions of strongly forced convection. Figure 12 depicts the dependence of the mass-transfer coefficient k_m on the Reynolds number or equivalently, the flow velocity [20]. Even at Reynolds numbers of 10 000, in 1-cm-wide flow channels corresponding to flow velocities of 1 m/s, the mass-transfer coefficient does not exceed 10^{-2} cm/s.

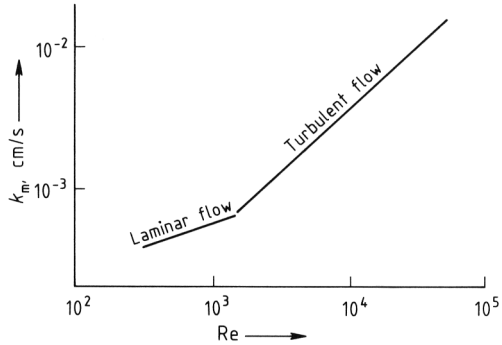

Figure 12. Dependence of mass-transfer coefficient k_m on Reynolds number for the flow of aqueous solutions in a $1\times1\,cm^2$ channel [20]

The current density and the limiting current density are proportional to the mass-transfer coefficient and the driving concentration difference.

$$j = k_m z F \left(c^{x=\infty} - c^{x=0}\right)$$
$$j_l = k_m z F c^{x=\infty} \tag{18}$$

For a one-electron charge transfer and one molar solution, limiting current densities j_l of no more than 1 A/cm^2 may be obtained with $k_m = 10^{-2}$ cm/s, corresponding to a surface specific conversion rate of only $10^{-5}\,mol\,cm^{-2}s^{-1}$. This conversion rate (and the space – time yield) is low and therefore electrochemical reactors are relatively expensive. All available and reasonable technical means must be used in order to enhance mass transfer. Forced convection, gas sparging, and gas evolution at the counterelectrode in an undivided cell are effective methods to enhance mass transfer. Another is the use of so-called turbulence promoters in narrow-gap cells (see Section 3.5).

The dependence of mass-transfer coefficients on fluid dynamic conditions is best presented in dimensionless form, which is independent of the geometry of the electrolyzer and the actual properties of the solvent – solute system, especially the solvent viscosity and solute diffusivity. Therefore, rather than correlate the mass-transfer coefficient k_m with the linear flow velocity (forced convection) or with the Grashof number *Gr* (free convection), the Sherwood number *Sh* is correlated with the Reynolds number *Re*.

$Sh = k_m\, d/D$ — d = electrode – electrode distance; D = diffusion coefficient

$Re = u\, d/\nu$ — u = linear flow velocity; ν = kinematic viscosity

If expressed in terms of the mass-transfer-limited current density,

$$Sh = \frac{j_l}{z F c^{x=\infty}} \frac{d}{D}$$

such correlations must include the Schmidt number, $Sc = \nu/D$, which accounts for the viscosity and diffusivity of a particular solvent – solute system.

Many industrial electrolyzers work under forced convection or under gas stirring. Free convection is important in metal refining and metal electrowinning. Table 2 summarizes important dimensionless correlations involving mass transport for free and forced convection for various types of electrode arrangements [21]. The most important for electrochemical engineering are certainly the correlations for the parallel plate cell with forced convection under laminar and turbulent conditions.

3. Electrochemical Cells

The parallel plate cell, undivided or divided by a diaphragm or a membrane, is the classical cell geometry for industrial electrolyzers. Such a parallel plate cell may constitute an entity on its own, and is operated in the *monopolar mode.* Several parallel plate cells may be put in series into a filter-press-like setup and are operated in the *bipolar mode* – each electrode operating as an anode on one side and as a cathode on the other side. Apart from the current-density distributions, the monopolar – bipolar options do not differ with respect to fluid dynamics and other operational characteristics and there is, in principle, no preference for either of these modes.

Construction of the bipolar electrolyzers is generally simpler than for monopolar electrolyzers, but material and corrosion problems are generally easier to solve in monopolar electrolyzers because stack shunting and stray currents are not encountered. Anode and cathode often are made of different metals (e.g., chlor-alkali cells have titanium anodes and steel or nickel cathodes), and generally different materials are not easily combined into one bipolar plate. Moreover, in bipolar electrolyzers all electrolysis cells are connected to each other by the electrolyte manifold, which supplies the cells with the circulated electrolyte. This electrolytic connection acts as an electrical shunt, which induces uneven current distributions along the stack. The highest current densities occur in the cells at the ends of the stack. The consequence is that effective current efficiencies decrease and energy consumption increases as the number of cells in the stack increases. If the current is interrupted, the cells in series act as galvanic cells, which can lead to reverse electrolysis, at least in some cells, possibly severely corroding and damaging these reversed cells.

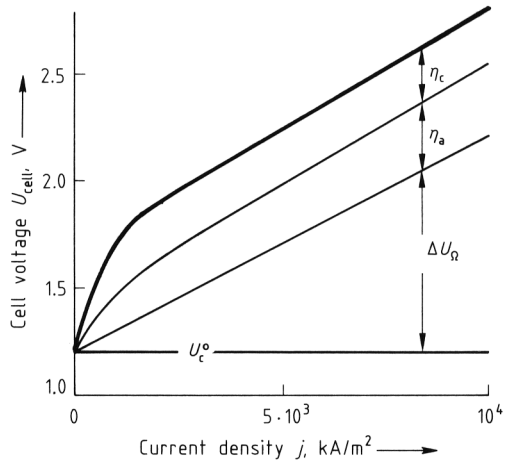

Figure 13. Schematic presentation of cell voltage (heavy line) versus current density
Cell voltage consists of decomposition voltage $U_{0,c}$, ohmic potential drop ΔU_Ω, anodic overpotential η_a, and cathodic overpotential η_c.

Because of these difficulties, bipolar electrolyzers have been developed up to commercial maturity only in a limited number of cases. In these cases the anode and the cathode material have usually been the same, and corrosion and destruction due to current reversal is of little importance due to insensitivity of the electrodes to reversed operation. These cases include alkaline water electrolysis with bipolar steel or nickel electrodes and electrolysis of hydrochloric acid with bipolar carbon electrodes. The construction problems have also been overcome for bipolar chlor-alkali electrolyzers.

Table 2. Dimensionless correlations involving mass transfer [21]

System	Correlation ∗	Characteristic length and velocity used for definition of dimensionless quantities	Validity limits ∗
Forced convection along parallel plate cell	1) $Sh = 2.54\ (ReSc)^{1/3}\ (d/L)^{1/3}$	electrode gap and length	1) $Re \leq 3000$, laminar flow
	2) $Sh = 0.025 Re^{0.8} Sc^{1/3}$		2) $Re > 3000$, turbulent flow
Free convection along vertical plate, laminar flow	$Sh = 0.66\ (ScGr)^{1/4}$	height of electrode	$500 < Sc < 80000$ $10^4 < Gr < 10^9$ $5 \times 10^6 < ScGr < 5 \times 10^{12}$
turbulent flow	$Sh = 0.31\ (ScGr)^{0.28}$	height of electrode	$4 \times 10^{13} < ScGr < 10^{15}$
Free convection across horizontal plate	$Sh = 0.19\ (ScGr)^{1/3}$	length of electrode	$10^8 < ScGr < 1.4 \times 10^{12}$ $2.1 \times 10^3 < Sc < 5.2 \times 10^4$
Free convection between vertical electrode and diaphragm	$Sh = 0.0225\ (ScGr)^{0.85}\ (e/h)^2$	electrode – diaphragm distance e	$2 \times 10^6 < ScGr < 2 \times 10^8$ $0.005 < e/h < 5.48\ (ScGr)^{-0.3}$
Flow along plate	$Sh = 0.66 Re^{1/2} Sc^{1/3}$	height of electrode in flow direction, linear flow velocity	$Re < 3 \times 10^5 - 10^6$
Axial flow through gap between two coaxial cylinders,		equivalent diameter, flow velocity	$300 < Sc < 3000$
laminar	$Sh = 1.62\ (ReSc)^{1/3} \left(\frac{d_e}{L}\right)^{1/3}$		$300 < Re < 2100$
turbulent	$Sh = 0.023 Re^{0.8} Sc^{1/3}$		$2100 < Re < 30000$
Rotating cylinder	$Sh = 0.079 Sc^{0.356} Re^{0.7}$	diameter of cylinder, circumferential velocity	$1000 < Re < 100000$ $835 < Sc < 11490$
Rotating cylinder with coaxial stationary hollow cylinder	$Sh = 0.079 Re^{0.7} Sc^{0.356}\ (d_2/d_1)^{0.7}$	diameter of counter-electrode, circumferential velocity	$1000 < Re < 53000$ $0.22 < d_2/d_1 < 0.68$

∗ Sh = Sherwood number = $j_l\ L/Dc^{x=\infty}$; Gr = Grashof number = $g\ L^3\ (\varrho_0 - \varrho_e)/\varrho_0\ \nu^2$; Re = Reynolds number = ud/ν; Sc = Schmidt number = ν/D ;
d = electrode gap; d_1 = diameter of stationary working electrode; d_2 = diameter of rotating counterelectrode; d_e = equivalent diameter = $(d_2 - d_1)$; e = diaphragm – electrode distance; h = height of electrode; L = electrode length

The engineering of single cells – irrespective of their incorporation into a bipolar stack or their monopolar operation – must incorporate a constructional design and mode of fluid dynamic operation that takes into account several phenomena, all of which might interfere with cell operation by increasing overpotentials and ohmic potential drops.

The most important of these phenomena are (1) uneven current distribution, which partially increases local current densities above mean current density, thus increasing the overpotential, (2) increase of the specific resistivity of the electrolyte by incorporation of dispersed gas bubbles, which can even partially block the electrodes by formation of gas blankets, and (3) excessive electrode distance or an inappropriate diaphragm or membrane with excessive electrical resistance. Figure 13 underlines the importance of these factors by presentation of the cell voltage of a water electrolyzer as a function of current density. For a low current density, increases in cell voltage are caused mainly by the increase of overpotential. For a high current density, it is to a large extent the internal resistance, ΔU_Ω, that determines the further increase of cell voltage with current density [22].

3.1. Optimized Current Density

The type of cell voltage – cell current density correlation in Figure 13 permits the use of a linear approximation for U_c vs. j in the domain of

higher current densities, which are the ones of commercial concern

$$U_{\text{cell}} = U_c^* + jR^*$$

where $U*_c$ is the apparent decomposition potential and $R*$ is the effective surface specific resistance of the cell ($\Omega\,\text{cm}^2$, also called the k-value).

The cell voltage, and hence energy consumption per unit mass of product, increases linearly with current density. Since for a fixed production per unit time the electrode area (or number of electrolyzers) varies inversely with current density, the investment costs are proportional to j^{-1}. Total specific product costs equal energy costs plus investment costs, in symbols:

$$C = Aj + B/\text{j}$$

C = total costs per unit mass of product
A = energy cost coefficient, mainly determined by effective cell resistance $R*$
B = investment cost coefficent

As illustrated in Figure 14, according to this approximation there is an optimal current density, which determines the lowest production costs and which depends on energy price.

$$j_{\text{optimal}} = \left(\frac{B}{A}\right)^{1/2}$$

High energy prices call for low current densities, whereas low energy prices allow the process to be designed to run with high current densities [22].

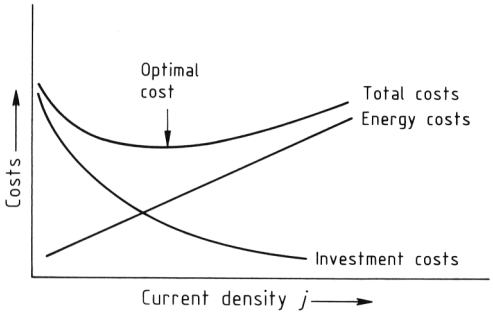

Figure 14. Total specific product costs versus current density (schematic)

3.2. Primary, Secondary, and Tertiary Current Distributions

The current distribution between any two parallel electrode surfaces is only uniform in cases where a nonconducting containment of the same cross section surrounds the interelectrode space. In all other cases, the current density is unevenly distributed, leading to higher current densities at the edge of the electrodes. If the electrode is not smooth, current densities are also enhanced on whiskers, steps, or studs that protrude from the electrode into the electrolyte toward the other electrode. Neglecting any polarization effects, in particular neglecting the dependence of current density on overpotential (Butler – Volmer equation, Eq. 2), permits calculation of the *primary current-density distribution* according to the Poisson equation, a fundamental principle of electrostatics. The primary current-density distribution neglects charge transfer kinetics.

However, the exponential dependence of current density on overpotential and the linear dependence of ohmic voltage drops on current density (Eq. 17) causes a notable smoothing and equalization, the *secondary current-density distribution*, which prevails whenever actual current densities everywhere at the electrode surface are much lower than mass-transfer-limited current densities ($j < 0.1\,j_l$): the secondary current-density distribution neglects mass transfer. Figure 15 depicts schematically the primary and secondary current-density distribution at a plane electrode. Wagner showed that geometric similarity can be applied to current density distributions of electrodes of similar form but different size whenever the Wagner number *Wa* is equal for two different electrodes or electrolysis cells.

$$Wa = \kappa \left(\frac{\text{d}U}{\text{d}j}\right) / d$$

$\frac{\text{d}U}{\text{d}j}$ = polarization resistance
d = characteristic length

The higher the Wagner number, the more uniform the current density distribution. Whenever mass-transfer limitations prevail ($j \approx j_l$), the secondary current-density distribution is outweighed by the current-density distribution as dictated mainly by the progressive depletion of

the depolarizer, which then defines the *tertiary current-density distribution* ($j = j\,(c^{x=\infty})$,

Eq. 18): the tertiary current distribution is observed for mass transfer-controlled currents and does not resemble the primary and secondary current distributions at all. Instead, it reflects only the locally variable mass-transfer rate of the particular electrode at the particular convective-diffusion or nonsteady diffusive mass-transfer conditions.

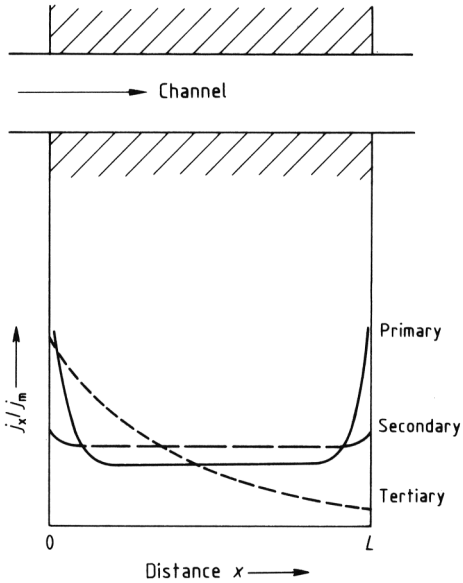

Figure 15. Primary, secondary, and tertiary current-density distribution for an electrolyte flowing between parallel electrodes
j_x/j_m = local current density/mean current density

Figure 16 shows some cases where tertiary current-density distributions are important whenever electrolyte solutions contain the depolarizer in low concentrations ($c < 10^{-2}$ mol/dm^3). A typical case is the current distribution in porous or packed-bed electrodes used for wastewater treatment (Fig. 16B) to extract noxious metal ions such as Cu^{2+} from galvanic effluents (see Section 4.6).

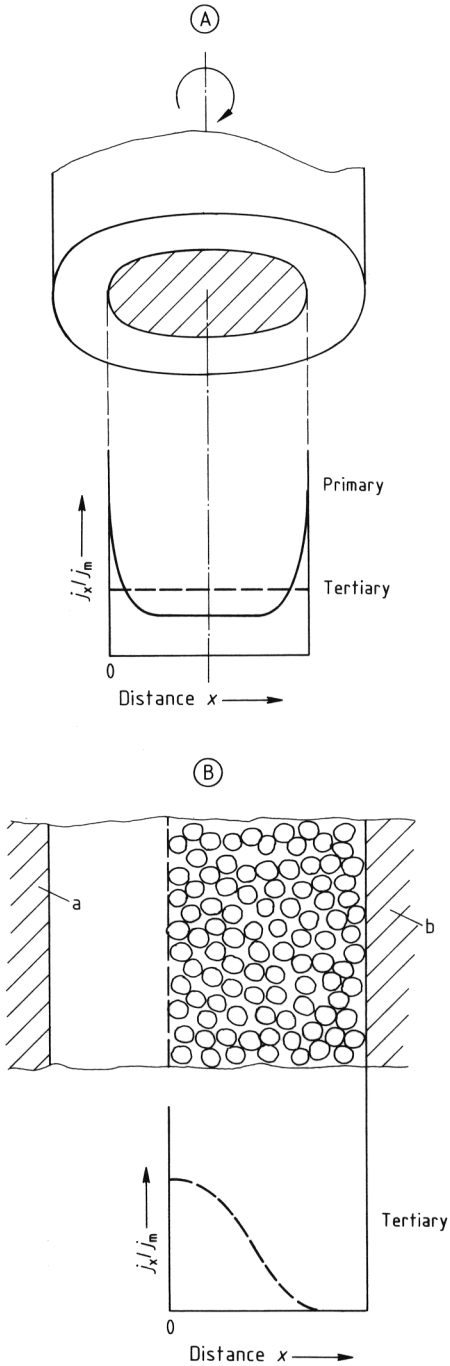

Figure 16. Schematic presentation of the primary and tertiary current-density distributions at (A) the rotating disk electrode and of tertiary current-density distribution in (B) a packed-bed electrode with (a) counterelectrode and (b) feeder electrode
Tertiary current density distribution is depicted for mass-transfer-controlled current densities in both cases.

3.3. Gas-Evolving Electrodes

Most of the important industrial electrochemical processes are accompanied by gas evolution, either at one electrode or both. The existence of a gas phase adhering to an electrode or in the electrolyte bulk significantly complicates operation of electrochemical reactors. Whenever gas bubbles form at an electrode, they interfere with the electrochemical and physical events. The interference is not restricted to the electrode surface or an adjacent electrolyte layer, but spreads throughout the whole interelectrode gap. To ensure optimal operation of these electrochemical reactors requires an understanding of the causes and effects of the manifold phenomena associated with gas evolution [23, 24, 25, 26].

The gas does not form at the electrode as an immediate consequence of an electrode reaction. The substance is formed on the electrode-electrolyte interface and appears first dissolved in the solution. For this reason, the highest concentration of dissolved gas occurs at this interface, and the concentration polarization caused by this may significantly increase the cell voltage.

When the interfacial concentration of dissolved gas exceeds a certain value, stable gas bubbles start to grow at nucleation sites, i.e., preexisting gas residues trapped in pits, scratches, and grooves on the electrode surface. The bubbles grow from the supply of dissolved gas in the supersaturated electrolyte and induce microconvection, which in turn increases the mass-transfer rate of dissolved gas and thus increases the bubble growth rate. Simultaneously, microconvection affects mass transfer of any substance to or from the electrode, e.g., species of the main reaction. Therefore, electrodes covered with growing bubbles exhibit mass-transfer behavior significantly different from that of bubble-free electrodes.

The microevents at the electrode are stochastic in space and time. For this reason, and unlike the situation for electrodes without gas evolution, the current density can be considerably different at points close to one another and at the same time vary at one point extremely quickly. Measured values are commonly average values.

When the bubbles detach from the electrode, tehy rise in the electrolyte, imparting momentum to it. Electrolyte flow past the electrode develops. Depending on the type of industrial electrochemical reactor, this flow may be sufficient to obviate the need for a pump. This effect is most pronounced in interelectrode gaps between vertical electrodes, but by skillful design of the flow channels it can also be utilized with horizontal electrodes. Violent macroflow of the electrolyte contributes significantly to mass transfer.

Another aspect of the effects of gas evolution results from the fact that bubbles adhering to the electrode partly shield the electrode surface and thus diminish its effective area. The shielding of the electrode results in an increase of the local current density and consequently increases the local overpotential. Gas evolution is seen to have two antagonistic effects on mass transfer: adhering gas bubbles interfere with mass transfer by decreasing the available electrode area, while their growth and departure from the electrode enhances convective mass transfer.

After detachment from the electrode, the bubbles remain for some time in the interelectrode gap where they increase the effective resistivity of the electrolyte. Since an increasing fraction of the dispersion consists of gas bubbles as the dispersion flows through the cell, the current distribution is nonuniform.

Regimes in Gas Evolution. Whether the substance generated at the electrode is transferred into the gas phase or stays dissolved depends on the conditions. For low current densities, the concentration gradient at the electrode – electrolyte interface is large enough to transfer the generated molecules dissolved to the bulk electrolyte by convective diffusion. The interfacial concentration of dissolved gas is too low to activate nucleation sites at the electrode surface, and gas bubbles are not formed. Increasing the current density increases the interfacial concentration of dissolved gas to such an extent that nucleation sites located on the electrode surface are activated. Stable bubbles form and grow by diffusion of dissolved gas from the surrounding supersaturated electrolyte. When the bubble attains a sufficient size, it detaches from the electrode. After a time, a new bubble forms at the same nucleation site, and the cycle restarts.

An important anomaly is coalescence of bubbles on the electrode surface, which may occur within 10^{-4} s when bubbles touch. A peculiar

phenomenon is the jump-off upon coalescence of two bubbles of nearly equal size, characterized by a departure of the united bubble from the electrode, followed by an immediate return to the electrode, where this bubble remains attached for some time [27]. Another type of bubble coalescence is the sliding of large bubbles upward along vertical electrodes while incorporating small bubbles. Coalescence also occurs after departure of bubbles from the electrode. Departing bubbles induce a wake flow that makes smaller bubbles slip across the electrode surface [28]. Under conditions that are not fully understood, tiny bubbles can be rapidly ejected from the electrode [29], [30]. These irregularities greatly complicate reliable predictions of the behavior of gas-evolving electrodes.

Large current densities, with associated high gas evolution rates, increase the number of active nucleation sites on the electrode surface and favor coalescence of bubbles at the electrode surface. A more or less extended gas film forms, initially covering discrete areas of the electrode, but which can finally cover almost the complete electrode area, interrupting the current. In the Hall-Héroult process for aluminum, this is known as the anode effect.

Rate of Gas Evolution. Figure 17 shows an overview of possible paths of hydrogen between electrochemical generation and the gas phase. After desorption from the electrode surface, the hydrogen may dissolve partly in the electrode material, partly in the electrolyte, and diffuse in both directions.

The hydrogen may partly be trapped in the electrode material (5), partly permeate through the electrode (6), or enter an adhering bubble by crossing the metal – gas interface (7). For most electrode materials these three paths are unimportant; however, they may be relevant for metals possessing high hydrogen solubility, such as palladium and some of its alloys.

Usually, almost all of the generated hydrogen dissolves in the electrolyte. The supersaturation of the electrolyte with gas at the electrode – electrolyte interface causes hydrogen to diffuse partly to bubbles adhering to the electrode, where it enters the gas phase (8), and partly to the electrolyte bulk where it may be incorporated into freely moving bubbles (9), or hydrogen may leave the interelectrode gap still in solution (10).

The fraction attributable to each depends on the operational conditions, the material and state of the electrode, and the kind of electrolyte, and may differ significantly from case to case. Usually path 8 is considered to be dominant, and the rate of gas evolution can be directly calculated from Faraday's law. However, for many cases this is an inadmissible simplification. Generally, only a fraction of the product generated is transferred to bubbles adhering to the electrode. This fraction, the gas evolution efficiency, is small for low current densities and continues to be markedly lower than unity for the current densities used for most industrial processes (500 – 10 000 A/m^2) [31, 32, 33, 34, 35].

The rate of gas evolution is even lower when the electrolysis product reacts either with the electrode (1, 2) or with the electrolyte (3, 4), e.g., in the production of hypochlorite, where no gaseous chlorine evolves.

Mass Transfer at Gas Evolving Electrodes. Gas bubbles are already formed at electrodes above very low values of the current density (about 1 A/m^2) and are generally able to affect mass transfer of any dissolved species. However, below about 1000 A/m^2, the effect of forced convection and/or single-phase natural convection usually predominates although gas bubbles are present at the electrode. Forced convective mass transfer may be estimated from well-known mass transfer equations. Single-phase free convective mass transfer is affected by concentration gradients rather than by temperature gradients [36]. An equation particularly adapted to the conditions of gas evolving electrodes is available [37].

Conditions change significantly as the current density is raised above about 1000 A/m^2. Not only the number of bubbles adhering to the electrode increases substantially as the current density is increased, but also their growth and thus the frequency of detachment from the electrode increases. Rapidly growing and detaching bubbles affect the mass transfer by vigorously stirring the boundary layer. The superimposed bubble-induced mass transfer becomes strong enough to govern overall mass transfer at large current densities. The action of bubble-induced mass transfer is rather complex and depends on

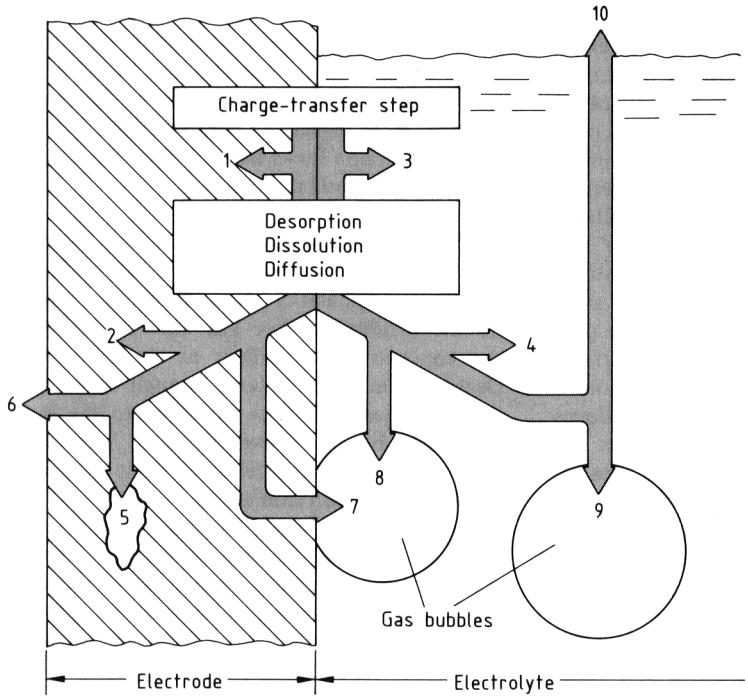

Figure 17. Schematic overview of the paths of hydrogen formed at an electrode
The various paths (1 – 10) are described in the text.

the pressure and temperature of the electrolyte; the rate of gas evolution (a function of the current density); and position, polarity, and surface of the electrode. Various models have been developed for bubble-induced mass transfer, each considering particular effects to be dominating [24].

In the *penetration model* [42], the core concept is that a bubble departing from the electrode induces a wake flow, which provides replacement of the boundary layer and cyclic penetration of fresh electrolyte to the electrode. The resulting equation is

$$Sh = \frac{2.76}{C}\,(ReSc)^{0.5}\,(1-\theta)^{0.5}$$

The dimensionless groups are

Sherwood number	$Sh = \frac{k_2 d_b}{D}$
Reynolds number	$Re = \frac{u_G^* d_b}{\nu}$
Schmidt number	$Sc = \frac{\nu}{D}$

The geometry of the adhering bubble is taken into account by the value of C, e.g., $C = 1.59$ for a hemispherical bubble and $C = 2$ for a spherical bubble. The parameter θ denotes the fraction of the electrode surface shielded by adhering bubbles [42], [43]. The Reynolds number is formed with the volume flow density u_G* of gas evolved at the electrode. The mean diameter of bubbles departing from the electrode d_b is of the order of 50 µm in aqueous electrolytes, although it can differ from this substantially, especially in molten salts [26].

The *microconvection model* takes account of the growth of bubbles adhering to the electrode surface. Growth pushes electrolyte over the electrode in a radial direction. A microconvective flow regime develops in the vicinity of the bubble. The resulting equation [44]

$$Sh = 1.89\,\left(Re^{0.5}\,Sc\right)^{0.487} \cdot \left[\theta^{0.5}\,\left(1-\theta^{0.5}\right)\right]^{0.5}$$

can be approximated in the range of mean current density values ($0.15 < \theta < 0.4$) by

$$Sh = 0.93 Re^{0.5} Sc^{0.487}$$

which exhibits fair agreement with experimental data.

The *hydrodynamic model* considers the flow of bubbles after detachment from the electrode as the predominant effect on mass transfer [45]. Adhering bubbles are disregarded. The fundamentals of the model have been discussed [24], but practical application is more problematic than for the other models. A promising attempt was made to develop a combined convection–penetration model [46].

The combined action of various mass transfer mechanisms, separately calculated for the mass transfer coefficients k_1, k_2, and k_3 may be roughly estimated from [38]:

$$k = \left(k_1^2 + k_2^2 + k_3^2\right)^{0.5}$$

The vector addition of the coefficients gives fair correlation with experimental results [39], [40] and has been confirmed theoretically [41].

Heat Transfer. The overtemperature of gas-evolving electrodes was long considered to be quite high. However, investigations have shown that the surface temperature of electrodes operating properly is only a few degrees above the electrolyte bulk temperature. Whenever an estimate of the electrode surface temperature is required, the bubble-induced rate of heat transfer may be estimated from [47]:

$$Nu = 0.93 Re^{0.5} Pr^{0.487}$$

where the additional dimensionless groups are

Nusselt number	$Nu = \frac{h\, d_b}{\lambda}$
Prandtl number	$Pr = \frac{\eta\, c_p}{\lambda}$

where h is the heat-transfer coefficient, λ is the thermal conductivity, η is the dynamic viscosity, and c_p is the specific heat capacity at constant pressure.

Electric Conductivity of Gas–Electrolyte Dispersions. The effective electric conductivity of gas–electrolyte dispersions may be substantially smaller than the conductivity of the bubble-free electrolyte. The conductivity is correlated to the volume fraction v of the gas in the dispersion (void fraction) and in an electrolyzer generally varies in the direction perpendicular to the electrode surface as well as in the main flow direction of the electrolyte. The effective conductivity depends not only on cell geometry, but also on the flow rates of evolved gas and the electrolyte.

The problem of the effective conductivity of a dispersion with uniformly distributed gas bubbles was first treated more than 100 years ago. There are numerous theoretical and some empirical equations for the effective conductivity κ of a dispersion relative to the conductivity κ_L of bubble-free electrolyte. The most important equations are those derived by Maxwell [48] and Rayleigh [49]

$$\frac{\kappa}{\kappa_L} = \frac{1-v}{1+v/2} = \left(1 + 1.5\frac{v}{1-v}\right)^{-1}$$

by Bruggeman [50]

$$\frac{\kappa}{\kappa_L} = (1-v)^{1.5}$$

and by Prager [51]

$$\frac{\kappa}{\kappa_L} = (1-v)\,(1 - v/2)$$

Series expansion shows that all three equations coincide for $v \to 0$. Further equations have been reviewed [26]. Which of these equations is the most satisfactory one is difficult to decide since their deviation for usual values of v are smaller than the scatter of experimental data and small enough to make each of the equations useful in practice.

The major practical problem is the assessment of v under real conditions. In most cases electrolyte flow is forced in industrial cells. Large electrolyte velocities are an efficient means to keep the gas fraction low.

The absolute velocity of gas bubbles in a swarm depends on the volume flow rates of gas and liquid, $\dot{V}_G$ and $\dot{V}_L$, and on the cross section S.

$$u_G = \frac{\dot{V}_G}{Sv} = \frac{\dot{V}_L}{S\,(1-v)} + \frac{u_{rs}}{1-v} \tag{19}$$

The last term represents the relative rising velocity of both phases considered as continua [52]. The rising velocity u_{rs} of a bubble swarm depends on the rising velocity u_r of a single bubble of diameter d_b in vastly extended stationary liquid,

$$u_r = \frac{g d_b^2}{C\nu}$$

and on void fraction v according to

$$u_{\mathrm{rs}} = u_{\mathrm{r}}\,(1-v)^n$$

where $n = 4.5 - 4.65$ [53], [54] and the values of C are 18 for a rigid bubble surface and 12 for a mobile one, particularly depending on the presence of surface-active substances in the electrolyte [55]. Equation (19) is accurate for large values of the ratio $\dot{V}_{\mathrm{L}}/\dot{V}_{\mathrm{G}}$ only, but fails for stationary electrolyte, $\dot{V}_{\mathrm{L}} = 0$, since it allows values of v equal to or near unity, which are unrealistic in bubble – electrolyte dispersions. Various modifications of Equation (19) to extend its range are available [56], [57].

If the gas fraction v of the dispersion flowing between planar electrodes is approximately considered constant in the direction perpendicular to the main flow direction, the mean relative conductivity of the dispersion is

$$\frac{\kappa}{\kappa_{\mathrm{L}}} = 1 + 0.75\frac{\dot{V}_{\mathrm{Gt}}}{\dot{V}_{\mathrm{L}} + u_{\mathrm{rs}}\,(1-v)^n\,S}$$

where $\dot{V}_{\mathrm{Gt}}$ denotes the total gas flow rate at exit of the interelectrode gap [58, 59, 60, 61, 62]. Neglecting the effect of the relative velocity u_{rs}, i.e., setting the slip ratio

$$\sigma = \frac{\dot{V}_{\mathrm{G}}}{\dot{V}_{\mathrm{L}}}\,\frac{1-v}{v} = 1 + \frac{u_{\mathrm{rs}}}{\dot{V}_{\mathrm{L}}/S}$$

approximately equal to unity [63], yields the simple, practical equation [64]

$$\frac{\kappa}{\kappa_{\mathrm{L}}} = 1 + 0.75\dot{V}_{\mathrm{Gt}}/\dot{V}_{\mathrm{L}}$$

Frequently, an older stagnant-electrolyte model [65] is used to assess the ohmic voltage drop, but such a model may involve heavy deviations from reality for all cases with forced flow.

The assumption of uniform bubble distribution in each area perpendicular to the main flow direction is applicable to the bulk of flowing dispersion, but fails for the nearly stagnant bubble layer close to the electrode, where bubbles are crowded. The assessment of this stagnant bubble layer resistance has been the object of several investigations [66, 67, 68]. Models taking account of the resistance of a stagnant layer together with the resistivity of the flowing central part are available [69, 70, 71].

Increasing the interelectrode distance of gas-evolving electrochemical reactors lowers the frictional pressure drop of the dispersion, and provides an optimization problem with respect to the voltage drop. Various systems with natural and with forced electrolyte circulation were studied with regard to the energy requirements of the total hydraulic system of electrolyte inside and outside the cell [72, 73, 74, 75, 76].

All these models refer to planar, nonperforated electrodes. For perforated electrodes, adequate models are not yet available.

3.4. Cells with Two-Dimensional Electrodes

The simplest electrode shapes are the plate and the cylinder. Owing to their generally low production costs, they have been used frequently and on an industrial scale. Due to increasing demands in economy and maintenance of industrial processes, electrodes of various special shapes have been introduced into industrial practice to an increasing degree. The perforated plate, extended metal, slitted, grooved, louvered, or other regularly deformed electrode surfaces do not really increase the geometric surface of the electrode. These electrodes are active on their outer surface only and are often termal two-dimensional to distinguish their action from that of the three-dimensional electrodes (see Section 3.6.).

Vertical Electrodes. Beginning many decades ago the form of vertical planar electrodes was modified by holes or slits [77], the action of which seems to have not always been fully understood. On one hand, the effective electrode area is diminished, but on the other hand, there is a gain in surface area due to accessibility of the rear side and the inner edges, although these are positioned at a greater distance from the counterelectrode. These modifications are advantageous in all cases of gas evolution at the electrode because release of bubbles at the rear side of the electrode is facilitated. As shown by experiment [78], the gas – electrolyte dispersion flows through the holes of a planar electrode – except for a rather short zone at the entrance side – and leaves the interelectrode gap (Fig. 18). Thus the effective ohmic resistance of the dispersion is reduced significantly. Attention must be paid to the size of the holes in

the electrode because holes too small, although desirable with respect to uniformity of current distribution, can result in gas bubble trapping [79], [80].

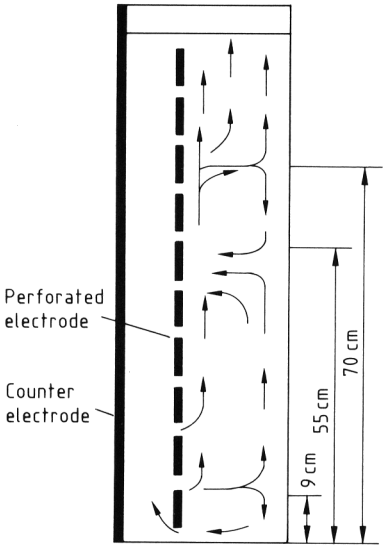

Figure 18. Flow pattern in a cell with vertical perforated electrodes [78]

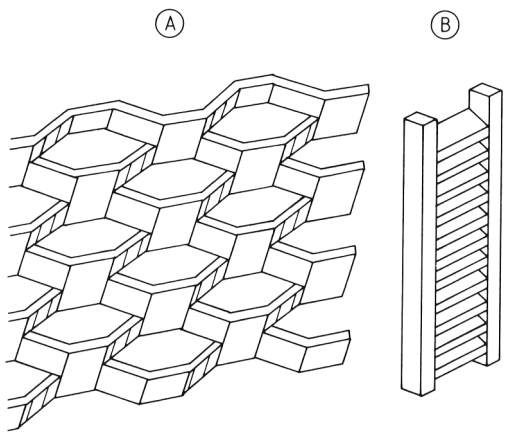

Figure 19. Gas-diverting electrodes
A) Expanded metal electrode; B) Louvered electrode [84], [86]

Unflattened expanded metal gauze electrodes and electrodes fitted with louvers or slots (Fig. 19) are especially effective as electrodes favoring the release of gas – electrolyte dispersion to the rear of the electrode [81, 82, 83, 84, 85]. Mesh electrodes provided with inclined ramps at regular intervals or louvered mesh electrodes serve the same purpose. The inclination angle of the fins has an optimum value [87].

However, gas diversion to the rear side of the electrode can make the design of the electrochemical reactor as a whole problematic. In cells with central monopolar (unipolar) electrodes, where gas evolution occurs equally at the front and rear sides, gas removal through the electrode is not feasible, and in cells with bipolar electrodes, the catholyte and anolyte compartments on opposite sides of the electrode ordinarily are separated hydraulically. In these two cases gas-diverting preelectrodes, which are electrochemically active and at the same time separate the interelectrode gap from a chamber where the gas – electrolyte dispersion is collected without unduely increasing the interelectrode resistance, are a possibility (Fig. 20). The dispersion essentially flows parallel to the electrodes through the rear chamber; the interelectrode distance between preelectrode and counterelectrode can be kept small. However, the use of preelectrodes conflicts with the need for compact cell design and often leads to bulky electrochemical reactors.

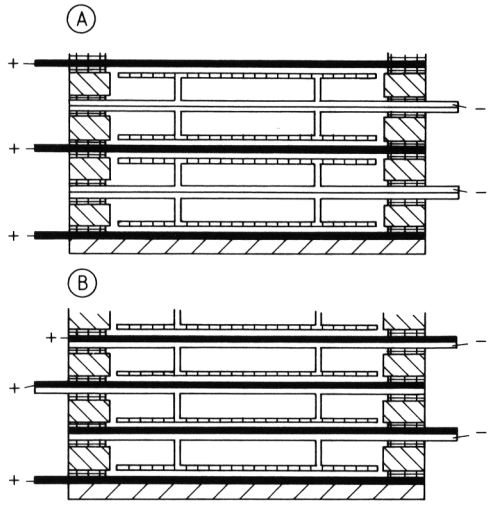

Figure 20. Preelectrode
A) Monopolar; B) Bipolar

Horizontal Electrodes. In the case of horizontal electrodes facing downwards, a design producing effective gas release usually gets highest priority, since gas bubbles can accumulate easily and coalesce to form abnormally large

bubbles or extended gas films, which insulate the electrode at least partially [88], [89]. An effective remedial measure is to impart a large velocity to the electrolyte removing the gas. However, this cannot be done in all processes.

Special problems are encountered in the electrolysis of alumina (Hall – Héroult process), where carbon anodes of considerable cross section are immersed from above into the molten cryolite electrolyte. Owing to the continuous combustion corrosion of these anodes inherent in the process, there is no possibility of shaping the electrode with a geometry that allows fast gas removal and that at the same time is sufficiently durable. The active surface of this electrode is essentially planar. The problem has encouraged extensive research, the outcome of which is more related to the extent of the shielding effect than to corrective measures. A gas bubble layer of considerable thickness is formed even in medium-sized electrode areas [90], [91]. However, a small inclination of the electrode surface from the horizontal induces a flow of gas – electrolyte dispersion that removes the gas safely [92]. Angles as small as 1° seem to be sufficient [89]. This possibility has scarcely been used systematically in the aluminum industry so far because of the shape instability of the carbon anodes.

In other cases, the most common way to ensure gas removal is the use of specially shaped surfaces. This technique has long been used on an industrial scale. Grooves in the electrode surface accumulate gas bubbles, which are removed by vigorous electrolyte flow. Of course, this electrode design causes a loss in active electrode area, as compared to planar surfaces, often resulting in considerable increase in effective current density and overpotential. However, this effect is more than offset by the beneficial action of smooth gas removal and can be especially advantageous for small interelectrode widths. The grooves are usually oriented parallel to the main flow direction, or if necessary are widened to balance the increasing gas volume transported [93]. With appropriate connecting vertical flow channels, the electrolyte flow is sufficient to dispense with any mechanical pump [94].

Wherever possible, vertical holes are used. This measure is easily able to prevent excessive accumulation of gas below the electrode. A typical example was the graphite anode that was used up to a decade ago in the chlor-alkali mercury process [95], [96]. The introduction of solid metal electrodes allowed the perforated graphite electrodes to be replaced by nets, expanded metal gauze, or parallel wires or rods, which allow the efficient release of chlorine [97], [98].

In special cases, e.g., in the electrochemical production of permanganate or in anodic oxidation of wastewaters, it is expedient to operate anode and cathode with distinctly different current densities. Instead of partially covering one of the electrodes [99], the electrode can be formed from a base plate covered by numerous layers of nets electrically connected with the base plate and thus active as the anode [100].

Microporous Electrodes. Fuel cell technology has shown that gases evolved at both electrodes of a cell can be separated without the use of a diaphragm if microporous electrodes are used. These electrodes were originally developed to meet the requirements of a stable gas – liquid – solid interface, but were found also to have favorable characteristics as gas-evolving electrodes. One of these is the so-called *valve electrode*, which is composed of two layers of markedly different porosity, each homoporous in itself (Fig. 21A). One layer, which is electrocatalytically inactive and in contact with the electrolyte, allows the electrolyte access to the other layer, which is electrocatalytically active and where gas is evolved. Because of its finer porosity, the first layer is permanently filled with electrolyte, and the gas formed can only be released on the rear of the electrode, which has the coarser porosity [101]. Major industrial importance is expected for another type of porous electrode, the *double-skeleton electrode* (Fig. 21B). Unlike the system just described, it utilizes a diaphragm. The electrode consists of a single porous layer that, however, incorporates two mutually piercing pore systems of different porosity. If the pressure difference between electrolyte and gas compartment is appropriately adjusted, the fine pore system in the electrode is filled with electrolyte, while the coarser one is not and serves for gas release. The concept dispenses with separate gas channels inside the cell block and allows compact design, which also does not require separate electrolyte compartments [102], [103]. To avoid mass-transport problems,

these electrodes must be thin. With respect to geometry, therefore, these electrodes are classified as two-dimensional, while operationally they form the transition to the three-dimensional electrodes.

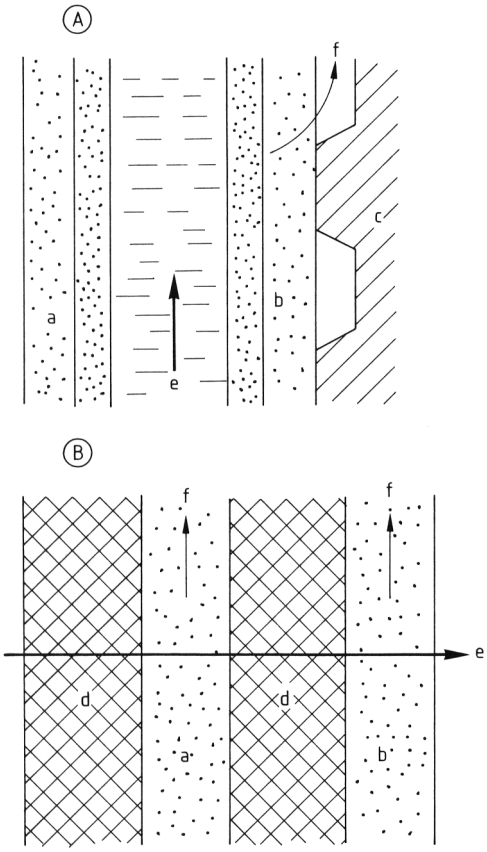

Figure 21. Cells with microporous electrodes
A) Valve electrode; B) Double-skeleton electrode
a) Anode; b) Cathode; c) Current collector; d) Diaphragm; e) Electrolyte; f) Gas phase

3.5. Special Cell Designs

During the past two decades, many specialized electrochemical cells have been designed. For example, the electrolytes used in organic electrochemistry often have conductivities much lower than those of aqueous solutions; therefore, the ohmic voltage drop constitutes a considerable part of the total cell voltage, thus resulting in high specific energy consumption [104].

A suitable cell design is required to offer a low ohmic voltage drop even with poorly conducting electrolytes. This objective can be achieved by minimization of the interelectrode gap. For the economics of an electrochemical process not only is the specific energy consumption important, but also the space – time yield [104],

$$STY = a_e \frac{j M_r \phi^e}{z F} \tag{20}$$

The space – time yield is proportional to the specific investment cost, which is the product of a design factor (the specific electrode area a_e) and a reaction factor (current density and current efficiency).

In many electroorganic reactions (see Chap. 5) and in the electrochemical deposition of metals from wastewaters (see Chap. 4.6), the current density is limited either by slow reaction or by diffusion. In such cases, improvement of the space – time yield requires a cell design incorporating large specific electrode area.

In Figure 22, the specific electrode area of various electrochemical cells is shown as a function of characteristic length (interelectrode distance or particle diameter); it is seen that available electrochemical cells differ in specific electrode area and characteristic length by orders of magnitude. As can be seen from Equation (20), for a current efficiency of 1 the space – time yield is determined by the product $a_e j$

$$a_e j = \frac{I}{V} = j^*$$

The quantity $j*$, in units of A/cm^3, is sometimes referred to as current concentration. This quantity is shown in Figure 23 as a function of the characteristic length. To estimate these $a_e j$ values, typical current densities for known applications of these cells were chosen. Extremely high current densities, e.g., 10 kA/m^2 in the chlor-alkali electrolysis, can be applied only in a limited number of cases, and the use of cells with high specific electrode area is restricted to cases of low current densities.

The special cell concepts are listed in Table 3, where each cell design is qualitatively characterized with respect to ohmic voltage drop, mass-transfer rate, and specific electrode area.

In the classical parallel-plate cells, which are usually applied as filter-press-type arrangements, *turbulence promoters* [3], [105], [106] can be introduced easily into the interelectrode

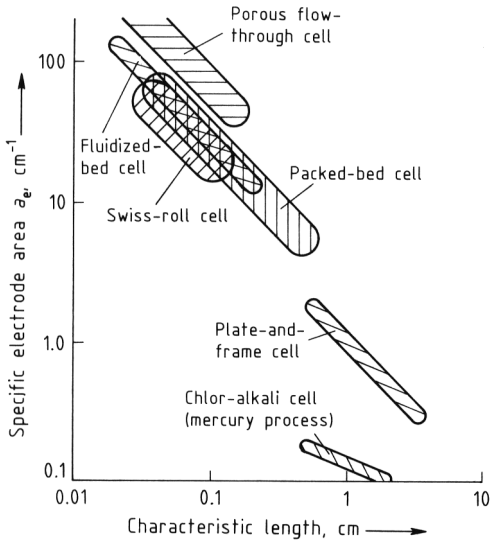

Figure 22. Specific electrode area as a function of characteristic length for various electrochemical cells

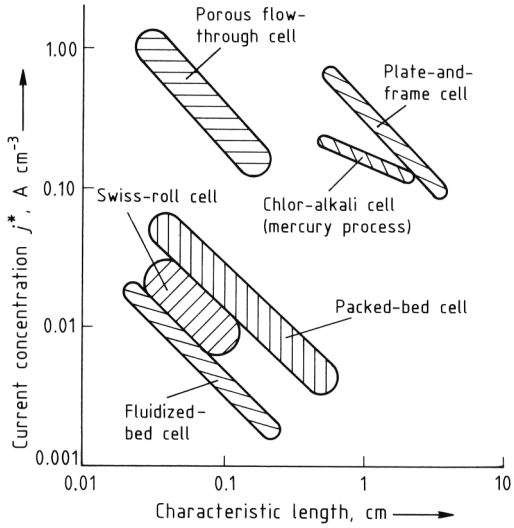

Figure 23. Typical current concentrations as a function of characteristic length for various electrochemical cells

gap or into the diaphragm – electrode gap to enhance mass-transfer rate. In Figure 24 a schematic view of such a cell is shown, with rods acting as turbulence promoters mounted between the electrodes of a parallel-plate cell, their axes perpendicular to the electrolyte flow. For channel flow, the regime changes from laminar to turbulent at a critical Reynolds number of ca. 2300. This critical Reynolds number is drastically decreased by the turbulence promoters, so that a higher mass-transfer coefficient providing a higher space – time yield, is achieved, at lower flow rates. A further advantage of turbulence promoters is the independence of the flow length and the diffusion layer thickness and thus the diffusion-limited current density. For laminar flow at a plate electrode within a channel, the integral mean diffusion-limited current density is [107]

$$j_l = 2.45 z F D^{2/3} L^{-1/3} d_e^{-1/3} u^{1/3} c_0$$

i.e., current density decreases as electrode length increases. This effect, which limits the extension of the electrode, does not occur under turbulent conditions, where the local current density is site-independent. Instead of rods, plastic tissues or nonconducting particles in the form of a packed or fluidized bed may also serve as turbulence promoters.

Table 3. Modern concepts of cell design

Cell	Ohmic voltage drop	Mass transfer rate	Specific electrode area
Cells with turbulence promoters	high	high	small
Rotating cylinder electrode cell	high	high	small
Trickle tower cell	high	high	large
Capillary gap cell	low	low	small
Pump cell	low	high	small
Swiss-roll cell	low	high	large
Zero-gap cell	low	low	small
Solid polymer electrolyte cell	low	low	small

Another possibility to enhance mass transfer is provided by *rotating electrodes.* As an example, the rotating cylinder electrode cell [108] is shown schematically in Figure 25. In this case the steep velocity gradient near the electrode surface, which induces a high mass-transfer rate, is not obtained by rapid motion of the fluid, but by rotation of the electrode. The mass-transfer equation of the rotating cylinder electrode is [109]

$$Sh = 0.079 Sc^{0.356} Re^{0.7}$$

$$Sh = \frac{k_m r}{D}, \; Sc = \frac{\nu}{D}, \; Re = \frac{r^2 \omega}{\nu}$$

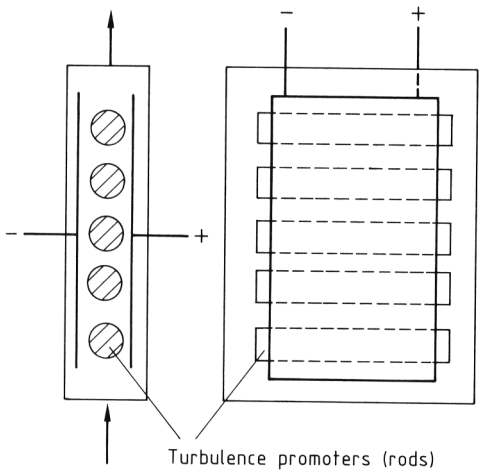

Figure 24. Parallel-plate cell with turbulence promoters

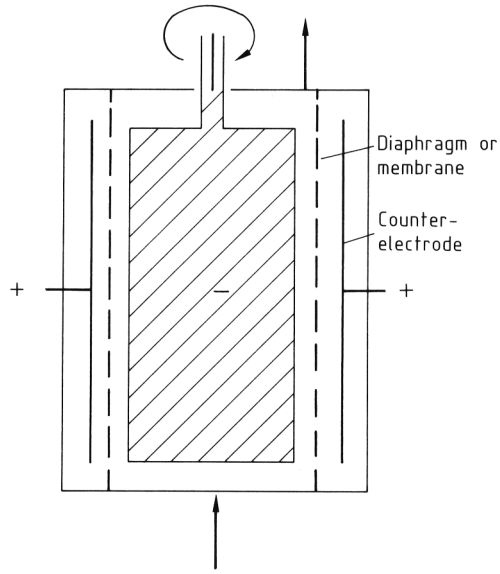

Figure 25. Cell with rotating cylinder electrode

The mass-transfer coefficient increases with the 0.7th power of the rotation velocity and the 0.4th power of the cylinder radius. Rotating cylinder electrode cells are used to recover silver from photographic effluents (see Chap. 4.6).

The *trickle tower cell* [110, 111, 112], depicted in Figure 26, is easy to construct and scale up. It is composed of several layers of conducting disks or particles separated from each other by insulating mesh. Each particle layer acts as a bipolar electrode. If the space between the particles is large enough, the trickle tower can also be used to produce gases, since the electrolyte flows as a film around the particles, and the gas may be removed via the electrolyte-free space in the tower. Rinsing a thin electrolyte film around the particles ensures bipolar operation, even for highly conducting electrolytes, because of the large electrolyte resistance parallel to the electric field. However, the trickle tower cell is restricted to cell reactions that can be carried out in an undivided cell, and is especially useful for cases involving a three-phase (gas – liquid – solid electrode) reaction.

A cell design combining ease of construction and scale up with a small interelectrode gap is the *capillary gap cell* [113], which is shown in Figure 27. It consists of a stack of bipolar circular electrode disks separated by nonconducting spacers. The electrolyte enters via a central channel and is radially distributed between the electrodes.

Although this design is an undivided cell, it can also be used to avoid reconversion of the products at the counterelectrodes. This can be achieved by a careful tuning of disk geometry and flow velocity to ensure that the diffusion layer thicknesses, which increase along the radius, do not reach the counterelectrodes [114]. By this means the products formed at the electrodes are prevented from contacting the counterelectrode. However, the products must be removed, e.g., by extraction, before the electrolyte can be recycled. Whereas the scale up of the stack height and the number of electrode disks is simple, the outer disk radius cannot be enlarged without limit. The slower linear flow velocity along the radius lowers the local mass-transfer coefficient [115], which lowers the mean current density and therefore the space – time yield as the outer radius is scaled up. The decreasing mass-transfer coefficient along the radial flow direction is avoided by the *pump cell* [116], [117], which is shown in Figure 28. A disk rotating around a central shaft is placed between two stationary electrode disks. By this means a tangential velocity component, which is caused by disk rotation and which increases along the

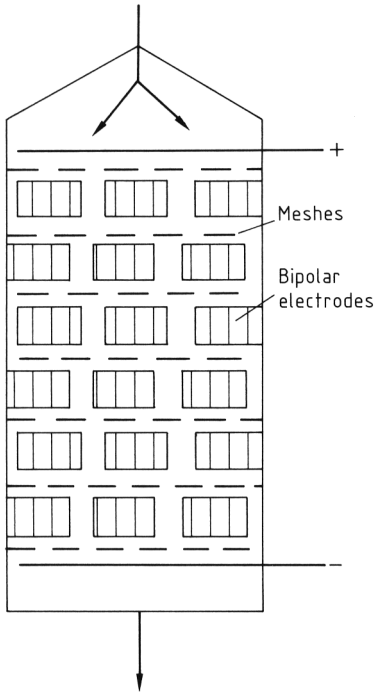

Figure 26. Trickle tower cell

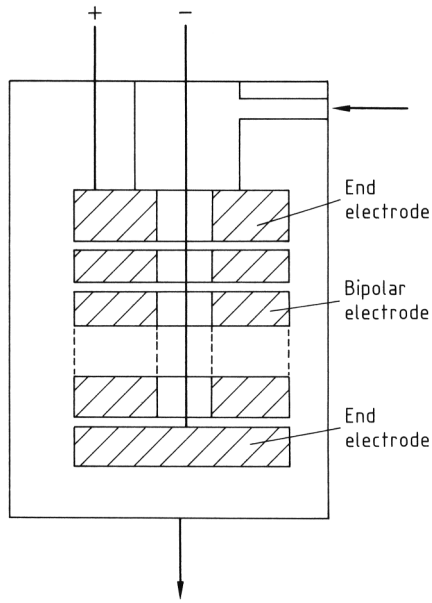

Figure 27. Capillary gap cell

radius, is superimposed on the radial flow velocity component, which decreases along the radius. This modified flow pattern keeps variation of the local mass-transfer coefficient and current density with the radius small [118]. The number of electrodes can be easily increased by always placing one stationary electrode between two rotating electrodes. Operation is bipolar, alleviating the need to contact rotating electrodes. The rotation pumps the liquid through the gaps of the stack.

An effective, surprisingly simple cell design for undivided cells is offered by the *Swiss-roll cell* [119], [120]. Two metal foils acting as electrodes are separted from each other by two insulating plastic meshes or tissues that act as turbulence promoters and are rolled up (Fig. 29). This cell simultaneously provides large specific electrode area and low ohmic voltage drop. Because of the thinness of the electrodes and their resistance a nonuniform current distribution results if the electrode contacts are positioned at the same end. However, this can be avoided by contacting the electrodes at opposite ends: one electrode contact is in the center of the roll, and the other is at the periphery. Scaling up of the electrode lengths then does not cause any trouble, but this is at the expense of a higher total cell voltage [121].

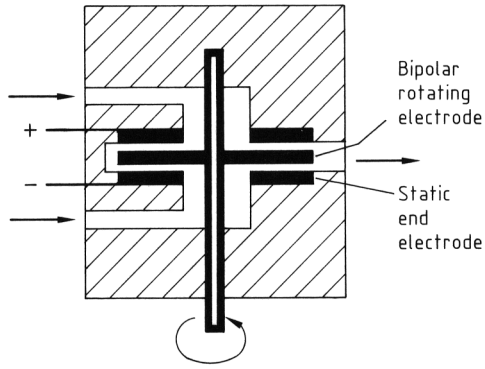

Figure 28. Pump cell

A cell design strictly directed toward minimization of the ohmic voltage drop in the electrolyte, especially if gases are evolved at the electrode, is the *zero-gap cell* [122], [123]. As

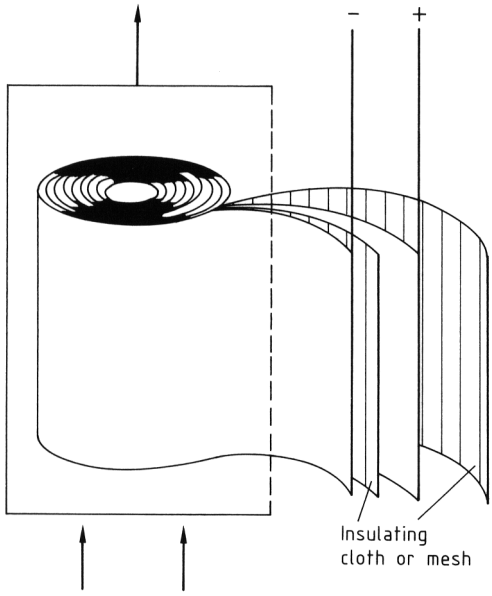

Figure 29. Swiss-roll cell

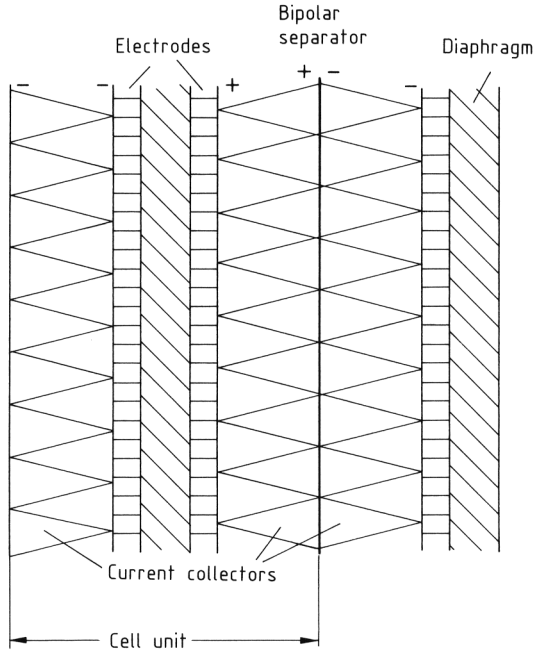

Figure 30. Electrode arrangement of a zero-gap cell

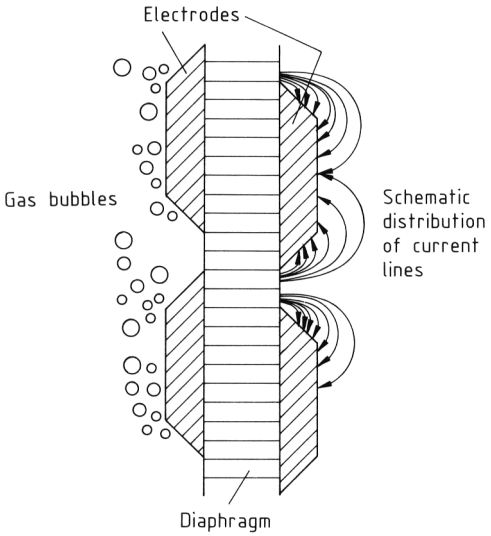

Figure 31. Gas evolution and current distribution at perforated-plate zero-gap electrodes

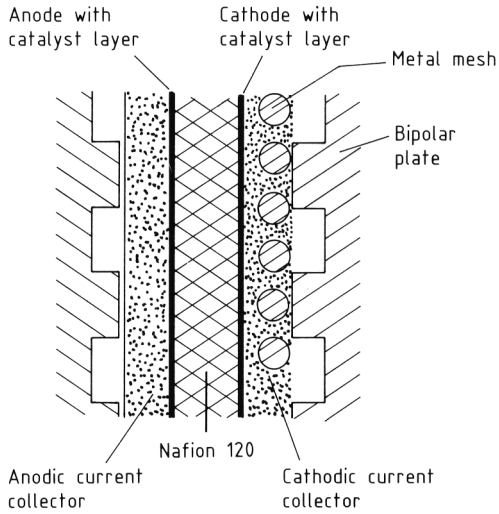

Figure 32. Solid polymer electrolyte (SPE) cell

shown in Figure 30, electrode meshes or perforated electrodes are pressed directly onto the diaphragm or the membrane by suitable current collectors, which simultaneously provide optimum electrode contact across the whole electrode area. This expedient avoids larger current crossways, which would give rise to an additional ohmic voltage drop. However, in such zero-gap cells, especially if equipped with perforated electrodes, an uneven current distribution arises, which is illustrated by Figure 31: the diaphragm cross section available for current flow is decreased to the vacant area of the electrodes, which increases the ohmic voltage drop across the diaphragm. The current must pass through the less conductive gas-containing electrolyte. A nonuniform current distribution is established around the electrode holes, requiring optimization of hole diameter and spacing, which depends on the current density. Zero-gap cells have been successfully used for alkaline water electrolysis [122], [123].

Ion-exchange membranes can be modified by putting the electrocatalyst directly onto the membrane surface [124]. Because of the surface resistance of the catalyst layers, the catalyst must be contacted by conducting, porous current collectors allowing the transport of dissolved reactants and gaseous products (Fig. 32). Such a cell – called the *solid polymer electrolyte* (SPE) *cell* – is used for water electrolysis (see page 48) and chlor-alkali electrolysis and has also been investigated for organic electrosyntheses (see page 102).

3.6. Cells with Three-Dimensional Electrodes

The rate of conversion in electrochemical cells is proportional to the electrode surface area. This is the incentive for designing electrochemical cells with a high electrode surface-to-volume ratio as actualized in porous electrodes, packed-bed electrodes, or fluidized-bed electrodes, which are usually referred to as three-dimensional electrodes [125, 126, 127, 128, 129, 130]. In these three-dimensional cells, not only the mass-transfer problem but also the current-distribution problem must be solved.

The current and potential distribution is mainly affected by the cell geometry, especially by the relative arrangement of the three-dimensional electrode and the counterelectrode. In regard to the relative directions of current and electrolyte flow there are two electrode arrangements: flow-through (current and electrolyte flow parallel) and flow-by (current and electrolyte flow perpendicular). For industrial applications, flow-by electrodes are preferred since they provide a longer residence time and therefore a greater degree of conversion per cycle. These two possibilities are illustrated in Figure 33.

In Figure 33, configuration A acts approximately as a differential reactor and has been mainly used for fundamental kinetic studies [131]. Arrangement B has been investigated as a rotating packed-bed electrolyzer [130] and as a radial flow-through electrode [132]. The cubic flow-by design (Fig. 33D) has been applied for fluidized-bed electrodes [126], [132, 133, 134, 135]. For packed-bed electrodes, the same design may be used with the modification of a variable depth (parallel to current flow) along the direction of electrolyte flow [136, 137, 138, 139]. Configuration F is the elementary unit of a design successfully applied in industry for the removal of copper from dilute acidic electrolytes [140, 141, 142].

The kinetic characteristics of a three-dimensional electrode are illustrated in Figure 34, which compares polarization curves for silver deposition from 10^{-3} m Ag^+ at identical flow velocities u for planar electrodes in a channel and for a flow-through packed-bed electrode. The curves for the planar electrodes indicate diffusion-controlled silver deposition. For the bed electrode, the diffusion-limited current density is ca. 100 times greater than that for the planar electrode, due to the larger actual electrode surface as well as to the enhanced local mass-transfer rate within the packed-bed electrode.

The mass-transfer coefficient can be obtained from

$$Sh = C\,Sc^{1/3}\,Re^{m}$$

In this case

$$Sh = \frac{k_{\mathrm{m}}\,v\,d_{\mathrm{p}}}{(1-v)\,D},\quad Sc = \frac{\nu}{D},\quad Re = \frac{u\,d_{\mathrm{p}}}{(1-v)\,\nu}$$

where v is the bed voidage and d_{p} the particle diameter

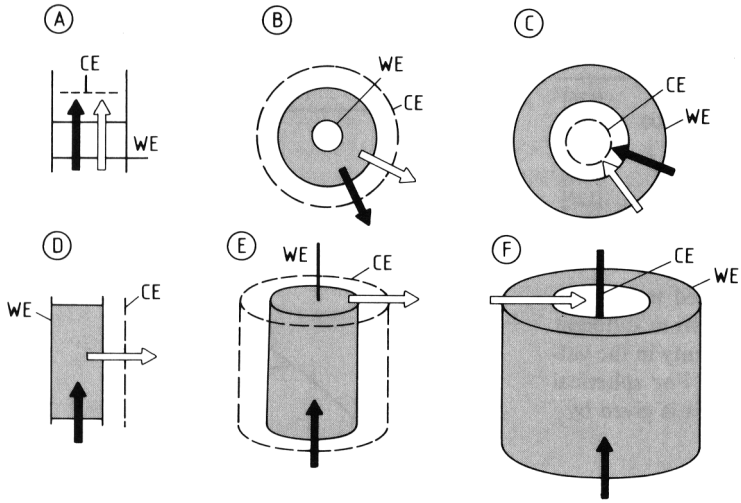

Figure 33. Geometrical arrangements of three-dimensional electrodes
A) Cubic flow-through electrode; B) Concentric flow-through electrode with outer counterelectrode; C) Concentric flow-through electrode with inner counterelectrode; D) Cubic flow-by electrode; E) Concentric flow-by electrode with outer counterelectrode; F) Concentric flow-by electrode with inner counterelectrodeWE labels the working electrode, CE labels the counterelectrode, → indicates electrolyte flow, and ⇒ indicates current flow.

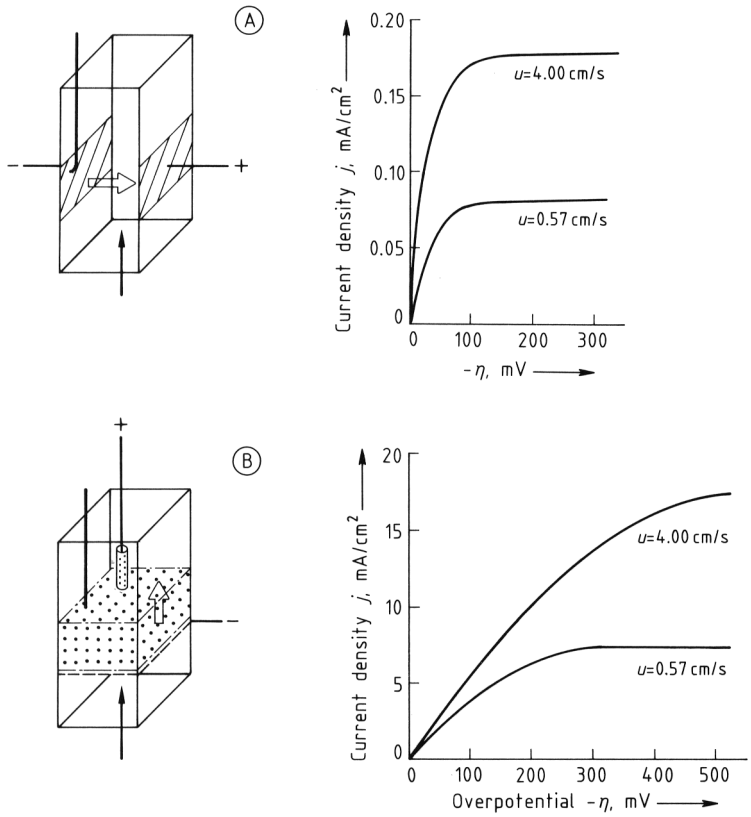

Figure 34. Polarization curves for silver deposition [33]
A) Plane electrode in a channel; B) Flow-through 4.6-mm packed-bed electrode

Some values for C and m are listed in Table 4. The variation in these constants from different references emphasizes the uncertainty in the values of mass-transfer coefficients. For spherical particles the specific electrode area is given by

$$a_{\mathrm{e}} = \frac{6\,(1-v)}{d_{\mathrm{p}}}$$

Table 4. Constants for mass-transfer correlations

System	C	m	Re	Ref.
Packed bed	0.50	0.61	50 – 5000	[143]
	0.32	0.66	100 – 10 000	[144]
Fluidized bed	5.7	0.22	1 – 30	[145]
	1.52	0.5	2 – 7	[146]
	0.29	0.75	100 – 5000	[144]

The macrokinetic behavior of three-dimensional electrodes expressed by the packed-bed polarization curves in Figure 34B can be explained by the potential distributions shown in Figure 35.

The *bed current density* plotted in Fig. 34 is defined as

$$j_{\mathrm{b}} \equiv a_{\mathrm{e}} \int_0^h j\,[\eta\,(x)]\,\mathrm{d}x$$

where h is the bed height and $j\,[\eta\,(x)]$ in general is represented by the expression

$$j\,[\eta\,(x)] = \frac{j_0\,[\exp\,(\alpha_{\mathrm{a}}\eta\,(x)) - \exp\,(-\alpha_{\mathrm{c}}\eta\,(x))]}{1 + \frac{j_0}{j_{\mathrm{l}}^{\mathrm{a}}}\exp\,(\alpha_{\mathrm{a}}\eta\,(x)) - \frac{j_0}{j_{\mathrm{l}}^{\mathrm{c}}}\exp\,(-\alpha_{\mathrm{c}}\eta\,(x))}$$

with

$$\alpha_{\mathrm{a}} = \frac{\beta^{+}F}{RT}\,,\ \alpha_{\mathrm{c}} = \frac{\beta^{-}F}{RT}\,,$$
$$j_0 = j_0' \prod_i c_i^{\nu_i}\,,\ j_{\mathrm{l}}^{\mathrm{a,c}} = k_{\mathrm{m}} z F c$$

For a mass-transfer-controlled reaction, the general expression simplifies to

$$j\,[\eta\,(x)] = j_{\mathrm{l}}\left[1 - \exp\left(-\frac{zF}{RT}\eta\,(x)\right)\right]$$

and for an activation-controlled reaction the Butler – Volmer equation (Eq. 2) can be applied. The local overvoltage is

$$\eta\,(x) = \phi_{\mathrm{p}}\,(x) - \phi_{\mathrm{s}}\,(x) - \Delta\phi_{\mathrm{eq}} \text{ with } \eta\,(0) = \eta_{\mathrm{B-F}}$$

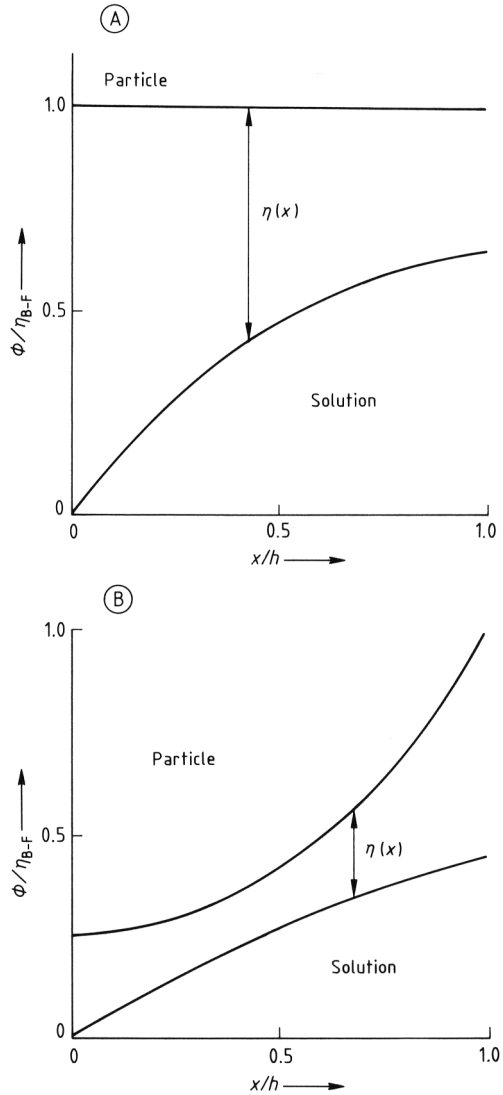

Figure 35. Schematic potential distributions
A) Packed-bed electrodes and B) Fluidized-bed electrodes [33]

For a packed-bed electrode, $\eta\,(x)$ decreases along the bed height x as can be seen in Figure 35A. For diffusion-controlled silver deposition, 99 % of the limiting current density is reached at an overvoltage of 0.120 V; i.e., the bed electrode reaches its limiting current density at a depth h where

$$\eta\,(h) = \phi_{\mathrm{p}}\,(h) - \phi_{\mathrm{s}}\,(h) \geq 0.120\,\mathrm{V} \tag{21}$$

However, $\eta_{\mathrm{B-F}}$, which serves as the reference for the potential axis in Figure 35, is already larger than 0.120 V, which explains that with a

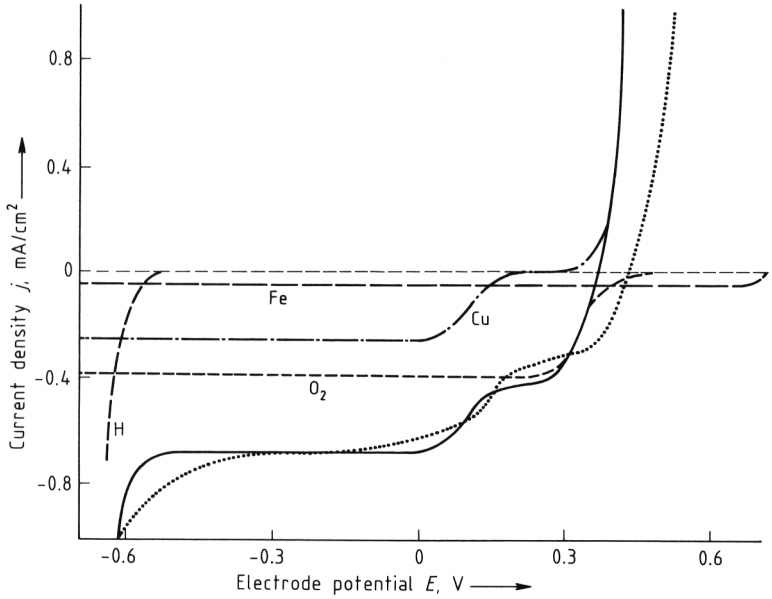

Figure 36. Partial and overall (— theoretical, ····· experimental) polarization curves for the electrolyte 10 ppm Cu^{2+}, 5 ppm Fe^{3+}, and 3 ppm O_2 at pH 3.5
The electrode potential is versus the NHE.

bed electrode the limiting current density is approached at higher potentials than with a plane electrode. If the flow rate is increased, the mass transfer coefficient increases, which results in a larger diffusion-limited current density j_l. For a given local overvoltage, a larger j_l yields a higher microkinetic current density $j\,[\eta\,(x)]$. Because of Ohm's law the $\phi_s\,(x)$ curve in Fig. 35 becomes steeper, and the η_{B-F} value that corresponds to Equation 21 becomes still larger.

The calculation of the macrokinetic bed current density requires quantitative knowledge of the potential functions $\phi_p\,(x)$ and $\phi_s\,(x)$.

According to Ohm's law, for three-dimensional electrodes of cubic geometry (Fig. 33A and D) the following pair of second-order differential equations are valid [4], [125], [131], [139], [147], [148]:

$$\frac{d^2\phi_s}{dx^2} = -a_e \frac{j\,[\eta\,(x)]}{\kappa_s}$$

$$\frac{d^2\phi_p}{dx^2} = a_e \frac{j\,[\eta\,(x)]}{\kappa_p}$$

For electrodes of concentric geometry the corresponding differential equations are

$$\frac{d^2\phi_s}{dr^2} + \frac{1}{r}\frac{d\phi_s}{dr} = -a_e \frac{j\,[\eta\,(x)]}{\kappa_s}$$

$$\frac{d^2\phi_p}{dr^2} + \frac{1}{r}\frac{d\phi_p}{dr} = a_e \frac{j\,[\eta\,(x)]}{\kappa_p}$$

To solve these systems of differential equations two boundary conditions are necessary for each equation. The boundary conditions for cell geometries A – F depicted in Fig. 33 are as follows:

A and D:

$$\phi_s\,(0) = 0 \qquad \phi_s\,(h) = \eta_{B-F} \qquad \frac{d\phi_p(0)}{dx} = \frac{d\phi_s(h)}{dx} = 0$$

B and E:

$$\phi_p\,(r_1) = \eta_{B-F} \qquad \phi_s\,(r_2) = 0 \qquad \frac{d\phi_s(r_1)}{dr} = \frac{d\phi_p(r_2)}{dr} = 0$$

C and F:

$$\phi_s\,(r_1) = 0 \qquad \phi_p\,(r_2) = \eta_{B-F} \qquad \frac{d\phi_p(r_1)}{dr} = \frac{d\phi_s(r_2)}{dr} = 0$$

In the general case the differential systems of equations can only be solved numerically. For limiting current electrodes – at each point of the electrode the microkinetic limiting current density is established – the substitution

$$j\,[\eta\,(x)] = k_m z F c = j_l$$

permits an analytical solution for the cubic geometry.

For cubic electrodes, the bed depth corresponding to a given potential variation $\Delta\eta$ can be expressed [131] for a fluidized-bed electrode

$$\kappa_p \geq \kappa_s \quad h = \left[\frac{2\Delta\eta}{a_e j_l \left(\frac{1}{\kappa_s} - \frac{1}{\kappa_s + \kappa_p} \right)} \right]^{0.5}$$

$$\kappa_p \leq \kappa_s \quad h = \left[\frac{2\Delta\eta}{a_e j_l \left(\frac{1}{\kappa_p} - \frac{1}{\kappa_s + \kappa_p} \right)} \right]^{0.5}$$

and for a packed bed electrode

$$\kappa_p \gg \kappa_s \quad h = \left[\frac{2\kappa_s \Delta\eta}{a_e j_l} \right]^{0.5} \tag{22}$$

For flow-through electrodes acting as differential reactors, the concentration of the reacting species may be considered as constant within the electrode. Modeling of flow-by electrodes requires an additional consideration of the concentration change along the electrolyte flow direction, which is given by the ideal tube-reactor equation:

$$\frac{dc}{dy} = -a_e \frac{k_m c}{u}$$

In fluidized-bed electrodes, the effective particle conductivity κ_p is determined by collision statistics since charge can be transferred through the particle phase only during particle collisions. Several theories for this have been discussed in the literature [149], [150] but the following formula shows reasonable accuracy [151]:

$$\varrho_p = \frac{u_p C}{0.159 \varepsilon^{1/3} \left(\varepsilon^{1/3} - 1 \right)}$$

where u_p represents the mean particle velocity, C is the double-layer capacity, and ε represents the height of the fluidized bed divided by that of the packed bed.

As an example for the complexity of fluidized-bed electrode behavior, the deposition of copper from an electrolyte containing 10 ppm Cu^{2+}, 5 ppm Fe^{3+} and 3 ppm O_2 at pH 3.5 may be considered [152]. The partial and the overall polarization curves are shown in Figure 36.

For a cubic flow-by fluidized-bed electrode (Fig. 33D) and the same electrolyte, the potential distributions along the flow length y ($y = 0$ corresponds to the electrolyte inlet) and the current-flow direction x (bed depth, $x = 0$ corresponds to

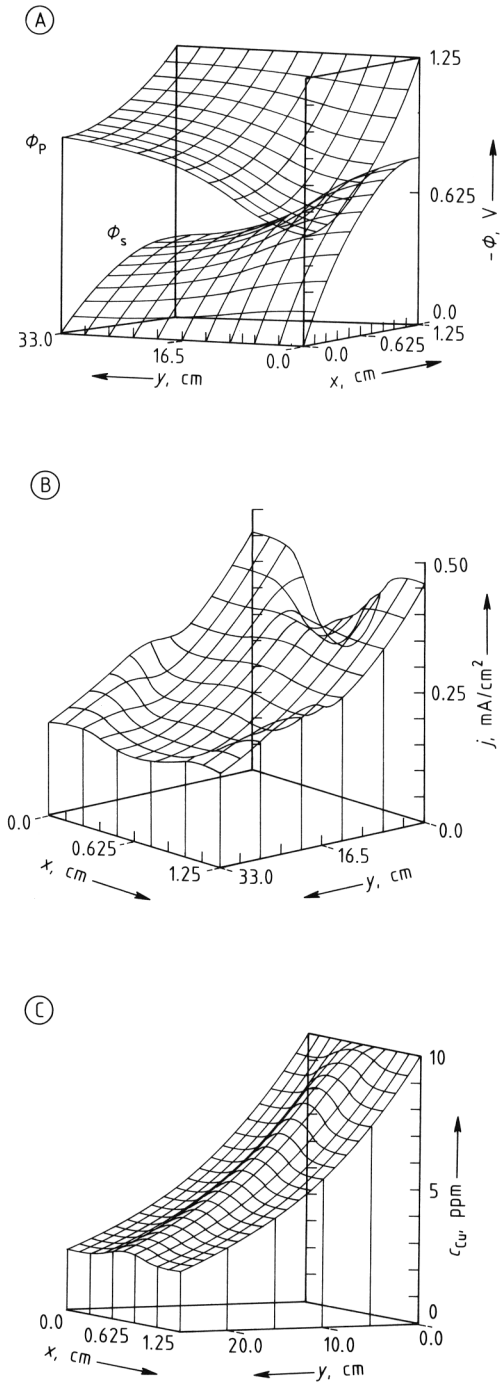

Figure 37. Electrolysis in a cubic flow-by fluidized-bed electrode
A) Potentials; B) Total current density; and C) Copper concentrationThe electrolyte contains 10 ppm Cu^{2+}, 5 ppm Fe^{3+}, and 3 ppm O_2 at pH 3.5.

the diaphragm plane) are shown in Figure 37A. The low overpotentials at the center of the inlet are responsible for low local current densities in this region (see Fig. 37B). The increasing current density tails at the electrode outlet are due to the beginning hydrogen evolution in these regions. The current density minimum at the inlet causes delayed copper depletion in the middle of the electrode as can be seen in Figure 37C.

Three-dimensional electrodes have industrial applications [153]. Packed-bed electrodes are mainly used for removal of metals from wastewaters (see Chapter 4.6). Fluidized-bed electrodes are more suitable for metal winning from solutions containing ca. 0.5 – 10 g of metal per liter, i.e., dilute solutions, since they allow the deposition of larger amounts of metal than packed-bed electrodes and the degree of conversion per pass of electrolyte is usually in the range of 90 %. Such applications have been described [130], [132, 133, 134, 135], [152], [154, 155, 156, 157, 158, 159]. This process has been used industrially by AKZO [141], [142].

Insight into the economic situation of metal winning by fluidized-bed electrolysis is given by Figure 38. For a specific investment cost of an available fluidized bed electrolysis plant of $0.142\,\mathrm{DM\,h^{-1}\,L^{-1}}$ ($1.8\,\mathrm{DM} \approx \$\,1$) the economic concentration limit is plotted as a function of the normalized space velocity ϱ_s^n [104], [160] and the metal price [152].

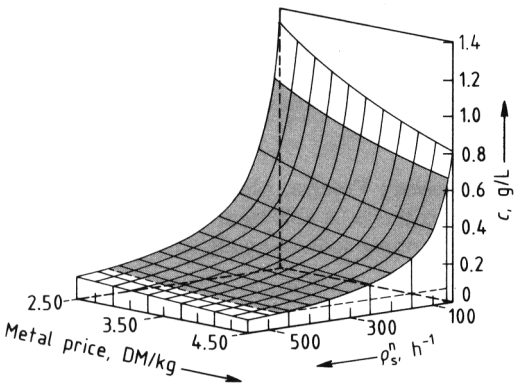

Figure 38. Economic lower limits of metal-ion concentration as a function of metal price and normalized space velocity ϱ_s^n for fluidized-bed electrolysis

3.7. Electrochemical Nanostructuring with a Scanning Tunneling Microscope

With the invention of the scanning tunneling microscope (STM) in the early 1980s by G. Bining and H. Rohrer [161], [162] a structure-sensitive technique for studies of the electrode – electrolyte interface with atomic-scale resolution has become available for electrochemists.

Besides imaging surface structures, the STM can also be used to manipulate surfaces on a nanometer scale [163], [164]. This so-called nanostructuring, the miniaturization of surface structures, is becoming increasingly important in areas such as electronic devices, sensors, and the tayloring of catalysts. Electrochemists have started to put into practice an electrochemical nanotechnology using the STM for nanostructuring electrode surfaces. This can be achieved, e.g., by creating tip-induced surface defects which then act as nucleation centers for metal deposition at preselected positions [165], [166].

In 1997 a novel process has been explored for making metal nanoclusters with the help of an STM [167, 168, 169]. The suggested mechanism for this tip-induced cluster generation is sketched in Figure 39. By applying a tip potential which is more negative than the bulk deposition potential for the metal ions in solution, metal is deposited from the electrolyte onto the tip (Fig. 39 B). The metal-loaded tip is then made to approach the surface for a short time so that a so-called jump-to-contact [170] can occur and a metal bridge between tip and surface is formed (Fig. 39 C). During the retreat of the tip the connective neck breaks and leaves a small cluster on the substrate (Fig. 39 D).

Figure 40 shows 12 Cu clusters, which were arranged in a circle on a Au(111) surface (the Au surface is covered with a monolayer of Cu at $E_{\mathrm{sample}} = 10\,\mathrm{mV}$ vs. $\mathrm{Cu/Cu^{2+}}$ due to underpotential deposition). Each of the clusters is about 0.8 nm in height (4 monolayers of Cu), the full width at half maximum (FWHM) of the clusters as imaged with the STM is about 3.5 nm, which yields an estimated 100 Cu atoms per cluster.

The electrochemical stability of these clusters is remarkable: at an electrode potential of $+10\,\mathrm{mV}$ vs. $\mathrm{Cu/Cu^{2+}}$ they remain practically unchanged for at least 1 h [168], [171]. By use of a microprocessor-controlled tip movement a large amount of clusters can be generated au-

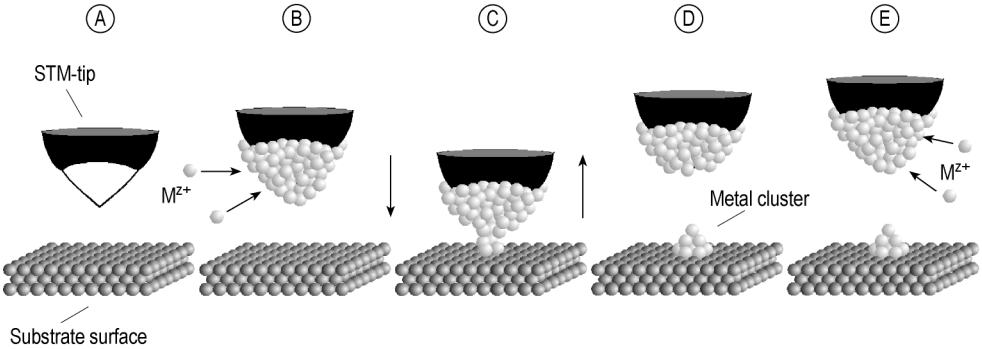

Figure 39. Suggested mechanism for the tip-induced metal deposition
A) Tip and metal surface; B) Deposition of metal from the electrolyte on the tip; C) Formation of a metal bridge between tip and surface; D) Formation of a metal cluster; E) Continuing metal deposition on the tip

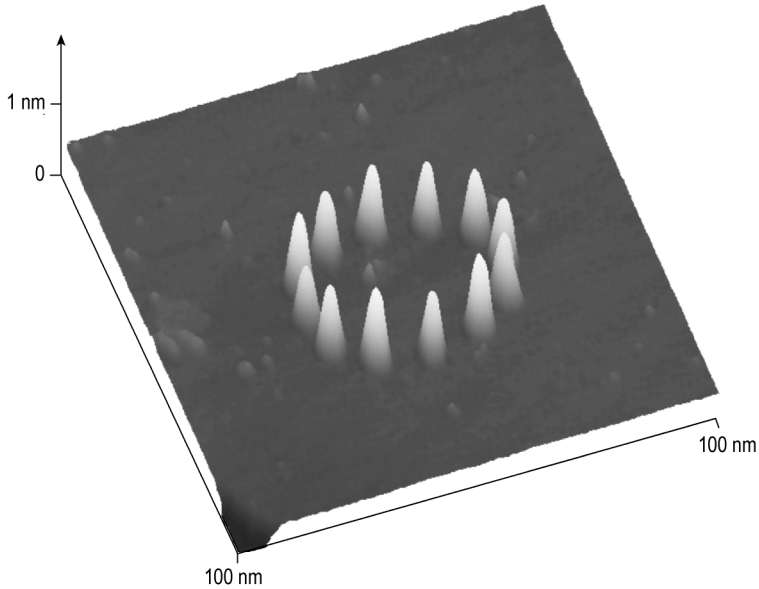

Figure 40. STM image of 12 Cu clusters on a Au(111) electrode in 0.05 M H_2SO_4 + 1 mM $CuSO_4$
E_{sample} = + 10 mV vs. Cu/Cu^{2+}; E_{tip} = − 50 mV vs. Cu/Cu^{2+}; I_{tunnel} = 2 nA [169]

tomatically, e.g., an array of 10 000 Cu clusters was generated with a rate of 50 Hz [172]. One advantage of this method is that no depletion of the deposited metal occurs, even when a large amount of clusters is produced, because the tip potential guarantees a continuous refill of metal on the tip.

It is not only possible to position the Cu clusters exactly, but also their size can be varied within certain limits [169], [171]. This is done by varying the approach of the tip: the closer the tip approaches the surface the more material is transferred (up to the point where tip and sample get into mechanical contact); if the tip approach is not close enough, no material is transferred, because a jump-to-contact cannot occur.

This method is not restricted to the system Cu/Au(111). Figure 41 shows an array of 400 Pd clusters on Au(111), the height of the clusters ranging between 0.2 and 0.6 nm. The generation of nanometer-sized particles of catalytically active metals such as Pd may be important for studying relations between catalytic activity

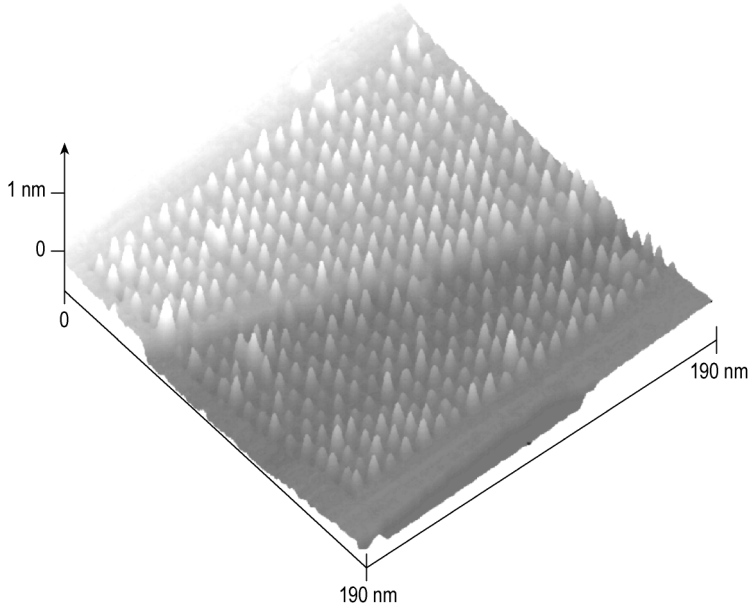

Figure 41. STM image of 400 tip-generated Pd clusters on Au(111) in 0.05 M H_2SO_4 + 1 mM H_2PdCl_4 E_{sample} = + 0.66 V; E_{tip} = + 0.48 V vs. SCE; I_{tunnel} = 2 nA [173]

and size of the clusters. Other systems are discussed in [171].

Present research centers on the nanostructuring of semiconductor surfaces with small metal clusters [174], because of the tremendous importance of small metal-semiconductor junctions for use in electronic devices, and the possibility of generating artificial structures from magnetic metals like Ni or Co is investigated.

4. Inorganic Electrochemical Processes

A number of inorganic substances are produced by electrochemical processes. Typical examples are the chlor-alkali process, which produces chlorine and caustic soda or chlorine oxides, chlorine oxygen acids, and their salts and hydrogen, the electrolysis of water to produce hydrogen and oxygen, and the anodic oxidation of aqueous sulfuric acid solutions to produce peroxodisulfuric acid and its salts. Electrochemical processing of metals is used to produce (electrowinning) and purify (electrorefining) copper, nickel, zinc, lead, aluminum, sodium, and magnesium, among others. Nuclear fuels and wastes can be processed electrochemically, and wastewaters can be purified by electrodeposition of heavy metals.

Anodic dissolution is the basis of electrochemical shaping, in which a cathodic tool is used to give a metal a particular shape (electrochemical machining or grinding) or a smooth surface (electropolishing).

4.1. Chlor-Alkali Electrolysis

The products of industrial chlor-alkali electrolysis are of great importance for the chemical industry: chlorine [*7782-50-5*], caustic soda [*1310-73-2*], hypochlorite, chlorate, and perchlorate. The rate of chlorine production is considered one of the measures of the industrial standing of a nation. Of the electrochemical processes, chlor-alkali electrolysis holds second place in total electrical energy consumption, aluminum production consuming more.

In industrial chlor-alkali electrolysis, the electrolyte is generally aqueous sodium chloride solution, to a much lesser extent potassium chloride solution. For the production of sodium

metal, fused sodium chloride is used. In all cases, the primary anodic reaction is the oxidation of the chloride ion, in itself a relatively simple reaction,

$$Cl^- \longrightarrow \tfrac{1}{2}\,Cl_2 + e^-$$

but this reaction can be accompanied or followed by other reactions. At the cathode, various reactions can take place: the products depend on the operating conditions. Chlor-alkali electrolysis justifies a detailed survey, not only on account of its economic importance, but also because of the interesting reaction technology. It is an excellent example of the way in which one anode reaction, from one raw material, can lead to a great variety of products, according to the way in which secondary reactions are prevented, encouraged, or controlled by the design of the equipment and choice of the reaction conditions. The introduction of dimensionally stable anodes based on ruthenium oxide – titanium oxide [175], which replaced the conventional graphite anodes in all aqueous chloride electrolyses, has contributed to far-reaching process modifications within the last three decades.

4.1.1. Molten Salt Electrolysis

In molten salt electrolysis, the cathodic reaction

$$Na^+ + e^- \longrightarrow Na$$

yields sodium [*7440-23-5*], which is of lower density than the melt, floats to the surface, and is then easily separated. Futher reaction of the two products does not take place, provided recombination is prevented. The high cell voltage makes the process useful only for winning sodium (→ Sodium und Sodium Alloys). Chlorine is in this sense a byproduct.

4.1.2. Production of Chlorine

Processes designed to produce chlorine are always based on electrolysis of aqueous chloride solutions, below 100 °C. The decomposition voltage is much lower than that for molten salt electrolytes. This technology is, however, considerably complicated by the secondary reactions of the primary electrolysis products and by side reactions.

The dissolved chlorine formed at the anode reacts with water to form hypochlorite,

$$Cl_2 + H_2O \longrightarrow HClO + Cl^- + H^+$$

The electrolyte near the anode becomes acidic. If the electrolyte is not replaced, or not replaced quickly enough, the supersaturation of chlorine at the anode increases to such an extent that stable chlorine bubbles form. Acidic electrolyte is the precondition of chlorine gas formation in aqueous solution. If the anode boundary layer is neutralized, the formation of gaseous chlorine is prevented. This situation arises if no steps are taken to prevent mixing of anolyte and catholyte.

At the cathode, hydrogen gas and hydroxyl ions are liberated simultaneously from aqueous solution:

$$2\,H_2O + 2e^- \longrightarrow H_2 + 2\,OH^-$$

There are two ways of ensuring the formation of gaseous chlorine: one is to prevent the hydroxyl ions from reaching the anode region, the other is to prevent the formation of these ions at the cathode in the first place. Both approaches are used in industry and have led to two technologies. One of these, the mercury process, uses mercury as the cathode material. The other technology uses cells divided by either a diaphragm or a membrane.

Mercury Process. Mercury [*7439-97-6*], the liquid cathode, is able to alloy the liberated sodium forming sodium amalgam (Fig. 42).

$$Na^+ + nHg + e^- \longrightarrow NaHg_n$$

The evolution of hydrogen is prevented by the exceptionally high overpotential of hydrogen at mercury surfaces. Hydroxyl ions are not formed. The amalgam increases in viscosity rapidly, so the sodium concentration must be kept low (< 0.3 wt % Na). High-load mercury cells require increased mercury throughputs. Outside the electrolysis cell, the amalgam is treated in a decomposer, a short-circuited galvanic cell. Here the sodium is liberated, and sodium hydroxide solution and hydrogen are formed. The caustic soda solution is both extremely pure and highly concentrated (50 wt %), but the high cell voltage, more than 1 V higher than that required

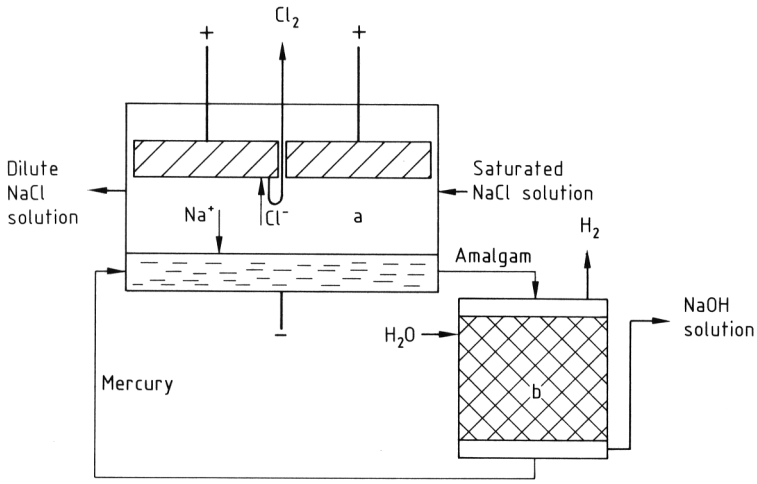

Figure 42. Mercury Process
a) Cell; b) Decomposer

for the competitive processes, and the toxicity of mercury have discouraged use of the highly developed mercury process.

Diaphragm Process. The diaphragm process allows the formation of hydroxyl ions but prevents their movement towards the anode region. The process uses cells divided into compartments by diaphragms made of microporous materials such as asbestos or synthetic polymers. These materials permit the flow of electrolyte but do not allow the OH^- ions to leave the cathode region if there is a sufficient flow of electrolyte from the anode into the cathode chamber, i.e., if the linear velocity of the electrolyte within the diaphragm is somewhat greater than the migration velocity of the OH^- ions in the electric field (Fig. 43A). This method allows high current efficiencies, but the sodium hydroxide that is formed is contaminated with chloride. Further treatment, such as evaporation and crystallization to remove sodium chloride, is unavoidable and expensive, which reduces the economy of the process [176]. And, unless special purification procedures are used, the alkali product (50 wt % NaOH) is contaminated with 1.3 wt % NaCl. The low cell voltage is an advantage, but the electrical resistance of the diaphragm produces an additional potential difference.

Membrane Process. The development of efficient cation-exchange membranes has in general made the diaphragm process obsolete. Fluorinated cation exchange membranes are used. Unlike diaphragms, membranes prevent the hydrodynamic flow of water but permit cation transfer by cation migration, and they enable the almost complete separation of anolyte and catholyte (Fig. 43B). However, the highly mobile hydroxyl ions cannot be completely prevented from crossing the membrane. High alkali concentration in the cathode region favors OH^- diffusion, contributes to buffering of the anolyte, and promotes hypochlorite formation and its anodic oxidation. The extent of these effects may be gauged by monitoring the oxygen content of the chlorine. As a result, efforts are being made to counteract hydroxide migration by improvements in plant operation, plant design, and membrane quality.

The membrane is not completely impermeable to water: the hydrated alkali-metal ion transports a part of its water through the membrane and water arrives at the catholyte side, preventing the concentration of the alkali from attaining the values reached in the mercury process.

Oxygen-Reducing Cathodes. Instead of a hydrogen-evolving electrode, an oxygen-reducing cathode can be used. This alternative allows substantial reduction (ca. 1 V) in cell voltage. Due to its purity, the catholyte of membrane cells is particularly appropriate for the use of air instead of pure oxygen. The high cost of

the membranes calls for high current density, ca. 3000 A/m^2, which today cannot yet be met by oxygen-reducing cathodes. The potential cell voltage reduction should stimulate further development [177].

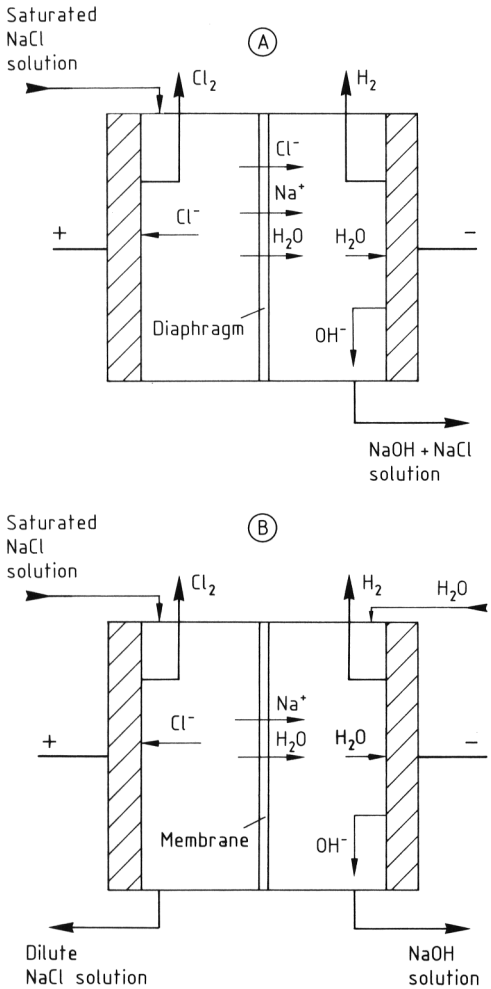

Figure 43. Diaphragm and Membrane Processes
A) Diaphragm cell; B) Membrane cell

For a detailed description of chlorine production → Chlorine.

4.1.3. Electrosynthesis of Hypochlorite

In undivided cells hypochlorite is formed instead of chlorine.

$$2Cl^- \longrightarrow Cl_2 + 2e^-$$
$$Cl_2 + H_2O \longrightarrow HClO + Cl^- + H^+$$

Hypochlorite is, however, not stable. In addition, it can be converted by anodic oxidation to chlorate or cathodic reduction to chloride. Both reaction rates are mass-transfer controlled. Therefore, high hypochlorite concentrations are impossible to obtain under the conditions for anodic chloride oxidation. The upper concentration limit is ca. 10 g/L. High production rates are possible only if the chloride solution is passed through the cell at high velocity, which also favors mixing of the electrolyte and prevents formation of chlorine gas. But, unfortunately, high electrolyte velocity favors mass transfer of hypochlorite already formed, thus lowering the current efficiency. An understanding of these relationships is necessary to optimize the production of hypochlorite by correct design and operation of the plant.

4.1.4. Electrosynthesis of Chlorate

Higher world tonnages are produced by this process than by any other inorganic electrosynthesis. There is no separator between the electrodes in the reactors, and mixing of the electrolyte is an integral feature of the process. The electrolyte, a concentrated sodium chloride solution similar to that used in chlorine production, flows at a high rate through the region between the electrodes. The formation of chlorate, rather than hypochlorite, is encouraged by design and operation of the installation. Chlorate is formed from hypochlorite by two mechanisms, by anodic oxidation

$$3\,ClO^- + 1.5\,H_2O \longrightarrow ClO_3^- + 3\,H^+ + 2\,Cl^- + 0.75\,O_2 + 3\,e^-$$

and by partial disproportionation of hypochlorous acid and subsequent autoxidation

$$2\,HClO + ClO^- \longrightarrow ClO_3^- + 2\,Cl^- + 2\,H^+$$

This chemical formation of chlorate is the desired reaction in industrial production because, in contrast to anodic oxidation, additional electrical energy is not required.

The anodic oxidation of hypochlorite is mass-transfer limited, and the rate is therefore proportional to the electrode surface and the residence time of electrolyte in the interelectrode

gap. The chemical formation, however, is homogeneous, and today all chlorate plants are provided with a comparatively large holding volume in addition to the space between the electrodes, this volume being either integral with the electrolyzer or constructed as a separate unit through which the electrolyte passes. The velocity between the electrodes is high, of the order of 1 m/s. Although this favors the mass transfer of the hypochlorite to the anode, it reduces the residence time in the interelectrode gap. Further conditions favorable to the chemical formation of hypochloric acid are temperatures of 80 – 90 °C, as attainable in modern cells with dimensionally stable anodes, and an optimum pH of 6.1 – 6.4.

4.1.5. Electrosynthesis of Perchlorate

In perchlorate electrosynthesis, the cell is undivided. Perchlorate is formed solely by anodic oxidation of chlorate. There is, however, a competing reaction, namely the anodic oxidation of water, and this has almost the same standard electrode potential. The relative rate of these two reactions is not significantly affected by pH or other operating conditions. However, the anode potential, which should be high, is of great importance. The process may be optimized only by suitable choice of the material of construction and physical state of the anode. Not surprising then, in contrast to the other process discussed, there has been no significant recent improvement in perchlorate manufacturing technology.

For a detailed description of hypochlorite, chlorate, and perchlorate production → Chlorine Oxides and Chlorine Oxygen Acids.

4.2. Water Electrolysis for Hydrogen Production

Electrolytic splitting of water [*7732-18-5*] ($\Delta G^0 = 241.9$ kJ/mol, $\Delta H^0 = 286.66$ kJ/mol) corresponds to a decomposition voltage of only 1.2 V and electrical energy consumption per 1 m^3 of H_2 (STP) of only 2.95 kWh, whereas the combustion enthalpy per 1 m^3 of H_2 (STP) of 3.68 kWh corresponds to a "thermoneutral cell voltage" of ca. 1.5 V. In this section, production in m^3 always refers to 1 m^3 of H_2 (STP). The present technology, with cell voltages of ≥ 1.9 V and energy consumption of 4.7 – 4.8 kWh/m^3, is still far greater. Water electrolysis is a long-established technique for producing hydrogen [*1333-74-0*] on a commercial scale. Nevertheless, it has gained only limited acceptance for large-scale production. In general, electrical energy, the cost of which constitutes more than half the cost of hydrogen production by water electrolysis, is too expensive to make electrolytic hydrogen competitive with hydrogen obtained by steam reforming of natural gas or oil.

Table 5 presents some information on the few large-scale water electrolysis plants; however, the facilities in Norway are closed now because even there hydroelectricity became too expensive. Nonetheless, the socalled energy crisis of 1972, together with potentially more extensive use of nuclear and renewable energies, has led to remarkable efforts to improve the established techniques – in particular alkaline water electrolysis – and to develop new ones – in particular membrane or solid polymer electrolyte (SPE) and high-temperature steam electrolysis.

Table 5. Large-scale water electrolysis plants for H_2 production

Location	Manufacturer	Capacity, m^3 of H_2 at STP
Aswan, Egypt	Brown Bovery	33 000
Nangal, India	De Nora	30 000
Ryukan, Norway	Norsk Hydro	28 000
Ghomfjord, Norway	Norsk Hydro	27 000
Trail, Canada	Electrolyser Inc.	15 200
Cuzco, Peru	Lurgi	4 500

Hydrogen production is also treated under → Hydrogen.

Alkaline Water Electrolysis. With the exception of the Noranda electrolyzer, all electrolyzers of substantial power (> 0.1 MW) currently on the market (Tables 6 and 7) are bipolar electrolyzers with forced convection of the electrolyte. All, with the exception of the newly developed Hydrogen Systems (H. S.) and the MTU electrolyzers, use asbestos diaphragms as separators, and most electrolyzers on the market approach zero-gap geometry.

All electrolyzers are operated with 25 – 35 wt % KOH (with the exception of the H. S. electrolyzer, which uses 14 wt % NaOH), and

Table 6. Large-scale commercial ater electrolyzerwater electrolyzers

Characteristic	Manufacturer					
	Electrolyser Corp. (Noranda)	ABB & Cie	Norsk Hydro	De Nora	(Lurgi) Giovanola∗	MTU∗∗
Cell type	monopolar tank	bipolar	bipolar	bipolar	bipolar	bipolar
Operating pressure, MPa	ambient	ambient	ambient	ambient	3	3
Operating temperature, °C	70	80	80	80	90	130
Electrolyte	KOH	KOH	KOH	KOH	KOH	KOH
Electrolyte concentration, wt %	28	25	25	29	25	30
Current density, A/m^2	1340	2000	1750	1500	2000	7000 – 10000
Cell voltage, V	1.90	2.04	1.75	1.85 – 1.95	1.86	1.65 – 1.8
Current efficiency, %	99.9	99.9	98	98.5	98.75	> 99.5
Power consumption (d.c.), kW h per 1 m^3 of H_2 at STP	4.9	4.9	4.1	4.6	4.3	4 – 4.4

∗ Technology transferred from Lurgi to Giovanola s.p.A.
∗∗ Maschinen und Turbinen Union Friedrichshafen GmbH.

Table 7. Smaller water electrolyzers (100 – 300 m^3/h)

Characteristic	Manufacturer		
	Hydrogen Systems	Hydrotechnik (System Demag)	Krebskosmo
Cell type	bipolar filter press	bipolar filter press	bipolar filter press
Operating pressure, MPa	0.7	ambient	ambient
Operating temperature, °C	120	80	80
Electrolyte	NaOH	KOH	KOH
Electrolyte concentration, wt %	14	25	25
Current density, A/m^2	0 – 5000	1500 – 2500	1000 – 3000
Cell voltage, V	1.5 to 1.7	1.9	1.65 – 1.9
Current efficiency, %	98	99	98.5
Power consumption (d.c.), kW h per 1 m^3 of H_2 at STP	≤3.9	4.9	3.9 – 4.6

the process temperature is generally above 75 °C, with an upper limit of 90 °C for the Lurgi and of 130 °C for the MTU electrolyzer.

Most of the electrolyzers work under ambient pressure, leaving compression of the product gases, if necessary, to compressors. Only the Lurgi and MTU electrolyzers, which in some ways represent the most ambitious technique, work under medium pressure, 3 MPa, whereas the newly developed H. S. electrolyzer operates at an only moderately enhanced pressure of 0.7 MPa. Low-temperature operation (< 80 °C) under ambient pressure permits the use of light equipment such as normal steel for electrodes, cells, pipes, pumps, etc., and helps keep investment costs low. Higher temperature and pressure call for heavier construction and for more care in particular to selection of steels with reliable corrosion stability and insensitivity towards stress corrosion cracking.

The advantages of higher temperatures and pressures are, however, convincing with respect to power consumption. Higher temperatures reduce the cell resistance because of improved electrolyte conductivity and reduce the electrode overpotentials due to thermal activation of the electrode process. High pressure reduces the volume of the evolved gases and gas holdup in the cell and thus reduces the effective electric resistance of electrolyte.

Figure 13 depicts a cell voltage – current density curve that is typical for the present electrolyzer generation having effective surface specific resistance around $1\,\Omega\,cm^2$. For improved pressurized electrolyzers this resistance is some 70 to 80 % less (Fig. 44). The steep increase of cell voltage with current density – mainly due to the relatively high cell resistance – is a good reason to keep current densities in conventional electrolyzers low, 0.2 – 0.3 A/cm^2, thus conserving expensive electrical energy. Only the

Lurgi electrolyzer in its most modern version, the MTU and the H. S. electrolyzers aim at current densities upwards of 0.5 A/cm^2 and can do so because of their relatively low surface specific resistances.

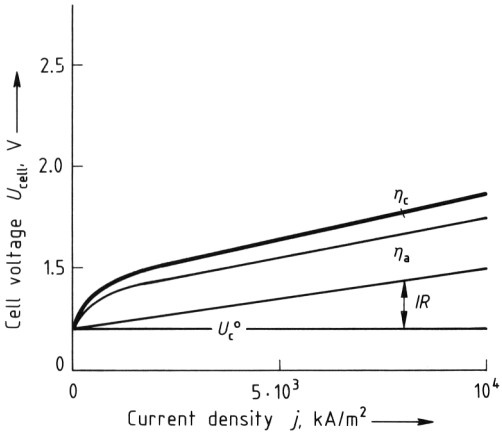

Figure 44. Current – voltage curves for advanced alkaline water electrolysis [178]

The three most advanced electrolyzer technologies allow hydrogen production with an energy consumption of 3.8 kWh/m^3 at current densities as high as 0.8 A/cm^2. This energy consumption approaches the water-splitting enthalpy of 3.68 kWh/m^3.

Electrocatalysis. Relatively little effort is needed to catalyze the cathode reaction. Most commonly used are high surface area nickel layers, either Raney nickel [*7440-02-0*] or porous nickel coatings produced in situ by cathodic reduction of nickel sulfide coatings.

$$Ni_2S_3 + 6\,e^- \longrightarrow 2\,Ni + 3\,S^{2-}$$

Leachable Ni_2S_3 coatings are obtained either by cathodic nickel deposition from galvanic Ni^{2+} baths containing sulfur donors (thiosulfate, thiourea, thiocyanate, and similar compounds) or by sulfidizing nickel electrodes in an atmosphere containing hydrogen sulfide. Raney nickel coatings can be applied on steel or nickel electrodes by cold-rolling [179]: the Raney nickel (Ni – Al prealloy containing 50 wt % Ni, corresponding approximately to stoichiometry $NiAl_3$) is bonded by pressure and shear to the electrode and is stabilized by subsequent annealing on the electrode. A second method consists of cathodically depositing Ni – Zn alloys containing Ni : Zn ratios in the range 1 : 1 to 2 : 1 [180]. Such coatings are relatively thin, between 50 and 150 µm. The well-established theory of porous electrodes indicates that the penetration depth of the current in these highly porous electrocatalysts is ca. 100 µm for current densities around 1 A/cm^2.

Improvements for advanced electrolyzers aim at decreasing cell resistances, the objective being to increase current densities above 0.5 A/cm^2 and still decrease cell voltages below the values used in the established technology. Electrocatalysis, especially at the anode, is now being introduced for the established electrolyzers as well.

Prior to public disclosure of the results of the recent research and development almost the only anodic electrocatalyst was nickel or nickel-coated steel anodes. NiO, or rather NiOOH, is by itself – compared to other metal oxides – a relatively effective electrocatalyst to enhance anodic oxygen evolution, and allows the use of an anodic overpotential lower than + 280 mV at current densities of 0.1 – 0.2 A/cm^2 at 90 °C in 30 wt % KOH.

Coatings of mixed oxides containing cobalt oxide such as cobalt nickel calcium perovskite and lanthanum nickelate ($Ca_{0.2}Co_{0.8}LaO_3$–$NiLaO_3$), cobalt nickel spinel (Ni_2CoO_4), and cobalt spinel (Co_3O_4), and iron – cobalt mixed spinel have proved to be relatively efficient electrocatalysts for anodic oxygen evolution. Indeed, the new H. S. electrolyzer is known to use Co_3O_4 coatings at the anode.

Diaphragms. Most impressive are the advances made in replacing the thick asbestos diaphragm – its thickness causing excessive electrolyzer resistances – by much thinner, mechanically strong diaphragm structures with electrical resistances approximately one order of magnitude lower than those of asbestos diaphragms.

The new separators make use of composite materials because one material seldom is able to fulfill the many opposing demands simultaneously – for instance, high porosity small thickness, and high mechanical strength. The diaphragm of the H. S. electrolyzer, for instance, is based on polysulfone, a mechanically strong, relatively easily processed, but hydrophobic polymer. In the form of highly porous

sheets it is hydrophilized internally by finely dispersed zirconia, an inorganic oxide of low water solubility. Similarly, porous PTFE sheets can also be hydrophilized by dispersed zirconia.

Quite different and comparably successful is the approach of using nickel net-supported porous ceramics such as NiO or $CaTiO_3$–Ni cermets, yielding mechanically stronger and more robust, yet flexible, diaphragms.

Water Electrolysis with Cation Exchange Membranes (SPE Electrolysis). The aqueous electrolyte may be completely dispensed with by use of a sheet made of proton-loaded water-swollen cation-exchange polymer as electrolyte (the solid polymer electrolyte, SPE, Fig. 45A, also see Fig. 32), which was first introduced by General Electric for fuel cells in space missions (Gemini flights). This fuel-cell option has successfully been converted into electrolyzer technology and offers a promising solution for small electrolyzers: any dangers and difficulties involved in handling hot concentrated caustic alkali solutions are avoided.

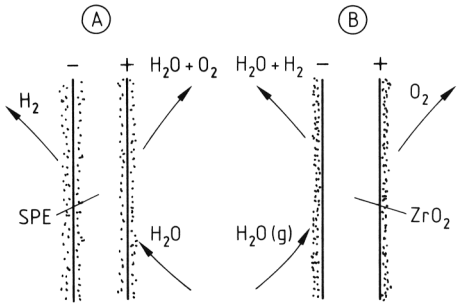

Figure 45. Nonconventional electrolyzers
A) Solid polymer electrolyte (SPE) electrolyzers decomposing deionized water at ca. 100 °C; B) Steam electrolyzers with zirconia electrolyte decomposing water vapor at 800–900 °C

The electrodes are formed within the outer layers of Nafion membranes by interdiffusing two solutions, one containing salts of the platinum group metals, the other an appropriate reducing agent, and precipitating in situ a coherent highly porous catalyst layer on either side of the membrane. The primary electrode deposits are then reinforced by further metal deposition. Current is transmitted through porous plates (platinized titanium anodes and carbon cathodes) and dense bipolar metal sheets. The porous metal plates permit simultaneous feeding of the highly purified water to be decomposed and removal of the evolved gases. Fig. 46 compares the current voltage curve for the solid polymer electrolyte electrolyzer with that of advanced alkaline electrolyzers and the steam electrolyzer. Typically, SPE electrolyzers are operated at current densities around 1 A/cm^2. At lower current densities, current efficiencies decrease substantially below 100 % because of counterdiffusion of the generated gases through the < 50-µm ion-exchange membrane.

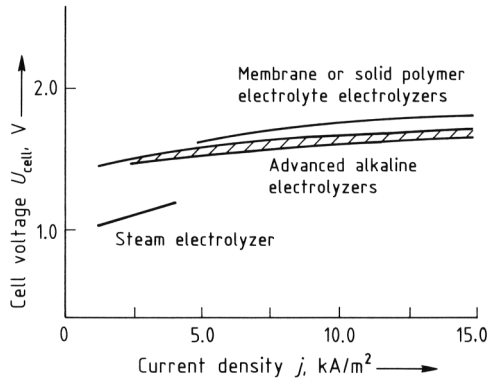

Figure 46. Comparison of current – voltage curves of solid polymer electrolyte, advanced alkaline, and steam electrolyzers

High-temperature Steam Electrolysis. The ionic conductivity of zirconia, which between 800 and 900 °C becomes comparable to that of 10^{-1} M aqueous electrolyte solutions at ambient temperature, permits high-temperature steam electrolysis. This technique is particularly appealing because the decomposition voltage, as calculated from $\Delta G°$, decreases as the temperature increases and drops to a value of only 0.9 V around 850 °C. The cell scheme (Fig. 45B) resembles the SPE cell. The cathodes are made of nickel cermet and the anodes consist of semiconducting Mn-containing perovskite layers. Although much effort has gone into the development of this electrolyzer, which is realized in the form of small cylindrical elements of 1 cm × 1 cm size, the technique is still far from competitive with conventional or advanced alkaline water electrolysis. The operating cell voltage of 1.3 – 1.4 V is still far from the theoretical 0.9 V, and quite a number of material problems are also unsolved. In particular, the thickness of the electrolyte is still too high, so that the internal

resistance of the cell does not permit current densities of more than a few hundred milliamperes per square centimeter. It might be possible that the development of solid oxide fuel cells leads to substantial improvements of steam electrolysis. At present, SPE and steam electrolysis are not commercially available.

4.3. Anodic Generation of Peroxodisulfuric Acid and Peroxodisulfates

Peroxodisulfuric acid and its salts – in contrast to the salts of most other inorganic peroxo acids, which are prepared chemically – are produced exclusively by anodic oxidation of aqueous sulfuric acid solutions:

$$2\,SO_4^{2-} \longrightarrow S_2O_8^{2-} + 2\,e^-$$

Preconditions for attaining high current yields in this process are the use of high concentrations of sulfuric acid (ca. 500 g H_2SO_4 per liter), high current densities (0.5 – 0.7 A/cm^2), low process temperature (18 – 20 °C), the addition of O_2-evolution inhibitors such as F^- anions, and the use of smooth platinum anodes, which favor adsorption of SO_4^{2-} anions. Reduction of the product is prevented by porous coatings on the cathode, which build up a diffusion barrier for peroxodisulfate anions.

Potassium, sodium, and ammonium peroxodisulfates are prepared with similar electrolysis technology from solution of the alkali-metal or ammonium sulfates and sulfuric acid (→ Peroxo Compounds, Inorganic).

4.4. Electrowinning and Electrorefining of Metals

4.4.1. Aqueous Electrolytes

Pyrometallurgical winning of metals is less expensive by far than electrowinning of metals by cathodic deposition from aqueous solution. But the pyrometallurgical processes have slight or only partial purification effects and yield impure metals that do not satisfy quality standards. Therefore, electrowinning and electrorefining have become important for the production of high-quality copper, zinc, lead, nickel, and cobalt. Electrorefining is based on the potential series of metals [181]. Anodic dissolution of impure metals if the current densities are small and therefore the overpotential negligible leaves the noble components of raw, unrefined metals undissolved. These noble components precipitate in the so-called anode slime or sludge, whereas the metals that are less noble than the metal being electrorefined dissolve with the main metal but are not redeposited on the cathode. Therefore, the less noble metals accumulate in the aqueous electrolyte. For any two metals M_1 with standard electrode potential E_1^0 and M_2 with E_2^0, the voltage difference $(E_1^0 - E_2^0)$ establishes the anodic dissolution equilibrium and defines the concentrations c_1 and c_2 of the two metal ions in equilibrium at a given electrode potential E according to the Nernst equation

$$E = E_1^0 + (RT/z_1 F)\ln c_1 = E_2^0 + (RT/z_2 F)\ln c_2$$

$$\frac{(c_1)^{1/z_1}}{(c_2)^{1/z_2}} = \exp\left\{-\frac{F\left(E_1^0 - E_2^0\right)}{RT}\right\}$$

The second equation neglects the activity coefficients of the metals and is applicable only if electrochemical dissolution/deposition is a fast process for both metals. This is the case for the metals Cu, Zn, and Pb but not for Ni, Co, and Fe. The deposition/dissolution of nickel is particularly sluggish, which diminishes the efficiency of an electrorefining process involving nickel because enhanced overpotentials are generated at both the anode and the cathode.

Alloy formation, which may change the chemical activity of the minor components greatly, is also not allowed for in the equation. For instance, zinc alloyed in nickel ultimately codeposits with nickel, although according to the difference in standard electrode potentials, $E_{Ni}^0 - E_{Zn}^0 = 0.53\,V$,

the equilibrium concentration of the two ions should differ by a factor of 10^9. However, zinc becomes much more noble upon alloying.

Such effects and the accumulation of impurities in the recirculated electrolyte make chemical treatment and purification of the electrolyte indispensable for all electrowinning and electrorefining processes. Since electrowinning with insoluble anodes starts from cruder materials than electrorefining, the chemical purification of

the electrolyte must be even more elaborate and expensive for electrowinning than for electrorefining.

Therefore, the important purification steps for electrowinning of nonnoble metals, like zinc, are hydrometallurgical rather than electrochemical. This is exemplified by the flow sheet in Fig. 47. The process begins with roasted galena, which is leached with recirculated electrolyte (cell acid). The silver and lead remain in the sludge. All elements nobler than zinc are then precipitated by cementation with zinc dust, and the cemented sludge is treated further to recover Cd, Ni, Co, and Fe. Iron can disturb zinc electrowinning seriously by electrochemical short circuiting:

$$\text{Cathode } Fe^{3+} + e^- \longrightarrow Fe^{2+}$$
$$\text{Anode } Fe^{2+} \longrightarrow Fe^{3+} + e^-$$

Therefore, iron(III) ions are precipitated as jarosite, a crystalline iron(III) sulfate oxide hydrate of extremely low solubility. As a second example, in copper refining accumulated nickel and other contaminants must be removed in a bleed stream.

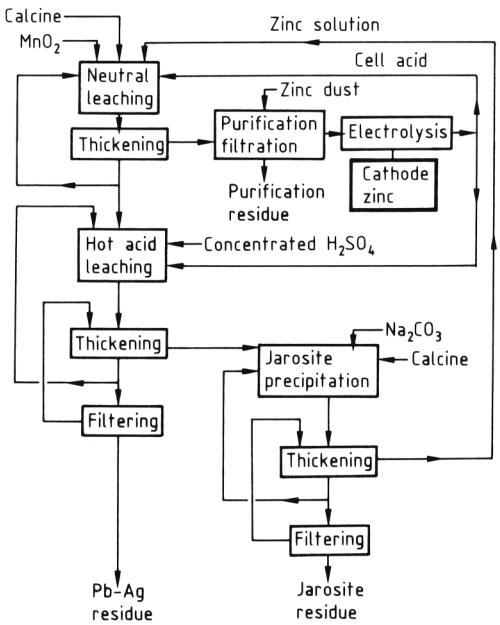

Figure 47. Flow sheet of the jarosite process [182]

For electrowinning and electrorefining of metals, mass transfer towards the cathode is of supreme importance. Convective diffusion driven by free (only gravity) convection caused by depletion of the solution at the cathode dominates mass transfer in these electrolysis processes. Anodes and cathodes, typically $0.8 \times 1.2\,m^2$, are suspended vertically at a distance around 10 cm from each other. They are arranged alternately one after the other in large rectangular tanks, which contain several hundred electrodes. Although there is a steady flow of electrolyte from one end of these huge tanks to the other, the flow velocity is small and electrolyte flow is not directed parallel to the electrode surface and thus does not contribute to convective mass transfer towards the electrodes to a significant degree.

The electrolyte for copper electrorefining contains ca. 40 g of Cu per liter (a concentration somewhat less than 1 M) and total depletion of copper from the solution decreases the density by several grams per 100 mL, i.e., by several percent. The Grashof number under these conditions

$$Gr = \left(x^3 g \Delta\varrho / \varrho\right) / \nu^2$$

would exceed the critical value of 10^{15} for x, the distance from the lower edge of the cathode, where the electrolyte is depleted of Cu^{2+} ions by electrodeposition of copper for turbulence; this distance x exceeds several centimeters. Therefore, for the greater part of the cathode length, usually of the order of 1 m, turbulent flow, with correspondingly intense mass transfer, prevails. But since mean mass-transfer coefficients do not exceed ca. 10^{-3} cm/s, mean current densities for cathodic metal deposition in electrowinning and electrorefining processes must be kept low, on the order of $200\,mA/cm^2$, to stay well below mass-transfer-limited current densities. Higher current densities cause the quality of the deposited metal to deteriorate: the deposit becomes coarse, forms whiskers, and begins to occlude or incorporate electrolyte, floating anode slime, and other impurities.

Even at current densities significantly lower than mass-transfer-limited values, deposit quality must be improved with surfactants. These additives increase the crystallization overpotentials for the deposited metal – inducing frequent nucleation, which gives rise to smooth cathode surfaces with little inclusion of foreign substances. Such surfactants are absolutely neces-

sary for obtaining highly purified metals and preventing the formation of dendrites, which can short-circuit the cell. Cheap, widely used additives are animal glue (gelatin), lignin sulfonates, and aloe extract, the last mainly polyphenols. All these additives are continuously consumed because of hydrolysis, oxidation by air, and adsorptive incorporation into the deposited metal and the anode slime.

Anodes for electrowinning processes are conventionally made of lead with few percent of silver. They have remarkably high oxygen-evolution overpotentials. In the future, activated titanium anodes will become more important.

Unlike all other electrochemical and chemical processes, the cost factor labor is extraordinarily high for electrowinning and electrorefining of metals because of the frequent electrode handling. In a process with anode lives of ca. 30 days (Cu, Ni) or only 10 – 14 days (Pb), and even shorter life spans for cathodes, there is the constant necessity of anode casting and the exchange of anodes and cathodes. Continuous progress in automatization is vital for the survival of these processes.

Aside from the labor costs, the cost structure of these processes are unique because of the investment costs for solution processing and the electrolyzer tank house and the capital required for the metal inventory in electrodes and circulated electrolytes.

Copper Electrowinning and Electrorefining. Copper [*7440-50-8*] is almost always electrorefined, less frequently electrowon. The raw material is the raw copper obtained by reduction of the oxidic ores or concentrates with carbon in a blast furnace. Raw copper is cast into anodes containing roughly 98 % Cu, the rest being mainly nickel and lead, with a fraction of a percent of noble metals, especially silver.

Table 8 summarizes the operational conditions. Nickel sulfate removal is accomplished in a bleed stream from which copper is electrowon prior to evaporation and crystallization of $NiSO_4$. The removal of Cu in *liberator cells* serves to balance the copper content of the bath and offset the additional chemical dissolution of copper due to oxidation by dissolved oxygen, which otherwise tends to increase the Cu^{2+} content of the electrolyte steadily (→ Copper, Chap. 6.2.).

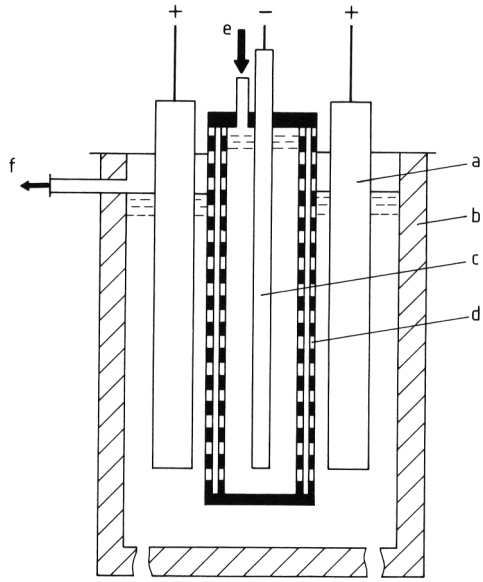

Figure 48. Hybinette cell for nickel electrowinning [182], [183]
a) Anode; b) Cell wall; c) Cathode; d) Diaphragm; e) Catholyte inlet; f) Anolyte exit

The cost contribution of electrical energy is an insignificant fraction of the whole process cost. Therefore, increasing the space – time yield is of special concern. Increasing the current density by at least 50 % by periodic current reversal is one approach now being used. The current is reversed for 10 s out of every 200 s, and correspondingly 5 % of the deposited copper is redissolved. This causes electropolishing, which produces high-quality, smooth deposits at high current densities. Another approach (Onahama refinery, Japan) consists of casting thinner anodes, which reduces the copper inventory by 50 %, and increases the space – time yield by ca. 20 % because the electrodes can be spaced closer together without short-circuiting. However, labor costs are increased by this measure, as does the amount of recycled anode scrap.

Electrowinning of copper is not fundamentally different from electrorefining, except that the depleted electrolyte is recirculated to leach copper ores. Purification is today mainly based on solvent extraction. The current densities can be greater for electrowinning because the stirring by the oxygen evolved at the anode enhances bath convection and mass transfer.

For more details see → Copper.

Table 8. Typical energy requirements in electrorefining and electrowinning

	Copper		Nickel refining		Zinc
	Refining	Winning	Metal anode	Sulfide anode	electrowinning
Current density, mA/cm^2	210	300	200	200	570
Current efficiency, %	97	85	96	96	90
Cell voltage, mV	280	2000	1900	3700	3500
Reversible cell voltage	0	900	0	350	2000
Cathodic overpotential	80	50	250	1500	150
Anodic overpotential	30	600	300	250	600
Ohmic voltage drops	100	400	1050 *	1250 *	500
Cell hardware IR	70	500	300	350	250
Electric energy requirement, kW h/kg	0.25	2	1.9	3.5	3.3

* With diaphragm.

Nickel Electrowinning. Nickel [*7440-02-0*] electrowinning begins with either crude nickel or nickel matte (Ni_2S_3) anodes [182, p. 357 ff]. In both cases the metallic contaminants to be removed are the same: mainly copper, cobalt, and iron. In both cases the anolyte must not be allowed to enter the cathode chamber prior to purification. In the Hybinette cell (Fig. 48), the cathodes are surrounded by a cathode bag. The depleted catholyte flows out of the cathode bag into the anode chamber because of a small hydrostatic head. In the anode chamber, nickel ions pass into the electrolyte:

$$Ni \longrightarrow Ni^{2+} + 2e^-$$

or

$$Ni_2S_3 \longrightarrow 2\,Ni^{2+} + 3S + 4e^-$$

Then the recycled electrolyte is purified. Copper(II) is extracted by cementation with nickel powder, and iron and cobalt are precipitated together with lead by oxidation with chlorine and addition of basic $NiCO_3$. If nickel matte is the starting material, purified sulfur is recovered from the anode sludge. The operational conditions for both processes are shown in Table 8. Starting from nickel matte, rather than raw nickel, requires higher cell voltages, much higher than would be expected from the difference between the two standard electrode potentials, which differ by only ca. 0.3 V. This much higher cell voltage is caused by electrode kinetic hindrance and additional ohmic potential drop on account of an adherent porous layer of sulfur that covers the anode.

For more details see → Nickel.

Zinc Electrowinning. The chemical purification procedures involved in electrowinning zinc [*7440-66-6*] were discussed, as an example, at the beginning of this section. The peculiar circumstance of zinc electrowinning is the negative standard electrode potential of zinc (− 0.76 V), which would make zinc deposition impossible except for the fact that hydrogen evolution at zinc cathodes has a very high overpotential ($j_0 = 10^{-11}$ A/cm^2). Nevertheless, care must be taken to keep the current efficiency for zinc deposition as high as ca. 90 %. The process is carried out at the rather low temperature of 35 °C. Ten percent of the current goes into hydrogen evolution, which improves mass transfer enough that current densities can exceed 0.5 A/cm^2. The operational conditions for zinc electrowinning are listed in Table 8 [182, p. 363 ff].

For more details see → Zinc.

Lead Electrorefining. The electrorefining of lead [*7439-92-1*], introduced in 1903 as the Betts process, is especially effective in separating bismuth, which goes into the anode sludge. Chemical treatment of bullion is used to remove copper, and subsequently tin (together with As, Sb, and Bi) by precipitation with caustic soda. The crude anodes consist of (wt %) ca. 98 % Pb, 0.5 % Bi, 0.01 % As, 1 % Sb, 0.02 % Cu, and 0.02 % Sn. The electrorefining process uses either hexafluorosilicic acid (H_2SiF_6) or sulfamic acid (HSO_3NH_2) as

the electrolyte. Current densities are somewhat higher for the processes working with H_2SiF_6 (160 – 200 mA/cm^2) than for the sulfamic acid process (120 – 160 mA/cm^2). In both cases a relatively low temperature is maintained to retard hydrolysis of the electrolyte [182, p. 356 ff].

For more details see → Lead.

4.4.2. Melts

Manganese ($E^0 = -1.05$ V) is the most reactive metal that can be electrodeposited from aqueous solution. Metals even more reactive – aluminum, magnesium, and the lanthanides – are electrodeposited from salt melts. Cathodic hydrogen evolution, due to water decomposition, cannot interfere with cathodic metal deposition. Normally melt electrolysis is not used as a purification step because of the difficulty of purifying and recycling contaminated melts. A thoroughly purified oxide is the starting material for electrowinning. The Bayer process, for instance, is an integrated step in aluminum production since it produces aluminum oxide of very high purity (impurities < 0.1 wt %) from crude bauxite. The Bayer process is responsible for 40 – 50 % of the production cost of aluminum [184, p. 296].

Typical problems posed by melt electrolysis are first of all material problems: the *stability* of (1) cell linings against chemical attack of the usually agressive melt, (2) anodes against the corrosiveness of the anode gases (Cl_2, O_2), and (3) cathode material against reaction with the liquid metal. These three problems are especially severe because of the high process temperature, usually above the melting point of the metal: 660 °C (Al), 648 °C (Mg), and 97 °C (Na).

A second typical problem is dispersion or dissolution of the liquid metal into the electrolyte, which allows partial reoxidation due to convective transport of the metal mists toward the anode. There the metal reacts with evolved gas, or it is directly reoxidized at the anode. This phenomenon is caused by (1) the low surface tension of molten metals in contact with the molten electrolyte, (2) the relatively small density differences between molten metal and molten salt, and (3) strong convection due to magnetohydrodynamic effects and gas stirring. Cell geometry must be designed to suppress excessive convection or to provide a sufficiently large interelectrode gap (ca. 8 cm in aluminum smelters) to minimize metal reoxidation.

A further engineering aspect that is unique and typical for all high-temperature electrolytic processes is the careful calculation of heat balances and heat isolation devices for electrolyzers and potlines, which is indispensable for energy conservation and for steady, reliable operation of molten-salt electrolyzers over months and years.

Aluminum Production [184, 185, 186].
The electrolytic production of aluminum [*7429-90-5*] by electrolysis of Al_2O_3 dissolved in cryolite (Hall–Héroult process) is by far the most important electrochemical process based on the use of a molten-salt electrolyte.

$$Al_2O_3 + 3\,C \xrightarrow{6F} 2\,Al + 3\,CO$$

Six faradays of electricity are required for each mole of Al_2O_3.

The addition of Al_2O_3 to cryolite (*mp* ca. 1000 °C) decreases the melting point as the alumina content is increased up to 10 wt %, where the eutectic melts at 960 °C [186].

Figure 49 depicts the construction of a modern Hall–Héroult cell schematically. The cell is composed of an outer carbon or graphite lining and a row of adjustable prebaked carbon anodes. At the bottom of the pot there is the cathode, a pool of liquid aluminum, which is emptied periodically by suction and which is covered by the molten electrolyte into which are immersed the gas-evolving anodes. The melt is covered by a crust of solid electrolyte and a supply of fresh, loose Al_2O_3. Periodically crushing this crust serves to replenish the Al_2O_3 content of the bath. The cathodic reaction consists of reducing fluoro-oxo-aluminum complexes:

$$AlOF^{(n-1)}{}_n{}^- + 3\,e^- \longrightarrow Al + O^{2-} + nF^-$$

A mixture of carbon monoxide and carbon dioxide is evolved at the anode: the depolarizing carbonaceous anode material reacts with the anodically generated oxygen,

$$(x+y)\,C + (x+2y)\,O^{2-} \longrightarrow x\,CO + y\,CO_2 + (2x+4y)\,e^-$$

In accord with this anode reaction, the Hall – Héroult process consumes significant amounts

of extra energy due to the consumption of the carbon anode.

There is violent stirring of the molten metal and the electrolyte due to magnetohydrodynamic effects and anodic gas evolution. Vigorous stirring may cause severe losses in current yields and energy efficiency. The reoxidation of metal mists at the anode or the reoxidation of the dispersed metal by the anode gases causes a loss of energy. These energy losses are avoided to a certain extent by increasing the anode – cathode distance. Aluminum smelters operated with optimal interelectrode distances of a little less than 8 cm achieve current efficiencies exceeding 93 % and electrical energy consumptions of optimally 13 kW h/kg (theory: 6.34 kW h/kg). The relatively low energy efficiency of the Hall–Héroult process is the reason more energy-efficient electrolysis processes have been sought.

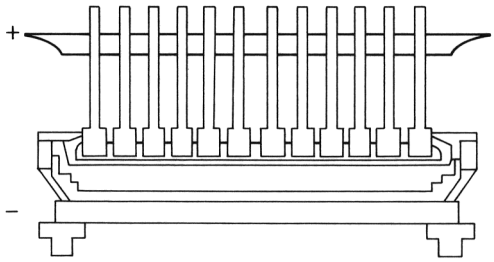

Figure 49. Hall – Héroult cell

In 1954, Alcoa resumed the work of Bunsen [187], who was the first to produce aluminum metal by electrolysis of alkali chloride – aluminum chloride melts. Alcoa developed the technology of aluminum deposition from alkali chloride – aluminum chloride melts up to a commercial scale. The Alcoa electrolyzer uses stacks of bipolar carbon electrodes. The upper parts serve as cathodes, and the lower grooved parts serve as chlorineevolving anodes. This electrolyzer is reported to produce aluminum metal with a specific energy demand of 10 kW h/kg, offering a 30 % saving in electrical energy over the Hall–Héroult process. In LiCl melts of low $AlCl_3$ content the energy savings seem to be still more promising [188]. However, further development of chloride electrolysis technology also requires the development of a commercially attractive route for $AlCl_3$ production from bauxite [188]. However, the Alcoa process eventually failed, because the anodically generated chlorine was contaminated with perchlorinated aromatic compounds which could not be eliminated at reasonable costs.

For more details on aluminum production, especially the Hall–Héroult process, see → Aluminum.

Magnesium and Alkali Metal Production. Figure 50 shows the decomposition potentials of the alkali-metal chlorides as functions of temperature. Magnesium chloride, with a decomposition voltage of 2.6 V at 700 °C is the most easily decomposed chloride among the Group 1 and 2 chlorides.

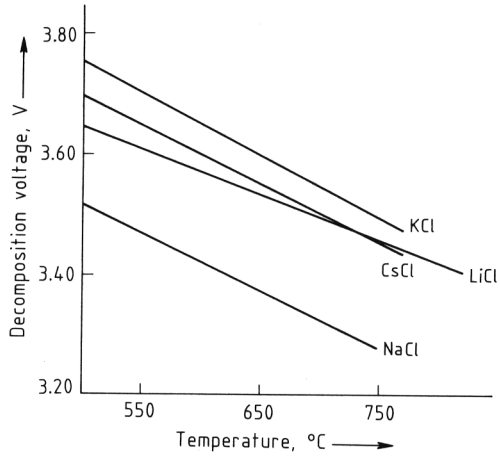

Figure 50. The decomposition potentials of the alkali-metal chlorides as functions of temperature [189]

The light alkali metals and alkaline-earth metals magnesium, lithium, sodium, and potassium have lower densities than the melts of their chlorides. Therefore, the electrodeposited metals ascend through and collect on top of the chloride melts. They must be prevented from coming into contact with the electrogenerated chlorine.

Interelectrodic gaps of several centimeters together with steel diaphragms between the electrodes and steel collectors that gather the ascending metal droplets above the cathode allow effective separation of chlorine and liquid metal – but at the expense of relatively high cell voltages and relatively low energy efficiencies. Nevertheless, metal dispersion and the solubility of metal in the melt are reasons for low current efficiency (Table 9). The trade-off between high current efficiency and low cell voltage results in energy

yields between 30 % and 40 %, which are typical for salt-melt electrolysis but which are low compared to large-scale electrolysis processes with aqueous solutions.

Table 9. Working temperatures, cell voltages, and current efficiencies in the production of light alkali metals and alkaline-earth metals

Cell operating characteristics	Metal		
	Lithium	Sodium	Magnesium
Working temperature, °C	400 – 420	590 – 600	660 – 670
Cell voltage, V	8 – 9	6.5 – 7	5.5 – 6
Current efficiency, %	85 – 90	90 – 95	78 – 80

For more details see → Lithium and Lithium Compounds, → Magnesium, → Potassium and Potassium Alloys, and → Sodium and Sodium Alloys.

4.5. Electrochemical Processes in Nuclear Fuel Reprocessing

Oxide fuel light water reactors (LWR) are dominant for nuclear electricity production worldwide. For this type of fuel an aqueous medium process to recover the fissionable material is well established. The zirconium alloy-clad fuel elements containing uranium oxide – plutonium oxide fuel are cut up, and the uranium and plutonium are leached out of the cladding material with nitric acid. Afterwards the solution is purified by multiple extractions using *n*-tributylphosphate (TBP) diluted with kerosene (→ Nuclear Technology, Chap. 5.4.1.). This process, the Purex process, was first developed in the late 1940s and industrially used in the USA in the early 1950s.

Determined by the demand of the following processes for reenrichment and fuel fabrication to minimize shielding and to facilitate the handling of the refabricated fuel elements, it is a particular requirement for the reprocessing process to guarantee an extreme product purity and to keep the contents of radioactive contaminants as low as possible. The permissible contamination limit for the long lived fission product cesium, after reprocessing, is, for example, about 3 µg/kg and the residual content for plutonium is limited to about 15 µg/kg in the uranium product. Furthermore, the process calls for a high product yield of 99 %. For the extraction process about 99.5 % is achievable in a modern plant. On average, this corresponds to a content of about 1 mg/kg plutonium in the extraction raffinates.

Another requirement of commercial reprocessing is minimization of the radioactive waste volume in order to reduce costs. The costs of radioactive waste treatment and storage are incomparably higher than the actual costs for conventional chemical wastes. The major part, more than 50 % of reprocessing costs are capital costs. This is the reason why current R&D activities are directed to a cost reduction by a simplification of the process.

Figure 51 shows a schematic diagram of the Purex process (see also → Nuclear Technology, Chap. 5.4.1.7.). Five extraction cycles are used to meet the product specifications. In the first high-active cycle the major portion of fission products is separated from uranium and plutonium. A radioactivity portion of about 10^{-5} of the 'nonproblematic' fission products and actinides (alkali, alkaline-earth, rare earth, transplutonium elements, etc.) leave this cycle with the products uranium and plutonium. The products are separated later by reducing tetravalent plutonium to Pu(III) which has a very low distribution coefficient and is backextracted into an aqueous phase. Between each cycle plutonium has to be reoxidized and hydrazine, which is necessary to stabilize plutonium in the trivalent state, must be destroyed by oxidation.

One reason for incomplete separation is the degradation of the solvent by radiolysis and hydrolysis [190]. Some fission products and plutonium form complexes with the TBP degradation products. Thus the degradation products must be removed by an alkaline wash.

Both laboratory and industrial experiences with the conventionally designed high-active first Purex cycle show that the product purity specifications cannot be met, even for the 'nonproblematic' fission products [191]. Therefore, two additional refining cycles are used for each product, uranium and plutonium (see Fig. 51).

4.5.1. Electroredox Separation Processes

In order to separate uranium from plutonium, Pu(IV) is reduced to virtually inex-

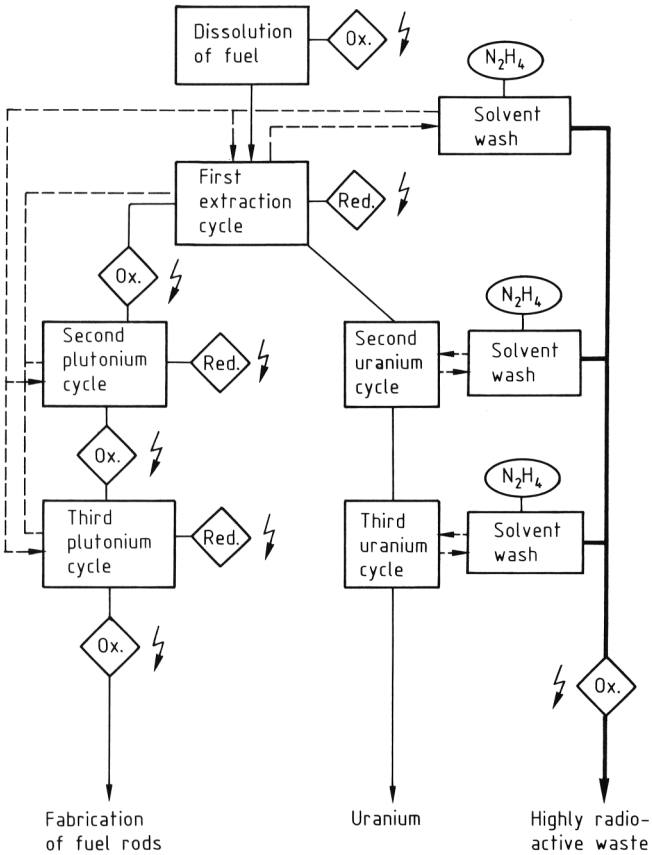

Figure 51. Schematic diagram of the PUREX process
The jagged arrow indicates an electrochemical process.

tractable Pu(III). For the reduction of plutonium, four processes have been developed to technical maturity with the following four reductants: iron(II)sulfamate [192], uranium(IV)nitrate [193], hydroxylammonium nitrate [194], and electroreduction within the extractor [195]. Generally U(IV) is added as a reducing agent during extraction according to the following equation

$$2\,Pu^{4+} + U^{4+} + 2\,H_2O \longrightarrow 2\,Pu^{3+} + UO_2^{2+} + 4\,H^+$$

The external feed U(IV) process, which is usually applied in current industrial reprocessing has at least two drawbacks: (a) demand for a high U(IV) excess, and (b) occasionally failure of the process caused by the autocatalytic reoxidation of Pu(III) and U(IV) starting in the organic phase.

These facts initiated activities in different countries to develop the in situ *electroreduction* process [196, 197, 198, 199]. A comparison of the chemical process with the electrochemical process was made in [200] and resulted in a preference for the electroreduction process. Compared with chemical separation processes for Pu, electroreduction is simpler from both engineering and operational points of view. The process has all the advantages of the externally fed U(IV) process but avoids its drawbacks.

Because of its simplicity and compactness, also *electrooxidation* has significant advantages when compared to the chemical oxidation process.

Titanium, platinum, and hafnium show sufficiently low corrosion rates to guarantee a long life-time of the electroredox equipment. Since the first electroreduction experiments performed

at Forschungszentrum Karlsruhe, efforts to design simple, compact and reliable electroredox equipment for industrial applications led to the following three design criteria [195]: (a) no diaphragm, (b) casing of apparatus used as cathode, and (c) operation at constant current or voltage.

Basic chemical investigations and corrosion tests of different materials showed that these criteria can be fulfilled for both the electroreduction extractors and the electrooxidation cell [195], [201]. Diaphragms are not necessary because of the irreversible character of the U(VI) electroreduction in the used electrolyte [202] and of the hydrazine electrooxidation. Titanium was found to be well suited as the container and cathode material, having corrosion rates of less than 50 µm per year under process conditions. Experiments with constant current or voltage operation showed that side-reactions such as hydrogen formation and nitrate reduction can be suppressed to a tolerable level, so that reference electrodes are not required for constant voltage operation.

4.5.2. Electroreduction Processes

In the electroreduction processes plutonium is reduced by electrolysis at the cathode

$$Pu^{4+} + e^- \longrightarrow Pu^{3+} \quad 0.92\,V$$

and by U(IV) generated in-situ.

The local uranium : plutonium ratio in the extractor determines which reaction is dominant. The Pu(III) is reoxidized in both phases by nitrite in an autocatalytic reaction [205]:

$$Pu^{3+} + HNO_2 + H^+ \longrightarrow Pu^{4+} + NO + H_2O$$

For this reason, the aqueous phase is stabilized with hydrazine, which reacts with nitrous acid

$$N_2H_5^+ + HNO_2 \longrightarrow HN_3 + 2\,H_2O + H^+$$
$$N_2H_5^+ + 2\,HNO_2 \longrightarrow N_2 + N_2O + 3\,H_2O + H^+$$

Since hydrazine is not extractable, a part of the Pu(III) is oxidized by organic nitrite, which results in an excessive U(IV) consumption, [214].

$$N_2 + 5\,H^+ + 4\,e^- - N_2H_5^+ \quad -0.23\,V$$

Hydrazine itself slowly reduces Pu(IV):

$$4\,Pu^{4+} + N_2H_5^+ \longrightarrow 4\,Pu^{3+} + N_2 + 5\,H^+$$
$$2\,Pu^{4+} + 2\,N_2H_5^+ \longrightarrow 2\,Pu^{3+} + 2\,NH_4^+ + N_2 + 2\,H^+$$

The last reaction is catalyzed by platinum, which is available because of corrosion of the anode [212].

Mixer-Settler. Figure 52 shows schematically the construction of the electroredox mixer-settler developed in Karlsruhe. The mixer-settler is made of titanium and the containment acts as cathode. The anodes are installed in the settler chambers where the electroreduction takes place. The most stable anode material is platinum with corrosion rates of a few micrometers per year. For industrial electroredox equipment, platinized Ta, Ti or other materials are used.

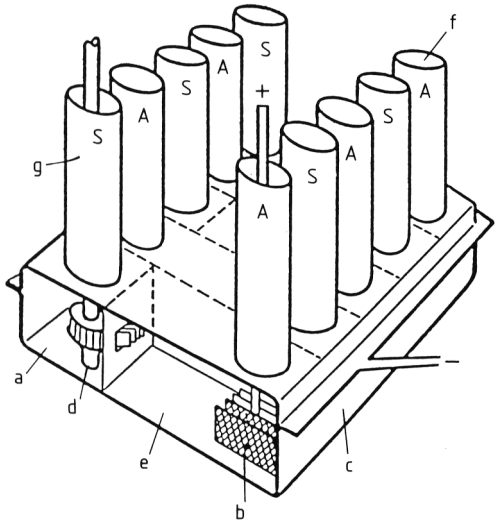

Figure 52. Electroreduction mixer – settler (1 B unit) at Karlsruhe Reprocessing Plant
a) Mixing chamber; b) Anode; c) Cathode; d) Stirrer; e) Settler; f) Anode stack (A); g) Stirrer stack (S)

Table 10 summarizes the results of countercurrent experiments with the electroreduction mixer-settlers, called EMMA [195], [204], [205]. The 16-stages MILLI-EMMA was used for uranium-plutonium separation, as well as for plutonium purification, but without a uranium scrub in the latter case. Because of the miniature size of the apparatus, the extraction

Table 10. Results of test runs with electroreduction mixer-settlers

	Purpose of test	Flow ratio org./aq.	Pu product concentration, g/L	Separation		Voltage, V
				Pu-DF[a]	U-DF[a]	
MILLI-EMMA	U/Pu split FBR model fuel[c]	5 – 6.7	≤ 22	≤ 4500[b]	≤ 3000[b]	2.8 – 8.0
	U/Pu split LWR model fuel	7.1 – 9.3	≤ 4.8	≤ 2000[b]	≤ 3800[b]	2.8 – 8.0
	Pu purification cycle	2.0 – 3.5	≤ 46	$\leq 20\,000$[d]		≤ 6.0
WAK-2B-EMMA	Pu purification cycle	2.0 – 3.0	≤ 36	$\leq 100\,000$[e]		2.0 – 8.0

[a] DF = decontamination factor. [b] 9 practical stages for Pu strip, 7 for U scrub. [c] 20 vol % TBP, 30 vol % for all other experiments. [d] 16 practical stages used for Pu strip, without U scrub. [e] 12 practical stages used for Pu strip, without U scrub.

stage efficiency was low, especially in experiments with a high organic to aqueous flow ratio. The major parameter variations were: (a) TBP concentration, (b) plutonium content, (c) current density, (d) HNO_3 concentration in the aqueous strip solution, (e) dibutylphosphoric acid, $(C_4H_9O)_2POOH \cdot$ (HDBP) concentration in the organic solvent, (f) flow ratios, and (g) total volume flow rate (residence time).

The most significant effect on the U/Pu separation was due to variations of the total volume flow rate. A decrease in the total volume flow rate causes complex interactions in the countercurrent extractor due to the increase in residence times for both phases. In the case of the miniature mixer-settler, the increase in extraction stage efficiency seems to be the dominant factor for separation improvement. In the Pu purification experiments an upper limit for the organic to aqueous flow ratio of 3 : 1 was found at Pu product concentrations of about 40 g/L.

The first industrial application was performed in the Pu purification cycle of the Wiederaufarbeitungsanlage Karlsruhe (WAK). The plant was equipped with an electroreduction mixer-settler (WAK-2B-EMMA, 12 stages) and an electrooxidation cell (WAK-2B-ROXI) in 1980 [205], [206]. In addition, in 1984, the first highly active cycle of WAK was equipped with an electroreduction mixer-settler.

In the WAK-2B-EMMA the flow ratio could not be raised above 3 : 1, corresponding to a Pu product concentration of about 36 g/L, although this apparatus has a significantly higher stage efficiency. One reason for the limitation of the Pu decontamination factor (DF) to a maximum of 10^5 are the very long residence times (volume flow limitations for technical reasons) [205], [206].

Pulsed Column. The extractor type preferred today for industrial reprocessing plants are pulsed sieve plate columns (Fig. 53). The main reason for this is the less complicated design of a critically safe pulsed column in comparison to a mixer-settler bank, at least for the U/Pu separation step. The pulsed column is constructed in a similar way to the mixer-settler. The column and the sieve plates, made of titanium, function as the cathode, while the platinized central rod is the anode. In the top decanter of the column, separator shields are installed to promote the separation of electrolytic gas from the organic liquid. The ratio of cathode to anode areas is made as large as possible to reduce anodic reoxidation of Pu(III). To eliminate any risk of a hydrogen explosion, the gas volumes of the mixer-settler and of the top decanter of the pulsed column are sparged with air to dilute the hydrogen below the explosive limit. The hydrogen current yield can be kept below 10 % under typical conditions.

The first experiments with the electroreduction column ELKE done in the plutonium test facility PUTE under the conditions of a Pu purification cycle yielded decontamination factors of up to 10^6 depending on the residual uranium content, flow ratio, nitric acid concentration, etc. [207]. The effective column length was about 8 m. The following experiments were carried out to study the electroreduction feasibility in the first extraction cycle, i.e., for U/Pu separation. Based on the results achieved in the purification cycle a comparable outcome was expected for separation. However, the experiments showed a

surprisingly poor separation efficiency: decontamination factors of only 200 to 2000 were achieved [208]. This was due to the fact that in the middle region of the column unusually large organic phase drops appeared with diameters of up to 20 mm and with inclusions of aqueous phase. This latter effect explained the observed insufficient U/Pu separation by heterogeneous contamination, in this case aqueous drops.

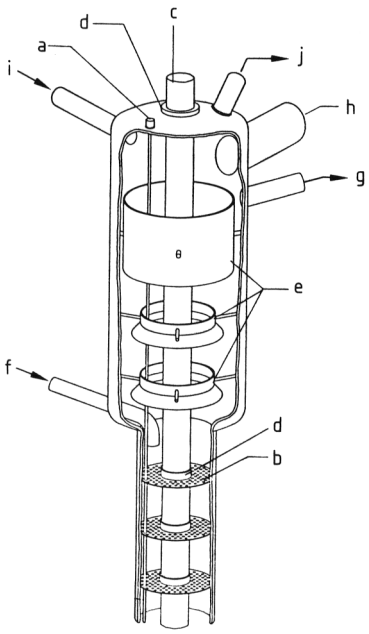

Figure 53. Pulsed electroreduction column
a) Cathode (outer casing); b) Cathode (perforated plates); c) Anode; d) Insulators; e) Gas – liquid separator plates; f) Aqueous-phase inlet; g) Organic-phase outlet; h) Overflow; i) Gas inlet; j) Off-gas outlet

The first way to reduce the amount of entrained aqueous phase transported with the organic phase drops is an increase in pulse energy [209]. The plutonium decontamination factor increases steeply when increasing the pulse energy $A \times f$ (amplitude × frequency) from 90 to 180 cm/min. Applying a pulse energy of $A \times f = 180$ cm/min small organic drops form a homogeneous dispersion without any inclusions of aqueous phase and a maximum Pu-DF of about 12 000 is achieved with a column length of only 5 m for the reductive Pu strip (BX part, see Fig. 51). To meet the final product specification for U, a Pu-DF of about 500 000 is necessary, however. To further increase separation, two consecutive columns were used in order to extend the reductive strip height and to minimize the entrainment effect by the introduction of an intermediate decantation. The results are summarized in Table 11 [209]. Row 2 shows that the desired Pu-DF can be reached with a flow ratio of 7.25 (which is usual in this process step for LWR fuel) by the use of two combined columns with 12.6 m total length.

Furthermore, it was also shown in the experiments, that in the case of a maloperation (interrupted hydrazine feed), no inadmissibly high Pu accumulation appears. The measured maximum plutonium concentration (first column) was always less than 20 g/L within 30 h after interruption in hydrazine supply.

In the electroreduction process only small amounts of the byproducts hydrazoic acid and ammonium are formed, thus reducing the hydrazine consumption. As was demonstrated in the PUTE facility [210], [211], electroreduction has the potential to be operated without the addition of hydrazine, if the nitric acid concentration is kept below 0.7 M within the whole extractor.

4.5.3. Electrooxidation

Between each cycle of the Purex process, plutonium has to be reoxidized to the tetravalent state and hydrazine must be destroyed. For this purpose an electrooxidation cell, called ROXI, was developed (Fig. 54). The anodic destruction of hydrazine is aided by the reactions of hydrazine with Pu(IV) formed by anodic oxidation and with HNO_2, which is cathodically produced in small amounts. The electrooxidation cell is a typical electrochemical flow reactor designed to minimize axial mixing. The electrolytic gas leaves the cell diluted with sparging air. The containment (cathode) is again made of titanium. The anode stacks are made of platinized metal sheets. The major advantage of electrooxidation compared to chemical oxidation with nitrogen oxides is the compactness of the apparatus. The chemical oxidation process requires one absorption column for the complete oxidation of hydrazine and Pu(III) and one additional column to desorb the excess of nitrogen oxides [203].

The yield of the direct anodic oxidation of Pu(III) depends on the nitric acid concentration. Under typical conditions, the major part of the

Table 11. Results of U/Pu separation with two electroreduction columns in series*

Volume flow ratio org./aq. in BX part	Pulsation I. column, cm/min	$A \times f$ 2. Column, cm/min	[Pu] in raffinate, mg/L	Pu-DF	U-DF
6.00	135	180	$\ll 0.01$	$\gg 73\,000$	143
7.25	135	180	0.0015	507 000	345
9.33	135	180	0.005	116 000	311

* Org. feed: 88 to 90 g U + Pu per liter, [U]/[Pu] = 100, $[HNO_3] = 0.2$ M, strip solution: $[HNO_3] = 0.1$ M, $[N_2H_5^+] = 0.05$ to 0.15 M, extraction height: BX part: 6.6 m + 6 m, BS part: 1 m, volume flow ratio in BS part: 1, cathodic current density: 0.5 to 0.8 mA/cm^2.

Pu(III) is oxidized autocatalytically by HNO_2. The accompanying U(IV) is consumed by reduction of Pu(IV).

The formation of Pu(VI) is suppressed at a HNO_3 concentration higher than 2 M. The formation of ammonium by cathodic reduction of HNO_3 depends on the electrolyte composition and was measured to be 0.03 to 0.46 mmol $A^{-1}\,h^{-1}$ for an electrolyte without hydrazine [212].

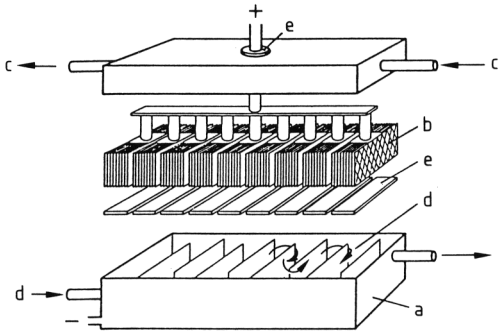

Figure 54. Electro-oxidation cell for plutonium(III) and hydrazine
a) Cathode and housing; b) Anode; c) Gas flow; d) Electrolyte flow; e) Insulation

The average values achieved with the two electrooxidation cells instal1ed in the PUTE facility are given in Table 12. For typical feed compositions, current consumption of about 100 A h per mole $N_2H_5^+$ were measured in the Ti-ROXI. For the Hf-ROXI, a higher current demand was measured, which is probably caused by a smal1er fraction of cathodic reaction supporting the hydrazine destruction. The ammonium formation is low but in some cases still higher than expected. This can be explained by the formation of ammonium in parallel to the electroreduction of HNO_3.

4.5.4. Corrosion

The corrosion of titanium under cathodic conditions depends on the electrolyte composition and on the current density. For the relevant and maximum HNO_3 concentration of 2.5 M, gravimetric experiments show a drastic decrease in corrosion in the presence of reducible cations; no definite effects of current density are detected in this case. Traces analysis in the product solutions confirm this effect for various pieces of equipment. Only a small increase in corrosion is observed at higher ROXI current densities. Thickness measurements of the PUTE-ROXI channel sheets show, after an operation period of 1950 h, comparably small corrosion rates of 23 µm/a even at current densities up to 135 mA/cm^2. The corrosion rate of hafnium, estimated from traces analysis, amounts to 7.7 µm/a at current densities of 50 mA/cm^2.

The anodic corrosion of platinum increases with decreasing nitric acid concentration and with increasing current density and temperature [213]. The corrosion rates obtained by gravimetric, trace and neutron activation analysis for typical electrolyte compositions and current densities amount to a few micrometers per year.

4.6. Electrochemical Water and Effluent Treatment

4.6.1. Cathodic Treatment

Wastewater from different industrial sectors, such as electroplating, photographic development, printed circuit board production, or battery technology, requires special treatment to remove toxic metal ions or to recycle valuable material. Due to increasingly stricter regulations for

Table 12. Average results of electrooxidation cells in PUTE

Apparatus	$N_2H_5^+$ current consumption, Ah/mol	Pu(VI) formation, % of Pu	NH_4^+ formation, mmol A^{-1} h^{-1}
Titanium-ROXI	≤ 100	≤ 2	≈ 1
Hafnium-ROXI	≤ 170	≤ 2	0.1 – 2
Hafnium-ROXI	≈ 150	~ 1	≈ 0.6

Table 13. Effluent limits and metal concentrations obtained by hydroxide precipitation at pH 8

Metal	Effluent limits, ppm				Metal concentration
	Germany (as of anuary 1983)	Germany (as of January 1991)	United States	Switzerland (1991)	after precipitation, ppm
Pb	2	0.5	0.5	0.5	21
Cd	0.5	0.2	0.3	0.1	1500
Cu	2	0.5	0.5	0.5	1
Ni	3	0.5	0.5	2	340
Hg	0.05	0.05		0.01	
Ag	2	0.1		0.1	
Zn	5	2	0.5	2	2.6
Sn	5	2		2	

the discharge of effluents, the permissible concentration of metal ions in effluents has been strongly decreased in recent years [218, 219, 220]. Table 13 shows the development of the limiting values for different metal ions in Germany from 1983 until 1991 and for comparison the latest values in Switzerland. The concentrations which can be obtained by the conventional technique of hydroxide precipitation are also listed in Table 13. Obviously, the limits for effluent discharge can only be met for a few metals such as Cu or Zn by the conventional technique. Here, the precipitated metal ions are replaced by alkali or alkaline-earth metal ions and in many cases the salt content prevents the recycling and reuse of the water. Furthermore, the cost of disposing sludge containing toxic metals has also been drastically increased during recent years [221]. Costly measures are also necessary to avoid the contamination of ground water with metal ions leached out by the rain.

As an alternative to hydroxide precipitation, ion exchangers are gaining in importance; however, ion exchangers are too costly for many types of effluents, and regenerable exchange resins are not available for all metals. As a result, there is incentive to develop electrochemical processes for pollution control [222, 223, 224, 225, 226, 227, 228, 229, 230].

4.6.1.1. Optimization of Cell Design

Since most of the metal ions can be removed by cathodic deposition, electrochemical processes have been developed, some of which are already commercialized and now being used in the industry [222], [225], [227], [228]. The removal of metal ions Me^{z+} from wastewater is based on cathodic metal deposition:

$$Me_{sol}^{z+} + ze^- \leftrightarrow Me$$

From a thermodynamic point of view, the Nernst equation predicts that it should be possible to decrease the Me^{z+} concentration in solution to an arbitrarily low value, if the potential E of the Me/Me^{z+} electrode is maintained sufficiently negative with respect to the standard potential $E^{o}_{Me/Me^{z+}}$.

$$c_{Me^{z+}} = c^{+} \exp\left[\frac{zF}{RT}\left(E - E^{o}_{Me/Me^{z+}}\right)\right] \quad (23)$$

However, at extremely low concentrations, the rate of the mass transport controlled process strongly decreases. In practice, electrolysis of metals at concentrations below 0.05 ppm (mL/m^3) is no longer economical due to increasing electrolysis time and unacceptably low space — time yield. Since the current densities at low metal ion concentrations are small, the specific energy demand for an electrochemical wastewater purification process is generally

quite low, typically in the order of 0.025 Euro/m^3 [218]. More relevant are the specific investment costs, which are inversely proportional to the space — time yield *STY* of the reactor. For the cathodic metal deposition in an electrochemical reactor of volume V, the space — time yield is defined as the amount of metal dm deposited in a time interval dt:

$$STY = \frac{1}{V}\frac{\mathrm{d}m}{\mathrm{d}t} \tag{24}$$

According to Faraday's law, dm is proportional to the electrolysis charge $A\ \phi^{\mathrm{e}}\ j\ \mathrm{d}t$:

$$\mathrm{d}m = A\phi i \mathrm{d}t \frac{M}{zF} e \tag{25}$$

The optimal process condition are met when the rate of the heterogeneous reaction attains its maximum at the limiting diffusion current density j_{l}. For a given metal ion concentration $c_{\mathrm{Me}^{z+}}$ a high mass transfer coefficient k_{m} and a large specific electrode area are essential to obtain high space — time yields (see Section 3.5).

The reactor performance is not independent of the wastewater properties. Therefore, a normalized space velocity $\varrho_{\mathrm{s}}^{\mathrm{n}}$ was introduced for the characterization of the reactor performance for a given degree of conversion [104], [231]:

$$\varrho_{\mathrm{s}}^{\mathrm{n}} = \frac{I\phi e}{(c_{\mathrm{i}} - c_{\mathrm{e}})\, VzF} \log \frac{c_{\mathrm{i}}}{c_{\mathrm{e}}} \tag{26}$$

This figure gives the volume of wastewater in cubic centimeters treated by reducing the inlet concentration c_{i} by a factor of 10 (c_{e} = 0.1 c_{i}) within 1 s in a reactor volume of 1 cm^3.

4.6.1.2. Electrochemical Reactors and Their Applications

Different types of cell constructions have been designed during recent years. Efficient cell design has been directed towards optimizing the space – time yield, using a high specific electrode area and/or a large mass transport coefficient. With respect to these criteria, the electrochemical cells may be classified into the following three groups [218]:

Improved mass transport and thus increased current density by setting the electrodes in motion or by applying turbulence promoters, but with a relatively small electrode area in a given cell volume. Examples are the pump cell [116], [117] (see Section 3.5) the Chemelec cell [232], the ECO cell [233], [234], the beat rod cell [235], [236] and cells with vibrating electrodes or electrolytes [237].

Attempts to accommodate *large electrode area* in a small cell volume resulted in developments such as the multiple cathode cell [238], the Swiss roll [239], or the extended surface electrolysis (ESE) cells [240].

Improved mass transfer coefficients and enlarged specific electrode area are provided by the use of three-dimensional electrodes. Examples are the porous flow-through cell [241], the RETEC cell [242], the packed-bed cell [243,244,245] the fluidized bed cell [246,247,248,249,250], and the rolling tube cell [251], [252]. Due to the specific fluid dynamic conditions inside these three-dimensional bed electrodes, they provide not only large specific electrode areas but also high mass transport coefficients.

Chemelec Cell. The Chemelec cell (manufactured by Bewt Water Engineers Ltd., Alcaster, UK), depicted schematically in Figure 55 uses a fluidized bed of glass spheres as turbulence promoters to improve the mass transfer to the electrodes consisting of a series of closely spaced gauze or expanded metal sheet electrodes. [232], [253]. Since the electrolyte flow velocity must exceed the minimum fluidization velocity, the residence time and the degree of conversion per pass are limited. Therefore, this cell is suitable for pretreatment or recycling operation and is commonly used in the electroplating industry for maintaining a moderate metal ion concentration of approx. 50 ppm in a recirculated wash water of a rinsing tank.

Cells with Rotating Cylinder Electrodes. High mass transport coefficients are also achieved in cells with rotating cylinder electrodes (RCE), and a small gap between the cylindrical cathode and anode. High rates of mass transport are experienced in the turbulent flow regime, so that RCE reactors allow metal deposition processes to take place at high speeds, even from dilute solutions. RCE reactors have been operated at diameters from 5 – 100 cm, with rotation speeds from 100 – 1500 rpm and currents

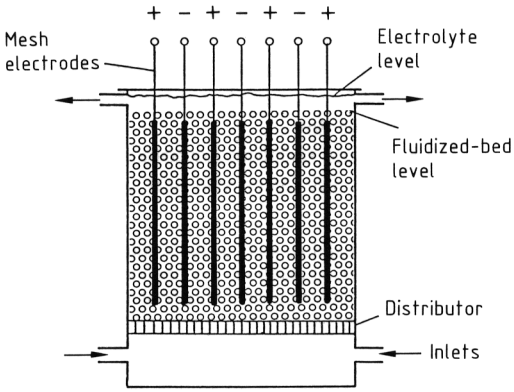

Figure 55. The Chemelec cell, with inert fluidized bed

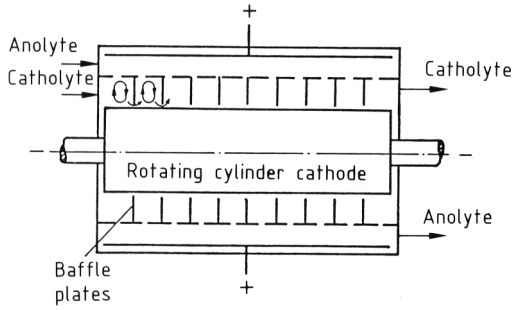

Figure 56. The ECO-cell cascade

from 1 A to 10 kA [254]. It is possible to design RCE reactors with scrapping devices to remove the deposited metal continuously in powder form from the cathode surface.

The *ECO cell* (Fig. 56) designed in the U.K., has already been produced and applied industrially. The cell usually has a segmented elctrolyte chamber (Fig. 56) to improve the residence time distribution by cascading. It has been used to recover silver and copper from wastewaters. The copper concentration in an 8-m^3/h effluent stream, passed over a 1.7-m^2 electrode, is reduced from 100 ppm to 2 ppm with a current of 1000 A at a current efficiency of 65 %. In this case, the power consumption per 1 m^3 of wastewater is 4.5 kW h, including the electrical energy necessary for electrode rotation.

Beat Rod Cell. The *beat rod cell* [251], [252], which is mainly used in small electroplating shops, is shown in Figure 57. Metal rods mounted into a rotating cylindrical carrier act as cathodes. Because of the rotation of the carrier, these rods beat against one another, causing mechanical release of metal, which can be collected as a powder at the bottom of the cell. Such a cell can reduce the silver content of a cyanide solution from 7.05 g/L to 3.5 mg/L in a batch operation mode.

Swiss Roll Cell. An efficient way of fitting a large planar electrode area in a small volume is the design of the Swiss roll cell, (see Fig. 29, Section 3.5). When the cell is filled with deposited metal, it is regenerated by treatment with acid, similar to common ion exchanger resins. Large space – time yields are realized and mass transport conditions are favorable since the separator functions as a turbulence promoter. The capacity of the cell can be further increased by using perforated electrodes, thus achieving a truly three-dimensional electrode structure. Mesh electrodes wrapped around a perforated winding core, lead to another version of the Swiss roll cell with radial flow.

ESE Cells. Another design which uses mesh electrodes instead of foils is known as ESE cells. These cells have been successfully applied on an industrial scale for the purification of copper containing effluents. A quite efficient commercial system for wastewater treatment is the *RETEC cell* (Eltech System Corp., Cardon, Ohio). The design is similar to a simple tank house cell, but contains 6 – 50 three-dimensional cathodes constructed as flow-through metal sponge electrodes [242]. These cathodes have an active surface area of approximately 15 times their geometrical area. Between each pair of cathodes is placed an inert, dimensionally stable anode (DAS) which is an oxide-coated titanium mesh. The cell is used in closed water recycling systems in the electroplating industry with applications similar to those of the Chemelec cell.

Packed-bed cells with three-dimensional electrodes are obtained when using electrodes consisting of electron conducting particulate material through which the electrolyte can flow. Numerous versions of these porous cells have

Table 14. Applications of the enViro cell [219]

Application	Metal	Throughput, m^3/h	Concentration, ppm Inlet	Concentration, ppm Outlet	Energy consumption, $kW\,h/m^3$	Anode area, m^2
Production of measuring instruments	Hg	0.3	300	0.05	1.2	1
Film processing	Ag	0.2	15	1.0	0.15	1
Salt production	Pb	0.5	2	0.1	0.07	1
Electroplating	Cd	0.2	20	1.0	0.18	1
Battery production	Hg/Cd	0.08	500	0.01	1.7	3
Cellulose acetate production	Cu	20	20	1.9	0.08	40
Pickling (recycling of solution)	Cu	3	150	50	0.19	5
Dye production	Cu	6	400	2.0	4.0	90
Dye production	Hg	2	4	0.05	2.5	15

been described in the literature [227], [228], [241], [255,256,257] (see Section 3.6. In a three-dimensional electrode the penetration depth of the current is restricted in the direction parallel to current flow. According to Equation (22), see Section 3.6, the penetration depth of the limiting current density increases as the concentration, which is proportional to j_l, decreases. The application of this principle led to the *enViro cell.* shown in Figure 58, which can operate satisfactorily at low metal ion concentrations. For example, in a single-pass operation, the concentration can be reduced from 10 – 50 mL/m^3 to less than 1 mL/m^3. This cell has a high space–time yield at a residence time of a few minutes only. The metal ion concentration can be reduced by up to a factor of 1/1000 in a single pass. The enViro cell has found wide spread industrial application (see Table 14).

Fluidized-Bed Electrolysis. The principle of fluidized bed electrolysis originates from FLEISCHMANN and GOODRIDGE [247]. The electrolyte flows from bottom to top through a loose bed of particles (Fig. 59). The advantage is that the particles are held in suspension, thus avoiding bed blockage as found in fixed-bed cathodes and allowing continuous metal recovery. The fluidized particles are cathodically charged via a feeder electrode. The growing particles are gathering at the lower part of the cell where they can be collected and replaced by fresh particles from the top. Because the height of the cell is restricted to about 2 m for hydraulic reasons and the fluidization velocity has to be maintained at a relatively high level, short residence times and thus only limited degree of conversion per pass can be achieved. Continuous operation with solution recirculation and cascade arrangements of cells are therefore used in practical applications. Fluidized-bed electrolysis has been examined for different applications in many laboratories, [133], [134], [155], [246], [255], [256] and has also been applied on an industrial scale for metal recovery [249], [250]. A typical example is the removal of copper from the process liquor in cellulose acetate production for membrane fabrication of artificial kidneys [218].

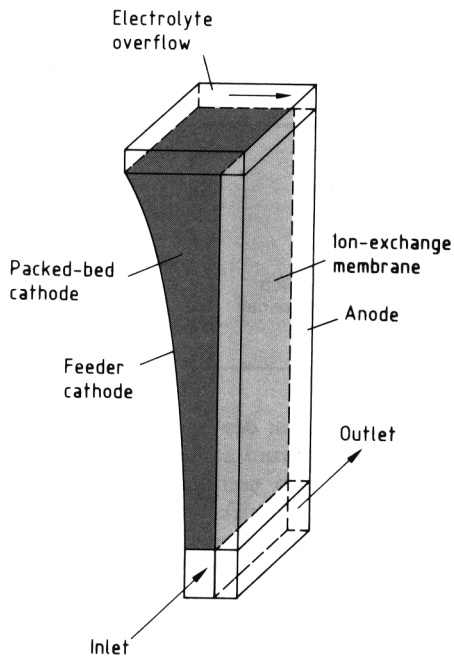

Figure 58. The enViro cell

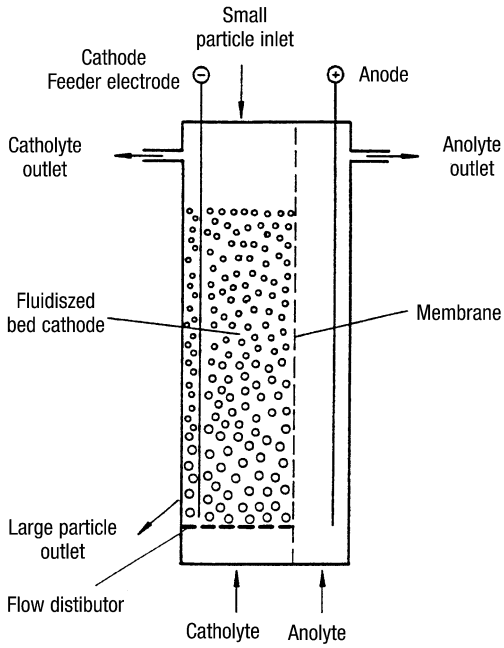

Figure 59. Fluidized-bed electrolysis according to FLEISCHMANN and GOODRIDGE

Moving and Circulating Bed Electrodes [257]. Figure 60 shows an example of the *rolling tube cell* [251], [252]. In principle this cell resembles the one used in the well-known plating process for piece-goods, using slowly rotating barrels. The rotating drum is only partially filled with graphite particles to achieve thorough agitation. The cathode has the shape of a perforated annular-type drum and is surrounded inside and outside by anodes to render the current distribution as homogeneous as possible. This cell is especially useful for silver and gold recovery. Further developments of moving bed cells and their application can be found in [258]. Utilization of the spouted bed, vortex bed, and moving bed electrodes for copper recovery from dilute solutions has been described in [257].

Cementation Processes. [259, 260, 261, 262]. An alternative to electrically driven cells using high surface area electrodes such as packed or fluidized-bed electrodes is a cementation system where electrochemical activity takes place spontaneously and a more noble metal may be deposited onto a less noble substrate which itself progressively dissolves [259]. Thus, no electrical energy is consumed in the precipitation reaction although power is required for electrolyte circulation. Well known examples of commercial cementation processes are the final purification step in leach liquor treatment prior to electrolysis in zinc electrowinning, where zinc dust is used as precipitant, and the recovery of silver from spent photographic solutions using steel substrates.

4.6.1.3. Operation Data of Electrochemical Cells

A summary of operation data of the different reactors and cell design described above is given in Table 15. Since these data are obtained from cells of different size and different process conditions, they give a rough guide rather than a reliable basis for assessing the individual reactors [218].

The data presented in Table 16, i.e., the specific electric energy demand and the normalized space velocity, are closely related to the eco-

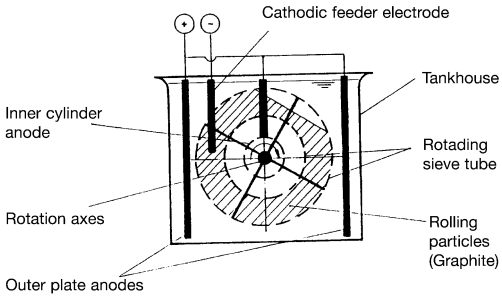

Figure 60. Rolling tube cell

Table 15. Operation data of different types of wastewater electrolysis cells [219]

Process	Scale	Removed metal	Concentration, ppm		V_r, L	V_D, L/h	j, A/cm^2	I, A	U_c, V	ϕ^e
			Inlet	Outlet						
ECO cell	Ind.	Cu	100	2	275	8000	1000	0.059	12	0.65
Beat rod cell	Ind.	Ag	6000	4	200	12.5	125		5	0.15
Swissroll cell	Lab.	Cu	380	25	0.3	0.7	0.55	0.46×10^{-3}	1.56	0.38
Porous flow-through cell	Lab.	Cu	800	0.2	2.6	0.64	0.45	3.5×10^{-3}	1.46	0.95
Packed-bed cell	Ind.	Cu	50	0.1	4.8	50	3.2	0.76×10^{-3}	1.9	0.66
Fluidized-bed cell	Ind.	Cu	77	5	192	7000	600		3.1	0.71
Rolling tube cell	Ind.	Ag	81	0.4	40	6.9	15		6.05	0.005

Table 16. Specfic and total energy consumption and normalized space velocity of different electrolysis cells for industrial applications [219]

Process	E_s^e, kWh/m^3	E_s^t, kW h/m^3	v_s^n
ECO cell	1.5	3	20
Beat rod cell	50	60	0.2
Swiss roll cell	1.23		20
Porous flow-through cell	1.03		0.9
Packed-bed cell	0.12		28
Fluidized-bed cell	0.27		30
Rolling tube cell	14.2		0.4

nomic efficiency [218], [220]. The specific electric energy consumption E_s^e for the treatment of 1 m^3 of effluent is relevant for the calculation of the energy costs. For cells containing rotating parts, the total energy consumption E_s^t is also specified. Considering both the specific electrical energy consumption and the normalized space velocity of processes which have already found industrial application, the packed-bed and fluidized-bed electrolysis seem to perform best.

4.6.2. Electrodialysis

For desalination or removal of ionic pollutants from aqueous wastewaters, electrodialysis is also useful [263, 264, 265, 266, 267, 268, 269]. The scheme for recovery of nickel salts from electroplating rinsing waters is shown in Figure 61. An electrodialysis cell is constructed like a bipolar filter-press electrolyzer, but only the two end chambers of the stack are fitted with electrodes. The compartments are sepa-

rated from each other alternately by cation (c) and anion (a) exchange membranes. The separation is based on the migration of ions in the electric field established by the cell voltage between the end electrodes. The material flux of a species i is $t_i j / z_i F$. The alternating cation and anion exchange membranes cause the contaminated electrolyte to be depleted, while the concentration, in this case of nickel and counterions, is enhanced in the concentrate. This separation is due to the fact that the transport numbers are close to unity for anions in the anion exchange membrane and for cations in the cation exchange membrane:

free electrolyte $t_i^+ \approx t_i^-$
cation membrane $t_i^+ \gg t^-{}_i$
anion membrane $t_i^+ \ll t_i^-$

Industrial applications have been reported, especially for nickel recovery [95], [96].

Electrodialysis does not result in the recovery of solid metals, but concentrates salt solutions. Although the concentration requirements for effluent treatment cannot be met, the concentration can be maintained on a level suitable for recirculation. Electrodialysis is generally not applicable to mixed electrolytes since suitable ion selective membranes are not available.

4.6.3. Anodic Treatment

By an electrochemical treatment of wastes either a partial (reduction of toxicity) or a complete decomposition of the pollutants can be achieved. In case of organic materials complete decomposition means the oxidation to carbon dioxide and as consequence a relatively high energy consumption for large organic molecules. The attempts for an electrochemical oxidizing treatment of wastewater or wastes can be classified into two categories:

Direct oxidation at the anode
Indirect oxidation using appropriate anodically formed oxidants (Cl_2, hypochlorite, peroxide, ozone, Fenton's reagent, peroxodisulfate)

A special case of indirect oxidation – frequently called *mediated electrochemical oxidation* (MEO) – is the use of metal ions with high oxidation potential (e.g., Ag(II), Co(III)) which can be electrochemically generated in a closed cycle largely avoiding emissions.

4.6.3.1. Direct Oxidation at the Anode

A suitable anode for direct oxidation has to meet mainly two requirements: high oxygen overpotential and corrosion stability.

A model described in [270] for the course of the anodic oxidation of an organic molecule assumes three major steps: (S indicates surface) Discharge of water forming an adsorbed hydroxyl species:

$$S[\ ] + H_2O = S[OH] + H^+ + e^-$$

The adsorbed OH is the "activated state" of water in oxygen-transfer reactions to the organic molecule R:

$$S[OH] + R = S[\ ] + RO + H^+ + e^-$$

Coevolution of O_2 by oxidation of water diminishing the current efficiency:

$$S[OH] + H_2O = S[\ ] + O_2 + 3\,H^+ + 3\,e^-$$

Phenol and its derivatives are the most frequently investigated molecules of electrochemical oxidation studies. Different anode materials such as platinum as reference electrode [271], nickel, glassy graphite, and titanium-supported oxide electrodes (IrO_2, RuO_2, PbO_2, and SnO_2 [272], [273]) and Bi as well as Fe-doped PbO_2 [270] were tested. A high oxidation rate was achieved with lead dioxide anodes using a packed-bed cell [274], [275]. The oxidation of phenol was complete at optimum conditions of pH and temperature in these experiments; the main product was carbon dioxide, benzoquinone and maleic acid were identified as byproducts.

The oxidation of numerous aromatic compounds has been investigated and for industrial wastewater treatment the use of a pilot plant equipped with Ti/SnO_2 anodes is reported [276].

Other applications that have been studied are direct anodic oxidation of sugars, alcohols, distillery effluents, dyes, aromatics, etc. [277].

For the direct oxidation of SO_2 in solution a kinetic study has been carried out using Pt, Pd, Ru, Ir, Rh, Re, and Au as anode materials [279]. The DSA PdO_x/Ti showed the highest electrocatalytic activity in these experiments

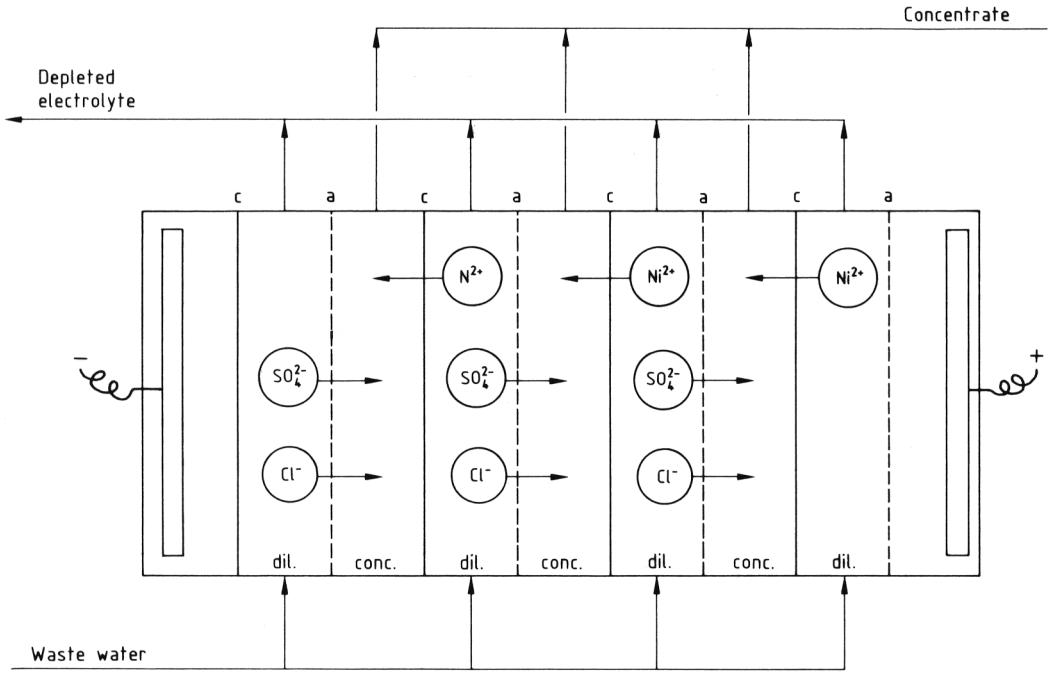

Figure 61. Nickel recovery from electroplating rinsing water by electrodialysis [267], [269]: a) Anion exchanger; c) Ccation exchanger

whilst $RuO_x - TiO_2/Ti$ and $IrO_x - TiO_2/Ti$ were inactive. A problem for the oxidation of SO_2 is to avoid the formation of elemental sulfur at the cathode.

The direct anodic oxidation of SO_2 as a method for flue gas treatment has been studied at several electrodes (see, e.g., [280]). At graphite electrodes sulfate and dithionate are formed as products. A continuous neutralization of the generated acid is required in the process. Numerous fundamental investigations and even bench scale experiments have been performed for the anodic treatment of aqueous effluents containing further inorganic pollutants, e.g., cyanide [281], thiocyanate [278], and sodium dithionite [282].

4.6.3.2. Indirect Oxidation

The most used oxidant electrochemically generated belonging to this category is chlorine or hypochlorite, ozone is the other prominent oxidant which can be produced electrochemically. In addition peroxide, Fenton's reagent [283], and peroxodisulfate are used occasionally. ABB Ltd. and OxyTech Ltd. use hollow, cylindrical fluorocarbon-impregnated carbon anodes for ozone generation in a solid-polymer electrolyte cell [277].

As already mentioned the use of redox couples with high oxidation potential such as Ag(II)/Ag(I) (1.98 V vs. NHE) and Co(III)/Co(II) (1.82 V vs. NHE) can completely oxidize organic compounds to CO_2. Therefore this *mediated indirect oxidation method* was developed for the treatment of hazardous organic wastes containing only a limited amount of water. The oxidant is continuously generated electrochemically in a closed cycle. The method was originally used in the nuclear industry to dissolve refractory plutonium dioxide in nitric acid [284], [285]. This experience initiated the treatment of mixed (radioactive) wastes (e. g., tributyl phosphate, solvent of the Purex Process) at the Savannah River Laboratory, USA and at AEA Dounreay, UK. The successful experiments launched process development activities at the Lawrence Livermore Laboratory [286], at AEA Dounreay [287], and at the Forschungszentrum Karlsruhe [288], and the method was ap-

plied to a plurality of hazardous wastes. The principle of this process is shown in Figure 62. Ag(II) is generated anodically from Ag(I) in an aqueous nitric acid solution using an electrochemical cell which is divided by a membrane. In the cathodic compartment nitric acid is reduced [289] to NO which can be regenerated to nitric acid by oxidative absorption in columns [290].

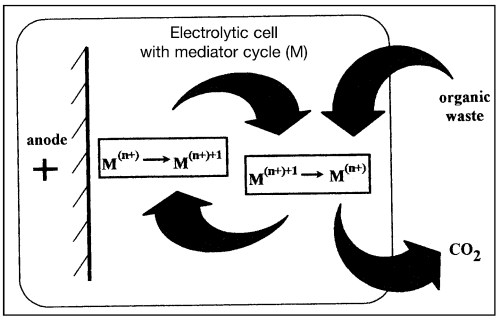

Figure 62. Principle of the MEO process

In this procedure, using two electrochemical cells, one for destruction of organics and the other for nitric acid recovery, hydrogen is the final cathodic reaction product. Investigations of Forschungszentrum Karlsruhe together with Eilenburger Elektrolyse- und Umwelttechnik have combined these two electrochemical cells into one, using silver containing nitric acid as anolyte and sulfuric acid as catholyte, thus avoiding NO formation at the cathode.

For the process development destruction experiments with different model substances and real wastes were performed both in a laboratory and in a bench-scale plant (Fig. 63). The experiments on a laboratory scale were caried out in a small filter press cell. Anode and cathode compartment were divided by a cation exchange (Nafion 450) membrane. Both electrode areas were 11 cm^2 and the volume of anode and cathode compartment was 6 cm^3 each [291]. The bench sale plant is designed similar to the laboratory plant with a scale up factor of 100. The central part of this plant consists of a monopolar electrochemical cell (ELECTROCELL AB, Sweden) or a bipolar cell from Eilenburger Elektrolyse- und Umwelttechnik GmbH (Germany). Both cells have anode and cathode areas of 0.12 m^2 each. A platinum foil anode is used, titanium or carbon serve as cathode. The maximum current density is limited by the Nafion 450 membrane to 5 kA/m^2 corresponding to a maximum current of 600 A. Liquid organic waste is fed into the anodic cycle by a pump, solid waste is fed into the anodic glass vessel from the top. Due to high flow rates of the electrolyte (up to 1500 L/h) the organic waste is finely dispersed in the electrolyte to minimize the transport inhibition at the aqueous – organic phase boundary layer [292].

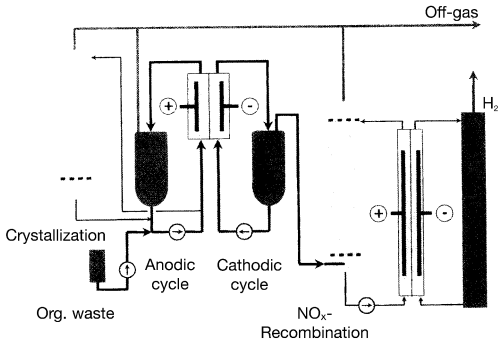

Figure 63. Bench scale MEO plant at Forschungszentrum Karlsruhe

In case of degradation of chlorinated organic compounds intermediate precipitates of silver chloride occur which are redissolved in presence of an excess of Ag(II) by oxidation of chloride to perchlorate, which is enriched in the anolyte and can be separated from time to time as potassium perchlorate by cooling crystallization [293].

In the laboratory plant the destruction of a large variety of different, mostly chlorinated, model substances were investigated. In all experiments the electrolyte consisted of 0.5 M $AgNO_3$/7 M HNO_3. The experiments were carried out with current densities between 1 and 5 kA/m^2 and reaction temperatures of 40 – 90 °C. All selected model substances with the exception of hexachlorobenzene could be destroyed to more than 99 %. As reaction products always CO_2 and small amounts of CO were formed.

Moreover, several organic multicomponent waste mixtures from the chemical industry were treated in the laboratory and bench scale plants. The main components were benzoyl chloride, chlorobenzoyl chlorides, chlorobenzaldehyde, benzotrichloride, and benzoic anhydride. The

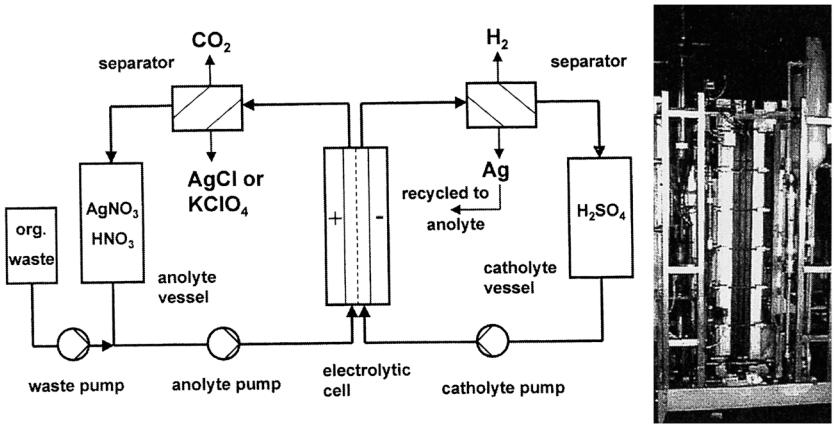

Figure 64. Mobile demonstration plant for destruction of chlorinated organic wastes

experiment was performed at 60 °C using a current density of 4 kA/m^2. At the start of the experiment CO_2 evolution occurred and current efficiencies up to 100 % were achieved. After 60 min CO_2 evolution decreased and O_2 formation, caused by water oxidation, increased.

Based on the described experiences and founded by the European Commission in the frame of a CRAFT project, a mobile demonstration plant was designed and built, which has been operated on an industrial site since March 1999 (Fig. 64). In case of destruction of chlorinated organic wastes, the plant offers both possibilities, separation of silver chloride from the anolyte and recovery of silver or oxidizing the organic chlorine to perchlorate and winning of potassium perchlorate. Sulfuric acid is used as catholyte, generating hydrogen at the cathode as a valuable byproduct. The migration of silver ions from the anolyte through the membrane into the catholyte was determined to be only 0.03 g $A^{-1}\,h^{-1}$. These silver ions are reduced and deposited at the cathode, flushed out periodically, separated from the catholyte and recycled to the anolyte.

4.7. Electrochemical Gas Purification

4.7.1. General Aspects

An increasing demand for off-gas purification, particularly for smaller scale power plants, heating combustion units, or chemical plants, has encouraged the development of new concepts of electrochemical gas purification techniques. Many gaseous pollutants, such as chlorine, hydrogen sulfide, nitrous oxides, or sulfur dioxide permit electrochemical conversion in an aqueous environment, since the standard potentials of the corresponding reactions are all within the stability range of aqueous electrolytes:

$$2\,Cl^- \rightleftharpoons Cl_2 + 2\,e^- \quad E^o = 1.360\,V$$
$$H_2S \rightleftharpoons S + 2\,H^+ + 2\,e^- \quad E^o = 0.142\,V$$
$$NO + 2\,H_2O \rightleftharpoons NO_3^- + 4\,H^+ + 3\,e^- \quad E^o = 0.957\,V$$
$$NO_2 + H_2O \rightleftharpoons NO_3^- + 2\,H^+ + 3\,e^- \quad E^o = 0.775\,V$$
$$SO_2 + 2\,H_2O \rightleftharpoons SO_4^{2-} + 4\,H^+ + 2\,e^- \quad E^o = 0.138\,V$$

Different concepts for electrochemical gas purification can be found in the literature [219]. In any case, the initial step is absorption of the pollutant species into a liquid phase. Since the solubility of several gases in aqueous solutions is too low, the transfer from the gas phase into the aqueous phase must be supported by a reaction, which converts the primarily dissolved species permanently to a more soluble one. In the simplest case, this can occur either by direct conversion at the electrode of an electrochemical cell [293, 294, 295], or indirectly via chemical reaction with a redox mediator, which can be electrochemically regenerated in a subsequent step. Homogeneous and heterogeneous catalysts have been applied as redox mediators. The advantage of heterogeneous mediators, such as oxides, is that a separation of the reaction products from

the mediator is not necessary. Both, the electrochemical conversion of the pollutant and the electrochemical regeneration of the redox mediator can be achieved either by an inner cell or an outer cell process, as illustrated in Figure 65 [297].

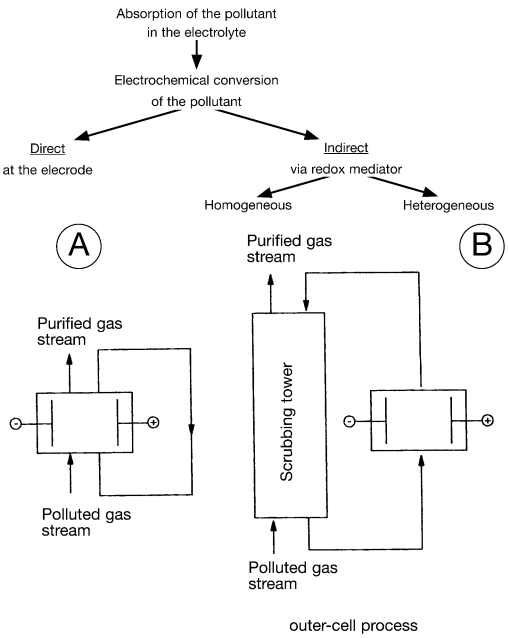

Figure 65. Electrochemical gas purification by inner cell and outer cell processes

Direct electrochemical conversion of gases by an inner cell process can be carried out in a specially designed electrochemical absorption column, which consists of a three-dimensional packed bed electrode of conducting particles in contact with a cylindrical feeder electrode and a counterelectrode separated by a porous diaphragm or ion exchanger membrane [293, 294, 295].

Figure 66 shows schematically the construction of such a device, which usually operates under countercurrent flow conditions of gas and liquid phases. The dissolved pollutant is directly converted at the surface of the packed bed electrode. Such a device exhibits high space – time yield and has been successfully tested for electrochemical absorption of sulfur dioxide and chlorine [295], [298], [299].

Indirect electrochemical processes with *homogeneous redox mediator* are the so-called "Peracidox"process developed by Lurgi [300] using peroxodisulfate, and the modified "Mark 13 A" process developed at the Joint European Research Center Ispra [301], using bromine as mediator as the redox mediator for indirect SO_2 oxidation. In both cases the electrochemical regeneration of the redox mediator is performed by an outer cell process in a separate electrochemical cell.

Processes with *heterogeneous redox mediators* are the lead dioxide catalyzed oxidation of SO_2, suggested by Strafelda [302] and the combined electrochemical and catalytic removal of SO_2 at a Cu_2O/Cu^{2+} catalyst [303]. Porous gas diffusion electrodes have been employed for the conversion of SO_2 using pyrolyzed Co phthalocyanines as electrocatalyst [304]. Electrode technology adopted from the fuel cell technology have also been used for effluent gas treatment. A cell construction similar to that of a molten carbonate fuel cell with porous electrodes and a molten $Na_2S - Li_2S$ electrolyte at 350 °C has been examined for removal of SO_2 and H_2S [305, 306, 307, 308]. High-temperature solid-state electrochemical cells based on Y_2O_3-stabilised ZrO_2(YSZ) and porous palladium electrodes have been studied for the successful electrochemical removal of both NO and CH_4 at 650 – 750 °C in an oxidizing atmosphere [309]. NO is reduced to N_2 and CH_4 is oxidized to CO_x.

4.7.2. New Process Developments

Since flue gases not only contain SO_2 but to a certain extent also NO_x, the development of processes for the simultaneous removal of both components has been the subject of recent investigations [299], [310, 311, 312, 313].

The cerium(IV)-assisted process [310], [311] is an indirect outer cell process with Ce^{4+} as homogeneous redox mediator for the simultaneous oxidation of SO_2 and NO_x to sulfuric acid and nitric acid. The acidic solution is continuously fed down to the anode compartment

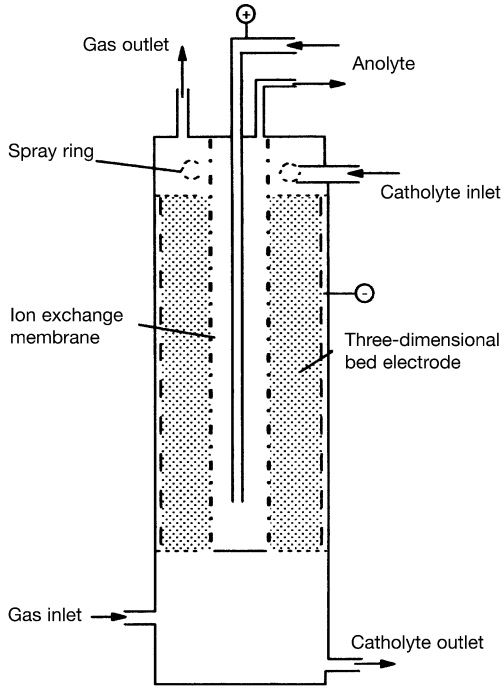

Figure 66. Scheme of an electrochemical absorption column with three-dimensional particle bed electrode

of an electrochemical cell where Ce^{4+} is regenerated. When the solution has reached a critical concentration of approx. 30 – 40 % HNO_3 and H_2SO_4, separation of the nitrate effluent from the sulfate solution and recycling of Ce^{3+} are performed in a separate unit, where the nitric acid is first separated by distillation and the excess of sulfuric acid is then continuously removed by liquid – liquid extraction [314], [315].

The lead dioxide dithionite process [299], [312], [313] combines direct and indirect conversion of SO_2 and NO_x, respectively, in two steps with dithionite as homogeneous redox mediator for the indirect reduction of NO_x and lead dioxide as heterogeneous catalyst and mediator for the direct oxidation of SO_2. The gas mixture first enters an absorption column, where NO is absorbed by the complexing agent Fe(II)EDTA and simultaneously reduced with dithionite. Dithionite is continuously regenerated by cathodic reduction of SO_3^{2-} in an electrochemical cell. The second component, SO_2, passes the NO absorption column without reaction and enters an electrochemical cell, for example as shown in Fig. 66, where it is oxidized directly to sulfuric acid at the anode. A pilot plant for the treatment of 100 m^3/h (STP) of flue gas with 600 ppm NO inlet concentration had been tested on an industrial site (Brite Project 2026).

Microkinetic studies of NO_x absorption revealed that the degree of NO conversion can be significantly improved, if Fe(II)EDTA is present in the absorption solution at a low content. Another interesting aspect of electrochemical promotion of environmentally important heterogeneous gas phase reactions is the so called NEMCA effect (non-faradaic electrochemical modification of catalytic activity). NEMCA significantly increases the catalytic activity and selectivity of metal and metal oxide catalysts deposited on solid electrolytes such as YSZ or β"-Al_2O_3 by electrochemical polarization of the catalyst electrode [316, 317, 318]. Technical ap-

plications are still in the initial stage of development.

4.8. Electrochemical Shaping

Shaping pieces of metal by controlled anodic dissolution – controlled with respect to dissolution depth in three dimensions – is the essential aspect of electrochemical metal shaping [319, 320, 321]. This technique includes electropolishing, electrolytic or electrochemical grinding, and electrochemical machining. The last, electrochemical machining, is the fastest, most vigorous form of spatially controlled anodic metal dissolution.

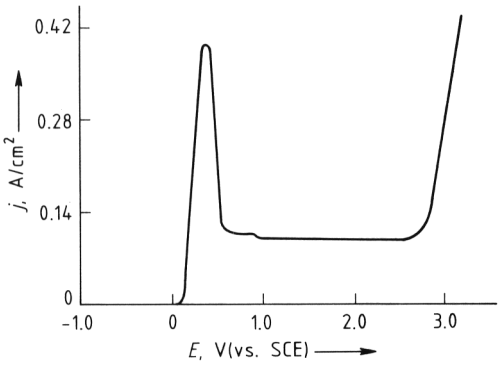

Figure 67. Voltammogram of nickel, a passivating metal, in 9 M HCl [320]

Anodic corrosion, the fundamental phenomenon behind these three techniques, is characterized by the shape of the current – voltage curve in Figure 67. Anodic dissolution, which proceeds at a sizable rate in the subpassive voltage region, increases with anodic polarization. Its rate drops sharply, by several orders of magnitude, at the passivation potential, at which a dense semiconducting oxide hydrate film forms, and prevents further dissolution. At much higher potentials, however, this film breaks down, and vigorous transpassive dissolution of the metal is observed [319], [322]. If a compound of low solubility precipitates on the surface as a more or less porous deposit, instead of a passivating film, then anodic dissolution is inhibited, but the current density remains several orders of magnitudes higher than in the case of passivation. In such cases anodic dissolution usually is limited by mass transfer of the corrosion products through the pores of the deposited layer.

This mass tranfer-controlled dissolution through layers of precipitates is the principle governing electropolishing (Fig. 68A), whereas transpassive high current density corrosion is used for electrochemical machining and electrochemical grinding (Fig. 68B and C). Electrochemical grinding, which combines anodic dissolution with limited mechanical grinding is advantageous whenever a component of the alloy is not dissolved electrochemically but sticks to the surface and must be removed continuously along with the base metal. This is the case for cemented carbides as well as nitrides and borides of transition metals or refractory metals.

For processing superalloys and refractory metals, which came into use in the aerospace industry and for the construction of steam and gas turbines in the 1960s and early 1970s, the electrochemical methods – electrochemical machining, electrochemical grinding, and electrochemical polishing – are especially suitable. These methods, with the exception of electropolishing, are not likely to replace mechanical processing of conventional metals and alloys because the electrochemical techniques are expensive and too sophisticated to be used for softer alloys.

Electropolishing. Figure 67 shows the current – voltage curve for anodic dissolution of nickel in 9-N HCl solution with a limiting current density of the order of 100 mA/cm^2. This mass transfer-controlled dissolution rate is caused by formation of a relatively thin layer of porous oxide hydrate. The metal cations that are formed by anodic dissolution can leave the metal surface only by diffusion through pores. Since there is good reason to believe that the pores are randomly distributed and formed and closed at random, dissolution should also be random with no preference for particular crystal faces. Consequently, etch patterns are not observed.

The polishing effect is attributed to higher local dissolution rates of the elevated parts of the initially rough surface because of the lesser thickness of the porous oxide hydrates there. Table 17 lists the conditions and electrolyte compositions for electropolishing of metals. All the electrolytes contain components or establish a pH that causes the formation of only moderatey soluble oxide hydrates or salt deposits.

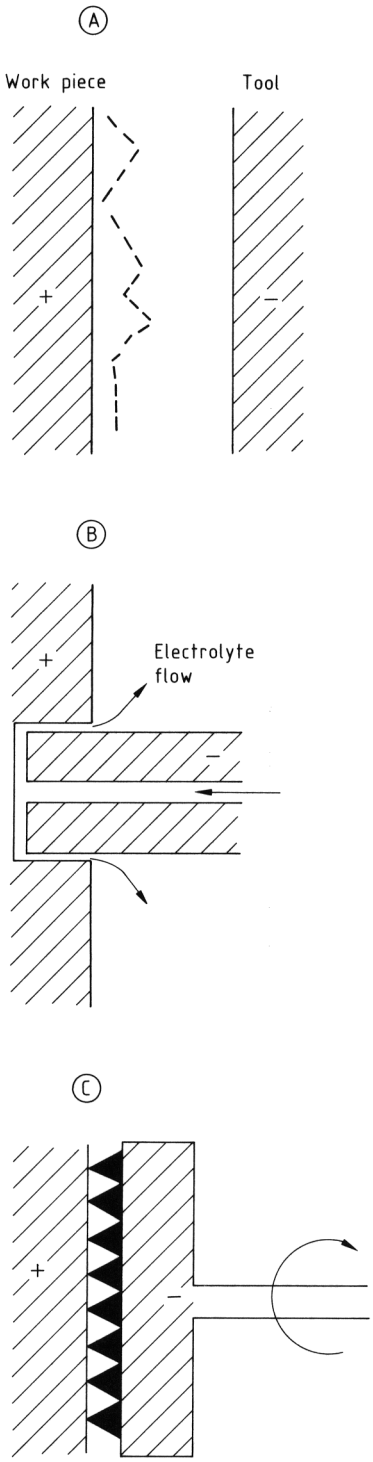

Figure 68. Schematic electrochemical shaping
A) Electrochemical polishing (– – – initial contour); B) Electrochemical machining; C) Electrochemical grinding

Such deposits form on the metal surface under steady-state conditions at current densities of ca. 300 – 3000 mA/cm^2.

Electrochemical Machining (ECM). For a practical electrochemical machining process, the metal must be removed at a cutting rate of 0.25 cm/min or more, corresponding to current densities > 160 A/cm^2. At these current densities, the electrolyte gap between the cathode and the workpiece must be adjusted continually to 0.005 – 0.13 cm to conserve electric energy. In addition, the anodically generated metal ions, hydroxides, and debris, the cathodically evolved hydrogen, and the dissipated Joule heat must be removed efficiently by vigorously pumping the electrolyte through the narrow gap between the cathode and workpiece. In order to keep ohmic voltage drop across the interelectrodic gap as low as possible, the electrolytes must have high conductivity.

Table 18. Electrolytes for electrochemical machining [320]

Metal	Electrolyte	Remarks
Aluminum and aluminum alloys	$NaNO_3$ (100 – 400 g/L)	excellent surface finish
Cobalt and cobalt alloys	$NaClO_3$ (100 – 600 g/L)	
Molybdenum	NaOH (40 – 100 g/L)	NaOH consumed and must be added continuously
Nickel and nickel alloys	$NaNO_3$ (100 – 400 g/L)	
	$NaClO_3$ (100 – 600 g/L)	good surface finish, good dimensional control, low metal removal rate
Titanium and titanium alloys	NaCl (180 g/L) + NaBr (60 g/L) + NaF (2.5 g/L)	good surface finish, good dimensional control, good machining rate
	$NaClO_3$ (100 – 600 g/L)	
Tungsten	NaOH (40 – 100 g/L)	NaOH is consumed and must be added continuously
Steel and iron alloys	$NaClO_3$ (100 – 600 g/L)	excellent dimensional control, brilliant surface finish, high metal removal rate, fire hazard when dry
	$NaClO_3$ (100 – 400 g/L) $NaNO_3$ (100 – 400 g/L)	

Commercial ECM machines apply a maximum voltage. This is distributed between the sum of the anodic and the cathodic overpoten-

Table 17. Electrolytes for electropolishing metals [320]

Metal	Electrolyte	Current density, mA/cm^2	Temperature, °C	Polishing time, min
Aluminum	20 wt % $HClO_4$ + 80 wt % acetic anhydride *	200 – 400	≤38	15
Copper	63 wt % H_3PO_4 + 37 wt % H_2O	20 – 50	22	5
	65 wt % H_3PO_4 + 15 wt % H_2SO_4 + 6 wt % CrO_3 + 14 wt % H_2O	200 – 250	22 – 25	2 – 5
Lead	30 vol % $HClO_4$ + 70 vol % glacial acetic acid *	1300 – 1600	22	5
Nickel	73 wt % H_2SO_4 + 27 wt % H_2O	1000 – 3000	22	2
	65 wt % H_3PO_4 + 15 wt % H_2SO_4 + 6 wt % CrO_3 + 14 wt % H_2O	200 – 250	22 – 25	2 – 5
Tin	20 vol % $HClO_4$ + 80 vol % acetic anhydride	600 – 1000	32	8 – 10
Tungsten	10 wt % $NaNO_3$ + 90 wt % H_2O	200 – 400	22	20 – 30
	Na_3PO_4 (160 g/L)	600	38 – 49	10
Zinc	25 wt % KOH + 75 wt % H_2O	1000	22	15
	17 wt % CrO_3 + 83 wt % H_2O	2000		
Brass	16 wt % CrO_3 + 84 wt % H_2O	2000		
	70 wt % H_3PO_4 + 30 wt % H_2O	500	26	5
Steel, carbon	48 wt % H_3PO_4 + 40 wt % H_2SO_4 + 12 wt % H_2O	3000	35 – 49	10
	10 wt % $HClO_4$ + 90 wt % glacial acetic acid	1500 – 2500	26	0.5 – 2
Steel, stainless	30 wt % H_3PO_4 + 60 wt % H_2SO_4 + 10 wt % H_2O	1800	49	2
	42 wt % H_3PO_4 + 45 wt % glycerol + 13 wt % H_2O	100 – 600	93 – 149	8 – 15

* Especially hazardous.

tials, which determines the current density and hence the dissolution rate, and the ohmic voltage drop across the interelectrodic gap. The ohmic voltage drop increases with gap width and current density. Therefore, for every cutting rate and voltage there is one gap width that is self-adjusting and that can be kept steady during the process, the so-called equilibrium gap.

Nevertheless, short circuits, which can severely damage the tool, must be avoided, and most ECM machines are protected against such damage by special sensing devices and ultrafast electronic switches.

Figure 69 shows how electrochemical machining works in the case of electrochemical drilling or sinking. The tool is a cylindrical cathode, the wall of which is covered with an electrically insulating coating. Current densities around 300 A/cm^2 flow between the bottom of the hole and the cathode, and a vigorous stream of electrolyte flushes the dissolved metal ions and solids out of the hole.

Immediately behind the cathode face there is a region where electropolishing takes place because of the reduced current densities. Consequently, electrochemical machining in most cases produces an excellent furnace finish.

Most important for successful electrochemical machining is the choice of the electrolyte. Its composition should allow metal dissolution with current densites up to several hundred amperes per square centimeter while preventing the precipitation of insoluble salts and oxides. Alkali-metal chlorates and perchlorates, which can be hazardous, have proved to be suitable ECM electrolytes. Table 18 lists the electrolyte and its concentration in typical ECM electrolytes.

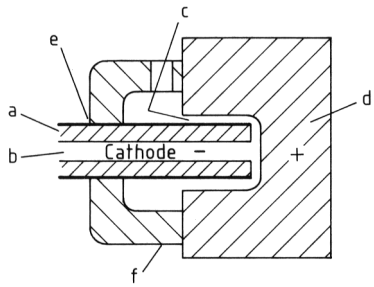

Figure 69. Schematic electrochemical sinking machine [320]
a) Cathodic tool; b) Electrolyte inlet; c) Electrolyte outlet; d) Anodic workpiece; e) Insulating coating; f) Shield

Electrochemical machining became popular for processing high-strength metals and alloys and refractory metals. Some of the procedures for which ECM can be used are shown schematically in Figure 70. Die sinking is also possible. Sinking the cathode tool into the workpiece produces the cavity used for casting.

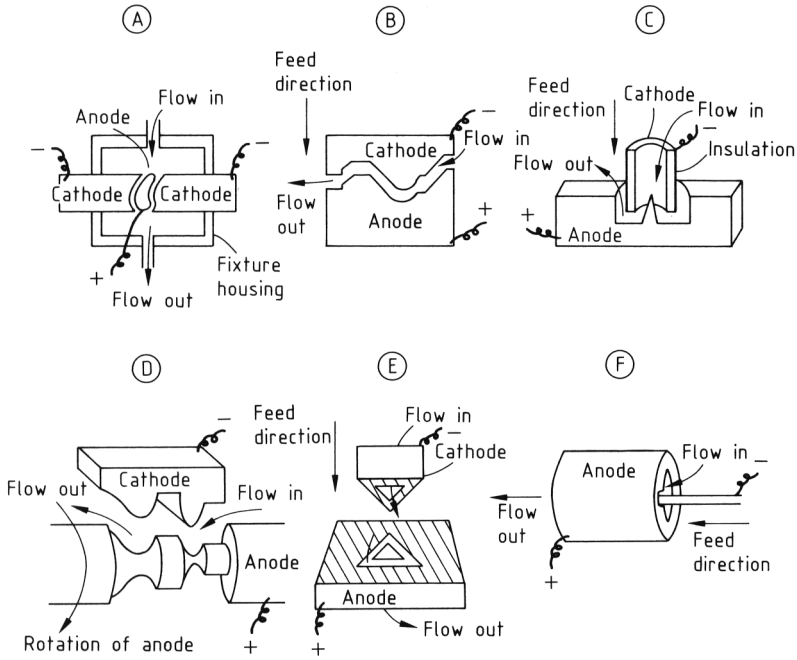

Figure 70. Sketches of various electrochemical machining operations [320]
A) External shaping; B) Cavity sinking; C) Plunge cutting; D) Turning; E) Trepanning; F) Internal grooving

In all these operations overcuts due to the action of stray currents cannot be avoided. Computed current-density distributions have always turned out to be insufficiently accurate to prevent the overcuts. Shaping is carried out by empirically changing the tool dimensions, current density, electrolyte concentration, electrolyte flow velocity, etc.

Electrochemical Grinding. Diamonds, α-alumina (corundum), or another hard material, the grinding powder, is embedded into the surface of a grinding wheel of copper or copper bronze, which serves as the cathode. The grinding powder is pressed into the metal surface and held in place by plating the studded surface with a layer of nickel. With an optimum load of 3.0 carats/cm^2, the grinding powder is exposed to a height of only 10 to 60 µm by etching the nickel away in an electropolishing bath.

The diamond (or corundum) grains do not serve themselves for grinding, but rather define and maintain the width of the electrolyte gap between the grinding wheel and the workpiece. Therefore, only light pressure is used to press the grinding wheel and the workpiece together. Around 90 % of the metal is removed by anodic dissolution. Fully hardened steels can be processed without difficulty and without detectable wear of the tools because nearly 100 % of the metal is removed by anodic dissolution. Current densities of 12 – 40 A/cm^2 approach those used for ECM; mixtures of chlorides and nitrates serve as electrolytes. Vigorous pumping and copious electrolyte exchange is vital for electrochemical grinding.

Unlike mechanical grinding, electrochemical grinding leaves no scratches, grooves, or tool registry marks. Current densities must not be too high if increased surface roughness is to be avoided.

5. Organic Electrochemistry

5.1. Introduction

Organic electrochemistry [323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354] provides a very versatile and powerful means for the

selective reagent-free reduction or oxidation of organic compounds. Organic substrates are activated by addition (reduction) or removal (oxidation) of electrons by a heterogeneous electron transfer reaction with the electrode. Starting from neutral compounds, either radical anions (reduction) or radical cations (oxidation) are formed (Fig. 71). Therefore, organic electrochemistry is regarded as the chemistry of radical ions. In contrast to the use of particular redox reagents, the selectivity of the electrochemical reaction can be determined by the selection of the electrode potential (see Section 5.2). Furthermore, as the electrons are mass-free, no waste from spent reagents is produced. The ecologically and economically problematic separation and treatment of waste materials can be avoided to a large extent.

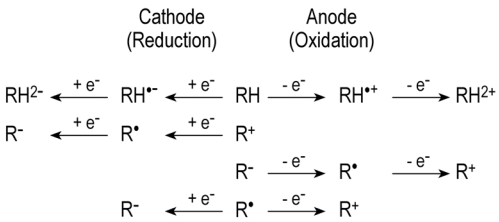

Figure 71. Electrochemically accessible reactive intermediates

From the retro-synthetic point of view, electrochemical redox reactions provide easy pathways to accomplish the principle of redox-umpolung (polarity reversal). As can be seen in Figure 71, oxidation of an electron-rich neutral compound will lead to an electrophilic radical cation, or starting with a nucleophilic anion, anodic oxidation may lead to an electrophilic cation. In the mirror image reductions, an electron-poor neutral compound is transformed to a nucleophilic radical anion, or an electrophilic cation will end up as a nucleophilic anion.

Accordingly, organic electrochemistry is an inherently non-pollutant technology. Electrochemical activation of organic molecules supplements the conventional chemical or catalytic techniques, which involve thermal or photochemical activation.

Organic electrochemistry is a mature field of organic synthesis. A very broad variety of reactions can be performed in all areas of organic chemistry [323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336, 337, 338, 339, 340, 341, 342, 343, 344, 345, 346, 347, 348, 349, 350, 351, 352, 353, 354]. Although the electrochemical step usually is only one step in multi-step transformations, it is often the key-step generating an activated intermediate. The electrochemical reaction is mostly followed by other chemical steps leading to the product. Chemical side reactions must be suppressed by proper selection of the reaction conditions. Product separation from the solvent – electrolyte system is very often a limiting factor which has to be optimized by means of chemical engineering.

5.2. Perspectives of Electroorganic Reactions in Industry

Although, electroorganic syntheses are among the oldest reactions in organic chemistry, e.g., the Kolbe dimerization of carboxylic acids [355], and although the electrochemical methodology has quite attractive features for technical use (see Section 5.3), up to now their application in industry is still limited. Only the tetraethyllead formation (Nalco) [356], [357] which seems to be discontinued, and the cathodic hydrodimerization of acrylonitrile to give adiponitrile [357, 358, 359, 360, 361, 362], which is the largest commercial electroorganic process with a total output of about 300 000 t/a (30 % of the total world production of adiponitrile), have found industrial realization on the scale of commodity chemicals. In contrast, a considerable number of small and medium size processes are performed in industry especially in the fine chemicals area and more are expected to come. Because of secrecy requirements in industry there is a lack of exact data. Nevertheless, in the field of medium scale productions, especially the anodic formation of aromatic aldehydes by side-chain oxidation of alkyl aromatics developed by BASF [363], [364] and the production of perfluorinated organics performed by 3M and Phillips Petroleum [365], [366] are successful by industrial means and thereby are impressive examples. The understanding of the underlying mechanisms and the preparative potential of electroorganic techniques has furthered the interest in this methodology also in industrial laboratories. A good overview of the prepar-

ative potential is found in a number of excellent monographs and reviews [323,324,325,326,327,328,329,330,331,332,333,334,335,336,337,338,339,340,341,342,343,344,345,346,347,348,349,350,351,352,353,354]. The patent literature on organic electrosyntheses in industry up to 1986 has also been reviewed intensively [363].

Organic electrochemical transformations must compete with other chemical redox processes, i.e., mainly catalytic processes using hydrogen (reduction) or oxygen (oxidation) as reagents. Therefore, it is necessary to give an overview on the advantages and risks which are encountered when electrochemical reactions are to be implemented in industry. The advantages of electroorganic reactions may be summarized as follows:

1) The selectivity of the reaction can be influenced by the applied electrode potential which can be controlled by the current density. In contrast to the common redox reagents (potential range mostly between + 1 to − 1 V vs. SHE), the potential can be varied continuously and extended up to 2 V in both directions.
2) The selectivity can also be enhanced by careful selection of the reaction conditions at the phase boundary, e.g., by selection of the electrode material and the electrolyte composition.
3) The high selectivity of electroorganic reactions allows for the selective transformation of a certain functional group within a polyfunctional molecule. This is especially important for the synthesis of high-value specialty products.
4) The reaction rate can be controlled by the applied current density.
5) The turnover degree of the starting material can be fixed by the charge consumption.
6) The reaction conditions are very mild. Usually, electrolysis is performed at ambient temperature and pressure. The rate of the electron transfer is not strongly influenced by the temperature. Therefore, lowering the reaction temperature will not diminish the electron transfer rate but eventually lower the tendency for chemical side reactions.
7) The pollution of the environment by spent reagents can be avoided. The possibility of regenerating redox reagents electrochemically in situ (indirect electrolysis, see Section 5.3.1) allows the use of environmentally unfriendly systems such as manganese(III) or chromium(VI) in a closed loop. Thus, the product quality can be enhanced. The control of the reaction and its automation in terms of a continuous processing can be achieved easily by proper selection of the electrical parameters.

On the other hand, some problems have to be solved for a successful application of electroorganic processes in industry:

1) For electrochemical reactions, special reactors (electrolysis cells) are required. Fortunately, a large number of electrolysis cells are commercially available nowadays for almost all types of electrochemical processes ranging from production scale down to laboratory size (see Section 5.4.3).
2) In principle, electrochemical reactions are heterogeneous in nature. They take place at the electrode/electrolyte phase boundary. Therefore, for technical applications optimization of the space – time yield is particularly important. This can be achieved by using a high electrode surface-to-volume ratio by application of electrodes with very high specific surface area (see Section 5.4.2). In other cases, a mediator system which acts as an electron shuttle between electrode and substrate may accelerate the reaction rate and avoid electrode fouling by shifting the rate-determining redox reaction into the homogeneous solution (indirect electrolysis, see Section 5.3.1).
3) The relatively low conductivity of organic solvents makes the addition of a supporting electrolyte necessary. An exception can only be made when solid polymer electrolytes are used. The supporting electrolyte must dissociate in the solvent. At the end of the electrolysis, it must also be separated from the product which can cause problems and can be costly. This problem, however, can be diminished by using electrolysis cells with very small gaps between the electrodes (capillary gap cells) so that only very low concentrations of supporting electrolyte are needed (see Section 5.4.3). Typical solvents for performing electroorganic processes at indus-

trial scale are mainly alcohols (especially methanol), acetic acid, or dimethyl formamide. However, the most important solvent for industrial use is water. Very often, co-solvents such as alcohols or acetonitrile are used to increase the solubility of the organic substrate. Aqueous sulfuric acid is also a traditional electrolyte. However, the acid has to be neutralized during work-up leading to the problem of waste salts. Emulsions or two-phase systems of water and an organic solvent are often more effective.

More than 200 electroorganic processes may have reached the pilot plant stage. The number and scale of processes which are really commercialized is lower and exceeds probably 100. The exact number cannot be determined because of commercial secrecy of industry [367]. The more detailed data given in [368,369,370,371,372] are also influenced by this lack of knowledge and should therefore only be used with care. Companies and research centers which are strongly involved in the development and application of electroorganic processes are:

- BASF in Germany as the leading company in the application of electrochemical processes including the manufacturing of adiponitrile, the formation of a variety of substituted benzaldehydes by side-chain oxidation of alkyl-substituted aromatics, the preparation of succinic dialdehyde and the application of electrodialysis.
- Hoechst in Germany has developed a number of bench scale and pilot plant scale processes. The epoxidation of hexafluoropropene was commercialized. For 90 years the company has been running the electrochemical regeneration of chromic acid on a large technical scale for the bleaching of montan waxes [373]. After take-over by Clariant 11 000 t/a of bleached montan waxes are produced at the plant in Gersthofen. A membrane process using dimensionally stable titanium anodes with lead dioxide coatings has been developed for this purpose and tested in pilot plants [374]. After restructuring of Hoechst, research in electroorganic chemistry seems to have been discontinued or taken over by Clariant.
- Hydro Quebec in Canada has developed a number of electrooxidations mainly using cerium(IV) as electrogenerated oxidizing agent. The manufacturing of anthraquinone via naphthoquinone seems to go in production at a cooperating company. Side-chain oxidations of alkyl aromatics seem to be also close to commercialization.
- Electricité de France (EDF) mostly in cooperation with Société Nationale des Poudres et Explosifs (SNPE) is very active in the development of electroorganic processes like the reductive carboxylation of benzylic halides to form the anti-inflammatory arylpropionic acids.
- Otsuka Chemical Co. in Japan is producing maltol (3-hydroxy-2-methyl-4-pyrone) by an electrochemical process and is very active in the study of a large number of other electrochemical processes.
- Asahi Chemical Co. in Japan is a producer of adiponitrile using the old Monsanto process in divided cells. It is also a manufacturer of dialkyl sebacate by Kolbe electrolysis of dialkyl adipate.
- Monsanto in the USA had been very active in the field of electroorganic synthesis. After restructuring the company, the research in this area was practically discontinued. The modern adiponitrile synthesis facilities in Great Britain have been sold to BASF and the production unit in the USA is now run by Monsanto's spin-off company Solutia Inc.
- Electrosynthesis Co. in East Amherst, New York, is very active in research, development and consulting in the area of industrial application of organic electrochemistry. Similarly, in India the Central Electrochemical Research Institute (CECRI) in Karaikudi and in Great Britain the Capenhurst Electricity Council Research Center (ECRC) are active research centers.

Other active companies are mentioned in Tables 20, 21, 22, 23, 24, 25, 26, 27.

5.3. Syntheses

5.3.1. Direct vs. Indirect or Mediated Electrochemical Reactions

In *direct electrochemical reactions* the substrate undergoes a heterogeneous redox reaction at the

electrode surface within the Helmholtz layer. The thus formed reactive intermediate, i.e., a radical ion, undergoes the chemical follow-up reaction to the product in the reaction layer. The steep concentration gradients near the electrode are one important factor in the chemical follow-up reaction of the electrogenerated intermediates. Second-order reactions of the intermediates can thus be obtained at high current densities.

In *indirect electrochemical reactions* [375, 376, 377], the heterogeneous reaction between the substrate and the electrode is replaced by a homogeneous redox reaction in solution between the substrate and an electrochemically activated and regenerated redox catalyst (Fig. 72). First-order reactions of the electrogenerated intermediates may be favored under these conditions. In addition, overpotentials for the reaction of the substrate at the electrode may be avoided and reactions may be accelerated. If electrode passivation causes problems for direct electrolyses, an indirect pathway could be an alternative because the electrode reaction of the substrate can thus be avoided. In many other cases, the selectivity properties of the redox catalysts can be used profitably in indirect processes. On the other hand, direct electrochemical reactions are less complicated because no additional catalysts must be added and separated.

Two types of redox catalysts are used in indirect electrolyses: (A) pure, outer-sphere or non-bonded electron transfer agents, and (B) redox reagents which undergo a homogeneous chemical reaction which is intimately combined with a redox step. This may be called inner-sphere electron transfer or bonded electron transfer mechanism.

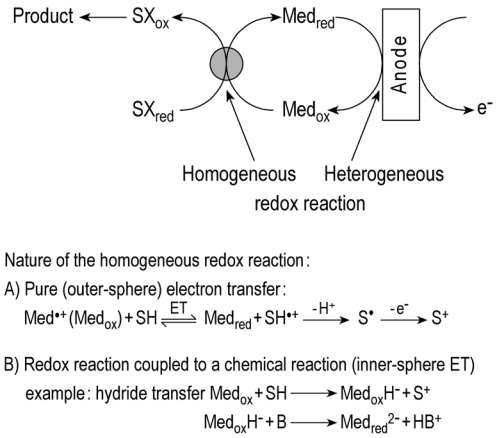

Figure 72. Schematic representation of an indirect electrochemical process in the in-cell mode, given for an oxidation Med_{ox}, Med_{red} = Oxidized and reduced form of the redox catalyst = (mediator); SX_{ox}, SX_{red} = Oxidized and reduced form of the substrate; ET = electron transfer

In case (A), the mediator catalyzes the electron exchange between the electrode and the substrate. As compared with direct electrolyses, overvoltages may be reduced and passivation of electrode surfaces may be avoided. If the homogeneous electron transfer equilibrium is followed by a fast and irreversible chemical step, mediators can be used with redox potentials which are up to 500 mV lower than the standard potential of the substrate (electron transfer against the standard potential gradient). Also, slow heterogeneous electron transfers at the electrode surface can be replaced by fast homogeneous electron transfers in solution. This is especially valuable, if large biomolecules are the substrates. Typical examples of such electron transfer mediators are stable radical anions or dianions of mostly aromatic or heteroaromatic systems or the stable radical cations of substituted triarylamines [378], [379].

Table 19. Advantages and disadvantages of indirect (mediated) electrochemical reactions with regard to the type of electron transfer

Mechanism	Advantages	Disadvantages
A) outer-sphere pure ET	no overpotential no passivation electron transfer against the potential gradient is possible (up to ca. 500 mV) increase of the reaction rate by replacement of a slow heterogeneous ET step by a fast homogeneous one	selectivity is only determined by the potential difference between the mediator and the substrate and the rate of the follow-up reaction
B) inner-sphere	very large potential differences can be overcome (up to 2 V) selectivity is controlled by the chemical step which is accompanied with the ET	regeneration is often difficult because an additional chemical reaction is necessary to transform the reacted mediator into a form from which the active catalyst can be regenerated

In case (B), the homogeneous redox reaction of the electrogenerated and regenerated redox catalyst consists of a chemical reaction. For oxidations, these reactions may be hydride ion or hydrogen atom abstraction, oxygen transfer, or an intermediate complex or bond formation. For reductions, hydride or carbanion transfer from a metal complex is often observed. In all these cases, very large potential differences between the standard potential of the substrate and the redox catalyst may be overcome. The selectivity can be very high and may be tuned by the proper selection of the metal ions and the ligands in the case of metal complexes as redox catalysts. All chemical steps involved must be very fast for an efficient redox catalytic system. Most of the inorganic redox catalysts, especially transition metal complexes, are acting in this way. For oxidations, not only transition metal salts or complexes but also halide ions are applied. The advantages and disadvantages of mediated electrochemical reactions are summarized in Table 19.

If the electrochemical generation and regeneration of the redox catalyst and the homogeneous redox reaction with the substrate take place in separate reactors, this is called an *ex-cell process*. If both steps are performed in-situ in the same reactor, this is called an *in-cell process*. While the in-cell process requires a lower investment, the ex-cell process is easier to optimize because both steps, the electrochemical one and the chemical one, can be treated separately.

Typical examples of inorganic redox couples which are used for indirect electrolyses are

for oxidations:
Ce(III)/Ce(IV); Cr(III)/Cr(VI); Mn(II)/Mn(III); Mn(II)/Mn(IV); $Ni(OH)_2$/NiOOH; I^-/I_2; Br^-/Br_2; Cl^-/ClO^-;
for reductions:
Sn(IV)/Sn(II); Cr(III)/Cr(II); Ti(IV)/Ti(III); Zn(II)/Zn; Na(I)/NaHg.

5.3.2. Technically Important Processes Ordered by Reaction Classes

Technically important processes which have been developed in industry to the pilot plant stage (P) or have been commercialized (C) are presented in the order of reaction classes. Bench scale processes (B) are only reported in a few cases. As mentioned before, the status of the processes is not always exactly known to the public.

5.3.2.1. Electrochemical Substitution

Substitution at the Anode.

$$R\text{–}E + Nu^- \rightarrow R\text{–}Nu + E^+ + 2e^-$$

Principle: Substitution of an electrophile (E), mostly a proton or in rare cases a trialkylsilyl group by a nucleophile (Nu), mostly from the solvent such as alcohol, water, or acetic acid, or by a halide ion.

Anodic substitution reactions are useful for nuclear or side-chain aromatic substitutions or for substitutions in the α-position to heteroatoms like nitrogen, oxygen, or sulfur.

Several industrial processes make use of anodic side-chain substitutions of alkyl aromatics. Thus, aromatic aldehyde acetals are generated mostly at graphite anodes in the presence of methanol in a direct electrochemical process in

undivided bipolar capillary gap cells (see Table 20, examples no. 1 – 3). This process is commercialized and used intensively by BASF in scales of several thousand tons per year.

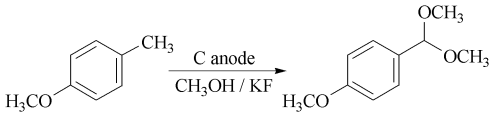

Other companies use indirect electrochemical processes applying electrogenerated manganese(III) (Table 20, no. 4) [380,381,382,383] or cerium(IV) (Table 20, no. 5) [384,385,386,387,388,389,390,391] as chemical oxidants in an ex-cell mode. The better solubility of cerium(IV) in aqueous methanesulfonic acid is advantageous as compared to the less soluble cerium ammonium nitrate (CAN), used for conventional oxidations. Based on developments by W. R. Grace, Hydro Quebec is currently intensively involved in the commercialization of nuclear and side chain substitutions of aromatic compounds using the system Ce(IV)/methanesulfonic acid.

Industrially important anodic aromatic substitution reactions are listed in Table 21. The formation of anthraquinone using electrogenerated dichromate in an ex-cell process has been developed in the UK (Table 21, no. 11). The electrochemical process is carried out in a filterpress type plate and frame membrane cell [377]. Hydro Quebec has developed a process for the nuclear (Table 21, no. 11 – 13) and side-chain oxidation of aromatics using electrogenerated Ce(IV) in aqueous methanesulfonic acid in an undivided bipolar plate and frame cell [384,385,386,387,388,389,390,391], [397], [398]. It is expected that a plant with a capacity of several thousand t/a will be constructed for the production of anthraquinone from electrogenerated naphthoquinone by reaction with butadiene.

The formation of hydroquinone from benzene via benzoquinone has been studied intensively (Table 21, no. 14). However, the process could not yet compete with the established methods and has therefore not been commercialized.

Perfluorination of hydrocarbons can only be achieved electrochemically in HF solution at nickel electrodes. A typical example for this technically extremely important process (Simons process developed by 3M Co.) is given below in Table 22, no. 17.

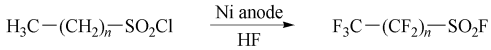

Partial fluorination can be obtained by the Phillips Petroleum process as indicated in Table 22, no. 18). Also a number of Japanese companies seem to perform electrochemical fluorinations on a commercial basis (Central Glass, Co.; Toyo Tannso Ind. Co.; Tokem Co.).

In a different way, perfluorinated functionalized compounds can be synthesized using electrogenerated peroxodisulfuryldifluoride which starts an in-situ additive telomerization with tetrafluoroethylene (Table 22, no. 19).

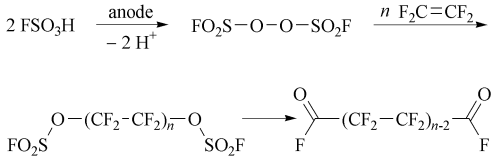

Selective monofluorinations especially in the α-position of heteroatoms or in benzylic positions by anodic oxidation in acetonitrile in the presence of $Et_3N \cdot nHF$ are promising for industrial application [410].

Anodic substitution can also occur easily in the α-position to heteroatoms like nitrogen or oxygen. Especially the α-methoxylation or acetoxylation of amides like *N*-acyl and *N*-alkoxycarbonyl derivatives of amines or amino acid derivatives leads to the synthetically very useful *N,O*-acetals which are powerful amidoalkylation reagents [411,412,413,414]. No practically useful chemical alternative exists. The formation of 2-methoxy-*N*-formylpiperidine from *N*-formylpiperidine has been reported to be developed to the pilot plant stage in Europe [369]. The *N,O*-acetals are starting materials for the synthesis of complex organic molecules like alkaloids. By elimination of methanol an alternative pathway to *N*-vinyl amides is opened [415].

Table 20. Industrially important anodic substitution reactions: aromatic side chain substitutions

No.	Substrate	Product	Status	Conditions	Remarks and references
1	4-Methoxytoluene	4-methoxybenzaldehyde dimethyl acetal	C (BASF)	direct electrolysis; MeOH/KF or other supporting electrolyte; undivided capillary gap disk-stack cell	[363], [364] several thousand t/a
2	4-*tert*-Butyltoluene	4-*tert*-butylbenzaldehyde dimethyl acetal	C (BASF)	direct electrolysis; MeOH/KF or other supporting electrolyte; undivided capillary gap disc-stack cell	[363], [364]
3	4-*tert*-Butyloxytoluene	4-*tert*-butyloxybenzaldehyde dimethyl acetal	C (BASF)	direct electrolysis; MeOH/KF or other supporting electrolyte undivided capillary gap disc-stack cell	[363], [367]
4	4-*tert*-Butyltoluene	4-*tert*-butylbenzaldehyde	C (ETH Zürich, Givaudan-Roure)	indirect electrochemical ex-cell process via electrogenerated Mn^{3+}; undivided bipolar plate and frame cell	[380,381,382,383]
5	Div. substituted toluenes	div. substituted benzaldehydes	P/C (Hydro Quebec)	indirect electrochemical ex-cell process via electrogenerated Ce^{4+} in aq. methane sulfonic acid; undivided bipolar plate and frame cell	[384,385,386,387,388,389,390,391] originally developed by W.R. Grace & Co.; commercial processes for anisaldehyde considered
6	*p*-Xylene	tolualdehyde	P/C (India)		
7	3,4,5-Trimethoxytoluene or 3,5-dimethoxy-4-hydroxytoluene	3,4,5-trimethoxybenzaldehyde by solvolysis of the intermediate quinone methide followed by air oxidation of the benzylic alcohol	C (Otsuka)	direct electrochemical oxidation in MeOH/NaOH	[392,393,394]
8	2-Methylpyridine (4-methylpyridine)	picolinic acid (isonicotinic acid)	C (Reilly Industry)	PbO_2 anode, aq. H_2SO_4, divided cell	[395] may be discontinued
9	4-Nitrotoluene	4-nitrobenzoic acid	C (India)	indirect electrochemical process via electrogenerated Cr^{6+}	
10	Nitrotoluenes	aminobenzoic acids	B/P (EDF)	paired indirect electrosynthesis using anodically generated Cr(VI) for the side chain oxidation and cathodically generated Ti(III) for the nitro group reduction (divided cell)	[396]

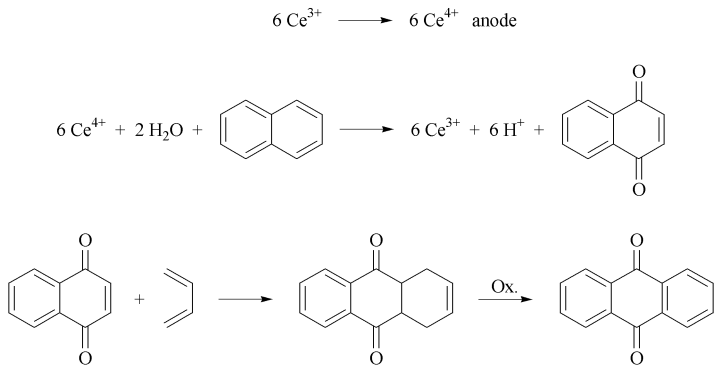

Table 21. Industrially important anodic substitution reactions: aromatic nuclear substitutions

No.	Substrate	Product	Status	Conditions	Remarks and references
11	Anthracene	anthraquinone			
			1) C (ECRC; L. B. Holliday with Cr^{6+}) 2) C (Hydro Quebec with Ce^{4+})	1) indirect electrochemical ex-cell process via electrogenerated Cr^{6+} in a divided parallel plate cell 2) via electrogenerated Ce^{4+} in methane sulfonic acid in an undivided bipolar plate and frame cell	1) [346], [377] 2) [384,385,386,387,388,389,390,391], [397]
12	Naphthalene	naphthoquinone	P/C (Hydro Quebec)	indirect electrochemical ex-cell process via electrogenerated Ce^{4+} in methanesulfonic acid in an undivided bipolar plate and frame cell	[384,385,386,387,388,389,390,391], [397] commercialization expected
13	1-Nitronaphthalene	5-nitronaphthoquinone	P (Hydro Quebec)	indirect electrochemical ex-cell process via electrogenerated Ce^{4+} in methanesulfonic acid in an undivided bipolar plate and frame cell	[398] formation of 1-aminoanthraquinone after reaction with butadiene and nitro group reduction
14	Benzene	benzoquinone	P	PbO_2 anode, aq. H_2SO_4, divided cell	[399,400,401,402] formation of hydroquinone possible by cathodic reduction in a paired electrosynthesis
15	Naphthalene	1-naphthylacetate	P/C (BASF, Union Carbide)	direct electrolysis; $HOAc/Me_3HN^+OAc^-$; undivided capillary gap disk-stack cell; PTFE-modified graphite anode	[403,404,405] formation of 1-naphthol by hydrolysis
16	2-Methylnaphthalene	2-methyl-1-naphthylacetate	P (BASF)	direct electrolysis; HOAc/trimethyl ammonium acetate; undivided capillary gap disk-stack cell; PTFE-modified graphite anode	[403,404,405] formation of 2-methyl-1-naphthol by hydrolysis

Table 22. Industrially important anodic substitution reactions: halogenations

No.	Substrate	Product	Status	Conditions	Remarks and references
17	Alkanoyl chlorides and fluorides; alkanoic sulfuric acid chlorides	perfluorinated alkanoic acid fluorides and sulfuric acid fluorides	C (3M, Bayer)	Simons process; Ni anode in liqu. HF	[365], [366], [406], [407]
18	Alkanes, chloroalkanes, carboxylic acids	partially fluorinated alkanes, chloroalkanes and alkanoic acid fluorides	P (Phillips Petroleum)	porous carbon electrodes; KF-2HF melt	[366]
19	Fluorosulfuric acid	peroxodisulfuryl difluoride	P (Hoechst)	undivided tank cell; glassy carbon anode, steel cathode; in fluorosulfonic acid/0.1 M alkali fluorosulfonate	[409] electrotelomerization of tetrafluoroethylene by electrogenerated peroxodisulfuryl difluoride

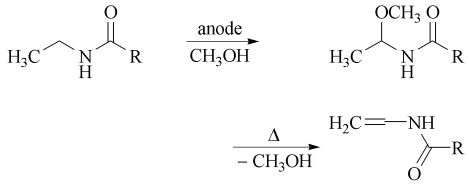

A phosgene-free synthesis of alkylisocyanates makes use of the indirect electrochemical oxidation in the α-position to nitrogen of formamides. Bromide in methanol solution acts as the redox catalyst which, presumably, is oxidized to the methyl hypobromite [416].

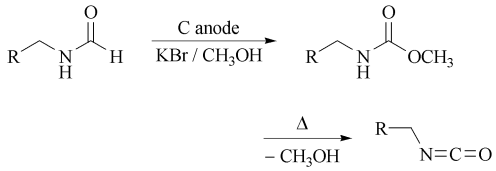

A different phosgene-free synthesis of isocyanates or carbamates starting from oxamates has been developed by Shell via substitution of a carboxylic function in the α-position to nitrogen by decarboxylative methoxylation (Non-Kolbe reaction) [417].

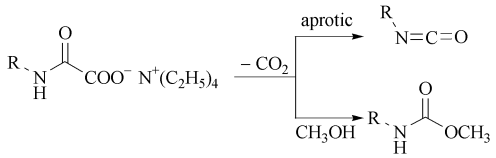

α-Hydroxylation of ketones can be performed anodically using iodide as a mediator in methanol solution leading to α-hydroxy acetals. These offer an industrially interesting way to acyloins [418]. Recently, BASF has commercialized the production of acetoin and adipoin according to this procedure [419].

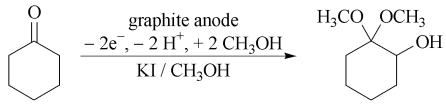

Substitution at the Cathode.

$$R{-}X + E^+ + 2e^- \rightarrow R{-}E + X^-$$

Principle: Substitution of a nucleophile (X^-), mostly halogen, by an electrophile (E^+), mostly an alkyl group from alkyl halides, carbon dioxide, a carbonyl compound, a Michael acceptor or a silyl group from silyl halides.

Cathodic substitutive carboxylation starting from benzylic halides using carbon dioxide as electrophile has been developed industrially for the synthesis of the anti-inflammatory arylpropionic acids naproxen, fenoprofen, and ibuprofen.

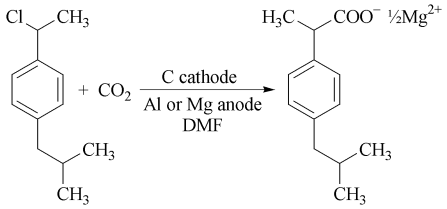

In France, SNPE in cooperation with EDF has developed this process to the pilot plant stage using a sacrificial magnesium anode (see Section 5.4.2.2) in a so-called pencil-sharpener cell (see Section 5.4.3.2; Fig. 80) [420, 421, 422, 423]. Alternatively, these products can also be obtained by reductive carboxylation of acetyl-substituted aromatics (see below).

5.3.2.2. Electrochemical Addition

Addition at the Anode.

$$R{-}CH{=}CH{-}Y + 2\,Nu^- \rightarrow Nu{-}CRH{-}CHY{-}Nu + 2e^-$$

Principle: Addition of two nucleophiles (Nu^-), mostly from the solvent to an electron-rich double-bond.

Anodic addition reactions have been developed on a technical scale for the dimethoxylation of furan and furan derivatives using bromide as redox catalyst to give 2,5-dimethoxy-2,5-dihydrofurans as synthetic intermediates. This process is commercialized by BASF for the formation of succinic dialdehyde and by Otsuka, Japan, for the synthesis of maltol and ethyl maltol (Table 23, no. 1,2).

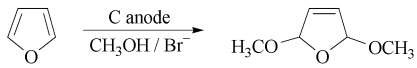

The epoxidation of double bonds has also been studied intensively in case of the bromide or chloride mediated formation of propylene oxide from propene via the intermediate formation

of BrO^- or ClO^- at the anode and of hydroxide at the cathode (Table 23, no. 3). Due to the formation of small amounts of halogenated byproducts and a limited space-time yield because of the low solubility of propene in the medium, this process never became commercialized. The synthesis of hexafluoropropylene oxide, however, has been carried out on a small scale by Hoechst (Table 23, no. 4). An electrochemical variant of the asymmetric bishydroxylation introduced by Sharpless has been developed by Engelhard Ltd. and ChiRex Ltd. in the UK, using a double mediator system consisting of ferricyanide and osmium tetroxide in the presence of chiral ligands (Table 23, no. 6) [433,434,435,436,437]. As an alternative to ferricyanide, electrogenerated iodine may be used [435], [436]. The double mediator system of Os(VIII)/ferricyanide without the use of the chiral ligands was already described in the late 1960s [437].

Addition at the Cathode.

$$-C{=}X + 2\,E^+ + 2e^- \rightarrow E{-}C{-}X{-}E$$

Principle: Addition of two electrophiles (E^+) to an electron-poor C-C-double-bond (C=C) or a C-heteroatom double-bond (C=X with X=N or O). Protons are mostly acting as electrophiles but also carbonyl compounds, carbon dioxide or alkyl and silyl halides are employed. Industrially important cathodic addition reactions are summarized in Table 24.

The electrochemical hydrogenation of double bonds can either be performed electrocatalytically at Raney nickel, palladium, or platinum modified electrodes [443], via electron transfer under Birch conditions to give the intermediate radical anion [444], or by direct electron transfer commonly using lead cathodes in sulfuric acid as electrolyte.

Other typical examples for the addition of protons are the reductions of carbonyl compounds such as aldehydes, ketones, or carboxylic acids and their derivatives to the corresponding alcohols or aldehydes (Table 25).

Cathodic hydrogenation reactions to form amines can be usefully applied to C=N double bonds like in imines and oximes or proceed via electroreductive amination of carbonyl compounds.

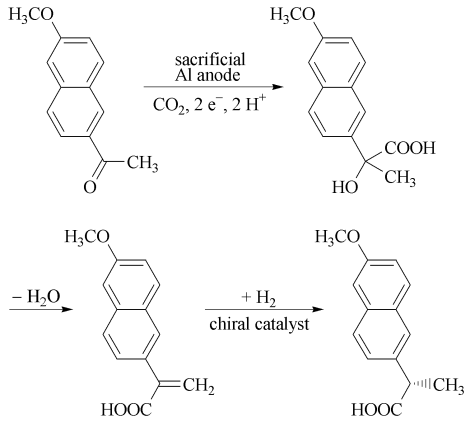

A novel synthesis of enantiomerically pure (*S*)-naproxene, an anti-inflammatory drug, has been developed by Searle and Solutia Inc. (Monsanto group). The first step is the cathodic addition of carbon dioxide as electrophile to an arylketone.

The formed α-hydroxy acid is subsequently dehydrated and the double bond homogeneously hydrogenated using chiral catalysts. The electrolysis takes place in an undivided cell using a lead cathode, a sacrificial aluminum anode and a tetraalkylammonium halide in DMF as electrolyte [464,465,466,467].

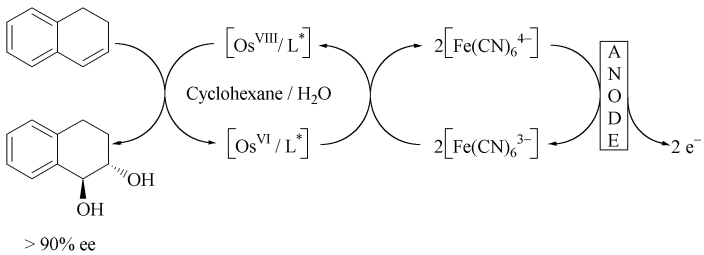

Table 23. Industrially important anodic addition reactions

No.	Substrate	Product	Status	Conditions	Remarks and references
1	Furan	2,5-dimethoxy-2,5-dihydrofuran	C (BASF)	indirect electrochemical in-cell process via the bromide/bromine couple in MeOH/NaBr in an undivided capillary gap disc-stack cell	[403]
2	Furfuryl-1-ethanol (also furfuryl-1-propanol)	2,5-dimethoxy-2,5-dihydrofurfuryl-1-ethanol and 1-propanol	C (Otsuka Chem. Co.)	indirect electrochemical in-cell process via the bromide/bromine couple in MeOH/NaBr	[424] part of the maltol and ethyl maltol synthesis
3	Propene	propylene oxide	P (Bayer, BASF, Olin, Dow, Univ. Dortmund)	indirect electrochemical in-cell process via the Cl^-/OCl^- or Br^-/BrO^- couple in H_2O/NaCl or H_2O/NaBr	[425] water reduction at the cathode to form hydroxide ions maintains the necessary pH value
4	Hexafluoropropene	hexafluoropropene oxide	C (Hoechst)	PbO_2 anode, divided cell, Nafion membrane, H_2O/HOAc/HNO_3	[426,427,428] status not known
5	Acetic acid and 1,3-butadiene	sorbic acid via γ-vinyl-γ-butyrolactone	B/P (Monsanto; Asahi Chem.)	indirect electrochemical process via electrogenerated $Mn(OAc)_3$	[429,430,431] addition of a carboxymethyl radical to 1,3-butadiene
6	Prochiral olefins	chiral 1,2-diols	P (Engelhard, ChiRex)	indirect electrolysis; ferricyanide/osmium tetroxide double mediator system in the presence of chiral ligands	[433,434,435,436,437] ferricyanide may be replaced by iodine electrogenerated from iodide

Table 24. Industrially important cathodic addition reactions: hydrogenation of aromatics and heteroaromatics and alkenes

No.	Substrate	Product	Status	Conditions	Remarks and references
1	Pyridine	piperidine	C (Robinson Bros.)	direct process; Pb cathode in H_2SO_4; divided cell	
2	2-Methylindole	2-methylindolene	C (L. B. Holliday; BASF)	direct process; Pb cathode in H_2SO_4; filterpress bipolar membrane divided cell	[377], [438,439,440]
3	Tetrahydrocarbazole	hexahydrocarbazole	C (L. B. Holliday; BASF)	direct process; Pb cathode in H_2SO_4; filterpress bipolar membrane divided cell	[377], [438,439,440]
4	Phthalic acid	dihydrophthalic acid	C (BASF)	direct process; Pb cathode in H_2SO_4; filterpress divided cell; dioxane as cosolvent	[403], [404] process discontinued; can be paired with the anodic oxidation of butynediol or propargyl alcohol
5	Naphthalene	1,4-dihydronaphthalene	P (Hoechst)	quaternary ammonium salts in water containing electrolytes; Hg or amalgam cathodes; divided cells	[363]
6	2-Methoxynaphthalene	2-methoxy-3,4-dihydronaphthalene	P (Hoechst)	quaternary ammonium salts in water containing electrolytes; Hg or amalgam cathodes; divided cells	[441] formation of β-tetralone
7	Maleic acid	succinic acid	C (India, CECRI)	direct process; Pb cathode in H_2SO_4	[442]

Table 25. Industrially important cathodic addition reactions: hydrogenation of carbonyl groups including the conversion of carboxylic acids and esters to aldehydes and alcohols

No.	Substrate	Product	Status	Conditions	Remarks and references
8	Glucose or mannose	sorbitol or mannitol	C (India, USA)	Pb cathode in acidic medium	[445], [446] probably discontinued; can be paired with the anodic formation of gluconate [447]
9	Oxalic acid	glycoxylic acid	P/C (Rhône-Poulenc; Steetley Chem.)	direct process; Pb cathode in H_2SO_4	[448,449,450] may also be paired with the oxidation of glyoxal to glyoxylic acid [451]
10	Anthranilic acid	*o*-aminobenzyl alcohol	P/C (BASF)	direct process; Pb cathode in acidic medium; divided membrane cell	[452]
11	2-Hydroxybenzoic acid (salicylic acid)	2-hydroxybenzaldehyde (salicyl aldehyde)	C (India)	Pb cathode in acidic medium	[453,454,455] process discontinued
12	3-Hydroxybenzoic acid	3-hydroxybenzyl alcohol	C (Mitsui Toatsu Chem. Co.)	Pb cathode in $H_2O/MeOH/H_2SO_4$; divided cell	[456] precursor for the synthesis of ethofenprox
13	Dimethyl terephthalate	4-hydroxymethyl benzoic acid methyl ester	P (Hoechst)	Pb or Hg cathode in MeOH/NaOAc	[457], [458]
14	1) Ammonium phthalamate from phthalic anhydride 2) Dimethyl phthalate	phthalide	B/P (BASF)	1) Pb cathode; pH 7 – 9 2) undivided capillary gap disk-stack cell with carbon electrodes in methanol/DMF	1) [459,460,461] 2) [462] *N*-methoxy-*N*-methylformamide is formed anodically as a couple product
15	4-Methylimidazole-5-carboxylic acid	4-methyl-5-hydroxymethylimidazole	P (Smith, Kline & French)	Pb or Hg cathode in H_2O/H_2SO_4	[463]

5.3.2.3. Electrochemical Transformations of Functional Groups

Functional Group Transformations at the Anode. Functional group transformations carried out technically include oxidations of alcohols and aldehydes to the corresponding carboxylic acids (Table 26). For montan wax bleaching, the crude montan wax, consisting mainly of esters of long chain alcohols and long chain acids is hydrolyzed under the acidic reaction conditions and the alcohol components are oxidized to the corresponding carboxylic acids by electrogenerated and regenerated Cr(VI). This is one of the most important electroorganic processes (see Section 5.2).

The indirect electrochemical oxidation of the aldehyde groups in aldoses to the corresponding aldonic acids (Table 26, no. 1) has been carried out industrially since 1930 and is still used today by Sandoz (Novartis) and in India. The bromide/hypobromite couple is used as the redox catalyst.

Functional Group Transformations at the Cathode. Cathodic functional group transformations are the reduction of cyano, nitro or amide functions to amino and aminomethyl groups. Reductive transformations of carbonyl compounds are treated as addition reactions of protons (see Section 5.3.2.2).

The cathodic reduction of aromatic nitro compounds to aniline derivatives is possible. Generally lead cathodes in acidic media are used. However, this process can usually not compete with the catalytic pathway [453], [475]. In developing countries the electrochemical method may be advantageous under special conditions. Thus, a considerable number of such processes are commercialized in India [368]. Sometimes, the processes are performed as indirect reactions using electrogenerated Ti(III) as the redox catalyst to lower the reduction potential and thereby decreasing the energy consumption [476]. The electrochemical methodology is technically more important for partial reduction of nitro groups to the hydroxylamine function in

Table 26. Industrially important anodic functional group transformations: oxidation of alcohol or aldehyde functions to carboxylic acids

No.	Substrate	Product	Status	Conditions	Remarks and references
1	Glucose or lactose	calcium gluconate or lactobionate	C (Sandoz; Chefaro, India)	indirect electrochemical in-cell process via the Br^-/BrO^- couple at C-anodes in the presence of NaBr and $CaCO_3$	[328], [442] can be paired with the cathodic formation of sorbitol [447]
2	Diacetone-L-sorbose	diacetone-2-keto-L-gulonic acid	P (Hoffmann-La Roche) C (Merck, Darmstadt)	on high-surface area Ni-anodes in aqueous base via the surface formation of NiOOH (indirect process); undivided Swiss-Roll cell	[469,470,471,472,473] intermediate for vitamin C synthesis; pilot plant by Hoffmann-La Roche using Swiss-Roll cell; commercialization by Merck
3	2-Butyne-1,4-diol	acetylene dicarboxylic acid	P (BASF)	PbO_2 anode, aq. H_2SO_4, divided cell	[403] can be paired with the cathodic reduction of phthalic acid
4	Propargyl alcohol	propiolic acid	P (BASF)	PbO_2 anode, aq. H_2SO_4, divided cell	[403] can be paired with the cathodic reduction of phthalic acid
5	Long chain alcohols from hydrolysis of montan wax	long chain carboxylic acids	C (Clariant, formerly Hoechst)	1) Pb/PbO_2 anode, aq. H_2SO_4, divided cell 2) cation-exchange membrane cell with dimensionally stable Ti/PbO_2 anodes	1) [373] large scale indirect electrochemical ex-cell process 2) [374]
6	Galacturonic acid	mucic acid	C (EDF)	indirect electrochemical process via the Br^-/Br_2 couple in 0.5 M HBr in a divided cell	[474] a modified Grignard reactor is used as electrolysis cell

buffered media ($HOAc/NH_4OAc/H_2O/EtOH$) [477] or for the reductive rearrangement to give *p*-aminophenol derivatives. Starting from nitrobenzene, *p*-aminophenol can be generated at copper or Monel cathodes in aqueous sulfuric acid at elevated temperature via the intermediate hydroxyl amine which rearranges under the reaction conditions (Bamberger rearrangement). If the cathodic reduction is carried out in the presence of alkanols, the corresponding alkoxy substituted anilines (*p*-aminophenol ethers) can be obtained. This reaction is unprecedented by other methods [478], [479]. Many companies have studied it including Bayer, Miles Laboratories (now Bayer), L. B. Holliday, PPG, and BASF. In India, the *p*-aminophenol formation seems to be carried out on a commercial scale [442].

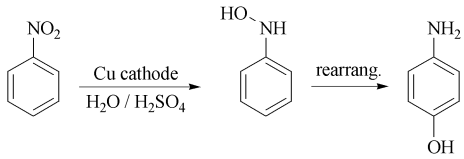

Cathodic reduction of 3-nitropyridines in hydrochloric acid using a membrane-divided flow-cell leads to 3-amino-2-chloropyridines also via a Bamberger rearrangement. Thus, the formation of 3,4-diamino-2-chloropyridine from 4-amino-3-nitropyridine has been developed on the miniplant scale by Merck, Darmstadt [480].

Cyano groups can be electrochemically reduced to give aminomethyl functions. This reaction has mainly being applied technically for the reduction of cyanopyridines to give the aminomethylpyridines using lead cathodes in acidic media in a divided membrane cell. This process was commercialized by Reilly Tar but seems to be discontinued [346], [395], [481].

Electrochemical carboxamide reductions have been applied in the formation of isoindoles from phthalinides by Ciba-Geigy.

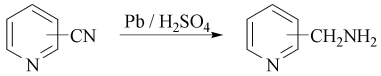

5.3.2.4. Electrochemical Cleavage and Elimination

Cleavage at the Anode. Cleavage reactions at the anode are until now in most cases of little interest for technical applications. They are mostly used for the cleavage of protecting groups. Oxidation in the α-position of heteroatoms can lead to the cleavage of carbon – oxygen, carbon – sulfur, carbon – nitrogen, and carbon – silicon bonds [482,483,484,485,486,487,488,489,490]. By cleavage of the carbon-nitrogen bond in amines, the formation of metabolites is possible because the anodic oxidation of nitrogen compounds equals the microsomal oxidation to a large extent [488]. This method has been applied to gain easy access to metabolites of pharmacological compounds to be studied toxicologically.

The anodic C–C bond cleavage of the 1,2-diol structures in starch to give dialdehyde starch has been studied intensively by the industry and has been performed technically via an indirect electrochemical process involving formation of electrogenerated IO_4^- at lead dioxide anodes. This reaction has been studied by Miles Laboratories (now Bayer) and is used commercially in India [442], [491], [492].

Starting from D-gluconic acid, decarboxylative cleavage of the C_1-C_2 bond leads to the formation of D-arabinose (Non-Kolbe electrolysis). D-glucose can also be used directly as the starting material [493].

Cleavage at the Cathode.

$$R\text{–}X + e^- \rightarrow R\text{–}X^{\cdot -} \rightarrow R^{\cdot} + X^-$$
$$R^{\cdot} + e^- \rightarrow R^-$$

In the cathodic cleavage, halogen, phosphonium, ammonium, nitrile, thioalkyl, or sulfonate groups can act as leaving groups and are replaced by hydrogen. A special case is the cleavage of the disulfide bond to form thiols. Selectivity between different leaving groups can be obtained by potentiostatic reduction. The application to deprotection has been used successfully [494], [495]. The selectivity in the cathodic cleavage of phosphonium salts has been used very early for the formation of optically active phosphine and phosphonium ions [496].

Industrially, besides the cleavage of carbon – halogen bonds (see Table 27) the cathodic cleavage of the disulfide group in cystine or cystine derivatives to give the corresponding cysteine or its derivatives has been studied and applied commercially (e.g., by Isochem, Nippon Chem., Degussa, Diamalt, Tanabe, Takeda, Ajinomoto, Electrosynthesis Co., Universidad de Alicante). The cleavage usually either takes place at lead cathodes in acidic media using a divided membrane cell or at carbon felt cathodes and a DSA anode with an aqueous ammonium hydroxide catholyte and a 3 M aqueous sulfuric acid anolyte [497,498,499,500,501,502,503,504]. If the cleavage of cystine is performed in the presence of chloroacetic acid in basic medium, *S*-carboxymethyl-L-cysteine can be obtained [503].

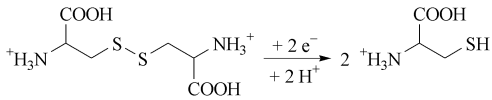

3-Exomethylene-cephem derivatives have been generated via reductive cleavage of the carbon – halogen or carbon – acetoxy function (Table 27, nos. 3, 4) [509,510,511].

Table 27. Industrially important cathodic dehalogenations

No.	Substrate	Product	Status	Conditions	Remarks and references
1	Di- and trichloroacetic acid	monochloroacetic acid	C (Hoechst)	indirect electrochemical process via the Pb^{2+}/Pb^0 couple in $ClCH_2COOH/H_2O/HCl$ (56:33:11) in the presence of 10 – 100 ppm $Pb(OAc)_2$	[426] reduction of overchlorinated products to the desired product in the synthesis of monochloroacetic acid
2	3,4,5,6-Tetrachloropicolinic acid	3,6-dichloropicolinic acid	C (Dow Chem.)	high-surface area (spongy) Ag-catodes activated by anodization in aqueous alkaline media in an undivided cell	[505, 506, 507, 508] product is a potent herbicide; production may be discontinued
3	2,2,2-Trichloro-1-(4-hydroxyphenyl)ethanol	2,2-dichloro-1-(4-hydroxyphenyl)ethene (elimination process)	C (Japan)	divided filter-press cell; Pb cathode in acidic $MeCN/H_2O$	[509] formation of p-hydroxyphenylacetic acid by hydrolysis
4	3-Chloromethyl-cephem derivative	3-exomethylene-cephem derivative	C (Japan)	at Pb cathodes in THF/ $H_2O/NH_4ClO_4/LiClO_4$	[510], [511]

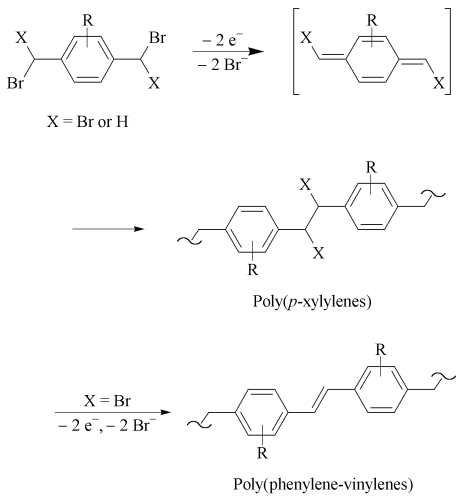

By reductive elimination of halogen from 1,4-bis(bromo or dibromomethyl)benzene, *p*-quinodimethanes are formed. These polymerize in situ to give either poly-(*p*-xylylene) or by further reductive elimination poly-(*p*-phenylene-vinylene), which are both important conducting polymers [512], [513].

The cathodic cleavage of the silicon-chlorine bond in dichlorosilanes to synthesize polysilanes has been intensively studied. The commercialization by Osaka Gas Co., Japan, has been reported. This process is best performed in $THF/LiClO_4$ using Mg cathodes and sacrificial Mg anodes [514, 515, 516, 517, 518].

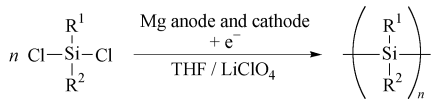

The cathodic cleavage of tetraalkylammonium groups is applied industrially in the formation of 1,4-dimethyl-2,3,5,6-tetrafluorobenzene from 1,4-bis(trimethylammoniomethyl)-2,3,5,6-tetrafluorobenzene dichloride making use of the selective cleavage of the benzylic carbon – nitrogen bond as compared with the methyl groups.

5.3.2.5. Electrochemical Coupling

C–C coupling reactions are in the center of organic syntheses to build up larger structures. Electrochemistry provides simple and efficient procedures for coupling reactions. The anodic Kolbe dimerization of carboxylates is one of the oldest electrochemical reactions [355] which is still important under technical aspects.

Coupling at the Anode.

Type 1:
$2\,RH \rightarrow R\text{–}R + 2\,H^- + 2\,e^-$ via cation radicals
Type 2:
$2\,R^- \rightarrow R\text{–}R + 2\,e^-$ via radicals

Type 1 describes the coupling of two electron-rich components like alkenes, aryl

olefins, enolacetates, enaminoesters, or enolethers [519]. Type 2 is the coupling of radicals obtained by electrooxidation of carbanions or carboxylates (Kolbe dimerization) [520].

The anodic coupling of compounds with an acidic methylene hydrogen has been developed technically by Monsanto Co. in the case of the tetramethyl ethane tetracarboxylate formation from dimethyl malonate. This reaction takes place in methanol in the presence of methoxide and of sodium iodide which acts as a redox catalyst [521].

The Kolbe dimerization of half esters (Brown – Walker reaction) has been applied on a pilot plant scale and also commercially in a number of cases by companies like BASF, Asahi Chem., Soda Aromatic, and other companies in Russia. Examples are the formation of dimethyl sebacate, dimethyl tetradecandioate and dimethyl hexadecanoate from monomethyl adipate, monomethyl suberate and monomethyl azelate, respectively. The reaction takes place at about 10 to 30 % neutralization of the acid at platinum anodes. The slow corrosion of the platinum is the major problem [403], [522].

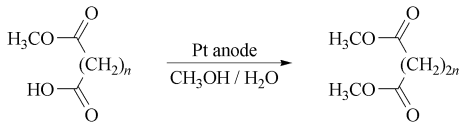

Coupling at the Cathode.

Type 1:
$2\,–CH{=}CZ– + 2\,e^- + 2\,H^+ \rightarrow –CHZ–CH_2–CH_2–CHZ–$ via anion radicals
or
$2\,–CR{=}X + 2\,e^- + 2\,H^+ \rightarrow HX–CR–CR–XH$ via anion radicals
Type 2:
$2\,R^+ + 2\,e^- \rightarrow R–R$ via radicals

At the cathode, two electron-poor components can be coupled. While the dimerization of cationic species (type 2) has industrially only some importance in the coupling of pyridinium salts to give 4,4′-bispyridinium salts, type 1, the electrohydrodimerization (EHD) of electron-poor alkenes (–CH=CZ–) and carbonyl compounds (–CR=X with X = O; pinacolization), has found important applications.

The most important example of type 1 is the *electrohydrodimerization* of acrylonitrile to give adiponitrile developed by M. M. Baizer at Monsanto [357,358,359,360,361,362]. This process is commercialized by several companies on an overall scale of more than 300 000 t/a. While the early process required a divided membrane cell and a high concentration of tetraalkylammonium salts as supporting electrolyte, the modern process is run in a bipolar stack of carbon steel sheets, the cathodic side cadmium plated with a 2 mm interelectrode gap. Thereby, the concentration of the supporting electrolyte could be reduced to 0.4 % and the energy consumption was lowered to 40 % [346], [357,358,359,360,361,362].

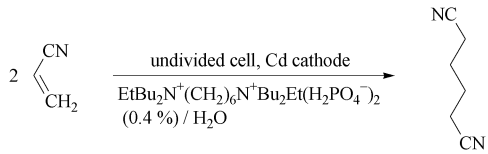

Similarly, the highly selective hydrodimerization of maleic acid esters to yield butanetetracarboxylic acid esters or the corresponding tetracarboxylic acid has been developed by Monsanto [523].

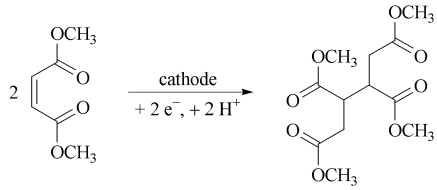

Other EHD reactions which have been developed to the pilot plant stage are the *pinacolizations* of formaldehyde to give ethylene glycol (Electrosynthesis Co.) [524,525,526], acetone to yield pinacol (Diamond Shamrock) [527], *p*-hydroxybenzaldehyde to yield bis(4-hydroxyphenyl)ethane (Monsanto) [528], [529] and of *p*-hydroxypropiophenone to give 3,4-dihydroxy-3,4-bis(4-hydroxyphenyl)hexane (Sorapec, Glaxo, Reilly Ind.) [438]. The production of ethylene glycol by cathodic coupling of formaldehyde is piloted by two Canadian companies. It takes place in a divided cell at modified graphite electrodes in the presence of tetraalkylammonium salts. The economics of the chemical

and electrochemical alternatives are similar and depend strongly on the prices of the different feedstocks.

5.3.2.6. Bioelectrosynthesis

Since the mid-1990s, the field of bioelectrosynthesis has been growing considerably. In this approach, the chemo-, regio-, stereo- and enantioselectivity of redox enzymes is combined with the convenient supply of the redox equivalents by the electrode [530,531,532,533,534,535,536,537]. Thus, the redox state of the enzymes or their cofactors is controlled either by direct or mediated electrolysis allowing the development of continuous processes in flow-through systems. The enzymes, mediators (sometimes bound to water soluble polymers) and cofactors can be held back by ultrafiltration membranes or they may be immobilized at the electrode surface. Although these reactions are still developed only on the laboratory scale, industrial applications would have an interesting potential.

5.3.3. Organic Electrochemistry for a Cleaner Environment

The aspect of environmental protection by using electrochemical methods has gained more and more importance since the 1980s. Electrochemistry does not only offer new methods for waste treatment (curative environmental protection) but, even more important, also the possibility of preventing waste formation by using electrochemical production processes (preventive environmental protection). As the electron is mass-free, no spent reagents are formed. Additionally, regeneration of redox reagents can be performed electrochemically either in situ (in-cell method) or in a two-step process (ex-cell method) (see Section 5.3.1). This allows the use of environmentally unfriendly systems so that the pollution of the environment by spent reagents can be avoided. A typical example is the ex-cell electrochemical chromium(VI) regeneration in montan wax bleaching (see Section 5.3.2.3 [373], [374]). The control of the reaction and its automation in form of a continuous process can be achieved easily by proper selection of the electrical parameters. Thereby the development of closed systems is feasible. These aspects have been reviewed in [538], [539]. Three typical examples for preventive environmental protection will demonstrate the power of electrochemical procedures in this area:

1) The electroreductive restoration of monochloroacetic acid from di- and trichloroacetic acid (see Table 27, no. 1) [426] eliminates the necessity to treat the hazardous waste of these side products.
2) In the generation of methoxyacetic acid by anodic oxidation of methoxyethanol at a nickel oxide hydroxide anode the methoxyacetic acid can be liberated from its salt and the strongly alkaline aqueous electrolyte can be recovered by electrodialysis using bipolar membranes thus avoiding a high salt freight [540].
3) In the dying with vat dyes, the reducing agents, usually sodium dithionate, cannot be regenerated. Thus, a recycling of the dyebaths is not possible and the spent reducing agents lead to a high salt freight. This can be avoided if complexes of iron with triethanolamine or tri(isopropanol)amine are used as regenerable redox mediators. Processes especially for dying with indigo have been developed [541,542,543,544,545].

5.4. Experimental Factors and Techniques

The following experimental factors have to be taken into account for the application of an electrochemical reaction in synthesis:

- The choice of the electrolysis procedure- either working at controlled potential (potentiostatic or controlled potential electrolysis, CPE) or at a fixed current density (galvanostatic)
- The type of electron exchange
- The redox potential applied at the working electrode, if controlled potential electrolysis is selected
- The current density, if galvanostatic conditions are selected
- The charge consumption
- The composition of the electrolyte
- The type of electrode material

– The type of the electrolysis cell, divided or undivided, flow-through or beaker-type.

5.4.1. Solvent – Electrolyte Systems

The electrolysis medium (solvent and supporting electrolyte) influences the electrochemical process in many ways. The choice of the medium does not only largely determine the solubility of the compounds and the conductivity, but also the potential window for the desired transformation, the selectivity of the reaction, the chemical follow-up reactions, the separation of the products, the reaction at the counter electrode, the energy consumption and thereby also the economics of the process.

In general, the following requirements should be met by the electrolyte system:

1) It should have a high conductivity so that low cell voltages are possible at low concentrations of the supporting electrolyte. For this purpose, solvents of high dielectric constant are desirable.
2) The supporting electrolyte salts should be totally dissociated in the applied solvent. The dissociation of salts of larger ions is usually more readily obtained than that of salts of smaller ions.
3) The solubility of the components in the solvent must be sufficient. For technical applications, concentrations of the reactants of more than 0.5 M are desired to reach an economical current density. To improve the solubility, co-solvents can be used. Another possibility is the application of a two-phase system or an emulsion in the presence of phase transfer catalysts. A two-phase system has also advantages in product isolation and continuous electrolysis procedures. A typical example is the synthesis of *p*-methoxybenzonitrile by anodic substitution of one methoxy group in 1,4-dimethoxybenzene by the cyanide ion. The homogeneous cyanation system (acetonitrile, tetraethylammonium cyanide) [546] can be efficiently replaced by a phase-transfer system (dichloromethane, water, sodium cyanide, tetrabutylammonium hydrogensulfate) [547].
4) The stability of the electrolyte system at the applied potential or current density has to be high enough to ensure high current yields and to prevent electrolyte destruction. The stability of an electrolyte can be determined by measuring the foot potential in current – voltage curves. Limiting values can be found in most books on electroorganic chemistry [548].
5) For application in industry, the work-up of the product containing electrolyte must be easy and efficient. With capillary gap cells, the concentration of a supporting electrolyte may be so low that its recovery is not necessary from an economical point of view.

Other properties of the solvent-supporting electrolyte system may influence the reaction directly. For example, in aprotic solvents the cathodic limiting potential can be much more negative than in protic media, thus allowing the reduction of compounds with very negative potentials. However, under these conditions many compounds and electrogenerated intermediates tend to polymerize. Much cleaner reductions can often be obtained if small amounts of a proton donor are present to trap the reactive anionic intermediates. The nucleophilicity or electrophilicity of the electrolyte has also to be taken into account. Oxidations are very often running more smoothly and give cleaner products if at least small amounts of a nucleophile are present to trap cationic intermediates. In the case of reductions, it can be very helpful to trap the intermediates by the addition of electrophiles such as trimethylsilyl chloride, alkyl halides, carbonyl compounds, or carbon dioxide.

The electrolyte may also influence the product distribution. This can be demonstrated in the case of the side-chain oxidation of alkyl aromatics (see Fig. 73) [549].

Another well known example for the influence of the supporting electrolyte is found in the adiponitrile formation from acrylonitrile. If, for example, alkali metal salts are used as supporting electrolyte in the aqueous solution, the main product is propionitrile formed by hydrogenation of the double bond. This is favored over the dimerization because the metal ions which are accumulated within the Helmholtz layer are strongly hydrated so that the electrochemically formed

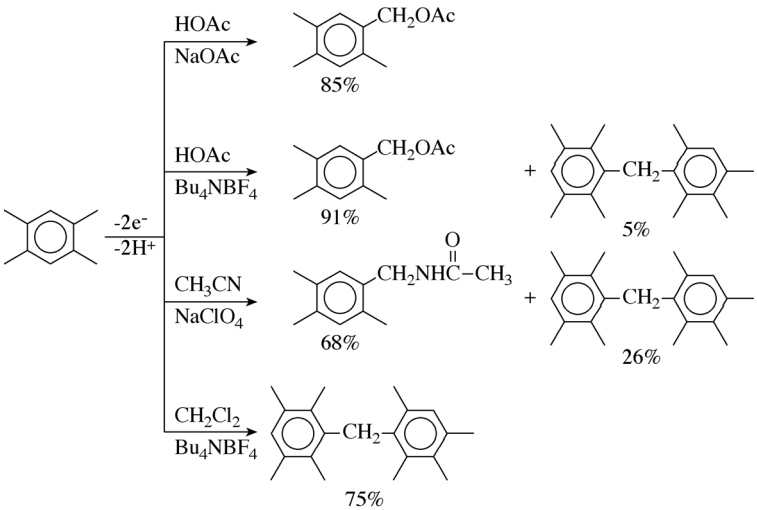

Figure 73. Influence of the product distribution by the nucleophilicity of the electrolyte

acrylonitrile anion radical is immediately protonated. To favor the dimerization, a tetraalkylammonium salt like $Et_4N^+EtOSO_3^-$ (early process) or $EtBu_2N^+(CH_2)_6N^+Bu_2Et(HPO_4^-)_2$ (modern process) has to be applied which is bulky and not hydrated, thus lowering the amount of water in the Helmholtz layer [357, 358, 359, 360, 361, 362].

It is obvious that water is the solvent of choice for industrial applications. Reasons besides the low price are its low impact on the environment, the good solubility of salts leading to highly conducting solutions, the easy control of the pH and the ease of controlling the reaction at the counter electrode (hydrogen or oxygen formation). On the other hand, because of solubility and selectivity reasons, organic solvents are often superior. Therefore, compromises have to be made. Thus, protic solvents, especially methanol, other alcohols, acetic acid, and mixtures of water or alcohols with other organic solvents are the next best choice. Dry aprotic solvents, often used in laboratory experiments, are generally not useful in larger scale applications. This is due to their high resistance, the instability of most membranes to these solvents, the difficulty in controlling the pH and the problem of controlling the reactions at the counter electrode for which purpose additives have to be used. Otherwise, polymerization, electrode passivation and the formation of side products are encountered. Sometimes these problems can be overcome by the addition of 5 to 10 % of water or methanol or by using emulsions or two-phase systems. Because of their value and their environmental impact, these solvent have to be recycled.

5.4.2. Electrodes

The choice of the electrode materials for the working electrode and also for the counter electrode influences the electroorganic process in many respects [550, 551, 552]. The material of the working electrode very often determines the selectivity of the electrode reaction – depending on the mechanism of the electron transfer (inner sphere or outer sphere) – and also the service life. The material of the counter electrode influences the type of the electrochemical process taking place at this electrode and thereby determines the electrode potential which develops at the counter electrode at a given current density. This affects directly the overall cell voltage and, thereby, the energy cost of the process. In addition, a controlled process at the counter electrode also influences the selectivity of the process by preventing undesired reactions of the products or starting materials at this electrode. By proper selection of the counter electrode and, thereby, the electrode process proceeding at this electrode, the application of a divided cell can be avoided in many cases. Thereby, also the cost

of the process can be diminished. In general, the electrode materials should meet the following requirements.

- Physical stability
- Chemical stability
- High conductivity
- Suitable physical form
- Suitable electrocatalytic properties
- Long service life
- Nonpolluting/noncontaminating
- Low cost
- No passivation/deactivation

Physical properties like physical stability, suitable physical form, and especially low electrical resistivity are essential. Massive, monopolar electrodes can be used in the case of good conducting metals (10^{-6} to $10^{-5}\,\Omega^{-1}\,cm^{-1}$) like Ag, Cu, Al, Zn, Ni, Cd, Fe, Pt, Pb, Ti, or PbO_2. For less conductive materials (10^{-4} to $10^{-3}\,\Omega^{-1}\,cm^{-1}$) like stainless steel, Hastelloy-C or carbon materials, a bipolar electrode arrangement reduces the current paths through the electrodes. Conducting polymers (conductance 2×10^{-3} to $10^{-2}\,\Omega^{-1}\,cm^{-1}$) have to be used as thin coatings on good conducting substrates.

The chemical stability criteria do not apply in cases where a dissolving anode (sacrificial anode) is selected to control the electrochemical process at the counter anode or to stabilize the products.

It is also necessary to divide the electrode materials into two categories:

1) Electrodes which are acting as pure electron acceptors or electron donors (outer-sphere electron transfer);
2) Electrodes which act as reagents or as heterogeneous redox catalysts (inner-sphere electron transfer).

The first category is made up mainly by the different carbon materials and the precious metals, if applied as working anodes. The second category contains metal and precious metal oxides mostly in the form of coatings on conductive substrates acting for example via oxygen transfer or hydride or hydrogen abstraction.

5.4.2.1. Cathodes

Cathodes as Working Electrodes. In aqueous media, cathodes used as working electrodes should have a hydrogen overvoltage which is high enough to prevent hydrogen generation. Table 28 presents the hydrogen overpotentials of some common electrode materials in 0.5 M aqueous sulfuric acid.

In nonaqueous, aprotic media the hydrogen overvoltage does not play a role. Commonly used cathode materials are lead, aluminum, copper, titanium, cadmium, silver, steel, stainless steel, Hastelloy-C (Ni – Mo – Fe or Ni – Mo – Cr) and all kinds of carbon.

On the other hand, if an electrocatalytic hydrogenation is the desired process, platinum group metal electrodes or coatings or high surface area nickel or nickel alloys can behave as hydrogen transfer catalysts via adsorbed hydrogen. This methodology has been widely used for the reduction of ketones, aldehydes, nitriles, alkenes, alkynes, Schiff bases, aromatics, nitro compounds, and carbohydrates [551], [553, 554, 555].

Cathodes as Counter Electrodes. In anodic oxidations, a controlled process has to take place at the counter cathode. Mostly the hydrogen evolution is selected. Hydrogen evolution can be supported by the application of precious metal electrodes or coatings thus lowering the electrode potential and, thereby, the cell voltage. To lower the cost of the materials, also Ni electrodes and coatings or different kinds of steel and also carbon of different types are commonly employed. Another possibility for controlling the process at the counter cathode and to lower the cathode potential is to use an oxygen consuming cathode (dispersed Pt on high area carbon). Nonstoichiometric titanium oxide (approximate composition Ti_4O_7), called Ebonex, has been proposed as a counter cathode. It is a very stable material and appears to be a good substrate for electrocatalytic layers [556], [557].

5.4.2.2. Anodes

Anodes as Working Electrodes. As more or less inert anode materials for working electrodes, platinum (Kolbe electrolysis) or all types of carbon and graphite are applied. Among carbon materials especially the three-dimensional

Table 28. Hydrogen overpotentials of common electrode materials in 0.5 M sulfuric acid

Electrode material	Pd	Fe	Pt	Ni	Sn	Pb	Zn	Hg
Hydrogen overpotential, V	0.00	0.08	0.09	0.21	0.53	0.64	0.70	0.78

materials like carbon felt and reticulated vitreous carbon are of interest for applications in which the current density has to be kept low.

Very often, metal oxide electrodes, mostly supported on a conductive carrier, can be applied as anodes. In most of these cases, the metal oxides are not inert but act as electrocatalysts for instance, via oxygen transfer (PbO_2, mostly supported on Ti or on Pb) or hydrogen abstraction (NiOOH on Ni) [469,470,471,472,473], [558]. Other electrocatalytic electrodes are Fe_3O_4, RuO_2 on Ti (dimensional stable anode, DAS), Co_3O_4, IrO_2 on Ti, MnO_2 on Ti or Ni/Co spinels [559]. Conducting ceramics may also be applied as catalytic anodes as it has been shown for Cr_2O_3 on Ti [560]. However, in most of these cases, because of the limited number of active centers on these metal oxide electrodes, the current densities have to be kept low and therefore high surface area electrodes have to be developed. The application of such oxide electrodes has by no means yet been fully developed to their expected potential. One major problem with anodes is their possible corrosion.

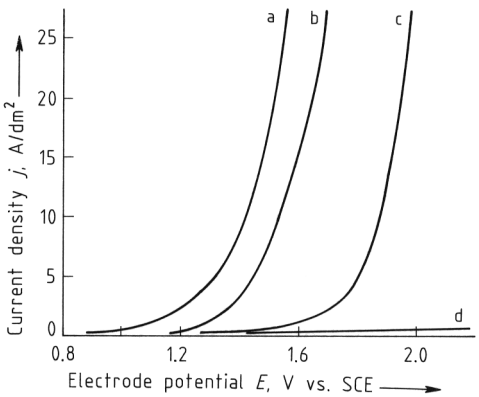

Figure 74. Anodic current – voltage curves for four anode materials in aqueous Na_2HPO_4 with 25 mM quaternary ammonium salt, equilibrated with AN/ADN (25 : 75)
a) DSA (RuO_2 and TiO_2 on titanium metal); b) Carbon steel; c) Pb/PbO_2; d) Pt
Courtesy of *Hydrocarbon Processing* [397].

Anodes as Counter Electrodes. For cathodic reductions, a controlled electrode process has to take place at the counter anode. This allows very often the application of an undivided cell and the lowering of the cell voltage. Most commonly, oxygen evolution is selected as the process at the counter anode. Thus, all materials with a low oxygen overvoltage are favorable. These include IrO_2 on Ti, PbO_2 on Pb or Ti, RuO_2 on Ti (DSA), and carbon steel (mild steel) or iron oxide on iron (low cost alternative to DSA as developed for the adiponitrile process [357,358,359,360,361,362] as shown in Figure 74. Also the nonstoichiometric titanium oxide, Ebonex, has been proposed as counter anode (see page 96).

Another possibility especially when aprotic solvents have to be used is to choose electrode materials that dissolve in the process (sacrificial anodes like Al, Mg, Zn, Sn [420,421,422,423], [561], [562]) with or without the accompanying formation of organometallic intermediates. Such a process clearly adds to the cost of the process and is, therefore, most likely only justified for high value added conversions.

5.4.2.3. Other Factors Influenced by the Electrode Material

The electrode material can also influence the product distribution as shown in the Kolbe electrolysis of carboxylates. With platinum anodes, the Kolbe dimerization of the intermediate radicals predominates strongly. At carbon anodes, however, further oxidation to the carbenium ion (Non-Kolbe reaction or Hofer – Moest reaction) becomes the main pathway.

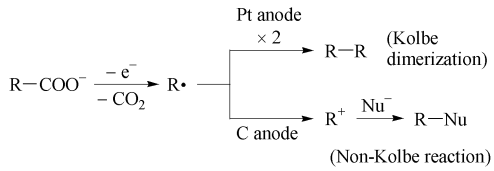

In the anodic olefin oxidation the electrode material plays an important role as well. The oxidation of enolethers at platinum anodes mainly leads to the dimethoxylation product [563] while the application of graphite anodes favors the additive dimerization [519].

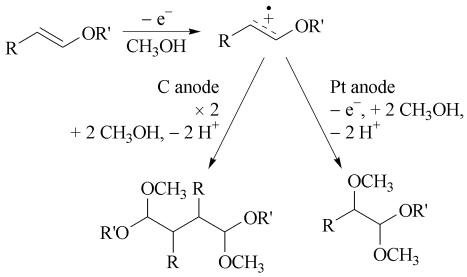

Another typical example can be found in the cathodic reduction of acetone, the outcome of which is dramatically changed when different electrode materials are used.

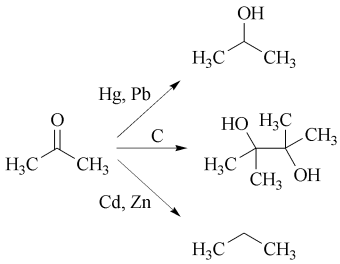

The adsorption of intermediates is also a function of the electrode material. All these effects of electrode materials will strongly influence the outcome of a reaction.

Three factors may influence the long-term behavior of electrodes:

1) Poisoning of the electrodes like, for example, deposition of traces of Fe, Cu or noble metals on cathodes of high hydrogen overpotential, thus lowering the potential of hydrogen evolution. This can be avoided by careful purification of the electrolyte.
2) Filming of the electrode by deposition of insulating polymeric byproducts. This may be avoided by interruption or reversal of the current, by mechanical or chemical cleaning of the electrode surface, or by introduction of a redox catalyst (mediator, indirect electrochemical process).
3) Corrosion of the electrodes. This may be avoided by proper selection of the electrode material, the electrolyte or by periodic replacement of the electrode stacks.

5.4.3. Electrochemical Cells

The electrochemical reactors are in the heart of an electrochemical process. The fundamentals of the cell design have been discussed in detail [346], [564,565,566,567,568] (see Chap. 3).

5.4.3.1. General Considerations

For larger scale application, especially in industry, it is necessary to lower the energy consumption as much as possible. The energy consumption (kW h/kg) is the product of the cell voltage U_C and the consumed charge:

$$\text{Energy Consumption} = (nF/3.6\,\Phi M)U_C\ 10^{-6}$$

where M is the molar mass of the product in kilograms per mole, Φ is the fractional current efficiency for the desired product, n is the number of electrons consumed for the formation of the product, and F is the Faraday constant in coulombs per mole.

The cell voltage (for a reduction) is given by the following equation:

$$U_C = E_C - E_A - U_\Omega$$

E_A and E_C are the active potentials at the electrodes (sum of the thermodynamic potential and the required activation overpotential), while U_Ω is the sum of the ohmic voltage drops IR at the resistances of the electrolyte, the diaphragm, and the circuit. The solution ohmic voltage drop is given by the equation

$$U_{\Omega\,\text{Soln}} = IR = r\,d\,j$$

Here, r is the resistivity of the electrolyte, d the interelectrode gap and j the current density (current divided by the electrode area).

Thus, for minimizing the energy consumption, it is mainly necessary to reduce the cell voltage. This can be done by the following measures:

1) To keep the $U_{\Omega\,\mathrm{Soln}}$ as low as possible, the interelectrode or electrode-membrane gap should be minimized. The electrode may even touch the membrane as in zero-gap or solid polymer electrolyte (SPE) cells. Because the resistivity r of organic electrolytes is much higher than that of inorganic electrolytes, the minimization of the ohmic drop is especially important in these cases. The electrolyte resistance is linearly dependent on the electrode distance d. Therefore, this distance should be as small as 0.1 – 1.0 mm and a diaphragm should be avoided (undivided cell), if possible. These conditions are best fulfilled by capillary gap cells with gaps of 0.1 – 0.2 mm. In this case, also the concentration of the supporting electrolyte can be very small.
2) The reaction at the counter electrode should be chosen in such a way that the reversible cell voltage is minimized (see Section 5.4.2).
3) The overpotentials at both electrodes should be reduced for example by using electrocatalysts or adequate electrode materials.
4) If possible, the cell should be undivided to minimize the construction cost and also the energy consumption due to the resistivity of the separator. The application of a controlled reaction at the auxiliary electrode taking place at low potential allows for the use of undivided cells in many cases. For oxidations, the cathodic process at the auxiliary electrode may be a proton reduction under formation of hydrogen or the reduction of oxygen. For reductions, the anodic process may be the formation of oxygen from water, for organic solvents the application of sacrificial anodes, or the oxidation of formate or oxalate to produce carbon dioxide [569] (see also Section 5.4.2). Besides this "least damage" approach, the possibility exists to carry out a synthesis at each electrode. This so-called "*paired electrosynthesis*" seems to be very attractive since the process cost can be spread over two products. However, it seems to be very difficult to find optimal conditions for both electrode processes. Also, work up is often more difficult and the two products are always formed in the same amount which may cause marketing problems. While many laboratory studies have been performed (for examples see Section 5.3) the commercial application has been established only recently. The problems which are involved are described by Baizer [447] for the paired synthesis of sorbitol (cathode) and gluconic acid (anode) from glucose. Other examples have been indicated in Tables 20, 21, 22, 23, 24, 25, 26, 27. A special case of the paired electrosynthesis is the formation of the same product at anode and cathode (200 % electrolysis cell). This principle can be accomplished by the direct or indirect electrochemical generation of hydrogen peroxide at the cathode which then is used as a chemical oxidant (oxidation at the cathode) forming the same product which is also obtained at the anode (Fig. 75). Several laboratory examples have been published [570, 571, 572] but practical applications have not yet been accomplished.
5) The electrode surface should be as large as possible to increase the space – time yield.

Besides the energy consumption, the electrolysis cell must be optimized for a high space – time yield or weight of product per unit time per unit volume of the reactors. It is given by the following equation:

$$STY = (M/Fn)\Phi\, j\, A/V$$

in which M is the molar mass of the product, n the total number of the exchanged electrons for the formation of the product, F the Faraday constant, j the current density, Φ the current efficiency, A the electrode area and V the cell volume. Thus, the space – time yield can be optimized by using a high area per unit cell volume (high surface area electrodes such as three-dimensional electrodes, capillary gap cells). To obtain a high current density, not only a high concentration of reactant is necessary but also an optimum mass transport. The latter may be obtained by using high flow rates, turbulence promoters, and/or electrode movement.

5.4.3.2. Cell Design

For laboratory cells, the minimization of the energy consumption and the optimization of the space – time yield are not equally as important. It is more important that the different reaction parameters like electrode materials, diaphragms,

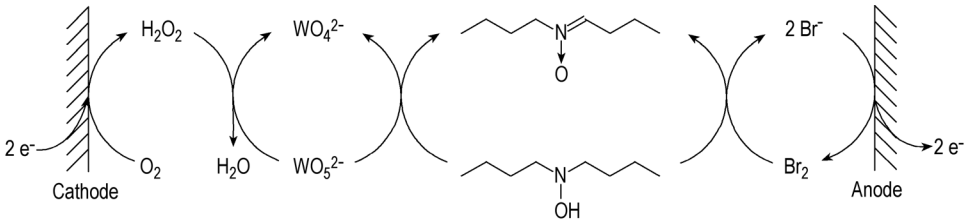

Figure 75. Paired electrosynthesis, special case of the formation of the same product at anode and cathode

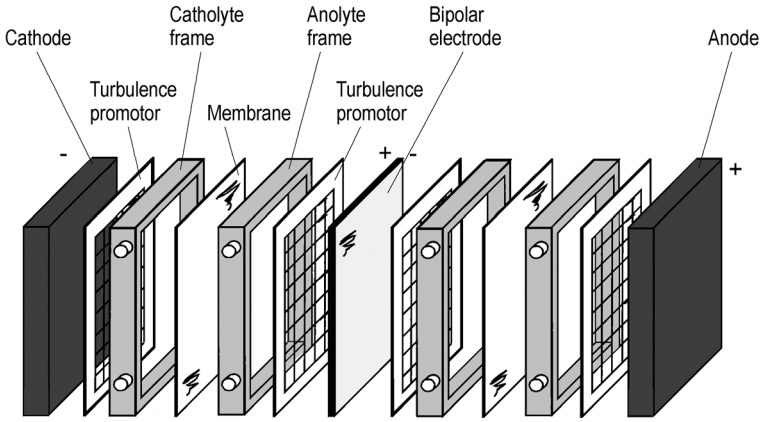

Figure 77. Schematic diagram of a parallel plate and frame cell with bipolar electrode arrangement

and the working potentials can be varied easily. For electrolyses under potential control, a three-electrode construction has to be used, which is schematically shown in Figure 76.

For technical applications, CPEs are undesirable because of technological problems, high cost and a nonstationary product formation due to changing current densities. Instead, in industry controlled current electrolyses using normal and less expensive d.c. power supplies are the only established methods. In this way, the current density can be fixed leading to a stationary formation of the product. The selectivity can be controlled by the proper selection of the current

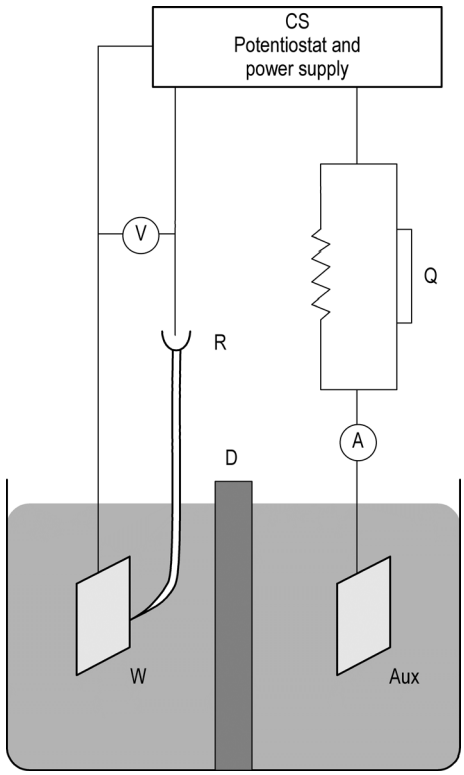

Figure 76. Schematic representation of an electrolysis cell (three-electrode construction)
W = working electrode; Aux = auxiliary or counter electrode; R = reference electrode; Q = coulombmeter; D = diaphragm, if necessary; CS = current source;

density and the substrate concentration which can be fixed to the desired value by using a continuous process. For this purpose, it is desirable that the whole electrode surface operates at the same optimum current density so that also an even potential distribution at the electrode surface is obtained. Thus, the whole electrode surface is productive and has the same rate of chemical change. At the same time, the formed intermediates find the same chemical environment for follow-up reactions. This condition is most easily fulfilled by using parallel plate cells which is therefore the most popular cell design. To obtain optimum space – time yield (see above), it is necessary to optimize the ratio of electrode area to cell volume. The three most important cell designs with parallel arrangement of the electrodes are arranged in the order of minimizing the interelectrode gap: the parallel plate and frame cells (Fig. 77); the capillary gap cell (Fig. 78) and the SPE cell, which is constructed like a proton exchange membrane fuel cell (PEMFC).

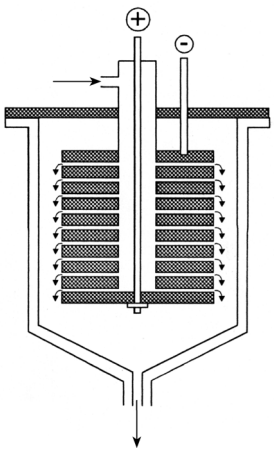

Figure 78. Capillary gap bipolar disc – stack cell

Parallel Plate and Frame Cells. The parallel plate and frame cells have interelectrode gaps of between 1 and 12 mm. The electrodes can be connected either in a monopolar or in a bipolar fashion (Fig. 79). For *monopolar connections*, the electrical contact has to be made to each electrode and thus both faces of the electrode will have the same polarity. The d.c. rectifier has to supply the cell voltage plus the cell current multiplied by the number of cells (interelectrode gaps). Therefore, a low voltage and a high current has to be supplied. In addition, as the current path through the electrodes is rather long, electrodes with a high conductivity have to be used. In *bipolar connections*, contact is only necessary to the end plates. The other electrodes are polarized in the electric field. The two faces of the electrodes, therefore, have different polarities. The d.c. rectifier now has to supply the cell current plus the cell voltage multiplied by the number of cells (interelectrode gaps). Thus, the current is used n times with n being the number of cells and, therefore, a lower current but a high voltage has to be supplied. As the current path through the electrodes is small, also less conducting electrode materials can be used.

Due to an active development since the late 1970s, reliable parallel plate and frame electrochemical cells of all sizes from laboratory scale to pilot scale up to production scale in divided or undivided form and with monopolar or bipolar connection are now commercially easily available. Suppliers are companies such as ElectroCell AB [573] (Micro Flow laboratory cell with 10 cm^2 electrode area for one electrode; MP Cell with 0.01 to 0.02 m^2; ElectroSyn Cell with 0.04 to 1.04 m^2; ElectroProd Cell with 0.4 to 16 m^2), ICI [574] (FM01-LC laboratory cell with 64 cm^2 for one electrode; FM21-SP Cell with 0.21 to 12.6 m^2 electrode area; FB cell family for bipolar electrode arrangements); Electrocatalytic Ltd. [575] (Dished Electrode Membrane Cell = DEM with single electrode areas of 0.05 m^2, 0.175 m^2, and 1.0 m^2) and Reilly Tar and Chemical [576]. Therefore, at least in the case of specialty chemicals or fine chemicals, the development of new cells for a process is not anymore necessary. The Mark II process for the formation of adiponitrile [361, 362, 363, 364, 365] from acrylonitrile is now performed in an undivided parallel plate cell with a cadmium plated steel cathode and an interelectrode gap of 2 mm.

Capillary Gap Cells. To reduce the interelectrode gap below 1 mm, undivided bipolar capillary gap cells are the best choice. In this way, poorly conducting solutions or low concentrations of the supporting electrolyte can easily be used. The most successful design has been developed by BASF (Fig. 78) and consists of a bipolar stack of horizontal circular rounds usually machined from graphite or graphite-filled

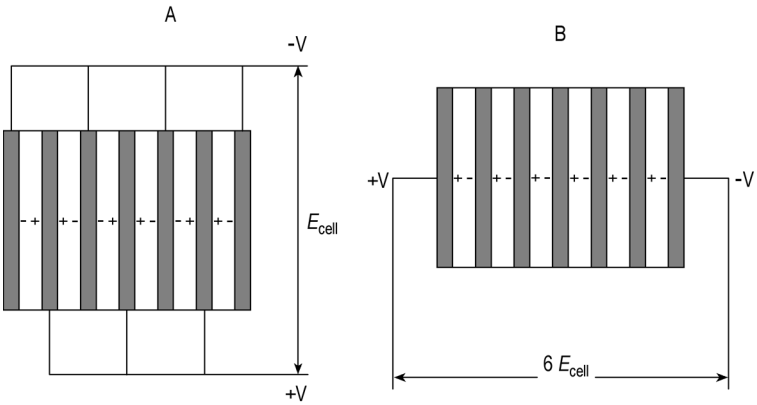

Figure 79. Electrical connection for monopolar (A) and bipolar (B) electrode arrangements

plastic discs separated by thin insulating spacers. The liquid is pumped radically in the 0.5 – 1 mm gaps [577,578,579]. Such cells are easily machined and are available in sizes from the laboratory scale up to the technical scale. The commercial production of benzaldehyde dimethylacetals from toluene derivatives and also the dimethoxylation of furan are performed in these cells at BASF. A concentric capillary gap cell made from a stainless steel tube (pipe) as cathode and an inner concentric graphite rod as anode was also developed [580]. It is very easy to construct but has a rather low space – time yield. Capillary gap cells can also be constructed in a parallel plate version using square vertical electrode plates [581].

Solid Polymer Electrolyte Cell. The SPE cell reduces the interelectrode gap even further. In this design, two porous electrode layers are in direct contact with a solid polymer electrolyte, usually an ion-exchange membrane like Nafion with a thickness of 20 to 50 µm. Thus, a supporting electrolyte is not required and product separation is made very easy. This concept has been suggested earlier [582], [583]. However scale-up and stability were problematic. More recently, it was shown that the pretreatment of the ion-exchange membrane is very important for the success of this cell design. For example, the pretreatment of Nafion membranes in DMF at elevated temperature leads to a successful application for the anodic α-methoxylation of DMF [584,585,586,587,588,589]. Besides amides, also furan could be methoxylated efficiently. Even cycloaliphatic ethers such as tetrahydrofuran can be methoxylated at a high current efficiency using only methanol as solvent [589]. The successful application of an electrode stack similar to that of the concentric capillary gap cell seems to open up the way to a technical application.

Cells with Sacrifical Anodes. A technical fenoprofen synthesis on the pilot scale was developed by SNPE based on the reductive carboxylation of chloroethyl diphenyl ether under cleavage of the carbon-halogen bond using a sacrificial magnesium anode [420,421,422,423], [561], [562]. For this purpose, a special cell for the application of the sacrificial anode was designed. This so-called "pencil-sharpener electrochemical cell" (Fig. 80) consists of a cylindrical magnesium bar of 40 cm diameter as anode. The cathode is a stainless steel cone which is 1 m high. The anode is pressed onto the conical cathode by its own weight. To prevent contact between the two electrode, the cathode surface is covered by a conical polyethylene grid. Due to the different current densities, the electroerosion is higher where the interelectrode gap is narrower. Thus, the anode is automatically electromachined into a pencil-like shape fitting closely to the cathode with an interelectrode gap of a few millimeters. The electrolyte is pumped between the two electrodes with the polyethylene grid acting as a turbulence promoter. The pilot unit has a cathode area of 1.2 m^2.

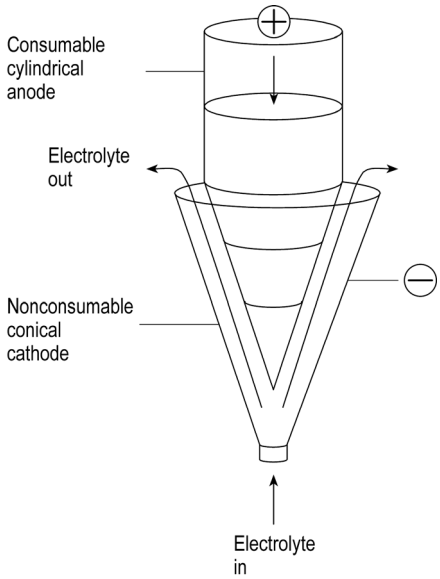

Figure 80. Pencil-sharpener electrolysis cell with sacrificial anode

Cells with Three-Dimensional Electrodes. If dilute solutions have to be electrolyzed at an acceptable rate or if reactions which show selectivity only at low current densities have to be carried out, three-dimensional electrodes must be considered. They give a high area:unit volume ratio under good mass transport conditions [590], [591]. However, the current distribution is not always homogeneous and has to be taken into account. Three-dimensional electrodes can be formed by using packed or fluidized beds of particulate electrodes or by using three-dimensional electrode materials such as carbon felts, reticulated materials, metal foams etc.

A special type of three-dimensional cell is called "Swiss Roll" cell. It has been developed for the oxidation of diacetone-L-sorbose to diacetone-2-keto-L-gulonic acid as a vitamin C precursor. It consists of a stainless steel net cathode and a nickel net anode separated by a spacer material such as a polypropylene net. The entire arrangement is rolled up and forced into a metal pipe (Fig. 81). Cells with a diameter of 1 m and a total anode and cathode area of 200 m^2 have been used in a pilot plant by Hoffmann-La Roche [469,470,471].

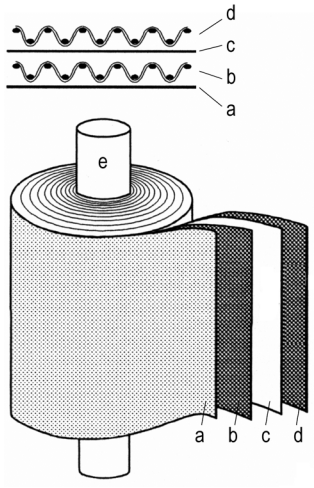

Figure 81. "Swiss-Roll Cell" for the application of NiOOH-electrodes (a = steel net cathode; b + d = polypropylene net; c = nickel net anode; e = current feeder)

Table 29. Cost of redox agents in \$/mol

Reagent	Cost, \$/mol
Electrons (0.06 \$/kWh · 3.5 V) [a] similar H_2, O_2	0.006
Hydrogen peroxide	0.038
Sodium metal	0.075
Magnesium metal	0.079
Zinc metal	0.086
Hydrazine	0.14
Sodium hydrosulfite	0.25
Sodium dichromate	0.39
Potassium permanganate	0.45
Sodium borohydride	1.70
Stannous chloride	2.80

[a] Estimated price of electricity for industry.

5.5. Economic Aspects

Electrons are by far the least expensive reagents. On a molar scale they are in the same low cost range as hydrogen or oxygen in an integrated chemical plant [592]. The costs of various reducing agents are compared in Table 29.

It is frequently said that the fact that only a relatively small number of electroorganic processes has been realized in industry is due to the high investment and energy costs. Degner [363] and others point out that this is a prejudice. For processes of up to 10 000 t/a, the energy costs for the electrolysis are in most cases between 6 and 15 % of the total manufacturing costs, in rare cases like in processes using low molecular mass starting materials, they can reach up

to 25 %. The investment for the electrochemical part of the equipment, usually, is only 30 to 40 % of the total investment, the proportion for the cell being only about 2 % of the total investment if an undivided cell is used and may increase to a maximum of 15 % for a divided cell.

The prerequisites for the industrial application of an electroorganic process are listed in Table 30 [363], [403]. However, these data may be too general. To evaluate the different cost factors for electroorganic processes more exactly, it is necessary to divide the products into three classes [367].

1) Commodity chemicals (above 10 000 t/a)
2) Fine chemicals (100 – 10 000 t/a)
3) Specialty chemicals (< 100 t/a)

As Figure 82 shows energy consumption is a dominant factor only in very large scale processes. In these cases, also the cost of electrodes and membranes (cell components), respectively their service lives and reliabilities are very important. Also, the cell costs are a contributing factor. In contrast, for specialty chemicals the cost of the starting material is dominating while the energy cost is almost negligible. Thus, the selectivity of the process is important. Therefore, even more costly cell components or their early replacement are easily tolerated if this helps to increase the selectivity. For fine chemicals the cost of the starting materials and the investment in cells is usually more important than the energy cost. In all cases, the ease of the product isolation and electrolyte recovery together with the nature and level of byproducts are critical.

Therefore, in contrast to some frequently heard conclusions, large scale electroorganic processes are not ruled out. To meet the requirements, however, is not easy. The Monsanto process for the electrohydrodimerization of acrylonitrile to adiponitrile demonstrates that it is possible. Another potential, large scale process which has been developed up to the pilot plant is the electrohydrodimerization of formaldehyde to give ethylene glycol. The economics of the chemical pathway via ethylene oxide and the electrochemical routes are similar and the preference for the chemical or the electrochemical pathway is mainly a result of the relative prices of the different feedstocks, ethylene or formaldehyde, which might differ from company to company.

It has been pointed out [592] that the ratio of the total manufacturing costs to the cost of the d.c. power of successful electrochemical processes relates to the plant capacity in tons per year in a somewhat hyperbolic fashion. Thus, for commodities in the range of 100 000 t/a the ratio of manufacturing costs to d.c. power costs is about 2, for 5000 t/a capacity it goes up to 3, for 1000 t/a to 4 and for 100 t/a it reaches 8.

The following factors influence the commercial success of an electrosynthetic process. Only some of the factors are inherent to the electrochemical methodology and they have been discussed in more detail above.

Factors which are common to all processes but are influenced by the electrochemical methodology:

1) The cost and availability of the starting materials
2) The sale price for the product
3) The market for the product
4) The patent situation
5) The performance of alternative processes
6) Safety considerations (hazards are usually low with electrochemical processes due to their mild conditions)
7) Environmental considerations (electrochemical processes are usually environmentally friendly)
8) The quality of the product depending on the nature and quantity of impurities (electrochemical processes often produce purer products)
9) The ease of product isolation (separation form supporting electrolyte sometimes causes problems)
10) The material yield and/or selectivity of the reaction

Factors which are inherent to the electrochemical methodology:

1) The energy consumption in kW h/t (optimization has been discussed in Section 5.4.3)
2) The cost and lifetime of the cells and cell components like electrodes and membranes
3) The strategy for the counter electrode reaction leading to a clean and balanced overall cell chemistry and also to lower the energy consumption
4) The current density which determines the space – time yield and thereby the plant production rate

Table 30. Preconditions for the industrial application of an electroorganic process [363], [592]

	1988	1995
High selectivity and product yields	> 80 %	no environmentally problematic byproducts
High current efficiencies	> 50 %	> 70 %
Energy (electricity) consumption per kilogram of desired product	< 8 kWh/kg of end product	
Concentration of the end product in the electrolyte	> 10 %	
Service life of the electrode	> 1000 h	> 4000 h
Membrane lifetime (divided cell)	> 2000 h	> 8000 h
Simple work-up of the electrolysis products		
Simple removal and recycling of the electrolyte		low emissions

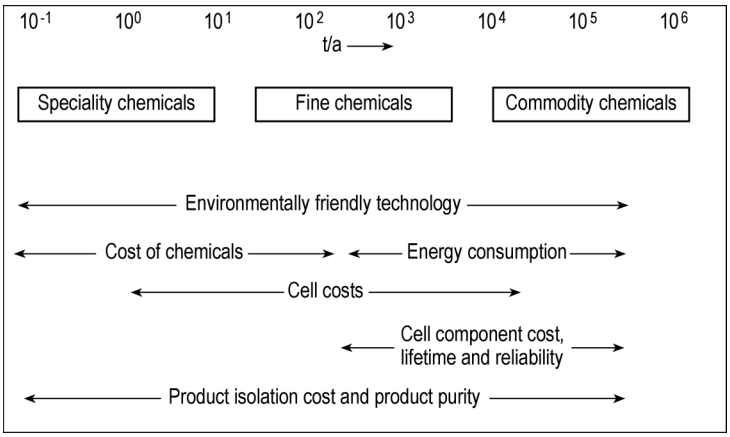

Figure 82. Importance of factors influencing the cost of electroorganic processes in dependence of production scale

5) The stability and reproducibility of the process (electrode poisoning and passivation, electrode corrosion)
6) Company experience of electrolytic processes
7) Company interest in performing related electrochemical reactions in the same plant

6. References

General References

1. G. Kreysa, H. Wendt: *Electrochemical Engineering,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1998.
2. F. Hine: *Electrode Processes and Electrochemical Engineering,* Plenum Press, New York-London 1985.
3. F. Coeuret, A. Storck: *Elements de Génie Electrochimique,* Tecdoc, Paris 1984.
4. J. Newman, W. Tiedemann, *Adv. Electrochem. Electrochem. Eng.* **11** (1978) 353 – 438.
5. E. Heitz, G. Kreysa: *Grundlagen der technischen Chemie,* Verlag Chemie, Weinheim-New York 1977, pp. 53 – 65. E. Heitz, G. Kreysa: *Principles of Electrochemical Engineering* , VCH Verlagsgesellschaft, Weinheim-New York 1986, pp. 53 – 66.
6. J. S. Newmann: *Electrochemical Systems,* Prentice Hall, Englewood Cliffs, N. Y., 1973.
7. K. Vetter: *Elektrochemische Kinetik,* Springer Verlag, Berlin-Göttingen-Heidelberg 1961.
8. A. J. Appleby: “Electrocatalysis,” *comp. Treatise Electrochem.* **7** (1983) 137–239.
9. S. Trasatti in ”Advances in Electrochemical Science and Technology” (H. Gerischer, W. Tobias, eds.), vol. 2, p. 1, VCH Verlagsgesellschaft, Weinheim, 1990.
10. S. Trasatti in H. Wendt (ed.): ”The oxygen evolution reaction” *Electrochemical hydrogen technologies,* Elsevier, Amserdam 1990.
11. H. Wendt, Th. Brenscheit, A. Fischer, *Phil. Trans. R. Soc.,London, A* **354** (1996) 1627 – 1641.
12. V. deNora in F. Hine et al. (eds.): *Proceedings, Symposium on Performance of Electrodes for Industrial Electrochemical Processes,*

PV-89–10, Electrochem. Society, Pennington 1989, p. 1.
13. F. Hine et al. (eds.): *Proceedings, Symposium on Performance of Electrodes for Industrial Electrochemical Processes,* **PV-89–10**, Electrochem. Society, Pennington 1989, p. 1.
14. S. Trasatti, *Electrochim. Acta* **29** 1503 (1984)369.
15. S. Trasatti, *Electrochim. Acta* **29** 1503 (1984)
16. A. I. Krasil'schchiv, *Z.h. Fiz. Khim.* **37** (1963) 531.
17. E. Justi, M. Pilkuhn, W. Scheibe A. Winsel, *Abhdlg. Math. Nat. Klasse* **8** (1959).
18. K. Kinoshita: *Carbon–Electrochemical and Physicochemical Properties,* John Wiley & Sons, New York 1987.

Specific References
19. H. Wendt, H. Hofmann, H. Berg, V. Plzak, J. Fischer : "Alkaline Water Electrolysis at Enhanced Temperatures; Basic and Material Studies, Engineering and Economies," in G. Imarisio, A. S. Strub (eds.): *Hydrogen as an Energy Carrier,* D. Riedel Pub., Dordrecht-Lancaster-Boston 1983, pp. 267 –285.
20. H. Wendt, *Ber. Bunsenges. Phys. Chem.* **81** (1977) 864 – 869.
21. N. Ibl, *Chem. Ing. Tech.* **35** (1963) 353 – 361.

Specific References
22. N. Ibl, E. Adam, *Chem. Ing. Tech.* **37** (1965) 573 – 581.
23. S. van Stralen, R. Cole: *Boiling Phenomena,* vol. 1, Hemisphere Publ., Washington 1980.
24. H. Vogt in E. Yeager, J. O. M. Bockris, B. E. Conway, S. Sarangapani (eds.): *Comprehensive Treatise of Electrochemistry,* **vol. 6**, Plenum Press, New York 1983, pp. 455 – 489.
25. H. Vogt, "Elektrochemische Reaktoren mit Gasentwicklung," *Fortschr. Verfahrenstech.* **16** (1978) 297 – 332.
26. H. Vogt, "Electrochemical Reactors with Gas Evolution," *Fortschr. Verfahrenstech.* **20** (1982) 369 – 404.
27. D. E. Westerheide, J. W. Westwater, *AIChE J.* **7** (1961) 357 – 362.
28. L. J. J. Janssen, C. W. M. P. Sillen, E. Barendrecht, S. J. D. van Stralen, *Electrochim. Acta* **29** (1984) 633 – 642.
29. A. Coehn, *Z. Elektrochem. Angew. Phys. Chem.* **29** (1923) 1 – 5.
30. F. A. Lewis, R. C. Johnston, M. C. Witherspoon, W. F. N. Leitch, S. G. Thompson, A. Obermann, S. F. Deane, *Surf. Technol.* **4** (1976) 89 – 94.
31. H. Vogt, *Electrochim. Acta* **29** (1984) 167 – 173, 175 – 180.
32. L. Müller, M. Krenz, R. Landsberg, *J. Electroanal. Chem.* **180** (1984) 453 – 462.
33. H. Vogt, *Electrochim. Acta* **30** (1985) 265 – 270.
34. B. Krause, H. Vogt, *J. Appl. Electrochem.* **15** (1985) 509 – 515.
35. L. J. J. Janssen, E. Barendrecht, *J. Appl. Electrochem.* **15** (1985) 549 – 555.
36. H. Vogt, *Int. J. Heat Mass Transfer* **36** (1993) 4115 – 4121.
37. H. Vogt, *Electrochim. Acta* **38** (1993) 1427 – 1431.
38. H. Vogt, *Electrochim. Acta* **23** (1978) 203 – 205.
39. B. Roald, W. Beck, *J. Electrochem. Soc.* **98** (1951) 277 – 290.
40. G. Bendrich, W. Seiler, H. Vogt, *Int. J. Heat Mass Transfer* **29** (1986) 1741 – 1745.
41. H. Vogt, *Electrochim. Acta* **32** (1987) 633 – 636.
42. N. Ibl, J. Venczel, *Metalloberfläche* **24** (1970) 365 – 374.
43. H. Vogt, *Electrochim. Acta* **25** (1980) 527 – 531.
44. K. Stephan, H. Vogt, *Electrochim Acta* **24** (1979) 11 – 18
45. L. J. J. Janssen, E. Barendrecht, *Electrochim. Acta* **24** (1979) 693 – 699.
46. L. J. J. Janssen, E. Barendrecht, *Electrochim. Acta* **30** (1985) 683 – 694.
47. H. Vogt, *Electrochim. Acta* **23** (1978) 1019 – 1022.
48. C. J. Maxwell: *A Treatise on Electricity and Magnetism,* 2nd ed., **vol. 1,** Clarendon Press, Oxford 1881, p. 435.
49. Lord Rayleigh, *Philos. Mag.* **34** (1892) 481 – 502.
50. D. A. G. Bruggeman, *Ann. Phys. (Leipzig)* **24** (1935) 636 – 664.
51. S. Prager, *Physica (Amsterdam)* **29** (1963) 129 – 139.
52. H. Behringer, *Z. gesamte Kälte Ind.* **43** (1936) 55 –58.
53. N. Zuber, *Chem. Eng. Sci.* **19** (1964) 897 – 917.
54. J. F. Richardson, W. N. Zaki, *Trans. Inst. Chem. Eng.* **32** (1954) 35 – 53.
55. V. G. Levich: *Physicochemical Hydrodynamics,* Prentice-Hall, Englewood Cliffs, N. Y., 1962.

56. G. Kreysa, M. Kuhn, *J. Appl. Electrochem.* **15** (1985) 517 – 526.
57. H. Vogt, *J. Appl. Electrochem.* **17** (1987) 419 – 426.
58. I. Roušar, *J. Electrochem. Soc.* **116** (1969) 676 – 683.
59. I. Roušar, A. Regner, V. Cezner, *Collect. Czech. Chem. Commun.* **31** (1966) 4193 – 4208.
60. I. Roušar, V. Cezner, *Collect. Czech. Chem. Commun.* **32** (1967) 1137 – 1151.
61. I. Roušar, V. Cezner, J. Hostomský, *Collect. Czech. Chem. Commun.* **33** (1968) 808 – 828.
62. I. Roušar, V. Cezner, J. Nejepsová, M. M. Jakšić, M. Spasojević, B. Z. Nikolić, *J. Appl. Electrochem.* **7** (1977) 427 – 435.
63. J. E. Funk, J. F. Thorpe, *J. Electrochem. Soc.* **116** (1969) 48 – 54.
64. J. Hopenfeld, R. R. Cole, *J. Eng. Ind. – Trans. ASME, Ser. B* **91** (1969) 755 – 765.
65. C. W. Tobias, *J. Electrochem. Soc.* **106** (1959) 833 – 838.
66. P. J. Sides, C. W. Tobias, *J. Electrochem. Soc.* **129** (1982) 2715 – 2720.
67. O. Lanzi, R. F. Savinell, *J. Electrochem. Soc.* **130** (1983) 799 – 802.
68. P. Gallone, G. Modica, S. Maffi, *J. Electroanal. Chem.* **180** (1984) 421 – 431.
69. F. Hine, M. Yasuda, R. Nakamura, T. Noda, *J. Electrochem. Soc.* **122** (1975) 1185 – 1190.
70. H. Vogt, *Electrochim. Acta* **26** (1981) 1311 – 1317; **27** (1982) 1157.
71. B. E. Bongenaar-Schlenter, L. J. J. Janssen, S. J. D. van Stralen, E. Barendrecht, *J. Appl. Electrochem.* **15** (1985) 537 – 548.
72. M. Schleiff, W. Thiele, H. Matschiner, *Chem. Technik (Leipzig)* **34** (1982) 18 – 22.
73. M. Schleiff, W. Thiele, H. Matschiner, *Chem. Technik (Leipzig)* **34** (1982) 250 – 252.
74. W. Thiele, M. Schleiff, H. Matschiner, *Chem. Technik (Leipzig)* **34** (1982) 576 – 579.
75. W. Thiele, M. Schleiff, *Chem. Technik (Leipzig)* **36** (1984) 95 – 97.
76. I. Roušar, *Electrochim. Acta* **31** (1986).
77. *Ullmann,* 3rd ed., **5,** p. 535.
78. F. Hine, K. Murakami, *J. Electrochem. Soc.* **128** (1981) 64 – 68.
79. A. T. Kuhn, J. Bin Yusof, P. Hogan, *J. Appl. Electrochem.* **9** (1979) 765 – 775.
80. J. Fischer, H. Hofmann, G. Luft, H. Wendt, *AIChE J.* **26** (1980) 794 – 802.
81. O. De Nora, *Chem. Ing. Tech.* **42** (1970) 222 – 226.
82. K. R. Koziol et al., *Chem. Ing. Tech.* **49** (1977) 288 – 293.
83. J.-W. Kühn-von Burgsdorff, *Chem. Ing. Tech.* **49** (1977) 294 – 298.
84. J. Jorne, J. F. Louvar, *J. Electrochem. Soc.* **127** (1980) 298 – 303.
85. C. Elsner, F. Coeuret, *J. Appl. Electrochem.* **15** (1985) 567 – 574.
86. L. J. J. Janssen, J. J. M. Geraets, E. Barendrecht, S. J. D. van Stralen, *Electrochim. Acta* **27** (1982) 1207 – 1218.
87. G. Kreysa, H.-J. Külps, *J. Electrochem. Soc.* **128** (1981) 979 – 984.
88. V. L. Kubasov, G. I. Volkov, *Sov. Electrochem. (Engl. Transl.)* **2** (1966) 665.
89. F. N. Ngoya, *Electrochim. Acta* **28** (1983) 1865 –1868.
90. W. E. Haupin, *J. Met.* **23** (1971) no. 10, 46 – 49.
91. G. M. Tkachenko, N. S. Siraev, *Tsvetn. Met. (Moscow)* **50** (1977) no. 5, 33 – 35; *Sov. J. Non-Ferrous Met. (Engl. Transl.)* **18** (1977) no. 5, 40 – 41.
92. J. Müller, *Chem. Ing. Tech.* **49** (1977) 326 – 327.
93. R. Dworak, H. Wendt, *Chem. Ing. Tech.* **49** (1977) 349.
94. Alcoa, US 3 755 099, 1973 (W. E. Haupin).
95. S. Okada, S. Yoshizawa, F. Hine, Z. Takehara, *J. Electrochem. Soc. Japan* **26** (1958) E66 – E69.
96. P. A. Danna, *J. Electrochem. Soc.* **121** (1974) 1286 – 1289.
97. A. Nidola in S. Trasatti (ed.): *Electrodes of Conductive Metallic Oxides,* part B., Elsevier, Amsterdam 1981.
98. O. De Nora, *Chem. Ing. Tech.* **43** (1971) 182 – 184.
99. M. D. Birkett, A. T. Kuhn, *Electrochim. Acta* **27** (1982) 263 – 268.
100. A. T. Kuhn (ed.): *Industrial Electrochemical Processes,* Elsevier, Amsterdam 1971.
101. W. Vielstich, *Chem. Ing. Tech.* **33** (1961) 75 – 79.
102. A. Winsel, *DECHEMA Monogr.* **92** (1982).
103. P. Brennecke, H. Ewe, *Chem. Ing. Tech.* **52** (1980) 426 – 428.
104. G. Kreysa, *J. Appl. Electrochem.* **15** (1985) 175 –179.
105. D. Hutin, A. Storck, F. Coeuret, *Inf. Chim.* **189** (1979) 227 – 235.
106. A. Storck, F. Coeuret, *Electrochim. Acta* **22** (1977) 1155 – 1160.
107. D. J. Pickett, B. Stanmore, *J. Appl. Electrochem.* **2** (1972) 151 – 156.
108. D. R. Gabe, F. C. Walsh, *J. Appl. Electrochem.* **13** (1983) 3 – 22.

109. D. R. Gabe, *J. Appl. Electrochem.* **4** (1974) 91 – 108.
110. M. Fleischmann, R. E. W. Jansson, *Chem. Eng. (London)* 1975, no. 302, 603 – 606.
111. R. E. W. Janssen, N. R. Tomov, *Electrochim. Acta* **25** (1980) 497 – 503.
112. C. Oloman, *J. Electrochem. Soc.* **126** (1979) 1885 – 1892.
113. F. Beck, H. Guthke, *Chem. Ing. Tech.* **41** (1969) 943 – 950.
114. R. Dworak, H. Wendt, *Ber. Bunsenges. Phys. Chem.* **80** (1976) 77 – 82.
115. G. A. Ashworth, R. E. W. Jansson, *Electrochim. Acta* **22** (1977) 1295 – 1299.
116. R. E. W. Jansson, R. J. Marshall, *Chem. Eng. (London)* 1976, no. 315, 769 – 772.
117. R. E. W. Jansson, N. R. Tomov, *Chem. Eng. (London)* 1977, no. 327, 867 – 869.
118. R. E. W. Jansson, G. A. Ashworth, *Electrochim. Acta* **22** (1977) 1301 – 1306.
119. P. M. Robertson, F. Schwager, N. Ibl, *J. Electroanal. Chem.* **65** (1975) 883 – 900.
120. P. M. Robertson, N. Ibl, *J. Appl. Electrochem.* **7** (1977) 323 – 330.
121. P. M. Robertson, *Electrochim. Acta* **22** (1977) 411 – 419.
122. J. Fischer, H. Hofmann, G. Luft, H. Wendt, *AIChE J.* **26** (1980) 794 – 802.
123. J. Divisek, H. Schmitz, J. Mergel, *Chem. Ing. Tech.* **52** (1980) 465.
124. S. Stucki, A. Menth, *Ber. Bunsenges. Phys. Chem.* **84** (1980) 1008 – 1013.
125. J. S. Newman, C. W. Tobias, *J. Electrochem. Soc.* **109** (1962) 1183 – 1191.
126. J. R. Backhurst, J. M. Coulson, F. Goodridge, R. E. Plimley, M. Fleischmann, *J. Electrochem. Soc.* **116** (1969) 1600 – 1607.
127. J. R. Backhurst, M. Fleischmann, F. Goodridge, R. E. Plimley, GB 1 194 181, 1970.
128. D. N. Bennion, J. Newman, *J. Appl. Electrochem.* **2** (1972) 113 – 122.
129. G. Kreysa, E. Heitz, *Chem. Ing. Tech.* **48** (1976) 852 – 860.
130. G. Kreysa, *Chem. Ing. Tech.* **55** (1983) 23 – 30.
131. G. Kreysa, *Electrochim. Acta* **23** (1978) 1351 – 1359.
132. S. Germain, F. Goodridge, *Electrochim. Acta* **21** (1976) 545 – 550.
133. H.-D. Steppke, R. Kammel, *Erzmetall* **26** (1973) 533 – 538.
134. A. J. Monhemius, P. L. N. Costa, *Hydrometallurgy* **1** (1975) 183 – 203.
135. G. Kreysa, V. Gundelach, G. Linzbach, *Chem. Ing. Tech.* **55** (1983) 952 – 958.
136. G. Kreysa, *Chem. Ing. Tech.* **50** (1978) 332 – 337.
137. G. Kreysa, *Metalloberfläche* **34** (1980) 494 – 501.
138. G. Kreysa, C. Reynvaan, *J. Appl. Electrochem.* **12** (1982) 241 – 251.
139. P. S. Fedkiw, P. Satemazandarani, *Chem. Eng. Commun.* **36** (1985) 27 – 43.
140. H. Scharf, DE 2 227 084, 1972.
141. C. Raats, H. Boon, W. Eveleens, *Erzmetall* **30** (1977) 365 – 369.
142. G. v. Heiden, C. Raats, H. Boon, *Chem. Ind. (London)* 1978, 465 – 468.
143. I. Colquhoun-Lee, J. Stepanek, *Chem. Eng. (London)* 1974, no. 282, 108 – 111.
144. G. A. Hughmark, *AIChE J.* **18** (1972) 1020 – 1024.
145. J. C. Chu, J. Kalil, W. A. Wetteroth, *Chem. Eng. Prog.* **49** (1953) 141 – 149.
146. D. J. Pickett, *J. Appl. Electrochem.* **5** (1975) 101 –103.
147. M. Fleischmann, J. W. Oldfield, *J. Electroanal. Chem.* **29** (1971) 211 – 230.
148. J. Newman, W. Tiedemann, *AIChE J.* **21** (1975) 25 – 41.
149. M. Fleischmann, J. W. Oldfield, L. Tennakoon, *J. Appl. Electrochem.* **1** (1971) 103 – 112.
150. A. A. C. M. Beenackers, W. P. M. van Swaaij, A. Welmers, *Electrochim. Acta* **22** (1977) 1277 – 1281.
151. G. Kreysa, *Electrochim. Acta* **25** (1980) 813 – 818.
152. G. Linzbach, Ph. D. Thesis, University of Dortmund, 1985.
153. R. E. Sioda, G. M. Cook, K. B. Keating, J. M. Williams, G. H. Kelsall, J. L. Anderson; *Chem. Eng.* (*N.Y.*) **90** (1983) 57 – 67.
154. I. F. Masterson, J. W. Evans, *Metall. Trans.* **13B** (1982) 3 – 13.
155. D. S. Flett, *Chem. Ind. (London)* 1972, 983 – 988.
156. J. A. E. Wilkinson, K. P. Haines, *Trans. Inst. Min. Metall. Sect. C* **81** (1972) 157 – 162.
157. F. Goodridge, C. J. Vance, *Electrochim. Acta* **22** (1977) 1073 – 1076.
158. R. L. LeRoy, *Electrochim. Acta* **23** (1978) 815 – 825.
159. R. L. LeRoy, *Electrochim. Acta* **23** (1978) 827 – 834.
160. G. Kreysa, *Electrochim. Acta* **26** (1981) 1693 – 1694.
161. G. Bining, H. Rohrer, C. Gerber, E. Weibel, *Phys. Rev. Lett.* **49** (1982) 57.
162. G. Bining, H. Rohrer, *Helvetica Physica Acta* **55** (1982) 726.

163. D. M. Eigler, E. K. Schweizer, *Nature* **344** (1990) 524.
164. M. F. Crommie, C. P. Lutz, D. M. Eigler, *Science* **262** (1992) 218.
165. W. Li, J. A. Virtanen, R. M. Penner, *Appl. Phys. Lett.* **60** (1992) 1181.
166. W. Li, J. A. Virtanen, R. M. Penner, *J. Phys. Chem.* **96** (1992) 6529.
167. R. Ullmann, T. Will, D. M. Kolb, *Chem. Phys. Lett.* **209** (1993) 238.
168. R. Ullmann, T. Will, D. M. Kolb, *Ber. Bunsenges. Phys. Chem.* **99** (1995) 1414.
169. D. M. Kolb, R. Ullmann, T. Will, *Science* **275** (1997) 1097.
170. U. Landman, W. D. Luedtke, N. A. Burnham, R. J. Colton, *Science* **248** (1990) 454.
171. D. M. Kolb, R. Ullmann, J. C. Ziegler, *Electrochim. Acta* **43** (1998)2751.
172. G. E. Engelmann, J. C. Ziegler, D. M. Kolb, *Surf. Sci,* **401** (1998) L 420.
173. G. E. Engelmann, J. C. Ziegler, D. M. Kolb, *J. Electrochem. Soc.* **45** (1998) L33.
174. J. C. Ziegler, G. E. Engelmann, D. M. Kolb, *Z. Phys. Chem.,* **208** (1999) 151.

Specific References

175. H. B. Beer, *J. Electrochem. Soc.* **127** (1980) 303 C – 305 C.
176. D. L. Caldwell, *Compr. Treatise Electrochem.* **2** (1981).
177. E. Yeager, P. Bindra, *Chem. Ing. Tech.* **52** (1980) 384 – 391.
178. H. Wendt, *Chem. Ing. Tech.* **56** (1984) 265 – 272.
179. J. Müller, K. Lohrberg, H. Wüllenweber, *Chem. Ing. Tech.* **52** (1980) 435 – 436.
180. T. A. Liederbach, A. M. Greenberg, V. H. Thomas: "Commercial Application of Cathode Coatings in Electrolytic Chlorine Cells," in M. O. Coulter (ed.): *Modern Chloralkali Technology,* Ellis Horwood Ltd. 1980, pp. 145 – 149.
181. D. Dobos: *Electrochemical Data,* Elsevier, Amsterdam-Oxford-New York 1975, p. 247.
182. V. A. Ettel, B. V. Tilak: "Electrolytic Refining and Winning of Metals," *Compr. Treatise Electrochem.* **2** (1981) 327 – 380.
183. A. Schmidt: *Angewandte Elektrochemie,* Verlag Chemie, Weinheim 1976, p. 207.
184. *Elektrolyse der Nichteisenmetalle,* Verlag Chemie, Weinheim-New York 1986.
185. W. E. Haupin, W. B. Frank: "Electrometallurgy of Aluminum," *Compr. Treatise Electrochem.* **2** (1981) 301 – 325.
186. K. Grjotheim, C. Krohn, M. Malinovsky, K. Matiasovsky, J. Thonstad: *Aluminium Electrolysis,* 2nd ed., Aluminium Verlag, Düsseldorf 1982, pp. 28 – 29.
187. R. Bunsen, *Poggendorf Annalen* **92** (1954) 648 –651.
188. H. Wendt, K. Reuhl, *Aluminium (Düsseldorf)* **61** (1985) 518 – 521, 592 – 597.
189. W. E. Cowley: "The Alkali Metals," in D. Lovering (ed.): *Molten Salt Technology,* Plenum Press, New York 1982, pp. 57 – 90.
190. L. Stieglitz, R. Becker, KFK report **3740** (1984).
191. Z. Kolarik, H.-J. Bleyl, H. Schmieder, Proceedings of International Conference on 'Solvent Extraction', ISEC'86, Munich (1986).
192. C. S. Schlea, M. R. Caverny, H. E. Henry, W. J. Jenkins, USAEC-Report, DP808 (1963).
193. H. A. C. McKay, R. J. W. Streeton, A. G. Wain, UKAEA report, AERE-R 4381 (1963).
194. M. J. McKibber, J. E. Bercaw, USAEC report, DP1284 (1970).
195. H. Schmieder, F. Baumgärtner, H. Goldacker, H. Hausberger, KFK report **2082** (1974).
196. F. Baumgärtner, E. Schwind, P. Schlosser, *German Patent 1905 51* (1970).
197. A. Schneider, A. L. Ayers, *US patent 3616 276* (1971).
198. M. Krumpelt, J. Heiberger, M. J. Steindler, USAEC report ANL-7799 (1971).
199. J. P. Charvillat, J. J. Fabre, M. Le Bouhellec, M. Henry, Proc. Extraction'84, Dounreay, 27 – 29 Nov. (1984), EFCE Publ., series no. **43**.
200. G. Petrich, H. Schmieder, Proceedings of the International Conference on 'Solvent Extraction', ISEC'83, Denver, Co. Aug. (1983).
201. F. Baumgärtner, H. Schmieder, *Radiochim. Acta* **25** (1978) 191 – 220.
202. B. McDuffie, C. N. Reilly, *Anal. Chem.* **38** (1966) 1881.
203. H. Schmieder, H. Goldacker, M. Heilgeist, G. Petrich, *Fast Reactor Fuel Cycle,* **vol. 47,** British Energy Society, London 1981.
204. H. Schmieder, Ph. D. Thesis, Institut National Polytechnique de Grenoble (1984).
205. H. Schmieder, H. Goldacker, M. Heilgeist, M. Kluth, H. Hausberger, L. Finsterwalder, KFK report **2957** (1980).
206. H. Goldacker, H. Hausberger, H. Schmieder, H. Wiese, Proc. ANS-Conf., Sept. (1980), Gatlinburg, USA.
207. H. Schmieder, H. Goldacker, Proceedings of the International Conference on 'Solvent Extraction', ISEC'86, Munich, Sept. (1986).

208. U. Galla, H. Schmieder, KFK report **4177** (1987).
209. U. Galla, H. Goldacker, R. Schlenker, H. Schmieder, *Separa. Sci. & Technol.* **25** (1990) 1751.
210. G. Petrich, H. Schmieder, Proceedings of the International Conference on 'Solvent Extraction', ISEC'88, Moscow, July (1988).
211. M. Heilgeist, K. Flory, U. Galla, H. Schmieder, Proceedings of the International Conference on 'Solvent Extraction', ISEC'86, Munich, Sept. (1986).
212. M. Heilgeist, KFK report **3517** (1983).
213. V. P. Caracciolo, A. A. Kishbaugh, USAEC report, DP-896 (1964).
214. H. Schmieder, H. Goldacker, G. Petrich, *DECHEMA Monogr.* **94** (1983).
215. E. K. Dukes, *J. Am. Chem. Soc.* **82** (1960) 9.
216. W. Ochsenfeld et al., KFK-911, 1970
217. M. Heilgeist, Dissertation, University of Heidelberg, 1982.
218. G. Kreysa, in S. Stucki (ed.): *Process Technologies for Water Treatment,* Plenum Publishing Corporation, 1988, p. 65.
219. G. Kreysa, K. Jüttner, in F. Lapique, A. Storck, A. A. Wragg (eds.): *Electrochemical Engineering and Energy,* Plenum Press, New York, 1994.
220. K. J. Müller, G. Kreysa, *Dechema Monographien* **98** (1985) 367.
221. B. von Felten, *Oberfläche Surface* (1990) no. 6, 8.
222. A. T. Kuhn, *Chem. Ind. (London)* 1971, 946 – 950.
223. D. S. Flett, D. Pearson, *Chem. Ind. (London)* 1975, 639 – 645.
224. R. Kammel, H.-W. Lieber, *Galvanotechnik* **68** (1977) 883 – 886.
225. R. Kammel, H.-W. Lieber, *Galvanotechnik* **69** (1978) 317 – 324, 624 – 630.
226. A. T. Kuhn, *Chem. Ind. (London)* 1978, 447 – 453.
227. G. Kreysa, *Chem. Ing. Tech.* **50** (1978) 332 – 337.
228. G. Kreysa, *Metalloberfläche* **34** (1980) 494 – 501.
229. C. Fabjan, *Oberfläche-Surf.* **21** (1980) 283 – 292.
230. W. Samhaber, *Chem. Ing. Tech.* **56** (1984) 246 – 247.
231. G. Kreysa, *Electrochim. Acta* **26** (1981) 1693.
232. C. L. Lopez-Cacicedo, *Inst. Chem. Eng. Symp. Ser.* **42** (1975) 29 – 31.
233. D. R. Gabe, *J. Appl. Electrochem.* **4** (1974) 91.
234. L. J. Ricci, *Chem. Eng.* (1975) 29.
235. R. Kammel, H.-W. Lieber, *Galvanotechnik* **68** (1977) 710.
236. W. Götzelmann, *Galvanotechnik* **68** (1977) 789.
237. EPA 86109265.8, 1986 (D. Bruhn, W. Dietz, K.-J. Müller, C. Reynvaan).
238. A. Storck, P. M. Robertson, N. Ibl, *Electrochim. Acta* **24** (1979) 373 – 380.
239. P. M. Robertson, B. Scholder, G. Theis, N. Ibl, *Chem. Ind. (London)* 1978, 459 – 465.
240. K. B. Keating, J. M. Williams, *Resour. Recovery Conserv.* **2** (1976) 39 – 50.
241. D. N. Bennion, J. Newman, *J. Appl. Electrochem.* **2** (1972) 113.
242. D. Pletcher, F. Walsh (eds.): *Industrial Electrochemistry,* Chapman and Hall, London – New York, 1990.
243. G. Kreysa, *Chem.-Ing.-Tech.* **55** (1983) 23.
244. G. Kreysa, C. Reynvaan, *J. Appl. Electrochem.* **12** (1982) 241.
245. DE 2 622 497, 1976 (G. Kreysa).
246. J. R. Backhurst, J. M. Coulson, F. Goodbridge, R. E. Plimley, M. Fleischmann, *J. Electrochem. Soc.* **116** (1969) 1600.
247. GB 1 194 181, 1970 (J. R. Backhurst, M. Fleischmann, F. Goodbridge, R. E. Plimley).
248. DE 2 227 084, 1972 (H. Scharf).
249. C. Raats, H. Boon, W. Eveleens, *Erzmetall* **30** (1977) 365.
250. G. v. Heiden, C. Raats, H. Boon, *Chem. Ind.* (1978) 465.
251. R. Kammel, H.-W. Lieber, *Galvanotechnik* **69** (1978) 687 – 693.
252. W. Götzelmann, *Galvanotechnik* **70** (1979) 596 – 603.
253. C. L. Lopez-Cacicedo, GB 1 423 369, 1973.
254. D. R. Gabe, F. C. Walsh, *Trans. IchemE* **68 B** (1990) 107.
255. D. S. Flett, *Chem. Ind.* (1971) 300.
256. G. Kreysa, *Erzmetall* **28** (1975) 440.
257. K. Scott, *J. Appl. Electrochem.* **18** (1988) 504.
258. H. Bergmann, H. Hertwig, F. Nieber, *Chem. Eng. Proc.* **31** (1992) 195.
259. A. A. Wragg, F. N. Bravo de Nahui, in Electrochemical Engineering and the Environment, *IChemE Symposium Series* No. 127, p. 141, Rugby, UK, 1992.
260. G. P. Power, I. M. Ritchie, in B. E. Conway, J. O'M Bockris (eds.): *Modern Aspects of Electrochemistry,* **No. 11,** Prentice Hall, New York 1975, p. 199.
261. R. S. Rickard, M. C. Fuerstenau, *Trans A.I.M.E.* **242** (1968) 1485.
262. Ch. Schlimm, E. Heitz, *Environmental Progress* **15** (1996) 37.

263. L. Hartinger: *Taschenbuch der Abwasserbehandlung,* **vol. 1,** Hanser Verlag, München 1976, pp. 54 – 57.
264. R. Weiner: *Die Abwässer der Galvanotechnik und Metallindustrie,* 4th ed., Eugen G. Leuze Verlag, Saulgau 1973, pp. 301 – 305.
265. F. Nohse : "Elektrodialyse," in L. Hartinger (ed.): *Taschenbuch der Abwasserbehandlung,* vol. 2, Carl Hanser Verlag, München 1977, pp. 165 – 175.
266. H. Behret, H. Binder, R. Eggersdorf, H. J. Hampel. A. Köhling, *Ber. Bunsenges. Phys. Chem.* **83** (1979) 1094 – 1097.
267. H. Binder, H. Behret, *DECHEMA Monogr.* **97** (1984) 289 – 303.
268. F. Linnhoff, *DECHEMA Monogr.* **97** (1984) 305 –307.
269. S. Itoi, J. Nakamura, T. Kawahara, *Desalination* **32** (1980) 383 – 389.
270. D. C. Johnson, N. Popović, J. Feng, L. L. Houk, K. T. Kawagoe, The Electrochem. Society, Proc. Volume 95 – 26, 176 – 188 (1996).
271. Ch. Comminellis, C. Pulgarin, *J. Appl. Electrochemistry* **21** (1991) 703.
272. Ch. Comminellis, C. Pulgarin, *J. Appl. Electrochem.* **23** (1993) 108.
273. Ch. Comminellis, *IChemE Symposium Series 127* (1992) 189 – 201.
274. V. Smith De Sucre, A. P. Watkinson, *Can. J. Chem. Eng.* **59** (1981) 52.
275. H. Sharifian, D. W. Kirk, *J. Electrochem. Soc.* **133** (1986) 921.
276. Ch. Comminellis, *Gas Wasser Abwasser* **72** (1992) no. 11.
277. K. Scott, The 1994 IChemE Research Event, Vol. 1, 344 (1994).
278. K. Scott, The Electrochemical Society, Proc. Vol. 94 – 19, p. 51 (1994).
279. P. W. T. Lu, R. L. Ammon, *J. Electrochem. Soc.* **127** (1980) 2610.
280. T. Hunger, F. Lapique, A. Storck, *J. Appl. Electrochem.* **21** (1991) 588.
281. G. H. Kelsall, S. Saage, D. Brandt, *J. Electrochem. Soc.* **138** (1991) 117.
282. P. Meszaros et al., *Hungarian J. Ind. Chem.* **12** (1984) 163.
283. M. Sudoh et al., *J. Chem. Eng. Japan* **21** (1988) 198.
284. S. D. Fleischmann, R. A. Pierce, US-Report WRSC-MS-91-192, 180th Meeting of The Electrochemical Society, Phoenix Arizona, 1991.
285. C.E.A. France, private communication.
286. J. C. Farmer, F. T. Wang, R. A. Hawley-Fedder, P. R. Lewis, L. J. Summers, L. Foiles, *J. Electrochem. Soc.* **139** (1992) 654.
287. F. Steele, *Platinum Met. Rev.* **34** (1990) 10.
288. U. Leffrang, K. Ebert, K. Flory, U. Galla, H. Schmieder, Conf. Separation Sci. & Techn., Gatlinburg, Tennessee, USA, 24. – 28. Oct. 1993.
289. R. D. Lide (eds.): *CRC Handbook of Chemistry and Physics,* 73rd ed., Chap. 8, CRC Press, Boca Raton 1992 – 1993, pp. 17 – 27.
290. W. J. Plieth, *Encyclopedia of Electrochemistry of the Elements,* **Vol. 8,** 440.
291. U. Galla, J. Bringmann, H. Schmieder, GDCh Monografie, Vol. 14, 270 (1998).
292. J. Bringmann, U. Galla, H. Schmieder, 5th International HCH and Pesticides Forum, Leioa, Basque Country, Spain, 25 – 27 June, 1998.
293. J. Bringmann, K. Ebert, U. Galla, U. Leffrang, H. Schmieder, ASME Heat Transfer Division, HTD-Vol. 317-2, (1995) IMECE 289.
294. G. Kreysa, H.-J. Külps, *Chem.-Ing.-Tech.* **55** (1983) 58.
295. G. Kreysa, H.-J. Külps, *Ger. Chem. Eng.* **6** (1983) 352.
296. G. Kreysa, H.-J. Külps, C. Woebcken, *Dechema-Monographien* **94** (1983) 199.
297. G. Kreysa, *Dechema-Monographien* **109** (1987) 9.
298. G. Kreysa, *Chem.-Ing.-Tech.* **62** (1990) 357.
299. R. Rottmann, Fortschr.-Ber. VDI, Reihe 15, Nr. 111, VDI-Verlag, Düsseldorf 1993.
300. Lurgi Schnellinformation C 1217/12.76 (1976).
301. D. van Velzen, H. Langenkamp, A. Moryoussef, *J. Appl. Electrochem.* **20** (1990) 60.
302. CZ 153 372, 1974 (F. Strafelda, J. Krofta).
303. G. Kreysa, J. M. Bisang, W. Kochanek, G. Linzbach, *J. Appl. Electrochem.* **15** (1985) 639.
304. P. Iliev, I. Nikolov, T. Vitanov, E. Budevski, *J. Appl. Electrochem.* **22** (1992) 425.
305. D. Townley, J. Winnick, *Electorchim. Acta* **28** (1983) 389.
306. H. S. Lim, J. Winnick, *J. Electrochem. Soc.* **131** (1984) 562.
307. D. Weaver, J. Winnick, *J. Electrochem. Soc.* **134** (1987) 2451.
308. J. Winnick, in H. Gerischer, C. W. Tobias (eds.): *Advances in Electrochemical Science and Engineering,* **Vol. 1,** VCH-Verlagsgesellschaft, Weinheim 1990, p. 205.

309. T. Hibino, *J. Appl. Electrochem.* **25** (1995) 203.
310. M. Aurousseau, C. Roizard, A. Storck, F. Lapicque, *Ind. Eng. Chem. Res.* **35** (1996) 1243.
311. J. M. Nzikou, M. Aurousseau, F. Lapicque, *J. Appl. Electrochem.* **25** (1995) 967.
312. K. Jüttner, G. Kreysa, K.-H. Kleifges, R. Rottmann, *Chem.-Ing.-Tech.* **66** (1994) 82.
313. K.-H. Kleifges, Fortschr.-Ber. VDI, Reihe 15, Nr. 153, VDI-Verlag, Düsseldorf 1996.
314. F. Lapicque, M. Aurousseau, C. Roizard, A. Storck, P. Dyens, *Dechema-Monographien* **128** (1993) 487.
315. M. Aurousseau, T. Hunger, A. Storck, F. Lapicque, *Chem. Eng. Sci.* **48** (1993) 541.
316. C. G. Vayenas, S. Bebelis, S. Ladas, *Nature* **343** (1990) 625.
317. C. G. Vayenas, M. M. Jaksic, S. Bebelis, S. G. Neophytis, *Modern Aspects of Electrochemistry* **29** (1995) 57.
318. J. Nicole, D. Tsiplakides, S. Wodiunig, C. Comninellis, *J. Electrochem. Soc.* (1977) L 312.
319. D. Pletcher: *Industrial Electrochemistry,* Chapman and Hall, London-New York 1982, pp. 200 – 216.
320. J. P. Hoare, M. A. LaBoda: "Electrochemical Machining," *Compr. Treatise Electrochem.* **2** (1981) 399 – 520.
321. J. P. Hoare, M. A. LaBoda : "Electrochemical Machining," in E. Yeager, A. J. Salkind (eds.): *Techniques of Electrochemistry,* J. Wiley & Sons, New York-Chichester-Brisbane-Toronto 1978, pp. 48 – 141.
322. A. Rahmel, W. Schwenk: *Korrosion und Korrosionsschutz von Stählen,* Verlag Chemie, Weinheim-New York 1977, p. 63.

General References

323. F. Fichter: *Organische Elektrochemie,* Verlag Steinkopf, Leipzig – Dresden 1942.
324. S. Swann, Jr., in A. Weissberger (ed.): *Technique of Organic Chemistry,* 2nd ed., **vol. 2,** Wiley-Interscience, New York 1956.
325. M. J. Allen: *Organic Electrode Processes,* Reinhold, New York 1958.
326. R. N. Adams: *Electrochemistry at Solid Electrodes,* M. Dekker, New York 1969.
327. C. K. Mann, K. N. Barnes: "Electrochemical Reactions" in *Nonaqueous Systems,* M. Dekker, New York 1970.
328. A. T. Kuhn: *Industrial Electrochemical Processes,* Elsevier, Amsterdam 1971.
329. L. Eberson, H. J. Schäfer, *Fortschr. Chem. Forsch. (Top. Curr. Chem.)* **21** (1971) 1.
330. N. S. Hush (ed.): *Reactions of Molecules at Electrodes,* Wiley-Interscience, New York 1971.
331. F. Beck, *Elektroorganische Chemie,* Verlag Chemie, Weinheim 1974.
332. N. L. Weinberg: Technique of Electroorganic Synthesis, Part I and II, in A. Weissberger (ed.): *Techniques of Chemistry,* **vol. V,** Wiley-Interscience, New York, 1974/1975.
333. M. Rifi, F. H. Covitz: *Introduction to Organic Electrochemistry,* M. Dekker, New York 1974.
334. S. D. Ross, M. Finkelstein, E. J. Rudd: *Anodic Oxidation,* Academic Press, New York 1975.
335. A. J. Bard, H. Lund (eds.): *Encyclopedia of Electrochemistry of the Elements,* **vol. XI – XIV,** M. Dekker, New York 1976 – 1980.
336. L. Eberson, "Electroorganic Synthesis," in R. Scheffold (ed.): *Modern Synthetic Methods,* **vol. 2,** Salle Sauerländer, Frankfurt/M. 1980.
337. D. Pletcher: *Industrial Electrochemistry,* Chapman & Hall London 1982.
338. N. L. Weinberg, B. V. Tilak, "Technique of Electroorganic Synthesis," in A. Weissberger (ed.): *Techniques of Chemistry,* **vol. 5,** Part III, Wiley-Interscience, New York 1982.
339. K. Yoshida, "Electrooxidation", in *Organic Chemistry – The Role of Cation Radicals as Synthetic Intermediates,* J. Wiley & Sons, New York 1984.
340. T. Shono: *Electroorganic Chemistry as a New Tool in Organic Synthesis,* Springer-Verlag, Berlin 1983.
341. S. Torii: *Electroorganic Syntheses, Methods and Applications,* Part I, Oxidations, Kodansha, Tokyo 1985.
342. H. L. Chum, M. M. Baizer: *The Electrochemistry of Biomass and Derived Materials,* ACS Monographie, 1985.
343. S. Torii (ed.): *Stud. Org. Chem., 30, Recent Advances in Electroorganic Synthesis,* Kodansha – Elsevier, Tokyo – Amsterdam1987.
344. "Organische Elektrochemie – Angewandte Elektrothermie", *DECHEMA-Monogr.* **112** (1989).
345. A. J. Fry: *Synthetic Organic Electrochemistry,* 2nd ed., Wiley-Interscience, New York 1989.
346. D. Pletcher, F. C. Walsh: *Industrial Electrochemistry,* 2nd ed., Chapman & Hall, London 1990.
347. T. Shono: *Electroorganic Synthesis,* Academic Press, London 1991.
348. R. D. Little, N. L. Weinberg (eds.): *Recent Advances in Electroorganic*

Synthesis – Festschrift for Manuel M. Baizer, M. Dekker, New York 1991.
349. H. Lund, M. M. Baizer (eds.): *Organic Electrochemistry, An Introduction and a Guide,* 3rd ed., Marcel Dekker, New York 1991.
350. T. Shono (ed.): *Modern Methodology in Organic Synthesis,* Kodansha – VCH, Tokyo – Weinheim1992 (Proceedings).
351. "Elektrochemische Stoffgewinnung – Grundlagen und Verfahrensstechnik", *DECHEMA-Monogr.* **125** (1992).
352. D. Kyriacou: *Modern Electroorganic Chemistry,* Springer Lab Manual, Springer-Verlag, Berlin – Heidelberg1994.
353. S. Torii (ed.): *Novel Trends in Electroorganic Synthesis,* Kodansha, Tokyo 1995 (Proceedings).
354. E. Steckhan (ed.): *Topics in Current Chemistry, Electrochemistry Series,* **vol. I – VI, 142 (1987), 143 (1987), 148 (1988), 152 (1989), 170 (1994), 185 (1997),** Springer-Verlag, Berlin – Heidelberg.

Specific References
355. M. Faraday, *Pogg. ann. Phys. Chem.* **33** (1834) 438; H. Kolbe, *Liebigs Ann. Chem.* **69** (1849) 275.
356. L. L. Bott, *Hydrogen Process.* **1965,** 115.
357. D. E. Danly, C. J. H. King, in [349], p. 1285.
358. M. M. Baizer, *Chem. Tech.* **10** (1980) 161.
359. D. E. Danly, *J. Electrochem. Soc.* **131** (1984) 435c.
360. D. E. Danly, *Hydrocarbon Process.* **60** (1981) 161.
361. M. M. Baizer, D. Danly, *Chem. Ind. (London)* (1979) 435.
362. D. Danly in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot, to Production,* Electrosynthesis Co., Inc., East Amherst, NY 1990, p. 147.
363. D. Degner, *Top. Curr. Chem.* **148** (1988) 1 – 95.
364. BASF Intermediates, BASF company brochure, 1997, 84.
365. N. L. Weinberg, in *Techniques of Electroorganic Synthesis,* Part II, John Wiley & Sons, New York 1975, pp. 1 – 82.
366. W. V. Childs, in V. Tilak, N. L. Weinberg (eds.): *Techniques of Electroorganic Synthesis,* Part III, John Wiley & Sons, New York 1982, 341 – 383.
367. *Electrolytic Processes, Present and Future Prospects,* EPRI Report No. TR-107022, Dec. 1997, Electrosynthesis Company Inc. and Dextra Associates.
368. R. Jansson, *Chem. Eng. News* **62** (1984) no. 47, 43 – 57.
369. N. L. Weinberg, in J. D. Genders, D. Pletcher (eds): *Electrosynthesis – From Laboratory, To Pilot, To Production,* The Electrosynthesis Company, Inc., East Amherst, NY 1990, p. 1.
370. D. E. Danly, *Emerging Opportunities for Electroorganic Processes,* Marcel Dekker, New York 1984.
371. J. H. Utley, *Chem. Ind. (London)* (1994) 215.
372. J. D. Genders, D. Pletcher, *Chem. Ind. (London)* (1996) 682.
373. M. Käppel, *Chem. Ing. Tech.* **35** (1963) 386.
374. M. Herbst, K. Petz, J. Stenzel, *DECHEMA-Monogr.* **125** (1992) 243.
375. E. Steckhahn, *Topics Curr. Chem.* **142** (1987) 1.
376. E. Steckhahn, *Angew. Chem.* **98** (1986) 681; *Angew. Chem. Int. Ed. Engl.* **25** (1986) 693.
377. R. L. Clark, in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – from Laboratory, To Pilot, To Production,* Electrosynthesis Company, Inc., East Amherst, NY 1990, p. 93.
378. W. Schmidt, E. Steckhahn, *Chem. Ber.* **113** (1980) 557.
379. S. Dapperheld, E. Steckhahn, K.-H. Grosse-Brinkhaus, T. Esch, *Chem. Ber.* **124** (1991) 2557.
380. P. Vaudano, E. Plattner, C. Comninellis, *Chimia* **49** (1995) 12.
381. P. Vaudano, *Chimia* **46** (1992) 103.
382. Givaudan & Cie, EP 131 001; (C. Comninellis, E. Plattner); PCT Int. Appl. WO 8 402 522, 1984; *Chem. Abstr.* **102** (1985) 5934s.
383. C. Comninellis, E. Plattner, *J. Appl. Electrochem.* **17** (1987) 1315.
384. P. M. Bersier, L. Carlsson, J. Bersier, *Top. Curr. Chem.* **170** (1994) 152 – 156.
385. P. Bouchard, A. Velin, in S. Torii (ed.): *Novel Trends in Electroorganic Synthesis,* Kodansha, Tokyo 1995, p. 271.
386. R. P. Kreh, R. M. Spotnitz, J. T. Lundquist, *Tetrahedron Lett.* **28** (1987) 1067.
387. R. P. Kreh, R. M. Spotnitz, J. T. Lundquist, *J. Org. Chem.* **54** (1989) 1526.
388. R. M. Spotnitz, R. P. Kreh, J. T. Lundquist, R. J. Press, *J. Appl. Electrochem.* **20** (1990) 209.
389. W. R. Grace & Co, US 4 639 298, (R. P. Kreh, R. M. Spotnitz); *Chem. Abstr.* **106** (1987) 127956d.

390. R. P. Kreh, R. M. Spotnitz, J. T. Lundquist, in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot to Production,* Electrosynthesis Company, Inc., East Amherst, NY 1990, p. 187.
391. G. Kreysa, O. U. Langner, *DECHEMA-Monogr.* **120** (1989) 127.
392. Otsuka Pharm. Co., JP 82 02 882, 1982, JP 82 02 881, 1982; *Chem. Abstr.* **96** (1982) 151409y and 151410s.
393. Otsuka Pharm. Co., JP 81 69 385, 1981, *Chem. Abstr.* **95** (1981) 158792f.
394. M. Taniguchi, in Recent Advances in S. Torii (ed.): *Electroorganic Synthesis,* Elsevier, Amsterdam 1987, p. 436.
395. Reilly Tar & Chem. Co., US 4 482 439, 1984, (J. E. Toomey, Jr.); *Chem. Abstr.* **102** (1985) 102453s.
396. R. Jansson, *Chem. Eng. News* (1984), Nov. 19, 43.
397. S. Harrison, A. Théoret, *Int. Congr. Electr. Appl., 13th* **2** (1996) CC81; *Chem. Abstr.* **127** (1997) 177948g.
398. S. Harrison, K. Amouzegar, 10th International Forum of Electrolysis in the Chemical Industry, Clearwater Beach, Florida, USA November 1996; *Proc. – Electrochem. Soc.* **97-6** (1997) 29.
399. D. Danly, in *Emerging Opportunities for Electroorganic Processes,* Marcel Dekker, New York 1984, p. 229.
400. Carus Co., DE 2 221 691, 1972; US 3 897 319, 1972 (F. J. Anderson, J. L. Kessler); *Chem. Abstr.* **78** (1973) 142367e.
401. The Electricity Council, DE 2 201 017, 1972 (J P. Millington); *Chem. Abstr.* **77** (1972) 121544g.
402. Union Rheinische Braunkohlen Kraftstoff, DE 2 108 623, 1972 (M. Fremery, H. Höver, G. Schwarzlose); *Chem. Abstr.* **77** (1972) 139606k.
403. D. Degner, in N. L. Weinberg, B. V. Tilak (eds.): *Technique of Electroorganic Synthesis,* in A. Weissenberger (ed.): *Techniques of Chemistry,* **vol. 5,** Part III, Wiley-Interscience, New York 1982, p. 251.
404. BASF, DE 2 434 845, 1976 (D. Degner, J. Haufe, C. Rentzea); *Chem. Abstr.* **84** (1976) 142633s.
405. BASF, DE 2 659 148, 1978 (K. Boehlke, D. Degner, W. Treptow); *Chem. Abstr.* **89** (1978) 97149s.
406. A. J. Rudge, in A. T. Kuhn (ed.): *Industrial Electrochemical Processes,* Elsevier, Amsterdam 1971, p. 71.
407. J. H. Simons, *J. Electrochem. Soc.* **95** (1949) 47, 53, 55, 59, 64.
408. M. Sander, W. Blöchel, *Chem. Ing. Tech.* **37** (1965) 7.
409. H. Millauer, *DECHEMA-Monogr.* **112** (1988) 429.
410. T. Fuchigami, *Top. Curr. Chem.* **170** (1994) 1.
411. S. D. Ross, M. Finkelstein, R. C. Petersen, *J. Am. Chem. Soc.* **86** (1964) 2745; *J. Org. Chem.* **31** (1966) 128; *J. Am. Chem. Soc.* **88** (1966) 4657.
412. T. Shono, *Tetrahedron* **40** (1984) 824.
413. T. Shono, *Electroorganic Chemistry as a New Tool in Organic Synthesis,* Springer-Verlag, Berlin 1984, pp. 63 – 77.
414. E. Steckhahn, in H. Lund, M. M. Baizer (eds.): *Organic Electrochemistry: An Introduction and a Guide,* 3rd ed., Marcel Dekker, New York 1991, pp. 601 – 607.
415. Ref. [254] – [272] in [363]
416. D. Degner, *DECHEMA-Monogr.* **112** (1989) 375.
417. Shell Int. Res., EP 67 463, 1982 (R. Engels, J. H. H. Meurs), *Chem. Abstr.* **98** (1983) P 178994p.
418. T. Shono, Y. Matsumura, K. Inoue, F. Iwasaki, *J. Chem. Soc., Perkin 1* **1986,** 73.
419. H. Püttner, J. Botzem, H. Hannebaum, D. Hermeling, *Annual Meeting 1998 of the Fachgruppe Angewandte Elektrochemie der GDCh, GDCh-Monographie,* **14** (1999) 458.
420. J. F. Fauvarque et al., *J. Organomet. Chem.* **264** (1984) 273.
421. J. Chaussard et al., *J. Appl. Electrochem.* **19** (1989) 345.
422. J. Chaussard, in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot to Production,* Electrosynthesis, Co., Inc., East Amherst, NY 1990, p. 165.
423. M. Troupel, J. Perichon, *Tetrahedron Lett.* **26** (1985) 1509; SNPE, FR 2 566 434, 1986; *Chem. Abstr.* **104** (1986) 158203u.
424. M. Taniguchi, in [343], p. 433.
425. Ref. [49] – [58] in [363].
426. S. Dapperheld, H. Millauer, in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot to Production,* Electrosynthesis, Co., Inc., East Amherst, NY 1990, p. 115.
427. H. Millauer, *Chem. Ing. Tech.* **52** (1980) 53.
428. Ref. [59] – [61] in [363].
429. US 3 992 417, 1976; (R. M. Dessau, E. I. Heiba); *Chem. Abstr.* **86** (1977) 54999.

430. Monsanto Co., EP 55 934, 1982; *Chem. Abstr.* **97** (1982) 146549k; EP 55 936, 1982 (J. P. Coleman, R. C. Halcher, D. E. McMackins); *Chem. Abstr.* **97** (1982) 146550d.
431. Monsanto Co., US 4 560 775, 1985; (J. P. Coleman et al.); *Chem. Abstr.* **104** (1986) 207132f.
432. I. J. M. Girvan, W. N. Brooks, [343], p. 445.
433. Sepracor Inc., US 5 302 257, 1994, WO 93 17 150, 1993 (Y. Gao, C. M. Zepp); *Chem. Abstr.* **120** (1994) 119437j.
434. A. R. Amundsen, E. N. Balko, *J. Appl. Electrochem.* **22** (1992) 810.
435. S. Torii, *Interface* **6** (1997) no. 4, 46.
436. S. Torii et al., *J. Org. Chem.* **61** (1996) 3055.
437. J. S. Mayell, *Ind. Eng. Chem., Prod. Res. Develop.* **7** (1968) 129; American Cyanamid Co., US 3 555 084, 1971; *Chem. Abstr.* **74** (1971) 60256a.
438. J. E. Toomey, Jr. et al., in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot to Production,* Electrosynthesis, Co., Inc., East Amherst, NY 1990, p 131.
439. BASF, DE 2 403 446, 1975 (H. Hannebaum, H. Nohe, H. R. Müller); *Chem. Abstr.* **83** (1975) 210783x.
440. BASF, DE 2 658 951, 1978 (H. Nohe, H. Hannebaum); *Chem. Abstr.* **89** (1978) 109902v.
441. Hoechst, DE 2 618 276, 1977 (D. H. Skaletz); *Chem. Abstr.* **88** (1978) 62213x.
442. H. V. K. Udupa, in S. Torii (ed.): *Recent Advances in Electroorganic Synthesis,* Elsevier, Amsterdam 1987, p. 423.
443. K. Junghans, *Chem. Ber.* **107** (1974) 3191; T. Matsue, T. Yansada, H. Takahashi, T. Osa, *Bull. Chem. Soc. Jpn.* **59** (1986) 3690.
444. K. Junghans, *Chem. Ber.* **106** (1973) 3465.
445. R. L. Taylor, *Chem. Met. Eng.* **44** (1937) 588.
446. Electricité de France, DE 3 108 350, 1981 (D. Doniat); *Chem. Abstr.* **96** (1982) 43136y.
447. P. N. Pintauro, N. Phan, M. M. Baizer, K. Nobe, *AIChE Symp.* **254** (1987) 34.
448. Ref. [513] – [522] in [363].
449. L. Carlsson, H. Holmberg, B. Johansson, A. Nilsson, in N. L. Weinberg, B. V. Tilak (eds.): *Technique of Electroorganic Synthesis,* in A. Weissenberger (ed.): *Techniques of Chemistry,* **vol. 5,** Part III, Wiley-Interscience, New York 1982, p. 192.
450. BASF, DOS 1 950 282, 1971 (F. Beck, P. Jäger, H. Gutschke); *Chem. Abstr.* **74** (1971) 140980x.
451. Y. Hou, H. Huixin, R. Tong, D. Gu, *Huaxue Shijia* **33** (1992) 37.
452. G. C. Coleman, H. L. Johnson, *Org. Synth. Coll.* **Vol. 3,** 1955, p. 60.
453. T. R. Beck (ed.): *A Survey of Organic Electrolytic Processes,* **ANL/OEPM-79-5,** Argonne Natl. Lab. Argonne 1979.
454. K. S. Udupa, G. S. Subramaniam, H. V. K. Udupa, *Industr. Chemist* **39** (1963) 238.
455. J. C. Hofmann, P. M. Robertson, N. Ibl, *Tetrahedron Lett.* **1972,** 2433.
456. S. Takenaka, R. Oi, C. Shimakawa, Y. Shimakawa, [343], p. 215; Mitsui Toatsu Chem., JP 62 127 487, 1987; *Chem. Abstr.* **107** (1987) 245188y; JP 63 83 289, 1988; *Chem. Abstr.* **109** (1988) 118234b; JP 63 103 093, 1988; *Chem. Abstr.* **109** (1988) 63272e; JP 63 157 891, 1988; *Chem. Abstr.* **109** (1988) 239327t; JP 63 157 892, 1988; *Chem. Abstr.* **109** (1988) 239328u; JP 63 62 890, 1988; *Chem. Abstr.* **109** (1988) 239329v; JP 63 206 488, 1988; *Chem. Abstr.* **110** (1989) 47421t; JP 206 489, 1988; *Chem. Abstr.* **110** (1989) 47422u; S. Takenaka, *Chem. Lett.* **1988,** 899; C. Shimakawa, Y. Shimakawa, S. Takenaka, *Bull. Chem. Soc. Jpn.* **60** (1987) 4193; JP 34 275, 1995 (N. Kazunari, H. Matsuno, C. Shimakawa, H. Gamachi); *Chem. Abstr.* **122** (1995) 250508s.
457. Hoechst, DE 2 428 878, 1976 (L. Horner, D. H. Skaletz, H. Hönl); *Chem. Abstr.* **84** (1976) 105228t.
458. Osaka Soda Co., JP 8343 66, 1961.
459. BASF, DE 2 144 419, 1973 (F. Beck, E. Schefczik, F. P. Woerner); *Chem. Abstr.* **78** (1973) 147778d.
460. BASF, DE 2 510 920, 1976 (H. Hannebaum, H. Nohe); *Chem. Abstr.* **86** (1977) 16430b.
461. BASF, DE 2 630 927, 1978 (H. Hannebaum, H. Nohe); *Chem. Abstr.* **88** (1978) 105119r.
462. BASF, DE 19 618 854, 1997 (H. Pütter, H. Hannebaum); *Chem. Abstr.* **127** (1997) 338559y.
463. Smith, Kline & French, DE 2 538 621, 1976 (R. J. King, G. R. White); *Chem. Abstr.* **85** (1976) 70091j.
464. J. H. Wagenknecht, A. S. C. Chen, T. T. Huang, R. E. Miller, *J. Org. Chem.* **60** (1995) 345.
465. A. K. Datta, P. A. Marron, C. J. H. King, J. H. Wagenknecht, *193rd Meeting of The Electrochemical Society,* San Diego, 1998, Abstr. No. 922.
466. Monsanto, US 4 582 577, 1986 (J. H. Wagenknecht); *Chem. Abstr.* **105** (1986) 50861p.

467. Monsanto, US 4 601 797, 1986 (J. H. Wagenknecht); *Chem. Abstr.* **106** (1986) 24982e.
468. D. Pletcher, L. Slevin, *J. Chem. Soc., Perkin 2,* **1996,** 217.
469. P. M. Robertson et al., *J. Electrochem. Soc.* **130** (1983) 591.
470. P. M. Robertson, N. Ibl, *J. Appl. Electrochem.* **7** (1977) 323.
471. P. Seiler, P. M. Robertson, *Chimia* **36** (1982) 305.
472. Merck, EP 40 331, 1981 (R. Wittmann, W. Wintermeyer, J. Butzke); *Chem. Abstr.* **96** (1982) 69373y.
473. Merck, EP 40 709, 1981 (F. Reiff, W. Wintermeyer, R. Wittmann, J. Butzke); *Chem. Abstr.* **96** (1982) 26350w.
474. H. Marzouk et al., EDF, *10th International Forum of Electrolysis in the Chemical Industry,* Clearwater Beach, Florida, U.S.A., November 1996, p. 309, Electrosynthesis Company, Lancaster, N.Y.; Agro Ind. Rech. & Dev. Ard., FR 2 699 937 1992.
475. Ref. [566] – [568] in [363]
476. Ref. [201] – [212] in [375]
477. BASF, DE 2 262 851, 1984 (P. Jäger, B. Zeeh); *Chem. Abstr.* **81** (1974) 120636f.
478. G. Trümpler, P. Rüetschi, *Helv. Chim. Acta* **36** (1953) 1649.
479. Ref. [574] – [582] in [363]
480. Merck, DE 4 333 697, 1995 (U. Fechtel, K. Wembach, H.-H. Bokel); *Chem. Abstr.* **122** (1995) 290722p.
481. J. E. Toomey, Y. C. Yu, *Chem. Eng. (N.Y.)* **96** (1989) 140; *Chem. Abstr.* **111** (1989) 176632g.
482. Q. N. Porter, J. H. P. Utley, *J. Chem. Soc., Chem. Commun.* **1978** 255.
483. M. Platen, E. Steckhahn, *Tetrahedron Lett.* **21** (1980) 511; *Chem. Ber.* **117** (1984) 1679.
484. L. L. Miller, J. F. Wolf, E. A. Mayeda, *J. Am. Chem. Soc.* **93** (1971) 3306, **94** (1972) 6812.
485. R. F. Garwood, N. ud Dim, B. C. L. Weedon, *J. Chem. Soc., Perkin. Trans. 1* **1975,** 2471.
486. S. M. Weinreb, G. A. Epling, R. Comi, M. Reitano, *J. Org. Chem.* **40** (1975) 1356.
487. W. Schmidt, E. Steckhan, *Angew. Chem.* **90** (1978) 717; **91** (1979) 850; **91** (1979) 851; *Angew. Chem. Int. Ed. Engl.,* **17** (1978) 673; **18** (1979) 802.
488. T. Shono: *Electroorganic Chemistry as a New Tool in Organic Synthesis,* Springer-Verlag, Berlin 1984, p. 54.
489. T. Shono, T. Toda, N. Oshino, *J. Am. Chem. Soc.* **104** (1982) 2639; *Drug Metab. Dispos.* **9** (1981) 481.
490. S. Torii, T. Inokuchi, N. Takahashi, *J. Org. Chem.* **43** (1978) 5020.
491. H. F. Conway, E. B. Lancaster, V. E. Sohns, *Electrochem. Tech.* **2** (1964) 43.
492. R. Ramaswamy, M. S. Venkatachalapathy, H. V. K. Udupa, *Ind. J. Technol.* **1** (1963) 115.
493. N. L. Weinberg, *10th International Forum of Electrolysis in the Chemical Industry,* Clearwater Beach, FL Nov. 1996.
494. V. G. Mairanovsky, *Angew. Chem.* **88** (1976) 283, *Angew. Chem. Int. Ed. Engl.* **15** (1976) 281.
495. M. I. Montenegro, *Electrochim. Acta* **31** (1986) 607.
496. L. Horner, *Pure Appl. Chem.* **9** (1964) 225.
497. J. D. Genders, N. L. Weinberg, C. Zawodzinski, in [348], p. 273.
498. F. C. Walsh, *A First Course in Electrochemical Engineering,* The Electrochemical Consultancy Ltd., Roomsey, 1993, p. 327.
499. Electrosynthesis Co., EP 436 055, 1991 (J. D. Genders, D. J. Mazur, N. L. Weinberg); *Chem. Abstr.* **115** (1991) 169016p.
500. Electrosynthesis Co., US 5 106 463, 1992 (J. D. Genders, N. L. Weinberg, D. J. Mazur); *Chem. Abstr.* **117** (1992) 16094d.
501. Isochem, Information Chimie No. 266, Nov. 1985, p. 167.
502. Diamalt, DE 1 643 285, 1967, (F. Bayerlein, W. Kohlbeck, H. D. Kessler).
503. A. Aldaz et al., WO 96 38 601, 1996; *Chem. Abstr.* **126** (1997) 81285s; WO 97 42 358, 1997; *Chem. Abstr.* **128** (1998) 8221e.
504. G. Sanchez Cano, V. Montiel, A. Aldaz, *Tetrahedron* **47** (1991) 877.
505. K. Krumel, S. Phillips, B. Wu, *193rd Meeting of The Electrochemical Society,* San Diego, 1998, Abstr. No. 921.
506. Dow Chem., US 4 217 185, 1980 (F. Y. Edamura, D. Kyriacou, J. Love); *Chem. Abstr.* **94** (1981) 22193h.
507. Dow Chem., EP 18 069, 1980 (D. Kyriacou); *Chem. Abstr.* **94** (1981) 54933r.
508. Dow Chem., US 3 694 332, 1972.
509. Otsuka Chem., DE 3 207 506, 1982 (T. Shono); *Chem. Abstr.* **97** (1982) 215767u.
510. Otsuka Chem. DE 3 507 592, 1985 (S. Torii, H. Tanaka, M. Sasaoka, Y. Kameyama); *Chem. Abstr.* **104** (1986) 148628x; S. Torii, H. Tanaka, T. Oshima, S. Sasaoka, *Bull. Chem. Soc. Jpn.* **59** (1986) 3975.

511. G. Zappi, *Proceedings 12th Intern. Forum on Electrolysis in the Chemical Industry, Clearwater,* FL 1998; C. Martin, V. Menon, H. Zang, P. Zelenay, D. Walker, J. Lee, *Proceedings 10th Intern. Forum on Electrolysis in the Chemical Industry,* Clearwater, FL 1996, publ. by Electrosynthesis Co.
512. J. H. P. Utley, Y. Gao, J. Gruber, R. Lines, *J. Mater. Chem.* **5** (1995) 1297.
513. J. H. P. Utley et al., *J. Mater. Chem.* **5** (1995) 1837.
514. T. Shono, S. Kashimura, M. Ishifune, R. Nishida, *J. Chem. Soc., Chem. Commun* **1990,** 1160.
515. S. Kashimura et al., *Tetrahedron Lett.* **38** (1997) 4607.
516. S. Kashimura et al., *J. Org. Chem.* 1999, in print.
517. Osaka Gas Co., JP 03 264 683, 1991 (T. Shono, S. Kashiwamura, R. Nishida); *Chem. Abstr.* **116** (1992) 161234h; Osaka Gas Co., JP 04 348 128, 1992; *Chem. Abstr.* **119** (1993) 73387s.
518. Osaka Gas Co., JP 09 241 267, 1997 (T. Shono, S. Kashiwamura, T. Ishifune, R. Nishida, H. Murase); *Chem. Abstr.* **127** (1997) 307816a.
519. R. Engels, H. J. Schäfer, E. Steckhan, *Liebigs Ann. Chem.* **1997,** 204; H. Baltes, E. Steckhan, H. J. Schäfer, *Chem. Ber.* **111** (1978) 1294; D. Koch, H. J. Schäfer, E. Steckhan, *Chem. Ber.* **107** (1974) 3640; D. Koch, H. J. Schäfer, *Angew. Chem.* **85** (1973) 264; *Angew. Chem. Int. Ed. Engl.* **12** (1973) 245.
520. H. J. Schäfer, *Topics Curr. Chem.* **152** (1990) 91.
521. Monsanto Co., US 4 076 601, 1978 (D. A. White); *Chem. Abstr.* **88** (1978) 160738s.
522. Ref. [219] – [224] in [363].
523. E. A. Casanova, M. C. Dutton, D. J. Kalota, J. H. Wagenknecht, *J. Appl. Electrochem.* **25** (1995) 479; Monsanto, WO 97 26389, 1997; *Chem. Abstr.* **127** (1997) 141984. Monsanto, EP 0816 533 A2; *Chem. Abstr.* **115** (1991) 217299g. Monsanto, EP 0433 260 B1, 1991. N. L. Weinberg, D. J. Mazur, in R. D. Little, N. L. Weinberg (eds.): *Recent Advances in Electroorganic Synthesis – Festschrift for Manuel M. Baizer,* M. Dekker, New York 1991, p. 55.
524. N. L. Weinberg, G. D. Zappi, J. D. Genders, D. Pletcher, *Novel Trends in Electroorganic Synthesis,* Kodansha, Toxyo 1995, p. 437.
525. US 4 478 694, 1984 (N. L. Weinberg); *Chem. Abstr.* **102** (1985) 35368c.
526. N. L. Weinberg et al., *Recent Advances in Electroorganic Synthesis,* S. Torii, Ed., Elsevier, Amsterdam, 1987, 441; M. I. Montenegro et al., *J. Appl. Electrochem.* **20** (1990) 54.
527. Diamond Shamrock, US 3 992 269, 1976 (T. T. Sugano, B. A. Schenker, J. A. Walborg, N. Schuster); *Chem. Abstr.* **86** (1977) 23591h.
528. Monsanto, US 4 087 336, 1978 (J. H. Wagenknecht); *Chem. Abstr.* **89** (1978) 119771e.
529. Monsanto, US 4 133 729 and 4 157 286, 1979 (C. J. H. King); *Chem. Abstr.* **90** (1979) 112138r and **91** (1979) 80947f.
530. E. Steckhan, *Top. Curr. Chem.* **170** (1994) 83.
531. E. Steckhan, B. Brielbeck, M. Frede, G. Hilt, *10th International Forum of Electrolysis in the Chemical Industry,* Clearwater Beach, FL, Nov. 1996.
532. B. Brielbeck, M. Frede, E. Steckhan, *Biocatalysis* **10** (1994) 49.
533. A. Bergel, R. Devaux-Basséguy, *J. Chim. Phys.* **93** (1996) 753.
534. H. Simon, J. Bader, H. Günther, S. Neumann, J. Thanos, *Angew. Chem.* **97** (1985) 541, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 539; H. Simon, *Pure Appl. Chem.* **64** (1992) 1181.
535. A. E. Biade et al., *J. Am. Chem. Soc.* **114** (1992) 893.
536. S. B. Sobolov et al., *J. Org. Chem.* **61** (1996) 2125.
537. G. Hilt et al., *Liebigs Ann./Reccueil* **1997,** 2289.
538. J. D. Genders, N. L. Weinberg (eds.): *Electrochemistry for a Cleaner Environment,* The Electrosynthesis Company, East Amherst, NY 1992.
539. P. M. Bersier, L. Carlsson, J. Bersier, *Top. Curr. Chem.* **170** (1994) 113.
540. BASF, DE 3 620 013, 1988 (M. Steiniger, H. Voss); *Chem. Abstr.* **108** (1988) P 175945n.
541. T. Bechtold, E. Butschler, A. Amann, O. Bobleter, *Angew. Chem.* **104** (1992) 1046; *Angew. Chem. Int. Ed. Engl.* **31** (1992) 1068.
542. T. Bechtold, E. Burtscher, A. Amann, O. Bobleter, *J. Chem. Soc. Faraday Trans.* **89** (1993) 2451.
543. T. Bechtold, E. Burtscher, A. Turcanu, O. Bobleter, *J. Appl. Electrochem.* **27** (1997) 1021.
544. T. Bechtold, E. Burtscher, A. Turcanu, O. Bobleter, *J. Appl. Electrochem. Soc.* **143** (1996) 2411.

545. BASF, DE 19 513 839, 1996 (M. Frede, E. Roske, H. Pütter); *Chem. Abstr.* **125** (1996) P 287307p.
546. S. Andreades, E. W. Zahnow, *J. Am. Chem. Soc.* **91** (1969) 4181.
547. L. Eberson, B. Helgee, *Acta Chem. Scand.* **B 29** (1975) 451.
548. L. Eberson, *Modern Synthetic Methods,* **vol. 2,** Otto Salle Verlag, Frankfurt, 1980, p. 401.
549. K. Nyberg, *J. Chem. Soc. Chem. Commun.* **1969** 774; *Acta Chem. Scand.* **24** (1970) 1609.
550. D. Pletcher, *J. Appl. Electrochem.* **14** (1984) 403.
551. D. Pletcher, F. Walsh, A. M. Cooper, *Chem. Rev.* **90** (1990) 837.
552. D. Genders, in J. D. Genders, D. Pletcher (eds.): *Electrosynthesis – From Laboratory, to Pilot, to Production,* Electrosynthesis Co., Inc. West Amherst, NY, 1990.
553. K. Junghans, *Chem. Ber.* **107** (1974) 3191.
554. T. Matsue, T. Yamada, H. Takahashi, T. Osa, *Bull. Chem. Soc. Jpn.* **59** (1986) 3690.
555. T. Nonaka, M. Takahashi, T. Fuchigami, *Bull. Chem. Soc. Jpn.* **56** (1983) 2584; *J. Electrochem. Soc. Jpn.* **51** (1983) 129.
556. P. C. S. Hayfield, R. L. Clark, *Proceedings of the Electrochemical Society Meeting,* Los Angeles, 1989, Abstr. No. 369.
557. R. R. Miller-Folk, R. E. Noftle, D. Pletcher, *J. Electroanal. Chem.* **274** (1989) 257.
558. H. J. Schäfer, *Topics Curr. Chem.* **142** (1987) 101.
559. P. Cox, D. Pletcher, *J. Appl. Electrochem.* **20** (1990) 549; **21** (1991) 11.
560. F. Beck, H. Schulz, *Electrochim. Acta* **29** (1984) 1569.
561. J. Fischer, Th. Lehmann, E. Heitz, *J. Appl. Electrochem.* **11** (1981) 743.
562. G. Silvestri, S. Gambino, G. Filardo, A. Gulotta, *Angew. Chem.* **98** (1984) 978; *Angew. Chem. Int. Ed. Engl.* **23** (1984) 979.
563. B. Belleau, Y. K. Au-Yang, *Canad. J. Chem.* **47** (1969) 2117.
564. D. J. Pickett: *Electrochemical Reactor Design,* Elsevier, Amsterdam 1979.
565. J. S. Newman: *Electrochemical Systems,* Prentice-Hall, London 1973.
566. E. Heitz, G. Kreysa: *Principles of Electrochemical Engineering,* VCH Verlagsges., Weinheim 1986.
567. D. Pletcher, in: *Electrosynthesis – From Laboratory, to Pilot, to Production,* Electrosynthesis Co., Ltd., East Amherst, NY , 1990, p. 15; D. Genders in [552] p. 65.
568. F. Walsh, D. Robinson, *Interface (The Electrochem. Soc.)* **7** (1998) no. 2, 40.
569. R. Engels, C. J. Smit, W. J. M. van Tilborg, *Angew. Chem.* **95** (1983) 502; *Angew. Chem. Int. Ed. Engl.* **22** (1983) 492.
570. R. J. Chan, C. Ueda, T. Kuwana, *J. Am. Chem. Soc.* **105** (1983) 3713.
571. J. Brodesser, Diploma Thesis, University Bonn 1985; T. Arns, Diploma Thesis, University Bonn 1998.
572. W. Li, T. Nonaka, *Chem. Lett.* **1997,** 1271.
573. L. Carlsson, B. Sandegren, D. Simmonsson, M. Rihovsky, *J. Electrochem. Soc.* **130** (1982) 342.
574. W. N. Brooks, *I. Chem. E. Symp. Ser.* **98** (1986) 1.
575. Electricity Council, EP-A 64417, 1983 (J. P. Millington, I. M. Dalrymple); *Chem. Abstr.* **98** (1983) 62142p.
576. Reilly Tar and Chemical Corp., US 4 589 9688, 1984 (J. E. Toomey); *Chem. Abstr.* **102** (1985) 14136k; J. E. Toomey, *Chem. Tech.* **15** (1985) 738; EP-A 122 736, 1984 (J. E. Toomey).
577. F. Beck, H. Guthke, *Chem. Ing. Tech.* **41** (1969) 943.
578. D. Degner, in N. L. Weinberg, B. V. Tilak (eds.): *Techniques of Chemistry,* **vol. 5,** Part III, John Wiley & Sons, New York 1982, p. 251.
579. F. Beck, *Proc. IUPAC Congress, 24th 1973,* **vol. 5,** Butterworth, London 1974, p. 111.
580. L. Eberson, K. Nyberg, H. Sternerup, *Chem. Scripta* **3** (1973) 12.
581. Hoechst, GB 2 003 506, 1979 (J. Cramer, W. Lindner).
582. Z. Ogumi, S. Ohashi, Z. Takehara, *Electrochim. Acta* **30** (1985) 1850.
583. J. Sarrazin, A. Tallec, *J. Electroanal. Chem.* **137** (1982) 183.
584. J. Jörissen, *Electrochim. Acta* **41** (1996) 553.
585. J. Jörissen, M. Held, D. Klein, K. H. Simmrock, *DECHEMA-Monogr.* **125** (1992) 693.
586. R. Fabiunke, J. Jörissen, *Chem. Ing. Tech.* **62** (1990) 400.
587. J. Jörissen, *GDCh-Monographie* **2** (1995) 41.
588. D. Hoormann et al., *Annual Meeting 1998 of the Fachgruppe Angewandte Elektrochemie der GDCh, GDCh-Monographie,* **14** (1999) 537.
589. L. Kröner et al., *Annual Meeting 1998 of the Fachgruppe Angewandte Elektrochemie der GDCh, GDCh-Monographie,* **14** (1999) 484.
590. G. Kreysa, *Chem. Ing. Tech.* **50** (1976) 332.
591. D. Simonsson, *J. Appl. Electrochem.* **14** (1984) 595.

592. H. Pütter, *9th International Forum of Electrolysis in the Chemical Industry,* Clearwater Beach, FL, Nov. 1995.

Electroless Plating → Electrochemical and Chemical Deposition

Electroluminescent Materials and Devices

Semiconductors is a separate keyword.

GÜNTER WINSTEL, Siemens AG, München, Federal Republic of Germany

1. Introduction 1
2. Physics of Light Emission in Semiconductors 3
3. Manufacture of Crystal Layers for Luminescent Devices 5
4. Practical Luminescent Diodes 6
4.1. Diodes Producing Red, Yellow, and Green 6
4.2. Blue Light Diodes 7
4.3. Infrared Diodes 7
4.4. Optoelectronic Couplers 8
5. The Laser Diode as a Source of Coherent Light 9
6. References 10

1. Introduction

Two luminescence phenomena are of major technological significance: cathodoluminescence, which is widely used in cathode ray tubes (e.g., television screens), and *electroluminescence.* The latter owes its importance to two semiconductor photoemission applications, namely, *light emitting diodes* (*LEDs*) and *laser diodes* (*LDs*). Electroluminescence refers to the process of radiation emission by a suitable material, a luminophor, following excitation induced by electronic processes, such as an applied voltage.

Electroluminescence was first observed in 1907 by H. J. ROUND at the electric point contact of a silicon carbide crystal. This effect was rediscovered in 1921 by O. V. LOSSEV, and the excitation of zinc sulfide in an alternating electric field, carried out in 1935 by G. DESTRIAU, revealed a photoemission process that is still of in significance today for display elements. These applications of electroluminescence and the material questions are treated under → Display Technology.

The monochromatic radiation discovered by LOSSEV had an energy hc/λ, close to that of the applied voltage, which enabled him to classify this process as the inverse of the Einstein photoelectric effect. In 1951, K. LEHOVEC explained the radiation observed as a result of the recombination of injected minority carriers – electrons and holes – in semiconductor diodes, an effect predicted in 1947 by W. SHOCKLEY in connection with the transistor. This theory laid the basis for the use of special semiconductor diodes as photoemissive luminescence diodes.

Normally semiconductor diodes are used as rectifiers, since they allow or inhibit the flow of current, an effect based on the properties of the *p-n* junction (→ Semiconductors), which consists of two adjacent regions of a semiconductor crystal doped with donors or acceptors, thus allowing conduction by electrons (*n*) or holes (*p*). *Donors* are impurity atoms on lattice sites that, compared with the replaced atom, have an extra electron in their outer shell. *Acceptors* lack one electron and therefore are able to capture a valence electron. For details, see → Semiconductors.

Only specific energy states are permitted for the electrons in a semiconductor. Because the distances between the atoms in the solid semiconductor are small, the electrons of the outer shells interact – both ground state and excited states degenerating into broad energy bands. The ground-state band contains the electrons that bind the atoms in the crystal lattice; therefore, it is called the *valence band.* The excited states contain the additional electrons supplied by the donors. These electrons move freely in the crystal; therefore, the corresponding band is called the *conduction band.* In a similar way, the valence electrons captured by the acceptors produce in the valence band empty electron states that can move freely in the crystal as positively charged *holes.*

A *barrier layer* without charged carriers is set up between the *n* and *p* regions of a *p-n* junc-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_255

tion on account of diffusion equilibrium. Holes or electrons can be injected through the barrier layer with a forward voltage, i.e., by applying the positive pole of a potential to the *p* region and the negative pole to the *n* region (Fig. 1). The result is a *transition zone* that simultaneously contains electrons in the conduction band and holes in the valence band.

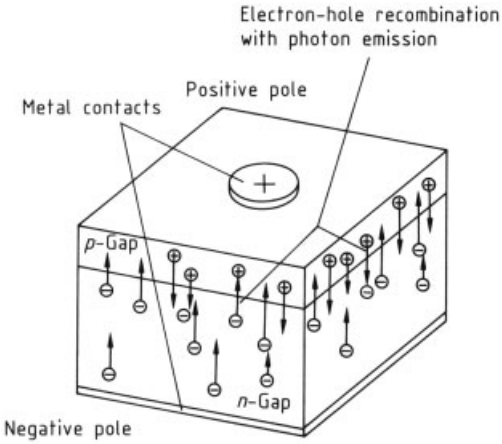

Figure 1. Structure of a simple semiconductor luminescent diode: ⊖ electrons, ⊕ holes

This state is naturally unstable, and the system tends to return to equilibrium. In this equilibration process, the electrons of the conduction band recombine with the holes of the valence band, and the energy corresponding to the gap E_g between the bands is usually absorbed by the crystal lattice as vibrational motion. In some binary and quasibinary mixed semiconductors composed of atoms of valencies III and V or II and VI, both with tetrahedral structure, or IV and VI with an octahedral structure, the recombination process for additional carriers leads directly to the emission of a light quantum (photon) without involvement of the lattice atoms. Especially for some compositions of III – V materials, this process takes place with a high quantum efficiency, i.e., the number of light-emitting processes per injected electron is close to unity. In this case, the photon energy is given by

$$h\nu = hc/\lambda = E_g$$

h = Planck's constant, ν = frequency of the light, c = speed of light, and λ = wavelength of the light

If these processes run spontaneously, i.e., in an uncorrelated way, the crystal or diode emits *incoherent light.* In this context, the term light also includes the electromagnetic radiation in the infrared and ultraviolet regions adjoining the visible spectrum insofar as such radiation can be generated by an electroluminescent mechanism. If a Fabry-Perot resonator is created, e.g., by arranging two plane crystal surfaces parallel, as a reflector pair, then a large part of the generated light is reflected back into the excitation region of the crystal, where it builds up such a high intensity between the reflecting surfaces that further radiative recombination processes are induced. The light generated is in phase with the existing light wave, thus amplifying its intensity and stimulating further radiative recombination. The resulting laser light is *coherent,* extremely monochromatic, and may have an exceptionally high energy intensity (→ Lasers).

Table 1. Direct semiconductor materials for electroluminescent and laser diodes: band-to-band transitions

Crystal system	Emission wavelength, μm
Binary III – V compounds	
GaAs [*1303-00-0*]	0.84 – 0.90
InP [*22398-80-7*]	0.91
GaSb [*12064-03-8*]	1.50
InAs [*1303-11-3*]	3.10
InSb [*1312-41-0*]	5.40
Ternary III – V compounds	
InN–GaN	0.36 – 0.62
GaAs–AlAs, Al : Ga<0.85	0.63 – 0.90
GaAs–GaP, P : As<0.82	0.64 – 0.90
InP–GaP, Ga : In<1.63	0.67 – 1.10
GaAs–GaSb	0.90 – 1.50
InP–InAs	0.91 – 3.10
GaAs–InAs	0.84 – 3.10
InAs–InSb	3.20 – 5.40
Quaternary III – V compounds	
(Ga, In)(P,As)	0.58 – 1.30
(Ga, In)(As, Sb)	0.90 – 2.20
(Ga, In, Al)P	0.62
IV – VI compounds	
PbS [*1314-87-0*]	4.30
PbSe–PbS	4.30 – 8.50
PbTe [*1314-91-6*]	6.80
PbSe [*12069-00-0*]	8.50
PbTe–SnSe	10 – 12
PbTe–SnTe	6.5 – 28

Ever since the discovery in 1962 of the high quantum efficiency, almost 100 %, of gallium arsenide diodes, which emit monochromatic infrared radiation, the use of electroluminescent semiconductors has made rapid strides, leading to the rapid development of *light emitting diodes* (*LED*), which supply incoherent light, and *laser diodes* (*LD*), which supply coherent light. These light emitting diodes and laser

diodes play a major part in implementing communication between man and machine with luminous red, yellow, green, and blue indicator lamps and optoelectronic displays for numerals and other characters. Semiconductor diodes operating at room temperature have been successfully used to generate light of wavelengths between 5.4 and 0.36 μm (Table 1). If lower operating temperatures are possible, this range can be extended to 28 μm using the chalcogenides of lead and tin.

The development of special laser diodes is still in full swing. Their principal uses in optical signaling and communications are as coherent light sources for light transmission at 0.8 μm and 1.55 μm for the absorption windows in quartz fibers and at 1.3 μm for the dispersion windows in quartz fibers (→ Glass). For absorption windows, interest centers on monomode-emitting laser diodes adapted to monomode fibers, which are preferred because of their high transmission bandwidth. Outside the dispersion window, diodes with a corrugated waveguide structure are required – they have a very narrow emission line because of their internal distributed feedback (DFB), thus experiencing little spread when traversing the fiber route. Such DFB lasers are needed particularly in integrated optics (IO) and optoelectronics (OEIC) for monolithic integration with detector, amplifier, and switching elements, which are usually developed on the basis of gallium arsenide or indium phosphide substrates. The lacking long-term stability of most semiconductor lasers, which was initially a major problem, has now been largely solved or at least is seen to be solvable.

2. Physics of Light Emission in Semiconductors

The recombination of electrons and holes can take place with or without photoemission, corresponding to the various permitted transitions between energy states.

The electronic states of the pure semiconductor are degenerated to state bands. These bands comprise the permitted energy states with various electron momenta *p*. An important role is also played by more or less localized states that result from lattice defects or the inclusion of impurity atoms. Figure 2 shows regions in the band structures (of gallium arsenide, GaAs, and gallium phosphide, GaP) that are important for the emission of light through recombination of electrons and holes. *Gallium arsenide* and *gallium phosphide* are representative materials for the different light-producing processes involved in electroluminescent devices. The permitted energy states are plotted against the momentum vector p for characteristic crystal orientations. Excess electrons are introduced into the conduction band E_c and holes into the valence band E_v, e.g., by optical excitation of valence electrons or by current carrier injection via a *p-n* junction. In GaAs, these can recombine *directly* where the momenta coincide (Γ_1 on E_c and Γ_8 on E_v); otherwise, a third collision partner is necessary. The direct process, which is always associated with the emission of light, takes place only in *direct semiconductors*, i.e., those having a band structure like GaAs (see Table 1).

Indirect transition, demonstrated in Figure 2 for GaP, can take place in all materials, usually yielding no light. If the energy is initially transferred to a free carrier in the conduction or valence band or to a weakly localized charged imperfection, lattice vibrations are generated through interaction with the lattice atoms. So far, radiative indirect transitions of high efficiency have been achieved primarily by the inclusion of isoelectronic imperfections or by use of ultrapure GaP crystals. Isoelectronic impurity atoms, e.g., nitrogen, can replace the phosphorus atoms, or special binary "molecules," e.g., zinc – oxygen pairs, can replace adjacent gallium – phosphorus pairs without affecting the charge equilibrium (Table 2). Such uncharged imperfections may lead to a more localized binding of an electron whose momentum, due to the uncertainty principle, is blurred over the entire momentum range (E_i in Fig. 2). Radiant electron transitions from the conduction (X_1 on E_c) to the valence band (Γ_8 on E_v) are now possible via these or other states that lie in the forbidden energy band.

Which of these two recombination processes predominates in a given case is determined by the transition probability. For nonradiative transitions via charged imperfections, this probability increases with the density of the imperfections, lowering light efficiencies at high doping density even in direct semiconductors. In semiconductors with a small gap between the valence

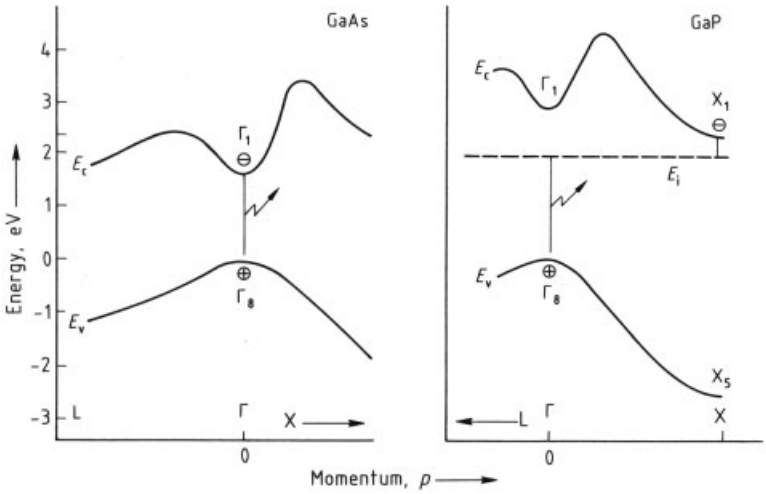

Figure 2. Important photoemission regions of the energy band structure for GaAs, a direct semiconductor, and GaP, an indirect semiconductor, showing $\ominus \rightarrow \oplus$ light emissive recombination processes

Table 2. Indirect semiconductor materials for luminescent diodes: term transitions

Crystal	Doping	Wavelength, µm	Emission color	Luminous efficacy, lm/W
GaP [*12063-98-8*]	Zn, O pairs	0.69	red	3.0
GaP	N	0.57	green	4.5
GaP	N, N pairs	0.59	yellow	0.5
GaAs–GaP, P : As>0.82	N	0.55 – 0.67		
$GaAs_{0.35}P_{0.65}$	N	0.63	orange	0.9
$GaAs_{0.15}P_{0.85}$	N	0.59	yellow	0.9
GaN [*25617-97-4*]	Zn		blue	0.03
ZnS [*1314-98-3*]	Al		blue	0.05
SiC [*409-21-2*]	Al, N pairs		blue	0.01

and conduction band, nonradiative recombination processes predominate even at room temperature, so that luminescent diodes with long-wavelength emission have practical efficiencies only at low temperature.

A laser effect occurs at very high excitation, i.e., when the densities of the electrons and holes exceed a threshold in the excited region. The interaction between the crystal and the emitted light changes from the normally predominant absorption effect to a state of stimulated emission.

The wavelength of light attainable with luminescent diodes would be restricted by the intrinsic values of the bandgaps E_g of known semiconductor materials and by the attainable electronic states in the forbidden band caused by known dopants were it not possible to realize continuous series of mixed crystals with continuous values of the bandgap E_g, as is shown in Figure 3 for III – V semiconductor materials.

Table 1 lists the most important photoemissive semiconductor materials and their mixed crystal series along with the light wavelength attainable in the direct range of the band gaps. Indirect materials (Table 2) can be used for other colors when isoelectronic dopants are present. In addition, the emission lines of GaP-based diodes can be continuously shifted toward longer wavelengths by partial replacement of P by As.

3. Manufacture of Crystal Layers for Luminescent Devices

The most important aspect of manufacturing luminescent diodes is to set the desired wavelength. In practice, however, the wavelengths in Table 1 may be set only under certain conditions.

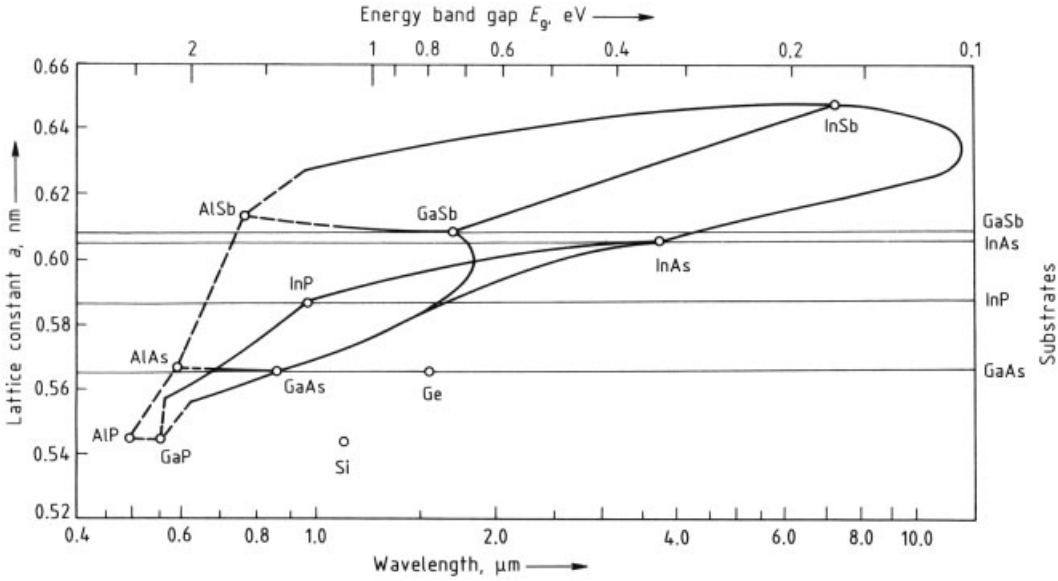

Figure 3. Lattice constant vs. wavelength (bandgap energy) for important III – V semiconductors
Solid curves show compounds with direct recombination; dashed curves show regions of indirect transition. Binary substrates that can be used for lattice-matched layer growth are indicated by the straight horizontal lines.

The nonbinary crystals, i.e., mixed crystals, cannot be obtained by the procedures used for manufacturing bulk binary crystals. Instead, mixed crystals must be grown by epitaxial processes on binary substrates with optimally matched lattice constants, this being especially important for laser diodes. However, the choice of possible substrates is extremely limited. In Figure 3, the lattice parameters for GaAs, InP, InAs, and GaSb are plotted as horizontal lines. The concentration of Al can be freely selected only in the AlAs – GaAs system; in all other ternary systems, only small regions in the vicinity of one of the binary components are permissible. The only way of obtaining other emission wavelengths is to use quaternary mixed crystals to avoid the enormous lattice distortions that are the main cause of aging, especially in laser diodes, and thus the drastic reduction of the emission efficiency during operation. A number of deposition procedures are available for epitaxial growth of binary, ternary, and quaternary layers on suitable substrate crystals.

Epitaxy from a metal melt containing the required components in the correct composition is the most advantageous approach in terms of equipment outlay. This process, known as *liquid-phase epitaxy* (*LPE*), uses the trivalent metal components, e.g., Ga or Ga – Al, as solvents for the pentavalent components. Thus, a Ga – Al melt saturated with As or P is applied onto a GaAs or GaP substrate at a high temperature or generated by dissolving a thin layer of the substrate. If the system is allowed to cool slowly, an epitaxial semiconductor layer grows until the remaining melt is pushed away. This process can be carried out several times in succession, with variously doped melts, or the various doping materials can be successively introduced via the gas phase.

Liquid-phase epitaxy has proved effective for Si-doped GaAs diodes, in which the doping character of the silicon changes at a specific temperature during the cooling process from a predominantly donor mode to a predominantly acceptor mode, resulting in the *p-n* junction required for charge carrier injection. Liquid-phase epitaxy is also used predominantly for production of laser diodes with several heterojunctions between GaAs and (GaAl)As, the individual layers having different Al contents. For N-doped and ZnO-doped GaP diodes that radiate in the visible region, LPE is the usual method if maximum efficiency is a prime concern.

Because of its simpler handling and greater versatility with respect to specific compositions, *vapor-phase epitaxy* (*VPE*) is being used increasingly for mixed crystals, especially if components with high vapor pressure are involved. However, VPE makes considerable demands in terms of equipment sealing and purity of the gases used where the metal components, e.g., Al, are highly sensitive to oxygen. Vapor-phase epitaxy is also used for the manufacture of Ga(As, P) LEDs, since the As : P ratio in the growing layer can be regulated via the AsH_3 and PH_3 concentration in the gas phase.

Gallium is introduced by reacting liquid gallium with HCl at high temperature, which preferentially produces GaCl. On the substrate crystal, which is at low temperature, GaCl disproportionates to Ga and $GaCl_3$. On the substrate the resulting gallium combines with the arsenic and phosphorus molecules (As_n and P_n) present in the gas. Instead, $AsCl_3$ and PCl_3 can be reacted with gallium to provide GaCl, As_n, and P_n. Gaseous dopants such as H_2S, H_2Se, and H_2Te are suitable for producing *n*-type conduction; zinc or an organozinc compound produces *p*-type conduction. Nitrogen is introduced as an active isoelectronic center by means of NH_3. Frequently, however, *n*-type layers produced by VPE are converted to a *p-n* junction by diffusion treatment in zinc vapor or by introduction of zinc ions by ion implantation and subsequent annealing. Both processes can be carried out either across the whole area or locally by means of a window etched into an SiO_2 cover layer similar to the planar technology processes used in microelectronics.

In VPE, increasing use is being made of gases to transport the metal components. Organometallic compounds such as trimethyl- or triethylgallium and the analogous aluminum or indium compounds are especially suitable for this purpose. The designation VPE then becomes *MOVPE.*

Since optoelectronic integrated circuits and integrated optics require, in the limiting case, atomically smooth and thin layers and layer sequences (superlattice structures), a new technology was developed for their production. *Molecular beam epitaxy* (*MBE*) or *organometallic molecular beam epitaxy* (*MOMBE*) operates at ultrahigh vacuum (UHV) with Knudsen evaporators, which can be adjusted extremely accurately, and liquid or gaseous compounds of the components, e.g., metal alkyls, as the material source. Local doping can be implemented by ion implantation or, for fine structures, ion lithography. The necessary equipment can be integrated directly into the UHV unit. In this way, a fully automated UHV technology is available for the most advanced electroluminescence components.

4. Practical Luminescent Diodes

4.1. Diodes Producing Red, Yellow, and Green

The main use for luminescent diodes is emission of visible light. Here they have replaced earlier incandescent filaments and gas discharge lamps, attaining good color contrast and intensities adapted to eye sensitivity. Systems emitting red, yellow-orange, and green light have been selected for practical applications. Red diodes (Table 3) consist of $GaAs_{0.6}P_{0.4}$ on opaque GaAs or, better still, N-doped $GaAs_{0.35}P_{0.65}$ on a transparent GaP substrate. Yellow-orange is obtained optimally with $GaAs_{0.15}P_{0.85}$ and green with GaP, both on GaP substrate and N-doped. Table 2 also lists the color and the best light yield, i.e., the number of lumen for every watt of input power, for the most important materials. The values for commercial luminous diodes are around 0.2 lm/W for GaAs substrate; however, for GaP substrate they are better than 0.5 lm/W.

After the epitaxial layers have been produced, contact metals are vaporized onto the crystals: an Au or Ag alloy containing Sn, Si, or Ge for the *n* side, and an Au or Ag layer containing Zn, Be, or Mg for the *p* contact. After the crystal has been split by diamond cleavage – splitting along (110) crystal faces – or by sawing with diamond-impregnated rotary disks or wire saws, the diode systems are alloyed onto a conductive band or glued with a conductive adhesive. Gold wire is used for the electrical connection on the upper side.

The diode is subsequently encapsulated with a plastic that has two main functions: (1) to protect the system against mechanical damage, and (2) to gradually adapt the refractive index to air, which increases the external light yield.

Table 3. Direct semiconductor materials for commercial red and infrared light emitting diodes: band-to-band transitions

Crystal system	Emission wavelength, μm	Luminous efficacy, lm/W	Efficiency∗, %
Red light emitting			
GaAsP on GaAs substrate	0.66	0.4	
GaAlAs on GaAs substrate	0.65	5.0	
GaAlAs on GaAlAs substrate	0.65	15.0	
Infrared emitting			
GaAs : Si	0.94		6
GaAs : Zn	0.91		1
GaAlAs	0.88		12
	0.82		2
GaInAsP	1.30		0.5
GaAsP	0.70		0.5

∗ Photons emitted per injected current carrier (electron).

The semiconductor crystal itself has a refractive index n of 3 – 4, and without plastic encapsulation a large part of the light generated would be lost because of total internal reflection.

Special shaping of the plastic surface improves the focusing of the emitted light, and finely dispersed light-scattering particles of SiO_2 or TiO_2 in the plastic produce a uniformly luminescent surface. In addition, this plastic can be tinted to improve the contrast and change the color. For example, N-doped GaP diode systems that have a natural yellow-to-green emission can be made into yellow or green diodes. However, there is an accompanying loss in efficiency.

4.2. Blue Light Diodes

The indirect semiconductor silicon carbide, which has a bandgap of 2.98 eV for the polytype 6H, is of practical importance for blue light-emitting diodes. This material may be produced in n- and p-conducting forms by doping, which distinguishes it from other materials, such as GaN, ZnS, or ZnSe, that might be used for emission in the blue region of the spectrum. The effective photoemission mechanisms have not been fully explained. Transitions between the conductance and valence bands, possibly including transitions via donors and acceptors (e.g., N – Al pairs emitting at 0.5 μm) have been proposed as possible mechanisms. The optical yield of these devices is small, with an external efficiency <0.01 %. However, SiC has a high current carrying capacity, making high luminescence values attainable.

Blue-light diodes of this kind are not yet available commercially. Their production involves considerable difficulties, since large SiC crystals must be used as the substrate for the doped p-n layer sequences, which are manufactured by liquid-phase epitaxy.

The requisite crystals and p-n layer sequences must be crystallized out from silicon melts containing carbon and dopants (Al, N) above 1800 °C. The problems associated with this technology have been solved, but this high-temperature method proceeds slowly and is costly. For the time being blue-light diodes can be introduced only for special applications.

4.3. Infrared Diodes

For some applications, e.g., optical communications and light barriers, GaAs diodes emitting between 0.8 and 1.1 μm are used (see Table 3). Since high output power is often important, the high refractive index and the self-absorption of these devices call for special measures. All light that strikes the boundary with an angle of reflection greater than that of total reflection returns to the crystal interior and is lost by self-absorption. A hemispherical diode surface provides a good solution, allowing an output power exceeding 1 W at 30 % efficiency. Since optical paths in the crystal are quite long, the absorption must be reduced by shifting the absorption edge to shorter wavelength by high n-doping and by red-shifting the light emission by simultaneous high p-doping using amphoteric silicon or a combination of donors and acceptors as dopants.

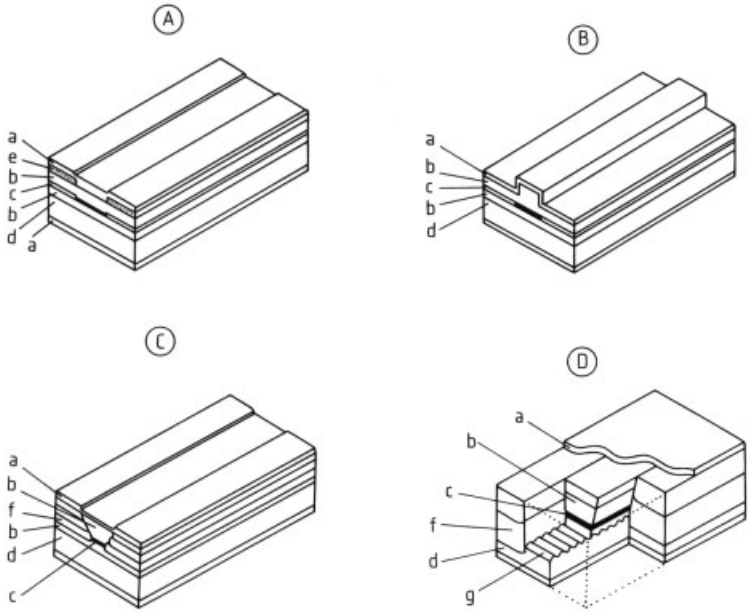

Figure 4. Important diode laser structures
A) Oxide stripe; B) Metal-clad ridge waveguide (MCRW); C) Buried heterostructure (BH); D) Distributed feedback (DFB)
a) Contact metallization; b) Cladding layer, e.g., InP; c) Active layer, e.g., InGaAsP; d) Substrate, e.g., InP; e) Oxide mask layer; f) Burying layer, e.g., InP; g) Grating

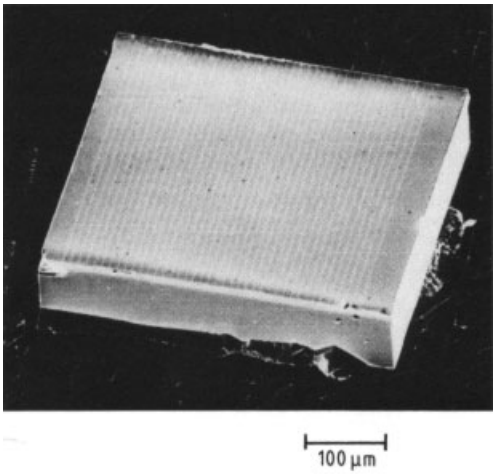

Figure 5. High-power parallel stripe laser (MCRW type structure)

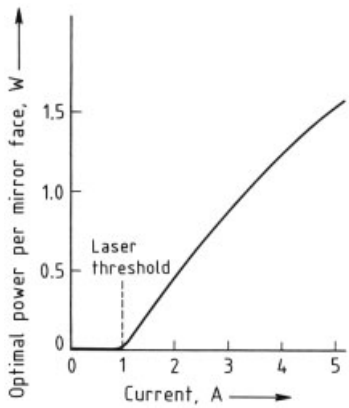

Figure 6. Continuous output power on each side of the high-power parallel stripe laser as a function of current

4.4. Optoelectronic Couplers

If the light is required only for electrical decoupling within an electrical circuit or for simple coupling of optoelectronic circuits, an emission diode and a receiver (photodiode or transistor) can be connected via a highly refractive medium having high insulation strength and long-term stability, to produce a practical coupling device.

Apart from plastics with a refractive index of 1.4 – 1.8, semiconductive glasses, e.g., the (As, S, I) and (Ge, P, Se) compounds with refractive

Table 4. Typical operating parameters for laser structures

Structure	Driving cw∗ current, mA	Emitted cw∗ power, mW	Pulse power (length <100 ns), mW	Emission spectrum	Far field pattern ∗∗	
					⊥	‖
Oxide stripe	100	5 – 10	300	multimode 30 lines 200 nm wide	50 °	30 °
Metal-clad ridge waveguide (MCRW)	80	5 – 10	300	multimode 5 lines 2 nm wide	50 °	15 °
Buried heterostructure (BH)	70	5 – 10	300	multimode 5 lines 2 nm wide	50 °	20 °
Distributed feedback laser (DFB)	90	5 – 10	300	monomode 1 line <0.01 nm wide	50 °	20 °
High power arrays	1000	50 – 200	2000	multimode 5 lines 2 nm wide	50 °	5 °

∗ cw, continuous wave operation.
∗∗ Emission angle perpendicular (⊥) or parallel (‖) to the laser chip.

indexes up to 2.4 (→ Glass) are candidates to connect the emitter with the detector.

5. The Laser Diode as a Source of Coherent Light

Laser diodes can be realized with various semiconductors. Lasing can be recognized by the rapid increase in emission accompanied by a decrease in line width when a specific threshold current, corresponding to a current density of ca. 1 kA/cm^2, is exceeded.

The active volume, i.e., the layer excited by carrier injection, is limited by two adjacent layers with somewhat higher aluminum content, and thus greater bandgaps, which reduce the threshold current density. Since the refractive index also drops at both inner junctions, the emitted light is not deflected into these regions, and wave conduction occurs in the active layer. This reduces the optical losses and reduces the threshold current density even more. The threshold current is reduced in absolute terms even further by a reduction of the width of the excited area of the diode and by extremely effective heat dissipation (e.g., by diamond). This stripe geometry also allows monomode operation, i.e., only a single oscillation mode is excited in the diode. Since most of the generated light is available at the front windows, the diode has high external efficiency, up to 50 %, and high radiation densities, $10^4 - 10^6$ W/cm^2. The maximum operating temperature increases as the threshold current is decreased.

Figure 4 schematically represents the four most important structures involving GaAs lasers (emission at ca. 850 µm) [17]: the oxide stripe laser, the metal-clad ridge waveguide (MCRW), the buried heterostructure (BH) laser, and a planar buried ridge structure (PBRS) laser with a distributed feedback (DFB) lattice structure. The first three are Fabry-Perot resonators, while the last is a grating resonator. In the oxide stripe laser, the current is confined by a window stripe in an insulating oxide cover layer. In the MCRW laser, the thickness of the light-emitting layer is limited by two heterojunctions. The area of the active layer is determined by the width of a narrow ridge, which simultaneously confines lateral current flow and the propagating optical wave. In addition, the active area of the buried heterostructure (BH) laser is limited by two vertical heterojunctions provided by the burying layer. In the DFB laser, a buried heterostructure, the active region, is confined vertically and horizontally, like the BH laser. The grating serves for feedback of radiation along the resonator axis and provides for stable single-mode operation. In all these cases, to allow steady operation at room temperature, the operating current must be kept as small as possible, and heat dissipation must be good. Available power and emission wavelength are the most important parameters for laser-induced chemistry, whereas narrow line width is important for chemical analysis.

Table 4 lists operating data for the laser structures in Figure 4 as well as for high-power arrays. Particularly interesting is the monolithic

integrated laser diode array produced from many (40 in this case) parallel stripe lasers with a spacing of about 0.01 mm, as shown in Figure 5. It is a GaAs power laser with a total output power for both mirrors of 3 W and a permissible constant operating current of 5 A at 300 K (Fig. 6).

Uses. Laser diode devices are used in optical communications for local-area networks and in wideband communications (integrated services digital network, ISDN). Long distances, over 100 km without repeaters, and high data rate, exceeding 1 Gbit/s, have been realized. In computer technology, power lasers are used to write or read information on optical memories, the most important example being the digital compact disk (CD), while applications for video systems and large memories (100 Gbyte) are in the offing. Even more powerful arrays will be required for laser printing, material machining, and microsurgery. Laser diodes provide more efficient resonant pumping for solid-state lasers, an efficiency increase by a factor of 10^3. Laser chemistry, i.e., the induction of specific chemical reactions, should extend the applications of semiconductor lasers significantly. The 1985 sales figure for laser-linked products, \$ 3.5×10^9 (€ 3.3×10^9), is expected to grow by 50 % per year.

6. References

General References

1. H. K. Henisch: *Electroluminescence,* Pergamon Press, London 1962.
2. C. Hilsum : "Gallium Arsenide," in A. F. Gibson, R. E. Burgess (eds.): *Progress in Semiconductors,* vol. 9, Temple Press Book, London 1965, pp. 137 – 177.
3. N. Riehl: *Einführung in die Lumineszenz,* Thiemig, München 1971.
4. J. I. Pankove: *Optical Processes in Semiconductors,* Prentice-Hall, London 1971.
5. C. H. Gooch: *Injection Electroluminescent Devices,* J. Wiley & Sons, London 1973.
6. I. Runge : "Halbleitertechnologie," in W. Heywang, R. Müller (eds.): *Reihe Halbleiter-Elektronik*, vol. 4, Springer Verlag, Berlin-Heidelberg-New York 1975.
7. M. Bleicher: *Halbleiter-Optoelektronik,* Hüthig Verlag, Heidelberg 1976.
8. A. A. Bergh, P. J. Dean: *Light-Emitting Diodes,* Clarendon Press, Oxford 1976.
9. H. Kressel, J. K. Butler: *Semiconductor Lasers and Heterojunction LED,* Academic Press, London-New York 1977.
10. H. C. Casey, M. B. Panish : Heterostructure Lasers, Parts A and B, Academic Press, London-New York 1978.
11. S. E. Midwinter: *Optical Fibers for Transmission,* Wiley Interscience, New York-Chichester-Toronto 1979.
12. C. P. Sandbank: *Optical Fiber Communication Systems,* Wiley Interscience, New York-Chichester-Toronto 1980.
13. G. Winstel C. Weyrich : "Optoelektronik I (Lumineszenz und Laserdioden)," in W. Heywang, R. Müller (eds.): *Reihe Halbleiter-Elektronik*, vol. 10, Springer Verlag, Berlin-Heidelberg-New York 1980.
14. K. Ploog : *Crystal Growth, Properties and Applications,* vol. 3, Springer Verlag, Berlin-Heidelberg-New York 1980.
15. R. T. Kersten: *Einführung in die optische Nachrichtentechnik,* Springer Verlag, Berlin-Heidelberg-New York 1983.
16. *Londolt-Börnstein,* New Series Group III, vol. 17 d, "Technology of III – V, II – VI and Nontetrahedrally Bonded Compounds," 1984.

Specific References

17. D. Botez: "Laser diodes are power packed," *IEEE Spectrum* **22** (1985) no. 6, 43 – 53. T. E. Bell: "Single-frequency semiconductor laser," *IEEE Spectrum* **20** (1983) no. 12, 38 – 45.

Electronic Ceramics → Ceramics, Electronic

Electronic Chemicals

STEVE DAIGLE, Charles H. Kline & Co., Fairfield, New Jersey 07006, United States

ERIC VOGELSBERG, Charles H. Kline & Co., Fairfield, New Jersey 07006, United States

BIRON LIM, Charles H. Kline & Co., Fairfield, New Jersey 07006, United States

IAN BUTCHER, Charles H. Kline & Co., Fairfield, New Jersey 07006, United States

1. Introduction 2
2. Substrates 3
2.1. Silicon Wafers 3
2.2. Other Semiconductor Substrates . . 4
2.3. Printed Circuit Board Laminates . . 4
3. Packaging Materials 5
3.1. Ceramic Packaging Materials 5
3.2. Plastic Packaging Materials 6
4. Plating Chemicals 7
4.1. Electrolytic Gold Plating Chemicals 7
4.2. Electroless Copper Plating Chemicals 8
4.3. Electrolytic Copper Plating Chemicals 9
4.4. Electrolytic Tin – Lead Plating Chemicals 10
4.5. Other Electrolytic Plating Chemicals 10
5. Photoresists 11
5.1. Dry Film Photoresists 11
5.2. Liquid Positive and Negative Photoresists 12
5.3. Screening Inks 13
6. Acids and Etchants 13
6.1. Sulfuric Acid 13
6.2. Hydrofluoric Acid 14
6.3. Ammoniacal Solutions 15
6.4. Nitric Acid 15
6.5. Hydrogen Peroxide 16
6.6. Proprietary Peroxy and Sulfuric Etchants 17
6.7. Ammonium Fluoride 17
6.8. Persulfates 18
6.9. Other Acids and Etchants 18
7. Atmospheric Gases 20
8. Photoresist Adjuncts 21
8.1. Photoresist Strippers 21
8.2. Dry Film Photoresist Developers . . 22
8.3. Positive Photoresist Developers . . . 22
8.4. Negative Photoresist Developers . . . 23
8.5. Photoresist Adhesion Promoters . . . 23
9. Cleaners and Specialty Materials . . 24
10. Specialty Gases 25
10.1. Silicon Gases 25
10.2. Etchant Gases 26
10.3. Light-Emitting Diode Gases 27
11. Solvents 28
12. Masks 29
13. Dopants 30
13.1. Liquid Dopants 30
13.2. Gas Dopants 31
13.3. Solid Dopants 31
14. Auxiliary Materials 32
14.1. Ceramic Substrates 32
14.2. Liquid Crystals 32
14.3. Conductive Inks and Pastes 32
14.4. High-Purity Metals 32
14.5. Passivation Materials 33
14.6. Wafer Cleaners 33
15. References 33

1. Introduction

The chemicals and materials used to manufacture and package semiconductors and printed circuit boards are considered electronic chemicals. The more than 100 different products used in these applications are grouped into the following 13 major product categories:

acids and etchants (Chap. 6)
atmospheric gases (Chap. 7)
auxiliary materials (Chap. 14)
cleaners (Chap. 9)
dopants (Chap. 13)
masks (Chap. 12)
packaging materials (Chap. 3)
photoresists (Chap. 5)
photoresist adjuncts (Chap. 8)
plating chemicals (Chap. 4)
solvents (Chap. 11)
specialty gases (Chap. 10)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_265

substrates (Chap. 2)

Semiconductors are devices having a specific electronic circuit fabricated onto a single square or rectangular piece of semiconducting material such as silicon (→ Semiconductors). The two main types of semiconductor devices are *discrete devices* and *integrated circuits* (IC). Discrete devices contain only one active element, whereas integrated circuits contain multiple elements that act together to form a complete electronic circuit. Since the semiconductor industry began in 1947 with the invention of the transistor and was further enhanced by the invention of the integrated circuit in 1958, the business has grown rapidly.

More than 100 manufacturers of semiconductor devices exist in the United States, including both merchant and captive producers. International Business Machines (IBM), a captive producer, is believed to be the largest semiconductor manufacturer worldwide. Other important semiconductor manufacturers include Texas Instruments, Motorola, National Semiconductor, Intel, Fairchild, Signetics, AT & T, and Advanced Micro Devices in the United States; Philips (Netherlands); Siemens (Federal Republic of Germany); and NEC, Hitachi, Toshiba, and Fujitsu (Japan).

Printed circuit boards are plastic sheets onto which conductive paths are formed in a specific pattern for the purpose of interconnecting electronic components such as semiconductor devices, resistors, and capacitors. The circuitry can be formed on either one side or both sides of the board, and several boards can be connected to produce a multilayer board. The need for printed circuit boards is widespread, and they are either produced captively by large consumers or purchased on the merchant market.

Of the over 1000 individual manufacturers of printed circuit boards in the United States, two companies, IBM and Western Electric, are by far the largest producers. Other important manufacturers include Hewlett-Packard, Digital Equipment, Delco, Tyco Laboratories, Hadco, and Photocircuits (United States); GEC (United Kingdom); Telefunken (Federal Republic of Germany); Olivetti (Italy); Thompson-CSF (France); and NEC, Hitachi Chemical, CMK, and Fujitsu (Japan) [4].

2. Substrates

A substrate is the base material in the manufacture of semiconductor devices and printed circuit boards. The semiconductor industry uses materials, such as silicon and gallium arsenide, which are produced in single crystal form. Thin circular slices or wafers of these single crystal materials serve as the substrate in manufacturing semiconductor devices.

The printed circuit board industry uses materials known as laminates to act as the base in the production of circuit boards. Combinations of a resin with either paper or glass cloth form the laminate, which is produced in thin sheets.

2.1. Silicon Wafers

Types and Grades. Silicon [*7440-21-3*] wafers are sold as either polished wafers or epitaxial wafers, i.e., wafers having an added silicon layer over the surface, that has the same crystallographic orientation as the wafer itself. All wafers are manufactured with dopants embedded in the silicon. Wafers positively doped with boron or negatively doped with either antimony or phosphorus are available. Silicon wafers are also sold in two types of crystal lattices: (100) and (111) orientations.

Production. The two major methods of producing silicon wafers are the Czochralski method and the Floating Zone technique (→ Crystal Growth).

The *Czochralski method* is the most widely used crystal growing process. Pure polycrystalline silicon is placed in a fused silica quartz crucible. The material is heated to 1400 °C in a protective atmosphere, usually argon, which melts the silicon and keeps it molten. A seed crystal of the desired crystal orientation is dipped into the melt. The single silicon crystal is grown by rotating the seed and the crucible in opposite directions while withdrawing the seed.

The *Floating Zone technique* uses a rod of polycrystalline silicon attached to a chuck at the top of an inert atmosphere chamber and a seed crystal attached to a shaft at the bottom of the chamber. A heating coil located at the base of the

rod forms a molten zone, which is fused with the seed to begin crystal growth. Movement through the molten zone purifies the polysilicon and establishes the crystal structure. The molten silicon is kept in place by surface tension, which limits the diameter of the pure silicon.

The major drawback of the Floating Zone method in comparison to the Czochralski process is the limited wafer diameter. The Floating Zone process can produce a maximum wafer diameter of 7.6 cm, whereas wafer diameters up to 15.2 cm are feasible with the Czochralski method. Both processes add controlled amounts of impurities or dopants during the growth process to create the necessary electrical properties within the silicon crystal structure. Dopants, such as boron, are added to establish a positive charge within the silicon; phosphorus and antimony are used to create a negatively charged material. In the Czochralski process, dopants are typically added to the molten silicon as pure elements or concentrated silicon solutions. The Floating Zone method uses dilute phosphine or diborane gas, which is added when the rod is in the molten zone or when the rod is initially manufactured.

Quality Specifications. The dimension and thickness specifications for polished silicon wafers have been established by the Semiconductor Equipment and Materials Institute (SEMI) [5].

Uses. Silicon is used as a substrate for discrete semiconductor devices, such as diodes, rectifiers, and transistors, as well as for all types of integrated circuits (IC). Wafers produced by the Czochralski process are used primarily for ICs and transistors, whereas those manufactured by the Floating Zone method are used mainly for rectifiers and thyristors.

Economic Aspects. The consumption of silicon wafers in the United States was ca. $3\times10^9\,cm^2$ in 1980. The demand in Western Europe during the same period was 0.742 $\times10^9\,cm^2$. The leading suppliers worldwide of silicon wafers are Wacker (Federal Republic of Germany), Monsanto (United States), Shin-Etsu Handotai (Japan), and Dynamit Nobel (Italy).

2.2. Other Semiconductor Substrates

Gallium arsenide [*1303-00-0*], GaAs, is a substrate produced by the Czochralski process from a molten mix of gallium and arsenic metals. It is sold as wafers, with diameters of 5.08 and 7.62 cm being the most popular. Most GaAs is used as the base material for light-emitting diodes, although it does have applications with other discrete devices, such as field-effect transistors, Gunn diodes, and Schottky diodes. Gallium arsenide is also an electrically fast substrate material and is of interest as a potential source for producing very-high-speed integrated circuits.

Some of the more important suppliers of gallium arsenide include Sumitomo Electric Industries and Mitsubishi Monsanto (Japan), Morgan Semiconductor (United States), and Wacker (Federal Republic of Germany).

Silicon-on-sapphire (SOS) is produced by epitaxially depositing a thin silicon layer on a sapphire substrate. It is sold as wafers, with the 7.62 cm diameter variety being the most common. As a substrate material, SOS is capable of producing dense, high-speed devices that are resistant to radiation. However, its manufacture is both difficult and expensive. Despite attempts to develop SOS for commodity type devices, it remains mainly as a material for specialty devices in military applications.

The United States is the main market for SOS substrates as a result of its military applications. Union Carbide and Sappikon are the major U.S. suppliers of SOS substrates. Small quantities of SOS are used in Japan mainly by Toshiba, which supplies some of its internal requirements.

Gadolinium gallium garnet (GGG) [*12024-36-1*], $Gd_3Ga_5O_{12}$, is manufactured from gadolinium oxide and gallium oxide by the Czochralski method. Single-crystal GGG is available in wafer form with a typical diameter of 7.62 cm and thicknesses of 5 – 6 µm. The main application for GGG is as the substrate material for magnetic bubble memories, a relatively new device that can store $>10^6$ bits of information. Gadolinium gallium garnet (GGG) is electrically involatile; however, it is a relatively slow electrical semiconductor. Unfortunately, the de-

mand for magnetic bubble memories has been limited.

2.3. Printed Circuit Board Laminates

Laminates are used as the substrate for printing electrical circuitry. They are composed of layers of material compressed under heat and pressure to form a consistent electrical insulator. The basic raw materials of the laminate include paper and glass cloth, or glass mat bonded with one of several resins, mainly phenolic, epoxy, or polyester. Circuit boards processed by the substractive method use laminates clad with copper foil, whereas the additive and semiadditive methods use unclad laminates,

Types and Grades. Laminates are manufactured specifically for circuit printing on both sides (double-sided board) or an only one side (single-sided board) and can be either rigid or flexible in form. The types of laminates available commercially are given in Table 1 [6].

Uses. Paper-based phenolic and composite laminates are used as substrates for single-sided and simple double-sided circuit boards. These boards are typically used for less critical consumer items, such as electronic games, televisions, appliances, calculators, and automotive circuits. Glass-based epoxy and polyimide laminates are favored for complex double-sided and multilayer boards that have applications with computers, communication equipment, and other electronic devices requiring a high degree of reliability.

Economic Aspects. Total demand for printed circuit board laminates in the United States was $20\times10^6\,m^2$ in 1980. The demand in Western Europe was $15\times10^6\,m^2$ during the same period. The major laminate suppliers include Norplex and Oak Laminates (both of which are part of Allied-Signal), Westinghouse, and General Electric (United States); Dynamit Nobel and Isola (Federal Republic of Germany); MAS (Italy); and Hitachi Chemical, Sumitomo Bakelite, and Matsushita Electric (Japan).

3. Packaging Materials

3.1. Ceramic Packaging Materials

Ceramic materials are used to encase semiconductor devices to protect the devices from the adverse effects of moisture, atmospheric salt, and high temperature (→ Ceramics, Electronic). Because ceramic packages provide a high degree of hermaticity and reliability, ceramics are generally used to package critical military, computer, and communication devices.

Most ceramics used in packaging are made from aluminum oxide [*1344-28-1*], Al_2O_3, although beryllium oxide [*1304-56-9*], BeO, is sometimes used. Aluminum oxide is very strong and refractory and has a low dielectric constant, which makes it suitable for most packaging applications. Beryllium oxide offers the unusual combination of high thermal and electric conductivity. However, the high cost of this material limits its use to the most critical applications.

Types and Grades. Most ceramic packages contain 94 wt % aluminum oxide, although other compositions are available. The major types of ceramic packages are (1) the metallized ceramic dual in-line package (DIP), (2) the ceramic dual in-line package (CERDIP), and (3) the chip carrier. Other types include specialty beryllium oxide packages.

Uses. The type of ceramic package used with a semiconductor device depends on (1) the level of hermetic seal desired, (2) the degree of reliability required, (3) the amount of space alloted to the device on the printed circuit board, and (4) the total package cost. A ceramic package that is metallized by cofiring the alumina powder with tungsten forms the tightest hermetic seal of all types of packages, whether ceramic or plastic. However, this package is also the most expensive, except for the seldom used beryllium oxide-type package. Because of their high reliability, DIPs are most often used for devices used by the military and for the more critical devices used in computers and communication equipment.

The *CERDIP,* which has no metallization, offers a compromise between the level of hermetic seal and package cost. As a result, the CERDIP is favored for many computer and communication applications where a more hermetic seal is

Table 1. Standard laminate materials

NEMA grade	MIL-P-13949E∗	Resin system	Base	Color	Description
XXXP	none	phenolic	paper	opaque brown	phenolic paper with no cold-punch stipulation
XXXPC	none	phenolic	paper	opaque brown	phenolic paper with punchability at or above room temperature
FR-2	none	phenolic	paper	opaque brown	phenolic paper, punchable, with flame-resistant (self-extinguishing) resin system
FR-3	PX	epoxy	paper	opaque cream	epoxy resin, paper base with flame-resistant resin system, cold punching, and high insulation resistance
none	none	epoxy	paper – glass composite	opaque cream	epoxy resin paper core with glass on the laminate surface, self-extinguishing, economical fabrication of paper base, mechanical characteristics of glass
FR-6	none	polyester	glass mat	opaque white	polyester, random glass fiber, flame-resistant, designed for low-capacitance or high-impact applications
G-10	GE	epoxy	glass	translucent green or white	epoxy – glass, general purpose
FR-4	GF	epoxy	glass	translucent green or white	epoxy – glass with self-extinguishing resin system
G-11	GP	epoxy	glass	translucent green or white	high-temperature epoxy – glass with strength and electrical rentention at elevated temperature
FR-5	GH	epoxy	glass	translucent green or white	high-temperature epoxy – glass with flame-resistant resin system; strength and electrical retention at elevated temperature
		polyimide	glass	translucent dark brown	polyimide resin, glass laminate with high continuous operating temperature and high property retention at high temperature, low z-axis dimensional expansion

∗ Military designation.

desired than plastic can be obtained using a package, but where cost limitations prevent use of a DIP.

The *chip carrier* was developed to fulfill the needs of systems manufacturers for (1) increased package density, (2) better electrical performance, and (3) more economical device packaging. It has a metallized aluminum oxide ceramic substrate similar to that of a DIP. Its shape is square, whereas both the DIP and CERDIP are rectangular. The square configuration of the chip carrier reduces the space requirements on a printed circuit board by 3 – 5 times compared to that of the DIP and CERDIP. The square shape also improves electrical performance by shortening the length of the electrical leads and by making the lead length more uniform over the entire package, thus permitting faster switching times and improved system performance. Overall, the chip carrier produces a hermetic seal nearly as tight as that of the DIP, but at < 50 % the cost.

Economic Aspects. In 1980, ca. 2.2×10^9 ceramic packages were used in the United States for packaging semiconductor devices. At the same time, ca. 87×10^6 ceramic packages were consumed. The two leading suppliers of ceramic packages are Kyocera and NTK Technical Ceramics, both Japanese companies.

3.2. Plastic Packaging Materials

Plastic packaging materials are used to encase the finished semiconductor device, thus protecting the device from moisture, atmospheric salt, and high temperature. The four basic types of plastic materials used for this purpose are as follows:

Epoxy molding resins are used primarily to package metal oxide semiconductor (MOS) and bipolar devices that require an inexpensive package while still receiving protection from the environment (→ Epoxy Resins). Epoxy resins offer low shrinkage and high strength, which help to prevent package cracking both on curing and from the thermal shock.

Silicone molding resins are favored for packaging devices, such as power transistors, which generate a great amount of heat. Silicone resins are among the best of all polymeric materials for operating at high temperature. They also have excellent electrical properties, which change little with temperature.

Silicone-epoxy resins are hybrids that attempt to combine the best qualities of each resin while maintaining a low package cost. So far, these resins have been used with some telecommunication devices that seek a better hermetic seal than that from epoxy resins while maintaining a package cost close to that of epoxy.

Liquid potting compounds are generally used to package discrete devices such as transistors and diodes. These potting compounds offer the lowest package cost, but also the least amount of protection from the environment.

Types and Grades. Plastic packaging materials are all sold as semiconductor-grade products. The molding resins are available in granular form, whereas potting compounds are sold as liquids. All products are marketed in ready-to-use form.

Uses. Two main processes are used for packaging semiconductor devices.

Transfer molding is the process of transferring a hardenable material, under pressure, from a pot or container into the mold that contains the semiconductor device. This process is economical for large-volume operations, but initial facility construction and mold costs are high. All molding resins use this process for packaging devices.

Potting consists of pouring a hardenable liquid into a shell or housing, which remains as an integral part of the package. This process is also excellent for large-volume runs and requires minimal tooling. However, this process is subject to limitations of both size and housing complexity. Liquid potting compounds are used exclusively in this packaging process.

Economic Aspects. Worldwide consumption of plastic packaging materials in 1980 is shown in Table 2. Epoxy-based molding compounds were the preferred product.

Table 2. Worldwide consumption of plastic packaging materials in 1980

Product	Consumption, 10^3 t		
	United States	Western Europe	Total
Epoxy	10.0	1.4	11.4
Silicone	1.1	0.2	1.3
Liquid potting	0.6		0.6
Silicone – epoxy	0.2		0.2
Total	11.9	1.6	13.5

The major suppliers of epoxy molding compounds are Rohm & Haas and the Hysol Division of Dexter, both in the United States. Dow Corning (United States) is one of the most important suppliers of silicone molding compounds.

4. Plating Chemicals

4.1. Electrolytic Gold Plating Chemicals

Gold [*7440-57-5*] plating on printed circuit boards functions mainly as the surface metal for contact areas because of its high electrical conductivity and resistance to tarnish. Gold can also function as a metal etch resist because of its resistance to all common circuit board etchants. Electrolytic gold plating chemicals consist of gold salts, gold strikes, conductivity salts, and pH-adjusting salts.

Gold salts, also known as gold plating processes, are proprietary formulations of either potassium gold cyanide or alkaline noncyanide gold in either powder or solution form. In addition to supplying gold metal for plating, these formulations often include brighteners to produce a mirror finish or complexes of cobalt, nickel, and indium for increased hardness and wear resistance.

Gold strikes are used to deposit a thin layer of gold on other plated metals prior to the use of a noncyanide gold plating solution or an electroless gold plating process. The thin gold layer

improves the bonding of the noncyanide or electroless gold layers to the circuit board.

Conductivity salts allow gold plating baths to operate at a lower voltage, thus increasing bath efficiency; *pH-adjusting salts* are used to raise or lower the pH of the bath.

Production. All electrolytic gold baths and strikes are prepared by mixing the conductivity salts in water and adding either the gold salt or gold strike to the solution. Some pH-adjusting salts may be added initially to obtain the proper pH. Once the bath is operating, only gold salts (and occasionally some pH-adjusting salts) are added to replenish the bath. Gold baths are long lasting and are reformulated only once or twice per year. Acid gold baths operate at a temperature of 21 – 49 °C with a pH of 3.5 – 4.0, whereas alkaline noncyanide gold baths operate at 32 – 54 °C with a pH of 9.0 – 10.0 [7].

Types and Grades. Gold salts are available in three types and four grades (see Table 3), as defined by the U.S. Department of Defense specification MIL-G-45204 B Amendment 2 for electrodeposited gold plating [7, p. 28]. The type of gold salt refers to the purity of the gold deposit, whereas the grade refers to the hardness of the deposit as measured by the Knoop method.

Table 3. Types and grades of gold salts

Type	Gold purity (min.), wt %	Grade	Knoop hardness
I	99.7	A	0 – 90
II	99.0	B	91 – 129
III	99.9	C	130 – 200
		D	≧ 201

Gold concentration in the salts varies with vendors, but most salts have concentrations of ca. 67 %.

Uses. Electrolytic gold plating is used on printed circuit boards for end uses such as computers, communications, and the military, all of which require a high degree of reliability. The thickness of plated gold varies among printed circuit board manufacturers. Western Electric has developed a specification requiring a plated gold thickness of 1.27 µm. Specifications for other communication and computer companies range from 0.762 to 2.03 µm and military requirements can typically vary from 2.54 to 3.81 µm.

Economic Aspects. An estimated 5.3 kg of gold metal and formulation chemicals was consumed in the United States for electrolytic gold plating on printed circuit boards. The major suppliers of gold plating processes are Oxy Metal Industries, Engelhard, and Lea Ronal, all in the United States.

4.2. Electroless Copper Plating Chemicals

Electroless plating is used to deposit thin metal layers on nonmetallic surfaces without the use of an electric current. Most electroless plating is used for depositing copper [*7440-50-8*], which functions to metallize nonconductive surfaces on printed circuit boards. Electroless copper plating systems consist of a solution, a catalyst, and an accelerator.

Electroless copper solutions are usually a proprietary two-component system. One component contains copper(II) salts and formaldehyde, whereas the other contains sodium hydroxide. When these solutions are combined in water, copper is reduced as follows [7]:

$$2\,Cu^{2+} + HCHO + 3\,OH^- \xrightarrow{Pd} 2\,Cu^+ + HCOO^- + 2\,H_2O$$
$$2\,Cu^+ \rightarrow Cu^0 + Cu^{2+}$$

Electroless copper catalysts are proprietary formulations containing both palladium chloride and tin(II) chloride. The palladium is essential for initiating the copper reduction reaction on a nonconductive surface. Once copper is deposited, the reaction is autocatalytic.

Electroless copper activators are proprietary formulations that increase the ability of the palladium catalyst to reduce copper more rapidly and uniformly. The activator also improves adhesion of the electroless copper to the base metal.

Production. All electroless or immersion plating is conducted by submerging the circuit board into a prepared bath. Most electroless copper baths operate at a temperature of 21 – 27 °C with a pH of ca. 12 [7, p. 6]. The difference between electroless solutions and immersion solu-

tions is subtle but important. Electroless solutions contain a reducing agent that supplies electrons to reduce metal ions to the metallic state. Theoretically, an electroless bath can proceed indefinitely. Immersion plating is accomplished by displacing one metal with another of lower oxidation potential. As a result, plating ceases when all the metal on a given substrate has been displaced.

Types and Grades. All electroless copper plating chemicals are sold as a concentrate that requires dilution with water before it is ready for use. Many suppliers have different dilution rates for the various electroless chemicals because of the wide range of concentrations available.

Uses. Electroless copper plating is used to deposit metal on the nonconductive walls of holes that are drilled in double-sided and multilayer boards. The metallized holes allow the printed circuitry to connect both sides of a board. Other electroless plating systems are employed only occasionally for specific functions. Electroless nickel solutions and immersion tin or tin – lead solutions can be used to inhibit corrosion and promote solderability on a completed circuit board. Electroless or immersion gold solutions are used when electroplating is impractical because of a lack of electrical continuity on the edge connectors.

Economic Aspects. In 1980, ca. 6.1×10^6 L of electroless copper plating chemicals was used in the United States; half that amount was consumed in Western Europe. MacDermid and Shipley both in the United States, are the two major suppliers of these plating chemicals.

4.3. Electrolytic Copper Plating Chemicals

Electrolytic copper plating chemicals are used to deposit copper onto printed circuit boards. Copper is widely accepted as a base conductor metal in circuit board fabrication because of its excellent electrical conductivity, low cost, high reliability, and good bonding with other metals. The two most commonly used copper processes are acid sulfate and pyrophosphate. The chemicals used in the *acid sulfate process* are copper sulfate, sulfuric acid, copper anodes, and proprietary brighteners, whereas the *pyrophosphate process* uses copper pyrophosphate, potassium pyrophosphate, copper anodes, and proprietary additives.

Copper(II) sulfate [*7758-98-7*], $CuSO_4$, is the source for copper ions, which are plated onto the board in an acidic copper bath. A chemically pure grade copper sulfate is usually recommended for plating.

The function of *sulfuric acid* [*7664-93-9*], H_2SO_4, is to increase the conductivity of the bath solution and to corrode the copper anodes. *Copper* serves as the anode for the electrical current and supplies copper to form copper sulfate as corrosion takes place in acid baths. Acid copper baths require phosphorized anodes, whereas oxygen-free high conductivity or rolled phosphorized anodes are used in pyrophosphate baths.

Copper and potassium pyrophosphate [*7320-34-5*], $K_4P_2O_7$, are combined to form potassium copper pyrophosphate, which serves as the source of copper ions in a pyrophosphate bath. *Proprietary brighteners* are added to the bath to increase grain refinement, hardness, distribution, and brightness of the plated copper. These additives are used to a much greater extent with acid copper systems than with pyrophosphate systems.

Types and Grades. Copper sulfate is available in both technical and chemically pure grades for use in electroplating and can be purchased as a commodity or as a trade name in either liquid or crystalline form. Copper sulfate is also sold in a dilute form mixed with sulfuric acid and water for a ready-to-use bath solution. Copper anodes are sold either as chips or rods. Pyrophosphates are available as both solutions and crystals and are usually bought as a trade name product. Proprietary brighteners are sold as ready-to-use liquid concentrates.

Uses. Both the acid copper and pyrophosphate processes are used to deposit copper on the printed circuit board to form the circuitry. The thickness of the plated copper varies among circuit board manufacturers, but the typical range is 2.54 – 5.08 μm, regardless of the plating process. Acid copper operates at a temperature of 18 – 32 °C, whereas pyrophosphate

operates at 43 – 60 °C and requires a pH of 8.2 – 8.8. Typical bath compositions for acid copper and copper pyrophosphate are as follows [7, pp. 14, 16]:

Acid copper bath	
Copper(II) sulfate	25 vol %
Sulfuric acid	10 vol %
Chloride ion	50 ppm
Brightener	0.6 vol %
Deionized water	74.4 vol %
Copper pyrophosphate bath	
Copper(II) pyrophosphate	33.3 vol %
Potassium pyrophosphate	3.5 vol %
Ammonium hydroxide	0.5 vol %
Brightener	0.09 vol %
Deionized water	62.6 vol %

Economic Aspects. In the United States, ca. 4.8×10^3 t of copper anodes and 5.7×10^3 m^3 of copper plating chemicals were consumed for electrolytic copper plating in 1980. The major suppliers of the plating chemicals are Sel-Rex and Lea Ronal in the United States. Copper anodes are available from many suppliers, with such companies as AMAX, ASARCO, and Univertical among the more important.

4.4. Electrolytic Tin – Lead Plating Chemicals

Tin – lead or solder is electroplated onto the printed circuit board by using electrolytic tin – lead plating chemicals. The plated solder serves two functions: as a metal etch resist for the copper circuitry and as a solderable substrate for subsequent soldering of components to the board. Tin – lead plating baths consist of anodes, lead and tin tetrafluoroborates, fluoroboric acid, boric acid, and brighteners. Fluoroboric and boric acid are discussed further in Chapter 6.

Lead (II) tetrafluoroborate [*13814-96-5*], $Pb(BF_4)_2$, and *tin(II) tetrafluoroborate* [*13814-97-6*], $Sn(BF_4)_2$, are the sources of metal for deposition onto the board. Variations in the metal content of the bath can affect the content of the deposition alloy.

Fluoroboric acid [*16872-11-0*], HBF_4, increases bath conductivity and corrodes the tin – lead anodes. *Tin – lead* (Sn [*7440-31-5*], Pb [*7439-92-1*]) serves as the anode for the electrical current and supplies tin and lead to form $Sn(BF_4)_2$ and $Pb(BF_4)_2$ as corrosion takes place.

Boric acid [*10043-35-3*], H_3BO_3, is added to maintain bath stability by inhibiting the decomposition of tetrafluoroborate. Two types of *brighteners* are available: peptones and nonpeptone proprietary brighteners promote smooth, fine-grain deposits.

Types and Grades. Lead and tin tetrafluoroborates are available as liquid concentrates with 50 % active material content, usually of a chemically pure grade. Tin – lead anodes are sold in both rod and nugget form with a minimal purity of 99.9 %. Brighteners are sold as ready-to-use liquid concentrates.

Uses. Solder plating is typically used on double-sided boards and outer layers of multilayer boards. Solder is plated on top of the electroplated copper to serve mainly as an etch resist for the copper. The thickness of most tin – lead deposits is 7.62 – 12.7 µm. Tin – lead baths usually operate at a temperature of 21 – 27 °C [7, p. 21].

Economic Aspects. In 1980, ca. 1.4×10^3 m^3 of tetrafluoroborates, acids, and brighteners was consumed in the United States for electrolytic tin – lead plating in electronics, along with 1.1×10^3 t of tin – lead anodes. The more important suppliers of the plating chemicals are General Chemical (formerly Allied), Sel-Rex, Lea Ronal, and Harshaw. By far, the most important supplier of tin – lead anodes is Alpha Metals.

4.5. Other Electrolytic Plating Chemicals

Several other electrolytic plating processes are employed by the printed circuit board industry. Among the most important processes are nickel, tin, and rhodium plating.

Nickel electroplating uses such products as nickel sulfamate or nickel sulfate, boric acid, nickel bromide, nickel anodes, and proprietary additives. *Nickel sulfamate* [*13770-89-3*], $Ni(SO_3NH_2)_2$, is the primary source of nickel metal ions in a nickel sulfamate bath, whereas

nickel sulfate [*15244-37-8*], $NiSO_4$, serves the same function in a modified watts nickel bath. *Boric acid* functions as a buffer and reduces burning at high current densities. *Nickel bromide* [*13462-88-9*], $NiBr_2$, is used to minimize internal stress and cause corrosion of the nickel anodes. *Nickel* [*7440-02-0*] serves as the anode for the electric current and provides additional nickel ions as corrosion takes place. *Proprietary additives* are used to reduce stress and impart a semibright appearance to the deposit.

Tin electroplating requires the use of tin sulfate, sulfuric acid, tin anodes, and proprietary brighteners. *Tin(II)* sulfate [*7488-55-3*], $SnSO_4$, acts as the source of tin ions. *Sulfuric acid* increases bath conductivity, solubilizes the tin salts, and corrodes the tin anodes. *Tin* [*7440-31-5*] supplies additional tin ions through corrosion and functions as the anode for the electric current. *Proprietary brighteners* serve to brighten and level the distribution of the deposit.

Rhodium electroplating uses proprietary rhodium salt solutions, sulfuric acid, and platinized titanium anodes. The *proprietary rhodium solutions* supply the metallic rhodium as an ion and contain additives that reduce the tensile stress of the deposit. *Sulfuric acid* increases the conductivity of the bath and the *platinized titanium* serves as the anode for the electric current.

Types and Grades. All anodes are available as either rods or nuggets and have a minimal purity of 99.9 %. The proprietary additives for nickel, tin, and rhodium are all supplied as ready-to-use liquid concentrates. The remaining commodity chemicals are usually sold in liquid form as either a technical or chemically pure grade.

Uses. Nickel plating is used on printed circuit boards as an undercoat for both precious and nonprecious metals. Its main function is surfaces, such as switch and edge contacts, that normally receive heavy wear. The nickel is plated underneath a gold or rhodium plate and greatly increases wear resistance. There are two basic nickel plating baths, the nickel sulfamate bath and the modified watts nickel bath, which differ primarily on the source of nickel ions. Typical compositions for both baths are as follows:

Nickel sulfamate bath	
Nickel sulfamate	28 vol %
Boric acid	4 vol %
Nickel bromide	1 vol %
Proprietary additive	as needed
Deionized water	67 vol %
Modified watts nickel bath	
Nickel sulfate	23 vol %
Boric acid	4 vol %
Nickel bromide	1 vol %
Proprietary additive	as needed
Deionized water	72 vol %

Tin plating on circuit boards has limited application, mainly to promote solderability. The increase in the cost of precious metal plating has caused some renewed interest in tin plating as a metal etch resist.

The use of rhodium plating on circuit boards is limited to edge connectors where a low-resistance, long-wearing, oxide-free contact is required. In this type of application, rhodium replaces gold as the plated metal [7, pp. 23 – 24].

Economic Aspects. The United States printed circuit board industry consumed 45×10^3 kg of nickel anodes and 230×10^3 L of nickel plating chemicals in 1980 for electrolytic nickel plating. An additional 190×10^3 L of tin and rhodium plating chemicals was also consumed that year. The largest nickel anode supplier is International Nickel, and the two largest suppliers of nickel plating chemicals are Sel-Rex and Lea Ronal.

5. Photoresists (→ Imaging Technology)

5.1. Dry Film Photoresists

Dry film photoresists all work negatively and consist of an acid-based photopolymer layer sandwiched between a layer of polyester film on one side and either a polyolefin or a polyethylene sheet on the other side. Both cover sheets protect the photopolymer from damage by abrasion. The polyolefin or polyethylene sheet is removed as the resist is laminated to the printed circuit board. An ultraviolet source is used to expose the resist; the polyester sheet is then removed, and the resist is developed.

Ultraviolet light causes the photopolymer to polymerize and become insoluble in the developing solutions. After the dry film is developed, the remaining material acts as an etch resist with print- and etch-type circuit boards and as a plating resist with print-, plate-, and etch-type circuit boards. Once the dry film is no longer needed, it is stripped from the board by a methylene chloride solution for solvent type film, or it is stripped by a highly alkaline solution (pH 13) at ca. 49 °C for aqueous and semiaqueous type films. This process is performed only once on a printed circuit board.

Types and Grades. Three types of dry film photoresists are employed commercially: (1) solvent, (2) semiaqueous, and (3) aqueous. The solvent types are developed by using 1,1,1-trichloroethane, and semiaqueous resists are developed by using proprietary water-based solutions. Aqueous dry films are developed by using solutions of sodium carbonate or trisodium phosphate.

Uses and Economic Aspects. Dry film photoresists are used exclusively by the printed circuit board industry and are favored for both multilayer and double-sided boards with dense circuitry.

In 1980, the United States printed circuit board industry consumed ca. 1.9×10^6 m^2 of dry film photoresists while Western Europe used 9.1×10^6 m^2. The two main suppliers of this product are Du Pont and Dynachem.

5.2. Liquid Positive and Negative Photoresists

Positive photoresists are based on the photochemical conversion of an ortho diazoketone to a carboxylic acid. On exposure to ultraviolet light, these resists become less hydrophobic and are more easily attacked by the alkaline developer. The remaining unexposed areas act as an etch resist for the manufacture of integrated circuits. One of the significant advantages of this type of photoresist is that instead of swelling the film and leaching out the solubilized materials, as with negative resists, the developer removes the exposed areas of the positive resist. As a result, extremely-high-resolution images are possible because no swelling induced pattern deformation occurs.

Negative photoresists are azide-based photopolymers that polymerize when exposed to ultraviolet light. The polymerized areas become insoluble in the organic developer solutions while the nonexposed areas are dissolved. The remaining material acts as either an etching resist or a plating resist, depending on the application.

Types and Grades. All liquid photoresists are currently exposed using ultraviolet light. They are generally supplied as 100 % active materials with standard kinematic viscosities of (2, 6, 18, 30, and 70)$\times10^{-6}$ m^2/s for positive photoresists, and for negative ones, (5, 30, 45, 50, 110, and 540)$\times10^{-6}$ m^2/s.

Uses. Positive and negative photoresists are applied to the *semiconductor substrate* as a liquid that is spun to form a uniform coating and subsequently baked to a dry film. Once the positive resist is developed, the remaining material acts as an etch resist to form the desired pattern. After the etching is complete, the resist is stripped from the semiconductor wafer. Once the negative photoresist is developed, it protects the area that it covers from being etched. After the etching process is complete, the developed photoresist is stripped from the semiconductor wafer. This process is typically repeated 8 – 9 times for metal oxide semiconductor (MOS) devices and 10 – 12 times for bipolar devices.

Liquid photoresists can also be used as *etching resists* for the manufacture of thin-film circuits. However, these circuits are not classified as semiconductor devices; therefore, the resists used for this application are not included in the economic data.

Liquid photoresists can be used as either a plating or an etching resist on the inner layer of multilayer *printed circuit boards.* The developed resist prevents the etching or plating of base and precious metals on the areas that it covers. Typically, these resists are applied only once to the circuit board.

Economic Aspects. Semiconductors are the major application for positive and negative photoresists, with the United States using 300×10^3

and 800×10^3 L, respectively, and Western Europe using 76×10^3 and 95×10^3 L in 1980. The three major suppliers of positive resists are Shipley, Olin-Hunt, and Hoechst, with Kodak and Tokyo Ohka also active. The two leading suppliers of negative resists are Olin-Hunt and Kodak, with E. Merck also active.

5.3. Screening Inks

Screening inks are based primarily on vinyl – polymer resins, although other resins are also available. These inks are screened onto a printed circuit board by forcing it through a stainless steel or polyester monofilament screen that has a specific pattern that determines where the ink is deposited. The inks act as either a plating resist or an etching resist. Because of this end-use application, the screening ink must be chemically resistant to the various cleaning, plating, and etching solutions used in actual processing operations, while maintaining viscosity and thixotropic properties suitable for screen printing. When the printing process is complete, the ink is cured by either thermal drying or ultraviolet light before being used as a resist.

Types and Grades. There are two basic types of screening inks, thermally curable and ultraviolet curable, both of which can be soluble in either alkali or solvent. The inks are supplied as 100 % active material and ready for immediate use.

Uses. Screening inks are used exclusively with *printed circuit boards* as either a plating resist or an etching resist. Alkali-soluble inks are used primarily on print- and etch-type circuit boards and can withstand acid-plating baths, but not alkaline cleaning steps. These inks are removed from the board by alkaline solutions such as sodium hydroxide. Solvent-soluble inks are more versatile because they can withstand all usual plating baths and cleaning steps. As a result, these inks are more likely to be used with print, plate, and etch boards. Solvent inks are removed from the board by the use of chlorinated solvents such as 1,1,1-trichloroethane and methylene chloride.

Economic Aspects. In 1980, the United States used a total of 590×10^3 L of both thermal and ultraviolet curing screening inks for printed circuit board production. At the same time, Western Europe consumed 380×10^3 L of screening inks. Hysol and Dynachem are suppliers in both regions, with Colonial, Nuz-Dor, London Chemical, and Grace active in the United States; Coates, Sericol, and Wiederhold are active in Europe.

6. Acids and Etchants

Acids and etchants are used extensively by manufacturers of semiconductors and printed circuit boards for operations such as cleaning and etching either wafers or copper-clad laminates. Some of the most important products and their applications are listed in Table 4.

6.1. Sulfuric Acid

Sulfuric acid (H_2SO_4) is a corrosive, oily liquid that is clear, cloudy, or fuming, depending on concentration (→Sulfuric Acid and Sulfur Trioxide). Sulfur, the principal raw material employed in the manufacture of sulfuric acid, can be obtained from various sources, including (1) elemental sulfur, (2) stack gas emissions, (3) sulfuric acid sludge, and (4) pyrites and hydrogen sulfide gas.

New sulfuric acid is commonly manufactured by the contact process. This procedure involves catalytic oxidation of sulfur dioxide (SO_2) to sulfur trioxide (SO_3), which is absorbed in water. Other manufacturing methods include the chamber process, which offers slower rates of production and a lower strength and quality acid, compared to that produced by the contact process.

Grades and Properties. Sulfuric acid is produced commercially in a variety of strengths for various industrial applications. Acids containing up to 93.3 wt % H_2SO_4 are measured in degrees Baume, while acid strengths of 93.4 – 100.0 wt % are measured in terms of percent H_2SO_4. Sulfuric acid purchased by manufacturers of *semiconductors* is usually a semiconductor grade product assayed at 96 – 97 wt %

Table 4. Functions of acids and etchants

Products	CAS registry number	Formula	Applications: Semiconductors	Applications: Printed circuit boards
Acids				
Acetic acid	[*64-19-7*]	CH_3COOH	wafer etching of metals	
Boric acid	[*10043-35-3*]	H_3BO_3		acid bath stabilization
Fluoroboric acid	[*16872-11-0*]	HBF_4		tin – lead cleaner
Hydrochloric acid	[*7647-01-0*]	HCl	wafer etching	metal cleaner and regeneration of copric chloride etchants
Hydrofluoric acid	[*7664-39-3*]	HF	wafer etching	
Nitric acid	[*7697-37-2*]	HNO_3	wafer cleaning and etching	laminate cleaning
Phosphoric acid	[*7664-38-2*]	H_3PO_4	wafer etching of metals	metal cleaner
Sulfuric acid	[*7664-93-9*]	H_2SO_4	wafer cleaning and resist stripping	cleaning and etching of laminate
Etchants				
Ammoniacal solutions		NH_4^+ salts		final copper etching
Ammonium fluoride	[*16099-75-5*]	NH_4F	wafer etching	
Buffered oxide etchants			wafer etching of MOS devices	
Copper(II) chloride	[*7447-39-4*]	$CuCl_2$		copper etching of inner layer boards
Iron(III) chloride	[*7705-08-0*]	$FeCl_3$		copper etching
Others				
Hydrogen peroxide	[*7722-84-1*]	H_2O_2	wafer etching and photoresist stripping	
Ammonium persulfate	[*7727-54-0*]	$(NH_4)_2S_2O_8$		copper foil etching to provide better adhesion of electroless copper
Sodium persulfate	[*7775-27-1*]	$Na_2S_2O_8$		same as above

acid. The SEMI (Semiconductor Equipment and Materials Institute) specifications for sulfuric acid are available [5].

In addition to semigrade sulfuric acid, extremely low particulate products are also available to semiconductor manufacturers. These materials meet or exceed specifications for semigrade sulfuric acid while being virtually free of suspended matter.

Manufacturers of *printed circuit boards* generally purchase technical grade sulfuric acid. Specifications for particulate content and purity of these products are not typically as high as specifications for acid used by the semiconductor industry.

Uses. During the manufacturing of *semiconductors,* the wafer is cleaned numerous times with sulfuric acid to remove microscopic dust and other contaminants from the substrates surface. For example, during the masking operation alone, wafers may be cleaned several times before progressing to the next manufacturing step. Other applications for sulfuric acid include mixture with hydrogen peroxide to form Caro's acid, which is a wafer cleaner and resist stripper.

Printed Circuit Boards. Sulfuric acid is used as a (1) cleaner, (2) etchant, and (3) constituent of electrolytic copper plating baths. As a cleaner, the acid removes foreign contaminants formed from contact of an etchant of other material with the laminate surface. For example, an ammonium persulfate etch in contact with copper can produce copper ammonium sulfate, which can form a film on the board's surface. This film is readily removed by dipping the board in a sulfuric acid bath prior to the next manufacturing step.

Other applications for this electronic chemical include acid – copper electrolytic plating, whereby sulfuric acid is added to a copper sulfate bath for anode corrosion to replenish the copper sulfate level of the solution while also increasing bath conductivity.

6.2. Hydrofluoric Acid

Hydrofluoric acid (HF) in the anhydrous form is a fuming, corrosive liquid that can cause serious

burns on skin contact. The acid is manufactured by the reaction of sulfuric acid and fluorspar (calcium fluoride) to produce hydrogen fluoride gas and calcium sulfate. After absorption in water, the acid is concentrated by distillation, forming anhydrous hydrogen fluoride. Aqueous grades are then made by dilution with water.

Grades and Properties. Hydrofluoric acid purchased by manufacturers of semiconductors is often a semigrade product assayed at 48.8 – 49.2 % HF. Specifications for the semiconductor grade chemical according to SEMI are available [5].

Low particulate HF is also available to manufacturers of semiconductors. This acid meets or exceeds the semiconductor grade specifications.

Uses. Hydrofluoric acid is employed primarily to etch conductive paths on the surface of the wafer. The acid is often mixed with ammonium fluoride by the customer to form a buffered oxide etch, particularly for MOS devices. Various concentrations of ammonium fluoride and hydrofluoric acid are used for this purpose; however, a solution containing six parts of ammonium fluoride to one part of hydrofluoric acid is preferred by most semiconductor fabricators.

6.3. Ammoniacal Solutions

Ammoniacal solutions are proprietary formulations that typically contain a blend of various ammonium-based compounds. For example, these proprietary solutions are generally composed of hydroxides, chlorides, carbonates, phosphates, and nitrates of ammonia. The exact proportion of the individual chemicals to form the finished ammoniacal solution is kept confidential by the specialty suppliers.

Grades and Properties. Ammoniacal solutions are specifically formulated to meet the demands for copper etching set forth by the printed circuit board industry. These products generally differ in formula mass, appearance, and other properties, depending on the proprietary product analyzed. However, most blends have a pH of 8 – 9 and a density of 12 g/cm^3.

Uses. Ammoniacal solutions are used only as final copper etchants by manufacturers of printed circuit boards. Early versions of alkaline etchants, such as the ammoniacal solutions, were used for batch etching; the etch rates decreased as the copper content in the etching solutions increased. However, newer products were developed by adding oxidizing agents to speed the etching rate and increase the copper capacity of the solution at a constant temperature. Although batch operations are still employed, continuous systems have become popular. A typical alkaline etching installation operates continuously and in a closed loop. System components include (1) a standard spray etcher, (2) metering pumps, (3) control modules, and (4) storage tanks.

As previously mentioned, various ammonium compounds are important in formulating ammoniacal solutions (Table 5). These ammonium compounds serve to hold the copper being etched in solution and to increase the etch rate. Other important components of these proprietary solutions include copper ions, which react with and dissolve the metal by oxidation, as indicated in Table 5. Generally the solutions formulated with these chemicals are compatible with metals such as gold, nickel, silver, tin, tin – nickel, and particularly tin – lead.

6.4. Nitric Acid

Nitric acid (HNO_3) is strong, reactive, corrosive, and colorless in pure (100 %) form. Pure nitric acid is unstable and decomposes to nitrogen dioxide, which gives concentrated solutions a yellow-brown color; thus, commercial products typically contain various proportions of water.

Commercial production of nitric acid involves oxidation of ammonia first to form nitric oxide, and then nitrogen dioxide, which is absorbed in water to give the acid. Concentrated nitric acid is manufactured primarily by dehydration with sulfuric acid. In this procedure sulfuric acid (98 – 99 %) is mixed with sulfur trioxide (SO_3) and 60 % nitric acid in a concentrator. The sulfuric acid absorbs the water, resulting in nitric acid of ≥ 95 % strength. The concentration of the remaining sulfuric acid ranges from 60 to 70 %.

Table 5. Compounds important in formulating ammoniacal solutions

Component	CAS registry number	Formula	Function
Ammonium hydroxide	[*1336-21-6*]	NH_4OH	holds copper in solution
Ammonium chloride	[*12125-02-9*]	NH_4Cl	increases etch rate, copper holding capacity, and solution stability
Ammonium carbonate	[*506-87-6*]	$(NH_4)_2CO_3$	acts as a buffer, retains clean solder holes
Ammonium phosphate	[*7783-28-0*]	$(NH_4)_2HPO_4$	retains clean solder
Ammonium nitrate	[*6484-52-2*]	NH_4NO_3	increases etch rate and retains clean solder
Copper ion		Cu^{2+}	reacts with and dissolves metallic copper by acting as an oxidizing agent
Sodium chlorate	[*7775-09-9*]	$NaClO_3$	acts as an oxidizing agent that reacts with and dissolves metallic copper

Grades and Properties. Nitric acid is manufactured in several concentrations. Acids containing up to 67.2 wt % nitric acid are described in degrees of Baume, whereas acids of 70 – 95 wt % (or greater) concentration are commonly referred to in terms of percent HNO_3. The semiconductor grade of nitric acid used by manufacturers of semiconductors contains ca. 70 wt % HNO_3; the SEMI specifications for this acid are available [5].

Grades of nitric acid with a low particulate content are also available to the semiconductor industry. Manufacturers of printed circuit boards often buy an electronic- or technical-grade of nitric acid. Specifications of particulate content for these products are generally inferior to those of acid purchased by the semiconductor industry.

Uses. Nitric acid is used to clean foreign contaminants from the wafer's surface during selected steps in the manufacture of *semiconductors.* In addition, this material is also blended with other acids, such as acetic and hydrofluoric acid, to etch away metals, such as tantalum, simply by reacting with them.

During the manufacture of *printed circuit boards,* the board itself is often cleaned. Nitric acid is frequently used for this operation.

6.5. Hydrogen Peroxide

Hydrogen peroxide (H_2O_2) is a nonviscous, clear liquid with a faint blue-green color. Hydrogen peroxide has a slightly sharp odor and is miscible in all proportions with water.

Several manufacturing processes exist to make hydrogen peroxide; however, the current route involves successive hydrogenation and oxidation of organic diketones. For example, ethyl anthraquinone is hydrogenated to ethyl anthrahydroquinone, which is then oxidized to regenerate the ethyl anthraquinone and hydrogen peroxide.

Grades and Properties. Commercially, aqueous solutions of hydrogen peroxide are available in concentrations of 27 – 70 wt %. Most of this material purchased by the semiconductor industry contains 30 – 32 wt % hydrogen peroxide. The SEMI specifications for hydrogen peroxide are available [5].

Grades of hydrogen peroxide with a low particulate content are also available to semiconductor manufacturers. Technical grades are often preferred by the producers of printed circuit boards.

Uses. Hydrogen peroxide is commonly used by *semiconductor* manufacturers to clean wafer surfaces of foreign contaminants or to remove a photoresist. Quite often sulfuric acid is mixed with hydrogen peroxide (4 : 1 ratio) to form Caro's acid [*7722-86-3*], peroxymonosulfuric acid, for maximum cleaning efficiency during these operations.

Hydrogen peroxide is often purchased individually by manufacturers of *printed circuit boards* to regenerate etching solutions of copper chloride. In this procedure, the laminate is dipped into a copper chloride bath. The hydrogen peroxide acts as an oxidizing agent to free chlorine ions, which etch copper from the laminate.

In addition, manufacturers of printed circuit boards add hydrogen peroxide to proprietary etchants based on sulfuric acid to etch copper before electroless copper is deposited.

6.6. Proprietary Peroxy and Sulfuric Etchants

Proprietary etchants are based on either hydrogen peroxide or sulfuric acid. The formulations contain specialty additives for stability.

Proprietary peroxy etchants are composed primarily of hydrogen peroxide. Special additives act as stabilizers to retard degradation of the base chemical. The concentration of these additives is generally 1 – 4 wt % of the total product.

Proprietary sulfuric etchants usually contain sulfuric acid. Specialty additives, primarily stabilizers, extend the life of the product during etching procedures.

Grades and Properties. Properties of the proprietary peroxy and sulfuric etchants depend primarily on the base chemicals. For example, properties of the peroxy etchants are directly related to hydrogen peroxide, the base component. The relative molecular mass of hydrogen peroxide is 34 and its relative density is ca. 1.5 times that of water. Sulfuric acid, the base chemical for the sulfuric etchants, has a substantially higher molecular mass of 98.1 and a density of ca. 1.83, depending on the concentration.

Uses. Before being used as an etchant by the printed circuit board industry, these proprietary products may be further compounded by the customer, depending on the product purchased. For example, a stabilized peroxy etchant requires the addition of water and sulfuric acid (technical or reagent grade) at the customer's facility. A typical formulation of the finalized peroxy product may contain 82 vol % water (deionized), 11 vol % sulfuric acid, and a 7 vol % proprietary peroxy etchant.

Conversely, a manufacturer of printed circuit boards buying stabilized sulfuric acid would add water and hydrogen peroxide to get the final formulation.

In addition to stabilized hydrogen peroxide or sulfuric acid, specialty additive formulations are also available. These products consist primarily of the same materials used to give stability to the preceding base chemicals; however, these products require addition of sulfuric acid and hydrogen peroxide by the customer.

Once the etchant is formulated by the manufacturer of printed circuit boards, it is then used primarily to etch copper before electroless copper is deposited. Because of the etching action of these products, electroless copper adheres more strongly to the copper foil on the laminate simply by its abrasive action. Sodium and ammonium persulfate etchants also compete with the peroxy products in this end use.

In addition to acting as an etch, proprietary peroxy etchants are used less extensively to remove unwanted copper from the board in a later production process, leaving the final circuitry untouched. Other etchants used more frequently in this application include ammoniacal solutions and iron(III) and copper(II) chlorides.

Once the customer formulates the proprietary etchant with sulfuric acid or hydrogen peroxide (or both), the copper is etched by formation of copper sulfate:

$$Cu + H_2O_2 + H_2SO_4 \rightarrow CuSO_4 + 2\,H_2O$$

Cooling the etching solution to room temperature allows a pure grade of copper sulfate to be extracted for reclamation of copper.

6.7. Ammonium Fluoride

Ammonium fluoride (NH_4F) is a white crystalline salt. It often has a strong ammonia odor because it tends to decompose to ammonia and the more stable ammonium hydrogen fluoride ($NH_4F \cdot HF$). Ammonium fluoride can be produced easily by mixing aqueous ammonia with ammonium hydrogen fluoride in equimolar proportions.

Grades and Properties. Ammonium fluoride employed by manufacturers of semiconductors contains 40 wt % NH_4F. The SEMI specifications for this material are available [5].

Uses. Ammonium fluoride is generally mixed with hydrofluoric acid by manufacturers of semiconductors to etch the surface of wafers, particularly those used to make MOS devices. These buffered oxide etches, made captively, often contain six parts of ammonium fluoride to one part of hydrofluoric acid.

6.8. Persulfates

Ammonium persulfate $(NH_4)_2S_2O_8$ and sodium persulfate $(Na_2S_2O_8)$ employed by manufacturers of printed circuit boards are white crystalline materials with strong oxidizing characteristics. These products are typically produced by electrolytic oxidation of the appropriate hydrogen sulfate salt.

Grades and Properties. Both ammonium and sodium persulfates are high-purity salts easily solubilized in water and other materials. The molecular mass of sodium persulfate is 238, whereas that of ammonium persulfate is somewhat less at 220. The amount of active oxygen differs between each product type, but generally is 6 – 7 wt %.

Uses. Persulfates are used by the *printed circuit board* industry mostly as a cleaner and etchant prior to the deposition of electroless copper. Because they are strong oxidizing agents, these products roughen or abrade the surface of the board, enabling electroless copper to adhere more strongly without attacking resists or plastics used as the circuitry base. Persulfates are applied in either continuous or batch methods, depending on the customer's operation.

Solutions based on persulfates have an etch rate of 0.76 – 2.5 nm/min, depending on the temperature and copper content of the solution, which can be $\geq$ 7.8 mg/L. The persulfate concentration can also affect the etch performance of the finished persulfate solution.

Ammonium persulfate can be used to etch other metals besides copper, including materials such as aluminum, beryllium, and brass. The ratio of the compounds in the mixture depends on the customer's operation. A 20 % solution is made simply by adding 0.2 kg of ammonium persulfate per liter of water in a sump or dissolving tank.

Solutions are often catalyzed with materials such as mercuric chloride to achieve optimal etch rates or phosphoric acid to eliminate incomplete etching of copper next to solder-plated conductors.

Sodium persulfate solutions are also prepared according to the customer's preference. For example, solutions often contain 0.15 kg of $Na_2S_2O_8$ per liter of warm water. Small quantities of phosphoric or sulfuric acid can be added on initial makeup for more aggressive cleaning solutions.

6.9. Other Acids and Etchants

A variety of other acids and etchants are employed by manufacturers of semiconductors and printed circuit boards in their production processes. Among the more important are acetic, phosphoric, and hydrochloric acids, whereas other material, such as formulated buffered oxide etches (BOE) and iron(II) chloride, are also significant.

Acetic Acid (→ Acetic Acid). The three significant processes employed to produce acetic acid are (1) methanol carboxylation, (2) butane oxidation, and (3) acetaldehyde oxidation. Acetic acid purchased by manufacturers of semiconductors is usually a semigrade product assayed at 99.7 wt % CH_3COOH. The SEMI specifications for acetic acid as used in phosphoric etchants are available [5].

Acetic acid is used by manufacturers of semiconductors only where it is mixed with nitric and phosphoric acids to serve as a metal etchant. Semigrade acetic acid sells for ca. \$ 2.20/kg (\$ 1/lb.), and it is usually shipped in 2.3-kg (5-lb.) bottles, with six to a case.

Phosphoric Acid. (→ Phosphoric Acid and Phosphates) Phosphoric acid is a corrosive, syrupy liquid that can cause burns on contact with the skin. It can be made by reacting elemental phosphorus with oxygen to form phosphorus oxides, which are then dissolved in water to form the acid. An alternate process involves the reaction of acids such as sulfuric or nitric acid with phosphate rock to dissolve the calcium phosphate. Phosphoric acid purchased by manufacturers of semiconductors is usually a semigrade product assayed at 85 – 87 wt % H_3PO_4. The SEMI specifications for this chemical are available [5].

Technical grades of phosphoric acid have lower purity and a higher particulate content than the semigrade products and as a result are used less frequently.

Manufacturers of semiconductors mix phosphoric acid with nitric and acetic acids and use the mixture as a metal etchant, whereas the printed circuit board industry employs the acid by itself as a metal cleaner. Semigrade phosphoric acid sells for ca. $ 1.50/kg ($ 0.68/lb.), and it is usually packaged in 6.35-kg (14-lb.) bottles, with four to a case. Technical grade acid is somewhat less expensive at $ 0.46/kg ($ 0.21/lb).

Hydrochloric Acid. (→ Hydrochloric Acid) Hydrochloric acid is a strong, highly corrosive acid, primarily produced as a byproduct from the manufacture of organochlorine chemicals. The semiconductor industry buys a semigrade product assayed to contain a minimum of 37 wt % acid. The SEMI specifications are available [5].

In addition to semigrade acid, products with an extremely low particulate content are also available to manufacturers of semiconductors. Technical- or reagent-grade hydrochloric acid is commonly employed by the printed circuit board industry.

Hydrochloric acid is used primarily as a wafer etchant by semiconductor manufacturers. The printed circuit board industry employs HCl as a metal cleaner and as a source of chlorine ions for copper(II) chloride etchants. Semigrade acid sells for ca. $ 1.28/kg ($ 0.58/lb.) in 4.54-kg (10-lb.) containers, with four to a case, whereas technical-grade acid is considerably less expensive at $ 0.18/kg ($ 0.08/lb).

Fluoroboric Acid. (→ Boron Compounds, Chap. 2.) Fluoroboric acid with a density of ca. 1.38 g/cm^3 has properties similar to those of hydrofluoric acid, including a corrosive effect on glass. This acid is used primarily by the printed circuit board industry as a tin – lead cleaner, and the cost is ca. $ 1.32/kg ($ 0.60/lb).

Mixed Acid Etchants. Mixed acids are specially blended products usually containing nitric, hydrofluoric, and acetic acids. Properties of these products depend on the proprietary proportions of each component. A mixture of acids is used by the semiconductor industry as a metal etchant; cost ranges from $ 1.52 to $ 2.29/kg ($ 0.69 to $ 1.04/lb). The SEMI specifications for mixed acid etchants are available [5].

Chromic – Sulfuric Acid Etchants. Chromic – sulfuric etchants are specially blended products containing chromic and sulfuric acids with the exact proportion of each component determined by the formulator. These products are used sparingly by printed circuit board manufacturers to remove epoxy smears in the etch-back process and as a final copper etchant. Chromic – sulfuric blends usually cost ca. $ 0.53/L ($ 2/gal).

Boric Acid. (→ Boric Oxide, Boric Acid, and Borates, Chap. 3.) Boric acid is used by printed circuit board manufacturers for acid bath stabilization. The cost of this product is ca. $ 0.14 kg ($ 0.30/lb).

Buffered Oxide Etchants. Buffered oxide etchants (BOE) are specially blended products often containing six parts of ammonium fluoride to one part of hydrofluoric acid. These products are used primarily as etchants, particularly for MOS devices, by semiconductor manufacturers. Formulated BOE etchants purchased by this industry cost ca. $ 1.83/kg ($ 0.83/lb.) in 1980. The SEMI specifications for this product are available [5].

Copper(II) Chloride. (→ Copper Compounds, Chap. 4.) Copper(II) chloride etching solutions contain components such as hydrochloric acid, sodium chloride, ammonium chloride, water, and copper(II) chloride crystals. Specific formulations of these commodity chemicals depend on performance results experienced by the individual manufacturer of printed circuit boards during final etching operations. For example, chloride ions can be provided by several sources, including hydrochloric acid, sodium chloride, or ammonium chloride, to solubilize the relatively insoluble copper(I) chloride formed by the etching action of the copper (II) chloride solution on copper. The use of such chloride ions assists in maintaining a stable etch rate. In addition, copper(I) chloride can be oxidized to copper(II) chloride by electrolytic processes or by direct chlorination, as follows:

$$Cu_2Cl_2 + Cl_2 \longrightarrow 2\,CuCl_2$$

Copper(II) chloride solutions are used mostly to etch inner layers of multilayer boards. These

solutions are also used to process print and etch boards. Resists, such as screening inks, dry film, and tin – nickel, are compatible with this etchant, whereas other materials, such as tin – lead and tin boards, are unsuitable.

Table 6. Consumption of acids and etchants in electronics in the United States, 1980

Product	Consumption, 10^3 t		
	Semiconductors	Printed circuit boards	Total
Acids			
Sulfuric	16.8	27.2	44.0
Nitric	7.7	2.0	9.7
Hydrochloric	1.4	7.7	9.1
Hydrofluoric	8.2		8.2
Proprietary		3.1	3.1
Phosphoric	1.7	1.3	3.0
Acetic	2.3		2.3
Fluoroboric		2.1	2.1
Chromic – sulfuric		1.0	1.0
Mixed	0.5		0.5
Boric		0.4	0.4
Subtotal	38.6	44.8	83.4
Etchants			
Ammoniacal solutions		48.1	48.1
Copper(II) chloride		12.2	12.2
Iron(III) chloride		8.1	8.1
Hydrogen peroxide	5.4	1.4	6.8
Persulfates		5.3	5.3
Ammonium fluoride	3.7		3.7
Buffered oxide etchants	2.0		2.0
Subtotal	11.1	75.1	86.2
Total	49.7	119.9	169.6

Iron(III) Chloride (→ Iron Compounds). Iron(III) chloride etchants consist primarily of $FeCl_3$ in water with concentrations ranging from 28 to 42 wt %; wetting and antifoam agents also are typically added. These solutions are widely used in the printed circuit board industry as final etchants for copper, copper alloys, and steel. In addition the iron(III) chloride etchant is compatible with screening inks and photoresists, but not with solder or tin-plated boards.

Solutions of iron(III) chloride formulated with additives by specialty suppliers provide fast and even etching, partly because of the addition of strong oxidizers and surface wetting agents. Over the years, suppliers of these products have greatly extended the useful life and uniformity of etch rates provided by these products. Iron(III) chloride etching solutions sold for ca. \$ 0.18/L (\$ 0.70/gal) or ca. \$ 0.13/kg (\$ 0.06/lb.) in 1980.

Economic Aspects. Total demand for acids and etchants by the electronics industry in the United States was 169.6×10^3 t in 1980. The printed circuit board industry used 71 % of the total (119.9×10^3 t), and the semiconductor industry consumed the balance. A breakdown of consumption by product is given in Table 6.

The leading suppliers of semiconductor-grade acids and etchants in the United States are General Chemical, Ashland Chemical, and Mallinckrodt. Technical-grade chemicals are provided by numerous bulk chemical manufacturers. Ammoniacal solutions are sold mainly by Olin-Hunt, MacDermid, and Southern California Chemical. The three major suppliers of proprietary acid mixtures are Electrochemicals, Shipley, and MacDermid. The major supplier of persulfates in the United States is FMC.

7. Atmospheric Gases

Nitrogen [*7727-37-9*], N_2, *bp* −196 °C, is a chemically inert, odorless, colorless, tasteless, diatomic gas (→ Nitrogen). It makes up 80 vol % of the air we breathe. Nitrogen reacts with other elements only at high temperature and pressure.

Nitrogen is produced in air-separation plants by liquefaction and subsequent fractionation of air into its major components, i.e., nitrogen, oxygen, and small amounts of argon. In some plants small quantities of neon, krypton, and xenon may also be recovered, whereas other gases in the air, such as carbon dioxide, helium, and hydrogen, are usually vented. Generally, the most economical production ratio is 2.4 parts of nitrogen to 1 part of oxygen, with excess N_2 used to clean alternate compressors within the plant.

Grades and Properties. Nitrogen has a molecular mass of 28.01 and a specific volume of 0.862 m^3/kg. Typically, gas sold to the semiconductor industry exceeds 99 % in purity. The SEMI specifications for liquid nitrogen used in bulk by the semiconductor industry are available [5].

Uses. Nitrogen is essential in many production processes and is by far the most widely used atmospheric gas among manufacturers of semiconductors in the United States. For example, nitrogen is commonly found in furnaces, epi (epitaxial) reactors, and ion implanters as an inert atmosphere to protect against impurities such as

oxygen. Additional uses for this versatile gas include serving as the transportation medium in air tracts and the driving force in air guns. Nitrogen is supplied to customers by several means, including (1) pipelines serving either single or multiple facilities, (2) air-separation plants located on site, and (3) merchant bulk deliveries.

Hydrogen [*1333-74-0*], H_2, is a colorless, odorless, and tasteless gas that is lighter than air and practically insoluble in water (→ Hydrogen). Commercial processes for hydrogen production include electrolysis, ammonia dissociation, propane re-forming, partial oxidation of hydrocarbons, and steam re-forming of natural gas.

Types and Grades. Hydrogen purchased by manufacturers of semiconductors generally has a minimal purity of 99.9 % The SEMI specifications for impurity levels in bulk hydrogen are available [5].

Uses. Hydrogen is used by the semiconductor industry to create a reducing atmosphere, particularly in manufacturing bipolar devices. For example, with metals such as aluminum, any oxide present can be reduced to seal the metallic surface before becoming a component of the integrated circuit. An additional, but less common, application for hydrogen involves mixture of gas with carbon tetrafluoride for reactive ion etching (RIE).

Argon [*7440-37-1*], Ar, *bp* −186 °C, *fp* −198.4 °C is a colorless and odorless gas (→ Noble Gases). Virtually all commercial production of argon is the result of oxygenic separation of atmospheric gases.

Types and Grades. Argon purchased by manufacturers of semiconductors is almost 100 % pure. The SEMI specifications for the allowable impurity levels are available [5].

Uses. Argon is used primarily by the semiconductor industry as an inert atmosphere for growing single silicon cystals (→ Crystal Growth; → Silicon). Argon is also used as an atmosphere during ion-implant procedures and during annealing, a process to repair substrate damage after ion implantation takes place.

Oxygen [*7782-44-7*], O_2, is a colorless gas that forms oxides with all elements except some of the noble gases (→ Oxygen). Almost all commercial oxygen is produced by fractional distillation of air.

Types and Grades. Oxygen purchased by manufacturers of semiconductors is high in quality with a minimal purity of 99.6 %.

Uses. Oxygen is used by manufacturers of semiconductors to produce passivation layers of silicon dioxide on silicon substrates, particularly for MOS devices. Additional uses of this gas include plasma stripping of photoresists.

Economic Aspects. Consumption of the four atmospheric gases discussed in this chapter is shown in Table 7.

Table 7. Consumption of atmospheric gases by the semiconductor industry, 1980

Gas	Consumption, 10^6 m^3	
	United States	Western Europe
Nitrogen	1076.0	138.8
Hydrogen	31.2	6.4
Argon	7.1	2.0
Oxygen	21.5	2.6

These gases are typically sold by the same companies. Several major suppliers are active in both the United States and Western Europe, including Air Products, BOC/Airco, and L'Air Liquide. The Linde Division of Union Carbide is an important supplier of atmospheric gases in the United States, and Messer Griesheim is a big supplier in the Federal Republic of Germany.

8. Photoresist Adjuncts

8.1. Photoresist Strippers

Strippers are chemicals used to remove photoresist after the etching or plating process is complete. Unlike developers, which remove solubilized photoresist, strippers remove the hardened unsolubilized photoresist without damaging the underlying material.

Types and Grades. The three types of strippers for *semiconductor photoresists* are (1) oxygen in a plasma state, (2) phenolic-based formulations, and (3) nonphenolic formulations, which are based on either sulfonic acid or chromic acid. All of these products are sold as electronic-grade material. Strippers for *printed*

circuit board photoresists include commodity chemicals, such as methylene chloride and sodium hydroxide, as well as proprietary formulations based on these chemicals. These printed circuit board strippers are available as technical-grade materials.

Uses. *Semiconductor* photoresist stripping with oxygen is accomplished by placing the wafers in a vacuum chamber, where the gas passes through a radio frequency electrode to form a highly ionized gas or plasma. The plasma contacts the photoresist, causing a chemical reaction that oxidizes the resist to ash; the resist ash is then rinsed from the water surface. On the other hand, chemical resist stripping involves immersion of the wafer in a stripping bath for 30 – 60 s.

All *printed circuit board* resists are stripped by immersion techniques. Solvent-based strippers usually operate at room temperature and require 1 – 2 min for complete photoresist removal, while alkaline-based strippers operate at 49 – 60 °C and require from 30 s to 3 min for complete photoresist removal.

Economic Aspects. The total consumption of strippers for electronics in the United States was 12.8×10^6 L in 1980. Of this total, 10.9×10^6 L was used by the printed circuit board industry, as indicated in Table 8.

Table 8. Consumption of strippers by the electronics industry

Product	Consumption, 10^6 L		
	Printed circuit boards	Semiconductors	Total
Solvent	8.3		8.3
Alkaline	2.6		2.6
Phenolic		1.5	1.5
Nonphenolic		0.4	0.4
Total	10.9	1.9	12.8

Many companies supply photoresist strippers for printed circuit boards, but the three largest suppliers are Chemical Solvent, J & S Laboratories, and Inland Specialty Chemical. For semiconductor photoresist strippers, Industri-Chem is the major supplier of phenolic-based products, General Chemical is an important supplier of both phenolic and nonphenolic products, and EKC Technology specializes in nonphenolic products.

8.2. Dry Film Photoresist Developers

Dry film photoresist developers are either solvent- or alkaline-based products that remove the soluble, nonexposed areas of the resist prior to the plating process.

Types and Grades. The two types of dry film developers are commodity solvents and proprietary formulations. Usually technical-grade 1,1,1-trichloroethane is the solvent used to develop solvent-processable dry films. Proprietary developers are used with aqueous- and semiaqueous-processable dry films.

Uses. Dry film photoresists are developed by immersing the printed circuit board in a solution of aqueous developer. The time required for optimum development varies with the thickness of the resist and the temperature of the bath, but normally 30 – 60 s is needed.

Economic Aspects. In 1980, the United States printed circuit board industry used a total of 14.8×10^6 L of dry film photoresist developers. Approximately 11.4×10^6 L consisted of 1,1,1-trichloroethane, and the balance was proprietary formulations. Dow and PPG are the major suppliers of 1,1,1-trichloroethane, whereas Du Pont and Dynachem are the major suppliers of proprietary developers.

8.3. Positive Photoresist Developers

Positive photoresist developers are alkaline-based solutions that remove the soluble exposed areas of the resist by an etching process. Development of the resist is accomplished by either a spray or immersion technique.

Types and Grades. The two types of positive photoresist developers are commodity chemicals, such as sodium and potassium hydroxide, and proprietary alkaline-based formulations. Electronic grades of sodium and potassium hydroxide are normally used, while proprietary developers are available only in an electronic grade.

Uses. *Semiconductor* wafers can be either immersed in a developer solution or sprayed with the solution. Immersion is favored with positive photoresist developers because line resolution is usually better with this technique. The developing time required for immersion is ca. 1 min. Spray techniques offer faster speed than that obtained with the immersion method, with a typical development time of 15 s.

Positive photoresists used on *printed circuit boards* are usually developed by immersion techniques.

Economic Aspects. In 1980, a total of 704×10^3 L of positive photoresist developers was consumed in the United States. Of the total, 643×10^3 L consisted of proprietary developers and the rest were either sodium or potassium hydroxide. Proprietary developers are sold by the photoresist supplier, Shipley, Hoechst, Olin-Hunt, and Kodak; firms such as Ashland Chemical, J. T. Baker, and General Chemical supply semiconductor grade hydroxides.

8.4. Negative Photoresist Developers

Negative photoresist developers are either solvent or solvent-based products that swell the photoresist film after the imaging process. The developer removes the soluble, unexposed area of the resist by a leaching process. The resist is developed by either a spray or immersion technique. A solvent rinse is used after the film is developed to remove any residual developer and to reduce swelling of the remaining photoresist.

Types and Grades. The two types of negative photoresist developers are commodity solvents, such as xylene, 1,1,1-trichloroethane, trichloroethylene (TCE), and stoddard solvent proprietary formulations. Both electronic-grade and reagent-grade xylene is used by the electronic industry, while only electronic-grade proprietary formulations are available. The proprietary products are based on xylene, with stabilizers added. The printed circuit board industry usually uses technical-grade commodity solvents. The SEMI specifications for xylene are available [5].

Uses. Usually electronic-grade xylene or proprietary formulations are used to develop photoresists on *semiconductor wafers,* although some reagent-grade xylene is also used. Semiconductor wafers can either be immersed in a developer solution or have the solution applied by spraying. Spray systems offer the benefit of speed and provide a continuous flow of clean solvent, while batch immersion tends to give finer resolution. The developing time with spray systems ranges from 15 to 30 s compared with 60 – 90 s for batch immersion.

Printed circuit boards using a Kodak or Hunt negative resist can be developed either in a vapor degreasing chamber using trichloroethylene or by immersion in a proprietary developer. Dynachem resists are attacked by trichloroethylene and must be developed by immersion in either 1,1,1-trichloroethane or a proprietary solution. Development time with a vapor degreaser ranges from 3 to 4 min, while immersion in a formulated developer usually requires 20 – 30 s. A vapor degreaser can develop many boards simultaneously, but immersion is usually done one board at a time.

Economic Aspects. The use of negative photoresist developers in the United States was 1.89×10^6 L in 1980, the vast majority being used in semiconductor applications. Xylene was the preferred product, representing 90 % of all consumption, with proprietary formulations accounting for most of the balance. Firms such as Ashland Chemical, J. T. Baker, and General Chemical supply semiconductor-grade xylene. Photoresist manufacturers, including Olin Hunt, Kodak, and E. Merck, supply the proprietary formulations.

8.5. Photoresist Adhesion Promoters

1,1,1,3,3,3-Hexamethyldisilazane [*999-97-3*], (HMDS), $(CH_3)_3SiNHSi(CH_3)_3$, is an organosilicon compound that is used to promote adhesion of photoresists on silicon and silicon oxide surfaces during manufacture of semiconductor devices. 1,1,1,3,3,3-Hexamethyldisilazane accomplishes increased photoresist adhesion by reacting with surface hydroxyl groups, which normally repel photoresists, to form hex-

amethyldisiloxane, which does not repel photoresists.

The three methods for applying HMDS to the surface of the semiconductor substrate are evaporation, spinning it on, and complete immersion. In the *evaporative method,* elevated temperature causes HMDS to vaporize and subsequently contact the wafer surface. This technique generally uses 100 % active HMDS, although less concentrated solutions can also be used; a tray of wafers can be treated simultaneously. Several industry sources believe this method to be the most cost effective of the three.

The *spin-on method* applies the liquid HMDS directly to the wafer just prior to applying the photoresist. This approach uses solvent mixtures of HMDS, requiring that the solvent be compatible with the photoresist thinner. Usually solvents such as xylene and stoddard solvent are used with negative resists, while ethylene glycol, fluorocarbon 113, or cellusolve acetate are used with positive resists. Carbon tetrafluoride is compatible with both types of resists. The spin-on technique provides a uniform coating of HMDS on the wafer, but this method can treat only one wafer at a time.

Complete immersion requires dipping the wafer into a bath of HMDS and solvent. This process is not widely used because it offers little control over the amount of HMDS coated on each wafer.

Types and Grades. 1,1,1,3,3,3-Hexamethyldisilazane is available in concentrations ranging in activity from 5 to 100 %. In solutions, HMDS is normally mixed with solvents such as xylene, stoddard solvent, ethylene glycol, and carbon tetrafluoride. All HMDS is sold as electronic-grade material, indicating that the level of each metal impurity is < 1 ppm.

Uses. The semiconductor industry uses HMDS exclusively as an adhesion promoter. The use of HMDS to treat oxide surfaces is covered by a patent assigned to IBM [8]. Licenses under this patent are available from IBM on reasonable terms and conditions.

Economic Aspects. In 1980, ca. 303×10^3 L of 100 % active HMDS was consumed by the semiconductor industry in the United States. The leading suppliers are PCR, KTI, Mallinckrodt, and Ashland Chemical.

9. Cleaners and Specialty Materials

Tin – Lead Strippers. Tin – lead strippers are based mostly on hydrogen peroxide, a clear liquid with a slightly sharp odor. Although not specifically identified, stabilizers are usually included in the formulation to prevent degradation of the base chemical. The average density of tin – lead strippers based on hydrogen peroxide is estimated to be 1.16 g/cm^3. Compositions of the less frequently used products that are not based on peroxide are unavailable.

Uses. Manufacturers of printed circuit boards frequently employ tin – lead stripping to remove tin – lead from edge connectors or tops without damaging the copper or laminate itself. This procedure promotes adhesion of the nickel plate on the copper surface prior to gold striking. Strippers currently available are capable of removing 0.013 mm of tin – lead in < 1 min or up to 135 g of tin – lead per liter. Other products are specifically designed just to remove tin to prevent subsequent poisoning of the nickel plate by enhancing adhesion.

Copper Cleaners. Copper cleaners are generally acid- or alkaline-based products sold in concentrate form. As expected, properties such as density and pH are directly related to the base chemical. For example, the density of acid concentrates is ca. 1.08 g/cm^3 and the pH is usually < 4. The exact amount of water or other solvents mixed with copper cleaners depends on operating results experienced by the individual operator at each company.

Uses. The printed circuit board industry employs copper cleaners for a variety of jobs including removal of organic and inorganic soils, such as fingerprints, light oils, or tarnish, from copper surfaces. Boards are generally sprayed or soaked with the copper cleaner followed by a water rinse.

Tin – Lead Cleaners. Tin – lead cleaners are generally acid-based products containing additives to reduce degradation of components during adverse operating conditions. Some products are employed in full strength, and others

are frequently diluted with water before being applied.

Properties of tin – lead cleaners vary by product, depending on the base chemicals. For example, densities of acid-based products can range from 1.08 to 1.14 g/cm^3 and the pH is typically < 1.

Economic Aspects. In 1980, the printed circuit board industry in the United States consumed 9.0×10^6 L of the following cleaners and specialties (in 10^6 L):

Tin – lead strippers	3.4
Tin – lead cleaners	3.0
Copper cleaners	2.6

Numerous companies supply these products in the United States. The leading ones are Chemelex, J & S Laboratories, MacDermid, Enthone, Argus International, and Electrochemicals.

10. Specialty Gases

10.1. Silicon Gases

Silicon gases are used to deposit silicon and silicon compounds onto silicon substrates by chemical vapor deposition. These deposits or epitaxial layers are part of the finished semiconductor device. Four silicon gases are used to form the epitaxial layers: silicon tetrachloride, trichlorosilane, dichlorosilane, and silane.

Silicon(IV) tetrachoride [*10026-04-7*], $SiCl_4$, tetrachlorosilane, silicon(IV) chloride, is manufactured as a liquid by an exothermic reaction of silicon carbide with chlorine as given in the following equation:

$$SiC + 2\,Cl_2 \rightarrow SiCl_4 + C$$

The major use of $SiCl_4$ is production of fumed silica, but other applications, include (1) synthesis of silicon esters, (2) production of special glass for the electronics industry, and (3) production of high-purity silicon.

Trichlorosilane [*10025-78-2*], $SiHCl_3$, is made by passing hydrogen chloride through silicon powder at a moderate temperature (280 – 300 °C):

$$2\,Si + 6\,HCl \longrightarrow 2\,SiHCl_3 + 2\,H_2$$

The manufacturing equipment used in this process is similar to that for the production of silicon tetrachloride.

The major application for trichlorosilane is as a raw material for producing high-purity polycrystalline silicon, which eventually becomes the polished silicon wafers used in semiconductor production.

Dichlorosilane [*4109-96-0*], SiH_2Cl_2, is a byproduct in the manufacture of trichlorosilane:

$$Si + 2\,HCl \rightarrow SiH_2Cl_2$$

The only significant application for dichlorosilane is as a source of polycrystalline silicon and silicon nitride for epitaxial deposition in the manufacture of semiconductors. However, Hemlock Semiconductor is conducting experiments using dichlorosilane as a raw material for large-scale production of high-purity polycrystalline silicon.

Silane [*7803-62-5*], SiH_4, is produced by combining an ether solution of silicon tetrachloride with an ether solution of lithium aluminum hydride, as shown in the following equation:

$$SiCl_4 + LiAlH_4 \longrightarrow SiH_4 + LiCl + AlCl_3$$

Silane is used largely as a raw material to form other silicon-containing compounds, but it also serves as a source of silicon for deposition of epitaxial layers.

Types and Grades. The four silicon products are all sold as compressed gases. Purity levels of the gases vary among suppliers, but all are available as semiconductor-grade material (> 99 % pure). The SEMI specification for silicon tetrachloride and for silane (for both polysilium – silicon dioxide and silicon nitride applications) are available [5].

Uses and Toxicology. All four of the silicon gases can serve as a source for epitaxial silicon and silicon dioxide by chemical vapor deposition. Silane forms silicon by simple pyrolysis at temperatures of 800 – 1000 °C, as follows:

$$SiH_4 \rightarrow Si + 2\,H_2$$

The other gases form silicon deposits by hydrolysis and release of hydrogen chloride. Silicon tetrachloride requires a temperature > 1200 °C. The reactions are as follows:

$$SiCl_4 + 2\,H_2 \longrightarrow Si + 4\,HCl$$

$$SiHCl_3 + H_2 \longrightarrow Si + 3\,HCl$$

$$SiH_2Cl_2 \longrightarrow Si + 2\,HCl$$

Epitaxial silicon dioxide layers are formed by oxidation of silane; the other gases hydrolyze in a moist atmosphere. Reactions for these processes are as follows:

$$SiH_4 + O_2 \longrightarrow SiO_2 + 2\,H_2$$

$$SiHCl_3 + 2\,H_2O \longrightarrow SiO_2 + 3\,HCl + H_2$$

$$SiH_2Cl_2 + 2\,H_2O \longrightarrow SiO_2 + 2\,HCl + 2\,H_2$$

$$SiCl_4 + 2\,H_2O \longrightarrow SiO_2 + 4\,HCl$$

Silane and dichlorosilane are also used to form epitaxial silicon nitride deposits for the passivation of semiconductor devices. In each case, the silicon gas reacts with ammonia as the following equations show:

$$3\,SiH_4 + 4\,NH_3 \longrightarrow Si_3N_4 + 12\,H_2$$

$$3\,SiH_2Cl_2 + 4\,NH_3 \longrightarrow Si_3N_4 + 6\,HCl + 6\,H_2$$

Silicon tetrachloride is the easiest of the silicon gases to handle. Although corrosive, it does not ignite on contact with the atmosphere, nor is it classified as toxic. Silicon tetrachloride has the fastest layer growth rates and forms the thickest layers in the deposition of silicon and silicon dioxide, which is important in the manufacture of many linear type devices. It does cause distortion pattern shift in the deposition of epitaxial layers because of the chlorine radicals and requires the highest operating temperatures (> 1200 °C) of all silicon gases.

Silane is considered one of the most dangerous of the silicon gases to handle because it ignites on contact with air. However, the flame is relatively cool and no toxic fumes are known to be released. The ACGIH has established a TLV-TWA of 5 ppm, 7 mg/m^3 (no MAK has been established) [9]. The advantages of silane are that it causes little or no pattern shift during formation of epitaxial layers and that it operates at the lowest temperature (800 – 1000 °C) of all silicon gases. Silane also exhibits a layer growth rate that is the easiest to control. As a result, silane is important in the semiconductor industry as the thickness of epitaxial layers decreases with more sophisticated devices.

Both dichlorosilane and trichlorosilane offer a compromise in characteristics between silane and silicon tetrachloride. Dichloro- and trichlorosilane are corrosive and toxic, but do not ignite when in contact with air. Both cause some pattern shift, but not as much as silicon tetrachloride, and both operate at temperatures between 1100 and 1200 °C. Dichlorosilane is particularly effective in producing thick epitaxial layers and requires less processing time than silane.

Economic Aspects. The total United States consumption of silicon gases for semiconductor fabrication was 108.9 t in 1980. Silicon tetrachloride represented the majority of the volume as shown (in t):

Silicon tetrachloride	45.4
Silane	34
Dichlorosilane	22.7
Trichlorosilane	6.8

The major suppliers of these products in the United States are the Linde Division of Union Carbide, Liquid Carbonic, Matheson, Air Products, Scientific Gas Products, and Synthaton.

10.2. Etchant Gases

Etchant gases are used to remove selected areas of materials such as silicon, silicon dioxide, silicon nitride, glass, aluminum, and tantalum during the processing of semiconductor wafers. The most common gas used for etching is carbon tetrafluoride [*75-73-0*], CF_4, tetrafluoromethane. It is a colorless, nontoxic,

stable gas, which can be used in pure form or mixed with oxygen to etch semiconductor wafers. Other gases that can be used as an etchant include silicon tetrafluoride (SiF_4), fluoroform (CHF_3), hexafluoroethane (C_2F_6), perfluoropropane (C_3F_8), and chlorine (Cl_2).

Types and Grades. Carbon tetrafluoride is sold as a compressed gas. It is usually available with a minimum purity level of 99.7 %.

Uses. The etchant gases are employed in a process known as plasma etching. A plasma resembles a gas, but a fraction of the molecules or atoms are ionized and carry an electrical charge. The plasma also contains a substantial quantity of free radicals, which are electrically neutral but highly reactive. The presence of ions and free radicals makes the plasma very reactive and, thus excellent for an etching process.

A plasma etching system consists of three basic units: (1) a reactor, which holds wafers during etching, (2) a radio-frequency (RF) generator, which supplies energy for creating plasmas inside the reactor, and (3) control equipment for monitoring the etching process. Once the reactor is loaded with wafers, air is evacuated by a mechanical pump. An etchant gas then enters the reactor and is ionized to form an etching plasma. Fresh etchant gas flows into the reactor continuously during etching, while gaseous reaction products are withdrawn from the reactor by a vacuum pump.

The specific etchant gas or gas mixture used for plasma etching depends on the type of material to be processed. Examples of the various etchants that are used are listed in Table 9.

Table 9. Examples of etchants used for plasma etching

Etched material	Substrate material	Gas etchant
Polycrystalline silicon	silicon dioxide	CF_4/O_2 or SiF_4/O_2
Silicon nitride	silicon dioxide	CF_4/O_2 or SiF_4/O_2
Silicon nitride	silicon	SiF_4/O_2
Silicon	sapphire	SiF_4/O_2
Silicon nitride	gold	CF_4/O_2
Tantalum	silicon dioxide	CF_4/O_2
Silicon		CF_4
Silicon dioxide	silicon	CF_4

Economic Aspects. The United States semiconductor industry used 91 000 kg of carbon tetrafluoride as an etchant gas in 1980. An additional 9100 kg of miscellaneous etchant gases, including silicon tetrafluoride, chlorine, hexafluoroethane, and perfluoropropane, was also consumed. The only basic manufacturers of carbon tetrafluoride in the United States are Linde, Air Products, and Du Pont. The Du Pont material is sold through such gas suppliers as Matheson and Scientific Gas Products.

10.3. Light-Emitting Diode Gases

Light-emitting diode (LED) gases are used to deposit epitaxial layers of gallium arsenide, gallium phosphide, and gallium arsenide phosphide onto substrates of the same composition during the manufacture of LED and miscellaneous gallium-based semiconductor devices. Arsine [*7784-42-1*], AsH_3, and phosphine [*7803-51-2*], H_3P, are two gases that are used for this purpose. Both gases are extremely toxic, and exposure to concentrations of 1000 ppm for only a few minutes is lethal. The TLV-TWA and MAK for arsine are 0.05 ppm, 0.2 mg/m^3. For phosphine, the TLV-TWA is 0.3 ppm, 0.4 mg/m^3; the TLV-STEL is 1 ppm, 1 mg/m^3; and the MAK is 0.1 mL/m^3, 0.15 mg/m^3 [9], [10].

Types and Grades. Both gases are available as an LED-grade material, which has purity levels of 99.99 – 99.9995 %. The SEMI specifications for impurities in LED-grade phosphine are available [5].

Uses. Arsine and phosphine are used to form epitaxial devices as LEDs and field-effect transistors (FETs). These gases can also be used as dopants for the manufacture of semiconductor devices (see Chap. 13). Most LED manufacturers utilize the gases in high percentage concentrations, 5 – 10 % for arsine and 5 – 15 % for phosphine, mixed with extremely high-purity hydrogen as the background gas. Some manufacturers have begun to use a pure phosphine gas for growing epitaxial layers, but it is a relatively small percentage of the total demand.

Economic Aspects. In 1980, the United States semiconductor industry used 4500 kg of arsine and 1100 kg of phosphine for manufacturing light-emitting diodes. The most important

supplier of these LED-grade gases is Phoenix Research, now part of Union Carbide.

11. Solvents

Fluorocarbon solvents can be either a fluorinated product or a blend of the fluorochemical with alcohols such as ethanol or 2-propanol. One of the most commonly used materials is 1,1,2-trichlorotrifluoroethane [*76-13-1*], $CCl_2F–CClF_2$, which is a clear, dense, and colorless liquid having a faint solvent odor (→ Fluorine Compounds, Organic).

Manufacturers of semiconductors utilize fluorocarbon solvents to clean wafers; the printed circuit board industry employs these products to remove grease from the boards. In addition, manufacturers of printed circuit boards use fluorocarbon solvents in defluxing operations to remove solder flux after the components are attached.

Isopropyl alcohol [*67-63-0*], $CH_3CHOHCH_3$, 2-propanol, has many chemical and solvent properties similar to those of ethyl alcohol; it has a slight odor resembling a mixture of ethanol and acetone (→ Propanols). Production of isopropyl alcohol is based on the reaction of propene and sulfuric acid to form isopropyl hydrogen sulfate, with subsequent hydrolysis to yield the alcohol and sulfuric acid. Isopropyl alcohol purchased by manufacturers of semiconductors is usually a semigrade product assayed at 99.5 wt % $CH_3CHOHCH_3$.

Manufacturers of semiconductors use isopropyl alcohol to clean wafers; the printed circuit industry also employs this product in cleaning operations.

Acetone [*67-64-1*], CH_3COCH_3, is an inflammable liquid that is low in toxicity compared to other volatile solvents. Acetone is produced by either oxidation of cumene or catalytic dehydrogenation of isopropyl alcohol (→ Acetone). Acetone purchased by manufacturers of semiconductors is often a semigrade product assayed at 99.5 wt % CH_3COCH_3.

The semiconductor industry uses acetone for wafer cleaning, and manufacturers of printed circuit boards use the solvent to clean masks contaminated with dirt.

Trichloroethylene [*79-01-6*], TCE, $ClCH{=}CCl_2$, is a clear, water white, volatile liquid that is completely miscible with most organic solvents, but only slightly soluble in water (→ Chlorinated Hydrocarbons, Chap. 3.4.). The production of TCE is generally accomplished by chlorination of ethylene, forming ethylene dichloride, which is then oxychlorinated to form trichloroethylene. Manufacturers of semiconductors purchase this electronic chemical typically in a semigrade product assayed at 99 wt % TCE.

The semiconductor industry consumes limited quantities of trichloroethylene to clean wafers; however, this industry also uses this product as a chlorination source in crystal growing furnaces, whereby the chlorine minimizes crystal defects. Manufacturers of printed circuit boards use this material in metal cleaning and degreasing operations.

***n*-Butyl acetate** [*123-86-4*], $CH_3COO(CH_2)_3CH_3$, is produced by esterification of acetic acid with *n*-butyl alcohol, using a catalyst, typically sulfuric acid. Manufacturers of semiconductors generally purchase a semigrade product assayed at 99 wt %.

Manufacturers of semiconductors use *n*-butyl acetate as a rinse for negative resists to stop the developing action.

Methanol [*67-56-1*], CH_3OH, methyl alcohol, is produced from steam-re-formed natural gas, which is compressed, catalytically converted to crude methanol, and finally purified by distillation (→ Methanol). Purchases of methanol by the semiconductor industry are typically assayed at 99.9 wt % CH_3OH.

Manufacturers of semiconductors use methanol for wafer cleaning.

1,1,1-Trichloroethane [*71-55-6*], CH_3CCl_3, is one of the least toxic chlorinated hydrocarbons. The most popular route to manufacture this material is from vinyl chloride, which is hydrochlorinated with hydrogen chloride to form 1,1-dichloroethane and then thermally chlorinated to give 1,1,1-trichloroethane (→ Chlorinated Hydrocarbons, Chap. 2.4.).

Semiconductor companies use this electronic chemical for wafer cleaning and in crystal growing furnaces to reduce defects. Manufacturers of printed circuit boards use this material in cleaning operations; however, its more important application in this industry is to develop resists.

Types and Grades. The SEMI specifications for acetone, *n*-butyl acetate, methanol, trichloroethylene, and trichlorotrifluoroethane are available [5].

Economic Aspects. A total of 14.60×10^3 t of solvents was used in the United States for electronic applications in 1980. Of the total, 9.88×10^3 t was used in semiconductor applications and the balance was used in printed circuit boards, as shown in Table 10.

Table 10. Solvents used in printed circuit boards in 1980

Product	Consumption, 10^3 t		
	Semiconductors	Printed circuit boards	Total
Fluorocarbon solvents	0.68	1.13	1.81
Isopropyl alcohol	2.22	0.91	3.13
Trichloroethylene	1.59	0.91	2.50
Acetone	1.72	1.09	2.81
n-Butyl acetate	1.13	0.23	1.36
Methanol	1.81		1.81
1,1,1-Trichloroethane	0.73	0.45	1.18
Total	9.88	4.72	14.60

The major suppliers of semiconductor-grade solvents are Ashland Chemical, General Chemical, J. T. Baker, and Mallinckrodt. There are numerous suppliers of solvents to the printed circuit board industry, including Du Pont, Dow, Shell, Union Carbide, PPG, and Exxon.

12. Masks

Solder Masks. Most solder masks are permanent and are intended to remain on the board; thus, they are usually made of durable scratch-resistant material such as epoxy resins. Temporary solder masks are normally used to protect gold contact fingers from being soldered; the mask can be removed at a later date to expose the gold contacts.

Screening Masks. Screening masks are photosensitive polymeric resins that are screen printed onto a printed circuit board. Once printed, the mask is cured by using either heat or ultraviolet light. After the mask is cured, it acts as a solder resist, selectively protecting the circuitry during component soldering.

Types and Grades. Two types of screening masks are available. One type can be cured thermally, and the other type can be cured with ultraviolet radiation. Both can be either permanent or temporary resists. The masks are supplied as 100 % active material and are ready for use.

Uses. Screening masks are used exclusively by the printed circuit board industry as a solder resist. The mask is applied to the board by forcing it through a stainless steel or polyester monofilament screen. The screen pattern determines where the mask is deposited on the board. Once applied to the board, the mask is cured either by heat or by ultraviolet light. After curing, temporary masks can be removed from the board with hot alkaline solutions of sodium hydroxide and water. Permanent solder masks are not made to be removed after curing.

Economic Aspects. A total of 852×10^3 L of screening masks was consumed by the printed circuit board industry in the United States in 1980. Of this total, ca. 85 % or 719×10^3 L were thermally curing masks. The major suppliers of screening masks in the United States are MacDermid, Photocircuits, Mysol, Colonial, London Chemical, Grace, and Dynachem.

Dry Film Masks. Dry film solder masks consist of a photosensitive polymer layer sandwiched between a layer of polyester film on one side and a polyolefin sheet on the other side. Both cover sheets protect the photopolymer from damage by abrasion. The polyolefin sheet is removed as the solder mask is laminated to the printed circuit board. Ultraviolet light selectively images the mask by polymerizing the exposed photopolymer, causing these areas to be insoluble in the developing solution. The polyester sheet is removed just prior to immersion in the developing solution. After the developing process, the dry film acts as solder resist, protecting the circuitry during component soldering.

Types and Grades. Currently, only two types of dry film masks are available. One type has a heat-resistant coating, and the other has a flame-retardant coating. Both products work negatively and can be processed in solvent, usually with 1,1,1-trichloroethane.

Uses. Dry film masks are used exclusively by the printed circuit board industry and are designed for boards with tight circuitry spacings (≤ 0.3 mm). The developed dry film protects the circuitry from coming in contact with solder when components are assembled on the board.

Economic Aspects. Dry film solder masks were introduced in 1979 by Du Pont. Therefore, consumption of these products in 1980 was low, at 30 480 m^2. Du Pont, Dynachem, and Hercules all offer dry film masks.

13. Dopants

Dopants are materials that enhance the ability of a semiconductor to conduct electricity by selectively forming positively and negatively charged areas within the substrate.

As the dopant contacts the substrate, dopant atoms, such as phosphorus or boron, begin to replace semiconductor atoms, such as silicon, in the crystal structure. Electric charges form because the change of atoms also changes the number of valence electrons that are available in the substrate. Replacing a silicon atom, which has four valence electrons, with a boron atom, which has only three valence electrons, produces a positively charged area because of the absence of one electron. Conversely, replacing a silicon atom with a phosphorus atom, which has five valence electrons, produces a negatively charged area because of an excess of one electron. This imbalance of electrons allows better conduction of electricity through the semiconductor material.

13.1. Liquid Dopants

Liquid dopants are compounds that contact semiconductor substrates to produce charged areas. These liquids include compounds that are liquid at room temperature and solid compounds that are dissolved in solvents to form solutions. The main difference between the two is the method of application. The dopant – solvent solutions are applied by spin-on techniques, while the pure liquids are applied by chemical vapor deposition (CVD) techniques.

Orginally, all liquid dopants were formulated for application by *CVD techniques.* This technique uses a carrier gas, usually nitrogen, which bubbles through the liquid dopant and carries the dopant vapor mixture into a furnace, where the dopant diffuses into the wafer.

Later, *spin-on techniques* were developed. This method uses several drops of the dopant solution, which is placed directly on the wafer, spun to a uniform thickness, and dried to form a dopant film. Once dry, the wafers are placed in a furnace operating between 1100 and 1300 °C, where the diffusion process occurs.

Types and Grades. The two major classifications of liquid dopants are inorganic- and organometallic-based compounds. Within the *inorganic class,* the four main types of doping compounds are based on phosphorus, boron, antimony, or arsenic. The most common inorganic liquid dopants, with the method of application and available purity grades, are shown in Table 11.

The *organometallic dopants* consist of compounds such as diethylzinc, dimethylcadmium, diethyltelluride, and trimethylgallium. These products are available as electronic-grade materials that have a purity in excess of 99.9 % and are formulated for use by CVD techniques.

Uses. The spin-on type of liquid dopants are formulated into several different products by varying the dopant concentration of the solution. These variations cause different concentrations of dopant atoms per unit area in the semiconductor substrate.

Phosphorus, antimony, and arsenic compounds are examples of dopants that produce negatively charged sites (*n*-type dopant); boron, aluminum, indium, and gallium compounds are examples of dopants that produce positively charged sites (*p*-type dopant).

Table 11. Most common inorganic liquid dopants

Product	CAS registry number	Formula	Method of application	Purity grades, wt %
Phosphorus oxychloride	[*10025-87-3*]	$POCl_3$	CVD *	99.995, 99.999, 99.9999
Phosphorus tribromide	[*7789-60-8*]	PBr_3	CVD *	99.995, 99.9999
Phosphorus pentaoxide	[*1314-56–3*]	P_2O_5	spin-on	99.99
Boron tribromide	[*10294-33-4*]	BBr_3	CVD *	99.99, 99.995, 99.999, 99.9999
Antimony trioxide	[*1309-64-4*]	Sb_2O_3	spin-on	99.99
Arsenic trioxide	[*1327-53-3*]	As_2O_3	spin-on	99.99
Antimony trichloride	[*10025-91-9*]	$SbCl_3$	CVD *	99.995, 99.9999

* Chemical vapor deposition.

Economic Aspects. In 1980, almost 6300 kg of liquid dopants was consumed in the United States for semiconductor fabrication. Of this total, ca. 2300 kg was supplied as CVD dopants and the remainder as spin-on dopants. The two leading suppliers of CVD dopants are J. C. Schumacher, now part of Air Products, and Apache Chemical, now part of Olin-Hunt. The important suppliers of spin-on dopants are Emulstone, Diffusion Technology, and General Chemical.

13.2. Gas Dopants

The three primary doping gases are arsine [*7784-42-1*], AsH_3; diborane [*19287-45-7*], B_2H_6; and phosphine [*7803-51-2*], H_3P. All three are colorless, highly toxic, and flammable.

Traditionally, the gas mixture is pumped into a furnace operating at a temperature between 900 and 1300 °C and is introduced into the substrate by diffusion. A newer technique, called ion implantation, operates at room temperature and is gaining popularity for doping of the semiconductor substrate.

Types and Grades. An electronic grade of dopant is used, requiring a minimum of 99.995 % purity. Gas dopants are typically mixed, usually with nitrogen as the carrier gas, although 100 % phosphine is sometimes used. The concentration of dopant gas in the mixture ranges from 5 to 50 000 ppm with the exception of diborane, whose concentration only ranges from 5 to 10 000 ppm. The SEMI specifications for electronic-grade arsine and phosphine are available [5].

Uses. Dopant gases are used to produce selectively charged areas within the semiconductor substrate and within such epitaxial layers as polysilicon and silicon nitride. Arsine and phosphine are examples of dopants that produce negatively charged sites (*n*-type dopant); diborane is an example of a dopant that produces positively charged sites (*p*-type dopant).

Economic Aspects. The total consumption of gas dopants in the United States for semiconductor fabrication was 2300 kg as 100 % active material in 1980. The four largest suppliers of these products in the United States are Matheson, Linde, Scientific Gas Products, and Air Products.

13.3. Solid Dopants

Solid dopants, either in crystalline, powder, or wafer form, contact semiconductor substrates to produce positively or negatively charged sites. Usually these dopants are compounds of metals such as arsenic, antimony, boron, and phosphorus.

Types and Grades. The two main types of solid dopants are boron and phosphorus compounds. These materials are available in grades varying in purity from 99.9 to 99.999 % and are ready for immediate use.

Uses. The boron compounds are examples of dopants that produce positively charged sites (*p*-type dopant) and the arsenic, antimony, and phosphorus compounds are examples of dopants that produce negatively charged sites (*n*-type dopants).

Crystalline or powder-type solid dopants are sublimed and pumped into the diffusion reactor, where the material diffuses into the semiconductor substrate, doping the desired areas. Most boron solids, particularly boron nitride, are available in wafer form and are placed inside the diffusion reactor beside the semiconductor substrate. The boron sublimes and travels a relatively short distance to contact the substrate. Ion implantation may eventually be an alternate application method, and some semiconductor device manufacturers are experimenting with this process.

Economic Aspects. In 1980, ca. 860 kg of solid dopants were consumed by the United States semiconductor industry. Mallinckrodt is the leading supplier for the phosphorus, arsenic, and antimony compounds, and Carborundum is the leading supplier of boron nitride.

14. Auxiliary Materials

A host of auxiliary materials, including ceramic substrates for thick- and thin-film circuits, liquid crystals, conductive inks and pastes, high-purity metals, wafer cleaners, and passivation materials, are used in the manufacture of electronic devices.

14.1. Ceramic Substrates

Ceramic substrates of high-purity alumina (Al_2O_3) are used as the base material in the manufacture of thick- and thin-film circuits (→ Ceramics, Electronic). The alumina content of substrates used for thick films ranges from 94 to 96 %; most thin-film circuits use ceramics having an alumina content of > 99 %.

Because most thick- and thin-film circuits are designed for specific purposes, the shape and size of the ceramic base is made to meet consumer specifications. Consumers are typically semiconductor manufacturers that use thick- and thin-film circuits for resistor networks and hybrid circuits, thus augmenting the capabilities of the semiconductor devices.

14.2. Liquid Crystals

Liquid crystals are used in conjunction with semiconductor devices in applications such as watches, calculators, and electronic games (→ Display Technology). Although not a semiconductor device, the liquid crystal serves as a display or read-out that rivals the light-emitting diode, which is a semiconductor device. The liquid crystal operates by applying an electric charge, which changes the molecular orientation of the liquid. Light passing through the crystal is scattered or reflected where the charge is applied, thereby forming the display. This process allows very little current consumption by the liquid crystal display, which is its main advantage over the light-emitting diode.

14.3. Conductive Inks and Pastes

Conductive inks and pastes are screened onto a ceramic substrate to form the conductors, resistors, and capacitors in thick-film circuits. Thick-film conductors are made from precious metal pastes or inks, such as gold, silver, platinum, and palladium, either separately or in combination. Resistors and capacitors use pastes that consist of glass dielectric materials and metal or semiconducting compounds. The pastes and inks are deposited on the substrate by a stencil screen process and are then fired in a furnace, where the composition of the material determines the final characteristics of the circuit. Other components, such as integrated circuits or discrete semiconductor devices, can be added to the thick-film assembly to form a hybrid circuit device.

14.4. High-Purity Metals

High-purity metals have applications with both semiconductor devices and thin-film hybrid circuits. Aluminum and aluminum alloys (mainly tantalum) are used exclusively to form metallized areas on integrated circuit devices. Precious metals, such as palladium and platinum, are used to form conductive paths in thin-film hybrid circuits. All the metals used have purity levels ranging from 99.9999 to 99.99999 %.

The two methods for depositing the metal onto the substrate are sputtering and evaporation. *Evaporation* is the simpler technique and

involves placing the substrate and the metal source in a ball jar under vacuum. The metal is vaporized by heat and then moves in all directions within the chamber and deposits on the substrate.

Sputtering uses the metal source as a cathode target, which is bombarded by ions of an inert gas. This causes the metal to be deposited on the substrate surface.

14.5. Passivation Materials

Passivation materials protect the finished semiconductor device from damage by environmental factors such as dust, humidity, and temperature. The passivator is applied as the final layer on the semiconductor device. Several types of materials can be used to form passivation layers, such as silicon nitride formed by the reaction of silane and ammonia gases, and polyimide coatings. Only the polyimide material is discussed in this section; the previously mentioned gases are covered in Chapter 10.

The polyimide material is sold as a polyamic acid solution, which is converted by heat to a stable polyimide coating. The material can be applied to the wafer by screening, spraying, dipping, or a spin-on method. The polyimide forms a tough, temperature-resistant coating without pin holes and seems to have an advantage over silicon nitride with devices having many stepped layers. In addition, the polyimide can be used with a photoresist to expose metallized contact areas on the device while the polyimide protects the rest of the chip.

14.6. Wafer Cleaners

Wafer cleaners can be hydrophilic polymers or surfactant materials that are either sprayed, dipped, or spun onto the wafers. These materials remove particles from the substrate surface while such cleaners also protect against edge fracturing.

15. References

General References

1. E. D. Feit (Harris Semiconductor Group), C. W. Wilkins, Jr. (Bell Laboratories) (eds.): *Polymer Materials for Electronic Applications,* Am. Chem. Soc., Washington, D.C., 1982.
2. E. R. Howells (ed.): *Technology of Chemicals and Materials for Electronics,* Society of Chemical Industry, London 1984.
3. L. F. Thompson (Bell Laboratories), C. G. Willson (IBM Research Laboratory), M. J. Bowden (Bell Laboratories) (eds.): *Introduction to Microlithography,* Am. Chem. Soc., Washington, D.C., 1983.

Specific References

4. S. R. Daigle, M. S. Rappaport, J. Diffenthal, C. J. Listner in A. A. Boccone (ed.): *Electronic Chemicals, United States, Western Europe, and Japan,* Charles H. Kline & Co., Fairfield, N.J., 1981.
5. Semi-Standards Committee Chemical Section (ed.): *Book of Semi-Standards,* **vol. 1,** Semiconductor Equipment & Materials Institute, Mountain View, Calif., 1982, pp. C1. STD. 1 – 22, C2. STD. 1 – 3, C.3 STD 1 – 11.
6. C. G. Henningsen in C. F. Coombs (ed.): *Printed Circuits Handbook,* McGraw-Hill, New York 1979, Section 7, pp. 2 – 18.
7. R. J. Bahn in C. F. Coombs (ed.): *Printed Circuits Handbook,* McGraw-Hill, New York 1979, Section 7, p. 5.
8. IBM, US 3 549 368, 1970 (R. H. Collins, F. T. Deverse).
9. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* ACGIH, Cincinnati, Ohio, 1986 – 1987.
10. DFG (ed.): *Maximum Concentrations at the Workplace (MAK),* VCH Verlagsgesellschaft, Weinheim 1986.

Emulsions

RUDOLF HEUSCH, Bayer AG, Leverkusen, Federal Republic of Germany

1. Definitions 2
2. Properties of Emulsions 4
2.1. Droplet Formation 4
2.2. Enveloping of the Droplet Surface . 4
2.3. Electrical Charge of Emulsion Droplets 5
2.4. Interfacial Tension 5
2.5. Stability 5
2.6. Sedimentation and Coalescence . . 6
2.7. Phase Volume Ratio 6
2.8. Emulsion Type 6
2.9. Phase Reversal 7
2.10. Rheology 7
3. Basic Processes in the Preparation of Emulsions 8
3.1. Breaking up to Drops 8
3.2. Increase in Surface Area 9
3.3. Work of Division 9
3.4. Decreasing the Interfacial Tension 9
3.5. Other Factors Governing Emulsification 10
3.6. Mixed Emulsifiers 11
4. Basic Processes in the Stabilization of Emulsions 12
4.1. Electric Charge 12
4.2. Steric Factors 12
4.2.1. Molecular Mobility 12
4.2.2. Complex Formation 12
4.2.3. Polymers in Emulsions 13
4.3. Further Stability Factors 14
4.4. Long-Term Stability 14
4.5. Emulsion Thickening Agents 14
4.6. Cream Formation 14
5. Preparation of Emulsions 15
5.1. Theories Relating to Mechanical Disintegration 15
5.2. Emulsifying Machines 15
5.3. Use of Emulsifiers 17
5.3.1. General Rules 17
5.3.2. Preparation of Multiple Emulsions . . 17
5.3.3. Preparation of Microemulsions 17
5.3.4. Spontaneous Emulsification 21
5.3.5. Inorganic Emulsifiers 21
6. Breaking of Emulsions 21
7. Chemical Structure of Emulsifiers 22
7.1. Surface-Active Emulsifiers 23
7.2. Natural Products 23
7.3. Inorganic Emulsifiers 23
7.4. Sources of Supply 24
8. Chemical Structure of Emulsion Promoters 24
8.1. Polymers 24
8.2. Special Emulsion Promoters 24
9. Selection of Emulsifiers 26
9.1. Selection Criteria 26
9.2. Selection Systems 27
9.3. Assignment of HLB Values 29
9.4. Selection Procedure 31
10. Fields of Application 31
10.1. Cosmetics 31
10.2. Pharmaceutical Preparations 32
10.3. Agriculture 32
10.4. Crop Protection 32
10.5. Food Industry 32
10.6. Textile Industry 33
10.7. Plastics Industry 33
10.8. Building Industry 34
10.9. Paints and Coatings 34
10.10. Paper Industry 34
10.11. Leather Industry 35
10.12. Mining 35
10.13. Metal Industry 35
10.14. Electroplating 35
10.15. Polishes, Manufacture and Removal 35
10.16. Petroleum Industry 35
10.17. Mineral Oil Industry 36
10.18. Adhesives Industry 36
10.19. Photographic Industry 36
10.20. Chemical Industry 36
10.21. Detergents and Cleansers 36
10.22. Special Applications 36
11. Investigation of Emulsions 37
11.1. Stability 37
11.2. Emulsion Type 37
11.3. Droplet Size and Droplet Size Distribution 38
11.4. Interfacial Tension 38
11.5. Required HLB Value 39
11.6. Physical Parameters of Emulsions 39
12. References 40

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_297

1. Definitions [5], [6], [17], [22], [25], [26]

Emulsions are disperse systems consisting of two (or more) mutually insoluble or sparingly soluble liquids (→ Disperse Systems and Dispersants). The liquid usually present in excess is termed the *closed, continuous,* or *external phase*, while the liquid dispersed in it is termed the *internal* or *dispersed phase*. If the degree of dispersion is so high that the particles have colloidal dimensions, the term "phase" should no longer be applied to the particles.

The preparation of an emulsion is termed *emulsification* and the agents used for this purpose are termed *emulsifiers* In addition to the emulsifiers, other agents, such as *emulsion promoters* or *stabilizers* are often added to an emulsion to boost the emulsifying process, for example, by increasing the viscosity or providing a protective colloid action.

If the external phase consists of water and the internal phase of an organic liquid, e.g., mineral oil, the term *"oil-in-water"* (o/w) emulsion is used. On the other hand, if water is finely dispersed in a nonaqueous liquid, a *"water-in-oil"* (w/o) emulsion is produced. International abbreviations are being introduced, L – H for an o/w emulsion and H – L for a w/o emulsion, L and H being derived from the Greek "lipos" and "hygros" [39] (Fig. 1). If two nonaqueous liquids, e.g., ethylene glycol in a liquid polyether, are emulsified, the term *"oil-in-oil"* emulsions, abbreviated to o/o emulsion, is used. If two aqueous phases are emusified in each other, they are described as w/w emulsions (see Section 4.2.3).

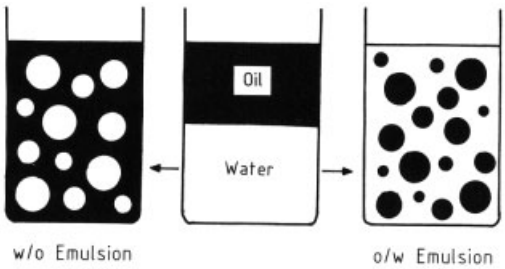

Figure 1. Types of emulsion in aqueous systems [5, p. 173]

For example, cow's milk is an o/w emulsion consisting of ca. 3.5 % fat dispersed in the aqueous phase. On the other hand, butter is a w/o emulsion containing up to 20 % water in the dispersion.

The formation of a natural emulsion in cow's milk, for example, results in an optimum dispersion and dilution of the nutrients necessary for feeding newborns. In an analogous manner, emulsions are prepared industrially to improve the effect of an active substance in a particular application by using it in a more or less finely dispersed form, or to achieve a convenient method of controlling chemical reactions such as, for example, chlorination or polymerization reactions. In such cases the external phase assumes the role of a transporting and/or diluting agent. Thus, in crop protection, for example, it is usually necessary to produce the active substances in diluted form as an emulsion or a suspension.

The terms *multiple-phase* or *polyphase emulsions* are used to describe systems in which water droplets are already dispersed in the oil droplets of an o/w emulsion (w/o/w emulsion) or systems in which oil droplets are dispersed in the water droplets in a w/o emulsion (o/w/o emulsion). The preparation sequence should be denoted by indices [40]. This, a $w_1/o/w_2$ emulsion is made by dispersing a w/o emulsion in water (Fig. 2).

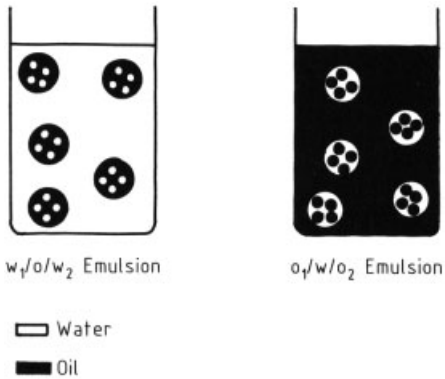

Figure 2. Polyphase emulsion [5, p. 175]

The term *microemulsion* [1], [3, p. 287], [41–43] is now used to describe an isotropic liquid phase with a specific microstructure that is formed spontaneously [44], [45] and in which the particle diameter is less than 10^{-8} m. These are usually intermediate states between micellar solutions and mesomorphic structures.

The term *atypical emulsion* is used if the emulsifier is no longer free to move in highly viscous systems but is more or less fixed crystallographically, usually as liquid crystals [46].

The term *spontaneous emulsion* is used to describe emulsions that can be formed without supplying substantial amounts of mechanical energy (mean particle diameter 10^{-7} m). These are primarily required in crop protection [47], [48]. For further definitions, see → Disperse Systems and Dispersants.

Solubilization. By *hydrotropy* or *solubility promotion,* a phenomenon is meant that involves a substance being rendered soluble in a pure solvent in which it is sparingly soluble by the addition of another substance. If this third substance is a surface-active compound, the term *solubilization* is used, because the sparingly soluble substances dissolve more easily in the surfactant micelles than in the aqueous solution of individual surfactant molecules [49].

In the case of aromatic or colored substances, it is easy to follow the solubilization process by spectrophotometry (Fig. 3).

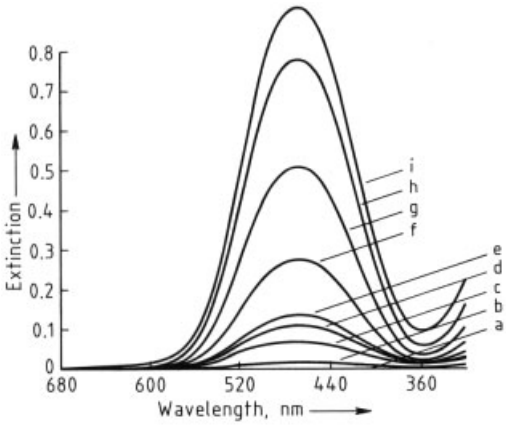

Figure 3. Solubilization of Resolin Brilliant Orange RS in water (500 mg/L) by emulsifier PV-15, a polyethoxylated condensed phenol [50], [51]
In order of increasing emulsifier concentration (g/L): a) Without; b) 0.1; c) 1; d) 5; e) 10; f) 20; g) 30; h) 50; i) 100

In the case of emulsions, solubilization becomes involved if the emulsifier is added to the system in a concentration that is greater than the critical concentration for micelle formation [52], [53].

2. Properties of Emulsions

2.1. Droplet Formation

An emulsion is generally imagined to be a milky, turbid liquid such as the familiar cow's milk or rubber latex. However, this milky and turbid appearance arises only in the case of a fairly coarse dispersion of the particles in the external phase (mean particle diameter 10^{-4} – 10^{-6} m). Emulsions with a mean particle diameter of 10^{-6} – 10^{-8} m have the typical appearance of colloidal solutions and exhibit the Tyndall effect [54, p. 176]. If the particles are smaller than 10^{-8} m, the emulsion is highly disperse and begins the transition to a micellar or true solution.

The color of the emulsion often gives a rough indication of the particle size. If the difference in refractive indices of the emulsified components is large enough, emulsions with a bluish appearance are usually more finely dispersed than those with a yellow or reddish appearance.

The particle size of most emulsions is not uniform. If the droplet size exhibits a wide statistical distribution, the emulsion is described as *polydisperse,* in contrast to *monodisperse* systems with a uniform droplet size [55] (→ Disperse Systems and Dispersants). The ideal particle size depends on the availably methods of preparation and industrial application in each case. Most emulsions prepared industrially are polydisperse.

2.2. Enveloping of the Droplet Surface

To prevent droplets from reuniting once they have been formed, energy barriers must be erected that either prevent the droplets from colliding or from fusing if they do collide. The barriers are erected by adding emulsifiers and can be maintained by adding stabilizers.

The polar emulsifier molecules accumulate at the interface between the external and internal phases so that their hydrophilic groups project into the water and their hydrophobic groups project into the organic phase. This produces an interface film, which becomes more extensive and adsorbs additional emulsifier molecules if the droplets are broken up mechanically, until all the droplets are enveloped by a surface film [56]. In the case of an o/w emulsion, the hydrophobic hydrocarbon chains are contained in

the internal dispersed phase (Fig. 4 A), whereas in the case of a w/o emulsion, they project into the external continuous phase (Fig. 4 B).

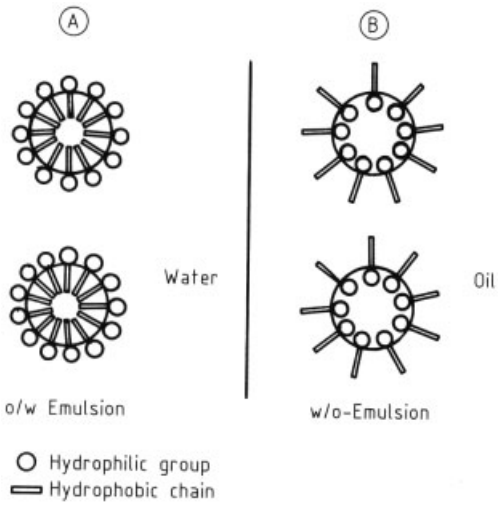

Figure 4. Envelopment of droplets with a surface film of emulsifier in an o/w emulsion (A) and a w/o emulsion (B)

The thickness of the adsorption layer surrounding the droplets of the emulsion depends on the system and the concentration of the ingredients. Only a few emulsifiers or emulsion promoters can be expected to result in a monomolecular boundary layer, and as a rule, multilayers are built up. Liquid – crystalline boundary layers may be produced with such thickness that the layers become microscopically visible. For information on the calculation of the equilibrium thickness of adhering and free lamellae, see Derjaguin [57], [58] and Frenkel [59].

2.3. Electrical Charge of Emulsion Droplets

In addition to the size of the droplets, the electrical charge of the droplets plays a decisive role in emulsification and stabilization processes (see Section 4.1).

2.4. Interfacial Tension

The accumulation of an emulsifier or an emulsion promoter at the interface between the internal and external phase results in a decrease of interfacial tension between the phases. Thus, the interfacial tension between paraffin oil and water normally is 38.7 mN/m, but after the addition of oleyl decaglycol ether (2 g/L), this tension drops to 5.2 mN/m (for further formation, see Section 3.4).

2.5. Stability [3, p. 129], [11, p. 76],[17, p. 522]

An emulsion is stable if fusion of the droplets is prevented by a sufficiently high energy barrier. In general, the energy barrier is built up by the film of emulsifier that forms at the surface of the droplets. In the absence of an emulsifier, triboelectric charging is probably the only stabilizing factor.

A maximally stable emulsion, however, is not always desirable; its stability should match its application. Thus, for a number of applications, the emulsion should be stable under very specific conditions, but it should then disintegrate into its components (break) after its purpose has been achieved.

Most of the emulsions used in the *textile industry* are designed to have an optimum effect at a particular temperature, at a particular pH, or at a particular concentration of dyes, salts, or the like. Thus, in waterproofing textile materials such as rain coats or tarpaulins, for example, the water-repellent finish is usually applied as an aqueous emulsion. This guarantees a uniform wetting and soaking of the fiber material and, consequently, a uniform application of the waterproofing agent to the fiber. The water subsequently must be removed from the textile material by drying. Breaking the emulsion at as low a temperature as possible facilitates the drying considerably.

Medicines are now increasingly supplied as emulsion concentrates. Among other things, this method of preparation enables the dispensing of dosages of sparingly soluble products to be improved. However, to achieve a rapid absorption of the drug, the retention of the active substance by emulsifiers should not be excessive. Breaking

of the emulsions, for example, by a change in the pH in the intestine causes the desired result.

In *crop protection* and in *pest control,* emulsion concentrates are used in fields, gardens, or greenhouses after dilution with water to form spray mixtures. Since complicated emulsifying equipment is not usually available, emulsifier mixtures, which facilitate spontaneous dispersion of the active substance in water and guarantee uniform distribution of the active substance over large areas, are often used. However, the stability of the spray mixtures should not be high, since they must not be kept for a long time. On the other hand, the emulsion concentrate that is delivered must be stable for some period of time.

Further requirements are very often imposed on the stability of industrial emulsions. For example, they should be stable to storage, heat and/or cold, and in some cases, they should be insensitive to the addition of acids, bases, and/or salts [60]. The production of a stable emulsion depends also on the rate at which the emulsifiers or emulsion promoters that are added diffuse to the freshly formed interfaces and become distributed and concentrated in them.

2.6. Sedimentation and Coalescence

The uniform dispersion of the droplets in the external phase may be destroyed reversibly by sedimentation or irreversibly by coalescence.

Sedimentation. In sedimentation, the uniform dispersion of the droplets is disturbed by aggregation, which leads to *settling* or *creaming* of the internal phase. To keep an emulsion stable, such aggregation must be prevented since droplet aggregates sediment faster than individual small droplets. Sedimentation is not always necessarily accompanied by coalescence (see next paragraph). Although the distribution has been altered, the original dispersion can be restored by shaking or stirring. Often such reversible sedimentation in an emulsion cannot be prevented; however, it is not necessarily troublesome.

Coalescence. In coalescence, the individual droplets fuse together. First, the smaller droplets are absorbed by the larger droplets, and then increasingly larger drops merge together until two continuous phases are finally formed. Two droplets can only coalesce if the intervening layer of liquid is pierced when they approach each other. Therefore, coalescence is opposed in two ways by the emulsifier film surrounding the droplets. First, as in the case of aggregation, the like charges of the electrical double layer prevent them from approaching each other. Second, the buildup of an elastic surface film causes the emulsion droplets to bounce off each other when they collide [61]. Coalescence is always followed by accelerated settling or creaming, which destroys the emulsion completely. The emulsion is then broken and cannot be reconstituted by shaking or stirring.

2.7. Phase Volume Ratio

The ratio of the volume of the internal phase V_i to the volume of the external phase V_e is called the *phase volume ratio* Φ:

$$\Phi = \frac{V_i}{V_e} \tag{1}$$

The corresponding ratio of masses is called the *phase mass ratio.* If $\Phi < 0.43$ ($V_i = 30$ % of total volume), the flow properties of the emulsion are determined primarily by the external phase. If $\Phi > 0.43$, the viscosity of the emulsion generally increases. As Φ increases, either phase reversal or cream formation occurs. A phase volume ratio of ca. 1.1 ($V_i = 52$ % of total volume) corresponds to a body-centered cubic (bcc) packing of spheres, one of ca. 1.8 ($V_i = 64$ % of total volume) to face-centered cubic (fcc) packing of spheres. At phase volume ratio of 2.6 ($V_i = 72$ % of total volume) the packing of spheres is hexagonal and maximum space filling is achieved [62], [10, vol. I].

2.8. Emulsion Type

Whether a w/o or an o/w emulsion is formed depends essentially on the stability of the emulsifier layer surrounding the droplets. If a water-stable emulsifier envelope is formed around the oil droplets in a system containing water, oil, and emulsifier, an o/w emulsion is produced. On the

other hand, the formation of an oil-stable envelope around the water droplets produces a w/o emulsion. The rate at which the droplets coalesce is the determining factor [63]. If the water droplets coalesce more rapidly than the oil droplets, an o/w emulsion is formed, and vice versa [64].

The selection of suitable emulsifiers is specially important in producing an emulsion of the required type. For the solubility of the emulsifier in the internal and external phases (Bancroft rule), see Section 9.1. DAVIES links the emulsion type that is formed with the HLB (hydrophilic – lipophilic balance) values (see Section 9.2). Emulsifiers with a fairly low HLB value form w/o emulsions, and those with a high HLB value tend to form o/w emulsions. Emulsifiers with a moderate HLB value have an equal tendency to form both types of emulsion [63].

The phase volume ratio may also affect the emulsion type. In aqueous systems, an excess of water favors the formation of an o/w emulsion, while a water deficiency favors a w/o emulsion. Moreover, interface films carrying a weak electric charge tend to form w/o emulsions, while o/w emulsions occur with highly charged films [65], [66]. In the phase diagram [36, p. 125], the o/w, and w/o emulsion regions are separated by a line with a constant water – emulsifier ratio [67].

The assumption that a given emulsifier can only form one type of emulsion is inadmissible [12, p. 327]. Although this basic tendency certainly exists, it is affected by the composition of the system, by the phase volume ratio, by the sequence in which the components are added, by the temperature, and by mechanical factors. If mixtures of emulsifiers are used, their mixing ratio may determine the type of emulsion, especially if emulsifiers with opposing tendencies (opposing HLB values) are combined.

2.9. Phase Reversal

The conversion of a w/o emulsion into an o/w emulsion, or vice versa, is called phase reversal. It may occur if the amount of internal phase of an emulsion is increased. However, the emulsion must not be broken in this process, which can be achieved by selecting a suitable emulsifier system. Phase reversal can be achieved relatively easily with a mixture of a water-soluble and an oil-soluble component. The phase reversal is linked to the solubility of the emulsifiers in the two phases [68].

If, for example, a mixture of 2 g of calcium alkyl benzenesulfonate and 3 g of a highly ethoxylated condensed phenol is dissolved in 50 cm^3 of xylene, a w/o emulsion is first obtained when water is added. If the mixture is not stirred too rapidly, phase reversal occurs when ca. 300 cm^3 of water has been added. As further water is added, the o/w emulsion is maintained.

Furthermore, phase reversal can be achieved by increasing the charge of a weakly charged w/o emulsion or by discharging a charged surface film in an o/w emulsion [65]. Phase reversal usually leads to the formation of a finely divided stable microemulsion (see page 13 and 5.3.3). Emulsions prepared by phase reversal are sometimes called inverted emulsions.

The formation of hydrate structures with different water contents supplies an explanation of phase reversal. It may occur if a water-soluble emulsifier hydrate loses so much of its water of hydration that an oil-soluble emulsifier or an oil-soluble emulsifier hydrate is formed. On the other hand, if water is added, phase reversal occurs if an emulsifier dissolved in a nonaqueous phase has formed a sufficiently water-soluble hydrate [69], [70]. In polyglycol ether solutions, phase reversal always occurs at the cloud point [71], [72].

Incomplete phase reversal may also result in oil droplets that are emulsified in water but in turn still have droplets of water in their interior, or conversely, water droplets that still contain oil droplets. In this way, so-called *double* or *multiple emulsions* may be prepared [73], [74] (Fig. 2).

2.10. Rheology [3, p. 405]

The consistency of an emulsion may vary considerably. Thus, in the case of o/w emulsions, the viscosity is usually low and hardly distinguishable from that of a true aqueous solution. On the other hand, in w/o emulsions, viscosity often reaches very high values. The emulsions then have a consistency like butter or ointment, which is often the result of liquid – crystalline gel structures [75], [76].

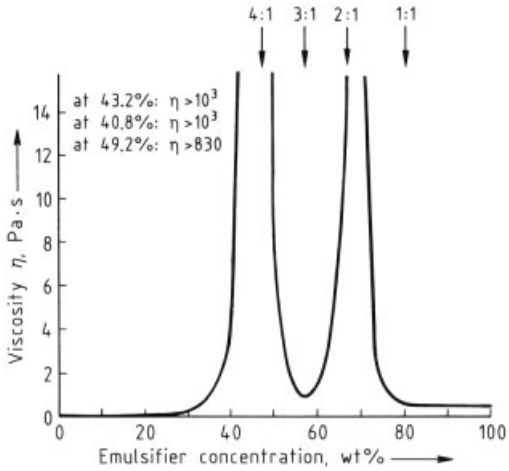

Figure 5. Viscosity of emulsifier – water mixtures at 30 °C (η) vs. emulsifier concentration (emulsifier made by reaction of phenol with styrene and subsequent ethoxylation [51])
Arrows point at ratios of water molecules to oxygen atoms of poly(ethylene oxide) chain.

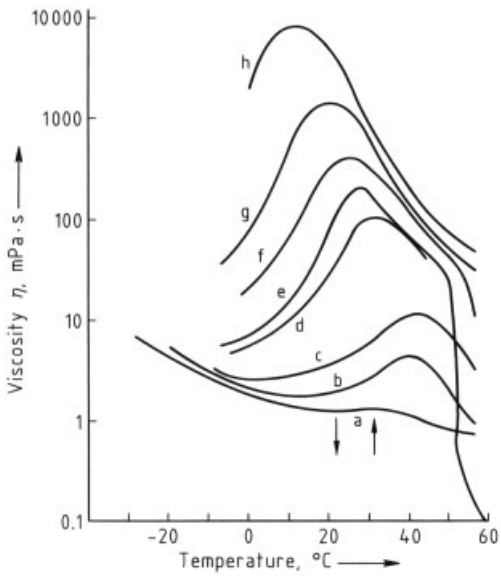

Figure 6. Viscosity vs. temperature at various concentrations of emulsifier OM-7-g (ethoxylated oleic acid amide) Concentrations in wt %: a) 0.5; b) 1.0; c) 2.0; d) 5.0; e) 6.0; f) 10.0; g) 15.0; h) 25.0

Figures 5 and 6 show viscosity abnormalities that may be significant in relation to the rheological behavior of emulsions. In Figure 5, the viscosity of an emulsion at 30 °C is plotted against its concentration. The two high-viscosity gel phases are the results of structures in which two and four water molecules are localized at every oxygen atom in the polymer chain [69].

Figure 6 is a plot of viscosities against temperature for an ethoxylation product of oleic acid amide (OM-7-g). The transition from spherical to rod-shaped micelles is assumed to be responsible for this anomalous behavior [77]. However, another group of authors make the flocculation of lamellae responsible for this phenomenon; at increased temperature, these lamellae obstruct each other [78].

The preparation of high-viscosity emulsifier structures is essential in the manufacture of creams, but such gel structures are troublesome in some industrial processes such as the transportation of petroleum in pipelines [1, p. 471].

3. Basic Processes in the Preparation of Emulsions

Different basic processes are operative in the preparation and stabilization of an emulsion, and therefore, a distinction should be made between them [79].

3.1. Breaking up to Drops

The preparation of emulsions always involves breaking up the internal phase–a process that is initiated by supplying mechanical or chemical energy. A geometrical series can be used as the simplest approximation for describing the size reduction of drops. The number of droplets Z_n after n divisions is then

$$Z_n = 2^n \tag{2}$$

Thus, ca. 2.4×10^{15} droplets with a total surface area of ca. 300 m^2 are produced, for example, in emulsifying 10 cm^3 of olive oil as droplets with a radius of 10^{-5} cm. This requires ca. 50 divisions.

If droplet division takes place according to a geometrical series, a monodisperse system is always approached. The polydisperse systems encountered as a rule in practice are due to the formation of smaller fragments, so-called secondary droplets, produced in addition to the large primary droplets in the course of droplet disintegration.

3.2. Increase in Surface Area

When an emulsion is formed, the interface between the phases is considerably increased as a result of the droplets formation. This is clearly illustrated by the emulsification of olive oil. As the disintegration of the phase to be emulsified continues, the interfaces determine the flow behavior of the emulsion and, consequently, the action of the emulsifiers at the droplets surface acquires a decisive importance.

3.3. Work of Division

The liquid always tends to reduce its surface or interface to a minimum; therefore, an increase in interface is possible only if energy is supplied. The work that must be expended on drop division is

$$dA = \mu dI \tag{3}$$

The work dA to be expended is proportional to the increase dI in interface. The proportionality factor is the interfacial tension μ between the phases to be emulsified. Thus, if the interfacial tension between the two phases is high, considerable mechanical energy is required for emulsification unless an emulsifier is added; if the interfacial tension is low, little mechanical energy is consumed.

3.4. Decreasing the Interfacial Tension

[5, p. 180]

The interfacial tension is decreased by adding emulsifiers. In a given two-phase system, an emulsifier is the more effective as it more rapidly decreases the interfacial tension at the lowest possible concentration. To achieve this, the emulsifier molecules must (1) reach the interface from the phase in which they are dissolved as fast as possible (diffusion effect) and (2) be absorbed at the interface (adsorption effect) to decrease the interfacial tension as much as possible. In Figure 7, the interfacial tension between xylene and water is plotted against time. The emulsifier PV-15 has the more beneficial effect in this case, since it decreases the interfacial tension much more rapidly than the emulsifier NP-10.

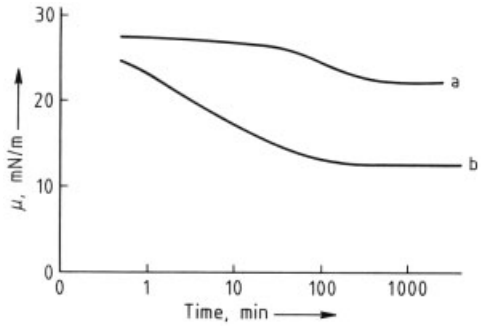

Figure 7. Interfacial tension μ between water and xylene as a function of time. Emulsifier concentration 0.01 g/L [6, p. 180]
a) Emulsifier NP-10; b) Emulsifier PV-15 [51]

Figure 8 is a plot of the interfacial tension between xylene and water after equilibrium has been reached (after 50 min for emulsifier NP-10 and after 12 min for emulsifier PV-15 in Figure 8) against emulsifier concentration. With the better emulsifier PV-15, low values of interfacial tension are achieved even at a concentration of 0.1 g/L. The poorer emulsifier NP-10 requires a concentration of 100 g/L for this purpose [6].

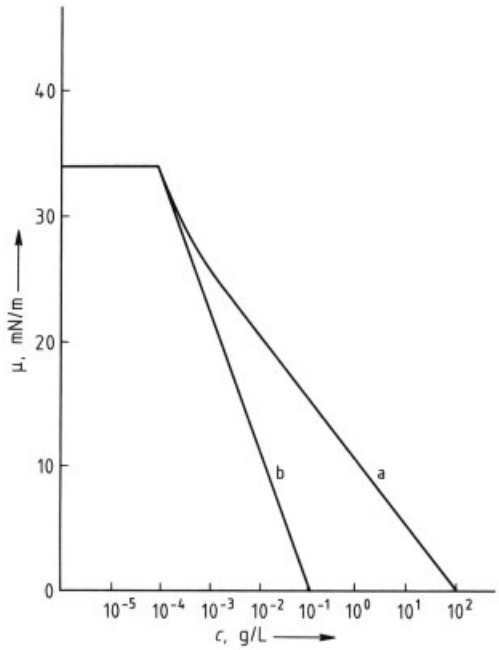

Figure 8. Interfacial tension μ between water and xylene with and without emulsifiers added at 20 °C [6, p. 181]
a) Emulsifier NP-10 dissolved in water; b) Emulsifier PV-15 dissolved in water or in xylene

The decrease of the interfacial tension is due to the accumulation of the emulsifier molecules at the interface between the phases to be emulsified. The area required by an emulsifier molecule in the interfacial film can be calculated from plots of μ vs. log c, as in Figure 8 using the Gibbs equation, which is [70], [81] as follows:

$$\tau = \frac{1}{RT} \cdot \frac{\mathrm{d}\mu}{\mathrm{dln}c} \tag{4}$$

where τ is the interface adsorption or accumulation in mol/cm^2, R the gas constant, and T the absolute temperature.

The area required, A (in m^2), can then be determined by using Avogadros number, N_A, from τ:

$$A = \frac{10^{-4}}{\tau \cdot N_\mathrm{L}} \tag{5}$$

A molecule of emulsifier PV-15 occupies an area of 88×10^{-20} m^2 (88 Å^2) in the interfacial film, whereas a molecule of emulsifier NP-10 occupies an area of 182×10^{-20} m^2 (128 Å^2). Therefore, the molecules of good emulsifiers only require a small area in the interface. The molecules arrange themselves at right angles to the interface and form a stable film between the phases whose structure is such that the hydrophobic hydrocarbon portions are tightly arranged in the oil phase, whereas the polar hydrophilic emulsifier groups project into the water phase. This structure of the interfacial film has been confirmed by measurements of interfacial tension. Different μ values are obtained, depending on whether the interface is approached from the water phase or from the oil phase [82]. From its importance, it appears correct to regard the boundary layer in an emulsifier system as a third phase.

3.5. Other Factors Governing Emulsification

The decrease in interfacial tension between two phases that are to be emulsified is a necessary, but insufficient condition for emulsification. In addition to the diffusion-dependent interface adsorption [49], [83], the following processes occur:

1) Spreading of emulsifier molecules in a freshly formed interface (Marangoni effect)
2) Flow and turbulence effects that occur when a drop is ruptured
3) Life of the droplet itself
4) In the case of polymer emulsifiers or stabilizers, the additional factor of untangling the polymer loops of the polymer molecules adsorbed at the interface [84], [85] (cf. Section 4.2.3).

Marangoni Effect. The spreading of the molecules in a freshly prepared interface is called the Marangoni effect [86–88]. This effect is important in the formation and merging of droplets [89].

If two emulsion droplets are formed from a larger droplet, e.g., by mechanical division or by Brownian motion, the point of rupture is initially free of emulsifier molecules. This causes an interfacial tension gradient to be produced between the occupied droplet surface and the unoccupied droplet surface; this gradient must be stabilized either by adsorption of emulsifier molecules from the solution or from the interior of the droplet at the interface (Fig. 9 A), or by the molecules present in the interface spreading out (Fig. 9 B). This smoothing out takes place more rapidly as the diffusion or spreading rate is larger.

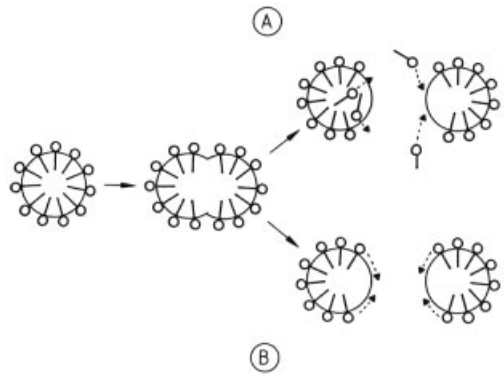

Figure 9. Emulsification mechanisms
A) Interfacial adsorption (diffusion effect); B) Interfacial spreading (Marangoni effect)

The diffusion and spreading rates depend on the steric structure of the emulsifier molecules. Thus, when emulsions are formed, steric effects become important; these effects are described in more detail under → Disperse Systems and Dispersants.

Flow and Turbulence Effects. When a droplet ruptures, flow and turbulence effects occur between the fragments produced; these effects are intensified by the spreading of the molecules already present in the drop surface. These Marangoni flow effects encourage the formation of new droplets and oppose their coalescence.

A Schlieren photograph (Fig. 10) reveals the flow and turbulence effects produced at the toluene – water interface when acetone passes from the toluene phase into the water [90]. When a spontaneous emulsifier passes from xylene into water, in addition to such flow and turbulence effects, large drop deformations also occur that result directly in the disintegration of the drop.

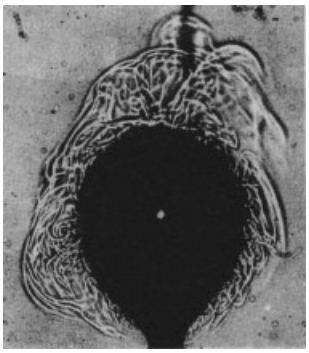

Figure 10. Schlieren photograph of a toluene drop from which acetone is escaping to the surrounding aqueous phase [90]

Life of the Droplet. The life of the droplet not only has a decisive effect on the stability of an emulsion, but it also plays a part in the preparation of an emulsion. If a freshly formed droplet surface is not occupied rapidly enough by emulsifier molecules, they encounter so little resistance while colliding that they coalesce again immediately after being formed. This always occurs if the diffusion, spreading, and absorption rate of the emulsifier molecules are too low. Even if only one of these rates is too slow, adequate emulsification is not achieved.

3.6. Mixed Emulsifiers [91], [92]

If a *nonionic polyglycol ether* is used as an emulsifier, stable o/w emulsions are generally obtained only at a high emulsifier concentration, which ensures the formation of a compact elastic surface film (Fig. 11 A). A reduction in emulsifier concentration can be achieved if two nonionic components that have markedly different HLB values are used [93].

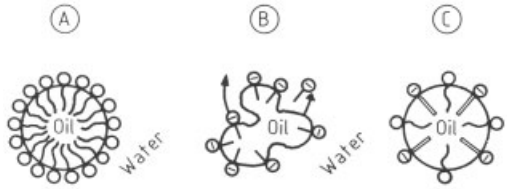

Figure 11. Action of mixed emulsifiers
A) Nonionic emulsifier, high emulsifier concentration; B) Ionic emulsifier, mutual repulsion of the individual emulsifier molecules; C) Mixed emulsifier, low emulsifier concentration necessary, no repulsion

On the other hand, *ionic emulsifiers* alone do not usually result in usable emulsions, since the like charges on the emulsifier ions prevent the formation of a stable surface film (Fig. 11 B). The ions adsorbed at the interface repel each other despite the fact that they are screened by their hydration envelopes.

However, both of these disadvantages are eliminated if a *mixture of a nonionic and anionic emulsifier* is used. In the ideal case, a film is produced at the interface between the phases in which ionic and nonionic emulsifier molecules alternate (Fig. 11 C) [91]. This combines the desirable elasticity of polyglycol ether films with the stabilizing effect of surface films containing like charges. However, since some of the mutual repulsion of like ions is retained in such films despite screening by the polyglycol ether molecules, the surface film is loosened. In this way, the amount of emulsifier required can be markedly decreased.

In practice, a mixture of 40 – 50 wt % of ionic component and 50 – 60 wt % of nonionic polyglycol ether has proved successful as a mixing ratio. A more precise assessment cannot be made since the interface film equivalents of the individual emulsifiers in the film surrounding the emulsion droplets are unknown.

4. Basic Processes in the Stabilization of Emulsions [79], [94]

4.1. Electric Charge

If an emulsion of two phases is formed mechanically, the triboelectric effects produce a measurable potential difference (zeta potential) between the internal and external phases. In most o/w emulsions, the oil droplets are negatively charged.

When emulsifiers are used, the accumulation of surfactant molecules in the droplet surface results in additional charging or in charge reversal. When ionic emulsifiers are used, this change of charge is particularly large. Anionic emulsifiers intensify the negative charge of oil droplets in an o/w emulsion, whereas cationic emulsifiers initially produce a charge reversal. In any event, like charges are finally produced on the surface of the droplets, which opposes any coalescence. In o/w emulsions, the electrical charge on the droplets is generally greater than that in w/o emulsions.

The charging of the drops is shown diagrammatically in Figure 12. In the solution the emulsifier is dissociated (Fig. 12 A) and the droplet surface of the emulsified oil is first charged negatively by the emulsifier anions. Cations are then attracted from the surfactant solution by the negatively charged droplet surface and bound in accordance with their charge. This produces an electrical double layer, called the Helmholtz layer [95] (Fig. 12 B). This extracts further ions from the electrically neutral solution so that a diffuse ion cloud, also called stern layer, is formed [96] (Fig. 12 C). Under these circumstances, the concentration of positive ions decreases as the distance from the phase boundary increases. The potential difference between mobile droplets and solution is symbolized in Figure 12 C by four electrons; it is called the zeta potential, which can be measured, for example, electrophoretically (see page 40) [97, p. 577], [17, p. 383], and [98].

For the mathematical treatment of electrical absorption forces and their interaction with the van der Waals forces of attraction, using the DLVO theory [99–101], see → Disperse Systems and Dispersants.

4.2. Steric Factors [7]

Electrical forces play a subordinate role in relation to stability in nonpolar or weakly polar media. Therefore, other physical parameters, such as density and viscosity of the internal and external pase, become more important. However, steric factors in particular oppose the van der Waals attractive forces. Thus, the steric factors acquire a dominating role in all w/o emulsions, but they may also be important in o/w emulsions.

4.2.1. Molecular Mobility

In the case of emulsifiers for o/w emulsions, the ability of the ethylene oxide chains to rotate freely impedes the approach of two droplets because this free rotation results in the chains becoming interlocked and additional work is required to overcome this. In a similar manner, hydrocarbon chains protect the water droplets in a w/o emulsion from coalescing.

An increase in the length of the hydrocarbon chains and/or the poly(ethylene oxide) chains makes an emulsion more stable. The mobility of the molecules increases with increasing temperature in the case of this "steric effect"; i.e., the disorder is increased and the term *entropy effect* is also used. This effect can be recognized from the increase in stability with rising temperature.

Viscosity Effect. In general, all of the factors that make droplet movements difficult favor the stability of an emulsion. An increase in viscosity, therefore, promotes emulsion stability. Consequently, the selection of a suitable thickening agent may be decisive for achieving the appropriate stability of an emulsion.

4.2.2. Complex Formation

Hydrate Formation. In emulsions, the solvating envelops have a stabilizing action on the droplet surfaces. Thus, the hydrated groups projecting into the water form a more or less continuous hydrate envelope that opposes any approach between droplets. Emulsifiers that are completely and stably hydrated stabilize an o/w emulsion better than those that have only a few

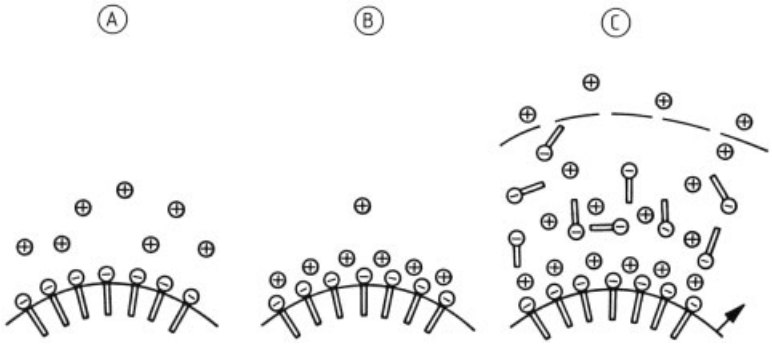

Figure 12. Charging of the emulsion droplets and generation of the zeta potential (diagrammatic)
A) Formation of the surface film; B) Formation of the Helmholtz electric double layer; C) Formation of the Stern diffuse ion layer (potential gradient perpendicular to interface)

localized water molecules. However, the latter are suitable emulsifiers for w/o systems.

Since hydrogen bonds are loosened or broken up as the temperature rises, the emulsion stability decreases with an increase in temperature, and this effect is referred to as an *enthalpy effect.*

Liquid – Crystalline Phases. In emulsions with a high emulsifier content, the stability is increased by the emergence of liquid – crystalline phases [102], [103] in the external phase [104–107]. The increase in viscosity associated with this also contributes to the improvement in stability [52], [53], [108]. Thus, in the system water –*p*-xylene – 4-isononylphenol-9-glycol ether (NP-9), the addition of even a small amount of emulsifier leads to the formation of stable emulsions as a result of the appearance of a liquid –crystalline phase. In the system water – *n*-hexadecane – NP-9, the stability is substantially less because liquid crystals are not observed [109], [110]. Liquid – crystalline phases often occur as gels in the hydration or solubilization of emulsifiers [111]. A different arrangement of the emulsifier molecules may also lead to such liquid – crystalline phases [64].

Microemulsions. The formation of micellar three-dimensional structures at the phase boundary is responsible for the stability of microemulsions [92]. Since the emulsifiers of these often highly viscous systems cannot move freely but are generally fixed crystallographically, they are also termed atypical emulsions, and in many preparations they have lost the true nature of an emulsion [112].

4.2.3. Polymers in Emulsions [6, p. 459], [113–115]

The chemical structure of polymers makes them particularly suitable for stabilizing emulsions. The essential precondition for steric stabilization is adsorption of macromolecules at the droplet surfaces [116], [117]. In addition to chain segments situated parallel to the phase boundary (trains; c in Fig. 13), tails (a) perpendicular to the boundary and loops (b) are formed in the course of such adsorption [118].

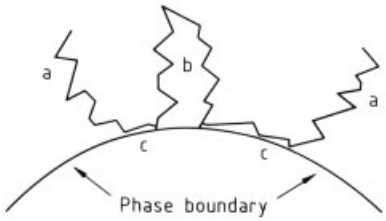

Figure 13. Steric emulsion stabilization by polymers
a) Tails; b) Loops; c) Trains

Figure 13 shows that relatively long tails, in particular, are responsible for steric stabilization. This phenomenon occurs in the following way: if droplets approach each other, the free mobility of the tails is restricted, and the entropy is reduced while the free energy increases. This is opposed by the system and the particles repel each other.

On the other hand, if the interfaces are not completely occupied, the tails of the macromolecules may form bridges between the droplets. This may produce *bridging flocculation,* which is promoted by a high degree of polymerization of the polymers and by a low solubility in the external phases (for more details, see → Disperse Systems and Dispersants).

If two polymer layers penetrate each other, the segment density increases and a stronger interaction is produced between the segments. In this case the free energy generally increases; i.e., the particles repel each other. Only when the polymer is sparingly soluble in the external phase is the effect reversed [119].

The manner in which a polymer behaves in an emulsion determines whether it stabilizes the system or destabilizes it. To achieve adsorption, the interaction between polymer and droplet (oil or water) must be high and that between polymer and solvent must be weak so that no desorption occurs in the event of a collision produced by Brownian motion. On the other hand, the interaction between polymer and solvent must be high for steric stabilization to occur. For these reasons, block or graft polymers are especially suitable as emulsions stabilizers. They have segments that are insoluble in the external phase and adhere to the droplets of the internal phase, as well as tails that are soluble in the external phase and produce the stabilization.

Not only can o/w and w/o emulsions be stabilized by such block and graft polymers, but also o/o and w/w emulsions. Thus, a dimlethylformamide – cyclohexane emulsion can be produced, for example, with a polystyrene – polyisoprene block polymer. A w/w emulsion can be obtained with a copolymer of poly(ethylene oxide) and poly(vinylpyridinium chloride) [114]. The w/w emulsions play a role in biological processes and are acquiring increasing significance in bioengineering and pharmacy, e.g., in vesicle formation.

4.3. Further Stability Factors

A rise in *temperature* as a rule opposes stability [120], entropy stabilization being an exception. The difference in *density* between the two phases should be as low as possible, and account should be taken of this factor in selecting suitable solubilizers. In stable emulsions, the droplets should be as small and uniform as possible. A high *zeta potential* should favor stability [121], and the *HLB values* (see Section 11.5) should be adjusted to optimum values [122]. The magnitude of the *dipole moment* of the oil to be emulsified is inversely proportional to the stability of the o/w emulsion [123].

4.4. Long-Term Stability

To ensure the long-term stability of an emulsion, recrystallization of an emulsifier must be prevented or suppressed. Liquid emulsifiers generally result in emulsions with good long-term stability. Complex formation between the emulsifier and the internal or external phase may, however, result in compounds that crystallize even when liquid emulsifiers are used.

Mixtures have proved valuable when crystalline or liquid – crystalline emulsifiers are used, since the melting point is reduced and the tendency to crystallization is repressed. Moreover, the addition of a small amount of solubilizer often performs a valuable service.

4.5. Emulsion Thickening Agents

The emulsion can also be stabilized by modifying the external phase. Emulsion thickening agents increase the viscosity of the external phase, which slows down the motion of the emulsion droplets and counteracts or even prevents any coagulation of the droplets. Since emulsion thickening agents in small concentrations often produce high viscosities in aqueous systems, polymer chains may be assumed to be "cross-linked" by water molecules.

4.6. Cream Formation

Creams or water-containing ointments contain large quantities of water that often exceed 80 %. They are the result of the formation of a lamellar structure by liquid – crystalline hydrate phases, which can in turn store additional water [76], [124]. Therefore, in the creams two different

types of bound water may be encountered, water of hydration and storage water. In the case of water of hydration, the water molecules are attached to the hydrophilic groups of the emulsifiers, and in the case of the storage water [125], the water molecules are stored more or less loosely between the hydrated lamellae.

5. Preparation of Emulsions

Emulsions can be prepared mechanically by using stirrers or emulsifying machines, or chemically by adding emulsifiers.

5.1. Theories Relating to Mechanical Disintegration

In mechanical emulsification, the interface between the two liquid phases is first deformed to such an extent that drops are produced. These drops are usually too large, and must then be broken up into smaller ones.

The mechanical deformation of drops in laminar shear flow was studied by LAMB and TAYLOR as early as 1932 [126]. There are now theoretical hypotheses and experimental studies relating to the disintegration of particles in turbulent flows and in shearing fields [3, p. 57], [127]. The difficulty in arriving at fundamental results is indicated by the work of KOGLIN, PAWLOWSKI, and SCHNÖRING [128].

5.2. Emulsifying Machines [129–131]

In mechanical emulsification, the work necessary to reduce the liquid to droplets and the associated increase in interface area is usually supplied to the system by frictional forces. They are produced by the liquids flowing past each other (shearing) and increase as the relative flow velocities and the viscosity of the system are higher. Contact with solid parts of the equipment reinforces the frictional forces. The following devices are available for mechanically preparing industrial emulsions:

1) High-speed stirrers
2) Agitation or impact machines
3) Emulsifier centrifuges
4) Colloid mills
5) Metering pumps (atomizers)
6) Vibrators
7) Ultrasonic generators
8) Homogenizers

The emulsifying machines under 1 – 4 all seek an ideal mixture of the phases to be dispersed. At the same time, excessively intensive shearing can result in breaking of the emulsions. In the case of *metering pumps*, emulsion formation only occurs above a critical injection rate when injection is carried out with or without pressure [93], [132]; the mean droplet size depends on the injection rate, the size of the nozzle orifice, the interfacial tension, and the viscosities of the two liquids. In *vibrators* the necessary frictional forces are produced by a considerable acceleration of the movements in the liquid that is achieved by cyclic oscillations [133]. Other devices make use of plates attached to rods, which are caused to vibrate by a.c. solenoids or by eccentric motion [13]. In *ultrasonic generators* the emulsifying action of ultrasound is attributed to the cavitation forces generated in the liquid. *Homogenizers* are intended to reduce the size of coarse polydisperse emulsions. To do this, the coarse emulsion is usually forced through narrow jets to minimize the effects of large differences in droplet size. Repeated homogenization initially improves the droplet distribution, but soon leads to a limiting value of the droplet size of ca. 1 – 2 μm [134], [135].

It is often necessary to prepare a preliminary solution of the liquid to be emulsified in a solvent that is miscible with the external phase (e.g., alcohol). When this solution is then stirred into the external phase (e.g., water) or injected through jets, very fine droplets are usually produced with a fairly uniform droplet size, which depends on the concentration of the internal phase in the auxiliary solvent. The general disadvantage of purely mechanical emulsification is that only emulsions of low concentration (0.1 %) can be prepared. To reduce the work involved in emulsification, small quantities of emulsifiers are usually added in most mechanical processes.

Table 1. Nonionic emulsifiers

ID*	Name	Structure	BD**
N 100	**fatty acid esters of**		
N 101	alcohols	~~~~~~ COO ~~~~~~	+
N 102	ethylene glycol	~~~~~~ $COOCH_2CH_2OH$	+
N 103	poly(ethylene glycol)	~~~~~~ $COO(CH_2CH_2O)_nH$	+
N 104	poly(propylene glycol)	~~~~~~ $COOCH(CH_3)CH_2OH$	+
N 105	glycerol, polyglycerol	~~~~~~ $COOCH_2CH(OH)CH_2OH$	+
N 106	sorbitol, e.g., "Span"	~~~~~~ $COOCH_2(CHOH)_4CH_2OH$	+
N 107	pentaerythritol	~~~~~~ $COOCH_2C(CH_2OH)_3$	+
N 108	glycerol esters (R = acetic, lactic, tartaric, and citric acid)	~~~~~~ $COOCH_2$ $HOCH$ $ROCH_2$	+
N 109	saccharose	~~~~~~ $COOC_{12}H_{21}O_{10}$	+
N 200	**fatty amines and fatty acid amides**		
N 201	fatty amines	~~~~~~ NH_2	+
N 202	fatty acid amides	~~~~~~ CON (~~~~)$_2$	+
N 203	fatty acid alkanolamides	~~~~~~ CONH ~~~~ OH	+
N 204	condensed amines	~~~~~~ CONH ~~~ NH ~~~ NHCO–	+
N 205	polyamines	~~~~~~ $NH(CH_2CH_2NH)_n$	+
N 300	**polyglycol ethers of** (PEO = poly(ethylene oxide, PPO = poly(propylene oxide)		
N 301	alcohols, e.g., "Peregal"	~~~~~~ OH · PEO	(+)
N 302	thioalcohols	~~~~~~ SH · PEO	+
N 303	fatty acids (ethers and esters), e.g., "Emulfor"	~~~~~~ COOH · PEO	+
N 304	fatty acid esters, such as glycerides or sorbitol esters, e.g., "Tween"	~~~~~~ $COOCH_2CHCH_2OH$ · PEO OH	(+)
N 305	fatty amines, e.g., "Genamine"	~~~~~~ NH_2 · PEO ~~~~ NH_2 · $(PEO)_2$	(+)
N 306	fatty acid alkanolamides	~~~~~~ $CONHCH_2CH_2OH$ · PEO	(+)
N 307	fats and oils (transesterified, intercalated)	~~~~~~ $COOCH_2$ ~~~~~~ COOCH · PEO ~~~~~~ $COOCH_2$	(+)
N 308	poly(propylene glycol), e.g., "Pluronic"	PPO · PEO	–
N 309	ethylenediamine poly(propylene glycols), e.g., "Tetronic"	PPO · NH_2CH_2 PEO · PPO · NH_2CH_2	–
N 310	alkylphenols, e.g., "Igepal C"	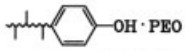	–
N 311	condensed phenols	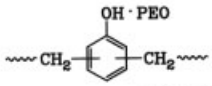	–
N 312	abietic acid	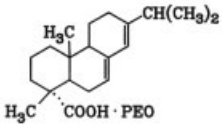	(+)
N 313	condensed amines	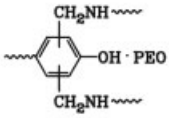	(+)

Table 1. Continued

N 314	lanolin		(+)
N 315	beeswax		(+)
N 316	cationic surfactants (cf. K100)		−(+)
N 317	phosphoric acid derivatives (cf. A 400)		+
N 318	polysiloxanes		−
N 400	**poly(propylene glycol) ethers of**		
N 401	alcohols	~~~~~~OH · PPO	(+)
N 402	fatty acids	~~~~~~COOH · PPO	(+)
N 403	fatty amines	~~~~~~NH_2 · PPO ~~~~NH_2 · $(PPO)_2$	(+)
N 404	lanolin		(+)
N 405	cationic surfactants (cf. K100)		(−)(+)

∗ Identification of compound, as in Chapter 10.
∗∗ Biologic degradability.

5.3. Use of Emulsifiers

5.3.1. General Rules

In the vast majority of cases, emulsions are prepared with the aid of one or more emulsifiers, usually with different HLB values. The emulsifier is either dissolved or pre-emulsified in the oil or water phase, and this solution is then worked into the external phase, or the external phase is slowly added to the emulsifier – oil mixture. Phase reversal may be used in this process. The following rules may produce the desired result:

Hydrophilic emulsifiers with a high HLB value favor the formation of o/w emulsions. Thus, hydrophilic alkali-metal soaps are used more often to obtain o/w emulsions than are the sparingly soluble alkaline-earth-metal soaps, which are more suitable emulsifiers for the w/o type. If the hydrophilic emulsifier is sufficiently soluble in the oil phase, the emulsifier should be dissolved in oil and both should be stirred into water. These emulsifiers diffuse rapidly to the interface. The high solubility in water facilitates the insertion of the hydrophilic groups into the water and the formation of a stable interfacial film.

Hydrophobic emulsifiers with a low HLB value are generally more suitable for preparing w/o emulsions. They are accordingly sparingly soluble in water and diffuse rapidly through the interface if they are dissolved in water. Their high solubility in oil results in rapid insertion of the hydrophobic parts of the molecules into the oil phase.

However, in practice the water solubility of oil emulsifiers or the oil solubility of water emulsifiers is often too low. As a result, compromise solutions must often be used. Preparation of the emulsifier may be expedient only at the mixing stage, since its formation results in stable films at the interface. When w/o emulsions are prepared with the aid of aluminum soaps, this method gives outstanding results. For this purpose, the aluminum salt is dissolved in the aqueous phase and the fatty acid in the oil. When the two phases are mixed, the aluminum soap forms at the interface where it is required as an emulsifier.

5.3.2. Preparation of Multiple Emulsions

Multiple emulsions are prepared by first producing a w/o emulsions, using a hydrophobic emulsifier such as sorbitan monooleate, and then mixing this emulsion with an aqueous solution of a hydrophilic emulsifier such as polyethoxylated sorbitan monooleate [136]. An $o_1/w/o_2$ emulsion can similarly be prepared by mixing an o/w emulsion with an oil solution of a hydrophobic emulsifier. A multiple emulsion may also be produced by incomplete phase reversal [137], [138].

5.3.3. Preparation of Microemulsions

As a rule, microemulsions consist of four components: (1) the substance to be emulsified, (2)

Table 2. Anionic emulsifiers

ID*	Name	Structure	BD**
A 100	**carboxylates** (A = ammonium salt; AA = alkanolammonium salt)		
A 101	soaps	~~~~~~~~ COONa (also K, Li, NH_4, A, AA, Morpholin)	+
A 101/S	sparingly soluble metal soaps	Ca, Al, Mg, Zn, and Pb soaps	+
A 102	fluorinated carboxylic acids	~~~~F~~~~ COONa	–
A 103	alkoxycarboxylates	~~~~~~~~ O ~~~~~~ COONa	+
A 104	sarkosides, e.g., "Medialan"	~~~~~~~~ $CON(CH_3)CH_2COONa$	+
A 105	sulfosoaps (sulfonamidecarboxylic acids)	~~~~~~~~ SO_2NHCH_2COONa	+
A 106	lactylates	~~~~~~~~ $CO \cdot OCH(CH_3)COONa$	+
A 107	salts of alkylmalonic and alkylsuccinic acid	~~~~~~~~ CHCOONa (COONa) ~~~~~~~~ CHCOONa (CH_2COONa)	+
A 108	disproportionated abietic acids, e.g., "Dresinates"		+
A 109	naphthenates (also as sulfonates, e.g., "Petrowet R")		+
A 110	protein–fatty acid condensation products, e.g., "Lampeon"	~~~~~~~~ $(CONHCH)_nCOONa$ (R)	+
A 200	**sulfonates**		
A 201	alkylsulfonates, e.g., "Mersolat" (K and NH_4 salts)	~~~~~~~~ SO_3Na ~~~~~~~~ SO_3Na	+
A 201/S	sparingly soluble alkylsulfonates	Ca, Ba salts	+
A 202	sulfonated esters, e.g., "Igepon A"	~~~~~~~~ $COOCH_2CH_2SO_3Na$	+
A 203	fatty acid sulfonates	~~~~~~~~ CHCOONa (SO_3Na)	+
A 204	fatty acid ester sulfonates (R = alcohol)	~~~~~~~~ CHCOO R (SO_3Na)	+
A 205	sarkoside	~~~~~~~~ $CONCH_2SO_3Na$ (CH_3)	
	tauride	~~~~~~~~ $CONHCH_2CH_2SO_3Na$	
	methyltauride, e.g., "Igepon T"	~~~~~~~~ $CONCH_2CH_2SO_3Na$ (CH_3)	
A 206	sulfosuccinic acid esters (of alcohols "Aerosol OT," and alkanesulfonamides, e.g., "Erkantol RN")	~~~~~~~~ $OOCCHCH_2COO$ ~~~~~~~~ (SO_3Na)	+
		~~ $SO_2NH-OCCHCH_2COO$ ~~~~~~~~ (SO_3Na)	+
A 207	perfluorinated alkylsulfonates	~~~~F~~~~ SO_3Na	–
A 208	alkylbenzenesulfonates, e.g., "Marlon" (also K, NH_4, Li, A, and AA salts)	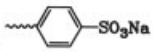	+
A 208/S	sparingly soluble alkylbenzenesulfonates	Ca, Mg salts	(+)
A 208/K	short-chain alkylbenzenesulfonates, e.g., of cumene, toluene, or xylenol		
A 209	alkylnaphthalenesulfonates, e.g., "Nekal BX"	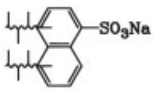	–

Table 2. Continued

A 210	condensed sulfonates	$[C_6H_4(SO_3Na)-CH_2]_n$	–
A 211	petroleum sulfonate, also some carboxylates (naphthenates)		
A 212	alkylbenzimidazoles, e.g., "Ultravon"		
A 300	**sulfates**		
A 301	sulfated primary and secondary alcohols (also Li, K, NH_4, A, AA, Mg salts)	$\sim\sim OSO_3Na$ $\quad$ $\sim\sim CH(OSO_3Na)\sim\sim$	+
A 302	sulfated soaps	$\sim\sim CH(OSO_3Na)-COONa$	+
A 303	sulfated esters (fats and oils), e.g., "Monopolseife," "Turkey Red Oil"	$\sim\sim CH(OSO_3Na)-COO\sim\sim$	+
A 304	sulfated fatty acid amides	$\sim\sim CH(OSO_3Na)-CONH\sim\sim$	+
A 305	sulfated alkanolamides	$\sim\sim CONH\sim\sim OSO_3Na$	+
A 306	sulfated mono- and polyglycerides	$\sim\sim COOCH_2-HOCH-NaO_3SOCH_2$	+
A 307	sulfated polyglycol ethers, e.g., of fatty alcohols or alkylphenols (also Na, K, A, Mg salts)	$\sim\sim OH \cdot PEO \cdot OSO_3NH_4$ $C_6H_4-OH \cdot PEO \cdot OSO_3NH_4$	+ –
A 400	**organic compounds of phosphoric acids**		
A 401	phosphates of alcohols, phenols, or their polyglycol ethers, e.g., "Hostaphat"	$\sim\sim OPO(OH)_2$	+
A 402	phosphoric acid esters of glycols and glycerides	$\sim\sim COOCH_2-HOCH-(OH)_2OPOCH_2$	+
A 403	esters of phosphorous acid	$\sim\sim OP(OH)_2$	+
A 404	phosphonates and phosphinates	$\sim\sim PO(OH)_2$ $\quad$ $\sim\sim P(OH)_2$	+
A 405	polyphosphates, such as trisodium phosphate, tetrasodium diphosphate, or sodium tripolyphosphate		eutro
A 406	complex phosphates	$\sim\sim O-P(=O)(OR)-OCH_2CH_2OP(=O)(OH)_2$	eutro
A 407	alkyl and aryldithiophosphates	$\sim\sim OPO(SH)_2$	+

∗ Identification of compound, as in Chapter 10.
∗∗ Biologic degradability.

water, (3) the emulsifier or the emulsifier mixture, and (4) a solubilizer [1], [24], [139].

For example, a cloudy emulsion of benzene in an aqueous potassium oleate solution becomes clear when alcohol is added. If short-chain alcohols are used, an o/w emulsion is produced, while alcohols with a chain length of more than six carbon atoms result in the formation of w/o emulsions [140].

Microemulsions containing 5 – 10 % of a suitable solubilizer can be readily prepared, as long as certain conditions are met [141]. If non-

Table 3. Cationic emulsifiers

ID*	Name	Structure	BD**
K 101	fatty amine salts (primary, secondary, tertiary), e.g., "Armac 12 to 18"	$\sim\!\sim NH_3^+ \cdot CH_3COO^-$	+
K 102	salts of alkylene diamines and polyamines	$\sim\!\sim(NH\sim\!\sim)_n NH_3^+\ Cl^-$	+
K 103	quarternary alkyl ammonium salts, e.g., "Hyamin"	$[\sim\!\sim N^+(CH_3)_2\sim\!\sim]\ Cl^-$	(−)
K 104	alkylbenzyl ammonium salts, e.g., "Zephirol"	$[\sim\!\sim N^+(CH_3)_2-CH_2-C_6H_5]\ Cl^-$	(−)
K 105	quaternary amidoamino compounds	$[\sim\!\sim CONH\sim\!\sim \overset{+}{N}H_3]\ Cl^-$	(−)
K 106	alkanolammonium salts of ethers and esters	$[(OH\sim\!\sim)_2\overset{+}{N}H\sim\!\sim OOC\sim\!\sim]\ Cl^-$ $[OH\sim\!\sim\overset{+}{N}H(\sim\!\sim O\sim\!\sim)_2]\ Cl^-$	(−)
K 107	alkylpyridinium salts	$[\sim\!\sim\overset{+}{N}C_5H_5]\ Cl^-$	(−)
K 108	alkylimidazolinium salts	$[C_3H_2N_2]\ Cl^-$	(−)
K 109	alkyloxazolinium salts	$[C_3H_2NO]\ Cl^-$	(−)
K 110	aminoxides, e.g., "Ammonyx"	$\sim\!\sim N\rightarrow O$	(+)
K 111	perfluorinated cationic compounds		(−)

∗ Identification of compound, as in Chapter 10.
∗∗ Biologic degradability.

Table 4. Ampholyte emulsifiers

ID*	Name	Structure	BD**
M 101	protein ampholytes, e.g., acyl derivatives of weakly degraded protein		+
M 102	substituted betaines and sulfobetaines, e.g., "Tego Betains"	$\sim\!\sim CONH-N^+(CH_3)_2-CH_2COO^-$	+
M 103	substituted aminocarboxylic, aminosulfuric, and aminophosphoric acids	$\sim\!\sim CH(NH_2)\sim\!\sim COOH$	+
M 104	ampholytes with cyclic structure	$[\sim\!\sim C(=N-C(H_2)-CH_2)-N^+(CH_2CH_2OH)-CH_2COONa]\ OH^-$	+

∗ Identification of compound, as in Chapter 10.
∗∗ Biologic degradability.

ionic surfactants are used, the solubilizer can be deleted [142].

A microemulsion is believed to consist of very small droplets surrounded by a monomolecular layer of alternating molecules of the emulsifier and the solubilizer [140]. NIKITINA detected three-dimensional micellar structures in microemulsions [92], and ADAMSON conceived microemulsions as expanded micelles [143]. The thermodynamic stability of microemulsions is demonstrated by their representation in phase diagrams [144], [145]. In the preparation of microemulsions, cloudy transitional states with marked birefringence occur and various stages of hydration have been detected in such microemulsions [146].

5.3.4. Spontaneous Emulsification

For the purpose of spontaneous emulsification, emulsion concentrates are prepared that emulsify automatically when they are introduced into water. The movement during mixing is sufficient to produce the emulsion, but occasional stirring is often advisable. The emulsion concentrates usually contain suitable combinations of emulsifiers [147], which reduce the interfacial tension between the phases to be emulsified virtually to 0. The addition of solubilizers such as, e.g., alcohols [67], [148], and/or xylenes is often unavoidable, since spontaneous emulsification takes place more readily as the components to be emulsified are more soluble in each other. In addition, the solvents, which are present often only in small quantities, reduce the viscosity of the interfacial film. Since microemulsions are often produced in spontaneous emulsification, pre-emulsification (by adding a little water to the emulsion concentrates followed by phase reversal) often produces the desired result.

5.3.5. Inorganic Emulsifiers

The emulsifying action of inorganic emulsifiers depends largely on the specific surface area of the solids used. If high-efficiency mixing equipment is employed (e.g., a high-speed stirrer), the solid emulsifier can even be added slowly during the mixing operation [16, Vol. I/2], [79]. Fine grinding is consequently necessary. The resultant powders may be introduced into one of the two phases, into both phases, or into the interface between the phases. Before use, they may be wetted with water or oil.

6. Breaking of Emulsions

Breaking is the term applied to the irreversible coalescense of the individual emulsion droplets to form a continuous phase [5, p. 189], [6, p. 453], [16, Vol. III, p. 213], [62], [149]. To break an emulsion, the stabilizing forces must be weakened or eliminated, and the forces of attraction and the bonding mechanisms between emulsion droplets must be reinforced; see also → Disperse Systems and Dispersants.

Temperature Increase. Breaking of an emulsion is usually favored by an increase in temperature. As the temperature rises, the kinetic energy of the individual droplets is increased. This facilitates coalescence during collisions since the protective envelopes of the droplets are more easily penetrated under these conditions. If an emulsion is stabilized sterically, a rise in temperature may favor stability (see Section 4.2.3.).

Addition of Electrolytes. The addition of electrolytes often causes an emulsion to break, since the electric charge of a boundary layer is very sensitive to the addition of foreign ions. The effectiveness of the added electrolyte increases with its valency. Thus, aluminum or iron sulfate is more effective than sodium sulfate, and the latter is in turn more effective than sodium chloride (Shulze – Hardy rule [14, p. 475]).

Electric Fields. Particular use is made of the discharge of the interfacial film at oppositely charged electrodes for breaking w/o emulsions of crude oil.

Removal or Destruction of Emulsifier. To reduce the stability of the interfacial film, substances are added to the emulsion that displace the film-forming components from the interface. The number of emulsion-breaking substances listed in the literature almost equals the number of emulsifiers, and includes additives that destroy the emulsifier directly. Acids or bases,

for example, destroy hydrolysis-sensitive emulsifiers. The addition of a cationic compound to an emulsion stabilized with an anionic emulsifier usually results in rapid breaking, since sparingly soluble complexes, which precipitate from the system, are produced [150].

Occasionally, adsorption of the emulsifiers on solids is also possible. All the well-known adsorption agents, such as activated carbon, silica gel, and fuller's earth, are suitable for this purpose. This method does not produce any appreciable loss in either phase. Adsorption breaking is particularly suitable for purifying condensation water, e.g., in the removal of trace amounts of oil from the cooling water of power plants.

An emulsion can also be broken by producing a small amount of a voluminous precipitate in the aqueous phase. Iron and aluminum oxide hydrates are particularly suitable for this purpose. For example, iron oxide hydrate may also be produced continuously at the anodes of electrolysis equipment.

Mechanical Stress. An emulsion is often broken by shaking, by impact, or by the action of pressure. Thus, the protein film of the fat droplets in milk is destroyed mechanically in the making of butter. At the same time, foam formation removes some of the proteins from the milk. In road construction, the breaking of the bitumen emulsion is promoted by the application of pressure.

Dilution, Evaporation, and Freezing. Diluting the external phase or concentrating it by evaporation also usually breaks the emulsion. Complete evaporation of the external phase leaves a residue consisting of the internal phase and the emulsifier. If the emulsifiers are thermally stable, the emulsion is reconstituted when the external phase is added again. A porous material may also be used to remove a phase from an emulsion system, as in the seepage of the water from a bitumen emulsion to the gravel bed.

Interaction of Several Factors. To break a stable emulsion, several factors usually must be exploited simultaneously. For example, to break natural rubber latex, the stabilizing protein film is destroyed by the action of heat, addition of acid or alkali, enzymatic degradation, and possibly by addition of fungi or bacteria. In the context of wastewater purification, extensive references to the cracking of industrial emulsions in the detergent, metal-processing, and oil industries, as well as in the food industry and related fields, are given in [151].

7. Chemical Structure of Emulsifiers

Any attempt to classify emulsifiers is made difficult by the fact that the active substances belong to chemically very different categories [6, p. 455]. A helpful distinction is made between surface-active emulsifiers (Table 1, 2, 3, 4), natural substances that are only partially surface-active (Table 5), and inorganic emulsifiers, which generally have low surface activity (Table 6).

Tables 1, 2, 3, 4, 5, 6 summarize the common emulsifiers. In addition to the chemical names, a trade name that has become established throughout the world for an emulsifier type is occasionally cited in quotation marks.

Table 5. Natural emulsifiers

ID *	Name	BD **
T 101	lanolin	+
T 102	lecithins and cholesterols, e.g., egg yolk	+
T 103	waxes, e.g., beeswax	–
T 104	ligninsulfonates, e.g., "Waldhoflauge" sulfite waste liquor	–
T 105	cellulose and cellulose derivatives, e.g., cellulose ethers	–
T 106	proteins, e.g., gelatin and casein	+
T 107	polysaccharides, e.g., starch, dextrin, and pectin	+
T 108	types of rubber, e.g., natural rubber latex, gum arabic, mastics, and tragacanth	+
T 109	mucilages from algae and lichen (alginates), e.g., Carragheenan, agar – agar	
T 110	saponins	+
T 111	locust bean flower, e.g., "Diagum S"	+

* Identification of compound, as in Chapter 10.
** Biologic degradability.

Table 6. Inorganic emulsifiers

ID *	Name	BD **
R 101	fuller's earth and bentonite	–
R 102	oxide hydrates, e.g., $Fe(OH)_3$, $Cr(OH)_3$	–
R 103	aluminum salts	–
R 104	heavy-metal salts, e.g., zirconium salts	–
R 105	activated carbon	–
R 106	water glass	–

* Identification of compound, as in Chapter 10.
** Biologic degradability.

In the tables the first letter of the identification (ID) classifies emulsifiers as follows:

N	nonionic	surface-active emulsifiers
A	anionic	surface-active emulsifiers
K	cationic	surface-active emulsifiers
M	ampholytic	surface-active emulsifiers
T	natural products	partially surface active
R	inorganic emulsifiers	surface activity generally low

In the structural formulas, the long wiggly lines denote a straight hydrocarbon chain. In most cases, branched chains may also be used. If only derivatives with branched hydrocarbon chains are available commercially, the wiggly lines bear vertical bonds (e.g., last entry in Table 2). If an F appears over the wiggly line (e.g., third entry in Table 2), the chain is perfluorinated. Since only the straight chains are biologically degradable, such products are increasingly important [156], [157].

The information on biologic degradability (BD) relates to compounds with straight hydrocarbon chains [158]:

+	More than 85 % biologically degradable
(+)	Partially degradable; generally the straight hydrocarbon chain is degraded, leaving behind, for example, the polyglycol ether groups, which have virtually no surface activity.
−	Not degradable
(−)	Only degradable in dilute solutions; above a certain threshold level, the mostly cationic emulsifiers have a bacteriostatic action and inhibit any degradation process
(eutro)	Accelerates the growth of algae in rivers and lakes so strongly that the biological equilibrium is disturbed

7.1. Surface-Active Emulsifiers

All surface-active emulsifiers molecules consist of a section that is hydrophobic (lipophilic) and sparingly soluble in the aqueous phase and another section that is hydrophilic (lipophobic) and highly soluble in the aqueous phase [16], [19], [20], [33], [152].

The hydrophobic section of the molecule is always a nonpolar aliphatic and/or aromatic hydrocarbon group, whereas the hydrophilic section contains either an ionic group or an accumulation of OH groups (saccharose derivatives) or of poly(alkylene oxide) groups, e.g., $(-CH_2-CH_2-O-)_n$.

The latter substances lack polar ionic groups and, therefore, are described as nonionic surfactants (Table 1). In the case of polar surfactants, a distinction is between anionic (Table 2) and cationic surfactants (Table 3), depending on whether the main entity that is left behind after dissociation is negatively or positively charged. The term "ampholyte" is applied to an ionic emulsifier that forms zwitterions (Table 4). These emulsifiers may act either as acids or as bases, depending on the pH [108]. All classes of surfactants find extensive use as emulsifiers.

7.2. Natural Products

Most natural products or their derivatives, such as cellulose ether, carboxylic acids, or ligninsulfonate, are emulsion promoters and have little or no surface activity (Table 5). They may be added to the emulsions as protective colloids [153] or to increase the viscosity and, consequently, the stability. Occasionally their object is also to give the emulsion certain mechanical, optical, or electrical properties. Thus, the lecithin and cholesterol content of egg yolk imparts the desired mechanical strength to mayonnaise prepared from it. In contrast to most surface-active synthetic emulsifiers, the natural substances are generally sensitive to changes in temperature and pH [16, p. 126].

7.3. Inorganic Emulsifiers

The chief inorganic emulsifiers are bentonite and fuller's earths, but china clay, activated carbon, or finely ground metal dust may also be used (see also Section 5.3.5). Freshly precipitated basic salts of aluminum, calcium, copper, nickel, or zinc [154], [155] and partially flocculated sols of $Fe(OH)_3$ or $Cr(OH)_3$ have proved very successful as o/w emulsifiers. Solid emulsifiers having a more hydrophobic surface, such as activated carbon, tend to form w/o emulsions. Oxide hydrates of aluminum or suitable heavy-metal salts such as those of zirconium are constituents of emulsions used as waterproofing agents (Table 6).

7.4. Sources of Supply

Emulsifiers are produced by many chemical factories, some of which are listed below (from [15, Vol. 2, p. 197]).

Aceto Chemicals, Flushing, United States
Akzo, Düren, F. R. of Germany
Albright & Drieson, Whitehaven, Great Britain
Armoir, Chicago, United States
Atlas Chemical, Wilmington, United States
BASF, Ludwigshafen, F.R. of Germany
Bayer, Leverkusen, F.R. of Germany
Beacon Chemicals, Cambridge, United States
C. Becker, Hamburg, F.R. of Germany
Carbide & Carbon Chem., New York, United States
Chem. Werke Hüls, Marl, F.R. of Germany
Chem. Y. Bodegrawn, Netherlands
Ciba–Geigy, Basel, Switzerland
Colgate, Jersey City, United States
Comp. Franc. Prod. Industriels, Asnières-sur-Seine, France
Croda, Yorkshire, Great Britain
Dow Chemical, Midland, United States
Drew, Booton, United States
Du Pont, Wilmington, United States
Dutton & Reinisch, London, Great Britain
Dynamit-Nobel, Köln, F.R. of Germany
Eastman Kodak, Rochester, United States
Emery, Mauldin, United States
Emulsion A/S, Palsgaad, Denmark
Emulsol Chem., Chicago, United States
Gilnow, Hamburg, F.R. of Germany
Goldschmidt, Essen, F.R. of Germany
Grünau, Illertissen, F.R. of Germany
Heftl, Zürich, Switzerland
Henkel, Düsseldorf, F.R. of Germany
Hercules Powder, Wilmington, United States
Hoechst, Frankfurt, F.R. of Germany
ICI, Manchester, Great Britain
Lion Fat, Tokyo, Japan
Marabishi, Osaka, Japan
National Anilin, New York, United States
Nihon Surfactant, Tokyo, Japan
Ninol, Chicago, United States
Nopco, Moristown, United States
Onyx Chemical Co., Jersey City, United States
Rewo, Steinau, F.R. of Germany
Rohm & Haas, Philadelphia, United States
Shell, London, Great Britain
Stockhausen, Krefeld, F.R. of Germany
Sudfeld, Melle, F.R. of Germany
Tensia, Liège, Belgium
Ugine Kuhlmann, Levallois, France
Unilever, Buurje, Netherlands
VEB Fettchemie, Karl-Marx-Stadt, German Democratic Republic
Witco, London, Great Britain
Zschimmer & Schwarz, Lahnstein, F. R. of Germany

A comprehensive list of manufacturing companies and their addresses can be found in [159]. *"McCutcheon's Functional Materials"* lists commercial products for various fields of application. The 1979 issue does not include products made by European companies. Further sources of supply for emulsifiers are given in [5, pp. 771 –977], [160], [161].

8. Chemical Structure of Emulsion Promoters [6, p. 459]

8.1. Polymers

Polymers of all types are now being used increasingly to stabilize emulsions (Table 7). A number of polymer compounds, for example, certain polyurethanes, are inherently dispersible without requiring any other additive. The interfacial activity of other polymers must be increased by adding emulsifiers and/or emulsion promoters [115, Vol. I, p. 55]. Copolymers are particularly suitable for preparing emulsions and suspensions because the mixing ratio of hydrophilic and hydrophobic components can be matched to almost any particular application [114]. The action of polymers in stabilizing emulsions is described in Section 4.2.3. If a mixture consisting of an emulsifier of low molecular mass and a highly polymerized stabilizer is used, the emulsifier molecules may be adsorbed on the polymer chains and transported by the latter to the interface [162].

Table 7 provides an overview of polymers that are used as emulsion promoters, for example, as stabilizers, as emulsion thickening agents, or as protective colloids.

8.2. Special Emulsion Promoters

In many cases other agents are necessary for preparing, stabilizing, and preserving emul-

sions. Typically, organic solvents are used as *solubilizers.* For example, the spontaneous emulsification of concentrated emulsions of crop protection agents usually requires the addition of a small quantity of an aliphatic alcohol as a solubilizer. The addition of urea and its derivatives, particularly to emulsions of textile or dyeing auxiliaries, is widespread; urea is generally considered to be a solubilizer.

For the *thickening of emulsions*, such natural products as alginates or locust bean gum are used, as well as large quantities of synthetic polymers and copolymers, e.g., styrene and acrylates or butadiene and acrylonitrile [163].

Protective colloids are occasionally added to counteract the coalescence of emulsions once they have been formed. These colloids surround the formed droplets, but unlike the emulsifiers, they do not invade the internal phase. They are added to the external phase, in which they are usually quite soluble, and are intended to keep the emulsifier drops in suspension once they have been formed. Thus, in the copolymerization of styrene and divinyl benzene for preparing ion-exchange resins, poly(vinyl alcohol) is used as an emulsifier and a cellulose ether is used as a protective colloid. Household detergents generally contain carboxymethyl cellulose as a protective colloid to prevent the dirt particles from being readsorbed on the substrate after they have been detached (graying inhibitors).

Biocides. Emulsions are readily attacked and destroyed by bacteria and fungi [164], and the use of biologically degradable emulsifiers promotes such destruction. For this reason, preservatives and disinfectants must be added to the emulsions [165]. Often the use of cationic emulsifiers suffices, since these generally have an inherent bactericidal or bateriostatic action. Aldehydes, such as formaldehyde or chloral hydrate, or aldehyde precursors are used as added preservatives. Phenol derivatives, such as 2-phenylphenol, 4-chloro-3-cresol or, in cases where it is physiologically safe, pentachlorophenol or its fatty acid esters, are also used. Finally, heterocyclic compounds such as 2-chloroanilinodichlorotriazine or derivatives of 1,2-benzoisothiazolone have been introduced [166]. Table 8 lists common special emulsion promoters.

Table 7. Polymeric emulsifiers

ID*	Name, Structure	BD**
P 100	**polymers**	
P 101	polyethylene compounds, e.g., polycyclobutadiene	–
P 102	poly(vinyl acetate), e.g., "Movilit," and poly(vinyl alcohol)	+
P 103	polycarboxylic acid, e.g., polyacrylates	+
P 104	polysulfonic acids, e.g., poly(styrene sulfonic acid)	–
P 105	polyesters, e.g., glycol polyacrylate	–
P 106	polyamides, e.g., polyacrylamide	–
P 107	polyurethanes, such as polyurethane elastomers that contain ionic groups (e.g., $-COO^-$, $-SO_3^-$ or $-N^+-$) or such as polyurethane elastomers that contain nonionic hydrophilic groups (e.g., $[-CH_2-CH_2-O-]_n$)	–
P 108	polyvinylpyrrolidone	–
P 109	poly(ethylene oxide), e.g., "Polyox"	–
P 110	polymeric cationic compounds, e.g., 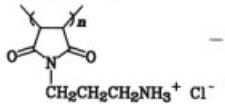	–
P 200	**polycondensates**	
P 201	ethoxylated phenol–formaldehyde resins, cf. N 311, N 313	–
P 202	sulfonated aromatic formaldehyde resins, cf. A 209	–
P 203	urea or melamine–formaldehyde compounds	–
P 204	polyamide, polyamine, and epichlorohydrin resins	–
P 300	**copolymers**	
P 301	styrene copolymers, e.g., styrene–acrylate polymers or styrene–ethylene oxide polymers	–
P 302	copolymers of polyvinyl and maleic acid compounds, e.g., styrene–maleic anhydride polymers or vinyl acetate–maleic acid ester polymers	–
P 303	polyvinyl–polyalkylene copolymers, e.g., vinylacetate–ethylene polymers, such as "Vinnapas", ethylene–acrylic acid–acrylic acid ester polymers or ethylene–acrylic acid–acrylonitrile polymers	–
P 304	other vinyl copolymers, e.g., vinyl acetate polymers, acrylic acid–acrylonitrile polymers, acrylic acid–acrylamide polymers, or butadiene–acrylonitrile polymers	–
P 305	acrylonitrile–butadiene–styrene polymers (ABS polymers)	–

* Identification of compound, as in Chapter 10.
** Biologic degradability.

Table 8. Special emulsion promoting agents

ID*	Name	Structure	BD**
H 100	**solubilizers**		
H 101	aliphatic alcohols, e.g., 2-ethylhexanol and dimethyloctyldiol		+
H 102	xylene		–
H 103	hydrocarbons, straight-chain branched		+ –
H 104	chlorinated hydrocarbons		–
H 105	amides, e.g., formamide, dimethylformamide		(–)
H 106	urea		+
H 107	urotropin		+
H 200	**emulsion thickeners,** see under P (Table 7) and T (Table 5)		
H 201	N 106, for example, for w/o systems		+
H 202	T 111, for example, for o/w systems		+
H 300	**protective colloids**		
H 301	natural substances, see T (Table 5)		
H 302	polymeric substances, see P (Table 7)		
H 400	**preservatives**		
H 401	aldehydes and aldehyde donors, e.g., hexamethylenetetramine derivatives; glyoxal		(–)
H 402	phenols, e.g.,	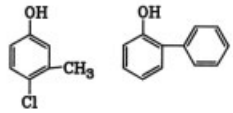	
H 403	organic mercury and tin compounds, e.g.,		
H 404	heterocycles, e.g.,	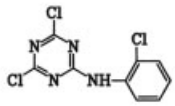	
H 500	**defoaming agents**		

∗ Identification of compound, as in Chapter 10.
∗∗ Biologic degradability.

9. Selection of Emulsifiers [5, p. 200], [6, p. 462], [167]

9.1. Selection Criteria

Emulsion manufacturers are generally confronted with the problem of selecting from hundreds of emulsifiers, only a few of which produce the desired dispersion of the components in each other. Which emulsifier or emulsifier mixture has an optimum effect in a given system depends, on the one hand, on the physical properties of the system and, on the other hand, on the function to be performed in the application. Thus, general practice includes trying out various emulsifiers. The following considerations may make selection simpler, but they also indicate how limited general rules are.

Chemical Structure. The chemical structure of the emulsifier should match the chemical structure of the internal phase. For example, fats can be emulsified with fatty acid esters or sulfated fats. As a rule, aromatic compounds are easier to emulsify with emulsifiers whose hydrophobic section contains aromatic rings. However, mineral oils are more readily emulsified with alkylphenol polyglycol ethers with a shorter polyether chain than with corre-

sponding alcohol polyglycol ethers. The choice of an emulsifier that has proved successful in emulsifying substances with a similar chemical structure is often advisable.

Solubility. The emulsifier should be sparingly soluble in the internal phase and as soluble as possible in the external phase, a requirement that in practice often results in compromises. As a rule, oil-soluble emulsifiers, such as lithium, calcium, magnesium, or aluminum salts of fairly long-chain alkyl carboxylic or sulfonic acids, primarily form w/o emulsions, whereas water-soluble emulsifiers such as the alkali-metal or ammonium salts of the previously mentioned acids are more prone to form o/w emulsions (Bancroft rule [168]).

Volume Ratio. The phase volume ratio (Section 2.7) usually determines the type of emulsion. In general, an excess of oil produces w/o emulsions, which necessitate oil-soluble emulsifiers. Similarly, with an excess of water, o/w emulsions are usually formed, for which water-soluble emulsifiers are used. However, in the manufacture of creams or in the production of water-in-hydrocarbon emulsions for textile printing [163], w/o emulsions with an excess of water are produced. Special oil-soluble emulsifiers such as "Span" types (N106, Table 1) are used [169].

Ionic or Nonionic Emulsifier. Whether an ionic or nonionic emulsifier is used for emulsifying depends on the system. As a rule, nonionic emulsifiers are used to prepare w/o emulsions. The addition of ionic emulsifiers has not proved effective. Nonionic emulsifiers are also often used to prepare o/w emulsions.

Usually, fairly large quantities of the uncharged emulsifier are needed to prepare stable emulsions (Fig. 11 A). In addition, the use of two nonionic components with polyglycol ether chains of different length generally decreases the required emulsifier concentration [170]. The addition of anionic emulsifiers has resulted in mixed emulsifiers, which have found widespread use (Section 3.6).

Several Ionic Emulsifiers. Ionic emulsifiers are usually not used alone (see Section 3.6), but they are used to produce emulsions for achieving special effects. The low stability of these emulsions is then often desirable. Thus, cationic emulsifiers produce a charge reversal in the gravel when they are used in tar emulsions in road paving. The subsequent spontaneous breaking of the emulsion is desirable so that water can rapidly evaporate from the road surface.

Mixed Emulsifiers. As a rule, the use of mixed emulsifiers requires a smaller amount of emulsifier than if nonionic emulsifiers are used alone (Section 3.6). If these emulsifiers are employed, the mixing ratio may determine the type of emulsion, especially if emulsifiers with opposite solubilities are added.

Acidic or Basic Emulsions. In acidic or basic environments, emulsifiers that are sensitive to hydrolysis cannot be used. In salt-sensitive emulsions, nonionic emulsifiers are more suitable than ionic ones, since no possibility of salt exchange exists. When emulsions are exposed to higher temperature, the cloud point of nonionic polyglycol ethers should be considered [70], [171], [172].

9.2. Selection Systems

HLB System. To make the selection of emulsifiers easier, GRIFFIN assigned a dimensionless number between 0 and 20 to each emulsifier to give information on water and oil solubility. Numbers between 0 and 9 characterize oil-soluble hydrophobic products, whereas numbers between 11 and 20 are used for water-soluble oleophobic compounds. The hydrophilic – hydrophobic equilibrium (termed "HLB," which stands for hydrophilic – lipophilic balance) is situated at the center (10) of this scale [169], [173], [174]. Substances with an HLB value of 10 are distributed between the two phases so that the hydrophilic group (molecular mass M_{H}) projects completely into the water while the hydrophobic hydrocarbon group (molecular mass M_{P}) is adsorbed in the nonaqueous phase.

In the ideal case, the HLB of an emulsifier with a molecular mass of M and a hydrophilic portion M_{H} is given by

$$\mathrm{HLB} = 20\frac{M_{\mathrm{H}}}{M} \tag{6}$$

Therefore, the HLB value indicates what mass fraction of the emulsifier projects into the aqueous phase in the ideal case. An advantage of the HLB system is that the HLB value of an emulsifier mixture (HLB_{mi}) is a linear combination of the contributions made by the individual components:

$$HLB_{mi} = g_1 \cdot HLB_1 + g_2 \cdot HLB_2 + \ldots \quad (7)$$

where g_1, $g_2, \ldots$ are the mass fractions of the individual components. If Equation (6) is multiplied by 5, the *Budewitz hydrophilic number,* HN, [175] is obtained:

$$HN = 5HLB \quad (8)$$

For fatty acid esters, the HLB value can be determined from the saponification index I_S of the ester and the acid index I_A of the fatty acid [176]:

$$HLB = 20\left(1 - \frac{I_S}{I_A}\right) \quad (9)$$

For *ethoxylated fatty* acid esters of polyhydric alcohols, (e.g., glycerol or sorbitol) the following simple formula can be used if the mass fractions of alcohol (w_A) and poly(ethylene oxide) (w_E) are known or can be determined [122].

$$HLB = \frac{w_A + w_E}{5} \quad (10)$$

If the hydrophilic portion of the emulsifier consists only of ethylene oxide (w_E), which is known or can be determined, the HLB values are given by Equation (11) [173]:

$$HLB = \frac{w_E}{5} \quad (11)$$

For polyglycol ethers Moore and Bell [177] give the following equation for determining the hydrophilic – lipophilic equilibrium *H/L*:

$$\frac{H}{L} = \frac{\text{number of ethylene oxide groups} \times 100}{\text{number of carbon atoms in the lipophilic part}} \quad (12)$$

For these emulsifiers, Equations (6) – (12) give similar HLB values. Table 9 lists HLB values of nonionic emulsifiers.

Table 9. HLB ranges of some nonionic emulsifiers , p. 202[5].

Emulsifier	HLB range
Fatty acid esters of [169]:	
Ethylene glycol	1.5 – 2.9
Poly(ethylene glycol)	1.8 – 18
Glycerol and polyglycerol	0.8 – 4.5
Sorbitol	1.8 – 8.6
Saccharose	7 – 15
Polyglycol ethers of [169], [174]:	
Alcohols	8.2 – 16.9
Fatty acids	8.1 – 15.0
Glycerides	8.0 – 18.0
Alkylphenols	9.9 – 16.5
Abietic acid	6.7 – 14.7

For *ionic emulsifiers,* no generally applicable method of determining reliable HLB values has been found. When Equation (6) is applied, the HLB values for ionic emulsifiers are too low; i.e., the ionic emulsifier is more hydrophilic than would be expected from the ideal distribution. To take account of this, a correction factor *C* has been included in (Eq. 6) [174]:

$$HLB = 20\frac{M_H}{M} + C \quad (13)$$

Table 10 shows tentative values of *C* for ionic and nonionic emulsifiers.

Table 10. Value of correction factors *C* for HLB values in Equation (13) [174]

Emulsifier	*C* in Equation (13)
Aliphatic polyglycol ethers	−1.2
Aromatic polyglycol ethers with one alkyl group	−1.9
Aromatic polyglycol ethers with two alkyl groups	−4.4
Ethanolamine salts of *n*-dodecylbenzenesulfonic acid	+2.1
Sodium salts of *n*-alkylsulfonates	+5.5
Sodium salts of *n*-alkylsulfates	+6.0

According to Davies [178], the HLB value of an emulsifier can be derived by combining increments of individual fragments, using the following equation (Table 11):

$$HLB = \sum mH_h + nH_l + 7 \quad (14)$$

Table 11. HLB increments according to DAVIES [178] and LIN [179]

Molecule fragment	HLB increment	
	H_h	H_l
$-SO_4Na$	38.7	
$-SO_3Na$	37.4	
$-COOK$	21.1	
$-COONa$	19.1	
$\equiv N$ (tertiary amine)	9.4	
Ester (sorbitan ring)	0.8	
Ester (free)	2.4	
$-COOH$	2.1	
$-OH$ (free)	1.9	
$-O-$	1.3	
$-OH$ (sorbitan ring)	0.5	
$-(CH_2CH_2O)-$	0.33	
$-(CH_2CH_2CH_2O)-$		-0.15
$-CH-$		-0.475
$-CH_2-$		-0.475
$-CH_3-$		-0.475
$=CH-$		-0.475
Benzene ring		-1.662
$-CF_2-$		-0.870
$-CH_3-$		-0.870

where m is the number of hydrophilic groups (increment H_h) and n is the number of CH_2 groups (increment H_l) that occur in the hydrophobic hydrocarbon part of the molecule.

For nonionic emulsifiers, the HLB values calculated by the Davies method usually agree with the Griffin HLB values. For ionic emulsifiers, values are obtained that are much in excess of 20 and, therefore, differ markedly from the Griffin HLB values.

P-Value. HANSCH suggests another method for emulsifier selection. The P-value is the distribution coefficient of compound R between octanol (o) and water (w):

$$P = \frac{c_R\,(o)}{c_R\,(w)} \tag{15}$$

Comparative studies have shown that this distribution coefficient P of compound R consists of increments π_x, in accordance with the following equation [180–182]:

$$\log P = \sum_{1}^{n} \pi_x \tag{16}$$

For hydrophobic parts of molecules, the logarithms of the P values are positive, whereas for hydrophilic portions, they are negative.

For $P = 1$, a substance is uniformly distributed between octanol and water. If $P > 1$ the substance is predominantly soluble in octanol, while if $P < 1$, the compound preferably dissolves in water.

Values of π_x are listed in Table 12 [182]; such values are unknown for ethylene oxide and propylene oxide, and the determination of values of π_x for ionic groups such as the $-SO_3Na$ group is impossible [182, p. 21].

Phase Inversion Temperature (PIT) . SHINODA uses the phase inversion temperature (PIT) to characterize an emulsion. The value of the PIT indicates the temperature at which a water-soluble emulsifier hydrate is transformed into an oil-soluble one [69], [75], [172]. When the PIT is lower, the emulsifier used is more hydrophobic, with a low PIT value corresponding to a low HLB value [183]. In the case of polyglycol ethers, the PIT is comparable to the cloud point [71], [184]. The PIT can be determined visually, electrically (see Section 11.5), or by differential thermal analysis [185]. A direct relationship between PIT and HLB values has only been established in homologous polyglycol ether series [183], [186]. No assessment of the HLB values of ionic emulsifiers has been provided by PITs thus far, although phase diagrams of water and oil with a mixed emulsifier containing an ionic component reveal a phase inversion.

Other Systems. Other classifying indices for emulsions include the R factor [187] and the cohesive energy ratio (CER) [188], [189]. Table 13 compares the numerical values of the different selection systems used in various fields of application [177], [182], [190], [191].

9.3. Assignment of HLB Values

According to experience with the HLB system, an HLB value should be assigned not only to the emulsifier, but also to the substances to be emulsified. The HLB values required for frequently emulsified products are listed in Table 14 [192]. The significance of this for the emulsion manufacturer is as follows: if the HLB value of the product to be emulsified is known or can be determined from Table 14, the emulsifiers or emulsifier mixtures selected for preparing an emulsion should have an HLB value which approximately matches that of the product to be emul-

Table 12. π_x values according to HANSCH [180]

Group	π_x	Group	π_x
Aliphatic group		$-CONH_2$	−1.71
$-CH_3$; $-CH_2-$ (chain)	0.50	$-COCH_3$	−0.71
$-CH_2-$ (cyclic)	0.41	$COOCH_3$, $-OCOCH_3$	−0.27
$-CH-$ (chain branching)	−0.20		
$-NH_2$	−1.19		
$-NHCH_3$	−0.67	**Aromatic group**	
$-N(CH_3)_2$	0.32	benzene ring	2.13
		$-CH_2-C_6H_5-$ (benzyl)	2.69
$-CH_2-NH-CH_2-$	−0.17		
$-CH_2-N(CH_3)-CH_2-$	0.18	$-CH_2-$ (on aromatic group)	0.56
		$-NH_2$	−1.23
$-OH$ (primary)	−1.16	$-OH$ (phenol)	−0.67
$-OH$ (secondary)	−1.39	$-OH$ (benzyl)	−1.59
$-OH$ (tertiary)	−1.49	$-OCH_3$	−0.02
$-OCH_3$	−0.47	$-OCCH$	0.28
$-SCH_3$	0.45	$-CH_2COOH$	−0.72

Table 13. Numerical values for different emulsion selection systems used in various fields of application , p. 205[5]

	HLB	R	CER	PIT	log *P*	H/L
Defoaming agents	0 – 3	$\gg 1$	up to 3.90		7.84 (stearyl alcohol)	
Emulsion, w/o	3 – 8	>1	3.90 to 1.25			20
Wetting agents	7 – 9	≈ 1	1.50 to 1.00			
Emulsion, o/w	8 – 18	<1	1.25 to 0.08			40 to 125
Detergents	11 – 15	<1	0.70 to 0.29	25 to 105		
Solution promoters	12 – 18	<1	0.58 to 0.08		0.32 (2-propanol) −0.66 (n-ethanol)	150

Table 14. HLB values of frequently emulsified products [192], [169]

Emulsified product	HLB	Emulsified product	HLB value
o/w emulsions			
essential oils	12 – 18	nonylphenol	14
avocado oil	9	dichlorobenzene	13
cotton seed oil	10	oleic acid	17
beeswax	12	paraffin oil	11
benzene	15	petroleum	14
gasoline	14	Pur-Cellin oil	9.5
butyl stearate	11	castor oil	14
ethyl benzoate	13	silicone oil 300 CP	11 – 12
carnauba wax	12	sunflower oil	8
chlorobenzene	13	stearic acid	15
chloroparaffin	8	stearyl alcohol	14
cyclohexane	15	styrene	15
decahydronaphthalene	15	table paraffin	11
diisooctyl phthalate	13	carbon tetrachloride	16
diisopropylbenzene	15	toluene	15
isopropyl myristate	9.5	trichlorotrifluoroethane	14
pine needle oil	16	tricresyl phosphate	17
lanolin (wool fat)	10 – 11	vaseline	12
lauryl alcohol	14.5	vitamin oils	12 – 18
laurylamine	12	wheat germ oil	9.5
menhaden oil	12	xylene	14
mineral oil	10 – 12		
w/o emulsions			
cotton seed oil	5	stearic acid	6
beeswax	4	table paraffin	4
lanolin (wool fat)	8	vaseline	5 – 7
paraffin oil	8		

sified. If the required HLB value is unknown, it can often be determined by a series of tests (see Section 11.5).

Various attempts have been made to determine HLB values experimentally [173], [178–182], [193–199]. For nonionic emulsifiers, all of these methods yield usable HLB values that are often only comparable in homologous series.

Classical emulsification experiments using a test series [169], [173] (see Section 11.5) often result in HLB values that can also be used for emulsifier selection. In such a case, the method of calculation provided is a valuable aid for many emulsion producers.

The HLB system and the other selection systems provide no information on the application properties of emulsions such as their long-term stability or their heat resistance. Therefore, any decision on the suitability of an emulsifier must be based on experimental data [100].

9.4. Selection Procedure

If no details of the phases to be emulsified are known, the following selection procedure can be adopted in practice:

1) Determination of the required HLB value of the substance to be emulsified either from a comparison with the substances in Table 14 or by emulsification experiments using a series of test emulsifiers (see Section 11.5)
2) Selection of emulsifiers or emulsifier mixtures having an HLB value that matches the required HLB value; if no HLB values are known, they can be estimated from the chemical structure of the available emulsifiers by using Equations (6) – (14).
3) Qualitative determination of the solubility of the selected emulsifiers in a test tube; low solubility in the internal phase and good solubility in the external phase are prime requirements for rapid adsorption at the interface between the phases to be emulsified.
4) Determination of the interfacial tension between the phases to be emulsified with and without 5 g/L of emulsifier added (Section 11.4); if the result indicates a marked lowering in the interfacial tension, a necessary condition for emulsification is fulfilled.
5) Emulsification experiments in a glass cylinder, as described in Section 11.1, to test the desired rate of emulsion formation and the required stability
6) If stability over a fairly long period is required, long shelf lives can be simulated in short periods by centrifuging.
7) A large-scale practical test under the required conditions

10. Fields of Application

The applications of emulsions in industry and technology are extremely varied [5, p. 389], [6, p. 465], [19], [20], [22], [29], [171]. Not only do they improve, simplify, or cheapen conventional methods, but emulsions often make new technologies possible. The emulsifiers that are used are of particular interest, since considerable effects are usually achieved with small quantities of substances. Often the addition of wetting agents, solubilizers, thickening agents, or preservatives is indispensable. Since the theoretical aids for preparing emulsions are still only at the initial stage, laborious testing of many emulsifiers is usually unavoidable. A collection of emulsion formulations for a wide variety of applications is found in [15, Vol. II].

The most important applications for emulsions and the assiociated emulsifiers and emulsion promoters are listed on the following pages. The emulsifiers specified are given under the corresponding key numbers in Tables 1, 2, 3, 4, 5, 6, 7, 8. Occasionally, commercial products are mentioned whose composition is not precisely known; these are shown in quotation marks. An F behind the emulsifier designation signals preference for emulsifiers that have double bonds in their hydrophobic part. In such cases, derivatives of oleyl alcohol or oleic acid are usually used. Occasionally substances that are to be emulsified are cited in brackets behind the emulsion type.

10.1. Cosmetics [19, p. 646], [22, p. 167], [29, p. 178], [134],[171, p. 608], [201–203]

Cream bases, w/o, HLB 4 – 6: T103, N315, T101 [204], N314, T102, T106, N301, A101 AA, A104, A401, N102, N104, M101, N105, N106, N107 N109, N304.

Fat and oil emulsions, w/o, HLB 5 – 7: N103, N105, N301, N304, N303; o/w, HLB 12 – 15: N308, N310, N404.
Essential oil emulsions, o/w, HLB 16 – 18: N301, N303, N304, N307, N310, A208 + N303, N303 + N310 + N102.
Paraffin oil emulsions, o/w, HLB 10 – 14: N301/F N105 + 304, N106 + N304, N301 + N303.
Stearic acid emulsions, o/w: N301, N303, N102, N105, or N106 with N304 or N307.
Wax emulsions, o/w, HLB 14 – 16: Borax, N106, N301, N310, N304, individual or mixed with each other.
Lanolin emulsions, o/w, HLB 12 – 14 and w/o: N301.
Silicone emulsions, N301, N310, N301 + N307 [205].
Toothpastes, o/w: A303.
Foam baths, o/w, HLB 14 – 18: A307, A101, A201, A208, A301.
Hair Lotion: K100 with N100 or N300 [206], [207].

10.2. Pharmaceutical Preparations [19, p. 638], [29, p. 178], [171, p. 673], [208],[209, vol. 7/3], [210]

Drug emulsions, o/w: T110, A301, N105, N106, N109 [210], [211], N301, N303, N304 [212], [213], K103, K104, K107, A106, A402.
Ointment bases, [214], [215], w/o, HLB 3 – 6: T101 and H101 with A301 added [164], N102/F, N105/F, alone or mixed with N304, N301, N303 or N307; mixture of T101, T102 or T103 with N301 and H100. o/w, HLB 8 – 14: N301, N303, N304, A301 alone or mixed with each other.
Suppository compound, w/o: N102, N105, N106 alone or mixed with N304 or N307.
Encapsulation: T106.
Injection preparation [216–218]: N307; especially ethoxylated castor oil.

10.3. Agriculture [19, p. 681], [171, p. 722], [219]

Soil improvement: P103, P302, A201 short-chain.
Fertilizer additive: N301
All purpose cleaners: N102, N105, N106, N301, N303, N307, alone or mixed with each other.
Cleaners for fruit and vegetables: N105, N106, N108, M102.
Cleaners for milk churns: K100 [22, p. 178].

10.4. Crop Protection [19, p. 684], [29, p. 182], [209, vol. 7/4],[219–222]

Naturally occurring insecticides (e.g., pyrethrum), o/w: N304.
Chlorinated hydrocarbons, o/w: 208/S, 208/A, or 208/AA with N311, N307, N105, or N106; N301 with A208, "Toxanon 500"; "Emcol N141 B," "Emcol N146," "Emcol N300 B," or "Emcol N500 B," "Emulsogen IC, IT, and 140."
Carbamates, sparingly soluble: A208/AA with N311 or N307.
Carbamates, readily soluble: A209 with A210 "Emulsogen IT."
Phosphoric acid esters [223–225], o/w, HLB 10 – 14: A208/S with N311, N307, N105, or N106; "Atlox 3403," "3404," or "2309"; "Toxanon 888" or "P900"; "Triton X151"; "Emcol N141 B," "Emcol N 146," "Emcol N 300 B;" "Emulsogen ITN."
Fungicides, o/w: A209 or A210 with N311; A208/AA with N311 or N310.
Herbicides, o/w: A208/S with N311 or N307; N304, N307, or N311 with A209, "Emcol N710," "Emcol N311" with A209 or "Tocanon HC-23," "Emulsogen 140 and IT."

10.5. Food Industry [2, p. 29], [19, p. 689], [22, p. 178], [29, p. 177],[34], [35], [171, p. 713], [219], [226]

Table 15 gives a survey of the methods of preparation and the stabilization mechanisms in emulsions used in the foods industry [2, p. 29].
Bread and cakes [207], [227–229], w/o, HLB 4 – 8: T102; T106, N108, N105, N106, N304, or T105, alone or mixed with each other.

Margarine, w/o: T102, T106, N105, N109, A402, A204.
Chocolate [220]: T102.
Fat bloom prevention, w/o, HLB 5 – 10: N105 or N106, alone or with N304.
Sugar frosting, o/w, HLB 14 – 16: N304.
Softeners for caramels and chewing gum, w/o, HLB. 2 – 4: N105.
Prevention of sticking, w/o, HLB 2 – 4: N105.
Ice cream additives [206], [231], w/o, HLB 4 – 6: N105, alone or mixed with N304 or T105.
Wetting of milk and baking powders [232], w/o, HLB 9 – 11: N304.
Custard powder, w/o, HLB 2 – 4: N105.
Drinks industry: T107; T106.
Fruit and vegetables: N 105.
Flavorings, w/o and o/w, HLB 10 – 12: N105 or 106, alone or mixed with N304.
Meat, salad, or other flavoring sauces, o/w:T105 [284].
Food dyes, w/o, HLB 2 – 4: N 105; o/w, HLB 8 – 18: N304.
Foam inhibitors: N109; N406 [201].

10.6. Textile Industry [19, p. 623], [20, p. 2557], [22, p. 160], [29, p. 189], [171, p. 567], [209, vol. 6/5], [233], [234]

Pretreatment agents for wool or jute [235], o/w (oleic acid – mineral oil): N106, N101, A201.
Pretreatment of cotton: o/w: A205, A401, A201, A208, or A201, A208 with N301 or N310.
Spinning preparations and textile lubricants [221], [222], [226] [236–240], o/w (mineral oil), HLB 8 – 12: N165, N106, N304, N307, A402, generally in mixtures, also A208 with N304.
Carrier concentrates [241–243], o/w: A208/S or A208/AA with N301 or N311; N310 with A307.
Level-dyeing assistants with dye affinity [216]: N301, N303, N201, N305.
Level-dyeing assistant with fiber affinity [217], [218]: A303, A301, A208, A213, A210.
Dye liquid formulations, o/w: N301, N310, A201, A301. Solvents: ethylene glycol, propylene glycol, acetic acid, and urea as solubilizer.
Avivages [244–246], o/w (paraffins, waxes, silicones, polyethylene): N100, N202, K110, N204, often mixed with N301, also A307, A402, M102, T101.
Sewing thread, o/w (paraffin): K101, K164, K108 [247].
Waterproofing agents, o/w, HLB 8 – 11 [248–250] (paraffin, waxes): N105, N106, N301, often with R102 or R104 added, and also A208 with N304 or N204 with A101/A.
Waterproofing agents, o/w (polyethylene), HLB 10 – 12: N301, alone or mixed with N106.
Waterproofing agents, o/w (silicones): K 104 and also N301 or N310 with longer ethylene oxide chain.
Fixing accelerators: H108 and also ethoxylated form N301, N310, N303, [18, p. 191], [252], [253].
Printing pastes: H100, A200: o/w or w/o (T109 – T106 – 111 – synthetic latices): N301, branched; N311, A200, A300 [163], [251].
Pad aids, w/o: T109, T111, T108, A201, generally in mixtures; o/w: A301 or N310, alone or in mixtures.
Antistatics: A303, A301, A201, N301, N310, N303, N202, N306, N305, A401, N317.
Flocculating and accelerating agents, o/w: P100, especially P102, P107, P303.

10.7. Plastics Industry [16, vol. 14/1], [19, p. 744], [20, p. 2886],[209, vols. 6/3, 6/4], [254–261]

Emulsion polymerization (EP) [115, vol. 1, p. 64] of butadiene, o/w: A101, A208, A209, A307, A201.
EP of chloroprene, o/w: A108; N310.
EP of styrene, o/w: A208, A301, A303, A307, N301, N303, N310.
EP of vinyl polymers, o/w: A101/short-chain, A404, A201, A206/short-chain.
EP of acrylonitrile – butadiene – styrene (ABS), o/w: A108, A101/S, A208 [262], [263].
EP of polyamide and polyester, o/w: A210 with N301 or N311.
EP of polyethylene, o/w: A301, N204, A207 [264], [265].
EP of acrylonitrile, o/w: A101, P110.

Table 15. Examples of typical food emulsions

Food type	Emulsion	Preparation method	Stabilisation mechanism
Milk	o/w	natural product	natural protein – lipid membrane
Cream	o/w	centrifugation	natural protein – lipid membrane
Ice cream	o/w	homogenization	reconstituted protein membrane
Butter and margarines	w/o	churn/votator	fat crystal network
Sauces	o/w	high-speed mixers/homogenizers	protein – polysaccharide membrane
Fabricated meat products	o/w	low-speed mixers, choppers	gelled protein matrix
Bakery products	o/w	mixers (various)	starch – protein network

EP of acrylonitrile, w/o: A101, N106, K107 [266].
EP of phenol formaldehyde resins, o/w, w/o: T106, A101/S.
EP of acrylamide, w/o: N106.
Bead polymerization, o/w: P102, T105.
Foam plastics, o/w: A202, N109; w/o: N318.
Release agents (silicone emulsions), w/o, HLB 8 – 12: A202, N301, N310, [267].
Microcapsules [268–274], o/w: T106, T102, T105, T107.
Adhesion improvement between fillers and plastics: potassium oleate, N311, A209 [275].
Emulsifiers for nonaqueous systems, o/o: P301, P201, P202, P104, alone or as copolymers with styrene [114].

10.8. Building Industry [13], [19, p. 673], [22, p. 180], [29, p. 186]

Asphalt emulsions [152], [276], [277], w/o, o/w: K110, K101 [278], K108 [279].
Cement thinners [262], [280], [281], w/o: A307, A208, T104 or A110, alone or mixed with N310 or A209, and also silicone emulsions; addition of H500.
Waterproofing of masonry, o/w: A101/barium salt.
Adherents for concrete, o/w (paraffin/mineral oils): N105, N106, N301, often with R102 or R104 added; also A208 with N204 or N204 with A101/A.
Plasters and paints (polymer dispersions) [115, vol. II, p. 32], o/w: N100, N301, N308, N309; mixed emulsifiers P102, P107, H100, H200, H400.
Filler putties (polymer dispersions) [115, vol. II, p. 80], o/w: A201, A208, A301, A401.
Wood impregnation, w/o, o/w: A208 or T104, alone or mixed with N301 or o/w: N310.

10.9. Paints and Coatings [19, p. 721], [20, p. 2880], [29, p. 192],[171, p. 746], [209, vol. 6/6], [211], [256],[282], [283]

Emulsion paints, nonaqueous systems: A108, A109/S, A303, N203 [114].
Emulsion paints, aqueous systems, o/w: N301, N310, N311, A108, A101, N100, T104, T106, A107, A210, A208, P103, A405, T105, N308.
Strippers (paint removers), w/o: N310 alone or mixed with A101 or A201.

10.10. Paper Industry [19, p. 697], [20, p. 2890], [22, p. 177],[29, p. 193], [171], [228–234], [247],[284–290]

Antiresin agents: A209, A201, N100, N300.
Antifroth agents: A101 and many others.
Dispersions of white pigments: A405, A406, P103.
Dispersions of colored pigments: A209, A201.
Coating binders: A201 and many others.
Surface finishing, o/w, conductive papers: K100; absorbent papers: A200; glossy papers: A303; waterproof papers [poly(vinyl chloride) – polyurethanes – silicone – oil]: A201; cardboard articles (paraffin): A201.
Improvement of wet-fastness: P204 [291].
Glueing, o/w: (natural resins – stearyldiketene – T103 – T107 – T106 – P302): A108, A303, alone or mixed with N300.
Sieve and felt cleaning: A201 and many others.
Fiber recovery (emulsion breaking): electrical with addition of A303, N205, P106 [289], [292–294].
Wastepaper recycling [292], [295], [296]: A209, A210.

10.11. Leather Industry [19, p. 693], [20, p. 2898], [171, p. 273],[297–305]

Degreasing, o/w (fats), HLB 2 – 6: N105 mixed with N 304, with solvents added.
Tanning, o/w: N310.
Binder: P103, P304 [298, pp. 750 – 753].
Retanning, o/w: N310.
Dyeing, o/w: P202, N310, N301, N305, N306, [306], [307].
Oiling, o/w, (sperm oil – neat's foot oil – lanolin – chloroparaffin): N105; N301; M100 [308–310].
Finishing [276], [277], o/w (collodium – T106 – T103 –P103 – P107): A303, A210, A301 alone or mixed with N300, N100.
Polishes, w/o: K197/F, alone or mixed with N100, N300.

10.12. Mining [22, pp. 174, 180], [209, vol. 5/4], [311]

For hydraulic fluids, see Section 10.13; for wood treatment, see Section 10.8
Flotation [20, p. 2956], [29, p. 193], [171, p. 836], [209, vol. 4/2], [312], [313].
Flotation of nonferrous metal ores: A407, xanthates.
Flotation of oxidic ores: A101, A201, A301.
Flotation of potassium salts: K103, K104, K107, N202.

10.13. Metal Industry [19, p. 658], [29, p. 2883], [22. pp. 172, 175],[29, p. 221], [171, p. 728]

Cutting, grinding, and drilling oils (cooling lubricants), o/w, HLB 2 – 6: A101/S, A108/S, A109/S, A201/S, N106, N203, N301, N303, N304, N310, N400, H100, K103, K108, H400, and also ionic/nonionic mixtures.
Rolling and drawing oils, o/w, HLB 2 – 6: N106, N203, N301, N303, N310, N400.
Hardening oils, o/w: N301, N310 [247].
Hydraulic oils: N301, N310, with fairly long polyether chain.
Anticorrosion agents, o/w, HLB 2 – 8: A104, A105, A203, A201, A400, K105, T101, [171, p. 858].
Polishing pastes, o/w (mineral oils): N301, N303, N310.
Release agents, o/w (mineral oils): N300, A203.
Metal cleansers [171, p. 550], aqueous, o/w (soda – sodium hydroxide solution – sodium metasilicate – sodium tripolyphosphate): N310, N305, N316, K102.
Metal cleansers, nonaqueous, w/o: A306.
Metal dryers: K100, M100.

10.14. Electroplating [19, p. 664], [22, p. 173], [29, p. 193], [314], [315]

Electrolytic baths: A207 [316], (Bayer FT 248) [317], A301, A201, A102, A208 [318], P109, N308, N301, N305, N316.
Polishing of steel, tantalum, or nickel: N-acylsulfonylureas A201, A301, P106, T105.

10.15. Polishes, Manufacture and Removal [171, p. 821], [247], [317], [319–321]

Wax emulsions, o/w: A102, A207, N301, N304, morpholine soaps: addition of A207, Bayer FT 248.
Polyethylene emulsions, o/w: A301, A207, K101, N105, N106, N301 or, N304 alone or mixed with A210, P103 or P104, P202 [322].
Silicone emulsions, o/w: A301, morpholine oleate, A103/A, oleic acid esters, A402, addition of A207 (Bayer FT 248).
Car washes: wax or silicone emulsions with K103 added.
Wax film removal (gasoline – water emulsion): N100, N300.

10.16. Petroleum Industry [323]

Drilling aids, o/w: A101, T104, T105, A208, K103, K108, K109.
Deemulsifiers [16, vol. I/1, p. 213], [324], w/o: A201/F, A301/F, A210, A201/A, A208/A, M100, K108, N300, generally mixed.
Additives to flood water, o/w: N201, K103, K104, N305, N306, N303, N310.

For major spills on transportation, [325], [326], o/w: N300/short polyether chain length "Dispersole."

10.17. Mineral Oil Industry [19, p.667], [29, p. 221], [171, p. 793], [323],[388]

Fuel additives, w/o, HLB 2 – 6: A402/A, N105, N202, succinimides, succinamides, K108, polyisobutaneamines, and other [29, p. 227], [327].
Heating oil additives: P303, ammonium salts of medialanic acid.
Lubricant additives, w/o, HLB 2 – 6: A101/S, A108/S, A109/S, A201/S, A207/S, A404, P105: polyisobutylsuccinimide, P100, and others [29, pp. 248 – 250], [215], [328], [329].
Lubrifying additives, w/o (gasoline, diesel fuel) [5, p. 418]: N310 with various chain lengths, mixtures of N106 with N304 [330], or mixtures of N203, N105 with N306 [331], [332].

10.18. Adhesives Industry [22, p. 177], [29, p. 193]

Rubber adhesives: N100, N300.
Vinyl resin adhesives: A301, A307, N308 [333].
Adhesive dispersions: [115, Vol. II, p. 82].
Furniture glue, o/w: P102, P106.
Packaging industry, o/w: P102, P303, P103, P301, P304, P302, polychloroprene, P107.
Building industry, o/w: P102, P103, P301 for tiles, polychloroprene.
Leather industry, o/w: P107, P304, polychloroprene.
Textile industry (lamination coating): P102, P303, P107.

10.19. Photographic Industry [171], [334–338]

Wetting agents in pouring solutions [339], o/w, HLB 8 [340]: T110, A206, A201, A208, A304, A209, N310, N109, N106, N310, N305, A207 (Bayer FT 248).
Antistatics [341]: A301, A206, M102, K110, mixtures of M102 or K110 with A102, A207, or K111.
Lubricants, o/w (stearic acid – wax – polyethylene or silicone oil): short-chain N301 with A209 added.
Color component emulsions [342], o/w or w/o: A201, A210, A301, A204 [343], A209, alone or mixed with T110 [344], [345].
Processing baths: N101 with A201 added, A301, Agepon (A301 + A205 + A206 + N301).

10.20. Chemical Industry [19, p. 744], [20, p. 2886], [254–256]

Reaction accelerators [16, vol. V/3, p. 516], o/w: A201, A208, A209, N301, N310.
Extracting agent additives [346], [347], w/o: A208, N301, N310 [246], [347].
Micellar catalysis [348–352]: A301, P104.

10.21. Detergents and Cleansers

Protective colloids, o/w: T105, A405.
Optical brighteners, o/w: N310.
Cleaning intensifiers [353], w/o: A101/A, A101/AA, A201/A, A201/AA, A208/A, A207/AA, A206/A, A206/AA, N301, N310, H400, alone or mixed.
Industrial and foam cleansers, o/w: A201, A208, N301, N310, alone or mixed.
Washing liquids, o/w: A201, A208, N301, N310, alone or mixed.
Low-phosphate detergents [354]: P302 + A405 + 308.

10.22. Special Applications [5, p. 436]

Fire fighting: A301, A307, N100, M101 with A209 added, H207.
Evaporation retarders: N301 with 1 – 3 ethylene oxide.
Glass fiber reinforcing agents: N103, P101, silicones.
Window cleaning agents: N310.
Antidust agents: A104/F, A301/F, alone or mixed with N301, N310.

11. Investigation of Emulsions [15, vol. I, p. 53]

11.1. Stability [355]

The simplest method of determining stability is to measure *settling or creaming time* in a graded cylinder [356]. Poor emulsions separate rapidly, whereas good ones remain stable for a prolonged period.

The ability to wet the glass wall provides information about the quality of the emulsion. Stable o/w emulsions leave a coherent film on a moist wall, whereas with unstable emulsions, the film breaks up.

When the *spontaneous emulsification* of an emulsion concentrate is tested, droplets are added to the water in a stationary graded cyclinder and should rapidly become finely dispersed without a supply of mechanical energy.

Microscopic investigation easily shows whether the instability of an emulsion is due to reversible sedimentation or to irreversible coalescence. With very small droplets, particularly in colloidal emulsions, electron microscopy is required. When stability is tested, the investigations should be repeated at specified time intervals [357].

Centrifugation is the most rapid method of testing the stability of an emulsion [358]. If it does not break when centrifuged at ca. 300 rpm for 30 min, an emulsion may be assumed to have long-term stability. On the other hand, if it breaks, it may nevertheless be stable on standing undisturbed for a prolonged period.

Increase in *turbidity* [359], [360] is an indication of an increase in droplet size, which can be measured with an nephelometer [361], [362]. In the case of colorless emulsions, particularly colloidal emulsions the light transmission, for example at 700 µm, is a measure of the cloudiness caused by scattering. This can be readily observed and is easy to record.

Simple physical parameters are also suitable for assessing the stability. Useful information on emulsion stability can be derived from regular determination of *density, viscosity,* and particular, of *droplet size* and *droplet size distribution* [357]. The coalescence time is measured directly by determining the time required by the oil droplets to coalesce at the water – oil interface [98].

The only method for determining the *long-term stability* is to store the emulsions as long as possible under the conditions to which they will later be subjected.

In the case of industrial emulsions, the following properties are important: shelf life [339], [363], heat resistance, cold resistance, resistance to change in pH, resistance to the addition of salt, and resistance to mechanical action. The test procedures for these depend on the particular application, and the specified test methods should be modified accordingly. For example, a stability test should be repeated after a specified storage time or heat treatment.

11.2. Emulsion Type

A number of methods have been developed for determining the emulsion type [364, p. 129 – 131], [365]. A very simple and reliable method is the dilution test, since only o/w emulsions can be diluted with water. This test can also be performed on a small scale, the drop dilution method [13]. In this method, a drop of the emulsion on a glass plate is brought in contact with a drop of water. Quick spreading of the emulsified particles indicates an o/w emulsion, whereas the formation of a clear phase boundary characterizes a w/o emulsion. The result can be checked by repeating the experiment with a drop of benzene, in which case the situation should be reversed.

The Roberton dyeing test utilizes the solubility of a dye in the emulsion phase [366]. Coloration is produced in o/w emulsions by water-soluble dyes such as methylene blue or malachite green, but in w/o emulsions by oil- or fat-soluble dyes. The procedure, which involves a mixture of two dyes, is as follows:

A red, fat-soluble and a green, water-soluble dye are separately triturated with talc in a ratio of 1 : 1 and then intimately mixed with each other. This mixture is dusted onto a filter paper in a layer so thin that it can scarcely be seen. The coated paper is colored green by a drop of water and red by a drop of oil. If a drop of an o/w emulsion is applied to the paper, the center is red and the outside rim green. This effect is reversed for a w/o emulsion. To examine pastes, creams, and other emulsions of similar consistency, a thin

coating is smeared onto a microscope slide. After sprinkling an extremely thin layer of the dye mixture onto it, the slide is heated to 50 °C and the colors are observed as described above.

The paper method is based on the different ways in which paper is wetted by o/w or w/o emulsions. The rate of wetting being determined by the external phase. Since water rapidly wets filter paper, but oil only wets it slowly, the external water is seen to spread out rapidly in an o/w emulsion while a drop of a w/o emulsion applied to the filter papter initially remains spherical. Instead of filter paper, an absorbent fabric or a porous earthenware plate may also be used.

The *electrical conductivity* of o/w emulsions is higher than that of w/o emulsions [367]. The method benefits from a low electrolyte content since the phase reversal point can easily be determined. The use of a microwave resonance dielectrometer for determining emulsion type has been described [368].

11.3. Droplet Size and Droplet Size Distribution [27, p. 136]

A microscope often suffices to determine the droplet size and distribution [369]. Agglomeration or coalescence of the emulsion droplets can easily be distinguished. Even a semiquantitative measurement with a microscope stage micrometer provides usable information on droplet distribution.

To eliminate Brownian motion, the emulsion is allowed to dry partially or completely on the slide. To the same purpose, the addition of thickening or gelling agents is often effective [370]. Microprojection [371] or microphotography are also used.

More precise counts may be made with the hemocytometer [91] or the Heller chamber [372]. Automatic counting equipment has been developed for counting good micrographs [373], [374].

With finely dispersed or colloidal emulsions, the resolution of the light microscope is inadequate [101]. This has been overcome by using either the Coulter counter [375–377] or the electron microscope [378].

A freeze-drying method has been developed for electron microscope studies [379], [380]. In this method, the emulsion is frozen very quickly, and the external phase is then allowed to sublime from the solidified emulsion. This method usually works well in the case of o/w emulsions.

Sedimentation methods with or without centrifugal acceleration can also be used to determine the droplet distribution. Large droplets fall or rise more rapidly than small droplets (Stokes's law) [38, p. 136].

The measurement of cloudiness and light transmission has already been referenced in Section 11.1. Computer programs have now been developed that can be used to determine the mean droplet radius from measurements of light transmission. Droplet size distributions have also been calculated for o/w emulsions and are said to agree well with experimental data [380], [381]. Density changes can only be determined for emulsions that are not too stable [373], [382].

11.4. Interfacial Tension [383]

The following methods may be used to measure the interfacial tension between the phases to be emulsified:

1) geometrical measurement of drops, bubbles, and meniscuses [384]
2) maximum bubble pressure method [385]
3) Lenard anchor or ring method [386]
4) plate method [387]
5) capillary wave method [388]
6) spinning drop method [389], [390]

The ring method has been automated [391]. The spinning drop method can only be applied to o/w emulsions. Small interfacial tensions can be measured.

The methods listed above cannot be used to measure the interfacial tension if the interface between the two phases is not sharp. The decrease in interfacial tension on addition of emulsifiers is also measured inaccurately when very low interfacial tensions are involved; this is the case when an efficient emulsifier is added to two phases both having a low surface tension. In such cases, measurements of the surface tension of the phases involved [392] with and without the addition of emulsifier provide a rough guide to the usability of an emulsifier.

11.5. Required HLB Value

To determine the required HLB value of a substance experimentally, two emulsifiers with known HLB values that are as different as possible are mixed in different ratios. This produces a series of mixed emulsifiers with continuously changing HLB values.

An example is the mixture of Span 60 (sorbitan monostearate) having an HLB value of 4.7 with Tween 60 (poly(ethylene oxide) – sorbitan monostearate) having an HLB value of 14.9 (Table 16). Mixtures with different HLB values are obtained [169, p. 7].

Emulsions of the substance to be emulsified are then prepared with these emulsifier mixtures to determine which emulsion is the most stable. The HLB value of the emulsifier mixture used for the most stable emulsion corresponds to the HLB value of the substance to be emulsified. Good quality emulsions of that substance should then be obtained with the aid of emulsifiers having comparable HLB values.

Table 16. Emulsifier mixtures with continuously changing HLB values

Emulsifier mixture		Calculated HLB
Span 60, %	Tween 60, %	value (Eq. 7)
100	0	4.7
87	13	6.0
68	32	8.0
48	52	10.0
28	72	12.0
6	94	14.0
0	100	14.9

The *water number* which is usually proportional to the HLB value for nonionic emulsifiers, can be determined simply and rapidly [197], [198].

In this method, 1 g of the substance whose HLB value is to be determined is dissolved in a mixture consisting of 4 % benzene and 96 % dioxane. The volume of water required to produce the first permanent cloudiness is proportional to the HLB value of the substance.

The *cloud point* [71] can also be determined rapidly and conveniently for polyglycol ether; the cloud point corresponds to the PIT (see Section 9.2), which in turn corresponds to the HLB value of polyglycol ethers.

To carry out the determination, 1 g of the polyglycol ether is heated in 100 mL of distilled water until a cloudy solution is obtained. By slow cooling, the temperature is determined at which the solution becomes clear. Semiautomatic [393] and automatic equipment [394], [395] is now available for determining the cloud point. If the cloud point is above 95 °C, pressurized equipment may be employed [27], [75]. The addition of salts (50 g of NaCl per liter of water) or solvents (such as ethylene glycol monobutyl ether) lower the cloud point, whereas acids, such as acetic acid, increase the cloud point [27], [396]. Figure 14 plots the cloud point of 4-isononylphenol against the HLB value as determined from Equation (13). Poly(glycol ethers) occasionally emulsify the precipitate; in such cases the addition of a trace of salt (e.g., soidum sulfate) is required, or the cloud point must be determined in soft water.

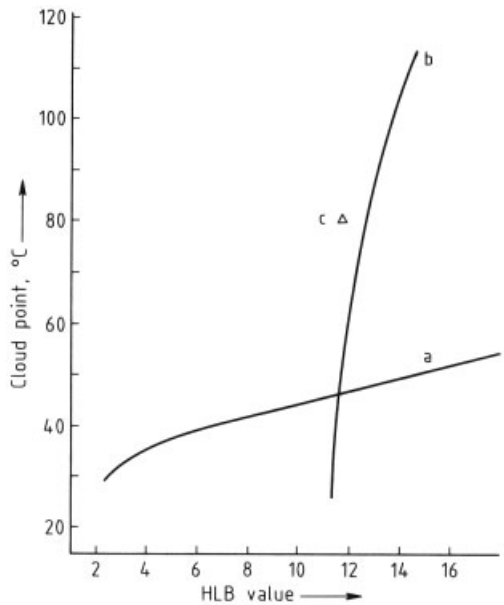

Figure 14. Cloud point vs. HLB value of 4-isononylphenol polyglycol ethers [6, p. 216]
a) 1 % solution in 25 % aqueous solution of ethylene glycol butyl ether; b) 1 % solution in water; c) 1 % solution in 10 % acetic acid (Δ)

11.6. Physical Parameters of Emulsions

Density. The simplest method of determining the density is to use a Mohr's balance. This can provide reasonable values even for viscous emulsions. The density of low-viscosity emul-

sions can also be determined with a pycnometer or a hydrometer.

Viscosity. The viscosity of emulsions is measured for such purposes as quality control [397] or rheological studies. It can be determined with conventional instruments such as the Höppler viscometer [398] or the rotational viscometer [4, p. 217], [398], [399]. Capillary viscometers can also be used [38, p. 483 and 478]. The surface viscosity can be measured with several methods [383]. The generation of low-frequency longitudinal waves on the surface of a film balance [185] produces usable, self-consistent results [402], [403].

Zeta Potential. Various methods are known for determining the zeta potential ξ [192], [404], [405]. A widespread method is to measure the migration velocity θ of an electrically charged droplet in an electric field (E). The emulsion droplets are usually negatively charged and migrate, with their electric double layer and diffuse ion cloud, toward the anode. The zeta potential is calculated from the following equation:

$$\zeta = \frac{4\pi\eta\vartheta}{\varepsilon E} \tag{17}$$

where ε is the dielectric constant and η the viscosity of the external phase.

This method can also be carried out on a microscopic scale [399]. In low-viscosity fine emulsions, the Brownian motion often interferes with the migration, thus making quantitative determination of the zeta potential inaccurate. However, determination of the sign of the charge from the direction of droplet migration is always easy.

Spectrometry. IR and UV spectrometry is used widely [406] for process control and for determining the chemical structure of unknown emulsifiers. Aromatic groups can readily be determined quantitatively by UV spectrometry, whereas ester and amine groups are easily determined quantitatively by IR spectrometry. For comparison, collections of emulsifier spectra are available [407], [408]. An example of a spectrometric determination of an increase in solubility caused by emulsifier addition is shown in Figure 3.

Conductivity. Conductivity measurements can be used to determine the emulsion type (Section 11.2), the PIT (Section 9.2) [409], or for quality control [410].

X-ray Diffraction. Both wide-angle and small-angle measurements are employed to investigate the crystalline [38, p. 227] or liquid – crystalline properties of emulsifiers or emulsifier – solvent mixtures. These methods allow the investigation of the mesomorphic phases that occur between an emulsifier and water or some other solvent [411], [412].

Phase Diagram. The theoretical basis for three-component phase diagrams has been developed predominantly by Kahlweit and coworkers [413], [414].

12. References

General References

1. D. O. Shah (ed.): *Macro- and Microemulsions, Theory and Application,* ACS Symp. Ser. 272, American Chemical Society, Washington, D.C., 1985.
2. Society of Chemical Industry (Colloid and Surface Chemistry Group): *Formation of Liquid-Liquid Dispersions – Chemical and Engineering Aspects,* Chameleon Press, London 1984.
3. P. Becher (ed.): *Encyclopedia of Emulsion Technology,* M. Dekker, New York and Basel 1983.
4. P. Becher: *Emulsion, Theory and the Technology,* Reinhold, New York 1965.
5. H. Stache (ed.): *Tensid Taschenbuch (Surfactant Pocket Book),* 2nd ed., C. Hanser, München 1981.
6. *Ullmann,* 4th ed., **10,** 449
7. A. W. Adamson: *Physical Chemistry of Surfaces,* 3rd ed., Wiley-Interscience, New York 1976
8. G. Gawalek: *Tenside,* Akademie-Verlag, Stuttgart 1975.
9. K. J. Lissant: *Demulsification,* M. Dekker, New York 1983.
10. K. J. Lissant: *Emulsions and Emulsion Technology,* vols. I and II, M. Dekker, New York 1974.
11. P. Sherman: *Emulsion Science,* Academic Press, London – New York 1968.

12. W. Clayton, C. G. Sumner: *The Theory of Emulsions and their Technical Treatment,* 5th ed., J. and A. Churchill, London 1954, 6th edition in preparation.
13. E. Manegold: *Emulsionen,* Verlag Straßenbau, Chemie und Technik, Heidelberg 1952.
14. G. M. Sutheim: *Introduction to Emulsions,* Chem. Publ. Co., New York 1947.
15. H. Bennett, J. L. Bishop, M. F. Wulfinghoff: *Practical Emulsions,* 3rd ed., vols. I and II, Chem. Publ. Co., New York 1968.
16. *Houben-Weyl, Methoden der organischen Chemie,* Thieme, Stuttgart 1959.
17. J. Stauff: *Kolloidchemie,* Springer, Berlin 1960.
18. J. P. Sisley: *Encyclopedia of Surface Active Agents,* Chem. Publ. Co., New York 1964.
19. A. M. Schwartz, J. W. Perry: *Surface Active Agents,* Interscience, New York 1958.
20. K. Lindner: *Textilhilfsmittel und Waschrohstoffe,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1971.
21. J. L. Moilliet, B. Collie, W. Black: *Surface Activity,* 2nd ed., van Nostrand, Princeton 1961.
22. H. Bueren, H. Grossmann: *Grenzflächenaktive Substanzen,* Verlag Chemie, Weinheim 1971.
23. A. Davidson, B. M. Milwidsky: *Synthetic Detergents,* Chem. Rubber Comp., Cleveland, Ohio, 1972.
24. K. L. Mittal (ed.): *Micellation, Solubilization and Microemulsions,* Plenum Press, New York 1977.
25. W. Scheer: *Zur Nomenklatur der Textilhilfsmittel, Leder- und Pelzhilfsmittel, Papierhilfsmittel, Gerbstoffe und Waschrohstoffe,* Verband Tegewa, Frankfurt/Main 1970.
26. C.I.D., Commission de Terminologie (ed.): *Vocabulary of Surface Active Agents,* 2nd ed.,1972.
27. H. J. Schick (ed.): *Nonionic Surfactants,* Surfactant Science Series, vol. 23, M. Dekker, New York 1967, new edition 1987.
28. E. Matijević (ed.): *Surface and Colloid Science,* vols. 1 – 6, Wiley – Interscience, New York 1969 to 1973.
29. J. Falbe, M. Hasselrodt: *Katalysatoren, Tenside und Mineralöladditive,* Thieme Verlag, Stuttgart 1978.
30. H. R. Kruyt (ed.): *Colloid Science,* vol. 2, Elsevier, Amsterdam 1969.
31. W. D. Harkins: *Physical Chemistry of Surface Films,* Reinhold, New York 1952.
32. J. W. McBain, E. Hutchinson: *Solubilization and Related Phenomena,* Academic Press, London – New York 1955.
33. H. Frotscher: *Chemie und physikalische Chemie der Textilhilfsmittel,* VEB Verlag Technik, Berlin 1955.
34. W. C. Griffin, M. J. Lynch in Th. E. Furia (ed.): *Handbook of Food Additives,* Chem. Rubber Comp., Cleveland, Ohio, 1968, pp. 413 – 446.
35. GDCh (Society of German Chemists) (ed.): *Specialist Group on Foodstuffs and Forensic Chemistry,* GDCh, Frankfurt/Main 1972.
36. K. L. Mittal, B. Lindmann (eds.): *Surfactants in Solution,* vols. 1 – 3, Plenum Press, New York 1984.
37. K. L. Mittal, E. J. Fendler (eds.): *Solution Behavior of Surfactants: Theoretical and Applied Aspects,* **vols. 1 and 2,** Plenum Press, New York 1982.
38. A. Kuhn: *Kolloidchemisches Taschenbuch,* Akademische Verlagsgesellschaft, Leipzig 1960.
39. DIN 53 900, July 1972; ISO, R 862, item 88, 89.

Specific References

40. E. Sheppard, N. Tcheurekdijian, *J. Colloid Interface Sci.* **62** (1977) 564.
41. J. Bokström, J. Sjobern, P. Stenius, T. Wärnheim, H. S. Fogler, *Tenside* **23** (1986) 125.
42. D. Langevin, *Mol. Cryst. Liq. Cryst.* **138** (1986) 259.
43. I. D. Robb (ed.): *Microemulsions,* Plenum Press, New York 1982.
44. S. E. Friberg, *J. Dispersion Sci. Technol.* **6** (1985) 317.
45. S. Friberg, K. Larsson, *Adv. Liq. Cryst.* 1976, no. 2, 173.
46. D. C. Führer, *Fette Seifen Anstrichm.* **75** (1973) 634.
47. A. Viallard, *Genie Chim.* **84** (1960) no. 3, 82.
48. S. A. Nikitina, O. S. Mochalova, *Kolloidn. Zh.* **30** (1968) 264; Ref.: *Kolloid Z. Z. Polym.* **234** (1969) 1067.
49. K. Rödel, H. Sonntag, *Tenside* **7** (1970) 135.
50. R. Heusch, *Progr. Colloid Polym. Sci.* **69** (1984) 29.
51. Bayer, DE 1 094 523, 1960.
52. G. Rimlinger, *Am. Perfum. Cosmet.* **82** (1967) no. 12, 31.
53. G. Rimlinger, *Parfums Cosmet. Savons* **13** (1970) no. 8, 605.
54. W. C. Vosburgh, G. R. Cooper, *J. Am. Chem. Soc.* **63** (1941) 437.

55. R. E. Wachtel, V. K. La Mer, *J. Colloid Sci.* **17** (1962) 531.
56. A. H. Brown, C. Hanson, *Nature (London)* **214** (1967) 77.
57. B. V. Deryagin, S. M. Levi: *Film Coating Theory,* Focal Press, London – New York 1964.
58. B. V. Deryagin, *Acta Physicochim. USSR* **20** (1945) 6.
59. J. J. Frenkel: *Kinetische Theorie der Flüssigkeiten,* VEB, Deutscher Verlag der Wissenschaften, Berlin 1957, pp. 353, 363.
60. P. H. Elworth, A. T. Florence, J. Rogers, *J. Colloid Interface Sci.* **35** (1971) 23.
61. P. Sherman, *J. Colloid Interface Sci.* **45** (1973) 427.
62. B. Dobias, *Tenside Deterg.* **15** (1978) 225; **16** (1978) 289.
63. J. T. Davies, *IInd Internat. Congress on Interfacially Active Substances,* London 1957, vol. I, p. 426.
64. H. Lange, C. P. Kurzendörfer, *Fette Seifen Anstrichm.* **76** (1974) no. 3, 120.
65. J. H. Schulman, E. G. Cockbain, *Trans. Faraday Soc.* **36** (1940) 651.
66. L. N. Mukerjee, S. D. Shukla, *Indian J. Appl. Chem.* **32** (1969) 130.
67. F. Lachampt, R. M. Vila, *Parfums Cosmet. Savons* **10** (1967) 372.
68. H. V. Tartar, R. E. Lothrop, G. F. Pettengill, *J. Phys. Chem.* **34** (1930) 375.
69. G. Boehmke, R. Heusch, *Fette Seifen Anstrichm.* **62** (1960) 87.
70. R. Heusch, *Tenside Deterg.* **21** (1984) 298.
71. DIN 53 917, Jan. 1981.
72. H. Stache, K. Kosswig, *Tenside* **18** (1981) 1.
73. W. Seifriz, *J. Phys. Chem.* **29** (1925) 738.
74. J. B. Parke, *J. Chem. Soc. II* 1934, 1112.
75. R. Heusch, *Ber. Bunsenges. Phys. Chem.* **82** (1978) 970; **83** (1979) 834; **88** (1984) 1093.
76. R. Heusch, *Tenside Deterg.* **21** (1984) 4.
77. H. Hoffmann in: *Aufarbeitung biologischer Medien* (Work-up of Biological Media), Publication No. 7 of GBF (German Society for Biotechnology Research), 1984, p. 95.
78. R. Heusch, F. Kopp, *Progr. Colloid Polym. Sci.,* in print (1987).
79. R. Heusch, *Chem. Ing. Tech.* **55** (1983) 608.
80. H. Lange, *IVth International Congress on Interfacially Active Substances,* Köln 1960, vol. I, p. 282.
81. G. Koerner, G. Rossmy, G. Säger, *Goldschmidt Informiert* **29** (1974) no. 2, 13.
82. A. M. Cazobal, D. Langevin, J. Meunier, A. Pouchelon, *VIIIth International Congress on Interfacially Active Substances,* Abstracts, Stockholm 1979, p. 84.
83. M. J. Schwuger, W. D. Hell, *Prog. Colloid Polym. Sci.* **65** (1978) 68.
84. J. M. G. Lankveld, Dissertation, Agricultive University Wageningen, Netherlands, 1970.
85. J. M. G. Lankveld, J. Lyklema, *J. Colloid Interface Sci.* **41** (1972) 454 – 483.
86. C. Marangoni, *Ann. Phys.* **143** (1871) 337. *Nuovo Chim.* **2** (1971) 239; **3** (1978) 97.
87. M. van den Tempel, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. 2, p. 573.
88. M. van den Tempel, *IInd Scandinavian Symp. Surface Active Processes,* (1964) p. 306.
89. J. T. Davies, E. K. Redeal: *Interfacial Phenomena,* 2nd. ed., Academic Press, New York 1963, pp. 410 – 413.
90. K. J. Hüttinger, private communication.
91. T. G. Sastry, S. N. Srivastava, *Z. Phys. Chem. (Leipzig)* **249** (1972) 101.
92. S. A. Nikitina, *VIth Intern. Congress on Interfacially Active Substances,* Zürich 1972, vol. II/2, p. 509.
93. E. G. Richardson, *J. Colloid Sci.* **5** (1950) 404.
94. H. Sonntag, K. Strenge: *Koagulation und Stabilität disperser Systeme,* VEB Deutscher Verlag der Wissenschaften, Berlin 1970.
95. H. von Helmholz, *Wiland American* **7** (1924) 508.
96. O. Stern, *Z. Elektrochem.* **30** (1924) 508.
97. W. J. Moore: *Physical Chemistry,* 5th ed., Longmans, London 1972. W. J. Moore, D. O. Hummel: *Physikalische Chemie,* 4th ed., de Gruyter, Berlin 1973, p. 548.
98. H. Sonntag, *Z. Phys. Chem (Leipzig)* **221** (1962) 365; **225** (1964) 284.
99. B. Koglin, *Chem. Ing. Tech.* **46** (1974) 720.
100. B. V. Derjaguin, L. D. Landau, *Acta Physicochim. USSR* **14** (1941) 633.
101. H. G. Elias, *Chem. Ing. Tech.* **36** (1964) 504.
102. R. Steinstrasser, *Kontakte* **1** (1971) 3.
103. G. H. Brown, J. W. Doane, *Appl. Phys.* **4** (1974) no. 1, p. 1.
104. S. Friberg, *J. Colloid Interface Sci.* **37** (1971) 291.
105. S. Friberg, *J. Am. Oil Chem. Soc.* **48** (1971) 578.
106. S. Friberg, L. Rydhag, *Kolloid Z. Z. Polym.* **244** (1971) 233.
107. V. N. Izmajlova, *Kolloidn. Zh. (USSR)* **34** (1972) 340; *Colloid Polym. Sci.* **252** (1974) 171.
108. C. D. Moore, *J. Soc. Cosmet. Chem.* **11** (1960) no. 1, 13.

109. S. Friberg, J. Wilton, *Am. Perfum. Cosmet.* **85** (1970) no. 12, 27.
110. L. Rydhag, S. Friberg, *VIth Intern. Congress on Interfacially Active Substances,* Zürich 1972, vol. II/2, p. 483.
111. R. Heusch, *Vth Intern. Congress on Interfacially Active Substances,* Barcelona 1968, vol. II, p. 1007.
112. D. C. Führer, *Fette Seifen Anstrichm.* **75** (1973) 634.
113. J. Swarbrick, *J. Soc. Cosmet. Chem.* **19** (1968) 187.
114. G. Reiss, G. Hurtvoz: *Encyclopedia of Polymer Science and Engineering,* 2nd. ed., Wiley-Interscience 1985, vol. II, pp. 324 – 434.
115. F. Hölscher, H. Reinhard: *Dispersionen synthetischer Hochpolymerer,* parts I and II, Springer, Berlin 1969.
116. R. H. Ottevile, C. H. Rocheser, B. R. Smith (eds.): *Adsorption from Solution,* Academic Press, New York – London 1983.
117. Th. F. Tadros (ed.): *The Effect of Polymers on Dispersion Properties,* Academic Press, New York 1982.
118. J. Lyklema, *Adv. Colloid Interface Sci.* **2** (1968) 65.
119. R. Menold, *Phys. Bl.* **33** (1977) 203.
120. R. D. Vold, K. L. Mittal, *J. Colloid Interface Sci.* **42** (1973) 436.
121. T. M. Riddick, *Chem. Eng. (N.Y.)* **68** (1961) 121, 141.
122. J. Boyd, C. Parkinson, P. Sherman, *J. Colloid Interface Sci.* **41** (1972) 359.
123. B. T. Lapan, R. W. Kucher, V. D. Enal'ev, L. G. Zhilina, *Colloid J. USSR (Engl. Transl.)* **35** (1973) 156.
124. T. de Vringer, J. G. H. Joosten, H. E. Junginger, *Colloid Polym. Sci.* **264** (1986) 623; **262** (1984) 56.
125. W. Noll, W. Büchner, H. H. Steinbach, C. Sucker, *Kolloid Z. Z. Polym.* **250** (1972) 9.
126. G. J. Tayler, *Proc. R. Soc. London Ser. A* **138** (1932) 41; **146** (1934) 501.
127. W. Holley: *Untersuchungen zum kontinuierlichen Emulgierprozeß in laminarer Schleppströmung unter Berücksichtigung der Emulgator – Adsorptionskinetik (Studies of the continuous emulsification process in laminar entrained flows allowing for the emulsifier adsorption kinetics),* Thesis presented to the Faculty of Chemical Engineering, University of Karlsruhe,1984.
128. B. Koglin, J. Pawlowski, H. Schnöring, *Chem. Ing. Tech.* **53** (1981) no. 8, 641.
129. A. Finiel, *Genie Chim.* **83** (1960) 183.
130. K. Saphir, *Seifen Öle Fette Wachse* **97** (1971) 811.
131. M. Seiller, T. Legrass, J. P. Treguier, F. Puisieux, J. Lo, *Pharm. Acta Helv.* **47** (1972) 516.
132. H. A. Troesch, *Chem. Ing. Tech.* **31** (1959) 667.
133. R. Rajagopalan, Ch. Tien, *Can. J. Chem. Eng.* **51** (1973) 272.
134. K. G. Ludwig, *Seifen Öle Fette Wachse* **98** (1972) 224.
135. L. N. Mukerjee, R. G. Sharma, *Indian J. Appl. Chem.* **31** (1968) 164.
136. S. Matsumoto, Y. Kita, D. Yonezawa, *J. Colloid Interface Sci.* **57** (1976) 353.
137. W. Seifriz, *J. Phys. Chem.* **29** (1925) 738.
138. J. B. Parke, *J. Chem. Soc. II* 1934, 1112.
139. J. H. Schulmann, E. G. Cockbain, *Trans. Faraday Soc.* **36** (1940) 651.
140. J. E. Bowcott, J. H. Schulman, *Z. Elektrochem.* **59** (1955) 283.
141. K. Shinoda, H. Kunieda, *J. Colloid Interface Sci.* **42** (1973) 381.
142. C. Solars, J. G. Dominguez, S. E. Friberg, *J. Dispersion Sci. Technol.* **6** (1985) 523.
143. A. W. Adamson, *J. Colloid Interface Sci.* **29** (1969) 261.
144. S. Friberg, *J. Am. Oil Chem. Sci.* **48** (1971) 578.
145. S. J. Ahmad, K. Shinoda, S. Friberg, *J. Colloid Interface Sci.* **47** (1974) 32.
146. D. O. Shah, R. M. Hamlin, Jr., *Science* **171** (1971) 483.
147. B. W. Barry, G. M. Saunders, *J. Colloid Interface Sci.* **41** (1972) 331.
148. P. H. Elworthy, A. T. Florence, J. Rogers, *J. Colloid Interface Sci.* **35** (1971) 34.
149. M. Dakkouri, *Maschinenmarkt* **91** (1985) 499.
150. K. D. Jain, M. K. Shamara, *J. Indian Chem. Soc.* **49** (1972) 302.
151. IWL Forum 1972/II – III: *Spaltung von Emulsionen (Breaking of Emulsions).* Institut für Gewerbliche Wasserwirtschaft und Luftreinigung e.V., D-5000 Köln 41, Oberländer Ufer 84 – 88.
152. W. Klin, *Angew. Chem.* **65** (1953) 201.
153. H. Thiele, J. Kowallik, *Kolloid Z. Z. Polym.* **234** (1969) 1071.
154. L. N. Mukerjee, S. N. Srivastava, *Kolloid Z.* **150** (1957) 144.
155. J. Leja, J. H. Schulman, *Trans. Faraday Soc.* **59** (1954) 598.
156. R. Wickbold, *Tenside* **11** (1974) 137.
157. W. K. Fischer, *Tenside* **8** (1971) 177.

158. R. D. Swisher: *Surfactant Biodegradation*, vol. 3 of *Surfactant Science Series,* M. Dekker, New York 1970.
159. *McCutcheon's Detergents and Emulsifiers,* published annually: 175 Rock Road, Green Rock NJ 07452, United States, telephone: (201) 652-2655.
160. *Tenside* **11** (1974) no. 1, A 14.
161. *J. Am. Assoc. Text. Chem. Col.* 1972, no. 4, 151.
162. Th. R. Paxton, *J. Colloid Interface Sci.* **31** (1969) 19.
163. W. Berlenbach, *Melliand Textilber.* **53** (1972) 207.
164. H. Kadmer, *Mineralöltechn.* **8** (1963) 1.
165. H. Reber, *Fette Seifen Anstrichm.* **76** (1974) 129.
166. O. Pauli, *TNO Nieuws* **22** (1967) 237.
167. M. Bierre, *Parfums Cosmet. Savons* **1** (1971) no. 1, 354.
168. W. D. Bancroft, *J. Phys. Chem.* **17** (1913) 501.
169. *Das Atlas HLB System,* Atlas Chemie GmbH, D-4300 Essen, Goldschmittstr. 100, 1968.
170. S. A. Nikitina, G. A. Simakova, *Colloid J. USSR (Engl. Transl.)* **31** (1969) 583.
171. N. Schönfeld: *Grenzflächenaktive Ethylenoxid-Addukte,* Wissenschaftliche Verlagsanstalt, Stuttgart 1976.
172. R. Heusch, *Makromol. Chem.* **182** (1981) 589.
173. W. C. Griffin, *J. Soc. Cosmet. Chem.* **5** (1954) 249.
174. R. Heusch, *Kolloid Z. Z. Polym.* **236** (1970) 31.
175. E. P. Budewitz, US 2 892 802, 1959.
176. L. Petkow, Seifen Öle Fette Wachse **106** (1980) 213.
177. C. D. Moore, M. Bell, *SPC Soap, Perfum. Cosmet.* **29** (1956) 893.
178. J. T. Davies, *IInd Internat. Congress on Interfacially Active Substances,* London 1957, vol. I, p. 426.
179. L. J. Lin, J. P. Friend, Y. Zimmels, *J. Colloid Interface Sci.* **45** (1973) 378.
180. C. Hansch, *Acc. Chem. Res.* **2** (1969) 232.
181. M. S. Tute in: *Advances in Drug Research,* vol. 6, Academic Press, London 1971.
182. J. M. Lehn, *Struktur Bindung* **16** (1973) 20.
183. K. Shinoda, *Vth Internat. Congress on Interfacially Active Substances,* Barcelona 1968, vol. II, p. 275.
184. H. Schott, *J. Pharm. Sci.* **58** (1969) 1443.
185. S. Matsumoto, P. Sherman, *J. Colloid Interface Sci.* **33** (1979) 294.
186. H. Saito, K. Shinoda, *J. Colloid Interface Sci.* **32** (1970) 647.
187. P. A. Winsor: *Solvent Properties of Amphiphilic Compounds,* Butterworths, London 1954.
188. A. Beerbower, J. Nixon, *Prep. Div. Pet. Chem. Am. Chem. Soc.* **14** (1969) 62.
189. A. Beerbower, M. W. Hill, *Am. Perfum. Cosmet.* **87** (1972) 85.
190. L. Marszall, *Riv. Ital. Essenze Profumi Piante Off. Aromi Saponi Cosmet. Aerosol* **15** (1973) 315.
191. R. Heusch, *Tenside Deterg.* **12** (1975) 81.
192. Don H. Moore: "Electrophoresis," in A. Weissberger (ed.): *Technique of Organic Chemistry,* vol. II, Interscience Publ., New York 1949, p. 168.
193. S. Ross, E. S. Chen, P. Becher, H. J. Ranauto, *J. Phys. Chem.* **63** (1959) 1681.
194. J. Bonadeo, *Tenside Deterg.* **13** (1976) no. 1, 53.
195. E. Orban, *Tenside* **7** (1970) 203.
196. N. Ohba, *Bull. Chem. Soc. Jpn.* **35** (1962) 1016.
197. H. L. Greenwald, G. L. Brown, M. N. Finemann, *Anal. Chem.* **28** (1956) 1693.
198. J. Middelton, *J. Soc. Cosmet. Chem.* **19** (1968) 129.
199. E. Götte, *Tenside* **4** (1967) 209.
200. E. Penzel, K. Oppenländer, *Tenside* **11** (1974) 129.
201. Bayer, DE-AS 1 419 677, 1960 (M. Quadvlieg, W. Hagge, G. Boehmke, H. Geffers).
202. *Fette Seifen Anstrichm.* **75** (1973) 65.
203. G. Lietz, *Tenside* **9** (1972) 76, 125.
204. *Eur. Chem.* 1969, no. 5, 4.
205. K. Schnurrbusch, *Seifen Öle Fette Wachse* **100** (1974) 173.
206. N. H. Nash, L. M. Brickman, *J. Am. Oil Chem. Soc.* **49** (1972) 459.
207. N. Krog, *J. Am. Oil Chem. Soc.* **54** (1977) 124.
208. W. Hansen, N. Zöllner, *Fette Seifen Anstrichm.* **73** (1971) 678.
209. *Winnacker-Küchler:Chemische Technologie,* 4th ed., **vol. 7/III,** C. Hanser, München 1981 – 86.
210. E. Ullmann, *Pharmazie* **20** (1965) 179.
211. H. Bertsch, F. Püschel, E. Ulsperger, *Tenside* **2** (1965) 397.
212. K. Thoma, *Goldschmidt Informiert* **26** (1973) no. 5, 8.
213. K. Thoma, *Goldschmidt Informiert* **32** (1975) no. 1, 22 – 31.
214. Z. Gruntova, *Farm. Obz.* **37** (1968) 273.
215. J. P. Redston, W. F. Bernholz, R. C. Natha, *J. Am. Oil Chem. Soc.* **48** (1971) 344.

216. K. Seib, *Seifen Öle Fette Wachse* **101** (1975) 299.
217. K. Thoma, *Goldschmidt Informiert* **1** (1975) no. 32, 22.
218. W. Hansen, N. Zöllner, *Fette Seifen Anstrichm.* **73** (1971) 678.
219. J. Schormüller, *IIIrd Intern. Congress on Interfacially Active Substances,* Köln 1960, vol. IV., p. 550.
220. R. Wegeler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel*, vol. I, Springer, Berlin 1970.
221. H. Niessen, *Pflanzenschutz Nachr. Bayer Ger. Ed.* **27** (1974) no. 1, 76.
222. O. Telle, *Pflanzenschutz Nachr. Bayer Ger. Ed.* **23** (1970) no. 2, 118.
223. H. L. Sanders, E. A. Knaggs, M. L. Nussbaum, US 2 696 453, 1954.
224. Bayer DE 1 094 523,1960.
225. Emulsol Chem. Corp., GB 815 001, 1959.
226. B. Kuchenbuch, *Fette Seifen Anstrichm.* **77** (1975) 407.
227. K. G. Ludwig, W. C. Gakenheimer, *Fette Seifen Anstrichm.* **70** (1968) 567.
228. W. Siebel, A. Menger, G. Hampel, H. Stephan, *Brot Gebäck* **22** (1968) 193.
229. R. Hüttinger, *Goldschmidt Informiert* **18** (1972) no. 1, 8.
230. K. G. Ludwig, *Fette Seifen Anstrichm.* **71** (1969) 672.
231. K. G. Ludwig, W. C. Gakenheimer, *Fette Seifen Anstrichm.* **69** (1967) 285.
232. G. F. Giannisi, *Tenside* **2** (1965) 40.
233. H. Forsters: *Chemie und physikalische Chemie der Textilhilfsmittel,* VEB Verlag, Berlin 1954 and 1955.
234. *Bayer Farben Rev., Special Issue* 1971, no. 13.
235. Bayer, DE-OS 2 142 318, 1971.
236. KAO SOAP, JP 48 053 099, 1971.
237. Allied Chem. Corp., US 3 781 202, 1972.
238. J. P. Redston, W. F. Bernholz, R. C. Natha, *J. Am. Oil Chem. Soc.* **48** (1971) 344.
239. H. G. Fröhlich, S. Müller, M. Rieber, *Chemiefasern Text. Anwendungstech.* **23** (1973) no. 8, 729.
240. H. Rupprecht, *Acta Pharmazeut. Tech. (Stuttgart)* **24** (1978) 105.
241. Hoechst, DE-OS 2 231 147, 1972.
242. Kanebo Co., JP 73 022 160, 1969.
243. K. Jakobs, *Text. Prax.* **28** (1973) 521.
244. US Defensive Publication T 912002,1972.
245. W. Dietsche, *Fette Seifen Anstrichm.* **75** (1973) 415.
246. Chem. Fabr. Pfersee, DE-OS 2 240 597, 1972.
247. A. Reng, R. Faber, J. M. Quack, *Tensid-Taschenbuch,* 1st ed., C. Hanser Verlag, München 1978, p. 153.
248. Atlas Kosmetik Kit, *Fette Seifen Anstrichm.* **75** (1973) 65.
249. J. M. Quack, A. K. Reng, *Parfüm. Kosmet.* **56** (1975) 157.
250. M. L. Schlossmann, *Soap. Cosmet. Chem. Spec.* **52** (1976) 33.
251. F. F. Hartmann, *Chemiefasern Text. Anwendungstech.* **24** (1974) 207.
252. W. Langmann, *Tenside* **7** (1970) 198.
253. W. Postman, *Text. Res. J.* **50** (1980) 444.
254. F. Püschel, *Tenside* **4** (1967) 1.
255. F. M. Menger, J. U. Rhee, H. K. Rhee, *J. Org. Chem.* **40** (1975) no. 25, 3803.
256. C. D. Moore, R. B. Hardwick, *Ind. Chem.* **40** (1964) 145.
257. Ch. P. Roe, *Ind. Eng. Chem.* **60** (1968) no. 9, 20.
258. A. S. Dunn, *Chem. Ind. (London)* **49** (1971) 1406.
259. H. Schuller, *Chem. Ing. Tech.* **45** (1973) 1123.
260. R. M. Panitsch, N. N. Iwanowa, S. S. Wojuzkij, *Kolloid Z. Z. Polym.* **232** (1969) 669.
261. J. W. Vanderhoff et al., *Adv. Chem. Ser.* **34** (1962) 32.
262. Hoechst, DE-AS 1 185 971, 1959.
263. V. T. Samorodov, L. V. Kosmodem'janskij, E. A. Lazurhnin, *Kolloid Zh. (Moscow)* **40** (1978) no. 6, 1206.
264. R. M. Panitsch, N. N. Iwanowa, S. S. Wojuzkij, *Koll. Z.* **232** (1969) 669.
265. Chem. Systems Inc., 747 Third Avenue, New York NY, 10017, report no. 74 – 4 (1975) p. 89.
266. J. W. Vanderhoff, E. B. Bradford, H. L. Tarkowski, J. B. Schaffer, R. M. Wiley, *Adv. Chem. Ser.* **34** (1962) 32.
267. K. Rombusch, G. Maahs, *Angew. Makromol. Chem.* **34** (1973) 55.
268. A. Courts, *Chem. Ind. London* **19** (1973) 943.
269. "Coacervates" in A. M. Schwartz, J. W. Perry: *Surface Active Agents,* vol. II, Interscience, New York 1949, pp.232 and 701.
270. Champion Paper Co., DE-OS 2 310 820, 1973.
271. C. Banse, J. J. Betz, *Sci. Techn. Pharm.* **7** (1978) no. 2, 73 and 97.
272. M. Koislu, *Kagaku Sochi* **21** (1979) no. 3, 11.
273. Fuji Photo-Film, DE-OS 2 434 406, 1974.
274. Campion Paper Co., DE-OS 2 310 820, 1973.
275. G. Perrault, G. Duchesne, *J. Appl. Polym. Sci.* **18** (1974) 1295.
276. G. S. Bakhrakh, Yu. M. Malinskii, *Colloid J. USSR (Engl. Transl.)* **35** (1973) 398.
277. F. Hess, *Goldschmidt Informiert* **14** (1971) no. 1, 16.

278. J. H. Nicholas, D. H. Madews, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. IV, p. 499.
279. H. Kraus, H. Kleinert, *Tenside* **5** (1968) 336.
280. Hoechst, DE-OS 2 210 085, 1972.
281. US Defensive Publication T 912002, dated 1. 5. 72.
282. *Ullmann,* 4th ed., **15,** 589.
283. F. Dambacher, *Tenside* **5** (1968) 303.
284. E. Matijević, *Mitt. Kolloid Ges. (Neue Folge)* **13** (1974) 3.
285. *Ullmann,* 4th ed., **17,** 581.
286. F. J. Poschmann, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. IV, p. 111.
287. F. Müller, *Bayer Farben Rev.* **21** (1972) 53.
288. W. Kolb, *Wbl. Papierfabrikant* **105** (1977) 117.
289. R. E. Läuchli, *Chem. Rundsch.* **35** (1972) 1086.
290. D. G. Williams, *Tappi* **56** (1973) no. 12, 144.
291. US 3 733 290, 1970.
292. H. Lang, S. Strobel, H. Selnders, *Wbl. Papierfabrikant* **106** (1978) 265.
293. D. G. Williams, *Tappi* **56** (1973) no. 12, 144.
294. F. Dambacher, *Tenside* **5** (1968) 109.
295. G. Bechstein, *Zellst. Pap. (Leipzig)* **22** (1973) 337.
296. H. Ortner, R. F. Wood, H. Gartemann, *Wbl. Papierfabrikant* **103** (1975) 597.
297. *Ullmann,* 4th ed., **15,** 166.
298. M. Bergmann in W. Grassmann (ed.): *Handbuch der Gerbereichemie und Lederfabrikation,* 2nd ed., **vol. 3,** parts 1 and 2, Springer, Vienna 1955 and 1961.
299. G. Otto, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. IV, p. 62.
300. H. Herfeld, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. IV, p. 79.
301. D. Burton, Robertschaw, *Ist Internat. Congress on Interfacially Active Substances,* Paris 1954, vol. 2, p. 547.
302. P. I. Levenko, Legkaya Industriya, Moscow 1974, from *J. Am. Leather Chem. Assoc.* **72** (1977) 459.
303. Bayer AG (ed.): *Gerben-Färben-Zurichten,* FB-A/Leder, 4th ed., Leverkusen 1978.
304. J. Eaglestone, *Leather* **179** (1977) 4424.
305. H. Keller, *Leather Manuf.* **90** (1973) 17.
306. G. Otto: *Das Färben von Leder,* E. Roether-Verlag, Darmstadt 1962.
307. *Ullmann,* 3rd ed., **11,** 577.
308. Henkel, DE-OS 2 231 162, 1972.
309. G. A. Nowak, *Leder Häutemarkt Gerbereiwiss. Prax.* **23** (1973) 264.
310. K. Eitel, K. Faber, *Bayer Farben Rev., Special issue* (1974) no. 14, p. 41.
311. *Ullmann,* 3rd ed., **1,** 667.
312. *Ullmann,* 4th ed., **2,** 110.
313. J. N. Plasking, A. M. Okolovich, G. N. Nazarova, *Tenside* **2** (1965) 61.
314. W. Strauss, *Tenside* **3** (1966) 144.
315. V. M. Shtanko, P. P. Karyazin, Yu. M. Volkov, *Khim. Prom. (Moscow)* 1976, no. 8, 603.
316. *Seifen Öle Fette Wachse* **101** (1975) 594.
317. H. G. Klein, J. N. Meussdörfer, H. Niederprüm, *Metalloberfläche* **29** (1975) 559.
318. H. Hadert, *Seifen Öle Fette Wachse* **94** (1968) 519.
319. Sinclair Oil Corp., US 3 766 113, 1969.
320. G. Y. Brokaw, *J. Am. Oil Chem. Soc.* **37** (1960) 523.
321. H. M. Truax, I. A. McDonald, *J. Am. Oil Chem. Soc.* 37 (1960) 651.
322. US Defensive Publication, T 912002, dated 1. 5. 72.
323. R. G. Schulze, *IIIrd Internat. Congress on Interfacially Active Substances,* Köln 1960, vol. IV, p. 617.
324. Ch. M. Blair, *Chem. Ind.* **20** (1960) 538.
325. G. Boardman, F. Dawson, *Chem. Ind.* **32** (1969) 1062.
326. W. B. Stead, *Seifen Öle Fette Wachse* **95** (1969) 716.
327. E. G. Nottes, *Erdöl Kohle Erdgas Petrochem.* **22** (1969) 404.
328. KAO Soap, JP 4 853 099, 1971, and Kanebo Co., JP 7 322 160, 1969.
329. Allied Chem. Corp., US 3 781 202, 1972.
330. Rex & Co., US 3 876 391, 1975.
331. Bayer, DE-OS 2 854 540, 1978.
332. Bayer, DE-OS 2 854 437, 1978.
333. Rewo Chem., DE-OS 2 063 424, 1970.
334. B. M. Deryagin, S. M. Levi: *Film Coating Theory,* Focal Press, London – New York 1964.
335. V. I. Zelikman, S. M. Levi: *Making and Coating Photographic Emulsions,* Focal Press, London – New York 1964.
336. H. M. Barchet: *Chemie photographischer Prozesse,* WTB Pocket Book 31, Akademie-Verlag, Berlin 1965.
337. T. H. James: *The Theory of the Photographic Process,* 4th ed., McMillen Comp., New York – London 1977.
338. P. Glafkides: *Chimie et Photografiques,* Paul Montel, Paris 1967, p. 420.

339. A. S. Yalabik-Kas, *Sci. Techn. Pharm.* **10** (1985) 978.
340. Agfa-Gevaert AG, DE 1 171 738, 1963.
341. 3 M Company, DE 2 337 638, 1973.
342. Eastman Kodak, US 2 949 360, 1960.
343. Eastman Kodak, US 2 322 027, 1943.
344. Eastman Kodak, US 2 801 170, 1957.
345. M. Oguchi, Y. Kuwabara, *Nippon Shushin Gakkai Kaishi* **26** (1963) no. 2, 73.
346. J. M. Martinez-Moreno, C. Gomez-Herrera, C. Janer-Del Valle, R. Carvaca-Barroso, *Grasas Aceites (Seville)* **13** (1962) 61.
347. S. Ikeno, K. Kaji, *Kogyo Kagaku Zasshi* **62** (1959) 47; Ref.: *Chem. Zentralbl.* 1960, 3662.
348. C. A. Bunton, R. J. Rubin, *Tetrahedron Lett.* 1975, no. 1, 55.
349. K. Arei, Y. Ogiwara, *Bull Chem. Soc. Jpn.* **51** (1978) no. 1, 182.
350. C. A. Bunton, Y. Ihara, J. L. Wright, *J. Org. Chem.* **41** (1976) no. 15, 2520.
351. J. M. Kapura, B. C. Gates, *Ind. Eng. Chem. Prod. Res. Dev.* **12** (1973) 62.
352. M. Grätzel, *Nachr. Chem. Tech. Lab.* **26** (1978) no. 8, 517.
353. S. Z. Monsa, D. B. Russel, E. F. Thoma, *Text. Chem. Color.* **5** (1973) 274.
354. BASF, US 4 608 188, 1985.
355. H. S. Yalabik-Kos, *Sci. Techn. Pharma* **10** (1985) 978.
356. WHO/M/13 (1965): ASTM-D 244 (1977): SAS-A 3742.
357. F. Gstirner, Ch. Gaudich, *Dtsch. Apoth. Ztg.* **112** (1972) 1583.
358. *Ullmann,* 4th ed., **2,** 205.
359. M. Kerker, *Am. Perfum Cosmet.* **85** (1970) no. 4, 43.
360. S. R. Reddy, H. S. Fogler, *J. Colloid Interface Sci.* **79** (1981) 101.
361. L. Marshall, *Estrato dalla Rivista Italiana, Essenze, Profumi, Piante Officinali, Aromi, Saponi, Cosmitici Aerosole* **50** (1974) no. 3, 133.
362. J. Georges, J. W. Chen, *Colloid Polym. Sci.* **264** (1986) 896.
363. P. Sherman, *SPC Soap Perfum Cosmet.* **44** (1971) no. II, 693.
364. E. A. Hauser, J. E. Lynn: *Experiments in Colloid Chemistry,* McGraw Hill, New York 1940.
365. H. Tronnier, H. Bussius, *Seifen Fette Öle Wachse* **86** (1960) 747.
366. T. B. Robertson, *Kolloid Z.* **7** (1910) 7.
367. T. Hamai, N. Koizumi, *Bull. Inst. Chem. Res. Kyoto Univ.* **40** (1962) 240.
368. J. P. Perl, H. E. Bussey, D. I. Sasan, *J. Colloid Interface Sci.* **108** (1985) 528.
369. E. K. Fischer, W. D. Harkins, *J. Phys. Chem.* **36** (1932) 98. H. Lotzkar, W. D. McLay, *Ind. Eng. Chem.* **35** (1943) no. 12, 1294.
370. Jb. Hvidberg, *Kolloid Z.* **72** (1935) 274. A. King, G. W. Wrzeszinski, *J. Chem. Soc. II* 1940, 1513.
371. S. Berkmann, *J. Phys. Chem.* **39** (1935) 527.
372. J. R. Cockton, J. B. Wynn, *J. Pharm. Pharmacol.* **4** (1952) 959.
373. H. Nassenstein, *Chem. Ing. Tech.* **29** (1957) 92.
374. IMANCO (Image Analysing Computers Ltd.), German Agency, Metalls Research GmbH, D-6050 Offenbach, Schweberstr. 18.
375. R. H. Berg, *Special Techn. Public. No. 234,* ASTM 1958.
376. R. E. Wachtel, V. K. La Mer, *J. Colloid Sci.* **17** (1962) 531.
377. *Ullmann,* 4th ed., **2,** 33.
378. W. J. Harris, *Br. J. Appl. Phys.* **12** (1961) 348.
379. R. Menold, W. Kaiser, B. Lüttge, A. Schmidt, *Chem. Ing. Tech.* **44** (1972) 1226.
380. L. Djaković, P. Dokić, F. Radvojecić, *Koll. Z. Z. Polym.* **244** (1971) 324.
381. P. Shermann, *Proc* . Soc. Anal. Chem. 6 (1969) 173.
382. N. Schwarz, C. Bezemer, *Kolloidchem. Makromol. Naturst. Verh. Diskuss. Hauptversamml. Koloid Ges.* **17** (1955) 139. R. D. Cadle: *Particle Size Determination,* Interscience, New York 1955.
383. R. Heusch, *Informationsdienst Arbeitsgem. Pharm. Verfahrenstech.* **18** (1972) 1.
384. F. Sauerwald, B. Schmidt, *Z. Anorg. Chem.* **213** (1933) 310.
385. K. H. Cuny, K. L. Wolf, *Ann. Phys. (Leipzig)* **17** (1956) Folge 6, no. 2 – 3, 57.
386. P. Lecomte du Noüy, *J. Gen. Physiol* **7** (1925) 525. G. Schwenker. *Ann. Phys.* (*Leipzig*) **11** (1931) no. 5, 525.
387. T. Kawanishi, T. Seimiya, T. Sasaki, *J. Colloid Interface Sci.* **32** (1970) 622. J. E. Lane, D. O. Jordan, *Austr. J. Chem.* **23** (1970) 2153.
388. L. Grunach, *Ann. Phys. (Leipzig)* **9** (1902) 4. Folge, no. 13, 1261.
389. C.I.E. 300-08-76 (French standard).
390. A. Capelle, *VIIIth Internat. Congress on Interfacially Active Substances,* Stockholm 1979, Abstracts, p. 103.
391. Manufacturer: Meßgeräte Werke Lauda, Dr. R. Wobster KG, D-6970 Lauda/Tauber, P.O. Box 40, F.R. of Germany.
392. J. Rahimian, H. J. Neumann, *Colloid Polym. Sci.* **256** (1978) 814.

393. H. Ackermann, *Fette Seifen Anstrichm.* **78** (1976) 37.
394. D. Harangozo, H. Suter, *Seifen Öle Fette Wachse* **96** (1970) 401.
395. M. Thomas-Collingnon, *Rev. Fr. Corps Gras* **22** (1975) 605.
396. *Ullmann,* 3rd ed., **7,** 79 *Ullmann,* 3rd ed., **5,**782, 793.
397. E. L. Roehl, *Seifen Öle Fette Wachse* **98** (1972) 568.
398. Gebrüder Hanke GmbH, D-7500 Karlsruhe 41, Dieselstr. 4, F.R. of Germany.
399. Contaves, P.O. Box 800 – 145, D-7000 Stuttgart 80, F.R. of Germany.
400. W. Flückiger, *Fette Seifen Anstrichm.* **69** (1967) 449.
401. H. Hühnerfuss, *J. Colloid Interface Sci.* **107** (1985) 84.
402. D. Langevin, *J. Colloid Interface Sci.* **80** (1981) 412.
403. P. Joos, R. Vanden Bogaert, *J. Colloid Interface Sci.* **56** (1976) 213.
404. T. M. Riddick, *Chem. Eng. (N.Y.)* **68** (1961) 121 and 141.
405. M. von Stackelberg, W. Kling, W. Benzel, F. Wilke. *Koll. Z.* **135** (1954) 67.
406. J. L. Carrion, S. Sagrado, M. de la Guardia, *Anal. Chim. Acta* **185** (1986) 101.
407. D. Hummel: *Analyse der Tenside,* Tafelband, Hanser Verlag, München 1962.
408. R. C. Gore, *I.R. Spectroscopy, Interpretation,* Work Book Hully, Merian, Pa., 1972; *Chem. Abstr.* **78** (1973) 166 873.
409. K. Kooswig, *Erdöl Erdgas Z.* **96** (1980) 318.
410. R. Brandau, K. W. Bold, *Fette Seifen Anstrichm.* **81** (1979) 366.
411. N. Krog, M. Sc. Brabrand, *Fette Seifen Anstrichm.* **73** (1971) 670.
412. St. Friberg, *Naturwissenschaften* **64** (1977) 612.
413. C. U. Herrmann, G. Klar, M. Kahlweit, *J. Colloid Interface Sci.* **82** (1981) 6.
414. M. Kahlweit, R. Strey, *Angew. Chem.* **8** (1985) 655; *Angew. Chem. Int. Ed. Engl.* **24** (1985) 654.

Enamel → Ceramics, Ceramic – Metal Systems

My Profile Log In

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Enzymes

Standard Article

Wolfgang Aehle[1], Richard N. Perham[48], Gerhard Michal[2], Albert Jonke[3], Gopal K. Chotani[4], Timothy C. Dodge[5], Andreas Herman Terwisscha van Scheltinga[21], Christian Gölker[6], Meng H. Heng[7], John Kan[8], Todd Becker[9], Saburo Fukui[10], Atsuo Tanaka[11], Hans de Nobel[12], Brian Jones[13], Rick Bott[14], Charlotte Horsmans Poulsen[15], Karsten M. Kragh[16], Jens Frisback Sorensen[17], Catherine Grassin[18], Marga Herweijer[19], Jan Wilms[20], André de Roos[22], Jorn Borch Soe[23], Ivan Herbots[24], Beatrix Kottwitz[25], Peter J. Reilly[26], Richard L. Antrim[27], Heidi Burrows[28], Herman B. M. Lenting[29], Liisa Viikari[30], Anna Suurnäkki[31], Marja-Leena Niku-Paavola[32], Johanna Buchert[33], Karl-Heinz Maurer[34], Andrea Saettler[35], Herbert Waldmann[36], Carsten Schultz[37], Harald Gröger[38], Carlo Dinkel[39], Karlheinz Drauz[40], Georg-Burkhard Kresse[41], Rainer Schmuck[42], Karl Wulff[43], Günther Henniger[44], Christoph Kessler[45], Alice J. Caddow[46], Beth Concoby[47]

[1]Genencor International B. V., Leiden, The Netherlands
[2]Tutzing, Germany
[3]Boehringer Mannheim GmbH, Penzberg, Germany
[4]Genencor International, Inc., Palo Alto, USA
[5]Genencor International, Inc., Palo Alto, USA
[6]Bayer AG, Wuppertal, Germany
[7]Genencor International, Inc., Palo Alto, USA
[8]Genencor International, Inc., Palo Alto, USA
[9]Genencor International, Inc., Palo Alto, USA
[10]Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
[11]Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
[12]Genencor International B. V., Leiden, The Netherlands
[13]Genencor International B. V., Leiden, The Netherlands
[14]Genencor International, Inc., Palo Alto, USA
[15]Danisco A/S, Brabrand, Denmark
[16]Danisco A/S, Brabrand, Denmark
[17]Danisco A/S, Brabrand, Denmark
[18]DSM Food Specialties, Seclin Codex, France
[19]DSM Food Specialties, Delft, The Nederlands, France
[20]Genencor International B. V., Leiden, The Nederlands
[21]Gibst-brocades N. V., Delft, The Nederlands
[22]DSM Food Specialities, Delft, The Netherlands
[23]Danisco A/S, Brabrand, Denmark
[24]Procter & Gamble Eurocor S.A., Strombeek-Bever, Belgium
[25]Henkel KGaA, Düsseldorf, Germany
[26]Iowa State University, Ames, USA
[27]Grain Processing Corporatio, Muscatine, Iowa, USA
[28]Finfeeds International, Marlborough, UK
[29]TNO Industry, Enschede, The Netherlands
[30]VTT Biotechnology, VTT, Finland
[31]VTT Biotechnology, VTT, Finland
[32]VTT Biotechnology, VTT, Finland
[33]VTT Biotechnology, VTT, Finland
[34]Henkel KGaA, Düsseldorf, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

[35]Henkel KGaA, Düsseldorf, Germany
[36]Max-Planck-Institut für molekulare Physiologie, Dortmund, Germany
[37]EMBL, Heidelberg, Germany
[38]Degussa AG, Hanau-Wolfgang, Germany
[39]SiChem GmbH, Bremen, Germany
[40]Degussa AG, Hanau-Wolfgang, Germany
[41]Roche Diagnostics GmbH, Penzberg, Germany
[42]Roche Diagnostics GmbH, Tutzing, Germany
[43]Boeringer Mannheim GmbH, Tutzing, Germany
[44]Roche Diagnostics GmbH, Penzberg, Germany
[45]Roche Diagnostics GmbH, Penzberg, Germany
[46]Genencor International, Inc., Palo Alto, USA
[47]Genencor International, Inc., Palo Alto, USA
[48]University of Cambridge, Department of Biochemistry, Cambridge, UK

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

Abstract

The article contains sections titled:

1. Introduction
1.1. History
1.2. Enzyme Nomenclature
1.2.1. General Principles of Nomenclature
1.2.2. Classification and Numbering of Enzymes
1.3. Structure of Enzymes
1.3.1. Primary Structure
1.3.2. Three-Dimensional Structure
1.3.3. Quaternary Structure, Folding, and Domains
1.3.4. The Ribozyme
1.4. Enzymes and Molecular Biology
1.4.1. Biosynthesis of Enzymes
1.4.2. Enzymes and DNA
2. Catalytic Activity of Enzymes
2.1. Factors Governing Catalytic Activity
2.1.1. Temperature
2.1.2. Value of pH
2.1.3. Activation
2.1.4. Inhibition
2.1.5. Allostery
2.1.6. Biogenic Regulation of Activity
2.2. Enzyme Assays
2.2.1. Reaction Rate as a Measure of Catalytic Activity
2.2.2. Definition of Units
2.2.3. Absorption Photometry
2.2.4. Fluorometry
2.2.5. Luminometry
2.2.6. Radiometry
2.2.7. Potentiometry
2.2.8. Conductometry
2.2.9. Calorimetry
2.2.10. Polarimetry

2.2.11. Manometry
2.2.12. Viscosimetry
2.2.13. Turbidimetry
2.2.14. Immobilized Enzymes
2.2.15. Electrophoresis
2.3. Quality Evaluation of Enzyme Preparations
2.3.1. Quality Criteria
2.3.2. Specific Activity
2.3.3. Protein Determination
2.3.4. Contaminating Activities
2.3.5. Electrophoretic Purity
2.3.6. High-Performance Liquid Chromatography
2.3.7. Performance Test
2.3.8. Amino Acid Analysis and Protein Sequence Analysis
2.3.9. Stability
2.3.10. Formulation of Enzyme Preparations
3. General Production Methods
3.1. Microbial Production
3.1.1. Organism and Enzyme Synthesis
3.1.2. Strain Improvement
3.1.3. Physiological Optimization
3.1.4. The Fermentor and its Limitations
3.1.5. Process Design
3.1.6. Modeling and Optimization
3.1.7. Instrumentation and Control
3.2. Isolation and Purification
3.2.1. Preparation of Biological Starting Materials
3.2.1.1. Cell Disruption by Mechanical Methods
3.2.1.2. Cell Disruption by Nonmechanical Methods
3.2.2. Separation of Solid Matter
3.2.2.1. Filtration
3.2.2.2. Centrifugation
3.2.2.3. Extraction
3.2.2.4. Flocculation and Flotation
3.2.3. Concentration
3.2.3.1. Thermal Methods
3.2.3.2. Precipitation
3.2.3.3. Ultrafiltration
3.2.4. Purification
3.2.4.1. Crystallization
3.2.4.2. Electrophoresis
3.2.4.3. Chromatography
3.2.5. Product Formulation
3.2.6. Waste Disposal
3.3. Immobilization
3.3.1. Definitions
3.3.2. History
3.3.3. Methods
3.3.3.1. Carrier Binding
3.3.3.2. Cross-linking
3.3.3.3. Entrapment
3.3.4. Characterization
3.3.5. Application

4. Discovery and Development of Enzymes
4.1. Enzyme Screening
4.1.1. Overview
4.1.2. Natural Isolate Screening
4.1.3. Molecular Screening
4.1.4. Environmental Gene Screening
4.1.5. Genomic Screening
4.1.6. Proteomic Screening
4.2. Protein Engineering
4.2.1. Introduction
4.2.2. Application of Protein Engineering in Academia and Industry
4.2.3. Outlook
5. Industrial Enzymes
5.1. Enzymes in Food Applications
5.1.1. Enzymes in Baking
5.1.1.1. Introduction
5.1.1.2. Amylases
5.1.1.3. Xylanases
5.1.1.4. Oxidoreductases
5.1.1.5. Lipases
5.1.1.6. Proteases
5.1.1.7. Transglutaminase
5.1.2. Enzymes in Fruit Juice Production and Fruit Processing
5.1.2.1. Introduction
5.1.2.2. Biochemistry of Fruit Cell Walls
5.1.2.3. Cell-Wall-Degrading Enzymes
5.1.2.4. Apple Processing
5.1.2.4.1. Apple Pulp Maceration
5.1.2.4.2. Apple Juice Depectinization
5.1.2.5. Red-Berry Processing
5.1.2.6. Tropical Fruit and Citrus Processing
5.1.2.7. Conclusion
5.1.3. Enzymes in Brewing
5.1.3.1. Introduction
5.1.3.2. Enzymes in Malting and Mashing
5.1.3.3. Enzymes for Problem Prevention or Solving
5.1.3.3.1. Bacterial *α*-Amylase in Mashing
5.1.3.3.2. Fungal *α*-Amylase in Fermentation
5.1.3.3.3. *β*-Glucanase in Mashing
5.1.3.3.4. Cysteine Endopeptidases (Postfermentation)
5.1.3.3.5. Glucoamylase in Mashing
5.1.3.4. Enzymes for Process Improvement
5.1.3.4.1. Adjunct Brewing
5.1.3.4.2. Improved Mashing Processes
5.1.3.4.3. Shelf-Life Improvement
5.1.3.4.4. Accelerated Maturation
5.1.3.4.5. Starch-Haze Removal
5.1.3.5. Special Brewing Processes
5.1.4. Enzymes in Dairy Applications
5.1.4.1. Introduction
5.1.4.2. Cheesemaking
5.1.4.2.1. Cheesemaking Process
5.1.4.2.2. Mechanism of Renneting

5.1.4.2.3. Types of Coagulants
5.1.4.2.4. Properties of Coagulating Enzymes
5.1.4.2.5. Cheese Ripening
5.1.4.2.6. Cheese Flavors and Ripening Acceleration
5.1.4.2.7. Lipase
5.1.4.2.8. Lysozyme
5.1.4.2.9. Milk Protein Hydrolysates
5.1.4.2.10. Transglutaminase
5.1.4.3. Milk Processing
5.1.4.3.1. *β*-Galactosidase
5.1.4.3.2. Other Enzymes
5.1.5. Other Food Applications
5.1.5.1. Introduction
5.1.5.2. Meat and Fish
5.1.5.2.1. Meat Processing
5.1.5.2.2. Fish Processing
5.1.5.3. Protein Cross-linking
5.1.5.4. Flavor Development
5.1.5.4.1. Protein Hydrolysis
5.1.5.4.2. Lipid Hydrolysis
5.1.5.5. Egg powder
5.1.5.6. Oils and Fats
5.1.5.6.1. Fat Splitting
5.1.5.6.2. Interesterification
5.1.5.6.3. Esterification
5.1.5.6.4. Oil Degumming
5.2. Enzymes in Nonfood Applications
5.2.1. Enzymes in Household Detergents
5.2.1.1. Historical Development
5.2.1.2. Laundry Soils
5.2.1.3. Detergent Composition and Washing Process
5.2.1.3.1. Washing Process
5.2.1.3.2. Detergent Compositions
5.2.1.4. Enzyme-Aided Detergency and Soil Removal
5.2.1.5. Detergent Enzyme Performance Evaluation and Screening
5.2.1.6. Enzyme Types
5.2.1.6.1. Proteases
5.2.1.6.2. Amylases
5.2.1.6.3. Lipases
5.2.1.6.4. Cellulases
5.2.1.6.5. Mannanase
5.2.1.7. Future Trends
5.2.2. Enzymes in Automatic Dishwashing
5.2.2.1. Introduction
5.2.2.2. Characteristics of Enzymes for ADDs
5.2.2.3. Proteases
5.2.2.3.1. Proteins: The Substrate of Proteases
5.2.2.3.2. Proteases for ADDs
5.2.2.4. Amylases
5.2.2.4.1. Starch: The Substrate of Amylases
5.2.2.4.2. Amylases for ADDs
5.2.2.5. Other Enzymes
5.2.2.6. Automatic Dishwashing Detergents

5.2.2.6.1. Composition of Automatic Dishwashing Detergents
5.2.2.6.2. Application of Enzymes in ADD
5.2.2.6.3. Stability and Compatibility
5.2.3. Enzymes in Grain Wet-Milling
5.2.3.1. Introduction
5.2.3.2. Overview of the Conversion of Corn to HFCS
5.2.3.2.1. Corn Steeping
5.2.3.2.2. Coarse Grinding and Germ Removal by Cyclone Separation
5.2.3.2.3. Fine Grinding and Fiber Removal by Screening
5.2.3.2.4. Centrifugation and Washing to Separate Starch from Protein
5.2.3.2.5. Hydrolysis with *α*-Amylase (Liquefaction)
5.2.3.2.6. Hydrolysis with Glucoamylase (Saccharification)
5.2.3.2.7. Isomerization with Glucose Isomerase
5.2.3.2.8. Fructose Enrichment and Blending
5.2.3.3. *α*-Amylase
5.2.3.3.1. Origin and Enzymatic Properties
5.2.3.3.2. Structure
5.2.3.3.3. Industrial Use
5.2.3.4. Glucoamylase
5.2.3.4.1. Origin and Enzymatic Properties
5.2.3.4.2. Structure
5.2.3.4.3. Industrial Use
5.2.3.5. Pullulanase
5.2.3.5.1. Origin and Enzymatic Properties
5.2.3.5.2. Structure
5.2.3.5.3. Industrial Use
5.2.3.6. Glucose Isomerase
5.2.3.6.1. Origin and Enzymatic Properties
5.2.3.6.2. Structure
5.2.3.6.3. Industrial Use of Glucose Isomerase
5.2.3.7. Use of Wheat Starch
5.2.4. Enzymes in Animal Feeds
5.2.4.1. Introduction
5.2.4.2. Enzymes Used in Animal Feed
5.2.4.2.1. Fiber-Degrading Enzymes
5.2.4.2.2. Phytic Acid Degrading Enzymes (Phytases)
5.2.4.2.3. Protein-Degrading Enzymes (Proteases)
5.2.4.2.4. Starch-Degrading Enzymes (Amylases)
5.2.4.3. Future Developments
5.2.5. Enzymes in Textile Production
5.2.5.1. Introduction
5.2.5.2. Cellulose Fibers
5.2.5.2.1. Desizing of Cotton Cellulose Fibers
5.2.5.2.2. Scouring of Cotton
5.2.5.2.3. Bleaching of Cotton
5.2.5.2.4. Removal of Hydrogen Peroxide
5.2.5.2.5. Cotton Finishing
5.2.5.2.6. Ageing of Denim
5.2.5.2.7. Processing of Man-Made Cellulose Fibers
5.2.5.2.8. Processing of Bast Fibers
5.2.5.3. Proteinous Fibers
5.2.5.3.1. Wool Processing
5.2.5.3.2. Degumming of Silk

5.2.5.4. Textile Effluent Treatment and Recycling
5.2.5.5. Outlook
5.2.6. Enzymes in Pulp and Paper Processing
5.2.6.1. Introduction
5.2.6.2. Enzymes
5.2.6.2.1. Cellulases
5.2.6.2.2. Hemicellulases
5.2.6.2.3. Lignin-Modifying, Oxidative Enzymes
5.2.6.3. Enzymes in Pulp and Paper Processing
5.2.6.3.1. Mechanical Pulping
5.2.6.3.2. Chemical Pulping
5.2.6.3.3. Bleaching
5.2.6.3.4. Papermaking
5.2.6.3.5. Deinking
5.3. Development of New Industrial Enzyme Applications
5.3.1. Introduction
5.3.2. Enzymes in Cosmetics
5.3.2.1. Hair Dyeing
5.3.2.1.1. Oxidases
5.3.2.1.2. Peroxidases
5.3.2.1.3. Polyphenol Oxidases
5.3.2.2. Hair Waving
5.3.2.3. Skin Care
5.3.2.4. Toothpastes and Mouthwashes
5.3.2.5. Enzymes in Cleaning of Artificial Dentures
5.3.3. Enzymes for Preservation
5.3.4. Enzymes in Hard-Surface Cleaning
5.3.4.1. Enzymes in Membrane Cleaning
5.3.4.1.1. Proteases
5.3.4.1.2. Hemicellulases
5.3.5. Enzymes Generating a pH Shift
5.3.6. Enzymes in Cork Treatment
5.3.7. Enzymes in Oil-Field Applications
5.3.8. Enzymes in Wastewater Treatment
5.3.9. Enzymes for Polymerisation: Wood Fiberboard Production
5.4. Overview of Industrial Enzyme Applications
6. Nonindustrial Enzyme Usage
6.1. Enzymes in Organic Synthesis
6.1.1. Introduction
6.1.2. Examples of Enzymatic Conversions
6.1.2.1. Syntheses by Means of Hydrolases
6.1.2.2. Reduction of CO and CC Bonds
6.1.2.3. Oxidation of Alcohols and Oxygenation of CH and CC Bonds
6.1.2.4. CC Coupling
6.1.2.5. Formation of Glycosidic Bonds
6.1.2.6. Enzymatic Protecting Proup Techniques
6.1.3. Enzyme-Analogous Catalysts
6.1.4. Commercial Applications
6.1.4.1. General
6.1.4.2. Nonstereoselective Biocatalytic Reactions
6.1.4.3. Biocatalytic Resolution Processes
6.1.4.3.1. Enzymatic Acylation of Amino Groups
6.1.4.3.2. Enzymatic Hydrolysis of Hydantoins

6.1.4.3.3. Enzymatic Hydrolysis of Lactams
6.1.4.3.4. Enzymatic Hydrolysis of CO Bonds
6.1.4.3.5. Enzymatic Hydrolysis of Nitriles
6.1.4.3.6. Enzymatic Cleavage of *threo* Aldol Products
6.1.4.4. Biocatalytic Asymmetric Synthesis
6.1.4.4.1. Biocatalytic Reductive Amination of CO Bonds
6.1.4.4.2. Biocatalytic Hydrocyanation of CO Bonds
6.1.4.4.3. Biocatalytic Addition of Water to CO Bonds
6.1.4.4.4. Biocatalytic Amination of CC Bonds
6.1.5. Outlook
6.2. Therapeutic Enzymes
6.2.1. Requirements for the Use of Enzymes in Therapy
6.2.2. Coping with Peculiar Protein Properties
6.2.3. Sources of Enzymes and Production Systems
6.2.4. Overview of Therapeutic Enzymes
6.2.4.1. Oxidoreductases
6.2.4.2. Transferases
6.2.4.3. Esterases
6.2.4.4. Nucleases
6.2.4.5. Glycosidases
6.2.4.6. Proteases
6.2.4.6.1. Pancreatic and Gastric Proteases
6.2.4.6.2. Plasma Proteases
6.2.4.6.3. Coagulation Factors
6.2.4.6.4. Plasminogen Activators
6.2.4.6.5. Proteases from Snake Venoms
6.2.4.6.6. Plant and Microbial Proteases
6.2.4.7. Amidases
6.2.4.8. Lyases
6.3. Enzymes in Diagnosis
6.3.1. Determination of Substrate Concentration
6.3.2. Determination of Enzyme Activity
6.3.3. Immunoassays
6.4. Enzymes for Food Analysis
6.4.1. Carbohydrates
6.4.2. Organic Acids
6.4.3. Alcohols
6.4.4. Other Food Ingredients
6.5. Enzymes in Genetic Engineering
6.5.1. Restriction Endonucleases and Methylases
6.5.1.1. Classification
6.5.1.2. Activity of Class II Restriction Endonucleases
6.5.1.2.1. Reaction Parameters
6.5.1.2.2. Additional Structural Requirements Influencing Activity
6.5.1.3. Specificity of Class II Restriction Endonucleases
6.5.1.3.1. Palindromic Recognition Sequences
6.5.1.3.2. Nonpalindromic Recognition Sequences
6.5.1.3.3. Isoschizomers
6.5.1.4. Changes in Sequence Specificity
6.5.1.5. Novel Class II Restriction Endonucleases
6.5.2. DNA Polymerases
6.5.2.1. *Escherichia coli* DNA Polymerase I
6.5.2.2. Klenow Enzyme

6.5.2.3. T4 DNA Polymerase
6.5.2.4. Reverse Transcriptase
6.5.2.5. Terminal Transferase
6.5.3. RNA Polymerases
6.5.3.1. SP6 RNA Polymerase
6.5.3.2. T7 RNA Polymerase
6.5.4. DNA Nucleases
6.5.4.1. DNase I
6.5.4.2. Exonuclease III
6.5.4.3. Nuclease S1
6.5.4.4. Nuclease Bal 31
6.5.5. RNA Nucleases
6.5.5.1. RNase H
6.5.5.2. Site-Specific RNases
6.5.5.2.1. RNase A
6.5.5.2.2. RNase CL3
6.5.5.2.3. RNase T_1
6.5.5.2.4. RNase U_2
6.5.5.2.5. Nuclease S7
6.5.5.2.6. Site-Specific RNases in RNA Sequence Analysis
6.5.6. Modifying Enzymes
6.5.6.1. Alkaline Phosphatase
6.5.6.2. T4 DNA Ligase
6.5.6.3. *Escherichia coli* DNA Ligase
6.5.6.4. T4 Polynucleotide Kinase
6.5.6.5. T4 Polynucleotide Kinase, 3′-Phosphatase-Free
6.5.6.6. Methylase *Hpa*II
7. Enzyme Safety and Regulatory Considerations
7.1. Safe Handling of Enzymes
7.1.1. Possible Health Effects
7.1.2. Control Technology
7.2. Product Regulatory Considerations
7.2.1. Food-Use Enzymes
7.2.2. Feed-Use Enzymes
7.2.3. Industrial-Use Enzymes

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Enzymes

Wolfgang Aehle, Genencor International B. V., Leiden, The Netherlands
Richard N. Perham, University of Cambridge, Department of Biochemistry, Cambridge, UK

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

1. Introduction

Enzymes are the catalysts of biological processes. Like any other catalyst, an enzyme brings the reaction catalyzed to its equilibrium position more quickly than would occur otherwise; an enzyme cannot bring about a reaction with an unfavorable change in free energy unless that reaction can be coupled to one whose free energy change is more favorable. This situation is not uncommon in biological systems, but the true role of the enzymes involved should not be mistaken.

The activities of enzymes have been recognized for thousands of years; the fermentation of sugar to alcohol by yeast is among the earliest examples of a biotechnological process. However, only recently have the properties of enzymes been understood properly. Indeed, research on enzymes has now entered a new phase with the fusion of ideas from protein chemistry, molecular biophysics, and molecular biology. Full accounts of the chemistry of enzymes, their structure, kinetics, and technological potential can be found in many books and series devoted to these topics [1-5]. This chapter reviews some aspects of the history of enzymes, their nomenclature, their structure, and their relationship to recent developments in molecular biology.

1.1. History

Detailed histories of the study of enzymes can be found in the literature [6], [7].

Early Concepts of Enzymes. The term “enzyme” (literally “in yeast”) was coined by KÜHNE in 1876. Yeast, because of the acknowledged importance of fermentation, was a popular subject of research. A major controversy at that time, associated most memorably with LIEBIG and PASTEUR, was whether or not the process of fermentation was separable from the living cell. No belief in the necessity of vital forces, however, survived the demonstration by BUCHNER (1897) that alcoholic fermentation could by carried out by a cell-free yeast extract. The existence of extracellular enzymes had, for reasons of experimental accessibility, already been recognized. For example, as early as 1783, SPALLANZANI had demonstrated that gastric juice could digest meat in vitro, and SCHWANN (1836) called the active substance pepsin. KÜHNE himself appears to have given trypsin its present name, although its existence in the intestine had been suspected since the early 1800s.

Enzymes as Proteins. By the early 1800s, the proteinaceous nature of enzymes had been recognized. Knowledge of the chemistry of proteins drew heavily on the improving techniques and concepts of organic chemistry in the second half of the 1800s; it culminated in the peptide theory of protein structure, usually credited to FISCHER und HOFMEISTER. However, methods that had permitted the separation and synthesis of small peptides were unequal to the task of purifying enzymes. Indeed, there was no consensus that enzymes were proteins. Then, in 1926, SUMNER crystallized urease from jack bean meal and announced it to be a simple protein. However, WILLSTÄTTER argued that enzymes were not proteins but “colloidal carriers” with “active prosthetic groups.” However, with the conclusive work by NORTHROP *et al.*, who isolated a series of crystalline proteolytic enzymes, beginning with pepsin in 1930, the proteinaceous nature of enzymes was established.

The isolation and characterization of intracellular enzymes was naturally more complicated and, once again, significant improvements were necessary in the separation techniques applicable to proteins before, in the late 1940s, any such enzyme became available in reasonable quantities. Because of the large amounts of accessible starting material and the historical importance of fermentation experiments, most of the first pure intracellular enzymes came from yeast and skeletal muscle. However, as purification methods were improved, the number of enzymes obtained in pure form increased tremendously and still continues to grow.

Methods of protein purification are so sophisticated today that, with sufficient effort, any desired enzyme can probably be purified completely, even though very small amounts will be obtained if the source is poor.

Primary Structure. After the protein nature of enzymes had been accepted, the way was clear for more precise analysis of their composition and structure. Most amino acids had been identified by the early 20th century. The methods of amino acid analysis then available, such as gravimetric analysis or microbiological assay, were quite accurate but very slow and required large amounts of material. The breakthrough came with the work of MOORE and STEIN on ion-exchange chromatography of amino acids, which culminated in 1958 in the introduction of the first automated amino acid analyzer [8].

The more complex question – the arrangement of the constituent amino acids in a given protein, generally referred to as its primary structure – was solved in the late 1940 s. The determination in 1951 of the amino acid sequence of the β-chain of insulin by SANGER and TUPPY [10] demonstrated for the first time that a given protein does indeed have a unique primary structure. The genetic implications of this were enormous. The introduction by EDMAN of the phenyl isothiocyanate degradation of proteins stepwise from the N-terminus, in manual form in 1950 and subsequently automated in 1967 [11], provided the principal chemical method for determining the amino acid sequences of proteins. The primary structures of pancreatic ribonuclease [12] and egg-white lysozyme [13] were published in 1963. Both of these enzymes, simple extracellular proteins, contain about 120 amino acids. The first intracellular enzyme to have its primary structure determined was glyceraldehyde 3-phosphate dehydrogenase [14], which has an amino acid sequence of 330 residues and represents a size (250 – 400 residues) typical of many enzymes. Protein sequencing is increasingly performed by liquid chromatography/mass spectrometry (LC/MS) techniques, and several tools and software packages are now available for protein identification and characterization. The methods of protein sequence analysis are now so well developed that no real practical deterrent exists, other than time or expense, to determination of the amino acid sequence of any polypeptide chain [9].

A more recent fundamental concept called proteome (*prote*in complement to a gen*ome*) will enable researchers to unravel biochemical and physiological mechanisms of complex multivariate diseases at the functional molecular level. A new discipline, *proteomics*, complements physical genome research. Proteomics can be defined as “the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes” [15].

Active Site. The fact that enzymes are highly substrate specific and are generally much larger than the substrates on which they act quickly became apparent. The earliest kinetic analyses of enzymatic reactions indicated the formation of transient enzyme – substrate complexes. These observations could be explained easily if the conversion of substrate to product was assumed to occur at a restricted site on an enzyme molecule. This site soon became known as the active center or, as is more common today, the active site.

Particular compounds were found to react with specific amino acid side chains and thus inhibit particular enzymes. This suggested that such side chains might take part in the catalytic mechanisms of these enzymes. An early example was the inhibition of glycolysis or fermentation by iodoacetic acid, which was later recognized as resulting from reaction with a unique cysteine residue of glyceraldehyde 3-phosphate dehydrogenase, which normally carries the substrate in a thioester linkage [16].

Many such group-specific reagents have now been identified as inhibitors of individual enzymes; often they are effective because of the hyper-reactivity of a functionally important side chain in the enzyme's active site. However, a more sophisticated approach to the design of enzyme inhibitors became possible when the reactive group was attached to a substrate; in this way, the specificity of the target enzyme was utilized to achieve selective inhibition of the enzyme [17]. Such active-site-directed inhibitors have acquired major importance not only academically in the study of enzyme mechanisms but also commercially in the search for a rational approach to selective toxicity or chemotherapy.

Three-Dimensional Structure. Chemical studies showed that the active site of an enzyme consists of a constellation of amino acid side chains brought together spatially from different parts of the polypeptide chain. If this three-dimensional structure was disrupted by denaturation, that is, without breaking any covalent bonds, the biological activity of the enzyme was destroyed. In addition, it was found that all the information required for a protein to fold up spontaneously in solution and reproduce its native shape was contained in its primary structure. This was part of the original “central dogma” of molecular biology.

The X-ray crystallography of proteins [18] demonstrated unequivocally that a given protein has a unique three-dimensional structure. Among the basic design principles was the tendency of hydrophobic amino acid side chains to be associated with the hydrophobic interior of the folded molecule, whereas charged side chains were almost exclusively situated on the hydrophilic exterior or surface. The first high-resolution crystallographic analysis of an enzyme, egg-white lysozyme, confirmed these principles and led to the proposal of a detailed mechanism [19]. The active site was located in a cleft in the structure (Fig. 1), which has subsequently proved to be a common feature of active sites. According to this, the enzymatic reaction takes place in a hydrophobic environment, and the successive chemical events involving substrate and protein side chains are not constrained by the ambient conditions of aqueous solution and neutral pH.

Figure 1. A molecular model of the enzyme lysozyme: the arrow points to the cleft that accepts the polysaccharide substrate (Reproduced by courtesy of J. A. RUPLEY)
[Full View]

1.2. Enzyme Nomenclature

Strict specificity is a distinguishing feature of enzymes, as opposed to other known catalysts. Enzymes occur in myriad forms and catalyze an enormous range of reactions. By the late 1950 s the number of known enzymes had increased so rapidly that their nomenclature was becoming confused or, worse still, misleading because the same enzyme was often known to different workers by different names; in addition, the name frequently conveyed little or nothing about the nature of the reaction catalyzed.

To bring order to this chaotic situation, an International Commission on Enzymes was established in 1956 under the auspices of the International Union of Biochemistry (IUB). Its terms of reference were as follows: “To consider the classification and nomenclature of enzymes and coenzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics.” The Commission's recommendations have formed the basis of enzyme nomenclature since its first report in 1961 [1].

Responsibility for enzyme nomenclature passed to the Nomenclature Committee of IUB in 1977, which has subsequently published several reports, e.g., [20] and supplements, e.g., [21]; it is expected that further supplements will be published from time to time in the *European Journal of Biochemistry*. The growth in scale can be appreciated from the fact that the 1961 *Report of the Enzyme Commission* listed 712 enzymes, whereas the 1992 version of Enzyme Nomenclature listed 3196. The most recent information about changes or additions to enzyme nomenclature is available at `http://www.chem.qmw.ac.uk/iubmb/`, which also offers also an up-to-date version of the *Enzyme Nomenclature* list.

1.2.1. General Principles of Nomenclature

The accepted system for classification and nomenclature of enzymes embodies three general principles.

The first is that enzyme names, especially those ending in *-ase*, should be used only for single enzymes, i.e., single catalytic entities. They should not be applied to systems containing more than one enzyme.

The second general principle is that an enzyme is named and classified according to the reaction it catalyzes. This refers only to the observed chemical change produced by the enzyme, as expressed in the chemical equation. The mechanism of action is ignored, and intermediate cofactors or prosthetic groups are not normally included in the name. Thus, an enzyme cannot be named systematically until the reaction it catalyzes has been identified properly.

The third general principle is that enzymes are named and classified according to the type of reaction catalyzed, which enables Enzyme Commission (E.C.) code numbers to be assigned to enzymes to facilitate subsequent unambiguous identification. For the purpose of systematic nomenclature, all enzymes in a particular class are considered to catalyze reactions that take place in a given direction, although only the

reverse direction may have been demonstrated experimentally. However, the recommended name for the enzyme may well be based on the presumed direction of the reaction in vivo.

Thus, a given enzyme often has two names, one systematic and the other recommended or trivial. The latter is generally the name in current usage, shorter and more readily applied. After its systematic name and E.C. code number have identified an enzyme, the recommended name can be used without fear of ambiguity. This practice is now generally followed in the literature.

1.2.2. Classification and Numbering of Enzymes

According to the report of the first Enzyme Commission in 1961, enzymes are divided into six main classes according to the type of reaction catalyzed. They are assigned code numbers, prefixed by E.C., which contain four elements separated by points and have the following meaning:

1. the number first indicates to which of the six classes the enzyme belongs,
2. the second indicates the subclass,
3. the third number indicates the sub-subclass, and
4. the fourth is the serial number of the enzyme in its sub-subclass.

The six classes are distinguished in the following manner:

1. *Oxidoreductases*
   This class encompasses all enzymes that catalyze redox reactions. The recommended name is *dehydrogenase* whenever possible, but *reductase* can also be used. *Oxidase* is used only when O_2 is the acceptor for reduction. The systematic name is formed according to *donor : acceptor oxidoreductase.*
2. *Transferases*
   Transferases catalyze the transfer of a specific group, such as methyl, acyl, amino, glycosyl, or phosphate, from one substance to another. The recommended name is normally *acceptor grouptransferase* or *donor grouptransferase.* The systematic name is formed according to *donor : acceptor grouptransferase.*
3. *Hydrolases*
   Hydrolases catalyze the hydrolytic cleavage of C —O, C —N, C —C, and some other bonds. The recommended name often consists simply of the substrate name with the suffix *-ase.* The systematic name always includes *hydrolase.*
4. *Lyases*
   Lyases catalyze the cleavage of C —C, C —O, C —N, and other bonds by elimination. The recommended name is, for example, *decarboxylase, aldolase, dehydratase* (elimination of CO_2, aldehyde, and water, respectively). The systematic name is formed according to *substrate group-lyase.*
5. *Isomerases*
   Isomerases catalyze geometric or structural rearrangements within a molecule. The different types of isomerism lead to the names *racemase, epimerase, isomerase, tautomerase, mutase,* or *cycloisomerase.*
6. *Ligases*
   Ligases catalyze the joining of two molecules, coupled with the hydrolysis of a pyrophosphate bond in ATP or another nucleoside triphosphate. Until 1983, the recommended name often included *synthetase,* but the current recommendation is that names of the type *X – Y ligase* be used instead, to avoid confusion with the name *synthase* (which is not confined to enzymes of class 6). The systematic name is formed according to *X : Y ligase* (ADP-forming).

A few examples will serve to illustrate how this system works. (The full list can be found in *Enzyme Nomenclature* 1992 [20].)

The enzyme alcohol dehydrogenase (recommended name) catalyzes the reaction

$$\text{Alcohol} + \text{NAD}^+ \rightleftharpoons \text{Aldehyde or Ketone} + \text{NADH} + \text{H}^+$$

The enzyme has been assigned E.C. number 1.1.1.1. It may also be called aldehyde reductase, but its systematic name is alcohol: NAD^+ oxidoreductase.

Similarly, the enzyme hexokinase (recommended name), which catalyzes the reaction

$$\text{ATP} + \text{D-Hexose} \rightleftharpoons \text{ADP} + \text{D-Hexose 6-phosphate}$$

has been given the E.C. number 2.7.1.1. It has such other names as glucokinase and hexokinase type IV, and its systematic name is ATP: D-hexose 6-phosphotransferase.

1.3. Structure of Enzymes

Enzymes are proteins (for an exception, see Section The Ribozyme) and, as such, are amenable to structural analysis by the methods of protein chemistry, molecular biology, and molecular biophysics.

1.3.1. Primary Structure

The primary structure of enzymes can be determined by direct chemical methods which, in sensitivity and automation, have reached very high levels of sophistication [9], [22]. However, for many proteins, particularly those with long polypeptide chains, direct sequence analysis would be very time-consuming; others may be available only in very small amounts. In these cases, a more profitable approach is to clone the relevant structural gene and determine its DNA sequence [9], [23], [24]. From this, the amino acid sequence can be inferred. Whenever possible, this sequence should be checked, e.g., for genetic reading frame, against whatever amino acid sequence information is available from direct methods. The recombinant DNA approach is so quick and so powerful, however, that amino acid sequence information about enzymes is growing much more rapidly from this source than from direct chemical analysis [25], [26]. Indeed, the information now available is so large in total that computer data banks are required to store it and make it available for systematic access [27].

1.3.2. Three-Dimensional Structure

The three-dimensional structure of an enzyme can be obtained at high resolution by X-ray crystallography [28] and, for molecules up to ca. 300 amino acids in length, by NMR spectroscopy. By this means, the detailed structures of many enzymes have been determined, and a broad understanding of the principles of protein structure has resulted [29], [30]. Proteins are generally well-ordered; their interiors are well-packed (comparable to other crystalline organic molecules) to produce a hydrophobic core with a dielectric constant similar to that of a hydrocarbon. Proteins vary in the amount of regular secondary structure (α-helix and β-sheet) they contain and can be grouped into four classes according to the combination and packing of these structural features [31]. Although the number of possible combinations of amino acids in a given protein is virtually unlimited, it is estimated that there are not more 1000 different families of folding patterns for protein structures [32].

Despite their close-packed and generally well ordered structure, enzymes are usually not entirely rigid molecules, and some conformational flexibility in solution is widely observed, particularly by NMR spectroscopy [33-37]. These conformational changes may be limited to a molecular “breathing” or flexing of the structure, they may involve various “hinge-bending” motions, or they may extend to more substantial conformational mobility in parts of the polypeptide chain. All such motions, contribute to the mechanisms of enzyme catalysis [2], [38].

As of August 28, 2001, 15 905 3D structures were freely accessible from the website of the Protein Data Bank (`http://www.rcsb.org/pdb/`) [39]. This site also offers several programs for analysis and visualization of protein (and DNA) 3D structures on numerous computer platforms.

1.3.3. Quaternary Structure, Folding, and Domains

Many enzymes consist of more than one polypeptide chain (or subunit), and these must form an aggregate, usually with relatively simple symmetry, before full (or even any) biological activity is conferred (Table 1). The subunits within an oligomer or multimer are often identical or at least limited to a few different types.

Aggregation is generally some form of self-assembly dictated by coherent binding patterns between the subunits, which provide the necessary recognition sites in sorting out the subunits required for assembly [29], [40].

Table 1. Quaternary structures of some typical enzymes

Enzyme	E.C. number [*CAS registry number*]	Source	Number of subunits	Point symmetry: Crystallographic symbol	Point symmetry: Schönflies symbol
Alcohol dehydrogenase	1.1.1.1 [9031-72-5]	horse liver	2	2	C_2
Glutathione reductase	1.6.4.2 [9001-48-3]	human red blood cells	2	2	C_2
Triose phosphate isomerase	5.3.1.1 [9023-78-3]	chicken muscle	2	2	C_2
Lactate dehydrogenase	1.1.1.27 [9001-60-9]	dogfish muscle	4	222	D_2
Glyceraldehyde 3-phosphate dehydrogenase	1.2.1.12 [9001-50-7]	*Bacillus stearothermophilus*	4	222	D_2
Pyruvate kinase	2.7.1.40 [9001-59-6]	cat muscle	4	222	D_2
Aspartate carbamoyltransferase	2.1.3.2 [9012-49-1]	*Escherichia coli*	6 + 6	32	D_3
Dihydrolipoamide acetyltransferase	2.3.1.12 [9032-29-5]	*Escherichia coli*	24	432	0
Dihydrolipoamide acetyltransferase	2.3.1.12 [9032-29-5]	*Bacillus stearothermophilus*	60	532	Y

The complexity of this sorting process in a cell becomes evident from the fact that many intracellular enzymes are dimers or tetramers. Increasingly more complicated structures are being recognized and their design principles analyzed. These range from enzymes with simple cyclic symmetry up to those with the most elaborate cubic point group symmetry, e.g., octahedral and icosahedral [29], [40].

The folding of polypeptide chains, along with their aggregation into ordered structures, is a spontaneous process in solution, and this implies that it is exergonic [39]. However, calculation of the time required for a protein to explore all possible structures during the folding process indicates that the search for the "right" structure cannot be entirely random. Thus, even for a small protein such as bovine pancreatic ribonuclease (124 amino acid residues), such a search might take around 10^{95} years, whereas the experimentally determined time in vivo is a few milliseconds. This dramatic discrepancy led to the concept of *kinetic pathways* during folding. Such pathways have been experimentally explored, and intermediates identified for various proteins. The stable structure of a protein in solution is therefore identified as the lowest free energy form of the kinetically accessible structures [29], [30], [40].

A typical enzyme is not an entity completely folded as a whole, as is evident from the growing catalogue of three-dimensional protein structures determined by X-ray crystallography. On the contrary, enzymes frequently consist of apparently autonomous or semiautonomous folding units, called *domains* (Fig. 2). Sometimes, these may be identified as products of limited proteolysis, i.e., regions of the polypeptide chain that can be excised from the chain with retention of their biological properties. Indeed, this has proved in many instances to be a valuable guide to the actual activity contributed by that part of the enzyme. Classical examples of such functional domains can be found in the study of muscle contraction and antibody-antigen recognition [29], [30].

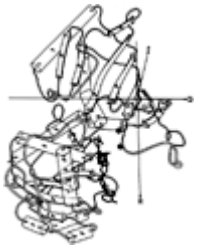

Figure 2. The domains in glyceraldehyde 3-phosphate dehydrogenase from *Bacillus stearothermophilus* [467]

Reproduced with permission
[Full View]

In other cases, domains are not readily released as biologically active entities, and their existence must be inferred from the three-dimensional structure of the enzyme. Most globular proteins can in fact be subdivided into such regions, which generally have molecular masses of 20 000 or less [29]. The active site of an enzyme is often located at the interface between two such domains as, for example, in the well-known cleft of lysozyme (Fig. 1) or in glutathione reductase. Other domains appear to represent favored folding patterns in the assembly of proteins, but biological activity associated with them can often be inferred from comparison of the structures of related proteins: a typical example is the NAD-binding domain present in dehydrogenases.

Structural domains may be regions of the polypeptide chain that fold independently of each other. Functional domains, as defined above, do indeed fold independently; and individual subunits of oligomeric enzymes appear to fold before association [29], [30], [40], [41].

1.3.4. The Ribozyme

Enzymes are proteins, but the specific involvement of RNA molecules in certain reactions concerned with RNA processing in vivo is worth noting. Until CECH *et al.* [42] and ALTMAN *et al.* [43] published their observations, it was generally accepted knowledge that the major duties in a biological system, namely, to encode information and to catalyze chemical reactions, are neatly split, one being performed by nucleic acids, the other by proteins. With the discovery of special RNAs which store genetic information and can also catalyze reactions on themselves or on other RNAs, this dogma was destroyed [42], [43]. Over the years, it has become evident, that group I and group II introns, catalyze various transesterifications. In cellular systems these reactions facilitate their excision from pre-RNAs and the ligation of flanking exons (self-splicing). In vitro these intron RNAs perform a variety of reactions in cis (i.e., on the same strand of

the RNA genome) and in trans (i.e., on another RNA), such as cleavage and ligation of RNAs, transfer of nucleotides between RNAs, polymerization, and editing-like reactions. These RNAs thus can act as enzymes and are therefore called “ribozymes” [44].

In *Escherichia coli*, tRNA precursors are cleaved by ribonuclease P to generate the correct 5'-ends of the mature tRNA molecules, and the enzyme contains an essential RNA moiety that can function in the absence of the protein. In fact, this RNA moiety fulfills all the criteria of an enzyme [45]. Similarly, the ribosomal RNA of *Tetrahymena thermophila* undergoes self-splicing to perform a highly specific intramolecular catalysis in the removal of an intervening sequence. A truncated version of the intervening sequence, lacking the first 19 nucleotides of the original excised RNA, can then behave as an enzyme in vitro, capable of acting as an RNA polymerase and a sequence-specific ribonuclease under appropriate conditions [46].

The structure of the ribosome's large subunit has since been solved. This largest unique structure established that the ribosome is a ribozyme in which the ribosomal RNA, and not the protein, performs catalytic functions, including the peptidyl transferase reaction that forms the peptide bond [47], [48]. One of the most remarkable findings to emerge from this is that although enzymes composed entirely of protein promote virtually all chemical reactions that occur in living organisms, the protein synthesis reaction that occurs on the ribosome is due to the two-thirds of its mass that is RNA, not the one-third that is protein. In addition to enhancing the understanding of protein synthesis, this work will have significant medical implications, because the ribosome is a major target for antibiotics [49].

Ribozymes also offer an excellent opportunity to compare and contrast the behavior of RNA enzymes with that of protein enzymes. The differences between the RNA and protein enzymes highlight features that are distinct and thus enable a better understanding of each of these classes of biological macromolecules. On the other hand, the features of protein and RNA enzymes that are similar may represent aspects that are fundamental to biological catalysis. Indeed, these studies have suggested that RNA enzymes, like their protein counterparts, can use binding interactions remote from the site of bond transformation to facilitate that transformation [50]. Beyond this, recent results suggest that RNA enzymes are ideally suited for exploration of the energetic origins of this interconnection between binding and catalysis [51]. This use of binding energy provides a natural connection between rate enhancement and specificity, the two hallmarks of biological catalysis. Finally, ribozymes will not only offer new clues about evolution [52], but also offer the potential for specific inactivation of disease-associated mRNAs or viral RNA genomes that, unlike conventional therapeutics, require no knowledge of the structure or function of proteins that target RNAs encode [53].

1.4. Enzymes and Molecular Biology

1.4.1. Biosynthesis of Enzymes

Enzymes are synthesized in cells by the normal machinery of protein synthesis. The structure of any given enzyme is encoded by a structural gene, whose DNA base sequence is transcribed into a messenger RNA, and the mRNA is translated from its triplet code into the amino acid sequence of the desired protein by the ribosomes and associated factors [54], [55]. The enzyme then folds spontaneously into its active conformation. Posttranslational modifications may be required to target an enzyme to its ultimate intracellular or extracellular location.

1.4.2. Enzymes and DNA

For many years, the chemical manipulation of DNA lagged behind that of proteins. The chemical complexity and variety of proteins, with up to 20 different naturally occurring amino acids, served to make them more amenable to increasingly sophisticated methods of analysis. On the other hand, DNA, composed of only four different nucleotides, appeared dauntingly large, with few structural features to make it yield to available methodology.

Paradoxically, this very lack of variety in the nature of the constituent nucleotides of DNA has permitted the revolution in genetic engineering, in which the enzymology of DNA [56] has played a prominent part. For example, the discovery and purification of restriction enzymes enabled DNA to be cleaved selectively into defined fragments; phosphatases and ligases permit the fragments to be rejoined selectively; and DNA polymerases allow DNA to be synthesized and sequenced at astonishing speed, all in vitro [23], [54-56].

PREVIOUS NEXT

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Gerhard Michal, Tutzing, Germany
Albert Jonke, Boehringer Mannheim GmbH, Penzberg, Germany

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

2. Catalytic Activity of Enzymes [57, 58, 60]

The theory of enzyme-catalyzed reactions proposed by MICHAELIS and MENTEN in 1913 [61] is based on the assumption that the enzyme (the catalyst, E) and the substrate (the reactant, S) form a complex (ES) by a reversible reaction. The complex is then converted into the product (P) with the reaction rate k_2, when practically no product is present.

$$S + E \underset{k_{-1}}{\overset{k_1}{\leftrightarrows}} ES \xrightarrow{k_2} P + E \tag{2.1}$$

Under commonly used conditions of enzyme activity measurement, c_{ES} can be considered to be sufficiently constant during the observed reaction period (steady-state assumption, BRIGGS and HALDANE [104]):

$$\frac{\mathrm{d}(c_{\mathrm{ES}})}{\mathrm{d}t} = 0 = (k_1 \cdot c_{\mathrm{E}} \cdot c_{\mathrm{S}}) - (k_{-1} \cdot c_{\mathrm{ES}}) - (k_2 \cdot c_{\mathrm{ES}}) \tag{2.2}$$

By using a term for the total concentration of enzyme

$$c_{\mathrm{Et}} = c_{\mathrm{E}} + c_{\mathrm{ES}} \tag{2.3}$$

one obtains

$$0 = k_1 \cdot (c_{\mathrm{Et}} - c_{\mathrm{ES}}) \cdot c_{\mathrm{S}} - (k_{-1} \cdot c_{\mathrm{ES}}) - (k_2 \cdot c_{\mathrm{ES}}) \tag{2.4}$$

or

$$c_{\mathrm{ES}} = \frac{k_1 \cdot c_{\mathrm{Et}} \cdot c_{\mathrm{S}}}{k_{-1} + k_2 + k_1 \cdot c_{\mathrm{S}}} \tag{2.5}$$

By introducing the Michaelis constant K_M

$$K_{\mathrm{M}} = \frac{k_{-1} + k_2}{k_1} \tag{2.6}$$

the reaction rate as a function of c_{ES}

$$v = k_2 \cdot c_{ES} \tag{2.7}$$

and the maximum reaction rate V, which is reached, when all of the enzyme is saturated with substrate ($c_{ES} = c_{Et}$)

$$V = k_2 \cdot c_{Et} \tag{2.8}$$

one obtains the so-called Michaelis–Menten equation

$$v = \frac{V \cdot c_S}{K_M + c_S} \tag{2.9}$$

which shows the dependency of the reaction rate on the substrate concentration (first-order reaction). The plot of this relationship is given in Figure 3.

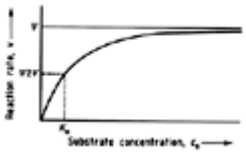

Figure 3. Reaction rate as a function of substrate concentration (enzyme at constant concentration)

For explanation of symbols, see text
[Full View]

In the case of frequently occurring two-substrate reaction, a similar derivation leads to the formula:

$$v = \frac{V}{1 + \frac{(K_M)_{S1}}{c_{S1}} + \frac{(K_M)_{S2}}{c_{S2}} + \frac{(K_M)_{S1S2}}{c_{S1} \cdot c_{S2}}} \tag{2.10}$$

This is called a second-order reaction. If, however, the concentration of the second substrate is kept at a level many times of the respective Michaelis constant, then the third and fourth term of the denominator are practically zero and the equation is identical with Equation (2.9), allowing the same evaluation as with a one-substrate reaction.

Michaelis Constant. As can be derived from Equation (2.9), the Michaelis constant equals the substrate concentration at half the maximal reaction rate. The value K_M can be obtained by plotting the experimentally measured reaction rate against the various substrate concentrations (Fig. 3). A more convenient way is the plot according to LINEWEAVER and BURK [105], using a reciprocal of the Michaelis – Menten equation

$$\frac{1}{v} = \frac{k_M}{V} \cdot \frac{1}{c_S} + \frac{1}{V} \tag{2.11}$$

A schematic plot is shown in Figure 3. The intersections with abscissa and ordinate allow the determination of the values for K_M and V. For further discussion, see [62]. The Michaelis constant approaches the dissociation constant K_s of the enzyme – substrate complex and is therefore valuable for estimating individual reaction kinetics.

Michaelis constants for enzymes usually range from 10^{-2} to 10^{-5} mol/L; a low K_M indicates a high affinity between enzyme and substrate.

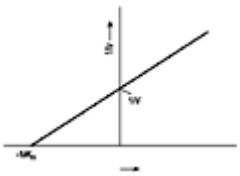

Figure 4. Lineweaver – Burk graph for determination of the kinetic parameters of an enzyme.

For explanation of symbols, see text
[Full View]

Molar or Molecular Activity (Turnover Number). The efficiency of an enzyme-catalyzed reaction is indicated by the molar activity, formerly called turnover number, which is defined as the number of substrate molecules converted in 1 min by one enzyme molecule under standardized conditions. This can be calculated from the specific activity of a particular enzyme if its molecular mass is known (cf. Section Absorption Photometry). The average molar activity ranges from 10^3 to 10^4; peak values have been measured for acetylcholinesterase (E.C. 3.1.1.7) [9000-81-1] at 1×10^6 and for catalase (E.C. 1.11.1.6) [9001-05-2] at 5×10^6.

2.1. Factors Governing Catalytic Activity [63]

2.1.1. Temperature

The temperature dependence of enzyme-catalyzed reactions exhibits an optimum because the thermodynamic increase of reaction rate (1 in Fig. 5) is followed by a steep drop caused by thermal denaturation of the enzyme (2 in Fig. 5). The optimum is generally between 40 and 60 °C. Some temperature-insensitive enzymes may exhibit an optimum at almost 100 °C. Data on various frequently used enzymes are given in [64]. Figure 5 illustrates temperature dependence [60].

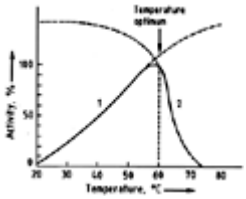

Figure 5. Temperature optimum of enzyme activity
[Full View]

2.1.2. Value of pH

All enzymes have an optimum pH range for activity. The optimum depends not only on pH but also on ionic strength and type of buffer. It may also be influenced by temperature, substrate, and coenzyme concentrations. For most enzymes, the pH optimum lies in the range from 5 to 7. Extreme values of 1.5 and 10.5 have been found for pepsin (E.C. 3.4.23.1) [9004-07-3] and for alkaline phosphatase (E.C. 3.1.3.1) [9001-78-9], respectively. Figure 6 shows some examples [60].

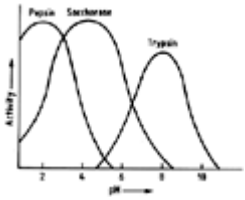

Figure 6. Activity of various enzymes as a function of pH

[Full View]

2.1.3. Activation

Many chemical effectors activate or inhibit the catalytic activity of enzymes. In addition to substrates and coenzymes many enzymes require nonprotein or, in some cases, protein compounds to be fully active. Enzyme activation by many inorganic ions has been adequately described. The activating ion may be involved directly in the reaction by complexing the coenzyme or cosubstrate (e.g., Fe ions bound to flavin or the ATP – Mg complex). In other cases, the ion is part of the enzyme and either acts as a stabilizer for the active conformation (e.g., Zn ions in alkaline phosphatase) or participates directly at the active site (e.g., Mn

ions in isocitrate dehydrogenase (E.C. 1.1.1.42) [9028-48-2] and Zn or Co ions in carboxypeptidases).

2.1.4. Inhibition [65]

In vivo and in vitro inhibition studies of enzymatic reactions contributed important knowledge to various fields of biochemistry. For example, the mechanism of action of many toxic substances and antidotes has been found to affect enzymes directly. In many cases, the importance of these enzymes for metabolism has been revealed. On the other hand, discovery of end-product inhibition elucidated many metabolic pathways. Enzyme inhibition can be either reversible or irreversible. Depending on the type of inhibitory effect, the following mechanisms of enzyme inhibition may be distinguished.

Irreversible Inhibition. An irreversible inhibitor frequently forms a stable compound with the enzyme by covalent bonding with an amino acid residue at the active site. For example, diisopropyl fluorophosphate (DIFP) reacts with a serine residue at the active site of acetylcholinesterase to form an inactive diisopropylphosphoryl enzyme. Alkylating reagents, such as iodoacetamide, inactivate enzymes with mercapto groups at their active sites by modifying cysteine.

Reversible Inhibition. Reversible inhibition, in contrast, is characterized by an equilibrium between enzyme and inhibitor. Several main groups of reversible inhibitory mechanisms can be differentiated.

Competitive Inhibition. The inhibitor competes with the substrate or coenzyme for the binding site on the active center by forming an enzyme-inhibitor complex EI. In most cases, the chemical structure of the inhibitor resembles that of the substrate. Inhibition can be made ineffective by excess substrate, as is the case for inhibition of succinate dehydrogenase (E.C. 1.3.99.1) [9002-02-2] by malonate.

Noncompetitive Inhibition. The inhibitor decreases the catalytic activity of an enzyme without influencing the binding relationship between substrate and enzyme. This means that inhibitor and substrate can bind simultaneously to an enzyme molecule to form ES, EI, or ESI complexes. Noncompetitive inhibition is dependent solely on the inhibitor concentration and is not overcome by high substrate concentration. An example is the blocking of an essential cysteine residue by such heavy metals as copper or mercury.

Uncompetitive Inhibition. The inhibitor reacts only with the intermediary enzyme – substrate complex. An example is the reaction of azide with the oxidized form of cytochrome oxidase (E.C. 1.9.3.1) [9001-16-5]. Lineweaver – Burk plots of the reciprocal initial reaction rate 1/ v_0 versus the reciprocal substrate concentration for these three modes of reversible inhibition are shown in Figures 7 A – C [57]. The inhibitor constant K_I characterizes the inhibiting activity. The respective formulae (reciprocal from according to Lineweaver – Burk [105] are:

Competitive inhibition:

$$\frac{1}{v} = \frac{K_M \cdot \left(1 + \frac{c_I}{K_I}\right)}{V \cdot c_S} + \frac{1}{V} \tag{2.12}$$

Noncompetitive inhibition:

$$\frac{1}{v} = \frac{K_M}{V \cdot c_S} \cdot \left(1 + \frac{c_I}{K_I}\right) + \frac{1}{V} \cdot \left(1 + \frac{c_I}{K_I}\right) \tag{2.13}$$

Uncompetitive inhibition:

$$\frac{1}{v} = \frac{K_m}{V \cdot c_S} + \frac{1 + \frac{c_I}{K_I}}{V} \tag{2.14}$$

For details, see [62].

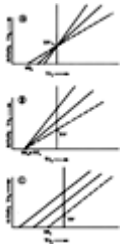

Figure 7. Lineweaver – Burk graphs of reversibly inhibited enzyme reactions

A) Competitive inhibition; B) Noncompetitive inhibition; C) Uncompetitive inhibition

– – –: Uninhibited reaction
[Full View]

Substrate Inhibition. High concentration of substrate (or coenzyme) may decrease the catalytic activity of an enzyme. Examples are the action of ATP on phosphofructokinase (E.C. 2.7.1.11) [9001-80-3] or of urea on urease (E.C. 3.5.1.5) [9002-13-5]. The Lineweaver – Burk plot [105] is given in Figure 7.

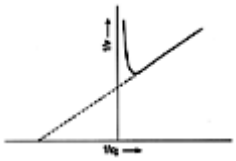

Figure 8. Lineweaver – Burk graph of a substrate inhibition of an enzyme
[Full View]

End-Product Inhibition. In many multienzyme systems, the end product of the reaction sequence may act as a specific inhibitor of an enzyme at or near the beginning of the sequence. The result is that the rate of the entire sequence of reactions is determined by the steady-state concentration of the end product (Fig. 9). This type of inhibition is also called *feedback inhibition* or *retroinhibition.*

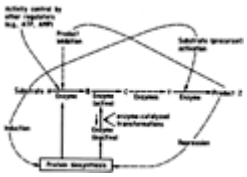

Figure 9. Natural mechanisms for regulating enzyme activity
[Full View]

2.1.5. Allostery [66]

Cosubstrates with a central role in metabolism, such as acetyl-CoA, ATP, or AMP, may also influence the rate of reaction sequences by allosteric regulation. For example, phosphofructokinase, the first enzyme in the energy-supplying Embden – Meyerhof – Parnass pathway, is inhibited by a high concentration of ATP (i.e., positive energy balance). A high concentration of AMP, on the other hand, (i.e., energy deficiency) terminates this inhibition. Allosterically regulated enzymes have a quaternary structure and are composed of two or more structurally similar or identical subunits (protomers), each with a binding site for the substrate and another independent binding site for the allosteric effector. The binding of the effector modifies the conformation of the subunit and its active center, which then affects the conformation and hence the catalytic activity of the entire molecule. The cooperation of substrate and effector regulates the overall catalytic activity of the enzyme depending on the concentration of metabolite.

Allosteric enzymes usually do not show the classical Michaelis – Menten kinetic relationship of c_s, V, and K_M. Many allosteric enzymes give a sigmoid plot of initial rate vs. substrate concentration rather than the hyperbolic plots predicted by the Michaelis – Menten equation. The shape of the curve is characteristically changed by an allosteric activator (positive cooperativity) or an allosteric inhibitor (negative cooperativity), as shown in Figure 10.

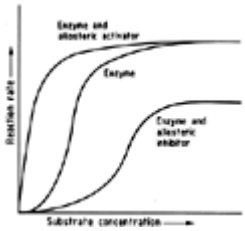

Figure 10. Activity of an allosteric enzyme as a function of substrate concentration in absence and presence of an allosteric activator or inhibitor
[Full View]

This sigmoidal curve implies that within a certain range of substrate concentration, the enzyme is able to respond to small concentration changes by great activity changes.

2.1.6. Biogenic Regulation of Activity

In principle, enzyme activity may also be controlled by regulating the amount of enzyme in the cell. This can be accomplished by regulating the biosynthesis of individual enzymes or of several functionally related enzymes by induction or repression, or by specific attack of proteolytic enzymes (Fig. 9).

Chymotrypsin (E.C. 3.4.21.1) [9004-07-3], for example, is synthesized as inactive zymogen and converted to the active enzyme by trypsin (E.C. 3.4.21.4) [9002-07-7], a typical protease. A large number of natural protease inhibitors have been isolated and characterized. They act as protease antagonists and are capable of selectively affecting many proteolytic reactions. These proteinase inhibitors are proteins with molecular masses of 5000 – 25 000; they inhibit proteases by specific complex formation [67].

Another important regulation mechanism is the modification of activity by phosphorylation and dephosphorylation, which allows quick adaptation to changing environmental conditions. For example, this takes place with hydroxymethylglutaryl-CoA reductase (NADPH) (E.C. 1.1.1.34), which is phosphorylated by [hydroxymethylglutaryl-CoA reductase (NADPH)] kinase (E.C. 2.7.1.109) [106] and dephosphorylated by [hydroxymethylglutaryl-CoA reductase (NADPH)] phosphatase (E.C. 3.1.3.47) [107]. In addition, this enzyme is regulated at the protein synthesis level and by controlled degradation.

2.2. Enzyme Assays

2.2.1. Reaction Rate as a Measure of Catalytic Activity

As biological catalysts, enzymes increase the rate of a reaction or permit it to proceed. Therefore, the conversion rate of substrates, v, is measured to determine the catalytic activity. The respective formulae for enzyme kinetics are given in Equations (2.9) and (2.10).

If in Equation (2.9) the substrate concentration is kept at a level considerably above the Michaelis constant ($c_S \gg K_M$), then all of the enzyme is saturated with substrate and the reaction proceeds with the constant and the maximum rate $v = V$ (zero order reaction). Consequently, catalytic activity is linearly dependent on the amount of enzyme used. This situation is attempted to be reached in enzyme activity measurements.

2.2.2. Definition of Units

Originally, units were defined by the investigator who first discovered and described an enzyme. Therefore, in the older literature, enzyme activity was expressed in arbitrary units, e.g., changes in absorbance, increase of reducing groups, amount of converted substrate expressed in milligrams or micromoles. These parameters were related to various time units, such as 1 min, 30 min, or 1 h.

To obtain standardized values for each enzyme, in 1961 the Enzyme Commission of the International Union of Biochemistry defined the *International Unit* U as the activity of an enzyme which, under optimized standard conditions, catalyzes the conversion of 1 µmol of substrate per minute. With respect to basic SI units, the Expert Panel on Quantities and Units (EPQU) of the International Federation of Clinical Chemists (IFCC) and the Commission on Quantities and Units in Clinical Chemistry (CQUCC) of IUPAC defined the base unit *katal* as the catalytic amount of any enzyme that catalyzes a reaction (conversion) rate of 1 mol of substrate per second in an assay system [68]. This unit, however, is not in common use.

The temperature must be stated for each assay. As a general rule, the rate of enzymatically catalyzed

reactions is about doubled by a temperature increase of 10 °C in the range of 0 – 40 °C. The temperature must be controlled precisely and kept constant to achieve reproducible results [63].

However, for practical reasons, many enzyme reactions cannot be monitored by measuring the stoichiometric amount of substrate consumed or product formed. Therefore, the catalytic activity of a particular enzyme may be impossible to express in International Units.

The definition of units for most enzymes used in molecular biology is rather arbitrary. For example, the unit of a restriction endonuclease (E.C. 3.1.21.3 – 5), is defined as the catalytic activity of the enzyme that yields a typical cleavage pattern, detectable after electrophoresis, with a precise amount (usually 1 µg) of a particular DNA under defined incubation conditions [69], [70]. Other parameters that define the activity of such enzymes are the degradation of a nucleic acid or the incorporation of nucleotides into a nucleic acid, expressed in micrograms, nanomoles, absorbance units, or number of base pairs. For practical reasons, different time periods such as 1, 10, 30, or 60 min, may be chosen as references.

2.2.3. Absorption Photometry [71]

Basic Considerations. Because of its simple technique and reliable, reasonably priced instruments, photometry is today one of the preferred methods of enzyme assay. It can be carried out most quickly and conveniently when the substrate or the product is colored or absorbs light in the ultraviolet region because the rate of appearance or disappearance of a light-absorbing product or substrate can be followed with a spectrophotometer.

According to the Bouguer – Lambert – Beer law, which is valid for very dilute solutions, the following relationship exists between absorbance *A* and concentration:

$$A = \log \frac{I_0}{I} = \varepsilon \cdot c \cdot d \tag{2.15}$$

and

$$c = \frac{A}{\varepsilon \cdot d} \tag{2.16}$$

where *c* is the concentration in millimoles per liter, ε the millimolar absorption coefficient in liters per mole and per millimeter, and *d* the path length in millimeters.

The catalytic activity *z* then corresponds to the absorbance change per minute.

$$z = \frac{\Delta A \cdot V \cdot 1000}{\varepsilon \cdot d \cdot \Delta t} \tag{2.17}$$

where *V* is the assay volume in liters and *t* the time in minutes. The unit of *z* is then micromoles per minute and corresponds to the definition of the International Unit U given in the previous section.

Example 1: Assay of Lactate Dehydrogenase (*LDH*).

In the reaction catalyzed by LDH (E.C. 1.1.1.27) [9001-60-9], hydrogen is transferred from NADH to pyruvate, to yield L-lactate and NAD [72]:

$$\text{Pyruvate} + \text{NADH} + \text{H}^+ \rightleftharpoons \text{L-Lactate} + \text{NAD}^+ \tag{2.18}$$

The reduced coenzyme NADH absorbs at 340 nm, whereas the oxidized form NAD, lactate, and pyruvate do not. Thus the progress of Reaction (2.18) can be followed by measuring the decrease in light absorption at

340 nm with a mercury line photometer emitting at 334 or 365 nm. The enzyme can also be measured by monitoring the reverse reaction under slightly alkaline conditions. However, the reverse reaction is much slower than the reaction starting with pyruvate [73].

The principle of the optical assay may also be used to follow an enzymatic reaction in which neither the substrate nor the product has any characteristic light absorption maxima. In that case, the reaction is coupled to some other enzymatic reaction which can be followed easily by photometry. The activity of the nonabsorbing enzyme system can then be measured if the enzyme considered is made the rate-determining component by appropriate choice of assay conditions.

Example 2: Assay of Pyruvate Kinase (*PK*). The reaction between phosphoenolpyruvate and ADP yields pyruvate and ATP by transfer of a phosphate group; it is catalyzed by pyruvate kinase (E.C. 2.7.1.40) [9001-59-6]:

$$\text{Phosphoenolpyruvate} + \text{ADP} \rightarrow \text{Pyruvate} + \text{ATP} \tag{2.19}$$

This reaction is easily measured when a large excess of LDH and NADH is added to the system, which couples Reactions (2.19) and (2.18) [74]. The formation of pyruvate in Reaction (2.19) is followed by the very rapid reduction of pyruvate to lactate in Reaction (2.18). For each molecule of pyruvate formed and reduced, one molecule of NADH is oxidized to NAD, causing a decrease in light absorption at 340 nm.

The catalytic activity of enzymes such as phosphatases [75], whose natural substrates do not have suitable spectral properties, can be determined by using a colorless synthetic substrate, which is split enzymatically to yield a colored product.

Example 3: Assay of Phosphatase. For the assay of phosphatases, glycosidases, and several other hydrolases, colorless 4-nitrophenyl compounds are incubated, and 4-nitrophenolate is formed under alkaline conditions with a characteristic maximum between 400 and 420 nm. In this way, the catalytic activity of such enzymes can be measured conveniently at this wavelength:

$$\text{4-Nitrophenyl phosphate} + H_2O \rightarrow \text{4-Nitrophenol} + \text{Phosphate} \tag{2.20}$$

$$\text{4-Nitrophenol} + OH^- \rightarrow \text{4-Nitrophenolate} \tag{2.21}$$

Another principle involves the absorption of a colored metabolite formed directly or indirectly, by the action of the enzyme being analyzed. Again, the activity of the enzyme can be determined only if it is made the rate-determining component by proper choice of conditions.

Example 4: Assay of Glycerol Phosphate Oxidase. Among the class of oxidoreductases utilizing O_2 as the sole final electron acceptor, many enzymes exist that generate H_2O_2 during oxidation of their individual substrates [76]. Highly sensitive and accurate colorimetric assays of H_2O_2 have been developed during the last two decades [77]; in one system, phenol gives a purple color with 4-aminoantipyrine in the presence of oxidizing agents. Therefore, this color reaction can be used for colorimetric determination of H_2O_2-generating enzyme reactions with peroxidase as oxidation catalyst and indicator enzyme.

Although this principle is used predominantly to determine metabolites in body fluids, it is often employed in the assay of such enzymes as glycerol phosphate oxidase (E.C. 1.1.3.21) or glucose oxidase (E.C. 1.1.3.4) [9001-37-0].

The catalytic activity of the first is determined according to the following reaction sequence:

$$2\,\text{L-}\alpha\text{-Glycerol phosphate} + 2\,O_2 \rightarrow 2\,\text{Dihydroxyacetone phosphate} + 2\,H_2O_2 \quad (2.22)$$

$$2\,H_2O_2 + 4\text{-Aminoantipyrine} + \text{Phenol} \rightarrow \text{Quinonimine dye} + 4\,H_2O \quad (2.23)$$

The broad absorption maximum of the quinonimine dye is centered around 500 nm, with an absorption coefficient of about 13×10^2 L mol^{-1} mm^{-1}. The increase in absorbance per minute at 500 nm is measured to determine the activity of glycerol phosphate oxidase.

2.2.4. Fluorometry [78]

The fluorometric method is rarely used for determining the catalytic activity of raw or purified enzyme preparations. Because of its high sensitivity, it permits the assay of small amounts of enzymes in organs or tissue sections [79]. For example, systems that depend on NAD and NADP can be measured by fluorometry; the reduced pyridine coenzymes exhibit a fluorescence of low intensity. To enhance sensitivity, the oxidized form is treated with alkali, to yield strongly fluorescing compounds; in addition, selective filters must be used. The overall sensitivity of this method is a thousand times that of absorption photometry.

2.2.5. Luminometry [80]

Luminometry uses fluorescence, phosphorescence, and chemiluminescence as detector systems. Chemiluminescence observed in living organisms is termed bioluminescence. Bioluminescence is catalyzed by enzymes called luciferases, whose substrates, known as luciferins, are converted to light-emitting products.

In luminometry, the number of photons emitted by the reaction system per unit time is measured with specially designed instruments called luminometers; they are based on single photon counting detectors, usually photomultiplier tubes.

An example is the reaction catalyzed by the luciferase from *Photinus pyralis*, *Photinus*-luciferin 4-monooxygenase (ATP-hydrolyzing) (E.C. 1.13.12.7) [61970-00-1]:

$$\text{ATP} + \text{D-Luciferin} + O_2 \rightarrow \text{Oxyluciferin} + \text{AMP} + \text{pp} + CO_2 + 0.9\,h\nu \quad (2.24)$$

In this reaction, ATP is consumed as a substrate, and photons at a wavelength of 562 nm are emitted. The quantum yield is 0.9 einstein per mole of luciferin, i.e., for one ATP molecule consumed, approximately one photon is emitted. This reaction is therefore suitable for the assay of ATP and, hence, of enzymes that catalyze ATP-consuming or ATP-producing reactions.

To monitor enzyme activity with the aid of firefly bioluminescence, the light intensity must increase linearly for several minutes and must be strictly proportional to the catalytic activity of the enzyme. Such measuring conditions have been realized for determination of the catalytic activity of creatine kinase (E.C. 2.7.3.2) [9001-15-4]:

$$\text{Creatine phosphate} + \text{ADP} \rightarrow \text{Creatine} + \text{ATP} \quad (2.25)$$

Other reactions that depend on NAD(P) can be followed by using the bioluminescence from lucibacteria.

2.2.6. Radiometry

When radioactively labeled substrates are used, the activities of some enzymes can be determined with high sensitivity. This technique is widely employed in the field of molecular biology to monitor (1) the incorporation of radioactively labeled nucleotides into acid-insoluble nucleic acids or polynucleotides (DNA and RNA polymerases), (2) the decomposition of radioactively labeled DNA (exonuclease III), (3) the

transfer of a radioactively labeled phosphate group from γ-^{32}P-ATP [2964-07-0] to the 5′-hydroxyl end of a polynucleotide (polynucleotide kinase), or (4) the exchange of radioactively labeled pyrophosphate on a carrier matrix (T_4 DNA ligase). The most common isotopes for labeling are ^{32}P, ^{14}C, ^{3}H, and to a minor degree, ^{35}S.

In experiments with radioactively labeled compounds, special safety and legal regulations must be observed. In some determinations, however, the amounts of radioactive material needed are below the limits regulated by law.

2.2.7. Potentiometry [81]

A pH-sensitive glass electrode can be used to measure reactions in which protons are produced or consumed. For this purpose, the pH is kept constant by countertitration, and the consumption of acid or base required to do this is measured. The electrode controls an automatic titrator, and this concept is the *pH-stat* technique described by BÜCHER [82]. A typical example is the determination of the catalytic activity of lipase (E.C. 3.1.1.3) [9001-62-1] [83]. A fat (triglyceride) is hydrolyzed by this enzyme, and the fatty acid formed is neutralized by countertitration with NaOH in a pH-stat mode:

$$\text{Triglyceride} + H_2O \rightarrow \text{Diglyceride} + \text{Fatty acid} \qquad (2.26)$$

The substrate, olive oil, is incubated with the diluted sample containing lipase, and the mixture is titrated at constant pH. A recorder plots the consumption of NaOH vs. time, and the resulting slope correlates with catalytic activity. Other examples are the assay of papain (E.C. 3.4.22.2) [9001-73-4] and glucose oxidase (E.C. 1.1.3.4.) [9001-37-0].

2.2.8. Conductometry

In principle, all enzymatic reactions that lead to a change in overall ionic mobility can be measured by conductometry. In this way, the elastolytic activity of elastase (E.C. 3.4.21.36) by using unmodified elastin as substrate has been determined [84]. Other applications have also been described [85]. In this reaction, protons are liberated by cleavage of peptide bonds.

2.2.9. Calorimetry

Many enzymatic reactions evolve heat; therefore, some interest in calorimetric (enthalpimetric) methods has developed. In a suitable experimental arrangement, a temperature sensor serves as a device for measuring the catalytic activity of the enzyme. A new area of analytical chemistry has developed from this approach. Previously called microcalorimetry, it is now commonly known as enthalpimetry [86]. The method is used mainly in research.

2.2.10. Polarimetry

Polarimetry is rarely used, partly because of the inconvenience involved. However, it is required in determining the catalytic activity of mutarotase (E.C. 5.1.3.3.) [9031-76-9], which catalyzes the equilibrium between α- and β-glucose [87].

2.2.11. Manometry

Manometry is one of the classical methods in biochemistry. It is no longer used for routine assay of enzymes. Formerly, the catalytic activity of glucose oxidase [88], arginase (E.C. 3.5.3.1) [9000-96-8] [89], and other enzymes was determined by this method.

2.2.12. Viscosimetry

Viscosimetry has been practically abandoned for enzyme assays. Formerly, e.g., cellulase (E.C. 3.2.1.4) [9012-54-8] activity has been determined by the change of viscosity per unit time [90]. Nowadays the cellulase assay is performed with a colorimetric reaction [108]:

$$\text{Cellulose} \rightarrow \text{Oligosaccharide} + n\,\text{Glucose} \rightarrow \text{Red dye} \qquad (2.27)$$

2.2.13. Turbidimetry

The turbidimetric method can be adapted for different enzyme assays [109]. As an example, lysozyme (muramidase, E.C. 3.2.1.17) cleaves bacterial cell walls. A standard substrate suspension (dried germs of *Micrococcus luteus)* is used. The decrease of absorbancy is measured at 450 nm and 25 °C.

2.2.14. Immobilized Enzymes [91]

Immobilized enzymes are used in analytical chemistry and as catalysts for the production of chemicals, pharmaceuticals, and food (Section Immobilization). They also serve as simple and well-defined models for studying membrane-bound enzymes.

Because of their particular structure, immobilized enzymes require specific assays. In addition to requiring optimal conditions different from those of soluble enzymes, particle size, particle-size distribution, mechanical and chemical structure, stability and structure of the matrix, and the catalytic activity used for immobilization must be considered.

At least two different assay procedures are used for insoluble enzymes: one employs a stirred suspension in a vessel; the other, a packed bed or a column reactor. The conditions of such assay are very close to those used in industrial applications. Enzyme activity can be assayed continuously or batchwise.

Conductometry, potentiometry, and polarimetry are better suited to detection than photometry, because they allow the reaction to be followed directly, without the need for additional indicator enzymes, coenzymes, or second substrates. An example is the assay of immobilized penicillin amidase (E.C. 3.5.1.11) [9014-06-6] [92]:

$$\text{Penicillin G} + H_2O \rightarrow \text{6-Aminopenicillanate} + \text{Phenyl acetate} + H^+ \tag{2.28}$$

2.2.15. Electrophoresis

Electrophoresis is an indispensable tool for determining the catalytic activity of nucleases, especially restriction endonucleases. It is also used for other important enzymes in genetic engineering, e.g., DNA methylases. Restriction enzymes catalyze the specific cleavage of DNA, e.g., that of *Escherichia coli* phage λ DNA, which is a typical substrate in genetic engineering (see Chap. Nonindustrial Enzyme Usage). The DNA is split into smaller fragments of defined lengths. Because the negative charge per base pair is the same for all fragments, these can be separated according to length. The products of cleavage can be separated from intact or incompletely split molecules by electrophoresis under conditions that are carefully optimized for each specific enzyme – substrate reaction. Electrophoresis is usually performed with high-quality agarose gels as carrier material. Each resolution problem requires a specific agarose concentration, e.g., a relatively low concentration (0.5 g per 100 mL) for large fragments and a relatively high concentration (1.6 – 2 g per 100 mL) for small fragments. When very small fragments must be separated, polyacrylamide gels may be used as an alternative carrier material. Separated DNA fragments are visualized by staining with the fluorescent dye ethidium bromide [1239-45-8].

The fragments can be labeled by staining the gel in a separate tank after completion of electrophoresis or, more easily, by electrophoresis in the presence of the dye in the gel and buffer (e.g., at 1 µg/mL). When the gel is illuminated with long-wavelength UV light (e.g., at 366 nm), the separated fragments become visible and can be photographed.

To calculate enzyme activity, the minimum amount of enzyme must be estimated which converts a given substrate completely to the fragment pattern typical for that enzyme. For practical reasons, this amount is divided by the volume of the enzyme solution.

As an example, electrophoretic assay of the restriction enzyme *Hin*dIII is described briefly (Fig. 11) [69].

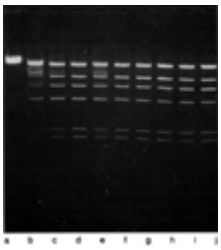

Figure 11. Determination of the catalytic activity of *Hin*dIII on λ DNA

a) λ DNA without enzyme; b – d) 1, 2, and 3 µL, respectively, 1 : 30 dilution; e – g) 1, 2, and 3 µL, respectively, 1 : 20 dilution; h – j) 1, 2, and 3 µL, respectively, 1 : 10 dilution
[Full View]

Definition of unit:

one unit is the catalytic activity of *Hin*dIII, which completely splits 1 µg of λ DNA in a total volume of 0.025 mL. The reaction is terminated after 60 min incubation at 37 °C in 0.025 mL of a defined buffer mixture.

Assay: different volumes (1 – 3 µL) of several enzyme dilutions (1 : 10, 1 : 20, 1 : 30) are incubated with 1 µg of λ DNA in a total volume of 0.025 mL. The reaction is terminated after 60 min by cooling with ice and adding 0.015 mL of a stop solution; the mixture (0.02 mL) is then placed in the slots of an agarose gel.

Agarose gel:

the gel consists of 1 g of agarose per 100 mL and 1 mg of ethidium bromide per liter. The gel dimensions are 200×200 mm; total volume is 250 mL; and slots of 1×7 mm are prepared with a comb.

Electrophoresis: the apparatus is designed for submarine electrophoresis (2 h at 100 V). The buffer system is tris(hydroxymethyl)aminomethane – acetate, 40 mmol/L, and disodium ethylenediamine-tetraacetate, 2 mmol/L; pH 8.2. The buffer contains 1 mg of ethidium bromide per liter.

Detection: the gel is illuminated directly after electrophoresis with long-wavelength UV light (366 nm) and photographed (Polaroid CU 5/film type 107); the amount of enzyme is estimated at which the complete typical fragment pattern is obtained [e.g., slot (c) in Fig. 11]. The activity is calculated according to

$$\text{Activity} = \frac{\text{Dilution}}{\text{Sample volume, mL}} \times \frac{\text{Units}}{\text{Volume, mL}} \tag{2.29}$$

(of original enzyme solution) In the example, complete digestion was obtained with a minimum of 0.003 mL at 1 : 30 dilution of the original enzyme solution, resulting in an activity of 10 000 U/mL.

2.3. Quality Evaluation of Enzyme Preparations

2.3.1. Quality Criteria

The quality of enzyme preparations is characterized by activity, purity, stability, formulation, and packaging. These parameters depend on each other, but the formulation and packaging are easy to control and keep constant. The other parameters influence each other in such a way that quality is considered to be a function of activity, purity, and stability [93].

2.3.2. Specific Activity

One of the most important quality criteria of an enzyme preparation is its *specific activity*, i.e., the catalytic activity related to its protein content. Specific activity is usually expressed as units per milligram or, for less purified products, units per gram (see Section Definition of Units).

Specific activity data can be evaluated correctly only if the specific activity can be compared with that of a highly purified enzyme of the same origin. For this purpose, catalytic activities must be measured under identical conditions including the determination of protein.

2.3.3. Protein Determination

Since protein content is the most important reference point for determination of the specific activity of an enzyme preparation, several methods of protein determination are briefly mentioned in the following paragraphs [94]. All of these procedures are based on different principles and depend on the amino acid composition of the enzyme proteins. They will, therefore, yield different values.

Ultraviolet Absorption. Because of their content of aromatic amino acids, proteins exhibit an absorption maximum at 270 – 280 nm. For many pure proteins, reference values have been established for the 280-nm absorbance of a solution containing 10 mg/mL ($A_{280}^{1\,\%}$). WARBURG and CHRISTIAN found a formula which accounts for the nucleic acid content [95]. For greater precision, absorbance is also measured at shorter wavelengths, e.g., 235 nm [96].

Biuret Method [97]. The reaction of peptide bonds with copper ions in an alkaline solution yields a purple complex which can be determined photometrically. The intensity is a linear function of protein concentration.

BCA Method [110]. The BCA method of the company Pierce is used for many protein samples. It combines the biuret method with the features of BCA:

$$\text{Protein} + \text{Cu}^{++} + \text{Bicinchoninic acid} \rightarrow \text{Red dye}\,(\text{Cu}^{+}\,\text{complex}) \qquad (2.30)$$

The complex allows the spectrophotometric quantitation of the protein in aqueous solutions.

Lowry Method [98]. The *Lowry* method combines the biuret reaction of proteins with reduction of the *Folin–Ciocalteu* phenol reagent (phosphomolybdic – phosphotungstic acid) by tyrosine and tryptophan residues. The reduction is promoted by the copper – protein complex. This method is very sensitive, but it is affected by many other compounds. The method has been modified to overcome these problems and to obtain a linear relationship between absorbance and protein content.

Protein-Dye Binding. Attempts have been made to determine protein concentration by using dyes. The method published by BRADFORD now predominates [99]. It is based on the shift of the absorption maximum of Coomassie Brilliant Blue G 250 [6104-58-1] from 465 to 595 nm, which occurs when the dye binds to the protein.

Kjeldahl Analysis. Before colorimetric procedures were established, protein concentration was calculated from nitrogen content by using an empirical factor [97].

2.3.4. Contaminating Activities

The content of contaminating activities, that is, the presence of other enzymes in the original material, is an important quality criterion for enzymes. This is usually related to the activity of the main enzyme and expressed in percent. Since the absolute amount is often very small, it cannot be expressed as protein mass (in milligrams) and, thus, does not influence the overall specific activity of the enzyme preparation. For example, lactate dehydrogenase from rabbit muscle has a specific activity of 500 U/mg and contains 0.001 % pyruvate kinase. Even if the content of pyruvate kinase increased tenfold to 0.01 %, the corresponding change in specific activity could not be measured. However, this contaminating activity is so high that such an enzyme preparation is useless for the determination of pyruvate kinase in blood.

Depending on the special applications of an enzyme, different impurities must be determined. For example, enzymes used in genetic engineering all act on a common substrate, nucleic acid; therefore, they must be free from impurities that also act on that substrate, such as specific or unspecific nucleases or phosphatases. Traces of unspecific endodeoxyribonucleases, for example, are detected routinely by incubating 10 – 100 U of an enzyme for a prolonged period (e.g., 16 h) with a susceptible substrate such as the supercoiled form of a plasmid DNA, e.g., pBR322. A minimum of 50 units should not influence the structure of that substrate, whereas another application may require that only 10 units do not change substrate structure. Such impurities are usually detected by electrophoresis (see Section Electrophoresis) or by radiometry (see Section Radiometry).

2.3.5. Electrophoretic Purity

Electrophoresis is important in the evaluation of purity because of its sensitivity (detection of less than 50 µg/mL contaminating protein is possible). It is far inferior to the determination of contaminating activities. Furthermore, enzymatically inactive proteins usually do not interfere with enzymatic analyses. Electrophoresis gained importance when isoenzymes had to be analyzed which could not be distinguished by their catalytic function but only by physical properties such as electric charge. It is an indispensable tool for the identification of lactate dehydrogenase isoenzymes.

Electrophoresis is also used in the isolation of various enzymes of RNA, DNA, and protein biosynthesis. For this purpose, the introduction of disc electrophoresis on polyacrylamide gels and the use of dodecyl sulfate for the separation of enzyme complexes became valuable tools [100]. The system developed by LAEMMLI [101] is widely used and exhibits very high sensitivity, especially if gradient gels are employed [102]. The sensitivity of the method has been enhanced considerably by silver staining [103].

Electrophoresis with 2-dimensional gels (2D PAGE) and isoelectric focussing (IEF), which separates the proteins according to their isoelectric points are commonly used methods [111], [112]. The Phast-system (Pharmacia) [113] and the electroblotting of biomolecules is applied for automatic gel electrophoresis. It consists in the transfer of electrophoretically separated proteins from the polyacrylamide matrix onto a “protein absorbing” membrane in an electrical field.

2.3.6. High-Performance Liquid Chromatography [111]

Since the 1980s the method has gained popularity for proteins. HPLC-assays are used to test enzymes for purity or to obtain information about isoenzymes. The pattern of the peaks is evaluated. Examples are peroxidase (E.C. 1.11.1.7), alkaline phosphatase (E.C. 3.1.3.1) and catalase (E.C. 1.1.1.6).

2.3.7. Performance Test

For many applications, partially purified enzyme preparations can be used, provided they do not contain any interfering contaminating activities. They are less costly and, therefore, preferred to highly purified products. However, they may contain unknown byproducts that can interfere with enzymatic analyses, for example. To avoid such problems, a performance test should be carried out. Examples are the determination of glucose with glucose oxidase and peroxidase (E.C. 1.11.1.7) [9003-99-0] or the determination of glycerol with glycerokinase (E.C. 2.7.1.30) [9030-66-4].

For some enzymes used in molecular biology, the determination of activity is not directly correlated to their application. In such cases, even highly purified enzymes must be analyzed for proper function in a typical performance experiment. An example is T_4 DNA ligase (E.C. 6.5.1.1) [9015-85-4] which functions properly by joining together fragments created by the action of a restriction endonuclease; in addition, the joined fragments can be recleaved by the same restriction endonuclease.

2.3.8. Amino Acid Analysis and Protein Sequence Analysis

Both methods are routinely used to estimate the amount and to determine the composition of enzymes. For amino acid analysis, the protein is first cleaved by complete hydrolysis (enzymatically or chemically). Then the released amino acids are assayed by quantitative chromatography. The amino acids require derivatization to improve their chromatographic behavior or their detectability. Although this is a laborious method, it is used frequently [115].

Sometimes it is necessary to characterize the enzyme sequence in detail, e.g., to check for purity this way. A useful procedure is the Edman degradation [116], [117]. With this method, the *N*-terminal amino acid is cleaved from a peptide or a protein, derivatized with phenylisothiocyanate and identified. Then the reactions are repeated with the next amino acid. Today, many instruments perform the Edman degradation automatically.

2.3.9. Stability [93]

A very important factor in the application of enzymes is their stability in concentrated or dilute form and after mixing with other substances. This applies to the manufacture of products for pharmaceutical purposes, food chemistry, or enzymatic analysis. Some enzymes can be stabilized by adding glycerol (50 vol %),

ammonium sulfate (ca. 3.2 mol/L), or sodium chloride (3 mol/L) to their aqueous solution. Furthermore, many enzymes can be kept in lyophilized form for a long period of time in the presence of stabilizers such as salts, preservatives, inert proteins (predominantly bovine serum albumin), or carbohydrates.

Most enzymes used in analysis are stored at ca. 4 °C; solutions of restriction endonucleases must be kept at – 20 °C or lower to maintain catalytic activity. To avoid degradation by moisture, the chilled enzyme preparation must be warmed before opening. Freezing and thawing may, in some cases, impair the activity of enzymes. Contamination by heavy metals or oxidants often inactivates enzymes, for example, by blocking mercapto groups. The activity of metalloenzymes or metal-dependent enzymes may be reduced by complexing agents such as ethylenediaminetetraacetate.

Especially for applications in commercial test kits some enzymes are modified by immobilization [118], for example, by covalent crosslinking with biopolymers like cellulose or dextrane. This improves the stability in solutions and/or the heat resistance.

2.3.10. Formulation of Enzyme Preparations

An enzyme preparation should be formulated according to its application. For analytical purposes, it should be easy to pipette and — if possible — free of stabilizers and preservatives that might impair its function. For example, glutamate dehydrogenase (E.C. 1.4.1.3) [9029-12-3] must not contain any traces of ammonia if it is to be used in the enzymatic determination of urea or ammonium.

In reagent kits employed for enzymatic analysis in clinical laboratories or for food analysis, the enzyme may be used preferably in lyophilized form. Compared to enzyme solutions, the solid material is in some cases easier to mix with other solid components and stable for a longer period of time, even at slightly elevated temperature. When immobilized enzymes are to be used in columns, their particle size must ensure fast flow.

Packaging. Careful selection of packaging materials is very important for handling enzymes. Bottles and stoppers used for lyophilized enzymes must be absolutely tight to prevent access of moisture. Glass or plastic bottles as well as stoppers (rubber or plastic) should not release any traces of heavy metals or other enzyme-inactivating substances into the enzyme solution or suspension. In some cases, enzymes must be protected from light and packaged in brown glass bottles.

PREVIOUS NEXT

[Top of page] [A to Z] [Authors] [Subjects] [Search] [HitList]

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Gopal K. Chotani, Genencor International, Inc., Palo Alto, USA
Timothy C. Dodge, Genencor International, Inc., Palo Alto, USA
Andreas Herman Terwisscha van Scheltinga, Gibst-brocades N. V., Delft, The Nederlands
Christian Gölker, Bayer AG, Wuppertal, Germany
Meng H. Heng, Genencor International, Inc., Palo Alto, USA
John Kan, Genencor International, Inc., Palo Alto, USA
Todd Becker, Genencor International, Inc., Palo Alto, USA
Saburo Fukui, Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
Atsuo Tanaka, Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
Rainer Schmuck, Roche Diagnostics GmbH, Tutzing, Germany

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

3. General Production Methods

3.1. Microbial Production

Use of an aerobic submerged culture in a stirred-tank reactor is the typical industrial process for enzyme production involving a microorganism that produces mostly an extracellular enzyme. This section concentrates on the typical process, with only short references to variations. Figure 12 shows the unit operations of the production process, and basic information can be found under → Biotechnology and in handbooks on microbiological principles and methods [119-121]. This section is concerned with fermentation itself, and three elements — the organism, the equipment, and the protocol for fermentation — are discussed. These elements must be arranged in the most effective way possible to realize economically optimal results in industrial processes. Optimization of biotechnological processes is a multidisciplinary effort; for further information, see [122-124].

Figure 12. Unit operations of a fermentation process
[Full View]

In the past, biotechnological process development may have looked somewhat like an art because adequate process control was lacking. This chapter should show that, based on the present state of the art, a production process can be highly controlled. To achieve this level of control in fermentation, the organism must be genetically adapted, and the process protocol must match the physiological possibilities of the organism and the limitations of the apparatus.

3.1.1. Organism and Enzyme Synthesis

A variety of different microorganisms are used for the industrial production of enzymes. They cover the taxonomic gamut from eukaryotic systems such as yeasts and fungi to prokaryotic systems from both the gram-negative and gram-positive families. When biopharmaceutical enzymes are considered as well, mammalian and insect cell lines also come into play. For most of the history of enzyme applications, production occurred in the strain known to make the enzyme of interest. This explains why so many different types of microorganisms have been employed to make enzymes. Alkaline protease is naturally secreted by *Bacillus licheniformis* to break down proteinaceous substrates and resulted in one of the first commercially produced enzymes, subtilisin Carlsberg, for use in detergents. Strains were selected that produced higher levels of protease, and an industry was born. A similar story was followed for α-amylase production. Again, *Bacillus licheniformis* naturally secreted a highly thermostable α-amylase capable of breaking down starch to more easily digestible oligosaccharides. Strains of *Bacillus* have been one of the

workhorses of enzyme production for decades, based mainly upon their ability to overproduce subtilisin and α-amylase [125].

Amylase from *Bacillus* has historically been used to liquefy starch. However, to break it down completely to single units of glucose, a second enzyme, namely, glucoamylase, is required. The most widely used enzyme for glucose production from starch is the glucoamylase from strains of the fungal genus *Aspergillus* [126]. Overproducing strains have been isolated over the years that have led to high production of glucoamylase.

An acid cellulase enzyme complex is found in the fungus *Trichoderma* [127]. This enzyme mixture was thought to be capable of breaking down cellulosic substrates to glucose, similar to the starch-degrading enzymes. This particular application was not initially commercialized. Instead, it has found applications in the treatment of textiles. New programs to improve the enzyme complex and its expression are resurrecting its potential use as an additive for the breakdown of cellulose. As with the other systems above, more productive strains have been isolated.

A final example is the production of glucose/xylose isomerase. Glucose isomerase catalyzes the rearrangement of glucose to fructose, yielding a product with a sweeter taste. The enzyme discovered to be capable of industrial conversion originated from a species of *Steptomyces*. As with the others, the native host strain has been used for production of glucose isomerase.

All of the strain types described in the examples above are capable of differentiation. For example, *Bacillus* thrives in nature due to its ability to survive various insults by forming a non growing, yet viable, spore. The spore remains dormant until it reaches an environment where it can germinate and divide. Regulation of protease production is associated with this differentiation. This highly complex behavior has been difficult to model for aiding process development. The same can be said of the fungal species *Aspergillus* and *Trichoderma*. Their differentiation makes process development and modeling difficult. Although *Streptomyces* does not truly sporulate, it differentiates by forming filaments as opposed to isolated individual cells. This property also has an effect on product formation and physical properties of the fermentation broth.

An organism can be viewed as a metabolic system converting substrates into cell mass and byproducts. Enzymes function in this system as catalysts for the different reactions. Each cell is equipped with mechanisms that regulate the synthesis and activity of the enzymes to enable the cell to respond adequately to environmental changes. Therefore, in its elementary form, an organism can be described as a set of metabolic components with a mechanism for enzyme synthesis and a regulatory apparatus. The kinetics of the process are determined by structural components of the organism and by various physical and chemical factors, as shown in Figure 13. The key role that enzymes play in biological processes has led to an extensive study of both the mechanism and the regulation of their synthesis [128], [129], and some of the results are related to the problem of process development.

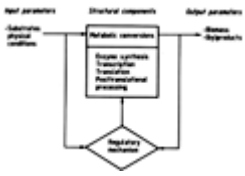

Figure 13. Scheme of the metabolic process
[Full View]

The basic mechanisms of enzyme synthesis, including transcription, translation, and posttranslational processing, seem to be highly conserved [124]. However, several differences exist between various classes of organisms, as well as some fundamental differences between prokaryotic and eukaryotic organisms. The enzymes themselves differ enormously in molecular mass, number of polypeptide chains, isoelectric point, and degree of glycosylation. In addition, a variety of enzyme-producing species exist [131]. Although all the differences may influence the characteristics of synthetic patterns, the basic mechanisms underlying enzyme synthesis are similar enough to allow a general treatment of the microbiological production process. However, the differences in production kinetics among various species are large enough to make individual optimization programs necessary.

Different organisms may also differ in their suitability for fermentation; such process characteristics as viscosity or recoverability, legal clearance of the organism, and knowledge available on the selected organism, must be considered.

Because of the action of the regulatory mechanism, enzyme synthesis rates range from no synthesis to maximum synthesis allowed by the synthetic apparatus, as in a normal control loop. The complexity of mechanisms ranges from relatively simple and well-understood induction and repression systems, to very complex global regulatory networks [132].

Process development must deal with the complexity of the enzyme synthesis system either by changing the structural characteristics, including structural elements of the regulatory systems (see Section Strain Improvement), or by selecting optimal environmental conditions (see Section Physiological Optimization).

3.1.2. Strain Improvement

Most of the strains used for enzyme production have been improved through classical selection. Mutagenesis by chemical agents or UV radiation has been used to more quickly find useful variants. Many cells must be subjected to a mutation or recombination procedure and then tested for the desired combination of characteristics by selection. The success of strain improvement programs often depends on development of an effective selection method for finding one mutant among 10 000–100 000 cells. Methods range from plate selection to the continuous culture technique [137], [138].

Mutation changes the protein structure and most probably results in a deterioration of function. Changes in structural components by mutation are therefore rarely improvements unless the specific loss of function is required for production purposes, e.g., when a loss of regulatory function results in enhanced enzyme production.

Mutation and selection are directed primarily toward higher overall productivity rather than mutation of a specific function, but a loss of regulatory function is highly probable. However, some studies describe screening for a mutation in a specific function [135].

Based on the complex interdependency of pathways and the competition for substrates in the organism, random mutation may also affect the rate of enzyme synthesis. However, only minor improvements should be expected, although a series of mutations can still result in an interesting degree of improvement. In vivo recombination techniques have been used to complement mutation techniques by bringing together mutations in different cell lines and cleaning strains of undesired deleterious mutations [122]. Strain improvement has been revolutionized by the advent of genetic engineering. The ability to specifically improve strains by the manipulation of the host genome or the addition of extra-chromosomal DNA elements has greatly increased the speed of strain improvement. In the early stages, most work was performed in strains for which methods to manipulate DNA were established. Attempts were made to express many different enzymes in *Escherichia coli*. However, *E. coli* typically does not excrete such proteins and instead deposits them in an insoluble form (called inclusion bodies) in the cytoplasm of the cell. It proved more effective to develop DNA manipulation techniques in the native host strain of the different enzymes. Today, nearly all production strains used commercially have been genetically engineered. Strains have been improved through extra copies of the gene of interest, removal of inhibitory regulation, and enhancement of positive regulation.

If the absence of regulation is assumed, considerable improvement can be realized by increasing the turnover rate of the limiting step in synthesis. Techniques of genetic engineering can be used to increase the rate of mRNA synthesis by constructing plasmids with the desired gene (gene multiplication). The number of plasmids per cell can be very high. The rate of mRNA synthesis should be related linearly to the number of gene copies, but the rate of enzyme synthesis may be limited again by the next step in the process. These techniques also allow replacement of the promoter by a more effective one [139]. The same methods allow replacement of the leader sequence of a gene, resulting in the excretion of a formerly intracellular enzyme [140].

Genetic engineering also uses microorganisms to produce enzymes of higher organisms by placement of the corresponding gene into the microorganism. The presence of introns may then prevent proper expression of the gene, but techniques have been developed to overcome this difficulty [136]. Chymosin (E.C. 3.4.23.4) [

9001-98-3], calf rennet, has been cloned by several groups either in prokaryotes or in yeast, one of the first cloned mammalian enzyme to be produced industrially by microorganisms [141].

Novel methods to improve strain performance are under development (ca. 2002). These include gene shuffling and directed evolution. Most of the literature regarding these techniques has focused on the improvement of specific enzyme properties (see Section Protein Engineering). However, directed evolution is now being applied to the overproduction of given enzymes as well. By mutagenizing a plasmid harboring a gene for subtilisin, enzyme production was increased five fold [142]. It was shown that the structural gene was unchanged and that the effect was most likely due to an improved promoter or higher plasmid copy number leading to higher rates of transcription.

3.1.3. Physiological Optimization

Figure 13 shows that not only the structural characteristics of the system, but also the input parameters, determine its metabolic activity. Once a suitable organism has been found, either genetically improved or not, the next task of process development is to define optimal input parameters.

The rate of enzyme synthesis r_e is defined by:

$$r_e = q_e \cdot c_x$$

where r_e is the enzyme synthesis rate in units per liter per hour (for definition of unit, see Section Definition of Units), q_e the specific enzyme synthesis rate in units per gram of biomass per hour, and c_x the biomass concentration in grams per liter.

In kinetic terms, optimization of the enzyme synthesis rate means finding the highest specific synthesis rate for a given amount of biomass. Because enzyme synthesis depends so much on primary metabolism, the complete cellular machinery must function at all times. Therefore, conditions favoring enzyme synthesis also favor growth, as suggested in Figure 13. The search for optimal conditions for enzyme synthesis can then be reduced to a search for (1) the conditions influencing growth rate and (2) the relation between specific enzyme synthesis rate and growth rate.

Growth can be expressed by a specific growth rate μ as the amount of biomass synthesized per unit biomass and unit time:

$$\mu = \frac{1}{c_x} \cdot \frac{dc_x}{dt}$$

where c_x is the biomass concentration in grams per liter and t the time in hours.

Specific enzyme synthesis rate and specific growth rate express the metabolic activities of the cell. These are abbreviated as synthesis rate q_e and growth rate μ. MONOD [143] formulated the relationship between growth rate and substrate concentration as:

$$\mu = \mu_{max} \cdot \frac{c_s}{K_s + c_s}$$

where μ_{max} is the maximum obtainable growth rate per hour; K_s is the saturation, and c_s the concentration of substrate, both expressed as moles per liter.

According to this relation actual growth rates at low substrate concentration are lower than the maximum possible value determined by the structural properties of the organism. When $c_s \gg K_s$, the growth rate reaches its maximum value. If all substrates are present in excess ($c_{s_i} \gg K_{s_i}$ for any i), the growth rate is limited by internal structure. The parameters μ_{max} and K_s are dependent on pH, temperature, osmotic pressure, and such factors as medium type (minimal vs. nutrient medium). When all substrates are present in

excess except one ($\mu < \mu_{max}$), that substrate is called the limiting factor; in this case, an external growth limitation exists. Nutrient-limited growth is considered the natural state for microorganisms [144].

The kinetics of enzyme synthesis cannot be expressed in one equation for all enzymes. Instead, a number of patterns of synthesis rate q_e versus growth rate μ are shown in Figure 14. The curves are somewhat idealized, but many examples can be found in the literature [145-148].

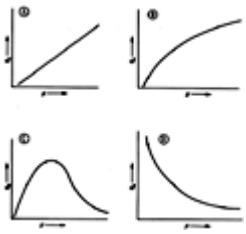

Figure 14. Types of relationship between specific synthesis rate q_e and specific growth rate μ

A): Growth-coupled synthesis [147]; B): Saturated synthesis [148]; C): Saturated synthesis with catabolic repression [151]; D): Repressed synthesis [154]
[Full View]

The importance of these physiological considerations is that a limiting factor can be used to control growth rate and thereby synthesis rate. If the q_e versus μ relationship is known, the value of μ that results in a maximum q_e corresponds to the optimal physiological condition for synthesis.

Complete physiological characterization includes knowledge of the relationship of q_e to μ, the yield on different substrates, the effect of the limiting factor, the physicochemical conditions, and the possible role of regulatory agents. Because these factors are interdependent, establishment of the complete physiological characteristics of an organism is a very extensive task. In industrial practice, experience, general knowledge, and use of well-known organisms or techniques are generally employed to find shortcuts.

The previous discussion emphasizes the importance of the limiting factor for control of growth and product synthesis. Control of growth is possible if the substrate concentration c_s is in the range of the saturation concentration K_s. The K_s values for most substrates (± 0.1 mmol/L) are rather low compared to the consumption rate (± 10 mmol L^{-1} h^{-1} for glucose), i.e., the response time constant (less than a minute) is low. Continuous feed is therefore necessary to control growth rate by limiting substrate concentration, as applied in fed-batch and continuous cultures. In batch cultures, all substrates are present at the start of fermentation, and the growth rate is maximal until the substrate is nearly consumed. Therefore, direct control of growth is not possible in a batch culture.

One of the applications of metabolic-pathway engineering to fermentation production technologies is the ever-expanding “toolbox” to enable mining of biodiversity, maximize productivity, enhance carbon efficiency, improve product purity, expand product lines, and broaden markets. Functional genomics, proteomics, metabolomics, fluxomics, physiomics, and cellomics applications holistically delineate metabolic pathways and environment within the host [149]. Functionomics-based optimization includes four key elements: (1) completion and optimization of the primary metabolic pathway to the targeted product, including removal of rate-limiting, transcriptional, and allosteric regulation; (2) genetic blockage of competing pathways; (3) enhanced carbon commitment to the primary metabolic pathway from central metabolism; and (4) modification of secondary metabolic pathways as necessary to enhance energy metabolism and availability of required enzymatic cofactors. Recently, strain-specific stoichiometry (metabolic genotype) has been woven into flux balance models to determine optimal flux distributions (metabolic phenotype) by using linear programming. A combination of such approaches may prove to be valuable for design and optimization of bioprocesses [150].

3.1.4. The Fermentor and its Limitations

Fermentation process design is interdisciplinary and uses concepts and methodologies of both chemical engineering and microbial physiology to accomplish scale-up. In practice, scale-up effects are more pronounced for aerobic, i.e., aerated and agitated environment, than for anaerobic fermentations. Therefore, as a rule of thumb, in an aerobic fermentor, the constant oxygen transfer rate and concentration of dissolved oxygen are generally maintained in the scale-up. While thermodynamic and kinetic phenomena are

independent of scale, momentum, mass and heat transfer are functions of scale. Mixing, aeration, and cooling are well controlled and uniform at 10 L scale, but not all transport parameters can be maintained in this way at large scale. Further scale-up complications arise from cell response to distributed values of dissolved oxygen, temperature, pH, and nutrients.

The development of entirely new processes or the improvement of existing processes requires the evaluation of a wide range of strains and cultivation conditions in a short period of time. Shake-flask fermentation studies have a cost advantage but lack process control options (pH, nutrient addition, aeration). This often leads to use of laboratory-scale (ca. 10 L) agitated, aerated fermentors with adequate instrumentation and control. Most fermentation processes can be translated to production scale by use of laboratory-scale fermentors. However, pilot-scale fermentors are often necessary for downstream process scale-up.

For enzyme production, economy of scale leads to the use of fermentors with a volume of 20–200 m^3. The higher energy yield from aerobic metabolism results in the use of aerobic processes which require continuous transfer of poorly soluble oxygen into the culture broth. The concomitant problems of mass and heat transfer are usually neglected in small fermentors and at low cell densities. However, in industrial microbiology, with the above-mentioned fermentor volumes and the economic necessity of using the highest possible cell densities, transport processes must be considered. Such processes can limit metabolic rates; for example, the oxygen supply may become limiting and the microorganism may respond by changes in its physiological pattern. Under these conditions, the desired control of microbial metabolism may be lost. In controlled operation of an industrial process, metabolic rates must be limited to a level just below the transport capacity of the fermentor. Therefore, the highest possible productivity in a fermentor is obtained at maximal transport capacity. This is reflected in low concentrations of dissolved oxygen during fermentation, as shown in Figure 15.

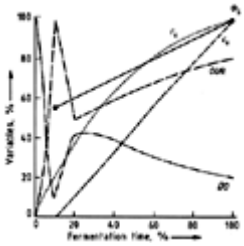

Figure 15. Schematic representation of a fermentation

This illustrates the production of amylase by *B. licheniformis*. After an inital exponential growth period, the process is controlled by an increasing feed rate which results in the oxygen uptake rate and biomass profile shown. The final result is an almost constant rate of enzyme synthesis during fermentation

Variables (expressed as percent of maximum value during fermentation):

c_x, biomass concentration; c_e, enzyme activity; *DO*, dissolved oxygen; *OUR*, oxygen uptake rate; Φ_s, carbon feed rate
[Full View]

Maximizing the transport processes is chiefly a problem of fermentor design and is generally treated in handbooks on bioengineering [121], [123], [151], [152], [153]. In the common stirred-tank reactor, agitation and injection of compressed air are used to mix the contents of the reactor and to transfer gases. The oxygen transport rate *n*, for example, can be expressed as:

$$n = k_1 a[c_0(\mathrm{g}) - c_0(1)]$$

where $k_1\ a$ is the transfer coefficient per hour; c_0 (g) is the oxygen concentration in the liquid phase in equilibrium with the gas phase, and c_0 (l) the actual oxygen concentration in the medium, both expressed in millimoles per liter.

The transfer coefficient $k_1\ a$ shows that the main resistance to oxygen transport is the gas – liquid interface, where k_1 is the resistance coefficient in meters per hour, and *a* the total gas – liquid interface area in square meters per cubic meter. The transfer capacity of a fermentor can be expressed as the $k_1\ a$ value. The $k_1\ a$ value is determined primarily by the power input of the agitator and the air jet system [154]. Bioengineering therefore aims at the design of a fermentor with the highest $k_1\ a$ value at the lowest power input.

The $k_1 a$ value is also influenced by factors such as viscosity, ionic strength, and presence of surfactants. The dependence of $k_1 a$ on viscosity explains the preference for using pellet growth to cultivate mycelium-forming microorganisms [155].

The complex relationships of factors determining transport processes make the scale-up of biotechnological processes a very difficult task [122].

3.1.5. Process Design

Process design entails the development of an optimum outcome for process performance. This typically begins at the fermentation stage, but must also take into account downstream operations such as cell harvesting, product isolation, enzyme purification, and final product formulation. The way in which these unit operations interact can have a profound effect upon optimum process performance.

The fermentation stage begins with culture storage. Cell lines can be stored for future use in a number of ways. These include cultures frozen in liquid media with an appropriate cryoprotectant, on or within solid media such as agar, as freeze-dried suspensions, and as spore suspensions. The process of reviving the stored culture is an oft-neglected phase of the fermentation process. Development time typically focuses on operation of the production fermentor, and the seed stages are run simply to grow the cells as quickly as possible. A typical seed train is shown in Figure 16. However, the physiological state in which the cells are cultivated during seed propagation can have a profound effect upon fermentor performance. This is especially true for microorganisms that can differentiate, either through filamentation or sporulation. For genetically engineered strains, maintaining strain stability is of utmost concern.

Figure 16. Typical fermentation train
[Full View]

The desired outcome of the fermentation process is often to maximize the enzyme synthesis rate, or volumetric productivity. This rate is the product of the specific enzyme production rate and the cell concentration (see Section Physiological Optimization). Both of these factors can be influenced by process design. If the specific production rate is a strong function of growth rate, i.e., growth-associated production, maintenance of high growth rate will be of high priority. In such cases, continuous-culture techniques are quite suitable. The constant removal of cell mass allows for sustained high growth rate and specific productivity. The high metabolic rate associated with high growth rate can limit the achievable cell mass concentration. Oxygen- and heat-transfer limitations of the fermentation equipment are usually the limiting factors.

Some enzymes are produced more as a secondary metabolite, and specific productivity may then be an inverse function of growth rate, i.e., non-growth-associated production. Here a recycling reactor may be most suitable. A recycling reactor is similar to the continuous culture, but a device is added to return a significant fraction of the cells to the reactor. Low growth rates with high cell concentration can often be achieved in such systems.

In many cases, however, the relationship between growth rate and specific productivity can be quite complex, so that neither the continuous-culture nor the recycling fermentor is a good choice. This is where fed-batch processes are most prevalent. Starting with a relatively low cell mass concentration, growth rate can be controlled throughout the course of the fermentation through feeding of a growth-limiting substrate.

The discussion above is focused on optimization of the fermentation process. Yet the goal of most processes is to produce a final formulated enzyme product and will likely include many post-fermentation unit operations. For the entire process, maximum production rate could still be the most important factor. However, lowest unit production cost could also be an important driving force. Optimization of each individual unit operation will not always lead to the optimal overall process performance, especially when there are strong interactions between unit operations [156]. Understanding of these interactions is crucial to overall process optimization. For instance, product concentration or purity in the fermentation broth can

significantly impact downstream purification unit operation. If the fermentation is optimized for productivity, without taking into account its effect on the purification step, the overall process productivity can be negatively affected. The use of antifoaming agents in the fermentation process is another example of such a trade-off. By reducing foaming in the fermentation, a higher working volume can be used to optimize the fermentation unit operations. However, many antifoaming agents negatively impact ultrafiltration membranes and reduce the capacity of this recovery unit operation [157]. Knowledge of how the fermentation process will affect other downstream unit operations is also of great importance.

3.1.6. Modeling and Optimization

Many enzyme production processes are carried out on a large scale (> 100 000 L). Small increases in performance can have a significant impact on economic viability. Therefore, process optimization is of great importance for enzyme production by fermentation. Process modeling can give a structured approach to the optimization problem so that optimal conditions can quickly be reached.

Relationships between controllable parameters and desired outputs are a prerequisite for successful process modeling. Balance equations combined with kinetic expressions have been the most widely used tools for such modeling. Monod-type kinetic expressions (see Section Physiological Optimization) are a typical starting point. This type of modeling has been successfully applied to the improvement of enzyme production processes. Kinetic expressions of cell growth, substrate utilization, and product formation are formulated. Whenever possible, independent experimental procedures are utilized to determine model parameters. Fermentation experiments are then performed to fit any remaining parameters. Mathematical simulations are then run to predict the performance of the system based upon known adjustable parameters, such as substrate feed rate. Optimization can be carried out “in silico” for any given objective function (e.g., maximum product concentration or maximum production rate). In fact, combinations of different parameters can simultaneously be optimized to meet conflicting constraints [158]. Experimental fermentations can then be used to calibrate and validate the model. Adjustments to the model architecture can be made to improve the predictive capability of the model. Continual iteration between model modification and experimental results is a trademark of such kinetic modeling.

The drawback of such models is the prerequisite for relationships between the controllable parameters and the desired outputs. One way around this impediment is the use of artifical neural networks (ANNs) [159], [160], [162]. These adaptive computer programs have the ability to learn and can be used for predictions if properly trained. Advantageously, many enzyme production processes have a large repository of past data that can be used to build and train such models. The ANN consists of a set of input and output variables determined by the process being investigated. A set of nodes is then established between the inputs and outputs, each with a different weighting factor. The optimal number of nodes is typically found through trial and error. The weighting factors are continually updated by “training” on actual process data. New inputs and outputs can be added to help improve the predictive power of the model. Such procedures have been used for enzymes and other microbially derived products.

ANNs can be extended by combining them with a more structured model to create hybrid models [161]. Input variables can be placed into an ANN and the output variables can then be used as the input to a set of balance equations. Such models have shorter training times as they combine the black-box approach of the ANN with known relationships between controllable parameters and desired outputs.

3.1.7. Instrumentation and Control

Like any chemical reactor, fermentors are controlled by monitoring and estimating several critical parameters such as pressure, temperature, pH, dissolved oxygen, redox potential, foam level, heat flux via coolant supply, respiration (carbon dioxide, oxygen) rate, mass- and heat-transfer coefficients (Figure 17). Mass spectrometry is commonly used for analyzing the off-gas composition for calculating oxygen uptake rate, carbon dioxide evolution rate, oxygen transfer coefficient k_1a, respiration coefficient, and carbon balance.

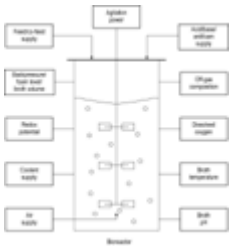

Figure 17. Bioreactor monitoring and control
[Full View]

Typically, fermentation is optimized by careful design of batch and feeding media, in order to regulate cell growth and maximize production. When an optimal trajectory of the feed rate has been found and desired values for physicochemical variables have been established, the process is ready for production scale. Elementary physical variables, pH, and feed rate are maintained by proper control. If process development is of reasonable quality, the resulting output will be somewhere in the expected range. In practice, however, deviations between actual and expected results may arise for several reasons: (1) the model and the experimental knowledge are never complete; (2) errors of measurement and set point always exist, along with deviations in the variables regulated; (3) raw materials may vary in composition; and (4) transport conditions can cause deviations in response.

If deviations from the expected results are minimal, the process can be operated with a minimum level of instrumentation and control. If the deviations cannot be ignored, output variables may be measured to correct input variables; for example, the carbon dioxide production rate can be used for feedback regulation of the feed rate. More advanced control can be realized when a model of the process is used to relate different input and output variables. The measured input and output variables of the process can be compared with the model, and some type of regulatory action can be derived from this comparison [168].

Fermentation processes are complicated, and in principle, control of the process can be improved by measuring as many variables as possible and using them for computer-based process control and optimization. A general limitation of such control systems is lack of adequate (cheap and sterilizable) sensors for measuring the chemical and biological variables [169]. Another problem is the lack of persons to develop the software and operate such systems [170]. Optimistically, reliable measurements and models will lead to on-line estimation and control of variables not measured, using balances for mass and elements, and phenomenological relations for mass transfer and acid-base equilibria in combination with the electroneutrality condition [171].

3.2. Isolation and Purification [172-187]

The degree of purity of commercial enzymes ranges from raw enzymes to highly purified forms and depends on the application. Raw materials for the isolation of enzymes are animal organs, plant material, and microorganisms.

Enzymes are universally present in living organisms; each cell synthesizes a large number of different enzymes to maintain its metabolic reactions. The choice of procedures for enzyme purification depends on their location. Isolation of intracellular enzymes often involves the separation of complex biological mixtures. On the other hand, extracellular enzymes are generally released into the medium with only a few other components. Enzymes are very complex proteins, and their high degree of specificity as catalysts is manifest only in their native state. The native conformation is attained under specific conditions of pH, temperature, and ionic strength. Hence, only mild and specific methods can be used for enzyme isolation. Figure 14 shows the sequence of steps involved in the recovery of enzymes.

3.2.1. Preparation of Biological Starting Materials

Animal Organs. Animal organs must be transported and stored at low temperature to retain enzymatic activity. The organs should be freed of fat and connective tissue before freezing. Frozen organs can be minced with machines generally used in the meat industry, and the enzymes can be extracted with a buffer solution. Besides mechanical grinding, enzymatic digestion can also be employed [188]. Fat attached to the organs interferes with subsequent purification steps and can be removed with organic solvents. However, enzymatic activity might be influenced negatively by this procedure.

Plant Material. Plant material can be ground with various crushers or grinders, and the desired enzymes can be extracted with buffer solutions. The cells can also be disrupted by previous treatment with lytic enzymes.

Microorganisms are a significant source of enzymes. New techniques, summarized under genetic and protein engineering, have much to offer the enzyme industry. A gene can be transferred into a microorganism to make that organism produce a protein it did not make naturally. Alternatively, modification of the genome of a microorganism can change the properties of proteins so that they may be isolated and purified more easily. Such modifications might, for example, cause the release of intracellular enzymes into the medium; change the net charge and, therefore, the chromatographic properties of proteins; or lead to the formation of fused proteins [189].

Most enzymes used commercially are *extracellular enzymes*, and the first step in their isolation is separation of the cells from the solution. For *intracellular enzymes*, which are being isolated today in increasing amounts, the first step involves grinding to rupture the cells. A number of methods for the disruption of cells (Table 2) are known, corresponding to the different types of cells and the problems involved in isolating intracellular enzymes. However, only a few of these methods are used on an industrial scale.

Table 2. Methods for disruption of cells

Mechanical methods	Nonmechanical methods	
High pressure (Manton – Gaulin, French-press)	Drying (freeze-drying, organic solvents)	
Grinding (ball mill)	Lysis	
Ultrasound	physical:	freezing, osmotic shock
	chemical:	detergents, antibiotics
	enzymatic:	enzymes (e.g., lysozyme), antibiotics

3.2.1.1. Cell Disruption by Mechanical Methods

High-pressure homogenization is the most common method of cell disruption. The cell suspension is pressed through a valve and hits an impact ring (e.g., Manton – Gaulin homogenizer). The cells are ruptured by shearing forces and simultaneous decompression. Depending on the type of machine, its capacity ranges from 50 to 5000 L/h. The rigid cell walls of small bacteria are only partially ruptured at the pressures up to 55 MPa (550 bar) achieved by this method. Higher pressures, however, would result in further heat exposure (2.2 °C per 10 MPa). Hence, the increased enzyme yield resulting from improved cell disruption could be counteracted by partial inactivation caused by heating and higher shearing forces. Therefore, efficient cooling must be provided.

The *wet grinding* of cells in a high-speed bead mill is another effective method of cell disruption [190-193]. Glass balls with a diameter of 0.2 – 1 mm are used to break the cells. The efficiency of this method depends on the geometry of the stirrer system. A symmetrical arrangement of circular disks gives better results than the normal asymmetrical arrangement [194]. Given optimal parameters such as stirring rate, number and size of glass beads, flow rate, cell concentration, and temperature, a protein release of up to 90 % can be achieved in a single passage [190].

3.2.1.2. Cell Disruption by Nonmechanical Methods

Cells may frequently be disrupted by *chemical*, *thermal*, or *enzymatic lysis*. The drying of microorganisms and the preparation of acetone powders are standard procedures in which the structure of the cell wall is altered to permit subsequent extraction of the cell contents. Methods based on enzymes or autolysis have been described in the literature [195-198]. Ultrasound is generally used in the laboratory. In this procedure, cells are disrupted by shearing forces and cavitation. An optimal temperature must be maintained by cooling

the cell suspension because heat is generated in the process. Additional problems may arise from generation of free radicals.

3.2.2. Separation of Solid Matter

After cell disruption, the next step is separation of extracellular or intracellular enzymes from cells or cellular fragments, respectively. This operation is rather difficult because of the small size of bacterial cells and the slight difference between the density of the cells and that of the fermentation medium. *Continuous filtration* is used in industry. Large cells, e.g., yeast cells, can be removed by *decantation.* Today, efficient *centrifuges* have been developed to separate cells and cellular fragments in a continuous process. Residual plant and organ matter can be separated with simpler centrifuges or filters.

3.2.2.1. Filtration

The filtration rate is a function of filter area, pressure, viscosity, and resistance offered by the filter cake and medium (→ Filtration). For a clean liquid, all these terms are constant which results in a constant flow rate for a constant pressure drop. The cumulative filtrate volume increases linearly with time. During the filtration of suspensions, the increasing thickness of the formed filter cake and the concomitant resistance gradually decrease the flow rate. Additional difficulties may arise because of the compressibility of biological material. In this case, the resistance offered by the filter cake and, hence, the rate of filtration depend on the pressure applied. If the pressure applied exceeds a certain limit, the cake may collapse and total blockage of the filter can result.

Pressure Filters. A *filter press* (plate filter, chamber filter) is used to filtrate small volumes or to remove precipitates formed during purification. The capacity to retain solid matter is limited, and the method is rather work-intensive. However, these filters are highly suitable for the fine filtration of enzyme solutions.

Vacuum Filters. Vacuum filtration is generally the method of choice because biological materials are easily compressible. A *rotary vacuum filter* (Figure 15) is used in the continuous filtration of large volumes. The suspension is usually mixed with a filter aid, e.g., kieselguhr, before being applied to the filter. The filter drum is coated with a thin layer of filter aid (precoat). The drum is divided into different sections so that the filter cake can also be washed and dried on the filter. The filter cake is subsequently removed by using a series of endless strings or by scraper discharge (knife). The removal of a thin layer of precoat each time exposes a fresh filtering area. This system is useful for preventing an increase in resistance with the accumulation of filter cake during the course of filtration.

Cross-flow filtration [198], [200]. In conventional methods, the suspension flows perpendicular to the filtering material (Fig. 18 A). In cross-flow filtration, the input stream flows parallel to the filter area (Fig. 18 B), thus preventing the accumulation of filter cake and an increased resistance to filtration. To maintain a sufficiently high filtration rate, this method must consume a relatively large amount of energy, in the form of high flux rates over the membranes. With the membranes now available, permeate rates of 30 – 50 $Lm^{-2}\ h^{-1}$ can be attained. Indeed, in many cases the use of a separator (see Section Centrifugation) is more economical [199]. The future of this method depends on the development of suitable membranes, but cross-flow filtration can be conveniently used in recombinant DNA techniques to separate organisms in a closed system.

Figure 18. Sequence of steps in the isolation of enzymes [Full View]

3.2.2.2. Centrifugation

The sedimentation rate of a bacterial cell with a diameter of 0.5 µm is less than 1 mm/h. An economical separation can be achieved only by sedimentation in a centrifugal field (→ Centrifuges, Sedimenting). The

range of applications of centrifuges depends on the particle size and the solids content (Table 3).

Table 3. Utilization of different centrifuges

Type of centrifuge	Solids content, %	Particle size, µm
Multichamber separator	0 – 5	0.5 – 500
Desludging disk separator	3 – 10	0.5 – 500
Nozzle separator	5 – 25	0.5 – 500
Decanter	5 – 40	5 – 50 000
Sieve centrifuge	5 – 60	5 – 10 000
Pusher centrifuge	20 – 75	100 – 50 000

The Σ value of a centrifuge is a good criterion for the comparison of centrifuges:

$$\Sigma = F \cdot Z$$

where $F = V/S$, with V the total volume of liquid in the centrifuge and S the thickness of the liquid layer in the centrifuge; i.e., F has the units of an area. $Z = r \cdot \omega^2/g$, where r is the radius of the centrifuge drum, ω the angular rotation speed, and g the gravitational constant.

Both sieve centrifuges and solid-wall centrifuges are available. Typical solid-wall centrifuges are shown in Figure 19.

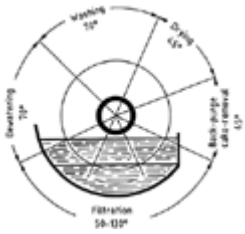

Figure 19. Rotary vacuum filter
[Full View]

Decanters (scroll-type centrifuges) work with low centrifugal forces and are used in the separation of large cells or protein precipitates. Solid matter is discharged continuously by a screw conveyer moving at a differential rotational speed.

Tubular bowl centrifuges are built for very high centrifugal forces and can be used to sediment very small particles. However, these centrifuges cannot be operated in a continuous process. Moreover, solid matter must be removed by hand after the centrifuge has come to a stop. A further disadvantage is the appearance of aerosols.

Separators (disk stack centrifuges) can be used in the continuous removal of solid matter from suspensions. Solids are discharged by a hydraulically operated discharge port (intermittent discharge) or by an arrangement of nozzles (continuous discharge). Bacteria and cellular fragments can be separated by a combination of high centrifugal forces, up to 15 000 × gravity, presently attainable, and short sedimentation distances. Disk stack centrifuges that can be sterilized with steam are used for recombinant DNA techniques in a closed system [201], [202].

3.2.2.3. Extraction

An elegant method used to isolate intracellular enzymes is liquid – liquid extraction in an aqueous two-phase system [203-207]. This method is based on the incomplete mixing of different polymers, e.g., dextran [9004-54-0] and poly(ethylene glycol) [25322-68-3], or a polymer and a salt in an aqueous solution [208]. The first extraction step separates cellular fragments. Subsequent purification can be accomplished by extraction or, if high purity is required, by other methods. The extractability can be improved by using affinity ligands [209], [210] or modified chromatography gels, e.g., phenyl-Sepharose [211].

3.2.2.4. Flocculation and Flotation

Flocculation. Separation of bacterial cells or cell debris by filtration or centrifugation can involve considerable difficulties due to their small size and physical properties. The compressible nature of the cells is the primary limiting factor for using filtration as a separation step to remove them. The low permeability of a typical cell cake results in a filtration rate that is often too slow to be practical. In cell removal by centrifugation, the small size and low density difference between the cells or cell debris and the medium results in a low sedimentation rate. Flocculation of cell suspensions has been reported to aid cell separation by both filtration and centrifugation [212-218].

Flocculation is the process whereby destabilized particles are induced to come together, make contact, and subsequently form larger aggregates. Flocculating agents are additives capable of increasing the degree of flocculation of a suspension (→ Flocculants) . They can be organic or inorganic, and natural or synthetic. A comprehensive review of various categories of flocculating agents can be found in [219], [220]

Synthetic organic flocculating agents are by far the most commonly used agents for cell flocculation in industrial processes. They are typically water-soluble, charged polymeric substances with average molecular weight ranging from about 10^3 to greater than 5×10^6 and are generally referred to as polyelectrolytes. The positively and negatively charged polymers are referred to as cationic and anionic polyelectrolytes, respectively. Polyelectrolytes containing both positive and negative charges are termed polyampholytes.

Flocculation of cells by polyelectrolytes is a two-step process. The first step is the neutralization of the surface charge on the suspended cells or cell debris. The second step involves the linkage of these particles to form large aggregates. The various mechanisms and theories of flocculation have been summarized [219-222].

Flocculant selection for a specific cell separation process is a challenge as many factors can impact flocculation. These factors can have their origin in the broth (cell surface charge and size, ionic strength, pH, cell concentration, and the presence of other charged matter), the polymer (molecular weight, charge and charge density, structure, type), and engineering parameters (mixing and mode and order of addition). The final criteria for flocculant selection should take into consideration all aspects of the flocculation process. These include the cost of the added flocculant, subsequent separation performance, process robustness, and yield. In some cases, flocculation can also provide purification by selectively removing unwanted proteins, nucleic acids, lipids and endotoxin from the cell broth [223-226].

Flotation. If no stable agglomerates are formed, cells can be separated by flotation. Here, cells are adsorbed onto gas bubbles, rise to the top, and accumulate in a froth. An example is the separation of single cell protein [227].

3.2.3. Concentration

The enzyme concentration in starting material is often very low. The volume of material to be processed is generally very large, and substantial amounts of waste material must be removed. Thus, if economic purification is to be achieved, the volume of starting material must be decreased by concentration. Only mild concentration procedures that do not inactivate enzymes can be employed. These include *thermal methods, precipitation,* and to an increasing extent, *membrane filtration.*

3.2.3.1. Thermal Methods

Only brief heat treatment can be used for concentration because enzymes are thermolabile. Evaporators with rotating components that achieve a thin liquid film (thin-layer evaporator, centrifugal thin-layer evaporator) or circulation evaporators (long-tube evaporator) can be employed.

3.2.3.2. Precipitation

Enzymes are very complex protein molecules possessing both ionizable and hydrophobic groups which interact with the solvent. Indeed, proteins can be made to agglomerate and, finally, precipitate by changing their environment. Precipitation is actually a simple procedure for concentrating enzymes [228].

Precipitation with Salts. High salt concentrations act on the water molecules surrounding the protein and change the electrostatic forces responsible for solubility. Ammonium sulfate [7783-20-2] is commonly used for precipitation; hence, it is an effective agent for concentrating enzymes. Enzymes can also be fractionated, to a limited extent, by using different concentrations of ammonium sulfate. The corrosion of stainless steel and cement by ammonium sulfate is a disadvantage, which causes additional problems in wastewater treatment. Sodium sulfate [7757-82-6] is more efficient from this point of view, but it is less soluble and must be used at temperatures of 35 – 40 °C. The optimal concentration of salt required for precipitation must be determined experimentally, and generally ranges from 20 to 80 % saturation.

Precipitation with Organic Solvents. Organic solvents influence the solubility of enzymes by reducing the dielectric constant of the medium. The solvation effect of water molecules surrounding the enzyme is changed; the interaction of protein molecules is increased; and therefore, agglomeration and precipitation occur. Commonly used solvents are ethanol [64-17-5] and acetone [67-64-1]. Satisfactory results are obtained only if the concentration of solvent and the temperature are carefully controlled because enzymes can be inactivated easily by organic solvents.

Precipitation with Polymers. The polymers generally used are polyethylenimines and poly(ethylene glycols) of different molecular masses. The mechanism of this precipitation is similar to that of organic solvents and results from a change in the solvation effect of the water molecules surrounding the enzyme. Most enzymes precipitate at polymer concentrations ranging from 15 to 20 %.

Precipitation at the Isoelectric Point. Proteins are ampholytes and carry both acidic and basic groups. The solubility of proteins is markedly influenced by pH and is minimal at the isoelectric point at which the net charge is zero. Because most proteins have isoelectric points in the acidic range, this process is also called *acid precipitation.*

Precipitation is usually carried out on a small scale. Problems can arise in scaling-up this process [229]. The mixing time, the residence time in the reactor (which affects agglomerate formation and enzyme activity), and the shearing forces generated by stirring (which affect the aggregates formed) are critical parameters. When the volume being processed is large, the mixing time is appropriately long and, especially with organic solvents, protein denaturation can occur. Experiments have been conducted to overcome difficulties of this kind by using a continuous process [230].

3.2.3.3. Ultrafiltration

A semipermeable membrane permits the separation of solvent molecules from larger enzyme molecules because only the smaller molecules can penetrate the membrane when the osmotic pressure is exceeded. This is the principle of all membrane separation processes (Table 4), including ultrafiltration. In reverse osmosis, used to separate materials with low molecular mass, solubility and diffusion phenomena influence the process, whereas ultrafiltration and cross-flow filtration are based solely on the sieve effect. In processing enzymes, cross-flow filtration is used to harvest cells, whereas ultrafiltration is employed for concentrating and desalting.

Table 4. Membrane separation processes

Process	Application	Separation range, M_r
Cross-flow microfiltration	concentration of bacteria, removal of cell debris	>1 000 000 (or particles)
Ultrafiltration	concentration of enzymes, dialysis, fractionation	>10 000 (macromolecules)
Reverse osmosis	concentration of small molecules, desalting	>200

Difficulties arise from *concentration polarization.* The semipermeable membrane excludes larger molecules, which tend to accumulate near the surface of the membrane because back-diffusion into the solution is limited. As a result of the different rates of diffusion of molecules of different sizes, the separating ability of the membrane changes. Thus, the membrane holds back small molecules more strongly than would be expected from its pore size. This effect limits the applicability of membrane separation. The formation of *gel layers* on the membrane is reduced by maintaining a turbulent flow or a laminar flow with high flow rate. Loss of permeability is also caused by *membrane fouling* or *deposition* on the membrane. In particular, antifoaming agents from fermentation solutions are deposited on the membrane and make concentration of enzymes more difficult.

Membranes are available for ultrafiltration which exclude molecules ranging from 1000 to 300 000 dalton [231], [232]. Anisotropic membranes consisting of a very thin membrane layer (0.1 – 0.5 µm) and a thicker, porous support layer [233] are generally used. The different membranes employed are flat membranes (plate and frame, cassette type, or spiral winding module), hollow fibers, and tubular modules (Fig. 20). Cellulose acetate [9004-35-7] and organic polymers such as polysulfone [25135-51-7], poly(vinylidene fluoride) [24937-79-9], and polypropylene [9003-07-0] have proved to be useful as membrane materials. With the exception of cellulose acetate, these membranes can be cleaned easily with alkali or acid and steam sterilized.

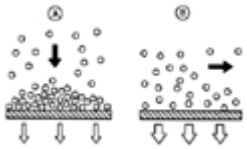

Figure 20. Principles of conventional, dead end filtration (A) and cross-flow filtration (B) [Full View]

Desalting. The desalting of enzyme solutions can be carried out conveniently by diafiltration. The small salt molecules are driven through a membrane with the water molecules. The permeate is continuously replaced by fresh water. In fact, the concentration of salt decreases according to the following formula:

$$\ln \frac{c_0}{c} = \frac{V}{V_0}$$

where

c_0 = starting salt concentration

c = final salt concentration

V_0 = starting volume

V = volume exchanged

Hence, a 99 % salt exchange can be achieved when the permeated volume is $5 \cdot V_0$, independent of the starting concentration.

3.2.4. Purification

For many industrial applications, partially purified enzyme preparations will suffice; however, enzymes for analytical purposes and for medical use must be highly purified. Special procedures employed for enzyme purification are crystallization, electrophoresis, and chromatography.

3.2.4.1. Crystallization

The rapid growth in the utilization of enzymes in commercial sectors such as agriculture and consumer products requires a cost-effective, industrial-scale purification method. Crystallization, one of the oldest chemical purification technologies, has the potential to fulfill these requirements. Enzyme crystallization is the formation of solid enzyme particles of defined shape and size. An enzyme can be induced to crystallize or form protein-protein interactions by creating solvent conditions that result in enzyme supersaturation. The theory and history of protein crystallization are well documented [234-237]. Despite its history extending more than 160 years, much of the emphasis in enzyme crystallization has focused on obtaining crystals for X-ray diffraction analysis rather than as a purification process.

However, many reports of enzyme crystallization from bulk fermentation for the purpose of purification have been published [238-240]. Enzymes that have been crystallized for commercial production include a cellulase [241], [242], glucose isomerase [243] subtilisin [243], [244], and alcohol oxidase [245].

Figure 21 shows the steps involved in the crystallization process for an industrial enzyme [244]. Typical crystallization process volumes can be as large as several thousands of liters of impure enzyme solution as the starting material. The desired characteristics of industrial scale enzyme crystallization are product purity, process yield, ease of crystal recovery, and short overall process time. To achieve these aims, the crystallization process must be carefully designed and developed to produce crystals with relatively large size and desired morphology.

Figure 21. Industrial-scale halide salt crystallization of subtisilin [244]
[Full View]

Many factors, including salt type and concentration, pH, temperature, the presence of variable amounts and types of impurities, mixing, and crystal seeding, can affect enzyme crystallization. Controlling the level of supersaturation throughout the crystallization process is essential for optimization of crystal size, which can be controlled by use of precipitants such as salt, pH, and temperature [235], [246]. Temperature plays a key role in the rate of enzyme crystallization. Cellulase [241], [244] and subtilisin [247], [248] were both reported to crystallize at a much higher rate with increasing temperature.

Some enzymes can be made to grow in a variety of different morphologies by simply changing the crystallization conditions during the course of crystallization. Subtilisin was shown to form rectangular plate-shaped crystals by first nucleating at low temperature followed by growing the crystals at elevated temperature. The rectangular plate-shaped crystal is a hybrid morphology between its high- and low-temperature morphologies, rods and square plates, respectively [248].

3.2.4.2. Electrophoresis

Electrophoresis is used to isolate pure enzymes on a laboratory scale. Depending on the conditions, the following procedures can be used: *zone electrophoresis*, *isotachophoresis*, or *porosity gradients*. The heat generated in electrophoresis and the interference caused by convection are problems associated with a scale-up of this method. An interesting contribution to the industrial application of electrophoresis is a continuous process in which the electrical field is stabilized by rotation [249], [250].

3.2.4.3. Chromatography

Chromatography is of fundamental importance to enzyme purification (Table 5). Molecules are separated according to their physical properties (size, shape, charge, hydrophobic interactions), chemical properties (covalent binding), or biological properties (biospecific affinity).

Table 5. Chromatographic methods

Type of chromatography	Principle	Separation according to
Adsorption	surface binding	surface affinity
Distribution	distribution equilibrium	polarity
Ion exchange	ion binding	charge
Gel filtration	pore diffusion	molecular size, molecular shape
Affinity	specific adsorption	molecular structure
Hydrophobic	hydrophobic chelation	molecular structure
Covalent	covalent binding	polarity
Metal chelate	complex formation	molecular structure

In *gel chromatography* (also called *gel filtration)*, hydrophilic, cross-linked gels with pores of finite size are used in columns to separate biomolecules. Concentrated solutions are necessary for separation because the sample volume that can be applied to a column is limited to ca. 10 % of the column volume. In gel filtration, molecules are separated according to size and shape. Molecules larger than the largest pores in the gel beads, i.e., above the exclusion limit, cannot enter the gel and are eluted first. Smaller molecules, which enter the gel beads to varying extent depending on their size and shape, are retarded in their passage through the column and eluted in order of decreasing molecular mass. The eluation volume of a globular protein is proportional to the logarithm of its molecular mass [251]. By varying the degree of cross-linking, gels of different porosities and with different fractionation ranges are obtained. Media that collectively cover all molecular sizes are available (Table 6). These include different types of Sephadex, which can be obtained by cross-linking dextran [9004-54-0] with epichlorohydrin [106-89-8], and Sephacryl, which is prepared by cross-linking allyldextran with *N,N´*-methylene-bisacrylamide [110-26-9]. Gel filtration is used commercially for both separation and desalting of enzyme solutions.

Table 6. Gel filtration media

Trade name (manufacturer)	Matrix	Fractionation range, M_r
Biogel (Bio-Rad)	polyacylamide (P-type)	100 – 400 000
	agarose (A-type)	$1000 – 150 \times 10^6$
Ultrogele (LKB)	agarose / polyacrylamide	$60\,000 – 1.3 \times 10^6$
	agarose	$25\,000 – 20 \times 10^6$
Fractogel (Merck)	vinyl polymer (various types)	$100 – 5 \times 10^6$
Sephadex (Pharmacia)	dextran (various types)	50 – 600 000
Sepharose	agarose	$10\,000 – 40 \times 10^6$
(Pharmacia)	cross-linked agarose	$10\,000 – 40 \times 10^6$
Sephacryl (Pharmacia)	sephacryl / bisacrylamide	$5000 – 1 \times 10^6$
Glycophase	surface-modified glass	1000 – 350 000

(Pierce)	(1,2-dihydroxypropyl-substituted)

Ion-exchange chromatography is a separation technique based on the charge of protein molecules. Enzyme molecules possess positive and negative charges. The net charge is influenced by pH, and this property is used to separate proteins by chromatography on anion exchangers (positively charged) or cation exchangers (negatively charged) (Table 7). The sample is applied in aqueous solution at low ionic strength, and elution is best carried out with a salt gradient of increasing concentration. Because of the concentrating effect, samples can be applied in dilute form.

Table 7. Ion-exchange resins

Trade name (manufacturer)	Matrix	Anion-exchange groups**	Cation-exchange groups**
Cellex (Bio-Rad)	cellulose	DEAE ECTEOLA QAE TEAE	CM phosphoryl
Bio-Gel A (Bio-Rad)	agarose	DEAE	CM
Trisacryl (LKB)	synthetic polymer	DEAE	CM
Fractogel (Merck)	vinyl polymer	DEAE	CM SP
Sephadex (Pharmacia)	dextran	DEAE QAE	CM SP
Sepharose (Pharmacia)	agarose	DEAE	CM
Sepharose FF *			
Q-Sepharose			
S-Sepharose			
Sephacel (Pharmacia)	cellulose	DEAE	

* FF = fast flow.

** CM = carboxymethyl: $–O–CH_2–COO^- Na^+$; DEAE = diethylaminoethyl: $–O–CH_2–CH_2–N^+(C_2H_5)_2H\ OH^-$; ECTEOLA = mixture of different amines; QAE = 2-hydroxypropylamine: $–O–CH_2–CH_2–N^+(C_2H_5)_2–CH(OH)–CH_3\ Br^-$; SP = sulfopropyl: $-O-CH_2-CH_2-CH_2-SO_3^-Na^+$; TEAE = triethylaminoethyl: $–O–CH_2–CH_2–N^+(C_2H_5)_3\ Br^-$.

The ability to process large volumes and the elution of dilute sample components in concentrated form make ion exchange very useful. The matrix used to produce ion-exchange resins should (1) be sufficiently hydrophilic to prevent enzyme denaturation and (2) have a high capacity for large molecules at fast equilibration. In addition, industrial applications require ion exchangers that give good resolution, allow high flow rates to be used, suffer small changes in volume with salt gradients or pH changes, and can be completely regenerated.

For *hydrophobic chromatography*, media derived from the reaction of CNBr-activated Sepharose with aminoalkanes of varying chain length are suitable [252]. This method is based on the interaction of hydrophobic areas of protein molecules with hydrophobic groups on the matrix. Adsorption occurs at high salt concentrations, and fractionation of bound substances is achieved by eluting with a negative salt

gradient. This method is ideally suited for further purification of enzymes after concentration by precipitation with such salts as ammonium sulfate.

In *affinity chromatography*, the enzyme to be purified is specifically and reversibly adsorbed on an effector attached to an insoluble support matrix. Suitable effectors are substrate analogues, enzyme inhibitors, dyes, metal chelates, or antibodies. The principle of affinity chromatography is shown in Figure 24. The insoluble matrix (C) is contained in a column. The biospecific effector, e.g., an enzyme inhibitor (I), is attached to the matrix. A mixture of different enzymes (E_1, E_2, E_3, E_4) is applied to the column. The immobilized effector specifically binds the complementary enzyme. Unbound substances are washed out, and the enzyme of interest (E_1) is recovered by changing the experimental conditions, for example by altering pH or ionic strength.

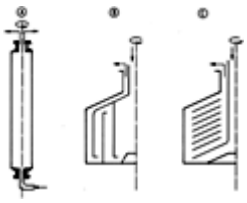

Figure 24. Solid-wall centrifuges

A): Tubular bowl; B): Multichamber solid bowl; C): Disk stack
[Full View]

Metal ions can also serve as effectors. They are attached to the matrix with the help of complexing agents, e.g., iminodiacetic acid [142-73-4]. Here, enzymes are separated on the basis of differing strengths of interaction with metal ions [253].

Immunoaffinity chromatography occupies a unique place in purification technology. In this procedure, monoclonal antibodies are used as effectors. Hence, the isolation of a specific substance from a complex biological mixture in one step is possible. In this procedure, enzymes can be purified by immobilizing antibodies specific for the desired enzyme. A more general method offers the synthesis of a fusion protein with protein A by “protein engineering.” Protein A is a *Staphylococcus* protein with a high affinity for many immunoglobulins, especially of the IgG class of antibodies. In this way, enzymes that usually do not bind to an antibody can be purified by immunoaffinity chromatography.

Covalent chromatography differs from other types of chromatography in that a covalent bond is formed between the required protein and the stationary phases. The principle is illustrated in Figure 25 for an enzyme containing a reactive mercapto group, such as urease (E.C. 3.5.1.5) [9002-13-5] [254].

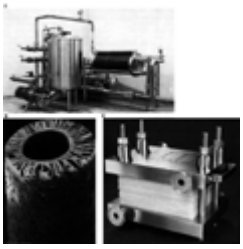

Figure 25. Different ultrafiltration modules

A): Plate-and-frame module (by courtesy of DDS, Ro-Division); B): Hollow-fiber module (by courtesy of Amicon); C): Cassette module (by courtesy of Millipore)
[Full View]

Industrial-Scale Chromatography. The main applications of industrial-scale chromatography in the 1960 s were the desalination of enzyme solutions by use of highly cross-linked gels such as Sephadex G-25 and batch separations by means of ion exchangers such as DEAE-Sephadex A-50. The stability and hydraulic properties of chromatographic media have been improved so that these techniques are now used on a production scale. Important parameters for the scale-up of chromatographic systems are the height of the column, the linear flow rate, and the ratio of sample volume to bed volume [255]. Zone spreading interferes with the performance of the column. Factors that contribute to zone spreading are longitudinal diffusion in the column, insufficient equilibration, and inadequate column packing. Longitudinal diffusion can be minimized by using a high flow rate. On the other hand, equilibration between the stationary and the mobile phases is optimal at low flow rates. Because good process economy depends to a large extent on the flow

rate, a compromise must be made. In addition, the flow rate is also dependent on particle size; the decisive factor is usually the pressure drop along large columns. Although optimal resolution is obtained only with the smallest particles, the gel must have a particle size that favors a good throughput and reduces processing times. The use of *segmented columns* prevents a large pressure drop in the column (Fig. 26) [256]. Above all, the column must be uniformly packed so that the particle-size distribution, porosity, and resistance to flow are the same throughout the column. If this is not done, viscous protein solutions can give an uneven elution profile, which would lead to zone bleeding. The design of the column head is important for uniform distribution of the applied sample. This is generally achieved by symmetrical arrangement of several inlets and perforated inserts for good liquid distribution. The outlet of the column must have minimal volume to prevent back-mixing of the separated components.

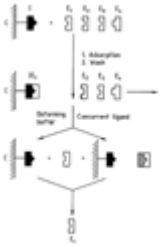

Figure 26. Affinity chromatography

For explanation of abbreviations, see text
[Full View]

Columns of glass, plastic, or in the case of large columns, metal (stainless steel) can be used. Metal columns are usually highly polished because a smooth surface is more resistant to corrosion. Viton or poly (tetrafluoroethylene) (PTFE) is applied as sealing material. Polyethylene tubing can be used to connect small columns. Larger tubing, with a diameter of more than 4 mm, must be made of metal, with special screw connections. No dead angles should be present in which liquid can collect and cause contamination by microbial growth.

Figure 27 shows a typical industrial chromatographic column. To avoid contamination of the column, sterile filters are placed at the entrance and exit. Optimal and efficient production systems not only must meet official standards such as those of the Food and Drug Administration (FDA) or Bundesgesundheitsamt (BGA), but must also comply with rules of good manufacturing practice (GMP), which makes cleaning-in-place (CIP) necessary. The system must be operated so that impurities from the process or from microbial growth can be removed completely. Chromatographic columns can be cleaned only by washing with chemicals. The reagents used for this purpose are described in Table 8. Figures 28 and 29 show the steps involved in processing extracellular and intracellular enzymes, respectively.

Table 8. Reagents for cleaning-in-place (CIP) of chromatographic media

Substance	Purification power (protein and fat)	Sterilization and removal of pyrogens
Sodium hydroxide	++	+
Acids	–	+
Sodium phosphate	+	–
Surfactants	++	–

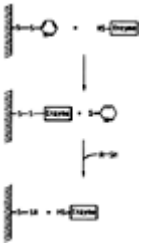

Figure 27. Covalent chromatography
[Full View]

Figure 28. Chromatographic KS 370 stack columns (by courtesy of Pharmacia)
[Full View]

Figure 29. Production-scale chromatography column (Pharmacia)
[Full View]

3.2.5. Product Formulation

Enzymes are sold as stabilized liquid concentrates or as particulate solids. The primary task of formulation is to minimize losses in enzymatic activity during transport, storage, and use. Enzymes are often exposed to humid, hot, or oxidative environments in industrial applications such as detergents, textile formulations, and food and beverage processing. Much can be done at the screening or research stage to identify enzymes that are structurally more stable or resistant to oxidation; chemical stabilizers can do little to protect thermally or chemically labile enzymes [257].

Formulations enhance stability by counteracting the primary forces of deactivation: denaturation, catalytic-site deactivation, and proteololysis [258]. Denaturation occurs by physical unfolding of an enzyme's tertiary protein structure under thermal or chemical stress. Once an enzyme begins to unfold it becomes dramatically more vulnerable to deactivation and proteolysis. To minimize unfolding, the formulator can alter the protein's environment so as to induce a compact protein structure; this is done most effectively by “preferential exclusion” of water from the protein surface by adding water-associating compounds such as sugars, polyhydric alcohols, and lyotropic salts [259]. The best ways to combat active site inactivation are to ensure sufficient levels of any required cofactors, to add reversible inhibitors, and to exclude oxidizing or reactive species from the formulation.

Besides enzymatic stability, a formulation should meet several key secondary requirements, including preservation against microbial contamination, avoidance of physical precipitation or haze formation, minimizing the formation of sensitizing dusts or aerosols, and the optimization of esthetic criteria such as color and odor. Many of these problems are best addressed by focusing as far “upstream” as possible, including the choice of raw materials in the fermentation or enzyme recovery process. Downstream operations such as diafiltration, adsorption, chromatography, crystallization, and extraction can be used to remove impurities responsible for color, odor, and precipitation [260]. The risk of physical precipitation is minimized by formulating near the isoelectric point of the enzyme with hydrophilic solvents such as glycerol or propylene glycol. One can effectively also add moderate levels of of solvating salts to avoid either salting-out or “reverse salting-in”. To prevent microbial contamination, one can use a combination of filtration, acidification, and the minimization of free water; biocides can be effective, but the range of

acceptable chemicals for controlling or killing microbes is increasingly circumscribed by health and safety regulations.

Worker safety has been a primary driver in the development of dry granular enzyme formulations for powdered laundry detergents and textile formulations. Since the 1970s, enzyme granules have become increasingly resistant to physical breakage and formation of airborne dusts upon handling. Two processes producing the most attrition-resistant granules to date are high-shear granulation and fluidized-bed spray coating. These processes use various binders, coatings, and particle morphologies to produce nonfriable particles which still protect enzymes during storage but allow for their ready release in solution during application.

3.2.6. Waste Disposal

Because of the generally low concentration of enzyme in the starting material, the volume of material that must be processed is large, and substantial amounts of waste accumulate. The spent fermentation medium can still contain large amounts of unused nutrients. However, recycling is generally not possible because of the presence of metabolites in the medium. Solid organ remains and mycelium, which are used as animal feed, can be separated. The latter must be carefully checked for undesired metabolites, e.g., antibiotics, before being fed to animals.

In recombinant DNA techniques, the need to maintain absolute containment is of great concern. Waste must be chemically or thermally inactivated before disposal, to ensure that no live organisms escape into the environment.

3.3. Immobilization

Synthesis, decomposition, and partial conversion of different compounds in biological systems are catalyzed by enzymes, water-soluble globular proteins. The catalytic activity of enzymes is brought about by their tertiary and quaternary (oligomeric enzymes) structures. All enzymes have a catalytic center, one or several substrate binding sites, and one or several regulatory sites.

The characteristics of enzymes as catalysts are (1) enzymes are able to catalyze reactions at ambient temperature and pressure and in a pH range around neutral, and (2) enzymes have strict substrate specificity, stereospecificity, regiospecificity, and reaction specificity. These facts suggest that energy-saving, resource-saving, and low-pollution processes could be designed by using enzymes, because such processes would operate at relatively low temperature and atmospheric pressure with little byproduct formation. However, the molecular structure of enzymes that is essential for their catalytic activity is liable to be destroyed under conditions such as high temperature, high or low pH, presence of organic solvents, or even conditions suitable for catalysis. The recovery of active enzymes from spent reaction mixtures is another problem when free (nonimmobilized) enzymes are used. Immobilization is one way of eliminating some of the disadvantages inherent to enzymes.

Therefore, the immobilization of biocatalysts – not only enzymes but also cellular organelles, microbial cells, plant cells, and animal cells – is attracting worldwide attention in the practical application of bioprocesses. In general, immobilized biocatalysts are stable and easy to handle compared to their free counterparts. One of their most important features is that they can be used repeatedly in a long-term series of batchwise reactions or continuously in flow systems. At present, applications of immobilized biocatalysts include (1) the production of useful compounds by stereospecific or regiospecific bioconversion, (2) the production of energy by biological processes, (3) the selective treatment of specific pollutants to solve environmental problems, (4) continuous analyses of various compounds with a high sensitivity and a high specificity, and (5) medical uses such as new types of drugs for enzyme therapy or artificial organs.

These processes require the immobilization not only of single enzymes but also of several different enzymes, organelles, or cells that catalyze more complex reactions. Various methods have been developed for the immobilization of biocatalysts, and they are being used extensively today.

3.3.1. Definitions

Immobilized enzymes are defined as “enzymes physically confined or localized in a certain defined region of space with retention of their catalytic activities, which can be used repeatedly and continuously” [261]. This definition is applicable to enzymes, cellular organelles, microbial cells, plant cells, and animal cells,

that is, to all types of biocatalysts. In some cases, these biocatalysts are bound to or within insoluble supporting materials (*carriers*) by chemical or physical binding. In other cases, biocatalysts are free, but confined to limited domains or spaces of supporting materials (*entrapment*).

The European Federation of Biotechnology, Working Party on Applied Biocatalysis (formerly, the Working Party on Immobilized Biocatalysts) has proposed that immobilized cells should be classified as *viable* and *nonviable*, where viable includes *growing* and *nongrowing* or *respiring.* According to the same definition, the terms *dead, resting*, and *living* should not be used [262]. However, in this section, the terms *treated, resting*, and *growing* — and in some cases, *living* — are used for classifying the condition of immobilized cells. Treated cells are subjected to chemical or physical treatment before or after immobilization, and resting cells do not show growth during utilization.

3.3.2. History

In 1916, NELSON and GRIFFIN observed that yeast invertase (E.C. 3.2.1.26) [9001-57-4] adsorbed on charcoal was able to catalyze the hydrolysis of sucrose [263]. After that, several reports were published on the immobilization of physiologically active proteins by covalent binding on several supports. However, immobilized enzymes were not used in practice until 1953, when GRUBHOFER and SCHLEITH immobilized several enzymes, such as carboxypeptidase, diastase, pepsin, and ribonuclease, on diazotized polyaminostyrene resin by covalent binding [264]. Thereafter, MITZ reported the ionic binding of catalase (E.C. 1.11.1.6) [9001-05-2] on DEAE-cullulose in 1956 [265]. BERNFELD and WAN described the entrapment of trypsin (E.C. 3.4.21.4) [9002-07-7], papain (E.C. 3.4.22.2) [9001-73-4], amylase, and ribonuclease in polyacrylamide gel in 1963 [266], and QUIOCHO and RICHARDS demonstrated the cross-linking of carboxypeptidase A (E.C. 3.4.17.1) [11075-17-5] with glutaraldehyde in 1964 [267]. Microencapsulation of carbonic anhydrase (E.C. 4.2.1.1) [9001-03-0] was reported by CHANG in 1964 [268] and the preparation of liposomes containing amyloglucosidase (E.C. 3.2.1.3) [9032-08-0] by GREGORIADIS in 1971 [269]; both were used in enzyme therapy. During this pioneering period, KATCHALSKI-KATZIR and co-workers made extensive contributions to the theoretical understanding of immobilized enzymes [270].

In 1969, CHIBATA and co-workers of Tanabe Seiyaku Co., Japan, were successful for the first time in the industrial application of immobilized enzymes. Fungal aminoacylase (E.C. 3.5.1.14) [9012-37-7] was immobilized on DEAE-Sephadex through ionic binding and used for the stereoselective hydrolysis of *N*-acyl-D,L-amino acids to yield L-amino acids and *N*-acyl-D-amino acids [271]. The first industrial application of immobilized microbial cells was also performed by CHIBATA and co-workers in 1973 to produce L-aspartate from ammonium fumarate by polyacrylamide gel-entrapped *Escherichia coli* cells containing a high activity of aspartase (E.C. 4.3.1.1) [9027-30-9]. Other processes that have been industrialized use immobilized enzymes and microbial cells. They are discussed briefly in Section Application.

3.3.3. Methods

For the application of immobilized biocatalysts, their screening to the desired activity and characteristics is most important. In addition, selection of the appropriate combination of supporting material and immobilization method, both of which should be suitable for each biocatalyst, is necessary. No systematic concept is available at present for design of the most appropriate method of immobilization for various biocatalysts. Optimization is carried out in general by trial and error.

At present, various immobilization techniques are available [272], and Figure 30 illustrates the principles of these methods. Because each method has its own merits and demerits, selection should be based on the intended purpose, including type of biocatalyst, type of reaction, and type of reactor.

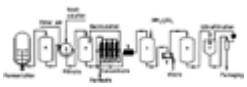

Figure 30. Isolation of extracellular enzymes

a) Hold tank; b) Filtration; c) Precipitation vessel; d) Centrifuge; e) Suspension vessel; f) Spray drier
[Full View]

Supporting materials should have adequate functional groups to immobilize the biocatalysts, as well as sufficient mechanical strength; physical, chemical, and biological stability; and nontoxicity. Easy shaping is

also important for applying immobilized biocatalysts to different types of reactors. Furthermore, the economic feasibility should be examined. Some supporting materials are described in the following sections.

3.3.3.1. Carrier Binding

Enzyme proteins have amino acid residues containing chemically reactive groups, ionic groups, and/or hydrophobic groups, as well as hydrophobic domains. These amino acid residues and the hydrophobic domains can participate in the immobilization of enzymes through covalent linkage, ionic binding, or physical adsorption. Various types of insoluble supports (carriers) are utilized as is or after proper modification or activation.

Covalent Binding. Amino acid residues that are not involved in the active site or substrate-binding site of the enzymes to be immobilized can be used for covalent binding with supports. These are the ε-amino group of lysine, the mercapto group of cysteine, the β-carboxyl group of aspartic acid, the γ-carboxyl group of glutamic acid, the phenolic hydroxyl group of tyrosine, or the hydroxyl groups of serine and threonine. Hydroxyl, carboxyl, and amino groups are especially excellent targets because of their relative abundance in enzyme molecules.

Enzymes immobilized by covalent binding have the following advantages: (1) because of the tight binding, they do not leak or detach from supports during utilization; (2) immobilized enzymes can easily come into contact with substrates because the enzymes are localized on the surface of supports; and (3) an increase in heat stability is often observed because of the strong interaction between enzyme molecules and supports.

On the other hand, disadvantages of covalent binding are: (1) active structures of enzyme molecules are liable to be destroyed by partial modification; (2) strong interaction between enzyme molecules and supports often hinders the free movement of enzyme molecules, resulting in decreased enzyme activity; (3) optimal conditions of immobilization are difficult to find; (4) this method is not suitable for immobilization of cells; and (5) supports, in general, are not renewable. Hence, this principle is well-suited to expensive enzymes whose stability is significantly improved by covalent binding.

Despite these disadvantages, covalent binding is often applied to the preparation of immobilized enzymes for analytical purposes. Some typical examples of covalent binding methods are described below, but many other techniques have also been reported.

Cyanogen Bromide. The cyanogen bromide method was first demonstrated by AXÉN and his co-workers [273]. It involves the activation of supports having vicinal hydroxyl groups (**1**), such as polysaccharides, glass beads, or ceramics, with cyanogen bromide [506-68-3] (**2**), to yield reactive imidocarbonate derivatives (**3**). The subsequent reaction between the activated supports and enzyme molecules gives *N*-substituted isourea (**4**), *N*-substituted imidocarbonate (**5**), and *N*-substituted carbamate derivatives (**6**).

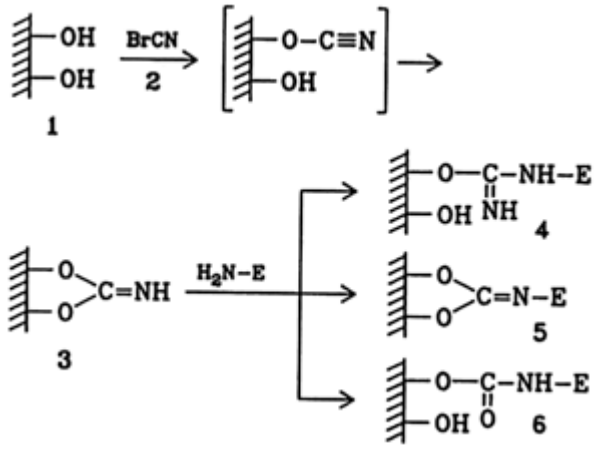

This method has been widely used for the immobilization of various enzymes, and CNBr-activated supports, such as CNBr-activated Sepharose, are available. Aminated glass beads and aminated polyacrylamide gel are also used as supports in this method. The insertion of spacers such as hexamethylenediamine [124-09-4] is also possible, to avoid strong interaction between enzyme molecules and supports.

Acid Azide Derivatives. The acid azide method, which is used for peptide synthesis, is also applicable to the immobilization of various enzymes [274]. For example, carboxymethyl cellulose [9004-42-6] (**7**) is converted to the methyl ester (**8**) and then to the hydrazide (**9**) with hydrazine. The hydrazide reacts with nitrous acid to form the azide derivative (**10**), which can then react with enzyme molecules at low temperature to yield immobilized enzymes (**11**). Several supports containing carboxyl groups can be used for this method.

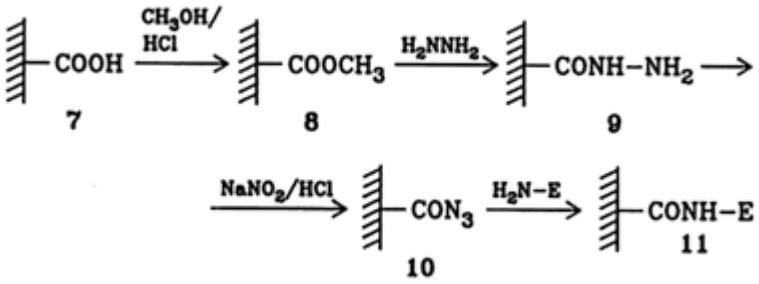

Condensing Reagents. Carboxyl groups or amino groups of supports and amino groups or carboxyl groups of enzyme molecules can be condensed directly through the formation of peptide linkages by the action of carbodiimide reagents or Woodward's reagent K (*N*-ethyl-5-phenylisoxazolium-3′-sulfonate) [4156-16-5]. Carboxymethyl cellulose is one of the supports used for this method [275].

Diazo Coupling. Supports having aromatic amino groups (**12**) are diazotized with nitrous acid to form the diazonium derivatives (**13**), which react with enzyme molecules to yield immobilized enzymes (**14**). 4-Aminobenzyl cellulose [9032-51-3] [276] and polyaminostyrene [9060-90-6] [277] are typical supports used for this method.

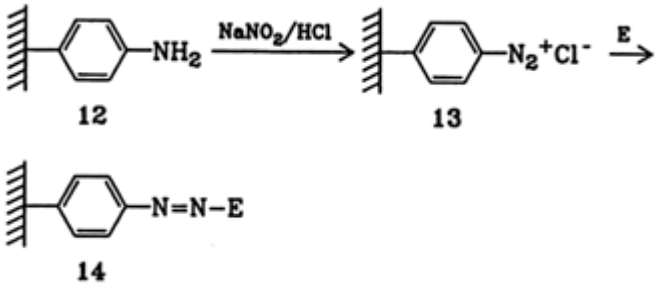

Alkylation. Alkylation groups on supports can easily react with amino groups, phenolic hydroxyl groups, and sulfhydryl groups of the enzyme molecules. For example, halogenated acetyl cellulose (**15**) [278] and triazinyl derivatives of ion-exchange resins (**16**) or cellulose [279] may be used for this method.

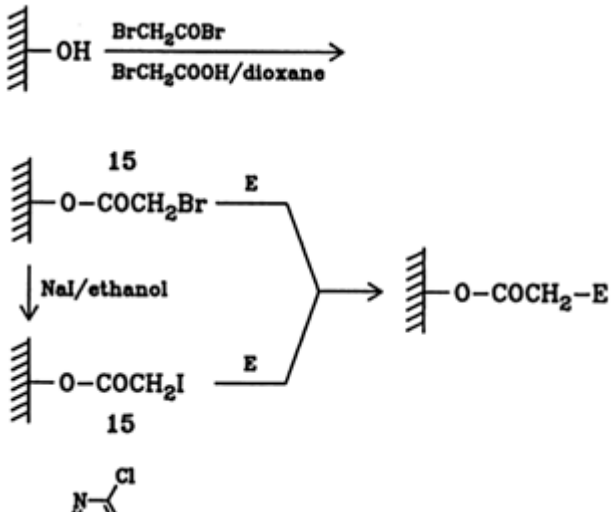

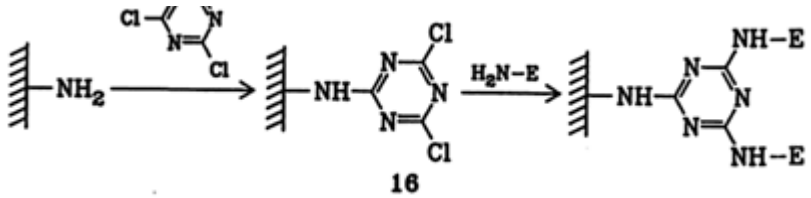

Carrier Cross-linking. Supports and enzyme molecules can be cross-linked with bi- or multifunctional reagents. For example, aminoethyl cellulose [9032-36-4] (**17**) and amino groups of enzyme molecules are combined through Schiff's base linkage (**19**) with glutaraldehyde [111-30-8] (**18**) [280]. Diisocyanates represent another group of cross-linking reagents. Several supports containing amino groups are also employed.

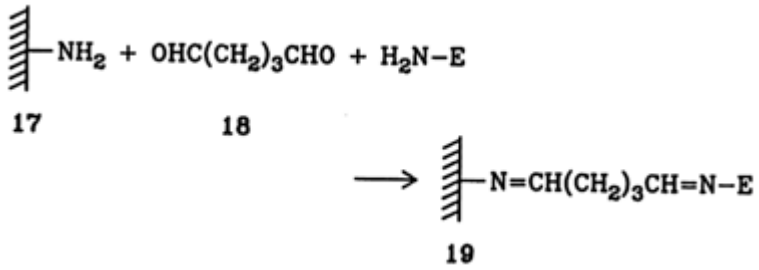

Ionic Binding. Since catalase was found to be bound to ion-exchange cellulose [265], this method has been applied for the immobilization of many enzymes, because the procedure is very simple, the supports are renewable, and the enzymes are not modified. The most notable example is the production of L-amino acids by aminoacylase (E.C. 3.5.1.14) [9012-37-7] immobilized on DEAE-Sephadex [271]. Binding of enzymes on supports is affected by the kind of buffers used, pH, ionic strength, and temperature. Several derivatives of cellulose and Sephadex, as well as various ion-exchange resins, can be utilized for immobilization.

Physical Adsorption. Biocatalysts often bind to carriers by physical interaction such as hydrogen bonding, hydrophobic interaction, van der Waal's forces, or their combined action. Although biocatalysts are immobilized without any modification, interaction between biocatalyst and support is generally weak and affected by such environmental conditions as temperature or concentration of reactants. Various inorganic supports are often used. Several synthetic resin beads and natural materials (e.g., chitosan beads with micropores of controlled size) having strong adsorption capacities are available. Adsorption followed by cross-linking with glutaraldehyde sometimes stabilizes the activity of immobilized enzymes. Phenoxyacetylated cellulose and glass beads are used as specific supports of a hydrophobic nature. Tannins, which interact strongly with proteins, are also applied as ligands after appropriate immobilization. Cellular organelles and various types of cells can be immobilized by physical adsorption. Supports are renewable under appropriate conditions.

3.3.3.2. Cross-linking

Bi- and multifunctional compounds serve as reagents for intermolecular cross-linking of enzymes. Cross-linked enzymes are then insoluble macromolecules. In addition to glutaraldehyde [111-30-8] (**20**) [267] which is the most popular cross-linking reagent, several compounds such as toluene diisocyanate [1321-38-6] or hexamethylene diisocyanate [822-06-0] (**21**) are used. The yield of enzyme activity is usually low. Microbial cells are also cross-linked with glutaraldehyde to yield cell pellets.

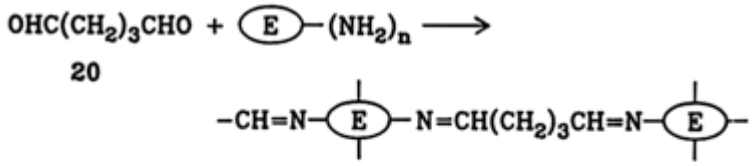

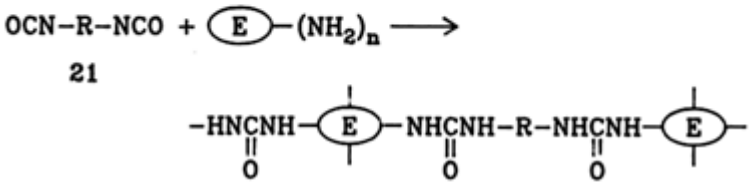

3.3.3.3. Entrapment

Entrapped biocatalysts are classified according to the following different types:

Lattice type: biocatalysts entrapped in gel matrices prepared from polysaccharides, proteins, or synthetic polymers.

Microcapsule type: biocatalysts entrapped in microcapsules of semipermeable synthetic polymers.

Liposome type: biocatalysts entrapped within liquid membranes prepared from phospholipids.

Hollow-fiber type: biocatalysts separated from the environment by hollow fibers.

Membrane type: biocatalysts separated from the spent reaction solution by ultrafiltration membranes.

The advantages of entrapping methods are that not only single enzymes but also several different enzymes, cellular organelles, and cells can be immobilized with essentially the same procedures. Biocatalysts are not subjected to serious modification, and immobilization eliminates the effect of proteases and enzyme inhibitors of high molecular mass. However, disadvantages are (1) substrates of high molecular mass can hardly gain access to the entrapped biocatalysts and (2) supports are not renewable. Entrapment within ultrafiltration membranes can avoid the disadvantages inherent in entrapping methods, although inactivated enzyme molecules often precipitate on the membrane surface, which results in decreased permeability to reaction solutions. The lattice-type method is most widely applied for preparing immobilized biocatalysts. Several examples of this technique are mentioned in the following paragraphs.

Polyacrylamide Gel. Since BERNFELD and WAN [266] reported the entrapment of several enzymes in polyacrylamide gels (**24**), different types of biocatalysts including cellular organelles and microbial cells have been immobilized by this method. This method was also applied to the industrial production of L-aspartate, L-malate, and acrylamide by immobilized microbial cells. In a typical procedure, acrylamide [79-06-1] (**22**) and *N,N′*-methylenebisacrylamide [110-26-9] (**23**) (cross-linking reagent) are mixed with biocatalysts and polymerized in the presence of an initiator (potassium persulfate [7727-21-1]) and a stimulator (3-dimethylaminopropionitrile [1738-25-6] (DMAPN) or *N,N,N′N′*-tetramethylenediamine [110-18-9]). Although this technique is used for various purposes, acrylamide monomer sometimes inactivates enzymes. Several analogues or derivatives of acrylamide can also be used in this method.

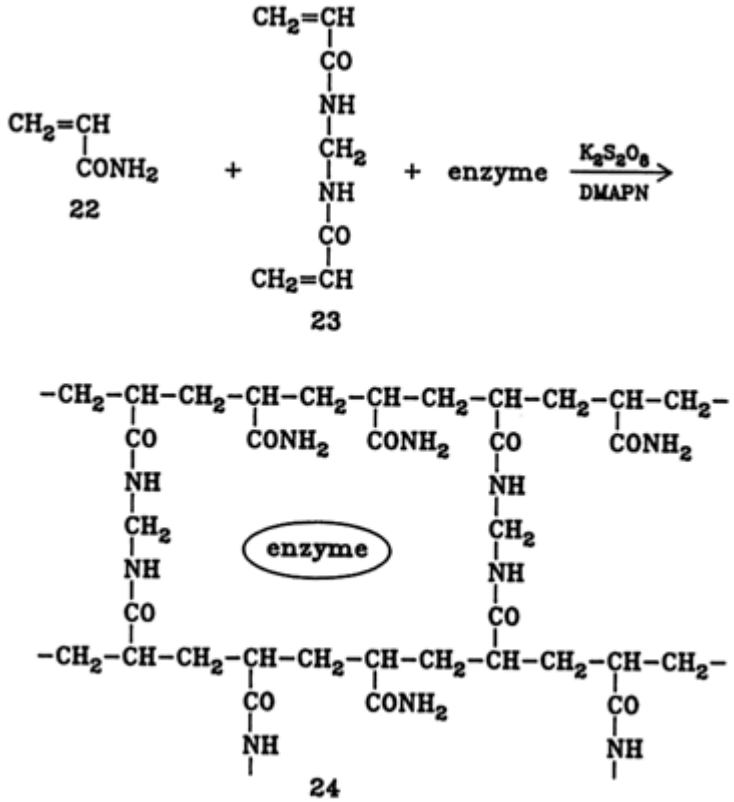

Alginate Gel. Several natural polysaccharides, such as alginate, agar, and κ-carrageenan, are excellent gel materials and used widely for the entrapment of various biocatalysts.

Sodium alginate I [9005-38-3], which is soluble in water, is mixed with a solution or suspension of the biocatalysts and then dropped into a calcium chloride solution to form water-insoluble calcium alginate gels that immobilize enzymes, cellular organelles, or microbial cells [281]. However, gels are gradually solubilized in the presence of calcium ion-trapping reagents such as phosphate ions. Aluminum ions or several divalent metal ions can be substituted for calcium.

This method is used widely for immobilization of various biocatalysts because of its simplicity and the availability of sodium alginate.

κ-Carrageenan Gel. CHIBATA and co-workers [282] have extensively screened gel materials for the immobilization of enzymes and microbial cells in industrial applications and have found κ-carrageenan (**25**) to be the best.

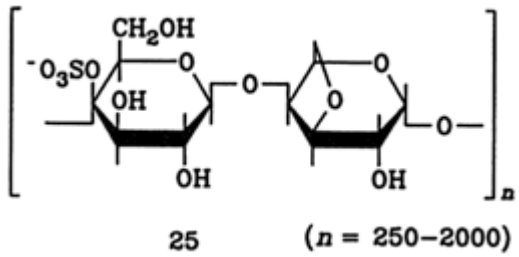

κ-Carrageenan in saline is mixed with a solution or suspension of biocatalysts and dropped into a solution of gelling reagent, such as potassium chloride [7447-40-7] [283]. Various cations, such as ammonium, calcium, and aluminum, also serve as good gelling reagents. Hardening of gels with glutaraldehyde and hexamethylenediamine often stabilizes the biocatalysts entrapped in κ-carrageenan gels.

This method has replaced polyacrylamide gels in the industrial production of L-aspartate [284] and L-malate

[285] by means of immobilized microbial cells.

κ-Carrageenan gels can be solubilized in saline or in water, which enables the number of cells in the gel to be counted.

Synthetic Resin Prepolymers. With the application of a variety of bioreactions, including synthesis, transformation, degradation, or assay of various compounds having different chemical properties, entrapment of biocatalysts in gels of controlled physicochemical properties has become desirable. Selection of suitable gels and modification of natural polymers to meet each purpose are usually difficult. For these reasons, a new entrapping method using synthetic resin prepolymers has been developed by FUKUI and co-workers [286].

Specific features and advantages of the prepolymer method are (1) entrapment procedures are very simple and proceed under very mild conditions; (2) prepolymers do not contain monomers that may have unfavorable effects on enzyme molecules; (3) the network structure of gels can easily be controlled by using prepolymers of optional chain length; and (4) the physicochemical properties of gels, such as the hydrophilicity – hydrophobicity balance and the ionic nature, can be changed by selecting suitable prepolymers synthesized in advance in the absence of biocatalysts.

Photo-cross-linkable resin prepolymers having hydrophilic (**26**) or hydrophobic (**27**) properties, cationic or anionic nature, and different chain lengths have been developed [286], [287].

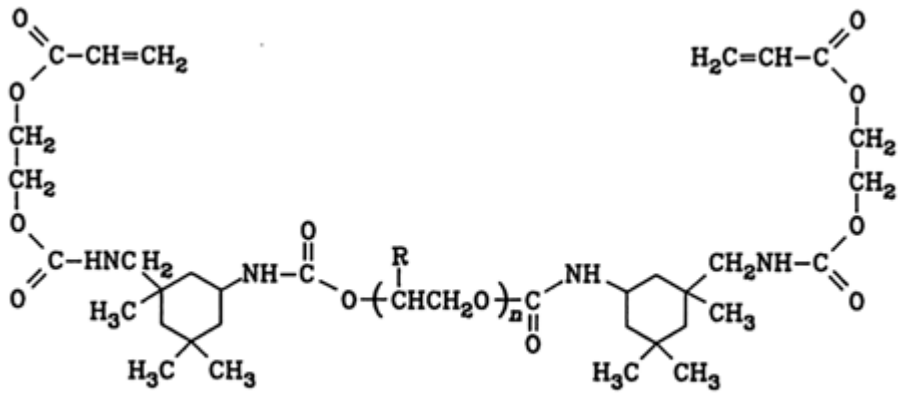

26: R = H
27: R = CH_3

Mixtures of prepolymers and biocatalysts are gelled by irradiation with long-wavelength UV light for several minutes in the presence of a proper sensitizer such as benzoin ethyl ether [574-09-4]. This method has been applied for the pilot-scale production of ethanol by immobilized growing yeast cells.

Entrapment by urethane prepolymers (**28**) with different hydrophilicity or hydrophobicity and chain length is much simpler [286], [288]. When the liquid prepolymers are mixed with an aqueous solution or suspension of the biocatalysts, the prepolymers react with each other to form urea bonds and liberate carbon dioxide.

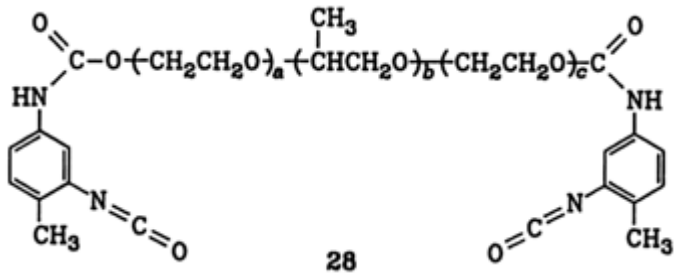

28

These prepolymers can immobilize not only enzymes but cellular organelles and microbial cells as well. The hydrophilicity – hydrophobicity balance of these gels has been demonstrated to affect especially the bioconversion of lipophilic compounds in organic solvent systems [289].

3.3.4. Characterization

The characteristics of immobilized biocatalysts should be described in the literature because different methods and preparations must be compared for their evaluation. However, determination of what kinds of description and what types of parameters should be given is difficult. Methods for estimating these parameters vary from group to group. Therefore, at present, the results described in the literature cannot be compared. The Working Party on Immobilized Biocatalysts within the European Federation of Biotechnology has made a proposal for the description of several properties of immobilized biocatalysts [290]. Although the proposal will be difficult to follow completely, it will be useful for the characterization of immobilized biocatalysts.

The minimum requirements recommended for characterization of an immobilized biocatalyst are as follows:

1. *General description*
   reaction scheme
   enzyme and microorganism
   carrier type
   method of immobilization
2. *Preparation of the immobilized biocatalyst*
   method of immobilization, reaction conditions, dry mass yield, activity left in supernatant
3. *Physicochemical characterization*
   biocatalyst shape, mean wet particle size, swelling behavior
   compression behavior in column systems, abrasion in stirred vessels, or minimum fluidization velocity and abrasion in fluidized beds
4. *Immobilized biocatalyst kinetics*
   Initial rates vs.
   - substrate concentration for free and immobilized biocatalyst, effect of pH and buffer
   - diffusional limitations in the immobilized biocatalyst system (effect of particle size or enzyme load on activity)
   - degree of conversion vs. residence time (points on a curve)
   - storage stability (residual initial rate after storage for different periods)
   - operational stability (residual initial rate or transforming capacity of reaction system) after operation for different periods

3.3.5. Application

Immobilized biocatalysts — enzymes, cellular organelles, microbial cells, plant cells, and animal cells — have been applied to the production or conversion of various compounds such as amino acids, peptides and enzymes, sugars, organic acids, antibiotics, steroids, nucleosides and nucleotides, lipids, terpenoids, fuels, or commodity chemicals. Some typical examples are described below.

A major industrial application of immobilized enzymes is the production of various L-amino acids, such as L-alanine, L-isoleucine, L-methionine, L-phenylalanine, L-tryptophan, and L-valine, by fungal aminoacylase (E.C. 3.5.1.14) [9012-37-7] immobilized on DEAE-Sephadex by ionic binding [271]. This enzyme hydrolyzes stereoselectively the L-isomer of *N*-acyl-D,L-amino acids (**29**) to yield L-amino acids (**30**) and unreacted *N*-acyl-D-amino acids (**31**); the latter are subsequently converted into the racemic form by heating. In this way, both the L- and D-isomers are converted completely to L-amino acids.

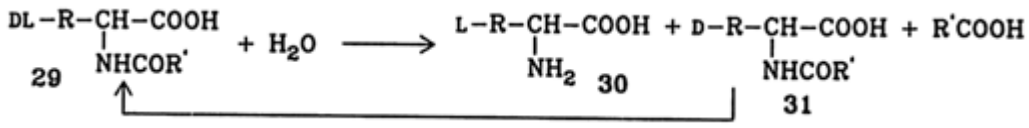

CHIBATA and co-workers also succeeded in producing L-aspartic acid [56-84-8] (**33**) from ammonium fumarate [14548-85-7] (**32**) by polyacrylamide gel-entrapped *Escherichia coli* cells containing a high activity of aspartase (E.C. 4.3.1.1) [9027-30-9].

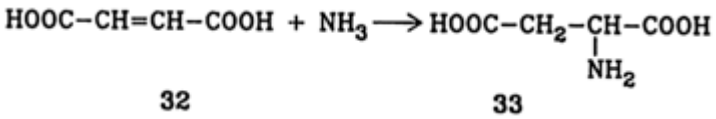

An active and stable preparation was obtained (half-life, 120 d) when the immobilized cells were incubated with a substrate solution for 48 h at 30 °C [291]. This process is the first example of an industrial application of immobilized cells and the enzyme partially isolated from *E. coli* cells is not as stable even after immobilization. At present, the process can be improved by entrapping the cells in κ-carrageenan gels, followed by hardening with glutaraldehyde and hexamethylenediamine; these immobilized cells have a half-life of 680 d [284].

L-Alanine [56-41-7] (**34**) is produced from L-aspartic acid [56-84-8] (**33**) by κ-carrageenan gel-entrapped cells of *Pseudomonas dacunhae* having a high activity of L-aspartate 4-decarboxylase (E.C. 4.1.1.12) [9024-57-1] [292]. This process was commercialized in 1982 by the Tanabe Seiyaku Co. in Japan.

Immobilized penicillin acylase (amidase) (E.C. 3.5.1.11) [9014-06-6] is now widely used in the production of 6-aminopenicillanic acid [551-16-6] (**36**) from penicillin G [61-33-6] (**35**). Glucose isomerase (E.C. 5.3.1.18) [9055-00-9] is used in the production of high-fructose syrup (**38**) from glucose [50-99-7] (**37**).

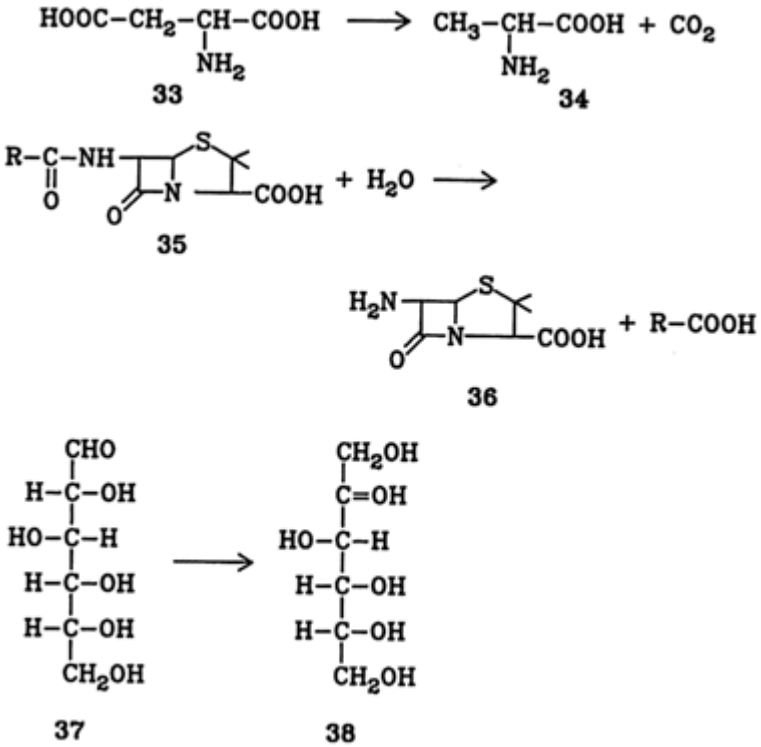

A process for production of L-malate (**40**) from fumarate (**39**) has been developed by CHIBATA and co-workers by using *Brevibacterium ammoniagenes* cells containing fumarase (E.C. 4.2.1.2) [9032-88-6] entrapped in polyacrylamide gels [293].

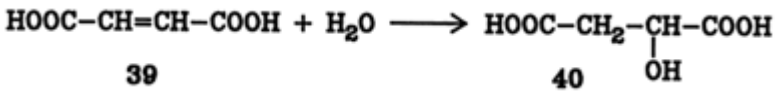

The entrapped cells were treated with bile extract to suppress the formation of byproduct succinate, which leads to the industrial application of this system in 1974. Later *B. flavum* was substituted for *B. ammoniagenes* and κ-carrageenan gels for polyacrylamide gels. κ-Carrageenan gel-entrapped cells showed a half-life of 160 d [285].

The latest use for immobilized biocatalysts is the production of cheap commodity chemicals such as acrylamide (**42**) from acrylonitrile (**41**).

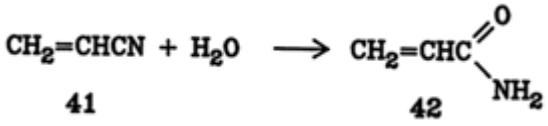

Polyacrylamide gel-entrapped cells of *Corynebacterium* species are used for the production of acrylamide at low temperature by Nitto Chemical Industry in Japan. YAMADA and his co-workers have found that in *Pseudomonas chlororaphis*, nitrile hydratase [82391-37-5] participates in this reaction [294].

In addition to the immobilized enzymes and treated cells mentioned above, other immobilized living or growing cells are now being investigated extensively for the production of more complex compounds by using the metabolic activities of these cells. Plant and animal cells as well as microbial cells are targets of immobilization because these cells have self-regenerating and self-proliferating catalytic systems [295]. Genetically improved cells are also useful biocatalysts.

Bioconversion of lipophilic compounds in organic solvent systems is another subject of interest industrially. Immobilization in or on proper supports stabilizes biocatalysts even in the presence of organic solvents [289].

Thus, the applications of immobilized biocatalysts are now expanding along with the development of immobilization techniques and the improvement of biocatalysts.

PREVIOUS NEXT

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Wolfgang Aehle, Genencor International B. V., Leiden, The Netherlands
Hans de Nobel, Genencor International B. V., Leiden, The Netherlands
Brian Jones, Genencor International B. V., Leiden, The Netherlands
Rick Bott, Genencor International, Inc., Palo Alto, USA

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

4. Discovery and Development of Enzymes

4.1. Enzyme Screening

4.1.1. Overview

Screening and selection are early processes in the research and development cycle of a product. In essence this means searching for a particular gene coding for the target enzyme. In the past this involved exclusively the screening of living microorganisms (classical microbial screening), but with the application of the modern tools of molecular biology, screening can be performed without the need to culture the organisms involved. As with all methods, there are limitations, and a combination of techniques (combinatorial screening) often provides the greatest success.

The most rewarding starting point for the process of search and discovery is the rich diversity of microorganisms in nature [296]. It is seldom appreciated that for 85 % of the Earth's history, life was restricted to microbial forms. The metabolic diversity this has created is truly immense. During more than four billion years of evolution, enzymes have emerged that are perfectly adapted for maintaining cellular processes. However, many enzymes are required to perform in a milieu that is far removed from the natural physiological conditions and may even involve exposure to chemicals known to degrade or denature enzymes, which leads to loss of activity. Usually, enzyme screening and selection strategies are based on knowledge of the application and the physical and chemical conditions under which the enzyme must operate. Therefore, selection for enzyme performance under application conditions is an essential early step in the screening process. Access to a large gene pool is another prerequisite for a successful screening program. The approaches can involve the indiscriminate examination of heterogeneous material (e.g., soil) gathered from the environment, or an ecological approach with a targeted examination of specific habitats where a particular activity is likely to be found. Strategies can be adopted to search specifically in poorly explored environments in order to access a greater variation in genomes from novel microbes. In the search for new genes, exploration of extreme environments such as volcanic hot springs, hypersaline lakes, and polar soils has yielded many novel microorganisms with peculiar properties [297-299]. Alternatively, a taxonomy-based approach involving a systematic study of known organisms can also be rewarding. By combining enzyme screening with modern techniques such as protein engineering and directed evolution, hybrid molecules with an improved performance under specific application conditions can be found. This chapter will describe the different screening methods currently used in industry to obtain enzymes with desired properties.

4.1.2. Natural Isolate Screening

The source material for enzyme discovery can be plant or animal matter and microbes, both prokaryotes (e.g., bacteria and archaea) and eukaryotes (e.g., yeasts and fungi). Microorganisms are the major source for industrial enzymes. The classical method of screening natural microbial isolates is a well-established process having its roots in antibiotic discovery (e.g., penicillins, streptomycin) in the decades after 1945. It involves the examination of thousands of samples of soils, plant material, etc., and the random isolation and screening of the resident microbial flora. Although capable of significant automation, the throughput capacity largely determines the speed of progress. In the past, screening strategies using soil samples as a starting point were highly empirical. But much can be done to reduce the need for extensive experimentation. These processes involve subjecting an environmental sample (e.g., soil) to selective pressure using the principles of enrichment culture by manipulating the growth conditions so that only microorganisms expressing a particular enzyme are able to grow and survive. For example, by providing only xylan as a source of carbon and energy, only microbes producing a xylanase constitutively or by induction will grow, and the culture

becomes enriched in these organisms. The system is capable of considerable refinement; for example, if an acidic xylanase is required, then the pH of the enrichment culture can be designed to permit growth of microbes producing the enzyme at pH 4 – 6. A successful outcome can be improved by a judicious choice of environment for examination, for example, acidic soils, peat bogs, and forest litter.

An ecological approach takes advantage of the natural selective pressure of the environment. For example, an alkaline environment such as a soda lake at pH 9 is inhabited only by alkaliphilic microorganisms. It is likely that these organisms have extracellular enzymes that are stable and perform optimally at high pH, and this is indeed the case [300]. For screening of feed enzymes such considerations are often less clear cut. What, for example, is the natural habitat for a phytase-producing microorganism? Phytate, a storage phosphate, is often a major component of some seeds, and screening microorganisms in intimate association with the roots of germinating seeds (rhizosphere) would be an appropriate niche to examine.

The time between collection of the environmental sample and its examination in the laboratory can be critical, since changes will inevitably occur. Even in the best-preserved samples, chemical and physical changes may lead to the death of some organisms and reduced viability of others. The longer the time lapse, the more serious these changes can be. There are many advantages in treating the sample immediately in the field, since this can lead to the isolation of a totally different population of microbes. It can be useful to provide “bait” for specific microbes or types of microbial activities. An example of this in situ enrichment is placing cotton linters in an environment such as lake sediment and retrieving the sample at a later date. The sample can then be taken back to the laboratory and screened for cellulolytic activity. Many other enrichment techniques have been developed empirically, such as the use of water- or chitin-agar for isolation of slow-growing actinomycetes, and such ideas are limitless [301].

In addition to enrichment for desired microorganisms, the numbers of unwanted microbes can be reduced. In fact, the first problem in processing of samples is often one of numbers. Composts and soils rich in organic matter contain enormous numbers of microbes. Some form of dilution is required, or targeting of specific groups. Selective techniques can involve, for instance, careful air-drying of the (soil) sample, which reduces the numbers of Gram-negative bacteria, or a brief heat treatment at 80 °C, which kills most microorganisms but leaves the resistant spores of bacteria such as *Bacillus* intact. Antibiotics may be incorporated into the growth medium. For example, Gram-positive bacteria are more sensitive towards penicillin than Gram-negative bacteria.

Whatever the screening strategy adopted, all these approaches rely on some form of selection criteria for the organism-activity combination being sought. The most widely applied methods involve growing the microbes on an agar plate containing the appropriate enzyme substrate for detecting the desired enzyme activity. Often the methods employed are proprietary and frequently constitute more of an art than an exact science. Some methods, however, are well understood. For instance, incorporating casein as sole carbon source into an agar medium with an opaque surface will reveal protease-secreting organisms as a clear halo of hydrolyzed casein around the growing colony. Similarly, lipase activity can be monitored by a change in color of an indicator that visualizes the drop in pH due to the release of fatty acids from a lipid substrate. Specific cleavage reactions can also be detected by release of a fluorescent product or a strongly absorbing (chromogenic) product from an appropriate substrate [302-304]. An interesting possibility is provided by display of a quenched fluorescent substrate on the cell surface of a recombinant host that after cleavage gives a fluorescent product which remains associated with the cell [305].

The selection of microorganisms is a very critical stage in the screening process and can be quite subjective, especially if the numbers involved are very large. Even though the screening has followed some form of selection process it is often not easy to distinguish the different types of microorganism growing on an agar medium. This is particularly true of bacteria, which, unless they are clearly colored, all look much the same. This gives rise to one of the biggest drawbacks of the random approach to isolation and screening of naturally occurring strains: the problem of redundancy. The same microbe will recur time and again from different samples collected from different locations, and this may explain why so many products come from a very limited range of microbes. Until relatively recently most screening strategies were fairly conservative. The tendency to use rich media that favor fast-growing, nutritionally undemanding organisms has lead to an overemphasis on particular groups of microorganisms. This is coupled to the fact that certain fungal genera such as *Aspergillus* and *Trichoderma*, and *Bacillus* and *Pseudomonas* species among the bacteria, have traditionally been a rich and proven source of commercial enzymes. To ensure greater diversity and as a

measure of how extensive a screening has been, some estimate of the biodiversity of the population screened is required. A decade ago this was a daunting task involving a detailed examination of each organism, but nowadays fast molecular methods are available. A range of techniques rapidly group collections of unknown strains into clusters of related organisms. Total cell protein patterns, amplified fragment length polymorphism, ribotyping, amplified ribosomal DNA restriction analysis, internally transcribed spacer region PCR, single-strand conformational polymorphism combined with sophisticated electrophoresis methods, and image analysis for data acquisition and computation can provide information on the diversity of a strain collection. Today, the ease with which small subunit ribosomal RNA (ssu rRNA) gene sequencing can be performed, combined with a large and growing database for comparison, means that specific organisms can often be quickly assigned to known taxonomic groups, or novel microorganisms can be recognized. The clustering of organisms into related groups of strains can be combined with screening of selection data, thus permitting the identification of specific attributes with recognized groups of organisms. This information allows one to devise particular procedures for the isolation of a specific target group and the imposition of more stringent selection criteria. Even so, it is still too easy to overlook potential novel isolates, especially when confronted by many hundreds of culture dishes, each one containing several hundred similar-looking microbial colonies. Nowadays, the task can be performed by colony-picking robots and visual imaging systems coupled to computers.

One problem often encountered early in the screening process is that natural microbial isolates usually produce commercially important enzymes in exceedingly low concentrations. Environmental manipulation that optimizes the growth and production conditions is limited by the intrinsic maximum ability of the wild-type organism. Classical genetic manipulation by selecting mutants with improved properties can lead to an increase of potential yield. The classical techniques of mutation and selection for microorganisms with improved attributes has been applied for many decades. Modern methods of molecular biology have accelerated the possibilities of creating in the laboratory microorganisms with new or improved activities.

4.1.3. Molecular Screening

If a suitable lead enzyme, for example, an acid xylanase, has been identified in a particular species then a search can be made for homologous enzymes in related species. The process requires the application of reverse genetics. Enzymes are polymers of amino acids covalently bonded in a defined sequence. The order in which the amino acids are arranged in the enzyme can be analyzed by a process known as N-terminal sequencing. Since the genetic code is universal it is possible to predict the probable sequence of nucleotide bases in the gene coding for the specific enzyme. Using this information a probe can be constructed consisting of 15 to 20 nucleotides (primer set) in the correct sequence, which, because of the double helix nature of DNA, will bind to the complementary strand. By using the primer set in the polymerase chain reaction (PCR) on chromosomal DNA from related organisms, a homologous gene can be amplified and transferred to a host organism such as *E. coli* for expression of the gene product, i.e., a homologous enzyme [306]. In practice the procedure can be a little more complex than is described here, but it is possible to track down enzymes with identical functionality (e.g., a xylanase) but with a significantly different sequence of amino acids, whereby only the crucial elements of the enzyme such as the active site have been conserved through evolution. In application tests these “new” enzymes may have significantly different properties, such as improved temperature or pH stability compared with the lead molecule. In this way whole families of related organisms can be screened. Alternatively, the microbial DNA can be probed with DNA carrying a radioactive or fluorescent label by a technique called hybridization. However, all molecular methods based on similarity screening have one big drawback. They rely on a comparison with data that has already been recorded.

4.1.4. Environmental Gene Screening

Even though molecular screening is a very powerful technique it seems limited by the ability to grow and extract suitable DNA from the target organisms. It is also confined to known microorganisms that have been isolated, described, and deposited in culture collections. However, it has been estimated that fewer than 1 % of all microorganisms that exist in nature have been isolated, and even fewer have been adequately characterized and described [307]. It is thought that fewer than 25 % of the microbes in a sample taken from the environment can actually be cultured in the laboratory. There may be many reasons for this “unculturable” portion of the microbial community. Some cells may be in a resting phase, damaged, or moribund; some may not be true residents and not form part of the active population; others may be opportunists waiting for a shift in environmental conditions in their favor; while others may have unknown requirements for culture that remain unmet in the laboratory.

Cultivation in the laboratory has a strong selective bias; the techniques are very conservative and have hardly changed for 100 years. Yet we are imprisoned within these constraints since we have no intellectual answers to this problem [308]. The main biodiversity of life is microbial, distributed among three primary relatedness groups or domains: Archaea, Bacteria, and Eukarya. The number of validly described species of prokaryotes is about 4500 and increases at a rate of 100–200 per year. In 1987, WOESE described the bacterial domain as comprised of about 12 naturally related groups (divisions) based on ssRNA sequence analysis of familiar cultivated organisms and showed that they displayed greater evolutionary depth (genetic biodiversity) than plants, animals, and fungi [309]. Phylogenetic studies of cultured and environmental sequences have since expanded substantially our appreciation of the scope of bacterial biodiversity. Three trail-blazing, seminal research reports, all published in 1990, indicated that bacterial biodiversity based on molecular studies was significantly larger than that known from pure cultures. From a DNA-DNA reassociation study it was estimated that a single sample of forest soil contains about 12 000 bacterial species [310]. The other two studies reported cultivation-independent retrieval of partial 16S rDNA sequences from marine bacterioplankton [311] and hot springs [312]. The retrieved 16S rRNA sequences did not match those of known cultured bacteria and thereby provided evidence for numerous as yet unknown bacteria. A more recent examination of environmental DNA derived from diverse habitats revealed, in addition to a vast biodiversity of yet to be cultivated microorganisms, that the cultivated microorganisms may not be the dominant flora in these environments, and this provides a further challenge to the screening by enrichment culture methods [313, 314]. Now, after nearly a decade of direct rRNA sequence retrieval from the natural environment, nearly 40 distinct bacterial divisions (kingdom-level phylogenetic divergences) are resolved. Just four of these divisions account for 90 % of all cultivated bacteria (Proteobacteria, Cytophaga-Flavobacterium-Bacteroides, and the high and low G+C Gram-positives). It is small wonder that most microbial biotechnology is devoted to organisms that fall into these groups. For over 15 of the bacterial divisions, not a single organism in culture is known, and for a further 15 divisions one or two cultured representatives are supported only by the environmental sequence footprint [315]. The phylogenetic differences between the bacterial divisions are probably reflected in substantial physiological differences [316].

Fortunately recent technological developments have permitted access to at least a portion of these uncultivated microbes [313, 317]. It is possible to extract DNA or even RNA directly from the microbial community present in a sample taken from the environment. Although there are many techniques for extracting chromosomal DNA from natural microbial communities, not all cells are equally susceptible to lysis, and the DNAis never recovered to 100 %. Another concern is the detection of organisms that form only a minor proportion of the total community. Their DNA will be under-represented unless efforts are made to normalize the sample. The environmental DNA can be tested for known genes by using the molecular techniques described above. This requires good quality DNA, since PCR screening is particularly sensitive to disturbance by contaminating substances in the environment. Alternatively, the environmental DNA can be cut into smaller fragments by using restriction enzymes, and the fragments cloned into an expression host such as *E. coli*. Likewise, environmental RNA can be used to construct cDNA expression libraries. The clones can then be screened for new enzyme-encoding genes in a manner similar to many of the techniques described for the screening of natural microbial isolates thus allowing the detection of desired activities encoded by genes that might show only limited or no similarity at all to known genes.

4.1.5. Genomic Screening

Spectacular progress has been made in the direct sequencing of the nucleotides of a DNA strand. Already there are millions of gene sequences in public databases. The whole genomes of over 800 organisms can be found in public databases such as Entrez Genomes [`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome`]. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life — bacteria, archaea, and eukaryota — are represented, as well as many viruses and organelles. At present (2003) 80 prokaryotic (66 bacteria and 13 archaea) genomes and several microbial eukaryotic genomes have been sequenced in their entirety, and the sequencing of more than 120 microbial genomes is in progress. Random shotgun sequencing allows rapid elucidation of microbial genomes [318], and technical advances are expected to further accelerate genome sequencing. Originally, pathogenic species were the main focus of microbial sequencing projects, but many microbes with industrial potential are being sequenced as well. For example, the genomes of thermophilic, alkaliphilic, acidophilic, and salt-tolerant microbes have been sequenced [319-323]. This illustrates the general belief that genome analysis will reveal

the mechanisms underlying the respective lifestyles of these microbes, aid in identifying industrially relevant extremophilic enzymes, and possibly allow the recognition of general protein rules that result in enzymatic performance under extreme conditions. Global analysis of thermophilic proteins shows, for instance, a tendency to reduced protein size; increased hydrogen bonding, beta-strand content, and helix stabilization through ion pairs; and preference for charged residues over uncharged polar residues [319], [324], [325].

When a microbial genome sequence is available, protein-coding genes can be identified computationally by essentially two different methods [326]. Ab initio searches use a template that integrates general knowledge about gene composition and signals to find novel genes. A more reliable computational method uses similarity to a protein-coding gene in another organism to identify novel genes. The latter method depends, however, on the availability of homologous genes, but it has as added benefit the ability to indicate gene function. Since prokaryotic genes frequently overlap with each other, exact prediction is notoriously difficult, and as a result the true number of protein-coding genes for sequenced microbial genomes remains debatable [327], [328]. A further complication that adds to the discrepancy between gene predictions and expressed proteins is presented by alternative splicing, which can occur in microbial eukaryotes [329]. Several experimental approaches, such as EST sequencing and scanning-tiling microarray analysis, allow validation of computationally identified genes [330].

The enormous amount of sequence data in the public domain has driven the emergence of bioinformatics, which can be defined as the interface between biological and computational sciences. Bioinformatics uses computational tools to organize and analyze biological data and aims to extract biological information and knowledge. Typical examples include identification of homologous genes or specific protein domains and transfer of functional knowledge associated with these homologues or domains to predict function for the new genes [331-333]. For example, similarity searching at the gene level allows in silico screening for enzymes [334]. Note, however, that sequence data are certainly not flawless, and a considerable error rate should therefore be considered in the analysis [331], [335]. Error frequency can be reduced by integration of additional information that can be extracted from genomic or experimental data [336]. Other computational methods that make use of available genomic sequences to infer function rely on, e.g., genomic clustering of functionally related genes [337], [338], phylogenetics [339], [340], metabolic pathway alignment [341], [342], and combined approaches [343]. Despite such powerful tools, about 25 % of the putative genes revealed by total genome sequencing projects cannot currently be assigned a function because they only show homology to other unannotated genes. Another 25 % of the genes are of unknown function and seem to be unique to their particular species [344]. These data indicate that the resources for screening of novel enzymes are still largely untapped. Sparked by these findings several new high-throughput experimental technologies have been developed that aim to identify functions for these orphan genes.

Many methods have been developed that use loss of function, cellular localization, protein interaction, or timing of expression to predict function for novel proteins and can be claimed as tools for so-called functional genomics. Essentially, most of these methods infer function indirectly as “guilt by association” in time or space with characterized proteins. For example, a wide variety of techniques are available to disrupt genes, random or directed, on a large scale in virtually any organism (Hensel and Holden, 1996; [345-350]). Subsequent phenotypic analysis, quite similar to natural isolate screening, allows determination of putative biological functions of the inactivated proteins. Some of the gene disruption methods also allow analysis of protein localization [351], which can aid functional annotation. When specifically screening for extracellular enzymes, secretion can also be predicted on the basis of sequence information [352] or transcript analysis [353]. Alternatively, the use of yeast-based signal sequence traps allows selection of secreted proteins from both eukaryotes and prokaryotes without prior sequence information [354]. Large-scale analysis of protein interactions by genetic 2-hybrid or biochemical/mass spectrometric approaches has also been used to assign function [355]. In addition, computational methods have proven successful in predicting functional interaction between proteins [357].

So far, however, genome-wide transcript analysis has emerged as most used tool for functional genomics. Based on the assumption that proteins with similar functions show similar regulation, transcript data can be used to functionally classify genes [356], [358], [359]. Microarrays of oligonucleotides or entire genes representing the complete genome can be hybridized with differentially labelled cDNAs obtained from separate cells or cell states. Comparison of transcript profiles indicates regulation of gene expression as a result of changes in the cell's conditions. Alternatively, differential expression of proteins can also be followed directly at the level of protein — albeit still at lower throughput — instead of indirect assessment

via the analysis of transcription [360]. Co-regulation of genes (or proteins) can then be used to identify genes that are relevant to a particular cellular process. For example, global transcript analysis in *Thermotoga maritima* is used to identify genes involved in the biodegradation of simple and complex plant polymers with potential as renewable energy sources, such as xylan and cellulose [361]. Designing a whole-genome DNA microarray for a fully sequenced microbe is conceptually straightforward [352]. Arraying multiple DNA arrays simultaneously can further increase throughput of expression analysis. For example, expression levels of more than 6000 human genes in 49 different cell types were recently detected in a single hybridization experiment [363]. Many other applications exist for the use of DNA microarrays in addition to transcript profiling [359].

4.1.6. Proteomic Screening

The above-mentioned functional genomics approaches are all clearly very valuable, but they still infer protein function indirectly. Direct assays for function require high-throughput expression of the identified genes, followed by functional analysis of purified proteins. This approach introduces new challenges to protein expression, including the need for high-throughput cloning and purification [364]. Many new technologies that facilitate such large-scale approaches are not limited to the full complement of sequenced genomes, the proteome, but can be adapted for the functional analysis of large numbers of recombinant proteins, for instance, those encoded by cDNA or environmental libraries.

Although any gene can be engineered to alter specific restriction sites and facilitate cloning into a suitable vector, this process cannot efficiently be performed when dealing with large sets of genes. To circumvent this, several recombination-mediated cloning strategies have been developed that allow flexible cloning into many different vectors [365-367]. High-throughput protein purification requires fusion with specific affinity tags that facilitate purification, immuno-localization, and/or immuno-precipitation [368], [369]. For example, genome-wide fusion to different tags has allowed parallel purification and characterization of thousands of different proteins. Protein fusions have also been used to improve secretion and solubility of small or low-solubility proteins. Although addition of tags might affect protein function, they strongly facilitate detection and purification, and can allow cell surface display or immobilization on beads or any other support [370-372].

Some exciting examples of proteomic screening were reported in the late 1990s and early 2000s. Expressed more than 6000 yeast genes fused to glutathione S-transferase (GST) were expressed [370]. Pools of purified fusion proteins were assayed for different enzymatic activities, and after deconvolution specified activities could be assigned to orphan genes. In a related approach nearly all protein kinases from baker's yeast were arrayed in microwells and simultaneously analysed by using 17 different substrates [373]. The same authors also expressed most yeast genes with a 6xHIS tag, which allowed binding of the fusion proteins to nickel-coated slides. The purified proteins were printed on these slides and screened for their ability to interact with proteins and phospholipids. Probing the proteome chips yielded many new functional interactions [371]. In two independent studies fusion proteins were also used to purify and characterize multiprotein complexes on an unprecedented scale [372], [374]. In a related study it was shown that an affinity purified TAP-tagged multiprotein complex retained its activity [375].

Many ideas to construct protein chips are being pursued [376] but from an enzyme screening perspective attention should focus on methods that are most likely to retain enzymatic activity. In an interesting alternative [377], instead of printing proteins, an expression library is printed and overlayed with competent cells, resulting in localized expression by transformants. The cell clusters can then be screened for clone-dependent properties. Cell surface display of protein libraries has been used to select novel or improved binders but can also be used to screen for enzymatic activity [378]. For example, mutants with improved catalytic activity have been identified by surface display of cellulase [379] and subtiligase [380]. Combined display of a quenched fluorescent substrate and a library of protease variants has been used to screen for novel substrate specificity [305]. Such approaches should also make it feasible to screen expression libraries of natural diversity.

4.2. Protein Engineering

4.2.1. Introduction

Protein engineering employs a combination of tool sets that allow any amino acid in the amino acid sequence of a protein to be replaced by one of the other 19 naturally occurring amino acids of proteins in order to change their properties. The genetic code can also be expanded by the introduction of nonnatural

amino acids into a protein [381-383]. These changes can be either directed to one position or randomly spread along the whole primary sequence.

The establishment of protein engineering as a new technology has been enabled by the advent of recombinant DNA technology, accompanied by the continuous improvement of 3D structure determination by X-ray crystallography and NMR spectroscopy.

Until the 1980s the possibilities for changing an enzyme were limited to chemical modifications, which are restricted to the reactive amino acid side chains and are neither regio- nor site-specific.

Recombinant DNA technology has revolutionized protein modification in a number of ways. It is now not only possible to replace one particular amino acid by one of the nineteen others, but these changes are also permanent. New material of the modified protein of the same quality is available simply by repeated fermentation of the production organism, followed by purification. In addition, enzymes produced in recombinant systems are available in larger quantities and higher purity. This in turn facilitates the determination of protein 3D structures and thus increases the knowledge base for further protein-engineering experiments.

Both technologies for protein structure determination, X-ray crystallography and NMR spectroscopy, have benefited from the availability of larger quantities of pure proteins, but without the simultaneous development of the physical methods the huge amount of structural data would not be available today (see Chap. Introduction). The quality of X-ray data has been improved by hardware developments of the traditional X-ray beam sources and, in particular, by the availability of coherent X-ray beams generated by synchroton radiation, which allow the resolution of crystallographic data to be brought down to down to even less than 1 Å (Fig. 22). At this level hydrogen atoms and hydrogen bonds in protein 3D structures can be directly observed [384], [385].

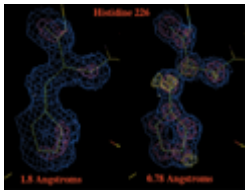

Figure 22. Information content of electron-density maps at different resolutions. The figure compares information available at 1.8 Å (left) with 0.78 Å (atomic resolution, right). Both maps were calculated from the same refined data set and phases derived from a refined model at 0.78 Å resolution. In both maps a standard $2F_o–F_c$ electron density was contoured at σ (blue) and 4 σ (magenta). At 0.78 Å, the 4 σ contour delineates individual atoms, and the relative dimensions indicate whether the atom is carbon, nitrogen, or oxygen having 6, 7, or 8 electrons, respectively. Contoring at 4 σ in the 1.8 Å plot does not convey this information. In the 0.78 Å map positive difference electron density (yellow) is observed for hydrogen electrons, which will be valence electrons participating in covalent bonds. As such, the positive electron density will not be centered about the hydrogen atom nucleus but rather about the covalent bond.
[Full View]

At ultra-high resolution of 0.54 Å it is even possible to study the valence electron density of proteins [386].

NMR spectroscopy has gained from the development of better magnets that allow better separation of signals and finer determination of the chemical shifts of the atoms. NMR spectroscopy is also influenced by the general developments in biotechnology that allow the effective production of the necessary ^{13}C- and/or ^{15}N-labeled proteins.

Both structure determination methods, however, were most influenced in the 1980s by the sudden availability of continuously increasing computational power. Since protein X-ray crystallography and NMR spectroscopy are indirect methods, they heavily depend upon computational power to process the data, and they require visual inspection of the results on high-end graphical computers.

The graphical computer systems paved the way for the establishment of the important knowledge base for protein engineering, the establishment of the structure-function relationship. Until then it was only possible to study the events around enzymatic catalysis by interpretation of the kinetic data under different conditions, but with the availability of structural data and cheaper graphics computers for visualization, it became possible to rationalize the observation at atomic detail by combining experimental and structural data. A very good example for this is the structural explanation for the reduction in the catalytic activity of

subtilisin due to oxidation. BOTT *et al.* [387] determined the structure of the oxidized form of subtilisin (EC 3.4.21.62) from *Bacillus amyloliquefaciens* and compared it to the nonoxidized native form (Fig. 23). In the oxidized form, the oxygen atom of the oxidized methionine 222 projects into the oxyanion hole. Thus, the substrate carbonyl atom will not bind in an identical manner as in the native enzyme. This may not only affect the relative binding energy of the substrate K_m, but the catalytic rate as well, since the oxyanion hole has been shown to be critical for catalysis. Substitution of Asn 155, for example, has been shown to reduce catalytic activity as measured by k_{cat} by 100 – 1000-fold [388].

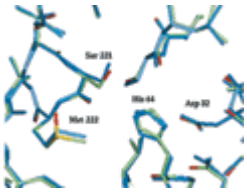

Figure 23. Comparison of the structure of native subtilisin (blue) with the structure of the enzyme with peroxide-oxidized met 222.
[Full View]

Besides structural knowledge, the generation of mutants is a major aspect of protein-engineering experiments. In older days, the majority of mutational experiments were performed by so-called cassette mutagenesis [389], [390], a technique with which all nineteen possible natural amino acids could be substituted at a given site. From these studies variants having altered substrate specificity were found that displayed increased specificity by as much as a 1000-fold by introduction of a single mutation. Moreover, the p*K* values of catalytic residues could be modified and the pH rate profiles could be tailored to the desired performance pattern. This technique allowed the generation of site-directed mutations and of small random mutagenesis banks in defined regions of the molecules. Using this technique, numerous sites near critical subsites were determined and evaluated in subtilisin. Three amino acid changes were sufficient to interchange specificities between related subtilisins [391], while in trypsin numerous changes in addition to the position 189 side chains were needed to interchange trypsin and chymotrypsin specificities [392].

These efforts are indicative of another recurring theme in modern protein engineering, namely, the exploitation of natural diversity for guidance in selecting sites and substitutions. Natural diversity is the result of millions of years of adaptation to various extremes of conditions such as temperature, pH, and substrates. Within a certain class of enzyme the overall tertiary folding pattern is often highly conserved, and thus it is possible to associate changes that can often be introduced at particular positions and associate specific substitutions with particular alterations in performance profiles and substrate specificities. The theme of exploiting natural diversity reappears in most commonly used techniques, either for rational selection of sites and substitutions or by recombination as a means of extending diversity.

Mutagenesis Techniques. After the recognition that recombinant DNA technology could be used to modify the sequence and thereby the function of an enzyme by altering the DNA sequence encoding for the protein, the next advancement came with the advent of cassette mutagenesis [389]. Complete sampling of all natural amino acid substitutions became realizable. With this tool, the full repertoire of diversity was now available for molecular evolution, at least one site at a time.

Around the same time, the idea of alanine scanning was introduced. In this model, many sites could be examined by replacement of existing amino acid residues with alanine. In this case, the diversity looks to identify sites that may play an important role for functioning of the enzyme. In this approach potential sites are screened that can later be explored by structural analysis and/or diversity generation.

The discovery of error-prone polymerase chain reaction (PCR) was the first revolution in enzyme-variant generation. In fact, PCR allows both the production of site-directed protein variants and the generation of random variant libraries over the whole gene or defined parts of it. Mutation frequency of the error-prone PCR experiments can be driven by the reaction conditions and the choice of the polymerase.

Another method for the generation of molecular diversity at the gene level is gene shuffling [393-395]. In gene shuffling experiments a set of homologous genes is pooled and cut into pieces. Finally the fragments are recombined randomly in a subsequent PCR. As a result an enzyme variant library is obtained that allows the combinations of all natural variations that were distributed across the genes in the input set to be

explored. The variations not only allow variations of individual amino acids to be tested, but also the effect of amino acid insertions or deletions among the different proteins.

The ability to identify the right variants for a given purpose determines the success of a protein-engineering experiment. A site-directed variant is of course easy to characterize, but the screening of a random library requires excellent knowledge about the assay that is used for screening.

Phage display of proteins has become an important tool for screening and selection of proteins. In a phage-display approach, the gene which encodes for the protein to be engineered is linked to one of the genes which encode for the phage's surface proteins. As a consequence the protein is expressed at the N- or C-terminus of one of the surface proteins. Since it is on the surface, the displayed protein can be selected for binding to a certain target. Phage display allows selection from more than 10^9 variants for binding in a single experiment. In addition the selected protein is always isolated in combination with the gene that codes for it. The gene can be multiplied by infection of a host organism with the phage, followed by either sequencing or a subsequent mutagenesis/selection cycle. Phage display has now been proven to be a good tool for selection of proteins that bind selectively to a target molecule.

Although active enzymes have been expressed on phages, phage display is not a suitable technique to change the catalytic properties of enzymes. Because the binding to a transition state analogue, which can be used for a selective binding screen, does not correlate well with improved catalysis [396], [397], it is necessary to design special capturing systems for the detection of phage-displayed enzymes with high catalytic activity [398].

For the improvement of an enzyme for technical application it is of outmost importance that the selective activity assay closely resembles the final application. Several examples have clearly demonstrated that activity assays with artificial substrates show no correlation with performance in real-world applications. This was shown by correlating wash performance of a protease with the enzyme's activity on casein [399]. The directed-evolution experiment for the improvement of the activity of galactose oxidase (EC 1.1.3.9) revealed a reasonable correlation between the activity of the different galactose oxidase variants on the model substrate methylgalactose and their performance on the technical substrate guar. Ironically the best performer in the technical application had to be excluded from the dataset to obtain the good correlation coefficient of r^2=0.92 (with the best performer included the correlation coefficient was r^2=0.72) [400].

It must be clear that none of the protein engineering tools is able to improve enzymes alone. The strength of protein-engineering projects lies in the combination of structural/functional knowledge with the right molecular biology approach for the generation of variants.

Finally, the importance of screening technology has grown during the last few years. Mainly, automation in combination with miniaturization of assays down to microscopic scales has lead to a shift from pure site-directed mutations to regio-directed random mutagenesis experiments followed by intensive variant library screening.

The most comprehensive use of protein engineering is known under the names directed evolution, molecular evolution, and molecular breeding. Common to these approaches is the iterative follow up of randomization and selection steps. Directed evolution mimics the natural evolutionary process in the laboratory. It can be very successful, because the laboratory allows a nonnatural selective pressure to be applied to direct the successive improvement of a particular enzyme to improve its performance in a technical environment. Directed evolution proved to be very successful. Often three or fewer iterative cycles of randomization and selection for a desired property can achieve enormous improvements. This success can mainly be explained by the fact that the directive influences of natural evolution do not force a molecule to perform well in, e.g., a laundry washing machine. In addition, the researcher can introduce a much higher mutation frequency than occurs in Nature.

4.2.2. Application of Protein Engineering in Academia and Industry

Both the academic scientific community and industrial scientists embraced the new protein engineering technologies when they became available in the second half of the 1980s. In academia, protein engineering is used extensively for elucidating intrinsic protein and enzyme properties such as stability, catalytic

mechanism, and activity under different conditions. Industrial scientists focus on engineering for improved performance of enzymes in industrial applications.

Deciphering the Catalytic Mechanism. Protein engineering can be used to identify amino acids which participate in the catalytic mechanism, before or even after the availability of 3D structural information. The most comprehensive studies on enzyme mechanisms by protein engineering were performed on tyrosyl-tRNA synthetase starting in the 1980s [401]. The complete investigation has been summarized elsewhere [402].

Protein engineering has also been used for the identification of the active sites in proteins, for example, in lipases. The principle nucleophile in hormone-sensitive lipase was identified by sequence analysis of the characteristic GXSXG segment of the α/ β-hydrolase fold and subsequent replacement of serine 423, by several other amino acids. Since all variants had no activity at all, the authors could conclude that serine 423 is indeed the active site serine of this lipase [403]. Other groups identified in similar experiments the whole catalytic triad of different lipases [404], [405]. The investigations of MARTINELLE, HOLMQUIST *et al.* [406], [407] expand this concept by investigating the role of amino acids in the proximity of the active site of *Humicola lanuginosa* lipase in substrate specificity. They combined mutagenesis results of interesting mutants with careful kinetic investigations of the variants under different conditions.

Stability of Enzymes. Fundamental Research. The intrinsic stability of proteins and the influence of single amino acids on the stability of proteins have been investigated with protein-engineering. Since the effect of exchanging an amino acid on protein structure depends on its position in the protein structure, it is necessary to use the right model system and to plan the mutation on the basis of the 3D structure of that protein.

One route to a more stable protein is the introduction of disulfide bonds by site-directed mutagenesis. In subtilisin, an enzyme having no cysteine residues, the introduction of cysteine for serine or alanine at positions 22 and 87 or 24 and 87 resulted in the formation of disulfide bridges [408] but yielded enzymes with little or no improvement in overall stability. CLARKE *et al.* showed in a study on barnase that engineered disulfide bonds can stabilize and destabilize proteins. The most stabilizing disulfide bridge does not change the overall structure of the molecule and leads to local stabilization according to hydrogen-exchange experiments. The destabilizing disulfide bridge induced considerable structural changes, including the displacement of a loop in the vicinity of the new disulfide bond [409], [410].

The influence of single amino acid replacements has been studied in detail with several model proteins. The studies clearly show that the influence of an amino acid on helix stability depends on its intrahelical position, the position of the helix in the globular protein structure, and the solvent accessibility of the mutation site [411-413].

The complexity of interaction between stabilizing and destabilizing mutations during evolution has been demonstrated in a comparative mutagenesis study with barnase and binase. Both are RNAses with 85 % sequence identity. Only 17 amino acids, scattered over the whole sequence, differ in these proteins, which have the same 3D structure but differ in stability. One-by-one replacement and combinations of the single mutations showed that each mutation has an effect on stability and that the effects are additive. However, a combination of only six of these mutations created a protein more stable than either of the parent molecules without affecting the activity [414]. This study shows that a simple comparison of sequences and 3D structures of enzymes from mesophilic and thermophilic organisms will not lead to the identification of the stabilizing amino acid changes, because nature balances several effects during evolution to obtain a set of enzymes which is good enough for its purpose.

Industrial Application. Enzymes have been successfully incorporated into industrial starch processing. In the first stage α-amylase is used to break down starch into short oligosaccharides. In the commercial process the enzyme must survive passage through a steam jet at 105 °C, which is necessary to explode the starch granule, and function at 90 °C for 1 – 2 h. With the natural amylase of choice, *B. licheniformis* α-amylase (BLA), the pH must be adjusted to 6 and then readjusted in subsequent stages. An engineered α-amylase that would operate at lower pH would save considerable time and money by eliminating the need for pH adjustment. Analysis of BLA as a function of pH indicated that stability of the enzyme was reduced at lower pH. Initially, random mutagenesis and screening was performed in the absence of a reliable model of its three-dimensional structure, and several variants were identified [415]. Once a structure became available it

was noticed that variants represented substitutions found to exist between related α-amylases. Surprisingly, the related α-amylases had lower overall stability than BLA. By using the three-dimensional structure as a guide it was realized that all of the variant substitutions were located on the surface and filled a cavity present on the surface of BLA. An additional five residues were identified, and each individually increased the stability of the enzyme and could be combined with a cumulative increase in stability by a factor of 23.

Subtilisin operates in an environment quite different from its natural milieu. As a detergent additive, subtilisin must perform at highly alkaline pH, high surfactant concentration, and high temperature. It must operate on a wide variety of proteinaceous soils and may encounter other agents such as oxidants like hydrogen peroxide, added as a bleaching agent. Thus the enzyme was engineered to be more active and stable under these conditions. One of the most beneficial results has been mutations from which susceptibility to oxidants found in commercial bleaching agents has been eliminated. This was primarily achieved by replacing a bleach-sensitive methionine residue [416].

Investigations on Protein Folding. The application of protein engineering for the dissection of protein folding pathways has been reviewed [417]. The number of different experimental techniques that have been used for this purpose clearly emphasizes the interdisciplinary nature of protein engineering.

Changing the Substrate Specifity of Enzymes. The use of enzymes as biocatalysts for chemical reactions has recently regained interest (see Enzymes in Organic Synthesis). Several protein-engineering projects have been successfully performed. They mainly applied different directed-evolution approaches to demonstrate their capabilities for the development and improvement of new biocatalysts.

The activity of subtilisin E in aqueous DMF was improved by protein engineering [418], [419]. Subtilisins have only low activity and reduced stability in DMF. The authors were able to enhance the stability in 40 % DMF more than twofold. But even more important is the activity enhancement that was achieved by subsequent evolution cycles. The final molecule showed a 471-fold activity improvement in 60 % DMF on a test substrate compared to the wild-type molecule.

The ability to change the substrate range and selectivity by protein engineering methods has been demonstrated by several authors. ZHANG *et al.* [420] used DNA shuffling to modify the substrate specificity of β-galactosidase (EC 3.2.1.23). Whereas the wild-type molecule showed no activity on test substrates with fucose moieties, the final enzyme was active against these substrates. The activity against galactose substrates dropped significantly.

The success of engineering of enantioselective enzymes finally demonstrates that protein-engineering technologies nowadays can even address the subtlest problems in chemical synthesis. In several studies on various enzymes it has been demonstrated that it is possible to protein-engineer enzymes to obtain improved enantioselectivity, mainly for hydrolysis reactions [421-423].

4.2.3. Outlook

There are similarities between natural and directed evolution, the difference being largely in the management of diversity. Protein engineering has always been directed evolution where screening and diversity management intervene in place of natural selection. Even now knowledge of the three-dimensional structure and its relation to function is often pivotal in locating critical sites and regions for focused diversity and recruitment. DNA shuffling techniques again use screening strategies to combine features between natural selected genes in an effort to recombine or recruit the desirable features in a family of enzymes.

Now with the knowledge of whole genomes and the expected comparison of the genomes of closely related species, it may be possible to perform in silico shuffling and perhaps understand inherent preferences on individual molecular evolution imposed by the organisms' metabolome — the ensemble of molecular functions necessary for the maintenance of metabolism and reproduction in the whole organism. Another element of in silico directed evolution is the requirement to be able to predict the three-dimensional structure and with this further be able to recognize the catalytic potentials and even to deduce the probable physiological role of that molecule within the metabolome.

PREVIOUS NEXT

[Top of page] [A to Z] [Authors] [Subjects] [Search] [HitList]
© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Charlotte Horsmans Poulsen, Danisco A/S, Brabrand, Denmark
Karsten M. Kragh, Danisco A/S, Brabrand, Denmark
Jens Frisback Sorensen, Danisco A/S, Brabrand, Denmark
Catherine Grassin, DSM Food Specialties, Seclin Codex, France
Marga Herweijer, DSM Food Specialties, Delft, The Nederlands, France
Jan Wilms, Genencor International B. V., Leiden, The Nederlands
André de Roos, DSM Food Specialities, Delft, The Netherlands
Jorn Borch Soe, Danisco A/S, Brabrand, Denmark
Ivan Herbots, Procter & Gamble Eurocor S.A., Strombeek-Bever, Belgium
Beatrix Kottwitz, Henkel KGaA, Düsseldorf, Germany
Peter J. Reilly, Iowa State University, Ames, USA
Richard L. Antrim, Grain Processing Corporatio, Muscatine, Iowa, USA
Heidi Burrows, Finfeeds International, Marlborough, UK
Herman B. M. Lenting, TNO Industry, Enschede, The Netherlands
Liisa Viikari, VTT Biotechnology, VTT, Finland
Anna Suurnäkki, VTT Biotechnology, VTT, Finland
Marja-Leena Niku-Paavola, VTT Biotechnology, VTT, Finland
Johanna Buchert, VTT Biotechnology, VTT, Finland
Karl-Heinz Maurer, Henkel KGaA, Düsseldorf, Germany
Andrea Saettler, Henkel KGaA, Düsseldorf, Germany

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

5. Industrial Enzymes

5.1. Enzymes in Food Applications

5.1.1. Enzymes in Baking

5.1.1.1. Introduction

Bread and baked products are among the main nutritional sources today. Over 600 × 10^6 t of wheat are grown in the world each year, making it the single most important crop [424]. Europeans are estimated to obtain approximately half their required carbohydrates and about one-third of their protein from bread [425].

Throughout the history of bread making, enzymes have always played an important role. Even the ancient Egyptians made use of enzymes present endogenously in the flour, although they may not have been aware of the effect. However, not until the 20th century were enzymes used as flour improvers [426]. The first application of enzymes in baked goods was supplementation of α-amylase by addition of malt to correct the concentration of endogenous α-amylase in the flour. Later malt was substituted by microbial α-amylases having a more suitable thermostability for baking. This was the first enzyme industrially produced as a tool to compensate for variations in natural flour. Today, a whole range of enzymes is available for end users of flour. These make it possible to correct suboptimal concentrations of endogenous flour enzymes. Furthermore, these biological catalysts are able to produce value-adding ingredients in situ from compounds present in the flour, and thus it is possible to upgrade lower grades of flour.

In many cases the beneficial effects of using enzymes in the baking industry are also obtainable by using chemical agents such as sodium metabisulfite, cystein, azodicarbonamide (ADA), potassium bromate, etc. However, the food market today shows a clear trend towards more “green” products, and this has clearly favored the use of the biological catalysts enzymes.

Baking enzymes are summarized in Table 9.

Table 9. E.C. numbers; reactions; and recommended, systematic, and other names of baking enzymes according to IUBMB (http://www.chem.qmul.ac.uk/iubmb/)

Recommended name	E.C. number	Systematic name	Reaction	Other name(s)
α-Amylase	E.C. 3.2.1.1	1,4- α-D-glucan glucanohydrolase	endohydrolysis of 1,4- α-D-glucosidic linkages in polysaccharides containing three or more 1,4- α-linked D-glucose units	glycogenase; α amylas amylase; endoamylase amylase A
Glucan 1,4- α-maltohydrolase	E.C. 3.2.1.133	1,4- α-D-glucan α-maltohydrolase	hydrolysis of (1 → 4)- α-D-glucosidic linkages in polysaccharides so as to remove successive α-maltose residues from the nonreducing ends of the chains	maltogenic α-amylase
Endo-1,4- β-xylanase	E.C. 3.2.1.8	1,4- β-D-xylan xylanohydrolase	Endohydrolysis of 1,4- β-D-xylosidic linkages in xylans	endo-1,4-xylanase; xyl 1,4-xylanase; endo- β- xylanase; endo-1,4- β- xylanase; 1,4- β-xylan xylanohydrolase; β-xy 1,4-xylan xylanohydro D-xylanase
Glucose oxidase	E.C. 1.1.3.4	β-D-glucose:oxygen 1-oxidoreductase	β-D-glucose + O_2 D-glucono-1,5-lactone + H_2O_2	glucose oxyhydrase; corylophyline; penatin aerodehydrogenase; m β-D-glucose oxidase; D oxidase; D-glucose-1-c β-D-glucose:quinone oxidoreductase; glucos oxyhydrase; deoxin-1;
Hexose oxidase	E.C. 1.1.3.5	D-hexose:oxygen 1-oxidoreductase	β-D-glucose + O_2 → D-glucono-1,5-lactone + H_2O_2	
Galactose oxidase	E.C. 1.1.3.9	D-galactose:oxygen 6-oxidoreductase	D-galactose + O_2 → D-galacto-hexodialdose + H_2O_2	D-galactose oxidase; β- galactose oxidase
Pyranose oxidase	E.C. 1.1.3.10	pyranose:oxygen 2-oxidoreductase	D-glucose + O_2 → 2-dehydro-D-glucose + H_2O_2	glucose 2-oxidase; pyr oxidase
Thiol oxidase	E.C. 1.8.3.2	thiol:oxygen oxidoreductase	4 R′C(R)SH + O_2 → 2 R′C(R)S-S(R)CR′ + 2 H_2O	sulfhydryl oxidase
L-Amino-acid oxidase	E.C. 1.4.3.2	L-amino-acid:oxygen oxidoreductase (deaminating)	L-amino acid + H_2O + O_2 → a 2-oxo acid + NH_3 + H_2O_2	ophio-amino-acid oxid
Lipoxygenase	E.C. 1.13.11.12	linoleate:oxygen 13-oxidoreductase	linoleate + O_2 → (9*Z*,11*E*)-(13*S*)-13-hydroperoxyoctadeca-9,11-dienoate	lipoxidase; carotene ox lipoperoxidase; fat oxi lipoxydase; lionoleate: oxidoreductase
Peroxidase	E.C. 1.11.1.7	donor:hydrogen-peroxide oxidoreductase	donor + H_2O_2 → oxidized donor + 2 H_2O	myeloperoxidase; lactoperoxidase; verdoperoxidase; guai peroxidase; thiocyanat peroxidase; eosinophil peroxidase; Japanese r peroxidase; horseradis peroxidase (HRP); ext

				peroxidase (MPO); chl peroxidase; heme pero MPO; oxyperoxidase; protoheme peroxidase; pyrocatechol peroxidas scopoletin peroxidase
Catechol oxidase	E.C. 1.10.3.1	1,2-benzenediol:oxygen oxidoreductase	2 catechol + O_2 → 2 1,2-benzoquinone + 2 H_2O	diphenol oxidase; *o*-diphenolase; phenolase polyphenol oxidase; ty pyrocatechol oxidase; oxidase; catecholase; *o* diphenol:oxygen oxidoreductase; *o*-diph oxidoreductase
Laccase	E.C. 1.10.3.2	benzenediol:oxygen oxidoreductase	4 benzenediol + O_2 → 4 benzosemiquinone + 2 H_2O	urishiol oxidase; urush oxidase; *p*-diphenol ox
Monophenol monooxygenase	E.C. 1.14.18.1	monophenol,L-dopa:oxygen oxidoreductase	L-tyrosine + L-dopa + O_2 → L-dopa + dopaquinone + H_2O	tyrosinase; phenolase; monophenol oxidase; catechol oxidase; polyphenolase; pyroca oxidase; dopa oxidase; chlorogenic oxidase; catecholase; polyphen oxidase; monophenola diphenol oxidase; chlo acid oxidase; diphenol o-diphenolase; tyrosine oxidase; o-diphenol:ox oxidoreductase; polyar oxidase; monophenol monooxidase; o-diphe oxidoreductase; monop dihydroxyphenylalanin oxidoreductase; *N*-acet hydroxytryptophan oxi monophenol, dihydrox phenylalanine oxygen oxidoreductase; *o*-diph oxidoreductase; pheno
Triacylglycerol lipase	E.C. 3.1.1.3	triacylglycerol acylhydrolase	triacylglycerol + H_2O → diacylglycerol + a carboxylate	lipase; triglyceride lipa tributyrase; butyrinase; ester hydrolase; tributy Tween hydrolase; stea triacylglycerol lipase; triacetinase; tributyrin Tweenase; amno *N*-AP 1969-4-9; Meito MY 3 Tweenesterase; GA 56 capalase L; triglyceride hydrolase; triolein hyd tween-hydrolyzing est amano CE; cacordase; triglyceridase; triacylg ester hydrolase; amano amano AP; PPL; glyce hydrolase; GEH; meito OF lipase; hepatic lipa

				lipazin; post-heparin p protamine-resistant lip resistant post-heparin l heparin releasable hep lipase; amano CES; an tributyrase; triglycerid liver lipase; hepatic monoacylglycerol acyltransferase
Phospholipase A2	E.C. 3.1.1.4	phosphatidylcholine 2-acylhydrolase	phosphatidylcholine + H_2O → 1-acylglycerophosphocholine + a carboxylate	lecithinase A; phospha phosphatidolipase; phospholipase A
Phospholipase A1	E.C. 3.1.1.32	phosphatidylcholine 1-acylhydrolase	phosphatidylcholine + H_2O → 2-acylglycerophosphocholine + a carboxylate	
Phospholipase C	E.C. 3.1.4.3	phosphatidylcholine cholinephosphohydrolase	A phosphatidylcholine + H_2O → 1,2-diacylglycerol + choline phosphate	lipophosphodiesterase lecithinase C; *Clostrid welchii* α-toxin; *Clostr oedematiens* β- and γ- lipophosphodiesterase phosphatidase C; heat-hemolysin; α-toxin
Phospholipase D	E.C. 3.1.4.4	phosphatidylcholine phosphatidohydrolase	A phosphatidylcholine + H_2O → choline + a phosphatidate	lipophosphodiesterase lecithinase D; choline phosphatase
Peptidase	E.C. 3.4.x.xx	Depending on specific peptidase	Endo- or exo- hydrolysis of peptide bonds	proteases
Protein-glutamine γ-glutamyl-transferase	E.C. 2.3.2.13	protein-glutamine:amine γ-glutamyltransferase	protein glutamine + alkylamine → protein N^5-alkylglutamine + NH_3	transglutaminase; Fact fibrinoligase; fibrin sta factor; glutaminylpepti glutamyltransferase; p transglutaminase; tissu transglutaminase; *R*-gl peptide:amine γ-glutar transferase

5.1.1.2. Amylases

Amylases for Flour Standardization. The most widely used enzyme in bread baking in terms of amount dosed and value is so-called fungal α-amylase or Taka-amylase from *Aspergillus oryzae* (E.C. 3.2.1.1). This is supplemented to sound flours to optimize their α-amylase activity with regard to final volume and crumb structure of the baked product. The primary effect of amylase supplementation was previously believed to be securing adequate gassing power by degradation of damaged starch granules in the dough, which facilitates maltose production by endogenous β-amylase. The resulting maltose can serve as fermentable sugar for yeast fermentation [427]. This mechanism may be important in some very lean dough systems without added sugars, but generally the primary effect is reduction of dough viscosity during initial starch gelatinization [428], which provides a prolonged oven spring leading to an increased volume and a more homogenous crumb structure.

Taka-amylase is named after Takamine, who in 1894 obtained US patent 525 823 on the use of mold fungi for production of amylase. The use of fungal amylase in bread baking was patented in 1928 [429] but first approved in 1955 for use in bread baking in the USA. The advantage of Taka-amylase is a very low protease

side activity and an adequate thermostability, so that inactivation of the enzyme occurs after initial starch gelatinization (70 – 80 °C). The most common alternatives — malt amylases from barley or wheat malt — have higher protease side activities and higher thermostability, and this makes them more prone to negative side effects when overdosed [430].

Antistaling Amylases. The second important goal for use of amylases in baking is antistaling, i.e., improving the fresh keeping of baked goods. Staling is a highly complex phenomenon [431-434], but it is generally accepted that retrogradation of amylopectin is the main contributor to bread staling. The original model for the role of starch was developed by SCHOCH in 1965 [431]. Upon cooling in the first hours after baking the initial crumb structure is set by amylose gelation, which creates a network in which the gelatinized starch granules are embedded. Recrystallization of amylopectin side chains leads to rigidification of the starch granules and to an overall strengthening of the crumb structure, measured as an increase in crumb firmness.

Different types of amylases used for antistaling provide different functionalities. First, a limited effect of fungal α-amylase is caused indirectly by volume increase. Secondly, endo-amylases weaken the amylose network, leading to reduced crumb firmness immediately after baking and during storage. This effect is more pronounced when endo-amylases with high thermostability are used. An example of the most effective type is the *B. amyloliquefaciens* α-amylase, which has been used in the USA since the 1950s [435]. This amylase survives baking and is probably active during cooling of the bread and during storage [436]. Consequently, only a very narrow window of optimal dosage exists, and even moderate overdosing can lead to keyholing of the bread and a gummy crumb structure with strongly reduced elasticity (resilience). As an alternative, intermediately thermostable amylases have been developed. They have lower risks of overdosing, but are also limited in their antistaling effects [436].

Compared to endo-amylases, exo-amylases such as the so-called maltogenic amylase (E.C. 3.2.1.133) offer clear improvements for antistaling applications. By shortening the amylopectin side chains and releasing maltooligosaccharides, they efficiently reduce the rate of amylopectin retrogradation without disadvantageous side effects caused by weakening of the amylose network [437]. In a model system the shortening of amylopectin side chains has been shown to reduce retrogradation substantially [438].

5.1.1.3. Xylanases

Xylanases (E.C. 3.2.1.8, β-1,4-D-xylan xylanohydrolase), traditionally named pentosanases or hemicellulases, catalyze the endo-hydrolysis of β-1,4-D-xylosidic linkages in xylans. In dough, xylanases reduce problems associated with variations in flour quality, improve dough handling and dough stability, and give improved crumb structure and bread volume [439], [440].

The target substrates in cereals used for baking are arabinoxylan (AX; Fig. 31), one of the major nonstarch polysaccharides in cereals, making up 60-70 % of wheat cell walls [441] and 2 – 3 % of cereal flour [440].

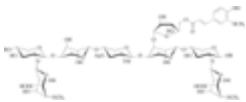

Figure 31. Model structure of arabinoxylan
[Full View]

Traditionally, the arabinoxylans have been divided into two fractions, a not extractable with water arabinoxylan fraction (WU-AX) and a water-extractable arabinoxylan fraction (WE-AX). Why some AX are water-extractable and others are not has been the topic of many studies. Early hypotheses explaining the differences were based on structure analysis of the two fractions, which showed a difference in the degree of substitution [442]. However, recent literature points more in the direction of both fractions having a similar degree of substitution [443], but the molecular weights of the two AX fractions differ, with WU-AX having the higher molecular weight. Furthermore, the concentrations of diferulates in the two fractions differ, and this may at least partly explain the difference in molecular weight and extractability of the two AX populations. The WU-AX has a much higher concentration of diferulates than the WE-AX [444]; this higher concentration of cross-linking between AX molecules will increase the molecular weight and probably make part of the AX water-unextractable.

Some of the best known characteristics of cereal AX are the water-holding capacity of WU-AX and the

ability of AXs to produce viscosity. GIRHAMMER [445] showed that WU-AX could absorb water corresponding to 10 times its own weight, and IZYDORCZYK and BILIADERIS [446] showed that AX has a very high intrinsic viscosity. These characteristics are probably the main factors determining the functionality of xylanases in the baking industry. In some baked products these phenomena are beneficial, but in others they are not.

Depending on the nature of the baked product, the flour AX may influence, in different ways, dough processing, dough characteristics, and the quality of the final product. This calls for the use of different, specific xylanases. In proofed and raised bread types, literature characterizes the WU-AX as detrimental for gluten strength and hence also for the stability of the dough [447], [448]. By influencing the dough stability, WU-AX also influences dough machinability and the quality of the final bread. Without xylanases the final products will have a more open crumb, and the volume will be smaller compared to dough in which the WU-AX fraction has been modified by a xylanase. Solubilizing the WU-AX into high molecular weight solubilized arabinoxylan (S-AX) by using a specific xylanase opens the possibility of converting WU-AX into a functional polymer. The high molecular weight S-AX will act as a hydrocolloid and increase the viscosity in the dough, giving a more stable dough system. The use of WU-AX-specific xylanases in such baked products will typically improve the dough stability, dough machinability, and improve crumb structure and volume (10 – 30 %) of the final product [439].

In dry cereal products such as crispbread, crackers, etc., AX influences the dough or batter characteristics through its viscosity-enhancing properties [449]. Furthermore, AX influences the evaporation of water during baking of these products, due to its high water-binding capacity, which prevents even and fast evaporation of water from the product. This may result in longer baking time, checking, and darker products. In contrast to the specificity of the xylanase mentioned above, the xylanase used in dry cereal products should be specific for WE-AX and S-AX, and thus decrease dough resistance and dough liquid viscosity.

5.1.1.4. Oxidoreductases

Oxidative enzymes have a beneficial effect on dough development and dough quality. Dough quality influences other quality parameters such as volume, texture, and crumb structure of the baked product [450]. During dough mixing and molding, disulfide linkages are formed and broken between gluten proteins. When this reaction takes place optimally, it leads to a good dough development, that is, a strong gluten network is formed. Chemical oxidants such as ascorbic acid and bromate are used in breadmaking worldwide to support and promote the formation of disulfide linkages. Today consumers in the Western world prefer food without added chemical ingredients. Due to regulatory problems with bromate in the USA and the E.C. countries, the wish to replace bromate exists. Hence, substitution with oxidative enzymes is a desirable alternative.

Glucose oxidase (E.C. 1.1.3.4) is usually obtained from *Aspergillus niger*. Glucose oxidase catalyzes the oxidation of glucose with consumption of oxygen to give gluconolactone and hydrogen peroxide. The generated hydrogen peroxide reacts with free thiol groups in the gluten with formation of disulfide linkages, which give increased gluten strength [451]. Glucose oxidase is known to give a dry and nonsticky dough, the latter characteristic being very useful in combination with xylanases, which may cause some stickiness when used alone [452].

Hexose oxidase (E.C. 1.1.3.5) is an enzyme originating from carrageenan seaweed, *Chondrus crispus*. This enzyme has broad substrate specificity, being able to oxidize glucose, galactose, maltose, etc. [453]. The enzyme oxidizes the endogenous substrates of the flour, mainly glucose and maltose, to the corresponding lactones with consumption of oxygen and formation of hydrogen peroxide. The hydrogen peroxide formed then oxidizes the thiol groups in the gluten proteins with formation of disulfide bonds [453].

Lipoxygenase (E.C. 1.13.11.12) enters bread production through the use of soybean flour or bean flour (France) that has not been heat-treated, i.e., is still enzymatically active. It improves the whitening of the breadcrumb. Lipoxygenase catalyses the oxidation of unsaturated fatty acids containing *cis*-*cis*-1,4-pentadiene groups. The resulting fatty acid hydroperoxide can react with carotenoids in the dough. According to CASEY [454] the molecular basis of bread improvement through the action of lipoxygenase is likely to involve the formation of intermolecular disulfide bridges between the high molecular weight gluten proteins [454].

Sulfhydryl oxidase (E.C. 1.8.3.2), isolated from *A. niger*, catalyzes the oxidation of free sulfhydryl groups to

disulfide groups in a variety of reduced thiol systems. A discussion continues as to whether sulfhydryl oxidase has only limited affinity for SH groups in high molecular weight substrates such as gluten [450]. SZALKUCKI wrote: “further testing is necessary to demonstrate the commercial utility of sulfhydryl oxidase in baking” [455].

Peroxidase (E.C. 1.11.1.7, horseradish peroxidase) is the general name for a group of enzymes with different substrate specificity. All peroxidases act on hydrogen peroxide as electron acceptor and oxidize a multitude of donor compounds [456].

During the peroxidase reaction a free radical is produced that reacts nonenzymatically with other compounds present in the dough [457]. It is well documented that certain peroxidases have a dough-improving effect [457], but different theories exist concerning the mechanism of function. The peroxidase might form diferulic acid bridges and thus lead to covalent cross-linking of arabinoxylans [458]. Coupling of arabinoxylans to proteins has also been demonstrated in model systems; this proceeds by oxidative cross-linking between ferulic acid and cysteine and/or tyrosine residues in proteins [457]. Finally, model studies also indicated the action of peroxidases in dough strengthening to be the result of protein – protein cross-linking through lysyl residues. Sulfhydryl groups may also be involved [457].

Polyphenol oxidase oxidizes diphenolic groups to quinones. Three types of enzymes are included in the group of polyphenol oxidases. Tyrosinase (monophenol monooxygenase, E.C. 1.14.18.1), diphenol oxidase or catechol oxidase (E.C. 1.10.3.1), and laccase (E.C. 1.10.3.2). Two mechanisms for the action of polyphenol oxidases in dough are postulated. One involves oxidation of ferulic acid esterified to arabinoxylan, which induces their cross-linking, and the other assumes cross-linking of gluten proteins by oxidation of tyrosyl groups or by conjugation between tyrosyl and thiol groups [459].

A number of enzymes that generate hydrogen peroxide by oxidation of various substrates with consumption of oxygen have been described as having dough-improving effects: glycerol oxidase [460], pyranose oxidase (E.C. 1.1.3.10) [461], L-amino acid oxidase (E.C. 1.4.3.2) [462], galactose oxidase (E.C. 1.1.3.9) [463].

5.1.1.5. Lipases

Lipases (E.C. 3.1.1.3) can be defined as carboxylesterases which catalyze the hydrolysis of long-chain acylglycerols to glycerol, free fatty acids, and mono- and di-glycerides. The substrate for lipases in wheat flour is the 2–3 % of endogenous wheat lipids. The composition of the free lipids (extractable with water-saturated butanol at room temperature) is shown in Figure 32.

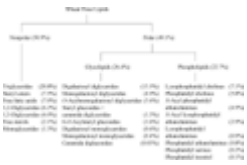

Figure 32. Composition of total wheat flour lipids, extracted with water saturated butanol [464] [Full View]

Lipids are probably the longest known and used bread improvers, whereas the use of lipase as a bread improver is relatively new. The understanding of lipase functionality is continuously being studied, but is mostly based on knowledge of the effects of lipids in dough and baking. The desirable effects obtained by lipases are improved dough stability and better dough handling properties, and improved stability of the gas cells in dough, yielding a nice, homogenous crumb structure and increased bread volume.

Lipase modifies flour lipids or added lipids, and the beneficial effect of an added lipase is dependent on the effect of the hydrolysis products [465]. Some of these hydrolysis products are quite surface-active or more surface-active than the unmodified lipid and therefore better stabilize the gas cells in the dough. This again leads to better crumb structure, increased dough stability, and increased volume [466]. Among the potential reaction products, monoglycerides are well-known crumb softeners in baking, as they form a complex with starch and thus decrease retrogradation [452].

Phospholipases are a class of four enzymes hydrolyzing specific bonds in phospholipids (see Fig. 33): phospholipase A1 (E.C. 3.1.1.32), phospholipase A2 (E.C. 3.1.1.4), phospholipase C (E.C. 3.1.4.3), and

phospholipase D (E.C. 3.1.4.4).

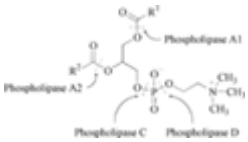

Figure 33. Phospholipase specificity (modified from A. L. Lehninger Biochemistry, Worth Publishers, Inc. 1970, p. 199, fig 10-4.)
[Full View]

The release of one fatty acyl chain from phosphatidylcholine yields 1-acylglycerophosphorylcholine, more commonly known as lysolecithin, which is an excellent emulsifier with a very positive effect in baking [467].

5.1.1.6. Proteases

Proteases (E.C. 3.4.X.X) catalyze the hydrolysis of peptide bonds in proteins [468]. Depending on the protease used, this modification may lead to a minor modification that gives more extensible dough [469], [470] or to an extensive hydrolysis of the gluten structure with loss of the viscoelastic properties [470]. Apart from the impact on the physical and rheological properties of the dough, some proteases are known to influence the flavor and color of the final product as well.

As proteins are a highly heterogeneous population of molecules, due to the 20 amino from which they are composed, the population of enzymes capable of hydrolyzing proteins is also very diverse [471]. This population of enzymes is divided into two categories: the peptidases and the polypeptidases (synonymous with proteases). The peptidases are exo-acting enzymes that remove one amino acid or a peptide from the C- or N-terminus of the protein, and the proteases are endo-acting

Wheat flour is unique in its ability to form a viscoelastic gluten matrix when mixed with water [472]. This gluten matrix can retain gas in the dough, giving wheat flour its unique baking performance. The quality and quantity of protein are therefore major quality parameters for wheat flour [424], [473].

In normal raised bread products too strong a flour may prohibit a good volume, as proper expansion of the dough will not be possible. Application of a moderate specific protease (typically a fungal protease) modifies the gluten, giving a better extensibility, faster development, and thereby better structure and volume of the final product [470]. The same functionality can be obtained by using chemical reducing agents, such as L-cysteine or sodium metabisulfite (Fig. 34). However, the biological, and labeling-free enzyme solution is often preferred.

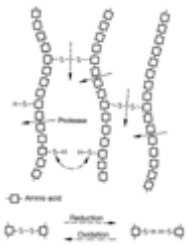

Figure 34. Gluten protein structure and modification by proteolytic hydrolysis or chemical reduction, respectively [After [425] (H. Uhlig: “Enzymes in flour processing and baking”, in: *Industrial Enzymes and their Application*. John Wiley & Sons, Inc. (1998). fig 5.36, pp 284)]
[Full View]

In cracker, biscuit, wafer, and pizza dough weaker flour having less viscoelastic properties is preferred [474]. A strong gluten network would prohibit the right shape and texture of the final product. Applying less specific proteases (typically a *Bacillus* protease) to these types of products allows the manufacturer to be more independent of flour quality. It is possible to obtain the desired dough rheological properties by using most flours in combination with protease. Again, the same solution is obtainable with chemical reducing agents, but the demand for biological and labeling-free solutions favor enzyme solutions.

Application of peptidases to bread might influence the flavor and crust color of the product [470]. The products of the peptidase, amino acids or peptides, influence the flavor. The amino acids are intermediate products in the production of aromatic components, and peptides are potential oxidants, taste enhancers, sweeteners, and bitter agents [475]. Furthermore, amino acids released from gluten proteins can react with

sugar through the Maillard reaction, thereby contributing to flavor and crust color.

5.1.1.7. Transglutaminase

Transglutaminase (protein-glutamine ϒ-glutamyltransferase, E.C. 2.3.2.13) introduces covalent cross-links in proteins by forming a bond between lysine and glutamine residues. Based on its ability to cross-link gluten proteins it can be used as a dough strengthener for bread and other baked products [476]. As further advantages reduction of the needed work input as well as increased water binding have been reported [477].

5.1.2. Enzymes in Fruit Juice Production and Fruit Processing

5.1.2.1. Introduction

The first application of enzymes in the fruit juice industry was the use of pectinases for apple juice clarification in the 1930s [478]. The fast clarification of juice after breakdown of pectin by pectinases and the decrease in juice viscosity resulted in a shorter process and greatly improved the quality of industrial apple juice. Later, pectinases were applied to depectinization of red-berry juice. The use of pectinases and amylases to degrade apple pectin and starch during the hot clarification stage prevents post-bottling haze formation, and thus concentration of the juice by a factor of six has become possible. This results in a smaller storage volume, cheaper transportation, and better concentrate stability without spoilage. Treatment of apple pulp with pectinases and hemicellulases was introduced later: by lowering the viscosity of the pulp, the press capacity and the yield were significantly improved. Depectinization of juice with pectinases with high arabanase activity improved product quality further by preventing araban haze formation after concentration of the juice. In the production of clear and stable red-berry juice or concentrate, hot maceration with enzymes ensures a higher juice yield and a better color extraction [479]. Nowadays, enzyme suppliers provide fruit juice producers with tailor-made enzyme preparations optimally blended on the basis of fruit composition for improved quality and stability of finished products, together with shorter process duration and larger plant capacity. In association with new equipment and processing technologies, industrial enzymes allow processors to add value to raw material for food and feed and to reduce waste quantity. Industrial processes are very diverse and numerous. Here the most common are described, but differences still exist depending on the company and plant.

5.1.2.2. Biochemistry of Fruit Cell Walls

The pulp of fruits and vegetables is composed of cells. They are surrounded by a cell wall, which resists internal pressure and external shock (Fig. 35). Polysaccharides constitute 90 – 100 % of the structural polymers of walls of growing plant cells, known as primary cell walls. Secondary cell walls develop from primary cell walls during cell growth. Cell wall composition depends on fruit species and evolves in dependence on agronomic and climatic conditions, fruit ripeness, the type and duration of storage of the fruit (cell growth and cell senescence). The composition of plant cell walls has been widely studied, and numerous models of the three-dimensional structure have been proposed [480], [481]. Three major independent domains are distinguished: the xyloglucan network, the pectin matrix, and the structural proteins. The cellulose – xyloglucan network is embedded in the pectin matrix. Pectin is the major structural polysaccharide component of fruit lamellas and cell walls. Three pectic polysaccharides are present in all primary cell walls: homogalacturonan and rhamnogalacturonans I and II [482] (Fig. 36). Recent models divide pectin into so-called smooth regions of unbranched homogalacturonan (60 – 90 %) and hairy regions of highly branched rhamnogalacturonan I (10 – 40 %) [484], [485]. Homogalacturonan is a homopolymer of (1→4)- α-D-galactosyluronic acid residues, capable of forming gels. Carboxyl groups of the galactosyluronic acid residues of primary cell wall homogalacturonan can be methyl-esterified at the C-6 position and acetyl-esterified at the C-2 or C-3 position. Helical chains of homogalacturonan that are less than 50 % methyl-esterified can form a gel-like structure and condense by cross-linking with calcium ions, which are present in primary cell walls. Degree of methylation, molecular weight, and pectin content are specific to fruit species (Table 10).

Table 10. Fruit pectin composition

Fruit	Pectin, wt %	Methylation wt %
Apple	0.5 – 1.6	80 – 92
Blackcurrant	1.0 – 1.2	50 – 80
Grape	0.1 – 0.4	50 – 65
Orange peel	3.5 – 5.5	65
Pear	0.7 – 0.9	50 – 70
Pineapple	0.04 – 0.1	22 – 40
Strawberry	0.5 – 0.7	20 – 60

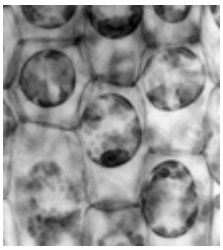

Figure 35. Fruit cells (fc) and middle lamellae (ml) connecting the cell walls
[Full View]

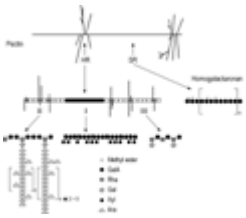

Figure 36. Model of pectin smooth regions (SR) and hairy regions (HR) model [483]. Ara = arabinose, Gal = galactose, GalA = galacturonic acid, Rha = rhamnose, Xyl = xylose.
[Full View]

These characteristics must be taken into consideration in choosing the right pectinase balance for best fruit processing. During ripening, the protopectin in primary cell wall is slowly transformed into soluble pectin by the action of endogenous pectinases present in the fruit. The decrease in molecular weight is due to pectin depolymerases, and the decrease in degree esterification mainly to pectin methylesterase. However, these activities are very low. Homogalacturonan contains chains of up to about 200 galacturonic acid units, 100 nm long, with some rhamnose. A xylogalacturonan subunit (XGalA), substituted with xylose at the C-3 position of the galacturonic acid residue, has been identified as part of the galacturonan backbone. It can be methyl-esterified. Rhamnogalacturonan I (RGI) has a backbone of up to 100 repeats of the disaccharide rhamnose-galactose. The side chain can vary in size from a single glycosyl residue to 50 or more glycosyl residues [482]. In general, about half of the rhamnosyl units of RGI have sidechains, but this can vary with cell type and physiological state.

Arabinans are mostly 5-linked arabinofuranosyl units forming helical chains, but arabinosyl units can be interconnected at each free O-2, O-3, and O-5 position to form a diverse group of branched arabinans [486]. The RGI of primary cell walls is branched with (1→5)- α-L-arabinan, (1→3)- or (1→2)- α-L-arabinosyl residues, and arabinogalactan (AG) type II with a (1→3),(1→6)- β-D-galactan backbone, with (1→3)- α-L-arabinosyl residues [487]. Other macromolecules such as cellulose, xyloglucan, and arabinogalactan proteins (AGP) are associated with the plasma membrane [488]. Rhamnogalacturonan II (RGII) is a low molecular weight (ca. 4.8 kdalton) complex polysaccharide with a backbone of nine (1→4)- α-D-galactosyluronic acid residues and four side chains attached to O-2 or O-3 of the backbone.

The side chains are composed of twelve different sugars [489] — apiose, 2-*O*-methyl-l-fucose, 2-*O*-methyl-D-xylose, aceric acid (3-*C*-carboxy-5-deoxy-l-xylose), Kdo (3-deoxy-D-manno-octulosonic acid), Dha (3-deoxy-D-lyxo-heptulosaric acid) — bound in more than 20 different linkages. The two main hemicelluloses of all primary cell walls are xyloglucan and arabinoxylan. Hemicelluloses bind tightly via hydrogen bonds to the surface of cellulose linking or cross-linking microfibrils to create a cellulose – hemicellulose network.

Interconnections with the pectic polysaccharides are of primary importance for the integrity of the pectin network. Cellulose is a (1→4)- β-D-glucan which accounts for about 20 – 30 % of the dry matter of most primary cell walls and is particularly abundant in secondary cell walls. In 1993-1994, VINCKEN and VORAGEN demonstrated that xyloglucan was a key structure of apple cell walls for the degradation of cell-wall-embedded cellulose (around 57 % of the apple cell-wall matrix) [490]. The cellulose – xyloglucan network determines the strength of the cell wall and is embedded in an independent pectin matrix, hemicelluloses, and proteins. In apples the xyloglucan fraction makes up about 24 % of the total amount of sugar. About 5 wt % of some primary cell walls is made up of the hydroxyprolin-rich structural glycoprotein extensin, which binds some polysaccharides together. Fractions of the different macromolecules in fruit are given in Table 11.

Table 11. Fruit cell wall composition in g/kg fresh matter

Fruit	EIR*	Pectin	Hemicellulose	Cellulose	Lignin	Protein	Total
Apple	20	272	169	349	2	76	868
Pear	15	281	148	267	69	82	847
Mango	25	408	91	236	27	127	889
Pineapple	13	163	267	210	85	94	819
Strawberry	12	411	66	232	11	255	975
Raspberry	20	168	89	177	73	277	784
Cherry	13	396	49	130	169	244	988
Papaya	26	364	165	124	4	127	784

*Ethanol-insoluble residue.

Starch is present in unripe apples in amyloplasts. This is the largest biological molecule, with a molecular weight of 10^5 – 10^9 dalton. Starch is composed of two components: amylose and amylopectin [491]. Amylose, the minor component, is a (1→4)- α-D-glucan and has a linear structure. Various degrees of polymerization have been ascribed to this fraction, with chain lengths in the range of 100 – 1000 glucose units. Amylopectin contains α-(1→6) and α-(1→4) glucose linkages. Amylopectin exhibits branching at the 1→6 position, and its degree of polymerization is far higher than that of amylose. The ratio amylose/amylopectin can vary in natural starches in the general range 1/3 to 1/4.

5.1.2.3. Cell-Wall-Degrading Enzymes

Pectinases. Progress in enzymology has been so fast that certain activities are not yet described in the International Enzyme Classification (E.C. no.). Many microorganisms produce enzymes that degrade fruit cell walls. Commercial pectinases for the fruit juice industry come from selected strains of *Aspergillus* sp. Enzymes are produced during fungal growth, purified, and concentrated. Pectinases are defined and classified on the basis of their action toward pectin (Fig. 37). Pectin lyase (PL, E.C. 4.2.2.10) is a pectin depolymerase of the endo type which has a great affinity for long, highly methylated chains and acts by β-elimination of methylated α-1,4 homogalacturonan with the formation of C4-C5 unsaturated oligo-uronides [492]. Pectin methylesterase (PME, E.C. 3.1.1.11) removes methoxyl groups from pectin, and at the same time decreases the affinity of PL for this substrate. This results in the formation of methanol and less highly methylated pectin. PME from *Aspergillus* has a strong affinity for highly methoxylated pectin such as apple pectin and acts according to a multichain mechanism [493]. Demethylation with PME generates free carboxylic acid groups and the pectin becomes negatively charged. Polygalacturonase (PG, E.C. 3.2.1.15) exists in two forms: endo-PG and exo-PG. Both types act only on pectin with a degree of esterification of less than 50 – 60 %. Endo-PG acts randomly on the α-1,4-polygalacturonic backbone and results in a pronounced decrease in viscosity. Exo-PG acts at the nonreducing end of the chain. Exo-PG releases small fragments from the chain and does not significantly reduce the viscosity.

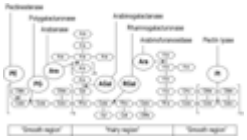

Figure 37. Fruit pectin and pectin-degrading enzymes. Ara = arabinose, Gal = galactose, GalA = galacturonic acid, OMe = methyl ester, OAc = ethyl ester, Rha = rhamnose, Xyl = xylose. [Full View]

Seven endo-PGs, two exo-PGs, and seven PL isoenzymes from *Aspergillus niger* have been described [494], [495]. Different enzymes acting on rhamnogalacturonan I were identified and purified from *Aspergillus* sp. [496]. RGase A was identified as a hydrolase that splits the α-D-GalA*p*-(1→2)- α-L-Rha*p* linkage of RGI, while RGase B appeared to be a lyase that splits the α-*l*-Rha*p*-(1→4)- α-D-GalAp linkage by β-elimination. Two novel enzymes were also identified: a rhamnogalacturonan rhamnohydrolase and a rhamnogalacturonan galacturonohydrolase. As an accessory enzyme for the RGases, rhamnogalacturonan acetyl esterase (RGAE) was also described. Although the structure of RGII substrate has been described, enzymes able to hydrolyse it are still unknown and have not been described yet (2003).

Arabanases are pectinases, since they remove arabinose covalently bound to the homogalacturonan backbone. Three enzymes have been described: an endo-arabinanase (α-1→5; ABFA) and two arabinofuranosidases, namely, exo-arabinofuranosidase A (α-1→2; α-1→3) (ABFA) and exo-arabinofuranosidase B (α-1→3; α-1→5; ABFB; E.C. 3.2.1.55). All three are produced by *Aspergillus niger* [497]. High activities are required for apple and pear processing. In general, pectinase activity and ratio of the different pectinases can vary in different commercial preparations.

Hemicellulases. Hemicellulases are enzymes that hydrolyze arabinogalactans, galactans, xyloglucans, and xylans. Arabinanases are classified as pectinases when they act on pectin side arabinans. They are also classified as hemicellulases when acting on arabinogalactans or arabinoxylans. *Aspergillus* sp. produces enzymes that hydrolyse arabino-(1→4)- β-D-galactans type I and arabino-(1→3)-(1→6)- β-D-galactans type II [487], [498], [499].

The exo-(1→3)- β-D-galactanase is able to release galactose and (1→6)- β-D-galactobiose. *Aspergillus* enzyme is able to bypass a branch point in a β-(1→3) backbone. The action of this enzyme is enhanced by the presence of ABFB. Xylanases hydrolyse the (1→4)- β-D-xylans in synergy with ABFs [500].

Amylases. In fruit juice industry, fungal acid amylase and amyloglucosidase are used to process fruits which contain starch. This is the case for unripe apples at harvest time. *Aspergillus niger* produces acid α-amylase and amyloglucosidase [501]. These exogenous enzymes are applied after starch gelatinization (occurring above 75 °C) for preventing post-bottling haze formation (starch retrogradation). Acid endoamylase (AA) acts on amylose and amylopectin. It produces dextrins, which are substrates for glucoamylase or amyloglucosidase (AG) α-1→4,1→6 exo-hydrolase, which release glucose from the non-reducing end of the chain.

5.1.2.4. Apple Processing

After oranges, apples are the most important raw material worldwide for the production of clear juice and clear concentrate. About 54 × 10^6 t/a of apples are gathered, and about 10 % are processed to produce mostly clear concentrate, and 2 % as cloudy juice, cider, and apple sauce [502]. The major producers of apple juice concentrate are located in the USA, Poland, Argentina, Italy, Chile, Germany, and China.

5.1.2.4.1. Apple Pulp Maceration

The current trend is to process apple juice from table varieties having defects or not sold for table consumption (Golden Delicious, Granny Smith, Jonagold, Red Delicious). Apples processed in the crop period are easily pressed with a relatively high yield. They are stored at low temperature in controlled atmosphere for several months and processed according to market demand. During storage, the insoluble protopectin is slowly transformed into soluble pectin by endogenous apple pectinases (protopectinase type), and starch is slowly degraded by endogenous apple amylases into glucose and consumed during post-harvest metabolism. Soluble pectin content can increase from 0.5 up to 5 g per kilogram of over-ripe apple [503]. Apples become difficult to press unless macerated with pectinases (Fig. 38). In the traditional process,

enzymes are used at two different stages (Fig. 39). Application of commercial pectinases from *Aspergillus* sp. in apple mash is necessary because activity of endogenous enzymes is too low to cause an immediate noticeable effect. Because apple pectin is highly methylated, commercial enzyme preparations must contain a high concentration of pectinlyase or pectin methylesterase in association with polygalacturonase and arabanase, together with side activities such as rhamnogalacturonase and xylogalacturonase. Rapidase Press (DSM), Pectinex UltraSP (Novozymes), and Rohapect MA+ (Röhm) pectinases sold for apple pulp maceration contain enzymes necessary to obtain a good pressability and high yield throughout the entire season. Commercial products contain more or less the enzyme activities described below, but with different ratios from non-GM and GM organisms.

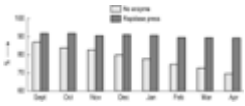

Figure 38. Effect of pectinases on percentage apple juice yield during the processing season
[Full View]

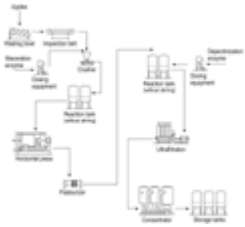

Figure 39. Production of apple concentrate
[Full View]

Pulp enzyming results in a fast decrease in pulp viscosity, a large volume of free-run juice, and fast pressing. The yield is over 90 % with hydraulic press and pomace leaching, compared to 75 – 80 % maximum without enzyme treatment.

5.1.2.4.2. Apple Juice Depectinization

Pectin is the main cause of juice turbidity. One liter of juice with 13 % dry matter can contain 2 – 5 g of pectin after pressing, depending on the ripeness of the fruit. Pectinase dosage is determined in the laboratory by performing the alcohol test. Use of acidified ethanol (2 volumes of ethanol containing 0.5 % HCl + 1 volume of apple juice) to precipitate residual pectin with a molecular weight of up to 3000 dalton. It is important to use acidified ethanol; otherwise, precipitation of organic acids or calcium pectate occurs because of the pH change after mixing ethanol with the juice, and thus a a false positive result is obtained. Pectinlyase or pectin methylesterase plus polygalacturonase and arabanase are the most important enzymes. Arabinose represents 55 % of neutral sugars in hairy regions of apple pectin. Arabanase activity prevents haze formation after juice concentration. The juice clarification is carried out after enzymatic depectinization [504]. The first stage consists of destabilization of the cloud by PL, which has no visible effect but results in a strong decrease in the viscosity of the juice (Fig. 40). PL cuts pectin at random, and cutting 1 – 2 % of the linkages is enough to reduce apple juice viscosity by 50 %. PG becomes active only after the action of PME. Because of its high molecular weight, PG cannot hydrolyse pectin with a degree of methylation greater than 50 – 60 % due to steric hindrance. At this stage, PL is no longer active and pectin hydrolysis is due to the system PME/PG. The second stage is cloud flocculation. The sedimentation necessary for the clarification of apple juice occurs only after enzymatic degradation of pectin and starch. The cloud is composed of proteins that are positively charged at the pH of the juice (3.5 – 4.0) since their isoelectric point lies between pH 4.0 and 5.0. These proteins are bound to hemicelluloses that are surrounded by pectin as a negatively charged protective colloid layer. Pectinases partially hydrolyze the pectin gel, and thus results in the electrostatic aggregation of oppositely charged particles (positive proteins and negative tannins and pectin), flocculation of the cloud, and then clarification of the juice [504]. The optimal pH for this mechanism is 3.6. The alcohol test shows whether juice depectinization is complete.

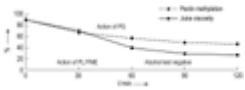

Figure 40. Apple juice depectinization
[Full View]

At the beginning of the processing season, unripe apples contain 5 – 7 g of starch per liter of juice. At this point, the iodine test gives a dark blue color and starch can be precipitated with iodine. Starch is present as granules 2 – 13 µm in size, composed of 30 % amylose and 70 % amylopectin chains. The latter can fix 20 wt % iodine. Apples contain endogenous amylases, but the process is too short and activities are too low to degrade the starch. When the juice is heated to 75 – 80 °C, starch changes from an insoluble form to a soluble form (gelatinization with water). If no amylase is added, these molecules can subsequently undergo retrogradation and form large aggregates that are difficult to hydrolyze and clog filters. The addition of fungal amylase and amyloglucosidase results in the hydrolysis of soluble starch into glucose monomers. Retrogradation and post-bottling haze formation are prevented, and clarification and filterability are improved. Preparations such as Hazyme DCL (DSM) and AMG (Novozymes) contain α-amylase and amyloglucosidase from *Aspergillus niger*. The dosage is determined by the iodine test. The third stage is clarification with finings and sheet filtration, which is being replaced by micro- or ultrafiltration (UF). Inorganic membranes have cut-offs from 20 000 to 500 000 dalton. Insoluble particles and soluble undegraded molecules such as parts of RGI, RGII, proteins including enzymes, and polyphenols, remain to a greater or lesser extent in the retentate, depending on the cut-off of the filtration membrane. Recently, it was found that undegraded RGI, RGII, and dextrins can foul UF membranes. Rapidase UF (DSM) and Novoferm 43 (Novozymes) contain rhamnogalacturonases and other side activities to improve ultrafiltration flow rate.

In conclusion, apple juice is easily produced after pulp maceration and juice depectinization with enzymes, with possible concentration of up to 70° brix (percentage by weight of soluble solids in a syrup at 68 °F), without risk of gel or haze formation. A few producers make “natural” or cloudy apple juice, e.g., in Germany and the USA. Until now, they did not add enzymes, because the existing commercial pectinases induce fast clarification of the juice. DSM sells a purified pectin methylesterase without pectin depolymerase activity. It can be used in the cloudy-juice process at the maceration stage to improve the yield, and also in the French cider process with calcium for defecation with flotation techniques.

5.1.2.5. Red-Berry Processing

The production of clear juice and concentrate from blackcurrant, raspberry, or strawberry requires enzymatic maceration and depectinization [505], [506]. Clarification, filtration, and concentration are difficult because these juices have high pectin content (typical content of residual low molecular weight pectin of 7 g L^{-1} compared to 0.5 g L^{-1} in apple juice). It is assumed that pectin hairy regions remain as soluble colloid in the juice and hemicelluloses tend to bind to phenolics and proteins during processing and storage. The result is the formation of irreversibly linked brown complexes that enzymes can no longer break down. An additional problem is related to the frequent contamination of red berries, mainly strawberry and raspberry, with *Botrytis cinerea*. This parasitic fungus, growing on rotten berries, secretes a β-1,3-1,6-linked glucan into the berries with a molecular weight of ca. 10^6 dalton [507]. This gum reduces the filterability and the clarity of the juice. It is possible to hydrolyze this glucan with β-glucanases with Glucanex (Novozymes).

Single-Stage Red-Berries Process. The single-stage process consists of simultaneous blackcurrant or blackberry maceration and depectinization (Fig. 41). Pectinases are used to improve juice and color extraction while retaining the organoleptic properties of the fruit. However, the extracted color is sometimes partly destabilized by anthocyanases (side activities of pectinases) or by oxidation. Oxidation can be chemical or enzymatic, due to the endogenous polyphenol oxidase (PPO) of the fruit, and is catalyzed by metal ions. It is therefore recommended that the pulp be heated to 90 °C to inhibit fruit oxidases prior to maceration with enzymes. Some red berries are very acidid (pH 2.6 – 2.8) and have high contents of phenolics and anthocyanins, which are inhibitors of pectinases. Hence, red-berry processing requires commercial pectinases that are especially stable under these conditions.

Figure 41. Production of blackcurrant concentrate
[Full View]

Two-Stage Red-Berries Process. The two-stage process consists of enzymatic maceration of fruit pulp, followed by a second addition of enzyme for juice depectinization at low temperature. Raspberries or gooseberries are heated to 90 °C for at least two minutes to increase color extraction and to destroy fruit polyphenol oxidase. The pulp is then cooled to 20 – 25 °C for enzymatic maceration. After pressing, the extracted juice is depectinized with pectinases. The low processing temperature after heating prevents aroma losses, and high-quality juices and concentrates can be produced. In the case of strawberries, it is appropriate to heat the pulp because it would create an unpressable purée, aromas loss, and juice browning. Hence, strawberries are processed at ambient temperature [506].

5.1.2.6. Tropical Fruit and Citrus Processing

Tropical Fruit. Tropical fruit is mainly processed to purée and stored before further processing to cloudy or clear juice. Fruit purée, cloudy or clear juice from apricot, peach, kiwi, mango, guava, papaya, and banana are often processed without enzymes. The main problem is viscosity, which can be decreased with pectinases [502]. Pectinases and amylases can be later used for clear juice production.

Citrus Fruits. The development of frozen concentrated orange juice started in 1940. Orange juice is the most consumed fruit juice in the world. Citrus fruit processing includes exploitation of all byproducts, among which pectin and essential oils are the most significant (Fig. 42). Although in certain countries it is not permitted to use enzymes in the production of premium orange juice, they can be used in other applications. Enzymes can increase the yield of solids recovery during pulp washing, facilitate the production of highly concentrated citrus base, improve recovery of essential oil from peel, debitter juice, and clarify lemon juice [502]. An example of citrus processing using enzymes is fruit peeling. The whole orange or grapefruit is treated with pectinases for digestion of albedo (the white, inside part of the peel), which binds flavedo (the orange or yellow, outside part of the peel). The fruit peel is scored and whole fruits are treated with a 2 % pectinase solution by vacuum infusion technology. After vacuum break, fruits are maintained at 40 °C for 15 – 60 min for albedo digestion inside the fruit. Peeled fruits are then rinsed, cooled, and packed.

Figure 42. Citrus processing
[Full View]

Another example is the production of clear lemon concentrate. Cloudy lemon juice coming from the extractor contains particles composed of pectin and proteins remaining in suspension and binding citral aroma (predominant lemon flavor). Clarification of lemon juice was achieved in the past by addition of large amounts of bentonite or sulfur dioxide. Today, clarification can be carried out with the very acidic pectinase Clarex Citrus 8XL (DSM), which degrades the pectin part of the cloud. Enzymes are added to the juice in amounts of 10 g hL^{-1} at 8 – 10 °C to avoid juice oxidation. The insoluble solids are removed by filtration or ultrafiltration. The brilliant lemon juice can be concentrated at 65° brix.

5.1.2.7. Conclusion

Nowadays, enzyme producers provide a wide range of pectinases for processing of fruit to give various products such as puree, cloudy juice, clear juice, and concentrate. The evolution of technology and processes including enzymes allow processors to obtain higher juice yield (productivity) together with higher quality of finished products. The use of specific pectinases adapted to the fruit process improves the shelf life of juices and concentrates (stability of color and freedom from turbidity). Apart from juice processing, the wide range and the high specificity of commercial enzymes open the way to new processes and new types of fruit-derived products. The trend is to process fruit under milder and more strictly controlled conditions to obtain new fruit products with sensory characteristics closer to those of fresh fruit.

However, because processing technology evolves faster than regulation, it is necessary to define food standard values and process references (equipment, process stages, type of enzymes . . .). In Europe, a Code of Practice has been developed to maintain quality and authenticity standards [508]. Members of

Association of the Industry of Juices and Nectars from fruits and vegetables (AIJN) in parallel with the Association of Microbial Food Enzyme Producers (AMFEP) have established references for fruit juice composition and enzyme specifications, in line with commercial standards and regulations of different countries.

5.1.3. Enzymes in Brewing

5.1.3.1. Introduction

Brewing is an old and traditional process that uses a variety of raw materials; in centuries past, many brewing processes have evolved, and they show the versatility and craftmanship of historic brewmasters. By far the most important brewing style is the “lager” type, characterized by the use of very pure water, low-temperature fermentation, and a long maturation period. Even in this relatively young beer style, many variations have developed; however, the variety of temperature programs during mashing, types of filtration, etc. are treated elsewhere (see → Beer). The common denominator in the brewing of beer is that part of the starch of a cereal is converted into alcohol. In the first stage enzymes hydrolyze starch into fermentable sugars. In a second process stage, these sugars are converted to alcohol and carbon dioxide by yeast. (Not all beers use yeast for fermentation; e.g., in the case of gueuze beers and weissen, bacteria are involved.)

The traditional source of enzymes used for the conversion of cereals into beer is malted grain, and malted barley is a key ingredient in brewing. Malting depends on many factors, including the barley variety and quality, the malt kilning regime and the skill of the malter. The extent and success of the malting process depends on the development, distribution, and survival of indigenous barley enzymes, and in the case of the notoriously heat labile malt enzymes, quality varies from season to season and from region to region.

During the malting of barley a series of enzymes — proteases, amylases, glucanases, and cellulases — developes in the malt. Enzymes such as β-amylase, exo-peptidase, and carboxy-peptidase are present in the starchy endosperm of the barley, and are activated during malting. In the aleurone layer of the barley, β-glucanase, endo-proteases, α-amylase and pentosanases are formed during malting. The formation and activation of the enzymes is promoted by increasing moisture and oxygen concentrations during the steeping process, and is then fixed by the kilning of the green malt. At this stage each enzyme with its individual properties is present at its individual concentration.

Detailed steps of the brewing process are never universal due to their wide variety in both starting material, brewing style, and technical preferences and possibilities. In this chapter a simplified brewing scheme is adopted (Fig. 43):

- Milling: reducing the size of the dry malt
- Mashing: adding water to the malt; controlled rise of temperature to allow enzymes to degrade proteins and carbohydrates
- Brewing: boiling the mash
- Filtration:separate spent grain and most (lautering)
- Cooling of the most
- Fermentation: convert sugars into alcohol and carbon dioxide
- Filtration: separate yeast from bright beer
- Maturation: settling of insolubles, reduction of vicinal diketones
- Packaging: in casks, kegs, bottles, cans
- Pasteurization: improving microbiological stability

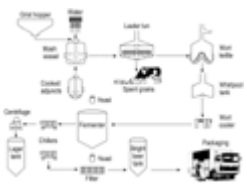

Figure 43. Schematic of the brewing process
[Full View]

Of the steps in this schemes, four are directly affected by enzymes (mashing, fermentation, maturation, and

packaging); indirectly, the several filtration steps are strongly affected, too. Enzymes have an impact on several qualities of the final product, such as appearance, taste, and economics.

Starting from malt, the concentration of each enzyme will change to some extent during storage, milling, and steeping, i.e., before the mashing step. Each enzymes set of intrinsic properties (temperature stabilities, substrate affinities, pH dependences, etc.) does not change, but the expression of enzyme activity is dependent on the reaction conditions (time, temperature, pH [510], salts [511], other enzymes, etc.), i.e., the mashing conditions.

Since brewing, and therefore mashing, is a highly optimized process [512], there are several undesirable consequences if too little of an enzyme activity is expressed in the mash; for example:

- The extract yield is too low
- Wort separation takes too long
- Fermentation is too slow
- Too little alcohol is produced
- The beer filtration rate is reduced
- Off-flavour of the beer (diacetyl)
- Inferior stability of the beer

Auxiliary enzymes are used to supplement endogenous malt enzymes in order to solve or prevent these problems (see Section Enzymes for Problem Prevention or Solving). Furthermore, enzymes can be used improve the brewing process (e.g., better adjunct liquefaction, shortening of the beer maturation time, production of beer from cheaper raw materials; see Section Enzymes for Process Improvement). Finally, by using new enzymes it has become feasible to produce “nontraditional” types of beer (low-calory “light beer”, IMO beer, “sparkling wine”; see Section Special Brewing Processes)

5.1.3.2. Enzymes in Malting and Mashing

During the malting process, enzymes migrate from the aleurone layer into the starchy endosperm. Cell-wall material is degraded by hemicellulases and β-glucanase, and breakdown of proteinaceous material by proteases enables amylases to act on the starch granules. Each malt enzyme has its own set of intrinsic properties The most notable property is the dependency of both activity and stability on a host of physicochemical parameters [temperature, water quality (pH, salts), malt and adjunct concentration and composition, etc.]. The dependency is different for each enzyme. Figure 44a shows enzyme initial activity and stability (half-life time) as functions of temperature at two pH values. The action of an enzyme is determined by its capability to convert a substrate into a product (starch to maltodextrins, or protein to oligopeptides), that is, the integrated enzyme action over a period of time during which the enzyme simultaneously loses its enzymic function due to (e.g., thermal) instability. As a consequence, somewhere between temperatures too low for the enzyme to be highly active, and temperatures too high for the enzyme to survive, an optimum for the enzymic effect will occur. It is dependent on many physicochemical parameters, for example, pH, as illustrated by Figure 44b, or time (Figure 44c). Clearly, the optimum temperature is a derived variable and not an intrinsic property of an enzyme. In the end, it is the optimum overall process that will prevail over the fate of a single enzyme.

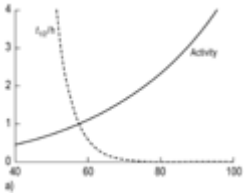

Figure 44. Enzyme activity as a function of various parameters
[Full View]

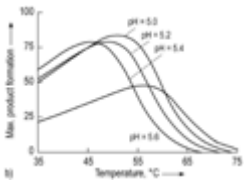

Figure 44. (*continued*) Enzyme activity as a function of various parameters
[Full View]

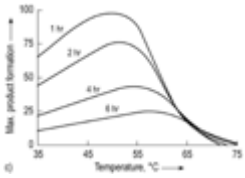

Figure 44. (*continued*) Enzyme activity as a function of various parameters
[Full View]

During mashing the situation is complicated, since some substrate molecules are formed by the action of one enzyme, and competed for by another enzyme, and these figures are based on a highly simplified model.

Figure 44d shows such a simulation of product formation with time by a typical enzyme at constant temperature, and Figure 44e over a temperature course typically used in a double-decoction mashing procedure. These graphs illustrate how conditions (e.g., temperature) for this enzyme might not completely ideal during mashing, and that a shorter mashing time would optimize the economics of this enzyme.

As a further illustration, Figure 44f and Figure 44g show that for another typical enzyme, a longer mashing period at lower temperature would be better suited.

Since many enzymes act simultaneously, and the results [e.g., high FAN (free alpha amino nitrogen) vs stable foam] often contradict on the molecular scale, the mashing process is a balancing act between obtaining enough activity from one enzyme without destroying another before the goal of its action is achieved [513]. The simulation model in Figure 44e applies to a highly simplified case. In contrast to the simulation, the composition of the solution changes rapidly during real mashing, one enzyme uses the product of another's action as its substrate, proteases also destroy enzymes, and temperature and viscosity change. Even when using identical malt and mashing conditions, relatively small differences between brewing water qualities of two regions can already give rise to substantial differences in mashing results.

Although the mashing process may be too complicated to translate into a kinetic model, brew masters achieve excellent mashing results by skillfully balancing and adapting the many process parameters [514]. In mashing, a premeasured amount of grain is infused with a specific volume of strike water at a predetermined temperature, and the mash is stepped through a series of so-called temperature rests. It is the method used to reach the temperature of these rests which distinguishes the infusion from the decoction mashing methods, and for simplification each rest is given a name according to its major designated effect (Table 12).

Table 12. Rest periods in mashing

Temperature	Enzyme	Rest period	Duration
30 – 40 °C	phytase	acid rest	10 – 30 min
40 – 50 °C	protease	protein rest	20 – 40 min
60 – 66 °C	β-amylase	conversion rest	40 – 70 min
68 – 70 °C	α-amylase	conversion rest	20 – 40 min

The acid rest (see Section Improved Mashing Processes) is rarely employed and has been replaced by the use of brewing salts. The protein rest is used primarily with larger amounts of incompletely modified malts to produce amino acids (as nutrients for the yeast) and to break down other proteins (e.g., those that cause chill haze) without destroying the proteins that provide the beer with a good frothy head.

In some cases it can be economically or technically feasible to use exogenous enzyme activities to assist the endogenous malt enzymes, e.g., for the production of a special beer, or to cope with the effects of poor malt modification. A new balance has to be established between parameters such as time and temperature on the

one hand, and the combined effects of both endogenous and exogenous enzymes on the other. The combination of different enzymes can result in a more robust, faster, or more economical over-all production process. If malt has been replaced to a large extent by brewing adjuncts such as unmalted barley, corn grits, rice, or sugar syrups, the level of endogenous enzymes will be proportionally reduced, and the additional exogenous enzyme activities become an intrinsic part of the production process.

Sources of exogenous enzymes may be vegetal (papain), fungal (glucanase), or bacterial (protease). Industrial producers of enzymes have been using many sources from as early as the 1960s.

The wide variety of exogenous enzymes range from close-to-malt enzymes to enzymes originating from extromophile bacteria. With the use of bacterial enzymes, the latest achievements of modern biotechnology become available to the brew master. Enzymes hitherto unavailable are now economically feasible for use on an industrial scale, with only personal or commercial sentiments standing between the brew master and the deployment of enzymes from GM organisms.

A traditional all-malt lager brewing process is used as starting point where possible, and the next sections address three areas for using exogeneous enzymes:

1. Solving or preventing problems occurring during the brewing process (Section Enzymes for Problem Prevention or Solving),
2. Improving the brewing process (Section Enzymes for Process Improvement),
3. Enabling a nontraditional brewing process (Section Special Brewing Processes).

5.1.3.3. Enzymes for Problem Prevention or Solving

5.1.3.3.1. Bacterial α-Amylase in Mashing

The most important change brought about in mashing is the conversion of insoluble starch molecules into soluble, fermentable sugars and unfermentable dextrins. The principal enzymes responsible for starch conversion are α- and β-amylase. α-Amylase (E.C. 3.2.1.1) rapidly splits insoluble and soluble starch into many shorter chains that can be attacked by β-amylase (E.C. 3.2.1.2). In all-malt brewing, a maximum final mash temperature of 78 °C results in the best separation of wort from spent grain. By adding a bacterial α-amylase, the temperature can be increased to 85 °C in the final mashing phase. The exogenous enzyme extends the degradation of remaining starch into soluble oligosaccharides, continues to do so in the lauter tun, and thus prevents an increase in viscosity due to gelatinization of starch during wort separation. Adding thermostable bacterial α-amylase just before the mash is pumped into the lauter tun degrades any remaining starch. As the mash in the lauter tun is at a higher temperature, it is thinner and passes through the filter medium more quickly and smoothly.

Adding bacterial α-amylases becomes essential when a high fraction of adjuncts has diluted the concentration of endogenous enzymes in the mash tun. Since this dilution factor applies to all endogenous enzymes, additional enzyme dosage or enzyme cocktails containing, e.g., β-glucanase and protease in addition to amylase, are used.

If a thermostable bacterial α-amylase is used, brew masters often prefer separate addition of the β-glucanase and protease to add flexibility (compared to an enzyme cocktail with fixed composition) to adapt to their particular malt/adjunct situation.

5.1.3.3.2. Fungal α-Amylase in Fermentation

Slow fermentation can result from incomplete saccharification in the mashing process. If the problem is diagnosed in an early phase, direct addition of fungal α-amylase (E.C. 3.2.1.1) in the fermenter can correct the problem. The enzyme action at relative low temperatures lowers the attenuation limit (degree of attenuation is the percentage of fermentable carbohydrates in the wort extract), increases fermentability, and more alcohol is produced, resulting in a “drier” more attenuated beer.

5.1.3.3.3. β-Glucanase in Mashing

The β-glucanase (E.C. 3.2.1.4) from barley is very heat labile and survives for only a very short time at mashing temperatures. If not enough β-glucan is broken down, the remaining glucans will partly dissolve, bind water, and cause problems later in the process by increasing viscosity (and thus extending wort runoff

times) and by causing haze.

Optimizing the mashing process by varying temperature/time regimes often improves the performance of amylases and increases FAN values. If the new regime decreases β-glucanase performance, this may cause several related problems:

- Poor runoff,
- Poor extract recovery,
- Poor spent-grains drainage,
- Slow yeast sedimentation, poor centrifuge efficiency,
- Poor beer-filter performance, problems with haze.

β-Glucanases from, e.g., fungal origin, selected for improved thermostability under mashing conditions, result in a more robust and stable mashing process. They assist in the mashing of grists containing insufficiently modified or unevenly modified grains, aid runoff from the mash tun by reducing wort viscosity, and improve beer finability and filterability.

5.1.3.3.4. Cysteine Endopeptidases (Postfermentation)
If insufficient protein is removed by hot and cold precipitation (protein is needed throughout as, e.g., yeast nutrient, FAN, and for foam), polypeptides can cross-react with polyphenols during the final stages of beer production (conditioning, lagering) to cause an undesirable, so-called chill haze. Soluble proteolytic enzymes, such as papain (E.C. 3.4.22.2) and, to a lesser extent, bromelain (E.C. 3.4.22.32) and ficin (E.C. 3.4.22.3), are commonly used (mostly in conjunction with other stabilization agents such as poly(vinyl pyrrolidone) (PVP) or silica gels (e.g., Lucilite) before final filtration to improve the colloidal stability of the beer, and thus control chill haze in and increase the shelf-life of packaged beer. Papain, extracted from the fruit of *Carica papaya*, is effective in the temperature range of 30 – 45 °C and the pH range of 4.0 – 5.5. Added to conditioning tank, this a-specific endo-peptidase breaks down high molecular weight proteins that react with polyphenols before being destroyed during pasteurization.

The use of immobilized proteases has been studied, but has not been widely adopted.

5.1.3.3.5. Glucoamylase in Mashing
If the amount of dextrins remaining in the beer as nonfermentable sugars is too high, the addition of exogeneous glucoamylase (E.C. 3.2.1.3) to the mash tun results in cleavage of terminal α-1,6 glycosidic bonds of the oligosaccharides, and thus in an additional amount of fermentable glucose in the wort. Since glucoamylase only reacts with oligosaccharides of maximal 10 – 15 glucose units in length, this enzyme does not prevent starch hazes caused by high molecular weight amylose and amylopectin molecules.

5.1.3.4. Enzymes for Process Improvement
5.1.3.4.1. Adjunct Brewing
In adjunct brewing, malt is partly replaced by another starch source (e.g., corn, rice) for economical reasons or to produce a lighter taste. With malt as the sole enzyme source, the malt enzymes have a more difficult job since their relative concentration during mashing is lower. When higher percentages of adjunct are used a separate pre-mashing step with exogenous enzymes is introduced to enable a more predictable and simpler production process. As heat-stable amylases are much more stable than malt amylases, simpler liquefaction, shorter process times, and an overall increase in productivity can be achieved. Eliminating the malt from the adjunct cooker means less adjunct mash and thus more freedom in balancing volumes and temperatures in the mashing program.

Traditionally, the use of barley has been limited to 10 – 20 % of the grist when using high-quality malts. At higher contents of barley (> 30 %, or with low-quality malts) processing becomes more difficult, and the mash must be supplemented with extra enzyme activity (apart from α-amylase, some extra β-glucanase and endo-peptidase are needed) if the brewer uses unmalted barley while maintaining brewing performance. Other raw starch adjuncts are used to partially replace malt as the carbohydrate source. Adding thermostable bacterial α-amylase to the mash allows extension to the higher temperatures needed for liquefying starches from, e.g., rice or maize. Brewing systems that use raw starch or grits from rice, maize, or sorghum will have a separate cooking stage for these materials, preferably at elevated temperatures of up to 108 °C (jet cooking). Malt amylases are not suitable for this, and thermostable bacterial α-amylase, or a protein-

engineered more robust version thereof, is needed. Pre-gelatinized adjuncts, such as micronized cereals added to the mash need (nonthermostable) bacterial α-amylase to ensure that no residual starch is present in the worts. The enzyme hydrolyzes both malt and adjunct starches, liberating soluble dextrins. This action is supplemental to the action of the natural malt α- and β-amylases.

The presence of approximately 200 ppm Ca^{2+} is important when using nonthermostable α-amylase, especially when hydrolysis occurs at higher temperatures. Enzyme inactivation occurs when the temperature is raised to ca. 100 °C for 1 – 20 min. For all practical purposes the enzyme will be destroyed in the brewing kettle during wort boiling.

Liquid adjuncts as carbohydrate source include cane and beet sugar syrups, and the corn based DE (see Section Industrial Use) syrups produced by the corn starch processing industry. “Brewer's syrup” (a maltose syrup produced from, e.g., corn, with a carbohydrate spectrum close to that of sweet wort) has become popular as adjunct in the UK, South Africa, and some Asian countries. The solubilization and partial hydrolysis of corn starch is performed outside of the brewery by starch processors using modern industrial enzymes such as thermostable (protein-engineered) bacterial α-amylase, pullulanase, and β-amylase extracted from malt or barley. By using different saccharification reaction conditions (time, temperature, enzyme), by blending or by introducing fungal α-amylase and glucoamylase, starch processors can nowadays meet any brewer's syrup specification, both in composition and in economics.

5.1.3.4.2. Improved Mashing Processes

Proteases. The endogenous endo- and exo-proteases are highly heat labile and act mainly in the malthouse. Carboxy-peptidases are a little less heat sensitive, and continue to operate for a while in the mash. Proteases (and also β-glucanase) are quickly destroyed in an infusion mash at 63 – 66° C. When using decoction mashing techniques and a reduced initial mash temperature, significant enzyme activity can occur in the early stages of mashing. Therefore, malts for use in decoction mashing systems need not be so well modified as malts in infusion mashes. During the so-called protein rest (30 min at 40 – 50 °C) protease reduces the overall length of the high molecular weight proteins — the cause of foam instability and haze — to low molecular weight proteins in the mash. Endo-proteases split high molecular weight proteins into simpler peptides by breaking the peptide bonds between amino acids. Endo-proteases are responsible for degrading insoluble globulins, and albumins already dissolved in the wort, into medium-sized polypeptides. The reduction in content of albumins and globulins is important for reducing the haze caused both by proteins and polyphenols (tannins derived from the malt husks and hops). As a rule, reducing the number of large protein molecules in beer makes it less prone to haze. The medium-sized proteins are not useful yeast nutrients, but are important for foam stability and thus head retention, as well as for body or palate fullness. Some brewers prefer to limit the duration of the protein rest to improve the foam quality of the beer. A wide variety of endopeptidases are readily available; historically, the proteases originating from *Bacillus amyloliquefaciens* have proved adequate for in assisting the available malt proteases.

Pullulanase. Endogenous β-amylase (1,4- α-D-glucan maltohydrolase, E.C. 3.2.1.2) is an exo-enzyme that cleaves external α-1,4-glucosidic links to form maltose molecules and β-limit dextrins. The latter contain α-1,6-glycosidic links and cannot be cleaved by either α- or β-amylase; β-limit dextrins remain in the wort throughout fermentation as nonfermentable sugars. The natural enzyme produced in the wort that can cleave this link is limit dextrinase (dextrin α-1,6-glucanohydrolase, E.C. 3.2.1.142), a heat-labile enzyme that is readily inactivated at mashing temperatures. The exogeneous enzyme pullulanase (pullulan 6-glucanohydrolase, E.C. 3.2.1.41) hydrolyses α-1,6 glycosidic linkages in branched polysaccharides, e.g., amylopectin. The enzyme requires at least two α-1,4 glucose units to be on either side of the α-1,6 link, and thus maltose is the main final reaction product. The activity and stability of exogenous pullulanase must be compatible with mashing conditions, and not all industrial enzymes used in the starch processing industry qualify. The tolerance for fermentation pH values and limited thermostability prevent carry-through risk for pullulanase, and limit its action to achieving a set degree attenuation.

Phytase. The traditional acid rest used to be responsible for reducing the initial mash pH for traditional decoction mashing of lager beers starting from under-modified malt. The enzyme phytase from barley malt is active at 30 – 53 °C and breaks down insoluble phytin (a complex organic phosphate containing both calcium and magnesium) to phytic acid. The phytase reaction releases hydrogen ions in the process and may be accelarated and extended by adding a more thermostable phytase of bacterial origin. Highly modified

malts contain too little endegenous phytase because of the high kilning temperatures and depend fully on exogeneous phytase to achieve phytase-induced acidification, whether during the acid rest or during the intitial mashing phase at higher temperature. However, the acidity of highly kilned malts is normally sufficient to lower the mash pH without an acid rest.

Diastase/ β-Amylase. The most obvious method to enrich the enzymatic activity of malt is to add an extract of malt enzymes. Composition (concentrations of various enzymes) may vary between the various products, dependent on malt type and extraction processes. For economical reasons, addition of malt or barley extracts is only used if no high-quality malt is available and the use of bacterial and fungal enzyme is undesired.

5.1.3.4.3. Shelf-Life Improvement

During fermentation, yeast takes up all dissolved oxygen, and in the following processing steps the atmosphere in tanks and equipment is pure carbon dioxide. Typically the oxygen concentration in beer is below 200 ppm. After packaging, the concentration of oxygen in beer may vary from less than 500 to 1000 ppm. Traces of entrapped oxygen (and polyphenols such as delphinidin acting as pro-oxidants), lead to formation of the volatile aldehydes responsible for stale flavor in beer. Antioxidants, such as sulfite, ascorbic acid, and catechin, can thus protect beer from becoming stale in the presence of oxygen [517].

Various antioxidants have been added to green beer to either remove oxygen or to negate its effects. Ascorbic acid (vitamin C) at 1.5 g/hL reduces oxidation haze and the effect, similar to reducing dissolved oxygen. The reduction of agents containing sulfur can reduce chill hazes. Sodium hyposulfite has some effect on chill haze when used in amounts of 20 ppm, and sodium metabisulfite and ascorbic acid (10 – 20 ppm each) have a synergistic action in protecting the activity of papain in beer during and after pasteurization. In contrast to the reducing agents, which are needed in equimolar amounts to the dissolved oxygen, enzymatic oxygen removal uses the small concentration of available glucose as the electron donor to scavenge oxygen. The removal of oxygen by glucose oxidase (β-D-glucose:oxygen 1-oxidoreductase, EC 1.1.3.4) plus catalase (H_2O_2 oxidoreductase, EC 1.11.1.6) is the sum of two reactions: glucose oxidase converts glucose and oxygen to gluconic acid and hydrogen peroxide, which is transformed by catalase into water and oxygen (net reaction: glucose +1/2 $O_2 \rightarrow$ gluconic acid).

In practice, a system with only enzymic deoxygenation is less efficient than a combination of enzymes and chemical reductants. The concentration of free glucose in beer may be too low for effective removal of oxygen, but it was demonstrated [518] that by adding only glucose oxidase and sulfite successfully suppressed the flavor deterioration of the beers. Another possibility is that the peroxide formed in the first reaction, and/or intermediates leading to oxygen formation in the second reaction, are reactive and responsible for oxidative deterioration of beer flavor [519].

5.1.3.4.4. Accelerated Maturation

Lagering involves secondary fermentation of remaining fermentable extract at a reduced rate at low temperatures and low yeast count. The low temperature promotes the settling of remaining yeast and precipitation of haze-forming material (protein/polyphenol complexes). The maturation period or diacetyl rest reinvigorates the yeast culture so that it metabolizes byproducts such as diacetyl and 2,3-pentanedione that were excreted early in the fermentation. During the maturation period 96 % of the diacetyl and 2,3-pentadione is used in biosynthesis (i.e., amino acid valine/leucine synthesis in particular) by active yeast and 4 % of α-acetolactate formed in beer is oxidized to diacetyl. The flavor threshold for diacetyl is 0.10 mg/L, and this small molecule causes an unpleasant buttery or butterscotch taste, considered a major off-flavor for lager-type beer.

Depending upon the yeast type, the physical environment, etc., this process takes 5 to 7 weeks with traditional lagering. Employing a diacetyl rest is particularly important when producing beers with high adjunct fraction due to the high levels of diacetyl produced [520]. It is also important in brewing lagers, as they do not have a heavy flavor impression.

In accelerated lagering the beer is fully attenuated, virtually free of yeast, and stored at higher temperatures to reduce the concentration of vicinal diketones, which are responsible for off-flavors in beer. Accelerated lagering reduces the time needed to achieve a product similar to beer brewed with a cold-lagering process to 7 to 14 d, with the same clarity and flavor stability [521]. Sometimes freshly fermenting wort is added to

diacetyl-laden beer in cold storage (kräusen) so that active yeast absorbs diacetyl.

With the exogenous enzyme α-acetolactate decarboxylase (ALDC, (*S*)-2-hydroxy-2-methyl-3-oxobutanoate carboxylyase, E.C. 4.1.1.5) the excess α-acetolactate can be directly converted into harmless butanediol, bypassing the formation of diacetyl.

By eliminating one of the major reasons for an extended lagering period, ALDC enables brewers to extend their peak lagering capacity.

5.1.3.4.5. Starch-Haze Removal
Fermentation problems, such as sticking fermentations or unacceptably low attenuation limits, are only noticed after mashing, when certain notable fermentation parameters are not met. Undegraded starch or high molecular weight carbohydrates that will neither fine nor ferment out, rearrange into insoluble complexes that cause beer with a starch or carbohydrate haze. Under these circumstances it is important to apply corrective action swiftly to avoid atypical flavor profiles. At the lower temperatures common in fermentation, fungal α-amylases are capable of rapidly hydrolyzing the interior α-1,4-glucosidic linkages of barley, malt, and cereal starch and the formation of maltose and a carbohydrate profile similar to that derived from indigenous malt amylases.

5.1.3.5. Special Brewing Processes
Low-calory beer (diet/light lagers) is based on the American lager style of brewing. Corn is the major adjunct, amounting to 50 – 65 % of the total grain bill, and processed with added enzymes, e.g., glucoamylase to break down nonfermentable carbohydrates and attenuation is higher than normal beers. *Dry beer*, *super dry beer*, and *sparkling wine* are highly carbonated fermentation products with almost complete conversion of all carbohydrates into alcohol and CO_2 assisted by bacterial and fungal enzymes. The main starch source in *sorghum beer* (bantu or kaffir beer, pombe, burukutu, pito), is unmalted sorghum, supplemented with maize. After acidification with lactic acid to lower the pH to 4, bacterial α-amylase may be added, and the cooker is heated to boiling for 90 – 120 min. After cooling to 60 – 62°C, malted sorghum and/or bacterial glucoamylase is added for a partial saccharification. After coarse filtration and cooling to 30 – 35 °C, yeast is added. The fermenting turbid mash is packed in open bottles, jerrycans, or vessels, ready for consumption after 16 – 24 h. Still experimental is the production of *IMO beer*. The consumption of isomaltooligosaccharides (IMO: glucose oligomers with α-1,6-glucosidic linkages) is claimed to stimulate the health-promoting activity of colonic bacteria of genus *Bifidobacterium*, as well as producing a mild sweet taste and low cariogenic properties. IMO-containing syrups are usually produced from starch by means of cooperative reactions, one of which is transglycosylation catalyzed by (immobilized [522]) enzymes of microbial origin, by conversion of high-maltose syrup into IMO-containing syrup. The transglycosylation product contains about 38 % of panose, 4 % of isomaltose, 28 % of glucose and 23 % of maltose, and the usage of IMO syrup in the brewery to replace maltose imparts functional properties to traditional food products with minimal changes in production technology and product taste. By introducing transglucosidase (EC 2.4.1.) during an enzyme assisted mashing process, it is also possible to produce IMO *in situ*.

High-gravity brewing involves worts of up to 18 °P and even higher. After fermentation and maturation, the beer is diluted with cold carbonated water to the designated gravity or to a prescribed alcohol concentration. Advantages associated with high-gravity brewing are that it results in beers that are more consistent (alcohol content, original gravity, etc.) and more physically stable because the compounds responsible for haze are more easily precipitated at the higher concentrations [523]. Handling more concentrated wort results in increased utilization of equipment and lower energy costs The disadvantages are problematic mashing, longer fermentation times, different flavor characteristics, and poorer hop utilization than normal-gravity fermentations [524]. Exogenous enzymes are used to assist in mashing (neutral protease, bacterial α-amylase) and fermentation (fungal α-amylase).

5.1.4. Enzymes in Dairy Applications
5.1.4.1. Introduction
Dairy products are among the classic examples of fermentation-derived foodstuffs, and their history of development goes back several millennia. Through the ages a huge variety of dairy products has been developed based on empirical experimentation. Profound understanding of biochemical, microbiological,

and physicochemical processes has mainly been achieved during the second half of the 1900s. Economical interest and scientific progress have led to the development of dairy food science as a special branch within food science, and it is the basis of control in manufacturing a huge variety of dairy products. Emphasis on understanding and controlling taste and texture development has led to a still growing assortment of cheeses and desserts. Enzymes originating from raw materials, microbial starter cultures, or other sources are prime tools in improving existing and creating novel dairy products.

5.1.4.2. Cheesemaking

5.1.4.2.1. Cheesemaking Process

A basic element in the process of cheesemaking is the clotting of the milk. Milk turns into a gel structure through the action of coagulating enzymes, briefly called coagulants. The casein proteins in milk lose their colloidal stability and aggregate to form a gel structure. After cutting of the gel, the liquid whey fraction, containing whey proteins, minerals, and lactose, separates from the casein chunks. The casein material, the curd, is collected in molds. After pressing and brining, the cheese is kept under conditioned storage for ripening. Storage time may last from weeks to months, even years, depending on the type of cheese.

5.1.4.2.2. Mechanism of Renneting

About 80 % of milk protein consists of casein, which is hydrophobic by nature. The bovine caseins may be subdivided in four species of phosphoproteins that exist, due to their poor solubility in water, in agglomerates. The four species αs_1-, αs_2, β- and κ-casein occur in relative molar concentrations of about 4:1:4:1.6. Caseins aggregate to form submicelles, which together form the casein micelle [525]. The outer surface of the casein micelle consists of submicelles containing a relatively high content of κ-casein molecules. Hydrophilic, negatively charged parts of the κ-casein molecules protrude from the periphery of the micelle and guarantee micelle stability due to electrostatic and entropic repulsion. The coagulating enzymes specifically split off a distinct part of these κ-casein hairs, hydrolyzing the Phe-Met bond (at amino acid 105 – 106) and thus provoke destabilization of the whole micelles, which aggregate with each other. The part of κ-casein split off is called the caseino macro peptide. Calcium ions facilitate casein micelle aggregation but do not affect the enzyme reaction. The enzyme reaction is very sensitive to pH but less to temperature. The aggregation rate is strongly affected by temperature between 25 and 35 °C. Renneting is thus described by the enzymatic reaction and by Casein Micelle Aggregation [526], [527].

5.1.4.2.3. Types of Coagulants

Commercially available coagulants used in the cheesemaking industry consist of animal rennet and microbial coagulants from non-GM and GM microorganisms. Animal rennet is obtained from the gastric mucosa of young mammals, usually of bovine origin. Animal rennet is a mixture of chymosin and pepsin (E.C. 3.4.23.1). The ratio of chymosin to pepsin depends on the age of the animal, being highest (about 95 %) for unweaned calves. The high ratio of chymosin to pepsin in the calf stomach is explained by the fact that chymosin does not hydrolyze immunoglobulins, which are present in the colostrum and needed to build up immunity, whereas pepsin does.

Due to the limited availability of calf rennet, substitutes from microbial sources have been sought, and nowadays microbial coagulants cover a considerable part of the demand for of coagulating enzymes worldwide. Microbial coagulants are produced by fermentation. Most widely applied are enzymes from *Rhizomucor miehei*, *Rhizomucor pussilus*, and *Cryphonectria parasitica*.

A third group of coagulants is made by genetically modified microorganisms. The bovine gene encoding for chymosin type B is cloned in the genome of fungi and yeasts such as *Aspergillus oryzae* and *Kluyveromyces lactis*. Chymosin type A cloned and expressed in *E. coli* is no longer commercially available. Several efforts have been made to clone microbial genes from *R. miehei* or *R. pussilus* in other microbial expression hosts, but, although technically feasible, none of these have reached commercial application.

In some minor cheese types plant proteases from cardoon and *Solanum dobium* are used.

Coagulants are commercially available in liquid, powdered, tabletted, and paste formulations of varying strengths. Animal rennets are made in standardized ratios of chymosin to pepsin.

5.1.4.2.4. Properties of Coagulating Enzymes

All coagulating enzymes are aspartic endoproteases. Two aspartic acid residues are located in the active center of the enzyme and play an essential role in the proteolytic mechanism. All coagulating enzymes are more or less inhibited by pepstatin, a microbial oligopeptide from strains of *Streptomyces* [528]. Coagulants are of comparable size and three-dimensional bilobal, structure and contain highly conserved regions. The amino acid homology between microbial coagulants is high, but they diverge extensively from mammalian ones [529]. The three-dimensional structures have been elucidated by crystallographic studies [530]. Chymosin and pepsin are secreted as inactive zymogens, whereas the microbial enzymes are not. In low-pH environments, as in the stomach, the propeptides are split off to activate the enzyme. Table 13 lists some biochemical characteristics of coagulating enzymes.

Table 13. Biochemical characteristics coagulating enzymes

Coagulant	Source	E. C. no.	M, kDa	IEP	Glycosylation	T_{opt}. °C (milk clotting)
Chymosin	bovine	3.4.23.4	35.7	4.98	no	40–44
Mucorpepsin	*R. miehei*	3.4.23.23	38	4.58	yes	58–62
Mucorpepsin	*R. pussilus*	3.4.23.23	30–39	4.41	yes	42–45
Endothiapepsin	*C. parasitica*	3.4.23.22	33. 8	4.89	no	42

Chymosin in animal rennet consists of types A and B, which are point-mutated at position 290, where chymosin A has Asp and chymosin B has Gly [531]. Chymosin A is said to have a higher specific activity for renneting but is less stable [532]. Due to its high specificity for cleaving the Phe-Met bond between amino acid residues 105 and 106 of κ-casein at the natural pH of cheese milk, chymosin is considered the standard renneting enzyme to which others are compared. At lower pH, aspartic proteases lose their specificity and aspecifically hydrolyze proteins randomly, even themselves. Apart from the coagulant from *C. parasitica*, which cleaves between amino acid residues 104 and 105 in κ-casein [529], the other coagulants also hydrolyze the Phe-Met bond of κ-casein. In general, microbial coagulants are more proteolytic than animal rennet. There are differences between the various coagulants in terms of the ratio of their milk clotting to their overall proteolytic activity, whereby a high ratio is important. The milk clotting to proteinase (also called caseinase) ratio (MC/PA) is used as a quality characteristic of coagulants [533]. These ratios vary with pH. At pH 6.5 chymosin derived through fermentation has the highest ratio, followed in order of decreasing ratio by calf rennet, bovine pepsin, and proteases from *R. pusillus*, *R. miehei*, and *C. parasitica*. The MC/PA ratio is directly related to cheese yield, one of the most important economical characteristics in the cheesemaking process. A high caseinase activity results in a relatively large amount of casein protein fragments that are washed away with the whey fraction and thus do not end up in the cheese curd. There is a huge, but also, conflicting amount of scientific literature describing comparison studies of cheese yield with different coagulants. In an elaborate study [534] the most relevant commercial coagulants were studied in a single comparable process (Table 14).

Table 14. Performance of commercial coagulants

Coagulant	Yield reduction, %
Calf rennet	0.0
Calf rennet/pepsin 1/1	0.09
Bovine rennet	0.14
Mucorpepsin (*R. pusillus*)	0.45
Mucorpepsin (*R. miehei*)	0.59
Endothiapepsin (*C. parasitica*)	1.11

Although the differences appear to be small, considering the huge amounts of milk processed, the economical relevance is considerable.

Heat lability is another important characteristic of coagulants [535]. For Emmental cheese, the cheesemaking process of which involves a cooking step at 50 °C to inactivate all residual coagulant in the cheese, mostly coagulant from *C. parasitica* is used because of its low heat stability. Any proteolytic activity of coagulants in whey, which is nowadays used in many food applications, should disappear following normal pasteurization (12 s, 72 °C). The following order of heat lability is known: porcine pepsin > bovine pepsin > *C. parasitica* > chymosin > *R. pusillus* > *R. miehei*. Commercial types of *R. miehei* coagulants have been chemically modified in order to lose stability and to meet application demands. The heat stability of all coagulants increases with decreasing pH.

5.1.4.2.5. Cheese Ripening

Cheese ripening [536] is mainly due to proteolytic breakdown of the casein protein. Endoproteases such as the endogenous milk enzyme plasmin and coagulating enzymes are responsible for the generation of casein polypeptide fragments. Plasmin is a trypsin-like alkaline endoprotease that occurs in milk also in its inactive plasminogen form. It is mainly associated with the casein micelle. Plasmin may become more or less active when plasmin- and plasminogen-activation inhibitors were separated with the whey fraction. Exoproteases, originating from the bacterial starter cultures, further break down part of these peptides to the level of free amino acids. These amino acids function as precursors for flavor compounds to be synthesized by the catabolic enzyme systems of the starter cultures. In this way the various types of starter cultures are responsible for the typical characteristics of cheese varieties. In this simplified description it is clear that coagulants play an important role in the ripening mechanism. At lower pH, as in cheese, coagulants become more and more aspecific. The various coagulants differ considerably in their ability to breakdown casein proteins. For that reason not all coagulants are suitable for all types of cheeses. Microbial coagulants are believed to be less suitable for long-ripening cheeses and are found to develop bitter off-flavors, although chemical modifications and processing optimization result in improved performance.

The separation of whey from curd is a crucial moment in the cheesemaking process in which retention of milk constituents in the curd is determined. Minerals and lactose are not retained in the curd. Particles such as bacteria and fat droplets are entangled in the casein network and end up almost entirely in the curd. The water present in cheese is mainly protein-bound and can not be regarded as a solvent. On the basis of partition calculations only a few percent of a water-soluble component present in milk would end up in the cheese. Retention of proteins is determined by the extent and nature of interaction of the proteins and casein. Whey proteins are completely expelled, but a lysozyme is fully retained. Retention of coagulants in the curd depends on the pH of the renneted milk at the moment of whey separation. Chymosin is retained to about 15 % at pH 6.5, which is a common pH in semihard cheesemaking (Gouda, Cheddar). In the Camembert cheesemaking process whey separation occurs at pH 6.2, and about 50 % of chymosin is retained. Under the same conditions retention of porcine pepsin is 65 %. Chymosin is electrostatically bound to the para-κ-casein part of κ-casein after having split off the caseinomacropeptide. This interaction, based on differences in isoelectric points of the two proteins, becomes stronger at lower pH [537]. The microbial coagulants are retained to about 18 % in curd, independent of pH [538]. As coagulants participate in the first stage of proteolysis in cheese the extent of retention is of significant importance for the rate of cheese ripening.

5.1.4.2.6. Cheese Flavors and Ripening Acceleration

For several semihard and hard cheese types long ripening, and consequently long storage times under strict temperature and humidity conditions, contribute substantially to the cost of cheese. For this reason, ripening acceleration has long been a topic of study. Acceleration of ripening can be achieved by changes in the process, addition of enzymes to the cheese milk, or by manipulation of lactic acid bacteria (LAB) from the so-called starter culture.

The most obvious way to accelerate ripening is by elevating the storage temperature. However, this results in unbalanced flavor and texture characteristics.

As cheese ripening is, for the most part, due to proteolysis of casein, addition of proteases would be an effective way to enhance ripening. The problem of addition of specific proteases, apart from finding the proper proteolytic specificities, is that proteases added to the cheese milk most often do not end up in the curd. Like whey proteins, extra added enzymes usually end up in the whey fraction. For this reason, specific formulation techniques, such as encapsulation in lyposomes, cross-linking, and entrapment of the proteases for targeting into the curd have been sought. In the Cheddar cheesemaking process, in which salt is added to the curd, addition of enzymes along with the dry salt presents an opportunity for addition, in contrast to many other cheese types. An uneven distribution of enzymes, resulting in “hot spots” in the cheese matrix may, however, result.

In practice the costly methods of formulation mean that the desired cost reduction is not achieved. Moreover, ripening acceleration results in a mature flavor that often does not match with the textural characteristics. Another approach to accelerate cheese ripening is by modification of starter cultures, both genetically or by composition. Addition of extra LAB is not feasible due to uncontrolled acidification of the cheese milk resulting in unwanted cheese characteristics. An elegant way may be the addition of attenuated starter cultures [539] in which the LAB have been killed by hot/cold temperature shock or by exposure to microwaves, but catabolic enzyme activities are still present. The problems of fast acidification of the cheese milk and low retention in curd have been solved in this way.

In the past the application of lysozyme (to lyse LAB in cheese) and lactase (to convert lactose into glucose and galactose, thereby offering other nutriants to LAB) have been studied, but did not result in commercial application.

Proteases and lipases have found commercial application in the production of spray-dried cheese-flavor protein derived from enzyme-modified cheese (EMC) processing. EMC is made of short-ripened, enzyme-inactivated cheese to which additives such as bile salts, flavors, and enzymes are added. Maturation proceeds at relatively high temperatures and high water content in a short time. Lipases and proteases of microbial origin are preferably used.

5.1.4.2.7. Lipase

Milk fat consists mainly of triglycerides. Lipases (triacylglycerol acylhydrolase, E.C. 3.1.1.2) hydrolyze triacylglycerols to di- and mono-acylglycerols, free fatty acids, and glycerol.

Lipolysis plays a significant role in the flavor formation of mold-ripened cheeses, blue cheeses, and some Italian types [540]. In other types of cheese the contribution of lypolysis to flavor formation is still obscure. Milk contains the endogenous lipoprotein lipase (LPL, E.C. 3.1.1.34). This LPL is associated with casein, is consequently incorporated in the curd, and may partially survive pasteurization. In milk this lipase is relatively inactive due to its immobilization on the casein micelle and due to the milk fat being present in milk fat globules surrounded by a protective proteinaceous membrane. Severe agitation or homogenization may damage these structures, after which lypolysis may occur, and the milk becomes rancid.

Pregastric esterases are present in rennet paste and are held responsible for the “piccante” taste of several Italian hard cheeses. Pregastric lipase powders are now used in Italian types of cheese along with a milk coagulant in view of the poor microbiological quality of rennet pastes. Due to pasteurization of raw milk, endogenous lipases are inactivated. Cheese types that, in the past, were made from raw milk are now made from pasteurized milk and added lipases.

Replacement of mammalian lipases by microbial ones is a topic of interest. Lipases from *R. miehei* and Aspergilli strains are commercially available and useful for the manufacture of Italian cheese types. These

enzymes have a preference for hydrolyzing fatty acids located at the 1,3-positions of the glycerol group. The short-chain fatty acids are flavorful and volatile and contribute most to lypolitic flavor development.

Lipases in blue cheeses originate from the *P. roqueforti* surface mold.

5.1.4.2.8. Lysozyme
Hen's egg lysozyme (muramidase, E.C. 3.2.1.17) is an enzyme consisting of 129 amino acids, cross-linked by four S —S bridges and able to split β-(1,4) linkages between *N*-acetylmuramic and *N*-acetylglucosamine residues that constitute the cell wall of Gram-positive bacteria. Lysozyme is found in secretory fluids such as human and cow's milk, tears, and cervical mucus. Lysozyme is active in acidic conditions and is able to withstand 100 °C with little loss of activity.

In the cheesemaking industry lysozyme and nitrate are used in the prevention of the defect of “late blowing”, which is caused by gas (carbon dioxide and hydrogen) produced by outgrowing butyric acid bacteria, in particular *Clostridium tyrobutyricum*. At the same time organic volatile acids are formed, which negatively affect the taste and flavor.

Lysozyme does not seem to affect lactic acid bacteria when applied in low concentrations, although successful application has been reported in the use of lysozyme to accelerate St. Paulin cheese ripening through enhanced lysis of lactic acid bacteria [541].

In its lytic action against the *Micrococcus lysodeikticus* substrate the enzyme is inhibited by the presence of NaCl. This inhibition is stronger at higher pH [542]. For this reason lysozyme is most effective in stages before brining and before the salt has reached the interior of the cheese by diffusion. Lysozyme associates strongly with αs-casein and β-casein but not with κ-casein. This strong association is responsible for almost full retention of lysozyme in cheese curd.

Lysozyme has found many other applications [543] in the preservation of food (meat, fruits, vegetables, seafood, and wine).

5.1.4.2.9. Milk Protein Hydrolysates
A relatively new area of applications has developed for protein hydrolysates. The milk derived caseins and whey proteins are one of the important substrates available in nature. Hydrolysis of proteins has major impact on their functionalities. Fragmentation of the protein due to proteolysis strongly affects its physicochemical characteristics such as molecular shape and size, electrochemical properties, and hydrophobicity, and hence solubility, foaming, and emulsifying and gelation properties. Allergenicity is decreased and taste is severely changed, most often to bitterness. All these characteristics are very much controlled and determined by the degree of hydrolysis of the protein, which in turn is determined by the processing conditions and by the type of proteolytic enzymes used. The degree of hydrolysis can vary from less than 10 to more than 40 %.

Endo- and exoproteases with defined specificities may be employed to control the degree of hydrolysis, the concentration of free amino acids, the peptide size distribution, and even to control the taste, whereby the prevention of bitterness is often a challenge.

Numerous proteases from animal (e.g., pancreatic pepsin, trypsin) or plant (papain, bromealin) origin, but also industrial enzymes from microbial sources (*B. subtilis*, *B. licheniformis*) are used. Aminopeptidases of microbial origin are often used to improve taste, by removing hydrophobic residues from bitter peptides.

Milk protein hydrolysates are used in health and fortifying sports drinks, in infant and low-digestible enteral nutrition, and dietetic foods.

5.1.4.2.10. Transglutaminase
An enzyme which is e to play an important role in the texture and rheological properties of dairy food is transglutaminase (E.C. 2.3.2.13). This enzyme catalyzes the acyl transfer reaction between the Ψ-carboxyamide group of glutamine residues and ε-amino groups of lysine between proteins. In this way high-molecular proteinaceous polymers are built and influence the hydration, gelation, emulsification, and foaming properties of food substrates such as whey protein, casein, soy, and fish meat proteins. The degree

of cross-linking depends on the three-dimensional structure of the protein substrates. The nonglobular structure of caseins or denatured proteins facilitates this cross-linking [544].

In the dairy, applications in the manufacturing of yogurt, preventing syneresis, and in the renneting ability of skim milk in fresh cheesemaking and in cream whipping have been described [545]

5.1.4.3. Milk Processing

5.1.4.3.1. β-Galactosidase

β-Galactosidase (E.C. 3.2.1.23) or lactase is an enzyme that has found ample application in the hydrolysis of lactose, a disaccharide consisting of a glucose and a galactose moiety, in cow's milk. The consumption of milk is hampered for some people due to a deficiency of lactase in the digestive system. Lactose intolerance in these people, predominantly in Asian countries but also in Africa and South and Middle America, results in symptoms such as bloating, diarrhea, and flatulence [546]. In many cases these symptoms occur after consumption of 200 –300 mL of milk.

Hydrolysis of lactose initially results in cleavage next to a glucose molecule, which is released, and an enzyme/galactose complex is formed. Cystein and histidine play an essential role in the active center. The SH group donates a proton to the galactose oxygen atom, and the nucleopholic imidazole group of histidine attacks the C-1 carbon atom of galactose. In the next phase an acceptor ROH (e.g., water) protonates the sulfhydryl anion and liberates free galactose. When the acceptor is another sugar a di-, tri-, or even a polysacharide is synthesized. This activity, also called transferase activity, is made use of in the production of nondigestible oligosacharides. Transferase activity may vary between types of different sources, and oligosaccharide yield is affected by equilibrium concentrations of reactants and reaction conditions.

Ions such as Ca^{2+} and Na^{+} may inhibit the reactions, whereas K^{+}, Mg^{2+}, and Mn^{2+} are activators. The reaction is product-inhibited. Especially galactose strongly inhibits the enzyme-catalyzed reaction.

Table 15 lists microbial sources and characteristics of β-galactosidases [547].

Table 15. Microbial sources and characteristics of β-galactosidases

Origin	$pH_{opt.}$	T_{opt}	K_m (mM) lactose*	*M*, kDa	Activator Inhibitor
Fungal					
Aspergillus niger	3.5	58	85	124	
Aspergillus oryzae	5.0	55	50	90	
Yeast					
Kluyveromyces lactis	6.5	37	35	115	K^{+}, Mg^{2+} Ca^{2+}, Na^{2+}, Zn, Cu
Bacterial					
E. coli	7.2	40	2	540	Na^{+}, K^{+}
B. subtilis	6.5	50	700		
B. stearothermophilus	6.2	55	2	220	Mg^{2+}
L. thermophilus	6.2	55	6	540	

The specific applications of lactases are determined by the optimum pH and pH-dependent stability characteristics of the different lactases. Neutral lactase is predominantly used in the production of low-lactose sterilized milk. Another concept includes lactase tablets to be added to the milk by the consumer a

prescribed period of time before consumption. Acid lactase is able to withstand the acid environment of the stomach and can be consumed with the milk and help the consumer hydrolyze lactose in the digestive tract.

Neutral lactase has also found application in lactose conversion in sweet whey (pH 6.1) and in the manufacture of lactose-free pet milk. Fungal lactase is more suitable in acid whey. Minor applications [547] are found in the manufacturing of yogurt (different outgrowth of cultures), in the sweetening of dairy-based beverages (lactose is less sweet than glucose and galactose), and in preventing the perception of sandiness due to crystallized lactose in ice cream.

5.1.4.3.2. Other Enzymes
Other enzymes such as glucose oxidase (E.C. 1.1.3.4), catalase (E.C. 1.11.1.6), superoxide dismutase, lactoperoxidase (E.C. 1.11.1.7), and sulfydryl oxidase have limited applications. Glucose oxidase, catalase, and lactoperoxidase are involved in the preservation of raw milk. Sulfhydryl oxidase is used to eliminate off-flavors generated due to the UHT treatment of sterilized milk.

5.1.5. Other Food Applications
5.1.5.1. Introduction
The use of enzymes for food applications has increased steadily over the past two decades, not only in traditional application areas such as starch processing, brewing, fruit processing, and dairies. In applications such as baking, the use of enzymes has grown even more [548].

The appearance of new enzyme applications is due to the increasing diversity of the enzymes available, the majority based on GM technology. In future new enzymes for food applications are expected to lead to major developments in the use of industrially produced enzymes [549].

5.1.5.2. Meat and Fish
5.1.5.2.1. Meat Processing
Cooked meat is considered tender if it is easy to chew and, at the same time, retains the desired texture. Tenderness is influenced by a number of factors, which are as yet not well understood. This applies not only to the biochemical reactions involved in rigor mortis, but also to their termination. Indeed, numerous endogenous enzymes take part in this process, including endogenous proteinases, particularly the cathepsins. These enzymes change muscle protein during maturation or aging of meat.

The mechanism of the maturation process has been studied in detail [550]. Natural maturing of carcasses or pieces of meat takes approximately 10 d at 1 – 2 °C. Slow maturation has the advantage of producing very tender meat, but also the disadvantage of moisture loss and shrinkage. Optimal meat maturation in special cold-storage depots (1 – 2°C, 83 – 86 % humidity) leads to a water loss of up to 7 % over three to four weeks. Since 1940, attempts have been made to use exogenous enzyme preparations as meat tenderizers. Proteinases capable of digesting connective tissue and muscle protein have been chosen for this purpose.

Papain (E.C. 3.4.22.2) [9001-73-4], bromelain (E.C. 3.4.22.33) [9001-00-7], and fiacin (E.C. 3.4.22.3) [9001-33-6] are the enzymes used on a commercial scale. The main problem associated with the use of enzymes is their even distribution in the tissue. Factors influencing this distribution are diffusion, time, salt content, and enzyme concentration.

If enzyme preparations are only sprinkled onto the surface of meat (as recommended for kitchen use) or if pieces of meat are dipped into an enzyme solution, only the surface is tenderized, and the interior remains tough. After enzyme application in the kitchen, e.g., using 2 % NaCl with 0.002 % bromelain, the meat is repeatedly poked to make it easier for the enzyme to penetrate. The main effect of the proteinase, though, is exerted during cooking.

Commercial methods can be divided into premortem and postmortem procedures. In postmortem treatment, a proteinase solution is spread in the carcass by repeated injections, possibly under pressure. In the premortem method (the Swift technique developed in 1960) a very pure, sterilized papain solution is injected intravenously 2 – 10 min. before the animal is slaughtered.

Yet another method involves injecting a papain solution that has been reversibly inactivated by oxidation. In the last stages of cooking the enzyme is reactivated by the liberation of mercapto groups.

Pancreatic proteinases are used in the maturation of fish. Bacterial proteinase is employed to dissolve bone meat or segments of meat.

5.1.5.2.2. Fish Processing
Until recently, industrial enzymes have only been used in limited amounts in the fish industry [551]. Today interest in controlling or aiding traditional fish processes has increased, and uses have emerged for enzymes in deskinning, membrane removal, and roe purification.

Fish sauce is a traditional product in some Asian countries, with an annual production of about 250 000 t. The production procedure consists of mixing small, uneviscerated fish with high concentrations of salt and storing the mixture in sealed vessels at ambient temperature for 6 – 18 months. During this process hydrolytic enzymes slowly degrade the fish tissue by proteolytic action with the formation of peptides and amino acids. There is some interest in accelerating fish sauce processing. Plant enzymes such as bromalein, fiacin, and papain digest fish tissue over a shorter period [552], and fungal proteinase increases the initial rate of protein hydrolysis.

Enzyme processes are also important in the traditional *salt curing* of herring, proteolytic enzymes having a favorable effect on both the flavor and texture of salted herring.

Another application of proteolytic enzymes is the *deskinning* of fish such as tuna, which are difficult to skin by manual or mechanical means. This process is based on placing the fish in a warm water bath containing a proteolytic enzyme for 10 – 90 min. After immersion, water jets can remove most of the skin. Enzymes are also used for deskinning and tenderizing squid.

Another use of proteolytic enzymes is in *fish roe purification*. Here the enzyme hydrolyzes the supportive and connective tissue that covers salmon or trout roe eggs and the roe sack, leading to a yield of 85 % compared to 50 – 65 % when using conventional processing. In Scandinavia and Canada 40 – 50 t of caviar (1986) are produced by this method.

5.1.5.3. Protein Cross-linking
The quality of meat protein is an important factor for meat texture and water binding. The addition of ingredients such as vegetable proteins to meat products is well known to improve water binding and other characteristics. It is, however, also possible to improve the meat product by modifying the meat protein by cross-linking. One such modification uses the protein-cross-linking enzyme transglutaminase, which in nature is known to cross-link proteins. The physiological function of transglutaminase cross-linking is to enhance the strength of the molecular structure of the protein network [553]. Substrate specificity is a key aspect in developing protein–protein cross-linking.

Transglutaminase (E.C. 2.3.2.13) catalyzes acyl-transfer reactions, introducing covalent cross-links between proteins, as well as peptides and primary amines. In protein cross-linking, ε-amino groups of lysine residues and ϒ-carboxyamide groups of glutamine residues act as acyl acceptors and acyl donors, respectively (Fig. 45).

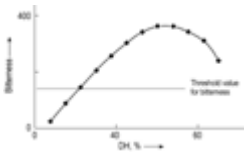

Figure 45. Quality relation between bitterness and DH of proteins

Source: Novo, Enzyme Information. IB-number 282-IB. Date 1982-12-23.
[Full View]

Transglutaminase of microbial origin is now commercially available [558] and has found several applications in the meat industry, both as a gelling agent and in connection with improving water binding.

Transglutaminase can also produce restructured meat by binding meat pieces [559], [560]. The meat binding system is based on the use of transglutaminase and caseinate simultaneously. Caseinate treated with transglutaminase acts as a glue to hold food components together. Using this system, a larger piece of

restructured meat or fish fillet can be prepared from smaller pieces. Minced meat can also be bound together without salt and phosphates to give more consumer acceptable products.

5.1.5.4. Flavor Development

5.1.5.4.1. Protein Hydrolysis

Hydrolyzed protein has been used for centuries to improve the taste of food products. The use of soy protein is well known in the production of the soy sauce by enzymatic hydrolysis during microbial fermentation. Soy and other vegetable proteins can also be used to produce a flavoring called HVP (hydrolyzed vegetable protein) by chemical hydrolysis. This chemical hydrolysis, however, has a number of disadvantages, including very high salt concentration and formation of unacceptable components. More recent developments have shown that better results can be obtained by using enzymatic reactions to produce HVP and other flavor-enhancing proteins.

There are many applications for enzyme-enhanced flavors, including soy and wheat hydrolysates for flavoring soups and sauces, and meat hydrolysates for flavor improvement of meat products, soups, and sauces. Milk protein hydrolysis and lipid hydrolysis contribute to flavor improvement in cheese products.

Naturally occurring food proteins do not contribute chemically to flavor formation in foods. It is, however, well known that peptides and amino acids, which are hydrolyzed protein products, do have a flavor. Amino acids taste either bitter or sweet, with the exception of aspartic and glutamic acid which have a sour taste [554]. Traditionally, these hydrolyzed products, available as flavor enhancers for soups and stock cubes etc., have been produced by acid hydrolysis of proteins during heat treatment. Flavors produced in this way are widely used in the food industry. Increasing concern about the less healthy compounds that develop during acid hydrolysis has led to alternative flavor production processes based on enzymatic hydrolysis of protein.

Protein hydrolysates produced by enzymes often have a bitter taste. This bitterness may be more or less pronounced depending on the type of amino acid and the length of the peptides. The development of bitterness depends on the degree of protein hydrolysis. As the degree of hydrolysis (DH) increases, the level of bitterness rises to a maximum before falling again (Fig. 46) [555].

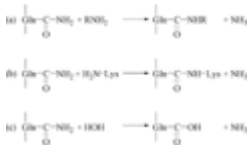

Figure 46. Reaction catalyzed by Transglutaminase (a) Acyl-transfer reaction (b) Cross-linking reaction (c) Deamidation

Source: Ajinomoto information http://www.transglutaminase.com/english/frame/ajinomoto1.htm [Full View]

Bitter peptides are characterised by their high content of hydrophobic amino acids, peptides with a high content of hydrophilic amino acids generally having a bland taste. From present knowledge of the amino acids in common food proteins, it can be predicted that casein-, maize-, and hemoglobin-derived hydrolysates have a tendency to be more bitter, while protein hydrolysates from meat, fish, and gelatin are less bitter.

Flavors produced by enzymatic hydrolyses of protein depend on the type of protein used (animal or vegetable), the type of enzyme, and the combination of enzymes used. By selecting special enzymes, the formation of bitter peptides can to a certain extent be avoided [556].

The ability of protease to develop flavor is used to accelerate the ripening of fermented sausages [557]. The ripening time can be reduced by 30 – 50 % using a combination of endo- and exoproteases (Table 16). The use of proteases in fermented sausages also has an impact on the microorganism used in the fermentation process.

Table 16. Proteases used to accelerate ripening of dry fermented sausages [557]

Enzyme	Type	Source
NCCDO151 proteinase	Serine-proteinase	*Lactobacillus paracasei* subsp. *paracasei*
Pronase E	Mixture of proteinases, amino-and carboxypeptidase	*Streptomyces griseus*
Aspartyl-proteinase	Acid proteinase	*Aspergillus oryzae*
Papain	Thiol-proteinase	*Carica papaya*
Alcalase	Serine-proteinase	*Bacillus lichiniformis*
Bromelain	Endoprotease	*Ananas comosus*
Neutrase	Metalloproteinase	*Bacillus subtilis*
Fungal Protease (Protease P)	Mixture of acid, neutral and alkaline proteinase	*Aspergillus oryzae*
Flavozyme	Mixture of proteases with both exo- and endopeptidase	*Aspergillus oryzae*

5.1.5.4.2. Lipid Hydrolysis

Traditionally, lipase has been used to enhance the flavor of cheese, especially in strongly-flavored products such as Parmesan [557]. Lipase degrades triglycerides and liberates fatty acids, which are generally not flavors themselves (with the exception of short chain fatty acids) but precursors for flavor substances.

Lipases from different sources (Table 17) have been used to enhance flavor development in cheese and dry-fermented sausages.

Table 17. Lipase for flavour development [557]

Enzyme	Type	Source
Lipase	–	Throat glands of kid goat
Lipid esterase	–	*Rhizopus arrhizus*
Lipase	1,3-specific	*Lactobacillus plantarum* MF 32
Pancreas lipase	1,3-specific	Porcine pancreas
Lipase	–	*Mucor miehei*
Lipase	–	*Candida cylindracea*
Palatase M®	1,3-specific	*Rhizomucor meihei*
Lipozyme®	1,3-specific	*Rhizomucor meihei*
Lipase AP6®	1,3-specific	*Aspergillus* spp.
Novozym® 667BG	1,3-specific	*Thermomyces lanuginosus*

Lipases show specificity for fatty acid chain length and esterification position, but it is not completely clear which fatty acid profile is ideal for the flavor profile.

Animal lipases release mainly short-chain fatty acids, whereas microbial lipases release long chain fatty acids. The unsaturated fatty acids released by lipase act as substrates for oxidation reactions and are thus precursors for flavor compounds such as aldehydes and ketones.

5.1.5.5. Egg powder

Egg powder has been produced since the early 1900s. One of the problems faced in connection with egg powder production is the Maillard reaction of protein and glucose in eggs. As early as the 1950s, an enzymatic process was developed to remove glucose from liquid whole egg or egg white (albumin) based on the use of glucose oxidase (E.C. 1.1.3.4) and catalase (E.C.1.11.1.6) [561]. The desugaring of egg albumin with catalase and glucose oxidase is performed at approx. 10 °C, and the pH is adjusted to 6.8 – 7.0 to optimize the performance of the enzymes. This pH adjustment is carried out by slowly adding citric acid solution (to prevent local protein precipitation). Acidification is followed by the addition of hydrogen peroxide. The addition of catalase converts H_2O_2 to oxygen and water (Fig. 47). Glucose is oxidized to glucone δ-lactone by glucose oxidase with oxygen consumption.

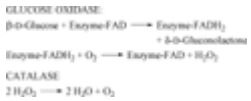

Figure 47. Oxidation of glucose by glucose oxidase. Reduction of hydrogen peroxide by catalase. [Full View]

If oxygen becomes the limiting factor, more hydrogen peroxide should be added until the desired glucose concentration (0.1 %) is reached. The desugaring procedure for whole egg or egg yolk is similar to the procedure for albumin, but no acid is added to adjust the pH.

Lipase and phospholipase have recently found applications in egg processing. Egg yolk contains phosphatidylcholine (PC), an emulsifier commonly known as egg lecithin. The enzyme-modified product is derived from the catalytic hydrolysis of a free fatty acid from the PC molecule, which yields lysophosphatidylcholine (Fig. 48). This structural modification improves the emulsification properties of the yolk without altering appearance or nutritional profile. Compared to traditional egg yolk, the structure of emulsions prepared with the enzyme-modified product is more stable and heat-resistant. Generally, the emulsions have greater viscosity and body and are able to emulsify more oil, often at reduced usage levels.

Figure 48. Hydrolysis of Phospholipid by a Phospholipase A2
[Full View]

5.1.5.6. Oils and Fats

The commercial application of enzymes in oil and fat processing is still rather limited. This situation is expected to change over the next few years due to growing interest in milder processes and the availability of new enzymes.

Lipases are by far the most widely used enzymes in oil processing. Their natural ability to hydrolyze oils and fats can be used to enrich oil and fat products with a specific fatty acid. Lipases can also be used to rearrange fatty acids and produce fats and oils with other physical characteristics.

Phospholipase (E.C. 3.1.1.4) is commercially used for removing phospholipids during oil refining (degumming).

Lipoxygenase (E.C. 1.13.11.12) can catalyze the oxidation of polyunsaturated fatty acid.

5.1.5.6.1. Fat Splitting

The hydrolysis of fats and oils is of major importance to the oleochemical industry in the production of fatty acids and their derivatives. These products are used in soaps, detergents, and pharmaceuticals as well as in food applications. Hydrolysis or fat splitting is normally carried out under pressure at a high temperature (typically 3 – 6 MPa and 250 °C). However, this process is not suitable for oils with a high content of polyunsaturated fatty acids. For this type of oil, hydrolysis by an enzymatic process is of great advantage. The enzymatic process is easily carried out at 40 °C in a 50 – 70 % oil-in-water emulsion. To enable reuse of

the enzyme, a number of processes based on immobilized lipase have been developed, making the process more economically feasible.

This principle is of special interest in the enrichment of long-chain polyunsaturated fatty acids from fish oil. Enzymatic hydrolysis of fish oil leads to partial glycerides high in eicosapentaoenic acid (EPA) and docosahexaenoic acid (DHA). *Candida rugosa* lipase is of special interest in this process, because this enzyme shows relatively low activity on DHA compared to EPA [562], [563]. If required, the partial glyceride high in DHA fatty acids can be re-esterified with DHA or EPA to generate triglycerides enriched with these fatty acids (Fig. 49) [564].

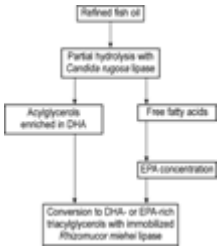

Figure 49. Reaction scheme for the enrichment of fish oil

Source: R. Diks, J. Bosleve in T. Uwe, T. Bornscheuer (eds.): *Enzymes in Lipid Modification,* Wiley-VCH, Weinheim 2000.
[Full View]

5.1.5.6.2. Interesterification

It is well-known that polyunsaturated fatty acids (PUFA) play an essential role in human nutrition and have important biomedical properties. Therefore, many physical methods have been developed for extracting and enriching these fatty acids, including crystallization, distillation, and the use of supercritical carbon dioxide. Lipase now offers an alternative, having the advantage of mild processing conditions that minimize fatty acid degradation and oxidation. These reactions rely on the principle that some lipases are able to discriminate between certain fatty acids, and that other lipases are selective for the position of the fatty acid in triglyceride moiety.

A special feature of lipases is their ability to work in nonaqueous environments. This property has been used to develop lipase-catalyzed rearrangement processes to generate tailored triglycerides that would be difficult to obtain by conventional physical and chemical methods. Lipases in nonaqueous media utilise the “reverse” reactions, whereby lipases catalyze esterification and rearrangement reactions under low-water conditions (Fig. 50).

Figure 50. Lipase catalysed transesterification reactions

Source: R. Rastall (ed.): Ingredients Handbook, Enzymes, Leatherhead Food International, Surrey 1999.
[Full View]

These systems, however, must contain small amounts of water to keep the enzyme hydrated. Lipases from different sources differ considerably in their ability to work under low-water conditions. Lipase from *Rhizomucor miehei* displays good activity in low water environments ($a_w < 0.2$) and is often used for these types of applications.

Lipase is normally immobilized for this type of reaction. Immobilization presents opportunities for reusing the enzyme, easy separation of the product from the enzyme, and running a continuous process.

The current main application of immobilized lipase in nonaqueous media is the production of tailored triglycerides for use in confectionery fat and nutritional supplements (Fig. 51).

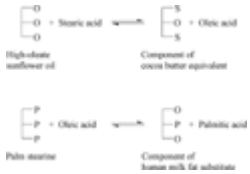

Figure 51. Reactions catalysed by a 1-,3-specific lipase

Source: R. Rastall (ed.): Ingredients Handbook, Enzymes, Leatherhead Food International, Survey 1999.
[Full View]

In the production of cocoa butter substitute (Fig. 52), a mixture of high-oleate sunflower oil (OOO) and stearic acid is passed through a packed bed reactor with immobilised *R. miehei* lipase. The fatty acid exchange occurs in the sn-1 and sn-3 positions of the triglyceride to generate SOS, which is a cocoa butter equivalent. The mixed fatty acid (oleic and stearic acid) is removed and can be hydrogenated and reused. Unwanted triglycerides can be removed by crystallisation and recycled.

Figure 52. Schematic production of cocoa butter equivalent (CBE)

Source: R. Rastall (ed.): *Ingredients Handbook, Enzymes*, Leatherhead Food International, Surrey 1999.
[Full View]

Lipase-catalysed interesterification can also be used to improve the nutritional properties of dietary fat. An example of this is the production of tailored triglyceride for use as a nutritional supplement in infant formula [566]. Human milk fat contains a high proportion of unsaturated fatty acids at the sn-1 and sn-3 positions, with palmitic acid in the sn-2 position. Lipase from the human pancreas releases the fatty acid in the sn-1 and sn-3 positions and produces 2-monoacylglycerol. The body readily absorbs these components.

Human milk substitute based on vegetable oils mainly has saturated fatty acid in the sn-1 and sn-3 positions, which, when released by pancreatic lipase, tends to form calcium soaps and is poorly absorbed by the body. Enzymatic interesterification makes it possible to produce a triglyceride with unsaturated fatty acid in the sn-1 and sn-3 positions and palmitic acid in the sn-2 position (OPO). In this process, palm oil high in palmitic acid is treated with oleic acid to form OPO triglyceride.

Similar techniques can be used to incorporate medium-chain fatty acids ($C_8 – C_{10}$) in specific positions in a triglyceride to generate lipids with improved nutritional properties.

Enzymatic interesterification of fats and oils for margarine and shortening has now become an attractive alternative to the traditional interesterification reaction using high temperature and inorganic catalyst. The enzymatic process is more attractive because it is carried out at 40 – 70 °C, a temperature range in which no *trans* fatty acids are formed.

5.1.5.6.3. Esterification
Enzymatic esterification for the production of diglyceride has attracted interest due to the nutritional aspects of 1,3-diglycerides. Commercial products containing high levels of diglyceride are on the market in Japan, where they have increased their share of the cooking oil market. Use of 1,3-diglyceride in cooking oil contributes to a lower serum lipid level, as pancreatic lipase hydrolyzes 1,3-diglyceride into free fatty acids and glycerol, while no 2-monoglyceride (a building block for serum lipid) is formed.

5.1.5.6.4. Oil Degumming
Crude vegetable oil such as soy oil contains phospholipids, which are normally removed by chemical and physical methods. During the last decade there has been a tendency to move away from chemical (caustic) refining processes and use only physical processes to reduce the amount of effluent. For reliable physical processing, it is essential that the phospholipid content of vegetable oil is reduced to a very low level (< 15 ppm P) before bleaching and deodorization take place.

Crude vegetable oil has a phospholipid content of up to 3 %, depending on the type of seed and extraction

procedure. The first stage of the refining process, water degumming, removes most of these phospholipids, but about 0.6 % nonhydrating phospholipid still remains in the oil. A number of physical methods have been developed to lower the phospholipid content further.

A newer method based on enzymatic hydrolysis of phospholipids by phospholipase A2 has, however, provided an interesting alternative. Phospholipase A2 catalyzes the hydrolysis of fatty acids in the sn-2 position and forms the more water-soluble lysophospholipid, which is more easily washed out of the vegetable oil [567]. This method gives an oil with a residual phosphorus content of less than 10 ppm.

Phospholipase A2 is also used for the production of lysolecithin, which has superior emulsification properties [565]. Lysolecithin is used in the food, cosmetic and pharmaceutical industry.

5.2. Enzymes in Nonfood Applications

5.2.1. Enzymes in Household Detergents

Household detergents need to remove a broad range of complex soil from different fiber surfaces. Soil and stain components with good water solubility are easily removed during the cleaning process. Most other stains are partially removed by the surfactant/builder/bleach system of a detergent, although the result is often unsatisfactory, depending on the washing conditions.

In most cases a suitable detergent enzyme aids the removal of soils and stains. Whereas the detergent components have a purely physicochemical action, enzymes act by degrading the dirt into smaller and more soluble fragments. However, to remove a stain totally still requires the joint action of the enzyme, the surfactant system, and mechanical agitation.

5.2.1.1. Historical Development

The first enzyme-containing detergent was introduced to the household market as early as 1913 when Röhm & Haas in Germany added the protease trypsin, extracted from pig pancreas, to their detergent Burnus, utilizing a patent of OTTO RÖHM [568].

Because the activity and stability of trypsin are moderate in the presence of typical detergent ingredients, the detergent enzyme concept did not really catch on until 1963, when Novo developed and marketed a much more alkali- and builder-tolerant bacterial protease called Alcalase [569].

Small detergent producers in Switzerland (BIO-40 from Gebrueder Schnyder) and the Netherlands (Biotex from Korman and Schulte) were the pioneers in the commercial utilization of Alcalase, which initially was only considered useful for washing bloodstained laundry from hospitals and slaughterhouses.

Gist-Brocades followed a few years later with the alkaline Maxatase protease (1965).

Within five years of their introduction into detergents, more than 50 % of all heavy-duty laundry detergent for domestic use in Europe contained protease enzymes, compared to only 15 % in the USA.

1970 – 1980. The rapid growth of the enzyme-containing detergents was temporarily discontinued in the early 1970s when industrial hygiene and safety problems became evident. Significant improvements in detergent manufacturing techniques, the development of low-dust encapsulated enzyme prills, and improvements in industrial hygiene practices and procedures resolved the issue. Since then enzyme sales have again increased steadily.

The first α-amylase for detergents (Termamyl, from Novozymes) was developed in the early 1970s and has increased tolerance to temperature and alkali.

In the mid-1970s the first liquid detergents were introduced in Europe and the USA using specially developed liquid enzyme preparations.

1980 – 1990. Washing conditions, especially in Europe, gradually shifted further to lower temperatures (from 60 to 40 °C), reduced water quantities, and shorter washing times, and major changes in formulation took place (phosphate-free detergent, activated bleach). New proteases better adapted to these working conditions and new enzyme activities became available.

The first detergent cellulase Celluzyme (Novozymes), a multicomponent cellulase, was developed for improved cleaning and stain removal.

New genetic technology and protein engineering techniques have been introduced for designing further optimized enzymes for detergent conditions or to achieve acceptable production economy. In this way the first bleach-compatible protease (Maxapem, Genencor) and the first detergent lipase (Lipolase, Novozymes) have been developed.

1990 – 2000. New detergent markets opened up in India, South America, Eastern Europe, the former Soviet Union, and China. Because of the specific washing conditions in these regions, such as cost constraints, low detergent concentrations, very low washing temperatures, and presoaking, new enzyme developments were again needed to cope with these difficult washing conditions.

Second-generation detergent proteases, amylases, and lipases were developed by using protein engineering and genetic techniques for further improved performance under difficult washing conditions (low temperature, single wash effects (cleaning, stain removal, whitening, softening, other, achieved with one single wash treatment.)) and with improved compatibility with new detergent technology (improved activated bleach) and for broader application (e.g., in automatic dish washing, liquid detergents, tablets, gel pads).

Up to the 1980s enzymes were primarily developed and used for aiding stain removal and cleaning of fabrics. Since the mid-1990s also enzymes with specific fabric-care effects (color brightening and fabric softening) appeared on the market. (Carezyme and Endolase monocomponent cellulases from Novozymes) as well as enzymes delivering new benefits for stain removal and cleaning (Mannaway, Novozymes).

Table 18 lists enzyme products for detergents by type and manufacturer

Table 18. Enzyme products for detergents by type and manufacturer (2002)

Protease						**Amylase**			**Lipase**	**Cellulases & hemicellulases**
Type	Neutral	Alkaline	Highly alkaline	Bleach-stable	Cold-water	Conventional	Bleach-stable	Cold-water		
Merchant market										
Novozymes	X	X	X	X	–	X	X	X	X	X
Genencor International	–	–	X	X	X	–	X	–	–	X
Captive market										
Kao (Japan)	–	–	X	–	–	X	–	–	–	X
Henkel	–	X	X	X	–	–	–	–	–	–

5.2.1.2. Laundry Soils

Laundry detergents need to remove a broad range of complex soil from different fabrics and fiber types under varying washing conditions. The most frequent soiling on garments and fabrics include all kind of body soils on shirts, underwear, and socks; a variety of food and beverage stains; oily soils; and particulate stains.

The body soils are complex mixtures secreted by sebaceous and sweat glands in the skin or can stem from

fecal, urinal, or vaginal secretions. These soils are mixtures of lipids, proteins, glycoproteins, blood and blood degradation components, skin epidermal debris, salts, and particulate materials collected from the environment. The fresh lipid fraction contains polyunsaturated fatty acids (linoleic and linolenic acid) and their triglycerides and unsaturated hydrocarbons (squalene), which can be rapidly autoxidized when exposed to air, heat, and moisture. The lipid autoxidation reactions yield high molecular mass complex compounds with lipids, proteins, and carbohydrates. This results in a gluelike matter that entraps skin debris and particulate soil components. This aged body soil is tough to remove from fabrics and tends to build up on the fabric during the lifetime of the garment.

The food stains encountered in laundry are also complex soils. They contain substantial amounts of animal and vegetable fats and oils; proteins, e.g., from eggs and meat products; carbohydrates such as sugars, starches, and other complex carbohydrates; residues from plants, e.g., from grass or spinach; and mixed oxidation products. Chemical and enzymatic browning of food products occurs during storage and preparation of foods, mostly by formation of dark-colored Maillard reaction products from carbohydrates and proteins at higher temperatures. Natural dyes and pigments present in foodstuffs such as carotenes and chlorophylls add to the deep coloring of some food stains. In particular carotenes are highly soluble in oils and fats, e.g., in tomato-based sauces, and this adds to the difficulty of removing these stains from fabrics. Certain dairy-based foodstuffs can contain substantial amounts of less common carbohydrate-based gums for stabilizing and texturizing the food, such as carrageenan, guar, and locust bean gums, and these are often difficult to remove from fabrics by detergents and even by current amylases.

Colored stains from beverages such as wines, vegetable and fruit juices, and tea contain complex mixtures of natural dyes and pigments such as anthocyanine and colored polyphenolic compounds, sugars, and organic acids. These materials can eventually be polymerized to higher molecular mass substances that are more difficult to remove from fabrics and fibers. Coffee can in addition contain substantial amounts of browning products formed during roasting of the coffee beans.

Particulate soiling can consist of insoluble metal carbonates, oxides, silicates (clay), carbon black, dust, and humus collected from the environment, e.g., airborne soot on shirt collars or garden soil on socks.

Organic soils on fabrics also can stem from the laundry process and fabric care treatments such as surfactant and builder residues, fabric softeners, and from ironing aids, e.g., starches.

However, the frequency and type of soiling on garments is also geographically different, owing to different consumer habits. Figure 53 exemplifies the differences in soiling incidences in a typical north and south European country.

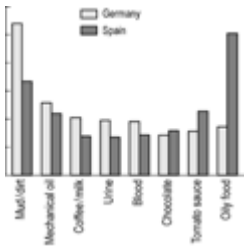

Figure 53. Soiling in Germany and Spain: incidence ranking [570]
[Full View]

5.2.1.3. Detergent Composition and Washing Process

5.2.1.3.1. Washing Process

Any washing process is an interplay between the washing equipment used; the materials entering the process, i.e., the detergent components and additives; water quality; the wash load composition; and the washing procedures followed, e.g., time, temperature, agitation. The equipment and wash procedures for laundering in three principal geographical areas are summarized in Table 19.

Table 19. Washing equipment and procedures [571]

	USA/Canada	Japan	Western Europe
Machine type	Top-loaded agitator	Top-loaded impeller	Front-loaded rotating drum
Wash time, min	12	10	20–90
Wash load, kg	2–3	1–1.5	3–5
Main wash water volume, L	35–80	30–45	8–15
Main wash temperature, °C	10–50	10–40	30–90
Water hardness, ppm $CaCO_3$	low (100)	very low (50)	high (250)
Recommend product usage, g regular detergent/L	1–5	1–3	5–10

As can be seen the average laundry conditions can vary significantly across the main geographical areas. Within each region also local washing habits exist such as separate addition of hypochlorite-based bleaches in southern Europe; extended boil washing as in Turkey; and multicycle use of the washing water in Japan.

All of this does not facilitate the formulation of detergent effective in the broad range of usage conditions. Often detergent compositions are adapted to these local washing habits.

Not all detergent components are effective over the whole range of temperatures from 4 °C up to boil wash and over the range of detergent concentrations. Here enzymes can become important detergency boosters especially for washing in the low temperature range and under conditions of low product usage.

5.2.1.3.2. Detergent Compositions
Detergent compositions also vary by geography and from country to country.

The household laundry detergents available to consumers can be roughly divided into five segments according to the physicochemical properties of the wash solutions prepared from the detergents:

- Near neutral pH, low ionic strength detergents are mostly liquid detergents having solution pH from 7.5 to 9. They contain no bleach and only low levels of salts. Dilute and concentrated liquid detergents are marketed. The latter type of liquid detergent contains only small amounts of water (< 10 %).
- Mildly alkaline pH, medium ionic strength detergents are typically compact powder detergents from Japan and regular powder detergents from the USA. Their solution pH is about 9 and they contain no bleach.
- Alkaline pH, high ionic strength detergents with a solution pH from 9.5 to 11, and with an activated bleaching system, e.g., European regular powder detergents, They also contain sodium sulfate as a filler, and builder systems, e.g., sodium triphosphate or zeolites and sodium carbonate. High dosages are used, which give a wash liquor with a high ionic strength.
- Alkaline pH, medium ionic strength detergents with bleach are represented by European compact powder detergents, which have gained a large market share during the early 1990s They are compact as a result of removal of most or all of the sodium sulfate; solution pH is from 9.5 to 11.

The household detergents are mostly delivered as low-dusting powders or granulates or as liquids with low or high viscosity (gels). Laundry bar soaps are still popular in certain countries (South Europe, Latin America).

More recently (2000) detergents in tablet form or as a single-dose pouch containing liquid detergent have been launched with success in Europe and North America by Unilever and P & G [572-574].

Practically all of these laundry detergents contain enzymes, although some specialty flanker brands are free of any enzymes. Some specialty laundry detergents combine fabric-cleaning and fabric-care technologies

delivered during the laundering process without use of additives (two-in-one products).

Examples of laundry detergent compositions are given in Table 20.

Table 20. Examples of laundry detergent composition (in wt %) [575]*

	Liquid	Liquid, concentrated	Granular low-bleach, USA/JP	Granular high bleach, Europe	Granular compact, Europe	Tablet
Surfactants	10–50	10–30	7–22	10–20	10–20	15–25
Builders	0–10	5–15	20–70	20–40	20–45	15–30
Bleaches	0	0	0–5	11–27	13–28	10–25
Enzymes	0–6	0–1	0–1	0.3–0.8	0.5–1.5	0–3
pH of 1 % solution	7.5–9	7.5–9	9.5–11	9.5–11	9.5–11	9.5–11

*Surfactants: linear alkyl benzene sulfonates, alcohol sulfates, alcohol ether sulfate, alcohol ethoxylates, cationic surfactants. Builders: zeolites, sodium carbonate, sodium silicate, sodium citrate, sodium tripolyphosphate (STTP), sodium nitrilotriacetate (NTA), polycarboxylates. Bleaches: sodium perborate, sodium percarbonate, tetraacetylethylenediamine (TAED), nonanoyl benzene sulfonate (NOBS) and others. Additives (not listed): foam regulators, stabilizers, processing aids, optical brighteners, antiredeposition agents, corrosion inhibitors, perfumes, colorants, fillers such as sodium sulfate in powders or water in liquid detergents.

5.2.1.4. Enzyme-Aided Detergency and Soil Removal

Soil and stain components with good water solubility are easily removed during the cleaning process. All other stains are partially removed by the surfactant/builder/bleach system of a detergent, although the result is often unsatisfactory, depending on the washing conditions.

In most cases a suitable detergent enzyme may help the removal of soils and stains. Contrary to the purely physical action of the surfactant system enzymes work by degrading the dirt into smaller and more soluble fragments. However, to remove a stain totally still requires the joint effects of the enzyme, surfactant system, and mechanical agitation.

Proteases and amylases can remove protein and starch stains, respectively. Fats and oils are generally difficult to remove at low wash temperatures (< 40 °C) by conventional detergents. By using lipases, it is possible to improve the removal of fats/oils of animal and vegetable origin even at temperatures at which the fatty material is in a solid form.

Particulate soils can be difficult to remove, especially if the particle size is small. Removal of particulate soil from cotton fabric can be improved by use of a cellulase that removes cellulose fibrils from the surface of the yarn.

Various kinds of dirt may adhere to textile surfaces via a glue of proteinaceous, starchy, or fatty material. In such cases of anchored dirt, an enzyme may assist in removing the dirt even though it does not attack the dirt directly. Compounds from several of these classes are intimately mixed in combined soiling, e.g., human sebum on shirt collars, cocoa milk, gravy, or chocolate.

Bleachable stains are the only group of stains for which no enzyme product for detergent applications has been marketed so far. However, the patent literature indicates that efforts are being made by the major enzyme suppliers and detergent producers to develop an enzymatic bleaching system for detergents.

5.2.1.5. Detergent Enzyme Performance Evaluation and Screening

Not all enzymes with a potential for stain degradation and/or removal are suitable for inclusion in detergent products. Enzyme performance is influenced by factors like detergent solution pH, ionic strength, wash

temperature, washing time, detergent composition, and mechanical handling. Detergent surfactants, bleach systems, and other enzymes can influence both performance and stability.

The enzymes must have a broad enough substrate specificity because the average load of dirty laundry contains the enzyme substrate in a multitude of forms. For example, in the case of proteases the protein substrate may have been oxidatively modified or complexed with carbohydrates.

A detergent enzyme must be active at the pH of detergent solutions (between 7 and 11) and at the relevant wash temperatures (4 to 60 °C). Importantly, the enzyme must be compatible with detergent components such as surfactants, builders, bleaches, and other enzymes, not only during the wash process but also upon storage when incorporated in the detergent. In particular, such an enzyme must be resistant toward denaturing surfactants and bleach and towards protease degradation under these conditions.

Enzymes can be screened by assay of their activity on soluble substrates under conditions different from those encountered in a realistic wash. Enzyme screening often uses spectrophotometric assays with chromophoric soluble substrates, and the procedure can be fully automated for high-throughput analysis. The assay results can indicate the enzyme's performance on soluble substrates with respect to pH and temperature variations and with respect to enzyme compatibility with detergent components or enzyme-inhibiting conditions, etc. However the analytical data thus obtained with soluble substrates are not necessarily representative of the wash performance of the enzyme, and often low correlation is achieved with performance results obtained with enzymes under realistic laundry conditions (Fig. 54).

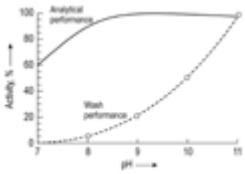

Figure 54. Example of deviating analytical activity on soluble substrates and laundry scale cleaning performance data [576]. Savinase protease: analytical performance on denatured hemoglobin, 25 °C, 10 min.; wash performance on grass stain, 30 °C, 10 min.
[Full View]

It is also important to realize that enzymes can have significant performance effects on substrates that apparently are unrelated to their target substrates. For example, proteases can deliver significant performance benefits on removal of protein-based bloodstains but also on grass stains on fabrics. Therefore real wash trials using a broad range of consumer relevant stains remain necessary to completely evaluate the wash performance of detergent enzymes.

Wash-performance assessments often use pre-soiled test pieces. For example the commonly used stains for protease evaluation are milk, blood, and grass applied to cotton or synthetic fabrics. These stains also may contain particulate matter, e.g., carbon black, as part of the stain matrix. These standardized test materials are available ready-to-use from a number of research and testing institutes in Europe and the USA, e.g., Center for Test Materials, Vlaardingen, Holland; Wascherei Forschungs Institut, Krefeld, Germany [577]; EMPA St. Gallen, Switzerland [578] (Table 21); Instituut voor Reinigingstechnieken TNO, Delft, Holland [579]; Test Fabrics, Middlesex, New Jersey [580] and SDL Textile Innovators [581]. Alternatively, enzyme manufacturers can supply stain preparation procedures, or specific stain material can be made in-house by the enzyme formulator. Finally, naturally soiled and stained fabrics obtained from households can be used for more realistic performance assessment.

Table 21. Examples of standard performance test materials from EMPA [578]

Soil	Cotton fabrics	Polyester/cotton blend
Carbon black/olive oil	101	104
Blood, milk, carbon black	116	117
Lipstick	141	142
Make-up	143	144
Starch	161	162
Tea	167	168
Grass	164	
Blood	111	
Wine (red)	114	

Laboratory wash trials are usually conducted in small-scale models of washing machines. The Terg-o-tometer [582] simulates the top-loaded U.S. type of washing machine, and the Launder-o-meter [583] or Linitest [584] simulates the European drum-type machine [585]. These small-scale models use 0.2 to 1 L of wash solution. Both have temperature-controlled heating and a timer. The wash solution is agitated with an impeller in the Terg-o-tometer and by rotation of a closed pot in the Launder-o-meter or Linitest equipment. The test stain of defined size is laundered in the wash solution with controlled detergent concentration and solution pH and with or without the enzyme under test. The enzyme concentration in the initial screening tests can be in the range of 1 - 100 mg/L of enzyme degradable protein.

The evaluation of the enzyme effects on the test pieces can be made visually by grading by an expert panel under controlled illumination conditions and by instrumentally measuring the reflectance of light under specified conditions. Typically, the intensity of light remitted R at 460 nm when illuminating the test pieces with a standardized daylight source is expressed as a percentage, ΔR or ΔE, of the intensity of incident light at the same wavelength. The ΔR or ΔE value is then a measure of the enzyme effect; it is defined as the difference in R between fabrics washed with and without enzyme. The R value is known to correlate well with the visual impression of whiteness of the fabric. Depending on type and degree of soiling, differences in R or E of 2 – 3 units are detectable by the human eye.

Digital imaging systems can be used for the instrumental evaluation of the performance effects in larger routine performance screening of enzymes using standardized stains and small-scale laboratory performance test methods.

For the complete evaluation of detergency effects of an enzyme two-steps can be considered: (1) “Primary washing effects” (cleaning efficacy) refer to the removal of soil and stains after one wash. Testing is carried out as described above using either artificially soiled test fabrics or naturally soiled laundry. (2) “Secondary washing effects” (effects after repeated washing) refer to potential multiwash performance (e.g., with Lipase) or to the detection of damage such as loss of tear strength, incrustations (ash residues), and graying. The latter evaluations are usually based on 25 or 50 washes, including a control fabric.

A convenient way to report detergent enzyme performance is the use of dose – response curves. Typically, the enzyme-related performance increases significantly in the lower enzyme concentration range, but reaches a maximum level at higher enzyme doses.

Figure 55 compares protease performance as a function of enzyme dosage. The performances at 60 °C reaches a plateau at an enzyme dosage of approximately 0.04 wt % of the enzyme granulate in the detergent. At this enzyme concentration, the fabric is clean and the stain can be totally removed, i.e., the measured R value is identical to that of an unsoiled test piece washed under the same conditions.

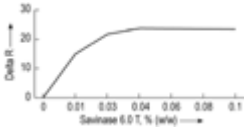

Figure 55. Example of a dose response curve for Savinase 6.0 T protease (Novozymes) in a powder detergent, 4 g/L in a Launder-o-meter test at 60 °C for 40 min. Source: Novozymes.
[Full View]

The extent to which the enzyme is able to remove stains and soils from the fabric depends on the detergent composition, the wash temperature, wash solution pH, washing time, wash load, etc. Enzyme wash performance varies between liquid and powder detergents and with the composition of the soiling. Figure 56 shows that the fabric composition also influences the stain-removal profile for the same soiling (EMPA 117 on polyester – cotton and EMPA 116 on cotton only).

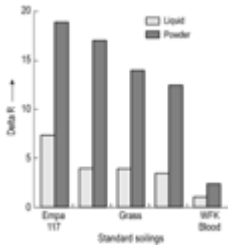

Figure 56. Stain removal performance of a U.S. liquid detergent (0.4 % Savinase 16L) and a U.S. powder detergent (0.8 % Savinase 8.0T) on different standard soilings in a Terg-o-tometer operating at 6°dH water hardness, 20 °C, 10 min. EMPA 117 (milk, blood, and ink on polyester/cotton); AS 10 (milk, oil, and pigments on 100 % cotton); grass (on 100 % cotton); EMPA 116 (milk, blood, and ink on 100 % cotton); WFK blood soiling (on 100 % cotton). Source: Novozymes.
[Full View]

Further scale reduction of the performance evaluation of detergent enzymes has been achieved, and high-throughput performance screens for enzymes are being developed. This type of performance assessment uses multiwell titer plates, each holding milliliter volumes of wash solution under controlled temperature conditions (Figure 57). A realistic level of agitation of the solution is assured with miniaturized stirrers. Using optimized test conditions a good correlation with real laundry conditions can be achieved. The performance can be evaluated visually or preferably instrumentally using titer plate readers and software that are commonly used in protein chemistry. Such miniaturized performance assessment has many advantages such as very low consumption of enzyme, on the order of micrograms versus milligrams in other small-scale tests, and the high-speed evaluation of many performance conditions in a multiwell experiment.

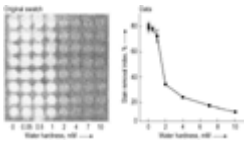

Figure 57. High-throughput performance screen for detergent enzymes. Enzyme performance on grass stain is evaluated at one enzyme level versus the hardness profile of the wash solution (0 – 10mmol Ca^{2+}). Left: the multiwell grass stain after treatment; right: instrumental evaluation of the multiwell grass stain. Source: unpublished data courtesy of G. Bechmann, C. Joos, S. Cooremans (Procter and Gamble Eurocor).
[Full View]

The final conclusion on the performance of a detergent enzyme candidate must be confirmed in a realistic laundry appliance using realistic wash loads and with wash programs at various wash temperatures and detergent usage versus water hardness. Finally these performance tests can be repeated in all important geographical or local washing conditions to confirm the robustness of the enzyme to the wash conditions.

5.2.1.6. Enzyme Types

5.2.1.6.1. Proteases

Proteases hydrolyze proteins into smaller fragments, i.e., peptides or amino acids. In a detergent context, protein stains on fabric (laundry) or hard surfaces (e.g., in dishwashing) are degraded into fragments which can be removed or dissolved by other components of the washing liquor.

Proteases are well-established ingredients in most household laundry detergents worldwide, except in some developing countries. They are used to remove stains such as blood and grass. The effects of proteases are very clear, and the enzyme cost, considering the effect, is relatively low. On some food stains, e.g., cocoa, a combined effect with amylases or lipases may be observed.

Most of the commercial detergent proteases are subtilisins and are classified as such in the Chemical Abstract Service. The Enzyme Classification number is E.C. 3.4.21.62. The name subtilisin derives from the name of the bacterial species, *Bacillus subtilis*, from which the protease was first isolated. Subtilisins are extracellular serine endo-peptidases, and the term also now covers proteases isolated from related *Bacillus* species.

The first protease of this type was subtilisin Carlsberg from *Bacillus licheniformis*, isolated in 1947 (also known as subtilisin A, Alcalase Novo). Later, in 1954, subtilisin was also isolated from *B. amyloliquefaciens*, also known as subtilisin BPN′, Nagase, Primase, Bacterial Proteinase Novo, Sub Novo.

The commercial detergent proteases are very similar in structure, and there are no fundamental differences between them. They are all characterized by very broad substrate specificity and mainly differ in temperature optimum, pH optimum, bleach sensitivity, and Ca ion demand. Some are highly alkaline, i.e., have a maximum activity in the high pH range, such as Maxacal and Savinase. Others are low-alkaline, e.g., Maxatase, Alcalase, and Subtilisin Novo (BPN′). Table 22 presents examples of commercially available protease and application ranges. They all have a molecular weight between 20 000 and 30 000, and have a serine residue in the active site of the enzyme. Figures 58 and 59 illustrate the pH and temperature profiles of some commercial low- and high-alkaline protease in analytical assay on dimethylcasein (DMC) substrate

Table 22. Examples of commercially available detergent proteases (Novozymes) [587-591]

Product name	Microorganism	pH application range	Temperature application range
Alcalase	*Bacillus* species	6 – 10	10 – 80
Esperase	*Bacillus* species	7 – 12	10 – 80
Everlase	GM *Bacillus* spp.	8 – 11	15 – 80
Savinase	GM *Bacillus* spp.	8 – 11	15 – 75
Durazym	GM *Bacillus* spp.	8 – 11	15 – 70

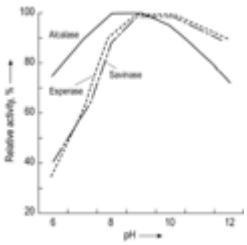

Figure 58. pH-dependent activity of different detergent proteases at 25 °C, 10 min reaction time, DMC substrate [588-590].
[Full View]

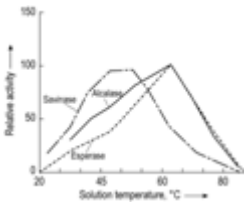

Figure 59. Temperature-dependent activity of different detergent proteases in solution at pH 8.5 (Alcalase) and 10.1 (Savinase, Esperase), 10 min reaction time, DMC substrate [588-590].
[Full View]

Protease performance can be strongly influenced by detergent pH and ionic strength. Surfactants influence both protease performance and stability in the wash solution. In general, anionic surfactants are more aggressive than amphoteric surfactants, which in turn are more aggressive than nonionic surfactants.

All detergent proteases are destabilized by linear alkylbenzenesulfonate (LAS) and alkyl sulfates (AS), the most common types of anionic surfactants in detergents. The higher the LAS concentration and wash temperature, the greater the inactivation of the enzyme. The presence of nonionic surfactants or ethoxylated alkyl sulfates, however, counteracts to some extent the negative effect of LAS. Since almost all detergents contain some nonionic surfactant or ethoxylated AS, the stability of proteases in a washing context is not problematic.

Phosphate builders such as sodium triphosphate and non-phosphate builders such as zeolite and citrate remove free calcium from the washing solution. Co-builders such as nitrilotriacetic acid and polycarboxylates also may be incorporated into the detergent formulation. Wash performance of detergents decreases with increasing calcium concentration. Protease performance varies, but high calcium concentrations tend to reduce protease performance. Therefore, it is an advantage to add a builder system to the detergent. Proteases need a small amount of calcium for the sake of stability, but even with the most efficient builder systems, stability during washing is generally not a problem.

Bleach systems in laundry products can also oxidize proteinaceous stains on fabric, often making them more difficult to remove. Detergent proteases can partly counteract this negative effect of the bleach system. Most detergent proteases are stable during the wash cycle in the presence of such oxygen-based bleach systems.

However, storage stability in detergents containing bleach may be a problem with the established detergent proteases. In particular, the amino acid methionine in the protease molecule can be oxidized during storage by the bleach system in powder detergents, which leads to inactivation of the protease. For example the Savinase detergent protease contains methionine residues in three positions, one of which is next to the active site serine. In new protein-engineered proteases introduced onto the market the most bleach-sensitive amino acid, i.e., methionine close to the active site, is replaced with other amino acids insensitive toward oxidation. This slight change in the molecular structure significantly increases the storage stability in detergents containing bleach. Bleach-stable proteases developed by using protein engineering include Everlase (one amino acid change from Savinase) and Purafect OxP, which are superior to Durazym and Purafect with respect to washing performance in bleach-containing detergents. In addition, the bleach-stabilized proteases also usually have better storage stability when formulated in bleach-containing detergents (Fig. 60).

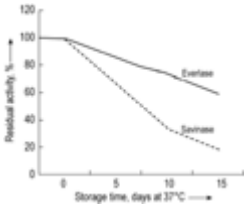

Figure 60. Storage stability of bleach-stable protease (Everlase, Novozymes) in European powder detergent with activated bleach system. Storage at 37 °C, 70 % R.H., open vial [587].
[Full View]

Chlorine bleach (sodium hypochlorite, NaOCI) is not incorporated into laundry detergents themselves, but can be used as a separate additive to the wash, as is common practice in some parts of the world. Hypochlorite bleach in excess of 200 ppm in the normal wash can quickly oxidize enzymes, resulting in significant loss of enzyme activity. Bleach-stabilized proteases can survive low concentrations of NaOCl, as depicted in Figure 61.

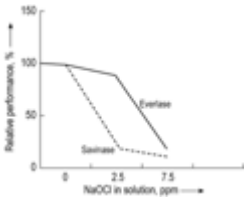

Figure 61. Performance activity of a bleach-stable protease (Everlase, Novozymes) versus parent protease (Savinase, Novozymes) in the presence of chlorine bleach. Grass stain removal, Terg-o-tometer, U.S. HDP at 1 g/L, pH 9.3, 40 °C, 10 min, 10° dH waterhardness [587].
[Full View]

5.2.1.6.2. Amylases

Commercial laundry amylases comprise the α-amylase from *Bacillus amyloliquefaciens* and the heat-stable α-amylase from *Bacillus licheniformis* and from *Aspergillus* species.

α-Amylase (E.C. 3.2.2.1) catalyzes the endo-hydrolysis of 1,4-alpha-D-glycosidic linkages in polysaccharides containing three or more 1,4- α-linked glucose units. The enzyme acts on starches, glycogen, and oligosaccharides in a random manner, liberating reducing groups. α-Amylases are characterized by attacking the starch polymer in an endo fashion, randomly cleaving internal α-1,4 bonds to yield shorter, water-soluble dextrins. They are the preferred type of amylase for laundry detergents, and are included in both powder and liquid formulations in many countries. α-Amylases boost overall detergent performance at lower wash temperatures and with milder detergent chemical systems.

Other starch-degrading enzymes of potential interest are β-amylases (E.C. 3.2.1.2) and pullulanase (E.C. 3.2.1.41), but these have not been commercialized on industrial scale for detergent application. β-Amylases differ from α-amylase in that they hydrolyze 1,4- α-D-glycosidic linkages, removing successive maltose units from the nonreducing of the polysaccharide chain. This exo-type activity is not as efficient for stain removal. Pullulanase is a starch-debranching enzyme and hydrolyzes the 1,6- α-D-glycosidic bonds in amylopectin and pullulan, debranching the amylopectin. A mixed amylase and pullulanase enzyme system can accomplish a more complete degradation of starch.

Table 23. lists some application conditions for α-amylases for detergent applications.

Table 23. Commercial amylases (Novozymes) [592], [595], [597], [598]

Product name	Microorganism	pH application range	Temperature application range
Termamyl	GM *Bacillus* species	6 – 11	25 – 100
Duramyl	GM *Bacillus* species	6 – 10	25 – 100
Natalase	GM *Bacillus* species	5 – 10	10 – 60
BAN	*Bacillus* species	5 – 8	15 – 90
Fungamyl	*Aspergillus* species	4 – 7	15 – 60

α-Amylases catalyze the degradation of starch-containing stains, and improve cleaning by hydrolyzing the starchy glue that binds other dirt and stains to fabric. However native starch is only slowly degraded by α-amylases. Gelatinization and swelling are needed to make the starch susceptible to enzymatic breakdown. For most food stains, various degrees of gelatinization result from cooking of starch-containing foodstuffs, e.g., pasta, potato, gravy, and baby food.

If the food processes did not exceed the starch gelatinization temperature, the starch may be in the form of partly or nongelatinized granules, or it may be partly retrograded. Such starch may be amorphous and is usually difficult to remove from surfaces without boil washing. The presence of an amylase renders boil wash superfluous in laundry process and enhances laundry-cleaning performance of detergents at lower wash temperature.

Gelatinized starch may form a film on fabric that can result in an increased pick-up of particulate soil after washing [593]. Starch stains combined with particulate soiling are more difficult to remove than starch alone. As a result white laundry items turn increasingly gray after repeated wash cycles, an effect that has been demonstrated by adding about 0.5 g starch per kilogram of cotton fabric [594]. Starches may react differently depending on their amylose content, which is thought to be the film-forming component of the starch. Film formation is favored under European laundry conditions, in which the temperature may be

closest to the starch gelatinization temperature.

In laundry detergents, amylases may maintain or even contribute to increased whitening of dingy fabrics [593] and inhibit the graying of white fabrics resulting from a combination of starch and particulate soiling, as illustrated in Figure 62. A noticeable amylase effect is obtained with pre-spotting (at high concentration), and with presoaking (prolonged reaction time).

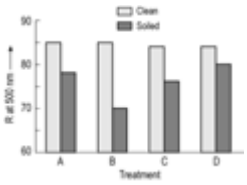

Figure 62. Effects of added starch in solution on particulate soiling and removal with amylase and European granular detergent. Treatments: A) no starch added and washed with amylase-free detergent; B) starch added and washed with amylase-free detergent; C) starch added, washed with 0.4 KNU/L (Kilo Novo Units per Liter) Termamyl and detergent, D) starch added, washed with 1.2 KNU/L Termamyl and detergent [594].
[Full View]

Examples of artificially soiled test pieces used to test the performance of amylases include cocoa/milk/sugar, cocoa/sugar/potato starch, cocoa/milk/sugar/potato starch, and starch/carbon black, all on cotton or polyester/cotton.

Bacterial α-amylases used in laundry detergents are fully compatible with detergent proteases, i.e., the two enzymes work together in the wash process. During storage in both powder and liquid detergents, the amylases are stable in the presence of proteases.

The stability of some calcium-dependent α-amylases can be enhanced by addition of small amounts of calcium salts to the detergent composition.

For bleach-containing detergents, bleach-resistant α-amylases have been developed by using protein engineering and genetic techniques, as indicated above for proteases. Figure 63 illustrates the gain in amylase enzyme stability, as indicated by the improved residual enzyme activity over time in a bleach-containing solution.

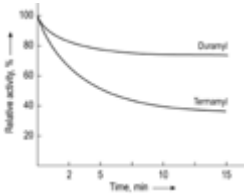

Figure 63. Residual activity of bleach-stable α-amylase (Duramyl) compared to standard amylase (Termamyl) in 100 mM hydrogen peroxide at pH 9.0, 40 °C, 0.1 mM Ca^{2+} [597]
[Full View]

Protein-engineering and genetic techniques have been used to develop α-amylase with lowered optimum temperature for enzyme activity, specifically for low-temperature wash conditions with heavy-duty liquid detergents. Figure 64 illustrates the reduced optimum working temperature for such a low-temperature amylase developed by Novozymes. The significantly improved performance of such low-temperature α-amylase in a cold-water wash (25 °C) in a Mexican powder detergent is shown in Figure 65.

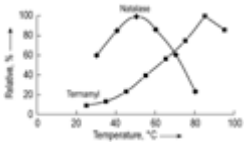

Figure 64. Optimum temperature of low-temperature amylase (Natalase) versus standard (Termamyl) amylase [595], [598]
[Full View]

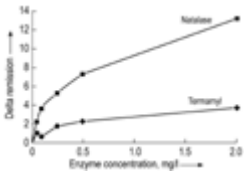

Figure 65. Stain-removal performance of low-temperature amylase (Natalase) versus standard amylase (Termamyl) at 25 °C in Mexican commercial detergent (P&G). Orange-dyed starch swatch, small-scale wash test [595], [598].
[Full View]

5.2.1.6.3. Lipases

Lipases (E.C. 3.1.1.3) hydrolyze triacylglycerol (triglyceride) substrates, as are present in fats and oils. Because of their strong hydrophobicity, fats and oils are difficult to remove from laundry at low temperatures. A triglyceride molecule is composed of three fatty acid moieties linked to a glycerol backbone by ester bonds. Lipases hydrolyze triglycerides to more hydrophilic mono- and diglycerides, free fatty acids, and glycerol. These hydrolysis products are all soluble under alkaline wash conditions.

Many patents have demonstrated that lipases can improve the removal of fatty stains when used in powder and liquid detergents, special pre-soakers, or other cleaning agents. Intense research activity is also reflected in the literature.

Lipolase was the first commercially available lipase for detergents and the first detergent enzyme produced by genetic engineering to achieve an acceptable production economy. Leading brands in the USA and Europe included lipase from 1990/1991 [599] in powder and in liquid detergents.

The slow development of a commercial detergent lipase was due to low fermentation yields and to difficulties in finding lipases with the appropriate characteristics for application in household detergent products. Patent literature indicates the only feasible way to produce lipases at an acceptable cost/performance ratio is by genetic engineering [600-602]. Lipolase was originally isolated from the fungus *Thermomyces lanuginosus* (formerly *Humicola lanuginosa*) with low levels of enzyme expression. By using rDNA techniques the lipase is expressed in acceptable yields in the harmless host microorganism *Aspergillus oryzae* [598, 603] in the current commercial production process.

Protein engineering and genetic techniques delivered improved lipase enzymes for detergent applications. For example Lipolase Ultra by Novozymes is a protein-engineered variant of Lipolase that has improved washing performance at lower temperatures, i.e., below 20 °C. The change made by protein engineering involves replacing a negatively charged amino acid in the lipid-contact zone of the Lipolase molecule. More precisely, aspartic acid (negatively charged) in position 96 has been replaced with leucine (neutral, hydrophobic). The change reduces the repellent electrostatic forces between the soil and the enzyme, making the active site more hydrophobic so that the affinity to a lipid contact zone on the textile surface was improved.

Another protein-engineered variant of Lipolase is LipoPrime for better first-wash effect, especially under European washing conditions. Further efforts are now being made to develop more “first-wash lipases” [604].

Surface-active molecules such as surfactants and fatty acids/soaps can strongly inhibit lipase enzymes [605]. Because of the presence of free fatty acids in the mix of hydrolysis products, pH strongly influences lipase activity and the removal of decomposed stains. The best rate of removal requires a pH above 8 [606]. Above pH 8 the hydrolysis reaction may be favored by small amounts of free Ca ions due to the formation of Ca soap, although lipases are also effective at low concentrations of free calcium [607]. The more the fatty stain is hydrolyzed by the lipase, the easier it is to remove the stain from the fabric due to its increased hydrophilicity.

However the effects of conventional lipases are seen only after several wash cycles (“multi-cycle wash performance”), as illustrated in Figure 66.

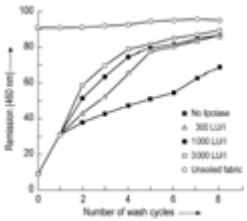

Figure 66. Multi-cycle wash performance of lipase in detergents. Stain removal as function of number of wash cycles. European wash conditions 5 g/L powder detergent, 30 °C, 20 min, wash at pH 9.7. Polyester swatch soiled with lard fat plus Sudan red [610].

LU/L = Lipase Units per Liter
[Full View]

Surprisingly, lipases are also active during a certain period of the drying step [604], [610] during line drying, and in tumble dryers, Lipolase displays maximum activity when the moisture content of the fabric is 20 – 30 wt %. This means that significant decomposition of any residual fatty matter will take place while the laundry is drying. This hydrolytic activity does not result in an immediate advantage in terms of fat removal, but the next time the stained fabric is washed the stain will be removed more effectively. Figure 67 illustrates lipase enzyme activity in the drying step.

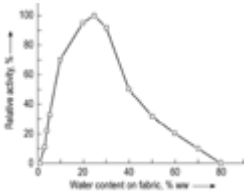

Figure 67. Lipase performance in the laundry drying step. Relative enzyme activity during drying after a European wash and one rinse with tap water [612].
[Full View]

Lipases have proven to be effective in pre-spotters and other liquid detergent formulations when used in undiluted form for pretreatment of tough fatty stains. The low water content on the fabric in this situation is believed to be responsible for the high lipase activity and stain-removal performance in these conditions.

In *performance testing* of lipase detergent enzymes, three consecutive washing/drying cycles are usually employed. Washing temperature, time, and Ca^{2+} concentration depend on the geographical area where the detergent is used. After washing and drying, the reflectance of the colored, lard-containing test swatches is measured, and the remaining lard may be extracted for quantitative determination.

The multiple-wash performance effects of a detergent lipase are shown in Figure 66.

Lipase can cause *malodors* on butter- or fat-stained fabrics by releasing short-chain free fatty acids [605]. Lipases also may indirectly create malodor from laundry items such as sweaty sports clothes. The main reason for malodor (at least in the case of sports clothes) is the survival of microflora. The problem is therefore most pronounced in detergents without bleaching agents. Under normal drying conditions, lipases do not give rise to additional malodor, but if the drying takes several days some lipase-associated malodor may develop. In laundry detergents, the dosage of lipase is often limited by the maximum acceptable malodor.

5.2.1.6.4. Cellulases

Cellulases cleave β-1,4-glycosidic bonds in cellulose and operate directly on the natural cotton fibers or cotton/flax blends and on the cellulose portion in blended cotton/synthetic fibers. This enzyme class is divided into endo-cellulases (endo-glucanase, EG; E.C. 3.2.1.4) and exo-cellulases (cellobiohydrolase, CBH; E.C. 3.2.1.91). EG-type cellulases catalyze the endo-hydrolysis of 1,4- β-D-glycosidic linkages in cellulose, lichenin, and cereal β-D-glucans. CBH-type cellulases hydrolyze the 1,4- β-D-glycosidic link in cellulose and cellotetraose, releasing cellobiose from the nonreducing ends of the cellulose chain. For both EG and CBH cellulases, many iso-enzymes with slightly different substrate specificities and optimum working conditions have been identified.

The cellulase enzyme molecule is composed of up to three types of functionally different domains, as illustrated schematically in Figure 68: (1) the catalytically active core, which is a large, spherical domain; (2) the linker domain, which is an elongated and flexible spacer; and (3) a spherical cellulose-binding domain (CBD).

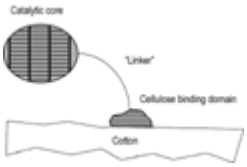

Figure 68. Schematic presentation of a multi-domain cellulase adsorbed to cellulose substrate. The enzyme having a catalytic core, a linker, and a cellulose-binding domain [592].
[Full View]

The nature of the core determines catalytic properties such as endo activity versus exo activity, substrate specificity, and the type of reaction products that are formed. The presence of a CBD is of particular importance for binding of the enzyme on insoluble and crystalline cellulose and for hydrolytic effects. Both EG- and CBH-type cellulases can contain linkers and cellulose-binding domains.

Table 24 lists some application conditions and a few comparative characteristics for some commercially available detergent cellulases from Novozymes. These cellulases can be used in heavy-duty liquid detergents, in bleach-containing powder detergents, and in a number of “color” and compact detergent powders.

Table 24. Examples of commercially available cellulases from Novozymes and application conditions. Source: Novozymes cellulase application sheets

Product name	Microorganisms	Cellulase type(s)	pH range	Temperature range
Celluzyme	*Thermomyces lanuginosus*	CBH 1, CBH II, EG I, EG II, EG III, EG V, and EG VI	4 – 10	25 – 70
Carezyme	*Thermomyces lanuginosus*	EG V	5 – 10.5	25 – 70
Endolase	*Thermomyces lanuginosus*	EG II	5 – 9	25 – 70

As an example the pH and temperature profiles of Carezyme are presented in more detail in Figures 69 and 70.

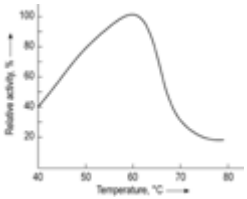

Figure 69. Temperature-dependent activity of EG V cellulase (Carezyme, Novozymes) at solution pH 8.5 and 100 ECU(CP) [Endo Cellulose Unit (Carezyme Product)] enzyme concentration, 20 min. reaction time [616].
[Full View]

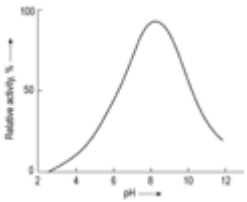

Figure 70. pH-dependent activity of EG V cellulase (Carezyme, Novozymes) at 40 °C and 100 ECU (CP) enzyme concentration, 20 min. reaction time [616].
[Full View]

Cleaning by removal of particulate soils, softening, and improved color brightness are the three basic benefits obtained from cellulases. Cellulases are unique in providing these effects.

When a textile is exposed to shear stress, either during wear or washing and tumble drying, the surface becomes slightly damaged. The worn yarns form fibrils and “pills” ranging in size from a few micrometers to a few millimeters on the surface. The pills scatter light, giving the fabric a grayish or dull appearance, and affecting color brightness and contrast. Dust particles also tend to stick to the areas of damaged cellulose fibers, and this further adds to the gray appearance. Damaged fibers are also thought to be responsible for making the fabric surface more rigid by entanglement, thereby reducing softness and altering size and shape

of a garment.

Cellulases hydrolyze the exposed β-1,4 bonds in the cellulose fibrils and pills, which leads to their removal during the treatment. This is believed to be the mechanism behind the softening and color-brightening effects. Published literature indicates that little is known about the mechanism behind the cleaning action of cellulases. Possible explanations for the cleaning effect are that by removing the fibrils, the soil attached to them is released, and that the enzymatic action facilitates cleaning by exposing dirt trapped in the fiber matrix to the washing solution [613], [614].

Extremely high dosages of detergent cellulases can cause fabric damage in some cotton products after repeated laundering. Damage may appear as loss of fabric strength and excessive softening of the mechanically exposed parts of laundry items, such as hems and edges. These effects may be eliminated by balancing the cellulase dosage to manage the desired benefits. Application tests include small-scale laundering in Terg-o-tometers and Launder-o-meters and full-scale multicycle laundering in commercial washing machines.

5.2.1.6.5. Mannanase

The mannanase of interest is a mannan endo-mannosidase, which degrades the β-1,4-mannose linkage of galactomannans, e.g., in neutral pectins such as guar gum. The hydrolysis products are mannose oligosaccharides. Galactomannans such as guar gums are widely used as thickeners and stabilizers in foods, household agents, cosmetics, and toiletries. Because of high affinity to the cotton surface due to hydrogen bonding effects, stains containing guar gum can be difficult to remove. Guar gum also has a glueing effect on particulate soil and can flocculate dispersed clay particles in the wash. Hence, invisible residual spots of guar gum can bind particulate soil released during the wash cycle, so that the stain re-appears.

Mannaway [37288-54-3] (E.C. 3.2.1.78) is a mannanase for use in detergents commercialized by Novozymes and developed in close cooperation with Procter and Gamble [617]. It was first introduced in Procter and Gamble's HDL Ariel Liquid in 2000. The mannanase in this enzyme product is a 33 kDa hemicellulase (pI 4.7) that was cloned into *Bacillus licheniformis* from alkalophilic *Bacillus* strain I633. The optimal temperature for the enzyme at pH 10 in an analytical system is 50 °C (Fig. 71). In the same system the optimal pH is determined to be 6 – 8 (Fig. 72).

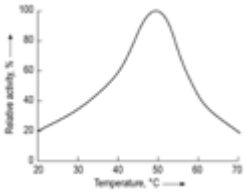

Figure 71. Temperature profile of Mannaway at pH 10. Determined by spectrophotometer at 600 nm, AZCL galactomannan in 0.1 M glycine [618].
[Full View]

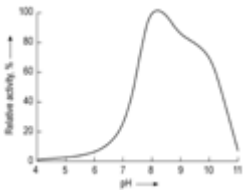

Figure 72. pH profile of Mannoway at 40 °C. Determined by spectrophotometry at 600 nm, AZCL (Azurine Cross Linked) galactomannan in 0.1 M glycine [618].
[Full View]

The enzyme activity of Mannaway is expressed in mannanase immuno units (Mannaway), MIU(M). One unit is equal to one milligram of pure enzyme protein and is defined relative to an enzyme standard under specific conditions. A specific enzyme activity of 1 MIU/M) mg is achieved in the commercial liquid preparation with a total of 25 MIU(M)/g.

Using Mannaway in a detergent formulation thus improves its overall cleaning efficacy as well as its stain-removal effect on stains like BBQ sauce, ice cream, gravy, and fudge.

5.2.1.7. Future Trends

With water becoming an increasingly precious commodity, increased environmental awareness and care for

continued energy conservation will impact consumer washing habits. Detergents will have to become even more efficient and effective for maintaining high performance standards at further reduced dosage using less wash liquor at lower wash temperatures in short wash cycles. In addition altered consumer demands and new fiber developments will require further specialized detergents. All of the above dictate development of smart detergents and wash processes, which certainly will include new enzymes [570].

The contribution of enzymes will continuously increase as more cost-effective enzyme production processes become available as well as new enzyme types that can deliver unique performance benefits in detergents. Increased effort to develop low-allergenicity enzymes, redox enzymes; targeting enzymes, smart enzyme delivery and new improved enzymes from psychrophilic and extremophilic organisms are underway, as indicated by many patent applications. Further progress in protein engineering, genomics, and proteomics will allow the discovery of new and better adapted enzymes for detergent applications.

New hydrolases for broadening the range of difficult to remove stains and for cleaning and whitening benefits are being investigated by the major detergent manufacturers and enzyme suppliers across the world. Many hemicellulases such as xylan-degrading enzymes [619], pectin-degrading enzymes [620], and dextran-degrading enzymes [621] are of interest for specific stain-removal benefits.

Redox enzymes, mainly peroxidases, haloperoxidases, and laccases, and other oxidases are being intensively researched as potential components for novel bleaching systems that can provide bleaching effects at lower washing temperatures than current bleaching systems. This would provide energy savings apart from utilizing less harsh chemicals and also would ensure enhanced fabric care of garments being washed [622].

Until now a peroxidase (Guardzyme, Novozymes) and a laccase have been commercially developed by Novozymes for use in dye-transfer inhibition (DTI) in textile processing in combination, respectively, with a low concentration of hydrogen peroxide and oxygen and with a polyphenolic mediator [623]. Also the potential for removing bleach-sensitive stains in the wash using these peroxidase- and laccase-based enzymatic bleach systems and sanitization effects have been reported in patents from Novozymes, Genencor, Unilever, and Procter and Gamble [624-626]. These enzymatic bleach systems have not been applied in detergents so far.

More recently mono- and dioxygenase enzymes have been investigated for bleaching performance according to patents by Unilever, Procter and Gamble [627], and Novozymes. These enzymes incorporate oxygen directly into the substrate to be oxidized or bleached. The dioxygenases such as lipoxygenases do not require added cofactors or mediators for achieving bleachlike performance effects. However, many, if not all, oxygenase enzymes are intracellularly active and are difficult to express as extracellular enzymes, as are preferred for economical and industrial production.

Also enzymes will start to appear in other household detergents (apart from automatic dishwashing detergents) as developments proceed. Protease enzymes have already been introduced in hand dishwashing detergent by Procter and Gamble for providing skin-care benefits to consumers. Household cleaners also can benefit from enzymes. Many enzyme-susceptible soils can be found on hard surfaces of floors and workbenches in the kitchen (dried-in fats, sticky sugars, dairy products, dust, and so on), and in bathrooms on tiles and bath tubs (calcium soaps, sebum, hair, and others).

5.2.2. Enzymes in Automatic Dishwashing

5.2.2.1. Introduction

Proteases and α-amylases have become state-of-the-art worldwide for automatic dishwashing detergents, one of the last remaining growing markets in the field of detergents. Because modern automatic dishwashing detergents are required to be safe for consumers and to take account of environmental aspects, proteases and α-amylases are now indispensable ingredients. Their incorporation made possible the development of automatic dishwashing detergents with reduced alkalinity, lower dosage, and active-oxygen-based bleach systems, as well as lower application temperatures. The enzymes offered on the market for this application are the same as for laundry detergents. But due to differences in the performance profile some proteases and α-amylases are preferred. Therefore the search for better performing enzymes by microbiological screening or protein engineering, such as site-directed mutagenesis or DNA shuffling, is still going on. Although an official, well-established method for testing the performance of enzymes does not yet exist, there are some methods that are commonly used.

The concept of using enzymes in detergents first appeared in 1913 in German Patent DRP 283 923, assigned to OTTO RÖHM and describing the introduction of a pancreatic protease into a laundry detergent [628]. The first enzymatic product, a soaking detergent launched in Germany under the trade name Burnus, was essentially made of soda ash and pancreatin, a protease mixture, mainly of trypsin and chymotrypsin, obtained from crude extracts of animal pancreatic glands. Since then protease, like surfactants and builders, has become a standard ingredient of detergents. Later, protease was joined by other enzymes such as amylase and lipase. Whereas enzymes had been used successfully in laundry detergents for many years as an aid to remove tough stains, the incorporation in automatic dishwashing detergents (ADD) only began in the early 1990s [629].

Though washing clothes and cleaning dishes are two different tasks, both laundry detergents and dishwashing detergents share similar concerns, such as removal of stains derived mostly from foods such as egg and milk, as well as starch-based soils. The main difference between the two areas and thus the reason for the late use of enzymes in ADD is the formulation and application conditions, which have a strong influence on the performance of enzymes. The development of ADD with reduced alkalinity (pH < 11.5) and lower dosage, the use of active-oxygen-based bleach systems instead of hypochlorite, and the trend to lower temperatures led to the incorporation of first protease and, shortly afterwards, amylase to maintain the high performance level [630].

5.2.2.2. Characteristics of Enzymes for ADDs

Proteases and α-amylases have become state-of-the-art worldwide for automatic dishwashing detergents. Both belong to the group of hydrolases because they degrade their substrates by hydrolysis, using water as co-substrate. The enzymes offered on the market for this application are the same as for laundry detergents [631], but due to differences in the performance profile some proteases and amylases are preferred. Sometimes the use of lipase has been discussed [632], as ADDs contain only a low content of surfactants compared with laundry detergents. Furthermore, enzymes belonging to the group of hemicellulases [633] are also described in patent literature for use in ADDs. However, so far no product containing lipase and/or hemicellulase has been offered on the market.

Natural enzymes which can be found in the stomach, saliva, or pancreas of animals, as well as in some fruits, are not really suitable for detergents because they only work under acidic or neutral conditions. Moreover, they are extremely specific, usually able to degrade only one particular type of bond, e.g., in protein molecules. However, enzymes used in ADDs must fulfill several requirements set by the environment (detergent and cleaning liquor) and by the soil substrate, which consists of a huge variety of foodstuffs and is therefore not very specific.

The currently used enzymes are usually produced by high-performance bacterial strains, conditioned to resist alkaline pH and therefore to produce alkaline-resistant enzymes. The enzymes are extracellular enzymes, secreted by the bacteria into the surrounding medium. Thus, they can be isolated without breaking the cells. This makes the purification process easier and more economical. In case of proteases, these bacterial enzymes are less specific and will degrade almost all kinds of proteins.

Enzymes are sensitive to some ingredients of automatic dishwashing detergents, both during the shelf life of the product and in the cleaning liquor. Anionic surfactants, especially alkylbenzene sulfonates, can degrade enzymes, whereas nonionic surfactants do not destabilize them [634]. Enzymes are also deactivated by oxidizing agents such as hypochlorite, which was the commonly used bleaching agent in ADDs for a long time, and to a lesser extent by hydrogen peroxide, peroxides, and peroxyacids. Application conditions are also important. For instance, any temperature increase favors the enzymatic reaction, as for most chemical reactions, but at the same time the degradation of the enzymes will increase. Heat degradation of proteases occurs above 55 °C in pH 9 buffers and sometimes at lower temperatures when surfactants and builders are present.

To be well suited for the use in ADDs, enzymes must exhibit the following properties:

- Activity optimum at alkaline pH,
- Efficacy at temperatures of 20–70 °C,
- Stability at temperatures up to 60 °C,
- Stability in the presence of other detergent ingredients, such as surfactants, builders (e.g., phosphates, silicates, etc.), and activated bleach, both during storage and use,
- Low specificity to soils, i.e., a specificity broad enough to enable the degradation of a large variety of proteins and starches.

Enzymes are commercially available as dust-free granulates for solid products and as yellowish to brown liquids for liquid products. Because they are biocatalysts, they act at very low concentrations. Their efficacy does not increase linearly with concentration, but levels off above a certain concentration.

5.2.2.3. Proteases

5.2.2.3.1. Proteins: The Substrate of Proteases

A fresh proteinaceous stain on textile or hard surfaces is generally not stubborn and can often be removed simply by cold water. It is only when the proteinaceous soil is dried, aged, or heated, for instance, after cooking, that it becomes difficult to remove, even with surfactants, as it coagulates and hinders the penetration of the cleaning liquor. Moreover, protein residues may be oxidized and thus denatured upon aging due to presence of oxygen or a bleach system as part of an ADD, so that the soil/stain becomes permanent.

Furthermore, there are some stains which contain only small amounts of protein and are not completely removed by surfactants and bleaching systems. Proteins are present in small amounts in many natural oils as well as starchy stains from different foodstuffs, such as gravy or cocoa [635]. In these types of stains, the coagulated proteins act as a glue fixing the soil to the surface.

5.2.2.3.2. Proteases for ADDs

Proteases catalyze the breakdown of large and complex protein molecules into peptides and amino acids (Fig. 73), which are more easily removed by the detergent. According to the site of catalytic hydrolysis of the protein, proteases can be classified in two groups: endopeptidases, which hydrolyze peptide bonds within the protein chain, yielding water-soluble peptides, and exopeptidases, which split terminal peptide bonds only, releasing free amino acids.

Figure 73. Schematic of enzymatic protein hydrolysis
[Full View]

Proteases for detergents are endopeptidases belonging to the class of serine proteases, and within this class to the subtilases (subtilopeptidases [9014-01-1], E.C. 3.4.21.62) [636-640]. Similar to another large class of proteases, the trypsin family, these enzymes all have the same active-site residues, i.e., a serine, a histidine, and an aspartate residue. For example, in subtilisin BPN′ the catalytic triad is formed by Asp-32, His-64, and Ser-221 [641]. Structurally, the positions of the active-site serine and histidine residues are very well conserved, whereas the location of the carboxylate function of the aspartate residue shows more scatter.

Subtilisins have several characteristics that make them well suited for detergents: stability in the presence of detergents, broad specificity for the amino acid sequence of the peptides to be hydrolyzed, high specific activity at alkaline pH, and the possibility for production in industrial amounts. Due to a pH profile ranging from 9 to 11, the subtilisin from *Bacillus lentus* shows superior performance to subtilisin BPN′ or Carlsberg and is thus the preferred protease for solid detergents, both laundry and dishwashing (Table 25). The *Bacillus lentus* subtilisin (same as *Bacillus alcalophilus* and *Bacillus clausii*) consists of 269 amino acid residues, slightly shorter than subtilisin BPN′ and Carlsberg, consisting of 275 and 274 amino acid residues, respectively.

Table 25. Proteases available on the market

Supplier: Trade name	Origin	Optimum pH*	Optimum temperature, °C*	Characteristics*
Genencor:				
Purafect®	*B. lentus*	10	60	highly alkaline
Purafect® OxP	GM Purafect	10	50	highly alkaline, oxidatively stable
Properase®	GM *B. alcalophilus*	11	50	highly alkaline, low temperature
Novozymes:				
Alcalase®	*B. licheniformis*	8–9	60	moderate alkalinity
Savinase®	*B. clausii*	9–11	55	highly alkaline
Everlase®	GM Savinase	10	50	highly alkaline, oxidatively stable
Esperase®	*B. lentus*	9–12	60	highly alkaline

*All data according to supplier.

Modification of enzymatic properties of the subtilisins by protein engineering started as soon as the technique became available in 1984. One objective is to stabilize the enzyme against harsh application conditions such as bleach and detergents. Surface-exposed methionine residues belong to the group of amino acids that are likely to be oxidized by peroxides and peracids. The subtilisins contain a methionine residue next to the active site serine-221. Also other methionine or aromatic residues may be oxidized, leading to destabilization and thus to increased autocatalytic breakdown. Oxidation of the exposed methionine-222 drastically lowers catalytic activity. This methionine residue was replaced by several other residues to give increased oxidative stability, but also relatively poor catalytic power on small peptides as well as protein substrates [642], [643].

The other objective is to improve the performance of the enzyme in general by increasing the turnover on given protein substrates or by improved or modified targeting of the enzyme. Besides site-directed mutagenesis, fusion techniques, e.g., by means of DNA shuffling, have also been described for this purpose [644], [645]. The use of these techniques led to the development of proteases with modified substrate specificity, e.g., higher performance on proteins derived from egg or reduced inhibition by substances present in egg, such as trypsin inhibitor. This was achieved by changing the charge ratio within the catalytic domain [646], [647] or by insertion of at least one amino acid in the active loop site region [648-651]. Both methods resulted in a protease with a modified substrate specificity. The protease according to refs. [646], [647] can be found in European ADD, but is not a commercial product. The other protease has been introduced to the market under the trade name Ovozyme®.

5.2.2.4. Amylases

5.2.2.4.1. Starch: The Substrate of Amylases

Starch (→ Starch) is a carbohydrate food reserve in plants and the major energy source for nonphotosynthetic organisms. It is abundant in nature and is composed of two distinct polysaccharides, amylose and amylopectin. Amylose is essentially a linear, water-soluble polymer consisting of α-1,4-linked D-glucopyranose, and amylopectin is a highly branched, water-insoluble polymer consisting of amylose chains linked at branch points by α-1,6-bonds (Fig. 74). The relative amount of amylose and amylopectin depends on the source of the starch. The major component is usually amylopectin (73 – 86 %) and the minor component is amylose (14 – 27 %). Some cultured variants, such as waxy corn or rice, can deviate from this general ratio of amylopectin and amylose.

Figure 74. Structure of amylose and amylopectin
[Full View]

As a major component of foodstuffs, starch is frequently found in food residues on dishes. Such starch-containing residues are derived from pasta, potatoes, oat meal, and sauces, etc. [652].

Household cooking as the final step of starch modification leads to the formation of gels with different viscosity and solubility in water. When starch is heated in water, it swells when the specific gelatinization temperature is reached. The concomitant water adsorption leads to a viscosity increase and to a change in the chemical and enzymatic degradability of starch. During cooling, the amylose fraction recrystallizes. This process is referred to as retrogradation. Foodstuffs may contain starch both in retrograded and in nongelatinized granular forms.

Before cooking, however, starches for food application are modified to facilitate handling during food manufacturing or to improve the taste and texture of the final product. The simplest starch modification is the separation of amylose and amylopectin. Enriched amylose is used for flexible films, production of capsules for food use, and in sweets. Enriched amylopectin can be used as a shape stabilizer in prepared food such as skinless sausages. Starches can also be modified by chemical and enzymatic methods. Oxidation by hydrogen peroxide or sodium hypochlorite gives starches that are good aids for the production of sweets and suspensions. Phosphorylated starches and enzymatically degraded starches like dextrins are used as emulsifiers, binders, or thickeners for the preparations of mayonnaise, dressings, mustard, gravy, and similar food products.

5.2.2.4.2. Amylases for ADDs

Amylases are specific starch-cleaving enzymes [653], [654]. Amylases degrade starch and related compounds (Fig. 75) by hydrolyzing the α-1,4 and/or α-1,6 glucosidic linkages in an endo- or an exo-acting fashion [655-657]. The variety of enzymes catalyzing such reactions include α-amylases (α-1,4-D-glucan glucanohydrolases [9000-90-2], E.C. 3.2.1.1), β-amylases (α-1,4-D-glucan maltohydrolases, E.C. 3.2.1.2), amyloglucosidases (α-1,4-D-glucan glucohydrolases [9032-08-0], E.C. 3.2.1.3), α-glucosidases (α-D-glucoside glucohydrolases, E.C. 3.2.1.20), pullulanases (α-dextrin 6-glucanohydrolases [9075-68-7], E.C. 3.2.1.41), and isoamylases (glycogen 6-glucanohydrolases, E.C. 3.2.1.68). Each of these enzymes can be distinguished from the others on the basis of reaction specificity.

Figure 75. Schematic of enzymatic starch hydrolysis
[Full View]

α-Amylases catalyze the hydrolysis of the amylose fractions of starch with cleavage of the α-1,4-glycosidic bonds in the interior of the starch chain. In the first step, this endo reaction leads to oligosaccharides and further to short-chain, water-soluble dextrins. The dextrins can be further degraded by β-amylases by splitting maltose from the reducing end of the chain. Additionally, there are amyloglucosidases that release glucose and pullulanases, or isoamylases that degrade starch directly into linear dextrins, for they also attack α-1,6-glycosidic bonds.

Of these different amylases, only α-amylases are used in automatic dishwashing detergents, although recently other carbohydrate-cleaving enzymes (e.g., pullulanases or isoamylases) have also been described for this application [658-661]. Usually, the primary hydrolysis of starch by α-amylases into oligosaccharides

and dextrins is sufficient for easier removal of starch from hard surfaces.

α-Amylases for detergents are obtained from bacteria by fermentation. Submerged cultures of *Bacillus subtilis*, *Bacillus licheniformis*, and *Bacillus amyloliquefaciens* are the common production method. The α-amylase that is predominantly used today for automatic dishwashing detergents is a genetically modified α-amylase produced from *B. licheniformis*, available under different trade names (Table 26). The primary target of the modification of the traditionally used α-amylase was improved resistance against oxidants, especially active-oxygen-based systems. This was achieved by replacement of methionine-197 in the α-amylase [662], [663]. Besides this improvement, performance advantages on particular starch-containing stains led to the preferred use of this modified amylase in ADD.

Table 26. Amylases available on the market

Supplier: Trade name	**Origin**	**Optimum pH***	**Optimum temperature,* °C**	**Characteristics***
Genencor:				
Purastar®	*B. licheniformis*	6–9	75–90	thermostable
Purastar® OxAm	GM Purastar	6–9	75–90	thermostable, oxidatively stable
Novozymes:				
BAN®	*B. amyloliquefaciens*	6–6,5	70–90	
Termamyl®	*B. licheniformis*	6–9	70–90	thermostable
Duramyl®	GM Termamyl	6–9	65–85	thermostable, oxidatively stable
Termamyl® ultra	GM Termamyl			reduced Ca dependancy

*All data according to supplier.

The α-amylases of *B. licheniformis* are more suited for the use in ADDs than the formerly used types from *B. amyloliquefaciens*. The activity optimum of these thermally stable bacterial α-amylases strongly depends on the pH, the temperature, and the presence of calcium ions as stabilizer [664], [665]. The molecular weight is about 58 000. At neutral pH the optimum temperature is near 90 °C. A certain concentration of calcium ions is required to maintain the activity of α-amylases. Calcium(II) stabilizes the enzyme against denaturation and the attack of proteases. In the range of 40 – 60 °C the optimum activity lies in a relatively broad pH range from 5 to 9. In the pH range of 9 – 11, which is relevant for ADD, the activity decreases; however, it is still sufficient to give a good cleaning effect.

The builder environment of detergent applications sometimes causes a problem for calcium-dependent enzymes like amylases [666, 667]. These builders, such as phosphates, citrate, silicates, or chelating agents, form stable complexes with ions of the alkaline earth group such as calcium(II). As a result, the concentration of free calcium in the liquor decreases. If the calcium binders have a higher complex-forming constant than the enzyme they can even deplete the structural calcium ion of the α-amylases to give a destabilized enzyme. The destabilized enzyme is inactivated early in the cleaning cycle and, as a result, the performance decreases. Therefore the development of α-amylases with less calcium dependancy has been started, and the first products have just entered the market (ca. 2002).

Optimization of α-amylases by using modern technics of protein engineering is also under progress [668].

Furthermore, new α-amylases derived from other bacilli than *B. amyloliquefaciens* or *B. licheniformis* are described in the patent literature [669], [670]. Some of these amylases show additional activities, such as cyclodextrin glycosyltransferase (CGTase) activity, besides their amylolytic properties.

5.2.2.5. Other Enzymes

Usually, all enzymes claimed for detergency are also described for the use in ADDs. Consequently, besides protease and amylase, which are state-of-the-art in ADD, other enzymes belonging to the group of hydrolases can be found in the literature for this purpose.

There is a variety of patent applications dealing with the use of esterases, especially lipases, in ADDs [632], [671-674]. Due to a low content of surfactants compared with laundry detergents the improved cleaning of fatty stains by using lipase was the target, especially with regard to lower application temperatures and reduced alkalinity. On the other hand, the use of lipase requires the reformulation of the ADD to avoid the formation of calcium soap (limesoap) on dishware.

The use of hemicellulases for ADD is also mentioned in the patent literature. Hemicellulases such as β-glucanases [633] or mannanases [675] deal with specific parts of starch-containing stains, such as β-glucan or mannan. For instance, β-glucan can be found in food residues with a high content of oatmeal, and mannan is used as binder/thickener in some milk products and special diet food. Therefore, compared to broad-application α-amylase, both hemicellulases can be regarded as enzymes with specific effects.

5.2.2.6. Automatic Dishwashing Detergents

During the 20th century electrical domestic appliances reduced the effort of housework in a dramatic way. The washing machine is considered the most important household appliance, because, in the past, laundry washing was the toughest housework job. However, dishwashing is also very time-consuming as it is done several times a day. In the USA the automatic dishwasher had already become fairly common in the household by the early 1960s, and the penetration of dishwashers exceeded the 50 % mark by about 1990. In Europe, dishwasher penetration differs dramatically from country to country [676]. For example, in northern Europe the penetration is quite high at approximately 60 % in Norway and Sweden. In countries like Switzerland, Germany, and Austria more than 40 % of households own an automatic dishwasher. However, in southern and eastern Europe the penetration is still quite low. Thus, substantial growth has been observed, especially in the last-mentioned regions. Therefore, automatic dishwashing is considered one of the last remaining growing markets for producers of electrical household appliances and ADDs.

5.2.2.6.1. Composition of Automatic Dishwashing Detergents

With a dishwasher cleaning can be done easier and even more economically than by hand [677] [678]. The amount of water necessary for a cleaning program in an automatic dishwasher has been drastically reduced in the last decade from 45 to about 20 L. In the cleaning cycle, water consumption has been reduced from about 10 to 4 – 5 L. On the other hand, the amount of crockery, and therefore the amount of soil, remained unchanged. Today modern European dishwashers consume, for the cleaning of 12 table settings, 18 L of water, 1.6 kWh of electrical energy, 20 – 30 mL of detergent, 2 – 3 mL of rinse aid, and 20 – 30 g of salt. These values for the standard 65 °C program can be lowered by using the 55 or even 45 °C program.

The first automatic dishwashing detergents contained almost 100 % sodium tripolyphosphate. Until the end of the 1980s, sodium tripolyphosphate remained the main ingredient, but in combination with sodium metasilicate. In addition, a chlorine-based bleaching agent such as sodium trichloroisocyanurate was incorporated to achieve an visible removal of bleachable stains like tea.

Modern detergents are required to be safer for consumers and to take concern of environmental aspects (e.g., chlorine bleach, phosphates, energy consumption). In some countries, household products containing “corrosive” metasilicate are no longer allowed or have strong restrictions for sale to the public. Consequently, automatic dishwashing detergents have been undergoing an evolutionary process similar to that of laundry detergents. Pancreas extracts have been described for dishwashing powders, but the use of enzymes was taken into consideration for this product category only after alkali-stable and thermostable enzymes became commercially available in granulated form. Dishwashing detergents based on metasilicates and thus of high alkalinity can very efficiently remove most types of soils that can swell under alkaline conditions. This effect decreases dramatically if the pH of the cleaning solution falls below 11.5, leading to an unacceptable loss in performance. By incorporating enzymes such as α-amylases and proteases into

products with reduced alkalinity, the performance loss can be partially or completely compensated. It can even lead to an improved performance on specific soils, e.g., dried-on starch.

The availability of relatively alkali-stable, granulated, and dust-free enzymes was the basis for the development of a new generation of enzymatic automatic dishwashing detergents which are called “compacts” in contrast to the “conventionals” and whose composition is totally different from that of the conventionals (Table 27).

Table 27. Types of automatic dishwashing formulations (in wt %)

Ingredient	Conventional detergents (powder/tablet)		Compact detergents (powder / tablet)		
				Phosphate	Non-phosphate
Alkaline carriers	metasilicate/disilicate	30–70	sodium carbonate	0–40	0–40
	sodium carbonate	0–10	sodium bicarbonate		0–40
			disilicates	0–40	0–40
Complexing and dispersing agents	phosphate	15–40	phosphate	> 30	
	polymers	0–10	citrate		> 30
			phosphonate	0–2	0–2
			polycarboxylate	0–5	0–15
Bleach system	chlorine carrier	0–2	oxygen carrier	3–20	3–20
			TAED	0–6	0–6
			Mn accelerator		< 1
Wetting agents	surfactants	0–2	surfactants	0–4	0–4
Enzymes				< 6	< 6
Auxiliaries	paraffin oil	ca. 1	perfume	< 0.5	< 0.5
			paraffin oil	< 1	< 1
			silver-protection agent	< 1	< 1
pH of 1 % solution		12–13		< 11	11–12

5.2.2.6.2. Application of Enzymes in ADD

The main task of protease and α-amylase in automatic dishwashing detergents is the cleaning of dishes by removal of tough stains, mainly consisting of proteins and starches, without damaging the surface and/or décor of china, glass, and metal. To evaluate this performance and to differentiate between various enzymes of a class like subtilisins, the producers of enzymes and manufacturers of ADDs apply different test methods for protease and amylase performance, but only a few have been published [629], [630], [679], [680]. Official methods exist in different countries for testing the performance of dishwashers, for example, IEC 436 and DIN 44990, the American AHAM, and the Australian Standard.

Some commonly used methods for testing protease and amylase are based on dried-on protein or starch,

respectively, and are classified in terms of their soil (Table 28):

Table 28. Soils for performance tests of protease and amylase

Enzyme	Soil	Evaluation	Method
Protease	egg yolk	gravimetric	[681]
	egg/milk	gravimetric	[681]
	minced meat on porcelain plates	visual (color reaction)/photo catalogue	[681] by analogy with EN 50242
Amylase	starch mix consisting of potato, maize, rice, and wheat starch	gravimetric	[681] by analogy with IEC SC 59A/WG II
	porridge	visual (color reaction)/photo catalogue	[681] in accordance with EN 50242

By using these soils and their corresponding methods differentiation between different proteases and amylases is possible. The influence of an increasing enzyme content can also be detected. For soils like minced meat on porcelain plates and porridge a minor contribution of the alkalinity must be considered.

The following figures illustrate the efficiency of different proteases and amylases dependant on concentration, given as protein content per liter and temperature. The performance of the different enzymes were tested according to the aforementioned methods (Figs. 76, 77, 78) in a compact phosphate-based European ADD with a dosage of 4 g/L in a 55 °C program at 16° German water hardness (corresponds to 286 ppm $CaCO_3$ with a ratio of Ca:Mg of 5:1). Within the group of proteases and amylases the enzymes were compared on equal protein concentration.

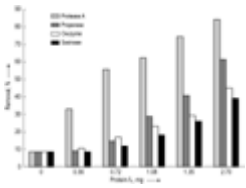

Figure 76. Efficiency of different proteases on egg yolk at 55 °C
[Full View]

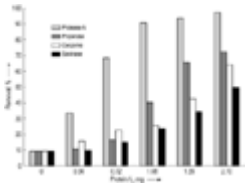

Figure 77. Efficiency of different proteases on egg/milk at 55 °C
[Full View]

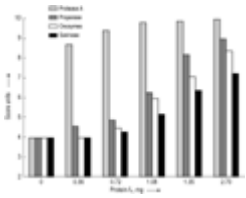

Figure 78. Efficiency of different proteases on minced meat at 55 °C
[Full View]

In Figures 76, 77, and 78 the preference for some proteases is evident. Protease A — a protease in accordance with WO 92/21760 and WO 95/23221 [646], [647] — achieves the best performance in all three criteria compared with the other proteases. Furthermore, the fact that this high performance level is already reached at low protein contents indicates a high turnover. This effect grows even stronger when the temperature is reduced from 55 to 45 °C (Fig. 79).

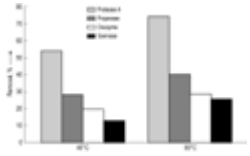

Figure 79. Performance of different proteases (1,35 mg protein / L) on egg yolk at 45 and 55 °C [Full View]

Moreover, the efforts to improve protease performance in ADD by using protein engineering can also be recognized, as three of the four proteases are genetically modified (see Section Proteases for ADDs), and the wild-type protease gives the poorest results.

In the field of amylases the picture does not appear so clear at first glance (Fig. 80 and Fig. 81). But taking all results into consideration there are advantages for the modified α-amylases according to WO 94/02597 [662] and WO 94/18314 [663], which are sold under the trade names Duramyl and Purastar OxAm, respectively. With the starch mix the modified α-amylase convincingly surpasses the wild-type amylase, whereas the superiority of this amylase on porridge only becomes evident with increasing protein content.

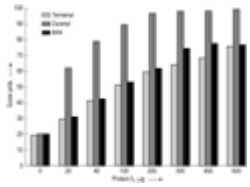

Figure 80. Efficiency of different amylases on starch mix at 55°C [Full View]

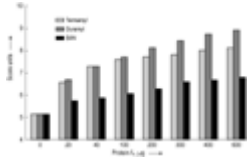

Figure 81. Efficiency of different amylases on porridge at 55 °C [Full View]

Summarizing all these results a clear statement can be given:

- A significant increase of the performance of modern automatic dishwashing detergent is detectable when using protease and α-amylase. This advantage is especially apparent when comparing ADDs with and without enzymes.
- A major contribution to this effect is based on the modification of protease and amylase by protein engineering and thereby improvement of their performance profile with regard to application and requirements in ADDs.

5.2.2.6.3. Stability and Compatibility

In addition to the aforementioned performance aspects, stability during storage is a further important aspect in the application of enzymes in automatic dishwashing detergents. For enzyme granulates, it is influenced by the formulating process of the granulate and of detergent in powder or tablet form. Here, especially the presence of bleaching systems plays an important role. The developments in ADDs, starting with a conventional powder, the one-phase tablet as the next step, and finally the two-phase ADD tablets, which are nowadays the leading ADD product in Europe, makes the separation of enzymes and active oxygen carrier possible and thereby improves the storage stability of enzymes. However, this requires enzyme granulates which are resistant to pressure. A further possibility of this two-phase concept [676] is the time-shifted release of enzymes and bleach, which is also in favor of the enzymes and their performance.

5.2.3. Enzymes in Grain Wet-Milling

5.2.3.1. Introduction

Wet milling of grain, in which the kernels are soaked and then separated into their individual components by grinding and centrifugation, is one of the highest volume processes for converting agricultural raw materials to finished products. The vast bulk of the grain that is wet-milled is maize (corn), and most wet milling occurs in the American Midwest, primarily in the states of Iowa, Illinois, and Indiana, where most of the corn is produced. The main products made in the USA and sold to consumers outside the corn-milling industry, including exports, in 10^6 t/a on a dry or pure basis in 2002 are fuel ethanol (10.4), high-fructose corn syrup (HFCS, 8.3), starch (3.8), and glucose syrup (2.4) [682].

Corn wet milling requires a complex series of processes, as described in Section Overview of the Conversion of Corn to HFCS. Important among them are three enzymatic steps: the hydrolysis of starch to shorter maltooligosaccharide chains (dextrin) with α-amylase, the further hydrolysis of dextrin to glucose with glucoamylase, often with some pullulanase added, and the isomerization of glucose to a near-equilibrium mixture of glucose and fructose with glucose isomerase. Because of the large scale of these processes, the enzymes used in them, along with the proteases and amylases used in synthetic detergents, are produced in tonnages higher than any other enzymes.

Several other enzymes, such as β-amylase and maltogenic α-amylase, are used in various peripheral processes but are not described here.

5.2.3.2. Overview of the Conversion of Corn to HFCS

There are six main steps in the conversion of corn to starch by wet milling and a further five to convert starch to HFCS (Fig. 82):

1. Steeping of corn kernels in a sulfur dioxide solution to soften them,
2. Coarse grinding of the kernels to liberate the germ, which contains oil, from starch, fiber, and protein,
3. Cyclone separation to separate the germ,
4. Fine grinding to liberate starch from fiber and protein,
5. Screening to separate fiber from protein and starch,
6. Centrifugation to separate protein from starch,
7. Hydrolysis of starch with α-amylase to form dextrin,
8. Hydrolysis of dextrin with glucoamylase and often with a little pullulanase to form glucose,
9. Isomerization of glucose to a glucose – fructose mixture with glucose isomerase,
10. Separation of fructose from glucose and other products by liquid chromatography,
11. Blending of the fructose with some of the glucose and byproducts to give HFCS 55 (55 % fructose).

Figure 82. Schematic of the main steps of the corn wet milling process.
[Full View]

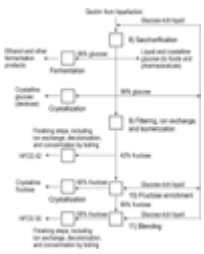

Figure 82. (*continued*) Schematic of the main steps of the corn wet milling process.
[Full View]

5.2.3.2.1. Corn Steeping

Yellow No. 2 corn, the chief grade grown in the U.S. Midwest, is soaked for 24 – 36 h in large tanks filled with a weak aqueous sulfur dioxide solution (Step 1). A fermentation that produces lactic acid occurs in the tanks. In general several tanks are used, with those containing the corn steeped the shortest length of time having the oldest solution, and those having the corn that has been steeped the longest having fresh solution. This is arranged by having each tank hold corn soaked for different lengths of time and by passing the steeping solution from the tank holding corn for the longest time to the tank holding it for the shortest time.

The goal of steeping is to soften the kernels, which originally contain between 10 and 15 % water. Steeping also loosens the components of the kernel, which is composed of the germ, a small oil- and protein-rich structure; the endosperm, the largest part of the kernel, which contains mainly microscopic granules of starch; and the pericarp, or hull, which is composed mainly of fiber and protein. The steeping solution, known as steepwater or corn steep liquor, is a valuable addition to many fermentation processes, since it is rich in vitamins and growth factors.

5.2.3.2.2. Coarse Grinding and Germ Removal by Cyclone Separation
Coarse grinding (Step 2) of the softened corn is designed to break the kernel into many pieces without destroying the germ. The germ is separated in cyclones (Step 3), since it is lighter than the other kernel components. The oil it contains is pressed out and refined, and the remainder of the germ, which is protein-rich, usually is combined with the gluten feed and meal produced in Step 6.

5.2.3.2.3. Fine Grinding and Fiber Removal by Screening
Fine grinding of the remaining contents of the kernel (Step 4) is designed to separate the starch and protein as much as possible from the fiber. Filtering over screens (Step 5) removes fiber, largely the remains of the corn hulls, from starch and protein, since it is in larger pieces. The fiber always contains some starch and a little protein that cannot easily be separated from it. This fraction usually is combined with the gluten feed and meal produced in the next step.

5.2.3.2.4. Centrifugation and Washing to Separate Starch from Protein
The starch and protein fractions, neither of which dissolve in water, can be separated from each other by centrifugation (Step 6). The starch is washed many times by successive water addition followed each time by dewatering in cyclone separators. This brings its purity to above 98 %, the remaining impurities mainly being contained within the starch granules themselves.

The protein, or gluten, fraction is used along with the fiber and that part of the germ from which the oil has been pressed for animal feed, mainly in two different products, gluten meal and gluten feed, of different protein content.

The starch can be dried and sold at this point, to be used as it is or after being modified by chemical processes, or it can be further processed as described below to give dextrin, glucose, ethanol, and HFCS. Some of the starch may be acid-hydrolyzed to make a series of corn syrups. How much of each product is made is at the discretion of the producer.

5.2.3.2.5. Hydrolysis with α-Amylase (Liquefaction)
Starch is contained in microscopic granules and is composed of long chains of glucosyl residues linked by α-(1→4) glycosidic bonds, with branches in its amylopectin fraction being initiated by α-(1→6) glycosidic bonds. Although α-amylases are capable in nature of attacking starch granules, in industrial practice steam under pressure is used to dissolve (liquefy) them as, simultaneously, highly stable bacterial forms of α-amylase hydrolyze α-(1→4)-glycosidic bonds in the now greatly increased number of accessible chains (Step 7). Since α-amylases in general attack interior bonds of these chains faster than those near chain termini, a mixture of products with intermediate numbers of glucosyl residues (dextrin) is formed in preference to glucose or maltose.

The two reasons for using α-amylase are (1) to make the starch chains shorter so that they will not form a gel if the temperature is decreased, and (2) to make more chain ends, since glucoamylase, the enzyme used in the next step, cleaves glucose molecules only from the nonreducing ends of the chains. A more detailed description of this process can be found in Section α-Amylase.

At this point the producer can sell the dextrin as a commercial product after drying or further process it, as described in the following sections.

5.2.3.2.6. Hydrolysis with Glucoamylase (Saccharification)
Glucoamylase, often called amyloglucosidase or AMG in the industry, and often pullulanase are added to dextrin, which is composed of chains of moderate length, mainly between one and twenty glucose molecules, linked by α-(1→4) glycosidic bonds (Step 8). There are occasional branch points initiated by α-(1→6) glycosidic bonds. The glucoamylase breaks the main bonds in the chain, and both glucoamylase and pullulanase break the branch points. This hydrolysis leads to about 96 % yield of glucose, with about 4 % of the original dextrin being found as other products, some coming from the liquefaction step and the majority from the saccharification step.

A more detailed description of the process is given in Sections Glucoamylase and Pullulanase.

The glucose solution is subjected to ultrafiltration or rotary precoat filtration to remove solids, while soluble impurities and salts added during the α-amylase step and formed during the change of pH in the glucoamylase step are removed by carbon and ion exchange, respectively. The producer has a choice at this point to divert the glucose solution to fermentation processes to make ethanol or other products. The glucose solution may also be evaporated and crystallized, giving a solid often sold under the name dextrose. Alternatively it can be converted to HFCS, as described for Step 9 (Section Isomerization with Glucose Isomerase).

The glucose solution diverted for ethanol production is fermented with yeast. The liquid fermentation mixture is distilled successively in three columns, each increasing the ethanol concentration by separating water, yeast cells, and dissolved matter, so that the ethanol finally reaches almost complete purity. This can be mixed with gasoline to make automotive fuel, serve as raw material for a number of industrial products, or used in alcoholic beverages. Corn millers making both HFCS and ethanol have more process flexibility than those making only one or the other.

5.2.3.2.7. Isomerization with Glucose Isomerase
Glucose solution is passed through a column containing glucose isomerase immobilized on porous solid particles (Step 9). Since the enzyme is much more expensive than α-amylase or glucoamylase, it must be immobilized so that it remains in the column to be used for long periods rather than being dissolved in the sugar solution. However, glucose isomerase does lose activity over time, so often several columns are used side-by-side, with the column holding the freshest glucose isomerase, that with the highest activity, having the highest rate of flow through it, and that with the oldest glucose isomerase having the lowest flow rate. This is done so that each column gives about the same fructose yield.

Glucose isomerase converts glucose to fructose by isomerization. The reaction does not yield only fructose, since glucose and fructose have a roughly equimolar equilibrium. Instead the product is a mixture of about 50 – 53 % glucose, 42 – 45 % fructose, and 5 % other products.

A more detailed description of process kinetics and conditions can be found in Section Glucose Isomerase

The producer can either subject this mixture to finishing steps including ion exchange, decolorization, and evaporation to give HFCS 42, or can enrich it to increase its fructose content.

5.2.3.2.8. Fructose Enrichment and Blending
The sugar mixture produced in the previous step can be passed through a column containing strongly acidic ion-exchange particles with Ca^{2+} counterions (Step 10). These separate the fructose from the glucose and other products, with the fructose-rich stream, whose sugars are 90 % or more fructose, exiting the column last. This mixture can be evaporated and crystallized to give pure fructose crystals. That part of the fructose-rich stream that is not subjected to crystallization is sent to the next step. Streams rich in glucose and other products can either be recycled to earlier steps or sent to Step 11.

The high-fructose stream that is not crystallized is blended with streams rich in glucose and other products from Steps 8, 9, and 10 to bring its sugar content to about 55 % fructose (Step 11). Various finishing steps

then make HFCS 55. They include ion exchange to remove salts, decolorization with charcoal, and evaporation of water to bring the sugar concentration from about 45 % to over 70 %, so that the finished product is microbially stable and can be shipped with less water.

HFCS 55 is virtually identical in sweetness, taste, and mouth feel to an equivalent concentration of sucrose, and since under present U.S. trade policy it is nearly always cheaper than sucrose, it has captured over half the U.S. natural sweetener market.

5.2.3.3. α-Amylase

5.2.3.3.1. Origin and Enzymatic Properties

α-Amylases (1,4- α-D-glucan glucanohydrolase, E.C. 3.2.1.1) are found throughout the living world. All forms cleave α-(1→4) glycosidic bonds in starch, glycogen, and maltooligosaccharides to give shorter maltooligosaccharides and little glucose, so α-amylases are classified as endo-hydrolases. Hydrolysis rates are faster when these bonds are located away from chain termini. This is because α-amylase active sites are clefts without barriers at either end, so that substrate chains can be bound with different α-(1→4) glycosidic bonds adjacent to the enzyme cleavage point. Productive binding near chain termini is less likely since fewer glucosyl residues are bound, resulting in lower binding energies and hence lower binding equilibrium constants. Longer maltooligosaccharide chains are hydrolyzed faster than shorter ones, with higher values of v_{max} and lower values of K_M [683-686], for the same reasons.

In general α-amylases can cleave α-(1→4) glycosidic bonds sufficiently distant from an α-(1→6)-initiated branch on a maltooligosaccharide chain and on either side of it, but they cannot attack α-(1→6) glycosidic bonds. With *Bacillus amyloliquefacians* α-amylase, a close relative of the *Bacillus licheniformis* α-amylase presently used in industry, the smallest molecule remaining after prolonged incubation is 6^2- α-maltosylmaltotriose (a maltosyl residue linked through an α-glycosidic bond to the 6-OH group of the middle glucosyl residue of maltotriose) [687]. This indicates that the three α-(1→4) glycosidic bonds immediately adjacent to the branch point are totally resistant to attack by this α-amylase. Shorter, but still lengthy, incubation leaves the hexaoses 6^2- α-maltosylmaltotetraose (linkage through the second glucosyl residue from the reducing end), 6^3- α-maltosylmaltotetraose, and 6^2- α-maltotriosylmaltotriose [688]. Attack on any chain is processive (multiple attack); a single chain is cleaved several times before being released [689]. All α-amylases yield products by the retaining mechanism [690], [691], which means that the anomeric hydroxyl group that is formed upon hydrolysis has the same α configuration as the bond that was hydrolyzed. Ca^{2+} is required for activity and stability.

5.2.3.3.2. Structure

α-Amylases are the most numerous members of glycoside hydrolase Family 13 [690]. There are references to enzymes with α-amylase activity in other families, but they are not of commercial interest, and therefore they are not discussed here. Although most enzymes in Family 13 are hydrolases that cleave α-(1→4) glycosidic bonds or in a few cases α-(1→6) bonds between glucosyl residues (E.C. 3.2.1.–), some are glucosyltransferases (E.C. 2.4.1.–), and one is a trehalose synthase (E.C. 5.4.99.16) [690]. Considering that glycoside hydrolase families are classified on the basis of similarity in amino acid sequences (primary structures), it follows that these different enzymes have arisen through divergent evolution.

Tertiary structures have been determined for α-amylases from the bacteria *B. licheniformis* (Protein Data Bank structures 1BPL [692] and 1VJS (Fig. 83) [693]), *Bacillus* (recently reclassified as *Geobacillus*) *stearothermophilus* (1HVX) [694], *Bacillus subtilis* (1BAG) [695], *Pseudoalteromonas haloplanctis* (1AQM) [696], and *Thermoactinomyces vulgaris* (1JI1) [697]; the fungi *Aspergillus niger* (2AAA) [698] and *Aspergillus oryzae* (7TAA) [699]; the plant *Hordeum vulgare* (1AMY) [700]; the insect *Tenebrio molitor* (1JAE) [701]; and the mammals *Homo sapiens* [both pancreatic (1JXJ) [702] and salivary forms (1HNY) [703]] and *Sus scrofa* (1PPI) [704] (other structures and articles can be accessed elsewhere [690]). Tertiary structures for maltogenic α-amylases (EC 3.2.1.133) from *G. stearothermophilus* and *Thermus* sp. and the maltotetraose-forming exo-amylase (E.C. 3.2.1.60) from *Pseudomonas stutzeri*, as well as from enzymes with less closely related specificities, are also available.

Figure 83. Stereo view of the tertiary structure of unligated *B. licheniformis* α-amylase. Domains B and C are gray, with B to the left and C to the right (1VJS) [693].
[Full View]

All Family 13 α-amylases so far known have three domains: A, the catalytic domain, containing a $(\beta\ \alpha)_8$ barrel structure, as their cores are composed of eight β-strands surrounded by eight α-helices; B, a long loop with a β-strand between β-strand 3 and α-helix 3 of Domain A that binds Ca^{2+}; and C, a Greek key β-sheet. Family 13 members all have aspartate residues as their catalytic bases/nucleophiles and glutamate residues as their catalytic acids/proton donors [690]. The two catalytic residues in hydrolases with retaining mechanisms, such as α-amylases, are approximately 6.5 Å apart, except in very unusual cases.

The *B. licheniformis* α-amylase, the one most used in industry, contains 483 amino acid residues and has a protein molecular mass of 55 268 Da. Its Domain A encompasses residues 1 – 103 and 207 – 396, while Domain B comprises residues 104 – 206. Residues 397 – 483 are in Domain C [693], which in this case is composed of eight β-strands that appear to bind granular starch [693]. A comprehensive account of the alignment of Domains A and B across different species and their evolution can be found elsewhere [705].

Enzyme domains that bind carbohydrates rather than having catalytic activity have been classified by their amino acid sequences into different families of carbohydrate-binding modules, of which four encompass starch-binding domains (SBDs) [690]. α-Amylases from different sources have SBDs in each of the four families, with many, including that from *B. licheniformis*, still unclassified. Those from the bacterial genera *Bacillus*, *Streptomyces*, *Thermoanaerobacter*, and *Thermomonospora* and the fungal genera *Aspergillus* and *Cryptococcus* are members of Family 20, along with SBDs from glucoamylases, cyclomaltodextrin glucanotransferases, β-amylases, maltotetraose-forming exo-amylases, maltogenic α-amylases, and other hydrolases. These domains are found at the C-termini of the catalytic domains. Although no tertiary structure of an α-amylase SBD is available, there are sufficient structures available from other Family 20 members to strongly imply that those from α-amylases are β-sandwich folds composed only of β-strands and associated loops.

Tertiary structures have not yet been determined for any members of the other three SBD families. Family 21 contains SBDs from an α-amylase from the fungus *Lipomyces kononenkoae*, along with those from two glucoamylases and several different protein phosphatases [690]. Family 25 has SBDS from *Bacillus* and *Clostridium* α-amylases and from various β-amylases and maltopentaose- and maltohexaose-forming amylases, while Family 26 has SBDs from α-amylases from the bacterial genera *Bacillus*, *Butyrivibrio*, *Lactobacillus*, and *Streptococcus*, and from maltopentaose- and maltohexaose-forming amylases.

The fact that closely related strains produce α-amylases with SBDs from different families (for instance, different *Bacillus* α-amylases have SBDs from three different families) suggests that the acquisition of SBDs is a relatively recent occurrence, more recent than the evolution of α-amylases themselves.

5.2.3.3.3. Industrial Use

A slurry of starch granules is mixed with *B. licheniformis* or *G. stearothermophilus* α-amylase, and then steam is added through an orifice at high velocity into a tube through which the slurry is flowing so that mixing of the two streams is complete. The mixture is held at about 105 °C and pH 5.6 – 6.5 for several minutes as it passes through the tube. This dissolves the granules and exposes the starch chains to α-amylase, which breaks them into shorter maltooligosaccharide chains. This dextrin mixture is often held in a tank for another hour or more at about 90 °C, sometimes with more α-amylase added.

If glucose or HFCS is the desired final product, then the reaction mixture is hydrolyzed to a DE (dextrose equivalent) value of 10 to 15, where DE is defined as a percentage of reducing power to that of glucose, glucose and starch therefore having DE values of 100 and close to zero, respectively. Hydrolysis to DE < 10 leaves the mixture susceptible to retrogradation, in which the resulting long maltooligosaccharide chains form a pseudocrystalline mass. On the other hand, if the mixture is hydrolyzed for a sufficiently long time at

high temperature so that its DE exceeds 15, various other products, especially maltulose [α-D-glucopyranosyl α-(1→4)-D-fructose], accumulate through base-catalyzed rearrangement of the reducing-end glucosyl residue to a fructosyl residue. The combined rate of acid-, base-, and water-catalyzed glucose rearrangement is lowest at approximately pH 4.5, but the α-amylases used industrially have optimal pHs for activity and thermostability between 6 and 7, so a balance between these two optimal pHs is chosen to best minimize rearrangement and enzyme inactivation and to maximize enzyme activity, so that the reaction time is minimized to reduce later losses of glucose and fructose yield.

If the desired final product is a solid dextrin mixture or a moderate-DE syrup, up to DE 42, then various mixtures of α-amylase and acid are used to achieve this. Total reaction times increase with increasing DE.

5.2.3.4. Glucoamylase

5.2.3.4.1. Origin and Enzymatic Properties

Glucoamylases (1,4- α-D-glucan glucohydrolase, E.C. 3.2.1.3) are found only in bacteria, archaea, and fungi. They attack the nonreducing ends of maltooligosaccharide chains to produce glucose. Glucoamylases can cleave both the α-(1→4) and α-(1→6) glycosidic bonds in these chains, as well as every other α-glycosidic bond between two glucosyl residues except that in α, α-trehalose (α-D-glucopyranosyl α-D-glucopyranoside) [706]. Hydrolysis rates increase with increasing substrate chain lengths [707], [708], but given sufficient time and sufficiently low concentrations of dissolved solids, glucoamylases will completely convert maltooligosaccharides to glucose. Glucoamylases are very slow enzymes; this is compensated by the large amounts produced by many fungi. Their initial product is β-glucose, so they act by an inverting mechanism.

At the high concentrations of dissolved solids that are present in industrial saccharifications, glucoamylases will produce substantial amounts of di-, tri-, and tetrasaccharides by condensation of glucose to the nonreducing ends of glucose and di- and trisaccharides [709]. Although glucoamylases hydrolyze α-(1→4) glycosidic bonds and therefore form them at much higher rates than the cleavage or formation of any other glycosidic bonds, thermodynamics favor formation of α-(1→6) glycosidic bonds because of their one extra degree of freedom. For that reason and because during most of the saccharification glucose is the major component of the reaction mixture, isomaltose [α-D-glucopyranosyl-(1→6)-D-glucose] is the chief condensation product, followed by isomaltotriose [709].

5.2.3.4.2. Structure

Glucoamylases are essentially the sole members of glycoside hydrolase Family 15 [690]. Only three tertiary structures of the catalytic domain have been described, one each from the filamentous fungus *A. niger* (1GAH) [710], [711], the yeast *Saccharomycopsis fibuligera* (1AYX) [712], and the bacterium *Thermoanaerobacterium thermosaccharolyticum* (1LF9) [713]. All glucoamylase catalytic domains have (α, α)$_6$ barrel structures, with six inner α-helices being surrounded by six outer ones [710], [714] (Fig. 84). In eukaryotic forms a peripheral thirteenth α-helix is present. The active site is a conical well, and this explains why glucoamylases act as exo-hydrolases and why they are more active on longer rather than on shorter substrates, since several glucosyl residues at the nonreducing end of the substrate are strongly bound. This also explains why glucoamylases act by a single-attack mechanism and why they are so slow, with each chain being released from the active site after a single glucosyl residue is cleaved from it. All glucoamylases have glutamate residues as both their catalytic nucleophiles/bases and as their catalytic proton donors [690].

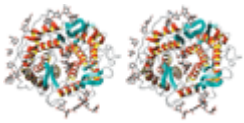

Figure 84. Stereo view of the tertiary structure of the *A. awamori* (now classified as *A. niger*) var. X-100 glucoamylase catalytic domain with acarbose ligand (1GAH) [711].
[Full View]

Most glucoamylases produced by filamentous fungi have SBDs connected to the C-termini of their catalytic domains by O-glycosylated linkers of variable lengths, with O-glycosylation extending into a thirty-residue belt at the C-terminus of the catalytic domain [714]. These SBDs are members of carbohydrate-binding modulus Family 20 [690], as are those of many α-amylases, with a β-sandwich fold composed of β-strands and loops but no helices [715] (Fig. 85). The protein molecular mass of the *A. niger* glucoamylase used

industrially is 65 790 Da, with that of the catalytic domain, residues 1 – 480, being 50 456 Da, that of the linker, residues 481 – 508, 3 454 Da, and that of the SBD, residues 509 – 616, 11 880 Da. N- and O-glycosylation of the enzyme varies among strains, and can add a further 12 – 15 % to the total molecular mass. The catalytic acid/proton donor of *A. niger* glucoamylase is Glu179 [716] and its catalytic base/nucleophile is Glu400 [717].

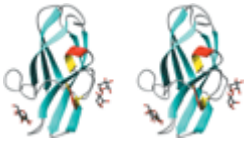

Figure 85. Stereo view of the tertiary structure of the *A. niger* glucoamylase starch-binding domain with two maltose ligands (1KUM) [715].
[Full View]

A. niger glucoamylase used industrially is a mixture of GAI and GAII species, the first containing all 616 amino acid residues, and the second only the first 512 residues. GAI can bind starch granules much better [718], since it contains the SBD, but in industrial practice this confers no advantage, since the two forms encounter soluble dextrin molecules, which they hydrolyze at the same rate [706], [707].

SBDs of glucoamylases produced by the filamentous fungus *Rhizopus oryzae* and the yeast *Arxula adeninivorans* are connected by O-glycosylated linkers to the N-termini of catalytic domains and are members of carbohydrate-binding modulus Family 21 [690], for which no tertiary structures are available. A recently determined tertiary structure of *T. thermosaccharolyticum* glucoamylase has an N-terminal domain of unknown function with eighteen antiparallel β-strands in two β-sheets arranged in a super- β-sandwich [713].

5.2.3.4.3. Industrial Use
Glucoamylase is used industrially to hydrolyze maltooligosaccharide chains of moderate length produced by α-amylase hydrolysis of starch. The enzyme is most stable and active around pH 4.0 – 4.5, and since this is also the pH of highest glucose stability, no compromises need be made to operate there. Glucoamylase is appreciably less stable than α-amylase, rapidly unfolding above 60 °C, so it is used industrially at 55 – 60°C in soluble form, since it is quite inexpensive. This temperature is also sufficiently high to prevent microbial contamination of the reaction mixture.

The saccharification step is conducted over roughly 48 h at 27 – 30 % dry solids and is stopped when glucose concentration reaches its highest value at about 96 % of the total sugar content. Further operation leads to lower glucose yield, since condensation reactions continue until equilibrium is attained. Since these reactions are essentially second-order (first-order each in glucose and the acceptor), while hydrolysis reactions are essentially pseudo-first-order in maltooligosaccharide (the molar concentration of water is very high and is almost invariable), increased concentrations of dissolved solids lead to higher concentrations of condensation products. However, the lower concentrations of dissolved solids that would be necessary to appreciably increase the glucose yield would result in much more water having to be removed from the reaction mixture before shipping, an uneconomical task.

5.2.3.5. Pullulanase
5.2.3.5.1. Origin and Enzymatic Properties
Pullulanase (α-dextrin endo-1,6- α-glucosidase or pullulan 6-glucanohydrolase, E.C. 3.2.1.41) is named for its ability to hydrolyze the α-(1→6) glycosidic bonds in pullulan, a polysaccharide composed of repeating α-D-glucopyranosyl-(1→4)- α-D-glucopyranosyl-(1→4)- α-D-glucopyranosyl-(1→6)- α-D-glucopyranosyl units. However, it also attacks the α-(1→6) glycosidic bonds in β-limit dextrin, the α-(1→6)-bond-rich residue left upon extended hydrolysis of maltooligosaccharides, starch, or amylopectin by β-amylase. The enzyme also attacks glucoamylase-trimmed dextrins. Pullulanase is not effective at even moderately high dosages in hydrolyzing α-(1→6) glycosidic bonds in starch, amylopectin, or in starch hydrolyzed by α-amylase (liquefied starch). The exact chain length of glucoamylase-trimmed dextrins necessary to provide an effective substrate for pullulanase is not known, but based upon known substrates for pullulanase (e.g., β-limit dextrins), it is safe to assume that it is no more than two to four glucosyl residues. The smallest molecule pullulanase can produce when it cleaves α-(1→6) glycosidic bonds is maltose [719]. Pullulanases are found in many bacteria, especially those in the *Bacillus*/*Clostridium* group, in occasional archaea and

yeasts, and in some plants, especially cereals [690].

5.2.3.5.2. Structure
Pullulanases are members of glycoside hydrolase Family 13, which is dominated by α-amylases. This classification is determined by primary sequence homology, and in fact no tertiary structure is available for any pullulanase. However, one would expect pullulanases to resemble α-amylases, with $(\beta, \alpha)_8$ barrel folds, but with the active site modified by divergent evolution so that α-(1→6) rather than α-(1→4) glycosidic bonds in molecules with a number of glucosyl residues would be hydrolyzed. There are reports of Family 20 SBDs being found in various bacterial amylase-pullulanases, but not yet in enzymes with strictly pullulanase function.

5.2.3.5.3. Industrial Use
Glucose yield from the saccharification step is limited by several factors:

1. The liquefaction step produces some materials, such as maltulose, that glucoamylase cannot further cleave;
2. Glucoamylase catalyzes condensation reactions that lead to formation of a number of di-, tri-, and tetrasaccharides, most containing α-(1→6) glycosidic bonds;
3. Glucoamylase rather slowly cleaves the α-(1→6) glycosidic bonds arising from the amylopectin fraction of starch and remaining in the maltooligosaccharide chains arriving from the liquefaction step. The maximal glucose yield occurs when condensation reactions have not yet substantially progressed but when incoming α-(1→6) glycosidic bonds have been largely hydrolyzed.

Pullulanase, since it hydrolyzes α-(1→6) glycosidic bonds in certain molecules containing a number of α-(1→4) bonds faster than does glucoamylase, allows a higher glucose yield than would be achieved by glucoamylase alone. It has no effect on the extent of condensation reactions, since it cannot attack isomaltose or isomaltotriose, the chief condensation products.

The addition of pullulanase to saccharification mixtures does not require a change in pH or temperature. However, its addition may allow higher contents of starch dry solids and shorter saccharification times.

Although the increase in glucose yield using pullulanase in saccharifications is only 0.3 to 0.5 %, pullulanase is widely used due to the competitive nature of commercial dextrose and HFCS production. Pullulanase is usually added to the glucoamylase preparation by the enzyme manufacturer.

5.2.3.6. Glucose Isomerase
5.2.3.6.1. Origin and Enzymatic Properties
A number of enzymes isolated from microbial sources over the past forty years have been called glucose isomerases. Technically however, they are D-xylose isomerases (E.C. 5.3.1.5, D-xylose ketol-isomerase) due to their substantially higher affinity for xylose than for glucose. All known xylose isomerases, in addition to isomerizing D-xylose to D-xylulose, also isomerize D-glucose to D-fructose. Since their only industrial application is the latter, they are referred to as glucose isomerases throughout this section.

Most if not all glucose isomerases require metal ions such as Co^{2+}, Mn^{2+}, Mg^{2+}, or Cr^{2+} for their catalytic activity [720]. A number of inhibitors of the enzyme are known, such as Ca^{2+}, Cu^{2+}, Zn^{2+}, Ni^{2+}, Ag^{+}, sugar alcohols, and various sugar breakdown products. Both Ca^{2+} and the sugar breakdown products are present from previous steps, and it is commercially important to minimize their concentrations by ion exchange and adsorption processes.

The reaction catalyzed by soluble glucose isomerase follows simple reversible Michaelis – Menten kinetics, since the equilibrium state is approximately equal in glucose and fructose concentrations (at an initial glucose concentration of 2 M, equilibrium fructose concentrations expressed as a percentage of glucose plus fructose range from about 48 % at 45 °C to 55 % at 85 °C), and because the Michaelis constants for glucose and fructose are also almost equal (0.25 M and 0.26 M for glucose and fructose, respectively, at 60 °C) [721]. Rate equations have been derived relating rate of fructose formation to rate coefficients, substrate and

enzyme concentrations, and Michaelis constants for fructose and glucose [721]. The soluble *Streptomyces rubiginosus* enzyme is relatively fast, having a turnover number of approximately 7300 min^{-1}.

Kinetic equations have also been derived for use of an immobilized form of glucose isomerase in packed-bed reactors [721]. The equations relate enzyme performance to a reaction rate constant, and express enzyme stability as a function of operational variables such as substrate and product composition, flow rate through the reactor, residence time of the substrate in the reactor, and amount of enzyme in the reactor. A basic working equation can be derived relating glucose isomerase performance (productivity) to measurable operating parameters. Enzyme efficiency can be defined as the weight of fructose syrup solids produced per unit of enzyme when the degree of isomerization (fructose concentration) is held constant and the enzyme is used for a given number of half-lives. Efficiency can therefore be calculated as

$$\mathrm{Eff} = \frac{1.44 k_f \tau \left(1 - e^{-0.693H}\right)}{d \ln \frac{X_{eq} - X_0}{X_{eq} - X}}$$

where k_f is the rate coefficient expressed in terms of g (glucose + fructose) $IU^{-1}\ h^{-1}$, τ is the enzyme half-life in hours, H is the number of half-lives the reactor is on stream, d is the fraction of glucose and fructose in the total sugar solids in the substrate, and X is the fractional conversion of glucose to fructose:

$$X = \frac{[F]}{[F] + [G]}$$

X_{eq} is the fractional conversion at equilibrium, and X_0 is the fractional conversion at $t = 0$ [721].

5.2.3.6.2. Structure

Tertiary structures of many glucose isomerases are now known, including those from *Actinoplanes missouriensis* (1BHW, 1XIM – 9XIM, 1XIN – 3XIN, 5XIN) [722-726], *Arthrobacter* sp. (1XIA, 4XIA, 5XIA, 1XLA – 1XLL) [727-730], *G. stearothermophilus* (1A0D) [731], *Streptomyces albus* (6XIA) [732], *Steptomyces diastaticus* (1CLK, 1QT1) [733], *Streptomyces murinus* (1DXI) [734], *Streptomyces olivochromogenes* (1XYA – 1XYC, 1XYL, 1XYM, 2GYI) [735-737], *S. rubiginosus* (1XIB – 1XIJ, 8XIA, 9XIA, 1XIS – 4XIS; Fig. 86) [738-740], *Thermoanaerobacterium thermosulfurigenes* (1A0C) [731], *Thermotoga neapolitana* (1A0E) [731], *Thermus caldophilus* (1BXC) [741], and *Thermus thermophilus* (1BXB) [741]. So far all have been from bacteria, and it does not appear that glucose isomerases are found in any other part of nature.

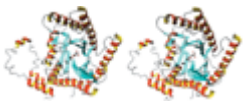

Figure 86. Stereo view of the tertiary structure of *S. rubiginosus* glucose isomerase with glucose, Mg^{2+}, and Mn^{2+}, ligands (1XIF) [739].
[Full View]

Glucose isomerases are members of two different families. Those in Family I have about 390 residues, while members of Family II have about 440 residues. For instance, the industrially important glucose isomerase from *S. rubiginosus* has a monomeric form with 388 residues and a molecular mass of 43 199 Da. All glucose isomerases for which tertiary structures are available, except those from *G. stearothermophilus*, *T. thermosulfurogenes*, and *T. neapolitana*, are members of Family I. This is not an indication that the extra residues confer thermostability, since glucose isomerases of several *Thermus* species belong to Family I.

Most if not all forms of glucose isomerase are homotetramers composed of two tightly bound dimers. Monomers have $(\beta, \alpha)_8$ barrels in which eight β-strands are surrounded by eight α-helices. The active site is in a deep pocket near the C-terminal ends of the β-strands. A C-terminal loop of about 65 residues with strong α-helical character binds to an adjacent subunit. Two divalent cations are required for activity; depending on the form, Co^{2+}, Mg^{2+}, or Mn^{2+} have different effects on activity, as well as on specificity between xylose isomerase and glucose isomerase activities.

The glucose isomerase mechanism consists of ring opening, with a cation hydrogen-bonded to O2 and O4 of the substrate, while the other cation binds to O1 after a substantial change of position. Isomerization occurs through a cation-mediated hydride shift from C1 to C2 and then ring closure to give the α-ketol [724], [725], [730], [737], [740].

5.2.3.6.3. Industrial Use of Glucose Isomerase
As mentioned in Section Isomerization with Glucose Isomerase, enzymatic conversion of glucose syrup to an equilibrium mixture of glucose and fructose is carried out by passing glucose syrup through immobilized glucose isomerase in controlled-temperature reactors.

In the 1970s six immobilized glucose isomerases derived from *Streptomyces*, *Bacillus*, *Actinoplanes*, and *Arthrobacter* were being used commercially in the USA. Immobilization techniques included electrostatic adsorption, cross-linked lysed cells, whole cells entrapped in glutaraldehyde-cross-linked gelatin, flocculated whole cells, and glutaraldehyde-cross-linked whole cells. Today (ca. 2002) there are only four immobilized glucose isomerases, all derived from *Streptomyces*, in use in the USA. Two of the enzymes are immobilized electrostatically on charged carriers (granular DEAE – cellulose and a silica-based material) and two are covalently cross-linked with whole or macerated *Streptomyces* cells by glutaraldehyde. This is, by a large margin, the largest use of immobilized enzyme in the world.

A reactor, usually a large column, is filled with the immobilized enzyme, kept in place by a perforated floor. Glucose solution containing a minimum of 93 % glucose based on total dry solids, which has been exhaustively refined by ion-exchange and carbon adsorption, is adjusted to 40 – 50 % dry solids, pH 7.5 – 7.8, 1.5 mM Mg^{2+}, and about 2 mM HSO_3^-. The solution is passed downwards through the reactor at a predetermined flow rate to produce 42 – 45 % fructose. Commercial glucose isomerases are relatively thermostable, being used at temperatures ranging from 50 to 65°C, and more preferably about 58 °C, to provide optimal enzyme performance and stability and to prevent microbial growth. Typical reactor lifetime is on the order of twelve months, allowing it to produce many tonnes of HFCS.

Use of immobilized glucose isomerase in large reactors today (2003) is relatively problem-free. In the early days of HFCS production there were occasional problems with blinding, channeling, and high pressure drops in the deep-bed reactors due to breakdown of the enzyme particles or to insoluble impurities in the feed syrup. Reactor life could be shortened by improperly refined syrup that allowed inactivators to contact the enzyme, or by temperature or pH excursions. Today, consistent quality of the commercial enzymes and over 30 years of user experience in handling them have greatly reduced the cost of producing HFCS.

After exiting the reactor, the fructose syrup is immediately adjusted to a lower pH in the range of 4 – 5 to prevent unwanted formation of base-catalyzed byproducts and color. The syrup is refined by ion exchange and carbon adsorption before it is fractionated into glucose, oligosaccharides, and fructose.

5.2.3.7. Use of Wheat Starch
Due to various governmental policies, wheat starch rather than corn starch was in the 1990s, and still is, enzymatically converted to syrups in Europe. The amount of syrup made there is small relative to that produced in the USA.

Although wheat is intrinsically more expensive than corn, the production of vital wheat gluten destined for human consumption rather than corn gluten used for animal consumption at least partly compensates for the extra cost. However, to obtain this material, wheat must be dry milled rather than wet milled, as wet milling accompanied by steeping renders the gluten unusable by humans [742].

Starch from dry-milled wheat is obtained by washing it out of a thick flour water dough, leading to a quite pure A fraction and a much less pure B fraction contaminated with a nonstarch polysaccharide, together with a buoyant mixture containing gluten, bran, and fiber [742]. The polysaccharide contributes to high viscosity during starch hydrolysis, so commercial hemicellulases or xylanases may be used to reduce the viscosity. Furthermore, an accompanying lipid, lysophospholecithin, acts as an emulsifier in saccharified starch and prevents a clean break of the insoluble protein/lipid fraction from the soluble glucose fraction. The enzyme lysophospholipase is used, at least in some plants, to break the complex.

5.2.4. Enzymes in Animal Feeds

5.2.4.1. Introduction

The use of enzymes in animal feed is now well accepted throughout the industry. In simple terms it has long been recognized that the addition of enzymes to animal feed will improve the efficiency with which animals digest the raw material.

The largest component of animal feed is in the form of cereal, and depending on the local availability this can be either a viscous cereal (wheat, barley, rye, or triticale) or nonviscous cereal (maize or sorghum). These cereals, with the addition of soybean meal and other protein sources, contain many antinutritional factors (ANFs), which restrict their value to the animal. Although animals produce their own endogenous enzymes, these are not sufficient to digest all the feed; for example, swine are unable to digest approximately one-fifth of their daily ration. Feed supplemented by the addition of commercial enzyme products is used to overcome this problem, and as the understanding of the levels of ANFs present in the feedstuff and the effect they have on the animal performance increases, enzymes can be targeted more effectively.

In general terms there are a wide variety of reasons for using commercial enzyme products in animal feed:

1. To break down the ANFs inherent in the feedstuff, and therefore to improve the performance of the animal, either through a higher weight gain, or better feed utilization.
2. To break down the chemical structure of the cereal base of the diet, and thus release extra nutrients, which will increase the feeding value of the cereal.
3. To increase the availability to the animal of proteins, starches, and minerals from sources that are not usually accessible.
4. To supplement the endogenous enzymes of the young animal.
5. To reduce the variability inherent in feedstuffs, and increase the uniformity of the animals, which in turn increases profitability.
6. To reduce the environmental pollution caused by the excretion of undigested feed, especially phytase, which reduces phosphorus excretion.

Enzymes were first used commercially in animal feed in Finland in the 1980s when a β-glucanase was added to a barley-based feed for poultry, and it has been the poultry industry that has seen the biggest use of feed enzymes. The swine industry is following this trend, and to a lesser extent the ruminant sector. The size of the animal feed production industry means that there is huge potential for enzymes, and the current penetration rate of approximately 10 % of all monogastric feeds [743] shows that there is room for growth. However, in the EU the level of use is already very high, with the majority of poultry feed already containing enzyme supplements.

Enzymes used in animal feed must be able to attack the target substrate relatively quickly, as the passage through the animal's digestive system is rapid, particularly for broiler chickens. The enzymes must also be able to work under the conditions of pH associated with the gastrointestinal tract, as well as being able to withstand the processing regime of the feed before it is fed to the animals. Most feed is pelleted, a process that involves conditioning of the feed with steam, followed by pressing through a die. The temperature at which this process occurs varies between 65 and 100 °C. Enzymes can be added directly to the feed as a dry product, in a premix containing vitamins and minerals, or applied after pelleting as a liquid if the regime of the feed mill is too harsh and may cause a loss of enzyme activity.

This section concentrates on the enzymes currently available (ca. 2003) for use in animal feed, their mode of action, and methods of analysis. There is still some debate as to the exact method by which the enzymes bring about the benefit to the animals, but that is outside the scope of this article. A more detailed treatment of the use of particular enzymes in animal diets is given in recent comprehensive reviews [744], [745].

5.2.4.2. Enzymes Used in Animal Feed

For the purposes of this section the enzymes currently used are discussed on the basis of their target substrates.

5.2.4.2.1. Fiber-Degrading Enzymes

In viscous diets that contain a large proportion of wheat, barley, rye, or triticale the main antinutritive factor is the high fiber content of the diet. Poultry and swine do not have the endogenous enzymes needed to degrade fiber and thus rely on supplementation of their diet. The fiber present consists mainly of nonstarch polysaccharides — arabinoxylan (pentosan) in wheat and rye, and β-glucan in barley and oats — with the levels of each depending on many factors, including environment, temperature, and species.

The nonstarch polysaccharides are antinutritive because in the soluble form they can increase the viscosity of the digesta in the animals' small intestine and thus reduce the degree and rate of nutrient digestion and decrease the animals' performance.

Enzymes targeted against these antinutritional factors formed the initial basis of commercial products, with xylanases aimed at the arabinoxylan backbone of wheat-based diets, and β-glucanases at the β-glucan backbone of barley.

Xylanases. Xylanases or endo-1,4- β-xylanases (E.C. 3.2.1.8) are the most commonly used feed enzymes [746]. Their target substrate is the xylan structure found in plant cell walls (Fig. 87). This consists of a β-1,4-linked D-xylose backbone that has side chains of α-arabinose, and varying degrees of substitution, for example, with ferulic acid, galactose, or coumaric acid [747]. The breakdown products of the xylanase are short-chain sugars.

Figure 87. Schematic xylan structure
[Full View]

The mode of action by which xylanases increase the digestibility of cereals has been the subject of debate. One theory is that xylanases release starch and protein molecules that are held within the xylan structure, and these are then utilized by the animal. The other theory is that the cell wall components are dissolved under conditions present in the digestive tract of poultry and swine, and then form viscous aggregates. This increase in viscosity in the intestine is a problem particularly in poultry [746], and leads to a reduction in digestive efficiency. In this theory xylanase works by reducing the viscosity in the gastrointestinal tract, as the oligomers that are produced by the action of the enzyme are less viscous, and therefore allow an increased rate of digestion.

Xylanases are produced by many bacteria and fungi. The source of xylanases for use in animal feed is mainly fungal, with species of *Trichoderma* and *Aspergillus* being among the most commonly used. *Trichoderma* xylanases have been shown to have optimum temperature ranges between 45 and 65 °C and pH ranges between 3.5 and 6.5 [748]. Bacterial xylanases, mainly from *Bacillus* species, have been investigated and some are available commercially. The catalytic activity of the xylanases differs from source to source, and in vitro data will depend on the substrate used for characterization.

The benefits of adding a xylanase to a wheat-based diet can be seen in Table 29, which demonstrates the addition of a *Trichoderma* xylanase to a wheat-based piglet diet at a rate of 5000 units per kilogram of feed. The results show an increase in the final weight of the piglet, and a reduction in the feed:gain ratio, both of which have large economic implications for the farmer. In broilers, increased weight gain, an improvement in feed conversion, and a reduction in gut viscosity are routinely observed.

Table 29. Piglet trial to show the effects of a xylanase product in wheat-based diets (Finnfeeds International)*

	Wheat control	Wheat + *Trichoderma xylanase*
Initial weight, kg	10.3	10.3
Final weight, kg	22.6	26.6
Daily gain, g	352	466
Daily feed intake, g	724	794
Feed:gain ratio	2.05	1.71

*Pietran x DL pigs were used in the growth trial, with 9 replicates per treatment and 2 pigs/pen. Castrates and gilts were evenly represented. All diets (Table 30) were pelleted (ca. 80 °C) and offered *ad libitum*.

Table 30. Diet composition for trial shown in Table 29

Diet	kg/t
Wheat	400
Rye	200
Barley	200
Soybean meal	100
Fishmeal	70
L-lysine HCl	2
DL-Methionine	1
L-Threonine	1
Tryptophan	0.5
Vitamins and minerals	25
Protein, %	19.9
Total lysine, %	1.12
Digestible lysine %	0.94
Total NSP* (soluble) %	11.3 (6.4)
DE,** MJ/kg	13.6

*NSP = non-starch polysaccharide.
**DE = digestible energy

Assaying feed xylanases is important both in the product and after addition to the feed. The analysis is needed for quality control of the products and to allow traceability through the feed manufacturing process.

As a number of xylanase products are available, there are a number of assays, each with their own definition of a xylanase unit. For product analysis the most common method is the colorimetric dinitrosalicyclic acid (DNS) assay [749] or the dye-linked method of [750]. Within these the pH, temperature, and substrate used all differ. Therefore, it is difficult to compare two xylanases using a standard assay.

Measuring the enzyme after it has been added to the feed and the feed has been processed is more difficult. Any method used will have to overcome the following problems:

1. Low activity in the finished feed. A typical xylanase product (e.g., Avizyme 1302, Finnfeeds International) is added to the feed at a rate of 2.5×10^6 units of xylanase per tonne. When assaying the feed a routine 10 g sample would therefore contain only 25 units.
2. Xylanase inhibitors are known to exist in wheat [751].
3. As the matrix of the feed is complex, binding of the enzyme to the feed components may occur, and specific extraction may be needed to release the xylanase.

One of the most commonly used methods for in-feed analysis is the Xylazyme dyed cross-linked wheat arabinoxylan method [752], which can be measured spectrophotometrically. This method is reported to be more sensitive than assays that use chromogenic substrates, for example, azo-wheat arabinoxylan [750].

β-Glucanases. Barley contains approximately 3 – 4 % of β-glucan, which acts as an antinutritional factor for poultry and swine by increasing intestinal viscosity, and has been shown to increase the incidence of sticky droppings in poultry [753]. It is present in the aleurone layer and the endosperm of barley and oats. The β-glucan consists of linear glucose units linked by β-1,3 and β-1,4 linkages (Fig. 88), which are cleaved by β-glucanases (E.C. 3.2.1.6) to produce free glucose and oligosaccharides.

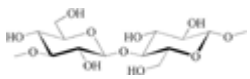

Figure 88. Schematic β-glucan structure
[Full View]

The use of β-glucanase in animal feeds is long established and the benefits are well recorded. In general terms the use of β-glucanase improves the weight gain and the feed efficiency of both poultry and swine fed diets that contain 30 % or more barley. Broiler chickens are the largest target market, with improvements of 50 % in weight gain reported [754]. A typical trial result using β-glucanase in broiler diets is shown in Table 31, with a significant increase in weight gain, and an improvement in the feed:gain ratio.

Table 31. The use of β-glucanase for broiler chickens fed barley-based diets

	Control	+ Avizyme
Weight gain, g	2238	2305
Feed:gain ratio	1.91	1.86
Hock lesions, %	0.72	0.33
Mortality, %	4.7	4.4

Design: A total of 720 one-day-old male Ross chicks were assigned to 12 pens of 60 each, with 6 pen replicates per diet. The pelleted (80 °C) diets (Table 32) were fed *ad libitum*. Mortality and hock lesions were recorded to 42 d.

Table 32. Diet composition for the trial reported in Table 31

Formulation	Starter (0–21 d)	Grower (21–42 d)
Barley	584	608
Soybean meal 48	291	279
Fishmeal 60	25	
Maize oil	59	78
Amino acids	3.8	1.6
Vitamins & minerals	36	32

AME,* MJ/kg	12.5	13.0
Lysine, %	1.31	1.09
Met + cys, %	0.95	0.81

*AME = apparent metabolisable energy

As with xylanases most commercial β-glucanases are either fungal, mainly *Trichoderma* or *Aspergillus* species, or bacterial, usually *Bacillus* species, in origin.

The problems associated with the in-feed analysis of β-glucanase are similar to those of xylanases, with the lower levels of activity used creating further problems with the detection limits. In a commercial β-glucanase product for piglets (Porzyme 8100) the recommended level of inclusion is 600 units of β-glucanase per kilogram, while for broilers it tends to be even lower. The most common assays for β-glucanase in feed utilize dye-labeled substrates, either soluble or insoluble [755], although the problem of sensitivity at lower levels remains. Insoluble dye-labeled substrates are more sensitive and are based on cross-linked polysaccharides.

As with xylanase assays, different companies use different conditions to assay their products, with a range of pH levels, temperatures, and substrates, and different definitions of activity. For example, Finnfeeds International uses pH 5.0 and 30 °C to assay their β-glucanase and defines one unit as the amount that will liberate 1.0 µmol of glucose per minute. Novozymes uses the same definition but the conditions used in their assay are pH 5.5 and 37 °C, while Roche uses pH 5.0 and 40 °C and defines their unit as the amount of β-glucanase needed to liberate 0.10 mmol of glucose per minute.

5.2.4.2.2. Phytic Acid Degrading Enzymes (Phytases)
Phytic acid (*myo*-inositol hexakis-dihydrogenphosphate) is present as the storage form of phosphorus in many of the cereals and legumes fed to animals and acts as the primary source of inositol. The contents vary from plant to plant, with phytate forming 60 % of the total phosphorus in soybean meal, 72 % in corn, and 77 % in wheat [756]. Monogastric animals can only partly digest it as they lack the endogenous enzymes required. This leads to the excretion of the undigested phytic acid, which is responsible for phytate phosphorus pollution [756], most notably water pollution after runoff from agricultural land.

Phosphorus is a vital part of the diet for poultry and swine as it is used in the formation and the maintenance of the skeletal structure and is implicated in many metabolic processes. Therefore, due to the unavailability of the phosphorus in the form of phytic acid, and the effects of a deficient level of phosphorus, including a loss of appetite leading to a low weight gain, leg disorders, and reduced fertility, when poultry and swine diets are formulated they have to be supplemented with inorganic phosphorus sources. However, the level of phosphorus in the diet then often exceeds the amount needed by the animal due to the estimate of the availability of the plant phosphorus. Therefore much of the additional phosphorus is also excreted by the animals, again leading to increased environmental pollution.

Another disadvantage of high levels of phytic acid in the diet is its ability to chelate potassium, calcium, copper, zinc, and magnesium [757], thus decreasing their availability to the animal and making phytic acid an ANF.

As poultry and swine have very little endogenous phytase activity, supplementation of the feed with phytase is now widely accepted within the animal feed industry and it is currently the fastest growing product sector. The addition of phytase makes the phosphorus stored as phytic acid available, and thus it helps to reduce the amount of phosphorus excreted and is thought to increase the nutrient availability from the feed. By replacing added inorganic phosphorus it also achieves an economic benefit for the animal feed producer.

Phytase, also known as *myo*-inositol hexaphosphate phosphohydrolase, acts by selectively removing orthophosphate from phytic acid, thus producing inositol phosphate intermediates, and *myo*-inositol. Phytases occur in two forms, either as 3-phytase (E.C. 3.1.3.8), with which the phosphate group at position C3 is the first to be hydrolyzed, or 6-phytase (3.1.3.26), which begins by hydrolyzing the phosphate group at

the C6 position. They both belong to the histidine acid phosphatase subclass of phosphatases, and as such have a histidine residue in the highly conserved active center. The histidine residue is involved in nucleophilic attack on the phytate phosphorus atom, while the overall reaction yields inorganic phosphate that can then be utilized by the animal.

In 1984 the Dutch government introduced legislation that restricted the application of phosphorus per hectare of land, either from bought sources or home-produced. Failure to comply with the regulations resulted in a fine. This lead has been followed by various European countries and U.S. states, producing a large market for phytase products.

The improvement in phytate phosphorus utilization has been estimated to be between 20 and 45 % [756] in a review of the available literature. The absorption of zinc, copper, iron, and magnesium has also been shown to increase in pigs fed phytase-supplemented feed [758].

In nature phytases are found in plants, animals, and microorganisms. The main source of commercially available phytase is fungal, although bacterial and yeast phytase have also been identified. Currently phytases from *Aspergillus* species account for the largest phytase market and are the best characterized phytases [759], and the majority of the literature on the use of phytase in animal feed is concerned with them.

The *assay for phytase* in the feed is among the most consistent between the different commercial producers. It is based on the method of [760] and is a colorimetric assay. The assay involves the incubation of the phytase with its substrate, sodium phytate, and the subsequent release of inorganic phosphate. This will then produce a yellow complex with an acid molybdate vanadate reagent, which can be measured spectrophotometrically at 415 nm. The main variation in the assay between producers is in terms of pH, with the optimum pH for the specific phytase being used: pH 5.5 (Novozymes and Roche) or pH 5.0 (Röhm).

5.2.4.2.3. Protein-Degrading Enzymes (Proteases)
Aside from the cereals used in animal feed the next largest component of the diets is the protein source. This can be vegetable or animal, and as with the cereals is of variable quality.

A number of proteases have been developed for animal feed, all of which are members of the peptidase family of enzymes (E.C. 3.4) and act on peptide bonds. The reactions of the proteases differ in their specificity for the peptide bonds. The most commonly used protease, subtilisin (E.C. 3.4.21.62), a serine endopeptidase produced by *Bacillus* species, has broad specificity and hydrolyzes peptide amides. Other proteases that are available are bacillolysin (E.C. 3.4.24.28), again produced by *Bacillus* species, but having a more specific reaction, preferring peptides that contain leucine or phenylalanine residues. Aspergillopepsin (E.C. 3.4.23.18), produced by species of *Aspergillus*, and oryzin (E.C. 3.4.21.63), also produced from fungal sources, both have broad specificity, but the latter does not hydrolyze peptide amines.

Generally the proteases are added to feed in combination with other enzymes in order to overcome antinutritional factors associated with the protein sources in the diet. The main protein source tends to be soybean meal, which contains lectins and protease inhibitors. These ANFs can reduce the digestion of the nutrients, and both compounds have also been implicated in physiological damage [761]. The proteases can break down the protein sources into smaller units that are easier for the animal to digest.

Proteases act by breaking down protein into its constituent peptides and amino acids. The animal can then more readily absorb these smaller molecules. It is not used as a single addition to animal feeds, but in combination with xylanase or β-glucanase for use in viscous diets, or with xylanase and amylase for use in nonviscous diets.

Due to the high concentration of protease inhibitors in cereals, especially wheat, the *analysis of protease* in feed preparations is very difficult. The inhibitors must be separated from the enzyme, and conditions for this are by necessity very harsh and can result in the denaturation of the protease. Commercial substrates are available for in-feed analysis, but these have proved to be very variable, and the results are not reproducible. New methods that overcome the problem of protease inhibitors still need to be developed.

5.2.4.2.4. Starch-Degrading Enzymes (Amylases)

α-Amylases (1,4- α-D-glucan glucanohydrolase; E.C. 3.2.1.1) are used in maize-based diets for poultry in combination with other enzymes to increase starch digestibility. Although maize is widely regarded as a highly digestible ingredient, evidence is emerging that there is great variability between different strains, and that starch digestibility is lower than previously thought [762]. The use of amylase helps to overcome these differences. Amylase is also an accepted addition to the diets of young pigs to help them adjust to a new feeding regime at weaning.

A typical broiler trial is shown in Table 33, in which the use of an enzyme product that contains amylase, protease, and xylanase (Avizyme 1500, Finnfeeds International) with a maize-based diet, gives an increased weight gain and an improved feed conversion ratio.

Table 33. The use of amylase, protease, and xylanase in broiler diets containing maize (Finnfeeds International)*

		Control + xylanase, amylase, protease
Weight gain, g	1702	1901
Feed intake, g	3557	3525
Feed:gain ratio	2.09	1.86

*A total of 1280 (sex-separated) Ross broilers were assigned to 16 pens of 80 chicks per pen, with 8 pen replicates (4 males, 4 females) per diet. Mash diets were fed for *ad libitum* consumption, the starter from 1 to 21, and the finisher from 22 to 42 d (Table 34).

Table 34. Diet composition for the trial reported in Table 33

Formulation	Starter (1 to 21 d)	Finisher (22 to 42 d)
Corn	570	606
Soybean meal-48	295	261
Fishmeal Peru-63	30	20
Wheat middlings, 15 % CP	50	50
Tallow	17	26
Vitamins & minerals	38	37
Avizyme 1500	0/1	0/1
AME,* MJ/kg	12.4	12.8
Crude protein, %	21.0	19.0
Lysine, %	1.21	1.07
Met & cys, %	0.90	0.82

*AME = apparent metabolisable energy.

The main source of commercially used α-amylases in animal feed are *Bacillus* species. Amylases act by breaking down the starch present in the diet to produce dextrins and sugars. They catalyze the endohydrolysis of 1,4- α-D-glycosidic linkages in which three or more glucose units are present.

The assay for α-amylase is based on the release of reducing sugars from a starch substrate. For *in-feed analysis* a colorimetric test is employed using insoluble dyed starch polymers as substrates.

5.2.4.3. Future Developments

The future developments in the feed enzyme area are likely to include a focus on the increased temperature stability needed to combat the processing regimes that feeds encounter. At present liquid application can be used to overcome the problems of a loss in enzyme activity, but this is expensive as it requires specialist equipment, and is not practical in many feed mills.

The major disadvantage for feed enzymes at present is the variability in the substrate and the interaction of the enzymes with other feed ingredients, both of which lead to a variation in the response. By increasing our understanding of the substrates, the enzymes can be targeted more successfully. Methods are available that optimize the enzyme dose in wheat-based diets, and these are also required for use with corn-based diets and for phytase dosing.

The production of feed enzymes is also an area in which new developments are to be found. At present the enzymes are produced commercially by traditional fermentation methods, through expression in a host system, usually bacterial or fungal. Researchers are looking at the possibility of expressing the enzymes in plants, although it is questionable whether such products would be accepted in the EU due to the current climate surrounding genetically modified organisms. If the enzymes were produced at sufficient levels of expression in plants already used in animal feed, then the costs of downstream processing compared to traditional methods would be greatly reduced.

To date phytase has been the enzyme of choice for research into production of enzymes from transgenic plants, although none have been commercialized. Plants investigated have included soybeans [763], wheat [764], tobacco [765], alfalfa, and oil seed rape.

Finally, at present the poultry and swine industries dominate the market but there is an increased profile of enzymes for use in the ruminant and aquatic sectors, and these markets have a potential for growth.

5.2.5. Enzymes in Textile Production

5.2.5.1. Introduction

The use of biotechnology in textile processing has been known for more than 2000 years. The first application known is the retting of bast fibers with the use of microorganisms [766]. Other early examples are the removal of starch by soaking starch-sized cloth with water liquor containing barley (1857) and the use of amylases in the same desizing process in 1912 [767]. Since ca. 1990, the use of enzyme technology has increased substantially, especially in the processing of natural fibers. A major reason for embracing this technology is the fact that application of enzymes is regarded as environmentally friendly and the reactions catalyzed are very specific with a focussed performance as a consequence. In contrast, chemical processes are less specific and often result in side effects, mostly undesired, for example, a reduction in the degree of polymerization of cellulose. Other potential benefits of enzyme technology include cost reduction, energy and water savings, improved product quality and potential process integration. Application and potential of enzyme technology have been reviewed [768-772]. Table 35 gives an overview of enzymes used in textile processing. Applications treated in this chapter have not all been implemented on an industrial scale yet (ca. 2002), but this is expected within the next five years.

Table 35. Enzymes used in textile processing. Applications mentioned are industrially implemented or in the research phase

Enzyme	Substrate	Application	Technical benefit	Performance
Amylase	amylose	desizing cotton	increased removal of starch	removal of starch from fiber
Cellulase	cellulose	1. biopolishing cotton	removal of fuzz	maintaining the new look
		2. defibrillation Lyocell	removal of microfibers	improve processibility/generate peach-skin feel

		3. stone washing of denim	removal of indigo selectively	creation of ageing effects
Pectinase	pectin	1. scouring cotton	destabilization of outer cell layer	removal of noncellulosics
		2. retting bast fibers	destabilization of outer fiber layer	improvement of fiber extraction
Catalase	hydrogen peroxide	bleaching	peroxide breakdown	neutralisation of bleaching agent
Laccase	mediator	1. bleaching	creation of bleaching agent	improvement of whiteness
	indigo dye	2. bleaching dye	degradation of chromophore	ageing of denim
	mediator	3. effluent treatment	oxidation of dyes	decolorization of effluent
Glucose oxidase	glucose	bleaching	creation of bleaching agent	improvement of whiteness
Protease	protein	1. scouring wool	removal of scales	improve anti-shrinkage
	sericin protein	2. degumming silk	degradation of sericin	removal of sericin
	cellulase protein	3. stone washing of denim	prevent indigo binding via protein	reduction of backstaining
Hemi-cellulase	hemi-cellulose	retting bast fibers	destabilization of outer fiber layer	improvement of fiber extraction
Xylanase	lignin	processing jute	degradation of lignin	improvement of bleachibility

5.2.5.2. Cellulose Fibers

5.2.5.2.1. Desizing of Cotton Cellulose Fibers

The first application of enzymes in textile processing was the desizing of cotton fabric. Prior to weaving, the warp yarn is sized to strengthen the yarn and prevent breakage in the weaving process. The natural sizes, which are of starch origin, can be removed after enzymatic degradation with α-amylases (E.C. 3.2.1.1) by simply washing out the water-soluble products. Besides starch, several synthetic polymers are available for strengthening the warp yarn, for example, poly(vinyl alcohol), PVA, and poly(vinyl pyrolidone), PVP. Nevertheless, starch is still the favored sizing agent for economic reasons [773], and α-amylase is preferred for desizing, since the other available processes require harsh chemicals such as acids and oxidizing agents. The enzymes used are mainly of bacterial origin, especially from *Bacillus subtilis*. Their pH optimum is between 5 and 7, while the optimal operating temperature ranges from around 65 °C, for natural, unmodified enzymes to beyond 100 °C for engineered enzymes. The stability of those enzymes is improved when some calcium is present in the application system (ppm level).

5.2.5.2.2. Scouring of Cotton

Before gray cotton fabric can be dyed and finished in water-based processes, it must be cleaned in order to make it hydrophilic and white. The convential industrial method includes a treatment with sodium hydroxide at high temperature followed by extensive washing. In this process hydrophobic (waxes and fats) and other noncellulosic compounds (pectin, protein, and organic acids) are removed in a nonselective way. Much research has focused on replacing this process by an enzymatic one. Enzymes such as pectinases [lyases, E.C. 4.2.2.2; polygalacturonases, E.C. 3.2.1.15 (endo-acting type) and E.C. 3.2.1.67 (exo-acting type)], proteases (E.C. 3.4.21 – 25), cellulases (endoglucanases, E.C. 3.2.1.4; cellobiohydrolases, E.C. 3.2.1.91; xylanases, E.C. 3.2.1.8), and lipases (E.C. 3.1.1.3) have been examined for degradation and subsequent

removal of the different natural components present in the outer layer of cotton fibers [774-780]. The best results have been obtained with a pectinase or a pectinase in combination with a cellulase. A bacterial alkaline pectinase (a lyase, E.C. 4.2.2.2) [781] proved to be especially effective. It is thought that pectin acts as cement in the primary cell wall of cotton fibers. After enzymatic destabilization of the pectin structure, the different components present in the primary cell wall layer can be removed easily in a subsequent washing procedure. Bioscouring processes, integrated in the desizing process in batch mode, have been performed successfully on an industrial scale [782, 783]. However, the bioscouring process based on pectinase does not remove motes or seed coat fragments, whereas the nonspecific alkaline scouring process removes these fragments. This may be a benefit when a “natural look” is desired, but is a problem when dyeing in lighter shades. Enzymatic solutions for this problem are currently (ca. 2001) under investigation using cellulases [784, 785].

5.2.5.2.3. Bleaching of Cotton

Scoured cotton fabric must be bleached prior to dyeing, especially with lighter shades. This is conventually performed with chlorinating agents or hydrogen peroxide. Alternative enzymatic processes are being studied. One process is the enzymatic generation of hydrogen peroxide by using a glucose oxidase (E.C. 1.1.3.4), after which the bath is made alkaline for bleaching the fabric. This so-called indirect enzymatic bleaching process makes use of the glucose produced in the desizing bath when an α-amylase is used for starch degradation [786, 787]. This process of peroxide generation requires oxygen as reactant, so aeration is required in the process. Benefits of this enzymatic process are:

- Consumption of the glucose generated in the desizing bath results in reduced BOD in wastewater treatment
- The gluconic acid produced in this process is a good chelator and can be used as a sequestering agent for metal ions
- The process can be integrated easily in the enzymatic desizing and scouring process to form a single integrated cotton pretreatment process

The integrated enzymatic desizing, scouring, and bleaching process yields products with acceptable levels of hydrophilicity and whiteness [788], and a considerable reduction in the required quantity of rinse water has been realized. The glucose oxidase enzyme has been immobilized in order to make repeated use possible [789].

Attention is also focused on a direct enzymatic bleaching process using oxidoreductases (E.C. 1.10.3.2) and different mediators. Much work has been devoted to the development of a system in which the oxidative power of oxygen or peroxide is increased. However, so far an industrially feasible enzyme-based bleaching process is not available, although the technology is steadily improving [790].

5.2.5.2.4. Removal of Hydrogen Peroxide

After the bleaching process, excess peroxide must be removed prior to dyeing to prevent oxidation of the dye. Conventionally this is realized by washing the fabric with large amounts of water or by adding a reducing agent [791]. Catalases (E.C. 1.11.1.6) have been successfully introduced in this process phase. The excess peroxide is rapidly and effectively decomposed into oxygen and water. This method reduces both process time and the consumption of water and energy [792]. An additional advantage is that the subsequent dyeing process can be performed in the same bath when production is carried out in the batch mode. A case study [793] on an industrial scale in an Egyptian textile mill revealed that applying catalase technology achieved the following reductions:

- Energy consumption by 24 %
- Costs for chemicals by 83 %
- Water consumption by 50 %
- Processing time by 33 %

5.2.5.2.5. Cotton Finishing

Cellulases are used to create different effects on cotton. This class of enzymes can be used for introducing an aged look (denim) or a renewed look on fabrics. These apparent contradictionary performances can be

realized by using the appropriate type of cellulase enzyme(s) in combination with a certain degree of agitation or shear force during application [794-799]. In a cotton fabric short (micro)fibers protrude from the fabric surface (fuzz). When these fiber ends become entangled, pills are formed. The process of renewing the look of cotton, known as biopolishing [800], [801] or depilling, is based on the removal of fuzz and pills to give an improved surface appearance [771], [802] which can be described as:

- Cleaner surface with a cooler feel
- Increased softness and improved drapebility
- Brighter colors (color revival or disappearance of the greyish look)
- Improved resistance to fuzzing and pilling

Additionally, a high dyeing yield is claimed [803], although a later study revealed no significant impact on uptake of Green 26 and Red 81 dyes [804]. This study also demonstrated only limited impact of cellulase action on cellulose pore size and volume. The biopolishing process has been introduced on industrial scale in the batch mode only, because of the high shear force required in the process and the lack thereof in continuous processing equipment. Besides the above-mentioned benefits of this enzymatic biopolishing process, a major disadvantage is the potential tensile strength loss (TSL). Much research has focused on minimizing TSL. It was found that use of different individual cellulases (endoglucanases and cellobiohydrolases) leads to different TSL values [805]. This observation is the key to the assumption that using the right mixture of different cellulases (especially certain endoglucanases) will lead to optimal biopolishing performance with a minimum TSL [806-809].

Since cellulases degrade cellulose, application of this type of enzyme on cotton fabric poses a risk of destroying the fabric itself. Therefore, attention has focused on the elucidation of the performance mechanism, especially in relation with the required shear force. It was suggested that high shear forces lead to higher rates of desorption of endoglucanases and hence more random cellulose degradation and shorter processing times for mechanical breakdown of the weakened fibers [699], [810]. Another suggestion is that cellulases first attack the cellulose in the microfibrils protruding from the surface of the fabric because they are more externally exposed than the cellulose of the original fabric structure [811]. Mechanical breakdown of the enzyme-weakened microfibrils can take place before the cellulose in the main fabric has been degraded, and thus the TSL is limited. A more plausible explanation for the relation between enzymatic cellulose degradation and shear force is based on the cellulose structure [812]. Cotton contains both crystalline and amorphous cellulose distributed in a homogenous fashion [813]. The amorphous structure is more flexible due to its lower ordering and water binding [814], while the crystalline structure is rather rigid and therefore sensitive to physical stress. Microfibrils protruding from the fabric surface have a damaged, and therefore amorphous, flexible cellulose structure at the point where they emerge from the fabric surface. Since cellulase activity is highest on amorphous cellulose, all amorphous cellulose will be degraded preferably, including the amorphous fraction at the connection point of the emerging microfibril. The consequence of this cellulase degradation is that crystalline cellulose remains mainly unmodified, and thus the connection point becomes a rather rigid physical structure. Because this cellulase-modified rigid microfibril protrudes from the fabric surface, it is most sensitive to shear force, which breaks it from the fabric (see Fig. 89). Based on this mechanism, guidelines have been published for minimizing TSL in the biopolishing process [815].

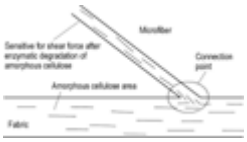

Figure 89. Schematic presentation of a microfiber emerging from the planar surface as side-view. [Full View]

5.2.5.2.6. Ageing of Denim

Today, the old fashioned pumice stone process to obtain a well-worn appearance on denim jeans has largely been replaced by a cellulase process. In this enzymatic process, in which the shear force also plays an important role, both soluble (less than six residues) and insoluble sugars are produced with indigo dye trapped inside the fiber [816, 817]. This enzymatic process can be used to create a worn appearance of textile goods dyed with surface dyes like indigo, sulfur dyes, pigments, and vat dyes [768]. Different

cellulases have been examined for performance, and an endoglucanase is the best performing enzyme. Especially the EGII from *Trichoderma reesei* is most effective, both in removing color from denim and producing a satisfying stone-washed effect with the lowest degree of hydrolysis [818]. Preference for this enzymatic process is due to the fact that the use of pumice stones is eliminated or reduced, thus protecting the machines and avoiding the occurrence of pumice dust in the laundry environment.

A major drawback of the enzymatic process is backstaining, that is, the redeposition of indigo dye on the undyed white weft yarns during washing. This results in an undesired lower contrast between warp and weft yarns. A recent study showed that the adsorbed indigo can be washed out to an extent of about 80 % at maximum by using chemicals such as linear ethoxylated fatty acids and modified poly(acrylic acids) [819]. An enzymatic solution for this backstaining problem has also been developed. Adsorption of the indigo dye proceeds via enzyme binding to the fabric [820, 821]. Therefore, the use of cellulases with weak fabric-binding properties and a low affinity for indigo reduces the problem of backstaining substantially [670, 820]. Also the addition of a protease is of benefit in reducing this drawback [822].

In 2001 a new enzymatic process for indigo dye bleaching was introduced to the market, based on the use of a laccase (E.C. 1.10.3.2) and mediator [823, 824]. In this bleaching process, indigo — both attached to the fabric as well as in dispersion in water — is oxidized to soluble and colorless indigo degradation products [825]. In this way, an attractive bleaching effect with improved contrast is produced, while maintaining the original ecru color of the weft yarns. The overall look of laccase-treated denim is therefore darker than that obtained by bleaching with NaOCl. In this nonspecific chemical bleaching process, the weft yarns are also bleached. The following performance benefits of this enzymatic process over the conventional chemical bleaching process are reported:

- Creation of new looks, fashions, and finishes (e.g., grey cast/antique finishes)
- Enhancement of denim abrasion
- Maximal strength retention at high abrasion levels (maintenance of cellulose structure)
- Prevention of backstaining [826] and therefore improved garment contrast

No information is available on the penetration of this new process into the denim market.

5.2.5.2.7. Processing of Man-Made Cellulose Fibers

Man-made cellulose fibers are obtained from wood. A relatively new fiber is Lyocell (→ Blood – Plasma Proteins), which entered the market in the late 1990s [827]. Total production of Lyocell fibers in 1998 was already over 100 000 t. In contrast to the production of viscose, for which a cellulose derivative is used, Lyocell is produced in a spinning process using a genuine solution of cellulose in *N*-methylmorpholine *N*-oxide (NMMO) monohydrate. Because of the totally dissolved state of the cellulose in the Lyocell process, used textiles can be also used as cellulose source [828]. Cellulase technology plays a key role in the Lyocell production process. Lyocell tends to fibrillate upon physical stress during production due to its high crystallinity index, swelling characteristics, and low degree of interfibril interaction [829]. Although this fibrillation results in a good hand performance, the attendant properties of high pilling tendency and perceived color changes make it essential to remove the fibrillation enzymatically. The raised long fibers and fibrils (the so-called primary fibrillation) are removed by the use of cellulases in equipment that also produces mechanical stress. After this removal, a uniform secondary fibrillation is produced, consisting of relatively short fibrils (the secondary fibrillation) which give the fabric its so-called peach-skin feel [830]. While the enzymatic removal of the primary fibrillation is essential for Lyocell staple fibers, it is not for Lyocell filament [831].

Despite its essential role in Lyocell processing, only a relatively small number of studies have been published concerning the optimal cellulase for the defibrillation application. Studies with a whole acid cellulase composition from *Trichoderma reesei* and different endoglucanase-enriched compositions, using Lyocell and blends thereof with cotton and linen, revealed that no essential differences occur between the various cellulase compositions [832, 833]. The intensity of mechanical action of the garment-processing equipment plays a more significant role in removing the fibrillation. However, when the performance of these enzyme compositions is evaluated on related parameters such as TSL and fabric hand, the endoglucanase-enriched compositions are preferred.

5.2.5.2.8. Processing of Bast Fibers

Flax contains fibers in the cortex region of the bast stem [834]. These fibers are separated from nonfiber tissue by retting and subsequently mechanically cleaned to remove nonfiber materials. Retting is usually a microbial process and is a crucial step in the production of linen fibers. The oldest known retting processes are dew and water retting. Both processes suffer from many disadvantages, including high labor costs, variations in fiber quality, and limited suitable geographical area [835]. Therefore, many studies have focused on replacing the above-mentioned processes by a more controllable enzymatic one. Despite the progress made, at present use of an enzymatic method is hindered mainly by cost. Mixtures of pectinases, (hemi)cellulases (E.C. 3.2.1.78), and xylanases (E.C. 3.2.1.8) have been investigated [834, 836]. The presence of a substantial amount of an endopolygalacturonase appeared to be essential [834]. Apparently, degradation of the pectin structure is the most effective route to make the cellulosics (over 50 % of the stem material) available for harvesting, a process comparable to the bioscouring of cotton (see Section Scouring of Cotton). The newer enzymatic spray-retting process uses pectinase-rich enzyme mixtures and crimp stems [837], and is claimed to have potential for reducing costs. Its economical viability must first be evaluated on a plant scale before implementation can occur.

5.2.5.3. Proteinous Fibers

5.2.5.3.1. Wool Processing

Wool (→ Wool) is a natural fiber with a rather complex architecture. It is mainly composed of proteins (ca. 97 %) and lipids (1 %) [768]. Morphologically, two major parts can be distinguished: the cuticle composed of overlapping cells, and the cortex composed of spindle-shaped cells separated by a cell membrane complex. At the exo side of the cuticle, the fatty acid 18-methyleicosanoic acid is bound to the protein matrix. These modified proteins can be removed by means of alcoholic alkaline or chlorine solutions. The protein present in wool is keratin. Two types are known: the acidic type I and the more neutral to alkaline type II keratin, both with a molecular weight of 40 – 60 kDa. [838]. Especially in the cuticle, the protein is cross-linked by disulfde (between cystine residues) and isodipeptide [ε-(ϒ-glutamyl)lysine] bonds, and thus forms an effective barrier to the environment. Consequently, processing of wool must start with modification of the composition and morphology of the surface.

The intrinsic wool properties of felting and tendency to shrink are a consequence of the arrangement of the cuticle scales with their outer edges towards the fiber tip, so that the friction of a wool fiber in the scale direction is lower than that against the scale direction [839]. This phenomenon, known as the differential frictional effect, causes preferred fiber movement towards the fiber root when mechanical action is applied in the wet state (see Fig. 90).

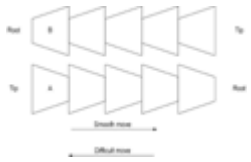

Figure 90. Schematic presentation of wool fiber shapes. When fiber A is moved once to the right (smooth move direction), it cannot be moved back as a consequence of hooking in of the fibers A and B. This phenomenon is called felting.
[Full View]

To obtain machine-washable wool products, the wool is treated with an antifelt or antishrink finish. Conventional commercial shrinkproofing processes can be divided into three categories:

1. Chemical modification of the fiber surface by oxidative or reductive degradation
2. Application of a polymer resin to mask the scale structure
3. A combination of the first two methods

The best known and most frequently used commercial process is the Chlorine-Hercosett process, in which strongly acidic chlorine treatment is followed by neutralization and application of a polymer resin. Disadvantages of this process are that the polymer finishing changes the natural wool handle to a more synthetic one, and the formation of AOX, which release in wastewater is under pressure from environmental legislation.

Much effort has been devoted to developing alternative processes. Enzyme technology, especially protease

technology, has been explored intensively [768, 770], as such or in combined processes. Protease activity towards wool is very slow due to the protection of the cross-linked cuticle cells. Once those cross-links have been eliminated by oxidation or reduction, the reaction rate is improved substantially [840, 841]. The major problem in the application of protease to wool is damage caused by enzyme penetration between cuticle scales into the interior of the fibers and break down of the cell membrane complex. Diffusion of the protease into the interior of the wool fiber must be prevented. However, reaching sufficient levels of antishrinkage behavior to obtain machine-washable wool according to Woolmark TM 31 leads to severe damage to the fibers and nonuniform performance [842].

To improve the proteolytic antishrinkage process, different types of proteases have been studied. A thermo- and alkali-stable protease has been applied on an industrial scale in the prewashing step of a wool dyeing process [843]. Whereas the degree of whiteness is enhanced and improved dye uptake and penetration are obtained with a more even distribution, the felting tendency is still too high. A protease from the fungus *Aspergillus flavus*, isolated from wool, has been examined [844]. Although the enzyme aggressively attacks the cuticle proteins and modifies the scales, the decrease in mechanical properties that is also observed indicates penetration of the enzyme into the interior of the fiber. Application of another enzyme, protease SZ, in a process starting with a hydrogen peroxide pretreatment followed by enzymatic breakdown resulted in the degradation of the cuticle and pronounced antifelting behavior. This enzyme-treated wool is claimed to be machine-washable, while the loss of mechanical properties is limited. Additionally, excellent low-temperature dyeing behavior is noted [845].

In a newer combined enzymatic process the enzyme is incorporated in the alkaline peroxide treatment, after which chitosan is applied to the wool fiber [846]. The tight ionic bonding between sulfuric groups generated on the wool scales and chitosan resulted in good shrink resistance. It is claimed that the material treated in this way is shrink-resistant to the level of machine-washable. Additionally, it is observed that the whiteness and wettability are enhanced.

The above-mentioned enzymatic performance benefits of enhanced whiteness, dyeability, and shrinkage resistance to a machine-washable level, together with maintained handle and improved pilling resistance, are claimed for a different enzymatic process in which Perizym AFW, a product containing a protease of unknown origin, is used [839]. This process has been introduced on industrial scale. In combination with the subsequent application of a polymer resin, the obtained antishrinkage level meets the criteria of the Super-Wash Standard.

The use of enzyme technology in the wool carbonization process has been reviewed [768]. In the carbonization process natural soils on wool such as vegetable matter and skin flakes are removed by impregnation with sulfuric acid and subsequently baking in order to char the cellulosic impurities. Enzymes such as hydrolases, pectinases, and oxidoreductases have been studied in this process, with the objective of reducing wool-fiber damage, effluent load, and energy consumption. Although some enzyme performance has been observed, for example, weakening of the binding of burrs to the wool fiber, the expectation is that it will not come to an industrial process because of the required time window [770].

5.2.5.3.2. Degumming of Silk

Silk fibers (→ Silk) consist of two fibroin filaments surrounded by the protein sericin. The diameter of silk fibers is in between 10 and 14 µm. Contrary to wool, the silk protein contains hardly any methionine or cystein residues [847]. The molecular weight of the fibroin protein is ca. 370 kDa. Its primary structure contains the frequently repeated hexapeptide unit Ser-Gly-Ala-Gly-Ala-Gly. Fibroin contains mainly glycine (44 mol %), alanine (29 mol %), and serine (12 mol %) residues, whereas in cericin the main amino acid residues are serine (37 mol %), aspartic acid (15 mol %), and glycine (15 mol %). The content of sericin is around 17 – 25 wt %. The presence of sericin causes brittleness, a harsh handle, and yellowness of silk fibers [847]. Therefore, this protein is removed in a so-called degumming process to give the soft handle and typical lustre of silk. Removal of sericin leads to a tensile strength loss of ca. 20 %. Methods to remove sericin include the use of Marseilles soap at high temperature, an alkaline solution of sodium hydrogencarbonate, or enzymes (proteases). The ideal enzymatic method would selectively cleave proteins after aspartic acid and/or threonine residues, which are more than nine times more abundant in sericin than in fibroin.

Much effort has been invested in the search for the best performing protease enzyme in the degumming

process [848-851]. In general, alkaline-operating proteases show a good performance with respect to sericin degradation and removal, handle, and lustre, while tensile strength loss is limited or not observed. The efficiency of degumming varies when different varieties of silk are used [850]. Proteolytic enzymes like trypsin, papain, and pepsin have a relative poor performance in this application. Also, a combination of a protease and a lipase has been tried with results better than were obtained with the conventual Marseilles soap [852]. The enzymatic degumming process is only used in China.

5.2.5.4. Textile Effluent Treatment and Recycling

Wet processing of textiles is a major source of water pollution. In general, starch or poly(vinyl alcohol) from the desizing process and a range of different dyes are found in mill effluents. Dyes are regarded as the most important contaminants to be removed from the effluent. The chromophoric compounds are quite stable, and conventional wastewater treatment is not efficient in removing recalcitrant dyes. For example, azo dyes, which are not easily biodegraded, are toxic and also become harmful for the environment upon degradation by formation of aromatic amines. However, most of the methods developed for the removal of dyes from wastewater, such as adsorption, ion exchange, electrokinetic coagulation, membrane filtration, and ozonation, are not implemented on a large scale due to high implementation cost and primarily because of the narrow application range: the methods are applicable to selected dyes only. Due to the emerging environmental awareness and the increasing pressure of legislative restrictions on colored effluent, the potential of biotechnology in this application area has been recognized (it is known that certain enzymes such as laccases, lignin peroxidases, and manganese peroxidases are capable of degrading aromatic rings [853]) and explored.

Several studies have reported complete or almost complete decolorization (dye degradation) using different fungal microorganisms. Among others, complete degradation of the dyes Acid Yellow 17, Reactive Blue 21, poly R-478, and Remazol Brilliant Blue R has been reported using the fungus *Pleurotus ostreatus* [854, 855] and of various azo dyes by *Phanerocheate chrysosporium* [856, 857]. In the latter study the fungus is used in a bioreactor operated in a continuous mode with 95 % decolorization at 0.2 g/L dye load. In these studies a correlation was observed between decolorization and the presence of peroxidase in the effluent, which was secreted by the fungus. A laccase system has been developed for the decolorization of water polluted with the anthraquinone dye Remazol Brilliant Blue R [858]. This system makes use of a redox mediator, with violuric acid being the most effective. The above studies have been concentrated on the use of single enzymes or fungi and isolated dye classes. In a more practically oriented study, a microbial reactor containing different, but not identified, immobilized microbes has been used for treating a number of different textile effluent samples. The samples were collected from different mills, including bleacheries, bleaching and dyeing plants, screen-printing plants, and carpet dyeing mills, employing a variety of dyes, chemicals, and processes. In batch runs of 24 h for most samples a color removal of over 75 % was reached, together with a COD reduction of the same degree [859]. The study has been carried out on both bench and pilot scale, yielding water that can be reused at various points in the plant. Although the system has not been defined very well, the results are encouraging and is the basis for the expectation that implementation on relative broad industrial scale of a biotechnological system for textile effluent will be realized by ca. 2006.

The vat dye indigo, present in wastewater from stone-washing processes for denim, can be degraded by the use of different fungi [860], or the use of a laccase enzyme, both in the absence and presence of a redox mediator [861]. Laccase has also been immobilized prior to application in the decolorization process to reduce enzyme costs by reuse [862].

In the textile finishing industry, for processing of 1 kg of textiles over 100 L of water is consumed. Since the availability of water is a growing problem and the costs of disposal are increasing, new ways are being explored for reuse of water. One possibility is the reuse of (peroxide) bleaching process water in the dyeing process after degradation of the excess of hydrogen peroxide. The reuse of bleaching water after treatment with a catalase was studied in a dyeing process using Reactive Blue 198 [863]. Unacceptable color changes on the dyed fabrics were obtained. It appeared that the presence of the enzyme protein, which undergoes inactivation and unfolding, interferes in the dyeing procedure by interacting with the dye [864]. By immobilization of the catalase on a carrier, this drawback has been eliminated, and a dyed fabric is obtained with an acceptable color difference (not visually detectable) when compared to fabric dyed in fresh water [865]. Although no information is available on the possibility for multicycle reuse, the potential for reuse, and thus a substantial reduction of fresh water requirement, has been shown. The impact of accumulating processing agents may be a major hurdle for implementation of this water-recycling process. This

development can be regarded as a first step towards a closed-loop production process.

As an example of achieving such a closed-loop water system, a study was performed for the desizing step in the textile production chain. In this model study, synthetic wastewater mimicking that from the desizing process was treated in an anaerobic (or methanogenic) bioreactor. This unit for biological wastewater treatment is integrated in the desizing phase of the textile production process to form a closed water loop [866]. The feasibility of such integrated water system within a particular part of the overall production process has been demonstrated. Further research is required to identify possible drawbacks for other parts of the overall process before such process can be implemented. Nevertheless, it is predicted that in future textile production mills will have their own wastewater-purification systems to keep the costs of waste treatment low and to minimize the use of fresh water.

5.2.5.5. Outlook

First, the topics described in this chapter that are in the investigation phase will be brought to implementation on an industrial scale, for example, biofinishing of wool and enzymatic effluent treatment. Second, different bioprocesses, especially the pretreatment processes for fabrics made from natural fibers, will be integrated in a single process with benefits in time, energy, water and chemical consumption, as well as increased production capacity. Integration of enzyme-catalyzed processes is rather easily due to the comparatively mild process conditions and the fact that enzymes are very specific. One major prerequisite is that all integrated enzyme applications should be run in the same pH and temperature range. In case of disharmony, one can screen for an identical enzyme type with different pH and/or temperature profile. Third, enzymatic processes will also be developed for production in a continuous mode. At present, most implemented enzymatic processes are operated in the batch mode, often for reasons of time (most enzyme systems) and/or shear force (cellulase technology). For continuous application, the enzymatic incubation time should be reduced substantially to prevent fabric pile up. A different application opportunity for enzyme technology, and especially cellulase technology, is its use for subtle design development. When biopolishing can be applied in certain patterns by, for example, foam technology [815], different color shades can be realized.

Enzyme applications in the textile sector initially focused on natural fibers. It can be expected that the focus will be broaden in the near future to synthetic fibers like polyester and nylon. Already in 1998 positive performance results were reported for the use of lipases to increase the water-wetting and absorbent properties of regular polyester fabrics [867]. Likewise, it was shown that the culture filtrate of a white-rot fungus can substantially change the surface properties of polyamide fabrics made of PA66 and PA6, apparently via action of a manganese peroxidase (E.C. 1.11.1.13) [868]. Proteases were shown to have some impact (hydrolysis) on nylon oligomers [869]. Finally, the hydrolysis of polyacrylonitrile fibers and granulates by nitrilases (E.C. 3.5.5.1 & 7) or nitrile hydratase (E.C. 4.2.1.84)/nitrile amidase (EC 3.5.1.4) enzyme system has been proven [870]. In this study it was shown that enzymatic surface modification of acrylic fibers results in improved dyeing efficiency.

All these are initial studies. In general, the enzymatically catalyzed reactions on synthetic fibers are rather slow, and the methods as such cannot be implemented on large scale. Nevertheless, the potential of enzyme technology for synthetic fibers has been shown. A search for more catalytically active enzymes for these substrates, both by screening and genetic engineering, is necessary. Enzyme technology is also expected to become a valuable tool for environmentally friendly production of synthetic fabrics with improved comfort characteristics, comparable to those of natural fibers.

5.2.6. Enzymes in Pulp and Paper Processing

5.2.6.1. Introduction

Enzymatic applications in the pulp and paper industry have been developed since ca. 1980. Enzymes were first introduced on a mill scale in the 1980s, following the discovery of xylanase-aided bleaching. Most of the biotechnical applications proposed for the pulp and paper industry are based on the use of enzymes. The specificity of enzymes makes them unique tools for targeted modification of specific components in pulp or process waters and their catalytic nature makes them efficient even in small dosages. Enzymatic applications are generally considered to be environmentally benign. The limitations of the use of enzymes in pulping and papermaking processes are related to the size and properties of the enzymes, as well as to the relative high price of these catalysts. The conditions used in many process stages in pulping and papermaking have led to the development of enzyme preparations having pH and temperature ranges suitable for these target

processes. Today (2003), the most important commercialized applications of enzymes in the pulp and paper are bleaching, deinking, improvement of paper-machine runnability by hydrolysis of extractives or enhanced drainage, and fiber modification for speciality products.

Enzymatic applications for the main processes in the pulp and paper industry, i.e., pulping, bleaching, and papermaking, have been intensively studied and developed since ca. 1980. However, the implementation of enzymatic process stages into industrial processes has not been as extensive as expected in the 1990s. Nevertheless, xylanase-aided bleaching, currently (2002) the main biotechnical application in the pulp industry, developed unusually rapidly from laboratory to industrial practise. This was mainly because the implementation of an enzyme stage did not require any major changes in the process and because cost-efficient commercial enzyme preparations were rapidly developed for industrial-scale use. The successful introduction of enzymes into the pulp and paper processes apparently requires at least the following elements:

- Clear economical benefits should be obtainable from the enzyme stage,
- The quality and processing of fiber material should be maintained or improved,
- No unwanted changes in the process runnability should occur,
- A suitable enzyme preparation should be available in large quantities and at a reasonable price for mill scale trials and thereafter for industrial use.

Due to the complex chemistry involved in wood materials, a number of different enzymes can be used to improve the pulp and paper processes (Table 36). As effective and specific catalysts enzymes can be used in any process phase having appropriate conditions. The harsh process conditions, i.e., mainly high or low pH and high temperature, used in some process stages, focus the use of enzymes to less extreme process stages. To overcome the limits related to pH and temperature, novel thermophilic and alkalophilic enzymes are, however, being developed.

Table 36. Technical benefits obtained with enzymes acting on different components in pulp and papermaking. Process stage:

Component	Enzyme	Physicochemical modification	Process stage*	Technical benefit
Cellulose	cellobiohydrolase	microfibrillation	1	energy saving in tmp refining, increased flexibility of fibers
	endoglucanase	depolymerization	2	increased paper machine runnability, speciality products/high density paper
	mixed cellulases	depolymerization	3	increased release of ink particles
Xylan	endoxylanase	depolymerization	4	increased extractability of lignin
Glucomannan	endomannanase	depolymerization	4	increased extractability of lignin
		decreased colloidal stability	5	increased runnability of paper machine

				paper machine
	acetyl glucomannanesterase	decreased solubility of glucomannan	5	increased pulp yield, increased strength properties
Pectins	polygalacturonase	depolymerization	6	energy saving in debarking
			5	decreased cationic demand in papermaking
	galactanase	depolymerization of galactan	4	improved extractability of lignin
Lignin	laccase	polymerization	5	polymerization of lignans in process water
	laccase mediator Mn-peroxidase	depolymerization	4	increased brightness in chemical pulps
Extractives	lipase	increased fiber hydrophilicity	5	increased strength properties of mechanical pulps, improved runnability of paper machine
Biofilm proteins and polysaccharides	mixed enzymes	depolymerization	1 – 5	slime control, improved runnability of paper machine

*1. Mechanical pulping, 2. paper manufacturing, 3. deinking, 4. kraft pulp bleaching, 5. mechanical pulping/paper manufacturing, 6. debarking.

5.2.6.2. Enzymes

The enzymatic degradation of lignocellulosic material involves a set of different enzymes. Cellulases have been studied intensively for over two decades, and their reaction mechanisms have been revealed on the molecular level for many enzyme components. In spite of the long history, new interesting enzymes are still being found. The prices of commercial cellulases have decreased several fold during the last two decades, and further improvements can be expected. Today commercial, tailored cellulase preparations can be designed according to process needs. The enzymology of hemicellulases is also well established, with the same general trends. Oxidative enzymes have been studied as well, but detailed knowledge on their mechanisms of action in the fiber matrix remains to be elucidated.

5.2.6.2.1. Cellulases

Cellulose is the main carbohydrate in lignocellulosic materials. It is a chemically simple homopolymer consisting of up to 1000 β-1,4-linked anhydroglucopyranoside units [871]. However, its physical state makes it a challenging substrate for enzymes. Single glucose polymers are packed onto each other to form highly crystalline microfibrils in which the individual cellulose chains are held together by hydrogen bonds. Cellulose microfibrils also contain some amorphous regions. In wood fibers the winding direction of cellulose microfibrils varies in different cell wall layers, and this gives the fiber its unique strength and flexibility.

Efficient degradation of cellulose requires a mixture of different cellulases acting sequentially or in concert. Cellulases have traditionally been categorized into two different classes by the IUB. Endoglucanases (EG,

E.C. 3.2.1.4) cleave the cellulose chains internally. Exoglucanases (cellobiohydrolases, CBH, E.C. 3.2.1.91) degrade cellulose starting from free chain ends, producing cellobiose. Endoglucanases act mainly on the amorphous parts of the cellulose fiber, whereas cellobiohydrolases are also able to degrade crystalline cellulose. However, this classification is not very satisfactory, as some exoglucanases have also been claimed to possess endo activity [872, 873]. The classification of cellulases becomes even more complicated, as some enzymes also have activity on other polysaccharides, such as xylan [874]. Therefore, a new classification of these enzymes has been adopted. Today, glycosyl hydrolases are grouped according to the structures of their catalytic domains into more than 70 families (`http://afmb.cnrs-mrs.fr/~pedro/CAZY/db.html`).

Most cellulases have a multidomain structure consisting of a core domain separated from a cellulose-binding domain (CBD) by a linker peptide. The core domain contains the active site, whereas the CBD interacts with cellulose by binding the enzyme to it [875], [876]. The CBDs are particularly important in the hydrolysis of crystalline cellulose. It has been shown that the ability of cellobiohydrolases to degrade crystalline cellulose clearly decreases when the CBD is absent [877]. However, the exact role and action mechanism of CBDs is still a matter of speculation.

5.2.6.2.2. Hemicellulases

The two most common wood hemicelluloses are xylans and glucomannans. Hardwoods contain mainly xylan, whereas the amount of glucomannan in softwoods is approximately twice the amount of xylan. Hardwood xylan is composed of β-D-xylopyranosyl units which may contain 4-*O*-methyl- α-D-glucuronic acid and acetyl side groups. 4-*O*-Methylglucuronic acid is linked to the xylan backbone by O-(1,2) glycosidic bonds and the acetic acid is esterified at the C2 and/or C3 hydroxyl group. Softwood xylan is arabino-4-*O*-methylglucuronoxylan in which the xylan backbone is substituted at C2 and C3 with 4-*O*-methyl- α-D-glucuronic acid and α-L-arabinofuranosyl residues, respectively. Softwood galactoglucomannan has a backbone of β-(1,4)-linked β-D-glucopyranosyl and β-D-mannopyranosyl units, which are partially substituted by α-D-galactopyranosyl and acetyl groups. The distribution of carbohydrates in the wood fibers varies depending on the wood species and growing conditions [871]. In softwoods, the xylan content of the innermost cell wall layer is generally very high, but otherwise xylan is relatively uniformly distributed throughout the fiber walls in native wood. In hardwoods, the outermost layers of fiber walls are rich in xylan. In softwoods the glucomannan content increases steadily from the outer parts to inner parts of cell walls. The amount and composition of hemicellulose components is drastically altered in kraft cooking. All acetyl groups are degraded, and most of the glucuronic acid groups in xylan are converted to hexenuronic acid groups [878]. Softwood glucomannan is deacetylated, and a considerable part is solubilized in the highly alkaline and hot cooking liquor [871].

Xylanases. The majority of published work on hemicellulases deals with the production, properties, mode of action, and applications of xylanases [879-884]. Endoxylanases (E.C. 3.2.1.8) catalyze the random hydrolysis of β-D-1,4-xylosidic linkages in xylans. Most xylanases belong to the two structurally different glycosyl hydrolase groups (Families 10 and 11). The three-dimensional structures of xylanases in both of these families are available, and they show clear differences in size and structure. *Trichoderma reesei* xylanase from Family 11 is an small, ellipsoidal enzyme with a diameter between 32 and 42 Å. It does not have any separate substrate-binding domain [885]. Some xylanases have been reported to contain either a xylan-binding domain [886] or a cellulose-binding domain [887], [888]. Some of the binding domains have been found to increase the degree of hydrolysis of fiber bound-xylan, whereas others have no effect thereon. Neither a xylan- nor a cellulose-binding domain was found to have any significant role in the action of xylanases in pulp fibers [889]. Most of the xylanases characterized are able to hydrolyze different types of xylans, showing only differences in the spectrum of end products. However, no systematic studies on the substrate specificity of xylanases belonging to different families on fiber-bound substrates have been carried out. From a practical point of view, the most important characteristics of xylanases are their pH and temperature stability and activity. A number of enzymes produced by extremophilic organisms have been characterized [890-892], but surprisingly few have reached commercial use, due to problems related to their efficient production in heterologuous host strains.

Mannanases. Endomannanases (E.C. 3.2.1.78) catalyze the random hydrolysis of β-D-1,4 mannopyranosyl linkages within the main chain of mannans and various polysaccharides consisting mainly of mannose, such as glucomannans, galactomannans, and galactoglucomannans. Mannanases seem to be a more heterogenous group of enzymes than xylanases. The mannanase of *Trichoderma reesei* has a similar multidomain structure

to several cellulolytic enzymes; i.e., the protein contains a catalytic core domain which is connected by a linker to a cellulose-binding domain [893], [894]. The CBD increases the action of *T. reesei* mannanase on fiber-bound glucomannan, even though the catalytic domain can efficiently degrade crystalline mannan [895]. The hydrolysis yield of glucomannans is dependent on the degree of substitution and on the distribution of the substituents [896]. The hydrolysis of glucomannans is also affected by the glucose/mannose ratio. Some mannanases are able to hydrolyze not only the β-1,4-bond between two mannose units but also the bond between the mannose and glucose units [897], [898].

Other Hemicellulases. The side groups connected to xylan and glucomannan main chains can be cleaved by α-glucuronidase (E.C. 3.2.1.131), α-arabinosidase (EC 3.2.1.55), and α-D-galactosidase (E.C. 3.2.1.22). Acetyl substituents bound to hemicellulose are removed by esterases (E.C. 3.1.1.72) [881]. Most of the side-group-cleaving enzymes are able to attack only oligomeric substrates produced by backbone-depolymerizing endoenzymes, i.e., xylanases and mannanases. Only few enzymes can attack intact polymeric substrates. However, most accessory enzymes of the latter type prefer oligomeric substrates.

5.2.6.2.3. Lignin-Modifying, Oxidative Enzymes

In nature, lignin (→ Lignin) is mostly found as an integral part of the cell wall, embedded in a carbohydrate polymer matrix of cellulose and hemicellulose. The exact composition of lignin varies widely with wood species, but as a rule, softwood contains mainly guaiacyl units whereas hardwood also contains syringyl building blocks. It has been suggested that both the chemical and three-dimensional structure of lignin is determined by, or at least strongly influenced by, the polysaccharide matrix. For example, the lignin near the cellulose surface resembles a glassy or low-order structure, rather than a random one [899].

Reactions catalyzed by oxidative enzymes play a significant role in the complete degradation of lignocellulosic biomass. The enzymatic modification and degradation of lignin, the third major component of lignocellulosic materials, is still only partially understood, because of the oxidoreductive and radical reactions involved.

The enzymology of lignin has been the focus of enzymologists for more than 30 years, and has resulted in the identification of the major enzyme systems believed to participate in delignification; lignin peroxidase, (LiP, E.C. 1.11.1.14.), manganese-dependent peroxidase, (MnP, E.C. 1.11.1.13), and laccase (E.C. 1.10.3.2). In addition, other types of peroxidases have been identified, such as manganese-independent and LiP-like peroxidases [900]. In vitro the oxidative reactions of these enzymes lead mainly to polymerization of lignin. Purified ligninolytic enzymes have, however, been shown to cause limited delignification provided that additives, such as veratryl alcohol and H_2O_2 for LiP [901] and manganese, H_2O_2, organic acids, and surfactants for MnP [902], [903] are supplemented. Due to the promising results in using additives (coordinating/chelating compounds and electron mediators), laccases and manganese-dependent peroxidases have recently been the most extensively studied groups of enzymes in the area of lignin degradation.

Lignin Peroxidases (LiP). Fungal peroxidases are heme proteins which oxidize their substrates with hydrogen peroxide as electron acceptor in two one-electron steps, like all other peroxidases [904]. LiPs have high redox potentials and can oxidize non-phenolic lignin model compounds, aromatic ethers, and polycyclic aromatics. Cation radicals formed by the action of the enzyme cause several different reactions and finally result in ring cleavages. LiPs have been characterized from several white-rot fungi. The three-dimentional structure of LiP from *Phanerochaete chrysosporium* has been solved and shows that the active-site environment is completely conserved, as in all peroxidases studied so far. Attempts to overproduce LiP in heterologuous hosts have mostly been unsuccesful, and this has hindered the commercial exploitation of LiP.

Manganese Peroxidases (MnP). Manganese peroxidase (MnP) is a heme protein which oxidizes Mn^{2+} to Mn^{3+}. Chelation of Mn^{3+} by organic acids is necessary to stabilize the ion and to promote its release from the enzyme. The Mn^{3+} · chelate may function as a diffusible oxidant and degrade the phenolic moieties of lignin. Phenoxy radicals are formed and undergo a variety of reactions that result in products quite similar to those released by LiPs. Interestingly, MnP has been shown to promote peroxidation of unsaturated lipids without added H_2O_2. MnP genes have been isolated from several white rot fungi [905-908]. Multiple sequence alignments have revealed specific differences between LiPs and MnPs, in spite of the general

sequence conservation. The three-dimensional structure of MnP from *Phanerochaete chrysosporium* has been published. MnP has also been successfully expressed in *Aspergillus oryzae*, and the secreted MnP was fully active and identical to the native enzyme [909].

Laccases. Laccases belong to the copper metalloenzymes and to the blue oxidase subgroup. Laccases contain four copper atoms per molecule. Three types of copper atom can be distinguished by their spectroscopic and paramagnetic properties: type 1 (T1), type 2 (T2), and type 3 (T3). Laccases contain one T1 Cu bound as a mononuclear center, one T2, and two T3 Cu atoms. These three T2 and T3 Cu atoms form a trinuclear center. In the resting enzyme all four copper atoms are likely to exist in the CuII state. The three-dimensional structure of a laccase from *Coprinus cinereus* was first elucidated [910].

The typical reaction of laccase is oxidation of a phenolic compound with the concurrent reduction of molecular oxygen to water. After four cycles of single-electron oxidations forming free radicals, the enzyme reduces one molecule of oxygen, generating two molecules of water. This mechanism requires the protein to “store” four electrons like a “molecular battery” before reduction of oxygen. The mononuclear T1 Cu site functions as the primary electron acceptor, extracting electrons from the reducing substrate and delivering them to the trinuclear T2/T3 site. The trinuclear T2/T3 center, the binding site for the second substrate, dioxygen, accepts electrons from the T1 site for reduction. Reduction of oxygen most likely takes place in two steps, since bound oxygen intermediates are involved [910].

In laccase-catalyzed oxidation the substrate loses a single electron and forms a free radical. The unstable radical may undergo further laccase-catalyzed oxidation or nonenzymatic reactions such as hydration and polymerization. Aryloxy radicals formed by laccases may undergo further nonenzymatic oxidation/reduction or couple to other phenolic structures and produce intensely colored products. Laccases have very broad substrate specificities and can couple four one-electron oxidations to a variety of substrates, such as di- and polyphenols, aromatic amines, and a considerable range of other compounds, to the irreversible four-electron reduction of dioxygen to water.

5.2.6.3. Enzymes in Pulp and Paper Processing

5.2.6.3.1. Mechanical Pulping

Mechanical pulping (→ Paper and Pulp – Mechanical Pulp) involves the use of mechanical force combined with pressure and temperature and in some cases also with chemicals to separate the wood fibers. Mechanical pulps, such as pressurized ground wood (PGW) or thermomechanical pulp (TMP) have a high yield (up to 95 %) and can be used to produce paper with high bulk, good opacity, and excellent printability. The drawbacks of these processes are the high energy intensity and the resulting fiber and paper quality, with lower strength, higher pitch content, and higher color reversion rate as compared to chemical pulps. Energy consumption is thus a key issue when developing new technical processes for high-yield pulping. The demand for improving paper quality has been the main reason for the increased specific energy consumption in thermomechanical pulping during the last few years. One way of reducing the high energy consumption of mechanical pulping is to modify the raw material by biotechnical means prior to refining.

The low accessibility of mechanical pulp fibers to enzymes limits fiber modification to the accessible pulp surfaces (outer fiber surface, fines) and to dissolved and colloidal material solubilized to process water [911]. Currently, the main biotechnical applications related to mechanical pulping investigated in various research institutes around the world are biomechanical pulping [912], microbial reduction of pitch components prior to pulping [913], and enzyme-aided refining of coarse mechanical pulp fibers [914].

Incorporation of an enzymatic step in a mechanical pulping process can be expected to be succesful only after primary refining, mainly due to the low accessibility of wood chips for enzymatic modification. A new process concept based on the treatment of coarse mechanical pulp fibers with monocomponent cellulases has been developed [914], [915]. Experiments with different enzymes have demonstrated that a slight modification of cellulose by the *T. reesei* cellobiohydrolase I (CBH I) results in energy saving of 20 % in a laboratory-scale disk refiner. Interestingly, when an unoptimized cellulase mixture was used, no positive effect on energy consumption was detected [915], [916]. In pilot-scale, two-stage secondary refining with a low-intensity refiner (wing defibrator), an energy saving of 10 – 15 % was obtained with CBH I [914]. The method has been tested on an industrial scale and is under further development.

Neither the cellulase mixture nor CBH I induced evident morphological modifications in the coarse and

rigid TMP fibers during a short incubation [913]. Based on the fibrillation index analyses it has been speculated that the action of CBH I possibly induces decreased interfibrillar cohesion inside the fiber wall and thus results in loosening and unraveling of fiber structure. Cellulase treatment with CBH I did not have any detrimental effects on pulp quality. In fact, the tensile index was even higher for the CBH I-treated pulp than for the reference. The increase in tensile index could be explained by the intensive fibrillation induced by the CBH I treatment. The good optical properties were also maintained after CBH I treatment.

5.2.6.3.2. Chemical Pulping

The main aim in chemical pulping (→ Paper and Pulp – Pulping Processes) is to remove lignin and to separate the wood fibers from each other to render them suitable for the papermaking process. In the pulping process, the lignified middle lamellae located between the wood fibers is solubilized by various chemicals. Today, the predominant pulping method is the kraft process, in which the cooking liquors are incinerated and the cooking chemicals recycled. Extensive modifications of hemicelluloses take place during pulping processes. During conventional kraft cooking, part of the hemicelluloses is first solubilized in the cooking liquor. In the later phases of the process, when the alkalinity of the cooking liquor decreases, part of the solubilized xylan is redeposited onto the cellulose fibers. Although glucomannan is the main hemicellulose in softwood, the bulk of glucomannans are dissolved and degraded during kraft pulping. Thus, the relative amount of xylan is increased in pine kraft pulp as compared to pine wood. In addition to xylan, lignin is also partially readsorbed on the fibers. Lignin has been reported to be linked to hemicelluloses in lignin – carbohydrate complexes (LCC). Furthermore, hemicelluloses seem to physically restrict the passage of high molecular mass lignin out of the cell walls of the pulp fiber, and hence the removal of hemicelluloses, especially xylan, can be expected to enhance the extractability of residual lignin from pulps.

Several methods have been studied for increasing the diffusion of cooking chemicals into wood to improve the efficiency of chemical pulping processes. A long-term goal would be to totally replace sulfur chemicals. Impregnation of chemicals into wood and removal of dissolved lignin are governed by diffusion and sorption phenomena, by the porosity and structure of the cell wall matrix, as well as by the molecular size of extractable molecules. In sulfate, or kraft, cooking wood fibers are liberated and partly delignified by combination of high alkalinity (pH 12 – 14), sulfidity, and high temperature (165 – 170 °C). The kraft fibers are more flexible and thus have more binding capacity than mechanical pulp fibers. The residual lignin that causes a dark brown color in kraft fibers must be removed by bleaching.

As in the case of mechanical pulping, prior to chemical pulping the ability of individual enzymes to penetrate into wood chips is limited by the low porosity of wood. It has, however, been reported that enzymes, including hemicellulases, pectinases, and cellulases, can increase the diffusivity of sodium hydroxide in Southern pine sapwood [916]. This result was attributed to the dissolution of pit membranes, which represent the main resistance to flow of liquids in wood. After acetone extraction and enzyme treatment the pulps were reported to be more uniform, to have higher viscosity and yield, and lower rejects [917]. The enhanced pulp uniformity was attributed to more uniform delignification owing to improved diffusion. The role of the degree of hydrolysis, i.e., loss of carbohydrates during hydrolysis, and the need for acetone extraction in softwood remain to be clarified.

5.2.6.3.3. Bleaching

In bleaching (→ Paper and Pulp – Pulp Bleaching) , the primary goal is to remove the small amount of residual lignin present in the pulp after cooking without decreasing the molecular weight of cellulose. Lignin in unbleached pulps represents typically only about 1 % of the dry weight. During pulping, however, lignin is chemically modified and condensed into poorly degradable structures. Cooking and bleaching are separate process phases, differing from each other with respect to the selectivity of the chemicals used. In the bleaching processes, lignin is sequentially degraded and extracted in several phases. Bleaching sequences are generally composed of at least five phases. Previously, the bleaching of chemical pulps was carried out with elemental chlorine and chlorine dioxide. Today, in Europe, pulp is bleached by ECF (elemental-chlorine free) or TCF (totally chlorine free) bleaching processes, in which oxygen, ozone, or peroxides are used.

Xylanase-aided bleaching of chemical pulp is the main biotechnical application used today in the pulp and paper industry (2003). Xylanase-aided bleaching is an indirect method, which does not directly degrade lignin and has thus a limited effect. Today, the most promising novel enzymatic process stage is laccase-

mediator bleaching, which directly oxidizes and degrades lignin (Table 37). Novel enzyme-assisted systems such as HOS and other slow-release systems are, however, under development [918], [919].

Table 37. Enzymatic bleaching concepts for chemical pulps

Concept	Mechanism	Benefits	Problems	Status
Xylanase-aided bleaching	degradation of redeposited xylan and lignin-carbohydrate complexes	10 – 20 % saving in chemical consumption, increased brightness	indirect method, chemical bleaching needed for delignification	commercialized
Mannanase-aided bleaching	degradation of residual glucomannan	savings in chemical consumption, increased brightness	effect limited to certain pulp types	commercialized (limited use)
Laccase-mediator system	degradation of lignin in the presence of mediators (nitrogen-containing aromatic compounds, ABTS, transition metal complexes)	up to 50 % delignification when combined with alkaline extraction stage	specificity, recyclability of mediator, bleaching costs higher than in conventional bleaching	close to commercialization
Mn peroxidase bleaching (with additives)	degradation of lignin in the presence of additives	good performance	costs of enzymes, cost of additives	not commercialized
Other enzyme-assisted oxidative systems (e.g., HOS, lipase with ketone, fatty acid and H_2O_2)	degradation of lignin by active compounds, such as dioxirane (HOS)	good performance claimed	system needs to be verified, costs of components	not commercialized

Xylanase-Aided Bleaching. The effect of xylanase in bleaching is based on the modification of pulp xylan to give enhanced extractability of lignin in subsequent bleaching stages [882]. Several alternative and perhaps concurrent mechanisms have been proposed to be involved in xylanase-aided bleaching. The enhanced leachability of lignin in the fiber wall has been suggested to be due to hydrolysis of reprecipitated xylan or to removal of xylan from the lignin – carbohydrate complexes (LCC) in fibers [921], [922]. Removal of xylan by xylanases from softwood kraft fibers was found to uncover surface lignin [923]. The action of xylanases on both reprecipitated and LC xylan suggests that it is probably not only the type but also the location of the xylan that is important in xylanase-aided bleaching. The role of hexenuronic acid in the effect

of xylanase treatment on kappa number (measure of lignin content) can be estimated to be relatively small. The hexenuronic acid, containing a double bond, gives rise to the consumption of bleaching chemicals and permanganate, increasing the apparent kappa number of pulp [924]. The partial removal of hexenuronic acid substituted xylan by xylanase treatment consequently results in a slightly lower kappa number [925].

Xylanases seem to be efficient on all types of kraft fibers, whereas the effects of mannanases depend on the type of fibers used [926]. The effect of xylanase on bleachability in most cases was independent of the origin of the enzyme, and both fungal and bacterial xylanases have been reported to act on pulp xylan and result in enhanced bleachability [927]. The efficiency of xylanases from Family 10 and Family 11 in bleaching has been compared, and it has been proposed that some xylanases from Family 11 could be more effective in bleach boosting as compared with the Family 10 xylanases [928]. The effect of cellulose- and xylan-binding domains (CBD and XBD) of xylanases on bleaching efficiency has also been investigated, and so far neither an XBD nor a CBD has been reported to have any significant role in the bleach-boosting efficiency of xylanases [889].

The use of xylanases in different bleaching sequences leads to a reduction in chemical consumption. The benefits obtained by enzymes are dependent on the type of bleaching sequence used and on the residual lignin content of the pulp. In chlorine bleaching an average reduction of 25 % in active chlorine consumption in prebleaching or a reduction of about 15 % in total chlorine consumption has been reported both in laboratory-scale and in mill trials. As a result, the concentration of chlorinated compounds, measured as AOX, in the bleaching effluent during mill trials was reduced by 15 – 20 %. Today, xylanases are industrially used both in ECF and TCF sequences. In ECF sequences, the enzymatic step is often implemented due to the limiting chlorine dioxide production capacity. The use of enzymes allows bleaching to higher brightness values when chlorine gas is not used. In TCF sequences, the advantage of the enzymatic step is due to improved brightness, maintenance of fiber strength, and savings in bleaching costs [882]. As of 2000, about 20 mills in Northern America and Scandinavia use enzymes [929].

Thermostable xylanase products for pulp bleaching have been on the market since 1995 [930]. The thermostabilities of new enzyme products are continuously improving, and research on xylanases acting at both high pH and temperature (pH 10 and 90 °C) is in progress. The approximate price of xylanase treatment in 2000 was less than $2 per tonne of pulp. Calculations of the economic benefits in an ECF sequence reveal that reduction in the chlorine dioxide consumption leads to savings of at least $2 per tonne of pulp. The costs of oxygen-based chemicals (ozone, peroxide) are even higher, and the respective savings even more pronounced. Thus, the potential economic benefits of enzyme-aided bleaching are significant to the user.

Laccase-Mediator Concept in Bleaching. In the laccase-mediator concept, the electron-transfering molecule (mediator) oxidized by laccase acts directly on lignin and results in efficient delignification. In the initial study, the common substrate of laccases, the diammonium salt of 2,2′-azino-di(3-ethylbenzothiazolinsulfonate), ABTS, was used as the mediator [931]. The search for more suitable mediators resulted in the discovery of 1-hydroxybenzotriazole (HBT), violuric acid (VIO), and *N*-hydroxy-*N*-phenylacetamide, (NHA) [932-934]. The most effective mediators in delignification usually contain N —OH — functional groups. The high cost due to the mediator dosage needed and the potential of generating toxic byproducts in a batch process are, however, concerns when the system is incorporated in mill-scale bleaching. Therefore, a continuous method for slow-dosing of the mediator using a precursor of NHA (i.e., DiAc, *N*-Acetoxy-*N*-phenylacetamide) has been developed [935]. In addition to nitrogen-based mediators, inorganic mediators, such as transition metal complexes, containing preferably molybdenum, have recently been successfully tested for laccase-mediator bleaching [936].

Several studies on the mechanisms of laccase-mediated delignification of pulps have been published [937-940]. The LMS system has been shown to be able to replace either the oxygen delignification or ozone stage [940], [941].

The combination of xylanase and laccase-mediator bleaching systems in sequence has been shown to result in additional enhancement of pulp bleachability [925]. The application of the LMS system employing HBT as mediator with xylanase treatment in a single stage was, however, found to be ineffective, apparently due to the inactivation of xylanase by HBT. This inactivating effect of HBT has also been observed towards laccases [934], as HBT radicals undergo chemical reactions with the aromatic amino acid side chains of

many laccases. Studies on new mediators have revealed that NHA caused less damage to enzymes [925], [942]. In practice, it would be beneficial to combine delignifying laccase-mediator treatment and indirect xylanase treatment, as the target substrate of the treatments are different and thus the maximum effect of both treatments could be gained.

5.2.6.3.4. Papermaking

In papermaking (→ Paper and Pulp – Paper and Board Manufacturing) , paper is formed from pulp, various papermaking chemicals, and pigments in a paper machine. Following stock preparation the paper machine normally performs the unit operations of web formation, pressing, drying, sizing, and calendering. A large amount of water is needed for the formation of a uniform paper web from the pulp slurry in the paper machine. High efficiency is required from all unit operations due to the high speed of the paper machine.

Papermakers are currently searching for options to fulfill the legislative and environmental demands to reduce fresh water usage. One major problem in water-circuit closure is the accumulation of dissolved and colloidal substances (DCS) in the process waters. These substances consist mainly of hemicelluloses, pectins, dispersed wood resin, lignans, and dissolved lignin. The DCS composition has a direct impact on paper-machine runnability and paper quality, as well as on the demand for and efficiency of purification of process water.

The composition and structure of extractives, glucomannan, pectin, and lignans can be modified by suitable enzymes such as lipases, mannanases, pectinases, and oxidative enzymes [943]. The small chemical changes achieved by enzymes may result in significant modifications in the behavior of DCS in white waters and, more significantly, may lead to potential improvements in technical parameters such as yield, strength, and brightness of pulp.

Pitch problems originating from lipophilic extractives can be reduced by lipases hydrolyzing triglycerides to glycerol and free fatty acids in mechanical pulps. A commercial lipase preparation, Resinase (Novozymes), has been industrially used in production of groundwood pulps for several years in Japan [944]. Triglycerides from various types of pulps have been reported to be efficiently hydrolyzed by lipases, reducing the stickiness and pitch problems [945-947]. Lipase treatment allows savings in the consumption of additives and surface-active chemicals and results in improved fiber properties [944], [947]. The improved tensile strength apparently is due to the increased hydrophilicity of the fibers [947]. In addition to hydrolytic enzymes, oxidative enzymes such as laccases have been used to modify the composition and structure of lipophilic and hydrophilic extractives [948-950]. Laccase treatment polymerizes lignans to fibers and could also slightly modify lipophilic extractives [951].

Enzymatic modification of the papermaking properties of chemical, mechanical, and recycled pulps were widely studied in the late 1990s. The main ideas have been to enhance the beatability and thus the strength properties of fibers through increased fibrillation or to affect the amount of fines in pulp in order to increase its drainage or water retention. “Fiber engineering”, i.e., the directed modification of fiber surface components for improved fiber properties, mainly binding, has been carried out with chemical pulps. Successful fiber engineering of mechanical pulps would, however, probably result in clearer changes in pulp properties than in the case of chemical pulps, as fewer chemical additives are involved in mechanical pulping.

The effects of individual cellulases on the properties of unbleached or bleached kraft pulp have been studied in detail [952-954]. *T. reesei* cellobiohydrolases (CBH) have been found to have only a modest effect on pulp viscosity, whereas endoglucanases (EG), and especially EG II dramatically decrease pulp viscosity and thus the strength properties after refining. Treatment of ECF-bleached pine kraft pulp with *T. reesei* endoglucanases EG I and EG II has been reported to enhance the beatability considerably [953], [954]. The strength properties of the pulp were simultaneously impaired, presumably due to the attack of the endoglucanase on the amorphous cellulose present, especially in the defects and irregular zones of the fibers [952]. However, in Douglas fir chemical pulp, cellulases and cellulase – hemicellulase mixtures have been reported to enhance the beatability of coarse fibers and thus to improve the paper properties [955]. The positive effect of *T. reesei* CBH I on beatability and thus on the development of binding properties of ECF-bleached spruce kraft pulp in refining was reported [954].

Commercial cellulase – hemicellulase mixtures are available for improving the drainage (freeness) and

runnability of paper machines. This is, however, an application area suitable for speciality paper grades or occasional special needs in the paper mills. The successful use of cellulase preparations also needs special attention regarding the enzyme dosage and treatment time.

5.2.6.3.5. Deinking
Recycled pulp is increasingly used in newsprint, tissue paper, and in higher grades of graphic papers. Fibers in recovered paper must be deinked, i.e., repulped and cleaned from dirt and ink before they can be used again in papermaking. In deinking ink particles are detached and removed from fibers by combined mechanical and chemical action. When aiming at more efficient and environmentally friendly deinking processes, enzyme-aided deinking is a potential alternative. The application of cellulase and hemicellulase mixtures in deinking has been studied on laboratory, pilot, and mill scales [956-964]. The use of enzymes in deinking of recovered paper is one of the highest potential enzymatic applications in the pulp and paper industry and is already used on the mill scale.

There are two principal approaches for the use of enzymes in deinking: the enzymatic liberation of ink particles from fiber surfaces by carbohydrate-hydrolyzing enzymes such as cellulases, hemicellulases, or pectinases, or the hydrolysis of the ink carrier or coating layer. The suitability of enzymes acting on soya-oil-based ink carriers, lignin, and starch have been tested. The hypothesis on enzyme-aided deinking is that the enzymatic hydrolysis of the ink carrier, starch coating, or fiber surface liberates ink particles that are large enough to be removed by flotation deinking. Enzyme mixtures designed for target paper grades are, however, needed to increase ink detachment and flotation of ink particles.

5.3. Development of New Industrial Enzyme Applications

5.3.1. Introduction

In addition to the long-established fields of technical enzyme applications such as detergents, cleansers, textiles, foods, feeds, and pulp and paper, a number of new applications have been developed since ca. 1990. Since most of these applications result in smaller market volumes for the enzyme industry and are less well-known as enzyme use in laundry detergents, they are often neglected. Whenever a new enzyme application is developed it must be elaborated for optimum enzyme effect, safety, economy, and stability of the product. Safety evaluation and risk assessment must take into account the fact that enzymes act as sensitizers when inhaled as dust or aerosol. This safety aspect especially limits the application of enzymes in cosmetic products and hard-surface cleaning, where their performance potential is clear, but where aerosol or dust generation must be avoided even under foreseeable conditions of misuse. As of ca. 2002 the application of enzymes in cosmetics is just emerging from development and the future is not yet clear. In a series of other industrial application fields enzymes are already established, e.g., in specific areas of hard-surface and membrane cleaning. Enzymes are also being used in quite exotic industrial processes, such as cork treatment, and the patent literature indicates that many processes are under evaluation which might in future be carried out by enzymes.

This chapter focuses on both new applications for known enzymes and on new enzymes in existing but less well known application fields. The industrial application potential of enzymes is mainly based on their specific reaction mechanisms and good ecological compatibility. Nevertheless, many good ideas for enzyme applications have not been realized. The reasons why these ideas did not go to market must be discussed against the background of the requirements for enzyme-based products: In the eyes of the customer the enzyme-containing product must show improved or unique performance compared to a conventional product without enzymes. It must be price competitive with existing products and safe to humans and the environment. From the point of view of the manufacturer the enzymatic product must be cost competitive with existing products, display a unique performance characteristic, be stable during handling, and must be compatible with other components and conditions of the product. It must be safe to humans during industrial handling of large quantities. All these parameters must be considered during product development.

Enzyme-containing products or applications have shown in the past that they can result in unique performance characteristics and that their use can be made economically attractive once an application is established. In all cases of new application ideas the safety aspect is probably the most important. Enzymes have been known as sensitizing agents since the 1970s, when FINDT discovered that enzyme dust leads to allergic symptoms in workers who inhaled it during production of biological detergents [965]. These symptoms range from rhinitis or other hay-fever symptoms to asthma. In the area of detergents this discovery lead to technical measures such as ventilation in the production areas and enzyme granulation and

coating procedures, which in combination with other precautionary measures resulted in safe enzyme usage in detergent production [965]. An additional consequence was the declaration of enzyme products under risk phrase R42 (causing sensitization by inhalation).

Certain new ideas for enzyme application, for example, carpet cleaning and malodor removal by spraying enzymes into the air, are not in compliance with these requirements of user safety and avoidance of human contact with the treated product unless additional measures could ensure this safety target is met. A special requirement of the safety evaluation for industrial products is the need to consider potential false application or a foreseeable misuse of the product. Safety evaluations have been performed for enzyme applications in institutional dish washing and in enzymatic stain pretreatments. These were organized by the U.S. and European detergent trade associations (SDA, AISE). The industry association of enzyme manufactures and formulators (AMFEP) informs on its web site on the necessary safety evaluations for enzyme-containing consumer products (see www.amfep.org).

Apart from the most essential safety evaluations and precautions, a number of additional enzyme-specific parameters must be considered by the manufacturer of enzymatic products: for all industrial applications enzyme formulations must be stable, both with respect to storage and process stability. Enzymes, as proteins, are easily degradable, and their degradation by microbes, proteases, and other chemical or physical influences such oxidation, light, heat, removal of essential metal ions or coenzymes must be prevented during industrial application. Starting with the stability aspect, the search for an optimal formulation represents a crucial step in product development.

A further critical aspect of technical enzyme use is the amount of time available for an application. Compared to most technical processes the time needed for enzyme-catalyzed reactions is long, ranging from a minimum of 15 min up to several hours. It depends on other application parameters such as temperature, pH, and the presence of chelating or chaotropic agents. In many cases such a prolonged reaction time is tolerated neither in technical nor in household applications.

Whereas good stability and short application time are a target for all technical products, there is, especially in the consumer product area, the additional requirement for an appearance of the enzyme product that is in compliance with the expectations of consumers, especially concerning color and odor. The crude enzyme material is usually associated with specific color and odor of a fermented product. The purification of an enzyme to fulfill these requirements is technically feasible for most enzymes, but it also is associated with higher production costs.

Finally, but generally of extreme importance in the industrial field, intellectual property rights must be observed, and in many cases these restrict the transfer of known enzymes and their reaction mechanisms to novel product applications.

5.3.2. Enzymes in Cosmetics

Enzymes have been investigated for their potential in hair dyeing, tooth cleaning, and skin treatment [967-973]. Cosmetic products containing enzymes are still relatively rare but their number has increased since ca. 1990. Due to the higher price of enzymes compared to conventional synthetic raw materials their use is restricted to products with unique performance characteristics which cannot be achieved by conventional compounds, or to products with a “natural” image relying on bio-ingredients and promising gentle and nature-compatible cosmetic efficacy.

In general, that the number of patent applications claiming the use of enzymes in cosmetics still exceeds the number of corresponding enzymatic products. The safety aspect is especially important in cosmetic products. Depending on the application form (rinse off, leave on) the risk of inhalation varies and must be assessed for each product category.

5.3.2.1. Hair Dyeing

Permanent hair dyeing processes (see also → Hair Preparations – Hairstyling Preparations) are based on the oxidative polymerization of small aromatic dye precursors [974]. The initial oxidation is usually performed by hydrogen peroxide. The dye precursors are colorless and consist of two components, the so-called developers and couplers. Commercial developers used in hair dyeing products are phenols or aminophenols which are oxidized in the initial step of the hair dyeing process to the corresponding quinones or imines. At

alkaline pH the developers and their initial oxidation products diffuse effectively into the hair, where they polymerize with the couplers to give higher molecular weight oxidation products. These polymerization products remain permanently in the hair. Both the chemical nature of the developers and couplers and their composition leads to the colors and shades offered to customers. Hydrogen peroxide not only initiates the polymerization reaction of dye precursors but also bleaches the natural hair pigment melanin, so that changes from darker to lighter colors can be achieved and the color of the polymerized dye precursors is not shifted by the natural hair color. Commercial hair dyeing products both for home and professional use are usually formulated as two-component systems, i.e., the dye precursors and the hydrogen peroxide are packaged separately to guarantee a stable product.

The concentrations of hydrogen peroxide needed for oxidation of dye precursors are in the range of 3 % in the final application formulation. Special bleaching products contain up to 6 % hydrogen peroxide to achieve bright blond shades. These concentrations, when applied repeatedly, can cause hair damage. The alkaline pH used to swell the hair and make it more accessible for the dye precursors may enhance this effect.

To achieve more gentle dyeing, a milder oxidation process is of interest, even if such products so far are only suitable for hair dyes that change the color to a darker color tone, because they are not strong enough to bleach natural melanin.

Since the 1960s a broad patent literature has claimed the application of several oxidative enzymes such as oxidases, peroxidases, and tyrosinases as substitutes for hydrogen peroxide to give a milder hair coloring process. Mildness is related to avoiding hydrogen peroxide and working at neutral pH, which is compatible with the enzymes. The oldest patent application in this field appeared as early as 1917 [975].

However, up to now no enzymatic hair dyeing product has appeared on the market. The main reason is that the performance, even of the more recent enzyme hair dyes, cannot compete with conventional dyes with respect to permanence, antigraying performance, and color diversity. Since 1999 on average more than 40 new patent applications concerning enzymatic hair dyeing processes appeared each year, and this can be considered to be a sign that the applying companies (e.g., L'Oréal [976], Wella [977], Lion [978], Henkel [979], Kao [980]) are continuously working on the improvement of enzyme performance in their products.

Several different types of oxidative enzymes have a potential for different types of dyeing products. Figure 91 summarizes schematically the different enzymatic mechanisms.

Figure 91. Principles of chemical and enzymatic oxidative hair dyeing processes
[Full View]

5.3.2.1.1. Oxidases

Oxidases such as glucose oxidase and uricase generate hydrogen peroxide in the presence of oxygen by oxidizing their substrates in situ. When these enzymes are used for hair dyeing the conventional oxidizing agent hydrogen peroxide is produced directly on the hair, i.e., the location where it is needed. The process is claimed to be milder than the conventional process due to the lower concentrations of hydrogen peroxide needed. In addition the oxidizing agent is produced continuously at its location of action and consumed by the dyeing process, so no excess of hydrogen peroxide can damage the hair. The coloring of hair with such oxidases was shown to be satisfactory with a number of aromatic amines, e.g., *p*-phenylendiamine [981]. Uricase produced hydrogen peroxide concentrations up to 0.06 % after 5 min of reaction at pH 7 [982]. Since L'Oréal's patent applications from 1999, the hydrogen peroxide generating oxidases are referred to as two-electron transfer oxidases due to their reaction mechanism, which transfers two electrons from the substrate to molecular oxygen with formation of hydrogen peroxide [983].

5.3.2.1.2. Peroxidases
Peroxidases (e.g., isolated from horseradish or soy bean) generate more potent oxygen species from hydrogen peroxide, which are able to oxidize dye precursors at a significantly lower concentration than hydrogen peroxide itself. Patent applications claim the combined use of oxidases and peroxidases to achieve dye formation with very small amounts of hydrogen peroxide [984-986]. These two enzyme types perform under identical reaction conditions (ionic strength, pH, temperature, reaction time). The disadvantage is that the price of a dye product would be increased not only by one additional component but by two, both of which are more expensive than hydrogen peroxide.

5.3.2.1.3. Polyphenol Oxidases
Polyphenol oxidases are copper-containing enzymes that reduce molecular oxygen to water and oxidize a polyphenolic substrate directly by a radical mechanism to the corresponding quinone [987]. They are divided into four subclasses: laccases (E.C. 1.10.3.2; [80498-15-3]), catechol oxidases (E.C. 1.10.3.1), ascorbate oxidases (E.C. 1.10.3.3), and tyrosinases (E.C. 1.14.18.1 and 1.10.3.1). The E.C. nomenclature is not totally consistent concerning the subclasses; for example, it classifies laccases and ascorbate oxidases differently due to their substrate profiles, although both enzyme types oxidize all kinds of polyphenols and also ABTS, albeit with different affinity. The E.C. system also classifies tyrosinases both as E.C. 1.14.18.1 and E.C. 1.10.3.1. The characterization of tyrosinases as E.C. 1.14.18.1 is due to their hydroxylation activity, which is not observed in laccases and which is catalyzed by the same copper centers as the polyphenol/quinone oxidation.

Whereas tyrosinases have long been considered for enzymatic hair dyeing processes, laccases have been the focus of more recent patent literature. Both enzymes display the above-mentioned rather complex oxidation mechanism of polyphenols by transfering a total of four electrons in two oxidation cycles from their corresponding substrates to molecular oxygen, which is thereby reduced to water. Due to this mechanism laccases and tyrosinases can use commercial dye precursors as substrates, which is an advantage over the two-electron-transfering oxidases. The latter need additional substrates such as glucose or uric acid for the catalytic cycle, which leads to a more complex end formulation.

For dye formation from polyphenolic precursors laccases do not need a so-called mediator, as is usually required for bleaching reactions in well-established laccase-based processes such as denim treatment.

Laccases that are useful for hair dyeing have first been isolated from the fungus *Myceliophtora* by Novozymes [988]. L'Oréal has claimed a series of laccases in various patent applications for use in a broad variety of hair dyeing products (see e.g., [989]).

Up to now only few data on safety evaluation and risk assessment of oxidative enzymes in hair dyes have been published [990]. As soon as studies with nondusting and non-aerosol-forming enzyme formulations result in safe products, enzymatic hair dyes are likely to appear on the market.

5.3.2.2. Hair Waving
A number of patent applications deal with the use of enzymes belonging to the classes of oxidases, transferases, or isomerases for hair waving [991-995]. Hair waving is conventionally achieved by forming hair waves mechanically on rolls, then reducing disulfide bonds in hair keratin (e.g., with thioglycolic acid) to break the natural hair structure, and finally oxidizing thiol groups to new disulfide bonds with hydrogen peroxide. The oxidation step fixes the new structure of hair keratin induced by the rolls. This process damages the hair by swelling it and by disturbing the natural structure of disulfide bonds without completely rearranging them. In addition an unpleasant odor results from the reduction step with thioglycolic acid. The enzyme mentioned in the patent literature for waving by rearranging disulfide bonds without using thioglycolic acid or hydrogen peroxide is protein disulfide isomerase (E.C. 5.3.4.1) [996]. For oxidizing thiol groups in hair keratin the enzyme glutathione sulfhydryl oxidase (E.C. 1.8.3.3) is mentioned [995]. Transglutaminases (E.C. 2.3.2.13) are also described as potential catalysts for waving [994]. These enzymes cross-link peptide chains by forming isopeptide bonds between glutamin and lysin residues, which are present in the keratin fibers. A recent patent also describes laccases as potential catalysts in waving [997]. The radical intermediates produced by the laccase reaction are claimed to be able to oxidize thiol groups in hair keratin to form new disulfide bonds.

Although at least transglutaminase is available on a bulk scale, due to its applications in the food industry

for meat treatment, none of these enzymes has reached the product phase in hair-waving applications. The main reason seems to be the small market segment represented by waving products. Waving went out of fashion in the western world in the 1990s, and enzymes would represent a smaller niche of this small segment, i.e., milder or odorless waving products. In addition enzymes would increase the price of the product, which makes the acceptance of a niche product in the market even more difficult.

5.3.2.3. Skin Care

Enzyme baths containing bacteria and/or enzymes are well known in traditional Japanese culture for giving smooth skin. As there is no major market potential for bathing products in the western world, enzyme applications are restricted to leave-on products of the mid to premium price range. Proteases from plant origin such as papaya (papain, E.C. 3.4.22.2) and pineapple (bromelain, E.C. 3.4.22.4) are used in western face-care products to provide gentle peeling effects. Up to the early 1990s such products were used mainly in professional cosmetic applications, but more recently an increasing number of consumer products have emerged containing plant and bacterial proteases for peeling of facial skin, for example, Juvena and Eucerin (Beiersdorf) in Germany. Cosmetic manufacturers seem to be switching from plant to bacterial proteases, which can be produced more economically in fermentation processes as opposed to plant-extraction processes. The bacterial enzyme in this context is subtilisin (E.C. 3.4.21.62), an enzyme type well known for its use in laundry detergent. These proteases hydrolyze to a limited extent proteins of epidermal cells to give a peeling effect similar to that of fruit acids, which are also used for this treatment.

The application of enzymes in shower gels, which have a much bigger market potential than bathing products, was not realized, due to safety assessments performed by Procter & Gamble in the 1990s [998]. The enzyme formulation tested in this study on atopic persons led to sensitization to the enzymes. A possible explanation was based on the idea, that under the conditions of the shower application, the water formed an enzyme-containing aerosol after impinging on the soap-covered skin. The inhalation of the formed aerosol would then lead to the observed sensitization.

5.3.2.4. Toothpastes and Mouthwashes

The use of enzymes in toothpastes or dentifrice has been an existing field of enzyme application in personal care products for several decades. Some products use a combination of oxidative enzymes to fight pathogenic bacteria causing plaque, caries, and gingivitis in the mouth. Others use starch- or protein-hydrolyzing enzymes to remove food residues and prevent tooth staining. Yet most of the experience is based on the product Zendium, produced under the umbrella label “Braun Oral-B” from The Gillette Company. This toothpaste uses the enzymes amyloglucosidase (E.C. 3.2.1.3) and glucose oxidase (E.C. 1.1.3.4). Amyloglucosidase splits starch into glucose, and glucose oxidase oxidizes glucose to gluconolactone and hydrogen peroxide. Hydrogen peroxide is then used by saliva peroxidase to produce hypothiocyanite, which is a strong antibacterial agent. This hypothiocyanite-mediated antibacterial mechanism is also part of the physiological protection system of human saliva and is claimed to be enforced by the Zendium enzyme system.

The toothpastes and dry mouth treatment products Infa-Dent, Biotène, and Oralbalance (Laclede, USA) contain the enzymes glucose oxidase, lactoperoxidase, and lysozyme as well as the protein lactoferrin as antibacterial system. All these types of proteins are also part of natural saliva and develop their antibacterial effects by the production of hypothiocyanite ions from glucose, similar to the Zendium mechanism. In addition lysozyme can hydrolyze the cell walls of pathogenic bacteria such as *Streptococcus mutans* and other gram-positive bacteria involved in formation of plaque, gingivitis, and even pariodontitis. The hydrolytic action of lysozyme on cell walls is considered to be facilitated by the hypothiocyanite ions. The fourth protein of the Laclede system is lactoferrin, which is an iron-binding protein able to inhibit pathogenic bacteria by depleting them of essential iron.

The toothpaste “Rembrandt” contains the protease papain in its whitening product. Papain is isolated from papaya fruit and is claimed to remove stains from teeth together with citric acid in the patented complex Citroxain [999].

As the oldest enzyme containing brand Zendium represents a benchmark for the use of enzymes in dentifrice. Many patents have since claimed further fields of enzyme application. The patent literature refers to the use of proteases, amylases, and cellulases to degrade food residues [1000], the application of lytic or cell-wall-degrading enzymes like lysozyme and enzyme preparations from *Streptomyces* strains [1001], and

the application of mutanase and dextranase enzymes to degrade the mutane or dextrane matrices which protect *Streptococcus mutans* on the teeth [1002, 1003].

Some additional beneficial effects for the Zendium system were described in the literature with respect to prevention of recurrent aphthous ulcers when applied as mouthwash or dentifrice [1004]. Although there are clearly established oxidative and hydrolytic enzyme systems on the market and a number of additional ideas published, so far no major dentifrice product lines have been launched by companies such as Colgate and Procter & Gamble.

This also holds true for mouthwashes and chewing gums, which are included in a number of the above-mentioned patents. One reason might be that a number of enzymes with beneficial effects have been identified in the meantime in human saliva, among them lactoperoxidase, lysozyme, and lactoferrin. These findings lead to a critical evaluation of the benefit of added enzymes, unless a clear indication is given that an individual is not producing sufficient amounts of saliva or the corresponding enzymes [1005].

With regard to safety aspects the use of enzymes in mouthwash or dentifrice can be seen as a benchmark of safe enzyme usage, where no consumer problems have become known so far.

5.3.2.5. Enzymes in Cleaning of Artificial Dentures

Considering the patent literature, cleaning of artificial teeth or dentures is a known field of applications for proteases, amylases, and other glucosidases, as well as oxidative enzymes. In analogy to dishwashing products proteases, amylases, and endoglucanases hydrolyze food residues and help dissolve them from the dentures [1006, 1007]. Dextranases, mutanases, and levanases are claimed to dissolve the special sugar components of dental plaque [1008, 1009]. Oxidases are considered to help bleaching stained dentures [1010]. The denture cleaning process usually takes place for several hours or overnight at moderate temperatures, that is, good conditions for enzymatic reactions. As the enzymes do not come into direct contact with the human body, adequate safety is achieved by a careful rinsing step. For the formulation of an enzyme-containing denture-cleaning product the same safety measures apply as for formulation of enzymatic automatic dishwashing detergents.

However, apparently Corega Tabs (GlaxoSmithKline, USA) is the only product on the market at present (2002) which contains a protease to support the cleaning step.

5.3.3. Enzymes for Preservation

The use of enzymes for preservation is a known field of application in the food area. The use of lactoperoxidase (E.C. 1.11.1.7) is a typical example. Lactoperoxidase is a natural antimicrobial system in bovine milk. The structure, functions, and applications of this enzyme system have been described in detail [1011]. The interest in using such natural systems that are harmless to mammalian cells has grown beyond the use in the food area [1012]. The number of products in the industrial field is still small, where such systems are used for prevention of proliferation of contaminating microorganisms, but their importance is expected to increase.

5.3.4. Enzymes in Hard-Surface Cleaning

The use of enzymes in hard-surface cleaning is limited by safety considerations, since most of the obvious application fields for enzymatic hard-surface cleaning, e.g., in slaughterhouses, would be associated with release of enzyme aerosols. But even with these limitations some enzyme applications could be shown to be safe and feasible, because the enzymes are used in closed systems or in baths.

In the field of meat treatment, equipment for personal protection can be treated with enzymatic cleaning baths, where the equipment is dipped into or soaked in the enzyme-containing bath. The enzymes relevant for such treatment are usually proteases of the subtilisin type (E.C. 3.4.21.62), which can be identical with enzymes used in automatic dish washing detergents.

Clogging of sinks can be avoided or cleared by using proteases in combination with surfactants and potentially reducing agents in the household setting. In restaurant kitchens the use of lipases is accepted in some countries to clean fat separators by degradation of the fat into free fatty acids [1013].

5.3.4.1. Enzymes in Membrane Cleaning

With increasing use of membrane technology in the food industry and biotechnology, an increasing number of enzyme applications has arisen. In many applications enzymes were used directly in food processing to improve the membrane filtration performance or yield. The applications in membrane cleaning are focused only on the separate membrane cleaning step, with no contact between enzyme and filtered material or product. The fruit juice industry is the preferred target for such processes [1014], but over the whole range of food industries enzymes are described in cleaning of cross-flow filtration membranes [1015]. In some cases, e.g., milk or beer, contact could even lead to negative sensoric evaluation. Thus, the process has to guarantee the effectiveness of the enzymatic cleaning product and at the same time the complete absence of contact of the relevant product with the cleaning agent. Initially, the benefits of enzymatic cleaners were primarily seen on organic ultrafiltration membranes with their limited stability towards harsh cleaning conditions. In the meantime even for more stable inorganic membranes the use of enzymes like the subtilisin protease Alcalase has been described [1016].

When the process is not linked to treatment of product streams, but to wastewater treatment it must meet different requirements [1017] (see Section Enzymes in Wastewater Treatment).

5.3.4.1.1. Proteases
Subtilisin proteases (E.C. 3.4.21.62) clearly improve the performance of alkaline membrane cleaning formulations when protein is involved in fouling or blocking films on membranes. Such applications include quite different fields ranging from the food industry to biotechnological downstream processing [1018].

5.3.4.1.2. Hemicellulases
The term hemicellulases describes enzymes with activity on polysaccharides like xylan, xyloglucan, arabinan, arabinogalactan, and galactomannan. Hemicellulases are thus a group of enzymes able to degrade these gel-forming polysaccharides from plants. The most common hemicellulases are galactanases (β-1,4-galactanases, E.C. 3.2.1.89; β-1,3-galactanases, E.C. 3.2.1.90), xylanases (endoxylanases, E.C. 3.2.1.8; exoxylanases, E.C. 3.2.1.37), xyloglucanases, and mannanases (E.C. 3.2.1.78) [1019]. Hemicellulases are used in the food industry for processing of fruit and vegetables to produce juice or concentrates. They are also used for maceration of plant material for improved filtration processes.

The same enzyme preparations used for direct fruit treatment can also be used to clean membranes in processes in which plant-derived material or juices are filtered. Large molecules of hemicellulosic or pectic nature can be associated with polyphenols or proteins. Because of their size and physicochemical properties they can block the ultrafiltration membranes used for concentration of fruit juices and thus reduce the flux rates. By hydrolyzing these polymers, enzymes can result in increased filtration periods, higher flux rates, and decreased costs. Examples of such membrane cleaning applications are known in fruit juice filtration [1019] and beer filtration before bottling. Especially the beer industry offers a broad field for applications of proteases, cellulases, and hemicellulases for cleaning of membranes used for filtration of beer. In contrast to other applications, the enzymatic treatment of beer-filtration membranes profits from the fact that there is little to no variation in the composition of the membrane-blocking material [1020]. Thus, an optimized enzymatic composition for this application can be used for extensive periods of time and over a broad geographic range.

5.3.5. Enzymes Generating a pH Shift
The activity of many enzymes is linked with pH changes based on consumption or release of acid or base by the reaction catalyzed. The consequent change of pH is often an unwanted side effect with respect to the application and is normally counterweighed by carrying out the reaction under buffered conditions. In some cases this side effect is considered the primary effect that changes the pH n situ during the reaction to modify the conditions for the next reaction step. In most industrial process which require pH shifts (e.g., pulp and paper treatment) acids or bases are added due to their low price and high effectiveness. pH-shifting enzymes include hydrolases such as protease, lipase, and, most effectively, urease. The last-named hydrolyzes urea and releases ammonia, leading to a pH increase, whereas esterases usually release acids from esters, resulting in a drop in pH. The efficiency of the enzymatic reactions is often limited to shifts of 1 or 2 pH units. pH-shifting enzymes have therefore only been investigated for hair treatment processes requiring relatively gentle pH shifts.

In hair waving, urease (E.C. 3.5.1.5) is claimed to increase the pH from neutral up to 9 over several minutes. Alkaline conditions improve the reducing efficacy of thioglycolic acid, which is used for reducing disulfide

bonds [1021]. To stop this reducing reaction in hair waving, lipases have been described. They hydrolyze esters and release acids, which leads to a drop in pH. At pH values below 8 the reduction is no longer effective [1022]. Although patent literature concerning pH shifts in hair waving appeared as early as 1957 no commercial waving product contains an enzymatic pH shifting system to date. As mentioned in Section Hair Waving the hair-waving market at this moment is rather small and would need further development to allow a quite expensive enzymatic system to be introduced into products.

5.3.6. Enzymes in Cork Treatment

The traditional stopper of wine bottles consists of cork made from the bark of *Quercus suber* L. Since the supply of high-quality material is limited, an increasing quality problem with cork is observed in the market. The problem is the generation of a specific cork taste leading to a wine quality that is rejected by the consumer. The replacement of cork by artificial material is possible, but is not generally accepted by customers, who prefer the traditional methods.

Chemically the cork taste is based on the generation of trichloroanisole (TCA). TCA is not a natural compound of cork, but is derived from natural polyphenols that become chlorinated to 2,4,6-trichlorophenols during chlorine bleaching of cork. The trichlorophenol is methylated to TCA by the metabolic action of bacteria in the cork. The TCA is finally extracted from the cork by the wine during storage. This extraction efficiency depends on the type of wine and a number of physicochemical parameters related to the wine.

Although the ability to identify TCA is not evenly distributed among consumers, the substance is usually smelled and thus tasted at extremely low concentrations in the range of 3 ng/L. There are more than hundred patents and patent applications claiming to prevent the generation of TCA in cork.

The application of laccases from *Myceliophtora* (Suberase, Novozymes, Denmark) leads to oxidation and subsequent polymerization of the polyphenols and thus removes them as reaction partners for the further steps of TCA generation [1023]. The oxidative polymerization also modifies the surface of the cork, generating an increasingly hydrophobic surface with increased surface tension. Thus, the water-repellent properties of the cork are increased, and extraction of substances into the wine reduced. By this method it is possible to significantly reduce the number of bottles with wine tasting of cork.

The potential of using the polymerization reaction of phenol-oxidizing enzymes in processing and preparation of cork material has been evaluated with a focus on replacing environmentally harmful binders. This technology is closely related to the use of laccase in fiberboard production (see Section Enzymes for Polymerisation: Wood Fiberboard Production) [1024].

The bleaching of cork with hydrogen peroxide and subsequent removal of the hydrogen peroxide with catalase has also been described [1025]. In the context of this article this must be considered a classical application of enzymes in production processes.

5.3.7. Enzymes in Oil-Field Applications

Polysaccharides such as starch, β-glucan, galactomannan, mannan, or polyanionic cellulose derivatives are used in oil fields for improved removal of drilled material by the drilling equipment. The filtration-control polymer solution is designed for low viscosity. The control of the viscosity of these drilling fluids is the important factor for achieving a rapid decrease in viscosity of the resulting filter cakes. This step is often catalyzed by enzymes. The specific requirement for this application is the high temperature, which can reach 100 °C. Thus, high-temperature stability of enzymes is required for this application.

The use of cellulases and amylases has been described for drilling fluid based on polyanionic cellulose and starch [1026]. Galactomannan and mannan are also used for drilling fluids. Hence, mannanases, with their ability to hydrolyze these polysaccharides, have also been tested for oil field application. With increasing industrial and food relevance of guar gums, an increasing number of mannanases (E.C. 3.2.1.25, E.C. 3.2.1.78) has become available, and for these enzymes an increasing number of applications is claimed in patents [1027].

In deep and high-temperature wells, as well as in water-based reservoir drilling fluids, thermostable mannanases have shown their potential for control of viscosity of mannan-based formulations, thus allowing

for improved extraction of oil and gas from existing wells, resulting in greater production and increased revenue per well. A typical example are the mannanases distributed under the trade name Pyrolase. They were isolated from natural diversity and screened for high-temperature stability by Diversa (San Diego, USA). Diversa has claimed a whole range of thermostable glycosidases from *Thermococcus*, *Staphylotermus*, and *Pyrococcus* organisms for a broad range of applications [1028].

5.3.8. Enzymes in Wastewater Treatment

In many wastewater treatment processes the use of intact microorganisms producing the relevant enzymes in situ is the preferred solution over the use of isolated enzymes. Only in industrial processes with specific wastewater problems is the use of enzymes relevant. A typical example is the treatment of industrial wastewater containing hydrogen peroxide, for which catalase (E.C. 1.11.1.6) treatment is the method of choice to degrade the hydrogen peroxide fast and efficiently.

Another typical example is the use of lipases in grease separators, which is not generally allowed in all countries, since it leads to the inflow of long-chain fatty acids into the biological sewage treatment plants with consequent formation and the stabilization of foam in activated-sludge treatment tanks [1013]. The application of enzyme-containing surfactants in grease traps in restaurants also showed that such products can have detrimental effects [1029]. This leads to a situation in which lipase-containing products must be seen not as a general solution, but as a product with the need for a case by case evaluation. Industrial processes, with their more stable and reliable wastewater situation, thus seem to be the preferred site for enzyme application.

5.3.9. Enzymes for Polymerisation: Wood Fiberboard Production

Polyphenol oxidases such as laccases catalyze the oxidation of a phenolic substrates by formation of a free radical. This radical can itself represent a further substrate of laccase action, but it can also react nonenzymatically with water or other phenols with formation of polymers. The ability to form polymers has been exploited to use laccases in fiberboard production from wood.

State-of-the-art fiberboard production is based on pressing wood chips with binders like urea, melamine, or phenol – formaldehyde resin to achieve a stable structure. This process is regarded as unfavorable for humans due to the emission of formaldehyde. The phenol oxidase based process uses enzymatic oxidation of lignin phenolic hydroxyl groups to form free radicals in lignin. By further reaction of these free radicals the auto-adhesion between wood fibers can be enhanced siginificantly to form fiberboards of good mechanical stability [1030]. Beech fibers were suspended in an aqueous solution of laccase, then dried, and pressed into fiberboards. The amount of water introduced in the enzymatic process was described as the critical parameter for transfer into an industrial process.

The pretreatment of wood chips with peroxidase and hydrogen peroxide before conventional pressing has also been described as effective for obtaining fiberboards with a stability close to medium quality standards [1031].

5.4. Overview of Industrial Enzyme Applications

This chapter summarizes enzyme use in industrial applications, as presented in Sections Enzymes in Food Applications to Development of New Industrial Enzyme Applications. Table 38 gives an overview of enzymes which are commercially used in the applications described in these Sections. It is limited to commercial applications presented by the authors and does not include applications which are under development in 2003 or for which development has ceased. Although this limits the scope, the table gives a good overview of the industrial enzyme usage as presented in these chapters. A closer look at the table and analysis of enzyme usage in industrial applications in more detail reveals a total of 35 different enzyme functionalities, which are used in almost 80 different technical applications. The 35 enzyme functionalities account for numerous different products, whose number goes far into the hundreds or even thousands, worldwide. Each enzyme company certainly has more than 35 products in its portfolio.

Table 38. Enzyme applications as described in Sections Enzymes in Food Applications, Enzymes in Nonfood Applications, and Development of New Industrial Enzyme Applications

E.C. number	Name	Section	Application area	Function
		Class 1: Oxidoreductases		
		Enzymes in Baking	baking	increase gluten strength
		Enzymes in Brewing	brewing	shelf life improvement
1.1.3.4	glucose oxidase	Enzymes in Dairy Applications	dairy	milk preservation
		Other Food Applications	textile	indirect enzymatic bleaching
		Development of New Industrial Enzyme Applications	new industrial use	toothpaste
1.1.3.5	hexose oxidase	Enzymes in Baking	baking	increase gluten strength
		Enzymes in Textile Production	textile	prevention of backstaining in enzymatic stone washing
1.10.3.2	laccase	Enzymes in Pulp and Paper Processing	pulp & paper	pulp bleaching
		Development of New Industrial Enzyme Applications	new industrial use	cork treatment
		Enzymes in Brewing	brewing	shelf life improvement
1.11.1.6	catalase	Enzymes in Dairy Applications	dairy	milk preservation
		Enzymes in Textile Production	textile	hydrogen peroxide removal
1.11.1.7	peroxidase, lactoperoxidase	Enzymes in Baking	baking	dough improvement
		Enzymes in Dairy Applications	dairy	milk preservation
1.13.11.12	lipoxygenase	Enzymes in Baking	baking	whitening of breadcrumb
		Class 2: Transferases		
2.3.2.13	transglutaminase	Enzymes in Dairy Applications	dairy	texture improvement in yoghurt, whipped cream
2.4.1.5	dextransucrase	Enzymes in	brewing	production of isomalto-

		Brewing		oligosaccharide beer
		Class 3: Hydrolases		
		Enzymes in Baking	baking	bread improvement
		Enzymes in Dairy Applications	dairy	cheese clotting
3.1.1.3	triacylglycerol lipase	Enzymes in Household Detergents	laundry detergent	removal of greasy stains
		Enzymes in Pulp and Paper Processing	pulp and paper	pitch removal
3.1.1.11	pectin methylesterase	Enzymes in Fruit Juice Production and Fruit Processing	fruit juice	apple and red berry juice, citrus fuit peeling
3.1.3.8	3-phytase	Enzymes in Brewing	brewing	mashing
		Enzymes in Animal Feeds	animal feed	Phosphate realease from phytic acid in animal feed
3.1.3.26	6-phytase	Enzymes in Brewing	brewing	mashing
		Enzymes in Animal Feeds	animal feed	Phosphate realease from phytic acid in animal feed
		Enzymes in Baking	baking	antistaling in dough
		Enzymes in Fruit Juice Production and Fruit Processing	fruit juice	apple juice production
		Enzymes in Brewing	brewing	mashing, fermentation
		Enzymes in Household Detergents	laundry detergent	removal of starch-containing stains
3.2.1.1	α-amylase	Enzymes in Automatic Dishwashing	automatic dishwashing	removal of starch-containing stains
		Enzymes in Grain Wet-Milling	grain wet milling	starch hydrolysis
		Enzymes in Animal Feeds	animal feed	Improved digestion of starch in maize feed
		Development of New Industrial Enzyme Applications	new industrial use	viscosity control in oil drilling

		Applications		
		Enzymes in Textile Production	textile	textile desizing
3.2.1.2	β-amylase	Enzymes in Brewing	brewing	mashing
		Enzymes in Fruit Juice Production and Fruit Processing	fruit juice	apple juice production
3.2.1.3	glucoamylase, amyloglucosidase	Enzymes in Brewing	brewing	mashing
		Enzymes in Grain Wet-Milling	grain wet milling	hydrolysis of maltooligopsaccharides
		Development of New Industrial Enzyme Applications	new industrial use	tooth paste
		Enzymes in Brewing	brewing	fermentation
3.2.1.4	endo-1,4- β-glucanase, cellulase	Enzymes in Household Detergents	laundry detergent	removal of particular soils, softening, improved color brightness
		Enzymes in Textile Production	textile	cotton finishing, denim ageing
3.2.1.6	endo-1,4(3)- β-glucanase, cellulase	Enzymes in Animal Feeds	animal feed	improved weight gain and feed efficiency in poultry and swine
		Enzymes in Baking	baking	improved dough handling, dough stability
3.2.1.8	endo-1,4- β-xylanase	Enzymes in Animal Feeds	animal feed	improved digestibility of cereals
		Enzymes in Textile Production	textile	flax retting
		Enzymes in Pulp and Paper Processing	pulp and paper	pulp bleaching
3.2.1.55	arabinosidase	Enzymes in Fruit Juice Production and Fruit Processing	fruit juice	apple juice production
3.2.1.67	exopolygalacturonase	Enzymes in Textile Production	textile	cotton scouring
		Enzymes in	laundry	removal of guar-gum-

		Household Detergents	detergent	containing stains
3.2.1.78	endo-1,4- β-mannanase	Enzymes in Textile Production	textile	flax retting
		Enzymes in Pulp and Paper Processing	pulp and paper	pulp bleaching
		Development of New Industrial Enzyme Applications	new industrial use	viscosity control in oil drilling
		Enzymes in Household Detergents	laundry detergent	removal of particular soils, softening, improved color brightness
3.2.1.91	exo-cellobiohydrolase	Enzymes in Textile Production	textile	cotton finishing
		Enzymes in Pulp and Paper Processing	pulp and paper	mechanical pulping
3.4.X.X	protease	Enzymes in Textile Production	textile	silk degumming, wool antishrinking
		Development of New Industrial Enzyme Applications	new industrial use	artificial-denture cleaning
		Enzymes in Household Detergents	laundry detergent	removal of proteous stains
3.4.21.62	subtilisin	Enzymes in Automatic Dishwashing	automatic dishwashing	removal of proteous stains
		Enzymes in Animal Feeds	animal feed	improved digestibility of proteins in animal feeds
		Development of New Industrial Enzyme Applications	new industrial use	membrane cleaning
3.4.21.63	oryzin	Enzymes in Animal Feeds	animal feed	Improved digestibility of proteins in animal feeds
3.4.22.X	cysteine endopeptidases	Enzymes in Brewing	brewing	filtration aid
3.4.23.4	chymosin	Enzymes in Dairy Applications	dairy	cheese clotting

3.4.23.18	aspergillopepsin I	Enzymes in Animal Feeds	animal feed	improved digestibility of proteins in animal feeds
3.4.23.22	endothiapepsin	Enzymes in Dairy Applications	dairy	cheese clotting
3.4.23.23	mucorpepsin	Enzymes in Dairy Applications	dairy	cheese clotting
3.4.24.28	bacillolysin	Enzymes in Animal Feeds	animal feed	improved digestibility of proteins in animals feed
		Class 4: Lyases		
4.1.1.5	α-acetolacetate decarboxylase	Enzymes in Brewing	brewing	diacetyl removal for flavor enhancement
4.2.2.2	pectate lyase	Enzymes in Textile Production	textile	cotton scouring
4.2.2.10	pectin lyase	Enzymes in Fruit Juice Production and Fruit Processing	fruit juice	apple and red berry juice, citrus fuit peeling
		Class 5: Isomerases		
5.3.1.5	xylose isomerase	Enzymes in Grain Wet-Milling	grain wet milling	fructose production

The discrepancy in number of enzyme functionalities and enzyme products can be explained by the fact that most of the enzymes are tailored for one specific application. Enzyme products appear in different formulations, each designed so that the enzyme performance is optimized for usage in a particular application.

The differentiation of a small number of functionalities into a broad range of products is also achieved by selecting and producing enzymes from different sources (e.g., animals, bacteria, yeast, fungi) with a wide variety in intrinsic properties such as pH/activity profile or stability under a certain set of conditions. The need for such differentiation becomes evident when looking at the different conditions used in the industrial enzyme field. The fruit juice industry uses processes in which enzymes have to operate at an acidic pH and rather low temperature. On the other hand, the detergent industry uses alkaline conditions up to pH 11, a temperature range between 40 °C and 60 °C, and a denaturing regime of surfactants and sequestrants that would be fatal for most enzymes. So far, the most extreme temperature stability has been obtained for α-amylases in the starch liquefaction, which is performed under pressure above the boiling point of water. The α-amylase has to operate for a few minutes at around 106 – 108 °C followed by a 2 h cooling period to 90 – 95 °C.

Table 38 also shows the wide chemical diversity in terms of the catalyzed reactions, as illustrated in Figure 92: the majority of the industrial enzymes that are described in the previous chapters are from the class of the hydrolytic enzymes. Almost 70 % of industrial enzymes belong to this group, class 3 of the IUB enzyme classification system.

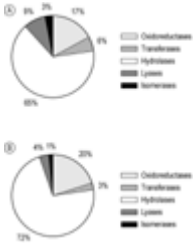

Figure 92. A) Distribution of technical enzyme functionalities over IUB enzyme classes. B) Distribution of technical enzyme applications over IUB enzyme classes.
[Full View]

The oxidoreductases (class 1) are the second largest contributor to the group of industrial enzymes (15 %). Classes 2, 4, and 5 are almost unused in industrial applications, and there is no example yet of an industrial enzyme of class 6. From this it is clear that the huge diversity of catalysts for chemical reactions which is available in nature's laboratory is completely under-utilized in industrial enzyme applications. There is thus a huge range of natural catalytic potency that awaits exploitation by humans.

A closer look at new but as-yet uncommercialized applications reveals a change in trend. The need to use new, unexploited functionalities, especially in new developments, is evident. The use of oxidoreductases in particular seem to be receiving more attention.

Another observation is that some enzymes are used in mixtures rather than as individual enzyme preparations. A good example of these is cellulases, which are used in many textile applications as complex mixtures of different types of cellulases. The basic chemical reaction is of course the same in all cellulases: the cleavage of the β-1,4 bonds between the sugar moieties. The difference lies, however, in the regioselectivity (exo or endo) and in the difference in enzyme binding to different regions of the variety of complex cellulase substrates. In most textile applications of cellulases, the synergistic action of the enzyme mixtures is needed to obtain the desired effect.

α-Amylases are good example for the use of one catalytic chemical reaction in numerous industrial applications. α-Amylases seem to be the most versatile enzymes in the industrial enzyme sector, with eight different applications, no doubt due to the abundance of starch. The widespread occurrence of this biopolymer and the ability of α-amylases to catalyze the hydrolytic degradation of amylose or amylopectin by cleavage of the α-1,4 bond between the glucose molecules results in a multitude of applications for this enzyme. Applications of α-amylase include starch degradation in the liquefaction step of glucose, maltose, or high-fructose corn syrup production; the degradation of starch-containing stains in detergent and dishwashing applications; and the clarification of fruit juice and beer. The use of α-amylase as an antistaling agent in baking is slightly different. In this application, the action of α-amylase modifies the three-dimensional polymer structure of amylose and thus changes the material properties of bread by influencing the crumb structure. Since starch is also used to service oil wells, thermostable α-amylases are used to control the viscosity of oil well fluids.

The beauty of the selectivity of enzyme use in industrial enzyme applications is again best exemplified by the use of α-amylase in textile desizing. In this particular application α-amylases are used to remove a protective layer of starch which is applied to the cellulose fibers to avoid damage during mechanical treatment in the manufacturing process. α-Amylases break down the amylose and amylopectin of starch, in which the glucose monomers are linked together by α-1,4 bonds, whereas the glucose moieties of cellulose, which are linked through β-1,4 bonds, are not hydrolyzed. In effect the starch size can be removed in a mild process without disrupting the integrity of the cellulose polymer. The cellulose polymer can subsequently be attacked by cellulases to give, for example, a stonewashed look. In chemical terms the stereochemistry of one carbon atom in the monomer is enough to direct the selectivity of the enzymes which attack the polymers.

It is impossible for this chapter to cover all uses of all enzymes in all industrial areas: many of the enzymes listed in Table 38 are also used as digestives, nutraceuticals, and probiotics; in the upgrading of waste materials; in the bating of hides; to accelerate silage; or in the recovery of silver from spent films, just to summarize some of the more exotic uses of enzymes in industrial applications.

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Wolfgang Aehle, Genencor International B. V., Leiden, The Netherlands
Herbert Waldmann, Max-Planck-Institut für molekulare Physiologie, Dortmund, Germany
Carsten Schultz, EMBL, Heidelberg, Germany
Harald Gröger, Degussa AG, Hanau-Wolfgang, Germany
Carlo Dinkel, SiChem GmbH, Bremen, Germany
Karlheinz Drauz, Degussa AG, Hanau-Wolfgang, Germany
Georg-Burkhard Kresse, Roche Diagnostics GmbH, Penzberg, Germany
Karl Wulff, Boeringer Mannheim GmbH, Tutzing, Germany
Günther Henniger, Roche Diagnostics GmbH, Penzberg, Germany
Christoph Kessler, Roche Diagnostics GmbH, Penzberg, Germany

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

6. Nonindustrial Enzyme Usage

6.1. Enzymes in Organic Synthesis

6.1.1. Introduction

A modern survey of catalysis would not be complete without a discussion of enzyme-catalyzed transformations. The fact that enzymes represent the first chiral catalysts should ensure them a place next to the most sophisticated man-made catalysts. Although it may seem odd that the catalyst must be isolated from cells, tissue, or organs and that sometimes whole organisms are employed, the rapid development and success of biocatalytic processes [1032-1052] for the synthesis of natural products, pharmaceuticals, and agrochemicals [1053] has immensely increased their acceptance among organic chemists. They have in part found industrial applications, especially since many technical problems, such as enzyme stabilization, enzyme immobilization, and cofactor regeneration, have been solved. Also, the availability of enzymes has been greatly improved: of the more than 3000 enzymes known, several hundred are commercially accessible or can be obtained in sufficient purity by easy, well-developed procedures. Moreover, enzymes can in principle be produced in large quantities by recombinant DNA techniques, and their properties can be manipulated in favor of a desired performance by directed evolution or DNA shuffling [1054], [1055]. Furthermore, enzyme techniques can be combined with high-throughput screening [1054] to give very powerful tools for testing for improved efficacy.

Due to the characteristic properties of enzymes compared with the chemical reactions usually employed, enzymatic transformations are exceptional in several respects:

1. They operate under mild conditions in a pH range of 5 – 8 at temperatures around 20 – 40 °C in aqueous media.
2. Many of them tolerate organic solvents [1041], [1048], [1056-1058].
3. Enzymes are highly efficient catalysts, accelerating reactions by factors of 10^5 – 10^{12} compared to the corresponding uncatalyzed reactions.
4. Enzymes often combine a high selectivity for the reactions they promote and the structures they recognize with a broad substrate tolerance.
5. Besides their chemoselectivity many enzymes exhibit a high degree of diastereo-, regio-, and/or enantioselectivity.

As a result of the increasing demand for complex, biologically relevant substances and the problems connected with their efficient preparation, it is mainly the enantio-discriminating properties of enzymes that have led to the exponential development of the use of such complex molecules in organic synthesis.

6.1.2. Examples of Enzymatic Conversions

6.1.2.1. Syntheses by Means of Hydrolases

Due to their ready availability and the ease with which they can be handled, hydrolytic enzymes have been

widely applied in organic synthesis. They do not require coenzymes, are reasonably stable, and often tolerate organic solvents. Their potential for regioselective and especially for enantioselective synthesis makes them valuable tools [1046], [1051], [1053], [1056-1058].

Acylation. Regioselective acylations of polyhydroxylated compounds such as carbohydrates, glycerols, steroids, or alkaloids have been carried out with lipases, esterases, and proteases [1049, 1050], [1059-1063]. One example is the *Candida antarctica* lipase B (CALB, E.C. 3.1.1.3, immobilized on acrylic resin) catalyzed monoacylation of the signalling steroid ectysone (**1**) giving selectively the 2-*O*-acetate **2**. Using vinyl acetate for this transesterification irreversibly pushed the reaction to the product side, since the liberated enol instantaneously isomerizes to acetaldehyde [1064]. The sometimes unfavorable aldehyde is avoided when 1-ethoxyvinyl acetates [1065], [1066], trichloro- or fluoroethyl esters [1067], [1068], oxime esters [1069], or thioesters [1070] are employed for the "quasi-irreversible" reaction courses.

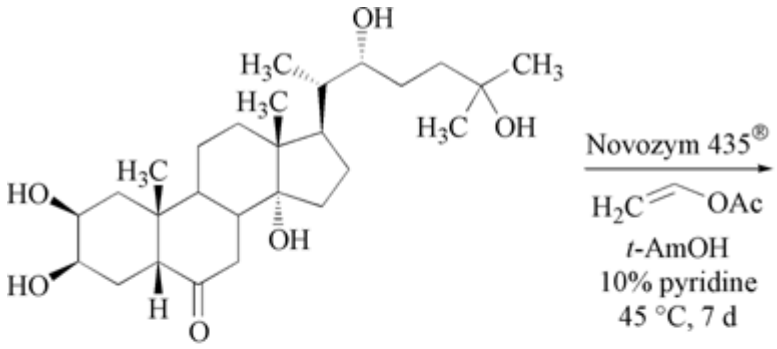

1

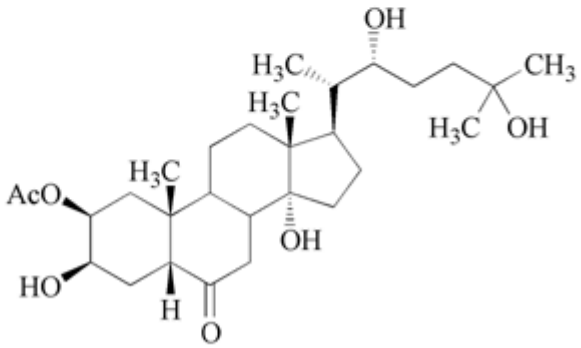

2

95%

Deacylation. In the synthesis of highly phosphorylated phosphoinositide derivatives, regioselective hydrolysis of one of three butyrate groups by a lipase from *Candida rugosa* (CRL, formerly *C. cylindracea*, E.C. 3.1.1.3) in pH 7.8 buffer containing 5 % methanol gave the key intermediate **3**, unfortunately without yielding enantiomeric excess (Scheme 1) [1071].

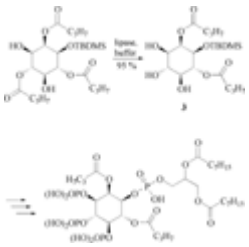

Scheme 1. Regioselective hydrolysis of the less hindered butyrate by a lipase
[Full View]

Enantioseparation. An intensively utilized strategy is the kinetic resolution of racemic mixtures [1056-1058], either by stereoselective (*trans*)-esterification or -amidation reactions in organic solvents or hydrolysis in

aqueous media. This approach is particularly attractive when the unconverted enantiomer is racemized in the process and reconsumed. In the example shown in Scheme 2 ethyl acetate was used as the acylating agent and triethylamine as the solvent [1072]. The enzyme was again the immobilized lipase from *Candida antarctica* B (Novozym SP 435). Racemization of the unconverted enantiomer (*S*)-**4** was achieved with palladium on charcoal. Isolated yields of **5** were 64 % with an ee of 99 %.

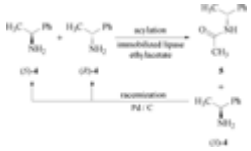

Scheme 2. Enantioselective acylation of racemic amines and in situ racemization of the unconverted enantiomer
[Full View]

Another advantageous use of hydrolytic enzymes is the enantioselective hydrolysis of prochiral substrates, making use of the ability of these biocatalysts to discriminate between enantiotopic groups. An elegant way to obtain both enantiomers (*R*)-**7** and (*S*)-**7** from the prochiral diol **6** is the combination of porcine pancreatic lipase (PPL, E.C. 3.1.1.3) mediated hydrolysis and esterification (Scheme 3) [1073].

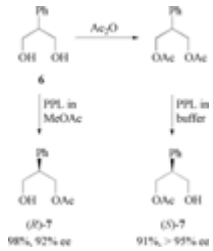

Scheme 3. Preparation of 2-phenylpropane-1,3-diol monoacetates using enantiotopic hydrolysis versus transesterification
[Full View]

Pig liver esterase (PLE, E.C. 3.1.1.1) has found extensive use for the hydrolytic cleavage of methyl or ethyl esters of prochiral carboxylic acids.

The usefulness of hydrolases such as the immobilized lipase from *Pseudomonas species* (LIP, E.C. 3.1.1.3) to provide chiral substances for elaborated synthetic strategies is demonstrated by the quantitative acetylation of *meso*-Diol **8** to the monoacetate **9**, an important intermediate in the preparation of synthon **10** (Scheme 4) used for prostaglandin syntheses [1074].

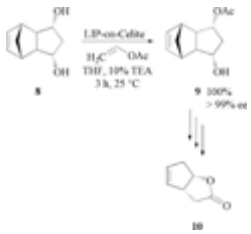

Scheme 4. Efficient enantioselective monoacetylation of a meso-diol by an immobilized lipase in organic media
[Full View]

As can be deduced from the examples above, availability and use of immobilized enzymes has significantly increased over the past years. When used in organic solvents immobilized enzymes often show a greatly augmented catalytic activity due to increased surface area and reduced denaturation [1057]. Three techniques are frequently employed to immobilize enzymes: simple absorption, entrapment, and covalent immobilization. Especially lipases, with their need for an aqueous – lipid interface proved to be favorably altered in their activities. Lipases entrapped in hydrophobic sol – gels made from alkylsilanes and $Si(OMe)_4$ showed up to 100-fold increased activity in organic solvents [1075-1077]. Their use in water has also been demonstrated [1078]. Covalent fixation of enzymes to a matrix is often achieved by cross-linking an absorbed biocatalyst with glutaraldehyde [1079], [1080]. Surprisingly, this methodology is particularly powerful when small cross-linked enzyme crystals (CLEC) are formed [1081-1083].

Several hydrolytic enzymes other than esterases have been applied for synthetic purposes. One important

subject is the chemoenzymatic preparation of amino acids. An industrial method for the synthesis of non-natural D- or L-amino acids employs the enzymatic hydrolysis of hydantoins, prepared by Bucherer – Bergs condensation using D- or L-hydantoinase (see Section Commercial Applications) [1084], [1085]. Another efficient method to prepare natural and non-natural amino acids is a two-step synthesis involving a palladium-catalyzed amidocarbonylation (Eq. 1) to afford racemic *N*-acyl amino acids, followed by enantioselective hydrolysis with various acylases [1086].

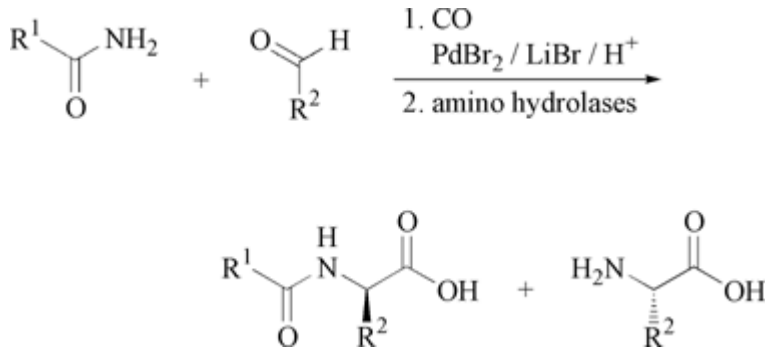

The enzymatic hydrolysis of a broad range of nitriles to the corresponding amides and acids is documented [1087-1089]. These conversions are effected directly by nitrilases (E.C. 3.5.5.1) or by successive action of a nitrile hydratase (E.C. 4.2.1.84) and an amidase (E.C. 3.5.1.4). Most of these enzymes are unstable, and whole-cell preparations are preferred. However, recently a purified nitrile hydratase preparation without amidase activity was shown to convert several 2-arylpropionitriles enantioselectively to the corresponding optically active amides (Eq. 2) [1090].

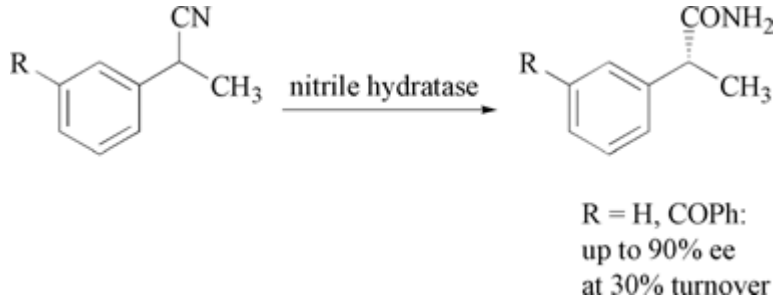

The important use of penicillin acylases (E.C. 3.5.1.11) is discussed in Section Commercial Applications.

6.1.2.2. Reduction of C ═O and C ═C Bonds

Carbonyl reductases utilizing NADH or NADPH as cofactor have been successfully applied for the asymmetric reduction of carbonyl groups. These enzymes deliver a hydride from the reduced cofactor in accordance with Prelog's rule to the *re* face or in rare cases to the *si* face of the carbonyl group. The most commonly used alcohol dehydrogenase for this purpose is that from horse liver (HLADH, E.C. 1.1.1.1), which preferentially reduces cyclic carbonyl compounds [1091]. The completely regioselective reduction of the 3,5-dioxocarboxylate **11** using the alcohol dehydrogenase from *Lactobacillus brevis* (E.C. 1.1.1.2) and employing substrate-coupled recycling of the cofactor gave the alcohol **12** (Scheme 5) [1092] in excellent enantiomeric excess and good yields.

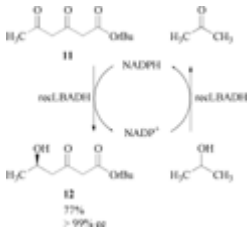

Scheme 5. Recombinant *Lactobacillus brevis* alcohol dehydrogenase (recLBADH) catalyzes the stereoselective reduction of 3,5-dioxocarboxylates with in situ NADPH regeneration [Full View]

Lactate dehydrogenases are another important subgroup of carbonyl reductases. They reduce α-oxo acids enantiospecifically to α-hydroxy acids. Both L- and D-selective enzymes (E.C. 1.1.1.27 and E.C. 1.1.1.28,

respectively) are available, giving access to both enantiomers of various α-hydroxy acids [1093-1095].

The use of purified nicotinamide cofactor-dependent dehydrogenases in preparative reduction reactions requires an efficient regeneration of the expensive and unstable NADPH cosubstrate. For the recycling of NADH the formate/formate dehydrogenase system is well established, which combines the use of an inexpensive substrate (HCOONa) with the formation of the volatile coproduct CO_2 [1096], [1097]. The use of cofactor-bound CLECs is an elegant way to avoid the instabilities of enzyme and cofactor [1083]. An example for an alternative single-enzyme system is shown in (Scheme 5). Through oxidation of the added propan-2-ol the dehydrogenase (E.C. 1.1.1.2), which also catalyzes the reduction of the carbonyl compound, regenerates NADPH [1098], [1099]. To avoid troublesome and often problematic cofactor regeneration, whole-cell-mediated reductions are currently utilized extensively. By this means the necessary cofactors and their regeneration is provided by immanent pathways. For example, baker's yeast catalyzed reduction of ketone **13** was a key step for the preparation of phorocantholide 1 **14** (Scheme 6) [1100].

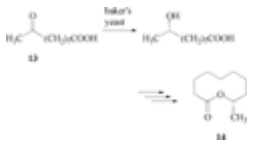

Scheme 6. Ketone reduction using baker's yeast as the key step in the synthesis of phorocantholide 1 [Full View]

Baker's yeast has also been used for the highly stereoselective reduction of prochiral 2,2-dimethyl-1,3-cyclohexanedione [1101]. The resulting *S*-configured hydroxycyclohexanone served as a versatile starting material for the synthesis of various natural products [1046], [1102].

NADH-dependent enoate reductases are also interesting tools for synthetic applications. They stereoselectively reduce conjugated carbonyl double bonds and those activated by electron-withdrawing substituents, as shown in Equations (3) and (4) [1103-1105].

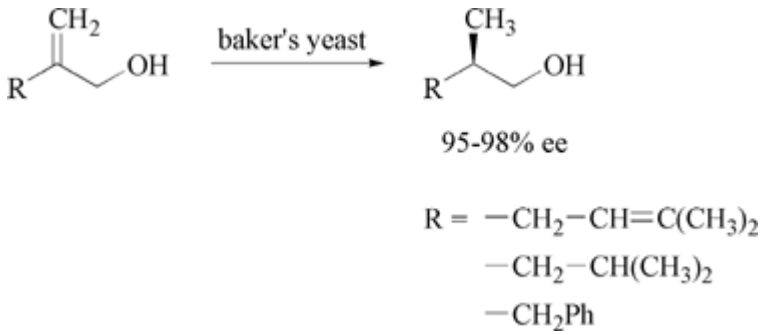

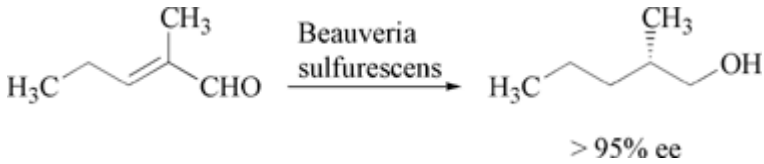

6.1.2.3. Oxidation of Alcohols and Oxygenation of C —H and C ═C Bonds

Horse liver alcohol dehydrogenase (HLADH, E.C. 1.1.1.1) has been used for the enantiotopic oxidation of *meso*-diols, preferably by applying an ammonium α-ketoglutarate/glutamate dehydrogenase recycling system to regenerate NAD^+. The initially produced asymmetric hydroxyaldehyde cyclizes to the hemiacetal, which is further oxidized to the lactone **15** [1106]. Scheme 7 presents an illustrative reaction with an alternative NAD^+ regeneration path [1107].

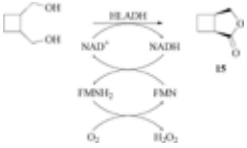

Scheme 7. Horse-liver alcohol dehydrogenase (HLADH) catalyzed stereoselective oxidation of a meso-diol with in situ NAD^+ regeneration through a flavin mononucleotide (FMN) oxygen recycling system
[Full View]

Mutants of *Pseudomonas putida* were found to exhibit arene dioxygenase activity, which has been exploited in whole-cell reactions for the regio- and enantioselective preparation of cis-dihydrodiols starting from benzene, substituted benzenes, and polycyclic or heteroaromatic compounds [1108-1111]. The products are invaluable precursors for natural product synthesis, as exemplified in Scheme 8 [1112], [1113].

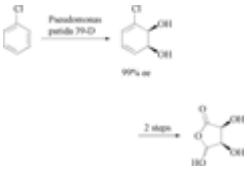

Scheme 8. Enzymatic dihydroxylation of chlorobenzene using whole cells give access to enantiomerically pure *cis*-dihydrodiols
[Full View]

The enzymatic counterpart of the Baeyer – Villiger reaction is catalyzed by a flavine-containing monooxygenase (cyclooxygenase) in the presence of oxygen and NAD(P)H. Using whole cells, racemic mixtures of cyclic ketones were kinetically resolved or enantio- and regioselectively oxidized [1114]. Recently, a cell-free system with cyclohexanone monooxygenase (CHMO, E.C. 1.14.13.22) and recombinant $NADP^+$-dependent formate dehydrogenase (FDH, E.C. 1.2.1.4) for cofactor regeneration was successfully introduced to prepare the chiral β-lactone **17** from prochiral 4-methylhexanone **16** (Scheme 9) [1115].

Scheme 9. Baeyer – Villiger oxidation of a prochiral cyclic ketone to a lactone with in situ cofactor regeneration (CHMO = cyclohexanone monooxygenase, FDH = formate dehydrogenase)
[Full View]

Regio- and stereospecific monohydroxylations of elaborate molecules have been achieved with monooxygenases that are in most cases membrane-bound cytochrome P450 enzymes. Because these are coupled with a complex redox system for electron transport from NADPH, the reactions are preferably carried out with whole cells [1116], [1117]. The same is true for enantioselective enzymatic epoxidations. An impressive example is the preparation of **18** in 78 % yield and high enantiomeric excess [1118].

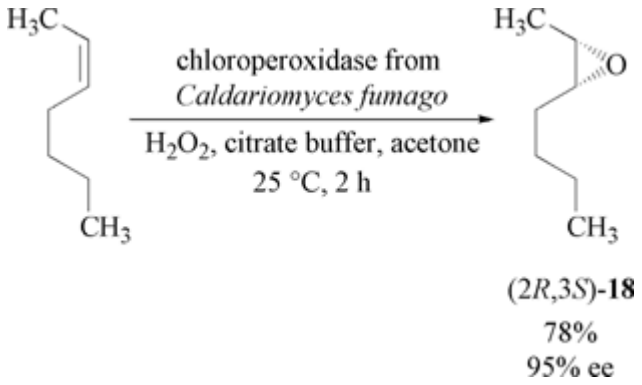

6.1.2.4. C — C Coupling

A wide range of natural and non-natural monosaccharides has been generated by exploiting the catalytic capacity of aldolases, which perform reactions equivalent to nonenzymatic aldol additions [1119-1124].

More than 20 aldolases have been identified so far and can be divided into three main groups, accepting either dihydroxyacetone phosphate (DHAP), acetaldehyde, or pyruvic acid, and phosphoenolpyruvate as nucleophilic methylene component. A common feature is their high stereocontrol in the formation of the new C —C bond. As presented in Scheme 10 all four possible *vic* diols are accessible by selection of the appropriate DHAP aldolase [1033], [1034], [1125-1127], all of which show a distinct preference for the two stereocenters and a broad substrate tolerance for the aldehyde component.

Scheme 10. Stereoselective C —C bond formation by means of DHAP-utilizing aldolases [Full View]

In a particularly elegant example (Scheme 11), 5-deoxy-5-ethyl-D-xylulose (**19**) was prepared from glycerol by a four-step, one-pot chemoenzymatic reaction sequence. The key step was the aldol reaction catalyzed by fructose 1,6-bisphosphate aldolase (FruA, E.C. 4.1.2.13) from *Staphylococcus carnosus* [1128].

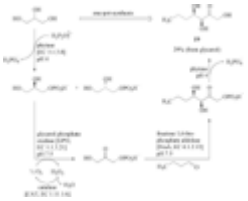

Scheme 11. Careful pH control can switch the enzyme activities for phosphorylation, oxidation/aldol addition and dephosphorylation. Only readily available starting materials are necessary [Full View]

Other C —C coupling reactions have been published [1032-1034]. A prominent example is the enzymatic Diels – Alder reaction which affords (–)-solanapyrone A (**20**). The enzyme solanapyrone synthase from extracts of the phytopathogenic fungus *Alternaria solana* has hence been named “Diels-Alderase” [1129].

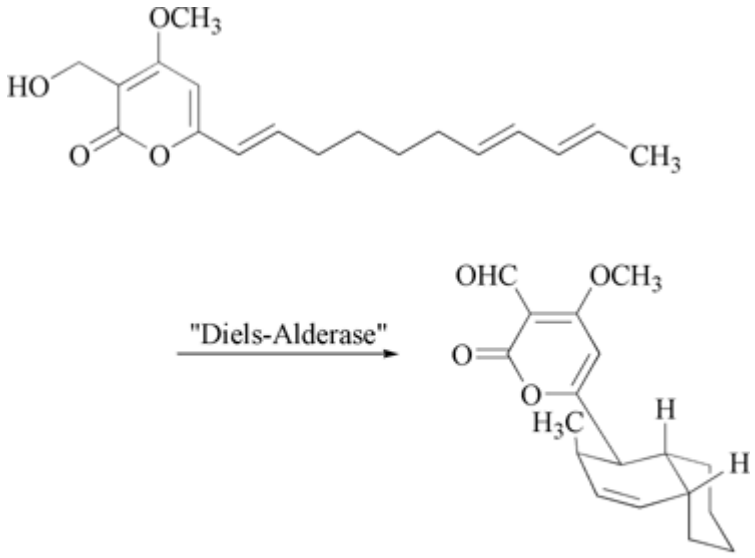

20
51%, > 98% ee

6.1.2.5. Formation of Glycosidic Bonds

The most frequently applied concepts for the enzymatic formation of oligosaccharides and glycoproteins involve the utilization of glycosyl transferases or *exo*- and *endo*-glycosyl hydrolases (glycosidases) [1130-

1134]. Glycosyl transferases require expensive and often unstable sugar nucleotides as activated donor substrates, a major problem besides the poor availability of the enzymes. However, transferase-catalyzed glycoside synthesis leads to regio- and stereochemically pure products in high yields. Glycosidases, on the other hand, are inexpensive and easy to handle. The reactions are mostly carried out under kinetic control using simple glycosides as donors. Generally, one or more regiomers are obtained in moderate yields, all of them having the same anomeric configuration of the newly formed glycosidic linkage.

For instance α-fucosides and α-sialosides, which are both difficult to produce by means of classical carbohydrate chemistry, have been synthesized by use of the corresponding nitrophenyl glycoside donors by applying pig liver α-fucosidase (E.C. 3.2.1.51) [1135] and the sialidase (E.C. 3.2.1.18) from *Vibrio cholerae* [1136], respectively.

Starting from lactose as donor and *N*-acetyl glucose as acceptor substrate the β-galactosidase from *Bacillus circulans* (E.C. 3.2.1.23) gives exclusively *N*-acetyl lactosamine **21** [1137], which has been subjected to sialyltransferase (E.C. 2.4.99.6) catalyzed conversion to the trisaccharide Neu5Ac- α(2,3)Gal- β(1,4) GlcNAc (**22**; Scheme 12). By combining these two enzymatic steps the reaction is pushed toward the product, and the problem of low yields related with the galactosidase-mediated reaction could be overcome [1138].

Scheme 12. Enzymatic synthesis of a sialyl trisaccharide by combination of glycosidase- and glycosyl-transferase-catalyzed reactions
[Full View]

To avoid problems associated with transferase-catalyzed reactions, i.e., the high price of the sugar nucleotides and possible inhibition by the released nucleotide phosphates, schemes for the regeneration of the sugar nucleotides were developed [1139-1141]. Thus, multienzyme systems have been successfully used in one-pot reactions for the large-scale synthesis of several oligosaccharides [1130-1134], [1139-1141].

A particular challenge is enzymatic in vitro glycoprotein synthesis. Initially, the way was paved for the preparation of several glycoforms of ribonuclease B (RNase B) (**24** and **25**). First, a monoglycosylated form of RNase B (**23**) was prepared by deglycosylation of the heterogeneous sugar moities [1142]. Subsequently, the homogeneously modified sialyl Lewis X RNase B was generated with the help of a series of glycosyl transferases (Scheme 13). In an alternative approach *endo*-glycosidase A (*endo* A, E.C. 3.2.1.96) transferred a mannose-rich glycoside core from an *N*-linked aspartate donor to the mono-glycoside of RNase B to give the homogeneous protein [1143].

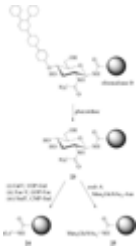

Scheme 13. Glycosylation of ribonuclease B using either glycosyltransferases or endo-glucosidase (endo-A) systems results in unnatural or homogeneously glycosylated proteins, respectively
[Full View]

6.1.2.6. Enzymatic Protecting Proup Techniques

Often, not only highly selective but also extremely mild methods to introduce and remove protecting groups are needed. This applies in particular for the construction of complex, polyfunctional molecules, e.g., oligosaccharides, peptides, nucleotides and their conjugates, as well as for the synthesis of alkaloids, macrolides, and other natural products. In this respect enzymes offer valuable opportunities complementing the already elaborated classical techniques. The chemo-, regio-, and stereoselectivity of enzymes allows

them to establish otherwise unattainable “quasi-orthogonal” protecting group arrangements.

The aforementioned hydrolase-catalyzed regioselective acylations and deacylations of a wide array of polyhydroxylated compounds has been addressed in numerous cases [1049], [1050].

In a recent advanced application of enzymatic protecting group techniques for the construction of complex multifunctional and sensitive targets, the synthesis of the characteristic *S*-palmitoylated and *S*-farnesylated lipohexapeptide of human *N-Ras* protein **26** was constructed. Compound **26**, which would not withstand standard procedures under basic or acidic conditions, could be obtained by following two different routes, both employing enzyme-labile protecting groups. The choline ester group was employed for C-terminal chain elongation [1144], while the *p*-acetoxybenzyloxycarbonyl (AcOZ) urethane group served well in a N-terminal coupling strategy [1145] (Scheme 14).

Scheme 14. Biocatalytic syntheses of a *S*-farnesylated and *S*-palmitoylated *N*-Ras fragment
[Full View]

The spectrum of enzymatic deprotecting tools (Fig. 93) was further broadened by the introduction of the *p*-phenylacetoxybenzyloxycarbonyl (PhAcOZ) group for N-terminal protection [1146], [1147] and the *p*-phenylacetoxybenzyl ester (PAOB) group for C-terminal protection of peptides [1148], respectively. Both groups are readily cleavable by penicillin G acylase (E.C. 3.5.1.11) and have been successfully used in lipopeptide and glycophosphopeptide synthesis [1148], [1149]. The glucosyloxycarbonyl (Gloc) group is a suitable substrate for α- and β-glucosidases (E.C. 3.2.1.20 and 3.2.1.21) [1150].

Figure 93. Enzyme-cleavable protecting groups for carboxy acids (PAOB) and amines (PhAcOZ and Gloc)
[Full View]

6.1.3. Enzyme-Analogous Catalysts

The capability of enzymes to transform organic substrates smoothly, in good yields, and highly selectivity, have stimulated the desire of organic chemists to tailor catalysts with similar features for specific reactions.

In this respect, reactions catalyzed by monoclonal antibodies have been the focus of much interest [1151-1153]. One of the first antibody-supported syntheses on a gram scale and with an enantiomeric access of up to 99 % was the kinetic resolution of aldol adducts. The products resulting from the retro-aldol reaction served as precursors in the synthesis of epothilones A and B [1154], [1155].

For the production of an effective monoclonal antibody it is crucial to raise it from a well-designed transition-state analogue. Besides not resembling the product too much in order to avoid product inhibition, this hapten should mimic the transition state as closely as possible to maximize its stabilization by the antibody generated. Utilizing the extraordinary specificity of the immune system by this means, the combining site of the antibody is programmed to exhibit a specific catalytic activity. This technique has been directed toward the development of selective reactions which are chemically difficult to achieve, for example, “disfavored”, and for which no suitable enzyme is available.

Therefore, several reactions were subjected to various antibody catalyzes, e.g., ester and enol ether cleavage,

transesterification, ketone reduction, Cope rearrangement, ring-closure reaction via epoxide opening, and Diels – Alder cycloaddition [1156-1158], [1159]. An exceptional reaction is the antibody-catalyzed Robinson annulation of triketone **28** to the Wieland – Miescher ketone **29** on a preparative scale. Surprisingly, even the alkylation of diketone **27** with methyl vinyl ketone was catalyzed by the same antibody, but at moderate rates (Scheme 15) [1160].

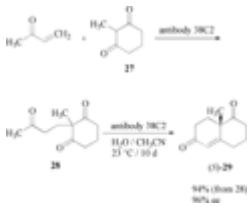

Scheme 15. Preparative enantioselective synthesis of a Wieland – Miescher ketone by an antibody [Full View]

Ribonucleic acids (RNAs) were employed as catalysts in the synthesis of amides and esters, peptide bond formation, and Diels – Alder reactions. The approach suffers from the fact that it requires the reactants to be either RNA itself or a compound covalently tethered to RNA. However, there is one prominent example in which RNA acts as a true catalyst in a bimolecular Diels – Alder cycloaddition reaction without tethering the substrates [1161-1163].

The chemical modeling of active sites of enzymes can help to gain deeper insight into their catalytic mechanisms and might also lead to enzyme substitutes which are reliable and easy-to-handle tools for organic synthesis. Recently, it was discovered that a single proline can mimick the active site of an aldolase. High ee values were obtained with acetone or hydroxyacetone, with generation of one (e.g., **30**) or two (e.g., **31**) new stereocenters in fair to excellent yields, respectively. This reaction seems to be well suited to industrial application, because of its simplicity and low cost [1164-1168].

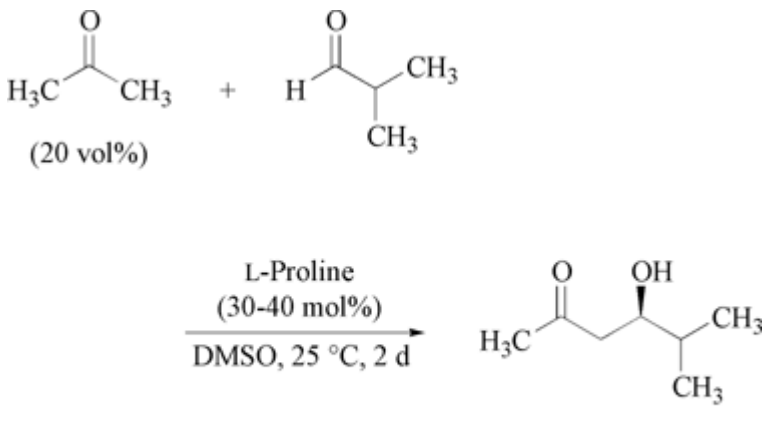

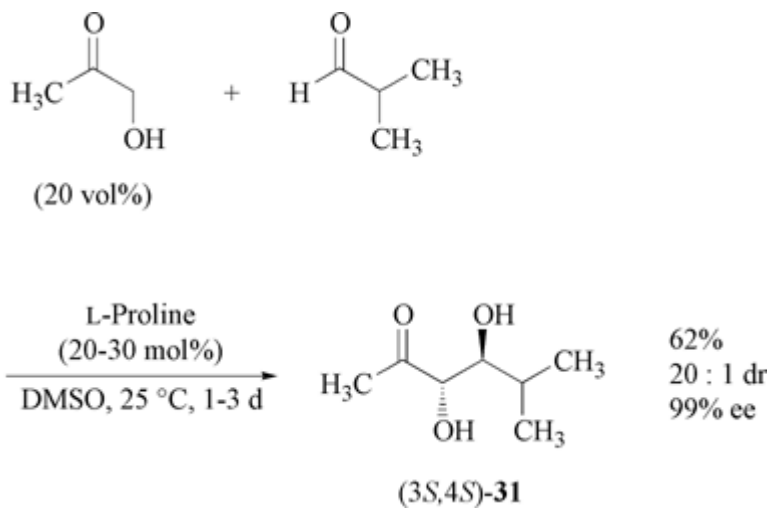

A more recent topic has so far received little attention, but will most probably develop to great importance in bioorganic chemistry: the use of enzymes for ligating peptides [1169], [1170]. Peptide cyclization to form the antibiotics tyrocidine A and gramicidin S was achieved with the thioesterase domain of tyrocidine synthetase [1169]. α-Chemotrypsin (E.C. 3.4.21.1) catalyzed the condensation of larger peptide fragments to form an active sequence of the human thyroid anchoring protein Ht31 [1170]. The synthetic peptide was able to abolish forskolin-induced outward chloride currents in guinea pig ventricular myocytes.

6.1.4. Commercial Applications

6.1.4.1. General

Since the mid-1970s biotransformations have become a well-established tool in the fine-chemicals industry. Biocatalytic systems, including crude and purified enzymes as well as whole-cell systems performing highly selective reactions under mild conditions, are widely used, especially in synthesis and production of biologically active compounds in the agrochemical and pharmaceutical sectors.

Biocatalysis has already been proven in many cases to overcome specific synthetic problems. Thus, the need for hazardous reagents or the formation of specific impurities or byproducts can be avoided. Furthermore, biocatalysis can provide excellent regio- and stereoselectivity, and much higher turnover numbers than with chemical catalysts are possible. Therefore, it is not surprising that biocatalysis has also found an important place among industrial methods for the production of enantiomerically pure compounds (EPC; see Figure 94). The reported sales of EPCs exceeded $\$100 \times 10^9$ in 1999 for the first time.

Figure 94. Industrially applied approaches to enantiomerically pure compounds (EPCs) [Full View]

In large-scale synthesis, cost-effectiveness and minimization of environmental problems are the most important factors in successful process development. The costs for the catalytic system, its separation, and in many cases its recovery are very important and have to be optimized. A critical figure in cost calculation is the “catalyst consumption number” (comparable with the TON in pure homogeneous catalysis) per kilogram of product produced. In the case of an enzyme, this number expresses how many units of enzyme or grams of biomass are deactivated and have to be replaced during production of 1 kg of product. Therefore, stability is often much more important than activity. The most suitable biocatalyst, which can be a native or an immobilized enzyme, or a whole-cell system, must be chosen carefully. Investigation of the total reaction engineering (including the reaction system) for optimal reaction conditions, reaction and reactor kinetics, as well as good downstream processing, offers the highest chance for success.

The examples selected here are commercially relevant and focus on specific topics discussed above. They demonstrate the advantage of biocatalysis in the conversion of functional groups and configuration, resolution of racemates, and asymmetric synthesis.

6.1.4.2. Nonstereoselective Biocatalytic Reactions

Even for simple, nonstereoselective conversions of functional groups, biocatalytic methods can be highly competitive (and superior) compared with chemical standard processes.

This has been impressively demonstrated by Nitto Chemical Industries (now Mitsubishi Rayon) with the biocatalytic manufacture of acrylamide, an important building block for polymers and copolymers, produced in quantities of over 200 000 t/a [1171]. The chemocatalytic route to acrylamide (**32**) uses a reduced Raney copper catalyst for hydration. This metal-catalyzed process has been shown to be superior to the acid-catalyzed hydration, but catalyst poisoning and wastewater problems due to heavy-metal content cause some problems [1172], [1173].

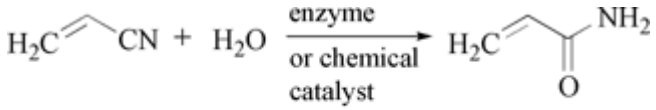

32

Nitto Chemical Industries developed and established an alternative way to produce acrylamide by an enzymatically catalyzed addition of water to acrylonitrile, using immobilized cells of *Rhodococcus* sp. N-774 and later an optimized mutant strain of *Pseudomonas chloraphis* B 23 [1172], [1174], [1175]. Using a highly improved cobalt-containing nitrile hydratase from *Rhodococcus rhodochrous* J1, it is possible to reach up to 700 g L^{-1} acrylamide in the reaction solution. The reaction temperature for the copper-catalyzed hydration is around 100 °C, whereas the enzymatic process is carried out below 10 °C. Nitto acrylamide production using bacterial nitrile hydratase is estimated to produce around 30 000 t/a, consumes less energy, and avoids heavy-metal problems in the wastewater. In Figure 95, the chemo- and biocatalytic production processes for acrylamide are compared.

Figure 95. Comparison of the microbial and conventional chemical process for acrylamide production [Full View]

In addition, the conversion of aromatic nitriles to carboxylic acids or amides has been applied for the production of fine chemicals. For example, Lonza reported the biocatalytic manufacture of niacinamide starting from 3-cyanopyridine [1176]. It is said that this process will be carried out in the future on a multithousand tonnes per year scale.

Another interesting commercial nonstereoselective application is the biocatalytic conversion of methyl groups to carboxylic acids. This new technology, also developed at Lonza, is already successfully applied on an industrial scale in the production of 5-methylpyrazine-2-carboxylic acid, a versatile building block for pharmaceuticals [1176].

It is noteworthy that the biocatalytic formation of a methyl group via decarboxylation can also be realized on a technical scale. Decarboxylation of L-asparagine in the presence of an aspartate-4-decarboxylase (E.C. 4.1.1.12) allows an efficent synthesis of L-alanine [1177].

6.1.4.3. Biocatalytic Resolution Processes

On an industrial scale, resolution of racemates is still in many cases the preferred way of producing enantiomerically pure compounds. If chiral-pool compounds are not available as optically active starting materials, a cost comparison of asymmetric and resolution processes often reveals advantages for the latter. Among the top 15 optically active pharmaceuticals and their intermediates, a chemical asymmetric process could be established only in the case of naproxen. For all others, diastereomeric and biocatalytic resolution are the methods of choice, apart from fermentation, for creating optically pure compounds.

Normally, only one enantiomer is needed; therefore, highly effective methods for racemizing the undesired one must be integrated into the process. Besides the well-known examples of separating diastereomeric salt pairs by crystallization, biocatalytic resolution is gaining more and more interest. There are a number of key factors which must be optimized for successful industrial development; for example, substrate and enzyme costs (including recovery); chemical, optical, and space-time yields; number of reaction steps and positioning of the resolution step within the whole process; and downstream processing, together with the necessary equipment.

In biocatalysis, hydrolases are the most important class of enzymes for carrying out enzymatic resolutions. Many hydrolases, such as esterases, lipases, epoxide hydrolases, proteases, peptidases, acylases, and amidases, are commercially available; a substantial number of them are bulk enzymes [1178-1180]. Resting-cell systems, if they are not immobilized, are used in dilute suspensions and could be handled as quasi homogeneous catalysts.

In the following examples some industrial applications of biocatalytic resolutions are decribed.

Enzymatic Hydrolysis of Acyclic Amides. Production of Amino Acids Using Amino Acylases: The Degussa Process

Enantiospecific hydrolysis of *N*-acylamino acids by amino acylase I from *Aspergillus oryzae* (E.C. 3.5.1.14) is the best established large-scale process for L-amino acids. Both L- and D-amino acids could be produced according to this technology. Tanabe has developed a process using an immobilized enzyme [1181], [1182], whereas at Degussa a process using native amino acylase I from *Aspergillus oryzae* in homogeneous solution is applied. For this purpose, a new technology using an enzyme membrane reactor (EMR) has been designed, based mainly on the work of Kula and Wandrey [1183-1185].

The substrate specificity of the amino acylase from *Aspergillus oryzae* is very broad, and a wide range of proteinogenic and nonproteinogenic *N*-acetyl and *N*-chloroacetyl amino acids are transformed by the enzyme. The enzyme membrane reactor (Figure 96) is operated continuously as a recycle reactor, and the enzyme is retained by a hollow-fiber ultrafiltration membrane (molecular weight cut off: 10 000 Da).

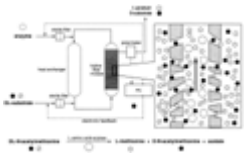

Figure 96. Recycle reactor used for amino acylase resolution of *N*-acetyl-D,L-amino acids
[Full View]

Since the early 1980s this process has been scaled up to a production level of many hundreds of tonnes per year. Scheme 16 shows the Degussa process for manufacturing L-methionine (**33**). The biocatalyst is produced in bulk quantities, and its operational stability is high, so this continuous EMR acylase process demonstrates high efficiency, especially on a large scale [1186], [1187].

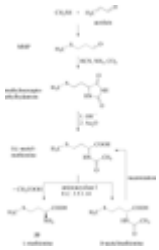

Scheme 16. Degussa L-methionine process using acylase
[Full View]

Production of Amino Acids Using Amidases: The DSM Process. DSM has developed a widely applicable industrial process for production of enantiomerically pure amino acids by enantioselective hydrolysis of racemic amino acid amides. These precursor compounds can easily be obtained by alkaline hydrolysis of α-aminonitriles. The L-specific amidase (aminopeptidase) was found in *Pseudomonas putida* (ATCC 12633), and shows a broad substrate specificity. Methods for racemizing the undesired enantiomer are also integrated in the process (Scheme 17) [1188].

Scheme 17. DSM amidase process for enantiomerically pure α-amino acids
[Full View]

Production of Carboxylic acids Using Amidases: (*S*)-2,2-Dimethylcyclopropanecarboxylic Acid [(*S*)-Ibuprofen]. (*S*)-2,2-Dimethylcyclopropanecarboxylic acid is an intermediate for the synthesis of cilastatin, a

dehydropeptidase I inhibitor. Lonza has developed an industrial two-step biotransformation process starting from racemic 2,2-dimethylcyclopropane nitrile (**34**; Scheme 18). The stereoselective amidase was found in *Comomonas acidovorans* and cloned into a fast-growing *E. coli* strain which produces large amounts of the biocatalyst [1189], [1190].

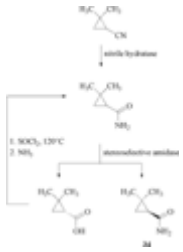

Scheme 18. Enzymatic production of (S)–2,2–dimethylcyclopropane carboxylic acid [Full View]

For the production of the analgesic (*S*)-ibuprofen from racemic ibuprofen amide, (*S*)-selective amidase-containing microorganisms belonging to the genera *Brevibacterium* and *Corynebacterium*, in particular the species *Brevibacterium* R312, *Corynebacterium* N771, and *Corynebacterium* N774, have been found to be suitable biocatalysts [1191].

Production of 6-APA, 7-ADCA, and 7-ACA Using Amidases: 6-Aminopenicillanic acid (6-APA, **35**), 7-aminocephalosporanic acid (7-ACA, **36**), and 7-aminodesacetoxycephalosporanic acid (7-ADCA, **37**) are the most important precursors for the highly active semisynthetic β-lactam antibiotics. The worldwide capacity thereof exceeds 20 000 t/a.

Before 1980, 6-APA and 7-ADCA were produced only by chemical deacylation of penicillin G and phenacetyl-7-ADCA, which could be obtained by chemical ring expansion of penicillin G (Scheme 19). Since 1980, enzymatic processes avoiding the use of environmentally critical reagents in large amounts have been developed [1192-1194]. It is noteworthy that, in addition to environmental advantages, the enzymatic route was found to be considerably cheaper than the chemical process [1195]. The most prominent enzyme is penicillin G amidase (Pen G amidase, E.C. 3.5.1.11) [1195], [1196], which preferentially hydrolyzes penicillin G and penicillin V as well as phenacetyl-7-ADCA without affecting the β-lactam ring. A broad variety of cultures can be used for large-scale production of this enzyme, including *E. coli* [1197]. The penicillin G amidase from *E. coli* enzyme is one of the best studied acylases, and has been characterized as a heterodimer with molecular weights of 19 500 and 60 000 Da for the two subunits α and β, respectively [1198]. Nowadays, Pen G amidase (E.C. 3.5.1.11) is applied in immobilized form. A preferred solid support is Eupergit C, a bead polymer containing reactive epoxide functions which covalently bind to the enzyme [1199]. It is noteworthy that Eupergit-immobilized penicillin G amidase can be reused up to 600 times. The enzymatic hydrolytic reactions on an industrial scale are performed at 30 – 40 °C and pH 7.5 – 8.0 [1200].

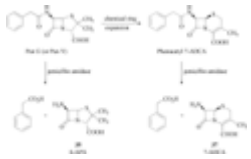

Scheme 19. Enzymatic production of 6–APA and 7–ADCA [Full View]

For the industrial production of 6-APA and 7-ADCA (see Scheme 19) excellent conversion rates of at least 98 % were reported. In addition, yield of isolated product and selectivity are high. Typically the industrial reaction is carried out batchwise. Detailed summaries of production data are available in the literature [1200].

7-ACA (**36**) [1201] is produced in a three-step process using two enzymes, starting from cephalosporin C (Scheme 20). After deamination by a D-amino acid oxidase, the α-ketoadipyl-7-ACA spontaneously decarboxylates, and the remaining glutaryl side chain is hydrolyzed by glutaryl-7-ACA-acylase.

Scheme 20. 7–ACA production starting from cephalosporin C
[Full View]

6.1.4.3.1. Enzymatic Acylation of Amino Groups
Production of Ampicillin and Cephalexin. Besides being used for the production of the precursors 6-APA and 7-ADCA, penicillin amidases are also able to couple 6-APA and 7-ADCA with D-phenylglycine ester or amide in a kinetically controlled enzymatic peptide synthesis to form ampicillin or cephalexin (**38**, **39**). These reactions (Scheme 21) have a great potential for being commercialized in the near future [1202].

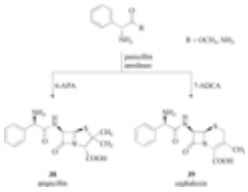

Scheme 21. Enzymatic synthesis of ampicillin and cephalexin
[Full View]

Production of Amines Using Lipases: The BASF Process. Biocatalytic resolution has been applied efficiently by BASF for the manufacture of optically active amines, such as phenylethylamine [1203], [1204]. The process is based on a highly stereoselective resolution of racemic amines by means of an acylation reaction in the presence of a lipase as catalyst (Scheme 22). The products are obtained in high yields and with excellent enantioselectivities. The undesired enantiomer can be racemized subsequently. Thus, starting from a racemate, efficient access (with theoretically 100 % yield) to the (*R*)- and (*S*)-amines is available. This technology, which is said to be carried out on a greater than 1000 t scale, has been extended recently by BASF to the production of chiral alcohols.

Scheme 22. BASF process for the production of amines
[Full View]

6.1.4.3.2. Enzymatic Hydrolysis of Hydantoins
Production of D-Amino Acids. Amino acids with D configuration play an important role in synthesis and production of biologically active compounds. The most prominent D-amino acids are D-phenylglycine and 4-hydroxyphenylglycine, which form the side-chain of the semisynthetic β-lactam antibiotics ampicillin, amoxicillin, and cephalexin. For production of both of these amino acids, resting-cell biocatalysts have been developed, containing a D-hydantoinase and sometimes also a D-carbamoylase and a racemase [1205]. The D-hydantoinase is able to transform a racemic 5-monosubstituted hydantoin into the D-hydantoic acid; the L-hydantoin can racemize at the alkaline pH of the reaction mixture. When the carbamoylase and racemase are both present, a complete transformation to the D-amino acid occurs. The racemase can assist hydantoin racemization if the process itself is too slow under the prevailing reaction conditions. Without the carbamoylase the liberation of the D-amino acid must be performed by nitrosation of the hydantoic acid. This is not a highly desirable reaction; additionally, HNO_2-sensitive functional groups are degraded.

Another example of the biocatalytic production of D-amino acids involves the ureido amino acid D-citrulline (D-Cit, **40**), which is needed for the luteinizing hormone-releasing hormone (LHRH) antagonist Cetrorelix, first synthesized by A. SCHALLY *et al.* and further developed by Asta Medica [1206-1208]. Cetrorelix consists of five D-amino acids, among which D-citrulline (D-Cit) is particularly difficult to synthesize. For the synthesis of D-Cit (**40**), a hydantoinase/carbamoylase system from *Agrobacterium radiobacter* was chosen. The synthesis starts from the inexpensive proteinogenic amino acid, L-arginine (L-Arg, see Scheme 23), which has the appropriate carbon chain length and also a nitrogen atom attached in the 5-position. In the presence of arginase, L-Arg is deguanylated quantitatively to L-ornithine (L-Orn). This reaction is much more effective than any chemical hydrolysis.

Scheme 23. Production of D-Cit starting from L-Arg
[Full View]

L-Orn then is then treated with two equivalents of cyanate to give dicarbamoylated L-Cit, which can easily be cyclized to give the L-Cit hydantoin. When the *Agrobacterium* biocatalyst with hydantoinase and carbamyolase activity is added to an aqueous alkaline solution of this hydantoin, complete transformation to D-Cit takes place. The ee of 99.8 % demonstrates that the enzymes act stereospecifically [1209]. At Degussa D-Cit (**40**) has been produced in quantities of several tens of kilograms. This hydantoinase/carbamoylase system offers an excellent method for producing D-amino acids starting from racemic or suitable L-precursors without additional racemization or subsequent chemical steps. Particularly for synthesis of side-chain-functionalized D-amino acids, this method is superior to other alternatives [1210], [1211].

6.1.4.3.3. Enzymatic Hydrolysis of Lactams

Production of L-Lysine: The Toray-Process. For some years Toray's enzymatic process for L-lysine (L-Lys, **41**) was competitive with fermentation. This chemoenzymatic L-Lys production was established with a capacity of 5000 – 10 000 t/a. The key intermediate is α-amino- ε-caprolactam (ACL), produced from cyclohexanone in a modified Beckmann rearrangement. The enantiospecific hydrolysis forming L-Lys is based on two enzymes: L-ACL hydrolase opens the ring of ACL to give L-Lys, and in the presence of the ACL racemase D-ACL is racemized. By incubating D,L-ACL with cells of *Cryptococcus laurentii* having L-ACL lactamase activity together with cells of *Achromobacter obae* with ACL racemase activity, L-Lys could be obtained in a yield of nearly 100 % (Scheme 24) [1212-1215].

Scheme 24. Toray enzymatic L-lysine process using α-amino-caprolactam as starting material
[Full View]

This process has now been abandoned because, on a 100 000 t/a scale, fermentation of L-Lys based on sugars has proven to be superior to the Toray process.

Production of 2-Azobicyclo[6.2.2.1]hept-5-en-3-one. A highly selective enzymatic hydrolysis of racemic 2-azobicyclo[6.2.2.1]hept-5-en-3-one has been developed by Chirotech [1216], [1217]. The selective hydrolytic cleavage of the (+)-lactam led to the desired, unhydrolyzed (–) enantiomer with a high ee of >98 %. The single (–) enantiomer (**42**) functions as an intermediate in the synthesis of the Glaxo Wellcome anti-HIV drug Abacavir (Scheme 25). It is noteworthy that this process operates at a substrate concentration of > 500 g/L.

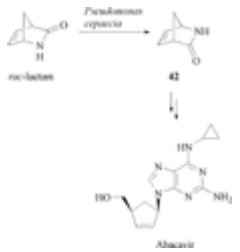

Scheme 25. Resolution of a racemic lactam
[Full View]

6.1.4.3.4. Enzymatic Hydrolysis of C —O Bonds

Esterases, lipases, and proteases are widely used in enzymatic resolution of racemic carboxylic ester substrates. Industrially relevant examples are the manufacture of (D)- β-acetylmercaptoisobutyric acid, the optically active side chain of the angiotensin-converting enzyme (ACE) inhibitor captopril (**43**; Scheme 26) [1218], [1219], and synthesis and production of (*S*)-2-arylpropionic acids such as (*S*)-naproxen and (*S*)-ibuprofen, which are very important nonsteroidal antiphlogistic and analgesic drugs [1220-1224].

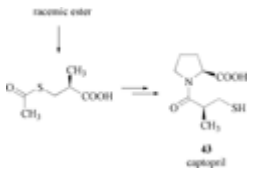

Scheme 26. Production of captopril
[Full View]

In Andeno's diltiazem synthesis (Scheme 27), an enantioselective enzymatic hydrolysis of an epoxy ester by a lipase is the key step, creating the necessary optically active intermediate **44** [1225-1227].

Scheme 27. Production of diltiazem
[Full View]

The enzymatic aspartame process of Tosoh can also be regarded as a reactive deracemization of D,L-phenylalanine methyl ester (D,L-Phe-OCH_3). In the key step the metalloproteinase thermolysin (E.C. 3.4.24.27) couples only the L isomer with (*Z*)-aspartic acid in a thermodynamically controlled peptide synthesis; the remaining D-Phe-OCH_3 is racemized and reused [1228].

Many more examples can be found in the literature [1229], [1230]. Another esterase process is described in more detail here. It is the optical resolution of D,L-pantoyl lactone (D,L-PL) by a fungal lactonase. The D isomer is a key component in the vitamin pantothenic acid and in coenzyme A. D,L-PL is easily produced by adding HCN to the aldol product of formaldehyde and isobutanal with subsequent acidic lactonization. For commercial production of D-pantoyl lactone (D-PL) the racemate can be resolved by chiral amines.

Recovery of the chiral auxiliary and racemization of the L isomer are troublesome. Therefore, much work has been done on developing more efficient resolution processes or asymmetric syntheses. YAMADA *et al.* screened new fungal aldonolactonases which are able to hydrolyze D-PL specifically to D-pantoic acid (D-PA, see Figure 97).

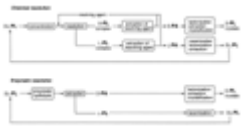

Figure 97. Comparison of resolution of diastereomeric salt-pairs and enzymatic resolution for producing D,L-pantoyl lactone
[Full View]

This D-acid can easily be separated from the remaining L-PL, which can be racemized and recycled. Relactonization of D-PA yields the desired D-PL.

Among various lactonase-containing fungal microorganism, *Fusarium oxysporum* AKU 302 showed the highest activity. Compared with asymmetric chemical or microbial reduction methods or conventional chemical resolution, this lactonase process exhibits superior performance [1231-1233].

6.1.4.3.5. Enzymatic Hydrolysis of Nitriles
Production of (*R*)-Mandelic acid: The BASF Process. That the conversion of nitriles into carboxylic acids also represent a suitable industrial resolution process has been reported by BASF [1036], [1234]. Based on racemic mandelonitrile, a stereoselective hydrolysis led to (*R*)-mandelic acid (**45**) in high yield and with high enantioselectivity (Scheme 28). It is noteworthy that during the process the cyanohydrin enantiomers are racemized in situ when the appropriate pH conditions are chosen. Thus, in spite of starting from a racemate a (theoretically) quantitative enzymatic conversion to the optically active (*R*)-mandelic acid can be achieved. The production of (*R*)-mandelic acid (**45**) is carried out on a scale of several tons. The hydrolysis of nitriles for the stereoselective manufacture of fine chemicals has also been applied by Celgene [1235].

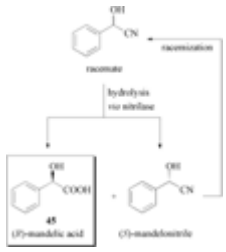

Scheme 28. BASF process for the production of (*R*)–mandelic acid
[Full View]

6.1.4.3.6. Enzymatic Cleavage of *threo* Aldol Products
A chemoenzymatic process for the preparation of D- and L-*threo*-phenylisoserine and derivatives thereof has been developed by Celgene. This process integrates an Erlenmeyer condensation as an initial chemical step which produces racemic *threo* compounds, and a subsequent enzymatic cleavage of the undesired enantiomer with regeneration of the starting materials [1235]. As a biocatalyst, a novel D- or L-aldolase enzyme has been used (Scheme 29). Enantiomerically pure *threo*-phenylisoserines find a wide range of applications, e.g., for the preparation of the taxol side chain or a variety of phenicol antibiotics.

Scheme 29. Celgene-process for the production of L-threo–phenylisoserine
[Full View]

6.1.4.4. Biocatalytic Asymmetric Synthesis
Enzymatic reduction, oxidation, and ligase or lyase reactions, provide numerous examples in which prochiral precursor molecules are stereoselectively functionalized. Ajinomoto's β-tyrosinase-catalyzed L-dopa process [1236], [1237], the formation of L-carnitine from butyro- or crotonobetaine, invented by Lonza [1238], and the IBIS naproxen route oxidizing an isopropylnaphthalene to an (*S*)-2-arylpropionic acid are

representative, examples for many successful applications of enzymatic asymmetric synthesis on an industrial scale. A selection of recent industrial contributions in this field are summarized below.

6.1.4.4.1. Biocatalytic Reductive Amination of C ═O Bonds

A very effective method for synthesizing α-amino acids with bulky side chains is the cofactor-dependent enzymatic reductive amination of α-keto acids, which is carried out by Degussa. For example, bulky-side-chain amino acids such as L-*tert*-leucine (L-Tle, **46**) and L-neopentylglycine (L-Npg) could be produced starting from the corresponding α-keto acid. In this reaction two enzymes, namely, leucine dehydrogenase (LeuDH, E.C. 1.4.1.9) and formate dehydrogenase (FDH, E.C. 1.2.1.2), and the cofactor NAD^+ act simultaneously in the transformation of the keto acid to the L-amino acid. The cofactor is used in catalytic amounts and is recycled during the reaction (Scheme 30). The use of LeuDH as amino acid dehydrogenase guarantees L stereospecificity, as no D enantiomer could be detected.

Scheme 30. L-Tle production by cofactor-dependent reductive amination
[Full View]

This type of reaction forming non-natural L-amino acids uses enzymes from the metabolism of proteinogenic amino acids, which additionally have an unexpectedly versatile substrate specificity by also accepting highly sterically hindered α-keto acids. L-Tle and L-Npg are of growing importance because of their extensive use as a building blocks in pharmaceuticals and as chiral auxiliaries in asymmetric synthesis [1239-1241].

The synthesis of chiral α-amino acids starting from α-keto acids by means of transamination has been reported by NSC Technologies [1242], [1243]. In this process, L-aspartate serves as the amino donor. The broad substrate specificity allows the conversion of numerous keto acid substrates.

Another type of asymmetric reduction of a broad range of carbonyl compounds was developed by Celgene [1235]. Using ketones as substrate in the presence of a highly efficient transaminase as biocatalyst gave the corresponding amines in high yield and with high enantiomeric excess. It is noteworthy that transaminases have been found for the synthesis of (*R*)-amines as well as (*S*)-amines.

6.1.4.4.2. Biocatalytic Hydrocyanation of C ═O Bonds

A further important asymmetric biocatalytic synthesis is the hydrocyanation of aldehydes [1244-1246] for the production of cyanohydrins, which are intermediates for a broad variety of pharmaceuticals and agrochemicals. For example, (*R*)-mandelonitrile is a versatile intermediate for the synthesis of (*R*)-mandelic acid, and (*S*)-*m*-phenoxybenzaldehyde cyanohydrin is a building block for the preparation of pyrethroids.

For the production of (*S*)-*m*-phenoxybenzaldehyde cyanohydrin (**47**) DSM established an enzymatic hydrocyanation process on an industrial scale (Scheme 31). An efficient (*S*)-oxynitrilase biocatalyst (E.C. 4.1.2.11) has been developed. This enzyme is derived from the plant *Hevea brasiliensis*, and has been cloned and overexpressed in a microbial host organism [1247]. In the presence of this biocatalyst the desired product (**47**) has been obtained with high enantioselectivity.

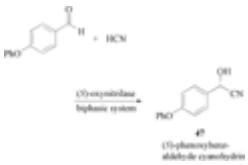

Scheme 31. Asymmetric hydrocyanation using (*S*)-oxynitrilases
[Full View]

6.1.4.4.3. Biocatalytic Addition of Water to C ═O Bonds

The enantioselective addition of water to fumaric acid gives malic acid [1248]. This highly efficient

biocatalytic reaction is catalyzed by a fumarase (E.C. 4.2.1.2). Although substrate tolerance of this enzyme is narrow, high enantioselectivities are obtained.

6.1.4.4.4. Biocatalytic Amination of C ═C Bonds
A biocatalytic enantioselective addition of ammonia to a C ═C bond of an α, β-unsaturated compound, namely, fumaric acid, makes the manufacture of L-aspartic acid possible on industrial scale. This process, which is applied by, e.g., Kyowa Hakko Kogyo and Tanabe Seiyaku, is based on the use of an aspartate ammonia lyase (E.C. 4.3.1.1) as a biocatalyst [1249]. Another comparable reaction is the asymmetric biocatalytic addition of ammonia to *trans*-cinnamic acid, which is a technically feasible process for the production of L-phenylalanine [1250].

6.1.5. Outlook
The increasing use of biocatalysis in research and industrial production, especially in the developing field of chirotechnology is foreseeable. Advances in the following areas seem most promising:

1. Elucidation of the active site and the reaction mechanism of enzymes by a combination of X-ray analysis and computer-aided modeling may allow precise prediction of enzyme – substrate interaction.
2. Screening of microorganisms from exotic environments may lead to the discovery of promising new enzyme activities.
3. Using modern gene technology the catalytic properties of enzymes could be improved or custom-tailored for specific reactions. Even the design and engineering of synthetic catalysts may become feasible.
4. Considering the significant progress in the field of microbial pathway engineering, it is expected that the importance of this technology will also grow rapidly in the future. Microbial pathway engineering leads to biocatalysts which allow the production of desirable metabolites. Thus, similarly to fermentation, this method represents a biocatalytic one-pot synthesis in which several consecutive enzymatic steps are involved. Numerous important industrially relevant contributions in this field have been already made, for example, successful pathway engineering for carotenoids and polyketides [1251]. Further representative examples are the production of indigo, vitamin C, and 1,3-propandiol [1252].

There is a strong need for new biocatalytic methods producing enantiomerically pure compounds which are the basis for synthesis and production of biologically active compounds of agrochemical and pharmaceutical importance.

6.2. Therapeutic Enzymes
6.2.1. Requirements for the Use of Enzymes in Therapy
There is a long history of successful use of enzymes and other proteins as drugs [1253]. Although pharmaceutical research and development activities in the field of recombinant therapeutic proteins are mainly focused on antibodies, cytokines (such as interferons), and hormones (such as erythropoietin) rather than on enzymes, there are nevertheless three broad areas where enzymes are successfully employed in therapy: (1) to replace enzymes which are missing or defective as a consequence of inherited genetic disease; (2) to replace enzymes that are deficient or present only in inadequate quantity as a consequence of acquired disease in the organ(s) where they are normally synthesized; and (3) to provide a specific biological effect which is dependent on the catalytic activity of the enzyme [1254]. Furthermore, some enzymes are used in a supportive manner in combination with other therapies, or for the removal of toxic or otherwise undesirable substances from the bloodstream or a body tissue.

To be effective as therapeutic agents, the enzymes must meet a number of requirements:

- They must reach their site of action in the body and tissue compartment.
- They must be active under the conditions present at the intended site of action. This includes substrate and cosubstrate/coenzyme availability, appropriate redox potential, adequate pH value for activity, absence (or saturation) of inhibitors.
- They must be sufficiently stable to ensure adequate pharmacokinetics, i.e., the required concentration/activity level for the necessary time period.
- They must be sufficiently soluble to allow application as a solution if administration is via the intravenous, intramuscular, or subcutaneous route.

- They must be sufficiently pure to avoid eliciting unwanted side reactions caused by contaminants, e.g., microbial endotoxins, pyrogens, or other harmful materials.
- There must be a therapeutic effect that depends on the particular activity of the enzyme applied.

Of course, enzymes as active pharmaceutical ingredients of drugs must fulfill a number of further requirements which are equally relevant for low molecular weight compounds used as drugs:

- Safety depending on the indication/disease. This means that the therapeutic benefit must surpass any adverse reactions, especially immmunological complications,
- Efficacy, i.e., the therapeutic effect must be established in controlled clinical trials according to FDA/EMEA regulations,
- Convenience of the application form, again depending on the indication or disease: in the treatment of life-threatening diseases, convenience is clearly of less importance than for, e.g., digestion aids.

6.2.2. Coping with Peculiar Protein Properties

Enzymes have unsurpassed substrate and reaction specificity, as well as exceptional efficacy as biological catalysts. Furthermore, they have been optimized by evolution for action in living organisms. Especially in the case of genetically encoded enzyme defects, enzyme therapy will make rational therapy possible.

However, due to their specific molecular properties, enzymes (or, in general, proteins) would by no means be considered “ideal” therapeutic agents for several reasons:

- Instability: enzymes are polypeptides and therefore inherently labile with respect to heat, extreme pH values, denaturants, and biological degradation. This may lead to limited shelf life and short half-lives in the human body, e.g., due to proteolysis in the stomach and intestine or receptor-mediated clearance from the blood followed by proteolytic degradation in the liver.
- Presence of endogenous inhibitors which are especially common in blood [1255] and may destroy the catalytic activity of the enzyme.
- Bioavailability: since the molecular surface of soluble enzymes is hydrophilic, proteins generally cannot pass through biological membranes. Thus enzyme drugs are not orally available, and, unless the enzyme is intended to act in the oral cavity or in the gastrointestinal compartment, an intravenous, intramuscular, or subcutaneous route of administration must be chosen. In some cases administration by inhalation using a nebulizer has proved possible for enzymes acting in the respiratory tract.
- Cell permeability: Since enzymes will not usually enter tissue through the intestinal wall or into human cells from the bloodstream, protein drugs are unsuitable for intracellular action. Therefore, at present many of the most successful applications of enzymes as drugs are extracellular, such as purely topical uses, the removal of toxic substances, and the treatment of life-threatening disorders within the bloodstream. Most enzymes presently used in therapy belong to the class of hydrolases, but there are also examples of oxidoreductases, transferases, and lyases.
- Immunogenicity: nonhuman proteins are foreign to the human body and therefore are expected to be immunogenic. When injected into the bloodstream, they may induce the formation of antibodies and cellular immune response [1256]. This might prevent repeated or prolonged application of the same protein drug (in contrast, immunogenicity is even desired when proteins are used as vaccines [1257]). In some cases, it appears that the immunogenicity of nonhuman enzymes produced by recombinant DNA technology is lower than that of their natural forms, probably because of the higher purity and less stressful purification process. However, even with recombinant human enzymes, immunogenicity may be an issue if muteins are applied, or if the production process does not ensure the absence of misfolded enzyme species.
- Allergic reactions may occur on inhalation of enzyme powders, especially with proteases.
- Manufacturing and characterization: Enzymes are complex macromolecules that require difficult production processes, leading to high cost of the drug. Furthermore, in contrast to small molecules, only limited analytical characterization of the final enzyme product is possible, even with the highly sophisticated protein-analytical techniques available nowadays. This makes quality control at the end-product level difficult, so that process changes will require bioequivalence testing.

On the other hand, there are some important advantages in the use of enzymes in therapy:

- Fewer adverse reactions (including teratogenic effects) for protein drugs than for chemically synthesized low molecular weight compounds, for which the safety profile of new chemical entities (NCEs) cannot be predicted.
- Shorter development times: since the advent of recombinant protein drugs in the 1980s, it has initially been found that the total time needed for research and development of a new drug from project initiation until marketing approval generally was significantly shorter, and the attrition rate, i.e., the fraction of development projects which must be terminated during preclinical or clinical development, is lower than for traditional low molecular weight compounds. However, the mean development times for biopharmaceuticals rose dramatically during the 1990s, and seem to be converging for biopharmaceuticals and NCEs since 1996 [1258], [1259].

Therefore, a number of challenges remain for the future development of enzymes for therapeutical use:

- Improve stability: Approaches to prolong the half-life in the circulation include coupling of enzymes to polyethylene glycol (PEGylation) [1260], [1261]; chemical cross-linking; the development of sustained-release forms and depot systems, e.g., by microencapsulation of the enzyme in polymers such as poly (lactide) – poly(glycolide) copolymers [1262], or by entrapment within artificial liposomes or red blood cell ghosts; as well as the removal of receptor binding sites by designing appropriate enzyme muteins.
- Diminish immunogenicity, e.g., by chemical (PEGylation) or genetic modification to cover or remove epitopes on the protein surface recognized by the immune system.
- Achieve specific targeting to tissues/cells [1263]. This may be possible by modification with sugar residues which are recognized by hepatocyte receptors, or by genetic fusion to “targeting devices” such as recognition peptides or antibody fragments [1264]. On the other hand, another use of enzymes in tumor therapy which is presently under study is antibody-directed enzyme prodrug therapy [1265], [1266]: an antibody with an attached enzyme that is not normally present in extracellular fluids or on cell membranes, is administered to achieve specific binding to the surface of cancer cells. This is followed by administration of a prodrug, which is a substrate of the enzyme and on enzymatic cleavage releases the physiologically active cytotoxic entity in a site-specific manner. This is expected to reduce the exposure of normal tissue to the active drug while maximizing concentrations in the tumor tissue.
- Improve tissue penetration by decreasing molecule size (muteins) or by fusion to sequences mediating membrane permeation, such as penetrating (a peptide derived from the third helix of the DNA-binding domain of the Antennapedia homeodomain protein) [1267], TAT protein from human immunodeficiency virus-1 [1268], VP22 protein from Herpes simplex virus [1269], or Pep-1, a designed 21-residue peptide carrier [1270]. However, these approaches are at present (ca. 2003) in the experimental stage and appear to be still far from being useful in clinical practice.

6.2.3. Sources of Enzymes and Production Systems

For many decades, enzymes obtained from natural nonhuman sources such as bacteria, fungi, animals, and snake venoms have been used for therapeutic purposes. Due to their immunogenicity, these enzymes usually are suitable only for external applications (e.g., digestive or topical). In these cases, purity, source, and dosage of these enzyme preparations are not considered critical.

The purification of enzymes from human sources in a form and quantity which would allow their administration to patients, at least for correctly organized clinical trials, has been a major challenge [1271]. Nevertheless, a number of human enzymes obtained from blood, urine, placenta, or human cell culture have been introduced into clinical practice, such as blood fractions and coagulation factors, urokinase, imiglucerase, and others. Besides the limited supply of starting material, a major problem with enzymes obtained from human sources is the reliable elimination of viral (or maybe viroid) contamination. Therefore, since the 1980s, human enzymes are preferentially obtained by recombinant DNA technology.

Whereas initially basic replacement protein products, i.e., recombinant proteins identical with the native human enzymes were developed, there is now a trend towards engineered “second-generation” protein products whose amino acid sequence has been deliberately altered to improve properties relevant for

therapeutic efficacy or safety, or to improve the manufacturing process [1272].

An important aspect of the production of recombinant enzymes for therapy is the choice of the expression system. Most or perhaps all human enzymes in their native form are subject to posttranslational modifications such as glycosylation, phosphorylation, acetylation, myristoylation, or proteolytic processing. If the host cell used for recombinant production is not identical to the cell in which the protein is produced naturally, posttranslational modifications may deviate from the authentic pattern. This may influence the pharmakokinetic, pharmakodynamic, and immunological properties of the enzyme. Whether this is of importance for use in therapy must be studied experimentally in each case.

At present, various expression systems are used to obtain therapeutic proteins [1273]:

- Production from human cell lines clearly will provide the best fit to the natural modification pattern, but may give rise to problems of viral contamination.
- Production from mammalian cells [mostly chinese hamster ovary (CHO) and baby hamster kidney (BHK) cell lines] is often used, especially for glycosylated proteins in cases where glycosylation affects therapeutic efficacy [1274]. Glycosylation provided by mammalian host cells may be somewhat different in detail but is generally similar to the human glycosylation pattern.
- Production from yeast (*S. cerevisiae*) results in posttranslational modifications different from the human pattern; hence, it has been used until now primarily for fungal (nonhuman) proteins [1275].
- Production from *E. coli* results in nonglycosylated products. In principle, bacterial production systems have significant advantages concerning lower development cost, improved process economics, and absence of viral contamination [1276], [1277]. However, since eukaryotic proteins are often not folded correctly in bacteria, refolding of the target protein from “inclusion bodies” may be necessary. This can be done on an industrial scale [1278] but may require extensive process optimization to obtain good product yield.
- Production from transgenic plants [1279], [1280] or the milk of transgenic animals [1281]: these production systems are under study and may provide interesting options, especially for proteins needed in very large amounts or for very complex proteins such as coagulation factor VIII. No therapeutic enzyme product produced in transgenic animals or plants has yet been approved for marketing [1282].

For several years it has been discussed whether gene therapy will be a future third generation of protein therapy following the first (therapy with naturally occurring proteins) and second (engineered protein muteins) generations. In this case, the proteins are no longer produced in a fermenter but by the patients themselves after application of the protein-coding gene by means of a suitable gene-transfer system. This approach appears especially attractive for long-term therapies, such as enzyme replacement therapy of hereditary defects (e.g. ADA deficiency or lysosomal storage diseases). Many preclinical and clinical trials have already been performed in this field, but given the progress of gene therapy to date, this alternative apparently does not pose an immediate threat to the enzyme drug market.

6.2.4. Overview of Therapeutic Enzymes

Many of the requirements listed in Section Requirements for the Use of Enzymes in Therapy are not met by older enzyme products, i.e., enzymes from nonhuman sources approved prior to ca. 1980 which are still in use as generics in 2003 (some of them also in nutrition additives, oral hygiene products, or cosmetics; these applications are not treated here). On the other hand, in the field of modern innovative drugs, only a few recombinant enzymes have been developed until now as the main active principles of therapeutics, in contrast to the increasing number of recombinant cytokines, hormones, and antibodies which have already been approved or are in clinical development. The following overview includes older enzyme preparations still in use, even if they would not meet the requirements of modern drugs, as well as newer enzyme drugs developed and approved according to present-day standards.

Furthermore, in the development of therapeutic enzymes, as with other biopharmaceuticals, special attention must be given to regulatory requirements, as well as to aspects of quality assurance, assay methods, and standardization issues. These topics are not treated here in detail.

Vaccines such as diphtheria toxin (an ADP ribosyltransferase) or tetanus toxin (an endopeptidase) are not

discussed here. These are proteins whose toxic properties indeed rely on their enzymatic activity. However, they are used as immunogens for vaccination after chemical treatment with formaldehyde to convert them into enzymatically inactive toxoids. Thus, their medical use does not rely on their enzymatic activity but rather on their immunogenic properties.

The overview of enzymes (Table 39) used in therapy is ordered according to the E.C. numbering scheme of enzyme classes. The table includes enzymes presently marketed, as well as several enzymes which are at present in advanced clinical studies but have not yet been approved. For more information, the reader is referred to comprehensive books and reviews [1283-1287].

Table 39. Enzymes used in therapy

E.C. number	Enzyme	Source	Trade name (s), (examples)	Therapeutic use (examples)
1.7.3.3	urate oxidase, recombinant	recombinant *Saccharomyces cerevisiae*	Elitek, Fasturtec	hyperuricemia in TLS patients
1.7.3.3	urate oxidase	*Aspergillus flavus*	Uricozyme	gout; hyperuricemia in TLS patients
1.15.1.1	superoxide dismutase	bovine liver or erythrocytes	Peroxinorm, Ontosein, Oxinorm	inflammatory arthrosis, polyarthritis, asthma, oxygen toxicity in premature infants
1.15.1.1	superoxide dismutase, recombinant	recombinant *E. coli* or yeast	OxSODrol	prevention of bronchopulmonary dysplasia
(1.15.1.1)	PEG-superoxide dismutase	bovine liver	Dismutec	closed head injury
(2.3.2.13)*	coagulation factor XIII	human plasma	fibrogammin-P	bleeding, factor XIII deficiency; fibrin glues
3.1.1.3	lipase	*Rhizopus arrhizus*		digestion aid
3.1.1.47	platelet-activating factor acetylhydrolase, recombinant	recombinant *E. coli*	Pafase	severe sepsis, ARDS, pancreatitis
3.1.4.12	sphingomyelin phosphodiesterase 1 (acid sphingomyelinase), recombinant	recombinant CHO cells		type B Niemann – Pick disease
3.1.6.12	*N*-acetylgalactosamine-4-sulfatase (arylsulfatase B), recombinant	recombinant CHO cells	Aryplase	Maroteaux – Lang syndrome (mucopolysaccharidosis type VI)
3.1.6.13	iduronate-2-sulfatase, recombinant	recombinant CHO cells		Hunter syndrome (mucopolysaccharidosis type II)
3.1.21.1	desoxyribonuclease, recombinant	recombinant CHO cells	Pulmozyme	chronic obstructive pulmonary disease
3.1.21.1	desoxyribonuclease	*Streptococcus*	Varidase	gastric ulcers (together

	(streptodornase)	*haemolyticus*		with streptokinase)
3.1.21.-	T4 endonuclease V	bacteriophage T4-infected *E. coli*	Dimericine T	xeroderma pigmentosum
3.1.26.4	ribonuclease	*Rana pipiens*	Onconase	malignant mesothelioma, refractory breast cancer, renal cell cancer, antiviral
3.2.1.1	α-amylase	porcine pancreas		digestion aid
3.2.1.2	β-amylase	*Aspergillus oryzae*		digestion aid
3.2.1.3	acid α-1,4-glucosidase, recombinant	recombinant CHO cells		Pompe's disease
3.2.1.4	cellulase	*Trichoderma viride*		digestion aid
3.2.1.17	lysozyme	hen's egg white		infectious diseases (antibiotic)
3.2.1.22	α-galactosidase, recombinant	recombinant human cell line or CHO cells	Replagal, Fabrazyme	Fabry's disease
3.2.1.23	β-galactosidase (Lactase)	*Kluyveromyces fragilis*, *Aspergillus oryzae*, or *Aspergillus niger*	Lactaid, Lactrase, SureLac	lactose intolerance
3.2.1.35	hyaluronidase	bovine testes	Hylase, Vitrase	adsorption adjunct; heart attack
3.2.1.45	glucosylceramidase (β-glucocerebrosidase)	human placenta	Ceredase	Gaucher's disease
3.2.1.45	glucosylceramidase (β-glucocerebrosidase), recombinant	recombinant CHO cells	Cerezyme	Gaucher's disease
(3.2.1.45)	PEG- β-glucocerebrosidase	human placenta	Lysodase	Gaucher's disease
3.2.1.76	α-L-iduronidase, recombinant	recombinant CHO cells	Aldurazyme	mucopolysaccharidosis I
–	pankreatin (mixture of pancreatic enzymes, e.g. trypsin, chymotrypsin, lipase, α-Amylase)	porcine pancreas	Cotazym, Kreon, Mezym F, Nutrizym, Pancrex, Pangrol, Pankreatan, Pankreon, Panzytrat	digestion aid
3.4.21.1	chymotrypsin	bovine pancreas		wound healing, digestive aid
3.4.21.4	trypsin	bovine pancreas	Leukase, Granulex	debridement of wounds, digestion aid

			Grandfex	digestion aid
3.4.21.5	thrombin	human plasma	Thrombinar, Velyn	superficial bleeding; fibrin glues
3.4.21.7	plasmin	bovine or human plasma	Fibrolan, Lysofibrin, Actase, Elase	debridement of wounds
(3.4.21.21) *	coagulation factor VII	human plasma	Factor VII S-TIM 4	bleeding, factor VII deficiency
3.4.21.21	coagulation factor VIIa, recombinant	recombinant BHK cells	NovoSeven	hemophilia A and B
–	coagulation factor VIII	porcine plasma	Hyate C	hemophilia A
–	coagulation factor VIII	human plasma	Autoplex, Factorate, Profilate HS, Monoclate-P, Hemate P, Hemofil M, Monarc-M, Beriate P	hemophilia A
–	coagulation factor VIII (containing von Willebrand factor)	human plasma	Alphanate, Humate-P, Koate DVI, Immunate STM	von Willebrand disease
–	coagulation factor VIII, recombinant	recombinant CHO cells	Recombinate, Bioclate	hemophilia A
–	coagulation factor VIII, recombinant	recombinant BHK cells	KoGENate, Helixate	hemophilia A
–	coagulation factor VIII, recombinant (B-domain deleted)	recombinant CHO cells	ReFacto	hemophilia A
(3.4.21.22) *	coagulation factor IX	human plasma	Mononine, AlphaNine, Berinin HS, Immunine, Prothromplex-T	hemophilia B
(3.4.21.22) *	coagulation factor IX, recombinant	recombinant CHO cells	BeneFIX	hemophilia B
3.4.21.34	kallikrein	porcine pancreas	Padutin, Bioactin, Prokrein, Onokrein P	peripheral vascular and coronary disease, fertility disturbances
3.4.21.63	oryzin (*Aspergillus* alkaline protease)	*Aspergillus sp.*		inflammation
3.4.21.68	tissue plasminogen activator, recombinant	recombinant CHO cells	Activase, Actilyse	acute myocardial infarction; acute pulmonary embolism; ischemic stroke
(3.4.21.68)	tissue plasminogen activator mutein, recombinant	recombinant *E. coli*	Rapilysin, Retavase, Ecokinase	acute myocardial infarction
(3.4.21.68)	tissue plasminogen activator mutein,	recombinant CHO cells	Metalyse	acute myocardial infarction

	recombinant			
3.4.21.69	activated protein C, recombinant	recombinant CHO cells	Xigris, Zovant	severe sepsis
3.4.21.73	urokinase	human urine or human kidney cell culture	Abbokinase, Actosolv, Alphakinase, Rheothromb, Ukidan	acute myocardial infarction
3.4.21.73	urokinase, recombinant	recombinant *E. coli*	Saruplase	acute myocardial infarction
–	streptokinase	C β hemolytic *Streptococci*	Kabikinase, Streptase, Kinalysin, Varidase	acute myocardial infarction; gastric ulcers (together with ribonuclease)
–	APSAC (*p*-anisoylated derivative of plasminogen-streptokinase complex)	human plasma (plasminogen)/C β hemolytic *streptococci* (streptokinase)	Eminase	acute myocardial infarction
3.4.21.74	ancrod	*Agkistrodon rhodostoma* venom	Arwin, Venacil	improvement of blood rheology; lymphoblastic leukemia, lymphosarcoma
3.4.21.74	batroxobin	*Bothrops atrox* venom	Defibrase, Botropase, Reptilase, Vivostat	improvement of blood rheology; chirurgic tissue glues and hemostyptics
3.4.21.-	sfericase	*Bacillus sphaericus*		chronic bronchitis, acute pneumonia, chronic sinusitis
3.4.22.2	papain	*Carica papaya*	Panafil	digestion aid, debridement of wounds
3.4.22.6	chymopapain	papaya latex	Discase, Chymodiactin	chemonucleolysis of prolapsed invertebral disk
3.4.22.32	bromelain	*Ananas comosus* var.	Traumanase, Phlogenzym	digestion aid; inflammation
3.4.23.1	pepsin	porcine stomach	Enzymorn, Sentinel	support of gastic function; gastric ulcers
3.4.24.3	collagenase	*Clostridium histolyticum*	Iruxol, Biozyme C, Santyl	debridement of wounds; skin ulcers
3.4.24.28	protease	*Bacillus subtilis*	Travase	debridement of wounds
	extracts containing proteases, cellulases, RNases, α- and β-amylases	*Aspergillus oryzae*	NORTASE, COMBIASE	digestion aid
3.4.24.40	serrapeptase (serralysin)	*Serratia* E15	Danzen	inflammations
3.5.1.1	asparaginase	*E. coli* or *Erwinia carotovora*	Crasnitin, Elspar, Kidrolase, Leunase, Erwinase	acute lymphocytic leukemia

			Erwinase	
(3.5.1.1)	PEG-asparaginase	*E. coli*	Oncaspar	acute lymphocytic leukemia
(3.5.4.4)	PEG-adenosine deaminase	bovine intestine	ADAGEN	SCID
4.3.1.8	uroporphyrinogen-1-synthase (porphobilinogen deaminase), recombinant	recombinant *E. coli*		acute intermittent porphyria

*E.C. number of activated form.

6.2.4.1. Oxidoreductases

Urate oxidase (E.C. 1.7.3.3), an enzyme which catalyzes the oxidation of uric acid to the more soluble allantoin, is present in many organisms (including mammals) but not in higher primates. A nonrecombinant urate oxidase product isolated from Aspergillus flavus has been used in some countries for many years (France, since 1975; Italy, since 1984) for prophylaxis or treatment of hyperuricemia in patients with hematological malignomas which are prone to the risk of tumor lysis syndrome (TLS). This is an oncologic emergency most commonly occurring spontaneously or after chemotherapy in patients with large tumor burdens such as acute leukemia, aggressive lymphoma, or bulky solid tumors. TLS is characterized by hyperuricemia leading to the precipitation of uric acid and to intraluminal tubular obstruction, hyperphosphatemia, and hypocalcemia. The consequences of TLS may be life-threatening and include acute renal failure and multiorgan failure.

Several factors, including the complex manufacturing process, have limited the availability of A. flavus urate oxidase. Therefore, the cDNA coding for the enzyme was cloned and expressed in Saccharomyces cerevisiae. The purified recombinant enzyme (rasburicase) was approved for marketing in Europe in 2001, and in the USA in 2002. Detailed protein-chemical studies demonstrated a higher purity of rasburicase compared to the natural A. flavus enzyme; in particular, the fungal oxidase contained a cysteine adduct to Cys-103, likely attributable to the purification process, which was absent in the recombinant enzyme product [1288], [1289]. Because urate oxidase is a nonhuman protein, the drug is antigenic and potentially may induce hypersensitivity reactions. While nonrecombinant urate oxidase has a rate of acute hypersensitivity reactions of up to 5 %, the incidence appears to be much lower with the recombinant form.

Superoxide ion (O_2^-) and hydrogen peroxide are crucial starting points in the formation of other reactive oxygen species (ROS) which may cause oxidative damage to cells and tissue destruction. This is specially evident in endothelial cells, damage of which appears to be involved in pathological situations such as myocardial infarction, lung injury and inflammation, ischemia-reperfusion damage, hypoxia, and hyperoxia. Superoxide dismutase (orgotein, E.C. 1.15.1.1), a copper – zinc protein identified in bovine liver in the 1960s, removes the superoxide radical by catalyzing its dismutation to molecular oxygen and hydrogen peroxide, and thus acts as part of the cellular ROS detoxification pathway. Bovine superoxide dismutase has been used since the early 1980s as an anti-inflammatory drug, and proved to be therapeutically effective in the treatment of osteoarthritis, chronic polyarthritis, radiation cystitis, and interstitial cystitis. However, superoxide dismutase is susceptible to the known limits of native protein administration such as unsatisfactory pharmakokinetics and immunogenicity, although the incidence of allergic and anaphylactic reactions seems to be surprisingly low. Recombinant human superoxide dismutase (sudismase) preparations obtained from recombinant E. coli or yeast are presently in clinical development for treatment of amylotrophic lateral sclerosis, bronchopulmonary dysplasia in premature neonates, and respiratory distress syndrome.

To improve the blood residence time and simultaneously reduce immunogenicity, bovine superoxide dismutase has been coupled to polyethylene glycol. This conjugate (pegorgotein) indeed was found to have greatly increased clearance time that depends on the molecular weight of the PEG used for modification [1290]. The therapeutic efficacy of pegorgotein has been studied in several indications, and the product is presently in phase III clinical development for treatment of closed head injury.

6.2.4.2. Transferases
The only enzyme of this class which is used in therapy is coagulation factor XIII (fibrin-stabilizing factor) obtained from human plasma. Factor XIII is a glycoprotein which circulates in the blood as a proenzyme. During the final stages of blood coagulation, thrombin converts it to an active form called factor XIIIa (E.C. 2.3.2.13). This is a transglutaminase which, in the presence of Ca^{2+}, catalyzes the cross-linking of fibrin monomers through the formation of intermolecular ε-(γ-glutamyl)lysine bonds.

Factor XIII deficiency is a rare inherited disease characterized by poor wound healing and abnormal scar formation, with bleeding from the umbilical stump after birth as the most common clinical symptom. Deficiency of factor XIII can be corrected with infusions of fresh frozen plasma or human factor XIII concentrates. Although the cloning and recombinant expression of factor XIII has been reported in several publications since the 1980s, no recombinant factor XIIIa preparation has yet been approved for therapeutic use.

6.2.4.3. Esterases
Platelet-activating factor (PAF, 1-*O*-alkyl-2-acetyl-*sn*-glycero-3-phosphatidyl choline) is a potent phospholipid mediator with pleiotropic effects, produced by a wide variety of cells. Multiple substances stimulate its production, including lipopolysaccharide and tissue factors released after endothelial disruption. It has been found that serum levels of PAF are elevated in human sepsis, and indeed, when PAF is injected into animals, it produces physiological dysfunction similar to that seen in sepsis. These effects are mediated by various cytokines and other molecules [1291]. Two phase II trials in the sepsis indication have been reported which used recombinant human PAF acetylhydrolase (E.C. 3.1.1.47, pafase), a phospholipase A_2 specifically hydrolyzing phospholipids in which the *sn*-2 fatty acid is an acetyl group. In one of these trials, which was a multicenter prospective, randomized, double-blind, placebo-controlled trial with 240 patients with severe sepsis or major trauma, there was a statistically significant reduction in mortality associated with the use of the recombinant PAF acetylhydrolase. Validation of these results was sought in a large phase III multicenter trial [1292] which, however, failed to demonstrate clinical benefit. Pafase is also in phase II clinical trials for patients at risk of developing acute respiratory distress syndrome and acute pancreatitis. Another important medical area in which PAF acetylhydrolase may play a protective role is in oxidative processes that are thought to be a component of early vascular disease [1293].

Lipase (E.C. 3.1.1.3) from porcine pancreas (see Section Proteases) or from *Rhizopus arrhizus* (rizolipase) is used as a digestion aid.

6.2.4.4. Nucleases
In lung diseases characterized by persistent airway infection such as cystic fibrosis (CF, mucoviscidosis) and chronic bronchitis, respiratory distress and progressive lung destruction are strongly associated with bacterial persistence and the accumulation of viscous purulent secretions in the airways. Infected-airway secretions contain mucus glycoproteins and abundant amounts of DNA which, as an extremely viscous polymer polyanion, is a major cause of the increased viscosity.

Recombinant human desoxyribonuclease I (dornase alfa, E.C. 3.1.21.1) reduces the viscosity of the sputum by hydrolyzing DNA. Dornase alfa produced from recombinant CHO cells is virtually identical to DNase I, an extracelluar glycoprotein produced in human pancreas and salivary glands.

The clinical efficacy and safety of dornase alfa was well established in trials on clinically stable CF patients. Dornase alfa, administered through an inhalation device, was found to reduce CF-related morbidity, including the risk of respiratory tract infectious exacerbations requiring treatment with parenteral antibiotics, and to improve pulmonary function and quality of life. It was approved for treatment of CF in 1993.

Streptodornase, a desoxyribonuclease obtained from *Streptococcus* bacteria, is used together with streptokinase for the debridement of wounds, as well as to treat ulcers in order to liquefy blood clots and dead tissue so that these can be easily removed.

DNA damage plays an important role in the development of skin cancer, which is the most common cancer in white people in the USA and other countries. Patients with the rare genetic disease xeroderma

pigmentosum have an autosomal recessive genetic defect in the pathway that repairs sun-induced damage of DNA. As a result, the rates of malignant disorders of the skin are 1000 times higher than in the general population. The DNA repair enzyme, T4 Endonuclease V (E.C. 3.1.21.-) is at present in phase III trials as a liposomal lotion (T4N5 liposome lotion). It is a viral DNA glycosylase which specifically excises pyrimidine dimers in DNA. The enzyme binds to UV-irridiated DNA, progressively scans the DNA until a pyrimidine dimer is encountered, and then cleaves the glycosyl bond of the 5′-pyrimidine of the dimer as well as the 3′-phospho diester bond, resulting in breakage of the DNA strand. In a randomized study, it was shown that topical application of the endonuclease lowered the rate of development of new skin lesions (actinic keratoses and basal-cell carcinomas) during a year of treatment [1294].

Following studies which demonstrated inhibition of tumor growth after injection of bovine pancreatic ribonuclease (RNase) into tumor-bearing mice, the potential of ribonucleases (E.C. 3.1.26.4) to inhibit tumor growth has been investigated since 1955. However, at present only an amphibian RNase isolated from the leopard frog Rana pipiens (ranpirnase) is in phase II/III clinical trials in the indications malignant mesothelioma, refractory breast cancer, renal cell cancer, and prostate cancer [1295]. Ranpirnase, which is not inhibited by endogenous mammalian RNase inhibitors, is also being studied preclinically for use in antiviral therapy.

6.2.4.5. Glycosidases

α-Amylase (E.C. 3.2.1.1) from porcine pancreas is part of pancreatin preparations (cf. Section Proteases) and is used to support the digestion of dietary carbohydrate. This is also the case for extracts from Aspergillus oryzae containing cellulases (E.C. 3.2.1.4), proteases, and α- and β-amylase (E.C. 3.2.1.1/3.2.1.2). In fact, an enzyme preparation from Aspergillus oryzae cultures (Takadiastase) developed by JOKICHI TAKAMINE and used as a digestive aid in the 1890s, probably represented the first use of a microbial enzyme preparation for therapeutic purposes.

Lysozyme (E.C. 3.2.1.17) was discovered in 1922 by ALEXANDER FLEMING as a bacteriolytic substance present in many tissues and secretions. Its antibacterial effect is based on hydrolysis of the bacterial cell walls by cleaving the β-(1,4) glycosidic bonds between *N*-acetylmuraminic acid and *N*-acetylglucosamine residues. Lysozyme also has chitinolytic activity, which contributes to the cell-wall lysis of numerous gram-positive bacteria. The explanation for a reported antiviral action of lysozyme is elusive. Lysozyme, obtained from hen egg white, is still used for treatment of infectious and inflammatory diseases, for the treatment of viral diseases, for eye therapy, and within the oral cavity and throat in anti-angina preparations. Chitinases (E.C. 3.2.1.14) are also under study for treatment of fungal infections.

Hyaluronidases (E.C. 3.2.1.35) randomly hydrolyze the β-*N*-acetyl hexosaminic bonds in hyaluronate, chondroitin 4- and 6-sulfate and their desulfated derivatives to yield even-numbered oligosaccharides with glucuronic acid residues at their nonreducing end. By hydrolysis of the connective tissue glycosaminoglycans, they facilitate the absorption of injected fluids and resorption of exudates. Therefore, hyaluronidase obtained from bovine testes (a mixture of several enzymes) is used to promote absorption and dispersion in the subcutaneous or intramuscular injection of relatively large volumes of fluids. Ophthalmic injectable formulations have also been developed for use in the treatment of vitrous haemorrhage in diseases such as diabetic retinopathy.

Diseases which are caused by genetically encoded single-enzyme defects generally are candidates for rational therapy by enzyme replacement. Until now, this has been realized in severe combined immunodeficiency (see Section Amidases) and in two rare glycolipid storage diseases, Gaucher's disease and Fabry's disease. For some other indications, clinical development of enzyme replacement therapies is in progress [1296], especially for lysosomal storage diseases such as recombinant human α-L-iduronidase (laridonase, E.C. 3.2.1.76) for mucopolysaccharidosis type I (Hurler's syndrome) [1297], recombinant acid sphingomyelinase (E.C. 3.1.4.12) for type B Niemann – Pick disease, recombinant lysosomal acid maltase (α-glucosidase, E.C. 3.2.1.3) for Pompe's disease, recombinant human iduronate-2-sulfatase (E.C. 3.1.6.13) for mucopolysaccharidosis type II (Hunter syndrome), and recombinant human *N*-acetylgalactosamine 4-sulfatase (E.C. 3.1.6.12) for mucopolysaccharidosis type VI (Maroteaux – Lang syndrome).

Development of these recombinant enzyme-based drugs has been supported greatly by the U.S. and European “orphan drug” regulations allowing drugs directed to life-threatening diseases which affect less

than 200 000 persons in the USA, or occur with a frequency of less than 5 in 10 000 persons in Europe, to obtain tax incentives and grant funding, as well as seven (USA) or ten (Europe) years market exclusivity. This gives the originator companies a good chance to pay off the development costs of these drugs in spite of the small market size.

Gaucher's disease is the most prevalent lysosomal storage disorder, characterized by the accumulation of glycolipid (glucosylceramide) primarily in peripheral tissues. It is inherited in an autosomal recessive manner, and the incidence is 1/600 – 1/2500 in ashkenazic jews and 1/50 000 in caucasians (Norrbottnian region of Sweden). This disease is caused by deficiency in the enzyme β-glucocerebrosidase (glucosylceramidase, E.C. 3.2.1.45) which catalyzes the removal of glucose from glucosylceramide to form ceramide. In 1991, alglucerase (glucocerebrosidase derived from human placenta) was launched as the first enzyme replacement therapy for Type 1 Gaucher's disease. Meanwhile, the human enzyme has been replaced by recombinant glucocerebrosidase (imiglucerase; U.S. approval 1994, European approval 1997) produced in CHO cells. This was critical for overcoming the limitations of availability: 20 000 placentas were required to provide a one-year supply for a single patient. Imiglucerase differs from the native human enzyme by one amino acid, Arg-495 being substituted by His, and has a modified oligosaccharide component. A PEGylated form of glucocerebrosidase is also in clinical studies.

Fabry's disease is an X-chromosomal, recessively inherited lysosomal glycosphingolipid storage disease which leads to the abnormal deposition of neutral glycosphingolipids (mainly globotriaosylceramide) in many tissues and cell types, such as kidney, heart, skin, and the cells lining the blood vessels. In the USA, it occurs with an incidence of 1 in 40 000, mostly in caucasians, but is also found in african americans, hispanics, and asians. The disease is caused by a genetically encoded defect of the lysosomal α-galactosidase A (E.C. 3.2.1.22). Two recombinant human α-galactosidases (agalsidase alfa produced in a human cell line, and agalsidase beta obtained from recombinant CHO cells) have been shown to mobilize and clear the sphingolipid deposits and have been approved for long-term enzyme-replacement therapy of Fabry's disease [1298], [1299]. The phase III clinical trial preceding the approval of agalsidase beta enrolled 58 patients at eight centers and was the largest trial ever conducted for Fabry's disease.

Lactose is present in concentrations of about 4.7 vol % in milk and the whey left after the coagulation stage of cheesemaking. Its presence makes milk unsuitable for the majority of the world's adult population: the young of all mammals are able to digest milk but in most cases, this ability declines after weaning. Of the Thai, Chinese, and african american populations, 97, 90, and 73 %, respectively, are reported to be lactose intolerant, whereas 84 and 96 % of the U.S. white and swedish populations, respectively, are tolerant. For people who have an impaired ability to digest lactose, lactase (β-galatosidase, tilactase, E.C. 3.2.1.23) is available as a nonprescription enzyme as a dietary supplement to alleviate the symptoms of lactose intolerance.

6.2.4.6. Proteases

6.2.4.6.1. Pancreatic and Gastric Proteases

Gastric and pancreatic insufficiency is frequently treated by substitution with crude enzyme preparations applied orally. These unspecific hydrolases are able to degrade plasma proteins and cell membrane constituents in vitro. Nevertheless, orally administered proteases appear to be nontoxic, which indicates that there is no significant absorption of the active enzymes from the gastrointestinal tract.

Pepsin (E.C. 3.4.23.1) from porcine stomach is used for support of gastric function to increase the digestive power of gastric juice. In combination with bismuth complexes it is also used for treatment of gastric ulcers.

Treatment of disturbances of the exokrine pancreas function leading to maldigestion, such as chronic pancreatitis, partial or total pancreatic ectomy, or gastrogenic maldigestion, requires not only replacement of proteolytic, but also of amylolytic and lipolytic activities. Therefore, pancreatin, a hog (or bovine) pancreatic powder containing trypsin (E.C. 3.4.21.4), chymotrypsin (E.C. 3.4.21.1), α-amylase (E.C. 3.2.1.1), and lipase (E.C. 3.1.1.3) activities, is frequently used. The digestion of neutral fats is usually promoted by the addition of bile acids. To protect the pancreatin enzymes against the acid pH of the stomach, they are administered preferentially as enterocoated formulations which release the enzymes at the alkaline pH of the duodenum.

6.2.4.6.2. Plasma Proteases

Kallikrein (kallidinogenase, E.C. 3.4.21.8) was discovered in 1926 as a substance in human urine exhibiting hypotensive properties. Serum kallikrein plays an important physiological role in maintaining low blood pressure due to its ability to cleave L-kininogen, a serum α-globulin protein, and release vasoactive kinin peptides (bradykinin or lysyl-bradykinin). Kinins exert several biological activities such as vasodilation, blood pressure reduction, smooth muscle relaxation and contraction, pain induction, and inflammation. Therapeutically, kallikrein is used in the treatment of peripheral vascular and coronary arterial diseases, as well as in delayed fracture and wound healing.

The removal of purulent exudate, necrotic tissue, or hard and resistant eschar is a prerequisite for the effective cleaning and sterilization of burns, ulcers, traumatic and surgical wounds, and for the initiation of the healing process. A variety of proteolytic enzymes such as bovine plasmin (fibrolysin, E.C. 3.4.21.7), trypsin (E.C. 3.4.21.4) from bovine pancreas, clostridial collagenase (E.C. 3.4.24.3), and a protease from *Bacillus subtilis*, which can remove fibrin layers from wounds, are used as debriding agents for the cleaning of necrotic wounds, ulcers, abcesses, empyemas, sinuses, and fistulas. The enzymes are usually combined with antibiotics and are topically applied as ointments or as a wet or dry dressing. Dry enzyme powder may also be insufflated into wounds or cavities, or inserted into fistulas by means of small gelatin capsules. Because of the presence of trypsin and plasmin inhibitors in serum, the proteases are expected not to attack living tissue. However, due to their heterologous source, the risk of allergenic reactions must be considered.

6.2.4.6.3. Coagulation Factors

α-Thrombin (Fibrinogenase, E.C. 3.4.21.5) is a trypsin-like serine proteinase which causes blood clotting through proteolytic cleavage of fibrinogen and protease-activated receptors. It also amplifies its own generation from the inactive precursor prothrombin by activating the essential clotting factors V and VIII. Thrombin from bovine plasma is used to treat superficial bleeding resulting from trauma or surgical intervention, and ultrasound-guided thrombin injection has recently emerged as a treatment of iatrogenic femoral artery pseudoaneurysms. Another source of exposure to exogenous thrombin is the use of commercial “fibrin glue” preparations which contain a human or bovine thrombin component together with a gelatin-based matrix containing fibrinogen, fibronectin, and coagulation factor XIII in surgical procedures, especially during cardiac surgery, and to support wound healing. Bovine thrombin preparations represent potential antigenic stimuli, and an immunologic response may occur in up to 30 % of patients [1300]. Thrombin cannot be used intravascularly because of the danger of inducing intravascular coagulation.

Thrombin can both promote and prevent blood clotting (thrombin paradox). Its anticoagulant effect is caused by its ability to complex with an endothelial cell membrane protein, thrombomodulin, and then process the endogenous circulating serine protease zymogen Protein C into activated Protein C [APC, drotregocin alfa (activated), E.C. 3.4.21.69]. APC cleaves and inactivates factor Va and factor VIII and thereby inhibits the generation of thrombin, which prevents clotting, and also inhibits plasminogen activator inhibitor-1 (PAI-1), leading to a profibrinolytic effect. Furthermore, APC was reported to exert an anti-inflammatory effect by inhibiting production of tumor necrosis factor by monocytes and by blocking leukocyte adhesion to selectins. A recombinant version of naturally occurring human activated Protein C produced in CHO cells was approved in 2001 for treatment of severe sepsis [1291]. This condition results from an out-of-control host response to invading microbes, which gives rise to a dramatic increase of cytokine levels as well as escalation of the clotting cascade throughout the body, leading to inflammation, septic shock, and multiple organ failure.

Hemophilia is the most common congenital disorder of coagulation and affects approximately 1 in 10 000 males around the world [1301]. Hemophilia A, the oldest recognized hereditary bleeding disorder, is due to a deficiency of the antihemophiliac factor (AHF, factor VIII) in the circulating blood, whilst hemophilia B (also known as Christmas disease), a clinically identical, but less common disorder, is caused by factor IX deficiency. Both factor VIII, which is not an enzyme but a protein cofactor required for the activation of factor X by factor IXa, and factor IX (a zymogen of the serine protease factor IXa) are essential glycoproteins in the clotting cascade. Patients with hemophilia A and B require life-long substitution therapy, which was introduced about a century ago using fresh blood. A major therapeutic advance was the preparation of bovine factor VIII in 1954, followed by the introduction of cryoprecipitates and then lyophilized concentrates from human plasma. However, the pooling of plasma donations introduced the risk of transmission of viral infections (e.g., hepatitis, HIV). Although pretesting of the plasma used in the manufacture of coagulation factor products, as well as physical and chemical process steps of viral inactivation, have been introduced, recombinant products are believed to offer the greatest margin of safety

for hemophiliac patients. Recombinant expression of factor VIII and IX was a great technical challenge, since both are glycoproteins which undergo several posttranslational modifications, and factor VIII comprises no less than 2332 amino acid residues. Nevertheless, various human coagulation factor preparations obtained from recombinant mammalian cells have been developed and approved in the 1990s: recombinant factor VIII (octocog alfa) from CHO or BHK (1992/93); a second-generation, B-domainless factor VIII (moroctocog alfa), which retains activity in clotting assays but can be formulated without the addition of human albumin, which further reduces the theoretical risk of transmission of human viruses (1999); and recombinant factor IX (nonacog alfa) obtained from CHO cells (1997).

However, highly purified factor VIII and factor IX preparations obtained from human plasma are still available. Furthermore, concentrates of human plasma-derived factor VIII containing von Willebrand factor are used to treat von Willebrand disease, which is the most common inherited disorder of hemostasis, with an incidence in the population of about 1 %.

One complication of replacement therapy, however, is the development of inhibitory antibodies directed against factor VIII or factor IX in a significant percentage of severe hemophilia patients. For this type of patients, porcine factor VIII isolated from porcine plasma as well as prothrombin complex concentrates are available. Furthermore, recombinant human factor VIIa (niastase, E.C. 3.4.21.21) has been obtained from BHK cells for treatment of these patients. Factor VIIa possesses very little inherent proteolytic activity, but when bound to tissue factor on the subendothelial cell surface, is able to convert factor X into factor Xa (and factor IX into factor IXa) directly in the absence of factors VIII and IX, and therefore bypasses factors VIII and IX in the coagulation cascade. Recombinant factor VIIa has been approved (EU, 1996; USA, 1999) for treatment of hemophilia A patients with factor VIII inhibitory antibodies and hemophilia B patients with factor IX inhibitory antibodies. Factor VII concentrates obtained from human plasma are used to treat patients with factor VII deficiency, which is a rare disorder occurring in 1 in 500 000.

6.2.4.6.4. Plasminogen Activators

Acute myocardial infarction and ischemic stroke are two main causes of death and disability in the Western societies. Fibrinolytic therapy of these diseases is based on the premise that their proximal cause is thrombosis, i.e., occlusive thrombi prevent blood flow to vital organs, resulting in oxygen deprivation, cell necrosis, and loss of organ function [1302]. Early and sustained recanalization is expected to prevent cell death, reduce infarct size, preserve organ function, and reduce early and late mortality. Thrombolytic agents dissolve blood clots by activation of the inactive circulating zymogen plasminogen to the active protease plasmin, which then will degrade fibrin, the main constituent of blood clots.

The first (and, in Europe, still most widely used) used plasminogen activator, streptokinase (SK), produced by several strains of hemolytic *Streptococci*, is not an enzyme, but activates plasminogen by forming an equimolar complex with plasminogen, which then undergoes a conformational change resulting in the exposure of the active site of the plasminogen molecule. In a second step, this active site catalyzes the activation of residual plasminogen to plasmin, which then again converts the plasminogen – SK complex to a plasmin – SK complex by proteolytic cleavage. Since streptokinase generates free circulating plasmin, its use is associated with a systemic lytic state. Furthermore, streptokinase acts as a strong antigen in humans, which is even more critical because preexisting antibodies from previous streptococcal infections are common. Nevertheless, prospective multicenter trials with streptokinase have demonstrated that lytic therapy significantly improves the one-year survival rate if acute myocardial infarction is treated within the first 3 h after the onset of symptoms.

To control the enzymatic activity of the plasmin(ogen) – SK complex, anistreplase (anisoylated plasminogen – streptokinase activator complex, APSAC) has been developed. In APSAC, the catalytic center of the plasminogen moiety is specifically and reversibly modified with an anisoyl residue. This should delay the formation of plasmin but has no influence on the lysine-binding sites involved in the (albeit weak) binding of the complex to fibrin, so that a certain degree of localization of protease activity at the clot would be expected. However, the deacylation of the complex occurs both in circulation and at the fibrin surface, and the additional fibrin specificity of thrombolysis by anistreplase appears to be marginal [1302].

Urokinase (urokinase-type plasminogen activator, u-PA, E.C. 3.4.21.73) is an enzyme of human origin, secreted by the kidney and eliminated in the urine. Urokinase is formed in vivo as a single-chain form (scu-PA, prourokinase, Mr 54 000) which is then processed at the Lys-158–Ile-159 bond into a two-chain

derivative (tcu-PA or HMW urokinase). Further proteolytic cleavage result in a form with Mr = 33 000 (LMW urokinase). Whereas the two-chain forms HMW and LMW show similar activity in vitro towards chromogenic substrates, scu-PA displays very low activity towards low molecular weight substrates. In plasma, in the absence of fibrin, scu-PA does not activate plasminogen; in the presence of the fibrin clot, scu-PA (but not tcu-PA) induces fibrin-specific clot lysis due to enhanced activity versus fibrin-bound plasminogen. u-PA is a serine protease with a high substrate specificity for plasminogen, and since it is a human enzyme, does not cause any immunological complications. For several years, LMW-urokinase obtained from human urine or from human neonatal kidney cell culture has been used for thrombolytic therapy. However, in the past few years, some manufacturers have run into problems with raw material supply (human urine) and uncertainties about possible viral contamination (human cell culture), leading to temporary unavailability of urokinase in some countries. Recombinant versions of single-chain urokinase (saruplase), obtained from recombinant *E. coli*, have been in clinical studies for a number of years, but no recombinant urokinase products have yet been approved for marketing.

A major advance in thrombolytic therapy was made when recombinant tissue-type plasminogen activator (t-PA, alteplase, E.C. 3.4.21.68) became available. t-PA is a Mr = 70 000 serine protease which also exists in single- and two-chain forms, which are both proteolytically active, and comprises three N-glycosylation sites. t-PA has little plasminogen activator activity in the absence of fibrin, but its activity is strikingly enhanced in the presence of fibrin. Thus, when introduced into the systemic circulation, t-PA will initiate local fibrinolysis at the site of blood clots, but only limited systemic proteolysis. Recombinant t-PA obtained from CHO cells was approved for the treatment of acute myocardial infarction (AMI) in 1987, and for acute massive pulmonary embolism in 1990. In large clinical trials, it was conclusively demonstrated that fibrin-selective recombinant t-PA combined with intravenous heparin is superior to nonselective streptokinase for early mortality reduction.

Like many plasma proteases, t-PA consists of several discrete structural modules, i.e., a “finger” domain, a “growth-factor-related” domain, two “kringle” domains, as well as a protease domain, structurally similar to trypsin, which contains the enzymatically active center. In the 1990s, many approaches were reported to design t-PA mutants with improved properties, e.g., higher fibrin specificity, greater zymogenicity (activity ratio of two-chain vs single-chain form), slower clearance from the circulation, and resistance to plasma proteinase inhibitors, but only two of these mutants have been clinically developed and approved for marketing: reteplase [1303] is a single-chain, unglycosylated deletion variant consisting only of the second kringle and the protease domains of human t-PA. The plasminogenolytic activities of alteplase and reteplase in the absence of a stimulator do not differ, but in the presence of stimulating fibrinogen fragments, reteplase activity is about four times lower as compared to t-PA, and its binding to fibrin is five times lower. The reteplase molecule no longer contains the structural elements responsible for the rapid clearance of t-PA from the bloodstream (finger domain and carbohydrate residues), and indeed shows significantly prolonged half-life in the circulation (14 – 18 min in humans, as compared to 3 – 4 min for wild-type t-PA). This allows application as a double bolus, instead of a bolus followed by a 90 min infusion, as is necessary with t-PA. Reteplase is produced in recombinant E. coli by a process involving protein refolding on an industrial scale. Several large clinical studies have demonstrated that a clinical benefit of reteplase over alteplace could not be shown, and the two agents can be considered equivalent [1304], [1305]. Reteplase was approved for treatment of AMI in 1996.

In the second t-PA mutant which is commercially available, tenecteplase (TNK-t-PA), the glycosylation site in kringle-1 was deleted by substitution of Asn-117 by Gln, whereas a new glycosylation site was introduced, at a different locus, by replacement of Thr-103 by Asn. Furthermore, the amino acids 296 – 299 were replaced by Ala residues, which confer resistance against inhibition by PAI-1. Tenecteplase, which is produced in recombinant CHO cells, has a plasma half-life of 17 min, similar to that of reteplase, and has a similar ability to wild-type t-PA to bind to fibrin. Large clinical trials with AMI patients have demonstrated that tenecteplase, given as a single bolus, was equivalent to alteplase in terms of 30-d mortality. Tenecteplase was approved for marketing in 2000.

Other plasminogen activators, e.g., the plasminogen activator from vampire bat (DSPAa-1), other t-PA muteins, a chimeric t-PA Kringle 2/u-PA plasminogen activator (amediplase, [1306]) and staphylokinase from *Staphylococcus aureus* (obtained from recombinant *E. coli*) have been intensively studied in preclinical and even clinical studies, but until now, none of them has been approved for routine use [1307].

6.2.4.6.5. Proteases from Snake Venoms

Snake venoms contain two types of thrombin-like proteinases, including thrombocytin and crotalocytin, which preferentially act on substrates other than fibrinogen and are used for research purposes only, as well as enzymes (E.C. 3.4.21.74) such as Ancrod, obtained from the Malayan pit viper *Calloselasma* (*Agkistrodon*) *rhodostoma*, and Batroxobin, obtained from *Bothrops atrox moojeni* or *Bothrops jararaca* [1308]. These latter enzymes act very specifically on the *A* α chain of fibrinogen, which is transformed into fibrinopeptide A and fibrin I, but do not promote cross-linking of the resulting fibrin gel. Fibrin I stimulates release of plasminogen activators from the endothelium and is degraded by plasmin into noncoagulating degradation products. This leads to a decrease of fibrinogen plasma levels and to reduced viscosity and coagulability of the blood. Therefore, Ancrod and Batroxobin are used for the prevention of thromboses, improvement of blood rheology, and support of drug-induced fibrinolysis, as well as for anticoagulation in patients with heparin-induced thrombocytopenia (HIT syndrome). Loss of activity after prolonged therapy, resulting from antibody formation, is observed with both snake venom proteases. Batroxobin is also used in chirurgic tissue glues and hemostyptics in order to obtain fibrin I preoperatively from the patient's own blood, which is then used post-operatively for wound closure.

6.2.4.6.6. Plant and Microbial Proteases

In the same way as pancreatin, lytic enzymes from plants or microorganisms are used as digestive aids, e.g., papain (papayotin, E.C. 3.4.22.2), a cysteine endopeptidase obtained from the latex of *Carica papaya*, and bromelain (E.C. 3.4.22.4), obtained from the stem and fruit of pineapple (*Ananas comosus* var.).

In Europa and Japan (but not in the USA), enzyme preparations containing bromelain together with papain, trypsin, or chymotrypsin, as well as serrapeptase (E.C. 3.4.24.40), a protease from the nonpathogenic enterobacterium *Serratia* E15, have been widely used in rheumatic disorders [1309] and generally for anti-inflammatory treatment. A protease from *Aspergillus* (promelase, E.C. 3.4.21.63) has also been studied as an anti-inflammatory agent. However, enzyme use in this market segment is declining versus other drugs such as salicylates, nonsteroid anti-inflammatory agents, and steroids.

Intravenous administration of brinase, a proteolytic enzyme preparation from *Aspergillus oryzae*, was found to be beneficial in the treatment of chronic arterial obstruction. Sfericase, the main intracellular protease of *Bacillus sphaericus* was found to reduce the viscosity of human pathological bronchial secretions while not modifying their elasticity, and has been studied for use in chronic bronchitis, acute pneumonia, and chronic sinusitis. Sutilain, a proteolytic enzyme isolated from *Bacillus subtilis*, has been used for the treatment of burns.

Chemonucleolysis is now an established treatment modality in the management of herniated or prolapsed lumbar intervertebral disk disease as an alternative to surgery. The enzyme chymopapain (E.C. 3.4.22.6), a proteolytic cysteine protease from papaya latex, is injected directly into the center of the affected disk (nucleus pulposus), where it digests the protein part of the insoluble mucopolysaccharide – protein complex. The fragmented proteoglycan diffuses from the disk, causing a reduction in hydrostatic pressure on the nerve root and thus an easing of pain and disability. This procedure has been in widespread use since its introduction in 1963, and its efficacy has been established in a placebo-controlled, double-blind study. However, the injection technique is complicated, and a relatively high complication rate is observed.

6.2.4.7. Amidases

Certain tumor cells may require unusual or specific nutrients derived from the bloodstream, or may require a nutrient in higher concentration than normal cells. Thus, continued growth of these tumors might be prevented by restricting their nutrient supply by administration of an appropriate enzyme. In this area, microbially derived L-asparaginase (E.C. 3.5.1.1), isolated from various microorganisms such as *E. coli* (colaspase) and *Erwinia carotovora* (crisantaspase), has been approved for the treatment of some forms of leukemia, and glutaminase is also under study for treatment of neoplastic diseases. Asparaginase has proved to be particularly promising for treatment of acute lymphocytic leukemia. Its action depends on the fact that tumor cells are deficient in aspartate ammonia ligase activity, which restricts their ability to synthesize the normally nonessential amino acid L-asparagine. Asparaginase deprives the circulating blood of asparagine. This does not affect the functioning of normal cells, which are able to synthesize enough for their own requirements, but induces a state of starvation in the susceptible tumor cells. The enzyme is administered intravenously and shows a half-life of about one day in dogs. This half-life can be increased 20-fold by use of PEG-modified asparaginase (Peg-asparaginase) which is likely to expand the scope of application and the

therapeutic efficacy of this enzyme. However, the main problems of this therapy are the rapid development of resistance by the tumor cells, disturbances in the homeostatic system, as well as immunological complications.

Severe combined immunodeficiency (SCID) is a primary immune deficiency which usually results in the onset of one or more serious, and even life-threatening, infections within the first few months of life. There are several forms of SCID; one of them is linked to a chronic deficiency of the enzyme adenosine deaminase (ADA, E.C. 3.5.4.4). For these patients, the only available alternative to bone marrow transplant is enzyme replacement with PEGylated bovine ADA (pegademase bovine) in infants from birth or children of any age at the time of diagnosis. Injection of unmodified ADA is not effective because of its short circulating life (less than 30 min) and the potential for immunogenic reactions to a bovine-sourced enzyme. The attachment of PEG to ADA increases its circulating life and masks epitopes on the surface of the ADA molecule to avoid immunogenic reactions.

6.2.4.8. Lyases
The only lyase presently under study for therapeutic application is recombinant human porphobilinogen deaminase (PBGD, E.C. 4.3.1.8), an enzyme involved in the heme-synthesizing pathway. In 2002, this enzyme obtained orphan designation for the treatment of acute intermittent porphyria, a clinical manifestation of an autosomal dominant genetic disorder caused by mutation in the PBDG gene and resulting in a 50 % reduction of enzymatic activity. This may lead to an increase in blood and urinary concentrations of heme precursors such as porphobilinogen, and a deficit in heme synthesis. The effects of recombinant human PBDG, until now, have only been demonstrated in experimental models. Administration of PBDG is expected to decrease plasma concentrations of heme precursors.

6.3. Enzymes in Diagnosis
This section describes the principles of enzyme use in diagnosis; the reader who wishes to get a deeper insight into this field is referred to special monographs, e.g., [1310]. In clinical chemistry, enzymes are used as auxiliaries for determination of *substrate concentration* or *enzyme activity* and as labels for enzyme *immunoassays*. Reactions are monitored photometrically by absorption change, or they are measured by fluorescence. In addition, light-emitting reactions catalyzed by enzymes are of increasing practical importance, especially in immunoassays.

6.3.1. Determination of Substrate Concentration
Substrate concentrations are determined enzymatically in two ways: (1) end-point methods and (2) measurement of reaction rate [1311], [1312].

End-Point Methods. In the simplest case, the enzyme-catalyzed substrate conversion is practically complete, and either the decrease in substrate concentration or the increase in product concentration can be measured, e.g., by light absorption at a given wavelength. For substrate concentrations well below the value of the Michaelis constant (K_m), the reaction rate obeys the following equation [1311]:

$$v = c_S \cdot V / K_m \tag{6.1}$$

Hence, the time required for completion (e.g., 99 % conversion) of the enzymatic reaction strongly depends on the maximum reaction rate (V) and the Michaelis constant (K_m) of the enzyme used.

Determination of Glucose with Glucose Oxidase. In the first reaction, glucose is oxidized with glucose oxidase (E.C. 1.1.3.4) [9001-37-0] at 10 U/mL and hydrogen peroxide is formed:

$$\text{Glucose} + O_2 + H_2O \rightarrow \text{Gluconate} + H_2O_2$$

In the second reaction, hydrogen peroxide is used to oxidize a chromogen to form a dye with the aid of

horse-radish peroxidase (E.C. 1.11.1.7) [9003-99-0] at 0.8 U/mL:

$$H_2O_2 + \text{chromogen} \rightarrow \text{dye} + H_2O$$

2,2′-Azino-bis(3-ethyl-2,3-dihydrobenzthiazolsulfonate) [30931-67-0] (ABTS) is used as a chromogen [1315].

Determination of Urea. Urea is hydrolyzed with urease (E.C. 3.5.1.5) [9002-13-5] at 0.7 U/mL and the ammonia formed is assayed by use of glutamate dehydrogenase (E.C. 1.4.1.3) [9029-12-3] at 6.2 U/mL:

$$\text{Urea} + H_2O \rightarrow 2\,NH_3 + CO_2$$

$$2\text{-Ketoglutarate} + 2\,NH_4^+ + 2\,NADH \rightarrow 2\,\text{L-Glutamate} + 2\,NAD^+ + 2\,H_2O$$

The reaction is followed photometrically (decrease in NADH concentration).

Kinetic Methods. The drawback of end-point methods is that a considerable amount of time may be required before the reaction comes to completion; this limits the sample frequency. In addition, many time- and reagent-consuming sample blanks must be determined, to eliminate interference from light-absorbing serum constituents. These problems can be overcome by assaying the substrate concentration via the rate of the enzyme-catalyzed reaction [1312]. As shown in Equation (6.1), the kinetics is of pseudo-first order, and the rate is proportional to substrate concentration at concentrations well below K_m.

Such assays are most conveniently run on automatic analyzers using the so-called kinetic fixed-time approach, in which the absorbance is read at two constant points of time, and the slope of the absorbance change is calculated. Only one calibration is required.

These methods work only for substrate concentrations below the K_m value, which sets an upper limit on the dynamic range. This upper concentration limit is increased in the presence of competitive inhibitors that increase the apparent K_m of the enzyme. Kinetic substrate assays are described for many analyses.

Urea concentration is determined according to the reaction sequence mentioned previously.

Determination of Glucose.

$$\text{D-Glucose} + ATP \rightarrow \text{D-Glucose 6-phosphate} + ADP$$

$$\text{D-Glucose 6-phosphate} + NADP^+ \rightarrow \text{D-Glucono-}\delta\text{-lactone 6-phosphate} + NADPH + H^+$$

In the first reaction, glucose is phosphorylated by hexokinase (E.C. 2.7.1.1) [9001-51-8]; then glucose 6-phosphate is dehydrogenated by the action of glucose 6-phosphate dehydrogenase (E.C. 1.1.1.49) [9001-40-5]. The NADPH production is monitored photometrically.

Determination of Triglycerides. Triglycerides (fats) are hydrolyzed by lipase (E.C. 3.1.1.3) [9001-62-1] and carboxylesterase (E.C. 3.1.1.1) [9016-18-6]. Glycerol is then phosphorylated by glycerol kinase (E.C. 2.7.1.30) [9030-66-4]. The ADP formed in this reaction is rephosphorylated to ATP with

phosphoenolpyruvate and pyruvate kinase (E.C. 2.7.1.40) [9001-59-6]. Finally, the pyruvate is hydrogenated by L-lactate dehydrogenase (E.C. 1.1.1.27) [9001-60-9] and the decrease in NADH concentration is followed:

$$\text{Triglyceride} + 3\,H_2O \rightarrow \text{Glycerol} + 3\,\text{Fatty acid}$$

$$\text{Glycerol} + \text{ATP} \rightarrow \text{Glycerol 3-phosphate} + \text{ADP}$$

$$\text{ADP} + \text{Phosphoenolpyruvate} \rightarrow \text{ATP} + \text{Pyruvate}$$

$$\text{Pyruvate} + \text{NADH} + H^+ \rightarrow \text{L-Lactate} + NAD^+$$

6.3.2. Determination of Enzyme Activity [1314]

Ideally, the substrate concentration must be so high for an enzyme assay that the value of K_m can be neglected. Then $v = V$, and the reaction rate is equal to the maximum rate defined by the Michaelis – Menten equation. The reaction obeys pseudo zero-order kinetics. In many cases, the enzyme to be assayed cannot be determined directly but must be coupled to a second enzyme-catalyzed reaction that serves as the indicator reaction. In this case, the conditions chosen must make the enzyme reaction which is to be determined the rate-limiting step.

Activity of Alkaline Phosphatase (E.C. 3.1.3.1) [9001-78-9].

$$\text{4-Nitrophenylphosphate} + H_2O \rightarrow \text{Phosphate} + \text{4-Nitrophenolate}$$

This is the easiest case. A chromogenic substrate is cleaved by the enzyme whose activity is to be determined. At the optimum pH of this reaction, 9.8, the product is dissociated, and its rate of formation can be followed by the increase in light absorption at 405 nm.

Activity of Creatine Kinase (E.C. 2.7.3.2) [9001-15-4].

$$\text{Creatine phosphate} + \text{ADP} \rightarrow \text{Creatine} + \text{ATP}$$

This is an example of a reaction coupled with an auxiliary enzyme. The ATP formed in the reaction catalyzed by creatine kinase is used to phosphorylate glucose to glucose 6-phosphate, which is then dehydrogenated, as described previously. The rate of NADPH formation is a measure of the rate of the reaction catalyzed by creatine kinase.

6.3.3. Immunoassays

Enzyme immunoassays use enzymes as labels to determine the amount of immunocomplex (antigen – antibody) formed; for a review, see [1314], [1315]. Depending on the chemical reaction catalyzed by the respective enzyme, its activity can be measured photometrically, fluorometrically, or luminometrically. Enzymes can be bound to antibodies using such bifunctional coupling reagents as glutaraldehyde [111-30-8] [1315], 3-maleinimidobenzoyl-*N*-hydroxysuccinimide [15209-14-0] [1316], or bis(maleido)-methyl ether [1317], depending on the nature of the reactive groups on the enzyme. Horseradish peroxidase is bound

directly to the antibody by reaction of carbonyl groups in the carbohydrate part of the enzyme with amino groups of the antibody via Schiff base formation and subsequent reduction with sodium borohydride [1318].

Alkaline phosphatase (E.C. 3.1.3.1) [9001-78-9], used in enzyme immunoassays, should be homogeneous on electrophoresis, and its specific activity should be above 2500 U/mg when 4-nitrophenylphosphate is used as substrate at 37 °C. For conjugate synthesis, glutaraldehyde is recommended [1315]. Alkaline phosphatase activity can also be assayed fluorometrically by using 4-methylumbelliferylphosphate [3368-04-5] as substrate. For use in immunohistochemistry, 5-bromo-4-chloro-3-indolylphosphate [6578-06-9] is especially suitable. A photometric amplification system for the immunoassay of alkaline phosphatase has been published [1319].

β-Galactosidase (E.C. 3.2.1.23) [9031-11-2] should have a specific activity of more than 250 U/mg when 4-nitrophenyl- β-D-galactoside is used as substrate at 37 °C. The enzyme should not be present in the aggregated form. One enzyme molecule contains 12 mercapto groups; these are linked to antibodies by maleimide bifunctional reagents. The activity of β-galactosidase can be measured photometrically with 4-methylumbelliferyl- β-galactoside [39940-54-0]; in histochemistry, the substrate of choice is 5-bromo-4-chloro-3-indolyl- β-galactoside [7240-90-6].

Horseradish peroxidase (E.C. 1.11.1.7) [9003-99-0] contains 2 – 3 reactive amino groups per molecule and 12 – 14.5 wt % of carbohydrates, which can both be used for coupling to antibodies. The enzyme activity can be assayed photometrically in different ways. The chromogen 2,2′-azino-bis[3-ethylbenzothiazoline-sulfonate] (ABTS) is especially suited as a substrate, and the peroxidase should have a specific activity of ca. 1000 U/L at 25 °C. 3,3′,5,5′-Tetramethylbenzidine [54827-17-7] is recommended as a substrate for use in immunohistochemistry.

Luminescence Reactions. *Chemiluminescence* occurs when part of the free energy produced in a chemical reaction is emitted as light, because a reaction product is generated in its singlet excited state. *Bioluminescence* occurs in living organisms, where these reactions are catalyzed by enzymes. The corresponding enzymes are called *luciferases*, and the substrates being converted to light-emitting products are called *luciferins*; for a review see [1320]; see also Section Luminometry. Because light intensities can be measured with very high sensitivity, luminescence reactions are of increasing interest as indicator systems for immunoassays, giving detection limits comparable to those achieved with radioactive labels, e.g., see [1321], [1322].

Firefly Bioluminescence. For a long time, bioluminescence was considered to be something exotic. However, several groups have shown that these reactions follow the same rules as other enzyme-catalyzed reactions [1320]. An example is the reaction catalyzed by luciferase (E.C. 1.13.12.7) [61970-00-1] of the American firefly *Photinus pyralis*, as mentioned in Section Luminometry. When Michaelis – Menten kinetics are applied to this reaction, the light intensity is:

$$I = \frac{\mathrm{d}N}{\mathrm{d}t} = \frac{1}{\Phi} \cdot v = \frac{1}{\Phi} \frac{V \cdot c_{\mathrm{ATP}}}{K_{\mathrm{m}}}$$

where

v = reaction rate

V = maximum reaction rate

I = light intensity

Φ = quantum yield of the reaction

N = number of photons emitted

Hence, the light intensity is proportional to the reaction rate, and substrate assays using this reaction follow the rules of kinetic substrate determination. These considerations also apply to other enzyme-catalyzed chemiluminescence reactions. Bioluminescence catalyzed by firefly luciferase has been used in

immunoassays in two different ways: (1) the luciferase can be used directly as a marker, e.g., [1323]; or (2) pyruvate kinase can be used as a label, and the ATP formed in the rephosphorylation of ADP by phosphoenolpyruvate (see Phytase.) is then assayed by the reaction catalyzed by firefly luciferase [1324].

Chemiluminescence Catalyzed by Horseradish Peroxidase. The peroxidase from horseradish (E.C. 1.11.1.7) [9003-99-0] catalyzes the oxidation of luminol (5′-amino-2,3-dihydro-1,4-phthalazinedione [521-31-3] and luminol derivatives by hydrogen peroxide, with concomitant light emission:

$$\text{Luminol} + H_2O_2 + 2\,OH^- \rightarrow \text{Aminophthalate} + N_2 + 2\,H_2O$$

Under normal conditions, this reaction is rather slow and, depending on the substrate concentrations, a lag phase of varying length is observed. Therefore, this reaction is not suitable for routine assay of peroxidase activity. However, several groups of compounds have been found recently that activate this reaction by eliminating the lag phase, which results in an increase in light intensity by a factor of 100 – 1000. The light signal is stable at a given set of concentrations. Firefly luciferin [2591-17-5] [1325], 6-hydroxybenzothiazoles [1326], and phenol derivatives such as 4-iodophenol [540-38-5] and 4-phenylphenol [92-69-3] are used as activators [1327]. At pH 6, the phenol derivatives also act as activators of a photometric reaction catalyzed by peroxidase [1328]. Hence, these compounds probably activate the peroxidase.

Immunoassays are described for the determination of several antibodies and metabolites of diagnostic use, based on this chemiluminescence reaction stimulated by peroxidase activator.

6.4. Enzymes for Food Analysis [1329], [1330], [1331-1333]

Use of the catalytic activity of enzymes for analytical determination dates from the middle of the 19th century. Nevertheless, the use of enzymes for food analysis in its present form was established only about 20 or 25 years ago. The factors responsible for this achievement were the large-scale preparation of pure enzymes, the availability of suitable and inexpensive photometers, and the elaboration of both enzymatic procedures and methods for sample preparation.

In *food research*, the amounts of various food constituents are determined, and changes occurring during technological processing and subsequent storage are monitored. Because of the great specificity manifest by enzymes, enzymatic methods are now used for this purpose.

In *industry*, raw materials are checked, production controls are carried out, and both the end products and the competing products are analyzed. Enzymatic methods were accepted quickly by industry because of the obvious advantage of using easy, rapid, and highly specific methods of analysis.

Official inspections involve regular analysis of samples and checking that laws, regulations, and guidelines are being observed. Checks are also made to determine if the food industry is fulfilling its obligation to make complete declarations according to the law. Enzymatic methods were used very early for these purposes not only because of the accurate results obtained from enzymatic analysis but also because of the great flexibility of the methods and their applicability to different types of sample materials. Because of tariffs on food, *customs officials* routinely analyze food by enzymatic methods. *Military inspectors* also use enzymes in the general control of food as well as in stability and storage investigations.

The use of enzymes for analytical determinations is recommended today by many national and international commissions; in fact, standardized enzymatic methods are widely available.

6.4.1. Carbohydrates

Carbohydrates are determined routinely in food analysis. The sugar components are an important part of the overall analytical picture: they constitute a large part of the caloric value of food, and the authenticity of food samples can be deduced from the relative amounts of different sugars. The presence of lactose indicates the use of milk. Starch serves as a raising agent in meat products and as a thickening agent.

Glucose [50-99-7] is by far the most abundant sugar and is usually determined by using the hexokinase method [1329, vol. VI, pp. 163 – 172]:

$$\text{D-Glucose} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.1}]{\text{Hexokinase}} \text{ADP} + \text{Glucose 6-phosphate} \tag{6.2}$$

$$\text{Glucose 6-phosphate} + \text{NADP}^+ \xrightarrow[\text{E.C. 1.1.1.49}]{\text{Glucose 6-phosphate dehydrogenase}} \text{D-Gluconate 6-phosphate} + \text{NADPH} + \text{H}^+ \tag{6.3}$$

Other enzymatic methods are not recommended for a variety of reasons: e.g., the enzymes are nonspecific (glucose dehydrogenase), or reducing substances in the sample interfere with the reaction (glucose oxidase – peroxidase method). In addition, other methods do not allow the simultaneous determination of fructose in the same cuvette. Glucose is usually determined together with other carbohydrates such as fructose, sucrose, maltose, and starch.

Fructose [57-48-7]. The enzyme hexokinase also acts on fructose, which is determined after glucose (Reactions 6.4, 6.5, and 6.3) [1329, vol. VI, pp. 321 – 327]:

$$\text{D-Fructose} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.1}]{\text{Hexokinase}} \text{ADP} + \text{Fructose 6-phosphate} \tag{6.4}$$

$$\text{Fructose 6-phosphate} \xrightarrow[\text{E.C. 5.3.1.9}]{\text{Glucose phosphate isomerase}} \text{Glucose 6-phosphate} \tag{6.5}$$

Glucose and fructose are determined in wine and fruit juices, e.g., to detect the prohibited addition of sugar. Fructose and glucose are also determined together with sucrose.

Galactose [59-23-4] is easily determined by using the following reaction [1329, vol. VI, pp. 104 – 112]:

$$\text{D-Galactose} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.48}]{\text{Galactose dehydrogenase}} \text{D-Galactonic acid} + \text{NADH} + \text{H}^+ \tag{6.6}$$

This determination is of interest only for certain milk products (yogurt, soft cheese). The method can also be applied to acid hydrolysates of thickening agents like agar, guar, carrageenan, gum arabic, locust-bean gum, and tragacanth gum.

Mannose [3458-28-4]. The determination of free mannose is also of little interest. However, like galactose, it is an important component of thickening agents and can be estimated after subjecting these substances to acid hydrolysis according to Reactions (6.7), (6.8), (6.5), and (6.3) [1329, vol. VI, pp. 262 – 267]:

$$\text{D-Mannose} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.1}]{\text{Hexokinase}} \text{ADP} + \text{Mannose 6-phosphate} \tag{6.7}$$

$$\text{Mannose 6-phosphate} \xrightarrow[\text{E.C. 5.3.1.8}]{\text{Phosphomannose isomerase}} \text{Fructose 6-phosphate} \tag{6.8}$$

Sucrose [57-50-1]. The disaccharide sucrose does not occur in animals but occurs in many plants in varying amounts. On hydrolysis, sucrose yields an equimolar mixture of glucose and fructose (Reaction 6.9) and is estimated via glucose (Reactions 6.2 and 6.3) [1330, vol. 3, pp. 1176 – 1179]:

$$\text{Sucrose} + H_2O \xrightarrow[\text{E.C. 3.2.1.26}]{\beta\text{-Fructosidase}} \text{D-Glucose} + \text{D-Fructose} \tag{6.9}$$

In addition to measurement of glucose, fructose produced in Reaction (6.9) can also be measured according to Reactions (6.4), (6.5), and (6.3) to ensure accurate results or increase the sensitivity of the determination.

Sucrose, glucose, and fructose are determined in fruit and vegetable products (e.g., juices, jams, tomato puree, or potatoes), bread and biscuits, honey, ice cream, confectionery, desserts, diet food, sugar refinery products, or beverages. A large excess of glucose interferes with the accuracy of sucrose and fructose analysis. Therefore, excess glucose should be removed as completely as possible with a mixture of glucose oxidase and catalase.

Maltose [69-79-4]. Enzymatic hydrolysis of maltose yields two glucose molecules which are determined as described above [1329, vol. VI, pp. 119 – 126]:

$$\text{Maltose} + H_2O \xrightarrow[\text{E.C. 3.2.1.20}]{\alpha\text{-Glucosidase}} 2\ \text{D-Glucose} \tag{6.10}$$

Sucrose, which also contains an α-glucosidic bond, and maltotriose are hydrolyzed as well.

Maltose is determined in malt products, beer, baby food, and often together with partial hydrolysates of starch (glucose syrup, starch sugar; dextrins in beer).

Lactose [63-42-3]. The carbohydrate of milk, lactose, is an important constituent of infant diets. It is usually determined after hydrolysis (Reaction 6.11) via galactose and indicates the use of milk or milk products in the production of food [1329, vol. VI, pp. 104 – 112]:

$$\text{Lactose} + H_2O \xrightarrow[\text{E.C. 3.2.1.23}]{\beta\text{-Galactosidase}} \text{D-Galactose} + \text{D-Glucose} \tag{6.11}$$

Analysis via galactose gives accurate results because very few samples contain large amounts of galactose. Lactose can also be estimated via glucose (Reactions 6.11, 6.2, and 6.3), but excess free glucose impairs the precision of the results and must be removed with a mixture of glucose oxidase and catalase (see lactulose). The amount of lactose is determined in milk (decreases with mastitis), milk products, and other food as a measure of the milk content, e.g., baby food, bread, biscuits, ice cream, chocolate, desserts, and sausages.

Lactulose [4618-18-2] is formed when lactose is heated. It is measured by means of Reactions (6.11), (6.4), (6.5), and (6.3) [1329, vol. VI, p. 111] and is used to check the heating of milk.

Raffinose [512-69-6] occurs in relatively high concentrations in sugar beet. It accumulates in molasses during the production of sugar. Raffinose, which also occurs in soybean, is determined by using Reactions (6.12) and (6.6) [1329, vol. VI, pp. 90 – 96]:

$$\text{Raffinose} + H_2O \xrightarrow[\text{E.C. 3.2.1.22}]{\alpha\text{-Galactosidase}} \text{D-Galactose} + \text{Sucrose} \tag{6.12}$$

Raffinose is measured at different stages in the production of beet sugar. The addition of soybean protein to food is detected indirectly by measuring raffinose, because soybean flour can contain up to 10 % of this carbohydrate (addition of soybean is often subject to declaration).

Starch [9005-25-8]. The analysis of starch has played an important part in food analysis for over 100 years. It is determined via glucose after hydrolysis. The use of acid hydrolysis for this purpose is not advisable because other glucose polysaccharides and oligomers are also hydrolyzed by acids; in addition, undesired reactions can take place (e.g., conversion of glucose to fructose). Amylase is less suited for enzymatic hydrolysis because it does not break down starch to glucose completely. Amyloglucosidase is the enzyme of choice; however, in addition to starch, maltose and oligoglucosides are also hydrolyzed to glucose by this enzyme. Analysis is carried out by using Reactions (6.13), (6.2), and (6.3):

$$\text{Starch} + (n-1)\,H_2O \xrightarrow[\text{E.C. 3.2.1.3}]{\text{Amyloglucosidase}} n\,\text{D-Glucose} \tag{6.13}$$

Starch in the sample must be dissolved, which is achieved by heating in an autoclave or by treating with concentrated hydrochloric acid and dimethyl sulfoxide. All steps must be performed carefully according to instructions. Oligoglucosides (maltodextrins, glucose syrup, starch sugar) are separated by extraction with ethanol – water mixtures which do not dissolve starch.

Starch is determined in products such as flour, bread, biscuits, and meat products, which contain starch as a raising agent. Oligoglucosides are estimated in fruit juices, beverages, confectionery, jam, and ice cream. Dextrins (low molecular mass starch) are determined in beer; they indicate the presence of fermentable carbohydrates.

6.4.2. Organic Acids

Organic acids and their salts frequently appear in metabolism and are of varying importance. Their presence in food has a considerable effect on taste and may indicate fermentation.

Acetic Acid [64-19-7]. Acetic acid, a component of the “volatile acids,” is determined specifically by using Reactions (6.14) – (6.16) [1329, vol. VI, pp. 639 – 645]. The use of acetate kinase is not suitable because this enzyme also acts on propionate.

$$\text{Acetate} + \text{ATP} + \text{CoA} \xrightarrow[\text{E.C. 6.2.1.1}]{\text{Acetyl-CoA synthetase}} \text{Acetyl-CoA} + \text{AMP} + \text{Pyrophosphate} \tag{6.14}$$

$$\text{Acetyl-CoA} + \text{Oxaloacetate} + H_2O \xrightarrow[\text{E.C. 4.1.3.7}]{\text{Citrate synthetase}} \text{Citrate} + \text{CoA} \tag{6.15}$$

$$\text{L-Malate} + \text{NAD}^+ \underset{\text{E.C. 1.1.1.37}}{\overset{\text{L-MDH}}{\rightleftharpoons}} \text{Oxaloacetate} + \text{NADH} + \text{H}^+ \tag{6.16}$$

Reaction (6.16) is a preceding indicator reaction; the equilibrium must be considered when the calculations are being made.

Acetate is measured in wine, fruit and vegetable products, cheese, dressings, and vinegar as a check for fermentation.

Ascorbic Acid [50-81-7]. As a vitamin, ascorbic acid is of great biological importance in humans; it is also often used as a food additive in industry, e.g., for fruit and vegetable products. A chemical – enzymatic process (Reaction 6.17) is applied to the quantitative determination of ascorbic acid [1329, vol. VI, pp. 376 – 385]:

$$\begin{aligned}&\text{L-Ascorbic acid } (\text{XH}_2) + \text{MTT} \xrightarrow{\text{PMS}} \\ &\qquad \text{Dehydroascorbic acid (X)} + \text{Formazan}^- + \text{H}^+\end{aligned} \tag{6.17}$$

where MTT = 3-(4,5-dimethylthiazolyl-2)-2,5-diphenyl-tetrazolium bromide; PMS = 5-methylphenazinium methyl sulfate.

Ascorbate is oxidatively removed in a sample blank reading in order to increase specificity:

$$\begin{aligned}&\text{L-Ascorbic acid} + 1/2\,\text{O}_2 \\ &\qquad \xrightarrow[\text{E.C. 1.10.3.3}]{\text{Ascorbate oxidase}} \text{Dehydroascorbic acid} + \text{H}_2\text{O}\end{aligned} \tag{6.18}$$

Ascorbic acid is determined in fruit and vegetable products, meat products, milk, beer, wine, and flour. The vitamin content is often declared.

Dehydroascorbic acid [490-83-5] can be determined by using Reactions (6.17) and (6.18), after it has been chemically oxidized to ascorbic acid [1329, vol. VI, pp. 376 – 385]:

$$\text{Dehydroascorbic acid} + \text{Dithiothreitol(reduced)} \rightarrow \text{L-Ascorbic acid} + \text{Dithiothreitol (oxidized)} \tag{6.19}$$

Aspartic Acid [56-84-8]. Aspartic acid is estimated primarily in apple juice [1329, vol. VIII, pp. 350 – 357]:

$$\begin{aligned}&\text{L-Aspartate} + \alpha\text{-Oxoglutarate} \\ &\qquad \xrightarrow[\text{E.C. 2.6.1.1}]{\text{GOT}} \text{Oxaloacetate} + \text{L-Glutamate}\end{aligned} \tag{6.20}$$

$$\begin{aligned}&\text{Oxaloacetate} + \text{NADH} + \text{H}^+ \\ &\qquad \xrightarrow[\text{E.C. 1.1.1.37}]{\text{L-MDH}} \text{L-Malate} + \text{NAD}^+\end{aligned} \tag{6.21}$$

Citric Acid [77-92-9]. Citric acid is a key substance in metabolism. It occurs abundantly in plants and also in milk. It is determined by using Reactions (6.22) and (6.21) [1329, vol. VII, pp. 2 – 12]:

$$\text{Citrate} \xrightarrow[\text{E.C. 4.1.3.6}]{\text{Citrate (}pro\text{-}35\text{)-lyase}} \text{Oxaloacetate} + \text{Acetate} \tag{6.22}$$

The decarboxylation of oxaloacetic acid to pyruvic acid can occur either chemically or enzymatically (oxaloacetate decarboxylase present as an impurity in citrate lyase). However, loss of oxaloacetate does not lead to inaccurate results because the pyruvate formed is also estimated by using the L-lactate dehydrogenase reaction (6.38).

Fruit and vegetable products, bread, cheese, meat products, beverages, wine, tea, and confectionery are all analyzed for citric acid.

Formic Acid [64-18-6]. An end product of the metabolism of bacteria and fungi, formic acid is employed as a preservative in a variety of foods. However, the use of this additive must comply with the law. Formic acid is determined according to [1329, vol. VI, pp. 668 – 672]:

$$\text{Formate} + \text{NAD}^+ + \text{H}_2\text{O} \xrightarrow[\text{E.C. 1.2.1.2}]{\text{Formate dehydrogenase}} \text{Hydrogen carbonate} + \text{NADH} + \text{H}^+ \tag{6.23}$$

Gluconic Acid [526-95-4]. Oxidation of glucose yields gluconic acid. δ-Gluconolactone is employed as a ripening agent in the production of sausages. The enzyme gluconate kinase is used to determine gluconic acid [1329, vol. VI, pp. 220 – 227]:

$$\text{D-Gluconate} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.12}]{\text{Gluconate kinase}} \text{D-Gluconate 6-phosphate} + \text{ADP} \tag{6.24}$$

$$\text{D-Gluconate 6-phosphate} + \text{NADP}^+ \xrightarrow[\text{E.C. 1.1.1.44}]{\text{6-PGDH}} \text{D-Ribulose 5-phosphate} + \text{NADPH} + \text{H}^+ + \text{CO}_2 \tag{6.25}$$

where 6-PGDH = 6-phosphogluconic acid dehydrogenase.

δ-Gluconolactone is also determined by using Reactions (6.24) and (6.25) after conversion to gluconic acid by alkaline hydrolysis (pH 10 – 11):

$$\text{D-Glucono-}\delta\text{-lactone} + \text{H}_2\text{O} \rightarrow \text{D-Gluconate} \tag{6.26}$$

Glutamic Acid [6899-05-4]. The enzyme glutamate dehydrogenase is applied to the quantitative determination of glutamic acid [1329, vol. VIII, pp. 369 – 376]:

$$\text{L-Glutamate} + NAD^+ + H_2O \xrightarrow[\text{E.C. 1.4.1.3}]{\text{Glutamate dehydrogenase}} \alpha\text{-Oxoglutarate} + NADH + NH_4^+ \tag{6.27}$$

The equilibrium of this reaction strongly favors the reductive synthesis of glutamate. Originally, hydrazine was used to trap α-oxoglutarate to achieve quantitative conversion. This is more conveniently carried out today by using the enzyme diaphorase and iodonitrotetrazolium chloride (INT):

$$NADH + H^+ + INT \xrightarrow[\text{E.C. 1.8.1.4}]{\text{Diaphorase}} NAD^+ + \text{Formazan} \tag{6.28}$$

This color reaction is very sensitive; all reducing substances in the sample must be removed during its preparation to avoid interference.

3-Hydroxybutyric Acid [300-85-6]. The presence of hydroxybutyric acid in an egg indicates that fertilization has taken place and that the egg has been incubated for more than 6 d. Hydroxybutyric acid is estimated by using the following reaction:

$$\text{D-3-Hydroxybutyrate} + NAD^+ \xrightarrow[\text{E.C. 1.1.1.30}]{\text{3-HBDH}} \text{Acetoacetate} + NADH + H^+ \tag{6.29}$$

where 3-HBDH = 3-hydroxybutyrate dehydrogenase.

To ensure that this reaction is quantitative and to increase sensitivity, Reaction (6.29) is coupled to the color reaction (6.28).

Isocitric Acid [6061-97-8]. The determination of isocitric acid, an intermediate in the citric acid cycle, is of considerable importance for the analysis of fruit juice. The ratio of citrate to isocitrate is very constant; hence, analysis of isocitric acid can be used to show if citric acid has been added to the juice (adulteration). Isocitrate dehydrogenase is used to measure this acid quantitatively [1329 vol, VII, pp. 13 – 19]:

$$\text{D-Isocitrate} + NADP^+ \xrightarrow[\text{E.C. 1.1.1.42}]{\text{Isocitrate dehydrogenase}} \alpha\text{-Oxoglutarate} + NADPH + CO_2 + H^+ \tag{6.30}$$

The lactone or esters of isocitrate are first subjected to alkaline hydrolysis (Reaction 6.31) and then determined by using Reaction (6.30):

$$\text{D-Isocitric acid lactone (esters)} + H_2O \xrightarrow{\text{pH 9–10}} \text{Isocitric acid (+alcohol)} \tag{6.31}$$

Isocitric acid is too expensive to be used to adulterate citrus fruit juices.

Lactic Acid [598-82-3]. One of the most important examples of the use of enzymes in food analysis is the

enzymatic determination of the stereoisomeric forms of lactic acid. Only L-(+)-lactate is found in animals; it accumulates as the end product of glycolysis, e.g., in muscle. D-(–)-Lactate is formed by certain types of Lactobacilli. There are now no reservations about the intake of D-(–)-lactate in the diet. Indeed, the ADI value (acceptable daily intake) proposed by the WHO was relatively quickly withdrawn. The analysis is carried out by using the following reactions [1329, vol. VI, pp. 582 – 592]:

$$\text{L-Lactate} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.27}]{\text{L-LDH}} \text{Pyruvate} + \text{NADH} + \text{H}^+ \tag{6.32}$$

$$\text{D-Lactate} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.28}]{\text{D-LDH}} \text{Pyruvate} + \text{NADH} + \text{H}^+ \tag{6.33}$$

The equilibria of the lactate dehydrogenase reactions favor formation of lactic acid. Originally, hydrazine was used to trap pyruvate in order to achieve quantitative conversion. Today, use of an enzymatic reaction, such as the following, is more convenient:

$$\text{Pyruvate} + \text{L-Glutamate} \xrightarrow[\text{E.C. 2.6.1.2}]{\text{GPT}} \text{L-Alanine} + \alpha\text{-Oxoglutarate} \tag{6.34}$$

L-Lactate is determined in fruit and vegetable products, meat additives, and baking agents. Both D- and L-lactate are measured in such milk products as yogurt and cheese as a check for microbial activity, and in wine and beer where the lactate content correlates with sensory results.

Malic Acid [6915-15-7]. L-Malic acid, an intermediate in the tricarboxylic acid (TCA) cycle, is found in fruit (e.g., grapes) and vegetables. In analytical determination, it is oxidized in the presence of malate dehydrogenase and NAD^+ to yield oxaloacetic acid: the generated oxaloacetic acid is no longer trapped with hydrazine, but with an enzymatic reaction [1329, vol. VII, pp. 39 – 47]:

$$\text{L-Malate} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.37}]{\text{L-MDH}} \text{Oxaloacetate} + \text{NADH} + \text{H}^+ \tag{6.35}$$

$$\text{Oxaloacetate} + \text{L-Glutamate} \xrightarrow[\text{E.C. 2.6.1.1}]{\text{GOT}} \text{L-Aspartate} + \alpha\text{-Oxoglutarate} \tag{6.36}$$

L-Malic acid is analyzed in fruit products and wine. In comparison with chemical procedures which determine total malic acid content only, enzymatic methods can detect even small amounts of cheap D,L-malic acid added to the product. An enzymatic reaction for the determination of D-malate has been described in the literature, but the required enzyme is not available on the market.

Oxalic Acid [144-62-7]. More than half of all kidney stones are composed of calcium oxalate. Oxalic acid is also important because it influences the absorption of calcium in the intestine. It is determined by using Reactions (6.37) and (6.23) [1330, vol. 3, pp. 1551 – 1555]:

$$\text{Oxalate} \xrightarrow[\text{E.C. 4.1.1.2}]{\text{Oxalate decarboxylase}} \text{Formate} + \text{CO}_2 \tag{6.37}$$

Oxalic acid is determined in beer (calcium oxalate is responsible for the Gushing effect), fruit, vegetables, and cocoa products. The oxalate oxidase method [1329, vol. VI, pp. 649 – 656] is not recommended for food analysis because too many factors interfere with the reaction.

Pyruvic Acid [127-17-3] is a key intermediate in metabolism. It is determined by using the following reaction [1329, vol. VI, pp. 570 – 577]:

$$\text{Pyruvate} + \text{NADH} + \text{H}^+ \xrightarrow[\text{E.C. 1.1.1.27}]{\text{L-LDH}} \text{L-Lactate} + \text{NAD}^+ \tag{6.38}$$

Wine is analyzed for pyruvic acid, which is an SO_2-binding component. In evaluating the quality of raw milk (and pasteurized milk), the determination of pyruvate is used as an alternative to microbial counting.

Succinic Acid [110-15-6]. Succinic acid is also an intermediate in the citric acid cycle; it is determined by using Reactions (6.39), (6.40), and (6.38) [1329, vol. VII, pp. 25 – 33]:

$$\text{Succinate} + \text{ITP} + \text{CoA} \xrightarrow[\text{E.C. 6.2.1.4}]{\text{Succinyl-CoA synthetase}} \text{IDP} + \text{Succinyl-CoA} + \text{p}_i \tag{6.39}$$

$$\text{IDP} + \text{PEP} \xrightarrow[\text{E.C. 2.7.1.40}]{\text{Pyruvate kinase}} \text{ITP} + \text{Pyruvate} \tag{6.40}$$

Fruits and fruit products are analyzed for succinic acid, which indicates the degree of ripeness. The presence of this acid in whole eggs is a sign of microbial activity following contamination. Other items sampled are cheese, soybean products, and wine.

6.4.3. Alcohols

Enzymes are also used for the analytical determination of alcohol, such as ethanol, glycerol, the sugar alcohols sorbitol and xylitol, or cholesterol, in food.

Ethanol [64-17-5]. Anaerobic metabolism by many microorganisms, notably yeast, results in the formation of ethanol. Reaction (6.41), employed in the determination of ethanol, was one of the first routinely used enzymatic analyses. The acetaldehyde generated was originally trapped with semicarbazide. However, its enzymatic oxidation (Reaction 6.50) is quicker and more efficient [1329, vol. VI, pp. 598 – 606]:

$$\text{Ethanol} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.1}]{\text{ADH}} \text{Acetaldehyde} + \text{NADH} + \text{H}^+ \tag{6.41}$$

The ethanol concentration serves as an index of the quality of alcoholic drinks, such as wine, champagne, spirits, and beer. The ethanol content of products made from fruit indicates either the use of spoiled raw materials or the presence of yeast (e.g., in the case of kefir). The maximum ethanol concentration in drinks that are classified as “low in alcohol” or “nonalcoholic” has been established by law.

Glycerol [56-81-5] is widely distributed in nature, the bulk of it being bound as lipid. Glycerol is determined

by using Reactions (6.42), (6.43), and (6.38) [1330, vol. IV, pp. 1825 – 1831]:

$$\text{Glycerol} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.30}]{\text{Glycerol kinase}} \text{Glycerol 3-phosphate} + \text{ADP} \tag{6.42}$$

$$\text{ADP} + \text{PEP} \xrightarrow[\text{E.C. 2.7.1.40}]{\text{Pyruvate kinase}} \text{ATP} + \text{Pyruvate} \tag{6.43}$$

The determination of glycerol in wine is of considerable importance because the ratios of ethanol to glycerol and of glycerol to gluconic acid indicate immediately whether the wine has been adulterated by adding glycerol. Beer, spirits, and marzipan are also analyzed for glycerol.

Sugar Alcohols. *Sorbitol* [50-70-4] is obtained by reduction of fructose; it is of interest as a sugar substitute for people suffering from diabetes. The enzymatic determination of sorbitol can be carried out by using the following reaction, which is not specific, because the enzyme also acts on other polyols [1329, vol. VI, pp. 356 – 362]:

$$\text{D-Sorbitol} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.14}]{\text{Sorbitol dehydrogenase}} \text{D-Fructose} + \text{NADH} + \text{H}^+ \tag{6.44}$$

This analysis can be made specific for sorbitol by quantitatively measuring the fructose produced, via Reactions (6.4), (6.5), and (6.3). Another possibility is to couple Reaction (6.44) to Reaction (6.28). Under these conditions, *xylitol* [87-99-0], which also takes part in the sorbitol dehydrogenase reaction, can be measured as well [1329, vol. VI, pp. 484 – 490]:

$$\text{Xylitol} + \text{NAD}^+ \xrightarrow[\text{E.C. 1.1.1.14}]{\text{Sorbitol dehydrogenase}} \text{Xylulose} + \text{NADH} + \text{H}^+ \tag{6.45}$$

The difference between the quantitative determination of sorbitol via fructose and via INT – diaphorase gives the amount of xylitol present in the sample. Diet foods, pomaceous fruit products, ice cream, confectionery, and biscuits are analyzed for sorbitol. Xylitol is estimated in diet foods, chewing gum, and confectionery.

6.4.4. Other Food Ingredients

Cholesterol [57-88-5] is an important steroid with diverse physiological functions; e.g., it is a component of the plasma membrane. At the same time, excessive deposition of cholesterol in vascular tissue leads to atherosclerosis. Cholesterol is determined according to Reaction (6.46) coupled to Reactions (6.47) and (6.48) [1329, vol. VIII, pp. 139 – 148]:

$$\text{Cholesterol} + \text{O}_2 \xrightarrow[\text{E.C. 1.1.3.6}]{\text{Cholesterol oxidase}} \Delta^4\text{-Cholestenone} + \text{H}_2\text{O}_2 \tag{6.46}$$

$$H_2O_2 + \text{Methanol} \xrightarrow[\text{E.C. 1.11.1.6}]{\text{Catalase}} \text{Formaldehyde} + 2\,H_2O \tag{6.47}$$

$$\text{Formaldehyde} + NH_4^+ + 2\ \text{Acetylacetone} \longrightarrow \text{Lutidine dye} + 3\,H_2O \tag{6.48}$$

Cholesterol is used as a measure of the egg content of food, e.g., noodles and eggnog. In general, the cholesterol content of food made from animal fats is very important. When food containing plant materials is analyzed for cholesterol, phytosterols with a 3 β-hydroxyl group (other than lanosterol) interfere with the reaction.

Triglycerides, an important group of lipids containing glycerol, are normally determined in clinical chemistry. Low-fat food is also analyzed enzymatically for triglycerides, which are hydrolyzed with esterase or lipase to yield fatty acids and glycerol (Reaction 6.49). The glycerol generated is determined by Reactions (6.42), (6.43), and (6.38) [1330, vol. IV, pp. 1831 – 1835]:

$$\text{Triglyceride} + 3\,H_2O \xrightarrow[\text{E.C. 3.1.1.1/E.C. 3.1.1.3}]{\text{Esterase and lipase}} \text{Glycerol} + 3\ \text{Fatty acids} \tag{6.49}$$

Acetaldehyde [75-07-0] is a flavor substance in beer, yogurt, and spirits. It is bound by SO_2 in wine. Acetaldehyde dehydrogenase linked to NAD is used to determine acetaldehyde [1329, vol. VI, pp. 606 – 613]:

$$\text{Acetaldehyde} + NAD^+ + H_2O \xrightarrow[\text{E.C. 1.2.1.5}]{\text{Aldehyde dehydrogenase}} \text{Acetic acid} + NADH + H^+ \tag{6.50}$$

Ammonia [7664-41-7]. The simplest compound containing nitrogen and hydrogen is ammonia. Fruit juice, milk, biscuits, cheese, diet food, and meat products are analyzed for ammonia [1329, vol. VIII, pp. 454 – 461]:

$$\alpha\text{-Oxoglutarate} + NADH + H^+ + NH_4^+ \xrightarrow[\text{E.C. 1.4.1.3}]{\text{Glutamate dehydrogenase}} \text{L-Glutamate} + NAD^+ + H_2O \tag{6.51}$$

Nitrate. The determination of nitrate by use of a stable enzyme was described for the first time in 1986:

$$\text{Nitrate} + NADPH + H^+ \xrightarrow[\text{E.C. 1.6.6.2}]{\text{Nitrate reductase}} \text{Nitrite} + NADP^+ + H_2O \tag{6.52}$$

The determination of nitrate is of great importance. This compound is the precursor of nitrite and nitrosamines; hence, it is considered hazardous. Samples of water, beverages, meat, milk products, fruits and

vegetables, and baby food are among the items analyzed for nitrate. Indeed, fertilizers, which are primarily responsible for the nitrate in food, are also subject to analysis.

Sulfite. The enzymatic determination of sulfite was first described in 1983 [1329, vol. VII, pp. 585 – 591]:

$$SO_3^{2-} + O_2 + H_2O \xrightarrow[\text{E.C. 1.8.3.1}]{\text{Sulfite oxidase}} SO_4^{2-} + H_2O_2 \tag{6.53}$$

$$H_2O_2 + NADH + H^+ \xrightarrow[\text{E.C. 1.11.1.1}]{\text{NADH peroxidase}} 2\,H_2O + NAD^+ \tag{6.54}$$

Sulfite is often used as a preservative in food technology and has diverse functions; e.g., it inactivates enzymes, prevents browning, and binds acetaldehyde in wine. Usually, too high concentrations are obtained when sulfur-containing samples, such as cabbage, leek, onion, garlic, and horseradish, are analyzed by either conventional distillation or the enzymatic procedure. However, use of the enzymatic method in wine analysis gives accurate results.

Creatine and Creatinine. Both creatine [57-00-1] and creatinine [60-27-5] are found in muscle. The amount of meat contained in soup, sauces, and meat extracts is determined by analyzing these foods for creatine and creatinine (Reactions 6.55, 6.56, 6.43, and 6.38) [1329, vol. VIII, pp. 488 – 507]:

$$\text{Creatinine} + H_2O \xrightarrow[\text{E.C. 3.5.2.10}]{\text{Creatininase}} \text{Creatine} \tag{6.55}$$

$$\text{Creatine} + \text{ATP} \xrightarrow[\text{E.C. 2.7.3.2}]{\text{Creatine kinase}} \text{Creatine phosphate} + \text{ADP} \tag{6.56}$$

Lecithin [8002-43-5] phosphatidylcholine) is the most important phospholipid occurring in plants and animals. The plural term lecithins refers to emulsifying agents, such as soybean preparations (soy lecithin contains 18 – 20 % phosphatidylcholine). Lecithin can be determined by using phospholipase D, which catalyzes its hydrolytic cleavage to form choline. The choline generated is subsequently measured as Reinecke salt. The method of choice, however, involves the use of phospholipase C from *Bacillus cereus* and of alkaline phosphatase to hydrolyze lecithin (Reactions 6.57 and 6.58), followed by heat inactivation of alkaline phosphatase and determination of choline (Reactions 6.59, 6.43, and 6.38) [1329, vol. VIII, pp. 87 – 104]:

$$\text{Lecithin} + H_2O \xrightarrow[\text{E.C. 3.1.4.3}]{\text{Phospholipase C}} \text{1,2-Diglyceride} + \text{Phosphorylcholine} \tag{6.57}$$

$$\text{Phosphorylcholine} + H_2O \xrightarrow[\text{E.C. 3.1.3.1}]{\text{Alkaline phosphatase}} \text{Choline} + p_i \tag{6.58}$$

$$\text{Choline} + \text{ATP} \xrightarrow[\text{E.C. 2.7.1.32}]{\text{Choline kinase}} \text{Phosphorylcholine} + \text{ADP} \tag{6.59}$$

The sample can be prepared for analysis in various ways. An aqueous suspension can be made by ultrasonic treatment, or a sample solution can be prepared in *tert*-butyl alcohol and water (e.g., egg products, like eggnog). Alternatively, the sample can be dissolved in *tert*-butyl alcohol after extraction with organic solvents, or it can be prepared by alkaline hydrolysis with methanolic potassium hydroxide (e.g., cocoa products). Interpretation of the results can cause further problems, especially if acetone-soluble phosphorus compounds are used as comparison and if no difference is made between “lecithin” and “lecithins.”

Urea [57-13-6] is the most important end product of protein metabolism, most of it being formed in the liver from ammonia. Urea is excreted via the kidneys. The enzyme urease is used for determination of urea (Reactions 6.60 and 6.51) [1329, vol. VIII, pp. 444 – 449]:

$$\text{Urea} + H_2O \xrightarrow[\text{E.C. 3.5.1.5}]{\text{Urease}} 2\,NH_3 + CO_2 \qquad (6.60)$$

Samples analyzed for urea are meat products and milk, where it serves as an index of protein in the diet of the cow. The urea content of wastewater and swimming pool water indicates the urine concentration in these waters.

6.5. Enzymes in Genetic Engineering [1334-1394]

Cloning techniques provide the basis for recombinant DNA technology and its applications in genetic engineering (Figs. 98 and 99), see also → Genetic Engineering. The term “genetic engineering” includes methods for the formation of new combinations of genetic material and for reintroducing and multiplying the recombinant nucleic acid molecules in a new and nonnatural environment.

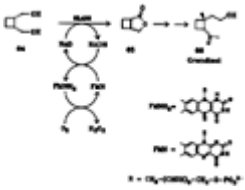

Figure 98. Horse liver alcohol dehydrogenase (HLADH) catalyzed stereoselective oxidation of a meso diol with in situ NAD regeneration.
[Full View]

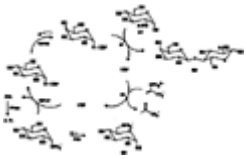

Figure 99. Enzymatic synthesis of lactosamine

GT = galactosyltransferase; PK = pyruvate kinase; PGM = phosphoglucomutase; UDPGP = UDP – glucose pyrophosphorylase; UDPGE = UDP – glucose epimerase; PPase = inorganic pyrophosphatase
[Full View]

The availability of a large variety of class II restriction endonucleases with different sequence specificities is a prerequisite for recombinant DNA technology (Section Restriction Endonucleases and Methylases). *Analytical applications* of this enzyme class can be classified roughly as follows (see Fig. 99):

1. Molecular analysis of chromosome and genome structure (mapping, degree of methylation),
2. Characterization of phenotypically manifest or latent hereditary genetic defects on the DNA level (polymorphism analysis),
3. Evidence of cell degeneration at an early stage by alteration of certain restriction sites in oncogene sequences,
4. Taxonomy of viruses or other disease-causing organisms by the correlation of characteristic fragment patterns,
5. Phylogenetic relationships by comparison of selected restriction sites in essential genes such as the hemoglobin gene.

In addition, DNA- and RNA-modifying enzymes are essential tools for *cloning technology*; these enzymes are discussed in Sections DNA Polymerases, Enzymes in Diagnosis, Enzymes for Food Analysis, Enzymes in Genetic Engineering , Safe Handling of Enzymes.

Cloning technology allows the identical multiplication of specific genes and fragments thereof to high copy numbers. From these overproducing cells — the clones — the coding DNA sequences and, after expression, the corresponding proteins can subsequently be isolated in large quantities and high purity.

The starting materials for gene cloning are first of all complementary mRNA molecules which are being retro-transcribed by the retroviral enzyme reverse transcriptase (E.C. 2.7.7.49) into complementary double-stranded DNA molecules. The starting material may also be the structural gene itself which must first be isolated from the genome, i.e., the entire chromosomal DNA. Furthermore, in addition to the structural gene, essential *regulatory elements* must be available. Isolation of both specific gene fragments and regulatory sequences is achieved by fragmentation of the DNA with sequence-specific class II restriction endonucleases. This requires that the structural gene and the regulatory elements can be excised from the chromosome at exact nucleotide positions.

The cDNA or the fragments obtained by restriction endonucleases can then be modified further by DNA-modifying enzymes such as exonucleases, kinases, or phosphatases. The fragment ends can also be altered in their sequence specificity by attachment of synthetic oligonucleotides, i.e., *linkers* or *adapters*.

With the help of T 4 DNA ligase, the isolated fragments are inserted into special cloning vehicles, the *vectors*. These vectors may, for example, be extrachromosomal circular plasmids (e.g., pBR 322), appropriate derivatives of specific bacteriophages (e.g., λ, M 13), or eukaryotic viruses (e.g., SV 40).

The chimeric molecules obtained in this way, i.e., the recombinant DNA, are then inserted into new host cells, such as *Escherichia coli* bacteria which are easy to cultivate. All vectors carry the *origin element ori*, a DNA sequence for the autonomous replication of the DNA hybrid. The recombinant DNA is thus able to amplify to high copy numbers in transformed cells irrespective of the regular cell cycle. Growth of the host cells under selective conditions, e.g., on culture media containing specific antibiotics, ensures that only transformed bacteria replicate. Bacterial cells that do not take up recombinant DNA molecules die after a short period in such selective media. In this way, only transformed cells are able to grow, and the introduced vectors, and thus the inserted DNA fragments, are selectively amplified to high copy numbers. This is referred to as *cloning* because under selective growth conditions, all bacteria can be traced back to a homogeneous population of identically transformed cells, the cell clone.

Large quantities of identical recombinant DNA molecules can be purified easily from the cultivated bacteria. Subsequently, the inserted DNA fragment can again be excised from the vector with the help of the initially used class II restriction endonucleases. This DNA fragment that contains the gene to be cloned can be characterized, e.g., by physical mapping and sequencing.

In the final stage, the cloned DNA fragment is inserted into efficient expression vectors which have been provided with regulatory control elements. After renewed introduction into appropriate host cells, the regulatory elements are triggered by growth conditions in such a way that primarily the cloned genes are transcribed into complementary mRNA. In prokaryotes, the synthesized mRNA is directly translated into colinear proteins by the ribosomes. In eukaryotes, the mRNA is first synthesized as a precursor mRNA which is processed into mature mRNA by splicing. The structure and sequence of the various RNA species can be analyzed by using RNA-modifying enzymes. As in prokaryotes, the mature mRNA is then translated into colinear protein. Often the overproduced protein is synthesized in such high concentrations that it precipitates in the cell and can, therefore, be isolated from the cell mass in high yield and with high purity.

6.5.1. Restriction Endonucleases and Methylases [1395-1402]

6.5.1.1. Classification

In the search for enzymes that are the basis of restriction modification systems [1403], [1404], class I enzymes were discovered in 1968 [1405], [1406]. The first member of a class II restriction endonuclease, *Hin*dII, was found in 1970 [1407], [1408]. Later, enzymes of class III were also isolated [1409], which have properties between those of class I and class II. A selection of important class II restriction endonucleases

and methylases is given in Table 43.

Table 43. Recognition sequences of important class II restriction endonucleases and methylases

Position no.	Recognition sequence	Enzyme	Number of recognition sites λ	Ad2	SV 40	ΦX 174	M 13mp7	pBR 322	pBR 328	Microorg
A. Class II enzymes with palindromic recognition sequences										
A.1. Tetra-, hexa-, or octanucleotide recognition sequences										
A.1.1. Internal AT palindriomes										
1	AATT	–	188	0	39	25	65	8	13	
2	AAATTT	–	16	13	4	5	3	0	0	
3	$G/\overset{0}{A}\overset{+}{A}TT\overset{+}{C}$	*Eco*RI	5	5	1	0	2	1	1	*Escherich* 13
	$GA\overset{M}{A}TTC$	M · *Eco*RI	5	5	1	0	2	1	1	*Escherich* 13
4	CAATTG	–	8	4	4	1	0	0	0	
5	TAATTA	–	8	4	2	1	3	0	0	
6	$G\overset{m}{A}/TC$	*Dpn*I	116	87	8	0	8	22	27	*Diplococc pneumoni*
	$/G\overset{+}{A}T\overset{0}{C}$	*Mbo*I	116	87	8	0	8	22	27	*Moraxellc*
	$G\overset{+}{A}TC$	*Nde*II	116	87	8	0	8	22	27	*Neisseria denitrificc*
	$/G\overset{0}{A}T\overset{+}{C}$	*Sau*3AI	116	87	8	0	8	22	27	*Staphyloc aureus*
	$G\overset{M}{A}TC$	M · *Eco dam*	116	87	8	0	8	22	27	*Escherich* HB 101
7	$A/G\overset{0}{A}T\overset{+}{C}T$	*Bgl*II	6	11	0	0	1	0	0	*Bacillus g*
8	$G/G\overset{0}{A}T\overset{+}{C}\overset{0}{C}$	*Bam*HI	5	3	1	0	2	1	1	*Bacillus amyloliqu*
	$GGATC\overset{M}{C}$	M · *Bam*HI	5	3	1	0	2	1	1	*Bacillus amyloliqu* RUB 500
	$\binom{A}{G}/GATC\binom{T}{C}$	*Xho*II	21	22	3	0	4	8	7	*Xanthomc holcicola*
9	$CG\overset{0}{A}T/\overset{+}{C}G$	*Pvu*I	3	7	0	0	1	1	1	*Proteus vu*
10	$T/G\overset{+}{A}T\overset{0}{C}A$	*Bcl*I	8	5	1	0	0	0	1	*Bacillus caldolytic*
11	CATG/	*Nla*III	181	183	16	22	15	26	27	*Neisseria*
12	ACATGT	–	2	9	0	0	3	1	0	
13	$GCATG/\overset{0}{C}$	*Sph*I	6	8	2	0	0	1	1	*Streptomy phaeochrc*
14	$\overset{+}{C}/CATGG$	*Nco*I	4	20	3	0	0	0	1	*Nocardia*
	$C/C\binom{A}{T}\binom{T}{A}GG$	*Sty*I	10	44	8	0	0	1	2	*Escherich* KM201 [p S-a]
15	TCATGA	–	8	3	2	3	1	4	4	

15	TCATGA	–	8	3	2	3	1	4	4	
16	TATA	–	113	65	17	11	16	8	13	
17	ATATAT	–	11	3	0	0	4	1	3	
18	GTATAC	*Sna*I	3	3	0	0	0	1	0	*Sphaeroti* C
19	CTATAG	–	0	4	2	1	0	0	0	
20	TTATAA	–	12	4	3	1	2	0	0	
A.1.2. Internal GC palindromes										
21	$\mathrm{AG/\overset{+}{C}T}$	*Alu*I	143	158	34	24	24	16	14	*Arthrobac*
	$\mathrm{AG\overset{M}{C}T}$	M · *Alu*I	143	158	34	24	24	16	14	*Arthrobac*
22	$\mathrm{\overset{+}{A}/AG\overset{+}{C}TT}$	*Hin*dIII	6	12	6	0	0	1	1	*Haemoph influenzae*
23	$\mathrm{G\overset{0}{A}G\overset{+}{C}T/C}$	*Sac*I	2	16	0	0	0	0	0	*Streptomy achromog*
24	$\mathrm{CAG/\overset{+}{C}TG}$	*Pvu*II	15	24	3	0	3	1	1	*Proteus vu*
25	TAGCTA	–	2	4	2	0	3	0	0	
26	$\mathrm{GG/\overset{+}{C}\overset{0}{C}}$	*Hae*III	149	216	18	11	15	22	30	*Haemoph aegyptius*
	$\mathrm{GG\overset{M}{C}C}$	M · *Hae*III	149	216	18	11	15	22	30	*Haemoph aegyptius*
27	$\mathrm{AGG/\overset{0}{C}CT}$	*Stu*I	6	11	7	1	0	0	0	*Streptomy tubercidic*
28	$\mathrm{GGG\overset{+}{C}C/C}$	*Apa*I	1	12	1	0	0	0	0	*Acetobact pasteurian pateurian*
	$\mathrm{G\binom{A}{G}GC\binom{T}{C}/C}$	*Ban*II	7	57	2	0	1	2	2	*Bacillus aneurinol*
29	$\mathrm{C/GG\overset{+}{C}CG}$	*Xma*II	2	19	0	0	0	1	2	*Xanthomo malvacear*
	$\mathrm{GC/GG\overset{+}{C}CG\overset{0}{C}}$	*Not*I	0	7	0	0	0	0	0	*Nocardia caviarum*
	$\mathrm{\binom{T}{C}/GG\overset{+}{C}C\binom{A}{T}}$	*Eae*I	39	70	0	2	3	6	7	*Enterobac aerogenes*
30	$\mathrm{TGG/\overset{+}{C}CA}$	*Bal*I	18	17	0	0	1	1	1	*Brevibact albidum*
31	CG/CG	*Fnu*DII	157	303	0	14	18	23	24	*Fusobace nucleatum*
32	$\mathrm{\overset{0}{A}/CGCGT}$	*Mlu*I	7	5	0	2	0	0	0	*Micrococ*
	$\mathrm{A/C\binom{A}{G}\binom{T}{C}GT}$	*Afl*III	20	25	0	2	3	1	0	*Anabaena aquae*
33	G/CGCGC	*Bss*HII	6	52	0	1	0	0	0	*Bacillus stearother* H3
34	$\mathrm{\overset{+}{C}CGC/GG}$	*Sac*II	4	33	0	1	0	0	0	*Streptomy achromog*
35	$\mathrm{TCG/CG\overset{+}{A}}$	*Nru*I	5	5	0	2	0	0	1	*Norcadia*
36	TGCA		272	206	36	18	15	21	19	
37	ATGCA/T	*Nsi*I	14	9	3	0	0	0	0	*Neisseria*
38	G/TGCAC	*Apa*LI	4	7	0	1	1	3	2	*Acetobact pasteurian*

	$G\binom{A}{T}G\overset{+}{C}\binom{T}{A}/C$	*Hgi*AI	28	38	0	3	2	8	7	*Herpetosi giganteus*
	$G\binom{A}{G}{T}G\overset{+}{C}\binom{T}{C}{A}/C$	*Bsp*1286	38	105	4	3	4	10	10	*Bacillus s*
39	$\overset{+}{C}TGC\overset{+}{A}/G$	*Pst*I	28	30	2	1	1	1	1	*Providenc* 164
	CTGCAG	M · *Pst*I	28	30	2	1	1	1	1	*Providenc* 164
40	TTGCAA	–	13	23	3	2	2	1	1	
A.1.3. Internal CG palindromes										
41	$A/\overset{+}{C}GT$	*Mae*II	143	83	0	19	22	10	12	*Methanoc aeolicus* F
42	AACGTT	–	7	3	0	3	2	4	5	
43	GACGT/C	*Aat*II	10	3	0	1	0	1	1	*Acetobact*
44	CACGTG	–	3	10	0	0	0	0	0	
45	TAC/GTA	*Sna*BI	1	0	0	0	1	0	0	*Sphaeroti*
46	$G\overset{+}{C}G/\overset{+}{C}$	*Cfo*I	215	375	2	18	25	31	25	*Clostridiu fromicoac*
	$G\overset{+}{C}G/\overset{+}{C}$	*Hha*I	215	375	2	18	25	31	25	*Haemoph haemolyti*
	$G\overset{M}{C}GC$	M · *Hha*I	215	375	2	18	25	31	25	*Haemoph haemolyti*
47	AGC/GCT	*Eco*47III	2	13	1	0	2	4	3	*Escherich* RFL47
48	$GG/CGC\overset{0}{C}$	*Nar*I	1	20	0	2	1	4	5	*Nocardia argentien*
	$\binom{A}{G}G\overset{+}{C}GC/\binom{T}{C}$	*Hae*II	48	76	1	8	6	11	9	*Haemoph aegyptius*
	$G\binom{A}{G}/\overset{+}{C}G\binom{T}{C}\overset{+}{C}$	*Aha*II	40	44	0	7	1	6	7	*Aphanoth halophytic*
	$G/G\binom{T}{C}\binom{A}{G}CC$	*Ban*I	25	57	1	3	6	9	12	*Bacillus aneurinol*
49	CGCGCG	–	1	48	0	0	1	1	0	
50	TGCGCA	*Fsp*I	15	17	0	1	0	4	3	*Fischerell*
51	$\overset{+}{C}/\overset{+}{C}GG$	*Hpa*II	328	171	1	5	19	26	33	*Haemoph parainflue*
	$C\overset{M}{C}GG$	M · *Hpa*II	328	171	1	5	19	26	33	*Haemoph parainflue*
	$\overset{+}{C}/\overset{0}{C}GC$	*Msp*I	328	171	1	5	19	26	33	*Moraxella*
	$\overset{M}{C}CGG$	M · *Msp*I	328	171	1	5	19	26	33	*Moraxella*
52	ACCGGT	–	13	5	0	0	0	0	1	
53	$G\overset{+}{C}C/GG\overset{+}{C}$	*Nae*I	1	13	1	0	1	4	6	*Nocardia aerocolon*
	$\binom{A}{T}CCGG\binom{T}{C}$	*Cfr*10I	61	40	1	0	1	7	10	*Citrobacte*
54	$CC\overset{+}{C}/GGG$	*Sma*I	3	12	0	0	0	0	0	*Serratia m* S_b
	$C/\overset{+}{C}\overset{0}{C}GGG$	*Xma*I	3	12	0	0	0	0	0	*Xanthomo malvacear*
	$C/\binom{\overset{0}{T}}{C}\overset{+}{C}G\binom{\overset{+}{A}}{G}G$	*Ava*I	8	40	0	1	1	1	1	*Anabaena*

55	T/CCGGA	*Bsp*MII	24	8	0	0	0	1	1	*Bacillus s*
56	$\mathrm{T/\overset{0}{C}G\overset{\dagger}{A}}$	*Taq*I	121	50	1	10	14	7	13	*Thermus* YTI
	$\mathrm{TCG\overset{M}{A}}$	M · *Taq*I	121	50	1	10	14	7	13	*Thermus* YTI
57	$\mathrm{AT/CG\overset{\dagger}{A}T}$	*Cla*I	15	2	0	0	2	1	1	*Caryopha* L
	$\mathrm{ATCG\overset{M}{A}T}$	M · *Cla*I	15	2	0	0	2	1	1	*Caryopha* L
58	$\mathrm{G/T\overset{\dagger}{C}G\overset{\dagger}{A}C}$	*Sal*I	2	3	0	0	2	1	1	*Streptomy* G
	$\mathrm{GT}\binom{\mathrm{T}}{\mathrm{C}}/\binom{\overset{\dagger}{\mathrm{A}}}{\mathrm{G}}\mathrm{AC}$	*Hinc*II	35	25	7	13	2	2	2	*Haemoph influenzae*
	$\mathrm{GT}\binom{\mathrm{T}}{\mathrm{C}}/\binom{\mathrm{A}}{\mathrm{G}}\overset{\dagger}{\mathrm{A}}\overset{0}{\mathrm{C}}$	*Hind*II	35	25	7	13	2	2	2	*Haemoph influenzae*
	$\mathrm{GT}/\binom{\mathrm{A}}{\mathrm{C}}\binom{\mathrm{T}}{\mathrm{G}}\overset{\dagger}{\mathrm{A}}\mathrm{C}$	*Acc*I	9	17	1	2	2	2	1	*Acinetoba calcoacet*
59	$\mathrm{C/T\overset{\dagger}{C}G\overset{\dagger}{A}G}$	*Xho*I	1	6	0	1	0	0	0	*Xanthomo holcicola*
60	TT/CGAA	*Fsp*II	7	1	0	0	0	0	2	*Ficherell*
A.1.4. Internal TA palindromes										
61	ATAT	–	230	81	21	20	42	9	19	
62	AAT/ATT	*Ssp*I	20	5	6	1	6	1	2	*Sphaeroti*
63	$\mathrm{G\overset{\dagger}{A}T/ATC}$	*Eco*RV	21	9	1	0	0	1	1	*Escherich* [pLG74]
64	CA/TATG	*Nde*I	7	2	2	0	3	1	0	*Neisseria denitrific*
65	TATATA	–	4	5	0	0	1	1	1	
66	GT/AC	*Rsa*I	113	83	12	11	18	3	4	*Rhodopse sphaeroid*
67	AGT/ACT	*Sca*I	5	5	0	0	0	1	2	*Streptomy caespitos*
68	$\mathrm{G/GTACC\overset{\dagger}{C}}$	*Asp*718	2	8	1	0	0	0	0	*Achromob* species 71
	$\mathrm{GGTACC\overset{0}{C}/\overset{0}{C}}$	*Kpn*I	2	8	1	0	0	0	0	*Klebsiella pneumoni*
69	CGTACG	–	1	4	0	2	0	0	0	
70	TGTACA	–	5	5	2	0	1	0	0	
71	C/TAG	*Mae*I	13	54	12	3	4	5	4	*Methanoc aeolicus* F
72	A/CTAGT	*Spe*I	0	3	0	0	0	0	0	*Sphaeroti*
73	G/CTAGC	*Nhe*I	1	4	0	0	0	1	1	*Neisseria*
74	C/CTAGG	*Avr*II	2	2	2	0	0	0	0	*Anabaena* uw
75	$\mathrm{T/\overset{\dagger}{C}TAG\overset{\dagger}{A}}$	*Xba*I	1	5	8	0	0	0	0	*Xanthomo*
76	TTAA	–	195	115	47	35	61	15	17	
77	ATTAAT	–	17	3	3	2	6	1	1	
78	$\mathrm{GTT/A\overset{\dagger}{A}\overset{0}{C}}$	*Hpa*I	14	6	4	3	0	0	0	*Haemoph parainflue*

										paradoxus
79	C/TTAAG	*Afl*II	3	4	1	2	0	0	0	*Anabaena aquae*
80	TTT/AAA	*Dra*I	13	12	12	2	5	3	5	*Deinococc radiophilu*
A.2. Class II enzymes with penta- or heptanucleotide recognition sequences										
81	$G/G\binom{A}{T}C\overset{+}{C}$	*Ava*II	35	73	6	1	1	8	7	*Anabaena*
	$\binom{A}{G}G/G\binom{A}{T}CC\binom{T}{C}$	*Ppu*MI	3	25	1	0	0	2	0	*Pseudomo putida* M
	$CG/G\binom{A}{T}CCG$	*Rsr*II	5	2	0	0	0	0	0	*Rhodopse sphaeroid*
82	$\overset{+m}{C}\overset{m}{C}/\binom{A}{T}GG$	*Apy*I	71	136	17	2	7	6	10	*Arthroba pyridinoli*
	$\overset{0}{C}\overset{0}{C}/\binom{A}{T}GG$	*Bst*NI	71	136	17	2	7	6	10	*Bacillus staerother*
	$/\overset{0}{C}\overset{+}{C}\binom{A}{T}GG$	*Eco*RII	71	136	17	2	7	6	10	*Escherich* R245
83	$C\overset{+}{C}/\binom{G}{C}GG$	*Nci*I	114	97	0	1	4	10	10	*Neisseria*
84	$G/\overset{+}{A}NT\overset{+}{C}$	*Hin*fI	148	72	10	21	26	10	10	*Haemoph influenzae*
85	$G/GN\overset{+}{C}\overset{+}{C}$	*Cfr*131	74	164	11	2	4	15	16	*Citrobact* RFL13
	$G/GN\overset{+}{C}\overset{+}{C}$	*Sau*96I	74	164	11	2	4	15	16	*Staphyloc aureus* PS
	$\binom{A}{G}G/GNC\overset{+}{C}\binom{T}{C}$	*Dra*II	3	44	3	0	0	4	2	*Deinococc radiophilu*
86	$G\overset{+}{C}/NGC$	*Fnu*4HI	380	411	24	31	15	42	37	*Fusobacte nucleatum*
87	$\overset{0}{C}\overset{+}{C}NGG$	*Scr*FI	185	233	17	3	11	16	20	*Streptoco cremonis*
88	/GTNAC	*Mae*II	156	118	14	17	25	17	18	*Methanoc aeolicus* F
	G/GTNACC	*Bst*EII	13	10	0	0	0	0	0	*Bacillus stearother* ET
89	$\overset{+}{C}/TNAG$	*Dde*I	104	97	20	14	29	8	9	*Desulfovi desulfuric* Norway
	$CC/TNA\overset{0}{G}\overset{0}{G}$	*Sau*I	2	7	0	0	1	0	1	*Streptomy aureofaci* IKA18/4
A.3. Class II enzymes with recognition sequences containing internal $(N)_x$ sequences										
90	GGN/NCC	*Nla*IV	82	178	16	6	6	24	26	*Neisseria*
91	CACNNN/GTG	*Dra*III	10	10	1	1	1	0	0	*Deinococc radiophilu*
92	GAANN/NNTTC	*Asp*700	24	5	0	3	2	2	1	*Achromob* species 70
93	$G\overset{+}{A}\overset{0}{A}NN/NNTTC$	*Xmn*I	24	5	0	3	2	2	1	*Xnathomo manihotis*

94	$GC\overset{+}{C}\overset{0}{C}(N)_4/NGGC$	*Bgl*I	29	20	1	0	1	3	5	*Bacillus g*
95	$\overset{+}{C}C\overset{+}{A}(N)_5/NTGG$	*Bst*XI	13	10	1	3	0	0	0	*Bacillus stearother* X1
96	$GG\overset{0}{C}C(N)_4/NGG\overset{0}{C}\overset{0}{C}C$	*Sfi*I	0	3	1	0	0	0	0	*Streptomy fimbriatus*
B. Class II enzymes with nonpalindromic recognition sequences										
B.1. Class II enzymes										
97	GAATG-CN/N CTTAC/GN-N	*Bsm*I	46	10	4	3	1	1	3	*Bacillus stearother* NUB36
98	TGACN/N-NGTCN ACTGN-N/NCAGN	*Tth*111I	1	6	0	0	0	2	0	*Thermus thermoph*
B.2. Class IIS enzymes										
99	$G\overset{+}{C}AGC(N)_8$ $CGTCG(N)_{12}$	*Bbv*I	199	179	22	14	8	21	11	*Bacillus b*
100	$ACCTGC(N)_4$ $TGGACG(N)_8$	*Bsp*MI	41	39	0	3	4	1	2	*Bacillus* s
101	$G\overset{(+)}{G}ATG(N)_9$ $CCTAC(N)_{13}$	*Fok*I	150	78	11	8	4	12	11	*Flavobact okeanoko*
102	$GACG\overset{+}{C}(N)_5$ $CTGCG(N)_{10}$	*Hga*I	102	87	0	14	7	11	12	*Haemoph gallinaru*
103	$GGT\overset{(+)}{G}\overset{+}{A}(N)_8$ $CCACT(N)_7$	*Hph*I	168	99	4	9	18	12	18	*Haemoph parahaem*
104	$GAAG\overset{+}{A}(N)_8$ $CTTCN(N)_7$	*Mbo*II	130	113	16	11	12	11	12	*Moraxell*
105	$CCTC(N)_7$ $GGAG(N)_7$	*Mnl*I	262	397	51	34	61	26	30	*Moraxell nonliquef*
106	$GCAT\overset{0}{C}(N)_5$ $CGTAG(N)_9$	*Sfa*NI	169	84	6	12	7	22	17	*Streptoco faecalis*

Class I enzymes are coded by the three gene segments *hsd*S, *hsd*R, and *hsd*M, expressing the functions for sequence specificity, restriction, and modification methylation. The corresponding endonucleases and DNA methyltransferases are enzyme complexes of high molecular mass. These enzyme complexes require magnesium ions, ATP, and SAM as essential cofactors. Class I enzymes bind to and methylate specific sequences, but cut unspecifically after translocation into the 3′-direction [1381], [1490-1492]. An example of a class I system is the *Eco*B restriction — modification system. The early steps of SAM-binding, i.e., enzyme activation and formation of the initial enzyme – DNA complex, are common to both restriction and modification; however, these reaction pathways diverge after the stable complex has been formed. In the case of recognition, ATP leads to the release of the enzyme from modified DNA; it stimulates translocation in unmodified recognition complexes. This reaction is coupled to ATP hydrolysis. Sequential cleavage of both DNA strands occurs in a limited sequence stretch some 100 to 1000 base pairs away from the binding site.

Class II enzymes are expressed as two single proteins from two different genes acting as separate enzymes. Class II endonucleases hydrolyze both DNA strands at specific phosphodiester bonds within, or very near, the recognition sequence. Their activity depends only on magnesium ions. The corresponding class II DNA methyltransferases methylate both DNA strands, independent of the presence of the respective nuclease in an SAM-dependent reaction. An example of a class II system is the *Eco*RI restriction and modification system. Class II restriction endonucleases are by far the most important tools for the synthesis of recombinant DNA because of their absolute sequence specificity for both binding and cleavage [1493], [1494], [1495], [1496], [1497], [1498], [1372].

Class III enzymes are also expressed from two genes. However, like class I enzymes, the corresponding endonucleases and DNA methyltransferases form enzyme complexes in which both activities act simultaneously. Class III enzymes recognize specific sequences, but cleave rather specifically 25 to 27 base pairs downstream from this region. The endonuclease activity depends on magnesium ions and ATP, whereas methylation occurs in the presence of SAM [1490], [1499], [1500], [1409], [1492].

Class II restriction endonucleases and methylases are found in all prokaryotic cells [1396], [1398], [1400]. In accordance with the proposals made by SMITH and NATHANS, the various enzymes are abbreviated by taking the first letters of the bacterial name [1501]. The various endonucleases and DNA-methyltransferases occur in almost all genera of the two kingdoms of eu- and archaebacteria. In particular, type II enzymes are known in both gram-positive and gram-negative eubacteria, in cyanobacteria, as well as in thermophilic, halophilic, methanogenic, and sulfur-dependent archaebacteria. In contrast, eukaryotic restriction – modification systems have not been described yet. Specifically cleaving eukaryotic endonucleases, such as the HO-nuclease in yeast or the retroviral endonucleases, have other functions and are characterized by more complex recognition sequences [1502], [1503].

In addition to class II restriction – modification systems of single specificity, a whole series of multiple component systems were discovered to contain several restriction – modification systems of different specificity. Particularly complex systems, which contain up to seven different specificities, are found in certain species of *Nostoc* (*Nsp* 7524I – V) and in *Neisseria gonorrhoeae* (*Ngo*I – VII) [1504], [1505].

Methods for separating the various enzyme species include size fractionation, column chromatography with anion and cation exchangers, and hydrophobic or affinity materials with DNA or DNA-like heparin as ligands [1506], [1507]. At the end of 1986 as many as 634 different class II restriction endonucleases with at least 111 different specificities were known, which have been isolated from 564 different bacterial strains (Table 43). The corresponding restriction methylases have been characterized only for the most important systems. Enzymes are described completely in [1396], [1398], [1400].

The active enzymes of the class II restriction endonucleases consist of protein dimers of identical subunits, whereas the corresponding DNA methyltransferases act as monomeric proteins. The molecular masses under denaturing conditions range from 20×10^3 to 70×10^3 dalton for both class II endonucleases and DNA methyltransferases. The only protein of higher molecular mass known so far is *Bst*I which is about 400×10^3 dalton [1494].

As an example, the *Eco*RI endonuclease monomer is a 277 amino acid protein with a molecular mass of 31 063 dalton, which is processed after translation by removing the terminal *N*-formylmethionine. The corresponding methylase protein consists of 326 amino acids with a molecular mass of 38 048 dalton. This protein is also processed by removing the N-terminal dipeptide fMet-Ala [1510], [1512].

Examples of cloned restriction modification systems are *Bsu*RI from *Bacillus subtilis* [1433], [1361]; *Eco*RI [1508], [1509], [1510], [1511], [1513], [1512]; *Eco*RII [1417]; and *Eco*RV [1514], [1515], [1516] from *Escherichia coli*; *Hha*II from *Haemophilus parainfluenzae* [1517], [1518]; *Pae*R7 from *Pseudomonas aeruginosa* [1519], [1356]; *Pst*I from *Providencia stuartii* [1520], [1369]; and *Pvu*II from *Proteus vulgaris* [1521]. According to gene mapping of the various systems, the proteins for both endonuclease and DNA methyltransferase may be plasmid or chromosomally coded. Further analysis of the two genes shows that they are either arranged in tandem to form transcription units or expressed as independent genes [1510], [1433], [1512], [1361]. When the known protein and genetic sequences of corresponding endonucleases and

DNA methyltransferases were compared, no statistically significant sequence homology could be detected [1433], [1512]. Circular dichroism studies also indicate extensive differences in secondary structure [1512]. According to these observations, the two proteins have different evolutionary origins and interact with DNA in different ways.

6.5.1.2. Activity of Class II Restriction Endonucleases

An enzyme unit, U, is normally defined as the amount of protein that completely digests 1 µg of the DNA from bacteriophage λ in 1 h at 37 °C. To analyze the completeness of the reaction, the fragment mixture is separated in 0.5 – 2 % agarose gels by horizontal or vertical electrophoresis (see Section Electrophoresis). The use of λ DNA as the standard test substrate has two advantages. This DNA has 48 502 base pairs and contains cleavage sites for almost all known restriction endonucleases, which allows a comparison of activities [1396], [1523], [1524], [1525]. Because the complete sequence of λ DNA is known, any cleavage pattern can be calculated in advance by computer.

6.5.1.2.1. Reaction Parameters

The activity of *Eco*RI and other class II restriction endonucleases not only depends on the presence of specific recognition sites and essential divalent cations, but also very strongly on such reaction parameters as temperature, ionic strength, and pH [1372]. The activity of class II restriction endonucleases is also frequently stimulated by the presence of reducing agents such as 2-mercaptoethanol or dithiothreitol.

On the plasmid pBR322, *Eco*RI has a temperature optimum of 40 °C. The apparent activation energy for the overall enzymatic reaction is 5.1 kJ/mol. The enzyme is inactivated irreversibly above 45 – 50 °C [1526]. The optimal pH for a specific *Eco*RI cleavage is 7.5; the optimal NaCl concentration, 150 mmol/L [1377].

The optimal temperature for enzymatic activity often corresponds to the optimal growth temperature of the organism from which the enzyme is isolated [1372]. Accordingly, enzymes such as *Bcl*I, *Bst*EII, *Mae*I – III, or *Taq*I with high temperature optima are isolated from thermophilic organisms [1528], [1529], [1531], [1527]. Enzymes are also known which require higher ionic strength (e.g., *Mae*III with 350 mmol/L NaCl) or higher pH values (e.g., *Bgl*I pH 9.5) for maximal activity [1532], [1527].

6.5.1.2.2. Additional Structural Requirements Influencing Activity

The structure of the substrate also strongly influences the activity of class II restriction endonucleases at individual cleavage sites. Important parameters are:

1. G–C content and base distribution in natural DNA;
2. Size of the DNA and cleavage frequencies;
3. Length and base composition of the flanking sequences;
4. Position of the cleavage sites with respect to each other;
5. DNA tertiary structure;
6. Modification of DNA and protein attachment.

Guanosine – Cytidine Content. In natural DNA, the G–C content varies broadly. The DNA from *Micrococcus lysodeikticus* contains 72 % G–C residues, but the DNA from *Euglena gracilis* chloroplasts is composed of only 25 % G–C residues [1822], [1365]. Although this second DNA consists of about 130×10^3 base pairs, there is not a single cleavage site for *Sma*I recognizing only G–C base pairs [CCC/GGG]. Furthermore, the base distribution in natural DNA is nonstatistical. In eukaryotic SV40 DNA, the dinucleotide CpG occurs very rarely [1534]. As a result, the cleavage frequency for such class II restriction endonucleases as *Mae*II, *Cfo*I, *Hpa*I, *Taq*I, or *Fnu*DII, which contain the CpG dinucleotide in their tetranucleotide recognition sequences, is extremely low [1396], [1463]. By contrast, the number of cleavage sites on prokaryotic pBR322 DNA of similar size is about ten times higher, with values like those obtained from enzymes such as *Alu*I or *Hae*III, which have no CpG dinucleotides in their tetranucleotide recognition sequences [1396].

Length of DNA. The influence of DNA length can be demonstrated by the *Sal*I cleavage of pBR322 DNA [1535]. The length of this plasmid is only about one-tenth the length of the bacteriophage λ DNA [1537], [1525], [1538]. The λ DNA has only two cleavage sites, but the much smaller pBR322 DNA still has one

cleavage site. Because the number of cleavage sites per microgram of DNA increases fivefold with pBR322 DNA, digestion of the same amount of λ DNA with *Sal*I requires about five times more enzyme as compared with the plasmid DNA.

Flanking Sequences. The influence of flanking sequences is demonstrated by the observation that short oligonucleotides consisting, for example, only of the hexanucleotide *Eco*RI recognition sequence GAATTC are not cleaved. Even if several hexanucleotides were ligated to longer DNA molecules to stabilize the double strand — which would be completely denatured in the short hexanucleotide at 37 °C — they could not be digested with *Eco*RI. Only if the recognition sequence is extended on both ends by one or more flanking nucleotides the enzyme can exert its activity. However, in these cases, higher quantities of enzyme and longer reaction times are required. Normal reaction rates can be achieved only by further increasing the length of the flanking sequences [1539].

Base Composition of Flanking Sequences. The base composition within the flanking sequence distinctly influences the actual cleavage rate at the individual recognition site. This effect was first observed for *Eco*RI which digests the various sites on λ or Ad2 DNA with rates differing by a factor of ten [1540], [1495]. A similar preferential digestion has also been observed with *Hin*dIII on λ, *Hga*I on Φ X 174, and *Pst*I on plasmid pSM1 DNA [1541], [1542]. In the latter case, significant resistance to cleavage was conferred on G – C-rich flanking regions.

More significant differences in individual cleavage rates were observed on pBR322 DNA with *Nae*I which recognizes the G – C-rich sequence GCC/GGC [1543]. The enzyme *Nae*I cuts pBR322 DNA at four different sites, two of which are digested rapidly and the third 5 – 10 times more slowly. Complete cleavage at the fourth site, however, proceeds 50 times slower [1544]. Similar slow cleavage sites were also observed for the enzymes *Nar*I [GG/CGCC], *Sac*II [CCG/CGG], and *Xma*III [C/GGCCG] which recognize sequences with alternating G and C residues [1396], [1545], [1546]. In all these examples, the slower reaction rates at particular sites are attributed to the sequences within the flanking regions. However, whether a particular sequence element or an altered DNA conformation is responsible for the strong decrease in cleavage rate is still unknown.

Relative Position of Cleavage Sites. A similar decrease in cleavage rate is observed if two or more recognition sites that would be digested readily alone are only a few nucleotides apart from each other. After digestion at one of the two sites, the essential flanking sequences for efficient cleavage at the neighboring position are missing. Therefore, after the first cleavage, the DNA can only be hydrolyzed very slowly at the second site. This is the case for the multiple cloning sites of the M13mp derivatives that are used frequently for sequencing studies. In these polylinker regions, the recognition sequences for several restriction endonucleases are located directly next to each other [1379].

6.5.1.3. Specificity of Class II Restriction Endonucleases

The specificity of the presently known class II restriction endonucleases is defined by three criteria:

1. recognition sequence,
2. position of cleavage site, and
3. influence of methylation.

However, all three characteristics of enzyme specificity are known only for a limited number of enzymes. For many class II restriction endonucleases, only the recognition sequences have so far been determined [1396], [1400].

6.5.1.3.1. Palindromic Recognition Sequences

For the majority of class II enzymes, the recognition sequence on the double-stranded DNA is characterized by a two-fold axis of rotational symmetry. Such symmetric recognition sequences are called palindromes. In a palindromic sequence, a horizontal complementary arrangement exists in addition to the normal base pairing between A – T and G – C. For example, the *Eco*RI recognition sequence is

G/AATT-C

C-TTAA/G

Most palindromic recognition sequences are tetra-, penta-, hexanucleotides, but defined octanucleotides also exist [1547]. In addition, palindromic recognition sequences of extended length exist, e.g., for *Bst*XI, in which only the flanking nucleotides are specifically recognized. In most of these cases, the flanking trinucleotides are unambiguously defined. However, one case of nonequivocally defined flanking trinucleotides is known with *Dra*II [1548], [1549]:

$$\binom{A}{G}G/GNCC\binom{T}{C}$$

An example of an enzyme that recognizes flanking tetranucleotides is *Sfi*I [GGCCNNNN/NGGCC] [1550].

Palindromic recognition sequences can be arranged according to homologies within their recognition sequences (Table 44). In this approach, the sequences are ordered by starting from the shortest central palindromes AT, GC, CG, and TA. Tetranucleotide palindromes are created for each dinucleotide sequence by adding one nucleotide on the left side and its complementary nucleotide on the right side. Repeated application of this cycle generates hexa- and octanucleotide palindromes. Out of the 16 possible tetranucleotides, 11 sequences are covered by corresponding enzymes. For the 64 possible hexanucleotides, 49 enzyme specificities are known [1396].

The cleavage sites for class II restriction endonucleases lie within or directly next to the recognition sequence; they are marked by a slash (/) in Table 45. As in recognition, the positions in the two complementary strands have a two-fold rotational symmetry. Either *blunt* double-stranded ends or *protruding* single-stranded 5′- or 3′-ends are formed. The lengths of the protruding terminal single-stranded regions depend on the position of the cleavage sites within the recognition sequence. Independent of cleavage positions, all class II restriction endonucleases form fragments that have 5′-phosphate and 3′-hydroxyl fragment ends. The only exception known so far is *Nci*I, which produces 3′-phosphorylated fragment termini [1551].

Table 44. Palindromic recognition sequences of restriction endonucleases

	AT	GC	CG	TA	
A		*Alu*I	*Mae*II		T
G	*Sau*3AI	*Hae*III	*Cfo*I	*Rsa*I	C
C	*Nla*III	*Fnu*DII	*Hpa*II	*Mae*I	G
T			*Taq*I		A
AA		*Hin*dIII		*Ssp*I	TT
GA	*Eco*RI	*Sac*I	*Aat*II	*Eco*RV	TC
CA		*Pvu*II	*Pma*Cl	*Nde*I	TG
TA			*Sna*Bl		TA
AG	*Bgl*II	*Stu*I	*Eco*47III	*Sca*I	CT
GG	*Bam*HI	*Apa*I	*Nar*I	*Asp*718	CC
CG	*Pvu*I	*Xma*III		*Spl*I	CG
TG	*Bcl*I	*Bal*I	*Mst*I		CA
AC		*Mlu*I		*Spe*I	GT
GC	*Sph*I	*Bss*HII	*Nae*I	*Nhe*I	GC
CC	*Nco*I	*Sac*II	*Sma*I	*Avr*II	GG
TC	*Bsp*HI	*Nru*I	*Bsp*MII	*Xba*I	GA

AT		*Nsi*I	*Cla*I		AT
GT	*Sna*I	*Apa*LI	*Sal*I	*Hpa*I	AC
CT		*Pst*I	*Xho*I	*Afl*II	AG
TT			*Mla*I	*Dra*I	AA
	AT	GC	CG	TA	

Table 45. Cleavage positions of restriction endonucleases and generated fragment ends

Enzyme	Recognition sequence	End produced
*Hae*III	GG/CC CC/GG	blunt end
*Eco*RI	G/AATT-C C-TTAA/G	5′-protruding
*Pst*I	C-TGCA/G G/ACGT-C	3′-protruding
*Not*I	GC/GGCC-GC CG-CCGG/CG	5′-protruding
*Eco*RII	/CC(A/T)GG GG(A/T)CC/	5′-protruding
*Sau*I	CC/TNA-GG GG-ANT/CC	5′-protruding
*Xmn*I	GAANN/NNTTC CTTNN/NNAAG	blunt end
*Sfi*I	GGCCN-NNN/NGGCC CCGGN/NNN-NCCGG	3′-protruding
*Bst*XI	CCAN-NNNN/NTGG GGTN/NNNN-NACC	3′-protruding
*Mnl*I	CCTCNNNNNNN/N GGAGNNNNNNN/N	blunt end
*Hga*I	GACGCNNNNN/NNNNN-N CTGCGNNNNN-NNNNN/N	5′-protruding
*Mbo*II	GAAGANNNNNNN-N/N CTTCTNNNNNNN/N-N	3′-protruding

Some class II restriction endonucleases are able to digest DNA · RNA hybrids. The enzymes *Alu*I, *Hae*III, *Hha*I, and *Taq*I cut the DNA strand of the hybrid at the correct recognition sequences [1552].

Sequence-specific digestion of *single-stranded* DNA has also been shown for a number of enzymes. However, like double-stranded DNA, base-paired recognition sequences are necessary for cleavage of single-stranded DNA. Therefore, refolding effects between different recognition sites, which result in transiently formed secondary canonical structures, play a decisive role. The cleavage rates depend mainly on the stability of these double-stranded structures [1553], [1554], [1380].

6.5.1.3.2. Nonpalindromic Recognition Sequences
In addition to class II enzymes that act on palindromic sequences, class II S enzymes like *Fok*I, *Hga*I, *Mbo*II, or *Mnl*I have been found, which recognize nonsymmetric sequences. These enzymes digest the double-stranded DNA at precise distances downstream from their recognition sites [1555], [1556], [1557], [1382].

The spatial separation between recognition and cleavage has been utilized to digest any predetermined

sequence in single-stranded DNAs like those of the various M13mp derivatives [1558], [1559]. Universal cleavage is achieved by applying a specifically designed *adapter*. This synthetic DNA element is constructed from a constant double-stranded domain harboring (1) the recognition sequence of the class II S enzymes and (2) a variable single-stranded domain complementary to the sequence to be cleaved. The linearized single-stranded target DNA can be converted to double-stranded DNA by a Klenow-catalyzed polymerization reaction using the synthetic deoxyoligonucleotide as a primer. By this approach, double-stranded DNA molecules are created that can be cleaved precisely at any position. Thus, the new method permits the design of novel enzyme specificities by combining class II S restriction endonucleases with bivalent DNA adapter oligonucleotides mediating the novel sequence specificities.

6.5.1.3.3. Isoschizomers

Isoschizomers are class II restriction endonucleases that are isolated from various organisms but recognize identical sequences. Isoschizomeric enzymes, however, may have various cleavage sites within the recognition sequence or may show different sensitivity to particular modified bases. The great variability of isoschizomers is known with 36 different enzymes all recognizing the sequence GATC [1396].

In this collection of isoschizomers, well-known examples for enzymes with different methylation sensitivities are *Dpn*I, *Mbo*I, and *Sau*3AI [1560], [1561], [1522]. Whereas *Sau*3AI is insensitive to methylation of the internal adenine residue, *Mbo*I is completely inhibited by this methylation. In contrast, *Dpn*I is fully dependent on this modification because the enzyme does not cleave the unmethylated GATC sequence. The DNA from pBR322, isolated from wild-type *Escherichia coli* cells, is highly methylated at adenosine residues within GATC sequences by *E. coli dam* methylase that specifically recognizes GATC [1562], [1362]. Consequently, *Sau*3AI and *Dpn*I will cleave this DNA, whereas *Mbo*I is unable to digest this substrate.

Another important pair of isoschizomers, *Hpa*II and *Msp*I, shows a different sensitivity with respect to cytosine methylation. Modification of the central cytosine residue within their common tetranucleotide recognition sequence CCGG renders the cleavage site resistant to digestion with *Hpa*II but sensitive to cleavage with *Msp*I [1420], [1563]. This ability to discriminate between methylated and unmethylated cytosine residues is used to detect methylated CpG dinucleotides which are thought to be involved in eukaryotic gene regulation [1564], [1565], [1375].

An example of an isoschizomeric enzyme pair with different cleavage sites is *Asp*718 and *Kpn*I. Both enzymes recognize the sequence GGTACC but cleave between either 5′-terminal guanosine residues or 3′-terminal cytosine residues [1566], [1360]. The 5′-terminal protruding ends of *Asp*718 fragments can be efficiently labeled with T4 polynucleotide kinase, whereas the 3′-ends are suitable substrates for 3′-endlabeling reactions with Klenow enzyme. Labeling of the 3′-protruding ends of *Kpn*I fragments can be obtained via the tailing reaction with terminal transferase.

The most commonly used cloning vehicles often have no restriction sites for particular enzymes. In such cases, cloning can be achieved by the use of enzymes from an enzyme family. Members of an enzyme family all produce identical single-stranded fragment ends, whereas the complete recognition sequences of individual enzymes are different in the flanking nucleotides. The best known enzyme family is the GATC family, whose members are *Sau*3AI, *Bgl*II, *Bam*HI, *Bcl*I, and *Xho*II [1396]. All these enzymes produce single-stranded ends of the sequence GATC. The respective hexanucleotide recognition sequences are lost after recombination of fragments produced by different enzymes of this family, but the common internal nucleotides remains the same in all possible combinations. In this way, for example, *Bgl*II fragments which are cloned into the *Bam*HI site of pBR322 DNA can be recovered from the vector by cutting with *Sau*3AI. Another enzyme family exists for the sequence CTAG with the enzymes *Mae*I, *Spe*I, *Nhe*I, *Avr*II, and *Xba*I [1396].

6.5.1.4. Changes in Sequence Specificity

Relaxation. For many enzymes such as *Eco*RI, relaxation of the specificity occurs when the solvent is changed [1493], [1567], [1568], [1569], [1358], [1377]. This altered activity of *Eco*RI is known as *Eco*RI * activity. Relaxed specificity markedly increases the number of small fragments, whereas the large fragments disappear. According to sequencing data, the subcanonical sequences differ from the canonical recognition sites by at least one base pair [1570]. Relaxation of both flanking positions yields the tetranucleotide AATT

as the predominant *Eco*RI * specificity. However, relaxation of internal positions has also been observed [1571].

Relaxation of the specificity of class II restriction endonucleases has been attributed generally to the effects of altered reaction environment on the nature of the complexes formed between the DNA helix and the *Eco*RI dimer. Because *Eco*RI * is an inherent enzyme property, it cannot be eliminated by more extensive purification. However, relaxation specificity can be suppressed by using suitable buffers. Relaxation can be prevented by the following conditions [1568], [1358], [1377]:

- high ionic strength,
- decreased pH,
- use of Mg^{2+} ions rather than Mn^{2+}, Co^{2+}, or Zn^{2+} ions,
- low enzyme concentration,
- low concentration of glycerol or other organic solvents such as dimethyl sulfoxide, which destabilize the double-stranded DNA helix, or
- short incubation periods.

In addition, incubation *temperature* strongly affects relaxation of specificity. As shown for *Pst*I, an increase in incubation temperature from 37 to 42 °C suppresses the relaxed specificity almost completely even at higher concentrations and longer incubation periods. In contrast, lowering the incubation temperature to 30 ° C strongly favors *Pst*I * activity. When the incubation temperature is further lowered to 25 or 20 °C, the relaxed activity becomes predominant even under normal incubation conditions. Analogous observations have been made for *Bam*HI, *Bst*EII, *Eco*RI, *Sph*I, *Sal*I, and *Taq*I [1572].

Modification of DNA by Methylation. Additional changes in the activity and specificity of class II restriction endonucleases are attributed to DNA modifications such as methylation [1398]. Certain enzymes are inhibited not only by the corresponding methylase counterpart but also by DNA methylation mediated by *Escherichia coli dam* [GATC] or *dcm*I methylase [$CC\binom{A}{T}GG$] [1573], [1574], [1362]. These two methylases are not part of the restriction – modification systems but have other biological functions such as the regulatory role of *dam* methylase in mismatch repair and DNA replication [1575], [1576], [1367].

When the recognition sequences of *dam* or *dcm*I methylase and the particular class II restriction endonuclease overlap totally, the activity on DNA isolated from *E. coli* is completely inhibited. Well-known examples of different *dam* influences at N^6-methylated adenine residues in GATC sequences are the isoschizomers *Dpn*I, *Mbo*I, and *Sau*3AI mentioned previously [1560], [1561], [1558]. Similar considerations apply to the *dcm*I methylase that inhibits *Eco*RII but activates *Apy*I by 5-methylation of the internal cytosine residue [1577], [1578]; however, *Bst*NI is not affected by this particular methylation [1578], [1447]. The inhibitory effects of *dam* or *dcm*I can be eliminated by isolation of the DNA from *E. coli* cells in which both methylase genes have been inactivated by mutation [1573], [1562], [1574].

Partial overlapping with the *dam* recognition sequence is observed for a variety of enzymes and frequently inhibits their activity [1580], [1581]. As an example, after *dam* methylation, *Cla*I digestion is inhibited specifically at the heptanucleotide sequences **GATC**GAT or ATC**GATC** [1579]. All the other *Cla*I sites are unaffected. Therefore, sequential action of isolated *dam* methylase and *Cla*I endonuclease generates the following new sequence specificity:

$$\binom{A}{G}ATCGAT\binom{T}{G}$$

In contrast, the heptanucleotides mentioned above are cleaved specifically with endonuclease *Dpn*I after methylation of DNA with *Cla*I methylase [1582]. In this approach, the other *Dpn*I cleavage sites remain

stable, and both enzyme pairs complement each other in creating new sequence specificities.

Permutation of Fragment Ends. Another approach for creating novel cleavage sites applies the permutation of generated fragment ends with 5′-protruding termini [1583]. Sites generated by such enzymes as *Eco*RI, *Bam*HI, or *Sal*I can be rendered blunt-ended by filling in the protruding ends with deoxynucleoside triphosphates and Klenow enzyme. This ligation creates a new symmetry center which corresponds to the ligation point of the two filled-in ends. If, for example, a *Sal*I site [G/TCGAC] is treated in this way, a 10 base pair palindromic sequence GT**CGATCG**AC with a novel *Pvu*I cleavage site [(CGAT/CG)] is created. After *Pvu*I cleavage, a third restriction site may be generated by treatment with nuclease S1. After ligation, another novel cleavage site is created because the resulting octanucleotide G**TCGCGA**C contains a new *Nru*I cleavage site [TCG/CGA]. Analogous conversion of restriction sites can be accomplished for all the other enzymes that generate fragments with 5′-protruding ends. Interconversion of restriction sites could generate unique novel cloning sites without the need of synthetic linkers. This should improve the flexibility of genetic engineering experiments.

Modification by Bacteriophages. Bacteriophage-induced DNA modification is also known to strongly influence the activity of class II restriction endonucleases. The respective phage genomes contain all or most of the cytosine or thymine residues substituted at the 5-position [1584], [1585]. Substitution of cytosine by 5-methylcytosine (*Xanthomonas oryzae* phage XP12) or glycosylated 5-hydroxymethylcytosine (*Escherichia coli* phage T4), or of thymine by phosphoglucuronated and glycosylated 5-(4′,5′-dihydroxypentyl)uracil (*Bacillus subtilis* phage SP15) renders the DNA resistant to almost all known class II restriction endonucleases. The only exception is *Taq*I, which fragments the modified DNAs extensively. Substitution of 5-hydroxyuracil or uracil for thymine (*B. subtilis* phages SPO1 or PBS1) results only in a decreased cleavage rate but not in complete inhibition of the various enzyme activities.

Endonuclease Inhibitors. Other inhibitors of class II restriction endonucleases are DNA-binding agents such as ethidium bromide, actinomycin D, proflavine, distamycin A, or neotrypsin [1586], [1587], [1588]. Spermine or spermidine also have an inhibitory effect at high concentrations, whereas low concentrations of these substances have a stimulatory effect. This stimulation is similar to that observed after addition of proteins such as *E. coli* RNA polymerase or T4 gene 32 protein which bind tightly to the single-stranded termini generated by many class II restriction endonucleases [1592], [1591], [1589], [1590].

6.5.1.5. Novel Class II Restriction Endonucleases

New sequence specificities are also obtained by screening bacteria for novel class II restriction endonucleases [1401]. The first method for finding new enzymes consists of examining those bacterial strains that are resistant to certain bacteriophages. As an alternative procedure, representatives of previously unexamined families from the wide range of eu- and archaebacteria are screened systematically for the presence of new class II restriction endonucleases. For example, in a total of 348 microorganisms screened, 105 novel enzymes were found, predominantly in gram-negative bacteria [1593]. To detect the novel specificities, aliquots of bacterial extracts are incubated with a DNA of known sequence such as λ DNA for various periods under standard incubation conditions. The fragment mixtures are analyzed by electrophoresis on agarose gels.

The new sequence specificity is determined with computer assistance by correlating experimentally observed and predicted fragment patterns. In a following step, the corresponding fragments obtained with other DNAs, e.g., from pBR322 or SV40, are analyzed. In a third step, the precise cleavage positions with the recognition sequence are fixed in sequencing gels by accurate fragment length determination of both strands of a particular DNA fragment created with the novel enzyme.

The three steps in characterization of the specificity and cleavage positions of a new enzyme may be examplified by the class II restriction endonuclease *Mae*I. This new enzyme can be isolated from the archaebacterium *Methanococcus aeolicus* PL15/H, along with *Mae*II and *Mae*III which are also new enzymes with novel specificities [1527]. The enzyme *Mae*I recognizes the tetranucleotide sequence C/TAG and cleaves both DNA strands specifically between the 5′-flanking cytosine and thymine nucleotides. The enzyme *Mae*II also recognizes a tetranucleotide sequence [A/CGT], whereas *Mae*III acts on the pentanucleotide sequence /GTNAC.

Novel class II restriction endonucleases with more complex recognition sequences have also been found. Examples are *Dra*II and *Dra*III which were isolated recently as minor species from *Deinococcus radiophilus* in addition to the already known main activity *Dra*I, an isoschizomer of *Aha*III [1373]. The enzyme *Dra*II recognizes a novel type of heptanucleotide with ambiguities in the flanking nucleotides:

$$\binom{G}{A}G/GNCC\binom{G}{T}$$

The enzyme *Dra*III is characterized by the nonanucleotide CACNNN/GTG with three internal undefined nucleotides [1548], [1549].

6.5.2. DNA Polymerases

Table 47 lists the properties of important DNA polymerases under A.

6.5.2.1. Escherichia coli DNA Polymerase I

The enzyme *E. coli* polymerase I, also called deoxynucleoside triphosphate: DNA deoxynucleotidyltransferase (DNA-directed) (E.C. 2.7.7.7) [9012-90-2], is obtained from *Escherichia coli.*

Properties. *Escherichia coli* DNA poylmerase I consists of a single polypeptide with a molecular mass of 109×10^3 dalton [1595] and contains one Zn^{2+} ion per molecule [1594]. The enzyme consists of two domains joined by a protease-sensitive peptide linker [1596], [1597], although the complete protein is approximately spherical with a diameter of 6500 pm [1598], [1599]. *Escherichia coli* DNA polymerase I contains a single free mercapto group able to form a dissociable complex with Hg^{2+}. The protein can be carboxymethylated without loss of activity [1600].

One unit of *E. coli* DNA polymerase I is defined as the enzyme activity that incorporates 10 nmol of total nucleotide into acid-precipitable material in 30 min at 37 °C [1601].

Uses. *Escherichia coli* DNA polymerase I is used predominantly for the in vitro labeling of DNA by *nick translation* [1392]. Several published procedures all follow a common principle (Fig. 100). Trace amounts of DNase I (see Section DNase I) introduce nicks into unlabeled DNA, thus exposing internal 3′-hydroxyl groups. These nicks move toward the 3′-terminus by the subsequent sequential action of *E. coli* DNA polymerase I. First, the enzyme removes the nucleotide on the 5′-side of the nick. In the coupled reaction that follows, the excised nucleotide is replaced with a labeled nucleotide at the 3′-terminus of the nick. Repeated action of *E. coli* polymerase I replaces all nucleotides with labeled nucleotides downstream from the nick. Below 22 °C, the reaction will not proceed further than one complete exchange of the existing unlabeled DNA strand for a labeled one, as shown in Figure 100. At the right ratio of DNA, DNase I, and *E. coli* DNA polymerase I, about 20 – 40 % of the labeled nucleotides are incorporated into the product, with a yield in label density of 10^8 – 10^9 dpm per microgram of DNA. Only 20 – 30 % of this incorporated label is commonly found in foldback DNA [1602]. The amount of foldback DNA is often higher if the unlabeled DNA fragment is relatively small (<1000 nucleotides) [1603]. In this case, more 3′-ends are probably used for DNA synthesis than internal 3′-termini at nicks.

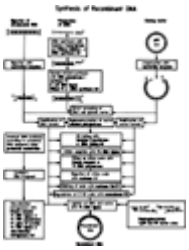

Figure 100. Use of DNA- and RNA-modifying enzymes in recombinant DNA synthesis

For abbreviations, see text
[Full View]

6.5.2.2. Klenow Enzyme

Klenow enzyme is the largest fragment obtained by proteolysis of *Escherichia coli* DNA polymerase I with

subtilisin; it is also called deoxynucleoside triphosphate: DNA deoxynucleotidyltransferase (DNA-directed) (E.C. 2.7.7.7) [9012-90-2].

Properties. Klenow enzyme has a molecular mass of 75×10^3 dalton. It can be obtained by subtilisin treatment, which digests the protease-sensitive peptide linker between the two functional domains of the native enzyme [1604]. The large fragment carries the 5′ – 3′ polymerase and the 3′ – 5′ exonuclease activities of intact DNA polymerase I, but lacks the 5′ – 3′ exonuclease activity located on the small fragment of 36×10^3 dalton. Klenow enzyme catalyzes the addition of deoxynucleotides to the 3′-hydroxyl terminus of a primer annealed to the template DNA.

One unit of Klenow enzyme is defined as the enzyme activity that incorporates 10 nmol of total nucleotide into an acid-precipitable fraction in 30 min at 37 °C [1601].

Uses. Klenow enzyme is used in a wide variety of techniques, such as the following:

1. Conversion of 3′-recessed ends of restricted DNA fragments to blunt ends by the fill-in reaction [1605];
2. 3′-Endlabeling of DNA fragments by use of α-^{32}P-deoxynucleotides [1605];
3. Second-strand synthesis of cDNA [1606];
4. Homogeneous labeling of DNA fragments with random oligodeoxynucleotides as primers (oligolabeling) [1608];
5. Sequencing of DNA by the dideoxy chain termination method [1607] (Fig. 101);
6. Elongation of oligonucleotides mediating mismatch formation for site-directed mutagenesis [1609].

Figure 101. Use of RNA- and DNA-modifying enzymes in recombinant DNA analysis

For abbreviations, see text.
[Full View]

6.5.2.3. T4 DNA Polymerase

The enzyme T4 DNA polymerase, also called deoxynucleoside triphosphate: DNA deoxynucleotidyltransferase (DNA-directed) (E.C. 2.7.7.7) [9012-90-2], is obtained from phage T4-infected *Escherichia coli*.

Properties. The enzyme T4 DNA polymerase consists of a single polypeptide with a molecular mass of 114×10^3 dalton [1610]. This polymerase displays a relatively broad pH optimum ranging from 8 to 9; at pH values of 7.5 and 9.7, ca. 50 % of the optimal activity is observed. Maximal polymerase activity requires 6 mmol/L of Mg^{2+}; this activity is decreased to one-fourth if Mg^{2+} is replaced by Mn^{2+} at its optimal concentration of 0.1 mmol/L.

One unit of T4 DNA polymerase is defined as the enzyme activity that incorporates 10 nmol of ^{3}H-dTTP into acid-precipitable DNA products in 30 min at 37 °C [1610].

The T4 DNA polymerase catalyzes the addition of mononucleotides onto the 3′-hydroxyl terminus of a primer annealed to a single-stranded region of a DNA template. Because fully duplex DNA cannot serve as a template primer, DNA must first be made partially single-stranded by digestion of the 3′-termini with *E. coli* exonuclease III (see Section Exonuclease III) [1601].

Single-stranded DNA can also serve as a template primer for T4 DNA polymerase. This polymerase is unable to use a DNA duplex that contains a nick as template primer. However, addition of the T4 gene 32 protein facilitates strand displacement, which allows the T4 DNA polymerase to replicate even the nicked

duplex DNA in vivo and in vitro [1611], [1612]. Although T4 DNA polymerase lacks 5′ – 3′ exonuclease activity, it contains an extremely active 3′ – 5′ exonuclease that shows a strong specificity for single-stranded DNA. Its turnover number is about 250 times greater than that of the 3′ – 5′ exonuclease associated with *E. coli* DNA polymerase I and about 3 times greater than the turnover number of its own 5′ – 3′ polymerase activity [1613].

Uses. The enzyme T4 DNA polymerase can act as either a 5′ – 3′ polymerase or a 3′ – 5′ exonuclease. Polymerization occurs when all three of the following substrates are available to the enzyme: a polynucleotide template, a 3′-hydroxyl primer at least one residue shorter than the template, and the appropriate dNTPs or NTPs complementary to the template. In the absence of any one of these three components, the enzyme functions as an 3′ – 5′ exonuclease.

The T4 DNA polymerase can be used to label blunt or recessed 3′-termini of DNA by its 5′ – 3′ polymerase activity. In the case of blunt ends, the 3′ – 5′ exonuclease activity of the enzyme is responsible for digestion of duplex DNA to produce molecules with recessed 3′-termini. On subsequent addition of labeled dNTPs, the partially digested DNA molecules serve as primer –templates that are regenerated by the polymerase into intact, double-stranded DNA. Molecules labeled to high specific activity by this technique may be used as sensitive hybridization probes. They have two advantages over probes prepared by nick translation. First, they lack the artifactual hairpin structures that may be produced during nick translation. Second, they can be converted easily into strand-specific probes by cleavage with suitable restriction endonucleases in the unlabeled central region of the DNA molecule [1614]. In combination with T4 gene 32 protein, the T4 DNA polymerase is used for gap filling in site-directed mutagenesis experiments in which short oligonucleotides are employed for mismatch formation [1615].

6.5.2.4. Reverse Transcriptase

The enzyme AMV reverse transcriptase, also called deoxynucleoside triphosphate: DNA deoxynucleotidyltransferase (DNA-directed) (E.C. 2.7.7.49), is obtained from avian myeloblastosis virus.

Properties. The enzyme AMV reverse transcriptase is one of the gene products of the RNA genome of avian myeloblastosis virus existing in two copies within the core structure of the retrovirus particle. Retroviruses which are capable of replicating without helper viruses have at least three genes arranged in the following order:

1. *gag* (encoding structural proteins of the inner coat and core),
2. *pol* (encoding reverse transcriptase and DNA endonuclease), and
3. *env* (encoding glycoproteins of the envelope).

A *gag – pol* precursor protein, $p180^{gag–pol}$, is one of the primary protein products; it is processed into a smaller protein, $p130^{gag–pol}$ [1616], [1617]. Further proteolysis releases two polypeptides designated as α and β, the latter being phosphorylated [1618]. The enzymatically active forms of reverse transcriptase exist as α, β β, and α β. The molecular mass of the α-subunit is 68×10^3 dalton and of the β-subunit 92×10^3 dalton.

One unit of reverse transcriptase is defined as the enzyme activity that incorporates 1 nmol of ^{3}H-dTMP into acid-precipitable products in 10 min at 37 °C with poly(dA) · $p(T)_{15}$ as template primer [1619].

Because the amino acid sequence of α is a subset of the β-subunit, the mature form α β is considered to be formed by proteolytic cleavage of the β β dimer into α β and p32, containing DNA endonuclease activity [1620]. Enzymatic activities associated with α β are (1) RNA-dependent DNA polymerase, (2) DNA-dependent DNA polymerase, (3) RNase H, and (4) an unwinding activity normally attributed to single-strand binding proteins [1621]. The α-subunit contains both the polymerase and the RNase H activity which has the unique ability to degrade the RNA in DNA · RNA hybrids in an exonucleolytic manner [1622], [1623]. This exonucleolytic activity of RNase H can proceed from either the 5′- or the 3′-terminus. However, it cannot act on RNA linkages in covalently closed duplex circles. Reverse transcriptase was

shown to synthesize poly(dT) transcripts in vitro in a progressive way with poly(rA)$_{1100}$ · oligo(dT)$_{12–18}$ as template and primer [1619]. In addition, the capacity of reverse transcriptase to use poly(rC) · oligo(dG)$_{12}$ as substrate is a useful feature for discrimination between the viral polymerase activity and the various host DNA polymerases [1624], [1625]. Effective priming is observed if the primer is more than eight nucleotides in length [1626].

Uses. The enzyme AMV reserve transcriptase is widely used for synthesis of cDNA transcripts of specific RNA sequences in vitro (Fig. 102). Synthesis of the primary DNA strand is catalyzed by the RNA-dependent DNA polymerase activity. The DNA-dependent DNA polymerase activity is responsible for second-strand synthesis in cDNA formation when reverse transcriptase is used as polymerizing agent. The DNA-dependent DNA polymerase activity can be inhibited by addition of actinomycin D [1627]. Transcripts of cDNA are employed for analysis of structure, organization, and expression of eukaryotic genes. Comparison between cDNA and genomic DNA sequences elucidates genomic rearrangements, the existence of intervening sequences, and details of the splicing events [1628], [1629], [1630]. Synthesis of cDNA is also a powerful tool for the isolation of functional DNA sequences and the generation of corresponding hybridization probes. More recently, cloning of cDNA copies of the RNA virus genomes has opened another way to study details of their genomic structure [1631], [1632], [1633]. Several protocols for cDNA synthesis have been described [1634], and typical procedures have been reviewed recently [1635], [1637], [1638]. In addition, new methods have been developed for highly efficient cloning of full-length cDNA and sequencing of RNA [1639] by using AMV reverse transcriptase [1640].

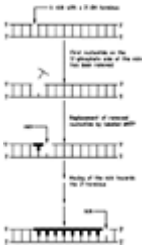

Figure 102. Sequential action of DNase I and *Escherichia coli* DNA polymerase I during nick translation
[Full View]

6.5.2.5. Terminal Transferase

Terminal transferase, also called nucleoside triphosphate: DNA deoxynucleotidylexotransferase (E.C. 2.7.7.31) [9027-67-2], is obtained from calf thymus.

Properties. Terminal transferase from calf thymus catalyzes the addition of deoxynucleotides to the 3′-hydroxyl terminus of a DNA primer [1641].

One unit of terminal transferase is defined as the enzyme activity that incorporates 1 nmol of dAMP into acid-insoluble products in 60 min at 37 °C with d(pT)$_6$ as primer [1642].

Terminal transferase from calf thymus has a molecular mass of 32×10^3 dalton. Under denaturing conditions, two nonidentical subunit structures are observed with molecular masses of 26.5×10^3 and 8×10^3 dalton. The isoelectric point is at pH 8.6, and the pH optimum at 7.2. At 35 °C, terminal transferase has an absolute requirement for an oligodeoxynucleotide primer containing at least three deoxynucleotides and a free 3′-hydroxyl end. However, the reaction is not template-dependent. Any dNTP, including substituted deoxynucleotides, can be polymerized by the enzyme [1643]. The highest polymerization rate for purine deoxyribonucleoside triphosphates is obtained when cacodylate buffer and Mg^{2+} are used, whereas the use of Co^{2+} instead of Mg^{2+} increases the polymerization of pyrimidine bases [1644].

Uses. Because terminal transferase requires single-stranded DNA as primer, the incorporation efficiency is highest for dsDNA with 3′-protruding ends. However, blunt ends and 3′-recessive ends are also tailed by reverse transcriptase, with reduced efficiency [1645], [1633]. Terminal transferase is used mainly to add homopolymer tails to DNA fragments for the construction of recombinant DNA. By use of this tailing method, any double-stranded DNA fragment can be joined to a cloning vehicle. Optimal lengths for G–C

tails are about 20 – 25 nucleotides, for A–T tails about 100 nucleotides [1392]. In addition to the tailing reaction, terminal transferase is used for 3′-end-labeling of DNA fragments with α-^{32}P-labeled dNTPs or ddNTPs [1646], [1647] and for the addition of a single nucleotide to the 3′-ends of DNA for in vitro mutagenesis experiments [1645].

6.5.3. RNA Polymerases

Table 47 lists important RNA polymerases under B.

6.5.3.1. SP6 RNA Polymerase

The enzyme SP6 RNA polymerase, also called nucleoside triphosphate: RNA nucleotidyltransferase (DNA-directed) (E.C. 2.7.7.6) [9014-24-8], is obtained from phage SP6-infected *Salmonella typhimurium* LT2.

Properties. The enzyme SP6 RNA polymerase consists of a single polypetide chain of 96×10^3 dalton [1648].

One unit of SP6 RNA polymerase is defined as the enzyme activity that catalyzes the incorporation of 1 nmol of GTP into acid-precipitable RNA products in 60 min at 37 °C [1648].

The SP6 RNA polymerase requires a DNA template and Mg^{2+} as a cofactor for RNA synthesis, and is strongly stimulated by bovine serum albumin or spermidine [1648]. The enzyme is not inhibited by the antibiotic rifampicin. This inhibitor affects bacterial RNA polymerases but not bacteriophage-coded RNA polymerases such as SP6 RNA polymerase [1648], [1649]. SP6 RNA polymerase possesses a very stringent promoter specificity which is distinct from that of analogous enzymes induced by bacteriophages T7, T3, gh1, or other T7-like bacteriophages [1650], [1651], [1652]. Thus SP6 RNA polymerase is able to transcribe only bacteriophage SP6 DNA [1653] or DNA cloned downstream from SP6 promoter as, for example, in the plasmids pSP64 or pSP65 [1654].

Uses. The enzyme SP6 RNA polymerase is used for specific in vitro synthesis of RNA transcripts obtained from distinct DNA sequences cloned into the multiple cloning site of the vectors pSP64 or pSP65 [1654]. In these vectors, the SP6 promoter located in front of the multiple cloning site directs specific initiation of the transcription reaction. Restriction enzyme cleavage of the vector at a unique restriction site downstream from the inserted DNA sequences provides a discrete terminus for the run-off RNA transcript. Sense or anti-sense RNA transcripts of unique length can be generated, depending on the orientation of the integrated DNA (Fig. 103)

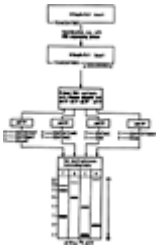

Figure 103. DNA sequencing with the dideoxy chain-termination method
[Full View]

Splicing Analysis. When this vector system is applied, SP6 RNA polymerase can be used to synthesize large amounts of precursor RNA for in vitro and in vivo studies of the splicing reaction [1655], [1656], [1654].

RNA Probes. The enzyme SP6 RNA polymerase is also useful for the generation of homogeneously labeled, single-stranded RNA probes of highest specific radioactivity. These probes may be applied in Northern- and Southern-blot hybridization techniques [1657]. After DNase treatment and removal of the nonincorporated nucleotides by ethanol precipitation, the RNA probes generated with the pSP64/65 vectors are ready for hybridization without prior purification by gel electrophoresis. A tenfold increase in sensitivity may be observed when RNA probes are used instead of nick-translated DNA probes [1654].

Genomic DNA Sequencing. Radioactively labeled RNA probes derived from SP6 RNA polymerase are also applied in genomic DNA sequencing [1658] and in situ hybridization experiments.

Gene Mapping. M13 libraries may be screened with RNA probes synthesized by SP6 RNA polymerase in a one-step procedure that identifies the orientation of cloned DNA fragments.

Synthesis of Specific RNA. An additional application of SP6 RNA polymerase is the synthesis of specific RNA strands for use in the RNase protection technique. After hybridization to total cell RNA, nonhomologous sequences are digested with a mixture of RNase A and RNase T_1. Only the homologous double-stranded regions will give stable hybrids of characteristic length [1657].

6.5.3.2. T7 RNA Polymerase
The enzyme T7 RNA polymerase, also called nucleoside triphosphate: RNA nucleotidyltransferase (DNA-directed) (E.C. 2.7.7.6) [9014-24-8], is obtained from phage T7-infected *Escherichia coli*.

Properties. The enzyme T7 RNA polymerase consists of a single polypeptide chain of 98×10^3 dalton [1659].

One unit of T7 RNA polymerase is defined as the enzyme activity that incorporates 1 nmol of GMP into acid-precipitable RNA products in 60 min at 37 °C [1660].

The enzyme T7 RNA polymerase is highly specific for a T7 promoter (17 base pairs) including the transcription start site [1661]. The promoter sequence differs from that of bacteriophage SP6-coded RNA polymerase (Section *Escherichia coli* DNA Polymerase I) by only two nucleotides. Although the sequences are similar, T7 RNA polymerase specifically transcribes only T7 DNA or DNA cloned downstream of the T7 promoter. Appropriate cloning vectors are pT7–1 and pT7–2 containing the strong T7 promoter **Φ**10 in front of a polylinker region with eight unique restriction sites [1662]. In pT7–1 and pT7–2, the polylinker regions are integrated in opposite orientation. In addition to a DNA template with a T7 promoter, T7 RNA polymerase requires Mg^{2+} as cofactor and the four NTPs as substrates. Spermidine or bovine serum albumin stimulates the activity of the enzyme. In contrast to bacterial RNA polymerases, T7 RNA polymerase is not inhibited by rifampicin [1648], [1649].

Uses. The T7 transcription system consisting of T7 RNA polymerase and vectors pT7–1 and pT7–2 is used for in vitro synthesis of specific RNA. For this purpose, the template DNAs are cloned into the polylinker regions of these vectors downstream from the T7 promoter. By cutting the cloned DNA at the 3′-end with an appropriate restriction enzyme before the transcription reaction has started, run-off RNA transcripts of defined size are obtained. By using the T7 vectors pT7–1 and pT7–2 with oppositely oriented polylinker regions, DNA can be transcribed into sense and anti-sense mRNA.

The T7 system allows in vitro synthesis of either homogeneously labeled (in the presence of radioactive or biotinylated NTPs) or unlabeled RNA molecules. Since the ratio of the produced RNA to template DNA reaches 30 : 1, elimination of the DNA is usually not necessary but can be achieved by treatment with RNase-free DNase I. The RNA probes are highly specific for the complementary DNA strand; therefore, they are especially useful for different RNA or DNA blotting techniques [1663], [1664], in situ hybridization [1665], genomic sequencing [1658], or nuclease S1 studies. This property also allows the screening of M13 gene banks for clones containing DNA with a defined orientation. With unlabeled primary mRNA transcripts, splicing of RNA may be studied in vivo or in vitro. With the aid of in vitro synthesized anti-sense mRNA, expression of the corresponding genes can be suppressed after introduction of this RNA into eukaryotic cells [1666]. The T7 system also allows the synthesis of pure RNA transcripts for in vitro translation studies.

6.5.4. DNA Nucleases
Table 47 lists important DNA nucleases under C.

6.5.4.1. DNase I
The enzyme DNase I, also called deoxyribonuclease I (E.C. 3.1.21.1), is obtained from bovine pancreas.

Properties. DNase I from bovine pancreas is a glycoprotein with a molecular mass of 31×10^3 dalton. The pH optimum of the enzyme is 7.0. One unit of DNase I is defined as the enzyme activity that produces after incubation of calf thymus DNA an absorbance increase of 0.001 in one minute at 25 °C [1667], [1668] .

DNase I is normally isolated as a mixture of four isoenzymes: A, B, C, and D. The complete amino acid sequence of isoenzyme A, which is the main component, has been determined [1669]. Isoenzymes A and B differ only in the composition of the carbohydrate side chain attached to asparagine 18, but isoenzymes C and D have the histidine at position 118 replaced by a proline [1670]. All DNase I isoenzymes contain two disulfide bridges, one of which can be reduced in the presence of Ca^{2+} without loss of activity [1671]. DNase I degrades double-stranded DNA in an endonucleolytic way to yield 5′-oligonucleotides [1672], [1673]. DNase I requires divalent metal ions for DNA hydrolysis, with maximal enzymatic activity in the presence of Ca^{2+} and Mg^{2+} or Ca^{2+} and Mn^{2+} [1674]. The mode of action and the specificity of the enzyme depend on the type of divalent cations present [1675], [1676]. DNase I appears to be primarily a digestive enzyme. However, DNase I-like activities have also been found in nondigestive tissues, which suggests that the enzyme might have other functions [1673]. It may play a role in the regulation of actin polymerization, because it forms a very stable complex with monomeric actin and causes depolymerization of filamentous actin [1455], [1677], [1678].

Uses. DNase I is used for labeling dsDNA by nick translation [1392]. RNase-free DNase I is also applied to template digestion after in vitro synthesis of RNA with bacteriophage SP6-, T7-, or T3-coded RNA polymerases [1648].

6.5.4.2. Exonuclease III

Exonuclease III, also called exodeoxyribonuclease III (E.C. 3.1.11.2), is obtained from *Escherichia coli*.

Properties. Exonuclease III is a monomeric protein with a molecular mass of 28×10^3 dalton [1679].

One unit of exonuclease III is defined as the enzyme activity that releases 1 nmol of acid-soluble nucleotides from calf thymus DNA treated with ultrasonic waves in 30 min at 37 °C [1680].

The enzyme is a multifunctional protein which catalyzes the hydrolysis of several types of phosphoester bonds in dsDNA. In addition, the following minor activities are found [1681], [1682], [1683]:

an endonuclease activity, specific for apurinic and apyrimidinic sites,
a 3′-phosphatase activity, and
an RNase H activity, which degrades the RNA strand in DNA · RNA hybrids.

The pH optima for the various endo- and exonucleolytic activities are between 7.6 and 8.5. A lower pH optimum between 6.8 and 7.4 has been reported for phosphatase activity. For optimal exonucleolytic activity, Mg^{2+} or Mn^{2+} is required, whereas Zn^{2+} inhibits exonuclease III [1601], [1683]. In addition, the activity depends strongly on temperature, salt concentration, and ratio of DNA to enzyme concentration [1684], [1685]. Also, the rate at which mononucleotides are released decreases in a base-dependent mode in the following order $C \gg A \sim T \gg G$ [1686].

Uses. Exonuclease III is used mainly for the generation of single-stranded terminal regions from dsDNA by acting as a 3′ – 5′ exonuclease. The DNA modified in this way is used as a substrate for labeling DNA with Klenow enzyme or for sequencing studies [1687], [1684], [1685]. Another application of exonuclease III is to shorten progressively both strands of dsDNA by additional digestion of exonuclease III-generated single-stranded tails with nuclease S1 [1688]. Exonuclease is also used to determine stable lesions in defined sequences of DNA [1689].

6.5.4.3. Nuclease S1

Nuclease S1, also called *Aspergillus* nuclease S1 (E.C. 3.1.30.1), is obtained from *Aspergillus oryzae.*

Properties. Nuclease S1 is a monomeric protein with a molecular mass of 38×10^3 dalton [1690], [1691]. The

enzyme is a glycoprotein containing 18 % carbohydrate residues [1690]. Its isoelectric point is at pH 4.3 – 4.4 [1692].

One unit of nuclease S1 is defined as the enzyme activity that releases 1 µg of acid-soluble deoxynucleotides from denatured DNA in 1 min at 37 °C [1691].

The pH optimum is 4.0 – 4.3, with half-maximal rates at pH 3.3 and 4.9. Higher pH values (4.6 – 5.0) have been used to avoid possible nicking of DNA substrates by depurination. The enzyme requires Zn^{2+} for optimal activity. The ions Co^{2+} and Hg^{2+} can replace Zn^{2+}, but they are less effective [1691]. Nuclease S1 is optimally active at NaCl concentrations of 0.1 mol/L and is relatively unaffected by ionic strengths between 0.01 and 0.2 mol/L of NaCl [1693]. The enzyme is remarkably stable to such denaturing agents as urea, sodium dodecyl sulfate, or formamide [1694]. When the enzyme is used in high concentrations, the salt concentration must be high enough (>0.2 mol/L) to suppress nonspecific nicking [1695].

Uses. Under optimal conditions, the rates of nuclease S1-catalyzed hydrolysis of single- and double-stranded nucleic acids have been estimated to differ by a factor of 75 000 [1696]. Thus, the main application of nuclease S1 is to trim single-stranded protruding ends of DNA or RNA without significant nibbling of the duplex ends [1697], [1698]. Mapping of transcripts is performed by analyzing S1-resistant RNA · DNA hybrids in alkaline agarose or denaturing polyacrylamide gels [1699], [1700]. High-resolution mapping is obtained after end-labeling the DNA with ^{32}P by analyzing the S1-resistant DNA fragments on sequencing gels [1702]. Use of this method allows mapping of the 5′- and 3′-ends of mRNA [1703]. In contrast to exonuclease VII, single-stranded loops in dsDNA or RNA · DNA hybrids are digested by nuclease S1 [1704], [1705]. Thus, mapping of introns in eukaryotic genes is possible by digesting the hybrid of mature mRNA and the ^{32}P-labeled codogenic DNA strand with nuclease S1 (Fig. 104).

Figure 104. Use of AMV reverse transcriptase for in vitro synthesis of cDNA [Full View]

A further application that takes advantage of this property is digestion of hairpin structures during cDNA synthesis [1635]. An S1 mapping based on analysis by two-dimensional gel electrophoresis is also described [1706]. The methodology of S1 mapping has been used successfully to study the regulation of cloned genes reintroduced into eukaryotic cells [1707].

6.5.4.4. Nuclease Bal 31

Nuclease Bal 31, also called exodeoxyribonuclease (E.C. 3.1.11), is obtained from *Alteromonas espejiani*.

Properties. Nuclease Bal 31 progressively degrades both strands of dsDNA from the ends inward. The enzyme also has a ssDNA-specific endonuclease activity which cleaves supercoiled DNA to yield, for example, the linear form. No dsDNA-specific endonuclease activity is associated with the enzyme.

One unit of nuclease Bal 31 is defined as the enzyme activity that releases 600 base pairs in 10 min at 30 °C from 2 µg of linearized pUR222 DNA [1708].

Enzyme activity requires Ca^{2+} and Mg^{2+}, and is optimal at an NaCl concentration of 0.6 mol/L. Nuclease Bal 31 is not inhibited by sodium dodecyl sulfate or urea. Fragments of DNA shortened with nuclease Bal 31 can be ligated readily with T4 DNA ligase either directly or after filling in any single-stranded end with Klenow enzyme.

Uses. In molecular cloning experiments, various specific manipulations of DNA sequences must be

accomplished: for example, coding sequences must be joined to a promoter in the proper reading frame, deletions of predetermined DNA segments are required to map functional regions, or transcription units must be localized by trimming to the minimum essential size. If suitable restriction sites do not exist at the proper location on the DNA sequence, nuclease Bal 31 can be applied.

Nuclease Bal 31 can also be used for mapping DNA with restriction endonucleases [1709]. The DNA segment to be mapped is shortened to various lengths with the nuclease. Each sample is then digested with the corresponding restriction endonuclease, and the resulting fragments are analyzed by gel electrophoresis. Along with shortening of the original DNA segment, restriction fragments disappear from the gel in exactly the same order in which the restriction sites are arranged along the DNA strand.

Another use of nuclease Bal 31 is to transfer any restriction site located adjacent to a preselected position [1710]. A linearized plasmid containing a cloned DNA sequence is digested with nuclease Bal 31 to a predetermined length, cut with a single cutting restriction enzyme, circularized with T4 DNA ligase, and transformed into *Escherichia coli* cells. Clones are selected which are sensitive to digestion with the restriction enzyme used before.

The ssDNA-specific endonuclease activity of nuclease Bal 31 has also been exploited to investigate the secondary structure in supercoiled DNAs or an altered helical structure in nonsupercoiled DNA produced by carcinogenic or mutagenic agents [1454].

6.5.5. RNA Nucleases

Table 47 lists important RNA nucleases under D.

6.5.5.1. RNase H

The enzyme RNase H, also called endoribonuclease (E.C. 3.1.26.4), is obtained from *Escherichia coli.*

Properties. The enzyme activity found in *E. coli* which hydrolyzes the RNA of RNA · DNA hybrids is designated RNase H, whereas the activity against the RNA in RNA · RNA duplexes is called RNase III [1711].

One unit of RNase H is the enzyme activity that produces 1 nmol of acid-soluble ribonucleotides from ^{3}H-poly(A) · poly(dT) in 20 min at 37 °C [1712].

RNase H acts as an endoribonuclease and degrades the RNA strand of RNA · DNA hybrids of natural origin, and of synthetic complexes like poly(A) · poly(dT). RNase H produces ribonucleotides with 5′-phosphate and 3′-hydroxyl termini. Almost no activity is detected with polyribonucleotides alone or polymers annealed to their complementary ribopolymer. For optimal activity, RNase H requires Mg^{2+} which can be replaced only partially by Mn^{2+}. The enzyme has its maximum activity in the presence of compounds containing mercapto groups and is inhibited by *N*-ethylmaleimide. The pH optimum is between 7.5 and 9.1. RNase H activity is relatively insensitive to salt; 50 % of its activity is retained in the presence of NaCl at 0.3 mol/L [1713].

Uses. In addition to the use of *E. coli* RNase H for the study of in vivo functions such as RNA-primed initiation of DNA synthesis [1712], RNase H is applied in the synthesis of cDNA by combination of the classical first-strand synthesis with the novel second-strand synthesis mediated by *E. coli* DNA polymerase I, *E. coli* RNase H, and *E. coli* DNA ligase [1714], [1640]. Furthermore, RNase H is used to detect RNA · DNA regions in dsDNA of natural origin [1715], [1623]. A further application of the enzyme is the removal of poly(A) sequences of mRNA, which leads to increased electrophoretic homogeneity of mRNA in gel electrophoresis [1716]. RNase H may also be used for the site-specific enzymatic cleavage of RNA. With this method, a synthetic DNA oligomer will hybridize only to complementary single-stranded regions of an RNA molecule, which are therefore digested by RNase H in a site-specific manner [1717].

6.5.5.2. Site-Specific RNases

6.5.5.2.1. RNase A

The enzyme RNase A, also called pancreatic ribonuclease, (E.C. 3.1.27.5), is obtained from bovine pancreas.

Properties. RNase A is an endonuclease that cleaves RNA but not DNA. The enzyme specifically attacks pyrimidine nucleotides by cleaving the 3′-adjacent phosphodiester bond Py/pN. The molecular mass is 13.7×10^3 dalton and the pH optimum is 7.0 – 7.5 [1783], [1718]. Cyclic 2′,3′-pyrimidine nucleotides are obtained as intermediates. Pyrimidine 3′-phosphates and oligonucleotides with terminal pyrimidine 3′-phosphate groups are the final products.

One unit of RNase A is defined as the enzyme activity that produces total conversion of RNA as substrate in 1 min at 25 °C. The decrease in absorbance A_0 to A_1 corresponds to total conversion, A_1 being the final absorbance [1719]. RNase A is used as A- and G-specific RNase in RNA sequencing [1720], [1722].

6.5.5.2.2. RNase CL3

The enzyme RNase CL3, also called endoribonuclease (E.C. 3.1.27.1), is obtained from chicken liver.

RNase CL3 cleaves RNA predominantly at Cp/N bonds and produces fragments with 3′-terminal cytidine phosphate. Bonds of the type Ap/N and Gp/N are cleaved less frequently, and Up/N bonds are cleaved to a very small extent [1720]. The molecular mass is 16.85×10^3 dalton; the pH optimum is at 6.5 – 7.5 [1721].

One unit of RNase CL3 is defined as the enzyme activity that releases a sufficient quantity of acid-soluble oligonucleotides to produce an absorbance increase of 1.0 at 260 nm in 15 min at 37 °C [1721]. One unit corresponds to the decomposition of ca. 40 µg of poly(C).

RNase CL3 is used as minus-U-specific RNase in RNA sequencing [1720], [1722]. The Cp/N cleavage is more efficient in the presence of urea and at 50 °C, i.e., under conditions in which the secondary structure of RNA is destroyed [1722], [1723]. Under these denaturing conditions, approximately five to ten times as much enzyme must be used, but predominantly C-specific cleavage is obtained.

6.5.5.2.3. RNase T_1

Ribonuclease T_1 (E.C. 3.1.27.3), also called endoribonuclease, is obtained from *Aspergillus oryzae*.

RNase T_1 is an endonuclease that cleaves RNA but not DNA. The enzyme specifically attacks the 3′-adjacent phosphodiester bond Gp/N. RNase T_1 is therefore used as G-specific RNase in RNA sequencing [1720], [1722]. The molecular mass is 11.1×10^3 dalton, the pH optimum is at 7.4 [1724], [1725]. Cyclic 2′,3′-guanosine nucleotides are obtained as intermediates, and guanosine 3′-phosphates and oligonucleotides with guanosine 3′-phosphate terminal groups are the final products.

One unit of RNase T_1 is defined as the enzyme activity that releases a sufficient quantity of acid-soluble oligonucleotides to produce an absorbance increase of 1.0 at 260 nm with denatured DNA in 15 min at 37 ° C [1724]. RNase T_1 is used especially for RNA sequencing and RNA fingerprinting.

6.5.5.2.4. RNase U_2

Ribonuclease U_2 (E.C. 3.1.27.4), also called endoribunuclease, is obtained from *Ustilago sphaerogena.*

RNase U_2 is an endonuclease that cleaves RNA but not DNA. The enzyme specifically attacks purine nucleotides by cleaving the 3′-adjacent phosphodiester bond Pu/pN [1745]. The molecular mass is 10×10^3 dalton, and the pH optimum is 3.5. Cyclic 2′,3′-purine nucleotides are obtained as intermediates. Purine 3′-phosphates and oligonucleotides with purine 3′-phosphate terminal groups are the final products. Reversal of the final step can be used for the synthesis of ApN or GpN [1726].

One unit of RNase U_2 is defined as the enzyme activity that releases a sufficient quantity of oligonucleotides to produce an absorbance increase of 1.0 at 260 nm with denatured DNA in 15 min at 37 °C [1726]. RNase U_2 is used as A- and G-specific RNAse in RNA sequencing [1720], [1722].

6.5.5.2.5. Nuclease S7

Nuclease S7, also called endonuclease (E.C. 3.1.31.1), is obtained from *Staphylococcus aureus.*

Nuclease S7 is an endonuclease that cleaves both DNA and RNA. The products obtained after cleavage of RNA or DNA with nuclease S7 are oligo- and mononucleotides with terminal 3′-phosphate groups. The enzyme manifests A or U specificity in RNA digestion at pH 7.5 [1727].

One unit of nuclease S7 is defined as the enzyme activity that releases a sufficient amount of acid-soluble oligonucleotides to produce an absorbance increase of 1.0 at 260 nm with denatured DNA in 30 min at 37 ° C [1727].

The enzyme is used widely to optimize the standard rabbit reticulocyte lysate. The mRNA-dependent protein synthesis system with only low amounts of endogenous RNA is obtained by incubation of the lysate with the enzyme in the presence of $CaCl_2$ [1728]. Because of the strict dependence of the nuclease on Ca^{2+} ions, it can easily be inactivated by a specific chelating agent such as ethylenediaminetetraacetic acid. During incubation at 50 °C and pH 7.5 in the presence of $CaCl_2$ and urea, NpA and NpU bonds are cleaved. Nuclease S7 is therefore also used as A- and U-specific RNase in RNA sequencing [1722]. This reaction generates RNA fragments with 3′-phosphate termini. Thus, in sequencing gels, the generated RNA fragments are shifted downstream by one nucleotide compared with RNase T_1 or RNase U_2 fragments of identical length.

6.5.5.2.6. Site-Specific RNases in RNA Sequence Analysis

Base-specific RNases and nucleases with RNA and DNA specificity can be used for enzymatic sequencing of RNA molecules [1745], [1746], [1747], [1748] (Fig. 105). Because the sequence of RNA molecules larger than some hundred nucleotides cannot be determined directly, two alternative approaches are possible:

1. Fingerprint analysis, for example, of RNase T_1-resistant fragments [1729], [1730];
2. Sequencing RNA after partial digestion of large RNA chains (e.g., with RNase T_1) to produce a set of fragments covering the whole molecule [1731].

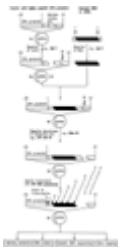

Figure 105. Use of SP6 RNA polymerase for in vitro synthesis of mRNA
[Full View]

In both approaches, the RNA fragments must be terminally labeled with ^{32}P in vitro prior to separation. End-labeling of RNA can be achieved in two ways:

1. Labeling the 5′-terminus by using T4 polynucleotide kinase and γ-^{32}P-ATP as substrate (left branch, Fig 105); [1733], [1734], [1735].
2. Labeling the 3′-terminus by using T4 RNA ligase and 5′-^{32}P-Cp as substrate (right branch, Fig. 105) [1736].

In addition, 3′-end-labeling of tRNA may be obtained with tRNA nucleotidyltransferase. Any RNA molecules that fail to be polyadenylated can be labeled at the 3′-terminus with *E. coli* poly(A) polymerase [1737], [1738], [1739]. After isolation of the various RNase T_1 subfragments, RNA sequencing consists of three analytical steps, as shown in Figure 105, bottom part.

Determination of 5′- and 3′-Terminal Nucleotides. Information about the nature of the 5′-terminally labeled nucleotide is obtained by complete digestion of the fragment with nuclease P1, resulting in 5′-^{32}P-nucleotides [1672], [1740], [1748]. Excessive digestion of 3′-terminally labeled fragments with RNase T_2 permits the identification of the 3′-termini by analysis of the liberated 3′-^{32}P-nucleotides.

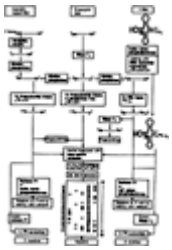

Figure 107. Use of nucleases with RNA and RNA – DNA specificity in sequencing RNA

H^+ = acid hydrolysis
[Full View]

Figure 108. Chemical DNA sequencing

[Full View]

Sequence Determination of the Terminal Nucleotides by Mobility-Shift Analysis. Sequence information concerning limited regions of the 5′- and 3′-termini is obtained by partial enzymatic digestion of alternatively 5′- or 3′-labeled oligonucleotides with randomly acting endonuclease P1 or snake venom phosphodiesterase, acting as 3′-exonucleases; by randomly acting endonuclease S1 or calf spleen phosphodiesterase, acting as 5′-exonucleases; [1672], [1741], [1742], [1743], or by chemical treatment with $NaHCO_3$ or ethylenediaminetetraacetic acid [1739].

Complete Sequence Analysis by Partial Digestion with Base-Specific Nucleases. In addition to chemical RNA sequencing [1744], RNA sequencing by reverse transcription in cDNA [1639], enzymatic RNA sequencing by partial digestion with the base-specific RNases mentioned in the previous sections (RNase T_1, RNase U_2, nuclease S7, and RNase CL3) is a well-established procedure [1720], [1722], [1745], [1748]. Reaction details are given in Table 46.

Table 46. Partial digestion of RNA with base-specific RNases

Enzyme	Conditions	Specificity
RNase T_1	urea, 8 mol/L, pH 3.5; 50 °C	$^{5'}Gp/N^{3'}$
RNase U_2	urea, 8 mol/L, pH 3.5; 50 °C	$^{5'}Ap/N^{3'}$ and ($^{5'}Gp/N^{3'}$)
RNase S7	urea, 8 mol/L, pH 7.5; 50 °C	$^{5'}Np/A^{3'}$ and $^{5'}Np/U^{3'}$
RNase CL3	urea, 8 mol/L, pH 8.0; 50 °C	$^{5'}Cp/N^{3'}$ and ($^{5'}Ap/N^{3'}$ + $^{5'}Gp/N^{3'}$)

Identification of purine bases within the RNA chain is achieved by digestion of RNA with RNase T_1 and RNase U_2 [1745], [1746], [1747], [1748]. Digestion with RNase T_1 (G-specific) results in fragments with 3′-

terminal Gp residues. Digestion with RNase U_2 (A- and G-specific) under appropriate conditions results in fragments with mainly 3′-terminal Ap residues (A-specific). Pyrimidine residues are localized by digestion of RNA with nuclease S7 at pH 7.5 and RNase CL3 at pH 8.0. Digestion with RNase CL3 results mainly in fragments with 3′-terminal Cp residues, less frequently in fragments with 3′-terminal Ap and Gp residues (minus-U-specific = C-, A-, and G-specific). Digestion with nuclease S7 at pH 7.5 results in fragments with 5′-terminal A and U residues (A- and U-specific) [1722].

Care must be taken in reading the sequence ladder, because — compared with T_1 or U_2 fragments of corresponding length — the electrophoretic mobility of 5′-endlabeled S7 fragments is shifted downward by exactly one nucleotide; 3′-endlabeled S7 fragments are shifted upward by one nucleotide. This difference in eletrophoretic mobility compared with T_1 and U_2 fragments is caused by the cleavage site of the enzyme upstream from A or U residues. Normal electrophoretic mobility is observed with RNA fragments obtained with RNase CL3 [1720], [1721]. Digestion with RNase CL3 results mainly in fragments with 3′-terminal Cp residues. The formation of fragments with 3′-terminal Ap and 3′-terminal Gp residues is much less frequent. 3′-Terminal Up residues are generated to only a very small extent. Thus the information obtained by digestion of RNA with nuclease S7 and RNase CL3 is useful for unambiguous discrimination between both pyrimidine residues.

6.5.6. Modifying Enzymes

Table 47 lists important DNA-modifying enzymes under E.

Table 47. Properties of important DNA- and RNA-modifying enzymes

Enzyme	Organism	M_r, 10^3 dalton	Sub-units	References (isolation)	Specificity[a]	Substrates[b]	P c
A. DNA polymerases							
A.1. Prokaryotic DNA polymerases							
***E. coli* DNA polymerases**							
DNA polymerase I	*Escherichia coli*	109 [Zn^{2+}]	α	[1844], [1847], [1855]	3′–5′ polymerase	DNA matrix, 3′-OH DNA or RNA primer, dNTP	7
(Kornberg enzyme)					3′–5′ exonuclease	ssDNA	
					5′–3′ exonuclease	dsDNA, DNA/RNA hybrid	
					pyrophosphorylase	DNA matrix, 3′-OH DNA or RNA primer, pp_i	
					pyrophosphate exchange	DNA, dNTP, pp_i	
DNA polymerase I,	*Escherichia coli*	76 [Zn^{2+}]	αC	[1596], [1597], [1912]	3′–5′ polymerase	DNA matrix, 3′-OH DNA or RNA	7

						primer, dNTP	
large fragment (Klenow enzyme)					3′–5′ exonuclease pyrophosphorylase pyrophosphate exchange	ssDNA dsDNA DNA matrix, 3′-OH DNA or RNA primer, pp_i	7
DNA polymerase I, small fragment	*Escherichia coli*	36	α_N	[1596], [1597], [1732]	5′–3′ exonuclease	dsDNA	7
DNA polymerase II	*Escherichia coli*	120	α	[1984], [2059]	3′–5′ polymerase	DNA matrix, 3′-OH DNA or RNA primer,	7
					3′–5′ exonuclease	ssDNA	
DNA polymerase III	*Escherichia coli*	[(175_{core}) 379_{holo}]	[(α, ε, θ) core τ, γ, δ, β] holo	[1857], [1896], [1897]	3′–5′ polymerase	ssDNA matrix, 3′-OH RNA primer, dNTP, ATP	7
					3′–5′ exonuclease	ssDNA	
					5′–3′ exonuclease	ssDNA	
Phage-coded DNA polymerases							
T4 DNA polymerase	*Escherichia coli*, phage T4-infected	114	α	[1610], [1909]	3′–5′ polymerase	DNA matrix, 3′-OH DNA or RNA primer, dNTP	8
					3′–5′ exonuclease	ssDNA	
T5 DNA polymerase	*Escherichia coli*, phage T5-infected	96	α	[1919], [2055]	3′–5′ polymerase	DNA matrix, 3′-OH DNA primer, dNTP	8
					3′–5′ exonuclease	ss or dsDNA	
T7 DNA polymerase	*Escherichia coli*, phage T7-	96	α, β	[1833], [1999]	3′–5′ polymerase	DNA matrix, 3′-	7

	infected					OH DNA primer, dNTP	
					3′–5′ exonuclease	ss or dsDNA	
A.2. Eukaryotic DNA polymerases							
Type α DNA polymerases							
DNA polymerase α	*Drosophila melanogaster*	280	4 sub-units	[2008]			
	mouse myeloma cells	190	2 sub-units	[1893], [2032]			
	rat liver	155–250	5 sub-units	[1898], [2046]	3′–5′ polymerase	DNA (RNA) matrix, 3′-OH DNA primer, dNTP	7
	calf thymus	210–230	2 sub-units	[1835], [1838], [1842]	pyrophosphorylase		
	human HeLa cells	320 (α_1)		[1869]	pyrophosphate		
		600 (α_2)	5 sub-units	[1869]	exchange		
		220 (α_3)	2 sub-units	[1869]			
	human KB cells	150–160	2 sub-units	[1854], [2047]			
Type β DNA polymerases							
DNA polymerase β	calf thymus	44	α	[2027], [2029]			
	chicken embryo	40	α	[1993]	3′–5′ polymerase	DNA or RNA matrix,	8
	human KB cells	43	α	[1981]		3′-OH DNA primer, dNTP	
	Novikoff hepatoma cells	31	α	[1965]			
Type γ DNA polymerases							
DNA polymerase γ	chicken embryo	180	α_4	[1992]			
	mouse myeloma cells	140	α_4	[1894]			
	human HeLa cells	110		[1851], [1961]	3′–5′ polymerase	RNA (DNA)	8

						matrix, 3′-OH DNA primer, dNTP	
	human lymphoblasts	120		[1940]			
Type δ DNA polymerases							
DNA polymerase δ	calf thymus	140–200	α, β	[1456], [1870]	3′–5′ polymerase	RNA matrix, 3′-OH RNA primer, dNTP	
					3′–5′ exonuclease		
	rabbit bone marrow	122		[2021], [2022]	3′–5′ polymerase	RNA matrix, 3′-OH RNA primer, dNTP	
					3′–5′ exonuclease		
Virus-coded DNA polymerases							
Reverse transcriptase	avian myeloblastosis virus (AMV)	160 [Zn^{2+}]	α, β	[1621], [2062]	3′–5′ polymerase	RNA or DNA matrix, 3′-OH DNA or RNA primer, dNTP	8
					RNase H	RNA/DNA hybrid	
					DNA endonuclease	dsDNA	
					pyrophosphorylase		
					pyrophosphate exchange		
Reverse transcriptase	Moloney murine leukemia virus (Mo–MLV)	80 [Zn^{2+}]	α	[1906], [1978], [1621], [2061]	3′–5′ polymerase	RNA or DNA matrix, 3′-OH DNA or RNA primer, dNTP	8
					RNase H	RNA or	

						DNA hybrid	
					DNA endonuclease pyrophosphorylase pyrophosphate exchange	dsDNA	
HSV-1 DNA polymerase	RC-37 cells, HSV-1-infected	144–150	α	[1850], [1873], [2017], [2044]	3′–5′ polymerase	DNA matrix, 3′-OH DNA primer	8
					3′–5′ exonuclease pyrophosphate exchange	ssDNA	
Vaccina virus DNA polymerase	human HeLa cells, vaccina virus-infected	110–115	α	[2023]	3′–5′ polymerase	DNA matrix, 3′-OH DNA primer, dNTP	8
					3′–5′ exonuclease pyrophosphate exchange	ssDNA	
Terminal transferases							
Terminal transferase (Bollum enzyme)	calf thymus	62 [Co^{2+}]	α	[1642], [1824]	3′–5′ polymerase	3′-OH DNA primer, dNTP	7
B. RNA Polymerases							
B.1. Prokaryotic RNA polymerases							
Bacterial RNA polymerases							
RNA polymerase	*Escherichia coli*	454 [Zn^{2+}]	*	[1575], [2018], [2025], [1920], [1921], [1922], [1996]			
	Acetobacter vinelandii	400–500 [Zn^{2+}]	*	[1862], [1863]	3′–5′ RNA polymerase	DNA matrix, NTP	7
	Bacillus subtilis	400–500 [Zn^{2+}]	*	[2005], [1954]	de novo RNA synthesis	ATP and UTP or ITP and CTP	7
	Caulobacter crescentus	400–500 [Zn^{2+}]	*	[2011]	pyrophosphate exchange	poly[d(AT)], ATP, UTP, pp_i	7
	Lactobacillus curvatus	400–500 [Zn^{2+}]	*	[1970]			
	Micrococcus luteus	400–500 [Zn^{2+}]	*	[1839], [1877]			
Phage-coded RNA polymerases							
T7 RNA polymerase	*Escherichia coli*, phage T7	98	α	[1651], [1659]			7

	coli, phage T7-infected			[1659]			
T3 RNA polymerase	*Escherichia coli*, phage T3-infected	105	α	[2010], [1650], [1895]			7
N4 RNA polymerase	*Escherichia coli*, phage N4-infected	350	α	[1923]			7
SP6 RNA polymerase	*Salmonella typhimurium*, phage SP6-infected	96	α	[1648], [1653]	3′–5′ RNA polymerase	DNA matrix, NTP	7
PBS2 RNA polymerase	*Bacillus subtilis*, phage PBS2-infected	260	α, β, γ, δ, ε	[2034]			7
Poly(A) polymerases							
Poly(A) polymerase	*Escherichia coli*	50	α	[1738]	terminal riboadenylate transferase	3′-OH RNA	7
Poly(A) polymerase	calf thymus	60	α	[1975]	terminal riboadenylate transferase	3′-OH RNA	7
		120–140	α_2				
Poly(A) polymerase	human HeLa cells	63–75	α	[1915]	terminal riboadenylate transferase	3′-OH RNA	7
		50–58	α				
Polynucleotide phosphorylases							
PNPase	*Escherichia coli*	252	α_3	[1926]	polyribonucleotide nucleotidyltransferase ADP-p_i exchange	NDP	
PNPase	*Micrococcus luteus*	237		[2009]	polyribonucleotide nucleotidyltransferase ADP-p_i exchange	NDP	
tRNA nucleotidyltransferases							
tRNA nucleotidyltransferase	*Escherichia coli* B	45–54	α	[1901]	tRNA nucleotidyltransferase	tRNA, ATP and GTP	9

nucleotidyltransferase	*coli* B				nucleotidyltransferase	and CTP	
tRNA nucleotidyltransferase	baker's yeast	70–71	α	[1967]	tRNA nucleotidyltransferase	tRNA, ATP and CTP	9
B.2. Eukaryotic RNA polymerases							
RNA polymerases							
RNA polymerase I	yeast	500–600 [Zn^{2+}]	11 sub-units	[1946], [1977]			
	wheat germ	400–500	7 sub-units	[1997], [1998]			
	Acanthamoeba castellanii	500–600	11 sub-units	[2001], [2002]			
RNA polymerase II	yeast	500–600 [Zn^{2+}]	9 sub-units	[1946], [1977]			
	wheat germ	500–600 [Zn^{2+}]	11 sub-units	[1997], [1998]	3′–5′ polymerase	ss (ds) DNA matrix, NTP	7
	Acanthamoeba castellanii	500–600	10 sub-units	[2001], [2002]			
RNA polymerase III	yeast	500–600 [Zn^{2+}]	10 sub-units	[1946], [1977]			
	wheat germ	500–600	14 sub-units	[1997], [1998]			
	Acanthamoeba castellanii	500–600	14 sub-units	[2001], [2002]			
Organelle-specific RNA polymerases							
RNA polymerase	maize chloroplasts	500	α, β, γ, δ	[1452], [1848]	3′–5′ polymerase	DNA matrix, NTP	
RNA polymerase	yeast mitochondria	100–150	α_{2-3}	[1874]	3′–5′ polymerase	DNA matrix, NTP	
2′,5′-Oligoadenylate synthetases							
2′,5′-Oligoadenylate synthetase	chicken embryonic cells	50–60		[2006]	2′,5′-oligoadenylate adenyltransferase	dsRNA, ATP	7
2′,5′-Oligoadenylate synthetase	Ehrlich ascites tumor cells	85		[2006], [1828]	2′,5′-oligoadenylate adenyltransferase	dsRNA, ATP	7
C. Nucleases							
C.1. Restriction endonucleases (see Table 43)							
C.2. Endonucleases							
Double-strand specific endonuclease							
DNase I	bovine pancreas	31	α	[1667], [1669]	endonuclease	dsDNA	7

Single-strand specific endonucleases							
Nuclease S1	*Aspergillus oryzae*	38 [Zn^{2+}]	α	[2003], [1691]	endonuclease	ssDNA (RNA), partially denatured dsDNA, DNA/RNA hybrid	4
P. citrinum nuclease	*Penicillium citrinum*			[2053]	endonuclease	ssRNA (DNA)	5
U. maydis nuclease	*Ustilago maydis*	42 [Zn^{2+}]	α	[1841]	endonuclease	ssDNA (RNA)	8
Mung bean nuclease	mung beans	39 [Zn^{2+}]	α	[2003]	endonuclease	ssDNA or ssRNA	5
WS nuclease	wheat germ	43 [Zn^{2+}]	α	[1864]	endonuclease	ssDNA (RNA)	4
AP endonucleases							
Endonuclease III	*Escherichia coli*	27	α	[2058]	AP endonuclease DNA glycosylase	dsAP DNA thymine dimers	7
Endonuclease IV	*Escherichia coli*	33	α	[1886]	AP endonuclease	dsAP DNA	8
Endonuclease V	*Escherichia coli*	20	α	[2057]	AP endonuclease	dsAP DNA	9
Endonuclease VII	*Escherichia coli*	45	α	[2051]	AP endonuclease	ssAP DNA	7
Mlu AP endoA	*Micrococcus luteus*	35	α	[1925]	AP endonuclease	dsAP DNA	7
Mlu AP endoB	*Micrococcus luteus*	35	α	[1925]	AP endonuclease	dsAP DNA	6
Hin AP endo	*Haemophilus influenzae*	30	α	[2035]	AP endonuclease 3′-exonuclease 3′-phosphatase	dsAP DNA	7
Bsu AP endo	*Bacillus subtilis*	56	α	[2014]	AP endonuclease		
CL AP endo	calf liver	28	α	[1865]	AP endonuclease		9
HP AP endo	human placenta	27–31	α	[1880]	AP endonuclease		
C.3. Exonucleases							
***E. coli* exonucleases**							
Exonuclease I	*Escherichia coli*	70–72	α	[1888], [1930]	3′–5′ exonuclease	ssDNA	9
Exonuclease III	*Escherichia*	28	α	[1844],	3′–5′ exonuclease	dsDNA	7

	coli			[1982]			
					DNA 3′-phosphatase	3′-P DNA	
					AP endonuclease	apurinic–apyrimidinic DNA	
					RNase H	RNA/DNA hybrid	
Exonuclease IV	*Escherichia coli*			[2015]	3′–5′ exonuclease	ssDNA	8
Exonuclease V	*Escherichia coli*	270	α, β	[2065]	3′–5′ exonuclease 5′–3′ exonuclease	dsDNA	9
Exonuclease VII	*Escherichia coli*	88	α	[1704]	3′–5′ exonuclease 5′–3′ exonuclease	ssDNA	7
Exonuclease VIII	*Escherichia coli*	140	α	[2064]	5′–3′ exonuclease	dsDNA	8
Exonuclease V-analogous enzymes							
*Hin*d exonuclease V	*Haemophilus influenzae*	290	α, β, γ	[2033], [1690], [1987]	3′–5′ exonuclease 5′–3′ exonuclease	dsDNA	9
Bsu exonuclease V	*Bacillus subtilis*	300	α, β	[1913], [1949]	3′–5′ exonuclease 5′–3′ exonuclease	dsDNA	9
Pae exonuclease V	*Pseudomonas aeruginosa*	300		[1900], [1913]	3′–5′ exonuclease 5′–3′ exonuclease	dsDNA	9
Sin exonuclease V	sea urchin intermedius	450		[2056], [1913]	3′–5′ exonuclease 5′–3′ exonuclease	dsDNA	9
Other exonucleases							
Nuclease Bal 31	*Altermonas espejiani*	73	α	[2068], [1832]	3′–5′ exonuclease 5′–3′ exonuclease	ss or dsDNA	8
N. crassa nuclease	*Neurospora crassa*	55 [Zn^{2+}]	α	[1878], [1879]	exonuclease endonuclease	linear ss or dsDNA circular ssDNA or ssRNa	7
Phage-coded exonucleases							
λ exonuclease	*Escherichia*	52	α	[2066],	5′–3′ exonuclease	dsDNA, 5′-	8

	coli, phage λ-infected			[1928]		P terminus	
D. RNA nucleases							
D.1. *E. coli* ribonucleases							
RNase I	*Escherichia coli*	low	α	[1962]	endonuclease	ssRNA, (A/U)p/N	8
RNase II	*Escherichia coli*	68–85	α	[1834]	3′–5′ exonuclease	ssRNA, 3′-OH terminus	7
RNase III	*Escherichia coli*	50	α_2	[2041], [1830]	endonuclease	ss or dsRNA, secondary structures	7
RNase IV	*Escherichia coli*			[1797]	endonuclease	ssRNA secondary structures	
RNase H	*Escherichia coli*	40	α	[2038], [1712], [1902]	endonuclease	DNA/RNA hybrid	7
D.2. Other ribonucleases							
RNase A	bovine pancreas	13.7	α	[1867]	endonuclease	ssRNA, (C/U)/pN	7
RNase CL3	chicken liver	16.85	α	[1720], [1721]	endonuclease	ssRNA, C (A/G)p/N	8
RNase T_1	*Aspergillus oryzae*	11.1	α	[2043], [1725]	endonuclease	ssRNA, Gp/N	3
RNase U_2	*Ustilago spherogena*	10.0	α	[1726]	endonuclease	ssRNA, Ap/N	3
Nuclease S7	*Staphylococcus aureus*	16.8	α	[1728], [1974]	endonuclease	ssRNA, Np/(A/U)	7
D.3. tRNA-processing enzymes							
RNase P	*Escherichia coli*	17.5 + 120	α, Ml RNA	[1852]	endonuclease	tRNA or rRNA precursor	8
RNase D	*Escherichia coli*	40		[2040]	exonuclease	tRNA precursor, 3′-terminus	9
RNase P_2	*Escherichia coli*			[1592]	endonuclease	tRNA precursor	
RNase O	*Escherichia coli*			[1952]	endonuclease	tRNA precursor	7
E. DNA-modifying enzymes							
E.1. Restriction methylases (see Table 42)							

E.1. Restriction methylases (see Table 43)							
E.2. Phosphatases							
Alkaline phosphatase	calf intestine	140 [Zn^{2+}]	α_2	[1733]	phosphatase	5′-P or 3′-P DNA or RNA	7
Alkaline phosphatase	*Escherichia coli*	80 [Zn^{2+}]	α_2	[1636]	phosphatase	5′-P or 3′-P DNA or RNA	7
Phosphodiesterase	calf spleen			[2016]	phosphodiesterase	5′-OH DNA or RNA	5
Phosphodiesterase	*Crotalus durissus*			[1934]	phosphodiesterase	3′-OH DNA or RNA	5
E.3. DNA ligases							
Bacterial DNA ligases							
DNA ligase	*Escherichia coli*	77	α	[1905], [1772]	polydeoxyribonucleotide synthase	5′-P DNA or RNA, 3′-OH DNA or RNA (overlapping single-stranded ends of dsDNA or dsRNA)	7
Phage-coded DNA ligases							
T4 DNA ligase	*Escherichia coli*, phage T4-infected	68	α	[1761], [1763]	polydeoxyribonucleotide synthase	5′-P DNA or RNA, 3′-OH DNA or RNA (overlapping or blunt, single-stranded or double-stranded ends of ds DNA or dsRNA)	7
					pyrophosphate exchange	ATP, pp_i	
E.4. Kinases							
T4 polynucleotide kinase	*Escherichia coli*, phage T4-infected	140	α_4	[1777], [1763]	5′-DNA kinase	5′-OH DNA or RNA, NTP or dATP	7
					3′-phosphatase	3′-P DNA or RNA	5
T4 polynucleotide kinase, 3′-phosphatase-free	*Escherichia coli*, phage T4*pse*T1-infected	140	α_4	[2019], [1798]	5′-DNA kinase	5′-OH DNA or RNA, NTP or	7

	infected					dATP	
DNA kinase	rat liver	80		[1875]	5′-DNA kinase	5′-OH DNA, ATP	5
DNA kinase	calf thymus	70		[2004]	5′-DNA kinase	5′-OH DNA, ATP	5
DNA kinase	human HeLa cells			[1955]	5′-RNA kinase	5′-OH RNA, ATP	5
E.5. DNA topoisomerases							
Type I DNA topoisomerases							
ω protein	*Escherichia coli*	100–120	α	[1826]	transient ssDNA breakage:		7
M. luteus	*Micrococcus luteus*	100–120	α	[1866]	ccDNA relaxation	dsDNA	
topoisomerase I					DNA–catenane	dsDNA	
HeLa topoisomerase I	human HeLa cells	100	α	[1882]	knot formation		
Type II DNA topoisomerases							
E. coli DNA gyrase	*Escherichia coli*		α, β	[1904]	transient dsDNA breakage:		7
					cccDNA formation	dsDNA, ATP	
M. luteus DNA gyrase	*Micrococcus luteus*	420–430	$(\alpha\beta)_2$	[1883]	DNA-dependent ATPase cccDNA relaxation	dsDNA, ATP dsDNA	
						dsDNA	
T4 topoisomerase I	*Escherichia coli*, phage T4-infected		α, β	[1884], [1969]	DNA–catenane knot formation	dsDNA	
E.6. DNA helix unwinding enzymes							
rep protein	*Escherichia coli*, phage ΦX174-infected	65	α	[1944]	DNA unwinding	dsDNA, *E. coli* ssb protein	
					DNA–dependent ATPase	ssDNA, ATP	
Helicase III	*Escherichia coli*	40	α_2	[1829]	5′–3′ DNA unwinding DNA–dependent ATPase	dsDNA, *E. coli* ssb protein ssDNA, ATP	
E.7. Single-strand DNA-binding proteins							
E. coli ssb protein	*Escherichia coli*		α_4	[2031]	ssDNA binding	ssDNA, 8 nucleotides per monomer	6
T4 gene 32 protein	*Escherichia coli*, phage T4-infected		$\alpha_{1-2/n}$	[2012]	ssDNA binding	ssDNA, 7 nucleotides per monomer	5

T7 DNA-binding protein	*Escherichia coli*, phage T7-infected		α	[1933]	ssDNA binding	ssDNA	7
fd gene 5 protein	*Escherichia coli*, phage fd-infected		$\alpha_{1–2}$	[2000]	ssDNA binding	ssDNA, 4 nucleotides per monomer	8
CT HDP-I protein	calf thymus		α	[1837]	ssDNA binding	ssDNA, 7 nucleotides per monomer	7
F. RNA-modifying enzymes							
F.1. RNA ligases							
T4 RNA ligase	*Escherichia coli*, phage T4-infected	43	α	[2037], [1794]	polyribonucleotide synthase	5′-P RNA (DNA), 3′-OH RNA (DNA)	7
					pyrophosphate exchange	ATP, pp_i	
F.2. CAP-forming enzymes[c]							
Capping enzyme	smallpox virus	127	α, β	[1891], [1957]	RNA triphosphatase	pppRNA	7
					RNA guanyltransferase	GTP + ppRNA	
					RNA (guanine-7) methyltransferase	GpppRNA + SAM	
					GTP–pp_i exchange nucleoside	GTP + pp_i	
					NTP triphosphate phosphohydrolase		
F.3. CAP-splitting enzymes[d]							
Decapping enzyme (TAP)[e]	*Nicotiana tabacum* var. Wisconsin 38	280	α_4	[1953]	acid pyrophosphatase	7-methyl-GpppN	6

[a]AP = apurinic.
[b]cc = closed circular; ccc = covalent closed circular.
[c]IP = isoelectric point.
[d]CAP = 7-methyl-GpppN.
[e]TAP = tobacco acid pyrophosphatase.
*$[(\beta'\beta\alpha_2)_{core}{}^{a}]_{holo}$

6.5.6.1. Alkaline Phosphatase
Alkaline phosphatase, also called orthophosphoric monoester phosphohydrolase (alkaline optimum) (E.C. 3.1.3.1) [9001-78-9], is obtained from calf intestine.

Properties. The molecular mass of alkaline phosphatase from calf intestine is 140×10^3 dalton. The enzyme is a dimeric glycoprotein composed of two identical or similar subunits with a molecular mass of 69×10^3 dalton each. Alkaline phosphatase contains four Zn^{2+} ions per molecule [1733].

One unit of alkaline phosphatase is defined as the enzyme activity that hydrolyzes 1 µmol of 4-nitrophenyl phosphate in 1 min at 37 °C in a diethanolamine buffer (1 mol/L); 4-nitrophenyl phosphate concentration: 10 mmol/L; $MgCl_2$ concentration: 0.25 mmol/L; pH 9.8 [1749].

The purified enzyme is most stable in the pH range 7.5 – 9.5 and is inactivated rapidly at lower pH. Protection against acidic denaturation is observed in the presence of inorganic phosphate [1750]. Alkaline phosphatase can also be inactivated either by treatment with NaOH or by heating for 45 min at 65 °C in the presence of nitrilotriacetic acid or ethylenediaminetetraacetic acid to chelate the essential Zn^{2+} ions [1751]. Alternatively, alkaline phosphatase can be inhibited by inorganic phosphate. In the first two methods, RNA or DNA may be damaged. However, treatment with the chelating agents completely inactivates alkaline phosphatase without damage to the polynucleotide.

Uses. Alkaline phosphatase catalyzes the hydrolysis of numerous phosphate esters, such as esters of primary and secondary alcohols, sugar alcohols, cyclic alcohols, phenols, and aminoalcohols. Phosphodiesters do not react. The enzyme is used preferentially to selectively cleave terminal phosphate groups from oligonucleotides and monophosphate esters. In molecular biology, alkaline phosphatase is used primarily for dephosphorylation of 5′-phosphorylated DNA or RNA ends. These 5′-hydroxylated substrates can be effectively 5′-endlabeled with T4 polynucleotide kinase and γ-^{32}P-ATP as substrate [1733], [1635], [1752], [1753], [1734], [1754], [1755]. 3′-Phosphorylated ends of RNA are also dephosphorylated with alkaline phosphatase [1754], [1747]. Either DNA or RNA that is ^{32}P-labeled at the 5′-terminus is most frequently used in chemical DNA sequencing [1753], [1734] or in RNA sequencing by degradation of end-labeled RNA with base-specific RNases [1754], [1747]. However, alkaline phosphatase is also involved in the labeling of DNA and RNA fragments used for mapping and fingerprinting studies [1756], [1757]. The enzyme is further used in the construction of recombinant DNA molecules. Self-annealing of the vector DNA can be suppressed considerably by dephosphorylation of the linearized vector molecule prior to insertion of the DNA fragment to be cloned [1758], [1759], [1760].

6.5.6.2. T4 DNA Ligase
The enzyme T4 DNA ligase, also called poly(deoxyribonucleotide) : poly(deoxyribonucleo-tide) ligase (AMP-forming) (E.C. 6.5.1.1) [9015-85-4], is obtained from bacteriophage T4-infected *Escherichia coli*.

Properties. The molecular mass of T4 DNA ligase is 65×10^3 dalton [1761], [1762]. The enzyme is composed of a single subunit [1763]. The T4 DNA ligase depends on ATP as a cofactor in the joining reaction [1764]. The energy of ATP hydrolysis to yield AMP and pyrophosphate is used to form a phosphodiester linkage between polynucleotide chains. The enzyme also catalyzes an exchange reaction between pyrophosphate and ATP [1765].

One unit of T4 DNA ligase is defined as the enzyme activity that converts 1 nmol of ^{32}P from pyrophosphate into material that can be absorbed by Norit in 20 min at 37 °C [1765]. The optimal pH range for T4 DNA ligase is 7.2 – 7.8; at pH 6.9 and 8.0, the enzyme has 46 and 56 %, respectively, of its activity at pH 7.6 [1766]. The enzyme requires Mg^{2+} for activity; Mn^{2+} is only 25 % as effective as Mg^{2+}. Low concentrations of NH_4^+ ions have no effect on the T4 DNA ligase reaction. Higher levels of monovalent cations inhibit the enzyme completely [1767].

Uses. The T4 DNA ligase is used to ligate DNA fragments with either 5′- or 3′-protruding or blunt ends. This enzyme is the only DNA ligase known that can catalyze blunt-end joining [1768]. An important use of

T4 DNA ligase is the preparation of recombinant DNA molecules for cloning experiments. Hydrogen-bonded recombinant DNA molecules can be generated by annealing two DNA fragments containing cohesive ends; T4 DNA ligase is the enzyme of choice for joining such cohesive ends since it requires a smaller overlapping sequence than *E. coli* DNA ligase.

Cohesive ends can be generated by cleavage of DNA with class II restriction endonucleases. An alternative approach is the addition of complementary homopolymer tails to the appropriate fragments with terminal transferase (G—C tailing). Cohesive ends can also be generated by blunt-end ligation with a synthetic DNA linker that contains the recognition sequence for a restriction endonuclease producing cohesive termini.

After labeling internal 5′-ends with T4 polynucleotide kinase, T4 DNA ligase can also be used to identify 3′- and 5′-end groups at single-stranded interruptions by nearest neighbor analysis [1766]. Further, T4 DNA ligase can be applied to determine the ability of other enzymes to act at nicks and gaps in duplex DNA molecules. In addition, T4 DNA ligase can be used to study the primary and secondary structure of DNA molecules [1769] and may be applied to the chemical synthesis of double-stranded DNAs with specific nucleotide sequences [1770].

6.5.6.3. Escherichia coli DNA Ligase

The enzyme *E. coli* DNA ligase, also called poly(deoxyribonucleotide) : poly(deoxyribonucleotide) ligase (AMP-forming, NMN-forming) (E.C. 6.5.1.2) [37259-52-2], is obtained from *Escherichia coli*.

Properties. *Escherichia coli* DNA ligase consists of a single polypeptide chain with a molecular mass of 74×10^3 dalton [1771], [1772]. The bacterial enzyme catalyzes phosphodiester-bond synthesis coupled to cleavage of the pyrophosphate group in NAD.

One unit of *E. coli* DNA ligase is defined as the enzyme activity that converts 100 nmol poly[d(AT)] to an exonuclease III-resistant form in 30 min at 30 °C [1773].

The NAD-dependent *E. coli* DNA ligase is much more specific than the ATP-dependent T4 DNA ligase. A number of ligations catalyzed by T4 DNA ligase are not catalyzed by *E. coli* DNA ligase, e.g., blunt-end ligation [1774], [1775]. The *E. coli* DNA ligase also does not act as an RNA ligase joining RNA molecules or DNA molecules containing RNA primer sequences.

Uses. The *E. coli* DNA ligase is used in full-length cDNA synthesis [1640]. This method complements the present technique that employs nuclease S1 digestion of the hairpin loop prior to the second-strand synthesis. With this new technique, second-strand synthesis is mediated by the synchronous action of the enzymes *E. coli* DNA polymerase I, *E. coli* RNase H, and *E. coli* DNA ligase after first-strand synthesis with reverse transcriptase. The procedure uses a plasmid DNA vector which itself serves as a primer for the first- and, ultimately, the second-strand cDNA synthesis. Both steps are designed to enrich for recombinants containing full-length cDNAs over those with truncated cDNAs.

6.5.6.4. T4 Polynucleotide Kinase

The enzyme T4 polynucleotide kinase, also called ATP : 5′-dephosphopolynucleotide-5′-phosphotransferase (E.C. 2.7.1.78) [37211-65-7], is obtained from phage T4-infected *Escherichia coli*.

Properties. The T4 polynucleotide kinase catalyzes the transfer for the terminal ϒ-phosphate group of ATP to the 5′-hydroxylated termini of polynucleotides like DNA or RNA. It also catalyzes the exchange of 5′-terminal phosphate groups [1776], [1755].

One unit of T4 polynucleotide kinase is defined as the enzyme activity required for the formation of 1 nmol of acid-precipitable ^{32}P in 30 min at 37 °C [1755].

The enzyme T4 polynucleotide kinase migrates as a single species in sodium dodecyl sulfate – polyacrylamide gel electrophoresis. The single band with a molecular mass of 33×10^3 dalton represents one of the four identical subunits of the active enzyme complex with a total molecular mass of 140×10^3 dalton [1777], [1763]. Maximum activity is obtained at pH 7.6 at 37 °C and requires Mg^{2+} ions and such reagents

as dithiothreitol or 2-mercaptoethanol [1763]. The reported stimulating effect of higher ionic strength or polyamides like spermine or spermidine [1752], [1763] results from stabilization of the active oligomeric tertiary structure of the enzyme [1777]. Concentrations of ATP of at least 1 µmol/L and a ratio of 5 : 1 between ATP and protruding 5′-hydroxyl ends are required for optimal phosphorylation [1778]. The T4 polynucleotide kinase is inhibited to 50 % by 7 mmol/L of sodium or potassium phosphate and to 75 % by 7 mmol/L of ammonium sulfate [1763]. In addition to its kinase activity, T4 polynucleotide kinase also exhibits 3′-phosphatase activity [1779], [1780]. The pH optimum of this activity is between 5.0 and 6.0 and, thus, different from that of the kinase activity at pH 7.6.

Uses. The enzyme T4 polynucleotide kinase is used to label DNA and RNA at its 5′-termini with ^{32}P residues by using γ-^{32}P-ATP as substrate [1781], [1778], [1755]. The 5′-terminus of DNA can be labeled with T4 polynucleotide kinase either by direct phosphorylation of 5′-hydroxyl groups or by exchange of DNA-bound, nonradioactively labeled 5′-phosphoryl groups and ^{32}P molecules [1782], [1784], [1734]. Alternative conditions for both reactions are also described for 5′-endlabeling of RNA [1745], [1747]. The T4 polynucleotide kinase is most commonly used in chemical sequence determination of DNA [1753], [1734] and RNA [1745], [1785], [1747], [1748]. Chemical DNA sequencing can determine the sequence of both the coding and the noncoding strands. Sequencing of the coding strand in the 5′ – 3′ direction is possible after labeling the 5′-end of DNA with T4 polynucleotide kinase and γ-^{32}P-ATP (Fig. 106). Sequencing of the complementary noncoding strand in the 3′-5′ direction is possible if a recessed 3′-end of dsDNA is elongated with Klenow enzyme and suitable α-^{32}P-dNTPs as substrates. 5′-Endlabeling is also used for mapping restriction sites by partial digestion [1786], [1787]; DNA [1647] or RNA fingerprinting [1788], [1789]; DNA footprinting via DNase protection [1790] or methylation protection [1791] ; hybridization studies [1792]; synthesis of substrates for DNA or RNA ligase [1793], [1794]; and sequence analysis of DNA [1795], [1931]. In addition, the 3′-phosphatase activity of T4 polynucleotide kinase may be used as a specific 3′-phosphatase [1779].

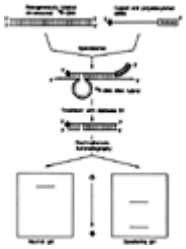

Figure 106. Mapping of introns in eukaryotic genes by digesting a hybrid of mature mRNA and the coding strand of DNA with nuclease S1
[Full View]

6.5.6.5. T4 Polynucleotide Kinase, 3′-Phosphatase-Free

3′-Phosphatase-free T4 polynucleotide kinase, also called ATP : 5′-dephosphopolynucleotide 5′-phosphotransferase (E.C. 2.7.1.78) [37211-65-7], is obtained from phage T4 *pse*T1- infected *Escherichia coli.*

Properties. 3′-Phosphatase-free T4 polynucleotide kinase is an altered T4 *pse* T1 gene product in which only the 3′-phosphatase activity but not the 5′-kinase activity has been affected [1779], [1780], [1796], [1797].

One unit of 3′-phosphatase-free T4 polynucleotide kinase is defined as the enzyme activity required for the formation of 1 nmol of acid-precipitable ^{32}P in 30 min at 37 °C [1755].

Maximum activity of the 5′-kinase reaction is obtained under the same conditions as those for wild-type T4 polynucleotide kinase. However, even under optimal conditions of the wild-type 3′-phosphatase activity, no removal of 3′-phosphatase groups of 3′-AMP is observed with the mutant T4 polynucleotide kinase at pH 5.5 – 6.0.

Uses. 3′-Phosphatase-free T4 polynucleotide kinase is of special interest in RNA analysis. The absence of the 3′-phosphatase makes this enzyme extremely useful for the preparation of unique species of RNA phosphorylated at both the 5′- and the 3′-termini. Oligoribonucleotides with both 3′- and 5′-terminal

phosphates are used as donors in the T4 RNA ligase reaction [1798], [1799]. Although the 3′-phosphatase is not required for activity with T4 RNA ligase, it is a convenient blocking group to prevent cyclization or self-addition of the donor. Thus, a unique intermolecular product is ensured.

Another use of 3′-phosphatase-free T4 polynucleotide kinase is the labeling of CMP to give 5′-^{32}P-Cp, which is commonly used for 3′-endlabeling of RNA with T4 RNA ligase [1800]. The 3′-terminally labeled RNA is very useful for fingerprinting and sequencing studies [1745], [1733], [1747].

6.5.6.6. Methylase HpaII

Methylase *Hpa*II, also called *S*-adenosyl-L-methionine : DNA (cytosine-5)-methyltransferase (E.C. 2.1.1.37) [9037-42-7], is obtained from *Haemophilus parainfluenzae*.

Properties. Methylase *Hpa*II is isolated from *H. parainfluenzae* containing methylase *Hpa*I as well [1801], [1802]. Methylase *Hpa*I acts on double-stranded DNA by transferring methyl groups from *S*-adenosylmethionine to the palindromic recognition site

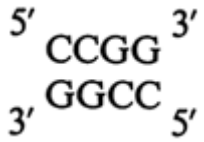

which results in 5-methylation of both internal cytosines [1801], [1803].

One unit of methylase *Hpa*II is defined as the enzyme activity required to protect 1 µg of λ DNA to > 95 % against cleavage by the restriction endonuclease *Hpa*II in 1 h at 37 °C [1801]. Single-stranded DNA is methylated by methylase *Hpa*II with very low efficiency. However, methylase *Hpa*II is capable of methylating hemimethylated recognition sequences [1927]. Since divalent cations such as Mg^{2+} are not essential for its activity, methylation of DNA with methylase *Hpa*II may be carried out in the presence of ethylenediaminetetraacetic acid. Site-specific methylation of DNA by methylase *Hpa*II protects the DNA against digestion by the restriction endonuclease *Hpa*II which recognizes and digests the identical tetrameric sequence CCGG [1804]. However, the activity of restriction endonuclease *Msp*I, an isoschizomer of *Hpa*II, is not affected by methylation of CCGG residues with methylase *Hpa*II because *Msp*I is only sensitive to methylation of the external cytosine residues [1805], [1806]. In addition, the action of the restriction endonuclease *Sma*I whose recognition site CCCGGG is a subset of the *Hpa*II sites, is also inhibited by methylation of the internal cytosine residue with methylase *Hpa*II. In contrast, the activity of *Xma*I (similar to *Msp*I) is not inhibited by methylation of *Hpa*II sites with methylase *Hpa*II [1807].

Uses. Methylase *Hpa*II is a useful tool for studying the in vivo effect of distinctly methylated $^{5'm}$CpG$^{3'}$ residues toward gene expression in eukaryotic cells. These 5-methylcytosine residues are located predominantly in CpG sequences [1808], [1809], [1810] at a level of about 1 % of the total nucleotides [1811]. An inverse correlation between the amount of CpG residues and gene activity was postulated [1812], [1813]. These results were confirmed by the gene-stimulating effect of azacytidine in vivo, because incorporation of this analogue into DNA mimics an undermethylated state of DNA [1814]. To study the in vivo effects of distinct methylation patterns, methylase *Hpa*II is used to introduce in vitro methylated CpG residues at the specific CCGG sites of isolated DNA [1815]. The DNA, which was previously methylated with methylase *Hpa*II to different degrees, is microinjected into frog oocytes [1816], or eukaryotic cells are transformed with the modified DNA by DNA-mediated gene transfer [1817]. The in vivo effects of the distinct *Hpa*II methylation patterns on gene activity can be studied by analysis of gene-specific transcripts in Northern blots. To correlate the level of transcriptional activity with changes in the degree of methylation, the level of *Hpa*II-specific methylation sites within the transformed eukaryotic cells can be demonstrated directly by comparison of the restriction activities of isoschizomeric nucleases *Hpa*II and *Msp*I on extracted DNA in Southern blots [1818], [1816], [1819].

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Alice J. Caddow, Genencor International, Inc., Palo Alto, USA
Beth Concoby, Genencor International, Inc., Palo Alto, USA

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

7. Enzyme Safety and Regulatory Considerations

This chapter reviews the safe handling of enzymes in the workplace and the regulation of enzyme-containing products. There are other aspects of both categories, including safety for the consumer of enzyme-containing products, labeling for safety and use, regulations pertaining to manufacturing, and some product regulatory categories that are not addressed. For information on these topics, the reader is referred to the trade association and other web sites listed.

7.1. Safe Handling of Enzymes

Much of our knowledge on the safe handling of enzymes has been developed and communicated by the detergent industry. In the late 1960s, proteases were first used on a large scale in detergent manufacturing. In the late 1960s and early 1970s there were reports of skin irritation and occupational asthma in detergent industry employees. These adverse health effects were attributed to enzyme exposure. In response to these health effects, the UK Soap and Detergent Industry Association (SDIA) formed a committee that developed industrial hygiene and employee health monitoring recommendations to reduce employee exposure to enzymes. The detergent industry and enzyme manufacturers worked together to reduce employee exposures to enzyme aerosols (dusts and mists) by the development of liquid and encapsulated enzymes [2070]. This has allowed the safe introduction and use of enzymes in many products and production processes, including detergents, processed foods, and textiles.

This chapter describes the elements of an enzyme safety program that are important for limiting exposure to enzymes and maintaining employee health and safety in the workplace. These include employee education, control measures, monitoring methods and medical surveillance. Many articles have been published on the various aspects of enzyme safety [2070-2073] and should be consulted for more detailed information.

7.1.1. Possible Health Effects

Data on the health hazards associated with enzymes substantiate a low degree of toxicity. The adverse effects induced by enzymes can be divided into two main types: (1) irritation and (2) respiratory allergy. There is no scientific evidence to suggest that exposure to enzymes can result in other manifestations of toxicity, including reproductive toxicity, developmental toxicity, mutagenicity, chronic toxicity, and carcinogenicity.

Respiratory Allergies. As with any protein that is foreign to the respiratory tract, repeated inhalation of enzyme-containing dust or mists can lead to the development of respiratory allergy or allergic asthma. There are two main stages in the development of respiratory allergy, which is also called type-one immediate hypersensitivity. The first stage is sensitization. This occurs when the individual is first exposed by inhalation to an allergen such as an enzyme, household dust, or pollen. If enough enzyme is inhaled the body will begin to recognize it as a foreign material and will produce allergic antibodies. Once allergic antibodies are produced, the individual is said to be sensitized. However, sensitization is not a disease, as there are no allergic symptoms at this stage.

In the second stage, when a sensitized individual is re-exposed to an allergen, such as an enzyme, clinical allergic symptoms can occur. Typical clinical allergic symptoms include watery eyes, runny nose, and coughing, as well as tightness of chest and shortness of breath, or occupational asthma. Not all sensitized individuals will develop allergic symptoms. The development of allergies depends on the susceptibility of the individual, and the magnitude and length of exposure. Symptoms will only occur if the susceptible individual inhales enzyme aerosols, and the symptoms will disappear within hours or a few days.

Skin and Eye Contact. Enzymes are not skin sensitizers. However, proteolytic enzymes cause skin irritation. Skin irritation usually only develops after prolonged exposure. This adverse manifestation is due to the irritant characteristics of proteases and is not an allergic response. Skin irritation is heightened by the presence of moisture and is most likely to appear on areas of the body where perspiration occurs: hands, armpits, and feet. Gloves should be worn when there is a potential for skin contact. Good personal hygiene is also essential for preventing skin irritation.

Proteolytic enzymes will also cause eye irritation. Appropriate eye protection should be worn to prevent contact when working with enzyme preparations.

7.1.2. Control Technology

The primary goal in an enzyme safety program is to maintain exposure below a level that would cause adverse health effects. This can be accomplished by paying attention to the product's physical form and the implementation of engineering controls, work practices, and personal protective equipment. The detergent industry has established the following key strategies to prevent employee exposure [2073]. These can be applied to other industries as well.

- Containment at source of any dust or liquid aerosol formation.
- The avoidance of recurring routine or uncontrolled spillages.
- The avoidance of personal contamination.
- The handling of empty containers with appropriate controls.

Product Form. The form of the product greatly influences the potential for aerosol formation. Because of this, product form often dictates the selection of engineering controls, handling procedures, and the type of protective equipment needed to provide adequate protection to the employee. Enzyme preparations are supplied in three main forms: liquids, granules, and powders. By their nature powdered enzyme preparations present the greatest chance for inhalation exposure because the fine dust can easily become airborne. When using powdered enzymes more stringent handling procedures and local exhaust ventilation are critical for ensuring safety in the workplace. Granular preparations are often provided with a protective coating which reduces the potential for formation of enzyme dust, but care must be given to avoid damaging and crushing the granules. When the granules have been crushed or damaged, the potential for exposure to enzyme-containing dust is increased. Enzymes in liquid form are nonvolatile; but there is still the chance of aerosol formation during material transfer, spills, mixing operations, and cleaning of equipment.

Engineering Controls. Engineering controls are the preferred method of preventing the release of enzyme aerosols from manufacturing equipment. Because the form of the product determines what kind of engineering controls to use in the production facility, engineering controls should be designed for the specific product form and production process. Engineering controls in the form of enclosures and local exhaust ventilation are very effective and are the most suitable methods to control enzyme exposure. The process should be enclosed as much as possible to contain enzyme aerosols that are generated during production operations. Local exhaust ventilation should be used to supplement the enclosures. Using enclosures and local exhaust ventilation in conjunction with each other assists in the isolation of the enzyme preparation from the employee. For this reason, they should be used in the following areas:

- Locations where enzyme preparations are added into the process.
- Material transfer points.
- Where the enzyme containing product is packaged into containers.

Design specification, performance verification, system maintenance, and process change management are key factors in designing and implementing engineering controls.

Work-Practice Controls. Work-practice controls include safe work practices, education, and good-housekeeping practices. Safe work practices should be instituted to prevent the generation of enzyme aerosols and prevent skin contact. Aerosols can form during cleaning and operations for cleaning up spills, such as sweeping and washing with high-pressure water. Vacuuming with a vacuum cleaner equipped with a

high-efficiency particulate air (HEPA) filter is the preferred method of cleaning. Washing with high-pressure water and steam cleaning should be avoided. Workers need to be aware of any work practices that may create aerosols and how to minimize aerosol formation.

Employee and contractor education is paramount to a safe working environment. To achieve good compliance with work practices and the use of control technology, employees need to know the reasons for these control measures. Employees need to be informed of the potential health hazards, how and when to use control measures, emergency procedures and reasons for a medical surveillance program.

Personal Protective Equipment. Personal protective equipment is utilized to supplement other control measures or as a primary control method in special situations such as spillage clean-up or equipment maintenance. Types of personal protective equipment include respiratory protection, protective clothes, and eye protection. A risk assessment should be conducted to determine when personal protective equipment should be worn and what type is necessary.

There are instances when engineering and work-practice controls need to be supplemented with respiratory protection. Cleaning operations and spillage clean-up may require respiratory protection, since it is difficult to control enzyme levels during these procedures. Important components of a respiratory protection program include selection of the appropriate respirator, training, fit testing, and medical surveillance.

Protective clothing and eye wear may be required in some operations such as maintenance operations, spillage clean-up, and cleaning operations. This skin and eye protection should be worn when there is a potential for skin and eye contact. This is especially true for proteolytic enzymes because of the irritation potential. Examples of eye and skin protection are safety glasses, goggles, face shields, gloves, and coveralls.

Air Monitoring. Air monitoring is often performed in work environments to evaluate airborne levels of a particular substance. This is important for assessing the efficiency of engineering controls, whether respiratory protection is required, and whether the airborne levels meet occupational exposure limits. Components of an air-monitoring program include an air-sampling plan, measurement methods, air-sampling equipment, and methods to evaluate the results.

Exposure Limits. The ACGIH has established a TLV for subtilisin of 60 ng/m^3. This is based on high-volume air monitoring for at least 60 min. Several countries have also established regulatory limits for subtilisin. Australia, Argentina, Canada, Denmark, Netherlands, and the UK have adopted the same level as the ACGIH for subtilisins.

Medical Surveillance. Medical surveillance programs are implemented to monitor the health of employees for early detection of any potential health effects. Elements of a medical surveillance program include a baseline exam, routine medical monitoring, and evaluation of employees with symptoms. Components of the medical exams could include medical history, respiratory questionnaire, medical examination, pulmonary function test, and determination of enzyme sensitivity.

Inhalation of enzymes can cause the production of allergic antibodies. An individual is sensitized once they produce allergic antibodies. Sensitization is not a disease, but an indication of exposure to a particular enzyme. Further inhalation exposure of a sensitized individual to the same enzyme type then can lead to allergic symptoms. There are medical tests to detect allergic antibodies, either by a laboratory blood test or by a skin prick test. These tests are used to detect sensitization or to aid in confirmation of an allergy to a particular enzyme.

Additional information can be obtained from enzyme suppliers and trade associations such as the Association of Manufacturers of Fermentation Enzyme Products (AMFEP) http://www.amfep.org/amfep.html, Association Internationale de la Savonnerie et de la Detergence (AISE) http://www.aise-net.org/, US Soap and Detergent Association (SDA) http://www.sdahq.org/, and the Enzyme Technical Association (ETA) http://enzymetechnicalassoc.org/.

7.2. Product Regulatory Considerations

Enzyme products are regulated according to their application, i.e., whether they are used in food manufacture, as digestive aids, in animal feed, in detergents, for textile processing, etc. Regulations also differ country by country. This section gives an overview of many of the regulations and gives references for more detailed information; it specifically covers enzymes used as food ingredients and processing aids, in feed, and for industrial or chemical use. The information contained herein is not meant to give the reader a step-by-step guide on how to register an enzyme product for a particular use in a particular country. In addition to the information here, both the Association of Manufacturers of Fermentation Enzyme Products (AMFEP) and the Enzyme Technical Association (ETA) maintain web sites with regulatory and safety information and links which may be of interest to the reader: http://www.amfep.org/amfep.html and http://enzymetechnicalassoc.org/. As with all of regulatory affairs, before seeking approval for a particular product in a country, it is important to contact the relevant authorities and discuss procedures and dossier content.

Enzyme products or preparations are generally characterized, and referred to according to their main enzymatic activity, i.e., as an α-amylase, subtilisin, glucoamylase, cellulase, etc. The IUB officially assigns E.C. numbers to enzymes (see http://www.chem.qmw.ac.uk/iubmb/enzyme/) as does the Chemical Abstracts Service (CAS, see http://www.cas.org/). Both the E.C. and CAS numbers are often referenced in specific regulations.

Whether there is specific legislation or not, it is prudent for the manufacturer of an enzyme preparation to assure that the potential product is safe for its intended use prior to introduction. For enzyme products intended for use in food, one needs to address the safety of the production organism (microorganism, plant, or animal), production process, and product. There are several publications that discuss these issues and provide a framework for this assessment, including [2074-2078]. In addition, the Food Chemicals Codex [2079] and WHO/JECFA [2080] (see http://www.fao.org/es/ESN/Jecfa) have specifications for food-grade enzymes that should be regarded as the minimum specifications required for the regions they cover; several countries impose stricter specifications. For enzyme products intended for use in feed, similar safety requirements apply; in addition, the safety issues discussed by by the Scientific Committee for Animal Nutrition (SCAN) and the Association of American Feed Control Officials (AAFCO) (see Section Feed-Use Enzymes) need to be addressed. For enzymes used in industrial processes, such as textile processing, leather tanning, paper production, and consumer products, such as detergents and automatic dish washing products, the principal safety issues are discussed in the Section Industrial-Use Enzymes.

7.2.1. Food-Use Enzymes

Table 40 lists approvals needed by application and country for enzymes used in food processing.

Table 40. Approvals needed by application and country for enzymes used in food processing

Country/region	Traditional production	GM production organism
EU additive	Approval, acc. 89/107 Directive	Approval, acc. 89/107 Directive
EU processing aid	Not harmonized	Not harmonized
France	Article 8 (1989) Notification or full submission	Article 8 (1989) Notification or full submission
Denmark	Approval needed	Approval needed
UK	Voluntary	Voluntary
Poland	Approval needed	Approval needed
Switzerland	No approval needed	Approval needed
Thailand	Registration	Registration
Korea	Registration	Approval needed
Taiwan	Registration	Registration
Japan	Approval needed if new	Approval needed
Australia/New Zealand	Approval needed	Approval needed

Zealand		
Canada	Approval needed if new	Approval needed
USA	Company GRAS assessment, GRAS Notice, or Food Additive	Company GRAS assessment, GRAS Notice, or Food Additive

The regulation of enzymes used in food processing is perhaps the most complex. Along with how enzymes are regulated in general country-by-country, one needs to be aware that enzymes produced using modern biotechnology techniques often have extra regulations, including the currently evolving ones for genetically modified organisms (GMO) food labeling. Because they are currently being developed and interpretation is key, GMO labeling regulations are not discussed in detail here. At present, the EU, Japan, and Australia/New Zealand, among others, have developed or are developing such regulations. The use of enzymes as digestive aids is also regulated specifically in some countries, e.g., the USA and the EU; these regulations are not discussed here.

In the EU, food enzymes are classified either as food additives or processing aids. If an enzyme has its technological function in the final food, it is considered a food additive and therefore regulated under the Food Additives Directive (95/2); there are only two enzymes currently regulated as such, invertase (E 1103) in confectionery and lysozyme (E 1105) in cheese. Certain other enzymes can be considered as food or food ingredients and therefore, if produced with novel methods, could fall under the Novel Foods Regulation 97/258. Most of the enzyme preparations used in food processing in the EU are considered processing aids, meaning that they have their technological function in the food-processing stage and not in the final food. They are excluded from the Food Additives Framework Directive according to its Article 1.3. Therefore, the use of most enzymes in food is not covered by a community regulation, but by widely divergent national provisions. In this context, it is important to mention that the European Commission, under pressure from some members who are not happy with the unharmonized situation, initiated a working group in 2000, the SCOOP Task Force 7.4. This Task Force 7.4 made an inventory of the enzymes available on the EU market, how they are regulated, how they are evaluated for safety, and how they should be classified. The final report of this Task Force was not published at the time when this paper was finalized.

To gain approval of an enzyme regulated as a food additive in the EU, a dossier needs to be submitted to the Scientific Committee for Food (SCF), which evaluates the safety of all food additives. The SCF issued Guidelines for the presentation of data on food enzymes on 11 April 1992 (see http://europa.eu.int/comm/food/fs/sc/scf/reports_en.html).

There are also two Vertical Directives in the EU that mention enzymes: the Fruit Juice and Wine Directives. In the Council Directive 93/77/EEC relating to fruit juices and certain similar products, pectolytic, proteolytic, and amylolytic enzymes are allowed. In the Council Regulations 822/87/EEC on the common organization of the market in wine, pectolytic enzymes are on the allowed ingredients list. If these Vertical Directives do not allow an enzyme, then the enzyme cannot be used in fruit juice or wine production.

Enzymes used as processing aids (not present with a technical effect in the final food) are regulated by national legislation in the EU. France, Denmark and the UK have legislation covering all food-use enzymes; France and Denmark require approval prior to use; in the UK, the approval is voluntary, but recommended.

The French Food Enzyme Law, Arrete du Septembre 1989 (J.O. du 01-10-89), stipulates that for a food enzyme to be used in France, either a Declaration de Vente (marketing notification) or a full dossier has to be submitted and approved. The responsibility of the review of the dossier lies with Agence Française de Sécurité Sanitaire des Aliments (AFSSA), the new Food Safety Agency in France established in 2000 (see: http: www.afssa.fr). The regulations list enzymes by principal activity and use, and the authorities also track them by trade name. So, if a manufacturer wishes to change the name or add a use for an already listed enzyme, a Declaration de Vente must be filed and approved prior to sale. One important nuance is that after a dossier is approved, a Declaration de Vente must also be filed to list the enzyme product as an approved food additive on the positive list.

For approval of GMO-derived enzymes, the gene construction description has to be given; the production

organism must be absent in the product and there should be no detectable (transformable) DNA in the product.

The French Enzyme law also sets purity requirements for food enzymes, as follows:

Cadmium	<0.5 mg/kg
Mercury	<0.5 mg/kg
Arsenic	<3 mg/kg
Lead	<10 mg/kg
Total microorganism count	<50.000/g
Salmonella	not present in 25 g
Coliforms	<30/g
Anaerobic SO_2 reducing	<30/g
Staphylococcus aureus	not present in 1 g
Antibiotic activity	not present
Aflatoxins	<0.005 mg/kg.

In the UK, the regulation of enzymes is under the Department of Health (http://www.doh.gov.uk/), which is part of the Department for Environment, Food and Rural Affairs (DEFRA) (http://www.defra.gov.uk/). The Department of Health bases its approval of a "new" enzyme on the basis of its need and safety. The decision of "need" for a new enzyme is made by the Food Advisory Committee (FAC) (http://www.foodstandards.gov.uk/), and the safety is evaluated by the Committee on the Toxicity of Chemicals in Foods, Consumer Products and the Environment (COT) (http://www.foodstandards.gov.uk/committees/cot/summary.htm). COT published guidelines for the safety assessment of microbial enzyme preparations in 1993.

In Denmark, food enzymes are regulated in "Bekendtgørelse om tils ætningsstoffer til levnedsmidler" (Order on Additives for Foodstuffs) No. 942, issued December 11, 1997 by the Danish Ministry of Food, Agriculture and Fisheries (with amendments). The user or importer of an enzyme preparation is responsible for a notification under this scheme. The information to be contained in the notification should follow the guidelines laid down by the SCF in 1992. Notification is valid only for a specific brand or trade name and must be approved by the authorities prior to sale.

In Japan, the Ministry of Health, Labour and Welfare (MHLW) (http://www.mhlw.go.jp/english/) regulates the use of food enzymes. There is a positive list of approved enzymes (for English version, see Food Sanitation Law, Japan Food Additives Association, 1999) and all recombinantly derived enzymes must be approved, except for those derived from self-cloned microorganisms. (The MHLW will need to confirm the self-cloned status and should be contacted to determine information requirements.) The list of currently approved recombinant foods and food additives, including enzymes can be found at http://www.mhlw.go.jp/english/topics/food/index.html. The requirements for the approval of recombinantly derived enzymes can also be found at this site.

A system similar to that in Japan is currently under development in Korea.

Several countries, including Poland, Hungary, Taiwan, and Thailand have registration processes for food enzymes. These processes differ widely in what information is required to be provided prior to importation or sale of the enzyme product. It is important to note that Poland has a lengthy list of specifications that the products must meet, and Korea is developing regulations that will require approval of recombinantly derived food enzyme products.

In the USA, food ingredients, including food enzymes, are either regulated as food additives by the FDA or

are Generally Recognized As Safe (GRAS). To be GRAS, an ingredient either needs to have been in commerce prior to 1958 or to have been determined to be safe for its intended use based upon scientific principles, and scientists knowledge about food safety need to acknowledge that it is safe. If an ingredient is not GRAS, it is regulated as a food additive. GRAS ingredients can be introduced into commerce without FDA approval, through what is commonly called a self-affirmation. FDA can also be notified of the GRAS determination through the voluntary GRAS Notification process (see: http://www.cfsan.fda.gov/~dms/opa-noti.html). FDA proposed the GRAS Notification process on April 17, 1997, and although the regulation has not been finalized, the GRAS Notice process is in use. When a GRAS Notice is filed, the FDA evaluates whether each submitted notice provides a sufficient basis for a GRAS determination and whether information in the notice or otherwise available to FDA raises issues that lead the agency to question whether use of the substance is GRAS. Following this evaluation, FDA responds to the notifier by letter, and posts the letter on the web site.

In general, FDA's response has been in one of three categories:

1. The agency does not question the basis for the notifier's GRAS determination.
2. The agency concludes that the notice does not provide a sufficient basis for a GRAS determination (e.g., because the notice does not include appropriate data and information or because the available data and information raise questions about the safety of the notified substance).
3. The response letter states that the agency has, at the notifier's request, ceased to evaluate the GRAS notice.

For a list of GRAS Notices and issued letters, see: http://www.cfsan.fda.gov/~rdb/opa-gras.html.

Food additives require prior approval by the FDA through a food additive petition (see: http://www.cfsan.fda.gov/~lrd/foodadd.html). The food additive petition process can take several years to result in a regulation indicating approval; food additives must be approved prior to introduction in commerce in the USA.

GMO-derived products are regulated by the same regulatory scheme as non-GMO-derived products. GMO-derived food enzymes can be introduced into commerce through self-affirmations, GRAS Notifications and/or food additive petitions, just as for non-GMO derived food enzymes.

Prior to the implementation of the GRAS Notice process, there was a petition process for GRAS substances in which the petitioner would ask FDA to affirm the product as GRAS. Several enzyme preparations have been affirmed as GRAS and most are listed in the Partial List of Enzymes that are Used in Foods, http://www.cfsan.fda.gov/~dms/opa-enzy.html, which also lists those regulated as food additives.

Products which are regulated by the Bureau of Alcohol, Tobacco, and Firearms (BATF; i.e., beer, wine, and distilled alcohol) and the U.S. Department of Agriculture (USDA; i.e., meat and poultry) require FDA sanction prior to the BATF or USDA approval. Historically, the only way to do this is through the petitioning process and having a regulation finalized on the product. Once the product is regulated by FDA, BATF will allow its use in wine (27 CFR §24.246) and distilled alcohol (27 CFR §24.247) without any further action, but it may need to be added to the Brewer's Adjunct Manual for use in beer. USDA will allow the use through a letter request. Both BATF and USDA have agreed to honor the GRAS Notification process, and that is currently the process a manufacturer would follow to gain the approval of a new GRAS food enzyme that is also regulated by BATF or USDA.

In Canada, Health Canada regulates food-use enzymes under the Food and Drugs Act. Each enzyme for sale in Canada must be listed in the regulation by enzyme activity, the specific source from which it is derived, and the allowed application(s) (see Part B, Div. 16, Table V, Food Additives that May be Used as Food Enzymes; http://www.hc-sc.gc.ca/food-aliment/english/publications/acts_and_regulations/food_and_drugs_acts/c-tables.pdf). Listing is achieved via a food additive petition. Once Health Canada's staff has completed their safety review and are satisfied, it is possible to obtain provisional authorization to sell the food enzyme through an Interim Marketing Authorization. The full regulation of an approved food additive takes several years, with listing of the proposed regulation in Gazette I for comment and publication of the regulation in Gazette II

after approval by Parliament. Submission requirements are similar to those in the USA. GMO regulations are currently being developed under Novel Food Legislation.

Australia and New Zealand regulate food enzymes through the Australia New Zealand Food Authority (ANZFA). Enzymes need to be approved prior to sale; for more information and how to seek approval of an enzyme, label it, etc., see: http://www.anzfa.gov.au/foodstandards/. Approved enzymes are listed in Standard 1.3.3, clauses 15 – 17 and are listed by enzyme activity and source (see: http://www.anzfa.gov.au/foodstandards/foodstandardscodecontents/standard13/standard133.c

7.2.2. Feed-Use Enzymes

Table 41 lists approvals needed by application and country for enzymes used in animal feed and feed processing.

Table 41. Approvals needed by application and country for enzymes used in animal feed and feed processing

Country/Area	
EU	Full dossiers required for new enzymes/products
	Notifications of changes for approved products
USA	GRAS or Feed Additive, AAFCO listing
Canada	Approval required

In the EU, enzymes used as feed additives are regulated under the Feed Additive Directive 70/524/EEC. Details on the criteria and the contents of dossiers to be submitted to achieve approvals for these enzymes are given in Directive 87/153/EEC, as amended by 94/40/EC and 95/11/EC (the so called guidelines directive). In addition to these criteria, a description of the production organism and process needs to be included and the production organism needs to be “on deposit” (i.e., the production organism needs to be deposited with a recognized culture collection); this requirement, along with others, then limits the amount of process improvement the manufacturer can implement without reverting to the authorities for approval. The Scientific Committee for Animal Nutrition (SCAN; see http://europa.eu.int/comm/food/fs/sc/scan/index_en.html) determines the requirements for approval and reviews dossiers for feed enzymes. Approved enzymes are added to Annex I of the directive, along with all feed additives; approval is considered production-strain-specific.

In a publication in August 1998, SCAN took the position that in the GMOs used for future feed enzyme products, antibiotic resistance markers (ARMs) should be avoided or removed. Therefore, it is prudent for industry developing microbial sources for the production of feed enzymes to avoid the use of ARMs. SCAN has also published several other opinions on enzymes used in feed that can be viewed at their web site.

In the USA, ingredients used in feed must be either GRAS and/or listed in the Official Publication of American Feed Control Officials [AAFCO Manual (see also http://www.aafco.org/)] in Table 30. Labels are approved state by state. Under the Enzyme Coordination Policy proposed in 1997 and adopted in 1998, FDA will review new enzyme products for use in feed and recommend listing in the AAFCO Manual via a Letter of No Objection. Requirements for listing are spelled out in the Coordination Policy in the AAFCO Manual. Other related materials that may consist of or contain enzymes are also listed in the AAFCO Manual, §§ 36.11 and 36.12.

In Canada, feed use enzymes are regulated by the Feed Section, Animal Health and Production Division, Canadian Food Inspection Agency. Addition of a new enzyme requires safety and efficacy data, intended labeling, product formulation, and analytical data and methods. Enzymes must be approved prior to use in feed. Instructions for how to apply for feed enzyme approval, which enzymes are approved, and activity-testing criteria can be found in several of the trade memoranda at

http://inspection.gc.ca/english/anima/feebet/trademem/trindxe.shtml.

7.2.3. Industrial-Use Enzymes

Table 42 lists approvals needed by application and country for enzymes used in industrial applications.

Table 42. Approvals needed by application and country for enzymes used in industrial applications

Country/region	
Australia, Canada, EU*, Japan, Korea, Philippines, USA	Listed on Inventories**
Germany, Austria	Dossier for detergent use to obtain registration number

*Enzymes for all uses (food, feed, chemical, etc.) are regulated through this system; some uses, as noted above under food and feed enzymes, also require specific approvals through other regulatory systems.
**For example, European Inventory of New and Existing Chemical Substances (EINECS) and the U.S. Toxic Substances Control Act (TSCA) Inventory, enzymes listed by activity (CAS or IUB No.).

Enzymes used in industrial processes such at textile processing and paper manufacture and as ingredients in laundry and dishwashing detergents are generally regulated as other chemicals are in the countries that have such regulations. They are usually included in an inventory listing, as discussed below. In addition, the ingredients used in such enzyme products also need to be listed on the respective inventories.

In the EU, this regulation is harmonized, and the enzyme activity in question needs to be listed in the European Inventory of Existing Chemical Substances (EINECS) or approved as new through the European List of New Chemical Substances (ELINCS). (See also: http://ecb.ei.jrc.it.) The regulation of these enzymes is covered under Council Directive 67/548/ EEC as amended (for the 7th time by 92/32/EEC), hereafter referred to as ELINCS directive, and implementing national regulations. There are more than 300 enzyme entries in EINECS. Enzymes for all uses, whether used as an ingredient or processing aid, if used in the EU are regulated through this regulatory scheme.

All the enzymes listed in EINECS are currently characterized solely on the basis of their catalytic activity and listed by CAS number; the source is not critical for determination of the EINECS status of any enzyme. This means that enzymes derived from GM microorganisms, including protein-engineered enzymes, are covered by the EINECS listing. The EU Commission is currently examining how enzymes are listed on EINECS, and this regulatory scheme may change in the future.

In the USA, enzymes and the intergeneric microorganisms (i.e., recombinant microorganisms created by combining DNA from more than one taxonomic genus) used to produce them are regulated under the EPA Toxic Substances Control Act (TSCA). EPA TSCA regulations are in 40 CFR parts 700 – 789 (see: http://www.access.gpo.gov/nara/cfr/waisidx_00/40cfrv23_00.html).

While there are over 100 enzymes explicitly listed on the TSCA Inventory by CAS number, current EPA policy (unwritten) is that enzymes are naturally occurring substances and are implicitly on the Inventory. If a manufacturer has intent to manufacture or import an enzyme product and wants to confirm with the Agency that the enzyme is on the Inventory, the manufacturer or importer can file a Bona Fide Intent to Manufacture, per 40 CFR §720.25. The EPA will assess the information provided and make a determination of whether or not the entity that is the subject of the Bona Fide is on the Inventory. This is also the procedure that is used to determine if an entity is in the confidential section of the Inventory.

Intergeneric microorganisms are regulated under the Microbial Products of Biotechnology; Final Regulation Under the Toxic Substances Control Act; Final Rule, published in 1997. For the text of the rule, guidance on how to use it, and a list of microorganisms cleared for use, see: http://www.epa.gov/biotech_rule/index.html. Research and development of enzymes conducted in

the USA is exempt from TSCA under 40 CFR §720.36 for chemicals and 40 CFR §720.234 for contained use intergeneric microorganisms. The intergeneric microorganism is only subject to regulation if its manufacture or R&D is conducted in the United States. If the enzyme product is manufactured outside the USA, and the product does not contain the microorganism when imported, only the enzyme product and formulation ingredients are subject to TSCA.

In Canada, enzymes for these uses need to be listed on the Domestic Substances List (DSL), and microorganisms, whether recombinant or not, used to produce them in Canada also need to be on the DSL. Notification requirements, procedures and the DSL can be found at `http://www.ec.gc.ca/cceb1/nsd/eng/index_e.htm`.

Japan, Australia, Korea, and the Philippines also have chemical inventories that include enzymes and their ingredients. In Japan, enzymes are considered natural and are not explicitly listed on the Inventory. Before import or manufacture of an enzyme product in any of these countries, the competent environmental regulatory authorities should be consulted.

PREVIOUS NEXT

[Top of page] [A to Z] [Authors] [Subjects] [Search] [HitList]

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Wolfgang Aehle, Genencor International B. V., Leiden, The Netherlands
Richard N. Perham, University of Cambridge, Department of Biochemistry, Cambridge, UK
Gerhard Michal, Tutzing, Germany
Albert Jonke, Boehringer Mannheim GmbH, Penzberg, Germany
Gopal K. Chotani, Genencor International, Inc., Palo Alto, USA
Timothy C. Dodge, Genencor International, Inc., Palo Alto, USA
Andreas Herman Terwisscha van Scheltinga, Gibst-brocades N. V., Delft, The Nederlands
Christian Gölker, Bayer AG, Wuppertal, Germany
Meng H. Heng, Genencor International, Inc., Palo Alto, USA
John Kan, Genencor International, Inc., Palo Alto, USA
Todd Becker, Genencor International, Inc., Palo Alto, USA
Saburo Fukui, Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
Atsuo Tanaka, Department of Industrial Chemistry, Faculty of Engineering, Kyoto University, Kyoto, Japan
Hans de Nobel, Genencor International B. V., Leiden, The Netherlands
Brian Jones, Genencor International B. V., Leiden, The Netherlands
Rick Bott, Genencor International, Inc., Palo Alto, USA
Charlotte Horsmans Poulsen, Danisco A/S, Brabrand, Denmark
Karsten M. Kragh, Danisco A/S, Brabrand, Denmark
Jens Frisback Sorensen, Danisco A/S, Brabrand, Denmark
Catherine Grassin, DSM Food Specialties, Seclin Codex, France
Marga Herweijer, DSM Food Specialties, Delft, The Nederlands, France
Jan Wilms, Genencor International B. V., Leiden, The Nederlands
André de Roos, DSM Food Specialities, Delft, The Netherlands
Jorn Borch Soe, Danisco A/S, Brabrand, Denmark
Ivan Herbots, Procter & Gamble Eurocor S.A., Strombeek-Bever, Belgium
Beatrix Kottwitz, Henkel KGaA, Düsseldorf, Germany
Peter J. Reilly, Iowa State University, Ames, USA
Richard L. Antrim, Grain Processing Corporatio, Muscatine, Iowa, USA
Heidi Burrows, Finfeeds International, Marlborough, UK
Herman B. M. Lenting, TNO Industry, Enschede, The Netherlands
Liisa Viikari, VTT Biotechnology, VTT, Finland
Anna Suurnäkki, VTT Biotechnology, VTT, Finland
Marja-Leena Niku-Paavola, VTT Biotechnology, VTT, Finland
Johanna Buchert, VTT Biotechnology, VTT, Finland
Karl-Heinz Maurer, Henkel KGaA, Düsseldorf, Germany
Andrea Saettler, Henkel KGaA, Düsseldorf, Germany
Herbert Waldmann, Max-Planck-Institut für molekulare Physiologie, Dortmund, Germany
Carsten Schultz, EMBL, Heidelberg, Germany
Harald Gröger, Degussa AG, Hanau-Wolfgang, Germany
Carlo Dinkel, SiChem GmbH, Bremen, Germany
Karlheinz Drauz, Degussa AG, Hanau-Wolfgang, Germany
Georg-Burkhard Kresse, Roche Diagnostics GmbH, Penzberg, Germany
Rainer Schmuck, Roche Diagnostics GmbH, Tutzing, Germany
Karl Wulff, Boeringer Mannheim GmbH, Tutzing, Germany
Günther Henniger, Roche Diagnostics GmbH, Penzberg, Germany
Christoph Kessler, Roche Diagnostics GmbH, Penzberg, Germany
Alice J. Caddow, Genencor International, Inc., Palo Alto, USA
Beth Concoby, Genencor International, Inc., Palo Alto, USA

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS NEXT

References

Previous Version (online posting date: June 15, 2000)

Specific References

1. M. Dixon, E. C. Webb: *Enzymes*, Longmans, London 1979.
2. A. R. Fersht: *Structure and Mechanism in Protein Science*, Freeman, New York 1999.
3. C. Walsh: *Enzymatic Reaction Mechanisms*, Freeman, San Francisco 1979.
4. N. C. Price, L. Stevens: *Fundamentals of Enzymology* 5th Edition, Oxford University Press, Oxford 1989.
5. T. Godfrey, S. West (eds.): *Industrial Enzymology*, 2nd ed., The Macmillan Press Ltd., New York, Stockton Press, London, 1996.
6. J. S. Fruton: *Molecules and Life*, John Wiley, New York 1972.
7. H. Gutfreund (ed.): *Enzymes: 100 years. FEBS Letters*, vol. 62, Suppl., North Holland, Amsterdam 1976.
8. D. H. Spackman, W. H. Stein, S. Moore, Anal. Chem. **30** (1958) 1190 – 1206. Link
9. M. Elzinga (ed.): *Methods in Protein Sequence Analysis*, Humana Press, Clifton, New Jersey, 1982.
10. F. Sanger, Science (Washington, D.C.) **129** (1959) 1340 – 1344. Link
11. P. Edman, G. Begg, Eur. J. Biochem. **1** (1967) 80 – 91. Link
12. D. G. Smyth, W. H. Stein, S. Moore, J. Biol. Chem. **238** (1963) 227 – 234. Link
13. R. E. Canfield, J. Biol. Chem. **238** (1963) 2698 – 2707. Link
14. B. E. Davidson, M. Sajgo, H. F. Noller, J. I. Harris, Nature (London) **216** (1967) 1181 – 1185. Link
15. `http://www.expasy.ch/proteomics_def.html`.
16. I. Harris, B. P. Meriwether, J. H. Park, Nature (London) **198** (1963) 154 – 157. Link
17. E. Shaw, Physiol. Rev. **50** (1970) 244 – 296. Link
18. M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead et al., Nature (London) **185** (1960) 416 – 422. Link
19. D. C. Phillips, Proc. Natl. Acad. Sci. U.S.A. **57** (1967) 484 – 495. Link
20. *Enzyme Nomenclature*, Academic Press, New York 1992 and `http://www.chem.qmw.ac.uk/iubmb/`
21. Eur. J. Biochem. **264** (1999) 610 – 650. Link
22. C. H. W. Hirs, S. N. Timasheff (eds.): *Enzyme Structure Part I, Methods in Enzymology*, vol. 91, Academic Press, New York 1983.
23. R. W. Old, S. B. Primrose: *Principles of Gene Manipulation*, Blackwell Scientific Publications, Oxford 1994.
24. T. Maniatis, E. F. Frisch, J. Sambrook: *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1982.
25. L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes et al., Nature (London) **321** (1986) 674 – 679. Link
26. A. Wada, Nature (London) **325** (1987) 771 – 772. Link
27. G. G. Kneale, M. J. Bishop, CABIOS **1** (1985) 11 – 17. Link
28. T. L. Blundell, L. N. Johnson: *Protein Crystallography*, Academic Press, New York 1976.
29. G. E. Schulz, R. H. Schirmer: *Principles of Protein Structure*, Springer Verlag, New York 1990.
30. T. E. Creighton: *Proteins*, Freeman, New York 1993.
31. M. Levitt, C. Chothia, Nature (London) **261** (1976) 552 – 558. Link
32. C. Chothia, Nature **357** (1992) 543 – 544. Link
33. R. Bott, R. Boelens, Curr. Opin. Biotechnol. **10** (1999) no. 4, 391 – 397. Link
34. G. M. Clore, C. D. Schwieters, Curr. Opin. Biotechnol. **12** (2002) no. 2, 146 – 153. Link
35. V. Kanelis, J. D. Forman-Kay, L. E. Kay, IUBMB Life **52** (2001) no. 6, 291 – 302. Link
36. G. M. Clore, A. M. Gronenborn, Curr. Opin. Biotechnol. **2** (1998) no. 5, 564 – 570. Link
37. U. Heinemann, G. Illing, H. Oschkinat, Curr. Opin. Biotechnol. **12** (2002) no. 4, 348 – 354. Link

38. R. Porter, M. O'Connor, J. Whelan (eds): *CIBA Foundation Symposium 93, Mobility and Function in Proteins and Nucleic Acids*, Pitman, London 1983.
39. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne: "The Protein Databank", Nucleic Acids Research, **28** (2000) 235 – 242. Link
40. R. N. Perham, Philos. Trans. R. Soc. London B **272** (1975) 123 – 136. Link
41. R. Jaenicke (ed.): *Protein Folding*, Elsevier, Amsterdam 1980.
42. K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, T. R. Cech, Cell **31** (1982) 147 – 157 Link
43. Guerrier-takada, K. Gardiner, T. Marsh, N. Pace, S. Altman, Cell **35** (1983) 849 – 857. Link
44. R. Saldanha, G. Mohr, M. Belfort, A. M. Lambowitz, FASEB J. (1993) 15 – 24. Link
45. C. Guerrier-Takada, S. Altman, Science **223** (1984) 285 – 286. Link
46. A. J. Zaug, T. R. Cech, Science **231** (1986) 470 – 475. Link
47. N. Ban, P. Nissen, J. Hansen, P. B. Moore, Steitz TA Science **289** (2000) 905 – 920. Link
48. P. Nissen, J. Hansen, N. Ban, P. B. Moore, T. A. Steitz, Science **289** (2000) 920 – 930. Link
49. E. C. Böttger, B. Springer, T. Prammananan, Y. Kidan, P. Sander, EMBO Rep. **4** (2001). Link
50. G. J. Narlikar, D. Herschlag, "Mechanistic Aspects of Enzymatic Catalysis: Lessons from Comparison of RNA and Protein Enzymes", Annu. Rev. Biochem. **66** (1997) 19 – 59. Link
51. K. J. Hertel, A. Peracchi, O. C. Uhlenbeck, D. Herschlag, "Use of Intrinsic Binding Energy for Catalysis by an RNA Enzyme." Proc. Natl. Acad. Sci. U.S.A. **94** (1997) 8497 – 8502. Link
52. J. W. Szostak, D. P. Bartel, P. L. Luisi, Nature **409** (2001) 387 – 390. Link
53. T. B. Campbell, B. A. Sullenger, Adv. Pharmacol. **33** (1995) 143 – 178. Link
54. J. Darnell, H. Lodish, D. Baltimore: *Molecular Cell Biology*, Scientific American Books, New York 1990.
55. D. Freifelder: *Molecular Biology*, Jones and Bartlett, Boston 1987.
56. A. Kornberg: *DNA Replication*, 2nd ed., W. H. Freeman, New York 1992.

Specific References

57. H. U. Bergmeyer, J. Bergmeyer, M. Graßl (eds.): *Methods of Enzymatic Analysis*, 3rd ed., **I – XII**, Verlag Chemie, Weinheim 1983 – 1986.
58. P. D. Boyer (ed.): *The Enzymes*, 3rd ed., Academic Press, New York 1970.
59. D. Glick (ed.): *Methods of Biochemical Analysis*, vols. 1 – 30 will be continued, Interscience Publ., New York 1954 – 1984.
60. K. Musil, O. Novakova, K. Kunz: *Biochemistry in Schematic Perspective*, Avicenum Czechoslovak Medical Press, Prague 1977.
61. L. Michaelis, M. L. Menten: "Die Kinetik der Invertinwirkung," Biochem. Z. **49** (1913) 333 – 369. Link
62. H. U. Bergmeyer, J. Bergmeyer, M. Graßl (eds.): *Methods of Enzymatic Analysis*, 3rd ed., **I** Verlag Chemie, Weinheim 1983, pp. 69 – 85
63. H. U. Bergmeyer, J. Bergmeyer, M. Graßl (eds.): *Methods of Enzymatic Analysis*, 3rd ed., **I** Verlag Chemie, Weinheim 1983, pp. 109 – 114
64. Boehringer Mannheim: *Biochemicals for the Diagnostics Industry* (1984).
65. J. L. Webb: *Enzyme and Metabolic Inhibitors*, vol. I, Academic Press, New York-London 1963.
66. P. D. Boyer (ed.): *The Enzymes*, 3rd ed., **1**, Academic Press, New York 1970. pp. 341 – 396
67. H. U. Bergmeyer, J. Bergmeyer, M. Graßl (eds.): *Methods of Enzymatic Analysis*, 3rd ed., **V**, Verlag Chemie, Weinheim 1983 – 1986. pp. 86 – 98
68. [57] vol. I, pp. 7 – 14
69. Boehringer Mannheim: *Biochemicals for Molecular Biology*, p. 83 (1985).
70. P. Modrich, R. J. Roberts in St. M. Linn, R. J. Roberts (eds.): *Nucleases*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1982, p. 109.
71. [57] vol. I, pp. 280 – 305

72. [57] vol. III,pp. 118 – 126
73. [57] vol. III, pp. 126 – 133
74. [57] vol. III, pp. 496 – 501
75. [57] vol. IV, pp. 75 – 86
76. [57] vol. I, pp. 210 – 221
77. P. Trinder: "Determination of Glucose in Blood Using Glucose Oxidase with an Alternative Oxygen Acceptor," Ann. Clin. Biochem. **12** (1969) 24 – 27. Link
78. [57] vol. I, pp. 326 – 340
79. [59] vol. 17, pp. 189 – 285
80. [57] vol. I, pp. 340 – 368
81. [57] vol. I, pp. 368 – 397
82. Th. Bücher, H. Hofner, J.-F. Ronayrenc: "Methods with Glass Electrodes," in H. U. Bergmeyer (ed.): *Methods of Enzymatic Analysis*, 2nd ed., vol. I, Verlag Chemie, Weinheim, and Academic Press, New York 1974, pp. 254 – 261.
83. [57] vol. IV, pp. 15 – 25
84. H. Bakala, J. Wallach, M. Hauss: "Determination of Elastolytic Activity Using a Conductometric Method," Biochimie **60** (1978) 1205 – 1207. Link
85. A. J. Lawrence, G. R. Moores: "Conductimetry in Enzyme Studies," Eur. J. Biochem. **24** (1972) 538 – 546. Link
86. J. K. Grime: "Biochemical and Clinical Analysis by Enthalpimetric Measurements – A Realistic Alternative Approach?" Anal. Chim. Acta **118** (1980) 191 – 225. Link
87. J. M. Bailey, P. H. Fishman, P. G. Pentchev: "Studies on Mutarotases," J. Biol. Chem. **242** (1967) 4263 – 4269. Link
88. K. Kusai, J. Sekuzu, B. Hagihara, K. Okunuki *et al.*, "Crystallization of Glucose Oxidase from Penicillium amagasakiense," Biochim. Biophys. Acta **40** (1960) 555 – 557. Link
89. S. J. Bach, J. D. Killip: "Purification and Crystallization of Arginase," Biochim. Biophys. Acta **29** (1958) 273 – 280. Link
90. [57] vol. IV, pp. 180 – 183
91. [57] vol. I, pp. 142 – 163.
92. Boehringer Mannheim: Assay Instruction Penicillin-G Amidase (carrier fixed) (1980). Link
93. [57] vol. II, pp. 102 – 115.
94. [57] vol. II, pp. 86 – 99.
95. O. Warburg, W. Christian: "Isolierung und Kristallisation des Gärungsferments Enolase," Biochim. Z. **310** (1941) 384 – 421. Link
96. J. R. Whitaker, P. E. Granum: "An Absolute Method for Protein Determination Based on Difference of Absorbance at 235 and 280 nm," Anal. Biochem. **109** (1980) 156 – 159. Link
97. G. Beisenherz, H. J. Boltze, Th. Bücher, R. Czok *et al.*, "Diphosphofructose-Aldolase, Phosphoglyceraldehyd-Dehydrogenase, Milchsäure-Dehydrogenase, Glycerophosphat-Dehydrogenase und Pyruvat-Kinase aus Kaninchenmuskulatur in einem Arbeitsgang," Z. Naturforsch. B Anorg. Chem. Org. Chem. **8 B** (1953) 555 – 577. Link
98. O. H. Lowry, N. J. Rosebrough, A. L. Farr, R. J. Randall: "Protein Measurement with the Folin Phenol Reagent," J. Biol. Chem. **193** (1951) 265 – 275. Link
99. M. M. Bradford: "A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding," Anal. Biochem. **72** (1976) 248 – 254. Link
100. H. R. Maurer: *Disk-Elektrophorese*, 2nd ed., De Gruyter, Berlin 1971.
101. U. K. Laemmli: "Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T 4," *Nature* (London) 227 (1970) 680 – 685.
102. M.-E. Mirault, K. Scherrer, L. Hansen: "Isolation of Preribosomes from Hela Cells and Their Characterization by Electrophoresis on Uniform and Exponential-Gradient-Polyacrylamide Gels," Eur. J. Biochem. **23** (1971) 372 – 386. Link

103. C. R. Merril, D. Goldmann, M. L. van Keuren: "Silver Staining Methods for Polyacrylamide Gel Electrophoresis," in *Methods in Enzymology*, vol. 96, Academic Press, New York 1983, p. 17.
104. G. E. Briggs, J. B. S. Haldane: Biochem J. **19** (1925) 338 – 339. Link
105. H. Lineweaver, D. Burk: J. Am. Chem. Soc. **56** (1934) 658 – 666. Link
106. Z. H. Beg, J. A. Stonik, H. B. Brewer, Proc Nat. Acad. Sci. (USA) **75** (1978) 3678 – 3682. Link
107. G. Gil, F. G. Hegardt, Arch. Biochem. Biophys. **214** (1982) 192 Link
108. H. U. Bergmeyer: *Methods in Enzymatic Analysis*, 3rd ed., vol. **2**, VCH Verlagsgesellschaft, Weinheim 1983 166 – 167.
109. H. U. Bergmeyer: *Methods in Enzymatic Analysis*, 3rd ed., vol. **4**, VCH Verlagsgesellschaft, Weinheim 1983 189 – 195.
110. M. Smith *et al.*, Anal. Biochem. **150** (1985) 76 – 85. Link
111. H. Schägger in R. Kellner (ed.): *Microcharacterization of Proteins*, VCH Verlagsgesellschaft, Weinheim 1994.
112. R. Aebersold: *Advances in Electrophoresis*, VCH Verlagsgesellschaft, Weinheim 1991 , pp. 81 – 86.
113. User manual for the Phast System. Pharmacia LKB Biotechnology, Freiburg.
114. R. M. Chiz in M. P. Deutscher (ed.): *Methods in Enzymology*, vol. **182**, Academic Press, London 1990, pp. 392 – 421.
115. P. R. Sibbald *et al.*, Anal. Biochem. **198** (1991) 330 – 333. Link
116. P. Edman, Acta Chem. Scand. **4** (1950) 283 – 290. Link
117. D. L. Crimmings *et al.*, *Techniques in Protein Chemistry* vol. **3**, Academic Press, London 1992, 35 – 43.
118. M. D. Trevan, *Immobilized Enzymes*, Wiley, Chichester 1980.

Specific References

119. S. J. Pirt: *Principles of Microbe and Cell Cultivation*, Blackwell Scientific Publ., Oxford 1975.
120. J. R. Norris, D. W. Ribbons: *Methods in Microbiology*, vols. 1 – 8, Academic Press, London 1969 – 1973.
121. H. J. Rehm: *Industrielle Mikrobiologie*, 2nd. ed., Springer Verlag, Berlin 1980.
122. W. Crueger, A. Crueger: *Biotechnology: A Textbook of Industrial Microbiology*, Science Tech., Madison 1984.
123. M. Moo-Young (ed.): *Comprehensive Biotechnology*, parts 1 – 4, Pergamon Press, Oxford 1985.
124. H. J. Rehm, G. Reed (eds.): *Biotechnology, A Comprehensive Treatise in Eight Volumes*, VCH Verlagsgesellschaft, Weinheim 1985.
125. A. L. Sonenshein, J. A. Hoch, R. Losick (eds.): *Bacillus subtilis and Other Gram-Positive Bacteria: Biochemistry, Physiology, and Molecular Genetics*, ASM Press, 1993.
126. J. E. Smith (ed.): *Aspergillus*, Plenum Press, 1994.
127. G. E. Harman, C. P. Kubicek (eds.): *Trichoderma and Gliocladium v.2: Enzymes, Biological Control and Commercial Applications*, Taylor & Francis, 1998.
128. B. Alberts, D. Bray, J. Lewis , M. Raff *et al.*: *Molecular Biology of the Cell*, Garland Publ., New York 1983.
129. R. Walker: *The Molecular Biology of Enzyme Synthesis*, J. Wiley & Sons, New York 1983.
130. L. L. Randall, S. J. S. Hardy, Microbiol. Rev. **48** (1984) 290 – 298. Link
131. B. P. Wasserman, Food Technol. **1984**, 78 – 98. Link
132. S. Gottesman, Annu. Rev. Genet. **18** (1984) 415 – 441. Link
133. G. Holt, G. Saunders in M. Moo-Young (ed.): *Comprehensive Biotechnology*, part 1, Pergamon Press, Oxford 1985, pp. 51 – 76.
134. K. Aunstrup, Abh. Akad. Wiss. DDR Abt. Math. Naturwiss. Tech. 1981, 445 – 457 (3N, Mikrob. Enzymprod.). Link
135. D. E. Eveleigh, B. S. Montenecourt, Adv. Appl. Microbiol. **25** (1979) 57 – 74. Link

136. T. J. White, J. H. Meade, S. P. Shoemaker, K. E. Koths, M. A. Innis, Food Technol. **1984**, 90 – 98. Link
137. A. L. Demain, Adv. Appl. Microbiol. **16** (1973) 177 – 202. Link
138. W. Harder, J. G. Kuenen, A. Matin, J. Appl. Bacteriol. **43** (1977) 1 – 24. Link
139. T. J. R. Harris, J. S. Emtage, Microbiol. Sci. **3** (1986) 28 – 31. Link
140. T. J. Silhavy, S. A. Benson, S. D. Emr, Microbiol. Rev. **47** (1983) 313 – 344. Link
141. Gist-brocades, EP 0 096 430, 1983 (C. P. Hollenberg, A. de Leeuw, S. Pas, J. A. van den Berg).
142. D. Naki, C. Paech, G. Ganshaw, V. Schellenberger, Appl. Microbiol. Biotechnol., **29** (1998) 290 – 294. Link
143. J. Monod in P. S. S. Dawson (ed.): *Microbiol. Growth*, Halsted Press, Dowden, Hutchinson and Ross, Inc., Strandsburg, Penn., 1974, pp. 88 – 110.
144. D. K. Button, Microbiol. Rev. **49** (1985) 270 – 297. Link
145. A. G. Williams, S. E. Withers, Appl. Microbiol. Biotechnol. **22** (1985) 318 – 324. Link
146. J. Greenman, K. T. Holland, J. Gen. Microbiol. **131** (1985) 1619 – 1624. Link
147. E. I. Emanuilova, K. Toda, Appl. Microbiol. Biotechnol. **19** (1984) 301 – 305. Link
148. J. Frankena, G. M. Koningstein, H. W. van Verseveld, A. H. Stouthamer, Appl. Microbiol. Biotechnol. **24** (1986) 106 – 112. Link
149. J. E. Bailey: “Reflections on the Scope and the Future of Metabolic Engineering and its Connections to Functional Genomics and Drug Discovery”, Metab. Eng. **3** (2001) 111 – 114. Link
150. A. Varma, B. Palsson, “Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use”, Biotechnology **12** (1994) 994 – 998. Link
151. S. Aiba, A. E. Humphrey, N. F. Millis: *Biochemical Engineering*, 2nd ed., Academic Press, New York 1973.
152. B. Atkinson, F. Mavituna, *Biochemical Engineering and Biotechnology Handbook*, John Wiley & Sons, New York 1992.
153. K. van't Riet, J. Tramper, *Basic Bioreactor Design*, Marcel Dekker, New York, 1991.
154. A. E. Humphrey in M. Moo-Young (ed.): *Comprehensive Biotechnology*, vol. 2, Pergamon Press, Oxford 1985, pp. 3 – 4.
155. J. C. van Suijdam: *Mycelial Pellet Suspensions*, Thesis, Delft University Press 1980.
156. M. E. Groep, M. E. Gregory, L. S. Kershenbaum, I. D. L. Bogle, Biotechnol. Bioeng. **67** (2000) 300 – 311. Link
157. J. M. B. Cabral, B. Casala, C. L. Cooney, Biotechnol. Lett. **7** (1985) 749 – 752. Link
158. J. L. Montesino, M. A. Gordillo, F. Valero, J. Lafuente, C. Sola, B. Valdman, J. Biotechnol. **52** (2000) 207 – 218. Link
159. S. Linko, J. Luopa, Y.-H. Zhu, J. Biotechnol., **52** (1997) 257 – 266. Link
160. K. Kovarova-Kovar, S. Gehlen, A. Kunze, T. Keller, R. von Daniken, M. Kolb, A. P. G. M. van Loon, J. Biotechnol. **79** (2000) 39 – 52. Link
161. J. Schuber, R. Simutis, M. Dors, I. Havlik, A. Lubbert, J. Biotechnol., **35** (1994) 51 – 68. Link
162. R. Simutis, A. Lubbert, Biotechnol. Prog., **13** (1997) 479 – 487. Link
163. T. Yamanè, S. Shimizu, Adv. Biochem. Eng. Biotechnol. **30** (1984) 147 – 194. Link
164. G. Skøt: “Proc. Conf. on Advances in Fermentation,” *Chelsea Coll.*, 21 – 23 Sept. 1983, University of London, Wheatland Journ. Ltd., Penn House 1983, pp. 154 – 159.
165. D. Herbert in P. S. S. Dawson (ed.): *Microbiol. Growth*, Halsted Press, Dowden, Hutchinson and Ross, Inc., Strandsburg, Penn., 1974, pp. 230 – 262.
166. S. Aiba, Biotechnol. Bioeng. Symp. **9** (1979) 269 – 281. Link
167. A. R. Moreira, G. van Dedem, M. Moo-Young, Biotechnol. Bioeng. Symp. **9** (1979) 179 – 203. Link
168. M. J. Rolf, H. C. Lim in M. Moo-Young (ed.): *Comprehensive Biotechnology*, Pergamon Press, Oxford 1985, pp. 165 – 174.
169. D. J. Clarke, M. R. Calder, R. J. G. Carr, B. C. Blake-Coleman, S. C. Moody, T. A. Collinge, Biosensors **1** (1985) 213 – 320. Link

170. C. L. Cooney, Biotechnol. Bioeng. Symp. **9** (1979) 1 – 11. Link
171. A. Rothberg A., J. Weegar, R. von Schalien, K. Fagervik, M. Rydström, K. Lind, Bioprocess Eng. **21** (1999) 307 – 312. Link

General References

172. S. Aiba, A. E. Humphrey, N. F. Mills: *Biochemical Engineering*, Chap. 13, Academic Press, New York-London 1973, pp. 346 – 392.
173. M. Cheryan: *Ultrafiltration Handbook*, Technomic Publishing Company, Inc., Lancaster, Pennsylvania, 1986.
174. C. L. Cooney (ed.): *Comprehensive Biotechnology* vol. **2**, Sect. 2, Pergamon Press, Oxford-New York 1985, pp. 231 – 590.
175. H. Determann: *Gelchromatographie*, Springer Verlag, Berlin-Heidelberg-New York 1969.
176. T. J. C. Gribnau, J. Visser, R. J. F. Nivard: "Affinity Chromatography and Related Techniques," Anal. Chem. Symp. Ser. **9**, 1982 . Link
177. J.-C. Janson, P. Hedman: "Large-Scale Chromatography of Proteins," Adv. Biochem. Eng. **25** (1982) 43 – 99. Link
178. T. Kremmer, L. Boross: *Gel Chromatography*, "Theory, Methodology, Applications," J. Wiley & Sons, New York 1979.
179. G. Kopperschläger, H.-J. Böhme, E. Hofmann: "Cibachrom Blue F 3 G-A and Related Dyes as Ligands in Affinity Chromatography," Adv. Biochem. Eng. **25** (1982) 101 – 138. Link
180. R. M. Lafferty (ed.): *Enzyme Technology*, Springer Verlag, Berlin-Heidelberg-New York 1983.
181. C. J. O. R. Morris, P. Morris: *Separation Methods in Biochemistry*, Pitman Publishing, 1976.
182. R. Scopes: *Protein Purification*, "Principles and Practice," Springer Verlag, New York-Heidelberg-Berlin 1982.
183. L. Svarovsky (ed.): *Solid-liquid separation*, Butterworths, London 1981.
184. J. Turkova: "Affinity Chromatography," J. Chromatogr. Libr. **12** (1978) . Link
185. I. C. D. Wang, C. L. Cooney, A. L. Demain, P. Dunnill, E. A. Humphrey, M. D. Lilly: *Fermentation and Enzyme Technology*, Chap. 12, J. Wiley & Sons, New York 1979, pp. 238 – 310.
186. A. Wisemann (ed.): *Handbook of Enzyme Biotechnology*, J. Wiley & Sons, New York 1985.
187. C. Yang, G. T. Tsao: "Affinity Chromatography," Adv. Biochem. Eng. **25** (1982) 19 – 42. Link

Specific References

188. H. Schutt, DE-OS 2 509 482, 1976.
189. Brewer *et al.*, Trends in Biotechnol. **3** (1985) 119 – 122. Link
190. J. A. Currie, P. Dunnill, M. D. Lilly, Biotechnol. Bioeng. **14** (1972) 725 – 736. Link
191. J. R. Woodrow, A. V. Quirk, Enzyme Microb. Technol. **4** (1982) 385 – 389. Link
192. H. Schütte, K. H. Kroner, H. Hustedt, M. R. Kula in R. M. Lafferty (ed.): *Enzyme Technology*, Springer Verlag, Berlin 1983, pp. 115 – 124.
193. H. Schütte, M.-R. Kula, Biotech-Forum **2** (1986) 68 – 79. Link
194. H. Schütte R. Kraume-Flügel, M.-R. Kula, Chem. Ing. Tech. **57** (1985) 626 – 627. Link
195. A. Wiseman, Process Biochem. **4** (1969) 63 – 65. Link
196. J. A. Asenjo, B. A. Andrews, J. B. Hunter, S. LeCorre, Process Biochem. **20** (1985) 158 – 164. Link
197. K. S. Lam, J. W. D. GrootWassink, Enzyme Microb. Technol. **7** (1985) 239 – 242. Link
198. G. B. Tanny, D. Mirelman, T. Pistole, Appl. Environ. Microbiol. **40** (1980) 269 – 273. Link
199. K. H. Kroner, Biotech-Forum **3** (1986) 20 – 30. Link
200. M. S. Lee, T. Atkinson, Process Biochem. **20** (1985) 26 – 31. Link
201. R. A. Erickson, Chem. Eng. Prog. **91** (1984) Dec, 51 – 54. Link
202. H. Hemfort, W. Kohlstette, Chem. Ind. London **37** (1985) 412 – 416. Link
203. M.-R. Kula, Appl. Biochem. Bioeng. **2** (1979) 71 – 95. Link

204. H. Hustedt, K. H. Kroner, U. Menge, M.-R. Kula, Enzyme Eng. **5** (1980) 45 – 47. Link
205. M.-R. Kula, K. H. Kroner, H. Hustedt, H. Schütte: "Biochemical Engineering II," Ann. N.Y. Acad. Sci. **369** (1981) 341 – 354. Link
206. A. Veide, A.-L. Smeds, S.-O. Enfors, Biotechnol. Bioeng. **25** (1983) 1789 – 1800. Link
207. H. Hustedt, K. H. Kroner, H. Schütte, M.-R. Kula, Biotech-Forum **1** (1985) 10 – 17. Link
208. P. A. Albertson: *Partition of Cell Particles and Macromolecules*, Wiley-Interscience, New York 1971.
209. S. D. Flanagan, S. H. Barondes, J. Biol. Chem. **250** (1975) 1484 – 1489. Link
210. G. Kopperschläger, G. Johansson, Anal. Biochem. **124** (1982) 117 . Link
211. A.-K. Frej, J. G. Gustafsson, P. Hedman, Biotechnol. Bioeng. **28** (1986) 133 – 137. Link
212. J. Baustita, E. Chico, A. Machado, Biotechnol. Lett. **8** (1986) 315 – 318. Link
213. I. Persson, B. Lindman, "Flocculation of Cell Debris for Improved Separation by Centrifugation", in Y. A. Attia (ed.): *Flocculation in Biotechnology and Separation Systems*, Elsevier, Amsterdam, 1987, pp. 457 – 466.
214. A. C. Bentham, J. Bonnerjea, C. C. Orsborn, P. N. Ward, M. Hoare, Biotechnol. Bioeng. **36** (1990) 397 – 401. Link
215. P. Bajpai, R. K. Gera, P. K. Bajpai, J. Ferment. Bioeng. **71** (1991) 284 – 285 Link
216. V. Sitkey, M. Minarik, P. Michalik, Biotechnol. Techn. **6** (1992) 49 – 52. Link
217. V. Sitkey, E. Kusmova, R. Horenitzky, P. Michalik, M. Minarik, Acta Biotechnol. **12** (1992) 299 – 302. Link
218. H. Theelen, H. Hustedt in *Downstream Processing in Biotechnol. II* (1989) 1.37 – 1.43.
219. F. Halverson, H. P. Panzer in: *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd ed., Vol. 10, John Wiley & Sons, New York 1978, pp. 489 – 522.
220. H. I. Heitner in: *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., Vol. 11. John Wiley & Sons, New York 1992, 61 – 80.
221. J. Gregory, J. Coll. Interface Sci. **42** (1973) 448 – 456. Link
222. G. P. Treweek, J. J. Morgan, J. Coll. Interface Sci. **60** (1977) 258 – 273. Link
223. P. Milburn, J. Bonnerjea, M. Giare, P. Dunnill, Enzyme Microbiol Technol. **12** (1990) 527 – 532. Link
224. J. Rumpus, M. Bulmer, N. Titchener-Hooker, M. Hoare, *Biochem. Eng. 2001, Proc. Asia-Pac. Biochem. Eng. Conf.* (1992) 513 – 516, Springer, Tokyo, Japan.
225. D. E. Salt, S. Hay, O. R. T. Thomas, M. Hoare, P. Dunnill, Enzyme Microbiol. Technol. **17** (1995) 107 – 113 Link
226. J. G. Shan, J. Xia, Y. X. Guo, X. Q. Zhang, J. Biotechnol. **49** (1996) 173 – 178. Link
227. M. Mayer, R. Woernle, Chem. Ing. Tech. **57** (1985) 152 – 153. Link
228. D. J. Bell, M. Hoare, P. Dunnill, Adv. Biochem. Eng. **26** (1982) 1 – 72. Link
229. M. Hoare, P. Dunnill, D. J. Bell, Ann. N.Y. Acad. Sci. **413** (1983) 254 – 269. Link
230. P. Dunnill, M. D. Lilly, Biotechnol. Bioeng. Symp. **3** (1972) 97. Link
231. H. Strathmann, Chem. Tech. **7** (1978) 333 – 346. Link
232. *Ullmann*, 4th ed., **16**, 515 – 535.
233. H.-D. Saier, H. Strathmann, Angew. Chem. Int. Ed. Engl. **14** (1975) 452 – 459. Link
234. P. A. Belter, E. L. Cussler, W.-S. Hu: *Bioseparations, Downstream Processing for Biotechnology*, Wiley-Interscience, New York, 1988.
235. R. K. Scopes: *Protein Purification, Principles and Practice*, Springer, New York, 1993.
236. A. McPherson: *Crystallization of Biological Macromolecules*, Cold Spring Harbor Laboratory Press, New York 1998.
237. M. R. Ladisch: *Bioseparations Engineering*, Wiley-Interscience, New York 2000.
238. J. Fukumoto, M. Iwai, Y. Tsujisaka, J. Gen. Appl. Microbiol. **9** (1963) 353 – 361. Link
239. A. Kitazono, T. Yoshimoto, D. Tsuru, J. Bacteriol. **174** (1992) 7919 – 7925. Link

240. J. E. Pitts, J. M. Uusitalo, D. Mantafounis, P. G. Nugent, D. D. Quinn, P. Orprayoon, M. E. Penttila, J. Biotechnol. **28** (1993) 69 – 83. Link
241. M. H. Heng, N. T. Becker, E. L. Braunstein, R. C. Fewkes, H. Gros, Abstr. Pap Am. Chem. Soc. **21** (1996) Biotechnology Abstr. 177. Link
242. PCT WO 97/15660, 1997 (N. T. Becker, E. L. Braunstein, R. Fewkes, H. Meng).
243. H. Gros. Abstr. Pap Am. Chem. Soc. **21** (1996) Biotechnology Abstr 144. Link
244. US 5041377, 1991 (T. Becker, V. B. Lawlis Jr).
245. P. M. Harrison, L. P. Nelles, Fermentations **113** (1990) 585. Link
246. R. A. Judge, M. R. Johns, E. T. White, Biotechnol. Bioeng. **48** (1995), 316 – 323. Link
247. M. H. Heng, *Abstr. Pap. Recovery of Biological Products IX*, May 23–28, 1999.
248. US 6207437, 2001 (E. H. Gros, J. L. Cunefare).
249. P. Mattock, G. F. Aitchison, EP 82/3 015 706, 1982.
250. D. E. G. Austen, T. Cartwright, C. H. Dickerson, Vox Sang. **44** (1983) 151 – 155. Link
251. P. Andrews in E. Glick (ed.): *Methods of Biochem. Anal.*, vol. **18**, Interscience, New York-London 1970, pp. 1 – 53.
252. S. Shaltiel, Z. Er-el, Proc. Nat. Acad. Sci. U.S.A. **70** (1973) 778 – 781. Link
253. J. Porath, J. Carlson, I. Olsson, G. Belfrage, Nature (London) **258** (1975) 598 – 599. Link
254. J. Carlsson, I. Olsson, R. Axen, H. Drevin, Acta Chem. Scand. Ser. B **30** (1976) 180 – 182. Link
255. J.-C. Janson, P. Dunnill, Proc. FEBS Meet. **30** (1974) 81 – 105. Link
256. J.-C. Janson, J. Agr. Food Chem. **19** (1971) 581 – 588. Link
257. K. Clarkson, B. Jones, R. Bott, B. Bower, G. Chotani, T. Becker: "Enzymes: Screening, Expression, Design and Production" in M. R. Bedford, G. G. Partridge (eds.): *Enzymes in Farm Animal Nutrition*, CAB International, pp. 315 – 352.
258. T. Becker, G. Park, A. L. Gaertner: "Formulation of Detergent Enzymes" in J. H. Van Ee, O. Misset, E. Baas (eds.): *Enzymes in Detergency*, Marcel Dekker, New York, pp. 299 – 325.
259. R. DePaz, C. C. Barnett, D. A. Dale, J. F. Carpenter, A. L. Gaertner, T. W. Randolph: "The Excluding Effects of Sucrose on a Protein Chemical Degradation Pathway: Methionine Oxidation in Subtilisin," Arch. Biochem. Biophys. **384** (2000) 123 – 132. Link
260. T. Becker: "Separation and Purification Processes for Recovery of Industrial Enzymes" in R. K. Singh, S. S. H. Rizvi (eds.): *Bioseparation Processes in Foods*, Marcel Dekker, New York, pp. 427 – 445.

Specific References

261. I. Chibata (ed.): *Immobilized Enzymes, Research and Development*, Kodansha, Tokyo 1978.
262. M. D. Lilly, Enzyme Microb. Technol. **8** (1986) 315. Link
263. J. M. Nelson, E. G. Griffin, J. Am. Chem. Soc. **38** (1916) 1109. Link
264. N. Grubhofer, L. Schleith, Naturwissenschaften **40** (1953) 508. Link
265. M. A. Mitz, Science **123** (1956) 1076. Link
266. P. Bernfeld, J. Wan, Science **142** (1963) 678. Link
267. F. A. Quiocho, F. M. Richards, Proc. Natl. Acad. Sci. **52** (1964) 833. Link
268. T. M. S. Chang, Science **146** (1964) 524. Link
269. G. Gregoriadis, P. D. Leathwood, B. E. Ryman, FEBS Lett. **14** (1971) 95. Link
270. E. Katchalski, I. H. Silman, R. Goldman, Adv. Enzymol. **34** (1971) 445. Link
271. I. Chibata, T. Tosa, T. Sato, T. Mori, T. Matuo, *Proc. IVth Int. Ferment. Symp.*, "Fermentation Technology Today," Soc. Ferment. Technol., Japan 1972, p. 383.
272. K. Mosbach (ed.): *Methods in Enzymology*, vol. **44**, Academic Press, New York 1976. K. Mosbach (ed.): *Methods in Enzymology*, vol. **135**, Academic Press, New York 1987.
273. R. Axén, J. Porath, S. Ernback, Nature (London) **214** (1967) 1302 Link P. Axén, S. Ernback, Eur. J. Biochem. **18** (1971) 351.

274. M. A. Mitz, L. J. Summaria, Nature (London) **189** (1961) 576. Link W. Brümmer, N. Hennrich, M. Klockow, H. Lang, H. D. Orth, Eur. J. Biochem. **25** (1972) 129.
275. N. Weliky, F. S. Brown, E. C. Dale, Arch. Biochem. Biophys. **131** (1969) 1. Link
276. P. Liu-Osheroff, R. J. Guillory, Biochem. J. **127** (1972) 419. Link
277. H. Filippusson, W. E. Hornby, Biochem. J. **120** (1970) 215. Link
278. T. Sato, T. Mori, T. Tosa, I. Chibata, Arch. Biochem. Biophys. **147** (1971) 788. Link
279. G. Kay, E. M. Crook, Nature (London) **216** (1967) 514. Link
280. R. D. Falb, J. Lynn, J. Shapira, Experientia **28** (1973) 958. Link
281. M. Kierstan, C. Bucke, Biotechnol. Bioeng. **19** (1977) 387. Link
282. I. Takata, T. Tosa, I. Chibata, J. Solid-Phase Biochem. **2** (1977) 225. Link
283. T. Tosa, T. Sato, T. Mori, K. Yamamoto, I. Takata, Y. Nishida, I. Chibata, Biotechnol. Bioeng. **21** (1979) 1967. Link
284. T. Sato, Y. Nishida, T. Tosa, I. Chibata, Biochim. Biophys. Acta **570** (1979) 179. Link
285. I. Takata, K. Yamamoto, T. Tosa, I. Chibata, Enzyme Microb. Technol. **2** (1980) 30. Link
286. S. Fukui, A. Tanaka, Adv. Biochem. Eng./Biotechnol. **29** (1984) 1. Link
287. S. Fukui, A. Tanaka, T. Iida, E. Hasegawa, FEBS Lett. **66** (1976) 179. Link
288. S. Fukushima, T. Nagai, K. Fujita, A. Tanaka, S. Fukui, Biotechnol. Bioeng. **20** (1978) 1465. Link
289. S. Fukui, A. Tanaka, Endeavour **9** (1985) 10. Link
290. The Working Party on Immobilized Biocatalysts, Enzyme Microb. Technol. **5** (1983) 304. Link
291. I. Chibata, T. Tosa, T. Sato, Appl. Microbiol. **27** (1974) 878. Link
292. K. Yamamoto, T. Tosa, I. Chibata, Biotechnol. Bioeng. **22** (1980) 2045. Link
293. K. Yamamoto, T. Tosa. K. Yamashita, I. Chibata, Eur. J. Appl. Microbiol. **3** (1976) 169. Link
294. Y. Asano, T. Yasuda, Y. Tani, H. Yamada, Agric. Biol. Chem. **46** (1982) 1183. Link
295. S. Fukui, A. Tanaka, Annu. Rev. Microbiol. **36** (1982) 145. Link

Specific References

296. A. T. Bull, “Isolation and Screening of Industrially Important Organisms”, in F. Vardar-Sukan, S. S. Sukan (eds.): *Recent Advances in Biotecnology*, Kluwer Academic, Boston, 1992, pp. 1 – 17.
297. K. Horikoshi, W. D. Grant: *Extremophiles: Microbial Life in Extreme Environments*. John Wiley & Sons, New York, 1998.
298. J. K. Kristjansson, G. O. Hreggvidsson: “Ecology and Habitats of Extremophiles”, World J. Microbiol. & Biotechnol. **11** (1995) 17 – 25. Link
299. K. Horikoshi: “Discovering Novel Bacteria, With an Eye to Biotechnological Applications”, Curr. Biol. **6** (1995) 292 – 297. Link
300. B. E. Jones, W. D. Grant, A. W. Duckworth, G. G. Owenson: “Microbial Diversity of Soda Lakes”, Extremophiles **2** (1998) 191 – 200. Link
301. S. T. Williams, M. Goodfellow, J. C. Vickers: “New Microbes From Old Habitats?” in D. P. Kelley, N. G. Carr (eds.): *The Microbe 1984, Part II: Prokaryotes Eukaryotes*. Cambridge University Press, UK, 1984, pp. 219 – 256.
302. C. A. Thomson, P. J. Delaquis, G. Mazza: “Detection and Measurement of Microbial Lipase Activity: A Review”, Crit. Rev. Food Sci. Nutrition **39** (1999) 165 – 187. Link
303. F. Beisson, A. Tiss, C. Riviere, R. Verger: “Methods for Lipase Detection and Assay: A Critical Review”, Eur. J. Lipid Sci. Technol. **2000** 133 – 153. Link
304. D. Wahler, J. L. Reymond: “High-Throughput Screening for Biocatalysts”, Curr. Opin. Biotechnol. **12** (2001) 535 – 544. Link
305. M. J. Olsen, D. Stephens, D. Griffiths, P. Daugherty, G. Georgiou, B. L. Iverson: “Function-Based Isolation of Novel Enzymes from a Large Library”, Nature Biotechnol. **18** (2000) 1071 – 1074. Link
306. H. Dalborge, L. Lange: “Using Molecular Techniques to Identify new Microbial Biocatalysts”, Trends Biotechnol. **16** (1998) 265 – 272. Link

307. R. I. Amann, Ludwig, W. K. H. Schleifer: “Phylogenetic Identification and in situ Detection of Individual Microbial cells without Cultivation”, Microbiol. Rev. **59** (1995) 143 – 169. Link
308. K. N. Timmis, A. L. Demain: “Ecology and Industrial Microbiology—Strange Bedfellows”, Curr. Opin. Microbiol. **1** (1998) 267 – 270. Link
309. C. R. Woese: “Bacterial Evolution”, Microbiol. Rev. **51** (1987) 221 – 271. Link
310. V. Torsvik, J. Goksoyr, F. L. Daae: “High Bio-Diversity of DNA of Soil Bacteria”, Appl. Environ. Microbiol. **56** (1990) 782 – 787. Link
311. S. J. Giovannoni, T. B. Britschgi, C. L. Moyer, K. G. Field: “Genetic Bio-Diversity in Sargasso Sea Bacterioplankton”, Nature **345** (1990) 60 – 63. Link
312. D. M. Ward, R. Weller, M. M. Bateson: “16S rRNA Sequences Reveal Numerous Uncultured Microorganisms in a Natural Community”, Nature **345** (1990) 63 – 65. Link
313. P. Hugenholtz, N. R. Pace: “Identifying Microbial Diversity in the Natural Environment: a Molecular Phylogenetic Approach”, Trends Biotechnol. **14** (1996) 190 – 197. Link
314. E. F. DeLong: “Marine Microbial Diversity: The Tip of the Iceberg”, Trends Biotechnol. **15** (1997) 203 – 207. Link
315. P. Hugenholtz, G. W. Tyson, R. I. Webb, A. M. Wagner, L. L. Blackall: “Investigation of Candidate Division TM7, A Recently Recognized Major Lineage of the Domain Bacteria with no Known Pure-Culture Representatives”, Appl. Environ. Microbiol. **67** (2001) 411 – 419. Link
316. P. Hugenholtz, B. M. Goebel, N. R. Pace: “Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Bio-Diversity”, J. Bacteriol. **180** (1998) 4765 – 4774. Link
317. J. M. Short: “Recombinant Approaches for Accessing Biodiversity”, Nature Biotechnol. **15** (1997) 1322 – 1323. Link
318. L. Frangeul, K. E. Nelson, C. Buchrieser, A. Danchin, P. Glaser, F. Kunst: “Cloning and Assembly Strategies in Microbial Genome Projects”, Microbiology **145** (1999) 2625 – 2634. Link
319. G. Deckert, P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, R. V. Swanson: “The Complete Genome of the Hyperthermophilic Bacterium Aquifex aeolicus”, Nature **392** (1998) 353 – 358. Link
320. H. Takami, K. Nakasone, Y. Takaki, G. Maeno, R. Sasaki, N. Masui, F. Fuji, C. Hirama, Y. Nakamura, N. Ogasawara, S. Kuhara, K. Horikoshi: “Complete Genome Sequence of the Alkaliphilic Bacterium Bacillus halodurans and Genomic Sequence Comparison with Bacillus subtilis”, Nucleic Acids Res. **28** (2000) 4317 – 4331. Link
321. A. Ruepp, W. Graml-Santos, M. L. Martinez, K. K. Koretke, C. Volker, H. W. Mewes, D. Frishman, S. Stocker, A. N. Lupas, W. Baumeister: “The Genome Sequence of the Thermoacidophilic Scavenger Thermoplasma acidophilum”, Nature **407** (2000) 508 – 513. Link
322. W. V. Ng et al.: “Genome Sequence of Halobacterium Species NRC-1”, Proc. Natl. Acad. Sci. USA **97** (2000) 12 176 – 12 181. Link
323. F. T. Robb, D. L. Maeder, J. R. Brown, J. DiRuggiero, M. D. Stump, R. K. Yeh, R. B. Weiss, D. M. Dunn: “Genomic Sequence of Hyperthermophile, Pyrococcus furiosus: Implications for Physiology and Enzymology”, Methods Enzymol. **330** (2001) 134 – 157. Link
324. S. Chakravarty, R. Varadarajan: “Elucidation of Determinants of Protein Stability through Genome Sequence Analysis”, FEBS Lett. **470** (2000) 65 – 69. Link
325. R. Das, M. Gerstein: “The Stability of Thermophilic Proteins: A Study Based on Comprehensive Genome Comparison”, Functional Integrative Genomics **1** (2000) 76 – 88. Link
326. J. W. Fickett: “Finding Genes by Computer: The State of the Art”, Trends Genetics **12** (1996) 316 – 320. Link
327. A. M. Shmatkov, A. A. Melikyan, C. F. Lhernousko, M. Borodovsky: “Finding Prokaryotic Genes by the ‘Frame-by-Frame’ Algorithm: Targeting Gene Starts and Overlapping Genes”, Bioinformatics **15** (1999) 874 – 886. Link
328. M. Skovgaard, L. J. Jensen, S. Brunak, D. Ussery, A. Krogh: “On the Total Number of Genes and Their Length Distribution in Complete Microbial Genomes”, Trends Genetics **17** (2001) 425 – 428. Link

329. B. R. Graveley: “Alternative Splicing: Increasing Diversity in the Proteomic World”, Trends Genetics **17** (2001) 100 – 107. Link
330. C. B. Burge: “Chipping Away at the Transcriptome”, Nature Genetics **27** (2001) 232 – 234. Link
331. P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, Y. Yuan: “Predicting Function: From Genes to Genomes and Back”, J. Molecular Biol. **283** (1998) 707 – 725. Link
332. R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant: “The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites”, Nucleic Acids Res. **29** (2001) 37 – 40. Link
333. M. Pagni, C. V. Jongeneel: “Making Sense of Score Statistics for Sequence Alignments”, Briefings Bioinformatics **2** (2001) 51 – 67. Link
334. K. Robison, W. Gilbert, G. M. Church: “Large Scale Bacterial Gene Discovery by Similarity Search”, Nature Genetics **7** (1994) 205 – 214. Link
335. U. Bhatia, K. Robison, W. Gilbert: “Dealing with Database Explosion: a Cautionary Note”, Science **276** (1997) 1724 – 1725. Link
336. D. Devos, A. Valencia: “Intrinsic Errors in Genome Annotation”, Trends Genetics **17** (2001) 429 – 431. Link
337. R. Overbeek, N. Larsen, G. D. Pusch, M. D'Souza, E. Selkov Jr., N. Kyrpides, M. Fonstein, N. Maltsev, E. Selkov: “WIT: Integrated System for High-Throughput Genome Sequence Analysis and Metabolic Reconstruction”, Nucleic Acids Res. **28** (2000) 123 – 125. Link
338. T. Ettema, J. van der Oost, M. Huynen: “Modularity in the Gain and Loss of Genes: Applications for Function Prediction”, Trends Genetics **17** (2001) 485 – 487. Link
339. M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates: “Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles”, Proc. Natl. Acad. Sci. USA **96** (1999) 4285 – 4288. Link
340. R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, E. V. Koonin: “The COG Database: New Developments in Phylogenetic Classification of Proteins from Complete Genomes”, Nucleic Acids Res. **29** (2001) 22 – 28. Link
341. R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, N. Maltsev: “The Use of Gene Clusters to Infer Functional Coupling”, Proc. Natl. Acad. Sci. USA **96** (1999) 2896 – 2901. Link
342. T. Dandekar, S. Schuster, B. Snel, M. Huynen, P. Bork: “Pathway Alignment: Application to the Comparative Analysis of Glycolytic Enzymes”, Biochem. J. **343** (1999) 115 – 124. Link
343. E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, D. Eisenberg: “A Combined Algorithm for Genome-Wide Prediction of Protein Function”, Nature **402** (1999) 83 – 86. Link
344. C. M. Fraser, J. A. Eisen, S. L. Salzberg: “Microbial Genome Sequencing”, Nature **406** (2000) 799 – 803. Link
345. M. Hensel, D. W. Holden: “Molecular Genetic Approaches for the Study of Virulence in both Pathogenic Bacteria and Fungi”, Microbiology **142** (1996) 1049 – 1058. Link
346. P. S. Coelho, A. Kumar, M. Snyder: “Genome-Wide Mutant Collections: Toolboxes for Functional Genomics”, Curr. Opin. Microbiol. **3** (2000) 309 – 315. Link
347. M. D. De Backer, B. Nelissen, M. Logghe, J. Viaene, I. Loonen, S. Vandoninck, R. de Hoogt, S. Dewaele, F. A. Simons, P. Verhasselt, G. Vanhoof, R. Contreras, W. H. Luyten: “An Antisense-Based Functional Genomics Approach for Identification of Genes Critical for Growth of Candida albicans”, Nature Biotechnol. **19** (2001) 235 – 241. Link
348. L. Hamer, K. Adachi, M. V. Montenegro-Chamorro, M. M. Tanzer, S. K. Mahanty, C. Lo, R. W. Tarpey, A. R. Skalchunes, R. W. Heiniger, S. A. Frank, B. A. Darveaux, D. J. Lampe, T. M. Slater, L. Ramamurthy, T. M. DeZwaan, G. H. Nelson, J. R. Shuster, J. Woessner, J. E. Hamer: “Gene Discovery and Gene Function Assignment in Filamentous Fungi”, Proc. Natl. Acad. Sci. USA **98** (2001) 5110 – 5115. Link

349. M. D. Gonzalez, C. A. Lichtensteiger, E. R. Vimr: “Adaptation of Signature-Tagged Mutagenesis to Escherichia coli K1 and the Infant-Rat Model of Invasive Disease”, FEMS Microbiol. Lett. **198** (2001) 125 – 128. Link
350. S. M. Hammond, A. A. Caudy, G. J. Hannon: “Post-Transcriptional Gene Silencing by Double-Stranded RNA”, Nature Rev. Genetics **2** (2001) 110 – 119. Link
351. A. Kumar, S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K. H. Cheung, P. Miller, M. Gerstein, G. S. Roeder, M. Snyder: “Subcellular Localization of the Yeast Proteome”,Genes Dev. **16** (2002) 707 – 719. Link
352. F. Eisenhaber, P. Bork: “Wanted: Subcellular Localization of Proteins Based on Sequence”, Trends Cell Biol. **8** (1998) 169 – 170. Link
353. M. Diehn, M. B. Eisen, D. Botstein, P. O. Brown: “Large-Scale Identification of Secreted and Membrane-Associated Gene Products Using DNA Microarrays”, Nature Genetics **25** (2000) 58 – 62. Link
354. K. A. Jacobs, L. A. Collins-Racie, M. Colbert, M. Duckett, M. Golden-Fleet, K. Kelleher, R. Kriz, E. R. LaVallie, D. Merberg, V. Spaulding, J. Stover, M. J. Williamson, J. M. McCoy: “A Genetic Selection for Isolating cDNAs Encoding Secreted Proteins”, Gene **198** (1997) 289 – 296. Link
355. A. Kumar, M. Snyder: “Protein Complexes Take the Bait”, Nature **415** (2002) 123 – 124. Link
356. A. Kumar, M. Snyder: “Emerging Technologies in Yeast Genomics”, Nature Rev. Genetics **2** (2001) 302 – 312. Link
357. I. Yanai, A. Derti, C. DeLisi: “Genes Linked by Fusion Events are Generally of the Same Functional Category: A Systematic Analysis of 30 Microbial Genomes”, Proc. Natl. Acad. Sci. USA **98** (2001) 7940 – 7945. Link
358. D. H. Kozian, B. J. Kirschbaum: “Comparative Gene-Expression Analysis”, Trends Biotechnol. **17** (1999) 73 – 78. Link
359. D. J. Lockhart, E. A. Winzeler: “Genomics, Gene Expression and DNA Arrays”, Nature **405** (2000) 827 – 836. Link
360. T. J. Griffin, D. K. Han, S. P. Gygi, B. Rist, H. Lee, R. Aebersold, K. C. Parker: “Toward a High-Throughput Approach to Quantitative Proteomic Analysis: Expression-Dependent Protein Identification by Mass Spectrometry”, J. Am. Soc. Mass Spectrom. **12** (2001) 1238 – 1246. Link
361. K. E. Nelson, I. T. Paulsen, J. F. Heidelberg, C. M. Fraser: “Status of Genome Projects for Nonpathogenic Bacteria and Archaea”, Nature Biotechnol. **18** (2000) 1049 – 1054. Link
362. C. A. Cummings, D. A. Relman: “Using DNA Microarrays to Study Host-Microbe Interactions”, Emerging Infectious Diseases **6** (2000) 513 – 525 Link
363. P. P. Zarrinkar, J. K. Mainquist, M. Zamora, D. Stern, J. B. Welsh, L. M. Sapinoso, G. M. Hampton, D. J. Lockhart: “Arrays of Arrays for High-Throughput Gene Expression Profiling”, Genome Res. (2001) 1256 – 1261. Link
364. S. A. Lesley: “High-Throughput Proteomics: Protein Expression and Purification in the Postgenomic World”, Protein Expression Purif. **22** (2001) 159 – 164. Link
365. J. R. Hudson, Jr., E. P. Dawson, K. L. Rushing, C. H. Jackson, D. Lockshon, D. Conover, C. Lanciault, J. R. Harris, S. J. Simmons, R. Rothstein, S. Fields: “The Complete Set of Predicted Genes from Saccharomyces cerevisiae in a Readily Usable Form”, Genome Res. **7** (1997) 1169 – 1173. Link
366. Q. Liu, M. Z. Li, D. Leibham, D. Cortez, S. J. Elledge: “The Univector Plasmid-Fusion System, a Method for Rapid Construction of Recombinant DNA without Restriction Enzymes”, Curr. Biol. **8** (1998) 1300 – 1309. Link
367. J. L. Hartley, G. F. Temple, M. A. Brasch: “DNA Cloning Using in vitro Site-Specific Recombination”, Genome Res. **10** (2000) 1788 – 1795. Link
368. N. Sheibani: “Prokaryotic Gene Fusion Expression Systems and Their Use in Structural and Functional Studies of Proteins”, Preparative Biochemistry Biotechnol. **29** (1999) 77 – 90. Link
369. M. Nasoff, M. Bergseid, J. P. Hoeffler, J. Heyman: “A. High-Throughput Expression of Fusion Proteins”, Methods Enzymol. **328** (2000) 515 – 549. Link

370. M. R. Martzen, S. M. McCraith, S. L. Spinelli, F. M. Torres, S. Fields, E. J. Grayhack, E. M. Phizicky: “A Biochemical Genomics Approach for Identifying Genes by the Activity of their Products”, Science **286** (1999) 1153 – 1155. Link
371. H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, M. Snyder: “Global Analysis of Protein Activities Using Proteome Chips”, Science **293** (2001) 2101 – 2105. Link
372. A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga: “Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes”, Nature **415** (2002) 141 – 147. Link
373. H. Zhu, J. F. Klemic, S. Chang, P. Bertone, A. Casamayor, K. G. Klemic, D. Smith, M. Gerstein, M. A. Reed, M. Snyder: “Analysis of Yeast Protein Kinases Using Protein Chips”, Nature Genetics **26** (2000) 283 – 289. Link
374. Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S. L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sorensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. Hogue, D. Figeys, M. Tyers: “Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry”, Nature **415** (2002) 180 – 183 Link
375. G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin: “A Generic Protein Purification Method for Protein Complex Characterization and Proteome Exploration”, Nature Biotechnol. **17** (1999) 1030 – 1032. Link
376. A. Abbott: “Betting on Tomorrow's Chips”, Nature **415** (2002) 112 – 114. Link
377. J. Ziauddin, D. M. Sabatini: “Microarrays of Cells Expressing Defined cDNAs”, Nature **411** (2001) 107 – 110. Link
378. K. D. Wittrup: “Protein Engineering by Cell-Surface Display”, Curr. Opin. Biotechnol. **12** (2001) 395 – 399. Link
379. Y. S. Kim, H. C. Jung, J. G. Pan: “Bacterial Cell Surface Display of an Enzyme Library for Selective Screening of Improved Cellulase Variants”, Appl. Environ. Microbiol. **66** (2000) 788 – 793. Link
380. S. Atwell, J. A. Wells: “Selection for Improved Subtiligases by Phage Display” Proc. Natl. Acad. Sci. USA **96** (1999) 9497 – 9502. Link

Specific References

381. C. J. Noren, S. J. Anthony-Cahill, M. C. Griffith, P. G. Schultz, Science **244** (1989) 182. Link
382. T. J. Maglieri, J. C. Anderson, P. G. Schultz, J. Mol. Biol. **307** (2001) 755. Link
383. Hirao, I. Ohtsuki, T. Fujiwara, T. T. Mitsui, T. Yokogawa, T. Okuni, H. Nakayaama, K. Takio, T. Yabuki, T. Kigawa, K. Kodama, T. Yokogawa, K. Nishikawa, S. Yokoyama, Nature Biotechnol. **20** (2002) 177. Link
384. S. Longhi, M. Czjzek, V. Lamzin, A. Nicolas, C. Cambillau, J. Mol. Biol. **268** (1997) 779. Link
385. P. Kuhn, M. Knapp, S. M. Soltis, G. Ganshaw, M. Thoene, R. Bott, Biochemistry **37** (1998) 13 446. Link
386. C. Jelsch, M. M. Teeter, V. Lamzin, Pichon-Pesme, V. R. H. Blessing, C. Lecomte, Proc. Natl. Acad. Sci. USA **97** (2000) no. 3, 171. Link
387. R. Bott, M. Ultsch, A. Kossiakoff, T. Graycar, B. Katz, S. Power, J. Biol. Chem. **263** (1998) 7895. Link
388. J. A. Wells, B. C. Cunningham, T. P. Graycar, D. A. Estell, Phil. Trans. Roy. Soc. London **A317** (1986) 415. Link
389. J. A. Wells, D. A. Estell, TIBS **13** (1988) 291. Link

390. A. J. Russel, A. R. Fersht, Nature **328** (1987) 496. Link
391. J. A. Wells, B. C. Cunningham, T. P. Graycar, D. A. Estell, Proc. Natl. Acad. Sci. USA **84** (1987) 5167 Link
392. L. Hedstrom, L. Szilagy, W. J. Rutter, Science **255** (1992) 1249. Link
393. W. P. Stemmer, Proc. Natl. Acad. Sci. USA **93** (1994) 10747. Link
394. W. P. Stemmer, Nature **307** (1994) 389. Link
395. W. M. Coco, W. E. Leveinson, M. J. Crist, H. J. Hektor, A. Darzins, P. T. Pienkos, C. H. Squires, D. J. Monticello, Nature Biotechnol. **19** (2001) 354. Link
396. M. Baca, T. S. Scanlan, R. C. Stephenson, J. A. Wells, Proc. Natl. Acad. Sci. USA **94** (1997) 10063. Link
397. M. R. Arkin, J. A. Wells, J. Mol. Biol. **284** (1998) 1083. Link
398. S. Atwell, J. A. Wells, Proc. Natl. Acad. Sci. USA, **96** (1999) 9497. Link
399. O. Misset in van den Tweel et al. (eds.): *Stability and Stabilization of Enzymes*, Elsevier, Amsterdam 1993, p. 111.
400. S. Delgrave, D. J. Murphy, J. L. Rittenhouse Pruss, A. M. Maffia, III, B. L. Marrs, E. J. Bylina, W. J. Coleman, C. L. Grek, M. R. Dilworth, M. M. Yang, D. C. Youvan, Protein Eng. **14** (2001) 261. Link
401. G. Winter, A. R. Fersht, A. J. Wilkinson, M. Zoller, M. Smith, Nature **299** (1982) 756. Link
402. A. R. Fersht, *Structure and Mechanism in Protein Science*, Chap. 15, W. H. Freeman, New York, 1999.
403. C. Holm, R. C. Davis, T. Osterlund, M. C. Schotz, G. Frederikson, FEBS Lett. **16** (1994) 234. Link
404. J. Emmerich, O. U. Beg, J. Petersen, L. Previato, J. D. Brunzell, H. B. Brewer Jr., S. J. Santamarina-Fojo, Biol. Chem. **267** (1992) 4161. Link
405. S. Jager, G. Demleitner, F. Gotz, FEMS Micobiol. Lett. **79** (1992) 249. Link
406. M. Holmquist, M. Martinelle, I. G. Clausen, S. Patkar, A. Svendsen, K. Hult, Lipids **29** (1994) 599. Link
407. M. Martinelle, M. Holmquist, I. G. Clausen, S. Patkar, A. Svendsen, K. Hult, Protein Eng. **9** (1996) 519. Link
408. B. A. Katz, A. J. Kossiakoff, Biol. Chem. **261** (1986) 15 480. Link
409. J. Clarke, K. Henrick, A. R. Fersht, J. Mol. Biol. **253** (1995) 493. Link
410. J. Clarke, A. M. Hounslow, A. R. Fersht, J. Mol. Biol. **253** (1995) 505. Link
411. L. Serrano, J. Sancho, M. Hirshberg, A. R. Fersht, J. Mol. Biol. **227** (1992) 544. Link
412. A. Horovitz, J. M. Matthews, A. R. Fersht, J. Mol. Biol. **227** (1992) 560. Link
413. Y. W. Chen, A. R. Fersht, FEBS Lett. **347** (1994) 304. Link
414. L. Serrano, A. G. Day, A. R. Fersht (1993) J. Mol. Biol. **233**, 305. Link
415. A. Shaw, R. Bott, A. G. Day, Curr. Opin. Biotechnol. **10** (1999) 349. Link
416. D. A. Estell, T. P. Graycar, J. A. Wells, J. Biol. Chem. **260** (1985) 6518. Link
417. L. Serrano, Curr. Opin. Struct. Biol. **4** (1994) 107. Link
418. K. Chen, F. H. Arnold, Proc. Natl. Acad. Sci. USA **90** (1993) 5618. Link
419. L. You, F. H. Arnold, Protein Eng. **9** (1996) 77. Link
420. Zhang, J.-H. G. Dawes, W. P. C. Stemmer, Proc. Natl. Acad. Sci. USA **94** (1997) 4504. Link
421. U. T. Bornscheuer, J. Altenbuchner, H. H. Meyer, Biotechnol. Bioeng. **58** (1998) 554. Link
422. Fong, S. T. D. Machajewski, C. C. Mak, C. Wong, Chem. Biol. (2000) 873. Link
423. M. T. Reetz, S. Wilensek, D. Zha, K.-E. Jaeger, Angew. Chem. Int. Ed. **40** (2001) 3589. Link

Specific References

424. P. R. Shewry, A. S. Tatham, *Wheat Gluten*, Royal Society of Chemistry, London 2000.
425. H. Uhlig: “Enzymes in Flour Processing and Baking”, in *Industrial Enzymes and their Application*, Wiley, New York 1998.

426. J. H. G. M. Mutsaers: "The Application of Enzymes in Baking", in S. A. G. F. Angelino, R. J. Hamer, W. van Hartingsveldt, F. Heidekamp, J. P. van der Lugt (eds.): *Proceedings of ESEGP-1, Noordwijkerhout, NL, December 2 – 4*, AACC, St. Paul, MN 1997, pp. 160 – 164.
427. O. Silberstein, The Bakers Digest **35** (1961) 44 – 48, 86. Link
428. P. E. Pritchard, FMBRA Bull. **5** (1986) 208 – 221. Link
429. US 1 676 446, 1928 (H. A. Kohman, R. Irwin, E. S. Stateler).
430. J. A. Shellenberg, Cereal. Sci. Today **16** (1971) 114 – 117. Link
431. T. J. Schoch: "Starch in Bakery Products", The Bakers Digest **39** (1965) 48 – 57. Link
432. J. A. Maga, CRC Crit. Rev. Food Technol. **5** (1975) 443 – 486. Link
433. J. Kulp, J. G. Ponte, Jr.: "Staling of White Pan Bread: Fundamental Causes". CRC Crit. Rev. Food Technol. **15** (1981) 1 – 48. Link
434. H. F. Zobel, K. Kulp: "The Staling Mechanism", in R. E. Hebeda, H. F. Zobel (eds.): *Baked Goods Freshness, Technology, Evaluation, and Inhibition of Staling*, Marcel Dekker, New York 1996, pp. 1 – 64.
435. A. S. Schultz, F. D. Schoonover, R. A. Fisher, S. S. Jackel, Cereal Chem. **29** (1952) 200 – 211. Link
436. L. K. Bowles: "Amylolytic Enzymes", in R. E. Hebeda, H. F. Zobel (eds.): *Baked Goods Freshness, Technology, Evaluation, and Inhibition of Staling*, Marcel Dekker, New York 1996, pp. 105 – 129.
437. K. R. Morgan, L. Hutt, J. Gerrard, D. Every, M. Ross, M. Gilpin, Starch **49** (1997) 54 – 59. Link
438. P. Würsch, D. Gumy, Carbohydr. Res. **256** (1994) 129 – 137. Link
439. J. Maat, M. Roza, J. Verbakel, H. Stam, M. J. Santos da Selva, M. Bosse, M. R. Egmond, M. L. D. Hagemans in J. Visser, G. Beldman, M. A. Kusters-van Someren, A. G. J. Voragen (eds.): *Xylanas and Xylanases*, Elsevier Science Publishers, Amsterdam 1992, pp. 349 – 360.
440. B. V. McCleary, Int. J. Biol. Macromol. **8** (1986) 349 – 354. Link
441. G. B. Fincher, B. A. Stone in Y. Pomeranz (ed.): Adv. Cereal Sci. Technol. **8** (1986) 207 – 295. Link
442. R. Montgomery, F. Smith, J. Am. Chem. Soc. **77** (1955) 3325 – 3328. Link
443. H. Gruppen, R. A. Hoffmann, F. J. M. Kormelink, A. G. J. Voragen, J. P. Kamerling, J. F. G. Vliegenthart: *Structural Characteristics of Wheat Flour Arabinoxylans*, Thesis, Wageningen 1992, pp. 69 – 84.
444. M. Bunzel, J. Ralph, J. M. Marita, R. D. Hatfield, H. Steinhart, J. Sci. Food Agric. **81** (2001) 653 – 660. Link
445. U. Girhammer, B. M. Nair, Food Hydrocoll. **6** (1992) 285 – 299. Link
446. M. S. Izydorczyk, C. G. Biliaderis, Carbohydr. Polym. **28** (1995) 33 – 48. Link
447. X. Rouau, M. L. El-Hayek, D. Moreau, J. Cereal Sci. **19** (1994) 259 – 272. Link
448. C. M. Courtin, A. Roelants, J. A. Delcour, J. Agric. Food Chem. **47** (1999) 1870 – 1877. Link
449. US 5 176 927, 1993 (S. Haarasilta, T. Pullinen, I. Tammersalo-Karsten, S. Vaisanen, H. Franti).
450. M. van Oort, IFI **4** (1996) 42 – 45. Link
451. S. Haarasilta, T. Pullinen, Agro Food Industry Hi-Tech **3** (1992) 12 – 13. Link
452. J. H. G. M. Mutsaers: "The Application of Enzymes in Baking", in S. A. G. F. Angelino, R. J. Hamer, W. van Hartingsveldt, F. Heidekamp, J. P. van der Lugt (eds.): *Proceedings from The First European Symposium on Enzymes and Grain Processing*, TNO Nutrition and Food Research Institute, Zeist, The Netherlands, 1997, pp. 160 – 163.
453. C. Poulsen, P. B. Hostrup, Cereal Chem. **75** (1998) 51 – 57. Link
454. R. Casey: "Lipoxygenases and Breadmaking", in S. A. G. F. Angelino, R. J. Hamer, W. van Hartingsveldt, F. Heidekamp, J. P. van der Lugt, (eds.): *Proceedings from The First European Symposium on Enzymes and Grain Processing*, TNO Nutrition and Food Research Institute, Zeist, The Netherlands, 1997, pp. 188 – 194.
455. T. Szalkucki: "Applications of Oxidoreductases" in T. Nagodawithana, G. Reed (eds.): *Enzymes in Food Processing*, 3rd ed., Academic Press, New York 1993, pp. 279 – 291.
456. F. E. Hammer, "Oxidoreductases" in T. Nagodawithana, G. Reed (eds): *Enzymes in Food Processing*, 3rd Ed. *Academic Press*, New York, 1993, pp. 221 – 277.

457. M. van Oort, H. Hennink, H. Moonen: "Peroxidases in Breadmaking", in S. A. G. F. Angelino, R. J. Hamer, W. van Hartingsveldt, F. Heidekamp, J. P. van der Lugt, (eds.): *Proceedings from The First European Symposium on Enzymes and Grain Processing*, TNO Nutrition and Food Research Institute, Zeist, The Netherlands, 1997, pp. 195 – 203.
458. M. C. Figueroa-Espinoza, X. Rouau, Cereal Chem. **75** (1998) 259 – 265. Link
459. R. Drapron, B. Godon: "Role of Enzymes in Baking", in J. E. Krüger, D. Lineback, C. E. Stauffer (eds.): *Enzymes and their Role in Cereal Technology*, AACC, St. Paul, MN, 1987, pp. 281 – 324.
460. WO 98/44804, 1998 (J. B. Soe, C. H. Poulsen, P. Rasmussen, M. R. Zargahi).
461. WO 97/22257, 1997 (P. Wagner, J. Q. Si).
462. WO 97/21351, 1997 (P. Wagner, J. Q. Si).
463. WO 99/03351, 1999 (X. Rouau, M. Schroder, J. B. Soe).
464. Y. Pomeranz: "Wheat Flour Components in Bread Making" in *Modern Cereal Science and Technology*, VCH Verlagsgesellschaft, Weinheim, 1987, pp. 165 – 200.
465. C. H. Poulsen, J. B. Soe, "Effect and Functionality of Lipases in Dough and Bread" in S. A. G. F. Angelino, R. J. Hamer, W. van Hartingsveldt, F. Heidekamp, J. P. van der Lugt (eds.): *Proceedings from The First European Symposium on Enzymes and Grain Processing*, TNO Nutrition and Food Research Institute, Zeist, The Netherlands, 1997, pp. 204 – 214.
466. T. L.-G. Carlson: "Law and Order in Wheat Flour Dough. Colloidal Aspects of the Wheat Flour Dough and its Lipid and Protein Constituents in Aqueous Media", Lund University, Sweden, 1981.
467. US 4 567 046, 1986 (S. Inoue, S. Ota).
468. J. Adler-Nissen: *Enzymic Hydrolysis of Food Proteins*, Elsevier, London 1986.
469. J. A. Johnson, B. S. Miller, P. D. Boyer, W. F. Geddes, Cereal Chem. **33** (1956) no. 1, 1 – 17. Link
470. O. P. Ward in H. W. Blanch, S. Drew, D. I. C. Wang (eds.): *Comprehensive Biotechnology*, Pergamon Press, Oxford, 1985, vol. 3, pp. 789 – 815.
471. `http://www.expasy.ch/enzyme/`.
472. R. Lásztity: *The Chemistry of Cereal Proteins*, CRC Press, Boca Raton 1996, pp. 19 – 139.
473. D. B. Pratt, in Y. Pomeranz (ed.): *Wheat Chemistry and Technology*, 2nd ed. Vol. 3, AACC, St. Paul, MN, 1976.
474. G. Oliver, R. J. Wheeler, J. Sci. Food Agric. **71** (1996) 337 – 344. Link
475. M. A. Martínez-Anaya, J. Agric. Food Chem. **44** (1996) no. 9, 2469 – 2480. Link
476. J. A. Gerrard, M. P. Newberry, M. Ross, A. J. Wilson, S. E. Fayle, S. Kavale, J. Food Sci. **65** (2000) 312 – 314. Link
477. J. A. Gerrard, S. E. Fayle, A. J. Wilson, M. P. Newberry, M. Ross, S. Kavale: "Dough Properties and Crumb Strength of White Pan Bread as Affected by Microbial Transglutaminase", J. Food Sci. **63** (1998) 472 – 475. Link

Specific References

478. A. Mehlitz, Biochem. **221** (1930) 217. Link
479. C. Grassin, P. Fauquembergue in J. Visser, A. G. J. Voragen (eds.): *Pectin and Pectinases*, Elsevier Science, Amsterdam (1996) 453.
480. M. C. Mac Cann, K. Roberts, Agro-Food Ind. Hi-Tech. **5** (1994) 43. Link
481. N. C. Carpita, D. M. Gibeaut, Plant J. **3** (1993) 1. Link
482. P. Albersheim, A. G. Darvill, M. A. O'Neill, H. A. Schols, A. G. J. Voragen in J. Visser, A. G. J. Voragen (eds.): *Pectin and Pectinases*, Elsevier Science, Amsterdam, 1996, p. 47.
483. H. A. Schols, A. G. J. Voragen: "Complex Pecitins: Structure Elucidation using Enzymes", in J. Visser, A. G. J. Voragen (eds.): Progress in Biotechnology, **14**: *Pectin and Pectinases*, Elsevier Science, Amsterdam, 1996, p. 793 – 798.
484. J. De Vries, A. Voragen, F. Rombouts, W. Pilnik in: "Chemistry and Function of Pectins", ACS Symp. Ser. **310** (1986) 38. Link
485. H. Schols, M. A. Posthumus, A. Voragen, Carbohydr. Res. **206** (1990) 117. Link
486. P. Dey and K. Brinson, Adv. Carbohydr. Chem. Biochem. **42** (1984) 265. Link

487. F. Will, *PhD Thesis*, University of Giessen, 1990.
488. P. Pellerin, S. Vidal, P. Williams, J. M. Brillouet, Carbohydr. Res. **277** (1995) 135. Link
489. T. Doco, P. Williams, S. Vidal, P. Pellerin, Carbohydr. Res. **297** (1997) 181. Link
490. J. P. Vincken, G. Beldman, A. Voragen, Plant Physiol. **104** (1994) 99. Link
491. I. S. Bentley, E. C. Williams in T. Godfrey, S. West (eds.): *Industrial Enzymology*, 2nd ed., Macmillan, London, 1996, 341.
492. A. Voragen, *PhD Thesis*, University of Wageningen, NL, 1972.
493. C. Versteeg, *PhD Thesis*, University of Wageningen, NL, 1979.
494. L. Parenicova, *PhD Thesis*, University of Wageningen, NL, 2000.
495. H. J. D. Bussink, *PhD Thesis*, University of Wageningen, NL, 1992.
496. G. Mutter, *PhD Thesis*, University of Wageningen, NL, 1997.
497. F. M. Rombouts, A. Voragen, M. Searle-van-leeuwen, C. Gereads, H. Schols, W. Pilnik, Carbohydr. Polym. **9** (1988) 5. Link
498. H. Schols, P. Veld, W. Van Deelen, A. Voragen, Z. Lebensm. Forsch. **192** (1991) 142. Link
499. L. Saulnier, J. M. Brillouet, M. Moutounet, C. Herve du Penhoat, V. Michon, Carbohydr. Res. **224** (1992) 19. Link
500. F. Kormelink, M. Searle-van-Leeuwen, T. Wood, A. Voragen, J. Biotechnol. **27** (1993) 249. Link
501. W. Fogarty, C. Kelly in A. Rose (ed.): *Economic Microbiology*, vol. V, Academic Press, NY, 1980, 115.
502. C. Grassin, P. Fauquembergue in T. Godfrey, S. West (eds.): *Industrial Enzymology*, 2nd ed., Mac Millan, London, 1996, 227.
503. "Apple and Pear", *The Fruit Enzyme Directory*, Gist-brocades, 1998.
504. C. Grassin, Flüssiges Obst. **57** (1992) 519. Link
505. F. Will, G. Hasselbeck, H. Dietrich, Flüssiges Obst, **59** (1992) 352. Link
506. "Grapes and Berries", *The Fruit Enzyme Directory*, Gist-Brocades, 1998.
507. D. Dubourdieu, *PhD Thesis*, University of Bordeaux, 1997.
508. "Tropical Fruit and Citrus", *The Fruit Enzyme Directory*, Gist-Brocades, 1998.
509. G. G. Martin, E. Jamin, J. Gonzales, G. Remaud, V. Hanote, P. Ströber, N. Naulet, Fruit Proces. **9** (1997) 344. Link

Specific References

510. G. J. Fix: "The Role of pH in Brewing", The New Brewer **12** (1995) no. 6. Link
511. A. A. D. Comrie: "Brewing Water", Brewer's Digest **42** (1967) no. 7. Link
512. J. S. Hough: *The Biotechnology of Brewing*, Cambridge University Press, London, 1991.
513. G. J. Fix: *Principles of Brewing Science*, Brewers Publications, Boulder, CO, 1989.
514. J. De Clerck: *A Textbook of Brewing*, **vol. 1**, Translated by K. Barton-Wright, Chapman-Hall Ltd., London, 1957.
515. J. D. McGuiness, R. Eastmond, D. R. J. Laws: J. Inst. Brew. **81** (1975) 287. Link
516. D. S. Ryder, J. Power in W. A. Hardwick (ed.): *Handbook of Brewing*, Marcel Dekker, New York, 1995 p. 203.
517. C. W. Bamforth, R. E. Muller, M. D. Walker, J. Am. Soc. Brew. Chem. **51** (1993) 79. Link
518. C. Blockmans, M. C. Heilporn, C. A. Masschelein: "Scope and limitations of enzymatic deoxygenating methods to improve flavor stability of beer", J. Am. Soc. Brew. Chem. **45** (1987) 85 – 90. Link
519. M. Uchida, M. Ono, J. Am. Soc. Brew. Chem. **57** (1999) no. 4, 145 – 150. Link
520. J. L. Owades, L. Maresca, G. Rubin, *Proceedings of the American Society of Brewing Chemists*, (American Society of Brewing Chemists, St. Paul, Minnesota, 1959).
521. I. C. Trelea, S. Landaud, E. Latrille, G. Corrieu, J. Am. Soc. Brew. Chem. **59** (2001) no. 2, 77 – 87. Link

522. J. W. Ahn, K. H. Park, J. H. Seo: "2. Production of isomaltooligosaccharides by immobilized transglucosidase in a packed-bed reactor", Korean J. Food Sci. & Technol. **30** (1998) no. 1, 110 – 117. Link
523. J. H. Munroe: "Fermentation", in W. A. Hardwick (ed.): *Handbook of Brewing*, Marcel Dekker, New York, 1995.
524. T. Wainwright: "Biochemistry of Brewing", in W. P. K. Findlay (ed.): *Modern Brewing Technology*, Macmillan Press, Cleveland, Ohio, 1971.

Specific References

525. P. Walstra, T. J. Geurts, A. Noomen, A. Jellema, M. A. J. S. van Boekel in: *Dairy Technology*, Marcel Dekker, New York, 1999, Chap. 3.
526. D. G. Dalgleish in P. F. Fox (ed.): *Cheese: Chemistry, Physics and Microbiology*, Vol 1, Elsevier Applied Science, London, 1987, Chap. 3.
527. M. L. Green, A. S. Grandison in P. F. Fox (ed.): *Cheese: Chemistry, Physics and Microbiology*, Vol. 1, Elsevier Applied Science, London, 1987, Chap 4.
528. T. Aoyagi, S. Kunimoto, H. Morishima, T. Takeuchi, H. Umezawa: "Effect of Pepstatin on Acid Proteases", J. Antibiot. **24** (1971) 687. Link
529. M. Baudys, S. Foundling, M. Pavlik, T. Blundell, V. Kostka: "Protein Chemical Characterization of Mucor pusillus Aspartic Protease. Amino Acid Sequence Homology with the Other Aspartic Proteinases, Disulfide Bond Arrangement and Site of Carbohydrate Attachment," FEBS Lett. **235** (1988) 271. Link
530. A. Sali, B. Veerapandian, J. B. Cooper, D. S. Moss, T. Hofmann, T. Blundell, Proteins: Struct. Funct. Gen. **12** (1992) 158. Link
531. B. Foltman: "Chymosin: A Short Review on Foetal and Neonatal Gastric Proteases", Scand. J. Lab. Invest. **52** (suppl. 210) (1992) 65. Link
532. B. Foltman, V. Barkholt Pedersen, H. Jacobsen, D. Kaufman, G. Wybrandt: "The Complete Amino Acid Sequence of Prochymosin", Proc. Natl. Acad. Sci. USA, **74**, (1977) 2321 – 2324. Link
533. R. Y. Yada, S. Nakai: "Use of Principal Component Analysis to Study the Relationship between Physical/Chemical Properties and the Milk-Clotting to Proteolytic Activity Ratio of Some Aspartyl Proteases", J. Agric. Food Chem. **34** (1986) 675. Link
534. D. B. Emmons, D. C. Beckett: "Milk-Clotting Enzymes: 1. Proteolysis During Cheese Making in Relation to Estimated Losses of Yield", J. Dairy Sci. **73** (1990) 2007 – 2015. Link
535. W. J. Harper, C. R. Lee: "Residual Coagulants in Whey", J. Food Sci. **40** (1975) 28. Link
536. P. L. H. McSweeney, M. J. Sousa: "Biochemical Pathways for the Production of Flavour Compounds in Cheeses during Ripening: A Review", Lait, **77**, 41. Link
537. A. L. de Roos, T. J. Geurts, P. Walstra: "On the Mechanism of Rennet Retention in Cheese", Bull. Int. Dairy Fed. (1998) No. 332, 15. Link
538. D. G. Holmes, J. W. Duersch, C. A. Ernstrom: "Distribution of Milk Clotting Enzymes Between Curd and Whey and their Survival during Cheddar Cheese Making", J. Dairy Sci. **60** (1977) 862 – 869. Link
539. N. Klein, S. Lortal: "Attenuated Starters: an Efficient Means to Influence Cheese Ripening - A Review", Int. Dairy J. **9** (1999) 751. Link
540. J. H. Nelson, R. G. Jensen, R. E. Pitas: "Pregastric Esterase and other Oral Lipases – A Review", J. Dairy Sci. **60** (1976) 327. Link
541. H. Goudedranche, P. Ducruet, J.-C. Vachot, R. Pannetier, J.-L. Maubois: "Utilisation du lysozyme en tant qu' agent regulateur de l' affinage en fromagerie MMV," Le Lait **66** (1986) 189. Link
542. A. L. de Roos, T. J. Geurts, P. Walstra: "The Association of Lysozyme with Casein Micelles", Int. Dairy J. **8** (1998) 319. Link
543. F. E. Cunningham, V. A. Proctor, S. J. Goetsch: "Egg-White Lysozyme as a Food Preservative: An Overview", World Poultry Sci. J. **47** (1991) 141. Link
544. E. Dickinson: "Enzymic Crosslinking as a Tool for Food Colloid Rheology Control and Interfacial Utilization", Trends Food Sci. Technol. **8** (1997) 334. Link

545. P. Chr. Lorenzen, E. Schlimme: "Properties and Potential Fields of Application of Transglutaminase Preparations in Dairying", Bull. Int. Dairy Fed. (1998), no. 332, 47. Link
546. H. T. Vesa, P. Marteau, R. Korpela: "Lactose Intolerance", J. Am. Coll. Nutrit. **19** (2000) 165S. Link
547. R. R. Mahoney: *Modification of Lactose and Lactose Containing Dairy Products with β-Galactosidase* in P. Fox (ed.): Developments in Dairy Chemistry **3**, Elsevier Applied Science, Amsterdam, 1985, Chap. 3, pp. 69 – 109.
548. T. Godfrey, 2nd European Symposium on Enzymes in Grain Prosessing ESEPG-2, VVT Symposium 207, Helsinki, Dec. 1999.
549. T. Godfrey, ET Consulting, The Industrial Enzymes Market: A Summary: http://www.biocatalysts.com/frames/biocatalysts_01.html.
550. S. Schwimmer: *Source Book of Food Enzymology*, The Avi Publ. Comp., Westport, Conn. 1981, pp. 461 – 510.
551. G. Stefansson, in T. Nagodawithan, G. Reed (eds.): *Enzymes in Food Processing*, 3rd ed., Academic Press, New York 1993, p. 459.
552. C. G. Beddows, A. G. Ardishir: "The production of soluble fish protein solution for use in fish sauce manufacture I. The use of added enzymes", J. Food Technol. **14** (1979) 603 – 612. Link
553. G. R. Lilley, N. J. Skill, M. Griffin, P. L. R. Bonner: "Detection of Transglutaminase in Vivia faba Cotyledons in Plant Proteins from European Crops", J. Guéguen (ed.): *Food and Non-food Applications*, Springer Verlag, Heidelberg 1998.
554. R. S. Shallenberger, *Taste Chemistry*, Aspen Publishers, Inc., New York, 1994.
555. P. M. Nielsen, D. Petersen, C. Dambmann: "Improved Method for Determining Food Protein Degree of Hydrolysis", J. Food Sci. **66** (2001) no. 5, 642 – 646.Link
556. M. R. Smith, G. N. Bruce in H. Gruppen, W. van Hartngsveldt (eds): *1. International Symposium on Enzymatic Protein Processing. Proceedings of ISEPP-1, Noordwijkerhout, The Netherlands, December 2 – 4, 1998.*
557. M. Fernández, J. A. Ordónez, J. M. Bruna, B. Herranz, L. de la Hoz: "Accelerated Ripening of Dry Fermented Sausages", Trends Food Sci. Technol. **11** (2000) 201 – 209. Link
558. http://www.transglutaminase.com/englisch/frame/ajinomoto1.htm
559. M. Motoki, K. Seguro: "Transglutaminase and its use for Food Processing", Trends Food Sci. Technol. **9** (1998) 204 – 210. Link
560. J. Kuraishi, H. Sakamoto, T. Soeda: "Application of Transglutaminase for Food Processing", Fleischwirtschaft Int. **3** (1998) 37 – 9. Link
561. T. Szalkucki, in T. Nagodawithan, G. Reed, (eds.): *Enzymes in Food Processing*, 3rd ed., Academic Press, New York 1993, p. 279 – 291.
562. Y. Tanaka, J. Hirano, T. Funada: "Concentration of Docosahexaenoic Acid in Glyceride by Hydrolysis of Fish Oil with Candida Cylendracea Lipase", J. Am. Oil. Chem. Soc. **69** (1992) 1210 – 1214. Link
563. G. P. McNiell, R. G. Ackman, S. R. More: "Lipase-catalyzed Enrichment of Long Chain Polyunsaturated Fatty Acids", J. Am. Oil. Chem. Soc. **73** (1966) no. 11, 1403 – 1407. Link
564. R. Diks, J. Boslev in T. Uwe, T. Bornscheuer (eds.): *Enzymes in Lipid Modification*, Wiley-VCH, Weinheim 2000.
565. W. van Nieuwenhuyzen: "The Industrial uses of Special Lecithins: A Review", J. Am. Oil. Chem. Soc. **58** (1981) 886 – 888. Link
566. R. Rastall (ed.): *Ingredients Handbook - Enzymes*, Ingredients Handbook Series Leatherhead Food International, Leatherhead 1999.
567. Röhm GmbH, DE 19527274, (B. Sprösler).

Specific References

568. DE 283 923,1913, (O. Röhm).
569. Novo Terapeutisk Laboratoriums AS, Copenhagen, Denmark, US 3 674 643, 1972 (K. Aunstrup; O. Andresen; H. Outtrup).
570. B. Brackmann; A. Viefhaus, SOFW J. **3** (2000) 2 – 12. Link

571. M. Nogueira, HAPPI (1995) no. 8, 42 – 46. Link
572. Procter & Gamble, EP 1048717 A1, 1999 (E. Tcheou).
573. Unilever, WO 00/2208, 2000 (R. J. A. Janssen).
574. Procter & Gamble, EP 09711028 A1, 2000 (J. L. Vega, J. Wevers).
575. R. F. Modler, S. Fenelon, Y. Yoshida, "Surfactants, Household Detergents and their Raw Materials", the *Chemical Economics Handbook*, SRI International, 1998.
576. D. Aaslyng, E. Gormsen, H. Malmos, J. Chem. Technol. Biotechnol. **50** (1991) 321. Link
577. EMPA Eldgenossische Materials Profungs Anstalt, Unterstrasse 11, Postfach, CH-9001, St. Gallen, Switzerland.
578. WFK Forschungsinstitut für Reinigungstechnologie (Germany); URL: htpp://www.gewebe@wfk.de
579. Testfabrics, Inc.; URL: http://www.testfabrics.com
580. TNO Instituut voor Reinigings Techniek-Delft-Netherlands; URL: http://www.tno.nl
581. SDL International Textile Innovators; URL; http://www.textileinnovators.com
582. Terg-o-tometer available from United States Testing Co. Inc., Hoboken, NJ, USA.
583. Launder-o-meter available from Atlas Electric Devices Comp., Chicago, IL, USA.
584. LiniTest available from Heraeus-Original Hanau, D-6460 Hanau, Germany.
585. M. Murata et al., JAOCS **68** (1989) no. 7, 189. Link
586. J. C. Harris, *Detergency Evaluation and Testing*, Interscience Publishers, New York, 1954.
587. Novozymes, Detergent Protease Application Sheet 2001 – 04366-02.
588. Novozymes, Alcalase Application Sheet 2001 – 04375-02.
589. Novozymes, Savinase Application Sheet 2001 – 04379-02.
590. Novozymes, Esperase Application Sheet 2001 – 04377-02.
591. Novozymes, Everlase Application Sheet 2001 – 04378-02.
592. H. S. Olsen, P. Falholt J. Surfactants Deterg. **1** (1998) 555 – 567. Link
593. Procter and Gamble, WO 9 704 067, 1997 (A. C. Baeck, T. W. Horner, R. Ohtani, K. Pramod, and S. Rai,).
594. K. Gibson, *The Importance of Detergent Amylases for Whiteness Maintenance*, Paper Presented at CID, Genoa, October 22 – 24, 1997.
595. Novozymes, Termamyl Application Sheet 2001 – 04368-02.
596. G. Jensen, SOFW J. **123** (1997) 723, 726, 728, 730 – 731. Link
597. Novozymes, Duramyl Application Sheet 2001 – 04367-02.
598. Novozymes, Natalase Application Sheet 2001 – 11810-01.
599. P. H. Nielsen, *Development of New Detergent Enzymes*, Paper Presented at ISSD96, China, 1996.
600. Novo Nordisk, EP 305 216, 1988 (E. Boel, B. Huge-Jensen).
601. Gist-Brocades, EP 334 462, 1989 (P. M. Andreoli et al.).
602. Unilever, EP 407 225, 1991 (A. M. Batenburg et al.).
603. E. Boel et al., *Lipases: Structure, Mechanism and Genetic Engineering*, Vol. 16, GBF Monographs, 1991, pp. 207 – 219.
604. E. Gormsen, JOCS/AOCS Conference in Kuala Lumpur, Malaysia, September 1997 (available from Novo Nordisk A/S as A-6537).
605. D. N. Rubingh, M. D. Bauer, *Tenside Surfactants Deterg 35* (1998) 254 – 256, 258 – 260.
606. D. Aaslyng, E. Gormsen, H. Malmos, J. Chem. Technol. Biotechnol. **50** (1991) 321 – 330. Link
607. J. A. C. Flipsen, A. C. M. Appel, H. T. W. M. Van Der Hijden, C. T. Verrips, Enzyme Microb. Technol. **23** (1998) 274 – 280. Link
608. H. Malmos et al., Chem. Ind. **6** (1990) 183 – 186. Link
609. D. Aaslyng, E. Gormsen, H. Malmos, J. Chem. Tech. Biotechnol. **50** (1991) 321 – 330. Link
610. P. N. Christensen, V. Donelyan, Chim. Oggi **8** (1990) 41. Link
611. S. Munk, P. Munch, Stahnke, Adler-Nissen, P. J. Schieberle, Surfactants Deterg. **3** (2000) 505 – 515. Link

612. P. N. Christensen, E. Gormsen, Chim. Oggi **11** (1993) 15. Link
613. Unilever, EP 0 177 165 B1, 1986 (J. R. Martin, J. R. Nooi).
614. A. Suzuki, *Research Report*, Kao Cooperation Research Laboratories, Beseibutsu, 1989, pp. 100, 104, 105.
615. T. S. Jakobsen, P. Lindegaard, M. Chan, INFORM **9** (1998) 788–792. Link
616. Novozymes, Carezyme Application Sheet 2001 – 04369-02.
617. Procter and Gamble, WO 99/09126, 1998 (J. L. Bettiol, C. Joos).
618. Novozymes, Mannaway Application Sheet 2001 – 04406-02.
619. Procter and Gamble, WO 9 839 402, 1997 (I. Herbots et al.).
620. Procter and Gamble, WO 9 806 805, 1997 (I. Herbots et al.).
621. Procter and Gamble, WO 9 731 999, 1996 (I. Herbots et al.).
622. H. A. Hagen, in R. T. Coffey (ed.) *New Horizons. An AOCS/CSMA Detergent Industry Conference*, AOCS Press, Champaign, 1996, pp. 57 – 62.
623. L. S. Conrad, T. Damhus, O. Kirk, P. Schneider, INFORM **8** (1997) 950 – 957. Link
624. T. Damhus, L. S. Conrad, T. T. Hansen, *An Enzymatic System for Dye Transfer Inhibition in Laundry Detergents, Presented at the 88th AOCS Annual Meeting & Expo*, Seattle, Washington, May 11 – 14, 1997.
625. Procter and Gamble Company, WO 9 731 090, 1996 (I. Herbots et al.).
626. Procter and Gamble Company, WO 9 743 384, 1996 (I. Herbots et al.).
627. Procter and Gamble Company, WO 9 902 632, 1998 (I. Herbots et al.).
628. DRP. 283 923, 1913 (O. Röhm).
629. L. H. Dalgaard, D. Kochavi, M. Thellersen, INFORM AOCS **2** (1991) 532. Link
630. D. H. Cater, M. J. Flynn, P. F. Plank, INFORM AOCS **5** (1994) 1095. Link
631. A. Crutzen, M. L. Douglass: "Detergent Enzyme: A Challenge!" in U. Zoller, G. Broze (eds.): *Handbook of Detergents: Part A: Properties, Surfactant Science Series Vol. 82*, Marcel Dekker, New York, 1999, pp. 639 – 690.
632. Henkel, WO 97/087281, 1995 (J. Haerer, P. Jeschke, Ch. Nitsch).
633. Henkel, WO 99/06515, 1997 (K. Maurer, Ch. Nitsch).
634. M. Bahn, R. D. Schmid, Biotechnology **1** (1987) 119. Link
635. D. Aaslyng, E. Gormsen, H. Malmos, J. Chem. Tech. Biotechnol. **50** (1991) 321. Link
636. R. Siezen, W. M. de Vos, J. A. M. Leunissen and B. W. Dijkstra, Protein Eng. **4** (1991) 719. Link
637. M. Ottesen, J. Svendsen, Methods Enzymol. **19** (1970) 199. Link
638. F. S. Markland, E. L. Smith, in *The Enzymes*, 3rd ed. B. Academic Press, New York, 1971, pp. 561 – 608.
639. I. Svendsen, Carlsberg Res. Commun. **41** (1976) 239. Link
640. R. Siezen: "Subtilases: Subtilisin-like Proteases" in R. Bott, C. Betzel (eds.): *Subtilisin enzymes*, Plenum Press, New York, 1996, 75 – 95.
641. R. Bott: "Development of New Proteases for Detergents", in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency, Surfactant Science Series*, Vol. 69, Marcel Dekker, New York, 1997, pp. 75 – 91.
642. M. R. Egmond: "Application of Proteases in Detergents", in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency, Surfactant Science Series*, Vol. 69, Marcel Dekker, New York, 1997, pp. 61 – 74.
643. H. Gron, L. M. Bech, S. Branner, K. Breddam, Eur. J. Biochem. **194** (1990) 897. Link
644. J. E. Ness, M. Welch, L. Giver, M. Bueno, J. R. Cherry, T. V. Borchert, W. P. C. Stemmer, J. Minshull, Nature Biotechnology **17** (1999) 893. Link
645. Ph. N. Byran, Biochemica et Biophysica Acta **1543** (2000) 203. Link
646. Henkel, WO 92/21760, 1992 (Ch. Paech, R. Wilson et al.).
647. Henkel, WO 95/23221, 1995 (Ch. Paech, K. Maurer, B. Kottwitz, R. Wilson et al.).
648. Novozymes, WO 01/44452, 2000 (T. S. Fano, F. F. Mikkelsen).

649. Novozymes, WO 01/68821, 2000 (T. S. Fano, L. B. Hansen, P. K. Hansen, M. N. Madsen, F. F. Mikkelsen).
650. Novozymes, WO 02/16547, 2000 (H. Outtrup, P. E. Pedersen, M. V. Sorensen).
651. Novozymes, WO 02/31133, 2000 (M. Norregaard-Madsen, L. B. Larsen, P. K. Hansen).
652. M. Zobel: *Die Anwendung der Stärke in der Lebensmittelindustrie*, Paul Parey, Berlin, 1975.
653. H. Uhlig, in: *Enzyme arbeiten für uns*, Carl Hanser Verlag, München, 1991, pp. 33 – 54.
654. H. Upadek, B. Kottwitz: "Application of Amylases in Detergency", in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency, Surfactant Science Series*, Vol. 69, Marcel Dekker, New York, 1997, pp. 203 – 212.
655. W. M. Fogarty, C. T. Kelly: "Starch Degrading Enzymes in Microbial Origin. Part I. Distribution and Characteristics" in M. J. Bull (ed.): *Progress in Industrial Microbiology*, Vol. 15, Elsevier Scientific Publishing, Amsterdam, 1979, pp. 87 – 150.
656. W. M. Fogarty, C. T. Kelly: "Amylases, Amyloglucosidases and Related Glucaneses" in A. H. Rose (ed): *Economic Microbiology*, Vol. 5, Academic Press, London, 1980, pp. 115 – 170.
657. W. M. Fogarty, C. T. Kelly: "Recent Advances in Microbial Amylases" in W. M. Fogarty, C. T. Kelly (eds): *Microbial Enzymes and Biotechnology*, Elsevier Science Publishers, London, 1990, pp. 71 – 132.
658. Kao Corp., WO 94/19468, 1994 (K. Ara, K. Igarashi, K. Saeki, S. Kawai, S. Ito).
659. Kao Corp., JP 6172792, 1994.
660. Kao Corp., JP 6172796, 1994.
661. Kao Corp., JP 6264094, 1994.
662. Novo Nordisk, WO 94/02597, 1994. (A. Svendsen, H. Bisgaard-Frantzen).
663. Genencor International, WO 94/18314, 1994 (R. L. Antrim, C. C. Barnett, C. Mitchinson, S. D. Power, C. A. Requadt, L. Solheim).
664. V. N. Ivanova, E. P. Dobreva, E. I. Emanuilova, J. Biotechnol. **18** (1993) 277. Link
665. T. V. Borchert, S. F. Lassen, A. Svendsen, H. B. Frantzen in S. B. Petersen, B. Svendsen, S. Pedersen (eds.): *Carbohydrate Bioengineering*, Elsevier, Amsterdam, 1995, 175.
666. W. Aehle: "Development of New Amylases" in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency, Surfactant Science Series*, Vol. 69, Marcel Dekker, New York, 1997, pp. 213 – 229.
667. M. Hashida, H. Bisgaard-Frantzen, Trends Glycosci. Glycotechnol. **12/68** (2000) 389. Link
668. J. E. Nielsen, T. V. Borchert, Biochemica et Biophysica Acta **1543** (2000) 253. Link
669. Henkel, WO 02/10356, 2000 (R. Breves, K. Maurer, B. Kottwitz, I. Schmidt et al.).
670. Henkel KGaA, WO 02/44350, 2000 (R. Breves, K. Maurer, B. Kottwitz, I. Schmidt et al.).
671. Procter & Gamble, EP 694058, 1994 (E. S. Sadlowski).
672. Unilever, EP 766726, 1995 (J. O. Carnali, J. W. Gordon).
673. Procter & Gamble, EP 662120, 1994 (B. de Smet, L. Jones, J. Pluyter, J. Christie et al.).
674. Unilever, EP 271155, 1986 (W. R. van Dijk).
675. Procter & Gamble, US 6376445, 2002 (J. L. Ph. Bettiol, M. S. Showell).
676. Th. Müller-Kirschbaum, J. Härer, Ch. Nitsch, *Lecture at the WfK-Conference 1999*, Krefeld, Germany.
677. ARGE Prüfgemeinschaft: Report-No. R 3001/09.94, 1994.
678. HEA (Hauptberatungsstelle für Elektrizitätsanwendung e.V.): *Informationsreihe, Elektriziätsanwendung im Haushalt, Geräte zur Geschirreinigung*, Energie-Verlag, Heidelberg.
679. H. Magg, G. Waeschenbach: "Application of Enzymes in Automatic-Dishwashing Detergents" in J. H. van Ee, O. Misset, E. J. Baas (eds.): *Enzymes in Detergency, Surfactant Science Series*, Vol. 69, Marcel Dekker, New York, 1997, pp. 231 – 250.
680. IKW-AK Automatic Dishwashing Detergents, *Seifen Oele Fette Wachse* 125/11 (1999) 52.
681. IKW-AK Automatic Dishwashing Detergents, *Seifen Oele Fette Wachse* 126/3 (2000) 40.
682. 2002 Corn Annual, Corn Refiners Association, Inc., Washington 2002.
683. S. Iwasa, H. Aoshima, K. Hiromi, H. Hitano, J. Biochem. **75** (1974) 969 – 978. Link

684. J. A. Thoma, J. D. Allen, Carbohydr. Res. **48** (1976) 105 – 124. Link
685. E. M. Torgerson, L. C. Brewer, J. A. Thoma, Arch. Biochem. Biophys. **196** (1979) 13 – 22. Link
686. H. Kondo, H. Nakatani, R. Matsuno, K. Hiromi, J. Biochem. **87** (1980) 1053 – 1070. Link
687. D. French, E. E. Smith, W. J. Whelan, Carbohydr. Res. **22** (1972) 123 – 134. Link
688. K. Umeki, T. Yamamoto, J. Biochem. **72** (1972) 101 – 109. Link
689. K. M. Bendetskii, V. L. Yarovenko, Biokhimiya **38** (1973) 568 – 572. Link
690. P. M. Coutinho, B. Henrissat: Carbohydrate-Active Enzymes Server at `http://afmb.cnrs-mrs.fr`, 1999.
691. P. M. Coutinho, B. Henrissat in H. J. Gilbert, G. Davies, B. Henrissat, B. Svensson (eds.): *Recent Advances in Carbohydrate Bioengineering*, The Royal Society of Chemistry, Cambridge 1999, pp. 3 – 12.
692. M. Machius, G. Weigand, R. Huber, J. Mol. Biol. **246** (1995) 545 – 559. Link
693. K. Y. Hwang, H. K. Song, C. Chang, J. Lee, S. Y. Lee, K. K. Kim, S. Choe, R. M. Sweet, S. W. Suh, Mol. Cells **7** (1997) 251 – 258. Link
694. D. Suvd, K. Takase, Z. Fijimoto, M. Matsumura, H. Mizuno, Acta Crystallogr. D **56** (2000) 200 – 202. Link
695. Z. Fujimoto, K. Takase, N. Doui, M. Momma, T. Matsumoto, H. Mizuno, J. Mol. Biol. **277** (1998) 393 – 407. Link
696. N. Aghajari, G. Feller, C. Gerday, R. Haser, Protein Sci. **7** (1998) 564 – 572. Link
697. S. Kamitori, A. Abe, A. Ohtaki, A. Kaji, T. Tonozuka, Y. Sakano, J. Mol. Biol. **318** (2002) 443 – 453. Link
698. E. Boel, L. Brady, A. M. Brzozowski, Z. Derewanda, V. J. Jensen, S. B. Petersen, H. Swift, L. Thim, H. F. Woldike, Biochemistry **29** (1990) 6244 – 6249. Link
699. A. M. Brzozowski, G. J. Davis, Biochemistry **36** (1997) 10 837 – 10 845. Link
700. A. Kadziola, J.-I. Abe, B. Svensson, A. Haser, J. Mol. Biol. **239** (1994), 104 – 121. Link
701. S. Strobl, K. Maskos, M. Betz, G. Wiegand, R. Huber, F. X. Gomis-Ruth, R. Glockshuber, J. Mol. Biol. **278** (1998) 617 – 628. Link
702. N. Ramasubbu, C. Ragunath, Z. Wang, PDB accession **1JXJ** (2001). Link
703. G. D. Brayer, Y. Luo, S. G. Withers, Protein Sci. **4** (1995) 1730 – 1742. Link
704. M. Qian, R. Haser, G. Buisson, E. Duée, F. Payan, Biochemistry **33** (1994) 6284 – 6294. Link
705. G. Pujadas, J. Palau, Mol. Biol. Evol. **18** (2001) 38 – 54. Link
706. M. M. Meagher, P. J. Reilly, Biotechnol. Bioeng. **34** (1989) 689 – 693. Link
707. M. M. Meagher, Z. L. Nikolov, P. J. Reilly, Biotechnol. Bioeng. **34** (1989) 681 – 688. Link
708. P. J. Reilly in J. R. Whitaker, A. G. J. Voragen, D. W. S. Wong (eds.): *Handbook of Food Enzymology*, Marcel Dekker, New York, 2003, pp. 727 – 738.
709. Z. L. Nikolov, M. M. Meagher, P. J. Reilly, Biotechnol. Bioeng. **34** (1989) 694 – 704. Link
710. A. Aleshin, A. Golubev, L. M. Firsov, R. B. Honzatko, J. Biol. Chem. **267** (1992) 19291 – 19298. Link
711. A. E. Aleshin, L. M. Firsov, R. B. Honzatko, J. Biol. Chem. **269** (1994) 15 631 – 15 639. Link
712. J. Šev čik, A. Solovicová, E. Hostinová, J. Gašperík, K. S. Wilson, Z. Dauter, Acta Crystallogr. D **54** (1998) 854 – 866. Link
713. A. Aleshin, P.-H. Feng, R. B. Honzatko, P. J. Reilly, J. Mol. Biol., **327** (2003) 61 – 73. Link
714. P. M. Coutinho, P. J. Reilly, Proteins **29** (1997) 334 – 347. Link
715. K. Sorimachi, A. J. Jacks, M.-F. Le Gal-Coëffet, G. Williamson, D. B. Archer, M. P. Williamson, J. Mol. Biol. **259** (1996) 970 – 987. Link
716. M. R. Sierks, C. Ford, P. J. Reilly, B. Svensson, Protein Eng. **3** (1990) 193 – 198. Link
717. E. M. S. Harris, A. E. Aleshin, L. M. Firsov, R. B. Honzatko, Biochemistry **32** (1993) 1618 – 1626. Link
718. B. K. Dalmia, Z. L. Nikolov, Enzyme Microb. Technol. **13** (1991) 982 – 990. Link

719. M. Abdullah, D. French, Arch. Biochem. Biophys. **137** (1970) 483 – 493. Link
720. R. van Tilburg in G. M. A. van Beynum, J. A. Roels (eds.): *Starch Conversion Technology*, Marcel Dekker, New York, 1985, pp. 175 – 236.
721. R. L. Antrim, W. Colilla, B. J. Schnyder in L. B. Wingard, Jr., E. P. Katchalski-Katzir, L. Goldstein (eds.): *Applied Biochemistry and Bioengineering*, Vol. 2, "Enzyme Technology", Academic Press, New York, 1979, pp. 97 – 155.
722. F. Rey, J. Jenkins, J. Janin, I. Lasters, P. Alard, M. Claeyssens, G. Matthyssens, S. Wodek, Proteins **4** (1988) 165 – 172. Link
723. N. T. Mrabet, A. Van den Broeck, I. Van den Brande, P. Stanssens, Y. Laroche et al., Biochemistry **31** (1992) 2239 – 2253. Link
724. J. Jenkins, J. Janin, F. Rey, M. Chiadmi, H. van Tilbeurgh et al., Biochemistry **31** (1992) 5449 – 5458. Link
725. A. M. Lambeir, M. Lauwereys, P. Stanssens, N. T. Mrabet, J. Snauwaert et al., Biochemistry **31** (1992) 5459 – 5466. Link
726. H. van Tilbeurgh, J. Jenkins, M. Chiadmi, J. Janin, S. J. Wodak, N. T. Mrabet, A. M. Lambeir, Biochemistry **31** (1992) 5467 – 5471. Link
727. K. Henrick, D. M. Blow, H. L. Carrell, J. P. Glusker, Protein Eng. **1** (1987) 467 – 469. Link
728. K. Henrick, C. A. Collyer, D. M. Blow, J. Mol. Biol. **208** (1989) 129 – 157. Link
729. C. A. Collyer, D. M. Blow, Proc. Nat. Acad. Sci. U.S.A. **87** (1990) 1362 – 1366. Link
730. C. A. Collyer, K. Henrick, D. M. Blow, J. Mol. Biol. **212** (1990) 211 – 235 Link
731. O. Gallay, R. Chopra, E. Conti, P. Brick, D. Blow, PDB accessions 1A0C, 1A0D, 1A0E (1997).
732. Z. Dauter, H. Terry, H. Witzel, K. S. Wilson, Acta Crystallogr. B **46** (1990) 833 – 841. Link
733. X. Zhu, M. Teng, L. Niu, C. Xu, Y. Wang, Acta Crystallogr. D **56** (2000) 129 – 136. Link
734. H. Rasumssen, T. La Cour, J. Nyborg, M. Schülein, PDB accession **1DXI** (1993). Link
735. K. N. Allen, A. Lavie, G. K. Farber, A. Glasfeld, G. A. Petsko, D. Ringe, Biochemistry **33** (1994) 1481 – 1487. Link
736. K. N. Allen, A. Lavie, A. Glasfeld, T. N. Tanada, D. P. Gerrity, S. C. Carlson, G. K. Farber, G. A. Petsko, D. Ringe, Biochemistry **33** (1994) 1488 – 1494. Link
737. A. Lavie, K. N. Allen, G. A. Petsko, D. Ringe, Biochemistry **33** (1994) 5469 – 5480. Link
738. K. Henrick, D. M. Blow, H. L. Carrell, J. V. Glusker, Protein Eng. **1** (1987) 467 – 469. Link
739. H. L. Carrell, J. P. Glusker, V. Burger, F. Manfre, D. Tritsch, J. F. Biellmann, Proc. Natl. Acad. Sci. U.S.A. **86** (1989) 4440 – 4444. Link
740. M. Whitlow, A. J. Howard, B. C. Finzel, T. L. Poulos, E. Winborne, G. L. Gilliland, Proteins **9** (1991) 153 – 173. Link
741. C. Chang, B. C. Park, D. S. Lee, S. W. Suh, J. Mol. Biol. **288** (1999) 623 – 634. Link
742. H. F. Zobel in F. W. Schenck, R. E. Hebeda (eds.): *Starch Hydrolysis Products*, VCH, New York, 1992, pp. 36 – 38.
743. C. Sheppy: "The Current Feed Enzyme Market and Likely Trends," in M. R. Bedford, G. G. Partridge (eds.): *Enzymes in Farm Animal Nutrition*, CABI Publishing, Bath 2001.
744. M. R. Bedford: "Exogenous Enzymes in Monogastric Nutrition - their Current Value and Future Benefits", Anim. Feed Sci. Technol. **86** (2000) 1 – 13. Link
745. M. R. Bedford, G. G. Partridge (eds.): *Enzymes in Farm Animal Nutrition*, CABI Publishing, Bath 2001.
746. M. R. Bedford, H. Schulze: "Exogenous Enzymes for Pigs and Poultry", Nutrit. Res. Rev. **11** (1998) 91 – 114. Link
747. A. Sunna, G. Antranikian: "Xylanolytic Enzymes from Fungi and Bacteria", Crit. Rev. Biotechnol. **17** (1997) 39 – 67. Link
748. K. K. Y. Wong, J. N. Saddler: "Trichoderma Xylanases, Their Properties and Application", in J. Visser, M. A. Kusters van Someren, G. Beldman, A. G. T. Voragen (eds.): *Xylans and Xylanases*, Elsevier Health Sciences, Amsterdam 1992.

749. G. M. Miller: “Use of Dinitrosalicyclic Acid Reagent for Determination of Reducing Sugar”, Anal. Chem. **31** (1959) 426 – 428 Link
750. B. V. McCleary, “Analysis of Feed Enzymes” in M. R. Bedford, G. G. Partridge (eds.): *Enzymes in Farm Animal Nutrition*, CABI Publishing, Bath 2001.
751. W. Debyser, W. J. Peumans, E. J. M. Van Damme, J. A. Delcour, “Triticum aestivum Xylanase Inhibitor (TAXI), a New Class of Enzyme Inhibitor Affecting Breadmaking Performance”, J. Cer. Sci. **30**, (1999) 39 – 43. Link
752. B. V. McCleary, “Problems in the Measurement of β-Xylanase, β-Glucanase and α-Amylase in Feed Enzymes and Animal Feeds”, *Proceedings of the 2nd European Symposium on Feed Enzymes*, Noordwijkerhout, 1995, pp. 135 – 141.
753. G. L. Campbell, M. R. Bedford, “Enzyme Applications for Monogastric Feeds: A Review”, Can. J. Anim. Sci. **72** (1992) 449 – 466. Link
754. G. L. Campbell, B. G. Rossnagel, H. L. Classen, P. A. Thacker, “Genotypic and Environmental Differences in Extract Viscosity of Barley and their Relationship to its Nutritive Value for Broiler Chickens”, Anim. Feed Sci. Technol. **26** (1989) 221 – 230. Link
755. T. Cosson, A. M. P. Vendrell, B. G. Teresa, D. Rene, P. Taillade, J. Brufau, “Enzymatic Assays for Xylanase and β-Glucanase Feed Enzymes” Anim. Feed Sci. Technol. **77**, (1999) 345 – 353. Link
756. V. Ravindran, W. L. Bryden, E. T. Kornegay, “Phytates: Occurrence, Bioavailability and Implications in Poultry Nutrition”, Poultry Avian Biol. Rev. **6** (1995) 125 – 143. Link
757. M. Torre, A. R. Rodriguez, F. Saura-Calixto, “Effects of Dietary Fiber and Phytic Acid on Mineral Availability”, Crit. Rev. Food Sci. Nutrit. **1** (1991) 1 – 22. Link
758. O. Adeola, “Digestive Utilization of Minerals by Weanling Pigs fed Copper- and Phytase-Supplemented Diets”, Can. J. Anim. Sci. **75** (1995) 603 – 610. Link
759. B.-L. Liu, A. Rafiq, Y.-M. Tzeng, A. Rob, “The Induction and Characterization of Phytase and Beyond”, Enzyme Microbial Technol. **22** (1998) 415 – 424. Link
760. A. J. Engelen, F. C. van der Heeft, P. H. G. Randsdorp, E. L. C. Smit, “Simple and Rapid Determination of Phytase Activity”, AOAC International **77** (1994). Link
761. I. E. Liener, “Implications of Antinutritional Components in Soybean Foods”, Crit. Rev. Food Sci. Nutrit. **34** (1994) 31 – 67. Link
762. Y. Noy, D. Sklan, “Digestion and Absorption in the Young Chick”, Poultry Sci. **74** (1995) 366 – 373. Link
763. D. M. Denbow, E. A. Grabau, G. H. Lacy, E. T. Kornegay, D. R. Russell, P. F. Umbeck, “Soybeans Transformed with a Fungal Phytase Gene Improve Phosphorus Availability for Broilers”, Poultry Sci. **77** (1998) 878 – 881. Link
764. H. Brinch-Pedersen, S. K. A. Olesen, Rasmussen, P. B. Holm, “Generation of Transgenic Wheat (Triticum aestivum L.) for Constitutive Accumulation of an Aspergillus Phytase”, Mol. Breeding **6** (2000) 195 – 206. Link
765. T. C. Verwoerd, P. A. Van Paridon, A. J. J. Van Ooyen, J. W. M. Van Lent, A. Hoekema, J. Pen, “Stable Accumulation of Aspergillus niger Phytase in Transgenic Tobacco Leaves”, Plant Physiol. **109** (1995) 1199 – 1205. Link
766. J. Neves, *Tecnologia Textil*, LLS Ed., Porto, Portugal, 1982, Vol. 1.
767. R. Peters, *Textile Chemistry*, Elsevier, Amsterdam, London, New York, 1963, Vols. 1–2.
768. E. Heine and H. Hocker, Rev. Prog. Coloration **25** (1995) 57 – 63. Link
769. J. Cegarra, J. Soc. Dyers Colourists **112**, (1996) 326 – 329. Link
770. A. Cavaco-Paulo in K.-E. Eriksson, A. Cavaco-Paulo (eds.): “Enzyme Applications in Fiber Processing”, *ACS Symposium Series 687*, 1998, Chap. 15, 180 – 189.
771. N. Sekar, *Colourage 1999*, no. 1, 27 – 32.
772. M. Van Parys, *Unitex 5* (1999) 50 – 56.
773. J. N. Etters, P. A. Annis, *Am. Dyestuff Reporter*, 1998, no. 5, 18 – 23.
774. U. Rößner, *Melliand Textilber. 2* (1993) 144 – 148.
775. J. Buchert, J. Pere, A. Puolakka and P. Nousiainen, *Book of Papers of the International Conference & Exhibition of the AATCC*, Philadelphia, 1998, pp. 493 – 499.

776. C.-H. Lin, Y.-L. Hsieh, Textile Res. J. **71** (2001) 425 – 434. Link
777. M. M. Hartzell, Y.-L. Hsieh, Textile Res. J. **68** (1998) 233 – 241. Link
778. Y. Li, I. R. Hardin, Textile Res. J. **68** (1998) 671 – 679. Link
779. M. K. Traore, G. Buschle-Diller, TCC & ADR **32** (2000) 40 – 43. Link
780. P. Husain, N. K. Lange, L. Henderson, J. Liu, B. Condon, *Book of Papers of the International Conference & Exhibition of the AATCC*, Charlotte, 1999, 170 – 182.
781. Nova Nordisk, WO 98/24965, 1998, (C. A. Miller, S. S. Jørgensen, E. W. Otto, N. C. Lange, B. Condon, J. Liu).
782. N. K. Lange, L. Henderson, J. Liu, P. Husain, B. Condon, *Book of the 18th IFATCC Congress*, 1999, pp. 4 – 8.
783. R. B. Waddell, AATCC Review **2** (2002) no. 4, 28 – 30. Link
784. E. Csiszar, G. Szakacs, I. Rusznak, Textile Res. J. **68** (1998) 163 – 167. Link
785. T. Takagishi, R. Yamamoto, K. Kikuyama, H. Arakawa AATCC Review **1** (2001) no. 8, 32 – 34. Link
786. H. Schacht, W. Kesting, E. Schollmeyer, Textilveredlung **30** (1995) 237 – 243. Link
787. T. Tzanov, M. Calafell, G. M. Guebitz, A. Cavaco-Paulo, Enzyme Microbial Technol. **29** (2001) 357 – 362. Link
788. G. Buschle-Diller, X. D. Yang, R. Yamamoto, Textile Res. J. **71** (2001) 388 – 394. Link
789. T. Tzanov, S. A. Costa, G. M. Guebitz, A. Cavaco-Paulo, J. Biotechnol. **93** (2002) 87 – 94. Link
790. N. K. Lange, J. Liu, P. Husain, B. Condon, *Novo Nordisk Enzyme Business Article A6611a-GB*, 1999.
791. G. K. Hall, *Book of Papers of the International Conference & Exhibition of the AATCC*, Atlanta, 1995, p. 581.
792. M. Schmidt, *Book of Papers of the International Conference & Exhibition of the AATCC*, Atlanta, 1995, pp. 248 – 255.
793. Biotimes **3** (2001) (Novozymes). Link
794. A. Cavaco-Paulo, L. Almeida, Textile Res. J. **66** (1996) 287 – 294. Link
795. A. Cavaco-Paulo, L. Almeida, Textile Chem. Colorist **28** (1996) 28 – 32. Link
796. A. Cavaco-Paulo, L. Almeida, D. Bishop, Textile Res. J. **68** (1998) 273 – 280. Link
797. M. K. Traore, G. Buschle-Diller, TCC & ADR **1** (1999) 51 – 56. Link
798. A. Cavaco-Paulo, J. Cortez, L. Almeida, J. Soc. Dyers Colorists **117** (1997) 17 – 21. Link
799. H. Azevedo, D. Bishop, A. Cavaco-Paulo, Enzyme Microbiol. Technol. **27** (2000) 325 – 329. Link
800. G. L. Pedersen, G. A. Screws, D. M. Cedroni, Can. Text. J. **109** (1992) no. 10, 31. Link
801. Novo Nordisk, WO 9 320 278, 1993 (T. Videbaek, L. D. Andersen).
802. G. L. Pedersen, G. A. Screws, D. M. Cedroni, Melliand Textilber. **74** (1993) 1277 – 1280. Link
803. Genencor, WO 9 206 183, 1992 (K. A. Clarkson, G. L. Weise, E. A. Larenas).
804. C. Li, C. M. Ladisch, M. R. Ladisch, Textile Res. J. **71** (2001) 407 – 414. Link
805. J. Buchert, L. Heikinheimo, Carbohydr. Eur. **22** (1998) 32 – 34. Link
806. A. Kumar, A. Harnden, *Book of Papers of the International Conference & Exhibition of the AATCC*, Philadelphia, 1998, pp. 482 – 492.
807. J. Liu, E. Otto, N. K. Lange, P. Husain, B. Condon, *Book of Papers of the International Conference & Exhibition of the AATCC*, Philadelphia, 1998, pp. 445 – 454.
808. J. Liu, E. Otto, N. K. Lange, P. Husain, B. Condon, H. Lund, TCC & ADR **32** (2000) 30 – 36. Link
809. A. Miettinen-Oinonen, L. Heikinheimo, J. Buchert, J. Morgado, L. Almeida, P. Ojapalo, A. Cavaco-Paulo, AATCC Rev. **1** (2001) no. 1, 33 – 35. Link
810. A. Cavaco-Paulo, Carbohydr. Polym. **37** (1998) 273 – 277. Link
811. J. Morgado, A. Cavaco-Paulo, M. A. Rouselle, Textile Res. J. **70** (2000) 696 – 699. Link
812. H. B. M. Lenting, M. M. C. G. Warmoeskerken, J. Biotechnol. **89** (2001) 217 – 226. Link
813. G. Buschle-Diller, S. H. Zeronian, N. Pan, M. Y. Yoon, Textile Res. J. **64** (1994) 270 – 279. Link

814. B. P. Hudson, A. C. Clapp, D. Kness in B. Rosenberg, K. Kaylor (eds.): *Joseph's Introductory in Textile Science*, 6th ed., Harcourt Brace, Orlando, FL, 1993, pp. 76 – 88.
815. H. B. M. Lenting, M. M. C. G. Warmoeskerken, J. Biotechnol. **89** (2001) 227 – 232. Link
816. J. N. Etters, Textile Chemist & Colorist **27** (1995) 17 – 22. Link
817. A. Cavaco-Paulo, L. Almeida, D. Bishop, Textile Chem. Colorist **28** (1996) 28 – 32. Link
818. L. Heikinheimo, J. Buchert, A. Miettinen-Oinonen, P. Suominen, Textile Res. J. **70** (2000) 969 – 973. Link
819. J. Andreaus, R. Campos, A. Cavaco-Paulo, Melliand Textilber. **82** (2001) 807 – 810. Link
820. A. Cavaco-Paulo, J. Morgado, L. Almeida, D. Kilburn, Textile Res. J. **68** (1998) 398 – 401. Link
821. J. Andreaus, R. Campos, G. Guebitz, A. Cavaco-Paulo, Textile Res. J. **70** (2000) 628 – 632. Link
822. Genencor, WO 94/29426, (K. Clarkson, P. Lad, M. Mullins, C. Simpsom, G. Weiss, L. Jacobs).
823. Novozymes, Productsheets from DeniLite II Base and Assist, document no. 2001 – 07252-01 and 02, 2001.
824. M. Mueller, C. Shi, *CD-rom of the Int. Conference & Exhibition of the AATCC*, Winston Salem, 2000.
825. M. van Parys, Unitex **5** (1999) 50 – 56. Link
826. R. Campos, A. Cavaco-Paulo, K.-H. Robra, M. Schneider, G. Guebitz, Textile Res. J. **71** (2001) 420 – 424. Link
827. Faserstoff-Tabellen nach P.-A. KOCH, Melliand Textilber. 78 (1997) 575 – 581.
828. H. J. Koslowski in *Dictionary of Man-Made Fibres*, Intern. Business Press, Francfustam Main, 1998, pp. 106 – 109.
829. T. Hohberg, S. Thumm, Melliand Textilber. **79** (1998) 253 – 256. Link
830. K. Gandhi, S. M. Burkinshaw, J. M. Taylor, G. W. Collins, AATCC Rev. **2** (2002) no. 4, 48 – 52. Link
831. S. Picht, Melliand Textilber. **78** (1997) 718 – 719. Link
832. A. Kumar, A. Harnden, *Book of Papers of the International Conference & Exhibition of the AATCC*, Philadelphia, 1998, pp. 482 – 492.
833. A. Kumar, A. Harnden, Textile Chem. Colorist **31** (1999) no. 9, 37 – 41. Link
834. D. E. Akin, D. Slomczynski, L. L. Rigsby, Textile Res. J. **72** (2002) 27 – 34. Link
835. D. E. Akin, L. L. Rigsby, Textile Res. J. **69** (1999) 747 – 753. Link
836. D. E. Akin, J. A. Foulk, R. B. Dodd, D. D. McAlister, J. Biotechnol. **89** (2001) 193 – 203. Link
837. D. E. Akin, R. B. Dodd, W. Perkins, G. Henriksson, K.-E. L. Eriksson, Textile Res. J. **70** (2000) 486 – 494. Link
838. Faserstoff-Tabellen nach P.-A. Koch, Chemiefasern/Textilind. **41** (1991) 521 – 553. Link
839. R. Breier, Textilveredlung **37** (2002) 5 – 10. Link
840. D. Bishop, J. Shen, E. Heine, B. Hollfelfder, J. Text. Inst. **89** (1998) 546 – 553. Link
841. J. Shen, D. Bishop, E. Heine, B. Hollfelfder, J. Text. Inst. **90** (1999) 404 – 411. Link
842. P. Jovancic, D. Jocic, P. Erra, R. Molina, M. R. Julia, Eur. Microscop. Anal. **51** (1998) 15 – 16. Link
843. K. Schumacher, E. Heine, H. Hocker, J. Biotechnol. **89** (2001) 281 – 288. Link
844. J. M. Cardamone, AATCC Rev. **2** (2002) no. 5, 30 – 35. Link
845. K. Sun, W. Zhou, J. Wang, *Book of Papers of the International Conference & Exhibition of the AATCC*, Philadelphia, 1998, 299 – 308.
846. P. Jovancic, D. Jocic, R. Molina, M. R. Julia, P. Erra, Textile Res. J. **71** (2001) 948 – 953. Link
847. Faserstoff-Tabellen nach P.-A. Koch, *Tafel Seide*, Ausgabe 1994, ed. by Institute of Textile Technology Aachen; distribution via Chem Fibres Int., Mainzer Landstrasse 251, D-60326 Frankfurt/Main, Germany.
848. S. Chopra, M. L. Gulrajani, Asian Textile J. **2** (1994) no. 3, 39 – 46. Link
849. M. L. Gulrajani, S. V. Gupta, A. Gupta, M. Suri, Ind. J. Fibre Textile Res. **21** (1996) no. 4, 270 – 275. Link

850. M. L. Gulrajani, S. Sen, A. Soria, M. Suri, Ind. J. Fibre Textile Res. **23** (1998) no. 1, 52 – 58. Link
851. M. L. Gulrajani, S. Agarwal, R. Agarwal, Ind. J. Fibre Textile Res. **24** (1999) no. 3, 226 – 228. Link
852. M. L. Gulrajani, R. Agarwal, A. Grover, M. Suri, Ind. J. Fibre Textile Res. **25** (2000) no. 1, 69 – 74. Link
853. I. R. Hardin, H. Cao, S. Wilson, *Book of Papers of the International Conference & Exhibition of the AATCC*, Charlotte, 1999, pp. 474 – 478.
854. H. Cao, I. R. Hardin, *CD-ROM of the International Conference & Exhibition of the AATCC*, Winston Salem, 2000.
855. C. Novotny, B. Rawal, M. Bhatt, M. Patel, V. Sasek, H. P. Molitoris, J. Biotechnol. **89** (2001) 113 – 122. Link
856. M. Adosinda, M. Martins, I. C. Ferreira, I. M. Santos, M. J. Queiroz, N. Lima, J. Biotechnol. **89** (2001) 91 – 98. Link
857. I. Mielgo, M. T. Moreira, G. Feijoo, J. M. Lema, J. Biotechnol. **89** (2001) 99 – 106. Link
858. G. M. B. Soares, M. T. Pessoa de Amorim, M. Costa-Ferreira, J. Biotechnol. **89** (2001) 123 – 129. Link
859. C. A. Metosh-Dickey, T. M. Davis, C. A. McEntire, J. Christopher, H. DeLoach, R. J. Portier, *Book of Papers of the International Conference & Exhibition of the AATCC*, Charlotte, 1999, pp. 334 – 339.
860. D. S. L. Balan, R. T. R. Monteiro, J. Biotechnol. **89** (2001) 141 – 145. Link
861. R. Campos, A. Kandelbauer, K. H. Robra, A. Cavaco-Paulo, G. M. Guebitz, J. Biotechnol. **89** (2001) 131 – 139. Link
862. T. Tzanov, A. Abreu, A. Carneiro, S. Costa, G. Guebitz, A. Cavaco-Paulo, *CD-ROM of the International Conference & Exhibition of the AATCC*, Winston Salem, 2000.
863. T. Tzanov, S. Costa, G. Guebitz, A. Cavaco-Paulo, *CD-ROM of the Int. Conference & Exhibition of the AATCC*, Winston Salem, 2000.
864. T. Tzanov, S. Costa, A. Cavaco-Paulo, G. M. Guebitz, Textile Chem. Colorist **1** (2001) no. 10, 25 – 28. Link
865. A. Paar, S. Costa, T. Tzanov, M. Gudelj, K. H. Robra, A. Cavaco-Paulo, G. M. Guebitz, J. Biotechnol. **89** (2001) 147 – 153. Link
866. H. Feitkenhauer, U. Meyer, J. Biotechnol. **89** (2001) 185 – 192. Link
867. Y. L. Hsieh, L. A. Cram, Textile Res. J. **68** (1998) 311 – 319. Link
868. T. Deguchi, Y. Kitaoka, M. Kakezawa, T. Nishida, Appl. Environ. Microbiol. **64** (1998) 1366 – 1371. Link
869. S. Kakudo, S. Negoro, I. Urabe, H. Okada, Appl. Environ. Microbiol. **59** (1993) 3978 – 3980. Link
870. M. Tauber, G. Guebitz, A. Cavaco-Paulo, *CD-ROM of the International Conference & Exhibition of the AATCC*, Winston Salem, 2000.
871. E. Sjöström, *Wood Chemistry, Fundamentals and Applications*. Academic Press, San Diego, CA, 1993, pp. 293.
872. J. Ståhlberg, G. Johansson, G. Pettersson, Biochim. Biophys. Acta **1157** (1993) 107 – 113. Link
873. P. Tomme, R. Warren, R. Gilkes, Adv. Microb. Physiol. **37** (1995) 1 – 81. Link
874. P. Biely, M. Vrsanska M. Claeyssens, Eur. J. Biochem. **200** (1991) 157 – 163. Link
875. H. Van Tilbeurgh, M. Clayssens, R. Bhikhabhai, G. Pettersson, FEBS Lett. **204** (1986) 223 – 227. Link
876. P. Tomme, H. van Tilbeurgh, G. Pettersson, J. van Damme, J. Vandekerckhove, J. Knowles, T. Teeri, M. Claeyssens, Eur. J. Biochem. **170** (1988) 575 – 581. Link
877. M. Linder, T. Teeri, J. Biotechnol. **57** (1997) 15 – 28. Link
878. A. Teleman, V. Harjunpää, M. Tenkanen, J. Buchert, T. Hausalo, T. Drakenberg, T. Vuorinen, Carbohydr. Res. **272** (1995) 55 – 71. Link
879. P. Biely, Trends Biotechnol. **3** (1985) 286 – 290. Link
880. K. Wong, J. Saddler, Critical Revs. Biotechnol. **12** (1992) 413 – 435. Link

881. M. Coughlan, G. Hazlewood, Biotechnol. Appl. Biochem. **17** (1993) 259 – 289 Link
882. L. Viikari, A. Kantelinen, J. Sundquist, M. Linko FEMS Microbiol. Rev. **13** (1994) 335 – 350. Link
883. A. Sunna, G. Antranikian, Crit. Rev. Biotechnol. **17** (1997) 39 – 67. Link
884. L. Viikari, M. Tenkanen, K. Poutanen in H.-J. Rehm, G. Reed (eds): *Biotechnology, Vol 10: Special Processes*, Wiley-VCH, Weinheim-New York, 2001.
885. A. Törrönen, A. Harkki, J. Rouvinen, EMBO J. **13** (1994) 2493 – 2501 Link
886. D. Irwin, E. Jung, D. Wilson, Appl. Environ. Microbiol. **3** (1994) 763 – 770 Link
887. H. Gilbert, G. Hazlewood J. Gen. Microbiol. **139** (1993) 187 – 194. Link
888. K. Sakka, Y. Kojima, T. Kondo, S. Karita, K. Ohmiya, K. Shimada, Biosci. Biotech. Biochem. **57** (1993) 273 – 277. Link
889. J. Rixon, J. Clarke, G. Hazlewood, R. Hoylamd, A. McCarthy, H. Gilbert, Appl. Microbiol. Biotechnol. **46** (1996) 514 – 520. Link
890. A. Lapidot, A. Mechaly, Y. Shohom, J. Biotech. **51** (1996) 259 – 264. Link
891. A. Garg, A. Mc Carthy, J. Roberts, J. C. Enz, Microb. Technol. **18** (1996) 261 – 267. Link
892. A. Shah, D. Cooper, R. Adolphson, K.-E. Eriksson, J. Pulp Paper Sci. **26** (2000) 8 – 11. Link
893. H. Stålbrand, A. Saloheimo, J. Vehmaanperä, B. Henrissat, M. Penttilä, Appl. Environ. Microbiol. **61** (1995) 1090 – 1097. Link
894. M. Tenkanen, J. Buchert, L. Viikari, Enzyme Microb. Technol. **17** (1995) 499 – 505. Link
895. E. Sabini, H. Schubert, G. Murshudov, K. Wilson, M. Siika-aho, M. Penttilä, M. Acta Cryst. **D56** (2000) 3 – 13. Link
896. B. McCleary, ACS Symp. Ser. **460** (1991) 437 – 449. Link
897. I. Kusakabe, G. Park, N. Kumita, T. Yasui, K. Murakami, Agr. Biol. Chem. **52** (1988) 519 – 524. Link
898. M. Tenkanen, M. Makkonen, M. Perttula, L. Viikari, A. Teleman J. Biotechnol. **57** (1997) 191 – 204. Link
899. R. Atalla, *ISWPC Proceedings Vol I* (1995) 77 – 84.
900. A. Heinfling, F. Ruiz-Duenas, M. Martinez, M. Bergbauer, U. Szewzyk, A. Martinez, FEBS Lett. **428** (1998) 141 – 146. Link
901. M. Arbeloa, J. de Leseleuc, G. Goma, J.-C. Pommier, Tappi J. **75** (1992) 215 – 221. Link
902. M. Paice, I. Reid, R. Bourbonnais, F. Archibald, L. Jurasek, Appl. Environ. Microbiol. **59** (1993) 260 – 265. Link
903. R. Kondo, K. Harazono, K. Sakai, Appl. Env. Microb. **60** (1994) 4359 – 4363. Link
904. L. Banci, J. Biotechnol. **53** (1997) 253 – 263. Link
905. D. Li, N. Li, B. Ma, M. Mayfield, M. Gold, Biochim. Biophys. Acta **1434** (1999) 356 – 364. Link
906. S. Lobos, L. Larraondo, L. Salas, E. Karahanian, R. Vicuna, Gene **206** (1998) 185 – 193. Link
907. M. Tello, G. Corsini, L. Larrondo, L. Salas, S. Lobos, R. Vicuna, Biochim. Biophys Acta **1490** (2000) 137 – 144. Link
908. P. Maijala, K. Kulju, M. Raudaskoski, Les Colloques **89** (1998) 325 – 340. Link
909. A. Conesa, C. van den Hondel, P. Punt, Appl. Environ. Microbiol. **66** (2000) 3016 – 3023. Link
910. V. Ducros, A. Brzozowski, K. Wilson, S. Brown, P. Ostergaard, P. Schneider, P. Yaver,
911. A. Mustranta, P. Spetz, R. Ekman, K. Luukko, J. Buchert, *J. Proc. EUCEPA* (1998) Symp. Chemistry in Papermaking, Florence, pp. 93 – 101.
912. M. Akhtar, R. Blanchette, G. Myers, T. Kirk in R. A. Young, M. Akhtar (eds.): *Environmentally Friendly Technologies for the Pulp and Paper Industry*, Wiley, New York, 1998, pp. 309 – 340.
913. R. Farrell, K. Hata, M. Wall, in T. Scheper (ed.): *Advances in Biochemical Engineering/Biotechnology*, Vol. 57, Springer-Verlag, Berlin, Heidelberg, 1997, pp. 197 – 212.
914. J. Pere, M. Siika-aho, L. Viikari, Tappi J. **83** (2000) 1 – 8. Link
915. J. Pere, S. Liukkonen, M. Siika-aho, J. Gullichsen, L. Viikari, *Proc. Tappi Pulping Conf*, Nashville, TN, Oct. 27 – 31, 1996, 693 – 696.
916. J. Richardson, K. Wong, T. Clark, J. Pulp Pap. Sci. **24** (1998) 125. Link

917. C. Jacobs, R. Venditti, T. Joyce, Tappi J. **81** (1998) 260 – 266. Link
918. C. Jacobs-Young, R. Gustafsson, J. Heitman, Paper and Timber, **82** (2000) 114 – 119. Link
919. H.-P. Call (2000) *TAPPI Pulping Conf.*, Nov. 5 – 9, Boston, USA.
920. H.-P. Call, *10th Int. Symp. on Wood and Pulping Chemistry*, Vol I, Yokohama, Japan, June 7 – 10, 1999, I-540.
921. A. Kantelinen, B. Hortling, J. Sundqvist, M. Linko, L. Viikari, Holzforschung **47** (1993) 318 – 324. Link
922. J. Yang, K.-E. Eriksson, Holzforschung **46** (1992) 481 – 488. Link
923. J. Buchert, G. Carlsson, L. Viikari, G. Ström, Holzforschung **50** (1996) 69 – 74. Link
924. T. Vuorinen, A. Teleman, P. Fagerström, J. Buchert, M. Tenkanen, *International Pulp Bleaching Conference*, Washington D.C., 1996, pp. 43 – 51.
925. L. Viikari, T. Oksanen, J. Buchert, M. Amann, A. Candussio, *Proceedings from the ISWPC*, Yokohama, I, (1999) 504 – 507.
926. A. Suurnäkki, T. Clark, R. Allison, L. Viikari, J. Buchert, Tappi J. **79** (1996) 111 – 117. Link
927. R. Patel, A. Grabski, T. Jeffries, Appl. Microb. Biotechnol. **39** (1993) 405 – 412. Link
928. J. Clarke, J. Rixon, A. Ciruela, H. Gilbert, Appl. Microbiol. Biotechnol. **48** (1997) 177 – 183. Link
929. J. Tolan, L. Thibault, *Tappi Pulping Conf.*, Boston, USA, 2000, pp. 318 – 324.
930. M. Paloheimo, A. Mäntylä, J. Vehmaanperä, S. Hakola, R. Lantto, T. Lahtinen, E. Parkkinen, R. Fagerström, P. Suominen in M. Claeyssens, W. Nerinckx, K. Piens (eds.), *Carbohydrases from Trichoderma reesei and other Micro-Organisms. Structures, Biochemistry, Genetics and Applications*, The Royal Society of Chemistry, Cambridge, UK, 1998, 255 – 264.
931. R. Bourbonnais, M. Paice, Appl. Microbiol. Biotechnol. **36** (1992) 823 – 827. Link
932. H.-P. Call, *Proceedings from the International Non-Chlorine Bleaching Conference*, Florida, USA, 1995, Paper 4 – 3.
933. M. Amann, *International Symposium on Wood and Pulping Chemistry*, Montreal, Quebec, Canada, 1997, CPPA. pp. F4-1 – F4-5.
934. J. Freudenreich, M. Amann, E. Fritz-Langhals, J. Stohrer, *Int. Pulp Bleaching Conf.*, Helsinki, Finland, 1998, 71 – 76.
935. M. Amann, A. Candussio, R. Müller, V. Frey, *TAPPI Pulping Conf.*, Nov. 5-9, Boston, USA, 2000.
936. R. Bourbonnais, D. Rochefort, M. G. Paice, S. Renaud, D. Leech, Tappi J. **83** (2000) no. 10. Link
937. M. Balakshin, E. Capanema, C.-L. Chen, J. Gratzl, A. Kirkman, H. Gracz, *10th Int. Symp. on Wood and Pulp Chem.*, Vol. 1, Yokohama, Japan, 1999, pp. 572 – 577.
938. J. Sealey, A. Ragauskar, *9th Int. Symp. Wood Pulp. Chem.*, Montreal, Canada, 1997, F11 – F14
939. K. Poppius-Levlin, W. Wang, T. Tamminen, B. Hortling, L. Viikari, M.-L. Niku-Paavola, J. Pulp Paper Sci. **25** (1999) 90 – 94. Link
940. F. Chakar, A. Ragauskas, *10th Int. Symp. Wood Pulp Chem.*, Vol. 1, Yokohama, Japan, 1999, pp. 566 – 570.
941. R. Pfaller, M. Amann, J. Freudenreich, *Proceedings from the 7th International Conference on Biotechnology in the Pulp and Paper Industry*, Vancouver, Canada, 1998, A99 – A102.
942. J. Buchert, A. Mustranta, H. Kontkanen, S. Karlsson, M. Tenkanen, B. Holmbom, *11th ISWPC International Symposium on Wood and Pulping Chemistry*, Nice, 11 – 14 June, 2001.
943. Y. Fujita, H. Awaji, H. Taneda, M. Matsukura, K. Hata, H. Shimoto, M. Sharyo, H. Sakaguchi, K. Gibson, Tappi J. **75** (1992) 117 – 112. Link
944. K. Fischer, K. Messner, Tappi J. **75** (1992) 130 – 134. Link
945. A. Mustranta, L. Fagernäs, L. Viikari, Tappi J. **78** (1995) 140 – 146. Link
946. A. Mustranta, P. Spetz, B. Holmbom, J. Buchert, J. Nordic Pulp Pap. Res. 2001, no. 2 Link
947. J. Buchert, A. Mustranta, P. Spetz, R. Ekman, K. Luukko, *Proc. Pre-symp. 10th ISWPC*, Seoul, 1999, pp. 115 – 119.
948. X. Zhang, Pulp Paper Can. **101** (2000) 59 – 62. Link

949. S. Karlsson, S. Holmbom, P. Spetz, A. Mustranta, J. Buchert, Applied Microbiol. Biotechnol. **55** (2001) 317 – 320. Link
950. J. Buchert, P. Spetz, A. Mustranta, R. Ekman, B. Holmbom, *Proc. 27th EUCEPA Conference. Grenoble*, France, 1999, pp. 191 – 194.
951. J. Pere, M. Siika-aho, J. Buchert, L. Viikari, Tappi J. **78** (1995) 71 – 78. Link
952. T. Oksanen, J. Pere, J. Buchert, L. Viikari, Cellulose **4** (1997) 329 – 339. Link
953. T. Oksanen, A. Suurnäkki, C. Schönberg, J. Buchert, *10th Int. Symp. Wood and Pulping Chemistry*, Yokohama, 1999, 7 – 10 June, Vol. 1, pp. 440 – 445.
954. A. Kantelinen, J. Sarkar, T. Oksanen, *Tappi Pulping Conf.*, Boston 2000, USA.
955. S. D. Mansfield, K. K. Y. Wong, E. de Jong, J. N. Saddler, Tappi J. **79** (1996) 125 – 172. Link
956. D. Y. Prasad, Appita **46** (1993a) 289 – 292. Link
957. D. Y. Prasad, J. A. Heitman, T. W. Joyce, Nordic Pulp Paper Res. J. **2** (1993) 284 – 286. Link
958. T. W. Jeffries, J. H. Klungness, M. S. Stykes, K. Rutledge-Cropsey, Tappi J. **77** (1994) 173 – 179. Link
959. T. Welt, R. Dinus, Prog. Pap. Recycl. **4** (1995) no. 2, 36 – 47. Link
960. P. Bajpaj, Adv. Appl. Microbiol. **45** (1997) 241 – 269. Link
961. A. Morbak, W. Zimmermann, Prog. Pap. Recycl. **7** (1998) no.2, 14 – 21. Link
962. A. Suurnäkki, H.-J. Putz, K. Renner, G. Galland, J. Buchert, *Proc. EUCEPA 1998 Symp. Chemistry in Papermaking*, Florence, 1998, pp. 215 – 224.
963. O. Knudsen, D. J. Young, J. L. Yang, *Proc. 7th Int. Conf. on Biotechnology in Pulp and Paper Industry*, Vancouver, Canada, 1998, A17 – A20.
964. H. K. Zollner, L. R. Schroeder, Tappi J. **81** (1998) 166 – 170. Link

Specific References

965. M. L. H. Flindt, Am. J. Ind. Med. **29** (1996) 99 – 110. Link
966. D. F. S. Flood, R. E. Blofeld, C. F. Bruce, J. I. Hewitt, C. P. Juniper, D. M. Roberts, Br. J. Ind. Med. **42** (1985) 43 – 50. Link
967. O. Kirk, T. V. Borchert, C. C. Fuglsang, Curr. Opin. Biotechnol. **13** (2002) 345 – 351. Link
968. H. Offredo-Delille, Fragrance J. **30** (2002) 49 – 53. Link
969. T. Masunaga, Cosmet. Sci. Technol. Ser. **25** (2002) 385 – 403. Link
970. L. M. Lods, C. Dres, C. Johnson, D. B. Scholz, G. J. Brooks, Int. J. Cosmet. Sci. **22** (2000) 85 – 94. Link
971. A. Charbonnelle, Parfuem. Kosmet. **80** (2000) 32 – 35. Link
972. A. Miller, Soaps, Deterg. Toiletries Rev. **7** (1977) 21 – 5. Link
973. J. L. Abegg, Parfums Cosmet. Savons Fr. **1** (1971) 84 – 6. Link
974. W. Umbach (ed.), *Kosmetik*, 2nd ed., Thieme, Stuttgart, 1995.
975. GB 109051, 1917 (J. Korselt).
976. L'Oréal, WO 9936041 A1, 1999 (G. Lang, J. Cotteret).
977. Wella, WO 0247633 A2, 2002 (D. Rozzell, G. Sauter, H.-J. Braun).
978. Lion, WO 0037030 A, 2000 (T. Onuki, M. Noguchi, J. Mitamura).
979. Henkel, WO 0021497 A1, 2000 (A. Saettler, A. Kleen, A. Weiss, D. Rose).
980. Kao, JP-Kokai 2 002 255 764 A2, 2002 (K. Koike).
981. Y. Tsujino, Y. Yokoo, K. Sakato, J. Soc. Cosmet. Chem. **42** (1991) 273 – 282. Link
982. Y. Tsujino, K. Kitayama, Y. Yokoo, K. Sakato, J. SCCJ **24** (1991) 220 – 223. Link
983. L'Oréal, WO 9917720 A1, 1999 (R. De La Mettrie, J. Cotteret, A. De Labbey, M. Maubru).
984. Wella, EP 795313 A2, 1997 (M. Kunz, D. Le Cruer).
985. Henkel, WO 0021497, 2000 (A. Saettler, A. Kleen, A. Weiss, D. Rose).
986. Goldwell, EP-A 548620, 1993 (K. Dornbusch, H. Lorenz, A. Olt, J. Tennigkeit).
987. M. G. Peter, U. Wollenberger, Front. Biosens. I (1997) 63 – 82. Link

988. Novo Nordisk, WO 9719998, 1997 (D. Aaslyng, N. H. Sorensen, K. Rorbaek).
989. L'Oréal, WO 9917720 A1, 1999 (R. De La Mettrie, J. Cotteret, A. De Labbey, M. Maubru).
990. T. Toyoshi, M. Yoshida, Y. Ohashi, N. Aoyama; J. Imai, M. Mitsuishi, K. Isowa, T. Mitsuzono, M. Aoki, Y. Tsujino, Oyo Yakuri, **55** (1998) 77 – 89. Link
991. Novo Nordisk, WO 9501425 A1, 1995 (C. M. Hjort, B. Diderichsen, D. Aaslyng, C. K. Hansen, J. T. Andersen).
992. University of Reading, UK, EP-A 272781 A1, 1989 (B. E. Brockway).
993. Henkel, WO 0121139, 2001 (A. Kleen, O. Brabaender, A. Saettler, H. Hoeffkes, R. Otto).
994. J. M. Gardner, P. E. Swanson, B. Torres-Lopez, Virg. J. Soc. Cosmet. Chem. **46** (1996) 11 – 28. Link
995. Mitsui Petrochemical Industries and Loumar Cosmetics, JP-Kokai 04005220, 1992 (S. Kise, Y. Ikuta, Y. Ueno).
996. Henkel, DE 2141764, 19730301, 1973 (W. Schreiber).
997. Novo Nordisk, WO 9958013, 1999 (N. H. Sorensen, J. P. McDevitt).
998. C. K. Kelling, R. G. Bartolo, K. D. Ertel, A. L. Smith, A. Laurence. D. D. Watson, K. Sarlo, J. Allergy Clin. Immunol. **101** (1998) 179 – 187. Link
999. Den-Mat Corp., US 4986981, 1991 (W. R. Glace, R. L. Ibsen).
1000. Investigations Scientifiques Pharmaceutiques S. A., FR 2132980, 1972.
1001. Lion Fats & Oils, JP-Kokai 50058243, 1975 (K. Shibuya, Y. Yamazaki).
1002. Aspro-Nicholas, DE 1965043, 1971 (S. A. Barker, L. F. Wiggins).
1003. Blendax, DE 1948468, 1971 (H. Harth, K. Rau, K. Scheer).
1004. G. Fridh, G. Koch, Swed. Dent. J. **23** (1999) 49 – 57. Link
1005. J. Tenovua, Oral Diseases **8** (2002) 23 – 29. Link
1006. Warner-Lambert Co., EP-A 258186, 1988 (A. R. Gallopo).
1007. Procter & Gamble, EP-A 425017, 1991 (R. S. Carpenter, A. M. Wolff, P. J. Lad).
1008. Lion, JP 2002193778 A2, 2002 (M. Ishikawa).
1009. E. Budtz-Jorgensen, J. Biol. Buccale **5** (1978) 239 – 44. Link
1010. Novo Nordisk, WO 9706775 A1, 19970227, 1997 (D. Aaslyng, R. Tsuchiya).
1011. J. N. de Wit, A. C. M. van Hooydonk, Neth. Milk Dairy J. **50** (1996) 227 – 244. Link
1012. V. Bussmann, Volker, Parfuem. Kosmet. **77** (1996) 538 – 539. Link
1013. H. Lemmer, Korrespondenz Abwasser **46** (1999) 1611 – 1615. Link
1014. R. Urlaub, Flüssiges Obst **67** (2000) 515 – 525. Link
1015. P. Blanpain-Avet, J. F. Maingonnat, Entropie **37** (2001) 2 – 18. Link
1016. M. A. Aaguello, S. Alvarez, F. A. Riera, R. Alvarez, J. Agric. Food Technol. **50** (2002) 1951 – 1958. Link
1017. E. P. Jacobs, P. Swart, C. J. Brouckaert, O. O. Hart, Water SA **19** (1993) 127 – 133. Link
1018. Henkel, WO 91/00333, 1991 (H. W. Wollenweber, R. Krack).
1019. T. Godfrey, S. West: *Industrial Enzymology*, MacMillan Press, London, 1996, p. 234.
1020. Henkel-Ecolab, DE 19503060, 1996 (R. Krack, H. Manners).
1021. Schwarzkopf, DE 1229980, 1967 (F. Keil, R. E. Diewas, K. R. Kloesel).
1022. DE 1124640, 1966 (T. Wajaroff).
1023. Novo Nordisk, WO 9958309, 1999 to (C. L. Sparre, W. R. Sponholz, O. Berker).
1024. Novozymes, WO 0110614, 2001 (C. L. Sparre, C. Felby, O. Berker).
1025. Novo Nordisk, WO 9713628, 1997, (R. Martinez, L. Sparre, M. Gomez, P. V. Neto).
1026. E. J. Hanssen, P. Jiang, H. H. Pedersen, J. F. Jorgensen, *Proceedings SPE International Symposium on Oilfield Chemistry*, Houston, Feb. 16 – 19, pp. 79 – 91, Publisher: *Society of Petroleum Engineers*, Richardson, Texas, 1999.
1027. Novo Nordisk, WO 94/25576, 1994 (S. Christgau, L. V. Kofod, L. N. Andersen, S. Kauppinen, H. P. Heldt-Hansen, H. Dalboege).
1028. Diversa Corporation, WO 98/24799, 1998, (E. J. Bylina, R. V. Swanson, E. J. Mathur, D. E. Lam).

1029. E. Gary, J. Sneddon, Microchem. J. **61** (1999) 53 – 57. Link
1030. C. O. Felby, P. O. Olesen, T. T. Hansen, *CELL-067 in 213th ACS Natl. Meeting*, San Francisco, April 13 – 17, 1997.
1031. A. Kharazipour, K. Bergmann, K. Nonninger, A. Huettermann, J. Adhes. Sci. Technol. **12** (1998) 1045 – 1053. Link

Specific References

1032. K. Drauz, H. Waldmann, eds., *Enzyme Catalysis in Organic Synthesis*, VCH, Weinheim, 1995.
1033. C.-H. Wong, G. W. Whitesides, *Enzymes in Synthetic Organic Chemistry*, Elsevier, Oxford, 1994
1034. W. D. Fessner (ed.): *Biocatalysis – From Discovery to Application*, Springer, Berlin, 1999.
1035. C.-H. Wong, R. L. Halcomb, Y. Ichikawa, T. Kajimoto, Angew. Chem. **107** (1995) 453; Link Angew. Chem., Int. Ed. Engl. **34** (1995) 521. Link
1036. C. Walsh, Nature **409** (2001) 226. Link
1037. K. M. Koeller, C.-H. Wong, Nature **409** (2001) 232. Link
1038. C. Khosla, P. B. Harbury, Nature **409** (2001) 247. Link
1039. F. H. Arnold, Nature **409** (2001) 253. Link
1040. A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, Nature **409** (2001) 258. Link
1041. A. M. Klibanov, Nature **409** (2001) 241. Link
1042. S. M. Roberts, *Biocatalysis for Fine Chemicals Synthesis*, Wiley-VCH, Weinheim, 1999.
1043. K. Faber, *Biotransformations in Organic Chemistry*, 4th edition, Springer, Berlin, 2000.
1044. W. Gerhartz, *Enzymes in Industry*, VCH, Weinheim, 1990.
1045. A. Zaks, D. R. Dodds, Curr. Opin. Drug Discov. Dev. **1** (1998) 290. Link
1046. K. Mori, Synlett (1995) 1097. Link
1047. J.-M. Fang, C.-H. Lin, C. W. Bradshaw, C.-H. Wong, J. Chem. Soc. Perkin Trans. **1** (1995) 967. Link
1048. J.-M. Fang, C.-H. Wong, Synlett (1994) 393. Link
1049. H. Waldmann, D. Sebastian, Chem. Rev. (1994) 911. Link
1050. T. Pathak, H. Waldmann, Curr. Opin. Chem. Biol. **2** (1998) 112. Link
1051. R. D. Schmid, R. Verger, Angew. Chem. **110** (1998) 1694; Angew. Chem., Int. Ed. Engl. 1998, **37**, 1608. Link
1052. D. G. Drueckhammer, W. J. Hennen, R. L. Pederson, C. F. Barbas III, C. M. Gautheron, T. Krach, C.-H. Wong, Synthesis (1991) 499. Link
1053. A. L. Margolin, Enzyme Microb. Technol. **15** (1993) 266. Link
1054. M. T. Reetz, Angew. Chem. 2001, **113**, 292; Link Angew. Chem., Int. Ed. Engl. **40** (2001) 284. Link
1055. H. Joo, Z. Lin, F. H. Arnold, Nature (1999) 399, 670. Link
1056. G. Carrea, S. Riva, Angew. Chem. **112** (2000) 2312; Link Angew. Chem., Int. Ed. Engl. **39** (2000) 2226.
1057. U. T. Bornscheuer, R. J. Kazlauskas, Wiley, Weinheim, 1999.
1058. A. M. P. Koshinen, A. M. Klibanov (eds.): *Enzymatic Reactions in Organic Media*, Chapman & Hall, Glasgow, 1991.
1059. M. Therisod, A. M. Klibanov, J. Am. Chem. Soc. **108** (1986) 5638. Link
1060. S. Riva, J. Chopineau, A. P. G. Kieboom, A. M. Klibanov, J. Am. Chem. Soc. **110** (1988) 584. Link
1061. S. Riva, R. Bovara, G. Ottolina, F. Secundo, G. Carrea, J. Org. Chem. **54** (1989) 3161. Link
1062. G. Ottolina, G. Carrea, S. Riva, Biocatalysis **5** (1991) 131. Link
1063. D. L. Delinck, A. L. Margolin, Tetrahedron Lett. **31** (1990) 3093. Link
1064. B. Danieli, G. Lesma, M. Luisetti, S. Riva, Tetrahedron **53** (1997) 5855. Link
1065. Y. Kita, Y. Takebe, K. Murata, T. Naka, S. Akai, J. Org. Chem. **65** (2000) 83. Link
1066. M. Schudok, G. Kretzschmar, Tetrahedron Lett. **38** (1997) 387. Link
1067. G. Kirchner, M. P. Scollar, A. M. Klibanov, J. Am. Chem. Soc. **107** (1985) 7072. Link

1068. A. Ghogare, G. S. Kumar, J. Chem. Soc, Chem. Commun. 1990, 134. Link
1069. R. V. Nair, M. M. Salunkhe, Synth. Commun. **30** (2000) 134. Link
1070. P.-J. Ulm, D. G. Drueckhammer, J. Am. Chem. Soc. **120** (1998) 5605. Link
1071. C. Dinkel, M. Moody, A. E.. Traynor-Kaplan, C. Schultz, Angew. Chem. **113** (2001) 3093; Link Angew. Chem., Int. Ed. Engl. **40** (2001) 3004. Link
1072. M. T. Reetz, K. Schimossek, Chimia **50** (1996) 668. Link
1073. G. M. Ramos-Tombo, H.-P. Shar, I. Fernandes, X. Busquets, O. Ghisalba, Tetrahedron Lett. **27** (1986) 5707. Link
1074. T. Sugahara, K. Ogasawara, Synlett (1996) 319. Link
1075. M. T. Reetz, Adv. Mater. **9** (1997) 943. Link
1076. M. T. Reetz, A. Zonta, J. Simpelkamp, Angew. Chem. **107** (1995) 373; Link Angew. Chem., Int. Ed. Engl. **34** (1995) 301.
1077. M. T. Reetz, A. Zonta, J. Simpelkamp, Biotechnol. Bioeng. **49** (1996) 527. Link
1078. M. T. Reetz, R. Wenkel, D. Avenir, Synthesis (2000) 781. Link
1079. V. M. Balcao, H. Akita, Biocatal. Biotransform. **13** (1996) 141; Link
1080. A. L. Paiva, F. X. Malcata, Enzyme Microb. Technol. **18** (1996) 392. Link
1081. A. L. Margolin, Trends Biotechnol. **14** (1996) 223. Link
1082. T. Zelinski, H. Waldmann, Angew. Chem. **109** (1997) 746; Link Angew. Chem., Int. Ed. Engl. **36** (1997) 722.
1083. N. St. Clair, Y.-F. Wang, A. L. Margolin, Angew. Chem. **112** (2000) 388; Link Angew. Chem., Int. Ed. Engl. **39** (2000) 380. Link
1084. A. Bommarius, M. Kottenham, H. Klenk, K. Drauz, in: *Microbial Reagents in Organic Synthesis*, S. Servi (ed.): Kluwer Academic Publishers, Dordrecht, 1992, p. 161.
1085. K. Drauz, M. Kottenham, K. Makryaleas, H. Klenk, M. Bernd, Angew. Chem. **103** (1991) 704; Link Angew. Chem., Int. Ed. Engl. **30** (1991) 712.
1086. M. Beller, M. Eckert, H. Geissler, B. Napierski, H.-P. Rebenstock, E. W. Holla, Chem. Eur. J. **4** (1998) 935. Link
1087. P. Hönicke-Schmidt, M. P. Schneider, J. Chem. Soc., Chem. Commun. (1990) 648. Link
1088. A. de Raadt, N. Kempier, K. Faber, H. Griengl, J. Chem. Soc., Perkin Trans. **1** (1992) 137. Link
1089. M. Kobayashi, N. Yanaka, T. Nagasawa, H. Yamada, Tetrahedron **46** (1990) 5587. Link
1090. R. Bauer, H.-J. Knackmuss, A. Stolz, Appl. Microbiol. Biotechnol. **49** (1998) 89. Link
1091. D. R. Dodds, J. B. Jones, J. Am. Chem. Soc. **110** (1988) 577. Link
1092. M. Wolberg, W. Hummel, C. Wandrey, M. Müller, Angew. Chem. **112** (2000) 4476; Link Angew. Chem., Int. Ed. Engl. **39** (2000) 4306 Link
1093. B. L. Hirschbein, G. M. Whitesides, J. Am. Chem. Soc. **104** (1982) 4458. Link
1094. C. W. Bradshaw, C.-H. Wong, W. Hummel, M.-R. Kula, Bioorg. Chem. **19** (1991) 29. Link
1095. M.-J. Kim, J. Y. Kim, J. Chem. Soc., Chem. Commun. (1991) 326. Link
1096. Z. Shaked, G. M. Whitesides, J. Am. Chem. Soc. **102** (1980) 7104. Link
1097. C. Wandrey, A. F. Bückmann, M.-R. Kula, Biotechnol. Bioeng. **23** (1981) 1789. Link
1098. C. W. Bradshaw, H. Fu, G.-J. Shen, C.-H. Wong, J. Org. Chem. **57** (1992) 1526. Link
1099. G.-J. Shen, Y.-F. Wang, C.-H. Wong, J. Chem. Soc., Chem. Commun. (1990) 677. Link
1100. Y. Naoshima, H. Hasegawa, T. Nishiyama, A. Nakamura, Bull. Chem. Soc. Jpn. **62** (1989) 608. Link
1101. K. Mori, H. Mori, Org. Synth. **68** (1989) 56. Link
1102. K. Mori, Bull. Soc. Chim. Belg. 1992, **101**, 393. Link
1103. M. Desrut, A. Kergomard, M. F. Renard, H. Veschambre, Tetrahedron **37** (1981) 3825. Link
1104. A. Kergomard, M. F. Renard, H. Veschambre, J. Org. Chem. **47** (1982) 792. Link
1105. P. Ferraboschi, S. Casati, E. Santaniello, Tetrahedron Asymm. **5** (1991) 19. Link
1106. N. S. Y. Ng, L.-C. Yuan, I. J. Jakovac, J. B. Jones, Tetrahedron **40** (1984) 1235. Link
1107. I. J. Jakovac, H. B. Goodbrand, K. P. Lok, J. B. Jones, J. Am. Chem. Soc. **104** (1982) 4659. Link

1108. J. Lehmann, A. Moritz, Liebigs Ann. Chem. (1991) 937. Link
1109. S. V. Ley, F. Sternfeld, Tetrahedron **45** (1989) 3463. Link
1110. M. E. De Luca, T. Hudlicky, Tetrahedron Lett. **31** (1990) 13. Link
1111. D. R. Boyd, N. D. Sharma, R. Boyle, B. T. McMurray, T. A. Evans, J. F. Malone, H. Dalton, J. Chima, G. N. Sheldrake, J. Chem. Soc., Chem. Commun. (1993) 49. Link
1112. T. Hudlicky, in: *Microbial Reagents in Organic Synthesis*, S. Servi ed., Kluwer Academic Publishers, Dordrecht, 1992, p. 123.
1113. T. Hudlicky, H. Luna, J. D. Price, F. Rulin, Tetrahedron Lett. **30** (1989) 4053. Link
1114. S. M. Roberts, P. W. H. Wan, J. Mol. Catal. B: Enzymat. **4** (1998) 111. Link
1115. S. Rissom, U. Schwarz-Linek, M. Vogel, V. I. Tishkov, U. Kragl, Tetrahedron Asymm. **8** (1997) 2523. Link
1116. A. Archelas, R. Furstoss, in W. D. Fessner (ed.): *Biocatalysis – From Discovery to Application*, pp. 159 – 191, Springer, Berlin, 1999.
1117. W.-D. Woggon, *Topics in Current Chemistry*, **184** (1996) 40.
1118. E. J. Allain, L. P. Hager, L. Deng, E. J. Jacobsen, J. Am. Chem. Soc. **115** (1993) 4415. Link
1119. W.-D. Fessner, Kontakte (Darmstadt) 1993 no. 1, 23. Link
1120. W.-D. Fessner, Curr. Opin. Chem. Biol. **2** (1998) 85. Link
1121. G. Guanti, L. Banfi, M. T. Zannetti, Tetrahedron Lett. **41** (2000) 3181. Link
1122. H.-L. Arth, W.-D. Fessner, Carbohydr. Res. **305** (1998) 313. Link
1123. F. Morís-Varas, X.-H. Qian, C.-H. Wong, J. Am. Chem. Soc. **118** (1996) 7647. Link
1124. C.-C. Lin, F. Morís-Varas, G. Weitz-Schmidt, C.-H. Wong, Bioorg. Med. Chem. **7** (1999) 425. Link
1125. W.-D. Fessner, G. Sinerius, A. Schneider, M. Dreyer, G. E. Schulz, J. Badia, J. Aguilar, Angew. Chem. **103** (1991) 596; Link Angew. Chem., Int. Ed. Engl. **30** (1991) 555.
1126. W.-D. Fessner, C. Walter, *Topics in Current Chemistry* **184** (1996) 97.
1127. R. Schoevaart, F. van Rantwijk, R. A. Sheldon, Biotechnol. Bioeng. **70** (2000) 349. Link
1128. R. Schoevaart, F. van Rantwijk, R. A. Sheldon, J. Org. Chem. **65** (2000) 6940. Link
1129. H. Oikawa, T. Kobayashi, K. Katayama, Y. Suzuki, A. Ichihara, J. Org. Chem. **63** (1998) 8748. Link
1130. K. M. Koeller, C.-H. Wong, Chem. Rev. **100** (2000) 4465. Link
1131. H. Herzner, T. Reipen, M. Schultz, H. Kunz, Chem. Rev. **100** (2000) 4495. Link
1132. O. Seitz, ChemBioChem **1** (2000) 214. Link
1133. P. M. St. Hilaire, M. Meldal, Angew. Chem. **112** (2000) 1210; Link Angew. Chem., Int. Ed. Engl. **39** (2000) 1162.
1134. B. G. Davis, J. Chem. Soc. Perkin Trans. **1** (1999) 3215. Link
1135. S. C. T. Svensson, J. Thiem, Carbohydr. Res. **200** (1990) 39. Link
1136. J. Thiem, B. Sauerbrei, Angew. Chem. **103** (1991) 1521; Link Angew. Chem., Int. Ed. Engl. **30** (1991) 1503. Link
1137. T. Usui, S. Kubota, H. Ohi, Y. Ishido, Glycoconjug. J. **8** (1991) 272. Link
1138. G. F. Herrmann, Y. Ichikawa, C. Wandrey, F. C. A. Gaeta, J. C. Paulson, C.-H. Wong Tetrahedron Lett. **34** (1993) 3091. Link
1139. C. Augé, R. Fernandez-Fernandez, C. Gautheron, Carbohydr. Res. **200** (1990) 257. Link
1140. Y. Ichikawa, R. Wang, C.-H. Wong, Methods Enzymol. **230** (1994) 300. Link
1141. T. Bülter, L. Elling, Glycoconjug. J. **16** (1999) 147. Link
1142. K. Witte, P. Sears, R. Martin, C.-H. Wong, J. Am. Chem. Soc. **119** (1997) 2114. Link
1143. K. Takegawa, M. Tabuchi, S. Yamaguchi, A. Kondo, I. Kato, S. Iwahara, J. Biol. Chem. **270** (1995) 3094. Link
1144. M. Schelhaas, S. Glomsda, M. Hänsler, W. Vogt, H.-D. Jakubke, H. Waldmann, Angew. Chem. **108** (1996) 82; Link Angew. Chem. Int. Ed. Engl. **35** (1996) 106. Link
1145. H. Waldmann, E. Nägele, Angew. Chem. **107** (1995) 2425; Link Angew. Chem. Int. Ed. Engl. **34** (1995) 2259. Link

1146. T. Pohl, H. Waldmann, Angew. Chem. **108** (1996) 1829; Link Angew. Chem. Int. Ed. Engl. **35** (1996) 1720.
1147. T. Pohl, H. Waldmann, J. Am. Chem. Soc. 1997, **119**, 6702. Link
1148. F. Eisele, H. Waldmann, Angew. Chem. **113** (2001) 382; Link Angew. Chem., Int. Ed. Engl. **40** (2001) 369. Link
1149. R. Machauer, H. Waldmann, Angew. Chem. **112** (2000) 1503; Link Angew. Chem. Int. Ed. Engl. **39** (2000) 1449. Link
1150. T. Kappes, H. Waldmann, Carbohydr. Res. **305** (1998) 341. Link
1151. P. G. Schultz, R. A. Lerner, Science **269** (1995) 1835. Link
1152. J. Hasserodt, Synlett (1999) 2007. Link
1153. D. Hilvert, Annu. Rev. Biochem. **69** (2000) 751 Link
1154. S. C. Sinha, C. F. Barbas III, R. A. Lerner, Proc. Natl. Acad. Sci. USA **95** (1998) 14603. Link
1155. G. Z. Zhong, D. Shabat, B. List, J. Anderson, S. C. Sinha, R. A. Lerner, C. F. Barbas III, Angew. Chem. **110** (1998) 2609; Link Angew. Chem., Int. Ed. Engl. **37** (1998) 2481.
1156. K. D. Janda, C. G. Shevlin; R. A. Lerner, Science, **259** (1993) 490. Link
1157. K. D. Janda, C. G. Shevlin; R. A. Lerner, J. Am. Chem. Soc. **117** (1995) 2659. Link
1158. K. Gruber, B. Zhou, K. N. Houk, R. A. Lerner, C. G. Shevlin, I. A. Wilson, Biochemistry, **38** (1999) 7062 Link
1159. V. E. Gouverneur, K. N. Houk, B. de Pascual-Teresa, B. Beno, K. D. Janda, R. A. Lerner, Science, **262** (1993) 204. Link
1160. G. Zhong, T. Hoffmann, R. A. Lerner, S. Danishefsky, C. F. Barbas III, J. Am. Chem. Soc. **119** (1997) 8131. Link
1161. B. Seelig, S. Keiper, F. Stuhlmann, A. Jäschke, Angew. Chem. **112** (2000) 4764; Link Angew. Chem. Int. Ed. Engl. **39** (2000) 4576.
1162. C. Frauendorf, A. Jäschke, Angew. Chem. **110**, (1998) 1449; Link Angew. Chem. Int. Ed. Engl. **37** (1998) 1378.
1163. G. F. Joyce, Proc. Natl. Acad. Sci. USA **95** (1998) 5845. Link
1164. H. Gröger, J. Wilken, Angew. Chem. **113** (2001) 545; Link Angew. Chem. Int. Ed. Engl. **40** (2001) 529.
1165. B. List, P. Pojarliev, C. Castello, Org. Lett. **3** (2001) 573. Link
1166. W. Notz, B. List, J. Am. Chem. Soc. **122** (2000) 7386. Link
1167. T. Bui, C. F. Barbas III, Tetrahedron Lett. **41** (2000) 6951. Link
1168. B. List, R. A. Lerner, C. F. Barbas III, J. Am. Chem. Soc. **122** (2000) 2395. Link
1169. J. W. Trauger, R. M. Kohli, H. D. Mootz, M. A. Marahiel, C. T. Walsh, Nature **407**, (2000) 215. Link
1170. V. Cerovský, J. Kockskämper, H. G. Glitsch, F. Bordusa, ChemBioChem **2** (2000) 126. %26SNM%3DKocksk%E4mper%26SNM%3DGlitsch%26SNM%3DBordusa%26JTL%3DChemBioChem%26PYR%3D2000%26VID%3D2%26PPF%3D126">Link
1171. *Chemical Economics Handbook*, SRI International, 1991.
1172. T. Nagasawa, H. Yamada, Tibtech **7** (1989) 153. Link
1173. Y. Asano, T. Yasuda, Y. Tani, H. Yamada, Agric. Biol. Chem. **46** (1982) 1183. Link
1174. I. Watanabe, Y. Satoh, K. Enomoto, S. Seki, K. Sakashita, Agric. Biol. Chem. **51** (1987) 3201. Link
1175. M. Kobayashi, T. Nagasawa, H. Yamada, Tibtech **10** (1992) 402. Link
1176. C. Chassin, Chimica Oggi – Chemistry Today **17** (1999 nos. 5/6, 48. Link
1177. An overview about this reaction is given in: A. S. Bommarius, M. Schwarm, K. Drauz, Chimia **55** (2001) 50. Link
1178. Ref. [1032], pp. 963 – 975.
1179. J. A. Jongejau, J. B. A. van Tol, J. A. Duine, Chem. Today (1994) 15. Link
1180. J. P. Rasor, *Speciality Chemicals* 1995, S5.
1181. T. Tosa, T. Mori, N. Fuse, I. Chibata, Agric. Biol. Chem. 1969, 1047. Link

1182. I. Chibata, T. Tosa, T. Sato, T. Mori, Methods Enzymol. (1976) 746. Link
1183. C. Wandrey, E. Flaschel, Adv. Biochem. Eng. (1979) 147. Link
1184. Degussa/GBF, C. Wandrey, R. Wichmann, W. Leuchtenberger, M. R. Kula, US 4.304.858 (1981).
1185. W. Leuchtenberger, M. Karrenbauer, U. Plöcker, Enzyme Eng. 7 Ann. N. Y. Acad. Sci. (1984) 434. Link
1186. A. S. Bommarius, K. Drauz, H. Klenk, C. Wandrey, Enzyme Eng. 11 Ann. N. Y. Acad Sci. (1992) 126. Link
1187. A. S. Bommarius in H.-J. Rehm, G. Reed, A. Pühler, P. Stadler (eds.): *Biotechnology*, Vol. 3, VCH Verlagsgesellschaft, Weinheim 1993, pp. 427.
1188. Ref. [1032], pp. 379.
1189. O. M. Birch, J. M. Brass, A. Kiener, K. Robins, D. Schmidhalter, N. M. Shaw, T. Zimmermann, *Chemistry Today* (1995) 9.
1190. C. Chassin, *Proc. Chiral USA* (1993) 137.
1191. E. Cerbeland, D. Petre, US 5.034.329, 1991.
1192. E. De Vroom, Org. Proc. Res. Developm. (1998) **2**, 128. Link
1193. E. J. A. X. Van de Sandt, E. de Vroom, Chimica Oggi – Chemistry Today **18** (2000) no. 5, 72. Link
1194. K. Matsumoto in A. Tanaka, T. Tosa, T. Kobayashi (ed.): *Industrial Application of Immobilized Biocatalysts*, Marcel Dekker Inc., New York 1993, pp. 67.
1195. A. Parmar, H. Kumar, S. S. Marwaha, J. F. Kennedy, Biotechnol. Advances **18** (2000) 289. Link
1196. A. Chadha, P. S. R. Babu, T. Panda, Int. J. Pharm. Adv. **1** (1995) no. 2, 216. Link
1197. Y. Chisti, M. Moo-Young in V. Moses, R. E. Cape (ed.): *Biotechnology: The Science and the Business*, Harwood Academic Publishers, New York 1991, pp. 167.
1198. I. S. Novella, C. Fargues, G. Grevillot, Biotechnol. Bioeng. **44** (1994) 379. Link
1199. E. Katchalski-Katzir, D. M. Kraemer, J. Mol. Catal. B: Enzym. **10** (2000) 157. Link
1200. A. Liese, K. Seelbach, C. Wandrey, *Industrial Biotransformations*, Wiley-VCH, Weinheim, 2000, pp. 281.
1201. A. Parmar, H. Kumar, S. S. Marwaha, J. F. Kennedy, Crit. Rev. Biotechnol. **18** (1998) no. 1, 1. Link
1202. Ref. [1032], pp. 386 and 458.
1203. F. Balkenhohl, K. Ditrich, B. Hauer, W. Ladner, J. Prakt. Chem. **339** (1997) 381. Link
1204. U. Vagt, Chimica Oggi – Chemistry Today **16** (1998) nos. 7/8, 38. Link
1205. Ref. [1032], pp. 409.
1206. A. Müller, E. Busker, J. Engel, B. Kutscher, M. Bernd, A. V. Schally, Int. J. Pept. Protein Res. (1994) 264. Link
1207. S. Bajusz, V. J. Csernus, T. Janaky, L. Bokser, M. Fekete, A. V. Schally, Int. J. Pept. Protein Res. (1988) 425. Link
1208. T. Reissmann, J. Engel, B. Kutscher, M. Bernd, P. Hilgard, M. Peukert, I. Szelenyi, S. Reichert, D. Gonzales-Barcena, E. Nieschiag, A. M. Comaru-Schally, A. V. Schally, Drugs Fut. (1994) 228. Link
1209. K. Drauz, M. Kottenhahn, M. Makryaleas, H. Klenk, M. Bernd, Angew. Chem. **103** (1991) 704; Link Angew. Chem. Int. Ed. Engl. **30** (1991) 712.
1210. K. Drauz, M. Kottenhahn, H. Klenk, J. Prakt. Chem. **334** (1992) 214. Link
1211. A. S. Bommarius, M. Kottenhahn, H. Klenk, K. Drauz, in S. Servi (ed.): *Microbial Reagents in Organic Synthesis*, Kluwer Academic, Dordrecht 1992, pp. 161.
1212. T. Fukumura, Agric. Biol. Chem. **40** (1976) 1687. Link
1213. T. Fukumura, Agric. Biol. Chem. **40** (1976) 1695. Link
1214. T. Fukumura, Agric. Biol. Chem. **41** (1977) 1327. Link
1215. J. Crosby, Tetrahedron 1991, 4789. Link
1216. S. J. C. Taylor, R. McCague, R. Wisdom, C. Lee, K. Dickson, G. Ruecroft, F. O'Brien, J. Littlechild, J. Bevan, S. M. Roberts, C. T. Evans, Tetrahedron: Asymmetry **4** (1993) 1117. Link
1217. S. J. C. Taylor, R. C. Brown, P. A. Keene, I. N. Taylor, Bioorg. Med. Chem. **7** (1999) 2163. Link

1218. A. Sakimae, A. Hosoi, E. Kobayashi, N. Ohsuga, R. Numazawa, I. Watanabe, H. Ohnishi, Biosci. Biotech. Biochem. **56** (1992) 1252. Link
1219. A. Sakimae, R. Numazawa, H. Ohnishi, Biosci. Biotech. Biochem. **56** (1992) 1341. Link
1220. Q.-M. Gu, C.-S. Chen, C. J. Sih, Tetrahedron Lett. (1986) 1763. Link
1221. P. Cesti, P. Picardi, EP 195.717, 1986.
1222. G. Fülling, M. Schlingmann, R. Keller, DE 3.919.029, 1990.
1223. J. G. H. M. Mutsaers, H. J. Kooremann, Rec. Trav. Chim. Pays-Bas (1991) 185. Link
1224. J. W. H. Smeets, A. P. G. Kienboom in S. Servi (ed.): *Microbial Reagents in Organic Synthesis*, Kluwer Academic, Dordrecht 1992, pp. 273.
1225. L. A. Hulshof, J. H. Roskam (Stamicarbon B. V.), EP 343.714, 1989.
1226. J. W. Young, R. L. Brataler, *Proc. Chiral 90 Symp. Spring Innovations UK 1990*, pp. 23.
1227. M. Villa, C. Pozzoli, *Proc. Chiral 95 USA Symp. Spring Innovations UK 1995*, pp. 13.
1228. K. Oyama in: *Chirality in Industry*, A. N. Collins, G. N. Sheldrake, J. Crosby (eds.), J. Wiley, New York 1992, pp. 237.
1229. Ref. [1032], pp. 178 – 261.
1230. R. Chênevert, R. Gagnon, M. Desiardins, M. Dickman, P. Bureau, G. Fortier, R. Martin, M. Létourheau, S. Thiboutot, R. Bel-Rhlid in S. Servi (ed.): *Microbial Reagents in Organic Synthesis*, Kluwer Academic, Dordrecht 1992, pp. 135.
1231. S. Shimizu, H. Yamada in E. J. Vandamme (ed.): *Biotechnology of Vitamins, Pigments and Growth Factory*, Elsevier Applied Science, London 1989, p. 199.
1232. H. Yamada, S. Shimizu in: *Enzyme Engineering 11*, D. S. Clark, D. Estell, J. Dordick (eds.), New York Academy of Science, New York, 1992, pp. 374.
1233. H. Yamada, *Chimia* (1993) 5.
1234. M. Ress-Löschke, T. Friedrich, B. Hauer, R. Mattes, D. Engels (BASF), DE 19848129, 1998.
1235. G. Matcham, A. Bowen, S. Lee, P. Pienkos, A. Zeitlin, *Proceedings of Chiral'94 USA Symposium*, Spring Innovations Ltd., 1994, p. 55.
1236. H. Yamada, H. Kumagai, Pure Appl. Chem. **50** (1978) nos. 9 – 10, 1117. Link
1237. H. Enei, H. Yamada, Prog. Ind. Microbiol. **24** (1986) 280. Link
1238. H. G. Kulla, Chimia, (1991) 81. Link
1239. A. S. Bommarius, K. Drauz, W. Hummel, M.-R. Kula, C. Wandrey, Biocatalysis **10** (1994) 37. Link
1240. K. Drauz, W. Jahn, M. Schwarm, Chem. Eur. J. (1995) 538. Link
1241. A. S. Bommarius, M. Schwarm, K. Stingl, M. Kottenhahn, K. Huthmacher, K. Drauz, Tetrahedron: Asymmetry **6** (1995) 2851. Link
1242. P. P. Taylor, D. P. Pantaleone, R. F. Senkpeil, I. G. Fotheringham, Trends Biotechnol. **16** (1998) 412. Link
1243. D. J. Ager, I. G. Fotheringham, T. Li, D. P. Pantaleone, R. F. Senkpeil, Enantiomer **5** (2000) 235. Link
1244. For reviews on the asymmetric synthesis of cyanohydrins and applications thereof, see: F. Effenberger, Angew. Chem. **106** (1994) 1609. Link
1245. Angew. Chem. Int. Ed. Engl. 1994, **33**, 1555; R. J. H. Gregory, Chem. Rev. **99** (1999) 3649. Link
1246. M. Schmidt, H. Griengl, Top. Curr. Chem. **200** (1999) 193. Link
1247. M. Bauer, H. Griengl, W. Steiner, Enzyme Microb. Technol. **24** (1999) 514, and references therein. Link
1248. R. L. Hill, J. W. Teipel in P. Boyer (ed.): *The Enzymes*, Vol. 5, Academic Press, New York 1971, pp. 539.
1249. U. T. Bornscheuer in H.-J. Rehm, G. Reed, A. Pühler, P. Stadler (eds.): *Biotechnology*, Vol. 8b, Wiley-VCH, Weinheim 2000, pp. 285.
1250. S. B. Mirviss, S. K. Dahod, M. W. Empie, Ind. Eng. Chem. Res. **29** (1990) 651. Link
1251. L. Rohlin, M.-K. Oh, J. C. Liao, Curr. Opinion Mircobiol. **4** (2001) 330. Link

1252. G. Chotani, T. Dodge, A. Hsu, M. Kumar, R. LaDuca, D. Trimbur, W. Weyler, K. Sanford, Biochimica et Biophysica Acta **434** (2000) 1543. Link

Specific References

1253. D. Blohm, C. Bollschweiler, H. Hillen, Angew. Chem. **100** (1988) 213 – 231. Link
1254. D. M. Goldberg, Clin. Chim. Acta **206** (1992) 45 – 76. Link
1255. A. J. Barrett, G. Salvesen (eds.), *Proteinase Inhibitors*, Elsevier Science Publishers, Amsterdam, 1986.
1256. M. Konrad, Trends Biotechnol. **7** (1989) 175 – 179. Link
1257. M. Hansson, P.-A. Nygren, S. Ståhl, Biotechnol. Appl. Biochem. **32** (2000) 95 – 107. Link
1258. G. Ashton, Nature Biotechnology **19** (2001) 307 – 311. Link
1259. J. M. Reichert, Trends Biotechnol. **18** (2000) 364 – 369. Link
1260. P. Bailon, W. Berthold, Pharmaceut. Sci. Technol. Today **1** (1998) 352 – 356. Link
1261. A. P. Chapman, Adv. Drug Delivery Revs. **54** (2002) 531 – 545. Link
1262. W. R. Gombotz, D. K. Pettit, Bioconjugate Chem. **6** (1995) 332 – 351. Link
1263. D. K. Pettit, W. R. Gombotz, Trends Biotechnol. **16** (1998) 343 – 349. Link
1264. M. Singh, A. J. Ferdous, M. Branham, G. V. Betageri, Drug Delivery Rev. **3** (1996) 289 – 304. Link
1265. K. D. Bagshawe, Adv. Pharmacol. **24** (1993) 99 – 121. Link
1266. R. F. Sherwood, Adv. Drug Delivery Rev. **22** (1996) 269 – 288. Link
1267. D. Derossi, G. Chassaing, A. Prochiantz, Trends Cell Biol. **8** (1998) 84 – 87. Link
1268. M. Becker-Hapak, S. S. McAllister, S. F. Dowdy, Methods **24** (2001) 247 – 256. Link
1269. G. Elliott, P. O'Hare, Cell **88** (1997) 223 – 233. Link
1270. M. C. Morris, J. Depollier, J. Mery, F. Heitz, G. Divita, Nature Biotechnol. **19** (2001) 1173 – 1176. Link
1271. M. C. Rattazzi, K. Dobrenis in: R. J. Desnick (ed.), *Therapy of Genetic Diseases*, Churchill-Livingstone, New York, 1991, pp. 131 – 152.
1272. G. Walsh, Nature Biotechnol. **18** (2000) 831 – 833. Link
1273. G.-B. Kresse, in: A. E. Kulozik, M. W. Hentze, C. Hagemeier, C. R. Bartram (eds.), *Molekulare Medizin*, W. de Gruyter Verlag, Berlin, 2000, pp. 421 – 434.
1274. S. Geisse, H. Gram, B. Kleuser, H. P. Kocher, Prot. Expr. Pur. **8** (1996) 271 – 282. Link
1275. S. M. Kingsman, A. J. Kingsman, J. Mellor, Trends Biotechnol. **5** (1987) 53 – 57. Link
1276. G. Hannig, S. Makrides, Trends Biotechnol. **16** (1998) 54 – 60. Link
1277. P. Balbás, Mol. Biotechnol. **19** (2001) 251 – 267. Link
1278. R. Rudolph, H. Lilie, FASEB J. **10** (1996) 49 – 56. Link
1279. L. Miele, Trends Biotechnol. **15** (1997) 45 – 50. Link
1280. A. R. Kusnadi, Z. L. Nikolov, J. A. Howard, Biotechnol. Bioeng. **56** (1997) 473 – 484. Link
1281. K. E. Van Cott, W. H. Velander, Exp. Op. Invest. Drugs **7** (1998) 1683 – 1690. Link
1282. R. G. Werner, Exp. Op. Invest. Drugs **8** (1999) 731 – 736. Link
1283. A. Lauwers, S. Scharpé (eds.): *Pharmaceutical Enzymes*, Marcel Dekker, New York, 1997.
1284. P. Buckel, ed. *Recombinant Protein Drugs*, Birkhäuser, Basel, 2001.
1285. F. M. Steinberg, J. Raso, J. Pharm. Pharmaceut. Sci. **1** (1998) 48 – 59. Link
1286. K. Dembowsky, P. Stadler, (eds.) *Novel Therapeutic Proteins*, Wiley-VCH, Weinheim, 2001.
1287. G. Walsh, D. R. Headon, *Protein Biotechnology*, Wiley, Chichester, 1994.
1288. A. Bayol, J. Capdevielle, P. Malazzi, A. Buzy, M. C. Bonnet, N. Colloc'h, J.-P. Mornon, D. Loyaux, P. Ferrara, Biotechnol. Appl. Biochem. **36** (2002) 21 – 31. Link
1289. J. M. Brant, Clin. J. Onc. Nursing **6** (2002) 12 – 16. Link
1290. F. M. Veronese, P. Caliceti, O. Schiavoni, M. Sergi, Adv. Drug Delivery Reviews **54** (2002) 587 – 606. Link

1291. S. Krishnagopalan, R. P. Dellinger, BioDrugs **15** (2001) 645 – 654. Link
1292. I. Horkheimer, D. P. Schuster, Adv. Sepsis **2** (2002) 2 – 7. Link
1293. D. M. Stafforini, T. M. McIntyre, G. A. Zimmerman, S. M. Prescott, J. Biol. Chem. **272** (1997) 17895 – 17898. Link
1294. D. Yarosh, J. Klein, A. O'Connor, J. Hawk, E. Rafal, P. Wolf, Lancet **357** (2001) 926 – 929. Link
1295. S. M. Rybak, D. L. Newton, Exp. Cell Res. **253** (1999) 325 – 335. Link
1296. E. D. Kakkis, Exp. Op. Invest. Drugs **11** (2002) 675 – 685. Link
1297. E. D. Kakkis, J. Muenzer, G. E. Tiller, L. Waber, J. Belmont, M. Passage, B. Izykowski, J. Phillips, R. Doroshow, I. Walot, R. Hoft, E. F. Neufeld, New Engl. J. Med. **344** (2001) 182 – 188. Link
1298. C. M. Eng, N. Guffon, W. R. Wilcox, D. P. Germain, P. Lee, S. Waldek, L. Caplan, G. E. Linthorst, R. J. Desnick, New Engl. J. Med. **345** (2001) 9 – 16. Link
1299. D. P. Germain, Exp. Op. Invest. Drugs **11** (2002) 1467 – 1476. Link
1300. N. Winterbottom, J. M. Kuo, K. Nguyen, C. J. Reich, K. J. Trent, J. F. Rondinone, N. S. Shargill, J. Appl. Res. Clin. Exp. Therapeutics **2** (2002) no. 1 Link
1301. D. A. Triplett, Clin. Chem. **46** (2000) 1260 – 1269. Link
1302. D. Collen, H. R. Lijnen, Fibrinol. & Proteol. **14** (2000) 66 – 72. Link
1303. U. Martin, R. Bader, E. Böhm, U. Kohnert, E. von Möllendorf, S. Fischer, G. Sponer, Cardiovasc. Drug Revs. **11** (1993) 299 – 311. Link
1304. C. Bode, K. Peter, T. K. Nordt, J. Ruef, R. W. Smalling, W. Kübler, Fibrinol. Proteol. **11** (1997) Suppl. 2, 29 – 33. Link
1305. J. R. Hampton, Fibrinol. Proteol. **11** (1997) Suppl. 2, 35 – 38. Link
1306. F. Vermeer, J. Pohl, K. Oldroyd, S. Giannelli, B. Charbonnier, J. Am. Coll. Cardiol. **37** (2001) Suppl. A, 322. Link
1307. P. Sinnaeve, F. Van de Werf, Thrombos. Res. **103** (2001) S71 – S79. Link
1308. K. Stocker, Schweiz. Med. Wochenschr. **129** (1999) 205 – 216. Link
1309. J. Leipner, F. Iten, R. Saller, BioDrugs **15** (2001) 779 – 789. Link

Specific References

1310. H. U. Bergmeyer (ed.): *Methods of Enzymatic Analysis*, 3rd ed., **vol. 1**, Verlag Chemie, Weinheim 1983.
1311. H. U. Bergmeyer, in [1310] pp. 163 – 181.
1312. J. Siedel, R. Deeg, J. Ziegenhorn in [1310] pp. 182 – 197.
1313. M. J. O'Sullivan, J. W. Bridges, V. Marks, Ann. Clin. Biochem. **16** (1979) 221 – 239. Link
1314. H. U. Bergmeyer, in [1310] pp. 233 – 260.
1315. M. Oellerich, in [1310] pp. 233 – 260.
1316. E. Ishikawa, M. Iamagawa, S. Hasida, S. Yoshitake, Y. Hamagushi, T. Ueno, J. Immunoassay **4** (1983) 209 – 327.Link
1317. P. D. Weston, J. A. Devries, R. Wrigglesworth, Biochim. Biophys. Acta **612** (1980) pp. 40 – 49. Link
1318. L. Hurn, S. M. Chantler, in S. P. Colowick, N .O. Kaplan, (eds.): *Methods in Enzymology*, **vol. 70**, Academic Press, New York 1980, pp. 104 – 142.
1319. J. Stanley, A. Johannsson, C. H. Self, J. Immunol. Methods **83** (1985) 89 – 95. Link
1320. K. Wulff, in [1310] pp. 340 – 368.
1321. Weeks, J. Beheshti, F. McCapra, A. K. Campbell, J. S. Woodhead, Clin. Chem. (Winston Salem N.C.) **29** (1983) pp. 1474 – 1479. Link
1322. Weeks. A. K. Campbell, J. S. Woodhead, Clin. Chem. (Winston Salem N.C.) **29** (1983) 1480 – 1483. Link
1323. J. Wannlund, M. DeLuca, in M. A. DeLuca, W. D. McElroy, (eds.): *Bioluminescence and Chemiluminescence*, Academic Press, New York 1981, pp. 693 – 696.
1324. H. Fricke, C. J. Strasburger, W. G. Wood, J. Clin. Chem. Clin. Biochem. **20** (1982) 91 – 94. Link

1325. Th. P. Whitehead, G. H. G. Thorpe, T. J. N. Carter, C. Croucutt, J. L. Kricka, Nature (London) **305** (1983) 158 – 159. Link

1326. G. H. Thorpe, L. J. Kricka, E. Gillespie, S. Moseley, R. Amess, N. Baggett, Th. P. Whitehead, Anal. Biochem. **145** (1985) 96 – 100. Link

1327. G. H. G. Thorpe, L. J. Kricka, S. Moseley, Th. P. Whitehead, Clin. Chem. (N.Y.) **31** (1985) 1335 – 1341. Link

1328. Enzyme Technology Company, EP-A 0 103 784, 1983 (R. C. Stout).

Specific References

1329. H. U. Bergmeyer, J. Bergmeyer, M. Graßl (eds.): *Methods of Enzymatic Analysis*, 3rd ed., vol**s. I – XII**, VCH Verlagsgesellschaft, Weinheim 1983 – 1986.

1330. H. U. Bergmeyer (ed.): *Methods of Enzymatic Analysis*, 2nd. ed., vol**s. 1 – 4**, Verlag Chemie, Weinheim, and Academic Press, New York 1974.

1331. Boehringer Mannheim GmbH: *Methods of Biochemical Analysis and Food Analysis*, Mannheim 1986.

1332. Boehringer Mannheim GmbH: Die enzymatische Lebensmittelanalytik (series). Link

1333. G. Henniger: "Enzymatische Lebensmittelanalytik – Enzymatic Food Analysis," Z. Lebensm.-Technol. Verfahrenstechn. **30** (1979), 137 – 144 and 182 – 185. Link

Specific References

1334. D. M. Glover: *Gene Cloning, The Mechanisms of DNA Manipulation*, Chapman and Hall, London 1984.

1335. D. M. Glover: *DNA Cloning, A Practical Approach*, vol. I, IRL Press Limited, Oxford 1985.

1336. D. M. Glover: *DNA Cloning, A Practical Approach*, vol. II, IRL Press Limited, Oxford 1985.

1337. L. Grossmann, K. Moldave: "Nucleic Acids," in S. P. Colowick, N. O. Kaplan (eds.): *Methods of Enzymology*, vol. 65, Part I, Academic Press, New York 1980.

1338. P. B. Hackett, J. A. Fuchs, J. W. Messing: *An Introduction to Recombinant DNA Techniques*, The Benjamin Cummings Publishing Company, Inc., Menlo Park, CA, 1984.

1339. B. D. Hames, S. J. Higgins: *Transcription and Translation. A Practical Approach.* IRL Press Limited, Oxford 1984.

1340. I. J. Higgins, D. J. Best, J. Jones: *Biotechnology, Principles and Applications*, Blackwell Scientific Publishers, Oxford 1985.

1341. A. D. B. Malcolm: "The use of restriction enzymes in genetic engineering," in R. Williamson (ed.): *Genetic Engineering*, vol. 2, Academic Press, New York 1981, p. 129.

1342. T. Maniatis, E. F. Fritsch, J. Sambrook: *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1982.

1343. J. H. Miller: *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1972.

1344. S. J. O'Brien: *Genetic Maps 1984*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1984.

1345. R. W. Old, S. B. Primrose in N. G. Carr, J. L. Ingraham, S. C. Rittenberg (eds.): *Principles of Gene Manipulation* Studies in Microbiology, vol. 2., Blackwell Scientific Publishers, Oxford 1985.

1346. R. H. Pritchard, I. B. Holland: *Basic Cloning Techniques*, Blackwell Scientific Publishers, Oxford 1985.

1347. A. Pühler, K. N. Timmis: *Advanced Molecular Genetics*, Springer Verlag, Berlin 1984.

1348. R. L. Rodriguez, R. C. Tait: *Recombinant DNA Techniques.* An Introduction, Addison-Wesley, Publishing Company, London 1983.

1349. S. M. Weissmann: *Methods of DNA and RNA Sequencing*, Praeger Publishers, New York 1983.

1350. E.-L. Winnacker: *Gene und Klone. Eine Einführung in die Gentechnologie*, VCH Verlagsgesellschaft, Weinheim 1985.

1351. R. Wu: “Recombinant DNA,” in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. 68, Academic Press, New York 1980.
1352. R. Wu, L. Grossmann, K. Moldave: “Recombinant DNA,” in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. 100, Part B, Academic Press, New York 1983.
1353. R. Wu, L. Grossmann, K. Moldave: “Recombinant DNA,” in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. 101, Part C, Academic Press, New York 1983.
1354. M. Zabeau, R. J. Roberts: “The role of restriction endonucleases in molecular genetics,” in J. H. Taylor (ed.): *Molecular Genetics.* Academic Press, New York 1979, p. 1.
1355. H. Takahashi, H. Kojima, H. Saito, Biochem. J. **231** (1985) 229. Link
1356. G. Theriault, P. H. Roy, K. A. Howard, J. S. Benner, J. E. Brooks, A. F. Waters, T. R. Gingeras, Nucleic Acids Res. **13** (1985) 8441. Link
1357. M. Thomas, R. W. Davis, J. Mol. Biol. **91** (1975) 315. Link
1358. T. I. Tikchonenko, E. V. Kramarov, B. A. Zavizion, B. S. Naroditsky, Gene **4** (1978) 195. Link
1359. J. Timko, A. H. Horwitz, J. Zelinka, G. Wilcox, J. Bacteriol. **145** (1981) 873. Link
1360. J. Tomassini, R. Roychoudhury, R. Wu, R. J. Roberts, Nucleic Acids Res. **5** (1978) 4055. Link
1361. C.-P. D. Tu, R. Roychoudhury, R. Wu, Biochem. Biophys. Res. Commun. **72** (1976) 355. Link
1362. S. Urieli-Shoval, Y. Gruenbaum, A. Razin, J. Bacteriol. **153** (1983) 274. Link
1363. H. van Heuverswyn, W. Fiers, Gene **9** (1980) 195. Link
1364. M. van Montagu, D. Sciaky, P. A. Myers, R. J. Roberts, cited in [1396], [1400]
1365. B. F. Vanyushin, A. N. Belozerskii, N. A. Kokurina, D. K. Kadirova, Nature (London) **218** (1968) 1066. Link
1366. L. P. Visentin, R. J. Watson, S. Martin, M. Zuker, cited in [1396], [1400]
1367. R. Wagner Jr., M. Meselson, Proc. Natl. Acad. Sci. USA **73** (1976) 4135. Link
1368. R. Y. Walder, J. L. Hartley, J. E. Donelson, J. A. Walder, Proc. Natl. Acad. Sci. USA **78** (1981) 1503. Link
1369. R. Y. Walder, J. A. Walder, J. E. Donelson, J. Biol. Chem. **259** (1984) 8015. Link
1370. R. J. Watson, I. Schildkraut, B. Q. Qiang, S. M. Martin, L. P. Visentin, FEBS Lett. **150** (1982) 114. Link
1371. R. Watson, M. Zuker, S. M. Martin L. P. Visentin, FEBS Lett. **118** (1980) 47. Link
1372. R. D. Wells, R. D. Klein, C. K. Singleton: “Type II Restriction Enzymes,” in P. D. Boyer (ed.): *The Enzymes*, vol. 14, Academic Press, New York 1981, p. 158.
1373. P. R. Whitehead, N. L. Brown, FEBS Lett. **143** (1982) 296. Link
1374. P. R. Whitehead, N. L. Brown, FEBS Lett. **155** (1983) 97. Link
1375. M. H. Wigler, Cell **24** (1981) 285. Link
1376. G. A. Wilson, F. E. Young, J. Mol. Biol. **97** (1975) 123. Link
1377. J. L. Woodhead, N. Bhave, A. D. B. Malcolm, Eur. J. Biochem. **115** (1981) 293. Link
1378. Y. Yamada, M. Murakami, Agric. Biol. Chem. **49** (1985) 3017. Link
1379. C. Yanisch-Perron, J. Vieira, J. Messing, Gene **33** (1985) 103. Link
1380. O. J. Yoo, K. L. Agarwal, J. Biol. Chem. **255** (1980) 10 559. Link
1381. R. Yuan, Annu. Rev. Biochem. **50** (1981) 285. Link
1382. M. Zbeau, R. Greene, P. A. Myers, R. J. Roberts, cited in [1396], [1400]
1383. M. Zabeau, R. J. Roberts, cited in [1396], [1400]
1384. B. S. Zain, R. J. Roberts, J. Mol. Biol. **115** (1977) 249. Link
1385. P. D. Boyer: *The Enzymes*, “Protein Synthesis, DNA Synthesis and Repair, RNA Synthesis, Energy-Linked ATPases, Synthetases,” vol. 10, Academic Press, New York 1974.
1386. P. D. Boyer: *The Enzymes*, “Nucleic Acids,” vol. 14, Part A, Academic Press, New York 1981.
1387. P. D. Boyer: *The Enzymes*, “Nucleic Acids,” vol. 15, Part B, Academic Press, New York 1981.
1388. J. G. Chirikjian, T. S. Papas: *Gene Amplification and Analysis*, vol. 2, Elsevier, Amsterdam 1981.

1389. S. T. Jacob in L. S. Hnilica (ed.): “Enzymes of Nucleic Acid Synthesis and Modification, DNA Enzymes,” *CRC Series in the Biochemistry and Molecular Biology of the Cell Nucleus*, vol. **I**, CRC Press, Inc., Boca Raton, Florida 1983.
1390. S. T. Jacob “Enzymes of Nucleic Acid Synthesis and Modification, RNA Enzymes,” in L. S. Hnilica (ed.): *CRC Series in the Biochemistry and Molecular Biology of the Cell Nucleus,* vol. II, CRC Press, Inc., Boca Raton, Florida, 1983.
1391. A. Kornberg: *DNA Replication*, Freeman and Co., San Fransisco 1980.
1392. T. Maniatis, E. F. Fritsch, J. Sambrook: *Molecular Cloning. A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1982.
1393. J. D. Watson: *Molecular Biology of the Gene*, 3rd ed., Benjamin, Inc., Menlo Park, CA 1977.
1394. R. B. Wicker: “DNA Replication,” in A. I. Laskin, J. A. Last (eds.): *Methods in Molecular Biology*, vol. 7, Marcel Dekker, New York 1974.
1395. J. G. Chirikjian: *Gene Amplification and Analysis,* vol. 1, Elsevier, Amsterdam 1980.
1396. C. Kessler, H.-J. Höltke: “Sequence Specificity of restriction endonucleases and methylases – a review (Edition 2),” Gene **47** (1986) 1. Link
1397. S. M. Linn, R. J. Roberts: *Nucleases*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1982.
1398. M. McClelland, M. Nelson: “The effect of site specific methylation on restriction endonuclease digestion,” Nucleic Acids Res. **13** (1985) r201. Link
1399. A. Razin, H. Cedar, A. D. Riggs: *DNA Methylation*, Springer Verlag, New York 1984.
1400. R. J. Roberts: “Restriction and modification enzymes and their recognition sequences,”, Nucleic Acids Res. **13** (1985) r165. Link
1401. I. Schildkraut: “Screening for and Characterizing Restriction Endonucleases,” in J. K. Setlow, A. Hollaender (eds.): *Genetic Engineering,* vol. 6, Plenum Press, New York 1984.
1402. T. A. Trautner: Methylation of DNA, in *Current Topics in Microbiology and Immunology,* Springer Verlag, New York 1984.
1403. G. Bertani, J. J. Weigle, J. Bacteriol. **65** (1953) 113. Link
1404. S. E. Luria, M. L. Human, J. Bacteriol. **64** (1952) 557. Link
1405. S. Linn, W. Arber, Proc. Natl. Acad. Sci. USA **59** (1968) 1300. Link
1406. M. Meselson, R. Yuan, Nature (London) **217** (1968) 1110. Link
1407. T. J. Kelly Jr., H. O. Smith, J. Mol. Biol. **51** (1970) 393. Link
1408. H. O. Smith, K. W. Wilcox, J. Mol. Biol. **51** (1970) 379. Link
1409. / L. Kauc, A. Piekarowicz, Eur. J. Biochem. **92** (1978) 417. Link
1410. I. G. Bogdarina, L. M. Vagabova, Ya. I. Burynaov, FEBS Lett. **68** (1976) 177. Link
1411. T. Shinomiya, S. Sato, Nucleic Acids Res. **8** (1980) 43. Link
1412. D. I Smith, F. R. Blattner, J. Davies, Nucleic Acids Res. **3** (1976) 343. Link
1413. H. Sugisaki, Y. Maekawa, S. Kanazawa, M. Takanami, Nucleic Acids Res. **10** (1982) 5747. Link
1414. I. Schildkraut, C. D. B. Banner, C. S. Rhodes, S. Parekh, Gene **27** (1984) 327. Link
1415. I. Schildkraut, D. Comb, cited in [1396], [1400]
1416. D. Kleid, Z. Humayun, A. Jeffrey, M. Ptashne, Proc. Natl. Acad. Sci. USA **73** (1976) 293. Link
1417. V. G. Kosykh, Ya. I. Buryanov, A. A. Bayev, Mol. Gen. Genet. **178** (1980) 717. Link
1418. D. Sciaky, R. J. Roberts, cited in [1396], [1400]
1419. J. Seurinck, A. van de Voorde, M. van Montagu, Nucleic Acids Res. **11** (1983) 4409. Link
1420. S. Jentsch, U. Günthert, T. A. Trautner, Nucleic Acids Res. **9** (1981) 2753. Link
1421. S. Kelly, D. R. Kaddurah, H. O. Smith, J. Biol. Chem. **260** (1985) 15 339. Link
1422. C. Kessler, unpublished results.
1423. C. Kessler, G. Nesch, R. Brack, Gene **16** (1981) 321. Link
1424. J. A. Langdale, P. A. Myers, R. J. Roberts, cited in [1396], [1400]

1425. G. V. Kholmina, B. A. Rebentish, Yu. S. Skoblov, A. A. Mironov, N. K. Yankovskii, Yu. I. Kozlov, L. I. Glatman, A. F. Moroz, V. G. Debabov, Dokl. Acad. Nauk SSSR **253** (1980) 495. Link
1426. R. J. Roberts, P. A. Myers, A. Morrison, K. Murray, J. Mol. Biol. **102** (1976) 157. Link
1427. R. J. Roberts, P. A. Myers, A. Morrison, K. Murray, J. Mol. Biol. **103** (1976) 199. Link
1428. R. J. Roberts, G. A. Wilson, F. E. Young, Nature (London) **265** (1977) 82. Link
1429. S. Sato, K. Nakazawa, T. Shinomiya, J. Biochem. (Tokyo) **88** (1980) 737. Link
1430. J. E. Brooks, R. M. Blumenthal, T. R. Gingeras, Nucleic Acids Res. **11** (1983) 837. Link
1431. A. Janulaitis, P. Povilionis, K. Sasnauskas, Gene **20** (1982) 197. Link
1432. S. A. Endow, R. J. Roberts, J. Mol. Biol. **112** (1977) 521. Link
1433. A. Kiss, G. Posfai, C. C. Keller, P. Venetianer, R. J. Roberts, Nucleic Acids Res. **13** (1985) 6403. Link
1434. J. S. Sussenbach, C. H. Monfoort, R. Schiphof, E. E. Stobberingh, Nucleic Acids Res. **3** (1976) 3193. Link
1435. M. Szekeres, cited in [1396], [1400]
1436. T. A. Bickle, K. Ineichen, Gene **9** (1980) 205. Link
1437. P. A. Sharp, B. Sudgen, J. Sambrook, Biochemistry **12** (1973) 3055. Link
1438. H. Shimotsu, H. Takahashi, H. Saito, Gene **11** (1980) 219. Link
1439. R. J. Roberts, J. B. Breitmeyer, N. F. Tabachnik, P. A. Myers, J. Mol. Biol. **91** (1975) 121. Link
1440. D. E. Garfin, H. M. Goodman, Biochem. Biophys. Res. Commun. **59** (1974) 108. Link
1441. D. W. Leung, A. C. P. Lui, H. Merilees, B. C. McBride, M. Smith, Nucleic Acids Res. **6** (1979) 17. Link
1442. B.-C. Lin, M. C. Chien, S.-Y. Lou, Nucleic Acids Res. **8** (1980) 6189. Link
1443. A. C. P. Lui, B. C. McBride, G. F. Vovis, M. Smith, Nucleic Acids Res. **6** (1979) 1. Link
1444. C. D. O'Connor, E. Metcalf, C. J. Wrighton, T. J. R. Harris, J. A. Saunders, Nucleic Acids Res. **12** (1984) 6701. Link
1445. R. Old, K. Murray, G. Roizes, J. Mol. Biol. **92** (1975) 331. Link
1446. J. A. Olson, P. A. Myers, R. J. Roberts, cited in [1396], [1400]
1447. M. Petrosyte, A. Janulaitis, Eur. J. Biochem. **121** (1982) 377. Link
1448. V. Pirrotta, Nucleic Acids Res. **3** (1976) 1747. Link
1449. K. Kita, N. Hiraoka, F. Kimizuka, A. Obayashi, H. Kojima, H. Takahashi, H. Saito, Nucleic Acids Res. **13** (1985) 7015. Link
1450. P. J. Greene, M. C. Betlach, H. W. Boyer, H. M. Goodman, Methods Mol. Biol. **7** (1974) 87. Link
1451. A. Dugaiczyk, J. Hedgpeth, H. W. Boyer, H. M. Goodman, Biochemistry **13** (1974) 503. Link
1452. C. H. Duncan, G. A. Wilson, F. E. Young, J. Bacteriol. **134** (1978) 338. Link
1453. M. S. Laskowski in P. D. Boyer (ed.): *The Enzymes*, vol. **4**, Academic Press, New York 1971, p. 313.
1454. P. P. Lau, H. B. Gray Jr., Nucleic Acids Res. **6** (1979) 331. Link
1455. E. Lazarides, U. Lindberg, Proc. Natl. Acad. Sci. USA **71** (1974) 4742. Link
1456. M. Y. W. T. Lee, C.-K. Tan, K. M. Downey, A. G. So, Biochemistry **23** (1984) 1906. Link
1457. K. Murray, S. G. Hughes, J. S. Brown, S. A. Bruce, Biochem. J. **159** (1976) 317. Link
1458. P. A. Myers, R. J. Roberts, cited in [1396], [1400]
1459. G. Nardone, J. George, J. G. Chirikjian, J. Biol. Chem. **259** (1984) 10 357.
1460. N. L. Brown, M. McClelland, P. R. Whitehead, Gene **9** (1980) 49. Link
1461. N. L. Brown, M. Smith, FEBS Lett. **65** (1976) 284. Link
1462. A. A. Janulaitis, P. S. Stakenas, J. B. Bitinaite, B. P. Jaskeleviciene, cited in [1396], [1400]
1463. A. R. Buchmann, L. Burnett, P. Berg in J. Tooze (ed.): *DNA Tumor Viruses*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1980, p. 779.
1464. D. G. Comb, I. Schildkraut, cited in .[1396], [1400]
1465. D. G. Comb. I. Schildkraut, G. Wilson, L. Greenough, cited in [1396], [1400]

1466. K. Mise, K. Nakajima, Gene **33** (1985) 357. Link
1467. S. Hattman, T. Keister, A. Gottehrer, J. Mol. Biol. **124** (1978) 701. Link
1468. R. Morgan, cited in [1396], [1400]
1469. R. Borsetti, R. Grandoni, I. Schildkraut, cited in [1396], [1400]
1470. J. Hedgpeth, H. M. Goodman, H. W. Boyer, Proc. Natl. Acad. Sci. USA **69** (1972) 3448. Link
1471. M.-T. Hsu, P. Berg, Biochemistry **17** (1978) 131. Link
1472. S. Bron, K. Murray, Mol. Gen. Genet. **143** (1975) 25. Link
1473. S. G. Hughes, K. Murray, Biochem. J. **185** (1980) 65. Link
1474. D. Jacobs, N. L. Brown, cited in [1396], [1400]
1475. A. Janulaitis, M. Petrusyte, V. Butkus, FEBS Lett. **161** (1983) 213. Link
1476. T. R. Gingeras, L. Greenough, I. Schildkraut, R. J. Roberts, Nucleic Acids Res. **9** (1981) 4525. Link
1477. T. R. Gingeras, J. P. Milazzo, R. J. Roberts, Nucleic Acids Res. **5** (1978) 4105. Link
1478. T. R. Gingeras, P. A. Myers, J. A. Olson, F. A. Hanberg, R. J. Roberts, J. Mol. Biol. **118** (1978) 113. Link
1479. S. P. Lynn, L. K. Cohen, S. Kaplan, J. F. Gardner, J. Bacteriol. **142** (1980) 380. Link
1480. R. A. Makula, R. B. Meagher, cited in [1396], [1400]
1481. J. R. Arrand, P. A. Myers, R. J. Roberts, cited in [1396], [1400]
1482. A. Pope, S. P. Lynn, J. F. Gardner, cited in [1396], [1400]
1483. I. J. Purvis, B. E. B. Moseley, Nucleic Acids Res. **11** (1983) 5467. Link
1484. B.-Q. Qiang, I. Schildkraut, L. Visentin, cited in [1396], [1400]
1485. S. A. Endow, J. Mol. Biol. **114** (1977) 441. Link
1486. V. M. Kramarov, V. V. Smolyaninov, Biokhimiya (Moscow) **46** (1981) 1526. Link
1487. R. Grosskopf, C. Kessler, unpublished results.
1488. J. H. Middleton, M. H. Edgell, C. A. Hutchison III, cited in [1396], [1400]
1489. J. H. Middleton, P. V. Stankus, M. H. Edgell, C. A. Hutchison III, cited in [1396], [1400]
1490. T. A. Bickle in S. M. Linn, R. J. Roberts (eds.): *Nucleases*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1982, p. 85.
1491. B. Endlich, S. Linn in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1981, p. 137.
1492. P. Modrich, Q. Rev. Biophys. **12** (1979) 315. Link
1493. R. Fuchs, R. Blakesley: "Guide to the Use of Type II Restriction endonucleases," in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. 100, Academic Press, New York 1983, p. 3.
1494. P. Modrich, R. J. Roberts in S. M. Linn, R. J. Roberts (eds.): *Nucleases*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1982, p. 101.
1495. K. Nath, B. A. Azzolina in J. G. Chirikjian (ed.): *Gene Amplification and Analysis*, vol. **1**, Elsevier, Amsterdam 1981, p. 113.
1496. D. Nathans, H. O. Smith, Ann. Rev. Biochem. **44** (1975) 273. Link
1497. R. J. Roberts, CRC Crit. Rev. Biochem. **4** (1976) 123. Link
1498. H. O. Smith, Science **205** (1979) 455. Link
1499. S. M. Hadi, B. Bächi, S. Iida, T. A. Bickle, J. Mol. Biol. **165** (1983) 19. Link
1500. S. Iida, J. Meyer, B. Bächi, M. Stählhammar Carlemalm, S. Schrickel, T. A. Bickle, W. Arber, J. Mol. Biol. **165** (1983) 1. Link
1501. H. O. Smith, D. Nathans, J. Mol. Biol. **81** (1973) 419. Link
1502. G. Duyk, J. Leis, M. Longiaru, A. M. Skalka, Proc. Natl. Acad. Sci. USA **80** (1983) 6745. Link
1503. R. Kostriken, M. Zoller, F. Heffron in L. Leive (ed.): *Microbiology*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1985, p. 295.
1504. C. Korch, P. Hagblom, S. Normark, J. Bacteriol. **155** (1983) 1324. Link
1505. J. Reaston, M. G. C. Duyvesteyn, A. de Waard, Gene **20** (1982) 103. Link

1506. P. J. Greene, H. L. Heyneker, F. Bolivar, R. L. Rodriguez, M. C. Betlach, A. A. Covarrubias *et al.*, Nucleic Acids Res. **5** (1978) 2373. Link
1507. V. Pirrotta, T. A. Bickle in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 89.
1508. J. Botterman, M. Zabeau, Gene **37** (1985) 229. Link
1509. S.-C. Cheng, R. Kim, K. King, S.-H. Kim, P. Modrich, J. Biol. Chem. **259** (1984) 11571. Link
1510. P. J. Greene, M. Gupta, H. W. Boyer, W. E. Brown, J. M. Rosenberg, J. Biol. Chem. **256** (1981) 2143. Link
1511. T. M. Haqqi, S. Ahmad, N. S. Ahmad, M. Ahmad, A. Hasnain, M. Siddigi, S. M. Hadi, Indian. J. Biochem. Biophys. **22** (1985) 252. Link
1512. A. K. Newman, R. A. Rubin, S. H. Kim, P. Modrich, J. Biol. Chem. **256** (1981) 2131. Link
1513. P. A. Luke, S. E. Halford, Gene **37** (1985) 241. Link
1514. L. Bougueleret, M. Schwarzstein, A. Tsugita, M. Zabeau, Nucleic Acids Res. **12** (1984) 3659. Link
1515. L. Bougueleret, M. L. Tenchini, J. Botterman, M. Zabeau, Nucleic Acids Res. **13** (1985) 3823. Link
1516. A. S. Kraev, A. N. Kravets, B. K. Chernov, K. S. Skryabin, A. A. Baev, Mol. Biol. (Moscow) **19** (1985) 278. Link
1517. M. B. Mann, H. O. Smith, cited in [1396], [1400]
1518. B. Schoner, S. Kelly, H. O. Smith, Gene **24** (1983) 227. Link
1519. T. R. Gingeras, J. E. Brooks, Proc. Natl. Acad. Sci. USA **80** (1983) 402. Link
1520. S. H. Lee, H. M. Rho, Korean J. Genet. **7** (1985) 42. Link
1521. R. M. Blumenthal, S. A. Gregory, J. S. Cooperider, J. Bacteriol. **164** (1985) 501. Link
1522. S. Lacks, B. Greenberg, J. Mol. Biol. **114** (1977) 153. Link
1523. D. L. Daniels, J. L. Schroeder, W. Szybalski, F. Sanger, F. R. Blattner in W. R. Hendrix, J. W. Roberts, F. W. Stahl, R. A. Weisberg (eds.): *Lambda II*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1983, p. 469.
1524. D. L. Daniels, J. L. Schroeder, W. Szybalski, F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, F. R. Blattner in W. R. Hendrix, J. W. Roberts, F. W. Stahl, R. A. Weisberg (eds.): *Lambda II*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1983, p. 519.
1525. F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, J. Mol. Biol. **162** (1982) 729. Link
1526. F. M. Pohl, R. Thomae, A. Karst, Eur. J. Biochem. **123** (1982) 141. Link
1527. K. Schmid, M. Thomm, A. Laminet, F. G. Laue, C. Kessler, K. O. Stetter, R. Schmitt, Nucleic Acids Res. **12** (1984) 2619. Link
1528. A. H. A. Bingham, T. Atkinson, D. Sciaky, R. J. Roberts, Nucleic Acids Res. **5** (1978) 3457. Link
1529. J. A. Lautenberger, M. H. Edgell, C. A. Hutchison III, Gene **12** (1980) 171. Link
1530. S. Lacks, B. Greenberg, J. Biol. Chem. **250** (1975) 4060. Link
1531. S. Sato, C. A. Hutchison III, J. I. Harris, Proc. Natl. Acad. Sci. USA **74** (1977) 542. Link
1532. T. A. Bickle, V. Pirrotta, R. Imber in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 132.
1533. R. P. Grandoni, I. Schildkraut, cited in [1396], [1400]
1534. R. Nussinov, Nucleic Acids Res. **8** (1980) 4545. Link
1535. J. R. Arrand, P. A. Myers, R. J. Roberts, J. Mol. Biol. **118** (1978) 127. Link
1536. D. G. Comb, P. Parker, R. Grandoni, I. Schildkraut, cited in [1396], [1400]
1537. K. W. C. Peden, Gene **22** (1983) 277. Link
1538. J. G. Sutcliffe, Cold Spring Harbor Symp. Quant. Biol. **43** (1979) 77. Link
1539. P. S. Miller, D. M. Cheng, N. Dreon, K. Jayaraman, L. S. Kan, E. E. Leutzinger, S. M. Pulford, P. O. P. Ts-O, Biochemistry **19** (1980) 4688. Link
1540. S. Forsblum, R. Rigler, M. Ehrenberg, U. Pettersson, L. Philipson, Nucleic Acids Res. **3** (1976) 3255. Link
1541. K. Armstrong, W. R. Bauer, Nucleic Acids Res. **10** (1982) 993. Link

1542. N. L. Brown, M. Smith, Proc. Natl. Acad. Sci. USA **74** (1977) 3213. Link
1543. D. G. Comb, G. Wilson, cited in [1396], [1400]
1544. G. F. Fitzgerald, C. Daly, L. R. Brown, T. R. Gingeras, Nucleic Acids Res. **10** (1982) 8171. Link
1545. M. Ehrlich, R. Y.-H. Wang, Science **212** (1981) 1350. Link
1546. L. M. Kunkel, M. Silberklang, B. J. McCarthy, J. Mol. Biol. **132** (1979) 133. Link
1547. I. Schildkraut, R. Wise, R. Borsetti, B.-Q. Quiang, cited in [1396], [1400]
1548. C. M. de Wit, B. M. M. Dekker, A. C. Neele, A. de Waard, FEBS Lett. **180** (1985) 219. Link
1549. R. Grosskopf, W. Wolf, C. Kessler, Nucleic Acids Res. **13** (1985) 1517. Link
1550. B.-Q. Qiang, I. Schildkraut, Nucleic Acids Res. **12** (1984) 4507. Link
1551. A. W. Hu, A. H. Marschel, cited in [1396], [1400]
1552. P. L. Molloy, R. H. Symons, Nucleic Acids Res. **8** (1980) 2939. Link
1553. B. Hofer, G. Ruhe, A. Koch, H. Köster, Nucleic Acids Res. **10** (1982) 2763. Link
1554. K. Nishigaki, Y. Kaneko, H. Wakuda, Y. Husimi, T. Tanaka, Nucleic Acids Res. **13** (1985) 5747. Link
1555. N. L. Brown, C. A. Hutchison III, M. Smith, J. Mol. Biol. **140** (1980) 143. Link
1556. H. Sugisaki, Gene **3** (1978) 17. Link
1557. H. Sugisaki, S. Kanazawa, Gene **16** (1981) 73. Link
1558. A. J. Podhaiska, W. Szybalski, Gene **40** (1985) 175. Link
1559. W. Szybalski, Gene **40** (1985) 169. Link
1560. B. Dreiseikelmann, R. Eichenlaub, W. Wackernagel, Biochim. Biophys. Acta **562** (1979) 418. Link
1561. R. E. Gelinas, P. A. Myers, R. J. Roberts, J. Mol. Biol. **114** (1977) 169. Link
1562. S. Hattman, J. E. Brooks, M. Masurekar, J. Mol. Biol. **126** (1978) 367. Link
1563. M. B. Mann, H. O. Smith, Nucleic Acids Res. **4** (1977) 4211. Link
1564. Y. Gruenbaum, R. Stein, H. Cedar, A. Razin, FEBS Lett. **124** (1981) 67. Link
1565. A. Razin, A. D. Riggs, Science **210** (1980) 604. Link
1566. B. J. Bolton, G. Nesch, M. J. Comer, W. Wolf, C. Kessler, FEBS Lett. **182** (1985) 130. Link
1567. J. George, J. G. Chirikjian, Proc. Natl. Acad. Sci. USA **79** (1982) 2432. Link
1568. E. Malyguine, P. Vannier, P. Yot, Gene **8** (1980) 163. Link
1569. T. Shinomiya, M. Kobayashi, S. Sato, T. Uchida, J. Biochem. (Tokyo) **92** (1982) 1823. Link
1570. R. C. Gardner, A. J. Horwarth, J. Messing, R. J. Shepherd, DNA **1** (1982) 109. Link
1571. J. M. Rosenberg, P. Greene, DNA **1** (1982) 117. Link
1572. B. Rexer, unpublished results.
1573. S. Hattman, in P. D. Boyer (eds.): "DNA Methylation," in *The Enzymes*, vol. 14, Academic Press, New York 1981, p. 517.
1574. M. S. May, S. Hattman, J. Bacteriol. **122** (1975) 129. Link
1575. M. Meijer, E. Beck, F. G. Hansen, H. E. N. Bergmans *et al.*, Proc. Natl. Acad. Sci. USA **76** (1979) 580. Link
1576. K. Sugimoto, A. Oka, H. Sugisaki, M. Takanami, A. Nishimura, Y. Yasuda, Y. Hirota, Proc. Natl. Acad. Sci. USA **76** (1979) 575. Link
1577. H. W. Boyer, L. T. Chow, A. Dugaiczyk, J. Hedgpeth, H. M. Goodman, Nature (London) **244** (1973) 40. Link
1578. Y. Gruenbaum, H. Cedar, A. Razin, Nucleic Acids Res. **9** (1981) 2509. Link
1579. H. Mayer, R. Grosschedl, H. Schütte, G. Hobom, Nucleic Acids Res. **9** (1981) 4833. Link
1580. M. McClelland, L. G. Kessler, M. Bittner, Proc. Natl. Acad. Sci. USA **81** (1984) 983. Link
1581. M. Nelson, C. Christ, I. Schildkraut, Nucleic Acids Res. **12** (1984) 5165. Link
1582. M. McClelland, Nucleic Acids Res. **9** (1981) 6795. Link
1583. N. Panayotatos, Gene **31** (1984) 291. Link
1584. L.-H. Huang, C. M. Farnet, K. C. Ehrlich, M. Ehrlich, Nucleic Acids Res. **10** (1982) 1579. Link

1585. D. A. Kaplan, D. P. Nierlich, J. Biol. Chem. **250** (1975) 2395. Link
1586. M. Goppelt, J. Langowski, A. Pingoud, W. Haupt, C. Urbanke, H. Mayer, G. Maass, Nucleic Acids Res. **9** (1981) 6115. Link
1587. M.-G. Nilsson, C. Skarped, G. Magnusson, FEBS Lett. **145** (1982) 360. Link
1588. M. Österlund, H. Luthman, S. V. Nilsson, G. Magnusson, Gene **20** (1982) 121. Link
1589. A. Pingoud, Eur. J. Biochem. **147** (1985) 105. Link
1590. A. Pingoud, C. Urbanke, J. Alves, H.-J. Ehbrecht, M. Zabeau, C. Gualerzi, Biochemistry **23** (1984) 5697. Link
1591. M. Kuosmanen, H. Pösö, FEBS Lett. **179** (1985) 17. Link
1592. D. F. Dombroski, A R. Morgan, J. Biol. Chem. **260** (1985) 415. Link
1593. C. Kessler, B. J. Bolton, M. J. Comer, J. Cell. Biochem. **10 D** (1986) 100, # 058. Link
1594. J. P. Slater, A. S. Mildvan, L. A. Loeb, Biochem. Biophys. Res. Commun. **44** (1971) 37. Link
1595. T. M. Jovin, P. T. Englund, L. Bertsch, J. Biol. Chem. **244** (1969) 6996. Link
1596. D. Brutlag, M. R. Atkinson, P. Setlow, A. Kornberg, Biochem. Biophys. Res. Commun. **37** (1969) 982. Link
1597. H. Klenow, I. Henningsen, Proc. Natl. Acad. Sci. USA **65** (1970) 168. Link
1598. P. T. Englund, M. P. Deutscher, T. Jovin, R. B. Kelly, N. R. Cozzarelli, A. Kornberg, Cold Spring Harbor Symp. Quant. Biol. **33** (1968) 1. Link
1599. J. Griffith, J. A. Huberman, A. Kornberg, J. Mol. Biol. **55** (1971) 209. Link
1600. T. M. Jovin, P. T. Englund, A. Kornberg, J. Biol. Chem. 244 (1969) 3009.
1601. C. C. Richardson, I. R. Lehman, A. Kornberg, J. Biol. Chem. **239** (1964) 251. Link
1602. H. M. Shepard, B. Polisky in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **68**, Academic Press, New York 1980, p. 503.
1603. P. W. J. Rigby, M. Dieckmann, C. Rhodes, P. Berg, J. Mol. Biol. **113** (1977) 237. Link
1604. H. Jacobsen, H. Klenow, K. Overgaard-Hansen, Eur. J. Biochem. **45** (1974) 623. Link
1605. J. L. Telford, A. Kressmann, R. A. Koski, R. Grosschedl, F. Müller, S. G. Clarkson, M. L. Birnstiel, Proc. Natl. Acad. Sci. USA **76** (1979) 2590. Link
1606. F. Rougeon, P. Kourilsky, B. Mach, Nucleic Acids Res. **2** (1975) 2365. Link
1607. F. Sanger, S. Nicklen, A. R. Coulson, Proc. Natl. Acad. Sci. USA **74** (1977) 5463. Link
1608. A. P. Feinberg, B. Vogelstein, Anal. Biochem. **132**, (1983) 6. Link
1609. M. Smith, S. Gillam in J. K. Setlow, A. Hollaender (eds.): *Genetic Engineering. Principle and Methods*, vol. **3**, Plenum Press, New York 1981, p. 1.
1610. M. Goulian, Z. J. Lucas, A. Kornberg, J. Biol. Chem. **243** (1968) 627. Link
1611. B. M. Alberts, L. Frey, Nature (London) **227** (1970) 1313. Link
1612. N. G. Nossal, J. Biol. Chem. **249** (1974) 5668. Link
1613. W. M. Huang, I. R. Lehman, J. Biol. Chem. **247** (1972) 3139. Link
1614. C. S. Craik, C. Largman, T. Fletscher, S. Roczniak, P. J. Barr, R. Fletterick, W. J. Rutter, Science **228** (1985) 291. Link
1615. K. C. Deen, T. A. Landers, M. Berninger, Anal. Biochem. **135** (1983) 456. Link
1616. R. N. Eisenman, W. S. Mason, M. Linial, J. Virol **86** (1980) 62. Link
1617. G. F. Gerard: “Reverse Transcriptase,” in S. T. Jacob (ed.): *Enzymes of Nucleic Acid Synthesis and Modification*, vol. 1, CRC Press, West Palm Beach, Florida, 1983, p. 1.
1618. R. D. Schiff, O. P. Grandgenett, J. Virol. **36** (1980) 889. Link
1619. D. S. Gregerson, J. Albert, T. W. Reid, Biochemistry **19** (1980) 301. Link
1620. J. Leis, G. Duyk, S. Johnson, M. L. Ongiaru, A. Skalka, J. Virol. **45** (1983) 727. Link
1621. I. M. Verma, Biochim. Biophys. Acta **473** (1977) 1. Link
1622. J. J. Champoux, E. Gilboa, D. Baltimore, J. Virol. **49** (1984) 686. Link
1623. W. Keller, R. Crouch, Proc. Natl. Acad. Sci. USA **69** (1972) 3360. Link
1624. A. A. Kiessling, M. Goulian, Biochem. Biophys. Res. Commun. **71** (1976) 1069. Link

1625. A. Ueno, A. Ishihama, J. Biochem. (Tokyo) **91** (1982) 323. Link
1626. T. Fukui, A. Ishihama, E. Ohtsuka, M. Ikehara, R. Fukuda, J. Biochem. (Tokyo) **91** (1982) 331. Link
1627. K. Fujinaga, J. T. Parsons, J. W. Beard, D. Beard, M. Green, Proc. Natl.. Acad. Sci. USA **67** (1970) 1432. Link
1628. C. Brack, M. Hirama, R. Lenhard-Schuller, S. Tonegawa, Cell **15** (1978) 1. Link
1629. R. Breathnach, J. L. Mandel, P. Chambon, Nature (London) **270** (1977) 314. Link
1630. S. M. Tilghman, P. J. Curtis, D. C. Tiemeier, P. Leder, C. Weissmann, Proc. Natl. Acad. Sci. USA **75** (1978) 1309. Link
1631. N. Kitamura, B. L. Semler, P. G. Rothberg, G. R. Larsen, C. J. Alder, A. J. Dorner *et al.*, Nature (London) **291** (1981) 547. Link
1632. J. K. Rose, Cell **19** (1980) 415. Link
1633. R. Roychoudhury, E. Jay, R. Wu, Nucleic Acids Res. **3** (1976) 863. Link
1634. A. F. C. Efstratiadis, L. Villa-Komaroff in J. K. Setlow, A. Hollaender (eds.): *Genetic Engineering*, vol. **1**, Plenum Press, New York 1979, p. 1.
1635. H. M. Goodman, R. J. McDonald in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **68**, Academic Press, New York 1980, p. 75.
1636. A. Torriani in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **12**, Academic Press, New York 1968, p. 212.
1637. D. C. Hinkle, C. C. Richardson, J. Biol. Chem. **250** (1975) 5523. Link
1638. W. Salser in A. M. Chakrabarty (ed.): *Genetic Engineering*, vol. **1**, CRC Press, West Palm Beach, Florida 1979, p. 53.
1639. P. K. Ghosh, V. B. Reddy, M. Piatak, P. Lebowitz, S. M. Weissman in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 580.
1640. H. Okayama, P. Berg, Mol. Cell. Biol. **2** (1982) 161. Link
1641. F. J. Bollum: "Terminal Deoxynucleotidyl Transferase," in P. D. Boyer (ed.): *The Enzymes*, vol. 10, Academic Press, New York 1974, p. 145.
1642. L. M. S. Chang, F. J. Bollum, J. Biol. Chem. **246** (1971) 909. Link
1643. C. Vincent, P. Tchen, M. Cohen-Solal, P. Kourilsky, Nucleic Acids Res. **10** (1982) 6787. Link
1644. R. L. Ratcliff in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1981, p. 105.
1645. G.–R. Deng, R. Wu in S. P. Colowick, N. O. Kaplan (eds.): *Methods in Enzymology*, vol. **100**, Academic Press, New York 1983. p. 96.
1646. U. Bertazzoni, S. D. Ehrlich, G. Bernardi, Biochim. Biophys. Acta **312** (1973) 192. Link
1647. C.-P. D. Tu, S. N. Cohen, Gene **10** (1980) 177. Link
1648. E. T. Butler, M. J. Chamberlin, J. Biol. Chem. **257** (1982) 5772. Link
1649. G. Hartmann, K. O. Honikel, F. Knüsel, J. Nüesch, Biochim. Biophys. Acta **145** (1967) 843. Link
1650. P. R. Chakraborty, P. Sarkar, H. H. Huang, U. Maitra, J. Biol. Chem. **248** (1973) 6637 Link
1651. M. Chamberlin, J. McGrath, L. Waskell, Nature (London) **228** (1970) 227. Link
1652. H. C. Towle, J. F. Jolly, J. A. Boezi, J. Biol. Chem. **250** (1975) 1723. Link
1653. G. A. Kassavetis, E. T. Butler, D. Roulland, M. J. Chamberlin, J. Biol. Chem. **257** (1982) 5779. Link
1654. D. A. Melton, P. A. Krieg, M. R. Rebagliati, T. Maniatis, K. Zinn, M. R. Green, Nucleic Acids Res. **12** (1984) 7035. Link
1655. M. R. Green, T. Maniatis, D. A. Melton, Cell **32** (1983) 681. Link
1656. A. R. Krainer, T. Maniatis, B. Ruskin, M. R. Green, Cell **36** (1984) 993. Link
1657. S. B. Zimmermann, B. H. Pheiffer in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 315.
1658. G.–M. Church, W. Gilbert, Proc. Natl. Acad. Sci. USA **81** (1984) 1991. Link
1659. S. Stahl, K. Zinn, J. Mol. Biol. **148** (1981) 481. Link
1660. P. Davanloo, A. H. Rosenberg, J. J. Dunn, F. W. Studier, Proc. Natl. Acad. Sci. USA **81** (1984) 2035. Link

1661. J. J. Dunn, F. W. Studier, J. Mol. Biol. **148** (1981) 303. Link
1662. S. Tabor, C. C. Richardson, Proc. Natl. Acad. Sci. USA **82** (1985) 1074. Link
1663. G. E. Smith, M. D. Summers, Anal. Biochem. **109** (1980) 123. Link
1664. E. M. Southern, J. Mol. Biol. **98** (1975) 503. Link
1665. K. Zinn, D. Dimaio, T. Maniatis, Cell **34** (1983) 865. Link
1666. D. A. Melton, Proc. Natl. Acad. Sci. USA **82** (1985) 144. Link
1667. M. Kunitz, J. Gen. Physiol. **33** (1950) 363. Link
1668. M. S. Laskowski in P. D. Boyer (ed.): *The Enzymes*, vol. **5**, Academic Press, New York 1961, p. 123.
1669. T.-H. Liao, J. Salnikow, S. Moore, W. H. Stein, J. Biol. Chem. **248** (1973) 1489. Link
1670. J. Salnikow, S. Moore, W. H. Stein, J. Biol. Chem. **245** (1970) 5685. Link
1671. P. A. Price, W. H. Stein, S. Moore, J. Biol. Chem. **244** (1969) 929. Link
1672. M. S. Laskowski in P. D. Boyer (ed.): *The Enzymes*, vol. **4**, Academic Press, New York 1971, p. 289.
1673. S. Moore in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 281.
1674. P. A. Price, J. Biol. Chem. **250** (1975) 1981. Link
1675. V. W. Campbell, D. A. Jackson, J. Biol. Chem. **255** (1980) 3726. Link
1676. E. Junowicz, J. H. Spencer, Biochim. Biophys. Acta **312** (1973) 85. Link
1677. S. E. Hitchcock, J. Biol. Chem. **255** (1980) 5668. Link
1678. H. G. Mannherz, R. S. Goody, M. Konrad, E. Nowak, Eur. J. Biochem. **104** (1979) 367. Link
1679. B. Weiss, J. Biol. Chem. **251** (1976) 1968. Link
1680. C. C. Richardson, A. Kornberg, J. Biol. Chem. **239** (1964) 242. Link
1681. R. N. Rao, S. G. Rogers, Gene **3** (1978) 247. Link
1682. S. G. Rogers, B. Weiss in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 201.
1683. S. G. Rogers, B. Weiss, Gene **11** (1980) 187. Link
1684. L.-H. Guo, R. Wu in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **100**, Academic Press, New York 1983, p. 60.
1685. A. J. H. Smith, Nucleic Acids Res. **6** (1979) 831. Link
1686. W. Linxweiler, W. Hörz, Nucleic Acids Res. **10** (1982) 4845. Link
1687. L.-H. Guo, R. Wu, Nucleic Acids Res. **10** (1982) 2065. Link
1688. L.-H. Guo, R. C. A. Yang, R. Wu, Nucleic Acids Res. **11** (1983) 5521. Link
1689. B. Royer-Pokora, L. K. Gordon, W. A. Haseltine, Nucleic Acids Res. **9** (1981) 4595. Link
1690. A. E. Oleson, M. Sasakuma, Arch. Biochem. Biophys. **204** (1980) 361. Link
1691. V. M. Vogt, Eur. J. Biochem. **33** (1973) 192. Link
1692. G. W. Rushizky, V. A. Shaternikov, J. H. Mozejko, H. A. Sober, Biochemistry **14** (1975) 4221. Link
1693. K. Horikoshi, Biochim. Biophys. Acta **240** (1971) 532. Link
1694. K. Zechel, K. Weber, Eur. J. Biochem. **77** (1977) 133. Link
1695. V. M. Vogt in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 248.
1696. R. G. Wiegand, G. N. Godson, C. M. Radding, J. Biol. Chem. **250** (1975) 8848. Link
1697. A. J. Berk, P. A. Sharp, Cell **12** (1977) 721. Link
1698. R. A. Hallewell, S. Emtage, Gene **9** (1980) 27. Link
1699. A. J. Berk, P. A. Sharp, Cell **14** (1978) 695. Link
1700. A. J. Berk, P. A. Sharp, Proc. Natl. Acad. Sci. USA **75** (1978) 1274. Link
1701. B. Weiss, S. G. Rogers, A. F. Taylor in P. C. Hanawalt, E. C. Friedberg, C. F. Fox (eds.): *DNA Repair Mechanisms*, Academic Press, New York 1978, p. 191.
1702. R. F. Weaver, C. Weissmann, Nucleic Acids Res. **7** (1979) 1175. Link
1703. R. Grosschedl, M. L. Birnstiel, Proc. Natl. Acad. Sci. USA **77** (1980) 1432. Link
1704. J. W. Chase, C. C. Richardson, J. Biol. Chem. **249** (1974) 4545. Link

1705. F. Gannon, J. M. Jeltsch, F. Perrin, Nucleic Acids Res. **8** (1980) 4405. Link
1706. J. Favaloro, R. Treisman, R. Kamen in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 718.
1707. U. Weidle, C. Weissmann, Nature (London) **303** (1983) 442. Link
1708. H. B. Gray Jr., W. B. Upholt, J. Vinograd, J. Mol. Biol. **62** (1971) 1. Link
1709. R. J. Legerski, J. L. Hodnett, H. B. Gray Jr., Nucleic Acids Res. **5** (1978) 1445. Link
1710. N. Panayotatos, K. Truong, Nucleic Acids Res. **9** (1981) 5679. Link
1711. H. D. Robertson, J. J. Dunn, J. Biol. Chem. **250** (1975) 3050. Link
1712. G. Hillenbrand, W. L. Staudenbauer, Nucleic Acids Res. **10** (1982) 833. Link
1713. I. Berkower, J. Leis, J. Hurwitz, J. Biol. Chem. **248** (1973) 5914. Link
1714. U. Gubler, B. J. Hoffman, Gene **25** (1983) 263. Link
1715. L. I. Grossman, R. Watson, J. Vinograd, Proc. Natl. Acad. Sci. USA **70** (1973) 3339. Link
1716. J. N. Vournakis, A. Efstratiadis, F. C. Kafatos, Proc. Natl. Acad. Sci. USA **72** (1975) 2959. Link
1717. H. Donis-Keller, Nucleic Acids Res. **7** (1979) 179. Link
1718. C. H. W. Hirs, S. Moore, W. H. Stein, J. Biol. Chem. **219** (1956) 623. Link
1719. M. Kunitz, J. Gen. Physiol. **24** (1940) 15. Link
1720. M. S. Boguski, P. A. Hieter, C. C. Levy, J. Biol. Chem. **255** (1980) 2160. Link
1721. C. C. Levy, T. P. Karpetsky, J. Biol. Chem. **255** (1980) 2153. Link
1722. G. Krupp, H. J. Gross, Nucleic Acids Res. **6** (1979) 3481. Link
1723. G. Krupp, H. J. Gross in P. E. Agris, R. A. Kopper (eds.): *The Modified Nucleosides of Transfer RNA*, vol. **2**, Lisa, New York 1983, p. 11.
1724. F. Egami, K. Takahashi, T. Uchida in J. N. Davidson, W. E. Cohen (eds.): *Progress in Nucleic Acid Research and Molecular Biology*, vol. **3**, Academic Press, New York 1964, p. 59.
1725. K. Sato, F. Egami, J. Biochem. (Tokyo) **44** (1957) 753. Link
1726. K. Takahashi, J. Biochem. **49** (1961) 1. Link
1727. J. N. Heins, H. Taniuchi, C. B. Anfisen in J. Cantoni, D. Davies (eds.): *Procedures in Nucleic Acid Research*, vol. **1**, Harper and Row, New York 1966, p. 79
1728. H. R. B. Pelham, R. J. Jackson, Eur. J. Biochem. **67** (1976) 247. Link
1729. G. Winter, G. G. Brownlee, Nucleic Acids Res. **5** (1978) 3129. Link
1730. C. R. Woese, G. E. Fox, Proc. Natl. Acad. Sci. USA **74** (1977) 5088. Link
1731. E. Stackebrandt, W. Ludwig, K. H. Schleifer, H. J. Gross, J. Mol. Evol. **17** (1981) 227. Link
1732. P. Setlow, A. Kornberg, J. Biol. Chem. **247** (1972) 232. Link
1733. A. Efstratiadis, J. N. Vournakis, H. Donis-Keller, G. Chaconas, D. K. Dougall, F. C. Kafatos, Nucleic Acids Res. **4** (1977) 4165. Link
1734. A. M. Maxam, W. Gilbert in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 499.
1735. D. A. Soltis, O. C. Uhlenbeck, J. Biol. Chem. **257** (1982) 11 332. Link
1736. T. E. Englund, O. C. Uhlenbeck, Nature (London) **275** (1978) 560. Link
1737. H. Sano, G. Feix, Eur. J. Biochem. **71** (1976) 577. Link
1738. A. Sippel, Eur. J. Biochem. **37** (1973) 31. Link
1739. G. Wengler, G. Wengler, Nature (London) **282** (1979) 754. Link
1740. Y. Yamada, H. Ishikura, Biochim. Biophys. Acta **402** (1975) 285. Link
1741. S. H. Chang, C. K. Brum, M. Silberklang, U. L. Rajbhandary, L. I. Hecker, W. E. Barnett, Cell **9** (1976) 717. Link
1742. M. Silberklang, A. M. Gillum, U. L. Rajbhandary, Nucleic Acids Res. **4** (1977) 4091. Link
1743. M. Silberklang, A. Prochiantz, A.-L. Haenni, U. L. Rajbhandary, Eur. J. Biochem. **72** (1977) 465. Link
1744. D. A. Peattie, Proc. Natl. Acad. Sci. USA **76** (1979) 1760. Link

1745. H. Donis-Keller, A. M. Maxam, W. Gilbert, Nucleic Acids Res. **4** (1977) 2527. Link
1746. U. L. Rajbhandary, Fed. Proc. Fed. Am. Soc. Exp. Biol. **39** (1980) 2815. Link
1747. M. Silberklang, A. M. Gillum, U. L. Rajbhandary in L. Grosman, K. Moldave (eds).: *Methods in Enzymology*, vol. **59**, Academic Press, New York 1979, p. 58.
1748. A. Simoncsits, G. G. Brownlee, R. S. Brown, J. R. Rubin, H. Guilley, Nature (London) **269** (1977) 833. Link
1749. E. Mössner, M. Boll, G. Pfleiderer, Hoppe-Seyler's Z. Physiol. Chem. **361** (1980) 543. Link
1750. M. Fosset, D. Chappelet-Tordo, M. Lazdunski, Biochemistry **13** (1974) 1783. Link
1751. M. Simsek, J. Ziegenmeyer, J. Heckman, U. L. Rajbhandary, Proc. Natl. Acad. Sci. USA **70** (1973) 1041. Link
1752. J. R. Lillehaug, K. Kleppe, Biochemistry **14** (1975) 1225. Link
1753. A. M. Maxam, W. Gilbert, Proc. Natl. Acad. Sci. USA **74** (1977) 560. Link
1754. K. Randerath, R. C. Gupta, E. Randerath in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 638.
1755. C. C. Richardson, Proc. Natl. Acad. Sci. USA **54** (1968) 158. Link
1756. D. Kamp, R. Kahmann, D. Zipser, R. J. Roberts, Mol. Gen. Genet. **154** (1977) 231. Link
1757. C.-P. D. Tu, R. Wu in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 620.
1758. F. Bolivar, K. Backman: “Plasmids of Escherichia coli as Cloning Vectors,” in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. 68, Academic Press, New York 1980, p. 245.
1759. D. Ish-Horowicz, J. F. Burke, Nucleic Acids Res. **9** (1981) 2989. Link
1760. A. Ullrich, J. Shine, J. Chirgwin, R. Pictet, E. Tischer, W. J. Rutter, H. M. Goodman, Science **196** (1977) 1313. Link
1761. N. E. Murray, S. A. Bruce, K. Murray, J. Mol. Biol. **132** (1979) 493. Link
1762. G. G. Wilson, N. E. Murray, J. Mol. Biol. **132** (1979) 471. Link
1763. A. Panet, J. H. van de Sande, P. C. Loewen, H. G. Khorana, A. J. Raae, J. R. Lillehaug, K. Kleppe, Biochemistry **12** (1973) 5045. Link
1764. S. Soderhall, T. Lindahl, FEBS Lett. **67** (1976) 1. Link
1765. B. Weiss, A. Thompson, C. C. Richardson, J. Biol. Chem. **243** (1968) 4556. Link
1766. B. Weiss, A. Jacquemin-Sablon, T. R. Live, C. G. Fareed, C. C. Richardson, J. Biol. Chem. **243** (1968) 4543. Link
1767. A. J. Raae, P. K. Kleppe, K. Kleppe, Eur. J. Biochem. **60** (1975) 437. Link
1768. V. Sgaramella, J. H. van de Sande, H. G. Khorana, Proc. Natl. Acad. Sci. USA **67** (1970) 1468. Link
1769. B. M. Olivera, I. R. Lehman, J. Mol. Biol. **36** (1968) 275. Link
1770. E. L. Brown, R. Belagaje, M. J. Ryan, H. G. Khorana in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **59**, Academic Press, New York 1979, p. 109.
1771. N. P. Higgins, N. R. Cozzarelli in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **68**, Academic Press, New York 1980, p. 50.
1772. S. M. Panasenko, R. J. Alazard, I. R. Lehman, J. Biol. Chem. **253** (1978) 4590. Link
1773. P. Modrich, I. R. Lehman, J. Biol. Chem. **245** (1970) 3626. Link
1774. H. Sano, G. Feix, Biochemistry **13** (1974) 5110. Link
1775. V. Sgaramella, H. G. Khorana, Mol. Biol. **72** (1972) 493. Link
1776. R. E. Depew, T. J. Snopek, N. R. Cozzarelli, Virology **64** (1975) 144. Link
1777. J. R. Lillehaug, Eur. J. Biochem. **73** (1977) 499. Link
1778. J. R. Lillehaug, R. K. Kleppe, K. Kleppe, Biochemistry **15** (1976) 1858. Link
1779. V. Cameron, D. Soltis, O. C. Uhlenbeck, Nucleic Acids Res. **5** (1978) 825. Link
1780. V. Cameron, O. C. Uhlenbeck, Biochemistry **16** (1977) 5120. Link
1781. W. R. Folk in J.-G. Chirikjian, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 299.

1782. K. L. Berkner, W. R. Folk: “Polynucleotide Kinase Exchange as an Assay for Class II Restriction Endonucleases,” in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. 65, Academic Press, New York 1980, p. 28.
1783. D. M. Brown, A. R. Todd in E. Chargaff, J. N. Davidson (eds.): *The Nucleic Acids*, vol. **1**, Academic Press, New York, 1955, p. 409.
1784. G. Chaconas, J. H. van de Sande in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 680.
1785. H. J. Gross, G. Krupp, H. Domdey, M. Raba, P. Jank, C. Lossow, H. Alberty, K. Ramm, H. L. Sänger, Eur. J. Biochem. **121** (1981) 249. Link
1786. J. Maat, A. J. H. Smith, Nucleic Acids Res. **5** (1978) 4537. Link
1787. H. O. Smith, M. L. Birnstiel, Nucleic Acids Res. **3** (1976) 2387. Link
1788. H. J. Gross, H. Domdey, C. Lossow, P. Jank, M. Raba, H. Alberty, H. L. Sänger, Nature (London) **273** (1978) 203. Link
1789. R. Reddy, D. Henning, P. Epstein, H. Busch, Nucleic Acids Res. **9** (1981) 5645. Link
1790. D. J. Galas, A. Schmitz, Nucleic Acids Res. **5** (1978) 3157. Link
1791. L. Johnsrud, Proc. Natl. Acad. Sci. USA **75** (1978) 5314. Link
1792. J. W. Szostak, J. I. Stiles, B.-K. Tye, P. Chiu, F. Sherman, R. Wu in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **68**, Academic Press, New York 1980, p. 419.
1793. H. G. Khorana, K. L. Agarwal, H. Büchi, M. H. Caruthers, N. K. Gupta, K. Kleppe *et al.*, J. Mol. Biol. **72** (1972) 209. Link
1794. R. Silber, V. G. Malathi, J. Hurwitz, Proc. Natl. Acad. Sci. USA **69** (1972) 3009. Link
1795. K. Randerath, M. V. Reddy, R. C. Gupta, Proc. Natl. Acad. Sci. USA **78** (1981) 6126. Link
1796. R. E. Depew, N. R. Cozzarelli, J. Virol. **13** (1974) 888. Link
1797. D. A. Soltis, O. C. Uhlenbeck, J. Biol. Chem. **257** (1982) 11 340. Link
1798. K. Sirotkin, W. Cooley, J. Runnels, L. R. Snyder, J. Mol. Biol. **123** (1978) 221. Link
1799. O. C. Uhlenbeck, V. Cameron, Nucleic Acids Res. **4** (1977) 85. Link
1800. E. Ohtsuka, S. Tanaka, T. Tanaka, T. Miyake, A. F. Markham, E. Nakagawa *et al.*, Proc. Natl. Acad. Sci. USA **78** (1981) 5493. Link
1801. M. B. Mann, H. O. Smith, Nucleic Acids Res. **4** (1977) 4211. Link
1802. O. J. Yoo, P. Dwyer-Hallquist, K. L. Agarwal, Nucleic Acids Res. **10** (1982) 6511. Link
1803. M. Wigler, D. Levy, M. Perucho, Cell **24** (1981) 33. Link
1804. H. O. Smith, Science **205** (1979) 455. Link
1805. S. Jentsch, U. Günthert, T. A. Trautner, Nucleic Acids Res. **9** (1981) 2753. Link
1806. C. Waalwijk, R. A. Flavell, Nucleic Acids Res. **5** (1978) 3231. Link
1807. H. Youssoufian, C. Mulder, J. Mol. Biol. **150** (1981) 133. Link
1808. M. Ehrlich, R. Y.-H. Wang, Science **212** (1981) 1350. Link
1809. Y. Gruenbaum, T. Naveh-Many, H. Cedar, A. Razin, Nature (London) **292** (1981) 860. Link
1810. Y. Gruenbaum, R. Stein, H. Cedar, A. Razin, FEBS Lett. **124** (1981) 67. Link
1811. A. Razin, H. Cedar, Proc. Natl. Acad. Sci. USA **74** (1977) 2725. Link
1812. A. Razin, A. D. Riggs, Science **210** (1980) 604. Link
1813. D. Sutter, W. Doerfler, Proc. Natl. Acad. Sci. USA **77** (1980) 253. Link
1814. P. A. Jones, S. A. Taylor, Nucleic Acids. Res. **9** (1981) 2933. Link
1815. L. Vardimon, I. Kuhlmann, W. Doerfler, H. Cedar, Eur. J. Cell Biol. **25** (1981) 13. Link
1816. R. M. Harland, Proc. Natl. Acad. Sci. USA **70** (1982) 2323. Link
1817. T. Naveh-Many, H. Cedar, Proc. Natl. Acad. Sci. USA **78** (1981) 4246. Link
1818. H. Cedar, A. Solage, G. Glaser, A. Razin, Nucleic Acids Res. **6** (1979) 2125. Link
1819. R. Stein, Y. Gruenbaum, Y. Pollack, A. Razin, H. Cedar, Proc. Natl. Acad. Sci. USA **78** (1982) 4246. Link
1820. L. Y. Fuchs, L. Covarrubias, L. Escalante, S. Sanchez, F. Bolivar, Gene **10** (1980) 39. Link

1821. R. E. Gelinas, P. A. Myers, G. H. Weiss, R. J. Roberts, K. Murray, J. Mol. Biol, **114** (1977) 433. Link
1822. P. W. Gray, R. B. Hallick, Biochemistry **16** (1977) 1665. Link
1823. D. Brutlag, A. Kornberg, J. Biol. Chem. **247** (1972) 241. Link
1824. M. R. Deibel Jr., M. S. Coleman, J. Biol. Chem. **254** (1979) 8634. Link
1825. B. Demple, S. Linn, Nature (London) **287** (1980) 203. Link
1826. R. E. Depew, L. F. Liu, J. C. Wang, J. Biol. Chem. **253** (1978) 511. Link
1827. M. P. Deutscher, A. Kornberg, J. Biol. Chem. **244** (1969) 3019. Link
1828. J. P. Dougherty, H. Samantha, P. J. Farrelland, P. Lengyel, J. Biol. Chem. **255** (1980) 3813. Link
1829. M. Duguet, G. Yarranton, M. Gefter, Cold Spring Harbor Symp. Quant. Biol. **43** (1979) 335. Link
1830. J. J. Dunn, J. Biol. Chem. **251** (1976) 3807. Link
1831. J. Dunn in P. P. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 485.
1832. H. B. Gray Jr., T. P. Winston, J. L. Hodnett, R. J. Legerski, D. W. Nees, C.-F. Wei, D. L. Robberson in J. G. Chirikjian, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 169.
1833. P. Grippo, C. C. Richardson, J. Biol. Chem. **246** (1971) 6867. Link
1834. R. S. Gupta, T. Kasai, D. Schlessinger, J. Biol. Chem. **252** (1977) 8945. Link
1835. F. Grummt, G. Walfl, H. M. Jantzen, K. Hamprecht, U. Hübscher, C. C. Kuenzle, Proc. Natl. Acad. Sci. USA **76** (1979) 6081. Link
1836. T. J. Guiltoyle in S. T. Jacob (ed.): *Enzymes of Nucleic Acid Synthesis and Modification*, vol. **2**, CRC Press, Boca Raton 1983, p. 1.
1837. G. Herrick, B. M. Alberts, J. Biol. Chem. **251** (1976) 2124. Link
1838. I. P. Hesslewood, A. M. Holmes, W. F. Wakeling, I. R. Johnston, Eur. J. Biochem. **84** (1978) 123. Link
1839. G. R. Hartmann, Angew. Chem. **88** (1976) 197. Link
1840. W. K. Holloman, J. Biol. Chem. **248** (1973) 8114. Link
1841. W. K. Holloman, R. Holliday, J. Biol. Chem. **248** (1973) 8107. Link
1842. A. M. Holmes, I. P. Hesslewood, I. R. Johnston, Eur. J. Biochem. **62** (1976) 229. Link
1843. K. Hori, D. F. Mark, C. C. Richardson, J. Biol. Chem. **254** (1979) 11 591. Link
1844. T. M. Jovin, P. T. Englund, J. Biol. Chem. **244** (1969) 2996. Link
1845. D. L. Kacian, K. F. Watson, A. Burny, S. Spiegelmann, Biochim. Biophys. Acta **246** (1971) 265. Link
1846. A. C. Kato, K. Bartok, M. J. Fraser, D. T. Denhart, Biochim. Biophys. Acta **308** (1973) 68. Link
1847. W. S. Kelley, K. H. Stump, J. Biol. Chem. **254** (1979) 3206. Link
1848. G. H. Kidd, L. Bogorad, Proc. Natl. Acad. Sci. USA **76** (1979) 4890. Link
1849. R. P. Klett, A. Cerami, E. Reich, Proc. Natl. Acad. Sci. USA **60** (1968) 943. Link
1850. K. Knopf, Eur. J. Biochem. **98** (1979) 231. Link
1851. K.-W. Knopf, M. Yamada, A. Weissbach, Biochemistry **15** (1976) 4540. Link
1852. R. Kole, S. Altman, Biochemistry **20** (1981) 1902. Link
1853. R. Kole, S. Altman in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 469.
1854. D. Korn, P. A. Fisher, J. Battey, T. S. F. Wang, Cold Spring Harbor Symp. Quant. Biol. **43** (1979) 613. Link
1855. A. Kornberg, I. R. Lehman, M. Bessman, E. S. Simms, Biochim. Biophys. Acta **21** (1956) 197. Link
1856. T. Kornberg, M. Gefter, Biochem. Biophys. Res. Commun. **40** (1970) 1348. Link
1857. T. Kornberg, M. L. Gefter, Proc. Natl. Acad. Sci. USA **68** (1971) 761. Link
1858. S. C. Kowalczykowski, D. G. Bear, P. H. von Hippel in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 373.
1859. T. Kozu, T. Kurihara, T. Seno, J. Biochem. (Tokyo) **89** (1981) 551. Link
1860. J. S. Krakow, E. Fronk, J. Biol. Chem. **244** (1969) 5988. Link

1861. J. Krakow, M. Karstadt, Proc. Natl. Acad. Sci. USA **58** (1967) 2094. Link
1862. J. S. Krakow, K. von der Helm, Cold Spring Harbor Symp. Quant. Biol. **35** (1971) 73. Link
1863. J. Krakow, G. Rhodes, T. M. Jovin in R. Losick, M. Chamberlin (eds.): *RNA Polymerase*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1976, p. 127.
1864. W. D. Kroecker, J. L. Fairley, J. Biol. Chem. **250** (1975) 3773. Link
1865. J. P. Kuebler, D. A. Goldthwait, Biochemistry **16** (1977) 1370. Link
1866. V. T. Kung, J. C. Wang, J. Biol. Chem. **252** (1977) 5398. Link
1867. M. Kunitz, M. R. McDonald, Biochem. Prep. **3** (1953) 9. Link
1868. S. R. Kushner, H. Nagaishi, A. J. Clark, Proc. Natl. Acad. Sci. USA **71** (1974) 3593. Link
1869. P. Lamothe, B. Baril, A. Chi, L. Lee, E. Baril, Proc. Natl. Acad. Sci. USA **78** (1981) 4723. Link
1870. M. Y. W. T. Lee, C.-K Tan, A. So, K. M. Downey, Biochemistry **19** (1980) 2096. Link
1871. I. R. Lehman, M. J. Bessman, E. S. Simms, A. Kornberg, J. Biol. Chem. **233** (1958) 163. Link
1872. I. R. Lehman, C. C. Richardson, J. Biol. Chem. **239** (1964) 233. Link
1873. S. Leinbach, J. M. Reno, L. Lee, A. F. Isbell, J. A. Baezi, Biochemistry **15** (1976) 426. Link
1874. D. Levens, A. Lustig, M. Rabinowitz, J. Biol. Chem. **256** (1981) 1474. Link
1875. C. J. Levin, S. Zimmermann, J. Biol. Chem. **251** (1976) 1767. Link
1876. M. K. Lewis, R. R. Burgess in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 109.
1877. U. I. Lill, E. M. Behrendt, G. R. Hartmann, Eur. J. Biochem. **52** (1975) 411. Link
1878. S. Linn, I. R. Lehman, J. Biol. Chem. **240** (1965) 1287. Link
1879. S. Linn, I. R. Lehman, J. Biol. Chem. **240** (1965) 1294. Link
1880. W. S. Linsley, E. E. Penhoet, S. Linn, J. Biol. Chem. **252** (1977) 1235. Link
1881. J. W. Little in J. Chirikjian, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 135.
1882. L. F. Liu: “Mechanism Studies of DNA Replication and Genetic Replication,” ICN-UCLA Symp. Mol. Cell. Biol. **19** (1980) 53. Link
1883. L. F. Liu, J. C. Wang, Proc. Natl. Acad. Sci. USA **75** (1978) 2098. Link
1884. L. F. Liu, C.-C. Liu, B. M. Alberts, Nature (London) **281** (1979) 456. Link
1885. D. Livingston, C. C. Richardson, J. Biol. Chem. **250** (1975) 470. Link
1886. S. Ljungquist, J. Biol. Chem. **252** (1977) 2808. Link
1887. R. E. Lockard, L. Riesner, J. N. Vournakis in J. G. Chirikijan, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 229.
1888. V. MacKay, S. Linn, Biochim. Biophys. Acta **349** (1974) 131. Link
1889. A. Maitra, D. Novogrodsky, D. Baltimore, J. Hurwitz, Biochem. Biophys. Res. Commun. **18** (1965) 801. Link
1890. S. Martin, B. Moss, J. Biol. Chem. **250** (1975) 9330. Link
1891. S. Martin, E. Paoletti, B. Moss, J. Biol. Chem. **250** (1975) 9322. Link
1892. A. Matsukage, E. W. Bohn, S. H. Wilson, Biochemistry **14** (1975) 1006. Link
1893. A. Matsukage, M. Sivarajan, S. H. Wilson, Biochemistry **15** (1976) 5305. Link
1894. A. Matsukage, T. Tanabe, M. Yamaguchi, Y. N. Taguchi, M. Nishizawa, T. Takahashi, T. Takahashi, Biochim. Biophys. Acta **655** (1981) 269. Link
1895. N. J. McGraw, J. N. Bailey, G. R. Cleaves, D. R. Dembrinski, C. R. Gocke, L. K. Joliffe *et al.*, Nucleic Acids Res. **13** (1985) 6753. Link
1896. C. McHenry, W. Crow, J. Biol. Chem. **254** (1979) 1748. Link
1897. C. McHenry, A. Kornberg, J. Biol. Chem. **254** (1977) 6478. Link
1898. M. Mechali, J. Abadiedebat, A.-M. de Recondo, J. Biol. Chem. **255** (1980) 2114. Link
1899. C. Milcarek, B. Weiss, J. Mol. Biol. **68** (1972) 303. Link
1900. R. V. Miller, A. J. Clark, J. Bacteriol. **127** (1976) 794. Link

1901. J. P. Miller, G. R. Phillip, J. Biol. Chem. **246** (1971) 1274. Link
1902. H. I. Miller, A. D. Riggs, G. N. Gill, J. Biol. Chem. **248** (1973) 2621. Link
1903. M. A. Minks, S. Benvin, P. A. Maroney, C. Baglioni, J. Biol. Chem. **254** (1979) 5058. Link
1904. K. Mizuuchi, D. H. O'Dea, M. Gellert, Proc. Natl. Acad. Sci. USA **75** (1978) 5960. Link
1905. P. Modrich, Y. Anraku, I. R. Lehman, J. Biol. Chem. **248** (1973) 7495. Link
1906. K. Mölling, J. Virol. **18** (1976) 418. Link
1907. K. Mölling, D. P. Bolognesi, W. B. Auer, W. Büsen, H. W. Plassmann, P. Hausen, Nature (London) **234** (1971) 240. Link
1908. E. Mössner, Z. Physiol. Chem. **361** (1980) 543. Link
1909. C. F. Morris, H. Hama-Inaba, D. Mace, N. K. Senha, B. Alberts, J. Biol. Chem. **254** (1979) 6787. Link
1910. M. I. Moseman-McCoy, T. H. Lubben, R. I. Gumport, Biochim. Biophys. Acta **562** (1979) 149. Link
1911. R. E. Moses, C. C. Richardson, Biochem. Biophys. Res. Commun. **41** (1970) 1557. Link
1912. R. E. Moses, C. C. Richardson, Biochem. Biophys. Res. Commun. **41** (1970) 1565. Link
1913. K. M. T. Muskavitch, S. Linn in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 233.
1914. K. Nath, J. Hurwitz, J. Biol. Chem. **249** (1974) 3680. Link
1915. J. Nevins, W. Joklik, J. Biol. Chem. **252** (1977) 6939. Link
1916. N. G. Nossal, M. S. Hershfield, J. Biol. Chem. **246** (1971) 5414. Link
1917. N. G. Nossal, M. F. Singer, J. Biol. Chem. **243** (1968) 913. Link
1918. M. Novogrodsky, A. Tal, A. Traub, J. Hurwitz, J. Biol. Chem. **241** (1966) 2933. Link
1919. C. W. M. Orr, S. T. Herriott, M. J. Bessman, J. Biol. Chem. **240** (1965) 4625. Link
1920. Yu. A. Ovchinnikov, V. M. Lipkin, N. N. Modyanov, O. Yu. Chertov, Yu. V. Smirnov, FEBS Lett. **76** (1977) 108. Link
1921. Yu. A. Ovchinnikov, G. S. Monasturskaya, V. V. Gubanov, S. O. Guryev, O. Yu. Chertov, N. N. Modyanov *et al.*, Eur. J. Biochem. **116** (1981) 621. Link
1922. Yu. A. Ovchinnikov, G. S. Monasturskaya, V. V. Gubanov, S. O. Guryev, I. S. Salomatina, T. M. Shuvaeva *et al.*, Dokl. Akad. Nauk SSSR **261** (1981) 763. Link
1923. A. Pesce, C. Casoli, G. Schito, Nature (London) **262** (1976) 412. Link
1924. B. H. Pheiffer, S. H. Zimmermann, Biochemistry **18** (1979) 2960. Link
1925. J. Pierre, J. Laval, Biochemistry **19** (1980) 5018. Link
1926. C. Portier, Eur. J. Biochem. **55** (1975) 573. Link
1927. A. Quint, H. Cedar, Nucleic Acids Res. **9** (1981) 633. Link
1928. C. M. Radding in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **21**, Academic Press, New York 1971, p. 273.
1929. M. Radman, J. Biol. Chem. **251** (1976) 1438. Link
1930. R. K. Ray, R. Reuben, I. Molineux, M. Gefter, J. Biol. Chem. **249** (1974) 5379. Link
1931. M. V. Reddy, R. C. Gupta, K. Randerath, Anal. Biochem. **117** (1981) 271. Link
1932. T. W. Reid, I. B. Wilson in P. D. Boyer (ed.): *The Enzymes*, vol. **4**, Academic Press, New York 1971, p. 373.
1933. R. C. Reuben, M. L. Gefter, Proc. Natl. Acad. Sci. USA **70** (1973) 1846. Link
1934. G. M. Richards, D. J. Tutas, W. J. Wechter, M. Laskowski, Biochemistry **6** (1967) 2908. Link
1935. C. C. Richardson, Proc. Natl. Acad. Sci. USA **54** (1965) 158. Link
1936. C. C. Richardson, J. Biol. Chem. **241** (1966) 2084. Link
1937. C. C. Richardson in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 299.
1938. C. C. Richardson, A. Kornberg, J. Biol. Chem. **244** (1964) 2996. Link
1939. C. C. Richardson, L. J. Romano, R. Kolodner, J. E. Leclerq, F. Tamanoni, M. J. Enlges *et al.*, Cold Spring Harbor Symp. Quant. Biol. **43** (1978) 449. Link

1940. M. Robert-Guroff, A. W. Schrecker, B. J. Brinkmann, R. C. Gallo, Biochemistry **16** (1977) 2866. Link
1941. H. D. Robertson, R. E. Webster, N. D. Zinder, J. Biol. Chem. **243** (1968) 82. Link
1942. H. Samanta, J. P. Dougherty, P. Lengyel, J. Biol. Chem. **255** (1980) 9807. Link
1943. P. Schedl, J. Roberts, P. Primakoff, Cell **8** (1976) 581. Link
1944. J. F. Scott, A. Kornberg, J. Biol. Chem. **253** (1978) 3292. Link
1945. G. Seal, L. A. Loeb, J. Biol. Chem. **251** (1976) 975. Link
1946. A. Sentenac, J. M. Buhler, A. Ruet, J. Huet, F. Iborra, P. Fromageot in B. F. C. Clark, H. Klenow, J. Zeuthen (eds.): *Gene Expression. Protein Synthesis and Control. RNA Synthesis and Control. Chromatin Structure and Function*, Pergamon Press, New York 1977, p. 187.
1947. P. Setlow, D. Brutlag, A. Kornberg, J. Biol. Chem. **247** (1972) 224. Link
1948. V. Sgaramella, Proc. Natl. Acad. Sci. USA **69** (1972) 3389. Link
1949. M. F. Shemyakin, A. A. Grepachevsky, A. V. Chestukhin, Eur. J. Biochem. **98** (1979) 417. Link
1950. V. Shen, D. Schlessinger in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 501.
1951. T. E. Shenk, C. Rhodes, P. W. J. Rigby, P. Berg, Proc. Natl. Acad. Sci USA **72** (1975) 989. Link
1952. Y. Shimura, H. Sakano, F. Nagawa, Eur. J. Biochem. **86** (1978) 267. Link
1953. J. Shinshi, M. Miwa, K. Kato, M. Noguchi, T. Matsushima, T. Sugimura, Biochemistry **15** (1976) 2185. Link
1954. R. G. Shorenstein, R. Losick, J. Biol. Chem. **248** (1973) 6163. Link
1955. S. Shuman, J. Hurwitz, J. Biol. Chem. **254** (1979) 10 396. Link
1956. S. Shuman, J. Hurwitz in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 245.
1957. S. Shuman, M. Surks, H. Furneaux, J. Hurwitz, J. Biol. Chem. **255** (1980) 11 588. Link
1958. D. Smith, R. Ratcliff, D. Williams, A. Martinez, J. Biol. Chem. **242** (1967) 590. Link
1959. H. J. Smith, L. Bogorad, Proc. Natl. Acad. Sci. USA **71** (1974) 4839. Link
1960. D. Smoler, I. Molineux, D. Baltimore, J. Biol. Chem. **246** (1971) 7697. Link
1961. S. Spadari, A. Weissbach, J. Biol. Chem. **249** (1974) 5809. Link
1962. P. F. Spahr, B. R. Hollingworth, J. Biol. Chem. **236** (1961) 823. Link
1963. K. S. Sriprakash, N. Lundh, M. M. Huh, C. M. Radding, J. Biol. Chem. **250** (1975) 5438. Link
1964. A. Srivastava, M. J. Modak, J. Biol. Chem. **255** (1980) 2000. Link
1965. D. M. Stalker, D. W. Mosbaugh, R. R. Meyer, Biochemistry **15** (1976) 3114. Link
1966. B. C. Stark, R. Kole, E. J. Bowman, S. Altman, Proc. Natl. Acad. Sci. USA **75** (1978) 3711. Link
1967. H. Sternbach, F. von der Haar, E. Schlimme, E. Gaertner, F. Cramer, Eur. J. Biochem. **22** (1971) 166. Link
1968. H. Sternbach, M. Srpinzl, J. B. Hobbs, F. Cramer, Eur. J. Biochem. **67** (1976) 215. Link
1969. G. L. Stetler, G. J. King, W. M. Huang, Proc. Natl. Acad. Sci. USA **76** (1979) 3737. Link
1970. K. O. Stetter, W. Zillig, Eur. J. Biochem. **48** (1974) 527. Link
1971. A. Sugino, H. M. Goodman, H. L. Heyneker, J. Shine, H. W. Boyer, N. R. Cozarelli, J. Biol. Chem. **249** (1977) 3680. Link
1972. A. Sugino, T. J. Snopek, N. Cozarelli, J. Biol. Chem. **252** (1977) 1732. Link
1973. K. S. Szeto, D. Soell, Nucleic Acids Res. **1** (1974) 171. Link
1974. H. Tanuchi, C. B. Anfinsen, A. Sodja, J. Biol. Chem. **242** (1967) 4752. Link
1975. C. Tsiapalis, J. Dorson, F. Bollum, Eur. J. Biochem. **250** (1975) 4486. Link
1976. O. C. Uhlenbeck, R. I. Gumpori in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 31.
1977. P. Valenzuela, G. I. Bell, F. Weinberg, W. Rutter in G. Stein, J. Stein, L. J. Kleinsmith (eds.): *Methods in Cell Biology*, vol. **19**, Academic Press, New York 1978, p. 1.
1978. I. M. Verma, J. Virol. **15** (1975) 843. Link

1979. G. C. Walker, O. C. Uhlenbeck, E. Bedows, R. I. Gumport, Proc. Natl. Acad. Sci. USA **72** (1975) 122. Link
1980. J. C. Wang, J. Mol. Biol. **55** (1971) 523. Link
1981. T. S.-F. Wang, W. D. Sedwick, D. Korn, J. Biol. Chem. **249** (1974) 841. Link
1982. B. Weiss, J. Biol. Chem. **251** (1976) 1896. Link
1983. B. Weiss in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 203.
1984. R. B. Wickner, B. Ginsberg, I. Berkower, J. Hurwitz, J. Biol. Chem. **247** (1972) 489. Link
1985. S. Wickner, Proc. Natl. Acad. Sci. USA **73** (1976) 3511. Link
1986. W. Wickner, A. Kornberg, Proc. Natl. Acad. Sci. USA **70** (1973) 3679. Link
1987. K. W. Wilcox, M. Orlosky, E. A. Friedman, H. O. Smith, Fed. Proc. Fed. Am. Soc. Exp. Biol. **34** (1975) 515. Link
1988. F. R. Williams, T. Godefroy, E. Mery, J. Yon, M. Grunberg- Manago, Biochem. Biophys. Res. Commun. **19** (1965) 25. Link
1989. M. Winters, M. Edmonds, J. Biol. Chem. **248** (1973) 4756. Link
1990. M. Wright, G. Buttin, J. Hurwitz, J. Biol. Chem. **246** (1971) 6543. Link
1991. D. M. Yaiko, M. C. Valentine, B. Weiss, J. Mol. Biol. **85** (1974) 323. Link
1992. M. Yamaguchi, A. Matsukage, T. Takahashi, J. Biol. Chem. **255** (1980) 7002. Link
1993. M. Yamaguchi, K. Tanabe, Y. N. Taguchi, M. Nishizawa, T. Takahashi, A. Matsukage, J. Biol. Chem. **255** (1980) 9942. Link
1994. J. Yarranton, R. H. Das, M. L. Gefter, J. Biol. Chem. **254** (1979) 12002 Link
1995. J. Yarranton, M. L. Gefter, Proc. Natl. Acad. Sci. USA **76** (1979) 1658. Link
1996. W. Zillig, P. Palm, A. Heil in R. Losick, M. Chamberlin (eds.): *RNA Polymerase*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1976, p. 101.
1997. J. J. Jendrisak in C. Leaver (ed.): *Genome Organization and Expression in Plants*, Plenum Press, New York 1980, p. 93.
1998. J. J. Jendrisak, Plant Physiol. **67** (1981) 438. Link
1999. S. Adler, P. Modrich, J. Biol. Chem. **254** (1979) 11 605. Link
2000. B. M. Alberts, L. Frey, H. Delius, J. Mol. Biol. **68** (1972) 139. Link
2001. J. M. D'Alessio, P. J. Perna, M. R. Paule, J. Biol. Chem. **254** (1979) 11 282. Link
2002. J. M. D'Alessio, S. R. Spindler, M. R. Paule, J. Biol. Chem. **253** (1978) 4669. Link
2003. T. Ando, Biochim. Biophys. Acta **114** (1966) 158. Link
2004. G. E. Austin, D. Sirakoff, B. Roop, G. H. Moyer, Biochim. Biophys, Acta **522** (1978) 412. Link
2005. J. Avila, J. M. Hermoso, E. Vinuela, M. Salas, Eur. J. Biochem. **21** (1971) 526. Link
2006. L. A. Ball, Virology **94** (1979) 282. Link
2007. A. L. Ball: "2′,5′-Oligoadenylate Synthetase," in P. D. Boyer (ed.): *The Enzymes*, vol. 15, Academic Press, New York 1982, p. 281.
2008. G. R. Banks, J. A. Boezi, I. R. Lehman, J. Biol. Chem. **254** (1979) 9886. Link
2009. E. K. Barbehenn, J. E. Craine, A. Chrambach, C. B. Klee, J. Biol. Chem. **257** (1982) 1007. Link
2010. H. Beier, R. Hausmann, Nature (London) **251** (1974) 538. Link
2011. I. K. Bendis, L. Shapiro, J. Bacteriol. **115** (1973) 848. Link
2012. M. Bittner, R. L. Burke, B. M. Alberts, J. Biol. Chem. **254** (1979) 9565. Link
2013. P. Blackburn, S. Moore: "Pancreatic Ribonuclease," in P. O. Boyer (ed.): *The Enzymes*, vol. 15, Academic Press, New York 1982, p. 317.
2014. T. Inove, F. Kada, J. Biol. Chem. **253** (1978) 8559. Link
2015. S. E. Jorgensen, J. F. Koerner, J. Biol. Chem. **241** (1966) 3090. Link
2016. A. Bernardi, G. Bernardi: "Spleen Acid Exonuclease," in P. D. Boyer (ed.): *The Enzymes*, vol. 4, Academic Press, New York 1971, p. 329.
2017. A. Bolden, J. Aucker, A. Weissbach, J. Virol. **16** (1975) 1584. Link

2018. Z. Burton, R. R. Burgess, J. Lin, D. Moore, S. Holder, C. A. Gross, Nucleic Acids Res. **9** (1981) 2889. Link
2019. V. Cameron, D. Soltis, O. C. Uhlenbeck, Nucleic Acids Res. **5** (1978) 825. Link
2020. D. S. Carre, F. Chapeville, Biochim. Biophys. Acta **361** (1974) 176. Link
2021. J. J. Byrnes, V. L. Black, Biochemistry **17** (1978) 4226. Link
2022. J. J. Byrnes, K. M Downey, V. L. Black, A. G. So, Biochemistry **15** (1976) 2817. Link
2023. M. D. Challberg, P. T. Englund, J. Biol. Chem. **254** (1979) 7812. Link
2024. M. D. Challberg, P. T. Englund, J. Biol. Chem. **254** (1979) 7820. Link
2025. M. Chamberlin in R. Losick, M. Chamberlin (eds.): *RNA Polymerase*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1976, p. 17.
2026. M. Chamberlin in P. D. Boyer (ed.): *The Enzymes*, vol. **15**, Academic Press, New York 1982, p. 61.
2027. L. M. S. Chang, J. Biol. Chem. **248** (1973) 3789. Link
2028. L. M. S. Chang, J. Biol. Chem. **248** (1973) 6983. Link
2029. L. M. S. Chang, F. J. Bollum, J. Biol. Chem. **256** (1981) 494. Link
2030. J. W. Chase, L. D. Vales in J. G. Chirikjian, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 147.
2031. J. W. Chase, R. F. Whittier, J. Auerbach, A. Sancar, W. D. Rupp, Nucleic Acids Res. **8** (1980) 3215. Link
2032. Y.-C. Chen, E. W. Bohn, S. R. Blanck, S. H. Wilson, J. Biol. Chem. **254** (1979) 11 678. Link
2033. A. V. Chestukhin, M. F. Shemyakin, N. A. Kalinina, A. A. Prozorov, FEBS Lett. **24** (1972) 121. Link
2034. S. Clark, J. Virol. **25** (1978) 224. Link
2035. J. E. Clements, S. G. Rogers, B. Weiss, J. Biol. Chem. **253** (1978) 2990. Link
2036. N. R. Cozzarelli, R. B. Kelly, A. Kornberg, J. Mol. Biol. **45** (1969) 513. Link
2037. J. W. Cranston, R. Silber, V. G. Malathi, J. Hurwitz, J. Biol. Chem. **249** (1974) 7447. Link
2038. R. J. Crouch, J. Biol. Chem. **249** (1974) 1314. Link
2039. R. J. Crouch in J. G. Chirikjian, T. S. Papas (eds.): *Gene Amplification and Analysis*, vol. **2**, Elsevier, Amsterdam 1981, p. 217.
2040. H. Cudny, M. P. Deutscher, Proc. Natl. Acad. Sci. USA **77** (1980) 837. Link
2041. J. L. Darlix, Eur. J. Biochem. **51** (1975) 369. Link
2042. S. K. Das, R. K. Fujimura, J. Biol. Chem. **252** (1977) 8700. Link
2043. F. Egami, T. Oshima, T. Uchida in F. Chapeville, A. L. Haenni (eds.): *Molecular Biology, Biochemistry and Biophysics*, vol. **32**, Springer Verlag, Berlin 1980, p. 250.
2044. B. Eriksson, A. Larsson, E. Helgstrand, N. G. Johansson, B. Öberg, Biochim. Biophys. Acta **607** (1980) 53. Link
2045. G. C. Fareed, E. M. Wilt, C. C. Richardson, J. Biol. Chem. **246** (1971) 925. Link
2046. O. Fichot, M. Pascal, M. Mechali, A. M. de Recondo, Biochim. Biophys. Acta **561** (1979) 29. Link
2047. P. A. Fisher, D. Korn, J. Biol. Chem. **252** (1977) 6528. Link
2048. M. J. Fraser in L. Grosman, K. Moldave (eds.): *Methods in Enzymology*, vol. **65**, Academic Press, New York 1980, p. 255.
2049. M. J. Fraser, R. Tjeerde, K. Matsumoto, Can. J. Biochem. **54** (1976) 971. Link
2050. G. D. Frenkel, C. C. Richardson, J. Biol. Chem. **246** (1971) 4839. Link
2051. E. C. Friedberg, C. T. M. Anderson, T. Bonura, R. Cone, E. H. Radany, R. J. Reynolds, Prog. Nucleic Acid Res. Mol. Biol. **26** (1981) 197. Link
2052. E. C. Friedberg, T. Bonura, E. H. Radany, J. D. Love in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 251.
2053. M. Fujimoto, A. Kuninaka, H. Yoshino, Agric. Biol. Chem. **38** (1974) 777. Link
2054. M. Fujimoto, A. Kuninaka, H. Yoshino, Agric. Biol. Chem. **38** (1974) 785. Link
2055. R. K. Fujimara, B. C. Roop, J. Biol. Chem. **251** (1976) 2168. Link
2056. Y. M. Gafurov, L. L. Terentev, V. A. Rasskazov, Biochemistry (Engl. Transl.) **44** (1979) 996. Link

2057. F. T. Gates III, S. Linn, J. Biol. Chem. **252** (1977) 1647. Link
2058. T. F. Gates III, S. Linn, J. Biol. Chem. **252** (1977) 2802. Link
2059. M. L. Gefter, I. J. Molineux, T. Kornberg, H. G. Khorana, J. Biol. Chem. **247** (1972) 3321. Link
2060. M. Gellert in P. D. Boyer (ed.): *The Enzymes*, vol. **14**, Academic Press, New York 1982, p. 345.
2061. G. F. Gerard, D. P. Grandgenett, J. Virol. **15** (1975) 785. Link
2062. G. F. Gerard, D. P. Grandgenett in J. R. Stephenson (ed.): *Molecular Biology of RNA Tumor Viruses*, Academic Press, New York 1980, p. 345.
2063. R. K. Ghosh, M. P. Deutscher in D. Söll, J. N. Abelson, P. R. Schimmel (eds.): *Transfer RNA: Biological Aspects*, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 1980, p. 59
2064. J. R. Gillen, A. E. Karu, H. Nagaishi, A. J. Clark, J. Mol. Biol. **133** (1977) 27. Link
2065. P. J. Goldmark, S. Linn, J. Biol. Chem. **247** (1972) 1849. Link
2066. S. Gottesman, M. Gottesman, Proc. Natl. Acad. Sci. USA **72** (1975) 2188. Link
2067. D. P. Grandgenett, A. C. Vora, R. D. Schiff, Virolog. **89** (1978) 119. Link
2068. H. B. Gray Jr., D. A. Ostrander, J. L. Hodnett, R. J. Legerski, D. L. Robberson, Nucleic Acids Res. **2** (1975) 1459. Link
2069. S. K. Das, R. K. Fujimura, J. Biol. Chem. **254** (1979) 1227. Link

Specific References

2070. The Soap and Detergent Industry Association (SDIA), *Work Practices for Handling Enzymes in the Detergent Industry*, New York, 1995.
2071. G. Peters, G. Johnson, A. Golembiewski: “Safe Use of Enzymes in the Workplace”, Appl. Occup. Environm. Hyg. **16** (2001) 389 – 396. Link
2072. G. Peters, D. P. MacKenzie: “Worker Safety: How to Establish Site Enzyme Capability”, in J. van Ee (ed.): *Enzymes in Detergency*, Marcel Dekker, New York, 1997 pp. 327 – 340.
2073. Association Internationale de la Savonnerie et de la Detergence (AISE), “*Guidelines for the Safe Handling of Enzymes in Detergent manufacturing*”, Brussels, 2001.
2074. M. W. Pariza and E. A. Johnson, *Evaluating the Safety of Microbial Enzyme Preparations Used in Food Processing: Update for a New Century, Regulatory Toxicology and Pharmacology*, **33**: 1 – 14, 2001.
2075. D. A. Kessler, M. R. Taylor, J. H. Maryanski, E. L. Flamm, L. S. Kahl, “The Safety of Foods Developed by Biotechnology”, Science **256** (1992) 1747 – 1749, 1832. Link
2076. SCF, Reports of the Scientific Committee for Food (Twenty-Seventh Series), Ref. No EUR 14181, EN Guidelines for the Presentation of Data on Food Enzymes, 1992, pp. 13 – 22.
2077. IFBC (International Food Biotechnology Council), Chapter 4: Safety Evaluation of Foods and Food Ingredients Derived from Microorganisms in Biotechnologies and Food: Assuring the Safety of Foods Produced by Genetic Modification, Reg. Toxicol. Pharmacol. **12** (1990) S1 – S196. Link
2078. M. W. Pariza, E. M. Foster, “Determining the Safety of Enzymes Used in Food Processing”, J. Food Protect. **46** (1983) 453 – 463. Link
2079. Committee on Food Chemicals Codex, Food and Nutrition Board, *Institute of Medicine, National Academy of Sciences, Food Chemicals Codex*, 4th ed., National Academy Press, Washington, D.C., 1996.
2080. Joint FAO/WHO Expert Committee on Food Additives (JECFA), *Compendium of Food Additive Specifications Volume 1, General Specifications for Enzyme Preparations Used in Food Processing*, 1 – 3; `http://www.fao.org/es/esn/jecfa/jecfa.htm`, 1992.

PREVIOUS NEXT

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Enzymes

Ullmann's Encyclopedia of Industrial Chemistry
Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_341.pub2
Article Online Posting Date: October 15, 2003

PREVIOUS

Abbreviations used in this article

A: adenosine

ACA: acetamidocinnamic acid

ACL: α-amino- ε-caprolactam

ADH: alcohol dehydrogenase

ADI: acceptable daily intake

ADP: adenosine 5′-diphosphate

Ala: alanine

Arg: Arginine

AMP: adenosine 5′-monophosphate

ATC: D,L-2-amino- Δ^2-thiazoline-4-carboxylic acid

ATP: adenosine 5′-triphosphate

C: cytidine

cDNA: copy DNA

CL: citrate lyase

CMP: cytidine 5′-monophosphate

CoA: coenzyme A

CS: citrate synthetase

CTP: cytidine 5′-triphosphate

d: deoxy

***dam*:** gene locus for E. coli DNA adenine methylase (N^6-methyladenine)

***dcmI*:** gene locus for E. coli DNA cytosine methylase(5-methylcytosine)

dd: dideoxy

ddNTP: dideoxynucleoside 5′-triphosphate

DE: dextrose equivalent

DEAE: diethylaminoethyl

DNA: deoxyribonucleic acid

DNase: deoxyribonuclease

dNTP: deoxynucleoside 5′-triphosphate

DOPA: 3-(3,4-dihydroxyphenylalanine) [3-hydroxy-L-tyrosine]

dpm: decays per minute

ds: double-stranded

E.C.: Enzyme Commission

F6P: fructose 6-phosphate

FAN: free alpha amino nitrogen, i.e., a measure of peptides/amino acids available for yeast to be used as nutrient

fMet: *N*-formylmethionine

FMN: flavin mononucleotide

$FMNH_2$: flavin mononucleotide, reduced

G: quanosine

GDP: guanosine 5′-diphosphate

Glu: glutamic acid

Gly: glycine

GMP: guanosine 5′-monophosphate

GOD: glucose oxidase

GOT: glutamate – oxaloacetate transaminase

G6P: glucose 6-phosphate

GPT: glutamate – pyruvate transaminase

GTP: guanosine 5′-triphosphate

3-HBDH: 3-hydroxybutyrate dehydrogenase

HFCS: high-fructose corn syrup

***hsd*M:** *E. coli* gene locus for methylation

***hsd*R:** *E. coli* gene locus for restriction

***hsd*S:** *E. coli* gene locus for sequence specificity

IDP: inosine 5′-diphosphate

Ile: isoleucine

INT: iodonitrotetrazolium chloride

ITP: inosine 5′-triphosphate

LDH: lactate dehydrogenase

Lys: lysine

^{m}(superscript): methylated

MDH: malate dehydrogenase

Met: methionine

M6P: mannose 6-phosphate

mRNA: messenger RNA

MTT: 3-(4,5-dimethylthiazolyl-2)-2,5-diphenyltetrazolium bromide

N: any nucleotide

NAD: nicotinamide – adenine dinucleotide

NADH: nicotinamide – adenine dinucleotide, reduced

NADP: nicotinamide – adenine dinucleotide phosphate

NADPH: nicotinamide – adenine dinucleotide phosphate, reduced

NMN: nicotinamide mononucleotide

NTP: nucleoside 5′-triphosphate

p: phosphate groups

^{32}P: phosphate groups containing ^{32}P phosphorus atoms

p_i: inorganic phosphate

°P: degree Plato; i.e., sugar content equivalent to 1 % sucrose by weight

PEP: phosphoenolpyruvate

6-PGDH: 6-phosphogluconate dehydrogenase

Phe: phenylalanine

PMS: 5-methylphenazinium methyl sulfate

poly(dA): poly(deoxyadenosine 5′-monophosphate)

pp_i: inorganic pyrophosphate

Pro: proline

PRPP: phosphoribosyl pyrophosphate

Pu: purine

Py: pyrimidine

r: ribo

RNA: ribonucleic acid

RNase: ribonuclease

SAM: S-adenosylmethionine

SMHT: serine hydroxymethyltransferase

ss: single-stranded

T: thymidine

TMP: thymidine 5′-monophosphate

tRNA: transfer RNA

TTP: thymidine 5′-triphosphate

U: uridine

UMP: uridine 5′-monophosphate

UTP: uridine 5′-triphosphate

Val: valine

: Plasmids:

pBR322:

pBR328:

pSM1:

pSP64:

pSP65:

pSPT18, pSPT19:

pT7–1, pT7–2:

pUC 18, pUC 19:

pUR222:

: Bacteriophages:

fd:

ghl:

M13:

N4:

PBS1:

PBS2:

SPO1:

SP6:

SP15:

T3:

T4:

T5:

T7:

XP12:

λ:

λgt11:

ΦSM11:

ΦX174:

: Eukaryotic viruses:

Ad2:

SV40:

PREVIOUS

© 2004 by Wiley-VCH Verlag GmbH & Co. KGaA All rights reserved.

Epoxides

Ethylene Oxide and Propylene Oxide are separate keywords.

Guenter Sienel, Peroxid-Chemie GmbH, Höllriegelskreuth, Federal Republic of Germany (Chaps. 2, 3, 4.2 – 7)

Robert Rieth, Peroxid-Chemie GmbH, Höllriegelskreuth, Federal Republic of Germany (Chaps. 2, 3, 4.2 – 7)

Kenneth T. Rowbottom, Laporte Industries Ltd., Widnes, Cheshire WA8 OJU, United Kingdom (Section 4.1)

1. Introduction 1
2. Reactions of Epoxides 1
2.1. Reactions with Compounds Containing Ionizable Hydrogen 1
2.2. Reactions with Nitrogen or Phosphorus Compounds 4
2.3. Rearrangements 4
2.4. Other Reactions 4
3. Production of Epoxides 4
3.1. Epoxidation with Percarboxylic Acids 4
3.2. Epoxidation with Hydrogen Peroxide 6
3.3. Epoxidation with Hydroperoxides . . 7
3.4. Epoxidation with Halohydrins 8
3.5. Epoxidation with Oxygen 8
3.6. Epoxidation by Other Methods 8
4. Industrially Important Epoxides . . . 9
4.1. Epichlorohydrin 9
4.2. Other Important Monoepoxides 10
4.3. Functional Epoxides 11
4.4. Diepoxides 12
4.5. Polyepoxides 12
5. Analysis 13
6. Economic Aspects 13
7. Toxicology 13
8. References 14

1. Introduction

Epoxides, also known as oxiranes, are characterized by the following group:

The epoxy group is a highly reactive moiety (see Chap. 2), which makes epoxides an important group of industrial organic intermediates. The most significant members of this group, ethylene oxide and propylene oxide, are treated in separate articles. Epichlorohydrin (Section 4.1) dominates among the raw materials for epoxy resins (→ Epoxy Resins).

2. Reactions of Epoxides

Polarity and ring strain make the oxirane ring highly reactive. Thus, epoxides participate in numerous reactions, which makes these compounds useful building blocks in organic synthesis (see Table 1) [1–6]. Often epoxides formed in an initial step react further to provide industrially important products, such as surfactants or detergents (tensides), antistatic- or corrosion-protection agents, additives to laundry detergents, lubricating oils, textiles, and cosmetics [1].

2.1. Reactions with Compounds Containing Ionizable Hydrogen

Reactions of epoxides with oxygen, sulfur, or carbon anions, usually in the presence of either acid or alkaline catalysts, affords β-hydroxy compounds (see Table 1). Unsymmetrically substituted epoxides may yield two isomers the ratio of which is controlled by pH [7].

The *base-catalyzed reaction* follows an S_N2 substitution mechanism; attack of the nucleophile X^- occurs predominantly at the sterically less hindered and more electron-deficient carbon atom. This substitution leads to Walden inversion at this carbon atom. The major product obtained from **1** is isomer **2**:

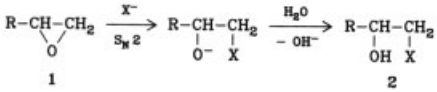

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_531

Table 1. Reactions of epoxides

Reaction number	Reagent*	Product	Name
1)	H_2O		vicinal diol
2)	ROH		β-hydroxyether
3)	RCOOH	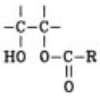	β-hydroxyester
4)	HX		halohydrin
5)	H_2S		β-hydroxythiol
6)	SO_3^{2-}		β-hydroxysulfonate
7)	$H_2N-C(=S)-NH_2$ or RSCN		episulfide
8)	HCN		β-hydroxynitrile
9)	$R-CH_2-CO_2Et$		lactone
10)	NH_3		monoalkanolamine
11)	NR_3/HX	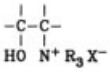	betaine
12)	PR_3		olefin
13)	$RP(=O)(OR')(OH)$		phosphonic acid diester
14)			aldehyde or ketone

Table 1. Continued

15)	(Heat)	$>C=C-CH_2OH$	allyl alcohol
16)	H_2/Cat.	$-C(HO)-C(H)-$	alcohol
17)	Oxidizing agent	$-C(OH)-COOH$	α-hydroxycarboxylic acid
18)	$CH_3\overset{O}{\overset{\|\|}{S}}CH_3$	$-C(OH)-C(=O)-$	α-hydroxyketone
19)	Ph-NCO		*N*-phenyloxazolidone
20)	CO_2		carbonate
21)	$R-\overset{O}{\overset{\|\|}{C}}-R'$		1,3-dioxolane
22)	$n(-C(O)-C-)$	$-C-C-O-(C-C-O)_{n-2}-C-C-O-$	polymers

* Reagent reacts with epoxide $-C(O)-C-$

The *acid-catalyzed reaction* proceeds via an intermediate oxonium ion (**3**), which may yield either the same isomer (**2**) obtained by the base-catalyzed reaction via the S_N2 substitution mechanism or isomer **4** via the more stable secondary carbocation intermediate by the S_N1 substitution mechanism with retention of configuration.

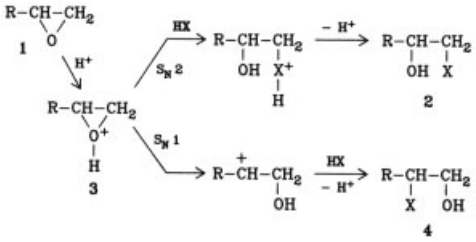

Thus, acid-catalyzed addition generally yields an isomeric mixture.

Hydrolysis of epoxides in the presence of acid or alkali leads to the corresponding vicinal diols (Table 1, no. 1). Monofunctional alcohols afford β-hydroxyethers (Table 1, no. 2), which may be further reacted to yield polyoxyethylated compounds. β-Hydroxyesters are formed with monocarboxylic acids (Table 1, no. 3). In the presence of basic aluminum oxide, these ring-opening reactions frequently proceed under mild conditions with a good yield [3], [8]. Halohydrins are obtained by reaction with hydrogen halides (Table 1, no. 4). This equilibrium reaction usually proceeds rapidly and smoothly. Base-catalyzed reaction with hydrogen sulfide yields β-hydroxythiols (Table 1, no. 5), which readily react further to bis(β-hydroxyalkyl) sulfides. Sulfites yield β-hydroxysulfonates (Table 1, no. 6). Episulfides (Table 1, no. 7) can be obtained readily with thioureas [9] or thio-

cyanates [10]. The reactions with hydrocyanic acid to provide β-hydroxynitriles (Table 1, no. 8) and with malonic acid derivatives or acetoacetic ester to form isomeric five-membered lactones (Table 1, no. 9) [11] are examples of the reaction of epoxides with carbanion species.

2.2. Reactions with Nitrogen or Phosphorus Compounds

Ammonia reacts with epoxides (Table 1, no. 10) to form mono-, di-, or trialkanolamines, depending on the molar ratio. Tertiary amines (R_3N) and tertiary phosphines (R_3P) are sufficiently nucleophilic to open the oxirane ring with initial formation of a betaine. Betaines of tertiary amines can be captured under acidic conditions (Table 1, no. 11), whereas the reaction of tertiary phosphines generates phosphine oxide and leads to olefins (Table 1, no. 12). Reaction with phosphonic acid or its esters leads to the formation of cyclic phosphonic acid diesters (Table 1, no. 13) [12].

2.3. Rearrangements

Epoxides are capable of undergoing intramolecular rearrangement to form carbonyl compounds (Table 1, no. 14) or allyl alcohols (Table 1, no. 15). The rearrangement resulting in carbonyl compounds is generally initiated by Lewis acids (e.g., $ZnCl_2$, $SnCl_2$, $AlCl_3$, $FeCl_3$, and $TiCl_4$) [13] or Brønsted acids (e.g., H_2SO_4, HCOOH, 4-toluenesulfonic acid, HF, HCl, and HI) [14], [15]. Epoxides on cyclic systems may lead to ring enlargement or contraction. Epoxides with a sufficiently acidic hydrogen atom at the oxirane ring may rearrange to carbonyl compounds in the presence of strong bases (e.g., lithium diethylamide, LiBr, LiI, and $LiClO_4$) or transition-metal complexes (e.g., $Mo(CO)_6$ [16] and $Co_2(CO)_8$ [17]). Rearrangement in the presence of strong nonnucleophilic bases (e.g., *n*-butyl lithium and lithium dialkylamide) may also lead to allyl alcohols [18].

2.4. Other Reactions

Reduction of epoxides to alcohols can be accomplished by using a variety of reducing agents. In addition to catalytic hydrogenation (Table 1, no. 16), which is frequently accompanied by deoxygenation of the epoxide, hydrogenolysis with complex metal hydrides (e.g., $LiAlH_4$) or borohydrides is also possible.

Catalytic oxidation with oxygen yields α-hydroxycarboxylic acids (Table 1, no. 17), which may be followed by C–C bond cleavage. Oxidation with dimethyl sulfoxide can be used to prepare α-hydroxyketones (Table 1, no. 18) [19].

If epoxides are reacted with a molecule containing polar double or triple bonds, either saturated or unsaturated five-membered heterocycles are obtained. Cycloaddition with phenylisocyanate affords *N*-phenyloxazolidones (Table 1, no. 19) [20]; cycloaddition with CO_2 yields alkylene carbonates (Table 1, no. 20), and cycloaddition with carbonyl compounds gives 1,3-dioxolanes (Table 1, no. 21) [21].

Because of the high activity of the oxirane ring, epoxides are readily polymerized (Table 1, no. 22). Polymerization of short-chain diepoxides is industrially important for the production of polymers cross-linked from an epoxide and a hardener. Examples of hardeners include di- and polyamines, polyphenols, di- and polycarboxylic anhydrides, and various di- and tricarboxylic acids [22].

3. Production of Epoxides

Epoxides are produced by the addition of oxygen to alkenes. The electrophilic agent may be either molecular or chemically bound oxygen such as percarboxylic acids, hydrogen peroxide, or hydroperoxides.

3.1. Epoxidation with Percarboxylic Acids [23]

Of all the available production methods for converting alkenes to epoxides, the most widely used method is the Prilezhaev (Prileschajew) reaction [24].

The *mechanism* proposed by P. D. Bartlett is generally accepted and involves a bicyclic transition state [25].

Epoxidation is a second-order reaction and is very exothermic (ca. 250 kJ/mol); great care must be taken at all times to ensure safe operation. Safety is especially critical when the speed of epoxidation is fast. Electron-donating groups, e.g., alkyl groups at the double bond carbon atoms, greatly enhance the reaction rate (see Table 2). Electron-withdrawing groups have the opposite effect and may sometimes stop the reaction entirely. 2-Butene reacts considerably faster than propene, whereas allyl chloride reacts slower [26]. Cyclic olefins react faster than dialkyl-substituted open-chain olefins, and substitution by a phenyl group exerts only a mild accelerating effect.

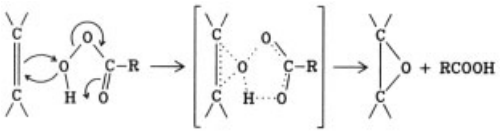

The reactivity of the percarboxylic acid is also subject to electronic effects. Electron-withdrawing groups increase the reaction rate of epoxidation by enhancing its electrophilicity.

Conversely, electron-donating groups reduce the reaction rate. 3-Chloroperbenzoic acid is a more rapid epoxidizing agent than perbenzoic acid, and trifluoroperacetic acid is similarly more reactive than peracetic acid.

Solvents also have a marked effect on the reaction rate. Hydrophilic solvents retard the rate by interfering with intramolecular hydrogen bonding of the percarboxylic acid. Chlorinated and aromatic solvents are more suitable for faster reaction rates.

Epoxidation is stereospecific; *cis* olefins give *cis* epoxides and *trans* olefins give *trans* epoxides. An excellent review of the stereochemistry of percarboxylic acid epoxidation is available [27].

The selection of a percarboxylic acid for epoxidation is not entirely dependent on the olefin and percarboxylic acid structure. The stability of the epoxide ring in the system that is chosen also influences product yield. Electronic effects of substituents attached to the epoxide ring markedly affect stability.

Carboxylic acids and water attack epoxides to produce mono esters and vicinal diols (Table 1, nos. 3 and 1). Strong acids have a marked catalytic effect on ring cleavage. Loss of epoxide in the presence of acid is generally inevitable, although careful choice of the percarboxylic acid and reaction conditions can minimize loss. A solvent is often employed to moderate the reaction and to facilitate subsequent isolation of the epoxide in high yield. Some evidence also suggests that use of a solvent reduces the extent of side reactions. Several excellent reviews on epoxidation with percarboxylic acids are available [28], [29].

The percarboxylic acid that is employed may be prepared in situ or made in advance (preformed). In both procedures, the percarboxylic acid or hydrogen peroxide is always added slowly to the olefin. The reverse procedure can be hazardous and is not recommended.

In Situ Percarboxylic Acids. In this simplest method of epoxidation, the olefinic compound is dissolved in the carboxylic acid and hydrogen peroxide is added. A low level of carboxylic acid (0.25 mol per mole of olefin) is required. Because of water from the aqueous hydrogen peroxide and from the reaction, the process occurs in two phases. Epoxidation with in situ performic acid or in situ peracetic acid (a strong acid catalyst is desirable) are established commercial processes. Major applications are in epoxidation of long-chain olefins [28], vegetable oils (e.g., soya bean oil), polybutadiene, natural and synthetic rubbers, and polyesters.

Preformed Percarboxylic Acid. Both equilibrium and solvent-extracted percarboxylic acids are used. Equilibrium peracetic acid contains 1 wt % sulfuric acid, which is neutralized shortly before use in epoxidation. The resulting solution, which is generally called buffered equilibrium peracetic acid, has been used to epoxidize stilbene [30]:

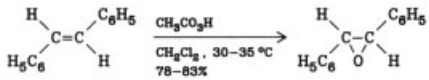

Greater selectivity is obtained by neutralization of acetic acid present in equilibrium peracetic acid, using stoichiometric quantities of sodium carbonate. The best method is to combine buffered peracetic acid with sodium carbon-

Table 2. Relative rates of epoxidation of selected olefins with peracetic acid at 25.8°C [23]

Olefin	CAS registry number	Formula	Relative rate of epoxidation
Ethylene	[74-85-1]	$CH_2=CH_2$	1
Propene	[115-07-1]	$CH_3CH=CH_2$	25
Hex-1-ene	[592-41-6]	$CH_3(CH_2)_3CH=CH_2$	
Oct-1-ene	[111-66-0]	$CH_3(CH_2)_5CH=CH_2$	
But-2-ene	[590-18-1]	$CH_3CH=CHCH_3$	500–600
Hex-2-ene	[7688-21-3]	$CH_3CH=CH(CH_2)_2CH_3$	
Methyl oleate	[2462-84-2]	$CH_3(CH_2)_7CH=CH(CH_2)_7COOCH_3$	
2-Methylbut-2-ene	[513-35-9]	$(CH_3)_2C=CHCH_3$	6000
Vinyl acetate	[108-05-4]	$CH_3COOCH=CH_2$	0.2
Allyl acetate	[591-87-7]	$CH_3COOCH_2CH=CH_2$	1–2
Cyclohexene	[110-83-8]		675
Styrene	[100-42-5]	$C_6H_5CH=CH_2$	60

ate. This technique has been applied to bicyclic terpenes, e.g., α-pinene [31], [32]:

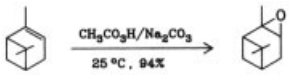

Other Percarboxylic Acid Systems. Distilled aqueous peracetic acid can be used in much the same way as equilibrium peracetic acid. Anhydrous peracetic and perpropionic acids (extracted percarboxylic acids) in aprotic solvents, such as ethyl acetate [33] or dichloropropane [34], have been developed. A wide range of other aliphatic and aromatic percarboxylic acids have been used in epoxidation reactions [27], [28].

The commercial availability of 3-chloroperbenzoic acid and the ecologically beneficial magnesium monoperoxyphthalate hexahydrate (INTEROX MMPP) has resulted in wide application [35], [36].

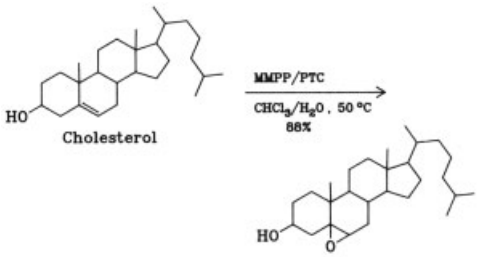

Unsaturated esters that are difficult to epoxidize have given good yields with trifluoroperacetic acid [37]. The use of this very strong epoxidizing agent is of academic interest.

Water-soluble olefins can be epoxidized readily in aqueous solution in the presence of tungstic or molybdic acids. The reaction is industrially important for the preparation of glycidol from allyl alcohol [38], [39], and *cis*-epoxysuccinic acid from maleic acid [40].

Benzene perselenic acid, $C_6H_5SeO_3H$, is produced in situ in two-phase systems from catalytic quantities (typically 1 %) of benzeneselenic acid and 30 % hydrogen peroxide [41], [42].

3.2. Epoxidation with Hydrogen Peroxide

Alkaline Hydrogen Peroxide [23]. Olefins with electron-withdrawing substituents may be epoxidized by alkaline hydrogen peroxide. The active epoxidizing agent is the hydroperoxy anion, and the generally accepted mechanism for the epoxidation of an α, β-unsaturated carbonyl compound is as follows:

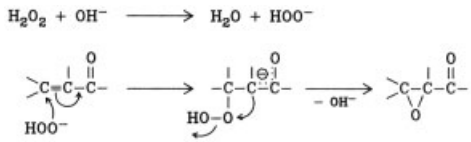

These reactions are stereoselective, in contrast to epoxidation with percarboxylic acids,

which is stereospecific. Only one isomer is obtained when both isomers of 3-methyl-2-penten-4-one are epoxidized by alkaline hydrogen peroxide [43].

Akaline Nitrile Hydrogen Peroxide [23]. In the absence of electron-withdrawing groups adjacent to the double bond, hydrogen peroxide must be activated for epoxidation to take place. Conversion to a percarboxylic acid is one method of achieving this. A similar method of activation is the reaction of nitriles in alkaline media to produce peroxyimidic acids (Payne system) [44].

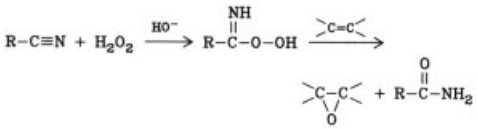

Acetonitrile and benzonitrile are generally used. The reaction is particularly applicable to acid-labile epoxides, e.g., glycidaldehyde diethylacetal [45]. The method has been adapted for large-scale production of hexafluoropropene oxide [46].

New Catalytic Systems. Although methods of epoxidation are versatile, highly efficient systems and simpler processes, particularly for industrial production, are desirable. Various catalytic systems and their reaction mechanisms have been compared with other oxidizing agents in a review article [47].

The catalytic tungstate/phosphate/tetraalkylammonium salt system (QX) is an example of a highly selective system that gives excellent yields [48].

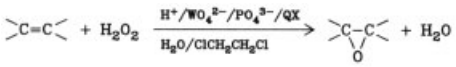

1-Dodecene can be epoxidized in 1 h at 70 °C in 87 % yield, using this system. Electron-rich double bonds can be converted with 85 % H_2O_2 and SeO_2 in an alcohol – alkali system with equally short reaction time and high selectivity [49].

For a catalyst to be effective in epoxidation reactions, it must be stable, highly reactive, and easily recyclable [50], [51]. Because many catalysts lose activity in the presence of water, water must be removed continuously during the process.

3.3. Epoxidation with Hydroperoxides

Catalytic epoxidation of olefins with hydroperoxides gained popularity only after the introduction of asymmetric epoxidation [52]. This method is restricted to allyl alcohols. In addition to reliability and ease of handling, it has the advantage of effective asymmetric induction and predictability of product. For instance, the reaction proceeds as follows depending on the stereochemistry of the diethyl tartrate employed [53]:

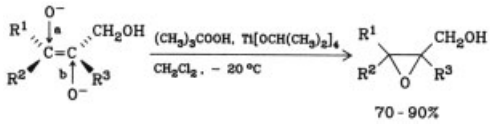

In the presence of (−)-diethyl D-tartrate (unnatural isomer), the active oxygen attacks from "above" (a) but attacks from "below" (b) in the presence of (+)-diethyl L-tartrate (natural isomer). The optical selectivity of the reaction is ≥ 90 %.

***tert*-Butyl hydroperoxide** [*75-91-2*] (TBHP) has been employed in catalytic epoxidation of linear and cyclic olefins [54]. High yields are obtained only if the reagents used in the reaction, especially the TBHP reagent, are absolutely dry [54], [55]. Epoxidation with TBHP has been used industrially to produce propylene oxide [56], [57].

Some of the advantages of TBHP as an oxidizing agent in the epoxidation of olefins include the following [47]:

1) high thermal stability
2) safer handling than H_2O_2 or percarboxylic acids
3) noncorrosive
4) selective oxidizing agent
5) good solubility in nonpolar solvents
6) neutral pH
7) the byproduct *tert*-butanol can be easily separated by distillation.

3.4. Epoxidation with Halohydrins

Hypohalous acids and their salts are suitable agents for the epoxidation of olefins with electron-deficient double bonds. Oxiranes are formed according to the following mechanism [58]:

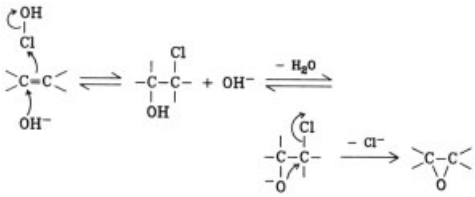

Dehydrohalogenation in the presence of base of the *trans* halohydrin formed by adding hypohalous acid to the olefin gives an epoxide with an inversion of configuration. Because addition of hypohalous acid already involves an inversion, the olefin undergoes at net double Walden inversion [24].

Aqueous chlorine is the preferred reactant for industrial application. Water-insoluble olefins are either emulsified in water or dissolved in a solvent. Alkali or alkaline-earth hydroxides are used as the base for the dehydrohalogenation reaction. Alkyl substituents on the olefin increase the rate of oxirane formation.

One difference between epoxide production by percarboxylic acids and synthesis via halohydrin is that dichlorides and halogen ethers are byproducts in the latter case. This restricts use of halohydrins to low molecular mass monoxides.

3.5. Epoxidation with Oxygen

In contrast to epoxidation processes described so far, selectivity and yield of epoxidation with oxygen are low. The reaction mechanism is presumed to be a radical-chain reaction. The actual oxidizing agent is an intermediate hydroperoxide.

The direct oxidation of olefins with oxygen can be carried out in the presence or absence of catalyst. Compounds from groups 4B – 6B of the periodic table (Mo, V, W, Cr, and Ti) are used as catalysts and are highly selective, but have little activity. On the other hand, compounds from groups 1B, 7B, and 8B (Co, Ni, Mn, Cu, Ir, Rh, Pt, and Ru) are more active, but show less selectivity [58].

The following process is important for the industrial production of ethylene oxide (→ Ethylene Oxide):

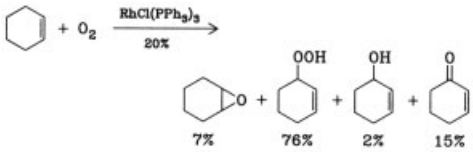

3.6. Epoxidation by Other Methods

Darzens Reaction. Esters of epoxycarboxylic acid can be prepared from esters of α-halocarboxylic acids and carbonyl compounds in the presence of alkaline condensing agents. The carbanion formed by reaction of the α-halocarboxylic acid ester with alkali reacts with the carbonyl moiety and then stabilizes by forming the glycidol ester [24]:

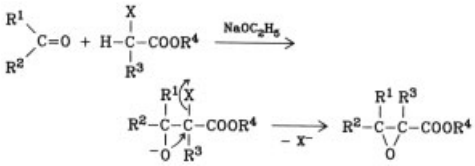

Esters of α-chloro- or α-bromo-α-phenylacetic acid and ketones in tautomeric keto–enol equilibrium do not undergo the Darzens reaction [24].

Reaction with Epichlorohydrin. Halohydrins are formed if compounds with active hydrogen atoms are reacted with epichlorohydrin [*106-89-8*] (see Section 4.1). From these halohydrins, subsequent treatment with alkali removes hydrogen halide, and a new epoxide ring is formed [24]:

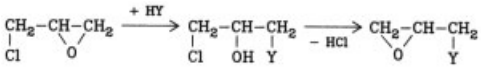

This method can be used, for example, to produce 2,3-epoxyalkyl ethers in a yield of 40 – 80 % [24].

Reaction with Glycidol. Glycidol [*556-52-5*] (**5**) is an important intermediate for the production of functional epoxides (see Section 4.3) [24]. For example, reaction of phosgene with glycidol yields 2,3-epoxypropyloxy chloroformate [24]:

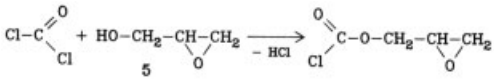

Reaction of glycidol (**5**) with an isocyanate affords the commercially important glycidyl urethanes (2,3-epoxypropyl urethanes) [59].

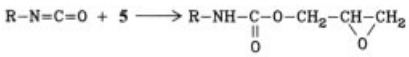

4. Industrially Important Epoxides

4.1. Epichlorohydrin

Epichlorohydrin [*106-89-8*], C_3H_5ClO, M_r 92.53, 1-chloro-2,3-epoxypropane, chloromethyl-oxirane, 1,2-epoxy-3-chloropropane, is the most important material for the production of epoxy resins. It is also used for the industrial production of glycerol (→ Glycerol).

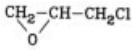

Physical Properties. Epichlorohydrin, *bp* 116.56 °C, *bp* 30–32 °C at 1.35 kPa, *mp* −48 °C, d_4^{20} 1.18066 g/cm^3, n_D^{20} 1.4382, viscosity 1.03 mPa · s at 25 °C, flash point 40.6 °C (Tagliabue Open Cup), autoignition temperature 415.6 °C, heat of vaporization (calculated) 37.9 kJ/mol, heat of combustion 18.943 kJ/mol, specific conductivity $34 \times 10^{-9}\,\Omega^{-1}\,cm^{-1}$, vapor density 3.29 (air = 1, at boiling point of epichlorohydrin), vapor pressure 1.333 kPa at 16.6 °C, is a colorless mobile liquid with a characteristic chloroform-like, irritating odor. The lower flammable limit of epichlorohydrin in air is 3.8 vol %; the upper limit is 21.0.

Epichlorohydrin is soluble in low molecular mass alcohols, esters, ethers, ketones, and aromatic hydrocarbons; it is sparingly soluble in water (6.6 wt % at 20 °C). Epichlorohydrin forms binary azeotropes with the following solvents (percent solvent; remaining percentage is epichlorohydrin):

water (25 %)	*bp* 88 °C
1-propanol (68 %)	*bp* 96 °C
isobutanol (54 %)	*bp* 105 °C
isoamyl alcohol (19 %)	*bp* 115 °C

Chemical Properties. As a pure compound, epichlorohydrin is stable. The epoxide and chlorine groups of epichlorohydrin are both potential reactive sites. In reactions with compounds containing active hydrogens, such as alcohols, acids, phenols, amines, and thiols, reaction occurs at the more reactive epoxide group:

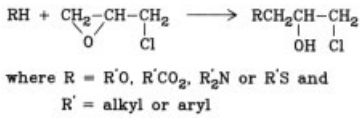

The monochlorohydrin can be converted to a glycidyl compound by reaction with sodium hydroxide or to the glycerol derivative by acid hydrolysis:

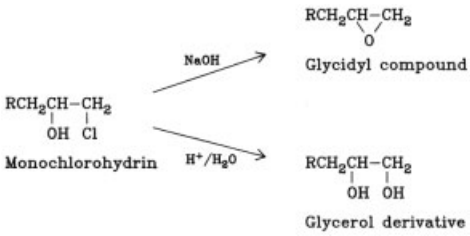

Uses. Epichlorohydrin is used mainly for the production of glycerol, unmodified epoxy resins (epichlorohydrin–bisphenol A resin), and elastomers. Several other products made from epichlorohydrin are glycidyl ethers, epichlorohydrin–polyamide resins, and alkyl glyceryl ether sulfonate salts.

Production [60]. Epichlorohydrin is made by chlorohydrination of allyl chloride (→ Allyl Compounds, Chap. 1.), which is obtained by high-temperature chlorination of propylene (Eq. 1). Byproducts of chlorination are *cis*- and *trans*-1,3-dichloropropene and 1,2-dichloropropane. Glycerol dichlorohydrins are

made from allyl chloride (Eq. 2), with 1,2,3-trichloropropane being obtained as a byproduct. Finally, epichlorohydrin is produced from the glycerol – dichlorohydrin mixture by treatment with base (Eq. 3). The reactions are as follows:

$$CH_2{=}CHCH_3 + Cl_2 \xrightarrow{500\,^\circ C} CH_2{=}CHCH_2Cl + HCl \qquad (1)$$

$$2\ CH_2{=}CHCH_2Cl + 2\ HOCl \longrightarrow \begin{matrix} HOCH_2CHClCH_2Cl \\ + \\ ClCH_2CHOHCH_2Cl \end{matrix} \qquad (2)$$

$$\xrightarrow{Base} \underset{O}{CH_2{-}CH}{-}CH_2Cl \qquad (3)$$

In 1985 Showa Denko KK (Japan) introduced an epichlorohydrin manufacturing process starting from allyl alcohol. Major producers include Shell Chemical and Dow Chemical.

Toxicology [61], [62]. Epichlorohydrin is intensely irritating and moderately toxic by oral, percutaneous, and subcutaneous routes as well as by inhalation of the vapor. The ACGIH has established a TLV-TWA for epichlorohydrin of 2 ppm, 10 mg/m^2, and a TLV-STEL of 5 ppm; the compound has a "skin" notation [63]. Direct contact of the liquid with skin or eyes causes severe burns and permanent injury. Inhalation of the vapor produces at least temporary sterility and causes lasting liver, lung, and kidney damage. A moderate degree of carcinogenic activity has been demonstrated in rats when exposed to epichlorohydrin. Persons working with epichlorohydrin require continuous medical supervision.

4.2. Other Important Monoepoxides

1,2-Epoxydodecane [*2855-19-8*], 1-dodecene oxide, decyloxirane, $C_{12}H_{24}O$, M_r 184.32, is a colorless liquid with a sweet, "soapy" odor; color (Hazen scale) $\leqq 10$; *bp* 250 °C (101.3 kPa), 125 °C (1.6 kPa), 80 °C (0.17 kPa); *mp* −5 °C; flash point (Seta flash) 104 °C; n_D^{20} 1.435; ϱ_4^{20} 0.840 g/cm^3; η^{20} 3.0 mPa · s; insoluble in water and soluble in most organic solvents; spontaneous exothermic reaction with strong acids and bases.

$$\underset{O}{CH_2{-}CH}{-}(CH_2)_9CH_3$$

Typical of long-chain α-epoxides, 1,2-epoxydodecane can enter into numerous reactions (see Chap. 2) [1]. It is produced industrially by established peracid processes either in the presence or absence of solvents [64], [65]. 1,2-Epoxydodecane is stable when stored in cool, dry conditions.

Long-chain α-epoxides are not used as such in most instances but rather as well-defined derivatives. Examples of direct application are as stabilizers for halogen hydrocarbons, as reactive diluents (e.g., for epoxy resins), as resin modifiers, or as a coating material. The various industrial applications of 1,2-epoxydodecane have been reviewed [1].

Isoamylene oxide [*5076-19-7*], 2-methyl-2-butene oxide, 2,3-epoxy-2-methylbutane, $C_5H_{10}O$, M_r 86.14, is a clear, colorless liquid with a stringent, ether-like odor; color (Hazen scale) 5; *bp* 73 °C (101.3 kPa); *mp* −83 °C; flash point (Seta flash) < 15 °C; n_D^{20} 1.3860; ϱ_4^{20} 0.805 g/cm^3; η^{20} 0.9 mPa·s; 8 % soluble in water (unstable) and soluble in polar organic solvents [66].

$$CH_3{-}\overset{CH_3}{\overset{|}{C}}{-}\underset{O}{CH}{-}CH_3$$

The major reaction of isoamylene oxide of industrial importance is ring-opening with acid. Isoamylene oxide isomerizes to methyl isopropyl ketone when heated to 250 °C over Al_2O_3.

Isoamylene oxide can be prepared by various methods, e.g., via the chlorohydrin route, with percarboxylic acid, or with hydrogen peroxide in a catalyzed reaction.

Isoamylene oxide is used commercially for stabilizing the chlorinated hydrocarbons 1,1,1-trichloroethane and trichloroethylene [67], [68]. These hydrocarbons are frequently used in metal cleaning (including light metal alloys) and must, therefore, be stabilized against decomposition and formation of acidic substances during storage and application. The basic stabilizer system consists of either an organic amine or a nitro compound and a low-molecular mass epoxide, to which additional components may be added.

α-Pinene oxide [*1686-14-2*], 2,3-epoxypinane, 2,7,7-trimethyl-3-oxatricyclo[4.1.1.0^{2,4}]octane, $C_{10}H_{16}O$, M_r 152.24, is a colorless liquid with a characteristic camphor – menthol odor; color (Hazen scale) 5 – 10; *bp* 185 °C (101.3 kPa), 82 °C (20 kPa); flash point (Seta flash) 62 °C; n_D^{20} 1.4690; ϱ_4^{20} 1.008 g/cm^3; η^{25} 4.01 mPa · s; $[\alpha]_D^{20}$ (+)-α-pinene oxide + 51 °, $[\alpha]_D^{20}$ (−)-α-pinene oxide −98 °; insoluble in water and soluble in most organic solvents; spontaneous exothermic reaction with concentrated acids.

6

α-Pinene oxide is prepared industrially by using percarboxylic acid [31], [32]. Processes based on air oxidation are also known.

An industrially important reaction of optically active α-pinene oxide (**6**) is rearrangement to campholene aldehyde (**7**) (e.g., in the presence of zinc bromide). Subsequent aldol condensation with lower aliphatic aldehydes or ketones (e.g., propionaldehyde, butyraldehyde, or acetone) provides unsaturated aldehydes (e.g., **8**) or ketones that can be reduced to the corresponding saturated alcohols (e.g., **9**) [69]. These are used in the fragrance industry as a sandalwood scent [70].

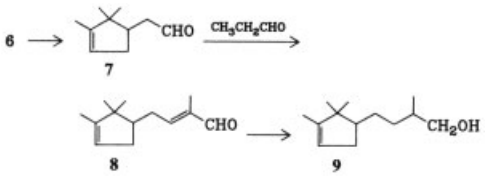

Hydrolysis of racemic α-pinene oxide and simultaneous rearrangement yields therapeutically active substances (e.g., **10**), which are used in compounding of pharmaceutical preparations [71].

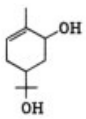

10

Styrene oxide [*96-09-3*], phenylethylene oxide, phenyloxirane, C_8H_8O, M_r 120.15, is a colorless liquid with an aromatic odor; *bp* 194 °C (101.3 kPa), 91 °C (3.33 kPa); *mp* −36.7 °C; ϱ_4^{25} 1.0469 g/cm^3; n_D^{20} 1.5350; vapor pressure < 1 kPa (20 °C); flash point 79 °C (open pan); soluble up to 0.3 % in water and miscible in all proportions with toluene, benzene, ether, chlorohydrocarbons, and ethyl acetate.

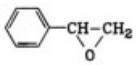

Similar to glycidol, styrene oxide is one of the most reactive epoxides [72]. In acid media, the rate of ring opening of the oxirane ring of styrene oxide is higher than for glycidyl ethers.

Styrene oxide is used industrially as a reactive diluent for epoxy resins. It is prepared either via the chlorohydrin route or catalytically with hydrogen peroxide.

4.3. Functional Epoxides

Glycidol [*556-52-5*] (**5**, page 8), glycidyl alcohol, epihydrine alcohol, glycide, glycerin anhydride, 2,3-epoxypropanol, hydroxymethyloxirane, $C_3H_6O_2$, M_r 74.081, is a colorless and odorless liquid; *bp* 162 °C (101.3 kPa), 61 °C (0.13 kPa); *mp* −53 °C; n_D^{20} 1.431; ϱ_4^{20} 1.115 g/cm^3; η^{20} 4.0 mPa · s; η^{60} 1.70 mPa · s; $[\alpha]_D^{21}$ L-(+)-glycidol +15°; flash point (DIN 51 758) 71 °C; ignition temperature (DIN 51 794) 415 °C; c_p (40 °C) 2.177 kJ kg^{-1} K^{-1}; miscible in all proportions with water, alcohols, ketones, esters, ethers, aromatics, etc. and almost insoluble in aliphatic hydrocarbons.

Because it is a bifunctional compound, glycidol tends to condense with itself to form glycerol glycidol ether, which can then react further with itself or with another glycidol molecule. The half-life for the polymerization of glycidol is 7.0 min (at 200 °C) and 4.0 min (at 250 °C). The polymerization energy at 200 °C has been reported to be 1170 kJ/kg. Glycidol is somewhat unstable during storage even in its purest form. The oxirane oxygen content decreases at 25 °C by 2 % and at 0 °C by 0.5 % per month. When heated, glycidol forms acetol and hydraacry-

laldehyde. The latter yields acrolein by elimination of water.

Solutions (up to 70 %) in toluene, ketones, esters, ethers, chlorohydrocarbons, and certain alcohols, such as 2-propanol and *tert*-butanol, are relatively stable up to 80 °C. For information on the numerous reactions and the preparation of glycidol, see [59].

Epoxidized soybean oil [*8013-07-8*] (ESBO) is a mixture of glycerol esters of epoxidized linoleic, linolenic, and oleic acid. It is a colorless to pale yellow oily liquid; *mp* 0 °C, ϱ_4^{20} ca. 0.99 g/cm^3, n_D^{20} ca. 1.473, $\eta^{20} < 600$ mPa · s; oxirane oxygen 6.5 %; iodine number < 6 (g of iodine/100 g); saponification number ca. 180 (mg of KOH/g); acid number < 0.4 (mg of KOH/g); iodine color number < 2; flash point < 300 °C (open pan); ignition temperature 430 °C (DIN 51 794); insoluble in water and miscible with toluene, benzene, chlorohydrocarbons, and ethyl acetate.

Epoxidized soybean oil is prepared commercially by the in situ peroxycarboxylic acid method either with or without a solvent. Flow charts for the continuous solventless processes used by Du Pont and Henkel may serve as representative examples of production methods [73], [74].

The ESBO market is second to that of epoxy resins; ESBO is used mainly as a plasticizer and stabilizer for poly(vinyl chloride) (PVC). Use as a costabilizer predominates in Europe.

4.4. Diepoxides

3,4-Epoxycyclohexylmethyl-3′,4′-epoxycyclohexane carboxylate [*2386-87-0*], (**13**), 3,4-epoxytetrahydrobenzyl-3′,4′-epoxytetrahydrobenzoate, 3,4-epoxycyclohexanecarboxylic acid-3′,4′-epoxycyclohexylmethyl ester, 7-oxabicyclo[4.1.0]heptane-3-carboxylic acid-7′-oxabicyclo[4.1.0]hept-3′-ylmethyl ester, $C_{14}H_{20}O_4$, M_r 252.3, is a clear, colorless, viscous liquid, almost odorless; color (Hazen scale) 5; *bp* 350 °C (10.13 kPa); glass transition temperature −16 to −18 °C; flash point (Seta flash) 197 °C; n_D^{20} 1.4977; ϱ_4^{20} 1.1680 g/cm^3; η^{20} 255 mPa · s, η^{40} 100 mPa · s, η^{80} 25 mPa · s; insoluble in water and miscible with polar organic solvents.

Production. If 1,2,5,6-tetrahydrobenzaldehyde (**11**) (obtained via the Diels–Alder reaction [75]) is used as the starting material, 3-cyclohexenylmethyl-3′-cyclohexene carboxylate (**12**) is formed in the presence of a catalyst (e.g., aluminum *sec*-butoxide) according to the Tishchenko reaction.

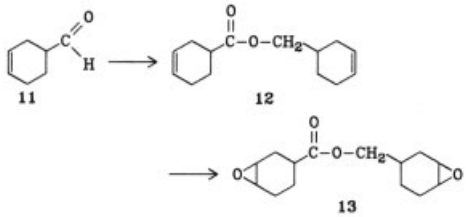

Compound **12** can be epoxidized to the product (**13**) by using the well-established peracid process [76].

This material is a low-viscosity bifunctional epoxy resin of great industrial importance because of its cross-linking capabilities, in most instances with cycloaliphatic dicarboxylic anhydrides. This diepoxide is used as a casting resin that can be hardened either thermally or by UV radiation. It includes among its distinguishing features good dielectric characteristics, excellent stability to environmental conditions, and shape retention when heated. It has also been used as a stabilizer in acid-releasing systems.

4.5. Polyepoxides

Polybutadiene oxide, epoxidized polyoils, M_r ca. 900 – 5000, are clear to pale yellow liquids, odorless to slightly sweet odor; degree of epoxidation 10 – 50 %; η^{20} 1300 – 17 000 mPa · s; flash point (Seta flash) 110 °C; insoluble in water, low solubility in toluene, benzene, and ethyl acetate, and miscible with tetrahydrofuran and chlorohydrocarbons.

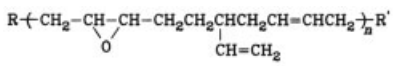

Polybutadiene oxides are prepared by in situ percarboxylic acid processes and are used to manufacture polymers that are used as architectural sealants and in the electronics industry.

5. Analysis

Titrimetric methods are generally used for the quantitative determination of epoxide groups. The method is based on the quick ring-opening reaction of the epoxide groups to form halohydrins when the epoxide is reacted with hydrohalides [77], [78].

When the direct titration method with hydrogen bromide in glacial acetic acid is used, maintenance of a constant HBr titer is difficult [79]. A variant of this method involves generating hydrogen bromide from a quaternary ammonium salt (e.g., tetraethylammonium bromide) during the titration by the action of perchloric acid in glacial acetic acid [80]. This very reliable method can also be used for functional epoxides. Peroxides do not interfere. Other methods include titration with hydrogen chloride in various solvents [81].

6. Economic Aspects

Reliable market data are available only for certain epoxides, such as epichlorohydrin and ESBO. In 1983, United States production of ESBO was 40 000 t, the same as that in Europe [82]. The bulk prices in the United States ranged between $ 1.52 and 1.59 per kg in 1984 [82] and were ca. DM 3.60 per kg in the Federal Republic of Germany in 1986.

The most important producers of ESBO are Argus, FMC, Ferro, Hall, Lankro, Rohm & Haas, and Viking (United States); and Air Liquide, AKZO, Ciba-Geigy, Henkel, PCUK, and VEB-Greiz (Europe).

The demand for other epoxides is difficult to estimate because much of the production never reaches the open market, but is used in-house to produce derived products.

7. Toxicology

Epoxides are not only extremely active in chemical syntheses, but are also equally active in the human body. By opening of the epoxide ring, they can bind in vivo to DNA, RNA, and proteins and are, thus, classified as alkylating agents. Low molecular mass linear epoxides are especially active and readily permeate to target areas in the cell where they can, for example, alkylate DNA [83].

Other chemicals can be metabolized to epoxides by the body. Epoxides of benzopyrenes and vinyl chloride formed in the body are considered to be the actual carcinogenic intermediates of these substances [84], [85].

In addition to the potential mutagenic properties of the epoxides [86], all but epoxidized fatty acids can cause skin inflammation [87] and allergic contact eczema.

Linear Epoxides. The mutagenicity of α-epoxides decreases with increasing chain length. For instance, propylene oxide is 20 times more mutagenic than 1-hexene oxide. No further increase takes place beyond 1-octene oxide [88]. However, electrophilic substituents increase the mutagenicity.

Cyclic Epoxides. The mutagenicities of a few cyclic epoxides have the following order: vinylcyclohexene monoxide $>>$ cyclopentene oxide $>$ cyclohexene oxide $>$ norbornene oxide [88], [89].

Diepoxides. The physiological effects of diepoxides were compared with those of nitrogen mustards [90]. Both 1,3-diepoxybutane and vinylcyclohexene dioxide gave rise to sarcoma in mammals following oral administration at 250 mg/kg [91]. The mutagenicities can be arranged in the following order: 1,3-diepoxybutane $>$ 1,7-diepoxyoctane $>>>$ 1,5-diepoxycyclooctane [88]. The last compound was shown to be neither carcinogenic [92] nor active in the mutagen assay [89].

Literature References.

Mutagenic activity: cyclohexene oxide [93]; styrene oxide and 4-methylstyrene oxide [94–97]; vinylcyclohexene monoxide and dioxide [95]; *cis*-stilbene oxide and 2-methyl-1,3-diepoxybutane [93].

Carcinogenic activity: 1,3-diepoxybutane [86], [87], [93]; 1,6-diepoxyhexane [87], [98]; 1,7-diepoxyoctane [94]; styrene oxide [87], [92], [98]; vinylcyclohexene dioxide [87], [92]; limonene dioxide [84], [87].

Specific Toxicity Data (threshold limit values) [87].

1-Butene oxide: LD_{50} 0.5 g/kg, $\leq$400 ppm (in humans).
1,3-Diepoxybutane: LD_{50} 0.078 g/kg (sarcoma in rats), $\leq$1 ppm (in humans).
Styrene oxide: LD_{50} 2 g/kg (lymphoma in rats).
Vinylcyclohexene monoxide: LD_{50} 2.83 g/kg, $\leq$50 ppm (in humans).
Vinylcyclohexene dioxide: LD_{50} 2.83 g/kg (mutagenic, teratogenic in animal experiments), $\leq$1 ppm (in humans).
Caryophyllene oxide: $LD_{50} > 5$ g/kg (negative in Ames test).
ESBO [99]: LD_{50} 22.5 mL/kg.

8. References

1. Peroxid-Chemie: *A 0.1.4., Langkettige Alpha-Epoxide,* Chapters 1 – 4, Höllriegelskreuth, 1981, pp. 1 – 29.
2. J. G. Smith, *Synthesis* 1984, 629 – 656.
3. G. H. Posner, D. Z. Rogers, *J. Am. Chem. Soc.* **99** (1977) 8208 – 8214.
4. G. H. Posner, D. Z. Rogers, *J. Am. Chem. Soc.* **99** (1977) 8214 – 8218.
5. M. Bischoff, U. Zeidler, H. Baumann, *Fette, Seifen, Anstrichm.* **79** (1977) 131 – 135.
6. B. Werdelmann, *Fette, Seifen, Anstrichm.* **76** (1974) 1 – 8.
7. R. E. Parker, N. S. Isaacs, *Chem. Rev.* **59** (1959) 737 – 799.
8. G. H. Posner, M. Hulce, R. K. Rose, *Synth. Commun.* **11** (1981) 737 – 741.
9. C. G. Moore, H. J. Porter, *J. Chem. Soc.* 1958, 2062 – 2064.
10. M. O. Brimeyer, A. Merota, S. Quici, A. Nigan et al., *J. Org. Chem.* **54** (1980) 4254 – 4255.
11. S. Torii, T. Okamoto, T. Oida, *J. Org. Chem.* **42** (1978) 2294 – 2296.
12. H. Gross, B. Costisella DD 108 3051972.
13. G. Kolaczinski, R. Mehren, W. Stein, *Fette, Seifen, Anstrichm.* **73** (1971) 553 – 557.
14. K. Wiechert, P. Mohr, *Z. Chem.* **7** (1967) 229 – 230.
15. H. Rutzen, *Fette, Seifen, Anstrichm.* **86** (1984) 109.
16. H. Alper, D. Des Roches, T. Durst, R. Legault, *J. Org. Chem.* **41** (1976) 3611 – 3613.
17. J. L. Eisenmann, *J. Org. Chem.* **27** (1962) 2706.
18. S. Murata, M. Suzuki, R. Noyori, *J. Am. Chem. Soc.* **101** (1979) 2738 – 2739.
19. T. Cohen, T. Tsuji, *J. Org. Chem.* **26** (1961) 1681.
20. D. Braun, J. Weinert, *Justus Liebigs Ann. Chem.* 1979, 200 – 218.
21. R. P. Hanzlik, M. Leinwetter, *J. Org. Chem.* **43** (1978) 438 – 440.
22. H. Batzer, F. Lohse, R. Schmid, *Angew. Makromol. Chem.* **29/30** (1973) 349 – 351.
23. Peroxid-Chemie: *A 0.1.2, Peroxygen Compounds in Organic Synthesis, Epoxidation,* Chapter 6, Höllriegelskreuth, 1981 pp. 11 – 17.
24. J. Falbe: "C–O Verbindungen," in F. Korte: *Methodicum Chimicum,* **5** G. Thieme Verlag, Stuttgart1975, pp. 131 – 201.
25. P. D. Bartlett, *Rec. Chem. Prog.* **11** (1950) 47 – 50.
26. D. Swern, *J. Am. Chem. Soc.* **69** (1947) 1692 – 1698.
27. B. Plesnicar: "Oxidation with Peroxy Acids and Other Peroxides," in N. S. Trahanovsky: *Oxidation in Organic Chemistry,* **5,** Part C, Chapter 3, Academic Press, New York 1978, pp. 211 – 294.
28. D. Swern: "Organic Peroxy Acids as Oxidizing Agent: Epoxidation,": in D. Swern: *Organic Peroxides,* **2,**5, Wiley-Interscience, New York 1971, pp. 355 – 533.
29. D. Swern: "Epoxidation and Hydroxylation of Ethylenic Compounds with Organic Peracids," in D. Swern: *Organic Reactions,* vol. 7, J. Wiley & Sons, New York 1953, Chapter 7, pp. 378 – 433.
30. D. J. Reif, H. O. House, *Org. Synth. Coll.* **4** (1963) 860.
31. Hoechst, DE-OS 2 835 940, 1978 (H. Häberlein, F. Scheidl).
32. Peroxid-Chemie, EP 55 387, 1981 (W. Edl., G. R. Sienel).
33. Union Carbide, DE-AS 1 216 306, 1959 (B. Phillips, P. S. Starcher).
34. Interox, GB-P 1 535 313, 1978.
35. R. N. McDonald, R. N. Steppel, J. E. Dorsey, *Org. Synth.* **50** (1970) 15 – 18.
36. V. I. Routledge, *Spec. Chem.* **6** (1986, May) 25 – 28.
37. W. D. Emmons, G. B. Lucas, *J. Am. Chem. Soc.* **77** (1955) 2287 – 2288.
38. Columbia-Southern, GB-P 837 464, 1957.
39. W. M. Weigert, A. Kleemann, G. Schreyer, *Chem. Ztg.* **99** (1975) 19 – 25.
40. Shell Development Co., US 2 833 787, 1955.
41. H. J. Reich, F. Chow, S. L. Peake, *Synthesis* 1978, 299 – 301.

42. M. Hori, K. B. Sharpless, *J. Org. Chem.* **43** (1978) 1689 – 1697.
43. H. O. House, R. S. Ro, *J. Am. Chem. Soc.* **80** (1958) 2428 – 2433.
44. G. B. Payne, P. H. Deming, P. H. Williams, *J. Org. Chem.* **26** (1961) 659 – 663. G. B. Payne, *Tetrahedron* **18** (1962) 763 – 765.
45. R. L. Rowland, A. Rodgman, J. N. Schumacher, *J. Org. Chem.* **29** (1964) 16 – 21.
46. Hoechst, DE-AS 2 557 655, 1975 (R. A. Sulzbach, F. Heller).
47. R. Sheldon, *Bull. Soc. Chim. Belg.* **94** (1985) 651 – 670.
48. C. Venturello, E. Alneri, M. Ricci, *J. Org. Chem.* **48** (1983) 3831 – 3833. Instituto Guido Donegani, GB-A 2 055 821, 1980.
49. Interox, EP-P 68 564, 1982 (A. Lecloux, F. Legrand, C. Declerck).
50. Ugine Kuhlmann, DE-OS 2 605 041, 1976 (M. Pralus, J.-P. Schirmann, S.-Y. Delavarenne); DE-OS 2 752 626, 1977 (J.-P. Schirmann, S.-Y. Delavarenne).
51. M. Pralus, J. C. Lecoq, J. P Schirmann, *Fundam. Res. Homogeneous. Catal.* **3** (1979) 327 – 343.
52. T. Katsuki, K. B. Sharpless, *J. Am. Chem. Soc.* **102** (1980) 5974 – 5976. K. B. Sharpless, *Chemtech.* **15** (1985) 692 – 700.
53. K. B. Sharpless, S. S. Woodard, M. G. Finn, *Pure Appl. Chem.* **55** (1983) 1823 – 1836.
54. K. B. Sharpless, T. R. Verhoeven, *Aldrichimica Acta* **12** (1979) 63 – 74. J. M. Klunder, M. Caron, M. Uchiyama, K. B. Sharpless, *J. Org. Chem.* **50** (1985) 912 – 915.
55. J. G. Hill, B. E. Rossiter, K. B. Sharpless, *J. Org. Chem.* **48** (1983) 3607 – 3608. Massachusetts Institute of Technology, EP-A 197 766, 1986 (R. M. Hanson, S. Y. Ko, K. B. Sharpless).
56. H. Waldman, *Compendium 75/76 Ergänzungsband der Zeitschrift Erdöl und Kohle, Erdgas, Petrochemie,* 1975 pp. 306 – 317.
57. R. Landau, G. A. Sullivan, D. Brown, *Chem. Technol.* 1979, 602 – 607.
58. M. Bartok, K. L. Lang: "Oxiranes," in S. Patai: *The Chemistry of Ethers, Crown Ethers, Hydroxyl Groups and Their Sulphur Analogues,* J. Wiley & Sons, New York 1980, Part 2, Chapter 14, pp. 609 – 879.
59. A. Kleemann, R. Wagner: *Glycidol,* Dr. Alfred Hüthig Verlag, Heidelberg 1981, pp. 81 – 83.
60. K. Wheeler, T. Ruess, S. Takahashi: "Epichlorohydrin," *Chemical Economics Handbook Marketing Research Report,* Stanford Research Institute International, Menlo Park, Calif., 1987, pp. 642.3021 A – 642.3021 W.
61. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., Van Nostrand Reinhold Co., New York 1984, pp. 709 – 710.
62. C. Hine, V. K. Rowe, E. R. White, K. I. Darmer et al., in G. D. Glayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd revised ed., **vol. 2 A,** Wiley-Interscience, New York 1981, pp. 2242 – 2247.
63. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* ACGIH, Cincinnati, Ohio, 1986 – 1987.
64. Degussa, EP 32 989, 1980 (G. Käbisch, R. Trübe, H. Wittmann, S. Raupach et al.).
65. Hoechst, DE-OS 2 436 817, 1974 (H. Häberlein, H. Korbanka, G. Nowy).
66. *Beilstein,* **17,** 13; **17 (2),** 22; **17 (3), (4),** 71.
67. Solvay, EP 62 952, 1982 (M. Servais, R. Crochet).
68. Solvay, EP 44 111, 1981 (R. Walraevens, M. Servais).
69. Dragoco, DE-OS 2 827 957, 1978 (E. J. Brunke, E. Klein).
70. VEB-Miltitz, DE-OS 1 922 391, 1969 (M. Mühlstädt, W. Dollase, M. Herrmann, G. Feustel).
71. C. Corvi, BE 764 323, 1971.
72. Dow Chemical, Styrene Oxide, Technical Service & Development, Midland, Mich., 1958.
73. A. F. Chadwick, *J. Am. Oil Chem. Soc.* **35** (1958) 355.
74. Henkel, DE-OS 3 320 219, 1983 (G. Dieckelmann, K. Echwert, L. Jeromin, E. Peukert et al.).
75. H. Batzer, E. Nikles, *Chimia* **16** (1962) 57.
76. Union Carbide, US 2 716 123, 1953.
77. B. Dobinson, W. Hofmann, B. P. Stark: *The Determination of Epoxide Groups,* Pergamon Press, Oxford 1969, pp. 1 – 79.
78. J. L. Jungnickel, E. D. Peters, A. Polgar, F. T. Weiss: *Organic Analysis,* **vol. 1,** Interscience, New York 1953, p. 127.
79. A. J. Durbetaki, *Anal. Chem.* **28** (1956) 2000 – 2001.
80. R. R. Jay, *Anal. Chem.* **36** (1964) 667 – 668.
81. W. B. Brown, *J. Soc. Chem. Ind.* **55** (1936) 321 T.
82. Predicast-Datenbank-Recherche, *Chem. Mark. Rep.* **225** (1984) 26.
83. L. Migliore, A. M. Rossi, N. Loprieno, *Mutat. Res.* **102** (1982) 425 – 437.

84. P. Kotin, H. L. Falk, *Radiat. Res.* 1963, Suppl. 3, 193 – 211.
85. W. K. Lutz: "*Chemische Karzinogenese: Biochemi-sche Aspekte*," Blockkurs in Toxikologie, Institut für Toxikologie der Eidgenössischen Technischen Hoch-schule (ETH) und Univ. Zürich, Schwerzenbach, Mar. 1987.
86. L. Ehrenberg, S. Hussain, *Mutat. Res.* **86** (1981) 1 – 113.
87. C. Hine, V. R. Rowe, E. R. White, K. I. Darmer et al. in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A,** Wiley-Interscience, New York 1981, pp. 2141 – 2257.
88. C. E. Voogd, J. J. von der Stel, J. J. J A. A. Jacobs. *Mutat. Res.* **89** (1981) 269 – 282.
89. S. W. Frantz, J. E. Sinsheimer, *Mutat. Res.* **90** (1981) 67 – 78.
90. H. P. Kaufmann, G. Hauschild, R. Schickel, *Fette, Seifen, Anstrichm.* **63** (1961) 239 – 241.
91. J. A. Hendry, R. F. Homer, F. L. Rose, A. L. Walpole, *Br. J. Pharmacol.* **6** (1951) 235 – 255.
92. M. M. Manson, *Br. J. Ind. Med.* **37** (1980) 317 – 336.
93. P. G. Gervasi, L. Citti, M. del Monte, V. Longo et al., *Mutat. Res.* **156** (1985) 77 – 82.
94. G. Turchi, S. Bonatti, L. Citti, P. G. Gervasi, *Mutat. Res.* **83** (1981) 419 – 430.
95. H. Norppa, K. Hemminki, M. Sorsa, H. Vainio, *Mutat. Res.* **91** (1981) 243 – 250.
96. S. de Flora, R. Koch, K. Strobel, M. Nagel: *Toxico-logical Environmental Chemistry,* **vol. 10,** Gordon & Breach Science Publ., London 1985, pp. 157 – 170.
97. K. Sugiura, M. Goto, *Chem.-Biol. Interact.* **35** (1981) 71 – 91.
98. B. L. van Duuren, *Ann. N.Y. Acad. Sci.* **163** (1969) 633 – 650.
99. C. S. Weil, N. Condra, C. Haun, J. A. Striegel, *J. Am. Ind. Hyg. Assoc.* **24** (1963) 305 – 325.

Epoxy Resins

Ha Q. Pham, Dow Chemical, Freeport, Texas,
Maurice J. Marks, Dow Chemical, Freeport, Texas,

1. Introduction 2
2. History 3
3. Industry Overview 5
4. Classes of Epoxy Resins and Manufacturing Processes 7
5. Liquid Epoxy Resins (DGEBA) 8
5.1. Caustic Coupling Process 9
5.2. Phase-Transfer Catalyst Process 9
6. Solid Epoxy Resins Based on DGEBA 10
6.1. SER Continuous Advancement Process 14
6.2. Phenoxy Resins 15
6.3. Epoxy-Based Thermoplastics . . 15
7. Halogenated Epoxy Resins 16
7.1. Brominated Bisphenol A Based Epoxy Resins 16
7.2. Fluorinated Epoxy Resins 16
8. Multifunctional Epoxy Resins . 17
8.1. Epoxy Novolac Resins 17
8.1.1. Bisphenol F Epoxy Resin 18
8.1.2. Cresol Epoxy Novolacs 18
8.1.3. Glycidyl Ethers of Hydrocarbon Epoxy Novolacs 19
8.1.4. Bisphenol A Epoxy Novolacs . . . 19
8.2. Other Polynuclear Phenol Glycidyl Ether Derived Resins 19
8.2.1. Glycidyl Ether of Tetrakis(4-hydroxyphenyl)ethane 19
8.2.2. Trisphenol Epoxy Novolacs 20
8.3. Aromatic Glycidyl Amine Resins 20
8.3.1. Triglycidyl Ether of *p*-Aminophenol 21
8.3.2. Tetraglycidyl Methylenedianiline (MDA) 21
9. Specialty Epoxy Resins 21
9.1. Crystalline Epoxy Resins Development 21
9.2. Weatherable Epoxy Resins 22
9.2.1. Hydrogenated DGEBA 22
9.2.2. Heterocyclic Glycidyl Imides and Amides 22
9.2.3. Hydantoin-Based Epoxy Resins . 23
9.3. Elastomer-Modified Epoxies . . 23
10. Monofunctional Glycidyl Ethers and Aliphatic Glycidyl Ethers . 23
11. Cycloaliphatic Epoxy Resins and Epoxidized Vegetable Oils 24
12. Epoxy Esters and Derivatives . . 26
12.1. Epoxy Esters 26
12.2. Glycidyl Esters 27
12.3. Epoxy Acrylates 28
12.4. Epoxy Vinyl Esters 28
12.5. Epoxy Phosphate Esters 29
13. Characterization of Uncured Epoxies 29
14. Curing of Epoxy Resins 32
15. Coreactive Curing Agents 34
15.1. Amine Functional Curing Agents 34
15.1.1. Primary and Secondary Amines . 34
15.1.1.1. Aliphatic Amines 36
15.1.1.1.1. Ketimines 36
15.1.1.1.2. Mannich Base Adducts 37
15.1.1.1.3. Polyetheramines 37
15.1.1.2. Cycloaliphatic Amines 37
15.1.1.3. Aromatic Amines 37
15.1.1.4. Arylyl Amines 38
15.1.2. Polyamides 38
15.1.3. Amidoamines 40
15.1.4. Dicyandiamide 41
15.2. Carboxylic Functional Polyester and Anhydride Curing Agents . 41
15.2.1. Carboxylic Functional Polyesters 42
15.2.2. Acid Anhydrides 43
15.3. Phenolic-Terminated Curing Agents 45
15.4. Melamine–, Urea–, and Phenol–Formaldehyde Resins 46
15.5. Mercaptans (Polysulfides and Polymercaptans) Curing Agents 47
15.6. Cyclic Amidines Curing Agents 47
15.7. Isocyanate Curing Agents 48
15.8. Cyanate Ester Curing Agents . . 48
16. Catalytic Cure 48
16.1. Lewis Bases 48
16.2. Lewis Acids 49
16.3. Photoinitiated Cationic Cure . . 50
17. Formulation Development With Epoxy Resins 51

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a09_547.pub2

17.1. Relationship Between Cured Epoxy Resin Structure and Properties 51
17.2. Selection of Epoxy Resins 52
17.3. Selection of Curing Agents 53
17.4. Epoxy/Curing Agent Stoichiometric Ratios 53
17.5. Catalysts 55
17.6. Accelerators 55
18. Epoxy Curing Process 56
18.1. Characterization of Epoxy Curing and Cured Networks 58
19. Formulation Modifiers 61
19.1. Diluents 61
19.2. Thixotropic Agents 63
19.3. Fillers 63
19.4. Epoxy Nanocomposites 65
19.5. Toughening Agents and Flexibilizers 65
20. Coatings Applications 66
20.1. Coatings Application Technologies 68
20.1.1. Low Solids Solventborne Coatings 68
20.1.2. High Solids Solventborne Coatings 68
20.1.3. Solvent-Free Coatings (100% Solids) 68
20.1.4. Waterborne Coatings 69
20.1.5. Powder Coatings 69
20.1.6. Radiation-Curable Coatings 71
20.2. Epoxy Coatings Markets 71
20.2.1. Marine and Industrial Maintenance Coatings 71
20.2.2. Metal Container and Coil Coatings 73
20.2.3. Automotive Coatings 75
20.3. Inks and Resists 77
21. Structural Applications 78
21.1. Structural Composites 78
21.1.1. Epoxy Composites 78
21.1.2. Epoxy Vinyl Ester Composites . . 79
21.1.3. Mineral-Filled Composites 79
21.2. Civil Engineering, Flooring, and Construction 79
21.3. Electrical Laminates 80
21.4. Other Electrical and Electronic Applications 83
21.4.1. Casting, Potting, and Encapsulation 83
21.4.2. Transfer Molding 84
21.5. Adhesives 84
21.6. Tooling 85
22. Health and Safety Factors 86
23. Acknowledgments 86
24. References 86

1. Introduction

Epoxy resins are an important class of polymeric materials, characterized by the presence of more than one three-membered ring known as the epoxy, epoxide, oxirane, or ethoxyline group.

The word "epoxy" is derived from the Greek prefix "ep," which means over and between, and "oxy," the combining form of oxygen [1]. By strict definition, epoxy resins refer only to uncross-linked monomers or oligomers containing epoxy groups. However, in practice, the term epoxy resins is loosely used to include cured epoxy systems. It should be noted that very high molecular weight epoxy resins and cured epoxy resins contain very little or no epoxide groups. The vast majority of industrially important epoxy resins are bi- or multifunctional epoxides. The monofunctional epoxides are primarily used as reactive diluents, viscosity modifiers, or adhesion promoters, but they are included here because of their relevance in the field of epoxy polymers.

Epoxies are one of the most versatile classes of polymers with diverse applications such as metal can coatings, automotive primer, printed circuit boards, semiconductor encapsulants, adhesives, and aerospace composites. Most cured epoxy resins provide amorphous thermosets with excellent mechanical strength and toughness; outstanding chemical, moisture, and corrosion resistance; good thermal, adhesive, and electrical properties; no volatiles emission and low shrinkage upon cure; and dimensional stability—a unique combination of properties generally not found in any other plastic material. These superior performance characteristics, coupled with outstanding formulating versatility and reasonable costs, have gained epoxy resins wide acceptance as materials of choice for a multitude of bonding, structural, and protective coatings applications.

Commercial epoxy resins contain aliphatic, cycloaliphatic, or aromatic backbones and are available in a wide range of molecular weights from several hundreds to tens of thousands. The most widely used epoxies are the glycidyl ether derivatives of bisphenol A (> 75 % of resin sales volume). The capability of the highly strained epoxy ring to react with a wide variety of curing agents under diverse conditions and temperatures imparts additional versatility to the epoxies. The major industrial utility of epoxy resins is in thermosetting applications. Treatment with curing agents gives insoluble and intractable thermoset polymers. In order to facilitate processing and to modify cured resin properties, other constituents may be included in the compositions: fillers, solvents, diluents, plasticizers, catalysts, accelerators, and tougheners.

Epoxy resins were first offered commercially in the late 1940s and are now used in a number of industries, often in demanding applications where their performance attributes are needed and their modestly high prices are justified. However, aromatic epoxies find limited uses in exterior applications because of their poor ultraviolet (UV) light resistance. Highly cross-linked epoxy thermosets sometimes suffer from brittleness and are often modified with tougheners for improved impact resistance.

The largest use of epoxy resins is in protective coatings (> 50 %), with the remainder being in structural applications such as printed circuit board (PCB) laminates, semiconductor encapsulants, and structural composites; tooling, molding, and casting; flooring; and adhesives. New, growing applications include lithographic inks and photoresists for the electronics industry.

2. History

The patent literature indicates that the synthesis of epoxy compounds was discovered as early as the late 1890s [2]. In 1934, Schlack of I.G. Farbenindustrie AG in Germany filed a patent application for the preparation of reaction products of amines with epoxies, including one epoxy based on bisphenol A and epichlorohydrin [3]. However, the commercial possibilities for epoxy resins were only recognized a few years later, simultaneously and independently, by the DeTrey Fréres Co. in Switzerland [4] and by the DeVoe and Raynolds Co. [5] in the United States.

In 1936, Pierre Castan of DeTrey Fréres Co. produced a low melting epoxy resin from bisphenol A and epichlorohydrin that gave a thermoset composition with phthalic anhydride. Application of the hardened composition was foreseen in dental products, but initial attempts to market the resin were unsuccessful. The patents were licensed to Ciba AG of Basel, Switzerland, and in 1946 the first epoxy adhesive was shown at the Swiss Industries Fair, and samples of casting resin were offered to the electrical industry.

Immediately after World War II, Sylvan Greenlee of DeVoe and Raynolds Co. patented a series of high molecular weight (MW) epoxy resin compositions for coating applications. These resins were based on the reaction of bisphenol A and epichlorohydrin, and were marketed through the subsidiary Jones-Dabney Co. as polyhydroxy ethers used for esterification with drying oil fatty acids to produce alkyd-type epoxy ester coatings. Protective surface coatings were the first major commercial application of epoxy resins, and they remain a major outlet for epoxy resin consumption today. Concurrently, epoxidation of multifunctional olefins with peroxy acids was studied by Daniel Swern as an alternative route to epoxy resins [6]. Meanwhile, Ciba AG, under license from DeTrey Frères, further developed epoxy resins for casting, laminating, and adhesive applications, and the Ciba Products Co. was established in the United States.

In the late 1940s, two U.S. companies, Shell Chemical Co. and Union Carbide Corp. (then Bakelite Co.), began research on bisphenol A based epoxy resins. At that time, Shell was the only supplier of epichlorohydrin, and Bakelite was a leading supplier of phenolic resins and bisphenol A. In 1955, the four U.S. epoxy resin manufacturers entered into a cross-licensing agreement. Subsequently, The Dow Chemical Co. and Reichhold Chemicals, Inc. joined the patent pool and began manufacturing epoxy resins.

In the 1960s, a number of multifunctional epoxy resins were developed for higher temperature applications. Ciba Products Co. manufactured and marketed *o*-cresol epoxy novolac resins, which had been developed by Koppers Co. Dow developed the phenol novolac epoxy

resins, Shell introduced polyglycidyl ethers of tetrafunctional phenols, and Union Carbide developed a triglycidyl *p*-aminophenol resin. These products continue to find uses today in highly demanding applications such as semiconductor encapsulants and aerospace composites where their performance justifies their higher costs relative to bisphenol A based epoxies.

The peracetic acid epoxidation of olefins was developed in the 1950s by Union Carbide in the United States and by Ciba AG in Europe for cycloaliphatic structures. Ciba Products marketed cycloaliphatic epoxy resins in 1963 and licensed several multifunctional resins from Union Carbide in 1965. The ensuing years witnessed the development of general-purpose epoxy resins with improved weathering characteristics based on the five-membered hydantoin ring and also on hydrogenated bisphenol A, but their commercial success has been limited because of their higher costs. Flame-retardant epoxy resins based on tetrabromobisphenol A were developed and commercialized by Dow Chemical for electrical laminate and composite applications in the late 1960s.

In the 1970s, the development of two breakthrough waterborne coating technologies based on epoxy resins helped establish the dominant position of epoxies in these markets: PPG's cathodic electrodeposition automotive primer and ICI-Glidden's epoxy acrylic interior can coatings.

While epoxy resins are known for excellent chemical resistance properties, the development and commercialization of epoxy vinyl ester resins in the 1970s by Shell and Dow offered enhanced resistance properties for hard-to-hold, corrosive chemicals such as acids, bases, and organic solvents. In conjunction with the development of the structural composites industry, epoxy vinyl ester resin composites found applications in demanding environments such as tanks, pipes and ancillary equipment for petrochemical plants and oil refineries, automotive valve covers, and oil pans. More recently, epoxy and vinyl esters are used in the construction of windmill blades for wind energy farms. Increasing requirements in the composite industries for aerospace and defense applications in the 1980s led to the development of new, high performance multifunctional epoxy resins based on complex amine and phenolic structures. Examples of those products are the trisphenol epoxy novolacs developed by Dow Chemical and now marketed by Huntsman (formerly Ciba).

The development of the electronics and computer industries in the 1980s demanded higher performance epoxy resins. Faster speeds and more densely packed semiconductors required epoxy encapsulants with higher thermal stability, better moisture resistance, and higher device reliability. Significant advances in the manufacturing processes of epoxy resins led to the development of electronic-grade materials with lower ionic and chloride impurities and improved electrical properties. Dow Chemical introduced a number of new, high performance products such as hydrocarbon epoxy novolacs based on dicyclopentadiene. The 1980s also witnessed the development of the Japanese epoxy resin industry with focus on specialty, high performing and high purity resins for the electronics industry. These include the commercialization of crystalline resins such as biphenol diglycidyl ether.

More recently, in order to comply with more stringent environmental regulations, there has been increased attention to the development of epoxy resins for high solids, powder, and waterborne and radiation-curable coatings. Powder coatings based on epoxy–polyester and epoxy–acrylate hybrids have continued to grow in the global markets, including new applications such as primer-surfacer and topcoats for automotive coatings. Radiation-curable epoxy–acrylates and cycloaliphatic epoxies showed tremendous growth in the 1990s in radiation-curable applications. These include important and new uses of epoxy resins such as the photoresists and lithographic inks for the electronics industry. Waterborne epoxy coatings are projected to grow substantially.

The continuing trend of device miniaturization in the computer industry, and the explosive growth of portable electronics and communications devices such as wireless cellular telephones in the 1990s demanded new, high performance resins for the PCB market. This has led to the development of new epoxies and epoxy hybrid systems having lower dielectric constants (D_k), higher glass-transition temperatures (T_g), and higher thermal decomposition temperatures (T_d) for electrical laminates. Environmental pressures in the PCB industry have fueled the development of a number of new

bromine-free resin systems, but their commercialization is limited because of higher costs.

Significant efforts have been directed toward performance enhancements of epoxy structural composites. Advances have been made in the epoxy-toughening area. Epoxy nanocomposites and nanotube systems have been studied and are claimed to bring exceptional thermal, chemical, and mechanical property improvements. However, commercialization has not yet materialized.

In 1999, Dow Chemical introduced a new epoxy-based thermoplastic resin, BLOX*, for gas barrier, adhesives, and coatings applications.

3. Industry Overview

From the first commercial introduction of diglycidyl ether of bisphenol A (DGEBA) resins in the 1940s, epoxy resins have gradually established their position as an important class of industrial polymers. Epoxy resin sales increased rapidly in the 1970s and continued to rise into the 1980s as new applications were developed (annual growth rate > 10 % in the U.S. market, Table 1). More recently, the slower growth rates (3–4 %) of the U.S., Japanese, and European markets in the 1990s were made up for by the higher growth rate (5–10 %) in the Asia-Pacific markets outside of Japan, particularly in Taiwan and China. Epoxy resin growth has historically tracked well with economic developments and demands for durable goods, and so the growth of the epoxy markets in Asia-Pacific is expected to continue into the next decade.

The global market for epoxy resins is estimated at approximately 1.15 million metric tons (MT) for the year 2000 [8]. This is an increase of 5 % over 1999 demands. The North American market consumed over 330,000 MT of epoxy resins, the European market is estimated at more than 370,000 MT, and the Asian market has surpassed both the North American and European markets by consuming 400,000 MT of epoxy resins. About 50,000 MT of epoxies were consumed in the South American markets. Imports of epoxy resins from Asia into North America has steadily grown to about 120,000 MT in 2000. Epoxy resins were used with over 400,000 MT of curing agents to produce an estimated 3 million MT of formulated compounds, worth over $20 billion.

Table 1. History of U.S. epoxy resin annual production[a]

Year	Production, 10^3 MT
1955	10
1960	30
1965	55
1970	79
1975	100
1980	201
1985	347
1990	475
1994	433

[a] Data from U.S. International Trade Commission, *Synthetic Organic Chemicals*. Data include modified and unmodified epoxy resins. Modified epoxy resins include solid epoxy resin (SER), vinyl ester resins, epoxy acrylates, etc. There appear to be some discrepancies in epoxy resin production and market data as reported by different publications and organizations [7]. This is primarily due to the fact that some epoxy resins such as liquid DGEBA resins and epoxy novolacs are used as raw materials to produce modified or advanced epoxy resins, which may be further converted to end-use products. Some publications report only unmodified epoxies.

Up until the mid-1990s, the major worldwide producers of epoxy resins were Dow Chemical, Shell, and Ciba-Geigy. However, both Shell and Ciba-Geigy have recently divested their epoxy resins businesses. Shell sold their epoxy business to Apollo Management LP (based in New York City) in the year 2000 and the company was renamed Resolution Performance Products. Similarly, Ciba's epoxy business was sold in 2000 to Morgan Grenfell, a London (U.K.)-based private equity firm, and the new company name was Vantico. More recently, in June 2003, the Vantico group of companies joined Huntsman. The Vantico business units are now named Huntsman Advanced Materials. The cycloaliphatic epoxy business of Union Carbide became part of The Dow Chemical Company after their merger in the year 2001. Together, these three producers continue to dominate the world market for epoxy resins, accounting for almost 65 % of the global market. However, this is a reduction from over 70 % of market shares owned by the three largest producers in the 1980s. Smaller producers of epoxy resins for the North American markets are Reichhold (owned by Dainippon Ink and Chemicals), CVC Specialty Chemicals, Pacific Epoxy Polymers, and InChem (phenoxy thermoplastic resins). Suppliers of epoxy derivatives include

Ashland Specialty Chemical, UCB Chemicals (Radcure), AOC LLC, Eastman Chemical, and Interplastic Corp.

The market in Europe is similarly dominated by the three big producers: Dow, Resolution, and Huntsman and their affiliated joint ventures. Other smaller epoxy producers include Bakelite AG, LEUNA-Harze, Solutia, SIR Industriale, and EMS-CHEMIE. Imports from Asia have become significant in recent years.

The last two decades marked the emergence of the Asian epoxy industry. In the 1980s, the Japanese epoxy industry was transformed from a number of joint venture companies with Dow, Shell, and Ciba into independent producers and the emergence of a high number of new producers. This coincides with the development of Japan as a world-class manufacturing base. The Japanese epoxy industry is known for their special focus on high performance, high purity resins for the electronics industry. According to data from the Japan Epoxy Resin Manufacturers Association, the total Japanese market demand is estimated at approximately 200,000 MT for the year 2000. The production capacity is estimated at 240,000 MT annually. Exports accounted for an estimated 40,000 MT in 2000. Major Japanese epoxy resin producers are Tohto Kasei, Japan Epoxy Resins Corp. (formerly Yuka-Shell), Asahi Kasei, Dai Nippon Ink and Chemicals, Dow Chemical Japan, Mitsui Chemicals, Nihon Kayaku, Sumitomo Chemical, and Asahi Denka Kogyo. In Japan, Tohto Kasei is a leading resin producer, with epoxy technology licensing arrangements with numerous resin producers in Asia.

Outside of Japan, there have been significant increases in epoxy market demands and capacity in the 1990s. This is due to the migration of many PCB, electronic, computer, and durable goods manufacturing plants into the region, which has considerably lower manufacturing costs. Nan Ya, a subsidiary of the Formosa Plastics Group based in Taiwan, is emerging as a major epoxy resin producer with some import presence in North America and Europe. Similarly, Kukdo of Korea also exports to the North American and European markets. The output of these two companies now account for an estimated 15 % of the world market. In China, there are numerous (more than 200) small domestic producers of epoxy resins. Recently, a number of major epoxy producers have announced joint ventures or plans to build manufacturing plants in China. These include a number of companies with integrated capacity into electrical laminates and PCB manufacturing, following the business model pioneered by the Formosa Plastics Group. Other notable Asian producers include Asia Pacific Resins, Chang Chun, and Eternal Chemical of Taiwan; Thai Epoxy of Thailand; Kumho, LG Chemical, and Pacific Epoxy Resins of Korea; and Guangdong Ciba Polymers, Sinopec Baling Petrochemical, Jiangsu Sanmu, and Wuxi DIC Epoxy Resin of China. The LG Chemical epoxy business was purchased by Bakelite in late 2002. A significant amount of resin produced in Taiwan and China is directed toward electrical laminates applications. The aggressive buildup of epoxy capacity in Asia has put significant pressures on resin prices, particularly the high volume products such as liquid epoxy resins based on bisphenol A (Table 2). But as of January 2004, the epoxy market demand in China alone has increased to more than 500,000 MT (Chinese Epoxy Industry Web site).

Table 2. Epoxy production capacity in Asia-Pacific[a] (2001)

Country	Existing capacity, 1000 MT/year	Announced capacity, 1000 MT/year
Japan	240	
Taiwan	239	70
China	100	255
Korea	180	
Thailand	30	
Malaysia	10	
Philippines	10	
Total	*809*	*325*

[a] Compilation of published data by Dow Chemical.

Estimated average prices for epoxy resin products in North America are given in Table 3. As with other petrochemical-based products, they depend on crude-oil prices. Prices of multifunctional resins are typically higher. They are based on more expensive raw materials than DGEBA resins and involve more complex manufacturing procedures. A listing of some major epoxy resin producers and the trade names of their products is shown in Table 4.

There are numerous suppliers of epoxy curing agents. Some of the major producers are Air Products and Chemicals, Cognis, Degussa, DSM, Huntsman, and Resolution.

Table 3. U.S. average epoxy resin prices and applications (2000)

Resin	$/kg	Applications
Liquid epoxy resins (Diglycidyl ether of bisphenol A, DGEBA)	2.2	coatings, castings, tooling, flooring, adhesives, composites
Solid epoxy resins (SER)	2.4	powder coating; epoxy esters for coatings; can, drum, and maintenance coatings
Bisphenol F epoxy	4.4	coatings
Multifunctional		
Phenol epoxy novolac	4.8	castings, coatings, laminates
Cresol epoxy novolac	8.8	electronics encapsulants, powder coatings, laminates
Other multifunctional epoxies	11–44	composites, adhesives, laminates, electronics
Cycloaliphatic epoxies	6.6	electrical castings, coatings, electronics
Brominated epoxies	3.3–5.5	printed wiring boards, composites
Epoxy vinyl esters	3.3	composites
Phenoxy resins	11–17	coatings, laminates, glass sizing
Epoxy diluents	4–11	

Table 4. List of some epoxy resin producers and their product trade names

Company	Trade name
Resolution Performance Products	Epon, Eponol, Eponex, Epi-Cure, Epikote
Dow Chemical	D.E.R., D.E.N., D.E.H., Derakane, E.R.L
Hunstman Advanced Materials (formerly Ciba, Vantico)	Araldite, Aralcast
Reichhold Chemical	Epotuf
Nan Ya	NPEL, NPES
Kukdo Chemical	YD
Dainippon Ink & Chemical (DIC)	Epiclon
Tohto Kasei	Epotohto
Japan Epoxy Resin (JER)	Epikote
Asahi Kasei	A.E.R.
Mitsui Chemical	Eponik
Sumitomo Chemical	Sumiepoxy
Thai Epoxy	Epotec
Chang Chun	
InChem	Paphen
Pacific Epoxy Polymers	PEP
CVC Specialty Chemicals	Erysis, Epalloy

4. Classes of Epoxy Resins and Manufacturing Processes

Most commercially important epoxy resins are prepared by the coupling reaction of compounds containing at least two active hydrogen atoms with epichlorohydrin followed by dehydrohalogenation:

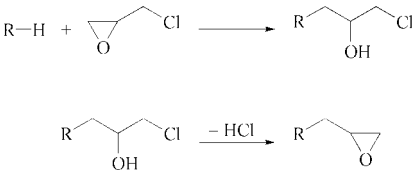

These included polyphenolic compounds, mono and diamines, amino phenols, heterocyclic imides and amides, aliphatic diols and polyols, and dimeric fatty acids. Epoxy resins derived from epichlorohydrin are termed glycidyl-based resins.

Alternatively, epoxy resins based on epoxidized aliphatic or cycloaliphatic dienes are produced by direct epoxidation of olefins by peracids:

Approximately 75 % of the epoxy resins currently used worldwide are derived from DGEBA. This market dominance of bisphenol A based epoxy resins is a result of a combination of their relatively low cost and adequate-to-superior performance in many applications. Figure 1 shows U.S. consumption of major epoxy resin types for the year 2000.

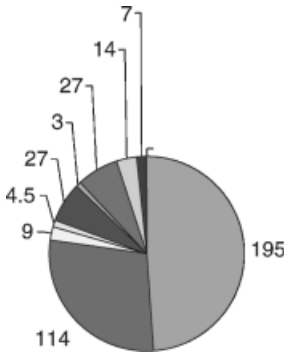

Figure 1. Major epoxy resin and derivatives markets (10^3 MT). LER; SER; epoxy novolacs; other multifunctional epoxies; brominated epoxies; cycloaliphatic; vinyl esters; and epoxy acrylates.

5. Liquid Epoxy Resins (DGEBA)

The most important intermediate in epoxy resin technology is the reaction product of epichlorohydrin and bisphenol A. It is often referred to in the industry as liquid epoxy resin (LER), which can be described as the crude DGEBA where the degree of polymerization, n, is very low ($n \cong 0.2$) (see below).

Pure DGEBA is a crystalline solid (*mp* 43°C) with an epoxide equivalent weight (EEW) of 170. The typical commercial unmodified liquid resins are viscous liquids with viscosities of 11,000–16,000 mPa · s (= cP) at 25°C, and an epoxide equivalent weight of ca 188.

EEW is the weight of resin required to obtain one equivalent of epoxy functional group. It is widely used to calculate reactant stoichiometric ratios for reacting or curing epoxy resins. It is related to the epoxide content (%) of the epoxy resin through the following relationship:

$$\mathrm{EEW} = \frac{43.05}{\%\mathrm{Epoxide}} \times 100$$

where 43.05 is the molecular mass of the epoxide group, $-C_2H_3O$. Other equivalent terminologies common in the industry include weight per epoxide (Wpe) or epoxide equivalent mass (EEM).

The outstanding performance characteristics of the resins are conveyed by the bisphenol A moiety (toughness, rigidity, and elevated temperature performance), the ether linkages (chemical resistance), and the hydroxyl and epoxy groups (adhesive properties and formulation latitude; reactivity with a wide variety of chemical curing agents).

LERs are used in coatings, flooring and composites formulations where their low viscosity facilitates processing. A large majority of LERs are used as starting materials to produce higher molecular weight (MW) solid epoxy resins (SER) and brominated epoxy resins, and to convert to epoxy derivatives such as epoxy vinyl esters, epoxy acrylates, etc. The bisphenol A derived epoxy resins are most frequently cured with anhydrides, aliphatic amines, phenolics, or polyamides, depending on desired end properties. Some of the outstanding properties are superior electrical properties, chemical resistance, heat resistance, and adhesion. Cured LERs give tight cross-linked networks having good strength and hardness but have limited flexibility and toughness.

Epichlorohydrin, or 3-chloro-1,2-epoxy propane (*bp* 115°C), is more commonly prepared from propylene by chlorination to allyl chloride, followed by treatment with hypochlorous acid. This yields glycerol dichlorohydrin, which is dehydrochlorinated by sodium hydroxide or calcium hydroxide [9].

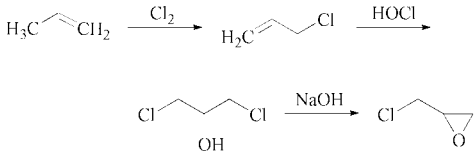

In industrial practices, epichlorohydrin is produced by direct chlorohydroxylation of allyl chloride in chlorine and water [10 – 13]. Alternatively, a new epichlorohydrin process has been developed and commercialized by Showa Denko [14] in Japan in 1985. It involves the chlorination of allyl alcohol as the precursor and is claimed to be more efficient in chlorine usage.

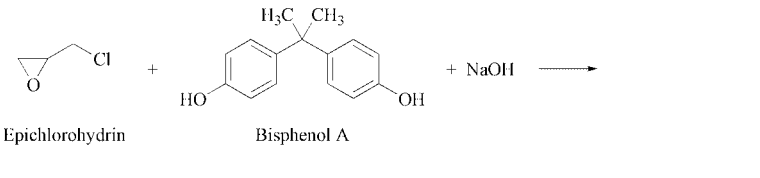

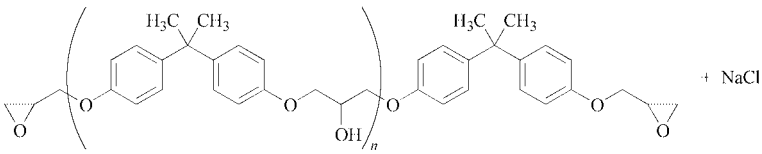

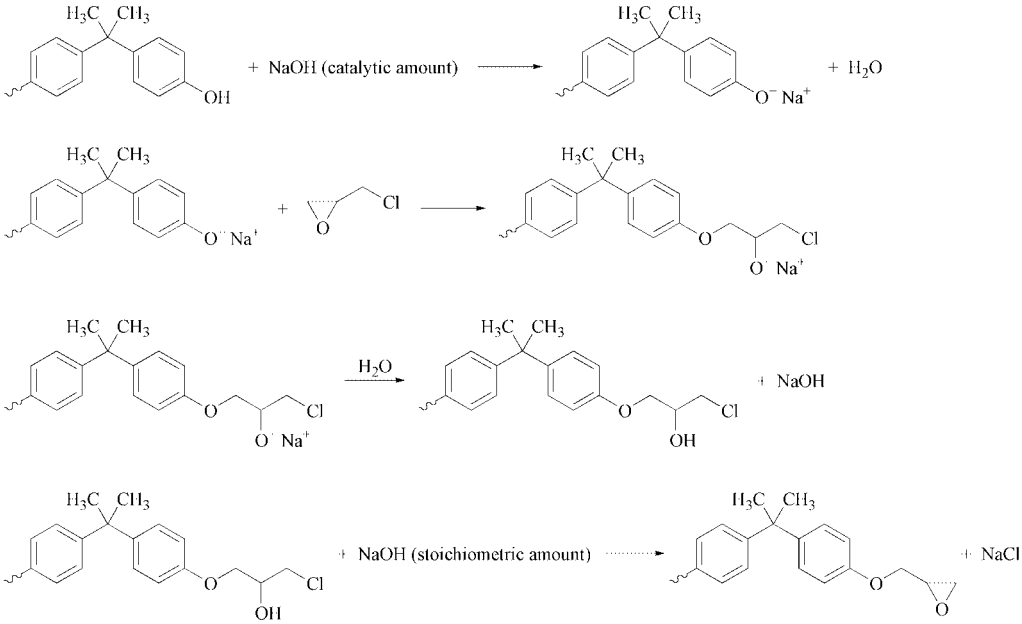

Bisphenol A (*mp* 153°C), or 2,2-bis(*p*-hydroxyphenyl)propane, is prepared from 2 moles of phenol and 1 mole of acetone [15, 16]

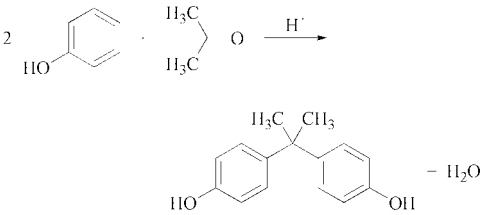

Bisphenol A based liquid epoxy resins are prepared in a two-step reaction sequence from epichlorohydrin and bisphenol A. The first step is the base-catalyzed *coupling* of bisphenol A and epichlorohydrin to yield a chlorohydrin.

Bases that may be used to catalyze this step include sodium hydroxide, lithium salts, and quaternary ammonium salts. Dehydrohalogenation of the chlorohydrin intermediate with a stoichiometric amount of base affords the glycidyl ether. Manufacturing processes can be divided into two broad categories according to the type of catalyst used to couple epichlorohydrin and bisphenol A [17, 18].

5.1. Caustic Coupling Process

In this process, caustic is used as a catalyst for the nucleophilic ring-opening (coupling reaction) of the epoxide group on the primary carbon atom of epichlorohydrin by the phenolic hydroxyl group and as a dehydrochlorinating agent for conversion of the chlorohydrin to the epoxide group (see above).

In caustic coupling processes, caustic (20–50 % sodium hydroxide in water) is slowly added to an agitated mixture of epichlorohydrin and bisphenol A. The highly exothermic coupling reaction proceeds during the initial stages. As the coupling reaction nears completion, dehydrochlorination becomes the predominant reaction. A high ratio (usually 10:1) of epichlorohydrin/bisphenol A is charged to the reactor to maximize the yield of monomeric ($n = 0$) DGEBA. At a 10:1 level of epichlorohydrin/bisphenol A, the $n = 0$ monomer comprises > 85 % of the reaction product mixture.

5.2. Phase-Transfer Catalyst Process

Alternatively, the coupling reaction and dehydrochlorination can be performed separately by using phase-transfer coupling catalysts, such as quaternary ammonium salts [19], which are not strong enough bases to promote dehydrochlorination. Once the coupling reaction is completed, caustic is added to carry out the dehydrochlorination step. Higher yields of the $n = 0$ monomer (> 90 %) are readily available via this method.

Many variations of these two basic processes are described in process patents [20, 21], including the use of co-solvents and azeotropic removal of water to facilitate the reactions and to minimize undesirablebyproducts such as insoluble polymers. The original batch methods have been modified to allow for continuous or semicontinuous production. New developments have been focused on improving manufacturing yield and resin purity.

The description of liquid DGEBA resins presented so far is oversimplified. In reality, *side reactions* result in the formation of low levels of impurities that both decrease the epoxide content from the theoretical amount of 2 per molecule and affect the resins properties, both before and after curing [22]. The five common side reactions are as follows:

1) *Hydrolysis of epoxy groups.* Unavoidable hydrolysis of the epoxide ring gives a small amount (0.1–5 %) of monohydrolyzed resin (MHR) or α-glycol. It has been reported that dispersability of pigments are enhanced and rates of epoxy resin curing with diamines can be dramatically increased by higher levels of MHR [23].

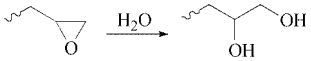

α-Glycol formation

2) *Incomplete dehydrochlorination* results in residual saponifiable or hydrolyzable chloride:

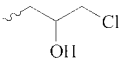

Incomplete dehydrochlorination increases the level of hydrolyzable chloride in the resin, which affects its suitablity for applications requiring superior electrical properties. In addition, hydrolyzable chlorides can affect reactivity by neutralizing basic catalysts such as tertiary amines. Many formulators adjust their formulations according to resin hydrolyzable chloride content. Typical hydrolyzable chloride contents of LERs range from < 100 ppm for electronic grade resin to 200–1000 ppm for standard grade resins.

3) *Abnormal addition of epichlorohydrin,* i.e., abnormal phenoxide attack at the central carbon of epichlorohydrin results in an end group that is more difficult to dehydrochlorinate:

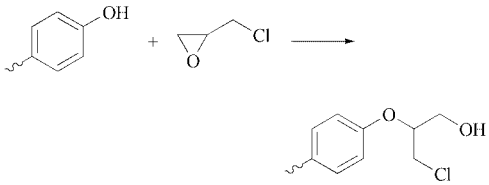

4) *Formation of bound chlorides* by reaction of epichlorohydrin with hydroxy groups in the polymer backbones:

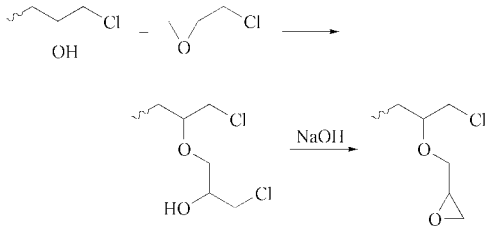

The bound chloride is not readily saponified with metal hydroxide solutions and is analyzed as part of the total chloride of the resin. Typical total chlorides values are 1000–2000 ppm.

5) *Higher oligomer formation.* Reaction of a phenolic terminal group with another epoxy resin molecule instead of an epichlorohydrin molecule gives epoxy resins with broader oligomer distribution and increased viscosity (n = 1 and higher oligomers). Typical LERs contain 5–15 % of the higher oligomers, mostly n = 1 and n = 2 compounds.

Pure DGEBA is a solid melting at 43°C. The unmodified commercial liquid resins are supercooled liquids with the potential for crystallization, depending on purity and storage conditions. This causes handling problems, particularly for ambient cure applications. Addition of certain reactive diluents and fillers can either accelerate or retard crystallization. Crystallization-resistant, modified resins are available. A crystallized resin can be restored to its liquid form by warming.

6. Solid Epoxy Resins Based on DGEBA

High molecular weight (MW) SERs based on DGEBA are characterized by a repeat unit con-

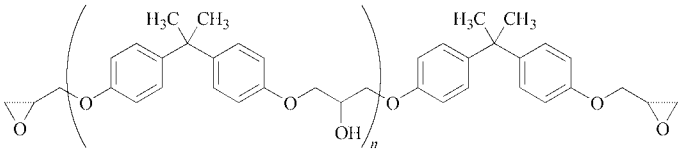

taining a secondary hydroxyl group with degrees of polymerization, i.e., *n* values ranging from 2 to about 35 in commercial resins; two terminal epoxy groups are theoretically present (see above).

The epoxy industry has adopted a common nomenclature to describe the SERs. They are called type "1," "2" up to type "10" resins, which correspond to the increased values of *n*, the degree of polymerization, EEW, MW, and viscosity. Examples of SERs are D.E.R. 661, 662, 664, 667, 669 resins from Dow Chemical, and Epon 1001 to 1009 series from Resolution. A comparison of some key properties of LERs and SERs is shown in Table 5.

SERs based on DGEBA are widely used in the coatings industry. The longer backbones give more distance between cross-links when cross-linked through the terminal epoxy groups, resulting in improved flexibility and toughness. Furthermore, the resins can also be cured through the multiple hydroxyl groups along the backbones using cross-linkers such as phenol–formaldehyde resoles or isocyanates to create different network structures and performance.

SERs are prepared by two processes: the *taffy* process and the *advancement* or *fusion* process. The first is directly from epichlorohydrin, bisphenol A, and a stoichiometric amount of NaOH. This process is very similar to the *caustic coupling* process used to prepare liquid epoxy resins. Lower epichlorohydrin to bisphenol A ratios are used to promote formation of high MW resins. The term taffy is derived from the appearance of the advanced epoxy resin prior to its separation from water and precipitated salts.

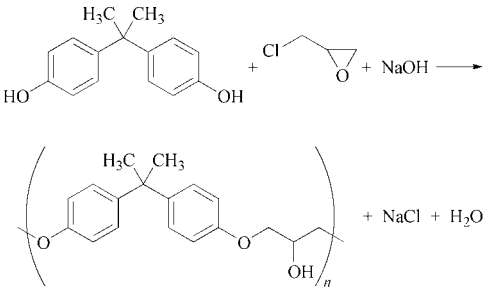

In the taffy process, a calculated excess of epichlorohydrin governs the degree of polymerization. However, preparation of the higher molecular weight species is subject to practical limitations of handling and agitation of highly viscous materials. The effect of epichlorohydrin–bisphenol A (ECH–BPA) ratio for a series of solid resins is shown in Table 6.

Table 6. Effect of epichlorohydrin–bisphenol a ratio on resin properties of taffy SERs

Mole ratio ECH/BPA	EEW	Softening point, °C
1.57:1.0	450–525	65–75
1.22:1.0	870–1025	95–105
1.15:1.0	1650–2050	125–135
1.11:1.0	2400–4000	145–155

In commercial practice, the *taffy* method is used to prepare lower MW solid resins, i.e., those with maximum EEW values of about 1000 (type "4"). Upon completion of the polymerization, the mixture consists of an alkaline brine solution and a water–resin emulsion. The product is recovered by separating the phases, washing the taffy resin with water, and removing the water under vacuum. One disadvantage of the taffy process is the formation of insoluble polymers, which create handling and disposal problems. Only a few epoxy producers currently manufacture SERs using the taffy process. A detailed description of a taffy procedure follows [24].

A mixture of bisphenol A (228 parts by weight) and 10 % aqueous sodium hydroxide solution (75 parts by weight) is introduced into a reactor equipped with a powerful agitator. The mixture is heated to ca 45°C and epichlorohydrin (145 parts by weight) is added rapidly with agitation, giving off heat. The temperature is allowed to rise to 95°C, where it is maintained about 80 min. for completion of the reaction. Agitation is stopped, and the mixture separates into two layers. The heavier aqueous layer is drawn off and the molten, taffy-like product is washed with hot water until the wash water is neutral. The taffy-like product is dried at 130°C, giving

Table 5. DGEBA-based epoxy resins

Resin type	*n* value[a]	EEW	Mettler softening point, °C	Molecular weight (M_w)[b]	Viscosity at 25°C, mPa · s (= cP)
Low viscosity LER	< 0.1	172–176		~ 350	4,000–6,000
Medium viscosity LER	~ 0.1	176–185		~ 370	7,000–10,000
Standard grade LER	~ 0.2	185–195		~ 380	11,000–16,000
Type 1 SER	~ 2	450–560	70–85	~ 1,500	160 – 250[c]
Type 4 SER	~ 5	800–950	95–110	~ 3,000	450 – 600[c]
Type 7 SER	~ 15	1,600–2,500	120–140	~ 10,000	1,500–3,000[c]
Type 9 SER	~ 25	2,500–4,000	145–160	~ 15,000	3,500–10,000[c]
Type 10 SER	~ 35	4,000–6,000	150–180	~ 20,000	10,000–40,000[c]
Phenoxy resin	~ 100	> 20,000	> 200	> 40,000	

[a] *n* value is the number-average degree of polymerization which approximates the repeating units and the hydroxyl functionality of the resin.
[b] Molecular weight is weight average (M_w) measured by gel-permeation chromatography (GPC) using polystyrene standard.
[c] Viscosity of SERs is determined by kinematic method using 40 % solids in diethylene glycol monobutyl ether solution.

a solid resin with a softening point of 70°C and an EEW of ca. 500.

Alternatively, epichlorohydrin and water are removed by distillation at temperatures up to 180°C under vacuum. The crude resin/salt mixture is then dissolved in a secondary solvent to facilitate water washing and salt removal. The secondary solvent is then removed via vacuum distillation to obtain the taffy–resin product.

Resins produced by this process exhibit relatively high α-glycol values, i.e., ca 0.5 eq/kg, attributable to hydrolysis of epoxy groups in the aqueous phase. Although detracting from epoxide functionality, such groups act as accelerators for amine curing. Resins produced by the taffy process exhibit *n* values of 0, 1, 2, 3, etc, whereas resins produced by the *advancement* process (described below) exhibit mostly even-numbered *n* values because a difunctional phenol is added to a diglycidyl ether of a difunctional phenol.

An alternative method is the chain-extension reaction of liquid epoxy resin (crude DGEBA) with bisphenol A, often referred to as the advancement or fusion process [25] which requires an advancement catalyst:

The advancement process is more widely used in commercial practice. Isolation of the polymerized product is simpler, since removal of copious amounts of NaCl is unnecessary. The reaction can be carried out with or without solvents. Solution advancement is widely practiced by coatings producers to facilitate handling of the high MW, high viscosity epoxy resins used in many coating formulations. The degree of polymerization is dictated by the ratio of LER to bisphenol A; an excess of the former provides epoxy terminal groups. The actual MW attained depends on the purity of the starting materials, the type of solvents used, and the catalyst. Reactive monofunctional groups can be used as chain terminators to control MW and viscosity build.

The following formula can be used to calculate the relative amount of bisphenol A that must be reacted with epoxy resin to give an advanced epoxy resin of predetermined EEW:

$$\text{Bis A} = \frac{\text{EEW}_i^{-1} - \text{EEW}_f^{-1}}{\text{EEW}_i^{-1} + \text{PEW}^{-1}}$$

where Bis A is the mass fraction of bisphenol A in the mixture prior to advancement, EEW_i is the

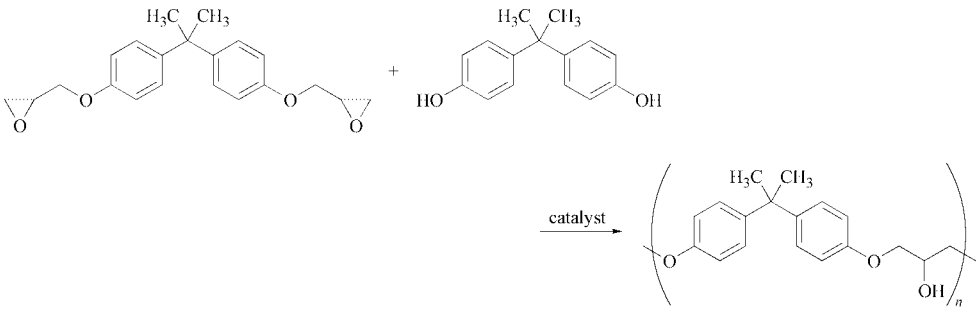

EEW of the epoxy resin that is to be advanced, EEW_f is the EEW of the advanced epoxy resin, and PEW is the phenol equivalent mass of the bisphenol, which is 114 g per equivalent for bisphenol A.

In a typical advancement process, bisphenol A and a liquid DGEBA resin (175–185 EEW) are heated to ca 150–190°C in the presence of a catalyst and reacted (i.e., advanced) to form a high MW resin. The oligomerization is exothermic and proceeds rapidly to near completion. The exotherm temperatures are dependent upon the targeted EEW and the reaction mass. In the cases of higher MW resins such as type "7" and higher, exotherm temperatures of $> 200°C$ are routinely encountered.

Advancement reaction catalysts facilitate the rapid preparation of medium and high MW linear resins and control prominent side reactions inherent in epoxy resin preparations, e.g., chain branching due to addition of the secondary alcohol group generated in the chain-lengthening process to the epoxy group [26, 27]. Nuclear Magnetic Resonance (NMR) spectroscopy can be used to determine the extent of branching [28] (see above).

Conventional advancement catalysts include basic inorganic reagents, e.g., NaOH, KOH, Na_2CO_3, or LiOH, and amines and quaternary ammonium salts. One mechanism proposed for the basic catalysts involves proton abstraction of the phenolic compound as the initiation step (see below).

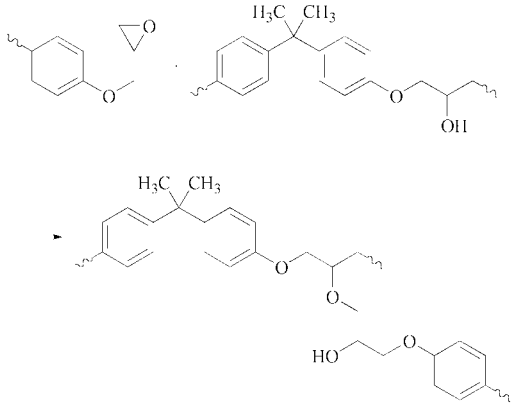

The phenoxide ion then attacks the epoxy ring, generating an alkoxide, which immediately abstracts a proton from another phenolic OH group. This is called the propagation step. Regeneration of the phenoxide ion repeats the cycle. The potential for side reactions increases after the phenolic OH groups have been consumed, particularly in melt (i.e., fusion) polymerization reactions.

One key disadvantage of catalysts based on inorganic bases and salts is the increased ionic

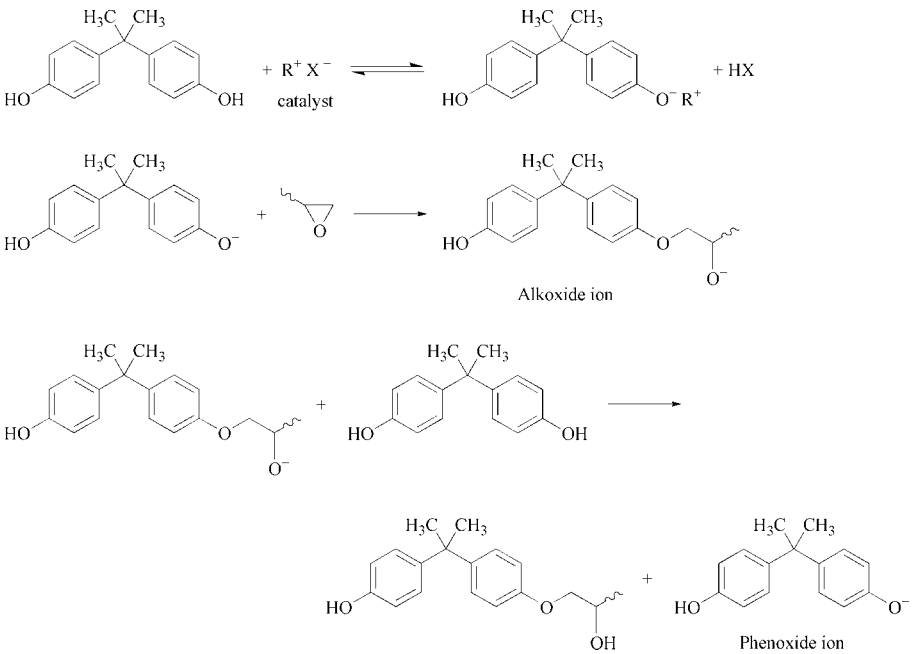

impurities added to the resin, which is not desirable in certain applications.

Imidazoles, substituted imidazoles, and triethanolamine have been patented as advancement catalysts [29]. However, most of the inorganic bases, salts, and amines produce resins with broad MW distribution and viscosity instability. This is due to poor catalyst selectivity and the continuing activity of the catalyst after completion of the advancement reaction.

Alternatively, a broad class of catalysts derived from aryl or alkyl phosphonium compounds were developed. Extensive patent literature claims a high order of selectivity [30, 31]. Selections of the phosphonium cation and counter ion have been shown to affect initiation rate, catalyst selectivity, catalyst lifetime, and, consequently, product quality and consistency. Some of the phosphonium salts are deactivated at high temperatures by the reaction exotherm, and are claimed to give better resin stability in terms of viscosity, EEW, and MW during the subsequent finishing steps [32 – 35].

Few mechanistic studies have been published on the selectivity of phosphonium compounds, but one publication describes the role of triphenylphosphine in advancement catalysis [36]. Nucleophilic attack by triphenylphosphine opens the epoxy ring, producing a betaine:

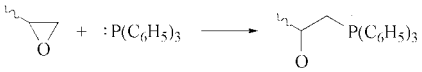

Proton abstraction from bisphenol A yields the phenoxide anion, forming a phosphonium salt. The phenoxide reacts with the electrophilic carbon attached to the positive phosphorus regenerating the catalyst:

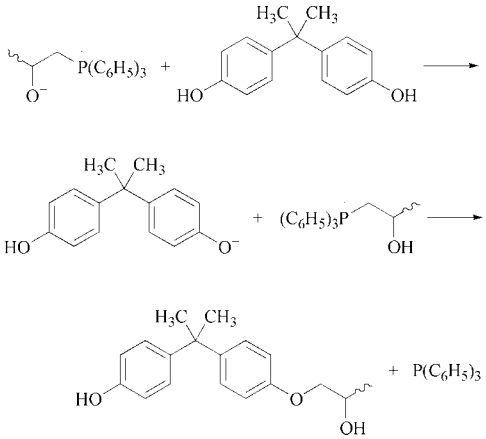

When the bisphenol A is consumed, the betaine decomposes into a terminal olefin and triphenylphosphine oxide:

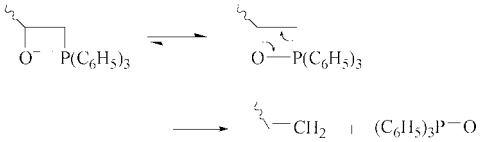

Branched epoxies [37] are prepared by advancing LER with bisphenol A in the presence of epoxy novolac resins. Such compositions exhibit enhanced thermal and solvent resistance.

SERs are available commercially in solid form or in solution. MW distributions of SERs have been examined by means of theoretical models and compared with experimental results [38]. Taffy-processed resins were compared with advancement-processed resins by gel-permeation chromatography (GPC) and high performance liquid chromatography (HPLC) [39] in conjunction with statistical calculations. The major differences are in the higher α-glycol content and the repeating units of oligomers. Resin viscosity and softening points are also lower with taffy resins. In addition, certain formulations based on taffy resins exhibit different behavior in pigment loading, formulation rheology, reactivity, and mechanical properties compared to those based on advancement resins.

6.1. SER Continuous Advancement Process

The recent literature review indicates efforts to develop continuous advancement processes to produce SERs. Companies seek to improve process efficiencies and product quality. One of the major deficiencies of the traditional batch advancement process is the long reaction time, resulting in EEW and viscosity drift, variable product quality, and gel formation. In addition, it is difficult to batch process higher MW, higher viscosity SERs such as types “9” and “10” resins. Shell patented several versions of the continuous resin advancement processes using modified reactor designs [40]. Dow Chemical received patents covering the uses of reactive extrusion [41] (REX) to produce SERs and other epoxy thermoplastic resins [42]. The latter process makes use of a self-wiping twin-screw ex-

truder. LER, bisphenol A, and catalyst are fed directly to the extruder to complete the resin advancement reaction in several minutes compared to the traditional several hours in a batch process. The process is claimed to be very efficient and is particularly suitable for the production of high molecular weight SERs, phenoxy resins, and epoxy thermoplastic resins. Compared to the traditional taffy processes used to produce phenoxy resins, the chemistry is salt-free, and the resins made via the REX process are fully converted in a matter of minutes, significantly reducing manufacturing costs. Additional benefits include reduced lot-to-lot variations in MW distribution, the flexibility to make small lots of varying molecular weights with minimal waste, and the ability to make custom resins with a variety of additives such as pigments and flow modifiers.

6.2. Phenoxy Resins

Phenoxy resins are thermoplastic polymers derived from bisphenol A and epichlorohydrin. Their weight-average molecular weights (M_w) are higher (i.e., > 30,000) than those of conventional SERs (i.e., 25,000 maximum). They lack terminal epoxides but have the same repeat unit as SERs and are classified as polyols or polyhydroxy ethers:

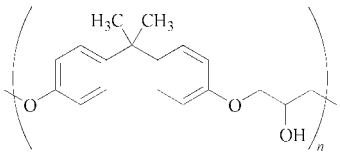

Phenoxy resins were originally developed and produced by Union Carbide (trade names PKHH, PKHC, PKHJ) using the taffy process. The process involves reaction of high purity bisphenol A with epichlorohydrin in a 1:1 mole ratio. Alternatively, phenoxy resins can be produced by the fusion process which uses high purity LER and bisphenol A in a 1:1 mole ratio. High purity monomers and high conversions are both needed to produce high MW phenoxy resins. The effects of monomer purity on phenoxy resin production are significant: monofunctional components limit MW, and functionality > 2 causes excess branching and increased polydispersity. Solution polymerization may be employed to achieve the MW and processability needed [43]. This, however, adds to the high costs of manufacturing of phenoxy resins, limiting their commercial applications.

The phenoxies are offered as solids, solutions, and waterborne dispersions. The majority of phenoxy resins are used as thermoplastics, but some are used as additives in thermoset formulations. Their high MW provide improved flexibility and abrasion resistance. Their primary uses are in automotive zinc-rich primers, metal can/drum coatings, magnet wire enamels, and magnetic tape coatings. However, the zinc-rich primers are being phased out in favor of galvanized steel by the automotive industry. Smaller volumes of phenoxy resins are used as flexibility or rheology modifiers in composites and electrical laminate applications, and as composite honeycomb impregnating resins. A new, emerging application is fiber sizing, which utilizes waterborne phenoxies. Literature references indicate their potential uses as compatiblizers for thermoplastic resins such as polyesters, nylons, and polycarbonates because of their high hydroxyl contents.

Current producers of phenoxy resins include the Phenoxy Specialties division of InChem Corp., Resolution, Huntsman, Tohto Kasei, and DIC.

6.3. Epoxy-Based Thermoplastics

Some of the new epoxy products developed in the past few years are the thermoplastic resins based on epoxy monomers. Polyhydroxy amino ether [44, 45] (PHAE) was commercialized by Dow Chemical in 1999 and trade named BLOX (Dow Chemical Co.). It is produced by the reaction of DGEBA with monoethanol amine using the reactive extrusion process. The high cohesive energy density of the resin gives it excellent gas-barrier properties against oxygen and carbon dioxide. It also possesses excellent adhesion to many substrates, optical clarity, excellent melt strength, and good mechanical properties. The product has been evaluated as a barrier resin for beer and beverage plastic bottles, as thermoplastic powder coatings, and as a toughener for starch-based foam [46]. Another epoxy thermoplastic resin under development by Dow is the polyhydroxy ester ether (PHEE). It is a reaction

product of DGEBA with difunctional acids. The ester linkage makes it suitable for biodegradable applications [47].

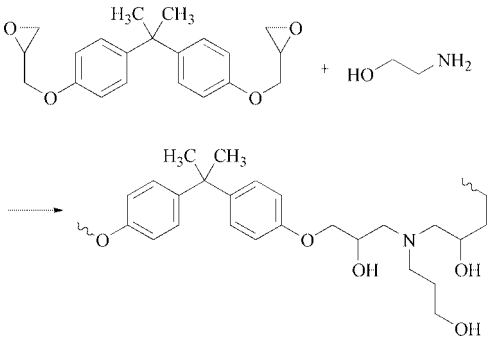

resin formulation. The diglycidyl ether of TBBA is produced via conventional liquid epoxy resin processes. Higher MW resins can be produced by advancing LERs or diglycidyl ether of TBBA with TBBA. The lower cost, advanced brominated epoxies based on LERs and TBBA containing ca. 20 wt. % Br are extensively employed in the PCB industry. The diglycidyl ether of TBBA (ca. 50 wt. % Br) is used for critical electrical/electronic encapsulation where high flame retardancy is required. Brominated epoxies are also used to produce epoxy vinyl esters for structural applications. Very high MW versions of brominated epoxies are used as flame-retardant additives to engineering thermoplastics used in computer housings.

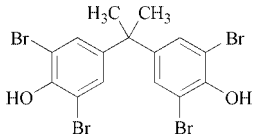

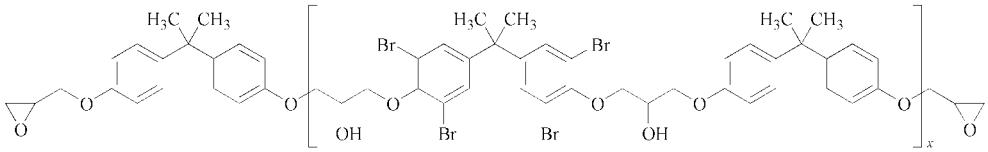

7. Halogenated Epoxy Resins

A number of halogenated epoxy resins have been developed and commercialized to meet specific application requirements. Chlorinated and brominated epoxies were evaluated for flame retardancy properties. The brominated epoxy resins were found to have the best combination of cost/performance and were commercialized by Dow Chemical in the late 1960s.

7.1. Brominated Bisphenol A Based Epoxy Resins

Many applications of epoxy resins require the system to be ignition-resistant, e.g., electrical laminates for PCBs and certain structural composites. A common method of imparting this ignition resistance is the incorporation of tetrabromobisphenol A (TBBA), 2,2-bis(3,5-dibromophenyl)propane, or the diglycidyl ether of TBBA, 2,2-bis[3,5-dibromo-4-(2,3-epoxypropoxy)phenyl]propane, into the

In order to meet increased requirements of the PCB industry for higher glass-transition temperature(T_g), higher thermal decomposition temperature (T_d), and lower dielectric constant (D_k) products, a number of new epoxy resins have been developed [48, 49].

7.2. Fluorinated Epoxy Resins

Fluorinated epoxy resins have been researched for a number of years for high performance end-use applications [50]. Fluorinated epoxies are highly resistant to chemical and physical abuse and should prove useful in high performance applications, including specialty coatings and composites, where their high cost may be offset by their special properties and long service life. The following fluorinated diglycidyl ether, 5-heptafluoropropyl-1,3-bis[2-(2,3-epoxypropoxy) hexafluoro-2-propyl] benzene, illustrates an example of fluoroepoxy resins [51] under development.

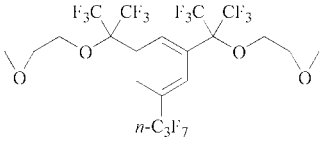

This resin is a viscous, colorless liquid (*bp* 118°C at 20 Pa · s) that contains 52 wt. % fluorine. It has a low surface tension, which makes it a superior wetting agent for glass fibers. The reactivity of this resin with amine or anhydride curing agents is comparable to epoxy resins based on bisphenol A and results in a thermoset that has a low affinity for water and excellent chemical resistance. Another fluorinated epoxy resin derived from hexafluorobisphenol A was introduced to the marketplace aiming at the anticorrosion coatings market for industrial vessels and pipes. The key disadvantages of fluorinated epoxies are their relatively high costs and low T_g, which limit their commercialization [52].

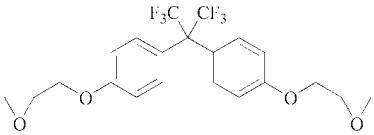

8. Multifunctional Epoxy Resins

The multifunctionality of these resins provides higher cross-linking density, leading to improved thermal and chemical resistance properties over bisphenol A epoxies.

8.1. Epoxy Novolac Resins

Epoxy novolacs are multifunctional epoxies based on phenolic formaldehyde novolacs. Both epoxy phenol novolac resins (EPN) and epoxy cresol novolac resins (ECN) have attained commercial importance [53]. The former is made by epoxidation of the phenol–formaldehyde condensates (novolacs) obtained from acid-catalyzed condensation of phenol and formaldehyde (see → Phenolic Resins). This produces random *ortho*- and *para*-methylene bridges.

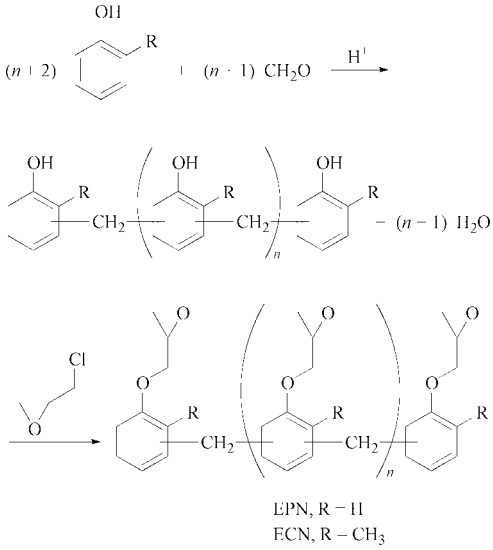

An increase in the molecular weight of the novolac increases the functionality of the resin. This is accomplished by changing the phenol or cresol to formaldehyde ratio. Epoxidation with an excess of epichlorohydrin minimizes the reaction of the phenolic OH groups with epoxidized phenolic groups and prevents branching. The epoxidation is similar to the procedure described for bisphenol A. EPN resins range from a high viscosity liquid of $n = 0.2$ to a solid of $n > 3$. The epoxy functionality is between 2.2 and 3.8. Properties of epoxy phenol novolacs are given in Table 7. When cured with aromatic amines such as methylenedianiline, the heat distortion temperatures (HDT) of EPN-based thermosets range from 150°C to 200°C, depending on cure and postcure schedules

Curing agents that give the optimum in elevated temperature properties for epoxy novolacs are those with good high temperature performance, such as aromatic amines, catalytic curing agents, phenolics, and some anhydrides. When cured with polyamide or aliphatic polyamines and their adducts, epoxy novolacs show improvement over bisphenol A epoxies, but the critical performance of each cure is limited by the performance of the curing agent.

The improved thermal stability of EPN-based thermosets is useful in elevated temperature services, such as aerospace composites. Filament-wound pipe and storage tanks, liners for pumps and other chemical process equipment, and corrosion-resistant coatings are typical applica-

Table 7. Typical properties of epoxy phenol novolacs

Property	D.E.N. 431,[a] EPN 1139[b]	D.E.N. 438,[a] EPN 1138[b]	D.E.N. 439[a]
n	0.2	1.6	1.8
EEW[c]	175	178	200
Viscosity, mPa · s (= cP)	1,400[d]	35,000[d]	3,000[e]
Softening point[f]			53
Color, Gardner	1	2	2

[a] The Dow Chemical Co.
[b] Huntsman.
[c] Epoxide equivalent weight.
[d] Temperature of measurement = 52°C.
[e] Temperature of measurement = 100°C.
[f] Durran's mercury method.

tions which take advantage of the chemical resistant properties of EPN resins. However, the high cross-link density of EPN-based thermosets can result in increased brittleness and reduced toughness.

8.1.1. Bisphenol F Epoxy Resin

The lowest MW member of the phenol novolacs is bisphenol F, which is prepared with a large excess of phenol to formaldehyde; a mixture of *o*,*o*′, *o*,*p*′, and *p*,*p*′ isomers is obtained (see below).

Epoxidation yields a liquid bisphenol F epoxy resin with a viscosity of 4000–6000 mPa · s (= cP), an EEW of 165, and $n \cong 0.15$.

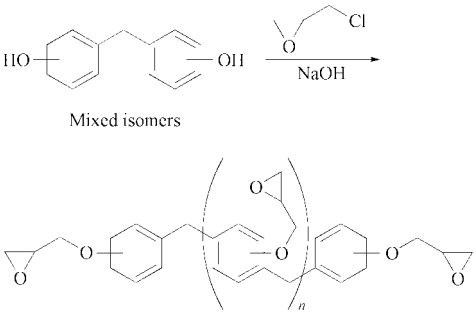

Bisphenol F resin
(**3**)

This unmodified, low viscosity liquid resin exhibits slightly higher functionality than unmodified bisphenol A liquid resins. Crystallization, often a problem with liquid bisphenol A resins, is reduced with bisphenol F resin. Consequently, noncrystallizing LERs which are blends of DGEBA and bisphenol F epoxy are available. Epoxy resins based on bisphenol F are used primarily as functional diluents in applications requiring a low viscosity, high performance resin system (e.g., solvent-free coatings). Higher filler levels and faster bubble release are possible because of the low viscosity. The higher epoxy content and functionality of bisphenol F epoxy resins provide improved chemical resistance compared to conventional bisphenol A epoxies. Bisphenol F epoxy resins are used in high solids, high build systems such as tank and pipe linings, industrial floors, road and bridge deck toppings, structural adhesives, grouts, coatings, and electrical varnishes.

8.1.2. Cresol Epoxy Novolacs

The *o*-cresol novolac epoxy resins (ECN) are analogous to phenol novolac resins. ECNs exhibit better formulated stability and lower moisture adsorption than EPNs, but have higher costs. ECN resins are widely used as base components in high performance electronic (semiconductors) and structural molding compounds, high temperature adhesives, castings and laminating systems, tooling, and powder coatings.

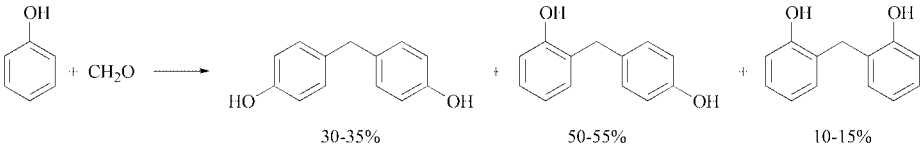

Increasing demands by the semiconductor industry has led to significant advances in ECN resin manufacturing technologies to reduce impurities, mainly the ionic content, hydrolyzable chlorides, and total chlorides. The use of polar, aprotic solvents, such as dimethyl sulfoxide (DMSO), as a co-solvent to facilitate chloride reduction has been patented [54]. Typical high purity ECN resins contain < 1000 ppm total chlorides and < 50 ppm hydrolyzable chlorides.

The melt viscosity of these resins, which are solids at room temperature, decreases sharply with increasing temperature (Table 8). This affords the formulator an excellent tool for controlling the flow of molding compounds and facilitating the incorporation of ECN resins into other epoxies, e.g., for powder coatings. While Ciba-Geigy was the first producer of ECN resins, many Japanese companies (Nippon Kayaku, Sumitomo Chemical, DIC, and Tohto Kasei) supply the majority of high purity ECN resins for the semiconductor industry today. Other suppliers are based in Korea and Taiwan.

8.1.3. Glycidyl Ethers of Hydrocarbon Epoxy Novolacs

In response to the increased performance demands of the semiconductor industry, hydrocarbon epoxy novolacs (HENs) were developed by Dow Chemical in the 1980s. HENs exhibit a much lower affinity for water compared to cresol or phenol epoxy novolacs. This translates directly into increased electrical property retention, which is important in the reliability of an electronic device encapsulated in the resin. An epoxy resin that is typical of this class is based on the alkylation product of phenol and dicyclopentadiene [55] (n = 0.1), 2,5-bis[(2,3-epoxypropoxy) phenyl]octahydro-4,7-methano-5*H*-indene (272 EEW; softening point 85°C; η at 150°C 0.4 Pa · s). The product is available from Huntsman as TACTIX* 556. Similar products based on *o*-cresol are commercialized in Japan by DIC (EPICLON HP-7200L).

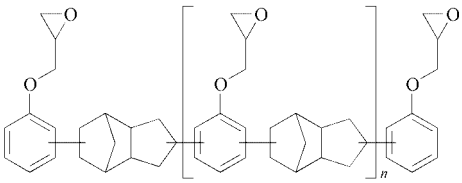

8.1.4. Bisphenol A Epoxy Novolacs

Bisphenol A novolacs are produced by reacting bisphenol A and formaldehyde with acid catalysts. Epoxidation of the bisphenol A novolacs gives bisphenol A epoxy novolac (BPAN) with improved thermal properties such as T_g, T_d of the epoxy-based electrical laminates.

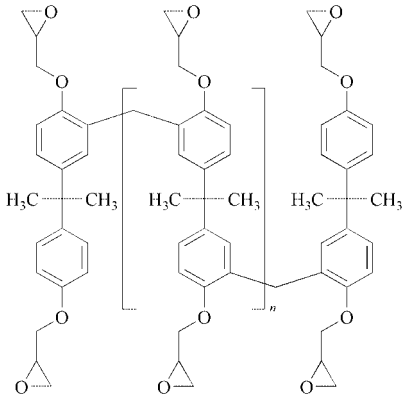

Bisphenol A epoxy novolac

8.2. Other Polynuclear Phenol Glycidyl Ether Derived Resins

In addition to the epoxy novolacs, there are other epoxy resins derived from phenol–aldehyde condensation products. New applications that require increased performance from the epoxy resin, particularly in the electronics, aerospace, and military industries, have made these types of resins more attractive despite their relatively high cost.

8.2.1. Glycidyl Ether of Tetrakis(4-hydroxyphenyl)ethane

One of the first polyfunctional resins to be marketed (by Shell) is based on 1,1,2,2-tetrakis[4-(2,3-epoxypropoxy)phenyl]ethane [56]. It is used primarily as additives in standard epoxy resin formulations for electrical laminates, molding compounds, and adhesives in which increased heat distortion temperature and

Table 8. Typical properties of epoxy cresol novolac resins[a]

Property	ECN 1235	ECN 1273	ECN 1280	ECN 1299
Molecular weight	540	1080	1170	1270
EEW[b]	200	225	229	235
Softening point, °C	35	73	80	99
Epoxide functionality	2.7	4.8	5.1	5.4

[a] Huntsman.
[b] Epoxide equivalent weight.

improved chemical resistance are desired. Tetrakis(4-hydroxyphenyl)ethane is prepared by reaction of glyoxal with phenol in the presence of HCl. The tetraglycidyl ether (*mp* ca. 80°C, and an EEW of 185–208) possesses a theoretical epoxide functionality of 4.

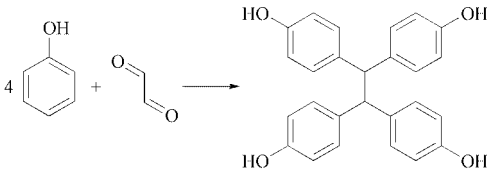

Glyoxal Tetrakis(4-hydroxyphenyl)ethane

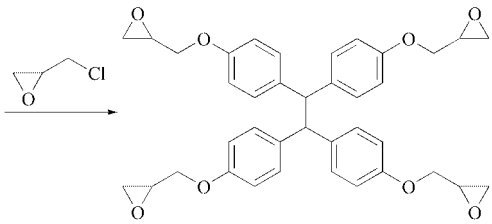

Tetraglycidyl ether of tetrakis(4-hydroxyphenyl)ethane

The commercial products Araldite 0163 (Huntsman) and Epon 1031 (Resolution) are tan-colored solids. They are widely used in high temperature resistance electrical laminates for high density PCBs and military applications. Their costs are typically higher than those of phenol and cresol epoxy novolacs.

8.2.2. Trisphenol Epoxy Novolacs

In the 1980s, new trifunctional epoxy resins based on tris[4-(2,3-epoxypropoxy)phenyl]methane isomers were introduced by Dow Chemical to help close the performance gap between phenol and cresol epoxy novolacs and high performance engineering thermoplastics [57]. These products were later sold to Ciba-Geigy and continued to be marketed under the TACTIX* 740 and XD 9053 trade names by Huntsman.

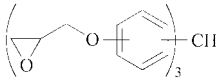

The resins are prepared via acid-catalyzed condensation of phenol and a hydroxybenzaldehyde, e.g., salicylaldehyde, to afford the trifunctional phenol, which is epoxidized with epichlorohydrin. These resins range from semisolids (162 EEW; Durran softening point 55°C; η at 60°C 11.5 Pa · s, at 150°C 0.055 Pa · s) to nonsintering solids (220 EEW; Durran softening point 85°C; η at 150°C 0.45 Pa · s).

The semisolid resins are used in advanced composites and adhesives where toughness, hot-wet strength, and resistance to high temperature oxidation are required. Their purity, formulated stability, fast reactivity, and retention of electrical properties over a broad temperature range make the solid resins suitable for use in the semiconductor molding powders industry. The trisphenol-based epoxies command significant high prices ($28–48/kg), limiting their uses.

8.3. Aromatic Glycidyl Amine Resins

Among the multifunctional epoxy resins containing an aromatic amine backbone, only a few have attained commercial significance. Their higher costs limit their uses to critical applications where their costs are justified. Glycidyl amines contain internal tertiary amines in the resin backbone, hence their high reactivity. Epoxy resins with such built-in curing catalysts are less thermally stable than nitrogen-free multifunctional epoxy resins.

8.3.1. Triglycidyl Ether of *p*-Aminophenol

The triglycidyl derivative of *p*-aminophenol was originally developed by Union Carbide [58] and is currently marketed by Huntsman under the designation MY 0500 and 0510. Epoxidation of *p*-aminophenol is carried out with a large excess of epichlorohydrin under carefully controlled conditions, since the triglycidylated resin exhibits limited thermal stability and polymerizes vigorously under the influence of its tertiary amine moiety.

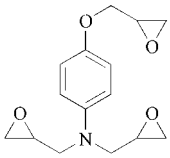

Triglycidyl-*p*-aminophenol

The resin exhibits a low viscosity, 2500–5000 mPa · s (= cP) at 25°C, and an EEW of 105–114; a molecularly distilled version (0510) has a viscosity of 550–850 mPa · s (= cP) at 25°C and an EEW of 95–107. It is considerably more reactive toward amines than standard bisphenol A resins. The trifunctional resin permits curing at low temperatures, i.e., 70°C, and rapidly develops excellent elevated-temperature properties. Used as additives to increase cure speed, heat resistance, and T_g of bisphenol A epoxy resins, it has utility in such diverse applications as high temperatures adhesives, tooling compounds, and laminating systems.

8.3.2. Tetraglycidyl Methylenedianiline (MDA)

These resins are used as binders in graphite-reinforced composites and are the binders of choice for many military applications. Epoxidation of MDA is carried out with stoichiometric excess of epichlorohydrin and under carefully controlled conditions to avoid rapid polymerization side reactions.

The tetrafunctional glycidylated MDA resins range in viscosity from 5000 to 25,000 mPa · s (= cP) at 50°C and have an EEW of 117–133; they are commercially available as Araldite MY 720 (Huntsman) and Epiclon 430 (DIC). When used in combination with the curing agent 4,4′-diaminodiphenylsulfone (DADS), it is the first system to meet the performance requirements set by the aerospace industry and is the standard against which other resin systems are judged [59]. Because of its outstanding properties, this resin is often used as the primary resin in high heat resistance formulations for military applications, despite its high costs (~ $22/kg). Among its attributes are excellent mechanical strength, high T_g, good chemical resistance, and radiation stability.

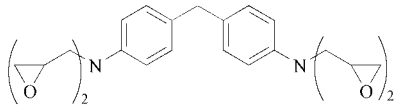

N,N,N',N'-Tetraglycidyl-4,4'-diaminodiphenylmethane

Another commercially important aromatic glycidyl amine resin is triglycidyl isocyanurate (TGIC), which is discussed in the "Weatherable Epoxy Resins" section.

9. Specialty Epoxy Resins

9.1. Crystalline Epoxy Resins Development

A number of new epoxy resins used in epoxy molding compounds (EMC) have been developed by Japanese resin producers in response to the increased performance requirements of the semiconductor industry. Most notable are the commercialization of crystalline epoxies based on biphenol by Yuka-Shell [60]:

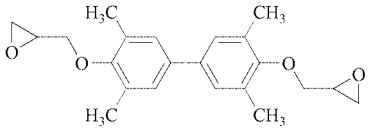

Diglycidyl ether of tetramethyl biphenol

The very low viscosity of these crystalline, solid epoxies when molten allows very high filler loading (up to 90 wt. %) for molding compounds. The high filler loading reduces the coefficient of thermal expansion (CET) and helps manage thermal shock and moisture and crack resistance of molding compounds used in new, demanding semiconductor manufacturing processes such as Surface Mount Technology (SMT). It should be noted that cured thermosets derived from these crystalline resins do

not retain crystallinity. Recently, a number of capacity expansions were announced for biphenol epoxies (sold as YX-4000 resin by Japan Epoxy Resins Corporation, formerly Yuka-Shell). DIC has developed dihydroxy naphthalene based epoxies [61] as the next generation product for this high performance market. Prices for crystalline epoxies are generally high ($22–26/kg), limiting their uses to high end applications.

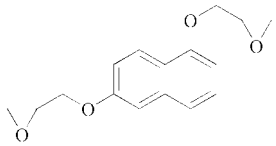

Diglycidyl ether of 1,5-dihydroxy naphthalene

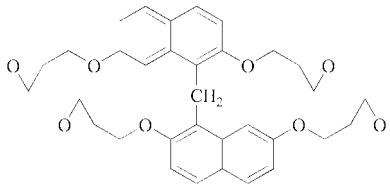

Tetraglycidyl ether based on dihydroxy naphthalene

Dow Chemical developed liquid crystalline polymers (LCP) based on diglycidyl ether of 4-4′-dihydroxy-α-methylstilbene in the 1980s [62, 63]. Liquid crystal thermoplastics and thermosets based on this novel chemistry showed excellent combinations of thermal, mechanical, and chemical properties, unachievable with traditional epoxies. However, commercialization of these products has not materialized.

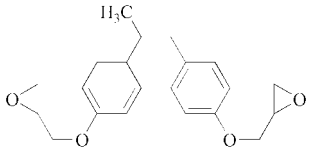

Diglycidyl ether of 4,4'-dihydroxy-α-methylstilbene

9.2. Weatherable Epoxy Resins

One of the major deficiencies of the aromatic epoxies is their poor weatherability, attributable to the aromatic ether segment of the backbone, which is highly susceptible to photoinitiated free-radical degradation. The aromatic ether of bisphenol A absorbs UV lights up to about 310 nm and undergoes photocleavage directly. This in turn produces free radicals that lead to oxidative degradation of bisphenol A epoxies, resulting in chalking. Numerous efforts have been devoted to remedy this issue, resulting in a number of new weatherable epoxy products. However, their commercial success has been limited, primarily because of higher resin costs and the fact that end users can topcoat epoxy primers with weatherable coatings based on other chemistries such as polyesters, polyurethanes, or acrylics. The following epoxy products when formulated with appropriate reactants can provide certain outdoor weatherability.

9.2.1. Hydrogenated DGEBA

In 1976, Shell Chemical Co. introduced epoxy resins based on the diglycidyl ether of hydrogenated bisphenol A, 2,2-bis[4-(2,3-epoxypropoxy)cyclohexyl]propane (232–238 EEW; η at 25°C 2–2.5 Pa · s).

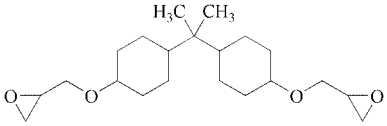

These resins resist yellowing and chalking because of their aliphatic structure. Epoxy resins based on hydrogenated bisphenol A are made by the epoxidation of the saturated diol, 2,2-bis(4-hydroxycyclohexyl)propane or 2,2-bis(4-hydroxycyclohexyl)propane with epichlorohydrin, or by the hydrogenation of a low molecular weight DGEBA resin [64]. Commercially available products include an Epalloy 5000 resin from CVC. One disadvantage is their much higher costs, and consequently, the products have not found broad acceptance in the industry. Furthermore, cross-linked networks based on hydrogenated bisphenol A epoxies lose some of the characteristic temperature and chemical resistances inherent with the bisphenol A backbone.

9.2.2. Heterocyclic Glycidyl Imides and Amides

In the 1960s, considerable work was devoted to preparing triglycidyl isocyanurate, 1,3,5-tris(2,3-epoxypropyl)-1,3,5-perhydrotriazine-2,4,6-trione [65]. The epoxidation of cyanuric

acid with epichlorohydrin gives triglycidyl isocyanurate (TGIC), marketed as PT 810 by Huntsman. It is a crystalline compound (*mp* 85–110°C) with an EEW of ca. 108. Miscibility with organic compounds is limited. Because of its excellent weatherability, TGIC is widely used in outdoor powder coatings with polyesters [66], despite its higher cost (∼ $12/kg).

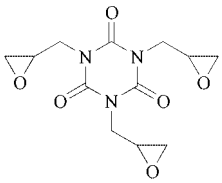

Triglycidyl isocyanurate

9.2.3. Hydantoin-Based Epoxy Resins

These resins were commercialized by Ciba-Geigy. Hydantoins are prepared from carbon dioxide, ammonia, hydrogen cyanide, and ketones via the Bucherer reaction and can be epoxidized with epichlorohydrin [67]. Cured and uncured resin properties depend greatly on the nature of the substituents R and R′. The hydantoin derived from acetone furnishes a low viscosity, water-dispersable epoxy resin, 5,5-dimethyl-1,3-bis(2,3-epoxypropyl)-2,4-imidazolidinedione (R = R′ = CH_3; 145 EEM; η at 25°C 2.5 Pa · s). A nonsintering solid epoxy resin is obtained if R = R′ = $-(CH_2)_5-$.

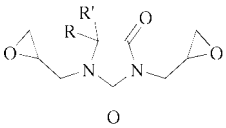

When cured with aromatic amines or anhydrides, these resins show high heat distortion temperatures and excellent adhesion and weatherability. A variety of applications are suggested for these new resins, particularly in applications in which a non-yellowing epoxy resin is desirable.

9.3. Elastomer-Modified Epoxies

Epoxy thermosets derive their thermal, chemical, and mechanical properties from the highly cross-linked networks. Consequently, toughness deficiency is an issue in certain applications. To improve the impact resistance and toughness of epoxy systems, elastomers such as BF Goodrich's CTBN rubbers (carboxyl-terminated butadiene nitrile) are often used as additives or pre-reacted with epoxy resins [68]. Most commonly used products are reaction adducts of liquid epoxy resins (DGEBA) with CTBN in concentrations ranging from 5 wt. % to 50 wt. %. They have been shown to give improved toughness, peel adhesion, and low temperature flexibility over unmodified epoxies. Primary applications are adhesives for aerospace and automotive and as additives to epoxy vinyl esters for structural composites. Formation of adducts of epoxy resins and CTBN is promoted by triphenylphosphine or alkyl phosphonium salts. Other elastomers used to modify epoxies include amine-terminated butadiene nitrile (ATBN), maleated polybutadiene and butadiene–styrene, epoxy-terminated urethane prepolymers, epoxy-terminated polysulfide, epoxy–acrylated urethane, and epoxidized polybutadiene.

10. Monofunctional Glycidyl Ethers and Aliphatic Glycidyl Ethers

A number of low MW monofunctional, difunctional, and mutifunctional epoxies are used as reactive diluents, viscosity reducers, flexiblizers, and adhesion promoters. Recent trends toward lower VOC, higher solids and 100 % solids epoxy formulations have resulted in increased utilization of these products. Most of these epoxies are derived from relatively compact hydroxyl-containing compounds, such as alcohols, glycols, phenols, and epichlorohydrin. Epoxidized vegetable oils, such as epoxidized linseed oils, are also used as reactive diluents. They are produced using a peroxidation process and are discussed in more detail in the cycloaliphatic epoxies and epoxidized vegetable oils section. Typically, these products have very low viscosity (1–70 cP at 20°C) relative to LERs (11,000–16,000 cP). They are often used in the range of 7–20 wt. % to reduce viscosity of the diluted system to 1000 cP. However, the uses of reactive diluents, especially at high levels, often result in decreased chemical resistance and

thermal and mechanical properties of the cured epoxies.

Important products include butyl glycidyl ether (BGE), alkyl glycidyl ethers of C8–C10 (Epoxide 7) and C12–C14 (Epoxide 8), *o*-cresol glycidyl ether (CGE), *p-tert*-butyl glycidyl ether, resorcinol diglycidyl ether (RDGE), and neopentyl glycol diglycidyl ether (Table 9). While BGE is the most efficient viscosity reducer and has been widely used in the industry for many years, it has been losing market share because of its volatility and obnoxiousness. Phenyl glicidyl ether (PGE) is no longer used by many formulators because of its toxicity. The industry trend is moving toward longer chain epoxies such as Epoxide 8 or neopentyl glycol diglycidyl ether.

Major suppliers of these products are Resolution, Air Products, Ciba Specialty Chemicals, Huntsman, CVC Specialty Chemicals, Pacific Epoxy Polymers, and Exxon.

Table 9. Some common commercial glycidyl ether reactive diluents

Name	Structure
n-Butyl glycidyl ether	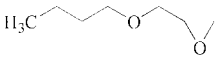
C12–C14 Aliphatic glycidyl ether	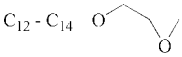
o-Cresol glycidyl ether	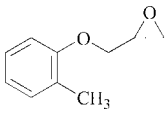
Neopentylglycol diglycidyl ether	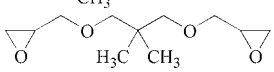
Butanediol diglycidyl ether	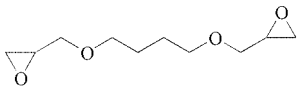

An example of multifunctional aliphatic epoxies is the triglycidyl ether of propoxylated glycerine (Heloxy 84) from Resolution. A similar product is based on epoxidized castor oil (Heloxy 505). These products are used primarily as viscosity reducers while increasing functionality and cross-linking density of the cured systems.

Epoxy resins based on long-chain diols, such as the diglycidyl ether of polypropylene glycol [α,ω-bis(2,3-epoxypropoxy)poly-(oxypropylene)] (305–335 EEW; η at 25°C 0.055–0.10 Pa · s), are used as flexibilizing agents to increase a thermoset's elongation and impact resistance. Because of the low reactivity of the aliphatic diols toward epichlorohydrin, these epoxies are produced by first coupling the diols with epichlorohydrin using phase-transfer catalysts such as ammonium salts or Lewis acid catalysts (boron trifluoride, stannic chloride), followed by epoxidation with caustic [69, 70]. A prominent side-reaction is the conversion of aliphatic hydroxyl groups formed by the initial reaction into chloromethyl groups by epichlorohydrin. The resultant epoxy resins are known to have lower reactivity toward conventional amine curing agents relative to bisphenol A epoxies. Dow Chemical manufactures D.E.R. 732 and D.E.R. 736 aliphatic epoxy resins. They are derived from polyglycols with different chain lengths.

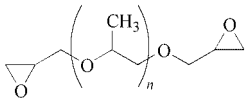

Polyglycidyl ethers of sorbitol, glycerol, and pentaerythritol are used as adhesion promoters for polyester tire cords. Their high chloride content improves adhesion to rubber.

11. Cycloaliphatic Epoxy Resins and Epoxidized Vegetable Oils

Resins based on the diepoxides of cycloaliphatic dienes were first commercialized in the 1950s by Union Carbide Corp. The combination of aliphatic backbone, high oxirane content, and no halogens gives resins with low viscosity, weatherability, low dielectric constant, and high cured T_g. This class of epoxy is popular for diverse end uses including auto topcoats, weatherable high voltage insulators, UV coatings, acid scavengers, and encapsulants for both electronics and optoelectronics. A comparison of some properties of two common aliphatic epoxies with those of LER (DGEBA) is shown in Table 10.

Table 10. Comparative Viscosities of cycloaliphatic epoxies, epoxidized oils, and DGEBA

Epoxy type	Viscosity, mPa·s(=cP)	EEW
Cyclo diepoxy ERL-4221[a] (3′,4′-epoxycyclohexylmethyl-3,4-epoxycyclohexanecarboxylate)	400	135
DGEBA	11000	190
Linseed oil epoxy	730	168

[a] Trademark of the Dow Chemical Co.

The preferred industrial route to cycloaliphatic epoxy resins is based on the epoxidation of cycloolefins with peracids, particularly peracetic acid [71]. Few side reactions are encountered. Some properties of various commercial products are given in Table 11. The peracid cannot be made in situ because the cyclic olefins are sensitive to impurities generated in this process.

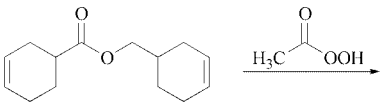

3'-Cyclohexenylmethyl 3-cyclohexenecarboxylate

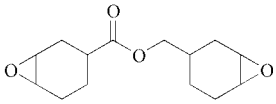

3',4'-Epoxycyclohexylmethyl 3,4-epoxycyclohexanecarboxylate

An important secondary reaction is the acid-catalyzed hydrolysis of the epoxide groups. The reaction is minimized at low temperatures and strongly depends on the constituents and the reaction medium.

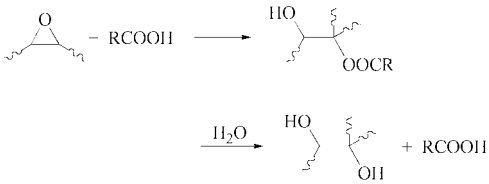

The cycloaliphatic epoxides are more susceptible to electrophilic attack because of the lower electronegativity of the cycloaliphatic ring relative to the bisphenol A aromatic ether group in DGEBA resins. Consequently, cycloaliphatic epoxies do not react well with conventional nucleophilic epoxy curing agents such as amines. They are commonly cured via thermal or UV-initiated cationic cures. In addition, cycloaliphatic epoxy resins are low viscosity liquids that can be thermally cured with anhydrides to yield thermosets having a high heat distortion temperature. They are often used as additives to improve performance of bisphenol A epoxies. Their higher prices ($6.60–8.80/kg) have limited their commercial applications to high end products.

The largest end uses of cycloaliphatic epoxies in order of volume are electrical, electronic components encapsulation, and radiation-curable inks and coatings. A potentially large volume application is UV-curable metal can coatings for beer can exterior and ends, but the market has not been growing significantly in recent years. Other uses include acid scavengers for vinyl-based transformer fluids and lubricating oils; filament winding for aerial booms and antennas; and as viscosity modifier for bisphenol A LERs in tooling compounds. An epoxy silicone containing cycloaliphatic epoxy end groups and a silicone backbone is used as radiation-curable release coatings for pressure-sensitive products. Dow Chemical is the largest producer of cycloaliphatic epoxies. Daicel of Japan has entered the cycloaliphatic epoxy resin market.

Epoxidation of α-olefins, unsaturated fatty acid esters, and glycerol esters is affected readily by peracids including in situ peracids generated from hydrogen peroxide and carboxylic acids.

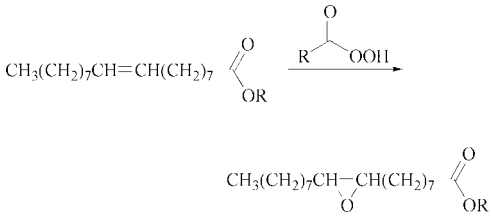

α-Olefin epoxies find utility as reactive diluents for coatings and as chemical intermediates for lubrication fluids. Larger volume epoxidized soybean and linseed oils are most frequently used as secondary plasticizers and co-stabilizers for PVC.

Table 11. Cycloaliphatic epoxy resins

Chemical name	Structure	Commercial products	EEW[a]	Viscosity, mPa · s (= cP) at 25°C[b]
3′,4′-Epoxycyclohexylmethyl-3,4-epoxycyclohexanecarboxylate	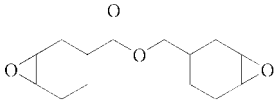	ERL-4221[c] UVR-6110 CY-179[d]	131–143	350–450
3,4-Epoxycyclohexyloxirane	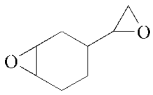	ERL-4206[c]	70–74	15
2–(3′,4′-Epoxycyclohexyl)-5,1″-spiro-3″,4″-epoxycyclohexane-1,3-dioxane	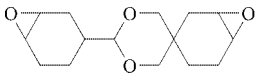	ERL-4234[c]	133–154	7,000–17,000[e]
Vinyl cyclohexene monoxide	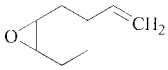	VCMX	124	5
3,4-Epoxycyclohexanecarboxylate methyl ester	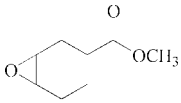	ERL-4140	156	6
Bis(3,4-epoxycyclohexylmethyl) adipate	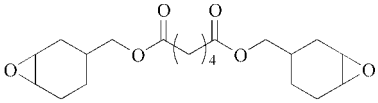	ERL-4299[c] UVR-6128	180–210	550–750

[a] Epoxide equivalent weight.
[b] Unless otherwise stated.
[c] Union Carbide, division of The Dow Chemical Co.
[d] Huntsman.
[e] At 38°C.

12. Epoxy Esters and Derivatives

12.1. Epoxy Esters

The esterification of epoxy resins with commercial fatty acids is a well-known process that has been employed for industrial coatings for many years. The carboxylic acids are esterified with the terminal epoxy groups or the pendant hydroxyls on the polymer chain.

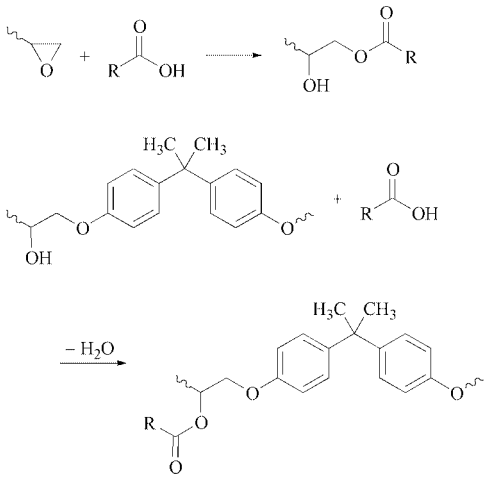

A wide variety of saturated and unsaturated fatty acids are utilized to confer properties useful in air-dried, protective, and decorative coatings. Typical fatty acids include tall oil fatty acids, linseed oil fatty acid, soya oil fatty acid and castor oil fatty acid. A medium molecular weight SER, a so-called 4-type, is commonly used. Catalysts such as alkaline metal salts (Na_2CO_3) or ammonium salts are essential to prevent chain branching and gelation caused by etherification of the epoxy groups.

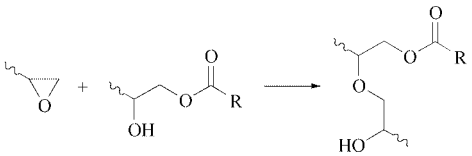

Esterification is generally conducted in an inert atmosphere at 225–260°C, with sparging to removebyproduct water. The course of the reaction is monitored by acid number to a specified end point and by viscosity. The product is then dissolved in a solvent [72].

Metallic driers are incorporated in unsaturated ester solutions to promote cure via air-drying, i.e., oxidative polymerization of the double bonds of the fatty acids. Chemical resistance is generally lower than that of unmodified epoxy resins cured at ambient temperatures with amine hardeners. Epoxy esters are also used to produce anodic electrodeposition (AED) coatings by further reaction with maleic anhydride followed by neutralization with amines to produce water-dispersable coatings. Epoxy esters were widely used as automotive primer-surfacer and metal can ends coatings for many years, but are being replaced by other technologies. Their high viscosity limited their uses in low solids, solventborne coatings. Waterborne epoxy esters are now available and are used in flexographic inks for milk cartons.

12.2. Glycidyl Esters

Glycidyl esters are prepared by the reaction of carboxylic acids with epichlorohydrin followed by dehydrochlorination with caustic:

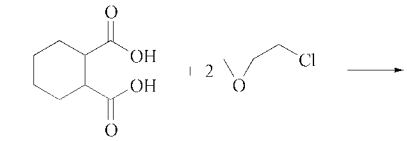

Hexahydrophthalic acid

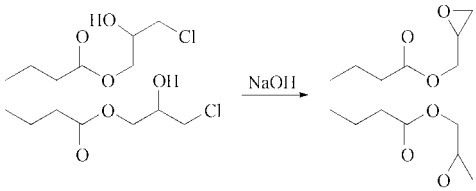

Diglycidyl ester of hexahydrophthalic acid

The viscosity of these esters is low, i.e., ca. 500 mPa · s (= cP), and their reactivity resembles that of bisphenol A resins. Similar epoxy resin derived from dimerized linoleic acid is also commercially available. They are often used as flexiblizing agents instead of epoxidized long-chain diols.

The *glycidyl ester of versatic acid* or neodecanoic acid is an example of high MW monoglycidyl aliphatic epoxy. The molecule is highly branched, thus providing steric effects to protect it from hydrolysis, resulting in good weatherability and water resistance. On the other hand, it suffers from high viscosity, and therefore it is not an effective diluent. It is often used to improve scrubability, chemical resistance, and weatherability of coatings. The product is commercially available from Resolution (Cardura E-10) and Exxon (Glydexx N-10).

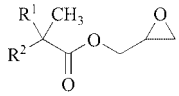

R^1, R^2 = alkyl or substituted alkyl
Glycidyl ester of versatic acid

A commercially important glycidyl ester is *glycidyl methacrylate (GMA)*, a dual functionality monomer, containing both a terminal epoxy and an acrylic C=C bond. It is produced by the reaction of methacrylic acid with epichlorohydrin. The dual functionality of GMA brings together desirable properties of both epoxies and acrylics, e.g., the weatherability of acrylics and chemical resistance of epoxies, in one product. GMA is useful as a comonomer in the synthesis of epoxy-containing polymers via free-radical

polymerization. The resultant epoxy-containing polymers can be further cross-linked. An example of such polymers is GMA acrylic, which is an acrylic copolymer, containing about 10–35 % by weight of GMA. Cure is by reaction with dodecanedioic acid. Its primary uses are in automotive powder coatings. GMA-containing polymers are also used as compatiblizers for engineering thermoplastics, in adhesives and latexes, and as rubber and asphalt modifiers [73]. Dow Chemical and Nippon Oils & Fats are two major producers of GMA.

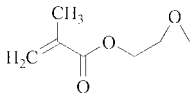

Glycidyl methacrylate (GMA)

12.3. Epoxy Acrylates

Epoxy resins are reacted with acrylic acid to form epoxy acrylate oligomers, curable via free-radical polymerization of the acrylate C=C bonds initiated by light [74]. UV lights are most commonly used, but electron beam (EB) curing is becoming more common because of its decreasing equipment costs. This is a fast-growing market segment for epoxy resins because of the environmental benefits of the UV cure technology: low to zero VOC, low energy requirements. Major applications include coatings for overprint varnishes, wood substrates, and plastics. Radiation-cured epoxy acrylates are also growing in importance in inks, adhesives, and photoresists applications. The 2001 global market for epoxy acrylates was estimated at 40,000 MT with an annual growth rate projected to be 8–10%.

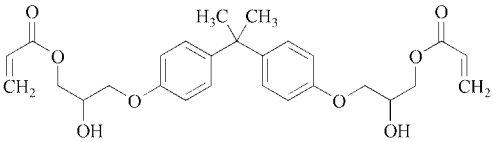

Epoxy acrylate based on DGEBA

Liquid epoxy resins such as DGEBA are most commonly used to produce epoxy acrylates. When higher thermal performance is required, multifunctional epoxies such as epoxy novolacs are used. Epoxy acrylates from epoxidized soybean oil and linseed oil are often used as blends with aromatic epoxy acrylates to reduce viscosity of the formulations. Major producers of UV-curable epoxy acrylates are UCB, Radcure, Dow Chemical, Sartomer, and Henkel.

12.4. Epoxy Vinyl Esters

A major derivative of epoxy resins is the epoxy vinyl ester resin. Originally developed by Dow Chemical [75, 76] and Shell Chemical in the 1970s, it is considered a high performance resin used in glass-reinforced structural composites, particularly for its outstanding chemical resistance and mechanical properties. The resins are made by reacting epoxy resins with methacrylic acid and diluted with styrene to 35–40 % solvent by weight. Liquid epoxy resins (DGEBA) are commonly used. Epoxy novolacs are used where higher thermal or solvent resistance is needed. Brominated epoxies are also used to impart flame retardancy for certain applications. In the final formulation, peroxide initiators are added to initiate the free-radical cure reactions of the methacrylic C=C bonds and styrene to form a random copolymer thermoset network.

The vinyl ester functionality of the epoxy vinyl esters provides outstanding hydrolysis and chemical resistance properties, in addition to the inherent thermal resistance and toughness properties of the epoxy backbone. These attributes have made epoxy vinyl esters a material of choice in demanding structural composite applications such as corrosive chemicals storage tanks, pipes, and ancillary equipment for chemical processing. Other applications include automotive valve covers and oil pans, boats, and pultruded construction parts. Significant efforts have been devoted to improve toughness and to reduce levels of styrene in epoxy vinyl ester formulations because of environmental concerns. In addition to Dow Chemical, other major suppliers of epoxy vinyl esters include Ashland, AOC, DSM, Interplastic, and Reichhold (DIC) (see top of next page).

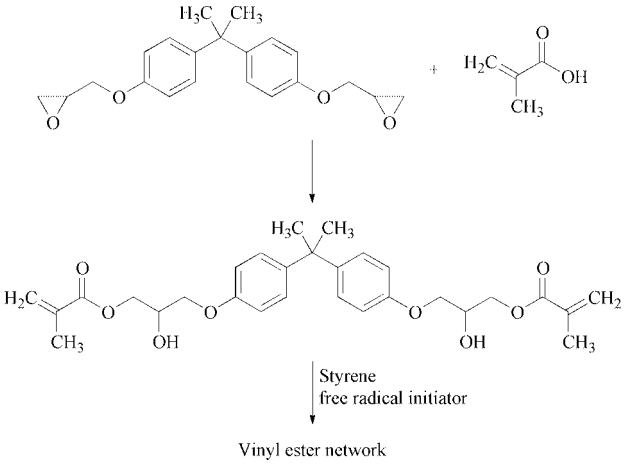

12.5. Epoxy Phosphate Esters

Dow Chemical developed the epoxy phosphate ester technology in the 1980s [77]. Epoxy phosphate esters are reaction products of epoxy resins with phosphoric acid. Depending on the stoichiometric ratio and reaction conditions, a mixture of mono-, di- and triesters of phosphoric acid are obtained. Subsequent hydrolysis of the esters is used as a way to control the distribution of phosphate and glycol end groups and product viscosity. Epoxy phosphate esters can be made to disperse in water to produce waterborne coatings. They are used primarily as modifiers to improve the adhesion property of nonepoxy binders in both solventborne and waterborne systems for container and coil coatings.

13. Characterization of Uncured Epoxies

Most industrial chemicals and polymers are not the 100 % pure, single chemicals as described in their general chemical structures. In the case of epoxy resins, they often contain isomers, oligomers, and other minor constituents. As a first requirement, one would need to know the epoxy content or EEW of the epoxy resin so the proper stoichiometric amount of cross-linker(s) can be calculated. However, a successful thermoset formulation must also have the proper reactivity, flow, and performance. Consequently, other epoxy resin properties are required by the formulators and supplied by the resin producers.

Liquid epoxy resins are mainly characterized by epoxy content, viscosity, color, density, hydrolyzable chloride, and volatile content [78]. Less often analyzed are α-glycol content, total chloride content, ionic chloride, and sodium. Solid epoxy resins are characterized by epoxy content, solution viscosity, melting point, color, and volatile content. Less often quoted are phenolic hydroxyl content, hydrolyzable chloride, ionic chloride, sodium, and esterification equivalent. Table 12 lists analytical methods adopted by ASTM [79] as standard testing methods for epoxy resins.

In addition, gel-permeation chromatography (GPC), high performance liquid chromatography (HPLC) [39, 80], and other analytical procedures such as nuclear magnetic resonance (NMR) [28] and infrared spectroscopy (IR) [81] are performed to determine MW, MW distribution, oligomer composition, functional groups, and impurities.

Resin components such as α-glycol content and chloride types and levels are known to influence formulation reactivity and rheology, depending on their interactions with the system composition such as basic catalysts (tertiary amines) and/or amine curing agents. Knowing the types and levels of chlorides guides formulators in the adjustment of their formulations for proper reactivity and flow.

Epoxide Equivalent Weight. The epoxy content of liquid resins is frequently expressed as epoxide equivalent weight (EEW) or weight per epoxide (WPE), which is defined as the weight in grams that contains 1 g equivalent of

Table 12. Uncured epoxy resin test methods

Test item	Unit	Condition	ASTM method
EEW			D1652-97
Viscosity, neat	cP[a]	25°C	D445-01
Viscosity, solution	cSt[b]	25°C	D445-01
Viscosity, melt	cSt[b]	150°C	D445-01
Viscosity, ICI Cone and Plate	Pa · s		D4287-00
Viscosity, Gardner–Holdt			D1545-98
Color, Co-Pt			D1209-00
Color, Gardner			D1544-98
Color, Gardner in solution			D1544-98
Moisture	ppm		E203-01
Softening point	°C		D3104-99

[a] cP = mPa · s.
[b] cSt = mm^2/s.

epoxide. A common chemical method of analysis for epoxy content of liquid resins and solid resins is titration of the epoxide ring by hydrogen bromide in acetic acid [82]. Direct titration to a crystal violet indicator end point gives excellent results with glycidyl ethers and cycloaliphatic epoxy resins. The epoxy content of glycidyl amines is determined by differential titration with perchloric acid. The amine content is first determined with perchloric acid. Addition of tetrabutylammonium iodide and additional perchloric acid generates hydrogen iodide, which reacts with the epoxy ring. The epoxy content is obtained by the second perchloric acid titration to a crystal violet end point.

In another procedure, a halogen acid is generated by the reaction of an ionic halide salt, e.g., tetraethylammonium bromide in acetic acid with perchloric acid with subsequent formation of a halohydrin; the epoxy group is determined by back-titration with perchloric acid using crystal violet indicator [83]. The end point can be determined visually or potentiometrically. A monograph on epoxide determinations was published in 1969 [84]. This is the method adopted by ASTM and is currently used by most resin producers.

Viscosity of epoxy resins is an important characteristic affecting handling, processing, and application of the formulations. For example, high viscosity LERs impede good mixing with curing agents, resulting in inhomogeneous mixtures, incomplete network formation, and poor performance. On the other hand, too low viscosity would affect application characteristics such as coverage and appearance.

Viscosities of liquid resins are typically determined with a Cannon–Fenske capillary viscometer at 25°C, or a Brookfield viscometer. The viscosity depends on the temperature, as illustrated in Figure 2. Viscosities of solid epoxy resins are determined in butyl carbitol (diethylene glycol monobutyl ether) solutions (40 % solids content) and by comparison with standard bubble tubes (Gardner–Holdt bubble viscosity). The Gardner color of the same resin solution is determined by comparison with a standard color disk. Recently, data have been reported for solid epoxy resins using the ICI Cone and Plate viscometers, which are much more time-efficient because they do not require sample dissolution.

Hydrolyzable chloride (HyCl) content of liquid and solid epoxy resins is determined by dehydrochlorination with potassium hydroxide solution under reflux conditions and potentiometric titration of the chloride liberated by silver nitrate. The solvent(s) employed and reflux conditions can influence the extent of dehydrochlorination and give different results. The "easily hydrolyzable" HyCl content, which reflects the degree of completion of the dehydrochlorination step in the epoxy resin manufacturing process, is routinely determined by a method using methanol and toluene as solvents. This is the method most commonly used to characterize LER and SER.

For epoxy resins used in electronic applications, such as cresol epoxy novolacs, more powerful polar aprotic solvents such as dioxane or dimethyl formamide (DMF) have been used to hydrolyze the difficult-to-hydrolyze Hy-

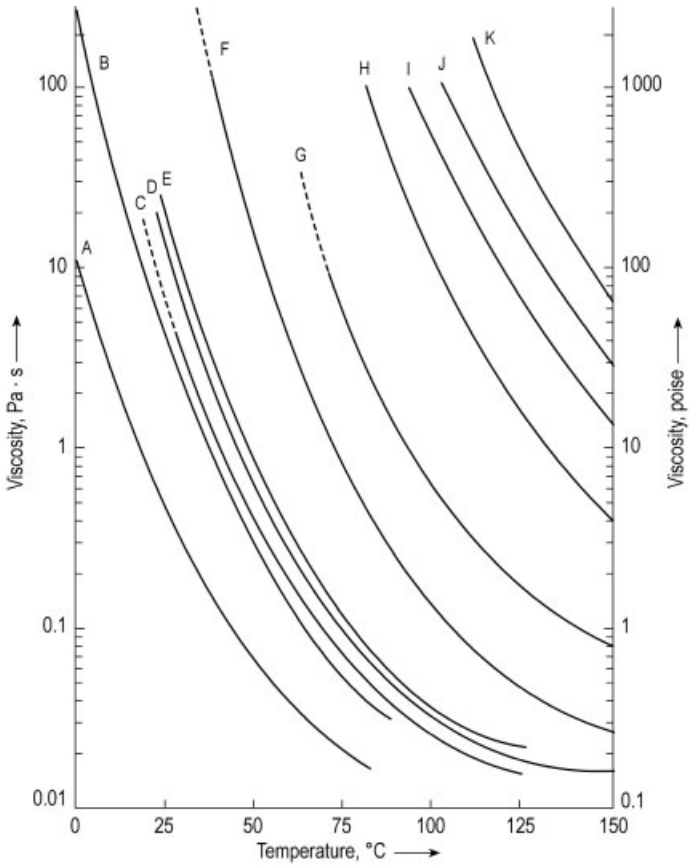

Figure 2. Viscosity–temperature profiles for bisphenol A epoxy resins with the following EEWs: A) 175–195 and 195–215 diluted resins; B) 172–178; C) 178–186; D) 185–192; E) 190–198; F) 230–280; G) 290–335; H) 450–550; I) 600–700; J) 675–760; K) 800–975.

Cls, such as the abnormal chlorohydrins and the organically "bound" chlorides. The issue here is the inconsistency in results obtained by different methods [78]. The presence of ionic hydrolyzable chlorides and total chlorides has been shown to affect electrical properties of epoxy molding compounds used in semiconductor encapsulation [85]. For these applications, producers offer high purity grade epoxy resins with low ionic, hydrolyzable and total chloride contents.

Total chloride content of epoxy resins can be determined by the classical Parr bomb method in which the sample is oxidized in a Parr bomb, followed by titration with silver nitrate [78]. The major disadvantage of this method is that it is time-consuming. Alternatively, X-ray fluorescence has been used successfully as a simple, nondestructive method to determine total chloride of epoxy resins. The method, originally developed by Dow Chemical, has been under consideration for adoption by ASTM.

The "ball and ring" and Durran's methods traditionally measure the softening point of SERs, which is important in applications such as powder coatings. The Durran's method involves heating a resin sample topped with a certain weight of mercury in a test tube until the resin reaches its softening point and flows, allowing the mercury to drop to the bottom of the test tube. The method is accurate but involves handling of highly hazardous mercury at elevated temperatures. The Mettlers' softening point method is more widely used recently because of its simplicity.

The esterification equivalent of solid resins is defined as the weight in grams ester-

ified by one mole of monobasic acid. This value includes both the epoxy and hydroxyl groups of the solid resin. It is determined by esterification of the sample with acetic anhydride in the presence of pyridinium chloride, followed by titration with sodium methoxide to a thymol blue–phenolphthalein end point.

Molecular structure of epoxy resins. Infrared spectroscopy (IR) is used to determine the epoxide content of resins as well as their structure. A compilation of IR spectra of uncured resins has been published [86] and their use in quality control and identification of components of resin blends has been described. Recently, near IR (NIR) has emerged as a useful tool to characterize epoxy resins [87].

NMR has been utilized to characterize epoxy resins, formulations and cured networks. It has been shown to be useful in determining the level of branching in epoxy resins and isomers distribution in epoxy novolacs [88, 89].

GPC and HPLC are utilized to characterize both liquid and solid epoxy resins [90]. MW and MW distributions are obtained from GPC measurements, but differences in chemical composition of resin samples are more apparent from HPLC chromatograms because of better resolution [91].

HPLC has proven to be a good fingerprinting tool to characterize LERs and SERs. Chromatograms of liquid epoxy resins (crude DGEBA) indicate a homologue distribution of $n = 0$, 85 %, and, in a specific case, $n = 1$, 11.5 %, although the values obtained depend on the source of the liquid resin. HPLC analysis of both liquid and solid epoxy resins has been studied in some detail using normal-phase and reversed-phase columns, respectively [39].

The difference between taffy-processed and fusion advancement solid resin can be noted in HPLC chromatograms. In the advancement process, the even-membered oligomers predominate, whereas taffy-produced resins exhibit both even- and odd-numbered oligomers. Compounds that contribute to hydrolyzable chloride and α-glycol content can be quantified by HPLC. The presence of branched chain components is detectable in studies using an improved reversed-phase gradient HPLC method [92, 93]. Excellent reviews of applications of chromatographic techniques to the analysis of epoxy resins are available [94].

14. Curing of Epoxy Resins

With the exception of the very high MW phenoxy resins and epoxy-based thermoplastic resins, almost all epoxy resins are converted into solid, infusible, and insoluble three-dimensional thermoset networks for their uses by curing with cross-linkers. Optimum performance properties are obtained by cross-linking (qv) the right epoxy resins with the proper cross-linkers, often called hardeners or curing agents. Selecting the proper curing agent is dependent on the requirements of the application process techniques, pot life, cure conditions, and ultimate physical properties. Besides affecting viscosity and reactivity of the formulation, curing agents determine both the types of chemical bonds formed and the degree of cross-linking that will occur. These, in turn, affect the chemical resistance, electrical properties, mechanical properties, and heat resistance of the cured thermosets.

Epoxy resins contain two chemically reactive functional groups: epoxy and hydroxy. Low MW epoxy resins such as LERs are considered difunctional epoxy monomers or prepolymers and are mostly cured via the epoxy group. However, as the MW of SERs increases, the epoxy content decreases, whereas the hydroxyl content increases. High molecular weight SERs can cross-link via reactions with both the epoxy and hydroxyl functionalities, depending on the choice of curing agents and curing conditions. Reaction of the epoxy groups involves opening of the oxirane ring and formation of longer, linear C−O bonds. This feature accounts for the low shrinkage and good dimensional stability of cured epoxies. The polycondensation curing via hydroxyl groups is often accompanied by generation of volatilebyproducts, such as water or alcohol, requiring heat for proper cure and volatiles removal.

It is the unique ability of the strained epoxy ring to react with a wide variety of reactants under many diverse conditions that gives epoxies their versatility [95]. Detailed discussions on the probable electronic configurations, molecular orbitals, bond angles, and reactivity of the epoxy ring are available in the literature [96].

Compared to noncyclic and other cyclic ethers, the epoxy ring is abnormally reactive. It has been postulated that the highly strained bond angles, along with the polarization of the C−C and C−O bonds account for the high reactivity of the epoxide. The electron-deficient carbon can undergo nucleophilic reactions, whereas the electron-rich oxygen can react with electrophiles. It is customary in the epoxy industry to refer to these reactions in terms of anionic and cationic mechanisms. The terminology was attributed to the fact that an anionic intermediate or transition state is involved in a nucleophilic attack of the epoxy while a cationic intermediate or transition state is formed by an electrophilic curing agent [97]. For the sake of clarity, the nucleophilic and electrophilic mechanism terminology is used in this article.

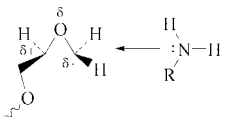

Curing agents are either catalytic or coreactive. A catalytic curing agent functions as an initiator for epoxy resin homopolymerization or as an accelerator for other curing agents, whereas the coreactive curing agent acts as a comonomer in the polymerization process. The majority of epoxy curing occurs by nucleophilic mechanisms. The most important groups of coreactive curing agents are those with active hydrogen atoms, e.g., primary and secondary amines, phenols, thiols, and carboxylic acids (and their anhydride derivatives). Lewis acids, e.g., boron trihalides, and Lewis bases, e.g., tertiary amines, initiate catalytic cures.

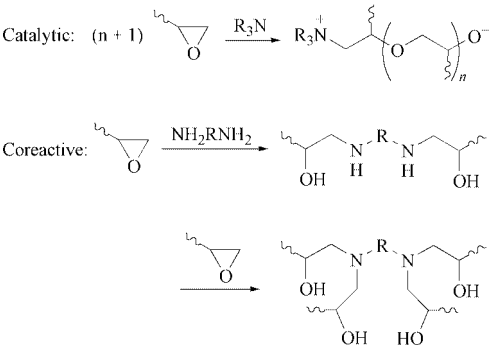

The functional groups surrounding the epoxide resin also affect the curing process. Steric factors [98, 99] can influence ease of cure. Electron-withdrawing groups adjacent to the epoxide ring often enhance the reactivity of the epoxy resin to nucleophilic reagents, while retarding its reactivity toward electrophilic reagents [98, 100, 101]. In general, aromatic and brominated aromatic epoxy resins react quite readily with nucleophilic reagents, whereas aliphatic and cycloaliphatic epoxies react sluggishly toward nucleophiles [102].

Figure 3 shows the pseudo first-order kinetic response for the disappearance of the epoxy in buffered methanol solutions (lines are for clarity only).

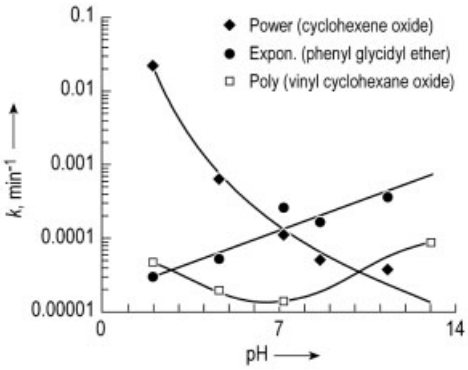

Figure 3. Effect of pH on reaction rates of epoxies

Clearly the epoxy structure dramatically influences the cure response of the epoxy as a function of pH. Cycloaliphatic epoxies are fast-reacting under low pH conditions. Aromatic glycidyl ethers are faster under high pH conditions. These results generally agree with "practical" cures: aromatic epoxies are easily cured with amines and amidoamines. Cycloaliphatics [102] are cured with acids and superacids. The behavior of the aliphatic epoxies is more complex but on balance is similar to that of cycloaliphatics.

In 2001, the U.S. market for epoxy curing agents was estimated at 165×10^3 MT (see Table 13), while approximately 318×10^3 MT of epoxy resins was consumed. The most commonly used curing agents are amines, followed by carboxylic-functional polyesters and anhydrides.

Table 13. U.S. consumption of curing agents for epoxy resins (2001)

Curing agents	Consumption, 10^3 MT	Market percentage
Amine functional compounds	50	48
Aliphatic amines and adducts	16	
Polyamides	14	
Amidoamines	9	
Cycloaliphatic amines	6.8	
Phenalkamines	1.8	
Dicyandiamide (DICY)	1.8	
Aromatic polyamines	0.9	
Carboxylics	37	36
Polycarboxylic polyesters	22	
Anhydrides	15	
Resole resins	9	9
Amino formaldehydes	4.5	
Phenol formaldehyde	4.5	
Novolacs and other phenolics	2.7	2.6
Polysulfides and polymercaptans	14	1.3
Catalysts	3.2	3
Anionic	3.1	
Cationic	0.1	
Others	0.9	< 1

A description of advantages, disadvantages and major applications of typical curing agents is given in Table 14.

15. Coreactive Curing Agents

Commercially, epoxy resins are predominantly cured with coreactive curing agents. Following are important classes of epoxy coreactive curing agents.

15.1. Amine Functional Curing Agents

This section describes one of the most important classes of epoxy coreactive curing agents.

15.1.1. Primary and Secondary Amines

Primary and secondary amines and their adducts are the most widely used curing agents for epoxy resins, accounting for close to 50 % of all the epoxy curing agents used in the United States in 2001. The number of amine hydrogen atoms present on the molecule determines the functionality of an amine. A primary amine group, one which has two hydrogens bound to it, will react with two epoxy groups while a secondary amine will react with only one epoxy group. A tertiary amine group, which has no active hydrogen, will not react readily with the epoxy group, but will act as a catalyst to accelerate epoxy reactions. Reactions of a primary amine with an oxirane group or an epoxy resin are shown in the following [103].

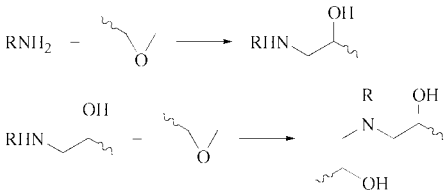

It has been reported that primary amines react much faster than secondary amines [101, 104]. Reaction of an epoxy group with a primary amine initially produces a secondary alcohol and a secondary amine. The secondary amine, in turn, reacts with an epoxy group to give a tertiary amine and two secondary hydroxyl groups. Little competitive reaction is detectable between a secondary hydroxyl group in the backbone and an epoxy group to afford an ether [100], provided a stoichiometric equivalent or excess amine is maintained. However, with excess epoxy, the secondary hydroxyl groups formed gradually add to the epoxide groups [105]. This reaction can be catalyzed by tertiary amines.

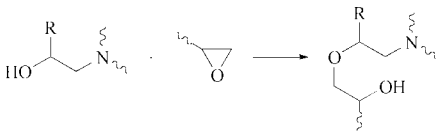

Hydroxyl compounds accelerate the rate of amine curing. A mechanism has been proposed [100] in which the hydroxyl group hydrogen bonds to the oxygen atom on the epoxy group, rendering the methylene group more susceptible to attack by the nucleophilic amine. Reactivity is proportional to the hydroxyl acidity and functionality; phenolics, aryl alcohols, and polyfunctional alcohols afford the best results.

Table 14. Curing agents for epoxy resins

Type	Advantages	Disadvantages	Major applications
Aliphatic amines and adducts	low viscosity; ambient cure temperature; little color; low cost	short pot life; rapid heat evolution; critical mix ratio; some are moderately toxic; high moisture absorption; blush; carbonation; limited high temperature performance ($< 100°C$)	flooring; civil engineering; marine and industrial coatings; adhesives; small castings
Cycloaliphatic amines	low viscosity; long pot-life; room temperature (RT) cure and heat-curable; adhesion to wet cement; good color; low toxicity; good electrical, mechanical, thermal properties (high T_g)	slower reactivity; high costs	flooring; paving; aggregate; industrial coatings; adhesives; tooling; composites; castings
Aromatic amines	excellent elevated temperature performance (150°C); good chemical resistance; long pot life; low moisture absorption	solids; incompatibility with resins; long cure cycles at high temperature (150°C); toxicity	high performance composites and coatings; adhesives; electrical encapsulation
Amidoamines	low viscosity; reduced volatility; good pot life; ambient cure temperature; convenient mix ratios; good toughness	poor performance at high temperature ($< 65°C$); some incompatibility with epoxies	high solids, solvent-free coatings; floorings; concrete bonding; troweling compounds
Polyamides	good mix ratios; pot life; RT cure; good concrete wetting; flexibility; low volatility and toxicity	high viscosity; low temperature performance; poor color; higher cost	marine and maintenance coatings; civil engineering; castings; adhesives
Anhydrides	low exotherm; good thermal (high T_g), mechanical, electrical properties; low shrinkage and viscosity; long pot life; little color	long cure cycles at high temperature (200°C)	composites; castings; potting; encapsulation
Catalytic	long pot life; high temperature resistance	brittle; moisture-sensitive	adhesives; prepregs; electrical encapsulation; powder coatings
Dicyandiamide	good electrical properties; high temperature resistance; latent systems	incompatibility with epoxy resins	electrical laminates; powder coatings; single-package adhesives
Carboxylic-terminated polyesters	good weatherability, corrosion resistance, and mechanical properties; low cost	poor chemical resistance	powder coatings
Isocyanates	fast cure at low temperature; good flexibility and solvent resistance	moisture-sensitive; toxic	powder coatings; maintenance coatings
Phenol–formaldehyde, novolacs	good chemical resistance, electrical properties, shelf stability, and compatibility with epoxies; high temperature resistance	high melting solids; high temperature cure; poor UV stability	molding compounds; powder coatings; electrical laminates
Polysulfides and polymercaptans	RT rapid cure times; flexible systems; moisture insensitive	poor performance at high temperature; odorous	consumer adhesives; sealants; traffic paints
Melamine–formaldehyde	good color and hardness; stable one-component systems	high temperature cure	stove paints; can coatings
Urea–formaldehyde	stable one-component systems; little color; low cost	high temperature cure; formaldehyde emission	fast-bake enamels; stove primers; can and drum coatings
Phenol–formaldehyde resoles	stable one-component systems; excellent chemical resistance	high temperature cure; brittle; gold color	baked enamels; can, drum and pail coatings; high temperature service coatings

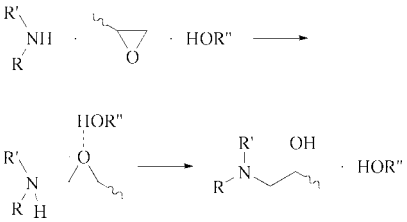

In general, reactivity of amines toward aromatic glycidyl ethers follows their nucleophilicity: aliphatic amines > cycloaliphatic amines > aromatic amines. Aliphatic amines cure aromatic glycidyl ether resins at room temperature (RT) without accelerators, whereas aromatic amines require elevated temperatures. However, with the help of accelerators, the cure rates of aromatic amines can approach those of some aliphatic amines. In general, the steric and electronic effects of substituents of the epoxy and

the amine influence the reaction rate of an amine with an epoxy resin.

15.1.1.1. Aliphatic Amines

The liquid aliphatic polyamines such as polyethylene polyamines (PEPAs) were some of the first curing agents used with epoxies. They give good RT cures with DGEBA-type resins. The low equivalent weights of the ethylene amines give tightly cross-linked networks with good physical properties, including excellent chemical and solvent resistance but limited flexibility and toughness. Good long-term retention of properties is possible at temperatures up to 100°C. Short-term exposure to higher temperatures can be tolerated. Certain aliphatic amines cured epoxies will blush (or bloom) under humid conditions. This undesirable property has been attributed to the incompatibility of some amine curing agents with epoxy resins. Incompatible amines can exude to the surface during cure and react with atmospheric carbon dioxide and moisture to form undesirable carbamates (carbonation). This, in turn, leads to gloss reduction and intercoat adhesion and recoatability problems in coating applications [106].

$$\underset{\text{Water}}{H_2O} + \underset{\text{Carbon dioxide}}{CO_2} \longrightarrow \underset{\text{Carbonic acid}}{H_2CO_3}$$

$$\underset{\text{Carbonic acid}}{H_2CO_3} + \underset{\text{Amine}}{RNH_2} \longrightarrow \underset{\text{Carbamic acid}}{RNHCOOH} + \underset{\text{Water}}{H_2O}$$

$$\underset{\text{Carbamic acid}}{RNHCOOH} + \underset{\text{Amine}}{R'NH_2} \longrightarrow \underset{\text{Carbamate}}{R'NH_3^{-}\ OCONHR}$$

Mixing ratios with epoxy resin are very critical, and working pot lives are too short for some applications. Aliphatic polyamines are hygroscopic and volatile, have bad odor, and cause dermatitis if improperly handled. Another disadvantage is high exotherm in thick sections or large mass parts that can lead to thermal decomposition. Consequently, significant efforts have been devoted toward remedying these shortcomings by modifications of the polyethylene polyamines. Adducts with epoxy resins (resin adducts), carboxylic acids (polyamides, amidoamines), ketones (ketimines), and phenols/formaldehyde (Mannich bases) [107] are widely used commercially. Longer chain alkylenediamines such as hexamethylenediamine (HMD) and polyetheramines (polyglycol-based polyamines) have also been developed. Currently, very small amounts of unmodified polyamines are used as curing agents for epoxies. They are primarily used to produce epoxy adducts (up to 90 %). Chemical modification by reaction with epoxy groups to yield epoxy adducts affords products with better handling properties, lower vapor pressure, reduced tendency to blush, and less critical mix ratio. For example, diethylenetriamine (DETA) readily reacts with ethylene oxide in the presence of water to give a mixture of mono- and dihydroxyethyl diethylenetriamine with a longer pot life and fewer dermatitic effects than free DETA.

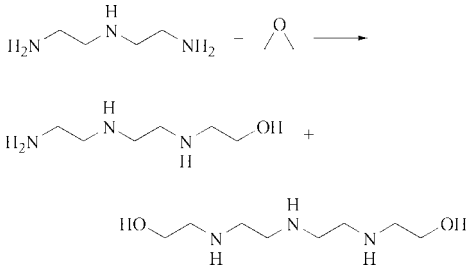

Resinous adducts are produced by reaction of excess diamine with epoxy resins.

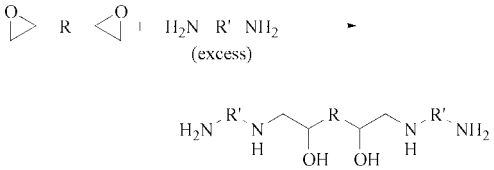

The higher molecular weight of the adduct affords a more desirable, forgiving ratio of resin to curing agent, lower water absorption, and better resin compatibility.

15.1.1.1.1. Ketimines

Ketimines are the reaction products of ketones and primary aliphatic amines. In the absence of reactive hydrogens, they do not react with epoxy resins. They can be considered blocked amines or latent hardeners, since they are readily hydrolyzed to regenerate the amines. They have low viscosity, cure rapidly when exposed to atmospheric humidity, and are useful in high solids coatings. Similar products have been obtained with acrylonitrile.

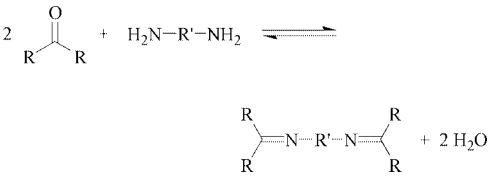

15.1.1.1.2. Mannich Base Adducts

Mannich base adduct is the reaction product of an amine with phenol and formaldehyde.

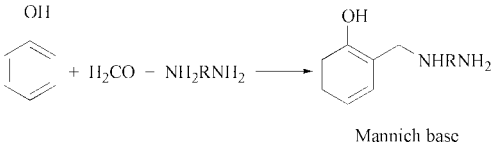

The resultant product has an internal phenolic accelerator. Compared to unmodified amines Mannich base adducts have lower volatility, less blushing and carbonation, and, despite their higher MW, faster reactivity.

15.1.1.1.3. Polyetheramines

Polyetheramines are produced by reacting polyols derived from ethylene oxide or propylene oxide with amines. The more commercially successful adducts are based on propylene oxide and are available in different MWs (JEFFAMINE* from Huntsman). The longer chain backbone provides improved flexibility but slower cure rate. Chemical and thermal resistance properties are also reduced. Polyetheramines are often used in combination with other amines in flooring, and adhesive and electrical potting applications.

15.1.1.2. Cycloaliphatic Amines

Cycloaliphatic amines were originally developed in Europe, where their use as epoxy curing agents is well established. Compared to aliphatic amines, cycloaliphatic amines produce cured resins having improved thermal resistance and toughness. Glass-transition temperatures (T_g) approach those of aromatic amines ($> 150°C$), while percent elongation can be doubled. However, chemical resistance is inferior to that of aromatic amines. Because cycloaliphatic amines are less reactive than acyclic aliphatic amines, their use results in a longer pot life and in the ability to cast larger masses. Unmodified cycloaliphatic amines require elevated temperature cure, but modified systems are RT-curable. Properly formulated, they can give an excellent balance of properties: fast cure, low viscosity, low toxicity, good adhesion to damp concrete, and excellent color stability. They are, however, more expensive than other types of curing agents.

Isophorone diamine (IPDA), bis(4-aminocyclohexyl)methane (PACM), and 1,2-diaminocyclohexane (1,2-DACH) are the principal commercial cycloaliphatic polyamine curing agents. IPDA is the largest volume cycloaliphatic amine. Commercial cycloaliphatic amines are formulated products. In addition to the cycloaliphatic amines, other components such as aliphatic amines and plasticizers are also included to improve RT cure speed and end-use properties. One popular formulation consists of IPDA used in combination with trimethylhexamethylenediamines (TMDA) or *meta*-xylenediamine (MXDA), and plasticizers/accelerators such as nonyl phenol or benzyl alcohol. In some ambient cure coating applications, cycloaliphatic amines can be reacted with phenol and formaldehyde to form the Mannich base products, which have an internal phenol accelerator and cure readily at ambient temperatures.

The largest market for cycloaliphatic amines is in flooring, followed by high solids coatings, composites, adhesives, castings, and tooling. Cycloaliphatic amines experienced significant growth in the early 1990s as replacements for more toxic aromatic amines such as MDA. However, anhydrides have been more successful at replacing aromatic amines in composite applications.

15.1.1.3. Aromatic Amines

Because of conjugation, aromatic amines have lower electron density on nitrogen than do the aliphatic and cycloaliphatic amines. Consequently, they are much less reactive toward aromatic epoxies. They have longer pot-lives and

usually require elevated temperature cures. Aromatic amines are usually solid at room temperature. These hardeners are routinely melted at elevated temperatures and blended with warmed resins to improve solubility. Eutectic mixtures of *meta*-phenylenediamine (MPD) and methylenedianiline (MDA or DDM) exhibit a depressed melting point resulting in an aromatic hardener that remains a liquid over a short period of time. MDA or 4,4′-diaminodiphenylmethane (DDM), 4,4′-diaminodiphenyl sulfone (DDS or DADPS), and MPD are the principal commercial aromatic amines. A new aromatic amine, diethyltoluenediamine (DETDA) has gained more significant uses in recent years.

Epoxies cured with aromatic amines typically have better chemical resistance and higher thermal resistance properties than products cured with aliphatic and cycloaliphatic amines. Their best attribute is their retention of mechanical properties at long exposures to elevated temperatures (up to 150°C). Consequently, they are widely used in demanding structural composite applications such as aerospace, PCB laminates, and electronic encapsulation. 4,4′-DDS is the standard curing agent used with a multifunctional amine epoxy (MY 720) for high performance aerospace and military composite application. 3,3′-DDS is used in aerospace honeycomb for its excellent peel strength. MDA, which has excellent mechanical and electrical properties, is the most widely used aromatic curing agent, but recently has been classified as a potential human carcinogen and its volume has been declining. Alkyl-substituted MDAs such as tetraethyl-MDA have been developed with lower toxicity and improved performance [108, 109]. However, none of the replacement products has the performance/cost combination of MDA. Anhydrides and cycloaliphatic amines have been used to replace aromatic amines in a number of composite applications. Efforts have been made to develop ambient-curable aromatic amines by adding accelerators such as phenols to MDA.

15.1.1.4. Arylyl Amines

These amines have cycloaliphatic or aromatic backbones, but the amine functional groups are separated from the backbone by methylene groups (benzylic amines and hydrogenated derivatives). Consequently, arylyl amines are much more reactive toward epoxies than aromatic amines while having improved thermal and chemical resistance over aliphatic amines. Fast cures at ambient and sub-ambient are possible with arylyl amines. These amines are more widely used in Japan and Europe than in North America *meta*-Xylylene diamine (MXDA) and its hydrogenated product, 1,3-bis(aminomethyl cyclohexane) (1,3-BAC) are popular arylyl amines.

The commercial polyamine curing agents are given in Table 15.

The stoichiometric quantity of polyamine used to cure an epoxy resin is a function of the molecular weight and the number of active hydrogens of the polyamine (amine equivalent weight, AEW) and the EEW or equivalent weight of epoxy resin; it is expressed as follows:

$$\left(\frac{\text{AEW}}{\text{EEW}}\right) \times 100 = \text{parts by weight polyamine per 100 parts by weight epoxy resin}$$

15.1.2. Polyamides

Polyamides are one of the largest volume epoxy curing agents used. They are prepared by the reaction of dimerized and trimerized vegetable-oil fatty acids with polyamines. Dimer acid is made by a Diels–Alder reaction between 9,12- and 9,11-linoleic acids. Subsequent reaction with diethylenetriamine or other suitable multifunctional amines yields the amine-terminated polyamides. They are available in a range of molecular weights and compositions.

Table 15. Commercial amine curing agents

Formula	Name	Abbreviation
Aliphatic		
$NH_2CH_2CH_2NHCH_2CH_2NH_2$	diethylenetriamine	DETA
$NH_2CH_2CH_2NHCH_2CH_2NHCH_2CH_2NH_2$	triethylenetetramine	TETA
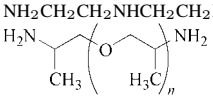	Poly(oxypropylene diamine)	
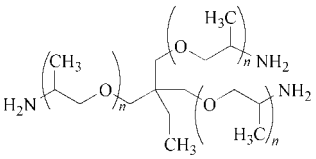	poly(oxypropylene triamine)	
$NH_2(CH_2)_3O(CH_2)_2O(CH_2)_3NH_2$	poly(glycol amine)	
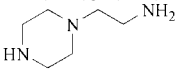	*N*-aminoethylpiperazine	AEP
Cycloaliphatic		
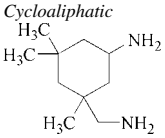	isophorone diamine	IPDA
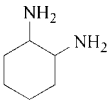	1,2-diaminocyclohexane	DACH
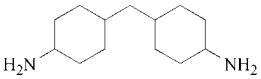	bis(4-aminocyclohexyl)methane	PACM
Aromatic		
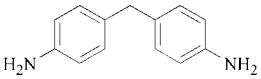	4,4′-diamino-diphenylmethane	MDA, DDM
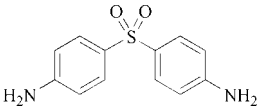	4,4′-diaminodiphenyl sulfone	4,4′-DDS
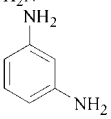	*m*-phenylenediamine	MPD
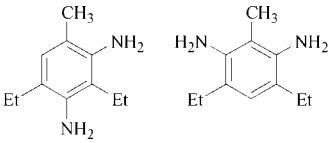	diethyltoluenediamine	DETDA
Arylyl amines		
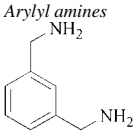	*meta*-xylene diamine	MXDA
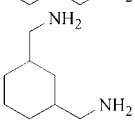	1,3-bis(aminomethyl cyclohexane)	1,3-BAC

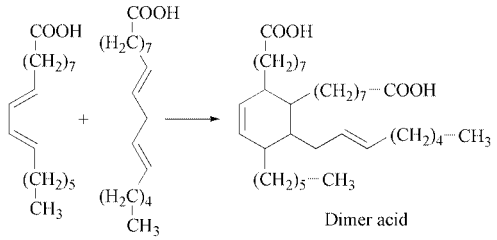

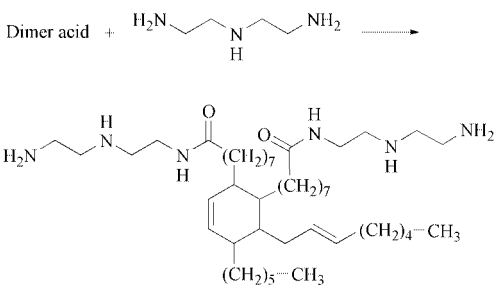

Polyamides are extremely versatile curing agents. The polyamides react with the epoxide group through the terminal amine functionality. The unreacted amide NH groups in the backbone provide good intercoat adhesion and the fatty acid structures provide good moisture resistance and mechanical properties. Wetting of cement surfaces is excellent. As a result of their relatively higher molecular weight, the ratio of polyamide to epoxy is more forgiving than with low MW polyamines. They are inexpensive, less toxic to handle; give no blushing; exhibit readily workable pot lives; and cure under mild conditions. Polyamides are mainly used in coating and adhesive formulations, mostly in industrial maintenance and civil engineering applications. The various MW polyamides exhibit different degrees of compatibility with epoxy resins. To ensure optimum properties, the polyamide/epoxy mixture must be allowed to react partly before being cured. This partial reaction assures compatibility and is known as the induction period.

Disadvantages of polyamides include slower cure speeds and darker color than polyamine-cured epoxies. Polyamide-cured epoxies lose structural strength rapidly with increasing temperature. This limits their use to applications not subjected to temperatures above 65°C. Formulations with tertiary amines, phenolic amines, or co-curing agents help to speed up cures at low temperatures. Alternatively, polyamides derived from polyamines with phenolic-containing carboxylic acids are called phenalkamines [110]. These curing agents have low viscosity and fast ambient cure speed and are widely used in on-site marine coatings and concrete deck applications.

The high viscosity of polyamides limits their uses primarily to low solids coatings, which have been losing ground to higher solids coatings. Waterborne polyamides have been developed for use with waterborne epoxies, but their growth has been modest over the past decade because the conversion to waterborne epoxy coatings has been slower than expected. Commercial polyamides include the Versamid resins from Cognis, Ancamide resins from Air Products, and Epicure resins from Resolution.

15.1.3. Amidoamines

Amidoamines have all the properties of polyamides, except for a significantly lower viscosity, which make them useful in high solids and solvent-free coating formulations. They are prepared by the reaction of a monofunctional acid like tall-oil fatty acid with a multifunctional amine such as DETA, resulting in a mixture of amidoamines and imidazolines.

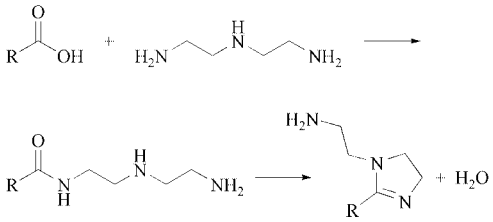

Imidazoline is formed by intramolecular condensation at high reaction temperatures. Commercial amidoamines are produced with different imidazoline contents to regulate reactivity and cured product performance. The pot life/reactivity of amidoamines varies with imidazoline content. High imidazoline content offers longer pot life and semi-latent curing activated by moisture. They are particularly useful in wet concrete applications. Like the polyamides, amidoamines can be used over a range of levels to enhance a specific property. However, amidoamines offer several advantages over aliphatic amines and polyamides. They offer more convenient mix ratios, increased

flexibility, and better moisture resistance than aliphatic polyamines, and they offer lower color and viscosity than polyamides. Consequently, the volume of amidoamines has grown significantly in the past decade.

15.1.4. Dicyandiamide

Dicyandiamide (DICY) is a solid latent hardener (*mp* 208°C). Its latent nature is due to its insolubility in epoxy resins at RT. DICY can be mixed with epoxy resins to provide a one-package formulation with good stability up to 6 months at ambient temperatures. Cure of epoxies with DICY occurs with heating to 150°C. It is often used with imidazoles as catalysts. DICY offers the advantage of being latent (reacts with epoxy resin upon heating and stops reacting temporarily when the heat is removed). This partially cured or "B-staged" state is ideal for prepreg applications. Typically, DICY is used at levels of 5–7 parts per 100 parts of liquid epoxy resins and 3–4 parts per 100 parts of solid epoxy resins.

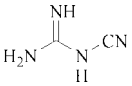

DICY is one of the first curing agents to be used with epoxy resins. It cures with epoxies to give a highly cross-linked thermoset with good mechanical strength, thermal properties, and chemical resistance, and excellent electrical properties. Because of its latency, low quantity requirements and excellent balance of properties, DICY is a widely used curing agent in powder coating and electrical laminate applications. These two applications account for 85 % of DICY consumption as epoxy curing agent.

The curing mechanism is rather complex, involving several simultaneous reactions. There are a number of conflicting proposed mechanisms in the literature. One study proposed the initial reaction of all four active hydrogens with epoxy resin catalyzed by tertiary amine catalysts followed by epoxy homopolymerization. The last step involves reactions between the hydroxyl groups of the epoxy resin with the cyano group [108, 109]. One of the more recent and plausible mechanisms of DICY cure with epoxies is that of Gilbert and co-workers [111]. The Gilbert mechanism is summarized in Figure 4. Gilbert and co-workers investigated the reaction of DICY with methyl glycidyl ether of bisphenol A (MGEBA). Products were analyzed using HPLC, NMR, and FTIR. On the basis of products that were isolated and characterized, Gilbert and co-workers proposed the mechanism shown in Figure 4.

The first step in the mechanism is the reaction of DICY with epoxy to form the alkylated DICY. This was confirmed by the imide IR peak at 1570 cm^{-1}. The second step involves further alkylation of the nitrogen that reacted in step 1, to form the *N*,*N*-dialkyldicyandiamide. No alkylation of the other amino group was suggested. The third step is the intramolecular cyclization step to form a zwitterionic five-membered intermediate. This involves the intramolecular reaction of the secondary alcohol formed in step 2 with the imide functionality (−C=N−). This is in contrast with the Zahir mechanism [112] where the intramolecular cyclization involves the hydroxy and the nitrile groups. The fourth step involves the elimination of ammonia and the formation of 2-cyanoimidooxazolidine. The formation of this heterocycle is consistent with the observed bathocromic IR shift from 1570 cm^{-1} to 1650 cm^{-1}. The ammonia that is eliminated can then react with epoxy to form a trifunctional cross-link. The last step involves the hydrolysis of the oxazolidine to form the oxazolidone and cyanamide. The hydrolysis step accounts for the formation of the carbonyl group.

15.2. Carboxylic Functional Polyester and Anhydride Curing Agents

Carboxylic polyesters and anhydrides are the second most important class of epoxy curing agent. Together, they constitute 36 % of the total curing agent volume used in the U.S. market (2001 data). Polyesters have been growing rapidly in powder coatings formulations with epoxy resins, consuming the highest tonnage of epoxy curing agents. This is driven in part by the conversion to the more environmentally friendly powder coating technologies, and in part by the versatility and cost efficiency of polyester–epoxy hybrid powder coatings. Anhydrides have been successfully replacing more toxic aromatic amines in composites. They account for 70 % of the volume of curing agents

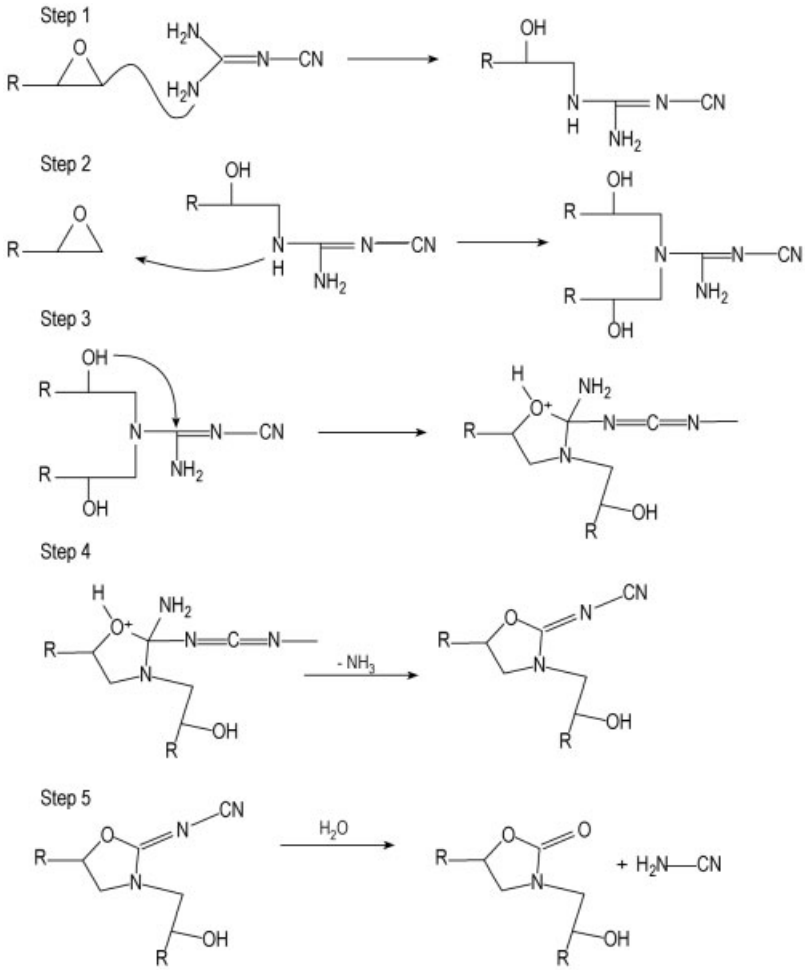

Figure 4. The Gilbert mechanism for the DICY curing of epoxy. From Ref. [111].

used in structural composite applications. Both polyesters and anhydrides are used in heat-cured applications only.

15.2.1. Carboxylic Functional Polyesters

The reaction of polyacids with polyalcohols produces polyesters. The terminal functionality is dictated by the ratio of the reactants. By virtue of their relatively cheap, widely available raw materials and good flexibility and weatherability, acid functional polyesters are used in hybrid epoxy powder coatings for a wide range of applications. For applications requiring good weatherability, triglycidyl isocyanurate (TGIC) is often used as curing agent for acid functional polyesters.

Terephthalic acid, trimellitic anhydride, and neopentyl glycol are commonly used raw materials to produce polyesters. Other acids, anhydrides, and glycols can also be used to modify functionality, MW, viscosity, and mechanical properties (after curing) of the polyesters. This versatility of the polyester building blocks allows many useful combinations of epoxy–polyester hybrid systems to be developed for a wide range of applications [113]. Major applications include coatings for metal furniture, general metal finishing, appliances, machinery and equipment, automotive, and wood. Auto-

motive is a new, large, and fast growing market with many car makers converting to primer-surfacer based on epoxy–polyester powder coatings. Wood coatings are a new, emerging market.

The curing mechanism of epoxy–polyester thermosets involves reaction of the acid functionality with epoxy followed by esterification of the epoxy hydroxyl groups with the acids [114]. Compounds such as amines and phosphonium salts catalyze these reactions. Water is a condensation reactionbyproduct that must be allowed to escape during the curing process to avoid coating defects.

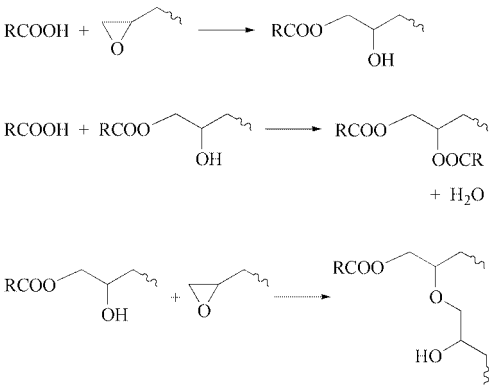

The first product is a β-hydroxypropyl ester, which reacts with a second mole of carboxylic acid to yield a diester. The hydroxyl ester can also undergo polymerization by reaction of its secondary hydroxyl group with an epoxy group.

15.2.2. Acid Anhydrides

Anhydrides are some of the very first epoxy curing agents used, and they remain a major class of curing agents used in heat-cured structural composites and electrical encapsulation. Their consumption volume equals that of all aliphatic amines and adducts in 2001 in the United States. While the carboxylic-terminated polyesters find widespread uses in coatings, anhydride use in coatings is minimal.

Epoxy–anhydride systems exhibit low viscosity and long pot life, low exothermic heats of reaction, and little shrinkage when cured at elevated temperatures. The low exotherm heat generation is a unique attribute of anhydrides, making them suitable for uses in large mass epoxy cures. Curing is slow at temperatures below 200°C and is often catalyzed by Lewis bases or acids. Postcure is often needed to develop optimum properties. Tertiary amines such as benzyldimethylamine, dimethylaminomethylphenol, tris(dimethylaminomethyl)phenol, boron trihalide amine complexes, stannic chloride, ammonium salts, phosphonium salts, and substituted imidazoles are effective catalysts. Proper catalyst concentration (0.5–2.5 % of resin weight) is critical, depending on the types of anhydrides and resins used and the cure schedules, and is known to affect high temperature performance.

Cured epoxy–anhydride systems exhibit excellent thermal, mechanical, and electrical properties, and are used in filament-wound epoxy pipe, PCB laminates, mineral-filled composites, and electrical casting and encapsulation applications. Anhydride-cured epoxies also have better aqueous acid resistance than similar amine-cured systems. Anhydrides are the principal curing agents for cycloaliphatic and epoxidized olefin resins in electrical casting and potting. Some key physical properties of exemplary epoxy resins cured with hexahydrophthalic anhydride are shown in Table 16.

The mechanism of anhydride cure is complex and controversial because of the possibility of several competing reactions. The uncatalyzed reaction of epoxy resins with acid anhydrides proceeds slowly even at 200°C; both esterification and etherification occur. Secondary alcohols from the epoxy backbone react with the anhydride to give a half ester, which in turn reacts with an epoxy group to give the diester. A competing reaction is etherification of an epoxy with a secondary alcohol, either on the resin backbone or that formed during the esterification, resulting in a β-hydroxy ether. It has been reported that etherification is a probable reaction since only 0.85 equivalents of anhydrides are required to obtain optimum cross-linked density and cured properties [103].

Table 16. Formulation and properties of epoxy resins cured with hexahydrophthalic anhydride

	DGEBA	3′,4′-Epoxycyclohexylmethyl 3,4-epoxycyclohexanecarboxylate	Hexahydrophthalic acid diglycidyl ester
Formulation			
Resin, pbw[a]	100	100	100
Hexahydrophthalic anhydride, pbw[a]	85	105	100
Accelerator type	tertiary amine	metal alkoxide salt	quaternary ammonium salt
pbw[a]	3	12	4
Cure schedule, h at °C	2 at 1001 at 150	4 at 120	4 at 804 at 140
Typical cured properties at 25°C			
Tensile strength, MPa[b]	65	68	83
Tensile modulus, MPa[b]	3400	3300	3000
Flexural strength, MPa[b]	131	89	127
Flexural modulus, MPa[b]	3400	3000	3000
Elongation, %	5.0	2.7	3.5
Compressive strength, MPa[b]	124	151	124
Heat-deflection temperature, °C	120	150	105
Water absorption, % weight gain[c]	0.5	0.4	0.4
Dielectric constant at 60 Hz	3.4	3.3	3.5
Dissipation factor at 60 Hz	0.006	0.005	0.007
Volume resistivity, $\Omega \cdot cm \times 10^{16}$	2.0	10.0	3.0

[a] Parts by weight.
[b] To convert MPa to psi, multiply by 145.
[c] After boiling for 1 h.

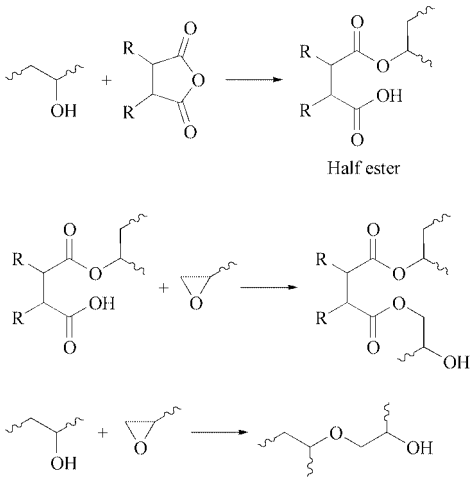

Lewis bases such as tertiary amines and imidazoles are widely used as epoxy–anhydride catalysts. Conflicting mechanisms have been reported for these catalyzed reactions [115]. The more widely accepted mechanism [103] involves the reaction of the basic catalyst with the anhydride in the initiation step to form a betain (internal salt). The propagation step involves the reaction of the carboxylate anion with the epoxy group, generating an alkoxide. The alkoxide then further reacts with another anhydride, propagating the cycle by generating another carboxylate which reacts with another epoxy group. The end result is the formation of polyester-type linkages. In practice, it has been observed that optimum properties are obtained when stoichiometric equivalents of epoxy and anhydride are used with high temperature cures, which is consistent with this mechanism and does not involve etherification reactions. At lower anhydride/epoxy ratios (0.5:1) and lower cure temperatures, some etherifications can take place by reaction of the alkoxide with an epoxy group.

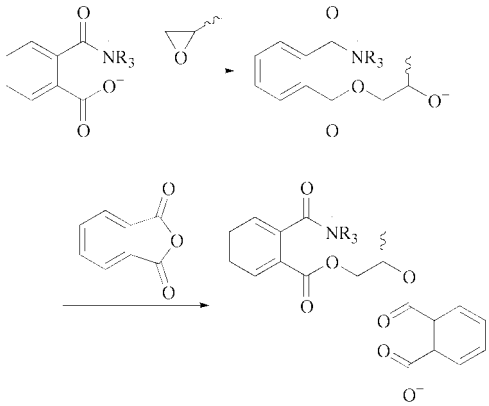

Numerous structurally different anhydrides can be used as epoxy curing agents, but the most widely used are liquids for ease of handling. The most important commercial anhydrides are listed in Table 17. Methyltetrahydrophathalic

Table 17. Commercially important anhydride curing agents

Name	Structure
Phthalic anhydride (PA)	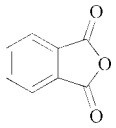
Tetrahydrophthalic anhydride (THPA)	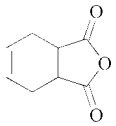
Methyltetrahydrophthalic anhydride (MTHPA)	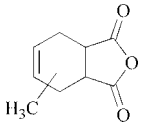
Methyl hexahydrophthalic anhydride (MHHPA)	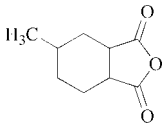
Hexahydrophthalic anhydride (HHPA)	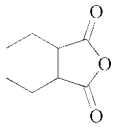
Nadic methyl anhydride or methyl himic anhydride (MHA)	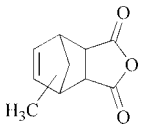
Benzophenonetetracarboxylic dianhydride (BTDA)	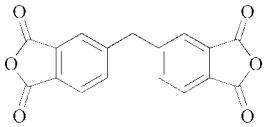
Tetrachlorophthalic anhydride (TCPA)	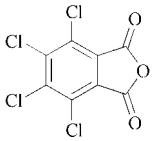

anhydride (MTHPA) is the largest volume anhydride, used in filament-winding composites. Phthalic anhydride (PA) is the next largest volume and is inexpensive; so it is used widely in mineral-filled laboratory bench top manufacturing, which requires low exotherm heat generation to avoid cracking. Dodecylsuccinic anhydride (DDSA) has a long aliphatic chain in the backbone and is used as blends to improve flexibility. Benzophenonetetracarboxylic dianhydride (BTDA) is a relatively new, multifunctional anhydride developed for high temperature applications, capable of achieving a high cross-linking density with a heat distortion temperature (HDT) of 280°C. It has been used as a replacement for more toxic aromatic amines. Tetrachlorophthalic anhydride (TCPA) is used in epoxy powder coatings for small electronic components with flame-retardancy requirements.

15.3. Phenolic-Terminated Curing Agents

Phenolics form a general class of epoxy curing agents containing phenolic hydroxyls capable of reacting with the epoxy groups. They include phenol-, cresol-, and bisphenol A terminated epoxy resin hardener. More recent additions include bisphenol A based novolacs. Cure takes place at elevated temperatures (150–200°C) and amine catalysts are often used.

The bisphenol A terminated hardeners are produced using liquid epoxy resins and excess bisphenol A in the resin advancement process. They are essentially epoxy resins terminated with bisphenol A. They are popular in epoxy powder coating applications for rebar and pipe, providing more flexible epoxy coatings than the novolacs.

The novolacs are produced via the condensation reaction of phenolic compounds with formaldehyde using acid catalysts. They are essentially precursors to epoxy novolacs. Novolacs are multifunctional curing agents and can impart higher cross-link density, higher T_g, and better thermal and chemical resistance than other phenolics. Cresol novolacs provide higher solvent and moisture resistance, but are more brittle than their phenol novolac counterparts. Recently, bisphenol A based novolacs have been used in electrical laminate formulations to improve thermal performance (T_g and T_d) [116]. Novolacs are widely used in composites, PCB laminates, and electronic encapsulation applications. Their uses in coatings are limited to high temperature applications such as powder coatings for down-hole oil-field pipe coatings.

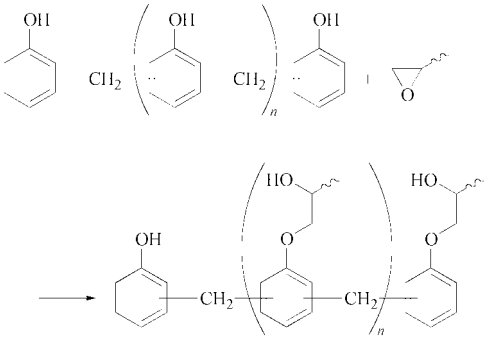

15.4. Melamine–, Urea–, and Phenol–Formaldehyde Resins

Melamine–formaldehyde, urea–formaldehyde, and phenol–formaldehyde resins react with hydroxyl groups of high MW epoxy resins to afford cross-linked networks [72].

The condensation reaction occurs primarily between the methylol or alkylated methylol group of the formaldehyde resin and the secondary hydroxyl group on the epoxy resin backbone. The high bake temperatures used in these applications drive off the condensationbyproducts (alcohol or water). Acids such as phosphoric acid and sulfonic acids are often used as catalysts.

There are two types of phenol–formaldehyde condensation polymers: resoles and novolacs [117]. Phenol–formaldehyde polymers prepared from the base-catalyzed condensation of phenol and excess formaldehyde are called resoles. In most phenolic resins commonly used with epoxies, the phenolic group is converted into an ether to give improved alkali resistance. At elevated temperatures (> 150°C), resole resins react with the hydroxyl groups of the epoxy resins to provide highly cross-linked polymers.

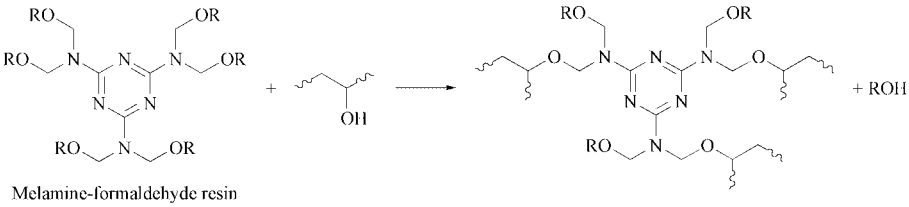

Melamine-formaldehyde resin

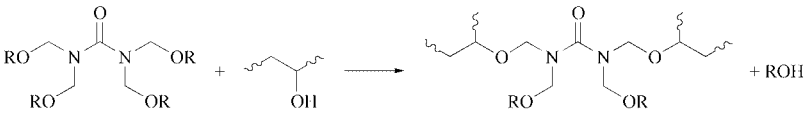

Urea-formaldehyde resin

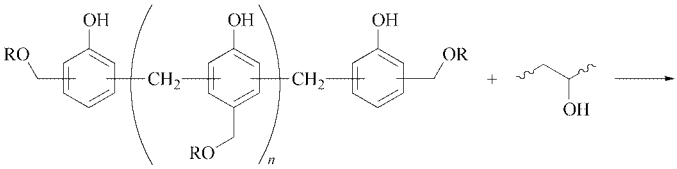

Phenyl-formaldehyde resin

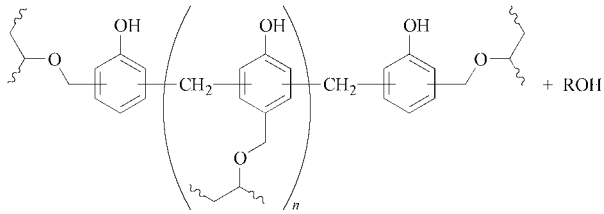

The melamine- and urea–formaldehyde resins are also called amino resins [118]. The phenol–formaldehyde resoles are often called phenolic resins, which is rather easily confused with phenolic-terminated cross-linkers such as novolacs and bisphenol A terminated resins.

These formaldehyde-based resins are widely used to cure high MW solid epoxy resins at elevated temperatures (up to 200°C) for metal can, drum, and coil coatings applications. The resultant coatings have excellent chemical resistance, good mechanical properties, and no effects on taste (adding or taking away taste from packaged foods or drinks). The vast majority of the food and beverage cans produced in the world today are coated internally with epoxy–formaldehyde resin coatings. The phenol–formaldehyde resoles are also used with epoxies in coatings for high temperature service pipes and to protect against hot, corrosive liquids.

15.5. Mercaptans (Polysulfides and Polymercaptans) Curing Agents

The mercaptan group of curing agents includes polysulfide and polymercaptan compounds which contain terminal thiols.

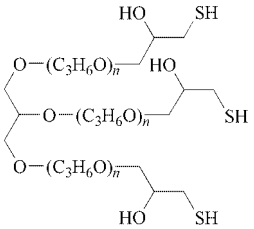

Polymercaptans

In the language commonly used in this industry, "polysulfides" typically have a functionality of 2, while "polymercaptans" have an average functionality of 3. By itself, the thiol or mercaptan group (SH) reacts very slowly with epoxy resins at ambient temperature. However, when converted by a tertiary amine to a mercaptide ion, they are extremely reactive [119].

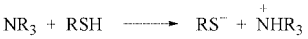

Increasing the basic strength of the amine increases the reaction rate. Polar solvents are also known to speed up these reactions. Fast curing at ambient conditions is the primary attribute of this class of curing agent, lending themselves to applications such as the "5-minutes" consumer adhesives, concrete road repairs, and traffic marker adhesives. In practice, they are often formulated with co-curing agents such as amines or polyamides to achieve a balance of fast cure with improved mechanical properties. The tertiary amine accelerated polymercaptan/epoxy systems exhibit good flexibility and tensile strength at ambient temperature. They are used in high lap-shear adhesion applications such as concrete patch repair adhesives. One disadvantage of polymercatans is their strong odor. Aliphatic amine/polysulfide co-curing agent systems yield improved initial elevated temperature performance and are widely used as building adhesives for their excellent adhesion to both glass and concrete. However, both systems lose some flexibility on aging.

15.6. Cyclic Amidines Curing Agents

Cyclic amidine curing agents are typically used in epoxy powder coating formulations and in decorative epoxy–polyester hybrid powder coatings to produce matte surface for furniture and appliance finishes. 2-Phenyl imidazoline has been used successfully to produce low gloss epoxy powder coatings. It is highly reactive, capable of curing at relatively low temperatures (140°C) and is suitable for curing of coatings on temperature-sensitive substrates such as wood and plastics. Other curing agents in this group include salts of polycarboxylic acids and cyclic amidines. Their volume is currently small but is expected to grow as the markets for low gloss and low temperature cure powder coatings develop. They can also be used as tertiary amine catalysts similar to imidazoles.

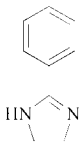

15.7. Isocyanate Curing Agents

Isocyanates react with epoxy resins via the epoxy group to produce an oxazolidone structure [120, 121] or with a hydroxyl group to yield a urethane linkage. The urethane linkage provides improved flexibility, impact, and abrasion resistance. The oxazolidone products have been successfully commercialized in high temperature resistance coating and composite applications. Blocked isocyanates are used as cross-linkers for epoxy in PPG's cathodic electrodeposition (CED) coatings. Isocyanates are also used to cure epoxies in some powder coatings, but their toxicity has limited their use.

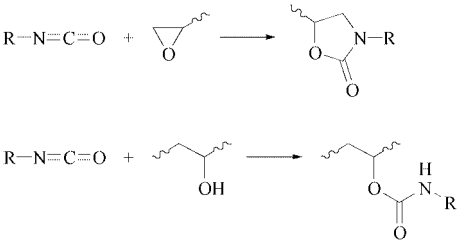

15.8. Cyanate Ester Curing Agents

Cyanate esters can be used to cure epoxy resins to produce highly cross-linked thermosets with high modulus and excellent thermal, electrical, and chemical resistance properties. They are used in high performance electrical laminate and composite applications. Cure involves oxazoline formation catalyzed by metal carboxylates in addition to homopolymerization of both cyanate ester and epoxy [122]. The high costs of cyanate esters however limit their uses.

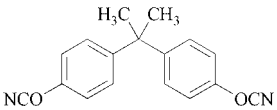

Bisphenol A based dicyanate ester

16. Catalytic Cure

The catalytic curing agents are a group of compounds that promote epoxy reactions without being consumed in the process. In some of the epoxy literature, catalysts are referred to as "accelerators"; the distinction of these two types of additives is discussed in later sections.

16.1. Lewis Bases

Lewis bases contain an unshared pair of electrons in an outer orbital and seek reaction with areas of low electron density. They can function as nucleophilic catalytic curing agents for epoxy homopolymerization; as co-curing agents for primary amines, polyamides, and amidoamines; and as catalysts for anhydrides. Tertiary amines and imidazoles are the most commonly used nucleophilic catalysts. Several different mechanisms are possible:

1) The catalytic curing reactions of tertiary amines with epoxy resins follow two different pathways, depending on the presence or absence of hydrogen donors, such as hydroxyl groups. In the absence of hydrogen donors [123], tertiary amines react with the electron-poor methylene carbon of the epoxy group to form an intermediate zwitterion. The zwitterion then attacks another epoxy group to continue homopolymerization via an anionic mechanism. In the presence of hydrogen donors such as alcohols, the zwitterion abstracts the proton from the alcohol to generate an alkoxide,

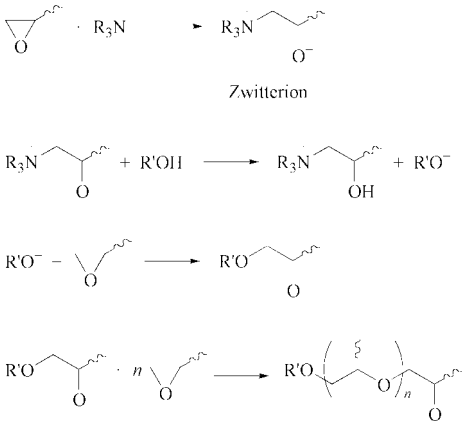

which further reacts with an epoxide group. Chain propagation continues by way of a polymeric anion mechanism.

2) With more acidic hydrogen donors such as benzyl alcohol, phenols, or mercaptans, the tertiary amine acts as a co-curing agent by

first abstracting the proton from the hydrogen donor:

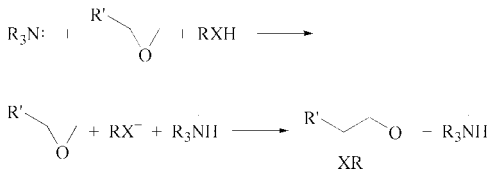

3) With anhydrides, the catalyst facilitates the anhydride ring opening:

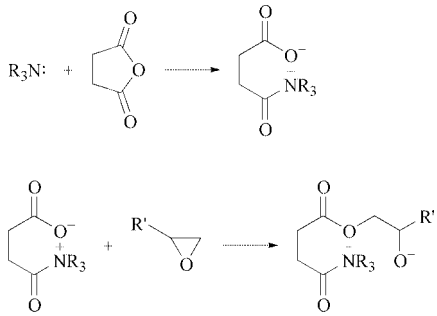

Commonly used tertiary amines include 2-dimethylaminomethylphenol (DMAMP) and 2,4,6-tris(dimethylaminomethyl)phenol (TDMAMP, trade name DMP-30 of Rohm and Haas), which contain built-in phenolic hydroxyl groups and can be used as a good catalysts and co-curing agents for room temperature cure of epoxies.

The rate of cure of epoxy resins with tertiary amines depends primarily upon the extent to which the nitrogen is sterically blocked. The homopolymerization reaction depends on the temperature as well as the concentration and type of tertiary amine. Benzyldimethylamine (BDMA) and TDMAMP are mainly used as accelerators for other curing agents, in the curing of anhydride- and dicyandiamide-based systems. Other tertiary amine catalysts include 1,4-diazabicyclo(2,2,2)octane (DABCO) and diazabicycloundecene (DBU).

Imidazoles such as 2-methylimidazole (2-MI) and 2-phenylimidazole (2-PI) contain both a cyclic secondary and a tertiary amine functional groups and are used as catalysts, catalytic curing agents, and accelerators [124, 125]. They are widely used as catalysts for DICY-cured epoxies in electrical laminates. For powder coatings, 2-MI adducts of LER are often used to facilitate dispersion of the components in powder coating formulations and to enhance shelf-life. Other modified imidazoles are also commercially available. The main advantage of imidazoles is the excellent balance of pot life and fast cure. 2-PI is used to increase T_g and thermal resistance.

Cyclic amidines such as 2-phenylimidazoline have also been used as a catalyst and co-curing agent in epoxy–polyester and epoxy powder coatings.

Substituted ureas are another group of epoxy nucleophilic catalytic curing agent, derived by blocking of isocyanates with dimethylamine. They are commonly used as catalysts for DICY cure of epoxies in adhesives, prepregs, and structural laminates. The ureas exhibit outstanding latency at room temperature and are widely used in one-pack adhesives. The catalytic mechanism of ureas is not well understood, but it has been postulated that DICY assists in deblocking of the urea to generate a tertiary amine, which in turns acts as epoxy curing catalyst. Commercially important substituted ureas are 3-phenyl-1,1-dimethyl urea (Amicure UR by Air Products), a reaction product of phenyl isocyanate with dimethylamine; and Amicure UR 2T, a reaction product of toluene diisocyanate (TDI) with dimethylamine.

Quaternary phosphonium salts such as the tetraalkyl and alkyl-triphenylphosphonium halides have been used as fast catalysts for curing of phenolics, carboxylic acid-terminated polyesters, or anhydrides with epoxies [126]. Used in powder coatings, they showed good latency and fast cure rates at moderate temperatures.

Air Products is a major epoxy catalyst supplier. Others include Huntsman, Cognis, and SKW Chemicals.

16.2. Lewis Acids

Lewis acids, e.g., boron trihalides, contain an empty outer orbital and therefore seek reaction with areas of high electron density. Boron trifluoride, BF_3, a corrosive gas, reacts easily with epoxy resins, causing gelation within a few minutes. Complexation of boron trihalides with amines enhances the curing action. Reasonable pot lives using these complexes can be achieved because elevated temperatures are required for cure. Reactivity is controlled by the choices of

the halide and the amine. The amine choice also affects other properties such as solubility in resin and moisture-sensitivity. Boron trifluoride monoethylamine ($BF_3 \cdot NH_2C_2H_5$), a crystalline material which is a commonly used catalyst, cures epoxy resins at 80–100°C. A chloride version is also commercially available. Other Lewis acids used in epoxy curing include stannic chloride and tin octanate.

Different mechanisms have been proposed for curing epoxy resins with BF_3 complexes or salts. In general, it is assumed that complexation with the oxirane oxygen is involved, facilitating proton transfer and ether formation. Thermal dissociation of the BF_3–amine complex may form a proton that further reacts with the epoxy group to initiate the curing process [127]. Another mechanism assumes an amine adduct or salt is solvated by the epoxy groups, resulting in an oxonium ion [128]. The curing reaction is initiated and propagated by attack of other epoxy groups on the oxonium ion.

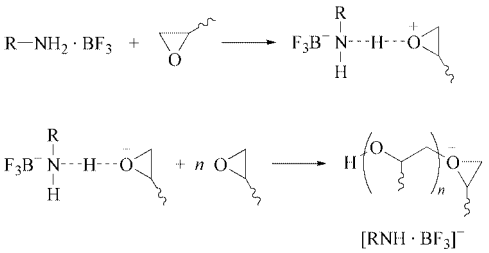

16.3. Photoinitiated Cationic Cure

Photoinitiated cationic curing of epoxy resins is a rapidly growing method for the application of coatings from solvent-free or high solids systems. This technology allows the formulation of epoxy coatings and adhesives with essentially "infinite" shelf life, but almost "instantaneous" cure rates. Cycloaliphatic epoxies are widely cured using photoinitiated cationic initiators.

Photoinitiators used for epoxy curing include aryldiazonium salts ($ArN_2^+X^-$), diaryliodonium salts ($Ar_2I^+X^-$), and onium salts of Group VIa elements, especially salts of positively charged sulfur ($Ar_3S^+X^-$). The anions must be of low nucleophilicity, such as tetrafluoroborates or hexafluorophosphates, to promote polymer chain growth rather than chain termination. Upon UV irradiation, photoinitiators yield a "super" acid, which polymerizes the epoxy resins by a conventional electrophilic mechanism.

The photolysis of diaryliodonium and triarylsulfonium salts may proceed via formation of a radical cation, which abstracts a hydrogen atom from a suitable donor.

$$Ar_3S^+ PF_6^- \xrightarrow{UV} Ar_2S^{\ddot{+}} + Ar\cdot + PF_6^-$$

$$Ar_2S^{\ddot{+}} + R{-}H \longrightarrow Ar_2SH^+ + R\cdot$$

$$Ar_2SH^+ + PF_6^- \rightleftharpoons Ar_2S + HPF_6$$

Subsequent loss of a proton yields the Brønsted acid HPF_6. Catalytic curing of the epoxy resin proceeds through an onium intermediate:

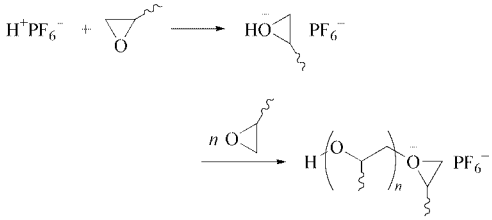

In the presence of triarylsulfonium and diaryliodonium salts, polymerization continues even if UV irradiation is terminated. This phenomenon is called "dark cure" and is due to the "living" nature of the "superacid" generated cation. The cure regime can be thought of as UV-initiated but "thermally cured." Thermally initiated cationic catalysts are also available [129].

In contrast, dialkylphenacyl sulfonium salts undergo reversible dissociation upon photolysis with formation of an ylid and a Brønsted acid. Cessation of UV activation results in termination of epoxy homopolymerization, since the acid is consumed in a reverse reaction with the ylid.

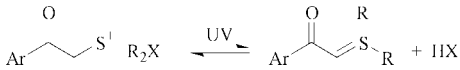

This type of behavior provides a means of controlling the degree of cure. Dialkylphenacyl sulfonium salts are thermally stable in epoxy resins at room temperature and up to 150°C for 1–2 h. Significant interest in thermal cationic cure of epoxies, especially cycloaliphatic epoxies, has developed [130].

17. Formulation Development With Epoxy Resins

The most important step in using epoxy resins is to develop the appropriate epoxy formulation since most are used as precursors to a three-dimensional cross-linked network. With the exception of the very high MW phenoxy resins and the epoxy-based thermoplastics, epoxy resin is rarely used by itself. It is usually formulated with modifiers such as fillers and used in composite structures with glass fiber or metal substrates (coatings). To design a successful epoxy formulation that will give optimum *processability* and *performance*, the following factors must be carefully considered:

1) Selection of the proper combination of epoxy resin(s) and curing agent(s) structures
2) Epoxy/Curing agent stoichiometric ratio
3) Selection of catalyst/accelerator
4) Curing/postcuring processes and conditions
5) Formulation modifiers such as fillers, diluents, toughening agents, etc
6) Interactions among the formulation ingredients and with the composite materials (fibers, metals, etc) on the system chemistry, adhesion, rheology, morphology, and performance

The development of an epoxy formulation containing a high number of components can be very resource and time-consuming. Techniques such as design of experiments (DOE) are useful tools to facilitate the formulation development process and to obtain optimum performance [131, 132]. Future developments should include application of high throughput techniques to epoxy formulation development and optimization.

17.1. Relationship Between Cured Epoxy Resin Structure and Properties

The following diagram illustrates the formation of cured epoxy networks using different ratios of a difunctional epoxy and a tetrafunctional hardener. The structures formed are significantly different, depending on the ratio used. Consequently, it is expected that performance of these networks will be quite different despite the fact that they are derived from identical building blocks (Fig. 5).

The structure between the cross-linking position and the distance between any two of these points are important characteristics. Molecular weight between cross-links (M_c) and cross-link density are terms developed to describe "distance" between cross-link points. The concept originated with the rubber elasticity theory developed for the lightly cross-linked elastomers and has been adopted for use with epoxy thermosets with mixed success [133, 134]. The cured epoxy system derives its properties mostly from a combination of cross-link density, monomer structure and the curing process.

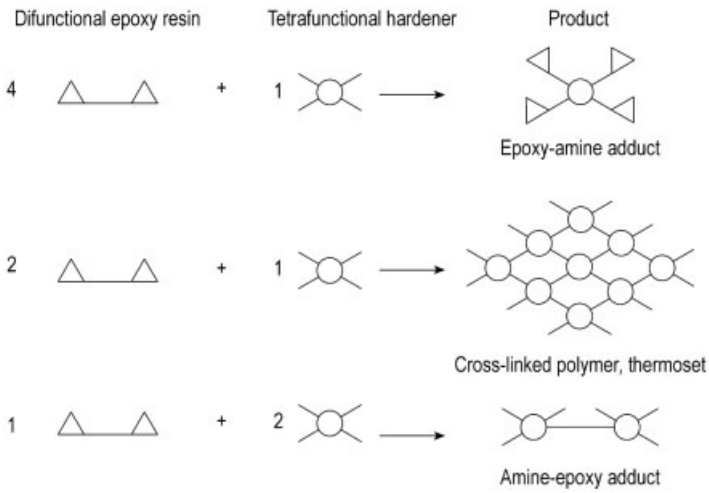

Figure 5. Formation of resin–hardener networks

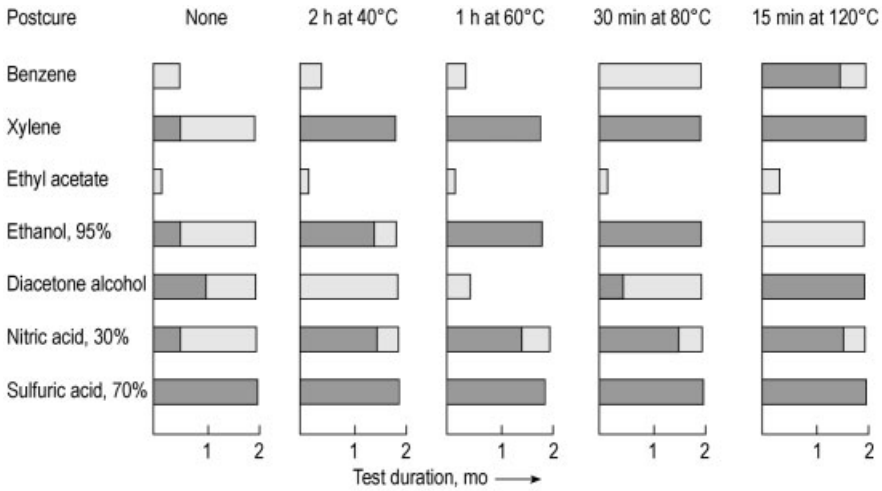

Figure 6. Chemical resistance of a DGEBA–aromatic polyamine adduct. Postcured substrate: sandblasted mild steel; film thickness: 300–350 µm; cure: 7 d at 20°C. degraded; resistant.

The two-dimensional schematic network structures do not represent spatial reality but have been devised to help understand the nature of the various structures [135]. A good understanding of the structure/property relationship is critical in designing the appropriate epoxy/curing agent combination. For example, cross-linking with dicarboxylic anhydrides yields polyesters that are resistant to oxidation, but less so to moisture, especially in the presence of basic components. Amine cross-linked systems are resistant to saponification but not to oxidation. There is a large body of specific structure/property relationship knowledge in the epoxy industry and literature, but only a few systematic treatments are available [136 – 138].

Cross-link density increases with degree of cure up to its limit at full conversion of the (limiting) functional groups. The curing temperature and process strongly influence cross-link density, molecular architecture, network morphology, residual stress, and the ultimate performance. The effects of degree of cure and subsequent cross-link density on the chemical resistance of a cured DGEBA–aromatic polyamine adduct system are depicted in Figure 6. The increase in chemical resistance properties after postcure also demonstrates the effects of increased cross-link density. The cross-link density of a cured epoxy system can be estimated by a number of different techniques as described in the characterization of cured epoxy section.

17.2. Selection of Epoxy Resins

Successful performance of epoxy-based systems depends on proper selection and formulation of components. The components that have the most significant influences are the epoxy resins and the curing agents. As discussed in earlier sections, there are numerous choices of epoxy resins and curing agents presenting a wide variety of structure and functionality. Figure 7 shows the general attributes of common types of epoxy resins.

Epoxy resins can be used separately or in combination, such that formulations can be designed to take advantage of the desirable characteristics of several components. Because combining resins from different families can result in certain trade-offs, a careful balance of components should be investigated to produce optimal performance for specific applications. Table 16 shows effects of different resin backbones on cured properties with formulations based on hexahydrophthalic anhydride as curing agent.

The difunctional DGEBA resins are offered commercially in a wide range of molecular weights. As the molecular weight increases, so does the chain length between the epoxy end groups. Table 5 shows the effects of increasing EEW and MW of bisphenol A based epoxy resins on resin properties. The cross-link density of a difunctional resin cured by way of the

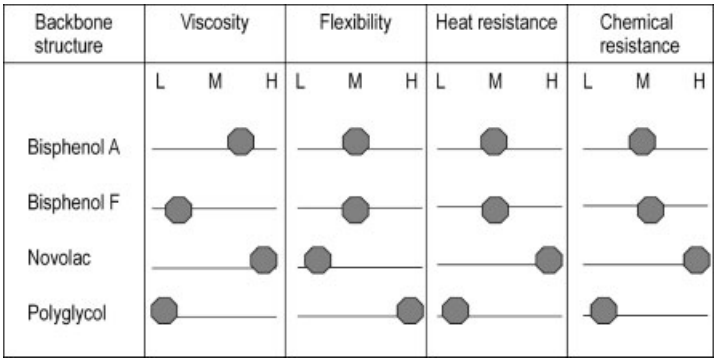

Figure 7. Comparison of relative properties of common epoxy resins
L, low; M, medium; H, high.

epoxy group decreases as the resin molecular weight increases. High molecular weight resins are frequently cured via the secondary hydroxyl group, chemistry that results in a different set of structure–property relationships.

Multifunctional epoxy resins are available with functionalities ranging from above 2 to about 5. When cured to the same degree using a given curing agent at stoichiometric ratios, they produce a higher cross-link density, higher glass-transition temperature, and better thermal and chemical resistance compared with difunctional epoxy resins.

17.3. Selection of Curing Agents

The selection of curing agents is just as critical as the selection of resins. As discussed in the Curing Agents section, there are numerous types of chemical reagents that can react with epoxy resins. Since coreactive curing agents become part of the network structure, careful consideration must be paid to their contributions. Besides affecting viscosity and reactivity of the formulation, curing agents determine both the types of chemical bonds formed and the functionality of the cross-link junctions that are formed. Table 18 show performance examples of a liquid epoxy resin cured with different curing agents. Several authors have attempted to rationalize the curing agent selection process for different applications [106, 139].

The effect of hardener structure on the heat resistance of a cross-linked DGEBA resin is shown in Table 19 [140]. Thermal stability is affected by the structure of the hardener. The heat resistance of aliphatic amine cured epoxy is low as measured by TGA. The nitrogen atoms are oxidized by atmospheric oxygen to amine oxides, which attack the polymer backbone. Anhydride systems tend to split off the anhydride at temperatures well below their decomposition point at about 390°C. The ether segments formed by 2-MI and phenolic cured epoxies have the highest thermal stability.

17.4. Epoxy/Curing Agent Stoichiometric Ratios

In addition to the choices of epoxy resins and curing agents, the stoichiometric ratio of epoxy/curing agent is another factor that has significant effects on the network structure and performance. A variety of products are obtained from different ratios. Network formations for a difunctional epoxy resin and a tetrafunctional amine are illustrated in Figure 5. The products range from an epoxy–amine adduct with excess epoxy to an amine–epoxy adduct with excess amine.

Theoretically, a cross-linked thermoset polymer structure is obtained when equimolar quantities of resin and hardener are combined. However, in practical applications, epoxy formulations are optimized for *performance* rather than to complete stoichiometric cures. This is especially true when curing of high MW epoxy resins through the hydroxyl groups.

In primary and secondary amines cured systems, normally the hardener is used in near stoichiometric ratio. Because the tertiary amine

Table 18. Typical properties, chemical resistance, and thermal degradation of liquid DGEBA resin (185 EEW) cured with common hardeners (Dow Chemical data)

	Curing agent				
	TETA[a]	MDA[b]	Polyamide[c]	Anhydride[d]	BF_3–MEA[e]
Property					
phr[f]	13	26	43	87.5	3
Formulation viscosity, Pa · s[g] (°C)	2.25 (25)	0.110 (70)	1.25(50)	0.038(80)	0.040 (100)
Cure schedule, h (°C)	16 (25) 3 (100)	16 (55) 2 (125) 2 (175)	16 (25) 3 (100)	4 (100) 4 (165) 16 (200)	4 (100) 16 (150)
Heat distortion temperature, °C	111	160	101	156	168
Strength, MPa[h]					
Compression	112	116	85.6	126	114
Flexural	96	93	67	97	100
Tensile	79	70.4	57.3	69	39.4
Modulus, GPa[i]					
Compression	3.05	2.6	2.6	3.04	2.3
Flexural	3.05	2.7	2.14	3.05	3.1
Textile elongation,%	4.4	4.4	3.9	2.5	1.6
Dielectric constant at 10^3 Hz	3.90	4.06	3.19	3.14	3.45
Dissipation factor at 10^3 Hz	0.020	0.015	0.0070	0.0054	0.0053
Resistivity at 25°C, 10^{-17} Ω · m					
Volume	6.1	12.2	12.2	6.1	8.6
Surface	7.8	> 7.9	5.5	> 7.3	> 7.9
Chemical resistance					
% Weight gain after 28 d					
50 % NaOH	0.04	−0.05	0.07	−0.12	−0.02
30 % H_2SO_4	1.8	1.6	1.9	0.83	1.1
Acetone	2.1	4.6	7.3	15.0	1.2
Toluene	0.07	0.13	3.7	0.09	0.17
Water	0.86	1.1	1.3	0.82	1.2
Thermal degradation					
% Weight loss after 300 h at 210°C	6.8	5.5	5.0	1.5	4.9

[a] Triethylenetetramine.
[b] 4,4′-Methylenedianiline.
[c] Versamide 140 (Henkel Corp.).
[d] Methylbicyclo[2.2.1]heptene-2,3-dicarboxylic anhydride catalyzed with 1.5 phr benzyldimethylamine.
[e] Methylethylamine.
[f] Parts per hundred epoxy resin.
[g] To convert Pa · s to P, multiply by 10.
[h] To convert MPa to psi multiply by 145.
[i] To convert GPa to psi, multiply by 145,000.

formed in the reaction has a catalytic effect on reactions of epoxy with co-produced secondary alcohols, slightly less than the theoretical amounts should be used. However, if substantially less than the theoretical amount of amine is used, the epoxy resin will not cure completely unless heat is applied (postcure). The use of excess amine will result in unreacted amine terminated dangling chain ends and reduced cross-linking, yielding a polymer that can be somewhat tougher but which is considerably more susceptible to attack by moisture and chemicals. In formulations containing anhydrides, less than stoichiometric ratios of curing agents normally are used (0.50 to 0.85 of anhydride to 1 epoxy stoichiometric ratio) because of significant epoxy homopolymerization.

Ladder studies are often conducted varying the stoichiometric ratios and other factors to determine the optimum formulations. Statistical

Table 19. Effect of hardener structure on reactivity and heat resistance of a cross-linked bisphenol a diglycidyl ether

Curing agent	T_{rmax}, °C	E_a, J/mol[a]	TGA, 4°C/min	
			Weight loss before decomposition	T_{dec}, °C
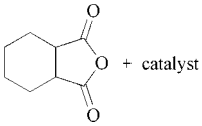 + catalyst	125	92	12	392
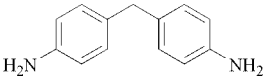	154	50	0	390
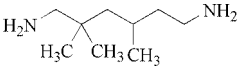	90	58		320
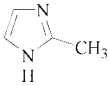 · catalyst	126	67	3.2	420
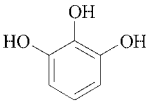 + catalyst	185	54	2.9	400
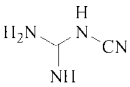	207	125	0	373

[a] To convert J to cal, divide by 4.184.

design of experiment (DOE) methodology has been used to efficiently carry out ladder studies [141]. Information concerning network structures can be obtained using dynamic mechanical analysis (DMA) [142, 143] and chemorheology to guide formulation development [144, 145].

17.5. Catalysts

The choice of a catalyst and of its amount is important. As discussed in previous sections, some tertiary amine catalysts can play multiple roles in the curing reaction. Anhydride cure in particular is highly sensitive to catalyst amount. Nucleophilic catalysts, used with acidic curing agents such as anhydrides and novolacs, can greatly reduce the gel time. In the case of anhydrides, a nucleophilic catalyst attacks the anhydride ring, causing the ring to open and promote bonding to the epoxy ring. Figure 8 shows the effect of BDMA and 1-propylimidazole levels on the pot life of a system combining D.E.R. 331 resin and nadic methyl anhydride at 90°C (194°F) [146]. Imidazoles are more efficient catalysts than tertiary amines; only half the concentration is required to produce the same catalytic effect.

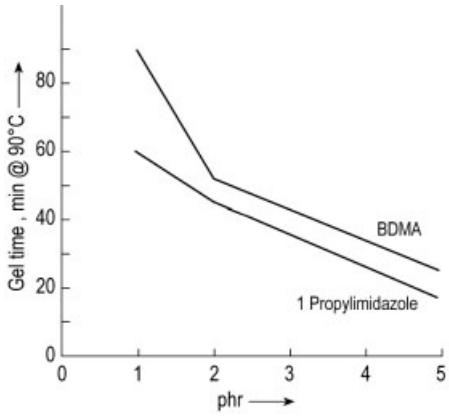

Figure 8. Effects of catalyst on epoxy/nadic methyl anhydride cure

17.6. Accelerators

Accelerators are commonly added to epoxy systems to speed up curing. This term should be used to describe compounds which increase the

rate of catalyzed reactions but which by themselves are not catalysts. However, the term accelerator is often used synonymously with catalyst in some of the literature. Hydrogen donors such as hydroxyl groups facilitate epoxy reactions via hydrogen bonding or reaction with the oxygen on the epoxide ring. More acidic donors such as phenols and benzyl alcohols increase the rate of acceleration. However, very strong acids can interfere with amine curing agents by protonation of the amine to form an amine salt, resulting in increased pot life. Figure 9 shows the effects of different accelerators on the rate of a DGEBA/triethylenetriamine formulation.

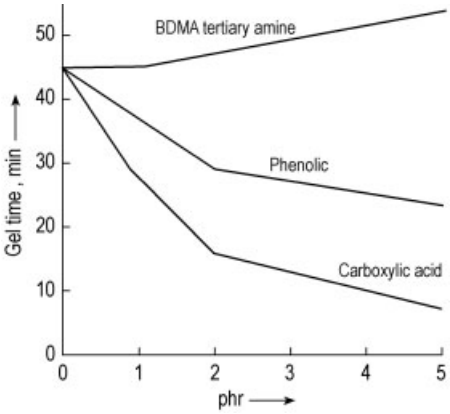

Figure 9. Effects of accelerator on epoxy/triethylene triamine cure

18. Epoxy Curing Process

The epoxy curing process is an important factor affecting the cured epoxy performance. Consequently, it is imperative to understand the curing process and its kinetics to design the proper cure schedule to obtain optimum network structure and performance. Excellent reviews on this topic are available in the literature [147, 148].

The curing of a thermoset epoxy resin can be expressed in terms of a time–temperature-transformation (TTT) diagram (Fig. 10) [149, 150]. Later, a CTP (cure-temperature–property) diagram was proposed as a modification of the TTT diagram [151]. For nonisothermal cure, the conversion-temperature-transformation (CTT) diagram has been shown to be quite useful [152]. In the TTT diagram, the time to gellation and vitrification is plotted as a function of isothermal cure temperature. Important features are the gel point and the onset of vitrification. The gel point is defined as the onset of the formation of insoluble, cross-linked polymer (gel fraction) in the reaction mixture. However, a portion of the sample may still be soluble (sol fraction). The onset of vitrification is when the glass-transition temperature (T_g) of the curing sample approaches the curing temperature T_c. Ideally, a useful structural thermoset would cure until all monomers are built into the network, resulting in no soluble fraction.

The S-shaped vitrification curve and the gelation curve divide the time–temperature plot into four distinct states of the thermosetting-cure process: liquid, gelled rubber, ungelled glass, and gelled glass. T_{g0} is the glass-transition temperature of the unreacted resin mixture; $T_{g\infty}$ the glass-transition temperature of the fully cured resin; and gel T_g the point where the vitrification and gellation curves intersect.

In the early stages of cure prior to gelation or vitrification, the epoxy curing reactions are kinetically controlled. When vitrification occurs the reaction is diffusion controlled, and the reaction rate is orders of magnitude below that in the liquid region. With further cross-linking of the glass, the reaction rate continues to decrease and is eventually quenched. In the region between gelation and vitrification (rubber region) the reaction can range from kinetic to diffusion control. This competition causes the minimum in vitrification temperature seen in the TTT diagram between gel T_g and $T_{g\infty}$. As the cure temperature is raised the reaction rate increases and the time to vitrification decreases until the decrease in diffusion begins to overcome the increased kinetic reaction rate. Eventually, slower diffusion in the rubbery region decreases the overall reaction rate and thus the increase in time to vitrify is seen. Below $T_{g\infty}$, the reaction does not go to completion. As curing proceeds, the viscosity of the system increases as a result of increasing molecular weight, and the reaction becomes diffusion-controlled and eventually is quenched as the material vitrifies [153]. After quenching, the cure conversion can be increased by raising the temperature. This is often practiced as postcure for certain epoxy systems to achieve maximum cure and performance. Postcure is only effective at temperatures higher than $T_{g\infty}$. However, it must be noted that

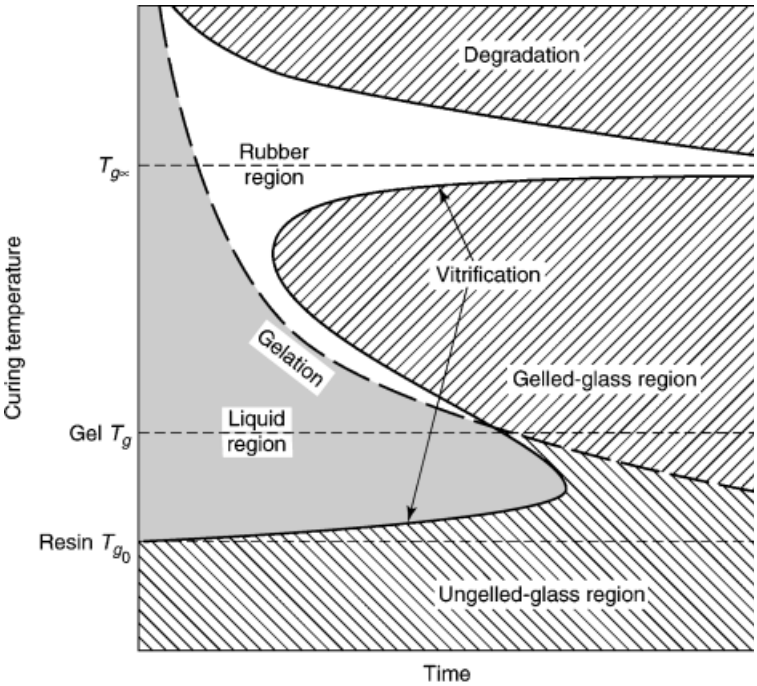

Figure 10. Time–temperature-transformation diagram

at temperatures sufficiently above $T_{g\infty}$, onset of network degradation can also be seen if sufficient time is involved. Thus one must be careful about potential "over-curing."

The TTT diagram is useful in understanding the cure kinetics, conversion, gelation, and vitrification of the curing thermoset. Gelation and vitrification times can be determined from the intersections of the storage and loss moduli and the maxima in the loss modulus of an isothermal dynamic mechanical spectrum, respectively. Recently, techniques have been developed using rheological and dynamic mechanical analysis instruments to determine the gel point and vitrification [154]. Understanding the gelation and vitrification characteristics of an epoxy/curing agent system is critical in developing the proper cure schedule/process to achieve optimum performance.

One important application is the management of cure temperatures (T_c) and heating rate: if T_c is too low, vitrification may occur before gelation and further reactions may not be completed, resulting in an incomplete network structure and poor performance. This is of particular relevance in ambient cures and radiation cures [155]. Furthermore, attention must be paid to the relationship between mixing of reactants and gel point. Epoxy resins and curing agents must be thoroughly mixed prior to the gel point since the rapid viscosity buildup at gel point inhibits homogeneous mixing of reactants, resulting in potential network and morphological inhomogeneities and defects [156].

Curing and quenching processes of epoxies have been reported to affect performance of certain epoxy coatings and composites. These effects have been attributed to phenomena known as internal or residual stress and physical aging of cured epoxies [147].

Internal stresses arise mainly because of the diminishing capacity of the cross-linked polymer to expand or contract to the same extent (volume) with the solid substrate to which it is adhered. This phenomenon is caused by mismatches of coefficients of thermal expansion (CTE) of the substrates (metal, glass, etc) and the cross-linked epoxies during nonisothermal cures; and cure shrinkage (solvent loss, cross-linking). The effect often contributes to adhesion failures and is more prominent in metal coatings and large composite parts manufacturing, especially when the T_g of the cross-linking polymer approaches T_c. As discussed previ-

ously, while curing of epoxy functional groups via polycondensation reaction results in relatively low shrinkage, failures attributable to internal stresses such as delamination have been observed in certain epoxy coatings of metal substrates, epoxy encapsulants for electronic devices and glass-fiber-reinforced composites [157]. The effect can be very severe in the case of photoinitiated curing of epoxy acrylates as well as free-radical curing of epoxy vinyl esters. Shorter bonds are formed during these free-radical curing processes, which result in significant shrinkage. Postcure with heat is often required to release some of the internal stresses and to improve adhesion. Efforts have been focused on understanding the mechanism of stress development, and stress minimization by modifications of the cure and postcure cycles [158].

Physical aging is a well-known phenomenon in glassy polymers and has been studied quite extensively in amorphous thermoplastics [159]. The term physical aging refers to the gradual changes in polymer physical properties with time after a glassy polymer is heated above its T_g and rapidly cooled (quenching) to temperatures below T_g. The physical aging process differs from chemical aging processes, in which breakage or formation of chemical bonds are involved such as continuing cure, hydrolytic aging, and photochemical and thermal degradation. The phenomenon has been attributed to the non-equilibrium state of the glassy polymer at temperatures below its T_g, in which the polymer contains excessive free volume as it is quenched. As the polymer recovers gradually over time to approach equilibrium, a reduction in free volume and an increase in density results. Consequently, the term densification is sometimes used to describe physical aging. For certain epoxy systems, physical aging has been reported to cause increases in stiffness and decreases of toughness [160, 161]. Hardening of certain baked epoxy coatings with time and failures of the coatings due to loss of ductility have been observed. However, physical aging has been reported to be reversible (erasable) by postheating above polymer T_g. Proper selection of the cure and postcure schedules including quenching cycle is important to minimize the potential detrimental effects of physical aging [162]. In some epoxy systems, it is difficult to distinguish physical aging from the effects of residual solvent loss and/or continuing cross-linking. They all can contribute to increases in stiffness of the system.

To develop a proper curing process, it is important to understand the reactivity of different curing agents toward the epoxy structure of interest. The effect of hardener structure on reactivity of a cross-linked DGEBA resin (determined by DSC) is shown in Table 19. Aliphatic amines show a maximum reaction rate, called T_{rmax}, at 90°C (heating rate 10°C/min). The same epoxy resin is somewhat less reactive (T_{rmax} = 126°C) when homopolymerized via initiators. Aromatic amines and phenols cure considerably more slowly, requiring higher curing temperatures. The highest temperatures are required for dicyandiamide curing, which can, however, be accelerated by basic components.

Relative reaction rates are often expressed in terms of the activation energy E_a (Arrhenius type relationship). E_a allows comparisons of reaction rates at different temperatures and is influenced by the type of chemical reactions involved in the cure. Curing of epoxy resins with phenols or aromatic and aliphatic amines proceeds with a fairly low activation energy of 50–58.5 kJ/mol (12–14 kcal/mol). Activation energies are higher when epoxy compounds having low hydroxyl content are cured alone in the presence of catalysts (92 kJ/mol = 22 kcal/mol) or with dicyandiamide (125.5 kJ/mol = 30 kcal/mol).

18.1. Characterization of Epoxy Curing and Cured Networks

Cured thermoset polymers are more difficult to analyze than thermoplastics since they are insoluble and generally intractable. Their properties are influenced by factors at the molecular level, such as backbone structures of epoxy resin and curing agent; nature of the covalent bond developed between the epoxy resin and the curing agent during cross-linking; and density and extent of cross-linking, i.e., degree of cure.

Epoxy resin formulators are concerned with formulation reactivity and flow during application. Reactivity tests or gel time tests are used to determine the proper reactivity of the formulations. Formulators also developed flow tests to check for the formulation rheology profile. The coatings industry widely uses MEK (methyl ethyl ketone) double rubs as an indication of

cure. While the test does give a relative indication of cure for a certain system, caution must be exercised when comparing different systems, which may have very different inherent resistance against MEK. In general, these end-use tests do not provide insights on the structure–property relationship of the system.

Epoxy curing process can be monitored by a number of different techniques:

1) Analysis of the disappearance and/or formation of functional groups
2) Indirect estimation of cure conversion
3) Measurements of changes in thermal, physical, and mechanical properties of the system

Comprehensive reviews of different techniques for epoxy cure monitoring are available [86, 94]. Wet chemical or physical analysis methods, such as solvent swell [163], titration of functional groups, IR, near IR [164], or NMR spectroscopy, are commonly used to monitor epoxy cure.

The thermal properties of the system reflect the degree of cure, and thermal analysis of polymers (DSC, DMA, TGA) has been used extensively in studies of epoxy resins [156]. Correlation between T_g and degree of cure has been well established for many systems.

Viscosity build is observed with increased reaction conversion in epoxy curing. More recently, chemorheology, which utilizes rheological measurement (qv) and thermal analysis such as DSC, has been applied to study epoxy cure [166, 167].

Since epoxy curing involves epoxy ring opening and the generation of polar groups, which have a high dipole moment, dielectric measurements have been applied to monitor cures. Dielectric methods [168, 169] encompass both macroscopic and microscopic features: the dipoles being oriented during dielectric measurements are on a microscopic scale, whereas the degree and rate of orientation may depend on macroscopic properties such as viscosity and density.

The mechanical properties of a resin system can also be used to estimate the degree of cure [170]. The methods range from hardness evaluation to complex static measurements or sensitive dynamic mechanical analysis (DMA). Table 20 gives ASTM standard procedures for measuring the properties of cured or partially cured epoxy resin systems.

Table 20. ASTM[a] procedures for cured or partially cured epoxy resin systems

Test	ASTM standard
Chemical	
Density by displacement	D792
Water absorption in plastics	D570
Moisture absorption properties in composites	D5229
Void content in composites	D2734
Electrical	
Volume resistivity	D257
Surface resistivity	D257
Dielectric strength	D149
Dielectric constant and dissipation factor	D150
Insulation resistance	D257
Thermal	
Heat-deflection temperature	D648
Glass-transition temperature	D696
Dynamic mechanical properties of plastics	D4065
Coefficient of thermal linear expansion	D296
Coefficient of linear thermal expansion by thermomechanical analysis	E831
Coefficient of thermal conductivity	C177
Mechanical	
Tensile strength	D638
Compressive strength (plastic)	D695
Compressive testing (composite)	D3410
Flexural strength	D790
Impact strength	D256
Fracture strength in cleavage of adhesives in bonded metal joints	D3433
Fracture strength in "T" peel of adhesives in bonded joints	D903
Fracture testing in 180° peel of adhesives	D5528
Mode I interlaminar fracture toughness of composites	D2344
Apparent interlaminar shear strength of composites	D5045
Plane strain fracture toughness of plastics	D4255
On-plane shear response of composites	
Hardness, Barcol	D2583
Hardness, Rockwell M	D785

[a] From Ref. [171].

Direct measurement of the cross-link density of thermoset polymers including those from epoxy resins remains one of the most difficult analytical challenges in the field. A far too common approach simply relates the rubbery mod-

ulus (G_r), the thermoset modulus above T_g, to the molecular weight between cross-links (M_c) using the theory of rubbery elasticity [133, 134]. Unfortunately thermoset networks have much more complex features than do true elastomers, including non-Gaussian chain behavior, interchain interactions, and entanglements [172]. These factors render rubbery elasticity theory inadequate as an absolute measure of M_c from G_r, and doing so can lead to totally erroneous conclusions on the network structure [173]. In a given family of thermosets, changes in G_r can be considered to reflect *relative* changes in M_c. Estimates of the expected M_c can be calculated from monomer MW and functionality for stochiometric systems [174]. More extensive network structure calculations including M_c are done using statistical relations developed by Miller and Makosco [175].

In many applications, epoxy systems derive their high thermal and mechanical performance characteristics from highly cross-linked network structures. However, this often results in brittleness of the epoxy thermosets and loss of end-use properties such as impact resistance. Elongation at break (% elongation) has been a popular test used in the industry for many years to measure toughness, ability to resist failure under tensile stress. While useful in certain applications, good correlations between elongation at break and end-use properties of cured epoxies are not always possible. The failure envelope concept has been useful in looking at the entire time–temperature failure spectrum of epoxies [176]. More recently, progress in the field of fracture mechanics [177, 178] has led to advanced fracture toughness tests that are more useful in characterizing cured epoxy performance. Examples of such tests are critical elastic strain release rate (G_{IC}) and critical stress intensity factor (K_{IC}) [179].

Dynamic mechanical analysis (DMA) of cross-linked epoxy resins typically shows, in order of decreasing temperature, an α transition corresponding to T_g, a β transition associated with relaxation of the glyceryl groups, and a γ transition due to methylene group motions [180]. Both the β and the γ transitions, which are typically observed at -30 to $-70°C$ and at about $-140°C$, respectively, are attributed to crankshaft motions of the polymer chain segments. The appearance of transitions between the α and β transitions is highly variable and has been attributed to segmental motions due to particular curing agents [181]. No definitive correlations between the appearance of sub-T_g relaxations and mechanical properties have been observed [182]. Like many other plastics, cross-linked epoxy resins undergo a change in fracture mechanism from brittle to ductile (T_b) with increasing temperature. The window between T_g and T_b has been shown to correlate well with the formability of epoxy can coatings in the draw-redraw (DRD) process [183, 184].

Adhesion is an important issue in epoxy applications since epoxy is almost always used as part of a composite system. Examples are epoxy coatings on metal substrates, epoxy adhesives for metal surfaces, and matrix resin in fiber-reinforced composites such as PCB laminates and aerospace composites. Consequently, optimum epoxy adhesion to the substrate is a prerequisite for good system performance in terms of static and dynamic mechanical properties and environmental durability. In rubber-toughened composite systems, it has been reported that a threshold of interfacial adhesion between both phases (rubber and resin matrix) is needed for maximum toughening by promoting the cavitation mechanism and by activating the crack-bridging mechanism [185]. Excellent review papers are available on the issue of adhesion of epoxy in composites [186], coatings, and adhesives [187]. Effects of internal stresses on coating adhesion failures including the role of coating defects and pigments as potential stress concentrators have been reported [188].

Surface properties such as dynamic contact angle and surface tension are used to ensure proper wetting of epoxy and the substrate. Microscopic techniques, such as scanning electron microscopy (SEM), transmission electron microscopy (TEM), and atomic force microscopy (AFM), are widely used to study morphology, fracture, and adhesion issues of cured epoxy systems. Chemical analysis techniques, such as micro-IR, X-ray photoelectron spectrometry (XPS), and secondary ion mass spectrometry (SIMS), are useful in providing functional group analysis at the interfaces.

Consumers of products which use epoxy resins have developed increasing expectations for longer and more reliable performance. In automobiles, for example, the coating is ex-

pected to maintain its initial "Class A" finish for 10 years and the composite leaf spring is designed to last for the life of the vehicle. To meet these expectations, the long-term durability of epoxy thermosets is a key material-specific and application-specific consideration. The durability of polymeric materials in general depends on phenomena such as physical aging, environmental exposure (such as weathering), and mechanical experience (such as impact and load). A detailed discussion of this topic is beyond the scope of this review; interested readers are referred to a leading reference [189].

In addition, the processing of epoxy formulations into their final thermoset structure and form has a major effect on ultimate performance. Material properties such as rheology and reaction kinetics interplay with processing variables such as temperature and shear rate to affect key properties of extent of cure, orientation, and residual stress. Design of the final form of the material also should incorporate fundamental thermoset properties using finite element analysis methods. Optimization of any given epoxy thermoset application is therefore very specific to formulation, processing conditions, and final form and use of the material, and involves the contributions from chemistry, engineering, and material science disciplines to be fully successful.

19. Formulation Modifiers

The processing behavior (mainly viscosity and substrate wetting) and other properties of an epoxy system can be modified by diluents, fillers, toughening agents, thixotropic agents, etc. Most commercial epoxy resin systems contain modifying agents.

19.1. Diluents

Diluents affect the properties of the cured resin system and, in particular, lower the viscosity in order to improve handling and wetting characteristics. They are often used in the range of 2–20 wt. % based on the epoxy resin. Diluents can be classified into *reactive* and *nonreactive* types.

The *reactive* diluents are products with low viscosity (1–500 cP at 25°C) used to lower the viscosity of standard epoxy formulations. The effect of reactive diluents on DGEBA viscosity is illustrated in Figure 11. Lower viscosity allows higher filler loading, lower costs, and/or improved processability. Because of the epoxy functionality, the diluents become part of the cured network. However, the reactive diluents can negatively impact properties, and so balancing of viscosity reduction and property loss is an important consideration. Decreases in tensile strength, glass-transition temperature, chemical resistance, and electrical properties are usually observed. Toxicity is another concern, particularly the aromatic mono glycidyl ethers such as phenyl glycidyl ether (PGE) and *o*-cresol glycidyl ether (CGE). *n*-Butyl glycidyl ether (BGE) is one of the most efficient viscosity reducers, but it has been losing favor because of its volatility and noxiousness. Longer chain alkyls, polyfunctional or aromatic glycidyl ethers such as bisphenol F epoxy, neopentylglycol diglycidyl ether, and triglycidyl ether of propoxylated glycerine are gaining popularity as epoxy reactive diluents. Cycloaliphatic epoxies and glycidyl esters of acids such as neodecanoic acid are also used as reactive diluents.

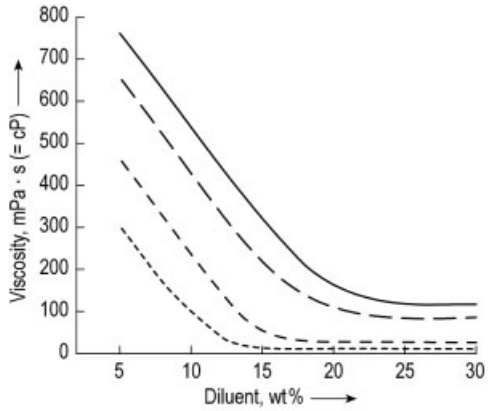

Figure 11. Reduction of DGEBA viscosity by reactive diluents: ——, *o*-cresol glycidyl ether; – – –, butanediol diglycidyl ether; ———-, C_{12}–C_{14} aliphatic glycidyl ether (Epoxide 8); ·····, *n*-butyl glycidyl ether.

Acrylics such as 1,6-hexanediol diacrylate and trimethylolpropane triacrylate are nonepoxy multifunctional diluents, which react readily with primary and secondary amines by means of Michael addition of the the amine to the acrylic double bond [190]. They have been used to increase cure speed or to lower cure temperature of

Table 21. Typical properties of fillers

Name	Composition	Particle shape	Surface area volume	Bulk density, kg/m^3	Characteristics and main use
Marble flour, dolomitic	magnesium–calcium carbonate	granular	medium	1120–1300	general-purpose fillers, particularly recommended for castings requiring machining
Chalk powder	precipitated calcium carbonate	crystalline	high	800–880	
Sand	quartz, feldspar, and subsidiary minerals	spheroidal	low	1500–1700	bulk filler giving high compressive strength and abrasion resistance; difficult to machine
Silica flour	ground quartz	granular	medium	1100–1150	standard filler for large electrical castings; high abrasion resistance; difficult to machine
Mica flour	muscovite[a]	lamellar	high	300–400	filler giving high crack resistance to castings exposed to mechanical and thermal shock
Slate powder	slatea	mainly lamellar	medium	700–900	general-purpose filler giving high abrasion resistance; difficult to machine
Vermiculite[b]	vermiculite[a]	exfolidated laminae	high	100–150	fillers giving lightweight bulk in cores or thick backing to increase the rigidity of thin sections
Phenolic microballoons	phenolic resins	hollow spheres	medium	100–150	
Zircon flour	zircon[a]	granular	medium	1700–1900	filler giving high abrasion resistance; difficult to machine
Aluminum powder	metallic aluminum	granular	medium	1000–1100	filler imparting thermal conductivity, e.g., to prevent excessive temperature buildup in electrical components or in tools for hot-forming plastics
Chopped glass strand[c]	low alkali glass	fibrous	medium	100–250	fillers improving the mechanical strength of prominent edges and thin sections
Hydrated aluminum oxide	alumina trihydrate	granular	medium	700–1300	filler improving wet and dry arc-track resistance and flame retardance

[a] Silicate.
[b] Grain size = 0.15–0.32 cm.
[c] Length = 0.60 cm.

epoxy–amine systems. Caprolactone acrylates have also been used for this application [191].

Solvents and plasticizers are *nonreactive* diluents. The most common nonreactive diluents are nonyl phenol, furfuryl alcohol, benzyl alcohol, and dibutyl phthalate. These materials have the advantage of being able to add to the amine side of the system to better balance mix ratios. Nonyl phenol and furfuryl alcohol also improve wet-out and accelerate cure slightly. They are also capable of reacting with the epoxy group under high temperature cure conditions. Benzyl alcohol is a popular diluent used with amine-cured systems. In addition to viscosity reduction,

it is also known to increase cure speed. Benzyl alcohol can be used up to 10 wt. % level without significant effects on cured properties. Dibutyl phthalate is widely used as a nonreactive diluent for liquid resins. However, performance properties will drop off more quickly with increasing levels of nonreactive diluents than with increasing levels of reactive diluents.

Aromatic hydrocarbons, such as toluene or xylene, significantly reduce the viscosity of liquid DGEBA resins, but their use can be accompanied by a 15–25 % decrease in compressive yield strength and a 10–20 % reduction in compressive modulus. If the solvent is trapped in the cured system, solvent resistance is reduced and cracks develop if the resin is used in heat-cured castings. The use of solvents and reactive diluents in epoxy systems is reviewed in References [192] and [193].

19.2. Thixotropic Agents

Thixotropy is the tendency of certain colloidal gels to flow when subjected to shear, and then to return to a gel when at rest. A thixotropic gel can be produced through the addition of either high surface area fillers such as colloidal silicas and bentonite clays or of chemical additives. Thixotropy is desirable in applications such as encapsulation where the coating is applied by dipping. The resin will wet out and coat the object being dipped, but will not run off when the object is removed from the dipping bath.

19.3. Fillers

Fillers are incorporated in epoxy formulations to enhance or obtain specific desired properties in a system. The type and amount of filler used are determined by the specific properties desired. Fillers can also reduce the cost of epoxy formulations. Inert commercial fillers (qv) can be organic or inorganic, and spheroidal, granular, fibrous, or lamellar in shape. The properties of commercial fillers are given in Table 21, and some effects on epoxy resins are shown in Tables 22 and 23. Some formulations contain up to 90 wt. % fillers. For certain applications, fillers can have significant effects on thermoset morphology, adhesion, and the resulting performance.

Table 22. Effect of fillers

Advantages	Disadvantages
Lower cost of product	increased weight[a]
Reduced shrinkage upon curing	loss of transparency
Decreased exothermic temperature rise on curing[a]	tendency to entrap air
Increased thermal conductivity[a]	difficulty of machining hard fillers
Reduced expansion and contraction with temperature change	decreased impact and tensile strengths
Higher deflection temperature	increased dielectric constanta and power factor[a]
Improved heat-aging properties[a]	
reduced water absorption[a]	
Improved abrasion resistance[a]	
Increased surface hardness[a]	
Increased compressive strength[a] and Young's modulus[a]	
Increased electric strength[a]	

[a] Certain fillers, such as vermiculite and phenolic microballoons, have the reverse effect.

Filler loading is often limited for a given application by the maximum viscosity allowable and/or the reduction in some mechanical properties such as tensile and flexural strength in the cured material. Viscosity can be modified by heat or by addition of a reactive diluent; heating is preferred since diluents affect overall system properties. Some of the major property enhancements affected by fillers are described below.

Fillers can increase pot life and lower exotherm of epoxy systems. Fillers reduce the reactant concentration in the formulation and act as a heat sink. Generally, they have higher heat capacities than the epoxy resins. They are also better heat conductors than the resins, and thus help to dissipate exotherm heat more readily. Commonly used fillers are silica, calcium carbonate, alumina, lithium aluminum silicate, and powdered metals.

Fillers help to increase thermal shock resistance and to decrease the thermal expansion coefficient of an epoxy system by replacing part of the resin with a material that does not change its volume as significantly with temperature variations. Such fillers are clay, alumina, wood flour, sawdust, silica, and mica. Epoxy molding compounds (EMC) can contain up to 90% of fused silica to manage the thermal stress experienced by encapsulated semiconductors. Powdered metals are

Table 23 Influence of Filler on Epoxy Resin Properties[a]

Filler	Ease of dispersion	Antisettlin	Cost	Shrinkage	Exotherm	Thermal conductivity	Coefficient of expansion	Deflection	Weight	Machinability	Abrasion resistance	Impact strength	Tensile strength	Flexural strength	Compressive strength	Thixotropy	Dielectric constant	Power factor	Protection against: Electric strength	Protection against: Mineral acids	Protection against: Caustic alkalies
Marble flour, dolomitic	P	N	–	–	–	+	–	+	+	O	O	–	–	–	+	O	+	O	O	– –	–
Chalk powder	N	·	–	–	–	+	–	O	+	O	O	– –	–	–	O	O	+	O	O	– –	–
Sand	P	N	– –	–	–	+	– –	+	+	–	+ +	–	O	O	+	O	+	+	+	O	O
Silica flour	·	N	–	–	–	+	– –	+	+	–	+ +	–	O	O	+ +	O	+	+	+	O	O
Mica flour	N	·	–	–	–	+	–	+	+	–	+	+	O	–	O	O	+	+	+	O	O
Slate powder	P	N	–	–	–	+	–	+	+	–	+ +	–	O	O	+	O	+	+	+	O	O
Vermiculite	N	P	– –	O	+	–	O	O	– –	+	– –	– –	– –	– –	– –	O	–	–	– –	[b]	[b]
Phenolic microballoons	P	P	– –	O	+	–	O	O	– –	+ +	– –	– –	– –	– –	– –	O	–	–	– –	[b]	[b]
Zircon flour	P	N	–	–	–	+	– –	+	+ +	– –	+ +	–	O	O	+	O	+	+	O	O	O
Aluminum powder	P	N	O	–	–	+ +	–	+	+	+	–	–	O	O	+	O	+ +	O	– –	–	–
Chopped glass strand	N	P	O	–	–	+	–	+	+	–	+	+	+	+	+	O	+	+	O	O	O
Hydrated aluminum oxide	N	P	–	–	–	+	–	+	+	O	O	– –	–	–	+	O	+	+	O	O	O

[a] Key: P = positive effect; N = negative effect; O = no significant effect; – = significant decrease; – – = large decrease; + = significant increase; + + = large increase; · = fillers taken for arbitrary standard for comparison of dispersibility and setting.

[b] Porosity of filler reduces protection provided by resin.

used when bonding metals together to better match the coefficient of thermal expansion of the bond with that of the metal, thus minimizing thermal stress.
Using fillers as a partial replacement for a reactive resin that shrinks on curing can reduce shrinkage of the system. Any inert filler will decrease shrinkage, but the most commonly used are silica, calcium carbonate, alumina talc, powdered metals, and lithium aluminum silicate.
The addition of fillers can increase the machinability and abrasion resistance of an epoxy resin system by increasing the hardness of the thermoset. Greater hardness leads to a higher energy required to scratch but cleaner cuts upon machining. Fillers used for this purpose are powdered metals, wood flour, calcium carbonate, sawdust, clay, and talc.
In certain applications, conducting fillers are added to epoxy formulations to reduce the good insulating properties of the epoxy systems. The most commonly used fillers are graphite and powdered metals.

Other properties that can be affected with the proper choice of fillers for a specific application include compressive strength, adhesion, arc and tracking resistance, density, and self-lubricating properties.

19.4. Epoxy Nanocomposites

Significant recent developments in polymer property enhancement involve polymer nanocomposites. This is a special class of fillers (mostly clay derivatives) in which the nanoscale, highly oriented particles are formed in the polymer matrix through monomer intercalation and particle aggregate exfoliation [194, 195]. The objective is to combine the performance attributes of both hard inorganic and plastic materials. Significant efforts have been dedicated to develop epoxy nanocomposites in the past decade. Improvements in electrical and mechanical properties, chemical resistance, high temperature performance, and flame retardancy have been reported. Other silica-based organic hybrids have been developed [196] for military and aerospace applications.

The emerging field of nanotechnology has produced new materials such as the carbon nanotubes, which are filaments of carbon with atomic dimensions. Recent publications claimed exceptional property enhancement from nanotube-laced epoxy (hardness, electrical and heat-conducting properties) [197]. However, cost remains a barrier for commercialization.

19.5. Toughening Agents and Flexiblizers

Some cross-linked, unmodified epoxy systems exhibit brittleness, poor flexibility, and low impact strength and fracture resistance. Modifiers can be used to remedy these shortcomings. However, there usually will be some sacrifices of properties. In general, there are two approaches used to modify epoxies to improve these features.

1) *Flexiblization.* Aliphatic diepoxide reactive diluents enhance the flexibility or elongation by providing chain segments with greater free rotation between cross-links. Polyaminoamide hardeners, based on aliphatic polyamines and dimerized fatty acids, perform similarly. Liquid polysulfide polymers possessing terminal mercaptan functionality improve impact properties in conjunction with polyamine hardeners.
   Flexible chain segments are incorporated in an epoxy resin by many means [189]. One approach is the incorporation of oligomeric aliphatic polyesters containing carboxylic acid end groups, forming an epoxy resin adduct. This is one of the reasons that epoxy–polyester hybrid powder coatings have become very popular. The effects of flexiblizers are shown in Table 24. Flexiblization can enhance elongation of the system but is often accompanied by a reduction of glass-transition temperature, yield stress, and elastic modulus. Other properties (e.g. water absorption and thermal and chemical resistance) may also be affected.
2) *Toughening* refers to the ability to increase resistance to failure under mechanical stress. Epoxies derive their modulus, chemical, and thermal resistance properties from cross-link density and chain rigidity. Increasing

cross-link density to meet higher thermal requirements (T_g) often comes at the expense of toughness. Toughening approaches for epoxies [199 – 202] include the dispersion of preformed elastomer particles into the epoxy matrix and reaction-induced phase separation of elastomers or thermoplastic particles during cure.

Elastomers such as carboxyl-terminated poly(butadiene-*co*-acrylonitrile)s (CTBN) have been popular tougheners for epoxies. Toughening by elastomers can be attributed to the incorporation of a small amount of elastic material as a discrete phase of microscopic particles embedded in the continuous rigid resin matrix. The rubbery particles promote absorption of strain energy by interactions involving craze formation and shear deformation. Craze formation is promoted by particles of 1–5 μm size, and shear deformation by particles > 0.5 μm. Systems possessing both small and large particles, i.e., bimodal distribution, provide maximum toughness [203]. The rubber is incorporated in the epoxy resin in a ratio of 1:8 in the presence of an esterification catalyst. The product is an epoxy ester capped with epoxy groups. The adduct is then formulated with unmodified resin and cured with standard hardeners and accelerators. Phase separation, of the adduct occurs during the curing process, resulting in the formation of segregated domains of elastomer-like particles covalently bound to the epoxy resin matrix. Optimum particle size and particle-size distribution, phase separation, and phase morphology are crucial for the development of desirable properties of the system. If the elastomer remains soluble in the epoxy matrix, it serves as a flexiblizer and reduces the glass-transition temperature significantly. Some reductions in T_g and modulus are typical of CTBN-modified epoxies. Amine-terminated poly(butadiene-*co*-acrylonitrile)s (ATBN) are also available [68].

Elastomer-modified epoxy resins are used in composites and structural adhesives, coatings, and electronic applications. A similar approach to toughen epoxy vinyl esters using other elastomeric materials has been reported [204]. Other elastomer-modified epoxies include epoxy-terminated urethane prepolymers, epoxy-terminated polysulfide, epoxy–acrylated urethane, and epoxidized polybutadiene. Preformed dispersions of epoxy-insoluble elastomers have been developed and reported to achieve toughening without T_g reduction [205, 206].

Other epoxy toughening approaches include chemical modifications of the system either through the epoxy backbone and/or crosss-linker. Dow Chemical developed a cross-linkable epoxy thermoplastic system (CET) [207]. The concept involves introducing stiffer polymer segments into the network structure to maintain the glass-transition temperature while allowing cross-link density reduction to improve toughness. Thermoplastics, core-shell rubbers (CSR), and liquid crystal polymers (LCP) have also been used. Semi-interpenetrating network (IPN) approaches involve formation of a dispersed, cross-linked epoxy second phase in a thermoplastic matrix. The systems were reported to have good combinations of toughness, high T_g, high modulus, and processability.

Incorporation of block copolymers has been shown to improve toughness of certain epoxy systems [208]. More recently, nanocomposites and self-healing epoxy systems [209] represent new approaches to develop more damage-tolerant epoxies.

Through the proper selection of resin, curing agent, and modifiers, the cured epoxy resin system can be tailored to specific performance characteristics. The choice depends on cost, processing and performance requirements. Cure is possible at ambient and elevated temperatures. Cured epoxies exhibit good combinations of outstanding properties and versatility at moderate cost: excellent adhesion to a variety of substrates; outstanding chemical and corrosion resistance; excellent electrical insulation; high tensile, flexural, and compressive strengths; good thermal stability; relatively low moisture absorption; and low shrinkage upon cure. Consequently, epoxies are used in diverse applications.

20. Coatings Applications

Commercial uses of epoxy resins can be generally divided into two major categories: protec-

Table 24. Effect of flexiblizers

Flexiblizers	Concentration, %	Advantages	Disadvantages
Poly(propylene glycol) diglycidyl ether	10–60	low viscosity, good flexibility	poor water resistance fair impact resistance
Polyaminoamides	30–70	good abrasion resistance, good flexibility	fair chemical resistance
Liquid polysulfides	10–50	good corrosion resistance, excellent flexibility	odor Poor heat resistance tendency to cold flow
Aliphatic polyester adducts	10–30	good water resistance fair flexibility over a range of temperatures	high viscosity

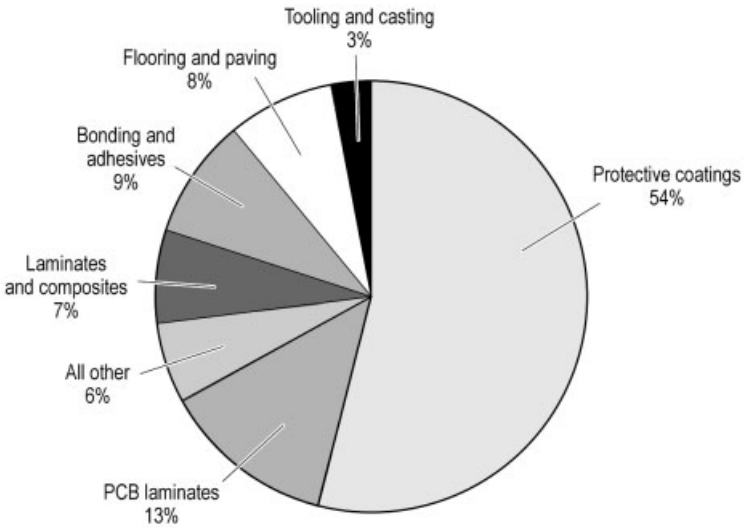

Figure 12. End-use markets of epoxy resins (U.S. data, 2000)

tive coatings and structural applications. U.S. consumption of epoxy resins is given in Figure 12. The largest single use is in coatings (> 50%), followed by structural composites. Among the structural composite applications, electrical laminates contribute the largest epoxy consumption. A similar trend is observed for the European market, but the Asian consumption is heavily tilted toward electrical laminate and electronic encapsulant applications [210]. Electrical and electronic applications account for the largest consumption of epoxy resins in Japan (> 40%). In 2000, it is estimated that the Asia-Pacific region consumed up to 70% of all epoxies used in electrical laminate production worldwide. While the overall epoxy markets continue to grow at a steady pace over the past two decades, more rapid growth has occurred in powder coatings, electrical laminates, electronic encapsulants, adhesives, and radiation-curable epoxies.

The majority of epoxy coatings are based on DGEBA or modifications of DGEBA. Chemical and corrosion-resistant films are obtained by curing at ambient and/or elevated temperatures. Ambient temperature cured coatings primarily involve cross-linking of the epoxy groups in mostly two-package systems, while elevated temperature cured coatings in one-package systems take advantage of the reactivity of both the epoxy and the secondary hydroxyl groups. As a class, epoxy coatings exhibit superior adhesion (both to substrates and to other coatings), chemical and corrosion resistance, and toughness. However, epoxy coatings have been employed mainly as primers or undercoats because of their tendency to yellow and chalk on exposure to sunlight.

Epoxy-based coatings are the preferred and dominant choices for cathodic electrodeposition of automotive primers, marine and industrial maintenance coatings, and metal container interior coatings. Use of epoxy flooring for institutions and industrial buildings has been growing at a steady rate as the industry becomes more aware of its benefits.

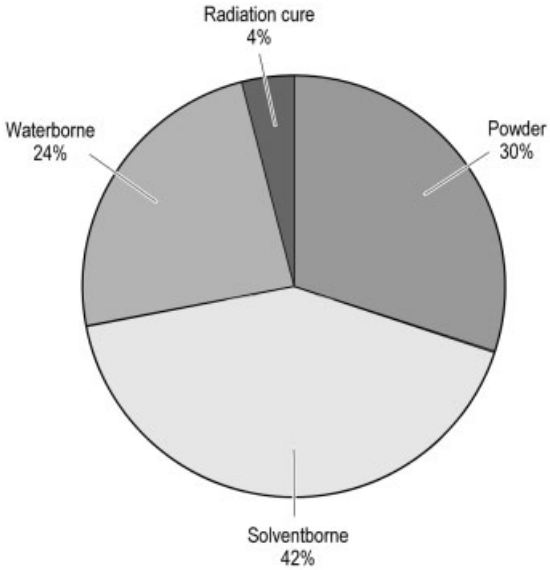

Figure 13. Global epoxy coating application technologies [211]

Solvents are commonly used to facilitate dissolution of resins, cross-linkers, and other components, and for ease of handling and application. Although most of the epoxy coatings sold in the 1970s were solventborne types, they made up only 40% of epoxy coating consumption in 2001 [211]. Economic and ecological pressures to lower the volatile organic content (VOC) of solventborne coatings have stimulated the development of high solids, solvent-free systems (powder and liquid), and waterborne and radiation-curable epoxy coatings technologies [212]. These environmentally friendly coating technologies have experienced rapid growth in the past decade. For example, epoxy powder coatings have been growing at rates exceeding those of other coating technologies as new applications such as automotive primer-surfacer and low temperature cure coatings for heat-sensitive substrates are developed. Radiation-curable liquid coatings based on epoxy acrylates and cycloaliphatic epoxies have also been growing significantly over the last decade. The current distribution of coating technologies is summarized in Figure 13.

20.1. Coatings Application Technologies

20.1.1. Low Solids Solventborne Coatings

These traditional low solids coatings contain less than 60% solids by volume (typically 40%). Their advantages include established application equipment and experience, fast drying and cure at ambient temperatures and excellent film formation at extremely fast cure conditions like those used in coil coatings (< 30 s, $> 200°C$). However, because of stricter VOC regulations, solvent-based coatings have been losing market share steadily to more environmentally friendly technologies.

20.1.2. High Solids Solventborne Coatings

High solids coatings contain 60–85% by volume of solids. They are mostly based on standard LERs or low molecular weight SERs modified by reactive diluents, low viscosity multifunctional aliphatic epoxies, or bisphenol F epoxy resins. High film build is one key advantage of high solids coatings. Examples include the coal-tar epoxy coatings that contain up to 85% solids used in industrial protective coatings.

20.1.3. Solvent-Free Coatings (100% Solids)

Ecological concerns have led to increasing uses of these materials. Low viscosity LERs based on bisphenol A and bisphenol F epoxies are often used in combination with reactive diluents. The advantages include high buildup in a single application, minimization of surface defects owing to the absence of solvents, excellent heat and chemical resistance, and lower overall application costs. Disadvantages include high viscosity, difficulties to apply and produce thin films, poor impact resistance and flexiblity, short pot life, and increased sensitivity to humidity. Weatherable cycloaliphatic epoxies can be used to formulate solvent-free thermally curable coatings because of their low viscosities [213].

20.1.4. Waterborne Coatings

In the switch from solventborne to waterborne systems, epoxies are successfully bridging the gap largely by adaptation of conventional resins. Waterborne coatings accounted for almost 25% of epoxy coating consumption in 2001.

In addition to the waterborne epoxy dispersions which are typically supplied by epoxy resin producers, significant advances in waterborne coatings have been made by coatings producers such as PPG Industries, ICI Paint, and others utilizing modified epoxies. PPG coatings are used in cathodic electrodeposition systems that are widely accepted for automobile primers. Many patents have been issued for this important technology [214]. The Glidden Co. (now ICI Paint) developed a waterborne system for container coatings based on a graft copolymerization of an advanced epoxy resin and acrylic monomers [215]. These two waterborne epoxy coatings were significant breakthroughs in the coatings industry in the 1970s and are still widely used today.

For ambient temperature cure applications such as industrial maintenance and marine coatings, LERs or low molecular weight SERs (type 1 resin) are dispersed in water with a surfactant package and small amounts of co-solvents [216]. Some producers offer waterborne curing agents that, typically, are salts of polyamines or polyamides. Key disadvantages include higher costs, slow cure at ambient and humid conditions, and tendency to cause flash rush. In addition, expensive stainless steel equipment are required for application. Recent developments include the elimination of co-solvents in some epoxy dispersions [217]. Custom synthesized acrylic latexes have shown promise when thermally cured with cycloaliphatic epoxies [218]. While the overall volume is still relatively modest (estimated at $< 20{,}000$ MT in 2000 for the global market), it is expected that future growth rate for this segment will be much higher than standard epoxy resins, particularly in Europe where environmental pressures are stronger.

20.1.5. Powder Coatings

Epoxy-based powder coatings exhibit useful properties such as excellent adhesion, abrasion resistance, hardness, and corrosion and chemical resistance. The application possibilities are diverse, including metal finishing, appliances, structural rebars, pipes, machinery and equipment, furnitures, and automotive coatings. Together, these applications accounted for 30% of epoxy coatings (Fig. 13) and 17% of epoxy resin consumption globally in 2001 (Fig. 14). This is a high growth segment of epoxy coatings (see → Paints and Coatings, Sect. 11.1).

The development of highly reactive powder systems which cure using low energy (150°C) and the possibility of economical thin films (30–40 μm) have made powder coatings competitive with waterborne and high solids systems. Powder coatings can be applied by fluidized-bed (thick films, 50–150 μm) or electrostatic spray (thin films, 30–40 μm).

In powder coatings, epoxies are continuing to grow at rates exceeding other technology segments mainly because of the 100% solids feature, improved coverage, and recyclability of overspray materials. Pipeline projects, important because of worldwide energy problems, are significant consumers of epoxy powder coatings. The value of improved service life is being increasingly accepted even at the somewhat higher material cost of epoxy systems. Four types of epoxy resin systems are commonly used as powder coatings.

1) *Epoxy powder coatings* are based on SERs of intermediate molecular weight (800–2000 EEW). They provide good flow and reactive terminal epoxy functionality. The properties of these thermoset coatings depend on the curing agents, which are friable solids such as dicyandiamide (DICY), phenol-terminated epoxy hardener, and anhydrides. Epoxy powder coatings are generally employed for interior or undercoat uses. Functional epoxy powder coatings are thick films (0.1–0.5 mm, 5–20 mil) used to protect automotive and truck parts, pipe, and concrete reinforcing bars. Fusion-bonded-epoxies (FBE), first developed by 3M Co., are epoxy powder coatings used to protect oil and gas pipelines where long-term corrosion protection under adverse conditions (for example, the Alaska oil and North Sea underwater pipelines) is critical. The performance requirements for FBEs are challenging as the hard, cross-

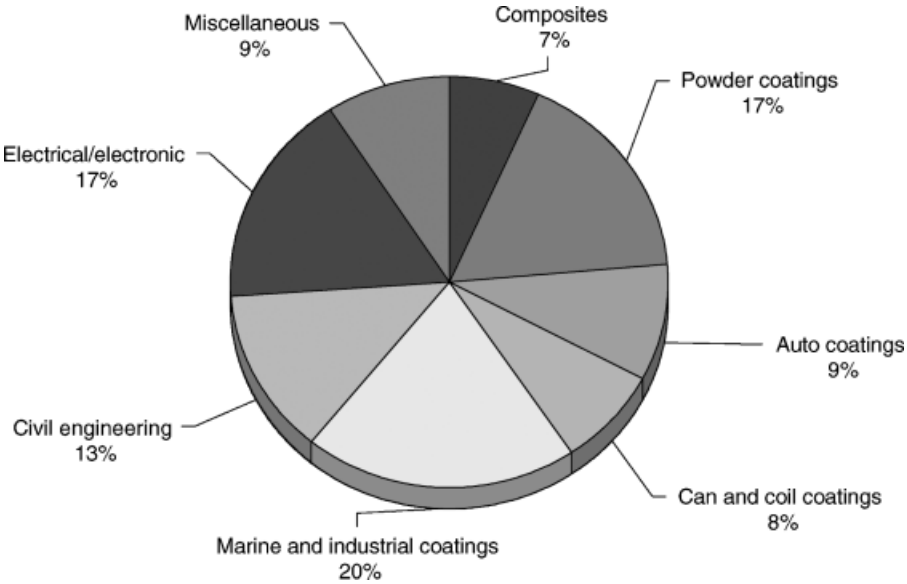

Figure 14. Major global epoxy resin markets (Dow Chemical data, 2001)

linked coatings are expected to survive the pipe bending/unwinding processes and handling abuses in the field. FBEs are also used to protect rebars embedded in critical concrete structures such as bridges, tunnels, and highways. Their primary function is to extend the lifetime of the concrete structures (5–10 years when built with uncoated rebars) to 20–30 years, reducing maintenance and repair costs. Epoxy powder coatings also serve as electrical insulation for bus bars, motor armatures, and similar articles. Decorative epoxy coatings are applied as a thin film (0.02–0.1 mm, 1–5 mil) and used mainly in appliance and general metal product applications. Coating for heat-sensitive substrates is an emerging market for epoxy powders.

2) *Epoxy–polyester hybrids* are mixtures of solid epoxy resins based on bisphenol A and acid-terminated polyester solid resin (25–85 acid equivalent weight). These hybrids are typically less expensive than the epoxy-based powder coatings and offer improved weatherability, and better resistance to overbake yellowing while retaining many of the properties of the standard epoxies. Corrosion resistance is equivalent to epoxy powders in most cases, although solvent and alkali resistance is inferior. One significant new application of the epoxy–polyester hybrids is the primer-surfacer coating for automobiles. Primer-surfacer coatings based on epoxy–polyester hybrid is applied in between the epoxy primer and the topcoats. Its functions are to provide intercoat adhesion and to improve the chip resistance properties of the coatings. Automakers have also found that the epoxy–polyester hybrid primer-surfacer give a smoother surface under the top coats, resulting in a better quality appearance [219]. Epoxy–polyester hybrids have experienced exceptional growth in the global market in the past decade.

3) *Polyester–TGIC,* a third type of epoxy powder coating, is based on a mixture of polyester polycarboxylic acids (18–37 acid equivalent mass) and triglycidyl isocyanurate (TGIC). The TGIC-cured powder coatings have excellent UV resistance, good gloss and color retention, as well as good adhesion and mechanical properties. They were originally developed in Europe for coating metal window frames and buildings, exterior siding, outdoor hardware, high quality outdoor furniture, and other articles requiring superior outdoor durability. These polyester–TGIC powder coatings have gained popularity worldwide.

   In recent years, there have been concerns over toxicity of TGIC, and a number of po-

tential replacement compounds have been developed. Among these, β- hydroxyalkylamide (HAA), trade named Primid by EMS-Chemie is gaining in popularity, particularly in Northern Europe.

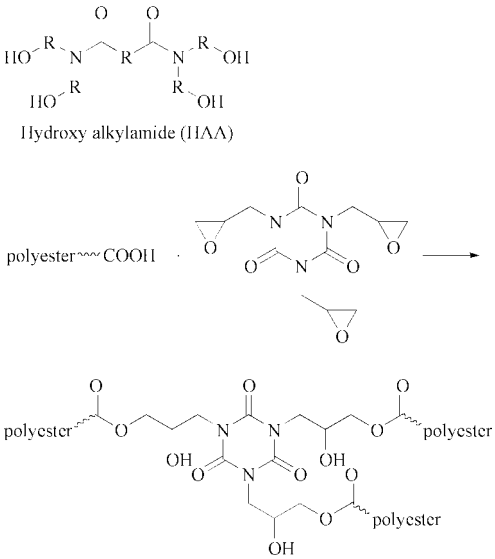

4) *Glycidyl methacrylate–acrylic.* These powder coatings are based on copolymers of glycidyl methacrylate (GMA) and acrylic monomers. They are often thermally cured with dodecanedioic acid (DDDA) and are used in automotive primer-surfacer and clear coats of luxury automobiles such as BMW [220, 221].

Recent developments in powder coatings based on epoxy include UV-curable GMA–acrylic coating for automotive parts; lower temperature cure coatings for heat-sensitive substrates such as wood [222, 223] and plastics; and dual cure (thermal/UV) systems.

20.1.6. Radiation-Curable Coatings

UV and electron beam (EB) radiation curable coatings [74] is a fast-growing segment of epoxy coatings, increasing at 8–10% annual growth rate. The technology is environmental-friendly. No solvent is used, and volatile emission is essentially eliminated. Cure is highly effective and energy-efficient at ambient temperatures, lending the technology highly applicable to heat-sensitive substrates such as wood, plastics, and paper. Capital cost requirements are low especially when compared to new thermal ovens for solventborne and waterborne coatings. EB cure is a relatively new technology which initiates cure via highly energetic electron beams, and unlike UV cure, no photoinitiator is needed. However, EB capital cost is higher.

Epoxy acrylates are widely used as the base resin in many UV-initiated free-radical cure varnish formulations. Epoxy acrylates provide varnishes with excellent scuff resistance, high gloss, and good adhesion. Major markets are overprint varnishes for papers (books, magazines, cards, labels, etc) and exterior can coatings. Wood furniture and particle board are new but growing markets for this technology. Alternatively, cycloaliphatic epoxies can be UV-cured via a photo-initiated cationic mechanism. They are used in metal container exterior overprint varnish and inks, and high performance electronic applications. While cycloaliphatic epoxies are more expensive than epoxy acrylates, they offer several advantages: better adhesion to metals, fewer hazards in handling, and continued curing in the dark (which is important in certain applications). A related and high value market is inks and resists, where radiation and heat-curable epoxies and epoxy acrylates are used. As discussed in the powder coatings section, radiation-curable epoxy powder coatings are being developed for a number of applications.

20.2. Epoxy Coatings Markets

The major global market segments of epoxies are represented in Figure 14. The marine and industrial protective coating is the largest market for epoxy coatings, followed by powder coatings, automotive, and container coatings.

20.2.1. Marine and Industrial Maintenance Coatings

The combined marine and industrial maintenance coatings application constitutes the largest epoxy coating market segment globally. While the end-use markets and application requirements are different, the basic epoxy systems

utilized in these two markets are quite similar. The basic function of these coatings is to protect metal and concrete structures from degradation in aggressive environments for extended periods of time. The long service life of the coating and/or extended intervals between repairs are critical requirements, especially for marine applications because of the high costs of dry-docking of ships for re-painting. The excellent corrosion, abrasion, and chemical resistance properties of epoxy coatings allow their dominant position in these markets. They are used in new construction as well as in maintenance and repair works. Examples are corrosion-resistant coatings for ships, shipping containers, offshore oil rigs and platforms, transportation infrastructures such as bridges, rail car coatings, coatings for industrial storage tanks, and primers for light industrial and agricultural equipment.

Most coatings used in these markets are two-component systems applied and cured at ambient conditions. LERs and low molecular weight SERs based on bisphenol A and bisphenol F epoxies are commonly used [224]. Aliphatic polyamines, adducts of epoxy resins with aliphatic and aromatic amines, ketimines, phenalkamines, amidoamide, and polyamide resins are employed as curing agents. The working pot life of the amine–epoxy resin systems depends on the curing agent, solvents, catalysts, and temperature. High solids solventborne coatings are most popular. Tighter VOC regulations have facilitated the development of lower VOC, 100% solids, and waterborne epoxy coating systems. Important types of epoxy coatings in this segment include the following.

1) *Two-component epoxy–amine coatings* are used primarily as a primer or mid-coat over the inorganic zinc-rich primer coating. High solids epoxy mastics can be applied over contaminated substrates and form thick, good barrier coatings. These coatings account for the majority of epoxies used in marine and industrial protective coatings markets.
2) *Organic (epoxy) zinc-rich primers* are used in place or to repair imperfections of the inorganic zinc-rich primer. Their advantages over inorganic zincs include improved adhesion to the epoxy primer coating and better tolerance to poor surface cleaning.
3) *Coal-tar epoxies* are historically some of the most popular high solids epoxy coatings, having excellent water barrier, chemical resistance properties, and low costs. They are typically cured with polyamides and are used as ship bottom or primer coatings for tanks, pipes, and steel pilings. However, their use has been declining or banned in certain countries because of concerns and regulations over the toxicity of coal-tar as a suspected carcinogen.
4) *Epoxy esters* are used as primers in less demanding applications. Their performance is inferior to epoxy–amine systems but their costs are lower.
5) *Waterborne coatings.* These are based on two-component epoxy–polyamine/polyamidoamine or epoxy–acrylic latex hybrids. One limitation of the waterborne systems is their poor cure in high humidity conditions. They have made some penetration in industrial maintenance coatings and are expected to grow more significantly in the future.

Epoxy coatings are known to have poor weatherability and often chalk when exposed to sunlight for long periods of time. Over the years, significant efforts have been dedicated to develop weatherable epoxy resin systems such as hydrogenated bisphenol A epoxy cured with siloxane-modified epoxy curing agent [225], but higher costs and compromises in curing characteristics and performance limited their commercialization. Today, most industrial structures are only coated with epoxy coatings which can last up to 10–15 years. When appearance is critical, epoxy primers are often top-coated with aliphatic isocyanate based polyurethane coatings. Marine coatings have very diverse requirements depending on the specific functions of the parts of the ship being coated. For example, ship decks are coated with antislip, abrasion- and corrosion-resistant epoxy coatings; the cargo tanks require highly chemical-resistant coatings; ship exteriors above the water line are coated with epoxy primers followed by urethane top coats for appearance; underwater ship bottoms are coated with multilayer coatings including a zinc-rich or epoxy primer, epoxy intermediate coats and antifouling top coats based on vinyls or acrylates.

Concerns over the safety of large tankers have led to regulations and construction of double-hull ships, increasing epoxy coatings consumption. While Japan was the center of the ship-building industry since the 1980s, Korea and China have emerged as major players in this market because of their low cost advantages. According to data from the Japanese Ship Building Industry Association, Korea has overtaken Japan as the global leader in ship building in the year 2000. The combined market shares of these three Asian countries now account for more than 80% of the global ship-building business. In addition, China already owns 80% of the world shipping container construction business.

The migration of ship building yards to Korea and China has led to significant increases in marine epoxy coatings consumption in that region, and has resulted in increased demands for lower temperature cure (LTC) epoxy systems. Traditional ambient-cure epoxy coatings do not cure well at temperatures below 10°C (50° F). They often require excessive cure time, affecting productivity and performance, and shorten the painting season in colder climates such as in Korea. A number of LTC epoxy coating technologies have been developed. Uses of accelerators such as tertiary amines, organic acids or alkyl-substituted phenols have allowed cures at 4.45°C (40° F) temperature range. However, shorter pot-life is a limitation of these systems. Newer epoxy coatings utilizing cycloaliphatic amines, phenalkamines as curing agents can cure at temperatures of about 0°C (30°F). For industrial protective coatings, systems developed by Ameron International (Amerlock 400) and ICI-Devoe (Bar-Rust 235) are claimed to achieve LTC down to −18°C (0°F).

Other new technology developments in this market segment include surface-tolerant epoxy coatings for aged or marginally-prepared surfaces, interval-free epoxy coatings to extend coating service life, mineral spirit-soluble epoxy coatings for shipping containers repairs, and styrene-free coatings to replace foul-smelling and regulated organic solvents such as toluene and xylene in coatings for new shipping containers.

20.2.2. Metal Container and Coil Coatings

Metal container and coil coatings represent a major outlet for epoxy resins considering there are more than 100 billion beverage cans and 30 billion food cans produced annually in the United States. Globally, the metal can market is estimated at over 300 billion cans. While the majority of metal containers coated with epoxy coatings are aluminum and steel food and beverage cans, coatings for drums, pails, and aerosol spray cans are included in this market segment. Coil coating is a highly efficient, automated coating process used to produce precoated metal coils, which are subsequently stamped and fabricated to parts. The majority of epoxy coil coatings are used to produce metal can ends and can bodies with smaller amounts going to building products, appliance panels, transportation, and metal furniture applications.

Higher molecular weight SERs (EEW = 2000–4000), which contain predominantly secondary hydroxyl groups, are used in these coatings where maximum resistance to chemicals, good flexibility, freedom from off-taste, good thermal stability, blush (hydrolysis) resistance, and the ability to hold corrosive foods and beverages are needed. In addition, compliance with food regulations such as the Food and Drugs Administration (FDA) rules is required for food and beverage interior can coatings. This application is where the unique combination of properties of epoxy resins stand out.

The can and coil coatings, generally, are cross-linked with phenol, melamine, or urea–formaldehyde condensation products at elevated temperatures (150–200°C) with acid catalysts. Normal epoxy–amino resin weight ratios are epoxy–urea, 70:30; epoxy–benzoguanamine, 70:30; epoxy–melamine, 80:20 and 90:10. Increasing cross-linker levels give improved thermal and chemical resistance at the sacrifice of coating flexibility and adhesion.

Phenol–formaldehyde resole cured epoxies have excellent chemical resistance and hardness and are the popular choices for drum coatings. Their golden color is affecting their uses in food can coatings because of the increasing popularity of the water-white coatings based on melamine–formaldehyde resins, which are perceived to be “cleaner” by the

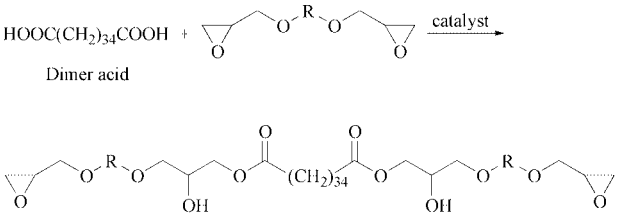

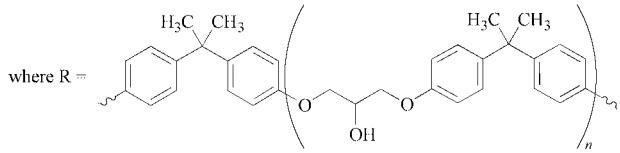

consumers and the food industry. Melamine–formaldehyde resins are the primary cross-linker for beer and beverage interior can coatings. Urea–formaldehyde resins can be cured at lower temperatures and faster speed than phenol and melamine–formaldehyde resins and are widely used in the coil coatings industry where cure schedules are extremely short. However, their use has been declining because of concerns over the release of formaldehyde fumes. Recently, there have been regulatory issues in the can industry concerning worker exposure to volatile formaldehyde emissions from the formaldehyde resins. A number of new, formaldehyde-free coating formulations have been introduced by coating suppliers [226].

High solids binders for metal can coatings have been developed on the basis of dimer acid modification of epoxy resins, whereby a flexible C_{34} difunctional acid is used to esterify a conventional diepoxide resin [227]. The resultant epoxy ester possesses a sufficiently lower viscosity to provide binders with solids contents >70 vol. %. Curing is accomplished by a melamine–formaldehyde resin (Cymel 303, from Cytec) in conjunction with phosphoric acid catalyst (see above).

In the 1970s, a waterborne coating system for aluminum beverage can coatings was developed by the Glidden Company (ICI Packaging Coatings) on the basis of a graft copolymerization of an advanced epoxy resin and acrylic monomers [228, 229]. The acrylic–vinyl monomers are grafted onto preformed epoxy resins in the presence of a free-radical initiator; grafting occurs mainly at the methylene group of the aliphatic backbone on the epoxy resin (see below).

The polymeric product is a mixture of methacrylic acid–styrene copolymer, SER, and graft copolymer of the unsaturated monomers onto the epoxy resin backbone. It is dispersible in water upon neutralization with an amine, and cured with an amino–formaldehyde resin. The technology revolutionized the can coatings industry in the 1970s which was primarily based on low solids, solventborne coatings. This waterborne epoxy coating system and its variations continue to be the dominant choices for inte-

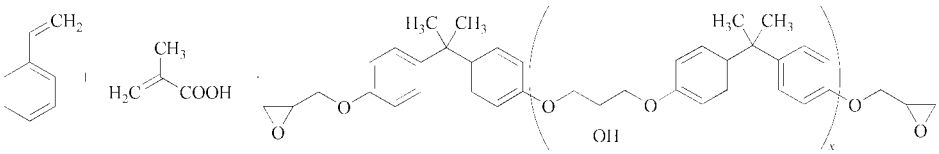

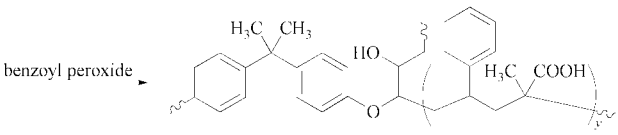

rior beer and beverage can coatings globally today. They are also formulated with phenol–formaldehyde resole and used as interior coatings in the new two-piece food can plants in the United States.

UV-curable coatings based on cycloaliphatic epoxies are used on the exterior of some beer, beverage, and food cans, as well as food and composite can ends. The technology is environmentally friendly and energy-efficient.

Coil coatings have been gaining in the appliance market. More OEMs have turned to precoated metal coils as an efficient manufacturing alternative to produce appliance panels, eliminating the needs for postformed coating processes. PVC organosol (copolymers of vinyl chloride and vinyl acetate) coatings for coil-coated can ends and bodies have been under environmental pressures and epoxy has been gaining as PVC coatings are replaced [230].

The growth of can coatings has been steady globally because of the expansion of new can plants in Asia-Pacific and South America in the 1980s and early 1990s, which made up for the stagnant growth of the U.S. market. However, growth of plastic bottles based on PET [poly(ethylene terphthalate)] has recently eroded the metal can position in beverage packaging, affecting epoxy can coatings growth. In addition, new can fabrication technologies utilizing other polymers are being developed which may challenge the dominant position of epoxy coatings in metal cans.

In the 1980s, Toyo Seikan Co. of Japan successfully developed and commercialized TULC (Toyo Ultimate Laminate Cans), a revolutionary technology in which cans are fabricated using a deep draw process from metal coils laminated with thermoplastic polyester films [231, 232]. No epoxy coating is used in this technology. Special polyester film combinations were used (in a much higher thickness than typical epoxy coatings) to facilitate the demanding deep draw process while maintaining all of the other requirements of can coatings. This technology is a significant breakthrough with claimed benefits such as no solvent emission, lower energy and water usage, and excellent quality cans. The costs however are significantly higher than those of conventional cans, and the technology has found widespread application only in Japan where higher packaging costs are acceptable. Other companies such as British Steel have been actively promoting laminated cans as a way to produce differentiable packaging like shaped cans with very limited success. Higher cost is the biggest barrier to their broad commercialization. Recent developments include attempts to fabricate can ends and bodies from extrusion-coated metals by companies such as Alcoa. Thermoplastics like modified polyesters are providing challenges to epoxies in these new technologies due to their excellent formability. However, their resistance against aggressive drinks, foods, and retort are inferior to those of epoxies.

More recently in the United States, Campbell Soup Co. has successfully launched a new line of microwaveable, ready-to-eat plastic cans. These cans are constructed from a molded thermoplastic can body (polypropylene, high density polyethylene) and an easy-open-end (EOE) of coated metal. In addition, flexible pouches have made inroad as an alternative for metal cans in certain markets such as packaged tuna fish.

Recently, there have been debates in the can coatings industry concerning the potential health effects of residual bisphenol A and DGEBA in epoxy can coatings. The resin suppliers, can coatings producers, and can makers have jointly formed an industry group to coordinate a number of studies on this issue. Results indicated that epoxy can coatings, when properly formulated and cured, are safe and in compliance with global food contact regulations. Current regulatory guidelines such as the Specific Migration Levels for Europe set extractable limits of 1 mg/kg for DGEBA and 3 mg/kg for bisphenol A. Additional information is available in the references [233, 234]. Some polyester coatings have been developed as epoxy coating alternatives, but high costs and inferior pasteurization-resistance limit their uses [235].

20.2.3. Automotive Coatings

Automotive coatings are another major application for epoxy resins. The excellent adhesion and corrosion resistance properties of epoxies make them the overwhelming choice for automotive primers. One new, growing application is the use of epoxy–polyester or acrylic–GMA powders in primer-surfacer coatings. In addition, glycidyl methacrylate (GMA) is used as a comonomer in

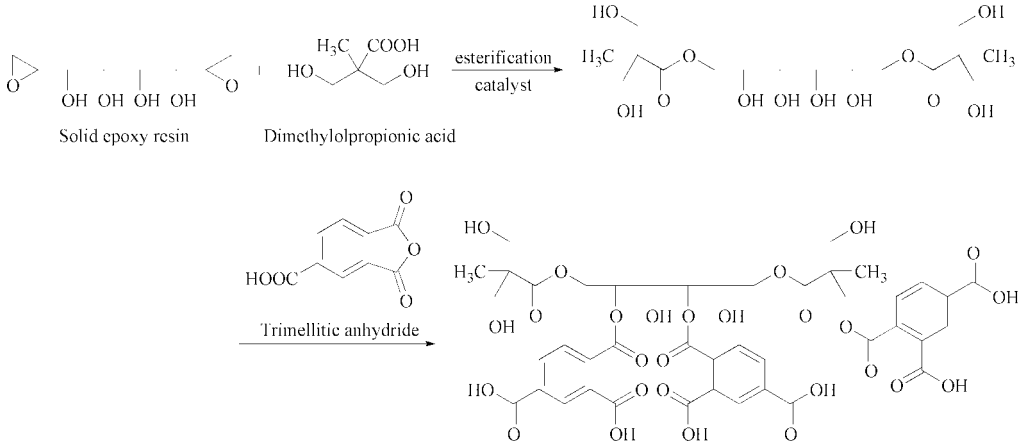

etch-resistant liquid top coats containing acrylic acid/anhydride [236] and in GMA-acrylic powder coatings for clear coats and automotive parts [220]. Epoxy powder coatings for automobiles are expected to grow significantly in the near future.

Electrodeposition processes using epoxy-based automotive primers were developed for anodic and cathodic systems. Anodic systems (AED) employ carboxylated epoxy resins neutralized with an amine. A typical binder is prepared by the esterification of the terminal epoxy groups of a solid resin (EEW = 500) with stoichiometric quantities of dimethylolpropionic acid to form a hydroxyl-rich resin. This intermediate is subsequently treated with a cyclic anhydride to form an acid functionalised polymer, which is then neutralized with the amine (see above).

Significant advances in waterborne automotive coatings have been made by PPG Industries and others utilizing epoxies as co-resins in the 1970s. These coatings are used in cathodic electrodeposited (CED) systems, which are widely accepted for automobile primers. Many patents have been issued for this important technology [214]. Cathodic systems, which have superior corrosion resistance, have replaced anodic systems. A typical epoxy binder for cathodic electrodeposition is prepared by first forming a tertiary amine adduct from an epoxy resin and a secondary amine, followed by neutralization with an acid to form a water-soluble salt:

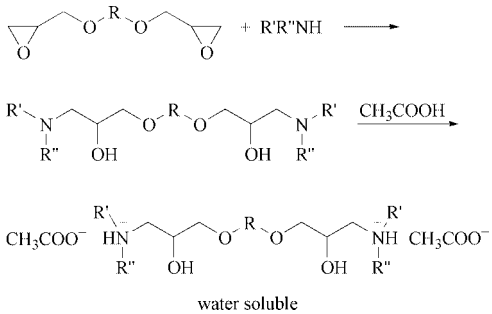

Cross-linking is achieved by reaction of the hydroxyl groups with a blocked isocyanate, which is stable at ambient temperature.

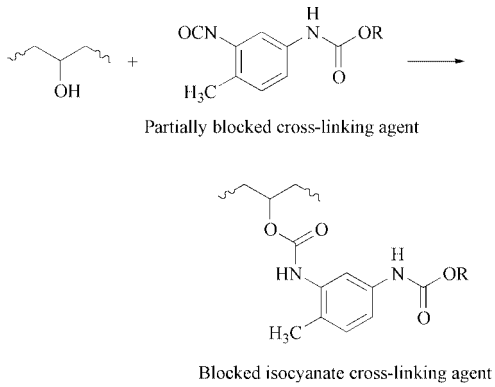

where R = 2-Ethylhexyl

The ability of the CED coating system to thoroughly coat all metal surfaces of the car and the resultant superior corrosion resistance was a significant breakthrough, enabling its dominant position in the global automotive industry.

PPG has continued to develop new generations of improved CED epoxy coatings [237]. Dupont, BASF, and a number of Japanese coating companies such as Nippon Paint and Kansai Paint have contributed to the epoxy primer coating technology by developing advanced coating systems to meet higher performance and regulatory requirements of the automotive industry [238 – 240]. The popular pigment systems based on heavy metals such as lead and chromium in primer coatings have been recently banned in certain countries, leading to efforts to develop new formulations with improved corrosion resistance. Nippon Paint has proposed pigment-free CED systems [241].

Epoxy–polyester and acrylic–GMA powder coatings have made significant advances recently in the area of primer-surfacer coatings. They offer better adhesion to topcoats and significantly improve chip resistance compared to the traditional liquid polyester and epoxy ester coatings. This translates to warranty cost reductions, leading many car manufacturers to convert to the powder coating technologies.

While epoxy coatings based on DGEBA and other aromatic epoxies are limited to undercoats and under-the-hood applications because of their poor UV resistance, GMA-based coatings have been developed for improved acid-etch performance automotive top coats. They compete with traditional acrylic polyol–melamine topcoats that are highly susceptible to acid rain-induced hydrolysis, and offer better mar resistance and less worker exposures than isocyanate-based topcoats [242, 243]. BMW has coverted to a GMA–acrylic powder clear coat developed by PPG.

20.3. Inks and Resists

Inks and resists comprise a relatively small but high value and growing market for epoxies and epoxy derivatives. In 2001, there were an estimated of 6800 MT of epoxies and epoxy derivatives used in this market to produce ink and resist formulations worth almost $400 million in the U.S. market. Epoxies are often used with other resins such as polyester acrylates and urethane acrylates in these formulations. The largest applications are lithographic and flexographic inks followed by electronic inks and resists.

Resist technology is widely used in the electronics industry to manufacture printed circuits. The resist (a coating or ink) is applied over a conducting substrate such as copper in a pattern to protect its surface during etching, plating, or soldering. Cure is either by radiation or heat. The uncured coating (or ink) is removed later by solvents. Solder masks perform similar functions in the manufacturing of printed circuit boards. The growth of the computer and electronics industries has fueled growth of epoxy-based inks and resists. The market is projected to grow at 10% annually.

The primary resins used in this market are the radiation-curable epoxy acrylates, accounting for 60% of the resins used. A small amount of cycloaliphatic epoxies are also used in UV-curable inks and resists. Phenol and cresol epoxy novolacs, and bisphenol A based epoxies are used in thermally cured formulations. The epoxy novolacs are used where higher heat resistance is needed such as in solder masks. Both free-radical and cationic-curable UV inks and colored base coats have grown rapidly because of the needs for higher line speeds, faster cleanup or line turnaround, less energy consumption, less capital for a new line, and fewer emissions.

A unique epoxy (epoxy chalcone) produced by Huntsman can be used for dual cure [244]:

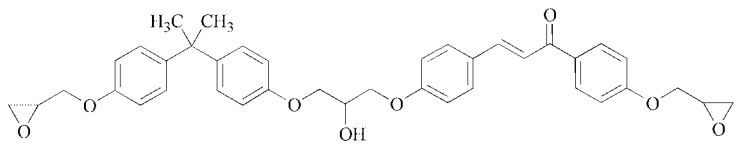

Radiation-initiated free-radical cure is possible via the double bonds, while the epoxy groups are available for thermal cure. Epoxy chalcone is used as a photopolymerizable solder mask and in photoresists.

21. Structural Applications

Next to coatings, structural applications account for the second largest share of epoxy resin consumption (~40%). Epoxy resins in structural applications can be divided into three major areas: fiber-reinforced composites and electrical laminates; casting, encapsulation, and tooling; and adhesives. Within this segment, the largest applications are electrical laminates for PCB and composites made of epoxy and epoxy vinyl ester for structural applications.

21.1. Structural Composites

Epoxy resins and epoxy vinyl ester resins are well suited as fiber-reinforcing materials because they exhibit excellent adhesion to reinforcement, cure with low shrinkage, provide good dimensional stability, and possess good mechanical, electrical, thermal, chemical, fatigue, and moisture-resistance properties. Epoxy composites are formed by aligning strong, continuous fibers in an epoxy resin-curing agent matrix. Processes currently used to fabricate epoxy composites include hand lay-up, spray-up, compression molding, vacuum bag compression molding, filament winding, resin transfer molding reaction, injection molding, and pultrusion (see → Composite Materials, Sect. 4.2).

Important fiber materials are surface-treated glass, boron, graphite (carbon), and aromatic polyaramides (e.g. DuPont's Kevlar). In most composites the reinforcement constitutes ca. 65% of the final mass. Orientation of the fibers is important in establishing the properties of the laminate. Unidirectional, bidirectional, and random orientations are possible. The characteristics of the cured resin system are extremely important since it must transmit the applied stresses to each fiber. A critical region in a composite is the resin–fiber interface. The adhesive properties of epoxy resins make them especially suited for composite applications.

The most important market for epoxy composites is for corrosion-resistant equipment where epoxy vinyl esters is the dominant material of choice. Other smaller markets are automotive, aerospace, sports/recreation, construction, and marine. Because of their higher costs, epoxy and epoxy vinyl esters composites found applications where their higher mechanical strength and chemical and corrosion resistance properties are advantageous.

21.1.1. Epoxy Composites

Composites made with glass fibers usually have a bisphenol A based epoxy resin–diamine matrix and are used in a variety of applications including automotive leaf springs and drive shafts, where mechanical strength is a key requirement. A large and important application is for filament-wound glass-reinforced pipes used in oil fields, chemical plants, water distribution, and as electrical conduits. Low viscosity liquid systems having good mechanical properties when cured are preferred. These are usually cured with liquid anhydride or aromatic–amine hardeners. Similar systems are used for filament-winding pressure bottles and rocket motor casings. Other applications that use fiber-reinforced epoxy composites include sporting equipment, such as tennis racquet frames, fishing rods, and golf clubs, as well as industrial equipment. The wind energy field is emerging as a potential high growth area for epoxy composites, particularly in Europe where a number of new wind energy farms are planned. With windmill blades increasing in lengths (up to 50 m), the strength and fatigue properties of epoxy composites provide benefits over competitive chemistries.

In the aerospace industry, particularly in military aircraft construction, the use of graphite fiber-reinforced composites has been growing because of high strength-to-weight ratios. Some newer commercial airliners now contain up to 10% by weight of composite materials. High performance polyfunctional resins, such as the tetraglycidyl derivative of methylenedianiline in combination with diaminodiphenylsulfone or nadic methyl anhydride, are used to provide good elevated temperature properties and humidity resistance. Handling characteristics are well suited to the autoclave molding technique primarily used in the manufacture of such components. The low viscosity and high T_g of cycloaliphatic epoxies has led to their use in certain aerospace applications. Newer resins such as diglycidyl ether of 9,9′-bis(4-hydroxyphenylfluorene) have been developed.

While the overall growth of composites in the aerospace industry is continuing, epoxy has been facing stiff competition from other materials and the growth rate has been relatively small (2% annually). While epoxies are still used in many exterior aircraft parts, carbon fiber composites based on bismaleimide and cyanate esters have shown better temperature and moisture resistance than epoxies in military aircaft applications. In the commercial aircraft arena, phenolic composites are now preferred for interior applications because of their lower heat release and smoke generation properties during fires. High performance thermoplastics, such as polysulfone, polyimides, and polyetherether ketone (PEEK), have also found some uses in aerospace composites.

21.1.2. Epoxy Vinyl Ester Composites

Epoxy vinyl ester composites are widely used to produce chemically resistant glass-reinforced pipes, stacks, and tanks by contact molding and filament-winding processes. Epoxy vinyl ester resins provide outstanding chemical resistance against aggressive chemicals such as aqueous acids and bases and are materials of choice for demanding applications in petrochemical plants, oil refineries, and paper mills. Epoxy vinyl ester composites are also used in demanding automotive applications such as engine and oil pan covers where high temperature performance is required. Exterior panels and truck boxes are also growth automotive applications for vinyl esters. However, in less demanding automotive applications, cheaper thermoplastics and thermosets such as unsaturated polyesters or furan resins are often used. In general, epoxy vinyl ester is considered to be a premium polyester resin with higher temperature and corrosion resistance properties at higher costs. It is used where the cheaper unsaturated polyesters cannot meet performance requirements. For the same reason, epoxy vinyl ester has not grown significantly in less demanding civil engineering applications. Other uses of epoxy vinyl ester composites include boat hulls, swimming pools, saunas, and hot tubs.

Improved versions of the high performance resin systems continue to be developed [245, 246]. Toughening of epoxies and epoxy vinyl esters has emerged as an area for investigation [247]. Lower styrene content vinyl esters have been developed to reduce worker exposure. Performance enhancements with epoxy and vinyl ester nanocomposites have been reported in the literature, but commercialization has not been yet realized.

21.1.3. Mineral-Filled Composites

Epoxy mineral-filled composites are widely used to manufacture laboratory equipment such as lab bench tops, sinks, hoods, and other laboratory accessories. The excellent chemical and thermal resistance properties of epoxy thermosets make them ideal choices for this application. Typically, liquid epoxy resins of bisphenol A are cured with anhydrides such as phthalic anhydride, which provide good exotherm management and excellent thermal performance. The systems are highly filled with fillers such as silica or sand (up to 70 wt.%). Multifunctional epoxy novolacs can be added when higher chemical and thermal performance is needed.

21.2. Civil Engineering, Flooring, and Construction

Civil engineering is another large application for epoxies, accounting for up to 13% of total global epoxy consumption. This application includes flooring, decorative aggregate, paving, and construction [248]. Key attributes of epoxies such as ease of installation, fast ambient cure, good adhesion to many substrates, excellent chemical resistance, low shrinkage, good mechanical strength, and durability make them suitable for this market. In the United States an estimated 20,000 MT of resins were used for flooring applications in 2000. The building boom in China has provided significant growth for this market during the past decade. Epoxy flooring compounds are expected to grow well as the construction industry becomes more aware of their benefits.

Epoxy resins are used for both functional and decorative purposes in monolithic flooring and in factory-produced building panel applications. Products include floor paints, self-leveling floors, trowelable floors, and pebble-

finished floors. Epoxy floorings provide wear-resistant and chemical-resistant surfaces for dairies and food processing and chemical plants where acids normally attack concrete. Epoxies are also used in flooring for walk-in freezers, coolers, kitchens, and restaurants because of good thermal properties, slip resistance, and ease of cleanup. In commercial building applications, such as offices and lobbies, terrazzo-like surfaces can be applied in thin layers. Continuous seamless epoxy floors are competitive with ceramic tiles. They are usually applied by trowel over a prepared subfloor. Semiconductive epoxy/carbon black floorings are used in electronics manufacturing plants because of their ability to dissipate electrical charges. Decorative slip-resistant coatings are available for outdoor stair treads, balconies, patios, walkways, and swimming-pool decks. Epoxy aggregates are highly filled systems, containing up to 90% of stones or minerals. They are used for decorative walls, floors, and decks.

Usually, two-component systems consisting of liquid epoxy resin, diluents, fillers (e.g. sands, stones, aggregates), pigments, thickening agents, and polyamine or polyamide curing agents are employed. Cycloaliphatic amines and their adducts are used when either better low temperature cure or adhesion to wet concrete is desired. The other components of the flooring formulation are as critical as the resin and hardener. Typical filler and pigment levels are 10% for paving, 30% for flooring, and 40% or higher for decorative aggregates. Self-leveling floors consist of resin-hardener mixtures with low filler content or unfilled compositions with high gloss. In epoxy terrazzo floors, an epoxy binder replaces the cement matrix in a marble aggregate flooring, providing impact resistance, mechanical strength, and adhesion.

Epoxy systems for roads, tunnels and bridges are effective barriers to moisture, chemicals, oils, and grease. They are used in new construction as well as in repair and maintenance applications. Typical formulations consist of liquid epoxy resins extended with coal tar and diethylenetriamine curing agent. Epoxy resins are widely used in bridge expansion joints and to repair concrete cracks in adhesive and grouting (injectable mortar) systems. Epoxy pavings are used to cover concrete bridge decks and parking structures. Formulations of epoxy resins and polysulfide polymers in conjunction with polyamine curing agents are used for bonding concrete to concrete. After cleaning the old surface, the epoxy adhesive is applied and good adhesion between the old and the new concrete is obtained.

Recent developments in the construction and civil engineering industry include the development of "intelligent concrete" with self-healing capability in Japan [249]. Some of the systems are based on epoxy resins encapsulated in concrete which when triggered by cracks open and cure to repair the concrete.

21.3. Electrical Laminates

Printed wiring boards (PWB) or printed circuit boards (PCB) are used in all types of electronic equipment. In noncritical applications such as inexpensive consumer electronics, these components are made from paper-reinforced phenolic, melamine, or polyester resins. For more critical applications such as high end consumer electronics, computers, complex telecommunication equipment, etc., higher performance materials are required and epoxy resin-based glass fiber laminates fulfill the requirements at reasonable costs. This application constitutes the single largest volume of epoxies used in structural composites. In 2000, an estimated 200,000 MT of epoxy resins were used globally to manufacture PCB laminates.

Systems are available that meet the National Electrical Manufacturers Association (NEMA) G10, G11, FR3, FR4, FR5, CEM-1, and CEM-3 specifications. Both low viscosity liquid (EEW = 180–200) and high melting solid (EEW = 450–500) epoxy resins are used in printed circuit prepreg manufacture. Currently, the most widely used boards (> 85%) are manufactured to the flame-retardant FR4 specification using epoxy thermosets. Flame retardance is achieved by advancing the liquid DGEBA epoxy resin with tetrabromobisphenol A (TBBA). This relatively low cost resin which contains about 20 wt.% bromine is the workhorse of the PCB industry. Epoxy resins based on diglycidyl ether of TBBA are also available, which allow the preparation of resins with even higher bromine content, up to 50 wt.%. Multifunctional epoxy resins such as epoxy novolacs based on phenol, bisphenol

A, and cresol novolacs or the tetraglycidyl ether of tetrakis(4-hydroxyphenyl)ethane are used as modifiers to increase the glass-transition temperature ($T_g > 150°C$), thermal decomposition temperature (T_d), and chemical resistance.

The most commonly used curing agent for PWBs is dicyandiamide (DICY) catalyzed with imidazoles such as 2-methylimidazole (2-MI), followed by phenolic novolacs and anhydrides.

The epoxy–DICY systems offer the following advantages:

1) Cost effectiveness (DICY is a low equivalent weight, multifunctional curing agent)
2) Stable formulations
3) Excellent adhesion to copper and glass
4) Good moisture and solder resistance
5) Good processability

The primary disadvantage of the standard epoxy–DICY systems is their relatively low thermal performance ($T_g < 140°C$, $T_d = 300°C$), which limits their uses in more demanding applications such as the FR-5 boards and other high density circuit boards. Specialty epoxy–DICY systems are available with T_g approaching 190°C but at higher costs. Alternatively, high temperature epoxy systems are obtained using diaminodiphenyl sulfone (DDS) as curing agent and boron trifluoride monoethylamine (BF_3/MEA) complex, benzyldimethylamine (BDMA), or various imidazoles as catalysts. However, concerns over the toxicity of DDS have led to significant decrease of its use. More recently, higher thermally resistant laminates using novolac curing agents, including bisphenol A based novolacs, have become popular in the industry. However, brittleness is a significant disadvantage of these systems.

Prepreg is commonly prepared by passing the glass cloth through a formulated resin bath followed by heat treatment in a tower to evaporate the solvent and partially cure the resin to an intermediate or B stage. Prepreg sheets are stacked with outer layers of copper foil followed by exposure to heat and high pressure in a laminating press. This structure is cured (C-staged) at high temperature (150–180°C) and pressure for 30–90 min. Attempts to develop continuous prepreg and laminating processes have only achieved limited commercialization. Laminate boards may be single-sided (circuitry printed on only one side), double-sided, or multilayered (3 to 50 layers) for high density circuitry boards. Electrical connections for mounted components are obtained via drilled holes which are plated with copper.

The 1990s witnessed the explosive growth of the personal computer, consumer electronics, and wireless telecommunication industries, resulting in significant demands for PWB based on epoxy resins. The PWB industry trends toward device miniaturization, multilayer laminates, high density circuitries, lead-free solder, and faster signal transmission speeds have resulted in increased performance requirements. For example, lead-free legislation which bans electronics containing lead in the European Union became law in 2003 with an implementation date of 2006. This legislation is expected to speed up the phase-out of lead-based solders globally, forcing the industry to use alternatives such as tin alloys which have much higher soldering temperatures, and thereby drives the need for epoxy systems with higher thermal performance.

The end-use industries' demands for PCB boards with better heat resistance [250, 251], higher glass-transition temperature (T_g), higher thermal decomposition temperatures (T_d), lower water absorption, lower coefficient of thermal expansion (CTE), and better electrical properties (dielectric constant D_k and dissipation factor D_f) have led to the development of new, high performance epoxies and cross-linker systems [252]. Toughness is also becoming an issue as electrical connection holes are drilled in the highly cross-linked, high T_g laminates.

Since reinforcing materials make up from 40 to 60 wt.% of the PCB laminates, their contributions to the laminate dielectric properties are significant. The standard reinforcing glass–cloth compositions in electrical laminates are designated E (electrical) glass. Woven E glass is most commonly used, but other reinforcing materials such as nonwoven glass mat, aramid fiber, S-2 glass, and quartz are available. In recent years, the PCB industry has been evaluating materials with better dielectric properties, but they are much more expensive than standard E glass (Table 25).

In recent years, environmental concerns over toxic smoke generation during fire and end-of-life incineration of electronic equipment containing brominated products, particularly in Eu-

Table 25. Reinforcing material comparison[a]

Reinforcing material	D_k (at 1MHz)	D_f (at 1MHz)	Relative cost
E glass	6.5	0.003	1
S-2 glass	5.3	0.002	4
D glass	3.8	0.0005	10
Quartz	3.8	0.0002	30
Aramid	3.8	0.012	10

[a] From Ref. [253].

rope and Japan, have driven development efforts on halogen-free resins. This has resulted in a number of alternative products such as phosphorous additives and phosphor-containing epoxies [254 – 256]. Some examples of these phosphorous compounds are as follows:

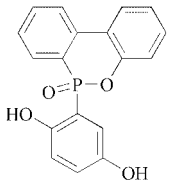

9,10-Dihydro-9-oxa-10-phosphathenanthrene-10-oxide

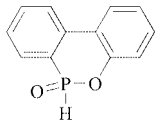

6*H*-Dibenz(*C*,*H*)(1,2)oxaphosphorin-6-oxide

However, commercialization of phosphor-containing epoxies has been limited because of higher costs and other disadvantages such as poorer moisture resistance and lower thermal performance. In addition, concerns over phosphine gas emission during fires and potential leakage of phosphorous compounds in landfills have raised questions about their long-term viability. Alternatively, the industry has been researching new epoxy resins based on nitrogen, silicon, sulfur-containing compounds, and new phenolic resins as potential halogen-free, phosphor-free replacements. Inorganic fillers such as alumina trihydrate, magnesium hydroxide, and zinc borate have also been evaluated as flame-retardant alternatives in epoxy systems.

While brominated epoxy resin remains the workhorse of the PCB industry (FR-4 boards) because of its good combination of properties and cost, it is facing competition from other thermoset and thermoplastic materials as industry performance requirements increase. Thermosets with higher temperature performance (> 180°C T_g) and lower dielectric properties include polyimides, cyanate esters, and bismaleimide–triazine (BT) resins. They are used alone or as blends with epoxies in high performance chip-packaging boards and military applications. GE's GETEK system is an interpenetrating network of polyphenylene oxide (PPO) in epoxy and has lower dielectric constant than standard epoxies. Polytetrafluoroethylene (PTFE) has a very low dielectric constant (Table 26) and is used primarily in high performance PCBs for military and high frequency (e.g. radars) applications. While these alternative materials offer certain performance advantages over standard epoxies, they are generally more expensive and more difficult to process. Thermosets such as polyimides, cyanate ester, and BT resins are very brittle and have higher water absorption than epoxies. PTFE has very poor adhesion to substrates, requiring special treatments. Consequently, they are limited to niche, high performance applications [250].

In flexible printed circuits, polyimide and polyester films are the preferred choices over epoxies. Molded interconnects based on heat-resistant thermoplastics such as polyether sulfone, polyether imide, and polyarylate have been developed to replace epoxy-based PCBs in certain applications. However, their uses are limited to special applications.

There has been a significant migration of the PCB laminate manufacturing capacity to Asia (mainly Taiwan and China) in the late 1990s. In 2001, 70% of epoxy resins used in PCB laminates was consumed in the region and the trend is expected to continue in the near future.

Table 26. Base resin systems used in PCB laminates

Resin system	T_g, °C	D_k (at 1MHz)	D_f (at 1MHz)	Estimated relative resin cost	Laminate cost
Standard Epoxies	135–140	4.6–4.8	0.015–0.020	1	1
High performance Epoxy	170–180	4.6–4.8	0.015–0.020	1.5–2	1.5
PPO/Epoxy	175–185	3.6–4.2	0.009–0.015	4–6	2–3
BT/Epoxy	170–220	3.9–4.2	0.008–0.013	8–15	2–5
Polyimide	260	3.9–4.4	0.012–0.014	5–16	3–6
Cyanate ester	230–260	3.5–3.7[a]	0.005–0.011[a]	5–16	4–8
Polyester	135–140	3.1–3.2[b]	0.004–0.014[b]	—	7–10
PTFE	NA	2.1–2.5[b]	0.0006–0.0022[b]	40	15–50

[a] Measured at 1GHz.
[b] Measured at 10 GHz.

21.4. Other Electrical and Electronic Applications

21.4.1. Casting, Potting, and Encapsulation

Since the mid-1950s, electrical-equipment manufacturers have taken advantage of the good electrical properties of epoxy and the design freedom afforded by casting techniques to produce switchgear components, transformers, insulators, high voltage cable accessories, and similar devices.

In casting, a resin-curing agent system is charged into a specially designed mold containing the electrical component to be insulated. After cure, the insulated part retains the shape of the mold. In encapsulation, a mounted electronic component such as a transistor or semiconductor in a mold is encased in an epoxy resin based system. Coil windings, laminates, lead wires, etc, are impregnated with the epoxy system. Potting is the same procedure as encapsulation except that the mold is a part of the finished unit. When a component is simply dropped into a resin-curing agent system and cured without a mold, the process is referred to as dipping. It provides little or no impregnation and is used mainly for protective coatings.

The choice of epoxy resin, curing agent, fillers, and other ancillary materials depends on factors such as cost, processing conditions, and the environment to which the insulated electrical or electronic component will be exposed.

The type and amount of filler that can be incorporated into the system are very important and depend on the viscosity of the resin at the processing temperature. Filler loading reduces costs, increases pot life, improves heat dissipation, lowers exotherms, increases thermal shock resistance, reduces shrinkage, and improves dimensional stability.

The exotherm generated during the resin cure must be controlled to prevent damage to the electrical or electronic component. The exotherm is easily controlled during the production of small castings, pottings, and encapsulations. In the production of large castings, the excess heat of reaction must be dissipated in order to prevent locked-in thermal stresses. During the 1970s, the pressure gelation casting process was developed [257]; this method provides better temperature control and reduces cycle times. The heat generated by polymerization is used to heat the resin mass and is not dissipated in the mold.

Both DGEBA and cycloaliphatic epoxy resins are used in casting systems. Most systems are based on DGEBA resins cured with anhydride hardeners and contain 60–65 wt.% inert fillers. The cycloaliphatic resin systems exhibit good tracking properties and better UV resistance than DGEBA resins, the latter of which causes crazing and surface breakdown. An electrical current is more likely to form a carbonized track in aromatic-based resins than in nonaromatic ones. Their lower viscosity also facilitates device impregnation. The cycloaliphatic epoxies are often used as modifiers for DGEBA resin systems. This application represents a significant outlet for cycloaliphatic epoxies.

Amine curing agents are used in small castings, and anhydrides are used in large castings. Anhydrides are less reactive and have lower exotherms than amines. In addition, their vis-

cosity and shrinkage are low and pot lives are longer.

21.4.2. Transfer Molding

Epoxy molding compounds (EMC) are solid mixtures of epoxy resin, curing agent(s) and catalyst, mold-release compounds, fillers, and other additives. These systems can be formulated by dry mixing or by melt mixing and are relatively stable when stored below room temperature. Molding compounds become fluid at relatively low temperatures (150–200°C) and can be molded at relatively low pressures (3.5–7.0 MPa) by compression, transfer, or injection molding. Advantages of molding over casting are elimination of the mixing step immediately before use, improved handling and measuring procedures, and suitability for high production quantities. A typical standard EMC formulation contains approximately 30% epoxies, 60% filler, and 10% of curing agents and other additives such as release agent.

An important application of epoxy molding compounds is the encapsulation of electronic components such as semiconductor chips, passive devices, and integrated circuits by transfer molding. Transfer molding is a highly automated, efficient method of encapsulation. High purity phenol and cresol epoxy novolacs and phenol and cresol novolacs and/or anhydride curing agents are used most often in semiconductor applications. For passive device encapsulation, standard epoxy novolacs can be used as blends with bisphenol A based solid resins. The ECN or EPN molding powders can be processed at relatively low pressures and provide insulation for the electronic components. Ionic impurities, i.e., NaCl or KCl, must be kept to a minimum, since trace quantities can cause corrosion and device failure. In addition, residual stress and thermal and mechanical shock resistances are issues that must be managed properly [258].

Efforts have been made to improve the high temperature performance of these systems by replacing the epoxy novolacs with other multifunctional epoxy resins. Hydrocarbon based epoxy novolacs (HEN) were developed to improve the moisture resistance of molding compounds. Crystalline epoxy resins derived from biphenol and dihydroxy naphthalenes were developed for high end semiconductor encapsulants using Surface Mount Technology (SMT). The emergence of SMT as a key semiconductor manufacturing technology requires epoxy molding compounds with a high filler loading capacity (up to 90 wt.%) to enhance solder crack resistance. SMT uses new solder alloys to attach components to the PCB board at high temperatures (215–260°C). Solder reflow, delamination, and package cracks are problems often encountered with conventional molding compounds based on cresol epoxy novolacs. The high filler content helps lower costs, reduces moisture absorption, and decreases the thermal expansion coefficient of the system. Crystalline products with very low melt viscosity such as biphenyl epoxies facilitate the processing of the high silica filler formulations while maintaining other critical requirements: moisture resistance and electrical, thermal, and mechanical properties [61]. The majority of high purity epoxies used in epoxy molding compounds (EMC) for semiconductor encapsulations are supplied by Japanese producers and a few Asian companies.

21.5. Adhesives

Epoxy-based adhesives provide powerful bonds between similar and dissimilar materials such as metals, glass, ceramics, wood, cloth, and many types of plastics. In addition, epoxies offer low shrinkage, low creep, high performance over a wide range of usage temperatures, and nobyproducts (such as water) release during cure. The epoxy adhesives were originally developed for use in metal bonding in the aircraft industry [259, 260]. In aircraft wing assemblies, high strength epoxy adhesives are used in place of metal fasteners to avoid corrosion problems inherent with metal fasteners, to reduce weight, and to eliminate "point" distribution by spreading the load over a large area. Today, epoxy is the most versatile engineering/structural adhesive, widely used in many industries including aerospace, electrical/electronic, automotive, construction, transportation, dental, and consumer. The market is of high value, consuming 25,000 MT of epoxies in North America in 2001 worth almost $500 million.

The broad range of epoxy resins and curing agents on the market allows a wide selection of system components to satisfy a particular application. Although the majority of epoxy adhesives are two-pack systems, heat activated one-pack adhesives are also available. Low molecular weight DGEBA liquid resins are the most commonly used. Higher molecular weight (EEW = 250–500) DGEBA epoxy resins improve adhesive strength because of the increased number of hydroxyl groups in the resin backbone. For applications requiring high temperature or improved chemical performance, the multifunctional epoxy phenol novolac and triglycidyl-*p*-aminophenol resins are employed. More recent products include vinyl epoxies. Adhesive systems modified with reactive diluents facilitate wetting of the substrate, allowing more filler to be added and modifying handling characteristics; however, adhesive strength is reduced. Toughened epoxy adhesives are available.

Polyamines or polyamides are the curing agents for ambient or slightly elevated temperature cures, and aromatic polyamines or anhydride hardeners are used for hot cures. These systems provide exceptional bonding strength but slower cure time. Boron trifluoride amine complexes and dicyandiamide are used in one-component adhesives. Polythiols (polysulfides, polymercaptans) are the fast-curing hardeners in "5-min." consumer epoxy formulations. The lap-shear strengths of a DGEBA epoxy cured with different hardeners are given in Table 27.

Table 27. Epoxy adhesive lap-shear strengths

Hardener	Lap-shear strength,[a] MPa[b]
Aliphatic polyamine	19
Polythiol cohardener	18
Aromatic diamine	24

[a] Adhesive strength.
[b] To convert MPa to psi, multiply by 145.

Cationically cured UV laminating adhesives based on cycloaliphatic epoxies are emerging as an alternative to solvent-based adhesives. The "dark cure" of cationics allows UV exposure and post lamination in line. This process does not require UV exposure "through" the plastic barrier material.

Epoxy adhesives are expected to grow at GDP (3–4%) over the next decade. Increased usage in the automotive and recreational markets, and replacement of mechanical fasteners help offset the slowdown in the aerospace industry (see also Adhesive Compositions).

21.6. Tooling

Tools made with epoxy are used for producing prototypes, master models, molds and other parts for aerospace, automotive, foundry, boat building, and various industrial molded items [261]. Epoxy tools are less expensive than metal ones and can be modified quickly and cheaply. Epoxy resins are preferred over unsaturated polyesters and other free-radical cured resins because of lower shrinkage, greater interlaminar bond strength and superior dimensional stability.

Most epoxy-based tooling formulations are based on liquid DGEBA resins. Aliphatic polyamines, amidoamines, or modified cycloaliphatic amines are used for ambient temperature cure, and modified aromatic diamines and anhydrides are used for high temperature cure. When high heat resistance is required ($> 350°F$) epoxy novolac resins can be employed. Reactive diluents such as aliphatic glycidyl ethers are often employed to permit higher filler load or to reduce the system viscosity for proper application. Fillers, reinforcing fibers, toughening agents, thixotropic agents, and other additives are often used depending on the desired application and final properties.

Tooling production uses four major processing methods: lamination, surface cast, splining, and casting. Lamination is made by alternating layers of glass cloth or fabric and formulated resin, usually on a framework of metal or plastic. Surface cast utilizes a filled resin compound that is applied onto the surface of a mold, which is later filled with a core material that adheres to the casting compound. Splining employs heavily filled formulations that are directly applied to a surface and manually molded or leveled to the desired shape, before or after curing, with the help of proper tools. Lastly, casting compounds are filled formulations that are directly poured or compressed into a mold coated with a release agent.

22. Health and Safety Factors

There have been many investigations of the toxicity of various classes of epoxy-containing materials (glycidyloxy compounds). The use and interpretation of the vast amount of data available has been obscured by two factors: (*1*) proper identification of the epoxy systems in question and (*2*) lack of meaningful classification of the epoxy materials. In general, the toxicity of many of the glycidyloxy derivatives is low, but the diversity of compounds found within this group does not permit broad generalizations for the class. Information on toxicity and safe handling of epoxy compounds are summarized in References [262, 263] and [264].

Bisphenol A based epoxies are the most commonly used resins. Although unmodified bisphenol A epoxy resins have a very low order of acute toxicity, they should be handled carefully and personal contact should be avoided. Prolonged or repeated skin contact with liquid epoxy resins may lead to skin irritation or sensitization. Susceptibility to skin irritation and sensitization varies from person to person. Skin sensitization decreases with an increase in MW, but the presence of low MW fractions in the advanced resins may present a hazard to skin sensitization. Inhalation toxicity does not present a hazard because of low vapor pressure. DGEBA-based resins have been reported to cause minimal eye irritation. Toxicological studies support the conclusion that bisphenol A based epoxy resins do not present a carcinogenic or mutagenic hazard. Because of the solvents used, solution of epoxy resins are more hazardous to handle than solid resins alone. Depending on the solvents used, such solutions may cause irritation to the skin and eyes, are more likely to cause sensitization responses, and are hazardous if inhaled.

Acute oral studies indicate low toxicity for these resins. Eye studies indicate only minor irritation in animals. The EPN resins have shown weak skin sensitizing potential in humans.

particularly the aromatic monoepoxides such as phenyl glycidyl ether (PGE) are known to have high toxicity and should be handled with care. They are capable of causing skin and eye irritation and sensitization responses in people. They may also present a significant hazard from inhalation.

In general, amine curing agents are much more hazardous to handle than the epoxy resins, particularly at elevated temperatures. Aliphatic amines and anhydrides are capable of serious skin or eye irritation, sensitization, and even burns. Other curing agents possess consideration variation in the degree of health hazards because of the variety of their chemical structures and it is impossible to generalize.

All suppliers provide material safety data sheets (MSDS), which contain the most recent toxicity data. These are the best sources of information and should be consulted before handling the materials.

23. Acknowledgments

The authors would like to acknowledge the contributions of Robert F. Eaton of the Dow Chemical Co. in Bound Brook, N.J., who contributed to the sections on cycloaliphatic epoxies and epoxidized vegetable oils and the cationic curing mechanism. We also would like to thank Timothy Takas of Reichhold who kindly reviewed the article. We are indebted to many colleagues in the Epoxy Products and Intermediates business at Dow Chemical for their assistance in many ways to make this article possible.

24. References

1. B. Ellis (ed.): *Chemistry and Technology of Epoxy Resins*, 1st ed., Blackie Academic & Professional, Glasgow, U.K. 1993.
2. H. Lee, K. Neville, *Handbook of Epoxy Resins*, McGraw-Hill, Inc., New York 1967. Reprinted 1982.
3. C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988.
4. B. Sedlacek, J. Kahovec (eds.): *Crosslinked Epoxies*, Walter de Gruyter, Berlin, 1987.
5. K. Dusek (ed.): *Epoxy Resins and Composites I–IV*, Advances in Polymer Science 72, 75, 78, 80, Springer-Verlag, Berlin 1986.

6. *Epoxy Resins*, Advances in Chemistry Series 92, American Chemical Society, Washington, D.C. 1970.
7. *Epoxy Resin Chemistry*, ACS Symposium Series 114, American Chemical Society, Washington, D.C. 1979.
8. *Epoxy Resin Chemistry II*, ACS Symposium Series 221, American Chemical Society, Washington, D.C. 1983.
9. *The Epoxy Resin Formulators Training Manual*, the Society of Plastics Industry, Inc., New York 1984.
10. E. O. C. Greiner, F. Dubois, M. Yoneyama, Epoxy Resins, Chemical Economics Handbook (CEH) Marketing Research Report, Stanford Research Institute (SRI) International, Menlo Park, CA 2001.
11. J. W. Muskopf, S. B. McCollister, Ullman's Encyclopedia of Industrial Chemistry, 5th ed., Vol. A9, 1987, pp. 547–563.
12. S. J. Hartman, *The Epoxy Resin Formulators Training Manual*, The Society of the Plastics Industry, Inc., New York 1984, p. 1.
13. H. Lee, K. Neville, *Handbook of Epoxy Resins*, McGraw-Hill, Inc., New York 1967, reprinted 1982.
14. I. G. Farbenindustrie, DE 676 117, 1938, US 2 136 928, 1938 (P. Schlack).
15. Ciba-Geigy Corp., US 2 324 483, 1943 (P. Castan).
16. DeVoe & Raynolds, US 2 456 408, 1948 (S. O. Greenlee).
17. D. Swern, *Chem. Rev.* **45** (1949) 1.
18. E. O. C. Greiner, F. Dubois, M. Yoneyama, *Epoxy Resins, Chemical Economics Handbook (CEH) Marketing Research Report*, Stanford Research Institute (SRI) International, Menlo Park, CA 2001.
19. Dow Chemical data, the Dow Chemical Co., Midland, Mich.
20. W. L. Faith, D. B. Keyes, R. L. Clark, *Industrial Chemistry*, 3rd ed., John Wiley & Sons, Inc., New York 1965, p. 404.
21. H. H. Szmant, *Organic Building Blocks of the Chemical Industry*, John Wiley & Sons, Inc., New York 1989, p. 281.
22. P. H. Williams, *Encyclopedia of Chemical Technology*, 1st ed., Vol. 3, John Wiley & Sons, Inc., New York 1965, pp. 857, 865.
23. W. F. Richey, *Encyclopedia of Chemical Technology*, 4th ed., Vol. 6, John Wiley & Sons, Inc., New York 1993, pp. 140, 155.
24. Shell Oil Co., US 2 714 123, 1955 (G. F. Johnson);
Solvay & CIE, BE 517 463, 1959.
25. Showa Denko K. K., JP 88 290 835, 1988 (N. Nagato, H. Mori, R. Ishioka);
Showa Denko K. K., US 4 634 784, 1987 (N. Nagato, H. Mori, K. Maki, R. Ishioka).
26. Bisphenol A, Chem Systems Report, Sept. 2002.
27. General Electric Co., US 4 400 555, 1983 (Mendiretta);
General Electric Co., WO 00/35847, 2000 (G. M. Kissinger, R. Sato).
28. W. G. Potter, *Epoxide Resins*, Springer-Verlag, New York 1970.
29. Shell Oil Co., US 2 575 558, 1951 (H. A. Newley, E. C. Shokal).
30. Union Carbide Corp., US 2 943 095, 1960 N. H. Reinking).
31. Shell Development, US 2 848 435, 1958 L. H. Griffin, J. H. Long);
Reichhold Chemicals, Inc., US 2 921 049, 1960 (H. L. Moroson);
Union Carbide Corp., US 3 069 434, 1962 (S. P. Spence, A. R. Grover, F. P. Klosek, R. E. Nicolson).
32. Dow Chemical Co., US 4 499 255, 1985 (C. S. Wang, H. Q. Pham, J. L. Bertram);
Tohto Kasei K. K., JP 61/195111A, 1986 (T. Ogata, H. Nakanishi, M. Aritomi).
33. N. S. Enikolopyan, M. A. Markevitch, L. S. Sakhonenko, S. Z. Rogovina, V. G. Oshmyan, *J. Polym. Sci., Chem. Ed.* **20** (1982) 1231–1245.
34. D. O. Bowen, R. C. Whiteside, in R. F. Gould (ed.): *Epoxy Resins*, Advances in Chemistry Series 92, American Chemical Society, Washington, D.C. 1970, p. 48.
35. Shell Oil Co., US 2 643 239, 1953 (E. C. Shokal, H. A. Newley, T. E. Bradley);
Shell Development Co.), US 2 879 259, 1959 (R. W. H. Tess).
36. Devoe & Raynolds Co., Inc., US 2 456 408, 1943 (S. O. Greenlee).
37. H. Batzer, S. A. Zahir, *J. Appl. Polym. Sci.* **19** (1975) 601.
38. W. Burchard, S. Bantle, S. A. Zahir, *Makromol. Chem.* **182** (1981) 145.
39. U. Fuchslueger, H. Stephan, H.-J. Grether, M. Grasserbauer, *Polymer* **40** (1999) 661–673.
40. Ciba-Geigy Corp., US 3 634 323, 1972 (R. M. Moran).
41. Shell Oil Co., US 3 477 990, 1969 (M. F. Dante, H. L. Parry).
42. Dow Chemical Co., US 3 948 855, 1976 (W. O. Perry).
43. Dow Chemical Co., US 4 302 574, 1981 (G. A. Doorakian, J. L. Bertram).

44. Shell Oil Co., US 4 358 578, 1982 (T. F. Brownscombe).
45. Dow Chemical Co., US 4 366 295, 1982 (M. C. Tyler Jr., A. L. McCrary).
46. Dow Chemical Co.), US 4 808 692, 1989 (H. Q. Pham, L. A. Ho).
47. W. A. Romanchick, J. F. Geibel, *Org. Coat. Appl. Polym. Sci. Proc.* **46** (1982) 410.
48. J. F. Geibel, *Org. Coat. Plast. Chem.* **43** (1980) 545.
49. H. Batzer, S. A. Zahir, *J. Appl. Polym. Sci.* **21** (1977) 1843.
50. D. P. Sheih, D. E. Benton, ASTM Special Technical Publication STP 1119, Analysis of Paints and Related Materials: Current Techniques for Solving Coatings Problems, 1992, pp. 41–56.
51. Shell International Research, WO 01/16204 A1, 2001 (S. MK. Li);
Resolution Research Nederland B. V., WO Appl. 01/46287 A1, 2001 (R. R. Dominquez, H. Frank, S. MK. Li).
52. Dow Chemical Co., US 4 612 156, 1986 (B. W. Heinemeyer, S. D. Tatum).
53. J. E. White, H. C. Silvis, M. S. Winkler, T. W. Glass, D. E. Kirkpatrick, *Adv. Mater.* **12** (2000) 1791, 1800.
54. Shell Oil Co., GB 980 509, 1965.
55. Dow Chemical Co., US 5 275 853, 1994 (H. C. Silvis, J. E. White).
56. H. C. Silvis, C. N. Brown, S. L. Kram, J. E. White, *Polym. Prepr.* **36** (1995) no. 2, 178–179.
57. T. Glass, H. Pham, M. Winkler, in Proceedings of the 58th SPE Annual Technical Conference and Exhibits (ANTEC 2000), Lauderdale, FL 2000.
58. Dow Chemical Co., US 5 134 201, 1992 (M. N. Mang, J. E. White).
59. Nan Ya Plastics Corp. Taiwan, US 6 512 075, 2003 (M. J. Tzou).
60. Tohto Kasei K.K., JP 08198949 A, 1996 (K. Ishihara, T. Sato, K. Aida, T. Hoshono).
61. J. R. Griffith, *CHEMTECH* **12** (1982) 290–293.
62. the United States of America, US 3 879 430, 1975 (J. R. Griffith, J. G. O'Rear).
63. T. E. Twardowski, P. H. Geil, *J. Appl. Polym. Sci.* **42** (1991) 69.
64. Devoe & Raynolds, US 2 521 912, 1950 (S. O. Greenlee).
65. Dow Chemical Co., US 4 785 061, 1988 (C. S. Wang, Z. K. Liao);
Yuka Shell Epoxy Co., JP 200239346, 2000 (Y. Murata, B. Shigeki).
66. Dow Chemical Co., US 4 394 497, 1983 (D. L. Nelson, B. A. Naderhoff).
67. Shell Development Co., US 2 806 016, 1957 (C. G. Schwartzer).
68. K. L. Hawthorne, F. C. Henson, in R. S. Bauer (ed.): *Epoxy Resin Chemistry II*, ACS Symposium Series 221, American Chemical Society, Washington, D.C. 1983, Chapt. 7, pp. 135–151.
69. Union Carbide Corp., US 2 951 825, 1960), (N. H. Reinking, B. P. Barth, F. J. Castner).
70. T. J. Galvin, M. A. Chaudhari, J. J. King, *Chem. Eng. Prog.* **81** (1985) no. 1, 45–48.
71. Shell Internationale Research B. V., JP 3 315 436, 2002 (Y. Murata, Y. Nakanishi, M. Yosumura).
72. I. Ogura, T. Imada, Dai Nippon Inks & Chemical (DIC) Technical Review No. 5, 1999.
73. Dow Chemical Co., US 5 463 091, 1995 (J. D. Earls, R. E. Hefner Jr., P. M. Puckett);
Dow Chemical Co., US 5 266 660, 1993 (R. E. Hefner Jr., J. D. Earls, P. M. Puckett).
74. H.-J. Sue, J. D. Earls, R. E. Hefner Jr., M. I. Villarreal, E. I. Garcia-Meitin, P. C. Yang, C. H. Cheatham, C. J. G. Plummer, *Polymer* **39** (1998) 4707–4714.
75. R. S. Bauer, in G. D. Parfitt, A. V. Patsis (eds.): *Organic Coatings, Science and Technology*, Vol. 5, Marcel Dekker, New York 1983, pp. 1–33.
76. Henkel & Cie. GmbH, US 3 288 789, 1966 (M. Budnowski, M. Dohr).
77. Internationale Octrool Maatschappij Octropa, GB 1 381 262, 1975 and
US 4 147 737, 1979 (A. J. Sein, J. Reitberg, J. M. Schouten).
78. E. H. Catsiff, R. E. Coulehan, J. F. Diprima, D. A. Gordan, R. Seltzer, in R. S. Bauer (ed.): *Epoxy Resin Chemistry*, ACS Symposium Series 114, American Chemical Society, Washington, D.C. 1979, Chapt. 10, pp. 115–156.
79. R. Y. Ting, in C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988, p. 551, 601.
80. H. Jahn, *J. Polym. Sci., Part C* **16** (1967) 1829, 1841.
81. Dow Chemical Co., US 4 273 921, 1980 (J. L. Bertram, P. S. Sheih).
82. Union Carbide Corp., US 2 716 123, 1953), (B. Phillips, F. Frostick Jr.).
83. P. K. T. Ording: *Waterborne and Solvent Based Epoxies and Their End User*

Applications, John Wiley & Sons, Inc., New York 1996, pp. 57, 100.
84. J. C. Kenny, T. Ueno, K. Tsutsui, *J. Coat. Technol.* **68**, (1996) 855;
T. Agawa, E. D. Dumain, *Proc. Waterborne High-Solids Powder Coat. Symp.* **24** (1997) 342–353;
B. V. Gregorovich, I. Hazan, *Prog. Org. Coat.* **24** (1994) 131.
85. G. Webster: *Chemistry & Technology of UV & EB Formulation for Coatings, Inks & Paints*, Vol. 2, John Wiley & Sons, Inc., New York 1997, pp. 41, 73.
86. Dow Chemical Co., US 3 367 992, 1968 (C. Bearden);
Dow Chemical Co., US 3 524 901, 1970 (D. J. Najvar).
87. T. F. Anderson, V. B. Messick, in G. Pritchard (ed.): *Developments in Reinforced Platics 1*, Allied Science Publishers Ltd., London 1980, pp. 29, 58.
88. Dow Chemical Co., US 4 397 970, 1983 (K. D. Campbell, H. G. Langer, P. H. Martin).
89. H. Jahn, P. Goetzky, in C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988, pp. 1049, 1087.
90. Annu. Book ASTM Stand. Section 8 (Plastics). Web site: http://www.astm.org.
91. H. Pasch, R. Unvericht, M. Resch, *Angew. Makromol. Chem.* **212** (1993) 191–200;
H. Pasch, J. Adrian, D. Braun, *GIT Spezial Separation* **21** (2001) no. 2, 104–108.
92. D. Crozier, G. Morse, Y. Tajima, *SAMPE J.* **18** (1982) no. 5, 17–22.
93. A. Durbetaki, *Anal. Chem.* **28** (1956) 2000.
94. R. Jay, *Anal. Chem.* **36** (1964) 667.
95. B. Dobinson, W. Hofmann, B. Stark, *The Determination of Epoxide Groups*, Pergamon Press, Elmsford, NY 1969.
96. G. H. Schneer, W. van Gilder, V. E. Hauser, P. E. Schmidt, IEEE Trans. Electron. Devices (1969) ED–15.
97. H. Lee, K. Neville, *Handbook of Epoxy Resins*, McGraw-Hill, Inc., New York 1967, reprinted 1982.
98. M. C. Paputa Peck, R. O. Carter III, S. B. A. Qaderi, *J. Appl. Polym. Sci.* **33** (1987) no. 1 77–86;
G. Lachenal, Y. Ozaki, *Macromol. Symp.* **141** (1999) 283–292.
13th European Symposium on Polymer Spectroscopy, 1998.
99. W. B. Moniz, C. F. Poranski Jr., *Org. Coat. Plast. Chem.* **39**, (1978) 99–102.
100. E. Mertzel, J. L. Koenig, in K. Dusek (ed.): *Epoxy Resins and Composites III, Advances in Polymer Science 75*, Springer-Verlag, Berlin 1986, pp. 73, 112.
101. W. A. Dark, E. C. Conrad, L. W. Crossman Jr., *J. Chromatogr.* **91** (1974) 247–60.
102. D. J. Crabtree, D. B. Hewitt, *Liq. Chromatogr. Polym. Relat. Mater.* **8** (1977) 63–77, Chromatographic Science Series.
103. G. Eppert, G. Liebscher, *J. Chromatogr.* **238** (1982) 399.
104. S. A. Zahir, S. Bantle, in R. S. Bauer (ed.): *Epoxy Resin Chemistry II*, ACS Symposium Series 221. American Chemical Society, Washington, D.C. 1983, pp. 245, 262.
105. D. K. Hadad, in C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988, pp. 1089, 1172.
106. W. Lwowski, in A. R. Katritsky, C. W. Rees (eds.): *Comprehensive Heterocyclic Chemistry*, Vol. 7, Pergamon Press, Oxford 1984, pp. 1, 16.
107. R. E. Parker, N. S. Isaacs, *Chem. Rev.* **59** (1959) 737, 799.
108. Y. Tanaka, R. S. Bauer, in C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988, pp. 465, 550.
109. L. Shechter, J. Wynstra, *Ind. Eng. Chem.* **48** (1956) 86.
110. I. T. Smith, *Polymer* **2** (1961) 95.
111. N. B. Chapman, R. E. Parker, N. S. Issacs, *J. Chem. Soc.* **2** (1959) 1925.
112. N. G. Rondan, M. J. Marks, S. Hoyles, H. Pham, paper presented at the 225th ACS National Meeting, New Orleans, LA, March 2003.
113. R. F. Eaton, Paint Coat. Ind. (1999) June, 76–80.
114. L. Shechter, J. Wynstra, R. P. Kurkjy, *Ind. Eng. Chem.* **48** (1956)).
115. K. Horie, H. Hiura, M. Sawada, I. Mita, H. Kambe, *J. Polym. Sci., A–1* **8** (1970) 1357.
116. K. Dusek, M. Ilavsky, S. Lunak, *J. Polym. Sci., Polym. Symp.* **53** (1975) 29.
117. C. V. Hare, *Protective Coatings: Fundamentals of Chemistry and Composition*, Technology Publishing Co., Pittsburg, PA 1994, pp. 187, 238.
118. CIBA (A.R.L) Ltd., GB 886 767, 1962 (P. Halewood).
119. W. R. Ashcroft, in B. Ellis (ed.): *Chemistry and Technology of Epoxy Resins*, 1st ed., Blackie Academic & Professional, Glasgow, U.K. 1993, pp. 37, 71.

120. M. Fedtke, F. Domaratius, A. Pfitzmann, *Polym. Bull.* **23** (1990) 381, 388.
121. Minnesota Mining & Manufacturing Co., US 3 390 124, 1968 (J. B. Kittridge, A. L. Michelli).
122. M. D. Gilbert, N. S. Schneider, W. J. McKnight, *Macromolecules* **24** (1991) 360.
123. S. A. Zahir, in G. D. Parfitt, A. V. Patsis (eds.): *Advances in Organic Coatings Science and Technology, Vol. IV: Sixth International Conference in Organic Coatings Science and Technology*, Technomic Publishing Co., Inc., Lancaster, PA 1982, p. 83.
124. M. J. Husband, in P. Oldring, G. Haywood, eds., *Resins for Surface Coatings, SITA Technology*, John Wiley & Sons Ltd., London 1987, pp. 63, 167.
125. P. J. Madec, E. Marechal, *Makromol. Chem.* **184** (1983) 323.
126. L. Matejka, J. Lovy, S. Pokorny, K. Bouchal, K. Dusek, *J. Polym. Sci., Polym. Chem. Ed.* **21** (1983) 2873.
127. H. Kunitomo, in Proceedings of the 26th Technical Conference of the Japan Society of Epoxy Resin Technology, Tokyo, July 2002, pp. 19, 33.
128. A. Gardziella, L. A. Pilato, A. Knop, *Phenolic Resins Chemistry, Applications, Standardization, Safety and Ecology*, 2nd ed., Springer, Berlin 1999.
129. A. J. Kirsch, *50 Years of Amino Coatings*, American Cyanamid Co., Wayne, NJ 1986.
130. T. M. Rees, *J. Oil Color Chem. Assoc.* **71** (1988) no. 2, 39, 41.
131. Jefferson Chemical, US 3 020 262, 1962 (G. P. Speranza);
Dow Chemical Co., US 4 658 007, 1987 (M. J. Marks, R. A. Plepys).
132. M. J. Marks, *Polym. Mater. Sci. Eng. (Am. Chem. Soc., Div. Polym. Sci. Eng.)* **58** (1988) 864.
133. D. A. Shimp, F. A. Hudock, S. J. Ising, paper presented at 33rd SAMPE, Anaheim, CA, March 7–10, 1988.
134. B. A. Rozenberg, in K. Dusek (ed.): *Epoxy Resins and Composites II*, Advances in Polymer Science 75, Springer-Verlag, Berlin 1986, pp. 146, 156.
135. F. Ricciardi, W. A. Romanchick, M. M. Joullie, *J. Polym. Sci., Polym, Chem. Ed.* **21** (1983) 1475.
136. M. S. Heise, G. C. Martin, *Macromolecules* **22** (1989) 99.
137. J. D. B. Smith, *Org. Coat. Plast. Chem.* **39** (1978) 42–46.
138. R. J. Arnold, *Mod. Plast.* **41** (1964) no. 4, 149.
139. J. J. Harris, S. C. Temin, *J. Appl. Polym. Sci.* **10** (1966) 523.
140. J. V. Crivello, J. H. W. Lam, *Macromolecules* **10** (1977) 1307;
ACS Symp. Ser. **114**, 1 (1979).
141. R. F. Eaton, K. T. Lamb, paper presented at the 23th International Waterborne, Higher Solids and Powder Coatings Symposium, New Orleans, LA, Feb. 1996.
142. J. M. Land, A. Aubuchon, C. Pundmann, W. L. Dechent, J. O. Stoffer, Book of Abstracts, 211th ACS National Meeting, New Orleans, LA, March 24–28, 1996, *PMSE-181*, American Chemical Society, Washington, D.C. 1996.
143. D. W. Brooker, G. R. Edwards, A. McIntosh, *J. Oil Color Chem. Assoc.* **52** (1969) 989–1034; G. K. Noren, *J. Coat. Technol.* **72** (2000) no. 905, 53–59.
144. J. E. Mark, B. Erman, *Rubber Elasticity: A Molecular Primer*, John Wiley & Sons, New York 1988.
145. T. I. Smith, *J. Polym. Sci., Polym. Symp.* **46** (1974) 97.
146. F. Lohse, R. Schmid, paper presented at the Fifth International Conference in Organic Coatings Science and Technology, Athens, FATIPEC, Liége, Belgium 1979.
147. I. Ogura, Dainippon Ink & Chemicals (DIC) Technical Review No. 7, 2001. Paper written in Japanese.
148. T. Kamon, H. Furukawa, in K. Dusek (ed.): *Epoxy Resins and Composites IV*, Advances in Polymer Science 80, Springer-Verlag, Berlin 1986, pp. 173, 202.
149. E. F. Oleinik, in Ref. 137, pp. 50, 99.
150. W. R. Ascroft, *Eur. Coat. J.* **4** (1991) 229, 241.
151. R. S. Bauer, *CHEMTECH* **10** (1980) Nov., 692.
152. C. Rooney, *Mod. Paint Coat.* **81** (1991) no. 5, 44–50.
153. J. K. Gillham, C. A. Grandt, in S. S. Labana (ed.): *Chemistry and Properties of Cross-Linked Polymers*, Academic Press, Inc., Orlando, FL 1977, pp. 491, 520.
154. J. D. Keenan, *J. Appl. Polym. Sci.* **24** (1979) 2375, 2387.
155. C. A. May (ed.): *Chemorheology of Thermosetting Polymers*, ACS Symposium Series 227, American Chemical Society, Washington, D.C. 1983.
156. D. Adolf, J. E. Martin, *Macromolecules* **23** (1990) 3700, 3704.

157. *General Guide to Formulating with Dow Epoxy Resins*, the Dow Chemical Co., Midland, MI 1983.
158. B. Ellis, in B. Ellis (ed.): *Chemistry and Technology of Epoxy Resins*, 1st ed., Blackie Academic & Professional, Glasgow, U.K. 1993, pp. 72, 116.
159. J. P. Pascault, H. Sautereau, J. Verdu, R. J. J. Williams, *Thermosetting Polymers*, Marcel Dekker, Inc., New York 2001.
160. J. C. Seferis, L. Nicolais (eds.): *The Role of Polymer Matrix in Processing and Structural Properties of Composites*, Plenum Press, New York 1983, pp. 127–145.
161. J. K. Gillham, *Encyclopedia of Polymer Science and Engineering*, 2nd ed., John Wiley & Sons, Inc., New York 1986, pp. 519, 524.
162. X. Wang, J. K. Gilham, *J. Coat. Technol.* **64** (1992) 37, 45.
163. H. E. Adabbo, R. J. J. Williams, *J. Appl. Polym. Sci.* **27** (1982) 1327, 1334.
164. J. B. Enns, J. K. Gillham, *J. Appl. Polym. Sci.* **28** (1983) 2567.
165. C. M. Tung, J. P. Dynes, *J. Appl. Polym. Sci.* **27** (1982) 569, 574;
M. E. Smith, H. Ishida, *J. Appl. Polym. Sci.* **73** (1999) 593, 600.
166. T. Glauser, M. Johansson, A. Hult, *Polymer* **40** (1999) 5297, 5302.
167. K. Dusek, in K. Dusek (ed.): *Epoxy Resins and Composites III*, Advances in Polymer Science 78, Springer-Verlag, Berlin 1986, pp. 1, 59.
168. S. G. Croll, in K. L. Mittal (ed.): *Adhesion Aspects of Polymeric Coatings*, Plenum Press, New York 1983.
169. D. B. Adolf, J. E. Martin, *J. Comp. Mater.* **30** (1996) no. 1, 13–34.
170. J. C. Arnold, *Polym. Eng. Sci.* **35** (1995) no. 2, 165–169.
171. E. S.-W. Kong, in K. Dusek (ed.): *Epoxy Resins and Composites IV*, Advances in Polymer Science 80, Springer-Verlag, Berlin 1986, pp. 125, 172.
172. R. S. Durran, *J. Non-Cryst. Solids* **131–133** (1990) 497–504;
D. J. Plazek, Z. N. Frund, *J. Polym. Sci., Bert. Polym. Phys.* **28** (1990) 431–448.
173. S. L. Maddox, J. K. Gillham, *J. Appl. Polym. Sci.* **64** (1997) 55–67.
174. L. W. Hill, *J. Coat. Technol.* **64** (1992) no. 808, 29.
175. C. Billaud, M. Vandeuren, R. Legras, V. Carlier, *Appl. Spectr.* **56** (2002) 1413–1421.
176. G. Wisanrakkit, J. K. Gilham, *J. Appl. Polym. Sci.* **41** (1990) 2885, 2929.
177. C. W. Macosko, D. R. Miller, *Macromolecules* **9** (1976) 199, 206;
J. Mijovic, C. H. Lee, *J. Appl. Polym. Sci.* **29** (1989) 2155, 2170.
178. C. A. May (ed.): *Chemorheology of Themosetting Polymers*, ACS Symposium Series 227, American Chemical. Society, Washington, D.C. 1983.
179. S. D. Senturia, N. F. Sheppard Jr., in K. Dusek (ed.): *Epoxy Resins and Composites IV*, Advances in Polymer Science 80, Springer-Verlag, Berlin 1986, pp. 1, 47.
180. G. P. Johari, D. A. Wasylyshyn, *J. Polym. Sci., Polym. Phys.* **38** (2000) 122–126.
181. J. H. Flynn, in H. F. Mark, N. M. Bikales, C. G. Overberger, G. Menges, eds. *Encyclopedia Polymer Science and Engineering*, 2nd ed., Suppl. Vol., John Wiley & Sons, New York 1989, pp. 715, 723.
182. Annu. Book ASTM Stand. Web site: http://www.astm.org
183. J. M. Charlesworth, *Polym. Eng. Sci.* **28** (1988) 230.
184. M. J. Marks, *Polym. Mater. Sci. Eng.* **66** (1992) 365.
185. A. J. Lesser, E. Crawford, *J. Appl. Polym. Sci.* **66** (1997) 387, 395;
E. Crawford, A. J. Lesser, *J. Appl. Polym. Sci., Part B: Polym. Phys.* **36** (1998) 1371, 1382.
186. C. W. Macosko, D. R. Miller, *Macromolecules* **9** (1976) 206, 211;
D. R. Miller, E. M. Valles, C. W. Macosko, *Polym. Eng. Sci.* **19** (1979) 272, 283.
187. E. D. Crawford, A. J. Lesser, *Polym. Eng. Sci.* **39** (1999) Feb., 385, 392.
188. J. G. Williams, *Fracture Mechanics of Polymers*, Ellis Horwood, Chichester, U.K. 1984.
189. A. J. Kinloch, in K. Dusek (ed.): *Epoxy Resins and Composites 1*, Advances in Polymer Science 72, Springer-Verlag, Berlin 1986, pp. 1, 59.
190. W. J. Cantwell, H. N. Kausch, in B. Ellis (ed.): *Chemistry and Technology of Epoxy Resins*, 1st ed., Blackie Academic & Professional, Glasgow, U.K. 1993, pp. 144, 174.
191. J. M. Charlesworth, *Polym. Eng. Sci.* **28** (1988) 221.
192. E. Urbaczweski-Espuche, J. Galy, J. Ferard, J. Pascault, H. Sautereau, *Polym. Eng. Sci.* **31** (1991) 1572.
193. J. Galy, J. Gerard, H. Sautereau, R. Frassine, A. Pavan, *Polym. Networks Blends* **4** (1994) 105.

194. R. A. Dubois, D. S. Wang, D. Sheih, in C. B. Arends (ed.): *Polymer Toughening*, Marcel Dekker, Inc., New York 1996, pp. 381, 409.
195. R. A. Dubois, D. S. Wang, *Prog. Org. Coat.* **22** (1993) 161.
196. J. P. Pascault, H. Sautereau, J. Verdu, R. J. J. Williams, *Thermosetting Polymers*, Marcel Dekker, Inc., New York 2001, p. 400.
197. L. T. Drzal, in K. Dusek (ed.): *Epoxy Resins and Composites II*, Advances in Polymer Science 75, Springer-Verlag, Berlin 1986, pp. 3, 32.
198. R. G. Schmidt, J. P. Bell, in K. Dusek (ed.): *Epoxy Resins and Composites II*, Advances in Polymer Science 75, Springer-Verlag, Berlin 1986, pp. 33, 71.
199. Z. W. Wicks Jr., F. N. Jones, S. P. Pappas, *Organic Coatings Science and Technology*, 2nd ed., Wiley-Interscience, New York 1999, pp. 77, 111.
200. W. Brostow, R. D. Corneliussen (eds.): *Failure of Plastics*, Hanser, Munich 1989.
201. Celanese Corp., US 4,051,195, 1977 (M. F. McWhorter).
202. R. F. Eaton, K. T. Lamb, in SPE Proceedings, Epoxy Resin Formulators Division of the Society of Plastic Industry, Inc., Aspen, CO, May 1996 (Paper 7).
203. M. DiBenedetto, Mod. Paint. Coat. (Jul. 1980) 39.
204. J. Melloan, paper presented at the Epoxy Resins Formulators Division of SPI Meeting, Atlanta, GA, Nov. 1983, The Society of Plastics Industry, New York.
205. Z. Wang, J. Massam, T. J. Pinnavaia, in T. J. Pinnavaia, G. W. Beall (eds.): *Polymer-Clay Nanocomposites*, John Wiley & Sons, New York 2000, pp. 127, 149.
206. J. H. Park, S. C. Jana, *Macromolecules* **36** (2003) 2758–2768.
207. Proceedings of POSS (Hybrid Plastics, Inc.) Nanotechnology Conference, Huntington Beach, CA, Sept. 2002.
208. M. J. Biercuk, M. C. Llaguno, M. Radosavljevic, J. K. Hyun, A. T. Johnson, J. E. Fischer, *J. Appl. Phys. Lett.* **80** (2002) 2767, 2769.
209. R. A. Dubois, P. S. Sheih, *J. Coat. Technol.* **64** (1992) 51.
210. B. L. Burton, J. L. Bertram, in C. B. Arends (ed.): *Polymer Toughening*, Marcel Dekker, Inc., New York 1996, pp. 339, 379.
211. H. S. Sue, E. I. Garcia-Meitin, D. M. Pickleman, in N. P. Cheremisinoff (ed.): *Rubber-Modified High Performance Epoxies*, CRC Press, Boca Raton FL 1993, Chapt. 18, pp. 661, 700.
212. A. F. Yee, J. Du, M. D. Thouless, in D. R. Paul, C. R. Bucknall (eds.): *Polymer Blends*, Vol. 2, John Wiley & Sons, Inc., New York 2000, pp. 226, 267.
213. Y. Huang, D. L. Hunston, A. J. Kinloch, C. K. Riew, in C. K. Riew, A. J. Kinloch (eds.): *Toughened Plastics 1*, Advances in Chemistry Series 233, American Chemical Society, Washington, D.C. 1993, pp. 1, 35.
214. C. K. Riew, E. H. Rowe, A. R. Siefert, *ACS Adv. Chem. Ser.* **154** (1976) 326.
215. H. Chen, N. Verghese, H. Pham, N. Jivraj, paper presented at the 9th Annual International Conference on Composites Engineering, San Diego, CA, July 2002.
216. Dow Chemical Co., US 4 708 996, 1987 (D. K. Hoffman, C. B. Arends).
217. H. J. Sue, *Polym. Eng. Sci.* **31** (1991) 275.
218. Dow Chemical Co., US 4 594 291, 1986; Dow Chemical Co., US 4 725 652, 1988 (J. L. Bertram, L. L. Walker, V. I. W. Stuart).
219. J. M. Dean, P. M. Lipic, R. B. Grubbs, R. F. Cook, F. S. Bates, *J. Polym. Sci., Part B: Polym. Phys.* **39** (2001) 2996, 3010; S. Ritzenthaler, F. Court, L. David, *Macromolecules* **35** (2002) 6245, 6254.
220. S. R. White, N. K. Sottos, P. H. Geubelle, J. S. Moore, M. R. Kessler, S. R. Sriram, E. N. Brown, S. Viswanathan, *Nature* **409** (2001) 794, 797.
221. E. O. C. Greiner, F. Dubois, M. Yoneyama, *Epoxy Resins*, Chemical Economics Handbook (CEH) Marketing Research Report, Stanford Research Institute (SRI) International, Menlo Park, CA 2001.
222. E. Linak, F. Dubois, M. Yoneyama, *Epoxy Surface Coatings*, Chemical Economics Handbook (CEH) Marketing Research Report, Stanford Research Institute (SRI) International, Menlo Park, CA 2001.
223. E. W. Flick, *Contemporary Industrial Coatings, Enviromentally Safe Formulations*, Noyes Publishers, Park Ridge, NJ 1985.
224. R. F. Eaton, K. T. Lamb, *J. Coat. Technol.* **68** (1996) 49.
225. PPG Industries, Inc., US 3 984 299, 1976 (R. D. Jerabek); PPG Industries, Inc., US 4 009 133, 1977 (J. E. Jones).
226. SCM Corp., US 4 212 781, 1980 (J. M. Evans, V. W. Ting).
227. E. C. Galgoci, P. C. Komar, Paints Coat. Ind. (1994) Aug., 50.

228. D. S. Kincaid, P. Komar, J. R. Hite, paper presented at the 24th International Waterborne, High Solids and Powder Coatings Symposium, New Orleans, LA, 2002.
229. M. D. Soucek, G. Teng, S. Wu, *J. Coat. Technol.* **73** (2001) Oct., 117, 125.
230. E. Linak, A. Kishi, Thermoset Powder Coatings, Chemical Economics Handbook (CEH) Marketing Research Report, Stanford Research Institute (SRI) International, Menlo Park, CA, 2002.
231. R. Amey, R. Farabaugh, Mod. Paint Coat. (1997) June, , 28, 30.
232. H. Nowack, Proc. PCE 2000 (2000) 365, 375.
233. P. Horinka, Powder Coating Magazine (Aug. 2002) 33, 48.
234. K. Buysens, K. Jacques, Eur. Coat. J. (Sept. 2001) 22, 26.
235. D. L. Steele, Surface Coat. Aus. (Oct. 1992) 6, 12;
W. Wood, J. Protective Coat. Linings (Apr. 1987) 32, 38.
236. US 5 275 645 1994 L. R. Ternoir, R. E. Foscante, R. L. Gasmens);
Ameron International Corp., US 5 618 860, 1997 (N. R. Mowrer, R. E. Foscante, J. L. Rojas).
237. Glidden Co., US 5 508 325, 1996 (G. P. Craun, D. J. Telford, H. J. DeGraaf).
238. Shell Oil Co., US 4 119 595, 1978 (R. S. Bauer, J. A. Lopez).
239. J. T. K. Woo, V. Ting, J. Evans, C. Ortiz, G. Carlson, R. Marcinko, in R. S. Bauer (ed.): *Epoxy Resin Chemistry 2*, American Chemical Society, Washington, D.C. 1983, pp. 283, 300.
240. Glidden Co., US 5 532 297, 1996 (J. T. K. Woo, G. C. Pompighano, D. E. Awarski, K. A. Packard).
241. P. Newman, The Canmaker (Apr. 2001) 57.
242. D. Hayes, The Canmaker (Nov. 1998) 37, 40.
243. Toyo Seikan Kaisha Ltd., Japan, EP Appl. 493133 A2, 1992 (N. Sato, K. Imazu).
244. Bisphenol A: Information Sheet from the Global Industry Group, Safety of Epoxy Can Coatings (Oct. 2002) at http://www.bisphenol-a.org;
Statement from CEPE (European Confederation of Paints, Printing Inks and Artists Colours Manufacturers), Feb. 2000.
245. P. Hitchin, The Canmaker (Oct. 2001) 43, 46.
246. BP Corp. North America Inc., US 6 472 480, 2002 (R. L. Anderson);
Valspar Corp., US 6 235 102 B1, 2002 (G. G. Parekh, L. P. Seibel).
247. PPG Industries, US 4 650 718, 1987 (D. A. Simpson, D. L. Singer, R. Dobenko, W. P. Blackburn, C. M. Kania).
248. L. Mauro, *25th FATIPEC Congr.* **4** (2000) 1, 10.
249. BASF, EU Appl. 581175 A2, 1994 (K. Huemke, C. Sinn).
250. Kansai Paint Co. Ltd; Nippon Packaging K.K., JP 08003483 A2, 1996 (H. Haishi, M. Kume, H. Ishiiand, K. Myawaki).
251. Nippon Paint, US Appl. 2002139673 A1, 2002 (Y. Kojima, M. Yamada).
252. Nippon Paint Co., Ltd., Japan, EU Appl. 0974623 A2, 2000 (H. Sakamoto, T. Kawanami, I. Kawakami, T. Kokubun, T. Saito, S. Yoshimatsu).
253. G. Mauer, D. Singer, PPG Technol. J (1999) 63, 72.
254. Y. Okude, S. Ishikura, *Prog. Org. Coat.* **26** (1995) 197, 205.
255. G. E. Green, B. P. Stark, S. A. Zahir, *J. Macromol. Sci. Revs. Macromol. Chem.* **21** (1981/1982) 187.
256. P. Kelly, G. Pritchard (eds.): *Reinforced Plastics Durability*, CRC Press, Boca Raton, FL 1999, pp. 282–321.
257. M. R. Thoseby, B. Dobinson, C. H. Bull, *Br. Polym. J.* **18** (1986) 286.
258. U. Helrold, J. Mason, N. Verghese, H. Chen, H. Reddy, in Proceedings of NACE Corossion 2003 Conference, Houston, TX., March 16–20, 2003.
259. D. Kriegh (ed.): Epoxies with Concrete, American Concrete Institute, Detroit, MI 1968; R. L. McGown, in Proceedings of the 1991 Steel, Structures, Painting Council (SSPC) National Conference and Exhibition, Long Beach, CA.
260. H. Mihashi, Y. Kaneko, *Transactions Mater. Res. Soc. Japan* **25** (2000) 557, 560.
261. S. Ehrler, PC Fab (Apr. 2002) 32, 38;
Part 2, PC Fab (May 2002) 32, 36.
262. W. Christiansen, D. Shirrell, B. Aguirre, J. Wilkins, in Proceedings of the Technical Conference, IPC Printed Circuits Expo 2001, Anaheim, CA, Apr. 2001, pp. SO3-1-1, SO3-1-7.
263. J. Sharma, M. Choate, S. Peters, in Proceedings of the Technical Conference, IPC Printed Circuits Expo 2002, Long Beach, CA, Mar. 2002, pp. SO5-1-1, SO5-1-8.
264. D. Sober, in Base Material Basics, IPC Printed Circuits Expo 2002, Long Beach, Ca.
265. Ciba-Geigy Corp., US 5 506 313, 1996 (P. Flury, C. W. Mayer, W. Scharf, E. Vanoli).

266. C. S. Cho, L. W. Chen, Y. S. Chiu, *Polym. Bull.* **41** (1998) no. 1, 45, 52.
267. D. K. Luttrull, F. E. Hickman III, Future Circuits International, Mar. 2001.
268. Ciba-Geigy AG, US 3 754 071, 1973 (O. Ernst, E. Kusenberg, E. Hubler, H. R. Aus Der Au).
269. K. Ito, Y. Nakamura, *IEEE Electrical Insulation Mag.* **6** (1990) no. 4, 25, 32.
270. T. M. Goulding, in A. Pizzi, K. L. Mittal (eds.): *Handbook of Adhesive Technology*, 2nd ed., Marcel Dekker, Inc., New York 2003, pp. 823, 838.
271. A. F. Lewis, in C. A. May, Y. Tanaka (eds.): *Epoxy Resins Chemistry and Technology*, 2nd ed., Marcel Dekker, Inc., New York 1988, pp. 653, 718.
272. J. Sheehan, *the Epoxy Resin Formulators Training Manual*, The Society of the Plastics Industry, Inc., New York 1984, Chapt. XV, p. 175.
273. J. Waechter, *Patty's Industrial Hygiene and Toxicology*, 5th ed., Vol. 6, John Wiley & Sons, Inc., New York 2001, Chapts. 82 and 83, pp. 993, 1145.
274. *Epoxy Resin Systems Safe Handling Guide*, the Society of Plastics Industry (SPI), Inc., New York, Sept. 1997. Publication No. AE-155. Web site: http://www.plasticsindustry.org/about/epoxy-/epoxy_guide.htm.
275. Epoxy Resins and Curing Agents, prepared by the Epoxy Resins Committee of the Association of Plastics Manufacturers in Europe (APME), Jan. 1996. Web site: http://www.apme.org/dashboard-/presentation_layer_htm/dashboard.asp.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Esters, Organic

Standard Article

Wilhelm Riemenschneider[1] and Hermann M. Bolt[2]

[1]from Hoechst AG, Frankfurt, Federal Republic of Germany

[2]Institut für Arbeitsphysiologie an der Universität Dortmund, Germany

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a09_565.pub2

Article Online Posting Date: April 30, 2005

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Physical Properties
3. Chemical Properties
3.1. Hydrolysis
3.2. Transesterification
3.3. Ammonolysis
3.4. Reduction
3.5. Claisen Condensation
3.6. Pyrolysis
4. Natural Sources
5. Production
5.1. Esterification of Carboxylic Acids
5.1.1. Catalysts
5.1.2. Industrial Processes
5.2. Alkylation of Metal Carboxylates
5.3. Acylation with Acyl Halides
5.4. Acylation with Carboxylic Anhydrides
5.5. Transesterification
5.6. Acylation with Ketenes
5.7. Carbonylation
5.8. Condensation of Aldehydes
5.9. Alcoholysis of Nitriles
5.10. Acylation of Olefins
5.11. Ethoxylation of Carboxylic Acids
5.12. Orthocarboxylic Acid Esters
5.13. Lactones
6. Environmental Protection
7. Quality Specifications and Analysis
8. Storage and Transportation
9. Uses and Economic Aspects
10. Toxicology and Occupational Health

[Top of Page]

1. Introduction

Organic esters are compounds with the general formula

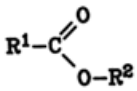

where R^1 and R^2 represent either the same or different aliphatic, aromatic, or heterocyclic groups. Esters are classified as carboxylic acid derivatives; they are prepared mainly by the reaction of a carboxylic acid and an alcohol (esterification).

Ortho esters are derived from ortho acids:

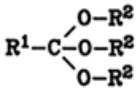

Ortho esters can be hydrolyzed to carboxylic acids and alcohols, but their properties are more similar to those of acetals. They are dealt with briefly in

Section Orthocarboxylic Acid Esters.

If a molecule possesses both a carboxyl and a hydroxyl group, it may form either polyesters:

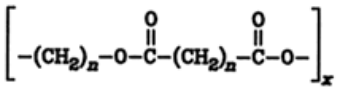

or cyclic internal esters (lactones) (Section Lactones):

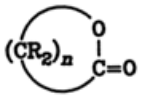

Ring opening reactions of lactones lead to the same type of polyesters. Another type of polyesters with different characteristics is usually produced through condensation of dicarboxylic or polycarboxylic acids with di- or polyhydric alcohols.

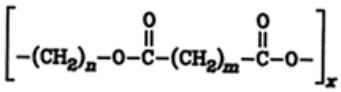

Organic esters are of considerable economic importance. Esters of glycerol with fatty acids (glycerides) occur naturally in large quantities in fats and fatty oils (see Chap. Natural Sources and ⟶ Fats and Fatty Oils); they are used predominantly in foods, but also as raw materials, especially in the production of surfactants (⟶ Surfactants; ⟶ Laundry Detergents; ⟶ Soaps; ⟶ Fatty Acids; ⟶ Fatty Alcohols).

Cellulose acetate is used on a large scale in fiber production. Synthetic esters are of increasing importance in many applications, e.g., fibers, films, adhesives, and plastics articles. Various synthetic esters have been developed for specific uses, e.g., solvents, extractants, plasticizers, lubricants, lubricant additives, and lacquer bases (see also ⟶ Solvents; ⟶ Plasticizers). A variety of volatile esters are used as aromatic materials in perfumes, cosmetics, and foods (⟶ Flavors and Fragrances).

This article gives a general survey of production methods, properties, and uses of esters. The commercially most important esters are treated under separate keywords: ⟶ Formic Acid; ⟶ Acetic Acid; ⟶ Propionic Acid and Derivatives; ⟶ Adipic Acid; ⟶ Acrylic Acid and Derivatives; ⟶ Methacrylic Acid and Derivatives; ⟶ Benzoic Acid and Derivatives; ⟶ Phthalic Acid and Derivatives; ⟶ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid; ⟶ Fats and Fatty Oils; ⟶ Waxes; ⟶ Cellulose Esters; ⟶ Polyesters.

Nomenclature. The term ester was coined by GMELIN (1850) from the German word **Es**sig-ä**th**e**r** (acetic ether, ethyl acetate). According to IUPAC rule C–463, esters of carboxylic acids are named by substituting the ending -ate for the -ic acid; the alkyl or aryl group is cited first. For example, the methyl ester of propionic acid is called methyl propionate [2]. Alternatively, the term propionic acid methyl ester is used.

Acid esters, such as the monoesters of dibasic acids, are named by inserting the word hydrogen between the names of the alcohol and acid (IUPAC rule C–463.2). Thus, methyl hydrogen succinate is the monomethyl ester of succinic acid.

Esters of glycerol such as fats and fatty oils are known as glycerides. Trivial names are usually preferred:

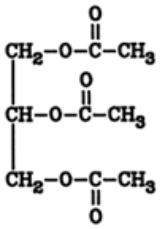

Triacetin

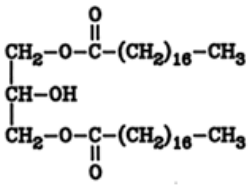

1,3-Distearin

Ortho esters are named as derivatives of ortho acids, e.g., trimethyl orthoacetate (IUPAC rule C–464).

In substitutive nomenclature, different prefixes are used for the ester group, depending on the position of the principal group R:

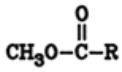

Methoxycarbonyl derivative

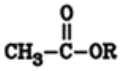

Methylcarbonyloxy derivative

[Top of Page]

2. Physical Properties

The physical properties of commercially important aliphatic and aromatic esters are listed in Table 1 [3]. The lower esters are colorless, mobile, and highly volatile liquids that usually have pleasant odors. As the molecular mass increases, volatility decreases and the consistency becomes waxy, then solid, and eventually even brittle, often with formation of lustrous crystals.

Table 1. Physical properties of the most important esters

	CAS registry number	M_r	n_D^{20} (DIN 51 423)	d_{20}^{20} (DIN 51 757), g/cm^3	*bp* at 101.3 kPa (DIN 53 171), °C	*fp* or *mp*, °C

Esters of aliphatic acids						
Methyl formate	[107-31-3]	60.05	1.344	0.975	32.0	– 99.8
Ethyl formate	[109-94-4]	74.08	1.3598	0.924	54.3	– 80
Methyl acetate	[79-20-9]	74.08	1.3594	0.933	57	– 98.1
Ethyl acetate	[141-78-6]	88.10	1.3723	0.900	77.06	– 83.6
Propyl acetate	[109-60-4]	102.13	1.3844	0.887	101.6	– 92.5
Isopropyl acetate	[108-21-4]	102.13	1.3773	0.872	90	– 73.4
Butyl acetate	[123-86-4]	116.16	1.3951	0.882	126	– 73.5
Isobutyl acetate	[110-19-0]	116.16	1.3902	0.871	117.2	– 98.6
Pentyl acetate	[628-63-7]	130.18	1.4023	0.876	149.25	– 70.8
2-Ethylhexyl acetate	[103-09-3]	172.26	1.4204	0.873	199.3	– 93
Vinyl acetate	[108-05-4]	86.10	1.3959	0.932	72.2 – 72.3	– 93.2
Ethylene glycol diacetate	[111-55-7]	146.15	1.415	1.128^{4}	191	– 31
Methoxyethyl acetate	[110-49-6]	118.13	1.4019	1.0067	145.0	– 65.1
2-Ethoxyethyl acetate	[111-15-9]	132.16	1.4058	0.975	156.4	– 61.7
2-Butoxyethyl acetate	[112-07-2]	160.21		0.943	187.8	
2-(2-Ethoxyethoxy)ethyl acetate	[111-90-0]	176.21		1.011	217.4	– 25
2-(2-Butoxyethoxy)ethyl acetate	[112-34-5]	204.27		0.981	247	– 32.2
Benzyl acetate	[140-11-4]	148.15	1.5232	1.055	215.5	– 51.5
Glyceryl triacetate	[102-76-1]	218.23		1.161	258	– 78
Methyl propionate	[554-12-1]	88.10	1.3775	0.915	79.8	– 87.5
Ethyl propionate	[105-37-3]	102.13	1.3839	0.892	99.10	– 72.6
Glyceryl tripropionate	[139-45-7]	260.3	1.4318	1.100	175 – 176 (2.67 kPa)	
Methyl butyrate	[623-42-7]	102.13	1.3878	0.898	102.3	– 84.8
Ethyl butyrate	[105-54-4]	116.16	1.4000	0.878	121.6	– 100.8
Butyl butyrate	[109-21-7]	144.22	1.4075	0.871	166.6	– 91.5
Methyl isobutyrate	[547-63-7]	102.13	1.3840	0.891	92.6	– 84.7
Ethyl isobutyrate	[97-62-1]	116.16	1.3903	0.870	110 – 111	– 88
Isobutyl isobutyrate	[97-85-8]	144.22	1.3999	0.875	148.7	– 80.7
Dimethyl adipate	[627-93-0]	174.20	1.4283	1.060	115 (1.73 kPa)	10.3
Diethyl adipate	[141-28-6]	202.25	1.4372	1.008	245	– 19.8
Bis(2-ethylhexyl) adipate	[103-23-1]	370.58		0.927	214 (0.67 kPa)	– 60
Methyl stearate	[112-61-8]	298.50	1.457^{40}	0.836^{60}	215 (2.0 kPa)	40
Ethyl stearate	[111-61-5]	312.52	1.429^{40}	0.848^{36}	213 – 215 (2.0 kPa)	33.7
Butyl stearate	[123-95-5]	340.58		0.855	343	27.5
Dodecyl stearate	[5303-25-3]	440.80	1.433^{50}			28
Hexadecyl stearate	[1190-63-2]	496.9	1.441^{70}			57
Methyl acrylate	[96-33-3]	86.09	1.4040	0.953	80.5	< – 75
Ethyl acrylate	[140-88-5]	100.11	1.4068	0.923	99.8	< – 72
Butyl acrylate	[141-32-2]	128.17	1.4185	0.898	69 (6.7 kPa)	– 64.6
2-Ethylhexyl acrylate	[103-11-7]	184.28		0.887	130 (6.7 kPa)	– 90
Methyl methacrylate	[80-62-6]	100.13	1.4119^{25}	0.944	100 – 101	– 48
Esters of aromatic acids						
Methyl benzoate	[93-58-3]	136.14		1.094^{15}	199.5	– 12.5
Ethyl benzoate	[93-89-0]	150.17		1.051^{15}	212.9	– 34.2
Methyl salicylate	[119-36-8]	152.14		1.184^{15}	223.3	–8.6
Ethyl salicylate	[118-61-6]	166.17		1.137^{15}	231.5	1.3
Phenyl salicylate	[118-55-8]	214.21			172.3 (1.6 kPa)	43
Dimethyl phthalate	[131-11-3]	194.19			282	0 – 2
Diethyl phthalate	[84-66-2]	222.23	1.4990	1.118	295	– 33
Bis(2-ethylhexyl)phthalate	[117-81-7]	390.57		0.986	231 (0.66 kPa)	– 46
Dimethyl isophthalate	[1459-93-4]	194.18			124 (1.6 kPa)	67 – 68
Dimethyl terephthalate	[120-61-6]	194.18			288	14
Trimethyl trimellitate	[2459-10-1]	252.30			194 (1.6 kPa)	– 13
Methyl anthranilate	[134-20-3]	151.16	1.584	1.168	132 (1.86 kPa)	24 – 25
Benzyl cinnamate	[103-41-3]	238.27			240 – 244 (3.3 kPa)	39

The melting point of an ester is generally lower than that of the corresponding carboxylic acid. The boiling point depends on the chain length of the alcohol component and may eventually exceed that of the acid. The lower esters are relatively stable when dry and can be distilled without decomposition.

With the exception of low molecular mass compounds formed by short-chain carboxylic acids, esters are insoluble in water (Table 2). By contrast, they are readily miscible with many organic solvents.

Table 2. Water solubility of lower esters

	CAS registry number	Solubility, g in 100 g of water	Temperature, °C
Methyl formate	[107-31-3]	21.3	20
Ethyl formate	[109-94-4]	11.0	18
Propyl formate	[110-74-7]	2.2	22
Butyl formate	[592-84-7]	1.0	22
Methyl acetate	[79-20-9]	25	20
Ethyl acetate	[141-78-6]	7.7	20
Propyl acetate	[109-60-4]	1.5	16
Isobutyl acetate	[110-19-0]	0.8	20
Butyl acetate	[123-86-4]	0.7	20
Methyl propionate	[554-12-1]	6.3	20
Ethyl propionate	[105-37-3]	1.7	20
Methyl butyrate	[623-42-7]	1.6	21
Ethyl butyrate	[105-54-4]	0.6	22
Methyl acetoacetate	[105-45-3]	40	20
Ethyl acetoacetate	[141-97-9]	9	20

[Top of Page]

3. Chemical Properties

Due to the large number of possible acid and alcohol moieties, the chemical properties of esters may differ considerably. Only typical reactions applicable to the majority of esters are described in the following sections.

3.1. Hydrolysis

Esters are gradually split into the acid and alcohol through the action of humidity:

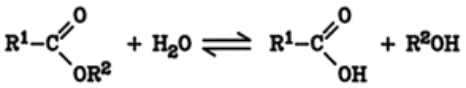

This hydrolysis is catalyzed by acids (protons) or bases (hydroxyl ions). The catalytic effect of bases is generally stronger. Industrially, the hydrolysis of esters is usually carried out with bases. However, acid hydrolysis is preferred for the production of pure acids and for alkali-sensitive compounds.

Hydrolysis of esters by use of water and a mineral acid leads to an equilibrium mixture of ester, alcohol, and free carboxylic acid. If bases are used, the salt of the carboxylic acid is formed; the carboxylic acid is thus removed from the equilibrium and the reaction becomes irreversible:

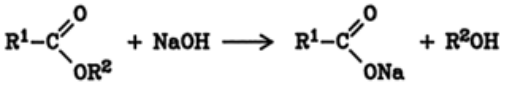

Higher temperatures accelerate the reaction. Table 3 gives the relative rates of hydrolysis for some esters: hydrolysis is hindered by bulky acyl or alcohol groups, whereas acidifying groups increase the reaction rate.

Table 3. Relative rates of acid-catalyzed hydrolysis of esters

	CAS registry number	$k_{ester}/k_{CH_3COOC_2H_5}$
$CH_3COOC_2H_5$	[141-78-6]	1
$C_2H_5COOC_2H_5$	[105-37-3]	0.47
$(CH_3)_2CHCOOC_2H_5$	[97-62-1]	0.10
$(CH_3)_3CCOOC_2H_5$	[3938-95-2]	0.01
$CH_3COOCH_2CH(CH_3)_2$	[110-19-0]	0.70
$CH_3COOCH_2C(CH_3)_3$	[926-41-0]	0.18
$CH_3COOCH_2C(C_2H_5)_3$	[10332-40-8]	0.03
CH_3COOCH_3	[79-20-9]	1.6
$(COOCH_3)_2$	[553-90-2]	284
$CH_2ClCOOCH_3$	[96-34-4]	1270
$CHCl_2COOCH_3$	[116-54-1]	27000
$CH_3COCOOC_2H_5$	[617-35-6]	17000

Esters of secondary alcohols are hydrolyzed more slowly than those of primary alcohols. Esters of tertiary alcohols, e.g., *tert*-butyl esters, are exceptional in that they are readily hydrolyzed in the presence of acids and then often yield α-olefins instead of alcohols. Some esters are hydrolyzed only under extreme conditions; for example, esters of pivalic acid [75-98-9] (2,2-dimethylpropionic acid) and ortho-substituted benzoic acids react only in concentrated H_2SO_4.

Basic Hydrolysis. Basic hydrolysis (saponification) is one of the earliest known chemical reactions. By this method, soaps are obtained from natural fats and oils, with glycerol as a byproduct. Saponification based on wood ash as the alkali was described as early as 2500 BC; the principle is still

employed today (⟶ Soaps). The reaction was investigated by SCHEELE in 1779.

Basic hydrolysis requires stoichiometric amounts of alkali. The reaction is carried out by boiling the ester in an aqueous alkaline solution. Under these conditions, the alkali salt of the carboxylic acid is dissolved.

Acidic Hydrolysis. If the goal is to obtain free carboxylic acids rather than soaps, acid hydrolysis is the preferred method. Free acids are needed for various industrial purposes (⟶ Carboxylic Acids, Aliphatic; ⟶ Fatty Acids). Complete hydrolysis can only be achieved by removal of the alcohol from the equilibrium. The reaction rate in dilute acids is generally fairly low, mainly because of the poor solubility of esters in water. Therefore, in practice, emulsifiers such as sulfonated oleic acid or sulfonated aromatic compounds (Twitchell reagent) are added and the reaction is carried out at ca. 100 °C under pressure.

Another industrial process involves high-temperature hydrolysis with steam at 170 –300 °C under pressure. Zinc, calcium, and magnesium compounds are used as catalysts [4]. Vinyl esters are readily hydrolyzed in the presence of palladium salts [5].

Enzymatic Hydrolysis. Specific lipolytic enzymes, the so-called esterases, effect hydrolysis at temperatures below 40 °C. Thus, lipase occurs in the digestive tracts of humans and most animals and in plants. Castor oil lipase is of commercial importance in hydrolysis of fats under mild conditions. The enzymes are especially effective in the weakly acidic range. Lipases may achieve importance as detergent additives; at low temperatures (20 °C) they facilitate the removal of fat-containing soils (⟶ Laundry Detergents).

3.2. Transesterification

When esters are heated with alcohols, acids, or other esters, the alcohol or acid groups are exchanged more or less completely. This process is called transesterification. It is accelerated in the presence of a small amount of acid or alkali.

Three types of transesterification are known:

1. Exchange of alcohol groups (alcoholysis)

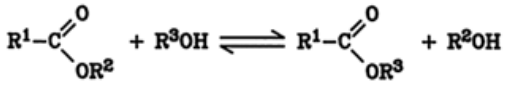

2. Exchange of acid groups (acidolysis)

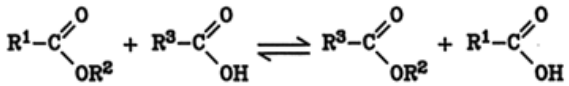

3. Ester – ester interchange

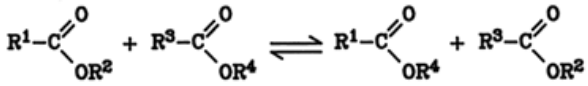

All three are equilibrium reactions and proceed to completion only if one component is removed from the reaction mixture.

Dispersed alkali metals, mainly sodium, are suitable catalysts for the transesterification of fats [6]. Tin salts have been used in the production of aryl esters, e.g., phenyl esters, at 190 – 250 °C. Methacrylic acid esters of higher alcohols, e.g., ethylhexanol, can be obtained by transesterification in the presence of tin compounds [7] or with magnesium alkoxides [8]. Other catalysts are discussed in Section Transesterification.

Transesterification reactions play a major role in industry and are important in laboratory practice and in analytical chemistry. They can be used to reduce the boiling point of esters by exchanging a long-chain alcohol group for a short one such as methanol, e.g., in the analysis of fats, oils, and waxes (see Chap. Quality Specifications and Analysis) and in vapor-phase reactions. Transesterification reactions can also be useful if direct esterification with the desired alcohol is technically difficult because of physical reasons (e.g., high boiling point, low solu-bility, or high viscosity). Examples are to be found in polymer chemistry. Thus, poly(ethylene terephthalate) is produced mainly by transesterification of dimethyl terephthalate with ethylene glycol. Propanediols and 1,4-butanediol are used in a similar manner. The reaction is catalyzed with calcium, antimony, titanium, and germani-um salts. A further example is the reaction of diols (e.g., bisphenol A) with carbonic acid esters to produce polycarbonates.

An industrial example of acidolysis is the reaction of poly(vinyl acetate) with butyric acid to form poly(vinyl butyrate). Often a butyric acid – methanol mixture is used and methyl acetate is obtained as a coproduct.

3.3. Ammonolysis

Ammonia and esters react to form alcohols and amides:

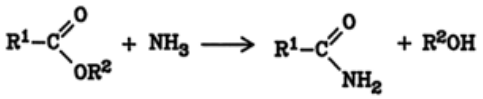

This reaction is similar to hydrolysis. It is carried out in aqueous or alcoholic ammonia. Lower esters give good yields even at room temperature; higher esters require higher temperature and pressure.

If primary or secondary amines are used, *N*-substituted amides are formed. This reaction is called aminolysis. Hydrazines yield the corresponding hydrazides.

If esters are passed with ammonia over a catalyst such as alumina at 400 – 500 °C, dehydration of the intermediate amides occurs and the corresponding nitriles are obtained directly.

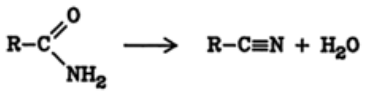

Fatty nitriles can be produced in this manner [9] (see also ⟶ Amines, Aliphatic).

3.4. Reduction

Under certain conditions esters can be reduced to alcohols.

Bouveault – Blanc Reduction. This classical method developed in 1904 uses metallic sodium in alcohol (⟶ Fatty Alcohols). However, the Bouveault – Blanc reduction has been replaced nearly completely by catalytic hydrogenation and remains important only for laboratory purposes. The advantage of the Bouveault – Blanc process is that only the ester group is hydrogenated, while carbon – carbon double bonds are not reduced.

Therefore, the reaction is suitable for the synthesis of unsaturated alcohols such as oleyl alcohol from unsaturated esters.

Catalytic Hydrogenation. A process of great industrial importance is the catalytic reduction of esters with hydrogen at 200 – 300 °C and 10 – 30 MPa (100 – 300 bar) in the presence of copper chromite catalysts [10]:

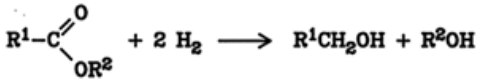

Sulfur and halogens are catalyst poisons. This method gives high yields and is used for the large-scale reduction of fats and fatty oils such as coconut oil and tallow, usually after transesterification to the methyl esters (⟶ Fatty Alcohols). The main products are dodecyl and decyl alcohols, which are important raw materials for the production of detergents. The method is also suitable for the conversion of 1,4-dimethyl terephthalate into 1,4-bis (hydroxymethyl) cyclohexane [105-08-8], an important intermediate in the production of polyesters.

Hardening of fats (hydrogenation of double bonds in unsaturated fats and oils) is carried out in the presence of nickel catalysts under slight hydrogen pressure and does not involve ester splitting.

Grignard Reaction. Esters react with alkylmagnesium halides under Grignard conditions. Formic acid esters give secondary alcohols; the other esters yield tertiary alcohols.

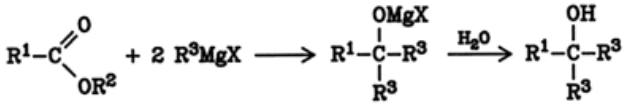

Yields are high, e.g., 95 % in the production of triphenylmethanol from ethyl benzoate and phenylmagnesium bromide. The method is used mainly on a laboratory scale.

Reduction with Complex Metal Hydrides. The reduction of esters with complex metal hydrides, especially lithium aluminum hydride, has only preparative impor-tance [11]. Its advantage is its broad applicability for all types of esters; double bonds are preserved. This method leads to very pure alcohols.

3.5. Claisen Condensation

The reaction of carboxylic acid esters with reactive methylene and methyl groups in the presence of alkaline compounds such as sodium alkoxides leads to the formation of β-oxocarboxylic acid esters (β-keto esters) [12]. The classical example is the condensation of two molecules of ethyl acetate to form acetoacetate:

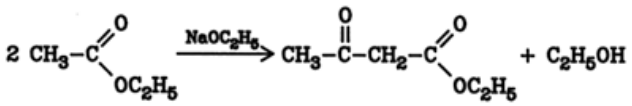

Oxocarboxylic acid esters are among the most important intermediates in the laboratory and in industry. During hydrolysis with dilute alkali, these esters readily decompose into ketones, alcohols, and carbon dioxide:

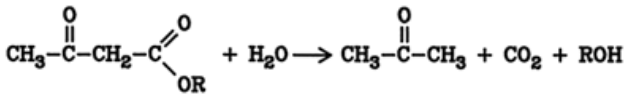

Almost any ketone can be synthesized by using this process.

In industry, the Claisen condensation has been replaced by the synthesis of acetoacetic acid derivatives from diketene [674-82-8], see also Section Acylation with Ketenes (⟶ Carboxylic Acids, Aliphatic; ⟶ Ketenes).

Condensation of esters in inert solvents in the presence of sodium rather than sodium ethoxide leads to formation of diketones and α-hydroxy ketones (acyloins) [13]. This reaction is utilized in the synthesis of large-ring compounds.

3.6. Pyrolysis

When passed at 300 – 500 °C over an inert heat-transfer agent such as quartz glass, esters decompose into carboxylic acid and the unsaturated compound that corresponds to the alcohol. Esters of primary alcohols generally lead to α-olefins:

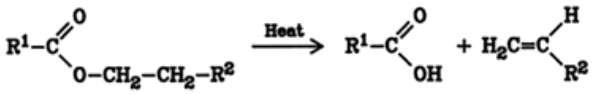

The double bond does not migrate along the carbon chain. Methyl esters and other esters without hydrogen atoms in the β-position do not yield uniform pyrolysis products.

When heated to high temperatures in the presence of metal oxides such as thorium oxide, calcium oxide, manganese chromite, or zinc chromite, fatty acid esters or free fatty acids give high yields of ketones:

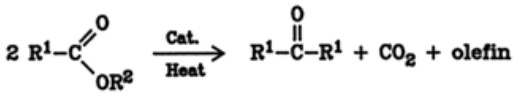

The alcohol moiety is split off as olefin.

[Top of Page]

4. Natural Sources

Esters occur naturally in large quantities in fats and fatty oils, waxes, and fruit ethers.

Fats and Fatty Oils. Fats and fatty oils (as opposed to mineral oils) are almost exclusively esters of glycerol with long- and medium-chain fatty acids.

About 1300 different fats and oils are known. They contain mainly even-numbered, straight-chain carboxylic acids (⟶ Fats and Fatty Oils). The most common are the triglycerides of stearic, palmitic, and oleic acid. The melting point of the natural products varies between 55 °C (mutton tallow) and –27 °C (nut oil) and decreases as the content of unsaturated and short-chain fatty acids increases.

Only a few of the animal fats and vegetable oils are suitable for commercial production. They are predominantly used as foods, but have also gained considerable industrial importance, e.g., in the production of lacquers, detergents, and solvents. The most frequently produced fats of animal origin are butter, beef and mutton tallow, lard, fish oil, and whale oil. Butter accounts for ca. 40 % of the world production of animal fats. The most important oils of plant origin are soybean oil, sunflowerseed oil, peanut oil, cottonseed oil, coconut oil, rapeseed oil, palm oil, and olive oil. Linseed oil, hempseed oil, and poppyseed oil contain a variable amount of unsaturated fatty acids, e.g., linoleic acid, linolenic acid, ricinoleic acid, and erucic acid. These oils gradually harden under the influence of oxygen in the air (⟶ Drying Oils and Related Products).

Waxes. Natural waxes (wax esters) are esters of monobasic carboxylic acids with higher monohydric and, less commonly, dihydric alcohols. They have been largely replaced by synthetic waxes, which are in most cases long-chain hydrocarbons, but not esters (⟶ Waxes).

The most important sources of wax esters are beeswax, whose main constituent is myricyl palmitate [6027-71-0]; spermaceti from the head oil of the sperm whale, which contains cetyl palmitate [540-10-3]; and carnauba wax from the Brazilian palm, containing myricyl cerotate [84324-99-2]. Fossil waxes, such as montan waxes in bituminous lignite and peat waxes, also have major wax ester components. Lanolin, the neutral part of wool grease, contains higher fatty acid esters of cholesterol.

Fruit Ethers. Fruit ethers, i.e., volatile esters of short- and medium-chain carboxylic acids with short- and medium-chain alcohols, are major constituents of essential oils. Their name derives from their pleasant, usually fruity odor. They are used as aromatic materials in fragrances, foods, beverages, cosmetics, and soaps (⟶ Flavors and Fragrances). Natural fruit ethers are obtained from plant materials by extraction, steam distillation, pressing, or a combination of these methods. The majority of simple, short-chain esters are currently produced synthetically.

[Top of Page]

5. Production

A great variety of production methods for carboxylic acid esters are known. However, many of these methods are valuable only for preparing special compounds [14]. In what follows, only those methods that have industrial significance are discussed.

5.1. Esterification of Carboxylic Acids

The simplest and most common method of esterification is the reaction of a concentrated alcohol with a concentrated carboxylic acid with the elimination of water:

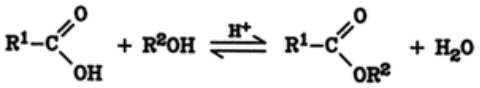

R^1, R^2 = alkyl or aryl

Esterification is the reverse of hydrolysis and leads to an equilibrium (see Section Hydrolysis). With acetic acid and ethanol, the molar equilibrium ratio of ethyl acetate to acetic acid is ca. 2 : 1 (65 mol % of ethyl acetate). The equilibrium may be shifted toward the ester by increasing the concentration of one of the reactants, usually the alcohol. However, quantitative esterification is possible only by continuous removal of one of the products, i.e., ester or water. Usually the water is distilled off as an azeotrope (see Section Industrial Processes).

Removal of the water by chemical or adsorptive binding is also possible. In preparative chemistry, calcium carbide, calcium hydride, and calcium and magnesium sulfate have proved successful; in industrial applications, activated bauxite is also used. The distilled, water-rich condensates are passed over the water-binding agents; alternatively, salts such as copper sulfate or calcium chloride can be added directly to the reaction mixture. Molecular sieves are also suitable [15]. Acetone dimethyl acetal, which binds water and simultaneously supplies methyl groups, may be used for the production of methyl esters [16].

Esterification is carried out at high temperature and in the presence of catalysts. The rate of ester formation depends on the carboxylic acid and the alcohol used. The lowest members, i.e., methanol and formic acid, react the most readily. Primary alcohols react faster than secondary alcohols, and the latter react faster than tertiary ones. Within each series, the reaction rate generally decreases with increasing molecular mass. Straight-chain acids react more readily than branched ones; the rate of esterification is lowered particularly by branching in the α-position. The relative rates of esterification are similar to those of hydrolysis (see Table 3). Esterification of aromatic acids, e.g., benzoic acid, is slow.

Experiments with primary and secondary alcohols containing ^{18}O have shown that the oxygen in the water formed during acid-catalyzed esterification originates from the acid, not the alcohol [17]. The reaction of tertiary alcohols is more complex and does not proceed clearly via a tertiary carbonium ion. The mechanism is discussed in [18].

5.1.1. *Catalysts*

Mineral Acids. Only strong carboxylic acids react sufficiently quickly without a catalyst. Generally, a strong mineral acid must be added. Suitable catalysts include sulfuric acid, hydrogen chloride, arylsulfonic acids such as *p*-toluenesulfonic acid, and chlorosulfuric acid. Phosphoric acid, polyphosphoric acids, and mixtures of acids are also recommended [19]. If the acids are adsorbed on a solid support, esterification can be carried out as a continuous process.

Ion Exchangers. The acid groups can be bound chemically to a polymeric material. Cat-ion exchangers such as sulfonated polystyrene permit esterification under mild reaction conditions. The resulting esters are generally very pure because acid-catalyzed side reactions such as dehydration, etherification, and rearrangement are almost completely suppressed [20], [21]. This method gives high yields and is widely used in industry. After the reaction water has been removed, the ion exchanger is filtered off and the ester is purified by distillation. Zeolites are also used as esterification catalysts [22].

Lewis Acids. Lewis acids such as boron trifluoride increase the reactivity of the carboxyl group toward alcohols. Boron trifluoride reacts with alcohols under proton release:

$$BF_3 + ROH \rightarrow [BF_3OR]^- + H^+$$

The proton acts as a catalyst. The complex with diethyl ether, boron trifluoride etherate [109-63-7], is generally preferred because it is easier to handle than gaseous boron trifluoride [23]. This method of esterification proceeds under mild conditions and is widely applicable. However, it is limited to the manufacture of high-value esters because of the relatively high price of BF_3. Each carboxyl group requires one equivalent of the boron trifluoride complex, which in most cases is destroyed during the further processing of the ester. All other groups that possess a free electron pair (e.g., hydroxyl, oxo, or amino groups) require an additional equivalent of BF_3.

Metal salts may also function as Lewis acids. For example, a complex acid is formed from the reaction of a zinc salt and an alcohol:

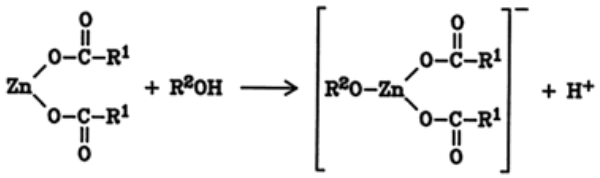

The actual catalysts are again the protons. Tin compounds have been used to prepare glyceryl esters. For the production of dimethyl terephthalate, silica treated with aluminum, titanium, zinc, or tin compounds has proved successful [24]; this esterification proceeds also in the absence of catalysts (12.5 – 14.0 MPa, 270 – 275 °C, 95 % yield).

Amphoteric Hydroxides. Esterification is catalyzed by amphoteric hydroxides of metals of groups I, II, III, and IV, especially aluminum hydroxide. The actual catalyst is probably the aluminate anion:

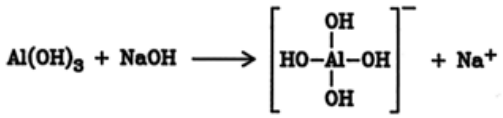

The aluminate may be generated in the reaction mixture from aluminum hydroxide and sodium hydroxide [25]. The esters produced by this method are light in color, and purification by distillation is generally unnecessary. Therefore, this process is especially suitable for preparing esters with high boiling points. Its main application is in the production of plasticizers, e.g., phthalates.

Other Compounds. A possible esterification catalyst is graphite hydrogen sulfate [12689-13-3], $(C_{24}HSO_4 \cdot 2\ H_2SO_4)_{57}$. It is obtained by electrolysis of 98 % sulfuric acid with a graphite anode.

Silanes and phosphoric acid derivatives are utilized for syntheses under mild conditions, the latter especially for phenyl carboxylates.

Acid esterification of amino acids and peptides is difficult because protecting groups and peptide bonds are easily hydrolyzed. High yields of esters of both protected and free amino acids or peptides are obtained in the presence of a mild dehydration catalyst such as sulfuryl chloride or thionyl chloride at room temperature [26].

Trifluoroacetic anhydride is a suitable catalyst for the esterification of sterically hindered acids [27].

5.1.2. *Industrial Processes*

Esterification is generally carried out by refluxing the reaction mixture until all the water has been split off. The water or the ester is removed from the equilibrium by distillation. The choice of method to achieve complete esterification depends on the boiling points of the alcohol, the acid, and the ester.

Basically, the following three possibilities exist:

1. The boiling point of the ester is lower than that of water. In this case the ester can be distilled off together with the alcohol. This method is used to produce methyl acetate, which forms an azeotrope with methanol.
2. Ester and water can be distilled off together, usually as an azeotropic mixture. On condensation, the mixture separates into an ester and a water phase. To achieve complete distillation of the ester, water or steam are often added to the reaction mixture. An example of this method is the production of *sec*-butyl acetate. Frequently alcohol, ester, and water form ternary azeotropes, e.g., during the esterification of acetic acid with ethanol.
3. The boiling point of the ester is higher than that of water. In this case, the water is distilled off, frequently as an azeotrope with the alcohol. Except for methanol, all alcohols of medium volatility form azeotropes with water. If the alcohol – water mixture separates into two phases during condensation, the alcohol can be recycled into the reaction mixture, as exem-plified in the production of *n*-dibutyl phthalate. If water and alcohol do not separate, fresh alcohol must be continuously introduced into the reaction mixture.

Removal of water usually involves the addition of entrainers, which form azeotropes with relatively low boiling points and high water content. Table 4 lists some frequently used entrainers. After condensation, the azeotrope separates into an aqueous phase and an organic phase, and the entrainer is recycled into the reaction mixture.

Table 4. Composition of common azeotropes

Entrainer	*bp*, °C		Water
	Entrainer	Azeotrope	content, wt %
Benzene	80.2	69.2	8.8
Toluene	110.7	84.1	13.5
Xylene	139	92	35.8
Cyclohexane	80.7	68.9	9.0
Tetrachloromethane	76.7	66.0	4.1

With ethanol, *n*-propanol, 2-propanol, allyl alcohol, and *tert*-butyl alcohol, a cosolvent such as benzene or toluene is often added to the condensate to achieve separation of the organic phase [28].

With high-boiling esters, e.g., esters of polyhydric alcohols, the water of reaction may be removed by means of steam or inert gases. Steam distillation can also be used to remove excess alcohol after the esterification is complete.

Esterification is usually carried out batchwise. In commercial production, continuous processes are preferred. Special procedures have been devised in some cases.

Methyl and ethyl esters of mono- and dicarboxylic acids are produced continuously by pumping the alcohol and the carboxylic acid melt (e.g., lauric acid or adipic acid) into a pipe from below and drawing off the ester at the top [29]. In a continuous process for the production of volatile esters, the mixture of acid, alcohol, and catalyst is introduced into the upper part of a distillation column and an excess of the alcohol is introduced into the bottom.

The column is heated so that the ester, water, and excess alcohol are distilled off [30].

Nonvolatile esters such as bis(2-ethylhexyl) phthalate [117-81-7] can be obtained continuously by the following procedure:

The mixture of acid or anhydride and alcohol together with sulfuric acid is introduced at the top of a plate column. An entrainer, e.g., benzene, is distilled through the column from below. The azeotrope of water and entrainer is drawn off the top and the ester accumulates in the sump.

Monochloroacetic acid esters of alcohols with three to five carbon atoms are obtained continuously by adding equimolar amounts of chloroacetic acid and alcohol to a chloroacetic acid melt and distilling off the ester and water.

Esterification can also be carried out continuously in the vapor phase by heating a mixture of acid, alcohol, ester, and catalyst to the reaction temperature. This process is called gas-phase esterification although one or more of the components may be present as liquids or solids.

Ion-exchange resins are especially suitable as catalysts in continuous processes. The reactants pass through or over the solid catalyst, and no separation or neutralization of the catalyst is necessary [21].

5.2. Alkylation of Metal Carboxylates

If metal carboxylates are treated with alky-lating agents, good yields of carboxylic acid esters result. A one-pot procedure is often used, in which the salt is first produced from the carboxylic acid and then directly alkylated to form the ester. Isolation of the salt is necessary only in the case of thermally unstable quaternary ammonium salts [31].

Alkali metal and alkaline-earth metal salts, as well as silver and copper salts, can be used. Dialkyl sulfates, aliphatic halides such as long-chain alkyl chlorides, and alicyclic halides are suitable alkylating agents. The method is used mainly in preparative chemistry.

5.3. Acylation with Acyl Halides

High yields of esters are obtained by acylation of alcohols with carboxylic acid halides:

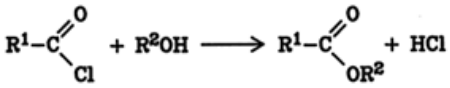

The acyl chloride is often produced in situ. The liberated hydrogen halide is generally bound to a base, e.g., pyridine or alkali hydroxide (Schotten – Baumann reaction).

This method is particularly important in preparative and analytical chemistry. A commercial application is the production of polycarbonates from phosgene and diols. Phenyl esters and esters of sterically hindered carboxylic acids can be produced from the corresponding acyl halides in the presence of Lewis acids, e.g., iron(III), aluminum, tin(IV), and titanium(IV) chlorides. Addition of phase-transfer agents such as tetrabutylammonium chloride has also proved effective [32]. An important application of this method is in the esterification of steroids.

5.4. Acylation with Carboxylic Anhydrides

Carboxylic anhydrides are suitable acylating agents for alcohols and phenols. With the exception of trifluoroacetic anhydride and mixed carboxylic anhydrides, they are less reactive than acyl halides.

The advantage of this method is that no water is formed; however, only one equivalent acyl per molecule of anhydride is available for esterification:

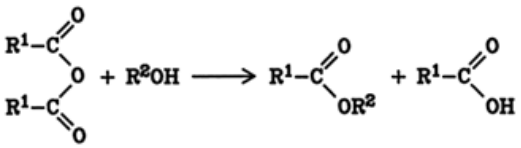

Therefore, the method is especially suitable for the esterification of rare or expensive alcohols with cheap anhydrides such as acetic anhydride.

Acylation can be accelerated by addition of protonic acids (sulfuric acid, hydrochloric acid, perchloric acid, *p*-toluenesulfonic acid) or Lewis acids (zinc chloride). A large-scale commercial process is the production of cellulose acetate from cellulose and acetic anhydride (⟶ Cellulose Esters – Cellulose Acetate).

Another important product is acetylsalicylic acid (aspirin), which is obtained from salicylic acid and acetic anhydride.

Base-catalyzed acylation with anhydrides is used to synthesize esters from sterically hindered or acid-sensitive alcohols [33], [34]. Common bases are triethylamine and pyridine.

Reaction with anhydrides is the basis of a quantitative analytical method for hydroxyl groups.

5.5. Transesterification

This method has already been described in Section Transesterification. Transesterifications take place under extremely mild conditions, which permits the reaction of components containing additional functional groups. Chiral centers can be preserved in many cases.

The most common transesterification catalysts are protonic acids (sulfuric acid, perchloric acid, *p*-toluenesulfonic acid, and molecular sieves). Alkaline catalysts such as alkoxides or anionic ion exchangers are especially effective in peptide chemistry. Potassium cyanide is recommended for unsaturated esters, which undergo cis – trans isomerization in the presence of acids. Thallium(I) salts are also employed. If titanium(IV) alkoxides or complexes of triphenylphosphine with trialkyltin alkoxides or with copper alkoxides are used, the alkoxide component need not be identical with the alcohol group that is exchanged.

5.6. Acylation with Ketenes

Ketenes react with alcohols or phenols to form carboxylic acid esters:

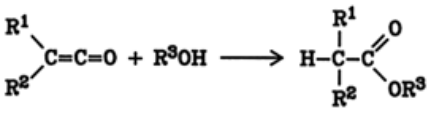

The addition is catalyzed by acids and bases (e.g., pyridine or sodium acetate). The use of ketenes is limited because they are difficult to handle and equally good results can be achieved with other acylating agents such as acyl halides and anhydrides.

The alcoholysis of diketene is an important method for the commercial production of β-oxocarboxylic acid esters (acetoacetates), especially ethyl acetoacetate [141-97-9] [35]:

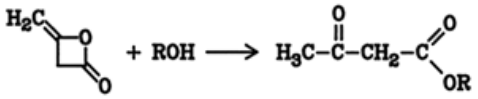

Acetoacetic acid anilides are produced in a similar way. These compounds are used in the synthesis of dyes, especially fast yellow dyes.

5.7. Carbonylation

The preparation of carboxylic acid esters by carbonylation of olefins in alcoholic solution (Reppe synthesis) is used on an industrial scale, e.g., in the production of methyl and ethyl propionate (⟶ Carbonylation) [36]:

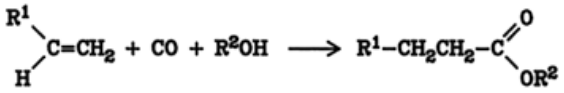

The linear ester is obtained in 85 % yield, with less than 10 % branched ester of the following structure:

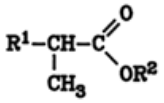

The reaction is carried out at a moderate temperature (ca. 100 °C) and a CO pressure of 20 – 25 MPa (200 – 250 bar). Iron, nickel, and cobalt salts, which are converted into the corresponding metal carbonyls during the reaction, and palladium or rhodium complexes are used as catalysts. A typical example is a bimetallic complex of bis(triphenylphosphine)palladium(II) chloride and tin(II) chloride [37]. Similarly, 1,3-dienes give γ, δ-unsaturated esters.

Alkyl, benzyl, vinyl, aryl, and allyl halides can also be carbonylated in the presence of suitable catalysts such as Mn – Fe – Ni powders. These reactions require only a low CO pressure [38].

Carbonylation of alcohols at a pressure up to 70 MPa (700 bar) in the absence of olefins gives formates. Thus, methyl formate is obtained by carbonylation of methanol at 70 °C and 2 – 20 MPa (20 – 200 bar) in the presence of sodium methoxide [39]:

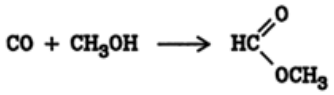

5.8. Condensation of Aldehydes

The condensation of two acetaldehyde molecules (Tishchenko reaction) is an important industrial process for the production of ethyl acetate:

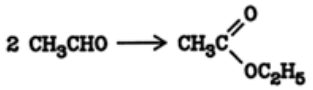

Aluminum ethoxide promoted with chloride ions serves as a catalyst. In practice, acetaldehyde is passed at 0 – 5 °C through a mixture of aluminum filings and traces of $AlCl_3$ in ethanol – ethyl acetate; the yield is 95 %.

5.9. Alcoholysis of Nitriles

Nitriles react directly with alcohols to produce the corresponding esters:

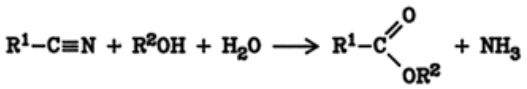

The ammonia that is formed must be removed from the equilibrium, preferably by using an excess of a strong mineral acid, e.g., sulfuric acid or *p*-toluenesulfonic acid. The acid also acts as a catalyst [40].

This method is used widely in industry for the production of esters from unsaturated nitriles such as acrylonitrile and methacrylonitrile. Alcoholysis of acrylonitrile is carried out with concentrated sulfuric acid. This leads initially to formation of the acrylamide sulfate, which then reacts with the alcohol (preferably methanol or ethanol) to produce the ester:

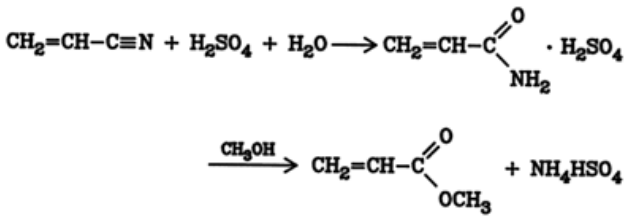

Selectivity for methyl acrylate is > 90 %. A disadvantage of this process is the formation of ammonium hydrogen sulfate, which cannot be used economically. An alternative process, which avoids formation of this byproduct, is the vapor-phase alcoholysis of acrylonitrile in the presence of boric oxide [41].

Methyl methacrylate is obtained by alcohol-ysis of the methacrylonitrile precursor acetone cyanohydrin. In this case, methacrylamide sulfate is first produced by using 98 % H_2SO_4 at 80 – 140 °C, which is then reacted with methanol at 80 °C to produce methyl methacrylate [42].

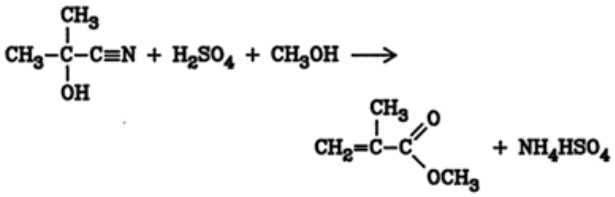

Selectivity based on acetone is 77 %. Other methods, which avoid the formation of NH_4HSO_4, have been developed on a pilot-plant scale.

5.10. Acylation of Olefins

Direct addition of olefins to carboxylic acids is used on a large scale for the production of vinyl acetate, the ester of the hypothetical vinyl alcohol (⟶ Vinyl Esters).

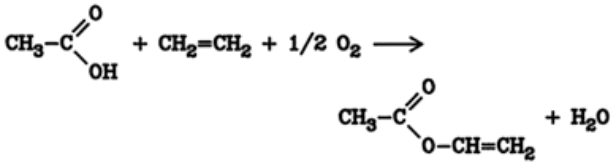

The presence of stoichiometric amounts of oxygen or air is necessary. Palladium or palladium salts are essential as catalysts; other group VIII metals are less effective.

The process can be carried out in the vapor phase or in the liquid phase. Only the vapor-phase process is used on an industrial scale. The reaction conditions are 175 – 200 °C and 0.5 – 1 MPa (5 – 10 bar); the catalyst consists of palladium metal and alkali acetate on a carrier. The carrier is necessary for activation and higher selectivity. During the reaction, the alkali acetate is slowly removed from the catalyst and has to be continuously replaced [43]. The vapor-phase process avoids the corrosion problems of the liquid-phase process and gives better yields of vinyl acetate (94 % based on ethylene or 98 – 99 % based on acetic acid).

The liquid-phase process is no longer used on an industrial scale. This process is closely related to the Wacker – Hoechst acetaldehyde process (⟶ Acetaldehyde – Direct Oxidation of Ethylene). The catalyst contains copper salts as a redox system. The advantages of the liquid-phase process are a better control of the highly exothermic reaction and the generation of acetaldehyde as a byproduct, which can be oxidized to acetic acid. In this way the necessary acetic acid can be entirely produced from ethylene. The disadvantages are corrosion problems and a lower yield of vinyl acetate, based on ethylene.

The olefin process has replaced the older process based on acetic acid and acetylene, because ethylene is cheaper and more readily available in larger quantities.

Allyl esters are produced from propene, carboxylic acids, and oxygen by using a vapor-phase process. Supported catalysts impregnated with palladium salts and bismuth or cadmium compounds are used [44]. Methallyl acetate is obtained from isobutene and acetic acid in the presence of palladium(II) chloride [45]:

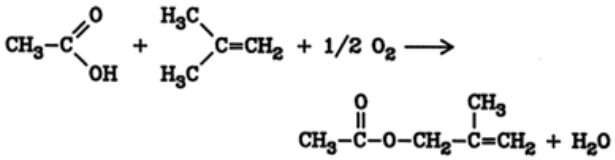

Glycol mono- and diesters can be produced from olefins and carboxylic acids by adding stoichiometric amounts of oxygen:

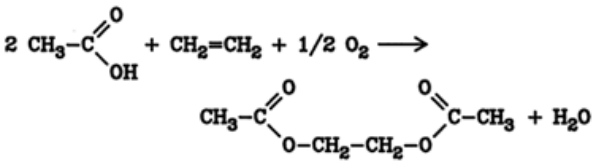

5.11. Ethoxylation of Carboxylic Acids

Addition of ethylene oxide to carboxylic acids generally proceeds smoothly and quickly. The low acidity of the carboxylic acids is sufficient to split the strained three-membered ring; no catalyst is required in the first addition step:

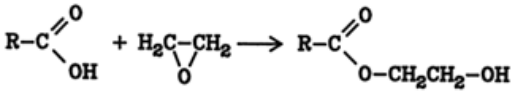

Further ethoxylation yields polyglycol esters; this reaction can be controlled or entirely suppressed by addition of bases such as alkali or amines.

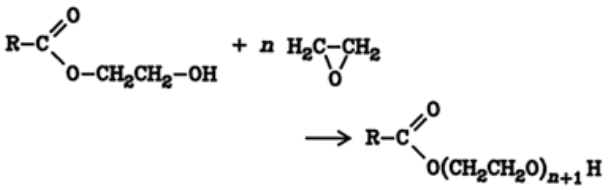

Ethoxylation reactions are widely used in industry. An example is the production of bis(2-hydroxyethyl) terephthalate [6422-86-2], by controlled ethoxylation of terephthalic acid:

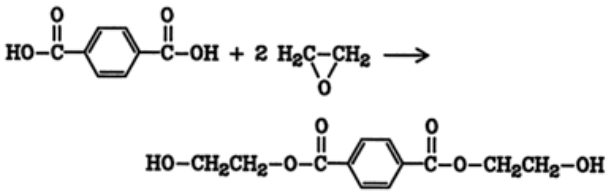

This ester is readily purified by crystallization from water, so that crude terephthalic acid can be used. The ethoxylation is carried out in the liquid phase at 2 – 3 MPa (20 – 30 bar) and 90 – 130 °C; no solvent is necessary. The presence of amines or quaternary alkylammonium salts is needed to avoid polyethoxylation. A considerable amount of poly(ethylene terephthalate) is produced by condensation of bis(2-hydroxyethyl) terephthalate.

Polyethoxylation of aliphatic carboxylic acids also is an industrial process. It is carried out under pressure at 120 – 220 °C in the presence of sodium hydroxide or sodium acetate. Addition of 10 – 30 mol of ethylene oxide per mole of acid causes fatty acids such as stearic acid to lose their hydrophobicity. The products have many applications, e.g., as surfactants. The degree of hydrophilicity depends on the number of glycol units. Ethoxylation of short-chain carboxylic acids yields hydrophilic lubricating oils.

5.12. Orthocarboxylic Acid Esters

Free orthocarboxylic acids are unstable. Their esters can be prepared by reaction of imido ethers with alcohols:

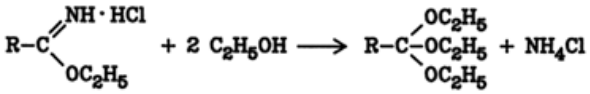

Orthoformic acid ethyl ester [122-51-0] is obtained from chloroform and sodium ethoxide:

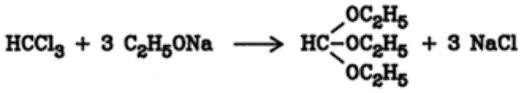

Orthoformates are relatively stable toward alkali; however, they are readily hydrolyzed by acids. In preparative chemistry, orthoformates are used to synthesize acetals and ketals. They also serve for the incorporation of methine groups into compounds with reactive methylene groups.

5.13. Lactones

Lactones are internal esters of hydroxycarboxylic acids. Of commercial interest are β-propiolactone [57-57-8] and γ-butyrolactone [96-48-0]. β-Propiolactone can be produced from ketene and formaldehyde in the presence of zinc chloride:

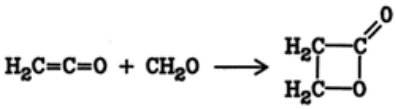

γ-Butyrolactone is produced mainly by dehydrogenation of 1,4-butanediol (⟶ Butyrolactone – Production):

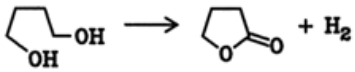

Lactones with five- and six-membered rings (γ- and δ-lactones) are the most stable. They readily form polyesters according to the following equation:

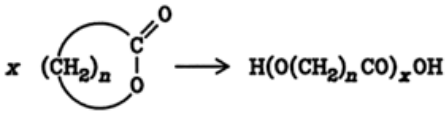

[Top of Page]

6. Environmental Protection

Most esters are synthesized by esterification of an acid with an alcohol, water being the only byproduct. Special environmental protection measures are usually unnecessary. Most esters possess low toxicity (Chap. Toxicology and Occupational Health). Because many of them are easily hydrolyzed on contact with water or moist air, the toxicological properties of the acid and alcohol component may need to be considered.

Waste Gases. Volatile esters with low carbon numbers are predominantly used as solvents and diluents for lacquers and plastics. Large-scale applications, e.g., in the auto and furniture industries, require excellent ventilation. The solvent vapors are suctioned off and burned in a flare or muffle furnace.

Waste Liquors. Short-chain esters and esters with hydrophilic groups are noticeably soluble in water; some results of biodegradability tests are listed in Table 5. The toxicity of some common esters for fishes and microorganisms is shown in Table 6.

Table 5. Biodegradation of organic esters

	Degree of degradation, *, %	BOD_5 **
Methyl acetate	> 95	ca. 500
Ethyl acetate	> 90	770
Vinyl acetate	> 90	810
Butyl acetate	> 95	1000

2-Methoxyethyl acetate	100	450
2-Butoxyethyl acetate	100	260
2-(2-Butoxyethoxy)ethyl acetate	100	380
Methyl acetoacetate	100	940
Ethyl acetoacetate	> 90	780
n-Butyl glycolate	93	570
Methyl crotonate	> 95	1050
Dimethyl acetylsuccinate	> 95	1100
Diethyl acetylsuccinate	> 95	1070
Dimethyl maleate	100	20
Monomethyl maleate	> 95	150
Diethyl maleate	> 90	200
Dibutyl maleate	99	630
Di(2-ethylhexyl)maleate	100	1450
Methyl 3-aminobenzoate	95	10
Methyl 4-hydroxybenzoate	100	1080
Methyl 4-hydroxyphenylacetate	98	320

* According to DIN 38 412, Part 25.
** Biochemical oxygen demand (milligrams of O_2 per liter of wastewater; sum over 5 d).

Table 6. Toxicity for fishes and bacteria [59]

	Fishes * LC_0, mg/L	Bacteria ** EC_0, mg/L
Methyl acetate		1500
Ethyl acetate	135 – 310	2000
Vinyl acetate	9	400
Butyl acetate	44 – 70	1200
2-Methoxyethyl acetate	150	1000
2-Butoxyethyl acetate	25	800
2-Ethoxyethyl acetate	90 – 116	
3-Methoxybutyl acetate		> 2500
2-(2-Butoxyethoxy)ethyl acetate	100	1000 – 2000
Methyl acetoacetate	250	500
Ethyl acetoacetate	173 – 412	3000
N-Butyl glycolate	50	100
Methyl crotonate		1250
Butyl acrylate	9	80
Methyl dimethylacrylate		200
Dimethyl acetylsuccinate	100	1500
Diethyl acetylsuccinate	50	2500
Dimethyl maleate	25	100
Monomethyl maleate	50	600
Diethyl maleate	25	800
Dibutyl maleate	12	40
Di(2-ethylhexyl) maleate	40	20
Methyl 3-aminobenzoate	100	1000
Methyl 4-hydroxybenzoate	100	1000
Methyl 4-hydroxyphenyl acetate		400

* DIN 38 412, Part 15.
** Measurement of CO_2 evolution in fermentation of bacterial sludge; EC_0 is the concentration at which CO_2 evolution is not suppressed.

[Top of Page]

7. Quality Specifications and Analysis

Table 7 shows some typical specifications of commercial products.

Table 7. Typical specifications of commercial esters

Color, max.[a]	n_D^{20}	d_{20}^{20}	Distillation range, °C (101.3 kPa)	Freezing point, °C	Flash point[b], °C	Autoignition temperature, °C	Lower flammable limit, vol % in air (*t*, °C)	Evaporation number (diethyl ether = 1.0)	Ignition group[d]

Methyl acetate	10	1.3601	0.930	56 – 58		– 13	455	3.1	2.2	IA
Ethyl acetate	10	1.371 – 1.373	0.889	76 – 78	– 84	– 4	460	2.1(38)	2.9	IA
Propyl acetate	15	1.3844	0.885	99.0 – 103.0	– 93	14	457	1.71(38)		IB
Isopropyl acetate	10	1.3772	0.872	85.0 – 90.0	– 62	6	479	1.76(38)		IB
Butyl acetate	10	1.394	0.880	124 – 127	– 74	25	422	1.38(38)	14	IC
Isobutyl acetate	10	1.390	0.870	115 – 118	– 99	18	478	1.27(93)	8	
2-Ethylhexyl acetate	15	1.4103	0.872	192.0 – 205.0	– 93	79	268	0.76(93)		IIIA
Vinyl acetate	5	1.369	0.93	72 – 74	– 93	– 8				IA
Ethylene glycol diacetate	15	1.4159	1.107	187.9 – 193.0	– 42	99	482	1.6(135)		IIIA
2-Methoxyethyl acetate	20	1.4602	1.003	142 – 148		44	400	1.7	34	II
2-Ethoxyethyl acetate	15	1.406	0.972	152 – 158	– 61	51	392	1.24(93)	60	II
2-Butoxyethyl acetate	15	1.4200	0.94	186.0 – 194.0	– 64	81	340	0.88(93)		IIIA
2-(2-Ethoxyethoxy) ethyl acetate	15	1.4230	1.011	214.0 – 221.0	– 25	107	360	0.98(135)		IIIB
2-(2-Butoxyethoxy) ethyl acetate	15	1.4265	0.978	244 – 250	– 32	116	290	0.76(135)	1200	IIIB
Glyceryl triacetate	5	1.4296	1.160	258.0	3.2	153	432	1.05(189)		IIIB
Glyceryl tripropionate		1.4314	1.09	285.0	–58	167	421	0.8(186)		IIIB
Isobutyl isobutyarate	15	1.3990	0.855	144.0 – 151.0	– 80	38	432	0.96(93)		II
Bis(2-ethylhexyl) adipate	20	1.4472	0.927	417	< – 70	206	377	0.38(242)		IIIB
Methyl acetoacetate	20	1.411	1.076	166 – 173		62	280	1.8	120	IIIA
Ethyl acetoacetate	20	1.420	1.030	175 – 184		65	350	1	135	IIIA
Butyl glycolate	10	1.425	1.023	183 – 197		75	404	0.6	460	IIIA
Methyl acrylate	10		0.9568	79.8 – 80.3		10				IA
Ethyl acrylate	10		0.9235	98.8 – 99.8		10				IA
Butyl acrylate	10		0.9009	145.7 – 148.0		49				II
2-Ethylhexyl acrylate	10		0.8862	214.8 – 218.0		91				IIIA
Dimethyl phthalate	5	1.513	1.192	284	– 1	157	490	0.94(181)		IIIB
Diethyl phthalate	10	1.4990	1.120	298	< – 50	161	457	0.75(187)		IIIB
Dibutyl phthalate	15	1.4905	1.048	340	– 35	190	404	0.47(236)		IIIB
Bis(2-ethylhexyl) phthalate	20 – 25	1.4836	0.9852	384	– 50	216	391	0.28(246)		IIIB
Dimethyl terephthalate	< 15			288	140.6	153	519			IIIB
Bis(2-ethylhexyl) terephthalate	15	1.4867	0.9825	383	– 48	238	399			IIIB
Bis(2-ethylhexyl) trimellitate	40	1.4832	0.989	600	– 38	263	410	0.26	2.5	IIIB

[a] Pt – Co scale.
[b] According to ASTM D 56–52, similar to DIN 51 755 and 51 758.
[c] ASTM D-2155, similar to DIN 51 794.
[d] According to NFPA (National Fire Prevention Association).

The most common analytical method is the determination of the saponification value, i.e., the number of milligrams of KOH necessary to hydrolyze 1.0 g of ester:

The ester (0.5 – 1 g) is refluxed for 30 min with 10 mL of 0.5 N alcoholic KOH. After cooling, the excess alkali is titrated with 0.5 N HCl using phenolphthalein as the indicator.

Before determination of the saponification value, determination of the free acid by titration with alkali in a water – alcohol mixture is often necessary.

The ester group can be readily identified by its characteristic absorption bands in the IR spectrum: the strong carbonyl absorption band at ca. 1740 cm^{-1} and two C–O stretching bands between 1300 and 1100 cm^{-1} [46].

Gas chromatography is suitable only for highly volatile esters. Esters with high boiling points must first be converted into low-boiling esters (usually methyl esters) either by transester-ification or by quantitative hydrolysis and sub-sequent reesterification. Quantitative conversion into the methyl ester can be achieved with diazomethane, methyl iodide – calcium oxide in dimethyl sulfoxide (DMSO), or with the boron tri-fluoride – methanol complex [47].

Unsaturated esters, e.g., natural fats and oils, are often sufficiently characterized by their iodine value. This is defined as the number of grams of iodine bound by 100 g of an unsaturated ester and is a measure of the number of double bonds in the molecule.

An overview of the different analytical methods is given in [48]. Physical properties such as boiling range, melting point, density, refractive index, residue or ash content, and color are also used to identify esters or mixtures of esters. Previous separation of the different components, (e.g., by high-vacuum distillation) is often unnecessary.

[Top of Page]

8. Storage and Transportation

All organic esters are flammable. The flash points, autoignition temperatures, and flammable limits of representative commercial esters are given in Table 7. The ignition group according to the NFPA (National Fire Prevention Association) corresponds approximately to the danger class of the German VbF (Verordnung über brennbare Flüssigkeiten of 27. Feb. 1980).

If esters are stored in tanks, a nitrogen blanket and a carbon filter is necessary. In some cases, the ventilation pipe must be linked to a flare.

Mild steel, aluminum, or other metallic materials can be used for storage and transportation. Plastics are unsuitable because the highly lipophilic esters may migrate into the container walls and cause them to soften or even dissolve. Esters with a high melting point, such as waxes, can also be transported in paper sacks or wooden barrels.

[Top of Page]

9. Uses and Economic Aspects

Solvents. Because of their highly lipophilic and hydrophobic nature and low polarity, esters are widely used as solvents, extractants, and diluents [49], [50]. Lower esters, especially the acetates of methanol, ethanol, propanols, and butanols, are good solvents for cellulose nitrate and other cellulose derivatives. Ethyl acetate is the most common technical solvent. Branched esters and esters of ethylene glycol also have good solvent properties. Considerable quantities of esters are used as diluents in paints, lacquers, and coatings. Methyl acetate is employed in fast-drying lacquers. Due to the development of low-solvent and solvent-free paints, e.g., for auto paints, the use of esters as diluents is decreasing. Table 8 gives production data for some acetates.

Table 8. Production of miscellaneous esters in 1985, t

	CAS registry number	United States	Western Europe	Japan
Methyl acetate	[79-20-9]		17 350 *	
Ethyl acetate	[141-78-6]	87 270	ca. 100 000 *	117 388
Propyl acetate	[109-60-4]	27 700		
n-Butyl acetate	[123-86-4]	81 430		29 942
Isobutyl acetate	[110-19-0]	34 880	ca. 50 000 *	
Vinyl acetate	[108-05-4]	960 200	> 200 000 *	402 930
2-Ethoxyethyl acetate	[111-15-9]	46 840		
Ethyl acetoacetate	[141-97-9]		19 450 *	
Acrylic acid esters, total			250 000	103 880
Butyl acrylate	[141-32-2]	192 560	88 000	33 000
Methyl acrylate	[96-33-3]	ca. 10 000	50 000	25 000
Ethyl acrylate	[140-88-5]	137 800	62 000	16 000
2-Ethylhexyl acrylate	[103-11-7]	36 020	50 000	29 000
Methyl methacrylate	[80-62-6]	390 060	ca. 200 000 *	236 690
Dibutyl maleate	[105-76-0]	1 540		
Dimethyl terephthalate	[120-61-6]	ca. 2×10^6	ca. 500 000 *	327 870

* Federal Republic of Germany.

Plasticizers. Large quantities of esters, especially phthalates, adipates, and fatty acid esters, are used as plasticizers [51], [52]. Production data for some important plasticizers are listed in Table 9. Isooctanoic acid esters and 2-ethylhexanoic acid esters of ethylene glycol are plasticizers for poly (vinyl choride).

Table 9. Production of plasticizers in 1985, t

	CAS registry number	United States	Western Europe	Japan
Phthalic acid esters, total		516 350	960 000	356 910
Dibutyl phthalate	[84-74-2]	9 880	22 000 *	16 900
Diethyl phthalate	[84-66-2]	7 800		
Diisodecyl phthalate	[26761-40-0]	66 500	70 000	10 510

			39 850 *	
Dimethyl phthalate	[131-11-3]	3 480		
Dioctyl phthalate	[117-81-7]	125 180	525 000	240 720
			222 320 *	
Ditridecyl phthalate	[119-06-2]	9 910		
Trimellitic acid esters, total		22 070		
Other cyclic esters, total		23 060		
Adipic acid esters, total		56 900	28 000	23 740
Bis(2-ethylhexyl) adipate	[103-23-1]	16 870		13 000
Complex linear polyesters, total		22 510		11 260
Ethoxylated esters, total		51 130	37 000	17 900
Oleic esters, total		5 240		
Butyl oleate	[142-77-8]	760		
Palmitic acid esters		1 550		
Sebacic acid esters		1 650	7 000	
Stearic acid esters, total		4 500		
Butyl stearate	[123-95-5]	3 500		

* Federal Republic of Germany.

Surfactants. Natural fats, oils, and waxes are used in considerable quantities as raw materials in the production of soap (alkali salts of fatty acids) and other detergents and surfactants (Table 10). Hydroxyethyl esters of long-chain fatty acids are sold as nonionic surfactants (nonionics), which are easily biodegradable. About 34 100 t of ethoxylated natural fats and oils were produced in the Federal Republic of Germany in 1985, as compared to only 1380 t in Japan. A defined number of ethylene oxide molecules can be added to the fatty acid by varying the process conditions; compounds with a specific water solubility — from moderate to excellent — can thus be obtained. Nonionics are important as low-foaming detergents and as emulsifiers. They are also used in the food, textile, cosmetics, and pharmaceutical industries. Because of their heat and salt resistance, they are suitable as lubricants in deep-well drilling.

Table 10. United States production of esters used as surfactants (1985)

	Production, t
Carboxylic acid esters, total	117 800
Anhydrosorbitol esters, total	15 320
Monostearate	8 080
Monooleate	3 080
Ethoxylated anhydrosorbitol esters, total	14 020
Monostearate	5 040
Monooleate	3 480
Glycerol esters, total	27 600
Monostearate	5 070
Monooleate	2 790
Glycerol esters of mixed acids	14 260
Diethylene glycol esters, total	8 900
Poly(ethylene glycol) esters, total	21 170
Monostearate	2 700
Monooleate	1 620
Monolaurate	2 290
Natural fats and oils, ethoxylated	15 000
Castor oil. ethoxylated	6 260

Polyesters. Higher esters derived from natural products, e.g., cellulose acetate and cellulose nitrate, are of considerable economic importance (⟶ Cellulose Esters). However, they are being gradually replaced by synthetic polyesters produced from monomers such as acrylates, terephthalates, and vinyl acetate (Table 8).

Vinyl acetate, with a worldwide production of 2×10^6 t/a, is the most important aliphatic ester (⟶ Vinyl Esters). Poly(vinyl acetate) is produced in large quantities for use in plastics, coatings, adhesives, and laminates (⟶ Poly(Vinyl Esters)). A considerable proportion is converted into poly(vinyl alcohol) [39]. Polymerization of acrylates such as methyl acrylate and methyl methacrylate yields homopolymers and mixed polymers used in paints, lacquers, and coatings. Polycondensation of dicarboxylic acids and diols gives polyester resins and plastics. Linear polyesters in particular are produced in large quantities.

Linear polyesters are used in the manufacture of synthetic fibers, packaging films, tapes, films for electric insulation, and a variety of plastics articles such as plastic bottles; they are also employed in hot-melt adhesives.

Polyesters are produced chiefly by transester-ification. The most important raw material is dimethyl terephthalate. This undergoes polycondensation with a diol (e.g., 1,2-ethanediol, 1,2-propanediol, 1,4-butanediol, 1,4-cyclohexanedimethanol), and methanol is released. However, direct esterification of terephthalic acid with ethylene oxide and polymerization of the resulting bis(2-hydroxyethyl) terephthalate is also used (see Section Ethoxylation of Carboxylic Acids). Copolymerization with unsaturated acids such as maleic acid yields polymers that can be cross-linked (thermosetting polymers).

Alkyd resins are an important group of polyesters used mainly in lacquers. The stability of the lacquer, e.g., its weatherability, can be improved by adding esters that are highly resistant to hydrolysis, such as pivalic acid esters.

Flavors and Fragrances. Certain esters with a pleasant odor are used in fragrances, flavors, cosmetics, and soaps. Production data for the most important flavor and perfume esters are given in Table 11. Although the quantities that are used are small, these esters are of great commercial importance because they are expensive [53]. Terpene esters as well as simple esters such as butyl and isobutyl acetate (fruity flavor) and benzyl acetate (jasmine odor) are widely used. These esters can be obtained from plant extracts, but many are produced by synthetic processes (⟶ Flavors and Fragrances).

Table 11. United States production of flavor and perfume esters (1985)

	CAS registry number	Production, t
Cedryl acetate	[77-54-3]	90
Geranyl acetate	[105-87-3]	60
Isoamyl acetate	[123-92-2]	51
Isoamyl butyrate	[106-27-4]	45
1,3-Nonanediol diacetate	[63270-14-4]	41
Citronellyl acetate	[150-84-5]	35
Benzyl propionate	[122-63-4]	20
2-Phenylethylphenyl acetate	[102-20-5]	15
Vetivenyl acetate	[117-98-6]	13
Isopentyl isovalerate	[659-70-1]	11
Allyl hexanoate	[123-68-2]	11
Necyl acetate	[141-12-8]	10
Citronellyl formate	[105-85-1]	9

Other Uses. Esters can be converted into various derivatives (see Chap. Chemical Properties) and are useful *intermediates* in preparative chemistry. Ethyl formate is used in the synthesis of vitamin B_1. Methyl formate is converted on a large scale into formamide and formic acid; about 30 000 t/a of formic acid are produced by this route.

Natural oils, fats, and waxes are used as *lubricants* for high-speed engines. High-boiling synthetic esters that experience little change in viscosity when exposed to extremes of temperature, either high or low, have been developed for use in auto and aircraft turboengines. Examples are bis(2-ethylhexyl) sebacate [122-62-3] and glycol or polyglycol esters of branched-chain carboxylic acids. These esters are also used as hydraulic fluids.

An example of the use of esters in *pharmaceuticals* is the common analgesic aspirin [50-78-2] (2-acetoxybenzoic acid. The active salicylic acid is released only after hydrolysis in the stomach. Benzocaine [94-09-7], the ethyl ester of 4-aminobenzoic acid, is a topical anesthetic. Phenyl salicylate [118-55-8], Salol, has antipyretic, antirheumatic, and antiparasitic properties.

[Top of Page]

10. Toxicology and Occupational Health

Hermann M. Bolt

With a few exceptions, esters have very low toxicity. Natural fats and oils are widely used in human nutrition. Table 12 lists toxicity data for commercially important esters [54-58].

Table 12. Toxicity data for various esters [54-58]

	LD_{50} (rat, oral), g/kg	Inhalation (rat, 4 – 6 h)	TLV ppm	TLV mg/m^3	Comments
Methyl formate	1.622[a]	10 000 ppm: lethal[c]	100	250	
Methyl acetate	3.7[a]	11 220 ppm: lethal[d]	200	610	irritation and burning of the eyes; narcotic; readily absorbed through the skin
Ethyl acetate	11.3	8 000 ppm: survival 16 000 ppm: lethal	400	1400	irritation of mucous membranes
Propyl acetate	>3.2	5 300 ppm: survival	200	840	irritation of the eyes
Isopropyl acetate	3.0 – 6.5	32 000 ppm: lethal	100	380	irritation of eyes, nose, and throat; defatting of skin
n-Butyl acetate	> 6.4	1 300 ppm: survival 14 000 ppm: lethal	200	470	irritation of eyes and respiratory tract
Isobutyl acetate	3.2 – 6.4	3 000 ppm: survival 23 000 ppm: lethal	150	700	irritation of eyes and respiratory tract
Pentyl acetate	16.6	5 000 ppm: survival	50	265	narcotic
2-Ethylhexyl acetate	3.2 – 5.9	> 1 100 ppm: survival			slightly irritating to the eyes
2-Methoxyethyl acetate			5	24[e]	
2-Ethoxyethyl acetate	5.1 – 11.9 4.94[b]	< 450 ppm: no effect	5	27[e]	slight skin irritation
Ethylene glycol mono- and diacetate	1.6 – 6.9				

Glyceryl triacetate	6.4 – 12.8	8 200 ppm : no effect
Glyceryl tripropionate	1.6 – 3.2	600 ppm: no effect
	6.4 – 12.8[b]	
Isobutyl isobutyrate	6.4 – 12.8	650 ppm: no effect
		5400 ppm: lethal

[a]Rabbit.
[b]Guinea pig.
[c]5 % (20 – 30 min) or 2.5 % (30 – 60 min) are also lethal to guinea pigs.
[d] Mouse.
[e] Skin notation (considerable cutaneous absorption).

In general, esters themselves are nontoxic. However, all esters are hydrolyzed more or less rapidly on contact with water; therefore, the toxicity of their hydrolysis products must be taken into account [54]. In the organism, esters may be split by esterases. In particular, this is important for acetate esters.

Most esters readily penetrate skin and mucous membranes. Considerable absorption through the skin can cause health problems. Exposed skin should be thoroughly washed with soap and plenty of water; clothing should be changed.

Inhalation of highly volatile esters in high concentrations can have a narcotic effect. Secondary effects include breathlessness and fits of choking. Ester vapors should not be allowed to contact the eye. Good ventilation of the work area and wearing of breathing apparatus are generally recommended [54].

Special toxicological consideration must be given to vinyl acetate [59] because of local carcinogenic properties found in animal experiments (rats and mice). Vinyl acetate is a colorless, volatile liquid with an odor threshold between 0.36 and 0.5 ppm. The reason for its carcinogenicity is that it is split at the site of entry into the organism into acetaldehyde and acetic acid. Acetaldehyde produces local genotoxic effects, while intracellularly formed acetic acid leads to acidification of the cell and subsequently to cell proliferation. Thus, the local induction of tumors (both on inhalation and oral administration) is explained. The local irritation threshold of vinyl acetate in humans has not been clearly defined, but two-year inhalation experiments in mice and rats have proven a ppm concentration to be a NOAEL (no observed adverse effect level); with respect to histopathological changes of nose and lungs it is 50 ppm. The American Conference of Governmental Industrial Hygienists had based its recommendation of a TLV (10 ppm) on this information, in order to avoid irritancy [55]. Vinyl acetate is of low acute toxicity after ingestion and dermal absorption, while inhalation of higher concentrations is acutely irritating and toxic. A NOAEL for sensory irritation of the respiratory tract in humans is reported to be on the order of 10 ppm. In rats, oral LD_{50} values in two studies were 3470 and 3500 mg/kg. A dermal LD_{50} value of 7440 mg/kg was determined from a range-finding study with rabbits. Vinyl acetate has proven to cause irritation or corrosion to the skin of rabbits depending on the duration of the skin contact. Exposed workers principally exhibit local irritant reactions of the skin, eyes, and respiratory tract [60]. Vinyl acetate was found to have genotoxic effects in vitro, e.g., chromosomal aberrations, micronuclei, and SCE, and gene mutations were observed in mammalian cell cultures.

[Top of Page]

References

General References

1. S. Patai (ed.): *The Chemistry of Carboxylic Acids and Esters*, Wiley-Interscience, New York 1969. J. D'Ans, E. Lax: *Taschenbuch für Chemiker und Physiker*, 3rd ed., Springer Verlag, Berlin – Heidelberg – New York 1970.

Specific References

2. J. D. Fletcher, O. C. Dermer, R. B. Fox (eds.): *Nomenclature of Organic Compounds*, Am. Chem. Soc., Washington D.C. 1974, pp. 137 – 145. IUPAC commission on nomenclature of organic chemistry: *Nomenclature of Organic Chemistry*, 4th ed., Pergamon Press, Oxford-New York 1979.
3. K. S. Markley (ed.): *Fatty Acids, Their Chemistry, Properties, Production and Uses*, 2nd ed., Interscience, New York 1960 /61.
4. D. Swern (ed.): *Bailey's Industrial Oil and Fat Products*, Wiley-Interscience, New York 1969, p. 937.
5. J. Smidt *et al.*, *Angew Chem.* **74** (1962) 93. Links
6. Unilever, DE 2 327 729, 1973 (J. J. Muller, T. J. Kock).
7. Rohm & Haas, DE 1 932 625, 1969 (W. R. Winslow).
8. Degussa, DE 1 568 376, 1966 (G. Morlock, H. Trageser).
9. A. M. Schwartz, J. W. Perry, J. Berch: *Surface-Active Agents and Detergents*, vol. **2**, Interscience, New York 1958, p. 104.
10. H. Adkins, *Org. React. (N.Y.)* **8** (1954) 1 – 27.
11. N. G. Gaylord: *Reduction with Complex Metal Hydrides*, New York 1956, pp. 391 – 543.
12. C. B. Hauser, B. E. Hudson, Jr., *Org. React. (N.Y.)* **1** (1947) 266, 302.
13. J. J. Bloomfield *et al.*, *Org. React. (N.Y.)* **23** (1976) 259 – 403.
14. *Houben-Weyl*, **E5**, 656 – 714.
15. H. R. Harrison, W. M. Heynes *et al.*, *Chem. Ind. (London)* 1968, 1568.
16. J. R. Rachele, *J. Org. Chem.* **28** (1963) 2898. Links
17. I. Roberts, H. C. Urey, *J. Am. Chem. Soc.* **60** (1938) 2391. Links
18. H. Zimmermann, J. Rudolf, *Angew. Chem.* **77** (1965) 65 – 74; Links *Angew. Chem. Int. Ed. Engl.* **4** (1965) 40.
19. R. Bader, A. D. Kontowicz, *J. Am. Chem. Soc.* **75** (1953) 5116.
20. Knapsack, DE 2 226 829, 1972 (E. Lohmar, F. X. Werber).
21. G. A. Olah *et al.*, *Synthesis* 1978, 929 ff. Links
22. H. J. Bergk, *Z. Chem.* **18** (1978) 22 ff.
23. G. Hallas, *J. Chem. Soc.* 1965, 5770. Links
24. DE 1 005 947, 1955 (W. L. Bears, F. X. Werber). B. F. Goodrich Comp., DE 1 103 335, 1956 (F. X. Werber, S. J. Averill). Reichhold Chemie AG, DE 1 173 473, 1962 (H. Dalibar). Reichhold Chemie AG, DE-AS 1 185 611, 1963 (H. Dalibar). ICI, DE-AS 1 292 140, 1966 (D. K. Steel). Scholven Chemie AG, DE-OS 1 593 503, 1966 (K. S. Schmitt, W. Flakus *et al.*). Mobil Oil Comp., DE-AS 1 618 639, 1967 (A. B. Gainor, L. McMakin) British Titan Products Comp., DE-OS 1 807 103, 1968 (R. H. Stanley, D. W. Brook, *et al.*). Hüls AG, DE-AS 2 243 240, 1972 (F. List, K. Wember) F. Brill, L. Ferry, I. T. Baker, *High Polym.* **27** (1972) 506.
25. Hüls, DE 914 006, 1951 (A. Coenen, F. Broich).
26. E. Taschner, C. Wasielewski, *Liebigs Ann. Chem.* **640** (1961) 136 – 139.
27. R. C. Parish, C. M. Stock, *J. Org. Chem.* **30** (1965) 927. Links
28. S. Natelson, S. Gottfried, *Org. Synth., Coll. Vol.* **3** (1955) 381.
29. Ruhrchemie, DE 1 171 897, 1959 (H. Feichtinger, H. Noerke).

30. BASF, DE 878 348, 1942 (H. Dierichs, E. Braun, *et al.*).
31. W. William, B. Halper, *Synthesis* 1974, 727.
32. G. Schwabe, G. Westphal, H. G. Hennig, *Z. Chem.* **20** (1980) 184. S. Szeja, *Synthesis* 1980, 402. Links
33. J. Cason, *Org. Synth., Coll. Vol.* **3** (1955) 169.
34. W. Steglich, DE 1 958 954, 1969 (W. Steglich, G. Höfle).
35. Hoechst, DE 1 956 558, 1969 (E. Fischer, O. Mautz).
36. J. Falbe: *New Syntheses with Carbon Monoxide*, Springer Verlag, Heidelberg 1980.
37. F. Knifton, *J. Org. Chem.* **41** (1976) 2885. Links
38. C. Cassar, G. P. Chiuroli, F. Guerrieri, *Synthesis* 1973, 509. Links
39. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie*, 2nd ed., Verlag Chemie, Weinheim – New York 1978, pp. 42 – 43, pp. 216 – 221.
40. F. L. James, W. H. Bryan, *J. Org. Chem.* **23** (1958) 1225. Links
41. Mitsui Toatsu, JP 4 725 120, 1971.
42. Röhm & Haas, DE 1 468 939, 1965 (G. Schröder, H. Fink).
43. W. Schwerdtel, *Chem. Ing. Tech.* **40** (1968) 781. Hoechst AG, DE 1 191 366, 1961 (W. Riemenschneider, T. Quadflieg). Hoechst AG, DE-AS 1 618 391, 1967 (L. Hörnig, H. Fernholz, et al.) DE-AS 1 643 999, 1967 (H. Fernholz, H.-J. Schmidt). DE-AS 1 667 140, 1967 (H. Erpenbach, H. Glaser, et al.). Knapsack AG, DE-OS 1 668 352, 1967 (K. Sennewald, W. Vogt, et al.). Asahi Kasei Kogyo, DE 1 768 643, 1968;DE-AS 1 793 362, 1968 (N. Kominami, H. Nakajima). Bayer AG, DE 1 793 519, 1968 (W. Krönig, G. Scharfe). Knapsack AG, DE-AS 1 808 610, 1968 (K. Sennewald, W. Vogt). Hoechst AG, DE-OS 2 315 037, 1973 (H. Fernholz, H. Krekeler, et al.). Celanese Corp., DE-OS 2 361 098, 1973 (G. M. Severs, Jr.). General Electric Co., DE-OS 2 503 748, 1975; 2 503 926, 1975 (W. E. Smith, R. J. Gerhart). H. Krekeler, H. Schmitz, *Chem. Ing. Tech.* **40** (1968) 785. Hoechst AG, DE-AS 1 668 063, 1967 (G. Roscher, H. Schmitz).
44. Hoechst AG, DE 1 903 954, 1969 (H. Fernholz, H.-J. Schmidt). Air Products & Chemicals, DE-AS 1 768 770, 1968 (J. G. Schukys). General Electric Co., DE-OS 2 504 230, 1975; 2 504 231, 1975 (W. E. Smith, R. J. Gerhart).
45. ICI, DE 1 468 086, 1965 (D. Clark, D. Wright).
46. L. J. Bellamy, *The Infrared Spectra of Complex Molecules*, 3rd ed., J. Wiley & Sons, New York 1975. L. J. Bellamy, *Advances in Infrared Group Frequencies*, Methuen & Co., London 1968, pp. 166 – 168.
47. L. D. Metcalfe, A. A. Schmitz, *Anal. Chem.* **33** (1961) 363. Links
48. T. S. Ma in S. Patai (ed.): *The Chemistry of Carboxylic Acids and Esters*, Wiley-Interscience, New York 1969, pp. 871 – 921.
49. E. H. Davies, T. H. Durrans: *Solvents*, 2nd ed., Chapman & Hall, London 1971, pp. 145 – 165.
50. C. Marsden, S. Mann (eds.): *Solvents Guide*, 2nd ed., Interscience Publishers, New York 1963.
51. D. N. Buttrey: *Plasticizers*, 2nd American ed., Franklin Publishing Co., Palisades, New York 1960.
52. I. Mellan: *Industrial Plasticizers*, Mac Millan, New York 1963.
53. D. L. J. Opdyke: "Monographs on Fragrance," *Food Cosmet. Toxicol.* **16** (1978) Suppl. 1, 839 – 841; **17** (1979) Suppl., 735 – 741, 841 – 843; **19** (1981) Suppl. 2, 237 – 245.
54. N. I. Sax: *Dangerous Properties of Industrial Materials*, 4th ed., Reinhold Publishing Corp., New York 1975.
55. *Threshold Limit Values for Chemical Substances in the Work Environment*, ACGIH, Cincinnati 2004.
56. M. Windholz (ed.): *The Merck Index*, 10th ed., Merck & Co, Rahway, N.J. 1983.
57. E. E. Sandmeyer, C. J. Kirwin in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, 3rd ed., vol. **2 A**, Wiley-Interscience, New York 1981, pp. 2259 ff.
58. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **12** (1974) 719.
59. J. G. Hengstler, M. S. Bogdanffy, H. M. Bolt, *Annu. Rev. Pharmacol. Toxicol.* **43** (2003) 485–520 Links
60. Deutsche Forschungsgemeinschaft: Vinylacetat. In: *Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten (Herausg.: H. Greim)*. VCi Publishers, Weinheim, Germany (2002).

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Ethanol

Standard Article

Naim Kosaric[1], Zdravko Duvnjak[2], Adalbert Farkas[3], Hermann Sahm[4], Stephanie Bringer-Meyer[5], Otto Goebel[6], Dieter Mayer[7]

[1]The University of Western Ontario, London, Ontario N6A 5B9, Canada
[2]University of Ottawa, Ottawa, Ontario K1N 9B4, Canada
[3]Consulting Chemist, Delray Beach, Florida 33483, United States
[4]Institut für Biotechnologie, Forschungszentrum Jülich GmbH, Jülich, Federal Republic of Germany
[5]Institut für Biotechnologie, Forschungszentrum Jülich GmbH, Jülich, Federal Republic of Germany
[6]CORA Engineering, Chur, Switzerland
[7]Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a09_587
Article Online Posting Date: June 15, 2001

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Physical Properties
3. Chemical Properties
4. Synthesis
4.1. Direct Catalytic Hydration of Ethylene
4.1.1. Chemistry
4.1.2. Catalysts
4.1.3. Production Process
4.2. Indirect Hydration of Ethylene
4.2.1. Chemistry
4.2.2. Production Process
4.3. Other Methods
4.3.1. Homologation of Methanol
4.3.2. Carbonylation of Methanol and Methyl Acetate
4.3.3. Conversion of Synthesis Gas
4.3.3.1. Synthesis by Heterogeneous Catalysis
4.3.3.2. Synthesis by Homogeneous Catalysis
5. Fermentation
5.1. Production by Yeast
5.1.1. Nutrients
5.1.2. Fermentation Pathways
5.1.3. Fermentation Variables
5.1.4. Direct Fermentation
5.2. Production by Bacteria
5.2.1. Ethanol Tolerance
5.2.2. Fermentation of Glucose by *Z. mobilis*
5.2.3. Fermentation of Other Sugars by Recombinant *Z. mobilis* Strains
5.2.4. Ethanol Production by Recombinant Enterobacteria
5.2.5. Comparison of Recombinant Bacterial Strains
5.3. Fermentation Modes

5.3.1. Batch Processes
5.3.2. Continuous Processes
5.3.3. Other Processes
5.4. Raw Materials and Processes
5.4.1. Readily Fermentable Carbohydrates
5.4.2. Starch
5.4.3. Lignocellulosic Materials
5.4.4. Waste Materials and Residues
6. Recovery and Purification
6.1. Distillation
6.1.1. Distillation of Azeotropic Ethanol
6.1.2. Dehydration by Azeotropic Distillation
6.1.3. Motor Fuel Ethanol
6.1.4. Reduction of Energy Costs
6.2. Nondistillative Methods
6.2.1. Solvent Extraction
6.2.2. Carbon Dioxide Extraction
6.2.3. Adsorptive Dehydration
6.2.4. Membrane Technology
6.2.5. Yarn-Filled Column
6.3. Storage and Transportation
7. Comparison of Process Economics for Synthetic and Fermentation Ethanol
7.1. Summary of Cost Analysis for Synthetic and Fermentation Ethanol
7.2. Production Costs of Synthetic Ethanol
7.3. Production Costs of Fermentation Ethanol
8. Analysis
9. Uses
10. Economic Aspects
10.1. Worldwide Production of Synthetic and Fermentation Ethanol
10.2. Major Producers of Fermentation Ethanol from Regenerable Resources
11. Toxicology

[Top of Page]

1. Introduction

Naim Kosaric and Zdravko Duvnjak

Ethanol [64-17-5] or ethyl alcohol (CH_3CH_2OH), M_r 46.7, is also referred to as alcohol spirit, spirit of wine, grain alcohol, absolute alcohol, and ethyl hydrate. Depending on its water content, preparation, and final use, several ethanol products exist on the market. The 99 % alcohol (often referred to as absolute alcohol) is used extensively for tinctures and pharmaceutical preparations, as a solvent and preservative, as an antiseptic, and in perfume. Ethanol is an important functional component of alcoholic beverages, which are produced by fermentation of fermentable carbohydrates. The fermentation broth itself may constitute (after processing and aging) a beverage, e.g., in the case of beer or wine, or the alcohol can be concentrated from the broth to produce high-alcohol-containing spirits. If the alcohol is used for purposes other than as a beverage, it is denatured by the addition of substances such as methanol, pyridine, formaldehyde, or sublimate. The denatured alcohol is then used by industry and commerce, principally as a solvent, as a raw material for manufacturing chemicals, or as a fuel.

Chemically synthesized ethanol is usually derived from petroleum sources by the hydration of ethylene (see Chap. Synthesis) and has found wide application as industrial alcohol. Various routes for the production of ethanol are depicted in Figure 1.

Figure 1. Industrial sources of ethanol

History. Ethanol can be considered to be one of the oldest human food products. Fermented beer was consumed in Babylon, and wine was produced as early as 3000 B.C. The distillation process probably orginated in the 10 – 14th centuries. At this time, the “spiritual” effect of ethanol was recognized; thus, the name *spiritus* was given to alcoholic drinks. The first wine distillates were used for medicinal rather than beverage purposes.

Until the 17th century, alcoholic fermentation was considered to be a spoilage process whereby the yeast produced was eliminated. The nature of fermentation was initially clarified in the 19th century with the discovery of the microscope, which showed that yeast cells were living organisms. However, recognition of the fact that these living organisms are responsible for the fermentation process took about 150 years.

In the 19th century, two theories were developed to explain the mechanism of fermentation: the “mechanistic” and the “vital” processes. LOUIS PASTEUR (1822 – 1895) promoted the vital theory, which stated that living organisms were responsible for the conversion of sugar to alcohol. The mechanistic theory was supported by JUSTUS FREIHERR VON LIEBIG (1803 – 1873) and by FRIEDRICH WÖHLER (1800 – 1882). A convincing proof of the mechanistic mechanism, by which physicochemical processes lead to chemical conversion of sugar to ethanol, came from EDWARD BUCHNER (1860 – 1917), who demonstrated that alcoholic fermentation is related not to the living cell but to a substance in the fermentation broth, which was later identified as an enzyme. As is now known, enzymes are ultimately responsible for the complex conversion of carbohydrates to ethanol.

[Top of Page]

2. Physical Properties

Naim Kosaric and Zdravko Duvnjak

Ethanol in its pure form (absolute alcohol) is a colorless liquid. It is miscible in all proportions with water and also with ether, acetone, benzene, and some other organic solvents. Anhydrous alcohol is hygroscopic; at a water uptake of 0.3 – 0.4 %, a certain stability does occur. Various physical properties of anhydrous ethanol are as follows:

bp	78.39 °C
fp	–114.15 °C
n^{20}	1.36048
d_4^{15}	0.79356
d_4^{20}	0.78942
d_{15}^{15}	0.79425
d_{20}^{20}	0.79044
Surface tension at 20 °C	22.03 mN/m
C_p (16 – 21 °C)	2.415 J g^{-1} K^{-1}
Heat of fusion	4.64 kJ/mol
Heat of evaporation	
At 70 °C	855.66 kJ/kg
At 80 °C	900.83 kJ/kg
At 100 °C	799.05 kJ/kg
Heat of combustion	
(at constant volume)	1370.82 kJ/mol
Thermal conductivity at 20 °C	18 μW m^{-1} K^{-1}
Dynamic viscosity η	1.19 mPa · s
Volumetric expansion coefficient	1.1×10^{-3} K^{-1} *
Heat of mixing 30 wt % ethanol	
and 70 wt % water at 17.33 °C	39.32 J/g
Flash point (in a closed vessel)	13 °C
Autoignition temperature	425 °C
Explosion limit (amount of	
ethanol in a mixture with air)	
Lower, 3.5 vol %	67 g/m^3
Upper, 15 vol %	290 g/m^3
Maximum explosion pressure	736 kN/m^2
Specific conductivity	135×10^{-11} Ω^{-1} cm^{-1}
Dilution number at 20 °C	
(diethyl ether = 1)	8.2
Diffusion coefficient for vapors	
at 20 °C and 101.3 kPa	0.12 cm^{-1}
Heating value	
Upper	29 895 kJ/kg

Lower	29 964 kJ/kg

* In practice, the volume increase for 1000 L is taken as 1 L/K.

Tables 1 and 2 show the freezing and flash points of ethanol – water mixtures. Table 3 shows the vapor – liquid equilibria of ethanol – water mixtures; the azeotropic mixture contains 95.57 wt % ethanol and 4.43 wt % water. Therefore, the highest concentration of ethanol, obtained by distillation from an ethanol – water mixture, is 95.57 wt %. Azeotropic distillation, with the help of a tertiary solvent (e.g., benzene), must be introduced to produce absolute (anhydrous) ethanol.

Table 1. Freezing points of ethanol – water mixtures

Ethanol, vol %	*fp*, °C	Ethanol, vol %	*fp*, °C
50	–36.9	30	–15.3
45	–28.1	25	–11.3
40	–24	20	–7.6
38	–22.3	15	–5.1
32	–16.8	10	–3

Table 2. Flash points of ethanol – water mixtures

Ethanol, wt %	Flash point, °C	Ethanol, wt %	Flash point, °C
100	13	50	24.5
94.5	16	40	26.5
80	19.5	30	30
70	21.5	10	46
60	22.6	5.5	56

Table 3. Vapor – liquid equilibria and boiling points of the ethanol – water system at 101.3 kPa

Ethanol in liquid, wt %	Ethanol in vapor, wt %	*bp*, °C
0	0.00	100.00
1	6.5	98.90
3	20.5	96.75
5	38.0	94.95
10	52.0	91.45
15	59.5	88.95
20	64.8	87.15
25	68.6	85.75
30	71.4	84.65
35	73.3	83.75
40	74.7	83.10
45	75.9	82.45
50	77.1	81.90
55	78.2	81.45
60	79.4	81.00

65	80.7	80.60
70	82.2	80.20
75	83.9	79.80
80	85.9	79.35
85	88.3	78.95
90	91.3	78.50
95	95.04	78.15
97	96.86	78.20
99	98.93	78.25
100	100.00	78.30

When ethanol is mixed with water, the volume contracts slightly, as shown in Table 4. For example, when 52 volumes of absolute ethanol and 48 volumes of water are mixed, 96.3 volumes of diluted ethanol result.

Table 4. Volume contraction for 1 mol of ethanol – water mixture at 15°C

x *	C_2H_5OH d_4^{15}	ΔV **
1.00000	0.79354	0.0000
0.97483	0.79666	–0.1063
0.92674	0.80270	–0.2903
0.88143	0.80851	–0.4429
0.81817	0.81688	–0.6250
0.74155	0.82756	–0.8028
0.65639	0.84032	–0.9495
0.56708	0.85510	–1.0540
0.47734	0.87186	–1.1123
0.35078	0.89980	–1.1116
0.22077	0.93482	–0.9695
0.15550	0.95355	–0.7864
0.07909	0.97311	–0.4030
0.05987	0.97790	–0.2869
0.03726	0.98435	–0.1586
0.01605	0.99196	–0.0582
0.00792	0.99542	–0.0267
0.00000	0.99913	0.0000

* Mole fraction of ethanol.
** Change in volume.

[Top of Page]

3. Chemical Properties

Naim Kosaric and Zdravko Duvnjak

The chemical properties of ethanol are dominated by the functional –OH group, which can undergo many industrially important chemical reactions, e.g., dehydration, halogenation, ester formation, and oxidation (⟶ Alcohols, Aliphatic).

Because ethanol can be produced efficiently not only by chemical synthesis from petroleum and coal-based feedstock, but also by fermentation of abundantly available organic materials, its commercial role as a raw material for various chemicals is of increasing importance.

Conversion of ethanol to “alkochemicals” is an entirely new approach to producing familiar petrochemicals (Fig. 2). Some of these routes are already being used industrially in large alcohol-producing countries.

Figure 2. Chemicals produced from ethanol

[Top of Page]

4. Synthesis

Adalbert Farkas

The formation of ethanol from ethylene [74-85-1] by reaction with sulfuric acid was demonstrated over a hundred years ago [1], [2]. Economic production of synthetic ethanol on an industrial scale was first accomplished in 1930 by the Carbide and Chemical Corporation (now Union Carbide). The process used involved absorption of ethylene in sulfuric acid and subsequent hydrolytic cleavage of the resulting ethyl sulfates. This indirect process has been gradually superseded by direct catalytic hydration of ethylene, introduced by the Shell Chemical Company in 1948 [3].

More recent routes for ethanol synthesis depend on coal-based starting materials such as methanol [67-56-1] or synthesis gas (a mixture of carbon monoxide and hydrogen). They include homologation and carbonylation processes and the direct conversion of synthesis gas to ethanol.

4.1. Direct Catalytic Hydration of Ethylene

4.1.1. *Chemistry*

In the temperature range at which industrial catalysts operate, the hydration of ethylene to ethanol is a reversible reaction controlled by the equilibrium

$$C_2H_4\,(g) + H_2O\,(g) \rightleftharpoons C_2H_5OH\,(g)\ \Delta H = -43.4\,\mathrm{kJ} \qquad (1)$$

Calculation of the equilibrium is described in [4-6]. The kinetics of Reaction (1) have been studied on a phosphoric acid – silica gel catalyst [7], [8] and on a blue tungsten oxide – silica gel catalyst (energy of activation = 125 kJ/mol) [9].

A nomogram for the correlation between ethylene conversion in this reaction and several process variables (pressure, temperature, and the water – ethylene molar ratio) has been developed [10]. Under usual reactor conditions (approximately equimolar ethylene – water feed, 250 –300 °C, 5 – 8 MPa), the equilibrium conversion of ethylene is 7 – 22 %.

A lower temperature favors higher ethylene conversion, but Reaction (1) is then accompanied by reversible formation of diethyl ether [60-29-7] according to Reaction (2):

$$C_2H_5OH + C_2H_4 \rightleftharpoons C_2H_5OC_2H_5 \qquad (2)$$

At high pressures, ethylene polymerizes to yield butenes and higher olefins, which are converted to higher alcohols by hydration.

The mechanism of ethylene hydration is assumed to consist of four steps [11].

1. Formation of a π-complex by the addition of a proton to the ethylene molecule:

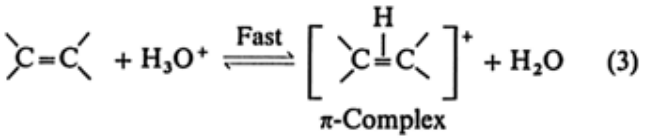

(3)

2. Conversion of the π-complex to a carbocation:

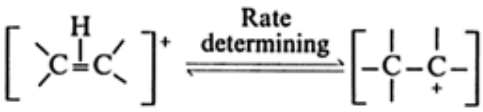

(4)

3. Addition of water to the carbocation:

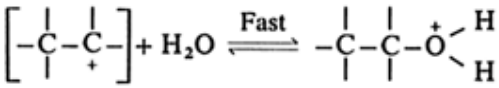

(5)

4. Removal of a proton from the protonated ethanol:

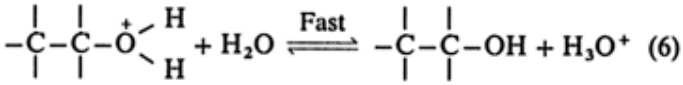

4.1.2. *Catalysts*

The technical and patent literature describes numerous catalysts for ethylene hydration (see ⟶ Heterogeneous Catalysis and Solid Catalysts), [12]. Most of them are acidic because the reaction involves carbocations. However, only phosphoric acid catalysts supported by diatomaceous earth (kieselguhr, Celite) [13-15], montmorillonite [16], bentonite [17], silica gel [18-21], or opoka (Volga sandstone) [22] are of industrial importance.

Improved mechanical properties are claimed for catalysts on a diatomaceous earth support with added silica gel [23] or bentonite [24]. Catalyst performance depends on the porosity of the support [21] which can be increased by incorporation of combustible material [24].

Under operating conditions, the support must be resistant to phosphoric acid but able to retain it in large amounts; it must also have adequate mechanical strength and a sufficiently long life and, ultimately, must provide an active, selective catalyst for the hydration [13]. The support is impregnated with 50 – 77 % phosphoric acid but is first treated with hydrochloric or sulfuric acid to remove most of its iron and aluminum. Otherwise, these materials are entrained on contact with phosphoric acid and tend to clog the heat exchangers. Heating the supported catalyst to 700 – 1100 °C has been recommended because even though this treatment impairs catalyst activity and selectivity, it ensures that the phosphoric acid binds firmly to the support [25-27].

The following compounds have also been suggested as catalysts for ethylene hydration: acidic oxides (such as tungsten trioxide) on silica [11], silicotungstic acid [28], [29], silicoborotungstic acid [30], or phosphoric acid on silica – alumina [31]; high-silica HZSM-5, dealuminated Y [32], HVSN-5 [33], or H-form [34] zeolites, as well as zeolites containing metal ions [35]; and salts such as boron phosphate [36-40], zirconium phosphate [41], other phosphates [40], cobalt sulfate [42], or magnesium sulfate [43].

4.1.3. *Production Process*

Ethanol production based on the catalytic hydration of ethylene is shown schematically in Figure 3. Ethylene and deionized water (molar ratio range 1 : 0.3 – 1 : 0.8) are heated to 250 –300 °C at 6 – 8 MPa by passage through a heat exchanger and a superheater. Since hydration is exothermic, the gaseous reaction products leave the reactor at a temperature 10 – 20 °C higher than when they entered and thus are used as a source of heat in the heat exchanger. Some phosphoric acid is entrained by the gas stream and is neutralized by injecting a dilute solution of sodium hydroxide.

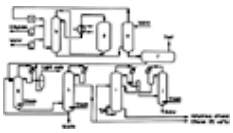

Figure 3. Synthesis of ethanol by direct hydration

a) Circulation compressor; b) Heat exchanger; c) Superheater; d) Reactor; e) Washer; f) Crude ethanol tank; g) Extractive distillation column; h) Concentration column; i) Dehydration; j) Concentration column

After condensation and separation of the liquid reaction products, the gas is freed from residual ethanol by water washing and then recompressed and recycled to the reactor. The phosphoric acid lost by entrainment and evaporation is replaced continuously or periodically by spraying it on the catalyst bed. Crude product collects in the sump of the washer and contains ca. 10 – 25 wt % ethanol. It is decompressed to recover the dissolved ethylene, which is recycled. The ethanol is then purified, ideally by extractive distillation followed by rectification, to obtain a 95 vol % ethanol – water azeotrope. The azeotrope can be dehydrated by azeotropic distillation to give anhydrous ethanol. For further details of distillation see Chapter Recovery and Purification.

Prior to distillation, the crude ethanol is often catalytically hydrogenated to convert acetaldehyde [75-07-0] and higher aldehyde impurities to the corresponding alcohols. Acetaldehyde is a highly undesirable impurity because it tends to form crotonaldehyde [4170-30-3].

Approximately 2 % diethyl ether is formed as a byproduct; it can be easily purified to a marketable product from the light ends fraction (⟶ Ethers, Aliphatic – Diethyl Ether) or recycled to the reactor where it is quantitatively converted to ethanol.

Ethylene is available in high purity (>99.9 vol %); therefore, only small amounts need to be vented even if an ethylene level of 85 % must be maintained in the reactor stream.

Cylindrical steel containers with a diameter up to 4 m and an internal volume of up to 150 m^3 are used as reactors. They are lined with copper or carbon bricks for protection against attack by phosphoric acid. The heat exchangers and the connecting piping exposed to phosphoric acid are made of, or lined with, copper or copper alloys. The remaining synthesis and distillation equipment is made of steel. Waste gas does not present any problems. Vented ethylene is either returned to the ethylene plant or burned. The synthesis step does not produce any liquid waste. Wastewater from distillation does not contain any organic compounds, and its low phosphate content (0.3 kg of Na_2HPO_4 per cubic meter) permits ready disposal.

Optimal reaction conditions depend on the correlation of activity and selectivity of the catalyst with temperature, pressure, residence time, and molar ratio of water to ethylene. These relationships have been studied extensively [44-48]. A rate equation for the decrease in catalyst activity with time has been derived and used to develop a mathematical model for computing optimum reaction parameters [49], [50].

4.2. Indirect Hydration of Ethylene

The indirect hydration, esterification – hydrolysis, or sulfuric acid process is based on the absorption of a large volume of ethylene in concentrated sulfuric acid; ethanol and some diethyl ether are formed when the sulfuric acid solution is diluted with water.

4.2.1. *Chemistry*

Ethylene is absorbed in two steps:

$$C_2H_4 + H_2SO_4 \rightleftharpoons C_2H_5OSO_3H \quad \Delta H = -60\,\mathrm{kJ} \tag{7}$$

$$C_2H_4 + C_2H_5OSO_3H \rightleftharpoons C_2H_5OSO_2OC_2H_5 \tag{8}$$

Kinetic studies have shown that this reaction also proceeds via the mechanism in Reactions (3) – (6) [51] but is complicated by the fact that only part of the ethylene is physically dissolved in sulfuric acid.

Hydrolysis involves three steps:

$$C_2H_5OSO_3H + H_2O \rightleftharpoons C_2H_5OH + H_2SO_4 \tag{9}$$

$$C_2H_5OSO_3C_2H_5 + H_2O \rightleftharpoons C_2H_5OH + C_2H_5OSO_3H \tag{10}$$

$$C_2H_5OSO_3C_2H_5 + C_2H_5OH \rightleftharpoons C_2H_5OC_2H_5 + C_2H_5OSO_3H \tag{11}$$

The products are ethanol and 5 – 10 % diethyl ether. The ether yield can be controlled by varying the reaction conditions, especially the ratio of ethylene to sulfuric acid, and the hydrolysis conditions (⟶ Ethers, Aliphatic – Diethyl Ether).

Transesterification. The ethyl sulfates formed in Reactions (7) and (8) can also be transesterified with acetic acid [64-19-7] at 104 °C to yield ethyl acetate [141-78-6], which is then recovered by distillation and hydrolyzed to ethanol and acetic acid [52]. This process avoids problems associated with reconcentration of sulfuric acid.

4.2.2. *Production Process*

Figure 4 shows the flow scheme for the production of ethanol by the sulfuric acid process. The feed gas must contain a minimum of 35 vol % ethylene and only inert gases such as methane and ethane. Higher homologues of ethylene cause resin formation and, therefore, must be removed from the feed.

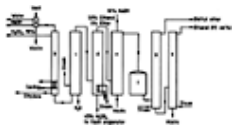

Figure 4. Synthesis of ethanol by the sulfuric acid method

a) Wash column for residual gas; b) Absorber; c) Hydrolyzer; d) Stripping column; e) Wash column for crude ethanol; f) Crude product tank; g) Diethyl ether distillation column; h) Ethanol distillation column

The absorption of ethylene increases almost linearly with pressure. Operating pressure is generally 1 – 3.5 MPa, the higher pressure being used when the ethylene level in the feed is low. Each mole of sulfuric acid absorbs up to 1.4 mol of ethylene. The absorption is carried out with 94 – 98 wt % sulfuric acid in wash towers at 65 – 85 °C. Temperatures above 90 °C lead to formation of resins and oils and, therefore, must be avoided. The resulting liquid is agitated for several hours under pressure to complete the reaction.

Hydrolysis is usually performed in two stages. First, diethyl sulfate is hydrolyzed at a low temperature (70 °C) in the presence of less than the equivalent amount of water. Then more water is added gradually, and the temperature is raised to ca. 100 ° C. Hydrolysis is complete within 1 h, and the sulfuric acid is then diluted to 40 – 55 wt %.

The ethanol formed is recovered, together with the diethyl ether byproduct, in a stripping column. The product mixture is then washed with sodium hydroxide solution to neutralize any acid, and the diethyl ether is removed in the ether distillation column. The ethanol is purified by distillation and concentrated to give a 95 vol % ethanol – water azeotrope. Anhydrous ethanol can be obtained as described in Chapter Recovery and Purification.

Reconcentration of the dilute sulfuric acid is the most expensive part of this process. It is encumbered by high energy consumption, corrosion problems, and oxidation of the organic compounds remaining in the sulfuric acid. Although tar formed in the absorption is separated as completely as possible, the sulfuric acid loss amounts to more than 5 % based on the ethanol formed.

To avoid these disadvantages, an operating pressure below 0.5 MPa has been recommended [53]; less than 0.5 mol of ethylene is absorbed per mole of sulfuric acid, which contains up to 7 % silver sulfate to accelerate the reaction. The sulfuric acid can then be reconcentrated by simple flash evaporation.

The following construction materials are used in this process: steel for absorption and distillation; lead and acid-resistant masonry for hydrolysis; silumin (an aluminum alloy), tantalum, cast iron, and lead for reconcentration of sulfuric acid.

4.3. Other Methods

Efforts have been made to develop methods for synthesizing ethanol that are not based on raw materials derived from petroleum and are likely to be of industrial importance in the future. Methanol [67-56-1] and synthesis gas, which are available from coal, can be used as starting materials.

4.3.1. *Homologation of Methanol*

The homologation (hydrocarbonylation) of alcohols discovered in 1949 [54], [55] allows the chain length of an alcohol to be increased by one $–CH_2–$ group by reaction with synthesis gas ($CO + H_2$). The reaction occurs in the presence of a cobalt carbonyl catalyst at elevated temperature and pressure:

$$ROH + CO + 2H_2 \xrightarrow{Cat.} RCH_2OH + H_2O \qquad (12)$$

When homologation was originally applied to methanol, the ethanol yield was relatively low and the reaction product contained higher alcohols, formate and acetate esters, and other oxygenated products. The complexity of the mixture formed was a result of further homologation of ethanol and other alcohols and of additional carbonylation reactions.

The selectivity for ethanol has been improved to up to 89 % by reaction in tetrahydrofuran [109-99-9] solution [56] or by addition of iodine or iodide together with organic phosphines [57-61]. High ethanol selectivities have also been claimed for combinations of cobalt catalysts with copper, iron, nickel, ruthenium, or rhodium compounds [62-70]. An alumina-supported rhodium – iron catalyst has been described as well [71]. Results of these studies are summarized in Table 5.

Table 5. Homologation of methanol to ethanol [a], [b]

Catalyst	Additive	Ethanol selectivity, %	Reference
$Co(OAc)_3$	tetrahydrofuran	72	[56]
$Bu_3PCo(CO)_3$	I_2	65	[57]
$Co(OAc)_3$	I_2, Bu_3P	70	[58]
$Co(OAc)_3$	I_2, Ph_3P	80	[59]
$Co(OAc)_3$	I_2, $[Ph_2P(CH_2)_3]_2$	65	[60]
CoI_3		89	[61]
$Co_2(CO)_8$, $RuCl_3$	I_2	79	[62]
$Bu_3PCo(CO)_3$, $Ru(acac)_3$	I_2	72	[63]
$Co(acac)_3$, $Ru(acac)_3$	I_2, $(Ph_2PCH_2)_2$	82 [c]	[64]
CoI_3, FeI_3, $RuCl_3$	I_2, MeOEt, MeOAc	86	[65]
CoI_3, RhI_3	*N*-methyl-2-pyrrolidone	62	[66]
$Co_{20}B_{10}RhH_6$		78	[67]
$Co_2(CO)_8$, $NiCl_2$	triphenylamine	82	[68]
CoS, Cu	Bu_3P, *N*-methyl-2-pyrrolidone	96	[69]
$Co(acac)_3$, $W(CO)_6$	HI, $(Ph_2PCH_2)_2$	89 [c]	[70]
$Co(acac)_3$, $Fe(acac)_3$	HI, $(Ph_2PCH_2)_2$	83 [d]	[70]
Rh – Fe/ γ – alumina		37 – 87 [e]	[71]

[a] Except where otherwise specified, homologation was performed at 150 – 250 °C and 20 – 35 MPa.
[b] Abbreviations: OAc = acetate; acac = acetylacetonate; Bu = butyl; Et = ethyl; Me = methyl; Ph = phenyl.
[c] Ethanol content of product.
[d] Selectivity for ethanol and acetaldehyde.
[e] Yield of ethanol and methyl acetate; conditions for homologation as described in [71].

4.3.2. *Carbonylation of Methanol and Methyl Acetate*

The first step of another method for converting methanol to ethanol is carbonylation of methanol to acetic acid [64-19-7] [72].

$$CH_3OH + CO \rightleftharpoons CH_3COOH \qquad (13)$$

This reaction is the basis of a commercial process developed by Monsanto for synthesis of acetic acid [73] by using a rhodium carbonyl – iodine complex as catalyst [74] (⟶ Acetic Acid – Production).

Acetic acid can be hydrogenated directly to ethanol, but this requires expensive highpressure equipment, and the process is corrosive [72]. According to a Davy McKee patent, carbonylation is carried out in a methanolic solution containing rhodium trichloride [10049-07-7], methyl iodide [74-88-4], and acetic acid at 175 °C and 7 MPa [75]. The acetic acid formed is esterified with ethanol in the presence of sulfuric acid to yield ethyl acetate [141-78-6]. Ethyl acetate is then converted to ethanol by hydrogenolysis in a tubular reactor at 200 °C over a copper – zinc oxide catalyst. Thus, the overall reaction sequence is:

$$CH_3OH + CO \rightarrow CH_3COOH \qquad (13)$$

$$CH_3COOH + C_2H_5OH \rightarrow CH_3COOC_2H_5 + H_2O \quad (14)$$

$$CH_3COOC_2H_5 + 2\,H_2 \rightarrow 2\,C_2H_5OH \quad (15)$$

The net reaction is:

$$CH_3OH + CO + 2\,H_2 \rightarrow C_2H_5OH + H_2O \quad (16)$$

HALCON has shown that carbonyls of non-noble metals, such as molybdenum, tungsten, and chromium, catalyze the carbonylation step [76], [77] and has developed an improved esterification procedure that is used commercially [72]. In a modification of this process, methyl acetate [79-20-9] rather than methanol is carbonylated [76], [78]:

$$CH_3COOCH_3 + CO \rightarrow (CH_3CO)_2O \quad (17)$$

Reaction of the resulting acetic anhydride [108-24-7] with ethanol and methanol then yields ethyl and methyl acetates:

$$(CH_3CO)_2O + CH_3OH + C_2H_5OH \rightarrow CH_3COOCH_3 + CH_3COOC_2H_5 + H_2O \quad (18)$$

After separation of the acetates, the ethyl acetate is hydrogenolyzed to ethanol (Reaction 15) and the methyl acetate is carbonylated (Reaction 17). The net reaction is given by Reaction (16).

Methyl acetate is carbonylated on a commercial scale in Eastman Kodak's acetic anhydride plant. Various techniques are available for both the esterification (Reaction 18) and the hydrogenolysis (Reaction 15) steps [72].

4.3.3. *Conversion of Synthesis Gas*

Soon after methanol had been synthesized from carbon monoxide and hydrogen (⟶ Methanol), and the Fischer – Tropsch synthesis had been discovered (⟶ Coal Liquefaction), a search started for related routes to ethanol and higher alcohols, based on synthesis gas and the use of heterogeneous [79] or homogeneous [80] catalysis.

Highly specific methods for producing ethanol from synthesis gas are not yet available, but some promising leads have been found.

4.3.3.1. *Synthesis by Heterogeneous Catalysis*

Attempts to improve the conversion of synthesis gas to ethanol have been based on modifying industrial catalysts containing oxides of zinc, chromium, or copper that are normally used for the synthesis of methanol. Early work showed that the incorporation of alkali [81] and cobalt [82] improved yields of ethanol and higher alcohols.

The Institut Français du Pétrole [83] has developed complex catalysts containing copper, cobalt, chromium, potassium, lanthanum, and other elements for conversion of synthesis gas to alcohols to be blended with gasoline. The product contains 50 – 70 % methanol, 16 – 23 % ethanol, 8 – 14 % propanol, and some higher alcohols. Institut patents report on two catalysts that can convert synthesis gas to an alcohol mixture containing 39 % ethanol, 24 % methanol, and 36 % propanols + butanols at 250 °C and 6 MPa [84], [85].

Use of a Cu – Ni – Ba catalyst is reported to give a product containing 46 % ethanol, 38 % methanol and 9 % propanol [86]. A Cu – Zn – Ni – Na catalyst showed a 55 % selectivity for alcohols, but the alcohol fraction contained only 29 % ethanol [87]. Iron and potassium have also been added to copper – zinc catalysts normally used for methanol synthesis; their effects on ethanol synthesis and the reaction mechanisms involved have been discussed [88], [89].

On certain supports, platinum-group metal catalysts prepared from carbonyl clusters convert synthesis gas to methanol – ethanol mixtures at 200 – 240 °C and atmospheric pressure [90]. Thus, a supported rhodium – lanthanum trioxide catalyst gives ethanol as the predominant product, with a total alcohol selectivity up to 80 %. More recent studies using supported rhodium catalysts promoted with cerium or europium report a similarly high selectivity for ethanol [91], [92].

The only catalysts derived from rhodium carbonyl clusters that produce ethanol as the major product are those which contain added cobalt [93]. A number of formulations containing rhodium and a variety of cocatalysts and promoters are claimed by the patent literature to be selective catalysts for the synthesis of ethanol, e.g., [94-96], acetaldehyde, and acetic acid, e.g., [97]. According to a Hoechst patent, a composition of cobalt, rhenium, gold, and barium oxides on a silica support produces ethanol from synthesis gas with 60 % selectivity [98].

4.3.3.2. *Synthesis by Homogeneous Catalysis*

The homogeneous catalytic conversion of synthesis gas to ethanol and ethylene glycol [107-21-1] has been reviewed with particular reference to the reaction mechanism, the nature of the active catalyst species, and their interaction with solvents, cocatalysts, and other additives [80].

Ruthenium is the only group 8 metal to show a high selectivity for the conversion of synthesis gas to ethanol when used in conjunction with organic phosphorus compounds. Synthesis gas can be converted to ethanol with a selectivity >50 % by using an iodine-promoted ruthenium carbonyl catalyst dissolved in tripropylphosphine oxide [99], [100]. Lower ethanol selectivities (20 – 30 %) have been reported for ruthenium catalysts in a melt of tetrabutylphosphonium bromide [101-103] or alkyltriphenylphosphonium bromides [104].

A combination of ruthenium and cobalt carbonyls dissolved in toluene is reported to convert synthesis gas to a mixture of ethanol, methanol, and acetic acid [105]. Ethanol is also reported to be the major product in a conversion catalyzed by rhodium carbonyl, potassium chloroplatinate, and stannous chloride [106].

[Top of Page]

5. Fermentation

Naim Kosaric, Zdravko Duvnjak, Hermann Sahm, Stephanie Bringer-Meyer

5.1. Production by Yeast

Yeasts are unicellular, uninucleate fungi that can reproduce by budding, fission, or both. They have been used for centuries to make alcoholic beverages.

Yeasts are the most commonly used organisms in the industrial production of ethanol. Some widely used, high-productivity strains are *Saccharomyces cerevisiae, S. uvarum* (formerly *S. carlsbergensis*), and *Candida utilis. Saccharomyces anamensis* and *Schizosaccharomyces pombe* are also used in some instances. *Kluyveromyces* species, which ferment lactose, are good producers of ethanol from whey.

Ethanol production by yeast is characterized by high selectivity, low accumulation of byproducts, high ethanol yield, high fermentation rate, good tolerance toward both increased ethanol and substrate concentrations, and lower pH value. Viability and genetic stability of yeast cells under process conditions and at high temperature are also desirable. Although finding a strain that has all these characteristics is difficult, some yeast strains can fulfill them to a great extent [107], [108].

5.1.1. *Nutrients*

Yeasts require the following for growth: carbon, nitrogen, phosphorus, sulfur, oxygen, hydrogen, minor quantities of potassium, magnesium, calcium, trace minerals, and some organic growth factors (vitamins, nucleic acids, and amino acids).

Various carbon compounds can serve as a carbon source for yeast (Table 6). Nitrogen can be supplied as ammonia, ammonium salts, urea, or amino acids. Orthophosphate salts and phosphoric acid are good sources of phosphorus. Sulfur, potassium, magnesium, and calcium can be supplied in the form of their salts. Table 7 shows the amount of these elements and trace minerals required by yeasts.

Table 6. Ability of *Saccharomyces* and *Kluyveromyces* species to ferment sugars [109]

Carbon number of basic subunit	Type of basic subunit	Sugar	Basic unit	Yeast [e]		
				S. cerevisiae	*S. uvarum (carlsbergensis)*	*K. fragilis*
		glucose	glucose	+	+	+
		maltose	glucose	+	+	–
		maltotitrirose	glucose	+	+	–
	aldose	cellobiose	glucose	–	–	–
	sugars	trehalose	glucose	+/–	+/–	–
		galactose	galactose	+	+	+
		mannose	mannose	+	+	+
		lactose	glucose, galactose	–	–	+
C_6 sugars		melibiose	glucose, galactose	–	+	
	ketose	fructose	fructose	+	+	+
	sugars	sorbose	sorbose	–	–	–
	aldoses	sucrose	glucose, fructose	+	+	+
	and ketoses	raffinose	glucose, fructose, galactose	+/–	+	+/–
	deoxy	rhamnose	6-deoxymannose	–	–	–
	sugars	deoxyribose	2-deoxyribose	+/–	+/–	+/–
C_5 sugars	aldose	arabinose	arabinose	–	–	–
	sugars	xylose	xylose	–	–	–

[*] + Indicates good ability to ferment sugars; – cannot ferment sugars; +/– low ability to ferment sugars.

Table 7. Growth optimum and calculated inlet concentrations of inorganic ions required for optimum growth of

yeast in a CSTR[a] [110]

Ion	Growth optimum, mmol/L	Cellular concentration, mmol/100 g dry mass	CSTR inlet concentration for a cell concentration of 10 g/L, mmol/L
B^+	0.0004 – 0.001	0.5×10^{-3}	0.0004 – 0.001
Ca^{2+}	0.5 – 5.0 [b]	0 – 1.5	0.5 – 5.0
Co^{2+}	0.0001 – 0.001[c]	$(0.03 – 1)\times10^{-3}$	0.0001 – 0.001
Cu^{2+}	0.001 – 0.01[c]	0.008 – 0.20	0.002 – 0.03
Fe^{2+}	0.001 – 0.01	0.036 – 0.18	0.005 – 0.02
K^+	2.0 – 10.0 [b]	8 – 56	3 – 15
Mg^{2+}	2.0 – 4.0	4 – 17	2 – 6
Mn^{2+}	0.002 – 0.01 [b]	0.007 – 0.055	0.003 – 0.01
Mo^{2+}	0.001 – 0.01	$(0.04 – 0.08)\times10^{-3}$	0.001 – 0.01
Ni^+	0.001 – 0.05	$(0.03 – 2)\times10^{-3}$	0.001 – 0.05
Zn^{2+}	0.005 – 0.015[b, c]	0.08 – 0.3	0.006 – 0.04
Cl^-	$\simeq$1	11 – 140	2 – 15
I^-	$\simeq$0.001	$0.04 – 1\times10^{-3}$	0.001
SO_4^{2-}	$\simeq$1	0.6 – 15.0	2 – 10
$H_2PO_4^-$	2 – 4	40 – 65	6 – 10
NH_4^+	$\simeq$1	$\simeq$700	$\simeq$100

[a] CSTR = continuous stirred tank reactor.
[b] Optimum concentration may be higher under conditions where specific inhibitory ions are present or any ion is present at an inhibitory level.
[c] These ions are most likely to be deficient in a complex industrial medium.

Yeast extract is a good source of trace minerals and organic growth factors. Many raw materials used for industrial ethanol production supply all the nutrients necessary for yeast growth; additional nutrient supplements are required in some cases.

5.1.2. *Fermentation Pathways*

Hexoses. Yeasts are able to metabolize various carbon compounds. Metabolic pathways differ under aerobic and under anaerobic conditions. Conversion of glucose by *Saccharomyces cerevisiae* under both conditions is shown in Figure 5 [111].

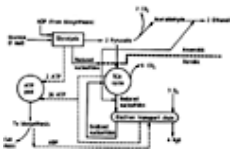

Figure 5. Simplified chart of anaerobic and aerobic catabolism of *Saccharomyces cerevisiae* [111]

ADP = adenosine diphosphate; ATP = adenosine triphosphate; TCA = tricarboxylic acid (citric acid).

Under *anaerobic conditions,* the yeast produces ethanol, according to the Gay – Lussac equation:

$$C_6H_{12}O_6 \rightarrow 2\,C_2H_5OH + 2\,CO_2$$

Therefore, from each gram of glucose consumed, 0.51 g of ethanol can be produced. However, some of the carbon source is used for biomass generation so that the actual ethanol yield is ca. 90 – 95 % theoretical. In addition to ethanol, carbon dioxide, and biomass, 2 mol of adenosine triphosphate (ATP) are produced per mole of glucose. A lower than theoretical yield of ethanol is also caused by the formation of small amounts of byproducts, such as glycerol and succinate, at the expense of the carbon source. If the carbon source is not used for the production of biomass and byproducts, the ethanol yield based on sugar would increase by 1.6 and 2.7 %, respectively [112]. In ethanol production, some higher alcohols (fusel oils) are formed partly from the carbon source and partly from deamination and subsequent conversion of certain amino acids.

Under *aerobic conditions* glucose is converted to carbon dioxide and water according to the equation:

$$C_6H_{12}O_6 + 6\,O_2 \rightarrow 6\,CO_2 + 6\,H_2O$$

In addition, biomass and energy are also generated.

The sequence of intermediates and reactions in the transformation of glucose (or a polysaccharide) to ethanol and CO_2 is

shown in Figure 6. The Embden – Meyerhof – Parnas or glycolytic pathway represents the major route of glucose catabolism in most cells that convert hexoses to pyruvate. This pathway, which must provide all the intermediates and most of the energy required for cell growth, is also used in microbial fermentation of glucose to ethanol, lactate, glycerol, glycols, and a variety of other products.

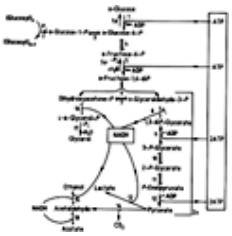

Figure 6. Embden – Meyerhof – Parnas scheme of glycolysis [113]

Glycolysis:	glucose + 2 P_i + 2 ADP	⟶ 2 lactate + 2 ATP (classical mammalian muscle or brain)
	glucose + P_i	⟶ α-glycerol phosphate + pyruvate (insect flight muscle, striated muscle)
Fermentation:	glucose + 2 P_i + 2 ADP	⟶ 2 ethanol + $2CO_2$ + 2 ATP + 2 H_2O (first form)
	glucose + HSO_3^-	⟶ glycerol + acetaldehyde · HSO_3 + CO_2(second form, no net ATP)
	2 glucose	⟶ 2 glycerol + ethanol + acetate + 2 CO_2 (third form, no net ATP)
	glucose + (P_i)	⟶ α-glycerol phosphate + acetaldehyde + CO_2 ⇓ glycerol + P_i (third form, no net ATP)

Enzymes, characteristic inhibitors, and standard free energy changes for each of the reactions are listed in Table 8. The reactions do not require oxygen and provide the cellular energy supply under completely anaerobic conditions.

Table 8. Reactions and thermodynamics of glycolysis and alcoholic fermentation [113]

Reaction number	Equation [a]	Name of enzyme	Characteristic inhibitor	$\Delta G^{0\prime}$ [b], kJ/mol
1	Glucose + ATP ⟶ glucose 6-*P* + ADP	hexokinase		–14.32
	cosubstrate: Mg^{2+}	glucokinase		
1a	Glucose 6-*P* + H_2O ⟶ glucose + P_i	glucose 6-phosphatase	glucose, orinase	–16.83
	cosubstrate: Mg^{2+}	nonspecific phosphatases		
2	Glycogen + *n* P_i ⇌ *n* glucose 1-*P*	α-1,4-glucan phosphorylase		3.06
3	Glucose 1-*P* ⇌ glucose 6-*P*	phosphoglucomutase	F, organophosphorus inhibitors	–7.29
	cosubstrate: glucose 1,6-di*P*			
4	Glucose 6-*P* ⇌ fructose 6-*P*	phosphoglucose (glucose phosphate) isomerase	2-deoxyglucose 6-phosphate	2.09
5	Fructose 6-*P* + ATP ⟶ fructose-1,6-di*P* + ADP + H^+	phosphofructokinase	ATP, citrate	–14.24
	cosubstrate: Mg^{2+}, activator: (ADP, AMP)K^+			
5 a	Fructose 1,6-di*P* + H_2O —Mg^{2+} ⟶ fructose 6-*P* + P_i	fructose diphosphatase	AMP, fructose 1,6-diphosphate,	–16.75
		nonspecific phosphatases	Zn^{2+}, Fe^{2+}	
	cosubstrate: Mg^{2+},			

6	Fructose 1,6-di*P* ⇌ dihydroxyacetone *P* + glyceraldehyde 3-P	(fructose phosphate) aldolase	chelating agents (microbial enzymes only)	23.99
7	Dihydroxyacetone *P* ⇌ glyceraldehyde 3-*P*	triose phosphate isomerase		7.66
8	2×(Glyceraldehyde 3-*P* + P_i + NAD^+ ⇌ 1,3-diphosphoglycerate + NADH + H^+	glyceraldehyde phosphate dehydrogenase; triose phosphate dehydrogenase	ICH_2COR D-threose 2,4-diphosphate	2×(6.28)
9	2×(1,3-Diphosphoglycerate + ADP + H^+ —Mg^{2+} ⟶ 3 phosphoglycerate + ATP)	phosphoglycerate kinase		2×(–28.39)
10	2×(3-Phosphoglycerate ⇌ 2-phosphoglycerate cosubstrate: glycerate 2,3-di*P*	phosphoglyceromutase		2×(4.43)
11	2×(2-Phosphoglycerate ⇌ phosphoenolpyruvate) cosubstrate: Mg^{2+} or Mn^{2+}	enolase (phosphopyruvate hydratase)	Ca^{2+} F^- plus P_i	2×(1.84)
12	2×(Phosphoenolpyruvate + ADP + H^+ ⟶ pyruvate + ATP) cosubstrate: Mg^{2+} activator: K^+(Rb^+, Cs^+)	pyruvate kinase	Ca^{2+} vs. Mg^{2+} Na^+ vs. K^+	2×(–23.95)
13	2×(Pyruvate + NADH + H^+ ⇌ lactate + NAD^+)	lactate dehydrogenase	oxamate	2×(–25.12)
14	2×(Pyruvate + H^+ ⟶ acetaldehyde + CO_2)	pyruvate (de) carboxylase		2×(–19.76)
15	2×(Acetaldehyde + NADH + H^+ ⇌ ethanol + NAD^+)	alcohol dehydrogenase	(HSO_3^-)	2×(–21.56)
Sums:	(glucose)$_n$ + H_2O ⟶ 2 lactate + 2 H^+ + (glucose)$_{n-1}$		–219.40	
	(glucose)$_n$ + 3 P_i + 3 ADP ⟶ 2 lactate + 3 ATP + (glucose)$_{n-1}$		glycolysis (muscle)	–114.38 [c]
	glucose ⟶ 2 lactate + 2 H^+		–198.45	
	glucose + 2 P_i + 2 ADP ⟶ 2 lactate + 2 ATP		glycolysis or lactate fermentation	–124.52[c]
	glucose ⟶ 2 ethanol + CO_2		–234.88	
	glucose + 2 P_i + 2 ADP ⟶ 2 ethanol + 2 CO_2 + 2 ATP		alcoholic fermentation	–156.92 [c]

[a] P = phosphate.
[b] $\Delta G^{0\prime}$ values refer to pH 7.0 with all other reactants including H_2O at unit activity; the free energy of formation of glucose in aqueous solution equals 910.88 kJ/mol, its ΔG^0 of combustion to CO_2 + H_2O is 2872 kJ/mol, and $\Delta G^{0\prime}$ for (glucose)$_n$ + H_2O ⟶ (glucose)$_{n-1}$ equals – 21.06 kJ/mol.
[c] From this table.

As shown in Table 8 and Figure 6, conversion of D-glucose to ethanol starts with the formation of D-glucose 6-phosphate (Robinson ester) (Reaction 1). This reaction is catalyzed by hexokinase (ATP : D-hexose 6-phosphotransferase). The enzyme uses ATP as a phosphate donor and requires a divalent cation (usually Mg^{2+}, but Mn^{2+} or frequently Ca^{2+} and Co^{2+} can replace the cofactor). This enzyme is relatively nonspecific and can catalyze the phosphorylation of a variety of hexoses of appropriate configuration including D-fructose, D-mannose, 2-deoxy-D-glucose, and D-glucosamine. The equilibrium of the reaction lies so far in the direction of product formation as to render the reaction virtually irreversible thermodynamically ($K_{eq} \simeq 6.5\times10^3$ at pH 7.4 and 25 °C) and of no practical importance in gluconeogenesis.

The conversion of starch and other starchlike glucose-containing polysaccharides is catalyzed by α-1,4-glucan phosphorylases (α-1,4-glucose : orthophosphate glycosyltransferases), which are widely distributed in a variety of organisms (Fig. 6, Reaction 2). This phosphorolysis commences at the free, nonreducing end of an amylose chain and removes one glucose unit at a time, yielding a total of (n + 1) glucose 1-phosphate molecules until the reducing end is reached. If the substrate is amylopectin, phosphorolysis continues only until the branch points are reached and the product is a limit dextrin. The 1,6-bond at the branch may be cleaved by an amylo-1,6-glucosidase, yielding free glucose, after which phosphorylase is able to act again until the next branch is reached, and so on.

Phosphoglucomutase (α-D-glucose 1,6-diphosphate : α-D-glucose 1-phosphate phosphotransferase) is responsible for the interconversion of D-glucose 1-phosphate (G1P) (Cori ester) and D-glucose 6-phosphate (G6P) (Fig. 6, Reaction 3). The enzyme is specific for α-D-pyranose phosphate and the K_{eq} is 19 at pH 7 and 25 °C for c_{G6P}/c_{G1P}. The reaction requires Mg^{2+} and is inhibited by fluoride.

The interconversion of D-glucose 6-phosphate and D-fructose 6-phosphate (Neuberg ester) is catalyzed by phosphoglucoisomerase (D-glucose 6-phosphate ketol-isomerase) (Fig. 6, Reaction 4). The enzyme is quite specific. It is competitively and specifically inhibited by 2-deoxy-D-glucose 6-phosphate, which is readily formed from the corresponding free deoxysugar in the presence of hexokinase.

Phosphofructokinase (ATP : D-fructose 6-phosphate 1-phosphotransferase) is a highly specific enzyme that controls the formation of D-fructose 1,6-diphosphate (Harden – Young ester) from fructose 6-phosphate. It requires ATP and Mg^{2+}. The enzyme can use phosphate from uridine triphosphate or from inosine triphosphate instead of from ATP. Its activity can be inhibited by high ATP concentrations, which occurs when lactate and pyruvate are oxidized aerobically to CO_2 via the Krebs cycle. In that case, glycolysis is blocked and glucose synthesis is favored. At very low ATP concentration, glycolysis is required for energy generation, and carbohydrate synthesis is shut off.

Fructose 1,6-diphosphatases (D-fructose 1,6-diphosphate 1-phosphohydrolases) are required for hydrolytic conversion of the diphosphate to the 6-monophosphate (Fig. 6, Reaction 5 a). The enzymes are specific and can be inhibited by high concentrations of their substrate and also by adenosine monophosphate (AMP).

The conversion of fructose 1,6-diphosphate to two molecules of triosephosphate [3-phosphoglyceraldehyde (Ficher – Bauer ester) and phosphodihydroxyacetone] (Fig. 6, Reaction 6) is carried out by aldolase (D-fructose 1,6-diphosphate : D-glyceraldehyde 3-phosphate-lyase). These two triosephosphate molecules are in equilibrium, supported by triosephosphate isomerase (D-glyceraldehyde 3-phosphate ketol-isomerase) (Fig. 6, Reaction 7). Only 3-phosphoglyceraldehyde undergoes further conversion to ethanol. In the course of cleavage of the hexose, triose products must be removed because the combined equilibrium lies in favor of hexose. The enzyme from yeast contains Zn^{2+} and, after removal or complexation of the metal, is activated equally well by Zn^{2+}, Fe^{2+}, and Co^{2+}. The activity of the enzyme is enhanced about 35 times in the presence of K^+ ions [112].

Glyceraldehyde 3-phosphate dehydrogenase [D-glyceraldehyde 3-phosphate: nicotinamide – adenine dinucleotide (NAD) oxidoreductase (phosphorylating)] catalyzes Reaction 8. Free sulfhydryl groups are required for its activity. The presence of NAD^+ and P_i (inorganic phosphate) is necessary. In the reaction, NAD^+ is reduced to NADH, which is then reoxidized in the reduction of acetaldehyde to ethanol or pyruvate to lactate. As a dehydrogenase, the enzyme is specific for D-glyceraldehyde 3-phosphate, the free dephosphorylated aldehyde, or acetaldehyde, and is strongly inhibited by D-triose 2,4-diphosphate.

The enzyme phosphoglycerate kinase (ATP : 3-phospho-D-glycerate 1-phosphotransferase) catalyzes the conversion of 1,3-diphosphoglycerate to 3-phosphoglycerate, with the generation of 1 mol of ATP (Fig. 6, Reaction 9). It requires a divalent metal cofactor (Mg^{2+}, Mn^{2+}, or Co^{2+}).

3-Phosphoglycerate, which is formed in Reaction (9) (Fig. 6), is isomerized to 2-phospho-D-glycerate by phosphoglyceromutase (2,3-diphospho-D-glycerate : 2-phospho-D-glycerate phosphotransferase) (Fig. 6, Reaction 10). The diester 2,3-diphosphoglycerate and Mg^{2+} are obligatory cofactors for the reaction.

The conversion of 2-phosphoglycerate to phosphoenolpyruvate (Fig. 6, Reaction 11) can be considered to be a dehydration reaction that is catalyzed by enolase (phosphopyruvate hydratase). This enzyme is widely distributed in living organisms. Divalent cations (Mn^{2+}, Mg^{2+}, Zn^{2+}, or Cd^{2+}) that can be antagonized by Ca^{2+} or Sr^{2+} are essential. Fluoride ion is an effective inhibitor, especially in the presence of P_i.

Pyruvate kinase (ATP : pyruvate phosphotransferase) catalyzes Reaction (12), (Fig. 6). For this reaction, Mg^{2+} or Mn^{2+} is required; Co^{2+} ions are competitive antagonists. Some enzymes of this type require K^+, Rb^+, or Cs^+ for full activity. The ions can be antagonized by Na^+ or Li^+. The reaction equilibrium is very favorable for the formation of pyruvate, and the turnover number is 6×10^{13} compared to 12×10^{13} for the opposite direction [113].

To produce ethanol from pyruvate, yeasts that operate anaerobically irreversibly decarboxylate pyruvate with the aid of pyruvate decarboxylase (2-oxo acid carboxylase) to yield acetaldehyde (Fig. 6, Reaction 14). Alcohol dehydrogenase (alcohol : NAD oxidoreductase) then catalyzes the reduction of acetaldehyde to ethanol (Fig. 6, Reaction 15). The enzyme responsible for the decarboxylation can satisfy its metal requirements by using either certain divalent ions (Mg^{2+}, Mn^{2+}, Co^{2+}, Ca^{2+}, Cd^{2+}, or Zn^{2+}) or trivalent ions (Al^{3+} or Fe^{3+}). Alcohol dehydrogenase from yeast contains Zn^{2+}. The enzyme has four catalytic sites and a molecular mass of 151 000. Both enzymes have been isolated from yeast and produced in crystalline form.

Pentoses. Yeasts metabolize aldopentoses by an oxidation – reduction reaction. They reduce D-xylose to xylitol, which is subsequently oxidized to D-xylulose (Fig. 7). The reduction is catalyzed by aldose reductase (alditol : NADP 1-oxidoreductase) [115]. The enzyme has a specificity for NADPH (reduced nicotinamide –adenine dinucleotide phosphate).

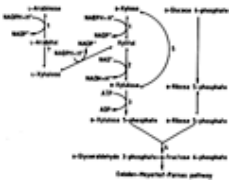

Figure 7. D-Xylose and L-arabinose metabolism in yeast [114]

1) Aldose reductase; 2) D-Xylulose reductase; 3) D-Xylulokinase; 4) Transaldolase and transketolase; 5) D-Xylose isomerase

D-Xylulose reductase (xylitol : NAD 2-oxidoreductase) is involved in the oxidation of xylitol to D-xylulose. The reaction is readily reversible, and NADH oxidation occurs with some ketoses including D-xylulose, D-fructose, and D-ribulose.

D-Xylulokinase catalyzes the phosphorylation of D-xylulose to D-xylulose 5-phosphate. The enzyme is found in bacteria [116], [117]. The presence of D-xylulokinase in yeast is indicated by the ability of many yeasts to use D-xylulose under both aerobic and anaerobic conditions [115], [118]. D-Xylulose 5-phosphate can be converted to pyruvate via both the pentose – phosphate and the Embden – Meyerhof – Parnas pathways.

Yeasts also metabolize L-arabinose by oxidation –reduction. The aldose is reduced to L-arabitol, which is converted to D-xylulose 5-phosphate, the key intermediate of the pentose phosphate pathway (Fig. 7) [114].

In the metabolism of D-xylose, the actual conversion of the sugar occurs after a lag period of adaption for the production of D-xylose-metabolizing enzymes [119], [120].

Some yeasts, such as *Candida utilis* and *Rhodotorula gracilis*, have inducible D-xylose isomerase in addition to the initial oxidation – reduction reaction and can isomerize D-xylose directly to D-xylulose [119], [121].

Ethanol can be produced readily from D-xylulose in high yield by yeasts that normally produce ethanol from glucose [118], [122].

About half of all yeasts can assimilate D-xylose under aerobic conditions [123]. Yeasts that could consume xylose or other pentoses under anaerobic conditions were unknown. But anaerobic conversion of D-xylose to ethanol has been demonstrated in the yeasts *Pachysolen tannophilus* [124] and *Pichia stipitis* [125]. Many yeasts can convert D-xylulose (which is a keto isomer of D-xylose) to ethanol under anaerobic conditions. The growth of yeasts on D-xylose under anaerobiosis has not been reported.

Certain yeasts can produce a high yield of ethanol directly from D-xylose (*Candida* sp. XF 217, *Pichia stipitis*), while others (*Saccharomyces cerevisiae, Schizosaccharomyces pombe*) can do so after D-xylose is isomerized to xylulose by a bacterial isomerase. Table 9 shows some advantages and disadvantages of these yeasts in the production of ethanol from xylose and xylulose.

Table 9. Pentose (xylose and xylulose) fermenting yeasts *

Yeast	Process	Advantage	Disadvantage
Candida sp. XF 217	batch	yield (0.42 g per gram of xylose), low xylitol formation	low ethanol tolerance
Kluyveromyces cellobiovorus	batch	cellobiose also	xylitol formation, low ethanol tolerance
Pachysolen tannophilus	batch	none	xylitol formation, low ethanol tolerance
Pichia stipitis	batch	yield (0.42 g per gram of xylose) low xylitol formation	low ethanol tolerance
Saccharomyces cerevisiae	simultaneous isomerization and xylulose fermentation	high ethanol tolerance, fast, yield (0.45 g per gram of xylulose)	isomerization, cost (xylitol formation)
Schizosaccharomyces pombe	simultaneous isomerization and xylulose fermentation	high ethanol tolerance, fast, yield (0.48 g per gram of xylulose)	isomerization, cost

* Reprinted with permission from [126].

5.1.3. *Fermentation Variables*

Effect of Substrate Concentration. The carbon substrate concentration has a significant effect on ethanol production. Previously, batch fermentation was carried out with an initial concentration of fermentable substrate of 14 – 18 wt %, which resulted in a final ethanol concentration of ca. 6 – 9 % (wt/vol) [127]. Increasing sugar concentration was an advantage because it led to higher final ethanol concentration and, subsequently, to reduction of distillation costs. At the same time, the growth of osmosensitive contaminants was suppressed. However, at sugar concentrations >14 %, plasmolysis of yeast cells begins. In addition, the initial rate of fermentation starts to decline before the ethanol concentration reaches a significant value [109].

Under anaerobic conditions and at low-to-moderate glucose concentrations, the rate of ethanol production may be represented by a Monod-type relationship [111]:

$$V = V_{max} S/(K_s + S)$$

where V is the ethanol production rate, g $g_{cell}^{-1} h^{-1}$; S is the glucose substrate concentration, g/L; and K_s is a constant with a low value of 0.35 g/L.

This relationship shows that under anaerobic conditions, maximum ethanol production per cell is achieved when sugar concentration is considerably > 3.5 g/L and <150 g/L. Inhibition of the ethanol production rate above 150 g/L of glucose is significant. If the inhibition caused by ethanol is considered, it appears that it limits the concentration of fermentable sugars.

Effect of Oxygen. Although biomass is produced under aerobic conditions and anaerobic conditions favor production of ethanol, oxygen was found to be essential for good fermentation. In addition to being a terminal acceptor of electrons from the respiratory chain, it also acts as a yeast growth factor and is involved in the synthesis of unsaturated fatty acids and ergosterol, which stimulate growth of yeast under anaerobic conditions and increase viability of cells [109], [128].

Oxygen is especially necessary when batch fermentation is carried out at high sugar levels requiring prolonged growth of yeast, or in continuous processes, because the yeast is unable to grow for more than four to five generations under fully anaerobic conditions. When oxygen disappears from the broth, the yeast continues to grow anaerobically. Cell division leads to a redistribution of the sterols and unsaturated fatty acids that have accumulated during aeration. After several generations the sterol and unsaturated fatty acid level is too low to enable normal functioning of the membrane causing a change in the physiology of the yeast [129], [130].

The necessary level of oxygen depends on the strain of yeast. Usually an oxygen tension of 6.7 – 13.3 Pa is required in the broth. If the concentration exceeds this value, cell growth is enhanced at the expense of ethanol production by promoting the Pasteur effect, i.e., the complete oxidation of glucose to CO_2 and H_2O via the tricarboxylic acid (TCA) cycle and the respiratory chain. In some yeasts, cell mass productivity increases drastically with a corresponding decrease in ethanol production at an oxygen tension >0.23 kPa [131]. This process is controlled by the concentration of fermentable sugar in the growth medium. If the concentration is high and conditions are aerobic, the Pasteur effect is likely to be diminished and glucose will be degraded by the aerobic pathway. This phenomenon, known as “aerobic” fermentation, results from the operation of the so-called reverse Pasteur or Crabtree effect, which represses both the synthesis and the activity of respiratory enzymes at glucose or sucrose concentrations greater than 0.02 – 0.1 % [109], [132].

Air sparging through the growth medium can be omitted if certain supplements, such as oleic acid, linolenic acid, tween 80, or ergosterol are present in the medium [129], [133]. Cultures enriched with these supplements can attain a high ethanol concentration (15.5 wt %) with a high substrate conversion efficiency (95 %) under anaerobic conditions.

Various sterols and fatty acids increase the viability of resting cells and prolong their fermentative activity [133], [134]. Incorporation of these compounds into the cell membrane appears to increase its permeability to ethanol and permits a high rate of ethanol transport out of the cells.

Effect of Ethanol. Ethanol is toxic to yeast. The general effect is most noticeable on the cell membrane; the major toxic effect has been postulated as membrane damage or a change in membrane properties [135]. Ethanol inhibits both growth and ethanol production in a noncompetitive manner [136]. When ethanol is present in concentrations of up to 2 %, the observed inhibition is almost negligible for most yeasts. With a higher concentration, the effect of ethanol becomes more evident (Figs. 8 and 9).

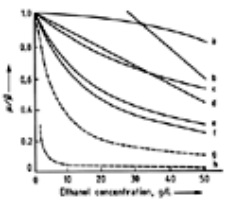

Figure 8. Comparison of the effect of various ethanol inhibition functions [137]

μ = specific growth rate; $\hat{\mu}$ = maximum specific growth rate; Source of ethanol: —— added, ------ autogenous

a) Continuous [138]; b) Batch [139]; c) Continuous [140]; d) Continuous [136]; e) Batch [141]; f) Batch [140]; g) Continuous [142]; h) Batch [143]

Figure 9. Effect of ethanol concentration on ethanol productivity [136]

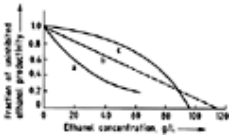

a) [144]; b) [136]; c) [138]

Concentrations of ethanol > 110 g/L stop both growth and ethanol production in most yeast strains. However, with most tolerant yeasts, ethanol production (not growth) is possible even in the presence of up to 20 % ethanol [109].

As shown in Figure 8, the effect of externally added ethanol on cell growth is much lower than the effect of autogenous ethanol. The effect of ethanol on cell viability increases at higher temperature [145] or if the cells are exposed to heat stress for a short period of time before coming into contact with ethanol [146].

Prolonged, continuous exposure of yeast to high ethanol concentration results in a loss of viability so that recycled cells must be continually replenished or discarded completely at intervals [147].

Effect of pH. The concentration of hydrogen ions in a fermentation broth affects yeast growth, ethanol production rate, byproduct formation, and bacterial contamination control.

Usually, industrial ethanol fermentation by yeast has an initial pH value of ca. 4 – 6, depending on the buffering capacity of the medium. In lightly buffered media, the initial pH is ca. 5.5 – 6 and in more highly buffered media, ca. 4.5 – 4.7 [127], [148]. If the pH value is <5 during fermentation, bacterial growth is severely repressed; the pH values for growth of most strains of *Saccharomyces cerevisiae* are 2.4 – 8.6, with an optimum of 4.5. Yeast sugar fermentation rates are relatively insensitive to pH values between 3.5 and 6 [109], [138].

Effect of Temperature. Most brewer's yeasts have a maximum growth temperature around 39 – 40 °C [126]. The maximum growth temperature reported for any species of yeast was 49 °C for *Kluyveromyces marxianus* [149]. Mesophilic strains of *Saccharomyces* have optimum cell yields and growth rates between 28 and 35 °C. The optimum and maximum temperatures for growth of thermophilic yeast are ca. 40 and 50 °C, respectively; these strains have a high maintenance requirement and more complicated nutritional requirements [109].

In batch processing, the optimum temperature for the complete utilization of glucose and the highest final ethanol concentration is generally slightly below the optimum growth temperature. This is attributed to enhanced ethanol inhibition at higher temperature [109], [150]. At higher temperature, the ethanol production rate is higher than its transport rate through the cell membrane. The difference in these rates results in an increase of ethanol concentration in the cells, a subsequent inhibition of some enzymes, and cell death. For continuous processes, the maximum process temperature (35 – 40 °C) in the absence of ethanol must be reduced by ca. 1 °C for each percentage increase in ethanol concentration [109].

Some yeasts have an optimum fermentation temperature of ca. 40 – 42 °C. They produce up to 12 % of ethanol with yields >90 % of theoretical [126], [151], [152]. Because sugar fermentation is exothermic (586 J of heat produced per gram of glucose consumed) [127], using yeasts that ferment at higher temperature substantially reduces cooling costs of fermentors.

5.1.4. *Direct Fermentation*

When polysaccharides (cellulose and starch) are used for ethanol production, they first must be hydrolyzed by means of different physicochemical or enzymatic methods and then converted to ethanol. Combining this two-step process into a single operation would be an advantage. This can be achieved by simultaneous saccharification and fermentation (SSF) processes using cocultures of microbes that are able to hydrolyze polysaccharides and convert sugars to ethanol, or by direct fermentation of polysaccharides to ethanol with monocultures.

In these cases the rate of enzymatic hydrolysis is increased by constant removal of the produced sugars through their conversion to the product. This is particularly important in the case of cellulose hydrolysis where glucose and cellobiose severely inhibit enzymes participating in the hydrolysis.

The coupled enzymatic hydrolysis of cellulose and simultaneous fermentation of the resulting sugars to ethanol by yeast have been investigated [153-157]. Table 10 shows simultaneous hydrolysis of ball-milled cellulosic substrates by cellulases of *Trichoderma reesei* QM 9414 and conversion of sugars to ethanol by *Candida brassicae.* Use of a thermotolerant yeast and cellulolytic enzymes in a combined saccharification fermentation process is an advantage because the optimum temperature for the hydrolysis of cellulose is 45 – 50 °C, and cooling problems can be simplified during large-scale fermentation [158].

Table 10. Simultaneous saccharification and fermentation (SSF) of various substrates with and without milling pretreatment (*Trichoderma reesei* QM 9414 cellulase with *Candida brassicae*) [157]

Feedstock	Conversion to ethanol			
	Unpretreated		Pretreated	
	24 h	48 h	24 h	48 h

Eucalyptus grandis sawmill waste	0	0	33.44	38.64
Acacia mearnsii sawmill waste	0	0	36.73	45.71
Southern pine sawmill waste	0	0	38.01	42.59
Digester rejects	41.76	54.55	40.31	72.86
Primary clarifier sludge	51.22	66.93	57.49	84.24
Newsprint	34.61	45.40	54.25	60.34
Cardboard	35.12	42.96	53.45	77.93
Air classified sawmill waste	49.58	53.00	59.05	75.13
Cotton linters	10.41	8.74	18.37	16.59
Rice straw	81.49	110.06	57.66	85.47
Rice hulls	45.99	50.19	65.57	96.05
Corn stillage	14.37	16.69	45.57	65.92

A simultaneous saccharification – fermentation process was also applied in the production of ethanol from starchy materials (cassava and corn). Enzymes from *Aspergillus niger, A. awamori,* and *Rhizopus* sp. were used for hydrolysis of starch with simultaneous yeast fermentation [159-161]. Ethanol yields in these processes ranged from 82 to 99 % of theoretical. In certain cases ethanol concentration reached a level of 20 vol % in five days [159].

The yeast *Schwanniomyces* is able to hydrolyze starch directly and partially ferment it to ethanol [162]. Amylolytic systems of *S. castellii* and *S. alluvis* have been reported [163-165]. Economic use of these and other amylolytic species is probably not feasible because of their limited ethanol tolerance [162], [166].

5.2. Production by Bacteria

Among bacteria the gram-negative *Zymomonas mobilis* occupies an outstanding position as an excellent producer of ethanol from glucose. This mesophilic organism was isolated at the beginning of the 20th century from palm wine and from the Mexican beverage pulque [167], [168]. The pioneering work of ROGERS et al. showed that this bacterium has a number of favorable characteristics for efficient production of industrial ethanol [169]. A major drawback of *Z. mobilis*, however, is its incapability to ferment pentoses such as xylose and arabinose [167]. The lack of microorganisms that can ferment pure ethanol from pentoses has given rise to a number of efforts to construct appropriate recombinant bacteria. By introduction of the genes for ethanol production from *Z. mobilis*, the native abilities of *Escherichia coli* for pentose fermentation were combined with an efficient ethanol-forming activity [170]. Alternatively, pentose-fermenting strains of *Z. mobilis* have been developed by genetic engineering [171-175]. A number of facultative and obligatory anaerobes are capabable of fermenting pentoses, but in addition to ethanol and carbon dioxide, they also synthesize large amounts of other fermentation products (e.g., acetic acid, formic acid, lactic acid, acetone, and hydrogen) [176-181].

5.2.1. *Ethanol Tolerance*

The growth of most bacteria is inhibited by ethanol concentrations of 10 – 20 g/L. However, both *Z. mobilis* and yeast can tolerate high ethanol concentrations. In fact, alcohol concentrations of 120 g/L have been attained by fermentations with *Z. mobilis* [169]. High ethanol concentrations generally destroy the structure and function of the cell membrane [182]. The cytoplasmic membrane of *Z. mobilis* contains various hopanoids **1** – **6** these membrane-stabilizing, pentacyclic triterpenoids are present at concentrations of 30 mg/g cell dry weight, i.e,. they make up 40 – 50 % of the organism's total lipid content [183-186]. Hopanoids play an important role in the ethanol tolerance of *Z. mobilis* by reinforcing the cytoplasmic membrane against the fluidizing effects of high concentrations of ethanol [184], [187]. Hopanoids are synthesized in bacterial cells under both aerobic and anaerobic conditions, whereas the biosynthesis of sterols in yeast requires oxygen. In yeast sterols are essential membrane-stabilizing triterpenoids.

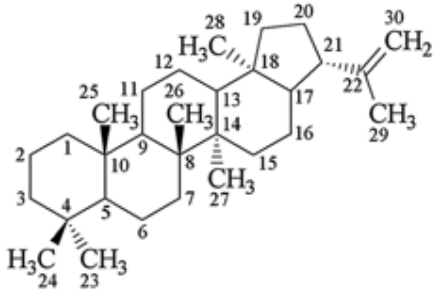

1

Diploptene

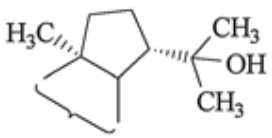

2

Diplopterol

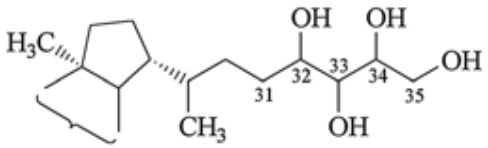

3

Tetrahydroxybacteriohopane

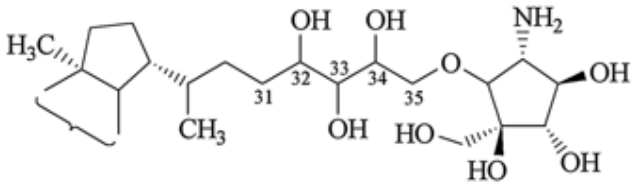

4

Bacteriohopanetetrol ether

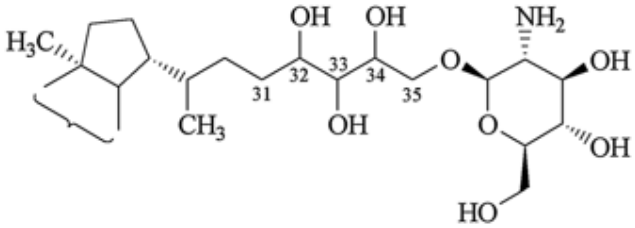

5

O- β-D-glucosaminylbacteriohopanetetrol

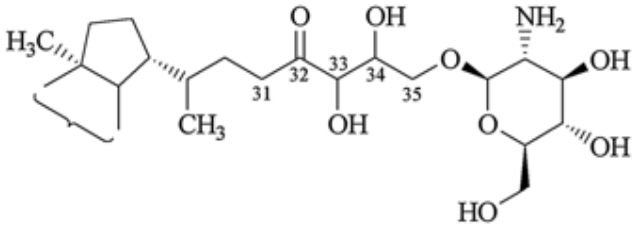

6

O- β-D-glucosaminyl-23-oxobacteriohopanetriol

5.2.2. *Fermentation of Glucose by Z. mobilis*

Zymomonas mobilis uses the 2-keto-3-deoxy-6-phosphogluconate or Entner – Doudoroff pathway to metabolize sugar, in which pyruvate is formed as an intermediate [167]. The formation of acetaldehyde by decarboxylation of pyruvate and its subsequent reduction to ethanol are identical to the reactions in yeast: 2 mol of ethanol and 2 mol of carbon dioxide are produced per mole of glucose. This homo-ethanol pathway involves two enzymes: pyruvate decarboxylase and alcohol dehydrogenase (see Figs. 10 and 11). However, the energy (ATP) yield of this bacterium, based on the amount of glucose metabolized, is only half of that obtained with yeast. Therefore, *Z. mobilis* can synthesize only half as much new cell mass per mole of glucose degraded as yeast does, but this results in a higher ethanol yield (Table 11).

Table 11. Fermentation properties of the bacterium *Zymomonas mobilis* and of the yeast *Saccharomyces carlsbergensis*

Property	*Zymomonas mobilis*	*Saccharomyces carlsbergensis*
Doubling time, h	2.51	5.64
Ethanol production rate $q_{P/X}$,* g g^{-1} h^{-1}	5.44	0.82
Cell yield $Y_{X/S}$,* g/g	0.028	0.043
Product yield $Y_{P/S}$,* g/g	0.465 **	0.460 **

* Explanation of symbols:

$$q_{P/X} = \frac{\text{g ethanol formed}}{\text{g cells formed} \cdot \text{h}}$$

$$Y_{X/S} = \frac{\text{g cells formed}}{\text{g substrate used}}$$

$$Y_{P/S} = \frac{\text{g ethanol formed}}{\text{g substrate used}}$$

** Max. theoretical product yield = 0.511 g/g.

Zymomonas mobilis, unlike yeast, does not require oxygen for growth, and this greatly simplifies fermentative ethanol production. In contrast to a number of strictly anaerobic bacteria, *Z. mobilis* is relatively insensitive to oxygen. Indeed, it possesses an electron-transport system similar to the respiratory chain of aerobic organisms [188-190] and can transfer hydrogen, formed in the conversion of glucose to pyruvate, to molecular oxygen. These reducing equivalents are then not available for reduction of acetaldehyde to ethanol; as a consequence, acetaldehyde accumulates and inhibits the growth of *Z. mobilis* [188], [191].

A major consideration in achieving economic ethanol production is the cost of the sugar used as a raw material. For this reason the fermentation of starch and waste starch has been investigated by two approaches: first, two-step processes, involving separate saccharification and fermentation stages were developed [192], [193]. For the fermentative production of ethanol from an industrial hydrolyzed waste starch fraction ("B-starch") *Z. mobilis* was used in two 50 m^3 continuous stirred-tank reactors with a residence time of 15 h, and production of 10 000 L/d of ethanol (96 vol %) was achieved [194]. If nonsterile B-starch hydrolysate was used, the process became unstable due to contamination by lactic acid bacteria. Lactic acid bacteria with an acidic pH optimum of growth were able to use the pentoses in the hydrolysed waste substrate, in contrast to *Z. mobilis*. It is known that lactic acid bacteria have a profoundly negative influence on the growth of *Z. mobilis*, although the inhibition mechanisms are unresolved [195]. To overcome this problem, fluidized-bed reactors with immobilized *Z. mobilis* cells in macroporous glass beads were used [193], [196]. The high concentration of immobilized *Z. mobilis* cells permitted unsterile fermentation of hydrolysed B-starch to ethanol at short residence times, so that contaminants were washed out. A cascade of two 55 L fluidized-bed reactors was operated stably at a mean residence time of 4.25 h and 99 % conversion of the unsterile hydrolysed B-starch, resulting in an ethanol concentration of 50 g/L and an ethanol productivity of 13 g L^{-1} h^{-1} [193].

A second approach is simultaneous starch saccharification and ethanol fermentation where *Z. mobilis* cells and amyloglucosidase are simultaneously present in the fermentor. With sago starch as substrate in a 500 L batch fermentor, the final concentration of ethanol obtained by this technique was 82 g L^{-1} with an ethanol yield of 0.457 g/g [197]. Whereas *Z. mobilis* generally produces high yields of ethanol from glucose and fructose, the conversion efficiency from sucrose-based substrates is reduced to 70 % due to the formation of byproducts such as levan and sorbitol [198]). A mutant of *Z. mobilis* lacking the property of levan synthesis has been shown to produce 73.5 g L^{-1} ethanol from 150 g L^{-1} sucrose in fermentation with immobilized cells [199]. The ethanol yield from sucrose by *Z. mobilis* could also be significantly enhanced (from 0.29 to 0.40 g/g) by addition of immobilized invertase to the fermentation broth [200]. Other co-immobilization systems of *Z. mobilis* with other microbes or enzymes for fermentation of different substrates have been described (for reviews, see [175], [201]).

5.2.3. *Fermentation of Other Sugars by Recombinant Z. mobilis Strains*

Up to now no yeast or bacterial strains are available that can ferment pentose to ethanol as efficiently as these organisms can ferment glucose or fructose. The price of the sugar substrate represents as much as 70 % of the total cost of ethanol fermentation. Efficiency could, however, be improved considerably by using both the hexoses and the pentoses of the plant biomass as substrates. A number of approaches have been made to genetically engineer *Z. mobilis* to broaden its range of utilizable substrates. Genes encoding enzymes of several metabolic pathways have been introduced and successfully expressed in *Z. mobilis*, and attempts have been made to construct new catabolic pathways (e.g. for L-arabinose, D-xylose) [171-175] .

For the conversion of the pentose sugars commonly found in agricultural residues and other lignocellulosic biomass to ethanol, the development of *Z. mobilis* strains with the capability of D-xylose and of D-arabinose fermentation was attempted. By introduction of four genes encoding xylose catabolism (i.e. xylose isomerase, xylulokinase, transaldolase, transketolase) into *Z. mobilis*, the organism was enabled to grow on xylose and ferment it to ethanol (Fig. 10) [202-204]. However, the xylose-fermenting recombinant strains showed decreased growth and ethanol production rates and increased byproduct formation [204], [205]. A ^{13}C NMR flux analysis indicated that ribose 5-phosphate is synthesized via the pentose phosphate pathway in *Z. mobilis*, and it identified a metabolic bottleneck in the recombinant xylose-fermenting *Z. mobilis* strain at the level of heterologous xylulokinase [204]. As was shown by ^{31}P NMR studies the levels of nucleoside triphosphates in cells grown with xylose were lower than those in cells grown in glucose-containing media. This energy limitation probably restricted the growth of the recombinant strain on xylose media [205].

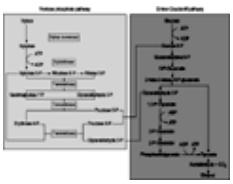

Figure 10. Catabolism of D-xylose in recombinant strains of *Zymomonas mobilis*

When grown with glucose – xylose mixtures, a recombinant, xylose-fermenting strain of *Z. mobilis* produced 62 g/L of ethanol from 65 g/L of glucose plus 65 g/L of xylose with a yield of 0.46 g ethanol per gram of sugar [206]. With substrates derived from lignocellulosic materials the ethanol-producing efficiency of *Z. mobilis* is substantially impeded by toxic substances present in the woody biomass after chemical acid hydrolysis, such as acetic acid and furfural (a pentose degradation product) [207], [208]. Nevertheless, it was shown that in a synthetic medium mimicing a dilute-acid hydrolyzate of yellow popular, a recombinant xylose-utilizing *Z. mobilis* strain produced 0.48 – 0.50 g of ethanol per gram of sugar, equivalent to a yield of 94 – 96 % of the theoretical maximum [209].

A second predominant pentose sugar found in agricultural residues (e.g., corn fiber) and other lignocellulosic biomass is *l*-arabinose. Five genes from *E. coli*, encoding *l*-arabinose isomerase, *l*-ribulokinase, *l*-ribulose-5-phosphate-4-epimerase, transaldolase, and transketolase, were introduced into *Z. mobilis* [210]. The engineered strain grew on and produced ethanol from L-arabinose as sole C source at 98 % of the maximum theoretical ethanol yield, but arabinose was not completely utilized [210]. A third abundant sugar in plant hemicelluloses, the hexose *d*-mannose, is not used as a growth substrate by *Z. mobilis*, but is taken up via the glucose uptake system (glucose facilitator protein) of the organism [211]. It was shown that *d*-mannose was phosphorylated by a side activity of the resident fructokinase to mannose 6-phosphate. By introduction of the *E. coli* gene encoding phosphomannose isomerase, *Z. mobilis* strains were enabled to utilize mannose as the sole carbon source [211].

5.2.4. *Ethanol Production by Recombinant Enterobacteria*

A variety of bacteria are able to metabolize and ferment both hexose and pentose sugars, but all produce a mixture of fermentation products. Metabolic studies have shown that these bacteria use both the pentose phosphate pathway and the Embden – Meyerhof pathway to metabolize hexoses and pentoses to a common intermediate, pyruvic acid, which is subsequently converted to a variety of fermentation products (Fig. 11). Members of the enteric group of bacteria, such as *E. coli*, *Klebsiella*, and *Erwinia*, perform a mixed-acid fermentation which yields a nearly equal mixture of ethanol, acetic acid, and formate (or carbon dioxide and hydrogen) to provide redox balance. INGRAM and collaborators focused on the genetic engineering of enteric bacteria using genes encoding *Z. mobilis* pyruvate decarboxylase (PDC) and alcohol dehydrogenase B (ADH B) [170], [212], [213]. The *adhB* and *pdc* genes were assembled to form an artificial operon (PET operon) and were expressed to high levels as plasmid-encoded genes or as chromosomally integrated genes in *E. coli* [214], [215]. These high levels of recombinant enzymes coupled with the low K_m of PDC for pyruvate resulted in a redirection of the carbon flux to ethanol and carbon dioxide, thus preventing the formation of lactic and acetic acids [216], [217].

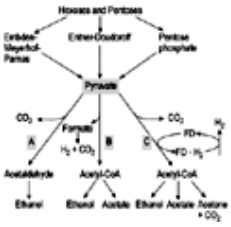

Figure 11. Microbial pathways of sugar fermentations

A) Pyruvate decarboxylase (EC 4.1.1.1; yeast and *Z. mobilis*); B) Pyruvate ferredoxin oxidoreductase (EC 1.2.7.1; Clostridia); C) Pyruvate formate lyase (EC 2.3.1.54; Enterobacteria)

FD = ferredoxin; P = phosphate.

Analogous plasmid constructs have been used to engineer *Klebsiella oxytoca* [218], [219] and *Erwinia chrysanthemi* [220]. In addition to a wide substrate range of monomeric sugars *K. oxytoca* has the native ability to transport and metabolize cellobiose and cellotriose, the soluble intermediates of the enzymatic hydrolysis of cellulose [221], [222]. During cellulose bioconversion in simultaneous saccharification and fermentation processes, the use of *K. oxytoca* eliminated the need to add supplemental cellobiase [223], [224]. Additional studies have added the PET operon to strains of *E. chrysanthemi* that have the native ability to secrete endoglucanase and possess PTS systems for cellobiose and cellotriose utilization [220].

5.2.5. *Comparison of Recombinant Bacterial Strains*

Bacterial ethanol production from pentose sugars can be approached by either exploiting the native characteristics of *Z. mobilis* for efficient ethanol fermentation, or the ability of *E. coli* to utilize a broad range of substrates. This can be brought about by genetic engineering of *Z. mobilis* for pentose utilization and of *E. coli* for ethanol production (Table 12).

Table 12. Xylose fermentation properties of of recombinant strains of *Z. mobilis* and *E. coli*

Property	*Z. mobilis* ZM4 (pZB5) [205] 60 g/L xylose	*E. coli* ATCC11303 (pLO1297) [215] 80 g/L xylose
Ethanol production rate $q_{P/X}$,* g g^{-1} h^{-1}	1.40	1.00
Cell yield $Y_{X/S}$,* g/g	0.010	0.050
Product yield $Y_{P/S}$,* g/g	0.410 **	0.470 **

* Explanation of symbols:

$$q_{P/X} = \frac{\text{g ethanol formed}}{\text{g cells formed} \cdot \text{h}}$$

$$Y_{X/S} = \frac{\text{g cells formed}}{\text{g substrate used}}$$

$$Y_{P/S} = \frac{\text{g ethanol formed}}{\text{g substrate used}}$$

** Max. theoretical product yield = 0.511 g/g.

5.3. Fermentation Modes

The two basic modes for production of ethanol by fermentation are the batch and continuous processes. A semicontinuous mode, which is a combination of the two, is also used.

5.3.1. *Batch Processes*

Most fermentation ethanol is produced by batch processes [111], [225]. A batchwise process for the fermentation of corn is shown in Figure 23. Typical kinetics in a batch process are shown in Figure 12.

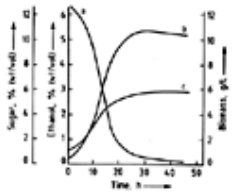

Figure 12. Ethanol production by *Kluyveromyces marxianus* 10606 [226]

a) Sugar; b) Ethanol; c) Biomass

Conversion of sugars in a simple batch system is ca. 75 – 95 % of theoretical, with a final ethanol concentration of 10 – 16 vol % [225], [227]. The productivity of simple, conventional batch processes is usually 1.8 – 2.5 g of ethanol per liter of fermentor volume per hour [111].

To increase productivity, a batch process with recycling of the yeast was developed; this is known as the “Melle Boinot” process. Recycling of yeast cells creates a high biomass concentration at the beginning of the process, which reduces the time for the conversion of substrate to ethanol (Table 13) [228]. This rapid fermentation also increases productivity. Increasing the initial concentration of yeast cells decreases yield. With a yeast cell concentration of ≥ 21 g/L, no growth was noticed; the existing cells used nutrients for both ethanol production and their own maintenance [228].

Table 13. Effect of initial cell concentration on time of fermentation during batch culture [228]

Initial cell concentration, g/L	Time to develop 12 vol % ethanol, h
21.4	7
23.6	6
27.2	5
26.13 *	4.5

* With yeast vitamins.

Under rapid fermentation conditions, a high ethanol concentration is attained quickly and the death rate of yeast cells is quite high. The death rate can be decreased by fermenting at lower temperature and by maintaining the dissolved oxygen concentration in the medium at 10 – 20 % of oxygen saturation [145], [229]. The tolerance of some yeasts to ethanol is also increased in the presence of yeast vitamins [228], [230].

Although classic batchwise processes for ethanol production are attractive because of their simplicity, they have many disadvantages such as low productivity, difficulty in automating, long and frequent downtime, and significant labor cost.

5.3.2. *Continuous Processes*

Without Cell Recycle. Generally, continuous processes eliminate most of the disadvantages that are inherent in the batch processes. They may be carried out for a long time period without shutdown, thus eliminating downtime between batches, increasing overall productivity. Because of increased productivity, smaller volume reactors are required than in batch production. Continuous processes can be fully automated and operated under conditions that give a uniform product. Continuous stirred tank reactors are used for these processes (Fig. 13).

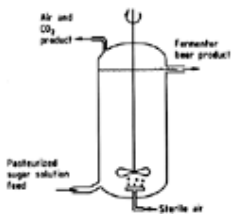

Figure 13. Continuous stirred tank reactor [111]

Simple continuous processes for ethanol production are characterized by continuous addition of feed (including oxygen) to a reactor in which a desirable steady state has been established. Beer is also removed continuously from the reactor. The beer contains ethanol, biomass, and unconsumed nutrients. The productivity of such a process (up to 6 g L^{-1} h^{-1} with highly productive strains) is about three times higher than that of a batchwise process. Therefore, the reactor volume need be only one-third that used to produce the same amount of ethanol in a batchwise manner. In this process, both specific ethanol productivity and total productivity of the reactor are limited by ethanol inhibition. In addition, total productivity is also limited by low biomass concentration (10 – 12 g/L) [111]. A sugar concentration of ca. 10 % in the feed gives the highest reactor productivity (Fig. 14). Economic analysis showed a 53 % reduction in operating costs and 50 % in capital costs for this process compared to classic batch processes [231].

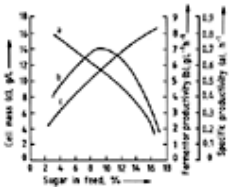

Figure 14. Effect of glucose concentration on continuous fermentation [231]

a) Specific productivity; b) Fermentor productivity; c) Cell mass

Sometimes two or more continuous stirred tank reactors can be joined in series to increase productivity. With two reactors, fresh medium is pumped to the first reactor; residence time is adjusted so that conversion of sugar is incomplete. This results in a lower ethanol concentration in the first reactor compared to the final concentration and, therefore, less ethanol inhibition. The outflow from the first reactor is fed to the second, where fermentation is completed. The productivity of the second reactor is lower than that of the first, but the overall productivity is increased compared to the system with one reactor [136], [232].

With Cell Recycle. Continuous systems with cell recycle were developed in order to increase productivity over systems that did not recycle biomass. The biomass from the stream leaving the reactor is separated by centrifugation and returned to the reactor. Yeast concentration as high as 83 g/L can be maintained during ethanol production [229]. Such high concentrations permit rapid and complete fermentation of concentrated sugar solutions. Fermentor ethanol productivities of 30 – 51 g L^{-1} h^{-1} have been achieved, which represents a more than tenfold increase over continuous fermentation without cell recycle [228], [229], [233], [234].

Both investment and operating costs of these processes are higher because of the required separation of cell mass from beer by centrifugation. Simpler and less expensive systems for cell separation are needed. Because of the low settling rate of yeast (<1 cm/h), shallow-depth sedimentation of biomass with the use of plate or tube settlers has been proposed [235]. The tube settler was found to be slower than the plate settler in reaching maximum effluent clarity. The plates and tubes are inclined to allow self-cleaning by the yeast sliding down the collection surfaces. The sedimentation rate of the yeast cells increases when evolution of CO_2 is repressed. This repression can be obtained by heat treatment for 20 min at 50 °C, which does not seriously impair cell viability. These settlers help reduce bacterial contamination because of their lower settling velocity.

Other separators, such as the whirlpool separator in which yeast cells are deposited in a central cone when the overflow is pumped tangentially into a vertical cylindrical vessel, have also been developed [236]. Another method involves a partial recycle fermentor that has a separate settling zone, which is not disturbed by agitation. This allows partial separation of cells [237].

5.3.3. *Other Processes*

Flocculating Cells (Internal Recycle). High-ethanol-producing and flocculating yeasts are of particular commercial interest. These cells can readily be concentrated and separated without the use of mechanical devices (e.g., centrifuges and conventional settlers). This simplifies the process and makes it more economic [234], [238-240]. A highly flocculating yeast, *Saccharomyces diastaticus,* was investigated for continous ethanol production from various substrates [234]. The process involved one fermentor with internal cell settling and no cell recycle or cell concentration by centrifuges or other such mechanical devices. A high specific ethanol productivity rate of up to 50 g L^{-1} h^{-1}, with an inflowing sucrose concentration of 112 g/L (using fodder beet juice), was obtained at a dilution rate of 1.033 h^{-1}.

A semicontinuous (fill-and-draw) process was also developed using the flocculating yeast, *S. diastaticus* [241].

The yeast is allowed to grow to a high biomass concentration in an agitated bioreactor; this assures good mixing and mass transfer during fermentation. The agitation is stopped after almost all the sugar has been consumed, and the yeast is allowed to settle rapidly (≤1 min) in the bioreactor. The clear supernatant, which is the alcohol-containing liquid broth, is decanted and sent to distillation. Next, fresh medium is added to the bioreactor, which starts a new fermentation batch. These cycles can be repeated (ten times or more) without loss in productivity and cell viability. High ethanol productivity is achieved with a very short fermentation time (Fig. 15).

Tower Reactor. The tower reactor is convenient for internal settling of flocculating yeast cells [242]. It consists of a vertical cylindrical tower with a conical bottom (Fig. 16). Above the cylindrical section is a large, baffled settling zone (d), which has the form of a yeast separator. The separator can have various forms and provides a space free from rising gas, which enables yeast cells to settle and be returned to the main part of the reactor. An overall aspect ratio of 10 : 1, with an aspect of 6 : 1 on the cylindrical section, is used. The reactor does not have a mechanical agitator. High cell concentrations in excess of 100 g/L are achieved without any auxiliary mechanical separator. Compared to simple batch processes, productivities are 32 – 80 times higher in the tower reactor [111], [239], [240]. Hydraulic retention time is less than 0.4 h, and conversion efficiency is up to 95 % of theoretical [240].

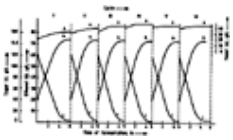

Figure 15. Semicontinuous fermentation of fodder beet juice with *Saccharomyces diastaticus* [241]

a) Yeast ; b) Ethanol ; c) Sugar

Figure 16. Tower fermentor (APV Co.) [111]

a) Flocculating yeast plug; b) Baffles; c) Attemperator jackets; d) Yeast settling zone; e) Clarifying tube

Flocculation and sedimentation characteristics remain unchanged over several months of continuous operation. Selective yeast retention is maintained even after moderate contamination, and any infecting organisms can be quickly washed out without any subsequent effect on reactor performance. Cell viability of 90 – 95 % is attainable at biomass concentrations in excess of 100 g/L and remains at this level when the substrate is limited. The viability drops to 70 – 80 % when the ethanol concentration is allowed to exceed 75 – 85 g/L [240].

A major disadvantage of this type of reactor is the long time required (ca. two weeks) for the initial start-up.

Membrane Bioreactor. Buildup of a high yeast concentration can be achieved in a simple continuous dialysis fermentor. The reactor has fermentation and nutrient zones to which fresh medium is fed and from which beer is removed. These two zones are separated by a membrane, which prevents biomass escaping from the fermentation zone and allows transfer of nutrients. At the same time ethanol passes through the membrane to the nutrient zone where it is recovered in an overflow [111], [243].

The rate of ethanol production in this bioreactor is limited by the rate of nutrient diffusion across the membrane. This problem can be overcome by using a pressure dialysis reactor (Fig. 17) in which a pressure differential is applied across the membrane (c) to force a flow of medium through the fermentation zone [111].

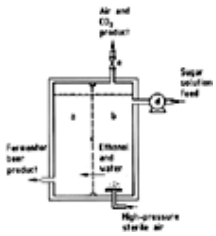

Figure 17. Pressure dialysis fermentor [111]

a) Low-pressure product recovery zone; b) High-pressure fermentation zone; c) Permeable dialysis membrane; d) High-pressure feed pump; e) Pressure-regulating valve

Membrane fouling by proteins, which results from cell lysis and can affect flow, is a problem with this type of bioreactor. The problem can be alleviated by replacing the fixed membrane with a rotating microporous membrane cylinder (rotofermentor) enclosed in a stationary fermentor vessel. The rotofermentor allows continuous removal of metabolic products (including ethanol) in the beer by filtration through the rotating membrane under a pressure of 115 – 170 kPa (Fig. 18). The cells are retained in the annular part of the reactor outside the membrane [244]. Large molecules that tend to accumulate on the membrane surface are sent back into the annular zone by the centrifugal force generated by rapid rotation of the membrane. This cleans the membrane and allows diffusion through it. In the rotofermentor, the ethanol production rate reaches 26.8 g L^{-1} h^{-1} at concentrations of 24.8 and 50.4 g/L of yeast cells and ethanol, respectively [245]. The reactor has several serious mechanical and operating disadvantages related to the complexity of the system and durability of certain parts of the reactor (membrane and seals).

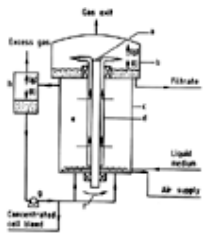

Figure 18. Schematic diagram of the rotofermentor assembly [244]

a) Rotor; b) Filtrate chamber; c) Fermentor; d) Rotating microporous membrane; e) Concentrated cell growth; f) Motor drive; g) Recycle pump; h) Gas – liquid separator

Membrane bioreactors in which the fermentation vessel is coupled in a semi-closed loop configuration to a membrane filtration unit are available [246], [247]. The fermentation broth is pumped through the filtration unit. Inhibitory products permeate through the membrane, while retained yeast cells are recycled to the fermentor. High cell concentrations (up to 100 g/L) and ethanol production rates (up to 100 g L^{-1} h^{-1}) are achieved [246].

Vacuum Bioreactor. Because high concentrations of ethanol are toxic to yeast, a vacuum process for continuous removal of ethanol from the beer has been developed [233], [248].

The bioreactor operates at a total pressure of 6.7 kPa. At that pressure, water and ethanol boil at a temperature that is compatible with the yeast. The liquid level in the bioreactor remains constant during operation of the bioreactor. An ethanol production rate of 82 g L^{-1} h^{-1} is achieved with a yeast density of 120 g/L. In the reactor, the ethanol concentration is kept at 35 g/L, while the ethanol concentration in the vapor product is > 200 g/L. Oxygen is sparged instead of air in order to satisfy

vacuum operation requirements [249]. Oxygen sparging can be omitted if the medium is supplemented with sterol and fatty acids [248].

The compressors used in the process must be large and must operate at low pressure. Capital cost of the compressors is quite high. Energy requirements exceed those of the normal ethanol fermentation process [228]. The probability of contamination is increased in this process.

A modification of the process exists (Fig. 19) in which the reactor operates at atmospheric pressure and CO_2 is released without compression from the reactor [249]. Beer is circulated continuously through a flash vessel (d) where ethanol – water solution is stripped under vacuum.

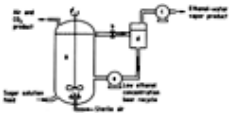

Figure 19. Continuous flash fermentation [111]

a) Atmospheric pressure fermentor; b) Flow-regulating valve; c) Vacuum compressor; d) Flash vessel; e) High-compression pump

Bioreactor with Solvent Extraction. The concentration of ethanol in beer can be reduced during fermentation by solvent extraction. This reduces inhibition caused by the product, increases productivity of the reactor, and allows a higher concentration of sugar to be fermented in a shorter period.

In some cases, after separation of the cells, the clear beer is treated with an immiscible solvent that removes most of the ethanol. The resulting beer containing only small amounts of ethanol is returned to the reactor, and the saturated solution undergoes ethanol separation [250].

In another system, both extractant and nutrients can be continuously fed to a reactor that contains immobilized cells. The ethanol produced is extracted, which reduces its concentration in the aqueous phase, and is later recovered from the extractant [251]. By using the technique of ethanol extraction from beer with a mixture of *n*-dodecanol (60 %) and *n*-tetradecanol (40 %), the ethanol production is five times greater than in the same system without extraction. A solution of 407 g/L of glucose can be totally fermented by a yeast that cannot normally transform more than 200 g/L glucose [251].

The extractant used in this technology must have the following properties [252]:

1. it must be nontoxic to yeast;
2. it should be selective for ethanol in comparison with water and secondary fermentation products;
3. it must have a high distribution coefficient for ethanol; and
4. it should not form emulsions with fermentation broth.

Further development of this extractive technology is necessary to avoid inhibitory effects of ethanol and attain high rates of ethanol production. If a good extractant is found, the cost of ethanol separation by means of a membrane technique will be considerably decreased compared to the classic distillation of ethanol from beer.

Immobilized Cells. A high yeast concentration (5.4×10^{10} cells per milliliter) [253] in the fermentor can be obtained by various cell immobilization techniques, e.g., by entrapment in a gel matrix, covalent binding to surfaces of various support materials, or adsorption on a support. These systems do not require agitation. The immobilized cells are retained in the reactor; therefore, cell separation devices and recycle are not needed. High dilution rates without cell washout can be achieved.

Immobilized cells can be used in fixed- and fluidized-bed reactors. In these cases, the substrate solution flows continuously through the reactor, and the immobilized cells convert available sugar to ethanol. Calcium alginate [9005-35-0] can be used to entrap yeast cells. In a system with calcium alginate beads, a maximum ethanol productivity of 53.8 g/L was achieved at a dilution rate of 4.6 h^{-1} and an initial glucose concentration of 127 g/L [254].

Carrageenan [9000-07-1] can also be used to immobilize yeast cells, giving ethanol production rates of 43 g L^{-1} h^{-1} at a dilution rate of 1.0 h^{-1} and sugar utilization of 86 % from the feed containing 100 g/L glucose [255].

Table 14 shows the ethanol production rates of *Saccharomyces cerevisiae* and *Kluyveromyces fragilis* cells immobilized by entrapment, covalent binding, and adsorption.

Table 14. Ethanol productivity from immobilized systems (*Saccharomyces cerevisiae* and *Kluyveromyces fragilis*)

System	Feed sugar	Feed sugar concentration, g/L	Feed sugar used, %	Dilution rate, h^{-1}	Maximum ethanol productivity, g L^{-1} h^{-1}	Reference
S. cerevisiae carrageenan	glucose	100	86	1.0	43	[255]
S. cerevisiae calcium alginate	glucose	127	63	4.6	53.8	[254]

S. cerevisiae calcium alginate	molasses	175	83	0.3	21.3	[256]
S. cerevisiae carrier A.	molasses	197	74	0.35	25	[257]
S. cerevisiae glass beads coated with gelatin and glutaraldehyde	glucose	150	99		43	[258]
S. cerevisiae anionic exchange resin XE-352	glucose	120	94		53.1	[260]
K. fragilis ceramic Intalox saddles coated with gelatin and glutaraldehyde (two fixed-bed reactors)	lactose	109	99		50.6	[259]

A comparison between batch and continuous (free and immobilized) processes for ethanol production is shown in Table 15. The ethanol production rate in an immobilized cell system had a lower optimum temperature (30 °C) than a system with free cells (37 °C) [254]. The downshift of the temperature optimum was possible because of the effect of ethanol inhibition (which increases with increased temperature) within the gel matrix.

Table 15. Ethanol productivity in free and immobilized systems (with *Saccharomyces cerevisiae*) [261]

Process	Substrate	Volumetric ethanol production rate, g $L^{-1}h^{-1}$
Batch	molasses	2.0 *
Continuous (free cell)	molasses	3.35
Continuous (immobilized)	molasses	28.6

* Excluding downtime of fermentor.

A diffusion limitation is imposed by the gel matrix, which produces a substrate and product gradient within the bead. Therefore, the immobilized system should be operated at lower temperature than the free cell system to minimize inhibition and maximize ethanol production.

The inhibition caused by high substrate concentration is excluded if the rate of sugar conversion to ethanol is higher or equal to the rate of sugar diffusion into the matrix.

A difference between the optimum pH values for the production of ethanol by free and immobilized cells was also found. Systems with free *Saccharomyces cerevisiae* cells have an optimum pH of 4.5, but the immobilized system showed a broad optimum spectrum with a maximum activity between pH 3.0 and 7.5. The reason for this difference is a pH gradient within the beads [254], [262].

5.4. Raw Materials and Processes

Raw materials for the production of ethanol by fermentation can be classified as:

1. readily fermentable carbohydrates that can be used directly, and
2. starch and other organic materials that must be converted to a fermentable form prior to fermentation.

The raw materials come from three major sources:

1. agricultural crops,
2. forest products, and

3. industrial and agricultural byproducts and residues.

Depending on need, end use, and availability, the choice of raw material varies for different regions, countries, and industries.

5.4.1. *Readily Fermentable Carbohydrates*

Various sugar crops, such as sugarcane, sugar and fodder beet, fruit crops, and crops based on crassulacean acid metabolism (CAM), are in this category.

Sugarcane. Sucrose [57-50-1] (α-D-glucopyranosyl- β-D-fructofuranoside) is the sugar obtained from cane or beet (⟶ Sugar). Sugarcane is a tropical crop whose successful cultivation is limited to an area spanning 37 °N to 31 °S.

Although sugarcane is grown mainly for production of table sugar and molasses, it is also an excellent raw material for the production of ethanol. The fermentable carbohydrates from sugarcane can be used either as the cane juice directly or as blackstrap molasses (a sugar byproduct). A material balance shows that 160 kg of fermentable solids can be obtained from 1 t of cane [263].

The cane juice is prepared by crushing raw cane and extracting the sugar with water, followed by clarification using milk of lime and H_2SO_4 to precipitate the inorganic materials [264]. The resulting extract is a green, sticky fluid, slightly more viscous than water, with an average sucrose content of 12 – 13 % [265].

Blackstrap molasses is the residue remaining after sucrose has been crystallized from cane juice. Molasses is a heavy viscous material, which contains sucrose, fructose, and glucose (invert sugar) at a total concentration of ca. 50 – 60 % (wt/vol) [266]. In contrast to cane juice, molasses is stable on storage and is usually diluted to the desired concentration just prior to fermentation.

A typical process for the production of ethanol from sugarcane is depicted in Figure 20. Production data are listed in Table 16. Ethanol production reaches a maximum after 14 – 20 h and then decreases until ca. 95 % of the available sugar is consumed. The process is usually batchwise, but some semicontinuous [232] and continuous operations [242], [267-269] are also used.

Table 16. Yields in production of ethanol from sugarcane [267]

	Alcohol, indirectly from molasses	Alcohol, directly from sugarcane juice
Sugarcane yield in 1.5 – 2-year cycle (south-central region), t/ha *	63	63
Average sucrose yield (13.2 wt %), t/ha	8.32	8.32
Crystal sugar production, t/ha	7.0	
Final molasses or cane juice production, t/ha	2.21	66.2
Fermentable sugar, molasses, or juice, t/ha	1.32	8.73
Alcohol yield at 100 % global efficiency, kg/ha	675	4460
Alcohol yield at reasonable 85 % global efficiency, L per ton of cane	11.5	75
or in L/ha	730	4800

* Hectare (ha) = 10^4 m^2.

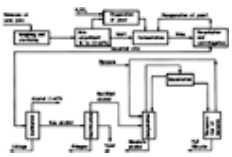

Figure 20. Typical process for the production of ethanol from sugarcane [267]

In the *batch process*, several fermentors are usually operating at staggered intervals to provide a continuous feed to the distillation columns. Overall productivity is ca. 18 – 25 kg of ethanol per cubic meter of fermentor volume per hour [225]. The “Melle Boinot” process is used in most Brazilian distilleries (see Section Batch Processes).

Ethanol has been produced in a *continuous process* (using continuously stirred tank reactors) from molasses by Danish Distilleries [268]. The process is shown in Figure 21, and performance data are given in Table 17.

Table 17. Performance data for the Danish Distilleries process * [268]

	Fermentor 1 (f_1) **	Fermentor 2 (f_2) **
Amount of yeast and dry matter per liter, g	10	10
pH	4.7	4.8
Alcohol, vol %	6.1	8.4
Residual sugar, %	1.0	0.1
Temperature, °C	35	35

* Residence time in each fermentor: 10.5 h; influx: 600 kg of molasses diluted in 22×10^3 L/h.
** See Figure 21.

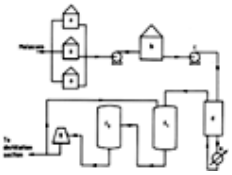

Figure 21. Continuous production of ethanol by Danish Distilleries [268]

a) Storage tank; b) Intermediate container; c) Metering pump; d) Regeneration section; e) Plate heat exchanger; f) Fermentor; g) Yeast separator

According to this process, the molasses is stored in two or three 1500-m^3 tanks from which it is pumped to intermediate containers. The material is adjusted for pH and nutrients (nitrogen and phosphorus), sterilized at 100 °C by using plate heat exchangers, and then introduced to three fermentors with a total volume of 170 m^3. The fermented wort is centrifuged after fermentation, and the live yeast returned to the first fermentor. At the start, sufficient yeast propagation must be accomplished by aeration (0.02 – 0.03 L of air per liter of liquid per minute). The yield is ca. 28.29 L of alcohol per 100 kg of molasses, or a maximum of ca. 65 L of alcohol per 100 kg of fermentable sugar.

A continuous process for production of beer from sugar by use of *tower fermentors* is shown in Figure 22 [270], [271].

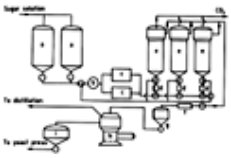

Figure 22. Commercial ethanol tower fermentation system (APV Company) [242]

a) Buffer storage tank; b) Divert line; c) Pasteurizer; d) Flow controller; e) Vertical cylindrical tower; f) Chiller; g) Yeast green beer buffer storage; h) Centrifuge; i) Separator

The key to the process is a vertical cylindrical tower fermentor with a conical bottom. A baffled yeast settling zone constitutes the upper part of the fermentor. The fermentor uses a flocculent yeast, which is pumped into the base of the tower. As the reaction proceeds, the beer rises, and the flocculating yeast settles and is retained in the reactor. High cell densities of 50 – 60 g/L are achieved without the use of mechanical cell concentration or separation devices. Short residence time (<4 h) with a sugar concentration of up to 12 % (wt/vol) sucrose, 90 % sugar utilization, and 90 % conversion to ethanol, produce up to 5 % ethanol in the final broth. The overall productivity of this system can be up to 80 times higher than that of the simple batch system.

Sugar Beet. Like sugarcane, sugar beet produces carbohydrates that consist primarily of sucrose (⟶ Sugar). Sugar beet is a more versatile crop than sugarcane. It can tolerate a wide range of soil and climatic conditions, and is grown throughout nearly half of the United States, Europe, Africa, Australia, and New Zealand.

In addition to sucrose, sugar beet contains sufficient nitrogen and other organic and microorganic nutrients [272] so that little, if any, fortification is required prior to fermentation. Another benefit is the high yield of coproducts such as beet tops and extracted pulp. The pulp has a high feed value, and the tops may be returned to the soil for erosion control and nutrient replacement. The yield of fodder beets is high (ca. 50 – 150 t/ha); their composition is described in [273].

A new fodder beet crop, produced in New Zealand through a genetic cross between sugar beets and marigolds, gives greater yields of fermentable carbohydrates per hectare than does sugar beet [274]. In addition, the sugar from fodder beet is reported to be more resistant to degradation over long storage.

Processes for the production of alcohol from sugar and fodder beets are basically the same as from sugarcane.

Fruit Crops. Many crops (grapes, plums, peaches, apricots, pineapples, etc.) contain variable proportions of sugars (sucrose plus fructose, usually 6 – 12 %). The fruit sugars can be readily fermented to alcohol, and this is done on a large scale for production of alcoholic beverages. The alcohol content of the product, which basically is the liquid after fermentation, separation of yeast, further treatment, and aging, depends on use and fermentation conditions. Table wines have <14 % alcohol, whereas wines with >14 % alcohol fall in the category of desert wines and aperitifs (⟶). Higher concentrations of alcohol are achieved by means of distillation to produce “strong” alcoholic beverages (e.g., brandy, whiskey, gin, vodka) (⟶).

Alcohol for industrial or fuel use is seldom produced from fruit and vegetable crops. However, some fruit from tropical and

semiarid climates, such as dates [275], mohwa flowers [276], and rain tree fruit [277], have been investigated for fuel alcohol production.

Crops Based on Crassulacean Acid Metabolism. Interest in using the agriculturally semi- or nonproductive regions of the world to grow alcohol-producing crops has increased [278]. These regions could be used to grow plants that utilize crassulacean acid metabolism (CAM) because their photosynthetic metabolism is extremely efficient with respect to irrigation requirements. These plants exhibit above-average productivity (expressed as a function of biomass production per unit of existing biomass) compared to other agricultural crops.

The CAM plants that are high in fermentable carbohydrates include various cacti (e.g., *Opuntia* sp.) and other plants such as *Euphorbia lathyrus* and *Agave* sp. Few data are available on potential ethanol production from these crops and its economic feasibility; however, an estimated 50 t/ha of these crops could be produced annually in subagricultural areas [279].

5.4.2. *Starch*

A variety of starch materials, such as grains, cassava, sweet potatoes, sweet sorghum, and Jerusalem artichoke, can be used for fermentation to ethanol (⟶ Starch). Selection depends on various factors, the major ones being climate and availability for large-scale production. Corn, wheat, potatoes, and Jerusalem artichokes are the most common raw materials in Europe and North America, whereas rice, cassava, sweet potato, and sweet sorghum are important in tropical countries.

Corn. Corn is the preferred raw material for conversion to alcohol in the United States and parts of Europe. It is available in large quantity, and its price (especially for low-grade or distressed corn) is thus acceptable for conversion to ethanol. Conversion to ethanol is efficient, and byproducts, such as corncobs, stalks, and leaves, are valuable as animal feed, energy source, or fertilizer. About 66 % of corn production is used for food and feed, and ca. 5 % is used to make alcohol (⟶ Cereals – Maize (Corn)).

A number of batch and continuous processes have been developed for production of ethanol from corn. A conventional fermentation plant producing 76×10^3 m^3 of anhydrous ethanol per year from 816.5×10^3 kg of corn per day is shown in Figure 23 .

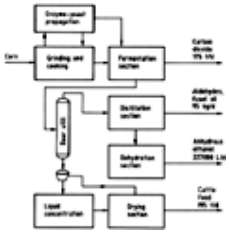

Figure 23. Flow diagram for a conventional fermentation plant producing anhydrous ethanol from corn [280]

In this process, corn is ground and cooked to dissolve and gelatinize the starch. The enzymes α-amylase and glucoamylase are then added to hydrolyze the starch to fermentable monosaccharides. After yeast fermentation for ca. 48 h at 32 °C, about 90 % of the starch is converted to ethanol. The fermentation broth is fed to the beer still where alcohol (ca. 50 vol %) is distilled. Subsequent distillation produces 95 % alcohol, which can be further concentrated by azeotropic distillation using benzene. After centrifugation, the stillage is concentrated to ca. 50 % solids in a multiple-effect evaporator, further concentrated in a fluidized-bed, transport-type dryer to ca. 10 % moisture, and then used as such for animal feed. This feed contains all the protein originally present in the grain, plus the additional protein from the yeast, resulting in a product containing 28 – 36 wt % protein.

In addition to alcohol and cattle feed, the original 816.5×10^3 kg of corn yields 175×10^3 kg of CO_2 and 95 kg of byproduct aldehydes, ketones, and fusel oils.

Alltech developed a method for an integrated grain-processing – fermentation route [281]. The grain pretreatment step prior to fermentation is shown in Figure 24.

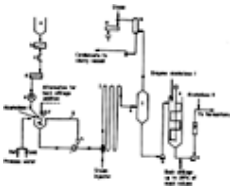

Figure 24. Alltech process for continuous whole mash cooking [281]

a) Grain hopper; b) Screen; c) Magnets; d) Continuous weigher; e) Hammer mill; f) Slurry vessel with agitator, temperature 50 – 70 °C; g) Rupture disk; h) Expansion vessel; i) Positive displacement pump; j) Continuous cooker tube, residence time 5 min, temperature variable up to 150 °C; k) Pressure valve; l) Flash vacuum cooler to 66 – 76 °C; m) Condenser; n) Open impeller pump; o) Converter, residence time 20 min, agitator 1 rpm; p) Wort cooler

Two enzymes, alcoholase I (from *Bacillus subtilis*) and alcoholase II (from *Aspergillus niger* and *Rhizopus niveus*) are used to hydrolyze the starch to fermentable sugars. Continuous whole mash cooking is applied. The ground starch is first mixed with water and alcoholase I at 60 °C, and then cooked at 93 – 165 °C in a batch or continuous cooker. The cooked mash is then cooled to ca. 66 – 76 °C, and a second portion of alcoholase I is added; 20 min is allowed for conversion. After this first hydrolysis step, the temperature is adjusted to 32 °C and the mash, supplemented with alcoholase II, is fermented with yeast.

Cassava. Cassava, also known as manioc, mandioc, aipum, yuca, cassada, and tapioca, is second in importance only to the sweet potato as a root crop throughout the tropics and in parts of South America where the plant originated. It was taken to West Africa by the Portuguese around 1914, where it now seems to have replaced yams and cocoyams because it adapts

easily and requires less labor than other crops. Cassava is one of the highest yielding plants of the vegetable kingdom (10 – 30 t/ha); it requires little cultivation and the tubers can be left in the ground until required without serious deterioration.

Cassava (genus *Manihot*) is in the family Euphorbiaceae, which belongs to the subdivision Angiospermae, class Dicotyledoneae, order Geraniales. This large, widely spread family comprises 283 genera including 7300 species, with an almost worldwide distribution [282].

Manihot esculenta, *M. utilissima,* and *M. dulcis* are some economically important members of a genus which includes over 150 species that are distributed throughout tropical countries. The species include herbs, shrubs, and trees, many of them producing latex and some yielding rubber. Brazil, Indonesia, and Zaire are the largest producers of cassava.

Roots are generally of interest for alcohol production. They contain 20 – 35 wt % starch and 1 – 2 wt % protein, although strains with up to 38 % starch have been developed [283]. The advantages of cassava for fuel ethanol production (which can amount to 7600 L/ha) have been assessed [267], [284], [285]. The process used to obtain ethanol from cassava is shown in Figure 25.

Figure 25. Production of ethanol from cassava root [267]

Fresh roots are washed, peeled, and ground into a mash. Part of this mash is dried; it can be stored in this form up to a year and is used for animal feed. For fermentation to ethanol, the starch is hydrolyzed with α-amylase, which is added in two steps. The first addition decreases the vicosity of the mash and facilitates cooking. In the second addition, the enzyme completes liquefaction of the starch. After that glucoamylase is added, which converts the liquefied starch to glucose and prepares the mash for fermentation. The fermentation process is the same as the one used for production of alcohol from sugarcane.

Alcohol yield from cassava is 165 – 180 L/t, which, on a mass basis, is higher than that obtained from sugarcane [285]. However, because sugarcane production can be as high as 90 t/ha, the alcohol yield per unit area is greater from cane under present cultivation conditions. Another advantage of cane is its dry fiber content, which equals the amount of total sugar present. This amount of fiber (bagasse) is sufficient to maintain the energy requirements of the plant; this is not the case with cassava, which only contains ca. 3.5 % fiber. Another disadvantage of cassava is that it does not contain readily fermentable sugars and, therefore, requires considerable processing of the roots prior to fermentation.

Sweet Sorghum. Sweet sorghum (*Sorghum sacchartum*) contains both starch and sugar. Its yield of ethanol from fermentable sugars is ca. 3500 – 4000 L/ha; an additional 1600 – 1900 L/ha can be produced from stalk fibers. There are more than 17 000 varieties of sorghum, and the yield is anticipated to increase by 30 % with some new hybrids [286]. The plant is adaptable to most of the world's agricultural regions; it is resistant to drought, and its nutrients are efficiently utilized by animals.

The fermentable sugars and starches are treated conventionally for ethanol production. The free sugars are fermented directly, whereas the starches are hydrolyzed by use of amylases, as is the case with cassava.

Potato. The potato is a common starch crop worldwide (⟶ Starch). The potato originated in South America (Chile and Peru) and came to Europe through Spain at the end of the 16th century. It is now grown in almost all climates and almost all types of soil, including dry and sandy soil [287].

Starch is the main carbohydrate component of potato (ca. 68 – 80 %). Depending on cultivation and variety of potato, starch content can vary between 12 and 21 % in raw potatoes. Only small quantities of soluble sugars are present (0.07 – 1.5 % sucrose, glucose, and fructose), as well as some rubber and dextrins (0.2 – 1.6 %) and pentosans (0.75 – 1.00 %).

The production of ethanol is based on fermentation of the available starch. A process developed by Danish Distilleries is shown in Figure 26 [268]. The process is semicontinuous and is applicable to both potatoes and grain.

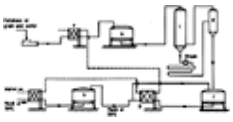

Figure 26. Danish Distilleries semicontinuous production of alcohol from potatoes or grain [268]

a) Preheater; b) Pulper; c) Enzyme treatment vessel; d) Flash cooler; e) Boiler tube; f) Holding tank; g) Condenser; h) Liquefaction vessel

Potatoes are mashed and then treated with amylases to hydrolyze the starch. The treatment section involves rapid steam treatment at 150 °C for ca. 3 min. The mash is cooled to 70 °C for liquefaction with commercial amylase preparations of bacterial origin; it is then cooled further to 30 °C and used for alcohol fermentation in the customary manner.

Jerusalem Artichoke. The Jerusalem artichoke (*Helianthus tuberosus*) is a member of the Compositae family and is closely related to the sunflower (*Helianthus annus*), earning it the nickname “wild sunflower”. About 102 different names are synonymous with the name *H. tuberosus*.

The plant is native in North America. It was originally grown by the Cree and Huron Indians who called it askipaw and skibwan, respectively. The plant was introduced to Europe at the beginning of the 16th century where it rapidly spread through the Mediterranean countries. The addition of “Jerusalem” to the name is most likely the result of an English version of *Girasole*, the Italian name for this plant [288].

The plant grows 1.5 – 2.5 m tall for wild strains and up to 3.7 m under cultivation. Top growth accounts for 40 – 56 % of the total plant biomass. The tubers are of greatest interest as a raw material for fermentation to ethanol.

The main soluble carbohydrate in the Jerusalem artichoke is inulin, which is composed of a homologous series of polyfructofuranose units. These units consist of linear chains of D-fructose molecules joined by β-2,1-linkages. The chains are terminated by a D-glucose molecule linked to fructose by an α-1,2-bond as in sucrose [289].

A process to produce 360×10^3 kg/a of ethanol from Jerusalem artichoke tubers is presented in Figure 27 [290].

Figure 27. Production of ethanol from Jerusalem artichoke tubers [290]

In this process, the juice is expressed from the tubers and extracted with water to obtain a carbohydrate concentration of about 20 %. The carbohydrates (predominantly inulin) are then hydrolyzed enzymatically by activating the endogenous inulinases at ca. 50 – 60 °C; acid hydrolysis (pH ca. 1) of inulin is also effective. The resulting fermentable sugars are then converted to ethanol by a conventional route.

Novel routes for conversion of the juice to ethanol have also been explored; the flocculating yeast *Saccharomyces diastaticus* has been used in semicontinous and continuous modes [234].

5.4.3. *Lignocellulosic Materials*

Lignocellulose is the largest terrestrial source of biomass that is renewably produced through photosynthesis (⟶ Biomass Chemicals; ⟶ Cellulose – Cellulose; ⟶ Lignin). The solar energy reaching the earth surface is 3.67×10^{21} kJ/a [291]. Gobal photosynthesis (with an efficiency of 0.07 %) could convert 2.57×10^{18} kJ of that energy to cellulose-containing biomass. This would result in a net production of 1.8×10^{11} t/a of biodegradable material, 40 % of which is cellulose [292]. Estimates are that 1 – 1.25×10^{11} t/a of terrestrial dry mass is produced together with 0.44 –0.55×10^{11} t/a in the oceans [293]. Present removal of this potential energy source is ca. 0.5 % of the total growing stock on a global basis [294].

The fermentation potential of lignocellulose is based mainly on the cellulose content of the biomass. Chemically, cellulose is similar to starch. It is a polymer of glucose in which the glucose units are linked by β-1,4-glucosidic bonds, whereas the bonds in starch are predominantly α-1,4-linkages. The degree of polymerization (DP) varies for different sources of cellulose; for example, newsprint cellulose has a DP of 1000, whereas cotton has a DP of ca. 10 000 [295].

The cellulose molecule is more resistant to hydrolysis compared to starch. This resistance is due not only to the primary structure based on glucosidic bonds, but also, to a great extent, to the secondary and tertiary configuration of the cellulose chain, as well as its close association with other protective polymeric structures such as lignin, starch, pectin, hemicellulose, proteins, and mineral elements.

The lignin molecule seems to be primarily responsible for difficulties in hydrolyzing the lignocellulosic material, because it forms a protective sheath around the cellulose microfibrils. Lignin is a macromolecule of phenolic character and can be viewed as a dehydration product of three monomeric alcohols: *trans*-4-coumaryl alcohol, *trans*-coniferyl alcohol, and *trans*-sinapyl alcohol. The relative amount of each varies with the source [296].

When cotton cellulose is treated with dilute acid, partial hydrolysis occurs rapidly, and ca. 15 % of the cellulose chain is degraded to glucose. The remaining 85 % is more resistant to hydrolysis, possibly because this portion of the cellulose exists in a highly crystalline order [295].

In order to use lignocellulosic materials for fermentation to ethanol, they must be pretreated and then hydrolyzed to fermentable sugars. Pretreatment may be physical or chemical, e.g., milling, steam explosion, or use of solvents and various swelling agents.

In *steam explosion* green wood chips are heated to ca. 180 – 200 °C for 5 – 30 min in a continuous operation (Stake process), or to a temperature of 245 °C for 0.5 – 2 min (Iotech process) [297]. The acids formed from hemicellulose under these high-temperature and high-pressure conditions start to “autohydrolyze” the cellulose and intact lignin. Lignin is sufficiently softened at the end of the steaming period, so that when the vessel is suddenly depressurized to atmospheric pressure, an explosion occurs within the woody cells. This partially disrupts the close association of cellulose with lignin and consequently increases the surface area available for further hydrolysis. The effect of steam pretreatment on the enzymatic hydrolysis of various cellulose-containing materials is shown in Table 18.

The pretreated lignocellulosic material is then subjected to further hydrolysis, which can be acidic or enzymatic (⟶ Cellulose – Cellulose; ⟶ Enzymes – Cellulases). A comparison of enzymatic and acid hydrolysis for cellulose degradation is given in Table 19. Cellulose and its degradation products are the only materials considered for fermentative purposes.

Table 18. Effect of steam pretreatment on the enzymatic hydrolysis [*] of cellulosic substrates [298]

Substrate	Pretreatment	Total reducing sugars, mg/mL	
		4 h	24 h
Hardwoods			
Poplar	none	1.4	2.4
	steam	15.3	25.8
Aspen	none	1.8	3.0
	steam	12.8	24.8
Agriculture residues			
Corn stover	none	4.9	7.8
	steam	15.7	22.5
Sugarcane bagasse	none	1.7	2.5
	steam	9.5	16.1
Urban waste	none	10.5	18.0
	steam	6.2	10.8
Softwoods			
Eastern spruce	none	2.0	3.8
	steam	3.5	6.4
Douglas fir	none	1.6	3.2
	steam	2.8	4.3

[*] *Trichoderma reesei* cellulase (QM9414), 19 IU (International Units) per gram of substrate; 5 % substrate slurries, pH 4.8, 50 °C; steamed substrates washed prior to enzymatic hydrolysis.

Table 19. Comparison between enzymatic and acid hydrolysis of cellulosic materials [227]

Acid	Enzyme
Nonspecific catalyst; therefore, will delignify material as well as hydrolyze cellulose	Specific macromolecular catalyst; therefore, extensive physical and chemical pretreatment of material necessary to make cellulose available for degradation
Decomposes hemicellulose to inhibitory compounds (i.e., furfural)	Produces clear sugar syrup ready for subsequent anaerobic fermentation
Harsh reaction conditions necessary; therefore, increased cost of heat- and corrosion-resistant equipment	Run under mild conditions (50 °C, atmospheric pressure, pH 4.8)
High chemical cost requires catalyst recovery and reuse	Cost of producing cellulases is the most expensive process step; therefore, recycle is necessary
High rate of hydrolysis	Lower rate of hydrolysis
Low overall yield of glucose because of degradation	High glucose yield depending on system and pretreatment

An example of a semicontinuous process for ethanol production from wood is shown schematically in Figure 28; this process uses dilute sulfuric acid for cellulose prehydrolysis.

Figure 28. Ethanol production from wood [299]

Optimum conditions for this process are: acid concentration in total water, 0.53 %; maximum temperature of percolation,

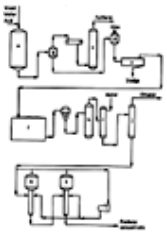

196 °C; rate of temperature rise, 4 °C/min; percolation time, 145 –190 min; ratio of total water to oven-dried wood, 10; percolation rate, 8.69 – 14.44 L min^{-1} m^{-3}

a) Digester; b) Flash evaporator; c) Furfural tower; d) Neutralization vessel; e) Clarifier; f) Fermentor; g) Yeast separator; h) Alcohol stripper; i) Extraction tower; j) Rectifying tower; k) Evaporator; l) Vapor compressor

The hydrolysate percolates through a bed of wood chips. Optimum conditions for the process are described in [299]. After acid digestion (a), the effluent passes through a flash evaporator (b), which separates the vapors containing furfural and methanol from the underflow condensate containing the sugar solution. The acid hydrolysate solution is further neutralized with a lime slurry, and the precipitated calcium sulfate is separated in a clarifier (d) as a 50 % solids sludge. The neutralized liquor is blended with recovered yeast (*Saccharomyces cerevisiae*) from previous fermention and is fermented to ethanol (e), which is further concentrated to 95 % by distillation (i).

The bottom material from the alcohol stripping column (g), which contains pentose sugars, is further concentrated in multiple-effect evaporators (j) to a 65 % solution, that can be used as animal feed or for chemical conversion to furfural.

Figure 29 shows a strong acid hydrolysis process.

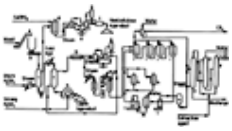

Figure 29. Ethanol production from wood by use of strong acid hydrolysis [300]

a) Feed hopper; b) Feeder; c) Digester; d) Neutralization vessel; e) Multiple-effect evaporators; f) Dryer; g) Electrodialysis membrane; h) Filter; i) Fermentor; j) Carbon dioxide scrubber; k) Seed fermentor; l) Centrifuge; m) Yeast wash vessel; n) Surge vessel; o) Beer still; p) Alcohol column

The air-dry wood is first pretreated with dilute sulfuric acid (c). Complete hydrolysis is accomplished in a subsequent strong acid cycle in which cellulose is hydrolyzed at room temperature with 70 – 80 % H_2SO_4. The glucose, retained by the dialysis membrane (g), is neutralized, deionized, and then sent to fermentation (i). The sulfuric acid permeate from the dialysis unit is evaporated and reconcentrated for recycle. Lignin is separated from the concentrated acid by filtration (h) and washing.

To illustrate enzymatic hydrolysis of cellulose for alcohol production, a process used by the Natick Development Center is shown in Figure 30. A large part of this process involves the preparation of the cellulase enzyme. Newspaper is the cellulose-containing substrate.

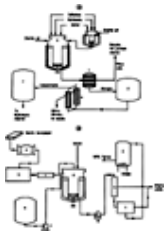

Figure 30. Enzymatic hydrolysis of newsprint by Natick Development Center (NDC) [298]

A) Pilot plant process for cellulase production:

a) Production vessel (vertical filters); b) Inoculum vessel; c) Filter; d) Harvest storage; e) Ultrafilter; f) Concentrate storageB) Pilot-plant process for newspaper hydrolysis: a) Ball mill; b) Solids metering; c) Solids transfer; d) Bioreactor; e) Enzyme storage; f) Metering pump; g) Harvest pump; h) Crude filter; i) Polish filter; j) Evaporator

5.4.4. *Waste Materials and Residues*

Various types of agricultural, industrial, or municipal refuse and waste can be used as substrates for ethanol fermentation. The fermentation is based on available sugar, starch, or cellulose in the waste material. The major advantage of this route lies in coupling waste treatment with the production of a higher value product. Both environmental pollution abatement and process economics are thus improved.

Cornstalks. Cornstalks are available in large quantities as a byproduct of corn agriculture. This material is predominantly composed of lignocellulose. A two-stage process using dilute acid treatment followed by concentrated acid impregnation of the lignocellulosic material is shown in Figure 31.

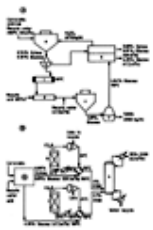

Figure 31. Production of ethanol from cornstalks [301]

A) Acid hydrolysis: a) Prehydrolysis tank, 4.4 % H_2SO_4; b) Filter; c) Rotary dryer; d) Impregnator; e) Hydrolysis tank, 8.0 % H_2SO_4; f) Filter; g) Electrodialysis unit

B) Fermentation: A) Acid hydrolysis process; h) Fixed film of *Fusarium oxysporum*; i) Centrifuge; j) Fixed film of *Saccharomyces cerevisiae*; k) Distillation column

In this process, ground corn stover (841 nm, 20 mesh) is treated with 4.4 % H_2SO_4 at 100 °C for 50 min (a). The mixture is then filtered (b) and the xylose-rich liquid is processed by electrodialysis (g) for acid recovery. The solids are dried further (c)

and impregnated with 85 % H_2SO_4 (d), followed by dilution with water to give a H_2SO_4 concentration of 8 % (e). Subsequent hydrolysis is carried out at 110 °C for ca. 10 min, and acid is again recovered by electrodialysis. The combined yield of xylose is 94 %, and the yield of glucose is 89 %. Glucose is converted to ethanol by *Saccharomyces cerevisiae* (j) and xylose by *Fusarium oxysporum* (h), both in immobilized cell reactors. The overall annual capacity of the plant is 17×10^3 m^3.

Domestic Refuse. Domestic refuse contains a complex variety of materials that come mainly from cellulosic-type residues. A large quantity of this material, ca. 1.3 – 2.2 kg per person, is generated daily. A process for the production of 36.5 t/d of ethanol from domestic refuse is illustrated in Figure 32 [302].

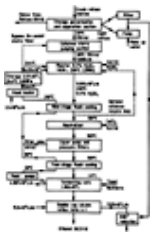

Figure 32. Flow diagram for continuous production of ethanol from refuse with 60 % content of cellulose [302]

* Biochemical oxygen demand.

The refuse is separated into dense and light fractions by the use of a flotation separator or a special pulper. The pulped fraction, which contains cellulose, is first subjected to removal of fines and plastics, and then introduced into a reactor where it is hydrolyzed with 0.4 % H_2SO_4 for ca. 1.2 min at 230 °C. This process is followed by flash cooling, neutralization with $CaCO_3$, and filtration. Fermentation of the sugar solution takes ca. 20 h at 40 °C and yields ca. 1.7 % aqueous ethanol solution, which is further concentrated by distillation to ca. 95 % ethanol.

Waste Liquor from the Pulp and Paper Industry. Two chemical pulping methods are predominant in the pulp and paper industry: the sulfate (Kraft) and the sulfite processes (⟶ Paper and Pulp). The basis of the these operations is treatment of the lignocellulosic material (wood, straw, etc.) with highly concentrated acid or base, which should dissolve the lignin portion of the wood and leave cellulose fibers that are processed into the final paper product. Depending on conditions (temperature, pressure, concentration of chemicals, chemical to wood ratio, and time of digestion), more or less delignification and breakdown of the original cellulose occur. As a result, a product pulp is produced as well as a waste chemical liquor, which basically consists of spent cooking chemicals. The more drastic the delignification conditions (low yield process), the better is the quality of the paper obtained. The high-yield process refers to milder delignification and a pulp that still contains a considerable amount of lignin. Low-yield processes are characterized by waste liquors with a high concentration of chemicals and a higher organic content.

The organic content of these liquors is primarily sulfonated lignin (e.g., 43 % of organic dry substance in a spent spruce sulfite liquor). However, cellulose and hemicellulose are also partially hydrolyzed during digestion so that waste liquors contain a certain proportion of free sugars (hexoses and pentoses in ca. 2 – 4 % concentration and ca. 14 % total solids) [303].

Tremendous quantities of waste liquor are generated in a pulp and paper mill; they amount to ca. 9180 L per ton of pulp produced [304]. Consequently, a chemical pulping process with a medium capacity of 500 t/d of pulp produces 4.59×10^6 L/d of waste liquors. Release of this liquor into natural waters is prohibited because both organic and toxic pollution result.

The sulfate (Kraft) process is designed so that the majority of the waste liquor can be recycled and its organic value converted to energy by combustion in a specially designed steam boiler (recovery furnace).

Previously, the majority of pulping mills worldwide were sulfite mills. Because of economic and environmental problems, sulfite pulping is gradually being phased out and the process converted to sulfate pulping or modified in other ways. Recovery of chemicals in the sulfite process is not as feasible as in the Kraft process, so large quantities of waste sulfite liquors (WSL) are discharged to the environment.

Because WSL contain fermentable sugars, these liquors have been used efficiently as fermentation substrates for alcohol production. The process is relatively old (1908 in Sweden) but is still in operation in some mills (e.g., the Ontario Paper Company, Canada). A typical process for fermentation of WSL is shown schematically in Figure 33.

Figure 33. Production of ethanol from waste sulfite liquors (WSL) [305] *

a) Digester; b) Blowpits; c) Storage; d) Stripper; e) Screen; f) Storage; g) Flash cooler; h) Barometric condenser; i) Ejectors; j) Fermentor; k) Yeast separator; l) Storage; m) Preheaters; n) Beer still; o) Rectifying column; p) Oil washer; q) Fusel oil; r) Purifying column; s) Vaporizer; t) Condenser; u) Alcohol; v) Heads

* Reprinted with permission of American Institute of Chemical Engineers.

The waste liquor is first stripped of SO_2 with a conventional steam stripper. This is necessary because SO_2 would inhibit subsequent fermentation. The liquor is adjusted to give a ca. 10 – 12 % concentration of sugars, the pH is adjusted to 4.5, and nitrogen and phosphorus nutrient sources are added (e.g., urea and phosphate). The fermentation is conventionally carried out with yeast (*Saccharomyces cerevisiae*) at 30 °C for ca. 20 h. The yeast is usually concentrated and recycled, and the broth containing ethanol is sent to the distillation section.

Cheese Whey. Whey is a byproduct of cheese production (⟶ Cheese, Processed Cheese, and Whey – Introduction, Processed Cheese, and Whey). An estimated 74×10^6 t of whey is produced annually worldwide. This amount of whey contains ca. 0.7×10^6 t of milk protein and 3.2 t of lactose [306].

Whey with its protein, carbohydrate, and vitamin content is a valuable, nutritious material; its composition is described in [307]. Whey is used in various forms as a component of either animal or human food. However, the utilization and recycling of whey nutrients depend on many factors, one of which is the size of the cheese factory. The smaller the factory, the less recycling of whey is practiced.

The alcohol produced from whey is derived mainly from the fermentation of available lactose. However, only a few microorganisms can convert lactose to ethanol, and conventional yeast (*Saccharomyces cerevisiae*) is not among them. The most efficient lactose-utilizing organism is reported to be *S. fragilis* [308].

A process using *Kluyveromyces fragilis* yeast was developed in Denmark (Fig. 34) [309-311]. The whey is first concentrated by reverse osmosis and ultrafiltration, and then introduced into fermentation vessels. The yield based on lactose is about 80 % of theoretical. About 42 L of whey, containing 4.4 % lactose, is required to produce 1 L of absolute ethanol.

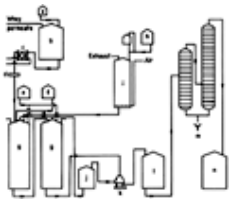

Figure 34. Continuous production of ethanol from whey [311]

a) Acid; b) Storage tank; c) Heat exchanger; d) Control; e) Antifoam; f) Chemicals; g) Fermentor; h) Substrate; i) Propagation plant; j) Storage; k) Separator; l) Buffer tank; m) Distillation; n) Alcohol storage

A better substrate for industrial fermentation of whey is enzymatically hydrolyzed lactose. β-Galactosidase-treated whey yields a mixture of monosaccharides, glucose, and galactose, which can be efficiently fermented by high-alcohol-producing yeasts [308], [312].

[Top of Page]

6. Recovery and Purification

Otto Goebel

Over the past 20 years a series of distillation systems have been developed for the efficient recovery of ethanol from synthetic and fermentation feedstocks. These units produce high-grade industrial alcohol, anhydrous alcohol, alcoholic spirits, and ethanol for motor fuels. Ethanol quality and recovery have been improved; energy consumption has decreased.

Distillation. Synthetic ethanol is purified in a simple three-column distillation unit (see Sections Production Process and Production Process). Recovery is 98 %, and the high-grade product contains less than 20 mg/kg of total impurities and has a permanganate time of over 60 min.

The following are key features of the efficient recovery of high-grade ethanol from fermentation feedstocks:

1. *Extractive distillation* results in a higher degree of purity than is possible in conventional purification columns; both investment and operating costs are reduced.
2. *Pressure-cascading installations* and *heat pumps* permit substantial heat recovery and recycling, thus minimizing heat loss and steam consumption. Virtually all (95 – 99 %) the ethanol in the crude feed is recovered as high-grade product.
3. *Advanced control systems* ensure stable operating conditions. Product quality can be maintained with a total impurity content of less than 50 mg/kg and a permanganate time of over 45 min.
4. *Energy requirements* are minimized. The flash heat recovered from the grain-cooking system is used to heat the ethanol distillation unit, thus reducing the energy consumption for ethanol production by ca. 10 %. Use of a vapor recompression technique can reduce the energy required for the evaporation of stillage to as little as one-tenth of that required in a triple- or quadruple-effect evaporator.

Dehydration. To produce anhydrous ethanol, the water – ethanol azeotrope obtained from distillation of the crude synthetic or fermentation feedstock must be dehydrated. For economic reasons, large distilleries rely mostly on azeotropic distillation for ethanol dehydration. Benzene has been used as an azeotropic dehydrating (entraining) agent in many plants, but some concern exists about its carcinogenicity and toxicity (⟶ Benzene – Toxicology and Occupational Health). However, proper design and control minimize benzene loss and exposure of operating personnel to this substance. Cyclohexane and ethylene glycol are used in some distilleries; they are also effective dehydrating agents.

Some smaller ethanol plants use molecular sieve adsorption techniques to dry the ethanol azeotrope. Pervaporation through semipermeable membranes or use of a solid dehydrating agent may reduce energy and equipment costs.

Concentration Units. Several systems are used to define ethanol concentration in addition to those based on conventional units (e.g., weight percent, volume percent).

Proof. In Canada, Great Britain, and the United States, the ethanol concentration of beverage spirits is expressed in terms of “proof”. United States law defines proof as follows: “Proof spirits shall be held to be that alcoholic liquor which contains one-

half of its volume of alcohol of a specific gravity of 0.7939 at 15.5 °C". More simply, the figure for proof is twice the percentage of ethanol content by volume. For example, 100 proof means 50 vol % alcohol.

In Great Britain as well as Canada, "proof spirit is such as at 10 °C weighs exactly twelve-thirteenths of the weight of an equal bulk of distilled water". A proof of 87.7 therefore signifies an ethanol concentration of 50 vol %. British proof can be converted into U.S. proof by multiplying by 1.11.

Degree Gay-Lussac. The degree Gay – Lussac (°GL) is measured with a hydrometer, which reads the volume percentage of ethanol in a mixture of ethanol and water at 15 °C.

Grades of Ethanol. Several grades of ethanol are available.

Industrial ethanol (96.5 vol %) is used for industrial and technical purposes as a solvent and a fuel, and is also converted into many other products. Industrial ethanol is usually denatured with 0.5 – 1 wt % of crude pyridine and is sometimes colored with 5×10^{-4} wt % methyl violet for easy recognition.

Denatured spirit (88 vol %) is a term used in some countries to describe industrial 88 vol % ethanol, which has been denatured and colored, and is generally used for heating and lighting.

Fine alcohol (96.0 – 96.5 vol %) is a purer type of ethanol used mainly for pharmaceutical and cosmetic preparations and for human consumption.

Absolute or anhydrous ethanol (99.7 – 99.8 vol %) is the term given to very pure, extremely dry ethanol, which is used in the food and pharmaceutical industries and also in the manufacture of aerosols. Anhydrous denatured ethanol mixed with 70 – 80 vol % of gasoline is used as a fuel for internal combustion engines. The specifications of anhydrous ethanol obtained by azeotropic distillation are given in Table 20.

Table 20. Specifications for anhydrous ethanol (99.6 vol %)

Parameter	Specification	Typical value	Test method
Purity, vol %	99.8 min.	>99.9	ETM 240.03 *
d_{20}^{20}	0.7911 max.	0.790	ASTM D 268
Acidity, wt % acetic acid	0.002 max.	0.001	ASTM D 1613
Color, Pt – Co	10 max.	5	ASTM D 1209
Nonvolatile matter, mg/100 mL	2.0 max.	1.0	ASTM D 1353
Distillation temperature, °C			
initial	77.5 min.	78.2	
dry point	79.0 max.	78.4	
Permanganate time, min	25 min.	45	ETM 110.77 *
Residual odor	nil	nil	ASTM D 1296
Appearance	clear and free of suspended matter	clear and free of suspended matter	ETM 80.65 *

* ETM = Ethanol testing method (CORA Engineering).

Motor fuel ethanol refers to fermentation ethanol that is used as an anhydrous or hydrous fuel or as a blending agent to improve the octane number of gasoline (⟶ Automotive Fuels).

6.1. Distillation

Ethanol is recovered as an azeotrope from ethanol – water mixtures by means of distillation [313], [314]. The boiling point diagram for this system is shown in Figure 35; the pure water azeotrope at the azeotropic point (a) contains 95.57 wt % (97.3 vol %) ethanol and has a boiling point of 78.15 °C.

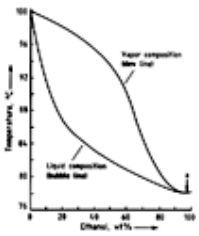

Figure 35. Boiling point diagram of ethanol – water mixtures

a) Azeotropic point

To obtain anhydrous ethanol, the pure ethanol – water azeotrope must be dehydrated. This is generally accomplished by azeotropic distillation with an entraining agent, usually benzene. The water is thus removed in the form of an overhead

ternary benzene – ethanol – water azeotrope.

6.1.1. *Distillation of Azeotropic Ethanol*

The crude ethanol synthesized by the direct or indirect hydration of ethylene contains ca. 50 vol % ethanol, whereas the crude product obtained after fermentation contains no more than 10 vol % ethanol. Consequently, the distillation of synthetic ethanol requires less energy and is less expensive than that of fermentation ethanol.

The distillation of synthetic azeotropic ethanol obtained by the direct and indirect hydration of ethylene is described in Sections Production Process and Production Process, respectively. The principles on which it is based are the same as those used in the distillation of fermentation ethanol described in this section.

The low-energy distillation of ethanol from a fermentation feedstock is shown schematically in Figure 36 [314-316]. Fermented mash containing ca. 10 wt % ethanol and ca. 10 wt % total solids is preheated to near saturation temperature, degassed to remove CO_2, and fed to the stripping column. The overhead product leaving the stripping column contains 75 – 85 vol % ethanol. The bottom liquid (stillage) contains less than 0.02 wt % ethanol and is sent either for disposal or for animal feed production. A vapor recompression system may be used to heat the reboiler of the stripping column for steam economy.

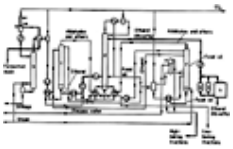

Figure 36. Distillation of 95 vol % ethanol from a fermentation feedstock

a) Degasser; b) Stripping column; c) Extractive distillation column; d) Rectifying column; e) Fusel oil washer; f) Concentrating column; g) Charcoal filtration; h) Storage tank

The overhead distillate from the stripping column is mixed with recycled ethanol from the concentrating column and fed into the extractive distillation column, which operates at a pressure of 0.6 – 0.7 MPa. This column removes essentially all fermentation byproducts, mainly aldehydes, ethers, methanol, and higher alcohols, from the ethanol. The aldehydes, ethers, and methanol are more volatile than ethanol and leave the top of the column. However, the higher alcohol byproducts, known collectively as fusel oil, are normally less volatile than ethanol. The average composition of fusel oil from fermentation ethanol derived from different raw materials is shown in Table 21 [317]. The technique used to extract fusel oil from ethanol exploits the fact that higher alcohols are more volatile than ethanol in solutions containing a high concentration of water. They can, therefore, also be steam distilled and removed in the overhead steam to leave a virtually pure ethanol – water mixture.

Table 21. Average composition of fusel oils (in wt %)

Fusel oil	1-Propanol [71-23-8]	1-Butanol (*n*-butyl alcohol) [71-36-8]	2-Methyl-1-propanol (isobutyl alcohol) [78-83-1]	2-Methyl-1-butanol (active amyl alcohol) [137-32-6]	3-Methyl-1-butanol (isoamyl alcohol) [123-51-3]
Molasses	13.2	0.2 – 0.7	15.8	28.4	37.4
Wheat – cereals	9.1	0.2 – 0.7	19.0	20.0	51.2
Potatoes	14.0	0.5	15.5	15.0	55.0
Sulfite waste liquor	7.0		22.0	13.0	55.0
Fruit	8.0	2.0	19.0	14.0	57.0

Dilute ethanol is then sent from the bottom of the stripping column to the rectifying column where it is brought up to strength. The rectifying column is heated by overhead vapors from the extractive distillation column and the concentrating column. The ethanol (95 vol %) is withdrawn as a side stream from one of the upper trays; it is then filtered through charcoal and stored. The water that collects at the bottom of the rectifying column contains traces of ethanol and is recycled to the extractive distillation column. Fusel oil is withdrawn as a side stream from one of the lower trays of the rectifying column and fed to the fusel oil washer.

In the concentrating column, overhead vapor from the extractive distillation column, which contains aldehyde, ether, and alcohol impurities, is separated into low-boiling and high-boiling fractions. Ethanol is also recovered and recycled to the extractive distillation column. A small stream is taken from the overhead condensate (low-boiling fraction), which contains acetaldehyde and a small amount of ethanol. It may be sold as a byproduct or burned as a fuel. A side stream containing a high concentration of fusel oil is sent through a cooler to the fusel oil washer. In the washer, ethanol is extracted from the fusel oil with water, and the washings are recycled to the concentrating column. High-boiling fusel oil is also run off from the bottom of the concentrating column. The combined fusel oil (high-boiling) fractions may be sold as a byproduct.

6.1.2. *Dehydration by Azeotropic Distillation*

Azeotropic distillation systems are designed for the production of pure, anhydrous ethanol (99.98 vol %), which contains <200 mg/kg of water and <20 mg/kg of total impurities (for other specifications, see Table 20) [314-316]. A flow scheme for

the two-column azeotropic distillation of anhydrous ethanol is shown in Figure 37. The dehydrating column and the hydrocarbon stripping column operate at atmospheric pressure. Therefore, they may be heated with low-pressure steam, hot condensate, or hot waste streams from other parts of the ethanol processing plant, thus minimizing steam consumption.

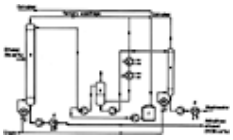

Figure 37. Production of anhydrous ethanol by azeotropic distillation

a) Dehydrating column; b) Decanter; c) Condenser; d) Cooler; e) Hydrocarbon stripping column; f) Entrainer tank

Water is removed from the ethanol – water azeotrope in the form of a ternary azeotrope which is produced by adding an entraining agent such as benzene, heptane (C_6– C_8 cut), or cyclohexane. The 95 vol % ethanol feed enters the dehydrating column near the midpoint. The anhydrous ethanol product collects at the bottom of the tower and is sent through a cooler prior to storage. The ternary azeotrope leaves the column as an overhead product which is condensed and then separated into an organic and an aqueous phase in the decanter. The upper organic layer containing the entrainer is returned to the top of the dehydrating column. The lower aqueous layer is pumped to the hydrocarbon stripper, where the hydrocarbon entrainer, the ethanol, and some water vapor are recovered overhead and sent to the condenser – decanter system. Water from the stripper is pumped off as waste; if it contains a substantial amount of ethanol, it may be recycled to the ethanol distillation unit.

The overall efficiency and reliability of the anhydrous ethanol system are the result of the following special features:

1. The use of a common condenser and decanter for the two columns reduces capital costs.
2. The use of very efficient column trays which can operate at a low throughput.
3. The BOD of the wastewater is low.
4. Consumption of the entraining agent is low.
5. Steam consumption is low: 1 – 1.5 kg is required per liter of anhydrous ethanol depending on the quality of the end product.

In some plants that produce motor fuel ethanol, gasoline is substituted for benzene as the entrainer. Since gasoline is required in the end product, the hydrocarbon stripping step is eliminated. However, gasoline dehydrating agents give products of varying composition because the gasoline contains a number of different components.

6.1.3. *Motor Fuel Ethanol*

Two types of ethanol for motor fuel are produced industrially [314-316], namely, anhydrous and hydrous motor fuel ethanol (⟶ Automotive Fuels).

Anhydrous Motor Fuel Ethanol. The distillation and dehydration of motor fuel ethanol from a fermentation feedstock are shown schematically in Figure 38.

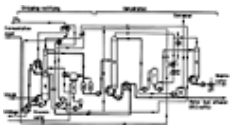

Figure 38. Production of anhydrous motor fuel ethanol

a) Degasser; b) Stripping – rectifying column; c) Fusel oil washer; d) Filter; e) Dehydrating column; f) Decanter; g) Entrainer tank; h) Entrainer stripping column; i) Cooler; j) Condenser

Stripping – Rectifying. Ethanol is distilled in a single stripping – rectifying column where the fermented mash is separated into an overhead, ca. 95-vol % ethanol fraction and a bottom liquid fraction (stillage) containing <0.02 wt % ethanol. The mash is preheated at its saturation temperature and degassed to remove residual CO_2 before it enters the stripping – rectifying column. The column operates at a pressure of ca. 0.3 MPa and is heated with steam by a forced circulation reboiler. This is the only part of the system that uses steam. The stillage leaving the bottom of the column is cooled to its boiling point at atmospheric pressure by heat exchange, the extracted heat being used to preheat the mash feed. The temperature of the resulting stillage is such that it can be subjected to vaporrecompression evaporation without preheating or flashing in an evaporator.

The pressurized overhead vapors from the stripping – rectifying column are used to preheat the mash feed and to heat the reboilers of the dehydration column and of the entraining stripping column. The condensed vapors are returned to the top of the stripping – rectifying column. A small overhead stream is drawn off to remove volatile acetaldehyde from the system; it contains <1 % of the product ethanol and is either added to the purified anhydrous ethanol product or burned as a fuel in the plant boiler.

A side stream containing fusel oil is also taken from the rectifying section of the column and sent to a fusel oil washer. The aqueous washings are returned to the stripping section of the column; the decanted, washed oil is combined with the ethanol product stream and fed to the dehydration column. Fusel oil has a higher energy value than ethanol and acts as an agent for blending ethanol with gasoline.

Dehydration. After leaving the stripper – rectifier, the 95 vol % ethanol and fusel oil are filtered and fed to a dehydrating column, which operates at atmospheric pressure. Here, water is removed from the feed by use of benzene, heptane (C_6– C_9 cut), cyclohexane, or some other entraining agent. The bottom stream from the dehydrating column consists of anhydrous, 99.5 vol % ethanol and is cooled prior to storage. The overhead fraction from the dehydrating column is a ternary azeotrope,

which is combined with overhead vapors from the entrainer stripping column. The combined vapors are condensed and cooled, forming two phases that are separated in a decanter. The upper entrainer layer from the decanter is pumped to the top of the dehydrating tower, while the lower aqueous layer is fed to the entrainer stripping column for recovery of the entrainer and ethanol.

Steam consumption in this system is 1.8 – 2.5 kg per liter of ethanol produced, depending on the alcohol concentration of the mash.

Hydrous Motor Fuel Ethanol. A low-energy system has been developed for distillation of the 85 – 95 vol % motor fuel ethanol used in engines that require neat alcohol and not gasoline. This distillation process is used mainly in Brazil; a flow scheme is shown in Figure 39. For maximum steam economy, the fermentation feed is fed into two stripper – rectifiers, the first of which operates at a pressure of ca. 0.4 MPa, the second at atmospheric pressure. Steam is used only in the high-pressure column, which takes 55 – 60 % of the fermentation feed. The steam consumption of this system is 1.2 – 1.5 kg per liter of 85 – 95 vol % motor fuel ethanol. The overhead and fusel oil byproducts are removed and processed in a fashion similar to that described for production of anhydrous motor fuel ethanol.

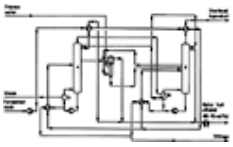

Figure 39. Distillation of hydrous motor fuel ethanol

a) High-pressure stripping – rectifying column; b) Reboiler; c) Fusel oil washer; d) Stripping – rectifying column (operating at atmospheric pressure)

6.1.4. *Reduction of Energy Costs*

Distillation of fermentation ethanol requires large amounts of energy. Energy costs have been cut by reducing steam consumption and by using vapor recompression systems.

Earlier columns for the distillation of ethanol from corn, wheat, or molasses were operated at atmospheric pressure. The development of new multiple-stage, high-pressure systems has reduced steam consumption by 40 % compared to previous systems. The new commercial installations are based on a pressure-cascading technique and consume 3.0 – 4.2 kg of steam for every liter of 96 vol % ethanol produced. Earlier distillation systems required 6 kg of steam per liter of ethanol.

A modern motor fuel ethanol plant has a total energy consumption of 1.1 – 1.4 MJ per liter of ethanol. Steam consumption figures for the systems described in Sections Dehydration by Azeotropic Distillation and Motor Fuel Ethanol are summarized in Table 22 [318-320].

Table 22. Steam consumption and ethanol concentration in the low-energy distillation of ethanol

Parameter	Distillation product			
	Ethanol – water azeotrope	Anhydrous ethanol	Anhydrous motor fuel ethanol	Hydrous motor fuel ethanol
Maximum ethanol concentration in the feed, vol %	10	96	10	10
Ethanol concentration of the product,				
vol %	96	100	99.5	95
proof (U.S.)	192	200	199	190
Steam consumption per liter of ethanol, kg	4.1	1.4	2.2	1.2

Vapor Recompression Systems. Energy costs can be reduced by up to 80 % in low-pressure distillation columns by using a vapor recompression system [321].

In this system, compressed overhead vapor is used as the heat source for the reboiler, instead of expensive steam. The temperature of the column overhead vapor is increased by compression.

Vapor recompression can be applied to distillation columns operating at or below atmospheric pressure. It is used mainly in the distillation of motor fuel ethanol. Investment costs for ethanol distillation units equipped with a vapor recompression system are almost 50 % higher than those for conventional distillation.

6.2. Nondistillative Methods

The energy required to remove water from ethanol can be reduced significantly by using methods that do not rely on distillation.

6.2.1. *Solvent Extraction*

Ethanol dissolves in some liquids that are virtually immiscible with water. These solubility differences can be exploited to recover ethanol from an aqueous solution by means of solvent extraction [322].

In the United States, the Energol Corporation employs liquid – liquid extraction with a proprietary solvent to separate ethanol from water. The solvent is then removed by distillation. This method does not require energy-intensive azeotropic distillation and thus has a low energy consumption. The energy budget for the entire plant is 3500 – 3700 kJ per kilogram of ethanol produced. In 1987, Energol's technique was used in four 40 000 to 45 000-L/d plants which came into operation in the mid 1980s.

The University of Pennsylvania and General Electric have developed a process that uses dibutyl phthalate [84-74-2] as a water-immiscible solvent for purifying ethanol. This solvent has a much higher boiling point than ethanol, and ethanol can therefore be separated in a single distillation step; solvent losses are low.

6.2.2. *Carbon Dioxide Extraction*

Another type of solvent extraction makes use of a critical fluid, i.e., a gas compressed to the point at which the distinction between gas and liquid disappears.

Ethanol is selectively extracted from grain mash with carbon dioxide close to its critical point of 7.3 MPa and 31 °C [321]. The ethanol-rich stream is flashed at ca. 4.8 MPa to remove the carbon dioxide, leaving ethanol separated in the liquid phase. This route requires a third to a half of the energy needed for conventional ethanol distillation, but the capital cost is 20 % higher. A further advantage of this method is that the carbon dioxide used as solvent is obtained as a low-cost byproduct of fermentation. This is important because some solvent inevitably escapes during ethanol recovery and has to be replaced.

6.2.3. *Adsorptive Dehydration*

Molecular Sieves. Ethanol azeotropes are dehydrated industrially by adsorption with molecular sieves whose pores are permeable to water but not to ethanol [323-325]. The molecular sieve may be a synthetic or naturally occurring zeolite (e.g., clinoptilolite) or a proprietary resin. The 95 vol % ethanol is dehydrated in molecular sieve columns; 75 % of the adsorbed material is water and 25 % ethanol. When the column is saturated, the stream is directed to a fresh column and the saturated column is regenerated. The regeneration stream containing 25 vol % ethanol is fed back to the ethanol distillation system.

Solid Agents. Ethanol can also be dehydrated by adsorption with solid agents [326]. Less energy is required to vaporize water from cellulose or corn-starch than from calcium hydroxide, because of their low heats of adsorption. Therefore, ground cornmeal is used as a dehydrating agent for removing water from an 85 % ethanol feed stream. The cornmeal adsorbent can be recycled ca. 20 times before being used as animal feed.

6.2.4. *Membrane Technology*

Pervaporation. A new method of ethanol purification based on pervaporation has been developed [327], [328] and is shown schematically in Figure 40. The pervaporator consists of a number of semipermeable membrane modules made of poly(vinyl alcohol) resins.

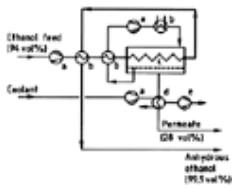

Figure 40. Production of anhydrous ethanol by pervaporation

a) Pump; b) Heater; c) Pervaporator; d) Condenser; e) Vacuum pump

The 94 vol % ethanol feed is preheated to 60 °C and pumped to the semipermeable membrane modules of the pervaporator. Water permeates the membrane down its concentration gradient; a phase change occurs from the liquid phase at the membrane inlet to the vapor phase in the permeate. Water is thus separated without azeotrope formation. The driving force for permeate flow is provided by a vacuum of less than 1 kPa at the permeate condenser inlet. The total energy consumption is the sum of the evaporation and the condensation enthalpies.

The condensed permeate contains a small amount of ethanol and can be recycled to a rectifying or distillation tower for recovery of ethanol. Pervaporation of 1096 L of a 91 wt % (94 vol %) ethanol feed yields 1000 L of anhydrous ethanol (99.85 wt %, 99.9 vol %) and 107.5 L of permeate byproduct containing 23 wt % ethanol. The production of 1000 L of anhydrous ethanol requires 135 kg of steam (200 kPa), 10 m^3 of cooling water (20 °C), and 15 kW · h of electricity.

Reverse Osmosis. Purification of ethanol by reverse osmosis employs membranes that are relatively impermeable to ethanol but permeable to water [329]. A pressure of 4 – 7 MPa is usually applied to remove the water by forcing it across the membrane. The ethanol retention of new noncellulosic membranes is much higher than that of the cellulose acetate membranes used earlier (80 % compared to 50 %). Reverse osmosis may prove useful for savings in energy costs by concentrating ethanol to about 10 % prior to distillation.

6.2.5. *Yarn-Filled Column*

Textile yarns, such as rayon, retard the movement of water vapor but allow ethanol vapor to pass. This phenomenon has been exploited to develop a technique for separating ethanol – water mixtures [330].

Water containing 12 wt % ethanol is vaporized by injecting it into a stream of air; the ethanol – water – air mixture is then sent through a yarn-filled column equipped with heating elements. A zone of water builds up at the beginning of the column; air and ethanol pass through and are cooled to condense the ethanol. Before the water zone moves too far into the column, the

direction of flow is reversed and the water is flushed out with air. The cycle takes about 25 s; flow occurs for 10 s in the forward direction and 15 s in the reverse direction.

A continuous version of the process is being developed in which the yarn is made into a continuous belt by tying the ends together. The belt moves slowly through the column, in the opposite direction to the flow of the ethanol – water – air vapor. Ethanol is then produced continuously at one end of the column. The yarn leaves the other end of the column and passes through a heating zone where water is driven off.

6.3. Storage and Transportation

Ethanol is shipped in railroad tank cars, 200-L tank trucks, 20-L drums, and smaller glass or metal containers. In the case of special quality requirements, the drums may be lined with phenolic resin.

The hazard classifications of ethanol are as follows:

GGVE/GGVS 3/3 b
IMDG Code 3.2, 3.3, UN No. 1170
United States CFR 49 : 172.101, flammable liquid

[Top of Page]

7. Comparison of Process Economics for Synthetic and Fermentation Ethanol

Otto Goebel

The figures and calculations in this chapter are based on data taken from [331-333].

7.1. Summary of Cost Analysis for Synthetic and Fermentation Ethanol

The cost of producing ethanol by ethylene hydration (synthetic ethanol) or fermentation (fermentation ethanol) depends primarily on the price of the raw material used. Fermentation ethanol became economically competitive with synthetic ethanol when crude-oil prices rose in the late 1970s and early 1980s, because ethylene is derived from refinery products. Production costs for fermentation ethanol may be further reduced by introduction of new processes and improved technologies that use raw materials such as wood, waste, and other low-cost feedstocks.

The feedstock price for synthetic ethanol production is high, but the capital cost requirement is low. Even if fermentation ethanol is produced from an inexpensive feedstock, the operating and capital costs are higher than those for synthetic ethanol.

The following analysis shows that the production costs for synthetic ethanol based on 1986 ethylene prices of $ 350/t are in the same range as those for fermentation ethanol produced from molasses costing $ 45/t.

Production figures are given in liters and metric tons. To convert liters to gallons, divide by 3.785; 1 t = 1260 L.

7.2. Production Costs of Synthetic Ethanol

The synthesis of ethanol from ethylene is divided into three processing steps: hydration, purification, and distillation (for details, see Chaps. Synthesis and Recovery and Purification). Approximately 87 major pieces of processing equipment are required.

Investment Costs. To calculate the production costs of synthetic ethanol, the number of installed pieces of major equipment and their costs were estimated in U.S. dollars based on 1986 cost levels. Figure 41 shows the estimated total investment costs for hypothetical plants and offsites (auxiliary equipment, primarily storage tanks) with capacities of 25 000, 50 000, 100 000, and 150 000 t/a. The breakdown of the investment for the main processing plant — the total inside-battery-limit investment — is as follows:

1. Equipment and machinery, including piping, insulation, painting, and electrical installations, 50 %
2. Construction and erection, 25 %
3. Civil engineering work, 10 %
4. Engineering, costs for equipment purchase and start-up, 15 %

License fees are considered separately.

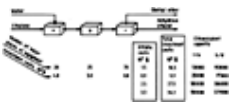

Figure 41. Investment costs for a plant producing synthetic ethanol

a) Hydration; b) Purification; c) Distillation

Production Costs. The annual costs (in 10^6 dollars) of synthetic ethanol produced in a 50 000-t/a plant from ethylene costing $ 350/t are as follows (with the assumption that 0.625 t of ethanol can be obtained from 1 t of ethylene):

Raw materials	**11.26**	**(59 %)**
Utilities	**4.18**	**(22 %)**

Labor	**0.24**	**(1 %)**
Depreciation, interest, maintenance	**3.59**	**(18 %)**
Total production costs	19.27	(100 %)

Payback Time. The payback time (in years) for a 50 000-t/a plant can be calculated by using the following figures (given in 10^6 $/a, with the assumption that ethanol sells at $ 450/t and that ethylene costs $ 350/t):

Annual turnover	22.50
Total production costs	19.27
Difference	3.23
Depreciation	1.65
Cash flow	4.88

$$\text{Payback time} = \frac{\text{Investment}}{\text{Cash flow}} = 3.4 \text{ years}$$

For a plant with double this capacity (100 000 t/a), the payback time is reduced from 3.4 to 2.5 years. Figure 42 shows the payback times for synthetic ethanol production in a 50 000-t/a plant as a function of the selling price for ethanol at various ethylene costs. The 1986 prices for naphtha and gasoline are indicated for comparison.

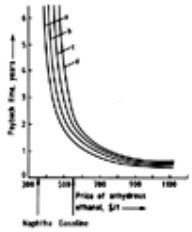

Figure 42. Price of anhydrous ethanol as a function of the payback time in a synthetic alcohol plant

Cost of ethylene (per ton): a) $ 300; b) $ 350; c) $ 400; d) $ 450Production capacity: 50 000 t/a (19 300 L/d)

7.3. Production Costs of Fermentation Ethanol

The processing scheme of ethanol fermentation from regenerable resources is divided into five major steps (for details, see Chaps. Fermentation and Recovery and Purification):

1. Handling of raw materials
2. Hydrolysis pretreatment
3. Fermentation
4. Distillation and dehydration
5. Processing of stillage and production of optional byproducts

Approximately 140 major pieces of process equipment are required for a molasses plant and almost 200 for a wood saccharification plant.

Investment Costs. Figure 43 shows the estimated investment costs for a fermentation-based ethanol plant with a capacity of 52 000 t/a (200 000 L/d); costs are given for each processing step and for the entire processing plant and offsites. The total investment costs are indicated for plants using sucrose, starch, or cellulose-containing raw materials such as molasses, cassava/manioc, or wood.

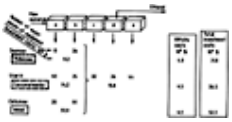

Figure 43. Investment costs for a plant producing fermentation ethanol from molasses, cassava, or wood *

a) Handling of raw materials; b) Hydrolysis pretreatment; c) Fermentation; d) Distillation and dehydration; e) Stillage processing

* Production capacity: 52 000 t/a (200 000 L/d).

The breakdown of the inside-battery-limit investment is similar to that for synthetic ethanol plants.

Production Costs for Three Different Raw Materials. The annual production costs of fermentation ethanol produced in a 52 000-t/a (200 000-L/d) plant from three basic raw materials (sugarcane molasses, cassava, and eucalyptus wood) are listed in Table 23.

Table 23. Production costs of fermentation ethanol obtained from sugarcane molasses, cassava, and eucalyptus

	Cost, 10^6 $/a		
	Molasses	Cassava	Eucalyptus wood
Raw materials	9.27	14.77	7.19
Utilities	3.77	7.66	12.32
Labor	0.76	1.10	1.18
Depreciation, interest, maintenance	7.06	7.75	11.58
Total production costs	20.86	31.28	32.27

Both wheat and corn (maize) are used by large ethanol-producing companies in the corn-belt region of the United States. The economics of ethanol production from these raw materials are similar to those of molasses-based ethanol production, but are strongly influenced by byproduct utilization and market values. The wet-milling process facilitates the production of high-quality ethanol, and retains the corn oil and corn protein in the food chain. In this process, the corn kernel is separated into starch, corn oil, gluten, and fiber, which are converted into a variety of other products. Utility costs are reduced by the use of coal-based energy. In some plants, the enzymatic hydrolysis is designed so as to permit production to be switched seasonally from ethanol to high-fructose corn syrup.

Payback Time. The payback time (in years) can be calculated by using the following figures (given in 10^6 $/a, with the assumption that ethanol sells at $ 450/t and that molasses cost $ 45/t):

Annual turnover	24.90
Total production costs	20.86
Difference	4.04
Depreciation	3.18
Cash flow	7.22

$$\text{Payback time} = \frac{\text{Investment}}{\text{Cash flow}} = 4.4 \text{ years}$$

If the selling price of ethanol is higher, e.g., $ 550/t, the payback time will be reduced to 2.8 years.

Figure 44 shows the payback times for fermentation ethanol production as a function of the selling price for ethanol. Sugarcane molasses, cassava, and eucalyptus wood have been considered as raw materials, their costs being $ 45, $ 30, and $ 15 per ton, respectively. In the early 1970s, synthetic ethanol was less expensive than fermentation ethanol because crude-oil prices, and thus ethylene prices, were low. Increases in the price of oil and, therefore, ethylene were accompanied by drops in the price of corn and sugarcane molasses; this meant that fermentation ethanol was less expensive to produce than synthetic ethanol. Fermentation ethanol can be expected to maintain a slight price advantage over synthetic ethanol as long as the raw material prices remain low. If ethanol is blended with gasoline for use as a motor fuel, i.e., gasohol (⟶ Automotive Fuels), a payback time of 3 – 4 years is acceptable. The minimum selling prices for ethanol which would economically justify its production from molasses, cassava, and eucalyptus wood would therefore be $ 500, $ 675, and $ 775 per ton, respectively (see Fig. 44). In 1986, Western European prices were $ 300 – 400 per ton for raw naphtha and $ 500 – 600 per ton for gasoline from the refinery (tax not included).

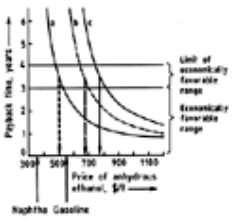

Figure 44. Price of anhydrous ethanol as a function of the payback time in a plant * producing fermentation alcohol

Raw material and price (per ton): a) Sugarcane molasses, $ 45; b) Cassava, $ 30; c) Eucalyptus wood, $ 15

* Production capacity: 52 000 t/a (200 000 L/d).

Stillage Processing Options. If stillage is used directly as a fertilizer without further processing (as in developing countries), the economic feasibility of fermentation ethanol production is increased by about 15 %.

Stillage can also be processed further in four main ways:

1. Multiple-effect evaporation, drying of the concentrated stillage, and production of distillers' dry grain (a byproduct, which is used as an animal feed)
2. Mechanical vapor recompression, syrup evaporation, drying of the concentrated stillage, and production of distillers' dry grain
3. Evaporation and syrup combustion
4. Anaerobic digestion with generation of biogas

Analysis of the economic merits and drawbacks of these four schemes with regard to the production of fermentation ethanol reveals that system (2) is the most advantageous; systems (1), (3), and (4) are less favorable, but equal. The disadvantage of schemes (1) and (2) is the unstable selling price of distillers' dry grain. Scheme (3) is expensive because an off-gas purification system is required. The advantage of scheme (4) is the production of energy in the form of biogas.

[Top of Page]

8. Analysis

Naim Kosaric and Zdravko Duvnjak

The classic *physical methods* for the determination of ethanol concentration in a sample are based on measurements of relative density (with a calibrated pycnometer, hydrometer, or density meter) [334]; boiling-point depression of alcohol – water mixtures (using an ebulliometer) [335]; and refractive index [334], [336]. In addition, gas chromatography is quick and reliable. Various internal standards, such as acetone, ethyl acetate, *n*-propanol, 2-propanol, and ethylene glycol monoethyl ether, have been used [334], [337], [338].

Chemical methods include oxidation to acetic acid by dichromate in the presence of sulfuric acid and titration of unreacted dichromate with ferrous ammonium sulfate (sodium diphenylamine 4-sulfonate or 1,10-*o*-phenanthroline are added as indicators):

$$2\,Cr_2O_7^{2-} + 3\,C_2H_5OH + 16\,H^+ \rightarrow 4\,Cr^{3+} + 3\,CH_3COOH + 11\,H_2O$$

$$Cr_2O_7^{2-} + 6\,Fe^{2+} + 14\,H^+ \rightarrow 2\,Cr^{3+} + 6\,Fe^{3+} + 7\,H_2O$$

The pH of the reaction is critical because ethanol could be oxidized either to acetaldehyde or to a mixture of acetaldehyde and acetic acid [335], [339], [340]. If the ethanol – water mixture is denatured, other compounds can interfere.

The chemical purity of ethanol is often determined using acetylation or phthalation reactions in which ethanol reacts with a defined amount of either acetic or phthalic anhydride in pyridine solution. The decrease in acidity of the anhydride solution (after hydrolysis and comparison with a standard solution) corresponds to the reaction of the hydroxyl groups. Any reactive hydroxyl group can take part in these reactions, which are not specific for ethanol. Other functional groups do not interfere [341].

Trace amounts of ethanol can be determined by colorimetric methods using compounds such as 8-hydroxyquinoline or vanadic acid, which form colored complexes with alcohols [342].

In addition to physical and chemical methods, a *biochemical method* is used to determine ethanol concentration [343]. This method is based on the oxidation of ethanol to acetaldehyde by nicotinamide – adenine dinucleotide (NAD^+). The enzyme alcohol dehydrogenase catalyzes the reaction:

$$CH_3CH_2OH + NAD^+ \rightarrow CH_3CHO + NADH + H^+$$

To bring the reaction to completion, semicarbazide, (aminooxy)acetic acid, or the enzyme aldehyde dehydrogenase is added to remove acetaldehyde from the reaction mixture. The reaction is followed spectrophotometrically at a wavelength of 340 nm where the absorbance is proportional to the increase in NADH concentration [335]. Despite the fact that some other alcohols can also react, the method is good and can be used to determine very small ethanol concentrations.

Some enzymatic methods use enzymatic probes with immobilized enzymes [344]. Electrode probes with whole microbial cells have also been developed for alcohol determination [344].

[Top of Page]

9. Uses

Naim Kosaric and Zdravko Duvnjak

Ethanol is an organic chemical with many applications, e.g.,

1. alcoholic beverage (⟶ Beer; ⟶ ; ⟶),
2. solvent,
3. raw material in chemical synthesis, and
4. fuel (⟶ Automotive Fuels).

In most countries, ethanol produced by fermentation has been used for beverages and specialty chemicals, and ethanol produced by chemical synthesis has been used for industrial purposes. Certain countries, such as Brazil and India, use fermentation ethanol for industrial purposes.

For many years, alcoholic beverages have been taxed worldwide. However, the price of taxed ethanol was too high for its use as an industrial raw material. To enable industry to obtain alcohol at a lower price, the Tax-Free Industrial and Denatured Alcohol Act of 1906 was passed in the United States. Similar laws have been enacted in other countries.

The U.S. government implemented financial, administrative, and chemical controls to prevent the utilization of tax-free ethanol in beverages. These regulations established the following four distinct categories of industrial ethanol:

1. Completely denatured alcohol
2. Proprietary solvents and special industrial solvents
3. Specially denatured alcohol
4. Pure (absolute) ethanol

Pure ethanol and slightly denatured ethanol are under strict control, whereas highly denatured ethanol has the fewest financial and administrative controls. A wide range of chemicals can be used for ethanol denaturation [345].

Solvent. Ethanol is the most important solvent after water. Its major commercial applications are in the manufacture of toiletries and cosmetics, detergents and disinfectants, pharmaceuticals, surface coatings, and in food and drug processing. Both synthetic and fermentation ethanol can be used for these purposes; however, fermentation ethanol is preferred (particularly in Europe) for applications involving human consumption (or body use) such as cosmetics, toiletries, and pharmaceuticals.

The amount of denatured ethanol used as a solvent is increasing. For example, in the United States 197 000 and 340 000 t of ethanol were used in 1960 and 1979, respectively. In Japan, between 1974 and 1978 the amount of ethanol consumed as a solvent increased by 28 % [346], [347]

Raw Material. Ethanol is also used to produce various chemicals (see Fig. 2):

1. Acetaldehyde (⟶ Acetaldehyde – Production)
2. Butadiene (⟶ Butadiene – Production)
3. Diethyl ether (⟶ Ethers, Aliphatic – Diethyl Ether)
4. Ethyl acetate (⟶ Acetic Acid – Ethyl Acetate)
5. Ethylamines (⟶ Amines, Aliphatic – Production from Alcohols)
6. Ethylene (⟶ Ethylene – Other Processes and Feedstocks)
7. Glycol ethers and other products formed by reaction with ethylene oxide or epoxides (⟶ Epoxides – Reactions with Compounds Containing Ionizable Hydrogen; ⟶ Ethylene Oxide – Chemical Properties; ⟶ Propylene Oxide)
8. Vinegar (⟶)

Table 24 shows the most important products derived from ethanol in the United States. The amount of specially denatured ethanol used as a raw material in the manufacture of chemicals was 37.6 % of the total amount of denatured ethanol used in the country during the specified period of time. In the United States, consumption of ethanol as a raw material reached a peak in 1960 (627 000 t) and then decreased substantially. The amount of ethanol used to manufacture chemicals in 1979 was only one-third of the amount used in 1960. The main reasons for this decline are the use of ethylene glycol rather than ethanol in antifreeze, replacement of ethanol by ethylene and ethane in the production of acetaldehyde, and production of ethylhexanol and butyraldehyde from other raw materials. Certain trends indicate that further decline in the use of ethanol to manufacture chemicals may be reversed [346].

In some countries, such as Brazil, the production of a variety of chemical products is based on ethanol. Among these products are those, such as acetaldehyde, that are made from petrochemicals in countries richer in crude oil. In the United States and elsewhere, ethylene is used to produce synthetic ethanol, whereas in Brazil, fermentation ethanol is used for ethylene production.

The example of Brazil is stressed because it is the world's largest producer of ethanol from sugar. Table 25 shows the amount of sugar used for ethanol production in the period from 1981 – 1982 to 1985 – 1986 in Brazil and in all other countries, except the United States in which production of ethanol is based on corn. In 1985 – 1986 Brazil's production represented 95 % of world production. Brazil almost doubled its ethanol production in the tabulated five year period [348]. Most of the ethanol is used as a fuel; only a small fraction is employed as a chemical feedstock for the production of acetaldehyde, acetic acid, butanol, octanol, chlorinated ethylenes, glycols, polyethylene, styrene, vinyl acetate, and other chemicals [346], [349].

In 1977, 158×10^6 L of ethanol was used as a raw material for chemical manufacturing in the EEC. A much larger amount was used as a solvent. In these countries, ca. 55 % of the ethanol was derived from fermentation [346].

More than 50 % of the total ethanol produced in India in 1978 was used to produce chemicals. The total production of acetaldehyde, acetic acid, acetic anhydride, and DDT (dichlorodiphenyltrichloroethane) was based on ethanol, as was a significant fraction of the production of organic acetates, acetone, butanol, and a certain amount of polyethylene and poly (vinyl chloride) and styrene [346].

Table 24. Ethanol used as a raw material in chemical manufacturing in the United States (from July 1, 1978 to June 30, 1979) [347]

Product	Ethanol consumption, L

Vinegar	73 853 912
Acetic acid	1 570 148
Ethyl acetate	33 190 215
Ethyl chloride	NA *.
Other ethyl esters	84 679 089
Sodium ethoxide	4 180 679
Ethylenamines for rubber processing	35 208 990
Dyes and intermediates	1 357 936
Acetaldehyde	NA *.
Ether, diethyl	327 524
Ether, glycol, and others (excluding diethyl ether)	11 419 306
Xanthanes	752 425
Drugs and medicinal chemicals	2 518 511
Organosilicon products	126 683
Other chemicals	9 733 531
Synthetic resins	12 609 456
Total	271 528 405

* Not available.

Table 25. Sugar consumed to produce ethanol [348]

Crop years	Sugar consumption, 10^6 t		
	Brazil	Other countries *	Worldwide
1981 – 1982	4.8	0 **	4.8
1982 – 1983	4.9	0 **	4.9
1983 – 1984	5.9	0.1	6.0
1984 – 1985	9.1	0.2	9.3
1985 – 1986	8.5	0.5	9.0

* Does not include the United States.
** Negligible amount.

[Top of Page]

10. Economic Aspects

Otto Goebel

The data given in this chapter are taken from [350-352].

10.1. Worldwide Production of Synthetic and Fermentation Ethanol

The capacities and annual production figures for already installed plants producing synthetic and fermentation ethanol in the developed and developing countries are listed in Table 26. Only industrial installations have been taken into consideration; small-scale production units with capacities of less than 5000 L/d (e.g., small units on farms and in local communities) have been ignored.

Table 26. Ethanol plant capacities and annual production

Synthesis			Fermentation		
Installed plant capacity, 10^6 L/d	Number of plants	Production, 10^6 L/a	Installed plant capacity, 10^6 L/d	Number of plants	Production, 10^6 L/a

Developed (industrialized) countries						
Australia, New Zealand				0.270	9	40
Eastern Europe				1.100	28	165
Western Europe	2.283	8	610	4.327	123	974
Israel, Japan, South Africa	0.231	4	69	0.918	31	138
North America	3.262	10	880	9.569	74	2153
Soviet Union	0.200	2	55	0.850	75	128
Total	5.976	24	1614	17.034	340	3598
Developing countries						
Africa				1.035	35	155
Asia	0.230	4	62	4.608	163	691
Asia (centrally planned economies)	2.240		605			
Central America				1.931	67	290
South America				7.704	116	1156
Oceania				0.036	2	
Total	2.470	4	667	15.314	383	2297
Total worldwide	8.446	28	2282	32.348	723	5895

The total worldwide production capacity for synthetic ethanol is 8.446×10^6 L/d and annual production is 2282×10^6 L. Most of the 28 plants producing synthetic ethanol are located in industrialized countries; the developing countries produce mainly fermentation ethanol. In Italy, the Federal Republic of Germany, the United Kingdom, and Canada, ca. 50 % of the ethanol is produced by ethylene hydration; the corresponding figure for the United States is ca. 35 %. The developed and developing countries produce approximately equal amounts of fermentation ethanol in a total of ca. 720 plants.

The total worldwide production capacity for fermentation ethanol far exceeds that for synthetic ethanol and amounts to more than 32×10^6 L/d; the average production is, however, less than 20×10^6 L/d. The reason is that fermentation plants in most of the developing countries are only supplied with raw materials derived from sugarcane for part of the year.

Fermentation ethanol producers in the industrialized nations import some of their raw materials in the form of molasses from the sugar-producing industries of developing countries.

10.2. Major Producers of Fermentation Ethanol from Regenerable Resources

More than 60 % of the total world production capacity for fermentation ethanol is shared by Brazil, the United States, and India. These countries all have favorable climates, good soil conditions, and enough land to produce the necessary raw materials.

Figure 45 shows the development of production capacities for fermentation ethanol in these countries and worldwide from 1970 to 1986. Since the first large increase in price of crude oil in the early 1970s, worldwide capacities of fermentation ethanol plants have more than doubled. The rate of capacity increase is, however, slowing down. Capacities of fermentation ethanol plants can be expected to decrease because of reduced investments and closure of obsolete plants. Drastic increases in the price of oil may, however, change this situation.

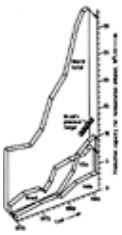

Figure 45. Development of production capacities for fermentation ethanol

Brazil. Brazil's 65 industrial fermentation ethanol plants have a total capacity of over 6.7×10^6 L/d but produce only 3.4×10

^{6} L/d. Raw materials for the production of fermentation ethanol are molasses and sugarcane juice; cassava and roots (manioc) have also been tested. Ethanol, therefore, originates from sugar factories and distilleries. The sugar industry has traditionally played a major role in the economic and social development of Brazil.

In 1975, the following two factors encouraged Brazil to set up a national program, called “Proalcool”, for the production of fermentation alcohol:

1. the fluctuations and sudden decline in international sugar prices, and
2. the increase in price of crude oil.

The establishment of new agroindustrial complexes requiring large investments and close coordination between agricultural and industrial planning is perhaps the major challenge facing the Proalcool program. The Brazilian automobile industry has already spent $ (30 – 40)×10^6 in development costs for the “alcohol engine”.

India. The production of ethanol from agricultural raw materials (primarily molasses) in India has a long tradition. The present “green revolution” is in favor of establishing more sugarcane plantations.

Until recently, India was the only country to use ethanol as a raw material for producing chemicals customarily derived from ethylene, but Brazil is now also adopting this approach.

The total capacity of India's approximately 80 industrial ethanol plants is over 3×10^6 L/d. Unfortunately, most of these plants operate at less than full capacity and for only five to six months of the year, so that actual production is ca. 1.2×10^6 L/d.

United States. Because of its vast cornfields, the United States uses corn as the primary raw material for production of fermentation ethanol. Some companies use corn to produce both ethanol and high-fructose corn syrup. Most fermentation ethanol producers are located in the corn-belt states of Iowa, Illinois, and Indiana.

Sugarcane molasses is used as a raw material for ethanol production primarily in the Southern States and the coastal areas. A few paper-manufacturing companies also use sulfite waste liquor as a raw material.

Tax reductions offered by the U.S. government as an incentive for gasohol utilization have led to the creation of more than 70 fermentation ethanol projects with an estimated total capacity of 13×10^6 L/d. The number of government-supported projects has, however, been reduced to about 45 because of budget cuts.

The total production capacity for fermentation ethanol in the United States is 9.5×10^6 L/d, and annual production is about 2×10^9 L.

[Top of Page]

11. Toxicology

Dieter Mayer

The acute oral toxicity of ethanol in rats (LD_{50}) is 11.5 [353] – 13.7 g/kg [354]. In mice [355], guinea pigs [353], rabbits [356] and dogs[355] the LD_{50} values are 9.5, 9.6, 9.9, and 6 g/kg, respectively.

Ethanol can cause mild to severe irritation in the rabbit eye depending on the concentration and amount used [357], but it is not appreciably irritating to the intact skin [357], [358].

Inhalation of ethanol causes irritation of the mucous membranes, excitation, ataxia, drowsiness, narcosis, and death due to respiratory failure. Lethal concentrations are in the range of 20 000 – 40 000 ppm in the mouse, guinea pig, and rat after inhalation for several hours. Concentrations of less than 6000 ppm can be tolerated without symptoms of intoxication [359], [360].

Repeated oral administration (12 weeks) of 10 g $kg^{-1}d^{-1}$ to juvenile rats resulted in low weight gains and fatty liver degeneration [361]. Triglyceride, cholesterol, and phospholipid metabolism were disturbed in monkeys fed a diet in which 40 % of the total calories were replaced by ethanol. Fatty degeneration occurred in the liver and the myocardium [362].

Ethanol increased embryolethality in pregnant rats and retarded fetal development when administered at levels of 5 mL/kg [363], [364]. Major malformations were not induced in the fetus.

Ethanol doses of 1 – 1.5 g/kg cause dominant lethal mutations in male mice [365], but not in female mice [366]. In the Ames test, ethanol does not exhibit mutagenic activity [367]. An increase of chromatid breakage was observed in human fibroblast cultures, but this is considered to be a cytotoxic effect rather than evidence of mutagenicity [368].

Ethanol is oxidized to acetaldehyde by means of the following enzyme systems [369]:

1. Alcohol dehydrogenase [370]
2. Catalase acting as a peroxidase and coupled to a system supplying oxygenated water [371]
3. A microsomal ethanol oxidizing system requiring NADPH as a coenzyme [372].

In the liver 80 % of the absorbed ethanol is metabolized first to acetaldehyde [371] and then to acetic acid. Acetic acid is

finally degraded to CO_2 and water. The oxidation to acetaldehyde seems to be much slower than that of acetaldehyde to acetic acid [373].

General symptoms produced by intake of ethanol in humans are listed in Table 27 [359]; however, considerable variations are observed between individual subjects.

Table 27. Symptoms produced after ethanol intake in humans

Symptom	Blood ethanol concentration, %
Beginning of uncertainty	0.06 – 0.08
Slow comprehension	0.10
Stupor	0.11 – 0.15
Drunkenness	0.16
Severe intoxication	0.2 – 0.4
Death	0.4 – 0.5

In humans suffering from alcoholism, fatty liver degeneration is thought to be due to increased fatty acid synthesis from acetate, increased transportation of lipids from peripheral fat depots to the liver, and decreased oxidation of fatty acids [374]. Fatty infiltration of the myocardium and chronic leptomeningitis are also reported in chronic alcoholism and are responsible for the well defined symptomatology of this condition [375].

The ACGIH and OSHA exposure limits for ethanol are both 1000 ppm; its MAK value is also 1000 ppm (1900 mg/m^3).

[Top of Page]

References

1. E. A. Cotelle, US 41 685, 1861.
2. W. Gorianoff, A. Butleroff, *Justus Liebigs Ann. Chem.* **180** (1876) 245.
3. C. R. Nelson, M. L. Courter, *Chem. Eng. Prog.* **50** (1954) 526. Links
4. Yu. M. Bakshi, A. I. Gelbshtein, M. I. Temkin, *Dokl. Akad. Nauk SSSR* **126** (1959) 314; *Chem. Abstr.* **55** (1961) 1944 f.
5. D. S. Tsiklis, A. I. Kulikova, *Zh. Fiz. Khim.* **35** (1961) 954; *Chem. Abstr.* **58** (1963) 6660g.
6. Z. Novosad, *Chem. Prum.* **5** (1955) no. 30, 72.
7. A. I. Gelbshtein, Yu. M. Bakshi, M. I. Temkin, *Dokl. Akad. Nauk SSSR* **132** (1960) 384; *Chem. Abstr.* **57** (1962) 6663d.
8. B. S. Bouden, A. I. Gelbshtein *et al.*, *Neftepererab. Neftekhim. (Moscow)* 1968, no. 5, 33.
9. C. V. Mace, C. F. Bomilla, *Chem. Eng. Prog.* **50** (1954) 385. Links
10. Ch.-K. Feng, *Hua Hsueh Tung Pao* 1978, no. 6, 367; *Chem. Abstr.* **90** (1979) 6558 f.
11. R. W. Taft, P. J. Riesz, C. A. Fazio, *J. Am. Chem. Soc.* **74** (1952) 5372. Links
12. F. Vachez, *Rev. Inst. Fr. Pet. Ann. Combust. Liq.* **18** (1963) 724.
13. De Bataafsche Petroleum Maatschappij, DE 1 042 561, 1955.
14. Eastman Kodak Co., GB 1 144 947, 1966.
15. *Hydrocarbon Process.* **46** (1967) no. 11, 168.
16. Veba Chemie, GB 1 201 181, 1967.
17. Hibernia-Chemie, BE 715 907, 1968.
18. National Distillers & Chemical Co., DE-OS 2 015 536, 1970.
19. BP Chemicals International, DE-OS 2 237 015, 1972.
20. C. L. Thomas: *Catalytic Processes and Proven Catalysts*, Academic Press, New York-London 1970, p. 230.
21. G. G. Eremeeva, V. M. Dronkin, G. I. Gagarina, *Neftepererab. Neftakhim (Moscow)* 1979, no. 16, 19 – 20; *Chem. Abstr.* **91** (1979) 107 638.
22. K. W. Toptschijewa, S. M. Rachowskaja, I. K. Kutschkajewa, *Neftekhimiya* **3** (1963) 271.
23. Chemopetrol Koncernovy Podnik Chemicke Zavody Ceskoslovensko-Sovetskeho Praptelstvi, BE 888 479, 1981 (V. Kadlec, V. Grosser, J. Rosenthal); *Chem. Abstr.* **96** (1982) 37 267.
24. Union Carbide, BE 854 606, 1977.
25. Kurashiki Rayon Co., DE-OS 1 280 833, 1964.
26. Kurashiki Rayon Co., DE-OS 1 543 130, 1965.
27. Esso Res. & Eng. Co., US 3 686 334, 1969.
28. Tokuyama Soda K. K., DE-OS 2 022 568, 1970.
29. Tokuyama Soda K. K., DE-OS 2 215 380, 1972.
30. G. K. Boreskow, W. A. Dsisjko *et al.*, *Khim. Promst. (Moscow)* 1961, no. 2, 97.
31. Eastman Kodak Co., US 3 554 926, 1968.

32. L. N. Tolkacheva, L. A. Novikova, V. A. Plakhotnik, V. Y. Danyushevski *et al.*, *Neftekhimiya* **23** (1983) no. 6, 803; *Chem. Abstr.* **100** (1984) 105 486.
33. Mobil Oil Co., US 4 214 107, 1980 (C. D. Cheng, N. J. Morgan).
34. Toray Industries, JP 7 245 323, 1968.
35. A. M. Tsybulevskii, L. A. Novikova, V. A. Kondrat'ev, L. N. Tolkacheva *et al.*, *Neftekhimiya* **19** (1979) no. 5, 771; *Chem. Abstr.* **92** (1980) 110 468 r.
36. G. F. Nencetti, R. Tartarelli, *Chim. Ind. (Milan)* **48** (1966) 30.
37. M. Baccaredda, G. F. Nencetti *et al.*, *Riv. Combust.* **22** (1968) 65.
38. R. Tartarelli, M. Giorgini *et al.*, *J. Catal.* **17** (1970) 41. Links
39. A. A. Kubashov, K. V. Topchieva, M. G. Mitichenko, G. G. Eremeeva *et al.*, *Zh. Prikl. Khim. (Leningrad)* **53** (1980) no. 5, 1090; *Chem. Abstr.* **93** (1980) 132 011 j.
40. M. Giorgini, R. Tartarelli, *Chim. Ind. (Milan)* **58** (1976) no. 9, 611; *Chem. Abstr.* **86** (1977) 11 804.
41. UOP, US 4 465 874, 1984 (J. A. Koca).
42. Mitsui Toatsu Chemicals, JP-Kokai 79 27 505, 1979 (F. Matsuda, T. Kato).
43. Mitsui Toatsu Chemicals, JP-Kokai 79 16 4 143, 1979 (F. Matsuda, T. Kato).
44. T. C. Carle, D. M. Stewart, *Chem. Ind. (London)* 1962, 830.
45. B. S. Bouden, G. G. Goryacheva, G. I. Gagarina, *Neftepererab. Neftekhim. (Moscow)* 1967, no. 12, 24.
46. Zh. A. Bril, V. M. Platonov, M. Ya. Klimenko, *Khim. Promst. (Moscow)* **45** (1969) 332.
47. B. S. Bouden, A. I. Gelbshtein *et al.*, *Neftepererab. Neftekhim. (Moscow)* 1968, no. 4, 17.
48. G. G. Goryacheva, B. S. Bouden, A. I. Gelbshtein, *Sov. Chem. Ind. (Engl. Transl.)* 1970, no. 6, 15.
49. J. Koutensky, *Sb. Pr. Vyzk. Chem. Vyuziti Uhli Dehtu Ropy* **15** (1978) 211; *Chem. Abstr.* **89** (1978) 75 211.
50. J. Koutensky, *Sb. Pr. Vyzk. Chem. Vyuziti Uhli Dehtu Ropy* **16** (1979) 183; *Chem. Abstr.* **93** (1980) 167 145.
51. H. G. Harris, D. M. Himmelblau, *J. Chem. Eng. Data* **9** (1964) 61. Links
52. UOP, US 4 374 286, 1983 (R. J. Schmidt).
53. Haldor F. A. Topsøe, DE-OS 2 158 795, 1971.
54. I. Wender, R. Levine, M. Orchin, *J. Am. Chem. Soc.* **71** (1949) 4160. Links
55. I. Wender, H. Greenfield, M. Orchin, *J. Am. Chem. Soc.* **73** (1951) 2656. Links
56. British Petroleum Co., EP 3 876, 1979 (B. R. Gane, D. E Stewart).
57. Gulf Res. & Dev. Co., US 4 239 924, 1980 (J. E. Bozik, T. P. Kobylanski, W. R. Pretzer).
58. British Petroleum Co., EP 1 937, 1979 (B. R. Gane, D. C. Stewart).
59. British Petroleum Co., GB 2 053 915, 1981 (W. J. Ball, D. G. Stewart).
60. British Petroleum Co., JP-Kokai 8 049 326, 1980.
61. Y. Sugi, K. Brando, Y. Takami, *Chem. Lett.* 1981, no. 1, 63. Links
62. Catalytica Associates, Review of PERC Program no. FE-7093; *Homogeneous Reactions-,* “The Homologation of Methanol”, Dept. of Energy Contract EF-77-C-01–2536, May 1978.
63. Gulf Res. & Dev. Co., US 4 239 924, 1980 (J. E. Bozic, T. P. Kobylinsky, W. R. Pretzer).
64. Gulf Res. & Dev. Co., DE 3 228 769, 1983 (M. M. Habib, W. R. Pretzer).
65. Agency of Industrial Sciences and Technology, JP-Kokai 59 110 637, 1984.
66. Ruhrchemie AG, DE 3 330 507, 1985 (W. Lipps, H. Bahrmann, B. Cornils, W. Konkal).
67. Air Products & Chemicals, US 4 171 461, 1979 (G. M. Bartish).
68. Agency for Industrial Sciences and Technology, JP-Kokai 59 110 637, 1982.
69. Mitsubishi Gas Chem. Co., JP-Kokai 82 80 334, 1982.
70. Union Chemische Braunkohlen Kraftstoff, DE 3 045 891, 1982 (K. H. Keim, J. Koff).
71. Ethyl Corp., US 4 309 314, 1982 (D. C. Hargis, M. Dubek).
72. B. Juran, A. V. Porcelli, *Hydrocarbon Process. Int. Ed.* **64** (1985) no. 10, Section 1, 85.
73. Monsanto Chemical Co., BE 713 296, 1968 (P. E. Paulik, J. F. Roth *et al.*).
74. J. F. Roth, J. H. Cradock, A. Hershman, P. E. Paulik, *Chem. Technol* . 1971, 600.
75. Davy McKee (London), WO 8 303 409, 1983 (M. W. Bradley, N. Harris, K. Turner).
76. Halcon SD Group, DE 3 335 594, 1984 (N. Rizkalla).
77. Halcon SD Group, DE 3 335 694, 1984 (N. Rizkalla).
78. Halcon SD Group, US 4 497 967, 1985 (C. G. Wan).
79. G. Natta, U. Colombo, I. Pasquon in P. H. Emmett (ed.): *Catalysis*, vol. **5**, Reinhold Publ. Co., New York 1957, p. 131.
80. B. D. Dombek, *Adv. Catal.* **32** (1983) 325.
81. D. A. Popeckov, A. A. Shokol, *Zap. Inst. Khim. Akad. Nauk Ukr. RSR* **4** (1937) 205.
82. R. Taylor, *J. Chem. Soc.* 1934, 1429.
83. P. Courty, J. P. Arlie, A. Convers, P. Mitchenko *et al.*, *Actual. Chim.* 1983, no. 9, 19.
84. Inst. Français du Pétrole, DE 2 748 097, 1978 (A. Suguir, E. Freund).
85. Inst. Français du Pétrole, DE 2 949 952, 1980 (A. Suguir, E. Freund).
86. New Fuel Oil Dev. Technology Res. Assoc., JP-Kokai 8 545 537, 1985.
87. Res. Assoc. for Petrol Alternative Dev., EP 110 357, 1984 (M. Shibata, Y. Aoki, S. Uchiyama).

88. I. A. Sibilia, J. M. Dominguez, R. G. Herman, K. Klier, *Prepr. Pap. Am. Chem. Soc., Div. Fuel Chem.* **29** (1984) no. 5, 261.
89. K. J. Smith, R. B. Anderson, *Prepr. Pap. Am. Chem. Soc., Div. Fuel Chem.* **29** (1984) no. 5, 269.
90. M. Ichikawa, *J. Chem. Soc. Chem. Commun.* 1978, 566; *Bull. Chem. Soc. Jpn.* **51** (1978) no. 8, 2268, 2273.
91. R. Bardet, J. Thivolle-Cazat, Y. Trambouze, *C. R. Seances Acad. Sci. Ser. 2* **292** (1981) no. 12, 883.
92. Societé Chimique de Grand Paroisse, Azote et Prod. Chim., FR 2 530 159, 1984 (R. Bardet, J. Thivolle-Cazat, Y. Trambouze, C. Harmon).
93. A. Cariotti, S. Martingo, L. Sanderighi, C. Tonelli *et al.*, *J. Chem. Soc. Faraday Trans. 1* **80** (1984) no. 6, 1605.
94. Sagami Chemical Research Center, JP-Kokai 82 108 026, 1982.
95. Agency for Industrial Sciences and Technology, JP-Kokai 8 532 730, 8 532 734, 8532736, 1985.
96. Agency for Industrial Sciences and Technology, JP-Kokai 8 525 943, 8 532 729, 8 532 731, 8 532 732, 8 532 733, 8 532 735, 1985.
97. Showa Denko K. K., JP-Kokai 8 262 231, 8 262 232, 8 262 233, 82 109 728, 82 109 729, 8 262 730, 82 126 432, 1982.
98. Hoechst, EP 21 241, 1981 (H. Hachenberg, F. Wunder, E. I. Leupold, H. J. Schmidt).
99. Union Carbide, US 4 301 253, 1981 (B. K. Warren).
100. B. K. Warren, P. D. Dombek, *J. Catal.* **79** (1983) 334. Links
101. J. F. Knifton, *Prepr. Pap. Am. Chem. Soc. Div. Petr. Chem.* **29** (1984) no. 2, 586.
102. J. F. Knifton, *J. Am. Chem. Soc.* **103** (1981) 3959. Links
103. Texaco Dev. Corp., US 4 265 828, 1981 (J. K. Knifton).
104. Texaco Dev. Corp., US 2 041 924, 1980 (J. F. Knifton).
105. Agency for Industrial Sciences and Technology, JP-Kokai 84 164 738, 1984.
106. Agency for Industrial Sciences and Technology, JP-Kokai 83 172 333, 1983.
107. S. E. Ferrari, J. J. C. Lopes, J. R. A. Leme, E. R. de Oliveira: *Proc. 4th Int. Symp. Alcohol Fuels Technology*, Sao Paulo, Brazil Oct. 1980, p. 139.
108. C. S. Gong, L. F. Chen, *Biotechnol. Bioeng. Symp.* **14** (1984) 257 – 268.
109. R. P. Jones, N. Pamment, P. F. Greenfield, *Process Biochem.* **16** (1981) no. 3, 42 – 49. Links
110. R. P. Jones, P. F. Greenfield, *Process Biochem.* **19** (1984) no. 2, 48 – 58. Links
111. B. Maiorella, Ch. R. Wilke, H. W. Blanck: “Alcohol Production and Recovery”, in A. Fiechter (ed.): *Advances in Biochemical Engineering*, **20** Springer-Verlag, New York 1981 pp. 43 – 92.
112. E. Oura, *Process Biochem.* **12** (1977) no. 3, 19 – 21, 35. Links
113. H. R. Mahler, E. H. Cordes: *Biological Chemistry*, Int. ed., Harper & Row, New York 1968, pp. 404 –435.
114. C. S. Gong, L. F. Chen, G. T. Tsao, M. C. Flickinger: “Conversion of Hemicellulose Carbohydrates”, in A. Fiechter (ed.): *Advances in Biochemical Engineering*, **20** Springer-Verlag, New York 1981 pp. 93 – 118
115. T. W. Jeffries: “Utilization of Xylose by Bacteria, Yeasts, and Fungi”, in A. Fiechter (ed.): *Advances in Biochemical Engineering/Biotechnology,* vol. 27, Springer-Verlag, Berlin-New York 1983, pp. 1 – 32.
116. S. Mitsuhashi, J. O. Lampen, *J. Biol. Chem.* **204** (1953) 1011 – 1018. Links
117. B. L. Wilson, R. P. Mortlock, *J. Bacteriol.* **113** (1973) 1404 – 1411. Links
118. P. Y. Wang, H. Schneider, *Can. J. Microbiol.* **26** (1980) 1165 – 1171. Links
119. M. Hoefer, A. Betz, A. Kotyk, *Biochim. Biophys. Acta* **252** (1971) 1 – 12. Links
120. R. W. Detroy, R. L. Cunningham, A. I. Herman, *Biotechnol. Bioeng. Symp.* **12** (1982) 81 – 89.
121. M. Tomoyeda, H. Horitsu, *Agric. Biol. Chem.* **28** (1964) 139.
122. C. S. Gong, L. F. Chen, M. C. Flickinger, L. C. Chiang *et al.*, *Appl. Environ. Microbiol.* **41** (1981) 430 – 436. Links
123. J. A. Barnett, *Adv. Carbohydr. Chem. Biochem.* **32** (1976) 125 – 234. Links
124. T. W. Jeffries, *Biotechnol. Bioeng. Symp.* **12** (1982) 103 – 110.
125. H. Dellweg, M. Rizzi, H. Methner, D. Debus, *Biotechnol. Lett.* **6** (1984) 395 – 400. Links
126. M. Korhola, I. Suomalainen, E. Vaisanen, H. Tuompo: “Distiller's Yeasts”, in M. Korhola, H. Tuompo, V. Kauppinen (eds.): *Proceedings of the 7th Conference on Global Impacts of Applied Microbiology: Symposia on Alcohol Fermentation and Plant Cell Culture,* vol. 4, Foundation for Biotechnical and Industrial Fermentation Research. Helsinki 1986, pp. 29 – 61.
127. A. H. Rose, J. S. Harrison (eds.): *The Yeasts*, vol. **3**, Academic Press, London 1970, p. 306.
128. G. K. Hoppe, G. Hansford, *Biotechnol. Lett.* **6** (1984) 681 – 686. Links
129. A. A. Andreasen, T. J. B. Stier, *J. Cell. Comp. Physiol.* **43** (1954) 271 – 281.
130. R. D. Tyagi, *Process Biochem.* **19** (1984) no. 4, 136 – 141. Links
131. G. R. Cysewski, C. R. Wilke, *Biotechnol. Bioeng.* **18** (1976) 1297 – 1313. Links
132. C. Beck, H. K. von Meyenburg, *J. Bacteriol.* **96** (1968) 479 – 486. Links
133. K. Watson, *Biotechnol. Lett.* **4** (1982) 397 – 402. Links
134. S. Laffon-Lafourcade, F. Larue, P. Ribereau-Gayon, *Appl. Environ. Microbiol.* **38** (1979) 1069 – 1073. Links
135. F. H. White, *Proc. Int. Brewing Conv. (Australia-New Zealand Section)* **15** (1978) 133.
136. T. K. Ghose, R. D. Tyagi, *Biotechnol. Bioeng.* **21** (1979) 1401 – 1420.
137. G. K. Hoppe, G. S. Hansford, *Biotechnol. Lett.* **4** (1982) 39 – 44. Links
138. C. D. Bazua, C. R. Wilke, *Biotechnol. Bioeng. Symp.* **7** (1977) 105 – 118. Links

139. I. Holzberg, R. K. Finn, K. H. Steinkraus, *Biotechnol. Bioeng.* **9** (1967) 413 – 427. Links
140. S. Aiba, M. Shoda, *J. Ferment. Technol.* **47** (1969) 790 – 794.
141. N. B. Egamberdiev, N. D. Ierusalimskii, *Microbiologiya* **37** (1968) 687.
142. G. K. Hoppe (1981), as cited in [137]
143. P. Strehaiano, M. Moreno, G. Goma, *C. R. Acad. Sci. Ser. D.* **286** (1978) 225.
144. S. Aiba, M. Shoda, M. Nagatani, *Biotechnol. Bioeng.* **10** (1968) 845 – 864.
145. T. W. Nagodawithana, C. Castellano, K. H. Steinkraus, *Appl. Microbiol.* **28** (1974) 383 – 391. Links
146. K. Watson, R. Cavicchioli, *Biotechnol. Lett.* **5** (1983) 683 – 688. Links
147. R. Espinosa, V. Cojulun, F. Maroquin, *Biotechnol. Bioeng. Symp.* **8** (1978) 69 – 74.
148. G. Reed (ed.): *Prescott and Dunns's Industrial Microbiology*, 4th ed., AVI Publishing Co., Westport, Conn., 1984, pp. 835 – 859.
149. D. B. Hughes, N. J. Tudroszen, C. J. Moye, *Biotechnol. Lett.* **6** (1984) 1 – 6. Links
150. H. Suomalainen, O. Kauppila, R. J. Peltonen, L. Nykanen: *Handbuch der Lebensmittelchemie*, vol. **7**, Springer-Verlag, Berlin-Heidelberg 1968, pp. 496 – 653.
151. T. Seki, S. Myoga, S. Limtong, S. Vedono *et al.*, *Biotechnol. Lett.* **5** (1983) 351 – 356. Links
152. A. J. Hacking, I. W. F. Taylor, C. M. Hanas, *Appl. Microbiol. Biotechnol.* **19** (1984) 361 – 363. Links
153. P. J. Blotkamp, M. Takagi, M. S. Pemberton, G. H. Emert: *Proc. 84th Natl. Meeting* AIChE, Atlanta, Ga. Feb. 1978.
154. Gulf Oil Corp., GB 2 036 074 A, 1980 (M. S. Pemberton, S. D. Crawford).
155. J. J. Savarese, S. D. Young, *Biotechnol. Bioeng.* **20** (1978) 1291 – 1293. Links
156. M. Mes-Hartree, C. Hogan, R. D. Hayes, J. N. Saddler, *Biotechnol. Lett.* **5** (1983) 101 – 106. Links
157. D. B. Rivers, G. H. Emert: *Proc. Bioenergy '80 World Congress and Exposition*, Atlanta, Ga., Apr. 1980, p. 157.
158. L. D. McCracken, C. S. Gong, *Biotechnol. Bioeng. Symp.* **12** (1982) 91 – 102.
159. S. Hayashida, K. Ohta, P. Q. Flor, N. Nanri *et al.*, *Agric. Biol. Chem.* **46** (1982) 1947 – 1950.
160. N. Matsumoto, A. Fukushi, M. Miyanaga, K. Kakihara *et al.*, *Agric. Biol. Chem.* **46** (1982) 1549 – 1558.
161. C. L. Weller, M. P. Steinberg, E. D. Rodda, *Biotechnol. Bioeng. Symp.* **13** (1983) 437 – 447.
162. J. J. Wilson, G. G. Khachatourians, W. M. Ingledew, *Biotechnol. Lett.* **4** (1982) 333 – 338. Links
163. A. M. Sills, G. G. Stewart, *J. Inst. Brew.* **88** (1982) 313 – 316.
164. K. Oteng-Gyang, G. Moulin, P. Galzi, *Z. Microbiol.* **21** (1981) 537 – 544.
165. J. J. Wilson, W. M. Ingledew, *Appl. Environ. Microbiol.* **44** (1982) 301 – 307. Links
166. R. De Mot, E. van Oudendijck, H. Verachtert, *Biotechnol. Lett.* **6** (1984) 581 – 586.
167. J. Swings, J. De Ley, *Bacteriol. Rev.* **41** (1977) 1 – 46. Links
168. H. Sahm, S. Bringer-Meyer, G. Sprenger, in A. Balows, H. G. Trüper, M. Dworkin, W. Harder, K. H. Schleifer (eds.): *The Prokaryotes* 2nd ed., Springer, New York 1992 , pp. 2287 – 2301.
169. P. L. Rogers, K. J. Lee, M. L. Skotniki, D. E. Tribe, *Adv. Biochem. Eng.* **23** (1982) 27 – 84.
170. L. O. Ingram, H. C. Aldrich, A. C. C. Borges, T. B. Causey, A. Martinez, F. Morales, A. Saleh, S. A. Underwood, L. P. Yomano, S. W. York, J. Zaldivar, S. Zhou, *Biotechnol. Progr.* **15** (1999) 855 – 866. Links
171. H. Yanase, N. Kato, K. Tonomura, *Bioprocess Technol.* **19** (1994) 723 – 729. Links
172. G. A. Sprenger, *J. Biotechnol.* **27** (1993) 225 – 237. Links
173. G. A. Sprenger, *FEMS Microbiol. Lett.* **145** (1996) 301 – 307. Links
174. G. A. Sprenger, M. A. Typas, C. Drainas, *World J. Microbiol. Biotechnol.* **9** (1993) 17 – 24.
175. P. Gunasekaran, K. C. Raj, *Curr. Sci.* **77** (1999) 56 – 68. Links
176. H.-J. Schepers, S. Bringer-Meyer, H. Sahm, *Naturforsch. C: Biosci.* **42C** (1987) 401 – 407.
177. J. S. Tolan, R. K. Finn, *Appl. Environ. Microbiol.* **53** (1987) 2033 – 2038. Links
178. R. K. Finn, S. Bringer, H. Sahm, *Appl. Microbiol. Biotechnol.* **19** (1984) 161 – 166. Links
179. S. Bringer, E. Durst, H. Sahm, R. K. Finn, *Biotechnol. Bioeng. Symp.* **14** (1984) 269 – 278.
180. L. H. Carreira, L. G. Ljungdahl, in D. L. Wise (ed.): *Liquid Fuel Developments*, CRC Press, Boca Raton, FL 1983, p. 1.
181. J. G. Zeikus: “Biology of the Spore-Forming Anaerobes”, in in A. L. Demain, N. A. Solomon (eds.): *Biology of Industrial Microorganisms*, Benjamin/Cummings, London 1985, p. 79.
182. L. O. Ingram, *CRC Crit. Rev. Biotechnol.* **9** (1990) 305 – 319.
183. M. A. Hermans, B. Neuss, H. Sahm, *J. Bacteriol.* **173** (1991) 5592 – 5595. Links
184. H. Sahm, M. Rohmer, S. Bringer-Meyer, G. A. Sprenger, R. Welle, *Adv. Microb. Physiol.* **35** (1993) 247 – 273. Links
185. R. A. Moreau, M. J. Powell, W. F. Fett, B. D. Whitaker, *Curr. Microbiol.* **35** (1997) 124 – 128. Links
186. E. L. Kannenberg, K. Poralla, *Naturwissenschaften* **86** (1999) 168 – 176. Links
187. S. Horbach, B. Neuss, H. Sahm, *FEMS Microbiol. Lett.* **79** (1991) 347 – 350. Links
188. S. Bringer, R. K. Finn, H. Sahm, *Arch. Microbiol.* **139** (1984) 376 – 381. Links
189. M. Strohdeicher, B. Neuss, S. Bringer-Meyer, H. Sahm, *Arch. Microbiol.* **154** (1990) 536 – 543. Links
190. U. Kalnenieks, N. Galinina, S. Bringer-Meyer, R. K. Poole, *FEMS Microbiol. Lett.* **168** (1998) 91 – 97. Links
191. L. M. Pankova, Y. E. Shvinka, M. E. Beker, E. E. Slava, *Mikrobiologiya* **54** (1995) 141 – 145.
192. S. Bringer, H. Sahm, W. Swyzen, *Biotechnol. Bioeng. Symp.* **14** (1984) 311 – 319.
193. D. Weuster-Botz, A. Aivasidis, C. Wandrey, *Appl. Microbiol. Biotechnol.* **39** (1993) 685 – 690. Links

194. H. Sahm, S. Bringer-Meyer, *Acta Biotechnol.* **7** (1987) 307 – 313. Links
195. H. G. Lawford, J. D. Rousseau, *Appl. Biochem. Biotechnol.* **34/35** (1992) 205 – 216.
196. D. Weuster-Botz, *Appl. Microbiol. Biotechnol.* **39** (1993) 679 – 684. Links
197. C. H. Kim, S. K. Rhee, *Proc. Biochem.* **28** (1993) 331–339.
198. L. Viikari, *Appl. Microbiol. Biotechnol.* **19** (1984) 252 – 255. Links
199. T. R. Kannan, G. Sangiliyandi, P. Gunasekaran, *Enzyme Microb. Technol.* **22** (1998) 179 – 184. Links
200. W.-C. Lee, C.-T. Huang, *Enzyme Microb. Technol.* **17** (1995) 79 – 84. Links
201. H. W. Doelle, L. Kirk, R. Crittenden, H. Toh, M. B. Doelle, *Crit. Rev. Biotechnol.* **13** (1993) 57 – 98. Links
202. M. Zhang, C. Eddy, K. Deanda, M. Finkelstein, S. Picataggio, *Science* **267** (1995) 240 – 243. Links
203. Midwest Research Institute, US 5 514 583, 1996 (S. K. Picataggio, M. Zhang, C. K. Eddy, K. A. Deanda, M. Finkelstein).
204. A. de Graaf, K. Striegel, R. M. Wittig, B. Lauffer, G. Schmitz, W. Wiechert, G. A. Sprenger, H. Sahm, *Arch. Microbiol.* **171** (1999) 371 – 385 Links
205. I. S. Kim, K. D. Barrow, P. L. Rogers, *Appl. Environ. Microbiol.* **66** (2000) 186 – 193. Links
206. E. Joachimsthal, K. D. Haggett, P. L. Rogers, *Appl. Biochem. Biotech.* **77 – 79** (1999) 147 – 157. Links
207. T. D. Ranatunga, J. Jervis, R. F. Helm, J. D. McMillan, C. Hatzis, *Appl. Biochem. Biotechnol.* **67** (1997) 185 – 198. Links
208. H. G. Lawford, J. D. Rousseau, *Appl. Biochem. Biotech.* **70 – 72** (1998) 161 – 172. Links
209. H. G. Lawford, J. D. Rousseau, A. Mohagheghi, J. D. McMillan, *Appl. Biochem. Biotech.* **77 – 79** (1999) 191 – 204. Links
210. K. Deanda, M. Zhang, C. Eddy, S. Picataggio, *Appl. Environ. Microbiol.* **62** (1996) 4465 – 4470. Links
211. P. Weisser, R. Krämer, G. A. Sprenger, *Appl. Environ. Microbiol.* **62** (1996) 4155 – 4161. Links
212. L. O. Ingram, J. B. Doran, *FEMS Microbiol. Rev.* **16** (1995) 235 – 241. Links
213. L. O. Ingram, P. F. Gomez, X. Lai, M. Moniruzzaman, B. E. Wood, L. P. Yomano, S. W. York, *Biotechnol. Bioeng.* **58** (1998) 204 – 214. Links
214. University of Florida, US 5 028 539, 1991 (L. O. Ingram, D. P. Clark).
215. K. Ohta, D. S. Beall, J. P. Mejia, K. T. Shanmugam, L. O. Ingram, *Appl. Environ. Microbiol.* **57** (1991) 893 – 900. Links
216. L. O. Ingram, T. Conway, *Appl. Environ. Microbiol.* **54** (1988) 397 – 404. Links
217. L. O. Ingram, T. Conway, D. P. Clark, G. W. Sewell, J. F. Preston, *Appl. Environ. Microbiol.* **53** (1987) 2420 – 2425. Links
218. US 5 424 202, 1995 (L. O. Ingram, D. S. Beall, G. F. H. Burchhardt, W. V. Guimaraes, K. Ohta, B. E. Wood, K. T. Shanmugam, D. E. Fowler, A. Ben-Bassat).
219. K. Ohta, D. S. Beall., J. P. Mejia, K. T. Shanmugam, L. P. Ingram, *Appl. Environ. Microbiol.* **57** (1991) 2810 – 2815. Links
220. D. S. Beall, L. O. Ingram, *J. Ind. Microbiol.* **11** (1993) 151 – 155.
221. X. Lai, F. C. Davis, R. B. Hespell, L. O. Ingram, *Appl. Environ. Microbiol.* **63** (1996) 355 – 363. Links
222. B. E. Wood, L. O. Ingram, *Appl. Environ. Microbiol.* **58** (1992) 2103 – 2110. Links
223. J. B. Doran, L. O. Ingram, *Biotechnol. Progr.* **9** (1993) 533 – 538. Links
224. J. B. Doran, H. C. Aldrich, L. O. Ingram, *Biotechnol. Bioeng.* **44** (1994) 240 – 247. Links
225. D. Rose, *Process Biochem.* **3** (1976) 10 – 12, 36. Links
226. A. Duvnjak, N. Kosaric, S. Kliza, *Biotechnol. Bioeng.* **24** (1982) 2297 – 2308. Links
227. N. Kosaric, A. Wieczorek, G. P. Cosentino, R. J. Magee et al.: “Ethanol Fermentation”, in H. J. Rehm, G. Reed (eds.): *Biotechnology,* vol. 3, Verlag-Chemie, Weinheim 1983, pp. 257 – 385.
228. T. K. Ghose, R. D. Tyagi, *Biotechnol. Bioeng.* **21** (1979) 1387 – 1400. Links
229. E. J. Del Rosario, K. J. Lee, P. L. Rogers, *Biotechnol. Bioeng.* **21** (1979) 1447 – 1482.
230. O. Rahn, *Growth* **16** (1952) 59. Links
231. G. R. Cysewski, C. R. Wilke, *Biotechnol. Bioeng.* **20** (1978) 1421 – 1444. Links
232. V. L. Yarovenko: “Theory and Practice of Continuous Cultivation of Microorganisms in Alcoholic Production”, in A. Fiechter, T. K. Ghose, N. Glakenbrough (eds.): *Advances in Biochemical Engineering,* vol. 9, Springer-Verlag, Berlin-Heidelberg-New York 1978, pp. 1 – 30.
233. G. R. Cysewski, C. R. Wilke, *Biotechn. Bioeng.* **19** (1977) pp. 1125 – 1143.
234. N. Kosaric, Z. Duvnjak, A. Wieczorek in, Foundation for Biotechnological and Industrial Fermentation Research, [126] vol. 4, pp. 101 – 120.
235. T. J. Walsh, H. R. Bungay, *Biotechnol. Bioeng.* **21** (1979) 1081 – 1084.
236. F. Jirmann, U. D. Runkel, *Brauwelt* **107** (1967) 1453.
237. J. Hough *et al.*, *J. Inst. Brew.* **68** (1962) 478.
238. C. B. Netto, A. Destrauhaut, G. Goma, *Biotechnol. Lett.* **7** (1985) no. 5, 355 – 360. Links
239. I. G. Prince, J. P. Barford, *Biotechnol. Lett.* **4** (1982) 263 – 268. Links
240. D. M. Comberbach, J. D. Bu'Lock, *Biotechnol. Lett*. (1984) no. 2, 129 – 134.
241. N. Kosaric, A. Wieczorek, Z. Duvnjak: *Proceedings of the First Bioenergy Specialists' Meeting on Biotechnology*, Waterloo, Ontario, Oct. 14 – 18, 1984, pp. 262 – 272.
242. R. N. Greenshields, E. L. Smith, *Process Biochem.* **9** (1974) no. 3, 11 – 17, 28. Links

243. S. J. Pirt: *Principles of Microbe and Cell Cultivation*, Blackwell Scientific Publications, Oxford, 1975, pp. 218 – 221.
244. A. Margaritis, C. R. Wilke, *Biotechnol. Bioeng.* **20** (1978) 709 – 726. Links
245. A. Margaritis, C. R. Wilke, *Biotechnol. Bioeng.* **20** (1978) 727 – 753. Links
246. M. Cheryvan, M. Mehaia, *Process Biochem.* **12** (1984) 204 – 208. Links
247. M. H. V. Mulder, C. A. Smolders, *Process Biochem.* **4** (1986) 35 – 39. Links
248. A. Ramalingham, R. K. Finn, *Biotechnol. Bioeng.* **19** (1977) 583 – 589. Links
249. B. Maiorella, H. W. Blanck, C. R. Wilke: “Rapid Ethanol Production via Fermentation”, Univ. Calif., Lawrence Berkley Lab. Report 10 219, *AIChE 72nd National Meeting,* San Francisco, Calif., Nov. 19, 1979.
250. E. K. Pye, A. E. Humphrey, “The Biological Production of Liquid Fuels from Biomass”, Univ. Penn. Interim Report to DOE, Washington D.C., Jun. – Aug. 1979, p. 79, Task 7.
251. M. Minier, G. Goma, *Biotechnol. Lett.* **3** (1981) 405 – 408. Links
252. E. K. Pye, A. E. Humphrey: “Production of Liquid Fuels from Cellulosic Biomass”, *Proc. 3rd. Ann. Biomass Energy Systems Conf.*, DOE, Solar Energy Res. Inst. Golden, Colo., Jun. 5, 1980, p. 69.
253. F. B. Kolot, *Process Biochem.* **2** (1984) 7 – 13. Links
254. D. Williams, D. M. Munnecke, *Biotechnol. Bioeng.* **23** (1981) 1813 – 1825. Links
255. M. Wada, J. Kato, I. Chibata, *Eur. J. Appl. Microbiol. Biotechnol.* **11** (1979) 67.
256. Y. Linko, P. Linko, *Biotechnol. Lett.* **3** (1981) 21 – 26. Links
257. T. K. Ghose, K. K. Bandyopadhyay, *Biotechnol. Bioeng.* **22** (1980) 1489 – 1496. Links
258. J. L. Vega, J. D. O'Malley, E. C. Clausen, J. L. Gaddy, *Biotechnol. Bioeng. Symp.* **15** (1985) 263.
259. M. C. Dale, M. R. Okos, P. C. Wankat, *Biotechnol. Bioeng.* **27** (1985) 932 – 942.
260. A. J. Dauglis, N. H. Brown, W. R. Cluett, D. B. Dunlop, *Biotechnol. Lett.* **3** (1981) 651 – 656.
261. R. D. Tyagi, T. K. Ghose, *Biotechnol. Bioeng.* **24** (1982) 781 – 795.
262. W. Halwacks, C. Wandrey, K. Schugerl, *Biotechnol. Bioeng.* **20** (1978) 541 – 554.
263. R. A. Nathan, “Fuels from Sugar Crops”, DOE Technical Information Center, Oak Ridge, Tenn., 1978.
264. J. L. Prouty in [107] p. 57.
265. H. W. Ockerman: *Source Book of Food Scientists*, AVI Publishing Co., Westport, Conn., 1978.
266. B. P. Baker: “The Availability, Composition and Properties of Cane Molasses”, in E. Sinda, E. Parkkinen (eds.): *Problems with Molasses in the Yeast Industry, Symposium,* Aug. 31 – Sept. 1, 1979, Helsinki, Finland, p. 126.
267. L. R. Lindeman, C. Rocchicciolo, *Biotechnol. Bioeng.* **21** (1979) 1107.
268. K. Rosen, *Process Biochem.* **13** (1978) no. 5, 25. Links
269. Alcon Biotechnology, London, 1980.
270. *Chem. Eng.* **25** (1980) Aug. 11.
271. W. M. Ingledew: *Proc. Canpac '81 Conf.*, Winnipeg, Manitoba Oct. 1981.
272. R. A. M. Ginuis: *Beet Sugar Technology*, 3rd ed., The Beet Sugar Development Foundation, Fort Collins, Colo., 1982.
273. J. K. Paul, *Ethyl Alcohol Production and Use as a Motor Fuel*, Noyes Data Corp., Park Ridge, N.J., 1979.
274. W. B. Earl, W. A. Brown: *3rd Int. Symp. Alcohol Fuels Technology*, Asilomar, Calif., May 1979.
275. A. A. Al-Talibi, N. D. Benjamin, A. R. Abboud, *Nahrung* **19** (1975) 335. Links
276. P. K. Agrawal: *Proc. 44th Am. Conv. Sugar Technologists Assn.*, Kanpur, India, 1980.
277. K. Nand, S. Srikanta, V. Screenivasa-Purthy, *J. Food Sci. Technol.* **14** (1977) 80.
278. J. A. Bassham: *2nd Int. Symp. Nonconventional Energy*, Trieste, Italy, Jun. 1981.
279. P. S. Nobel in [107] p. 13.
280. W. A. Scheller: Fermentation in Cereal Processing”, *61st National Meeting American Assn. Cereal Chemists,* New Orleans, Oct. 1976.
281. Alltech, Lexington, Ky. 1981.
282. E. L. Core: *Plant Taxonomy*, Prentice Hall, Englewood Cliffs, N. J., 1955, p. 346.
283. S. K. Chan, Investigation at the Federal Experiment Station, Serdang, Malaysia, 1969.
284. T. J. B. De Menezes, T. Arakaki, P. De Lamo, A. Sales, *Biotechnol. Bioeng.* **20** (1978) 555.
285. E. A. Jackson, *Proc. Biochem.* **11** (1976) no. 5, 29.
286. T. A. McClure, M. F. Arthur, S. Kresovich, D. A. Scantland in [107] p. 123.
287. T. Schormüller: *Lehrbuch der Lebensmittelchemie*, Springer-Verlag, Berlin-Heidelberg 1974, p. 506.
288. N. Kosaric, G. P. Cosentino, A. Wieczorek, Z. Duvnjak, *Biomass* **5** (1984) 1 – 36.
289. F. Schneider, F. W. Conti, *Stärke* **8** (1956) 269.
290. N. Kosaric, A. Wieczorek, Z. Duvnjak, S. Kliza: *Bioenergy Research and Development Seminar*, Winnipeg, Canada Mar. 29 – 31, 1982.
291. J. P. Holdren, P. R. Ehrlich, *Am. Sci.* **62** (1974) 282. Links
292. R. H. Whittaker: *Communities and Ecosystems*, MacMillan, New York 1970.
293. M. Slesser, C. Lewis: “Biological Energy Resources”, E & FN Spon Ltd., London, 1979.
294. T. K. Ghose, P. Ghosh, *J. App. Chem. Biotechnol.* **28** (1978) 309.
295. C. D. Callihan, J. E. Clemmer in A. H. Rose (ed.): *Microbial Biomass*, Academic Press, New York 1979, p. 271.
296. E. Adler, *Wood Sci. Technol.* **11** (1977) 169. Links

297. M. Wayman in [107]
298. L. Spano, T. Tassinari, D. D.Y. Ryu, A. Allen *et al.*, *29th Canadian Chem. Eng. Conf.*, Sarnia, Canada Oct., 1979.
299. A. E. Hokanson, R. Katzen, *Chem. Eng. Prog.* **74** (1978) no. 1, 67.
300. J. Yu, S. F. Miller, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 237.
301. G. L. Foutch, G. L. Margruder, J. L. Gaddy: "Agricultural Energy", *Biomass Energy Crop Production*, vol. **2**, American Society of Agricultural Engineers, St. Joseph, Mich., 1980, p. 299.
302. A. Porteous, *Pap. Trade J.* **156** (1972) no. 6, 30.
303. R. W. Detroy, C. W. Hesseltine, *Process Biochem.* **13** (1978) no. 9, 2. Links
304. A. J. Forage, R. C. Righelato, in [295] p. 289.
305. E. O. Ericsson, *Chem. Eng. Prog.* **43** (1947) no. 4, 165. Links
306. N. Kosaric, Y. J. Asher: "The Utilization of Cheese Whey and Its Components", in A. Fiechter (ed.): *Advances in Biochemical Engineering Biotechnology*, **32**, Springer-Verlag, Berlin-Heidelberg 1985, pp. 25 – 60.
307. S. M. Drews, *Ber. Landwirtsch. Sonderh.* **192** (1975) 599.
308. V. S. O'Leary, R. Green, B. C. Sullivan, V. H. Holsinger, *Biotechnol. Bioeng.* **19** (1977) 1019.
309. *Chem. Eng. News* **56** (1978) no. 17, 21.
310. S. Elias, *Food Eng.* **51** (1979) no. 1, 99.
311. L. Reesen, *Dairy Ind. Int.* **43** (1978) no. 1, 9. Links
312. E. T. Reese, *Biotechnol. Bioeng. Symp.* **5** (1975) 77. Links
313. G. T. Austin: *Shreve's Chemical Process Industries*, 5th ed., McGraw-Hill, New York 1984, pp. 581 – 590.
314. CORA Engineering: *Production of Ethanol from Renewable Resources*, Conceptual Engineering Manual, Chur, Switzerland, 1981, rev. 1986.
315. R. Katzen: "Low Energy Distillation Systems", *Bio-Energy Conference*, Atlanta 1985.
316. R. Katzen: "Large Scale Ethanol Production Facilities", *Bio-Energy '84 World Conference*, Gothenburg, Sweden 1984.
317. J. M. Paturau: *By-Products of the Cane Sugar Industry*, Elsevier, Amsterdam 1975, p. 183.
318. N. Kosaric *et al.* in H. J. Rehm, G. Reed (eds.): *Biotechnology*, vol. **3**, Verlag Chemie, Weinheim, Germany 1983, pp. 345 – 358.
319. Technipetrol, "Development and Progress in Ethylalcohol Technology", Rome 1982.
320. U. Tegtmeier, *Biotechnology Letters* **7** (1985) no. 2, 129 – 134. Links
321. *Chem. Eng.* **88** (1981) no. 11, 29.
322. *Chem. Eng.* **81** (1980) no. 18, 19.
323. *Chem. Eng.* **92** (1985) no. 10, 14.
324. M. A. Scharf: "Prism Separators Remove Water from Ethanol", *Monsanto Newsletter*, Jun. 1986, Monsanto Europe, Brussels.
325. *Chem. Eng.* **88** (1981) no. 19, 17.
326. *Chem. Eng.* **87** (1980) no. 23, 103.
327. J. Kaschemekat, J. Barbknecht, K. W. Böddeker, *Chem. Ing. Tech.* **58** (1986) no. 9, 740 – 742.
328. Vogelbusch Co.: *Alcohol Pervaporation*, Information 2/84, Vienna 1984.
329. Paterson Candy International, "Reverse Osmosis Concentration and Ultrafiltration Separation Applications in Biochemical Processes", Whitechurch, United Kingdom, 1985.
330. *Chem. Week* (1979) Jun. 6, 42 – 44.
331. A. Chauvel et al.: *"Procédés de pétrochimie – Charactéristiques techniques et économiques",* vol. 2, Editions Technip, Paris 1980.
332. J. A. Roels *et al.*: "Biotechnology and Base Chemicals", *Int. Conf. on Biomass*, Venice, Mar. 1985.
333. CORA Engineering: *Production of Ethanol from Renewable Resources*, Conceptual Engineering Manual, Chur, Switzerland, 1981, rev. 1986.
334. S. Williams (ed.): *Official Methods of Analysis of the Assn. of Official Analytical Chemists*, 14th ed., Assn. of Official Analytical Chemists, Arlington, Va., 1984.
335. M. A. Amerine, S. C. Ough: *Methods for Analysis of Musts and Wines*, J. Wiley & Sons, New York 1980.
336. *Pure Appl. Chem.* **17** (1968) 273 – 312. Links
337. "Petroleum Products, Lubricants and Fossil Fuels", *1985 Annual Book of ASTM Standards*, vol. 05.01 (1), ASTM, Philadelphia 1985, D56 – D1660.
338. T. S. Ma, R. E. Lang: "General Principles", *Quantitative Analysis of Organic Mixtures*, J. Wiley & Sons, New York 1979, Part 1, pp. 264 – 268.
339. H. W. Zimmermann, *Am. J. Enol. Vitic.* **14** (1963) 205 – 213.
340. H. Rebelein, *Allg. Dtsch. Weinfachztg.* **107** (1971) 590 – 594.
341. V. C. Mehlenbacher: "Determination of Hydroxyl Groups", in J. Mitchell, Jr., I. M. Kolthoff, E. S. Proskaner, A. Weissberger (eds.): *Organic Analysis,* vol. 1, Interscience, New York 1953, pp. 1 – 65.
342. M. Stiller, *Anal. Chim. Acta* **25** (1961) 85 – 89. Links
343. S. P. Colowichk, N. O. Kaplan, in D. B. McCormick, L. D. Wright (eds.): *Methods in Enzymology*, vol. **66**, Academic Press, New York 1980, pp. 41 –42.
344. G. G. Guilbault: *Analytical Uses of Immobilized Enzymes*, Marcel Dekker, New York 1984, pp. 112 – 230.
345. *Chem. Eng. News* **55** (1977) no. 2, 12.

346. Alcohol Production from Biomass in the Developing Countries, World Bank, Washington, D.C., Sept. 1980.
347. Statistical Release, Dept. of Treasury, Bureau of Alcohol, Tobacco and Firearms, Washington, D.C., Jun. 16, 1982.
348. United States Department of Agriculture, Foreign Agricultural Service, Foreign Agricultural Circular FS 1 – 86, May 1986.
349. H. Rothman, R. Greenshield, R. F. Calte: *Energy from Alcohol: The Brazilian Experience*, Univ. Press of Kentucky, Lexington, Ky., 1983, pp. 75 – 85.
350. "Ethanol – Chemical Profile", *Chem. Mark. Rep*. (1985), Feb. 25.
351. R. Katzen: "Large Scale Ethanol Production Facilities", *Bio-Energy '84 World Conference*, Gothenburg, Sweden 1984.
352. CORA Engineering: *Production of Ethanol from Renewable Resources*, Conceptual Engineering Manual, Chur, Switzerland 1981, rev. 1986.
353. H. F. Smyth Jr., *J. Ind Hyg. Toxicol.* **23** (1941) 253.
354. E. T. Kimura, D. M. Ebert, P. W. Dodge, *Toxicol. Appl. Pharmacol.* **19** (1971) 699. Links
355. W. S. Spector (ed.): "Acute Toxicities", *Handbook of Toxicol.*, vol. **1**, Saunders, Philadelphia-London 1956, p. 128.
356. J. C. Munch, E. W. Schwartze, *J. Lab Clin. Med.* **10** (1925) 985. Links
357. W. L. Guess, *Toxicol. Appl. Pharmacol.* **16** (1970) 382. Links
358. L. Phillips, M. Steinberg, H. Maibach, W. Akers *Toxicol. Appl. Pharmacol.* **21** (1972) 369.
359. K. B. Lehmann, F. Flury: *Toxicologie und Hygiene der technischen Lösungsmittel*, Springer-Verlag, Berlin 1938, p. 152.
360. A. Loewy, R. van der Heyde, *Biochem. Z.* **86** (1918) 125.
361. J. L. Hall, D. T. Rowlands, *Am. J. Pathol.* **60** (1970) 153. Links
362. S. C. Vasder, R. N. Chakravarti, D. Subrahmanyam, A. C. Jain *et al.*, *Cardiovasc. Res.* **9** (1975) 134. Links
363. A. M. Skosyreva, *Akush. Ginekol. (Sofia)* **4** (1973) 15.
364. W. J. Tze, M. Lee, *Nature (London)* **257** (1975) 479.
365. F. M. Badr, R. S. Badr, *Nature (London)* **253** (1975) 134.
366. L. Machemer, D. Lorke, *Mutat. Res.* **29** (1975) 209. Links
367. J. McCann, E. Choi, E. Yamasaki, B. Ames, *Proc. Nat. Acad. Sci. USA* **72** (1975) 5135.
368. L. F. Meisner, S. L. Juhorn, *Acta Cytologica* **16** (1972) 41. Links
369. V. K. Rowe, S. B. McCollister in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, 3rd ed., vol. **2 C**, Wiley-Interscience, New York 1982, p. 4549.
370. R. Derache, *Int. Encycl. Pharmacol. Ther.* **20** (1970) 507.
371. M. K. Roach, W. N. Reese, P. J. Creaven, *Biochem. Biophys. Res. Commun.* **36** (1969) 596. Links
372. E. Umdagard, *C. R. Trav. Lab. Carlsberg* **22** (1938) 333.
373. E. Jacobsen, *Pharmacol. Rev.* **4** (1952) 107. Links
374. O. Farsander, *Int. Encycl. Pharmacol. Ther.* **20** (1967) 117.
375. E. Petri in F. Hencke, O. Lubarsch (eds.): *Handbuch der Speziellen Pathologischen Anatomie und Histologie*, vol. **10**, Springer-Verlag, Berlin 1930, p. 226.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Ethanolamines and Propanolamines

Matthias Frauenkron, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chaps. 1 – 7)

Johann-Peter Melder, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chaps. 1 – 7)

Günther Ruider, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chaps. 1 – 7)

Roland Rossbacher, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 8)

Hartmut Höke, Weinheim, Germany (Chap. 8)

1. Introduction 2
2. Ethanolamines 2
2.1. Properties 2
2.1.1. Physical Properties 2
2.1.2. Chemical Properties 2
2.2. Production 4
2.3. Quality Specifications 5
2.4. Uses 5
2.5. Economic Aspects 7
3. *N*-Alkylated Ethanolamines 8
3.1. Properties 8
3.1.1. Physical Properties 8
3.1.2. Chemical Properties 8
3.2. Production 8
3.3. Quality Specifications 11
3.4. Uses and Economic Aspects 11
4. Isopropanolamines 11
4.1. Properties 12
4.1.1. Physical Properties 12
4.1.2. Chemical Properties 13
4.2. Production 14
4.3. Quality Specifications 15
4.4. Uses 15
5. *N*-Alkylated Propanolamines and 3-Alkoxypropylamines 16
5.1. Properties 16
5.1.1. Physical Properties 16
5.1.2. Chemical Properties 16
5.2. Production 18
5.3. Quality Specifications 18
5.4. Uses and Economic Aspects 19
6. Storage and Transportation 19
7. Environmental Protection 19
8. Toxicology and Occupational Health 19
8.1. General Aspects 19
8.2. Ethanolamines 20
8.3. *N*-Alkyl- und *N*-Arylethanolamines 24
8.4. Propanolamines 26
9. References 27

1. Introduction

Amino alcohols have been prepared industrially since the 1930s. However, large-scale production started only after 1945, when alkoxylation with ethylene oxide and propylene oxide replaced the older chlorohydrin route. In industry, amino alcohols are usually designated as alkanolamines. Ethanolamines (aminoethanols) and propanolamines (aminopropanols) are by far the most important compounds of this group. Both are used widely in the manufacture of surfactants and in gas purification [1–6].

2. Ethanolamines

Monoethanolamine [*141-43-5*] (**1**) (MEA; 2-aminoethanol), diethanolamine [*111-42-2*] (**2**) (DEA; 2,2′-iminodiethanol), and triethanolamine [*102-71-6*] (**3**) (TEA; 2,2′,2″-nitrilotriethanol) can be regarded as derivatives of ammonia in which one, two, or three hydrogen atoms have been replaced by a CH_2CH_2OH group.

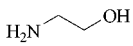

1

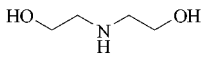

2

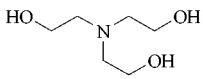

3

Ethanolamines were prepared in 1860 by Wurtz from ethylene chlorohydrin and aqueous ammonia [7]. It was only toward the end of the 19th century that an ethanolamine mixture was separated into its mono-, di-, and trieth-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_001

anolamine components; this was achieved by fractional distillation.

Ethanolamines were not available commercially before the early 1930s; they assumed steadily growing commercial importance as intermediates only after 1945, because of the large-scale production of ethylene oxide. Since the mid-1970s, production of very pure, colorless triethanolamine in industrial quantities has been possible [117–119]. All ethanolamines can now be obtained economically in very pure form.

The most important uses of ethanolamines are in the production of emulsifiers, detergent raw materials, and textile chemicals; in gas purification processes; in cement production, as milling additives; and as building blocks for agrochemicals [120]. Monoethanolamine is an important feedstock for the production of ethylenediamine and ethylenimine.

2.1. Properties

2.1.1. Physical Properties

Monoethanolamine and triethanolamine are viscous, colorless, clear, hygroscopic liquids at room temperature; diethanolamine is a crystalline solid. All ethanolamines absorb water and carbon dioxide from the air and are infinitely miscible with water and alcohols. The freezing points of all ethanolamines can be lowered considerably by the addition of water. Some physical properties of ethanolamines are listed in Table 1.

2.1.2. Chemical Properties

Because of their basic amino group and the hydroxyl group, ethanolamines have chemical properties resembling those of both amines and alcohols. They form salts with acids, and the hydroxyl group permits ester formation. When mono- and diethanolamine react with organic acids, salt formation always takes place in preference to ester formation. With weak inorganic acids, e.g., H_2S and CO_2, thermally unstable salts are formed in aqueous solution. This reaction of ethanolamines is the basis for their application in the purification of acidic natural gas, refinery gas, and synthesis gas [1], [8]. In the absence of water, monoethanolamine and diethanolamine react with CO_2 to form carbamates:

$$2\,HOCH_2CH_2NHR + CO_2 \longrightarrow$$
$$HOCH_2CH_2NRCOOH \cdot RHNCH_2CH_2OH$$

$R = H$ or CH_2CH_2OH

Triethanolamine does not form a carbamate.

By metal-catalyzed amination, the hydroxyl group of monoethanolamine can be replaced by an amino group to form ethyleneamines such as ethylenediamine [*107-15-3*], diethylenetriamine [*111-40-0*], piperazine [*110-85-0*], and aminoethylethanolamine [*111-41-1*] [9], [10]:

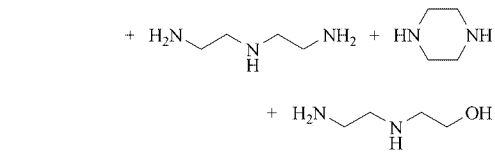

Considerable quantities of monoethanolamine are converted into ethylenimine (aziridine [*151-56-4*]) by adding sulfuric acid and cyclizing the hydrogensulfate with sodium hydroxide or by using heterogeneous catalysts [121] (→ Aziridines, Chap. 4.3.) [11], [12]:

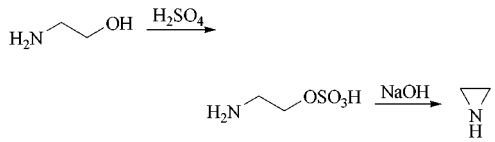

As primary and secondary amines, monoethanolamine and diethanolamine also react with acids or acid chlorides to form amides. The amine first reacts with the acid, e.g., stearic acid, to form a salt, which can be dehydrated to the amide by heating:

$$NH_2CH_2CH_2OH + C_{17}H_{35}COOH$$
$$\longrightarrow NH_2CH_2CH_2OH \cdot C_{17}H_{35}COOH$$
$$\longrightarrow HOCH_2CH_2\text{–}NH\text{–}CO\text{–}C_{17}H_{35}COOH + H_2O$$

When the ethanolamides of fatty acids are heated at fairly high temperature with removal of water, oxazolines are formed:

Table 1. Physical properties of ethanolamines

Compound	M_r	*mp*, °C	*bp* (101.3 kPa), °C	Density[a] ϱ (20 °C), g/cm^3	Heat of vaporization (101.3 kPa), kJ/kg	Specific heat c_p kJ kg^{-1} K^{-1}	Cubic expansion coefficient, K^{-1}
Monoethanolamine	61.08	10.53	170.3	1.0160	848.1	2.72	7.78×10^{-4}
Diethanolamine	105.14	28.0	268.5	1.0912 (30 °C)	638.4	2.73	5.86×10^{-4}
Triethanolamine	149.19	21.6	336.1	1.1248	517.8	2.33	4.82×10^{-4}

Compound	Viscosity (20 °C), mPa · s	n_D^{20}	Surface tension (20 °C), N/m	Flash point,[b] °C	Ignition temperature,[b], °C	Temperature class[d]
Monoethanolamine	23.2	1.4544	0.049	94.5	410	T2
Diethanolamine	389 (30 °C)	1.4747	0.0477	176	365	T2
Triethanolamine	930	1.4852	0.0484	192	325	T2

[a] According to DIN 53 217.
[b] According to DIN 51 758.
[c] According to DIN 51 794.
[d] According to VDE 0165. Ethanolamines do not belong to any hazard class according to VbF (regulation governing flammable liquids).

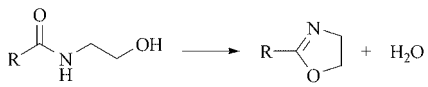

Formaldehyde reacts with monoethanolamine and diethanolamine to form hydroxymethyl compounds, which can be reduced to the *N*-methyl derivatives [14] (see Section 3.2).

Monoethanolamine reacts with carbon disulfide to form 2-mercaptothiazoline [*96-53-7*] [15]:

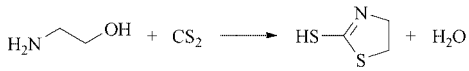

The hydroxyl groups of ethanolamines can be replaced with chlorine by reaction with thionyl chloride or phosphorus pentachloride. The chloroethylamines formed are hazardous because of their skin toxicity. For example, tris(2-chloroethyl)amine has been used as a gas in warfare (N-lost). Bis(2-chloroethyl)amine is obtained in good yield by reaction of diethanolamine with thionyl chloride:

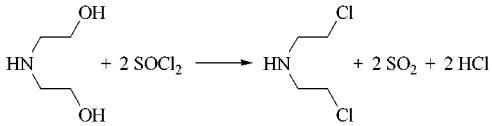

Reaction of triethanolamine with ethylene oxide gives unstable alkaline quaternary compounds and the corresponding stable ethers [122]. For example, undistilled triethanolamine prepared from mono- or diethanolamine always contains some triethanolamine monoglycol ether [16].

Mono- or diethanolamine can also be used as the amine component in aminoalkylation, the so-called Mannich reaction, which is very important in the biosynthesis of many alkaloids [17].

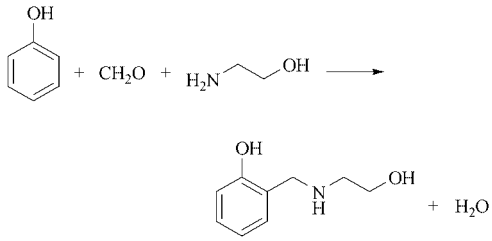

Catalytic dehydrogenation of diethanolamine leads to iminodiacetic acid, an important building block for agrochemicals [120].

N-Nitrosamines can be formed by treating diethanolamine or, under suitable conditions, triethanolamine with nitrous acid, nitrites, or oxides of nitrogen. Animal experiments have shown that *N*-nitrosamines are carcinogenic [18].

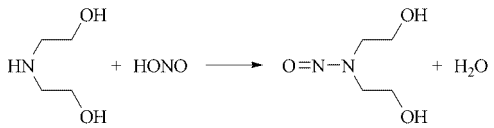

Monoethanolamine and triethanolamine can form complexes with transition metal ions (e.g., chromium, copper, nickel, cobalt, and iron); some of these complexes are water-soluble [123–125].

2.2. Production

Ethanolamines are produced on an industrial scale exclusively by reaction of ethylene oxide with excess ammonia, this excess being considerable in some cases [19].

The reaction of ethylene oxide with ammonia takes place slowly and is accelerated by water. Anhydrous procedures employ a fixed-bed catalyst consisting of an organic ion-exchange resin or thermally more stable acidic inorganic clays or zeolites [20–23].

In all conventional processes, reaction takes place in the liquid phase, and the reactor pressure is usually sufficiently large to prevent vaporization of ammonia and ethylene oxide at the reaction temperature. In current procedures, ammonia concentrations in water between 50 and 100%, pressures up to 16 MPa, reaction temperatures up to 150 °C, and an excess up to 40 mol of ammonia per mole of ethylene oxide are used. The reaction is highly exothermic; the enthalpy of reaction is about 125 kJ per mole of ethylene oxide [24]. The following competing reactions occur:

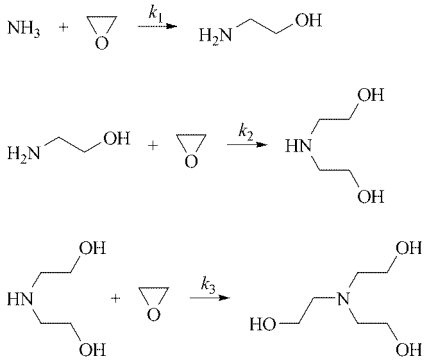

Side reactions forming tetrakis(hydroxyethyl)ammonium hydroxide or any ethers are of no importance in the synthesis [126]. The kinetics of these reactions have been studied [25].

All the reaction steps have about the same activation energy and show a roughly quadratic dependence of the reaction rate on the water content of the ammonia – water mixture used. Therefore, product composition depends solely on the molar excess of ammonia and not on water content, reaction temperature, or pressure [26]. The product distribution as a function of the molar ratio of the reactants is shown in Figure 1 [27].

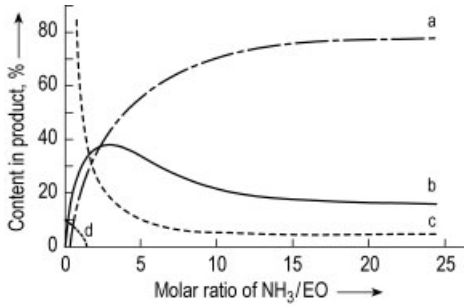

Figure 1. Product distribution of monoethanolamine (a), diethanolamine (b), triethanolamine (c), and glycol ethers of triethanolamine as a function of the molar ratio of ammonia to ethylene oxide (EO) in aqueous solution at 100 – 200 °C

Unconsumed ammonia and water are separated from the products in a distillation line downstream of the reactor and are recycled. In large-scale continuous single-line plants, the requirement for low energy use (i.e., operation with minimum steam consumption) determines both the transport of heat from the reactor and the design of the thermally integrated distillation line (Fig. 2).

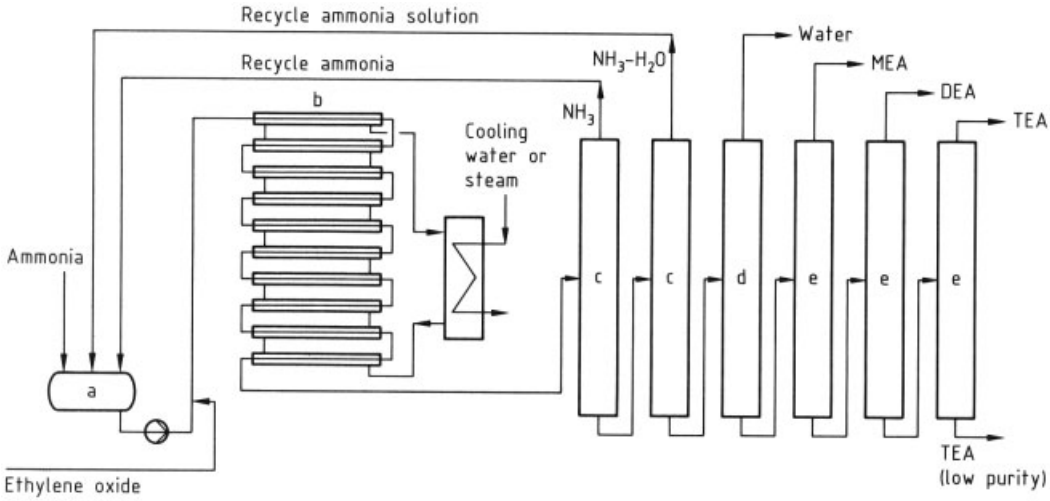

Figure 2. Flow sheet for the production of ethanolamines
a) Aqueous ammonia tank; b) Tubular reactor; c) Ammonia and ammonia solution columns; d) Dehydration column; e) Vacuum fractionation columns

Product distribution of the three ethanolamines can be controlled by appropriate choice of the ammonia : ethylene oxide ratio. A higher diethanolamine or triethanolamine content can also be obtained by recycling monoethanolamine or diethanolamine to the reactor or by treating them with ethylene oxide in a separate unit [127], [128].

For safety reasons, ethylene oxide must be metered into the ammonia stream; in the reverse procedure, ammonia or amines may cause ethylene oxide to undergo an explosive polymerization reaction.

In all industrial processes, complete conversion to the three ethanolamines, without significant formation of byproducts, is achieved. Therefore, feedstock costs are independent of the type of production process. On the other hand, manufacturing costs and, in particular, energy costs depend to a large extent on the product composition desired, plant design, and degree of thermal integration.

2.3. Quality Specifications

Today, all ethanolamines are prepared with $>99\,\%$ purity. Water, the two other ethanolamines, and a small amount of triethanolamine glycol ether are minor components; all other impurities are in the ppm range. Table 2 gives specifications for commercial ethanolamines.

Purity is determined by gas chromatography, and the residual water content by the Karl Fischer method. After pretreatment, the *N*-nitrosamine content can also be determined by gas chromatography with a thermal analyzer.

2.4. Uses

Surfactants. Ethanolamines are used widely as intermediates in the production of surfactants, which have become commercially important as detergents, textile and leather chemicals, and emulsifiers [2–6]. Their uses range from drilling and cutting oils to medicinal soaps and high-quality toiletries [28] (Fig. 3).

Properties of ethanolamine derivatives can be varied within wide limits by appropriate choice of the ethanolamine, the acid component, and their ratio. Synergistic effects in complex emulsifier systems can be obtained by simultaneous use of different combinations of amines and acids. Ethanolamines, particularly triethanolamine, which had presented problems with regard to color, are now produced in sufficient purity that no color problems arise when they are treated with acids or heated [29–32].

Table 2. Specifications of commercial ethanolamines

Compound	Purity, wt %	ϱ (DIN 51 757), g/cm^3	Color (DIN-ISO 6271), APHA
Monoethanolamine	>99	1.016 (20 °C)	≤ 10
Diethanolamine	>99	1.0912 (30 °C)	≤ 20
Triethanolamine	>99	1.1248 (20 °C)	≤ 30

Ethanolamine-based surfactants can be formulated as weakly basic or neutral products and are therefore particularly well tolerated by the skin; this is especially true for triethanolamine soaps. Moreover, they are noncorrosive and can be used on virtually all textiles without damage. Ethanolamine soaps prepared from oleic acid, stearic acid, lauric acid, or caprylic acid are constituents of many toiletries and medicinal soaps.

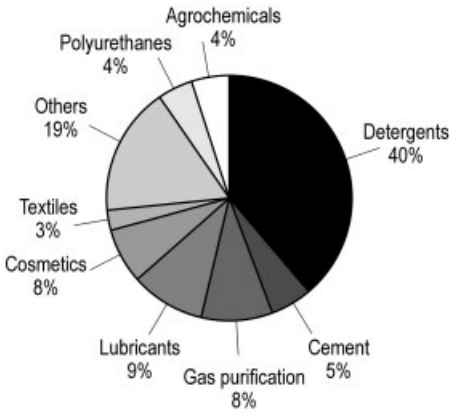

Figure 3. Percentage consumption of ethanolamines according to use (1996)

Ethanolamine soaps produced from fatty acids are among the most industrially important emulsifiers. They are used in cosmetics [33], polishes, shoe creams, car care products, drilling and cutting oils, and pharmaceutical ointments. Ethanolamine soaps combined with wax and resins are used as impregnating materials, protective coatings, and products for the care of textile and leather goods.

Ethanolamine soaps obtained from alkylarylsulfonic acids, preferably alkylbenzenesulfonic acids, or from alcohol sulfates are growing steadily in importance and dominate the market for household cleansers [34]. The use of linear alkyl groups instead of branched chains in these products has resulted in greater biodegradability.

Fatty acid ethanolamides and products obtained by further ethoxylation have foam-stabilizing ability and are therefore important additives for detergents [35], [36]. Diethanolamides obtained from coconut fatty acids and from oleic acid are used industrially. These amides were first described in 1937 [37]. Liquid detergents are based predominantly on triethanolamine [38–40].

In the manufacture of leather, ethanolamine-based chemicals are used for dressing, dyeing, and finishing. For paints and coatings, ethanolamines are employed both in production and in softeners and paint removers.

Corrosion Inhibitors. Diethanolamine and triethanolamine are important components of corrosion inhibitors, particularly in coolants for automobile engines, as well as in drilling and cutting oils. They are also employed as additives in lubricants.

When the corrosion inhibitor sodium nitrite and diethanolamine are used together, *N*-nitrosamines may form. These nitrosamines are carcinogenic; therefore, sodium nitrite and diethanolamine or triethanolamine should not be used simultaneously.

Gas Purification. Large amounts of ethanolamines, principally diethanolamine, are used in absorptive gas purification to remove weakly acidic components. For example, hydrogen sulfide and carbon dioxide are removed from natural gas, refinery gas, and synthesis gas [1].

Intermediates. A substantial portion of the monoethanolamine is used as a starting material in the preparation of ethylenimine and ethylenediamine. Ethylenimine is converted to polyethylenimine, an important chemical in paper technology. Diethanolamine is used as a building block for agrochemicals (e.g., Glyphosphate).

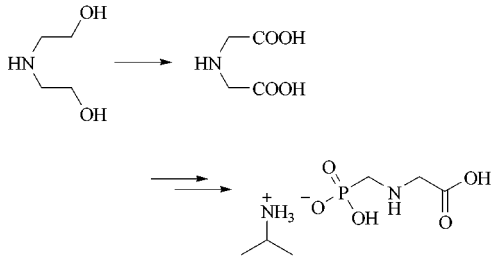

Cement Additives. Since the 1960s, aqueous solutions of triethanolamine and triethanolamine acetate have been used as milling additives in cement production [41]. During grinding of the clinker in ball mills, the small amount of triethanolamine added prevents agglomeration and cushioning of the grinding medium by saturating the surfaces of the freshly broken particles. Like other milling additives (e.g., various glycols), they reduce the power required for milling the clinker. The milling of virtually all high-quality portland cement involves milling additives. In addition to assisting the milling process, triethanolamine also improves the flow properties and setting behavior of the cement [42].

2.5. Economic Aspects

Large-scale, economical production of ethanolamines from ethylene oxide and ammonia in large, single-line plants has greatly promoted their use in many industrial sectors. In 1999, world ethanolamine capacity was about 1.09×10^6 t/a (Table 3). About 50 % of world production is monoethanolamine; 30 – 35%, diethanolamine; and 15 – 20%, triethanolamine.

Table 3. Approximate ethanolamine capacity and production in 1998

	Capacity, 10^3 t/a	Estimated production, 10^3 t/a
United States	569	456
Mexico	20	19
Western Europe	275	242
Southeast Asia	123	68
South America	30	16.8
Total	1017	759

3. *N*-Alkylated Ethanolamines

N-Alkylated ethanolamines are amino alcohols that contain a secondary or tertiary nitrogen atom and one or two hydroxyl groups.

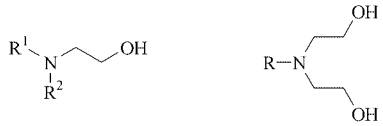

R^1 = alkyl, H
R^2 = alkyl, aryl, cycloalkyl

R = alkyl

Numerous compounds with different properties and uses can be synthesized by varying the alkyl group.

3.1. Properties

3.1.1. Physical Properties

N-Alkylated ethanolamines are generally liquid and have a faint amine-like odor. They tend to discolor on prolonged storage, especially if exposed to air, light, and temperatures above 40 °C. This tendency to discoloration can be reduced substantially by the use of additives when purifying the products by distillation, or by pretreating the crude products before distillation [43–45]. *N*-Alkylated ethanolamines are partially to readily soluble in alcohol, water, acetone, glycols, glycerol, and glycol ethers. They are only sparingly soluble or even insoluble in nonpolar solvents, e.g., aliphatic hydrocarbons. 2-Dimethylaminoethanol [*108-01-0*] (*N*,*N*-dimethylethanolamine) and 2-diethylaminoethanol [*100-37-8*] (*N*,*N*-diethylethanolamine) form azeotropes with water; *N*-methyldiethanolamine [*105-59-9*] does not form an azeotrope. All *N*-alkylethanolamines are hygroscopic and also absorb CO_2 from air. Physical properties of typical compounds are listed in Table 4.

3.1.2. Chemical Properties

The chemical properties of *N*-alkylated ethanolamines are very similar to those of unsubstituted ethanolamines (see Section 2.1.2). They

Table 4. Physical properties of *N*-alkylethanolamines

Compound	CAS registry number	M_r	*mp*, °C	*bp* (101.3 kPa), °C	Density[d] ρ, (20 °C) g/cm^3	Cubic expansion coefficient, K^{-1}	Heat of vaporization kJ/kg	Specific heat, kJ kg^{-1} K^{-1}	Specific electrical conductivity, Ω^{-1} cm^{-1}	Viscosity (20 °C), mPa · s	n_D^{20}	Surface tension (20 °C), mN/m	Flash point, °C	Ignition temperature, °C	Temperature class
Methylethanolamine	[*109-83-1*]	75.11	ca. −3	159	0.939 – 0.942	7.8×10^{-4}	564.4	2.55	3.33×10^{-6} (25 °C)	13.0	1.4389	34.4 (22 °C)	ca. 74	350	T2
Methyldiethanolamine	[*105-59-9*]	119.17	ca. −21	247	1.038 – 1.041	7.5×10^{-4}	418.7	1.72	8.1×10^{-7}	101	1.4694	40.9 (20.3 °C)	137	265	T3
Dimethylethanolamine	[*108-01-0*]	89.14	−70[a]	134	0.887	1.35×10^{-3}	454.3	2.30	3.8×10^{-7}	3.85	1.4296	27.1 (24.5 °C)	ca. 38	235	T3
Diethylethanolamine	[*100-37-8*]	117.19	−70[a]	162	0.883 – 0.889	1.07×10^{-3}	383.9	2.42	1.1×10^{-3}	5	1.4417	29.2	ca. 46	260	T3[b]
n-Butylethanolamine	[*111-75-1*]	117.19	−2.1	199	0.8917	7.6×10^{-4}	397.7	2.45	2.9×10^{-7}	19.7	1.4435	29.9 (22 °C)	85	265	T3
n-Butyldiethanolamine	[*102-79-4*]	161.25	−45[a]	ca. 270	0.970	7.7×10^{-4}	329.5	2.26	6.3×10^{-7}	75.9	1.462	33.9 (22 °C)	130	260	T3
Dibutylethanolamine	[*102-81-8*]	173.30	−75[a]	222 – 234	0.860	8.1×10^{-4}	277.6	2.32	9.5×10^{-7}	7.4	1.4444	26.3 (22 °C)	96	240	T3[c]
Cyclohexylethanolamine	[*2842-38-8*]	143.23	37.5	235	0.9797	7.6×10^{-4}	327.2	2.45	2.0×10^{-8}	317	1.4860	39.5 (25.1 °C)	122	270	T3
Cyclohexyldiethanolamine	[*4500-29-2*]	187.29		297	1.0339	6.6×10^{-4}	319.0	2.30	6.6×10^{-4}	815	1.4930	40.1	172	255	T3
1-(2-Hydroxyethyl)piperazine	[*103-76-4*]	130.19	24	112 (1.33 kPa)	1.0592	7.5×10^{-4}	411.1	2.00	6.1×10^{-7}	1040	1.110 (40 °C)	38.3	135	280	T3
4-(2-Hydoxyethyl)morpholine	[*622-40-2*]	131.18	1.5 – 2	224	1.0714	8.1×10^{-4}	342.7	1.91	4.7×10^{-5}	26.6	1.4783	40.8	113	205	T3
Hydroxyethylaniline	[*122-98-5*]	137.18	−30[a]	282 – 287	1.0954	1.14×10^{-3}	439.2	2.08		102	1.5796		ca. 148	410	T2
Ethylhydroxyethylaniline	[*92-50-2*]	165.24	37	270	1.030 (40 °C)	7.3×10^{-4}		2.14		86			139	335	T2
N-Ethylethanolamine	[*110-73-6*]	89.14	−5.8	167	0.914					15.1			78	330	T2
n-Propylethanolamine	[*16369-21-4*]	103.17	−0.5	180 – 181									84	290	T3
Hydroxyethylpiperidine	[*3040-44-6*]	129.21	16.7	198 – 203	0.973 (25 °C)					20.9	1.4804		83	240	T3
Dihydroxyethylaniline	[*120-07-0*]	181.24	52	346	1.0161 (200 °C)					1.27 (200 °C)			200	380	T2
tert-Butylethanolamine	[*4620-70-6*]	117.19	40 – 43	176 – 177	0.8576 (60 °C)					5.9 (60 °C)			81	345	T2
tert-Butyldiethanolamine	[*2160-93-2*]	161.25	44 – 47	136 – 139 (1.6 kPa)	0.9556 (60 °C)					21.0 (60 °C)			148	300	T3

[a] Pour point. [b] Hazard class AII (VbF). [c] Hazard class AIII (VbF).

Table 5. Specifications for commercial *N*-alkylethanolamines

Compound	Boiling range 5 – 95 mL [a], °C	Purity [c] (min.), %	Water content [d], %	Color [e], APHA
Dimethylethanolamine	133.5 – 135.5	99.8	< 0.05	15
Diethylethanolamine	161.5 – 163	99.5	< 0.05	15
Methylethanolamine	158.5 – 160.0	99.0	< 0.1	10
Methyldiethanolamine	115.5 – 118	99.0	< 0.2	50
4-(2-Hydroxyethyl)morpholine	223 – 225	99.0	0.2	40
1-(2-Hydroxyethyl)piperazine	121 – 123 [b]	99.0	0.2	40
1-(2-Hydroxyethyl)piperidine	92 – 96 [b]	99.5	0.1	40
Hydroxyethylaniline	282 – 287	99.0	0.3	100 – 150
Ethylethanolamine	167	99.5	0.2	15
Dibutylethanolamine	226 – 228	99.0	0.2	40
Butylethanolamine	196 – 198	99.5	0.1	15
n-Propylethanolamine	180 – 181	99.5	< 0.1	10
Butyldiethanolamine	264 – 278	99.0	< 0.3	10
tert-Butylethanolamine	178	99.0	< 0.3	
tert-Butyldiethanolamine	136 – 139 (1.6 kPa	99.0	< 0.3	

[a] DIN 51 751.
[b] DIN 53 406.
[c] Determined by gas chromatography.
[d] Determined by Karl Fischer method.
[e] According to DIN-ISO 6271.

form unstable salts with inorganic acids such as CO_2 and H_2S [129–131]. Salts with organic acids give a neutral to weakly basic reaction. *N*-Alkylated ethanolamines also react with acids or acid derivatives to give the corresponding esters. Important products are dimethylethanolamine acrylate [*2439-35-2*] and diethylethanolamine acrylate [*2426-54-2*], which are used as intermediates in the synthesis of flocculating agents.

Secondary *N*-alkylethanolamines also react with acids to form acid amides. The reaction of *N*-alkylethanolamines with ethylene oxide to give longer chain alkoxy derivatives gives low yields.

3.2. Production

Industrially, *N*-alkylated ethanolamines are produced almost exclusively by batchwise or continuous reaction of primary, secondary, or tertiary amines with ethylene oxide [46–48]. For safety reasons, ethylene oxide must be added to the amine, never vice versa (see Section 2.2). Depending on the product involved, reaction temperature varies from 50 to 170 °C and pressure from 0.2 to 4 MPa. Water accelerates the reaction.

Figure 4 shows a flow sheet for the *batchwise production* of alkanolamines. The procedure is also suitable for the manufacture of isopropanolamines (see Chap. 5); in this case, propylene oxide is used instead of ethylene oxide.

The oxygen-free reaction kettle (a) is filled with the amine raw material from the stock tank (b); water is added in an amount of 10 – 60 %, based on the amine; and the kettle is heated to a minimum temperature around 50 °C. Metering of ethylene oxide is then begun, the feed rate depending on the removal of heat and the operating pressure of the reaction kettle.

The optimum amount of ethylene oxide in relation to starting amine is determined by cost-effectiveness (byproducts, work-up costs). When addition of ethylene oxide is complete, the minimum reaction temperature is maintained for some time to ensure that the product discharged from the reactor is free of ethylene oxide. The reaction mixture is separated by fractional distillation. Unconverted amine and water are removed as the first fraction and recycled to the subsequent batch. Pure alkanolamines are obtained by distillation under reduced pressure.

When *primary amines* are used, a mixture of *N*-alkylethanolamines and *N*-alkyldiethanolamines is always formed:

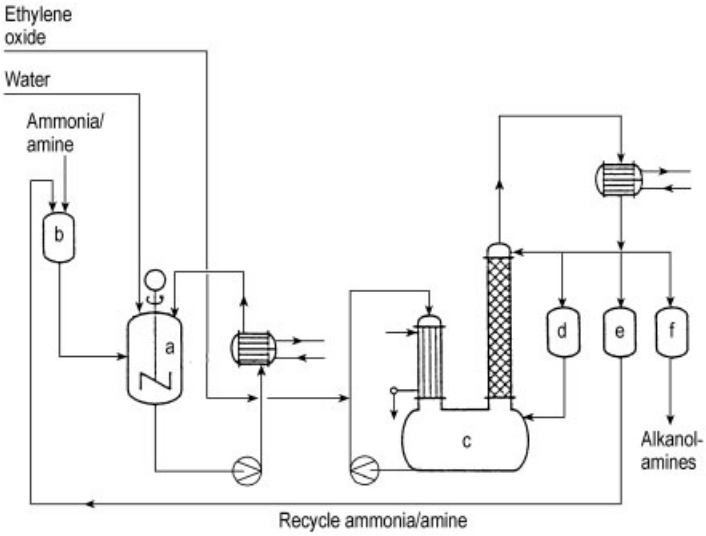

Figure 4. Flow sheet for the batch production of alkanolamines
a) Reaction kettle; b) Ammonia tank; c) Vessel; d, e, f) Storage tanks

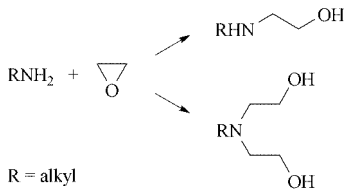

The larger the excess of amine, the lower is the proportion of *N*-alkyldiethanolamines. If *N*-alkylethanolamines are desired, a continuous process is usually preferred for economic reasons. Excess amine is distilled from the reaction product and recycled.

With *secondary amines,* only one product is formed:

Trialkylammonium chlorides yield the corresponding 2-hydroxyethylammonium chlorides. One of the commercially most important salts is choline chloride [*67-48-1*] (→ Choline):

Choline chloride can also be prepared from trimethylamine and ethylene chlorohydrin, but this route has no commercial significance.

Another common method of preparing *N*-alkylethanolamines is by *N*-alkylation of ethanolamines [49]. Methyl derivatives can also be obtained by reacting monoethanolamine or diethanolamine, or their homologues, with formaldehyde and then reducing the hydroxymethyl compound:

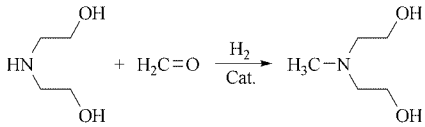

Pure components are obtained from the crude products by distillation.

The reaction of epichlorohydrin with amines to give *N*-alkylethanolamines is of only minor commercial importance.

3.3. Quality Specifications

Table 5 lists specifications of commercial products. Purity is almost always measured by gas chromatography.

3.4. Uses and Economic Aspects

N-Alkylethanolamines are used mainly as intermediates, especially in the production of pharmaceuticals such as the local anesthetics procaine (diethylethanolamine 4-aminobenzoate [*59-46-1*]) and tetracaine (dimethylethanolamine 4-butylaminobenzoate [*136-47-0*]), crop protection agents, and flocculants (dimethylethanolamine acrylate) [50]. They are also important in the preparation of chemicals for the paper and leather industries. Use of *N*-alkylethanolamines in the production of plastics has risen substantially in recent years.

Direct uses of *N*-alkylethanolamines, especially methyldiethanolamine [51], include gas-purification methods for removing acidic gases (CO_2, H_2S). Choline chloride is very important in the animal feedstuff industry (→ Choline) [52]. World production was ca. 160 000 t/a in 1999.

Annual worldwide demand for *N*-alkylethanolamines (excluding choline chloride) ranges from 2 to 15 000 t, depending on the product and application. Some products in the category of more than 1000 t/a are *N*,*N*-dimethylethanolamine, *N*,*N*-diethylethanolamine, 1-(2-hydroxyethyl)piperazine derivatives, *N*-methyldiethanolamine, and *N*-methylethanolamine.

4. Isopropanolamines

The propanolamines described here are propanol derivatives containing one amino group. They have the following general formulas:

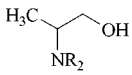

4

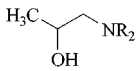

5

R = H, C_3H_6OH

2-Amino-1-propanols (**8**) and 1-amino-2-propanols (**9**) have been prepared industrially since 1937 by reacting propylene oxide with the appropriate amine. The products are mixtures which are processed further; the isomer mixtures are not usually separated on an industrial scale. The most important members of this class of compounds are the isopropanolamines, which are used as mixtures of various isomers. World capacity for isopropanolamines is estimated at 40 000 t/a.

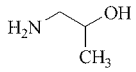

Isopropanolamine

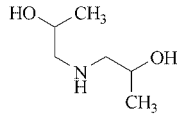

Diisopropanolamine

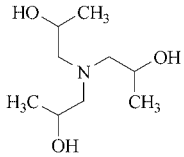

Triisopropanolamine

3-Amino-1-propanol, $H_2NCH_2CH_2CH_2OH$, which is obtained not from propylene oxide but from hydrogenation of ethylene cyanohydrin, is discussed in Chapter 5.

4.1. Properties

4.1.1. Physical Properties

Isopropanolamines are hygroscopic, virtually colorless substances with a slight amine-like odor. They are readily soluble in water, ethanol, glycol, and acetone but only slightly soluble in hydrocarbons and diethyl ether. Monoisopropanolamine is more soluble than monoethanolamine in organic solvents; it forms a miscibility gap with heptane and toluene.

Diisopropanolamine and triisopropanolamine undergo thermal decomposition above ca. 170 °C and can therefore be distilled only at reduced pressure. The freezing points of diisopropanolamine and triisopropanolamine can be greatly depressed by adding a small amount of water.

All isopropanolamines have at least one asymmetrically substituted carbon atom and thus can be optically active. The commercial products are racemic mixtures.

Important physical data are listed in Table 6 and safety data in Table 7.

Table 6. Physical properties of propanolamines

Compound	CAS registry number	M_r	*fp*, °C	*bp* (101.3 kPa), °C	*ϱ* (20 °C, g/cm^3	Cubic expansion coefficient (40–100 °C), K^{-1}	Heat of vaporization, kJ/kg	Specific heat, kJ kg^{-1} K^{-1}	Viscosity, mPa · s	Refractive index n_D
1-Amino-2-propanol (*d,l*)	[*78-96-6*]	75.11	−0.3	159.7						1.4466 (25 °C)
2-Amino-1-propanol (*d,l*)	[*6168-72-5*]	75.11	+7.5	168.5						1.4481 (25 °C)
Monoisopropanolamine [a]		75.11	+1.9	159	0.962 (20 °C)	8.2×10^{-4}	604	2.73	30 (60 °C)	1.4463 (45 °C)
Diisopropanolamine [b]	[*110-97-4*]	133.19	+38.9	244	0.9842 (60 °C)	8.2×10^{-4}	375	2.49	58 (60 °C)	1.4542 (60 °C)
Triisopropanolamine [b]	[*122-20-3*]	191.27	ca. 55	291	0.9904 (60 °C)	6.9×10^{-4}	309	2.47	141 (60 °C)	1.4555 (60 °C)

[a] Commercial isomer mixture consisting of ca. 94 % 1-amino-2-propanol and 6 % 2-amino-1-propanol.
[b] Commercial isomer mixture.

4.1.2. Chemical Properties

The properties of amines and alcohols are combined in isopropanolamines, and these compounds exhibit typical reactions of both functional groups; the amino group can be primary, secondary, or tertiary. In chemical behavior, isopropanolamines differ only slightly from the ethanolamines described in Section 2.1.2.

Isopropanolamines react with acidic gases (H_2S, CO_2) in aqueous solution to form salts, which decompose to the starting materials when heated. This effect is utilized in the purification of natural gas and synthesis gas. The salts of strong acids, such as HCl or H_2SO_4, are crystalline compounds.

Diisopropanolamine sulfate eliminates water at elevated temperature to form 2,6-dimethylmorpholine sulfate [54].

The sodium salt of the sulfuric acid half-ester of monoisopropanolamine undergoes cyclization when heated, and propylenimine is formed [55]:

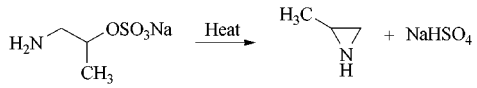

Fatty acids, such as stearic acid, react with isopropanolamine at room temperature to form neutral, waxlike isopropanolamine soaps. In reactions above 140 °C and with elimination of water, amides are formed preferentially, and minor amounts of fatty esters [56] are produced in a side reaction. Complete esterification can be effected by reaction with acyl chlorides in the presence of pyridine [57]. The reaction of esters with isopropanolamines above 100 °C gives amides [58]:

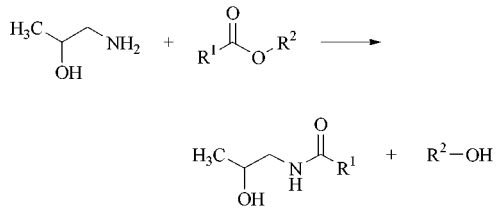

R^1, R^2 = alkyl

When thionyl chloride is used as a reactant and chloroform as the solvent, the hydroxyl group can be exchanged for chlorine. Isopropanolamine is converted to 2-chloropropylamine [59], which is obtained in good yield.

Isopropanolamine and diisopropanolamine react with epoxides to give mixed isopropanolamines. The amino group is several times more reactive than the hydroxyl group [60]. In the presence of suitable catalysts such as sodium hydroxide or sodium alkoxides, the epoxide reacts preferentially with the hydroxyl group to form polyethers [61]. Alkylating substances, e.g., alkyl halides or dimethyl sulfate, give *N*-alkylated derivatives [62].

Isopropanolamine can be converted to 1,2-diaminopropane by catalytic amination under pressure (→ Amines, Aliphatic, Chap. 8.1.2.).

Aldehydes and ketones react with the primary nitrogen atom of isopropanolamine to form a Schiff base.

$$CH_3CH(OH)CH_2NH_2 + RCHO \longrightarrow CH_3CH(OH)CH_2N{=}CHR + H_2O$$

Formaldehyde reacts with isopropanolamines to form the corresponding hydroxymethyl compounds, which can be converted to the methyl derivatives by catalytic hydrogenation.

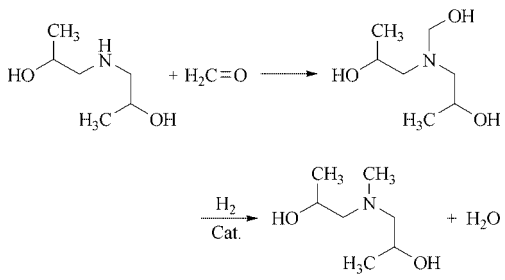

Compounds containing an active hydrogen atom undergo a Mannich reaction with formaldehyde and an isopropanolamine that has a primary or secondary amino group.

$$(CH_3)_2C{=}O + HCHO + HNR^1R^2 \longrightarrow CH_3C(=O)CH_2CH_2NR^1R^2 + H_2O$$

$R^1 = H, C_3H_6OH; R^2 = C_3H_6OH$

Diisopropanolamine reacts with nitrous acid or its salts to form *N*-nitrosamines. Isopropanolamines react with optically active acids, such as tartaric acid, to form diastereomeric pairs of salts; this enables the separation of isopropanolamines into their enantiomeric forms [63].

The use of triisopropanolamine as an esterification catalyst has been suggested [64].

4.2. Production

Isopropanolamines are produced like ethanolamines (see Section 3.2, Fig. 4), by treating propylene oxide with NH_3 in the liquid phase in a continuous or batchwise procedure [65], [66]. The product is always a mixture of 1-amino-2-propanols and 2-amino-1-propanols, with the 1-amino isomers being the major component.

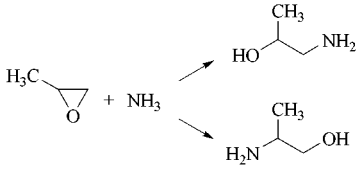

Industrially, the manner in which the propylene oxide ring is cleaved, and hence the isomer ratio of isopropanolamines, cannot yet be controlled.

Excess ammonia (about 2 – 20 mol per mole of propylene oxide) facilitates removal of the reaction heat (about 93 kJ/mol) and reduces the amount of polypropoxylated byproducts. This amount also depends on the relative reaction rates of propylene oxide with any amines present in the mixture.

Reaction temperatures can vary from 50 to 150 °C but are preferably around 100 °C. The resulting pressure can be reduced by adding water (ca. 10 – 60 %). Water and hydroxyl-containing compounds such as alcohols, phenols, and alkanolamines (autocatalysis) accelerate the reaction [67].

A *continuous procedure* is used for large-scale production, e.g., of monoisopropanolamine and diisopropanolamine [68], [69]. This procedure is similar to that used to produce ethanolamines (see Section 2.2).

The most important manufacturers of isopropanolamines are Air Products, BASF, Bayer, Sasol, and ICI in Western Europe, and Dow Chemical in the United States.

Table 7. Safety data for isopropranolamines [53]

	Monoisopropanolamine	Diisopropanolamine	Triisopropanolamine
Flash point, °C	71	135	160
Ignition temperature, °C	335	290	275
Explosion limits, vol %	1.9–10.4	1.6 – 8.0	0.8 – 5.8
Vapor pressure, hPa	1.9 (20 °C)	2.06 (100 °C)	1 (100 °C)

Table 8. Minimum quality requirements for commercial isopropanolamines (as of 2000)

Quality criterion	Method	Monoisopropanolamine	Diisopropanolamine	Diisopropanolamine containing 10 % H_2O	Triisopropanolamine	Triisopropanolamine containing 15 % H_2O
Water content, wt %	DIN 51 777	max. 0.2	max. 0.2	10 ± 1	max. 0.2	15 ± 1
Purity, % ∗	GC	min. 99.0	min. 99.0	88	min. 99.0	84
Low boilers (max.), % ∗ (excluding H_2O)	GC	0.3	0.3	0.3	0.3	0.3
High boilers (max.) % ∗	GC	0.5	0.5	0.5	0.5	0.5
Color (max.), APHA	DIN-ISO 6271	20	50	50	100	100
Acid consumption, mL of 1 N HCl per gram		13.15 – 13.45	7.35 – 7.65	6.60 – 6.90	5.10 – 5.40	4.30 – 4.60
Freezing point, °C	BS 523 : 1964	ca. 2	ca. 39	ca. −10	ca. 58	< -10

∗ Determined as area under GC peak.

4.3. Quality Specifications

Because isopropanolamines are manufactured, sold, and used worldwide, the commercial products differ only slightly in delivery specifications. Typical specifications are listed in Table 8.

Purity is usually determined by gas chromatography. Depending on analytical conditions, isomeric and enantiomeric forms of isopropanolamines exhibit different numbers of peaks, which can be identified by test analyses. Derivatives obtained by reaction with trifluoroacetic anhydride can be analyzed more easily and analysis time can be reduced.

4.4. Uses

Possible uses of isopropanolamines are similar to those of ethanolamines; as a rule, the isopropanolamine derivatives exhibit better solubility in hydrophobic media. For example, neutralization with fatty acids, such as stearic or lauric acid, gives the corresponding soaps which are used in lubricant additives, textile finishing, and coating paper and wood [70]. Salts of isopropanolamines with fatty acids are also used as ionic emulsifiers in emulsion paints and as dispersants for pigments. Salts of isopropanolamines with alkylarylsulfonic acids are employed as degreasers and cleaning promoters [71]. The fatty acid amides of isopropanolamines and mixtures of these with esters are frequently used in the cosmetics industry as shampoo thickeners and foam regulators [72].

Isopropanolamines are effective corrosion inhibitors in antifreezes [73] and cutting oils. Simultaneous use of nitrites should be avoided, since carcinogenic nitrosamines may form under certain circumstances. Isopropanolamines are important catalysts for urethane foams [74]. In epoxy resin systems, they have proved useful additives to curing agents [75].

Sulfur-containing isopropanolamine derivatives are used to extract metal ions, such as mercury, gold, and platinum, from salt solutions and wastewater [76].

A very important use of aqueous diisopropanolamine solutions is as a gas wash in the Adip and Sulfinol processes. Here, CO_2 and H_2S, in particular, are washed out of natural gas, synthesis gas, or coke oven gas and, after desorption, can be obtained in some cases in their pure form [77].

Monoisopropanolamine is employed extensively as an intermediate in the production of 1,2-diaminopropane (→ Amines, Aliphatic, Chap. 8.1.2.), while diisopropanolamine is used for the synthesis of 2,6-dimethylmorpholine [54].

5. *N*-Alkylated Propanolamines and 3-Alkoxypropylamines

N-Alkylated propanolamines can be divided into *N*-alkylated 1-amino-2-propanols, 2-amino-1-propanols, and 3-amino-1-propanols.

2-Amino-1-propanols and 1-amino-2-propanols are referred to industrially as isopropanolamines (Chap. 4). *N*-Alkylated isopropanolamines are prepared by reacting propylene oxide with the corresponding amines, which yields almost exclusively 1-amino-2-propanols [78]. Alkylated 2-amino-1-propanols have no industrial importance and are not discussed in this chapter.

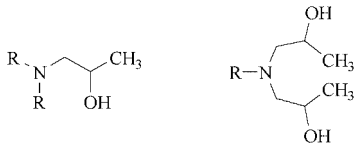

1-Amino-2-propanols

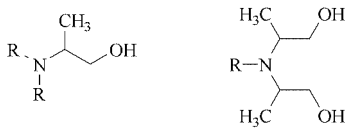

2-Amino-1-propanols

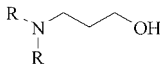

3-Amino-1-propanols

5.1. Properties

5.1.1. Physical Properties

With a few exceptions, *N*-alkylated isopropanolamines, 3-amino-1-propanols, and 3-alkoxypropylamines are colorless liquids at room temperature. Some of them, such as cyclohexylisopropanolamine, are crystalline solids. The compounds tend to discolor on prolonged storage, especially if exposed to light, air, and heat (> 40 °C). They have an amine-like odor, and absorb water and carbon dioxide. They are soluble in water, alcohol, acetone, glycol, glycerol, and glycol ethers, but only sparingly soluble in aliphatic or aromatic hydrocarbons and diethyl ether. Methyldiisopropanolamine does not form an azeotrope with water, but an azeotrope is formed upon distillation of aqueous dimethylisopropanolamine and 3-methoxypropylamine. Physical properties of *N*-alkylated isopropanolamines are given in Table 9 and those of 3-amino-1-propanols and 3-alkoxypropylamines in Table 10.

5.1.2. Chemical Properties

Because *N*-alkylated propanolamines contain hydroxyl groups and basic nitrogen, they possess the properties of both amines and alcohols. They form salts with organic and inorganic acids [79]. The sulfuric acid esters of propanolamines can be cyclized with sodium hydroxide to give propylenimines (see Section 4.1.2) [80]. Propanolamines with secondary amino groups form alkanolamides with fatty acids. Reaction of the amides with ethylene oxide gives water-soluble or easily dispersible nonionic surfactants [81]. 1-Amino-2-propanols are chiral and can be separated into enantiomers by means of appropriate reagents [63], [82]. *N*-Alkylpropanolamines with tertiary amino groups can be converted by means of benzoyl chloride, without the assistance of a base, into the amino alcohol ester hydrochlorides; with thionyl chloride, they give 2-chloropropylammonium chlorides.

The condensation of 3-amino-1-propanol with carbonyl compounds yields tetrahydro-1,3-oxazines [83], Schiff bases, or mixtures of both compounds. 3-Amino-1-propanol and 2-hydroxy-3,3-dimethylbutyrolactone react almost quantitatively to give panthenol [84].

Alkoxypropylamines react analogously to primary aliphatic amines. They form salts with inorganic or organic acids, and can be converted to acid amides, sulfonamides, urethanes, and Schiff bases. They can be alkylated with formaldehyde or formic acid to give dimethylamino compounds. 3-Methoxypropylamine catalyzes the addition of hydrogen cyanide to vinyl esters to give 1-cyanoethyl esters [85]. 3-Methoxypropylsulfanilylurea obtained from the 3-methoxypropylammonium salt of *N,N*-di(acetylsulfanilyl)urea is an oral antidiabetic agent [86].

Table 9. Physical properties of *N*-alkylated isopropanolamines

Compound	CAS registry number	M_r	*mp*, °C	*bp* (101.3 kPa), °C	ρ, (20 °C) g/cm^3	Cubic expansion coefficient, K^{-1}	Heat of vaporization kJ/kg	Specific heat, kJ kg^{-1} K^{-1}	Specific electrical conductivity, Ω^{-1} cm^{-1}	Viscosity (20 °C), mPa · s	n_D^{20}	Surface tension (20 °C), mN/m	Flash point, °C	Ignition temperature, °C	Temperature class
Monomethylisopropanolamine	[*16667-45-1*]	89.14	12.5	149	0.9062		455.6		4.8×10^{-7}		1.4340	28.7 (22 °C)	58.7	305	T2
Methyldiisopropanolamine	[*4402-30-6*]	147.22	−32	226	0.945		329.3		1.4×10^{-7}	47.3	1.4472 (25 °C)	31.0 (22 °C)	110.5	300	T3
Dimethylisopropanolamine	[*108-16-7*]	103.17	−85*	125.8	0.849 −0.852	7.0×10^{-4}	393.6	2.45 (25 °C)	2×10^{-6}	1.5	1.4189	24.5 (25 °C)	26	225	T3
Diethylisopropanolamine		131.22		157.5 −159	0.8570						1.4265				
Dibutylisopropanolamine		187.33	−80*	229.1	0.8419	7.8×10^{-4}	246.6			3.5	1.4361		205		
Cyclohexylisopropanolamine		157.26	45.1	238	0.9365 (40 °C)		329.3	2.47 (60 °C)	9.0×10^{-4}	11.5 (60 °C)	1.4732		111	270	T3
Cyclohexyldiisopropanolamine		215.34	2.0												
Cyclooctylisopropanolamine		185.31	−12.85*	275	0.953	7.6×10^{-4}	234.5	2.24 (20 °C)	5.1×10^{-4}	665	1.4902	37.0 (20.3 °C)	141	245	T3
Cyclooctyldiisopropanolamine		243.39	29.6	318	0.977	7.4×10^{-4}	233.95	1.98 (20 °C)	5.7×10^{-11}	10870	1.4890	38.0 (20.3 °C)	176	245	T3
Aminoethylisopropanolamine	[*123-84-2*]	119.19	−38*		0.9868					145	1.4750		123	345	T2
1-(2-Hydroxypropyl) piperazine		144.22	−17*		1.0108					423	1.4910		115	290	T3
1,4-Bis(2-hydroxypropyl)piperazine		202.30	97–98										157	250	T3
4-(2-Hydroxypropyl) morpholine (12)		145.21	−47*	218	1.0163		314.3	1.92		11.0	1.4642	34.5	92	215	T3
N,N,N′N′-Tetrakis(2-hydroxypropyl) ethylenediamine		292.43	+11*	190 (133 Pa)	1.03 (25 °C)	6.7×10^{-4}		2.297 (50 °C)	2.4×10^{-4}	45 (100 °C)	1.478 (25 °C)	35.1	201	285	T3

* Pour point.

5.2. Production

Industrially, *N*-alkylated isopropanolamines are prepared primarily from amines and propylene oxide according to the following reactions [67], [87]:

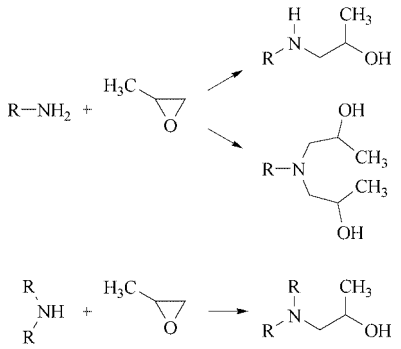

N-Alkylated 1-amino-2-propanols are formed predominantly [66], [88], [89]. The isomeric 2-amino-1-propanols are obtained only as byproducts (0.2 – 5 %, depending on the nature of the amino component). The reactions are carried out at 80 – 160 °C and pressures up to 5 MPa, batchwise or continuously. A detailed description can be found in Section 4.2. As in ethanolamine production, the oxide must be added to the amine, not vice versa (see Section 2.2). The crude reaction products are purified by continuous or batchwise distillation.

3-Amino-1-propanol is obtained by hydrogenating ethylene cyanohydrin. *N*-Methylated 3-amino-1-propanols are prepared from 3-amino-1-propanol and formaldehyde, with subsequent hydrogenation of the adducts [90]. 3-Alkoxypropylamines are prepared from acrylonitrile by addition to alcohols and subsequent hydrogenation [91].

5.3. Quality Specifications

Table 11 gives the specifications for commercial *N*-alkylated propanolamines and 3-alkoxypropylamines. Purity is generally determined by gas chromatography.

Table 10. Physical properties of 3-amino-1-propanols and 3-alkoxypropylamines

Compound	CAS registry number	M_r	*mp*, °C	*bp* (101.3 kPa), °C	ρ, (20 °C) g/cm^3	Cubic expansion coefficient, K^{-1}	Heat of vaporization, kJ/kg	Specific heat, kJ kg^{-1} K^{-1}	Viscosity, mPa · s	n_D^{20}	Flash point °C	Ignition temperature, °C	Temperature class
3-Amino-1-propanol	[*156-87-6*]	75.11	11.5	189	0.98 – 0.99	8.3×10^{-4}	641	2.71	36 (20 °C)	1.460 – 1.462	98 – 100	385	T2
3-Dimethyl-amino-1-propanol	[*3179-63-3*]	103.17	−32.7	162 – 163	0.882 – 0.885						57		
3-Methoxy-propylamine	[*5332-73-0*]	89.14	< −70	118	0.8729	1.2×10^{-3}	422.9		0.9 (25 °C)	1.4159 (25 °C)	27	270	T3
3-Ethoxy-propylamine	[*6291-85-6*]	103.17	< −70	134.7	0.861						33.7	245	T3
3-(2-Ethylhexyloxy)-propylamine	[*5397-31-9*]	187.33	< −70	237	0.8466		258.85	2.27			104	220	T3

5.4. Uses and Economic Aspects

N-Alkylated propanolamines and 3-alkoxypropylamines are used mainly as intermediates, especially in the preparation of crop-protection agents and pharmaceuticals. 3-Alkoxypropylamines have considerable commercial importance in the preparation of dyes. *N*-Alkylated propanolamines and their derivatives have been employed, in the form of soaps, as lubricating oil additives, mold release agents [92], [93], textile finishes, and coatings for paper and wood [94]. The propanolammonium salts of alkylarylsulfonic acids are used as cleansing boosters and degreasing agents [95]. 3-Alkoxypropylamines are utilized as catalysts in the production of polyurethane foams and epoxy resins.

3-Alkoxypropylamines and *N*-alkylpropanolamines are used as corrosion inhibitors and in the preparation of aqueous polymer solutions [96–100]. Anion exchangers can be prepared by reacting cellulose with dimethylisopropanolamine and epichlorohydrin [101].

World demand for *N*-alkylpropanolamines and 3-alkoxypropylamines is estimated at 5000 t/a. More than 100 t/a of 3-methoxypropylamine and dimethylisopropanolamine are used.

6. Storage and Transportation

Alkanolamines should be stored in stainless steel containers with exclusion of air (O_2, CO_2) and moisture, preferably under dry nitrogen. Storage temperature should not exceed 40 °C (50 °C for ethanolamines). Steel tanks may be used if absorption of iron (up to 10 ppm) is not important. Alkanolamines turn yellow on prolonged storage, especially in the presence of oxygen.

Depending on quality requirements and sensitivity of the products, steel, stainless steel, or polyethylene containers can be used for transportation. The containers must possess airtight closures to prevent absorption of water and carbon dioxide. Zinc and other nonferrous metals are attacked by alkanolamines. For transportation, particular national regulations, such as GGVS in Germany, should be observed [102].

Rubber gloves and safety goggles should be worn when handling alkanolamines. In some cases, especially with 3-alkoxypropylamines, a full face shield and rubber clothing are advisable.

For safety data of alkanolamines, see the sections on physical properties.

7. Environmental Protection

Ammonia- or amine-containing off-gases from alkanolamine production are either burned or purified by acid scrubbing. Very low amine concentrations (10^{-3} ppm) can be determined by modern analytical methods [103].

Wastewater from plant cleaning and acid scrubbing is treated in a sewage plant. When properly fed into a biological treatment plant, alkanolamines are readily degraded by appropriate bacteria.

Spilled product should be removed with an absorptive material such as urea resin foam (Hygromull) or peat dust, which is then incinerated.

8. Toxicology and Occupational Health

8.1. General Aspects

All alkanolamines considered here are compiled in Table 12 along with available data on acute toxicity, irritant/caustic properties, and occupational exposure limits.

Most prominent among the toxic effects of alkanolamines is their caustic/irritant action on exposed skin and mucous membranes. This effect decreases from primary to tertiary alkanolamines and is distinctly less pronounced or absent in their salts due to compensation of alkalinity. An increase in the number of hydroxyl groups further alleviates the irritant properties.

The low vapor pressure of these compounds at room temperature explains the generally low acute inhalation hazard. However, vapors formed at elevated temperatures may irritate eyes and respiratory tract. Aqueous solutions containing 1 – 10 % of monoalkanolamines such as monoethanol- and monoisopropanolamine are still irritating to eyes and skin. In principal, this also applies to the *N*-dialkanol derivatives,

Table 11. Specifications for commercial *N*-alkylated propanolamines and 3-alkoxypropylamines

Compound	Purity, %	ϱ (20 °C, DIN 51 757), g/cm^3	Boiling range 5 – 95 mL (DIN 51 751, DIN 53 406), °C	Color (DIN-ISO 6271), APHA	Water content (DIN 51 777), %
Monomethylisopropanolamine	98	0.906	146 – 152	≤ 50	
Methyldiisopropanolamine	98	0.954	226 – 230	≤ 50	
Dimethylisopropanolamine	> 99.5	0.849 – 0.852	125 – 127	≤ 50	< 0.2
Aminoethylisopropanolamine	≥ 98	0.9868	232.5 – 235.5	≤ 20	
N,N,N,N′-Tetrakis(2-hydroxypropyl)ethylenediamine	≥ 99	1.009 – 1.013		20	
3-Dimethylamino-1-propanol	99	0.882 – 0.885	162 – 163		
3-Methoxypropylamine	99	0.871 – 0.873	117 – 119	20	< 0.2
3-Ethoxypropylamine	99	0.855 – 0.861	130 – 135	5	< 0.2
3-(2-Ethylhexyloxy)propylamine	> 99	0.847 – 0.848	233 – 238	≤ 40	< 0.2

albeit to a lesser degree. Available data suggest that alkanolamines can penetrate the skin.

In general, the acute toxicity of the tested compounds is moderate to low and nonspecific; it is probably dominated by local irritant effects owing to their alkalinity. In a few cases, acute or long-term systemic symptoms may involve neuropharmacological effects that interact with acetylcholine synthesis and neurotransmitter function or neurotoxicity as a consequence of disturbed neural membrane structure.

The mutagenic potential of alkanolamines is very low. However, in the presence of nitrite or under other appropriate conditions, primary and secondary alkanolamines form nitroso compounds which may be mutagenic or carcinogenic. Existing regulatory limitations on the use of compounds need to be considered as well as work-place regulations.

The skin-sensitizing potential of alkanolamines is very low.

8.2. Ethanolamines

The chemistry, biochemistry, and toxicity of monoethanolamine (MEA), diethanolamine (DEA), and triethanolamine (TEA) have comprehensively been reviewed [142]. MEA is a natural building block of a group of important membrane components, the phospholipids, and the precursor in the de-novo synthesis of choline to form lecithin. DEA interferes with the MEA and phospholipid metabolism by replacing choline. The accumulation of atypical phospholipids appears to be the determinant in the systemic toxicity of DEA. TEA does not exhibit these properties.

There is no evidence that MEA and DEA have a sensitizing potential [142], while for TEA, occasional cases of allergic contact dermatitis were reported [108], [109]; however, no sensitizing effect was found in guinea pigs [110], [164], [142].

The majority of genotoxicity tests, including bacterial and mammalian in vitro test systems, indicate the absence of mutagenic activity in these ethanolamines [142].

Monoethanolamine. Early inhalation studies with various animal species demonstrated that the extreme irritancy of an MEA atmosphere to exposed skin and the respiratory tract resulted in pronounced discomfort and behavioral depression already at concentrations of 5 – 15 ppm (ca. 12.5 – 38 mg/m^3) [142]. Nonstop inhalation for about 3 – 6 weeks of 66 – 75 ppm MEA vapors was lethal to guinea pigs and rats; 100 ppm was lethal to dogs. The liver and the kidney were identified as target organs at high doses [142]. Oral administration of 1.28 g kg^{-1} d^{-1} in the feed to rats for 30 – 90 d was lethal, whereas a dose of 0.64 g kg^{-1} d^{-1} led to changes in the liver and kidney weights; a dose of 0.32 g kg^{-1} d^{-1} was tolerated without any toxic effects [105], [142].

MEA produced no signs of developmental toxicity in standard developmental toxicity tests when applied orally to pregnant rats [142], [143] and dermally to pregnant rabbits [144]. This refutes the putative increases in the incidence of fetal malformations, unrelated to dose, in a previous nonstandard study [142].

Diethanolamine. All toxicity data for repeated exposure suggest that DEA can be in-

Table 12. Acute toxicological data and occupational exposure limits of alkanolamines [a]

Substance [CAS no.]	Acute toxicity: LD_{50}, g/kg	Ref.	AIH (rat) [b]	Ref.	Local irritation: Skin [c]	Eye	Ref.	OEL [d], mL/m^3 (mg/m^3)
Ethanolamines								
Monoethanolamine	1.5 (o, rt)	[116],				c	[116]	MAK: 2 (5.1)
[*141-43-5*]	1.2 – 2.5 (o, rt)	[142],	> 8 h (520 ppm)	[116], [142]	1 – 5 min: c			TLV: 3 (7.5)
	1.0 – 2.5 (d, rbt)	[104]						
Diethylanolamine	ca. 1.6 (o,rt)	[116]	> 8 h (0.37 ppm)	[116]	se i; ≤15 min: n i	se i	[116]	OEL(NL): (15I) sk
[*111-42-2*]	1.7 – 2.8 (o, rt)	[142]		[142]				TLV: 0.46 (2) sk
	8.1 – 12.2 (d, rbt)	[142], [104]	1471 ppm (vapor/aerosol, 2 h)	[106]	m/sl i	i	[142]	
Triethanolamine	ca. 7.2 (o, rt)	[116]			4 h: n i	n i/sl i	[116]	OEL (S): (5)
[*102-71-6*]	8.4 – 11.3 (o, rt)	[142]	> 8 h (0.0047 ppm)	[116], [142]	20 h: i			TLV: (5)
	> 20 (d, rbt)	[142]						
2-(2-Aminoethoxy)ethanol	ca. 3.4 (o, rt)	[116],	> 8 h	[116]	c	se i	[116]	n.e.
	ca. 2.5 (o, rt)	[159],			≤ 3 min: n i	se i	[159]	
[*929-06-6*]	ca. 1.3 (d, rbt)	[181],			≤ 1 h: se i			
	> 3 (d, rbt)	[159]			4 h: c			
Alkylalkanolamines								
Monomethylethanolamine	1.39 – 2.34 (o, rt)	[114], [116], [132]	> 8 h	[114], [116]	3 min: i 4 h: c		[132], [133], [116]	n.e., sk
	> 2 (d, rbt)	[116]				se i	[114], [116]	
[*109-83-1*]	ca. 1.9 (d, rbt)	[132]						
	ca. 1.25 (d, rbt)	[133]						
Ethylethanolamine	1.1 (o, rt)	[116]	> 7 h	[116]	c	se i	[116]	n.e.
[*110-73-6*]	1.48 (o, rt)	[114]	> 8 h	[114]				
	3.67 (d, rt)	[116]						
n-Propylethanolamine	ca. 1 (o, rt)	[116]	> 8 h		3 min: sl i	c		n.e.
[*16369-21-4*]	> 0.4 (d, rbt.)	[116]	> 2.9 mg $L^{-1}(4\,h)^{-1}$	[116]	1 h: c		[116]	
n-Butylethanolamine [*111-75-1*]	ca. 1 (o, rt)	[116]	> 8 h	[116]	5 min: sei/c	c	[116]	n.e.
Hydroxyethylaniline	2.73 (o, rt)	[116]	> 8 h	[116]	n i	se i	[116]	n.e.
[*122-98-5*]	2.23 (o, rt)	[140]			i	i	[140]	
	0.06 (d, rbt)	[140]						
Dimethylethanolamine	≥ 2 (o, rt)	[116], [134]	LC_{50} ca. 6 mg L^{-1} $(4\,h)^{-1}$	[132], [136]	se i/c	se i/c	[116], [135], [132]	n.e.
	1.24 – 1.6 (o, rt)	[135]						
[*108-01-0*]	1.2 (d, rbt)	[132]	30 min	[116]				

Table 12. (Continued)

Substance [CAS no.]	Acute toxicity				Local irritation			OEL [d], mL/m^3 (mg/m^3)
	LD_{50}, g/kg	Ref.	AIH (rat) [b]	Ref.	Skin [c]	Eye	Ref.	
	1.7 – 2.4 (d, rbt)	[135]						
Diethylethanolamine	1.33 (o, rt)	[116], [113]	> 1 h	[116]	1 h: c	c	[116]	MAK: 10 (50) sk
[*100-37-8*]	ca. 0.89 (d, gp)	[113]						TLV: 2 (9.6)
Dibutylethanolamine	ca. 0.6 (o, rt)	[116]	> 7 h	[116]	> 5 min: i	se i	[116]	OEL (NL): (14) sk
[*102-81-8*]	1.68 (d, rbt)							TLV: 0.5 (3.5) sk
N-Hydroxyethylpiperazine	ca. 4.2 (o, rt)	[116]	> 8 h	[116]	se i	se i	[116]	n.e.
[*103-76-4*]	> 5 (d, rbt)	[166]						
N-Hydroxyethylmorpholine	ca. 6.8 (o, rt)	[116]	> 8 h	[116]	n i	sl i	[116]	n.e.
[*622-40-2*]	> 16 (d, rbt)	[141]			n i	n i	[141]	
N-Hydroxyethylpiperidine	ca. 1.1 (o, rt)	[116]	> 8 h	[116]	c (> 3 min)	c	[116]	n.e.
[*3040-44-6*]	1 – 2 (d, rbt)	[116]						
Methyldiethanolamine	ca. 4.7 (o, rt)	[116], [114]	> 8 h	[116], [114]	n i	i	[116], [114]	n.e.
[*105-59-9*]	1.95 (o, rt)	[132]						
	> 5 (d, rbt)	[132], [114]				mi	[132]	
n-Butyldiethanolamine	0.6 – 1.0 g/kg (o, rt)	[116]	> 8 h	[116]	15 min: sl i	c		n.e.
[*102-79-4*]					20 h: c		[116]	
Propanolamines and Isopropanolamines								
Monoisopropanolamine	4.26 (o, rt)	[104]	> 8 h	[116]	15 min: c	c	[116]	n.e.
[*78-96-6*]	2.7 (o, rt)	[116]			4 h (1-% sol.):		[138]	
[*6168-72-5*][e]	1.64 (d, rbt)	[105]			n i			
	2.7 (o, rt)	[116]			m i	se i	[104], [105]	
Diisopropanolamine	6.72 (o, rt)	[104]	> 8 h	[116]	se i	se i	[104]	n.e.
[*110-97-4*]	6 (o, rt)	[116]			15 min: n i	m i	[116]	
					20 h: m i		[116]	
Triisopropanolamine	6.5 (o, rt)	[104]	> 8 h	[116]		se i	[104], [105]	n.e.
[*122-20-3*]	> 10 (d, rbt)	[105]	2 h (vapors, 170 °C)	[137]				
	4 (o, rt)	[116]			15 min: n i	se i		
					20 h: sl i	m i	[116]	
3-Amino-1-propanol	2.83 (o, rt)	[104], [105]	> 8 h		se i		[104], [105]	n.e.
[*156-87-6*]	1.3 (o, rt)				5 min: se i	c		
	> 2 (d, rt)	[116]	LC_{50} > 16.4 mg L^{-1} h^{-1}	[116]	15 min: c		[116]	

Table 12. (Continued)

Substance [CAS no.]	Acute toxicity				Local irritation			OEL [d], mL/m^3 (mg/m^3)
	LD_{50}, g/kg	Ref.	AIH (rat) [b]	Ref.	Skin [c]	Eye	Ref.	
Alkylated Propanolamines/Isopropanolamines								
3-Dimethylamino-1-propanol [*3179-63-3*]	1.86 (o, rt)	[116]	8 h	[116]	15 min: se i/c	se i/c	[116]	n.e.
Monomethylisopropanolamine [*16667-45-1*]	1 (o, rt)	[116]	> 7 h	[116]	15 min: c	c	[116]	n.e.
Dimethylisopropanolamine [*108-16-7*]	1.36 (o, rt) 1.89 (o, rt)	[116] [104]	10 min	[116]	5 min: c	c	[116]	n.e.
Methyldiisopropanolamine [*4402-30-6*]	2.15 – 3.83 (o, rt)	[116]						n.e.
3-Alkoxypropylamines								
3-Methoxypropylamine [*5332-73-0*]	6.26 (o, rt, aq, sol., pH 7) 0.63 (o, rt)	[116] [139]	1 h	[116]	5 min: c	c	[116]	n.e.
3-Ethoxypropylamine [*6291-85-6*]	1.1 – 1.5 (o, rt)	[116]	1 h	[116]	1 min: c	c	[116]	n.e.
3-(2-Ethylhexyloxy)propylamine [*5397-31-9*]	0.85 (o, rt) 0.32 (o, rt) 0.36 (d, rbt)	[116] [104]	8 h	[116]	5 min: c	c	[116]	n.e.

[a] Symbols: o = oral; d = dermal; rt = rat; rbt = rabbit; gp = guinea pig; i = irritating (n = not, sl = slightly, m = moderately, se = severely); c = corrosive.
[b] AIH = acute inhalation hazard: rats inhale an atmosphere enriched with vapor at 20 °C (when not otherwise indicated, vapor-saturated air); time = duration of exposure required to cause the first death.
[c] The skin irritation for the specific time of exposure to the substance is given.
[d] OEL = occupational exposure limit: MAK (Germany, 1999), TLV (USA, 1997); NL = The Netherlands, S = Sweden; I = inhalable (total aerosol), sk = skin; n.e. = not established.
[e] Monoisopropanolamine (*d,l*) [*78-96-6*] contains 2-amino-1-propanol (*d,l*) [*6168-72-3*].

corporated in cell membranes and affects their stability [142]. The liver and kidneys are the major target organs, but demyelinization of nerve fibers and anemia, presumably due to altered composition of the erythrocyte membranes, is also possible. Toxic effects were observed after continuous inhalation of 25 ppm for 9 d or 6 ppm for 13 weeks (5 d/week), which was lethal to some of the exposed rats, and after a seven-week administration to rats of neutralized diethanolamine at a concentration of 4 mg/mL in drinking water [107], [142]. Oral administration of 0.02 g kg^{-1} d^{-1} to rats for 90 d was tolerated without any signs of toxicity, whereas a dose of 0.09 g kg^{-1} caused an increase in liver and kidney weights. A dose of 170 mg kg^{-1} $d^{-1}$was lethal [105], [142]. A subchronic no observed adverse effect level was established at a dose of 20 – 25 mg kg^{-1} d^{-1} for oral exposure in rats [145].

In two-year dermal carcinogenicity studies, increased incidence of liver tumors was observed in mice, but not in rats [146]. Follow-up studies in two transgenic mouse lines, which were expected to exhibit enhanced sensitivity to carcinogenic substances, clearly failed to elicit any carcinogenic potential of DEA [147]. Furthermore, the development of DEA-related neoplasia in mouse liver could reasonably be explained by a nongenotoxic mechanism expressed by disruption of choline metabolism [148].

DEA was maternally toxic and fetotoxic (i.e., retarded ossification) at oral doses of 0.2 g kg^{-1} d^{-1} [142], [149], and at dermal doses of 1.5 and 0.35 g kg^{-1} d^{-1} in rats and rabbits, respectively, but failed to produce structural malformations in any test [142].

Triethanolamine has a low systemic toxicity, mainly restricted to the kidney as target organ [112], [142]. In contrast to DEA it is readily eliminated via the urine without undergoing appreciable metabolism. Extensive mutagenicity testing did not reveal a genotoxic effect [142], [154]. In subacute inhalation studies with rats and mice exposed to TEA aerosols over a 16-d period (0.125 – 2 g/m^3, 6 h/d, 5 d per week), only minimal to slight acute inflammation of the submucosa of the larynx but no effects on nasal or lung tissue were noted; this indicates a lack of severe irritation from TEA exposure [142], [153]. Oral administration of 0.08 g kg^{-1} d^{-1} to rats in the feed for 90 d was tolerated without toxic effects. A dose of 0.17 g kg^{-1} d^{-1}caused changes in the liver and kidney weights, while 0.73 g kg^{-1} d^{-1} led to histopathological changes in both organs and deaths [181], [105]. High doses of ca. 1.4 g kg^{-1} d^{-1}administered in the drinking water for 14 d produced no significant signs of intoxication in rats and mice [142]. Two-year exposure to doses above 0.5 g kg^{-1} d^{-1} in the drinking water led to an acceleration of the chronic, age-related nephropathy usually observed in this rat strain [112], [142], while there were no morphological effects in mice under comparable conditions [150], [142]. No conclusions about apparently positive oncogenic findings can be drawn from another nonverifiable lifelong feeding study in mice [111], [142].

A dermal carcinogenicity study in mice suggested a significant increase in the spontaneous tumor rate of hepatocellular adenomas in the high-dose male and female groups (2 and 1 g kg^{-1} d^{-1}, respectively). However, because of an infection of the animals with *Helicobacter hepaticus*, the study could not be evaluated and was abandoned as being inadequate [165]. There was no significant carcinogenicity in rats epicutaneously administered TEA at doses from 30 to 250 mg kg^{-1} d^{-1} [165]. However, because of a marginally positive trend in the incidence of renal tubular adenomas in males, the Peer Review concluded on "equivocal evidence" in male rats [165].

2-(2-Aminoethoxy)ethanol (diglycolamine) applied as 10 % ethanolic solution showed no skin-sensitizing activity in the Buehler test [159], [167]. The Ames test [143], [159] and a DNA repair assay in rat hepatocytes failed to indicate a genotoxic potential [159]. A cell transformation study using mouse BALB/3T3 cells was also negative [159].

8.3. *N*-Alkyl- und *N*-Arylethanolamines

Methylethanolamine was not mutagenic in the Ames test [169]. Continuous inhalation exposure to 1.8 and 9.7 ppm (ca. 0.006 and 0.03 g m^{-3}, respectively) for 90 d resulted in local irritation (inflammatory effects on the nasal

epithelium), but no treatment-related systemically toxic findings in rats and guinea pigs. In male dogs, dose-related atrophic degenerations in the testes were noted under comparable test conditions [151]. After 32-d administration of 0.004, 0.04, and 0.4 g kg^{-1} d^{-1} in the diet to rats, relative liver weights were increased in all dose groups with slight hepatic vacuolization at the two upper levels, and slight testicular edema and atrophy in the high-dose male group [151]. There was no evidence of embryotoxic and teratogenic effects on fetuses after inhalation exposure to 150 ppm (ca 0.47 g/m^3) in pregnant rats [152].

Ethylethanolamine is reported to be negative in the Ames test [155]. After ingestion in drinking water for 30 d, mainly dose-related histopathological changes in the kidneys developed in rats, with diffuse edema in the tubular epithelia at about 0.023 g kg^{-1} d^{-1} and progressive degenerative lesions at higher doses; furthermore, counts of white and red blood cells were enhanced [156]. After 90-d administration, the above-mentioned effects appeared to be less pronounced at corresponding dosage [156].

In a specially designed test, ethylethanolamine applied at an oral dose of about 0.18 g kg^{-1} d^{-1} during the implantation period (day 1 through 6 of gestation), produced implantation losses of some 50 % in pregnant rats [157].

***n*-Propylethanolamine** was nonmutagenic in the *Salmonella typh./E. coli* reverse mutation test [158].

***n*-Butylethanolamine** was nonmutagenic in the *Salmonella typh./E. coli* reverse mutation test [158].

Hydroxyethylaniline was nonsensitizing in the guinea pig maximization test [143] and was mutagenic in a bacterial assay using only *Salmonella typh.* TA 98 [143]. Methemoglobinemia developed in cats after oral and dermal exposure (54.5 and 545 mg kg^{-1}, respectively), a known effect of aromatic amines [116].

Dimethylethanolamine occurs as intermediate in endogenous choline synthesis and therefore interferes with phospholipid and acetylcholine metabolism. It is said to have some stimulatory effect on the CNS by influencing neurotransmitter activity [163]. No skin sensitization was induced in guinea pigs [160]. A broad spectrum of various genotoxicity assays, including bacteria [135], [169], mammalian cells [135], [161] and an in-vivo micronucleus test in mice [135], [161] gave no evidence of mutagenic potential.

Inhalation toxicity is associated with strong local irritation in the nasal region and the eyes; no pathological changes of inner organs occurred in rats exposed to concentrations of 98 ppm (ca. 0.36 g/m^3) for two weeks. Thymus atrophy was seen at 288 ppm (ca. 1 g/m^3), but no histological lesions of the central or peripheral nervous system [136]. At 24 ppm (ca. 0.08 g/m^3) over 13 weeks, no systemic toxicity was noted, and transient corneal opacity was observed as a local adverse effect [136].

Following the administration of 0.2 – 0.3 g kg^{-1} d^{-1} in drinking water for more than 100 weeks, no increase in tumor incidence was observed in mice in a special test model using two mice strains infected with mammary-tumor provirus [162].

There was no evidence of developmental effects on fetuses after inhalation exposure to maternally toxic concentrations (up to 100 ppm or 0.48 g/m^3) during pregnancy of rats [135], [161].

Diethylethanolamine is also a metabolite of the local anesthetic procaine and has some pharmacological activity (vasodilatation, depression of nerve conductivity). No skin sensitization was induced in guinea pigs [168], and no mutagenicity in Ames tests [143], [169]. Inhalation toxicity is associated with strong local irritation of the nasal region and extreme discomfort of the exposed animals; no pathological changes of inner organs occurred in rats exposed to concentrations of up to 76 ppm (0.37 g/m^3) for 14 weeks [170], [171]. Likewise, there were no significant treatment-related findings in rats that received ca. 0.013 and 0.033 g kg^{-1} d^{-1} and successively increased doses of up to 0.67 g kg^{-1} d^{-1} in the diet for two years [172]. Marked neurological action was observed in dogs at repeated dietary doses of ca. 0.04 g kg^{-1} d^{-1} and above, expressed as tremor and head shaking followed by ataxia and convulsions at elevated doses [172]. There was no evidence of developmental effects

on fetuses after inhalative exposure to maternally toxic concentrations (up to 100 ppm or 0.48 g/m^3) during pregnancy of rats [173].

Dibutylethanolamine. Five-week administration of neutralized dibutylethanolamine in drinking water in doses of 0.43, 0.20, and 0.13 g kg^{-1} d^{-1}(male rats) and 0.33, 0.24, and 0.14 g kg^{-1} d^{-1}(female rats) produced no substance-related, histopathological effects; the lowest dose was devoid of any macroscopic findings [115]. Irritation of the eyes and nose was observed after inhalative exposure of rats (70 ppm or ca. 0.5 g kg^{-1} d^{-1}, 5 d per week, 6 h/d). This exposure was lethal to one of five animals and led to an increase in the relative liver and kidney weights and to liver damage. A concentration of 22 ppm (ca. 0.16 g kg^{-1} d^{-1}) for 6 h/d over a period of six months was tolerated by rats without any signs of toxicity [115].

Hydroxyethylpiperazine and Hydroxyethylmorpholine. Hydroxyethylpiperazine [175] and hydroxyethylmorpholine [176] were nonmutagenic in an Ames test. But the latter substance is reported to exhibit some genotoxic activity by inducing DNA damage and repair in rat liver cells and cell transformation in cultured mouse embryo fibroblasts [176].

***N*-Hydroxyethylpiperidine** (2-piperidinoethanol) was not mutagenic in an Ames test with *Salmonella typh.* TA100, TA98, and *E. coli* WP2 [177] and in a cytogenetic assay using Chinese hamster cells [178].

Methyldiethanolamine is reported to be nonsensitizing in a guinea pig maximization test [174] and proved to be nonmutagenic in various genotoxicity assays, including bacterial [161], [169] and mammalian cells and an in vivo micronucleus test [161]. After prolonged dermal administration to rats, no systemic effects were found up to the top dose of 0.75 g kg^{-1} d^{-1}; but local changes at the treated site, characterized as acanthosis and hyperkeratosis, were induced at doses of 0.25 g kg^{-1} d^{-1}and higher. No pathological changes were noted at 0.1 g kg^{-1} d^{-1} [174].

***n*-Butyldiethanolamine** was nonmutagenic in the *Salmonella typh./E. coli* reverse mutation test [158].

8.4. Propanolamines

Monoisopropanolamine induced no skin-allergic and photoallergic reactions in 150 humans tested in a repeated insult patch test [179]. Negative mutagenic results were obtained in Ames tests using *Salmonella* and *E. coli* tester strains [180], [169]. The ninefold inhalation of up to 0.24 g/m^3 (5 d/week, 6 h/d) failed to produce clinical signs of intoxication or macroscopic changes in rats and mice [179]. Likewise, after 90-d feeding of 0.14 to 2.2 g kg^{-1} d^{-1} to rats, no effects were reported, but weight changes of the liver and kidney occurred at high dosage [181]. No increase in tumor incidences occurred in rats that received 1 % in the feed (ca. 0.67 g kg^{-1} d^{-1}) for 94 weeks. However, in combination with 0.3 % sodium nitrite in drinking water, endogenously produced nitroso compounds induced tumors [182].

Diisopropanolamine applied as 2 % aqueous solution induced no skin-allergic and photoallergic reactions in a repeated insult patch test with 25 volunteers [179]. One case report of contact allergy has been published [183]. Dermatological testing programs using 1 % diisopropanolamine in facial sunscreen resulted in two positively responding subjects out of two random collectives of 424 volunteers in total [179]. Based on a nonverifiable test report, a weakly allergic potential was concluded in guinea pigs [184]. The substance was nonmutagenic in the Ames test and induced no chromosomal aberrations in rat lymphocytes [185], [186]. When it was administered to rats at doses of 0.1 – 3 g kg^{-1} d^{-1}in the drinking water for 2 weeks, no symptoms were observed up to 0.6 g kg^{-1} d^{-1}. The dose of 3 g kg^{-1} d^{-1} was lethal to two out of five male animals; the prominent effects for this dose group included a significant loss in body and organ weight, acute inflammation and degeneration of kidneys and urinary bladder, and general liver atrophy [138]. After four weeks of dermal exposure to up to 0.75 g kg^{-1} d^{-1}, no pathological changes were noted in rats, except for skin irritation at the treated site at 0.5 g kg^{-1} d^{-1}and higher [186]. No clinical and organotoxic findings, as well

as no increase in tumor incidences, were reported in rats that received 1 % in the feed (ca. 0.5 g kg^{-1} d^{-1}) for 94 weeks. However, in combination with sodium nitrite in drinking water, endogenously produced nitroso compounds induced tumors [187].

Triisopropanolamine is reported to be nonsensitizing in humans when patch-tested as a 1.1 % additive in a cosmetic lotion [179] and as such in a guinea pig maximization test [184]. It was nonmutagenic in the Ames test [169] and failed to induce chromosomal aberrations in a mouse micronucleus assay [143]. When it was administered to rats at doses of 0.1 to 0.2 g kg^{-1} d^{-1} in drinking water for two weeks, no effects were observed at 0.1 and 0.33 g kg^{-1} d^{-1}, with minor, but noncharacteristic findings at the higher doses [138]. No evidence of a fetotoxic and teratogenic potential was found in the progeny from pregnant rats exposed to oral doses reaching maternal toxicity [137], [143].

Propanolamine (3-amino-1-propanol) showed no skin-sensitizing potential in an early epicutaneous nonstandard test with guinea pigs [188] and no mutagenic activity in the Ames test [143], as well as in an SCE assay with CHO cells [189]. In a screening test, oral treatment of rats and cats with about 0.78 g kg^{-1} d^{-1} in 4 and 5 % aqueous solutions for two weeks (10 applications) caused no significant macroscopic and histological abnormalities [188].

3-Methoxypropylamine caused skin sensitization in the Buehler test (5 %in ethanol) [139]. There was no evidence of mutagenic activity in the Ames test [139], [143], two unspecified assays, a UDS assay, and a micronucleus test [139].

3-(2-Ethylhexyloxy)propylamine was reported to be nonmutagenic in the Ames test [19].

9. References

General References

1. A. L. Kohl, R. Nielsen: *Gas Purification,* 5th ed., Gulf Publishing Co., Houston 1997.
2. K. Lindner: *Tenside-Textilhilfsmittel-Waschrohstoffe,* 2nd ed., **vol. 1,** Wissenschaftliche Verlagsgesellschaft, Stuttgart 1964.
3. E. J. Fischer: *Triethanolamine und andere Alkanolamine,* 4th ed., Straßenbau, Chemie und Technik Verlagsgesellschaft, Heidelberg 1953.
4. N. Schönfeldt: *Grenzflächenaktive Alkylenoxid-Addukte,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1976.
5. A. N. Schwartz, J. W. Perry, J. Berch: *Surface Active Agents and Detergents,* vol. II, Interscience, New York 1958.
6. John W. McCutcheon, Inc.: *Detergents and Emulsifiers,* Morristown, published yearly.

Specific References

7. A. Wurtz, *Justus Liebigs Ann. Chem.* **114** (1860) 51 – 54; **122** (1862) 226 – 232.
8. C. M. Yon, G. R. Atwood, C. D. Swaim, *Hydrocarbon Process.* 1979, July, 197 – 200.
9. BASF, DE-OS 1 953 263, 1972; US 4 014 933, 1977 (H. Corr, G. Boettger).
10. L. P. Somogyi: "CEH Product Review Ethyleneamines", *Chemical Economics Handbook,* SRI International, Menlo Park 1996.
11. H. Kindler et al., *Chem.-Ing.-Tech.* **37** (1965) 400 – 402.
12. H. Wenker: "Preparation of Ethyleneimine from Ethanolamine," *J. Am. Chem. Soc.* **57** (1935) 2328.
13. W. Seeliger, H. Hellmann et al., *Angew. Chem.* **78** (1966) 913 – 927; *Angew. Chem. Int. Ed. Engl.* **5** (1966) 875.
14. BASF, DE-OS 2 618 580, 1977 (H. Hoffmann, K. Merkel, H. Toussaint, D. Voges).
15. L. Knorr, P. Rössler: "Zur Kenntnis des Ethanolamins," *Ber. Dtsch. Chem. Ges.* **36** (1903) 1281.
16. BASF, FR 2 307 902. 1976 (H. Bosche, H. Lüdemann).
17. H. A. Bruson, *J. Am. Chem. Soc.* **58** (1936) 1741-45.
18. H. Druckrey, R. Preussmann, S. Ivankovic, D. Schmähl, *Z. Krebsforsch.* **69** (1967) 103 – 201.
19. BASF, DE-OS 1 768 335, 1972 (R. Dahlinger, W. Goetze, G. Schulz, U. Sönksen).
20. Mo och Domsjö AB, US 3 697 598, 1972 (B. J. G. Weibull, L. U. F. Folke, S. O. Lindstrom).
21. Nippon Shokubai, EP 0 652 207, 1994 (A. Moriya et al.).
22. Texaco, US 4 438 281, 1984 (F. L. Johnson, Jr.).

23. Texaco, EP 0 375 267, 1990 (N. J. Grice et al.).
24. BASF, unpublished measurements.
25. *Yukagaku* **15** (1966) no. 5, 215 – 220.
26. G. Schulz, W. Goetze, P. Wolf: "Kinetische Untersuchungen zur Oxethylierung von NH_3 in konz. wässrigen Lösungen," *Chem.-Ing.-Tech.* **40** (1968) no. 9/10, 446 – 448.
27. Oxirane Limited, GB 0 760 215, 1956 (A. J. Lowe et al.).
28. W. Wirth: "Nichtionische Tenside: Bedeutung, Herstellungsverfahren, Eigenschaften und Einsatzmöglichkeiten," *Tenside Deterg.* **5** (1975) 245 – 247.
29. BASF, DE-OS 2 810 135, (H. Bosche, H. Hammer, G. Jeschek).
30. Texaco, US 4 673 762, 1987 (J. H. Paslean et al.).
31. Mitsui Toatsu Chemicals, JP 62 019 558, 1987 (H. Sugawara et al.).
32. Dow Chemicals, US 3 207 790, 1965 (D. N. Glew).
33. *Seifen, Öle, Fette, Wachse, Kosmet. Aerosole* **17** (1977) 487 – 491.
34. S. Scholz-Weigl, S. Holzmann, M. Schechter: *Tenside in unserer Welt – heute und morgen; Berichte vom Welt-Tensid-Kongress München 1984 (Proceedings of the World Congress on Surfactants, Munich 1984),* Kürle Druck und Verlag, Gelnhausen 1984, pp. 89 – 105 and 167 – 174.
35. H. Grossmann, *Fette, Seifen, Anstrichm.* **74** (1972) 58 – 63.
36. H. Manneck, *Seifen, Öle, Fette, Wachse* **4** (1959) 77 – 79.
37. W. Kritchevsky, US 089 212, 1937; US 2 094 608, 1937.
38. Procter & Gamble, DE-OS 2 527 793, 1976 (C. H. Nicol).
39. Procter & Gamble, GB 1527 – 141, 1975 (J. L. Arnan, C. R. Barrot, P. N. Crisp, I. Siklossy).
40. Procter & Gamble, EP 8 – 142, 1978 (P. Gosset, R. Ludewick).
41. H. Schneider: "Über die Verwendung von Mahlhilfen bei der Zementmahlung," *Zem.-Kalk-Gips* **22** (1969) 193 – 201.
42. W. Lieber, W. Richartz: "Einfluß von Triethanolamin, Zucker und Borsäure auf das Erstarren und Erhärten von Zement," *Zem.-Kalk-Gips* **25** (1972) 403 – 409.
43. Dow Chemical, US 146 489, 1971; DE 2 225 015, 1972 (W. Dowd).
44. Jefferson Chemical, DE 1 808 941, 1969 (V. A. Currier, A. Lichaa).
45. Yokkaichi Gosei, JP 35 515, 1969.
46. Hori Todashi: "Alkylamines and Derivatives V – Ethoxylated Chemicals," *Bosei Kanri* **20** (1976) no. 12, 37.
47. IG-Farbenindustrie, DE 650 574, 1928.
48. Carbides and Carbon Chem. Corp., US 1 904 013, 1927 (E. Reid, D. C. Lewis).
49. Du Pont, US 2 163 099, 1938 (R. Maxwell).
50. Texaco, US 3 784 566, 1970 (J. A. Patterson).
51. Dow Chemical, US 4289 738, 1980 (C. Martin, R. L. Pearce).
52. J. Gropp: *Chemie und Ernährung – BASF Forum Tierernährung am 28./29. 10. 1982,* Verlag Wissenschaft und Politik, p. 111.
53. BASF, *Safety data sheets,* Ludwigshafen 1985.
54. *Houben Weyl,* **6/4,** p. 511. CS 146 401(1972).
55. Hoechst, DE 832 152, 1952 (H. Bestian, W. Schuhmacher).
56. A. Davison, B. Milwidisky: *Synthetic Detergents,* 4th ed., C.R.C. Press, Cleveland, Ohio, 1968.
57. G. D. Jones, *J. Org. Chem.* **9** (1944) 491.
58. E. Jungermann, D. Taber: *Nonionic Surfactants,* Marcel Dekker, New York 1967.
59. US 3 025 325, 1959 (R. Velgos).
60. Dow Chemical, US 2 649 483, 1953 (M. Huscher, M. Long et al.).
61. Jefferson Chemical, US 3 317 609, 1967 (L. Sherman).
62. FR 2 103 968, 1972 (B. S. Polanskyi).
63. Merck & Co., US 2 189 808, 1940 (R. T. Major, H. T. Bonnett).
64. S. Hünig, *J. Prakt. Chem.* **5** (1958) 224 – 232.
65. Oxirane, DE 1 004 620, 1957 (A. Lowe, D. Butler et al.).
66. Progressive Products Co., US 3 454 647, 1966 (P. Kersnar).
67. Texaco, US 3 761 523, 1972 (R. E. Reid, T. P. Chen).
68. Shell, BE 834 514, 1975 (W. A. Gleich).
69. Mo och Domsjö AB, DE-OS 1 941 859, 1969 (B. J. Weibull, L. U. Thorsell et al.).
70. Chemische Fabrik Pfersee, GB 827 469, 1956.
71. US 2 673 215, 1952 (V. J. Keenan, F. J. Gozlow).
72. D. V. Beio, J. McCarthy: "The Fatty Alkanolamines," *Cosmet. Technol.* 1980, Mar.
73. Asahi Denka Kogyo, JP 56 032 581, 1979.
74. L. Thiele, *Acta Polym.* **30** (1979) 323 – 342.
75. H. John: *Epoxidharze,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1969.
76. Chemische Werke Buna, VEB, DD 240 648, 1984.

77. *Hydrocarbon Process.* 1975, no. 4, 79 – 111. E. E. Isaals, F. D. Otto et al., *Can. J. Chem. Eng.* **55** (1977) 210.
78. J. Rosen, US 2 993 934, 1956.
79. UCC, US 2 129 805, 1934 (A. Lamendin, B. Matel, J. Dhenin). Société des Produits Chimiques Marles-Kuhlmann, DE-OS 1 918 388, 1969 (A. L. Wilson).
80. H. W. Heine, R. W. Greiner, M. A. Boote, B. A. Brown, *J. Am. Chem. Soc.* **75** (1953) 2505 – 2506. Y. Minoura, T. Takebayashi, C. Price, *J. Am. Chem. Soc.* **81** (1959) 4589 – 4692.
81. A. Tettamanzi, *Atti Accad. Sci. Torino Cl. Sci. Fis. Mat. Nat.* **69** (1934) 369 – 377. B. Kirson, *Bull. Soc. Chim. Fr.* 1958, 223 – 226. ICI, GB 899 948, 1959 (R. Hulse, H. J. Twitchett). British Titan Products, BE 665 840, 1965.
82. Merck & Co., US 2 118 054, 1935 (R. Major, H. T. Bonnett).
83. E. D. Bergmann, A. Kaluszyner, *Recl. Trav. Chim. Pays-Bas* **78** (1957) 315 – 326.
84. Hoffmann – La Roche, GB 582 156, 1945.
85. American Cyanamid, GB 591 489, 1947.
86. C. H. Boehringer Sohn, DE 1 012 598, 1957 (E. Haak, A. Hagedorn).
87. H. Tiltscher, *Angew. Makromol. Chem.* **25** (1972) 1 – 14.
88. Oxirane, GB 763 434, 1951 (A. J. Lowe, D. Butler, M. Megde).
89. Carbon Chemicals Corp., US 2 137 627, 1938 (J. N. Wickert).
90. Atlantic Richfield, US 3 457 313, 1966 (T. Baker).
91. Union Carbide, US 2 813 877, 1954 (J. Lambrach).
92. Atlas Powder Corp., US 2 775 604, 1953.
93. Esso Res. & Eng. Corp., US 2 764 551, 1954 (S. Lippncott, C. Muessig, P. U. Smith).
94. Chemische Fabrik Pfersee, GB 827 469, 1956.
95. The Atlantic Refining Co., US 2 673 215, 1952 (F. Gozlow).
96. Hermann Wülfing Wings, BE 574 022, 1966.
97. Nox-Rust Chem. Corp., US 2 512 949, 1945 (E. Lieber).
98. Montclair Res. Corp., US 2744 883, 1956 (I. Hamer, J. Rust).
99. Carbide and Carbon Chemicals Corp., US 2 137 627, 1938 (M. Reed).
100. BASF, DE 1 052 685, 1956 (F. Ebel, E. Meyer).
101. Ciba, GB 770 570, 1955.
102. K. Ridder (ed.): *Gefahrengutverordnung Straße 1997 – ADR 1997,* Ecomed-Verlagsgesellschaft, Landsberg 1999.
103. P. W. Langvardt, R. G. Melcher, *Anal. Chem.,* **52** (1980) 669 – 671.
104. *NIOSH Registry of Toxic Effects of Chemical Substances,* US Department of Health, Education, and Welfare, Washington, D.C., 1986.
105. G. D. Clayton et al. (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2B,** John Wiley & Sons, New York 1981.
106. G. v. Foster, *Diss. Abstr. Int. B* **32** (1972) no. 11, 6549.
107. R. Hartung et al., *Toxicol. Appl. Pharmacol.* **17** (1970) 308.
108. G. Angelini et al., *Contact Dermatitis* **12** (1985) 263.
109. B. Scheurer, *Hautarzt* **34** (1983) 126.
110. Cosmetic Ingredient Review (CIR), *J. Am. Coll. Toxicol.* **2** (1983) 183 – 235.
111. H. Hoshino et al., *Cancer Res.* **38** (1978) 3918 – 3921.
112. A. Maekawa et al., *J. Toxicol. Environ. Health* **19** (1986) 345.
113. ACGIH: *Documentation of the Threshold Limit Values,* 6th ed., ACGIH, Cincinnati 1991, pp. 462/463.
114. H. F. Smyth et al., *Arch. Ind. Hyg. Occup. Med.* **10** (1954) 61.
115. H. H. Cornish et al., *Am. Ind. Hyg. Assoc. J.* **30** (1969) 46.
116. BASF, unpublished results, 1963 – 1983.
117. BASF, EP 0 036 152, 1980 (G. Boettger et al.).
118. BASF, EP 0 004 015, 1979 (H. G. Bosche et al.).
119. Dow Chemicals, US 3 207 790, 1965 (D. N. Glew).
120. G. M. Bohlmann: *Glyphosate,* PEP Review 97-10, SRI Consulting, Menlo Park 1999.
121. Nippon Shokubai, EP 228 898, 1987.
122. A. R. Doumaux, R. K. Barnes, *J. Org. Chem.* **38** (1973) 3630 – 3632.
123. B. E. Dixon et al., *J. Soc. Chem. Ind.* **69** (1950) 69 – 70.
124. K. Starke et al., *J. Inorg. Nucl. Chem.* **6** (1958) 130 – 133.
125. BASF, DE 3 010 105, 1980 (H. Hammer et al.).
126. Y. S. Ahn et al., *Hwahak Konghak* **18** (1980) 393 – 400.
127. Shell Internationale Research Maatschappij B.V., GB 1 529 193, 1975 (W. A. Gleich).
128. Shell Internationale Research Maatschappij B.V., DE 2 547 328, 1976 (E. Jung et al.).
129. H. Meissner, W. Hefner, BASF AG, CO_2 Removal by BASF's Activated MDEA Process, 1990 European Conference on Energy Efficient Production of Fertilizers, Bristol, UK, January 24 – 25, 1990.

130. R. Welker, V. Schuda, BASF AG, Conversion of the Gas Purification System to BASF aMDEA Solvent at Pt. Badak NGL, AVPG 2000, Caracas, May 10 – 12, 2000.
131. R. S. Hugo, BASF AG, Revamp of Hot Potassium Carbonate Units to BASF's aMDEA Process, NITROGEN '99, Caracas, February 28 – March 02, 1999.
132. B. Ballantyne, H. W. Leung, *Vet. Hum. Toxicol.* **38** (1996) 422 – 426.
133. Union Carb. Corp., TSCATS OTS0536032, 1990.
134. P. Schmidt et al., *Z. Ges. Hyg.* **20** (1974) 393 – 398.
135. Union Carb. Chem. and Plastics Comp.: Toxicology Summary -N,N-Dimethylethanolamine, Rep. No. **TS 212** (10/90).
136. D. R. Klonne et al., *Fund. Appl. Toxicol.* **9** (1987) 512 – 521.
137. BIBRA: Toxicity Profile Triisopropanolamine, 1990.
138. Dow Chem. USA, 1984/1985, cited in [179].
139. Texaco Inc., TSCATS OTS0529924, 1991.
140. H. F. Smyth et al., *Am. Ind. Hyg. Assoc. J.* **23** (1962) 103.
141. R. J. Papciak, V. T. Mallory, *Acute Tox. Data* **1** (1990) 16 – 17.
142. J. B. Knaak et al. in: G. W. Ware (ed.): *Rev. Environ. Contam. Toxicol.* **149** (1997) 1 – 87.
143. BASF, unpublished results, 1989 – 1996.
144. A. B. Liberacki et al., *Fundam. Appl. Toxicol.* **31** (1996) 117 – 123.
145. NTP (National Toxicology Program): Toxicity studies on diethanolamine, Tech. Rep. **20,** Publ. 92-3343, NIH 1992.
146. NTP (National Toxicology Program): Toxicity studies on diethanolamine, Tech. Rep. **478,** Publ. 99-3968, NIH 1999.
147. J. W. Spalding et al., *Toxicol. Sci.* **53** (2000) 213 – 223.
148. W. T. Stott et al., *Toxicol. Lett.* **114** (2000) 67 – 75.
149. NTP (National Toxicology Program): Developmental screen for diethanolamine. NTP Study No. TER-96-001, NIEHS 1999.
150. Y. Konishi et al., *Fundam. Appl. Toxicol.* **18** (1992) 25 – 29.
151. Eastman Kodak Co.: TSCATS, OTS0516742, 1982.
152. B. K. Nelson et al., *Environ. Health Persp.* **57** (1984) 261 – 271.
153. Mosberg et al. (1985a,b): Reports in inhalation studies. Battelle Columbus Lab. Ohio.
154. K. Inoue et al., *Mut. Res.* **101** (1982) 305 – 313.
155. Ishidate and Kada: *Environmental Mutagens Data Book* **1** (1980) 190.
156. R. Hartung, H. H. Cornish, *Fd. Cosmet. Toxicol.* **7** (1969) 595 – 602.
157. Bristol-Myers, DE-OS 2 105 232, 1971.
158. BASF, unpublished results, 1999 – 2000.
159. Texaco, unpublished results, 1982 – 1991.
160. H. W. Leung, *Vet. Human Toxicol.* **40** (1998) 65 – 67.
161. H. W. Leung, B. Ballantyne, *Mut. Res.* **393** (1997) 7 – 15.
162. F. Stenbaeck et al., *Mech. Ageing Dev.* **42** (1988) 129 – 138.
163. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 4th ed., **vol. II, Part B,** J. Wiley & Sons, New York 1994, p. 1158.
164. Hoechst AG, unpubl. results, 1988 (sponsored by BG Chemie/Germany).
165. NTP (National Toxicology Program) TR 449 (Draft) (1994).
166. Union Carbide Corp., unpublished results, 1970.
167. H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten,* Wiley-VCH, Weinheim, October 1995.
168. A. Nakamura et al., *Contact Derm.* **31** (1994) 72 – 85.
169. E. Zeiger et al., *Environ. Molec. Mutagen.* **9** (1987) (Suppl. 9), 1 – 110.
170. EXXON Biomed. Sci., Report 208618, 1990.
171. J. Hinz et al., *Fundam. Appl. Toxicol.* **18** (1992) 418 – 424.
172. Atochem North America: TSCATS OTS0530455-1, 1991.
173. EXXON Biomed. Sci., Report 108634, 1991.
174. M. S. Werley et al., *J. Toxicol. Cut. Toxicol.* **16** (1997) 157 – 171.
175. A. Takahashi, H. Ono, *Chem. Express* **8** (1993) 785 – 788.
176. C. C. Conoway et al., *Mut. Res.* **136** (1984) 153 – 157.
177. N. Lahitova et al., *Folia Microbiol.* **27** (1982) 269 – 271.
178. M. Karasova et al., *Bratisl. Lek. Listy* **82** (1984) 1423 – 1429. *Chem. Abstr.* 102178974h.
179. Cosmetic Ingredient Review (CIR), *J. Am. Coll. Toxicol.* **6** (1987) 53 – 76.
180. H. Shimizu et al., *Jpn. J. Ind. Health* **27** (1985) 400 – 419.
181. H. F. Smyth et al., *Arch. Ind. Hyg. Occup. Med.* **4** (1951) 119 – 122.

182. Y. Konishi et al., *IARC Sci. Publ.* **105** (1990) 318 – 321.
183. K. Fujimoto et al., *Contact Derm.* **21** (1989) 56.
184. E. N. Burkackaja et al., *Gig. i Sanit.* **3** (1986) 46 – 48.
185. K. Mortelsman et al., *Environ. Mutagen.* **8** (Suppl. 7) 1 – 119.
186. Dow Chemical Co., unpublished results, 1993.
187. K. Yamamoto et al., *Carcinogen.* **10** (1989) 1606 – 1611.
188. BASF, unpubl. results, 1961 – 1962.
189. M. Sorsa et al., *Mut. Res.* **204** (1988) 465 – 479.
190. E. Zeiger et al., *Environ. Molec. Mutagen.* **11** (1988) (Suppl. 12), 1 – 158.

Ethers, Aliphatic

The following are separate keywords: Cellulose Ethers, Crown Ethers, Dimethyl Ether, Dioxane, Epoxides, Ethylene Oxide, Furan and Derivatives, Methyl Tert-Butyl Ether, Poly(Vinyl Ethers), Propylene Oxide, Tetrahydrofuran, Vinyl Ethers. For allyl ethers → Allyl Compounds ; for aromatic ethers → Naphthalene Derivatives, → Phenol Derivatives ; for glycol ethers → Ethylene Glycol.

Wilhelm Heitmann, Hüls AG, Herne, Federal Republic of Germany (Chaps. 2 and 3, Sections 4.1, 4.3, and 4.4)

Günther Strehlke, Deutsche Texaco, Moers, Federal Republic of Germany (Section 4.2)

Dieter Mayer, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 5)

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 3
3. Synthesis 3
4. Individual Aliphatic Ethers 4
4.1. Diethyl Ether 4
4.2. Diisopropyl Ether 7
4.3. Other Aliphatic Ethers 8
4.3.1. Di-*n*-propyl Ether 8
4.3.2. Di-*n*-butyl Ether 9
4.3.3. Diamyl Ether 9
4.4. Chloroalkyl Ethers 10
4.4.1. Bis(chloromethyl) Ether 10
4.4.2. Bis(2-chloroethyl) Ether 10
4.4.3. Bis(2-chloroisopropyl) Ether 11
5. Toxicology 11
6. References 13

1. Introduction

Aliphatic ethers are compounds of the type R–O–R′, where R and R′ are alkyl radicals that may or may not be identical and may be substituted. Structurally, ethers may be regarded as oxides. However, as far as their chemical stability is concerned, ethers have little in common with organic oxides (basicity, see Section 2.1). They are generally regarded as anhydrides of alcohols since they can be prepared by dehydration of alcohols. Ethers are classified as either symmetrical or unsymmetrical, the latter being the anhydrides of two different alcohols. Ethers are named according to IUPAC rules as hydrocarbon derivatives, e.g., ethoxyethane.

Only the low molecular mass ethers are discussed in this article; for a list of separate articles dealing with other ethers, see under the main title of this article.

2. Properties

2.1. Physical Properties

The lower molecular mass aliphatic ethers have a pleasant and characteristic odor. They are highly volatile and highly mobile. Their boiling points are lower than those of the corresponding alcohols and are similar to those of hydrocarbons of equivalent molecular mass and structure. This indicates that ether molecules do not associate in the liquid phase. Ethers are insoluble or only poorly soluble in water but they are readily miscible with almost all organic liquids. Their solvent power for many organic compounds is considerable and is often enhanced in the presence of the corresponding alcohol. Some physical properties of commercially important ethers are summarized in Table 1 and the most important properties of others in Table 2.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_023

Table 1. Physical properties of commercially important ethers

Ether	CAS registry number	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}	Heat of vaporization, kJ/kg	Flash point, °C	Ignition temperature, °C	Explosive limits in air, vol % ether
Dimethyl	[*115-10-6*]	−138.5				−42	240	3.0–18.6	
		−24.8							
Diethyl	[*60-29-7*]	−116.3	34.5	0.7135	1.3527	362	−40	180	1.7–48
Di-*n*-propyl	[*111-43-3*]	−122	90.5	0.7360	1.3807		−21		
Diisopropyl	[*108-20-3*]	−86.2	68.5	0.7258	1.3682	285	−28	405	1.0–21
Di-*n*-butyl	[*142-96-1*]	−95.2	142.4	0.7704	1.3981	288	25	175	0.9–8.5
Di-*n*-amyl	[*693-65-2*]	−69.3	187.5	0.7751	1.4119	276 *	57 *	171 *	
Diisoamyl	[*544-01-4*]		173.2	0.7777	1.4085				
Bis(chloromethyl)	[*542-88-1*]		105	1.315	1.4346				
Bis(2-chloroethyl)	[*111-44-4*]	−50	178	1.2199	1.4572	268.4	55	365	
Bis(2-chloroisopropyl)	[*39638-32-9*]	<−20	187.1	1.1127	1.4505		85		

* For technical-grade amyl ether

2.2. Chemical Properties

Unsubstituted ethers are chemically very stable. At room temperature, they are neither attacked by alkali metals nor undergo base hydrolysis. They are also resistant to acids, with the exception of hydroiodic acid which cleaves ethers with formation of alkyl iodide; this reaction is the basis of the Zeisel method for the determination of alkoxy groups.

At higher temperature, ethers are also hydrolyzed by other acids, especially nitric and hydrobromic acids, acyl halides, and phosphorus pentachloride [10]. The rate of hydrolysis increases rapidly with the degree of branching of the alkyl group. Therefore, breakdown or removal of nonvolatile methoxy compounds or methyl-protecting groups by boiling with hydrobromic acid is advantageous. At higher temperature, ethers can be cleaved to produce olefins, water, and alcohol, especially in the presence of catalysts such as aluminum oxide.

Ethers react more quickly with chlorine or bromine than do the corresponding alkanes, substitution taking place mainly in the α-position. Mono- as well as bis(1-haloalkyl) ethers can be obtained by appropriate control of the reaction.

Ethers form peroxides in the presence of atmospheric oxygen, especially when exposed to light. This has long been known because violent explosions can occur toward the end of ether distillations due to the peroxides present in the distillation residue [11–14]. Peroxides can be detected with vanadium(V) sulfate or with iron(II) sulfate and potassium rhodanide [15]. They can be destroyed by treatment with an iron(II) salt solution [12], potassium pyrosulfite, or triethylenetetramine [16].

One of the important properties of ethers is their ability to form oxonium compounds with protons and Lewis acids, e.g., boron trifluoride, zinc chloride, or aluminum chloride. This is a result of the nucleophilic character of the ether oxygen:

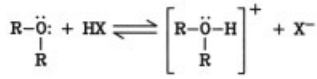

Therefore, ethers also dissolve in cold, concentrated sulfuric acid to form oxonium sulfates. This reaction is exploited to separate ethers from

hydrocarbons, whereby the ether is recovered by diluting the sulfuric acid. Organometallic compounds (in particular, organomagnesium Grignard reagents) dissolve in ethers to form complexes. The tertiary oxonium salts $[R_3O]^+X^-$ described by MEERWEIN are effective alkylating agents [1].

3. Synthesis

Ethers are generally prepared by catalytic dehydration of alcohols or by reaction of alkyl halides with alkoxides:

1) *Catalytic dehydration of alcohols in the liquid phase.* Acids (particularly sulfuric acid) and electrophilic metal and nonmetal halides are effective catalysts for dehydrating alcohols in the liquid phase to produce ethers. Moderately strong acids and salts of weak bases and strong acids are most effective [17] but favorable results are obtained only for the lower alcohols (1 – 3 carbon atoms). The reaction proceeds via intermediate formation of an oxonium compound:

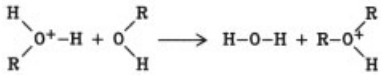

2) *Catalytic dehydration of alcohols in the gas phase* over, e.g., aluminum oxide, anhydrous alumina, or bauxite at a temperature of 180 – 250 °C gives ether yields of up to 75 %.
3) *Reaction of alkyl halides with alkali-metal alkoxides* by the Williamson synthesis [18]:
RONa + R'X ⟶ R–O–R' + NaX
This method is used for preparing most unsymmetrical and symmetrical aliphatic ethers. Symmetrical ethers can be prepared similarly from an alkyl halide and silver oxide. Thus, di-*tert*-butyl ether [*6163-66-2*], whose preparation is difficult, can be obtained by reacting *tert*-butyl chloride with silver carbonate [19].
Introduction of dimethyl sulfoxide as a solvent considerably improved yields in the Williamson ether synthesis [20]. A mixture of sodium hydroxide and alcohol can be used in place of the alkoxide.
4) *Other methods.* Ethers can also be prepared from secondary or tertiary alcohols [21] or by indirect addition of an alcohol to an olefin [22].

4. Individual Aliphatic Ethers

Dimethyl ether and Methyl *Tert*-Butyl Ether are separate keywords.

4.1. Diethyl Ether

Diethyl ether [*60-29-7*], also known as ethoxyethane or simply ether, CH_3CH_2–O–CH_2CH_3, M_r 74.12, is one of the most important ethers. It is a clear, mobile liquid with a sweetish, slightly pungent, characteristic odor.

Physical Properties (see also Table 1). The vapor – liquid equilibria of the diethyl ether – ethanol and diethyl ether – ethanol – water systems at different pressures are described in [23]. Binary azeotropes of diethyl ether are listed in Table 3. Diethyl ether is completely miscible with common organic solvents, but only partially miscible with water.

Vapor pressure of diethyl ether:

t, °C	0	10	20	30	40	60	80	100
p, kPa	18.89	38.88	56.28	86.29	122.8	231.1	396.4	647.2

Critical data t_{crit} 194 °C; p_{crit} 3.63 MPa; ϱ_{crit} 262.5 kg/m^3
Specific heat capacity c_p:

t, °C	0	30	120	180
c_p, kJ kg^{-1} K^{-1}	2.21	2.29	3.36	4.36

Heat of formation (25 °C): −273 kJ/mol
Heat of combustion: 36.85 MJ/kg

Density:

t, °C	0	10	15	20	30
p, kg/m^3	736.3	725.0	719.3	713.5	701.8

Electrical conductivity: 3.7×10^{-13} S
Dielectric constant (20 °C): 4.34; ether vapor (103.3 kPa, 100 °C): 1.0049
Viscosity
(20 °C): 0.245 mPa · s;
(25 °C): 0.223 mPa · s
Surface tension (20 °C): 17.1 mN/m

Table 2. Physical properties of some representative ethers

Ether	CAS registry number	Formula	*bp*, °C	d_4^{20}	n_D^{20}
Di-*sec*-butyl	[*6863-58-7*]	$[(CH_3)(C_2H_5)CH]_2O$	122	0.7590 *	1.3931
Diisobutyl	[*628-55-7*]	$[(CH_3)_2CHCH_2]_2O$	123	0.7600	1.4107
Di-*tert*- butyl	[*6163-66-2*]	$(CH_3)_3C\text{-}O\text{-}C(CH_3)_3$	108	0.7622	1.3946
Di-*n*-hexyl	[*112-58-3*]	$C_6H_{13}\text{-}O\text{-}C_6H_{13}$	226	0.7936	1.4204
Methyl *n*-propyl	[*557-17-5*]	$CH_3\text{-}O\text{-}C_3H_7$	38.9	0.738	1.3579
Methyl isopropyl	[*598-53-8*]	$CH_3\text{-}O\text{-}CH(CH_3)_2$	32	0.7347 **	1.3579
Methyl *n*-butyl	[*628-28-4*]	$CH_3\text{-}O\text{-}C_4H_9$	70	0.7443	1.3736
Methyl *tert*- butyl	[*1634-04-4*]	$CH_3\text{-}O\text{-}C(CH_3)_3$	55.1	0.7406	1.3690
Ethyl *n*-butyl	[*628-81-9*]	$C_2H_5\text{-}O\text{-}C_4H_9$	91.4	0.7490	1.3818
Bis(1-chloroethyl)	[*6986-48-7*]	$(CH_3CHCl)_2O$	113	1.1285	1.4250
1-Chloroethyl ethyl	[*7081-78-9*]	$CH_3CHCl\text{-}O\text{-}C_2H_5$	98 (decomp.)	0.9495	1.4040
2-Chloroethyl ethyl	[*628-34-2*]	$ClCH_2CH_2\text{-}O\text{-}C_2H_5$	106	0.9945	1.412

* d_4^{25}; ** d_{20}^{20}

Solubility limits in the system diethyl ether – water:

t, °C	0	10	20	30	60	100
Water in diethyl ether, wt %	1.00	1.09	1.19	1.30	1.80	
Diethyl ether in water, wt %	11.7	8.7	6.5	5.2	3.6	2.1

Table 3. Binary azeotropes of diethyl ether

Component	Component concentration, wt %	Azeotrope *bp*, °C
Water (98 kPa)	1.26	34.15
Water (231 kPa)	2.0	60
Water (1080 kPa)	4.5	114
Acetaldehyde	76.5	18.9
Methyl formate	50	28.2
Pentane	32	33.4
Propylene oxide	49.6	32.6

Chemical Properties (see also Section 2.2). Due to its high chemical stability, diethyl ether has found no applications as an intermediate. Although it can be used to prepare 1-chloroethyl ethyl ether and bis(1-chloroethyl) ether by chlorination at low temperature [24], the reaction is not exploited commercially.

Diethyl ether is oxidized to acetic acid by strong oxidants such as chromium trioxide or concentrated nitric acid. If nitric acid is used, a risk of explosion exists, especially if concentrated sulfuric acid is also present.

It must once again be stressed that peroxides may be formed when diethyl ether comes into contact with air. On distillation of the ether, the peroxides may be concentrated and then decompose violently [11–14]. Therefore, diethyl ether should always be checked for the presence of peroxides before use [15]. The peroxides can be removed by shaking the diethyl ether with reducing agents such as 5 % aqueous iron(II) sulfate solution [11], [25]. Peroxide formation can be prevented by the addition of radical scavengers such as phenols or amines. 2,6-Di-*tert*-butyl-*p*-cresol [*128-37-0*] is often added at a concentration of 1 – 30 mg/kg, but other stabilizers (e.g., sodium diethylthiocarbamate) are also used, depending on the purpose for which the ether is employed.

Production. In countries where ethanol is produced synthetically, diethyl ether is produced in sufficient quantities as a byproduct to make its synthesis unnecessary (→ Ethanol). Otherwise, ether is produced from ethanol by dehydration, either with sulfuric acid or catalytically in the vapor phase over alumn. The latter process is simpler, but gives lower yields.

In the *sulfuric acid process*, ethanol and sulfuric acid (96 %) are heated to ca. 125 °C in a ratio of 1 : 3. This is done in a lead-lined steel kettle that is fitted with a heating coil and further protected by an acid-resistant brick lining. The ether is distilled with part of the unreacted alcohol. Any sulfuric acid and sulfur dioxide (produced via the reduction of the sulfuric acid by ethanol with formation of acetaldehyde) present in the distillate are removed by washing with dilute sodium hydroxide solution.

Ethanol is introduced into the sulfuric acid either in vapor or in liquid form depending on the degree of conversion. The temperature in the reactor must be carefully controlled because

the reaction should not proceed below 125 °C or over 130 °C. The process runs continuously for several months and is then stopped to replace the sulfuric acid, which becomes heavily contaminated with tarlike polymers. The yield is ca. 95 %.

Catalytic dehydration over alumn is carried out from 180 to 230 °C, the higher temperatures being required toward the end of the life of the catalyst. The reaction is carried out in a steel tube furnace, which contains the alumn catalyst. The ethanol (94.5 wt %) is first passed through a vaporizer and then over the catalyst. The reaction heat (ca. 25 kJ/mol) is removed by pumping water under pressure to generate steam. The drum pressure is also utilized to control the temperature in the reactor. Ethanol conversion per pass is 60 – 80 %, depending on the catalyst activity. The vapor leaving the reactor is condensed and washed with dilute sodium hydroxide solution to remove traces of sulfuric acid. The maximum yield is 90 % due to a side reaction, which results in the formation of ethylene.

Water is removed from the alumn catalyst by heating crystalline potassium aluminum sulfate in flat iron pans until a viscous melt is obtained. After cooling, the resulting mass is ground to give pieces with a diameter of ca. 0.5 – 1 cm; more water is then removed by gentle heating to give a residual water content of 15 – 20 %.

During *ethanol synthesis* by the direct catalytic hydration of ethylene (→ Ethanol), about 2 % of the ethylene is converted to diethyl ether as a byproduct. Indirect hydration of ethylene with sulfuric acid also yields 10 % ether. The amount of ether thus obtained exceeds the demand. Therefore, the ether yield is matched to the demand by recycling it into the process (for reaction equilibrium, see [26]) or by appropriate adjustment of the reaction conditions.

Purification. Irrespective of the process used, the diethyl ether must be purified to eliminate small amounts of acetaldehyde, hydrocarbons, and varying amounts of ethanol and water. Removal of the acetaldehyde by distillation leads to loss of ether due to the formation of a binary azeotrope. Removal is also incomplete because part of the acetaldehyde is bound by ethanol as an acetal and is liberated during further processing. Therefore, high-quality ether cannot be obtained by this method.

Acetaldehyde can be removed by hydrogenation or by washing the ether vapor with sodium hydrogen sulfite solution; subsequent washing with sodium hydroxide solution is required to remove any sulfur dioxide present. Water can be removed from the ether azeotrope by passing the vapor over a desiccant such as sodium hydroxide flakes or through a molecular sieve or by high-pressure distillation. At a pressure of ca. 1.1 MPa, a binary azeotrope containing 4.5 wt % water leaves the top of the fractionating column and, after condensation and cooling, separates into two phases. The upper phase is returned to the column, whereas the aqueous phase is recycled into the low-pressure ether column.

For environmental reasons, hydrogenation is preferred for purification and high-pressure distillation for dehydration.

Quality Specifications and Analysis. Two grades of diethyl ether are commercially available: *technical-grade ether* and *anesthetic ether.* There are no universally accepted specifications for technical-grade ether, but it is conventionally free of peroxides. Anhydrous technical-grade ether has a maximum water content of 500 mg/kg. Anesthetic ether has a neutral reaction, is free from foreign odors, passes the tests for peroxides and aldehydes, and has a maximum water content of 0.2 % [27].

Potassium iodide – starch solution is used to test for traces of peroxide: no color should develop after 30 min in the dark.

Nessler's reagent is used to test for aldehydes and acetone: no more than a weak opalescence is permissible after 5 min in the dark. Water is determined by the Karl Fischer method.

Storage and Transportation. Diethyl ether is a hazardous, flammable liquid, it is subject to the Gefahrstoffverordnung of the Federal Republic of Germany [28] and must be appropriately labeled for transportation. Its hazard classifications are as follows:

VbF	A I
GGVE/GGVS	3/2 a
IMDG code	3.1, UN No. 1155
Ignition group (VDE 0171)	G 4
United States	CFR 49: 172.02, flammable liquid

Diethyl ether is stored in mild steel containers. Blanketing with nitrogen is recommended but not essential. Tin-plate cans, mild steel drums, and aluminum containers are suitable for transportation, but zinc-coated vessels are not.

Being heavier than air (ca. 2.5-fold), ether vapor creeps along the ground Therefore, good ventilation is necessary when ether is handled to prevent the formation of ignitable mixtures. Because of its low autoignition temperature (180 °C), diethyl ether vapor can ignite even if it comes into contact with a noninsulated 1.0-MPa steam pipeline.

The fact that ether is a poor conductor of electricity presents a further hazard because static electrical charge can be generated in flowing ether. Therefore, the flow rate of ether must be restricted; a limit of 1 – 1.5 m/s, for example, has been set for pipelines with a diameter of 12 mm [8], [10]. Pipelines and pumps must be grounded (earthed), and flanged joints must be bridged by a conductor.

When diethyl ether is poured or transferred from one container to another, both containers must be grounded to prevent the buildup of static charge and possible explosions. If glass or plastic containers are used, filling should be performed by using a metallic pipeline that is below the surface of the ether or by placing grounded chains or other metallic conductors in the free-flowing ether to remove static electrical charge.

Uses. Diethyl ether is a good solvent for many oils, fats, resins, alkaloids, odorants, and dyes and, therefore, is widely used as a solvent and as an extractant (→ Solvents). When mixed with ethanol, diethyl ether is used to gel nitrocellulose and to dissolve collodion wool.

Ether is used as a reaction medium in the laboratory and in industry due to its chemical stability, low boiling point, and solvent properties for organometallic compounds (Grignard reagents). It is also employed as an anesthetic in medicine, but this use is declining (→ Anesthetics, General, Chap. 2.4.).

Economic Aspects. Of the ethers discussed in this article, diethyl ether is the only one to be produced in quantities that warrant discussion. Western Europe an production is 13 000 – 14 000 t/a.

4.2. Diisopropyl Ether

Diisopropyl ether [*108-20-3*], also known as 2-isopropoxyisopropane or isopropyl ether, $(CH_3)_2CH{-}O{-}CH(CH_3)_2$, M_r 102.18, is a clear, colorless, mobile, highly flammable liquid with a characteristic ethereal odor. Its solubility in water is low, but it is soluble in organic solvents and readily forms explosive peroxides. It has a higher boiling point, a lower vapor pressure, and a lower water solubility than diethyl ether, which means that diisopropyl ether can be recovered without significant loss after use as a solvent or as an extractant.

Physical Properties [31] (see also Table 1). Azeotropes of diisopropyl ether are listed in Table 4. Diisopropyl ether is miscible with most organic solvents, including acetic acid, but it is not miscible with glycols, glycerol, glycerol monoacetate, or ethanolamines. Its mutual solubility with water is lower than that of diethyl ether.

Saturation concentration (20 °C): 734 g/m^3
Vapor density (air = 1): 3.5
Evaporation rate (ether = 1): 1.6
Critical data:
t_{crit} = 226.9 °C;
p_{crit} = 2.875 MPa;
ϱ_{crit} = 0.265 kg/m^3
Specific heat capacity (30 °C): 2.21 kJ kg^{-1} K^{-1}
Heat of formation: −3443 kJ/kg
Heat of combustion (25 °C): 39.279 MJ/kg
Vapor pressure:

t, °C	0	10	20	30	40	68.47
p, kPa	5.53	9.93	17.5	24.9	37.7	101.3

Coefficient of cubic expansion: 0.00144 K^{-1}
Surface tension (23 °C): 32 mN/m
Solubility limits in the system water – diisopropyl ether:

t, °C	10	20	30	50
Water in diisopropyl ether, wt %	0.50	0.55	0.61	0.84
Diisopropyl ether in water, wt %	1.35	0.88	0.70	0.59

Chemical Properties (see also Section 2.2). The chemistry of diisopropyl ether differs little from that of diethyl ether. However, it has a significantly greater tendency to form peroxides and must, therefore, be tested for the presence of peroxides before use. This can be done by adding potassium iodide solution in 1 mol/L sulfuric acid and a few drops of ammonium molybdate solution and then titrating the free iodine with sodium thiosulfate.

As with diethyl ether, peroxide formation can be prevented during long storage periods largely by adding stabilizers such as 1-naphthol or 2,6-di-*tert*-butyl-*p*-cresol. Methods for removing peroxides are described in [30].

Table 4. Azeotropes of diisopropyl ether

Component	Component concentration, wt %	Azeotrope *bp*, °C
Binary azeotropes		
Water	4.5	62.2 ∗
Water	7.6	92.0 ∗ (203 kPa)
Ethanol	17.1	64.0
2-Propanol	14.1	65.2
Acetonitrile	17.0	61.7
Propionitrile	>4.0	<67.5
Acetone	61.0	54.2
Hexane	47.0	67.5
Methylcyclopentane	<20.0	<68.0
Ternary azeotropes		
Water	1.9	
		} 53.8
Acetone	53.4	
Water	5.0	
		} 59.0 ∗
Acetonitrile	13.0	
Water	4.0	
		} 61.0 ∗
Ethanol	6.5	
Water	9.1	
		} 128.5 ∗ (700 kPa)
Ethanol	14.2	
Water	4.1	
		} 61.6 ∗
2-Propanol	7.1	

∗ Condensate separates into two phases

Production. Diisopropyl ether is obtained as a byproduct in the production of isopropyl alcohol from propylene and water (→ Propanols). The crude ether contains water, isopropyl alcohol, C_6 hydrocarbons, and other byproducts.

Purification of the ether involves extraction of isopropyl alcohol with water and drying with the binary azeotrope, diisopropyl ether – water. Further distillation stages afford diisopropyl ether as an overhead product with a purity >99 %.

Uses. Diisopropyl ether is used as a solvent for chemical reactions. Its volatility and water solubility are lower than those of diethyl ether, thus, diisopropyl ether is used for extracting substances from aqueous solutions, e.g., for extracting acetic acid, lactic acid, and phenols [31] and for removing phenols from wastewater [32–34]. Diisopropyl ether is also used as an extractant for essential oils, nicotine, and vitamins. As with diethyl ether, diisopropyl ether dissolves nitrocellulose when mixed with ethanol and, therefore, can be used in the explosives industry. Due to its high solvent power and volatility, diisopropyl ether is used as a stain remover, since it tends to leave no solvent marks.

Diisopropyl ether is employed for the dealkylation of mineral oil because the solubility of alkane wax in diisopropyl ether (unlike that of the other components of mineral oil) decreases greatly when it is cooled or when, for example, acetone is added. Large quantities of diisopropyl ether, usually still containing 2-propanol, are used in the fuel industry (→ Automotive Fuels); diisopropyl ether has good antiknock properties and helps to prevent the carburetor from icing up and deposits from forming in the engine. Although diisopropyl ether is less effective than benzene in removing water from ethanol (→ Ethanol) or 2-propanol (→ Propanols), it can be used for this purpose when the treated alcohol must be free from benzene (<0.5 mg/kg) for toxicological reasons.

Storage and Transportation. The requirements for the storage and transportation of diisopropyl ether are similar to those described for diethyl ether. The hazard classifications for diisopropyl ether are as follows:

GGVE/GGVS	3/3 b
IMDG code	3.1, UN No. 1159
United States	CFR 49: 172.101, flammable liquid
VbF	AI

4.3. Other Aliphatic Ethers

Only a limited market exists for the remaining aliphatic ethers.

4.3.1. Di-*n*-propyl Ether

Di-*n*-propyl ether [*111-43-3*], M_r 102.17, $CH_3(CH_2)_2{-}O{-}(CH_2)_2CH_3$, is a colorless, mobile liquid with an ethereal odor.

Properties (see also Table 1). Azeotropes of di-*n*-propyl ether are listed in Table 5. The solvent properties of di-*n*-propyl ether are similar to those of diisopropyl ether (see Section 4.2). At 25 °C the solubility of water in di-*n*-propyl

ether is 0.68 %, and that of the ether in water is 0.25 %.

Table 5. Azeotropes of di-*n*-propyl ether

Component	Component concentration, wt %	Azeotrope *bp*, °C
Binary azeotropes		
Water		75
Methanol	72	63.8
Ethanol	44	74.5
1-Propanol	32	85.8
2-Propanol	52	78.2
Isobutanol	10	89.5
2-Butanol	22	87.0
tert-Butanol	52	79.0
Ternary azeotropes		
Water	11.7	} 74.8
1-Propanol	20.2	

The danger of peroxide formation is considerably less for di-*n*-propyl ether than for diisopropyl ether.

Vapor pressure:

t, °C	0	21.6	50.4	70.0
p, kPa	2.67	8.00	26.7	53.3

Production and Uses. Di-*n*-propyl ether can be prepared from 1-propanol by dehydration with sulfuric acid. It is used for the same purposes as diethyl ether, advantages are offered by its considerably higher boiling point.

4.3.2. Di-*n*-butyl Ether

Di-*n*-butyl ether, also known as 1-butoxybutane, is the only butyl ether that is commercially available. Di-*n*-butyl ether [*142-96-1*], $CH_3(CH_2)_3{-}O{-}(CH_2)_3CH_3$, M_r 130.22, is a colorless liquid with a characteristic, fruity odor.

Properties (see also Table 1). Azeotropes of di-*n*-butyl ether are listed in Table 6.

Vapor pressure:

t, °C	11.3	29.7	66.8	127.7
p, kPa	1.0	2.0	7.6	76.0

Viscosity (30 °C): 1.6 mPa · s
Surface tension (20 °C): 22.9 mN/m

Production. Di-*n*-butyl ether is produced by dehydrating 1-butanol with sulfuric acid. It is also obtained as a byproduct during the synthesis of the butyl esters of higher acids; it occurs in the recovered alcohol and can be obtained by extractive distillation with water, drying, and rectification.

Typical *specifications* are: boiling range, 139 – 144 °C; d_{20}^{20}, 0.7690 – 0.7715; water content, max. 0.1 %; acid content as butyric acid, max. 0.02 %; peroxide content as dibutyl peroxide, max. 0.05 %.

Table 6. Azeotropes of di-*n*-butyl ether

Component	Component concentration, wt %	Azeotrope *bp*, °C
Binary azeotropes		
Water	33	92.9
1-Butanol	82.5	117.6
1-Pentanol	50	134.5
Isoamyl alcohol	65	129.8
Ethylene glycol	6.4	139.5
Butyl acetate	95	125.9
2-Methoxyethanol	68	122
2-Ethoxymethanol	50	127
2-Propoxymethanol	37	138.5
Ternary azeotropes		
Water	29.9	} 90.6
1-Butanol	34.6	
Water	31.2	} 45 (1.3 kPa)
1-Butanol	24.6	
Water	24.7	} 86.6
2-Butanol	56.1	
Quaternary azeotrope		
Water	30	
1-Butanol	13	} 90.6
n-Butyl acetate	51	

Uses. Di-*n*-butyl ether is an excellent solvent for many natural and synthetic resins, fats, oils, organic acids, and alkaloids, especially at high temperature (→ Solvents). When mixed with ethanol or 1-butanol, di-*n*-butyl ether also dissolves ethyl cellulose. The mutual solubility of di-*n*-butyl ether with water is low: at 20 °C, the solubility of the ether in water is 0.03 % and that of water in the ether is 0.19 %. Di-*n*-butyl ether is consequently used as an extractant for aqueous solutions. It is also employed as a reaction medium (Grignard reactions).

4.3.3. Diamyl Ether

Technical-grade diamyl ether is a mixture of di-*n*-amyl ether (di-*n*-pentyl ether) [*693-65-2*], $CH_3(CH_2)_4{-}O{-}(CH_2)_4CH_3$, M_r158.28, and

Table 7. Binary azeotropes of diamyl ethers

Component	Di-*n*-amyl ether azeotrope		Diisoamyl ether azeotrope	
	bp, °C	Amyl ether content, wt %	*bp*, °C	Amyl ether content, wt %
Water	98.4		97.4	46
Ethylene glycol	168.8		162.8	81
Ethylene glycol monoacetate	180.8	52	170.2	72
2-Butoxyethanol	169.0	33	164.95	46
Bis(2-chloroethyl) ether	<176.5		169.35	61

diisoamyl ether (diisopentyl ether) [*544-01-4*], $(CH_3)_2CH(CH_2)_2-O-(CH_2)_2CH(CH_3)_2$, with traces of other amyl ethers and amyl compounds.

Properties (see also Table 1). Di-*n*-amyl ether and diisoamyl ether are colorless to pale yellow liquids with a pearlike odor. Binary azeotropes of diamyl ethers are listed in Table 7.

Ternary azeotropes:

Water – *n*-amyl alcohol – di-*n*-amyl ether: *bp,* 95.94 °C

Water – isoamyl alcohol – diisoamyl ether: *bp,* 94.4 °C

Technical-grade diamyl ether

Vapor pressure:

t, °C	18.6	57	96	129	150
p, kPa	0.133	1.33	8.00	26.6	53.3

Specific heat capacity (20 °C): 2.15 kJ kg^{-1} K^{-1}
Surface tension (20 °C): 24.8 mN/m

Production and Uses. The production and uses of diamyl ether are similar to those of di-*n*-butyl ether.

4.4. Chloroalkyl Ethers

A few symmetrical chloroalkyl ethers are commercially important. They can be synthesized by direct chlorination of the appropriate ethers or by the sulfuric acid process (see Chapter 3); however, the former method is not used on a commercial scale. Chloroalkyl ethers can also be produced by reacting aldehydes with hydrogen chloride [35]:

$$RCHO + HCl \rightleftharpoons RCHClOH$$

$$2\,RCHClOH \rightleftharpoons RClCH-O-CHClR + H_2O$$

The yield of the bis(1-chloroalkyl) ether can be improved by removing the water formed, e.g., by using chlorosulfonic acid.

4.4.1. Bis(chloromethyl) Ether

Bis(chloromethyl) ether [*542-88-1*], also known as dichlorodimethyl ether, M_r 114.96, $ClH_2C-O-CH_2Cl$, is a clear liquid and a strong irritant to the eyes and the respiratory system. Its physical properties are listed in Table 1.

Production and Uses. Bis(chloromethyl) ether is produced by reacting paraformaldehyde and chlorosulfonic acid in sulfuric acid (65 %) under cooling at a temperature of 26 – 28 °C. The reaction is controlled to prevent any hydrochloric acid escaping.

Bis(chloromethyl) ether is employed for chloromethylating compounds that cannot be chloromethylated with formaldehyde and hydrochloric acid, e.g., aromatic compounds and anilides.

4.4.2. Bis(2-chloroethyl) Ether

Bis(2-chloroethyl) ether [*111-44-4*], also known as dichloroethyl ether, M_r 143.01, $ClCH_2CH_2-O-CH_2CH_2Cl$, is a colorless liquid with a chloroform-like odor that irritates the eyes and mucous membranes.

Properties (see also Table 1). Bis(2-chloroethyl) ether is miscible with most organic solvents except alkanes. The solubility of the ether in water at 40 °C is 1.1 %, and that of water in the ether is 0.5 %. Azeotropes of bis(2-chloroethyl) ether are listed in Table 8.

Vapor pressure:

t, °C	23.5	49.3	101.5	134.0
p, kPa	0.133	0.667	8.00	26.7

Specific heat capacity (20 – 30 °C): 1.55 kJ kg^{-1} K^{-1}
Viscosity (25 °C): 2.065 mPa · s
Surface tension (25 °C): 41.8 mN/m

Unlike bis(chloromethyl) ether, bis(2-chloroethyl) ether is somewhat resistant to hydrolysis. Heating with solid sodium hydroxide to 200 – 220 °C is necessary to produce 2-chloroethyl vinyl ether [*110-75-8*], and with solid potassium hydroxide to 200 – 240 °C in the presence of ammonia to produce divinyl ether [*109-93-3*] (→ Vinyl Ethers). Reaction of bis(2-chloroethyl) ether with primary aromatic amines leads to *N*-aryl morpholines; reaction with hydrazine leads to 4-aminomorpholine.

Table 8. Azeotropes of bis(2-chloroethyl) ether

Component	Component concentration, wt %	Azeotrope *bp*, °C
Binary azeotropes		
Water	65.5	98
Monoethylene glycol	12.5	170.5
Diethylene glycol	8	174.6
2-Ethylhexanol	10	96 (6.7 kPa)
Diisoamyl ether	61	169.35
Ternary azeotrope		
Water	53	} 97.5
2-Chloroethanol	25	

Production. Bis(2-chloroethyl) ether is obtained as a byproduct in the production of ethylene chlorohydrin (→ Chlorohydrins, Chap. 4.), where it collects in the distillation sump and is purified by vacuum distillation. Bis(chloromethyl) ether is also produced by reaction of ethylene chlorohydrin with sulfuric acid at 90 – 100 °C or by saturation of an aqueous ethylene chlorohydrin solution with ethylene and chlorine below 85 °C.

Typical *specifications* of commercially available bis(2-chloroethyl) ether are as follows: d_{20}^{20}, 1.219 – 1.224; *bp,* 175 – 181 °C; water content, max. 0.1 %; acid content as hydrochloric acid, max. 0.05 %; ethylene dichloride content, max. 1.0 %.

Uses. Bis(2-chloroethyl) ether is a good solvent for many natural and synthetic resins. It dissolves ethyl cellulose, but 10 – 30 % ethanol is also required to dissolve other cellulose esters.

Bis(2-chloroethyl) ether is used in the textile industry in dyeing, washing, and mercerizing solutions. It can be used for the isolation of butadiene from C_4 cuts by extractive distillation. Bis(2-chloroethyl) ether is also employed as an insecticide and an acaricide.

4.4.3. Bis(2-chloroisopropyl) Ether

Bis(2-chloroisopropyl) ether [*39638-32-9*], also known as dichloroisopropyl ether, $(CH_2Cl)(CH_3)CH{-}O{-}CH(CH_3)(CH_2Cl)$, M_r 171.07, is a colorless liquid.

Properties (see also Table 1). Except for its low volatility, the physical and chemical properties of bis(2-chloroisopropyl) ether are very similar to those of bis(2-chloroethyl) ether. Its solubility in water is slightly less: 0.17 % at 20 °C.

Bis(2-chloroisopropyl) ether does not form an azeotrope with propylene chlorohydrin, but does form an azeotrope with 62.6 wt % water, *bp* 98.5 °C.

Production and Uses. Bis(2-chloroisopropyl) ether can be prepared by dehydration of 2-chloroisopropyl alcohol with sulfuric acid.

Typical *specifications* are: Color, 40 APHA max.; d_{20}^{20}, 1.113 – 1.119; boiling range, 180 – 190 °C; water content, max. 0.1 %; acid content as hydrochloric acid, max. 0.01 %.

The uses of bis(2-chloroisopropyl) ether are the same as those of bis(2-chloroethyl) ether.

5. Toxicology

Diethyl Ether. The most important characteristic of diethyl ether is its anesthetic effect (→ Anesthetics, General, Chap. 2.4.). Diethyl ether is rapidly absorbed into the blood when it is inhaled and is equally rapidly exhaled when inhalation is stopped. Very little diethyl ether is absorbed through the skin, but contact with the skin and mucous membranes causes irritation.

A vapor concentration of ca. 10 vol % is necessary to induce anesthesia, whereas a concentration of ca. 5 vol % is required to maintain this state [36]. All the classical stages of narcosis are observed. Prolonged inhalation of over 10 vol % can cause death. In human beings, after intoxication has been overcome, reversible functional changes are observed in the kidneys and liver in the form of an increase in serum glutamic oxaloacetic transaminase and alkaline phosphatase [37].

The acute lethal dose has been found to be 6.4 vol % for rats, 10.6 vol % for dogs, and

7.16 – 19.25 vol % for monkeys [38]. The oral LD_{50} for rats is ca. 2 g/kg [39]. A small proportion of diethyl ether administered to rats is metabolized to CO_2 [40].

MAK:	400 ppm (1200 mg/m^3)
TLV:	400 ppm (1200 mg/m^3)

Diisopropyl Ether. The anesthetic effects of diisopropyl ether are less pronounced than those of diethyl ether; the local irritating effects on the skin and eyes are also milder.

In experiments on monkeys, rats, and guinea pigs, a vapor concentration of 6 vol % in the inhaled air was found to be lethal as a result of respiratory paralysis. A concentration of 3 vol % is tolerated without causing death, but anesthetic effects are then observed. Repeated inhalation (ten times) of diisopropyl ether vapor at a concentration of 3 vol % is tolerated by the animals without any permanent detrimental effects; this result implies that apparently no cumulative toxicity is to be found [41].

The acute oral toxicity (LD_{50}) of diisopropyl ether in adult rats is 16.0 mL/kg [39].

MAK:	500 ppm (2100 mg/m^3)
TLV-TWA:	250 ppm (1050 mg/m^3)
TLV-STEL:	310 ppm (1320 mg/m^3)

Di-*n*-butyl Ether. Dibutyl ether's threshold concentration for producing irritation of the eyes and nose in humans is ca. 200 ppm [42].

In rats, the acute oral toxicity (LD_{50}) is ca. 7.4 mL/kg. The dermal LD_{50} in rabbits is ca. 10 mL/kg. Inhalation of 4000 ppm for 30 min is lethal to rats; thus, the inhalation toxicity of dibutyl ether is higher than that of the other alkane ethers [43].

Bis(chloromethyl) Ether. Halogenated ethers generally exert stronger anesthetic effects than alkyl ethers.

Bis(chloromethyl) ether is known to display carcinogenic properties. Chronic dermal treatment of mice with a dose of 2 mg, three times a week, produces papillomas and squamous epithelium carcinomas within one year [44]. In chronic inhalation tests on mice employing nominal concentrations of ca. 0.005 mg/L, pulmonary tumors were found in almost all of the exposed animals [45]. Similar observations were made in tests on rats, where concentrations of 0.5 ppm were used [46]. Subcutaneous administration induced tumors in mice [47] and rats [48]. In the mice, an increase in pulmonary tumors was once again observed, but in the rats only fibromas and fibrosarcomas were formed at the site of injection.

Evidence also indicates that bis(chloromethyl) ether has carcinogenic effects in humans [49], [50]. Therefore, this ether should neither be manufactured as such nor used as an intermediate. However, if this is unavoidable, special precautions must be taken to prevent any form of contact such as inhalation or skin absorption. The same applies if chloromethylation is performed with formaldehyde and hydrochloric acid, since it has been reported that bis(chloromethyl) ether is then formed within less than 1 min at room temperature [51]. However, subsequent studies have failed to confirm this finding [52].

MAK:	class III A 1 carcinogen
TLV-TWA:	0.001 ppm (0.005 mg/m^3), class A1 a carcinogen

Monochlorodimethyl ether [*107-30-2*] is also classified as carcinogenic in humans (MAK, class III A carcinogen; TLV-TWA, class A 2 carcinogen) because it contains up to 7 % bis(chloromethyl) ether [50]. Monochlorodimethyl ether and bis(chloromethyl) ether can be chemically detoxified [53], [54]. Reaction with methanolic ammonia, secondary amines, sodium phenolate, or sodium methoxide is particularly recommended and destroys more than 99 % of these ethers.

Bis(2-chloroethyl) Ether. The position of the halogen atom is of considerable importance in determining the properties of halogenated ethers. Substitution at the α-position leads to higher reactivity than at the β- or γ-position. This obviously also affects the toxicological properties of these ethers.

In humans, bis(2-chloroethyl) ether vapor concentrations of 550 ppm cause significant irritation of the mucous membranes of the respiratory tract. The threshold value required to evoke a response appears to be between 260 and 100 ppm [55]. Acute oral toxicity is high, the oral

LD_{50} values being 75 mg/kg for rats [56] and 136 mg/kg for mice [57]. Inhalation of 250 ppm for 4 h is lethal to rats [58]. In chronic inhalation studies on rats and guinea pigs, a concentration of 69 ppm is tolerated without producing any symptoms [59].

Studies on mice [44], [60] and rats [61] do not reveal any carcinogenic effects. Mutagenic effects cannot be demonstrated in *Drosophila* [62] or in the Ames test [63]. Alkylation of DNA does not occur [64].

MAK:	10 ppm (60 mg/m^3)
TLV:	not specified

Bis-(2-chloroisopropyl)Ether. Bis(2-chloroiso-propyl) ether is less toxic than bis(2-chloroethyl) ether. In guinea pigs, the oral LD_{50} is ca. 450 mg/kg [57]. When subacute doses were given to rats, a concentration of 200 mg/kg of body mass caused changes in the liver and kidneys [65].

In acute inhalation experiments on rats [65], a vapor concentration of 700 ppm is lethal; lethal effects start to appear at a concentration of 175 ppm. Morphological investigations reveal hepatic cell necrosis and stagnation of blood in the lungs. Irritation of the skin and mucous membranes is low [65].

MAK and TLV values are not specified for bis(2-chloroisopropyl) ether.

6. References

General References

1. "Methoden zur Herstellung und Umwandlung von Ether" in *Houben-Weyl* **IV/3.**
2. "Eigenschaften und Reinigung der wichtigsten organischen Lösemittel" in *Houben-Weyl* **I/2.**
3. A. Weissberger: *Techniques of Organic Chemistry,* 3rd ed., **vol. VII,** Wiley Interscience, New York 1970.
4. F. Korte (ed.): *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart, Academic Press, New York-San Francisco-London 1975.
5. C. Mardsen, S. Mann: *Solvents Guide,* 2nd ed., Cleaver-Hume Press Ltd., London 1963.
6. E. W. Flick (ed.): *Industrial Solvents Handbook,* Noyes Data Corporation, Parke Ridge, N.J. 1985.
7. A. K. Doolittle: *The Technology of Solvents and Plasticizers,* J. Wiley & Sons, New York, Chapman & Hall Ltd., London 1954.
8. Hauptverband der gewerblichen Berufsgenossenschaften: *Richtlinien für die Vermeidung von Zündgefahren infolge elektrostatischer Aufladungen,* Carl Heymanns Verlag, Köln 1980.
9. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., van Nostrand Reinhold Comp., New York-Cincinnati-Toronto-London-Melbourne 1984.

Specific References

10. L. R. Burwell, *Chem. Rev.* **54** (1954) 615 – 685.
11. A. Rieche, *Z. Angew. Chem.* **44** (1931) 896 – 899.
12. R. Neu, *Z. Angew. Chem.* **45** (1932) 519 – 520.
13. A. Rieche, K. Koch, *Ber. Deutsch Chem. Ges.* **75** (1942) 1016 – 1028.
14. V. J. Karnojitzky, *Chim. Ind. (Paris)* **88** (1962) 233.
15. *DAB 7,* p. 293.
16. A. C. Hamstead, *Ind. Eng. Chem.* **56** (1964) no. 6, 37 – 42.
17. J. van Elphen, *Recl. Trav. Chim. Pays-Bas* **49** (1930) 754.
18. A. Williamson, *Justus Liebigs Ann. Chem.* **77** (1851) 37; **81** (1852) 373.
19. J. L. Erickson, W. H. Ashton, *J. Am. Chem. Soc.* **63** (1941) 1769.
20. R. G. Smith, A. Vanterpool, H. J. Kulak, *Can. J. Chem.* **47** (1969) 2015 – 2019.
21. B. Sjöberg, K. Sjöberg, *Acta Chem. Scand.* **26** (1972) 275.
22. G. L. Grady, S. K. Chokshi, *Synthesis* 1972, 483 – 84.
23. W. P. Moeller, S. W. Englund, Tsu Kan Tsui, D. F. Othmer, *Ind. Eng. Chem.* **43** (1951) 711 – 717.
24. G. E. Hall, F. M. Umbertini, *J. Org. Chem.* **15** (1950) 715 – 719.
25. *Houben-Weyl,* **I/2,** p. 811.
26. C. S. Cope, B. F. Dodge, *AIChE J.* **5** (1959) 10 – 16.
27. Europäisches Arzneibuch, Band III, Deutscher Apotheker Verlag Stuttgart, GOVI Verlag GmbH Frankfurt 1978, pp. 173 – 174.
28. Gefahrstoffverordnung vom 26. 08. 1986, Bundesgesetzbl. Teil 1, Nr. 47 vom 05. 09. 1986.
29. R. Kühn, K. Birett, *Merkblätter gefährlicher Arbeitsstoffe,* 22. Erg. Lieferung II/83-P 38-1, Ecomed Verlagsgesellschaft mbH.

30. A. C. Hamstead, *Ind. Eng. Chem.* **56** (1964) no. 6, 37.
31. V. I. Sushchenya, G. P. Solomakka, L. I. Blyakhman, *Zh. Prikl. Kleim (Leningrad)* **51** (1968) 437.
32. I. W. Filippow, S. S. Kagan, M. I. Kondratjewa, *Koks Khim.* **1963,** no. 12, 46;**1965,** no. 14, 2063.
33. E. Yu. Nyarep, V. V. Sheloumov, *Tr. Nauch.-Issled, Inst. Slantser* 1968, no. 17, 108.
34. R. G. Edmonds, G. F. Jenkins, *Chem. Eng. Progr.* **50** (1954) 111.
35. R. Summers, *Chem. Rev.* **55** (1955) 301 – 353.
36. H. L. Price, R. D. Dripps, L. S. Goodman, A. Gilman, eds.: *The Pharmacological Basis of Therapeutics,* 4th ed., Macmillan Publ. Co., New York 1970.
37. B. Dawson et al., *Mayo Clin. Proc.* **41** (1966) 599.
38. W. S. Spector, ed.: *Handbook of Toxicology,* **vol. 1,** Saunders, Philadelphia 1956.
39. E. T. Kimura, D. M. Ebert, P. W. Dodge, *Toxicol. Appl. Pharmacol.* **19** (1971) 699.
40. J. C. Krantz, jr., C. J. Carr: *The Pharmacologic Principles of Medical Practice,* 7th ed., Williams and Wikins, Baltimore 1969, pp. 97 – 112.
41. W. Machle, E. W. Scott, J. Treon, *J. Ind. Hyg. Toxicol.* **21** (1939) 72.
42. L. Silverman, H. F. Schulte, W. M. First, *J. Ind. Hyg. Toxicol.* **28** (1946) 262.
43. H. F. Smyth, jr., C. P. Carpenter, C. W. Weil, U. C. Pozzani, *AMA Arch. Ind. Hyg. Occup. Med.* **10** (1954) 61.
44. B. L. van Duuren, C. Katz, M. Goldschmidt, K. Frenkel, A. Sivak, *J. Nat. Cancer Inst.* **48** (1972) 1431.
45. B. K. J. Leong, H. N. MacFarland, W. H. Reese, jr., *Arch. Environ. Health* **22** (1971) 663.
46. S. Laskin, M. Kuschner, R. T. Drew, V. P. Cappiello, N. Nelson, *Arch. Environm. Health* **23** (1971) 135.
47. J. L. Gargus, W. H. Reese, jr., H. A. Rutter, *Toxicol. Appl. Pharmacol.* **15** (1969) 92.
48. B. L. van Duuren, A. Sivak, B. M. Goldschmidt, C. Katz, S. Melchionne, *J. Nat. Cancer Inst.* **43** (1969) 481.
49. A. M. Thiess, W. Hey, H. Zeller, *Zbl. Arbeitsmed.* **23** (1973) 97.
50. D. Heuschler: *Gesundheitsschädliche Arbeitsstoffe,* Verlag Chemie, Weinheim 1973.
51. *Am. Ind. Hyg. Ass. J.* **33** (1972) 381.
52. G. J. Kallos et al., *Am. Ind. Hyg. Ass. J.* **34** (1973) 469.
53. D. Martinez, *Z. Chem.* **26** (1986) 309 – 315.
54. M. Iovu, V. Radulescu, M. Iqbal, *Rev. Roum. Chim.* **30** (1985) 713 – 717; *Chem. Abstr.* **105** (1986) 114584.
55. H. H. Schrenk, F. A. Patty, W. P. Yant, *Public Health Rep.* **48** (1933) 1389.
56. H. F. Smyth, jr., C. P. Carpenter, *J. Ind. Hyg. Toxicol.* **30** (1948) 63.
57. H. F. Smyth, in W. S. Spector (ed.): *Handbook of Toxicology,* "Unveröffentlichte Untersuchung," vol. I, Saunders, Philadelphia 1956.
58. C. P. Carpenter, H. F. Smyth, jr., U. C. Pozzani, *J. Ind. Hyg. Toxicol.* **31** (1949) 343.
59. C. L. Hake, V. K. Rowe in F. A. Patty (ed.): *Industrial Hygiene and Toxicology,* "Ethers" 2nd rev. ed., **vol. II,** New York 1963, p. 1673.
60. J. R. M. Innes et al., *J. Nat. Cancer Inst.* **42** (1969) 1101.
61. B. Ulland, E. K. Weisburger, J. H. Weisburger, *Toxicol. Appl. Pharmacol.* **25** (1973) 446.
62. C. Auerbach, J. M. Robson, J. G. Carr, *Science* **105** (1947) 243.
63. J. Obermeier, H. Frohberg, unpublished results, Institut für Toxikologie, E. Merck, Darmstadt 1976.
64. F. P. Guengerich, P. S. Mason, W. T. Stoff, T. R. Fox, P. G. Watanabe, *Cancer Res.* **41** (1981) 4391.
65. Biochemical Research Laboratory, Dow Chemical Co., unpublished results.

Ethyl Alcohol → Ethanol

Ethylbenzene

VINCENT A. WELCH, Washington Group Intl., Inc. (formerly The Badger Company), Cambridge, Massachusetts, United States (Chap. 1, 2, 3, 4, 5, 6, 7, 8 and 9)

KEVIN J. FALLON, Washington Group Intl., Inc. (formerly The Badger Company), Cambridge, Massachusetts, United States (Chap. 1, 2, 3, 4, 5, 6, 7, 8 and 9)

HEINZ-PETER GELBKE, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 10)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Production 2
4.1. Alkylation with Non-Zeolite Lewis Acid Catalysts 3
4.2. Vapor-Phase Alkylation over Zeolites 5
4.3. Liquid-Phase Alkylation over Zeolites 7
4.4. Mixed-Phase Zeolite-Based Process . 8
4.5. Separation from Mixed C_8 Streams . 8
5. Environmental Protection 9
6. Quality Specifications 9
7. Handling, Storage, and Transportation 10
8. Uses 10
9. Economic Aspects 10
10. Toxicology 11
11. References 13

1. Introduction

Ethylbenzene [*100-41-4*], phenylethane, $C_6H_5CH_2CH_3$, M_r 106.168, is a single-ring alkylaromatic compound. It is almost exclusively (>99 %) used as an intermediate for the manufacture of styrene monomer [*100-42-5*], $C_6H_5CH{=}CH_2$, one of the most important large-volume commodity chemicals. Styrene production, which uses ethylbenzene as a starting raw material, consumes ca. 50 % of the world's benzene production. Less than 1 % of the ethylbenzene produced is used as a paint solvent or as an intermediate for the production of diethylbenzene and acetophenone.

Commercially, almost all ethylbenzene is produced by alkylating benzene with ethylene. The newest technologies utilize synthetic zeolites installed in fixed-bed reactors to catalyze the alkylation in the liquid phase. Another proven route uses narrower pore synthetic zeolites, also installed in fixed-bed reactors, to effect the alkylation in the vapor phase. A considerable quantity of ethylbenzene is still produced by alkylation with homogeneous aluminum chloride catalyst in the liquid phase, though the recent trend in the industry has been to retrofit such units with zeolite technology. The alkylation of aromatic hydrocarbons with olefins in the presence of aluminum chloride catalyst was first practiced by BALSOHN in 1879 [1]. However, FRIEDEL and CRAFTS pioneered much of the early research on alkylation and aluminum chloride catalysis.

Several facilities built during the 1960s recovered ethylbenzene by fractionation of mixed xylenes produced in refinery catalytic reforming units. This practice has largely been discontinued due to poor economics that result from high energy and investment costs, as well as small economies of scale in comparison to the conventional alkylation routes.

Ethylbenzene was first produced on a commercial scale in the 1930s by Dow Chemical in the United States and by BASF in Germany. The ethylbenzene/styrene industry remained relatively insignificant until World War II. The tremendous demand for synthetic styrene – butadiene rubber (SBR) during the war prompted accelerated technology improvements and tremendous capacity expansion. This considerable wartime effort led to the construction of several large-scale factories, and styrene production quickly became a significant industry. In 1999 world annual production capacity of ethylbenzene reached almost 25×10^6 t. Throughout the 1990s, most of the capacity increase oc-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_035.pub2

curred in the Far Eastern countries other than Japan, where the basic petrochemical industries have undergone considerable development and expansion.

2. Physical Properties

Under normal conditions, ethylbenzene is a clear, colorless liquid with a characteristic aromatic odor. Ethylbenzene is an irritant to the skin and eyes and is moderately toxic by ingestion, inhalation, and skin adsorption. Some physical properties of ethylbenzene follow [2, 3]:

Density	at 15 °C	0.87139 g/cm^3
	at 20 °C	0.8670 g/cm^3
	at 25 °C	0.86262 g/cm^3
mp		−94.949 °C
bp	at 101.3 kPa	136.2° C
Refractive index	at 20 °C	1.49588
	at 25 °C	1.49320
Critical pressure		3609 kPa (36.09 bar)
Critical temperature		344.02 °C
Flash point (closed cup)		15 °C
Autoignition temperature		460 °C
Flammability limit	lower	1.0 %
	upper	6.7 %
Latent heat	fusion	86.3 J/g
	vaporization	335 J/g
Heating value,	gross	42 999 J/g
	net	40 928 J/g
Kinematic viscosity	at 37.8 °C	0.6428×10^{-6} m^2/S
	at 98.9 °C	0.390×10^{-6} m^2/S
Surface tension		28.48 mN/m
Specific heat capacity		
ideal gas, 25 °C		1169 J kg^{-1} K^{-1}
liquid, 25 °C		1752 J kg^{-1} K^{-1}
Acentric factor		0.3026
Critical compressibility		0.263

3. Chemical Properties

The most important commercial reaction of ethylbenzene is its dehydrogenation to styrene. The reaction is carried out at high temperature (600 – 660 °C), usually over a potassium-promoted iron oxide catalyst. Steam is used as a diluent. Commercially, selectivities for styrene range from 90 to 97 mol % with per-pass conversions of 60 – 70 %. Side reactions involve mainly the dealkylation of ethylbenzene to benzene and toluene.

A reaction of increasing commercial importance is the oxidation of ethylbenzene by air to the hydroperoxide $C_6H_5CH(OOH)CH_3$ [*3071-32-7*]. The reaction takes place in the liquid phase without a catalyst. However, because hydroperoxides are unstable, exposure to high temperature must be minimized to reduce the rate of decomposition. The production of byproducts is reduced if the temperature is gradually lowered during the course of the reaction. The hydroperoxide is subsequently treated with propylene to give styrene and propylene oxide as co-products. In 1999 about 15 % of the ethylbenzene produced worldwide was used in the co-production of styrene monomer and propylene oxide.

Like toluene, ethylbenzene can be dealkylated catalytically or thermally to benzene. Ethylbenzene also undergoes other reactions typical of alkylaromatic compounds [4].

4. Production

Alkylation of benzene with ethylene is the source of nearly all ethylbenzene. For several decades most alkylation plants utilized dissolved Lewis acids, in most cases aluminum chloride, to catalyze the reaction in the liquid phase. About 40 % of worldwide ethylbenzene production still utilizes variations on this method. Although the aluminum chloride route generally provides competitive economics, disposal of the resulting waste streams has become increasingly costly. In addition, this route is susceptible to severe corrosion of equipment and piping. Since the early 1980s, technologies using heterogeneous zeolites, operating in the vapor phase and more recently in the liquid phase, have been most often selected for new grass-roots plants. More recently, increasing environmental pressures and improvements in the zeolite processes have provided incentives to a number of operators of aluminum chloride units to retrofit with zeolite technology. About 10^6 t of capacity in several plants were changed to zeolite technology from 1997 to 1999, and as of 2000, more conversions are in the engineering phase. It has been nearly ten years since the last significant grass-roots capacity utilizing aluminum chloride technology came on-stream.

A very minor amount of ethylbenzene is produced by superfractionation of mixed C_8 aro-

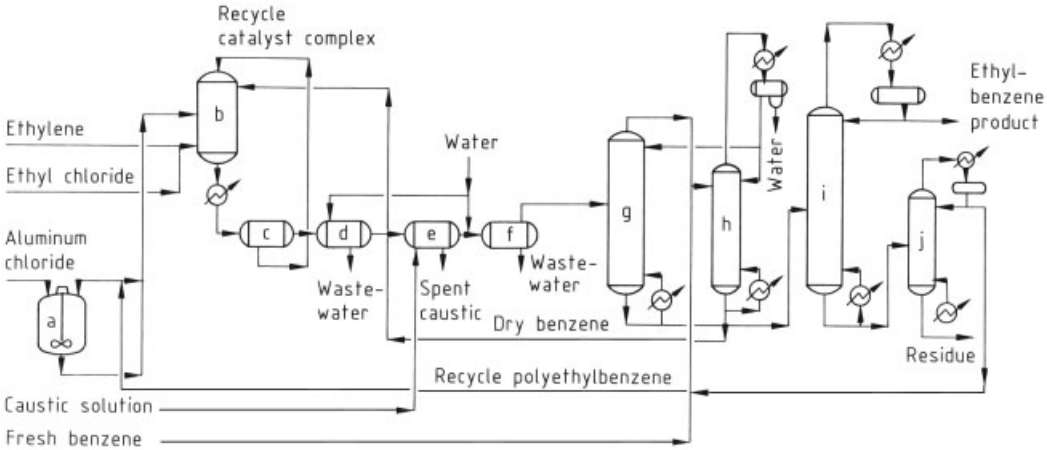

Figure 1. Aluminum chloride process for ethylbenzene production
a) Catalyst mix tank; b) Alkylation reactor; c) Settling tank; d) Acid separator; e) Caustic separator; f) Water separator; g) Benzene recovery column; h) Benzene dehydrator column; i) Ethylbenzene recovery column; j) Polyethylbenzene column

matic streams. Only a small number of plants in the world still utilize this method.

4.1. Alkylation with Non-Zeolite Lewis Acid Catalysts

Liquid-phase aluminum chloride processes were first commercialized in the 1930s. Several companies developed variations of this technology, including Dow Chemical, BASF, Shell Chemical, Monsanto/Lummus, Société Chimiques des Charbonnages, and Union Carbide/Badger. Of the aluminum chloride plants still in operation, the majority of them utilize the Monsanto/Lummus technology, widely considered the most advanced of the aluminum chloride processes. Lummus discontinued licensing this technology with the commercialization of their liquid-phase zeolite process around 1990.

Alkylation of benzene with ethylene is highly exothermic ($\Delta H = -114$ kJ/mol); in the presence of aluminum chloride the reaction is very fast and produces almost stoichiometric yields of ethylbenzene. In addition to $AlCl_3$, a wide range of Lewis acid catalysts, including $AlBr_3$, $FeCl_3$, $ZrCl_4$, and BF_3, have been used. Aluminum chloride processes generally use ethyl chloride or hydrogen chloride as a catalyst promoter. These halide promoters reduce the amount of $AlCl_3$ required. The reaction mechanism has been studied in detail [5].

In the older *conventional $AlCl_3$ process* (see Fig. 1) which is seldom practiced now, three phases are present in the reactor: aromatic liquid, ethylene gas, and a liquid catalyst complex phase (a reddish brown material called red oil). A mixture of catalyst complex, dry benzene, and recycled polyalkylbenzenes is continuously fed to the reactor and agitated to disperse the catalyst complex phase in the aromatic phase. Ethylene and the catalyst promoter are injected into the reaction mixture through spargers, and essentially 100 % of the ethylene is converted. Commercial plants typically operate at ethylene/benzene molar ratios of ca. 0.3 – 0.35. As this ratio is increased, more side reactions, such as transalkylation and isomeric rearrangement, occurr. Further alkylation of ethylbenzene leads to the reversible formation of lower molecular mass polyalkylbenzenes. The loss in yield due to residue formation is minimized by recycling this material to the alkylation reactor. In addition, because the reaction occurs close to thermodynamic equilibrium, the traditional processes use a single reactor to alkylate benzene and transalkylate polyalkylbenzenes.

The reaction temperature is generally limited to 130 °C; a higher temperature rapidly deactivates the catalyst and favors formation of nonaromatics and polyalkylbenzenes, which are preferentially absorbed by the highly acidic catalyst complex and form byproducts. Sufficient pressure is maintained to keep the reactants in the liquid phase. Because the reaction mixture

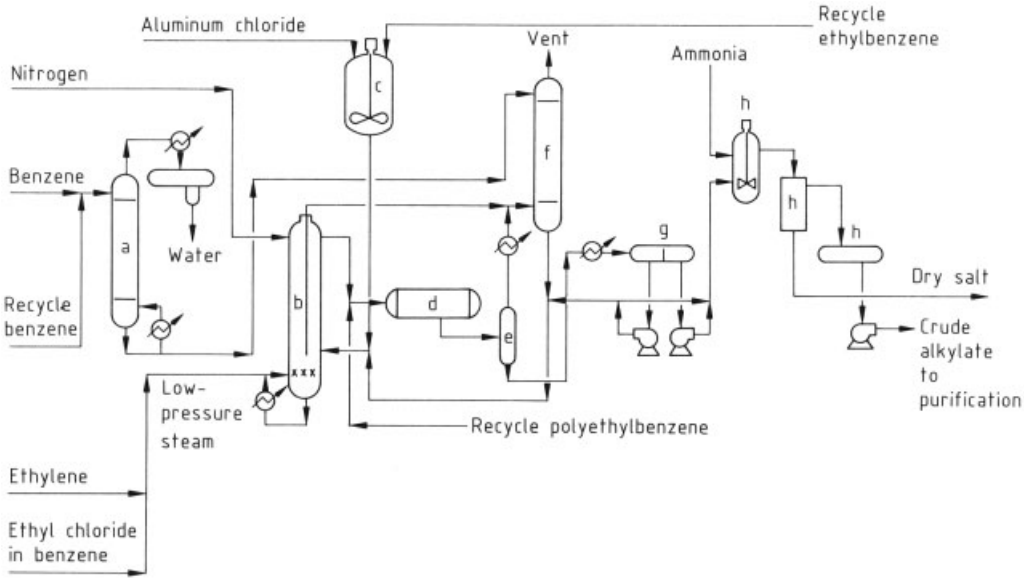

Figure 2. Homogeneous liquid-phase alkylation process for ethylbenzene production
a) Benzene drying column; b) Alkylation reactor; c) Catalyst preparation tank; d) Transalkylator; e) Flash drum; f) Vent gas scrubbing system; g) Decantor; h) Neutralization system

is highly corrosive, the alkylation reactors are lined with brick or glass. High-alloy construction materials are also required for the piping and handling systems.

The liquid reactor effluent is cooled and discharged into a settler, where the heavy catalyst phase is decanted from the organic liquid-phase and recycled. The organic phase is washed with water and caustic to remove dissolved $AlCl_3$ and promoter. The aqueous phase from these treatment steps is first neutralized and then recovered as a saturated aluminum chloride solution and a wet aluminum hydroxide sludge.

Removal of dissolved catalyst from the organic stream has long been a challenge for ethylbenzene producers. CdF Chimie found that more complete recovery of $AlCl_3$ could be achieved by first contacting the organic phase with ammonia instead of sodium hydroxide [6].

Purification of the ethylbenzene product is usually accomplished in a series of three distillation columns. The unconverted benzene is recovered in the first column as an overhead distillate. The second column separates the ethylbenzene product from the heavier polyalkylated components. The bottoms product of the second column is fed to a final column where the recyclable polyalkylbenzenes are stripped from nonrecyclable high molecular mass residue compounds. The residue stream, or flux oil, consisting primarily of polycyclic aromatics, is burned as fuel.

Because the alkylation mixture can tolerate only minor amounts of water, the recycle benzene and fresh benzene must be dried thoroughly prior to entering the reactor. Water not only increases corrosion, but also decreases catalyst activity. Benzene dehydration is accomplished in a separate column.

The improved *Monsanto process* (see Fig. 2) has several advantages over the conventional $AlCl_3$ processes. Over time many conventional $AlCl_3$ plants have been retrofitted with this technology. The most important advantage is a significant reduction in the amount of $AlCl_3$ catalyst used, thus lessening the problem and cost of waste catalyst disposal. Monsanto found that by increasing temperature and by careful control of ethylene addition, the required $AlCl_3$ concentration could be reduced to the solubility limit, thereby eliminating the separate catalyst complex phase [7]. Therefore, the alkylation occurs in a single homogeneous liquid phase instead of in the two liquid phases of earlier processes. Monsanto claimed that a separate catalyst complex phase may actually prevent the attainment

of maximum reactor yields. With a few exceptions, the flow scheme of the Monsanto process is nearly the same as that of more traditional processes. The process is also capable of operating with low-concentration ethylene feed. Typically, the alkylation temperature is maintained at 160 – 180 °C. This higher operating temperature enhances catalyst activity with the additional benefit that the heat of reaction can be recovered as low-pressure steam.

Whereas the traditional process accomplishes alkylation and transalkylation in a single reactor, the homogeneous catalyst system must employ a separate transalkylation reactor. At the lower catalyst concentrations, the recycle of substantial amounts of polyalkylbenzenes terminates the alkylation reaction. Therefore, only dry benzene, ethylene, and catalyst are fed to the alkylation reactor. The recycle polyethylbenzene stream is mixed with the alkylation reactor effluent prior to entering the transalkylation reactor. The transalkylation reactor is operated at much lower temperature than the primary alkylation reactor.

After transalkylation, the reaction products are washed and neutralized to remove residual $AlCl_3$. With the homogeneous process, all of the catalyst remains in solution. The catalyst-free organic reaction mixture is then purified by using the sequence described for the conventional $AlCl_3$ process. As with other $AlCl_3$ processes, the organic residue is used as fuel and the aluminum chloride waste streams are usually sold or sent to treatment facilities.

In 1999 about 40 % of the ethylbenzene manufactured worldwide used aluminum chloride technology. Another technology based on Lewis acids is the so-called *Alkar process*. Developed by UOP and based on boron trifluoride catalyst, this process had modest success in the 1960s, but fell from favor because of high maintenance costs resulting from the severe corrosion caused by small quantities of water. In the developed countries, only one Alkar plant is still in operation.

However, the process produced a high-purity ethylbenzene product and could use dilute ethylene feedstock. If the entry of water into the process was strictly prevented, the corrosion problems associated with aluminum chloride processes were avoided. However, even small amounts of water (< 1 g/kg) hydrolyzed the BF_3 catalyst.

The alkylation reaction took place at high pressure (2.5 – 3.5 MPa) and low temperature (100 – 150 °C). Dried benzene, ethylene, and makeup BF_3 catalyst were fed to the reactor. Typically, ethylene/benzene molar ratios between 0.15 and 0.2 were used. The reactor inlet temperature was controlled by recycling a small portion of the reactor effluent. Transalkylation took place in a separate reactor. Dry benzene, BF_3 catalyst, and recycled polyethylbenzenes were fed to the transalkylation reactor, which operated at higher temperature (180 – 230 °C) than the alkylation reactor. The effluent streams from the two reactors were combined and passed to a benzene recovery column where benzene was separated for recycle to the reactors. Boron trifluoride and light hydrocarbons were taken overhead as a vapor stream from which the BF_3 was recovered for recycle. The bottoms from the benzene recovery column was sent to a product column where ethylbenzene of > 99.9 % purity was taken overhead. A final column served to recover polyethylbenzenes for recycle to the transalkylation reactor.

The Alkar process could operate with an ethylene feed containing as little as 8 – 10 mol % ethylene, enabling a variety of refinery and coke-oven gas streams to be used. However, purification of these streams was necessary to remove components that poison the BF_3 catalyst, e.g., trace amounts of water, sulfur compounds, and oxygenates.

4.2. Vapor-Phase Alkylation over Zeolites

The Mobil – Badger vapor-phase technology was developed in the 1970s around Mobil's ZSM-5 synthetic zeolite and has been available in several different designs. The original first-generation design, commercialized by American Hoechst in 1980, carried out vapor-phase alkylation and transalkylation in a single reactor by recycling polyethylbenzene to the front end of the process, similar to conventional aluminum chloride technology. The newest, so-called third-generation technology performs transalkylation in a separate lower pressure reaction step. The third-generation technology of-

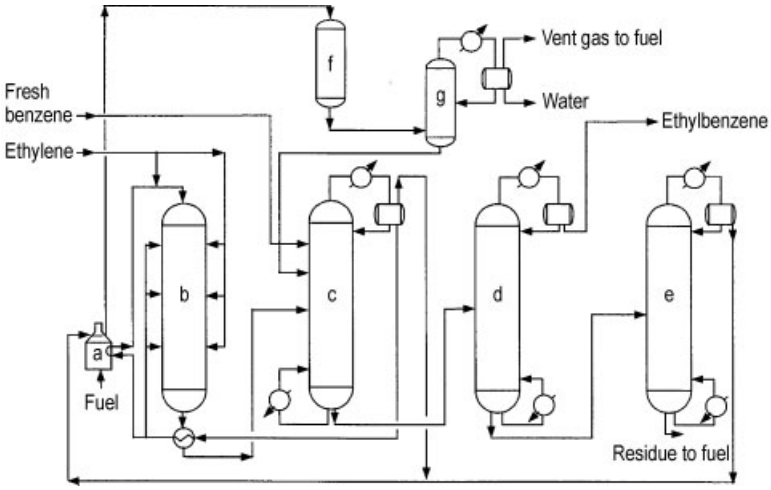

Figure 3. Third-generation Mobil – Badger ethylbenzene process
a)Reactor-feed heater; b) Alkylation reactor; c) Benzene recovery column; d) Ethylbenzene recovery column; e) Polyethylenebenzene recovery column; f) Secondary reactor; g) Stabilizer

fered significant benefits in yield, purity, and capital cost, and was widely used in the 1990s to debottleneck older vapor-phase plants.

The vapor-phase zeolite process is particularly suited to dilute ethylene streams, particularly refinery off-gas from fluid catalytic cracking (FCC) units. Until technologies using zeolites in the liquid phase were commercialized in the 1990s, the vapor-phase zeolite process was the dominant technology of new plants, primarily because it avoided the aqueous waste streams produced from aluminum chloride plants. Mobil – Badger licensed a total of 31 units since 1980, and the technology is still licensed for dilute-ethylene-based plants.

The fixed-bed ZSM-5 catalyst promotes the same overall alkylation chemistry as in the other processes; however, ethylene molecules are adsorbed onto the Brønsted acid sites within the catalyst, which activates the ethylene molecule and allows bonding with benzene molecules to occur. Hence, the range of higher alkylated aromatic byproducts formed by the Mobil – Badger process is somewhat different than those of the Friedel – Crafts processes.

Carbon steel is the primary material of construction; high-alloy materials and brick linings are not required. A diagram of the third-generation design is shown in Figure 3.

The alkylation reactor typically operates in the range of 350 – 450 °C and 1 – 3 MPa. At this temperature, > 99 % of the net process heat input and exothermic heat of reaction can be recovered as steam. The reaction section includes a multibed reactor, a fired heater and heat recovery equipment. The reactor operates with significant overall excess of benzene relative to ethylene.

Slow catalyst deactivation occurs as a result of coke formation and requires periodic regeneration. In situ regeneration takes ca. 36 h and may be necessary after 18 – 24 months of operation, depending on the operating conditions. The catalyst is less sensitive to water, sulfur, and other poisons than the Lewis acid catalysts and zeolites operating in the liquid phase.

The reactor effluent passes to the purification section as a vapor. The benzene from this stream is distilled overhead in the first distillation column and is subsequently recycled to the reactors. The ethylbenzene product is taken as the overhead product from the second column. The bottoms product from this column is sent to the last column where the recyclable alkylbenzenes and polyalkylbenzenes are separated from heavy nonrecyclable residue. The low-viscosity residue stream, consisting mainly of diphenylmethane and diphenylethane, is burned as fuel.

Recyclable higher alkylbenzenes and polyalkylbenzenes are sent to the vapor-phase transalkylator, where they are converted in the presence of excess benzene over zeolite catalyst. Because the transalkylator has lower pressure but higher temperature relative to the alkylator, higher alkylbenzenes are dealkylated while di-

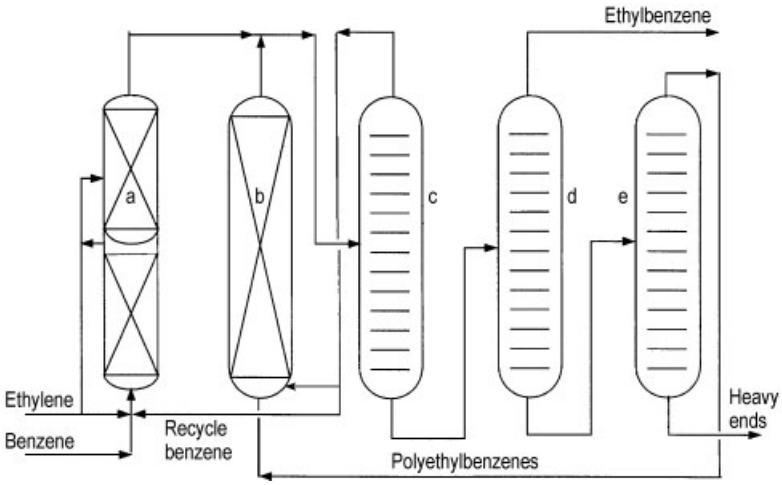

Figure 4. Lummus/UOP ethylbenzene process [9]
a) Alkylation reactor; b) Transalkylation reactor; c) Benzene column; d) Ethylbenzene column; e) Polyethylbenzene column

ethylbenzene is transalkylated to ethylbenzene. The ability to dealkylate higher alkylbenzenes, known to be residue precursors, serves to decrease overall residue production.

The older first- and second-generation processes were similar, the major difference being that the recycle polyethylbenzene stream was recycled to the alkylation reactor. For that reason the process had inferior performance relative to the third-generation design.

Though most often used with polymer-grade ethylene, this process is adaptable to dilute ethylene feedstocks. The process has operated on a mixed C_2 stream provided from a debottlenecked distillation train from a naphtha cracker. Additionally, and perhaps of more interest due to low feedstock cost, is the adaptability to dilute ethylene steams produced from FCC off-gas. Two world-scale Mobil – Badger units operate with FCC off-gas, one since 1991 and one since 1998 [8].

4.3. Liquid-Phase Alkylation over Zeolites

All-liquid phase processes using zeolite catalysts began to appear in commercial operation in 1990, the first plant being operated by Nippon SM in Oita, Japan under license from ABB Lummus Global and Unocal (UOP later became ABB's partner when they acquired Unocal's zeolite business). This process has utilized ultrastable zeolite Y or more recently zeolite beta. Another liquid-phase EB process, EBMax, became available from Mobil – Badger and is based on a Mobil MCM-22 catalyst and first operated at Chiba Styrene Monomer Corp., also in Japan. A total of twelve all-liquid-phase zeolite plants were operating at the end of 1999. Although there are differences between the two available liquid-phase technologies, the latest versions of both lead to lower investment cost and better product quality than was possible with previously available technologies using polymer-grade ethylene. The vapor-phase technology is now typically licensed by Mobil/Badger for dilute-ethylene applications.

These liquid phase processes all utilize wider pore zeolites than ZSM-5, which is necessary to overcome the diffusional limitations of the liquid-phase mechanism. Both licensors usually recommend off-site regeneration of catalyst because of long catalyst cycle times. Off-site regeneration results in less on-site equipment being required and hence in reduced investment.

The flowsheets of the two technologies are quite similar (Figs. 4 and 5). Ethylene is injected into a fixed bed alkylation reactor with multiple stages in the presence of excess benzene. Reactor temperatures vary from process to process, but they must be kept below the critical temperature of benzene (289 °C). Pressures must be high enough to keep the light gases in solution and are on the order of 4 MPa. Excess ben-

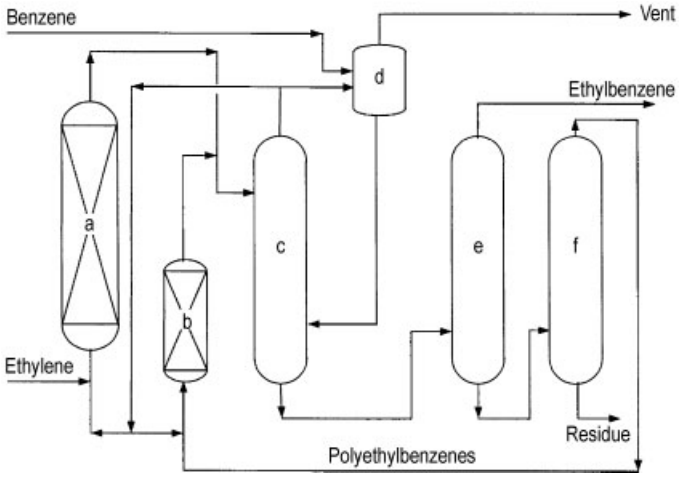

Figure 5. Mobil – Badger EBMax process
a) Alkylation reactor; b) Transalkylation reactor; c) Benzene column; d) Vent-gas column; e) Ethylbenzene column; f) Polyethylbenzene column

zene is distilled overhead from the effluent and is recycled to alkylation. The benzene column bottoms enter the ethylbenzene product column, from which the ethylbenzene is distilled overhead. The bottoms of the ethylbenzene column are fed to the polyethylbenzene column, which separates higher alkylbenzenes and polyethylbenzenes from the residue. This overhead stream is fed to the liquid-phase transalkylation reactor, where it is combined with excess benzene from the benzene column distillate. Effluent from the transalkylator is returned to the distillation train.

Besides benzene, major impurities in the ethylbenzene product can include nonaromatics (naphthenes), toluene, and higher alkylbenzenes. Depending on the operating conditions and technology, these components can originate from the feed benzene or be generated in the reactor. There is a wide variation in distillation column operating conditions, particularly in revamped facilities. There is also a wide variation in excess benzene from plant to plant. These details are held confidentially by the licensors.

4.4. Mixed-Phase Zeolite-Based Process

A mixed-phase ethylbenzene process is offered for license by CDTech, a partnership between ABB Lummus Global and Chemical Research and Licensing, Co. [10, 11] The first unit started in 1994, and as of 1999, three plants were operating. The main distinction of this process is the alkylation reactor, which contains bales of zeolite catalyst in a reactive-distillation column. Ethylene gas and benzene liquid are fed to the reactive-distillation column. Because of its ability to handle the ethylene feed in the vapor phase, the process has been applied to dilute ethylene streams produced from steam cracker distillation trains. It has also been applied to polymer grade ethylene. A flow scheme is shown in Figure 6. An alkylator and benzene stripper operate together as a distillation column. The overhead benzene and unconverted ethylene are in turn fed to a finishing reactor, which also utilizes zeolite. Bottoms from the benzene stripper are fractionated into ethylbenzene product in the overhead product column, and then transalkylatable polyethylbenzenes are distilled from the residue. The polyethylbenzenes are sent to a liquid-phase transalkylation reactor in the presence of excess benzene, and the transalkylation effluent is returned to the fractionation train.

4.5. Separation from Mixed C_8 Streams

Less than 1 % of worldwide ethylbenzene production is recovered from mixed xylene streams, usually in conjunction with xylene production from reformate. Although adsorption processes have been developed, most notably the EBEX process of UOP, ethylbenzene production from

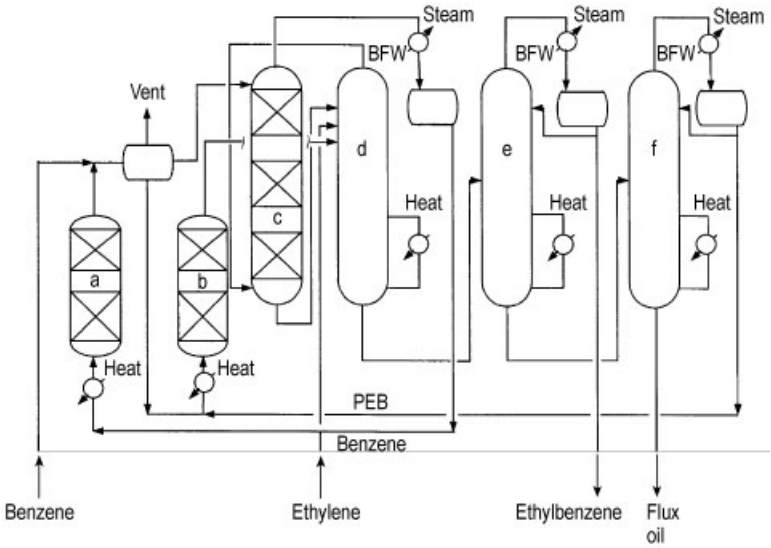

Figure 6. CDTech ethylbenzene process
a) Finishing reactor; b) Transalkylator; c) Alkylator; d) Benzene stripper; e) Ethylbenzene column; f) Polyethylbenzene column
BFW = boiler feed water, PEB = polyethylbenzene

these sources has been mainly performed by distillation. Because of the difficulty of the separation, the process is generally termed superfractionation. It was first undertaken by Cosden Oil & Chemical Company in 1957, using technology jointly developed with the Badger Company. The separation generally requires three distillation columns in series, each having over 100 stages. Several units were built during the 1960s in the United States, Europe, and Japan. However, the increased cost of energy and high capital cost has made this route noncompetitive.

5. Environmental Protection

In the United States ethylbenzene plants must conform to the requirements of U.S. EPA National Emission Standard Organic Air Pollutants from the Synthetic Organic Chemical Industry (40 CFR 63, Subparts F, G, H). Producers can comply by installing collection devices on process vents, improved seals on pumps and valves, and floating roof tanks.

Alkylation plants that use aluminum chloride technology produce an aqueous waste stream from the reactor effluent wash section. In the mid-1970s, plants produced a wet aluminum hydroxide sludge which was deposited in a landfill (Class 1). More modern plants recover a concentrated aluminum chloride solution that has found use in municipal water treatment or industrial floc applications. However, where demand from such applications does not exist, disposal can present a problem. Sometimes hazardous-waste incineration is required.

Studies have shown ethylbenzene to be toxic to aquatic life in relatively low concentration (10 100 mg/kg). Therefore, runoff from spills, fire control, etc. should be diked to prevent it from entering streams or water supplies.

6. Quality Specifications

The product specification on ethylbenzene is set to provide a satisfactory feedstock to the associated styrene unit. Objectionable impurities in the ethylbenzene can be grouped into two categories: those that are detrimental to the operation of the styrene unit and those that affect the purity of the styrene product. Impurities in product ethylbenzene that pose an operating problem in the conventional dehydrogenative styrene process are

1) Halides, which deactivate the dehydrogenation catalyst and contribute to downstream

equipment corrosion. Usually these are chlorides originating from an $AlCl_3$ alkylation section or fluorides from an Alkar unit.

2) Diethylbenzenes, which are dehydrogenated to divinylbenzenes in the styrene reactor section. The divinylbenzenes form insoluble cross-linked polymers in the downstream process equipment. A limit of less than 10 ppm of diethylbenzenes in the ethylbenzene product is usually imposed.

Ethylbenzene contaminants that can affect styrene purity are components having a boiling range between ethylbenzene and styrene. These include xylenes, propylbenzenes, and ethyltoluenes. The levels of cumene, *n*-propylbenzene, ethyltoluenes, and xylenes in the ethylbenzene is controlled to meet the required styrene purity specification.

A typical sales specification for a United States manufacturer is a follows:

Purity	99.5 wt % min.
Benzene	0.05 – 0.3 wt %
Toluene	0.1 – 0.3 wt %
Xylenes	0.2 wt % max.
Nonaromatics	0.05 wt % max.
Propylbenzenes	0.02 wt % max.
Diethylbenzenes	10 mg/kg max.
Total chlorides as Cl^-	1 – 3 mg/kg max.
Total organic sulfur	4 mg/kg max.
Relative density at 15 °C	0.869 – 0.872
APHA Color	15 max.

7. Handling, Storage, and Transportation

Ethylbenzene is a flammable liquid. It is stored and transported in steel containers and is subject to the control of the appropriate regulatory agencies. The U.S. DOT identification number is UNI 175 and the reportable quantity is 454 kg. Details on regulations concerning the transport of ethylbenzene can be found in the CFR or from the DOT Material Transportation Bureau. Other countries have regulations and safety practices similar to those of the United States.

Foam, carbon dioxide, dry chemical, halon, and water (fog pattern) extinguishing media are used in fighting ethylbenzene fires.

Adequate ventilation is necessary in handling and storage areas. The use of NIOSH-approved respirators is recommended at high concentration. Skin contact should be avoided. Chemical gloves and safety glasses should be worn if contact is possible.

Exposure of ethylbenzene to heat, ignition sources, and strong oxidizing agents should be avoided.

8. Uses

Essentially all commercial ethylbenzene production is captively consumed for the manufacture of styrene monomer or in the coproduction of styrene monomer with propylene oxide. Styrene is used in the production of polystyrene and a wide variety of other plastics (→ Styrene).

Of the minor uses, the most significant is in the paint industry as a solvent, which accounts for < 1 % of production capacity. Even smaller volumes go toward the production of acetophenone, diethylbenzene, and ethylanthraquinone.

9. Economic Aspects

Ethylbenzene production is linked directly to the styrene monomer market. A total of 99 % of the ethylbenzene produced worldwide is used to make styrene monomer. Through the 1960s and into the early 1970s annual growth rates for styrene and ethylbenzene averaged 10 %. During this period sustained growth was powered by the expanding polystyrene market. Subsequent growth since the early 1970s has been erratic, with the price of styrene going through four to five year cycles, and several new ethylbenzene/styrene plants were built each time the price started to rise. Since the mid-1980s the average growth in ethylbenzene capacity worldwide has been about 4 – 5 % per year, somewhat higher than the growth of the overall economy. As late as the mid 1980s most Far Eastern ethylbenzene/styrene was produced in Japan, and the emerging Far Eastern economies outside of Japan imported large quantities from the West, particularly the United States and Canada. This pattern changed in the late 1980s as capacity was added in Asia, most notably Korea, which today holds about 10 % of the world's

ethylbenzene capacity. Other Far Eastern countries installing significant new capacity have included Singapore, Thailand, and Taiwan. Over the next five years the expansion is likely to occur in mainland China, as multinational companies form joint ventures with local concerns to build world-scale plants. Today each region of the world is becoming relatively self-sufficient in ethylbenzene/styrene.

About 15 % of ethylbenzene is consumed in co-production of styrene and propylene oxide. These plants offer an economic advantage because of the value of the propylene oxide co-product, which has remained high in part because co-production technology has not been widely available. This situation is beginning to change, and today several producers are operating these plants in the Netherlands, Spain, Japan, Singapore, Korea and the United States. Lower prices for propylene oxide resulting from greater supply could slow the use of ethylbenzene in such plants, which have perhaps three to four times the investment cost of an ethylbenzene dehydrogenation facility.

The 1999 *worldwide* capacity (in 10^3 t/a) by region is as follows:

North America	7700
South America	330
Western Europe	5900
Eastern Europe	1300
Korea/Taiwan/China	3800
Japan	3500
Middle East and Africa	680
Southeast Asia/Australia	500
Total	24 700

Similar to the direct link of ethylbenzene production to that of styrene, ethylbenzene production cost is tied to feedstock cost. Modern processes all have raw material yields > 98 – 99 %. Integration of the ethylbenzene and styrene processes enables efficient energy recovery of heat from the exothermic alkylation reaction. With 90 – 99 % energy recovery of the heat of reaction plus the heat input to the process, production costs are directly related to benzene and ethylene feedstock prices. The U.S. unit sales values of ethylbenzene in $/kg from 1960 to 1997 follow [12]:

1960	0.13	1980	0.51
1965	0.09	1983	0.50
1970	0.09	1987	0.46
1973	0.11	1989	0.55
1974	0.37	1990	0.62
1975	0.20	1993	0.33
1978	0.24	1995	0.55
1979	0.35	1997	0.42

Price increases in 1974 and 1980 reflect the radical change in oil prices experienced during these periods. Price increases in 1989/1990 and 1995 were indicative of the cyclic tightness in supply of ethylbenzene and styrene.

10. Toxicology

The toxicology of ethylbenzene has been reviewed regularly in toxicological textbooks and by various scientific organizations. The most recent reviews, specifically relating to the exposure at the work place, are those of the German "MAK-commission" [17] and of the US TLV-Committee [18] which the reader is referred to for further details.

The acute toxicity in experimental animals is low; oral LD_{50} values in rats range from 3.5 to 4.7 g/kg body weight (b.w.). A 24-h dermal LD_{50} value of approximately 15 g/kg b.w. has been reported. The acute inhalative toxicity also is low. An 8-h exposure in experimental rats produced irritation of the respiratory tract starting at approximately 1000 ppm, and higher concentrations induced unsteadiness, staggering gait, and finally unconsciousness and death at 5000 ppm. Human volunteers exposed for 7.5 h at 25 ppm reported irritation of the mucous membranes, which was much more pronounced at 100 ppm.

Repeated dermal applications of undiluted ethylbenzene led to erythema, edema, and superficial necrosis. Such effects are also expected to occur in humans due to defatting of the skin after repeated exposures. Instillation of undiluted material into the eyes of rabbits produced a slight irritation of the conjunctival membranes but without corneal injury.

The low toxicity of ethylbenzene observed after a single exposure was also seen in repeated dosing studies, in which significant toxicity in experimental animals was produced only at relatively high doses. For example, when rats

and mice were exposed by inhalation over three months no toxicity was observed in rats at 100 ppm and in mice at 500 ppm. Some increases in lung, kidney, and liver weights were measured starting at 250 ppm in rats and 750 ppm in mice, but no histopathological changes were apparent in any of the tissues. When rats and mice were exposed to ethylbenzene by inhalation for two years (the major part of their natural life), no effects were observed in both species at 75 ppm. Some slight changes were found at 250 ppm, but even at 750 ppm there were no effects on overall survival in mice and female rats. Only male rats in the 750 ppm exposure group showed a reduced survival as compared to untreated controls. The reduction in survival was attributed to a rat-specific nephrotoxicity that can not be extrapolated to humans. While exposure to ethylbenzene at 750 ppm for two years had little effect on survival, it did produce an increase in hepatic and pulmonary tumors in mice and an increase in the incidence of renal tumours in male rats (see below).

Various toxicological investigations have considered the potential reproductive toxicity of ethylbenzene. In the three-month inhalation study described above no effects were found on the reproductive tissues of rats and mice at exposures up to 1000 ppm.

Other studies in which pregnant mice, rats, or rabbits were exposed to ethylbenzene have produced divergent results. For example, increases in fetal malformations have been reported when pregnant mice were exposed to 113 ppm ethylbenzene. The study did not find any other effects indicating fetotoxicity, but insufficient information was supplied by the investigators to enable an in-depth evaluation. Rats exposed continuously (24 h/d) up to 540 ppm showed slight maternal toxicity accompanied by fetotoxicity with some retardations and variations of fetal development. At the same concentrations, but with exposures lasting only 6 h/d, no adverse findings were noted on the dams and offspring. In another study in which pregnant rats were exposed to 1000 ppm ethylbenzene for 6 – 7 h/d, maternal toxicity was associated with slight fetotoxicity. In an investigation with pregnant rabbits no fetal effects occurred in animals continuously exposed to ethylbenzene (24 h/d) at approximately 110 ppm. Exposure to approximately 230 ppm was toxic to the dams, with an indication of fetotoxicity. In contrast, no clear effects were noted in pregnant rabbits and their offspring at 1000 ppm with a daily exposure of 7 h/d.

In vitro investigations examining the genotoxic potential of ethylbenzene have produced conflicting results. While studies in bacteria and yeast were negative, some studies with mammalian cells in culture have produced some weakly positive data. In contrast, in vivo investigations in drosophila and in mice [micronucleus test, test for unscheduled DNA synthesis (UDS)] were negative. In summary, while there may be some weak potential for ethylbenzene to interact with the DNA of isolated cells in vitro, the consistent lack of a response in various in vivo test systems indicates that ethylbenzene is not a mutagenic hazard. A limited in vitro mutagenicity database also is available for most of the metabolites of ethylbenzene (see below): 1-phenylethanol, acetophenone, ω-hydroxyacetophenone, phenyl glyoxylic acid, mandelic acid, benzoic acid, hippuric acid, and 2-, 3-, and 4-ethylphenol. Again, although there are some conflicting results the overall weight of evidence is that these metabolites do not exhibit a clear genotoxic potential.

A two-year inhalation carcinogenicity study has been carried out in rats and mice with exposures to ethylbenzene at 0, 250, and 750 ppm. Treatment at 750 ppm produced an increased incidence of kidney tumors in male rats, an increased incidence of lung tumors in male mice, and an increase in the incidence of liver tumors in female mice. At 250 ppm increased tumor rates were not observed. By comparison an oral two-year study with rats and mice in which the animals were exposed 250 or 500 mg/kg/d^{-1} of mixed xylenes (containing 17% ethylbenzene) showed no exposure-related tumor response.

As ethylbenzene does not interact with genetic material in vivo, various studies were carried out to investigate the mode of action responsible for the excess tumor formation seen in the two-year inhalation study. In rats ethylbenzene-induced nephrotoxicity has been shown to exacerbate chronic progressive nephropathy occurring in aging animals, which in association with sustained cell proliferation finally results in an increase in the background level of kidney tumors [19]. Similarly, ethylbenzene at 750 ppm caused an increase in cell proliferation in the lung and liver of mice, and again this prolifera-

tive stimulus may result in excess tumor formation [20]. Further mechanistic analysis is necessary to determine whether the observed cell replication identified in mice is likely to occur in humans and, if so, what exposure level is relevant for establishing a health-based industrial workplace limit. The lack of such data was the basis for the decision of the German MAK Commission to suspend its former workplace exposure limit of 100 ppm.

Ethylbenzene is readily absorbed after inhalative, oral, and dermal exposure. Metabolism proceeds mainly by oxidation of the side chain and to a minor extent by aromatic-ring hydroxylation with less than 5% being excreted in urine as 2- and 4-ethylphenol. Side-chain oxidation mainly leads to 1-phenylethanol, one of the major urinary excretion products in rats. Further oxidative metabolism results in the production of mandelic acid, phenyl glyoxylic acid, and benzoic acid. A second side-chain oxidative process producing 2-phenylethanol and ultimately leading to phenyl acetic acid is a minor pathway. In humans mandelic acid and phenyl glyoxylic acid are the major urinary metabolites, accounting for approximately 80% of the absorbed ethylbenzene.

Several regulatory and scientific commissions have evaluated ethylbenzene and the most important results are summarized below.

In the EU ethylbenzene is classified and labeled with:

- "F" and the risk phrase 11 — "highly flammable"
- "Xn" and the risk phrase R 20 — "harmful by inhalation"

An IARC evaluation of ethylbenzene carried out in 2000 concluded that while there was sufficient evidence for carcinogenicity in experimental animals there was inadequate evidence for cancer in exposed humans. On this basis it was classified as "possibly carcinogenic to humans" (Group 2B) [21].

The German MAK Commission formerly had assigned a MAK value of 100 ppm. In 2001, taking account of the carcinogenicity data, ethylbenzene was classified as a carcinogenic substance into category 3A. In the MAK process carcinogenic chemicals are assigned to category 3A if "the criteria for classification in category 4 or 5 are fulfilled but the database is insufficient for the establishment of a MAK or BAT value". For carcinogens of category 4 or 5 "no significant contribution to human cancer risk is to be expected" ... "provided the MAK and BAT values are observed". In other words, there is in principle an exposure level without an increased carcinogenic risk, but the data available currently do not allow such a specific workplace exposure limit to be assigned. The MAK value for ethylbenzene has therefore been suspended while the Commission awaits the results of further mechanistic investigations to establish an occupational exposure limit (see above) [17].

In 2002 ACGIH recommended a TLV-TWA of 100 ppm and a TLV-STEL of 125 ppm. While such exposure limits relate to inhalative uptake only, the ACGIH has also defined a biological exposure index (BEI) for mandelic acid in urine of 1.5 g/g creatinine [22]. As mandelic acid is a major urinary excretion product after exposure to ethylbenzene, the advantage of a BEI is that uptake by all exposure routes can be integrated. With regard to the carcinogenic effect, ethylbenzene was assigned to group A3: "confirmed animal carcinogen with unknown relevance to humans" [18].

11. References

1. R. H. Boundy, R. F. Boyer (eds.): *Styrene, Its Polymers, Copolymers, and Derivatives*, Reinhold Publ. Co., New York 1952, p. 16.
2. American Petroleum Institute (ed.): *Technical Data Book – Petroleum Refining*, 12th Revision Package to 5th ed., vol. **1**, American Petroleum Institute, Washington, D.C. 1997, pp. 1–78, 1–79, 1–112, 1–113.
3. American Petroleum Institute (ed.): *Technical Data Book – Petroleum Refining*, metric ed., American Petroleum Institute, Washington, D.C. 1981, pp. 1–56, 1–57.
4. *Beilstein* **5**, 776 – 786.
5. G. A. Olah (ed.): *Friedel – Crafts and Related Reactions*, vol. **2**, Wiley-Interscience, New York 1964, Part 1.
6. CdF Chimie, US 4 117 023, 1978 (P. J. Gillet, G. Henrich).
7. Monsanto, US 3 848 012, 1974 (F. Applegath, L. E. DuPree, Jr. A. C. MacFarlane, J. D. Robinson).
8. K. J. Fallon, H. K. H. Wang, C. R. Venkat, "UK Refinery Demonstrates Ethylbenzene Process", *Oil & Gas Journal*, April 17, 1995.

9. *Hydrocarbon Process.* **76** (1997) no. 3 126.
10. *Hydrocarbon Process.* **78** (1999) no. 3 110.
11. CDTech, *Ethylbenzene Technology, Highest Yield Using Zeolite Based Catalyst*, ABB Lummus Global, May 1997.
12. J. Surdyk, K. L. Ring, "Ethylbenzene" in *Chemical Economics Handbook*, SRI International, Menlo Park, CA 1999.
13. R. J. Lewis, Sr., *Sax's Dangerous Properties of Industrial Materials*, 10th ed., John Wiley & Sons, New York, 1999.
14. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices*, ACGIH, Cincinnati, Ohio 1997.
15. DFG (ed.): *MAK- und BAT-Werte-Liste*, VCH, Weinheim 1995.
16. F. W. Mackison, R. S. Stricoff, L. J. Partridge (eds.): "NIOSH/OSHA Occupational Health Guidelines for Chemical Hazards," U.S. Dept. of Health & Human Services (National Institute for Occupational Safety & Health) Publication no. 81 – 123, Washington, D.C., Jan. 1981.
17. H. Greim (ed.): *Gesundheitsschädliche Arbeitsstoffe; Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten (Maximale Arbeitsplatzkonzentrationen)*, Wiley-VCH, Weinheim 2001.
18. ACGIH (ed.): *Documentation of the Threshold Limit Values for Chemical Substances*, 7th ed., Cincinatti 2002.
19. G. C. Hard, *Toxicol. Sci.* **69** (2002) 30 – 41.
20. W. T. Stott, K. A. Johnson, R. Bahnemann, S. J. Day, R. J. McGuirk, *Toxicol. Sci.* **71** (2003) 53 – 66.
21. International Agency for Research on Cancer, *IARC Monographs on the Evaluation of Carcinogenic Risk to Humans*, Vol. 77, 2000, p. 227.
22. ACGIH (ed.): *Documentation of the Biological Exposure Indices*, 7th ed., Cincinatti 2001.

Ethylene

Heinz Zimmermann, Linde AG, Höllriegelskreuth, Germany
Roland Walzl, Linde AG, Höllriegelskreuth, Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Raw Materials 2
5. Production 5
5.1. Ethylene from Pyrolysis of Hydrocarbons 5
5.1.1. Cracking Conditions 5
5.1.2. Heat Requirements for Hydrocarbon Pyrolysis 9
5.1.3. Commercial Cracking Yields 11
5.1.4. Commercial Cracking Furnaces . . 16
5.1.5. Tube Metallurgy 25
5.1.6. Thermal Efficiency of Ethylene Furnaces 25
5.1.7. Coking and Decoking of Furnaces and Quench Coolers 27
5.2. Quenching of Hot Cracked Gas . 28
5.3. Recovery Section 33
5.3.1. Products 33
5.3.2. Cracked Gas Processing 34
5.3.2.1. Front-End Section 34
5.3.2.2. Hydrocarbon Fractionation Section 38
5.3.3. Utilities 47
5.3.4. Process Advances 49
5.3.5. Plant Size 49
5.4. Other Processes and Feedstocks . 49
6. Environmental Protection 51
7. Quality Specifications 52
8. Chemical Analysis 52
9. Storage and Transportation 53
10. Uses and Economic Aspects 53
11. Toxicology and Occupational Health 55
12. References 55

1. Introduction

Ethylene [*74-85-1*], ethene, $H_2C{=}CH_2$, M_r 28.52, is the largest-volume petrochemical produced worldwide. Ethylene, however, has no direct end uses, being used almost exclusively as a chemical building block. It has been recovered from coke-oven gas and other sources in Europe since 1930 [1]. Ethylene emerged as a large-volume intermediate in the 1940s when U.S. oil and chemical companies began separating it from refinery waste gas and producing it from ethane obtained from refinery byproduct streams and from natural gas. Since then, ethylene has almost completely replaced acetylene for many syntheses. Ethylene is produced mainly by thermal cracking of hydrocarbons in the presence of steam, and by recovery from refinery cracked gas.

In 2005 total worldwide ethylene production capacity was 112.9×10^6 t, with an actual demand of ca. 105×10^6 t/a [2], which has growth projections of 3.7 to 4.3 % per year worldwide for the period of 2005–2010 [3 – 5].

2. Physical Properties

Ethylene is a colorless flammable gas with a sweet odor. The physical properties of ethylene are as follows:

mp	−169.15 °C
bp	−103.71 °C
Critical temperature, T_c	9.90 °C
Critical pressure, P_c	5.117 MPa
Critical density	0.21 g/cm^3
Density	
at *bp*	0.57 g/cm^3
at 0 °C	0.34 g/cm^3
Gas density at STP	1.2603 g/L
Density relative to air	0.9686
Molar volume at STP	22.258 L
Surface tension	
at *bp*	16.5 mN/m
at 0 °C	1.1 mN/m

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_045.pub2

Heat of fusion	119.5 kJ/kg
Heat of combustion	47.183 MJ/kg
Heat of vaporization	
at *bp*	488 kJ/kg
at 0 °C	191 kJ/kg
Specific heat	
of liquid at *bp*	2.63 kJ kg^{-1} K^{-1}
of gas at T_c	1.55 kJ kg^{-1} K^{-1}
Enthalpy of formation	52.32 kJ/mol
Entropy	0.220 kJ mol^{-1} K^{-1}
Thermal conductivity	
at 0 °C	177×10^{-4} W m^{-1} K^{-1}
at 100 °C	294×10^{-4} W m^{-1} K^{-1}
at 400 °C	805×10^{-4} W m^{-1} K^{-1}
Viscosity of liquid	
at *mp*	0.73 mPa · s
at *bp*	0.17 mPa · s
at 0 °C	0.07 mPa · s
of gas	
at *mp*	36×10^{-4} mPa · s
at 0 °C	93×10^{-4} mPa · s
at 150 °C	143×10^{-4} mPa · s
Vapor pressure	
at −150 °C	0.002 MPa
at *bp*	0.102 MPa
at −50 °C	1.10 MPa
at 0°	4.27 MPa
Explosive limits in air at 0.1 MPa and 20 °C	
lower (LEL)	2.75 vol % or 34.6 g/cm^3
upper (UEL)	28.6 vol % or 360.1 g/cm^3
Ignition temperature	425–527 °C

3. Chemical Properties

The chemical properties of ethylene result from the carbon–carbon double bond, with a bond length of 0.134 nm and a planar structure. Ethylene is a very reactive intermediate, which can undergo all typical reactions of a short-chain olefin. Due to its reactivity ethylene gained importance as a chemical building block. The complex product mixtures that have to be separated during the production of ethylene are also due to the reactivity of ethylene.

Ethylene can be converted to saturated hydrocarbons, oligomers, polymers, and derivatives thereof. Chemical reactions of ethylene with commercial importance are: addition, alkylation, halogenation, hydroformylation, hydration, oligomerization, oxidation, and polymerization.

The following industrial processes are listed in order of their 2000 worldwide ethylene consumption [6]:

1) Polymerization to low-density polyethylene (LDPE) and linear low-density polyethylene (LLDPE)
2) Polymerization to high-density polyethylene (HDPE)
3) Addition of chlorine to form 1,2-dichloroethane
4) Oxidation to oxirane [*75-21-8*] (ethylene oxide) over a silver catalyst
5) Reaction with benzene to form ethylbenzene [*100-41-4*], which is dehydrogenated to styrene [*100-42-5*]
6) Oxidation to acetaldehyde
7) Hydration to ethanol
8) Reaction with acetic acid and oxygen to form vinyl acetate
9) Other uses, including production of linear alcohols, linear olefins, and ethylchloride [*75-00-3*], and copolymerization with propene to make ethylene–propylene (EP) and ethylene–propylene–diene (EPDM) rubber

4. Raw Materials

Table 1 lists the percentage of ethylene produced worldwide from various feedstocks for 1981 and 1992 [7]. In Western Europe and Japan, over 80 % of ethylene is produced from naphthas — the principal ethylene raw materials.

Table 2. World consumption of feedstocks for ethylene production in 2002 (as percentage of total ethylene produced) [6]

Ethane	29
Propane	8
Butane	3
Naphtha	53
Gas oil	6
Others	1

A shift in feedstocks occurred for the period from 1980 to 1991. In the United States and Europe larger amounts of light feedstocks(liquefied petroleum gas LPG: propane + butanes) and NGL (natural gas liquids: ethane, propane, butane) are used for ethylene production, whereas in Japan more naphtha was used in 1995 compared to 1981. The use of gas oils for ethylene production decreased slightly during the 1980s. Since 2000 a huge amount of new capacities have been installed in the Middle East based on ethane as feedstock, increasing the share of ethane-based production significantly (see Fig. 1). This trend will continue in the coming years

Table 1. Raw materials for ethylene production (as a percentage of total ethylene produced)

Raw material	USA 1979	USA 1995	Western Europe 1981	Western Europe 1995	Japan 1981	Japan 1993
Refinery gas	1	4		3		
LPG, NGL	65	76	4*	17	10*	3*
Naphtha	14	11	80	71	90	97
Gas oil	20	9	16	9	0	0

* Including refinery gas

due to favorable production costs for ethane-based crackers. As a result the production of propene from new crackers will decrease requiring different ways of production of this cracker coproduct.

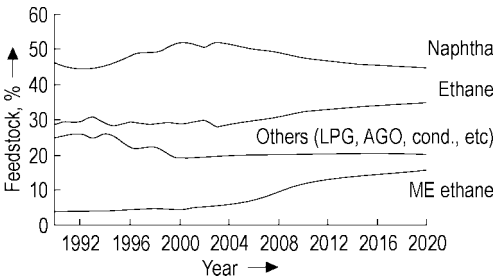

Figure 1. Feedstocks used for ethylene production

Ethane [*74-84-0*] is obtained from wet natural gases and refinery waste gases. It may be cracked alone or as a mixture with propane. Propane [*74-98-6*] is obtained from wet natural gases, natural gasolines, and refinery waste gases. Butanes are obtained from natural gasolines and refinery waste gases.

Naphthas, which are the most important feedstocks for ethylene production, are mixtures of hydrocarbons in the boiling range of 30–200 °C. Processing of light naphthas (boiling range 30–90 °C, full range (FR) naphthas (30–200 °C) and special cuts (C_6–C_8 raffinates) as feedstocks is typical for naphtha crackers.

A natural-cut full-range naphtha contains more than 100 individual components, which can be detected individually by gas chromatography (GC). Depending on the origin naphtha quality can vary over a wide range, which necessitates quality control of the complex feed mixtures. Characterization is typically based on boiling range; density; and content of paraffins (*n*-alkanes), isoalkanes, olefins, naphthenes, and aromatics (PIONA analysis) by carbon number. This characterization can be carried out by GC analysis or by a newly developed infrared method [8]. Full characterization of feedstocks is even more important when production is based on varying feedstocks, e.g. feedstocks of different origins purchased on spot markets.

The quality of a feedstock is depending on the potential to produce the target products (ethylene and propylene). Simple yield correlations for these products can be used to express the quality of a feedstock in a simple figure, the quality factor, which indicates whether yields of the target products are high or low, with aromatic feedstocks being poor and saturated feedstocks being good feedstocks.

Quality characterization factors for naphthas have been developed, which indicate the aromatics content by empirical correlation. Since aromatics contribute little to ethylene yields in naphtha cracking, a rough quality estimate can be made for naphthas with a typical weight ratio of *n*- to isoparaffins of 1–1.1. The *K* factor is defined as [9]:

$$K = \frac{(1.8T_{\mathrm{k}})^{1/3}}{d}$$

where T_{k} is the molal average boiling point in K. Naphthas with a *K* factor of 12 or higher are considered saturated; those below 12 are considered naphthenic or aromatic. The *K* factor does not differentiate between iso- and *n*-alkanes. The U.S. Bureau of Mines Correlation Index (BMCI) [10] can also be used as a rough quality measure of naphthas:

$$\mathrm{BMCI} = 48640/T + 473.7d - 456.8$$

where *T* is the molal average boiling point in K and *d* is the relative density $\mathrm{d}_{15.6}^{15.6}$. A high value of BMCI indicates a highly aromatic naphtha; a low value, a highly saturated naphtha.

Gas condensates are very similar to naphthas in composition with the main difference of having no sharp boiling endpoint. A typical boiling curve of gas condensates is shown in Figure

2. Gas condensates gain importance as raw material for ethylene production because they are cheaper than naphtha.

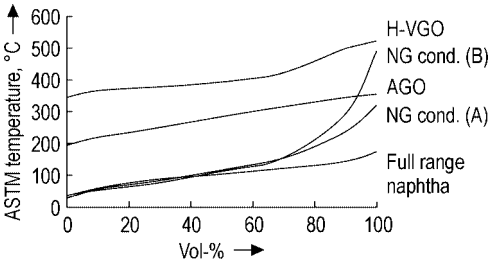

Figure 2. Boiling curves of liquid cracker feedstocks
VGO = Vacuum gas oil; AGO = Atmospheric gas oil

Gas oils are feedstocks that are gaining importance in several areas of the world. Gas oils used for ethylene production are crude oil fractions in the boiling range of 180–350 °C (atmospheric gas oils, AGO) and 350–600 °C (vacuum gas oils, VGO). In contrast to naphtha and lighter gas feeds, these feedstocks can not be characterized by individual components.

Gas chromatography coupled with mass spectrometry (GC–MS) or high performance liquid chromatography (HPLC) allow the analysis of structural groups, i.e., the percentage of paraffins, naphthenes, olefins, monoaromatics, and polyaromatics in the gas oil, and can be used to determine the quality of the hydrocarbon fraction. If this information is used together with data such as hydrogen content, boiling range, refractive index, etc., the quality can be determined quite accurately. A rough estimate of feed quality can be made by using the BMCI or the calculated cetane number of a gas oil. The cetane number, normally used to calculate the performance of diesel fuels, is an excellent quality measure, since it is very sensitive to the *n*-paraffin content, which is one of the key parameters for the ethylene yield. The cetane number CN is calculated as follows [11]:

$$\mathrm{CN} = 12.822 + 0.1164\,\mathrm{CI} + 0.012976\,\mathrm{CI}^2$$

where
CI = 0.9187 $(T_{50}/10)^{1.26687}$ $(n_{\mathrm{D}}^{20}/100)^{1.44227}$,
where T_{50} is the volume average boiling point in°C and n_{D}^{20} the refractive index at 20 °C.

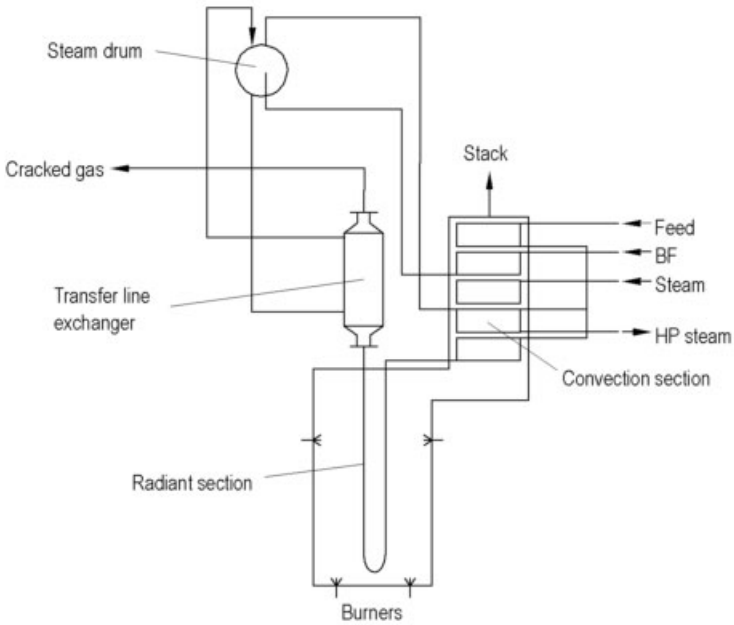

Figure 3. Principal arrangement of a cracking furnace

5. Production

5.1. Ethylene from Pyrolysis of Hydrocarbons

The bulk of the worldwide annual commercial production of ethylene is based on thermal cracking of petroleum hydrocarbons with steam; the process is commonly called pyrolysis or steam cracking. The principal arrangement of such a cracking reactor is shown in Figure 3, a photograph of the furnace section of a modern ethylene plant in Figure 4. The technology has been applied for more than 50 years with permanent improvements in details.

Figure 4. Furnace section of a modern ethylene plant

A hydrocarbon stream is heated by heat exchange against flue gas in the convection section, mixed with steam, and further heated to incipient cracking temperature (500–680 °C, depending on the feedstock). The stream then enters a fired tubular reactor (radiant tube or radiant coil) where, under controlled residence time, temperature profile, and partial pressure, it is heated from 500–650 to 750–875 °C for 0.1–0.5 s. During this short reaction time hydrocarbons in the feedstock are cracked into smaller molecules; ethylene, other olefins, and diolefins are the major products. Since the conversion of saturated hydrocarbons to olefins in the radiant tube is highly endothermic, high energy input rates are needed. The reaction products leaving the radiant tube at 800–850 °C are cooled to 550–650 °C within 0.02–0.1 s to prevent degradation of the highly reactive products by secondary reactions.

The resulting product mixtures, which can vary widely, depending on feedstock and severity of the cracking operation, are then separated into the desired products by using a complex sequence of separation and chemical-treatment steps.

The cooling of the cracked gas in the transferline exchanger is carried out by vaporization of high-pressure boiler feed water (BFW, p = 6–12 MPa), which is separated in the steam drum and subsequently superheated in the convection section to high-pressure superheated steam (HPSS, 6–12 MPa).

5.1.1. Cracking Conditions

Commercial pyrolysis of hydrocarbons to ethylene is performed almost exclusively in fired tubular reactors, as shown schematically in Figure 3. These furnaces can be used for all feedstocks from ethane to gas oil, with a limitation in the end point of the feedstock of 600 °C. Higher boiling materials can not be vaporized under the operating condition of a cracking furnace.

Increasing availability of heavy gas oil fractions, due to a shift in demand to lighter fractions, offers cost advantages for processing heavy feedstocks in some areas of the world. Furthermore, the availability of large quantities of residual oil have led some companies to investigate crude oil and residual oils as ethylene sources. Such feedstocks cannot be cracked in conventional tubular reactors. Various techniques employing fluidized beds, molten salts, recuperators, and high-temperature steam have been investigated, but none of these have attained commercial significance [12]. Recently cracking of gas condensates instead of naphthas has gained importance, as some gas condensates offer yield and cost advantages compared to naphthas. Some gas condensates are contaminated with Hg or As, which need to be removed prior to cracking, as these contaminants could lead to corrosion (Hg) of aluminum heat

exchangers used in the separation section or to poisoning (As) of hydrogenation catalysts [13].

Pyrolysis of hydrocarbons has been studied for years. Much effort has been devoted to mathematical models of pyrolysis reactions for use in designing furnaces and predicting the products obtained from various feedstocks under different furnace conditions. Three major types of model are used: empirical or regression, molecular, and mechanistic models [14].

Today, mechanistic computer models, which are available from various companies, are used for design, optimization and operation of modern olefin plants. Sophisticated regression models are also used, mainly by operators, and offer the advantage of a much lower computer performance requirements than mechanistic models.

The regression models are based on a data set, which can consist of historical data or calculated data. Depending on the quality of the data base the empirical regression models can be of sufficient accuracy for most operating problems, within the range of the data field. These models can be run on small computers and are well suited for process computer control and optimization.

Molecular kinetic models that use only apparent global molecular reactions and thus describe the main products as a function of feedstock consumption have been applied with some success to the pyrolysis of simple compounds such as ethane, propane, and butanes.

For example, cracking of propane can be described as

$$C_3H_8 \rightarrow a\,H_2 + b\,CH_4 + c\,C_2H_4 + d\,C_3H_6 + e\,C_4H_8 + f\,C_{5+}$$

where a, b, c, d, e, f are empirical factors depending on the conversion of propane.

Gross oversimplification is required if these models are applied to complex mixtures such as naphthas or gas oils, but some success has been attained even with these materials.

Advances have been made in mechanistic modeling of pyrolysis, facilitated by the availability of more accurate thermochemical kinetic and pyrolysis data and of high-speed computers. The major breakthrough in this area, however, has been the development of methods to integrate large systems of differential equations [15 – 17].

Mechanistic models need less experimental data and can be extrapolated. The accuracy of these models is very good for most components, but they require permanent tuning of the kinetic parameters, especially for computing the cracked-gas composition for ultrashort residence times. The main application for mechanistic models is the design of cracking furnaces and complete ethylene plants. The accuracy of the models has been improved, driven by the competition between the contractors for ethylene plants. A number of mechanistic models are used today in the ethylene industry, describing the very complex kinetics with hundreds of kinetic equations [18 – 20].

To demonstrate the complexity of the chemical reactions, the cracking of ethane to ethylene is discussed here in detail. A simple reaction equation for ethane cracking is:

$$C_2H_6 \rightarrow C_2H_4 + H_2 \tag{1}$$

If this were the only reaction, the product at 100 % conversion would consist solely of ethylene and hydrogen; at lower conversion, ethylene, hydrogen and ethane would be present. In fact, the cracked gas also contains methane, acetylene, propene, propane, butanes, butenes, benzene, toluene, and heavier components. This reaction (Eq. 1) is clearly not the only reaction occurring.

In the 1930s, the free-radical mechanism for the decomposition of hydrocarbons was established [21]. Although the free-radical treatment does not explain the complete product distribution, even for a compound as simple as ethane, it has been extremely useful. Ethane cracking represents the simplest application of the free-radical mechanism. Ethane is split into two methyl radicals in the chain initiation step (Eq. 2). The methyl radical reacts with an ethane molecule to produce an ethyl radical (Eq. 3), which decomposes to ethylene and a hydrogen atom (Eq. 4). The hydrogen atom reacts with another ethane molecule to give a molecule of hydrogen and a new ethyl radical (Eq. 5).

Initiation

$$C_2H_6 \rightarrow CH_3^{\bullet} + CH_3^{\bullet} \tag{2}$$

Propagation

$$CH_3^{\bullet} + C_2H_6 \rightarrow CH_4 + C_2H_5^{\bullet} \tag{3}$$

$$C_2H_5^{\bullet} \rightarrow C_2H_4 + H^{\bullet} \tag{4}$$

$$H^{\bullet} + C_2H_6 \rightarrow H_2 + C_2H_5^{\bullet} \quad (5)$$

If reactions (4) and (5) proceed uninterrupted, the molecular reaction in Equation (1) results. If only reactions (3)–(5) occurred, the cracked gas would contain traces of methane (Eq. 3) and equimolar quantities of ethylene and hydrogen with unreacted ethane. This is not observed.

Reactions (3) and (4) terminate if either an ethyl radical or a hydrogen atom reacts with another radical or atom by reactions such as:

Termination

$$H^{\bullet} + H^{\bullet} \rightarrow H_2 \quad (6)$$

$$CH_3^{\bullet} + H^{\bullet} \rightarrow CH_4 \quad (7)$$

$$H^{\bullet} + C_2H_5^{\bullet} \rightarrow C_2H_6 \quad (8)$$

$$C_2H_5^{\bullet} + CH_3^{\bullet} \rightarrow C_3H_8 \quad (9)$$

$$C_2H_5^{\bullet} + C_2H_5^{\bullet} \rightarrow C_4H_{10} \quad (10)$$

On termination of chain propagation, new methyl or ethyl radicals or a new hydrogen atom must be generated (Eqs. 2–4) to start a new chain. Thus, every time a new chain is initiated, a molecule of methane is formed (Eq. 3) and a molecule of ethylene is produced (Eq. 4). Other normal and branched-chain alkanes decompose by a similar, but more complex, free-radical mechanism [22]. The number of possible free radicals and reactions increases rapidly as chain length increases.

The free-radical mechanism is generally accepted to explain hydrocarbon pyrolysis at low conversion [21]. As conversion and concentrations of olefins and other products increase, secondary reactions become more significant. Partial pressures of olefins and diolefins increase, favoring condensation reactions to produce cyclodiolefins and aromatics. The cracking of heavy feed, such as naphthas or gas oils, often proceeds far enough to exhaust most of the crackable material in the feedstock.

The reaction scheme with heavier feeds is much more complex than with gaseous feedstocks, due to the fact the hundreds of reactants (feed components) react in parallel and some of those components are formed as reaction products during the reaction. Since the radicals involved are relatively short lived, their concentration in the reaction products is rather low.

Radical decomposition is one of the most important types of reaction and it directly produces ethylene according to the following scheme:

Radical decomposition

$$RCH_2CH_2CH_2^{\bullet} \rightarrow RCH_2^{\bullet} + C_2H_4 \quad (11)$$

This β-scission reaction produces a shorter radical ($RCH_2^{\bullet}$) and ethylene. Radicals normally decompose in the β-position, where the C–C bond is weaker due to electronic effects. Large radicals are more stable than smaller ones and can therefore undergo isomerization.

Radical isomerization

$$RCH_2CH_2CH_2^{\bullet} \rightarrow RCH_2\dot{C}\,HCH_3 \quad (12)$$

The free-radical decomposition of *n*-butane (Eqs. 12–14) results in the molecular equation (Eq. 15):

$$n\text{-}C_4H_{10} + H^{\bullet} \rightarrow n\text{-}C_4H_9^{\bullet} + H_2 \quad (14)$$

$$n\text{-}C_4H_9^{\bullet} \rightarrow C_2H_4 + C_2H_5^{\bullet} \quad (15)$$

$$C_2H_5^{\bullet} \rightarrow C_2H_4 + H^{\bullet} \quad (16)$$

$$n\text{-}C_4H_{10} \rightarrow 2\,C_2H_4 + H_2 \quad (17)$$

Reactions like (1) and (15) are highly endothermic. Reported values of ΔH at 827 °C are +144.53 kJ/mol for Equation (1) and +232.244 kJ/mol for Equation (15).

The mathematical description of these complex systems requires special integration algorithms [23]. Based on the pseudo steady state approximation, the chemical reactions can be integrated and the concentration of all components at each location of the reactor (cracking coil) can be computed [24, 25].

In a generalized and very simplified form the complex kinetics of cracking of hydrocarbons (ethane to gas oil) in steam crackers can be summarized as follows:

Primary reactions			Secondary reactions
Feedstock	→	ethylene	→ C_4 products
/steam		propene	C_5 products
		acetylene	C_6 products
		hydrogen	aromatics
		methane	C_7 products
		etc.	heavier products

The fundamentals of furnace design and the main influences of the different parameters can be understood even with this simplified mechanism:

- Residence time: From the above scheme it is clear that a long residence time favors the secondary reactions, whereas a short residence time increases the yields of the primary products, such as ethylene and propylene.
- Partial pressure: Since most of the secondary products result from reactions in which the number of molecules decreases, increasing the pressure favors the secondary products. One function of the steam present in the system is to reduce the hydrocarbon partial pressure and thus favor the formation of primary products.
- Temperature and temperature profiles: The oligomerization reactions involved in the formation of secondary products are favored by lower temperatures; therefore, special temperature profiles are applied along the cracking coil to avoid long residence times at low temperatures.

Most commercial pyrolysis to produce ethylene is carried out in fired tubular reactors in which the temperature of the reactant increases continuously from the inlet to the outlet. Typical inlet temperatures are 500–680 °C, depending on the material being processed. Typical outlet temperatures are 775–875 °C.

Modern cracking furnaces are designed for rapid heating at the radiant coil inlet, where reaction rate constants are low because of the low temperature. Most of the heat transferred simply raises the reactant from the inlet temperature to the necessary reaction temperature. In the middle of the coil, the rate of temperature rise is lower, but cracking rates are appreciable. In this section,the endothermic reaction absorbs most of the heat transferred to the mixture. At the coil outlet, the rate of temperature rise again increases but never becomes as rapid as at the inlet.

The designers of cracking coils try to optimize the temperature and pressure profiles along the radiant coils to maximize the yield of valuable products yields by special coil design that allows rapid temperature increase in the inlet section and low pressure drops in the outlet section of the cracking coils.

Typical process gas temperature profiles along the radiant coil of modern ethylene furnaces are shown in Figure 5 for ethane, propane, butane, and naphtha cracking.

The quantity of steam used, generally expressed as steam ratio (kilograms of steam per kilogram of hydrocarbon), varies with feedstock, cracking severity, and design of the cracking coil. Typical steam ratios used at a coil outlet pressure of 165–225 kPa (1.65–2.25 bar) for various feedstocks are:

Ethane	0.25–0.35
Propane	0.30–0.40
Naphthas	0.40–0.50
Gas condensates	0.40–0.60
Atmospheric gas oils (cut: 180–350 °C)	0.60–0.70
Hydrocracker bottoms (cut: 350–600 °C)	0.70–0.85

Steam dilution lowers the hydrocarbon partial pressure, thereby enhancing olefin yield. It also reduces the partial pressure of high-boiling, high-molecular-mass aromatics and heavy tarry materials, reducing their tendency to deposit and form coke in the radiant tubes and to foul quench-exchanger tubes. Steam reduces the fouling of the radiant tubes by reacting with deposited coke to form carbon monoxide, carbon dioxide, and hydrogen in a water gas reaction, but as can be seen from the concentration of carbon monoxide in the cracked gas (normally ca. 100 ppm), only to a limited extent.

The sulfur content of the feed is important, since sulfur passivates active Ni sites of the cracking coil material by forming nickel sulfides, which do not catalyze coke gasification, in contrast to nickel metal and nickel oxides. This explains why CO production is high (up to 1 % in the cracked gas) if the feedstock is free of sulfur. To prevent this effect, ca. 20 ppm of sulfur (e.g., as dimethyl sulfide) are added to sulfur-free feedstocks.

Feedstock composition is another important consideration in commercial production of ethylene. Ethylene-plant feedstocks generally contain straight- and branched-chain alkanes, olefins, naphthenes, and aromatics. Ethylene and other olefins are formed primarily from alkanes and naphthenes in the feed. *n*-Alkanes are the preferred component for high ethylene yields, and those containing an even number of carbon atoms give slightly better ethylene yields than odd-numbered ones. *n*-Alkanes also produce propene, whose yield decreases with increasing chain length.

Isoalkanes, in general, produce much smaller yields of ethylene and propene than *n*-alkanes. They give higher yields of hydrogen and meth-

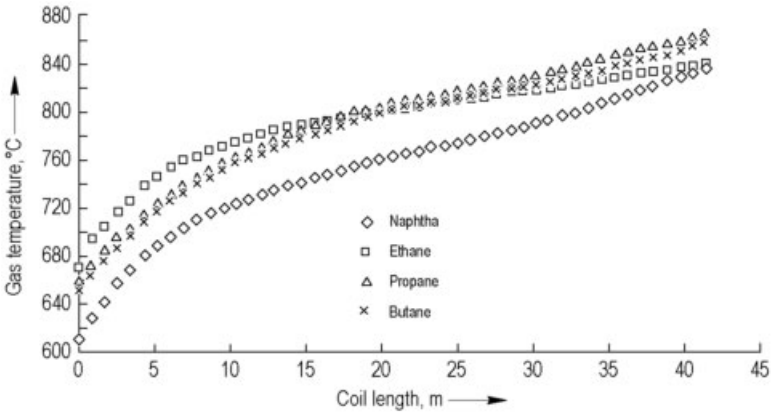

Figure 5. Process gas temperatures along radiant coils

ane, and of C_4 and higher olefins. Isoalkanes tend to produce more propene than ethylene compared to *n*-alkanes.

Olefin and diolefin yields from cyclopentane, methylcyclopentane, and cyclohexane have been reported [22]; methylcyclopentane and cyclohexane give substantial amounts of butadiene.

Simple and condensed ring aromatics produce no ethylene. Benzene is stable under normal cracking conditions. It is formed during cracking and remains unchanged in the feed. Other aromatics yield primarily higher molecular mass components. Aromatics with long side chains, such as dodecylbenzene, yield ethylene and other olefins by side-chain cracking. The aromatic nucleus still, however, generally produces pyrolysis fuel oil and tar.

In summary, maximum ethylene production requires:

– A highly saturated feedstock
– High coil outlet temperature
– Low hydrocarbon partial pressure
– Short residence time in the radiant coil
– Rapid quenching of the cracked gases

These conditions maximize the yield of olefins and minimize the yield of methane and high molecular mass aromatic components.

Cracking conditions and yield structure are often optimized for economic reasons in plant design and operation; even small changes in cracked gas composition influence the overall plant economics dramatically.

The dilution steam ratio is in most cases at the minimum of the range for the specified feedstock, due to the contribution of the dilution steam to the production costs.

5.1.2. Heat Requirements for Hydrocarbon Pyrolysis

Primary hydrocarbon pyrolysis reactions are endothermic; heat requirements are determined by feedstock and cracking conditions. The temperature required to attain a given reaction rate decreases as the carbon chain length in the feedstock increases. Thus, the enthalpy required to reach cracking temperature depends on feedstock and cracking conditions.

Heat requirements for hydrocarbon pyrolysis can be divided into three components:

1) The enthalpy required to heat the feedstock, including the latent heat of vaporization of liquids
2) The endothermic heat of cracking
3) The enthalpy required to heat the cracked gas from the radiant-coil inlet temperature to the radiant-coil outlet temperature

The first is accomplished in the convection section of the furnace; the last two, in the radiant section. Enthalpies (1) and (3) can be calculated from standard heat-capacity data; they represent 25–60 % of total process heat input in the furnace as the feedstock changes from ethane to gas oils [22].

A typical heat balance for a modern naphtha furnace including the transfer-line exchanger (TLE) is shown in Figure 6.

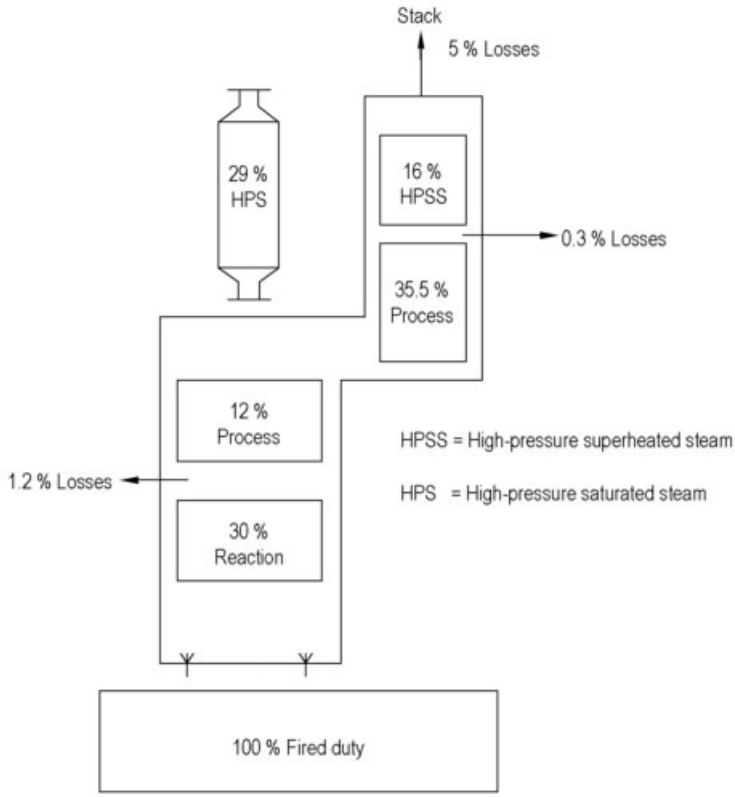

Figure 6. Heat balance of a naphtha cracking furnace

The endothermic heat of cracking is most simply calculated from heats of formation by the following equation:

$$\Delta H_r = \Delta H_p - \Delta H_f$$

where ΔH_r = endothermic heat of cracking, kJ/kg; ΔH_p = heat of formation of the cracked products, kJ/kg; and ΔH_f = heat of formation of the feed, kJ/kg.

Heats of formation for some alkanes, olefins, naphthenes, and aromatics are as follows (in kJ/kg) [26]:

Hydrogen	0.0
Methane	−5646.3
Ethane	−3548.1
Ethylene	1345.3
Acetylene	8540.7
Propane	−2949.9
Propene	−17.9
Methylacetylene	4317.1
Propadiene	4496.8
n-Butane	−2694.4
Isobutane	−2827.5
1-Butene	−459.4
cis-2-Butene	−649.5
trans-2-Butene	−668.2
Isobutene	−746.5
Ethylacetylene	2720.4
1,3-Butadiene	1721.8
Benzene	783.6
Toluene	243.4
Cyclopentane	−1611.3
Methylcyclopentane	−1722.6
Cyclohexane	−1846.4
Methylcyclohexane	−1890.3

As the alkane chain length increases, the heat of formation approaches 1800 kJ/kg asymptotically. Ethane, the most refractory alkane besides methane, has the most endothermic heat of cracking, +4893 kJ/kg. The corresponding figure for propane is +4295 kJ/kg and for a high-molecular-mass *n*-alkane, ca. 1364 kJ/kg.

Endothermic heat of cracking increases with degree of saturation of the feed; as conversion increases, secondary condensation products in-

crease. These products have less negative heats of formation than alkanes, or positive heats of formation.

5.1.3. Commercial Cracking Yields

Determining the complete yield patterns (gas and liquid products) for a commercial cracking furnace requires a standardized procedure, especially if the feedstock is changed from ethane to naphtha and gas oil. A homogeneous sample is withdrawn from the furnace effluent, quenched, and separated into gaseous and liquid fractions. These are analyzed separately and then combined to give a complete cracked-gas analysis.

The material balance is obtained by matching the hydrogen, carbon, and sulfur content of the feedstock with that of the cracked gas. This type of analysis is time-consuming. In a faster system, gas–liquid separation is eliminated [27] or only a gas-phase analysis is taken. Due to many sources of errors, series of samples have to be taken and analyzed. Statistical methods are used to check the quality of the results.

For gas feedstocks, conversion is used as a measure of the severity of cracking:

$$\text{Conversion} = (C_{n,\text{in}} - C_{n,\text{out}}/C_{n,\text{in}})$$

where $C_{n,\text{in}}$ = quantity of component n at inlet in kg/h and $C_{n,\text{out}}$ = quantity of component n at outlet in kg/h.

For liquid feedstocks the severity is defined as the ratio of propylene to ethylene (on weight basis) P/E or as ratio of methane to propylene M/P. Both ratios P/E and M/P are a linear function of temperature over a wide range and can be measured with much higher accuracy than the apparent bulk coil outlet temperature of a cracking furnace. The use of P/E and M/P ratios for severity definition is accepted in industry.

Ethane. Ethane is cracked commercially in all types of furnaces, from long residence time to short residence time. However, since ethane is a very stable paraffin, cracking temperatures at the coil inlet typically are higher than for other hydrocarbons. For ethane the coil inlet temperatures are in the range of 650–680 °C. Typical conversion ranges of commercial furnaces are 60–75 %, with 67–70% conversion being a typical figure for modern designs.

Table 3 lists yields for ethane cracking at various residence times. There is a slight increase in yield at short residence times. However, as the run length of short-residence-time ethane furnaces is low (Fig. 7) and yield improvement is moderate, application of short-residence-time furnaces for ethane cracking is not feasible in most cases. In certain economic scenarios with limited quantity of feedstock the application of short-residence-time coils can make sense, because additional ethylene is produced in a short-residence-time coil if a limited amount of feedstock is available. In such scenarios residence times in the range of 0.25 s are applied in order to give the advantage of high yield and sufficient run length.

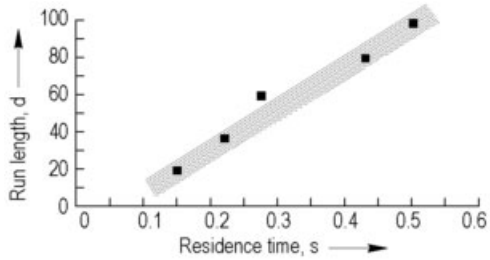

Figure 7. Run length versus residence time for ethane cracking

In naphtha plants the ethane contained in the cracked gas is separated and recycled to a segregated ethane-cracking furnace, in which cracking is carried out under typical ethane-cracking conditions. In case the capacity of the ethane furnace is not sufficient to crack all the ethane, co-cracking of ethane with naphtha is performed, with ethane conversions of 40–50 %. Due to the build up of large recycle streams in such operations, co-cracking should be avoided wherever possible.

Propane is cracked in all types of furnaces at typical conversions of 90–93 %. The application of short-residence-time furnaces for propane cracking has to be evaluated carefully, due to the reduction in run length between successive decokings. In many naphtha plants with propane recycle, propane is co-cracked together with naphtha, without a dramatic drop in propane conversion.

Table 3. Yields from ethane cracking with various residence times

Conversion, kg/kg		65.01	64.97	65.01	65.01
Steam dilution, kg/kg		0.3	0.3	0.3	0.3
Residence time, s		0.4607	0.3451	0.186	0.1133
H_2	wt %	4.04	4.05	4.09	4.12
CO	wt %	0.04	0.04	0.03	0.03
CO_2	wt %	0.01	0.01	0.01	0.01
H_2S	wt %	0.01	0.01	0.01	0.01
CH_4	wt %	3.75	3.52	3.19	2.84
C_2H_2	wt %	0.44	0.47	0.54	0.75
C_2H_4	wt %	51.88	52.31	52.85	53.43
C_2H_6	wt %	34.99	35.03	34.99	34.99
C_3H_4	wt %	0.02	0.02	0.02	0.02
C_3H_6	wt %	1.22	1.13	1.06	0.97
C_3H_8	wt %	0.12	0.12	0.12	0.13
C_4H_4	wt %	0.05	0.05	0.05	0.06
C_4H_6	wt %	1.80	1.80	1.79	1.65
C_4H_8	wt %	0.19	0.19	0.18	0.16
C_4H_{10}	wt %	0.21	0.21	0.21	0.22
Benzene	wt %	0.55	0.47	0.38	0.26
Toluene	wt %	0.08	0.07	0.06	0.04
Xylenes	wt %	0.00	0.00	0.00	0.00
Ethylbenzene	wt %	0.01	0.00	0.00	0.00
Styrene	wt %	0.03	0.02	0.02	0.01
Pyrolysis gasoline	wt %	0.35	0.32	0.29	0.24
Pyrolysis fuel oil	wt %	0.21	0.16	0.11	0.06
Sum	**wt %**	100.00	100.00	100.00	100.00

Table 4. Yields from propane cracking with various residence times

Conversion, kg/kg		90.020	90.035	89.926	89.983
Steam dilution, kg/kg		0.3	0.3	0.3	0.3
Residence time, s		0.4450	0.3337	0.1761	0.1099
H_2	wt %	1.51	1.55	1.61	1.68
CO	wt %	0.04	0.04	0.03	0.04
CO_2	wt %	0.01	0.01	0.01	0.01
H_2S	wt %	0.01	0.01	0.01	0.01
CH_4	wt %	23.43	23.27	22.82	22.40
C_2H_2	wt %	0.46	0.51	0.59	0.82
C_2H_4	wt %	37.15	37.51	38.05	38.59
C_2H_6	wt %	3.06	2.80	2.37	1.96
C_3H_4	wt %	0.52	0.57	0.65	0.89
C_3H_6	wt %	14.81	14.82	15.01	15.27
C_3H_8	wt %	9.97	9.96	10.07	10.01
C_4H_4	wt %	0.08	0.08	0.09	0.11
C_4H_6	wt %	2.85	2.9	2.98	2.99
C_4H_8	wt %	1.00	1	1.02	1.09
C_4H_{10}	wt %	0.04	0.04	0.05	0.05
Benzene	wt %	2.15	2.12	2.02	1.80
Toluene	wt %	0.43	0.4	0.36	0.28
Xylenes	wt %	0.05	0.05	0.04	0.03
Ethylbenzene	wt %	0.01	0.01	0.01	0.00
Styrene	wt %	0.21	0.2	0.18	0.15
Pyrolysis gasoline	wt %	1.27	1.26	1.27	1.24
Pyrolysis fuel oil	wt %	0.94	0.89	0.76	0.58
Sum	**wt %**	100.00	100.00	100.00	100.00

Typical yields for propane cracking for various residence times are listed in Table 4. For very mild propane cracking conditions (70 % conversion) yields of propylene show a maximum at 18–19 wt % based on propane feed.

Butane. All types of furnaces can be used for butane cracking, which typically is performed at 94–96 % conversion. Yields for *n*-butane and isobutane differ widely. For *n*-butane, ethylene yields in the range of 40 % are typical (Table 5), whereas for isobutane only 15 wt % ethylene yield is achieved. Due to the poor ethylene yield from isobutane, other applications such as dehydrogenation, oligomerization, and alkylation seem to be more feasible than cracking isobutane to produce ethylene. However, mild cracking at 70 % conversion and high pressures has been claimed to yield isobutene as primary product [28].

Naphtha. Naphtha, the refinery hydrocarbon fraction with the boiling range of 35–180 °C, can vary in composition and boiling range, depending on source and refinery conditions. Today naphtha cuts from 35 to 90 °C (light naphtha), 90–180 °C (heavy naphtha) and 35–180 °C (full-range naphtha) are processed. In Europe and Asia/ Pacific naphtha is the primary raw material for ethylene cracking, with full range naphthas being processed most.

The properties of the naphtha feedstock used to generate the yield patterns given in Tables 6, 7, and 8 are as follows:

Density at 20 °C	0.692 g/mL
S content	55 mg/kg
H content	15.17 wt %
C content	84.80 wt %
Average molar mass	92 g/mol
Boiling curve (ASTM D 86)	
initial boiling point	26 °C
5 vol %	49 °C
10 vol %	53 °C
20 vol %	58 °C
30 vol %	64 °C
40 vol %	70 °C
50 vol %	77 °C
60 vol %	86 °C
70 vol %	99 °C
80 vol %	116 °C
90 vol %	138 °C
95 vol %	152 °C
final boiling point	183 °C
PIONA analysis	
n-Paraffins	36.13 wt %
Isoparaffins	36.62 wt %
Olefins	0.21 wt %
Napthenes	21.06 wt %
Aromatics	5.98 wt %

Table 5. Yields from butane cracking with various residence times

Conversion, kg/kg		93.95	93.43	93.44	93.92
Steam dilution, kg/kg		0.35	0.35	0.35	0.35
Residence time, s		0.4775	0.3440	0.1781	0.1117
H_2	wt %	1.05	1.09	1.16	1.22
CO	wt %	0.04	0.04	0.03	0.04
CO_2	wt %	0.01	0.01	0.01	0.01
H_2S	wt %	0.01	0.01	0.01	0.01
CH_4	wt %	20.48	20.29	19.85	19.33
C_2H_2	wt %	0.37	0.42	0.51	0.71
C_2H_4	wt %	35.03	35.81	37.25	38.36
C_2H_6	wt %	4.57	4.16	3.47	3.02
C_3H_4	wt %	0.87	0.93	1.04	1.32
C_3H_6	wt %	17..37	17.24	17.03	16.99
C_3H_8	wt %	0.37	0.35	0.33	0.33
C_4H_4	wt %	0.09	0.1	0.11	0.13
C_4H_6	wt %	3.98	4.08	4.27	4.38
C_4H_8	wt %	2.85	2.83	2.78	2.85
C_4H_{10}	wt %	6.05	6.07	6.06	6.08
Benzene	wt %	2.78	2.69	2.52	2.18
Toluene	wt %	0.83	0.77	0.66	0.49
Xylenes	wt %	0.12	0.11	0.09	0.06
Ethylbenzene	wt %	0.01	0.01	0.01	0.01
Styrene	wt %	0.25	0.24	0.22	0.18
Pyrolysis gasoline	wt %	1.85	1.81	1.80	1.72
Pyrolysis fuel oil	wt %	1.02	0.94	0.79	0.58
Sum	**wt %**	100.00	100.00	100.00	100.00

Table 6. Yields from low-severity naphtha cracking

P/E * , kg/kg		0.65	0.65	0.65	0.65
Steam dilution, kg/kg		0.4	0.4	0.4	0.4
Residence time, s		0.4836	0.3526	0.1784	0.1096
H_2	wt %	0.82	0.83	0.85	0.86
CO	wt %	0.03	0.02	0.02	0.02
CO_2	wt %	0.00	0.00	0.00	0.00
H_2S	wt %	0.00	0.00	0.00	0.00
CH_4	wt %	13.56	13.2	12.44	11.83
C_2H_2	wt %	0.32	0.34	0.36	0.45
C_2H_4	wt %	25.23	25.6	26.08	26.43
C_2H_6	wt %	4.76	4.36	3.62	3.16
C_3H_4	wt %	0.59	0.63	0.70	0.92
C_3H_6	wt %	16.42	16.63	16.97	17.18
C_3H_8	wt %	0.68	0.65	0.61	0.60
C_4H_4	wt %	0.09	0.09	0.10	0.12
C_4H_6	wt %	4.63	4.86	5.27	5.68
C_4H_8	wt %	5.76	5.94	6.34	6.69
C_4H_{10}	wt %	0.72	0.75	0.80	0.85
Benzene	wt %	6.37	5.98	5.18	4.47
Toluene	wt %	3.08	2.9	2.60	2.26
Xylenes	wt %	1.17	1.15	1.09	0.96
Ethylbenzene	wt %	0.79	0.8	0.83	0.85
Styrene	wt %	1.08	1.01	0.87	0.77
Pyrolysis gasoline	wt %	11.25	11.81	13.19	14.09
Pyrolysis fuel oil	wt %	2.65	2.45	2.08	1.81
Sum	**wt %**	100.00	100.00	100.00	100.00

* Propene to ethylene ratio, a measure of severity.

Table 7. Yields from medium-severity naphtha cracking

P/E * , kg/kg		0.55	0.55	0.55	0.55
Steam dilution, kg/kg		0.45	0.45	0.45	0.45
Residence time, s		0.4840	0.3572	0.1828	0.1132
H_2	wt %	0.90	0.92	0.94	0.96
CO	wt %	0.04	0.04	0.03	0.04
CO_2	wt %	0.00	0.00	0.00	0.00
H_2S	wt %	0.00	0.00	0.00	0.00
CH_4	wt %	15.23	14.91	14.31	13.82
C_2H_2	wt %	0.44	0.47	0.53	0.71
C_2H_4	wt %	27.95	28.45	29.24	29.87
C_2H_6	wt %	4.59	4.23	3.57	3.14
C_3H_4	wt %	0.71	0.76	0.87	1.21
C_3H_6	wt %	15.38	15.64	16.09	16.43
C_3H_8	wt %	0.53	0.51	0.48	0.47
C_4H_4	wt %	0.12	0.13	0.16	0.20
C_4H_6	wt %	4.54	4.79	5.28	5.79
C_4H_8	wt %	4.41	4.52	4.75	4.95
C_4H_{10}	wt %	0.42	0.44	0.47	0.49
Benzene	wt %	7.45	7.12	6.5	5.85
Toluene	wt %	3.26	3.10	2.82	2.46
Xylenes	wt %	1.18	1.16	1.12	0.97
Ethylbenzene	wt %	0.62	0.62	0.64	0.63
Styrene	wt %	1.21	1.14	1.00	0.88
Pyrolysis gasoline	wt %	7.70	7.96	8.52	8.79
Pyrolysis fuel oil	wt %	3.32	3.09	2.68	2.34
Sum	**wt %**	100.00	100.00	100.00	100.00

* Propene to ethylene ratio, a measure of severity.

Table 8. Yields from high-severity naphtha cracking

P/E * , kg/kg		0.45	0.45	0.45	0.45
Steam dilution, kg/kg		0.5	0.5	0.5	0.5
Residence time, s		0.4930	0.3640	0.1897	0.1170
H_2	wt %	0.9	1.00	1.03	1.06
CO	wt %	0.06	0.06	0.05	0.06
CO_2	wt %	0.00	0.00	0.00	0.00
H_2S	wt %	0.00	0.00	0.00	0.00
CH_4	wt %	16.9	16.61	16.10	15.74
C_2H_2	wt %	0.64	0.69	0.80	1.14
C_2H_4	wt %	30.25	30.81	31.74	32.38
C_2H_6	wt %	4.32	4.00	3.43	3.00
C_3H_4	wt %	0.81	0.88	1.02	1.48
C_3H_6	wt %	13.63	13.88	14.29	14.55
C_3H_8	wt %	0.37	0.35	0.33	0.32
C_4H_4	wt %	0.18	0.2	0.24	0.32
C_4H_6	wt %	4.20	4.45	4.90	5.39
C_4H_8	wt %	3.11	3.17	3.26	3.28
C_4H_{10}	wt %	0.17	0.18	0.19	0.18
Benzene	wt %	8.32	8.06	7.63	7.15
Toluene	wt %	3.40	3.26	3.00	2.63
Xylenes	wt %	1.14	1.12	1.09	0.95
Ethylbenzene	wt %	0.44	0.44	0.44	0.40
Styrene	wt %	1.44	1.37	1.24	1.12
Pyrolysis gasoline	wt %	5.25	5.35	5.55	5.57
Pyrolysis fuel oil	wt %	4.38	4.12	3.67	3.28
Sum	**wt %**	100.00	100.00	100.00	100.00

* Propene to ethylene ratio, a measure of severity.

Table 9. Yields for AGO and HCR cracking

		AGO	**HCR**
P/E * , kg/kg		0.54	0.53
Steam dilution, kg/kg		0.8	1
Residence time, s		0.32	0.3
H_2	wt %	0.71	0.68
CO	wt %	0.01	0.01
CO_2	wt %	0.01	0.01
H_2S	wt %	0.00	0.00
CH_4	wt %	10.58	11.08
C_2H_2	wt %	0.38	0.60
C_2H_4	wt %	25.93	25.98
C_2H_6	wt %	2.82	2.81
C_3H_4	wt %	0.63	0.97
C_3H_6	wt %	14.07	16.02
C_3H_8	wt %	0.37	0.42
C_4H_4	wt %	0.12	0.04
C_4H_6	wt %	5.73	8.03
C_4H_8	wt %	3.61	4.14
C_4H_{10}	wt %	0.05	0.06
Benzene	wt %	5.44	5.64
Toluene	wt %	3.49	2.73
Xylenes	wt %	0.92	0.46
Ethylbenzene	wt %	0.37	0.34
Styrene	wt %	1.18	1.12
Pyrolysis gasoline	wt %	7.28	6.91
Pyrolysis fuel oil	wt %	16.30	7.95
Sum	**wt %**	100	100

* Propene to ethylene ratio, a measure of severity.

The yield increases with decreasing residence time. The maximum benefit of a selective cracking coil is achieved in high-severity cracking, in which a 5 % feedstock saving can be achieved by using short residence times. However, careful analysis is required since, similar to gas cracking, shorter run lengths are typical for short-residence-time cracking. Maximum annual furnace productivity is achieved with a combination of short residence time and long run length.

Gas Oil (AGO; Boiling Range 180–350 °C). Compositions and boiling ranges of gas oils vary greatly with source and refinery conditions. Saturated gas oils may have BMCI values as low as 20, whereas aromatic gas-oil BMCI values can exceed 45. Typical yield distributions are shown in Table 9 for an AGO with properties as listed in Table 10.

Gas oils produce much more C_{5+} material than the above feedstocks because they contain large amounts of condensed polynuclear aromatics. These components resist cracking and may remain unchanged or condense to higher molecular mass materials.

Table 10. Properties of AGO and HCR used for yield calculations in Table 5

	AGO	HCR
Density at 20 °C, g/mL	0.8174	0.8280
C, wt %	86.20	85.28
H, wt %	13.64	14.34
S, wt %	0.18	0.003
BMCI	24.8	9.37
Paraffins + naphthalenes, wt %	80.38	89.4
Monoaromatics, wt %	12.45	7.3
Polyaromatics, wt %	7.17	3.3
Boiling point analysis, °C		
ibp	170	313
10 %	203	351
30 %	227	393
50 %	248	405
70 %	273	422
90 %	316	451
95 %	340	460
fbp	352	475

Hydrocracker Residue (HVGO or HCR; Boiling Range 350–600 °C). Hydrocracker residues are the unconverted product fraction from hydrocrackers. Due to the severe operating conditions of the hydrocrackers, HCR is a highly saturated product, with only a limited content of aromatics and low contents of polyaromatics. HCR or HVGO is a feedstock that can result in ethylene yields as high as in naphtha cracking. Typical yields for a HCR are shown in Table 9 for an AGO with properties as listed in Table 10.

A comparison of the feedstock requirements as a function of residence time for a 1000 000 t/a plant for ethane, propane, butane and naphtha is shown in Figure 8. It can be seen that a naphtha cracker has about three times the feedstock throughput of an ethane cracker, and that naphtha cracking shows some sensitivity to short residence times, whereas ethane crackers are not very sensitive with respect to feedstock savings as short residence times are applied. A careful economic analysis is required to select the optimum residence time range for a given feedstock, since shorter residence times lead to reduced availability due to shorter run length, as shown for ethane cracking in Figure 7.

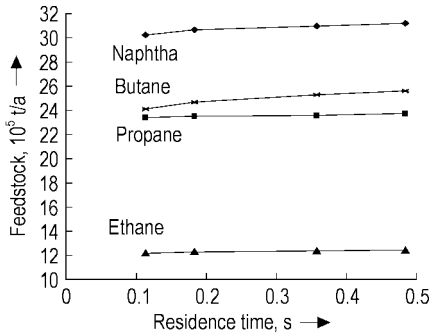

Figure 8. Feedstock requirements for a 1 000 000 t/a ethylene plant

5.1.4. Commercial Cracking Furnaces

Furnace Design. Modern cracking furnaces typically have one or two rectangular fireboxes with vertical radiant coils located centrally between two radiant refractory walls. Firebox heights of up to 15 m and firebox widths of 2–3 m are standard design practice in the industry. Firing can be performed with wall- or floor-mounted radiant burners, or a combination of both, which use gaseous or combined gaseous and liquid fuels. Fireboxes are under slight negative pressures with upward flow of flue gas. Flue gas flow and draft are established by induced-draft fans. In some furnace designs, burners are mounted on terraced walls.

Firebox length is determined by the total ethylene production rate desired from each furnace

and the residence time of the cracking operation. This also determines the number of individual radiant coils needed, with short residence times requiring many more individual coils than longer residence times for the same production capacity. This is due to the shorter lengths of the short-residence-time coils, which can be as short as 10–16 m per coil. Long-residence-time coils can have lengths of 60–100 m per coil. The number of coils required for a given ethylene capacity is determined by the radiant coil surface, which is in the range of 10–15 m^2 per tonne of feedstock for liquid feedstocks.

Production rate for each coil is determined by its length, diameter, and charge rate, which translates into a certain heat flux on the radiant coil. Values of 85 kW/m^2 for the average heat flux of a coil should be the maximum.

Radiant coils are usually hung in a single plane down the center of the firebox. They have also been nested in a single plane or placed parallel in a staggered, double-row tube arrangement. Staggered tubes have nonuniform heat-flux distributions because of the shadowing effect of one coil on its companion. If maximum capacity and full utilization of the firebox are desired, tube staggering can be adjusted so that the shadowing effect is only marginally greater than with a single-row arrangement.

Heat transfer to the radiant tubes occurs largely by radiation, with only a small contribution from convection. Firebox temperature is typically 1000–1200 °C.

The radiant coil of an ethylene furnace is a fired tubular chemical reactor. Inlet conditions must meet the required temperature, pressure, and flow rate. Hydrocarbon feed is heated from its convection-section entry temperature to the temperature needed for entry into the radiant coils. Dilution steam is introduced into the hydrocarbon stream in the convection section. The combined hydrocarbon and steam stream is then superheated to the target temperature at the radiant coil inlet (crossover temperature, XOT). Gaseous feeds require only the energy necessary to heat the feed and steam to the radiant coil inlet temperature. Liquid feeds require additional energy to heat the liquid to its vaporization temperature plus its latent heat of vaporization.

Energy is saved in the convection section by preheating the cracking charge; by using utility preheating, i.e., boiler-feedwater preheating, dilution-steam superheating, high-pressure steam superheating or by preheating the combustion air.

A typical arrangement of furnace elements is shown in Figure 9. In this case, the convection section contains six separate zones. As an example, the feed could be preheated in the upper zone, the boiler feedwater in the next, and mixed feed and dilution steam in the third zone. The process steam is superheated in the fourth zone, the HP (high-pressure) steam is superheated in the fifth zone and the final feed and dilution steam are preheated in the bottom zone. The number of zones and their functions can differ according to furnace design. For example, the upper zone could be used to recover heat for air preheating.

With the trend to higher capacities per individual furnace a new furnace set-up has been introduced to the industry with great commercial success. In this type two fireboxes are connected to one convection section. This design allows liquid feedstock cracking furnaces with capacities of 130 000 − 200 000 t/a to be installed. Capacities of gas furnaces are higher and can reach up to 250 000 to 300 000 t/a. This is a result of the general trend to built ethylene plant in much larger scale, which started with the number of gas-based projects erected in the Middle East in 2000.

The induced-draft fans allow complex convection sections to be designed with very high gas velocities to give optimum heat transfer in the convection section banks. Modern designs include integrated catalyst systems for NO_x removal by SCR (selective catalytic reduction) technology, whereby the NO_x present in the flue gas reacts with ammonia:

$$4\,NO_x + 4\,NH_3 + O_2 \rightarrow 4\,N_2 + 6\,H_2O$$

The temperature of the cracked gas leaving the radiant coils can range from 750 to 900 °C. Rapid reduction of gas temperature to 500–650 °C, depending on the feedstock, is necessary to avoid losses of valuable products by secondary reactions. This is accomplished by a transfer-line exchanger (see Fig. 9), which cools the cracked gases and recovers much of the heat contained in the cracked gas as high-pressure steam. The transfer-line exchanger can be a conventional or a linear exchanger as shown in Figure 10. Linear exchangers are preferred today for

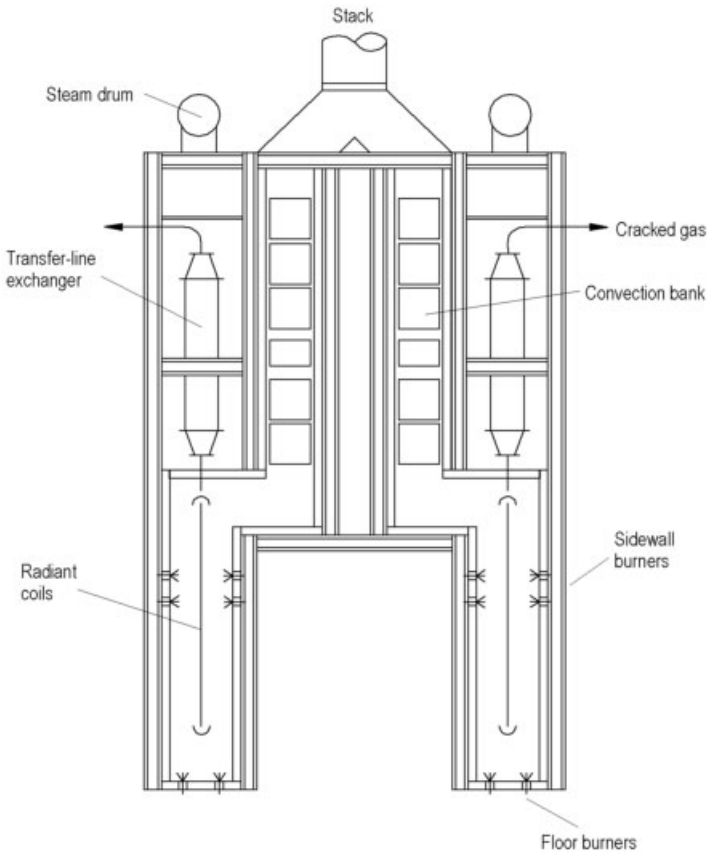

Figure 9. Typical arrangement of furnace elements

gas cracking, in order to avoid erosion problems with the very hard coke in gas cracking.

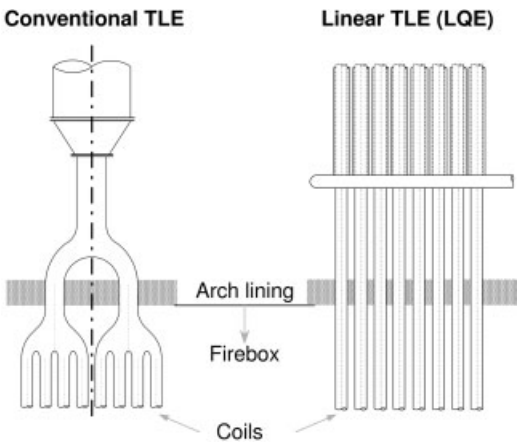

Figure 10. Conventional and linear transfer-line exchangers

In all modern furnaces the radiant coils exit the firebox at the top; the TLEs are top-mounted. In older designs coils also leave the bottom of the firebox, with TLEs mounted vertically along the side of the furnace or horizontally underneath it. Radiant-coil designs are generally considered proprietary information. All modern designs strive for short residence time, high temperature, and low hydrocarbon partial pressure. Most coil configurations have two features in common: coils are hung vertically, usually suspended on spring or constant hangers above the top of the radiant firebox, and they are fired from both sides of the radiant coil. Horizontal tubes are no longer used in ethylene furnaces.

Coils can range from a single, small-diameter tube with low feed rate and many coils per furnace, to long, large-diameter tubes, with high feed rate and few coils per furnace. Longer coils consist of lengths of tubing connected with return bends. Individual tubes can be the same

diameter or swaged to larger diameters one or more times along the complete coil; two or more individual tubes may be combined in parallel.

An example of a so-called split radiant coil arrangement is shown in Figure 11; here parallel small-diameter coils are combined into a larger diameter outlet coil. This arrangement allows advantageous temperature profiles to be applied, with a rapid temperature increase in the inlet section, and offers slightly higher yields than a uniform diameter coil of the same residence time.

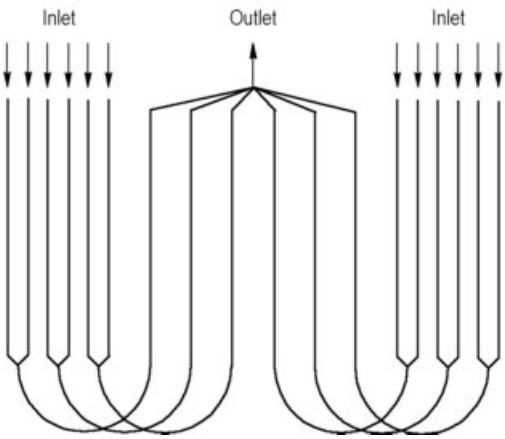

Figure 11. Short-residence-time coil arrangement

Coil lengths and diameters vary widely as can be seen from the following design characteristics:

Ethylene capacity (per furnace)	(50–250) × 10^6 t/a
Number of coils (per furnace)	4–200
Tube diameter	25–180 mm
Coil length	10–100 m
Coil outlet temperature	750–890 °C
Tube wall temperature	
clean	950–1040 °C
maximum	1040–1120 °C
Average heat flux	50–90 kW/m^2
Residence time	0.08–1.0 s

Cracking coils with internal fins have been offered to the industry, to increase the inner surface of a coil. These coils should offer the advantage of processing more feedstock in a coil of identical external dimensions. However, the fins result in a higher pressure drop of the coil, with disadvantages in yields, and coke spalling problems occur when the furnace is cooled down rapidly. Several new designs of fins and new processes for production of finned coils have been developed by different companies (Fig. 12).

Different developments are offered to the industry to reduce the coke deposit, resulting from catalytic coking and pyrolytic or thermal coking occurring during cracking. One approach is coating of the coil surface with inert materials to minimize the amount of catalytic coking (Fig. 13). Solutions are offered by several companies, but these technologies are rarely used, especially not in the large-scale new plants. For retrofits these coatings have been applied mostly in the USA. Coating can occur by, e. g., chemical vapor deposition (CVD) or physical vapor deposition (PVD).

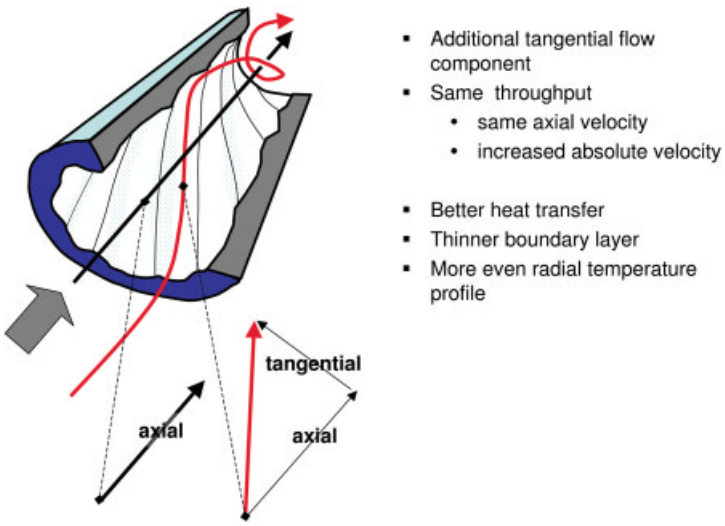

Figure 12. Coil with internal fins and improved heat transfer

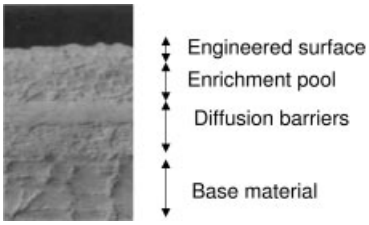

Figure 13. Coated radiant coil surface

The coils can also be passivated by chemical treatment during operation or prior to start-up. Passivation technologies are offered by various companies, including Atofina, ChevronPhillips, SK Corporation, Ondea Nalco, and Nova. Until today (2005) these technologies have not achieved a major breakthrough in new cracker projects. The advantage of chemical treatment is that the costly process of deposition by CVD or PVD can be avoided and the dosing can be adjusted during operation. A disadvantage is that chemicals are introduced that end up in the cracker products and should not influence the quality of the products significantly.

Another novel development in furnace technology is the application of new coil materials, especially 45Ni35Cr steels, offered by many suppliers. These high-alloy steels are superior to other HP modified materials and offer advantages with respect to carburization at similar creep rates as other HP modified materials.

An interesting new material is offered by Schmidt and Clemens (Kaiserau, Germany) under the designation Centralloy 60 HT. The product has an increased content of aluminum (≈6 %) and can be operated up to 1200 °C. Despite the high aluminum content the material can be machined, bent, and welded.

A different approach is the application of ODS (oxide dispersion strengthening) leading to "superalloys". ODS materials can be used up to 1300 °C with still good creep resistance but despite these ceramic-like properties are ductable and weldable like metallic materials. Due to the non-casting production process based on metal oxide powders, the costs for such materials are much higher than those for conventional HP modified or similar materials. The industrial breakthrough of ODS in ethylene plants has not been achieved until today [29].

ABB Lummus Global Furnace The coil design of Lummus furnaces has been modified in the light of pilot-plant investigations. Several generations of coils SRT-I (short residence time-first model) through SRT-X, have been produced.

The SRT-I model radiant coil is uniform in diameter. Many SRT-I heaters are still operating, primarily for cracking ethane, propane, or butane. All other coils are so-called split coils, with smaller inlet coils combined into larger outlet coils. The latest designs — SRT-V SRT-VI and SRT X — are short residence time coils with residence times in the range of 0.15–0.20 s. In the SRT-V coil, 8–12 small coils of ca. 40 mm internal diameter are headered into a single outlet coil of large diameter (150–180 mm). In SRT-VI designs a smaller number of parallel coils is headered into a larger outlet tube, which is connected to a linear quench exchanger.

ABB Lummus Global furnaces can be paired, as shown in Figure 9, or can operate alone. Stacks may be shared, with a common induced-draft fan; the flue gas flows upward. Some convection sections contain boiler feedwater preheating coils and steam superheating coils, along with steam desuperheat stations and process-heating coils. In some designs, steam is superheated in separate heaters. Lummus heaters can be fired by wall-mounted burners or by combinations of wall- and floor-mounted burners. Oil-fired floor burners are available.

Overall fuel efficiencies of 92–95 % net heating value (NHV) can be obtained, depending on feedstock, fuel sulfur content, firing control, and convection section type. Lummus furnaces can use any available transfer-line exchanger. Radiant-coil outlets may be paired, leading to a common exchanger. Capacities per furnace for full-range naphtha or atmospheric gas oil vary from 25 000 to 200 000 t/a.

Millisecond Furnace (KBR, Houston Texas). The Millisecond Furnace is designed for shortest residence times and low hydrocarbon partial pressures. It is recommended for gas and liquid feedstocks, offering the greatest advantages for the latter. This heater, developed in the early 1970s from a prototype installed by Idemitsu Petrochemical in Japan, reduces residence time to ca. 0.10 s. Kellogg successfully marketed the Millisecond concept in the mid-

1980s, but with the trend of the industry to robust furnaces rather than maximum selectivity, which started in the 1990s, Kellogg lost market shares.

KBR has enhanced its olefin technology portfolio through a licensing agreement with ExxonMobil Chemical. This agreement allows KBR to offer a combination of its own technology and that of a premier producer. Furnaces offered included the designs SC-1, SC-2, and SC-4 (Fig. 14).

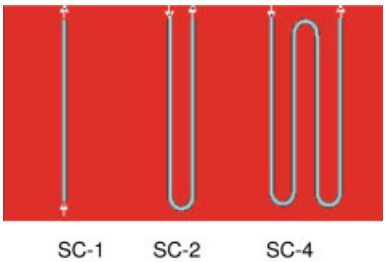

Figure 14. Radiant coils applied in KBR furnaces

Furnaces can be paired to use a common steel structure or they can stand alone. Paired furnaces can have a common stack and induced draft fan. Burners are located on opposite walls of the firebox at floor level; they use gas, oil, or combinations thereof with forced or induced draft. The design can be adapted easily for preheating combustion air. A Millisecond furnace is shown in Figure 15.

Flow is distributed to the many small-diameter radiant coils by pigtails connecting the individual radiant coils to their common feed header. Due to the short residence time and the resulting higher process temperatures, run lengths of Millisecond furnaces are shorter than those of competing furnaces and they thus require more decokes per year.

Shaw Furnace (The Shaw Group, Baton Rouge, USA). Shaw supplies an ultraselective cracking (USC) furnace coupled with a quench exchanger (see Section 5.2). Radiant coils have internal diameters of 50–90 mm. In contrast to many other contractors, Shaw is not offering split coils.

Their swaged USC-U (Super U Coil) coil configuration consists of two tubes connected with a return bend. The diameter of the outlet tube is larger than that of the inlet tube, allowing for thermal, molecular, and pressure expansion. Generally, outlet tubes from two USC-U coils are combined to feed a single linear exchanger.

The swaged USC-W coil configuration consists of four thin tubes, which differ in diameter. Individual diameters for each coil increase from inlet to outlet. A coil outlet can go directly to a LinEx (linear exchanger) quench system or two coils can share a quench system. The individual furnace capacity is obtained by installing 12–24 USC-W coils in the same firebox. Spring hangers support radiant coils from the top in the center plane of the firebox. Feed to the coils is equalized by critical-flow venturi nozzles.

The USC furnaces may be paired as shown in Figure 5, with a common stack and induced-draft fan, or they can stand alone. Burners are mounted on the wall, or on the wall and floor. Floor burners are gas- or oil-fired.

For furnaces using gas-turbine integration (GTI) or preheating of combustion air, a floor-fired, short-residence-time (FFS) design is available. This is completely floor fired, which reduces large-volume ductwork to a minimum. Coil geometry is similar to USC-U or USC-W designs, but to maintain a uniform firebox temperature, more and shorter coil legs are used. Overall fuel efficiencies of Shaw furnaces are competitive.

Technip Furnace (Technip, Paris, France). A number of radiant-coil designs (GK and SMK) are available, ranging from a simple serpentine coil of constant diameter to GK-1 (gradient kinetics, model 1) and GK-6 coils.

Gaseous feed, naphtha, and gas oil can be cracked with all coils. Residence times range from 0.15 to 0.5 s. Capacities range up to 250 000 t/a per furnace.

The GK radiant coil has two thin, parallel tubes from the coil inlet to ca. 66 % of the total coil length, where they combine to a single, larger tube. The thin tubes can be staggered or in-line; the larger diameter tubes must be in-line.

The Technip furnaces can be paired with a common stack and induced-draft fan, or they can stand alone. Top outlet coils are preferred. Burners are wall-mounted or wall- and floor-mounted.

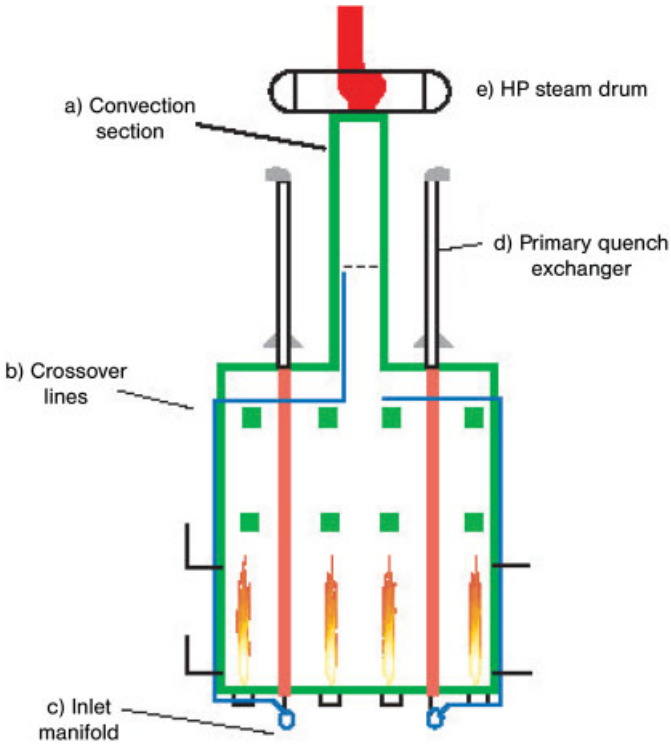

Figure 15. KBR SC furnace module

Linde-Pyrocrack Furnace (Linde AG, Munich, Germany). Linde-Pyrocrack furnaces provide capacities of 20 000–300 000 t/a. Feedstocks can be all straight-run or pretreated hydrocarbons from ethane, propane, butane, natural gas liquids, naphthas, raffinates, gas oils, gas condensates and hydrocracked vacuum gas oils. Furnaces are designed for segregated cracking or co-cracking and for maximum feedstock flexibility. Linde and its U.S. subsidiary Selas Fluid Processing have installed furnaces with a total ethylene capacity $> 25 \times 10^6$ t/a.

Pyrocrack coils are all split coils with small inlet coils headered into a larger outlet coil, with residence times from 0.15 to 0.5 s. The three basic Pyrocrack coil types are shown in Figure 16. Coil diameters range from ca. 40/55 mm (inlet/outlet) for Pyrocrack 1–1 coils to 100/150 mm for Pyrocrack 4–2 coils. Critical venturi nozzles are used to ensure equal distribution and residence times in all parallel coils of a furnace.

Today floor firing or a combination of sidewall and floor firing is used, the latter offering the best distribution of the released heat over the whole firebox and a excellent radiant efficiency. Especially for large-capacity furnaces with high fireboxes the combination of floor and sidewall firing offer the advantage of an even distribution of the heat release. Fuel for the radiant burners can be gas or oil or both. Oil firing offers the opportunity to use a fraction of the pyrolysis fuel oil (i. e., the liquid product of the pyrolysis boiling above 180 °C) as fuel and thus of increasing the amount of the pyrolysis methane fuel fraction available for export to other consumers. With today's requirements of low NO_x emissions, the application of liquid fuel burners in ethylene plants is the exception, as the NO_x emissions for liquid fuels are above the BAT levels (best available technique), forming the basis for local regulations in Europe. Present BAT levels in Europe are 80 to 100 mg/Nm^3 for furnaces in ethylene plants without catalytic DeNOx systems. Floor firing offers the advantage of somewhat lower NO_x levels than sidewall burners.

The convection section of Linde Pyrocrack furnaces contains process heating bundles for feed, dilution steam, boiler feedwater, and high-pressure steam superheater bundles. Overall efficiencies of up to 94–95 % (on lower heat-

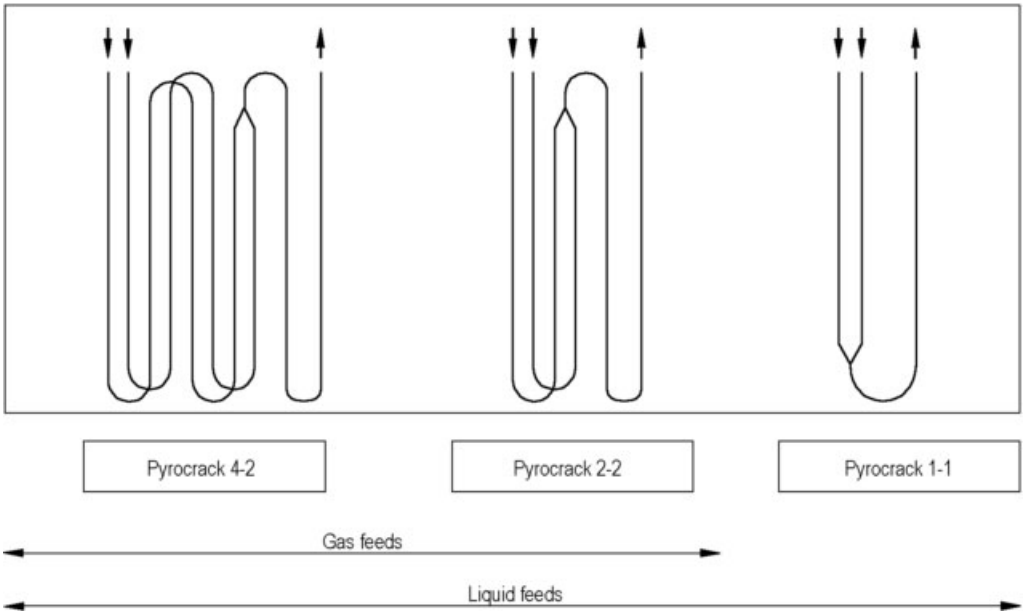

Figure 16. Linde-Pyrocrack coils

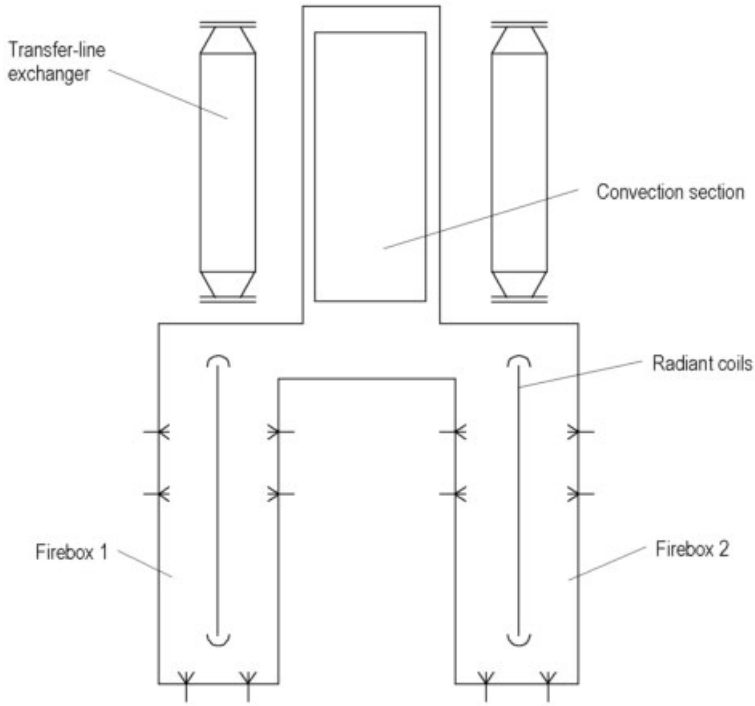

Figure 17. Twin radiant cell furnace

ing value, LHV) are typical for Linde furnaces. High-pressure steam systems for heat recovery have been developed by Linde for up to 14 MPa. Linde has built furnaces with SCR-based DeNOx technology integrated into the convection section. The design of the side and front walls of the convection section provides uniform heat distribution and prevents flue gas bypassing. Furnace draft is maintained by an induced-draft fan mounted over the convection section.

Linde has introduced with great commercial success the Twin Radiant Cell concept, in which a single high-capacity furnace has two radiant sections, which are connected to a common convection section (Fig. 17). The two radiant sections can be operated independently with different feedstocks or even with one firebox in cracking operation and the other in decoke operation. This offers the advantage of less capacity loss during decoke operation of a large capacity furnace. Other new features are a proprietary system for processing gas condensates, which have an extremely high and variable final boiling point. This system was developed from the well-established Linde technology for heavy hydrocracker bottoms and adapted to the requirements for gas condensates.

Other Furnaces. The furnaces listed below show some interesting features, but none is of commercial importance in today's worldwide olefin business. Only the furnace designs of the contractors listed above (ABB Lummus Global, KBR, Shaw, Technip, and Linde) are considered to be competitive.

The *terraced-wall furnace* (Foster & Wheeler Energy Corp., Livingston, N.J.) uses burners firing upward from shelves or ledges built into the walls of the firebox (Fig. 18). A similar firebox design is used for reformer heaters. This arrangement apparently gives uniform heating of the radiant coil. Furnaces are all-gas or all-liquid fuel fired.

The *Mitsui advanced cracker* (Mitsui Engineering & Shipbuilding Co., Tokyo, Japan) offers lower construction costs, smaller land area requirements, high thermal efficiency, and low NO_x emissions. The mass of steel used in a 400 000 t/a ethylene plant is 30 % lower than that used in a 300 000 t/a plant with conventional furnaces. This is achieved by a patented installation of three radiant coils in the firebox space normally occupied by one or two coils [30].

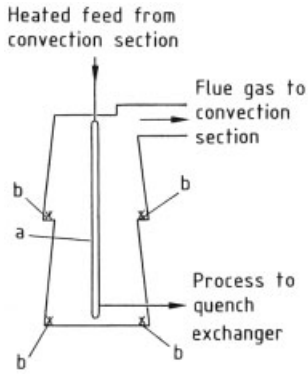

Figure 18. Foster–Wheeler terraced wall furnace
a) Radiant coil; b) Burners

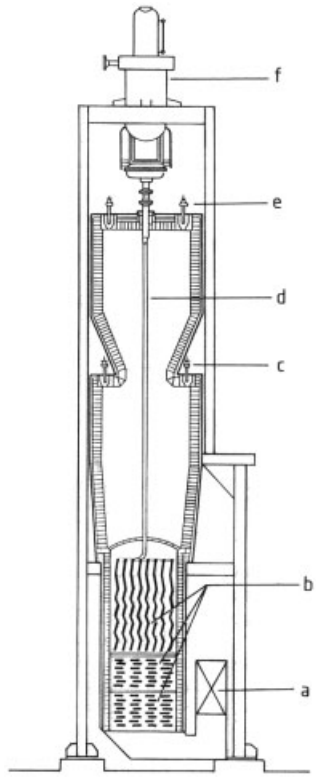

Figure 19. Section view of Mitsubishi M-TCF furnace *
a) Flue-gas duct; b) Convection tubes; c) Side burners; d) Cracking coil; e) Roof burner; f) TLE
* Reprinted with permission of Mitsubishi Heavy Industries.

The *Mitsubishi Furnace* jointly developed by Mitsubishi Petrochemical Co. and Mitsubishi Heavy Industries of Japan is quite different in design from other furnaces (Fig. 19). Known as the Mitsubishi thermal cracking furnace (M-TCF), it is a vertical, downdraft furnace with the convection section located below the radiant firebox. It is fired with gas or oil burners located in two rows mounted on the ceiling and two rows mounted on terraced walls. All burners are down-firing, with long flat flames. According to Mitsubishi, up to 70 % of the total heat can be supplied with any type of distillate fuel, including the cracked residue from an ethylene plant. Flue gas flows to a common stack by means of an induced-draft fan, which produces suction at the bottom of the convection section.

5.1.5. Tube Metallurgy

Ample evidence indicates that maximum yields of ethylene from tubular cracking of hydrocarbons require high temperature, short residence time, and low hydrocarbon pressure in the radiant coil. Modified coils can help achieve these conditions, but coil design is limited by metallurgy. Radiant coils must be selected in light of the following considerations:

- Operating temperature
- Tube service life
- Cost
- Carburization resistance
- Creep-rupture strength
- Ductility
- Weldability

An outside tube wall operating temperature up to and even exceeding 1100 °C is now possible. Coils are designed to operate at 975–1000 °C with an internally clean radiant coil. Tube wall temperature rises with time because of coke deposition inside the coil. Coke acts as a thermal insulator inside the coil, requiring increased tube wall temperature for a given furnace loading. Localized overheating must be avoided during removal of coke from the coil (when steam or a mixture of steam and air is used). With steam-air decoking, combustion occurs inside the coils.

Tube service life is economically important in plant operation. Furnace capital costs represent ca. 20 % of the total cost of an ethylene plant. About one-third of this is for radiant coils. With current metallurgy, a five-year service life in the hottest section and a seven-year service life of the inlet section of a furnace are typical. Service life is shortened by carburization of the inner tube surface, which is favored by the environment inside a cracking coil. Temperature is high, and feeds are usually highly carburizing; reaction occurs between the gaseous phase in the tube and its base metal.

Coil alloys are normally protected by an oxide layer on the inner surface. If this layer is destroyed and cannot be regenerated, carbon diffuses into the base metal, which itself diffuses toward the inside surface. Carburization may be more severe during decoking and can be reduced by lowering the percentage of air mixed with steam during this cycle. Aluminized coatings on the inner surface of the tube may reduce carburization [31 – 36]. The extent of carburization can be measured by magnetic tests [34].

The materials used for cracking coils are high-alloy steels, e.g. with 45 % Ni, 35 % Cr and several important trace metals (Si, Mn, Nb, Mo, W, Ti) [37]. The technical designation of this most common radiant coil material is HP 40 modified. The new materials offered by the leading vendors can be operated up to 1150 °C. Only a limited number of manufacturers worldwide are able to produce these materials according to very strict quality standards.

5.1.6. Thermal Efficiency of Ethylene Furnaces

Cracking furnaces represent the largest energy consumer in an ethylene plant; their thermal efficiency is a major factor in operating economics. New plants are designed for 93–95 % thermal efficiency, and revamping of older ones can increase efficiencies to 89–92 %. Only a limited number of heat sinks are available for heat released in the radiant firebox of a cracking furnace:

1) Process heat duty and duty for utility heating
2) Wall or radiation loss
3) Stack heat loss

Process heat duty is the energy required to heat the feedstock and dilution steam from the

temperature at which they enter the furnace convection section to the temperature at which they leave the radiant coil. Process heat duty varies with feedstock composition, feed rate, and cracking severity. The duty for heating the utilities is the heat requirement to heat boiler feedwater in the convection section (from 110 °C to ca. 180 °C), before it is used for the HP steam production in the TLEs, and for superheating the saturated HP steam produced in the TLEs in the convection section.

Wall or radiation loss is that portion of the heat released in the radiant firebox which is lost to the atmosphere through the refractory and outer steel shells of the radiant and convection sections. Wall loss is initially determined by the type and thickness of thermal insulation in the furnace. Typical wall losses in modern furnaces are 1.2 % of the fired duty for the radiant section and 0.3 % of the fired duty for the convection section.

Stack heat loss is that portion of the heat released in the radiant firebox which is discharged to the atmosphere in the flue gas. Stack heat losses are around 5 % of the fired duty. This is the only heat loss that furnace operators can control, given a fixed furnace design, quality of maintenance, feed rate, and coil outlet temperature. Stack heat loss is sensitive to excess combustion air in the furnace and, thus, to draft, fuel composition, and air leakage.

Radiative Heat Transfer and Furnace Thermal Efficiency. The heat absorbed by the radiant coils and lost through the walls of a radiant-section firebox must equal the heat given up by the flue gas as it drops from the adiabatic flame temperature to the temperature at which it leaves the firebox [38]. The heat absorbed by the radiant coils of a furnace is a function of coil geometry, burner location, flame emissivity, temperature of the flue gas leaving the radiant firebox, and tube wall temperature of the radiant coil [39]. Typical radiant efficiencies are in the range of 38–42 % for side wall firing, 40–45 % for a combination of sidewall and floor firing and 42–47 % for 100 % floor firing. The differences result from the direct radiation loss to the convection section if the burners are mounted in the upper section of the radiant box.

The effects of excess combustion air and combustion air preheating on radiative heat transfer have been determined [40, 41]. Excess air rates in modern furnaces are ca.10 %, but some plants are operated at rates as low as 5 % excess air, resulting in reduced fuel consumption.

At 20 % excess combustion air (for a given furnace configuration and a radiant-section process duty), a 21 % reduction in fuel consumption is possible if combustion air is heated to 270 °C. Fuel usage increases rapidly as excess combustion air increases. At 49 % excess air, fuel consumption is the same as if combustion air were not preheated.

Combustion air preheating may be used in a new ethylene plant or in a modernized furnace to increase thermal efficiency. Various heat-exchange techniques can be used, including direct exchange between flue-gas discharge and inlet combustion air by use of rotating refractory vane elements, and indirect exchange by means of heat pipes, high-pressure water, or circulating high-temperature thermal fluids.

The most common technique for combustion air preheating is the use of hot exhaust gas from gas turbines (GTs). Such plants co-produce electricity (in the range of 30–60 $mW_{el.}$), and this has a major impact on the plant economy. The GTs are exclusively used for electricity production. Direct driving of large compressors by the GTs is not feasible, due to the maintenance requirements for GTs, which require an annual shutdown. In modern plants the ethylene plant can run with and without the GT [42]. The economy of operation with an integrated GT is highly dependent on the credit for the electricity and the increased HP steam production. Scenarios that favor GT integration in Japan, Korea, and Brazil have existed in the eighties.

Radiant-coil inlet and outlet process temperatures in an ethylene furnace must be the same after application of preheated combustion air as they were before, to obtain the same cracking severity.

Convective Heat Transfer and Furnace Thermal Efficiency. Convective heat transfer is widely used in isolated green-fields ethylene plants to improve thermal efficiency by preheating boiler feedwater, generating steam, or superheating HP steam in the convection section of a furnace. Recovered energy must be considered in the overall energy balance for a plant of this

type. Such techniques can be employed in older plants if boiler feedwater, generated steam, or superheated steam can be used. Otherwise, preheating the combustion air is an alternative.

5.1.7. Coking and Decoking of Furnaces and Quench Coolers

Radiant-Coil Coking. Hydrocarbon pyrolysis produces acetylenic, diolefinic, and aromatic compounds, which are known to deposit coke on the inside surface of the radiant coil [43, 44]. This coke layer inhibits heat transfer from the tube to the process gas, raises tube wall temperature, and reduces the cross-sectional flow area of the tube, thus increasing the pressure drop along the radiant coil. Cracking yield is thus lowered because of increasing hydrocarbon partial pressure. When tube temperature limits are reached, or pressure limitations of the sonic venturi nozzles used for flow distribution in short-residence-time furnaces, the furnace must be shut down to remove the coke. During the course of one on-stream cycle, deposited coke can reduce the heat-transfer efficiency of the firebox by 1–2 %, resulting in a 5 % increase in fuel consumption [45]. As radiant efficiency decreases, convection section heat transfer increases. This results in an unwanted increase in process crossover temperature and a lower demand for radiant-coil heat input. After furnace firing equilibration, actual fuel demand increases by 1–3 %.

Gaseous feedstocks such as ethane yield a harder, denser coke with higher thermal conductivity than naphthas or gas oils. Highly aromatic feeds produce the most coke.

The mechanism of coil coking is not fully understood [46]. Two types of coke can be distinguished, the catalytic coke and pyrolytic coke (Fig. 20).

The possibility that tube metals such as iron and nickel catalyze coke formation is supported by the observation that sulfur in the feed inhibits coking. A sulfur content above 400 ppm, however, may increase coking.

Quench-Cooler Coking. Coke deposited in TLEs reduces both heat transfer and the amount of steam generated to such an extent that the equipment must be cleaned. Different mechanisms have been proposed for TLE coking or fouling [47]:

1) Spalled coke produced in the radiant coil is carried into the TLE inlet cone where it blocks some of the tubes on the TLE entry tube sheet. Such fouling is most serious in ethane cracking.
2) Poor flow distribution in the entry cone and at the tube sheet causes eddies and backmixing. Gas so trapped experiences long residence time at high temperature, which increases tar and coke production.
3) Heavy polynuclear aromatics and other high-boiling components in the cracked gas condense on the cool TLE tube walls. This is particularly true for gas oil and heavy naphtha feedstocks. The condensed high-boiling substances are gradually converted to cokelike materials. This mechanism is supported by the fact that fouling is fastest at the start of the run (SOR). Typically, TLE outlet temperatures increase by 50–100 °C during the first day or two in gas oil cracking. Deposits during this period reduce the heat-transfer rate, and inner surface temperature at a given location in a tube becomes higher than the original clean wall temperature. The fouling rate drops as the surface temperature increases; with time process gas temperature slowly increases.

Decoking. At some time, furnace operation that is affected by coke deposits must be stopped so that the coke can be removed. Mechanical techniques are not feasible. Usually, the coke is burned out with a mixture of steam and air (Fig. 21).

The furnace is taken off-line, the residual hydrocarbons are purged downstream with steam, and the process flow is rerouted to a special decoking system. A steam–air mixture is then introduced into the radiant coils to burn out the coke at coil outlet temperatures of ca. 800 °C. The air concentration is increased carefully to avoid overheating the radiant coil. In modern furnaces the CO_2 content is measured continuously during decoking and the air flow rate is adjusted accordingly. After ca. 20 h the decoking of the radiant coils is complete and the decoking conditions are adjusted to decoke the TLE. Total decoking of radiant coil and TLE takes ca.

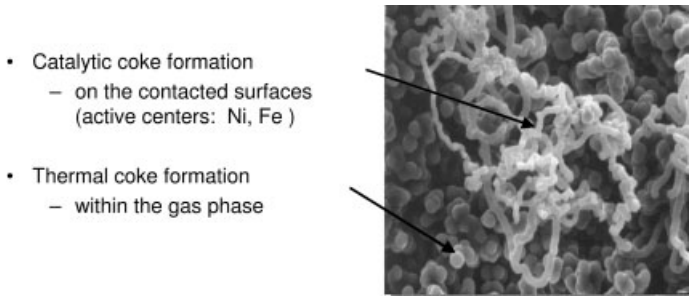

Figure 20. Catalytic and thermal coke

36 h. Mechanical cleaning of the TLE is not required in modern plant employing appropriate decoking procedures.

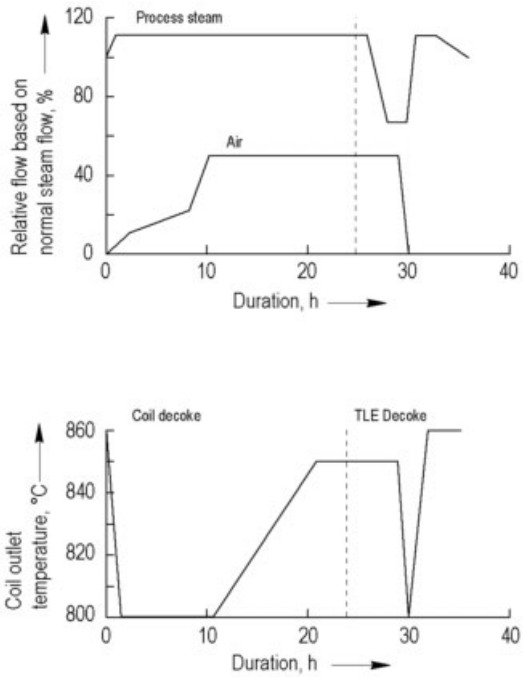

Figure 21. Steam–air decoking

Steam-only decoking makes use of the water–gas reaction, which is endothermic and less likely to burn out the radiant coils. This method does not require the furnace to be disconnected from downstream processing equipment, if that equipment is designed to accept the required amount of steam and handle the carbon monoxide produced. Temperatures as high as 950–1000 °C are required for steam-only decoking. The decoking gas from the radiant coils may be routed through the transfer-line exchangers, to partially decoke them.

Decoking time and radiant-tube wall temperature decrease as the amount of air mixed with steam increases [48]. The result of a steam-only decoking operation is poorer than that of a steam–air decoke.

Modern furnace quench systems are designed so that off-gas from radiant-coil decoking passes directly through the transfer-line exchangers, which are decoked as a result. About once a year a modern TLE must be opened for inspection and cleaned mechanically. This is carried out by shutting down the furnace, disconnecting the TLE, and removing the coke from the tubes mechanically or hydraulically.

The furnace should never be shut down and cooled without radiant-coil decoking. Differences between the coefficients of thermal expansion of the coke and of the metal tube wall are such that the radiant coils may rupture.

Off-gas from decoking is usually diverted to a decoking system containing a knockout pot equipped with a quench water inlet or a high-performance cyclone separator with dry or wet separation of the coke particles in the effluent. The cleaned, cooled gases are discharged to the atmosphere. Alternatively, the decoke off-gas can be rerouted to the firebox, where the coke particles are incinerated.

5.2. Quenching of Hot Cracked Gas

Cracked gases leave the radiant coil of an ethylene furnace at 750–875 °C. The gases should be cooled instantaneously to preserve their composition. However, this is not practical, so furnaces are designed to minimize residence time

of the hot gas in the adiabatic section between furnace outlet and quench-system cooling zone. Typically, residence time in this adiabatic zone should not exceed 10 % of the residence time in the radiant coil.

The cracked gas is generally cooled by indirect quenching in the TLEs, which permits heat recovery at a temperature high enough to generate valuable high-pressure steam. Direct quenching by injection of oil has been displaced almost totally by indirect quenching.

Design objectives for transfer-line exchangers are as follows:

1) Minimum residence time in the adiabatic section between furnace outlet and TLE cooling surfaces, limiting secondary reactions that reduce the value of the cracked gas
2) Uniform flow to all tubes across the surface of the tube sheet, preventing eddies in the entry cone, which prolong residence time of the cracked gas and favor coke deposition.
3) Low pressure drop across the exchanger; pressure drop across the TLE is directly reflected by the outlet pressure of the radiant coil; increasing pressure at the outlet of the radiant coil degrades the value of the cracking pattern.
4) High heat recovery in the cracked gas, which directly affects operating economics. The pressure for steam generation is affected by the cracking feedstock. For light hydrocarbon cracking, in which little TLE fouling from condensation results, steam of much lower pressure (60 bar) can be generated. With such feeds, two TLEs can be used in series, one for generating high-pressure steam, the other for low-pressure steam or other heat recovery. The TLE outlet temperature must be higher for liquid feeds (360 °C) because of condensation of heavy components in the cracked gas; therefore, the pressure of the generated steam can be higher (8–12 MPa). In these cases, because of the high outlet temperature, TLEs are generally followed by a direct oil quench to lower the gas temperature for entry into the downstream portion of the plant.
5) Reasonable TLE run lengths. Fouling of TLE tubes by condensation and blockage by coke increase outlet temperature and pressure drop. Ultimately, the system may have to be shut down and cleaned. The TLE outlet temperature usually increases rapidly for several hours from the start-of-run (SOR) value; the increase then falls to a few °C per day.

Figure 22 shows the TLE outlet temperature over the run length for ethane, naphtha, and heavy feedstock (AGO, HVGO) cracking in a modern highly selective furnace.

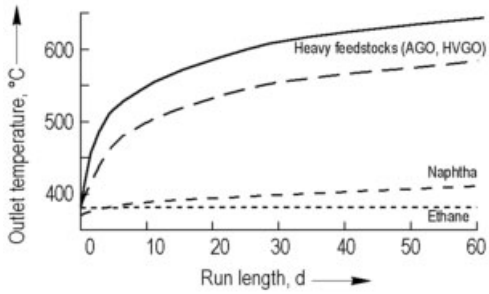

Figure 22. Transfer-line exchanger outlet temperatures

The length of run for a TLE in gas cracking service is limited by coke blockage of the TLE inlet tubes [46]. For liquid feeds, TLE run length is governed by fouling at the outlet end of the tubes resulting from condensation of heavy components in the cracked gas. Fouling rate increases as boiling range and relative density of the feed increases.

The transfer-line exchangers with commercial importance are the Borsig exchangers (shell-and-tube and a double pipe closed coupled linear exchanger), Alstom (multi double pipe and a double pipe linear exchanger), the Shaw USX exchanger and the KBR exchanger (similar to the USX).

Borsig Transfer-Line Exchangers (Borsig, Berlin, Germany). The Borsig TLE design has gained a large share of the world market (over 50×10^6 t/a ethylene capacity) for use with feedstocks ranging from ethane to atmospheric gas oils. Borsig offers two alternatives: a conventional shell-and-tube design and a linear exchanger. The latest series of the shell-and-tube exchanger, the tunnel-flow exchanger (Fig. 23 details of the TLE are shown in Fig. 24), which was introduced in 1989, is specially designed to operate with large temperature and pressure differences across the hot-end tube sheet. This sheet is thin (ca. 15 mm) but is supported by a studded stiffening system. The main feature of

the tunnel-flow exchanger is the controlled water flow pattern in the inlet section to avoid particle deposition on the tube sheet causing hot spots.

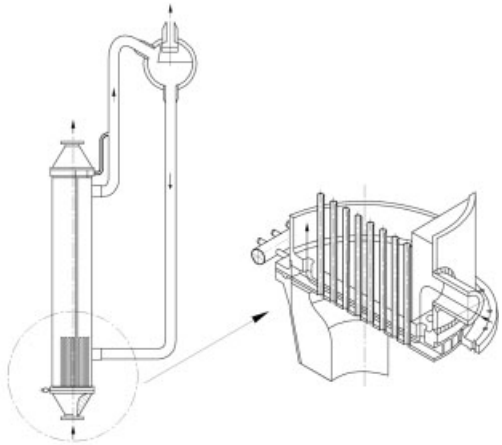

Figure 23. Borsig tunnel-flow exchanger

Water flows from downcomers (a) into a circumferential water-jacket annulus (b). From here water enters several parallel tunnels (c), which distribute the water perfectly over the bottom tubesheet and via annular gaps in the stiffening plate (e) to the individual tubes (d) containing the hot cracked gas, which is cooled down by the water, producing high-pressure saturated steam.

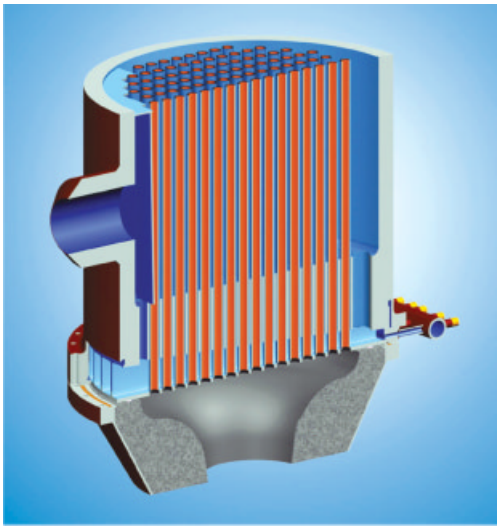

Figure 24. Details of Borsig TLE

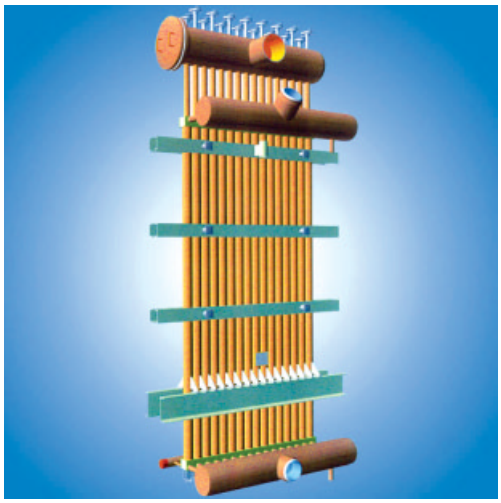

Figure 25. Details of the Borsig Linear Exchanger

Water flows through the tunnel system toward the opposite side at high velocity in intimate contact with the tube sheet, thereby preventing deposition of solids. On the opposite side, water flows upward through the circumferential annulus and enters the main shell volume (f). Debris may be deposited on the outer edges of the tube sheet as water flows from it to the main shell, where blowdown nozzles are present. Minor deposits here cause no problems because no heat load is present in this area.

Borsig exchangers can be mounted horizontally or vertically; the latter is the design for all modern furnaces with a top-out concept. Outlet tube sheets are generally thick and uncooled. Thin, studded tube sheets may be needed for TLEs with high outlet temperatures, e.g., in gas-oil cracking [49].

The is the Borsig Linear Quencher, is a double-pipe exchanger, connected directly to the coil outlet. This concept offers a minimum adiabatic volume of the cracking system. However, due to the larger tube diameter of this exchanger (up to 85 mm od), the heat transfer is not as efficient as in the conventional system with smaller tubes. This translates into tube lengths of ca. 20 m, which lead to the characteristic U shape of the linear exchangers, to keep the elevation of the steam drum within a reasonable range. Due to the length of 20 m, the pressure drop is higher than in a conventional system.

Figure 25 shows a Borsig Linear Quencher (BLQ);Figure 26. a Linear Exchanger Module.

The main features of the BLQ are the turboflow inlet chamber, with high water velocities to avoid solids deposition and a stepped bend at the end of the primary leg to resist erosion in the U-shaped return bend.

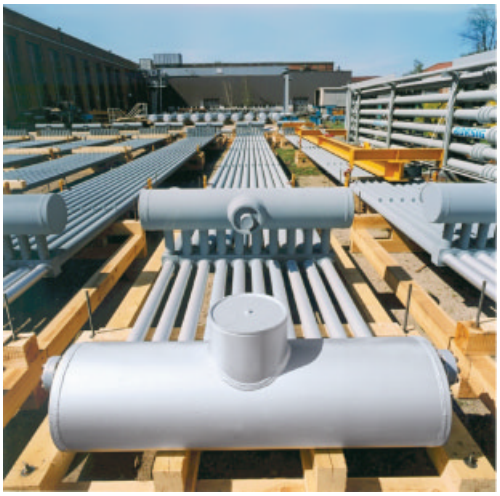

Figure 26. Borsig Linear Exchanger module
The ARG pipeline grid in Europe is shown in Figure 37.

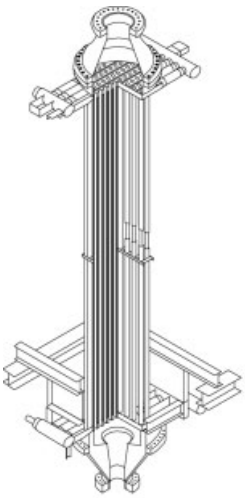

Figure 27. Alstom double-pipe transfer-line exchanger

Alstom Exchangers. Alstom Power and Energy GmbH, Kassel-Bettenhausen, Germany has been designing transfer-line exchangers for ethylene crackers since 1960 (formerly as SHG). The Alstom transfer-line exchanger shown in Figure 27 has separate exchangers, steam drums, risers, and downcomers. The basic heat-exchange element consists of concentric tubes (a, b) welded at each end to oval tubes (c). Process gas flows through the inner tube, and water circulates through the annulus between the tubes. Many double-tube elements are combined into a single exchanger by welding the oval headers to form a gastight tube sheet at each end. Process gas is delivered to the inner tubes of the basic tube element by a specially designed entry cone (h), which provides uniform process gas flow through all tubes and reduces the probability of eddy formation in the cone. Process gas leaving the exchanger is collected in a conical chamber and delivered to downstream equipment. Water flow to and from the oval tubes is provided by downcomers and risers (e) connected to the oval tubes by transition pieces (j).

The Alstom transfer-line exchangers can be mounted vertically with upward or downward flow of process gases. They may also be mounted horizontally with the gas-discharge end slightly elevated with respect to the gas-inlet end. They are normally cleaned by steam–air decoking and from time to time cleaned by hydrojetting.

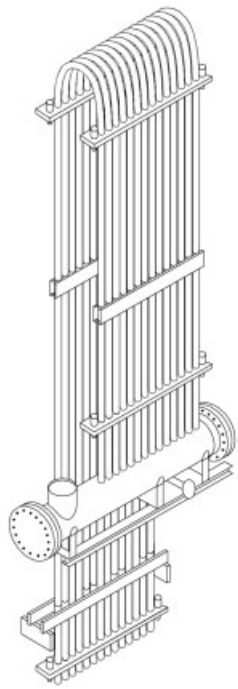

Figure 28. Alstom close coupled linear exchanger

Alstom offers also a closed coupled linear exchanger, with a reduced adiabatic reaction zone, but due to the increased length has a higher pressure drop than the conventional exchanger. Figure 28 shows an Alstom close coupled linear exchanger in the typical U-shape. Feeds for all Alstom exchangers range from ethane to gas oils [50].

Shaw Quench System (Shaw, Baton Rouge, USA). The Shaw proprietary quench system does not rely on a tube sheet. The USX (ultraselective exchanger) quench cooler shown in Figure 29 is a double-pipe exchanger, with the inner pipe slightly larger than the radiant-coil connecting piping. This design can quench the effluent from one to four radiant coils. The USX outlet temperature (500–550 °C) is higher than in most other transfer-line exchanger designs; this helps to reduce coking in the exchanger. Normally, effluents from all USX exchangers on a given furnace are combined downstream and fed to a conventional multitube exchanger where the gases are quenched to a more normal TLE outlet temperature.

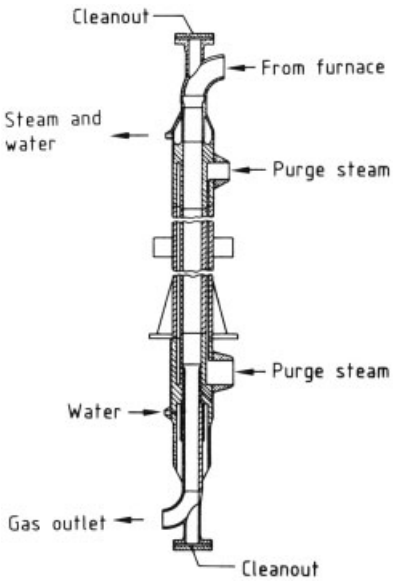

Figure 29. Ultraselective exchanger (USX) quench cooler*
* Reprinted with permission of Stone & Webster Engineering Corp.

This design, by eliminating the tube sheet and entry cone, eliminates turbulent eddies and backmixing between the coil outlet and the cooling surfaces, thereby reducing fouling in the TLE entry zone. The large USX gas flow tube reduces pressure drop across the exchanger and, thus, the probability of plugging by coke from the radiant coil.

The USX exchangers can be steam–air decoked; cleaning, when required, is facilitated by openings at each end of the exchanger.

KBR Millisecond Primary Quench Exchanger (KBR, Houston, Texas). A special primary quench exchanger (PQE) was developed for use with KBR's Millisecond furnace. The exchanger has a double-pipe arrangement (Fig. 30) capable of generating high-pressure steam.

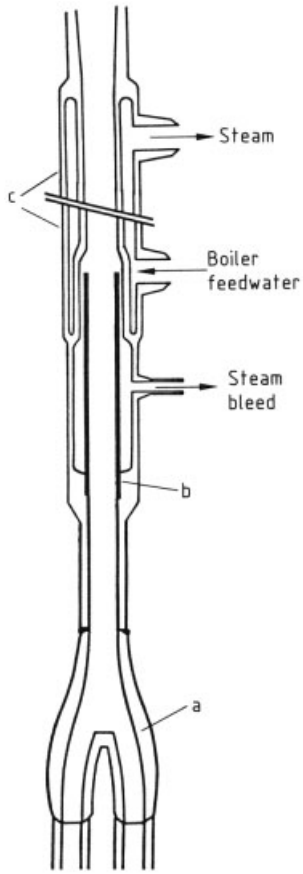

Figure 30. KBR millisecond primary quench exchanger *
a) Y-shaped fitting; b) High-alloy insert; c) Double pipe
* Reprinted with permission of KBR.

A special high-alloy section of tubing is inserted to avoid excessive thermal stress at the high-pressure steam transition piece. This section has the Y-fitting at its base and extends up to the boiler feedwater inlet nozzle. A steam bleed helps to control thermal gradients and to prevent coke deposition in the annulus between the high-alloy insert and the exchanger inner surface.

The steam system is a thermosiphon type, with individual risers and downcomers tied together in common headers connecting to a single steam drum. KBR's PQEs are normally fabricated in bundles for ease of shipping and installation. Typically, one PQE serves two coils in a Millisecond Furnace.

Other Quench-Cooler Designs. Proprietary quench coolers have been developed by ABB Lummus Global,KBR, Foster-Wheeler, and others.

5.3. Recovery Section

The cracked gas leaving the cracking furnaces must be cooled to recover the heat it contains in the cracked gas and render it suitable for further processing. The temperature of the cracked gas leaving the primary TLE depends on the feedstock type; it varies from 300 °C for gaseous feedstock, to 420 °C for naphtha feedstock, and up to 600 °C for AGO and hydrocracker residues. In ethane and propane cracking, cooling to ca. 200 °C is performed in heat exchangers by preheating feedstock or BFW. However, this step can not be done in heat exchangers for feedstocks heavier than propane, due to the condensation of the liquid portion contained in the cracked gas. Cooling such cracked gases to 200 °C would lead to excessive fouling of the heat exchangers by carbon deposits. Thus, cooling is performed by injection of oil into the cracked gas (oil quench) by means of special tangential-injection mixing devices. Oil quenching is performed individually for each furnace or in the combined cracked gas line after collection of the cracked gas from all furnaces. In modern designs, the oil injection is located in the furnace area, integrated into the cracked gas line downstream from the TLEs.

The cracked gas leaving the furnace section has a temperature of ca. 200–250 °C and is routed to the separation section via a large cracked gas line. The collected gas stream contains the effluents of all cracking furnaces, including those processing recycle streams.

Table 3 to Table 5 shows the yields of the product fractions for all typical feedstocks and for varying cracking conditions. Apart from the type of feedstock, the gas composition and, therefore, the design of the downstream recovery process also depends on the cracking severity and the applied coil design.

Further processing, i.e., separation into the main products or fractions, can be performed in different sequences and depends on the feedstock type and the number and the specification of the plant products. The main downstream processing steps are the removal of the heat contained in the cracked gas, condensation of water and heavy hydrocarbons, compression, washing, drying, separation, and hydrogenation of certain multiply unsaturated components. The main task of these steps is to recover the desired products at given specifications and at the desired battery-limit conditions.

Ethane, propane, and butane can be recycled to the furnace section to increase the overall yield of ethylene per tonne of feedstock processed. Processing differs for cracked gas derived from gaseous and liquid feedstocks, since the portion of heavy components increases, especially when processing naphtha or even heavier feedstocks such as gas oil and hydrocracker bottoms. When pure ethane is the feedstock, the amount of C_3 and heavier byproducts is small, and only rarely is the recovery of heavier products economically feasible. As the concentration of heavier components increases, recovering them for direct sale becomes feasible. Addition of significant amounts of propane to ethane feedstock makes a depropanizer necessary and hydrotreating and fractionation of the C_3 cut may be required; butane feeds make oil and gasoline removal from the cracked-gas compressor necessary; pentane-range feeds require fractionation of light from tarry oils, and so on.

5.3.1. Products

Local market conditions and the degree of integration of the ethylene units into refining or petrochemical complexes influence the desired

products and the feedstocks used. Many types of coproduct can be generated with different equipment [51 – 53]; only the major product options are discussed here.

The principal ethylene coproducts are as follows:

- Acetylenes (C_2 and C_3) are hydrogenated to ethane, ethylene, propane, and propene; or they may be recovered and sold as products.
- Aromatics (various fractions) can be recovered or they may remain in the hydrotreated pyrolysis gasoline.
- The C_4 fraction can be refined for butadiene, butene, isobutene, or mixtures thereof. If no C_4 products are desired, C_4 dienes and olefins are hydrogenated and recycled to cracking
- C_5 olefins can be either recovered and refined to give isoprene, piperylene, and cyclopentadiene, or hydrotreated in the pyrolysis gasoline fraction.
- Ethane is recycled as cracking feedstock or used as fuel.
- Fuel oil is used as such or to produce coke or carbon black.
- Hydrogen is purified and used for hydrogenation steps in the plant; excess hydrogen is sold or used as fuel in the plant.
- Methane is used as fuel or sold.
- Naphthalene is recovered for sale or left in the pyrolysis fuel oil fraction.
- Propane is recycled as a cracking feedstock, used as fuel, or sold.
- Propene is sold in various grades.
- Raw pyrolysis gasoline is sold as motor gasoline component after hydrotreatment, or it is used as feedstock for aromatics production.
- Tar is sold for fuel, used as coke feedstock, or diluted with hydrocarbon stocks to produce feedstocks for resins.
- Sulfur is recovered in some plants and sold.

5.3.2. Cracked Gas Processing

Many options are available for general process configuration [51, 54 – 60]. The sequence for isolation of hydrocarbon fractions can be varied, depending on plant size, amounts of ethylene and coproducts, impurities, product range, product purity desired, and other factors.

The following discussion of the recovery process is split into two sections:

1) Front-end section, including cracked gas compression and associated units
2) Hydrocarbon-fractionation section

Difference between the two sections is made for simplicity reasons. The basic process route for the front-end section is identical for all commercial processes, however major difference exists between gas- and liquid feedstock cracking. For the hydrocarbon fractionation section there are a large variety of process routes available, while the feedstock type results only in minor differences of the basic process route.

5.3.2.1. Front-End Section

The front-end section (Figs. 31, 32) receives the effluent stream from the cracking furnaces. Its main task is to cool the cracked gas stream, recovering waste heat and providing it to other plant sections, to condense and to produce dilution steam, and to remove heavy hydrocarbon components such as tar, oil, and heavy gasoline. The resulting purified gas is compressed in the cracked gas compression section and dried to prepare it for further cryogenic separation in the fractionation section. An acid-gas removal step that eliminates all CO_2 and H_2S from the cracked gas is incorporated between the final compressor stages.

The process configuration of this front-end section is fairly independent from licensors as they all follow the same process principals. However there are many variations in detailed design. Major differences are observed between plants processing gaseous feedstocks (ethane, propane, and butane) and those treating liquid feedstocks like naphtha and heavier fractions. Processing of cracked gas resulting from liquid feedstocks is more complex as higher amounts of heavy hydrocarbons are condensed and must be removed in this section.

Process Description. The main difference between the process for gaseous feedstock (Fig. 31) and liquid feedstock (Fig. 32) is the presence of a primary fractionator that by contacting with circulating oil removes tar and oily material ($bp > 200$ °C) from the cracked gas in the latter. In the case of cracking gaseous feedstock only,

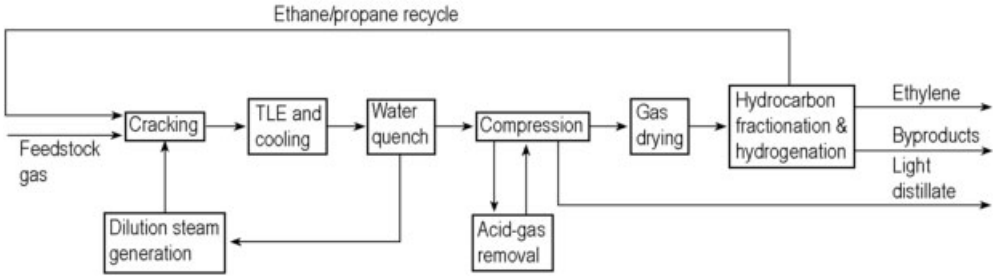

Figure 31. Simplified process flow diagram for the front-end process of an ethylene plant cracking gaseous feedstocks

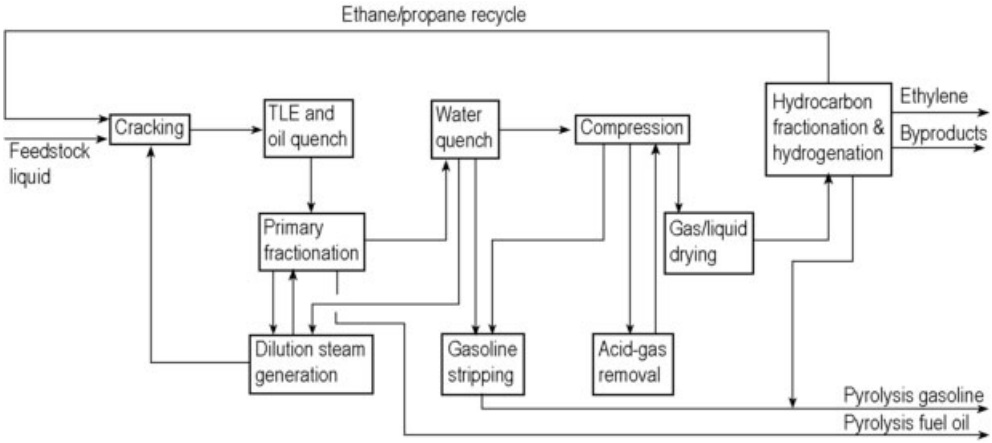

Figure 32. Simplified process flow diagram for the front-end process of an ethylene plant cracking liquid feedstocks

this step is not necessary, as the small amounts of heavy hydrocarbons present in the cracked gas can be removed in the water-quench column. Furthermore, additional installations are required in the water-quench system and in the compressor and drying units to treat the larger amount of pyrolysis gasoline that condenses in these units. The point when oil quench must be applied is reached, when substantial amounts of C_4 hydrocarbons are contained in the feedstock mixture.

Primary Fractionation. With liquid pyrolysis feedstocks, the primary fractionation column is the first step in the cracked gas processing route. Cracked gas enters the column typically at 230 °C. In the column it is contacted with circulating oil and, at the top of the column, with a heavy pyrolysis gasoline fraction obtained from the subsequent water-quench tower. Cracked gas leaves the top of the primary fractionator at ca. 100 °C, free of oil but still containing all the dilution steam. Hot oil, which functions as a heat carrier, is collected at the bottom of the column and, after cooling, recirculated as reflux to the middle section of the primary fractionator and to the quench nozzles downstream of the TLEs. Most of the heat contained in the cracked gas leaving the TLEs is removed by the circulating quench oil. Heat is reused for generation of process steam as well as for other purposes such as preheating of feedstock, preheating of process water, or for other energy consumers at an appropriate temperature level. Excess oil is removed from the primary fractionation system and sent out as product. Pyrolysis oil is generally produced as one fraction, containing all hydrocarbons boiling above 200 °C. However some plants produce two qualities of pyrolysis oil, one called PGO (pyrolysis gas oil) boiling in the range 200–400 °C and PFO (pyrolysis fuel oil) containing the heavier components.

Water-Quench Column. The overhead of the primary fractionator is routed to the water-quench column, in which the cracked gas is cooled to near ambient temperature by contacting it with a large stream of circulating quench water. In this cooling step most of the dilution steam and a heavy gasoline fraction are condensed and collected at the bottom. After gravity separation, heat is extracted in a circulating loop and most of the excess water stream is used to produce dilution steam. Large amounts of low-temperature heat (60–80 °C) are recovered and used for various heating purposes in the recovery section. The largest consumer of this low temperature heat is typically the reboiler of the propene–propane fractionation tower. Most of the condensed gasoline is recycled as reflux to the top of the primary fractionator. The excess gasoline fraction is sent to a gasoline hydrotreater or fractionated further into higher quality gasoline and heavier oil streams. Both water and gasoline fractions, except when fractionated further, are stripped of dissolved light gases [51, 54, 61, 62].

In the case of processing cracked gas from gaseous feedstock, the furnace effluent stream is directly fed to the water-quench tower after cooling in a secondary TLE to ca. 200 °C. In this case the system is simpler as no or only small amounts of tar and oil condense and must be removed. A general problem in all quench towers is the separation of hydrocarbons and water. Tarry material with a density similar to water, which is present in small quantities, tends to accumulate in the system and to stick to the surface of the column internals. On-line or off-line washing with high-boiling, aromatic-rich solvents is applied to extend the on-stream time of the system up to five years. There is furthermore a strong tendency for formation of emulsions, leading to fouling problems in the generation of dilution steam and in the processing of the wastewater from the system. Emulsification is avoided by careful control of the pH value in the quench-water system.

Numerous options exist in the design and operation of the quench column and primary fractionator units for removing pyrolysis tars, recovering various heavy hydrocarbon streams, and isolating quench oil for recycle [51, 61, 62].

Dilution-Steam System. The dilution steam required in the cracking process forms a closed loop over the cracking furnaces and the quench-water column. Excess water from the bottom of the quench column is stripped in a process water stripper to remove volatile hydrocarbons. Some 5–10 % of the process water is purged to avoid concentrating dissolved solids. The generation of the process steam is typically performed at 800 kPa. In liquid feedstock processes, 50–80 % of the dilution steam is generated by recovering heat from the hot quench oil cycle from the bottom of the primary fractionator. The balance heat duty is provided by condensing steam. Dilution steam is typically generated in dedicated heat exchangers and routed to the cracking furnaces where it is mixed to the feedstock at appropriated position in the convection section.

Some licensors use a different approach for the generation of dilution steam in gaseous feedstock processes. Ethane - propane feedstock is saturated in a dedicated column by direct contact with circulating water. Heat input is performed by heating the circulating water and the water - hydrocarbon ratio is adjusted by controlling pressure and temperature in this column [63].

Compression. After cooling and purification of the cracked gas in the primary fractionator and/or in the quench tower, further processing requires compression to ca. 3200–3800 kPa. Some condensation of water and hydrocarbons occurs during compression, and in the later stages, acidic gases are removed. Compression is typically performed by a turbine-driven centrifugal compressor in four to five stages with intermediate cooling. The number of stages depends primarily on the cracked-gas composition the temperature level of the cooling medium, and the highest temperature allowed for interstage discharge. Interstage cooling and temperature control keep the cracked gas below 100 °C to prevent diolefin polymerization and subsequent equipment fouling. Various injection systems, using aromatic rich oil or water, are applied to reduce fouling of compressor internals and intercooling equipment. Interstage cooling usually employs water coolers to minimize inlet temperature in the subsequent stages. First-stage compressor suction is normally maintained somewhat above atmospheric pressure to prevent oxygen intrusion. The operating pressure is set to economically balance yield in the cracking coils against higher energy consumption for compression [51]. The compressor dis-

charge pressure is set by the choice of process design and by the choice of refrigerant. With a high-pressure demethanization methane condenses in the demethanizer overhead condenser. Condensation temperature varies with hydrogen/methane ratio. The discharge pressure, with the coldest ethylene refrigerant level typically at ca. −100 °C, is normally 3200–3800 kPa. Processes involving low-pressure or medium-pressure demethanizers can use discharge pressures of 2000 kPa only, however, in such cases the total compression power is fairly constant, because the power is just shifted from cracked-gas compression to refrigerant compressors. The criteria used to determine suction pressure and discharge pressure are selectivity and energy consumption, and equipment costs. [51, 62]

Interstage cooler pressure drops must be kept low, especially in the initial stages, to reduce compressor and energy costs. The water and hydrocarbon condensates may be routed to another compressor stage, to the quench column preceding the compressor, or to a fractionation tower. In liquid cracking most of the gasoline fraction containing the C_6 to C_8 aromatics is condensed in the interstage coolers of the compressor. After gravity separation from process water this gasoline stream is fractionated in a stripper to remove C_4 and lighter components before routing it to a hydrotreating step. Drying is required for all liquid and gas streams subjected to temperatures below 15 °C [51, 54, 60, 61].

Acid-Gas Removal. Before further processing, carbon dioxide and hydrogen sulfide are removed from the cracked gas by once-through (caustic wash) and regenerative solvent (MEA) scrubbing. Sulfur may also be removed from the cracker feedstock. [51].

The acid-gas removal system is typically located between the 3rd and 4th, or between the 4th and 5th stages. In all process configurations acid-gas removal must be located upstream of the drying unit in order to avoid formation of ice and hydrates in the following fractionation steps.

Carbon dioxide is removed because it can freeze at low temperature in heat-exchange and fractionation equipment. Carbon dioxide can also be absorbed into ethylene, affecting product quality and further processing. Hydrogen sulfide is corrosive and a catalyst poison as well as a potential product contaminant [53, 54]. CO_2 and H_2S concentrations in the overhead stream of the caustic wash are typically below 0.2 ppm.

These acid gases are scrubbed with sodium hydroxide on a once-through basis or in combination with a regenerative chemical. Regenerative pre-scrubbing before final sodium hydroxide treatment is applied at high-sulfur feedstocks. This reduces sodium hydroxide consumption, which is desirable for economic and environmental reasons, however, it requires the handling of a sulfur-rich off-gas stream.

Regenerative scrubbing can employ alkanolamines, but it is not possible with this treatment alone to lower the concentration of CO_2 to that required to meet standard ethylene specifications(e.g., to 0.2 ppm). In any case a polishing scrubber consisting of a caustic wash unit is applied to achieve the required CO_2 specification in the effluent stream. Acid-gas removal systems are frequently subject to fouling, especially at high concentrations of C_4 and C_5 diolefins. Insiders talk about "yellow oil" or "red oil", an unpleasantly smelling polymeric material. The fouling process can be controlled by applying special inhibitors.

Acid gases liberated from the regenerative solvent can be incinerated or recovered [51, 54, 61, 62, 64]. Waste caustic, containing sulfides and carbonate in the range of a few % is usually oxidized and neutralized in a wet caustic oxidation unit or neutralized and stripped, before feeding it to the sewer system. Spent caustic is the most contaminated liquid effluent in the ethylene plant, because it is the outlet for various trace components accumulating in the process. In addition to the sulfur and carbonate load the stream further carries traces of dissolved hydrocarbons, organic acids, ketones, alcohols, and aldehydes.

Drying. The cracked gas is saturated with water before compression and after each intercooler stage. Moisture must be removed before fractionation to prevent formation of hydrates and ice. Typically, this is accomplished by chilling and by adsorption on molecular sieves. Older plants also used absorption by a glycol scrubbing system or adsorption on alumina [53]. Drying is arranged before the first fractionation step, typically after the last compression stage. Before feeding the compressor effluent to the molecular sieve dryer it is cooled in a water cooler and subsequently by propene refrigerant. A temperature

of 15 °C upstream of the dryer is desirable for removing as much water as possible to reduce load on the dryers. However, the temperature must remain above the hydrate-formation point, which lies in the range 10–15 °C [51, 61]. Water is adsorbed on molecular sieves, preferably after the highest compression pressure is reached to minimize adsorption costs. Higher pressure allows smaller dryer volume with lower adsorbent cost and less water removal because the water content in the gas stream decreases with increasing pressure. Multiple adsorption beds make continuous water removal possible. One or more adsorption beds are in operation while at least one unit is being regenerated. Regeneration requires the use of heated gas that is free of polymerizable components. Byproduct methane and hydrogen used as fuel are typically good choices for this gas [51 – 54, 61, 62]. In liquid-feedstock cracking processes large amounts of hydrocarbons are condensed at 15 °C. If these liquid hydrocarbons are to be fed to the fractionation column they are also dried over molecular sieve adsorbers.

5.3.2.2. Hydrocarbon Fractionation Section

The fractionation section receives the compressed cracked gas at a pressure of 3200–3800 kPa for further fractionation into different products and fractions at specified qualities and battery limit conditions. Over the history of design and construction of ethylene plants many different process schemes have been developed and implemented for this task.

Cryogenic separation is the predominant method for cracked-gas separation. Although gas separation processes via adsorption, absorption, or membrane technology has made remarkable progress in the recent past, it has not found major application in the ethylene industry. Rare applications of these methods can be found in very special cases like revamps or pretreatment of off-gases from polymerization plants or from refineries, containing valuable components, before being sent to the ethylene plant.

Today, three processing routes have gained commercial importance, with the main characteristics being the first separation step and the position of the hydrogenation of the acetylene contained in the cracked gas. These three process sequences are:

1) Demethanizer first with tail-end hydrogenation
2) Deethanizer first with front-end hydrogenation
3) Depropanizer first with front-end hydrogenation

These basic processes differ in the configuration of process steps for any given feedstock, but provide similar overall capabilities for fractionation and hydrogenation. The differences are primarily in the sequence of fractionation and hydrotreating steps downstream of cracked-gas compression [51].

Process Description. Compressed cracked gas can be:

1) Condensed and fed to a demethanizer column (front-end demethanizer), as shown in Figure 33
2) Condensed and fed to a deethanizer (front-end deethanizer), as shown in Figure 34
3) Removed from an intermediate compressor stage, condensed, and depropanized in a column, whereby the light gas is returned to the last compressor stage (front-end depropanizer), as shown in Figure 35

Combinations are possible; for example, intermediate compressor stage condensate and vapor are sent to different parts of the plant [57, 61].

Traditionally the demethanizer first/tail-end hydrogenation process sequence was used by American contractors. With the need for reduction of energy consumption in the separation section and process simplification for cost reasons, the deethanizer first/front-end hydrogenation process, which had long been used by Linde, was also applied by some American contractors. A comparison of all routes has shown that this sequence seems to have the best energy efficiency [65]. New developments in this area are the combination of the advantageous sequence with special equipment (dephlegmators) [66] and the application of washing steps for demethanization [67].

Front-End Demethanizer (Fig. 33). The hydrocarbon-separation sequence begins with removal of methane and lighter components, primarily hydrogen, from higher molecular mass components. Hydrogen may be removed before or after methane. The bottom product is routed to

a deethanizer column where acetylene, ethane, and ethylene are removed as overhead product, and C_3 and heavier components as bottom product.

Typically, the overhead product from the deethanizer is first hydrotreated to convert acetylene to ethylene and ethane. This treated overhead mixture is then fractionated to ethylene product and ethane recycle.

The bottom product from the deethanizer is depropanized to separate C_4 and heavier components as a bottom stream; methylacetylene, propadiene, propane, and propene are taken as an overhead stream. This overhead fraction is hydrogenated to remove methylacetylene and propadiene, which can also be recovered by distillation or absorption and stripping. In the propene fractionation step, propene is recovered for sale and propane for recycle. The propene product can be of chemical grade or polymer grade with additional purification. Depropanizer bottoms are sent to a debutanizer for separation into other product streams (C_4 materials, raw pyrolysis gasoline, C_5 materials, aromatics, etc.) [51, 53 – 55, 60, 61].

Front-End Deethanizer (Fig. 34). Deethanization of dried cracked gas is used as the first fractionation step to produce an overhead stream of ethane and lighter components, and a bottom stream of C_3 and heavier materials; subsequently, the lighter components are hydrogenated to remove acetylene and fractionated cryogenically. Because this overhead stream is hydrogen rich, no supplementation is necessary for hydrogenation. During the chilling and demethanizer operation, the overhead is separated into hydrogen and methane; the bottom stream is fractionated to yield ethylene product and ethane; and the ethane is recycled to the cracking furnaces.

The bottom stream from the deethanizer is treated in the same way as in the front-end demethanizer system described above.

Front-End Depropanizer (Fig. 35). Depropanization of dried cracked gas is used as the first fractionation step to produce a propane and lighter component overhead stream and a C_4 and heavier bottom stream. The depropanizer overhead stream is compressed and hydrogenated to remove acetylene, and to partially hydrogenate methylacetylene and propadiene. It is then chilled and demethanized, and the bottom fraction of demethanization is deethanized. The deethanizer overhead components are fractionated to ethylene product and ethane, which is recycled as cracking feedstock. Deethanizer bottom product is hydrotreated to remove methylacetylene and propadiene, and fractionated to propene product and propane, which is recycled to the cracking furnaces. The depropanizer bottom components are fractionated to produce C_4 materials and raw pyrolysis gasoline [54, 55, 60, 61].

For all three of the above-mentioned basic processes numerous process variants exist. For liquid cracking acetylene hydrogenation is always performed in the C_2 stream, in the C_{2-} or in the C_{3-} stream. In the case of cracking gaseous feedstock with little coproduction of C_4 hydrocarbons the hydrogenation step can also be arranged in the total cracked gas stream during or just after cracked-gas compression (after or before drying). This is not feasible for cracked gas streams resulting from liquid cracking because 1,3-butadiene, typically a valuable byproduct, would be cohydrogenated almost completely to butene and butane.

Acetylene may also be recovered for sale by an absorption process. This absorption process step is located in the gas phase upstream of the ethylene fractionation column. N-methyl-2-pyrrolidone, acetone, and N,N-dimethylformamide are typical solvents used for acetylene recovery. The acetylene absorption tower yields a very pure ethylene product, because acetylene is removed practically down to the detection limit.

Acetylene recovery, however, is becoming more and more a niche application due to the disappearing importance of the acetylene chemistry and due to the limited acetylene production rate that can be obtained from a large-scale plant (ca 10 000 t/a) This quantity, however, on the other hand is too large to be totally used by the welding industry.

Pre-fractionation Feed Chilling. Before being fed to any of the fractionation towers, dried cracked gas must be cooled and partially condensed. Dried cracked gas is cooled by the refrigerant systems and by the low-temperature product and recycle streams leaving the cold fractionation equipment down to a temperature of typically -140 to -160 °C. At this temperature C_2 components are entirely removed by partial condensation, and the noncondensed gas stream

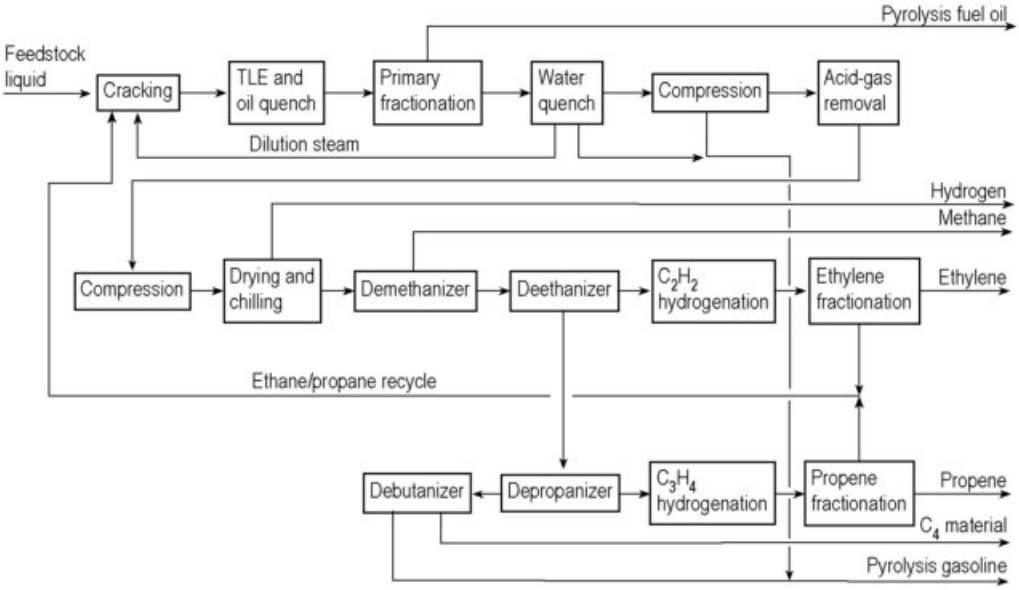

Figure 33. Simplified process flow diagram for the production of ethylene by liquid cracking with a front-end demethanizer

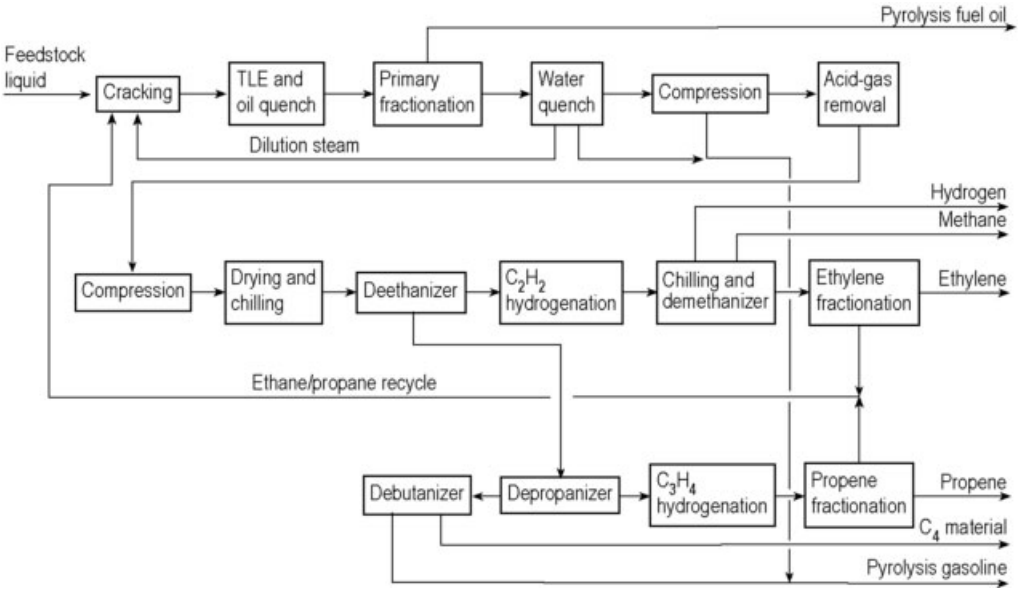

Figure 34. Simplified process flow diagram for the production of ethylene by liquid cracking with a front-end deethanizer

has a hydrogen concentration of 80–95 vol %. The type of refrigerant, number of refrigeration levels, and the design of the refrigeration and heat-exchange systems depend on the temperature and pressure required in the downstream fractionation equipment. Fractionation and purification consume much energy. Cryogenic purification of methane, ethane, and ethylene requires costly refrigeration. Design optimization is complex because:

1) The most efficient routing must be considered for the condensates and vapor streams from each stage of the cool-down sequence in the

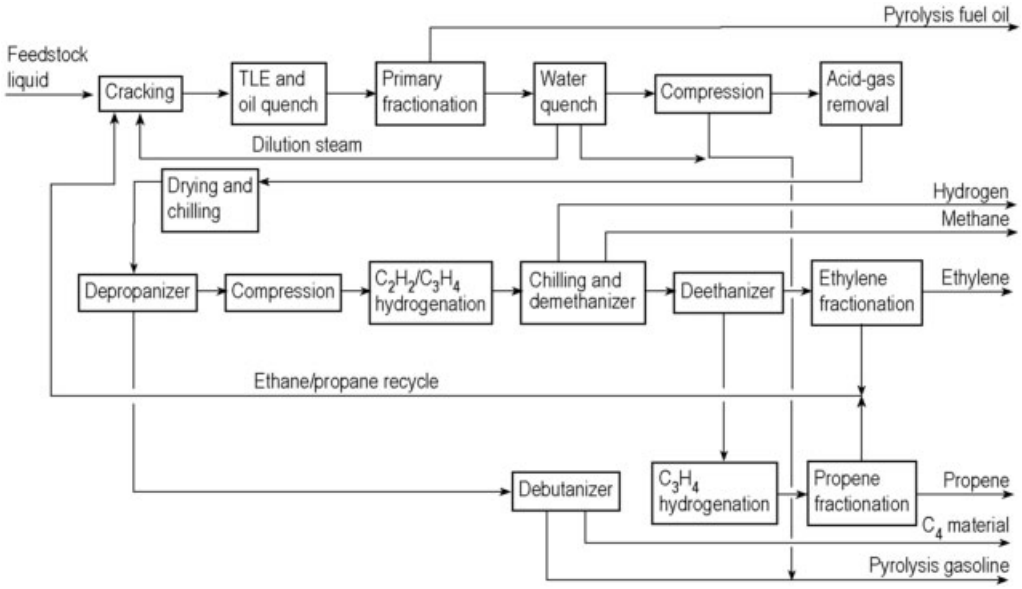

Figure 35. Simplified process flow diagram for the production of ethylene by liquid cracking with a front-end depropanizer

individual refrigeration loops (e.g., 2–6 stages each)

2) The energy requirements for fractionation must be determined
3) The freezing of components, such as acetylene, butadiene, and benzene, must be avoided

When acetylene content is high, prevention of freezing and safe hydrogenation are important. Also of importance from a safety standpoint for the chilling section and downstream equipment is the choice of appropriate construction materials for cryogenic conditions [51, 53]. Process equipment that may accidentally be exposed to cryogenic conditions should also be carefully studied. National standards are available in most countries to assist in making appropriate selections [51, 53].

Demethanization. Demethanization of cracked gas separates methane as an overhead component from C_2 and heavier bottom components; concurrently, hydrogen is removed from the cracked-gas stream and may be obtained as a product by purification before or after demethanization. In case of naphtha cracking hydrogen is recovered at high purity because it is needed in the plant itself for various hydrogenation steps and because any excess can typically be sent out when the plant is embedded in a refinery environment. In gas-cracking plants hydrogen is only required at low quantities. The bulk of hydrogen is recovered together with methane, forming the tail-gas stream, that is sent to the fuel system after expansion.

Methane is typically used as plant fuel or sold; C_2 and heavier components are sent to the recovery system. This overall fractionation is important from a design and operational standpoint because of its high share of the capital, energy, and fuel consumption of the ethylene plant. Demethanization typically consumes the greatest proportion of net energy from the refrigeration system (see Fig. 36 and discussion of refrigeration, Section page 47) [56, 64].

In the demethanizer system and the associated process for the isolation of high-purity hydrogen the lowest temperatures in the plant (down to −165 °C) prevail.

The industry offers a large variety of combinations of process steps to achieve the desired separation result. All processes consist of a combination of cooling steps, partial condensation, and a demethanizer tower system. Liquids, knocked out during cooling are fed to the demethanizer at an appropriate position. Some licensors apply special equipment like dephlegmators or recontactors in this process configuration to perform a preseparation of the material

before sending it to the column. Both dephlegmators and recontactors incorporate an additional number of theoretical equilibrium stages in the cooling train contributing to the demethanization.

Demethanizer systems operate in a wide pressure range between 1000 and 3500 kPa (LP, MP, and HP demethanizers), each of these having their specific characteristics.

A special challenge in this process is the provision of the respective chilling duty for individual cooling steps. The classical method for generating chilling duty down to a temperature of −100 °C is the refrigerant cascade of ethylene and propene. Temperatures below −100 °C are achieved by one of the following methods:

- Adding another element in the refrigerant cascade, namely a methane refrigerant cycle
- Generation of a light recycle stream, i.e., vaporizing condensed cracked gas fractions and recycling the vapors to the cracked gas compressor
- In the case of gas cracking, hydrogen is not recovered at high pressure and can be depressurized to fuel gas pressure in a turboexpander, generating chilling duty at very low temperature
- Applying a mixed refrigerant cycle that covers the full temperature range

[58, 62].

The demethanizer bottom stream contains C_2 and heavier components. With the front-end demethanizer, components from C_2 through pyrolysis gasoline may be present. However, with front-end deethanization, C_{3+} components are removed, and with front-end depropanization, C_{4+} components are removed. The demethanizer design changes significantly as heavier components are eliminated, especially with regard to the refrigeration needed to condense the incoming cracked-gas feed stream and to reboil the bottom stream.

Deethanization. Deethanization of cracked gas separates acetylene, ethylene, and ethane as overhead components from C_{3+} bottom components. In newer plants, low-pressure fractionation may be used, or feed streams from the cracked-gas chilling section can be supplied as multiple feeds to the deethanization tower. The deethanizer is usually the third largest user of refrigeration energy in a front-end demethanizer ethylene plant (Fig. 36) [51, 61, 62, 64].

With front-end demethanization and front-end depropanization, the feed stream to the deethanizer comes primarily from the demethanizer. With front-end deethanization, the deethanizer feed comes from the chillers that follow the gas dryers. In the simplest cases, single feed streams from these units are separated into overhead and bottom fractions for further processing. [51]. Dual-pressure deethanizer systems improve energy consumption and reduce the bottom temperature, thus reducing fouling in reboilers. As with the demethanizer, these designs conserve refrigeration energy but increase capital costs and operating complexity. The overhead product condensation temperature is typically ca. 0 to −50 °C, based on a column operating pressure of 1500–2700 kPa [54, 55, 60].

Depropanization. of cracked gas separates propane and lighter fractions as overhead components from C_{4+} fractions as bottom components. Feed streams from compression, and sometimes streams from the gas-chilling section, can be routed to the depropanizer, which, compared to other fractionation units, requires less refrigeration because of its higher operating temperature (Fig. 36) [64].

The feed stream to the depropanizer comes primarily from the deethanizer when front-end demethanization and deethanization are employed. With front-end depropanization, the feedstock is taken from the gas dryers. In the simplest case, the dried cracked-gas feed stream is separated into a propene–propane overhead stream, which contains small amounts of methylacetylene and propadiene, and a bottom stream which contains components from C_4 materials to raw pyrolysis gasoline. Energy requirement is reduced by mixing smaller feed streams removed from intermediate compression or gas-chilling stages with the main feed stream, or by having them enter separately at a point in the depropanizer consistent with the composition of column components. Condensate is usually stripped before entry into the depropanizer [51, 62].

Operating pressure and temperature of the depropanization step in ethylene plants, even downstream of the deethanizer, are kept low to reduce fouling of the fractionator by acetylenic

and diolefinic contaminants. Hydrogenation of cracked gas in the compressor zone lowers the amount of these polymerizing contaminants substantially, but also destroys most of the butadiene [51, 55, 60].

Debutanization. Debutanization of cracked gas separates C_4 materials as overhead components from C_5 and heavier fractions as bottom components. The feed stream comes typically from the depropanizer column or is a depropanized condensate stream from the compressor. The mixed C_4 overhead fraction may be sent to an extraction unit for removal of 1,3-butadiene and further processing. The bottom stream and the debutanized fraction from the compression section can be used as pyrolysis gasoline feedstock or processed to produce aromatics and specialty chemicals feedstocks [51].

Acetylene Hydrogenation. Cracked gas is selectively hydrogenated to remove acetylene from ethylene, except when acetylene is recovered for sale. Hydrogenation is performed with palladium-based catalysts. Ethylene specification requires less than 1 ppm of acetylene, thus acetylene content at the outlet of the hydrogenation must be < 0.5 ppm. Pressure ranges typically from 20 to 35 bar and the temperature from 25 to 100 °C. There are two basic processes for the hydrogenation of acetylene.

Tail-End Hydrogenation. Acetylene hydrogenation takes place in the gas phase of a pure C_2 fraction, consisting of ethylene, ethane, and acetylene only. The unit is typically incorporated after demethanization, upstream of the C_2 fractionation tower. Hydrogenation reactors consist of one or two adiabatic beds. Hydrogen is added upstream of the reactor at a molar H_2/C_2H_2 ratio of ca. 1.5–2, resulting in an ethylene gain of up to 50 % of the acetylene present. Hydrogen and small amounts of methane and CO introduced after demethanization must be removed by additional fractionation. Sometimes purging the light ends from the overhead ethylene fractionator condensate is possible, but additional multistage fractionation is usually required in the ethylene fractionator or in a small separate demethanizer. A useful design provides five to ten overhead stages above the product drawoff point in the ethylene fractionator column for further separation of the light ends. The overhead condensate is returned to the top of the tower, and the light ends are removed from the overhead partial condenser for reprocessing [51, 56, 61, 64].

Front-End Hydrogenation. Acetylene is hydrogenated:

1) In the total cracked gas stream after or before drying
2) In the C_{3-} stream with front-end depropanizer
3) In the C_{2-} stream with front-end deethanizer

In the case of (1) and (2) any methylacetylene, propadiene and butadiene, present in the gas stream is cohydrogenated. (methylacetylene ca. 80 %, propadiene 30 %, butadiene > 90 %).

Although hydrogenation in the cracked gas (1) may offer advantages for the lighter feedstocks, disadvantages exist with heavier feedstocks. In particular, the volume of gas treated at this stage is high; butadiene is hydrogenated and lost; and a single hydrogenation to concurrently remove acetylene, methylacetylene, and propadiene to the required levels without significant olefin loss is not possible. Hydrogenation of individual mixed C_2 and C_3 streams after deethanization and depropanization, respectively, appears to permit more efficient use of catalyst and result in higher product yields. However, proponents of the front-end deethanization and depropanization process with front-end acetylene hydrogenation report that bed sizes are equivalent for that reaction. Higher hydrogen concentration and improved catalysts allow higher gas space velocities than can be achieved downstream [61].

General advantages of the front-end hydrogenation versus the tail-end hydrogenation unit are: (1) no hydrogen make up is required as hydrogen is already present in the feed stream; (2) longer run length between catalyst regenerations; (3) No removal of light ends, as required in tail-end hydrogenation, is necessary.

Front-end hydrogenation is typically conducted in multistage adiabatic beds with intercooling, or less frequently in a single tubular reactor that allows an isothermal reaction.Due to the high concentration of hydrogen and ethylene in the process gas front-end hydrogenation requires a very selective catalyst, that promotes the hydrogenation of acetylene only.

Acetylene hydrogenation requires heat-exchange systems for incoming low-temperature process streams (2) and (3). Hydrogenation in the compressor section (1) does not require the heat-exchange system necessary for cryogenic streams. Most of the troublesome polymerizable components are removed before fractionation, although rapid fouling of the hydrogenation bed must be avoided. Hydrogenation may be necessary before chilling if a high acetylene concentration in the cracked gas would result in plugging or in unsafe downstream operation [51, 53, 61].

Although tail-end hydrogenation is the most common process in currently operating plants, it seems there is a strong tendency to increasingly apply the front-end hydrogenation system in new projects. This tendency is supported by a continuous improvement of front-end hydrogenation catalysts. New catalysts show an increased operating temperature window in which only acetylene or dienes are selectively hydrogenated. Furthermore, new catalysts are less sensitive toward CO. These improvements result in higher operational stability decreasing the danger of reaction runaway, where in addition to acetylene also ethylene participates in the reaction. The shapes of catalyst particle were typically little spheres, tablets, or extruded particles of a size of a few millimeters. New catalysts are available with different geometrical structures such as rings or other shapes that result in lower pressure drop. This contributes in overall energy saving, because the front-end hydrogenation is located in the main cracked gas stream where this pressure drop directly affects the cracked gas compressor outlet pressure.

Ethylene Fractionation. Ethylene fractionation separates ethylene as a high-purity overhead product ($>$ 99.9 wt %) from ethane, which is combined with propane and recycled for cracking. For this difficult fractionation, the net work done by the refrigeration system is high because of the high reflux and low temperature required (Fig. 36) [51, 64]. Fractionation requires a high reflux ratio (ca. 4) and as many as 90–125 separation stages [54, 55, 60, 61, 64]. The ethylene purity desired is typically $>$ 99.9 wt %. The feed stream comes from the acetylene-removal step that follows chilling, or from the demethanizer or deethanizer.

There are basically two processes in commercial use: high-pressure fractionation and heat-pumped C_2 fractionation.

High-Pressure Fractionation. High-pressure fractionation, operating in the pressure range of 1700–2800 kPa, is typically used in tail-end hydrogenation systems. Heating and cooling of the column is integrated into the propene refrigeration cycle. Low-pressure propene refrigerant is vaporized in the top condenser, providing chilling duty. Vaporized propene is typically compressed in two stages of the refrigerant compressor. After compression, propene vapors are condensed in the reboiler, providing heat duty. By means of this system the propene refrigerant compressor acts as an indirect heat pump, shifting heat from the condenser of the column to the reboiler. The top of the column is equipped with some pasteurization trays to remove light ends introduced when acetylene is hydrogenated in the feed stream to this column. Ethylene product is withdrawn as a liquid side stream some trays below the top, pumped to the required product pressure, and delivered to battery limits after vaporization and heating.

Heat-Pumped C_2 Fractionation. In heat-pumped C_2 fractionation, the column operates at a pressure of ca. 800 kPa. Overhead gases of the column, consisting of on-spec ethylene, are routed to a compressor after heating to ambient temperature in a feed–effluent exchanger and are compressed to ca. 2000 kPa. Part of the compressed ethylene vapor is cooled and routed to the reboiler where it is condensed, providing heat duty for reboiling. The condensed ethylene is recycled as reflux to the top of the column. Ethylene product is discharged to battery limits from the discharge side of the compressor. Heat-pumped ethylene fractionation can use a dedicated compressor. More common is the integration of the heat-pumped ethylene splitter in the ethylene compressor, using the 3rd stage of this compressor as a heat pump.

This process has some advantages over the high-pressure fractionation process with respect to investment and energy consumption. Less equipment is involved, and because of the better relative volatility of ethylene and ethane at lower pressure, a lower reflux ratio and less column trays are required,altogether resulting in lower power consumption, smaller column size, and generally less equipment. However it is not fea-

sible to use the heat-pumped process in combination with tail-end hydrogenation because of the light ends present in the feed stream.

Ethylene is generally discharged to battery limits as a gas. However, production of up to 50 % of production as the liquid phase and the use of storage facilities is also practiced.The ethylene plant is typically equipped with ethylene product storage capacity in order to balance production and consumption of downstream plants. Two storage types are commercial practice:

Semirefrigerated storage at about 1500 kPa and −40 °C, using spherical containers of a size of up to 3000 m^3 (for smaller storage inventories).

Atmospheric storage tanks, operated at atmospheric pressure and −103 °C. The size of these tanks is not limited, typically in the range of 10 000–20 000 m^3. Tank systems are usually equipped with liquefaction facilities in order to handle boil-off gases and incoming flash gases. Product vaporization facilities for the full capacity of the plant are provided in order to continue ethylene supply to downstream users in the case of a plant upset.

Hydrogenation of Methylacetylene and Propadiene (MAPD); C_4 Fraction; and Pyrolysis Gasoline.

C_3 Hydrogenation. The C_3 fraction coming from the top of the depropanizer typically contains 2–6 % of methylacetylene (MA) and propadiene (PD). For removal of these components from the propene product and for economic reasons MAPD is hydrogenated to propene and propane. Hydrogenation is performed in gaseous phase in multistage adiabatic bed reactors with intercooling steps or, more frequently, in liquid phase in a single trickle-bed reactor. Tubular reactors are applied in rare cases. The heat of reaction is removed in the trickle-bed reactor by partial vaporization of the processed liquid, this being condensed later by cooling water. Tubular reactors use a branch stream of the propane refrigerant compressor as coolant. Modern hydrogenation processes typically yield a net propene gain of 60 % of the MAPD present in the feed stream. The balance is converted to propane, which is recycled to the cracking furnaces as cracker feedstock.

MAPD concentration at the reactor outlet is typically 500–1000 ppm, when the propene product has polymer-grade quality. This figure results in high selectivity of the reaction. Residual MAPD is removed in the propene splitter along with propane by distillation. If only chemical-grade propene is required, the MAPD concentration after hydrogenation must be below 20 ppm as the propane–propene splitter will not be able to reduce MAPD down to the required level due to reduced reflex rate Recovery of MAPD is theoretically possible, applying a similar system as that one for acetylene recovery; however it is practically not used because of lacking demand for this product.

C_4 Hydrogenation. The C_4 cut from the top of the debutanizer typically contains small amounts of vinylacetylene and ethylacetylene, ca. 50 % 1,3-butadiene, and a few percent of butanes; the balance consists of various types of butenes, of which isobutene is the most important. Depending on the nature of the downstream processes it may be necessary to hydrogenate some of the unsaturated components. The following hydrogenation steps are in commercial use:

1) Selective hydrogenation of ethylacetylene and vinylacetylene to butadiene
2) Selective hydrogenation of butadiene to butenes
3) Full hydrogenation of all unsaturated to butane
4) Hydroisomerization to maximize 2-butene yield

Process (1) is applied to increase the 1,3 butadiene yield; process (2) is used if no use can be made of butadiene and to increase the butene yield; process (3) is employed if the C_4 cut is to be recycled as cracker feedstock to improve olefin yields or used as LPG product. Hydrogenation is performed in tubular reactor systems or, in more modern processes, in trickle-bed reactor systems. By selecting catalyst type and process parameters, the hydrogenation result with respect to yields of different components, mainly the various types of butenes, can be controlled in a certain range.

C_3 and C_4 hydrogenation units consist typically of adiabatic beds or tubular reactors, integrated in a system of coolers, recycle pumps and separation drums. New developments try to integrate the reaction step into the separation column, e.g., the depropanizer or debutanizer, ap-

plying the so called "Catalytic Distillation Process". In this application the catalyst is installed in the separation column and is acting not only as catalyst but also as separation element, providing some theoretical equilibrium stages. Process owners report a number of advantages for this process step. However, industry is still reluctant to apply this new process in large-scale plants [69, 70].

Pyrolysis Gasoline Hydrogenation. Pyrolysis gasoline — the C_5 to C_{10} hydrocarbon product fraction — consists mainly of aromatics. The nonaromatics are mainly unsaturated hydrocarbons with a high portion of acetylenes and dienes. This stream is unstable and cannot be stored, as the unsaturated components react further, forming polymers and gum. Depending on the downstream processes pyrolysis gasoline is hydrogenated and fractionated in different steps. The most common process route is as follows:

1) Selective hydrogenation of the total gasoline to hydrogenate acetylenes, dienes, and styrene to olefinic compounds. After stabilization and removal of oil this stream is suitable for use as motor fuel. The reaction is typically controlled in a way that the residual styrene value is ca 0.5 %.
2) Fractionation of the effluent of the 1st stage hydrogenation into a C_5 cut, a C_6–C_8 heart cut, and a C_{9+} cut.
3) The C_6–C_8 cut is further processed in a 2nd stage hydrogenation step to convert olefins to paraffins and naphthenes and to convert all sulfur to H_2S, which is removed from the product in a downstream stripper. this process is necessary to prepare the heart cut for aromatics recovery. The specification is controlled via the bromine number, which is typically 0.5.

C_2, C_3, and C_4 hydrogenation as well as 1st stage, selective gasoline hydrogenation use palladium-based catalyst. Palladium catalyst, however, are sensitive with respect to metallic poisons such arsenic or mercury, that may be present in the cracked gas in small traces, originating from feedstock contaminants. If such contaminants are to be expected, the Pd catalyst must be protected by guard beds, consisting of various metallic oxides like CuO, ZnO, PbO, impregnated activated carbon, and other materials. The 1st stage gasoline hydrogenation catalyst, however, should preferably be based on nickel metal if these contaminants are present, because nickel catalysts are much less sensitive to them.

Propene fractionation separates propene as a chemical-grade overhead product (typically 93–95 wt % min.) or more frequently as polymer-grade propene (> 99.5 wt %) from propane. This separation to polymer grade propene requires typically 150–230 stages and a reflux ratio of 20 because of the close boiling points of propene and propane. Two basic processes are applied for this difficult separation task: polymer-grade fractionators operate at ca 1800 kPa, with cooling water in the overhead condenser and hot quench water in the reboiler. In the case of naphtha cracking and where sufficient waste heat is available from the hot quench water cycle this is the most economic process. If no waste heat for reboiling is available, or where upstream cooling water temperature is higher than 35 °C a heat-pumped propene fractionator is applied. In this case the fractionator operates at ca. 800–1100 kPa, and the overhead gases are compressed and condensed in the reboiler. By means of this process the condenser heat is raised to a higher temperature level, and further use is made of it in the reboiler. The advantage of this system is that no external heat for reboiling and only small quantities of cooling water are required. The fractionation design depends on the concentration of propene in the feed, which varies from 70 % for propane cracking to 95 % for naphtha cracking [61, 62, 64].

The heat pump is typically driven by a dedicated compressor, in particular cases the heat pump is integrated into the propene refrigeration cycle combining heat pump and 3rd stage of this machine.

The feed stream to the propene fractionator usually comes from the methylacetylene and propadiene hydrogenation step. Excess hydrogen and methane are purged in an upstream stripper or as light ends in the column overhead or from the overhead condensate [54, 55, 60 – 62, 64].

Hydrogen Purification. Hydrogen is produced in the cryogenic section of the plant at a purity of typically 80–95 vol%. However, the product stream contains some 1000 ppm of carbon monoxide. Hydrogen product is required in

the ethylene plant as make-up stream for the hydrogenation units. Excess hydrogen is typically a desired product. However, hydrogen must be free of carbon monoxide as CO is a poison for all hydrogenation processes. Today two processes are in commercial use in ethylene plants to purify hydrogen: (1) methanation of CO in a catalytic process, converting CO to methane and water, after which the effluent must be dried in adsorber beds; (2) hydrogen purification by adsorption in a pressure-swing adsorption unit, producing CO-free high-purity hydrogen.

5.3.3. Utilities

Refrigeration. Refrigeration in ethylene plants is important and costly. Refrigeration optimization is vital in plant design. Typically, two different refrigeration systems are employed, namely, ethylene and propene (or propene–propane), each of which generates chilling duty at two to five different temperature levels. Propene refrigeration is designed as a closed-loop system in which the heat removed by cooling water or by air cooling by condensing the refrigerant after the final compression stage. Ethylene refrigeration systems are designed as closed- and open-loop systems. Open-loop cycles integrate the ethylene-fractionation tower in the refrigeration system. Propene and ethylene refrigerant systems act as a chilling cascade. Heat removed from the process by the ethylene cycle at temperatures below −50 °C is shifted to the propene cycle by condensing high-pressure ethylene refrigerant by means of vaporizing low-pressure ethylene refrigerant. Refrigerant temperatures are chosen to accommodate diverse plant needs most efficiently. Methane refrigeration by simple expansion (not employed as a refrigeration loop) generates additional cooling and especially the lower temperatures needed to purify hydrogen for special purposes.

Some processes add a third refrigeration cycle to the cascade, using methane as refrigerant for the generation of low-temperature chilling duty below −100 °C.

Ethylene, propene, propane and methane are used because of their physical properties and availability in the ethylene plant. [62]

One licensor recommends the application of a mixed refrigerant system, covering the full temperature range from −150 °C to +10 °C by one compressor only. Advantage of this system is the saving of one or even two compressor-turbine sets. This process applies a mixture of ethylene and propene or even ethylene, propene, and methane as refrigerant. Consequently the QT curve of the vaporizing coolant is not isothermal but follows on increasing line. [71].

Figure 36 shows the net heat and work required in a typical front-end demethanization plant with liquid feedstock. The less expensive refrigerant, propene, is used from + 25 to − 40 °C and the more expensive ethylene, from ca. −40 to −100 °C. The latter requires higher work input per unit of refrigeration;for example, Figure 36 shows that 41 % of the heat absorbed by the ethylene refrigeration system requires 58 % of the net work [64]. Refrigerants are usually generated by compression to 1600–2000 kPa in multistage units, followed by expansion to generate the lowest refrigeration temperature. Superheated ethylene is cooled and condensed with propene refrigerant. Propene, after compression, is cooled with process cooling water [54, 61, 64]. With open-loop systems, which are used less frequently, products are circulated in the refrigeration loops rather than isolated. Careful operation and reliable design of compressors (gas seals) reduces contamination [61].

Energy. Utility consumption, expressed in joules per kilogram, is the energy required to produce 1 kg of ethylene product. Feedstock, coproducts, and process design should be described fully for comparative purposes. Process optimization has to consider the cost of energy relative to capital costs, ease of process operation, product and byproduct quality control requirements, safety, and many other issues. The specific solution is always a compromise between the above mentioned parameters. Since the early 1970s until today, increased energy costs led to a reduction of energy consumption by nearly 50 %. This was accomplished by applying sophisticated process steps and designing the plants with a high degree of heat integration. Waste heat from cracked-gas cooling is utilized wherever possible, heat pumps, side reboilers, low-pressure distillation are applied. Modern design tools such as advanced process simulators or pinch analysis programs help the designer to find the optimum solu-

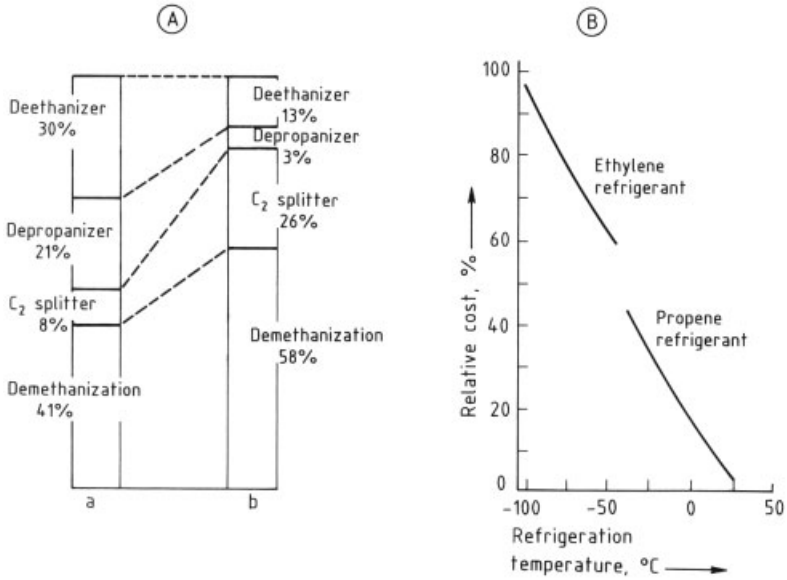

Figure 36. Distribution of heat load and work requirement of the refrigeration system in a typical front-end demethanizer ethylene plant *
A) Distribution of heat load and work requirement
a) Net heat absorbed by the refrigeration system; b) Net work done by the refrigeration system
B) Relative cost of heat absorbed by the refrigeration system at different temperatures
* Reprinted with permission [64].

tion for an individual case. Efficiencies of major energy consuming equipment such as cracking furnaces and compressor-turbines have continuously been improved.

Approximate energy consumption to produce 1 kg of ethylene from various feedstocks is in the range of 15 000 kJ/kg to 27 000 kJ/kg (see below)

A reduction of ca. 2000 J/ per kilogram ethylene product is possible by using hot exhaust from gas turbines in the pyrolysis area as combustion air to preheat combustion fuel, or in a heat-recovery boiler. The turbines may also optimize energy use to produce electricity or can drive fractionation compressors and heat pumps [51, 53, 58, 59, 61, 62, 64].The application of such a combined-cycle system, however, is costly and pays off only under special circumstances. Such installations are installed mainly in Japan and Korea, few of them in the rest of the world. In none of the new large-scale crackers built in the early 2000s such installations are applied.

Most of the energy consumed is introduced into the ethylene process via firing the cracking furnaces. The recovery process is driven by waste heat, recovered from the cracking furnaces via steam production in the TLEs or by recovering heat from the furnace effluent stream. The main energy consumers in the recovery section are the cracked-gas compressor and the refrigerant compressors, in which a total power of ca.0.6–0.7 kWh per kilogram ethylene is typically consumed. This power is usually provided by steam turbines, driven by steam generated and superheated in the cracking furnaces.Steam produced in the cracking furnaces and steam consumed in the ISBL ethylene plant is typically balanced. Depending on feed and byproducts some steam export or import is experienced.

Cooling. Ethylene production is a very energy-intensive process. Only some 20 % of the heat input is absorbed by the chemical reactions of the olefin production. The balance drives the process by driving turbines and pumps, heating columns, and process streams. At the end of the process the heat must be discharged to the ambient. Cooling is most commonly performed via a cooling-water cycle, discharging heat to the ambient via a cooling tower. In areas where cooling water is not available (e. g., new large petrochemical complexes in the Per-

sian gulf area) heat is discharged to the sea via the primary/secondary cooling-water cycle, or to the air via air-fin coolers. Both options require higher process temperatures leading to a slight reduction in energy efficiency and higher installation costs.

The ethylene process requires a number of utilities, some of them like fuel gas and steam being produced in the plant itself. Depending on feedstock and plant specifics fuel and steam must be supported by import or may be in excess. Typical consumption figures for various feedstocks for a standard-size 1000 t/a ethylene plant are summarized in Table 11.

Because feedstock composition, flow schemes, and product distribution are undefined, these values are approximate but they are representative.

5.3.4. Process Advances

Major advances have been made in feedstock availability and process design [51, 64],:

- Feedstock cost and availability, new cracking technology, and concerns for the environment and health have influenced changes in process technology.
- Rapid development has occurred in process control, data management, and optimization systems. Control of individual units, as well as of the overall process, has had a significant impact on operations. Improved analysis and cracking-simulation models are required to more fully apply state-of-the-art process control.
- Major improvements have been made in the selectivity and efficiency of cracking by operating at higher temperature. Coils with shorter residence time and faster quenching systems have been developed. Energy requirements have been reduced by heat recovery from furnace flue gas, economizer gas, and cracked gas; refrigeration energy requirements have been lowered by increased application and improved design of heat-exchange and distillation systems. Use of gas turbines has increased, with the sale of electricity generated. At the same time, control, reliability, and efficiency of operation have improved [61, 62].
- The reliability,plant availability service life, have considerably improved.Plants are designed today for uninterrupted operation between turnarounds of five years.
- Improved fractionation process designs provide additional flexibility and further reduce process energy consumption while meeting plant-reliability and product-quality demands [62].
- Improved hydrogenation catalysts allow more selective and more stable removal of impurities and better quality control.

5.3.5. Plant Size

At the beginning of industrial ethylene production, in the 1950s, first commercial ethylene plants had a standard size of some 20×10^3 t/a. In the next generation of plants, built in the 1960s, plant capacity was already expanded to $60 - 80 \times 10^3$ t/a. The era of large-scale plants started in the 1970s when the standard size reached some $300 - 400$–10^3 t/a. Since then until today typical plant capacity has continuously been increased to more than 10^6 t/a. Most of the large crackers, contracted during the boom period 2004–2006, most of them in the Middle East, have an ethylene capacity of $1 - 1.5 \times 10^6$ t/a.

The maximum plant capacity of a single-train ethylene plant is limited by the size of the cracked-gas compressor. Machinery manufacturers report that this capacity limit is most likely close to 2×10^6 t/a for a gas cracker considering current technology. However, it seems that technical constraints may not be the ultimate limiting factor. The industry will face the challenge of building larger plants if demanded by the market. Nontechnical issues such as logistic, feed availability, product offtake, and risk analysis will dominate this decision.

5.4. Other Processes and Feedstocks

Ethylene has also been produced by other processes and from other feedstocks [72]. Options for long-range feedstock availability and price, as well as in-place refining and chemical processes, influence process design and feedstock selection because of their high economic impact on manufacturing cost. Other drivers for new

Table 11. Typical consumption figures of a standard-size ethylene plant

Feedstock	Ethane	Propane	FR Naphtha	Gasoil
ethylene/propene product rate, t/h	125/0	125/50	125/55	125/75
Fired duty of cracking furnaces, MW	613	621	772	848
Fuel gas import (+) export (−), MW	28	− 540	− 142	40
SHP[a] steam production of the furnaces, t/h	450 at 6.5 MPa	454 at 7.5 MPa	458 at 11.5 MPa	463 at 11.5 MPa
HP Steam import (+) / export (−), t/h at 4.6 MPa	− 72	83	balanced	98
Electric power, kW_{el}	3.5	4.8	7.5	9
Circulating cooling water temperature rise 10 °C, m^3/h	28 000	42 000	47 000	60 000
Total compressor power, MW	72	90	85	93
Specific energy consumption, kJ/kg ethylene	16 000	20 000	23 000	27 000

[a] SHP: super high pressure

developments are the utilization of associated gas and natural gas as petrochemical feedstock, as the value of these fractions offers advantages with respect to costs compared to crude oil based feedstock. In addition as flaring of associated gas is no option anymore, other utlization is required for those streams.

Recovery from FCC Offgas. Recovery of ethylene and propene from FCC off-gas has gained importance, but due to the NO_x content, some steps have to be integrated to reduce the risk of formation of explosive resins from butadiene and N_2O_3.

Ethanol. Ethanol, an early ethylene feedstock, is dehydrated at high temperature over a solid catalyst. The catalyst is usually supported alumina, phosphoric acid, or silica. The ethanol can be obtained from fermentation, methanol homologation, syngas and other sources (→ Ethanol and → Ethanol, Chap. 5).

Crude Oil or Residual Oil. A number of nontubular reactors, developed to crack unprocessed hydrocarbon feedstocks, exist in various stages of development and commercial operation [72].

A process offered by Shaw (Deep Catalytic Cracking, DCC) employs catalysts particles, for cracking and heat transfer functions. The DCC process is suitable for cracking VGO and similar heavy feedstocks. Main target light product of the DCC process is propene, similar to FCC processes in which propene output can be increased significantly by catalyst modification.

Methanol can be catalytically dehydrated and partially converted to ethylene over alumina and zeolite catalysts. The process based on novel zeolite catalysts (ZSM-5) was developed by Mobil (MTO, Methanol to Olefins). A new development introduced by UOP/Norsk Hydro, converts methane to methanol in a first stage and then converts the methanol to olefins. Economics of this new process seem to be competitive with conventional processes [73]. The process is based on a fluidized-bed reactor for conversion of methanol. 80 % of the carbon content of methanol is converted into ethylene and propene. The process has been tested in a 0.5 t/d unit in Norway. Methanol obtained from syngas can be converted in high selectivity (but at a low rate) to ethanol with the help of catalysts and promoters.

In order to be competitive to the mega-scale ethane crackers in the Middle East with lowest feedstock costs worldwide, methanol has to be produced in huge quantities of approximately 5 $\times 10^6$ t/a leading to an olefins output of approximately 2×10^6 t/a of ethylene and propene. The ratio of propene to ethylene can be adjusted over a wide range and propene - ethylene ratios as required by the market (0.8 to 0.9) can be produced via this route.

Another development by Lurgi is the MTP process, similar to MTO, however, with a maximized propene output.

Syngas. Fischer–Tropsch and modified Fischer–Tropsch processes are used to produce (1) olefins directly as a byproduct of gasoline and diesel fuel production; (2) LPG and paraffin intermediates, which are cracked to produce olefins; and (3) intermediate oxygenated liquids, which are dehydrated and separated to produce olefins (→ Coal Liquefaction). Syngas is produced by gasification of coal and other materials (→ Gas Production).

Oxidative Coupling of Methane to Ethylene. Many researchers worldwide have been active in this field, using metal oxide catalysts at 700–900 °C for the oxidative coupling of methane. However, the yields obtained so far (20–25 %) are not competitive with conventional routes, since the ethylene concentration in the product gas is only ca. 10 vol % [74, 75]. Application of special separative reactors has led to improved yields up to 50–60 % [76, 77].

Dehydrogenation. Dehydrogenation of ethane over Cr or Pt catalyst is limited by equilibrium and allows only very poor yield of ethylene. This route is not competitive with conventional routes.

Metathesis. Metathesis of propene leads to ethylene and 1-butene:

$$2\,C_3H_6 \rightleftarrows C_2H_4 + C_4H_8$$

Since this reaction is reversible the ethylene/propene production of a plant can be adjusted to market conditions. So far this technology has been applied in a plant in Canada and at the U.S. Gulf Coast.

As the demand for propene is increasing faster than the ethylene demand and due to the lack of propene production in new ethane crackers in the Middle East, metathesis has gained importance for the production of propene based on ethylene and 2-butene feedstock. Several commercial units applying metathesis for propene production via this route are in operation. Contents of isobutene and butadiene in the feed create problems, as coking rates of the applied catalysts are increasing.

6. Environmental Protection

Ethylene plants with a production of up to 1.5 $\times 10^6$ t/a ethylene and a fired duty of up to 1000 mW require special measures for protection of the environment. Though national standards vary widely, the following general measures are required for every new plant:

Flue-Gas Emissions. NO_x emissions are limited by use of UltraLowNox burners or integrated SCR technology for catalytic reduction of NO_x. In some regions limits for NO_x emissions are already < 50 ppm. Particulate emissions during decoking can be reduced by incineration or separation in cyclones. Incinerators can achieve a reduction of 90 %, and high-performance separators, up to 96 %.

Fugitive Emissions and VOC. The equipment used in a modern olefin plant has to meet the standards for emission of critical components such as benzene. All components with a potential to emit critical components must have a certain tightness classification and have to be monitored (pump seals, valves, etc.). Storage tanks and piping are other sources of fugitive emissions.

Flaring. Flaring during start up should be avoided as much as possible, especially in highly populated areas. New procedures have been developed that allow for flareless start up of plant [78].

Noise Protection. Ethylene plants contain equipment that produces noise emission in the range above 85 dBA. These are the burners at the furnaces and the large compressors. Noise abatement measures are essential (silencers for burners, hoods for compressors) to bring the overall emissions into the range below 85 dBA. Other sources of noise emission can be high velocities in piping or vibrations. The detection of the source in this case is sometimes very difficult and requires much expertise.

Table 12. Typical ethylene specifications (in ppm vol unless otherwise noted) [79]

Component	Polymer Process				Polymer grade (today)
	Gas phase	Slurry	Solution	Metallocene [a]	
Ethylene	99.9 % min.	99.95 % min.	99.9 % min.	99.9 %	99.9 % min.
Methane	500	[b]	500 [b]	[e]	300
Ethane	500	[b]	500 [b]	[e]	500
Propene	20	20	50	20	10–15
Acetylene	0.1	2	55	0.1	2
Hydrogen	5	5	10	5	10
Carbon monoxide	0.1	1 [c]	2 [d]	0.1	2
Carbon dioxide	0.1	1 [c]	5 [d]	0.1	2
Oxygen	0.1	1	5	0.1	5
Water	0.1	5 [c]	5 [d]	0.1	2
MeOH		1 [c]	5 [d]	0.1	5
Sulfur as H_2S		1 (wt) [c]	2 (wt) [d]	0.1	2 (wt)
Other polar compounds		1	2	0.1	
DMF					2 (wt)

[a] Values for metallocene catalyst are expected values.
[b] Total combined not to exceed 0.125 %.
[c] Total not to exceed 1 ppm wt.
[d] Total must be less than 6 ppm.
[e] Total combined not to exceed 300 ppm wt.

Water Protection. The water emissions of the plant result from quench water, dilution steam, decoking water, and flare water discharges. These streams have to be treated properly before being fed to the wastewater plant. Streams from the caustic-scrubbing section require chemical treatment (oxidation) before discharge to the wastewater unit.

Solid and Hazardous Wastes. Examples of process wastes are cleaning acids; filter cartridges; catalysts; tars; polymers; waste oils; coke; sodium nitrite used in column passivation; *N,N*-dimethylformamide, acetone, or *N*-methyl-2-pyrrolidone used in acetylene removal; sulfuric acid from cooling towers; amine tars; and antifoaming agents. These wastes have to be treated according to the relevant regulations for disposal.

7. Quality Specifications

Ethylene quality worldwide is matched to customer requirements; no single chemical-grade ethylene exists. However, ethylene content normally exceeds 99.9 wt %. The reason for variation in ethylene specifications is due to the use of ethylene in downstream processes that differ in sensitivity to impurities in ethylene, due to different catalyst systems being used. In general, the trend is towards lower impurity contents as more active catalyst systems are developed. The largest application for ethylene is polymerization to HDPE, LDPE, and LLDPE. For these processes new catalyst systems have been developed, including the metallocene catalysts, which are more sensitive than other catalysts. Table 12 lists typical ethylene specifications for various polymerization processes.

Sulfur, oxygen, acetylene, hydrogen, carbon monoxide and carbon dioxide are the most troublesome and carefully controlled impurities, especially when ethylene from multiple sources is mixed in transportation.

8. Chemical Analysis

Laboratory analysis of ethylene is primarily by ASTM or modified ASTM gas chromatographic techniques.

Process control may use over 100 analyzers to monitor plant safety, operation, and optimization [80]. Measurements include personnel exposure, corrosion, furnace flue gas, steam systems, water systems, condensate systems, stack emissions, water emissions, feedstock quality, cracked gas analysis, distillation train product quality and moisture control, and, in some cases, heavy liquids product evaluation and control [81].

Process control instruments include gas chromatographs, combustible-gas detectors, conductivity analyzers, pH meters, densitometers, corrosion analyzers, oxygen and water analyzers, vapor pressure analyzers, and photoionization analyzers. Various automated wet methods are also used [80].

Process optimization is critical for ethylene production because cracking reactions change as the run proceeds. Operating costs for plants of this size are high [52, 82, 83]. Advances in computer control, data acquisition, and information analysis have given impetus to studies of process control, modeling, and optimization, especially in heavier liquid cracking operations.

Models for all kinds of feedstocks are applied, simulating all sections of the plant in detail, based on detailed chemical analysis of the various streams in the plant. With these tools operating strategies can be followed, e.g., production of a certain amount of ethylene, propene and other products at maximum profit, even if the feedstock quality or the type of feed change within short time. Such a computer controlled operation includes the individual control of cracking furnaces, columns, pumps, compressors, etc.

9. Storage and Transportation

Much of world ethylene production is consumed locally, requiring little storage and transportation. Interconnected pipelines and pressurized underground caverns in the United States and Europe have been developed because of the complex petrochemical infrastructure. They provide flexibility and prevent interruption of supply. The following pipeline systems have been developed:

United States:	Texas–Louisiana [85]
Canada:	Fort Saskatchewan–Sarnia [84]
Great Britain:	Grangemouth–Carrington Sevenside–Fawley [54]
Europe:	Northwest Europe (Frankfurt–Gelsenkirchen and Brussels–Rotterdam areas) and Spain [54, 86, 87]
Former Soviet Union:	Nizhnekamsk–Salavat and Angarsk–Zima [88, 89]

In the U.S. Gulf Coast area the pipeline grid is composed of common carriers and privately owned pipelines. Storage caverns are naturally occurring or artificially excavated salt domes; limestone caverns are less usual. In contrast to U.S. pipelines, which normally transport only ethylene, the 3000 km pipeline between Fort Saskatchewan and Sarnia transports ethane, propane, and condensate in batches [52, 84].

In pipelines, ethylene is normally under a pressure of 4–100 MPa. The upper end of the pressure range is significantly above the critical pressure. Below critical conditions, the temperature must be > 4 °C to prevent liquid ethylene from forming. If water is present, hydrates can form below 15 °C at normal operating pressure and can plug equipment. Special procedures and precautions are used during commissioning of pipelines to pressure test and dry the line and to provide appropriate draining and venting during operation [87].

Ethylene decomposition and spontaneous ignition can occur under certain conditions and can result in pressure and temperature increases capable of producing explosions. Such decomposition is a function of pressure, temperature, impurities (e.g., acetylene and oxygen), and catalysts (e.g., rust and carbon). Appropriate compression procedures, especially when nitrogen, oxygen, or other diatomic gases are present, help to prevent decomposition [90].

Ethylene is also transported by ship, barge, railcar, and tank truck. In the United States, these methods are normally limited to consumers not served by pipeline. In Europe, pipelines move a lower proportion of ethylene than in the United States.

Ethylene is also stored in underground high-pressure pipelines and tanks (e.g., at 10 MPa) and in surface refrigerated tanks. In the latter, compression and liquefaction of vapor from the refrigerated tank, followed by reinjection of the liquid, allow operation of the tank at pressures of only 7–70 kPa [52, 91].

10. Uses and Economic Aspects

More than 80 % of the ethylene consumed in 1993 was used to produce ethylene oxide; ethylene dichloride; and low-density, linear low-density, and high-density polyethylene. Significant amounts are also used to make ethylbenzene, oligomer products (e.g., alcohols and olefins), acetaldehyde, and vinyl acetate. LDPE

Figure 37. ARG pipeline grid in Europe

and LLDPE are the fastest growing outlets for ethylene.

Table 13 shows the uses of ethylene in 2000 [92].

Table 13. World consumption of ethylene in 2000 (in %)

Use	World
LDPE	18
LLDPE	14
HDPE	26
EO/EG	12
EDC	14
EB/SM	7
Others	9

Worldwide ethylene growth rate is projected at 4 to 5 % for the period of 2004 to 2009 [93] with the Asian market growing at 6 to 7 % per year.

The estimated 2005 world nameplate production capacity was over 112.9×10^6 t/a.

Table 14. Ethylene nameplate capacities by region in 2005 [94]

Region	Capacity, 10^3 t/a	% of World capacity
Europe (incl. Russia)	31 494	27.9
North America	33 830	30.0
South America	5669	5.0
Asia	39 880	35.3
Africa	1528	1.4
Australia	504	0.4

The principal influences on ethylene product value from the manufacturer's viewpoint are primarily the price of feedstock raw materials, chemical and byproduct credits, and expected return on capital costs associated with the facilities. The fixed cost component for ethylene production is relatively modest because of the high-volume commodity operation that has become the industry standard (e.g., up to 1500×10^3 single-train plants). The cost breakdown per tonne of ethylene produced is listed below for a new plant based on naphtha feedstock in the Asia/Pacific region:

Basis of economic analysis	
Capacity	1000×10^3 t/a
Location	Asia/Pacific
Feedstock	naphtha
Capital	10 years linear depreciation for ISBL plant 15 years linear depreciation for storage
Investment cost estimate	\$ 720×10^6 for ISBL plant \$ 144×10^6 for OSBL plant
Cost breakdown, %/t ethylene produced	
Feedstock costs	47.4
Catalysts and chemicals	0.5
Utilities	0.3
Cash cost	48.6
Operating costs	1.9
Overheads	1.9

11. Toxicology and Occupational Health [95]

Ethylene is not markedly toxic, but high concentrations may cause drowsiness, unconsciousness, or asphyxia because of oxygen displacement [96, 97]. The gas possesses useful anesthetic properties, with rapid onset, quick recovery, and minimal effects on the heart or lungs [98].

In animal studies, ethylene was not irritating to the skin or eyes [99]. It did not cause cardiac sensitization in a dog [100]. Mice and rats exposed repeatedly to ethylene showed minimal effects [101, 102]. In a chronic inhalation study with rats, no toxic effects were found by histopathology, hematology, clinical chemistry, urinalysis, eye examination, or mortality rate [102]. Ethylene improved wound healing following muscle injury to mice [103].

Studies with rodents have shown that ethylene caused increased serum pyruvate and liver weights [104], hypotension [105], decreased cholinesterase activity [102], hypoglycemia [106], and decreased inorganic phosphates [99]. No mutagenic activity was observed in *Escheria coli* B and *E. coli* Sd-4 exposed to ethylene [107].

Ethylene is a plant hormone and is useful for ripening fruits and vegetables [108, 109]. At high concentrations, it is phytotoxic [110]. Human exposure to ethylene is reported to cause anorexia, weight loss, insomnia, irritability, polycythemia, nephritis, slowed reaction time, and memory disturbances [111 – 113]. In the workplace, ethylene is treated as a simple asphyxiant [114]; the odor threshold is ca. 20 ppm [115]. NIOSH has published a procedure for evaluating ethylene exposure [116].

As with any compressed gas, skin and eye contact should be avoided. Appropriate precautions are required because the gas is highly flammable and explosive over a wide range of mixtures with air. Ethylene is also spontaneously explosive in sunlight in the presence of ozone or chlorine, and reacts vigorously with aluminum chloride and carbon tetrachloride [95].

12. References

Specific References

1. R. Becker, *Chem.-Tech. (Heidelberg)* **8** (1979) 6.
2. *Oil Gas J.* **104** (2006) March 27th special report.
3. *Oil Gas J.* **103** (2005) March 28th.
4. *SRI Consulting, Presentation at APIC*, May 19 (2005).
5. *ICRA Report* "Petrochemicals", June 2005.
6. M. Eramo: "Emerging Trends in Global Light Olefins," CMAI Paper presented at APTC 2002, Seoul.
7. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie*, 4th ed., VCH Verlagsgesellschaft, Weinheim, Germany 1994.
8. D. Lambert et al., *Analysis Magazine* **23** (1995) no. 4, M9–M14.
9. K. M. Watson, E. F. Nelson, *Ind. Eng. Chem.* **25** (1933) 880–887.
10. H. M. Smith: "Correlation Index to Aid in Interpreting Crude-Oil Analyses," U.S. Bureau of Mines Technical Paper no. 610, 1940.
11. J. M. Collins, G. H. Unzelman, *Oil Gas J.* **80** (1982) June, 148.
12. Y. C. Hu, *Hydrocarbon Process* **61** (1982, Nov.) 109, Part 1.
13. R. N. Bell: "Review of Impact of Mercury on Aluminum Heat Exchangers," Paper presented at EEPC conference Oct. 6, 2005, Prague..
14. M. E. Denti, E. M. Ranzi in L. F. Albright, B. L. Crynes, W. H. Corcoran (eds.): *Pyrolysis: Theory and Industrial Practice*, Academic Press, New York 1983, pp. 133–137.
15. D. Edelson, *J. Phys. Chem.* **81** (1977) 2309.
16. K. Ebert, H. Ederer, G. Isbarn, *Int. J. Chem. Kinet.* **15** (1983) 475.
17. M. Schärfe, H. J. Ederer, U. Stabel, K. Ebert, *Chem.-Ing.-Tech.* **56** (1984) no. 6, 488–489.
18. G. F. Froment, *Chem. Eng. Sci.* **47** (1992) 2163.
19. Q. Chen, *Chemtech* **26** (1996) Sept., 33.
20. M. Dente et al. *Chem. Eng. Sci.* **47** (1992) 2629.
21. F. O. Rice, *J. Am. Chem. Soc.* **53** (1931) 1959–1972.
22. S. B. Zdonik, E. J. Green, L. F. Hallee, *Oil Gas J.* **65** (1967) no. 26, 96–101; **66** (1968) no. 22, 103–108.
23. C. W. Gear: *Numerical Initial Value Problems in Ordinary differential Equations*, Prentice Hall, Englewood Cliffs, N.J. 1971.

24. M. Boudart: *Kinetics of Chemical Processes*, Prentice Hall, Englewood Cliffs, N.J. 1968.
25. N. N. Semenov (transl. by M. Boudart): *Some Problems in Chemical Kinetics and Reactivity*, Princeton University Press, N.J. 1959.
26. F. Rossini, K. Pitzer, R. Arnett, R. Braun et al.: *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds*, API Research Project 44 Report, Carnegie Press, Pittsburg, Pa. 1955, p. 706.
27. Kinetics Technology International, US 4 367 645, 1983 (G. F. Froment).
28. S. Barendregt, J. L. Monfils: "Steamcracking for Butane Upgrading," AlChE meeting, Houston, TX, 1991
29. D. Kaufmann: "New Developments for Radiant Coil Materials," Paper presented at EEPC HSE conference, Vienna. 2004
30. Mitsui Advanced Cracker, Technical Bulletin, Mitsui Engineering & Shipbuilding Co., Tokyo, Japan, pp. 1–5.
31. S. Ibarra, in [14], pp. 427–436.
32. G. E. Moller, C. W. Warren, *Mater. Perform.* **20** (1981) no. 10, 27–37.
33. J. A. Thuillier, *Mater. Perform.* **15** (1976) no. 11, 9–14.
34. R. H. Krikke, J. Hoving, K. Smit: "Monitoring the Carburization of Furnace Tubes in Ethylene Plants," International Corrosion Forum Paper no. 10, National Assoc. of Corrosion Engineers, Houston, Tex., Mar. 22–26, 1976.
35. D. E. Hendrix, M. Clark: "Contributing Factors to the Unusual Creep Growth of Furnace Tubing in Ethylene Pyrolysis Service," International Corrosion Forum Paper no. 21, National Assoc. of Corrosion Engineers, Boston, Mass., Mar. 25–29, 1985.
36. J. J. Jones, J. L. D. Steiner: "A New Generation of Heat Resisting Alloys, Paralloys SH24T and SH39T," International Corrosion Forum Paper no. 22, National Assoc. of Corrosion Engineers, Boston, Mass., Mar. 25–29, 1985.
37. S. B. Parks, C. M. Schimoller, *Hydrocarbon Process.* **75** (1996) 53.
38. L. A. Mekler, R. S. Fairall, *Pet. Refiner* **31** (1952) no. 6, 100–107, Part 1.
39. W. E. Lobo, J. E. Evans, *Trans. Am. Inst. Chem. Eng.* **35** (1939) 743–778
40. R. L. Grantom: "Combustion Air Should Be More Than Simply Recycling Energy," *Proc. 1980 Conference on Industrial Energy Conservation*, vol. **2**, Texas Industrial Commission, Austin, Tex. 1980, pp. 559–564.
41. R. L. Grantom: "Combustion Air Preheat and Radiant Heat Transfer in Fired Heaters–A Graphical Method for Design and Operating Analysis," *Proc. 1981 Industrial Energy Conservation Technology Conference and Exhibition*, vol. **1**, Texas Industrial Commission, Austin, Tex. 1981, pp. 151–159.
42. J. V. Albano, E. F. Olszewski, T. Fukushima, *Oil Gas J.* **90** 1992, Feb., 55.
43. M. J. Graff, L. F. Albright, *Carbon* **20** (1982) no. 4, 319–330.
44. J. C. Marek, L. F. Albright: "Coke Formation on Catalysts or in Pyrolysis Furnaces," *Am. Chem. Soc. Symposium, 182nd Am. Chem. Soc. Meeting*, New York 1981.
45. A. G. Goosens, M. Denti, E. Ranzi, *Hydrocarbon Process.* **57** (1978) no. 9, 227–236.
46. L. L. Ross, in [14], pp. 327–364.
47. B. Lohr, H. Dittmann, *Oil Gas J.* **70** (1978) no. 20, 63–68.
48. J. Chen, M. J. Maddock, *Hydrocarbon Process.* **52** (1973) no. 5, 147–151.
49. A. Mol, in [14], pp. 451–471.
50. "Transfer Line Exchangers in Ethylene Plants," Technical Bulletin, Schmidt'sche Heissdampf-Gesellschaft mbH, Kassel-Bettenhausen, Federal Republic of Germany, 1985.
51. E. S. Kranz, S. B. Zdonik, Personal Communication, Stone & Webster Engineering Corp., Boston, Mass., May 1985.
52. S. B. Zdonik, in [14], Chapter 15.
53. R. C. Updegrove, Personal Communication, E. I. du Pont de Nemours & Co., Chocolate Bayou, Tex., Jul. 1985.
54. *Ullmann*, 4th ed., **8**, p. 192.
55. H. E. Boyd, Y. H. Chen, O. J. Quartulli, *Oil Gas J.* **74** (1976, Sept. 27) 51–59.
56. C. W. Albers, T. A. Wells, *Oil Gas J.* **76** (1978) Sept. 4, 72–78.
57. K. W. Brooks, *Oil Gas J.* **62** (1964) Dec. 21, 70–75.
58. T. A. Wells, *Oil Gas J.* **80** (1982) Jun. 14, 63–68.
59. W. C. Petterson, T. A. Wells, *Chem. Eng.* **84** (1977) Sept. 26, 76–86.
60. L. K. Ng, C. N. Eng, R. S. Zach, *Hydrocarbon Process.* **62** (1983, Dec.) 99–103.
61. A. J. Weisenfelder, C. N. Eng., Personal Communication, C. F. Braun & Co., Alhambra, Calif., Aug. 1985.
62. C. Sumner, T. S. Williams, Personal Communication, Combustion Engineering, Bloomfield, N. J., Aug. 1985.

63. Greg Dunnels paper no 97C, presented at the Ethylene Producer Conference 2002, New Orleans, USA.
64. R. Orriss, Personal Communication, M. W. Kellogg Co., Houston, Tex., Jun. 1985.
65. R. Zeppenfeld, E. Haidegger, W. Borgmann, *Hydrocarbon Technology International* (1993) 129.
66. C. Bowen, R. McCue, C. Carradine, *Hydrocarbon Process.* **72** (1993) 129.
67. W. K. Lam, Y. R. Mehra, D. W. Mullins, AIChE Spring International Meeting, Houston 1993, session 18.
68. H. Zachariades, J. A. Reid, Article 19C, presented at the Ethylene Producer Conference 1996, New Orleans, USA.
69. ABB Lummus Global, US 5925799A 1999 (S. Stephen, C. Sumner).
70. G. Gildert, W. Groten, K. Rock, D. Weidert, Article 20B, presented at the Ethylene Producer Conference 1995, Houston, USA.
71. ABB Lummus Global, WO 03087022, 2003 (W. V. Tuan, M. Qi, WU J. Tzong-Chaur).
72. Y. C. Hu, *Hydrocarbon Process.* **51** (1972) Nov., 109–116, Part 1; **62** (1983) Apr., 113–116, Part 2; **62** (1983) May, 88–95, Part 3.
73. M. Heathcote, *Asian Chemical News* **3** (1996) June 24, 19.
74. N. A. Baronskaya, L. S. Woldman, A. A. Davydiv, O. V. Buyevskaya, *Gas Sep. Purif.* **10** (1996) no. 1, 86.
75. L. Guczi, R. A. van Santen, K. W. Sarma, *Catal. Rev. Sci. Eng.* **38** (1996) no.2, 249.
76. A. L. Tonkovich, R. W. Carr, R. Aris, *Science (Washington D.C.)* **262** (1993) 221.
77. T. Nozaki, O. Yamzaki, K. Omata, K. Fujimoto, *Chem. Eng. Sci.* **47** (1992) no.9, 2945.
78. A. Shaikh, C. J. Lee, *Hydrocarbon Process.* **74** (1995) 89.
79. J. A. Reid, D. R. McPhaul: " Contaminant Rejection Technology Update," presented at AIChE Spring National Meeting, 8th Annual Ethylene Producers Conference, New Orleans, LA, Feb. 28, 1996.
80. W. H. Delphin, J. C. Franks, Personal Communication, E. I. du Pont de Nemours & Co., Chocolate Bayou, Tex., Mar. 1985.
81. CFR, vol. 40, 1985, Part 61.
82. M. Nasi, M. Sourander, M. Tuomala, D. C. White et al., *Hydrocarbon Process.* **62** (1983) Jun., 74–82.
83. M. Sourander, M. Kolari, J. C. Cugini, J. B. Poje et al., *Hydrocarbon Process.* **63** (1984) Jun., 63–69.
84. *Pipeline Gas J.* **210** (1983) Jun., no. 7, 34–37.
85. S. D. Govrau, MAP Search Services, Houston, TX, personal communication, May 1985.
86. *Eur. Chem. News* **27** (1975) Feb. 28, 19.
87. L. Kniel, O. Winter, K. Stork: *Ethylene, Keystone to the Petrochemical Industry*, Marcel Dekker, New York 1975, Chapter 5.
88. *Oil Gas J.* **75** (1977) no. 39, Sep. 19, 108.
89. *Eur. Chem. News* **31** (1977) no. 813, Nov. 25, 56.
90. F. F. McKay, G. R. Worrell, B. C. Thornton, H. L. Lewis, *Pipe Line Ind.* **49** (1978, Aug.) no. 2, 77–78, 80, 82, 84.
91. E. L. Pawlikowski, Storing and Handling Liquefied Olefins and Diolefins, Exxon Chemical Co. USA, 1974.
92. M. Eramo: CMAI, APTC, May 7, 2002, Seoul, Korea.
93. M. Yoneyama: SRI Consulting, Asia Petrochemical Conference, May 19, 2005, Yokohama, Japan.
94. *Oil Gas J.* **103** (2005) March 28, 54.
95. W. D. Broddle, Personal Communication, Conoco, Ponca City, Okla., May 1985.
96. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, 3rd rev. ed., vol. **28**, Wiley-Interscience, New York 1981, pp. 3198–3199.
97. W. B. Deichman, H. W. Gerarde: *Toxicology of Drugs and Chemicals*, Academic Press, New York 1969, p. 316.
98. A. L. Cowles, H. H. Borgstedt, A. J. Gillies, *Anesthesiology* **36** (1972) no. 6, 558.
99. W. R. von Oettingen, Toxicity and Potential Dangers of Aliphatic and Aromatic Hydrocarbons, Public Health Bulletin no. 225, Washington, D.C., 1940.
100. J. C. Krantz, Jr., C. J. Carr, J. F. Vitcha, *J. Pharmacol. Exp. Ther.* **94** (1948) 315.
101. C. Reynolds, *Anesth. Analg. (New York)* **6** (1927) 121.
102. T. E. Hamm, D. Guest, J. G. Dent, *Fund.Appl. Toxicol.* **4** (1984) 473–478.
103. P. Pietsch, M. Chenoweth, *Proc. Soc. Exp. Biol. Med.* **130** (1968) 714.
104. R. B. Conolly, R. J. Jaeger, S. Szabo, *Exp. Mol. Pathol.* **28** (1978) 25.
105. M. L. Krasovitskaya, L. Malyarova, *Gig. Sanit.* **33** (1968) no. 5, 7.
106. P. Cazzamali, *Clin. Chirurg.* **34** (1931) 477.
107. M. M. Landry, R. Fuerst, *Dev. Ind. Microbiol.* **9** (1968) 370.
108. H. K. Pratt, J. D. Goeschl, *Am. Rev. Plant Physiol.* **20** (1969) 542.

109. R. E. Holm, J. L. Rey, *Plant Physiol.* **44** (1969) 1295.
110. G. D. Clayton, T. S. Platt, *Am. Ind. Hyg. Assoc. J.* **28** (1967) 151.
111. J. M. Arena (ed.): *Poisoning*, 4th ed., C. C. Thomas, Springfield, Ill. 1979, p. 426.
112. L. R. Riggs, *Proc. Soc. Exp. Biol. Med.* **22** (1924–1925) 269.
113. Chemical Safety Data Sheets, Chemical Manufacturers' Assoc. (formerly the Manufacturing Chemists' Assoc. Inc.,) Washington, D.C., 1979.
114. ACGIH (ed.): *Threshold Limit Values for Chemical Substances and Physical Agents in the Workroom Environment*, ACGIH, Cincinnati, Ohio 1986–1987.
115. M. L. Krasovitskaya, L. K. Malyarova, *Biol. Deistvie Gig Znach. Atmos. Zagryaz.* 1966, 74.
116. National Institute for Occupational Safety and Health Manual of Sampling Data Sheets, U.S. Dept. Health Education Welfare, U.S. Printing Office, Washington, D.C., 1977.

Ethylenediaminetetraacetic Acid and Related Chelating Agents

J. Roger Hart, W. R. Grace & Co., Lexington, Massachusetts 02173, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 4
5. **Chemical Analysis** 4
6. **Storage and Transportation** 5
7. **Uses** 5
8. **Trade Names** 6
9. **Economic Aspects** 6
10. **Toxicology and Occupational Health** . 6
11. **References** 7

1. Introduction

Ethylenediaminetetraacetic acid [*60-00-4*], {EDTA, *N,N′*-1,2-ethanediylbis[*N*-(carboxymethyl)glycine], edetic acid}, nitrilotriacetic acid [*139-13-9*], [NTA, *N,N*-bis(carboxymethyl)glycine], and their salts were first synthesized by Ferdinand Munz in the I.G. Farbenindustrie laboratories during the 1930s. Since that time, EDTA, hydroxyethylethylenediaminetriacetic acid (HEEDTA, *N*-{2-[bis(carboxymethyl)amino]ethyl}-*N*-(2-hydroxyethyl)glycine), diethylenetriaminepentaacetic acid (DTPA, *N,N*-bis{2-[bis(carboxymethyl)amino]ethyl}glycine, pentetic acid), and their salts have become important industrial chelating agents.

Aminopolycarboxylic acid chelating agents, like other amino acids, are capable of forming salts with strong acids or bases. In addition, these chelating agents can form strong, water-soluble metal complexes with di- and trivalent cations. These complexes greatly alter the reactivity of the metal ion, thus making them useful in many important industrial processes. The major volume usage of aminopolycarboxylic acid chelating agents is in preventing or removing scales and insoluble deposits or precipitates containing calcium, barium, iron, and other cations. These agents are also added to many processes, e.g., paper pulping, to prevent contained traces of metal ions from interacting in harmful ways during subsequent processing.

Chelating agents are also useful in altering the oxidation–reduction properties of transition-metal ions, such as iron and manganese, to increase or decrease the reactivity of these systems. In addition, they are widely used to provide water-soluble forms of metal ions that are resistant to anions, such as phosphate or carbonate, which can cause precipitation or inactivation. The agricultural use of water-soluble metal complexes of EDTA is an example.

2. Physical Properties

The structures and physical properties of the three most important aminopolycarboxylic acid chelating agents and their salts are given in Table 1.

3. Chemical Properties

Ethylenediaminetetraacetic acid, related aminopolycarboxylic acids, and their salts form water-soluble complexes with alkaline-earth and heavy-metal ions. These metal ions are incorporated into a ring structure called a *chelate*, after the Greek word $\chi\eta\lambda\acute{\eta}$ meaning "claw". This chelate structure is comprised of the dissociated ligand, e.g., $EDTA^{4-}$, and the metal ion, e.g., Fe^{3+}, to form the metal chelate $[Fe(EDTA)]^{-}$:

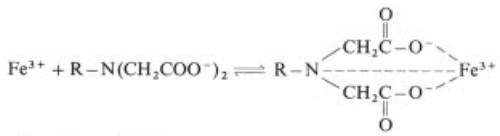

where R = $(^{-}OOCCH_2)_2NCH_2CH_2$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_095

Table 1. Physical properties of selected aminopolycarboxylic acid chelating agents

Compound	CAS registry number	Molecular formula	M_r	*mp*, °C	Solubility[a] in water (approximate), g/L at 20 °C	pH of 1 wt % solution
Ethylenediaminetetraacetic acid (EDTA), $(HOOCCH_2)_2NCH_2CH_2N(CH_2COOH)_2$	[*60-00-4*]	$C_{10}H_{16}N_2O_8$	292.24	245 (decomp.)	0.1 0.5 (90 °C)	
Na_2EDTA [b]	[*139-33-3*]	$C_{10}H_{14}N_2O_8Na_2$	336.20	0	105	5.0 – 5.5
Na_4EDTA [c]	[*64-02-8*]	$C_{10}H_{12}N_2O_8Na_4$	380.17		500	11.3
Hydroxyethylethylenediaminetriacetic acid (HEEDTA), $(HOCH_2CH_2)N(CH_2COOH)CH_2CH_2N(CH_2COOH)_2$	[*150-39-0*]	$C_{10}H_{18}N_2O_7$	278.26	212 – 214 (decomp.)	60	
$Na_3HEEDTA$	[*139-89-9*]	$C_{10}H_{15}N_2O_7Na_3$	344.22		480	11.2
Diethylenetriaminepentaacetic acid (DTPA), $[(HOOCCH_2)_2NCH_2CH_2]_2NCH_2COOH$	[*67-43-6*]	$C_{14}H_{23}N_3O_{10}$	393.35	220 (decomp.)	5	
Na_5DTPA	[*140-01-2*]	$C_{14}H_{18}N_3O_{10}Na_5$	503.26		500	11.3

[a] EDTA, HEEDTA, DTPA, and their salts are insoluble in most common organic solvents.
[b] Crystallizes from water as the dihydrate.
[c] Crystallizes from water as the tetrahydrate.

The metal ion is bound quite tightly to the ligand and the strength of this complex can be estimated through the stability constant, $K_{\mathrm{stab.}}$, which is a function of the equilibrium concentrations of the metal ion M and the chelating agent Y:

$$K_{\mathrm{stab.}} = \frac{c_{\mathrm{MY}}}{c_{\mathrm{M}} c_{\mathrm{Y}}}$$

The strengths of the various metal chelate complexes can be related by means of log $K_{\mathrm{stab.}}$ (see Table 2).

The *pH of the system* and the *presence of competing anions* have a decided effect on the strength of the metal chelate complex. The active chelating moiety is the fully dissociated chelate anion, e.g., $EDTA.^{4-}$ The concentration of this species is greater at higher pH; hence, chelating agents are more effective as alkalinity increases. However, this effect is offset by competing anions, which act as precipitating or complexing species. Increased hydroxyl ion concentration at higher pH values tends to disrupt the metal chelate complex if the metal ion forms an insoluble hydroxide. For example, at pH > 5.5, Fe^{3+} precipitates from its EDTA chelate. Likewise, carbonate, phosphate, oxalate, silicate, and other precipitating anions greatly alter the effective strength of the metal – EDTA complex. This occurs when the effective solubility product of the insoluble metal compound approaches the value of the effective stability constant of the metal chelate.

In general, chelating agents form 1 : 1 chelates with metal ions. The mass equivalents for various metal chelate complexes under ideal conditions are shown in Table 3. Increasing the stoichiometric concentration of the chelating agent to values greater than 1 : 1 increases the stability of the chelate complex. For example, Fe (III) (HEEDTA) remains in solution up to pH 7, whereas Fe (III) (HEEDTA), in the presence of a molar excess of HEEDTA, is stable to pH 12 [2].

In addition, converting a metal cation to a metal complex anion greatly alters the interaction of the metal ion with other species. For example, the movement of $[Fe(III)(EDTA)]^-$ through soil containing clay is facilitated by converting the easily absorbed Fe^{3+} cation to a more mobile anionic form.

Table 2. Stability constants of metal chelates of EDTA, HEEDTA, and DTPA [1]

Metal ion	$\log K_{\mathrm{stab.}}$		
	EDTA	HEEDTA	DTPA
Al^{3+}	16.3	14.3	18.6
Ba^{2+}	7.86	6.3	8.87
Ca^{2+}	10.69	8.3	10.83
Co^{2+}	16.31	14.6	19.27
Cu^{2+}	18.80	17.6	21.55
Fe^{2+}	14.32	12.3	16.5
Fe^{3+}	25.1	19.8	28.0
Mg^{2+}	8.79	7.0	9.30
Mn^{2+}	13.87	10.9	15.60
Ni^{2+}	18.62	17.3	20.32
Pb^{2+}	18.04	15.7	18.80
Sr^{2+}	8.73	6.9	9.77
Zn^{2+}	16.50	14.7	18.40

Table 3. Mass equivalents for 1 : 1 metal chelate complexes

Chelating agent, 1 g	Metal ion, mg						
	Ca^{2+}	Mg^{2+}	Cu^{2+}	Mn^{2+}	Fe^{2+}	Fe^{3+}	Zn^{2+}
Na_4EDTA	105	64	167	144	147	147	172
$Na_3HEEDTA$	116	70	184	159	162	162	190
Na_5DTPA	79	48	126	109	111	111	130

The effective oxidation potential, $E_{\mathrm{eff.}}$, is increased when the concentration of free metal ion, M, is lowered by the formation of a metal complex.

$$E_{\mathrm{eff.}} = E^0 - 0.05915 \log c_{\mathrm{M}}$$

The change in electrode potential resulting from chelation with EDTA, which forms the basis for a number of important industrial processes, is shown in Table 4 [3].

4. Production

The original commercial synthesis of EDTA was from ethylenediamine, chloroacetic acid, and caustic soda to form the tetrasodium salt [4]. Material produced in this way was contaminated with byproduct sodium chloride and had to be purified for many uses. This process is no longer employed commercially.

Other processes for the production of EDTA have been developed but not used commercially, for example, the catalytic oxidation of tetra(hydroxyethyl)ethylenediamine [*140-07-8*], 2,2′,2″,2‴-(1,2-ethanediyldinitrilo)tetrakis(ethanol) [5].

Today, the two principal manufacturing processes for EDTA and related chelating agents are both based on the cyanomethylation of the parent polyamine.

The most widely used synthesis is the *alkaline cyanomethylation of ethylenediamine* by means of sodium cyanide and formaldehyde [6]:

$$\begin{aligned} & H_2NCH_2CH_2NH_2 + 4\,CH_2O + 4\,NaCN + 4\,H_2O \\ \longrightarrow\ & (NaOOCCH_2)_2NCH_2CH_2N(CH_2COONa)_2 \\ & + 4\,NH_3 \end{aligned}$$

This method offers high yields (> 90 %) of the chelating agent. The principal byproduct is ammonia, which is continuously boiled off during the reaction. However, some of the ammonia is cyanomethylated to yield salts of NTA, *N*-(carboxymethyl)glycine, and of glycine. In addition, glycolic acid salts are formed from the reaction of sodium cyanide and formaldehyde. These impurities are not detrimental to most applications of chelating agents.

The second commercial method for producing EDTA is the two-step *Singer synthesis* [7]. In this process, the cyanomethylation step is separate from the hydrolysis. Hydrogen cyanide and formaldehyde react with ethylenediamine to form

Table 4. Standard electrode potentials for metals and their EDTA complexes

Reaction	Potential, V	Reaction	Potential, V
$Ce - 3e^- \rightleftharpoons Ce^{3+}$	−2.48	$Ce + EDTA^{4-} - 3e^- \rightleftharpoons [Ce\ (EDTA)]^-$	−2.78
$Co - 2e^- \rightleftharpoons Co^{2+}$	−0.28	$Co + EDTA^{4-} - 2e^- \rightleftharpoons [Co\ (EDTA)]^{2-}$	−0.76
$Cu - 2e^- \rightleftharpoons Cu^{2+}$	+0.34	$Cu + EDTA^{4-} - 2e^- \rightleftharpoons [Cu\ (EDTA)]^{2-}$	−0.22
$Fe - 2e^- \rightleftharpoons Fe^{2+}$	−0.44	$Fe + EDTA^{4-} - 2e^- \rightleftharpoons [Fe\ (EDTA)]^{2-}$	−0.87
$Fe - 3e^- \rightleftharpoons Fe^{3+}$	−0.04	$Fe + EDTA^{4-} - 3e^- \rightleftharpoons [Fe\ (EDTA)]^-$	−0.17
$Mn - 2e^- \rightleftharpoons Mn^{2+}$	−1.18	$Mn + EDTA^{4-} - 2e^- \rightleftharpoons [Mn\ (EDTA)]^{2-}$	−1.45

insoluble (ethylenedinitrilo)tetraacetonitrile [*5766-67-6*] (EDTN), 2,2′,2″,2‴-(1,2-ethanediyldinitrilo)tetrakis(acetonitrile) (**1**), in high yield (> 96 %). The intermediate nitrile is separated, washed, and subsequently hydrolyzed with sodium hydroxide to tetrasodium EDTA, with liberation of byproduct ammonia. Carrying out the synthesis in two stages eliminates most of the impurity-forming reactions and yields a very pure form of chelating agent.

$$H_2NCH_2CH_2NH_2 + 4\,CH_2O + 4\,HCN \longrightarrow \underset{\mathbf{1}}{(NCCH_2)_2NCH_2CH_2N(CH_2CN)_2} + 4\,H_2O$$

$$\xrightarrow{4\,NaOH} (NaOOCCH_2)_2NCH_2CH_2N(CH_2COONa)_2 + 4\,NH_3$$

This two-step reaction is also particularly well-suited to synthesizing pure NTA in high yield from ammonia as the parent amine.

5. Chemical Analysis

Ethylenediaminetetraacetic acid and related chelating agents are readily analyzed by *titration with standardized solutions of metal salts*, which react quantitatively to form metal chelate complexes. The end point is commonly determined by the precipitation and turbidity of an insoluble metal compound or the color change of a metal-sensitive dye used as an indicator.

Titration of chelating agents with calcium acetate in the presence of sodium oxalate is an example of the turbidity method [8]. The Schwarzenbach titration relies on the color change of Eriochrome Black T at the end point when the chelating agent removes calcium from the red calcium – Eriochrome Black T complex to liberate the blue dye [9].

Small amounts of EDTA may be determined by using copper ion titration with a copper ion specific electrode [10].

6. Storage and Transportation

Chelating agents, by their very nature, tend to be corrosive to metals such as copper, zinc, and iron. The alkaline salts corrode aluminum severely, and contact with these salts should be avoided. Stainless steel and most polymers or coatings that are able to withstand dilute alkali are acceptable as construction materials. Solid chelating agents should be stored in a cool, dry place to avoid caking.

7. Uses

Chelating agents, with their ability to solubilize and inactivate metal ions by complex formation, are used for a variety of purposes [11].

Textile dyeing and finishing have traditionally been a major market for EDTA-type chelating agents. Traces of metal ion contaminants from incoming fibers, chemicals, and process water will cause a shade change in many commercial dyes unless a chelating agent is present. In addition, DTPA is widely used to *stabilize hydrogen peroxide* bleaching liquors against decomposition by traces of manganese, iron, and copper.

The brightness of *paper pulp* bleached with hydrogen peroxide or dithionite is improved by the addition of chelating agents during the pulping stage and, later, to the bleaching solution where it acts as a stabilizer and metal ion control; DTPA is especially useful for peroxide stabilization [12]. Chelating agents aid especially in the removal of manganese from the wood source, which, in addition to decomposing the peroxide bleaching agent, later reacts with lignins to form dark-colored substances [13], [14].

Chelating agents are used to *clean scale deposits* from internal boiler surfaces and as additives to incoming boiler feedwater to prevent

the formation of calcium and magnesium scales. Care must be taken to control the ratio of chelating agent to water hardness, lest the amount of free chelating agent rises to corrosive levels. Polymers have been used in conjunction with chelating agents to reduce the corrosive potential of the treatment.

The *cleaning of metal surfaces* is aided by the use of chelating agents to solubilize adherent oxide films and provide complete rinsing. Metal-plating solutions often incorporate chelating agents to form complex metal ions, which improve the smoothness and integrity of the metal coating.

Cleaning compounds of many types, including laundry detergents and hard surface and bathroom cleansers, are improved by adding chelating agents [2]. The bactericidal activity of sanitizing agents, especially toward gram-negative bacteria, increases when EDTA is incorporated, because of its ability to destroy the outer cell wall of these often resistant species. This bactericidal enhancement by EDTA is also used to potentiate the preservative activity in many consumer preparations including cosmetics, shampoos, and ophthalmic products [2], [10], [15], [16].

Chelating agents are used to *stabilize systems* from breakdown by the catalytic action of trace metals. The oxidative rancidity of fatty emulsions, e.g., mayonnaise, salad dressings, cosmetic creams, and lotions, is prevented by adding EDTA as the disodium salt or disodium – calcium complex. *Discoloration* by iron and other metal ions may be inhibited by adding EDTA. Phenolic compounds present in potatoes form reddish brown iron complexes, which turn black during cooking; EDTA can be used to prevent this deleterious action [17].

The *change in oxidation – reduction potential* for metal ions on chelation is employed commercially in the emulsion polymerization of styrene – butadiene rubber in the "cold" process that uses Fe (II) (EDTA). The bleaching of color photographic film is accomplished with Fe (III) (EDTA), which oxidizes the metallic silver present in the exposed image to the ionic form for removal or "fixing" by thiosulfate. The control of plant emissions of hydrogen sulfide is accomplished on an industrial scale by using various iron chelate compounds to catalytically oxidize hydrogen sulfide to elemental sulfur. This process is especially useful for low concentrations of hydrogen sulfide and results in a salable grade of sulfur as a byproduct [18]. Iron chelates are also used as reversible absorbents for nitrogen oxide (NO). The activation of peroxygen bleaches in laundry detergents is also accomplished by metal chelate complexes [19].

Chelating agents are added to chemical processes to remove traces of iron and other metal ions, which either interfere with the process or must be reduced to meet stringent specifications.

Agricultural micronutrients or trace elements, such as iron, zinc, manganese, and copper, are added to fertilizers in the chelated form to prevent interaction and precipitation with contained phosphates and other precipitating anions. The conversion of these cationic metal ions to an anionic chelated form also allows the micronutrient to translocate through the soil medium to the root zone for uptake into the plant. Clays, carbonates, phosphates, and other soil constituents hinder the movement of cationic metal sources from the point of placement on the soil surface. These factors account for the increased efficiency of chelated micronutrients over inorganic sources [20].

Many techniques in *analytical chemistry,* especially titrimetric methods, that use EDTA have been developed [21]. The equivalence point is easily measured because of the abrupt reduction in free metal ion concentration when equimolar equivalence is achieved. Also EDTA is used as a masking agent to prevent interference by metal ions in some spectrophotometric methods.

8. Trade Names

The first trade name for EDTA was Trilon B from I.G. Farbenindustrie. Trade names are Trilon B (BASF), Versene (Dow), Sequestrene (Ciba-Geigy), Hamp-ene (Grace, USA), Rexene (Grace, Sweden), and Chelest (Chelest, Japan). Trade names for HEEDTA are Versenol (Dow), Chel DM (Ciba-Geigy), Hamp-ol (Grace, USA), Rexenol (Grace, Sweden), and Chelest H (Chelest, Japan). Major trade names for DTPA are Versenex (Dow), Chel DTPA (Ciba-Geigy), Hamp-ex (Grace, USA), Polychelate (Grace, Sweden), and Chelest P (Chelest, Japan).

Table 5. Production and capacities of EDTA and related chelating agents in 1985 [22]

	1985 Estimated production, 10^3 t			Approximate total capacity, 10^3 t
	EDTA	DTPA	HEEDTA	
Japan	4.6	1.1	0.3	9.0
Western Europe	14.3	0.5	0.3	21.0
North America	21.4	3.1	2.4	55.0

9. Economic Aspects

Consumption of EDTA and related chelating agents in Japan, North America, and Western Europe is shown in Table 5. Future U.S. market growth is estimated to be low, i.e., 3 – 4 % annually [23]. Production capacities are approximate because much of the tonnage of chelating agents is produced in flexible plant equipment capable of being used for other manufacturing processes.

The principal manufacturers in Japan are Chelest Chemical Co., Osaka and Teikoku Chemical Industries, Itami. Major Western European manufacturers include BASF, Ludwigshafen, Federal Republic of Germany; Akzo, Herkenbosch, the Netherlands, and Düren, Federal Republic of Germany; Rexolin (W. R. Grace & Co.), Helsingborg, Sweden, and Middlesborough, United Kingdom; Protex, St. Avold, France; and ABM Chemicals, Gloucester, United Kingdom. Three producers account for >90 % of U.S. production: Dow Chemical, Freeport, Texas; Ciba-Geigy, Mcintosh, Alabama; and W. R. Grace & Co. at Nashua, New Hampshire and Deer Park, Texas.

10. Toxicology and Occupational Health

For many years, EDTA and its salts have been used as food additives and preservatives in pharmaceuticals and cosmetics without incident. Early toxicological work indicated that EDTA is relatively nontoxic [24]. Acute oral LD_{50} value (rat) for Na_2EDTA is 2.0 – 2.2 g/kg. Subacute studies using 0.5 – 5.0 % Na_2EDTA in the diet found no toxic manifestations except diarrhea and lowered food consumption at the highest feeding level. Skin sensitization studies have shown that EDTA as the trisodium salt is not a skin sensitizer [25]. The genetic toxicology of EDTA has been reviewed extensively [26]. Because of its sequestering ability, EDTA is responsible for functional and structural alterations of genetic material, though it seems to be harmless to humans as far as genotoxicity is concerned. When Na_3EDTA was examined for possible carcinogenicity, no compound-related signs of clinical toxicity were found [27].

11. References

1. A. E. Martell, R. M. Smith: *Critical Stability Constants,* Plenum Press, New York 1974.
2. J. R. Hart, *Soap, Cosmet., Chem. Spec.* **56** (1980) no. 6, 39 – 40, 42, 74.
3. M. S. Antleman: *The Encyclopedia of Chemical Electrode Potentials,* Plenum Press, New York 1982.
4. I. G. Farbenindustrie, DE 718 981, 1935 (F. Munz). General Aniline Works, US 2 130 505, 1938 (F. Munz).
5. Carbide & Carbon Chemical Corp., US 2 384 818, 1945 (G. Curme, Jr., J. W. Clark).
6. Martin Dennis Co., US 2 387 735, 1945 (F. C. Bersworth).
7. Hampshire Chemical Corp., US 3 061 628, 1962 (J. J. Singer, Jr., M. Weisberg).
8. W. R. Grace & Co., Hampshire Chelating Agents Technical Bulletin, Lexington, Mass., 1985, p. 9.
9. G. Schwarzenbach, W. Biederman, *Helv. Chim. Acta* **31** (1948) 678.
10. J. R. Hart in J. J. Kabara (ed.): *Cosmetic and Drug Preservation – Principles and Practice,* Marcel Dekker, New York 1984, p. 333.
11. R. Hart, *Household Pers. Prod. Ind.* **16** (1979) no. 6, 54 – 55, 57, 79.
12. D. R. Bambrick, *Tappi* **68** (1985) no. 6, 96 – 100.
13. J. R. Hart, *Tappi* **64** (1981) no. 3, 43 – 44.

14. J. R. Hart, *Pulp & Pap.* **55** (1981) no. 6, 138 – 140.
15. J. R. Hart, *Cosmet. Toiletries* **98** (1983) no. 4, 54 –58.
16. R. G. Young, *Soap, Cosmet., Chem. Spec.* **60** (1984) no. 11, 37 – 38, 89 – 90.
17. T. E. Furia: *CRC Handbook of Food Additives,* CRC Press, Cleveland, Ohio, 1977, pp. 271 – 294.
18. W. R. Grace & Co., Chelating Agents in Oxidation-Reduction Reactions, Lexington, Mass., 1986.
19. J. R. Hart, *Soap, Cosmet., Chem. Spec.* **62** (1986) no. 5, 38, 41, 48.
20. J. R. Hart, *Solutions* **26** (1982) no. 7, 63 – 66, 68, 70, 72.
21. F. J. Welcher: *The Analytical Uses of Ethylenediaminetetraacetic Acid,* Van Nostrand, Princeton, N.J., 1958.
22. Personal communication, W. R. Grace & Co., Lexington, Mass., 1986.
23. P. Mann, *Chem. Mark. Rep.* **229** (1986) no. 18, 36.
24. S. S. Yang, *Food Cosmet. Toxicol.* **2** (1964) 763 –767.
25. J. W. Henck, D. D. Lockwood, K. J. Olsen, *Drug Chem. Toxicol.* **3** (1980) 99 – 103.
26. K. Heindorff, O. Aurich, A. Michalis, R. Rieger, *Mutat. Res.* **115** (1983) 149 – 173.
27. National Cancer Institute: "Bioassay of Trisodium Ethylenediaminetetraacetate Trihydrate (EDTA) for Possible Carcinogenicity", Technical Report Series no. 11, 1977, Report NCI-CG-TR-11.

Ethylenimine → Aziridines
Ethyl Ether → Ethers, Aliphatic

Ethylene Glycol

Ethylene oxide is a separate keyword

SIEGFRIED REBSDAT, Hoechst Aktiengesellschaft, Gendorf, Federal Republic of Germany (Chaps. 1 – 10)

DIETER MAYER, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 11)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 3
4. Production 5
4.1. Ethylene Oxide Hydrolysis 5
4.1.1. Current Production Method 5
4.1.2. Possible Developments 5
4.2. Alternative Methods of Ethylene Glycol Production 7
4.2.1. Direct Oxidation of Ethylene 7
4.2.2. Synthesis from C_1 Units 8
5. Environmental Protection and Ecology 8
6. Quality Specifications and Analysis 8
7. Storage and Transportation 9
8. Derivatives 9
8.1. Di-, Tri-, Tetra-, and Polyethylene Glycols 9
8.2. Ethers and Esters 10
9. Uses 11
10. Economic Aspects 13
11. Toxicology and Occupational Health 13
11.1. Ethylene Glycol 13
11.2. Ethylene Glycol Derivatives 14
12. References 15

1. Introduction

Ethylene glycol [*107-21-1,*] 1,2-ethanediol, $HOCH_2CH_2OH$, M_r62.07, usually called glycol, is the simplest diol. It was first prepared by WURTZ in 1859 [1]; treatment of 1,2-dibromoethane [*106-93-4*] with silver acetate yielded ethylene glycol diacetate, which was then hydrolyzed to ethylene glycol.

Ethylene glycol was first used industrially in place of glycerol during World War I as an intermediate for explosives (ethylene glycol dinitrate) [2], but has since developed into a major industrial product.

The worldwide capacity for the production of ethylene glycol via the hydrolysis of ethylene oxide [*75-21-8*] (→ Ethylene Oxide) is estimated to be ca. 7×10^6 t/a.

Ethylene glycol is used mainly as an antifreeze in automobile radiators (→ Antifreezes) and as a raw material for the manufacture of polyester fibers (→ Fibers, 4. Synthetic Organic; → Polyesters).

2. Physical Properties

Ethylene glycol is a clear, colorless, odorless liquid with a sweet taste. It is hygroscopic and completely miscible with many polar solvents, such as water, alcohols, glycol ethers, and acetone. Its solubility is low, however, in nonpolar solvents, such as benzene, toluene, dichloroethane, and chloroform. The UV, IR, NMR, and Raman spectra of ethylene glycol are given in [3]. Following are some of the physical properties of ethylene glycol [4–6]:

bp at 101.3 kPa	197.60 °C
fp	−13.00 °C
Density at 20 °C	1.1135 g/cm^3
Refractive index, n_D^{20}	1.4318
Heat of vaporization at 101.3 kPa	52.24 kJ/mol
Heat of combustion	19.07 MJ/kg
Critical data	
Temperature	372 °C
Pressure	6515.73 kPa
Volume	0.186 L/mol
Flash point	111 °C
Ignition temperature	410 °C
Lower explosive limit	3.20 vol %
Upper explosive limit	53 vol %
Viscosity at 20 °C	19.83 mPa · s
Cubic expansion coefficient at 20 °C	0.62×10^{-3} K^{-1}

Ethylene glycol is difficult to crystallize; when cooled, it forms a highly viscous, supercooled mass that finally solidifies to produce a glasslike substance.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_101

The widespread use of ethylene glycol as an antifreeze is based on its ability to lower the freezing point when mixed with water. The physical properties of ethylene glycol – water mixtures are, therefore, extremely important. The freezing points of mixtures of water with monoethylene glycol and diethylene glycol [*111-46-6*] are shown in Figure 1. The temperature dependencies of the thermal conductivity, density, and viscosity of ethylene glycol and ethylene glycol – water mixtures are shown in Figures 2, 3, and 4 respectively [7]. The Prandtl numbers (the ratio of the viscosity to the thermal conductivity) derived from these values are given in Figure 5 [7]. The vapor pressures of ethylene glycol – water mixtures have been obtained from [8] by interpolation and are listed in Table 1.

Table 1. Vapor pressure of ethylene glycol – water mixtures

Water content, wt %	Vapor pressure, in kPa at 65.1 °C	77.7 °C	90.3 °C
0	0.30	0.52	1.20
10	6.61	11.65	19.73
20	11.30	19.68	33.01
30	14.70	25.45	42.49
40	17.10	29.68	49.37
50	18.81	32.92	54.60
60	20.16	35.58	58.87
70	21.45	37.92	62.60
80	22.98	40.05	65.98
90	25.08	41.91	68.93
100	28.04	43.34	71.10

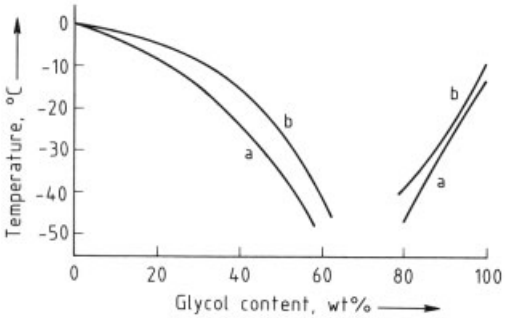

Figure 1. Freezing points of mono- and diethylene glycol – water mixtures
a) Monoethylene glycol; b) Diethylene glycol

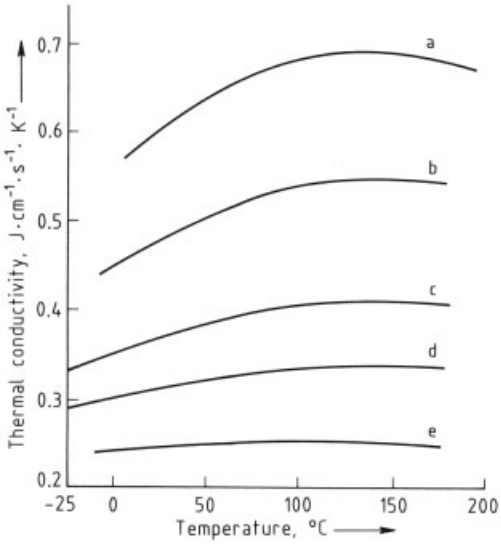

Figure 2. Temperature dependence of the thermal conductivity of ethylene glycol – water mixtures
Ethylene glycol content, mol %: a) 0; b) 25; c) 55; d) 75; e) 100

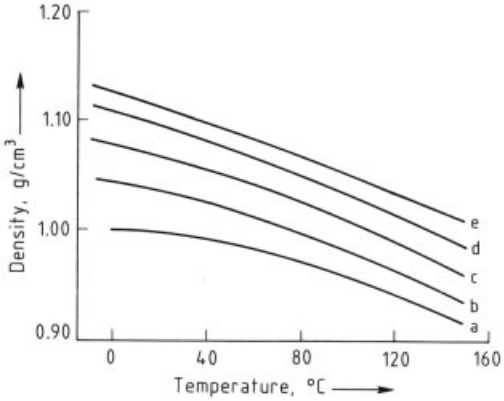

Figure 3. Temperature dependence of the density of ethylene glycol – water mixtures
Ethylene glycol content, mol %: a) 0; b) 26.1; c) 50.95; d) 76.9; e) 100

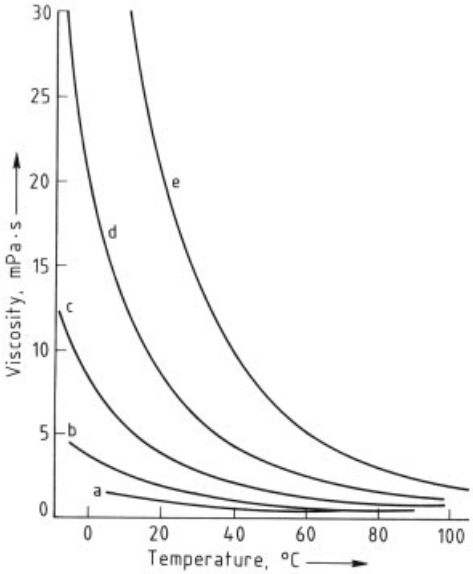

Figure 4. Temperature dependence of the viscosity of ethylene glycol – water mixtures
Ethylene glycol content, mol %: a) 0; b) 25; c) 49.90; d) 74.36; e) 100

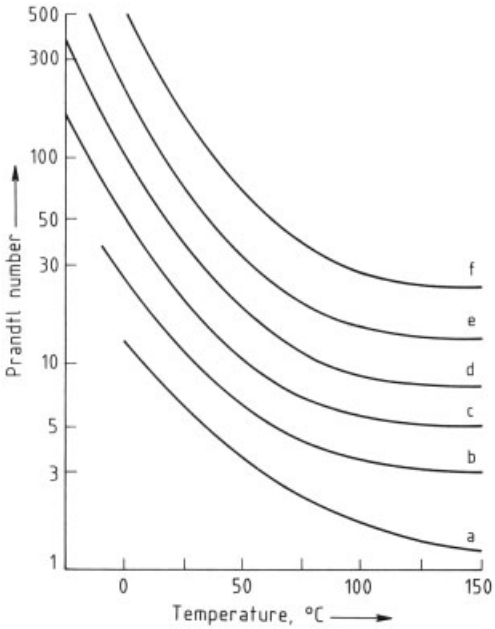

Figure 5. Temperature dependence of the Prandtl numbers of ethylene glycol – water mixtures
Ethylene glycol content, mol %: a) 0; b) 20; c) 40; d) 60; e) 80; f) 100

3. Chemical Properties

Ethylene glycol, like other alcohols, undergoes the reactions typical of its hydroxyl groups, which are described elsewhere (→ Alcohols, Aliphatic). Thus, only the special chemical characteristics and industrially important reactions of ethylene glycol are considered here. The two adjacent hydroxyl groups allow cyclization, and polycondensation; one or both of these functional groups may, of course, also react to give other derivatives.

Oxidation. Ethylene glycol is easily oxidized to form a number of aldehydes and carboxylic acids by oxygen, nitric acid, and other oxidizing agents. The typical products derived from the alcoholic functions are glycolaldehyde ($HOCH_2CHO$) [*141-46-8,*] glycolic acid ($HOCH_2COOH$) [*79-14-1,*] glyoxal (CHOCHO) [*107-22-2,*] glyoxylic acid (HCOCOOH) [*298-12-4,*] oxalic acid (HOOCCOOH) [*144-62-7,*] formaldehyde (HCHO) [*50-00-0,*] and formic acid (HCOOH) [*64-18-6*]. Many of these compounds are described in separate articles. Variation of the reaction conditions can lead to the selective formation of a desired oxidation product. Gas-phase oxidation with air in the presence of copper catalysts is of industrial importance for the production of glyoxal (→ Glyoxal and → Glyoxylic Acid). Glycol cleavage occurs in acidic solution with certain oxidizing agents such as permanganate, periodate, or lead tetraacetate. Cleavage of the C-C bond mainly produces formaldehyde, some of which is further oxidized to formic acid [9].

1,3-Dioxolane Formation. 1,3-Dioxolanes are formed by reacting ethylene glycol with carbonyl compounds [10]:

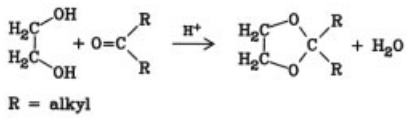

Acetalization to the cyclic 1,3-dioxolane proceeds more readily than acetal formation from straight-chain alcohols. If water is removed from the reaction mixture, an excellent yield can be

obtained. This reaction is used to protect carbonyl groups in organic syntheses.

1,3-Dioxolanes can also be formed from ethylene glycol by transacetalization. Examples are the reactions of ethylene glycol with orthoformates [11]:

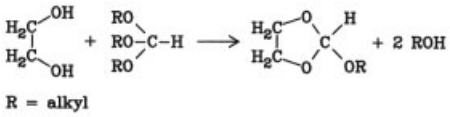

or with dialkyl carbonates:

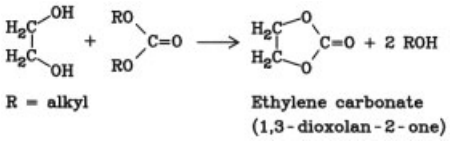

1,4-Dioxane Formation. Ethylene glycol can be converted to dioxane [*123-91-1*] by dehydration in the presence of acidic catalysts [12]:

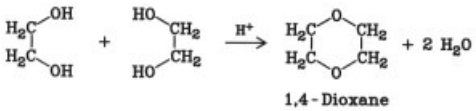

Ether and Ester Formation. Ethylene glycol can be alkylated or acylated by the customary methods to form ethers or esters, respectively. However, the presence of two hydroxyl groups leads to the formation of both mono- and diethers and mono- and diesters, depending on the initial concentrations of the individual reactants. The esterification of ethylene glycol with terephthalic acid [*100-21-0*] to form polyesters is especially important (→ Fibers, 4. Synthetic Organic; → Polyesters).

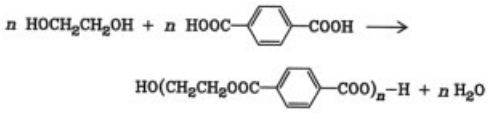

Ethoxylation. Ethylene glycol reacts with ethylene oxide to form di-, tri-, tetra-, and polyethylene glycols. The proportions of these glycols found in the reaction product are determined by the catalyst system that is used and the glycol excess. A considerable excess of glycol is required to obtain the lower homologues in a satisfactory yield. This reaction is rarely used industrially since these homologues are formed as byproducts during the production of ethylene glycol (cf. Section 4.1.1).

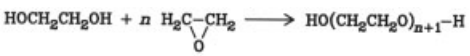

Decomposition with Alkali Hydroxide. Glycol is a relatively stable compound, but special care is required when ethylene (or diethylene) glycol is heated at a higher temperature in the presence of a base such as sodium hydroxide. Fragmentation of the molecule begins at temperatures above 250 °C and is accompanied by the exothermic evolution of hydrogen ($\Delta H = -90$ to -160 kJ/kg) [13]. This leads to a buildup of pressure in closed vessels.

4. Production

Although ethylene glycol has been known since 1859 (Wurtz) [1], it was not produced industrially until World War I. Its synthesis was then based on the hydrolysis of ethylene oxide [*75-21-8*] produced by the chlorohydrin process (→ Chlorohydrins, Chap. 4.). Production from formaldehyde [*50-00-0*] and carbon monoxide was also used commercially from 1940 to 1963 [14]. Neither of these methods is now used, however; the older literature should be consulted for details [2], [12]. Direct oxidation of ethylene [*74-85-1*] to ethylene glycol was also employed commercially for a short time [15], but was abandoned, probably due to problems caused by corrosion [16].

4.1. Ethylene Oxide Hydrolysis

4.1.1. Current Production Method

Only one method is currently used for the industrial production of ethylene glycol. This method is based on the hydrolysis of ethylene oxide obtained by direct oxidation of ethylene with air or

oxygen (→ Ethylene Oxide). The ethylene oxide is thermally hydrolyzed to ethylene glycol without a catalyst. Figure 6 shows a simplified scheme of a plant producing ethylene glycol by this method. The ethylene oxide – water mixture is preheated to ca. 200 °C, whereby the ethylene oxide is converted to ethylene glycol. Di-, tri-, tetra-, and polyethylene glycols are also produced, but with respectively decreasing yields (see also Chap. 3).

The formation of these higher homologues is inevitable because ethlyene oxide reacts with ethylene glycols more quickly than with water; their yields can, however, be minimized if an excess of water is used – a 20-fold molar excess is usually employed. Figure 7 shows the composition of the resulting product mixture as a function of the ratio of water to ethylene oxide. Although the values were determined by using sulfuric acid as a catalyst [17], they also apply as a good approximation for the reaction without a catalyst. Thus, in practice almost 90 % of the ethylene oxide can be converted to monoethylene glycol, the remaining 10 % reacts to form higher homologues:

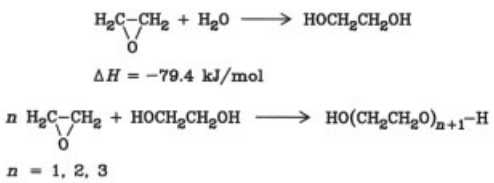

After leaving the reactor, the product mixture is purified by passing it through successive distillation columns with decreasing pressures. Water is first removed and returned to the reactor, the mono-, di-, and triethylene glycols are then separated by vacuum distillation. The yield of tetraethylene glycol is too low to warrant separate isolation. The heat liberated in the reactor is used to heat the distillation columns. A side stream must be provided to prevent the accumulation of secondary products, especially small amounts of aldehydes, which are produced during hydrolysis. The shape of the reactor affects the selectivity of the reaction. Plug-flow reactors are superior to both agitator-stirred tanks and column reactors [18].

4.1.2. Possible Developments

The glycol production method described in Section 4.1.1 is the only one of current industrial importance. It is simple, but has some major drawbacks:

1) The selectivity of the first step – the production of ethylene oxide – is low (ca. 80 %).
2) The selectivity of ethylene oxide hydrolysis is low – ca. 10 % is converted to di- and triethylene glycol.
3) Energy consumption for the distillation of the large amount of excess water is high.

Therefore, much research has been carried out to improve this process. The search for better silver catalysts is an objective for point 1 (→ Ethylene Oxide). Points 2 and 3 must be considered together, since higher selectivity for ethylene oxide hydrolysis automatically reduces the excess of water required.

Many catalysts have been described in the literature that are able to optimize selectivity or lower the reaction temperature and the required excess of water. Acids and bases are known to accelerate the reaction rate. The kinetics of the acid [19] and base [20] catalysis of ethylene oxide hydrolysis have been thoroughly investigated; mechanisms are discussed in [21]. The industrial feasability of catalysis with ion-exchange columns in the liquid phase [22], [23] and the gas phase [24] has been tested. Although the use of catalysts allowed the reaction temperature to be lowered, selectivity was not significantly enhanced. Furthermore, the catalyst needed to be separated and either fed back into the reaction mixture or replaced. As a result of these disadvantages, these types of catalysis have not proved to be of commercial use. However, catalysts that improve selectivity have been described in patents; they include molybdates [25], vanadates [26], ion exchangers [27], and organic antimony compounds [28]. However, their advantages do not yet seem to justify their use on an industrial scale.

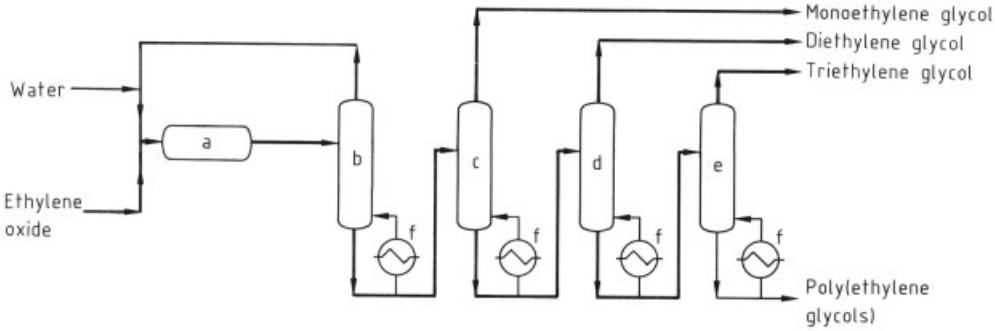

Figure 6. Flow diagram for a glycol plant
a) Reactor; b) Drying column; c) Monoethylene glycol column; d) Diethylene glycol column; e) Triethylene glycol column; f) Heat exchanger

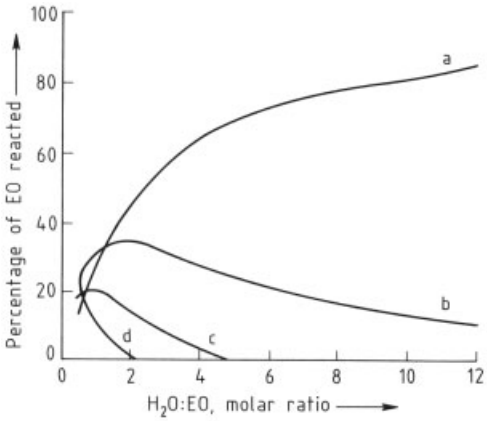

Figure 7. Composition of the product obtained on hydrolysis of ethylene oxide (EO) as a function of the water to ethylene oxide ratio
a) Monoethylene glycol; b) Diethylene glycol; c) Triethylene glycol; d) Higher poly(ethylene glycols)

The selective synthesis of ethylene glycol via the intermediate ethylene carbonate (1,3-dioxolan-2-one) [*96-49-1*] seems to be a promising alternative. This compound is obtained in high yield (98 %) by reacting ethylene oxide with carbon dioxide and can be selectively hydrolyzed to give a high yield of ethylene glycol. Only double the molar quantity of water is required for this reaction.

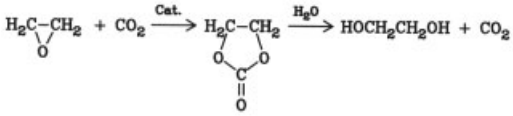

According to a Halcon patent, ethylene oxide can be extracted from the aqueous solution, formed during its production, with supercritical carbon dioxide [29]. An ethylene oxide – carbon dioxide solution is obtained, which reacts to form ethylene carbonate. Hydrolysis of the ethylene carbonate then yields ethylene glycol. Possible catalysts for this reaction are quaternary ammonium and phosphonium salts, such as R_4NHal, R_4PHal, or Ph_3PCH_3I. Problems such as product separation and catalyst feedback still need to be resolved, but this method for the selective synthesis of ethylene glycol from ethylene oxide seems to be the most promising for industrial-scale application.

4.2. Alternative Methods of Ethylene Glycol Production

The low selectivity of ethylene oxide production and increasing ethylene prices warrant the search for alternative ways of producing ethylene glycol.

4.2.1. Direct Oxidation of Ethylene

As mentioned earlier, catalytic oxidation of ethylene [*74-85-1*] with oxygen in acetic acid has already been used on an industrial scale, but this method was soon abandoned due to problems caused by corrosion. The yield of ethylene glycol (>90 %) was much higher than that obtained in the more indirect route via ethylene oxide [30].

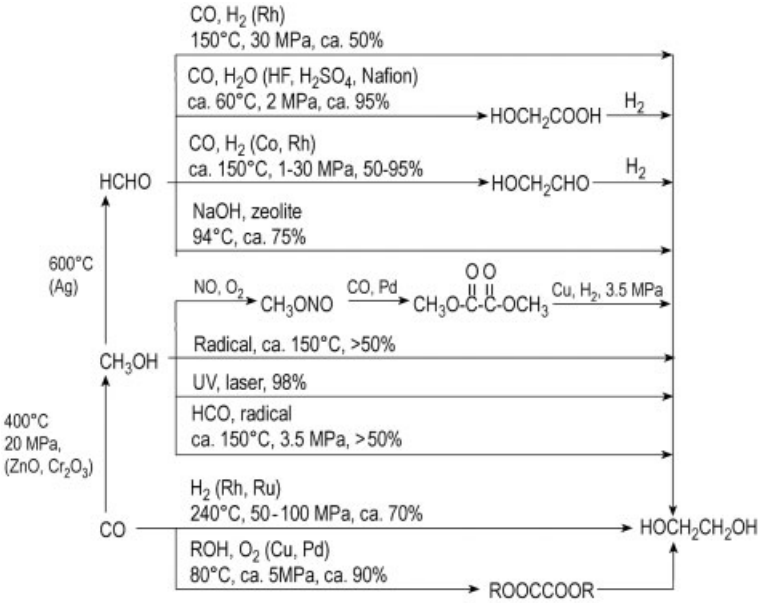

Figure 8. Production of ethylene glycol from carbon monoxide (percentages indicate approximate yields)

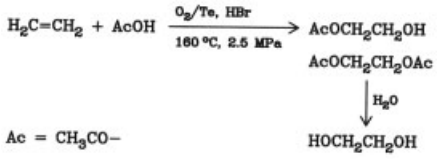

A more recently developed catalyst system is based on the use of Pd(II) complexes [31]. A mixture of $PdCl_2$, LiCl, and $NaNO_3$ in acetic acid and acetic anhydride has been shown to give a 95 % selectivity for glycol monoacetate and glycol diacetate formation (60 – 100 °C, 3.04 MPa) [32]. During this process, Pd(II) is reduced to Pd(0). The precipitation of Pd(0) is prevented because it is reoxidized to Pd(II) by the nitrate ions. The available oxygen finally regenerates the nitrate, thus providing a complete catalytic system.

The formation of ethylene glycol monoacetate [*542-59-6*] (50 % yield) and ethylene glycol diacetate [*111-55-7*] (7 % yield) has also been investigated using the catalyst system $PdCl - NO_2 - CH_3CN$ dissolved in acetic acid. Studies with radioactive isotopes showed that the NO_2 functions as an oxidizing agent [33]. Vinyl acetate is formed as a byproduct (20 % yield). However, the catalytic action of this system is quickly exhausted due to the precipitation of palladium compounds.

If a $PdCl_2 - CuCl_2 - CuOCOCH_3$ system is used, the reaction proceeds under mild conditions (65 °C, 0.5 MPa) without the formation of a precipitate; a yield of over 90 % is obtained [34].

In recent years, increasing attention has been paid to Pd(II) systems as catalysts for the direct oxidation of ethylene to ethylene glycol. In spite of the widespread interest in this alternative, industrial applications have yet to be realized.

4.2.2. Synthesis from C_1 Units

The long-term shortage and increased price of crude oil have led to an intensive search for methods of producing organic intermediates from C_1 units (i.e., methods based on coal). Many publications have appeared on the synthesis of ethylene glycol by this approach. Only the most important methods that rely on synthesis gas or carbon monoxide [*630-08-0*] are discussed here; they are summarized in Figure 8.

At a high pressure, carbon monoxide and hydrogen react directly to produce ethylene glycol [35], [36]. However, the reaction is slow and the catalyst is both sensitive and expen-

sive [37]. Other methods involve the formation of formaldehyde [*50-00-0*], methanol [*67-56-1*] [38], or esters of oxalic acid [*144-62-7*] as intermediates [37], [39]. The only method to attain industrial importance was that employed by Du Pont from 1940 to 1963, which used formaldehyde and glycolic acid [*79-14-1*] as intermediates. High operating pressure and temperature were required, however (48 MPa, 220 °C). This process was significantly improved by the introduction of hydrogen fluoride [*7664-39-3*] as a catalyst (1 – 2 MPa, 60 °C) [40].

At the present time, none of the described methods based on C_1 units can compete with the ethylene → ethylene oxide → glycol pathway. However, if crude oil prices increase, the synthesis of ethylene glycol from C_1 units will become more economically attractive [41].

5. Environmental Protection and Ecology

Ethylene glycol is readily biodegradable [42]; thus, disposal of wastewater containing this compound can proceed without major problems. The high LC_{50} values of over 10 000 mg/L [43], [44] account for its low water toxicity: LC_{50} crayfish (*Procambarus*) 91 000 mg/L, LC_{50} fish (*Lepomis macrochirus*) 27 540 mg/L.

6. Quality Specifications and Analysis

Since ethylene glycol is produced in relatively high purity, differences in quality are not expected. The directly synthesized product meets high quality demands ("fiber grade"). The ethylene glycol produced in the wash water that is used during ethylene oxide production is normally of a somewhat inferior quality ("antifreeze grade"). The quality specifications for mono-, di-, and triethylene glycols are compiled in Table 2 [45]. The UV absorption of fiber-grade glycol is often used as an additional parameter for quality control.

Table 2. Quality specifications of mono-, di-, and triethylene glycols [45]

Property	Method	Monoethylene glycol		Diethylene glycol	Triethylene glycol
		Antifreeze grade	Fiber grade		
Purity, %	Gas chromatography	>98.00	>99.80	99.60	99.50
Diethylene glycol content, wt %	Gas chromatography	<0.50.	<0.10		
Boiling range (at 101.3 kPa), °C	DIN 53171, ASTM D 1078	195 – 200	196 – 199	242 – 247	285 – 295
Density (20 °C), g/cm^3	DIN 51757, ASTM D 1122	1.113 – 1.115	1.1135 – 1.1140	1.1160 – 1.1175	1.1235 – 1.1245
Refractive index, n_D^{20}	DIN 53491, ASTM 1747	1.431 – 1.433	1.4315 – 1.4320	1.4460 – 1.4475	1.4555 – 1.4565
Water content, wt %	DIN 51777, ASTM E 203	<0.50	<0.100	<0.20	<0.10
Acid number, mg of KOH/g	DIN 53402, ASTM D 1613	<0.01	<0.005	<0.05	<0.05

Gas chromatography is commonly used for the quantitative determination of ethylene glycol. Monoethylene glycol can be detected by oxidation with periodic acid even if di- and triethylene glycols are also present; however, aldehydes, glycerol, and monopropylene glycol falsify the results [4].

7. Storage and Transportation

Pure anhydrous ethylene glycol is not aggressive toward most metals and plastics. Since ethylene glycol also has a low vapor pressure and is noncaustic, it can be handled without any problems; it is transported in railroad tank cars, tank trucks, and tank ships. Tanks are usually made of steel; high-grade materials are only required for special quality requirements. Nitrogen blanketing can protect ethylene glycol against oxidation.

At ambient temperatures, aluminum is resistant to pure glycol. Corrosion occurs, however, above 100 °C and hydrogen is evolved. Water, air, and acid-producing impurities (aldehydes) accelerate this reaction. Great care should be taken when phenolic resins are involved, since they are not resistant to ethylene glycol.

8. Derivatives

Only the most important of the many derivatives of ethylene glycol will be discussed in this section, namely di-, tri-, and polyethylene glycols, the methyl, ethyl, and butyl glycol ethers and the acetate esters.

8.1. Di-, Tri-, Tetra-, and Polyethylene Glycols

The higher homologues of ethylene glycol are formed as byproducts during the synthesis of ethylene oxide and monoethylene glycol or are prepared directly by reacting monoethylene glycol with ethylene oxide. Di-, tri-, and possibly tetraethylene glycol can be purified by distillation. For quality specifications of these compounds, see Table 2; Table 3 shows some of their physical properties. The temperature dependence of the thermal conductivity, density, viscosity, and Prandtl numbers of di- and triethylene glycol – water mixtures is described in [46].

Poly(ethylene glycols) (PEG) [*25322-68-3*] are produced as mixtures of higher molecular mass homologues (see also → Polyoxyalkylenes) and are characterized by their mean molecular masses (e.g., PEG 400). Physical properties of some poly(ethylene glycol) mixtures are listed in Table 4.

8.2. Ethers and Esters

Ethers. Ethylene glycol monoethers are usually produced by reaction of ethylene oxide with the appropriate alcohol. A mixture of homologues is obtained, all displaying the following general structure:

$RO(CH_2CH_2O)_n–H$

The glycol monoethers can be converted to diethers by alkylation with common alkylating agents, such as dimethyl sulfate or alkyl halides (Williamson synthesis). Glycol dimethyl ethers are formed by treatment of dimethyl ether with ethylene oxide [48]. The properties of some ethylene glycol mono- and diethers are shown in Tables 5 and 6, respectively; they are mainly used as solvents.

Esters. Ethylene glycol esters are produced from reaction of ethylene oxide and an appropriate acid and are used also as solvents. Use of a suitable catalyst can influence the outcome of the reaction so as to produce monoesters (sodium acetate) or diesters (sulfuric acid). Glycol monoethers can be converted to the corresponding ether esters by the usual methods. Some important properties of ethylene glycol esters and ether esters are listed in Table 7.

Trade Names. Some common trade names of ethylene glycol ethers and esters are as follows:

Monoethylene Glycol Derivatives. Monomethyl ether: Methyl Cellosolve Solvent, Dowanol EM Glycol Ether, Methyl Oxitol Glycol. Monoethyl ether: Cellosolve Solvent, Dowanol EE Glycol Ether. Monobutyl ether: Butyl Cellosolve Solvent, Dowanol EB Glycol Ether, Butyl Oxitol Glycol. Dimethyl ether (also

Table 3. Physical properties of mono-, di-, tri-, and tetraethylene glycols [4], [5], [46]

Property	Monoethylene glycol	Diethylene glycol	Triethylene glycol	Tetraethylene glycol
Molecular formula	$HOCH_2CH_2OH$	$H(OCH_2CH_2)_2OH$	$H(OCH_2CH_2)_3OH$	$H(OCH_2CH_2)_4OH$
CAS registry number	[*107-21-1*]	[*111-46-6*]	[*112-27-6*]	[*112-60-7*]
M_r	c62.7	106.12	150.17	194.23
bp(at 101.3 kpa), ºC	197.6	244.8	287.4	decomp.
fp, °C	–13	–8	–7	–4.1
Vapor pressure (20 °C), Pa	5.3	2.7	0.5	<1.3
Density (20 °C), g/cm^3	1.1130	1.1160	1.1230	1.1247
Refractive index, n 20 d	1.4318	1.4470	1.4560	1.4598
Heat of combustion, MJ/kg	19.07	22.32	23.68	
Heat of vaporization (101.3 kPa), kJ/mol	52.24	52.26	61.04	62.63
Viscosity (20 °C) mPa∗s	19.83	36.0	49.0	61.9
Surface tension, N/m^2	4.84 (20 °C)	4.85 (20 °C)	4.22 (25 °C)	
Flash point, °C	119	141	177	191
Ignition temperature, °C	410	390	370	
Lower explosive limit, vol %	3.2	0.7	0.9	

known as glyme): Dimethyl Cellosolve Solvent. Diethyl ether: Diethyl Cellosolve Solvent. Monomethyl ether acetate: Methyl Cellosolve Acetate. Monoethyl ether acetate: Cellosolve Acetate, Poly-solv EE Acetate.

Table 4. Physical properties of poly(ethylene glycols) [49]

Poly(ethylene glycol), mean M_r	*fp*, °C	Refractive index, n_D^{70}	Viscosity, mPa · s ∗
200	ca. −50	1.458 (25 °C)	60 – 70 (100 %)
600	17 – 22	1.452	16 – 19
1000	35 – 40	1.453	24 – 29
2000	48 – 52	1.454	50 – 60
10 000	55 – 60	1.456	530 – 1000
20 000	ca. 60	1.456	2700 – 3500
35 000	ca. 60	1.456	11 000 – 14 000

∗ Values refer to a 50 wt % aqueous poly(ethylene glycol) solution except where otherwise stated.

Diethylene Glycol Derivatives.Monomethyl ether: Methyl Carbitol Solvent, Dowanol DM Glycol Ether. Monoethyl ether: Carbitol Solvent, Dowanol DE Glycol Ether. Monomethyl ether acetate: Methyl Carbitol Acetate. Monoethyl ether acetate: Carbitol Acetate. Monobutyl ether acetate: Butyl Carbitol Acetate.

Diethylene glycol dimethyl ether is also known as diglyme.

Triethylene Glycol Derivatives. Monomethyl ether: Dowanol TM Glycol Ether.

9. Uses

Ethylene glycol lowers the freezing point of water (see Fig. 1). Its ease of handling makes it a perfect antifreeze [51], which accounts for over 50 % of its commercial uses (→ Antifreezes).

Commercial antifreezes based on glycol also contain corrosion inhibitors and are used, for example, in motor vehicles, solar energy units, heat pumps, water heating systems, and industrial cooling systems. Protection against freezing is directly related to the glycol concentration; 60 % glycol prevents freezing down to a temperature of −55 °C. There is little point using higher concentrations because the freezing point then starts to increase (Fig. 1).

Ethylene glycol is also a commercially important raw material for the manufacture of polyester fibers, chiefly poly(ethylene terephthalate) (→ Fibers, 4. Synthetic Organic). This application consumes ca. 40 % of the total ethylene glycol production. Polyesters are, however, used for other purposes, e.g., for producing recyclable bottles (→ Polyesters). Other minor uses of ethylene glycol are as a humectant (moisture-retaining agent), plasticizer, softener, hydraulic fluid, and solvent [52].

Ethylene glycol derivatives, mainly the ethers and esters, are frequently employed as reaction media, as absorption fluids, and as solvents for dyes, lacquers, and various cellulose products (cellulose esters, ethers, and nitrocellulose) [53].

Table 5. Physical properties of some ethylene glycol monoalkyl ethers, $RO(CH_2CH_2O)_n$–H [4], [5]

R	*n*	CAS registry number	M_r	*fp*, °C	*bp* (101.3 kPa), °C	Density (20 °C), kg/m^3	Refractive index, n_D^{20}	Specific heat (20 °C), J g^{-1} K^{-1}	Vapor pressure (20 °C), Pa	Flash point, °C	Ignition temperature, °C	Explosive limits, vol %
CH_3	1	[*109-86-4*]	76.094	− 85.1	124.50	964.6	1.4024	2.24	930	38	285	2.4 – 20.6
C_2H_5	1	[*110-80-5*]	90.120	−100.0	135.10	929.7	1.4079	2.32	530	40	215	1.8 – 15.7
C_4H_9	1	[*111-76-2*]	118.172	− 70.4	171.20	900.8	1.4194	2.44	80	66	225	1.1 – 10.6
CH_3	2	[*111-77-3*]	120.146	− 50.0	194.10	1021.0	1.4265	2.15	24	90	215	1.6 – 16.1
C_2H_5	2	[*111-90-0*]	134.172	− 54.0	201.90	989.0	1.4275	2.31	17	93	190	1.8 – 12.2
C_4H_9	2	[*112-34-5*]	162.224	− 68.1	230.60	953.0	1.4316	2.29	4	99	195	0.7 – 5.9
CH_3	3	[*112-35-6*]	164.198	− 38.2	249.00	1047.5	1.4381	2.18		118		
C_2H_5	3	[*112-50-5*]	178.224	− 18.7	255.50	1020.0	1.4380	2.17 (15 °C)	0.4	195		

Table 6. Physical properties of some ethylene glycol dialkyl ethers $RO(CH_2CH_2O)_n$-R [2], [4], [50]

R	*n*	CAS registry number	M_r	*fp*, °C	*bp* (101.3 kPa), °C	Density (20 °C), kg/m^3	Refractive index, n_D^{20}	Specific heat (20 °C), J g^{-1} K^{-1}	Vapor pressure (20 °C), Pa	Flash point, °C	Ignition temperature, °C	Lower explosive limit, vol %
CH_3	1	[*110-71-4*]	90.120	−58.0	85.2	867.0	1.3796		8100.0	−6	200	1.6
C_2H_5	1	[*629-14-1*]	118.172	−74.0	121.4	841.0	1.3920		1250.0	35	205	
C_4H_9	1	[*112-48-1*]	174.276	−69.1	203.6	836.0	1.4131	2.01	27.0	85		
CH_3	2	[*111-96-6*]	134.172	−64.0	163.0	944.0	1.4078	2.04	267.0	51		1.4
C_2H_5	2	[*112-36-7*]	162.224	−44.3	188.4	906.6	1.4115	2.10 (15 °C)	51.0	82		
C_4H_9	2	[*112-73-2*]	218.328	−60.2	254.6	883.8	1.4233	1.80 (13 °C)	1.3	118		
CH_3	3	[*112-35-6*]	178.224	−45.0	216.0	987.0	1.4233		120.0	113	195	0.7

Table 7. Physical properties of some ethylene glycol acetates and ether acetates [4]

Compound	CAS registry number	M_r	*fp*, °C	*bp* (101.3 kPa), °C	Density (20 °C), kg/m^3	Refractive index, n_D^{20}	Ignition temperature, °C	Flash point, °C	Explosive limits, vol %
Ethylene glycol monoacetate	[*542-59-6*]	104.10	−80.0	182.0	1110	1.4209		102	7.8 – 27.7
Ethylene glycol diacetate	[*111-55-7*]	146.10	−31.0	190.5	1106	1.4159		105	
Ethylene glycol methyl ether acetate	[*110-49-6*]	118.13	−65.1	145.0	1006	1.4019	410	49	1.7 – 8.2
Ethylene glycol ethyl ether acetate	[*111-15-9*]	132.16	−61.7	156.4	974	1.4058	390	54	3.18 – 10.1
Ethylene glycol butyl ether acetate	[*112-07-2*]	160.21	−63.5	192.3	942	1.4138	355	74	1.66 – 8.37
Diethylene glycol methyl ether acetate	[*629-38-9*]	162.19	−99.3	209.1	1041	1.4209		110	
Diethylene glycol ethyl ether acetate	[*112-15-2*]	176.21	−25.0	217.4	1011	1.4213		112	
Diethylene glycol butyl ether acetate	[*124-17-4*]	204.26	−32.2	246.7	981	1.4262	290	116	

Diethylene glycol serves as a solvent, softener (cork, adhesives, paper, etc.), dye additive (printing and stamping inks), deicing agent for runways and aircraft [54], and a drying agent for gases (e.g., natural gas) [55].

Triethylene glycol is used for the same purposes as diethylene glycol (e.g., as a solvent, plasticizer, and drying agent for gases) [56].

Poly(ethylene glycols) with varying molecular masses find numerous uses in the pharmaceutical industry (e.g., in ointments, liquids, and tabletting) and the cosmetic industry (e.g., creams, lotions, pastes, cosmetic sticks, and soaps). They are also used in the textile industry (e.g., cleaning and dyeing aids), in the rubber industry (e.g., lubricants and mold parting agents), and in ceramics (e.g., bonding agents and plasticizers) [57].

10. Economic Aspects

Ethylene glycol is one of the major products of the chemical industry. Its economic importance is founded on its two major commercial uses – as an antifreeze and for fiber production. Since ethylene glycol is currently produced exclusively from ethylene oxide (→ Ethylene Oxide), production plants are always located close to plants that produce ethylene oxide. The proportion of ethylene oxide that is converted to glycol depends on the local conditions, such as transport facilities and the market situation. Almost all of the ethylene oxide produced in recently installed plants in Saudi Arabia is, for example, converted to ethylene glycol because the domestic market for ethylene oxide is negligible and because ethylene glycol is easier to transport than ethylene oxide.

Detailed information about the glycol capacities of individual plants is difficult to obtain, but an estimated 60 % of the total world production of ethylene oxide is converted to ethylene glycol. Therefore, the worldwide ethylene oxide production capacity of 8×10^6 t corresponds to an annual ethylene glycol production of 6.7×10^6 t.

About 50 % of the ethylene glycol that is produced is used in antifreeze. Another 40 % is channeled into the fiber industry. Consequently, the ethylene glycol demand is closely connected to the development of these two sectors. Surplus capacities and growing competition for market shares are expected on account of increasing glycol capacities – not only in Saudi Arabia, but also in the Eastern Bloc. Japan has already reduced its glycol capacities considerably [58]. In view of the increasing price of crude oil, alternative production methods based on synthesis gas – hitherto still in the experimental phase – are likely to become more important and increasingly competitive.

11. Toxicology and Occupational Health

11.1. Ethylene Glycol

Oral Toxicity. The oral LD_{50} values of ethylene glycol are 14 – 15 g/kg [59], [60] for mice and ca. 8 g/kg [61] for rats. Oral toxicity in humans is higher, 1.4 g/kg can be lethal [59].

In rats, long-term administration of ethylene glycol in food at concentrations of 10.0 and 20.0 g/kg led to formation of bladder stones (calcium oxalate), damage to the renal tubules, and centrolobular fatty degeneration of the liver [62]. In other investigations, comparable changes were already apparent at a concentration of 5.0 g/kg, no changes were observed at 2.0 g/kg [63]. No changes were observed in the kidneys of monkeys fed a diet providing a maximal daily dose of 0.17 g/kg of body mass over a period of three years [64]. However, addition of ethylene glycol to the drinking water given to monkeys at a concentration corresponding to a daily dose of 0.24 g/kg of body mass led to the formation of oxalate crystals in the kidneys after about five months. Oxalate crystals were also formed in the brains of monkeys given very high doses (10 wt %) in drinking water [65].

Eye and Skin Irritation. Introduction of a single dose of ethylene glycol into the conjunctival sac does not affect the rabbit eye [66], but repeated administration leads to mild conjunctivitis [67]. Ethylene glycol vapor at a concentration of 17 mg/m^3 produces ocular damage in humans [68]. No effects were observed in the eyes of monkeys exposed to a vapor concentration of 265 mg/m^3 [68].

Ethylene glycol has not been shown to irritate mucous membranes. Repeated exposure of the skin to ethylene glycol can cause mild irritation, similar to that produced by glycerol.

Inhalation. The inhalative toxicity of an atmosphere saturated with ethylene glycol (0.5 mg/L) is low. Rats tolerated this concentration for four weeks without any permanent injuries, only mild narcosis was observed [69].

No adverse effects were detected in rats, rabbits, guinea pigs, dogs, and monkeys exposed to an ethylene glycol concentration of 57 mg/L for six weeks [70]. Monkeys survived after inhaling concentrations of 600 mg/m^3 administered as an aerosol over a period of several months.

Human volunteers exposed to maximal concentrations of 67 mg/m^3 for one month reported some irritation of the nasal mucous membranes and occasionally headache. The presence of ethylene glycol in the inhaled atmosphere was perceived at concentrations above 140 mg/m^3 [71]. A concentration of 200 mg/m^3 was found to be intolerable [72]. Rats, rabbits, dogs, and monkeys that inhaled pure oxygen containing 100 mL/m^3 of ethylene glycol at a pressure of 34.5 kPa for three weeks did not show any adverse effects apart from local irritation [73].

Mutagenicity and Carcinogenicity. Ethylene glycol gives a negative result in the Ames test [74], [75]. Many studies have been performed on its carcinogenicity [64], [76], [77]. Rats were fed with diets containing ethylene glycol concentrations corresponding to 1000, 200, and 40 mg/kg of body mass; no carcinogenic effects were observed. However, typical formation of oxalate crystals in the kidneys and other organs was found [78]. Similar experiments did not increase the incidence of tumor formation in mice [79].

Reproduction. In a study extending over three generations, ethylene glycol was added to the diet fed to rats at a concentration of 0.04, 0.2, and 1.0 g kg^{-1} d^{-1}. The fertility and viability of the offspring were not affected by this treatment [80].

Teratogenic effects were not detected in rats receiving maximal feed concentrations of 1000 mg/kg, only a mild fetal toxicity was observed [81]. Teratogenic and fetotoxic effects were produced in mice and rats after administration of very high doses (1.25, 2.50, and 5.00 g/kg of body mass) in the form of a bolus with the aid of an esophageal probe [82].

Metabolism. In the body, ethylene glycol is degraded by alcohol dehydrogenase to form glycolaldehyde, which is further degraded by an aldehyde oxidase to produce glyoxal. Glyoxal is subsequently metabolized to glycolic and glyoxylic acids, from which oxalic acid, formic acid, glycine, 2-oxo-4-hydroxygluconate, 2-hydroxy-3-oxoadipate and oxomalate are formed [83, p. 3827].

11.2. Ethylene Glycol Derivatives

Toxicity data and exposure limits for ethylene glycols and some monoethylene glycol derivatives are listed in Table 8.

Diethylene glycol has a low acute oral toxicity (see Table 8). A LDLo value of 1.0 mL/kg is often erroneously cited but this value was determined using a preparation that contained sulfanilimide [85].

In long-term studies performed over a period of two years, dietary levels of 1, 2, and 4 % diethylene glycol administered to rats produced calcium oxalate stones with associated kidney and liver damage [86]. Chronic mechanical irritation of the bladder epithelium by the resulting bladder stones led to the formation of benign tumors. Production of calcium oxalate stones is attributed to contamination of the diethylene glycol with monoethylene glycol (see Section 11.1) [87], [88]. The low cumulative toxicity of diethylene glycol has been confirmed [89–91].

Diethylene glycol does not irritate the eyes or skin [92]; absorption by the skin is extremely low [93]. Diethylene glycol does not affect the fertility of rats [94]. It does not display mutagenicity in the Ames test [95]; other studies also suggest that the compound is not carcinogenic [87], [88].

Triethylene glycol has an extremely low oral toxicity when given either in a single dose (see Table 8) or in repeated doses [86], [96]. This compound causes only slight irritation of the skin and mucous membranes, it is not teratogenic [97].

Table 8. Toxicity data and exposure limits for ethylene glycols and ethylene glycol derivatives

Compound	Single oral LD_{50} in rats, g/kg	Reference	MAK, ppm	TLV-TWA, ppm
Ethylene glycols				
Monoethylene glycol	8	[61]	10	50
Diethylene glycol	20.8	[61]	10	
Triethylene glycol	22.1	[61]		
Tetraethylene glycol	30.8 *	[83, p. 3843]		
Monoethylene glycol derivatives				
Monomethyl ether (methoxyethanol)	3.4	[84]	5	5
Monomethyl ether acetate (methoxyethyl acetate)	3.93	[61]	5	5
Monoethyl ether (ethoxyethanol)	5.5	[84]	5	5
Monoethyl ether acetate (ethoxyethyl acetate)	5.1	[61]	5	5
Monobutyl ether (butoxyethanol)	2.5	[84]	20	25
Monobutyl ether acetate (butoxyethyl acetate)			20	

* mL/kg

Tetraethylene glycol is, like triethylene glycol, nontoxic (see Table 8).

Poly(ethylene glycols) have an extremely low acute oral toxicity (LD_{50} in rats > 30 g/kg) which decreases as their molecular mass increases [98]. Studies on rats [98] and dogs [99] have shown that their cumulative toxicity is also very low.

Irritation of the skin and mucous membranes in humans and animals is slight. When applied to the skin, pure poly(ethylene glycols) do not cause sensitization [98], [100], [101] and are nontoxic [98]. Poly(ethylene glycols) with a molecular mass < 1000 are absorbed in the intestine and excreted in the urine and feces. Intestinal absorption decreases markedly as the molecular mass increases [102–104].

12. References

1. A. Wurtz, *Ann. Chem. Pharm.* **55** (1859) no. 3, 406.
2. G. O. Curme: *Glycols,* Rheinhold Publ. Co., New York 1952.
3. J. Ariddick, B. Bunger: *Techniques of Chemistry,* 3rd ed., **vol. II,** Wiley-Interscience, New York 1970, p. 198.
4. *Ullmann,* 4th ed., **vol. 8,** 200 – 210.
5. *Kirk-Othmer,* **9, 11,** 933 – 956.
6. Merkblatt Lösemittel Hoechst, Hoechst AG, Frankfurt, 1984.
7. D. Bohne, S. Fischer, E. Obermeier, *Ber. Bundesges. Phys. Chem.* **88** (1984) 739 – 742.
8. A. Nath, E. Bender, *J. Chem. Eng. Data* **28** (1983) no. 4, 370 – 375.
9. K. K. Sengupta, D. K. Khamrui, D. C. Mukherjee, *Indian J. Chem.* **13** (1975) 348 – 351. I. Bhatia, K. K. Banerji, *J. Chem. Soc. Perkin Trans. 2,* 1983 no. 10, 1577 – 1580.
10. F. A. J. Meskens, *Synthesis* **7** (1981) 501 – 522.
11. M. Ando, M. Ohhara, K. Takase, *Chem. Lett.* **6** (1986) 879 – 882.
12. S. A. Miller: *Ethylene and its Industrial Derivatives,* Ernest Benn Limited, London 1969.
13. P. Cardillo, A. Girelli, *Chim. Ind. (Milan)* **64** (1982) no. 12, 781 – 784.
14. DuPont, US 2 285 448, 1942 (D. J. Loder).
15. A. M. Brownstein, *Hydrocarbon Process.* **71** (1975) no. 9, 72 – 76.
16. G. E. Weismantel, *Chem. Eng.* (N. Y.) 86 (1978) no. 2, 67 – 70.
17. C. Matignon, H. Moureu, M. Dode, *Bull. Soc. Chim. Fr.* **5** (1934) no. 1, 1308 – 1317.
18. J. H. Miller, T. E. Corrigan, *Hydrocarbon Process.* **46** (1967) no. 4, 176 – 178.
19. J. Bronsted, M. Kilpatrick, *J. Am. Chem. Soc.* **51** (1929) 428.
20. H. J. Lichtenstein, G. H. Twigg, *Trans. Faraday Soc.* **44** (1948) 905.
21. G. Lamaty, R. Maleq, C. Selve, A. Sivade, J. Wylde, *J. Chem. Soc. Perkin Trans. 2,* 1975 no. 10, 1119 –1124.

22. D. F. Othmer, M. S. Thakar, *Ind. Eng. Chem.* **50** (1958) no. 9, 1235 – 1244.
23. L. Klein, H. Fehrecke, *Chem. Ing. Tech.* **52** (1980) no. 10, 818 – 820.
24. G. E. Mamilton, A. B. Metzner, *Ind. Eng. Chem.* **49** (1957) no. 5, 838 – 846.
25. UCC, PCT Int. Appl., WO 8 504 393, 1985 (B. T. Keen). Nippon Shokubai Kagaku, BE 878 902, 1979 (T. Kumarawa, T. Yamamoto, H. Odanaka).
26. UCC, PCT Int. Appl., WO 8 504 406, 1985(J. Robert, J. H. Robson).
27. Texaco, EP-A 0 123 709, 1983 (F. L. Jonson, L. W. Watts).
28. Mitsubishi Petrochem. KK, JP 6 0061-545-A, 1983.
29. Halcon, DE-OS 3 321 448, 1983 (V. S. Bhise).
30. Halcon, DE-OS 2 757 067, 1977 (A. Peltzman).
31. A. Mitsutani, *CEER Chem. Econ. Eng. Rev.* **5** (1973) no. 3, 32 – 35.
32. E. J. Mistrik, A. Mateides, *Chem. Tech. (Leipzig)* **35** (1983) no. 2, 90 – 94.
33. F. Mares, S. E. Diamond, F. J. Regina, J. P. Solar, *J. Am. Chem. Soc.* **107** (1985) no. 12 , 3545 – 52.
34. G. Lu, G. Huang, Y. Guo, R. Wang, Q. Bai, C. Li, H. Qi, *Shiyou Huagong* **14** (1985) no. 5, 262 – 266. *Chem. Abstr.* **104** (1986) 170513 c.
35. B. D. Dombek, *Adv. Catal.* **32** (1983) 325 – 416.
36. S. Nowak, I. Eichhorn, I. Ohme, B. Lücke, *Chem. Tech. (Leipzig)* **36** (1984) no. 2, 55 – 57.
37. J. Kollar, *CHEMTECH* **8** (1984) 504 – 511.
38. J. B. Saundby, *I. Chem. Symp.* (1986) 133 – 166.
39. S. Nowak, *Chem. Tech. (Leipzig)* **36** (1984) no. 4, 144 – 146.
40. Chevron, US 3 911 003, 1975 (S. Susuki).
41. A. Aquilo, J. S. Alder, D. N. Freeman, R. J. H. Voorhoeue, *Hydrocarbon Process.* **62** (1983) no. 3, 57 – 65.
42. J. Costa, J. Soley, J. Mata, J. Masides, *Effluent Water Treat. J.* **25** (1985) no. 12, 429 – 434. P. Schöberl, *Tenside Deterg.* **23** (1986) no. 5, 255 – 266.
43. R. A. Conway, G. T. Waggy, M. H. Spiegel, R. L. Berglund, *Environ. Sci. Technol.* **17** (1983) no. 2, 107 – 110.
44. G. A. Kolry, A. A. Abdelghani, A. C. Anderson, A. Monkiene, *Trace Subt. Environ. Health* **23** (1990)371 – 378.
45. Merkblatt Glykole, Hoechst AG, Frankfurt, 1972.
46. E. Obermeier, S. Fischer, D. Bohne, *Ber. Bunsenges. Phys. Chem.* **89** (1985) 805 – 809.
47. Merkblatt Glykole, Hoechst AG, Frankfurt, 1988.
48. Hoechst AG, DE 2 640 506, 1978 (G. Schefel, R. Obermeier).
49. Merkblatt Polyglykole, Hoechst AG, Frankfurt, 1986.
50. Merkblatt Ethylenglykoldimethylether, Hoechst AG, Frankfurt, 1986.
51. Merkblatt Genantin, Hoechst AG, Frankfurt, 1982.
52. Merkblatt Glykole, Hoechst AG, Frankfurt, 1972.
53. Merkblätter Hoechst AG, Frankfurt, Lösemittel, 1984; Methylglykol, 1978;Methyldiglykol, 1970; Butyldiglykol, 1970;Ethylenglykoldimethylether, 1986.
54. Merkblatt Hoechst 1704 LTV (Flugzeugenteisung) 1985; Merkblatt Hoechst 1678 (Landebahnenenteisung) Hoechst AG, Frankfurt, 1985.
55. Merkblatt Diethylenglykol, Hoechst AG, Frankfurt, 1986.
56. Merkblatt Triethylenglykol, Hoechst AG, Frankfurt, 1984.
57. Merkblatt Polyglykole Hoechst, Hoechst AG, Frankfurt, 1986.
58. *Chem. Week,* **134** (1984) March 7, 32 – 35.
59. E. P. Laug, H. O. Calvery, H. J. Morris, G. Woodard, *J. Ind. Hyg. Toxicol,* **21** (1939) 173.
60. G. Bornmann, *Arzneim. Forsch.* **4** (1954) 643.
61. H. F. Smyth, Jr., J. Seaton, L. F. Fischer, *J. Ind. Hyg. Toxicol.* **23** (1941) 259.
62. H. J. Morris, A. A. Nelson, H. O. Calvery, *J. Pharmacol. Exp. Therap.* **74** (1942) 266.
63. F. R. Blood, *Food Cosmet. Toxicol.* **3** (1965) 229.
64. F. R. Blood, G. A. Elliott, M. S. Wright, *Toxicol. Appl. Pharmacol.* **4** (1962) 489.
65. J. A. Roberts, H. R. Seibold, *Toxicol. Appl. Pharmacol.* **15** (1969) 624.
66. C. P. Carpenter, H. F. Smyth, Jr., *Am. J. Ophthalmol.* **29** (1946) 1363.
67. T. O. McDonald, M. D. Roberts, A. R. Borgmann, *Toxicol. Appl. Pharmacol.* **21** (1972) 143.
68. W. M. Grant: *Toxicology of the Eye,* 2nd ed., Charles C. Thomas, Springfield, Ill., 1974.
69. E. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, Amsterdam 1965.
70. R. A. Coon, R. A. Jones, L. J. Jenkins, Jr., J. Siegel, *Toxicol. Appl. Pharmacol.* **16** (1970) 646.

71. J. H. Wills, F. Coulston, E. S. Harris, E. W. McChesney, J. C. Russell, D. M. Serrone, *Clin. Toxicol.* **7** (1974) 463.
72. E. Harris: *Proceedings of the 5th Annual Conference on Atmospheric Contamination in Confined Spaces,* AMRL TR-69-130, paper no. 8, Aerospace Med. Res. Lab., Wright-Patterson Air Force Base, Ohio 1969, p. 9.
73. M. Felts: *Proceedings of the 5th Annual Conference on Atmospheric Contamination in Confined Spaces,* AMRL TR-69-130, paper no. 9, Aerospace Med. Res. Lab., Wright-Patterson Air Force Base, Ohio 1969, p. 105.
74. J. McCann, E. Choi, E. Yamasaka, B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.* **72** (1975) 5135.
75. E. H. Pfeiffer, H. Dunkelberg, *Food Cosmet. Toxicol.* **18** (1980) 115.
76. M. M. Mason, C. C. Cate, J. Baker, *Clin. Toxicol.* **4** (1971) 185.
77. P. H. Derse, *U. S. C. F. S. T. I. PBRep.* **195 153** (1969).
78. BRRC Project Report 44 – 109, 08. 03. 82.
79. BRRC Project Report 46 – 89, 11. 04. 84.
80. L. R. DePass, M. D. Woodside, R. R. Maronpot, C. S. Weil, *Fundam. Appl. Toxicol.* **7** (1986) 566.
81. BRRC Project Report 43 – 54, 25. 07. 80.
82. B. Ballantyne, A summary of studies conducted on the long-term toxicology of ethylene glycol, UCC, 4. 22. 1985.
83. V. K. Rowe, M. A. Wolf in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2C,** Wiley-Interscience, New York 1963.
84. C. P. Carpenter, U. C. Pozzan, C. S. Weil, J. H. Nair, III, G. A. Keck, H. F. Smyth, Jr., *AMA Arch. Ind. Health* **14** (1956) 114.
85. H. O. Calvery, T. G. Klumpp, *Southern Med. J.* **32** (1939) 1105.
86. O. G. Fitzhugh, A. A. Nelson, *J. Ind. Hyg. Toxicol.* **28** (1946) 40.
87. C. S. Weil, C. P. Carpenter, H. F. Smyth, Jr., *Arch. Environ. Health* **11** (1965) 569.
88. C. S. Weil, C. P. Carpenter, H. F. Smyth, Jr., *Ind. Med. Surg.* **36** (1967) 55.
89. G. Bornmann, *Arzneim. Forsch.* **5** (1955) 38.
90. A. Loesser, G. Bornmann, L. Grosskinsky, G. Hess, R. Kopf, K. Ritter, A. Schmitz, E. Stürner, H. Wegener, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **221** (1954) 14.
91. V. P. Plugin, *Gig. Sanit.* **33** (1968) 16.
92. C. P. Carpenter, H. F. Smyth, Jr., *Am. J. Ophthalmol.* **29** (1946) 1363.
93. P. J. Hanzlik, W. S. Lawrence, J. K. Fellows, F. P. Luduena, G. L. Lacqueur, *J. Ind. Hyg. Toxicol.* **29** (1947) 325.
94. H. Wegener, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **220** (1953) 414.
95. E. H. Pfeiffer, H. Dunkelberg, *Food Cosmet. Toxicol.* **18** (1980) 115.
96. W. M. Lauter, V. L. Vrla, *J. Am. Pharm. Assoc.* **29** (1940) 5.
97. E. G. Stenger, L. Aeppli, E. Peheim, F. Roulet, *Arzneim. Forsch.* **18** (1968) 1536.
98. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, *J. Am. Pharm. Assoc. Sci. Ed.* **39** (1950) 349.
99. H. F. Smyth, Jr., C. P. Carpenter, C. S. Weil, *J. Am. Pharm. Assoc. Sci. Ed.* **44** (1955) 27.
100. The Dow Chemical Company, *Dow Polyethylene Glycols* (1959).
101. C. P. Carpenter, M. D. Woodside, E. R. Kinkead, J. M. King, L. J. Sullivan, *Toxicol. Appl. Pharmacol.* **18** (1971) 35.
102. C. B. Shaffer, F. H. Critchfield, *J. Am. Pharm. As-soc. Sci. Ed.* **36** (1947) 152.
103. C. B. Shaffer, F. H. Critchfield, J. H. Nair, *J. Am. Pharm. Assoc. Sci. Ed.* **39** (1950) 340.
104. A. H. Principe, *J. Forensic Sci.* **13** (1968) 90.

Ethylene – Propene Rubber → Rubber, 3. Synthetic

Ethylene Oxide

Epoxides is a separate keyword.

SIEGFRIED REBSDAT, Winhöring, Federal Republic of Germany (Chaps. 1 – 10)

DIETER MAYER, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 11)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 4
4. Production 7
4.1. Catalysts 8
4.2. Mechanism of Catalysis 9
4.3. Technology 11
5. Environmental Protection and Ecology 15
6. Quality Specifications 16
7. Analysis 16
8. Handling, Storage, and Transportation 16
9. Uses 20
10. Economic Aspects 20
11. Toxicology and Occupational Health . 20
12. References 23

With kind support from Clariant GmbH, Werk Gendorf

1. Introduction

Ethylene oxide (oxirane) [*75-21-8*], M_r 44.05, is the simplest cyclic ether. It is a colorless gas or liquid and has a sweet, etheric odor. The ethylene oxide molecule with its short C–C bond and strained angles is shown in Figure 1 [6]. Theoretical calculations of the structure are in good agreement with experimental results [7].

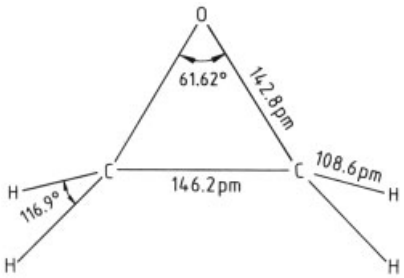

Figure 1. The ethylene oxide molecule

Ethylene oxide is very reactive because its highly strained ring can be opened easily, and is thus one of the most versatile chemical intermediates. Because of its reactivity and toxicity it is also a hazardous compound that has been involved in a number of serious incidents. A good understanding of its properties is a necessary prerequisite for its safe handling. Ethylene oxide was first described in 1859 by WURTZ [8], who prepared it by eliminating hydrochloric acid from ethylenechlorohydrin, using potassium hydroxide solution. Industrial production by the chlorohydrin process began in 1914 and was based on WURTZ's discovery. Since then the production and importance of ethylene oxide have steadily grown.

In 1931, LEFORT [9] discovered the direct catalytic oxidation of ethylene [*74-85-1*], which gradually superseded the chlorohydrin process. Currently, ethylene oxide is produced by direct oxidation of ethylene with air or oxygen; annual worldwide production capacity is ca. 15×10^6 t, making it an important industrial chemical. Ethylene oxide itself is used as a disinfectant, sterilizing agent, and fumigant. Its most important derivative is ethylene glycol [*107-21-1*], which is used in antifreeze (car radiators) (→ Antifreezes) and for the manufacture of polyester fibers (→ Fibers, 4. Synthetic Organic; → Polyesters). Other ethylene oxide derivatives, amines and poly(ethylene glycols), are used in surfactants, solvents, etc.

There is some endogenous formation from ethylene in humans [10], and it has been discovered in an interstellar source [11].

As atmospheric pollutant it is present in natural gas, cigarette smoke, and diesel exhaust.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_117

Though ethylene oxide in nature is rare its biological action may have a fundamental impact on DNA and RNA evolution of whole biological systems [12].

2. Physical Properties

Some physical properties of ethylene oxide are summarized in Table 1.

Table 1. Physical properties of ethylene oxide [5], [13], [14]

mp at 101.3 kPa	−112.5 °C
bp at 101.3 kPa	10.8 °C
Critical temperature	195.8 °C
Critical pressure	7.2 MPa
Critical density	314 kg/m^3
Refractive index, n_D^7	1.3597
Explosive limits in air at 101.3 kPa	
lower	2.6 vol %
upper	100.0 vol %
Electrical conductivity	4×10^{-6} S/m
Dielectric constant	
at −1 °C (liquid)	13.9
at 15 °C (vapor)	1.01
Heat of combustion at 25 °C, 101.3 kPa	29.648 kJ/kg
Entropy of the vapor at 101.3 kPa	
10.5 °C	5.439 kJ kg^{-1} K^{-1}
25.0 °C	5.495 kJ kg^{-1} K^{-1}
Ignition temperature in air at 101.3 kPa	429 °C
Ignition energy in air at 101.3 kPa and 25 °C	0.087 mJ
Minimum ignition energy of the gas at 220 kPa and 100 °C	64 mJ
Decomposition temperature of the vapor at 101.3 kPa	571 °C
Heat of polymerization	2091 kJ/kg
Heat of fusion	117.86 kJ/kg
Heat of decomposition of the vapor	1901 kJ/kg
Coefficient of cubic expansion	
at 22 °C	0.00161
at 55 °C	0.00170
Heat of solution in water at 25 °C	142.57 kJ/kg

In the pressure range 0 – 101.3 kPa and at 40 and 60 °C, the solubility of ethylene oxide in water obeys Henry's law. The Henry constants for these temperatures are 2.875 and 1.448, respectively. However, at 20 °C a more than proportional increase in solubility occurs with the partial pressure [15].

Liquid ethylene oxide and water are completely miscible in all proportions. Table 2 shows some of the properties of aqueous ethylene oxide solutions. Of particular note are the relatively high melting points, which are due to clathrate formation [17]. Clathrates consist of organic molecules that are enclosed in a cage structure. In this case, the cage is an ice lattice which is composed of unit cells, each containing 46 water molecules, and two types of cavities: six larger ones (14-sided, tetradecahedra) and two smaller ones (12-sided, dodecahedra). If the clathrates are made to crystallize from increasingly concentrated aqueous solutions of ethylene oxide, their density and ethylene oxide content also increase. Ethylene oxide first fills the six tetradecahedral cavities of the water lattice, followed by 20 – 40 % of the dodecahedral cavities [18].

The temperatures for the formation of ice and hydrate from mixtures of ethylene oxide and water for concentrations of 0 to 97.89 wt % ethylene oxide gave a highest hydrate formation temperature of 11.08 °C at an ethylene oxide concentration of 26.15 wt %. The hydrate corresponds to a composition of 6.91 molecules of water per molecule ethylene oxide. The unit cell consists of 46 water molecules and 6.66 ethylene oxide molecules. The eutectic point of the mixture has been determined to lie at 2.107 °C and 4.75 wt % ethylene oxide (Fig. 2) [19].

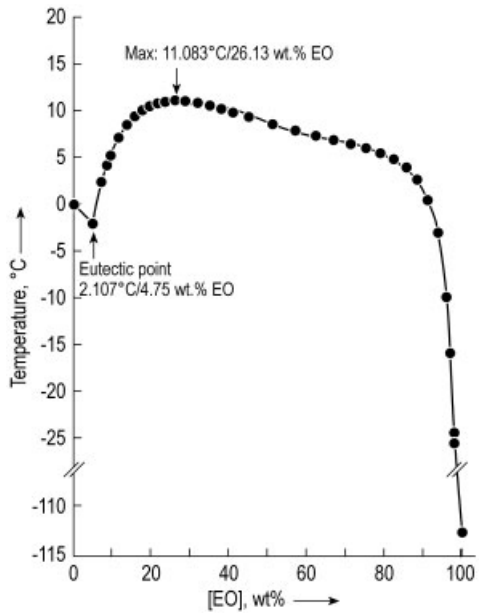

Figure 2. Hydrate-formation temperature as a function of ethylene oxide concentration

There are also mixed clathrates, in which the small cages are occupied by ethylene oxide and the big cages by molecules with an higher space requirement, e.g., tetrahydrofuran [20].

Table 2. Physical properties of aqueous ethylene oxide solutions [3], [5], [16]

Ethylene oxide content, wt %	*mp*, °C	*bp*, °C	Density at 10 °C, g/L	Flash point, °C
0	0.0	100	0.9991	
0.5				41.5
1	−0.4			31
2				3
3	−1.3			
5	−1.6	58	0.9988	−2
10	5.6	42.5	0.9980	
20	10.4	32	0.9945	−21
30	11.08 (max)	27	0.9882	−28
40	10.4	21	0.9792	−35
60	7.8	16	0.9534	−45
80	3.7	13	0.9194	−53
100	−112.5	10.4	0.8826	−57

The solubilities of gases in ethylene oxide vary, increasing in the order nitrogen, argon, methane, ethane. Earlier data [21] have been revised [22]. The Henry constants for these gases in ethylene oxide at different temperatures are given in Table 3.

Table 4 shows other temperature-dependent physical properties of gaseous and liquid ethylene oxide.

3. Chemical Properties

Ethylene oxide is a very reactive, versatile compound. Its reactions proceed mainly via ring opening and are highly exothermic. Explosive decomposition of the ethylene oxide vapor may occur at higher temperatures if heat dissipation is inadequate. Only the most important types of the large number of possible reactions are briefly discussed here. More detailed information can be found in [25–29].

Decomposition. Gaseous ethylene oxide starts to decompose at ca. 400 °C to form mainly CO, CH_4, as well as C_2H_6, C_2H_4, H_2, C, and CH_3CHO. The first step in the decomposition is presumed to be the isomerization of ethylene oxide to acetaldehyde [30]. Once the decomposition reaction has been initiated (ignition source), it can be propagated through the gas phase and, under certain conditions, may be explosive (see Chap. 8).

Addition to Compounds with a Labile Hydrogen Atom. Ethylene oxide reacts with compounds containing a labile hydrogen atom to form a product containing a hydroxyethyl group:

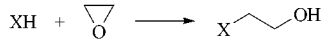

Examples of XH are: HOH, H_2NH, HRNH, R_2NH, RCOOH, $RCONH_2$, HSH, RSH, ROH, N≡CH, and B_2H_6 (R = alkyl, aryl). The reaction is accelerated by acids and bases. During acid catalysis, the ethylene oxide is first protonated to form an epoxonium ion (**1**), which is in equilibrium with the corresponding hydroxycarbenium ion (**2**). The anion X^- can then react with **1** or **2** in an S_N2 or S_N1 reaction, respectively. In an alkaline medium, the S_N2 mechanism is favored. A detailed discussion of reaction mechanisms can be found in [27], [31].

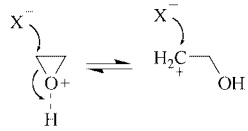

1 **2**

The protonation of ethylene oxide by various H-active compounds in tetrachloromethane can be detected by IR spectroscopy. Broadening and shift of the IR bands in the region of 530 to 740 cm^{-1} depend on the type of H-active compound (OH, NH, CH) [32]. Analogous protonated intermediates have been postulated in gas-phase S_N2 reactions of ethylene oxide on the basis of theoretical and experimental results [33].

Table 3. Solubility of gases in ethylene oxide (Henry constants in MPa) [22], [23]

Temperature, °C	Solubility in ethylene oxide			
	Nitrogen	Argon	Methane	Ethane
0	284	169	62.1	8.5
25	221	144	62.2	11.0
50	184	129	62.3	13.0

Table 4. Physical properties of ethylene oxide at various temperatures [3], [24]

Temperature	Density		Heat of Vaporization	Vapor pressure	Specific heat		Thermal conductivity		Surface tension	Dynamic viscosity	
	Vapor	Liquid			Vapor	Liquid	Vapor	Liquid		Vapor	Liquid
°C	g/L	kg/m^3	kJ/kg	kPa	kJ kg^{-1} K^{-1}		10^{-4} J cm^{-1} s^{-1} K^{-1}		mN/m	mPa · s	
−40	0.2	946	647	7.6		1.88		17.7	33.8		0.50
−30	0.3	931	634	16.2				17.2	32.3		0.44
−20	0.5	919	621	24.6		1.91		16.8	30.7		0.39
−10	0.9	909	607	40.3				16.4	29.1		0.34
0	1.3	899	592	65.4	1.00	1.95	1.03	16.0	27.6	0.0090	0.31
+10	1.8	888	578	99.2	1.06		1.10	15.6	26.7		0.28
+20	2.9	876	563	145.6	1.10	2.00	1.18	15.3	24.5		0.25
+30	3.9	862	547	208.0	1.15		1.25	15.0			0.23
+40	5.25	847	531	287.1	1.19	2.06	1.34	14.7	22.0		0.21
+50	6.85	831	515	391.0	1.23		1.43	14.3		0.0104	0.19
+60		814	498	534.6	1.28	2.15	1.52	14.0	17.5		0.18
+70		796	481	694.6	1.32		1.62	13.7			0.16
+80		779	463	912.8	1.36	2.27	1.72	13.4	14.5		0.15
+100		745	424	1464.0	1.45	2.40	1.93	12.5	12.0	0.0118	0.14

All common acids and Lewis acids as well as zeolites, ion exchangers [34], and aluminum oxide are effective catalysts. Solid polymeric acids (cation exchangers) are especially recommended for the reaction of ethylene oxide with methanol [35].

As the end product of the above reaction contains at least one hydroxyl group, it may react successively with further ethylene oxide molecules:

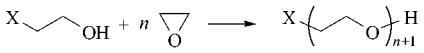

The molecular mass of the resulting polymers depends on the ratio of the reactants, the catalyst used, and the reaction conditions [1], [28].

A byproduct of the acid-catalyzed polymerization is dioxane, which may even be the sole product with a large excess of ethylene oxide. An explanation for this behavior is offered by the assumption that two different reaction mechanisms are effective. The first is the active chain end (ACE) mechanism, which leads to dioxane and poly(ethylene glycol).

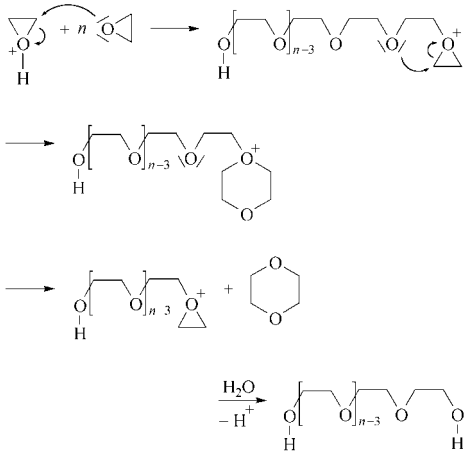

And the second is the activated monomer (AM) mechanism, which leads exclusively to poly(ethylene glycol).

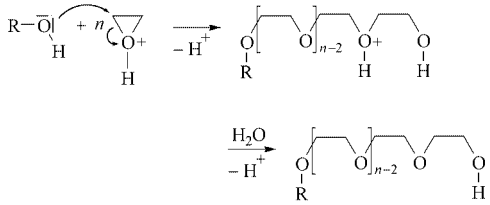

One of the conditions which favor the ACM mechanism and therefore the formation of dioxane is an increasing excess of ethylene oxide [36].

A large variety of reactions occur between ethylene oxide and compounds containing labile hydrogen atoms; therefore, such compounds are often used to produce derivatives.

Commercially, the most important of this type of reaction is the hydrolysis of ethylene oxide to ethylene glycol. About 60 % of the total ethylene oxide production is converted into ethylene glycol in this way (→ Ethylene Glycol). However, this reaction, which is often referred to as ethoxylation, is also used to produce the bulk of all the other commercially important ethylene oxide derivatives. The ethoxylation products of alkyl phenols, ammonia, fatty alcohols, fatty amines, and fatty acids have a variety of uses [28]. Although polymer formation according to the above equation is more likely to be an unwanted side reaction, the production of poly(ethylene glycols) of widely varying molecular mass by this route is of some importance [37]. Undesired polymerization may be catalyzed by rust in rusty containers [38].

Reaction with Water. The formation of hydrates is discussed in Chapter 2. Hydrolysis is the most important reaction of ethylene oxide: about 60% of the total ethylene oxide production is transformed into ethylene glycol. The reaction of ethylene oxide with water in the absence of catalysts at 20 °C is relatively slow, as shown by the half-life of 20 d. The dependence of the half-life of a 3 % solution of ethylene oxide in water on pH is shown in Figure 3. The dependence is characterized by a bell-shaped curve with a relatively broad plateau from pH 5 to 11 and a steep slope on both sides. A strong acceleration occurs only below pH 4 and above pH 12. Byproducts of the hydrolysis are diethylene glycol, triethylene glycol, and higher molecular mass poly(ethylene glycols). The percentage of ethylene oxide converted to these products depends almost solely on the ratio of water to ethylene oxide in the starting material and only marginally on pH and temperature.

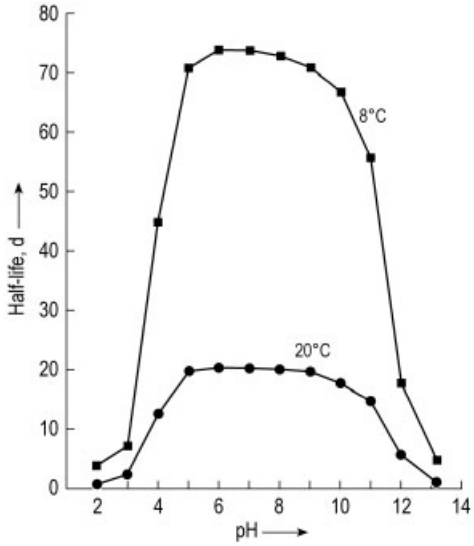

Figure 3. Half-life of 3 wt % ethylene oxide in water as a function of pH at 8 and 20 °C

Reaction with DNA, RNA and Proteins. The major product of reaction with DNA is 7-(2-hydroxyethyl) guanine. Further reaction products are (2-hydroxyethyl) guanine and 3-(2-hydroxyethyl) adenine [12].

In men exposed to ethylene oxide, adducts with hemoglobin can be found and its level may be evaluated by the concentration of (2-hydroxyethyl) valine [39].

Addition to Double Bonds. Ethylene oxide can add to compounds with double bonds, e.g., carbon dioxide [40], to form cyclic products:

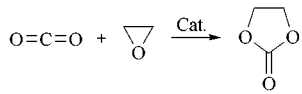

Ethylene carbonate
(1,3-dioxolan-2-one)

Quaternary ammonium compounds are suitable catalysts for this reaction, which then proceeds at 200 °C and 8 MPa [41]. Use of porphine complexes of aluminum in solvents such as chloroform allows the reaction to take place at nor-

mal pressure and at room temperature [42]. Hydrolysis of ethylene carbonate (1,3-dioxolan-2-one) [*96-49-1*] yields pure ethylene glycol [43]. Therefore, the reaction is of interest for the selective production of ethylene glycol from ethylene oxide, i.e., for avoiding the partly unwanted formation of poly(ethylene glycols) [44].

Ethylene oxide also adds to other double bond systems, e.g., to $R_2C{=}O$ [45], SC=S, $O_2S{=}O$, RN=CO, and OS=O.

Catalytic Isomerization to Acetaldehyde. Aluminum oxide (Al_2O_3) [46], phosphoric acid and phosphates [47], and under certain conditions silver [48] catalyze the isomerization of ethylene oxide to acetaldehyde.

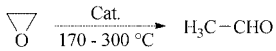

Reduction to Ethanol. Reduction of ethylene oxide to ethanol is catalyzed by Ni, Cu, and Cr on Al_2O_3 [49].

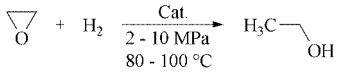

Reaction with Grignard Reagents. Grignard reagents react with ethylene oxide to produce compounds with a primary hydroxyl group [29], [50].

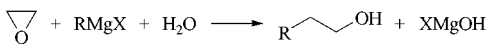

Reaction with Lithium Acetylide. Acetylene is treated with lithium in liquid ammonia to form the lithium salt, which reacts with ethylene oxide at −33 °C to give 3-butyne-1-ol in 80 % yield [51].

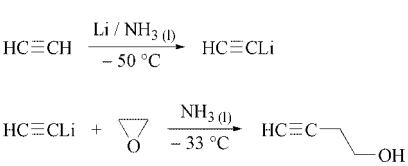

Reaction with Synthesis Gas. Ethylene oxide is carbonylated with CO/H_2 in the presence of rhodium catalysts in tetraethylene glycol dimethyl ether to give 1,3-propanediol in good yields [52].

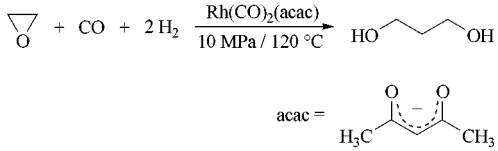

Oligomerization to Crown Ethers. Ethylene oxide oligomerizes to form cyclic polyethers (crown ethers) in the presence of a fluorinated Lewis acid catalyst (→ Crown Ethers).

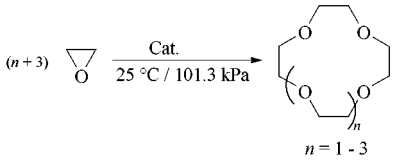

Synthesis is successful when groups that saturate potential terminal hydroxyl groups are absent [53]. Suitable catalysts are the BF_4^-, PF_6^-, or SbF_6^- salts of metal cations. The reaction can be directed to produce a particular oligomer by choosing an appropriate metal cation. The size of the cation determines the size of the crown ether ring by acting as a template [54]. For example, with $CsBF_4$, only the cyclic hexamer ($n = 3$), 1,4,7,10,13,16-hexaoxacyclooctadecane (18-crown-6) [*17455-13-9*], is produced; if $Cu(BF_4)_2$ is used, 90 % of the product consists of the pentamer ($n = 2$), 1,4,7,10,13-pentaoxacyclopentadecane (15-crown-6) [*33100-27-5*], and with $Ca(BF_4)_2$, a mixture of the tetramer (50 %, $n = 1$), 1,4,7,10-tetraoxacyclodecane (12-crown-4) [*294-93-9*], and the pentamer, 1,4,7,10,13-pentaoxacyclopentadecane (18-crown-6) [*17455-13-9*], is produced.

Reaction with Dimethyl Ether. Ethylene oxide reacts with dimethyl ether to produce poly(ethylene glycol) dimethyl ethers (→ Dimethyl Ether); this reaction is used for the industrial production of the lower molecular mass homologues, which are widely used as solvents [55].

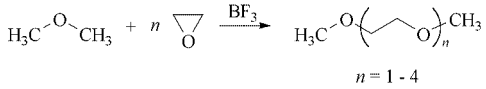

Reaction with Bromotrimethylsilane. Ethylene oxide adds to bromotrimethylsilane [*2857-97-8*] in a highly exothermic reaction with excellent yields [56].

Disproportionation in the presence of iron oxides (especially γ-Fe_2O_3 and γ-FeOOH) has been described as the cause of an explosion (see Chap. 8) [57].

$$4\,C_2H_4O \longrightarrow 3\,C_2H_4 + 2\,CO_2 + 2\,H_2$$

$$5\,C_2H_4O \longrightarrow 4\,C_2H_4 + 2\,CO_2 + H_2 + H_2O$$

$$6\,C_2H_4O \longrightarrow 5\,C_2H_4 + 2\,CO_2 + 2\,H_2O$$

4. Production

As mentioned in the introduction, ethylene oxide was produced formerly by the chlorohydrin process (→ Chlorohydrins, Chap. 4.). However, this method is no longer used on an industrial scale; a description is given in [58]. Although the selectivity of this process (80 %) was satisfactory, practically all of the chlorine that was used was lost as calcium chloride and unwanted chlorine-containing byproducts were generated. This not only was inefficient, but also caused pollution problems so that this method has now been replaced by the direct oxidation process.

WURTZ, who had discovered ethylene oxide, attempted as early as 1863 to produce it by direct oxidation of ethylene with oxygen, but did not succeed [59]. Many other unsuccessful attempts were made [60] before LEFORT made the crucial discovery in 1931 that the formation of ethylene oxide from ethylene and oxygen was catalyzed by metallic silver [9]. Since the catalyst plays a central role in the production process, it is discussed first.

4.1. Catalysts

To date no other metal has been found that can compete with silver in the catalysis of the direct oxidation of ethylene to ethylene oxide. However, the silver catalysts have been substantially improved since their discovery by LEFORT [9].

Figure 4. Silver catalyst for ethylene oxide production (SEM, × 10 000)

Only supported catalysts are used, the silver being deposited on a porous *support material* in concentrations of 7 – 20 %. The support material is of critical importance [61]; currently, preference is given to ultrapure (over 99 %) aluminum oxide [*57828-03-2*] fired at a high temperature that has a defined pore structure (pore diameter 0.5 – 50 µm) and low specific surface area (<2 m^2/g). Although supports with higher specific surface areas are very active, their selectivity is low, presumably because ethylene oxide can diffuse only slowly out of the smaller pores and, therefore, can be further oxidized [62]. Supports with chemical activity towards ethylene or ethylene oxide are unsuitable. All support containing hydroxyl groups catalyze the isomerization of ethylene oxide to acetaldehyde. As total oxidation of acetaldehyde is very fast on silver, catalysts consisting of silver and carriers like γ-alumina, SiO_2, MgO, SiC, TiO_2, Y_2O_3, and ZrO_2 give only poor selectivities or form only CO_2 and H_2O [63]. Silane treatment of supports

containing hydroxyl groups improves the performance in ethylene oxide formation, and this confirms the detrimental effect of hydroxyl groups, which result in isomerization of ethylene oxide to acetaldehyde [64].

The impregnation methods used to deposit the silver on the support are being constantly improved. Complexes of silver salts with amino compounds are used which decompose to give evenly and finely distributed silver particles with a diameter of 0.1 – 1 µm [65]. Figure 4 shows a scanning electron micrograph (SEM) of the inner surface of a silver catalyst particle.

In addition, 100 – 500 mg/kg of *promoters* such as salts or other compounds of alkali and alkaline earth metals are added to the catalyst, significantly improving the selectivity. Among the alkali metal salts, those of cesium are especially effective [66]. Promoters are often combined. An especially effective combination consists of rhenium, sulfur, tungsten, molybdenum and results in outstanding selectivities of up to 90 %, albeit at higher temperature and reduced catalyst life [67].

Addition of chlorine compounds to the reaction gases improves selectivity [68]. These *inhibitors* (e.g., 1,2-dichloroethane, vinyl chloride [69], ethyl chloride) [70] suppress the combustion of ethylene to carbon dioxide and water; they ensure that the silver surface is covered evenly with a supply of chlorine [71].

The advances made in silver-based catalysts since LEFORT's original discovery (i.e., optimized support materials, silver distribution, and use of promoters and inhibitors) have improved selectivity from 50 to about 90 %. Silver has maintained its position as the only known metal that can catalyze the oxidation of ethylene to ethylene oxide with a commercially viable selectivity.

The silver catalysts used in ethylene oxide production are not suited for the production of epoxides from olefins with allylic hydrogen atoms (e.g., propene); total oxidation is the prevailing reaction. Nonallylic olefins such as butadiene are reported to form epoxides in reasonable yields [72].

Aging of the Catalyst. Modern silver-based catalysts have an initial selectivity of 80 – 90 %; a maximum selectivity of 90 % is commercially feasible [44]. The disadvantages associated with highly selective catalysts are that they age relatively quickly and less heat is produced. As the catalyst is used, its selectivity and activity gradually deteriorate [73–75] due to

1) Abrasion, dust formation, and blocking of pores.
2) Accumulation of detrimental impurities introduced with the reaction gases (e.g., sulfur from ethylene or methane).
3) Changes in the silver particles, which enlarge, form agglomerates, and become unevenly distributed. Thus, the installed silver surface, which is about 20 km^2 for 100 t/a ethylene oxide capacity, drops to about 50 % of its initial value by sintering during two years on stream. To compensate this reduction in silver surface the reaction temperature (reaction velocity) has to be increased accordingly to maintain the production rate.

Generally applicable methods of regenerating such catalysts are unknown; the only method of any importance is the use of methanolic solutions of cesium salts [76]. If regeneration is impossible, it becomes economically necessary to replace the catalyst when selectivity has fallen to a critical level or becomes technically necessary because the reaction temperature has reached the maximum design value. The lifetime of a modern catalyst is two to five years, depending on the type of catalyst (high-selectivity catalysts have considerably shorter life times), the rate of ethylene oxide production, and the purity of the reaction gases (sulfur is extremely poisonous). From the spent catalysts silver is recovered in high purity and with only small losses, while the used supports are waste.

4.2. Mechanism of Catalysis

Two reactions take place simultaneously at the silver surface. In addition to ethylene oxide formation (partial oxidation, Eq. 1), complete combustion (total oxidation, Eq. 2) to CO_2 and water also takes place. Small amounts of acetaldehyde and formaldehyde are also formed [77]. The reactions given by Equations (1) and (2) are exothermic, their enthalpies being −106.7 and −1323 kJ/mol, respectively at 250 °C and 1.5 MPa [44].

$$H_2C{=}CH_2 + \tfrac{1}{2}\,O_2 \longrightarrow \triangledown_{O}$$

$$H_2C{=}CH_2 + 3\,O_2 \longrightarrow 2\,CO_2 + 2\,H_2O$$

Although many attempts have been made to discover the mechanism responsible for silver's unique action, opinions remain divided [78]. Evaluation of the various studies is difficult because often they were carried out under different conditions that were not always suitable for industrial purposes. It is agreed that silver can adsorb oxygen in a number of ways and that this phenomenon is the basis of its unequalled efficiency in catalyzing the oxidation of ethylene to ethylene oxide. The following adsorption forms of oxygen are of critical importance:

1) Atomic oxygen
2) Molecular oxygen
3) Subsurface oxygen, i.e., dissolved oxygen found below the surface

Opinions differ as to the role of these different types of adsorbed oxygen; two conflicting mechanisms have been proposed.

Mechanism 1. Only molecular oxygen reacts with ethylene to form ethylene oxide, whereas atomic oxygen only reacts to form carbon dioxide and water [79]. Chlorine (from the inhibitor) blocks the adsorption of atomic oxygen on the silver surface so that, in the ideal case, the optimally inhibited silver surface only adsorbs molecular oxygen (Eq. 3). The adsorbed molecular oxygen then reacts with ethylene to form ethylene oxide, leaving behind one oxygen atom after desorption (Eq. 4). This atomic oxygen then causes the combustion of ethylene to CO_2 and H_2O (Eq. 5).

$$6\,Ag + 6\,O_2 \longrightarrow 6\,Ag \cdot O_{2\,ads}$$

$$6\,Ag \cdot O_{2\,ads} + 6\,H_2C{=}CH_2 \longrightarrow 6\,\triangledown_{O} + 6\,Ag \cdot O_{ads}$$

$$6\,Ag \cdot O_{ads} + H_2C{=}CH_2 \longrightarrow 6\,Ag + 2\,CO_2 + 2\,H_2O$$

Since six oxygen atoms are needed for the complete oxidation of one ethylene molecule, six ethylene oxide molecules must be formed before one ethylene molecule can be completely oxidized. Therefore, if inhibition of atomic oxygen adsorption is optimal, the maximum selectivity is $6/7 \times 100 = 85.7\,\%$. This mechanism plausibly explains the way in which the inhibitor acts. For a long time the molecular oxygen mechanism was strongly supported by the fact that efforts to improve the catalysts and the process seemed to lead to selectivities approaching the limit of 85.7 % predicted by the mechanism, but not exceeding it.

A variant of this mechanism is based on the assumption that molecular oxygen can cause both ethylene oxide formation and total oxidation via a common intermediate [80], [81]:

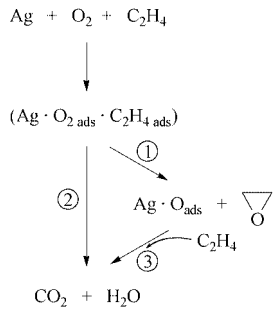

The effect of the inhibitor is interpreted differently here. Reaction 2 requires a greater space than reaction 1. The inhibitor favors reaction 1 by reducing the available space on the surface of the catalyst. On an optimally inhibited surface, reaction 2 is totally suppressed, so that total oxidation takes place only via reaction 3. This results again in a maximum selectivity of 85.7 %.

Mechanism 2. In this mechanism, atomic oxygen, possibly together with the subsurface oxygen, is thought to be responsible for both total and partial oxidation. The contribution of the molecular oxygen, if any, is indirect [82], [83]. The environment of the adsorbed oxygen atom determines whether the reaction with ethylene

leads to ethylene oxide or CO_2 and H_2O [84], [85]. Chlorine (inhibitor), promoters, and subsurface oxygen in the proximity of the reacting oxygen atom are thought to influence the latter to favor partial oxidation, e.g., by reducing the negative charge of the adsorbed oxygen [86]. Unlike mechanism 1, the theoretical upper limit of selectivity of 6/7 does not apply for mechanism 2.

Since with modern catalysts (containing Cs, W, Re, and S in promoting quantities on the silver) maximum selectivities of 90 % are feasible even on an industrial scale, mechanism 1 can at least not exclusively be valid with these catalysts. Also, it has been shown that atomic oxygen, formed on the silver surface by decomposition of N_2O, can produce ethylene oxide, though with low selectivities [87], [88]. For the epoxidation of nobornene and styrene it has been shown under ultrahigh-vacuum conditions that the atomic oxygen species on the silver is responsible for the epoxidation [89], [90]. Furthermore, theoretical calculations seem to favor mechanism 2 and can explain the action of inhibitor and promotor [91], [92].

Intermediates of the total oxidation are thought to include acetaldehyde, acetic acid, formic acid, and oxalic acid [84]. Desorption of intermediates from the silver surface produces acetaldehyde when a helium stream is used; ethanol and acetic acid are produced when a hydrogen stream is used [93].

Formate, acetate, and oxalate were detected on a silver surface by ^{13}C NMR spectroscopy after contacting the silver first with oxygen and subsequently with ethylene [94].

Kinetics. Three reactions are usually considered in kinetic models:

1) Partial oxidation
2) Total oxidation
3) Consecutive oxidation

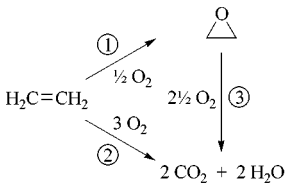

The consecutive combustion of ethylene oxide (3) is negligible only if temperatures are not too high and if the reaction takes place under chemical control. To avoid diffusion control low surface area catalyst supports with high pore diameters are required. The surface reaction leading to ethylene oxide is considered to be of the Langmuir – Hinshelwood type, that is, oxygen and ethylene are both adsorbed on the silver surface before reaction. Ethylene is not adsorbed on a clean, oxygen-free silver surface; this only occurs after preadsorption of oxygen. The reaction of ethylene with oxygen on the silver surface is also thought to be the rate-determining step in the absence of diffusion control. Kinetic data and models are available in the literature [95], [96], but these are usually restricted to very specific reaction conditions and catalysts, far from the complex situation of the industrial production. Ref. [97] describes kinetic studies with an industrial catalyst in different laboratory reactors to clarify the influence of the reaction products but does not take into account the influence of inhibitors, which is crucial to achieve high selectivities on an industrial scale.

4.3. Technology

An overview of the beginnings of commercial ethylene oxide production can be found in [1]. Production technologies for ethylene oxide plants based on the direct oxidation process are licensed by Shell, Scientific Design (SD), UCC, Japan Catalytic, Snam Progetti, and Hüls. Due to improved catalysts and production technology, large plants with capacities of up to 400 000 t/a can now be built.

The technologies are very similar, but differences exist, depending on whether air or pure oxygen is used for oxidation [44]. Shell plants use only pure oxygen, while Scientific Design and UCC have developed air-based oxidation plants as well. Figure 5 shows a simplified scheme for both the air- and oxygen-based processes. They both employ a recycle gas stream, which is continuously circulated through the reactors by compressors. The reactors consist of large bundles of several thousand tubes that are 6 – 13 m long and have an internal diameter of 20 – 50 mm. The catalyst is packed in the tubes

in the form of spheres or rings with a diameter of 3 – 10 mm; the initial selectivity of modern catalysts is ca. 80 – 90%, depending on the type of catalysts used. The catalysts are either very active, operating at an initial temperature of about 200 °C but with relatively low selectivities of 80 – 84 %, or they are highly selective but then need an initial temperature that is ca. 40 °C higher. Ethylene is converted at 200 – 300 °C and 1 – 3 MPa to produce ethylene oxide, CO_2, H_2O, and heat, as well as traces of acetaldehyde and formaldehyde; these products must be removed or separated from the recycle gas stream. The recycle gas is then reloaded with oxygen and ethylene and returned to the reactor.

Oxygen-Based Oxidation Process. At present, ethylene oxide is produced mainly by the oxygen-based process. The reactor tubes filled with the catalyst are surrounded by a coolant (water or a high-boiling hydrocarbon) that removes the reaction heat and permits temperature control. Heat is extracted either by pumping or by evaporating the coolant. If organic heat transfer media are used, the extracted energy is used to generate steam in secondary cycles, which is then employed for heating. The reaction heat also heats the recycle gas during its passage through the reactor. After leaving the reactor, the gas is likewise cooled by means of steam generation and/or used directly to heat the reactor inlet gases. Because of the different enthalpies of the partial and total oxidation reactions (see Section 4.1), the total available heat depends on the selectivity for the former (S) and amounts to $47\,250 - (434 \times S)$ kJ/kg ethylene.

Consequently, the quantity of energy released increases rapidly with decreasing selectivity. Therefore, the heat removal system must be sufficiently adaptable to cope with growing heat release in the reactor at rising reaction temperatures during the aging process of the catalyst, which is characterized by a decrease in its selectivity and activity in the course of its on-stream time. Temperatures on the coolant side (water or organic) of the catalyst tubes are restricted by the maximum design pressure, which usually does not allow temperatures higher than 300 °C. When the maximum temperature is reached the catalyst must be replaced.

After the gas from the reactor has been cooled, the ethylene oxide (1 – 2 %) and CO_2 (5 – 10 %) must be removed by scrubbing first with water and then with an aqueous potassium carbonate solution. In the ethylene oxide scrubber, virtually all of the ethylene oxide and small amounts of the other constituents of the recycle gas (CO_2, N_2, CH_4, CH_2CH_2, and aldehydes) dissolve in the water. The resulting aqueous ethylene oxide solution then passes to the ethylene oxide desorber. The ethylene oxide recovered as the head product in the desorber is subsequently stripped of its low-boiling components (i.e., the above-mentioned components of the recycle gas) and finally distilled, thereby separating into water and ethylene oxide.

A small proportion of the gas leaving the ethylene oxide scrubber (0.1 – 0.2 %) is removed continuously (combusted) to prevent the buildup of inert compounds (N_2, Ar, and C_2H_6), which are introduced as impurities with the reactants: ethylene (C_2H_6) and oxygen (Ar and N_2). The recycle gas from the scrubber is compressed, and a side stream is freed from carbon dioxide by further scrubbing with hot aqueous potassium carbonate solution. This leads to removal of the CO_2 not only by physical dissolution but also by reaction with potassium carbonate to give the hydrogencarbonate.

$$K_2CO_3 + CO_2 + H_2O \longrightarrow 2\,KHCO_3$$

The potassium carbonate solution, which is enriched with CO_2 is sent to the CO_2 desorber, where the CO_2 is stripped off at atmospheric pressure and is either released into the atmosphere or fed to a CO_2 utilization plant, after optional catalytic treatment (e.g., supported Pd/Pt) to remove residual hydrocarbons. The concentrations of the reactants in the bulk of the recycle gas, which is free from ethylene oxide and has a reduced CO_2 content, are restored to their starting levels by separate addition of oxygen, ethylene, inhibitor (1,2-dichloroethane, ethyl chloride or vinyl chloride), and if necessary, a diluent (CH_4). The gas is then returned to the reactor.

The oxygen that is used must be extremely pure (> 99 %) and is obtained by air separation. Nevertheless, a purge stream of recycle gas is still necessary due to the presence of traces of N_2 and Ar. Oxygen is added in a special mixing device that ensures rapid homogenization with the recycle gas. This is necessary because the explosive limit is locally exceeded at the mixing point.

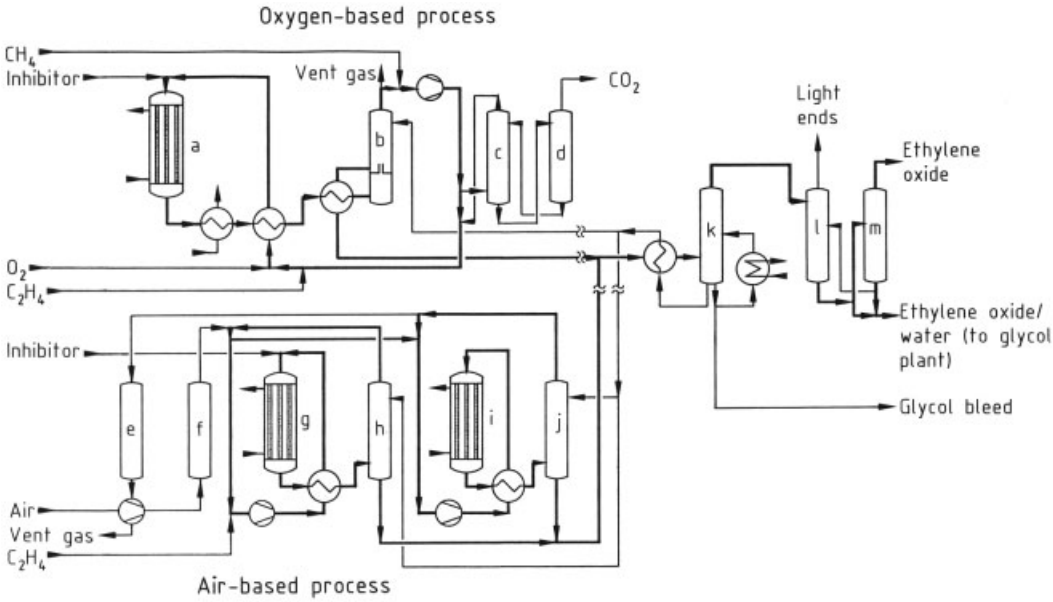

Figure 5. Flow scheme for ethylene oxide production by the oxygen-based or air-based oxidation of ethylene
a) Reactor; b) Ethylene oxide scrubber; c) CO_2 scrubber; d) CO_2 desorber; e) Off-gas purification; f) Air purification; g) Primary reactor; h) Primary ethylene oxide scrubber; i) Secondary reactor; j) Secondary ethylene oxide scrubber; k) Ethylene oxide desorber; l) Stripping column; m) Ethylene oxide distillation

The ethylene is also usually very pure (> 99.5 %) and must be free from the strong catalyst poisons sulfur and acetylene. Methane, used as a diluent, also must be free from sulfur compounds.

Natural gas is used as methane source and usually requires cleaning steps to attain the required purity. Sulfur compounds, which poison the catalyst irreversibly, are removed by adsorption beds. Higher hydrocarbons, which affect the reactor performance, mainly by removing the chlorine inhibitor from the silver surface, are removed by distillation or with molecular sieves.

Air-Based Oxidation Process. The air-based process is similar to the oxygen process, but some differences exist. Air introduces a large amount of nitrogen into the recycle gas, which means that a large amount of purge gas must be vented to maintain a constant nitrogen concentration in the recycle stream. The quantity of gas that is vented removes sufficient CO_2 to make CO_2 scrubbing unnecessary. However, the off-gas leaving the primary reactor still contains so much ethylene that it must be further converted in a subsequent secondary or purge reactor before it can be vented into the atmosphere.

The reaction conditions cannot be tailored to the needs of ethylene oxide formation as optimally as in the oxygen-based process. The conversion of ethylene is higher than in the oxygen-based process, especially in the secondary reactors, so as to obtain an acceptable level of ethylene loss in the purge gas. Since selectivity is inversely related to ethylene conversion, it follows that the air-based process has a lower selectivity.

The conditions and gas compositions used in the oxygen- and air-based processes are listed in Table 5.

Table 5. Operating parameters used in the air-based and oxygen-based production of ethylene oxide

Parameter	Air-based process	Oxygen-based process
C_2H_4 concentration, vol %	2 – 10	15 – 40
O_2 concentration, vol %	4 – 8	5 – 9
CO_2 concentration, vol %	5 – 10	5 – 15
C_2H_6 concentration, vol %	0 – 1	0 – 2
Ar concentration, vol %		5 – 15
CH_4 concentration, vol %		1 – 60
Temperature, °C	220 – 277	220 – 275
Pressure, MPa	1 – 3	1 – 2.2
GHSV ∗, h^{-1}	2000 – 4500	2000 – 4000
Pressure drop, kPa,	40 – 200	
C_2H_4 conversion, %	20 – 65	7 – 15
Selectivity, %	80	80 – 90

∗ GHSV = gas hourly space velocity.

Reactor. Since oxidation of ethylene is highly exothermic, reaction in a fluidized bed would appear to be appropriate. However, attempts to develop such a process on a commercial basis have not produced any advantages for selectivity [98], [99] and have led to problems due to abrasion and sintering. As a result, all ethylene oxide plants currently employ fixed-bed tubular reactors. The diameter of the tubes must not be excessive so as to ensure sufficient heat transfer to the heat transfer medium. A tube diameter of 20 – 40 mm is common. The diameter of the catalyst particles has an upper limit determined by the need for satisfactory gas mixing and a lower limit resulting from an increasing drop in pressure. A diameter of 3 – 8 mm is used.

Cooling. Each of the two possible heat extraction systems – circulation or evaporation of the coolant – has its advantages and disadvantages. Coolant circulation requires a large amount of liquid (pumping power), but allows a defined temperature increase in the direction of gas flow. Ethylene and oxygen concentrations in the cycle gas decrease on passage through the reactor; this leads to lower conversion, which can be compensated by controlled temperature increase toward the reactor outlet.

In evaporative cooling, an organic coolant (e.g., isododecane) or water is used. The reaction temperature is controlled very effectively by a control valve that keeps the pressure of the boiling coolant at the appropriate pressure. The whole temperature range from 190 to 290 °C, required by the aging catalyst, can be covered by evaporative cooling. Organic coolants with a boiling point of about 190 °C at atmospheric pressure are used. The hydrostatic height of the coolant in the reactor results in an unavoidable temperature increase towards the bottom of the reactor. Since evaporation extracts heat more efficiently than circulation, evaporative cooling requires a smaller volume of circulating coolant. The axial temperature differences can be minimized more easily. Large amounts of liquid coolant leave the reactor with the coolant vapor. The liquid is recovered in a separator and returned to the reactor. The coolant present in the reactor and in the separator constitutes a safety reserve during uncontrolled increases in ethylene conversion in the reactor ("runaway").

The flammability of organic coolants is a clear disadvantage compared to water, especially at the necessary high temperatures. However, water results in much higher pressures on the coolant side and hence higher apparatus costs.

When organic coolants are used, the coolant vapors are condensed in heat exchangers producing steam, after which the liquefied coolant is returned to the reactor. The steam serves to heat distillation columns in the plant. If water is used as coolant the steam leaving the reactors can be used directly. The higher the selectivity of the catalyst, the more steam must be imported into the plant.

Improved catalysts with high activities and selectivities operate at a lower temperature (ca. 220 °C) and allow smaller reactor volumes. This has encouraged the use of water-cooled reactors; however, organic heat transfer media (e.g., Dowtherm) are still widely employed. A detailed discussion of heat transfer with organic heat transfer media in terms of optimum selectivity and safety can be found in [100].

Ethylene Conversion. The selectivity of ethylene oxide formation depends on ethylene conversion. Selectivity decreases more or less linearly with increasing ethylene conversion [101]. Therefore, the highest selectivities are achieved with minimum conversions, but the resulting ethylene oxide concentrations are then too low for commercial purposes. Thus, ethylene conversion is chosen to achieve ethylene oxide concentrations of 1 – 3 vol % at the reactor outlet.

Byproducts. Besides large quantities of CO_2 produced by total oxidation of ethylene, ethylene glycols and small amounts of acetaldehyde and formaldehyde are formed and must be processed in ethylene oxide plants. The acetaldehyde stems from isomerization of ethylene oxide, catalyzed by the catalyst support and by rust on apparatus walls. Formaldehyde is produced by oxidation of ethylene oxide with and without participation of the wall surface [102]. Ethylene glycols are unavoidably produced when ethylene oxide is scrubbed from the recycle gas with water and is subsequently stripped from the aqueous solution by heating. Monoethylene glycol [*107-21-1*],

diethylene glycol [*111-46-6*], triethylene glycol [*112-27-6*], and higher poly(ethylene glycols) are formed when ethylene oxide comes into contact with water [103] (→ Ethylene Glycol). Some of the cycle water must be bled off continuously and the ethylene glycol removed to prevent it from accumulating (see Fig. 5). This glycol is of inferior quality to that synthesized by ethylene oxide hydrolysis. Special methods for purifying the resulting glycols have been described, e.g., treatment with ion exchangers and activated charcoal [104].

Materials. Since ethylene oxide is noncorrosive, the reactors and the sections of the plant that convey ethylene oxide are usually made of mild steel. To reduce the formation of acetaldehyde by isomerization of ethylene oxide on rust, stainless steel is increasingly being used for the reactors and the ensuing cycle gas lines, where temperatures are highest. The sealing materials for valves, flanges, and pumps must be chosen carefully [105]. Many materials used in chemical industry for O-rings, packings, and gaskets are not resistant to the influence of ethylene oxide and have been the cause of severe incidents [106]. The materials in the system used to remove the carbon dioxide must also be carefully selected because of possible CO_2 corrosion [107].

If rust is present in steel tubes or containers, polymer formation [38], increased viscosity, and brown discoloration are to be expected.

Current plant design is predominantly tailored to the oxygen-based process because this is normally more economical than the air-based process. In exceptional cases, however, the air-based process may be preferred depending on local factors (e.g., the availability of oxygen) [108].

Possible Developments. Alternative routes from ethylene to ethylene oxide are being studied, e.g., Tl(III)-catalyzed oxidation in solution [108], electrochemical oxidation [109], and enzymatic oxidation [110]. However, these processes appear to be far from industrial application. About half of the ethylene oxide produced is converted into ethylene glycol. Alternative syntheses for ethylene glycol based on carbon monoxide, formaldehyde, and ethylene are being developed [111–114] (→ Ethylene Glycol); these are more promising than those for ethylene oxide. With increasing raw material prices, the appearance of an economical alternative is foreseeable. Naturally, such an alternative would have an adverse effect on ethylene oxide capacities, but the economics of ethylene oxide production could be improved by

1) Selective production of monoethylene glycol from ethylene oxide, without the formation of higher ethylene glycol homologues, e.g., via glycol carbonate [44]
2) Reducing energy consumption
3) Improving the selectivity, capacity, life-span, and activity of the catalysts

5. Environmental Protection and Ecology

Ethylene oxide plants using the direct oxidation process represent a clear environmental improvement over the chlorohydrin process. Environmental and other important aspects of ethylene oxide production are discussed in EPA studies [115]. Both oxygen- and air-based plants emit two main off-gas streams: recycle vent gas and CO_2. In the oxygen-based process, ca. 70 % of the recycle vent gas consists of hydrocarbons (ethylene and methane), which can be combusted (by torch or in a steam generator). The corresponding off-gas in the air-based process is produced in considerably greater quantities, but contains only up to ca. 2.3 % hydrocarbons. It can be purified by catalytic oxidation. The CO_2-rich vent gas in the oxygen-based process contains ca. 0.2 % hydrocarbons. This concentration can be reduced substantially by depressurizing the CO_2-rich absorbent in two steps. The gas from the first step is rich in hydrocarbons and is recycled, only the gas from the second step is vented [116]. A catalytic treatment (e.g., supported Pd/Pt) may be required by environmental regulations to remove hydrocarbons before the CO_2 is released to the atmosphere but may also be necessary to ensure the purity of the CO_2 for further use. The CO_2-rich vent gas stream from air-based plants is produced in much smaller quantities and contains ca. 5 % hydrocarbons that can be removed by combustion. Ethylene oxide can be removed from waste gases, such as are generated during sterilization, by combustion with oxygen [117] or by an acid scrub [118],

[119], [121]. Biodegradation of the ethylene glycol generated during ethylene oxide hydrolysis presents no difficulties. If ethylene oxide occurs as a mixture with fluorohydrocarbons, recovery by drying and subsequent compression is recommended [120]. Methods for handling ethylene oxide reaction vessels without the release of vent gas have been described [122].

Ethylene oxide is toxic to microorganisms and fish [123]. The LC_{50} value for fish (*Pimephales promelas*) is 84 mg/L (exposure time 96 h). However, in free-flowing waters the ethylene oxide concentration decreases continually (after 4 h at 25 °C in moving water, the concentration drops by ca. 95 %) due to a combination of evaporation, hydrolysis (half-life at 25 °C ca. 14 d), and biodegradation. The ethylene glycol produced as a result of ethylene oxide hydrolysis is considerably less toxic to aquatic organisms ($LC_{50} > 10\,000$ mg/L) and is readily biodegradable. Effluents (e.g., from ethylene oxide production plants [115]) containing ethylene oxide are therefore treated biologically after the ethylene oxide has been converted into ethylene glycol.

Degradation of ethylene oxide in air is relatively slow and thought to proceed mainly by a slow reaction with hydroxyl radicals. Disregarding other removal mechanisms, this results in a lifetime of 330 d. The reaction with ozone is too slow to represent a significant removal pathway [124]. Another report indicates a lifetime of 100 – 200 d [125]

According to TA Luft [126], vent gas streams must not contain more than 25 g/h or 5 mg/m^3 ethylene oxide.

6. Quality Specifications

Ethylene oxide is a major industrial product that is obtained with a consistently high purity irrespective of the production process used. Typical specifications are given in Table 6.

Table 6. Typical specifications of ethylene oxide

Appearance	clear, colorless
bp at 101.3 kPa	10.8 °C
Water content ∗	50 mg/kg
CO_2 content ∗	10 mg/kg
Aldehyde content ∗	50 mg/kg
Ethylene oxide content (min.) ∗	99.5 %

∗ As determined by gas chromatography.

7. Analysis

An overview of analytical methods for ethylene oxide can be found in [127]. An established method for determining ethylene oxide is based on its reaction with $MgCl_2$ [128]. In view of the low TRK value (5 mg/m^3), accurate methods are needed to determine traces of ethylene oxide in air.

An adsorption chromatography method is reported to be suitable for concentrations exceeding 0.15 ppm [129]. The NIOSH method 1607 [134] with a working range of 0.04 – 1.7 ppm (0.09 – 3 mg/m^3) is based on adsorption with charcoal, derivatization with HBr to give 2-bromoethanol, and gas chromatography. The CS_2 used in the NIOSH method can be replaced by *N*,*N*-dimethylacetamide [130]. Methods using adsorption on charcoal are susceptible to the influence of temperature and humidity [131], [132].

A field comparison of four methods of monitoring ethylene oxide by portable devices – two passive monitors, one sorbent tube (OSHA), and one GC method (NIOSH) – showed their viability for measuring the workplace concentration for short-term and full-shift exposure [133].

Exposition to ethylene oxide can be monitored by the detection of (2-hydroxyethyl) valine in hemoglobin, the method resulting in accurate data of time-integrated exposure [39].

8. Handling, Storage, and Transportation

Ethylene oxide has repeatedly caused serious explosions, fires, and accidents [106], [135], [136]. It is an extremely hazardous substance because it can explode and is both highly flammable and extremely reactive (exothermic reactions). Aqueous solutions containing > 4 wt % ethylene oxide are flammable; flash points are given in Table 2. Furthermore, ethylene oxide is toxic and poses a danger both to health and to the environment. The inherent hazards of the product must be known; all prescribed safety measures and legal requirements must be observed. Personnel handling ethylene oxide must receive appropriate training.

Safety precautions for handling ethylene oxide and its dangerous properties are described in detail in the data sheets provided by the manufacturers (e.g., [106], [137]) and elsewhere (e.g., [105], [138–141]). Therefore, only the most important points are discussed here.

Explosion and Fire Control. Pure ethylene oxide vapor or ethylene oxide vapor mixed with air or inert gases can decompose explosively. Explosiveness depends on pressure, temperature, concentration, the type, form, and energy of the ignition source, and the type of container. High pressure can be generated on explosion of ethylene oxide; therefore, for safe handling the exact explosive limits must be known. Pure ethylene oxide vapor at 101.3 kPa decomposes when passed through a heated platinum coil [142]. The temperature at which decomposition starts (decomposition temperature) was calculated to be 571 °C. Calculations were based on the temperature at the tube exit and the thermodynamic data for ethylene oxide and its decomposition products. The thermal decomposition of ethylene oxide and ethylene oxide – nitrogen mixtures has been investigated [143]. If ethylene oxide vapor is introduced into a preheated vessel at 101.3 kPa, explosive decomposition takes place at ca. 500 °C [143]. The decomposition temperature is reduced by an increase in the pressure (e.g., ca. 450 °C at 1 MPa) but is increased by the addition of nitrogen. Explosive thermal decomposition is still possible below atmospheric pressure, but the necessary temperature then increases above 500 °C.

The ignition properties of ethylene oxide and its mixtures with air and inert gases have been studied by several authors. Pure ethylene oxide vapor can ignite even at a pressure as low as 48 kPa (electric spark) or 20.2 kPa (mercury fulminate) [144]. The maximum theoretical explosive pressure is ca. 10 times the inital pressure, but can increase to 20 times the initial pressure if liquid ethylene oxide is present. This phenomenon occurs because liquid ethylene oxide evaporates and participates in the decomposition reactions that take place in the vapor phase [145]. Mixtures of ethylene oxide with N_2, CO_2, methane, and air do not ignite over certain concentration ranges; the upper explosive limit is always 100 % ethylene oxide, since pure ethylene oxide also decomposes if ignited. In the presence of oxygen (air), combustion and decomposition take place simultaneously and the ignition temperature (or energy) is reduced. The ignition temperature of ethylene oxide in air at 101.3 kPa is 429 °C [146]. As the ethylene oxide concentration increases, the proportion destroyed by decomposition also increases [21], [147]. The minimum value cited for the lower explosive limit of ethylene oxide – air mixtures is 2.6 % [148]; the explosive range of ethylene oxide – air mixtures is accordingly 2.6 – 100 %. Figures for an upper explosive limit of ca. 80 % at 101.3 kPa are probably due to differences in the apparatus and methods used [149–151].

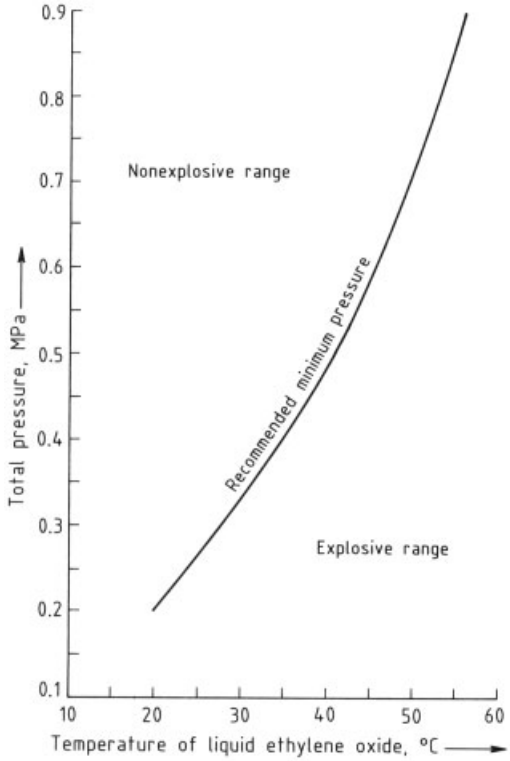

Figure 6. Recommended total pressure for the safe storage of ethylene oxide as a function of the temperature after blanketing with nitrogen

Explosive decomposition of ethylene oxide in closed containers can be suppressed by blanketing with an inert gas until the total pressure of the nonexplosive range is reached. Blanketing agents with higher thermal conductivities suppress explosions more effectively. Quotation of the minimum total pressure necessary for inert gas blanketing is important; values cited in [21]

were obtained without consideration of the pressure and temperature dependence and, therefore, do not apply at higher pressures. Figure 6 shows the total pressure currently recommended for inert gas blanketing of ethylene oxide with nitrogen as a function of temperature [105]. The explosive limits of other mixtures of ethylene oxide with inert gases and air can be found in the literature, e.g., ethylene oxide with H_2O [152]; N_2 [143], [152]; N_2-H_2O [144]; CO_2-H_2O [144]; CH_4 [145]; CO_2 [21], [152], [153]; C_3H_6 [154]; C_4H_9 [154]; N_2 – air [21]; CH_4 – air [21]; CO_2 – air [148]; CF_2Cl_2 – air [151], [155].

As has already been explained, liquid ethylene oxide can participate in the decomposition that is initiated in the vapor phase. Explosion of liquid ethylene oxide initiated by an ignition source within the liquid was first described in 1980 [156], [157], but once again, the decomposition reaction is thought to take place in the gas phase created by the source of ignition and is maintained by the subsequent evaporation of ethylene oxide. Figure 7 shows the temperature limits for ignition of liquid ethylene oxide as a function of the pressure. At pressures between 10 and 15 MPa, the ignition temperature decreases rapidly. This phenomenon is presumed to be due to the fact that the critical state is reached. The conditions under which liquid ethylene oxide is usually produced, stored, or processed lie within the nonexplosive range (Fig. 7).

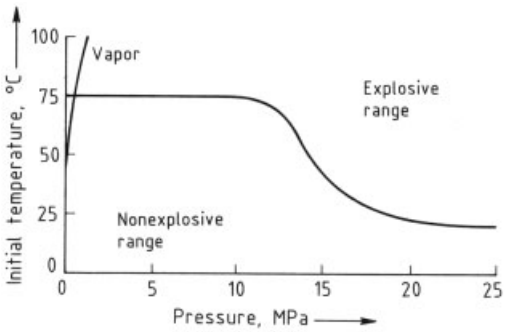

Figure 7. Explosive limits for pure liquid ethylene oxide as a function of pressure and temperature

The highly exothermic reactions of ethylene oxide represent a further hazard [135]. If traces of polymerization initiators (e.g., amines) find their way into a large supply of ethylene oxide (e.g., in a tank), polymerization starts slowly and then accelerates because of the resulting temperature increase. Polymerization proceeds quasi-adiabatically, leading to sudden and rapid increases in temperature and pressure that can rupture the container [158], [159], which may be followed by explosive decomposition of the released ethylene oxide vapor (see Fig. 8).

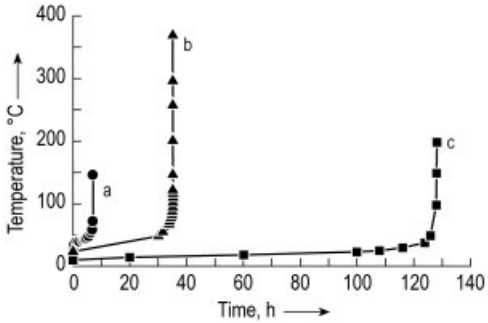

Figure 8. Runaway reactions of ethylene oxide (EO) [169] a) Dewar polymerizaation of EO (10 ml)/NaOH (0.02 g)/H_2O [167]; b) Calculated adiabatic reaction of EO (40 %) in H_2O (initial temperature 27 °C) [168]; c) Polymerization of EO/NMe_3 (0.184 %)

Although rust is only a mild catalyst for ethylene oxide polymerization, its presence in heat exchangers or other hot equipment can lead to ignition of the ethylene oxide vapor by a combination of polymerization and disproportionation to C_2H_4, CO_2, H_2, and H_2O (explosion in the UCC plant, Seadrift, 1991). γ-Fe_2O_3 and γ-FeOOH are the active species in this reaction [160].

Insulation fires can occur if ethylene oxide leaks into insulating material such as asbestos, magnesium silicate, calcium silicate, and mineral wool by a combination of effects and reactions. Like other organic compounds, ethylene oxide can have considerably lower autoignition temperatures in insulation material than otherwise [161], [162]. Isomerization of ethylene oxide to acetaldehyde in the presence of rust results in an additional decrease in the autoignition temperature [141]. In the presence of water, ethylene oxide may be hydrolyzed to ethylene glycol and subsequently polymerize to poly(ethylene glycols) with evolution of heat. Eventually, the organic material collected in the insulation (ethylene oxide, acetaldehyde, glycols) may be oxidized by air. The resulting hot spots on the surface of the apparatus can trigger an explosive decomposition of ethylene oxide vapor in the column (explosion in the BP plant, Antwerp,

1987 [163] and in the BASF plant, Antwerp, 1989 [164]). To avoid these risks, special insulating material with a low specific surface and closed cells (glass foam) is used, and special methods for insulation have been developed. A gap between the outer wall of the apparatus and the insulation prevents accumulation of organic material in the insulation and allows monitoring of organic compounds in the gap.

The accidents in ethylene oxide plants also initiated a fundamental reevaluation of the distillation section. The use of structured packing instead of trays can prevent the propagation of ethylene oxide decomposition. A beginning decomposition in the column can be recognized in time to be quenched with water [165], [166].

Because of the above-mentioned hazards, the following potentially dangerous situations must be prevented:

1) Leakage of liquid or gaseous ethylene oxide
2) Entry of air, oxygen, or reactive impurities into ethylene oxide containers
3) Ignition sources in danger areas
4) Overheating of ethylene oxide (chemical reaction, fire, etc.)

Leaks or spills should be promptly diluted with sufficiently large volumes of water to reduce the ethylene oxide concentration to less than 4 wt %. Mixtures of ethylene oxide and water will still burn if the ethylene oxide concentration exceeds 4 wt %. Fires should also be extinguished with large volumes of water, carbon dioxide extinguishers can be used for small fires.

Storage and Transportation. Ethylene oxide is transported in tank cars and containers and can be conveyed by rail or by sea and, in many countries, also by road. In Germany, conveyance by road is possible with special permission and under certain conditions (GGVS/ARD § 7). Hazard classifications for ethylene oxide follow:

Water hazard class	WGK 2 (no. 235)
CFR 49	no. 172.101, flammable liquid
GGVS/GGVE/RID/ADR	class 2, no. 4 ct
ADNR	class 2, no. 8 AF
EMS	2-06
MFAG	365
GGVSee, IMDG Code	class 2
UN no.	1040
JATA no.	722
Temperature class (DIN 57 165)	2
Explosion limit (DIN 57 165)	IIB

Storage and transport containers for ethylene oxide are usually made from steel, but steel is suitable only if special measures have been taken to prevent rust formation [38]. Rust acts as a mild polymerization catalyst and is dispersed in the ethylene oxide by polymerization starting on available surfaces. Polymer concentrations as low as ca. 110 mg/kg produce an increase in viscosity, may cause brown discoloration, may block filters, valves etc., and form deposits on the container walls. Therefore, steel containers should be blasted before first use, the blasting residues carefully removed, and the container flushed with nitrogen. Stainless steel is increasingly being used to avoid these problems. When aqueous solutions of ethylene oxide are handled, precipitation of solid hydrates (cf. Table 2) may cause blockages.

In Europe, tank cars are tested at a pressure of 1.6 MPa and they must be protected from the sun by a protective roof or insulation (GGVE). Ethylene oxide – nitrogen mixtures can be transported at a maximum pressure of 1 MPa at 50 °C (GGVS/RID). The proportion of nitrogen in the vapor phase must be high enough to ensure that explosion cannot occur at or below this temperature. In the United States railroad tanks are insulated and equipped with pressure-release valves.

Safety measures that should be observed when reactions with ethylene oxide are performed or when it is used for sterilization can be found in [105].

Personal Protective Equipment. To avoid skin contact with ethylene oxide, goggles as well as protective clothing (gloves, boots, suits, aprons) have to be used. As ethylene oxide permeates easily through most materials, the choice of an adequate material is crucial. Butyl rubber is normally a suitable material, but as quality and thickness play an important role, a test of the permeation resistance against ethylene oxide is recommended before use.

Leather clothing and footwear represent an often neglected risk when contacted with ethylene oxide or its aqueous solution, because the severe and slowly healing damage to the skin appears only after a induction period of many hours. Therefore leather articles must be discarded at once when contaminated with ethylene oxide. Other clothing contaminated with ethylene oxide must be taken off immediately and

discarded or decontaminated. Immediate flushing with copious water can avoid detrimental action on eyes and skin.

If ethylene oxide concentrations of 0.25 ppm in the air are reached, respiratory equipment must be used [170]. Gas masks are permitted only up to certain exposure limits, and their duration of use is restricted according to the ethylene oxide concentration [171]. Beyond a certain concentration level, equipment has to be used which is independent of the ambient atmosphere (e.g., positive-pressure self-contained breathing apparatus) [172]. Also above ethylene oxide concentrations of 0.25 ppm restrictions in the employment of juveniles and pregnant women are effective [170].

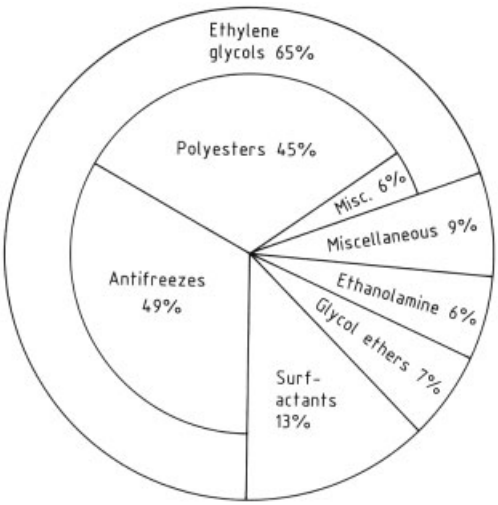

Figure 9. Uses of ethylene oxide

9. Uses

Ethylene oxide is an excellent disinfectant, sterilizing agent, and fumigant when it is used as a nonexplosive mixture with N_2, CO_2, or dichlorofluoromethane. The gas penetrates into pores and through packaging or clothing. It can be used, for example, to sterilize surgical instruments in hospitals or to remove pests and microorganisms from spices, furs, etc.

However, most ethylene oxide is converted into other products. Figure 9 shows the percentages of ethylene oxide used in these derivatives in the United States. Although this distribution was determined in 1978, it still applies today and is typical of the world market.

Products derived from ethylene oxide have many different uses, only the most important ones are listed here:

Monoethylene glycol [*107-21-1*] [173]: Antifreeze for engines, production of poly(ethylene terephthalate) (polyester fibers, foils, and bottles), and heat transfer liquids.

Diethylene glycol [*111-46-6*] [174]: Polyurethanes, polyesters, softeners (cork, glue, casein, and paper), plasticizers, gas drying, solvents, and deicing of aircraft and runways.

Triethylene glycol [*112-27-6*] [175]: Lacquers, solvents, plasticizers, gas drying, and humectants (moisture-retaining agents).

Poly(ethylene glycols) [176]: Cosmetics, ointments, pharmaceutical preparations, lubricants (finishing of textiles, ceramics), solvents (paints and drugs), and plasticizers (adhesives and priinks).

Ethylene glycol ethers [177]: Brake fluids, detergents, solvents (paints and lacquers), and extractants for SO_2, H_2S, CO_2, and mercaptans from natural gas and refinery gas.

Ethanolamine [*141-43-5*] : Chemicals for textile finishing, cosmetics, soaps, detergents, gas purification (H_2S, SO_2, and CO_2).

Ethoxylation products of fatty alcohols, fatty amines, alkyl phenols, cellulose, poly(propylene glycol) [28]: Detergents and surfactants (nonionic), biodegradable detergents, emulsifiers, and dispersants.

Ethylene carbonate [*96-49-1*] : Solvents.

10. Economic Aspects

In 1980, ca. 16 % of world ethylene production was used to synthesize ethylene oxide; this use was second only to polyethylene production (44 %). World production of ethylene oxide in 2000 was ca. 15×10^6 t/a. Ethylene oxide is an important raw material for major consumer goods in virtually all industrialized countries. Table 7 gives an overview of worldwide ethylene oxide capacities. Excess capacities were already apparent in 1984 [178].

11. Toxicology and Occupational Health

The toxicological properties of ethylene oxide are mainly determined by its reactivity with nu-

Table 7. World ethylene oxide capacities in 2000

Country	Company	Location	Licensor*	Capacity, 10^3 t/a
Australia	ICI	Botany	SD	**40**
Canada				**775**
	Dow	Saskatchewan	Dow	285
	UCC	Prentiss I	UCC	220
		Prentiss II	UCC	270
W. Europe				**2615**
Belgium	BASF	Antwerp	Shell, s	350
	Ineos	Antwerp	Shell, s	350
Germany	BASF	Ludwigshafen	Shell, s	215
	Erdölchemie	Dormagen	Shell, s	200
	Clariant	Gendorf	Shell, s	200
	RWE/DEA	Marl	Hüls, Shell, s	150
France	BP	Lavera	Shell, s	200
Great Britain	UCC	Wilton	Shell, s	300
Italy	Enichem	Priolo	SD	30
		Gela		40
Netherlands	Dow	Terneuzen	Dow, s	150
	Shell	Moerdjik	Shell, s	250
Sweden	Akzo Nobel	Stenungsund	Nippon Sh./SD, s	80
Spain	La Seda	Tarragona	Shell, s	100
E. Europe				**950**
Bulgaria	Neftochim	Burgas	SD	80
Slovakia	Slovnaft	Bratislava	Shell, s	40
Poland	Petrochemia	Plock	Shell, s	60
			Snam	30
Romania	Romchim	Brazi	SD	35
	Arpechim	Pitesti	SD	35
Ukrainia	Nephtekhim	Dzherzhinsk	SD	200
Russia	Techmashimp.	Khazan	Japan Catalytic	100
		Nishnekamsk	SD	200
		Nishnekamsk	SD	120
	Salavat Neft Org.	Salsvat		50
Asia				**3845**
China	CNPC	Liaoyang	Hüls, 1	60
		Fushun		50
		Jilian		40
		Jilin	SD	80
		Dushanzi		30
	Sinopec	Maoming		80
		Yanshan		65
		Yangzi		240
		Jinshan		130
		Tianjin		55
	Zhejiang Chem. Pet	Zhejiang		20
		Beijing		50
India	Glycols India	Kashipur		20
	Indian PC	Kojali	SD	20
		Vadodana		5
		Nagothane		45
	SM Deyechem Ltd.	Pune	SD/Toyo	20
	Reliance Ind.	Hazira		265
	NOCIL	Bombay	Shell, s	20
Indonesia	O.T. Yasa Ganesha Pura	Merak	SD, s	185
Japan	Nippon Shokubai	Kawasaki	Japan Catalytic	240

cleophilic groups such as carboxyl, amino, phenolic hydroxide, or sulfhydryl groups.

In humans, acute inhalative poisoning causes headache, nausea, and vomiting within a few minutes [179–181]. Local irritation results in dyspnea [182]. Myocardial damage [183], excitation, numbness, and finally coma [179] follow. After dermal exposure, blisters are formed on the skin, and symptoms similar to those found after inhalation appear due to skin absorption [184]. Sensitization has been reported after repeated dermal contact [185]. Repeated inhalation of ethylene oxide leads to sensory-motor polyneuropathy and impaired memory [186], [187]. Cataracts may be formed after chronic exposure [188].

Table 7. (Continued)

Country	Company	Location	Licensor*	Capacity, 10^3 t/a
	Mitsubishi PC	Yokkaichi	Shell, s	90
		Kashima	Shell, s	240
	Mitsui PC	Chiba	Shell, s	120
	Nisso Murazen	Chiba	SD, s	120
	Nisso Yuka	Yokkaichi	Shell, s	85
	Semppoku EO Daesan	Osaka	Shell, s	130
North Korea		Puyangtang	Japan Catalytic	10
South Korea	Honam PC	Yeochon	Shell, s	115
		Yeochon	Shell, s	100
		Yeochon	Shell, s	100
	Hyundai PC	Daesan	SD, s	90
		Daesan		160
	Samsung	Daesan	SD, s	80
Singapur	Ethylenegl. Sing.	Merbau	Shell, s	125
Taiwan	Orient. UCC	Koahsiung	UCC, s	190
	CMFC	Koahsiung	SD	40
		Koahsiung	SD	60
	Nan-Ya	Mailiao		240
Latin America				**720**
Brasil	Oxiteno	Camacari, Bahia	SD, s	130
			SD, s	105
		Maua, Sao Paulo	SD, s	45
Mexico	Pemex	La Cangrejera	SD	110
		Pajaritos	SD	40
		Morelos	SD	220
Venezuela	Pralca	Santa Rita	SD, s	70
Middle East				**1370**
Saudi Arabia	Sharq	Al Jubail	Shell, s	360
		Al Jubail	Shell, s	360
	Yanpet	Yanbu	SD, s	310
Kuwait	Equate	Shuiaba		270
Turkey	Petkim Petro.	Aliaga	Shell, s	70
United States				**4083**
	BASF Wyandotte	Geismar, LA	Shell, s	140
		Geismar, LA	Shell, s	160
	Celanese	Clear Lake, TX	Shell, s	305
	Dow	Plaqemine, TX	Dow, l	260
	Eastman	Longview, TX	Shell, s	100
	Huntsman	Port Neches, TX	SD, l	455
	Sun	Brandenburg, KY	Shell, s	55
		Marcus Hook, PA	Shell, s	55
	Equistar	Bayport, TX	Shell, s	340
		Beaumont, TX	SD	308
	Shell	Geismar, LA	Shell, s	165
		Geismar, LA	Shell, s	240
		Geismar, LA	Shell, s	180
	UCC	Seadrift, TX	UCC, l	420
		Taft, LA	UCC, s	350
		Taft, LA	UCC, l	310
	Nan-Ya	Point Comfort, TX	SD, s	240
Total				**14 396**

* l = air process; s = oxygene process; SD = Scientific Design.

Oral Toxicity. The acute oral toxicity (LD_{50}) of ethylene oxide in the rat is 330 mg/kg [189].

Inhalation. The 4-h LC_{50} value is 835 mL/m^3 in the mouse and 1460 mL/m^3 in the rat [190]. Acute symptoms largely resemble those observed in humans. Subchronic administration of ethylene oxide (100 and 200 mL/m^3) to cats by means of inhalation for a period of 22 days produced anorexia, apathy, atoxia, and paralysis of the hind quarters [191]. Lethalities also occurred, postmortem examination revealed liver and kidney damage accompanied by hyperemia and perivascular bleeding in various organs (e.g., the brain) [191]. Rodents that inhaled air containing ethylene oxide at a concentration of 300 – 400 mL/m^3 displayed additional

local irritation and severe, primary neurotoxic atrophy of the musculature in their hind quarters [190]. Hemotoxic effects have also been observed in rodents [192]. Monkeys exposed to inhalative concentrations of 204 mL/m^3 for up to 226 days showed impaired reflexes, reduced sensitivity to pain, and neurotoxic muscular atrophy of their hind quarters [193].

Teratogenicity. In studies on rats exposed to maximum inhalative concentrations of 100 [194] or 150 mL/m^3 [195], symptoms of intolerance were observed in the dams but no malformations were observed in the fetuses. Ethylene oxide does not produce teratogenic effects in rabbits [195].

Mutagenicity. Ethylene oxide is mutagenic due to its alkylating properties. It alkylates the germ cells of male rats [196] and produces mutations in *Drosophila melanogaster* [197], *Neurospora crassa* [198], and *Salmonella typhimurium* (Ames test) [199], [200]. Dominant lethal mutations and chromosome aberrations are induced in male rats [201]. Increased numbers of micronuclei are found in the polynuclear erythrocytes of rats [202] and mice [203], [204]. Chromosomal defects are induced in cultured amniotic cells from humans [205]. Increased exchange between sister chromatids is found in monkeys and rats [206], [207]; the same effect is observed in workers who have been exposed to ethylene oxide (cumulative dose 0.5 – 50 g) [208–211].

Carcinogenicity. A large number of reports have been published on the carcinogenic properties of ethylene oxide in animals. Dilute ethylene oxide applied to the skin of rats did not induce tumor formation [212]. Subcutaneous injection of ethylene oxide in rats did not have any systemic carcinogenic effects [213], [214]. However, intragastric administration led to an increase in the incidence of epithelial carcinomas in the gastric mucosa [215]. An increased incidence of brain tumors and mononuclear cell leukemia was found in rats that had inhaled ethylene oxide at concentrations of 10, 33, or 100 mL/m^3 over a period of two years [216]. An increased incidence of peritoneal mesotheliomas was also observed in the animals exposed to concentrations of 33 and 100 mL/m^3. Results of human epidemiological studies on workers exposed to ethylene oxide differ [217–220]. Ethylene oxide is classified as a "putative human chemical carcinogen" [221].

Permissible Exposure Limits. Ethylene oxide is classified as a class 2 carcinogen by the German MAK commission (TRK = 1 mL/m^3 = 2 mg/m^3) and as a class A2 carcinogen by the ACGIH (TLV-TWA = 1 mL/m^3 = 2 mg/m^3)[222].

12. References

General References

1. G. O. Curme: *Glycols,* Reinhold Publ. Co., New York 1952, pp. 74 – 113.
2. S. A. Miller: *Ethylene and its Industrial Derivatives,* Ernest Benn Limited, London 1969, pp. 513 – 638.
3. *Kirk-Othmer*, 4th ed., **9**, 915 – 959.
4. *Beilstein,* **XVII** 4 – 5, **XVII(E1)** 3, **XVII(E2)** 9 – 11, **XVII(E3/E4)** 3 – 10, **XVII (E5)** 3 – 7.
5. *Ullmann,* 4th ed., **8,** 215 – 221.

Specific References

6. C. Hirose, *Bull. Chem. Soc. Jpn.* **47** (1974) no. 6, 1311 – 1318.
7. J. A. Boatz, A. Jerry, M. S. Gordon, S. Mark, *J. Phys. Chem.* **93** (1989) 3025 – 3029.
8. A. Wurtz, *Justus Liebigs Ann. Chem.* **110** (1859) 125 – 128.
9. Société Française de Catalyse Generalisée, FR 729 952, 1931; 739 562, 1931 (T. E. Lefort).
10. J. G. Filser B. Denk, M. Tornquist, W. Kessler, L. Ehrenberg, *Arch. Toxicol.* **66** (1992) no. 3, 157 – 163.
11. J. E. Dickens, W. M. Irvine, M. Ohishi, M. Ikeda, S. Ishikawa, A. Nummelin, A. Hjalmarson, *Astrophys. J.* **489** (1997) 753 – 757.
12. B. Marczynski, W. Marek, X. Baur, *Medical Hypotheses* **44** (1995) 97 – 100.
13. A. S. Pell, G. Pichler, *Trans. Faraday Soc.* **61** (1965) 71 – 77.
14. C. J. Walters, J. M. Smith, *Chem. Eng. Prog.* **48** (1952) no. 7, 337 – 343.
15. W. Gestrich, W. Kraus, *Ber. Bunsenges. Phys. Chem.* **78** (1974) no. 12, 1334 – 1353.
16. D. N. Glew, N. S. Rath, *J. Chem. Phys.* **44** (1965) 1710 – 1711.
17. O. Maass, E. H. Boomer, *J. Am. Chem. Soc.* **44** (1922) 1709.

18. R. K. McMullan, G. A. Jeffrey, *J. Chem. Phys.* **42** (1965) 2725.
19. D. N. Glew, N. S. Rath, *Can. J. Chem.* **73** (1995) 788 – 796.
20. H. H. Richardson, P. J. Wooldridge, J. P. Devlin, *J. Chem. Phys.* **83** (1985) 4387 – 4394.
21. L. G. Hess, V. V. Tilton, *Ind. Eng. Chem.* **42** (1950) no. 6, 1251 – 1258.
22. J. D. Olson, *J. Chem. Eng. Data* **22** (1977) 326.
23. H. L. Clever, *Solubility Data Ser.* **27 – 28** (1987) 640 – 654.
24. C. L. Yaws: *Physical Properties,* McGraw-Hill, New York 1977, pp. 167 – 176.
25. A. Weissberger in A. Rosowski (ed.): *Heterocyclic Compounds,* **vol. 19,** Wiley-Interscience, New York 1964, pp. 1 – 523.
26. M. S. Malinovski: *Epoxides and their Derivatives,* Daniel Davey & Co, Inc., New York 1965.
27. I. Parker, *Chem. Rev.* **59** (1959) 737 – 797.
28. N. Schönfeld: *Grenzflächenaktive Äthylenoxidaddukte,* Wissensschaftl. Verlags GmbH, Stuttgart 1976, Ergänzungsband 1984.
29. J. Gorzinski Smith, *Synthesis* **8** (1984) 629 – 656.
30. R. R. Baldwin, A. Keen, R. W. Walker, *J. Chem. Soc. Faraday Trans. 1* **80** (1984) no. 2, 435 – 456. A. Lifshnitz, H. Ben Hanou, *J. Chem. Phys.* **87** (1983) 1782 – 1787.
31. D. N. Kirk, *Chem. Ind. (London)* 1973, no. 3, 109 – 116.
32. A. V. Iogansen, B. V. Rassadin, N. P. Sorokina, *Theor. Eksp. Khom.* **24** (1988) 461 – 466.
33. G. P. Ford, C. T. Smith, *Int J. Quantum Chem. , Quantum Bio. Symp.* **13** (1986) 107 – 119.
34. L. Vamling, L. Cider, *Ind. Eng. Chem. Prod. Res. Dev.* **25** (1986) no. 3, 424 – 430.
35. G. Olah, A. P. Fung, D. Meidar, *Synthesis* **4** (1981) 280 – 282.
36. P. Penczek, *Makromol. Chem. Rapid Commun.* **7** (1986) 1 – 14.
37. J. Farakawa, T. Saegusa: *Polymerisation of Aldehydes and Oxides,* Wiley-Interscience, New York 1963. G. Gee, W. C. E. Higginson, *J. Chem. Soc.* 1962, 231.
38. T. H. Baize, *Ind. Eng. Chem.* **53** (1961) 903.
39. N. J. Van Sittert, G. D. J. Beuling, E. W. N. Van Vliet, H. Van Der Waal, *Environ. Health Perspectives* **99** (1993) 217 – 220.
40. G. Hechler, *Chem. Ing. Tech.* **43** (1971) no. 16, 903 – 905.
41. H. Springmann, *Fette Seifen Anstrichm.* **73** (1971) no. 6, 396 – 398.
42. T. Aida, S. Inoue, *J. Am. Chem. Soc.* **105** (1983) no. 5, 1304 – 1309.
43. Halcon, EP-A 133 763, 1984 (V. S. Bhise, H. Gilman).
44. B. J. Ozero, J. V. Procelli, *Hydrocarbon Process.* **63** (1984) no. 3, 55 – 61.
45. F. A. Meskens, *Synthesis* **7** (1981) 501 – 522. D. L. Rakhmankulov, O. B. Chalova, T. K. Kiladze, E. A. Kantor, *Khim. Geterotsikl. Soedin* 1984, no. 3, 291 – 306.
46. Hoechst, DE 1 035 635, 1958 (K. Fischer, K. Vester).
47. R. F. Goldstein: *The Petroleum Chemicals Ind.,* 2nd ed., Wiley Interscience, New York 1958, p. 352.
48. G. Prauser, G. Fischer, K. Dialer, *Angew. Chem.* **92** (1980) no. 5, 389 – 390.
49. Nippon Shokubai Kagaku, JP 56 100 – 633, 1980 (Kishimoto Senji, Nakai Takahiko, Kumarawa, Toshihiko).
50. N. G. Gaylord, E. I. Becker, *Chem. Rev.* **49** (1951) 413. I. Mukerji, A. L. Wayda, G. Darbagh, S. H. Bertz, *Angew. Chem.* **98** (1986) no. 8, 756 – 757.
51. H. D. Verkruisse, L. Brandsma, *Synth. Commun.* **21** (1991) 235 – 237.
52. Celanese, EP 257 967, 1988 (M. A. Murphay, A. Aguilo, B. L. Smith).
53. J. Dale, G. Borgen, K. Daatsvatn, *Acta Chem. Scand. Ser. B* **B 28** (1974) 378 – 379.
54. J. Dale, *J. Chem. Soc. Chem. Commun.* 1976, 295 – 296.
55. Hoechst, DE 2 640 505, 1976 (G. Scheffel, R. Obermeier).
56. H. Kicheldorf, G. Mörber, W. Regel, *Synthesis* **5** (1981) 383.
57. G. A. Viera, L. L. Simpson, B. C. Ream, *Chem. Ing. Prog.* **89** (1993) no. 8, 66 – 75.
58. *Ullmann,* 3rd ed., **3,** 145.
59. A. Wurtz, *Ann. Chim (Paris)* **69** (1863) 355.
60. W. Bone, R. Wheeler, *J. Chem. Soc.* **85** (1904) 1637. L. Reyerson, L. J. Swearingen, *J. Am. Chem. Soc.* **50** (1928) 2872. R. Willstätter, M. Bommer, *Justus Liebigs Ann. Chem.* **422** (1921) 136.
61. S. A. Miller: *Ethylene and its Industrial Derivatives,* Ernest Benn Limited, London 1969, pp. 546 – 548.
62. R. Wolf, H. Götze, *Chem. Tech. (Leipzig)* **14** (1962) no. 10, 600 – 606. H. Kanoh, T. Nishimura, A. Ayame, *J. Catal.* **57** (1979) 372 – 379.
63. J. K. Lee, X. E. Verykios, R. Pitchai, *Appl. Catal.* **44** (1988) 1 – 2, 223 – 237.

64. K.-H. Jung, K.-H. Chung, M.-Y. Kim, J.-H. Kim, G. Seo, *Korean J. Chem. Eng.* **16** (1999) 396 – 400.
65. Shell, DE-OS 2 159 346, 1971 (R. P. Nielsen).
66. Shell, US 4 012 425, 1972 (R. P. Nielsen, J. H. La Rochelle). Shell, EP 0 266 015, 1987 (A. M. Lauritzen). UCC, EP 75 938, 1982 (M. M. Bashin, G. W. Warner).
67. Shell, EP 0 266 015, 1987 (A. M. Lauritzen).
68. CC, US 2 279 470, 1942 (G. H. Law, H. C. Chitwood). J. M. Berty, *Appl. Ind. Catal.* **1** (1983) 223 – 227.
69. G. L. Montrasi, G. C. Battiston, *Oxid. Commun.* **3** (1983) no. 3 – 4, 259 – 267.
70. J. M. Berty, *Chem. Eng. Commun.* **82** (1989) 223 – 227.
71. Bundesministerium für Forsch. und Technologie, *Forschungsbericht T 84 – 225,* Fachinformationszentrum Energie, Physik, Mathematik GmbH, Karlsruhe, Leopoldshafen 2, 7514 Eggenstein 1984, pp. 171 –241. J. M. Berty, *Appl. Ind. Catal.* **1** (1983) 223 – 227.
72. J. R. Monnier, *Stud. Surf. Sci. Catal.* **110** (1997) 135 – 149.
73. G. Leofontani, G. R. Tauszik, M. Padovan, *J. Therm. Anal.* **30** (1985) no. 6, 1267 – 1272.
74. G. L. Montrasi, G. R. Trauszik, M. Solari, G. Leofanti, *Appl. Catal.* **5** (1983) 359.
75. W. D. Mroß, *Ber. Bunsenges. Phys. Chem.* **88** (1984) 1042 – 1053.
76. S. Rebsdat, S. Mayer, J. Alfranseder, *Chem. Ing. Tech.* **53** (1981) no. 11, 850 – 854.
77. I. E. Wachs, *J. Catal.* **72** (1981) no. 1, 160 – 165.
78. W. M. H. Sachtler, C. Backx, R. A. Van Santen, *Catal. Rev. Sci. Eng.* **23** (1981) 127 – 149.
79. H. H. Voge, C. R. Adams, *Adv. Catal.* **17** (1967) 151. P. A. Kilty, W. M. H. Sachtler, *Catal. Rev. Sci. Eng.* **10** (1974) no. 1, 1 – 16.
80. M. Akimoto, K. Ichikawa, E. Echigoya, *J. Catal.* **76** (1982) 333 – 344. N. W. Cant, W. K. Hall, *J. Catal.* **52** (1978) 81.
81. C. T. Campbell, B. E. Koel, *J. Catal.* **92** (1984) no. 2, 272 – 283.
82. E. L. Force, A. P. Bell, *J. Catal.* **38** (1975) 440.
83. C. Backx, J. Moolhuysen, J. Geenen, R. A. van Santen, *J. Catal.* **72** (1981) 364.
84. R. F. Grant, R. M. Lambert, *J. Catal.* **92** (1985) 364 – 375; *J. Catal.* **93** (1985) 92 – 99.
85. R. A. van Santen, C. P. M. de Groot, *J. Catal.* **98** (1986) no. 2, 530 – 539.
86. R. B. Grant, R. M. Lambert, *Stud. Surf. Sci. Catal.* **19** (1984) 251 – 257. S. A. Tan, B. R. Grant, R. M. Lambert, *J. Catal.* **100** (1986) no. 2, 383 – 391.
87. S. A. Tan, R. B. Grant, R. M. Lambert, *J. Catal.* **104** (1987) 156 – 163.
88. Y. S. Yong, N. W. Cant, *Appl. Catal.* **48** (1989) 37 – 50.
89. N. W. Cant, E. M. Cennedy, N. J. Ossipoff, *Catal. Lett.* **141** (1993) no. 1-2, 133 – 134.
90. J. T. Roberts, J. R. Madix, W. W. Crew, *J. Catal.* **9** (1991) 300 – 307.
91. E. A. Carter, W. A. Goddard, *J. Catal.* **112** (1988) 80 – 92.
92. H. Kobayashi, K. Nakashiro, T. Iwakuwa, *Theor. Chem. Acc.* **102** (1999) 1 – 6, 237 – 243.
93. M. Kobayachi, T. Kanno, *Actas Simp. Iberoam. Catal.,* **9th** (1984) no. 2, 878 – 887. T. Kanno, M. Kobayashi, *Proc. Int. Congr. Catal.* **8th** (1984) pub. 1985, no. 3, III 277 – III 288.
94. J. K. Plischke, A. J. Benesi, M. A. Vannice, *J. Catal.* **138** (1992) 223 – 239.
95. P. D. Klugherz, P. Harriott, *J. AIChE* **17** (1971) 856.
96. R. A. von Santen, H. P. C. E. Kuipers, *Adv. Catal.* **35** (1987) 265 – 321.
97. E. P. S. Schouten, P. C. Borman, K. R. Westerp, *Chem. Eng. Process.* **35** (1996) 107 – 120.
98. K. Dialer, *Chem. Ing. Tech.* **54** (1982) no. 1, 18.
99. T. E. Corrigan, *Pet. Refiner* **32** (1933) no. 2, 87.
100. N. Piccinini, G. Levy, *Can. J. Chem. Eng.* **62** (1984) no. 4, 541 – 546, 547 – 558.
101. K. Klose, P. Kripylo, L. Mögling, *Chem. Techn. (Leipzig)* **37** (1985) no. 12, 506 – 508.
102. I. E. Wachs, C. C. Chersich, *J. Catal.* **72** (1981) 160 – 165.
103. Halcon, DE-OS 2 750 601, 1977 (B. J. Ozero); Halcon, US 3 964 980, 1975 (B. J. Ozero).
104. PPG Industries, US 3 904 656, 1973 (S. E. Broz).
105. Berufsgenossenschaft der chem. Industrie, Heidelberg, *Merkblatt M 045 7/85 ZH 1/54,* Jedermann-Verlag, Heidelberg 1985.
106. Ethylene Oxide Users Guide, 2nd ed., Celanese, Dow, Shell Chemicals, Sunoco Chemicals, Equistar, August 1999, http://www.ethyleneoxide.com
107. A. K. Nikitina, L. P. Vershinina, V. P. Khvatkova, V. M. Umnyashkina, *Khim. Tekhnol. Topl. Masel* **11** (1982) 27 – 28.
108. Halcon, DE 2 718 056, 1978 (A. N. Naglieri). S. E. Diamond, F. Mares, A. Szalkiewicz, D. A. Muccigrosso, J. P. Solar, *J. Am. Chem. Soc.* **104** (1982) no. 15, 4266 – 68.
109. M. Stoukides, G. G. Vayenas, *J. Catal.* **70** (1981) no. 1, 137 – 146.

110. Exxon, DE-OS 2 915 108, 1979 (C. T. Hou, R. N. Patel, N. J. Edison, A. I. Laskin). C. G. Van Ginkel, H. G. J. Welten, J. A. M. De Bont, *Eur. J. Appl. Microbiol. Biotechnol.* **24** (1986) no. 4, 334 – 337.
111. J. Kollar, *Chemtech* **14** (1984) no. 8, 504 – 511.
112. A. Aquilo, R. J. H. Vorhoeeve, D. N. Freeman, S. Alder, *Hydrocarbon Process.* **62** (1983) no. 3, 57 – 65.
113. B. D. Dombeck, *Adv. Catal.* **32** (1983) 325 – 416.
114. A. M. Brownstein, *Hydrocarbon Process.* **71** (1975) no. 9, 72 – 76.
115. D. E. Field: *Engineering and Cost Study of Air Pollution Control for the Petrochemical Industry,* **vol. 6,** EPA-450/3-73-006-f. EPA-450/4-84-007L, US Environmental Protection Agency Off. Air Qual. Plann. Stand. (Tech. Rep.), Research Triangle Park, N.C. 27711 1986.
116. Shell, US 3 867 113, 1975 (E. G. Foster, P. F. Russel, R. G. Vanderwater).
117. G. Groß, H. Kittner, *Gas Aktuell* **30** (1985) 7 – 13.
118. P. Oberle, *Chem. Anlagen + Verfahren* **1** (1983) 12 – 13.
119. Buonicore-Cashman Ass., PCT Int. Appl., WO 8 400 748, 1984, US-A 409 830, 1982 (A. J. Buonicore).
120. Leybold-Heraeus GmbH, EP-A 130 319, 1984 (H. Amlingek).
121. U. Müller, *Chem.-Ing.-Techn.* **62** (1990) 197 – 198.
122. O. Michel, R. Wiedemeyer, *Chem. Ing. Tech.* **49** (1977) no. 1, 57.
123. R. A. Conway, G. T. Waggy, M. H. Spiegel, R. L. Berglund, *Environ. Sci. Technol.* **17** (1983) 107 – 112.
124. A. M. Winer, R. Atkinson, J. Arey, S. M. Aschman, M. A. Goodman, PB88 – 215876, 165pp NTIS, Gov. Reap. Announce Index (U. S.) 88 (18) Abstr. No. 846,606 (1988).
125. D. Pierotti, S. Havlicek, L. Hilpert, P. Kelly, *Lab 2000* **5** (1991) 74 – 78.
126. Erste allgemeine Verwaltungsvorschrift zum Bundesimmissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft -TA-Luft) vom 27. 02. 1986. Gemeinsames Ministerialblatt 37. Jahrg., Nr. 7 vom 28. 02. 1986.
127. R. L. Anderson in F. D. Snell, L. S. Ettre (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 12,** J. Wiley & Sons, New York 1971, pp. 317 – 340.
128. F. W. Kerckow, *Z. Anal. Chem.* **108** (1937) 249.
129. A. H. Qazi, N. H. Ketcham, *Am. Ind. Hyg. Assoc. J.* **38** (1977) 635. Merkblatt ZH1/120.27, Hauptverband der gewerblichen Berufsgenossenschaften, Carl Heymanns Verlag, Köln 1985.
130. R. Binetti, S. Di Marzio, P. Di Prospero, L. Gramiccioni, G. Viviano, G. Ziemacki, *Chromatografia* **21** (1986) 701 – 704.
131. G. Greff, J. Delcourt, J. P. Guenier, B. Herve-Bazin, J. Muller, *Chromatografia* **21** (1986) 201 – 204.
132. C. Levevre, P. Ferrari, J. P. Guenier, J. Muller, *Chromatografia* **21** (1986) 269 – 273.
133. K. Cummins, G. E. Burroughs, J. Trembley, *Appl. Occup. Environ. Hyg.* **8** (1993) 463 – 471.
134. *NIOSH Manual of Analytical Methods,* DHHS (NIOSH) Publ. No. 84 – 100, U.S. Government Printing Office, Washington, D.C. 20 402, Method 1607, 5/15/85 pp. 1607-1 – 1607-6.
135. National Technical Information Service, Springfield, Virginia 22 151, *A Review of Violent Monomer Polymerisation,* Operations Research, Inc., prepared for the Coastguard, Oct. 1974.
136. R. Davies, AIChE Loss Prevention Symposium, Atlanta, Georgia, Feb. 27– Mar. 1, 1978.
137. Ethylene Oxide Product Guide, Union Carbide Corporation (UCC), Houston 2000.
138. *Codes of Practice for Chemicals with Major Hazards, Ethylene Oxide,* Brochure CPP4, The Chemical Industry Safety and Health Council of the Chemical Industries Assoc., London 1975.
139. *Properties and Essential Information for Safe Handling and Use of Ethylene Oxide,* Chemical Safety Data Sheet SD-38, Manufacturing Chemists Association, Washington, D.C., 1971.
140. P. Cardillo, N. Mazzei, A. Lunghi, *Riv. Combust.* **52** (1998) no. 3, 115 – 141.
141. L. G. Britton, *Plant/Operation Progress* **9** (1990) no. 2, 75 – 86.
142. E. Peytral, *Bull. Soc. Chim. Fr.* **39** (1926) 206 – 14.
143. J. Osugi, M. Okusima, M. Hamanoue, *Koatsu Gasu* **8** (1971) no. 4, 201 – 206.
144. S. N. Bajpai, *Chem. Eng. Prog. Loss Prev. Technical Manual* **13** (1980) 119 – 122.
145. B. Pesetsky, *Chem. Eng. Prog. Loss. Prev.* **13** (1980) 132 – 141.

146. W. H. Perkin, *J. Chem. Soc.* **63** (1893) 488.
147. F. A. Burden, J. H. Burgoyne, *Proc. R. Soc. London* **199** (1949) 328 – 351.
148. D. Conrad, *Bundesgesundheitsblatt* **9** (1963) 139 –141.
149. G. W. Jones, R. E. Kennedy, *Ind. Eng. Chem.* **22** (1930) no. 2, 146 – 147.
150. E. G. Plett, *Plant/Oper. Prog.* **3** (1984) 190 – 193.
151. A. Fiumara, N. Mazzei, *Chim. Ind. (Milan)* **65** (1983) no. 11, 683 – 687.
152. Y. Hashigushi, *Tokyo Kogyo Shikensho Hokoku* **60** (1965) no. 3, 85 – 91.
153. J. H. Burgoyne, K. E. Bett, *Inst. Chem. Eng. Symp. Ser.* **25** (1968) 1 – 7.
154. J. H. Burgoyne, K. E. Bett, R. Muir in J. M. Piric (ed.): *Symposium on Chem. Proc. Hazards (Inst. Chem. Eng.)* (1960), pp. 30 – 36.
155. E. O. Haenni, N. A. Affens, H. G. Lento, A. H. Yeomans, R. A. Fulton, *Ind. Eng. Chem.* **51** (1959) no. 5, 685 – 688.
156. J. N. Cawse, B. Pesetsky, W. T. Vyn, AIChE Loss Prevention Symposium, Houston, 1979.
157. L. D. Chen, G. M. Faeth, *Combustn. Flame* **40** (1981) 13 – 28.
158. A. K. Gupta, *J. Soc. Chem. Ind.* **68** (1949) 179 – 183.
159. H. Grosse-Wortmann, *Chem. Ing. Tech.* **40** (1968) no. 14, 689 – 692.
160. G. A. Viera, L. L. Simpson, B. C. Ream, *Chem. Ing. Progress* **89** (1993) no. 8, 66 – 75.
161. C. J. Hilado, F. H. Small, *J. Fire Flammability* **5** (1974) no. 6, 15.
162. L. Britton, *Plant/Operat. Prog.* **10** (1991) no. 1, 27.
163. B. E. Mellin, *A.I.Ch.E. Loss Prev. Bull.* **100** (1992) 1 – 12.
164. H. Köstner, *TU* **32** (1991) no. 9, 310 – 314.
165. B. Besling, U. Löffler, A. Polt, *Chem.-Ing.-Tech.* **67** (1995) 1614 – 1618.
166. U. Löffler, *Prax. Sicherheitstech.* **4** (1997) 31 – 40.
167. A. K. Gupta, *J. Soc. Chem. Ind.* **68** (1949) 179 – 183.
168. R. G. Vanderater, *Chem. Ing. Prog.* **12** (1989) 16 – 20.
169. H. Grosse-Wortmann, *Chem. Ing. Tech.* **40** (1968) 689 – 692.
170. Gefahrstoffverordnung GefStoffV, August 26, 1986 (BGBl. I S. 1470), Carl Heymanns Verlag, Köln.
171. *Sichere Chemiearbeit* **39** (1987) 11.
172. OSHA Regulation 29 CFR 1910.1047.
173. Merkblatt Glykole, Hoechst AG 1979; Merkblatt Genantin, Hoechst AG 1982; Merkblatt Antifrogen L, Hoechst AG 1985, Frankfurt am Main.
174. Merkblatt Diethylenglykol, Hoechst AG 1986; Merkblatt Hoechst 1704LTV, Hoechst AG 1985;Merkblatt Hoechst 1678, Hoechst AG 1985.
175. Merkblatt Triethylenglykol, Hoechst AG 1984.
176. Merkblatt Polyglykole Hoechst, Eigenschaften und Anwendungsgebiete, Hoechst AG 1986.
177. Merkblatt Ethylenglykoldimethylether, Hoechst AG 1986; Merkblatt Lösemittel Hoechst, Hoechst AG 1984.
178. *Chem. Week* **134** (1984) March 7, 32 – 35.
179. A. M. Thiess, *Arch. Toxikol.* **20** (1963) 127.
180. H. Encke, Inaug.-Diss., Köln (1936).
181. W. Tilling, *Ärztl. Wochenschr.* **9** (1954) 282.
182. R. S. McLaughlin, *Am. J. Ophthalmol.* **29** (1946) 1355.
183. E. Metz, *Ärztl. Sachverständ.-Ztg.* **44** (1938) 155.
184. P. D. Blandin, *Berufsdermatosen* **6** (1958) 217.
185. J. L. Shupack, S. R. Andersen, S. J. Romano, *J. Lab. Clin. Med.* **98** (1981) 723.
186. P. F. Finelli, Th. F. Morgan, J. Yaar, C. V. Granger, *Arch. Neurol. Chicago* **40** (1983) 419.
187. J. A. Gross, M. L. Haas, T. R. Swift, *Neurology* **29** (1979) 978.
188. V. F. Gray, J. Hozier, D. Jacobs, R. L. Wade, D. G. Gray, *Environ. Mutagen.* **1** (1979) 375.
189. H. F. Smyth, J. Seaton, L. Fischer, *J. Ind. Hyg. Toxicol.* **23** (1941) 259.
190. K. H. Jacobson, E. B. Hackley, L. Feinsilver, *Arch. Ind. Hlth.* **13** (1956) 237.
191. F. Koelsch, E. Lederer, *Zentralbl. Gewerbehyg. Unfallverhüt.* **7** (1930) 264.
192. W. M. Snellings, C. S. Weil, R. R. Maronpot, Project Rep. 44-20, Bushy Run Research Center, Pittsburgh, Pa., (1981).
193. R. L. Hollingsworth, V. K. Rowe, F. Oyen, D. D. McCollister, H. C. Spencer, *Arch. Ind. Hlth.* **13** (1956) 217.
194. W. M. Snellings, R. R. Maronpot, J. P. Zelenak, C. P Laffoon, *Toxicol. Appl. Pharmacol.* **64** (1982) 476.
195. P. L. Hackett et al., *"Teratogenic study of ethylene and propylene oxide and n-butylacetate,"* Study performed at Battelle. NIOSH contract no. 210-80-0013 (1982).
196. L. A. Rapoport, *Dokl. Akad. Nauk SSSR* **60** (1984) 469.
197. G. A. Sega, R. B. Cumming, J. G. Owens, C. Y. Horton, L. A. Lewis, *Environ. Mutagen.* **3** (1981) 371.

198. G. Kölmark, M. Westergaard, *Hereditas (Lund, Swed.)* **39** (1953) 209.
199. U. Rannug, R. Göthe, C. A. Wachtmeister, *Chem.-Biol. Interact.* **12** (1976) 251.
200. E. H. Pfeiffer, H. Dunkelberg, *Food Cosmet. Toxicol.* **18** (1980) 115.
201. J. W. Embree, J. P. Lyon, C. H. Hine, *Toxicol. Appl. Pharmacol.* **33** (1975) 172.
202. D. Jenssen, C. Ramel, *Mutat. Res.* **75** (1980) 191.
203. W. M. Generoso, K. T. Cain, M. Krishna, C. W. Sheu, R. M. Gryder, *Mutat. Res.* **73** (1980) 133.
204. L.-E. Appelgren, G. Eneroth, C. Grant, L.-E. Landström, K. Tenghagen, *Acta Pharmacol. Toxicol.* **43** (1978) 69.
205. V. Poirier, D. Papadopoulo, *Mutat. Res.* **104** (1982) 255.
206. D. W. Lynch, T. R. Lewis, W. J. Moormann, P. S. Sabharwal, J. A. Burg, *Toxicologist* **3** (1983) 60.
207. A. D. Kligerman, G. L. Erexon, M. E. Phelps, J. L. Wilmer, *Mutat. Res.* **120** (1983) 37.
208. J. W. Yager, C. J. Hines, R. C. Spear, *Science* **219** (1983) 1221.
209. C. Laurent, J. Fréderic, F. Maréchal: "Etude des effets cytogénétiques d'intoxication á l'oxyde d'éthyléne," *C. R. Séances Soc. Biol.* **176** (1982) no. 5, 733.
210. C. Laurent, J. Fréderic, A. Y. Léonard: "Sister chromatic exchange frequency in workers exposed to high levels of ethylene oxide in hospital sterilization service," *Int. Arch. Occup. Environ. Health*(1983), cited in [221]
211. C. Laurent, J. Fréderic, F. Maréchal: "Augmentation du taux d'échanges entre chromatides-soeurs chez des personnes exposées professionellement a l'oxyde d'éthyléne," *Ann. Genet.* (1983), cited in [221]
212. B. L. Van Duuren, L. Orris, N. Nelson, *J. Natl. Cancer Inst.* **31** (1965) 707.
213. H. Dunkelberg, *Br. J. Cancer* **39** (1979) 588.
214. H. Dunkelberg, *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Hyg.* **174** (1981) 384.
215. H. Dunkelberg, *Br. J. Cancer* **46** (1982) 924.
216. W. M. Snellings, C. S. Weil, R. R. Maronpot, *Toxicol. Appl. Pharmacol.* **75** (1984) 105.
217. C. Hogstedt, N. Malmqvist, B. Wadman, *JAMA J. Am. Med. Assoc.* **241** (1979) 1132.
218. C. Hogstedt, O. Rohlen, B. S. Berndtsson, O. Axelson, L. Ehrenberg, *Br. J. Ind. Med.* **36** (1979) 276.
219. R. W. Morgan, K. W. Claxton, B. J. Divine, S. D. Kaplan, V. B. Harris, *JOM J. Occup. Med.* **23** (1981) 767.
220. C. Hogstedt, L. Aringer, A. Gustavsson, *JAMA J. Am. Med. Assoc.* **255** (1986) 1578.
221. ECETOC Technical Report No. 11: *"Ethylene oxide toxicology and its relevance to man,"* European Chemical Industry Ecology and Toxicology Centre, Avenue Louise 250, Bte. 63, B-1050 Brussels (1984).
222. DFG *"MAK- und BAT-Werte-Liste 2001"* Wiley-VCH, Weinheim.

Ethylene Polymers → Polyolefins

2-Ethylhexanol

HELMUT BAHRMANN, Ruhrchemie Aktiengesellschaft, Oberhausen, Federal Republic of Germany (Chaps. 1–5)

HEINZ-DIETER HAHN, Ruhrchemie Aktiengesellschaft, Oberhausen, Federal Republic of Germany (Chaps. 1–5)

DIETER MAYER, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 6)

1. Introduction 1
2. Properties 1
3. Industrial Production 1
4. Uses 4
5. Economic Aspects 4
6. Toxicology 5
7. References 5

1. Introduction

2-Ethylhexanol (2-ethyl-1-hexanol) [*104-76-7*], $C_8H_{18}O$, M_r130.23, ranks after the lighter alcohols (methanol to butanol) as the most important synthetic alcohol (see also → Alcohols, Aliphatic) [1–3]. Approximately 2×10^6 t/a are produced worldwide. 2-Ethylhexanol is mainly used as the alcohol component for the manufacture of ester plasticizers for soft poly(vinyl chloride) (PVC) and has been produced for this purpose since the mid-1930s.

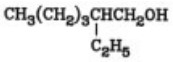

2-Ethylhexanol

The raw material for the synthesis of 2-ethylhexanol, namely butyraldehyde (butanal) [*123-72-8*], is made almost exclusively from the petrochemical feedstock propene via the oxo synthesis (→ Butanals). The route from ethylene via acetaldehyde is insignificant. This situation could change if acetaldehyde produced from ethanol (via fermentation or homologation of methanol) and coal-based synthesis gas (which is already used for the oxo synthesis) becomes available at a reasonable price.

Acetaldehyde was formerly produced from coal via acetylene. This synthesis route had no chance of competing successfully against the route based on petrochemical raw materials because of the large number of steps involved.

2. Properties

Physical Properties. 2-Ethylhexanol is a clear liquid with a characteristic odor and forms a transparent mixture with other alcohols, ethers, and most organic liquids. Its main physical properties are summarized in Tables 1 and 2.

Chemical Properties. 2-Ethylhexanol reacts in the typical manner of α-branched primary alcohols (→ Alcohols, Aliphatic, Chap. 2.2. and → Alcohols, Aliphatic, Chap. 2.3.).

3. Industrial Production

2-Ethylhexanol is produced in four steps:

1) Aldolization of butyraldehyde and subsequent dehydration
2) Separation of the aldolization solution
3) Hydrogenation of unsaturated 2-ethyl-2-hexenal as an intermediate product
4) Fractionation of 2-ethylhexanol

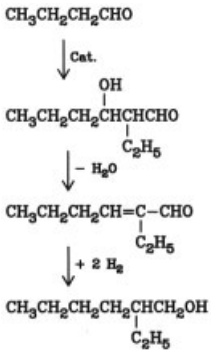

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_137

Table 1. Physical properties of 2-ethylhexanol [4]

Property	Value
Viscosity at 20 °C	9.8 mPa·s
Vapor pressure at 20 °C	ca. 0.03 kPa
Heat of vaporization at 184.8 °C	50.66 kJ/mol
Solubility in water at 20 °C	0.07 wt %
Solubility of water in 2-ethylhexanol at 20 °C	2.7 wt %
Surface tension at 20 °C	0.25×10^{-3} N/cm
Dielectric constant at 20 °C	7.7
bp at 101.3 kPa	184.6 °C
Critical temperature	339.8 °C
Critical pressure	2.76 MPa
Critical density	0.2636 g/cm^3
Critical compressibility	0.2670
Specific flow resistance at 20 °C	5.8×10^3 MΩ · cm
Azeotrope with water at 101.3 kPa, 20 % 2-ethylhexanol/ 80 % water, *bp*	183.5 °C

Table 2. Temperature dependence of some physical properties of 2-ethylhexanol

Property	*t*, °C	Value
Vapor pressure, kPa	78.7	1.333
	90.8	2.666
	104.6	5.332
	113.5	7.998
	125.5	13.33
	143.5	26.66
	164.2	53.32
	184.8	101.31
Relative density, $d^t{}_4$	10	0.8396
	15	0.8359
	20	0.8323
	25	0.8286
	30	0.8247
	50	0.8100
Refractive index, $n^t{}_D$	10	1.4356
	15	1.4337
	20	1.4317
	25	1.4298
	30	1.4278
	50	1.4199
Specific heat, c_p, J g^{-1} K^{-1}	0	2.22
	20	2.34
	50	2.51
	80	2.68

The butyraldehyde required for the aldolization stage is made by hydroformylation of propene. In addition to processes using cobalt as a catalyst [5–8], processes that employ rhodium catalysts [9–12] have gained increasing significance since the 1980s.

The demands placed on the purity of 2-ethylhexanol are so high that the purity of the starting product, butyraldehyde, is also extremely important. Isobutyraldehyde is formed to a greater (cobalt catalysts) or lesser (rhodium cata-lysts) extent during hydroformylation and must be separated to prevent mixed aldolization. Therefore, special processes in which aldolization and, in some cases, hydrogenation occur in parallel with hydroformylation are becoming less important [5], [13], [14].

Older processes based on acetylene that involve the intermediate stages acetaldehyde/crotonaldehyde would permit access not only to coal as a raw material, but also to petrochemistry via ethylene [15], [16] as well as to regenerable raw materials via ethanol. However, these processes are not currently important.

The major oxo producers [17–19] all offer processes for the manufacture of 2-ethylhexanol that differ only in certain details. Figure 1 shows the flow diagram for a 2-ethylhexanol plant.

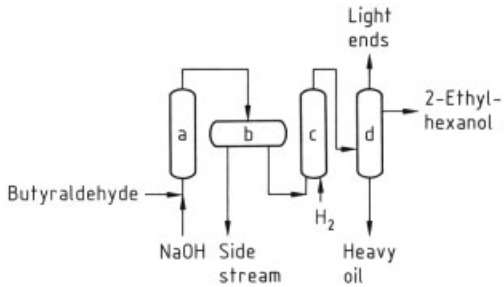

Figure 1. 2-Ethylhexanol plant
a) Aldolization; b) Phase separator; c) Hydrogenation; d) Distillation

In the aldolization reactor, butyraldehyde reacts very quickly to give 2-ethyl-2-hexenal, aqueous sodium hydroxide being employed as the standard industrial catalyst. Local overheating in the reaction mixture must be avoided, since this may cause secondary reactions and thus decrease yields. Thorough mixing of the two-phase system is necessary. The primary aldol addition must take place rapidly; the immediately ensuing dehydration of the hydroxyalde-

hyde must be conducted as quickly and quantitatively as possible because the aldol is unstable and can impair the product quality and yield.

The aldolization reactor may be a mixing pump [20], a packed column [7], or a stirring vessel [21], [22].

The various processes operate at a temperature of between 80 and 150 °C and pressures below 0.5 MPa. The ratio of aldehyde to aqueous sodium hydroxide solution is 1 : 10 – 1 : 20. Under these conditions, conversion rates >99 % are obtained. The heat of the aldolization reaction is used for steam generation.

After the aldolization stage in (a) (see Fig. 1), the mixture is separated in a phase separator (b) into an upper organic phase and a lower aqueous phase containing the aldolization solution. Part of the aldolization solution can be recycled, but the rest must be removed from the system via a side stream because the aldolization solution is diluted by the water that is produced in the reaction; the aldolization byproducts (Cannizzaro and Tischchenko reactions) must also be removed. Sodium hydroxide is added to maintain its concentration at 2 – 4 %.

The side stream has such a high COD value (e.g., due to its sodium butyrate content) that it should be subjected to preliminary purification before being conveyed to a biological treatment plant. Suitable purification methods are oxidation, acid treatment, and extraction which allows partial recycling of valuable products.

The organic product from the phase separator (b) can either be hydrogenated in a single stage on a fixed nickel catalyst [9] or in several stages either in a combination of gas – liquid phases [23] or liquid – liquid phases (sump-phase or trickle-bed reactor). The heat of reaction for the hydrogenation of the C = C double bond and the aldehyde group is relatively high, 178 kJ/mol. Temperature control problems may therefore arise as local overheating decreases yields and must be avoided.

With single-step hydrogenation, remixing with the hydrogenation product has been proposed to dissipate heat (150 – 200 °C) [7]; in contrast to other processes, medium pressure is initially necessary to ensure adequate conversion. Modern plants normally utilize two stages to remove residual amounts of carbonyl compounds and to ensure that high-grade 2-ethylhexanol is obtained [2]. Nickel [7], [24], [25], copper [26] or mixed systems [23], [27–30] are preferred as heterogeneous hydrogenation catalysts. The heat of the hydrogenation reaction is also used for steam generation. In the hydrogenation stage, a conversion of 100 % and a selectivity of >99 % is attained.

Fractional distillation of the hydrogenation product from (c) in (d) normally takes place in three stages: In the first stage, the light ends are separated at the head and can be employed for the manufacture of 1-butanol. In the second stage, pure 2-ethylhexanol is collected at the head. In the third stage, the recyclable intermediate fractions are separated from the heavy oil. The byproducts may be used for heating purposes.

Trade Names. 2-Ethylhexanol is usually marketed by the manufacturers (e.g., Ruhrchemie, BASF, Hüls, Union Carbide, Mitsubishi Chemical) under its chemical designation 2-ethyl-1-hexanol or as octyl alcohol. Typical *specifications* have been compiled in Table 3.

Table 3. Quality specifications for 2-ethylhexanol [3], [32]

Specification	Value
2-Ethylhexanol	≥99.5 wt %
Density at 20 °C	0.832 – 0.833 kg/L
Refractive index, n_D^{20}	1.431 – 1.432
Color, Pt – Co scale	max. 5
Distillation range (95 vol %) at 101.3 kPa	184 – 185 °C
Water content	max. 0.05 wt %
Pour point	< −60 °C

Storage and Transportation. Safety precautions for the shipment and storage of 2-ethylhexanol are determined by its combustibility, flash point, and ignition temperature (see Table 4). Dry 2-ethylhexanol does not corrode standard metals and, like other alcohols, can be stored in standard steel containers (e.g., ST 37). For particularly high-quality demands, use of stainless steel or aluminum containers is advisable. To prevent the entry of atmospheric oxygen or moisture, an inert gas cover should be provided. 2-Ethylhexanol can be transported in iron drums, tank trucks, or railway cars made of standard steel, aluminum, or stainless steel. Centrifugal pumps conforming to chemical standards with simple rotating mechanical seals and polytetrafluoroethylene (PTFE) flange

seals have proved to be suitable for use with 2-ethylhexanol.

Table 4. Safety information for 2-ethylhexanol

Safety parameter	Value
Flash point, closed cup	75 °C
Ignition temperature	270 – 305 °C
Lower and upper explosive limits in air at 101.3 kPa	1.1 – 7.4 vol %
Hazard classification for transport	
CFR 49: 172.101, flammable liquid	
RID/ADR class 3, number 32 c	

4. Uses

In the United States, 2-ethylhexanol is mainly used in the form of its secondary products for the following applications [1]:

Plasticizers	70 %
2-Ethylhexanol acrylate	15 %
Cetane improvers and lubricant additives	9 %
Surfactants	3 %
Miscellaneous	3 %

Worldwide, 2-ethylhexanol esters have >60 % share of the poly(vinyl chloride) (PVC) plasticizers market. In the United States, bis(2-ethylhexyl) phthalate [*117-81-7*], di-2-ethylhexyl phthalate (DEHP), also called dioctyl phthalate (DOP), dominates (70 %); this is followed by bis(2-ethylhexyl) adipate [*103-23-1*], di-2-ethylhexyl adipate (DEHA), also called dioctyl adipate (DOA), (16 %), and tris(2-ethylhexyl) mellitate, trioctyl trimellitate (12 %). In Europe, DEHP's share of the market is even higher. The outstanding importance of DEHP is due to its favorable property combination as a plasticizer (good gelling properties, low volatility, high resistance to heat and water, and excellent electrical properties) (→ Plasticizers). Comprehensive investigations have been conducted on the toxicology and ecology of the widely used chemical DEHP during the last few years. The results have shown that its health and ecotoxicological risks are extremely low [33], [34].

The second large field of application is the production of 2-ethylhexyl acrylate [*103-11-7*], which is used to manufacture coating materials (especially emulsion paints), adhesives, printing inks, impregnating agents, and reactive diluent/cross-linking agents. 2-Ethylhexyl nitrate [*27247-96-7*] serves as a cetane number improver [35]. 2-Ethylhexyl phosphates are used as lubricating oil additives [36]. Antifoaming agents (cf. [37], [38]), dispersants (cf. [39]), and flotation agents [40] made from 2-ethylhexanol deserve special mention as surfactants. Other applications of 2-ethylhexanol include its use as a solvent, above all for polymerization catalysts (cf. [41]) and in extracting agents (cf. [42]).

5. Economic Aspects

Low production costs for butyraldehyde in modern large-scale plants and the lower price of propene as compared with ethylene mean that the oxo route for the manufacture of 2-ethylhexanol will continue to be given preference for economic reasons. As the 2-ethylhexanol manufacturers have their own oxo and hydrogenation capacities, the 1-butanol formed as a byproduct via the hydration of butyraldehyde is also used. In modern plants, ca. 860 kg of 2-ethylhexanol and ca. 18 kg of 1-butanol are obtained from 1000 kg of butyraldehyde. Other production processes will be able to survive in the long term only in countries where the necessary raw materials are economically available. Table 5 lists worldwide capacities.

Table 5. Worldwide capacities for 2-ethylhexanol

Country/region	Capacity, 10^3 t/a
Western Europe	770
(Federal Republic of Germany)	(550)
Eastern Europe	490
United States	240
Central and South America	150
Asia	480
Total	2130

6. Toxicology

Oral LD_{50} values of 2-ethylhexanol are ca. 2.0 g/kg [43] and 3.7 g/kg in rats [44]. Gross necropsy revealed evidence of gastrointestinal

irritation. In the rabbit eye, the undiluted material produces an erythema, swelling of the conjunctiva, lacrimation, and mucous secretion [45].

After skin contact in rabbits, 2-ethylhexanol is mildly irritating [43] if the patch is not closed; otherwise moderate skin irritation occurs [46]. The lowest published LD_{50} after dermal application is 2 g/kg in rabbits [43].

Mice, rats, and guinea pigs survived inhalation of a nearly saturated atmosphere (227 mL/m^3) over a period of 6 h [43]. Typical signs of intoxication were central nervous depression, dyspnea, and irritation of mucous membranes. Repeated inhalation of a maximum dose of 110 mg/m^3 by rats caused dystrophic changes in parenchymatous organs [47].

The cumulative oral toxicity seems to be low [45]. A total of 12.5 g/kg in the diet fed to rats over a period of 90 d affected the kidneys and the liver [46]. An amount of 2 mL/kg fed to rabbits for 10 d caused signs of local irritation and systemic toxicity [43]. Peroxisomes were induced with liver enlargement when rats were fed with dietary levels of 2 % 2-ethylhexanol for a period of 3 weeks [48].

In rabbits [49] and rats [50], 2-ethylhexanol is conjugated to glucuronic acid and eliminated via the urine, its main metabolite being 2-ethylhexanonylglucuronide. A small amount of the radioactively labeled material is oxidized and expired as CO_2. Only 2 % is excreted unchanged.

Neither the ACGIH, OSHA, nor DFG have set standards for exposure levels.

7. References

General References

1. Chemical marketing reporter, New York 1984, 22. Oct., (Chemical profile).
2. B. Cornils, A. Mullen, *Hydrocarbon Process* **59** (1980) no. 11 93.
3. C. E. O'Rourke, P. R. Kavasmaneck, R. E. Uhl in E. J. Wickson: "Monohydric Alcohols," *ACS Symp. Ser.* **159** (1981) 71.

Specific References

4. J. Falbe, B. Cornils, *Fortschr. Chem. Forsch.* **11** (1968) no. 1, 101.
5. B. Cornils, J. Falbe (eds.): *New Synthesis with CO,* Springer Verlag, Heidelberg – Göttingen – New York 1980.
6. Ruhrchemie AG/Rhône-Poulenc S.A. *Hydrocarbon Process.* **56** (1977) no. 11, 163.
7. G. Dümbgen, D. Neubauer, *Chem. Ing. Tech.* **41** (1969) 974.
8. Farbwerke Hoechst, DE 838 746, 1953 (W. Berndt).
9. J. Matthey: "Low-pressure oxo process yields a better product mix," *Chem. Eng.* **12** (1977) 110.
10. *Chem. Ind. (London)* 1981, 169.
11. K. H. Schmidt, *Chem. Ind.* **108** (1985) 762.
12. Ruhrchemie/Rhône-Poulenc-Verfahren, *Hydrocarbon Process.* **64** (1985) no. 11, 153.
13. Esso, FR 1 324 873, 1962 (C. Roming); US 3 946 082, 1972.
14. Shell Int. Res. Matsch., DE-OS 1 468 603, 1972; DE-OS 1 468 615, 1972 (C. R. Greene).
15. H. Weber, W. Dimmling, A. M. Desai, *Hydrocarbon Process.* **55** (1976) no. 4, 127.
16. Kyowa Hakko Kogyo Co. *Hydrocarbon Process.* **46** (1967) no. 11,153.
17. BASF, *Hydrocarbon Process.* **52** (1973) no. 11, 107.
18. Ruhrchemie, *Hydrocarbon Process,* **64** (1985) no. 11, 140.
19. UCC, *Hydrocarbon Process.* **54** (1975) no. 11, 54, 117.
20. Ruhrchemie, DE 2 437 957, 1976 (G. Kessen, J. Meis).
21. Kuhlmann S.A., DE-OS 1 906 850, 1968 (H. Lemke, R. Duval); DE-OS 1 966 388, 1972.
22. BASF, DE 927 626, 1975 (H. J. Nienburg, E. Nicolai, W. Hagen).
23. G. S. Gurevich, L. A. Shapiro et al., SU 1 084 268, 1984;SU 1 010 052, 1983; *Chem. Abstr.* **101** (1984) 170 697 d; *Chem. Abstr.* **99** (1983) 53 115 s.
24. BASF, DE 1 277 232, 1968 (H. Corr, E. Haarer, H. Hoffmann, S. Winderl).
25. R. Kubicka, J. Veprek et al., CS 207 933, 1983; *Chem. Abstr.* **100** (1984) 191 365 p.
26. Ruhrchemie, DE 2 538 253, 1978 (G. Horn, C. D. Frohning).
27. S. G. Gurevich, L. A. Shapiro et al., SU 692 824, 1979;Chem. Abstr. **92** (1980) 58 222 q.
28. BASF, DE 1 949 296, 1974 (G. Boettger, H. Hoffmann, L. Schuster, H. Toussaint).
29. Chemische Werke Hüls, DE 1 643 856, 1971 (M. Reich).
30. BASF, DE 2 832 699, 1980 (G. Heilen, A. Nissen, W. Koernig et al.).
31. BASF, Data sheet for 2-ethylhexanol, 1984.

32. Hoechst AG, Firmenbroschüre über Alkohole, Hoechst 1984;Hoechst AG, Firmenbroschüre über 2-Ethylhexanol, 1982.
33. ECETOC, Technical Report No. 19, May 1985.
34. Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker, Stoffbericht 4, Jan. 1986.
35. Ethyl Corp., US 4 479 905, 1984 (G. G. Knapp, M. S. Marguerite, P. D. Seemuth),
36. Elco Corp., US 4 508 630, 1985 (F. A. Litt, T. R. Czernicki).
37. Ciba-Geigy, EP 96 806, 1983 (R. Defago, R. Baeuerle).
38. National Distillers and Chem. Corp., US. 4 540 511, 1985 (E. F. McCaffrey, A. J. Dieterman, L. Knazko).
39. Kao Corp., JP 60/55 089, 1985.
40. Petrofina S.A., CA 1 083 062, 1980 (P. E. A. Mortier).
41. Chisso Corp., JP 60 192 710, 1985 (M. Harada, M. Iijima, N. Saito).
42. General Electric Corp., US 4 547 596, 1985 (A. K. Mendiratta, J. J. Talley).
43. H. F. Smyth, Jr., C. P, Carpenter, C. S. Weil, U. C. Pozzani, J. A. Striegel, J. S. Nycum, *Am. Ind. Hyg. Assoc. J.* **30** (1969) 470.
44. R. A. Scala, E. G. Burtis, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 493.
45. P. Schmidt, R. Gohlke, R. Rothe, *Z. Gesamte Hyg. Ihre Grenzgeb.* **19** (1973) 485.
46. V. K. Rowe, S. B. McCollister in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2A,** Wiley-Interscience, New York 1981, p. 4620.
47. O. N. Mashkina, *Mater. Konf. Fiziol. Biokhim. Farmakol. Uchats. Prakt.* **168** (1966); *Chem. Abstr.* **67** (1967) 570 066.
48. D. E. Moody, *J. Cell Biol.* **70** (1976) 271.
49. J. A. Kamil, J. N. Smith, R. T. Williams, *Biochem. J.* **53** (1953) 137.
50. P. W. Albro, *Xenobiotica* **5** (1975) 625.

Europium → Rare Earth Elements
Eutrophication → Laundry Detergents

Explosives

Jacques Boileau, Consultant, Paris, France

Claude Fauquignon, ISL Institut Franco-Allemand de Recherches de Saint-Louis, Saint-Louis, France (Chaps. 2 and 3 except 2.6)

Bernard Hueber, Nobel Explosifs, Paris, France (Chap. 7)

1. Introduction 1
2. Physical Properties and Chemical Reactions 4
2.1. Detonation 4
2.1.1. Ideal Detonation 4
2.1.2. Deflagration and Detonation 5
2.2. Prediction of Detonation Data 5
2.2.1. Complete Calculation 5
2.2.2. Approximation Methods 5
2.3. Nonideal Detonation Waves and Explosives 7
2.3.1. Nonideal Explosive Compositions . . 7
2.3.2. Detonation of Cylindrical Cartridges . 7
2.3.3. Low- and High-Order Detonation Velocity 8
2.3.4. The Effect of Confinement 8
2.4. The Buildup of Detonation 8
2.4.1. Combustion – Deflagration – Detonation Transition (DDT) 8
2.4.2. Shock-to-Detonation Transition (SDT) 9
2.4.3. Shock and Impact Sensitivity 9
2.5. Classification of Explosives 10
2.6. Functional Groups 11
2.6.1. Nitro Group 11
2.6.2. Other Groups 11
3. Application 12
3.1. Energy Transfer from the Explosive to the Surroundings 12
3.1.1. Shock and Blast Waves 12
3.1.2. Casing and Liner Acceleration 12
3.2. High Compression of Solids 13
3.3. Metal Forming and Welding 13
3.4. Rock Blasting 13
3.5. Perforators, Shaped or Hollow Charges 13
4. Primary Explosives 14
5. Secondary Explosives 15
5.1. Production 15
5.2. Specific Secondary Explosives 17
5.2.1. Nitrate Esters 17
5.2.2. Aromatic Nitro Compounds 19
5.2.3. *N*-Nitro Derivatives 20
6. High Explosive Mixtures 22
6.1. Desensitized Explosives 23
6.2. TNT Mixtures 23
6.3. Plastic-Bonded Explosives (PBX) . . 24
7. Industrial Explosives 24
7.1. Dynamites 24
7.2. Ammonium Nitrate Explosives . . . 25
7.3. Ammonium Nitrate – Fuel Oil Explosives 25
7.4. Slurries and Water Gels 26
7.5. Explosive Emulsions 26
7.6. Uses 27
8. Test Methods 27
8.1. Performance Tests 28
8.2. Safety 28
9. Legal Aspects and Production 29
9.1. Safety Regulations 30
9.2. Production of Military Explosives . 30
10. Toxicology and Occupational Health 30
11. References 31

1. Introduction

An *explosion* is a physical or chemical phenomenon in which energy is released in a very short time, usually accompanied by formation and vigorous expansion of a very large amount of hot gas:

1) Mechanical explosions are caused by the sudden breaking of a vessel containing gas under pressure
2) Chemical explosions are caused by decomposition or very rapid reaction of a product or a mixture
3) Nuclear explosions are caused by fission or fusion of atomic nuclei

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_143

4) Electrical explosions are caused by sudden strong electrical currents that volatilize metal wire (exploding wire)

Only chemical explosions are treated in this article.

For an explosion to occur, the reaction must be exothermic; a large amount of gas must be produced by the chemical reaction and vaporization of products; and the reaction must propagate very fast. For example, gasoline in air burns at a rate of ca. 10^{-6} m/s; a solid propellant, at ca. 10^{-2} m/s; and an explosive detonates at a rate of ca. $10^3 - 10^4$ m/s (detonation velocity).

The two different modes of decomposition are deflagration and detonation. *Deflagration* exhibits two characteristics: 1) the combustion is very rapid (1 m/s up to a few hundred meters per second) and 2) the combustion rate increases with pressure and exceeds the speed of sound in the gaseous environment, but does not exceed the speed of sound in the burning solid. The materials are often powdered or granular, as with certain pyrotechnics and black powder. *Detonation* is chemically the same as deflagration, but is characterized by a shock wave formed within the decomposing product and transmitted perpendicularly to the decomposition surface at a very high velocity (several thousand meters per second) independent of the surrounding pressure (see Chap. 2).

Explosive substances can be divided into three classes. Members of the first class detonate accidentally under certain conditions. These are explosible substances, some of which are used in industry as catalysts (e.g., peroxides), dyes, and fertilizers. This class includes products or mixtures whose formation must be avoided or controlled, e.g., firedamp, or peroxides in ethers. In the second class are products normally used for their quick burning properties but which may detonate under some circumstances, e.g., pyrotechnic compositions, propellants, and some kinds of hunting powder. In the third class are substances intentionally detonated for various purposes.

For reasons of safety, acquaintance with the first group of materials is necessary. The second group is described elsewhere (→ Propellants; → Pyrotechnics). Pure substances and mixtures of the third class are described here.

The distinctions between these three classes are not clear-cut because most explosives burn smoothly if they are not confined. However, if some fine hunting powder burns under certain confined conditions, combustion may become detonation. Dry nitrocellulose fibers can easily detonate, but this tendency is significantly lower in the gelatinized form. Some compositions, such as mixtures of cyclotrimethylenetrinitramine (RDX) with a binder, can be used as a propellant, gunpowder, or high explosive, depending on the type of initiation.

The third class consists of primary and secondary explosives. *Primary explosives* (initiator explosives) detonate following weak external stimuli, such as percussion, friction, or electrical or light energy. *Secondary explosives* are much less sensitive to shock. However, they can detonate under a strong stimulus, such as a shock wave produced by a primary explosive, which may be reinforced by a booster composed of a more sensitive secondary explosive. The various secondary explosives are used militarily or industrially as shown in Figure 1.

Functions and Constraints. Explosives can be either pure substances or mixtures. They function in such systems as munitions, where they are a component of a complex firing system or as firing devices in quarries or metal forming equipment. With such systems, the ingredients must fulfill one or more functions, while meeting various constraints arising in manufacture or use. Therefore, tests that represent these functions and constraints are required.

When an explosive detonates, it generates a shock wave, which may initiate less sensitive explosives, cause destruction (shell fragments, blast effect, or depression effect), split rocks and soils, or cause formation of a detonation wave. A detonation wave of special geometry (hollow charge effect) may modify materials by very rapid generation of high pressure; for example, shaped charges, metal hardening, metal-powder compaction, or transformation of crystalline forms. A shock wave may be used to transmit signals, e.g., for safety devices or in seismic prospecting.

In general, constraints are related to safety, stability, compatibility with other elements of the system, vulnerability, toxicity, economics,

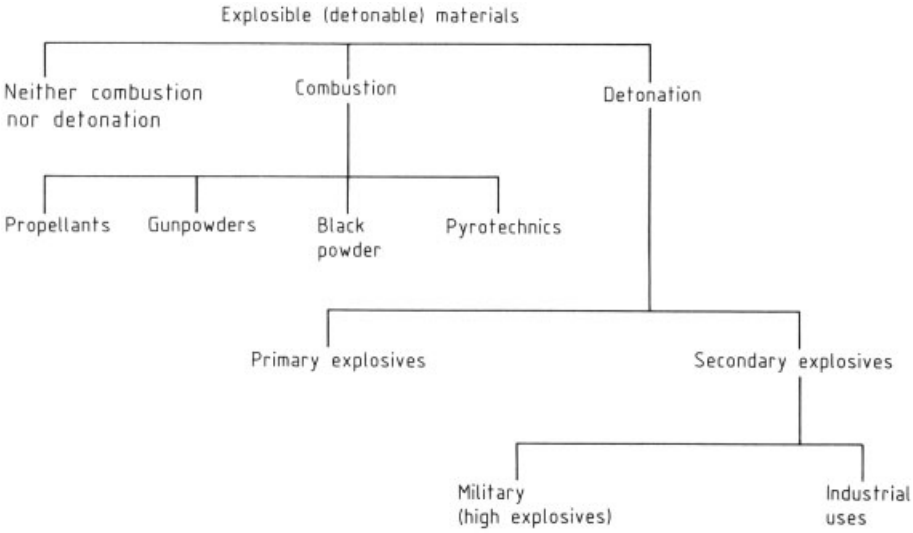

Figure 1. Differentiation of explosible materials

and, more recently, environmental and disposal problems [21].

History [1], [4], [87]. Explosives were probably first used in fireworks and incendiary devices. The admixture of saltpeter with combustible products such as coal and sulfur produced *black powder*, already known in China in the 4th century A.C., described in 808 A.C. by Qing-Xu-Zi, and used as a military gunpowder in a book published in 1044. The use in shells during the mongolian wars around 1270 and a severe explosion in a factory in 1280 were described. The first correct description of the phenomenon of shock waves in air seems to be in a book by the scientist Song-Ying-Xing in 1637. Around 1580 first descriptions in Europe are known (siege of Berg-op Zoom). However, the difficulties of initiation upon impact against the target were not overcome until 1820 when *fulminate caps* were developed. In the early 1600s, black powder was used for the first time to break up rocks in a mine in Bohemia. This technique spread throughout western Europe during the 1600s. Ammonium perchlorate was discovered in 1832.

The development of organic chemistry after 1830 led to new products, although their explosive properties were not always immediately recognized. These include nitrocellulose and nitroglycerin.

The very important contributions of Alfred Nobel (1833 – 1896) include the use of mercury fulminate for the initiation of explosives (1859 – 1861), the development of *dynamites* (1873), and the addition of 8 % nitrocellulose to nitroglycerin (blasting gelatine, 1876).

Many new products were developed between 1865 and 1910, such as nitrated explosives, mixtures in situ of an oxidizer and a fuel, explosives safe in the presence of firedamp, chlorate explosives, and liquid oxygen explosives. Organic nitro compounds for military uses included tetryl, trinitrophenol, and trinitrotoluene.

Between the two world wars, RDX, pentaerythritol tetranitrate (PETN), and lead azide were produced. After 1945 cyclotetramethylenetetranitramine (HMX), 1,3,5-triamino-2,4,6-trinitrobenzene (TATB), and hexanitrostilbene (HNS) were developed (see Chap. 5), as well as "fuel – air explosives". Ammonium nitrate – fuel oil (ANFO) explosives and slurries were developed for industrial uses; some were improved by adding bubbles or micropores.

In the 1880s Berthelot described the phenomenon of *detonation.* About the same time the hollow-charge effect was discovered, and the foundations of the hydrodynamic theory of detonation were established. In 1906, the first accurate measurements of velocities of detonation were made. After World War II, the science of

detonation was further developed and perfected [17].

Recently, new explosive compositions of low vulnerability (LOVA = low vulnerability ammunitions) are being developed. Major efforts are being made to use predicted properties (e.g., density, ΔH_f, sensitivity, themal stability) to avoid unnecessary experiments [18–20], [94], [102].

2. Physical Properties and Chemical Reactions

2.1. Detonation

The detonation process needed for most uses is characterized by a shock wave that initiates chemical reactions as it propagates through the explosive charge. The shock wave and reaction zone have the same supersonic velocity; a fraction of the chemical energy is used to support the shock.

2.1.1. Ideal Detonation

A model of ideal detonation (ID) is shown in Figure 2, with steady flow to the end of the reaction zone.

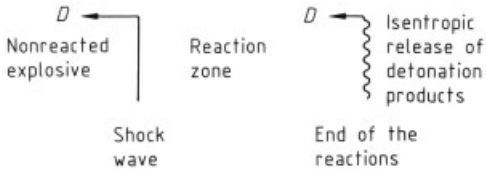

Figure 2. Ideal detonation

Stationarity requires the plane corresponding to the end of the reactions to be locally sonic. This condition, termed the Chapman – Jouguet (CJ) condition [23], [24], yields the relation $D = u + c$

needed to solve the equations of conservation of flow (where D = detonation velocity, u = particle velocity, and c = sound velocity).

The structure of the reaction zone of the ID plane detonation can be ignored, and the mechanical and thermodynamic data can be calculated by solving equations between the nonreacted and fully reacted states. The description of the ideal detonation is given by the model [25–27] represented in Figure 3 by the $p-V$ plane (Fig. 3A) and the pressure p vs. distance x profile at a given instant of time (Fig. 3B). This model relates the explosive at rest (V_0), the shocked nonreacted explosive (ZND spike ($*$)), and the end of the reaction zone (CJ plane ($\hat{}$)). This model has been recently experimentally ascertained by tests made by F. Bauer [28] with a PVDF piezogauge in cooperation with Lynn Barker by a VISAR system on a mixture PETN-binder.

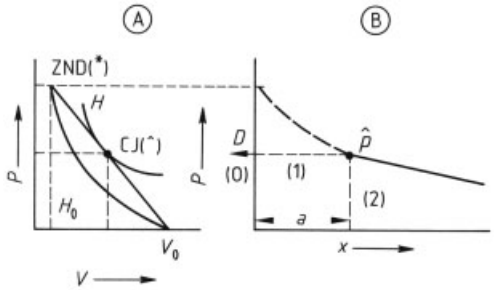

Figure 3. Development of the ideal detonation
A) The $p-V$ plane; B) The pressure vs. distance profile; p pressure; V specific volume; D detonation velocity; H_0 Hugoniot curve of nonreacted explosives; H Hugoniot curve of reaction products; (0) explosive at rest; (1) reaction zone of length a, an arbitrary quantity; (2) isentropic release of the detonation product

In Figure 3A the three states are located on a straight line (Rayleigh line), with slope equal to $-D^2/V_0^2$.

The loci of the shocked states are termed Hugoniot curves: H_0 for the unreacted explosives, and H for the completely reacted explosives.

Some relations at the CJ plane can be expressed as a function of D and the polytropic coefficent of the detonation products:

$$\Gamma = \left(\frac{\partial \log \hat{\varrho}}{\partial \log \hat{V}}\right)_s$$

The notation $(\,)_s$ represents the derivative along the isentrope at the CJ point.

$$\hat{u} = \frac{D}{\Gamma+1} \quad \hat{p} = \varrho_0 \frac{D^2}{\Gamma+1}$$
$$\hat{c} = \frac{\Gamma}{\Gamma+1} D \quad \hat{\varrho} = \varrho_0 \frac{\Gamma+1}{\Gamma}$$

where ϱ = density These relations are valid for most condensed organic explosives $\Gamma \approx 3$, with the assumptions that $p_0 = 0$ and $u_0 = 0$.

2.1.2. Deflagration and Detonation

In a thermodynamic diagram such as Figure 3A, there is another point that satisfies the Chapman – Jouguet condition in the region $p < p_0$ (not drawn in the figure). It represents the Chapman – Jouguet deflagration, in contrast to the detonation:

Detonation	Deflagration
$\hat{p} > p_0$	$\hat{p} < p_0$
$\hat{u} > 0$	$\hat{u} < 0$
$\hat{V} < V_0$	$\hat{V} > V_0$

Unlike the detonation wave, the deflagration wave is subsonic, and consequently a precursor shock is propagated in front of the reaction zone. Its intensity and velocity depend on the chemical energy released and on the boundary conditions; in contrast to detonation, a specific explosive does not provide a unique solution for deflagration.

A precursor shock that is strong enough can, in addition to compressing the explosive, also heat it sufficiently to initiate reactions just behind its front; a progressive buildup of a completely stationary process identical with the ZND model of the detonation is observed. Consequently, a detonation is equivalent to a shock followed by a deflagration.

2.2. Prediction of Detonation Data

2.2.1. Complete Calculation [29]

The quantities D, $\hat{p}$, $\hat{u}$, $\hat{V}$, and $\hat{T}$ in the Chapman – Jouguet state are needed to evaluate further the effectiveness of the explosive on the surroundings in a given action. The calculation of these quantities requires the equation of decomposition, the heats of reaction, and an equation of state for the reaction products, which may be theoretical (virial expansion) [30], [31], semiempirical [32], empirical [33], or a constant Γ law. In addition, the equilibrium constants for the reaction given as a function of V and T are needed.

For organic explosives, the distinction between oxygen-positive or weakly oxygen-negative explosives, which give only gaseous products, and the strongly oxygen-negative explosives, which also give free carbon, is important. In fact, the assignment to one class or the other is determined by the values of the equilibrium constants for $(\hat{V},\hat{T})$ determined after a first calculation.

Inasmuch as $(\hat{V},\hat{T})$ are a function of the loading density ϱ_0, some explosives, for instance, pentaerythritol tetranitrate (PETN), produce free carbon at high loading densities and only gaseous products at moderate values of ϱ_0. Because of the complexity of the calculation, the number and formulation of the equations depend on the final result.

The thermochemical equations are solved with *a priori* (V,T) pairs to give gas product composition. The mechanical equations are then applied at the Chapman – Jouguet plane with the geometric condition represented on Figure 3A: at the CJ point, the isentropic and the Hugoniot curves have the same tangent, with a slope equal to $-D^2/V_0^2$.

2.2.2. Approximation Methods

A first prediction of the CJ data can be given by using methods valid for condensed explosives.

Chemical Potential. For many organic explosives, D and $\hat{p}$ can be expressed simply as a function of a parameter Φ defined as [34]:

$$\Phi = N\sqrt{\bar{M}_r Q}$$

N = number of moles of gaseous products per gram

$\bar{M}_r$ = average molecular mass of these products, g/mol

Q = enthalpy of the explosive minus the enthalpy of the products, J/g

The calculation of Φ is made under the assumption that:

$$\mathrm{O + H \longrightarrow H_2O}$$
$$\mathrm{O + C \longrightarrow CO_2}$$

i.e., the formation of water is considered before the formation of carbon oxides.

The following relations are then proposed:

$$D = A\,(1 + B\,\varrho_0)\sqrt{\Phi}$$
$$\hat{p} = K\,\varrho_0^2\,\Phi$$

A good fit with experimental data is found for organic explosives by setting

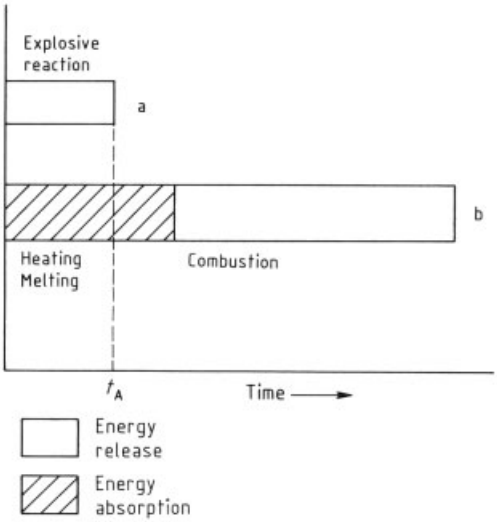

Figure 4. Sequence of reactions: explosive and combustible additive
a) Cast organic explosive; b) Aluminum grains or flakes

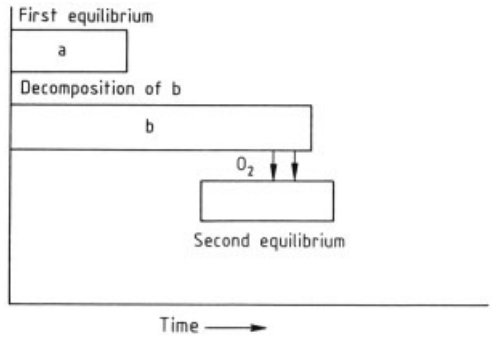

Figure 5. Sequence of reactions: explosive and oxygen-rich binder
a) Negative-oxygen-balance explosive; b) Oxygen-containing binder

$$A = 1.01 \quad B = 1.3 \quad K = 15.58$$

This method, which requires only a knowledge of the equation of decomposition, is useful for comparing organic explosives.

Rothstein Method. An empirical correlation has been found between the detonation velocity D and a parameter F, which is a function of the explosive molecule [35].

$$D(\varrho_0) = \frac{F - 0.26}{0.55} - 3(\varrho_{\mathrm{TM}} - \varrho_0)$$

ϱ_{TM} = theoretical density This correlation applies to both organic nitro and fluorinated nitro explosives.

$$F = 100\,\frac{n(\mathrm{O}) + n(\mathrm{N}) + n(\mathrm{F}) - \left(\frac{n(\mathrm{H}) - n(\mathrm{HF})}{2n(\mathrm{O})}\right)^{*}}{M_{\mathrm{r}}} + 100\,\frac{\frac{A}{3} - \frac{n(\mathrm{B/F})}{1.75} - \frac{n(\mathrm{C})}{2.5} - \frac{n(\mathrm{D})}{4} - \frac{n(\mathrm{E})}{5}}{M_{\mathrm{r}}} - G$$

where n(F), n(HF), and n(B/F) are the elemental terms for fluorinated explosives, and where $G = 0.4$ for each liquid explosive component, $G = 0$ for solid explosives, and $A = 1$ if the compound is aromatic; otherwise $A = 0$.

$*$ If $n(\mathrm{O}) = 0$ or if $n(\mathrm{HF}) > n(\mathrm{H})$, this term = 0.

For 1 mol of composition:

n(O) = number of oxygen atoms
n(N) = number of nitrogen atoms
n(H) = number of hydrogen atoms
n(F) = number of fluorine atoms
n(HF) = number of hydrogen fluoride molecules that can form from available hydrogen
n(B/F) = number of oxygen atoms in excess of those available to form CO_2 and H_2O or the number of fluorine atoms in excess of those available to form HF
n(C) = number of oxygen atoms doubly bonded to carbon as in a ketone or ester
n(D) = number of oxygen atoms singly bonded to carbon as in C–O–R where R = H, NH_4, C, etc.
n(E) = number of nitrato groups existing as a nitrate ester or as a nitric acid salt such as hydrazine mononitrate

Molecular masses and atomic composition for explosive mixtures must be derived as sums of mass-average molecular mass and of elemental mole fractions. The prediction of the detonation velocity is ca. 95 % accurate.

2.3. Nonideal Detonation Waves and Explosives

A detonation wave is nonideal if the geometry and dimensions of the charge are such that the reaction zone is affected by lateral shock or rarefaction waves. Consequently, the nonideality depends strongly on the dimensions of the charge and the explosive composition characterized by a given reaction zone length.

An explosive composition consisting of components with different reaction kinetics does not satisfy the ideal detonation model, which assumes all the exoenergetic reactions to be completed simultaneously in the sonic Chapman–Jouguet plane. Such compositions are nonideal explosives. The ideal detonation model applies only as an approximation to most multicomponent explosives; however, with data on reaction kinetics under pressure often lacking, calculations are commonly based on the ideal detonation model.

2.3.1. Nonideal Explosive Compositions

An example of an organic explosive with a metallic nonexplosive additive is a dispersion of aluminum grains or flakes in a cast organic explosive. The process sequence is shown in Figure 4. The energy effectively supporting the detonation wave has been delivered at time t_A. The absorption of energy by aluminum reduces $\hat{p}$ and D, but the late combustion of the aluminum allows the pressure to be sustained behind the CJ plane, thereby increasing the total impulse transmitted to the surroundings.

The possible sequence of an organic explosive (a) with a negative oxygen balance cast or pressed with a binder containing oxygen (b) is shown in Figure 5. The relatively slow decomposition of the binder produces oxygen gas, which shifts the reaction of the organic explosive toward better oxygen balance, thus increasing energy production.

A typical example of a mixture of two explosive components with different reaction kinetics is a mixture of trinitrotoluene (TNT) and ammonium nitrate (AN). The slower reaction of AN produces energy and an excess of oxygen, which shifts the equilibrium of the TNT reaction to a higher energy release. Because the detonation of the nonideal explosives involves heat transfer between components, the specific surface areas become important. Moreover, because the total reaction zone is relatively long, detonation depends on the dimensions of the charges.

2.3.2. Detonation of Cylindrical Cartridges

In the detonation of cylindrical cartridges, the reactive flow is two-dimensional, stationary, and relatively easy to model. Many military or engineering applications use cylindrical geometry.

The basic phenomenon is the interaction of the inwardly propagating rarefaction fan, originating at the outer surface as it is reached by the shock wave, with the reaction zone. The first consequence of this interaction is freezing of the reactions by cooling, which reduces the released energy. A second effect is a curvature of the detonation front and a loss in energy through a diverging flow. The reaction zone length and the radius of the cartridge are two quantities that play opposite roles in the process.

The experimental determination of the detonation velocity as a function of the diameter produces a curve limited in two ways:

1) As the diameter increases, the detonation velocity D approaches a constant value D_i equal to that of the ideal detonation wave.
2) Below a given diameter d, called the critical diameter, d_{cr}, the detonation is no longer self-supporting and fails.

Many analytical expressions $\frac{D}{D_i}(d)$ have been proposed in which the explosive is characterized by its ideal reaction zone length. The function $D(d)$ and the critical diameter d_{cr} depend on both the length and the structure of the reaction zone, i.e., on factors that affect this zone, such as the loading density ϱ_0, grain size g, initial temperature t_0, and composition, especially in the case of mixtures.

As an example of the effects of these factors on ideal explosive compositions, the influence of ϱ_0 is shown in Figure 6 for the conventional military composition 60 RDX/40 TNT (composition B) [36]. In the case of energy-rich explosives, the effect of the cylinder diameter increases with decreasing density, whereas the critical diameter decreases with increasing density and increasing

temperature [37]. The influence of mean grain size, e.g., that of powdered TNT, is shown in Figure 7 [29], and most nonideal explosive compositions and some pure explosives exhibit a particular behavior (see Chap. 6).

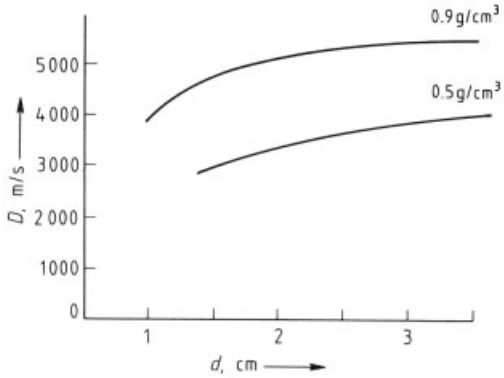

Figure 6. Detonation velocity D vs. diameter d for two densities ϱ of 60 wt % RDX – 40 wt % TNT

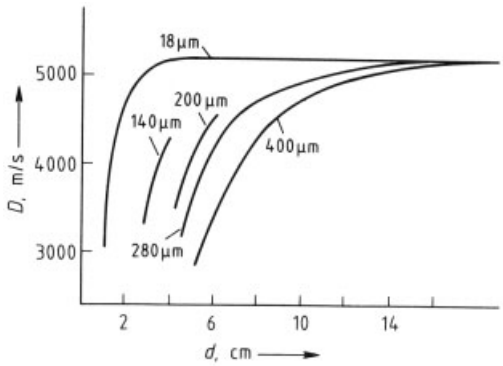

Figure 7. Detonation velocity D vs. diameter d for several grain sizes g of powered TNT

2.3.3. Low- and High-Order Detonation Velocity

Some explosives exhibit two different detonation velocities, depending on the diameter of the cartridge and the initial ignition energy. Such explosives always have a relatively low energy and loading density, with great sensitivity to shock waves [36]. Conventional examples are the dynamites and highly porous explosives [38].

The high sensitivity to shock waves tends to allow the detonation to be sustained. This is true even when the development of the reactions is seriously limited by the size of the cartridge. Thus, under steady conditions, a state of shock plus reaction would correspond to that appearing in a transient regime in the buildup of heterogeneous explosives to detonation.

2.3.4. The Effect of Confinement

Confinement usually delays the arrival of the expansion waves on the axis. A confined charge is equivalent to a charge having a large diameter. However, two anomalous effects should be noted: if the velocity of sound in the confinement exceeds the detonation velocity for the given diameter, a foreshock is propagated ahead of the wave. This foreshock can accelerate the wave (increase the density) or stop the detonation, i.e., desensitize a porous explosive by compaction, the explosive becoming homogeneous. An analogous desensitizing effect can be generated if there is an air gap between the explosive and the container. The expanding reaction products adiabatically compress the porous explosive. This effect can stop the detonation of mining charges embedded in boreholes.

2.4. The Buildup of Detonation

2.4.1. Combustion – Deflagration – Detonation Transition (DDT)

Deflagration can generate a shock wave, which is propagated in the unreacted medium, and if strong enough, can initiate reactions and become a detonation wave. This transition can occur only if the explosive charge is confined or is of such a size that expansion waves do not prevent formation of a shock wave. Accident reports have shown that DDT affects a variety of explosives and propellant charges.

However, different phenomena, which depend on the mechanical properties of the medium, may occur [39], [40]: if the explosive is porous or exhibits poor mechanical behavior, the gas formed and subjected to pressure is injected between the grains or into cracks, increasing the combustion surface. The resulting acceleration stops only when detonation occurs. The path versus time of the observed ionization front is continuous. In the case of explosives stiff

enough to withstand an elastic wave, a steady deflagration develops as soon as the pressure makes the medium impervious to the gas. The transition to detonation occurs as the pressure of the gas formed behind the deflagration zone becomes high enough to generate a shock wave. Detonation occurs when the shock wave reaches the front of the deflagration wave. As a consequence, a discontinuity is observed in the path versus time of the ionization front.

2.4.2. Shock-to-Detonation Transition (SDT)

If the end of an explosive charge is subjected to a shock wave, the steady detonation appears only at a distance s and with a delay τ inside the explosive charge [41]. For a given composition, s and τ are inversely proportional to the intensity of the shock wave. There are two SDT processes: one for homogeneous (for instance, a liquid) and one for heterogeneous composition.

Homogeneous Explosives. The shock wave (a) in Figure 8 propagated at constant velocity initiates decomposition reactions that are first completed along the entrance side (b) at time τ, the detonation appearing at point A. The detonation wave (c) travels in a compressed medium heated by shock wave (a). Therefore, the detonation wave (c) travels with a velocity exceeding the normal detonation velocity, which is attained at point B (wave d) as soon as wave c has reached wave a.

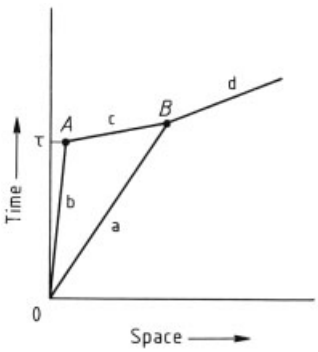

Figure 8. Detonation buildup (SDT), homogeneous explosive

Heterogeneous Explosives. In Figure 9, the shock wave (a) is gradually accelerated by the energy released upon its passage. Its acceleration ends with the appearence of a detonation wave at point B. This wave is characterized by its luminosity, as seen on optical records. Sometimes a second wave starts from point B and moves backward in the explosive, which has reacted only partially (retonation wave). The macroscopic analysis can be explained by the microscopic heterogeneities of the explosive [42].

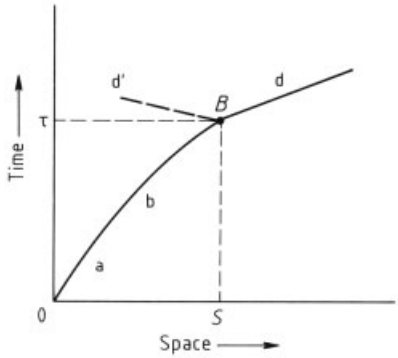

Figure 9. Detonation buildup (SDT), heterogeneous explosive

The energy of the shock wave is converted into heat energy by the implosion of occluded gas bubbles, the impact of microjets, or the adiabatic shearing of the powder grains. All the observations allow for two successive phases, ignition and buildup, that are accomodated by numerical models assuming inward or outward combustion of the grains [43].

2.4.3. Shock and Impact Sensitivity

The sensitivity of the explosives to an applied load is measured by the maximum load at which no detonation occurs, or at which there is a 50 % occurrence of detonation, i.e., a 50 % chance of failure. The sensitivity to shock loading depends on the pressure p^* and its action time τ^*. Sensitivity is defined with the aid of the curve p^* (τ^*), which separates detonation points from failure points.

For a variety of explosives and for a certain pressure range, the sensitivity is defined by the relation $(p^*)^2 \tau^* = \varepsilon$, where ε is a constant that depends on the composition and exhibits the dimension surface energy per unit area [44].

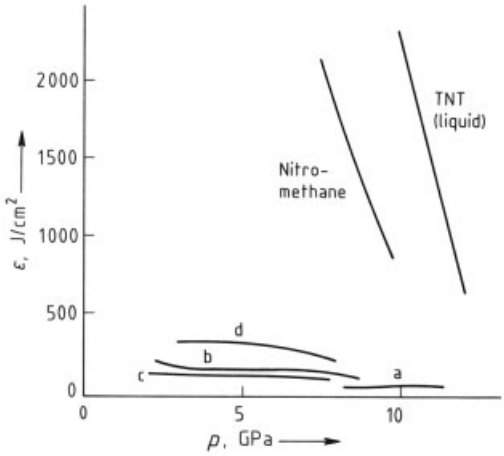

Figure 10. Initiation energy threshold vs. shock pressure a, b) HMX plus binder; c, d) Cast RDX compositions

Figure 10 shows that the concept of the energy threshold is an acceptable approximation for solid-state compositions [45]. For liquid explosives, which exhibit different behavior, the relation $\tau*$ ($p*$), based on an Arrhenius-type kinetic theory (with initial temperature $T*$ expressed as a function of pressure $p*$) accurately reproduces the experimental findings. The critical time $\tau*$ is equal, for a given shock pressure, to the time of the shock-to-detonation transition (see Fig. 8). Except for primary explosives, the sensitivity to shock-wave action and the shock-to-detonation transition are both dependent on the homogeneity of the composition. Sensitivity is affected by grain size [103] and internal grain defects [104]. The reaction of an explosive to impact loading depends on several factors. If the projectile is a large plate, the reaction is identical with that observed in the preceding case: the shock wave induced by the impact depends on the velocity and acoustic impedance, whereas the action time is proportional to the plate thickness. In tests of sensitivity to shock-wave action, the impacting plates are driven by explosives. If the projectile is a small sphere or cylinder, the interaction with the target becomes complex [41]. If the shock wave cannot induce a shock–detonation transition, it degenerates into an adiabatic compression. Frictional and shearing effects can cause ignition. Finally, a phenomenon analogous to the deflagration–detonation transition (see Section 2.4.1) is observed, in which the mechanical properties of the explosive play a part.

2.5. Classification of Explosives

For many years the behavior of cylindrical cartridges, as described in Section 2.3.2, was regarded as typical of ideal explosives. A detailed analysis has shown, however, that classification of explosives into two groups is possible, taking into account the influence of the loading density ϱ_0 on the detonation velocity as a function of the cartridge diameter and the critical diameter [46].

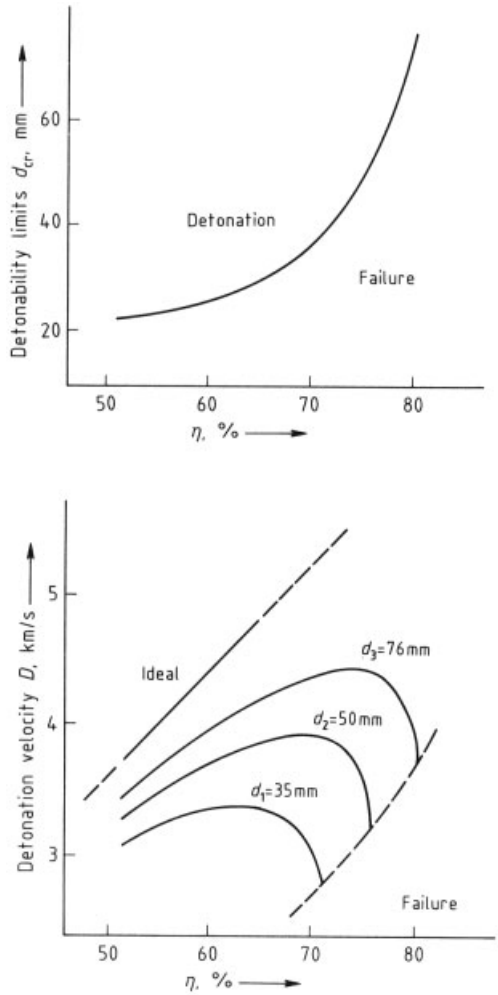

Figure 11. Detonation behavior of ammonium perchlorate; η = percent of the theoretical maximum density

The first group includes a variety of organic explosives: TNT, RDX, HMX, and their mixtures (see Chap. 5). The second group includes mixtures of an explosive and a combustible nonexploding constituent (nonideal composition). This group also includes some pure explosives such as hydrazine mononitrate (HN), nitroguanidine (NG), ammonium nitrate (AN), dinitrotoluene (DNT), dinitrophenol (DNP), and ammonium perchlorate (AP). The behavior of the second group is illustrated for ammonium perchlorate in Figure 11.

2.6. Functional Groups

Certain groups impart explosive potential.

2.6.1. Nitro Group

The nitro group is present in the form of salts of nitric acid, ONO_2 derivatives (nitrate esters), CNO_2 derivatives (aliphatic or aromatic nitro compounds), and NNO_2 derivatives (*N*-nitro compounds such as nitramines, nitroureas, etc.).

Salts of Nitric Acid. Salts of nitric acid include the alkali metal and alkaline earth metal nitrates, ammonium nitrate, and the nitrates of methylamine, urea, and guanidine. These are usually low-density, water-soluble, sometimes hygroscopic compounds. With the exception of some salts of hydrazines, e.g., triaminoguanidine and hydrazine nitrates, they are insensitive to impact and friction.

***O*-Nitro Derivatives, Nitrate Esters ($RONO_2$).** The low molecular mass representatives are liquids or low-melting solids. They are sensitive to impact when the number of carbon atoms and $-ONO_2$ groups are equal or nearly equal. Densities are in the medium range, except for symmetrical molecules such as pentaerythritol tetranitrate (PETN). Heat stability is moderate; they are subject to hydrolysis and autocatalytic decomposition.

***C*-Nitro Derivatives (CNO_2).** Aliphatic and cycloaliphatic nitro compounds differ greatly from the aromatic and heteroaromatic series. Members of the first group have limited explosive properties, except when the number of NO_2 groups equals or exceeds the number of carbon atoms, as in some *gem*-dinitro compounds and derivatives of nitroform, $HC(NO_2)_3$. The latter are often shock sensitive, with limited heat stability. The compounds $RCF(NO_2)_2$ are stable to heat and rather insensitive [6], [7]. Members of the second group containing two or three NO_2 groups per ring are valuable; they are often dense and insensitive to impact, with good hydrolytic and thermal stability.

***N*-Nitro Derivatives (NNO_2).** *N*-Nitro derivatives are often difficult to synthesize. They exhibit high densities and detonation velocities, with some sensitivity to impact. RDX and HMX are representative.

2.6.2. Other Groups

Organic chlorates and perchlorates, peroxides, metal salts of some organic compounds (acetylides and nitronates), and some organic compounds with three-membered rings or chains of nitrogen atoms or triple bonds have only limited application, except as primary explosives. Most are very dangerous to handle.

Nitroso (NO) compounds are usually unstable. So-called hexanitrosobenzene is an exception; however, it is actually not a nitroso compound but a furoxan derivative (benzotrifuroxan). It is of interest because it is free of hydrogen (zero-hydrogen explosive). Other furoxanes and also furazanes have explosive properties [88].

The difluoroamine group imparts explosive character, increased density and volatility, a lower melting point and detonation velocity, and often much higher impact sensitivity. These compounds resemble primary explosives.

Many metal azides are primary explosives. Some organic azido derivatives are being studied because of their high density and stability. However, polyazido compounds can be very sensitive.

3. Application

3.1. Energy Transfer from the Explosive to the Surroundings

The energy available in the gaseous reaction products is transferred to the surroundings by shock waves. The mechanical effects depend on the geometry of the charge and the surroundings, on the distance from the charge, and on the acoustic impedance of the media. The explosive energy is used either to create compression and tension for engineering applications or to accelerate projectiles for military applications.

3.1.1. Shock and Blast Waves

The highest dynamic pressures are produced at the exit end of the charge, where the detonation shock is transferred to the inert material. The shock pressure is a function of the shock impedance. It is defined by the so-called shock polar curve of the pressure (p) vs. particle velocity (u). By a graphical method (Fig. 12) the induced pressure can be determined at the intersection of a shock polar curve with the detonation gas-product curve passing through the CJ point.

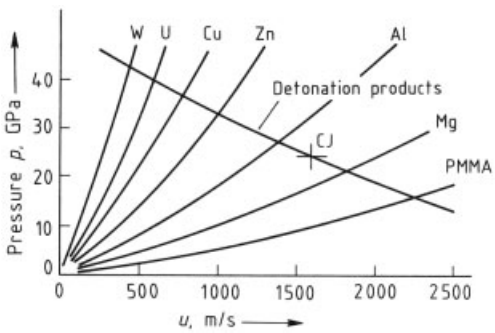

Figure 12. Determination of shock pressure p and particle velocity u induced in an inert material by an explosive; PMMA = poly(methyl methacrylate)

Even higher pressures can be obtained by using converging geometries or, indirectly, by a two-stage device. The explosive accelerates a thin metal plate, which generates an intense shock in the solid specimen as it strikes it in free flight. During wave propagation in the surrounding medium, the intensity of a shock wave decreases and the pressure profile changes. At a given distance from the charge, compression alternates with tension. The mechanical effects are a function of the maximum pressure and of the positive and negative impulses comprising the so-called blast wave. The maximum and minimum pressures are represented in Figure 13 as a function of the reduced distance $R/m^{1/3}$ from the charge.

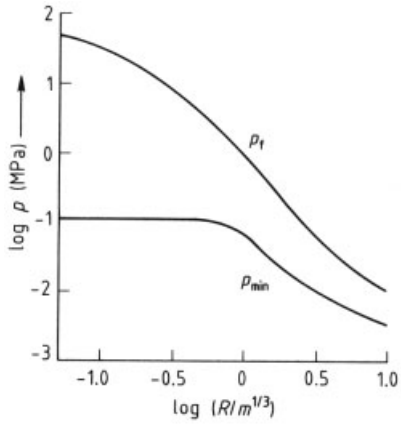

Figure 13. Front pressure p_f and minimum pressure p_{min} in air vs. reduced distance $R/m^{1/3}$ from TNT in air, where R is the distance in meters and m the mass in kilograms

3.1.2. Casing and Liner Acceleration

A casing or liner in contact with an explosive charge is accelerated by a three-step process:

1) u_1 given by the shock wave
2) $u_2 \approx 2\,u_1$ given by reflection of the shock wave in a release wave at the free surface
3) u_3 given by the further expansion of the gaseous detonation products

The Gurney formula gives an approximate u_3 for a casing filled with an explosive characterized by the quantity E, which is given in J/kg [47]:

$$u_3\,(\mathrm{m/s}) = (2E)^{1/2}\left(\frac{M}{C}+\alpha\right)^{-1/2}$$

where

M = mass of the casing, kg
C = mass of the explosive, kg
$\alpha = 0.5$ for a cylindrical charge and
$\alpha = 0.6$ for a spherical charge

3.2. High Compression of Solids

Explosives provide the most powerful means for compressing solids in spite of the fact that at high pressure the heating limits the volume reduction.

Production of New Crystal Phases. The large pressure generated by shock waves may be used to transform one crystal phase into another (polymorphic transformation). The best known example of this technique is the transformation of graphite into diamond. Unfortunately, the short duration of the shock limits this technique to the production of small crystals used as abrasives.

Powder Compaction. The compression of powders by shock waves produced by explosives creates high pressures and temperatures simultaneously, resulting in grain welding. The main problem, which has been solved only recently, is the explosion caused by the interaction of release waves, which follow the shock; special geometries are required [48]. Rapidly solidified amorphous and metastable microcrystalline materials and ultrahigh-strength ceramics are expected to be produced by this technique.

Shock Hardening. The detonation of a thin sheet of explosive covering a piece of steel creates great surface hardness by a sequence of rapid compression, heating, and cooling.

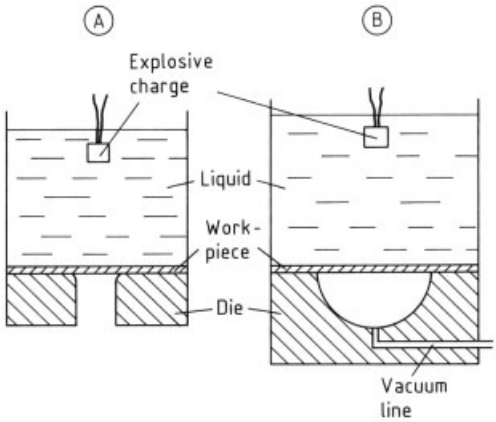

Figure 14. Setups for metal forming
A) Free forming; B) Bulkhead forming

3.3. Metal Forming and Welding

The detonation of an explosive charge is used to form a metal plate; the shock wave is moderated by a liquid transmitting medium (Fig. 14). Techniques of free forming (Fig. 14A) or bulkhead forming (Fig. 14B) may be used.

A grazing detonation may weld two metal plates with diffusion of metal through the interface (Fig. 15) [49]. The required collision velocity, v_c, depends on the materials. This process is called explosive welding or cladding.

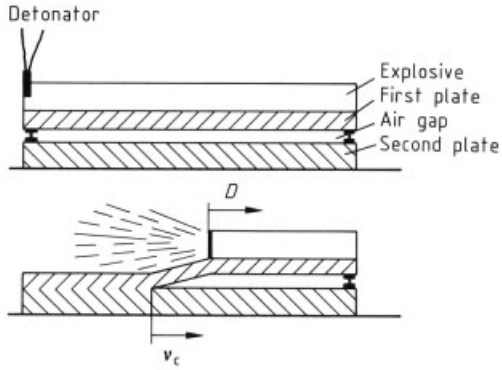

Figure 15. Explosive welding or cladding

3.4. Rock Blasting [36]

In rock blasting the charges are placed in drill holes, which are usually parallel to the rock surface. With strong confinement, most of the explosive energy is usefully employed, even though much of it is released by afterburning. Equally good results are given by nitroglycerin in high (6000 – 8000 m/s) or low (2000 – 3000 m/s) explosives. Various types of wedge cuts are used for tunnel blasting.

3.5. Perforators, Shaped or Hollow Charges

An explosive can be used to produce a thin, high-speed metallic projectile capable of perforating armor [50]. Nonmilitary applications include oil recovery.

A cavity in the explosive (Fig. 16) is lined with a metal, usually copper. The detonation divides the liner into two parts that move along

the axis at different velocities. Most of the mass of the liner forms the so-called slug at a velocity of several hundred meters per second. The remainder forms a thin projectile which is elongated because of the difference of velocity between the first formed elements near the apex (u_{jet} max. $\approx 8000-11\,000$ m/s) and those formed last (u_{jet} min. $= 1500-2000$ m/s). The ultimate length of the projectile depends on the ductility of the liner and can be > 10 times the diameter of the charge.

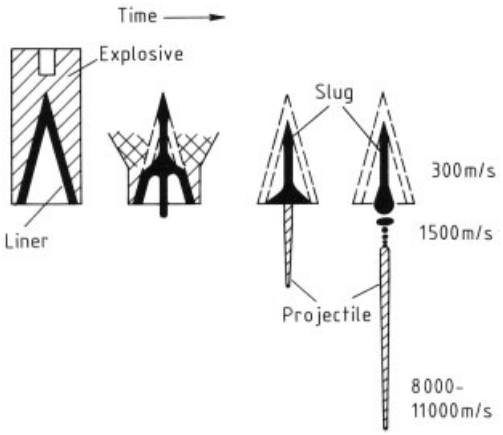

Figure 16. Projectile formation

When the projectile strikes a solid or liquid target, it drills a deep, narrow hole. The hole depth P is given approximately by the expression

$$P = L\sqrt{\frac{\varrho_{\mathrm{j}}}{\varrho_{\mathrm{t}}}}$$

where L is the length of the projectile and ϱ_{j} and ϱ_{t} are the densities of the projectile and target.

4. Primary Explosives

Under low-intensity stimulus of short duration, primary explosives, even in thin layers, decompose and produce a detonation wave; the activation energy is low. The stimulus may be shock, friction, an electric spark, or sudden heating. The deflagration – detonation transition occurs within a distance often too short to be measured. The released energy and the detonation velocity of primary explosives are small. Their formation is often endothermic.

The main function of primary explosives is to produce a shock wave when the explosive is stimulated by percussion, electrically, or optically (laser), thus initiating a secondary explosive. Primary explosives are the active detonator ingredients. Some are used in primer mixtures to ignite propellants or pyrotechnics. Because of their sensitivity, primary explosives are used in quantities limited to a few grams, and are manufactured under special precautions to avoid any shock or spark. To be used in industry, primary explosives must have limited sensitivity and adequate stability to heat, hydrolysis, and storage. Mercury fulminate, azides, and diazodinitrophenol are among the few products that meet these requirements and that can detonate a secondary explosive. Others, e.g., lead styphnate or tetrazene, initiate burning or act as sensitizers.

Properties of primary explosives are given in Table 1.

Mercury fulminate [*628-86-4*], $Hg(ONC)_2$, M_{r} 284.65, was first prepared in the 1600s and was used by Nobel in 1867 to detonate dynamites. It is prepared by the reaction of mercury with nitric acid and 95 % ethanol, in small quantities because the reaction is difficult to control. Mercury fulminate is a gray toxic powder, which lacks the stability for storage. It reacts with metals in a moist atmosphere; in most industrial countries its use has been abandoned.

Lead azide [*13424-46-9*], $Pb(N_3)_2$, M_{r} 291.26, discovered by Curtius (1891) [13], was developed after World War I and is now an important primary explosive. It is produced continuously by the reaction of lead nitrate or acetate with sodium azide in aqueous solution under basic conditions to avoid formation of hydrazoic acid, which explodes readily. The crystal size must be carefully controlled, large crystals being dangerous, by controlling the stirring and by using wetting agents. Demineralized water must be used. With thickeners such as dextrin, sodium carboxymethylcellulose, or poly(vinyl alcohol), purities from 92 % to 99 % are possible, the former containing 3 % dextrin and 4 – 5 % $Pb(OH)_2$. The lower sensitivity of the 92 % lead azide to impact and friction compared to purer lead azides facilitates detonator loading. Lead azide has good stability to heat and storage. Contact with copper must be avoided because *copper azide* is extremely sensitive;

Table 1. Properties of primary explosives

Property	Lead azide, pure	Silver azide	DDNP	Lead styphnate	Tetrazene
Crystal density, g/cm^3	4.8	5.1	1.63	3.0	1.7
Detonation velocity, km/s	6.1 ∗	6.8 ∗	6.9	5.2	
at density	4.8	5.1	1.6	2.9	
Impact sensitivity, Nm	2.5 – 4	2 – 4	1.5	2.5 – 5	1
Friction sensitivity, N	0.1 – 1		20	8	10
Lead block test, cm^3/10 g	110	115	325	130	130
Explosion temperature, ∗∗ °C	345	290	195	265 – 280	160

∗ Theoretical. ∗∗ After 5 s.

aluminum is preferred. *Silver azide* is used at high temperature or in miniaturized pyrotechnic devices. It is prepared from aqueous silver nitrate and sodium azide.

Diazodinitrophenol [*28655-69-8*], DDNP, $C_6H_2N_4O_5$, M_r 210.06, is obtained by diazotizing picramic acid and purified by recrystallization from acetone. It is sparingly soluble in water, nonhygroscopic, and sensitive to impact, but not as sensitive to friction or electrostatic energy. It is less stable to heat than lead azide. It is most often used in the United States.

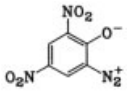

Lead styphnate [*15245-44-0*], lead trinitroresorcinate, $C_6HN_3O_8Pb$, M_r 450.27, is produced continuously by the aqueous reaction of the magnesium salt with lead acetate, sometimes in the presence of agents that promote the formation of the correct crystalline form. Especially sensitive to electrostatic discharge, it is most frequently used to sensitize lead azide and in primer compositions to initiate burning.

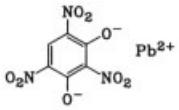

Tetrazene [*31330-63-9*], $C_2H_8N_{10}O$, M_r 188.07, is obtained by the reaction of sodium nitrite with a soluble salt of aminoguanidine in acetic acid at 30 – 40 °C. It decomposes in boiling water. Its greatest value is for the sensitization of priming compositions.

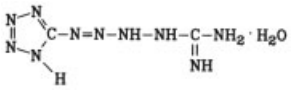

Now it is a trend in research of new primary explosives in order to replace the compounds containing lead (environmental problem).

5. Secondary Explosives

5.1. Production

Common high explosives are usually made by liquid-phase nitration. The overall mechanism is believed to be ionic, with NO_2^+ generally the reactive species. In some cases, N_2O_4 may be added to a double bond or to an epoxy group. It is also possible to nitrate gas-phase hydrocarbons. The most important nitrating agent is nitric acid → Nitric Acid, Nitrous Acid, and Nitrogen Oxides.

Numerous end products of nitration are soluble in concentrated HNO_3, and may be recovered by dilution. However, dilution below ca. 55 wt % acid is not economic.

To increase the NO_2^+ content (3 wt % in pure HNO_3), lower the solubility of end products, reduce oxidative side reactions, and facilitate the treatment of spent acids, mixtures of sulfuric and nitric acids (mixed acid) are used. The water content of a mixed acid may be reduced by adding oleum. In 50: 50 wt % mixed acid, the nitric acid is ca. 15 % dissociated.

However, sulfuric acid is difficult to remove from products by washing. Furthermore, some substances (e.g., nitramines, nitriles, etc.) are decomposed. Orthophosphoric acid or polyphosphoric acid may be used instead of sulfuric acid;

however, these phosphoric acids are more expensive and more difficult to recover.

Mild nitration or nitrolysis can be conducted in mixtures of nitric acid and acetic anhydride or acetic acid; the reactive species may be $CH_3COONO_2H^+$. Mixtures of nitric acid and acetic anhydride containing between 30 and 80 wt % HNO_3 can detonate [51]. A high concentration of acetic anhydride avoids this danger. More recent nitration methods use N_2O_5 as a solution in pure nitric acid ("nitric oleum") or in chlorinated solvents (e.g., CH_2Cl_2). Three processes for the production of nitric oleum are in operation or in development [90]:

1) Oxidative electrolysis of a $N_2O_4 - HNO_3$ mixture
2) Ozonation of N_2O_4
3) Distillation of an oleum $(H_2SO_4 + SO_3) - NH_4NO_3$ mixture [89]

Nitric oleum allows yields to be improved and permits some syntheses to be performed that are not possible by other routes; the organic N_2O_5 solutions allow nitration to be performed under very mild and selective conditions [16], [90]. These nitrating agents must be used at their site of production.

Nitration. Reaction is exothermic. Dilution of the nitric and mixed acids with water liberates heat [52]. Normally, nitration is rapid. However, 70 – 85 wt % nitric acid is also an oxidant, and for this reason the reaction is best conducted continuously to limit the contact time of the product with the reactive medium. If the medium is free of solid particles, a tubular reactor can be used; otherwise, reactors in cascade with efficient stirring are employed. These reactors are made of highly polished stainless steel. Reactors and stirrers must be carefully designed to avoid dead zones and friction with crusts of explosives. Reactors are often provided with a valve that opens quickly to discharge reactants into a dilution vessel in an emergency.

Product Isolation. If a precipitate is not highly sensitive to friction, it may be centrifuged. For safe continuous filtration, the product must be stable in its mother liquor, and the filter design must avoid introduction of the explosive between moving and fixed parts. The precipitate is washed in two stages. The first uses only a small amount of water, which is recovered and mixed with spent acid. The acid is recovered from this mixture.

Purification. Treatment with boiling water is sometimes sufficient to hydrolyze impurities and wash out nitric acid. Crystallization eliminates sulfuric acid and produces the desired grain size. The solvent can be diluted with water or removed by steam distillation. Very fine crystals are obtained by dilution with high-speed stirring or by grinding in water or an inert liquid. A fluid energy mill may also be used. Size separation can be effected by sieving under flowing water or in a classifier. Drying is usually done as late as possible in the process.

Recovery of Spent Acids. Acid recovery in an explosive plant is of great importance in controlling production costs. These acids include 55 wt % HNO_3, 63 – 68 wt % H_2SO_4 from nitric acid concentration, and spent H_2SO_4 containing some nitric acid and nitration products.

To concentrate nitric acid, water is removed by countercurrent extraction with 92 – 95 % H_2SO_4. At the top, 98 – 99 % HNO_3 is produced. At the bottom, 63 – 68 % H_2SO_4 is obtained, which can be concentrated to 93 % by stripping with combustion gases, or to 96 – 98 % by vacuum distillation. Nitric acid can also be concentrated by distillation over magnesium nitrate.

Sulfuric acid is freed of nitric acid and nitro compounds by heating or steam injection.

Pollution Problems. Gaseous pollution may occur during nitrations, with evolution of red nitrous fumes. These can be absorbed in columns by recycling water or dilute nitric acid to provide 50 – 55 wt % acid.

Liquid and solid pollution is created by washing. Usually acids are transferred to decanting and settling basins, where product and other solid particles settle and liquids are neutralized. Some liquid wastes require special treatment; for example, the red liquors from TNT are best destroyed by combustion.

Safety [9], [11]. Accidents caused by detonation are usually very severe. In addition to the normal precautions required in acid handling, the production and use of explosives obviously

involves special risks. Loss of life and destruction of property by an accidental explosion must be prevented at all costs, and the detonation of explosives stored near populated areas must be avoided. Minimum distances between plant and nearby structures are regulated.

The transmission of detonation by pipes or feeding devices must also be prevented by appropriate arrangements. The danger of detonation by an accidental fire can be mitigated by limiting vessel size. Detonations by shock and friction in pipes, pumps, or valves are prevented by suitable measures.

The following principles apply: partition of risks (e.g., specific buildings for specific operations); limitation of risks (limited number of persons present and limited quantities of explosives); and installation of two or three independent safety devices. Some of these measures are taken by the manufacturers under the supervision of professional organizations following government regulations.

5.2. Specific Secondary Explosives

Properties of secondary explosives are given in Tables 2 and 3.

Table 3. Impact sensitivity of secondary explosives *

Explosive	Value **
Insensitive	
TATB	> 2.5
TNT	1
Moderately sensitive	
DINGU	0.8
HNS	0.6
NTO	0.8
Sorguyl	0.15 – 0.2
CL20	0.15 – 0.2
Sensitive	
HMX, RDX	0.3
PETN	0.15 – 0.2 [a]
NG	0.1

* Drop-weight impact test [7].
** Based on TNT = 1.
[a] values are dependent on particle size

5.2.1. Nitrate Esters

Nitrate esters $RONO_2$ are prepared from alcohols and nitric acid, which may be mixed with sulfuric or acetic acid. The reaction is reversible:

$$NO_2^+ + ROH \rightleftharpoons RONO_2 + H^+$$

In an anhydrous medium the equilibrium shifts to the right; with dilution, hydrolysis occurs. In 60 – 80 % HNO_3, the unreacted alcohol may be oxidized. Furthermore, a catalytic effect of the nitrous acids produced causes rapid decomposition of the reaction medium. This decomposition either does not occur or occurs very slowly with an acid concentration below 55 %. Nitrate esters are usually less stable if traces of acid are present. Traces of sulfuric acid are difficult to remove from solid nitrate esters; therefore, mixed acid should not be used for their preparation. Acetic – nitric acid mixtures are used with sensitive or oxidizable alcohols.

Pentaerythritol tetranitrate [*78-11-5*], PETN, $C_5H_8O_{12}N_4$, M_r 316.15, is scarcely soluble in 82 % HNO_3 (0.7 wt %). The nitration of pentaerythritol is practically complete in a few seconds and is conducted continuously. Pentaerythritol and HNO_3 (ca. 1 : 5.5) are fed into a well-stirred, jacketed reactor, with overflow in cascade into a second reactor of the same design. The temperature is maintained below 30 °C. PETN precipitates and may be isolated in one of two ways:

1) In a third reactor the reaction mixture is diluted with water to 55 wt % acid (c_{PETN} = 0.1 wt %) and filtered. The PETN is washed with water
2) The reaction mixture is filtered immediately and then the 82 wt % spent nitric acid is concentrated, a cheaper process but less safe

Washed PETN is stable. It can be recrystallized by dilution of a hot acetone solution with water, which removes traces of acid and some organic impurities, and crystals of desirable size and flow properties for continuous feed purposes are obtained. Pentaerythritol tetranitrate is fairly stable above 100 °C and in alkaline medium. It is a powerful high explosive, with a small critical diameter, easy to initiate, and sensitive to impact and friction. But that sensitivity depends on the crystal size: PETN is more insensitive if ultrafine or in form of a very big monocrystal. Nanometric PETN was recently obtained by a sol – gel process. It is used in detonators, commercial detonating cord, boosters (with wax), sheet explosives (with an elastomeric binder),

Table 2. Properties of secondary explosives

Property	PETN	NG	TNT	HNS	TATB	RDX	HMX	DINGU	NTO	ADN	Sorguyl	CL20 (ϵ)
Melting point *mp*, °C	141.3	13.5	80.8	318 decomp.	>452 decomp.	204	283	260 decomp.	255 decomp.	92	190	
Vapor pressure, Pa	0.0011 at 100 °C	1 at 40 °C	14 at 100 °C	1.3×10^{-7} at 100 °C	0.4 at 175 °C	0.05 at 110 °C	4×10^{-7} at 100 °C					
Crystal density, g/cm^3	1.77		1.654	1.74	1.94	1.82	1.907 [b]	1.98 [e]	1.91 [d]	1.83 [d]	2.03–2.04	2.04
		1.591 [a]	1.47 [a] at 81 °C									
Detonation velocity at max. density, m/s [c]	8340	7700	6960	7120	7970	8850	9100	8450	8520	ca. 3900	9300	9300
Critical diameter, mm	0.4			0.4	6–8			6–8		4–8	<1	
Lead block test, 10 g, cm^3	520	520	300			480	480					
Explosion temperature after 5 s, °C	225	220–240	470	305	315 decomp. 520 explodes	301	330					

[a] Liquid. [b] β-form. [c] Calculated in some cases. [d] α-form. [e] by *x*-rays 1.99

and plastic explosives. In a mixture with TNT, it is employed in some warheads and in industrial explosives.

Nitroglycerin [*55-63-0*], NG, $C_3H_5O_9N_3$, M_r 227.09, unlike PETN, is a liquid at room temperature, totally miscible with HNO_3 of higher than 70 % concentration, sparingly soluble (3 wt %) in 50 % HNO_3, and sensitive to hydrolysis (any contact with hydrolytic medium must be very short) [53]. It is prepared by continuous two-phase nitration of glycerol in anhydrous mixed acid. The mass ratio of pure anhydrous glycerol ($H_2O < 0.5$ %) and 50: 50 mixed acid is 1: 5. The product is washed with aqueous sodium carbonate or bicarbonate.

In the Schmid–Meissner and Biazzi processes, the reactor is a vessel with overflow and good stirring and cooling. The reactants are fed continuously, while the temperature is kept at 10–20 °C. The emulsion produced flows into a separator, which in the Schmid process does not have moving parts. In the Biazzi process, the separator is conical with peripheral feeding, breaking the emulsion by rotation. The product is washed with alkaline water in columns (Schmid) or in mixer–settlers (Biazzi).

In the Gyttorp process, developed in the 1950s, the feeding of acid and alcohol can be interrupted simultaneously. The key feature is the use of an injector with an axial flow of acid; the alcohol is sucked through the neck with excellent mixing. For good suction and heat absorption, a volume ratio of acid: alcohol of ca. 8 is required; it is maintained by recycling part of the spent acid. The glycerol is heated to 48 °C to reduce viscosity. The emulsion that is formed is cooled to 15 °C, and centrifuged continuously to separate the nitroglycerin from spent acid. The product is emulsified with alkaline wash water by passage through an injector; it is separated and washed again until neutral. The final emulsion can be safely transmitted through pipes. The quantity of nitroglycerin present at any given time can be kept low, and safety is further increased by protective walls and the use of remote control.

Other liquid nitrates are produced by these processes; the Gyttorp process requires a liquid alcohol reactant. The spent acid may be purified by heating. Nitroglycerin is used in propellants, gunpowder, and dynamites.

5.2.2. Aromatic Nitro Compounds

The most common nitro-aromatic explosives contain the picryl (2,4,6-trinitrophenyl) group and are usually obtained by nitration. The Ingold mechanism involves nitronium ion, NO_2^+; electrophilic attack occurs with formation of a σ-complex intermediate [22].

The introduction of a second and third nitro group requires increasingly severe conditions, including more anhydrous mixed acids and elevated temperatures (above 100 °C for trinitration).

Some of these aromatic nitrocompounds are not used or are abandoned as explosives: among them:

1,3,5-Trinitrobenzene [*99-35-4*], TNB, benzite $C_6H_3O_6N_3$, M_r = 213.11 (I, R = H), attractive as high explosive, cannot be obtained by direct nitration of benzene, but only by indirect syntheses giving an expensive product used only for preparations of fine chemicals.

2,4,6-Trinitrophenol [*88-89-1*] $C_6H_3O_7N_3$, M_r = 229.11 (I, R = OH) TNP, picric acid, melinite, was extensively used in France during World War I because of lack of toluene, but was later replaced by trinitrotoluene (tolite), cheaper and non acidic. In some ammunitions, the ammonium picrate is still used.

N-Methyl N2,4,6-tetranitroaniline [*479-45-8*] Tetryl $C_7H_5O_8N_5$, M_r = 287 (I, R=N(CH_3)NO_2) has been used since 1905, noticeally for the production of boosters, after lubrication by graphite or calcium stearate. But it is now practically abandoned because some allergic incidents of dermatosis, and replaced mostly by RDX – wax mixtures.

Three nitroaromatic products with a picryl group remain for use as high explosives: TNT, HNS, and TATB.

2,4,6-Trinitrotoluene [*118-96-7*], TNT, tolite, $C_7H_5O_6N_3$, M_r 227.13 (**1**, R = CH_3), is produced by successive nitration of toluene, producing mono-, di-, and finally trinitrotoluenes. The mononitration products are 60 % *ortho*, 35 % *para*, and 3 – 5 % *meta* derivative.

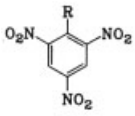

1

The *o*- and *p*-nitro compounds give 2,4- and 2,6-dinitrotoluenes and then 2,4,6-trinitrotoluene (*meta*-directing effect of NO_2). However, *m*-nitrotoluene does not give the 2,4,6-trinitro derivative; other products are formed that reduce the stability or lower the melting point of the TNT. These impurities can be removed by the so-called selliting process or by fractional distillation of the mononitration products. The *m*- and *p*-nitrotoluenes are used as chemical intermediates, and the *o*-nitrotoluene is further nitrated to TNT. This method gives a product that is more easily purified.

For reasons of safety, quality, and cost, modern high-capacity plants use a continuous process. The solubility of organic products in mixed acids can be quite low, 0.5 – 15 wt %, depending on temperature and sulfuric acid concentration. Nitration is believed to occur in the acid phase. Consequently, the rate of interphase transfer and the intensity of stirring are important.

The plant employs a series of 8 – 16 mixer – settlers. The organic phase is transferred from the preceding settler together with the acid phase from a following or preceding settler (countercurrent or parallel nitrators) into a mixer maintained at a constant temperature (50 °C for mono-, 80 °C for di-, and 100 °C for trinitration). Acids may be added to adjust the composition. From the last settler, molten TNT overflows into other mixer – settlers that are used for washing and selliting.

The spent sulfuric acid can be concentrated to ca. 96 % but not to oleum. Using oleum in the process leads to a higher yield, a shorter starting time, and a more rapid reaction; consequently, fewer mixer – settlers are required for a given output, but a use or customer must be found for the spent acid.

Selliting is a process of washing with aqueous Na_2SO_3. A preferential reaction with the unsymmetrical TNT isomers gives deep-red water-soluble products. Selliting is applied to crude molten TNT above 80 °C, but with some losses. A continuous process may also be employed at 50 °C, with TNT suspension produced by cooling an aqueous emulsion with vigorous stirring [54]. The resulting spherules contain pure 2,4,6-TNT at their center, and a mixture of isomers in the outer layer. The use of $MgSO_3$, rather than Na_2SO_3, has been proposed [55]. The most economical treatment of the red spent water from

selliting is concentration followed by combustion.

Trinitrotoluene is used alone or mixed with PETN, RDX, or ammonium nitrate. It is also used in some industrial explosives.

Hexanitrostilbene [*20062-22-0*], HNS, $C_{14}H_6O_{12}N_6$, M_r 450.24, is formed by oxidation of TNT in methanol – tetrahydrofuran solution with aqueous NaOCl [58]:

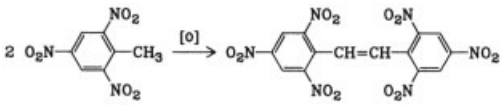

Yields of 40 – 50 % are obtained in a continuous tubular reactor [59], [60]. Hexanitrostilbene can be crystallized from HNO_3 or acetone.

Hexanitrostilbene has the following unusual properties: the presence of 1 % HNS in molten TNT prevents the formation of cracks and crazes during solidification. With its low critical diameter (0.4 mm) it can be used as the core in silver or other metallic explosive cords. It remains stable and explosive over a very wide temperature range (−200 to 250 °C) and is thus usable in space [61].

1,3,5-Triamino-2,4,6-trinitrobenzene [*3058-38-6*], TATB, $C_6H_6O_6N_6$, M_r 258.15, is obtained by the following sequence:

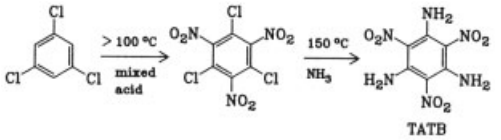

The proximity of nitro groups favors the displacement of chlorine by ammonia. The reaction is conducted under pressure at 150 °C in toluene containing a little water. The product is washed with water to remove NH_4Cl. It has low solubility and is difficult to recrystallize [62], [63]. Although discovered in 1887, it first attracted attention after World War II. Since the 1950s interest has increased because of its thermal stability and outstanding insensitivity. Mixed with HMX and a fluorinated binder, it is used as a booster. It is, however, expensive (ca. 65 $/kg in the mid-1980s in the USA) because of the high cost and low availability of trichlorobenzene. Therefore, alternative synthetic routes have been developed. An extensive review of TATB is given in [9].

5.2.3. *N*-Nitro Derivatives

N-Nitro derivatives include primary nitramines $RNHNO_2$, secondary nitramines $RR'NNO_2$, nitramides, nitroureas, nitrourethanes, and nitroguanidines. There are no general synthetic methods. Sometimes it is possible to nitrate the precursor with HNO_3 alone as with tetryl. Often an *N*-acetyl, *N*-*tert*-butyl, or *N*-nitroso derivative is used as starting material. In some cases, nitric oleum is effective. The nitration of hexamethylenetetramine (hexamine) to RDX and HMX is more complicated.

1,3,5-Trinitro-1,3,5-hexahydrotriazine [*121-82-4*], $C_3H_6O_6N_6$, M_r 222.12, cyclotrimethylenetrinitramine, Hexogen, RDX.

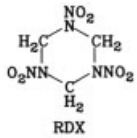

Both RDX and HMX are stable in nitric acid over a wide temperature and concentration range, but are destroyed by aqueous sulfuric acid and are attacked by alkaline substances. Sulfuric acid should not be present during the synthesis.

1,3,5-Trinitro-1,3,5-hexahydrotriazine is manufactured by nitrolysis of hexamine with nitric acid alone or with acetic anhydride. These processes were developed before and during World War II.

The main process, called the Woolwich process, which does not use acetic anhydride, was developed in the United Kingdom and is used there and in France. The raw materials are hexamine and 98 – 99 % HNO_3 free of sulfuric acid. The reaction is complex and not completely understood.

Hexamine dinitrate is probably formed first, followed by intermediates that are oxidized or decomposed, exothermically and often violently, to produce RDX, which precipitates. Large amounts of nitrous gases and CO_2 are evolved along with water. The process must be continuous with careful control of temperature and flow conditions. The plant consists of three groups of stainless steel cascade-type reactors

with efficient stirring and cooling devices (jacket or coil). Nitric acid and hexamine are fed into the first group at a temperature of 20 – 50 °C. The homogeneous mixture overflows from the last reactor into the first one of the second group (the decomposers); this reactor is equipped with a large pipe for the escape of gases and a water feeder. The temperature is maintained between 75 and 82 °C, and water is continuously added to maintain a HNO_3 concentration of 55 wt %. The gases evolved are condensed in towers, producing 50 – 55 wt % HNO_3, which is reconcentrated. This first decomposer is the key to the operation. Foam must be avoided as well as thick crusts on which the stirrer could scrape, causing an explosion. In the third group of reactors, the suspended product is cooled and then continuously filtered and washed. The resulting RDX contains HMX (1 % max.) and some nitric acid, which can be removed by heating with water at 135 °C for 3 h in an autoclave. The yield based on hexamine is 78 – 80 %.

The RDX can be recrystallized or used directly, after desensitizing by wax or 10 % TNT to permit drying and transportation. Recrystallization from cyclohexanone or methyl ethyl ketone removes traces of acidity (required, for example, for use with aluminum powder) and produces crystals of the desired size.

A process without acetic anhydride was also developed in Germany in 1940 (KH process). A modification is the addition of NH_4NO_3 to react with all CH_2 groups of hexamine (K process).

The KA process was developed during World War II in Germany, and the Bachman process was developed and is used in the United States. Both processes use acetic anhydride, and the reactions are very complex.

To a solution of hexamine (1 part) in glacial acetic acid (1.65 parts) are added a solution of NH_4NO_3 (1.5 parts) in concentrated HNO_3 (2 parts) and acetic anhydride (5.2 parts) at ca. 65 – 72 °C. The mixture is maintained at that temperature for at least 1 h, diluted with water, and heated to hydrolyze impurities. The yield is ca. 80 – 85 %; nitrous fumes are not evolved.

The product contains ca. 8 – 15 % HMX, depending on the reaction conditions. An adequate source of acetic anhydride must be nearby, where acetic wastes can be dealt with. This process is easier to carry out than the process employing nitric acid alone, but its environmental problems are more difficult to solve. The same plant can be adapted to HMX production.

Both RDX and TNT are the basic components of almost all modern high explosives. They are mixed together, alone or with wax, or formulated with plastic binders. For a powerful and heat-stable explosive, RDX is preferred.

1,3,5,7-Tetranitro-1,3,5,7-tetraazacyclooctane [*2691-41-0*], $C_4H_8O_8N_8$, M_r 296.16 (**2**, R = NO_2), octahydrotetranitrotetrazocine, cyclotetramethylenetetranitramine, octogen, HMX, was discovered shortly before World War II as a byproduct of RDX synthesis by a Bachman-type process. It exists in four crystalline forms [4], [66], [67]. The β-form has the highest density (ϱ = 1.907 g/cm^3) and a shock sensitivity similar to that of RDX. Large crystals of the α-form are shock sensitive, but microcrystals are not.

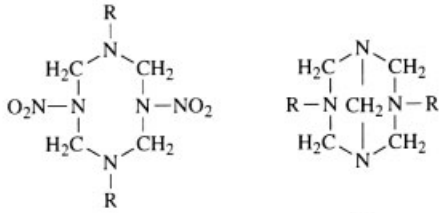

2 **3**

HMX is now obtained directly by a modified Bachman process. In this batch process, acetic anhydride and a mixture of NH_4NO_3 and HNO_3 are added simultaneously at 44 ± 1 °C to hexamine dissolved in glacial acetic acid. After 15 min, the operation is repeated. The reaction mixture is aged at 45 – 60 °C, diluted with hot water, and refluxed for up to several hours. The suspended HMX is filtered off and washed until neutral. The HMX is obtained as small crystals of the α-form, which are transformed into the β-form by recrystallization from acetone or DMSO.

HMX can also be obtained via dinitropentamethylenetetramine (DPT, **3**, R = NO_2) [68]. In another process starting from hexamine, the diacetyl intermediates DAPT (**3**, R = $COCH_3$) [69] and DADN (**2**, R = $COCH_3$) are isolated, and this is followed by nitration with nitric oleum. However, since demand for HMX is now very low, production by this route has ceased.

The price of HMX is usually 3 – 5 times that of RDX. It is used when high performance is required (e.g., shaped charges). Both RDX and HMX are also used in propellants.

Other High Explosives. Numerous nitramine derivatives have been studied. For example:

Dinitroglycoluril [*55510-04-8*] (DINGU or DNGU ; **4**, R = H) has a high density (1.99 g/cm^3, X-ray) [100] and explosive properties close to those of TATB [70], [71], but at a much lower price. It can replace part of the RDX in mixtures with TNT and can be introduced as a component in gunpowder. Some comparative thermodynamic properties are studied [72].

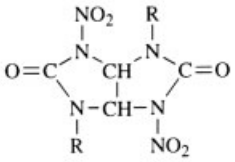

4

It is prepared by continuous nitration of glycoluril with 98 wt % HNO_3 at 30 – 60 °C in cascade-type reactors.

Tetranitroglycoluril (sorguyl ; **4**, R = NO_2) prepared by nitration of DINGU with a mixture $HNO_3 - N_2O_5$, has a high density (2.01 g/cm^3), and detonation velocity (9300 m/s). Only one crystalline form is known. It must be coated or plastic bonded because it is sensitive to hydrolysis [70], [74], [72].

Hexanitrazaisowurtzitane (HNIW, CL-20). One way to obtain explosives of high crystal density is the synthesis of cage compounds such as HNIW, which was discovered in 1986 by A. Nielsen[98].

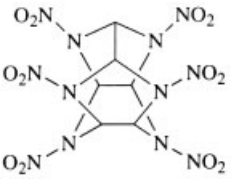

It is prepared by the reaction of benzylamine with glyoxal, followed by debenzylation and nitration. It exists in six crystalline forms [92], [93]; the ε-form has a density of 2.04 g/cm^3. Production has begun in pilot plants [91]. It can be used as a high explosives and in propellant formulations.

Oxynitrotriazole [*932-64-9*], ONTA, NTO, 3-nitro-1,2,4-triazol-5-one, was first reported in 1905, but its use as a low-vulnerability explosive was only discovered in 1979 [96],[73]. It has a high density (1.91 g/cm^3), good detonation velocity, low sensitivity to impact and friction, and good thermal stability (> 200 °C) [97].

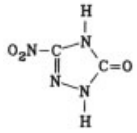

It is manufactured in two steps: reaction of the semicarbazide hydrochloride with formic acid to give the triazolone, which is nitrated with nitric acid. The NTO is recrystallized from water. NTO is used in LOVA formulations, some of which were tested in a joint LOVA programme of the USA, UK, Germany, and France.

Ammonium dinitramide (ADN), $NH_4^+N(NO_2)_2^-$, ϱ 1.83 g/cm^3, *mp* 92 °C, is a carbon-free ammonium salt that was discovered and developed in the former Soviet Union in the 1960s and has in the mid 1990s been disclosed [99]. ADN is almost as hygroscopic as ammonium nitrate and very insentive. It is of interest as a component of certain explosive mixtures and as a clean propellant.

6. High Explosive Mixtures

Explosives are used for global or directed destruction (bombs, torpedoes, mines, reactive armors, and warheads) or for pyrotechnics. Among pure compounds, only TNT and trinitrophenol are sufficiently insensitive to be loaded in large quantities by casting. Small quantities of other pure compounds are sometimes used for detonators or cutting or transmitting cords.

For flexible wrapped cords, the pure explosive (e.g., PETN) flows continuously through a funnel to form a flowing stream that is wrapped in a thin plastic band. Thread is immediately woven on the plastic-enclosed explosive, and this is again encased with plastic by drawing through a die, which also helps adjust the loading density.

For metallic cords, the high explosive (PETN, RDX, HMX, or HNS) is introduced into a metallic tube (lead, copper, or silver), which is drawn through dies until the desired size and cross section are obtained.

Loading Processes. In casting processes, the explosive is employed as a solution or a liquid suspension; it is cast into the shell or in the mold, where it crystallizes on cooling (physical process) or solidifies by the cross-linking of a polymer (chemical process). In pressing processes, the explosive is introduced in granular form into a mold or shell and pressed with a piston at a pressure of 10 – 300 MPa; conditions depend on the munitions size, the binder, and the required mechanical properties. Special shapes are prepared by laminating or calendering.

Flexibility is increased by the use of blends, which can enhance explosive performances while preserving safety and reducing cost. Special effects, such as delayed detonation, are possible with improved safety and high-temperature stability.

6.1. Desensitized Explosives

Because of their sensitivity and high melting points, RDX, PETN, and HMX must be desensitized before casting or pressing. This is achieved by coating the particles with wax, sometimes with addition of graphite as a lubricating and antistatic agent. The beeswax used in the past has been replaced by paraffin or synthetic waxes, which give more reproducible results. The wax content in the explosive mixture is usually 2 – 10 wt %.

In the simplest process, the particles are coated in hot water. The explosive (RDX, for example) is vigorously stirred in water and heated slightly above the melting point of the wax, which is introduced and dispersed on the particles. The mixture is cooled by the rapid introduction of cold water, without formation of crusts on the vessel walls. The suspension is filtered and dried. Coating agents melting above 100 °C are applied in a solvent (e.g., butyl acetate), which is removed by steam distillation. If the explosive is sensitive to hot water, the solvent is removed by filtration, possibly preceded by evaporation under vacuum. These processes are also used to prepare ternary mixtures of explosive aluminum and wax.

Detonation velocities can be obtained close to those of the pure explosives; e.g., for a mixture of 94.5 % RDX, 5 % wax, and 0.5 % graphite, $D = 8600$ m/s at $\varrho = 1.76$ g/cm^3 ($D = 8850$ m/s at $\varrho = 1.82$ g/cm^3 for pure RDX). However, large loadings are difficult to carry out because of problems with homogeneity, brittleness, porosity, and cost. The process is very useful for automatic filling of press-loaded small munitions.

6.2. TNT Mixtures

Molten TNT (*mp* 80 °C) is compatible with many explosives and other products, providing a castable melt above 80 °C. Among the products most frequently added are ammonium nitrate, to obtain very cheap munition loadings (amatols), and nitramines (RDX or HMX) or other high explosives (PETN, NTO), to improve performance. Mixtures containing 50 – 75 wt % RDX are used in bombs and warheads such as shaped charges. Some castable blends contain aluminum powder (tritonals, ammonals, Torpex, and HBX 1 and 3).

Wet and dry processes are employed. The former, the same as that used for desensitizing with low-melting wax, has cost and safety advantages. Almost the entire process is carried out under water. In France, RDX coated with 10 wt % TNT is produced in this manner, because its transportation as a dry product is allowed. However, air bubbles on the RDX crystals can create difficulties in loading.

In the dry process, the components are added dry to molten TNT with stirring, and the mixture is cast in shells. Wet RDX can be added to TNT at ca. 95 – 100 °C; the water is removed by decantation and evaporation.

Blends of TNT are used in the large-scale production of munitions. Shaped charges can be obtained by controlled sedimentation of the molten mixture. Crack formation in TNT during cooling can be controlled by the addition of 1 % HNS.

Explosive performance is restricted by the condition that the mixture must be pourable; for example, for a blend of RDX: TNT of 70: 30 (wt %), $D = 8060$ m/s at $\varrho = 1.73$. (Since $\varrho_{\mathrm{theor}} = 1.765$, the porosity is ca. 2 %.)

For blends with higher RDX or HMX content, a special apparatus such as a porous piston is used. Because of the danger of detonation by fire, mixtures of RDX with TNT are now forbidden on ships by some countries.

6.3. Plastic-Bonded Explosives (PBX)

Two different procedures are followed which improve the performance of new munitions.

Desensitized Explosives. Some new warheads require improved mechanical properties and stability at high temperature. Waxes are replaced by thermoplastics or by other plasticized polymers (nitrocellulose).

The manufacture of these plastic-bonded explosives resembles the use of waxes dissolved in a solvent immiscible with water. The molding powder is pressed, often in a heated and evacuated mold; for large charges, machining is necessary to obtain homogeneous charges of complex shapes with good mechanical properties at high temperature.

Examples include RDX or HMX boosters bonded with polyamide and mixtures of HMX and TATB containing 10 wt % fluorinated polymer [95]. Use is limited by the high cost of the pressing machines and the length of the processing cycle, especially for large charges.

Castable Explosives [75], [76]. For larger charges, mixtures are used of an explosive with a liquid binder, which is then cross-linked (composite explosives). A similar process is used for composite propellants. Composite explosives are being developed for uses requiring high accuracy and low vulnerability (LOVA explosives).

The liquid binder is introduced first into a mixer equipped with S-blades. A mixture of a diisocyanate and a diol (hydroxy-terminated polyether or polybutadiene) is usually employed in a precise ratio, with a precise quantity of a cross-linking agent such as a triol. The dry solid explosive (RDX, HMX, or PETN) is then added and mixed under vacuum with additives (e.g., surfactants) and a catalyst providing a pot life, i.e., the mixture remains pourable, of a few hours. During this time the mixture is cast, sometimes under pressure or by injection, in vibrating and evacuated molds or shells. The molds are cured in an oven for a few days. Two different compositions may be cast, one over the other (bicompositions), to produce special detonation waves.

The main advantage of these formulations is their low sensitivity, especially against bullets, impact, and friction, and the difficulty of the transition from combustion to detonation. The mechanical properties can be excellent, and thermal stability is fair.

Among the drawbacks are the slow curing and high cost when a multimodal particle-size distribution is used to increase the explosive content. The detonation velocity is diminished by the inert character of the binder; it can be estimated by the Urizar formula [7, pp. 8 – 10], [77]. Energetic binders improve the detonation velocity, but increase sensitivity. A review of energetic binders is given in [15], [79]. The use of thermoplastic elastomers as binders is of interest for environmental reasons since they facilitate the dismantling of ammunition.

7. Industrial Explosives

For more than 350 years explosives have been employed to mine ores and minerals. The annual consumption of explosives by each of the largest Western European consuming countries is around 40 000 t. World annual consumption of industrial explosives is at least 5×10^6 t, 75 % of which is ammonium nitrate – fuel oil (ANFO).

During the first 250 years of this period, only black powders were known and used, but fundamental changes occurred in the 1860s (invention of dynamite and blasting cap by Alfred Nobel), 1950s (ANFO), and 1980s (emulsions). The search continues for less expensive products and safer techniques for production and field use in mines, quarries, and road, tunnel, and dam construction. At the same time, the introduction of new products is restrained by the cost of existing investments and by safety and environmental regulations.

7.1. Dynamites

Gelatine dynamites are powerful explosives whose main ingredient is nitroglycerin. They

were invented opportunely to replace black powder in the construction of railroad and tunnels and gave great impetus to the use of explosives. Although diminished in importance since the 1950s and progressively replaced by new types of explosive, dynamites are still widely used because of their excellent qualities.

A detonation velocity range from 4300 to 7500 m/s provides high brisance. Sensitivity to initiation by cap or detonating cord is very good, as is the flashover coefficient. Density (ca. 1.5 g/cm^3), water resistance, and detonation pressure are high.

The composition of dynamites has changed only slightly since their origin. Nitroglycerin, gelatinized by nitrocellulose, has been totally or partially replaced by nitroglycol to reduce cost and permit lower operating temperatures. However, nitroglycol is more toxic than nitroglycerin, and its use is limited in some countries. Oxidizers such as ammonium nitrate reduce the cost of low-energy formulas. Combustible components (wood meal, peat, silicon, and aluminum) and other additives (sodium nitrate, ammonium chloride, and sodium chloride), provide formulations with specialized properties.

The consistency of such mixtures varies according to the nitroglycerin – nitroglycol content. Gelatin dynamites containing 20 – 40 wt % of nitroglycerin – nitroglycol form a plastic paste that may be made into cartridges after cutting or extrusion. Dynamites containing 10 – 20 wt % nitroglycerin – nitroglycol are powdery, and tampers are required to make cartridges. They have been largely replaced by the new explosives, except for use in coal mines, where the potential presence of dangerous dust and gas means that only very safe explosives with the lowest possible detonation temperature should be used. Permissible explosives for these applications usually contain a large amount of ammonium nitrate and sodium chloride, the endothermic melting of which absorbs much of the energy released by the detonation. Degrees of safety in coal mines are defined by national regulations. In France the classes are *rocher*, *couche*, and *couche améliorée*. The first refers to less safe explosives, permitted for underground use only. Dynamites are marketed in cylindrical paper, cardboard, or plastic cartridges. The characteristics of the leading products of Nobel Explosifs France are given in Table 4.

7.2. Ammonium Nitrate Explosives

Ammonium nitrate (AN) explosives were developed in European countries to replace explosives made with chlorates. They appeared later in the United States (Nitramon, DuPont, in 1935).

The French AN explosives, are made with ammonium nitrate (ca. 80 wt %) sensitized with a nitro compound (TNT, PETN, or their mixture, called pentolite). Formulas include aluminum powder and additives. These products are characterized by their good production safety and low production cost. They show good sensitivity when set off by caps or detonating cords. However, they are sensitive to humidity. Properties of the representative product Securex 80 are as follows:

Composition, wt %	
Ammonium nitrate	81.5
High explosive	16
Aluminum	0
Mass energy, kJ/g	3.51
Detonation velocity, m/s	4800
Flashover coefficient, cm	2
Detonating pressure	6.34
Density, g/cm^3	1.10

Ammonium nitrate explosives made in the United States contain dinitrotoluene but no nitro high explosive. They are produced by blending and as a result are not cap sensitive. The U.S. Government, followed by other governments, has established a category with relaxed rules for the transport and storage of these explosives. This category was first called NCN (nitrocarbonitrate) and is now known as blasting agents. They have been progressively replaced by water gels and emulsions.

7.3. Ammonium Nitrate – Fuel Oil Explosives

Ammonium nitrate – fuel oil (ANFO) explosives have been developed since 1955. They are made of porous prills of ammonium nitrate (ca. 94 wt %) soaked in mineral oil, usually domestic fuel oil (ca. 6 wt %). In some cases, aluminum powder is added to increase explosive strength.

Ammonium nitrate – fuel oil explosives are ordinarily used in bulk, although some countries

Table 4. Properties of dynamites

Property	F16	F19	GDC 16
Nitroglycerin/nitroglycol, wt %	32	40	12
Classification	gelatin	gelatin	powder
Application	general and underground	general and underground	permissible
Mass energy, kJ/g	3.97	4.15	1.78
Detonation velocity, m/s	6000	6500	2200
Flashover coefficient ∗, cm	8	10	10
Cartridging density, g/cm^3	1.45	1.45	1.10
Detonation pressure ∗∗	13.05	15.32	13.31

∗ Maximum distance at which a cartridge of diameter 30 mm and weight 50 g has a 50 % probability of iniating another cartridge.
∗∗ Calculated with the formula $\frac{1}{4}\,AD^2$.

limit their use to cartridges. They can be made on-site in a mobile unit, often mounted on a truck (ANFO truck). They are inexpensive and safe to handle, but low in strength, detonation velocity, and detonating pressure. In addition, they cannot be used in the presence of water. Because of their very low sensitivity, they require powerful primer charges, powerful detonating cords, dynamite relay cartridges, and boosters. The densities are low. Properties are listed in Table 5.

Table 5. Properties of ANFO explosives for general applications

Property	N 135	D7 fuel
Composition, wt %		
Ammonium nitrate	91	94.3
Fuel oil	4	5.7
Aluminum	5	0.0
Mass energy, kJ/g	3.43	2.74
Detonation velocity, m/s	3900	3700
Detonating pressure	3.42	2.84
Density, g/cm^3	0.9	0.83

7.4. Slurries and Water Gels

A new class of industrial explosives, known as water-explosives has been developed since 1956; they contain no nitroglycerin and are based on a solution of nitrates, thus containing considerable water (10 – 15 wt %). During the 1970s and 1980s, water explosives have took over much of the market, at the expense of dynamites and AN powders, because of their lower production cost and greater safety during production and handling.

Water explosives are used in cartridges or in bulk form. The bulk explosives are poured or pumped into the blast holes. Cartridges are formed with a continuous through-circulation device (CHUBPACK) in cylindrical plastic envelopes closed by metal clips.

Slurries and water gels are made of aqueous solutions of ammonium nitrate and sodium or calcium nitrate gelled by the addition of guar gum or cross-linking agents. They are sensitized by nitro explosives, organic amine nitrates, or by paint-grade aluminum. Combustible materials, such as aluminum, urea, sugar, or glycol, are mixed with these solutions either continuously or discontinuously.

The presence of reactive aluminum, water, and ammonium nitrate requires careful control to avoid chemical side reactions. The properties of water gel explosives and slurries marketed in France are listed in Table 6.

Table 6. Properties of water gel and slurry explosives for general applications

Property	Hydrolite AP, in bulk	Gelsurite 3000, in cartridges
Composition, wt %		
Nitrates	46	61.6
Water	15	13
Aluminum	0	15
High explosive	27	0
Mass energy, kJ/g	2.70	3.87
Detonation velocity, m/s	5100	4000
Flashover coefficient, cm	0	5
Detonating pressure	9.75	4.80
Density, g/cm^3	1.50	1.20

7.5. Explosive Emulsions

Since the cost of sensitizer for water gels is high, other formulas based on nitrate solutions but containing cheaper raw materials have been investigated. This led in 1962 to “water-in-oil” and “oil-in-water” emulsions, whose sensitivity

is due to the presence of air bubbles; these are most efficiently introduced by means of hollow glass bubbles or chemical gassing techniques.

Emulsions are produced, like water gels, by a continuous or a batch process and can be used in cartridges or in bulk form. Although the emulsions are neither gelled nor reticulated, their storage life has been considerably improved by using sophicated emulsifiers.

Emulsions differ from nitrate-fuel oil, nitrate, and water gels in their higher detonation velocity. The properties of typical emulsion-type products of Nobel Explosifs, France are listed in Table 7. The emulsions are sensitized by a chemical gassing process, apart from Iremite 2500s, the formulation of which includes microspherical glass bubbles.

7.6. Uses

The energy provided by explosives is used in a number of ways, including explosive cladding (see Section 3.3), metal working (forming, welding, and cutting), and shearing by pyrotechnics systems. However, the principal nonmilitary use of explosives is in mining. This application is governed by extensive safety regulations and is subject to competition from increasingly powerful mechanical means of extraction, such as mining machines, rippers, and tunnel-boring machines. Although the use of explosives entails such drawbacks as noise, vibration, and smoke, the technique is flexible, easy to use, and low in capital and material cost.

When using bulk explosives, cost is reduced because loading in cartridges is unnecessary. In addition, they fill the entire blast hole, allowing the possibility of enlarging the drilling pipe. These explosives make up half of those used in Europe, and over three-fourths of those used in the United States. They can be produced on-site in mobile units (ANFO truck, mix pump truck), which, however, are usually not economical for small-scale European operations. It is also possible to sensitize the bulk explosives just before filling in the blast holes. In this case, the unsensitized product is not entirely subject to the regulations for producing, handling, transporting, and storing dangerous materials.

The safety of explosives has been significantly improved by the use of nonelectric detonating devices with low-charge detonating cords or electronic detonators. The use of computers and sequential blasting machines has optimized firing systems, led to efficient use of explosive, and reduced environmental impacts such as noise and vibration.

8. Test Methods

Tests of secondary explosives are designed to study detonation phenomena, to approve a product, and to verify that manufactured lots meet the requirements (i.e., quality control).

Detonation phenomena (10^{-9} to 10^{-12} s) are observed by laser interferometry, streak cameras, and X-ray flashes, all electromagnetic methods. The very high instantaneous pressures (10 – 40 MPa) are measured, e.g., with low-impedance piezoresistive manganin gauges (Cu–Mn–Ni alloy) [78] or with fluorinated piezo polymers as shock sensors [101]. Phenomena are modelled based on Becker – Kistiakowsky – Wilson (BKW) [32] or Jones – Wilkins – Lee (JWL) [33] equations developed in the United States at the Los Alamos Scientific Laboratories (LASL) and the Lawrence Livermore National Laboratory (LLNL).

Tests to approve a product measure detonation velocity and pressure, energy released, ability to initiate or transmit detonation, and critical diameter. They also determine the safety characteristics in use, storage, and transportation.

The verification of a lot is usually based on the determination of physical properties (melting point, particle size, specific area) and chemical analysis. Safety and storage characteristics must match those of the approved product.

Codified tests for military use are published in the United States as military specifications (MIL), in France as *Manuel des modes opératoires,* in the Federal Republic of Germany as VTL, in the United Kingdom as DEF. STAN, and at NATO as STANAG. Japan issues *Japanese Industrial Standards* [80]. Unified and standardized European explosives tests (EXTEST) are published in *Explosifs* (Belgium, 1960 – 1970), in *Explosivstoffe* (FRG, 1971 – 1973), in *Propellants and Explosives* (FRG, 1977 to 1981), and in *Propellants, Explosives, Pyrotechnics* (1982 to date) (see also [7], [10]).

Table 7. Properties of explosive emulsions used for general and underground applications

Property	Iremite 1000 in cartridges	Iremite 4000 in cartridges	Iremite 2500s in cartridges	Gemulsite 100 for pumping
Mass energy, kJ/g	3.34	3.92	3.51	3.49
Detonation velocity m/s	5300	5300	5500	5100
Flashover coefficient, cm	>5	>5	5	
Detonating pressure	8.43	8.43	9.08	7.80
Density, g/cm^3	1.20	1.20	1.20	1.20

8.1. Performance Tests

Detonation Velocity. The *Dautriche method* is based on a comparison of the velocity in two parts of a circuit, composed in one section of a known detonating cord and in the other of a cartridge made with the test sample. The detonation wave fronts run in opposite directions in the two segments and collide at a precise point [81].

More sophisticated electrical or optical methods are also available. For example, electrical wires (switches), inserted into a cylindrical cartridge at precise locations, are short-circuited by the passage of the detonation wave, giving a signal that is recorded on a chronograph. Other methods use streak cameras. The detonation wave is built up at a distance from the initiation point, and this occurs only if the cartridge diameter is large enough. Therefore, this test requires an adequate quantity of explosive.

Energy Output [5], [81]. In the *lead block test*, strength or CUP (*coefficient d'utilisation pratique*) is measured by the expansion of a cavity in a lead block caused by the detonation of a sample of explosive, in comparison with a standard (PETN, picric acid, or TNT).

In the *ballistic mortar test,* the energy released by detonation gases is measured in a steel mortar consisting of two cavities: the first contains the explosive, and is connected to the second, which contains a projectile. The assembly is suspended by a pendulum, and when the projectile is driven out by the detonation, the recoil moves the mortar to an angle that is compared to that given by a standard explosive.

The *Kast method* is used to evaluate explosive brisance, a value related to the product ϱD^2 where $\varrho =$ density and $D =$ detonation velocity. The crushing of a copper cylinder by the detonation is compared to that obtained with a standard explosive.

The *cylinder test* has been studied at the Lawrence National Laboratory, Livermore, California, [82]. The velocity transmitted by a cylindrical rod of explosive to a tightly fitting copper tube is measured by a streak camera.

For more routine measurements, the *Défourneaux test* can be used (push-plate test). A metal plate lies on an explosive plate, which is detonated by a linear wave generator. During the detonation, the metal plate is continuously bent to an angle, which is observed, for example, by X-ray flash [83]. Among other tests is underwater explosion, with measurements of the bubble produced.

Other Performance Tests. In one of the critical diameter tests, the detonation propagation of metal cords (copper or silver) filled with an explosive of known density is observed as a function of the inner diameter. In another method, a conical charge is used.

Other tests include the determination of the coefficient of transmission of detonation, the sensitivity to initiation, the self-excitation coefficient (maximum distance between two cartridges loaded with the same explosive, allowing the initiation of one by the other), and the plate-dent test.

8.2. Safety

Stability. Aging effects are frequently determined by the *vacuum stability test*, in addition to the usual chemical methods. It can be used to study the compatibility between explosives and other materials.

The sample is placed in a test tube fitted with a manometric capillary glass tube; the other end of the tube dips into a cup of mercury. The tube is evacuated and the sample is heated, usually

between 80 and 130 °C. The variation of the inner pressure in the tube is shown by the mercury level. Pressure usually increases during the first few hours with the release of traces of moisture and volatile substances. Afterward, during a period of 100 h, the rise in pressure can be linear (constant decomposition rate) or more rapid (autocatalytic decomposition).

Sensitivity. The *impact sensitivity* of an explosive is the minimum energy that causes quick decomposition [5], [13], [84]. The test conditions strongly influence the results, which (given in Joules) are only orders of magnitude. The classification of explosives depends on the apparatus and the test conditions.

In the Bruceton test for impact sensitivity, the stimulus after a positive result is lowered by one unit and raised after a negative result. The value of the sensitivity is obtained as the mean of 30 – 50 trials. Probabilistic values are also determined.

The procedures differ in the size and placement of the sample and the use of tools exposed or covered with sandpaper. The reaction may be followed visually, aurally, audiometrically, or by examination of gas evolution.

Tests are prescribed by the U.S. Bureau of Mines, Picatinny Arsenal (United States), Explosives Research Laboratory (United States), BAM (Federal Republic of Germany), Julius Peters (Federal Republic of Germany, France), Rotter (United Kingdom, Canada), and Aubertein (France). Another test is the Susan projectile-impact test, where a projectile containing ca. 450 g of explosive is directed against a hard target at selected velocities.

The *friction sensitivity test* is even more difficult to reproduce than the impact sensitivity test. A thin layer of sieved explosive on a fixed surface is rubbed by a hard or abrasive surface moved by translation (Julius Peters apparatus) or by rotation (U.S. Bureau of Mines) under adjustable pressure.

Impact and friction are combined in the skid test (Pantex) and the Popolato test; an explosive sphere is projected against a sand-coated target plate at defined impact angles.

In the *electrostatic sensitivity test*, a small quantity of explosive is placed between two electrodes. A capacitor discharge causes a spark to pass between the electrodes through the explosive. The sensitivity is expressed by the minimum energy required to decompose the sample.

In addition to differential thermal analysis (DTA) and thermogravimetric analysis (TGA), the following *heat sensitivity tests* are employed:

1) The sample is subjected to a constant increase of temperature and the decomposition or ignition temperature is noted; the result depends on the quantity of explosive and the rate of heating.
2) The sample placed in a small bulb is dipped into a Wood's metal bath maintained at a fixed temperature. The time to decomposition is noted, or the temperature at which a decomposition occurs after 5 s.
3) The sample is ignited in a gutter and the burning rate measured.
4) The sample is externally heated in a closed or half-closed vessel. The tendency to burn or to detonate is observed (steel-sleeve test).
5) The sample is heated at a given rate in a closed vessel. The time of detonation or the temperature at which the sample detonates after a given time is noted (cook-off test).

To test the *sensitivity to initation* through a spacer (card-gap test), a sample in a steel tube is subjected to the detonation of a standard booster explosive separated from the sample by a barrier of cellulose acetate cards of 0.19 mm thickness [8]. The minimum number (N^-) of cards that prevent transmission of detonation fixes the legal class of the explosive for transportation and other regulations.

To obtain a LOVA agreement (MURAT in France), the following tests must be carried out on ammunitions: bullet and fragment impact, slow and fast cook-off tests, reaction towards a shaped-charge jet, and sympathetic reactions. The results depend on storage of the ammunitions, their design, and on the properties of the propellants and explosives they contain [94], [19, 20]. These investigations are supported by the NATO Insensitive Munitions Information Center (NIMIC, Brussels).

9. Legal Aspects and Production

Governments regulate the possesion, production, storage, packaging, shipping, and trading

of explosives. Periodic inspections and labeling prevent theft. Tagging the explosive permits tracing its origin.

9.1. Safety Regulations

Originally, each country established its own regulations; the standardization recommended by the United Nations was published in 1984 [85]. Substances are classified according to the type of risk involved, unless they are too dangerous to ship. Explosives are in Class 1, which contains five subdivisions [85]:

1) Substances and articles that present a bulk explosion hazard
2) Substances and articles that present a projectile hazard but not a bulk explosion hazard
3) Substances and articles that present a fire hazard and a minor blast and/or projectile hazard without a bulk explosion hazard
4) Substances and articles that present no significant hazard
5) Insensitive substances that present bulk explosion hazard, but very little probability of initiation or transition from burning to detonation under normal transport conditions

Class 1 is restricted, containing only registered substances; for new substances, test results must be presented. Simultaneous transport and storage are permitted for "compatibility groups".

The transportation regulations specify packaging, labeling, an identifying insignia on vehicles, compatibilities, and maximum load and safety devices on vehicles.

Manufacture and storage are governed by the TNT equivalency, which is defined as the quantity of TNT producing the same effects as 100 kg of the explosive [86].

The expression $d = k\,Q^{1/3}$ gives the safe distance from a load Q of explosive (expressed as TNT equivalent), where the constant k depends on the presence of walls or other protection.

9.2. Production of Military Explosives

High explosives are manufactured by private companies in many countries, e.g., United Kingdom, Germany, Norway, Sweden, Switzerland, and Japan.

In the United States, research and development involving high explosives are often undertaken in military laboratories. Production is conducted by arsenals or by private companies under the GOCO system (Government owned – contractor operated), in which investments and factories are owned by the government. The product is government property and cannot be sold by the operating company.

In Spain production is shared between a national company and private ones. In France military secondary explosives are manufactured by a national company; munition loading is carried out by a national and private companies.

Production figures and prices are generally not available.

10. Toxicology and Occupational Health

Raw Materials. Both acids (HNO_3 and H_2SO_4) and benzene derivatives (toluene and chlorobenzene) are of concern [4]. The usual safety measures must be followed to protect eyes, skin, and respiratory tract. Red fumes of nitrogen oxides, formed normally or by accident, can induce lung edema two days after exposure; the use of a gas mask is recommended.

Explosives. Some *primary explosives* are toxic (mercury fulminate and certain lead salts), but the toxic hazards are far below those arising from high sensitivities to detonation. Among the secondary explosives, *nitroglycerin* and other liquid nitrates exhibit hypotensive action accompanied by headaches and, in chronic cases, by methemoglobinemia. A tolerance to nitroglycerin can be developed. These substances are usually absorbed through the skin rather than by inhalation. For nitroglycerin and glycol dinitrate, TLV = 1.5 mg/m^3. *PETN,* which like nitroglycerin is prescribed for heart diseases, has a very low vapor pressure and is not readily absorbed by the skin.

The toxic effects of *TNT* include blood changes, cyanosis, methemoglobinemia, and toxic hepatitis. It is introduced by inhalation, ingestion, and skin absorption; at inhalation level,

TLV = ca. 1 mg/m^3. Dust in the plant is to be avoided.

Tetryl can induce an allergic dermatitis,therefore its use is practically abandoned.

Both *RDX* and *HMX* have negligible vapor pressures. They do not penetrate the skin and their toxicity is very low. However, there have been reports of epileptiform convulsions, without liver involvement, or carcinogenic or mutagenic effects. Dust formation should be avoided and masks should be worn.

11. References

General References

1. *Ullmann,* 4th ed., **21,** 637 – 697.
2. T. Urbanski: *Chemistry and Technology of Explosives,* Pergamon Press, Oxford 1983 – 1985.
3. A. Marshall: *Explosives,* 2nd ed., P. Blakiston's Son and Co., Philadelphia 1917.
4. B. T. Fedoroff et al.: *Encyclopedia of Explosives and Related Items,* U.S. Army and National Technical Information Service, Springfield, Va., 1960 – 1983.
5. R. Meyer: *Explosives,* 2nd ed., Verlag Chemie, Weinheim-New York 1981.
6. *Kirk-Othmer,* **9,** 561 – 620, and **15,** 841 – 853.
7. B. M. Dobratz: *LLNL Explosives Handbook,* UCRL Report 52 997, Livermore, Calif., 1981. B. M. Dobratz: *LLNL Explosives Handbook, UCRL Report 52 997, Change 2,* Livermore, CA 1985.
8. J. Quinchon et al.: *Les explosifs,* 2nd ed., Technique et Documentation, Paris 1987.
9. J. Quinchon, R. Amiable, P. Chereau: *Sécurité et hygiène du travail dans l'industrie des substances explosives,* 2nd ed., Technique et Documentation, Paris 1985.
10. L. Médard: *Les explosifs occasionnels, Technique et Documentation Paris 1979, (engl. Transl.: Accidental Explosives),* Ellis Horward Ltd (Distrib. J. Wiley & Sons) 1987.
11. J. Calzia: *Les substances explosives et leurs nuisances,* Dunod, Paris 1969.
12. S. Fordham: *High Explosives and Propellants,* 2nd ed., Pergamon Press, Oxford 1980.
13. H. D. Fair, R. F. Walker: *Energetic Materials. Inorganic Azides,* Plenum Press, New York 1977.
14. *Houben-Weyl,* **X1 and XI2,** 99 – 116.
15. *Structure and Properties of Energetic Materials, Symp. Proceedings, vol. 296,* Materials Research Society 1992. *Composition, Combustion and Detonation Chemistry of Energetic Materials, Symp. Proceedings, vol. 418,* Materials Research Society, Pittsburgh, USA 1995.
16. J. E. Field, P. Gray, *Energetic Materials,* The Royal Society, London 1992.
17. Approches microscopique et macroscopique des detonations, Journ. de Physique, Suppl., 1er atelier int. (31. May – 07. June 1987). Approches microscopique et macroscopique des detonations, Journ. de Physique, Suppl., 2e atelier int. (02. – 07. October 1994).
18. Proc. Int. Symp. Energetic Materials, Paris, May 1994; *International Défense et Technologie,* Sept. 1994.
19. *Europyro, 5e Congres Int. de Pyrotechnie,* 1993. *Europyro, 6e Congres Int. de Pyrotechnie,* 1995. *Europyro, 7e Congres Int. de Pyrotechnie,* 1999.
20. *Insensitive Munitions Technology Symposium,* ADPA (American Defense Preparedness Association) 1996.
21. *Life Cycle of Energetic Materials, Conferences Proceedings (1993, 1994, 1996),* Los Alamos Lab.
22. G. A. Olah et al., *Nitration. Methods and Mechanisms,* VCH Publishers 1989. Useful periodicals treating explosives include *ICT Intern. Jahrestg.* (1970 to date); *Propellants and Explosives* (1976 – 1981), continued as *Propellants, Explosives, Pyrotechnics* (1982 to date); *Symposia on Detonation (United States)*; and *Explosion and Explosives* (translation of Kogyo Kayaku (published by the Industrial Explosives Society of Japan) that is published by the NSF, Washington).

Specific References

23. D. L. Chapman, *Phil. Mag.* **47** (1899) 90.
24. E. Jouguet, *J. Math. Pures. Appl. Paris* **60** (1905) 347.
25. I. B. Zeldovich, *NACA Techn. Memo.* no. 1261 (1950).
26. V. von Neumann, OSRD Report 549 (1942).
27. W. Döring, G. Burkhardt, *Air Material Command Technical Report* F-TS-1227-IA (GDAM A9 – T46) (1949).
28. F. Bauer, MRS Fall Meeting Boston 2001. Session "Electroactive Polymers".
29. *Détonique Théorique,* Edition de l'Ecole Nationale Sup. des Techniques Avancées, Paris 1981.

30. J. O. Hirschfelder et al.: *Molecular Theory of Gases and Liquids,* J. Wiley & Sons, New York 1954.
31. M. Cowperthwaite, W. H. Zwisler, *Symposium (International) on Detonation, 6th,* Office of Naval Research Report ACR-221 (1976).
32. C. L. Mader, Los Alamos Laboratory Report CA-2900 (1963).
33. E. L. Lee, H. C. Hornig, J. W. Kury, UCRL Report 50 422.
34. M. J. Kamlet, S. J. Jacobs, *J. Chem. Phys.* **48** (1968) 23.
35. L. R. Rothstein, *Propellants Explos.* **4** (1979) 56 – 60; **6** (1981) 91.
36. C. H. Johansson, P. A. Persson: *Detonics of High Explosives,* Academic Press, London-New York 1970.
37. A. F. Belyaev: *Combustion, Detonation and Explosive Work of Condensed System,* USSR Academy of Science, Moscow 1968.
38. A. K. Parfenov, A. Y. Apin, *ATD Press* **4** (1965) no. 113.
39. R. R. Bernecker et al., *Symposium (International) on Detonation, 7th,* Office of Naval Research Report NSWC MP 82 – 334(1981).
40. M. Samirant, *Symposium (International) on Detonation, 7th,* Office of Naval Research Report NSWC MP 82 – 334 (1981).
41. C. Fauquignon, H. Moulard, *Sciences et Techniques de l'Armement* **1** (1984) 27.
42. F. P. Bowden, A. D. Yoffe: *Initiation and Growth of Explosives in Liquids and Solids,* Cambridge University Press, 1952.
43. C. M. Tarver, J. O. Hallquist, *Symposium (International) on Detonation, 7th,* Office of Naval Research Report NSWC MP 82 – 334 (1981).
44. F. E. Walker, R. J. Wasley, *Explosivstoffe* **17** (1969) 9.
45. Y. de Longueville, C. Fauquignon, H. Moulard, *Symposium (International) on Detonation, 6th,* Office of Naval Research Report ACR-221 (1976).
46. D. Price, *Symp. (Int.) Combust. 11th* (1967) 963.
47. S. J. Jacobs, Naval Ordnance Laboratory Technical Report 74 – 86.
48. M. L. Wilkins, UCRL Report 90 142 (1983).
49. V. D. Linse, H. E. Otto, P. Pocaliko: "Explosion Welding," *Welding Handbook,* vol. 3, Am. Welding Soc., Miami 1980.
50. M. Défourneaux, *Sciences et Techniques de l'Armement,* Paris 2 (1970) 293.
51. J. Dubar, J. Calzia, *C.R. Seances Acad. Sci. Sér. C* **266** (1968) 1114 – 1116.
52. T. Urbanski: *Chemistry and Technology of Explosives,* **vol. 1,** Pergamon Press, Oxford 1983, pp. 139 – 151.
53. P. Aubertein, *Mém. Poudres* **30** (1948) 7 – 42.
54. M. J. Barbiere, *Mém. Poudres* **26** (1935) 294 – 302.
55. E. E. Gilbert et al., *Propellants Explos. Pyrotech.* **7** (1982) 150 – 154.
56. J. P. Konrat, FR 1 258 323, 1961; *Chem. Abstr.* **57** (1962) P 3358 h.
57. M. L. Kastens, J. F. Koplan, *Ind. Eng. Chem.* **42** (1950) 402 – 413.
58. K. G. Shipp, *J. Org. Chem.* **29** (1964) 2620 – 2623.
59. SNPE, FR 2 256 144, 1973 (J. M. Emeury, L. Leroux).
60. E. E. Gilbert, *Propellants Explos.* **5** (1980) 168 – 172.
61. S. E. Klassen [Sandia ABQ] 32th Int. Ann. Conf. ICT (2001) 24 (1 – 12).
62. T. M. Benziger, *ICT Intern. Jahrestg.* 1981, 491 – 503.
63. U. S. Energy R. and D. Adm., US 4 032 377, 1977 (T. M. Benziger).
64. G. Desseigne, *Mém. Poudres* **28** (1938) 156 – 170.
65. J. Issoire, G. Burlet, *Mém. Poudres* **39** (1957) 65 – 95; **40** (1958) 47 – 76.
66. H. H. Cady, A. C. Larson, D. T. Cromer, *Acta Crystallogr.* **16** (1963) 617 – 623.
67. H. H. Cady, *ICT Intern. Jahrestg.* 1986, 17, 1 – 12.
68. J. A. Bell, I. Dunstan, *J. Chem. Soc. C* 1969, 1556 – 1562.
69. V. I. Siele et al., *Propellants Explos.* **6** (1981) 67 – 73.
70. J. P. Kehren, *ICT Intern. Jahrestg.* 1976, 47 – 58. J. Boileau, J. M. Emeury, Y. de Longueville, P. Monteagudo, *ICT Intern. Jahrestg.* 1981, 505 – 526.
71. M. Stinecipher, L.A. Sketg, 8th Int. Symposium on Detonation (1985) 351 – 360.
72. V. P. Sinditskii, V Y. Egorshev, M. V. Berezin, 32th Int. Ann. Conf. ICT (2001) 59 (1 – 13).
73. M. D. Coburn, B. W. Harris, K. Y. Lee, M. M. Stinecipher, H. H. Hayden, *Ind. Eng. Chem. Prod. Res. Dev.* **25** (1986) 68 – 72.
74. E. Wimmer: "Dérivés nitrés du glycolurile", Thèse, Université d'Aix-Marseille, 1985.
75. G. Roche, *ICT Intern. Jahrestg.* 1978, 235 – 242.
76. L. R. Rothstein, *ICT Intern. Jahrestg.* 1982, 245 – 256.

77. C. Gaudin, *ICT Intern. Jahrestg.* 1978, 243 – 254; 1981, 541 – 551.
78. M. J. Ginsberg, A. R. Anderson, J. Wackerle, *Symposium sur les jauges et matériaux piezoresistifs, 1st,* Arcachon, France, 1981. L. M. Erickson et al., *ibid.*
79. J. Boileau, MRS Symposium Proceedings 418 (1996) 91 – 102.
80. I. Fukiyama, *Propellants Explos.* **5** (1980) 94 – 95.
81. H. Ahrens, *Propellants Explos.* **2** (1977) 7 – 20.
82. J. W. Kury et al., *Symposium on Detonation, 4th* (1965) 3.
83. M. Défourneaux, *Sciences et Techniques de l'Armement* **47** (1973) 723 – 814, 847 – 930.
84. K. N. Bascombe, R. M. H. Wyatt, *ICT Intern. Jahrestg.* 1976, 211 – 231.
85. United Nations Organization: *Transport of Dangerous Goods,* 3rd ed.,1984.
86. D. M. Koger, *ICT Intern. Jahrestg.* 1980, 549 – 560.
87. J. Ding, *The discovery of gunpowder and shockwaves in China – CETR rep. A 12–89,* New Mexico Tech., Soccoro, 1989.
88. A. B. Sheremetev, T. S. Pivina, *ICT Int. Jahrestg.* 1996, 30, 1 – 11.
89. FR 1060425, 1952 (C. Frejacques). J. P. Passarello, Thesis (ENSTA-Paris), 1996.
90. R. W. Millar et. al., *ICT Int. Jahrestg.* 1996, 26, 1 – 14.
91. R. B. Wardle et. al., *ICT Int. Jahrestg.* 1996, 27, 1 – 10.
92. M. F. Foltz, *Prop. Exp. Pyro.* **19,** 199463 – 69 and 133 – 144.
93. N. V. Chugehov, F. Volk, ICT Intern. Jahrestg. 2001, 101, 1–9.
94. A. Sanderoon (NIMIC-NATO), *ICT Int. Jahrestg.* 1996, 18, 1 – 8, and references given in this paper.
95. B. M. Dobratz, *TATB, Development and Characterization 1888 to 1994, LA 13014 H, Aug. 1995, UC 741,* Los Alamos Nat. Lab. (NM 87545).
96. K. Y. Lee et al., *J. Energ. Mat.* **5** (1987) no. 1, 27 – 33. A. Becuwe, A. Delclos, *ITC Int. Jahrestg.* 1987, 27, 1 – 14. M. Piteau, A. Becuwe, B. Finck, *ADPA, Compatibility of Plastics with Explosives, Proceedings,* San Diego 1991, 69 – 79.
97. A. Becuwe, A. Delclos, *Prof. Exp. Pyro.* **18,** 1993, 1 – 10.
98. A. Nielsen, *J. Org. Chem.* **55** (1990) 1459 – 1466.
99. V. A. Tartakowsky, O. A. Lukyanov, *ICT Int. Jahrestg.* 1994, 13, 1 – 9. A. Davenas, [18], 69 – 71. H. Hatano et al., *Europyro (AFP), 6*^{e} *Congres Int. Pyrotechnie Tours,* 1995, 23 – 26
100. J. Boileau et al., *Acta Cryst. C* **44** (1988) 696.
101. F. Bauer, H. Moulard, *4*^{e} *Symposium Hautes Pressions Dynamiques (AFP),* 1995 333 – 340. R. A. Graham, L. M. Lee, F. Bauer, *Proceedings of Shock Waves in Condensed Matter,* Elsevier SC Publ. 1988, 619.
102. B. Blaive et al., *Europyro (AFP), 6*^{e} *Congres Int. Pyrotechnic Tours,* 1995, 413 – 419.
103. H. Moulard, *9th Int. Symposium on Detonation (OCNR 113291–7),* 1989, 18 – 24.
104. L. Borne, *10th Int. Symposium on Detonation (ONR 33395–12),* 1993, 286 – 293.

Extraction → Liquid–Liquid Extraction
Extraction → Liquid–Solid Extraction

Fats and Fatty Oils

ALFRED THOMAS, Unimills International, Hamburg, Federal Republic of Germany

1. Introduction . . . 3
2. Composition . . . 4
2.1. Glycerides . . . 4
2.2. Fatty Acids . . . 4
2.3. Phospholipids . . . 8
2.4. Waxes . . . 9
2.5. Sterols and Sterol Esters . . . 9
2.6. Terpenoids . . . 9
2.7. Other Minor Constituents . . . 11
3. Physical Properties . . . 12
3.1. Melting and Freezing Points . . . 12
3.2. Thermal Properties . . . 14
3.3. Density . . . 15
3.4. Viscosity . . . 15
3.5. Solubility and Miscibility . . . 15
3.6. Surface and Interfacial Tension 17
3.7. Electrical Properties . . . 17
3.8. Optical Properties . . . 17
4. Chemical Properties . . . 18
4.1. Hydrolysis . . . 19
4.2. Interesterification . . . 19
4.3. Hydrogenation . . . 19
4.4. Isomerization . . . 20
4.5. Polymerization . . . 20
4.6. Autoxidation . . . 21
5. Manufacture and Processing . . 22
5.1. Vegetable Oils and Fats . . . 22
5.1.1. Storage and Handling of Raw Materials . . . 22
5.1.2. Cleaning and Dehulling . . . 22
5.1.3. Expelling . . . 22
5.1.4. Extraction . . . 24
5.2. Land-Animal Fats . . . 28
5.3. Marine Oils . . . 29
5.4. Synthetic Fats . . . 29
6. Refining . . . 29
6.1. Degumming . . . 30
6.2. Deacidification (Neutralization) 30
6.3. Bleaching . . . 33
6.4. Deodorization . . . 34
7. Fractionation . . . 36
8. Hydrogenation . . . 38
9. Interesterification . . . 40
10. Environmental Aspects . . . 42
11. Standards and Quality Control . 42
11.1. Sampling . . . 43
11.2. Raw Materials . . . 43
11.3. Oils and Fats . . . 44
11.3.1. Physical Methods . . . 44
11.3.2. Chemical Methods . . . 45
12. Storage and Transportation . . . 49
13. Individual Vegetable Oils and Fats . . . 50
13.1. Fruit Pulp Fats . . . 50
13.1.1. Palm Oil . . . 50
13.1.2. Olive Oil . . . 51
13.1.3. Avocado Oil . . . 52
13.2. Seed-Kernel Fats . . . 52
13.2.1. Lauric Acid Oils . . . 52
13.2.1.1. Coconut Oil . . . 52
13.2.1.2. Palm Kernel Oil . . . 53
13.2.1.3. Babassu Oil and Other Palm Seed Oils . . . 54
13.2.1.4. Other Sources of Lauric Acid Oils 54
13.2.2. Palmitic – Stearic Acid Oils . . . 55
13.2.2.1. Cocoa Butter . . . 55
13.2.2.2. Shea Butter, Borneo Tallow, and Related Fats (Vegetable Butters) . 56
13.2.3. Palmitic Acid Oils . . . 56
13.2.3.1. Cottonseed Oil . . . 56
13.2.3.2. Kapok and Related Oils . . . 57
13.2.3.3. Pumpkin Seed Oil . . . 58
13.2.3.4. Corn (Maize) Oil . . . 58
13.2.3.5. Cereal Oils . . . 58
13.2.4. Oleic – Linoleic Acid Oils . . . 59
13.2.4.1. Sunflower Oil . . . 59
13.2.4.2. Sesame Oil . . . 60
13.2.4.3. Linseed Oil . . . 60
13.2.4.4. Perilla Oil . . . 61
13.2.4.5. Hempseed Oil . . . 61
13.2.4.6. Teaseed Oil . . . 61
13.2.4.7. Safflower and Niger Seed Oils . . 62
13.2.4.8. Grape-Seed Oil . . . 62
13.2.4.9. Poppyseed Oil . . . 62
13.2.5. Leguminous Oils . . . 63
13.2.5.1. Soybean Oil . . . 63
13.2.5.2. Peanut Oil . . . 63
13.2.5.3. Lupine Oil . . . 64
13.2.6. Cruciferous Oils . . . 64
13.2.6.1. Rapeseed Oil . . . 64
13.2.6.2. Mustard Seed Oil . . . 65
13.2.7. Conjugated Acid Oils . . . 65
13.2.7.1. Tung Oil and Related Oils . . . 65
13.2.7.2. Oiticica Oil and Related Oils . . . 66
13.2.8. Substituted Fatty Acid Oils . . . 66
13.2.8.1. Castor Oil . . . 66

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_173

13.2.8.2. Chaulmoogra, Hydnocarpus, and Gorli Oils 66
13.2.8.3. Vernonia Oil 67
14. Individual Animal Fats 67
14.1. Land-Animal Fats 67
14.1.1. Lard 67
14.1.2. Beef Tallow 68
14.1.3. Mutton Tallow 69
14.1.4. Horse, Goose, and Chicken Fat . . 69
14.2. Marine Oils 70
14.2.1. Whale Oil 70
14.2.2. Fish Oil 71
15. Economic Aspects 72
16. Toxicology and Occupational Health 73
17. References 73

1. Introduction

Naturally occurring oils and fats are liquid or solid mixtures consisting primarily of glycerides. Depending on whether they are solid or liquid at ambient temperature, they are referred to as *fats* or *oils*, respectively. Naturally occurring oils and fats always contain minor constituents such as free fatty acids, phospholipids, sterols, hydrocarbons, pigments, waxes, and vitamins. The nomenclature rules for glycerides have been summarized [18].

History. A century ago ecological, religious, and social factors still played a more important role than technology in the choice and utilization of oils and fats.

Primeval humans utilized animal fats, making cheese and butter from goat's milk. Oilseed plants were cultivated during the neolithic period. Poppy seeds have been found in remains of Bronze-Age bread; rapeseed and linseed, together with millstones, have been found in Bronze-Age dwellings.

Linseed, almonds, and sesame seed were part of Egypt's natural flora. Sesame oil had mythical significance. The oil-bearing safflower plant is still grown in Egypt. Olive oil came from Palestine, Syria, and Crete. The Phoenicians and Greek colonists introduced the olive tree to Sicily and Italy. Cotton is one of the oldest cultivated plants; it was grown 2600 years ago in India as a source of both oil and fiber. Soy and hempseed are mentioned as oilseed plants in a Chinese document of 2838 B.C.

The oilseeds were ground with a pestle and mortar or between stones. Simple mills of the type still being used in some developing countries – a concave stone rotating on a convex one – also evolved. The Egyptians developed the sack or expeller press. The Greeks and Romans used a grinding device known as the "trapetum". In North Africa, mechanical presses were used in processing plants that approached the size of modern factories.

The processing of oil fruits and seeds in Central and Northern Europe advanced more slowly. Oilseeds, primarily linseed, hempseed, and rapeseed, were pulped in hollowed stones, and the oil was expelled from the pulp by pressing between two cloth-covered frames. This domestic-type process was practiced up to the 16th century. Industrial oil milling developed primarily in regions where linseed was grown extensively. A wood engraving dated 1568 depicts an oil mill with a horse-drawn vertical millstone on a stone bed. The ground oilseed was heated in a kettle over an open fire and finally "beaten" in wedge presses.

Animal fats were obtained by rendering fatty tissue and by churning cream. Up to the middle of the 19th century, tallow and butter were the most important edible fats in Europe, lard and vegetable oils playing only a minor role.

Toward the end of the 19th century the production of oil by hydraulic pressing and solvent extraction was introduced. This process gave relatively high oil yields but necessitated posttreatment of the oils by neutralization, bleaching, and deodorization.

The invention of margarine in the 1870s gave further impetus to the oil-processing industry. With the discovery of oil hydrogenation (hardening) at the beginning of the 20th century, liquid oils could be converted into spreadable, consistent fats. In the 1930s, interesterification and fractionation were developed as further methods to modify the consistency of oils and fats.

Apart from being used for edible purposes, oils and fats are referred to in the Old Testament as cosmetic products and lamp fuel. Anointing with oil symbolized royal dignity. HOMER and HERODOTUS refer to the use of fats as "process-

ing aids" during weaving. The ability of fats to calm waves was studied by Indian scholars 3000 years ago. The Egyptians supposedly used fats as lubricants to transport stone blocks. They were also familiar with the use of drying oils in varnishes and paints.

The elucidation of the chemical nature of fats was initiated by SCHEELE, who produced glycerol from olive oil around 1780. CHEVREUL subsequently (ca. 1815) recognized that fats were predominantly esters of fatty acids and glycerol.

Modern research and development is focussed on the application of new biotechnological principles to the production and modification of oils and fats [19], as well as on nutritional aspects, problems of trace contaminants, and environmental pollution. The breeding of new plant varieties is an important method for increasing the types of oils and fats. Examples are new varieties of rapeseed (low erucic acid content), safflower, and sunflower (high in oleic acid). New sources of oils and fats are being exploited by cultivation of wild plants such as the jojoba and *Cuphea* shrubs.

2. Composition

2.1. Glycerides

Naturally occurring fats contain about 97 % triglycerides (triacylglycerides), i.e., triesters of glycerol with fatty acids; up to 3 % diglycerides (diacylglycerides); and up to 1 % monoglycerides (monoacylglycerides). Tri-, di-, and monoglycerides consist of 1 mol of glycerol esterified with 3 mol, 2 mol, or 1 mol of fatty acid, respectively. The triglycerides of naturally occurring oils and fats contain at least two different fatty acid groups. The chemical, physical, and biological properties of oils and fats are determined by the type of the fatty acid groups and their distribution over the triglyceride molecules. The melting point generally increases with increasing proportion of long chain fatty acids or decreasing proportion of short chain or unsaturated fatty acids. Milk fat (butterfat) and coconut oil, which contain a high proportion of C_6–C_{12} fatty acids, have lower melting points than fats such as tallow and lard, which contain predominantly C_{16} and C_{18} fatty acids. Vegetable oils are liquid at ambient temperature because of their high proportion of unsaturated fatty acids.

The properties of a triglyceride are also determined by the position of the various fatty acid groups in the triglyceride molecule (i.e., 1-, 2-, or 3-position). The total number N of possible triglycerides (including positional isomers) from x different fatty acids is

$$N = \frac{x^2 + x^3}{2}$$

However, the proportions of different triglycerides in a naturally occurring fat generally do not conform to a statistical distribution. In vegetable oils and fats, unsaturated fatty acids are linked preferentially to the 2-position of the glycerol group, whereas in animal fats they appear primarily in the 1- and 3-positions ("2-random" and "1,3-random" distributions) [20]. Extreme examples of nonrandom distributions of fatty acid groups over the triglyceride molecule are cocoa butter (ca. 40 % 1-palmito-3-stearo-2-olein) and lard (ca. 20 % 2-palmito-1,3-diolein).

2.2. Fatty Acids

The fatty acids that form the triglycerides of naturally occurring oils and fats are predominantly even-numbered, straight-chain, aliphatic monocarboxylic acids with chain lengths ranging from C_4 to C_{24}. Unsaturated fatty acids differ in number and position of double bonds and in configuration (i.e., cis or trans isomers). The more common fatty acids are known by trivial names such as butyric, lauric, palmitic, oleic, stearic, linoleic, linolenic (→ Fatty Acids). Crude oils contain significant amounts of free fatty acids.

The chief fatty acids in some commercial oils and fats are listed Tables 1 (saturated fatty acids) and 2 (unsaturated fatty acids). The fatty acid composition of most vegetable oils and fats is relatively simple; they consist predominantly of palmitic, oleic, and linoleic acids [20]. The fatty acids of land-animal fats mainly have a chain length of C_{16} or C_{18}. They are formed by biosynthetic conversion of carbohydrates, proteins, or fats, or originate directly from ingested fat.

Ruminant fats contain 5 – 10 % trans fatty acids, which are produced from linoleic and

Table 1. Saturated fatty acids in various oils and fats, (main sources: Unilever; Food RA, Leatherhead; ITERG [a])

	Saturated fatty acids [b], g/100 g fatty acids							
	C_{10} and lower	C_{12}	C_{14}	C_{16}	C_{18}	C_{20}	C_{22}	C_{24}
Liquid vegetable oils								
Almond			tr	6.5–7	1–2.5	tr	tr	
Avocado			tr	10–26	0.5–1			
Corn germ		tr–0.5	tr–0.3	9–12	1–3	ca. 0.5	tr–0.5	<0.5
Cottonseed		tr	0.5–2.0	21–27	2–3	<0.5	tr	tr
Grape-seed				4–11	2–5	tr	tr	tr
Linseed		tr	tr	5–6	3–5	<0.5	tr–0.2	tr
Olive				7–16	2–4	ca. 0.5	tr	tr
Peanut (Africa)			tr	7–12	1.5–5	ca. 1.5	2–4	1–2
Peanut (South America)			tr	10–13	1.5–4	ca. 1.5	3–4	1.5–2
Pumpkinseed				7–13	6–7	tr		
Rapeseed (high erucic)		tr	tr	2–4	1–2	0.5–1	0.5–2.0	0.5
Rapeseed (low erucic)		tr	tr	3–6	1–2.5	<1	tr–0.5	tr–0.2
Ricebran			ca. 0.5	13–18	ca. 2	0.5–1		ca. 0.5
Safflower		tr	tr	ca. 5	2–3	ca. 0.5	ca. 1	ca. 1
Sesame			tr	8–10	3–6	ca. 0.5		
Soybean			<0.5	8–12	3–5	<0.5	tr	
Sunflower		tr	tr–0.1	5.5–8	2.5–6.5	<0.5	0.5–1.0	<0.5
Wheat germ				12–14	ca. 1	0.5		tr
Consistent vegetable fats								
Babassu oil	ca. 12	42–44	15–18	8–10	2–3	tr		
Coconut oil	ca. 13	41–46	18–21	9–12	2–4	tr	tr	
Cocoa butter		tr	tr	23–30	32–37	<1	ca. 0.5	
Palm kernel oil	ca. 7	41–45	15–17	7–10	2–3	tr–0.3	tr–0.5	
Palm oil (Africa)		tr	1–2	41–46	4–6.5	ca. 0.5		
Palm oil (Indonesia)		tr–0.5	ca. 1	41–47	4–6	ca. 0.5		
Animal fats								
Beef tallow		tr	2–4	23–29	20–35	<0.5	tr	
Butterfat	7–9	2–5	8–14	24–32	9–13	2		
Chicken fat			ca. 1	20–24	4–7			
Goose fat				20–22	4–11			
Horse fat	tr	ca. 0.5	3–6	20–30	6–10	tr		
Lard	tr	<0.5	ca. 1.5	24–30	12–18	ca. 0.5		
Mutton tallow	tr	ca. 0.5	1–4	22–30	15–30	tr	tr	
Marine oils								
Fish oils								
Japanese	tr	tr	ca. 6	ca. 16	ca. 3	<0.5	tr	tr–1
Menhaden	tr	tr	ca. 9	ca. 20	ca. 4	tr–1	tr	tr–1
Scandinavian	tr	tr	6–8	11–15	1–3	tr–0.5	tr	tr–1
South American	tr	tr	ca. 7	17–19	2–4	ca. 0.5	tr	tr–1
Whale oil	tr	ca. 0.5	4–10	10–18	1–3	tr		

[a] ITERG = Institut des Corps Gras, Centre Technique Industrielle, Paris.
[b] tr = traces (<0.05 %).

linolenic acid in the rumen. Marine oils contain a high proportion of polyunsaturated fatty acids with a chain length of $C_{20}-C_{24}$. Land-animal fats and marine oils contain numerous odd-numbered and branched fatty acids in trace concentrations. More than 80 different fatty acids have been found in milk fat, and more than 40 in lard. However, most of these fatty acids occur only in traces. They can be of industrial significance for some oils and fats and they can play a role in the identification of fats and their detection in mixtures.

The fatty acid composition of a naturally occurring fat is determined genetically. The fatty acid composition of oleaginous seeds can be changed by developing new varieties. Examples are low-linoleic safflower and sunflower oils and low-erucic rapeseed oil. Environmental factors can influence the fatty acid composition within certain limits. The proportion of unsat-

Table 2. Unsaturated fatty acids in various oils and fats, (main sources: Unilever; Food RA, Leatherhead; ITERG [a])

	Unsaturated fatty acids [b], g/100 g fatty acids									
	$C_{14:1}$	$C_{16:1}$	$C_{18:1}$	$C_{18:2}$	$C_{18:3}$	$C_{20:1}$	$C_{22:1}$	$C_{20:x}$ [c]	$C_{22:x}$ [c]	$C_{24:1}$
Liquid vegetable oils										
Almond		<0.5	65–69	21–25	tr	tr–0.1				
Avocado		2–12	44–76	8–25	ca. 1					
Corn germ		<0.5	25–35	40–60	ca. 1	ca. 0.5	tr–0.1			
Cottonseed		<1	14–21	45–58	tr–0.2	tr				
Grape-seed			12–33	45–72	1–2					
Linseed		tr	18–26	14–20	51–56	<0.5				
Olive		1–2	64–86	4–15	0.5–1	0.5				
Peanut (Africa)		<0.5	50–70	14–30	tr	0.5–1.5	tr			
Peanut (South America)		<0.5	35–42	39–44	tr	0.5–1.5	tr			
Pumpkinseed			24–41	46–57						
Rapeseed (high erucic)		ca. 0.5	11–24	10–22	7–13	ca. 10	41–52			
Rapeseed (low erucic)		0.1–0.5	52–66	17–25	8–11	1.5–3.5	tr–2.5	tr–0.1		tr
Ricebran			ca. 44	30–40						
Safflower		tr	12–20	70–80	tr	tr				
Sesame		tr	35–46	40–48	tr–0.5	<0.5				
Soybean		tr	18–25	49–57	6–11	<0.5				
Sunflower		<0.5	14–34	55–73	tr–0.4	<0.5	tr–0.3			
Wheat germ			ca. 30	40–55	ca. 7					
Consistent vegetable fats										
Babassu oil			14–16	1–2						
Coconut oil	tr	tr	5–9	0.5–3	tr	tr				
Cocoa butter		ca. 0.5	30–37	2–4						
Palm kernel oil			10–18	1–3	tr–0.5	tr–0.5				
Palm oil (Africa)		<0.5	37–42	8–12	tr–0.5	tr				
Palm oil (Indonesia)		ca. 0.5	37–41	ca. 10	tr–0.5	tr				
Animal fats										
Beef tallow	ca. 0.5	2–4	26–45	2–6	ca. 1	<0.5		tr	ca. 0.5	
Butterfat	ca. 2	3	19–33	1–4	2–6				ca. 2	
Chicken fat		ca. 7	38–44	18–23	ca. 1				0.5–1	
Goose fat			41–74	7–19						
Horse fat	3–10		36–40	6–11	4–9	tr–0.5		tr	1–2	
Lard	tr	2–3	36–52	10–12	ca. 1	0.5–1		<0.5	tr	
Mutton tallow	ca. 0.5	3–4	31–56	3–7	1–2				ca. 0.5	
Marine oils										
Fish oils										
Japanese	tr	ca. 7	ca. 14	ca. 2	ca. 1	ca. 7	ca. 6	ca. 15	ca. 12	tr–1
Menhaden	tr	ca. 11	ca. 13	ca. 2	ca. 1	ca. 2	ca. 1	ca. 14	ca. 11	tr–1
Scandinavian	tr	6–11	12–15	1–2	0.5–1	9–16	14–20	6–10	5–11	tr–1
South American	tr	9–11	14–15	1–2	0.5–1	1–2	1–2	7–19	10–14	tr–1
Whale oil	1–3	13–20	24–33	1–2	tr	10–15	4–10	1–6	5–7	tr

[a] ITERG = Institut des Corps Gras, Centre Technique Industrielle, Paris.
[b] tr = traces (<0.05 %).
[c] $x > 1$.

urated fatty acids in the glycerides of linseed, soybean, and sunflower oils, for example, generally increases as the climate becomes colder and wetter. The subcutaneous fats of marine animals living in the colder parts of the oceans have a particularly high content of unsaturated fatty acids and thus a relatively low melting point. Linoleic acid cannot be synthesized by the animal or human organism and is hence referred to as an "essential" fatty acid.

Naturally occurring oils and fats are distributed homogeneously in varying concentra-

tions in vegetable and animal tissues. In plants they are found predominantly in the seeds and the fruit pulp where they serve as a source of energy.

Biosynthesis. The synthesis of fatty acids in plants and animals generally starts with "activated" acetic acid, i.e., acetyl coenzyme A, which is derived from carbohydrates via pyruvic acid. Animals also synthesize acetyl coenzyme A from amino acids.

A distinction is made between *de novo synthesis* of saturated fatty acids and *elongation* of a fatty acid chain. De novo synthesis leads primarily to palmitic acid with smaller amounts of lauric, myristic, and stearic acids.

De novo synthesis:

$$\underset{\text{Acetyl CoA}}{CH_3COSCoA} \xrightarrow{+CO_2} \underset{\text{Malonyl CoA}}{HOOCCH_2COSCoA} \xrightarrow[-HSCoA]{+CH_3COSCoA}$$

$$CH_3COCH(COOH)COSCoA \xrightarrow{-CO_2} CH_3COCH_2COSCoA$$

$$\xrightarrow{+H_2} CH_3CH(OH)CH_2COSCoA \xrightarrow{-H_2O}$$

$$CH_3CH{=}CHCOSCoA \xrightarrow{+H_2}$$

$$\underset{\text{Butyryl CoA}}{CH_3CH_2CH_2COSCoA} \xrightarrow[-HSCoA]{+\text{Malonyl CoA}}$$

$$CH_3CH_2CH_2COCH(COOH)COSCoA \text{ etc.}$$

The conversion of palmitic into stearic acid and the formation of longer chain saturated and unsaturated fatty acids presumably takes place by elongation, which involves addition of acetyl coenzyme A to activated fatty acids (RCOSCoA).

Elongation:

$$RCOOH \xrightarrow[-H_2O]{+HSCoA} RCOSCoA \xrightarrow[-HSCoA]{+CH_3COSCoA}$$

$$RCOCH_2COSCoA \longrightarrow \longrightarrow RCH_2CH_2COOH$$

The synthesis of fatty acids from C_2 units explains the predominance of even-numbered fatty acids in naturally occurring oils and fats. Bacterial degradation of feed in the rumen leads not only to acetic and butyric acids but also to propionic acid (C_3); incorporation of these C_3 units into fat synthesis in the udder explains the occurrence of branched and odd-numbered fatty acids in milk fat.

Oleic acid appears to play a key role in plants. It is probably synthesized from short-chain fatty acids and can be dehydrogenated into more highly unsaturated fatty acids; be converted into substituted acids such as ricinoleic acid; or elongated into erucic acid by specific enzyme systems (Fig. 1).

In the animal organism oleic acid is formed by dehydrogenation of stearic acid. Further dehydrogenation to linoleic and linolenic acid does not take place. The animal organism can, however, further desaturate linoleic and linolenic acid by introducing double bonds between the carboxyl group and the double bond nearest to it. Together with an elongation of the fatty acid chain this leads to the formation of γ-dihomolinolenic acid ($C_{20:3}$), arachidonic acid ($C_{20:4}$), and eicosopentaenoic acid ($C_{20:5}$), which are precursors of prostaglandins.

Three activated fatty acids (RCOSCoA) react successively with a molecule of phosphorylated glycerol to form a triglyceride molecule. The factors that determine the distribution of the fatty acids over the 1-, 2-, and 3-positions of the glycerides are not fully known. Phosphatides are also synthesized by this route.

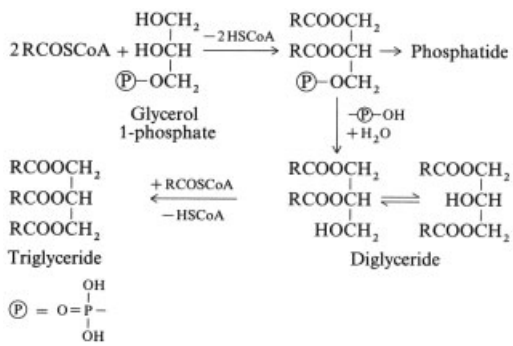

Biodegradation. During digestion and absorption of fats, triglycerides are successively split into di- and monoglycerides, glycerol, and fatty acids by lipases. The products are absorbed by the intestinal epithelial cells either as water-soluble complexes or as micelles. The fatty acids are biodegraded mainly via β-oxidation:

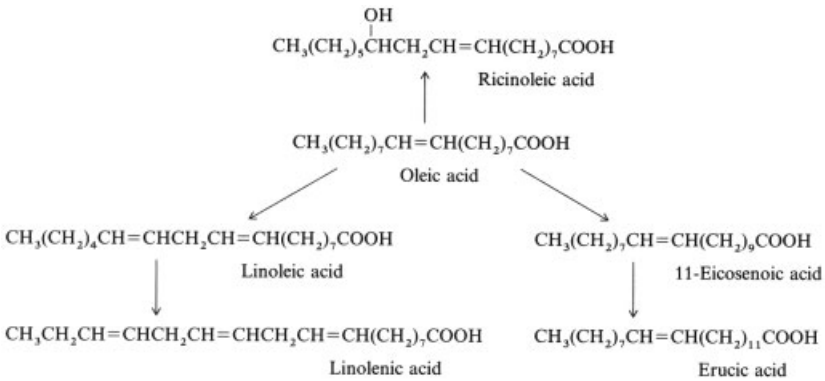

Figure 1. Unsaturated fatty acids derived from oleic acid

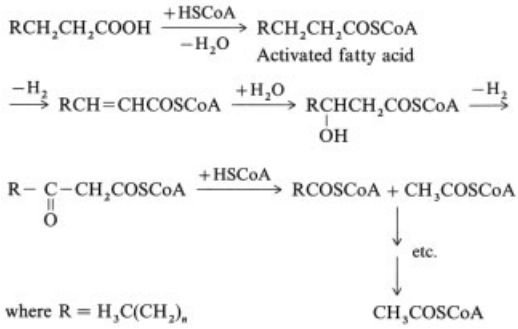

The bio-oxidation of unsaturated fatty acids involves additional steps.

2.3. Phospholipids [20]

Phospholipids are essential constituents of the protoplasm of animal and plant cells; they are mostly present as lipoproteins and lipidcarbohydrate complexes. Oilseeds, cereal germs, egg yolk, and brain are the richest sources of phospholipids. Esters of glycerophosphoric acid (glycerol 1-phosphate) are usually referred to as phosphatides (→ Lecithin).

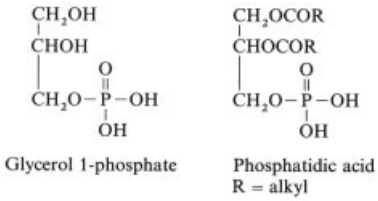

Some important phosphatides are phosphatidylcholine [*8002-43-5*] (lecithin), phosphatidylethanolamine [*5681-36-7*] (cephalin), phosphatidylinositol [*2816-11-7*], and phosphatidylserine.

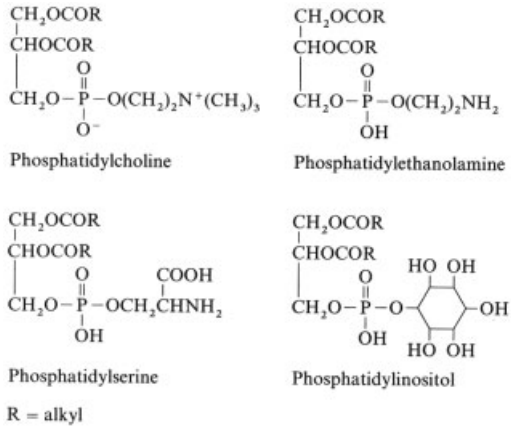

The plasmalogens are ethers of fatty alcohols and phosphatidic acid; they occur in animal tissue.

Sphingolipids are derivatives of the amino alcohol sphingosine. For example, sphingomyelins are constituents of the phospholipids in the brain, blood plasma, and erythrocytes.

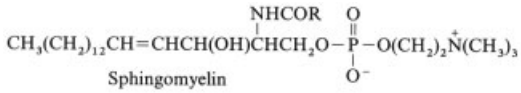

The cerebrosides are derivatives of sphingosine and either galactose (galactolipids) or glucose (glycolipids). They are important con-

stituents of the myelin nerve sheath. The gangliosides are based on neuraminic acid and are found in the ganglia cells of the brain.

Lecithin, cephalin, inositol phosphatides, and phosphatidic acid are the principal phospholipid components of plant origin. During pre-refining of crude vegetable oils, especially soybean and rapeseed oil, most of the phosphatides are removed as sludge by hydration with water. Drying of this sludge yields "lecithin," which is used, often after further modification, in the food industry as an emulsifier, antispattering agent, dispersant, or viscosity reducing agent. Lecithin is also used in pharmaceuticals, toiletries, animal feeds, and as a mold release agent and emulsifier – dispersant [21]. The phosphatide contents of crude oils and fats are listed in Table 3; the phosphatides are almost completely removed during refining of oils and fats.

2.4. Waxes

Waxes are esters of fatty alcohols and fatty acids (→ Waxes). Free and esterified fatty alcohols (e.g., cetyl, stearyl, oleyl alcohols) occur in considerable concentrations in marine oils. Ethers of glycerol and fatty alcohols (batyl, chimyl, and selachyl alcohols) are also found in animal tissues. The wax in the seed coat of sunflower seed causes the oil to become cloudy at refrigerator temperatures and is therefore removed by winterization (see Chap. 7).

2.5. Sterols and Sterol Esters

The major part of the nonsaponifiable matter of oils and fats consists of sterols present as such or as fatty acid esters and glycolipids. The most important sterol in animal fats is cholesterol [*57-88-5*]. β-Sitosterol [*83-46-5*] is the predominant sterol in vegetable oils and fats, although traces of cholesterol are also present. Total sterol concentrations are shown in Table 3, and sterol compositions in Table 4. Some of the sterols are removed during the deodorization step of refining oils and fats, without, however, changing their relative composition. Sterols are therefore a useful tool in checking authenticity. Vitamin D_2 [*50-14-6*] (calciferol) is present in milk and butter, and vitamin D_3 [*67-97-0*] (cholecalciferol) in cod liver oil. Other animal and vegetable oils contain hardly any vitamin D.

2.6. Terpenoids

The nonsaponifiable part of most fats also contains traces of terpenes and terpene alcohols. The triterpene squalene [*7683-64-9*] occurs in relatively high concentrations (up to 0.5 %) in olive oil. Shea butter contains 2 – 10 % of kariten, a rubber-like hydrocarbon.

Carotenoid pigments occur widely in oils and fats. Approximately 70 different carotenoids, ranging in color from yellow to deep red, are known. The most well-known carotenoids are the isomeric tetraterpenes ($C_{40}H_{56}$) α-carotene [*7488-99-5*], β-carotene [*7235-40-7*], and γ-carotene [*472-93-5*], lycopene [*502-65-8*], and xanthophyll [*127-40-2*]. Crude palm oil contains up to 0.2 % α- and β-carotene.

The bulk of the carotenoids are removed during refining, primarily during bleaching and deodorization. β-Carotene (provitamin A) can be oxidized to vitamin A [*68-26-8*] in the animal organism. Most vegetable oils and fats do not contain significant concentrations of vitamin A; in most countries margarine therefore contains added vitamin A. Vitamin A is present in high concentrations in fishliver oils. Butter contains ca. 0.003 – 0.0015 % vitamin A.

The most important compounds of the vitamin E [*59-02-9*] group are α-, β-, γ-, and δ-tocopherols:

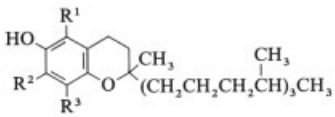

	R^1	R^2	R^3
α-Tocopherol	CH_3	CH_3	CH_3
β-Tocopherol	CH_3	H	CH_3
γ-Tocopherol	H	CH_3	CH_3
δ-Tocopherol	H	H	CH_3

These compounds act as a vitamin (rat fertility factor) and as antioxidants. α-Tocopherol has the highest biological activity.

Tocopherols occur only in traces in animal fats, whereas vegetable oils contain appreciable concentrations (see Table 5). Refined oils and

Table 3. Minor constituents of crude oils and fats

	Content, wt %		
	Phosphatides	Tocopherols	Sterols
Babassu fat		0.003	
Beef tallow	<0.07	0.001	0.08 – 0.14
Butterfat	<1.4	0.003	0.24 – 0.50
Castor oil			0.5
Cocoa butter	0.1	0.003	0.17 – 0.20
Coconut oil		0.003	0.05 – 0.1
Cod liver oil			0.42 – 0.54
Corn germ oil	1 – 2	0.1 – 0.3	0.8 – 2.2
Cottonseed oil	0.7 – 0.9	0.04 – 0.11	0.27 – 0.6
Fish oil			ca. 0.3
Lard	<0.05	0.003	ca. 0.1
Linseed oil	0.3	0.11	0.37 – 0.42
Olive oil		0.01 – 0.03	0.1 – 0.2
Palm oil	0.05 – 0.1	0.02 – 0.12	0.04 – 0.08
Palm kernel fat			0.08 – 0.12
Peanut oil	0.3 – 0.4	0.02 – 0.07	0.19 – 0.29
Rapeseed oil	2.5	0.07 – 0.08	0.5 – 1.1
Sesame oil	0.1	ca. 0.05	0.4 – 0.6
Soybean oil	1.1 – 3.2	0.09 – 0.12	0.2 – 0.4
Sunflower oil	<1.5	0.07 – 0.1	0.25 – 0.45
Wheat germ oil	0.1 – 2.0	ca. 0.28	1.3 – 1.7

Table 4. Sterol composition ∗ in crude oils (main sources: Food RA, Leatherhead; Unilever; ITERG ∗∗)

	Coconut	Corn germ	Cotton-seed	Olive	Palm	Palm kernel	Peanut	Rapeseed	Soybean	Sunflower
Cholesterol	0.6 – 2	0.2 – 0.6	0.7 – 2.3	0 – 0.5	2.2 – 6.7	1 – 3.7	0.6 – 3.8	0.4 – 2	0.6 – 1.4	0.2 – 1.3
Brassicasterol	0 – 0.9	0 – 0.2	0.1 – 0.9			0 – 0.3	0 – 0.2	5 – 13	0 – 0.3	0 – 0.2
Campesterol	7 – 10	18 – 24	7.2 – 8.4	2.3 – 3.6	18.7 – 29.1	8.4 – 12.7	12 – 20	18 – 39	16 – 24	7 – 13
Stigmasterol	12 – 18	4 – 8	1.2 – 1.8	0.6 – 2	8.9 – 13.9	12.3 – 16.1	5 – 13	0 – 0.7	16 – 19	8 – 11
β-Sitosterol	50 – 70	55 – 67	80 – 90	75.6 – 90	50.2 – 62.1	62.6 – 70.4	48 – 65	45 – 58	52 – 58	56 – 63
Δ5-Avenasterol	5 – 16	4 – 8	1.9 – 3.8	3.1 – 14	0 – 2.8	4 – 9	7 – 9	0 – 6.6	2 – 4	2 – 7
Δ7-Stigmastenol	2 – 8	1 – 4	0.7 – 1.4	0 – 4	0.2 – 2.4	0 – 2.1	0 – 5	0 – 5	1.5 – 5	7 – 13
Δ7-Avenasterol	0.6 – 2	1 – 3	1.4 – 3.3		0 – 5.1	0 – 1.4	0 – 5	0 – 0.8	1 – 4.5	3 – 6

∗ Sterols as percentage of total sterol fraction.
∗∗ ITERG = Institut des Corps Gras, Centre Technique Industrielle, Paris.

fats still contain ca. 80 % of the original tocopherols, the main losses occurring during deodorization. Deodorizer distillates are a valuable source of natural tocopherols.

2.7. Other Minor Constituents

Sesame oil contains 0.3 – 0.5 % sesamolin [*526-07-8*], a glycoside of the phenol sesamol [*533-31-3*], and 0.5 – 2.0 % sesamin [*607-80-7*]. These minor constituents give a characteristic color reaction (the basis of the Baudouin test) and impart stability to oxidation.

Gossypol [*303-45-7*], a toxic polyphenol with pronounced antioxidant activity, is found in crude cottonseed and kapokseed oils (0.5 – 1.5 %). Gossypol is removed during refining in the lye neutralization step.

Crude linseed, rapeseed, soybean, olive, avocado, and many other vegetable oils contain the green pigments chlorophyll and phaeophytin. The chlorophyll content is particularly high in oil from immature seeds and is generally determined by harvesting and climatic conditions. Chlorophyll is generally regarded to be indicative of inferior crude oil quality; it can be removed by treatment with acidic absorbents such as bleaching earth.

Table 5. Tocopherols (mg/kg) in crude oils (main sources: Food RA, Leatherhead; Unilever; ITERG ** [22])

	Coconut	Corn germ	Cotton-seed	Olive	Palm	Palm kernel	Peanut	Rapeseed	Safflower	Sesame	Soybean	Sunflower
α-Tocopherol	0–17	270–370	140–540	100–250	80–95 *	0–44	50–300	230–300	ca. 580	2–18	60–75	600–1000
β-Tocopherol	0–14	10–16		0–4	0–5 *	0–123	100–400	5–10	ca. 18	<0.5	10–20	15–35
γ-Tocopherol		640–860	160–625	8–14	5–15 *	0–6		400–550	ca. 16	500–540	580–740	3–35
δ-Tocopherol	0–2	30–50	0–17	<10	0–5 *		0–20	15–20	<0.5	<0.5	275–320	0–7
Total	0–31	1100–1300	410–1169	110–270	200–1200		175–700	670–800	ca. 610	500–550	900–1200	650–1000

* Tocopherols as percentage of total tocopherol fraction.
** ITERG = Institut des Corps Gras, Centre Technique Industrielle, Paris.

Crude oils and fats may contain traces of proteins, the concentration depending on processing conditions. In addition, vegetable oils may contain carbohydrates. These compounds are removed almost completely in refining.

Autoxidation of the fatty acid groups of triglycerides leads to the formation of volatile and nonvolatile oxidation products. The type and quantity of the volatile compounds (ketones, aldehydes, and alcohols) depend on the initial fatty acid composition and the oxidation conditions. These volatile products are responsible for the typical odor and taste of oils and fats. One of the main aims of refining is to remove these odoriferous compounds, which in some cases occur in concentrations of only 10^{-3} ppm. The nonvolatile oxidation products generally have little odor and taste, but they can act as oxidation promoters.

Most crude oils and fats contain traces of pesticide residues and metals (e.g., Fe, Cu, Pb, As, Cd, and Hg) as a consequence of crop treatment and environmental influences. Whereas phosphate-based pesticides used to treat oilseeds decompose with time, chlorinated pesticides are stable and gradually migrate into the oil.

Peanut oil may contain traces of aflatoxins produced by growth of *Aspergillus flavus* on the seed.

Most oils and fats, particularly coconut oil, contain varying concentrations of polycyclic aromatic hydrocarbons. These hydrocarbons are introduced during smoke-drying of the raw materials prior to storage and further processing.

Crude rapeseed oil contains up to 50 ppm of sulfur in the form of elemental sulfur, isothiocyanates, and 5-vinyl-2-oxazolidinethione [*500-12-9*] (goitrin) derived from glucosinolates (sulfur-containing glucosides) in the seed.

All of these undesirable contaminants are reduced to negligibly low levels in the course of refining, if necessary with the aid of additional steps such as adsorptive treatment with activated charcoal to remove polycyclic aromatic hydrocarbons.

Table 6. Melting points of fatty acids and glycerides in their stable polymorphic forms

	mp (slip point), °C			
	Fatty acid	1-Monoglyceride	1,3-Diglyceride	Triglyceride
Butyric acid	−7.9			
Caproic acid	−3.9	19.4		−25.0
Caprylic acid	16.3	40.0		8.3
Capric acid	31.3	53.0	44.5	31.5
Lauric acid	44.0	63.0	56.5	46.5
Myristic acid	54.4	70.5	65.5	57.0
Palmitic acid	62.9	77.0	72.5	65.5
Stearic acid	69.6	81.0	78.0	73.0
Arachidic acid	75.4	84.0		(72.0)
			(75.0)	
Behenic acid	79.9			81.0
Oleic acid	16.3	35.2	21.5	5.5
Elaidic acid	45.0	58.5	55.0	42.0
Erucic acid	34.7	50.0	46.5	30.0
Linoleic acid	−5.0	12.3	−2.6	−13.1
Linolenic acid	−11.0	15.7	−12.3	−24.2
Ricinoleic acid	5.0			

3. Physical Properties [23]

3.1. Melting and Freezing Points

The *melting point* of the even-numbered, saturated fatty acids increases with increasing chain-length, and decreases with increasing degree of unsaturation (see also → Fatty Acids, Chap. 2.1.). The glycerides show a similar behavior (Table 6). Since the naturally occurring fats are mixtures of glycerides, they melt over a wide range of temperature (for methods of melting point determination, see Section 11.3.1).

The melting point also depends on the polymorphic form of the glycerides, i.e., the crystalline structure (Table 7). The packing density and the spatial arrangement of the triglyceride molecules depend on the crystallization conditions. X-ray diffraction patterns (short spacing lines) can distinguish between three different polymorphic forms [25]. The lowest melting and most labile form is designated α. The most stable form is called β and the intermediate one β'. The α form is obtained by rapidly cooling the molten triglycerides, the β' form by suitable tempering. Polymorphic changes induced by processing conditions are of great practical importance since they can significantly influence the properties of, for example, a margarine (oral melting behavior, sandiness) or chocolate (fat bloom).

Table 7. Melting points of some polymorphic triglycerides

	mp, °C		
	α	β'	β
Trilaurin	15.0	35.0	46.5
Trimyristin	33.0	46.5	57.0
Tripalmitin	45.0	56.0	65.5
Tristearin	54.5	65.0	73.0
Triolein	−32.0	−12.0	5.5
Trielaidin	15.5	37.0	42.0
1,2-Dicapriolaurin	17.5	26.0	30.0
2-Capriodilaurin	23.0	33.0	38.5
1-Laurodimyristin	37.0	42.0	46.5
1-Laurodipalmitin	45.0	49.5	54.0
2-Laurodipalmitin	47.0	50.0	53.5
1,2-Dicapriostearin	32.0	38.0	41.0
1,3-Dicapriostearin	34.0	40.0	44.5
2-Palmitodistearin	56.0	64.0	68.5
1,3-Dicapriioolein	−10.2	0.6	6.2
1,2-Dilauroolein	−10.0	4.8	16.0
1,2-Dipalmitoolein	ca. 18.0	ca. 31.0	34.5
1,3-Dipalmitoolein	26.5	33.5	38.0
1,2-Distearoolein	ca. 30.0	ca. 40.0	
1,3-Distearoolein	37.0	41.5	44.0
1,3-Dipalmitoelaidin	ca. 40.0	53.0	54.0
1-Stearodibehenin	61.3	71.0	73.5
1-Lauro-2-myristo-3-palmitin	37.0	44.0	49.0
1-Lauro-2-myristo-3-stearin	27.5	45.5	49.5
1-Palmito-3-stearo-2-olein	18.2	33.0	38.0

The *congeal* or *set point* (point of solidification) is generally lower than the melting point.

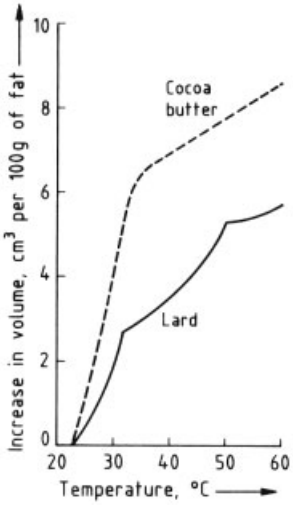

Figure 2. Dilatation curves for cocoa butter and lard

The *solids content* of a fat at different temperatures is normally determined by pulsed nuclear magnetic resonance [24]. Fats exhibit an increase in volume (dilatation) on melting that is disproportionately larger than that on heating a liquid fat; an obsolete method of measuring solids content at different temperatures was based on dilatation. *Dilatation* or *solids content curves* can be used to characterize a fat. Figure 2 reflects the large proportion of 1-palmito-3-stearo-2-olein in cocoa butter and the more complex glyceride composition of lard.

The *latent heat of fusion* increases with increasing chain length and increasing degree of saturation (see Table 8). Naturally occurring fats generally have a lower heat of fusion than simple glycerides.

Table 8. Latent heat of fusion of some fats

	Heat of fusion, J/g
Butterfat	81.6
Cottonseed oil	86.0
Fully hardened cottonseed oil ∗	185.0
Peanut oil	90.9
Partially hardened peanut oil ∗∗	103.4
Trilaurin (β-form)	193.5
Tripalmitin (β-form)	222.0
Tristearin (β-form)	228.0

∗ Iodine value ca. 1.
∗∗ Iodine value ca. 60.

Table 9. Specific heat of oils and fats

	Specific heat, J/g	Temperature, °C
Trilaurin	2.130	66.0
Trimyristin	2.152	58.4
Tripalmitin	2.173	65.7
Tristearin	2.219	79.0
Soybean oil	2.060	80.4
	2.000	60.0
Linseed oil	2.050	70.7
Cottonseed oil	2.200	90.0
Hardened cottonseed oil ∗	2.177	79.6
Olive oil	2.300	110
Palm oil	2.400	140
Sunflower oil	2.500	175.0

∗ Iodine value ca. 6.

Table 10. Vapor pressure of triglycerides (temperatures corresponding to a vapor pressure p of 6.7 Pa (0.05 mm Hg) and 0.13 Pa (0.001 mm Hg)

	Temperature, °C	
	$p = 6.7$ Pa	$p = 0.13$ Pa
Tributyrin	91	45
Tricaproin	135	85
Tricaprylin	179	128
Tricaprin	213	159
Trilaurin	244	188
Trimyristin	275	216
Tripalmitin	298	239
Tristearin	313	253
Soybean oil	308	254
Olive oil	308	253
1,3-Distearoolein	315	254
1-Myristo-2-palmito-3-stearin	297	237
1-Palmito-2-lauro-3-stearin	290	232
1-Myristo-2-lauro-3-stearin	282	223
1-Palmito-2-capro-3-stearin	280	223
1-Capro-2-lauro-3-myristin	249	189

3.2. Thermal Properties

The approximate *heats of combustion* of oils and fats can be calculated from the following formula:

$$\begin{aligned}&\text{Heat of combustion (J/g)}\\&= 47645 - 4.1868 \times \text{iodine value}\\&\quad -38.31 \times \text{saponification value}\end{aligned}$$

Using this equation, values ranging from 37 765 J/g (9020 cal/g) for coconut oil to 40 528 J/g (9680 cal/g) for a high-erucic rapeseed oil have been obtained.

The *specific heat* of liquid oils and fats increases with increasing chain length and degree of saturation (see Table 9); it also increases with temperature.

Triglycerides of long-chain fatty acids have extremely low *vapor pressures*; typical data for some triglycerides and oils are shown in Table 10. Monoglycerides have significantly higher vapor pressures (see Table 11). There is relatively little information on the *heats of vaporization* of fats (see Table 12).

Table 11. Boiling points of some monoglycerides

	Pressure, Pa	*bp*, °C
Monocaprin	133.3	175
Monolaurin	133.3	186
Monomyristin	133.3	199
Monopalmitin	133.3	211
Monostearin	26.7	190
Monoolein	26.7	186

Table 12. Heat of vaporization of some triglycerides calculated for 0.13 – 67 Pa

	Heat of vaporization, J/g
Tricaprylin	247.0
Tricaprin	226.1
Trilaurin	213.5
Trimyristin	205.2
Tripalmitin	201.0
Tristearin	188.4
Soybean oil	209.3

Oils and fats are relatively poor thermal conductors. Data for *thermal conductivity* are limited. Most data lie within ± 10 % of the following general relationship:

$$\text{Thermal conductivity } \left(\mathrm{W\,m^{-1}K^{-1}}\right) = 0.181 - 0.00025t$$

where t = temperature in °C.

The *smoke, flash,* and *fire points* of oils and fats are measures of their thermal stability when heated in air (Table 13).

3.3. Density

The density of fatty acids and glycerides decreases with increasing molecular mass and degree of saturation (Tables 14 and 15). Oxidation generally leads to higher densities. A high free fatty acid content tends to decrease the density of a crude oil. The following formula can be used to estimate the density of an oil:

$$d_{15}^{15} = 0.8475 + 0.0003 \times \text{saponification value} + 0.00014 \times \text{iodine value}$$

Up to 260 °C the density decreases by about 0.00064 g/cm^3 per temperature increase of 1 °C. The following equations apply for the density ϱ (in g/L) of commercial oils and fats (t = temperature in °C):

Soybean oil	$933.4 - 0.657\,t$
Sunflower oil	$932.7 - 0.680\,t$
Sesame oil	$933.0 - 0.700\,t$
Cottonseed oil	$931.7 - 0.755\,t$
Peanut oil	$927.0 - 0.642\,t$
Olive oil	$928.5 - 0.700\,t$
Palm oil	$925.0 - 0.655\,t$
Palm kernel oil	$940.0 - 0.740\,t$
Coconut oil	$932.0 - 0.745\,t$
Rapeseed oil	$925.5 - 0.700\,t$
Tallow	$956.8 - 0.898\,t$
Fish oil	$940.0 - 0.700\,t$

3.4. Viscosity

Oils tend to have a relatively high viscosity because of intermolecular attraction between their fatty acid chains. Generally, viscosity tends to increase slightly with increasing degree of saturation and increasing chain length (see Tables 16 and 17). There is an approximately linear relationship between log viscosity and temperature. The viscosity of oils tends to increase on prolonged heating due to the formation of dimeric and oligomeric fatty acid groups.

3.5. Solubility and Miscibility

Nearly all fats and fatty acids are easily soluble in common organic solvents such as hydrocarbons, chlorinated hydrocarbons, ether, and acetone. Castor oil is an exception in that it is only partially soluble in petroleum ether but easily so in ethanol. The solubility of fatty acids in ethanol is greater than that of the corresponding triglycerides. The solubility of fats in organic solvents decreases with increase in molecular mass and increases with degree of unsaturation.

Table 13. Smoke, flash, and fire points of refined oils and fats

	ffa ∗, %	Smoke point, °C	Flash point, °C	Fire point, °C
Rapeseed oil	0.08	218	317	344
Peanut oil	0.09	207	315	342
Peanut oil	0.11	198	333	363
Peanut oil	1.0	160	290	–
Cottonseed oil	0.04	223	322	342
Cottonseed oil	0.18	185	318	357
Soybean oil	0.04	213	317	342
Sunflower oil	0.1	209	316	341
Coconut oil	0.1	200	300	
Coconut oil	0.2	194	288	329
Coconut oil	1.0	150	270	
Palm oil	0.06	223	314	341
Hardened peanut oil (*mp* 32/34 °C)	0.04	226	314	340
Hardened soybean oil (*mp* 42/44 °C)	0.04	223	318	342
Beef tallow	0.4		316	344
Beef tallow	5.0		266	344

∗ ffa = free fatty acid.

The differences in solubility enable categories of glycerides to be separated by fractional crystallization although complete separation is rarely achieved because of mutual solubility effects.

The water solubility of fats is low and decreases with increasing chain length and with decreasing temperature. The solubility of water in cottonseed oil is, for example, 0.14 wt % at 30 °C and 0.07 wt % at 0 °C.

The solubility of gases in oils generally increases with increase in temperature, the reverse holding for carbon dioxide (Table 18).

Table 14. Density of triglycerides

	ϱ, g/cm^3		
	80 °C	15 °C	25 °C
Tricaprin	0.8913		
Trilaurin	0.8801		
Trimyristin	0.8722		
Tripalmitin	0.8663		
Tristearin	0.8632		
Triolein		0.9162	0.9078
Trilinolein		0.9303	
Trilinolenin		0.9454	

Table 16. Viscosity of triglycerides at 70°C

	η, mPa · s
Tributyrin	3.0
Tricaproin	5.9
Tricaprylin	8.8
Tricaprin	11.7
Trilaurin	14.6
Trimyristin	17.6
Tripalmitin	20.5
Tristearin	23.4

Table 15. Density of fats and oils

	ϱ_{15}, g/cm^3	d_{25}^{25}
Vegetable fats		
Babassu oil	0.9250	
Castor oil	0.950 – 0.974	0.945 – 0.965
Coconut oil	0.919 – 0.937	0.869 – 0.874[a]
Cocoa butter	0.945 – 0.976	0.856 – 0.864[a]
Corn germ oil	0.920 – 0.928	0.916 – 0.921
Cottonseed oil	0.917 – 0.931	0.916 – 0.918
Grape-seed oil	0.919 – 0.936	
Hempseed oil	0.924 – 0.932	0.923 – 0.925[b]
Linseed oil	0.930 – 0.935	0.931 – 0.936
Mustard seed oil	0.912 – 0.923	
Olive oil	0.914 – 0.925	0.909 – 0.915
Oiticica oil	0.9518 – 0.9694	0.978[b]
Palm oil	0.921 – 0.947	0.898 – 0.901[c]
Palm kernel oil	0.925 – 0.935	0.860 – 0.873[a]
Peanut oil	0.911 – 0.925	0.910 – 0.915
Perilla oil	0.927 – 0.933	min. 0.932[b]
Poppyseed oil	0.923 – 0.926	
Rapeseed oil	0.910 – 0.917	0.906 – 0.910
Ricebran oil		0.916 – 0.921
Safflower oil	0.923 – 0.928	0.919 – 0.924
Sesame oil	0.921 – 0.924	0.914 – 0.919
Shea butter	0.917 – 0.918	
Soybean oil	0.922 – 0.934	0.917 – 0.921
Sunflower oil	0.920 – 0.927	0.915 – 0.919
Tung oil	0.936 – 0.945	0.940 – 0.943[b]
Wheat germ oil		0.925 – 0.933
Animal fats		
Beef tallow	0.936 – 0.952	0.860 – 0.870[a]
Butterfat	0.935 – 0.943	
Herring oil	0.917 – 0.930	
Horse fat	0.915 – 0.932	
Lard	0.914 – 0.943	0.858 – 0.864[a]
Menhaden oil	0.925 – 0.935	
Mutton tallow	0.936 – 0.960	
Sperm oil	0.875 – 0.890	
Whale oil	0.914 – 0.931	0.910 – 0.920

[a] $d_{15.5}^{99}$. [b] $d_{15.5}^{15.5}$. [c] $d_{37.8}^{37.8}$.

Table 17. Viscosity of naturally occurring fats and oils

	η, mPa · s			
	20 °C	30 °C	40 °C	50 °C
Castor oil	1000	454	232	128
Coconut oil		39	26	19
Cottonseed oil	80	55	38	27
Fish oil	60	43	32	22
Lard			35	25
Linseed oil	48	33	25	18
Olive oil	80	55	40	30
Palm oil			40	28
Palmkernel oil		43	29	20
Peanut oil	78	50	32	23
Poppyseed oil	63			
Rapeseed oil (high erucic)	85	60	40	30
Sesame oil	65			25
Soybean oil	65	45	33	25
Sunflower oil	68	47	35	26
Tallow				25

Table 18. Solubility of gases in fats

	Gas solubility, vol %				
	t, °C	N_2	H_2	O_2	CO_2
Cottonseed oil	30.5	7.11	4.63		
	49.6	7.79	5.40		
	78.2	8.91	6.73		
	101.5	9.76	7.83		
	147.8	11.83	10.24		
Soybean oil	ca. 20.0	ca. 4.95		ca. 2.65	
Lard	41.5	7.65	5.218		
	73.2	8.79	6.58		
	111.3	10.38	8.50		
	147.3	12.06	10.35		
Hardened tallow (iodine value 1)	64.3		6.14		92.0
	67.0	8.44		14.50	
	84.7			15.35	
	88.0				79.1
	139.4	11.68	9.79		61.9

3.6. Surface and Interfacial Tension

The surface and interfacial tensions of some oils are shown in Table 19. The interfacial tension is markedly reduced by the presence of surface-active agents, e.g., phosphatides, monoglycerides, free fatty acids, and soaps.

Table 19. Surface and interfacial tension of some oils

	Surface tension, mN/m			Interfacial tension
	20 °C	80 °C	130 °C	at 70 °C, mN/m
Cottonseed oil	35.4	31.3	27.5	29.8
Coconut oil	33.4	28.4	24.0	
Castor oil	39.0	35.2	33.0	
Peanut oil				29.9
Soybean oil				30.6

3.7. Electrical Properties

Dry oils, fats, and fatty acids are poor conductors of electricity. Recorded values for the specific resistance of stearic acid are $0.6\times10^{11}\,\Omega$ at 100 °C and $22.3\times10^{11}\,\Omega$ at 186 °C, and for oleic acid 2×10^{11} and $83\times10^{11}\,\Omega$ at comparable temperatures. The dielectric constant of most oils lies in the range 3.0 – 3.2 at 25 – 30 °C. Castor and oiticica oil have dielectric constants of about 4 in this temperature range because of the hydroxyl and keto groups in the fatty acid chains. In emulsion systems, e.g., butter or margarine, the dielectric constant is affected by both the emulsion structure and the moisture content.

3.8. Optical Properties

The refractive index of oils, fats, and fatty acids generally increases with increasing chain length, number of double bonds, and extent of conjugation (see Tables 20 and 21). The refractive index of fatty acids is much lower than that of the corresponding triglycerides. Prolonged heating leads to an increase in refractive index due to the introduction of polar groups into the fatty acid chain.

Table 20. Refractive indices of glycerides

	n_D^{60}
Tricaprin	1.4370
Trilaurin	1.4402
Trimyristin	1.4428
Tripalmitin	1.4452
Tristearin	1.4471
Triolein	1.4548
Trilinolein	1.4645
1,2-Dilaurostearin	1.4437
1,3-Dilaurostearin	1.4442

Table 20. Continued

1,2-Dilauroolein	1.4456
1,3-Dilaurostearin	1.4459
1,2-Dipalmitoolein	1.4480
1,2-Distearoolein	1.4494
1-Stearodiolein	1.4524
Monocaprin	1.4443
Monolaurin	1.4462
Monomyristin	1.4480
Monopalmitin	1.4499
1-Lauro-3-olein	1.4472
1-Stearo-3-olein	1.4507

Table 21. Refractive indices of naturally occurring oils and fats

	n_D^{20}
Babassu oil	1.449 – 1.450∗
Beef tallow	1.454 – 1.459∗
Butterfat	1.452 – 1.457∗
Castor oil	1.477 – 1.479
Cocoa butter	1.453 – 1.458∗
Coconut oil	1.448 – 1.450∗
Corn oil	1.474 – 1.476
Cottonseed oil	1.472 – 1.477
Grape seed oil	1.474 – 1.478
Herring oil	1.470 – 1.475∗
Lard	1.458 – 1.461∗
Linseed oil	1.479 – 1.481
Menhaden oil	1.480
Mustardseed oil	1.470 – 1.474
Mutton tallow	1.455 – 1.458∗
Oiticica oil	1.4921 – 1.4945
Olive oil	1.467 – 1.471
Palm oil	1.453 – 1.456∗
Palmkernel oil	1.449 – 1.452∗
Peanut oil	1.460 – 1.472
Perilla oil	1.481 – 1.483
Rapeseed oil	1.472 – 1.476
Safflower oil	1.4754
Sesame oil	1.473 – 1.476
Soybean oil	1.470 – 1.478
Sunflower oil	1.474 – 1.476
Tung oil	1.517 – 1.526
Whale oil	1.463 – 1.471
Wheat germ oil	1.469 – 1.478∗

∗ n^{40}.

There are a number of equations depicting a relationship between refractive index and other data. The following equation has been suggested for fresh, nonhydrogenated oils and fats:

$$n_D^{40} = 1.4643 - 0.00066 \times \text{saponification value} - \frac{0.0096 \times \text{acid value}}{\text{saponification value}} + 0.0001711 \times \text{iodine value}$$

Pure glycerides do not absorb in the visible region of the spectrum (400 – 750 nm). However, naturally occurring oils and fats invariably contain pigments that have characteristic absorption bands (carotene at 450 nm, chlorophyll and phaeophytin at 660 nm). Most of these pigments are removed during refining. Unsaturated oils and fats absorb in the ultraviolet region between 200 and 400 nm. Conjugated double bonds show characteristic maxima at 232 nm (dienes, e.g., 9,11-*trans,trans*-linoleic acid) and at 268 nm (trienes, e.g., β-eleostearic acid). Conjugated tetraenoic acids absorb between 290 and 320 nm.

In the infrared region (0.075 – 1000 µm) chain substituents such as epoxy, hydroxyl, keto, and cyclopropene groups as well as trans double bonds exhibit specific absorption peaks:

monoglycerides	1.43 µm
hydroxy fatty acids	3.20 µm
ester carbonyl group	5.83 µm
trans double bonds	10.0 – 10.35 µm
isolated trans double bonds	10.35 µm

The fingerprint region around 8 µm can be used to determine the fatty acid chain length.

X-ray spectroscopy is employed to characterize the various polymorphic crystal structures of pure glycerides (see Section 3.1).

Nuclear magnetic resonance of hydrogen atoms is used in structural identification and to determine the solids content of fats at different temperatures. Nuclear magnetic resonance measurement of phosphorus and nitrogen atoms can be employed to analyze the composition of phosphatide mixtures.

The optical activity of enantiomorphic triglycerides with different fatty acid groups in the 1- and 3-positions is usually too small to be measured [20].

4. Chemical Properties

The chemical reactions of fats are basically those of esters and hydrocarbon chains (→ Esters, Organic). Only those reactions that are primarily relevant to the processing of edible oils and fats are dealt with in this chapter.

4.1. Hydrolysis

Glycerides can be hydrolyzed into fatty acids and glycerol:

$$\begin{array}{l} CH_2OCOR \\ | \\ CHOCOR + 3H_2O \rightleftharpoons \\ | \\ CH_2OCOR \end{array} \quad \begin{array}{l} CH_2OH \\ | \\ CHOH + 3RCOOH \\ | \\ CH_2OH \end{array}$$

The reaction is reversible; in practice the equilibrium can be shifted to the right by using a large excess of water, high temperatures, and high pressures.

Hydrolysis is catalyzed by inorganic and organic acids, e.g., sulfonated hydrocarbons. In the enzymatic hydrolysis of glycerides with pancreatic lipase, the fatty acid groups in the 1- and 3-positions, are split off preferentially.

A fat can also be hydrolyzed with alkali (saponification, see also → Soaps):

$$\begin{array}{l} CH_2OCOR \\ | \\ CHOCOR + 3NaOH \longrightarrow \\ | \\ CH_2OCOR \end{array} \quad \begin{array}{l} CH_2OH \\ | \\ CHOH + 3RCOONa \\ | \\ CH_2OH \end{array}$$

4.2. Interesterification

Like other esters, glycerides can be transesterified by acidolysis or alcoholysis (→ Esters, Organic, Chap. 3.2.). In the presence of an alkaline catalyst and an excess of glycerol, triglycerides form a mixture of mono- and diglycerides (alcoholysis).

The acyl groups of glycerides can also be exchanged inter- and intramolecularly without addition of acids or alcohols (interesterification).

$$\begin{array}{l} CH_2OCOR^1 \\ | \\ CHOCOR^1 \\ | \\ CH_2OCOR^1 \end{array} + \begin{array}{l} CH_2OCOR^2 \\ | \\ CHOCOR^2 \\ | \\ CH_2OCOR^2 \end{array} \rightleftharpoons$$

$$\begin{array}{l} CH_2OCOR^2 \\ | \\ CHOCOR^1 \\ | \\ CH_2OCOR^1 \end{array} + \begin{array}{l} CH_2OCOR^1 \\ | \\ CHOCOR^2 \\ | \\ CH_2OCOR^2 \end{array} \rightleftharpoons \text{etc.}$$

Even at 200 – 300 °C interesterification proceeds very slowly, but the reaction can be accelerated by using an alkaline catalyst such as a metal alkoxide. With such a catalyst the reaction is complete within one minute at 80 °C. Interesterification is of practical importance since it enables the physical properties of a fat, e.g., melting behavior and consistency, to be modified without changing the fatty acids chemically, as occurs in hydrogenation (hardening).

Interesterification may be either random or directed. *Random interesterification* leads to a random distribution of the fatty acid groups over the triglyceride molecules. This is demonstrated for interesterification of equal proportions of tristearin (S – S – S) and triolein (O – O – O):

$$\begin{array}{ll} S-S-S & + \; O-O-O \\ 50\% & \quad 50\% \end{array}$$

$$\downarrow$$

$$\begin{array}{cccccc} S-S-S & S-O-S & O-S-S & S-O-O & O-S-O & O-O-O \\ 12.5\% & 12.5\% & 25\% & 25\% & 12.5\% & 12.5\% \end{array}$$

S = stearic acid, O = oleic acid

In *directed interesterification*, the temperature is reduced to such an extent that the highest melting glycerides are continously frozen out of the reaction mixture, in turn continuously shifting the reaction equilibrium. In this way a fat can be separated into higher and lower melting fractions. The higher melting fraction contains the glycerides of saturated fatty acids (stearin fraction), whereas the glycerides of unsaturated fatty acids are found in the lower melting fraction (olein fraction). Directed interesterification of stearodiolein can yield 33.3 % of tristearin and 66.7 % of triolein:

$$O-S-O$$

$$\downarrow$$

$$\begin{array}{ll} S-S-S & O-O-O \\ 33.3\% & 66.7\% \end{array}$$

S = stearic acid, O = oleic acid

4.3. Hydrogenation

The double bonds in a fatty acid chain can be wholly or partially saturated by addition of hydrogen in the presence of a suitable catalyst such as nickel, platinum, copper, or palladium. Hydrogenation always leads to an increase in melting point and is therefore also called "hardening". Partial hydrogenation can lead to isomerization of cis double bonds to trans double bonds.

The catalyst, the oil, and the hydrogen must be brought into mutual contact under suitable

temperature and pressure conditions. The reaction rate depends on mixing intensity, the type of oil or fat, temperature, catalyst activity, and concentrations of catalyst and dispersed hydrogen. Hydrogenation is an exothermic process. Industrial nickel catalysts are generally obtained by precipitation of nickel hydroxide or carbonate on kieselguhr, silica gel, alumina, or similar carriers, followed by reduction to metallic nickel, or by in situ production of metallic nickel from nickel formate. Such heterogeneous catalysts have a large activated surface. During hydrogenation the double bonds form transient complexes with the active centers of the catalyst. These complexes disintegrate after reaction of the double bonds with hydrogen, leaving the catalyst in its original form [26]. The active centers of the catalyst can be inactivated or poisoned by a number of compounds such as phospholipids, sulfur compounds, organic acids, and oxidized lipids.

A fatty acid with several double bonds, such as linolenic acid ($C_{18:3}$), is hydrogenated more quickly to linoleic acid ($C_{18:2}$) or oleic acid ($C_{18:1}$) than is linoleic acid to oleic acid or oleic acid to stearic acid ($C_{18:0}$). The reaction sequence occurring during hydrogenation can be represented schematically as follows:

$$C_{18:3} \xrightarrow[+H_2]{k_1} C_{18:2} \xrightarrow[+H_2]{k_2} C_{18:1} \xrightarrow[+H_2]{k_3} C_{18:0}$$

The term selectivity is used to indicate which of these reactions is fastest. Selectivity I is defined as the ratio k_2/k_3; it is related to the proportion of saturated glycerides formed and to the melting behavior of the product. Selectivity II, expressed as the ratio k_1/k_2, must be as high as possible if the concentration of linoleic acid in the hydrogenated product is to be maximized [27].

Selectivity can be influenced by the catalyst type (surface area, pore size, etc.) and by altering the reaction conditions. An increase in selectivity, i.e., an increase in partial hydrogenation, promotes isomerization of cis to trans double bonds.

At temperatures above 200 °C and with a low hydrogen concentration, catalytic hydrogenation of polyunsaturated fatty acid groups can lead to the formation of traces of cyclic aromatic compounds [28].

The double bonds of substituted fatty acids such as ricinoleic acid can also be hydrogenated under suitable reaction conditions. Cyclopropane or cyclopropene groups behave as double bonds and lead to branched fatty acids on hydrogenation.

Iron pentacarbonyl and cobalt octacarbonyl are examples of homogeneous hydrogenation catalysts.

Reduction with hydrazine does not lead to isomerization; there is also no selectivity ($k_1 = k_2 = k_3$).

4.4. Isomerization

Naturally occurring fatty acids exist predominantly in the cis form. An equilibrium mixture in which the higher melting trans form predominates can be formed by heating to 100 – 200 °C in the presence of catalysts such as nickel, selenium, sulfur, iodine, nitrogen oxides, or sulfur dioxide.

If selenium or oxides of nitrogen and sulfur are used in the cis – trans isomerization (elaidinization) of oleic acid, there is virtually no positional isomerization. However, cis – trans isomerization of linoleic and linolenic acid leads to conjugated double bonds.

Nonconjugated systems can be isomerized into conjugated systems by heating in an alkaline solution at 200 °C (→ Fatty Acids, Chap. 2.1.). If reaction times and temperatures are extended, linolenic acid can be converted into cyclohexadiene and benzene derivatives:

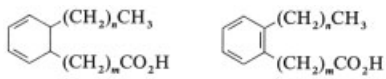

Isomerization can occur if oils and fats are heated at temperatures above 100 °C in the presence of bleaching earth, kieselguhr, or activated charcoal.

4.5. Polymerization

Dimeric, oligomeric, and polymeric compounds are formed by heating unsaturated fatty acids at 200 – 300 °C [29]. The rate of polymerization increases with increasing degree of unsaturation; saturated fatty acids cannot be poly-

merized. Thermal polymerization of polyunsaturated fatty acid groups is normally preceded by isomerization and conjugation of double bonds. Thermal polymerization involves formation of new carbon–carbon bonds by combination of acyl radicals and by Diels–Alder reactions, while oxidative polymerization involves formation of C–O–C bonds. Thermal dimerization is catalyzed by Lewis acids such as boron trifluoride; industrial processes for dimerizing oleic acid are based on this principle [30]. Heating of oils during refining or during household use does not lead to a significant increase in dimeric triglycerides. Up to 2 % dimeric triglycerides can be encountered in fresh raffinates; these dimers are not toxic, and are largely excreted as such.

4.6. Autoxidation

Autoxidation, the oxidation of olefins with oxygen, plays a decisive role in the development of rancidity, off-flavors, and reversion flavors in oils and fats during their production and storage. Autoxidation of oil-containing products such as oilseeds and spent bleaching earths can lead to their spontaneous combustion. Autoxidation of drying oils is an important initial stage of polymerization leading to stable surface films (→ Drying Oils and Related Products, Chap. 3.).

Autoxidation involves the formation of a hydroperoxide on a methylene group adjacent to a double bond; this step proceeds via a free-radical mechanism:

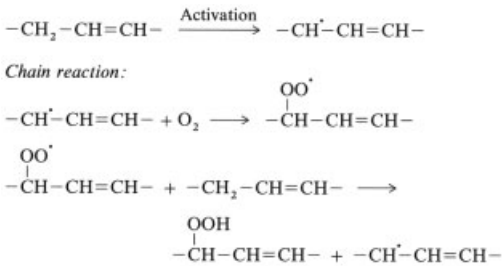

Autoxidation is characterized by an induction period during which free radicals are formed. This phase is triggered by light (photooxygenation), heat, and the presence of compounds that readily form free radicals (e.g., hydroperoxides, peroxides, and transition metals). Photooxygenation, i.e., light-induced oxidation, leads to a particularly fast buildup of radical concentration. The formation of singlet oxygen under the influence of short-wave radiation and a sensitizer such as chlorophyll or erythrosine probably plays a key role in this reaction.

The reactivity of a methylene group in forming a hydroperoxide is enhanced by a second adjacent double bond. Hence linoleic acid oxidizes 10 to 20 times faster than oleic acid. Linolenic acid reacts about three times faster than linoleic acid, since two doubly activated methylene groups are present. The olefin radical formed subsequently isomerizes; in a 1,4-diene such as linoleic acid, isomerization leads to conjugated hydroperoxides:

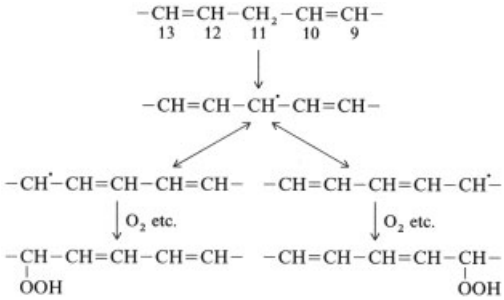

The intermediate hydroperoxides are labile compounds that decompose into a number of different products: epoxides, alcohols, diols, keto compounds, dicarboxylic acids, aldehydes, and isomerization and polymerization products. The volatile carbonyl compounds formed in this process are responsible for the taste and odor of oxidized oils and fats.

When the radical concentration has reached a certain limit, the chain reaction is gradually stopped by mutual combination of radicals.

Antioxidants prolong the induction period by reacting with the intermediate products of the chain reaction, forming inactive radicals. Tocopherols are naturally occurring antioxidants. Butylhydroxyanisole (BHA), butylhydroxytoluene (BHT), and propyl gallate are among the most effective synthetic antioxidants (→ Antioxidants). Certain organic or inorganic acids (citric, tartaric, ascorbic, phosphoric acids) have a synergistic effect without being true antioxidants; this effect is presumably based on the inactivation of trace metals or reduction of oxidized antioxidants.

5. Manufacture and Processing

Oils and fats are either of vegetable or animal origin. The approximate proportions of the corresponding world production are 55 % vegetable oils, 40 % land-animal fats, and 5 % marine oils.

5.1. Vegetable Oils and Fats

5.1.1. Storage and Handling of Raw Materials

The handling of oleaginous seeds and fruits during transport and storage has a decisive influence on the quality of the crude oils. The oils in fruit pulp (e.g., olive and palm) are very susceptible to enzymatic hydrolysis since they are finely dispersed in moist cell tissue. The oils contained in seeds are more stable but can also be attacked. A high moisture content accelerates the uptake of oxygen and the corresponding release of carbon dioxide due to degradation of starch. The heat generated in this reaction promotes lipolysis, growth of microorganisms (which can lead to the formation of mycotoxins such as aflatoxins), formation of undesirable color and odor, and – in extreme cases – may result in a coagulation of seed in a silo and even spontaneous combustion [31], [32]. Damage of seed cells, mechanically or by pests, also promotes lipolysis and lipoxidation. Free fatty acids, oxidation products, and coloring matter formed by lipolysis, lipoxidation, and degradation of protein and carbohydrate can in general be removed from the oil only with difficulty, with corresponding yield losses.

In order to minimize these effects, oilseeds must be dried before storage, preferably under mild conditions to avoid cellular damage. The critical moisture content roughly correlates with the hygroscopic equilibrium at 75 % relative humidity and varies between 6 and 13 %, depending on the protein and carbohydrate content. In the case of copra, rapeseed, and sunflower seed, the critical moisture content is 7 %; in the case of soybeans 13 %.

Transport and handling in bulk has almost entirely superseded that in bags. Seed is normally stored in concrete silos that are ventilated to avoid local generation of heat and water pockets; such silos can be up to 70 m in height and up to 12 m in diameter. The silos are filled from the top, emptied from the bottom, and normally equipped with special vibrating devices at the bottom to minimize bridging.

5.1.2. Cleaning and Dehulling

On arrival at the oil mill, oilseeds still contain plant residues, dust, sand, wood, pieces of metal, and foreign seed, which must be removed prior to further processing by screening, air classification, and passage over magnets.

Some oilseeds are dehulled (decorticated) before being further processed, especially if only the oil is to be expelled, since the hulls tend to retain part of the oil. Preexpelling and subsequent extraction with solvents is the most common process; a certain proportion of hulls in the material to be extracted can be desirable since they facilitate percolation. An excessively high proportion of hulls can, however, impair the quality of the crude oil since the solvent also dissolves wax from the seed coat. A further reason for dehulling can be the need to produce so-called high protein meal, for example from soybeans.

Dehulling is normally preceded by relatively intense drying of the seed material to help loosen the seed coat from the meats. Dehulling is normally performed by using impact disintegration or passage through rolls to break up the seeds. The hulls are separated by screening and air-classification.

5.1.3. Expelling

Modern processing of oilseeds generally involves a combination of expelling and solvent extraction. Seed is normally preexpelled to a residual fat content of ca. 20 %, and the expeller cake is then extracted to 1 – 2 % fat. Preexpelling is omitted if the fat content of the seed is 20 % or less. High-pressure expelling down to ca. 4 % residual fat is no longer important, mainly because of the higher processing costs.

Most of the fat present in seeds and fruit pulp is in the endosperm and hypocotyl cells, with much less in the seed coat. The processing conditions are largely determined by the size and stability of the oil-containing cells. With palm fruit and olives, mere heating or boiling with

water suffices to burst the membranes of the oil-containing cells to liberate the oil. Nearly all oleaginous seeds must, however, first be comminuted and thermally pretreated to isolate the fat in an acceptable yield. Comminution partially destroys the cellular structure, increasing the total internal surface and facilitating access to the oil within the seed.

Coarse materials such as copra are first broken by passage through interlocking rolls. Fluted rolls are used for nearly all other seeds. Smooth or so-called flaking rolls are used to complete seed comminution. These three types of rolls are shown in Figure 3; they can be arranged in parallel, diagonally, or sequentially. In the more modern, parallel arrangement, the rotational speed of the two rolls differs by 5 – 10 % to produce an additional shear effect. The separation of the rolls and their circumferential speed is adjusted to the type of seed and the degree of cell rupture required for the subsequent expelling and extraction.

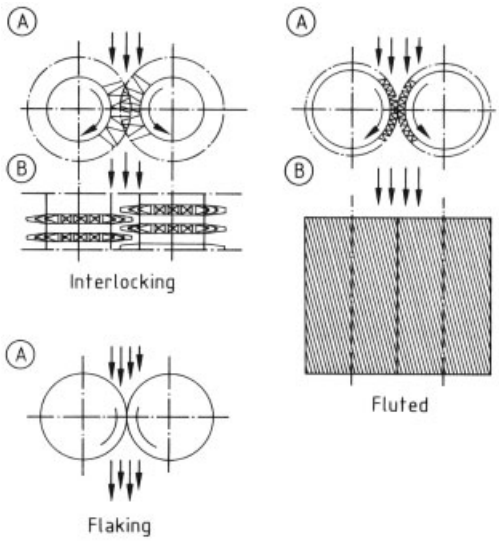

Figure 3. Schematic representation of various rolls used for expelling oilseeds
A) Side view; B) Top view

The comminuted seed is conditioned prior to expelling by moistening and heating in suitable equipment. During this treatment lipoproteins decompose, proteins coagulate, and intracellular oil bodies coalesce. The cell walls themselves are not macerated during conditioning. Coagulation of protein facilitates expelling and subsequent percolation in the extractor. Undesirable enzymes and microorganisms are deactivated at temperatures above 80 °C and in the presence of sufficient moisture [33].

By suitably conditioning cottonseed, the toxic gossypol can be deactivated by association with denatured protein. In rapeseed, suitably severe conditioning deactivates the enzyme myrosinase and hence improves the quality of the meal and the oil. However, excessive heating impairs the nutritive quality of the expeller cake as well as the color and taste of the expeller oil and must be avoided. An excessively high moisture content can reduce the yield of expeller oil. Each type of seed must therefore be adjusted to an optimal moisture content and temperature during conditioning.

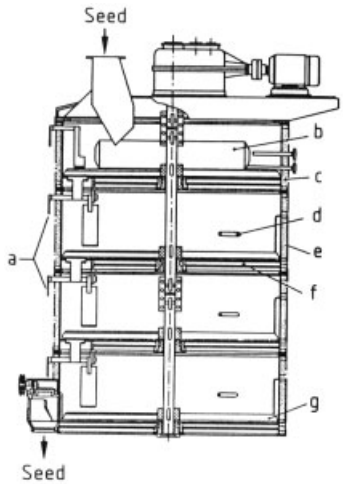

Figure 4. Vertical stack cooker for conditioning oilseeds (Krupp)
a) Level control; b) Steam heating; c) Insulation; d) Steam jet; e) Manhole; f) Steam heating; g) Sweep stirrer

Conditioning is performed in horizontal jacketed tubes (conditioners) or in vertical stack cookers. The stack cooker basically consists of up to five kettles arranged in a tier and through which the seed material successively moves from the top to the bottom; each kettle is equipped with a sweep stirrer and facilities for direct and indirect steam heating (Fig. 4). Generally, wet cooking with direct steam takes place in the top kettle, and drying, with increased temperatures and venting, in the bottom kettles.

Expelling (pressing) separates the oil from the solid phase, the so-called expeller cake, and

can be done in continuous or batch presses. Levers, wedges, or screws were used to apply pressure in the more primitive types of batch presses, but modern batch presses are almost invariably operated by a hydraulic system. They can be divided into the open type, in which the seed must be wrapped in press cloths, and the closed type, which dispenses with press cloths and confines the material in a cage-type construction. Open-type presses can be subdivided into plate and box presses, closed types into pot and cage presses.

Continuous expellers or screw presses (e.g., Simon – Rosedowns, Krupp) have generally replaced batch presses. Deoiling in these presses is efficient and uniform since the layer of material to be expelled is relatively thin and is continuously broken up. Several types exist, all with the common principle of a worm shaft rotating in a cylindrical drainage barrel (Fig. 5). The barrel is composed of armored, rectangular bars that fit into the barrel bar frame. The individual bars are separated by bar spacing clips, the specific spacings – ranging from 0.1 to 0.4 mm – depending on the type of seed and degree of expelling required. The required pressure is built up either by a gradual increase in the diameter of the worm shaft or by a gradual restriction of the barrel diameter. Specific pressure profiles and cake thickness can be further controlled by variations in the pitch of the worm and choke mechanisms in the barrel. The expelled oil is discharged through the barrel spacings while the expeller cake emerges at the end of the barrel.

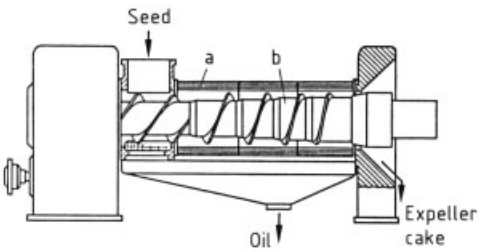

Figure 5. Screw press (Krupp)
a) Drainage barrel; b) Worm shaft

The capacity of screw presses varies widely, being dependent primarily on the cross-sectional ratio and the speed of the screw. The pressure should be built up fairly slowly. Initially, the channels in the expeller cake are sufficiently large for ready discharge of oil; in the latter sections of the screw press, the pressure is increased to about 300 MPa (3000 bar) to expel more tightly bound oil and to burst intracellular oil bodies.

Up to 200 t of seed per day per press is normally preexpelled to a residual fat content of 15 – 25 %. The expeller cake is then comminuted on fluted rolls, and possibly also pelletized, prior to solvent extraction.

Screw presses used to reduce the residual fat content to 3 – 6 % are similar in design to preexpellers but have a larger compression ratio (1 : 25 instead of 1 : 15) and have a lower worm speed (9 – 12 rpm versus 30 – 45 rpm).

A specific worm arrangement normally gives optimal results for only one particular type of seed. Difficulties encountered on changing the type of seed can often be overcome by modifying conditioning and comminution. Preexpellers that do not require seed conditioning of sunflower seed have been developed; however, they do not work as well with unconditioned rapeseed.

The expeller oil still contains so-called foots, which must be removed by filtration.

5.1.4. Extraction [34], [35]

Very much lower residual fat contents can be achieved by solvent extraction than by expelling. Seed material with an oil content of ca. 20 % and expeller cake with 15 – 25 % fat are usually subjected to solvent extraction. Both must be flaked prior to extraction. This pretreatment creates a large internal surface and yields thin (ca. 0.3 mm), firm flakes, which form a loose layer in the extractor and provide uniform, short diffusion pathways. The flakes must have a minimum moisture content and elasticity to prevent them from crumbling during transport to the extractor, which would lead to too dense a packing and hence poor percolation of solvent. However, a high moisture content also impairs percolation.

Ideally, the extraction solvent should dissolve only glycerides but not undesirable components such as coloring matter, gums, and phospholipids. The solvents must not contain toxic components, and should be recoverable with minimum loss, be safe in handling, and be readily removable from the extracted material. For these reasons aliphatic hydrocarbons, especially

hexane, are used almost exclusively. Technical hexane with a boiling point range of 55 – 70 °C has proved to be optimal. Hexane can be readily removed from the oil at temperatures below 100 °C in vacuo and can be stripped from the meal with steam. The solubility of hexane in the condensed water is only 0.1 %.

For special purposes, e.g., the production of heat-labile pharmaceuticals, lower-boiling hydrocarbons such as pentanes are sometimes used. Extraction with propane or carbon dioxide under supercritical conditions is reserved for special products of high intrinsic value because of the high equipment costs involved. Castor oil, being relatively polar, is preferably extracted with the higher-boiling heptane. Alcohols (methanol, ethanol, propanol, and butanol) and furfural are specially suited for the extraction of relatively wet materials. Extraction of oilseeds with alcohols leads to relatively high concentrations of phosphatides, glycolipids, carbohydrates, and similar constituents in the crude oil, although the glycerides can in principle be concentrated by cooling or extraction of the alcoholic solutions. Alcohols are generally not used as primary extraction solvents. However, they are occasionally used as secondary extraction solvents to remove gossypol from cottonseed meal, thioglycosides from rapeseed meal, sugars from soybean meal (to produce protein concentrates), and alkaloids from bitter lupine meal.

Chlorinated hydrocarbons such as trichloroethylene and dichloromethane are of interest because of their safety in handling and high extraction capacity but do not appear to be used for extraction of oilseeds because of the potential toxicity risk from residual solvent in the meal. Processes in which the oil is displaced by hot water are known but have not achieved industrial importance [36].

Oilseeds are generally extracted in a countercurrent process: pure solvent is contacted with material that has already been largely extracted, and the oil-rich solution is contacted with nonextracted material. Such extractions can be performed either continuously or in a batch process.

Batch processes. In a discontinuous process the solvent successively passes through a battery of up to batch extractors (e.g., from Buss). Each of these extractors is a cylindrical vessel with a large diameter : height ratio and has an opening for filling and emptying, a screen in the bottom part, and facilities for pumping solvent in and out, for direct and indirect steam heating, and a sweep stirrer to move the extracted meal during solvent stripping (desolventization) (Fig. 6).

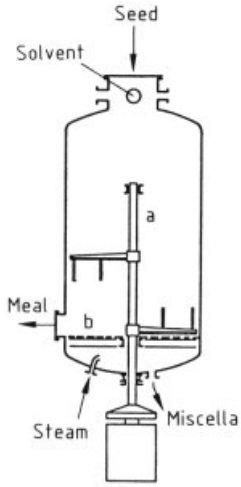

Figure 6. Batch extractor
a) Stirrer; b) Screen

After filling the extractor with comminuted seed material, the upper opening is closed and solvent pumped in. The displaced air escapes via a circuit; the hexane vapors in the displaced air are recovered. After the extractor has been filled with solvent, the hexane containing the extracted oil is pumped into a second extractor filled with seed material. This process is repeated until the hexane solution has been pumped through 3 or 4 extractors. The solvent strongly enriched with oil (the miscella) from the last extractor is filtered to remove seed particles and then distilled. In this type of process, exhaustive extraction can yield oil concentrations of at least 35 % in the final miscella.

When the contents of the first extractor have been extracted sufficiently, the solvent tap is closed; the pure extraction solvent is then passed directly to the second extractor while the fifth extractor becomes part of the battery. After discharging the solvent, remaining solvent is stripped from the extracted seed material by passing in steam. When the vapors leaving the extractor are free of solvent, the extractor is emptied via the lower opening.

The sequence of filling, extraction, desolventization and emptying is repeated for each ex-

tractor in a battery in such a way that while one extractor is being filled, the contents of three or four extractors are extracted, residual solvent is stripped in one or two extractors, and one extractor is emptied.

The vapors from the extractors and the distillation are condensed. The resulting hexane – water mixture is separated mechanically and the hexane recycled. Such a battery with 10 extractors of 7 m^3 each can reach a daily throughput of up to 1200 t seed, depending on the type of seed. The quantity of hexane circulated is ca. 100 t. The solvent loss is normally between 0.35 and 0.45 %, based on seed input.

These discontinuous plants can be automated, and have the principal advantage that special treatment of meal (e.g., detoxification of aflatoxin-containing peanut or cottonseed meal with ammonia) can be carried out fairly easily. Their capacities cannot, however, match those of fully continuous plants, which can process up to 200 t seed per hour.

Continuous Processes. Of the many fully continuous processes that have been proposed and developed, countercurrent gravity percolation has become the most popular because of its high capacity. The comminuted seed material is transported through the extractor and sprayed with solvent in the various extraction stations. The solvent becomes increasingly enriched with oil and is passed through the seed bed repeatedly according to the number of extraction stations. This arrangement has the advantage that the miscella is continuously cleaned by filtration through the seed bed.

The MIAG basket conveyor extractor (Fig. 7) uses a horizontal arrangement of baskets in which the flaked seeds are first sprayed with solvent or miscella and then passed through a solvent bath. The baskets containing the extracted and drained flakes are automatically inverted and discharged into a hopper from which the meal is conveyed to the desolventizers and dryers.

In modern plants with a capacity of up to 5000 t of seed per day, circular cell and traveling belt horizontal extractors are generally preferred. The Blaw – Knox Rotocel extractor is especially popular because it offers a large capacity within a relatively small space. It uses baskets that travel in a horizontal circle. After countercurrent extraction, the baskets are emptied through hinged sieve bottoms.

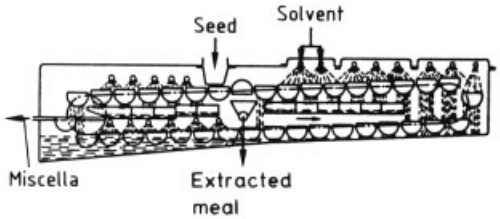

Figure 7. MIAG basket conveyor extractor

The carousel extractor from Extraktionstechnik operates on a similar principle (Fig. 8), but has a fixed discharge point instead of movable basket bottoms.

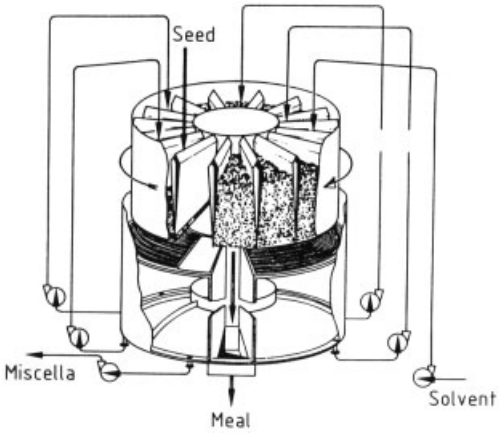

Figure 8. Carousel extractor

In traveling belt extractors, such as the De Smet extractor, the seed material is conveyed on an endless sieve belt, the height of the bed being regulated mechanically. Miscella and solvent are sprayed onto the seed material. Miscella percolates throught the belt, falls into compartments in the bottom of the extractor housing, and is picked up by a series of pumps and recirculated countercurrent to the flakes.

The Lurgi frame belt extractor works on a similar principle [37]. The endless frame belt runs over two bar-sieve belts which form the bottoms of the frames. The flaked oilseeds are filled into the frames of the upper belt, tipped onto the lower belt after half the extraction time has elapsed, and then discharged after completion of extraction and passage through a dripping zone. This separation into distinct extraction steps and

intermediate decompaction effects a particularly uniform extraction.

There are numerous other types of extractors which are of less practical importance. In Hildebrandt, Schlotterhose, Ford, Detrex, Adler, and Olier extractors, the comminuted seed material is transported through a solvent bath by a screw; Figure 9 shows the principle involved. In a Kennedy extractor (Fig. 10), the material to be extracted is conveyed to various sections of a trough by means of impeller wheels. In the Sherwin – Williams process the finely comminuted seed material is thoroughly mixed with solvent in three stages and then separated centrifugally from the miscella. The structure of the seed material is not as critical for these extractors as for basket or belt-type extractors. However, they require a far greater ratio of solvent to seed material. Their miscella are also contaminated with fines, which requires a separate filtration step. These types of extractors are therefore virtually obsolete.

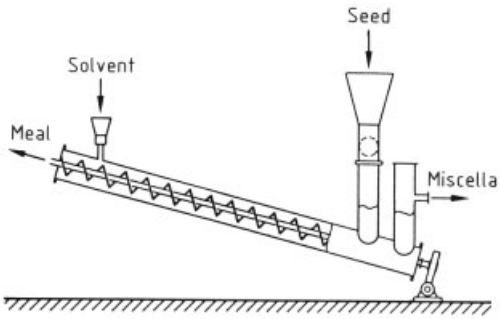

Figure 9. Screw extractor

Böhm, Tyca, Allis – Chalmers, Bonotto, and Anderson extractors are essentially columns divided into a number of sections in which the seed material moves from the top to the bottom, countercurrent to a rising flow of solvent. These extractors do not match the performance of the large basket or belt-type extractors.

Several attempts have been made to develop processes obviating preexpelling. In the Filtrex process, which is used to a limited extent, oilseeds such as rapeseed, sunflowerseed, or peanuts can be extracted directly to about 1 % residual fat without preexpelling, provided the seed material has been flaked to about 0.1 mm (the cell structure is probably destroyed) and subsequently conditioned by "crisping" [38]. A drawback is that emersion-type extractors are required. In the Direx process the seed material is only coarsely ground, extracted to 14 – 16 % residual fat with hexane, flaked, and then extracted to ca. 1 % residual fat in a second extraction step. The theory is that flaking the hexane-moist material bursts the oil cells. This process has been applied only on a small scale (ca. 5 t/h).

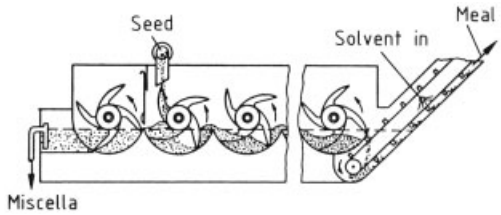

Figure 10. Kennedy extractor

A more recent development is the Alcon process [39] for soybeans, in which the essential step is to heat the flakes to ca. 100 °C with direct steam, followed by drying. The special flake conditioning prior to extraction deactivates enzymes such as phospholipases and lipoxygenase, and presumably dissociates lipoprotein complexes; this leads to lower residual fat contents in the meal and a more easily deslimable crude oil with improved oxidation characteristics.

Extracted oilseeds can be desolvized in a series of horizontal steam-jacketed tubes. The material is propelled through the tubes by screws and is then passed through toasters, similar to stack cookers, that are heated by both live and indirect steam. Alternatively, desolventization can be performed in one step in desolventizers – toasters. After drying and cooling, the meal is ready for use as an animal feed component. Toasting not only reduces residual solvent to levels below the lower explosion limit but also deactivates nutritionally undesirable enzymes such as urease and trypsin. In some cases, superheated solvent vapor is used for desolventization. This process and flash desolventization are used when minimal protein denaturation is desirable.

Solvent is recovered from the filtered miscella by distillation in multiple-stage evaporators, which are constructed to minimize thermal damage to the oil. The miscella is first concentrated to 96 % oil and finally stripped with injected steam in a falling-film evaporator. The vapors from the toaster, dryer, cooler, and dis-

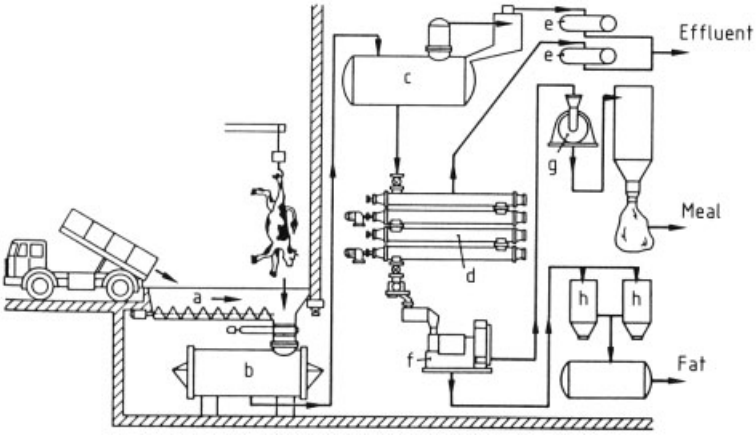

Figure 11. Kontipress system for working up carcasses (Krupp)
a) Raw material; b) Cooker – sterilizer; c) Buffer; d) Dryer; e) Condenser aggregate; f) Screw press; g) Meal grinder; h) Buffer

tillation unit are condensed and separated into water and hexane, which is recycled. Vented air is passed through special adsorbents such as activated charcoal or mineral oil to recover solvent. Total solvent loss is 0.2 – 0.3 %, based on seed input.

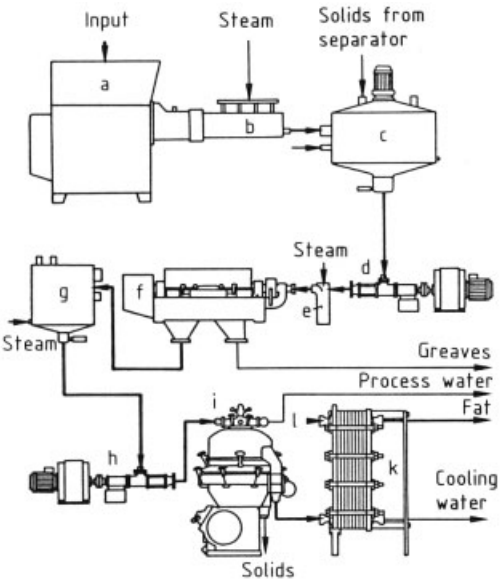

Figure 12. Alfa – Laval Centriflow rendering plant
a) Cutter; b) Renderer; c) Buffer; d) Propeller pump; e) Steam heater; f) Decanter; g) Buffer; h) Propeller pump; i) Separator (centrifuge); k) Plate heat exchanger; l) Cooling water

5.2. Land-Animal Fats

Animal fats can be readily isolated from tissue by heating since the cell membranes of animal cells are much weaker than those of plant cells. The intracellular fat expands on heating and bursts the membrane. Extraction with solvents is limited to offal. Fatty tissue is particularly susceptible to decomposition so that rapid processing is necessary if refrigeration is not possible. Microbial and autolytic degradation rapidly lead to oxidation, hydrolysis, color deterioration, and unpleasant odors and flavors, which are difficult to remove from the fat. The free fatty acid content therefore serves as a convenient, primary quality criterion.

In *dry rendering*, the comminuted fatty tissue is digested by cooking in steam-jacketed vessels equipped with agitators [40]. Local overheating of fatty tissue must be avoided since this can easily lead to unpleasant off-flavors. In modern automated plants for working up animal byproducts, the raw material is first heated and sterilized in a dry-rendering vessel. The resulting sludge is then separated, possibly after predrying, into crude meal (cracklings) and a liquid phase containing fat and water by a continuous screw press (see Fig. 11, Krupp Kontipress process) or by a decanter, i.e., a centrifuge with a built-in screw conveyor (Alfa–Laval Centrimeal process). The fat is finally separated from the aqueous phase by centrifuging. The 7 – 12 % residual fat con-

tent of the meal can be reduced to 1 – 4 % by solvent extraction.

Wet rendering, which involves treating the fatty tissue with direct steam, tends to give better yields and quality than dry rendering. In modern continuous rendering plants (e.g., as built by Alfa – Laval, Westfalia, and Sharples) selected tissue is coarsely comminuted and heated to 50 – 60 °C. It is then heated quickly to 80 – 90 °C with direct steam to deactivate oxidizing enzymes. The greaves are then separated in a decanter centrifuge, and the fat is clarified by further centrifuging, cooled, possibly texturized, and packed (Fig. 12). Fat obtained by this method in up to 99 % yield contains 0.1 – 0.2 % moisture; the content of free fatty acids is virtually identical with that of the raw material. This mild processing technique minimizes the formation of off-flavors. Further refining of lard or tallow for edible purposes is controlled by national legislation.

5.3. Marine Oils

Whale oil is isolated by rendering. The blubber is cooked with live and indirect steam, and the oil is then separated by centrifugation from insoluble tissue and water-soluble matter.

Similar processes are used to isolate *fish oils* [41]. Crushed fish (herring, menhaden, sardines, pilchards) with a dry matter content of 15 – 22 % is first cooked in closed vessels at relatively low pressure. This operation sterilizes the fish and coagulates the protein to produce a fish mass that can be easily pressed. Subsequent screw pressing separates most of the water and oil from the press cake. The press water is prepurified over screens or with a decanter (desludger) and then separated by centrifugation into crude fish oil and stickwater (or glue water). The press cake (ca. 50 % water) is dried, sometimes also extracted with a solvent, and ground into fish meal. Evaporation of the stickwater gives solubles which are added to the meal. Industrial plants are offered by Alfa – Laval (Centrifish) and Westfalia (see Fig. 13)

Liver oils are usually obtained on board ship by mild rendering of crushed liver, sometimes assisted by pressure-release disintegration of the oil cells, and subsequent separation in centrifuges. The Solexol process used in South Africa employs liquefied propane as the extraction solvent. This method can also be used to fractionate fishliver oils and to isolate vitamin concentrates.

5.4. Synthetic Fats

Pure synthetic glycerides [20] and phospholipids have few industrial uses. Synthetic fats for edible purposes are of little importance with the exception of specific dietary products such as medium-chain triglycerides of C_8 and C_{10} fatty acids (MCT fats), which are administered in some postsurgical cases [42].

6. Refining

Crude oils and fats obtained by expelling, extraction, or rendering contain trace components which are undesirable for taste, stability, appearance, or further processing. These substances include seed particles, dirt, phosphatides, carbohydrates, proteins, fatty acids, trace metals, pigments, waxes, oxidation products of fatty acids, and toxic components such as polycyclic aromatic hydrocarbons, gossypol, mycotoxins, sulfur compounds (in fish and cruciferous oils), and pesticide residues. The aim of refining edible fats and fats for industrial purposes is to remove these undesirable components as far as possible without significantly affecting the concentration of desirable constituents such as vitamins and polyunsaturated fatty acids, and without significant loss of the major glyceride components. Some oils, such as olive oil, butterfat, and cold-pressed sunflower oil or linseed oil, are not refined in order to preserve their typical flavors.

Refining usually involves the following stages:

1) Precleaning to remove phosphatides (degumming)
2) Neutralization by treatment with lye or by distillation
3) Decolorization by adsorptive treatments (bleaching)
4) Deodorization or stripping in vacuo

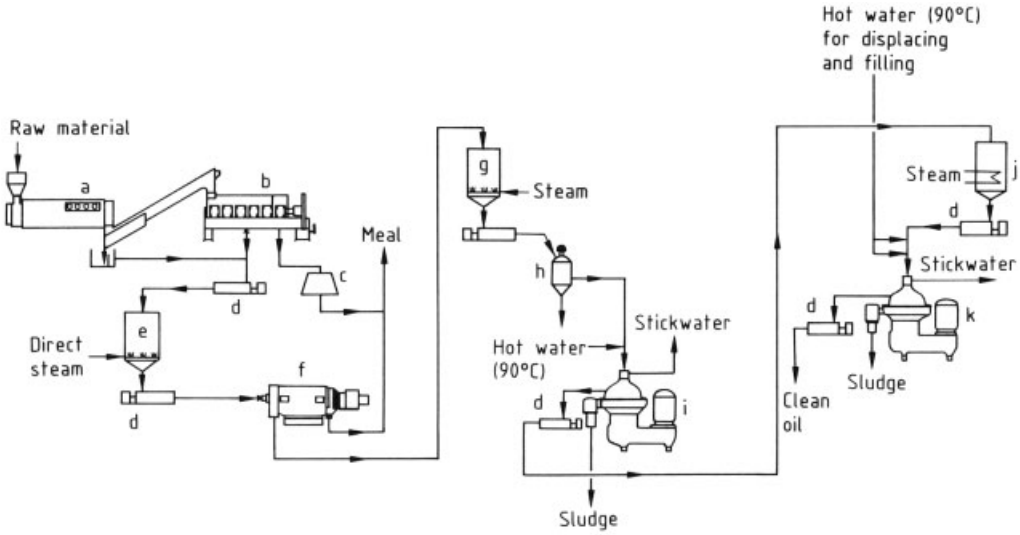

Figure 13. Fish meal and fish oil plant (Westfalia)
a) Cooker; b) Power press; c) Crusher; d) Eccentric screw pump; e) Preheater; f) Decanter for clarification of press water; g) Press water preheater (92 °C); h) Rotary brush strainer; i) Separator with self-cleaning bowl (1st stage, press water deoiling); j) Crude oil preheater (92 °C); k) Separator with self-cleaning bowl (2nd stage, oil polishing)

6.1. Degumming [43]

Phosphatides, gums, and other complex colloidal compounds can promote hydrolysis of an oil or fat during storage and also interfere with subsequent refining. They are therefore removed by degumming or desliming.

The degumming method depends on the type of the oil and the phosphatide content. In hydration degumming, which involves treating the oil or fat with water or steam, phosphatides absorb water to form a sludge that is insoluble in the oil. This process is used to remove soy lecithin (1 – 3 %) from soybean oil: 2 – 5 % water, based on oil, is intimately mixed with the crude oil at 70 – 80 °C; after a contact time of 1 – 30 min the phosphatide sludge is separated by centrifuging. The degree of hydration can be increased by using acids such as citric or phosphoric, bases, or salts but at the expense of lecithin quality. Such modified hydration processes are therefore confined to post-degumming and to oils whose lecithins have little commercial value. Degumming and neutralization can also be carried out together but usually with increased losses of neutral oil owing to the formation of emulsions. The use of sulfuric acid is generally restricted to industrial oils that are not to be further refined.

Some technical oils such as linseed oil can be degummed by heating to 240 – 280 °C; the precipitation of gums on heating is referred to as oil "breaking."

Precleaning of a crude oil by treatment with bleaching earth or other adsorbents, normally in combination with acids such as phosphoric or citric, can be advantageous.

Oils that are refined by distillation (physical refining, see page 32) should have a very low phosphatide content. This can be achieved by special degumming techniques involving the addition of citric acid followed by hydration at ca. 20 °C for several hours before centrifugal separation [44]. Another "superdegumming" technique involves separating the phosphatide micelles from a solution of the crude oil in hexane by ultrafiltration using special polysulfone or polyacrylonitrile membranes [45].

6.2. Deacidification (Neutralization)

Commercially available crude oils and fats contain on average 1 – 3 % free fatty acids. Whereas a soybean oil can have as little as 0.5 % free fatty acids, some palm and fish oils can contain up to 10 %. The free fatty acid content of refined fats should be below 0.1 %. Although longer

chain fatty acids do not generally affect taste, the short-chain acids can impart a soapy, rancid flavor. When the free fatty acids are removed as soaps by treatment with lye, other undesirable constituents such as oxidation products of fatty acids, residual phosphatides and gums, phenols (e.g., gossypol), and aflatoxins are also "washed out." In practice, deacidification is performed mainly by treatment with lye or by distillation. Deacidification by esterification with glycerol, by selective extraction with solvents, or by adsorbents is not industrially important.

Alkali Neutralization [46]. The usual method of alkali neutralization is treatment with weak lye, either batchwise or continuously. The more concentrated the lye, the more readily are undesirable constituents taken up in the soapstock. Dilute alkaline solutions (ca. 1 N) do not saponify the oil, but greater neutral oil losses occur due to occlusion in the soapstock. The losses increase with increasing content of free fatty acids. Concentrated lye (4 – 7 N) yields a relatively concentrated soap which takes up little neutral oil, but it can saponify substantial amounts of the neutral oil, especially in those oils that have a relatively high proportion of short-chain fatty acids. The optimum lye concentration therefore depends on the quality of the crude oil and the desired quality of the refined product. The neutralization process can be further influenced by temperature, residence time, and mode and intensity of mixing the lye and the fat phase.

Batch neutralization is done in an open or closed vessel (up to 75 t content) with a conical base. The reactor is fitted with a heating mantle, heating coils, and facilities for injecting live steam (Fig. 14). Lye and oil can be mixed without undue emulsification by using special coil and frame stirrers. Soap settles in the conical bottom of the neutralization vessel. Discharge of neutral oil with the soap can be minimized by using ultrasonic or dielectric measurement of the phases. Closed vessels can also be used for bleaching if they can be evacuated. Whereas weak lye is generally sprayed onto the oil at ca. 90 °C without stirring, stronger lye is normally stirred into the oil at 40 – 80 °C. Sodium carbonate is sometimes used instead of caustic soda; it does not saponify but tends to lead to foaming. Oils with a relatively high concentration of free fatty acids are neutralized with strong lye in order to limit the total volume. The required amount of lye, including a slight excess, is based on the free fatty acid content (ffa).

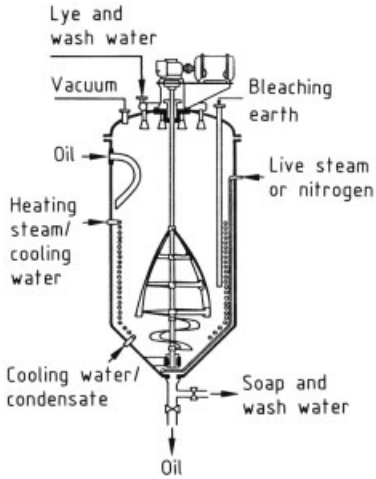

Figure 14. Combined neutralizer – bleacher

After neutralization and discharge of soapstock, the oil is washed with dilute lye (ca. 0.5 N) and then with water in order to reduce the soap content below 500 ppm. Incomplete removal of soap impairs the color of the oil and the efficacy of the following refining steps. Polyunsaturated oils such as soybean oil are sometimes subjected to posttreatment with sodium carbonate – silicate to remove residual traces of oxidized glycerides and phosphatides. Although soapstock can be utilized as such in chicken broiler feed, it is normally converted into "acid oil" (a mixture of fatty acids and neutral oil) by batch or continuous treatment with 30 % sulfuric acid at 70 °C.

The quantity of acid oils is a measure of the refining efficacy. A refining factor of 2 means that for each mass unit of free fatty acid in the crude oil, 2 mass units of acid oil are produced. The difference between acid oil and free fatty acid content is mainly due to occluded neutral oil. Acid oils are used for technical purposes and as energy carriers in animal feeds [47].

Batch neutralization is more flexible than continuous neutralization but gives lower refining yields.

Continuous processes achieve the same or greater capacities in a smaller space with better yields, provided long continuous runs can

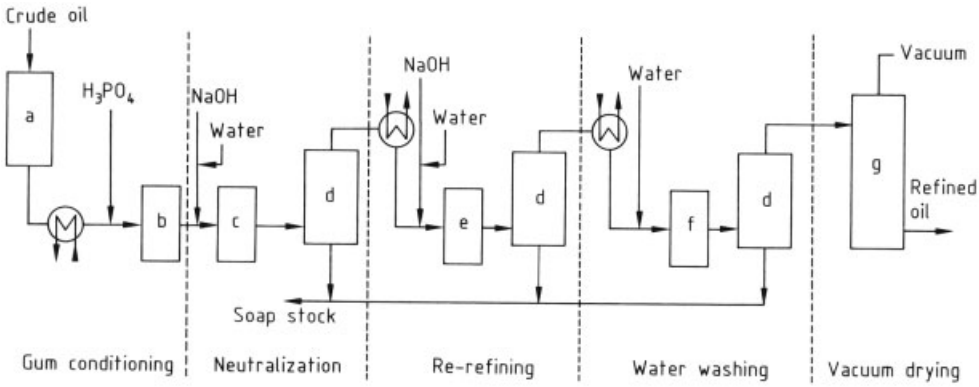

Figure 15. Flow diagram of continuous neutralization of oils and fats
a) Crude oil storage tank; b) Gum conditioning mixer; c) Neutralizing mixer; d) Separator (centrifuge); e) Re-refining mixer; f) Water-washing mixer; g) Vacuum dryer

be maintained. Continuous plants from Alfa–Laval, Westfalia, and Sharples are widely used (Fig. 15). In these plants, the oil is generally first conditioned by addition of citric or phosphoric acid in small continuous mixing chambers and then brought into contact with lye. Soap and gums are separated in centrifuges. After re-refining and passage through a washing centrifuge, the oil is practically free of fatty acid and soap. The Sharples plant uses high-speed open centrifuges with a high separation efficiency, while the Alfa–Laval short-mix lines employ completely closed centrifuges with a mean residence time of only a few seconds. The capacity can be increased considerably by using self-discharging centrifuges (e.g., from Westfalia).

The Zenith process employs a different principle: a fine stream of oil rises through a vessel filled with dilute caustic soda and is skimmed off continuously at the top. The spent lye must be replaced. This type of neutralization requires efficient pre-degumming and has the advantage of relatively low losses of neutral oil [48].

Neutralization with aqueous ammonia is an old process which is being reexamined because of environmental pollution problems with lye. The ammonium soaps can be separated by centrifuging and split into fatty acids and ammonia by heating, the ammonia being recycled [49].

Distillative Neutralization. In distillative neutralization (physical refining) the free fatty acids are continuously removed from the crude oil with water vapor in vacuo. Prior to this stripping, gums, phosphatides, and trace metals must be almost completely removed since these compounds impair the oil quality during distillation. They are preferably removed by treatment with acids and bleaching earth.

The economics of distillative neutralization improve with increasing content of free fatty acids. The fatty acid factor is approximately 1.1, thus the neutral oil losses are far lower than with alkali neutralization. An added advantage is the avoidance of soap-stock splitting and the consequent effluent problems. Distillative neutralization was formerly restricted to fairly saturated oils with a high free fatty acid content (e.g., coconut oil, palm oil, and tallow). However, it is now being extended to polyunsaturated oils such as rapeseed, soybean, and sunflower oil because techniques are available for reducing the phosphatide content of the crude oils to sufficiently low levels to obviate lye neutralization [50].

Temperatures between 240 and 270 °C are normally used to achieve free fatty acid contents of less than 0.1 %. At these temperatures decomposition and removal of carotenoid pigments result in a significant bleaching effect, this is particularly noticeable in palm oil. Older plants such as those built by Wecker and Feld & Hahn have now been generally superseded by modern plants (e.g., Lurgi, Girdler, HLS, and EMI), which have the advantage of a smaller spread in residence time and hence better process control (especially in combined deacidification, deodorization, and heat bleaching). The latter plants are very similar to the corresponding

deodorizer plants (see Figs. 20 and 21) except for an additional fatty acid condenser between the deodorizer and the vacuum assembly.

Other Processes. *Esterification* of the free fatty acids with glycerol may be economically feasible for highly acid oils. The esterification is carried out at 160 – 210 °C in the presence of catalysts such as magnesium oxide, alkali silicates, or metals (Zn or Sn).

Solvent extraction of free fatty acids is of interest only for highly acid oils. Olive oil with 22 % ffa can be deacidified to ca. 3 % ffa by liquid – liquid extraction with ethanol. Furfural can be used to dissolve polyunsaturated glycerides and fatty acids.

In the Solexol process, liquefied propane is used as solvent. Saturated glycerides are taken up in the propane phase while nonsaponifiable matter, fatty acids, oxidation products, and polyunsaturated glycerides remain largely undissolved. This process has been used to fractionate fish and fishliver oils and may increase in importance with the advent of supercritical fractionation and extraction techniques.

Various adsorbents such as silica gel and alumina have been proposed for the *selective adsorption* of free fatty acids [51]. Adsorption on strongly basic ion-exchange resins has also been used. Plant-scale application of such processes has, however, not been reported.

6.3. Bleaching

Degumming followed by alkali neutralization generally does not lead to significant decolorization of an oil or fat. A bleaching step with solid adsorbents such as bleaching earth or activated charcoal is normally employed. Bleaching with air or chemicals is not used for edible fats.

The aim of adsorptive treatment is to remove pigments such as carotenoids and chlorophyll but also residues of phosphatides, soaps, trace metals, and oxidation products such as hydroperoxides and their nonvolatile, polar decomposition products. These compounds can have a deleterious effect on the course of further processing (especially hardening and deodorization) and on the quality of the final product. Some of the oxidation products that are removed by adsorption can promote oxidation of the oil. Since some of these undesirable compounds can be removed by suitable degumming and neutralization, there is an interdependence between the pretreatment conditions and the treatment with adsorbents. This interdependence can also involve deodorization (heat bleaching) where, for example, the carotenoid content is reduced.

The adsorption of pigments follows Freundlich's isotherm: the amount of pigment adsorbed decreases with decreasing concentration of the pigment in the oil. Although adsorption should be most efficient if the oil is passed through a layer of adsorbent, in practice the adsorbent is normally stirred into the oil and then filtered off after a certain residence time.

The choice of the adsorption process and the type and concentration of adsorbent is governed by factors such as pretreatment, desired quality of the fully refined product, filtration speed of the oil, and oil retention by the adsorbent. Oil retention can be up to 50 wt % on bleaching earths and nearly 100 wt % on activated charcoal.

Natural and activated earths are the most frequently used adsorbents in the refining of oils and fats. Activated earths are obtained by leaching natural earths (predominantly aluminum silicates containing montmorillonite) with hydrochloric acid followed by washing with water, drying, and sizing. This process increases the internal surface of the bleaching earth by partial dissolution of the aluminum and iron oxides; calcium and magnesium ions are partially replaced by hydrogen ions.

The amount of bleaching earth to be used and the optimum bleaching conditions must be determined empirically. Lower concentrations (0.5 – 1.0 %) are required for activated earths than for natural earths. Activated earths catalyze oxidation of the oil. For this reason bleaching should be done below 100 °C in vacuo, and the subsequent filtration should be carried out with limited access of atmospheric oxygen. Bleaching is sometimes performed in two steps: a conditioning and adsorption stage at relatively low temperature (ca. 65 °C), and a second stage at elevated temperature (ca. 100 °C) to fix the adsorbed components onto the adsorbent surface [52].

In physical refining, bleaching is often preceded by addition of citric or phosphoric acid to chelate trace metals and to precipitate hydrated

phosphatides. This can also be achieved by using bleaching earths containing added acid.

The moisture content of bleaching earths lies between 5 and 10 %. Except for some types of natural earths, the moisture content does not appear to have a significant influence on bleaching capacity.

Activated charcoal (0.1 – 0.4 %) is sometimes used in combination with bleaching earth for oils that are difficult to bleach. Activated charcoal has also assumed importance as an adsorbent for removing polycyclic aromatic hydrocarbons from oils and fats [53]. It has been recommended for use in a fixed-bed process for the physical refining of soybean oil [54].

Bleaching and filtration is still mostly done *batchwise*. The oil is first dried in vacuo at ca. 30 kPa (30 mbar) to an optimum moisture content, frequently in the same closed vessels in which the oil was neutralized and washed (Fig. 14). The required amount of adsorbent is then drawn into the oil at 80 – 90 °C, preferably through a pipe reaching into the oil. After stirring for a few minutes to half an hour in vacuo, the oil – earth slurry is pumped over filter presses. Automated frame and plate presses or centrifugally dischargeable, closed-disk filters of the Funda (Fig. 16) and Schenk type are preferred; these also have the advantage of limiting access of air. The filter cake, containing 30 – 50 % oil, is often extracted with hot water (Thomson process) or with hexane to recover the bulk of the adsorbed oil. The quality of the recovered oil depends on the degree of unsaturation, the type of extraction, and the age of the spent earth before extraction. It is normally not economic to regenerate spent bleaching earth.

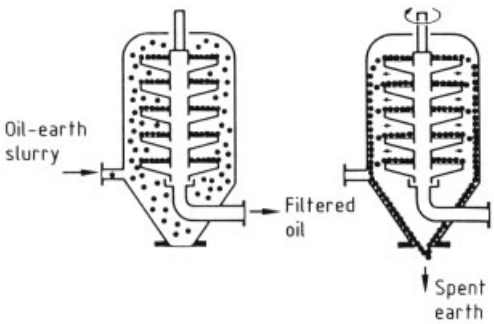

Figure 16. Closed disk filter with centrifugal discharge (Funda)

Continuous bleaching is especially relevant when the other refining steps are also performed continuously. The degummed or degummed and neutralized oil is normally dried and deaerated in vacuo and charged together with bleaching earth into an evacuated vessel where the oil – earth slurry is fed through various compartments, possibly at different temperatures, and is then filtered in a closed-disk system (Fig. 17).

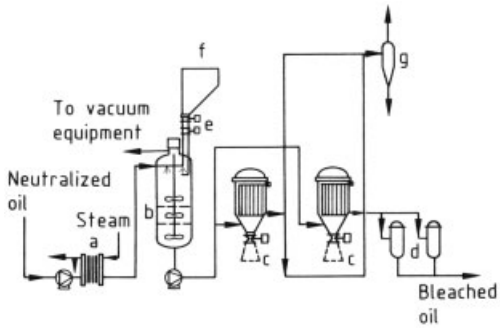

Figure 17. Alfa – Laval bleach system
a) Plate heat exchanger; b) Bleacher; c) Filters; d) Polishing filters; e) Bleaching-earth dosage equipment; f) Bleaching-earth storage tank, g) Cyclone

Bleaching with oxygen or chemicals is not practiced for edible fats and is used only occasionally for technical fats. Oxidizing agents such as hydrogen peroxide, sodium peroxide, benzoyl peroxide, potassium permanganate, chromium salts, hypochlorite, and chlorine dioxide, and reducing agents such as sulfurous acid and sodium dithionite have been proposed. Hydrogen peroxide is used when adsorptive methods do not suffice. Castor, coconut, peanut, and olive oils and certain types of tallow and bone greases can be successfully bleached with 0.05 – 0.1 % of a 20 % aqueous solution of sodium chlorite. Fats such as red palm oil are sometimes bleached by very slight hydrogenation which destroys carotenoid pigments.

6.4. Deodorization

Deodorization is the last step of the refining sequence, in which odors and flavors are removed from the bleached oil or fat. It is essentially a steam distillation process in which volatile compounds are separated from the nonvolatile glycerides. The odoriferous compounds are primarily aldehydes and ketones formed by autoxida-

tion during handling and storage and can have a flavor threshold value of a few ppm. Other volatile components such as free fatty acids, alcohols, sterols, or tocopherols are also partially removed by deodorization.

Deodorization is not, however, a purely physical process. During deodorization, flavor compounds can be formed by hydrolysis and thermal decomposition, and peroxides are decomposed by heat. These reactions presumably play a major role in the flavor stability improvement achieved by deodorization, especially in vegetable oils. In practice, therefore, the residence times and steam volumes required are significantly greater than those calculated for normal steam distillation. At the relatively high deodorization temperatures used, traces of air can initiate oxidation with severe impairment of taste and flavor stability. Because thermal decomposition products of proteins, carbohydrates, phosphatides, and soaps can also interfere with deodorization, the oil must be well prerefined, i.e., freed from phosphatides and soap.

Deodorization is performed in vacuo because of the low partial vapor pressure of the compounds to be removed. Consumption of stripping steam increases with decreasing vapor pressure of the volatile components and with increasing pressure in the deodorizer, it ranges from 5 to 10 wt % (based on the oil) for batch processes and from 1 to 5 wt % for continuous processes. Deodorization temperatures range from 190 to 270 °C and the pressure from 0.13 to 0.78 kPa (1 – 6 mm Hg). The volume of stripping steam required is 10 – 20 m^3 per kilogram of oil [24]. Carry-over losses of oil are minimized by limiting the velocity of the injected steam, by providing a sufficiently large headspace, and by baffles. The distillate removed during deodorization is about 0.2 % of the oil, and contains very little neutral oil. In a batch process a deodorizing time of at least 2 h is required for saturated or hydrogenated fats; about 4 h is needed for other oils and fats. Although attempts have been made to determine the end point of deodorization by analytical data, in practice this is done organoleptically.

The steam injected into the oil must be dry and oxygen-free. In order to deactivate prooxidative trace metals, aqueous citric acid solution (2 – 5 mg of citric acid per 100 g of oil) is often injected into the oil toward the end of the deodorization.

The deodorizer is normally constructed of stainless steel if temperatures of 190 °C are to be exceeded. Deodorizing temperatures up to 270 °C are employed, for example, to complete removal of pesticide residues [55], to heat-bleach palm oil, and to complete removal of free acids during physical refining. Since dimerization and isomerization of fatty acid groups increase with time and temperature [56], a deodorization at 270 °C should not last more than 30 min.

Batch deodorizers are still widely used (Fig. 18). They are about twice as high as they are wide and are normally half-filled to provide sufficient head space. Their capacity can vary from 5 to 25 t. The bottom part is fitted with heating coils and a steam distributor through which stripping steam is injected into the oil. The vacuum system (0.6 – 1.2 kPa) comprises a booster and ejector assembly. Barometric condensers can be eliminated by using dry condensing plants. When deodorization is completed, the oil is cooled to ca. 90 °C in the deodorizer and to 40 – 60 °C in a subsequent vacuum cooler; cooling minimizes oxidation that occurs on contact with atmospheric oxygen at elevated temperatures. The total cycle of charging, heating, deodorization, cooling, and discharging requires up to 8 h, the deodorization process itself up to 4 h. The deodorizer distillate has the appearance of a milky emulsion and contains calcium soaps, nonsaponifiable matter, and a small fraction of neutral oil.

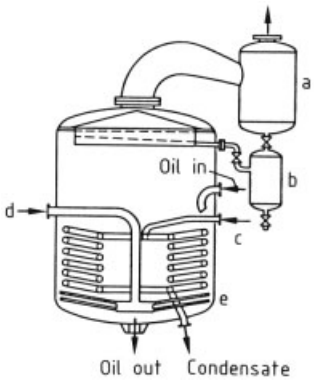

Figure 18. Batch deodorizer
a) Barometric condenser; b) Fat trap; c) Heating steam; d) Stripping steam; e) Steam distributor

Semicontinuous and continuous deodorization systems are replacing batch deodorizers because of the savings in steam resulting from more efficient stripping and heat recovery.

The semicontinuous Girdler deodorizer (Fig. 19) consists of an iron shell fitted with stainless steel trays. At defined time intervals (e.g., every 30 min), a certain quantity of oil flows into the top tray of the deodorizer where it is deaerated in vacuo and then heated to ca. 170 °C by exchange with the hot effluent oil. After 30 min the oil is discharged into the second tray where it is heated to the operating temperature (normally 230 – 250 °C). Injection of steam (at 0.6 – 1 kPa) takes place in the 3rd and 4th trays, and cooling in the 5th tray. Operation of the unit is made fully automatic by use of a timing device which opens and closes the oil valves. The semicontinuous Lurgi plant works on a similar principle (Fig. 20); the stripping steam is injected through special orifices and further distributed through the oil by steam-lift pumps. These types of deodorizers have capacities of up to 20 t/h.

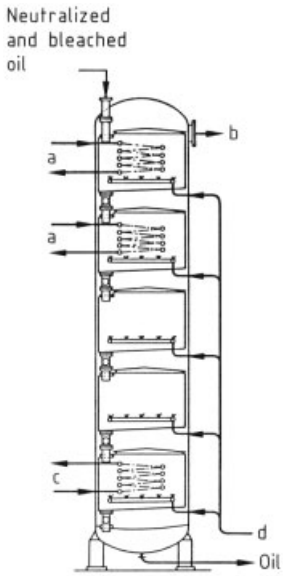

Figure 19. Semicontinuous deodorizer (Girdler)
a) Heating steam; b) Vacuum aggregate; c) Cooling water; d) Stripping steam

Fully continuous deodorizers generally comprise a high cylindrical column fitted with a number of trays with bubble caps or similar devices, down which the oil flows countercurrent to the stripping steam. The oldest continuous deodorizer is probably that built by Foster – Wheeler. Modern versions [57] are made by EMI, Gianazza, Mazzoni, Wurster & Sanger, Votator, de Smet, Ex-Technik, HLS, Lurgi, and Krupp (Fig. 21). An alternative construction is that of EBE in which the oil flows countercurrent to the stripping steam in horizontal tubes. The Campro deodorizer has a continuous tray-in-shell design that combines plug flow with a unique thin-film stripping concept.

7. Fractionation

The aim of fractionating fats is to remove either undesirable components or to isolate desired components with special properties. Edible oils can, for example, be winterized by removing waxes and saturated components; cocoa butter substitutes can be obtained from palm oil. Except for the molecular distillation of monoglycerides, distillative techniques cannot be employed. Industrial fractionation techniques are based on crystallization or liquid – liquid extraction. Separation of glycerides by selective adsorption is only used occasionally for purification purposes. Fractional crystallization can be subdivided into dry processes, processes employing solvents, and processes involving selective wetting of fat crystals. Directed interesterification combines selective fractionation with interesterification (see Section 4.2).

The oldest example of *dry fractionation* is the separation of tallow into stearin and olein by pressing; this separation came into use toward the end of the 19th century in connection with the production of margarine. The lower-melting glycerides (olein) were squeezed out of the crystal network of the higher-melting glycerides by pressing blocks of tallow wrapped in filter cloth.

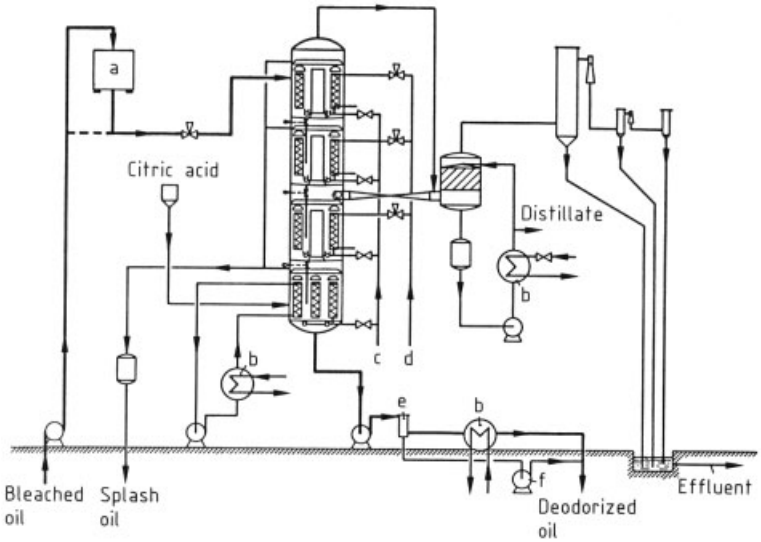

Figure 20. Flow diagram of a semicontinuous deodorization plant (Lurgi)
a) Measuring tank; b) Heat exchanger; c) Stripping steam; d) Heating steam; e) Filter; f) Pipe emptying

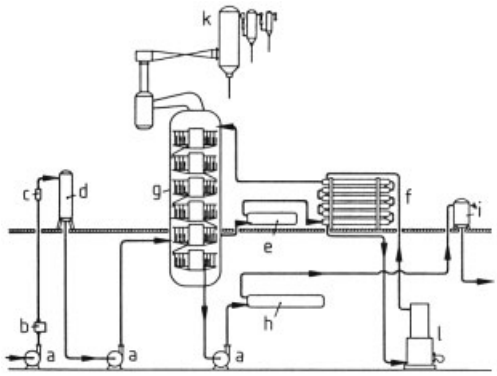

Figure 21. Continuous deodorizer (Krupp)
a) Pump; b) Filter; c) Meter; d) Deaerator; e) Preheater (low-pressure steam); f) Preheater (high-pressure steam); g) Kreuzstrom – Konti deodorizer; h) Cooler; i) Filter; k) Vacuum aggregate; l) Steam kettle

This method has been superseded by modern processes (Fig. 22) in which the fat is slowly chilled to the required temperature under controlled conditions in tempering vessels fitted with low-speed agitators, or in scraped coolers. Relatively large crystals of the higher-melting glycerides are obtained which can readily be filtered off in filter presses or drum filters. Slow cooling is important to obtain a relatively small number of nuclei which can then grow into large crystals. In addition, slow cooling promotes selective crystallization of the higher-melting glycerides and avoids the formation of mixed crystals containing both high- and low-melting glycerides. The difference in melting point must be at least 10 °C to achieve a sufficiently sharp, reproducible separation.

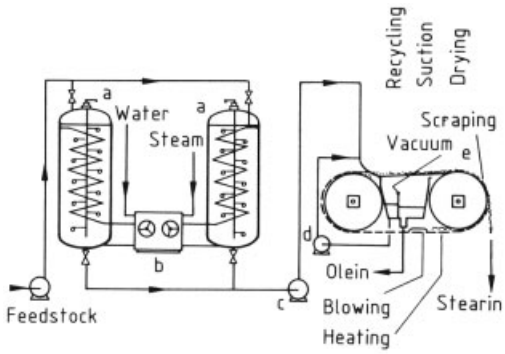

Figure 22. Dry fractionation plant
a) Crystallizers; b) Circulating unit; c) Feed pump; d) Recycling pump; e) Florentine filter

Winterization of edible oils such as sunflower or cottonseed oil involves removing waxes, which cause turbidity at refrigerator temperatures. After bleaching or deodorization, the oils are kept at ca. 10 °C for at least 1 h and then filtered. An alternative process is to combine dewaxing with superdegumming of the crude oil

(see also Section 6.1) [44]. Peanut oil is not normally winterized because its turbidity is caused by small glyceride crystals that are difficult to filter.

The separation of higher- and lower-melting glycerides can be improved by *fractionation using solvents* such as hexane, acetone, or methanol, in which the unsaturated fats are more readily soluble. Costs are much higher than for dry fractionation because of the need for solvent recovery. This relatively sharp fractionation method is generally only economical for preparing specialty products. For example, fractionation of palm oil with acetone gives a fraction (ca. 40 % of the feed) consisting of disaturated-monounsaturated glycerides, primarily 1,3-dipalmito-2-olein.

The separation of relatively small fat crystals can be facilitated by *selective wetting*, for example with sodium lauryl sulfate or decyl sulfate in aqueous solution. The crystal suspension, produced either batchwise or continuously, is mixed with an aqueous solution containing the wetting agent and normally also an electrolyte such as magnesium sulfate. The crystals covered with wetting agent migrate to the aqueous phase as a suspension that can be separated by centrifugation. Alfa – Laval recommends such a process (Lipofrac process) for the fractionation of palm oil and for the winterization of other oils.

Selective extraction [58] has long been used in refining mineral oil. Liquefied hydrocarbons or furfural are used as selective solvents in the separation of fats and fatty acids. This technique not only permits the separation of fats and fatty acids according to the degree of unsaturation or the molecular mass, but also effects the partial removal of coloring matter, gums, and vitamins.

Liquefied propane preferentially dissolves saturated components (Solexol process; see also page 33) [59]. This process has been used to fractionate tallow. The solubility of fats and fatty acids increases with decreasing molecular mass. Furfural preferentially dissolves unsaturated components. It is miscible with fatty acids and monoglycerides at room temperature, but with triglycerides only at elevated temperatures. Depending on the operating temperature, furfural is used alone or together with hexane. The selective solubility behavior is less pronounced for differences in molecular mass [60]. The furfural process has been used to produce high-performance drying oils from soybean and linseed oils.

8. Hydrogenation [61], [62]

Hydrogenation of the carbon – carbon double bonds of unsaturated glycerides leads to an increase in melting point (hardening, see also Section 4.3). The combination of hydrogenation, interesterification, and fractionation offers the opportunity of economically meeting the increasing demand for specialty products for foodstuff and technical applications from available raw materials.

Trans fatty acids produced by isomerization during hydrogenation also occur in animal fats such as butterfat and tallow. Although there are indications that trans fatty acids may be metabolized differently from their cis isomers [63], long-term studies have demonstrated their toxicological safety [64]. As far as their effect on blood lipid levels is concerned, they are comparable with saturated fatty acids.

The selectivity of hydrogenation is determined by the type and activity of the catalyst [65] (controlled poisoning and a large pore size increase selectivity), the operating temperature (both selectivity and cis – trans isomerization increase with increasing temperature), stirring – mixing efficacy (increasing the hydrogen concentration at the catalyst surface by intensive stirring – mixing reduces selectivity), and the hydrogen pressure (increasing the pressure reduces the selectivity).

Neither absolute selectivity nor complete isomerization (or complete suppression of isomerization) can be achieved. However, the process can be steered in a desired direction by suitable choice of the above parameters to produce fats with different degrees of saturation and melting properties from a given oil. For example, peanut oil can be hardened nonselectively to an iodine value of 70 – 90 yielding a soft, semiliquid fat. If the same oil is hardened to the same iodine value under conditions that promote isomerization, a firm, plastic fat with a melting point of ca. 33 °C is obtained. Palm oil is sometimes hydro-bleached, i.e., hydrogenated under very mild conditions that lead to selective hydrogenation of carotenoids.

The oil should be at least degummed and bleached prior to hydrogenation and preferably neutralized and bleached as well. Catalyst poisons include phosphatides, sulfur compounds, soaps, and oxidation products from poor handling and/or bleaching. Free fatty acids do not poison the catalyst but tend to slow down the hydrogenation reaction.

The hydrogen must be pure and dry and can be produced by various routes. The most important industrial process is decomposition of natural gas or propane by passage over a nickel catalyst in the presence of steam at 800 – 900 °C. Electrolysis of water is used to a lesser extent. The hydrogen must be specially purified to remove hydrogen sulfide and carbon monoxide, which are strong catalyst poisons.

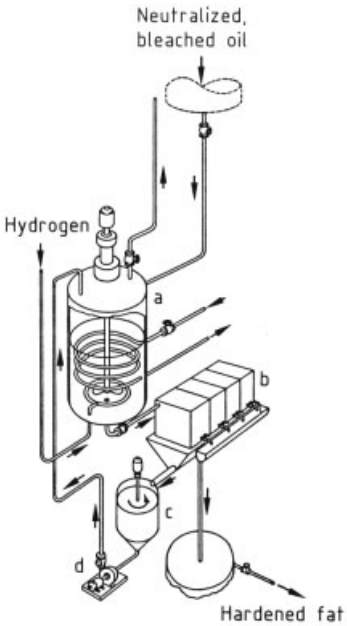

Figure 23. Flow diagram of a batch hardening plant
a) Converter; b) Filter press; c) Catalyst tank; d) Catalyst pump

Nickel, platinum, and palladium are preferred catalysts. Platinum and palladium are especially suited for laboratory-scale and low-temperature hydrogenations. Only nickel is used on a production scale. Many different types of nickel catalysts are available (→ Heterogeneous Catalysis and Solid Catalysts). They essentially consist of metallic nickel on a carrier such as silica gel, silicic acid, or alumina. The carrier is treated with a nickel salt such as the carbonate, hydroxide, or sulfate, which is converted into nickel oxide by roasting. The oxide is then reduced to metallic nickel in a stream of hydrogen at 300 – 400 °C. Other routes involve decomposition of nickel formate at 240 °C or dissolution of aluminum from a nickel – aluminum alloy with alkali (Raney catalyst). In addition, there are a number of mixed catalyst systems (copper, chromium, cobalt) with special selectivities.

Batch hydrogenation (Fig. 23) is normally performed on a scale of 5 – 25 t. The dry oil is pumped into the hydrogenation vessel that has been previously flushed with hydrogen, and heated to just below the required operating temperature (100 – 180 °C). The catalyst (0.01 – 0.1 % active nickel) is then introduced and the hydrogen pressure increased. Alternatively, the oil is vacuum-deaerated while it is heated before the catalyst and hydrogen are introduced. Hydrogenation is an exothermic reaction; a decrease in the iodine value of one unit causes a temperature increase of 1.5 – 2.0 °C. The required operating temperature is therefore maintained by controlled cooling. Hydrogen is introduced through a sparge ring at the bottom of the hydrogenation vessel, pumped out at the top, and recirculated. Sparging and fast stirring ensure adequate dispersion of the catalyst in the oil and optimum mass transport. The usual working pressure is 0.15 – 0.3 MPa (1.5 – 3 bar).

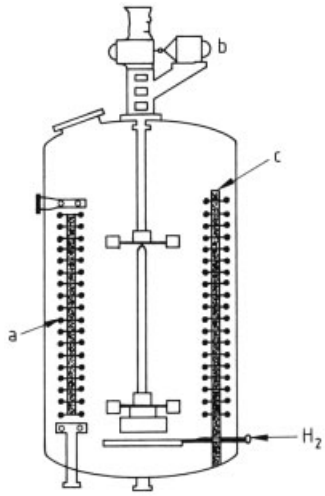

Figure 24. Hydrogenation converter (dead-end design)
a) Heating and cooling coil; b) Turbine agitator; c) Baffle

In hydrogenation autoclaves that operate without hydrogen recirculation ("dead-end" process), spent hydrogen is replaced at a somewhat higher pressure [66]. The intensive stirring required can be achieved with a turbine agitator or similar device (Fig. 24).

So-called loop hydrogenation reactors (e.g., Buss, Fig. 25) have become increasingly popular. The oil – catalyst mixture is constantly circulated through an external heat exchanger during heating and cooling. This loop reduces total cycle time; it also gives better temperature control and thus less product variation from batch to batch.

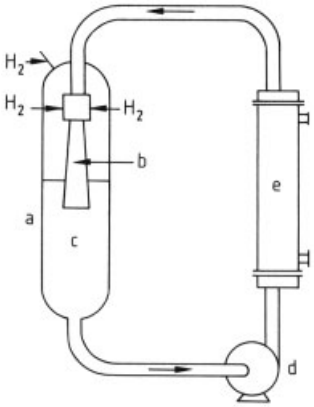

Figure 25. Loop hydrogenation reactor
a) Autoclave; b) Mixing and reaction zone; c) Reactant – solvent – catalyst suspension; d) Primary loop recirculation pump; e) Primary loop heat exchanger

When hydrogenation is complete, the product is cooled to 80 – 90 °C and the catalyst is filtered off. Since the acid value tends to increase slightly during hydrogenation, the crude hardened fat is usually postneutralized with lye and bleached before it is deodorized. Posttreatment with citric or phosphoric acid, followed by bleaching prior to deodorization, is an alternative to lye posttreatment. In either case the residual nickel content is reduced to <0.2 ppm. Deodorization of a hardened fat is essential since hydrogenation leads to typical hardening flavors which must be removed.

Continuous hydrogenation is primarily suitable for the production of one particular product over a longer period of time. Fully continuous processes are the exception rather than the rule. In semicontinuous processes several reaction vessels or autoclaves are arranged sequentially (e.g., Alfa – Laval system). In fully continuous plants, either the preheated oil and hydrogen are passed over a stationary catalyst bed, or all three components are mixed prior to entering the reactor (see Fig. 26). In the Procter & Gamble process, the reactor contains a special stirrer to ensure intensive mixing in a thin film. The Girdler Corporation developed a process in which three Votator heat exchangers are arranged sequentially; the first vessel serves as preheater, the second as a hydrogenator, and the third as a cooler. Continuous hydrogenation of miscella at low temperatures over a catalyst bed produces little isomerization but has not achieved industrial importance.

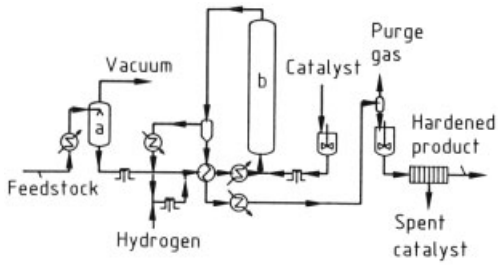

Figure 26. Continuous hydrogenation of fatty acids and fatty oils (Lurgi)
a) Dryer; b) Reactor

The course of hydrogenation is normally monitored by changes in analytical data such as iodine value or refractive index. The hydrogenated products are characterized primarily by melting point and solids content at different temperatures.

9. Interesterification

Prior to interesterification, fats must be neutralized to less than 0.1 % ffa and dried to avoid excessive deactivation of the catalyst, which is normally sodium ethoxide (see also Section 4.2). The catalyst (0.1 – 0.3 %) is finely dispersed in the fat. After completion of the interesterification reaction, which is carried out at 80 – 100 °C, excess catalyst is deactivated by addition of water. The resulting soap is removed by washing with water and bleaching. During interesterification, sodium ethoxide is converted into fatty acid ethyl esters, which are removed during the final deodorization step.

Interesterification can be performed batchwise (Fig. 27), usually in a neutralization–bleaching vessel, or continuously. In the continuous process, fat with suspended catalyst passes through a tubular reactor (Fig. 28) [67]. In directed interesterification (see Section 4.2) the fat is cooled, for example, in a scraped-surface cooler to crystallize the higher-melting glycerides.

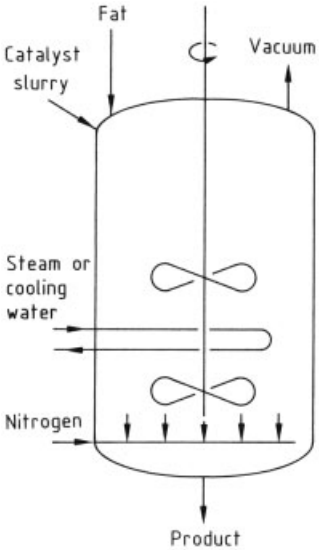

Figure 27. Batch reactor for interesterification

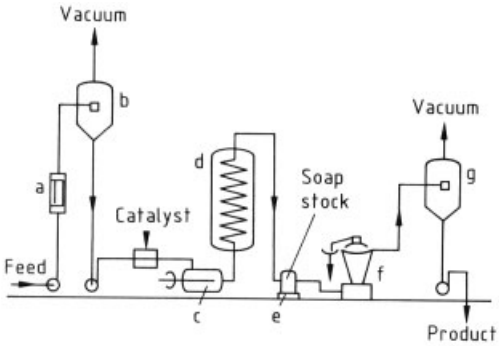

Figure 28. Continuous interesterification
a) Heat exchanger; b) Dryer; c) Homogenizer; d) Tubular reactor; e) Mixer; f) Centrifuge; g) Dryer

New types of fats can be created by suitable choice of interesterification components. Depending on the starting components, the melting point of the fat after interesterification can be lower or higher. Figure 29 shows the solids content (dilatation) curves of native and interesterified palm oil; the glyceride composition is given in Table 22.

Acidolysis. An example of the industrial application of acidolysis is the transesterification of coconut oil with acetic acid and the subsequent esterification of excess acid with glycerol. A mixture of aceto fats (laurodiacetin, myristodiacetin, etc.) is obtained which are used as plasticizers and coatings. Transesterification of coconut oil with higher fatty acids, employing continuous distillation of the liberated lower fatty acids ($C_6 - C_{10}$), yields fats with properties similar to those of cocoa butter. Acidolysis is usually performed in the presence of an acid or base catalyst; no catalyst is required at temperatures of 260 – 300 °C.

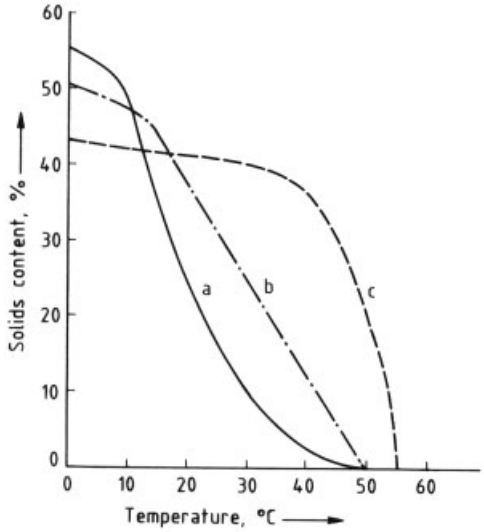

Figure 29. Solids content of native and interesterified palm oil (Alfa – Laval)
a) Native palm oil; b) Randomly interesterified palm oil; c) Directedly interesterified palm oil

Alcoholysis. The replacement of the glycerol moiety of a fat with lower monohydric alcohols proceeds at lower temperatures. The Bradshaw process [68] utilizes transesterification of fats with methanol as the first step in the continuous production of soap. Fat with an acid value of less than 1.5 is stirred with excess methanol for a few minutes in the presence of 0.1 – 0.5 % caustic soda at ca. 80 °C. On subsequent standing, practically dry glycerol settles at the bottom of the reaction vessel. The same principle is em-

Table 22. Fatty acid and glyceride composition of native and interesterified palm oil *

	Native palm oil	Interesterified palm oil	
		Random interesterification	Directed interesterification
mp, °C	42	47	52
Fatty acid composition, mol %			
S	51	51	51
U	49	49	49
Glyceride composition, mol %			
S_3	7	13	32
S_2U	49	38	13
SU_2	38	37	31
U_3	6	12	24

* S = saturated fatty acid. U = unsaturated fatty acid.

ployed in the preparation of fatty acid methyl esters for laboratory analyses. If the acid value is higher than 1.5, boron trifluoride is normally used as a catalyst.

The exchange of glycerol with other polyhydric alcohols, such as pentaerythritol, is best done under conditions at which glycerol can be distilled off, i.e., at 200 – 250 °C and 5 – 40 kPa (50 – 400 mbar), in the presence of an alkaline catalyst.

The production of mono- and diglycerides by reacting fats with an excess of glycerol is a special case of transesterification of fatty acid esters with polyhydric alcohols. These glycerides are especially important as emulsifiers. The production process consists of heating a fat with 25 – 40 % glycerol at 205 – 245 °C in the presence of sodium hydroxide or a sodium alkoxide. Part of the catalytic effect is probably based on soap enhancing the solubility of glycerol in fat. Industrial products contain 40 – 50 % monoglycerides. Oxidation and undesirable color effects can be minimized by working in vacuo or under an inert gas, and by using stainless steel reaction vessels. Monoglyceride concentrates can be produced by molecular or high-vacuum distillation.

10. Environmental Aspects

Crude oils and fats can be contaminated with traces of various substances such as metals (e.g., lead, arsenic, cadmium, selenium, mercury), chlorinated organic pesticides (DDT, dieldrin), mycotoxins (e.g., aflatoxins), and sulfur compounds. These trace contaminants are removed in the course of refining [55].

Wastewater from lye neutralization and from the barometric condensers of deodorizers can be discharged into communal sewage systems or surface waters only if certain limits for residual fat, pH, sediment, sulfate, and chemical and biological oxygen demand are met. Otherwise biological pretreatment is required. Since the cost of biological or other pretreatment of wastewater can be considerable, there has been a trend toward physical refining (i.e., refining without lye neutralization) by employing special degumming techniques, bleaching, and distillative deodorization.

Bleaching earth is normally extracted before being deposited or burnt. The extracted fats can be recycled.

Deodorizer distillates [55] can be used in animal feeds, as a source of vitamin E, for technical applications, or simply burnt.

11. Standards and Quality Control

Standardized methods for quality control have been issued by the following organizations:

Association Française de Normalisation (AFNOR), American Oil Chemists' Society (AOCS), Association of Official Agricultural Chemists (AOAC), American Society for Testing and Materials (ASTM), British Standards Institution (BSI), Deutsche Gesellschaft für Fettwissenschaft (DGF), Deutsches Institut für Normung (DIN), International Association of Seed Crushers (IASC), International Union of Pure and Applied Chemistry (IUPAC), International Standards Organization (ISO), the Netherlands Normalisatie Instituut (NNI), and the Fed-

eration of Oils, Seeds, and Fats Associations (FOSFA).

There is an increasing trend for the various national and international methods to be uniformly standardized as ISO procedures; these are generally adopted as official methods within the European Community.

The IASC (8 Salisbury Square, London EC 4P 4 AN, UK) has issued a list of associations dealing with or issuing standard forms of contract and/or quality standards for oilseeds, oils, and fats. This association has also drawn up a list of official institutions and associations that deal with or issue methods for grading, sampling, and preparing samples for analysis, as well as methods for analysis that are referred to in standard forms of contract or usually applied in the trade of oilseeds, oils, and fats.

11.1. Sampling

Correct sampling is a difficult process requiring careful attention; it must take into account that separation can occur during storage of fats and fatty products. Hulls, for example, tend to separate from oilseeds because of their lower density. Oils and fats tend to form a sediment of water and dirt; for this reason a representative average sample is often taken in bypass during pumping.

A special zone sampler consisting of a metal cylinder with a valve at both ends is often used for sampling the contents of a tank. The sampler is lowered into the tank, the valves being operated with a cord. Starting from the bottom of the tank, the oil is sampled at various depths up to the surface. Smaller containers (drums) are sampled with a tube or shutter scoop; solid fats in retail packs are sampled with a sampling scoop.

Special sample dividers are used for oilseeds. Sampling apparatus for bags includes sack-type spears or triers, cylindrical and conical samplers, and hand scoops. Shovels, scoops, cylindrical and conical samplers, or mechanical samplers are used for drawing small periodical samples from a flow of oilseeds. Shovels, quartering irons, or riffles are used for mixing and dividing.

Samples for analysis of oilseeds, expeller cakes, and meals should preferably be at least 2 kg, and of oils and fats at least 250 g. The samples should be homogenized immediately prior to analysis. Larger oilseeds are comminuted prior to analysis.

11.2. Raw Materials

Buying – selling contracts for oilseeds, crude oils, and fats normally specify a maximum permissible concentration of impurities. If this level is exceeded the buyer receives a discount. The key parameters in oilseeds are fat and moisture contents.

To determine the *fat content* of oilseeds, expeller cakes, and meals, the comminuted sample is extracted with a solvent and the fat taken up by the solvent is determined refractometrically, gravimetrically, densitometrically, or by the change in dielectric constant. The gravimetric methods are time-consuming but most reliable. Each method gives reproducible but different values, depending on the type of solvent (usually hexane or petroleum ether) and the extraction method. The fat content can also be determined directly without solvent extraction by nuclear magnetic resonance or by near infrared absorption.

According to ISO 659, the oilseeds are extracted with *n*-hexane or petroleum ether in a Bolton extractor. After extraction for several hours, the sample is removed from the extractor thimble, reground, and reextracted. This is repeated a third time. The extract is evaporated and the residual oil dried to constant weight at $103 \pm 2\,^{\circ}C$.

Unilever has developed a far quicker method which gives identical values to those obtained with the ISO method. The sample is first comminuted in a ballmill in the presence of petroleum ether, and then extracted for 2 h in a Bolton extractor.

When determining the fat content of fruit pulps and animal tissues, predrying of the substrate at 105 °C in vacuo is recommended. When extracting animal tissue polar solvents such as ether – HCl or ethanol – chloroform are used to facilitate liberation of fat from lipoprotein complexes present in aqueous dispersion.

Water and *volatile matter* are generally determined by weighing a sample before and after heating under standardized conditions, usually 103 °C. Crude oils are normally analyzed for water and volatile matter, color, particulate dirt,

phosphatides, and free fatty acids. Up to 0.5 % moisture and impurities are generally regarded as technically unavoidable. Inferior oils usually have a dark color and a high content of free fatty acids and oxidized components that is indicated by a relatively high absorption at 232 nm, high anisidine and totox numbers, and a high concentration of oxidized fatty acids. These analytical data give an approximate indication of the expected yield and quality of the fully refined product. Admixture can be detected within certain limits by determining the composition of fatty acids, sterols, and tocopherols. Very severe admixture can be detected by iodine value, refractive index, and saponification value.

11.3. Oils and Fats

11.3.1. Physical Methods

The *density* or *specific gravity* at a given temperature is determined by weighing a known volume, by determining the buoyancy, or by electromagnetic excitation of the natural frequency of the sample.

The *refractive index* is normally determined with an Abbe refractometer at 20 °C for oils which are liquid at this temperature and at 40, 60, or 80 °C for fats. The temperature coefficient is ca. 0.0036 K^{-1}. A correlation exists between refractive index and iodine value (see Section 3.8); for this reason the refractive index is often used to follow the course of hydrogenation.

Most oils and fats rotate *polarized light* by only a few tenths of a degree; higher values are observed with substituted fatty acids such a ricinoleic acid.

In melting a distinction is made between the slip point, the point of incipient fusion, and the point of complete fusion. The *slip point*, the usually cited melting point, is the temperature at which a sample plug placed in an open glass capillary begins to rise when gradually heated in a water bath or by other means. This method can be automated. The slip point value is influenced by the method of cooling used prior to melting. Whereas for conventional fats the sample is normally solidified by chilling the sample and capillary at − 10 °C for a few minutes, fats that exhibit distinct polymorphism, e.g., cocoa butter and substitutes, are conditioned at higher temperatures for up to 18 h.

The *point of incipient fusion* is the temperature at which a sample placed in a U-tube begins to flow on being warmed. The *point of complete fusion* is the temperature at which a sample becomes completely clear.

Flow and *drop points* are primarily of importance for technical greases and some plastic fats. They are determined in an Ubbelohde apparatus. Flow point is the temperature at which a drop begins to be clearly formed at the end of a thermometer, and drop point is the temperature at which the drop falls.

Dilatation, i.e., the increase in volume of a fat on passing from the solid to the liquid state at a given temperature, can be used to determine the concentration of solid components at a certain temperature (Solid Fat Index, SFI). This type of measurement has been replaced by the much quicker *nuclear magnetic resonance* method which is based on the difference between the response from the hydrogen nuclei in the solid phase and that from all the hydrogen nuclei in the sample.

Differential thermal analysis permits the quantification of changes in the solid phase of a fat on being heated. A sample is placed in a metal block which is heated at a constant rate. The difference between the temperature of the block and that of the sample is characteristic of crystal modifications occurring during melting [69].

On passing from the liquid to the solid state, the temperature remains constant for a certain period of time because of the liberated heat of fusion; with sufficient insulation an increase in temperature can be observed. The *titer value* of a fat is the maximum value to which the temperature rises in the cooling curve. Cooling curves are frequently used to check the purity of fats which consist predominantly of a few triglyceride types, e.g., cocoa butter.

The *cloud point* is the temperature at which an oil or molten fat begins to become turbid when cooled under controlled conditions. The temperature stability of an oil is normally assessed after keeping a carefully dried sample at 0 °C for 5 – 8 h.

The smoke, flash, and fire points of oils and fats are measures of their thermal stability when heated in air. Table 13 shows some typical data.

The *smoke point* is the temperature at which smoke is first detected in a standard apparatus with specially designed, draft-free illumination. The *flash point* is the temperature at which volatile products are evolved at such a rate that they can be ignited but do not continue to burn. The *fire point* is the temperature at which the production of volatile products is sufficient to support continuous combustion after ignition. The smoke point is important in the assessment of refined oils and fats used for deep frying. It is governed primarily by the free fatty acid content. The flash point is a measure of residual solvent in crude oils, although gas – liquid chromatographic methods are now generally preferred for this determination.

The *viscosity* of an oil can be determined in a flow (Engler), capillary (Ubbelohde), fall (Höppler), or shear (Haake) viscometer.

There are a number of different methods for determining the *color* of an oil or fat. In the Lovibond tintometer the color of the sample in a standardized cell with a length of 1″ (2.45 cm) or 5 1/4″ (13.3 cm) is matched – either manually or automatically – with the color of standard yellow, red, and blue color slides. Other colorimetric scales are based on standard iodine or potassium dichromate solutions. The most objective way of assessing color is to determine the absorption at specified wavelengths in the range 370 – 780 nm.

The characteristic absorption maxima of specific lipid groups in the infrared and ultraviolet ranges can be used qualitatively and quantitatively (see Section 3.8). Fatty acids with isolated double bonds such as linoleic, linolenic, and arachidonic acids do not exhibit specific UV absorption. After isomerization with alkali they can, however, be determined by the characteristic absorption maxima of the corresponding conjugated fatty acids. Infrared spectroscopy is used primarily for the identification of specific groups such as hydroxyl and trans double bonds.

Optical rotatory dispersion spectroscopy is used only in studies of biosynthesis and metabolism of lipids.

Mass spectroscopy, especially in combination with gas chromatography and other methods of separation, is often used to determine the structure of oxidation products and substituted fatty acids.

11.3.2. Chemical Methods

For preliminary identification and evaluation, determination of chemical data such as the iodine value (see below) and the fatty acid composition often suffice. In many cases, however, it is necessary to test for characteristic minor constituents directly or by specific chemical reactions. Table 23 gives some examples.

The complexity of oils and fats makes the determination of their precise composition difficult. For this reason analytical values are used to express certain properties. Analytical values in this sense are defined as the equivalent amounts of certain reagents which react with specific groups in the lipid molecules. Many of the older analytical values are no longer used because of the introduction of modern analytical techniques such as column, paper, gas – liquid, and high – performance liquid chromatography, and spectrophotometry. However, the principle of determining analytical values is still valid since it can authenticate and assess the quality of an oil or fat relatively quickly with modest laboratory facilities.

The *saponification value* is the number of milligrams of potassium hydroxide required to saponify (hydrolyze) 1 g of fat and is related to the molecular mass of the fat.

The *hydroxyl value* is expressed as the number of milligrams of potassium hydroxide required to neutralize the acetic acid needed to acetylate 1 g of fat.

The *carbonyl value* is defined as the number of milligrams of CO (carbonyl groups) per gram of fat. The basis for this value is the conversion of the carbonyl group into an oxime with hydroxylamine in alcoholic KOH:

$$RCO + NH_2OH \longrightarrow RC{=}NOH + H_2O$$

Hydroperoxides interfere with this determination.

The acidity of an oil or fat arises from free inorganic or organic acids. The *acid value* is defined as the number of milligrams of potassium hydroxide required to neutralize 1 g of fat. The content of free fatty acids (ffa) expresses how many parts of free fatty acids are contained in 100 parts of fat. This content is – depending on the type of oil or fat – expressed as % oleic acid

Table 23. Detection of oils and fats

Oil/Fat	Characteristic constituent	Specific reaction or detection procedure	Limit of detection	Remarks
Animal fats	cholesterol	isolation of sterols from the unsaponifiable matter by TLC or HPLC, and separation by HPLC, TLC, or GLC	1 % in vegetable oils and fats	Vegetable oils and fats contain traces of cholesterol in addition to phytosterols while in animal fats only cholesterol but no phytosterols have been found.
	branched $C_{13}-C_{17}$ fatty acids and odd-numbered, unbranched $C_{11}-C_{19}$ fatty acids	GLC of fatty acid methyl esters	ca. 2 % in vegetable oils and fats	The branched fatty acids may have to be concentrated, e.g., by urea adduct formation.
Lard	palmitic acid in the glyceride 2-position	isolation of the glyceride fraction with 1 saturated and 2 unsaturated fatty acid groups, followed by determination of the fatty acid composition in the glyceride 2-position	1 % lard in tallow	Determination of the fatty acid composition in the 2-position without prior glyceride fractionation gives a significantly higher limit of detection.
Partially hardened and unhardened marine oils	highly unsaturated $C_{20}-C_{24}$ fatty acids	Tortelli–Jaffe reaction (green coloration with bromine in chloroform/acetic acid) GLC of the fatty acid methyl esters	ca. 2 % in other oils and fats	Not strictly specific; highly unsaturated vegetable oils can also react.
Vegetable oils and fats	phytosterols	isolation of sterol esters and separation as sterols by TLC, HPLC, and GLC	ca. 2 % in animal fats	Sterols occur in vegetable oils and fats in the free and esterified form, while in animal fats they are only present in the free form.
Castor oil	ricinoleic acid	TLC; IR spectroscopy	ca. 1 % in other oils and fats	Hydroxy fatty acids can be formed during autoxidation of oils and fats.
Coconut oil and palm kernel oil	lauric acid	GLC of the fatty acid methyl esters	ca. 2 % in other vegetable oils except oils and fats rich in lauric acid, e.g., babassu fat	
Cottonseed oil and other oils of the malvaceae, tiliaceae, and bombaceae families	malvalic acid	Halphen reaction (red coloration on heating with sulfur in CS_2)	depending on extent of refining	Hardened cottonseed oil does not give this reaction; malvalic and sterculic acid are converted into saturated, branched fatty acids during hardening.

Table 23. Continued

Oil/Fat	Characteristic constituent	Specific reaction or detection procedure	Limit of detection	Remarks
Crude palm oil	α- and β-carotene	isolation from nonsaponifiable matter by HPLC		Indicative of distillative or other heat treatment.
Peanut oil	lignoceric and arachidic acid	GLC of the fatty acid methyl esters	4	Hardened marine oils and certain vegetable oils also contain higher molecular mass fatty acids.
Olive oil	α-tocopherol	isolation and separation by HPLC	ca. 5 % other oils in olive oil	
Partially hardened fats	trans fatty acids	IR spectroscopy	ca. 2 % in vegetable oils	Animal fats contain up to 10 % trans fatty acids.
Rapeseed oil	erucic acid	GLC of the fatty acid methyl esters	ca. 1 % of "high erucic acid" rapeseed oil in other vegetable oils	Admixture with "low erucic acid" rapeseed oil cannot be detected satisfactorily this way. Marine oils also contain erucic acid.
	brassicasterol	isolation and separation of sterols by TLC, HPLC, and GLC	ca. 5 % in other vegetable oils	Independent of erucic acid content.
Sesame oil	sesamol and sesamin	Baudouin reaction (red coloration with furfural and HCl)	ca. 0.5 % in other oils and fats	
Soybean oil	linolenic acid	GLC of the fatty acid methyl esters	ca. 10 % in e.g., sunflower oil	
	δ-tocopherol	isolation and separation by HPLC	ca. 5 % in e.g., sunflower oil	
Teaseed oil		Fitelson test (modified Liebermann – Burchard test)	e.g., 5 % in olive oil	
Tung oil	eleostearic acid	GLC of the fatty acid methyl esters	ca. 1 % in other oils and fats	

(M_r 282), % palmitic acid (M_r 256), or % lauric acid (M_r 200). If not specified, it is calculated as oleic acid.

Peroxides are determined iodometrically according to Wheeler (cold) or Sully (hot) and are expressed as milliequivalents of active oxygen per 1000 g of fat. The peroxide value can also be expressed in millimoles per kilogram (Lea value, 1 mequiv/kg = 0.5 mmol/kg) or in milligrams per kilogram (1 mequiv/kg = 8 mg/kg). It is important to use fresh reagents and to exclude oxygen as far as possible during the determination.

The reaction with anisidine is often used to determine the carbonyl compounds produced during autoxidation. The *anisidine value* is defined as 100 times the absorption of a 1 % sample solution in a 1-cm cell after reaction with anisidine. This value has been recommended for monitoring the refining of oils and fats. It is also used in combination with the peroxide value to characterize the overall autoxidative state of an oil or fat (*totox value*):

$$\text{Totox value} = 2 \times \text{peroxide value} + 1 \times \text{anisidine value}$$

The *iodine value* denotes the percentage by weight of iodine bound by 100 g of fat and is an approximate measure of the degree of unsaturation. The usual method is that of Wijs. The rate of addition of iodine to double bonds depends on their distance from the carboxyl group. Conjugated unsaturated fatty acids are only partially halogenated; substituted fatty acids such as keto acids and cyclic acids can also react. Trans fatty acids react more slowly than the cis isomers.

The *oxidative stability* of oils and fats is determined by their composition and their previous history. A number of empirical methods have been developed to express oxidative stability [70]. These tests are performed at elevated temperature and do not necessarily correlate with organoleptic stability at ambient temperature. The most frequently practiced methods are the Schaal test and the Swift test (or Active Oxygen Method); automated versions of the latter are popular, e.g., the Rancimat apparatus.

The *moisture content* can be determined by weighing before and after drying or by codistillation with toluene. It is most conveniently determined by the Karl Fischer titration, which is based on the following reaction:

$$2\,H_2O + SO_2 + I_2 \longrightarrow H_2SO_4 + 2\,HI$$

Impurities such as dirt and seed particles are not dissolved on treating the sample with petroleum ether or hexane and can be collected and weighed as filter residue.

Traces of *residual solvent* are best determined by gas chromatography, either by direct injection or after enrichment by the so-called headspace technique [71].

Nonsaponifiable matter is defined as that material which is a soluble constituent of the fat and which remains insoluble in the aqueous phase after saponification of the fat and can be extracted with petroleum ether or diethyl ether; diethyl ether tends to give higher values. Nonsaponifiable matter consists primarily of sterols, alcohols, hydrocarbons, and vitamins.

Oxidized fatty acids contain one or more hydroperoxide, hydroxyl, or keto groups in the fatty acid chain and can be determined by their reduced solubility in petroleum ether. Only part of the oxidized fatty acids are detected by this type of analysis. Thus, the separation of fats into polar and nonpolar fractions on silica gel tends to be preferred when, for example, characterizing used frying fats [72].

To determine *resin acids* in the presence of fatty acids, the sample is esterified with methanol in the presence of an acid catalyst; resin acids escape esterification.

Inorganic acids can be detected by extracting the sample with hot water and testing the aqueous extract with methyl orange.

Free alkali can similarly be detected by dissolving the sample in ether – ethanol and adding phenolphthalein (red coloration). Traces of soap can be detected down to 10 ppm by dissolving the sample in ether – ethanol, adding bromophenol blue (which gives a green-blue color in the presence of soap), and titrating with very dilute HCl.

Many crude seed oils contain considerable concentrations of *phosphatides*. These can be determined by dry digestion of the sample with magnesium oxide and photometric determination of the resulting phosphate after addition of molybdate [73]. An alternative is the direct determination of phosphatides by atomic absorption or plasma emission spectroscopy.

Methods for determining traces of *mineral oil* in edible oils are based on the fluorescence of mineral oils, their insolubility in acetic anhydride, and the fact that they are not retained on alumina. Headspace gas chromatography is also used.

The gas chromatographic determination of the *fatty acid composition* of an oil or fat has replaced the time-consuming, chemical techniques of fractionating fatty acids into single components or groups. Analytical values such as the Reichert–Meissl, the Polenske, the Kirchner, and the butyric acid number (as measures of the concentration of lower molecular mass or steamvolatile fatty acids) are now virtually obsolete. Prior to gas chromatographic analysis, the fatty acids are converted into their methyl esters either by direct transesterification of the sample with methanolic KOH (acid values <2) or by saponification (acid values >2) followed by esterification of the fatty acids with methanol in the presence of BF_3. Full details of these procedures and of the various types of columns, column packings, and detectors are given in ISO 5508/5509. If a marker fatty acid is used as an internal standard, quantitative determination of the total amount of monomeric, nonoxidized fatty acids is possible; nonsaponifiable matter, oxidized, and polymerized fatty acids are retained on the gas chromatographic column.

Polymeric fatty acids are conveniently determined by special high-performance liquid chromatography (HPLC) or gel permeation chromatography [74].

The *glyceride structure* may have to be determined to authenticate an oil or fat. The methods for determining positional isomers are based on fractional crystallization, selective oxidation, chromatographic separation, or enzymatic hydrolysis of the fatty acid in the 2-position; they are normally too expensive for routine control unless specialty products are involved.

Mono- and diglycerides are formed by hydrolysis of fats and therefore occur in higher concentrations in fats from damaged seeds. Mono-, di-, and triglycerides can be separated by column or thin-layer chromatography. 1-Monoglycerides can be determined directly in fats by oxidative fusion with periodic acid:

$$RCOOCH_2CHOHCH_2OH + H_5IO_6 \longrightarrow$$
$$HCHO + RCOOCH_2CHO + HIO_3 + 3\,H_2O$$

Monoglycerides can be isomerized into the 1-isomers prior to addition of periodic acid by treating the sample with perchloric acid.

12. Storage and Transportation [75]

Crude Oils and Fats. If crude oils and fats are to be stored for a long time, removal of insoluble impurities by sedimentation, filtration, or centrifugation is advisable. Seed particles and cell fragments contain fat-splitting enzymes, which can gradually hydrolyze glycerides in the presence of moisture and produce an increase in free fatty acid concentration. Proteins are good nutrients for microorganisms.

Further measures for avoiding undue quality deterioration include drying the crude oil to less than 0.2 % moisture and storage without heating; consistent fats should be stored at not more than 10–15 °C above the melting point. Frequent pumping should be avoided to prevent saturation of the oil with air and consequent autoxidation. The addition of antioxidants is useful primarily for animal fats; most vegetable oils contain sufficient concentrations of natural antioxidants. Most crude vegetable oils can be stored for some months without significant loss in quality if the above precautionary measures are observed. Animal fats can be stored for only relatively short periods of time without loss in quality. Continual monitoring of the acid value, peroxide value, and other data that characterize the state of oxidation during storage is recommended.

Raffinates. Since natural antioxidants, such as tocopherols, phenols, and synergists, are partially removed during the refining process, semi- and fully refined oils often have lower oxidation stability than the crude oils. This is especially true for oils with a relatively high content of polyunsaturated fatty acids. Therefore, raffinates must be protected as far as possible from light, air, moisture, relatively high temperatures, and trace metal prooxidants if they are to be stored for long periods.

Raffinates are best stored in closed containers with a relatively small ratio of surface area to volume in order to minimize access and infusion of air. Any heating should be uniform without local overheating; the temperature should not be

higher than 10 °C above the melting point. Aeration can be further reduced by using a bottom-loading pipe and a nitrogen atmosphere during both storage and transportation. Immediately after deodorization, nitrogen should be introduced into the suction side of the loading pump that feeds the storage or transport vessels.

The storage containers are preferably made of aluminum or stainless steel, although mild steel can be used if access of air is limited and if cooling is adequate. Raffinates can be kept stored in large tanks for several weeks without an increase in peroxide value; however, the peroxide value tends to increase each time the oil is pumped. Raffinates are transported in bulk in ships and trucks, in drums, tins, and bottles; consistent fats can also be transported in block or powder form if the melting point is sufficiently high.

13. Individual Vegetable Oils and Fats

Vegetable oils and fats are normally divided into fruit pulp fats and seed oils. Although there are few fruit pulp fats (e.g., palm oil and olive oil), the number of seed oils is considerable. Seed oils can be conveniently subdivided according to their characteristic fatty acids:

1) Lauric acid oils (e.g., coconut and palm kernel oil)
2) Palmitic – stearic acid oils (e.g., cocoa butter)
3) Palmitic acid oils (e.g., cottonseed and corn oil)
4) Oleic – linoleic acid oils (e.g., sunflower and linseed oil)
5) Leguminous oils (e.g., soybean and peanut oil)
6) Cruciferous oils (e.g., rapeseed oil)
7) Conjugated oils (e.g., tung oil and oiticica oil)
8) Substituted fatty acid oils (e.g., castor oil)

13.1. Fruit Pulp Fats

Palm oil and olive oil are the most important products within this group.

13.1.1. Palm Oil [76]

Palm oil [*8002-75-3*] is obtained from the fruit pulp of the oil palm (*Elaeis guineensis*). The fruit bunches, weighing 10 – 25 kg, contain 1000 – 2000 plum-sized fruits which are dark red because of their carotene content. The fruits contain 35 – 60 % oil, depending on the moisture content. The oil palm is native to West Africa, and is extensively cultivated in Malaysia, Indonesia, and Central Africa. The oil yield has been increased to 4.5 t of oil per hectare and year by the introduction of insect pollination and by selecting and cloning high-yielding cultivars. Malaysian palm oil production was 4.13×10^6 t in 1985 and is expected to reach 6×10^6 t/a by 1990.

Saponification value	195 – 205
Iodine value	44 – 58
Nonsaponifiable matter	0.5 %
mp	36 – 40 °C
n_D^{40}	1.453 – 1.456
Carotene	500 – 2000 ppm

For fatty acid composition see Tables 1 and 2.

Palm oil contains up to 40 % oleodipalmitin, consisting primarily of 2-oleodipalmitin.

Production. The fresh palm fruits contain very active fat-splitting enzymes (lipases), which must be deactivated as quickly as possible prior to isolating the oil. Since lipases are particularly active in damaged fruits, the whole fruit bunches are sterilized by heating with live steam in autoclaves; this deactivates the enzymes and loosens the fruit from the bunches.

The sterilized fruit bunches are threshed in a rotary drum stripper consisting of longitudinal channel bars; the fruit falls through whereas the empty bunch stem is retained in the drum. The separated, sterilized fruit is then converted into an oily mash by a mechanical stirring process known as digestion. The digested fruit mass is passed through screw presses, and the liquor from the press is clarified by static settling or by using decanter centrifuges. After washing with hot water and drying, the oil can be stored and transported.

High-grade palm oil contains ca. 3 % free fatty acids, inferior grades up to 5 %. Crude plantation oils obtained from palm fruits processed

under conditions that minimize autoxidation and lipolysis contain ca. 2 % free fatty acids and can be readily bleached; such oils are traded as super prime bleachable (SPB) grades.

Refining. Lye neutralization of palm oil may entail relatively high losses of neutral oil since inferior crude oils contain high concentrations of mono- and diglycerides, which exert a strong emulsifying effect. Palm oil is therefore often neutralized by distillation at temperatures up to 270 °C and 0.5 – 0.8 kPa (5 – 8 mbar) after pretreatment with phosphoric or citric acid and bleaching earth to remove foreign matter and traces of copper and iron compounds. During distillative neutralization (stripping), the oil is also bleached (i.e., heat-bleached), since carotenes are decomposed and removed. The carotenes in palm oil mainly consist of α-and β-carotene, with lower concentrations of γ-carotene, lycopene, and xanthophylls.

Distillative neutralization can be performed under conditions that simultaneously deodorize the oil. For very light raffinates, distillative neutralization is often followed by posttreatment consisting of a lye wash, bleaching with earth, and deodorization. Color and stability are largely a function of the quality of the crude oil. "White" palm oil, e.g., for biscuit fats, can generally only be made from crude SPB oils and by using fairly large amounts of bleaching earth. Bleaching with oxidizing agents, with air injected at 110 – 115 °C, or with activated earth at 140 – 160 °C is used only for technical applications. Palm oil can be effectively decolorized by mild hydrogenation at temperatures below 100 °C (hydrobleaching).

Palm oil can be separated into a solid and a liquid fraction. Two-stage cooling to 32 – 34 °C and 25 – 27 °C followed by separation over filter presses yields 20 – 25 % palm stearin, *mp* 50 – 52 °C, and 75 – 80 % liquid palm olein. The liquid glycerides can be separated from the solid fraction by centrifugation after addition of surfactant solutions. Fractionation in the presence of solvents (acetone, hexane) gives 10 % of a stearin melting at ca. 55 °C, 60 % olein, and 25 – 30 % of an intermediate fraction with properties similar to those of cocoa butter. The yield of the intermediate fraction can be increased to ca. 50 % by interesterification of the combined stearin and olein fractions followed by repeated fractionation.

Use. Palm oil is used predominantly for edible purposes. Refined palm oil is used extensively in shortenings and margarines. Carotene concentrates from crude oil have been largely replaced by synthetic carotene as coloring agents for margarines.

Palm stearin can be used in margarine and shortening blends whereas palm olein is used primarily in liquid frying fats and shortenings.

Technical applications of palm oil include tempering of metal and production of soap.

Other Palm Oil Varieties. These are commercially relevant only in their countries of origin. Tucum oil (saponifaction value 200, iodine value 40) is obtained from the fruit pulp of *Astrocaryum vulgare*, a palm growing in Central America.

13.1.2. Olive Oil

Olive oil [*68153-21-9*] is obtained from the fruit of olive trees (*Olea oleaster, Olea europaea oleaster*) and some related varieties such as *Olea americana* which is grown in South America.

Saponification value	185 – 196
Iodine value	80 – 88
Nonsaponifiable matter	<1.4 %
mp	−3 °C
n_D^{25}	1.466 – 1.468

For fatty acid composition see Tables 1 and 2.

Olive oil has a yellow to greenish-yellow color and becomes cloudy below −5 °C. Varieties with a higher cloud point can be made by winterization. The glycerides consist of 45 – 60 % monosaturated dioleins, 25 – 34 % dioleolinolein, and 4 – 29 % triolein. Saturated triglycerides are present at <1 %. The nonsaponifiable material contains saturated and unsaturated hydrocarbons, including 0.1 – 0.7 % squalene.

Production. The quality of olive oil depends on the ripeness of the olives, the type of harvesting (picking, shaking), intermediate storage, and type of processing. Olives contain 38 – 58 % oil

and up to 60 % water. Ripe olives should be processed as quickly as possible since lipases in the pulp cause rapid hydrolysis of the oil, impairing its quality for edible purposes. Top-grade oils are made from fresh, handpicked olives by comminution, pasting, and cold pressing.

Traditionally, olives were ground into a paste with stone mills; today modern milling equipment is used. Milling is followed by mashing, possibly with addition of salt. The pulp is then pressed and the press oil clarified by settling or centrifuging. Open-cage presses are being replaced by continuous screw expellers. The mashed pulp can also be separated in a horizontal decanter, the crude oil being recentrifuged after addition of wash water. An alternative is the use of machines to remove the kernels from the pulp; the residue is separated in self-discharging centrifuges.

Cold pressing, which yields vierge grades (also referred to as virgin, Provence, or Nizza oil), is generally followed by warm pressing at ca. 40 °C, which gives an oil with a less delicate flavor. The yields depend on the equipment used. The press cake (pomace) contains 8 – 15 % of a relatively dark oil, called Sanza or Orujo, which can be extracted with hexane and is used for technical purposes; after refining it is also fit for edible consumption.

Olive kernel oil is obtained by pressing and solvent extraction of cleaned kernels. It is similar to olive oil but lacks its typical flavor.

Use. Cold-pressed olive oil is a valuable edible oil. Trade specifications are based primarily on the content of free fatty acids and flavor assessment. In some countries, warm-pressed olive oil with a high acidity is refined by neutralization, bleaching, and deodorization, and flavored by blending with cold-pressed oil.

Testing for Adulteration. Because of its high price, olive oil is sometimes admixed with other oils. Some oils can be detected by specific color reactions: cottonseed oil by the Halphen test, sesame oil by the Baudouin reaction, teaseed oil by the Fitelson test (see Table 23). Gas chromatographic determination of the fatty acid and sterol composition is generally used to quantify the degree of adulteration. Olive-pomace oil and re-esterified olive oil, obtained by esterification of olive oil fatty acids or Sanza or Orujo oil with glycerol, can be detected by the glyceride composition.

13.1.3. Avocado Oil [77]

Avocado oil is isolated from the pulp of the pearlike fruit of the tree *Persea gratissima* L. growing in Southern Europe, South Africa, the Middle East, and Central America. Its composition and properties are similar to those of olive oil (see Tables 1 and 2). The oil is used in cosmetics; in California smaller quantities of avocado oil are used as special salad oil. The avocado fruit weighs up to 1.5 kg and contains 40 – 80 % oil in dry matter. World production of avocado fruit is ca. 1.5×10^6 t/a, but only a small fraction is used for oil production.

13.2. Seed-Kernel Fats

Oilseeds are the major source of oils and fats.

13.2.1. Lauric Acid Oils

Coconut, palm kernel, and babassu oil are the most important lauric acid oils. They contain >40 % of lauric acid and ca. 15 % of myristic acid.

13.2.1.1. Coconut Oil [76]

The coconut palm (*Cocos nucifera* or *Cocos butyracea*) is cultivated in coastal areas around the world within 20 ° of either side of the equator. The Philippines, Indonesia, Southern India, Sri Lanka, Equatorial Africa, and the West Indies are important producers. Around 1820 coconut oil [*8001-31-8*] was introduced to Britain (vegetable butter).

Saponification value	250 – 262
Iodine value	7 – 10
Nonsaponifiable matter	0.15 – 0.60 %
mp	20 – 28 °C
n_D^{40}	1.448 – 1.450

For fatty acid composition see Tables 1 and 2

The glycerides contain 50 – 60 % caprylolauromyristin and up to 20 % myristodilaurin in addition to smaller concentrations of laurodimyristin.

Production. In plantations the coconut palm reaches a height of 30 m and from its 6th to 30th years annually yields 50 – 70 coconuts with a diameter of 10 – 12 cm. The hard shell, covered by a fibrous husk, encloses the white endosperm tissue 1 – 2 cm thick, the copra. To obtain a light, flavor-stable coconut oil, the fresh copra with a water content of 60 – 70 % is dried in the sun or with hot air. This treatment prevents bacterial decomposition and lipolysis of the fat. Dry copra contains 60 – 67 % oil.

The dried copra is processed in an oil mill in two steps. About two-thirds of the oil is first obtained by expelling broken and rolled copra in continuous screw presses. The residual fat content of the expeller cake can be reduced to ca. 5 % by high-pressure expelling and to 2 – 4 % by subsequently extracting the expeller cake with hexane.

In order to obtain an edible oil of good quality, the crude coconut oil must be neutralized, bleached, and deodorized. Normally, the crude oil contains ca. 5 % free fatty acids but it can be lye-neutralized without great loss of neutral oil. The neutralized, washed, and dried oil contains only small amounts of pigments, phosphatides, and other constituents. It is decolorized with 1 – 2 % of bleaching earth and 0.1 – 0.4 % of activated charcoal. Activated charcoal also serves to remove polycyclic aromatic hydrocarbons deposited on the copra by drying with flue gases [53]. Crude coconut oil with a relatively high content of free fatty acids can be advantageously neutralized and deodorized by distillation after pretreatment with phosphoric acid and bleaching earth – activated charcoal.

Uses. Coconut oil is used for cooking, and is also an important component of vegetable margarines. Because of its high latent heat of fusion, it produces a pronounced cooling effect in the mouth on melting. This effect also makes it a valuable fat for biscuit filling, confectionery products, and confectionery coatings (couvertures). Coconut oil can undergo enzymatic hydrolysis in foods containing relatively high concentrations of water leading to so-called perfume rancidity. This reaction can be counteracted by incorporating at least 30 % sugar or by drying. Coconut oil has a high resistance to oxidative deterioration.

Coconut stearin, *mp* 27 – 32 °C, and coconut olein can be obtained by fractional crystallization and pressing of prerefined coconut oil. Coconut stearin is used in couverture for confectionery products.

Hydrogenated coconut oil (iodine value 2 – 4, *mp* 30 – 32 °C) is produced from prerefined coconut oil with fresh nickel catalyst at 140 – 180 °C. In the summer months it is often used in admixture with unhardened coconut oil to raise the melting point of couvertures and similar products.

Crude coconut oil, meeting certain color specifications, is also used in the manufacture of special soaps.

13.2.1.2. Palm Kernel Oil [76]

Palm kernel oil [*8023-79-8*] is derived from the kernels of the oil palm (*Elaeis guineensis*) (see Section 13.1.1).

Saponification value	242 – 254
Iodine value	16 – 19
Nonsaponifiable matter	0.2 – 0.8 %
mp	23 – 30 °C
n_D^{40}	1.449 – 1.452

For fatty acid composition see Tables 1 and 2

The properties of palm kernel fat are very similar to those of coconut oil. However, it contains more oleic acid and only half as much C_8 and C_{10} fatty acids. The glycerides of palm kernel oil consist of 60 – 65 % trisaturated, ca. 25 % disaturated – monounsaturated, and 10 – 15 % monosaturated – diunsaturated components. Its stability is similar to that of coconut oil.

Production. The palm nuts are separated by air classifiers from the fibers of the press cake obtained on pressing palm fruit (see Section 13.1.1). After drying in silo dryers, the nuts are cracked with centrifugal crackers, and the shell is separated from the kernel with air and water separation systems. The kernel, 1 – 2 cm in size, is dried from an initial moisture content of 20 – 25 % to about 4 – 7 %; it then contains 44 – 57 % oil. Following cracking and rolling, the palm kernels are passed through a screw expeller. The residual oil in the press cake can be further reduced by high-pressure expelling or, more commonly, by extraction with hexane.

Crude palm kernel oil often has a higher free fatty acid content than coconut oil (up to 15 %, calculated as oleic acid). Palm kernel oil with a free fatty acid content <5 % can be lye-neutralized without significant neutral oil losses. At higher contents of free fatty acids, distillative neutralization (possibly followed by lye post-neutralization) is generally preferred. Treatment with 1 – 2 % activated bleaching earth yields a light, yellowish oil. Deodorization generally requires a longer period of time than for coconut oil in order to obtain a stable, bland raffinate.

Use. Top-grade palm kernel raffinates are used in couvertures and margarines. Palm kernel stearin and olein are obtained by fractional crystallization of prerefined palm kernel oil at ca. 25 °C. Palm kernel stearin, *mp* 30 – 32 °C, is often used as such, or blended with coconut stearin, as a filling mass or couverture for confectionery products. Hydrogenated palm kernel oil (*mp* 39 °C, iodine value 1) is made in the same way as hydrogenated coconut oil and is mostly used in the confectionery trade. The melting point of hydrogenated palm kernel oil can be reduced from 39 °C to 33 – 35 °C by interesterification. Interesterified, hydrogenated palm kernel oil increases the range of special tailor-made products for application in confectionery products.

13.2.1.3. Babassu Oil and Other Palm Seed Oils

The babassu palm (*Orbygnia speciosa, Attalea funifara*) is a native of the great forest regions of the Brazilian states Maranhão, Piauí, Pará, and Minas Geraes. The number of babassu trees in these forests is estimated at $>10^9$. Collecting and transport of the babassu nuts to the oil mills and removal of the extremely hard shell prior to expelling make it difficult to exploit this interesting source of lauric oil. A bunch holds 200 – 600 fruits; 1 t of fruit yields about 125 kernels containing 63 – 70 % fat after drying to ca. 4 % moisture. The fat can be isolated by expelling and subsequent extraction with hexane.

Saponification value	242 – 253
Iodine value	10 – 18
Nonsaponifiable matter	0.2 – 0.8 %
mp	22 – 26 °C
n_D^{40}	1.449 – 1.451

For fatty acid composition see Tables 1 and 2

Use. Refined babassu oil is known in South America as an edible fat. In European countries it is used in place of palm kernel oil in margarine blends. The export of babassu kernels is only of minor economic importance; total production is likely to be around 150 000 t/a.

Other palm seed oils are used as edible fats in their countries of origin and do not play a role in international trade. Some of their properties are listed in Table 24. Cohune oil is obtained from the kernels of the cohune palm (*Attalea cohune*), a native of Mexico and Honduras. Its properties are similar to those of coconut oil. Murumuru oil, somewhat harder than coconut oil, is contained in the kernels of the palm *Astrocaryum murumuru*, which is found in the Northern provinces of Brazil. It contains ca. 40 % lauric acid, 35 % myristic acid, and only little caprylic and capric acids. Ouricuri oil, from the kernels of the Brazilian palm *Syagrus coronata*, has a lower melting point than murumuru oil. It contains 45 % lauric acid and 10 % myristic acid in addition to 20 % fatty acids with 6, 8, and 10 carbon atoms. The kernels of *Astrocaryum tucuma*, a native of Guyana, Venezuela, and Northern Brazil, yield tucum oil.

13.2.1.4. Other Sources of Lauric Acid Oils

Up to 43 % of lauric acid is contained in the seed oil of the evergreen laurel tree (*Laurus nobilis*), from which the name of this acid was derived. Laurel oil contains about 30 % trilaurin; it has no commercial importance.

The seed oils of *Myristica* species contain relatively high concentrations of myristic acid. Nutmeg butter [*8008-45-5*] is made by expelling ground and boiled seed kernels of *Myristica officinalis*; it contains 39 – 76 % myristic acid. *Myristica otoba* yields otoba fat. Ucuhuba fat is obtained from *Virola surinamensis*, a native to Brazil. Dika fat is contained in the seed kernels of *Irvingia gabonensis* and *I. barteri*, native to

Table 24. Properties and analytical data of seed oils of some South American palm varieties

	Cohune oil	Murumuru oil	Ouricuri oil	Tucum oil
Density (60 °C), g/cm^3	ca. 0.893	0.893	0.898	0.893
n_D^{60}	1.441	1.445	1.440	1.443
Saponification value	251 – 257	237 – 242	ca. 257	240 – 250
Melting point, °C	18 – 24	32 – 34	18 – 21	30 – 36
Iodine value	10 – 14	11 – 12	14 – 16	10 – 14
Nonsaponifiable matter, %	0.4	0.3	0.3	0.3

West Africa. None of these fats is involved in international trade.

Cuphea is a herbaceous, annual plant, native to Mexico and also found in Northern Brazil and Nicaragua. The seed oil contains various medium-chain fatty acids which, depending on the species, account for 40 – 80 % of the total fatty acids. Cuphea oil could be a substitute for coconut and palm kernel oil and serve as a natural source of capric acid; agronomic research is directed toward adaptation and yield improvement of *Cuphea* [78].

13.2.2. Palmitic – Stearic Acid Oils

13.2.2.1. Cocoa Butter

Cocoa butter is the seed fat of *Theobroma cacao* L., a tree reaching a height of 9 m, which is cultivated in many tropical countries (→ Chocolate, Chap. 6.1.). The cocoa tree is grown mainly in West Africa (Ivory Coast, Nigeria) and South Africa (Brazil). The world production of cocoa beans is in the order of 1.3×10^6 t/a.

Saponification value	190 – 200
Iodine value	35 – 40
Nonsaponifiable matter	0.2 – 0.5 %
mp	28 – 36 °C
n_D^{40}	1.453 – 1.458

For fatty acid composition see Tables 1 and 2

Cocoa butter has a pleasant aromatic flavor and keeps well. The cocoa butter glycerides are hard and brittle up to about 28 °C and melt in a range of 4 – 5 °C with a pronounced oral cooling effect. The melting point of cocoa butter depends on the type and length of tempering. After rapidly cooling to 0 – 5 °C, cocoa butter melts at 26 – 30 °C, whereas the melting point increases to 32 – 35 °C after tempering for 40 h at 28 °C. The dilatation curve (plot of solids content at different temperatures) is very steep (see Fig. 2).

The glycerides of cocoa butter consist roughly of 2 % trisaturated compounds, 14 % 2-oleodipalmitin, 40 % 1-palmito-3-stearo-2-olein, 27 % 2-oleodistearin, 8 % palmitodiolein, and 8 % stearodiolein.

Production. Cocoa butter is isolated from cocoa beans by fermentation, roasting, dehulling, grinding, and expelling (→ Chocolate, Chap. 6.1.). The fat content of dehulled cocoa beans is about 54 – 58 % in dry matter. In some countries the fat obtained from nondehulled cocoa beans by expelling or solvent extraction may also be called cocoa butter.

Use. Cocoa butter is used extensively in the production of chocolate and other confectionery products, and to a lesser extent in the pharmaceutical and cosmetics industries.

Test for Adulteration. Because of its high price, cocoa butter is sometimes admixed with press or extraction fat from damaged cocoa beans or byproducts, animal fats, coconut oil and stearin, palm oil and palm oil fractions, or hydrogenated vegetable or marine oils.

Extraction cocoa butter can be identified by its relatively dark color, by its strong fluorescence under ultraviolet light, by various color reactions, and by its behavior in simple tests such as the "smear test" (extraction fat tends to smear while authentic cocoa butter tends to disintegrate into small, hard pieces). In addition, extraction fat has a higher content of nonsaponifiable matter (cocoa butter contains 0.2 – 0.5 %, extraction fat 1.9 – 2.8 %, fat from cocoa bean hulls about 7.5 % nonsaponifiable matter) and a higher content of linoleic acid (10 – 29 % compared to 2 – 4 % in authentic cocoa butter).

Animal fats such as tallow and hydrogenated fish oils can be detected by gas chromatographic analysis of the fatty acid composition and by their cholesterol content. Hydrogenated vegetable oils can be recognized by their content of trans fatty acids. Coconut oil and similar oils can be detected by their concentrations of C_8, C_{10}, C_{12}, and C_{14} fatty acid groups. Addition of cocoa butter-like fats such as fractions of Borneo tallow, illipé fat, shea butter, and palm oil can sometimes be detected by their concentration of terpene alcohols. Gas chromatographic analysis of chocolate fat triglycerides can detect so-called cocoa butter equivalents in cocoa butter and certain types of chocolate; this technique can be augmented by analyzing the fatty acid composition at the 2-position [79].

13.2.2.2. Shea Butter, Borneo Tallow, and Related Fats (Vegetable Butters)

Shea butter, also known as Karité butter or Galam butter, is obtained from the seeds of a tree growing mainly in West Africa (*Butyrospermum parkii*). Plantations tend to be uneconomic since shea nuts can be harvested only after the tree is about 15 years old.

Shea butter is similar to cocoa butter (see Table 25). The content of nonsaponifiable matter, which contains rubber-like hydrocarbons such as kariten [$(C_3H_8)_n$, *mp* 63 °C], can be as much as 11 %.

The kernels can be readily separated from the fruit and the thin hulls. In Africa they are ground, boiled, and the fat is skimmed off. The commercial production of shea butter by expelling in continuous screw presses is difficult and requires special processing conditions. The green-brown crude fat can be neutralized only with considerable loss of neutral oil.

Refined shea butter can be used as an edible fat. The stearin fraction, primarily 2-oleodistearin, which can be obtained by crystallization and removal of the liquid fraction, is a valuable cocoa butter substitute. The export quantities (from Nigeria, Dahomey, Upper Volta) amount to ca. 50 000 t/a.

Borneo tallow (tengkawang tallow) is derived from *Shorea stenoptera*, a plant growing in the East Indies and Malaysia; there are about 100 different species including *S. gysbertsiana, S. palembassica,* and *S. seminis*. The proportions of component fatty acids in these species are very similar. The size of nut varies considerably from one variety to another, and only a few are worth harvesting. Borneo tallow resembles cocoa butter chemically and physically more closely than any other fat. For this reason it is used in preparing cocoa butter equivalents or extenders.

Sal butter is derived from *Shorea robusta*, commonly known as the sal tree, found extensively in parts of Northern and Central India. Illipé fat is derived from the group of Madhuca seed fats (also known as Bassia seed fat) and is itself produced from the Indian plant *Madhuca longifolia*. Another Indian plant, *Madhuca latifolia,* is the source of mowrah fat which is very similar in composition. Illipé and mowrah fats contain sufficient unsaturated fatty acids to make them resemble shea butter rather than the firmer cocoa butter and Borneo tallow. Illipé and mowrah fats are used principally in Asia for technical purposes but also for edible purposes when fully refined. Both fats are used to prepare cocoa butter extenders. Katiau fat is derived from the plant *Madhuca mottleyana,* and phulwara butter (Indian butter) from *M. butyracea.*

For properties of the various vegetable butters see Table 25 [80].

13.2.3. Palmitic Acid Oils

The oils in this group have a relatively high content of palmitic acid, mostly above 10 %. Cottonseed oil and some cereal oils are the most important representatives.

13.2.3.1. Cottonseed Oil

Cottonseed oil [*8001-29-4*] is obtained from the seeds of different varieties of cotton (*Gossypium*). The most important producers are the United States, the former Soviet Union, China, Pakistan, Brazil, Egypt, and India. *Gossypium barbadense* and *G. hirsutum* are preferred in the United States, while *G. herbaceum* is native to the Asian countries. The United States produce about one-third of the world's cottonseed oil (see also → Cellulose, Chap. 2.1.).

Table 25. Properties and analytical data of some vegetable butters

	Shea butter	Borneo tallow	Sal butter	Illipé butter	Mowrah butter	Katiau fat	Phulwara butter
Myristic acid, %		ca. 15					
Palmitic acid, %	5 – 6	18 – 22	8 – 9	ca. 28	16 – 27	ca. 10	54 – 57
Stearic acid, %	36 – 42	39 – 44	ca. 35	ca. 14	ca. 20	ca. 19	ca. 4
Arachidic acid, %		ca. 1	ca. 12				
Oleic acid, %	49 – 50	38 – 42	ca. 42	ca. 50	41 – 66	ca. 69	ca. 36
Linoleic acid, %	4 – 5	traces	ca. 3	8 – 9	9 – 14	2 – 3	3 – 4
Saponification value	178 – 196	189 – 200		186 – 200	187 – 195	189 – 192	188 – 200
Iodine value	55 – 67	29 – 38		50 – 60	58 – 63	53 – 67	40 – 51
mp, °C	32 – 42	34 – 39		25 – 29	23 – 31	30	38 – 43
n_D^{40}	1.4635 – 1.4668	1.4561 – 1.4573		1.459 – 1.462	1.458 – 1.461	1.461 – 1.462	1.455 – 1.458
Nonsaponifiable matter, %	2 – 11	<2		1.4 – 2.3	1.2 – 2.1	0.4 – 0.5	2 – 2.8
Fat content in seed kernel, %	45 – 55		20	50	50		60 – 65

Saponification value	190 – 198
Iodine value	100 – 117
Nonsaponifiable matter	0.5 – 1.5 %
mp	ca. 0 °C
n_D^{40}	1.464 – 1.468

For fatty acid composition see Tables 1 and 2

The oil contains 0.4 – 0.6 % of malvalic acid, which presumably is the basis of the Halphen color reaction. The glycerides of cottonseed oil consist of approximately 12 – 14 % disaturated – monounsaturated, 52 – 58 % monosaturated – diunsaturated, and 25 – 30 % triunsaturated compounds.

Production. Egyptian cottonseed (not dehulled) contains 22 – 24 % oil, the American varieties on average 19.5 %. About 50 cottonseeds are contained in a pod, which opens when it is ripe. After removal of the cotton, the seeds are still covered with fine hairs (linters), which are removed with delintering machines that essentially consist of rotating sawtooth discs [81]. The seeds are then dehulled by passage through "dehullers" (essentially rotating knives) and screens – air classifiers. The dehulled seeds (30 – 40 % oil content) are expelled in continuous screw presses and the expeller residue is extracted with solvent.

Gossypol is present in the seed in concentrations of 0.4 – 2.0 %. However, only traces are detectable in the crude oil, because gossypol is partially deactivated (presumably by interaction with protein) during conditioning and expelling the seed. Crude cottonseed oil has a dark color because of dissolved resins and pigments, and must therefore be refined before further use. The crude oil is first clarified by settling or centrifuging and then neutralized with lye, mostly in continuous centrifuge lines. The neutralized oil is free of gossypol, light yellow, stable during storage, and is usually traded in this form. A very stable oil can be obtained by postneutralization, bleaching, and deodorization. Stearin separates out at temperatures between 0 and 8 °C. An oil with a low cloud point is obtained by removing these solid glycerides by cooling and filtration (winterization).

Use. Winterized cottonseed oil is a very stable salad oil. Cottonseed oil is also used in margarine blends without being winterized. Lard-like fats for use in margarine blends and shortenings are obtained by hydrogenation under iomerizing conditions. Such fats have a melting point between 30 and 35 °C and an iodine value of 60 – 80. Cottonseed oil hardened to a melting point of 36 – 37 °C is suitable as a couverture fat for confectionery products. The characteristic Halphen color reaction disappears on hydrogenating the oil.

13.2.3.2. Kapok and Related Oils

Kapok oil is derived from the fruit of a tree belonging to the Bombaceae family that grows in Indonesia, Sri Lanka, the Philippines, and South America.

The seeds in the fruit pods are covered with fine hairs (kapok), which are used in upholstering and contain ca. 25 % oil. Crude kapok oil, produced by preexpelling and solvent extraction, has a reddish-brown color and is similar to cottonseed oil in composition and further processing. It contains little or no gossypol. Kapok oil contains ca. 15 % of a C_{18} cyclopropenoic acid and therefore responds strongly to the Halphen reaction.

Okra oil is obtained from the seeds of *Hibiscus esculentis,* a family related to Gossypium and found in the United States and some Mediterranean countries. Kenafseed oil is contained in the seeds of *Hibiscus canabinus,* a plant native to India where it is grown for its fibers.

The properties of kapok, okra, and kenafseed oils are listed in Table 26.

13.2.3.3. Pumpkin Seed Oil

Pumpkin seed oil is popular as a high-quality, edible oil in Southeast Europe. It is obtained from carefully decorticated pumpkin seeds by grinding, conditioning the moistened ground seeds, expelling in hydraulic presses, and clarification by settling in special tanks, mostly on smaller locations. It has a reddish-brown color and a nutty taste. Refining yields an oil with properties similar to those of sunflower oil.

Saponifaction value	185 – 198
Iodine value	117 – 130
Nonsaponifiable matter	0.6 – 1.5 %
Solidification point	ca. −15 °C
n_D^{40}	1.466 – 1.469

For fatty acid composition see Tables 1 and 2

Melon seed oil is very similar in composition to pumpkin seed oil [82], the major fatty acids being palmitic (ca. 12 %), stearic (ca. 11 %), oleic (ca. 11 %), and linoleic (ca. 65 %).

13.2.3.4. Corn (Maize) Oil

Corn oil (maize oil) [*8001-30-7*] is obtained from the germ removed during the processing of corn (*Zea mays* L.) into starch.

Properties. Winterized corn oil becomes cloudy at about −10 °C.

Saponification value	187 – 196
Iodine value	109 – 133
Nonsaponifiable matter	1.3 – 2.0 %
n_D^{40}	1.465 – 1.466

For fatty acid composition see Tables 1 and 2

Corn oil glycerides consist of ca. 2 % disaturated – monounsaturated, 40 % monosaturated – diunsaturated, and 58 % triunsaturated compounds.

Production. Corn contains 3.5 – 5 % oil, about 80 % of which is in the germ and about 20 % in the endosperm. The corn germ itself contains about 36 % oil. The germs can be separated by wet or dry processing. Wet processing tends to be preferred because it gives a higher oil yield. The cleaned corn is first conditioned by soaking in warm water and is then milled and slurried with water. The germs are collected by flotation, washed, and dried. In dry processing, the ground corn is separated into germ and endosperm fractions by screening and air classification; suitable moistening and conditioning facilitate this fractionation. The oil is isolated from the germs by preexpelling followed by extraction with hexane.

The golden yellow, crude corn oil contains up to 3 % free fatty acids. After lye neutralization, bleaching, and deodorization it yields a light, highly stable oil. The content of tocopherols, predominantly γ-tocopherol, is nearly 1000 ppm. Salad oils that remain clear at refrigerator temperatures are obtained by winterization, during which about 500 ppm wax is removed.

Use. Corn oil is a universal oil; it is used especially as a salad oil and in salad dressings.

13.2.3.5. Cereal Oils

Oils are generally not extracted from cereal grains except rice and wheat. The cereal oils are, however, very important as constituents of bread, other bakery products, and animal feeds.

Wheat germ oil is extracted from the wheat germ (8 – 11 % oil) obtained as a byproduct of

Table 26. Properties of kapok, okra, and kenafseed oils

	Kapok oil	Okra oil	Kenafseed oil
Myristic acid, %	ca. 0.5	0 – 4	
Palmitic acid, %	10 – 16	23 – 33	ca. 14
Stearic acid, %	3 – 8	0.5	ca. 6
Arachidic acid, %	ca. 1	traces	
Oleic acid, %	ca. 50∗	26 – 42	ca. 45
Linoleic acid, %	ca. 30	30 – 40	ca. 23
n_D^{40}	1.464 – 1.468	1.462 – 1.467	1.465 – 1.466
Saponification value	189 – 197	192 – 199	189 – 195
Iodine value	86 – 110	90 – 100	93 – 105
Nonsaponifiable matter, %	0.5 – 1.8	0.7 – 1.4	0.4 – 3.4

∗ Including 15 % malvalic acid.

milling wheat. It has a very high tocopherol content and is therefore popular as a dietetic oil.

Saponification value	180 – 189
Iodine value	115 – 126
Nonsaponifiable matter	3.5 – 6.0 %
mp	ca. 0 °C
n_D^{40}	1.468 – 1.478

For fatty acid composition see Tables 1 and 2

Rice bran oil [*68553-81-1*] is an important byproduct of rice processing. It is obtained by hexane extraction of the bran (brown outer coating) and the germ (8 – 16 % oil content) removed during grinding, dehulling, and polishing rice. The bran fraction must be extracted immediately, preferably after heat-sterilization, to destroy the very active lipases. Refined rice bran oil has a light yellow color and is a relatively stable edible oil. Rice bran oil hardened to 30 – 35 °C is often used in margarines and shortenings. In Japan and India ca. 100 000 t of rice bran oil are produced annually.

Saponification value	183 – 194
Iodine value	92 – 109
Nonsaponifiable matter	3.5 – 5.0 %
(including ca. 330 mg squalene per 100 g oil)	
mp	ca. −10 °C
n_D^{40}	1.466 – 1.469

For fatty acid composition see Tables 1 and 2

13.2.4. Oleic – Linoleic Acid Oils

This group encompasses a large number of drying and semidrying oils derived from plants growing in temperate and colder regions (see also → Drying Oils and Related Products). Some of these oils, such as hazelnut oil, poppyseed oil, walnut oil, and teaseed oil, only have regional significance, while sunflower oil, sesame oil, linseed oil, and safflower oil have considerable commercial importance.

13.2.4.1. Sunflower Oil

Sunflower oil [*8001-21-6*] is the seed oil of the sunflower, *Helianthus annuus,* that originated in America and is now grown extensively in East, West, and South Europe, the United States, Canada, South America, China, India, South Africa, and Australia. The annual plant reaches a height of 1 – 3 m, the short-stem varieties being preferred for ease of harvesting. The air-dry seeds contain about 45 % oil.

Properties. Sunflower oil begins to solidify at −16 °C; the cloud point is 0 – 5 °C.

Saponification value	188 – 194
Iodine value	125 – 144
Nonsaponifiable matter	0.4 – 1.4 %
n_D^{40}	1.466 – 1.468

For fatty acid composition see Tables 1 and 2

The linoleic acid content of sunflower oil is 50 – 70 % and tends to increase with decreasing temperatures during growing [83]. The glycerides contain 35 – 45 % diunsaturated and 56 – 63 % triunsaturated compounds. Linolenic acid is present to the extent of <0.4 %.

Production and Use. Sunflower oil is obtained by preexpelling followed by extraction of the expeller cake with hexane. Decortication is

practiced to a lesser degree; the hulls have to be burnt. The crude oil can be readily refined into a light, stable oil by lye neutralization, bleaching, and deodorization.

Cold-pressed sunflower oil is produced by expelling at $\leq 80\,°C$; posttreatment is limited to filtration and clarification. Such a "natural" oil is often recommended for dietetic purposes because of its relatively high concentration of phosphatides. Trace contaminants such as metals and pesticide residues, however, remain in the oil. Refined sunflower oil is used extensively in salad oils and margarine blends because of its oxidative stability. Winterization (removal of 500 – 2000 ppm of wax) of the deodorized or the neutralized and bleached oil at 10 – 15 °C yields an oil which does not cloud at refrigerator temperatures. Sunflower oil that is hydrogenated to a melting point of 32 – 35 °C (iodine value 70 – 75), under isomerizing conditions, is widely used in "one-oil" margarine blends.

Smaller quantities of a new sunflower variety are being grown in California; these plants yield an oil with an oleic acid content of about 80 % and hence a very high oxidative stability. This oil is recommended especially for salad and frying oils [84].

13.2.4.2. Sesame Oil

Sesame oil [*8008-74-0*] is contained in the seeds of *Sesamum indicum*, a plant resembling linseed and grown mainly in China, India, Africa, and Mexico. The flat seeds are 2 – 3 mm long and vary in color from white to brown-black; the oil content is 45 – 55 %. Harvesting is relatively difficult because of varying rates of ripening and the ripe seed pods shattering the seed. More uniform varieties with improved harvesting properties are being developed.

Properties. Sesame oil solidifies at −6 to −3 °C.

Saponification value	187 – 193
Iodine value	136 – 138
Nonsaponifiable matter	0.9 – 2.3 %
n_D^{40}	1.4665 – 1.4675

For fatty acid composition see Tables 1 and 2.

Crude sesame oil is yellowish and contains up to 3 % free fatty acids. The glycerides contain about 66 % dioleolinolein and oleodilinolein. Refined oil has a high stability due in part to the content of sesamol, a natural antioxidant formed by hydrolysis of sesamolin. The combined concentration of sesamol and sesamolin is 0.1 – 0.2 %. Sesamol gives a red color with an ethanolic solution of furfural in the presence of hydrochloric acid (color reaction of Baudouin and Villavechia); it also responds positively in color tests with an acidic solution of stannous chloride (Soltsien test), and with 3 % hydrogen peroxide – 75 % sulfuric acid (Kreis test).

Use. Sesame oil is widely used as an edible oil in the countries of origin. A small amount is exported for use in salad oils, salad dressings, and margarines. The addition of sesame oil to margarine is mandatory in some countries to provide a rapid means of identifying margarine by the above color reactions; for this application the oil may only be lightly bleached and deodorized at 150 °C.

13.2.4.3. Linseed Oil

Linseed oil [*8001-26-1*] is obtained from linseed or flax (*Linum usitatissimum*), which grows best in the temperate regions of Europe, Asia, and America. The flax fiber plant is a different variety of the same species. There is considerable variation in the iodine value of linseed oil from different regions. The iodine value of the oil tends to increase with severity of climate, but genetic and seasonal (e.g., rainfall) variations also have an influence [85]. While 180 – 185 is a typical iodine value range for individual lots, it may vary from 140 to 205.

The main producer countries are Canada, the United States, and Argentina. The brown, flat seeds contain ca. 40 % oil and 6 – 8 % water.

Properties. Fatty acid composition and other properties of linseed oil are given in Table 27 (see also Tables 1 and 2). Most applications require a linseed oil with a linolenic acid content of about 50 % and an iodine value of 170 – 190 [86], [87]. Linseed oil glycerides consist of about 5 % disaturated – monounsaturated, 43 % monosaturated – diunsaturated, and 52 % triunsaturated glycerides, which begin to solidify at

Table 27. Fatty acid composition and analytical data of vegetable oils of the oleic – linoleic group

	Linseed oil	Perilla oil	Hempseed oil
Myristic acid, %	ca. 0.5		
Palmitic acid, %	5 – 7	7 – 8	ca. 6
Stearic acid, %	3 – 5	traces	ca. 2
Arachidic acid, %	ca. 6		ca. 2
Lignoceric acid, %	traces		
Oleic acid, %	16 – 26	ca. 8	6 – 20
Linoleic acid, %	14 – 24	ca. 38	46 – 70
Linolenic acid, %	50 – 65	44 – 50	14 – 28
Saponification value	188 – 196	188 – 196	190 – 194
Iodine value	170 – 204	170 – 204	140 – 170
Solidification point, °C	10 – 21	12 – 17	15 – 17
n_D^{25}	1.4786 – 1.4815	1.4800 – 1.4820	
Nonsaponifiable matter, %	ca. 1.5	<1.5	<1.5

−18 to −27 °C. Most of the glycerides contain an average of seven double bonds per molecule, which is the reason for the excellent drying properties of linseed oil (→ Drying Oils and Related Products). The use of linseed oil for edible purposes is limited to smaller quantities of cold-pressed oil (for dietetic purposes) and to hydrogenated linseed oil, which can be used in low concentrations in shortenings, and margarine blends.

Some wild linseed varieties (e.g., *Linum capitatum* and *L. flavum*) have linolenic acid contents as low as 11 %. Modern biotechnological techniques are being used in Australia and Canada in an attempt to transfer this property to *L. usitatissium*. An alternative approach is γ-ray induced mutation of *L. usitatissimum*.

Production. Linseed oil is obtained from the seed by preexpelling followed by hexane extraction of the press cake. After refining it is used predominantly in the technical sector for paints and coatings. A small amount is used in East European countries as edible oil. Linseed oil has a characteristic odor and flavor which cannot be removed entirely by refining. The crude oil is rich in phosphatides and gums which sediment on storage. The crude oil is often pre-deslimed by treatment with hot water followed by centrifugation. Oils virtually free of phosphatides and gums are made by post-desliming with sulfuric or phosphoric acid. Very light drying oils can be made by a further posttreatment consisting of lye neutralization and earth bleaching. To obtain virtually odorless oils, deodorization is required as a final step. Heating refined linseed oil to 260 – 285 °C increases its viscosity and gives a so-called "boiled" linseed oil, which is used in special paints and coatings.

13.2.4.4. Perilla Oil

The seeds of *Perilla ocymoides*, an Asian plant, yield perilla oil (30 – 50 %), which resembles linseed oil in appearance and application. For properties see Table 27. World production is about 25 000 t/a. Perilla oil is sometimes blended with linseed oil; in Asia it is also used as an edible oil .

Lallemantia oil (Asia) and Chia seed oil (Mexico) are further Labiatae seed oils with similar compositions.

13.2.4.5. Hempseed Oil

The seeds of hemp (*Cannabis sativa* L.) contain 30 – 35 % oil. Hempseed is grown in the Soviet Union, India, China, Japan, and Chile, and yields fibers in addition to the oil. Hashish can also be obtained from Indian hempseed varieties. Hempseed itself can be used as a foodstuff. The oil is normally produced by pressing the seed. For properties see Table 27. Hempseed oil has been largely replaced by linseed oil for technical purposes.

13.2.4.6. Teaseed Oil

Teaseed oils are obtained in China, Japan, India, and Turkey from the hazelnut-like seeds of

tea shrub varieties, which contain about 60 % oil. Tea shrubs grown for oil are generally not suited for tea-leaf harvesting. Teaseed oil is produced primarily from the seeds of *Thea sasanqua* n. (sasanqua oil) and *Thea japonica* n. (tsubaki oil) by drying, grinding, conditioning, and pressing or solvent extraction. After refining the oil is used for edible purposes, as a special lubricant, and for toiletries.

The principal acids of teaseed oils [88] are palmitic (ca. 16 %), oleic (ca. 60 %), and linoleic acid (ca. 22 %). Teaseed oils can be detected in concentrations above 10 % in olive oil by the Fitelson color reaction. The expeller cake and extraction meal cannot be used as such in animal feeds because of their high saponin contents. Total world production is ca. 30 000 t, half of which is exported.

13.2.4.7. Safflower and Niger Seed Oils

Safflower oil [*8001-23-8*] is the seed oil of the thistle-like safflower plant (*Carthamus tinctorius*), thriving in the West of the United States and in Mexico, North Africa, and India. The plant can be grown under fairly arid conditions, the oil yield increasing with available moisture in the soil. The seeds resemble small sunflowerseed kernels and can be harvested mechanically. The oil content of the seed is 25 – 37 %.

Saponification value	180 – 194
Iodine value	136 – 152
Nonsaponifiable matter	0.3 – 2.0 %
Solidification point	−13 to −25 °C
n_D^{40}	1.467 – 1.469

The fatty acid composition is similar to that of sunflower oil (see Tables 1 and 2).

Production. Modern processing involves grinding the seed followed by conditioning, expelling, and hexane extraction of the expeller cake.

Use. Safflower oil has a high oxidative stability and is being used increasingly in salad oils and dietetic margarines because of its high content of linoleic acid. Some years ago special varieties with oleic and linoleic acid contents of 80 % and 15 % respectively ("high oleic" varieties) were developed and recommended as especially stable frying oils. World production amounts to about 200 000 t/a.

Niger seed oil, closely resembling safflower oil in its composition, is derived from the seed of the gingli plant, which is grown fairly extensively in Ethiopia, Togo, and India.

13.2.4.8. Grape-Seed Oil

Grape-seed oil [*60-33-3*] is a valuable oil obtained from the grapeseed left in winery pomace [89]. The oil content averages 15 % (dry weight basis).

Properties. The flavor of grape-seed oil resembles that of olive oil; however, its fatty acid composition is similar to that of sunflower oil (see Tables 1, 2).

Saponification value	180 – 196
Iodine value	124 – 143
Nonsaponifiable matter	0.3 – 1.6 %
Solidification point	−10 to −20 °C
n_D^{40}	1.464 – 1.471

Use. After suitable refining (removal of chlorophyll can be difficult), grape-seed oil keeps relatively well and is used as a salad and cooking oil and in special margarine blends. In France and Italy, 5000 t and 20 000 t, respectively, of grape-seed oil are produced annually. It has been estimated that the world production of grape seed is 1.4×10^6 t, with a potential oil yield of about 190 000 t [82].

13.2.4.9. Poppyseed Oil

Poppyseed oil is obtained from the seeds of different varieties of *Papaver somniferum*. The oil content of air-dry seeds is 44 – 50 %. Cold-pressed oil is suitable for edible purposes, as is preexpelled and extracted oil after refining. Poppyseed oil is also used in smaller quantities for superior oil paints. The major fatty acids are stearic (ca. 10 %), oleic (ca. 16 %), and linoleic acid (70 – 75 %).

Saponification value	ca. 193
Iodine value	ca. 140
n_D^{40}	ca. 1.468

The world production of poppy seed amounts to about 25 000 t/a.

13.2.5. Leguminous Oils

The commercially most important representatives of this group are peanut oil and soybean oil. Soybean oil [*8001-22-7*] holds first place in the worldwide production of vegetable oils.

13.2.5.1. Soybean Oil

The cultivation of the soybean (*Glycine maxima*) in the United States and South America has steadily increased. The soybean has been known in China for 5000 years. Main producing countries today are the United States, Brazil, and Argentina. Cultivation in the temperate regions of Europe is possible but the hectare returns are generally not sufficiently attractive. New, higher-yielding varieties are constantly being developed; in the United States they thrive best in the "corn-belt," i.e., Illinois, Minnesota, Iowa, Indiana, Ohio, and Missouri.

The soy pods contain up to four soybeans (5 – 10 mm in length). The oil content varies between 17 and 22 % and the protein content reaches 40 – 45 % in dry matter. Climatic conditions have a considerable influence on the oil and protein content; late sowing and a cool, wet climate tend to lower the oil content and to increase the protein content. Development of low linolenic soybeans in the United States has been hindered by lower seed productivity.

Properties. Refined soybean oil has a light yellow color and a bland flavor.

Saponification value	188 – 195
Iodine value	120 – 136
Nonsaponifiable matter	0.5 – 1.5 %
Solidification point	−15 to −8 °C
n_D^{40}	1.465 – 1.469

For fatty acid composition see Tables 1 and2.

The oil contains 40 – 60 % triunsaturated, 30 – 35 % diunsaturated, and up to 5 % monounsaturated glycerides. In spite of the high tocopherol content (up to 1200 ppm) the stability of the refined oil is limited, primarily because of the content of linolenic acid. Autoxidation leads to "green", "seedy" off-flavors. Soybean oil can be detected in other oils by its fatty acid, tocopherol, and sterol contents.

Production. Soybeans are normally cracked and flaked prior to extraction with hexane. Preexpelling is not normally performed. Soybean meal is a valuable animal feed. Crude soybean oil contains 0.5 – 1.0 % free fatty acids as well as up to 2.5 % phosphatides, which are removed by hydration with 2 – 3 % water at 70 – 80 °C followed by centrifuging. Drying of the resulting lecithin sludge yields lecithin. Soybean oil is often postdegummed with phosphoric or citric acid prior to neutralization with lye, earth bleaching, and deodorization.

The yield of lecithin can be increased by steam-heating the flakes prior to extraction [39]. With the aid of special degumming techniques [44], the phosphorus content of the crude oil can be reduced to 20 ppm, which facilitates subsequent physical refining (i.e., direct treatment with acids and earth followed by combined stripping of free fatty acids and deodorization).

Use. Soybean oil is used in almost all fatty products, from salad oils to margarines. Hydrogenation with fresh catalyst at ca. 100 °C yields a liquid oil with significantly improved stability due to hydrogenation of linolenic acid. Hydrogenation under isomerizing conditions to melting points ranging from 36 to 43 °C yields products for use in shortenings and margarines. Liquid soybean oil is a so-called semidrying oil; it is used in combination with tung oil and as a component of alkyd resins for coatings.

13.2.5.2. Peanut Oil

Peanut oil (groundnut oil) [*8002-03-7*] is obtained from the seed kernels of the peanut plant (*Arachis hypogaea*), native to South America. The peanut is now being grown in China, India, South Africa, West Africa, Argentina, and

the United States. Senegal and Nigeria are the main African export countries. The annual plant thrives best on light, sandy soils and reaches maturity in 4 – 5 months. Peanuts for further processing are mostly shipped after shelling, whereas peanuts for direct consumption are normally exported in the pods. Peanuts are often designated according to their shipping ports, e.g., Bombay, Casamance, Rufisque, Bissao nuts.

Properties. The iodine value of peanut oil is 84 – 105; Rufisque oil has an iodine value of 84 – 90 and Argentinian oil a value of 100 – 105 in line with the higher content of linoleic acid. The saponification value is 185 – 196, n_D^{40} 1.461 – 1.465. Tocopherols and other antioxidants as well as hydrocarbons and sterols are found in the nonsaponifiable matter (0.5 – 1.0 %). The oil solidifies below 0 °C. For fatty acid composition see Tables 1 and 2. The oil contains 36 – 46 % triunsaturated, 47 – 52 % diunsaturated, and 7 – 11 % monounsaturated glycerides.

Production. The peanut pods contain up to four hazelnut-sized kernels which are covered with a thin, reddish-brown skin and contain 5 – 12 % water, 45 – 50 % oil, and 23 – 35 % protein. The pods are first removed by passage through cracking rolls or disk mills and over screens. This step also loosens the skins, which are removed by aspiration. The kernels are then screened, cracked, and expelled in screw presses. The residual oil in the expelled cake is extracted with hexane. Crude peanut oil contains low concentrations of phosphatides and up to 1.5 % free fatty acids. Lye neutralization, earth bleaching, and deodorization yield a light, slightly yellowish oil with a very good oxidative and flavor stability. Peanut kernels to be processed into oil and meal can contain up to 0.5 ppm of aflatoxins. The bulk of the aflatoxins remain in the meal where they can be deactivated by treatment with, for example, ammonia; traces of aflatoxin that might enter the oil are removed during neutralization with lye.

Use. Peanut oil is used as a superior salad and cooking oil, less so in margarines. In shortenings it is used as such or after hardening (*mp* 31 – 38 °C). On cooling to 6 – 8 °C, peanut oil becomes cloudy due to the crystallization of glycerides containing $C_{20} - C_{24}$ fatty acid groups. The glycerides can be removed by filtration but only with a low yield of clarified oil.

Peanut butter is made by grinding specially roasted peanuts and homogenizing the mash with addition of liquid and possibly also hydrogenated peanut oil.

13.2.5.3. Lupine Oil

Lupins are attracting attention worldwide as an arable crop for possible use in foods and feeds. *Lupinus albus, L. angustifolius, L. luteus,* and *L. mutabilis* are potential sources of edible oil. Whereas the first three species (sweet varieties) contain only 4 – 9 % oil in the dried seed, the seed oil content of *L. mutabilis* approaches that of soybeans. The main fatty acid components are palmitic (7 – 13 %), stearic (2 – 7 %), oleic (27 –53 %), linoleic (18 – 47 %), linolenic (2 – 9 %); lower concentrations of $C_{20} - C_{22}$ fatty acids are found [90]. The meal of *L. mutabilis* (bitter lupins) cannot be used as an animal feed without prior extraction of the alkaloids.

13.2.6. Cruciferous Oils

The seed oils of the traditional cruciferous plants are mainly characterized by their content of glucosinolates (sulfur-containing glycosides) and erucic acid. Rapeseed oil is the most important cruciferous oil.

13.2.6.1. Rapeseed Oil [91]

Rapeseed oil [*8002-13-9*] is the seed oil of different varieties of *Brassica napus* and *B. campestris*. The air-dry seed contains about 40 % oil and 7 – 9 % water. The production of rapeseed, one of the oldest oil crops known, has increased dramatically over the past 10 years. The introduction of so-called single-zero rapeseed varieties, i.e., varieties with an erucic acid content of less than 5 % of the total fatty acids, has made rapeseed oil universally applicable as an edible oil. Generally, it can be used to replace soybean oil. Whereas single-zero, spring-sown rapeseed varieties are grown in Canada, the EEC countries (with the exception of Denmark)

generally prefer winterhardy single-zero varieties because of the higher crop yields. Within the European Community the traditional rapeseed varieties with up to 50 % erucic acid in the oil are now Within the European Community only single-zero varieties are eligible for subsidy, and the United States has recently permitted the use of rapeseed oil with max. 2 % erucic acid. Whereas single-zero, spring-sown rapeseed varieties are grown in Canada, the EEC countries (with the exception of Denmark) generally prefer winterhardy single-zero varieties because of the higher crop yielused only for technical and specialty purposes.

The increase in planted acreage has led to greater availability and to processing throughout the year; a significant improvement of the quality of the meal and the oil has been achieved [92]. Canada, the European Community, Poland, Sweden, China, India, and Pakistan are the main producing countries. To facilitate the marketability of meal, double-low varieties, i.e., low in erucic acid and glucosinolates, are grown in Canada (spring-type) and are being introduced in the EEC (winter-type). The winter-sown varieties are larger in size and generally easier to process than the smaller spring-sown varieties.

Properties [92]. For fatty acid compositions see Tables 1 and 2. Analytical data such as iodine value and saponification value are functions of the erucic acid content. The high level of erucic acid in traditional rapeseed oil has been associated with irreversible organ changes in animal tests [92].

Production. Modern rapeseed processing entails flaking and cooking the seed prior to preexpelling in continuous screw presses; the press cake is extracted with hexane. Crude rapeseed oil contains up to 2 % free fatty acids and ca. 2.5 % phosphatides; it is degummed with water or acids, lye-neutralized, bleached with earth, and deodorized, analogously to soybean oil. The stability is also comparable. The phosphatide content of the crude oil can be reduced to 0.1 % (i.e., 20 ppm phosphorus) by special degumming techniques [44], which enable physical refining to be carried out.

The enzymatic degradation products of sulfur-containing glycosides (glucosinolates) in rapeseed can influence the taste of the crude oil and impair the nutritional properties of the meal [92]. The enzyme responsible for glucosinolate degradation, myrosinase, is destroyed by adequate cooking of the seed prior to expelling. Sulfur compounds are removed from the oil during refining; residual traces can affect the course of hardening since they act as catalyst poisons. The residual sulfur compounds in the extracted meal can be deactivated by treatment with alkali and heating.

Use. Traditional rapeseed oil containing up to 50 % erucic acid can be hardened into steep-melting products, *mp* 32 – 34 °C, suitable as partial coconut oil substitutes in shortenings and couvertures. Modern single-zero varieties are used as such, after mild hardening (which eliminates the linolenic acid), and after hardening under isomerizing conditions to raise the melting point to 30 – 43 °C. The reduced complexity in fatty acid composition of the hardened oil leads to a different crystallization behavior, which can be overcome by interesterification. Treatment of rapeseed oil with sulfur or sulfur compounds is the basis of mastic production.

13.2.6.2. Mustard Seed Oil

Mustard seed oil [*8007-40-7*] is derived from the seeds of black mustard (*Brassica nigra* or *Sinapis nigra*), brown mustard (*Brassica juncea*), and white mustard (*Brassica alba* or *Sinapis alba*). The seeds contain 30–35 % oil and high levels of glucosinolates. The erucic acid content of the oil is 40 – 50 %; it is therefore of interest primarily for technical applications.

13.2.7. Conjugated Acid Oils

13.2.7.1. Tung Oil and Related Oils

Tung oil [*8001-20-5*] is obtained from the kernels (oil content about 50 %) of the fruit of the tung tree, *Aleuritis fordii* (Chinese tung oil) or *Aleuritis cordata,* syn. *vernica* and *verrucosa* (Japanese tung oil). The oil from *Aleuritis montana* is almost identical in composition. Oils similar to Chinese tung oil are obtained from the nuts of *Aleuritis trisperma* (kekuna oil) and *A. moluccana* or *A. triloba* (lumbang oil). Tung trees

are cultivated in Southern China, Southern Russia, the Southern United States, and Argentina. Tung oil contains about 80 % α-eleostearic, 4 % linoleic, 3 % linolenic, 8 % oleic, 1 % stearic, and 4 % palmitic acid. For properties, production, and use, see → Drying Oils and Related Products.

13.2.7.2. Oiticica Oil and Related Oils

Oiticica oil [*8016-35-1*] is obtained mainly from the nuts of the Brazilian oiticica tree, *Licania rigida*. This evergreen tree reaches a height of 20 m and produces 150 – 900 kg of nuts per year. The kernels contain 55 – 63 % oil, which is generally obtained by pressing. Fresh oiticica oil is yellowish and of lardlike appearance. Oiticica oil for export is heated for 30 min at 210 – 220 °C, after which it remains liquid. Licanic acid, 4-oxo-9,11,13-octadecatrienoic acid, is the principal fatty acid. The oil yields varnishes similar to those made from tung oil (→ Drying Oils and Related Products).

Boleko or isano oil (ongoke oil) is obtained from the nuts of *Ongokea gore* ENGLER, a tree growing in the Congo regions. It contains isanic and isanolic acids, which are fatty acids with conjugated triple bonds.

Parinarium oils are contained in the kernels of various tropical trees of the family Rosaceae and have little commercial significance. *Parinarium laurinum*, a tree growing in Japan and Oceania, yields a seed oil which contains parinaric acid, an acid with 4 conjugated double bonds.

Néou oil, obtained from *P. macrophyllum*, contains about 30 % eleostearic acid. Eleostearic acid, in addition to licanic acid, has been found in po-yoak oil (obtained from *P. sherbroense*).

13.2.8. Substituted Fatty Acid Oils

13.2.8.1. Castor Oil

The seeds of the castor tree, *Ricinus communis*, contain 45 – 50 % castor oil [*8001-79-4*]. The evergreen castor tree belongs to the family Euphorbiaceae and grows in tropical and subtropical countries. It can reach a height of 10 m.

Properties. The fatty acids of castor oil consist of 87 – 91 % ricinoleic acid, 2 % stearic and palmitic acid, 4 – 5 % oleic acid, 4 – 5 % linoleic acid, and 1 % dihydroxystearic acid. Pure castor oil can be readily identified by its hydroxyl value, viscosity, and specific gravity. The oil is soluble in ethanol at room temperature and in boiling hexane.

d_{25}^{25}	0.945 – 0.965
n_D^{40}	1.466 – 1.473
Solidification point	−12 to −18 °C
Viscosity at 20 °C	935 – 1033 mPa s
$[\alpha]_D$	+7.5 to +9.7 °
Saponification value	177 – 187
Iodine value (Wijs)	82 – 90
Hydroxyl value	161 – 169
Acetyl value	144 – 150
Nonsaponifiable matter	0.2 – 0.3 %

Production. Cold pressing of the spotted, oblong seeds yields a light, viscous oil which is used primarily in pharmaceutical products. Subsequent expelling at elevated temperatures and extraction with hexane yields a yellowish-brown oil which is used primarily in technical applications. The extracted meal contains ricin, a toxic protein, and hence cannot be used as an animal feed.

Use. A drying oil can be made by dehydrating castor oil (dehydrated castor oil, → Drying Oils and Related Products, Chap. 4.). Pyrolytic cleavage at 300 °C gives undecylenic acid and heptaldehyde, raw materials for polymers and perfumes. Alkali fusion of castor oil yields 2-octanol and sebacic acid.

Hydrogenation yields a fat with a melting point of 84 – 86 °C and waxlike properties, finding application as a special lubricant. Air-blown castor oil is used as plasticizer for varnishes and polymers. The oil is also used in the production of transparent soap, textile processing aids, special lubricants, and toiletries.

13.2.8.2. Chaulmoogra, Hydnocarpus, and Gorli Oils

The seed kernels of *Taractogenos kurzii*, a tree native to Southeast Asia, contain 48 – 55 %

Table 28. Properties of some "leprosy oils"

	Chaulmoogra oil (*Taractogenos kurzii*)	Hydnocarpus oil (*Hydnocarpus whigtiana*)	Gorli oil (*Oncoba echinate*)
Palmitic acid, %	ca. 4	1 – 2	7 – 8
Oleic acid, %	13 – 15	5 – 7	2 – 3
Lower homologues of hydnocarpic acid, %	ca. 0.5	3 – 4	
Hydnocarpic acid, %	ca. 35	ca. 50	
Chaulmoogric acid, %	ca. 23	27	ca. 75
Gorlic acid, %	ca. 23	12 – 13	ca. 14
n_D^{40}	1.471 – 1.473	1.472 – 1.473	
$[\alpha]_D^{25}$, °	+49.8	+55	+51.7
mp, °C	22 – 30	22 – 24	42 – 44
Saponification value	198 – 208	200 – 208	190 – 194
Iodine value	98 – 105	93 – 101	94 – 100
Nonsaponifiable matter, %	<0.5	<0.5	1 – 1.5

chaulmoogra oil, the chief fatty acid components of which are chaulmoogric acid, hydnocarpic acid, gorlic acid, and lower homologues of hydnocarpic, palmitic, and oleic acid. Chaulmoogra oil and related oils such as hydnocarpus oils, obtained from Southeast Asian *Hydnocarpus* varieties, and the African gorli oil have been used in the treatment of leprosy; for properties see Table 28.

13.2.8.3. Vernonia Oil [93]

The seeds of *Vernonia anthelmintica* (of the Compositae family) contain an oil with a high percentage of vernolic acid (*cis*-12,13-epoxy-*cis*-9-octadecenoic acid). This epoxy acid is used in protective coatings, plastics, and other industrial products. Efforts have been made to grow *Vernonia* varieties as new industrial crops in the United States and Asia.

14. Individual Animal Fats

Animal fats are obtained from the milk or the fatty tissue of certain groups of animals (see also → Milk and Dairy Products). The content of depot fat varies from 0 to 60 %. The concentration and composition of the fat depend on type of animal, age, sex, and diet.

14.1. Land-Animal Fats [94]

14.1.1. Lard

Lard [*61789-99-9*] is the fat rendered from fresh, clean, sound fatty tissues from pigs in good health at the time of slaughter. The tissues do not include bones, ears and tails, internal organs, windpipes, or large blood vessels. Rendered pork fat is obtained from the tissues and bones of pigs in good health at the time of slaughter; it may contain fat from bones, skin, ears and tails, and other tissues fit for human consumption. A range of different quality grades is available.

Properties. European lard (iodine value ca. 60) is often harder than America lard (iodine value ca. 70).

mp	30 – 40 °C
Solidification point	22 – 32 °C
Saponification value	193 – 202
Nonsaponifiable matter	0.1 – 1.0 %

Lard consists of 4 – 8 % trisaturated, 32 – 40 % disaturated – monounsaturated, 45 – 50 % monosaturated – diunsaturated, and 3 – 10 % triunsaturated triglycerides. The main fatty acid components are shown in Tables 1 and 2. Lard contains low concentrations of arachidonic acid (0.4 – 0.9 %), trans fatty acids, as well as traces of branched and odd-numbered fatty acids such as saturated C_{15}, C_{17}, and C_{19} fatty acids and monounsaturated C_{17} and C_{19} fatty acids. Triunsaturated C_{18}, C_{20}, and C_{22} fatty acids are

also present. These typical fatty acids, together with the cholesterol that occurs in relatively high concentrations in animal fats, are the basis for detecting lard and other animal fats in vegetable oils and fats.

Production. Lard is obtained by dry or wet rendering. The best quality lard (acid value <0.8; moisture <0.1 %) is made from selected tissue which has been washed and cooled immediately after slaughter. After comminution, the fatty tissue is wet-rendered by steam heating at 50 – 60 °C (for continuous wet-rendering plants see Section 5.2). Prime steam lard and leaf lard are obtained similarly from nonselected or partially selected tissue. Packer's lard retains a typical flavor; refiner's lard has to be refined. Inferior varieties, made in part from inedible offal, include so-called white, yellow, and brown greases, which are used primarily in technical applications.

Lard can be fractionated into a higher- and lower-melting fraction. In order to prevent the use of inferior raw materials, some European countries do not permit the refining of lard. The addition of antioxidants significantly improves the organoleptic and oxidative stability. Lard can be refined either by neutralization with lye, followed by bleaching with ca. 1 % earth and deodorization at 180 – 240 °C, or by simple pretreatment with phosphoric acid – bleaching earth and subsequent distillative stripping. Interesterification improves the performance of lard as a shortening in terms of creaming and cakemaking [95].

Use. Lard is used extensively as an edible fat (shortening, margarines) and as an energy source in animal feeds.

14.1.2. Beef Tallow

Properties. The main fatty acids of beef tallow [*61789-97-7*] are shown in Tables 1 and 2. In addition, it contains up to 4 % of characteristic branched and odd-numbered fatty acids. These, together with cholesterol, identify tallow in vegetable oils and fats. Tallow also contains 6 – 10 % of trans fatty acids, resulting from bacterial hydrogenation in the rumen. Top-grade tallow is white to greyish-white; in summer it is slightly yellow because of carotenes taken up with fresh roughage. Table 29 shows analytical data for various beef tallow products. Beef tallow contains 14 – 26 % trisaturated, 22 – 34 % disaturated – monounsaturated, and 40 – 64 % monosaturated – diunsaturated glycerides.

The stability of beef tallow and its products is relatively poor in the absence of an added antioxidant. Primary quality criteria are odor and taste, concentration of free fatty acids, and absence of impurities and of refined low-quality tallow (white grease).

Production. The choice of fatty tissue determines the quality of beef tallow. Fat rendered from the carcasses of cattle is generally not quite as hard as that obtained from sheep and goats. The softest, most unsaturated fat is found under the skin, and the firmer fat is located near the middle of the animal. The consistency of individual lots of tallow is therefore influenced by the procedure used in trimming the carcass prior to fat rendering.

Edible tallow (dripping, max. 1 % ffa) is obtained from clean, sound, fatty tissues of beef cattle in good health at the time of slaughter. Premier jus (max. 0.5 % ffa) is obtained by rendering the fresh fat (killing fat) of heart, caul, kidney, and mesentery, excluding cutting fats at low temperature (50 – 60 °C). Tallow to be incorporated into margarine blends must be refined. Fractionation at 26 °C yields a lower-melting fraction (olein) and a higher-melting fraction (stearin). Tallow olein is an important component of some margarines and shortenings. Stearin is used in shortenings and in specialty margarines. Bone grease is obtained by expelling and extracting of comminuted bones ; it is used primarily in the technical sector and for animal feeds. Neatsfoot oil [*8037-20-5*] is a low-melting, inedible fat rendered from the feet of cattle; it is used as a special lubricant and leather dressing aid.

Use. Only a small part of the annually produced beef tallow is used as edible fat. The concentration in margarine is limited because of its poor stability but it is used more widely in shortenings. The technical grades are important raw materials for soap and fatty acid derivatives and for use in animal feeds. Specifications for tech-

Table 29. Properties and analytical data of beef tallow and derived products

	Beef tallow	Stearin	Olein	Bone grease	Neatsfoot oil
Density, g/cm^3	0.898 – 0.908 (40 °C)		0.914 – 0.924 (15 °C)		
n_D^{60}	1.451 – 1.454	ca. 1.449		1.451 – 1.452	1.461 – 1.463
mp, °C	40 – 50	50 – 55	23 – 35	44 – 45	
Saponification value	193 – 200	190 – 198	193 – 198	190 – 200	188 – 198
Iodine value	32 – 47	14 – 25	40 – 53	49 – 53	67 – 80
Nonsaponifiable matter, %	0.3 – 0.8	0.3 – 0.8	0.1 – 0.5	0.5 – 0.6	0.1 – 0.6

Table 30. The Society of British Soap Makers specifications for tallows and greases (based on testing methods specified in B.S.3919 : 1976)

Grade	ffa ∗, %	Bleached oil colour (red)	Moisture and dirt, %	Nonsaponifiable matter, %	Titer, °C	Iodine value
Tallow 1	max. 3.0	max. 0.5, 5 1/4 inch cell	max. 0.5	max. 0.5	min. 40.0	max. 57
Tallow 2	max. 5.0	max. 1.0, 5 1/4 inch cell	max. 1.0	max. 1.0	min. 40.0	max. 57
Tallow 3	max. 8.0	max. 3.0, 5 1/4 inch cell	max. 1.0	max. 1.0	min. 40.0	max. 57
Tallow 4	max. 12.0	max. 4.0, 1 inch cell	max. 1.0	max. 1.5	min. 40.0	max. 60
Tallow 5	max. 15.0	max. 12.0, 1 inch cell	max. 1.0	max. 1.5	min. 40.0	max. 60
Tallow 6	max. 20.0	no limit	max. 1.0	max. 2.0	min. 40.0	max. 60
Grease	max. 20.0	–	max. 2.0	max. 2.0	36.0 – 40.0	max. 63

∗ ffa = free fatty acids.

Table 31. Analytical data for various fish oils [99]

	Herring oil	Sardine oil	Pilchard oil	Menhaden oil	South American oil
n_D^{40}	1.470 – 1.475	1.473 – 1.475	1.473 – 1.476	1.473 – 1.474	1.474 – 1.478
Saponification value	183 – 192	187 – 190	188 – 194	188 – 194	189 – 194
Iodine value	120 – 160	165 – 185	180 – 190	150 – 180	185 – 206
Nonsaponifiable matter, %	0.8 – 1.3	1.0 – 1.6	0.8 – 1.5	0.8 – 1.5	0.8 – 1.5

nical tallows used for soapmaking are shown in Table 30.

14.1.3. Mutton Tallow

Mutton tallow finds applications primarily in the technical sector. Superior grades are similar to corresponding beef tallows.

Saponification value	192 – 198
Iodine value	31 – 47
Nonsaponifiable matter	0.1 – 0.6 %
Density (40 °C)	0.896 – 0.898 g/cm^3
n_D^{40}	1.455 – 1.458
mp	44 – 55 °C

The main fatty acids are palmitic, stearic, and oleic acids (Tables 1 and 2). Up to 4 % of branched and odd-numbered fatty acids and traces of trans fatty acids are also present.

14.1.4. Horse, Goose, and Chicken Fat

Horse fat is softer than lard or rendered pork fat. The fatty acid composition (Tables 1, 2) is strongly influenced by diet. Branched and odd-numbered fatty acids occur up to about 2 %. The nonsaponifiable matter is <0.5 %, the melting point 29 – 40 °C. Horse fat is occasionally used in meat and sausage products in some countries.

Goose fat has a relatively low melting point, 25 – 35 °C; its consistency is often improved by addition of lard.

Chicken fat is also low-melting, *mp* 23 – 40 °C. The fatty acid composition (see Tables 1 and 2) is strongly influenced by diet.

Table 32. World production of oil fruits, oilseeds, oils, and fats according to countries (annual average, expressed as oil/fat in 1000 t)

	1934/38	1960	1970/71	1973/74	1980/81	1981/82	1982/83	1983/84	1984/85 [a]
Europe	3328 [b]		4790 [c]						
EEC		2300		4220	5615	5550	6065	6215	6080
EFTA		1135		710	855	850	930	910	930
Other West European countries		990	805	1015	835	650	1180	750	1285 [d]
Soviet Union	2000	3090	5335	6030	no data	4870	5340	5315	5105
Other East European countries	1231	1595	1975	2430	2920	2950	2975	3040	3290
Africa						3015 [e]			
West Africa	1132	2010	1845	1685	1380		2970	2800	2790
Other African countries	379	985	1420	1540					
America									
Argentina/Brazil	995	1525	2270	3080	5155	5100	5465	5935	6750
USA	3543	7470	11185	11945	14415	15620	16515	13535	14940
Other American countries	565	1480	3165	2605	3345	3210	3180	3340	3675
Asia									
Sri Lanka/Burma	216	240	320	215	265	290	330	365	375
India/Pakistan	2274	2725	3615	3620	3890	4750	4290	4875	4825
Indonesia/Malaysia	931	780	1425	1815	4510	5380	5355	5530	6800
Philippines	452	840	965	800	1585	1500	1345	1090	1140
Chinese Republic	3341	2905	3100	3070	4250	5685	6325	6160	6650
Other Asian countries	628	960	1460	1645	1105	1045	1045	1090	1120
Australia/Oceania	575	755	955	890	1100	1140	1115	1240	1270
World production	21590	31785	44630	46320	58880	62505	65485	63145	68155

[a] Preliminary data.
[b] $\sum$ EEC, EFTA, and other West European countries.
[c] $\sum$ EEC, EFTA.
[d] Including Spain.
[e] $\sum$ West Africa and other African countries.

14.2. Marine Oils [96]

Marine oils are the oils obtained from whales and fish. Fish oils have a high content of highly unsaturated (4 – 6 double bonds) C_{20}, C_{22}, and C_{24} fatty acids (cf. Tables 1 and 2). The oils from freshwater fish contain less C_{20} and C_{22} fatty acids and more oleic and linoleic acid than those from sea fish. These differences are caused primarily by differences in the composition of the feed. Only oils obtained from sea fish are commercially important.

14.2.1. Whale Oil

The whale population has been depleted to such an extent that whale oil is now completely overshadowed in commercial importance by the various fish oils. The Antarctic has long been the catching area. Whales are processed in modern factory ships (see also Section 5.3). The color of the oil can be very light, and the concentration of free fatty acids very low. Since 1913, when it was discovered that the oil can be converted into a stable edible fat by hydrogenation, many thousands of tons of hardened whale oil have been produced for margarine, cooking fats, and soapmaking.

Saponification value	185 – 205
Iodine value	110 – 135
Nonsaponifiable matter	<2 %
Slip point	22 °C
n_D^{65}	1.4554 – 1.4579

Whale oil has been graded according to free fatty acid content and color (apart from moisture and dirt):

Grade 0: pale yellow, max. 0.5 % ffa
Grade 1: pale yellow, max. 2.0 % ffa
Grade 2: amber yellow, max. 6.0 % ffa
Grade 3: pale brown, max. 15.0 % ffa

Grade 0 – 2 whale oils are hydrogenated for edible purposes. Prior to hydrogenation the oil

Table 33. World production of oil fruits, oilseeds, oils, and fats (annual average, expressed as oil/fat in 1000 t)

	1934/1938	1960	1970/71	1973/74	1980/81	1981/82	1982/83	1983/84	1984/85[a]
Vegetable oils									
Cottonseed	1453	2315	2450	2825	3035	3280	3025	3040	3770
Peanut[b]	1755	2760	3230	3965	2605	3455	2900	3190	3195
Corn[c]	95	185	325	390	665	695	745	785	805
Olive	950	1285	1605	1550	1980	1495	2025	1575	1785
Rapeseed	1273	1105	2220	2170	3675	4040	4905	4700	5500
Safflower	[d]	[d]	170	220	265	260	245	260	265
Sesame	563	485	755	720	615	730	620	700	695
Soybean	1263	3780	7700	9670	13275	13885	15255	13540	14675
Sunflower	435	1170	3335	4250	4695	5250	6000	5530	6165
Babassu	27	50	105	125					
Coconut	1633	2060	2375	2015	3150	2990	2860	2435	2640
Palm kernel	355	450	490	430	585	700	730	760	870
Palm	644	1100	1745	2260	4550	5410	5410	5645	6720
Linseed	1040	1025	1255	795	660	670	815	670	700
Castor	178	230	340	465	325	350	350	375	420
Tung	121	115	120	95	105	95	100	95	95
Others[e]	220	450	65	75	25	25	25	25	25
Animal fats									
Butterfat[f]	4160	4435	4950	5225	5450	5705	6200	6255	6135
Lard	2913	4370	4365	4180	4760	5050	5130	5185	5210
Tallow	1679	3455	5270	5320	6360	6200	6100	6110	6210
Fish oil	325	470	1170	970					
					1100[g]	1300[g]	1090[g]	1265[g]	1280[g]
Whale oil and sperm oil	507	490	190	140					

[a] Preliminary data.
[b] Until 1960 including 10 000 – 15 000 t teaseed.
[c] Including technical olive oil.
[d] Included under "others".
[e] Including oiticica, mowrah, niger, hempseed, perilla, and other commercially important oils.
[f] Including ghee fat.

[g] $\sum$ Fish oil, whale oil, and sperm oil.

is lye-neutralized and bleached. Whale oil can be hydrogenated into soft and hard fats. Soft fats with a melting point between 33 and 38 °C are produced by hydrogenation under isomerizing conditions with a partially inactivated nickel catalyst. Fats with a melting point between 40 and 45 °C are produced with relatively fresh catalyst. The crude hydrogenated fats are filtered to remove catalyst, treated with weak lye, bleached, and deodorized. In some cases, post-treatment can be restricted to treatment with citric acid – bleaching earth followed by distillative stripping. Hydrogenated whale oil is stable for about 6 months. The softer fats are used primarily for shortenings, the higher melting ones for preserves, dry soups, and special margarines.

The main components of sperm oil [*8002-24-2*] are esters of fatty acids and fatty alcohols (waxes). Sperm oil is a valuable raw material for high-quality cosmetic preparations. It is being replaced by jojoba oil, a seed oil from a shrub (*Simmondsia chinensis*) indigenous to the Sonoran desert and being cultivated in arid zones around the world. Its seeds contain 50 – 60 % of a liquid wax composed of esters of fatty acids and fatty alcohols [97].

14.2.2. Fish Oil [98]

The bulk of fish oils is produced in modern plants by cooking comminuted fish or fish offal, expelling in screw presses, and centrifugation; or by extraction of dried fish meal with hexane (see Section 5.3). Refining and hydrogenation yields fats of edible quality which are used in margarine and shortening blends. For historical reasons fish oils are not permitted as edible oils in the United States. The main crude oil origins are Japanese (iodine value ca. 180), U.S. (menhaden: iodine value 150 – 160), South American (iodine value

Table 34. World exports of oil fruits, oilseeds, oils, and fats according to countries (annual average, expressed as oil/fat in 1000 t)

	1934/38	1965	1970	1974	1980	1981	1982	1983
Europe								
Western Europe and Iceland		738	1020	1335	1693	1820	1658	2022
Eastern Europe and Soviet Union		315	783	740	532	458	548	466
Africa								
Former French Equatorial/West Africa	298	409	395	341	312	194	313	325
Nigeria	455∗	737	392	182	107	75	55	65
Former Portuguese Africa	62	93	111	96				
Sudan		115	100	105	85	91	93	52
Zaire	97	115	173	104	29	25	19	15
Other African countries	212	259	315	335	211	273	209	137
America								
Argentina/Uruguay	577	416	469	224	1426	1079	1243	1548
Canada		320	556	520	1071	1215	1012	1024
United States	100	3356	4479	5319	8246	7976	8534	7603
Other American countries	133	377	579	926	234	213	334	213
Asia								
India/Sri Lanka	589	142	105	123	75	138	166	163
Indonesia	529	260	317	304	610	201	323	449
Malaysia	132	164	435	883	2517	2827	3187	3346
Philippines	348	797	608	631	1002	1126	1080	1035
China	742	195	153	134	139	246	190	310
Other Asian countries		122	137	263	455	508	593	649
Australia/Oceania	363	540	640	530	755	749	790	804
Antarctic/Arctic	507	350	209	138	18	14	14	11
World exports	5829	8301	9820	13233	20819	20957	21589	21618

∗ Including Ghana, Gambia, Sierra Leone.

ca. 200), and Scandinavian (iodine value 130–140).

The oil content of fish is subject to seasonal variations and generally lies between 8 and 20 %. Typical properties of various fish oils are shown in Table 31 (for fatty acid composition see Tables 1 and 2). All fish oils contain small concentrations of branched and odd-numbered fatty acids.

Use. Fish oils are processed into hydrogenated fats analogously to whale oil [100]. However, due to their content of polyunsaturated fatty acids and sulfur compounds, hydrogenation requires more hydrogen and more catalyst. The melting points of hardened fish oils range from 31 to 45 °C. Organoleptic stability generally improves with decreasing iodine value of the starting oil and with increasing melting point of the hardened product. Hardened fish oils are used in shortenings and margarine blends. They have good creaming and cake-making properties but do not perform as well in frying.

Much of the oil of cod, halibut, and shark is stored in the liver. Cod-liver oil and halibut-liver oil are valued for pharmaceutical purposes because of their high content of vitamin A (1500–50 000 i.u./g) and vitamin D (40–200 i.u./g). Typical properties of cod-liver oil are as follows:

Saturated fatty acids, %	
C_{14}	6
C_{16}	6
C_{18}	8
Unsaturated fatty acids, %	
C_{14}	traces
C_{16}	20
C_{18}	29
C_{20}	26
C_{22}	10
Saponification value	180–197
Iodine value	150–175
Nonsaponifiable matter, %	ca. 1
n_D^{25}	1.481

15. Economic Aspects

Figures for the world production and exports of important fats and oils are shown in Tables 32, 33, 34, 35.

Table 35. World exports of oil fruits, oilseeds, oils, and fats (annual average, expressed as oil/fat in 1000 t)

	1934/38	1965	1970	1974	1980	1981	1982	1983
Vegetable oils								
Cottonseed	189	395	307	369	495	473	522	320
Peanut	826	967	817	677	678	519	641	669
Olive *	136	113	237	262	219	205	162	278
Rapeseed	51	272	481	670	1149	1364	1224	1274
Safflower		70	50					
Sesame	69	70	104					
Soybean	432	1802	2962	3903	6665	6757	7162	6770
Sunflower	26	263	700	629	1431	1535	1558	1832
Babassu	12	12	14					
Coconut	1057	1283	1087	923	1365	1455	1411	1376
Palm kernel	320	371	327	383	431	419	462	479
Palm	447	551	744	1327	2961	2818	3336	3567
Linseed	572	473	421	271	423	420	352	402
Castor	81	206	236	227	203	202	180	184
Tung	80	42	36					
Others	68	121	185	358	680	714	755	759
Animal fats								
Butterfat	500	499	641	704	775	740	671	608
Lard	173	290	347	372	432	439	389	366
Tallow	162	1225	1517	1575	2193	2181	2068	2025
Fish oil	121	445	554	500	701	702	682	698
Whale oil and sperm oil	507	350	209	138	18	14	14	11

* Including olive oil for industrial uses.

16. Toxicology and Occupational Health

With the exception of technical oils and fats such as castor and tung oil, oils and fats are classed as foods or food additives and do not entail toxicological or occupational health hazards.

17. References

General References

1. E. W. Eckey: *Vegetable Fats and Oils,* Reinhold Publ. Co., New York 1954.
2. H. P. Kaufmann: *Analyse der Fette und Fettprodukte,* **vols. I and II,** Springer Verlag, Berlin – Göttingen – Heidelberg 1958.
3. V. C. Mehlenbacher: *The Analysis of Fats and Oils,* The Garrard Press, Champaign, Ill., 1960.
4. J. Devine, P. N. Williams: *The Chemistry and Technology of Edible Oils and Fats,* Pergamon Press, Oxford 1961.
5. R. Lüde: *Die Raffination von Fetten und fetten ölen,* 2nd ed., Th. Steinkopff, Leipzig 1962.
6. A. J. C. Anderson: *Refining of Oils and Fats for Edible Purposes,* 2nd ed., Pergamon Press, London 1962.
7. W. Wachs: *öle und Fette,* **Part II,** Paul Parey, Berlin – Hamburg 1964.
8. "Fette und Lipoide," *Handbuch der Lebensmittelchemie,* **vol. IV,** Springer Verlag, Berlin – Heidelberg –New York 1965.
9. L. V. Cocks, C. V. van Rede: *Laboratory Handbook for Oil and Fat Analysis,* Academic Press, London –New York 1966.
10. F. D. Gunstone: *An Introduction to the Chemistry and Biochemistry of Fatty Acids and their Glycerides,* Chapman and Hall, London 1967.
11. H. A. Boekenoogen: *Analysis and Characterisation of Oils, Fats and Fat Products,* Interscience, New York 1968.
12. A. J. Vergroesen (ed.): *The Role of Fats in Human Nutrition,* Academic Press, London 1975.
13. H. Pardun: *Analysis of Edible Fats,* Paul Parey, Hamburg 1976.
14. D. Swern (ed.): *Bailey's Industrial Oil and Fat Products,* 4th ed., Wiley & Sons, New York 1979.
15. *Kirk-Othmer,* **9,** 795; **21,** 417; **23,** 717.
16. F. B. Padley, J. Podmore (eds.): *The Role of Fat in Human Nutrition,* VCH Verlagsgesellschaft, Weinheim 1985.

17. R. J. Hamilton, J. B. Rossell (eds.); *Analysis of Oils and Fats,* Elsevier Applied Science Publ., Barking,England 1986.

Specific References

18. J. Baltes: "Gewinnung und Verarbeitung von Nahrungsfetten," *Grundlagen und Fortschritte der Lebensmitteluntersuchung,* vol. 17, Paul Parey, Berlin, Hamburg 1975, p. 42.
19. J. B. M. Rattray, *JAOCS J. Am. Oil Chem. Soc.* **61** (1984) 1701.
20. F. D. Gunstone: *An Introduction to the Chemistry and Biochemistry of Fatty Acids and their Glycerides,* Chapman & Hall, London 1967.
21. M. Szuhaj, G. List: *Lecithins,* AOCS monograph, 1985.
22. A. J. Speek et al., *J. Food Sci.* **50** (1985) 121.
23. M. L. Meara: *Leatherhead Food RA, Scientific and Technical Survey no. 110,* July 1978.
24. K. van Putte, J. v. d. Enden, *Fette Seifen Anstrichm.* **76** (1974) 316.
25. E. S. Lutton, *J. Am. Oil Chem. Soc.* **49** (1972) 1.
26. B. G. Linsen, *Fette Seifen Anstrichm.* **73** (1971) 411, 753; **70** (1968) 8.
27. J. W. E. Coenen in I. Morton, D. N. Rhodes (eds.): *The Contribution of Chemistry to Food Supplies,* Butterworths, London 1974.
28. H. Wissebach, *Tenside* **3** (1966) 285.
29. A. K. Sen Gupta, H. Scharmann, *Fette Seifen Anstrichm.* **71** (1969) 873.
30. H. W. G Heynen et al., *Fette Seifen Anstrichm.* **74** (1972) 677.
31. E. W. Trautschold: *Die Lagerung von Sojabohnen unter Qualitätsaspekten,* DGF-paper, München 1970.
32. H. P. Kaufmann et al.: *Neuzeitliche Technologie der Fette und Fettprodukte,* Aschendorffsche Verlagsbuchhandlung, Münster 1956 – 1965.
33. I. E. Liener: *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York – London 1969.
34. "Oilseed Processing Symposium 1976," *J. Am. Oil Chem. Soc.* **54** (1977) 473A.
35. Deutsche Gesellschaft für Fettwissenschaft: *Gewinnung von Fetten und ölen aus pflanzlichen Rohstoffen durch Extraktion,* Industrieverlag von Hernhaussen KG, Hamburg 1978
36. S. Skipin, *Chem. Abstr.* **29** (1935) 7682. A. Carter et al., *J. Am. Oil Chem.* Soc. 51 (1974) 137.
37. P. König, *Fette Seifen Anstrichm.* **64** (1962) 23.
38. J. Furman et al., *J. Am. Oil Chem. Soc.* **36** (1959) 454.
39. G. Penk, *Proc. A.S.A. Symp. Soybean Process. 2nd* 1981.
40. W. H. Prokop, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 805.
41. S. M. Barlow, M. L. Windsor: *Techn. Bulletin no. 19,* Int. Ass. of Fish Meal Manufacturers, UK, Sept. 1984.
42. Th. Wieske, H.-U. Menz, *Fette Seifen Anstrichm.* **74** (1972) 133. V. K. Babayan, *J. Am. Oil Chem. Soc.* **51** (1974) 260.
43. O. L. Brekke in: *Handbook of Soy Oil Processing and Utilisation,* Amer. Soybean Ass., 1980.
44. H. J. Ringers, J. C. Segers, US 4 049 686, 1977.
45. A. K. Sen Gupta, US 4 062 882, 1977.
46. F. V. K. Young, *Proc. A.S.A. Symp. Soybean Process. 2nd* 1981.
47. A. N. Sagredos, K. Remse, *Fette Seifen Anstrichm.* **85** (1983) 185. P. Röttgermann, *Fette Seifen Anstrichm.* **85** (1983) 190.
48. Y. Hoffmann, *J. Am. Oil Chem. Soc.* **50** (1973) 260.
49. H. Pardun, *Fette Seifen Anstrichm.* **81** (1979) 297.
50. D. C. Tandy, W. J. Macpherson, *JAOCS J. Am. Oil Chem. Soc.* **61** (1984) 1253. A. Forster, A. J. Harper, *J. Food Sci.* **49** (1984) 23.
51. H. P. Kaufmann, D. Schmidt, *Fette Seifen* **47** (1940) 294.
52. M. Kock, *Proc. A.S.A. Symp. Soybean Process. 2nd* 1981.
53. G. Biernoth, H. E. Rost, *Chem. Ind.* 1967, 2002. G. Biernoth, *Fette Seifen Anstrichm.* **70** (1968) 217. A. N. Sagredos, D. Sinha-Roy, *Dtsch. Lebensm. Rundsch.* **75** (1979) 350.
54. D. B. Erskine, W. G. Schuliger, *Chem. Eng. Prog.* **67** (1971) 41.
55. A. Thomas, *Fette Seifen Anstrichm.* **84** (1982) 133.
56. J. B. Rossell et al., *Proc. A.S.A. Symp. Soybean Process. 2nd* 1981.
57. H. Stage: *Proc. A.S.A. Symp. Soybean Process. 2nd* 1981.
58. A. E. Rheineck, R. T. Holman et al.: *Progress in the Chemistry of Fats and Other Lipids,* **vol. 5,** Pergamon Press, New York1958.
59. H. J. Passino, *Ind. Eng. Chem.* **41** (1948) 280.
60. S. W. Gloyer, *J. Am. Oil Chem. Soc.* **26** (1949) 162.
61. Deutsche Gesellschaft für Fettwissenschaft: *Die Hydrierung von Fetten,* Industrieverlag von Herrnhaussen KG, Hamburg 1979.

62. H. B. W. Patterson: *Hydrogenation of Fats and Oils,* Elsevier Applied Science Publ., Barking, England, 1983.
63. F. A. Kummerow, *J. Am. Oil Chem. Soc.* **51** (1974) 255.
64. E. le Breton, P. le Marchal: *Riv. Ital. Sostanze Grasse* **47** (1970) 231. F. Camurati et al., *Riv. Ital. Sostanze Grasse* **47** (1970) 241.
65. P. v. d. Plank, *Fette Seifen Anstrichm.* **76** (1974) 337.
66. R. C. Hastert, *JAOCS J. Am. Oil Chem. Soc.* **58** (1981) 169.
67. B. Screenivasan, *J. Am. Oil Chem. Soc.* **55** (1978) 803.
68. G. Bradshaw, *Soap Sanit. Chem.* **18** (1941) no. 5, 23, 69.
69. A. J. Haighton, L. Vermaas, *Fette Seifen Anstrichm.* **74** (1972) 615.
70. J. R. Rossell: "Measurement of Rancidity in Oils and Fats," *Leatherhead Food RA, Scientific and Technical Survey no. 140,* Sept. 1983.
71. M. Arens, E. Kroll, *Fette Seifen Anstrichm.* **85** (1983) 307.
72. G. Guhr et al., *Fette Seifen Anstrichm.* **83** (1981) 373.
73. H. Karstens, *Fette Seifen Anstrichm.* **70** (1968) 400.
74. M. Unbehend et al., *Fette Seifen Anstrichm.* **75** (1973) 689.
75. K. G. Berger: *Recommended Practices for Storage and Transport of Edible Oils and Fats,* Palm Oil Research Institute of Malaysia, 1986.
76. "Proceedings of the World Conference, Kuala Lumpur, 1984, on Palm, Palm Kernel and Coconut Oils," *JAOCS J. Am. Oil Chem. Soc.* **62** (1985).
77. Y. Lozano et al., *Rev. Fr. Corps Gras* **32** (1985) 377.
78. F. Hirsinger, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 76. S. A. Graham, R. Kleiman, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 81.
79. D. Gegion, K. Staphylakis, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 1047.
80. R. Banerji et al., *Fette Seifen Anstrichm.* **86** (1984) 279.
81. A. E. Bailey: *Cottonseed and Cottonseed Products,* Interscience, New York 1948.
82. B. S. Kamel et al., *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 881.
83. W. Schuster, *Fette Seifen Anstrichm.* **74** (1972) 150.
84. R. Purdy, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 523.
85. R. Marquard et al., *Fette Seifen Anstrichm.* **80** (1978) 213.
86. W. Schuster, R. Marquard, *Fette Seifen Anstrichm.* **76** (1974) 207.
87. J. D. v. Mikusch, *Farbe + Lack* **7** (1952) 303.
88. T. Yaziciovglu et al., *Fette Seifen Anstrichm.* **79** (1977) 115.
89. G. W. Rohne, *Fette Seifen Anstrichm.* **86** (1984) 172.
90. B. J. F. Hudson et al., *J. Plant Foods* **5** (1983) 15.
91. L.-A. Appelquist, R. Ohlson: *Rapeseed,* Elsevier, Amsterdam 1972.
92. A. Thomas, *JAOCS J. Am. Oil Chem. Soc.* **59** (1982) 1.
93. M. Y. Raie et al., *Fette Seifen Anstrichm.* **87** (1985) 325.
94. O. Dahl: *Schlachtfette,* Fleischforschung und Praxis no. 10, Verlag der Rheinhessischen Druckwerkstätte, Alzey 1973.
95. E. S. Lutton et al., *J. Am. Oil Chem. Soc.* **39** (1962) 233.
96. "Fish Oils and Animal Fats," *Proc. Leatherhead Food RA,* Feb. 1986.
97. T. K. Miwa, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 377.
98. M. E. Stansby: *Fish Oils, Their Chemistry, Technology, Stability, Nutritional Properties and Uses,* The Avi Publishing Co., Westport, Conn., 1967.
99. F. V. K. Young: *The Chemical and Physical Properties of Crude Fish Oils,* Int. Ass. Fish Meal Manuf., Fish Oil Bulletin no. 18, June 1986.
100. F. V. K. Young: *The Refining and Hydrogenation of Fish Oils,* Int. Ass. Fish Meal Manuf., Fish Oil Bulletin no. 17, June 1986.

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Fatty Acids

Standard Article

David J. Anneken[1], Sabine Both[2], Ralf Christoph[3], Georg Fieg[4], Udo Steinberner[5], Alfred Westfechtel[6]

[1]Cognis Corporation, Cincinnati OH, United States of America, retired
[2]Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany
[3]Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany
[4]Technische Universität Hamburg-Harburg, Hamburg, Federal Republic of Germany
[5]Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany, retired
[6]Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a10_245.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Properties
2.1. Physical Properties
2.2. Chemical Properties
3. Production of Natural Fatty Acids
3.1. Resources and Raw Materials
3.2. Fat Splitting
3.2.1. Hydrolysis – Principles
3.2.2. Hydrolysis – Industrial Procedure
3.2.3. Hydrolysis – Enzymatic Processes
3.3. Separation of Fatty Acids
3.3.1. Distillation
3.3.2. Crystallization
3.4. Modification of Fatty Acids
3.4.1. Hydrogenation
3.4.2. Double Bond Isomerization
3.4.3. Dehydration
3.4.4. Dimerization
3.4.5. Ozonolysis
3.4.6. Thermal Decomposition
3.4.7. Bio-Oxidation of Fatty Acids
3.4.8. Enzymatic Esterification
4. Production of Synthetic Fatty Acids
4.1. Hydroformylation
4.2. Hydrocarboxylation
4.3. Other Commercial and Noncommercial Processes
5. Analysis
6. Storage and Transportation
7. Environmental Protection, Toxicology and Occupational Health
8. Uses
9. References

[Top of Page]

1. Introduction

Originally the term “fatty acids” was applied only to carboxylic acids separated from animal and vegetable fats. Today it includes all saturated and unsaturated aliphatic carboxylic acids with carbon chain lengths in the range of C_6 – C_{24}. Fatty acids obtained by splitting natural fats and oils are quantitatively far more important than synthetic fatty acids. Different processes for modification of natural fatty acids have been developed and are used to achieve special properties in the different fields of application. The principle production flow sheet in the oleochemistry chain is shown in Figure 1 [1].

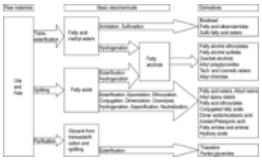

Figure 1. Oleochemical processes and products

The world production of oils and fats has grown from 60 × 10^6 t in 1980 to 124 × 10^6 t in 2003. Vegetable oils rose to a share of about 80 %, whereas animal fats decreased to 20 %. The majority of natural oils and fats is used for human nutrition (ca. 80 %) and animal feed (ca. 6 %). The remainder of ca. 14 % or 17 × 10^6 t is consumed by the chemical industry [2-5].

For technical applications, fatty acids play an important role since many years. World production capacity of fatty acids in 2000, not including tall oil fatty acids, was about 5.4×10^6 t and is estimated to grow to ca. $6.5 \times 10^6 – 8.1 \times 10^6$ t in 2005/2006. Expansion mainly took place and is planned in the Asian region. Based on the data for 2000 the contribution of Asia was about 50 %, Europe 30 %, and America 20 % [6], [7].

Although many fatty acids — often with unusual structures — occur naturally, with few exceptions (see Dimerization) only the straight-chain $C_6 – C_{22}$ carboxylic acids are of commercial importance. Moreover, in many fields of application, more or less broad cuts are employed, rather than chemically pure individual fatty acids.

[Top of Page]

2. Properties

2.1. Physical Properties

The face of chemical manufacturing is changing rapidly, shaped by such powerful forces as intensified global competition, increased stringency of safety and environmental regulations, and increased customer demand for consistent attainment of high product quality. In this context computer-aided process engineering (CAPE) plays a dominant role in the fast development of new or in revamping of existing processes. All process simulators use properties as intermediate variables. Their values are obtained from experiments or property models chosen to mimic the behavior of the process components and the operating conditions. For the process engineer, thermodynamic data form an essential basis for process development, design, and optimization. The thermodynamic data include thermal, transport, and caloric properties of components, such as density, heat capacity, and especially phase-equilibrium properties such as vapor – liquid phase equilibria. These data are necessary for the simulation of heat and mass transfer, design of thermal separation processes, and chemical reactors [8-10].

Especially important are physical data associated with a phase change from the solid to the liquid state or vice versa. The behavior on melting is highly specific for each fatty acid. For this reason, determination of the melting point is often used for identification and quality control. The melting point of a fatty acid depends on the number of carbon atoms, the degree of saturation, and the structure of the hydrocarbon chain.

With straight-chain fatty acids, the melting point increases alternatingly with increasing chain length, so that the melting points of fatty acids with an odd number of carbon atoms are somewhat lower than those of the adjacent even-numbered fatty acids. The behavior of fatty acids with chain lengths $C_6 – C_{22}$ is illustrated in Figure 2.

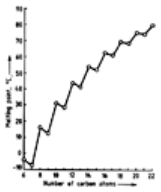

Figure 2. Melting point alternation of straight-chain saturated fatty acids

For unsaturated fatty acids, the relationship is more complex. In general, a decrease in the melting point is observed with increasing number of double bonds. However, naturally occurring *cis* isomers melt or solidify at lower temperatures than the corresponding *trans* isomers, as shown in Figure 3 for the C_{18} fatty acids. Other interesting parameters are the position of the double bond and, for polyunsaturated acids, the location of double bonds relative to each other. Fatty acids in which a conjugated arrangement (—CH ═CH —CH ═CH —) occurs have considerably higher melting points than nonconjugated acids. An example is given for eleostearic acid (Fig. 3).

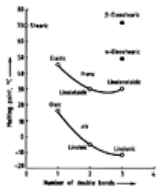

Figure 3. Effects of number of double bonds, cis – trans isomerism, and conjugation on melting points of C_{18} acids

A comprehensive survey of physical and chemical properties of straight-chain fatty acids is presented in Tables 1 and 2 [11-32].

Table 1. Physical properties of straight-chain saturated C_6–C_{30} fatty acids

Systematic name (trivial name)	CAS registry no.	Formula, $C_nH_{2n}O_2$	*M* kg/kmol	Neutralization value	*mp* °C	*bp*°C (*p*, kPa)[a]	Density ρ kg/m³ (*t*, °C)	Refractive index n_D (*t*, °C)	Molar heat capacity kJ kmol⁻¹ K⁻¹ (*t*, °C)	Molar heat of fusion kJ/mol	Molar heat of vaporization kJ/mol(*t*, °C)
Hexanoic (caproic)	[142-62-1]	$C_6H_{12}O_2$	116.16	483	–3	205.7 (101.3)	875.2 (80)	1.4170 (20)	261.41 (80)	15.20	63.01 (120)
Heptanoic (enanthic)	[111-14-8]	$C_7H_{14}O_2$	130.19	431	-7.5	223 (101.3)	867.0 (80)	1.4230 (20)	260.45 (15)	14.99	71.08 (120)
Octanoic (caprylic)	[124-07-2]	$C_8H_{16}O_2$	144.22	389	16.5	239.7 (101.3)	861.2 (80)	1.4280 (20)	332.72 (80)	21.39	71.19 (120)
Nonanoic (pelargonic)	[112-05-0]	$C_9H_{18}O_2$	158.24	354	12.4	255.6 (101.3)	855.6 (80)	1.4322 (20)	364.91 (80)	20.30	73.51 (120)
Decanoic (capric)	[334-48-5]	$C_{10}H_{20}O_2$	172.27	325	31.6	195.9 (10.13)	851.8 (80)	1.4269 (70)	390.09 (80)	28.05	74.78 (150)
Undecanoic	[112-37-8]	$C_{11}H_{22}O_2$	186.28	301	28.5	208.3 (10.13)	850.8 (80)	1.4202 (70)	404.56 (40)	25.07	78.72 (150)
Dodecanoic (lauric)	[143-07-7]	$C_{12}H_{24}O_2$	200.32	280	43.8	218.9 (10.13)	845.4 (80)	1.4230 (70)	454.66 (80)	36.03	85.52 (150)
Tridecanoic	[638-53-9]	$C_{13}H_{26}O_2$	214.35	261	41.4	230.9 (10.13)	845.8 (80)	1.4252 (70)	489.32 (50)	33.58	79.17 (200)

Tetradecanoic (myristic)	[544-63-8]	$C_{14}H_{28}O_2$	228.38	245		54.2	241.5 (10.13)	844 (80)	1.4273 (70)	522.43 (80)	45.05	83.21 (200)
Pentadecanoic	[1002-84-2]	$C_{15}H_{30}O_2$	242.41	231		52.1	251.4 (10.13)	842.3 (80)	1.4292 (70)	539.15 (60)	43.12	89.6 (200)
Hexadecanoic (palmitic)	[57-10-3]	$C_{16}H_{32}O_2$	256.43	218		62.8	261.7 (10.13)	839.7 (80)	1.4309 (70)	592.19 (80)	54.34	96.28 (200)
Heptadecanoic (margaric)	[506-12-7]	$C_{17}H_{34}O_2$	270.46	207		61.3	219.7 (1.33)	839.6 (80)	1.4324 (70)	634.13 (64)	51.16	91.53 (220)
Octadecanoic (stearic)	[57-11-4]	$C_{18}H_{36}O_2$	284.49	197		69.6	282.4 (10.13)	837.4 (80)	1.4337 (70)	660.60 (80)	62.51	93.13 (220)
Nonadecanoic	[646-30-0]	$C_{19}H_{38}O_2$	298.51	187		68.7	237.4 (1.33)	877.1 (24)	1.4512 (25)	686 (72)	57.80	
Eicosanoic (arachidic)	[506-30-9]	$C_{20}H_{40}O_2$	312.54	179		75.3	300.1 (10.13)	836.7 (100)	1.4250 (100)	717.78 (80)	70.92	61.76 (240)
Docosanoic (behenic)	[112-85-6]	$C_{22}H_{44}O_2$	340.60	164		80.0	263 (1.33)	822.1 (100)	1.4270 (100)	792.24 (80)	78.59	
Tetracosanoic (lignoceric)	[557-59-5]	$C_{24}H_{48}O_2$	368.65	152		84.2	257 (0.533)	820.7 (100)	1.4287 (100)	902.23 (98)	78.59	
Hexacosanoic (cerotic)	[506-46-7]	$C_{26}H_{52}O_2$	396.70	141		87.7	271 (0.533)	819.8 (100)	1.4301 (100)	947.17 (94)		
Octacosanoic (montanic)	[506-48-9]	$C_{28}H_{56}O_2$	424.76	132		90.9	285 (0.533)	819.1 (100)	1.4313 (100)			
Triacontanoic (melissic)	[506-50-3]	$C_{30}H_{60}O_2$	452.81	123		93.6	299 (0.533)		1.4323 (100)			

[a]kPa = 10 mbar

Table 2. Physical properties of straight-chain unsaturated C_{11}—C_{22} fatty acids

Systematic name (trivial name)	CAS registry no.	Formula	*M* kg/kmol	Neutralization value	Iodine value	*mp*, °C	*bp*, °C (*p*, kPa)	Density ρ ks/m^3 (*t*, °C)	Refractive index n_D (*t*, °C)
Unsaturated fatty acids with a single double bond									
10-Undecenoic	[544-64-9]	$C_{11}H_{20}O_2$	184.28	304	138	24.5	275 (101.3)	907.2 (24)	1.4486 (24)
cis-9-Tetradecenoic (myristoleic)	[373-49-9]	$C_{14}H_{26}O_2$	226.36	248	112	–4.0	183–186 (1.87)	903.2 (20)	1.4558 (20)
cis-9-Hexadecenoic (palmitoleic)	[593-39-5]	$C_{16}H_{30}O_2$	254.42	221	100	28–30	218–220 (2.0)	900.3 (15)	1.4586 (20)
cis-6-Octadecenoic (petroselinic)	[593-40-8]	$C_{18}H_{34}O_2$	282.47	199	90	28–30.1	237–238 (2.4)	879.4 (40)	1.4528 (40)
trans-6-Octadecenoic (petroselaidic)	[112-80-1]	$C_{18}H_{34}O_2$	282.47	199	90	54			
cis-9-Octadecenoic (oleic)	[112-79-8]	$C_{18}H_{34}O_2$	282.47	199	90	13.4 (α) 16.3 (β)	223 (1.33)	890.5 (20)	1.4582 (20)
trans-9-Octadecenoic (elaidic)	[29204-02-2]	$C_{18}H_{34}O_2$	282.47	199	90	45	225 (1.33)	873.4 (45)	1.4488 (50)
cis-9-Eicosenoic (gadoleic)	[112-86-7]	$C_{20}H_{38}O_2$	310.53	181	82	24–24.5		888.2 (25)	1.4597 (25)
cis-13-Docosenoic (erucic)	[112-38-9]	$C_{22}H_{42}O_2$	338.58	166	75	34.7	254.5 (1.33)	869.9 (55)	1.446 (57)
trans-13-Docosenoic (brassidic)	[506-33-2]	$C_{22}H_{42}O_2$	338.58	166	75	61.9	285 (4.0)	858.5 (57)	1.4472 (64)
Unsaturated fatty acids with two double bonds									
cis,cis-9,12-Octadecadienoic (linoleic)	[60-33-3]	$C_{18}H_{32}O_2$	280.45	200	181	–5.2	224 (1.33)	902.5 (20)	1.4699 (20)
trans,trans-9,12-Octadecadienoic (linolelaidic)	[506-21-8]	$C_{18}H_{32}O_2$	280.45	200	181	28–29	179–183 (0.11)		
Unsaturated fatty acids with three double bonds									

cis,cis,cis-9,12,15-Octadecatrienoic	[463-40-1]	$C_{18}H_{30}O_2$	278.44	202	274	–11	224.5 (1.33)	915.7 (20)	1.4800 (20)
trans,trans,trans-9,12,15,Octadecatrienoic (linolenic)	[28290-79-1]	$C_{18}H_{30}O_2$	278.44	202	274	30	358.9 (101.3)	915.9 (20)	1.4678 (50)
cis,trans,trans-9,11,13-Octadecatrienoic (a-eleostearic)	[4337-71-7]	$C_{18}H_{30}O_2$	278.44	202	274	49	235 (1.6)	898.0 (56)	1.5112 (50)
trans,trans,trans-9,11,13-Octadecatrienoic (b-eleostearic)	[544-73-0]	$C_{18}H_{30}O_2$	278.44	202	274	71–72		883.9 (80)	1.5002 (75)
Highly unsaturated fatty acids									
5,8,11,14-Eicosatetraenoic (arachidonic)	[7771-44-0]	$C_{20}H_{32}O_2$	304.48	184	334	–49.5	169–171 (0.02)	922 (20)	1.4870 (20)
4,8,12,15,19-Docosapentaenoic (clupanodonic)	[2548-85-8]	$C_{22}H_{34}O_2$	330.52	170	384	–78	236 (0.67)	935.6 (20)	1.5014 (20)

The behavior of fatty acid mixtures is characterized by a considerable lowering of the individual melting or solidification points. Most binary mixtures have phase diagrams with a eutectic point. Depending on the types of acids combined, three groups can be distinguished: mixtures of saturated fatty acids, mixtures of saturated and unsaturated fatty acids, and mixtures of unsaturated fatty acids. In Figures 4 and 5, typical solidification and melting point curves are illustrated. The values are plotted as a function of the concentration of the higher melting component [28], [33-35].

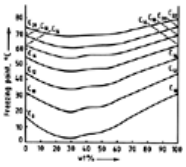

Figure 4. Freezing points of binary systems of saturated C_8 – C_{22} fatty acids

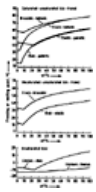

Figure 5. Freezing or melting points of binary systems of unsaturated fatty acids

The *liquid densities* of fatty acids are shown in Figure 6. The densities of C_6 – C_{20} fatty acids are all less than 1000 kg/m^3. The lower the molecular mass and the higher the degree of unsaturation of fatty acids, the higher is their density.

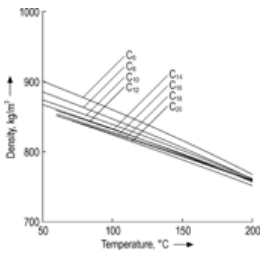

Figure 6. Densities of liquid straight-chain fatty acids as a function of temperature

Optical properties such as light refraction have long been used to identify liquids and establish their purity. The refractive index of pure fatty acids increases with increasing molecular mass, number of double bonds, and conjugation [30]. However, today measurement of the refractive index of fatty acids plays only a minor role in quality control. Modern gas chromatographs are the standard equipment for the determination of mixture compositions.

The *heat of fusion* is the energy consumption for phase transition from the solid crystal to the liquid state. Measured values of molar heat of fusion for straight-chain saturated fatty acids are listed in Table 1. Heats of fusion increase alternatingly with the length of the carbon chain. Fatty acids with an even number of carbon atoms have a higher heat of fusion than adjacent odd-numbered acids. Above a chain length of ten carbon atoms the heat of fusion can be determined by the following equations:

for even-numbered fatty acids (Table 1):

$$\Delta H_f = 4.2928 \cdot n - 14.99$$

for odd-numbered fatty acids (Table 1):

$$\Delta H_f = 4.152 \cdot n - 20.134$$

where ΔH_f is the heat of fusion (kJ/mol), and *n* the number of carbon atoms.

Table 3 lists heats of fusion for some C_{18} unsaturated fatty acids. They were derived from experimental data or calculated with the aid of theoretical relationships [28], [29], [36]. As expected, the values are lower than those for saturated fatty acids.

Table 3. Heat of fusion of some unsaturated C_{18} fatty acids

Component	Heat of fusion, kJ/mol
Oleic (9 c)	30.14
Elaidic (9 t)	38.94
Linoleic (9 c, 12 c)	51.50
Linolenic (9 c, 12 c, 15c)	39.36
α-Eleostearic (9 c, 11 t, 13 t)	43.96
β-Eleostearic (9 t, 11 t, 13 t)	55.27

The heat of vaporization and the related properties boiling point and vapor pressure are very important for the design and optimization of distillation columns by using simulation tools.

The *boiling point* and the molar latent heat of vaporization rise with increasing molecular mass (see Tables 1 and 2). The influence of the degree of saturation is of minor importance. As expected, fatty acids show an almost ideal equilibrium behavior, that means, fulfillment of Dalton's and Raoult's laws (especially under vacuum).

The *molar latent heat of vaporization* can be estimated by use of the integrated form of the Clausius – Clapeyron equation:

$$\Delta h_v = \frac{R T_1 T_2 \ln\left(\frac{p_1}{p_2}\right)}{T_1 - T_2}$$

where Δh_v is the molar latent heat of vaporization (J/kmol), R the universal gas constant (8314 J $kmol^{-1} \cdot K^{1}$), T_1, T_2 are temperatures (K), and p_1, p_2 pressures (Pa).

The vapor pressure curves for fatty acids of chain lengths C_6 to C_{20} are presented in Figure 7 [18], [22], [23].

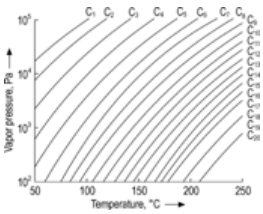

Figure 7. Vapour pressure of straight chain saturated fatty acids as a function of temperature

The following correlation relates vapor pressure and temperature of saturated, straight-chain fatty acids:

$$t = \frac{\sqrt{M} + 0.06075 \log^2 p + 1.30 \log p - 7.63}{\left(-3.54 \cdot 10^{-4} \log^2 p\right)\left(-3.91 \cdot 10^{-3} \log p\right) + 0.05191}$$

where t is the temperature (°C), p the vapor pressure (mbar), and M the molecular mass (kg/kmol).

The *heat of combustion* is of great practical and theoretical importance. It can be used to calculate the *heat of formation*, by means of which the enthalpy change in chemical reactions can be established. Heat of combustion and heat of formation increase with increasing number of carbon atoms. The data for unsaturated fatty acids, as far as they are known, are almost identical with those for the corresponding saturated fatty acids. For straight-chain saturated fatty acids the following linear correlations can be used (Tables 4 and 5):

$$-\Delta H_{comb} = 655.06 \cdot n - 447.71$$
$$\Delta H_{form} = 33.42 \cdot n + 368.33$$

where n is the number of carbon atoms (C_6 – C_{20}), ΔH_{comb} the heat of combustion (kJ/mol), and ΔH_{form} the heat of formation, kJ/mol

Table 4. Heat of combustion of straight-chain fatty acids (experimental values)[31, 33, 40]

Component – ΔH_{comb}, kJ/mol	
C_6	3 474
C_8	4 792
C_{10}	6 103
C_{12}	7 416
C_{14}	8 731
C_{16}	10 039
C_{18}	11 351
C_{20}	12 648
C_{22}	13 954
$C_{18:1\,c}$	11 168
$C_{18:1\,t}$	11 152
$C_{22:1\,c}$	13 799
$C_{22:1\,t}$	13 772

Table 5. Heat of formation of straight-chain fatty acids

Component	ΔH_{form}, kJ/mol
C_6	584.0
C_8	635.2
C_{10}	685.2
C_{12}	775.6
C_{14}	835.9
C_{16}	892.9
C_{18}	949.4
C_{20}	1064.0
$C_{18:1c}$	710.2
$C_{18:2c}$	603.1
$C_{18:3c}$	485.9
$C_{20:1c}$	763.9
$C_{22:1c}$	807.5

For straight-chain saturated fatty acids in the liquid phase, the *molar heat capacity* is shown in Figure 8. For solid fatty acids, the values are presented in Figure 9 [19]. The molar heat capacity is particularly important when considering heat-transfer problems (e.g., design of heat exchangers).

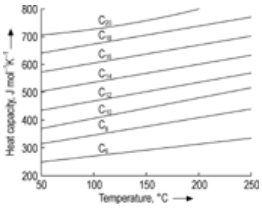

Figure 8. Molar heat capacity of straight-chain saturated fatty acids (liquid phase)

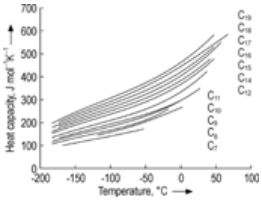

Figure 9. Molar heat capacity of straight-chain saturated fatty acids (solid phase)

Thermal conductivity and dynamic viscosity are important properties in the assessment of heat transfer and fluid dynamics. The thermal conductivity and dynamic viscosity of straight-chain saturated fatty acids as a function of temperature are shown in Figures 10 and 11, respectively.

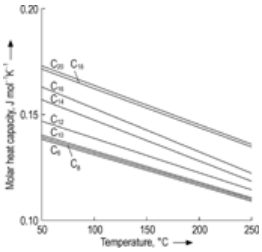

Figure 10. Thermal conductivity of straight-chain saturated fatty acids

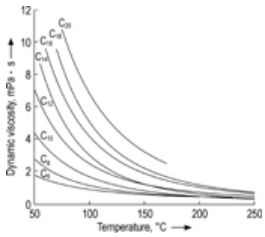

Figure 11. Dynamic viscosity of straight-chain saturated fatty acids as a function of temperature

At temperatures up to 100 °C, fatty acids and water have only low *mutual solubilities*, and the solubility of fatty acids in water is considerably lower than that of water in fatty acids. Both solubilities increase with increasing temperature and decrease with increasing chain length. Some data on the mutual solubilities of the binary water – saturated fatty acid system are given in Table 6. At high temperature and pressure, however, the solubility of water in fatty acids and in fatty acid – fat mixtures is considerable. At 250 °C and 5 MPa, for example, approximately 12 % of water dissolves in the fatty acid phase of a mixture of tallow fatty acids and water, while the value is 24 % for a mixture of coconut fatty acids. Complete miscibility with water is obtained at 320 and 290°C, respectively [37]. This fact must be considered in the high-pressure splitting of fats and oils (see Section Hydrolysis — Industrial Procedure).

Table 6. Mutual solubilities of saturated fatty acids with water

Component	Solubility of water in fatty acids, g water per 100 g acid (*t*, °C)	Solubility of fatty acid in water, g acid per 100 g water (*t*, °C)

C_6	4.96	(12.3)	0.968	(20)
	10.74	(46.3)	1.171	(60)
C_8	4.04	(14.4)	0.068	(20)
			0.113	(60)
C_{10}	3.22	(29.4)	0.015	(20)
			0.027	(60)
C_{12}	2.41	(42.7)	0.006	(20)
	2.93	(90.5)	0.009	(60)
C_{14}	1.73	(53.2)	0.002	(20)
			0.003	(60)
C_{16}	1.27	(61.8)	0.0007	(20)
			0.001	(60)
C_{18}	0.93	(68.7)	0.0003	(20)
	1.03	(92.4)	0.0005	(60)

Above their melting points, fatty acids are miscible in all proportions with many organic solvents such as hydrocarbons, ketones, esters, alcohols, and chlorinated or aromatic solvents. In the solid state, the solubility of saturated fatty acids depends on chain length and increases with increasing temperature. Unsaturated fatty acids, however, are completely miscible at room temperature and display only limited solubility at lower temperature [28], [33], [38], [39].

In Tables 7 and 8, the solubilities of saturated and unsaturated fatty acids in methanol are given as examples.

Table 7. Solubility of saturated fatty acids in methanol

Component	Solubility, g acid per 100 g methanol				
	0 °C	10 °C	20 °C	40 °C	60 °C
C_8	330	1300	∞	∞	∞
C_{10}	80	180	510	∞	∞
C_{12}	12.7	41.1	120	2250	∞
C_{14}	2.8	5.8	17.3	350	∞
C_{16}	0.8	1.3	3.7	77	4650
C_{18}			0.1	11.7	520

Table 8. Solubility of unsaturated fatty acids in methanol

Component	Solubility, g acid per 100 g methanol					
	–50 °C	–30 °C	–20 °C	–10 °C	0 °C	+ 10 °C
$C_{18:1c}$	0.1	0.86	4.02	31.6	250	1820
$C_{18:1t}$	0.01	0.06	0.18	0.48		
$C_{18:2c}$	3.3	48.1	233	1850		
$C_{22:1c}$	0.007	0.068	0.19	0.49		

The *surface tension* at the liquid – vapor interface and the interfacial tension in liquid – liquid systems are important properties for estimating the wetting behavior at solid surfaces and the coalescence behavior in liquid – liquid extraction. The surface tension of liquid fatty acids decreases with decreasing chain length and with increasing temperature. In Figure 12 and Table 9, some data are presented or listed for saturated fatty acids.

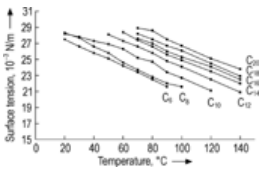

Figure 12. Surface tension of saturated fatty acids at 75 °C

Table 9. Interfacial tension of saturated fatty acids at 75 °C

Component	Interfacial tension*, mN/m
C_6	2.1
C_8	5.8
C_{10}	8.0
C_{12}	8.7
C_{14}	9.2

C_{16}	9.2
C_{18}	9.5

*System: Fatty acid - water

Dry fatty acids are poor conductors of electricity. The *specific electrical conductivity* is a function of chain length and temperature. Values vary from 0.6×10^{-9} to $83 \times 10^{-9}\ \Omega^{-1}\ m^{-1}$. The *relative dielectric constants* vary between 2 and 3, depending on the number of double bonds and the temperature [29].

The *flash point* and the *autoignition temperatures* of fatty acids are listed in Table 10 [41]. They rise significantly with increasing chain length.

Table 10. Average values for flash points and autoignition temperature of saturated and unsaturated fatty acids

Component	Flash point, °C	Autoignition temperature, °C
C_6	103	308
C_8	130	440
C_{10}	147,150	-
C_{12}	>160	-
C_{14}	>160	-
C_{16}	206	-
C_{18}	196	395
C_{22}(85%)	220	-
$C_{18:1}$(80%)	180	-
$C_{22:1}$(90%)	190	-

2.2. Chemical Properties

Fatty acids undergo the typical reactions of the carboxyl group (⟶ Carboxylic Acids, Aliphatic). Commercially important derivatives such as esters, amides, nitriles, and alkali or other metal salts are obtained by conventional methods (see also Uses). Hydrogenation of the carboxyl group to form fatty alcohols usually involves esterification and is discussed in detail under ⟶ Fatty Alcohols.

Unsaturated fatty acids are sensitive to atmospheric oxygen (see Storage and Transportation).

For industrially important reactions involving the double bond of unsaturated fatty acids (e.g., hardening, *cis – trans* isomerization, and dimerization) see Modification of Fatty Acids. Scientific approaches for the functionalization of fatty acids have been reviewed [42], [43]. Focus here is on the reaction of the double bonds of the fatty acids in order to achieve other or special properties.

[Top of Page]

3. Production of Natural Fatty Acids

3.1. Resources and Raw Materials

Fatty acids are widely distributed in nature as components of lipids such as oils and fats (esters of fatty acids with glycerol), waxes (esters of fatty acids with fatty alcohols), sterol esters, and other minor compounds.

Vegetable oils are obtained mainly from seeds and only to a small extent from fruit pulp; animal fats are obtained from fat deposits and from organs (⟶ Fats and Fatty Oils). Oils and fats are one of the most important renewable resources suitable to develop sustainable concepts. The term fats normally means materials of solid consistency, whereas liquid materials are referred to as oils. The definitions are indistinct, since the solid – liquid state of fats depends on temperature and many fats show a semisolid consistency at ambient temperature.

For the manufacturing of fatty acids on a commercial scale, only oils and fats which are regularly available in large quantities are used as raw materials. Much research in plant breeding has been done in the last few decades to adjust the fatty acid composition of vegetable oils to the needs of the food industry. The main commercial results were achieved with low erucic acid rapeseed oil (LEAR, Canola) and high oleic sunflower oil (HOSO) [44], [45]. By the use of transgenic approaches the fatty acid profile of soybean and rapeseed oil has been modified [46]. Though these developments have not been commercially used until now, they show that in future fatty acid compositions of vegetable oils may vary over broader ranges.

Fats and oils have characteristic fatty acid profiles [47], [48]. Tables 11 and 12 list the fatty acid composition of important fats and oils for fatty acid production.

Table 11. Fatty acid composition of important oils and fats (weight percent)* [continued in Tab. 12

Fatty acid(trivial name)	Carbon no.	No. of double bonds	Cotton seed oil	Peanut oil	Coconut oil	Linseed oil	Palm kernel oil	Olive oil	Rape-seed oil (C_{22}-low, LEAR, Canola)	Palm oil
Caproic acid	C_6	0			0 – 1		tr			
Caprylic acid	C_8	0			5 – 10		3 – 6			
Capric acid	C_{10}	0			5 – 10		3 – 5			
Lauric acid	C_{12}	0			45 – 53		40 – 52			
Myristic acid	C_{14}	0	0 – 2	0 – 1	15 – 21	tr	14 – 18	0 – 1	0 – 2	0 – 2
Palmitic acid	C_{16}	0	17 – 29	6 – 16	7 – 11	5 – 8	6 – 10	7 – 16	1 – 5	38 – 48
Stearic acid	C_{18}	0	1 – 4	1 – 7	2 – 4	2 – 4	1 – 4	1 – 3	1 – 3	3 – 6
Arachidic acid	C_{20}	0	0 – 1	1 – 3		tr		0 – 1	0 – 1	0 – 1
Behenic acid	C_{22}	0	tr	2 – 5					0 – 1	

Palmitoleic acid	C_{16}	1	0 – 2	0 – 1						
Oleic acid	C_{18}	1	13 – 44	36 – 72	6 – 8	15 – 25	9 – 16	65 – 85	50 – 66	38 – 44
Gadoleic acid	C_{20}	1	tr	0 – 2					1 – 3	
Erucic acid	C_{22}	1							0 – 5	
Linoleic acid	C_{18}	2	33 – 58	13 – 45	1 – 3	12 – 16	1 – 3	4 – 15	15 – 30	9 – 12
Linolenic acid	C_{18}	3		0 – 1		50 – 60		0 – 1	6 – 13	
Iodine value of oil			96 – 112	85 – 102	8 – 12	160 – 200	14 – 23	80 – 88	105 – 120	44 – 54
Saponification value of oil			190 – 198	188 – 195	250 – 264	188 – 195	245 – 255	188 – 196	185 – 198	194 – 206
Fatty acid titer, °C			30 – 36	26 – 32	20 – 24	19 – 21	20 – 28	17 – 26	5 – 15	40 – 47

*tr = traces; no value = no data in the literature or below precision of measurement.

Table 12. Fatty acid composition of important oils and fats, continued (weight percent)*

Fatty acid (trivial name)	Carbon no.	No. of double bonds	Castor oil	Rapeseed oil (C_{22}-rich)	Soybean oil	Sunflower oil (C_{18}:2-rich)	Sunflower oil HOSO(C_{18}:1-rich)	Sardine oil	Beef tallow
Caproic acid	C_6	0							
Caprylic acid	C_8	0							
Capric acid	C_{10}	0							
Lauric acid	C_{12}	0							
Myristic acid	C_{14}	0		0 – 1	tr	tr		4 – 6	1 – 6
Palmitic acid	C_{16}	0	2 – 3	2 – 5	7 – 12	3 – 10	3 – 5	9 – 11	20 – 37
Stearic acid	C_{18}	0	2 – 3	0 – 3	2 – 3	1 – 10	3 – 5	1 – 3	15 – 30
Arachidic acid	C_{20}	0		0 – 2	0 – 3	0 – 1			tr
Behenic acid	C_{22}	0		0 – 1	tr	0 – 1	1 – 2		
Palmitoleic acid	C_{16}	1		tr	tr	0 – 1	0 – 1	10 – 15	1 – 9
Oleic acid	C_{18}	1	4 – 9	11 – 24	20 – 30	18 – 28	82 – 86	15 – 25	20 – 50
Gadoleic acid	C_{20}	1		5 – 12	0 – 1	tr	tr		
Erucic acid	C_{22}	1		40 – 55		tr			
Ricinoleic acid	C_{18}	1	80 – 87						
Linoleic acid	C_{18}	2	2 – 7	11 – 29	45 – 58	60 – 68	4 – 7	3 – 8	0 – 5
Linolenic acid	C_{18}	3		6 – 13	4 – 10	0 – 1	0 – 1	1 – 3	0 – 3
Polyunsaturated fatty acid	C_{20}	2 – 6						15 – 30	tr
Polyunsaturated fatty acid	C_{22}	3 – 6						15 – 20	
Iodine value of oil			81 – 91	95 – 108	120 – 140	120 – 140	85 – 90	170 – 193	35 – 55
Saponification value of oil			174 – 186	170 – 180	190 – 195	186 – 194	185 – 194	189 – 193	190 – 200
Fatty acid titer, °C			13 – 16	20 – 24	16 – 20	8 – 20			39 – 43

*tr = traces; no value = no data in the literature, below precision of measurement.

Soap-stock fatty acids are byproducts in the processing of crude fats for human nutrition. The pulping of wood in the sulfate pulp process results in the digestion of the lipids composed of esters of triterpene alcohols that are present in the resin. The free fatty acids found as constituents of tall oil (⟶ Drying Oils and Related Products) represent another starting material for the commercial production of fatty acids.

Natural waxes, such as sperm oil or jojoba oil, are not significant sources of fatty acids.

3.2. Fat Splitting

Fats are fatty acid esters of glycerol and are also known as triglycerides. By splitting the glycerides, fatty acids and glycerol can be recovered (Fig. 13).

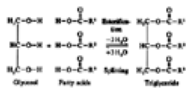

Figure 13. Formation and splitting of glycerides

The most important splitting agents are water (hydrolysis), methanol (methanolysis), caustic soda (saponification), and amines (aminolysis) (Fig. 14). Because of drawbacks in the subsequent purification of glycerol, saponification and aminolysis are of less commercial interest when full recovery of glycerol is important. Today, most soaps and fatty acid amides are manufactured via fatty acids or fatty acid methyl esters. Exceptions are ricinoleic acid and hydroxystearic acid, which are produced commercially by saponification of castor oil and hardened castor oil, respectively. In this way, the esterification reactions that occur with the hydroxyl groups of ricinoleic and hydroxystearic acids during hydrolysis are avoided.

Figure 14. Agents used to split triglycerides

3.2.1. *Hydrolysis — Principles*

The reactants in the hydrolysis of fats form a heterogeneous system made up of two liquid phases. The disperse aqueous phase consists of water and glycerol; the homogeneous lipid phase contains glycerides and fatty acids. The hydrolysis of fat to form glycerol and fatty acids occurs essentially as a homogeneous reaction by dissolved water in the lipid phase. The water – fat interfacial reaction is of minor significance [49]. The reaction takes place in several stages, via partial glycerides (diglycerides and monoglycerides), as shown in Figure 15.

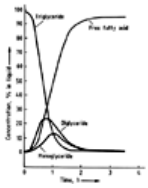

Figure 15. Typical course of hydrolysis reaction at 235 °C and 3.5 MPa using 1 part oil and 2 parts water

The total reaction can be summarized as follows (cf. Fig. 13):

$$\text{Triglyceride} + 3\,\text{Water} \rightleftharpoons 3\,\text{Fatty acid} + \text{Glycerol}$$

The reaction proceeds to a state of equilibrium. At higher temperature, the endothermic nature of fat hydrolysis results in displacement of the equilibrium in favor of the split products. The rate at which equilibrium is reached corresponds approximately to a first-order reaction [50]:

$$\frac{dc}{dt} = k \cdot c$$

where

k = rate constant
t = reaction time
c = concentration of triglyceride

However, at the beginning and the end, the reaction does not proceed according to this equation. This is because the necessary excess of water is present only after an induction period. Toward the end of the reaction, the glycerol concentration (which is not considered in the equation) has an increasing influence because of the reverse reaction.

Values of the rate constant k are highly dependent on temperature because of the increased solubility of water in the lipid phase [49], which leads to:

$$k = k_0 \cdot e^{-E/RT}$$

where

k_0 = statistical probability of reaction
E = energy of activation
T = absolute temperature
R = universal gas constant

An increase in temperature of 10 °C, for example, approximately doubles the value of k. Catalysts influence the rate constant k by changing the energy of activation, whereas k_0 is affected only slightly by temperature.

In the liquid two-phase system prevailing during hydrolysis of fats, the concentrations of the reactants depend primarily on the diffusion between the lipid and aqueous phases. Mass transfer can be influenced favorably by large, turbulent interfaces. During the entire reaction two mass transfer effects, which proceed in opposite directions, are of special importance for a high degree of conversion: (1) diffusion of water into the lipid phase and (2) removal of glycerol from the lipid phase into the aqueous phase by water extraction.

The overall hydrolysis reaction is a very slow process mainly controlled by temperature and duration. In comparison, mass transfer of glycerol and water across the phases is relatively fast [51]. LASCARAY [50]. summarized the complex reaction conditions in an empirical formula. He described the degree of splitting (hydrolysis) in the lipid phase as a function of the glycerol concentration in the aqueous phase:

$$H = 100 - 0.8 \cdot c_G$$
$$(c_G \text{ in the range from } 0 \text{ to } 20\ \%)$$

where

H = degree of splitting (% ffa = acid value/saponification. value × 100%)
c_G = glycerol concentration (wt % in the water phase)

Optimization of this process by chemical engineering largely determines the yields of fatty acid and glycerol. Temperature, duration, heat transfer, and hydrodynamic conditions, as well as the choice of catalyst, are the important parameters for approaching near-equilibrium conditions under the most favorable economic conditions, i.e., in the shortest time possible.

3.2.2. *Hydrolysis — Industrial Procedure*

Pretreatment. Crude commercial fats are sometimes purified to remove troublesome impurities, such as minerals, gums, soaps, and proteins, before they are subjected to splitting and following steps such as distillation or crystallization. The normal procedure is to heat the crude fat and then filter it or treat it with a small quantity of 0.3 – 1.0 % sulfuric acid. Animal fats from slaughterhouses contain varying amounts of dissolved packaging materials that consist mainly of polyhydrocarbons with differing rheological properties. These can best be separated from fats by filtration or centrifugation at ca. 70 °C with the use of filter aids. Also, settling of the fats over several days before splitting can be helpful for removing most of the impurities, which form deposits at the bottom of the tank. In this connection, one of the indicators for the quality of the fats and later yield of fatty acids is the share of total fatty material, which is defined as 100–M.I.U. (% moisture + % insolubles (dirt) + % unsaponifiables, e.g., polyethylene). Typical values for traded animal fats are a saponifiability of 97 % (0.5 % moisture, 0.5 %

dirt, 2 % unsaponifiables).

To suppress oxidation, degassing to remove atmospheric oxygen is recommended. All hydrolytic processes employ demineralized water as the splitting agent. The hardness of water reduces catalyst efficiency and increases the salt content of the glycerol obtained.

Catalysts. Up to 100 °C, fat hydrolysis is very slow. For palm oil with an initial content of 5 % free fatty acids (ffa), a degree of splitting of ca. 20 % is reached after 10 d at 100 °C [52]. TWITCHELL achieved a technical breakthrough in 1898 with an effective catalyst consisting of aromatic hydrocarbons, oleic acid, and concentrated sulfuric acid (Twitchell reagent). The original Twitchell batch process operates under atmospheric pressure at almost 100 °C in wooden tubs. The batch time is 12 – 24 h, and the degree of splitting without a change of water is about 80 – 85 % [53].

Lipases are specific biocatalysts for ester hydrolysis and exhibit optimum efficiency at temperature below 50 °C. Enzymatic splitting of fats is therefore of great interest today as a low-energy process combined with the advantage of low thermal stress to the heat-sensitive fats. Furthermore, enzymatic splitting can be applied to special fats which can be normally hydrolized only by saponification and neutralization (see Enzymatic Hydrolysis Processes). An improved fat-splitting process using lipase presplitting prior to pressure splitting is described by Henkel Corporation [54]. The partial hydrolysis achieved by the presplitting step mainly eliminates the induction period encountered during pressure splitting and leads to time savings or higher capacity of about 20 % compared to similar nonpresplit fats.

Acid catalysts are generally very effective for accelerating hydrolysis reactions. At higher temperature, however, material corrosion occurs. This problem does not arise when basic oxides are used as catalysts. Dibasic metal oxides have a higher activity than the more strongly alkaline monobasic metal oxides. In the literature, zinc oxide in its active form as zinc soap is described as the most effective accelerator for hydrolysis. The Zn^{2+} ion functions not so much as a Lewis acid, but rather as a phase-transfer catalyst which carries water in the form of hydrate into the lipid phase [55], [56].

Temperature. Without catalysts, the rate of fat hydrolysis is economical only above 210 °C (Table 13). The development of high-pressure techniques was the decisive step in making hydrolysis of fats a large-scale manufacturing process. Initially, batch operation in autoclaves at medium pressures of 0.6 – 1.2 MPa was employed.

Table 13. Influence of temperature on splitting time [57], [58]

t, °C	Time to reach equilibrium, min *		
	Beef tallow	Coconut oil	Peanut oil
225	156 (62)	158 (58)	156 (50)
240	82 (34)	85 (33)	85 (25)
260	47 (18)	46 (23)	53 (16)
280	34 (8)	33 (10)	33 (10)

* In parentheses: halftime of splitting.

Modern continuous splitting units operate at 210 – 260 °C and 1.9 – 6.0 MPa. The rate of splitting in high-pressure plants is sufficiently fast without any catalysts. With rising temperature and pressure, the mutual solubility of the two phases increases to a point where a homogeneous phase is formed. When coconut fat is used, this occurs at 293 °C and 8.0 MPa. For tallow, the values are 321 °C and 12.0 MPa [59].

Equipment. High pressure splitting can be carried out in autoclaves, which are employed mainly in batch operation. With a battery of autoclaves in parallel, a one-stage process can be carried out in cycles (Fig. 16). Autoclaves in series enable multistage processes to be operated batchwise or continuously with co- or countercurrent streams [60].

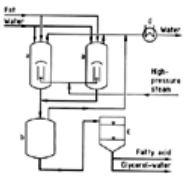

Figure 16. Batch fat splitting in two parallel autoclaves with direct heating

a) Splitting autoclaves; b) Flash tank; c) Separation tank; d) Condenser

For high throughput rates, spray columns operating countercurrently are used [60], [61-63]. The main feature of this system is the countercurrent movement of the two liquid phases as a result of their difference in density. Efficient mutual extraction is essential for fast hydrolysis of fat. This process, first described by H. MILLS [59], has been modified by several companies. Commercial countercurrent splitting towers with varying features (e.g., internal heat transfer between lipid and aqueous phases for energy conservation) were designed years ago, for example, by Colgate-Emery [64], Foster-Wheeler, Feld & Hahn [60], Crown Iron Works [65] (formerly Wurster & Sanger) and Lurgi [66]. An overview on plant and equipment suppliers is given in [67]. Based on the Crown Iron Works process [65] the main elements are shown in Figure 17.

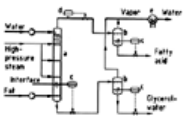

Figure 17. Continuous fat splitting (Foster – Wheeler)

a) Splitting column (250 – 260 °C, 5 – 6 MPa); b) Flash tank; c) Level control; d) Pressure control; e) Condenser

Fat and process water (mainly condensate from the concentration of glycerol water) are pumped into a splitting column by high-pressure metering pumps. Fat is fed in near the bottom of the column ca. 0.5 m below the interface; water is fed in at the top. The reaction zone, which lies between the upper and lower heat-exchange zones, is heated to 250 °C by direct feeding of high-pressure steam. The upper heat-exchange zone (water – fatty acid) is provided with baffles, whereas the lower heat-exchange zone (oil – glycerol/water) includes the interphase. In case of difficult fats (emulsion formation), feeding of the fat above the interface is recommended. A pressure regulator controls the discharge of fatty acids at the top, and an interface regulator automatically adjusts the flow of the glycerol-water mixture (12 – 18 % glycerol) from the bottom of the column.

Mass transfer between the lipid and aqueous phases, as described by LASCARAY [50], has a decisive influence on the final degree of fat splitting. Degrees of splitting >98 % are achieved today by applying principles of liquid – liquid extraction, especially at the top of the splitting tower (water against fatty acid).

Steam from expansion of the glycerol – water can be used as an energy source for the following evaporation of the glycerol – water, which comes out of the splitter with 15 – 20 % glycerol and which is generally concentrated to 88 % crude glycerol.

Continuous splitters have a consumption of high-pressure steam of about 200 kg steam per tonne of oil or lower. The ratio of the feeding rates of oil to water is normally about 2:1. In addition, the condensate of the direct-heating steam must be added to the water phase. The oil feeding rate per unit splitter volume (empty column) can be specified at about 200 kg h^{-1} m^{-3}, which corresponds to a residence time of the lipid phase of about 4 h.

Within the normal working range, the splitting degree (see Hydrolysis — Principles) shows a nearly linear dependence on the input flow rate of the oils. For the same column at constant splitting temperature, the plots for different oils and fats have different positions (see Figure 18). The splitting degree decreases with increasing input rate.

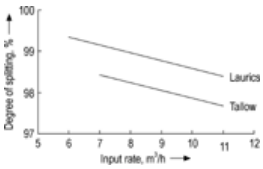

Figure 18. Splitting degrees of continuous splitting columns. Typical dependence from input rate.

The hydrolysis of fatty acid methyl esters into fatty acids and methanol is of special interest for generating short-chain fatty acids ($C_6 – C_{10}$) if the demand is higher than normally generated via oil splitting and fatty acid fractionation. The equipment and the conditions of normal direct steam splitting are not applicable. Therefore, formerly two-step processes were used, such as saponification and subsequent neutralization or transesterification with glycerol (glycerinolysis) and then normal splitting. Both are not very economical. Recent developments describe more efficient ways by direct splitting with the help of organic catalysts such as sulfonic acids [68] or metal soaps of organic acids [69]. The last process has been successfully used for some years on a industrial scale. Furthermore, enzymatic splitting of fatty acid methyl esters has been developed and can be recommended as suitable process (see Enzymatic Hydrolysis Processes).

3.2.3. *Enzymatic Hydrolysis Processes*

The high temperatures employed for steam splitting makes this process unsuitable for sensitive triglycerides, conjugated systems, hydroxylated fats and oils, and polyunsaturated oils with high iodine numbers (Table 14). There are also difficulties associated with alkaline hydrolysis, namely, high energy costs and the need to acidify the formed soaps in a second step to obtain the fatty acid products.

Table 14. Disadvantageous effects of chemical hydrolysis with sensitive triglycerides

Component	Disadvantage of chemical hydrolysis
Conjugated	thermal degradation, isomerization
Hydroxylated	dehydration, polymerization
Polyunsaturated	polymerization, coloration

A goal of technology in the 21st century is to become more environmentally friendly. For this reason, industrial enzyme catalysis technology based on green chemistry concepts is indispensable.

The enzymatic process is an alternative way to produce fatty acids, in which enzyme-catalyzed hydrolysis ⟶ Enzymes – Fat Splitting is carried out under milder conditions that restrict secondary reactions to a very low level; for example, coloration and the *trans* isomer content of the processed fatty acids are reduced due to the high enzyme selectivity.

Modification of starch and proteins by enzymatic processes is well established on an industrial scale. However, the use of enzymes to modify lipids, despite the significant research effort made in this field, is mostly still in the development stage.

The hydrolysis of oils and fats is catalyzed by lipase (EC 3.1.1.3), which breaks the ester bonds of acylglycerides. Lipases are produced by plants, animals, and microbes. However, plant and animal enzymes are not used commercially, while microbial enzymes (fungal, yeasts, molds and bacteria) are used extensively [70], [71].

Lipases have the particular characteristic of catalyzing the reaction only at the water/lipid interface. Thus, the surface area available for reaction is a crucial process efficiency factor, and an excess of water is essential. Both free enzymes in emulsion systems and immobilized enzymes have been extensively studied for performing the hydrolysis of fats and oils [72-77].

Lipases can be nonspecific enzymes that remove any fatty acid from the glycerol, e.g., lipases from *Candida cylindracea* [78]. Alternatively, they can be specific for positions in the triglyceride, e.g., 1,3-specific lipases from *Candida antarctica* preferentially release fatty acids from postions 1 and 3 of glycerol to give fatty acids and monoglycerides [79]. A lipase from *Geotrichum candidum* is specific for fatty acids with a *cis*-9 double bond [80].

Most microbial lipases for hydrolysis have a pH optimum of between 5 and 9 at the emulsion interface. The optimum temperature is 30 – 50 °C. Depending on enzyme concentration, stirring, water content, number of hydrolysis steps, and oil source, nearly complete hydrolysis can be performed in 8 – 72 h.

Prolonged reaction times can eventually yield complete hydrolysis of triacylglycerides to fatty acid and glycerol when a 1,3-specific lipase is used. To circumvent this issue, enzyme mixtures can be used in the process, as described in the enzymatic splitting process of castor oil to obtain high-quality ricinoleic acid [81]. The enzymatic hydrolysis can be conducted as a presplitting process that is combined with subsequent chemical hydrolysis to enhance the capacity of the overall fat splitting process [82].

As mentioned in Section Hydrolysis — Industrial Procedure methyl esters can be used instead of glycerides as further components to produce fatty acids. Most enzymes are inactivated in the presence of small amounts of methanol or they catalyze the esterification reaction to produce methyl esters. Therefore, the methanol content should be reduced by continuously removing methanol from the system, and water should be present in excess. There are two processes possible to obtain short-chain fatty acids in high quality: continuous removal of a methanol – water distillate and splitting and phase separation in several steps [83].

Lipase not only catalyzes the hydrolysis of glycerides, but can also catalyze the synthesis of esters or interesterification reactions by a reverse reaction (see Enzymatic Esterification and ⟶ Enzymes – Interesterification).

3.3. Separation of Fatty Acids

3.3.1. *Distillation*

Crude fatty acid mixtures from splitting contain a number of high-boiling impurities (e.g. partial glycerides, soaps, glycerol, sterols, phosphatides, pitch, color bodies) as well as low-boiling components (water, low molecular mass hydrocarbons, aldehydes, methyl ketones, odor bodies). These impurities must be removed by distillation to ensure the high quality of fatty acids demanded by customers. Processed crude fatty acids such as oleic acids obtained from crystallization (see Crystallization) and modified fatty acids (see Modification of Fatty Acids), as well as fatty acid bottom products of fractionation columns are also subjected to distillation.

On the other hand fatty acids are very reactive chemical components that can be damaged by oxidation, thermal decomposition, dehydration, polymerization, polycondensation, and other chemical reactions taking place in manufacturing plants. Typically, the reaction rate increases exponentially with increasing process temperature. The byproducts formed lead to a deterioration in color, color stability, odor, and other quality-control criteria. And from an economical point of view, byproducts lead to reduced product yields, which increase manufacturing costs.

Both mentioned points have a decisive influence on the engineering design of fatty acid manufacturing plants. The equipment development since ca. 1985 has concentrated on the chemical properties to produce fatty acids of high quality together with appropriate economic efficiency. The major requirement in the field of fatty acid distillation is to avoid product damage. This requires the liquid to be exposed to the lowest possible boiling temperature for the shortest period of time.

In the following two principle technologies for products of different quality and composition are considered in more detail: distillation and fractionation. Currently, both processes are only conducted continuously. Therefore, batch processes, which were often applied in the past, are not considered.

Due to their limited thermostability, fatty acids must be distilled or fractionated at low temperature under high vacuum. However, in conventional vacuum distillation systems the hydrostatic pressure of the liquid phase above the heating surfaces of the evaporator system partly destroys the effect of the vacuum system. As a solution, modern falling-film evaporators are widely used. Since ca. 1985 falling-film evaporators have practically replaced the forced-recirculation evaporators used until then. In most falling-film evaporators the heating medium is steam or heat-transfer fluids (e.g. Marlotherm, Dowtherm).

The falling-film evaporator is especially desirable from a product point of view, as it offers a short residence time. In falling-film evaporators, liquid and vapors flow downwards in parallel. The liquid to be partially evaporated is preheated to boiling temperature. An even thin film enters the heating tubes via a distribution device at the top of the evaporator, flows downward at boiling temperature, and is partially evaporated. This gravity-induced downward movement is increasingly augmented by the cocurrent vapor flow.

Falling-film evaporators can be operated with very low temperature differences between the heating media and the boiling liquid, and they also have very short product contact times, typically just a few seconds per pass. These characteristics make the falling-film evaporator particularly suitable for the partial evaporation of fatty acids as heat-sensitive products, and it is today the most frequently used type of evaporator in the field of fatty acid distillation. Falling-film evaporators must be designed very carefully for each operating condition; sufficient wetting of the heating surface by liquid is extremely important for trouble-free operation of the production plant. The proper design of the liquid-distribution system is critical to achieve full and even product wetting of the tubes. If the heating surfaces are not wetted sufficiently dry patches will occur; at worst, the heating tubes will be completely clogged. Thermal decomposition of fatty acids could occur.

Another advantage of falling-film evaporators is the low liquid holdup in this type of unit. As a result, the plant can be started up quickly.

The most important development in distillation and fractionation of fatty acid mixtures surely is the substitution of conventional trays and dumped packings by structured packings. The main advantage is that the pressure drop of a packed bed relative to the corresponding number of theoretical stages is much less than that of conventional trays. This pressure drop is less than 0.5 mbar per theoretical stage with a typical *F* factor of about 2 – 2.5 $Pa^{0.5}$. This is smaller by factor of 12 – 15 than that of conventional trays.

Another advantage of structured packing is the high separation efficiency. Typical established and experimentally investigated values for distillation of fatty acids in industrial plants are in the range 2 – 2.2 theoretical stages per meter packing (e.g., Mellapak 250, Sulzer AG) [84]). Latest developments in this field (Mellapak Plus, Sulzer AG; Montz-Pak Type M, Montz GmbH [85-88]) ensure higher *F* factors without simultaneous deterioration of the separation efficiency. This leads to distillation plants and fractionation columns that can be operated at lower pressure drop. The result is less thermal stress on the fatty acids. Therefore, the quantity of byproducts produced by thermal decomposition is reduced. This leads to better product quality and higher product yields.

Another benefit of structured packings is the extremely small liquid holdup in comparison to conventional trays. For this reason the residence time is lower than in the case of conventional trays. A typical holdup value of a structural packing is in the range of 2.5-3 % relative to the packing volume.

Additionally, structured packings permit a very wide range of liquid loads in the operating point of the production unit. For example, the minimum liquid load for Mellapak 250 (Sulzer AG) is in the range of 0.2 – 0.3 $m^3\ m^{-2}h^{-1}$, and the upper limit ranges up to 200 $m^3\ m^{-2}h^{-1}$.

Another gain of structured packings is the increased safety in scale-up in comparison to conventional trays. At present there is no longer need for experimental tests for the purpose of plant design. Modern flow-sheeting tools [89-91] guarantee a safe design.

To reduce the pressure drop in the distillation columns new types of condensers were developed and applied successfully in industrial practice. For example, special plate-type heat exchangers [92], [93] show extremely low pressure drop. This feature reduces the top pressure of a distillation plant and ensures the manufacture of high-quality fatty acids.

Distillation. Crude fatty acids from splitting contain a series of high-boiling impurities and lower boiling components. These impurities must be removed by distillation. In the following some industrial examples are presented and analyzed.

In a plant offered by Lurgi [94], the crude fatty acid is predried in the drier and degasser under vacuum (Fig. 19). It is then drawn into the distillation still, which is equipped with two sections of structured packing and internal chambers for separation of heavy ends. The chamber incorporates one or more heating elements and baffles. Propelling steam, which expands into a large volume in the vacuum space, circulates the liquid between heating elements and baffles by the principle of an air lift pump. The fatty acid is sprayed into a fine maze as it impinges on the baffles and heavy ends evaporate on the large liquid surface. The nonvaporized portion accumulates at the bottom and is circulated again, while the remaining undistilled residue is continuously discharged. A falling-film evaporator is used to reduce thermal stress on the product. An optional heat-recovery system can be applied in the top condenser.

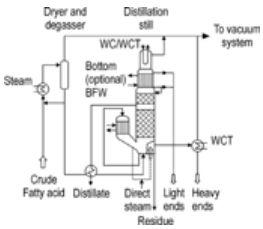

Figure 19. Fatty acid distillation plant (Lurgi AG)

At the top of the column fatty acid vapors are condensed in surface condensers. To remove low-volatility light-end components causing odor and color, the main distillate is taken out as a middle fraction, and the light ends as the top fraction. Heavy ends, primarily color carriers, can either be withdrawn separately or recycled directly for complete or partial redistillation.

The above-mentioned positive features of falling-film evaporators for the distillation of fatty acid mixtures are exploited in a manufacturing process of Cognis (Fig. 20). Crude fatty acid mixture is degassed and dried, preheated, and then evaporated in a falling-film evaporator. The vapors pass through the distillation column equipped with modern structured packings and condense stepwise in the upper part of the column in two condensers connected in series. The lower boiling components are removed from the second total condenser. The condensate from the first (partial) condenser is returned as undercooled reflux to the rectifying part of the distillation column. The main fraction is withdrawn above the stripping part of the distillation column. A second falling-film evaporator concentrates the residues in a side chamber of the still. With this arrangement, a strong fractionating effect is achieved, by which low-boiling odorous substances and high-boiling colored materials are removed.

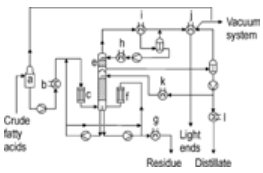

Figure 20. Continuous fatty acid distillation (Cognis Deutschland GmbH)

a) Dryer and degasser; b) Feed preheater; c) Falling-film evaporator; d) Distillation column; e) Internal condensation section; f) Falling-film evaporator (pitch stage); g) Residue cooler; h) Reflux cooler; i) Reflux condenser; j) Light-ends condenser; k) Cooler for main distillate reflux; l) Distillate cooler

Fractional Distillation. Fractional distillation is applied to obtain products of high purity. Typical raw materials are fatty acid mixtures with chain lengths of C_6 – C_{18} from the splitting of coconut or palm-kernel oil and unsaturated C_{18} – C_{22} fatty acids from the splitting of rapeseed oil. The purity required for the products

(referred to as fatty acid fractions in this section) is up to 99 %. Separation of fatty acids from resin acids in tall oil [95] is also accomplished by fractional distillation, while the fractionation of tallow, soybean, cottonseed, and linseed fatty acids is of lesser importance.

Separation usually takes place in vacuum distillation columns, column systems, or complex distillation arrangements with sophisticated internals (new generation of modern structured packings) to intensify mass and energy transport between the vapor and liquid phases and to minimize pressure drop in the columns. Before fractionation begins, the crude fatty acid must be degassed and dried. The final step is overhead distillation of the bottom product of the last column to improve the color.

In the bottom of the column, heat is supplied by means of falling-film evaporators. In the top of the column heat is removed by surface condensers. The vacuum system includes a multistage jet pump which permits a pressure of less than 2 – 3 mbar at the top of the column. Sealing the plant is particularly important because the presence of atmospheric oxygen results in rapid oxidation of fatty acids.

Column systems are being increasingly applied for the fractionation of fatty acid mixtures. The simplest case is a production plant with a system of several distillation columns connected in series (Fig. 21) [96]. The multicomponent mixture from splitting is first degassed and dried in a dryer. Then the mixture is fed to a system of distillation columns. Separation takes place in all columns in two product streams: top and bottom product streams.

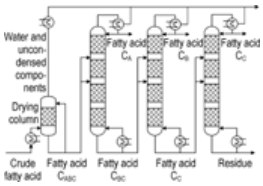

Figure 21. Sequence of fatty acid fractionation columns (Sulzer AG)

Figure 22 illustrates an improved alternative for the separation of crude coconut or palm kernel fatty acid [97]. Here, three products are simultaneously manufactured in a distillation column with a sidestream. The precut (C_6 – C_{10}) forms the top product. The bottom product contains the C_{16} – C_{18} components. The fraction C_{12} – C_{14} is produced as a sidestream. This sidestream is separated in two further columns to produce the C_{12} products and C_{14} products. The precut is specially treated and is used in a column system for the production of C_6, C_8, and C_{10} fatty acid fractions. For the separation of the C_{16} – C_{18} bottom product of the first column a sidestream column can be used again to obtain, e.g., a 16/70 % fraction as top, 18/92-94 % as sidestream, and a bottom fraction of C_{18} and higher.

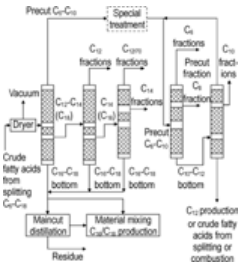

Figure 22. Fractionation of fatty acid mixture in a sequence of distillation columns with sidestreams

For all considered separation processes one distillation column can also be used several times, but this approach is not economical. The reason is repeated cooling and heating of the product streams, which raises energy costs and deteriorates the quality of the end products due to longer residence time. Furthermore, this technical solution demands many additional tanks for intermediate products.

A further application of a column with sidestream (Figure 23) permits production of the three products (C_{16} – C_{18}, C_{18} – C_{20}, and C_{22}) in one run [97]. The aim of the separation procedure is the production of erucic and behenic acid from the raw material after the splitting of erucic-rich rapeseed oil. In this case low top pressure, low pressure drop, and short residence time of the products in a distillation column due to the high temperature (ca. 270 °C) are important.

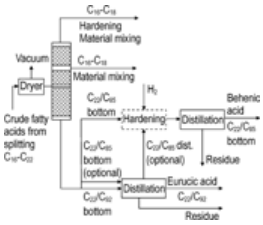

Figure 23. Manufacturing of erucic and behenic acids

The above-mentioned systems cover two or three product streams in distillation columns connected in series. Figure 24 shows two conventional arrangements for the separation of a typical four-component mixture (usually known as direct and indirect sequences). Process integration has proven to be very successful in reducing the energy costs for conventional distillation arrangements. One of the most significant nonconventional arrangements involves “thermal coupling”. Figure 25 shows a short survey of nonconventional arrangements which use thermal coupling. In thermal coupling part of the heat transfer necessary for separation is provided by direct contact via the product flows. Typically, the arrangement consists of a main distillation column surrounded by additional devices such as prefractionator, side stripper, and side rectifier.

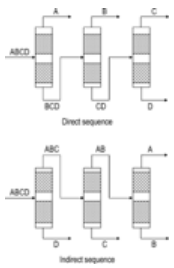

Figure 24. Conventional structures for the separation of a ternary system

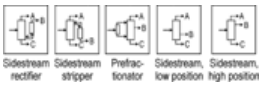

Figure 25. Complex distillation arrangements

The key advantage of the complex column over a sequence of simple columns is more compact equipment and substantial reduction of investment and manufacturing costs, instrumentation, man power, and maintenance. The application of these arrangements in industrial practice is changing rapidly because of recent advances in simulation, dynamics and control.

An interesting solution in the field of multicomponent distillation of oleochemical mixtures is a column with a vertical partition, proposed by Sulzer Chemtech for the separation of fatty acids [96]. This configuration is shown in Figure 26 and is often referred to as a dividing-wall column [98]. The advantage of this partitioned column is that a ternary mixture can be distilled into pure product streams with only one column, one reboiler, and one condenser. This reduces the cost of separation in comparison to sidestream columns or column systems [99-102]. It is expected that all above-mentioned complex arrangements will expand into the fractionation of fatty acid mixtures increasingly in the future because of clear economic advantages.

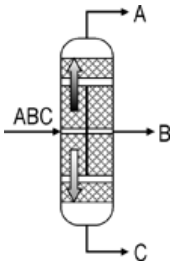

Figure 26. Dividing-wall column

Process Control. The increasing role of product quality control of distillation columns for manufacturing of fatty acids is significant. Driven not only by meeting fatty acid specifications and economic goals, distillation columns must also satisfy new customer demands. In many distillation columns product quality has been directly controlled by means of temperatures inside the column. However, since the feed to the distillation columns consists of multicomponent mixtures and they are often changed by operating conditions of the upstream units, rigid relations do not always exist between column temperatures and product qualities. In this context, it is difficult to control product quality by using column temperatures. In the modern oleochemical industry, there is a strong emphasis on the use of process analyzers to provide real-time process quality data for the purpose of maximum operating efficiency.

Fatty acids manufactured by using distillation columns with liquid or vapor sidestreams instead of two or more columns mostly require higher utility usage. Therefore, the application of these columns provides significant economic incentives of reduced capital and operating expenses. On the other hand, they also pose a more challenging control problem than standard distillation columns. Columns with a liquid or vapor sidestream have an additional degree of freedom and therefore an additional dimension of coupling. Therefore, these columns are generally more difficult to control than typical distillation units.

A newer process control concept for fatty acid qualities in distillation columns with a liquid sidestream has been presented [103]. The fatty acid composition is determined by an on-line gas chromatograph (Fig. 27). The determined product composition is then checked for consistency by an expert system embedded in the distributed control system (DCS). An additional supervision level is a fundamental element of the control system, implemented in the DCS and consisting of two basic parts: a diagnostic and an information level. The basic function guarantees an error-free operational sequence of the composition control by interception of all uncertain information due to the plant operator (errors in setpoints, product specifications) or the measurement procedure (errors of analyzing system). The next step of the control system covers composition control (Fig. 27). It was realized by using decentralized, time-discrete PI controllers [104]. A special user interface in the DCS for the plant operator was developed as a diagnosis, information and warning platform.

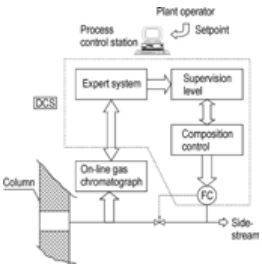

Figure 27. Process control system with on-line gas chromatograph

In summary, only the successful integration of all mentioned elements of the developed composition control system guarantees a stable operating point of the fatty acid fractionation plant.

3.3.2. *Crystallization*

Separation by crystallization means the separation of a fatty acid mixture into a higher melting fraction, consisting mainly of saturated acids, and a lower melting fraction, consisting mostly of unsaturated acids.

The prototype of this process is the separation of tallow fatty acids into stearic acid (stearin) and oleic acid (olein). The stearin obtained from this separation has an iodine value of less than 20 and is normally hardened subsequently to an iodine value of 1. The crude olein obtained has a cloud point below 3 °C. Apart from tallow fatty acid, distilled or crude fatty acids from the splitting of, bonefat, soybean, sunflower, palm, rapeseed oil or the C_{16} – C_{18} fraction of palm kernel fatty acid can be employed.

When molten fatty acids such as tallow fatty acid are cooled, the higher melting, saturated constituents crystallize first, while the appreciably lower melting, unsaturated constituents remain liquid over a defined temperature range. Because of the formation of mixed crystals, the solid phase still contains residual unsaturated fatty acids. Conversely, residual saturated fatty acids are found in the liquid phase because of their solubility in unsaturated fatty acids. Since various chain lengths and degrees of saturation are present in natural fatty acid mixtures, a theoretical discussion of this solid – liquid equilibrium is extremely difficult. From the experimentally determined phase diagram for tallow fatty acid, the composition and quantitative ratio of the two phases “stearin” and “olein” can be seen in relationship to temperature and to the composition of the starting material (Fig. 28).

Figure 28. Phase diagram of tallow fatty acids

A = saturated fatty acids; B = unsaturated fatty acids [105]

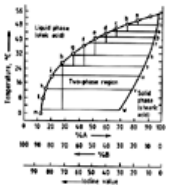

Olein must be removed at a temperature corresponding to the desired olein content. However, because the olein content of stearin increases with decreasing temperature, separation at higher temperature of a stearin with a low olein content may be advisable. Separation of the liquid phase can then be carried out in a second step at a lower temperature.

Fractional Crystallization without Additives. In the “panning and pressing process, ”the molten tallow fatty acid is cooled slowly to ambient temperature in flat aluminum pans. Fatty acid cakes are wrapped in cloths and separated from the liquid fraction in hydraulic presses. The liquid from the first pressing is cooled to 5 – 10 °C, and more stearin crystals are removed, which results in an olein with a cloud point of 5 – 10 °C. Today, the pressing process is no longer used because it is labor-intensive and separation is poor [31].

The fractional crystallization developed by Sulzer [106] is a solvent-free batch process divided into the three principal steps of 1) crystallization, 2) partial melting (sweating), and 3) melting. Depending on the purity of the final product the simpler process of static crystallization can be applied or, in the case of higher qualities, falling-film crystallization. The static crystallizer uses cooling plates immersed in a stagnant melt, whereas the falling-film crystallizer is equipped with a tube bundle that is cooled from outside, and the processed melt is pumped in a loop through the tubes. Both processes can be combined with each other or with other processes such as distillation. Typical applications are monomer purification and isomer separation to obtain purified products such as acrylic acid or paraffin waxes. In the case of fatty acids the static unit is used for enrichment of unsaturated acids based on rapeseed, soybean, or sunflower oil.

Fractional Crystallization with Addition of a Wetting Agent. Liquid and solid phases can be separated more easily and more completely if an aqueous solution of a wetting agent is added to the crystal paste. The stearin crystals are wetted by the wetting agent and form an aqueous suspension which is separated from the olein phase by centrifugation.

This principle was developed by Henkel as a continuous, two-stage hydrophilization process [107]. The mode of operation can be seen from the flow diagram (Fig. 29). Lurgi built in the past one- and two-stage plants based on this principle [108], [109].

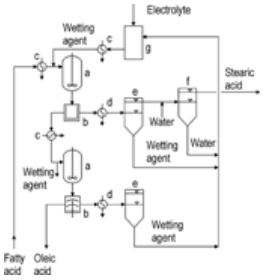

Figure 29. Flow diagram of two-stage hydrophilization process

a) Stirred vessel (wetting); b) Centrifuge; c) Cooler; d) Heater; Settling tank (separation of wetting solution); f) Settling tank (washing stage)

Stearin crystals form when molten tallow fatty acid is cooled to ca. 20 °C. The aqueous wetting agent solution containing a hydrophilization component such as sodium lauryl sulfate and an electrolyte such as magnesium sulfate is added to the mixture of stearin crystals and liquid olein, and the crystals are thus wetted. The stearin – water suspension is separated from olein in a centrifuge. Olein from the first stage is cooled in a second step to the desired cloud point, treated again with a solution of wetting agent, and separated in a centrifuge. The cold stearin suspension from the first stage is heated to above the melting point of the stearic acid and separated from the wetting agent solution. Before additional processing (e.g., final hardening), the stearin is washed and dried under vacuum. The stearin suspension from the second stage is recycled by adding it to the starting material for the first stage. The solution of wetting agent and electrolyte is also recycled; only a small amount of fresh material must be added continuously to maintain optimal concentration. Cooling is normally carried out indirectly by passing chilled brine through the outer jacket of a scraped surface crystallizer. Otherwise, the mixture can be cooled by evaporating part of the water under vacuum.

The hydrophilization is sometimes referred to as the Henkel [31], the Lipofrac [110], or the Lanza (Unilever) [111] process.

Fractional Crystallization from Solvents. Another method of crystallization is based upon the fact that saturated fatty acids are less soluble than their unsaturated counterparts in organic solvents such as methanol, hexane, or acetone. The first process of this type was put into operation by Emery (Emersol plant, Fig. 30) [31], [112].

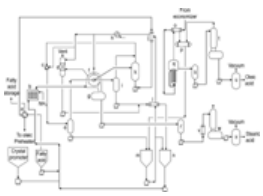

Figure 30. Flow diagram of Emersol® crystallization plant (Cognis Corporation)

(a)economizer (b)crystallizers (c)vacuum pump (d)vent scrubber (e)seal liquid receiver (f)filter (g)cake melt tank (h)wash chiller (i)wash receiver (j)filtrate economizer (k)filtrate receiver (l)wash economizer (m)wash alcohol receiver (n)process alcohol (o)oleic preheater (p)alcohol condenser (q)oleic evaporator (r) stearic surge tank (s)oleic surge tank (t)oleic still (u)oleic drier (v)preheater (w)stearic still (x)stearic drier

A ca. 30 % solution of, for example, tallow fatty acid in 90 – 95 % methanol is cooled continuously to –12 °C in a tubular crystallizer equipped with scrapers. Optionally, a crystal promoter or modifier may be added to the process to improve the separation efficiency [113]. From the resulting crystal paste, the mother liquor containing mainly unsaturated fatty acids is sucked through a rotating vacuum filter. The filter cake, which contains saturated fatty acids, is washed free of the mother liquor and melted. The solvent is recovered from both the filter cake and the mother liquor by distillation. The stearin obtained still contains ca. 5 – 15 % oleic acid and has an iodine value of ca. 5 – 15. Depending on the filtration temperature, the olein has a cloud point from 1 to 8 °C [114].

The corresponding Armour process applies acetone as a solvent [115], [116]. Indirect cooling can be replaced by direct cooling, e.g., by partial evaporation of the solvent at reduced pressure [117], [118].

The Bernardini [119] and Rau [120] processes also operate on the principle of fractional crystallization from solvents.

Common to all these processes is the need for extensive safety measures (explosion prevention) and for environmental protection with respect to the emission of solvents. Recovery of solvents requires additional equipment and energy.

Adductive Crystallization with Urea. Adductive crystallization with urea, that is, formation of urea inclusion compounds, is a well-known technique for fractionation of fatty acids. The method is useful to separate saturated fatty acids from unsaturated and to isolate selectively a variety of polyunsaturated fatty acids (PUFAs) from lipid mixtures, e.g., to increase the nutritional quality of the fatty acids.

For example 97 – 99 % oleic acid containing only 0.2 % polyunsaturated material has been obtained from olive oil fatty acid [121]. The fatty acids in soybean, fish, wheat germ, and grape-seed oils have also been separated by this method [122].

The main selectivities of this procedure are [123]:

- Enhanced selectivity against inclusion with increasing number of double bonds per fatty acid molecule
- Preferences for fatty acids of longer chain length
- Preference for *trans* versus *cis* double bonds in inclusion
- Sensitivity for double bond position

Processes for fatty acid fractionation [123], [124-127] are listed in Table 15.

Table 15. Fractionation of fatty acids by formation of urea inclusion compounds

Fatty acid source	Application
Fish oil	isolation of EPA* and DHA*
Linseed oil	isolation of ALA*
Rapeseed oil	removal of saturated fatty acid
Borage oil	isolation of GLA*
CLA reaction mixture	isolation of 18:2 (9c, 11 t) and 18:2 (10t, 12c)

*EPA: eicosapentenioc acid; DHA: docosahexenoic acid; ALA: α-linolenic acid; GLA: γ-linolenic acid CLA: conjugated linoleic acid.

The formation of urea inclusion compounds is an exothermic and equilibrium reaction. The equilibrium and stability constants of urea complexes for different fatty acids are listed in Table 16 [128].

Table 16. Equilibrium and stability constants of urea complexes for different fatty acids

Component	Equilibrium constant *K*	Stability constant 1/*K*
Caprylic acid	0.07	14
Caprinic acid	0.0185	54
Lauric acid	0.0063	159
Myristic acid	0.0029	345
Palmitic acid	0.0017	588
Stearic acid	0.0002	5000
Oleic acid	0.0024	476
Linoleic acid	0.019	53

The differences in the constants show that fractionation is possible under certain conditions.

A scheme of a urea complexation and crystallization process for separating saturated from unsaturated fatty acids is demonstrated in Figure 31.

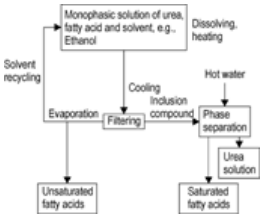

Figure 31. Process scheme for separation of saturated and unsaturated fatty acids with urea complexation [128]

Other Separation Methods.
Adsorption. For the separation of oleic and linoleic acids, UOP recommends an acid-resistant, hydrophobic, non-zeolite molecular sieve (Silicalite) [129]. Oleic acid is first selectively adsorbed and later displaced with the aid of a solvent. A nonionic, hydrophobic, spatially cross-linked styrene polymer has been used for the same purpose [130].

Extraction. Laboratory separation of stearic and oleic acids has been achieved by high-pressure extraction with a supercritical gas at 10 MPa and ca. 130 °C [131]. Benzene was used as solvent and ethylene as extractant.

The solubility of oleic, stearic, and behenic acids in supercritical carbon dioxide has been investigated at 40 and 60 °C and 10 – 25 MPa. A significant enrichment of oleic acid compared to stearic acid is observed only above 20 MPa [132]. Enrichment depends on the rate of material transfer and on the relative rates of movement of the phases, so that a separation can be expected after brief extraction [133].

High-purity oleic, linoleic, and linolenic acids have been prepared by means of liquid – liquid extraction, with solvent pairs such as furfural – hexane and ethanol – petrol ether [134].

Separation in the Presence of Acid Soaps. In the system soap – fatty acid – water, micelles known as “acid soaps ”form at certain concentrations [135]. Acid soaps of saturated acids crystallize from aqueous solution, whereas those of unsaturated acids remain in solution. Thus two fractions are obtained with differing degrees of saturation [136].

3.4. Modification of Fatty Acids

3.4.1. *Hydrogenation*

Catalytic hydrogenation of unsaturated fatty acids can be carried out in such a way that the carboxyl group is retained and saturated fatty acids are formed [137] by addition of hydrogen. Because of the increase in the degree of saturation, the melting point of the fatty acids increases. This process is generally known as hardening. Hardening of fatty acids usually means complete saturation of the carbon – carbon double bonds down to an iodine value less than 1. In this way, light-colored fatty acids are obtained which are thermally more stable and resistant to oxidation. Starting materials are, for example, split fatty acids from tallow, the stearin fraction from the olein – stearin separation process, fatty acid fractions from rapeseed oil (erucic acid), and lauric acid, and acid oils from the refining of edible oils. A short overview on fatty acid hydrogenation is given in [138].

Chemistry and Reaction Conditions. The hardening reaction, i.e., the addition of hydrogen to the double bonds of unsaturated fatty acids, is exothermic. Like other types of heterogeneous catalysis, the reaction proceeds in stages [49]:

$$\text{Linolenic acid}(C_{18:3}) \xrightarrow[k_3]{H_2} \text{Linoleic acid}(C_{18:2}) \xrightarrow[k_2]{H_2} \text{Oleic acid}(C_{18:1}) \xrightarrow[k_1]{H_2} \text{Stearic acid}(C_{18})$$

Highly unsaturated fatty acids are hydrogenated more rapidly than monounsaturated fatty acids ($k_3 > k_2 > k_1$). Hydrogenation is accompanied by isomerization processes in which double bonds migrate (conjugation) and geometrical changes occur (*cis – trans* isomerism) [139]. Because fatty acid hardening is carried out predominantly to achieve complete saturation, the selectivity of the reaction (i.e., the ratio k_3: k_2 (2 – 3) or k_2: k_1 (10 – 50) ⟶ Fats and Fatty Oils) is only of importance in some special cases [140]. It has more relevance in oil hardening, e.g., of soybean or rapeseed oil to achieve higher concentrations of monounsaturated acids by selective hydrogenation of the linolenic and linoleic content.

To some extent, the fatty acid reacts with the nickel catalyst, particularly in the absence of hydrogen and in the presence of water, to form nickel soaps [141], [142]:

$$Ni + 2RCOOH \rightleftharpoons Ni(OOCR)_2 + H_2 \quad \text{and} \quad NiO + 2RCOOH \rightleftharpoons Ni(OOCR)_2 + H_2O$$

Because nickel soaps are catalytically inactive, this reaction blocks the active surface of the catalyst. To reduce soap formation, the reaction with hydrogen is carried out at higher pressure [143], generally 2.0 – 3.0 MPa, and the fatty acids are predried. Despite these precautions, catalyst consumption is appreciably higher in the hardening of fatty acids than in hydrogenation of the corresponding triglycerides. The rate of hydrogenation increases initially with increasing temperature, then passes through a maximum in the range of 180 – 210 °C depending on the fatty acid used, and finally decreases [144]. Starting temperature of the reaction is in practice about 150 °C. Hydrogen transport through the liquid to the surface of the catalyst is dependent on the presence of large interfaces between the phases. Hardening is therefore carried out with intensive mixing by means of stirrers or circulation pumps. Newer developments show that sound waves also are suitable to accelerate the reaction [145]. The influence of various reaction parameters on the reaction rate of fatty acid hardening [146] is illustrated in Figure 32.

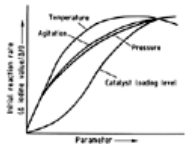

Figure 32. Effect of operating variables on the reaction rate of fatty acid hydrogenation

The average practical reaction rate corresponds to a reduction of the iodine value by 1 – 2 units per minute. A reduction of 1 in the iodine value corresponds to the consumption of ca. 1 m^3 (STP) of hydrogen per cubic meter of fatty acid and results in a temperature increase of ca. 1.5 °C.

It was further found that formation of nickel soap in the absence of hydrogen depends on temperature and is relatively fast, especially at temperatures normally used for hydrogenation [147]. This has an impact on the remaining nickel content of the hydrogenated fatty acid and subsequently of the distillation residue.

Catalysts. The hardening of fatty acids on a commercial scale is mostly carried out with nickel-based heterogeneous catalysts in slurry form. Dry reduced carrier catalysts (nickel on silica) with high activity and a good resistance to poisoning are used. The active surface of this type of catalyst has been increased to over 100 m^2/g. The pore diameter of the carrier must be large enough (ca. 3 nm) to allow transport of fatty acid molecules to the nickel crystallites. The nickel carrier catalyst is pyrophoric in its activated form. For safer transport and handling, the catalyst is passivated temporarily by a coating of hydrogenated fat and normally sold as pellets in drums. A typical catalyst has the following composition: 20 – 25 % nickel, 10 – 15 % silica, and 60 – 65 % fat. Catalyst consumption depends on the properties of the feedstock and the intended reduction in iodine number, as well as on the hardening conditions. Therefore prior laboratory trials are useful. A gross catalyst consumption of 0.2 – 0.4 % relative to the feedstock is the normal range. Detailed selectivity data of nickel catalysts are given in [148].

Higher selectivity is obtained when heterogeneous copper or palladium catalysts are applied [149]. Homogeneous catalysis by a liquid – liquid two-phase technique with a solvent-stabilized palladium colloid is described in [150].

A number of other catalysts have been described [151]. Acid- and poison-resistant catalysts in pellet form have been investigated for continuous hydrogenation of fatty acids in a fixed-bed process. Catalysts in pellet form based on noble metals are acid resistant, but they are deactivated rather quickly, especially when crude fatty acids are used as feedstock [152].

Manufacturing Procedure. In principle, the same plants can be used for the hydrogenation of fatty acids and for the hardening of oils. Plants in which fatty acids are processed, however, require stainless steel as construction material. The reactor is designed for pressures up to ca. 3.5 MPa. An increasing number of hardening plants operate continuously. The economics of batch operation have been improved significantly by heat recovery systems [153]. Despite good acid resistance of the catalyst, after filtration the hardened fatty acids still contain nickel soap. To minimize this content, the reactor should be discharged by the hydrogen pressure and the hardened fatty acid should be cooled (heat recovery) to about 80 – 120 °C before filtration under ambient conditions. The hardening step is usually followed by distillation to remove the nickel soap and to free the hardened fatty acids from the odor that arises during hardening.

Figure 33 shows a typical batch hardening plant. Hardening is carried out in batches of 5 – 30 t. Thorough mixing by effective agitators keeps the catalyst suspended and the hydrogen well dispersed.

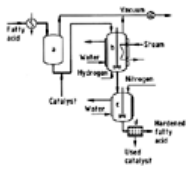

Figure 33. Batch hydrogenation plant

a) Dryer; b) Autoclave; c) Cooler; d) Filter

Particularly thorough mixing is achieved with the reaction loop designed by Lurgi or by Davy Process Technology (formerly Buss); see ⟶ Fats and Fatty Oils). This process can be carried out batchwise in a single reactor or nearly continuously by using several reactor loops in series [154], [155].The plant can also be equipped with a heat recovery system.

In the continuous slurry hardening process developed by Lurgi, a mixture of dried fatty acid, nickel catalyst slurry, and hydrogen is fed into the bottom of the reactor which consists of several chambers. The fatty acid passes through the reactor only once, while the excess of hydrogen is recycled (⟶ Fats and Fatty Oils).

Alternatively, continuous hardening can also be carried out in fixed-bed reactors in the presence of a noble metal (e.g., palladium) pellet catalyst [156]. The feedstock, together with the hydrogen, directly enters the trickle bed at the top of the mixed-phase reactor. The process is free of any slurry handling and eliminates metal soap formation and therefore nickel contamination and nickel residues. A process designed by Lurgi [157] is shown in Figure 34. Also fixed-bed reactors in form of tube-bundle reactors can be used. Such equipment was also developed for the hydrogenation of fatty methyl esters to fatty alcohols [158]. The major advantage is that the tubes filled with the catalyst pellets can be cooled very effectively to remove the exothermic heat and to avoid hot spots. A simulation program for the scale-up of trickle bed hydrogenation reactors are described by [159].

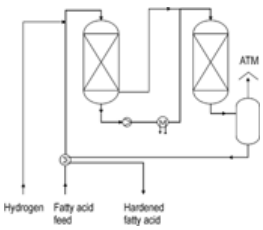

Figure 34. Continuous hydrogenation with fixed bed reactor (Lurgi AG)

3.4.2. *Double-Bond Isomerization*

Isomerization can be split in two main fields of commercial interest: isomerization of monoenes (*cis/trans*) and isomerization of dienes (conjugation).

Monoene Isomerization. Natural fatty acids occur mostly in the higher energy *cis* configuration. The conversion of oleic acid (*cis* configuration) into elaidic acid (*trans* configuration) is carried out industrially by treating oleic acid with sulfur dioxide, oxides of nitrogen, selenium, or acid-activated earths at 100 – 200 °C. The rearrangement reaches an equilibrium at 75 – 80 % elaidic acid [160-162].

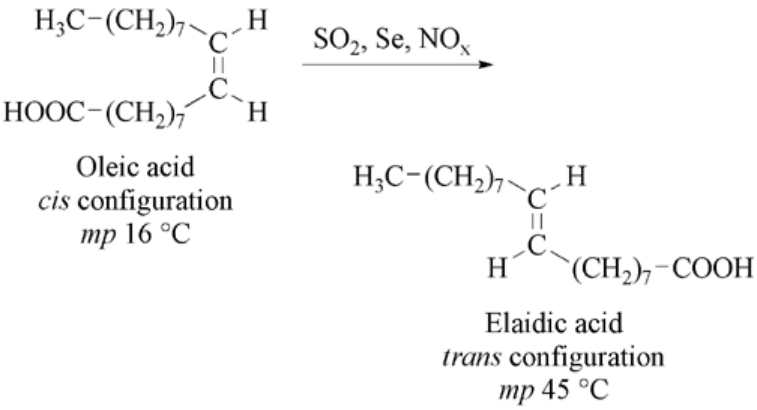

Partial hydrogenation of polyunsaturated fatty acids always increases the proportion of *trans* isomers. Therefore, this process can be used specifically for the preparation of technical elaidic acid. Selective partial hardening of oils for the food industry is carried out to increase the melting point of the oil. In this case low *cis-trans* isomerization is of interest. Today *trans* fatty acids are avoided in food applications because of negative health effects on blood cholesterol levels [163]. It was found that temperature, pressure, and catalyst type are the main parameters for formation of *trans* isomers. Operating at lower temperature and higher pressure with a noble metal catalyst results in a large drop in *trans* formation but also reduced selectivity [164-167].

Reduced formation of *trans* fatty acids was also achieved by using special types of zeolites suitable for triglycerides as well as for fatty acids [168], [169].

Another way to control of the melting point of fats without any problems with *trans* isomers is interesterification of completely saturated triglycerides with unhydrogenated oils (mostly *cis*).

Diene Isomerization. The conjugation of dienoic fatty acids has some commercial importance because of the higher reactivity of conjugated unsaturated fatty acids in technical applications such as alkyd resins and because of effects of single isomers in human nutrition [170]. Conjugated systems for technical applications are formed by alkali-catalyzed isomerization of nonconjugated fatty acids, e.g., fatty acids from sunflower or soybean oil:

$$-CH_2-CH=CH-CH_2-CH=CH-CH_2- \rightarrow -CH_2-CH_2-CH=CH-CH=CH-CH_2-$$

The conjugation of diene fatty acids is carried out by saponification with an aqueous alkali solution and heating with excess alkali to 200 – 300 °C. The soap paste is treated with mineral acid to recover the fatty acid mixture. The degree of conjugation achieved is 95 % [171], [172]. The conjugated diene fatty acids obtained are partially isomerized, so that a mixture of the *cis-cis*, *cis-trans*, *trans-cis*, and *trans-trans* configurations is formed.

For the more selective preparation of conjugated linoleic acid isomers, a catalytic isomerization process starting from the monoester of linoleic acid has been developed [173]. This process leads mainly to the 9-*cis*-11-*trans* and the 10-*trans*-12-*cis* isomers and avoids the formation of other species. Due to growing scientific interest in the biological effects of the single isomers [174], a synthesis of the pure 9-*cis*-11-*trans* isomer starting from ricinoleic acid has also been published [175], [176].

3.4.3. *Dehydration*

Fatty acids with a hydroxyl group in the carbon chain can be dehydrated. Ricinenic acid is obtained by dehydration of ricinoleic acid [177]. Ricinenic acid is a mixture of ca. 30 % conjugated 9,11-linoleic acid and ca. 50 % unconjugated 9,12-linoleic acid. The double bonds are partly in the *trans* configuration. By subsequent conjugation, a ricinene fatty acid with an increased content of conjugated double bonds can be prepared. A similar process is known for the dehydration of castor oil (⟶ Drying Oils and Related Products).

3.4.4. *Dimerization*

Unsaturated fatty acids can undergo self-condensation reactions to form high molecular weight dibasic and polybasic acids. Thermal condensation of unsaturated fatty acids was reported in the early 1900s [178], [179] and patents for commercial processes were issued to Emery Industries in 1949 [180] and 1953 [181]. This thermal condensation is believed to be a Diels – Alder type reaction, and the reaction rates are reported to be proportional to the degree of unsaturation [182], [183]. The products are known as dimer acids (Dicarboxylic Acids, Aliphatic – Dimer Acids). The feedstocks used are generally highly unsaturated C_{18} fatty acids derived from vegetable sources such as tall oil.

By the 1950s patents had been issued for catalytic condensation processes [184], [185]. The catalytic condensation reaction is believed to proceed via a carbonium ion intermediate [186] and is much less dependent on degree of unsaturation, which allows even mono-unsaturated fatty acids (e.g., oleic acid) to condense. The principal process flow sheet is given in Figure 35.

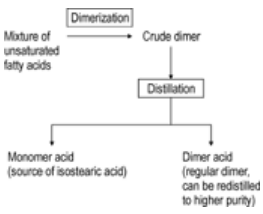

Figure 35. Principle process scheme for dimer production

Dimer acids are known to be a mixture of many chemical isomers with branching in their structure; therefore, they are liquid at room temperature. It was shown that they contain acyclic, cyclic, and aromatic structures, all of which show many isomers [187], [188].

The carbonium ion intermediate in this reaction can also undergo rearrangement to form a branched-chain monomeric acid, which is the source of commercial grades of isostearic acid [182]. The isostearic acid is obtained by hydrogenating the monomer acid and then separating into liquid and solid components. The liquid fraction is termed isostearic acid.

A special type of dimerization is the formation of so-called estolides, which are formed from two or more fatty acids when the carboxyl group of one fatty acid reacts with the double bond or hydroxyl group of another. Estolides are found naturally in some oils (e.g. Euphorbiaceae), but mostly are sythetically produced from ricinoleic acid (12-hydroxy-9-octadecenoic acid) or oleic acid [189], [190], as shown in Figure 36. Also the use of meadowfoam oil fatty acid (64 % C_{20}:1, *cis*-9-eicosenoic acid) has been reported [191]. In principal estolides can be characterized as fatty acid polymers with ester bonds and monocarboxylic acid functionality.

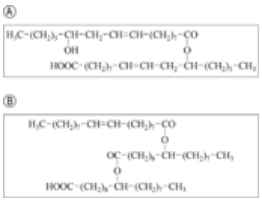

Figure 36. Formation of estolides (A: ricinoleic acid, B: oleic acid)

3.4.5. *Ozonolysis*

Oxidative cleavage of the double bond of oleic acid has been known for many years. This process leads to a C_9 dicarboxylic acid (azelaic acid, ⟶ Dicarboxylic Acids, Aliphatic – Degradative Methods) as main product and to C_9 monocarboxylic acid (nonanoic or pelargonic acid). A first commercial process was introduced in the late 1940s using chromium acid as oxidative agent [192]. Later the oxidative cleavage of oleic acid was carried out with ozone and an oxidative treatment [193], [194]. In the first stage, ozone is added to the double bond of oleic acid. The reaction is carried out preferably at 25 – 45 °C with oxygen containing 1 – 5 % ozone. In a second reaction stage, the ozonide is then split by oxidation with ozone-free oxygen to form crude azelaic and pelargonic acids. This step is carried out at 75 – 120 °C. Afterwards further steps for purification of the crude azelaic acid are necessary [195]. Azelaic acid is used for the production of specialty nylons, ester and polyester for elastomers, polyamide fibers, resins, lubricants, and also in cosmetic applications for skin care and an active ingredient for the treatment of acne vulgaris [196].

3.4.6. *Thermal Decomposition*

Heptanoic acid is produced by Atochem in France by thermal decomposition of esters made from castor oil. At 500 – 600 °C, ricinoleic acid methyl ester decomposes in the presence of water vapor to form oenanthal (heptanal) and undecenoic acid methyl ester. The oenanthal is then oxidized with oxygen to heptanoic acid [197]. The main product, undecenoic acid, is used as starting material in the manufacture of nylon 11 (Rilsan).

The alkaline cleavage of ricinoleic acid or castor oil at higher temperature leads to sebacic acid and 2-octanol as a byproduct [198] (⟶ Dicarboxylic Acids, Aliphatic – Degradative Methods).

3.4.7. *Biooxidation*

In the biooxidation process a terminal methyl group of the fatty acid is transformed into a carboxyl group without shortening the carbon chain. This fermentation method is useful for the production of long-chain dicarboxylic (dibasic) acids (carbon number > C_{12}). The microorganism used in the process is a genetically engineered yeast, *Candida tropicalis* [199]. The process has been developed since 1985, starting with modification of the *Canidida tropicalis* strain [200], [201]. The complex fermentation procedure is followed by extensive downstream processing to produce purified dicarboxylic acids (see Fig. 37) [202-204]. An overview on existing routes for the different carbon chain lengths together with the new possibilities via biooxidation is given in Figure 38. The long-chain dicarboxylic acids are mainly used in polymer and hot-melt adhesive applications.

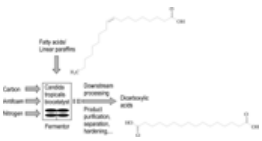

Figure 37. Process scheme for production of long chain dicarboxylic acids

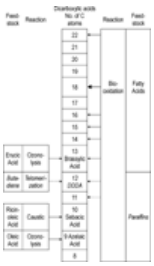

Figure 38. Existing routes and new route via biooxidation for production of dicarboxylic Acids [205]

DDDA: Dodecanediacid

3.4.8. *Enzymatic Esterification*

Lipases not only catalyze the hydrolysis of glycerides, they can also catalyze the esterification of fatty acids with alcohols to give different esters.

Most nonpolar organic liquids generally deactivate biocatalysts, though many remain active in polar solvents like acetone or water. Modification of the biocatalyst may stabilize it against deactivation. Immobilized lipases have been found to remain active in anhydrous organic solvents [206-211]. This property enables lipases to be widely used as biocatalysts in organic syntheses of various useful esters.

Esterification is the reaction between a fatty acid and an alcohol.

$$\mathrm{RCOOH + R'OH \overset{Lipase}{\rightleftharpoons} RCOOR' + H_2O}$$

Water as a product of the reaction has dramatic effects on shifting the reaction equilibrium. It must be continuously removed from the system to enhance the yield. On the other hand, a certain dynamic water environment should be maintained in order to have high enzyme activity. The removal rate of water from the reaction system should be carefully controlled in order not to dry out the lipase. The most common methods to remove water from the reaction system are vacuum and membrane evaporation, azeotropic solvent distillation, drying agents, and bubbling of dry nitrogen [212].

The development of enzyme immobilization enables the enzyme esterification process to be performed at low water contents. The immobilized enzyme is surrounded with water on the carrier (microaqueous reaction system) while the overall water content of the system is low (see Fig. 39).

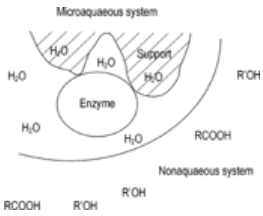

Figure 39. Scheme for immobilized enzyme and water content in the reaction system

The immobilization technique enables enzymatic esterification to compete with chemical esterification processes both with respect to operational and capital costs. Production rates of several thousands of kilograms of product per kilogram of immobilized catalyst and operational lifetimes of up to several months in packed-bed reactors or slurry systems encourage use of enzymatic catalysis [213].

At the same time the enzyme process produces pure products with respect to functionality, quality, and health aspects. Due to their high catalytic activity at ambient temperature and under mild reaction conditions it is possible to use their stereoselectivity toward a broad range of substrates and to synthesize enantiomerically pure intermediates [214], [215].

Owing to the solvent capability of most of the alcohols, esterification can be performed in solvent-free, homogeneous systems to increase the economic profitability of the process [216].

Meanwhile a large range of alcohols has been used for the synthesis of a variety of ester compounds. Table 17 lists some examples [212].

Table 17. Products from enzymatic esterification of C_6–C_{26} fatty acids

OH source	Products/applications
Straight-chain alcohols (C_1–C_{18})	biofuel, emollients, cosmetics
Straight chain alcohols (C_{19}–C_{28})	wax esters
Branched chain alcohols (C_3–2_8)	wax esters, lubricants
Glycerol and glycerides	triglycerides, 1,3- or 1,2-diglycerides, 1- or 2-monoglycerides, functional foods, emulsifiers
Glucose, fructose, etc	sugar and fructose esters, active cosmetics, emulsifiers
Sterols	sterol esters, functional foods
Ascorbic acid	ascorbyl esters, food preservative
Glycols	glycol esters, cosmetics

[Top of Page]

4. Production of Synthetic Fatty Acids

Mixtures of saturated fatty acids have been manufactured in the past by oxidation of alkanes, but have no economic relevance today. This process led to fatty acids with odd-numbered and branched carbon chains that were inferior in quality and were therefore mainly used in technical applications [217-220], [221].

Specific fatty acids for technical applications are produced from olefins by hydroformylation and hydrocarboxylation.

At present, no economical synthesis for unsaturated, even-numbered, long-chain fatty acids (e.g., *cis*-9-octadecenoic acid) exists.

4.1. Hydroformylation

In the oxo reaction discovered by ROELEN, carbon monoxide and hydrogen are added to the double bond of olefins to form aldehydes (⟶ Oxo Synthesis). This highly exothermic reaction takes place only in the presence of catalysts at temperatures above 180 °C and pressures of 8 – 30 MPa. Oxidation of the intermediate aldehyde yields the desired fatty acid. By starting with terminal linear olefins, both straight-chain and α-branched-chain fatty acids can be obtained [222]. The ratio of straight-chain to branched-chain fatty acids can be varied by changing the catalyst in the hydroformylation reaction. Suitable catalysts are cobalt compounds or rhodium phosphine complexes. The latter result in a higher percentage of straight-chain compounds. Celanese employs this process [223], [224]. Fatty acids with varying degrees of branching are produced in the range of heptanoic to nonanoic acid.

4.2. Hydrocarboxylation

In a similar reaction, the Koch – Haaf method, carbon monoxide is added to the double bond of olefins in the presence of strong acids. In a second step, the intermediate product reacts with water to form branched-chain fatty acids [225]. Concentrated sulfuric acid, phosphoric acid, hydrogen fluoride, or boron trifluoride can be used as catalyst. The reaction mechanism involves a carbonium ion, which reacts with carbon monoxide to form an acylium ion. This then reacts with water to form carboxylic acid. Fatty acids with highly branched chains result from rearrangement of the carbonium ion. The process operates under mild conditions up to 80 °C and 10 MPa. Shell manufactures Versatic acids by this method (for Koch process see ⟶ Carboxylic Acids, Aliphatic – Carboxylation of Olefins, Koch Process and for Versatic acids see ⟶ Carboxylic Acids, Aliphatic – Versatic Acids and Neo Acids).

4.3. Other Commercial and Noncommercial Fatty Acid Syntheses

A method used for manufacturing special fatty acids is the caustic fusion of alcohols. Hoechst (now Clariant) converts tridecyl alcohol to isotridecanoic acid, and the oleochemical part of Henkel (now Cognis) produces isopalmitic acid from the C_{16} Guerbet alcohol (⟶ Fatty Alcohols – Guerbet Alcohols) made from caprylic alcohol [226], [227].

Furtheron Akzo was investigating the addition of acetic anhydride to linear terminal olefins (telomerization). After hydrolysis of the intermediate product, branched-chain fatty acids with high molecular masses and low melting points are obtained, which are especially useful in the manufacture of synthetic lubricants and enamels [228], [229].

Many other processes for making synthetic fatty acids have been developed to the stage of technical feasibility. However, no synthesis can compete economically with the less expensive natural fatty acids. For example, hydrocarboxylation of olefins or acetylene with carbon monoxide and water in the presence of metal carbonyl catalysts (Reppe reaction, ⟶ Carbonylation – Introduction, ⟶ Carbonylation – Carbonylation of Alkynes, ⟶ Carbonylation – Carbonylation of Alkenes) has not yet been used for the commercial production of fatty acids, although this reaction has been the subject of intensive work by several firms [31], [57], [230], [231].

Likewise, oxidation of olefins with air, ozone, chromium trioxide, potassium permanganate, or other compounds has not resulted in an economical manufacturing process. Oxidation of fatty alcohols with air or oxygen has also not been commercially successful. Reaction of trialkylaluminum (obtained as an intermediate in the Ziegler alcohol synthesis, ⟶ Fatty Alcohols) with carbon dioxide under pressure has not led to a usable process.

[Top of Page]

5. Analysis

A number of wet chemical and physical methods are used for quality control of commercial products and identification of individual fatty acids. The composition of fatty acid mixtures and natural fats is best determined by gas chromatography, which allows the identification of individual fatty acids with great accuracy. Whereas fatty acid mixtures up to a chain length of 20 carbon atoms can be analyzed directly, fatty acids having a longer carbon chain must first be converted to their methyl esters to increase accuracy. Natural oils and fats must first be saponified and then derivatized or analyzed directly depending on the nature of the oil. Gas chromatography also allows determination of *cis/trans* contents or composition of conjugated isomers. Nuclear magnetic resonance (NMR) and infrared (IR) spectroscopy have been successfully applied to the determination of the molecular structure of fatty acids.

Many chemical data determined by wet analysis are used to characterize commercial fatty acids and to check their purity [49]. Important parameters are the acid value, saponification value, iodine value, and content of unsaponifiable matter (⟶ Fats and Fatty Oils – Chemical Methods). From the *acid and saponification values*, the average molecular mass and thus the average chain length of the fatty acids can be calculated. If a distilled fatty acid contains no esters or anhydrides components the saponification value is identical to the acid value. The difference between the saponification and acid values is the so-called *ester value*.

The *iodine value* is a measure of the total unsaturation of a fatty acid but does not permit conclusions to be made about the content of saturated fatty acids. Thus, oleic acid has the same iodine value (90) as a fatty acid consisting of a 1 : 1 mixture of stearic and linoleic acids.

Practically all technical-grade fatty acids contain small quantities of foreign substances which do not have carboxyl groups and are therefore not saponifiable. These constituents of natural fatty acids arise from foreign material in the fats, natural byproducts of the organisms that also generated the oil or fat, or from decomposition products. The *content of nonsaponifiable material* in technical-grade fatty acids is regarded as a quality criterion and is usually in the range of 0.5 – 2 %.

Based on the composition of the fatty acids determined by gas chromatography, analytical data, normally obtained by traditional wet chemical analysis as acid or saponification value or iodine value, can also be calculated. The calculation is the more exact the lower the content of unsaponifiables is.

Important physical data for characterizing fatty acids are the solidification point, sometimes called titer, the cloud point, and the color. The *solidification point* is the temperature at which a fatty acid begins to solidify and is defined as the temperature maximum resulting from the release of the heat of crystallization.

Low-melting technical fatty acids without sharp solidification temperatures (< ca. 15 °C) are characterized by the *cloud point*, i.e., the temperature at which a clearly perceptible turbidity occurs.

The measurement of *color* is also used to specify the quality of fatty acids. Color determination can be carried out by comparison with either standard solutions (APHA, Hazen) or colored glass standards (Lovibond, Gardner, F.A.C.). APHA/Hazen define their standards as aqueous solutions made of inorganic salts in given concentrations. Today for all methods the determination itself is mostly no longer carried out by subjective visual comparison but by means of automatic spectrometers. Instead of the Lovibond yellow and red colors it is also common to use the percentage transmission at defined wavelengths, e.g., at 440 and 550 nm.

Analytical methods are standardized and well documented in compendia issued by different associations. Important ones are ISO, AOCS, and DGF [232-234] (see Table 18).

Table 18. Common analytical methods for fatty acids

Method	ISO	AOCS	DGF
Acid value	660	Te 4a-70	C-V 2
Saponification value	3657	Tl 1a-64	C-V 3
Iodine value	3961	Tg 1a-64	C-V 11d
Unsaponifiables	3596-1	Tk 1a-64	C-III 1a
Titer	935	Tr 1a-64	C-IV 3c
Cloud point		Cd 15-78	D-III 3
Color (Lovibond scale)		Cc 13e-92	C-IV 4b
Gas chromatography	5508	Ce 1e-91	C-VI 10

[Top of Page]

6. Storage and Transportation

Storage. Fatty acids are unstable in the presence of atmospheric oxygen. As with fats, autoxidation of fatty acids leads to the formation of hydroperoxides, which decompose to oxygen-containing products such as aldehydes, ketones, and hydroxy compounds. Other secondary products, such as polymers and splitting products, can also be found [31], [235-238].

The effect of atmospheric oxygen on fatty acids depends primarily on the temperature, the number of double bonds, and the molecular structure. Saturated fatty acids show little tendency to undergo autoxidation, whereas unsaturated fatty acids, and especially polyunsaturated acids, are very susceptible in this respect. Higher temperatures favor the reaction. Small quantities of other substances present either naturally or adventitiously in the fatty acids can act as pro- or antioxidants. All vegetable oils and fats contain, for example, tocopherols as natural antioxidants. In general, vegetable oils have a higher tocopherol content than animal fats and therefore are usually more stable than animal fats with an equivalent degree of unsaturation. On the other hand, metals such as copper and iron (in the form of soaps) act as prooxidants [49].

Autoxidation products affect the color stability of fatty acids, produce a pungent, rancid odor, and alter characteristic analytical values. The content of nonsaponifiable matter and the ester value increase; the acid value and the iodine value decrease. The peroxide value is a measure of the degree of oxidation [239]. The stability of commercial fatty acids cannot be predicted reliably from the analytical data because the relationship of the raw material, manufacturing process, and storage conditions to the autoxidation behavior is very complex.

Synthetic antioxidants are not normally used in manufacturing, but they may be necessary for special applications of technical-grade fatty acids. Thus, up to 0.5 % 2-naphthol is added to oleic acid used as a lubricant in the textile industry. This reduces the tendency to self-heating as a result of autoxidation (Mackey test). Other widely used synthetic antioxidants for fatty acids are butylated hydroxytoluene (BHT), butylated hydroxyanisole (BHA), and *tert*-butyl hydroquinone (TBHQ), which are used for stabilization of oils and fats as well as for fatty acids. They are approved in many countries for food, drugs, and cosmetics [49], [238]. Concentrations in the order of 0.01 % have proved sufficient. With regard to health and safety, fatty acids can alternatively be stabilized by natural antioxidants such as tocopherols and their esters in concentrations on the order of 0.1 % [240-242].

Fatty acids are stored in liquid or solid form. In most cases, liquid fatty acids are kept in heated tanks. Stainless steel and aluminum are the construction materials normally used for tanks and pipelines. Storage temperature should be slightly above the solidification point.

Under these conditions, saturated fatty acids can be stored for several weeks without detrimental effects on quality; however, with some unsaturated fatty acids, the color deteriorates after a few days. Stability can be improved by storage under a nitrogen blanket.

Transportation. Railway tank cars, tank trucks, containers, ships, and aircraft are used for transporting liquid fatty acids. Materials and temperature correspond to those used for storage. Dispatch in steel drums lined with a protective, fatty acid-resistant coating or in plastic drums is less important.

Because of their corrosiveness, short-chain $C_6 – C_9$ fatty acids are subject to international regulations for the transportation of dangerous goods. This no longer applies to C_{10} or mixtures of C_{10} and C_{12} fatty acids, which must be labeled as irritant. Mixtures with main component C_8 such as $C_8 – C_{10}$ must be labeled as corrosive. Mixtures with main component C_{12} such as $C_8 – C_{14}$ or $C_{12} – C_{18}$ fatty acid fractions from palm kernel must be labeled as irritant. Special labeling and handling are not required for C_{14} and longer chain fatty acids. An overview on the classification for road (ADR), railroad (RID), inland water (ADNR), marine (IMDG), and air transport (IATA) is given in Table 19.

Table 19. Transport and safety classifications of fatty acids

Related to	Hazard symbol[a] ADR, RID, ADNR, IMDG, IATA	Tranport class no. ADR, RID, ADNR, IMDG, IATA	Transport UN no. ADR, RID, ADNR, IMDG, IATA	Hazard ident. no. ADR, RID	Packaging group IMDG, IATA	WGK[b]	Storage class VCI[c]
C_6	C	8	2829	80	III	1	8A
C_8	C	8	3265	80	III	1	8A
C_8–C_{10} (main C_8)	C	8	3265	80	III	1	8A
C_{10}	Xi					1	10
C_{12}	Xi					1	11
C_8–C_{14} (mainly C_{12})	Xi					1	10
C_{12}–C_{18} (mainly C_{12})	Xi					1	11
$\geq C_{14}$							11

[a]C = corrosive, Xi = irritant. WGK = Wassergefährdungsklasse (water-endangering class) according to German VwVwS of May 17, 1999, Appendix 1/2 (1 = slightly water-endangering product). VCI = Verband der chemischen Industrie (Association of the German Chemical Industry).

In compliance with EEC directives [243], members of the European Association of Fatty Acid Producing Countries (Association Européenne des Producteurs d'Acides Gras, APAG) have worked out a uniform method of labeling for Europe [244]. All available data on the transported product (including toxicity data), protective measures for storage and handling, and measures to be taken in case of accidents and fires are specified in the respective safety data sheets.

[Top of Page]

7. Environmental Protection, Toxicology, and Occupational Health

Organic substances such as fats and their components (fatty acids and glycerol) are major constituents of wastewater streams from industrial plants engaged in oleochemistry and in processing natural fats and oils or their fatty acids. In domestic wastewater streams, these substances are also ubiquitous components and originate from soaps, kitchen waste and foodstuffs.

Whereas saturated fatty acids of a chain length of 14 or higher are not classified as being a danger to water (German water-endangering class WGK 0: generally not water-endangering) because they have low toxicity and are easily biodegradable, saturated acids up to C_{12} and unsaturated fatty acids are classified as slightly water endangering (WGK 1). However none of the fatty acids is classified as environmentally dangerous with respect to the EEC directive [243]. When industrial plants discharge their wastewater into municipal sewage systems for joint treatment with domestic wastewater, a certain pretreatment may be required. Simple gravitational separation is often sufficient due to the low solubility of fatty acids in water. If necessary addition of traces of sulfuric acid and/or a surfactant greatly enhances the degree of separation. Emulsified fatty acids are separated and skimmed, whereas dissolved fatty acids remain in the wastewater. The latter fatty acids are mostly short-chain types originating from the vacuum systems of the distillation units.

The operation of sewage treatment plants can be badly disturbed by bulking sludge and scum caused by microorganisms (Type 021N, *Nocardioform actinomycetes*). It was found that native fats are primarily responsible for promoting the growth of the organisms. Therefore, the problems can be eliminated by reducing the load of native fats [245].

The entire organic load of the wastewater is established on the basis of the chemical oxygen demand (COD). The biological oxygen demand (BOD_n) within a certain number of *n* days at 20 °C is a measure of biodegradability achieved by the action of microorganisms [246]. The BOD_5 values of some fatty acids and related compounds are given in Table 20. Under both aerobic and anaerobic conditions, unsaturated fatty acids are degraded more rapidly than saturated acids. The rate of degradation of saturated fatty acids increases with decreasing chain length [31]. Fatty acids fulfil the ultimate (ready) biodegradability pass criterion of 60 % BOD/COD according to EEC directive [247] (method: OECD 301).

Table 20. Values of BOD_5 for fatty acids and related compounds [31]

Component	BOD_5, mg/g
Palmitic acid	1070
Stearic acid	786
Sodium palmitate	1020
Sodium stearate	1200
Sodium oleate	1430
Glycerol	780

In addition to the organic load, often also the heavy-metal load and concentration is limited for wastewater. In the case of fatty acid production Ni plays the main role. Nickel in wastewater results from cleaning processes in the hydrogenation plant or from treatment of distillation residues of hydrogenated acids with mineral acid and water. Nickel content can be reduced sufficiently by carefully handling of the spent catalyst and by final treatment of the contaminated wastewater streams. Precipitation with alkaline agents such as calcium hydroxide can achieve Ni contents of 1 ppm and lower.

Air pollution in the fatty acid industries, which results from the venting of storage tanks, reactors, or other process units, is objectionable due to the resulting odors mainly caused by unsaponifiables and short-chain fatty acids up to capric acid rather than toxic properties. The best way of dealing with odors is combustion. This is performed most economically by leading the exhaust air to the boiler house to replace part of the air normally used for combustion. Solid and liquid residues are processed to form a suspension which is then burned in the boiler house or utilized in special combustion plants for the production of low-pressure steam [248].

The toxicological properties of C_6 – C_{12} fatty acids can be summarized. Their acute oral toxicity LD_{50} in rats is > 2000 mg/kg (method: OECD 401). Therefore, they can be classified as nontoxic [243], [249-253]. Neither MAK nor TLV values have been established. However, fatty acids up to lauric acid are corrosive or irritant, and skin or eye contact should be prevented. The acute fish toxicity (LC_{50}) is about 100 mg/L or more (method: ISO 7346/2).

[Top of Page]

8. Uses

Fatty acids are used in various branches of industry as the acids themselves or in the form of derivatives such as soaps, esters, alcohols, or amides [49], [31], [254], [255], [256]. The shares of the main end-use sectors consuming fatty acids are given in Table 21 for Europe and worldwide consumption [255], [257-259].

Table 21. Main applications of fatty acids in 2003

	Europe %	World, %
Detergents, personal care, soaps, cosmetics	40	49
Plastics	14	10
Rubber and tire industry	7	6
Lubricants and greases	6	3
Paper chemicals	5	3
Protective coatings and resins	5	3
Others (food, feed, candles, textile, leather, adhesives, agrochemicals, construction . . .)	23	26
	100	100

The main end-use sectors as function of applied fatty acid C chain is shown in Table 22 [256].

Table 22. Main applications of fatty acids as function of C-chain

Product	C_8-C_{10}	C_8	C_{10}	C_{12}	C_{14}	C_{16}	$C_{18:1}$	C_{18}	C_{20}-C_{22}	C_{22}
	Coconut- and Palmkernel Oil					Palm oil	Tallow, soybean oil, rapeseed oil		Rapeseed oil (erucic rich)	
Soaps, detergents										
Cosmetics										
Textile										
Coatings and Inks										
Rubber, tyres										
Candles, waxes										
Plastics										
Paper, leather										
Lubricants										
Emulsifiers										
Food										

One of the oldest applications of fatty acids is in the manufacture of candles. Stearin (saturated C_{16} – C_{18} fatty acids) has been used for over 150 years as the basic material for this purpose. However, stearin has been replaced largely by paraffin (⟶ Candles).

Fatty alcohols, fatty amines, and fatty acid esters represent important intermediates in many different fields of application. The syntheses of these derivatives are analogous to those of other monocarboxylic acids, and reactivity decreases with increasing molecular mass.

The fatty acid esters include methyl esters, partial glycerides, wax esters (esters of fatty acids with long-chain fatty alcohols), and ester oils (esters of fatty acids with polyalcohols). Fatty acid methyl esters are intermediates in the manufacture of fatty alcohols and fatty acid alkanolamides. The partial glycerides (mono- and diglycerides) of hardened fatty acids in the palmitic – stearic acid range are used in the food industry as emulsifiers in cakes, pastries, and ice cream. Ester oils are used as lubricants for car engines and compressors, hydraulic fluids, and biogreases [260]. Epoxidized fatty acid esters are used as softeners or plasticizers in plastics such as poly vinyl chloride (PVC). Lead and cadmium stearates are good stabilizers for PVC. as are zinc, calcium, calcium – zinc, and barium – zinc stearates. Cadmium has been replaced because of its toxicity. Wax esters of long-chain fatty acids and fatty alcohols are used as external lubricants in PVC processing [261].

In detergents, soaps, and cosmetics, fatty acids are used primarily in the form of their sodium soaps. The sodium soaps used in the manufacture of soap bars are sometimes made from fatty acids or their methyl esters but are still obtained mainly by saponification of neutral oils many of the newer synthetic detergent (syndet) cleansing bars incorporate teh sodium salt of a fatty acid/isethionate ester. Fatty acid alkanolamides and quaternary fatty alkylammonium salts are used in detergents and cosmetics as foam stabilizer, booster, thickening agent, and consistency factor, and as base for shampoos and body cleansing products. Sulfonated fatty acid methyl esters are of interest as readily biodegradable detergents. Esters such as isopropyl myristate and the triglycerides of short-chain fatty acids are employed as the oil component of cosmetics and pharmaceutical products. Metal soaps, such as aluminum, magnesium, and zinc soaps, serve as thickening agents in cosmetic creams. These metal soaps are also used in powders, because of their lubricating properties. Sodium, lithium, and calcium soaps are employed as additives in mineral-oil lubricants for high-performance engines.

Fatty acid amides serve as slip agents and antiblock agents in plastic formulations [261], stabilizers of emulsions and dyes, builders in synthetic detergents, intermediates in the production of textile water repellent [49], and as components for agricultural chemicals such as pesticides.

Fatty acid amines produced from fatty acids via the nitrile route are used in the rubber industry as mold-release and accelerating agents and for reclamation. They are employed as corrosion inhibitors in the metal industry and for flotation of minerals. After quaternization they are applied for preparing cationic softeners in the textile industry, as antistatic agents, for producing cationic surfactants agents, and oil-well drilling auxiliaries.

Both fatty amides and amines can be transformed into nonionic surfactants by reaction with ethylene oxide, whereas processing with chloroacetates yields amphoteric surfactants with betaine structure [257].

Large quantities of fatty acids are required in the production of alkyd resins for surface coatings (e.g., paints and varnishes) used in all areas of coating applications. The polyunsaturated fatty acids obtained from soybean and sunflower oils are reacted with a polyhydric acid such as phthalic anhydride and polyalcohols to form alkyd resins and are especially important for this purpose [254]. Conjugated fatty acids, saturated fatty acids, and short-chain fatty acids are also used in making paints.

Dimer fatty acids from tall, soybean, or sunflower oil are are mainly employed as polyamides for the coatings and paint industry, as well as for hot-melt adhesives and printing inks. Reactive polyamides of dimer fatty acids are effective hardeners for epoxy resins. Isostearic acid as byproduct of the dimerization process is a monofunctional branched-chain fatty acid. Because of their low solidification points (titers) and high oxidative stability its esters and the conversion product isostearic alcohol are especially used for lubricants, cosmetics, and detergents. Isopalmitic acid has the same application areas. Estolides (also one acid function) and estolide esters have growing importance in cosmetics, coatings, and biogradable lubricants. Another special form of acid dimers that have been used for years in the paper industry for paper sizing (surface treatment) are alkyketene dimers (AKD) based on C_{16} – C_{22} saturated fatty acids, which are prepared by dehydrochlorination of the corresponding fatty acid chlorides (⟶ Ketenes – Production and Uses).

Dibasic fatty acids such as azelaic acid (C_9 from oleic acid, see chap. Ozonolysis), sebacic acid (C_{10} from ricinoleic acid, see chap. Thermal Decomposition) and brassilic acid (C_{13} from erucic acid, see Section Biooxidation) find important outlets in form of esters as plasticizers, modifiers, components for lubricants, and for polymers such as PVC and Nylon.

The C_{22} fatty acids erucic acid (monounsaturated) and behenic acid (saturated) from high-erucic rapeseed oil have applications as surfactants, plastics additives (erucaamide, behenamide), and lubricants [262], [263]. Fatty acid distillates of low-erucic rapeseed oil are used as vegetable-based unsaturated C_{18} source. Partial hydrogenation gives a high oleic acid with improved oxidation stability for production of diverse esters or for unsaturated fatty alcohols [264].

Ricinoleic acid as main natural hydroxy acid can be used for direct esters, estolides, fatty amides, or ethoxylates employed for cosmetics, coatings, and lubricants. Undecylenic acid (C_{11}, pyrolysis of ricinoleic methyl ester) is used as starting material in the manufacture of nylon 11 (Rilsan). The hydrogenation of ricinoleic acid leads to 12-hydroxy stearic acid, which is mainly used in lubricant greases [198]. The synthetic production of dihydroxy or polyhydroxy fatty acids involves as first reaction step the epoxidation (⟶ Epoxides) of the base fatty acid (e.g., oleic acid) or oil (e.g., rapeseed, soybean) followed by a catalytic epoxy ring opening in the presence of water or hydrogen donors. Applications exist in polymer additives (PVC), plastics, surfactants, lubricants, and the textile industry.

In the rubber industry, various types of stearin (mixture of C_{16} and C_{18}) are used as lubricants. In tire manufacture, stearic acid is also used as a separating agent during molding. Zinc and magnesium stearates act as accelerators in the vulcanization process. Soaps of diverse fatty acids are used as emulsifiers in emulsion polymerization for the manufacture of synthetic rubber. Stearin also serves as auxiliary for treatment of recycled paper (e.g., deinking) in the paper industry.

Olein, i.e., technical-grade oleic acid, has been used for many years as a lubricant in the textile industry and for leather auxiliaries. Saponification of oleic acid leads to soft soaps. Oleic acid derivatives are employed as wetting, leveling, and finishing agents in many other operations in textile manufacturing. Thus, *N*-methyloleyltaurine is used in dyeing textiles, and sulfonated monoethanolamides are used in washing printed cloths. Finishing agents contain esters and amides of longer chain fatty acids with polyglycols or polyamines. Melamine resins modified with fatty acids serve as impregnating agents. Oleic acids are used for producing fatty alcohols with high oleic content for the cosmetic, detergent, and plastics industries, where they are employed as solvents, oily components, super-fatting agents, defoamers, plasticizers, and emulsifiers. Esters of oleic acid with fatty alcohols such as oleic acid oleyl ester (C_{18} unsatd/C_{18} unsatd) or oleic acid decylester (18 unsatd/C_{10} satd) serve as oily components for cosmetics. Esters of oleic acid with di- or polybasic alcohols such as trimethylolpropane are used for lubricants.

Distillation and fractionation result in bottom products (residues) of 5 – 10 % in relation to the feed crude fatty acid [265]. Coproducts contain free fatty acids in the range of about 25 – 50%. The use as ingredients in feed compounds is limited according to regional laws with respect to heavy metal content and animal origin. APAG reports a maximum of 200 ppm Ni in the fat component and 5 – 20 ppm in the final feed [266]. Owing to mad cow disease (BSE) tallow-based coproducts are not permitted for feeding in Germany [267]. Therefore, an economical outlet for the coproducts in Europe is today mainly thermal utilization as fuel substitute in boiler houses and incinerators of the concrete and steel industry. The heat of combustion depends on the mix and achieves about 70 – 90 % of the value of light crude oil (42 MJ/kg). The allowed Ni content depends on the equipment and license of the end users.

[Top of Page]

References

1. G. Dieckelmann, H.J. Heinz, *The Basis of Industrial Oleochemistry*, Peter Pomp GmbH, Essen 1988.
2. S. Warwel, “Fettsäuren als Chemie-Grundstoffe”, *Fat Sci. Technol.* **94** (1992) 512–523.
3. Oil World, ISTA Mielke GmbH, Hamburg, Germany (www.oilworld.de).
4. J. Knaut, H. J. Richtler: “Oleochemicals Outlook until the 90s,” Proceedings of 2nd World Conference on Detergents, Oct. 5–10, 1986, Montreux (Switzerland), The American Oil Chemists’ Society, Champaign Ill.
5. V. V. S. Mani, A. D. Shitole, *Fats, Oleochemicals and Surfactants — Challenges in the 21st Century*, Science Publishers, Inc., Enfield, NH 1997.
6. A. Brunskill “Analysis of Fatty Acid Capacity and Demand on a Global and Regional Basis”, Proceedings of the World Oleochemicals Conference, Barcelona, April 2002, Ed. AOCS, www.aocs.org.
7. Cognis Company, own estimation.
8. R. Gani, J.P. O’Connell, *Comput Chem. Eng.* **25** (2001) 3–14.
9. W. Marquardt, *Chem.-Ing.-Tech.* **10** (1999) 1119–1137.
10. L. E. K. Achenie, R. Gani, V. Venkatasubramanian: *Computer Aided Molecular Design: Theory and Practice*, Elsevier, Amsterdam 2003.
11. F.D. Gunstone: *The Chemistry of Oils and Fats*, 1st ed, Blackwell, Oxford 2004.
12. F. D. Gunstone: *Fatty Acid and Lipid Chemistry*, Blackie Academic & Professional, London 1996.
13. J. Falbe, M. Regitz: *Römpp Lexikon Chemie, Georg Thieme Verlag*, Stuttgart 1996–1999.
14. H.-D. Belitz, W. Grosch, P. Schieberle: *Lehrbuch der Lebensmittelchemie*, 5th ed, Springer, Berlin 2001.
15. W. Heimann: *Grundzüge der Lebensmittelchemie*, 3rd ed, Dr. Dietrich Steinkopf, Darmstadt 1976.
16. J. F. Diehl: *Chemie in Lebensmitteln*, Wiley-VCH, Weinheim, Germany 2000.
17. D. R. Heldmann, D. B. Lund: *Handbook of Food Engineering*, Marcel Dekker, New York 1992.
18. *Physical and Thermodynamic Properties of Pure Chemicals (DIPPR project 801)* Taylor&Francis, Washington 1999.
19. M. Palczewska-Tulinska, D. Wyrzykowska-Stankiewicz, A.M. Szafranski, J. Cholinski: *Solid and Liquid Heat Capacity Data Collection*, Part 2: C_6-C_{33}, Dechema, Frankfurt 1997.
20. L. S. Cliffort, D. Ramjugernath, J.D. Raal, *J. Chem. Eng. Data* **49** (2004) 1189-1192.
21. D. Ambrose, N. B. Ghiassee, *J. Chem. Thermodyn.* **19** (1987) 505–519. Links
22. Korea Thermophysical Properties Databank.
23. Databank NIST Chemistry WebBook, http://webbook.nist.gov/.
24. Gmehlin, U. Onken, P. Grenzheuser: *Vapor/Liquid Equilibrium, Data Collection 5, Carboxylic Acids, Anhydrides, Esters*, Dechema Chemistry Series, Vol. 1, Part 5, Frankfurt 1982.
25. C. F. Beaton: *Physical Property Data for the Design Engineer*, Hemisphere Publishers, New York 1989.
26. C. L. Yaws: *Handbook of Transport Property Data*, Gulf Publishers, Houston 1995.
27. C. L. Yaws: *Chemical Properties Handbook*, McGraw-Hill, New York 1999.
28. A. E. Bailey: *Melting and Solidification of Fats*, Interscience, New York 1950.
29. K. S. Markley: *Fatty Acids*, 2nd ed, Parts 1–5, Krieger, Malabar 1983.
30. D. Swern, M. M. Chrysam, R. R. Allan: *Bailey’s Industrial Oil and Fat* Products, 4th ed, vols. 1–3, Wiley-Interscience, New York 1979/1982/1985.
31. E. H. Pryde: *Fatty Acids*, The American Oil Chemists’ Society, Champaign, ILL, 1979.
32. Henkel KGaA: *Fettchemische Tabellen*, 3rd ed 1971.
33. H. P. Kaufmann: *Analyse der Fette und Fettprodukte*, Part I, Springer Verlag, Berlin 1958.
34. H. A. Schnette, H. A. Vogel, *Oil Soap* **16** (1939) 209–212; **17** (1940) 155–157.
35. A. W. Ralston: *Fatty Acids and Their Derivatives*, Wiley, New York 1948.
36. R. R. Mod *et al.*, *J. Am. Oil Chem. Soc.* **30** (1953) 368–371.
37. V. Mills, H. K. Mc Clain, *Ind. Eng. Chem.* **41** (1949) 1982–1985.
38. H. D. Forman, J. B. Brown, *Oil Soap* **21** (1944) 183.

39. J. G. Atherton, R. A. Reck, *Toilet Goods Assoc.* **40** (1963) 47–52.
40. N. Adiaanse: Heats of Combustion and Physical Constants of Normal Saturated Fatty Acids and Their Methyl Esters, Dissertation, University of Amsterdam 1960.
41. E. Brandes, W. Möller: *Sicherheitstechnische Kenngrössen, Part 1, Brennbare Flüssigkeiten und Gase*, Wirtschaftsverlag, NW, Bremerhaven 2003.
42. G. Knothe, J. T. P. Derksen: *Recent Developments in The Synthesis of Fatty Acid Derivatives*, AOCS Press, Champaign, IL, 1999.
43. J. O. Metzger *et al.*, *Angew. Chem. Int. Ed.* **39** (2000) 2206–2224. Links
44. H. Eierdanz, F. Hirsinger, *Fett Wiss. Technol.* **92** (1990) no. 12, 463–467.
45. H. Eierdanz: *Perspektiven nachwachsender Rohstoffe in der Chemie*, Wiley-VCH Weinheim, Germany 1996.
46. S.J. Coughlan, A.J. Kinney in K. M. Oksman-Caldentey, W. H. Barz (eds.): *Plant Biotechnology and Transgenic Plants*, Marcel Dekker, New York 2002, pp. 305–321.
47. Y.H. Hui: *Bailey's Industrial Oil and Fat Products*, 5th ed, vol. 2, Wiley, New York 1996.
48. H. Zoebelein: *Dictionary of Renewable Resources*, 2nd ed, Wiley-VCH, Weinheim, Germany 2001.
49. D. Swern: *Bailey's Industrial Oil and Fat Products*, 4th ed, vols. 1 + 2, Wiley-Interscience, New York 1979/1982.
50. L. Lascaray, *J. Am. Oil Chem. Soc.* **29** (1952) 362–366.
51. T. A. Patil *et al.*, *Ind. Chem. Res.* **27** (1988), 727–743.
52. M. Loncin, *Fette Seifen* **55** (1953) 7–9.
53. E. Twitchell, *J. Am. Chem. Soc.* **22** (1900) 22–26; Links US 601 603 (1898).
54. Henkel Corporation, PCT/US 93/10133, 1993.
55. H. P. Kaufmann, M. C. Keller, *Fette Seifen* **44** (1937) 42–47.
56. L. Lascaray, *Fette Seifen* **46** (1939) 628–632.
57. Am. Oil Chem. Soc.: *Short Course on Fatty Acids*, Am. Oil Chem. Soc., Kings Island, Ohio, Sept. 23–26, 1984.
58. G. W. Eisenlohr, US 2 154 835, 1939.
59. V. Mills, H. K. Mc Clain, *Ind. Eng. Chem.* **41** (1949) 1982–1985.
60. D. Deublein, *Fette Seifen Anstrichm.* **87** (1985) 508–512.
61. M. D. Reinish, *J. Am. Oil Chem. Soc.* **33** (1956) 516–520.
62. V. J. Muckerheide, *J. Am. Oil Chem. Soc.* **29** (1952) 490–495.
63. Emery Ind., DE 880 463, 1950.
64. H. L. Barnebey, A. C. Brown, *J. Am. Oil Chem. Soc.* (1948) March, 95–99. Links
65. Crown Iron Works, Minneapolis, MN USA, company brochure.
66. Lurgi AG, Frankfurt, Germany, company brochure "Fatty Acid Technology".
67. Oils and Fats International, Plant & Equipment Survey 2004.
68. Cognis Deutschland GmbH, DE 199 07 458, 1999.
69. Haltermann-Ascot GmbH, EP 1 206 431, 2000.
70. K. Buchholz, V. Kasche: *Biokatalysatoren und Enzymtechnologie*, VCH, Weinheim, Germany 1997.
71. R. Sharma, Y. Chisti, U. C. Banerjee, "Production, Purification, Characterization and Applications of Lipases", *Biotechnol Adv.* **19** (2001) 627–662. Links
72. W. M. Linfield, R. A. Barauskas, L. Sivieri, S. Serota, R. W. Stevenson Sr., "Enzymatic Fat Hydrolysis and Synthesis", *J Am. Oil Chem. Soc.* **61** (1984) 191–202. Links
73. W. M. Linfield, D. J. O'Brien, S. S. Serota, R. A. Barauskas, "Lipid-Lipase Interactions. I. Fat Splitting with Lipase from Candida rugosa", *J Am. Oil Chem. Soc.* **61** (1984) 1067–1071.
74. M. Bühler, C. Wandrey, "Continuous Use of Lipases in Fat Hydrolysis", *Fat Sci. Technol.* **89** (1987) 598–605.
75. F. P. Cuperus, S. T. Bouwer, A. M. A. Krosse, J. T. P. Derksen, *Stabilization of Lipases for Hydrolysis Reactions on Industrial Scale*, International Conference on Membrane Process, Heidelberg 1993.
76. U.T. Bornscheuer, C. Bessler, Optimizing Lipases and Related Enzymes for Efficient Application, *Trends Biotechnol.* **20** (2002) No. 10. Links
77. R. V. Murty, J. Bhat, P. K. A. Muniswaran, "Hydrolysis of Oils by using Immobilized Lipase Enzyme: A Review", *Biotechnol. Bioprocess Eng.* **7** (2002) no. 2, 57–66.
78. Application and Product Sheet for Lipomod 34 P, from Biocatalysts, 2004.
79. Application and Product Sheet for Novozym CALB L, from Novozymes 2004.
80. DE 4124248, 1993 (A. Weiss, G. Lazar, K. Grundel).
81. Cognis Deutschland GmbH, DE 10054480 A1, 2002; WO 2003040091 A3, 2003 (M. W. Christensen, H. C. Holm, K. Abe).
82. Henkel Corporation, US 5677160, 1997.
83. Cognis Deutschland GmbH, EP 1501915, EP 1319714, 2003.
84. G. Fieg, unpublished experimental investigations on fatty acid manufacturing plants.
85. P. Bender, A. Moll, *Trans. IChemE Part A* **81** (2003) 58–67.
86. Z. Olujic, H. Jansen, B. Kaibel, T. Rietforf, E. Zich, *Ind. Eng. Chem. Res.* **40** (2001) 6172–6180. Links
87. L. Spiegel, W. Meier, Proceedings on the International Conference on Distillation & Absorption, Ed. GVC/VDI, Baden-Baden, Germany, 2002, www.vdi.de.
88. F. Moser, A. Kessler, *Vak. Forsch. Prax.* **12** (2000) 122–124.
89. AspenTech Company, www.aspentech.com.
90. Chemstations Company, www. chemstations.de.
91. Simulation Sciences, www.simsci-esscor.com.
92. Company Brochure of Buco-Wärmeaustauscher GmbH, Geesthacht, Germany.
93. Company Brochure of DEG Engineering GmbH, Gelsenkirchen, Germany.
94. Company Brochure: Fatty Acid Technology, mg engineering Lurgi Life Science 2000 Germany, www.lurgi.de.
95. W. Kehse, *Fette Seifen Anstrichm.* **78** (1976) 50–56.
96. P. Faessler, K. Kolmetz, K. W. Seang, S. H. Lee *(Sulzer Chemtech): Advanced Fractionation Technology*, presented on World Congress on Oleochemicals, Kuala Lumpur, Malaysia, Oct. 2004.
97. Henkel KGaA, DE 195 31 805, 1997 (J. Plachenka, J. Hammer, M. Lindemann, U. Steinberner).
98. www.montz.de
99. R. Agrawal, *Ind. Eng. Chem. Res.* **40** (2001) 4258–4266. Links
100. G. Kaibel, C. Miller, M. Stroezel, R. von Watzdorf, H. Jansen, *Chem.-Ing.-Tech.* **76** (2004) 258–263. Links
101. M.A. Schultz, D.G. Steward, J.M. Harris, S.P. Rosenblum, M.S. Shakur, D.E. O'Brien, *Chem. Eng. Prog.* 2002, no. 6, pp. 64–71.
102. H. Becker, S. Godorr, H. Kreis, *Chem. Eng.* 2001, Jan., pp. 68–74.
103. G. Fieg, *Chem. Eng. Processing* **41** (2001) no. 2, 123–133.
104. G. Fieg, B. Landwehr, G. Wozny, L. Jeromin, *Chem. Eng. Technol.* **19** (1996) no. 4, 299–307.
105. W. Stein, *J. Am. Oil Chem. Soc.* **45** (1968) 471–474.
106. Company brochure of Sulzer Chemtech: Fractional Crystallization, 1999, www.sulzerchemtech.com.
107. Henkel & Cie, DE 977 544, 1952; DE 970 292, 1953; GB 925 674, 1955; US 2 800 493, 1955; US 2 972 636, 1961; US 3 052 700, 1962; US 3 733 343,

1973; US 3 737 444, 1973.
108. Lurgi GmbH, Frankfurt (FRG): Fatty Acid Technology, 1984.
109. Lurgi GmbH, Frankfurt (FRG): Lurgi Handbuch, 1970.
110. G. Haraldson, *J. Am. Oil Chem. Soc.* **45** (1984) 219–222.
111. Unilever N.V., DE 1 617 023, 1966.
112. Emery Ind., US 2 298 501, 1942.
113. Chemetron Corporation, US 3 173 935, 1965 (W. A. Singleton).
114. R. L. Demmerle, *Ind. Eng. Chem.* **39** (1947) 126–131.
115. R. H. Potts, G. W. Mc Bride, *Chem. Eng. (N.Y.)* **57** (1950) no. 2, 124–127.
116. *Chem. Process. (Chicago)* **17** (1971) 14–16.
117. K. Zondeck, DE 2 126 969, 1971.
118. GHH-MAN-Technik, DE 2 142 134, 1971.
119. G. Coppa-Zuccari, *Oleagineux* **26** (1971) no. 6, 405–409.
120. Walter Rau Lebensmittelwerke, DE 3 111 320 1981.
121. D. Swern, W. E. Parker, *J. Am. Oil Chem. Soc.* **29** (1952) 614–615.
122. R. Rigamonti, V. Riccio, *Fette Seifen Anstrichm.* **55** (1963) 162, 164.
123. D. G. Hayes, *Inform* **13** (2002) 832–834.
124. Trikonex AB, WO 95/11216, 1995.
125. Societe des Produits Nestle S.A., EP 0682006 B1, 1999.
126. Nippon Oil and Fats Company Ltd., EP 0225946B1, 1991.
127. Norsk Hydro A.S., WO 87/03899, 1987.
128. H. Josten, G. Fieg, B. Gutsche, W. Johannisbauer, *Chem. Ing. Tech.* **76** (2004) 1700–1703.
129. UOP Inc. EP-A 134 357, 1983 (M. T. Cleary).
130. UOP Inc., EP-A 105 066, 1984 (M. T. Cleary).
131. G. Brunner, S. Peter, *Chem. Ing. Tech.* **53** (1981) 529–542.
132. J. Chrastil, *J. Phys. Chem.* **86** (1982) 3016–3021.
133. C. Tiegs, G. Peter, *Fette Seifen Anstrichm.* **87** (1985) 231–235.
134. R. E. Beal, O. L. Brekke, *J. Am. Oil Chem. Soc.* **36** (1959) 397–400. R. E. Beal *et al.*, *J. Am. Oil Chem. Soc.* **38** (1961) 524–527.
135. P. Ekwall, L. Mandell, *Kolloid Z. Z. Polym.* **233** (1969) 938–944.
136. W. O. Munns *et al.*, *J. Am. Oil Chem. Soc.* **39** (1962) 189–193.
137. W. Normann, DE 141 029, 1902.
138. R.C. Hastert, *Hydrogenating Fatty Acids*, World Conference on Oleochemicals, Into the 21th Century, Kuala Lumpur, Malaysia, Oct. 8–12, 1990, T. H. Applewhite (ed.), AOCS, Champaign, IL, 1991, pp. 165–171.
139. L. F. Albright, *Fette Seifen Anstrichm.* **87** (1985) 140–146.
140. Henkel KGaA, DE 4109246.5, 1992 (A. Behr).
141. H. Klimmek, *J. Am. Oil Chem. Soc.* **61** (1984) 200.
142. P. Blom, Chemical Catalyst News by Engelhard Corporation, Oct. 1991.
143. C.M. Lok, R.A. Joosten, *Am. Oil Chem. Soc. Meeting*, Chicago, May 1991.
144. J.W.E. Coenan, Unilever Research Laboratory, Klaardingen, The Netherlands.
145. A. Bartsch, *Z. Naturforsch. A* **50** (1995) Links Henkel kGaA, DE 44 36 064, 1996 (A. Bertsch).
146. R. C. Hastert, T. J. Selliom, J. S. F. Nebesh, *AOCS World Congress*, New York City 1980.
147. S. W. Turner, Henkel Corporation, Emery Group, unpublished report, 1995.
148. R. J. Grau, A. E. Cassano, M. A. Baltanas, *J. Am. Oil Chem. Soc.* **67** (1990) 226–232.
149. M. Münzing, O. M. Kut, G. Gut, *Fette Seifen Anstrichm.* **10** (1986) 387–391.
150. A. Behr *et al.*, *Fat Sci. Technol.* **95** (1993) 2–12.
151. J. I. Gray. L. F. Russel, *J. Am. Oil Chem. Soc.* **56** (1979) 36–43.
152. www.jmcatalysts.com, www.engelhard.com.
153. B. G. M. Grothues, *J. Am. Oil Chem. Soc.* **62** (1985) 390–399. Links
154. R.F. Duveen, G. Leuteritz, *Fette-Seifen-Anstrichm.* **84** (1982) 511–515.
155. D. Urosevic, *AOCS World Conference on Emerging Technologies*, Cannes, France, Nov. 3–8, 1985, pp 128–132.
156. F. Borninkhof, C. Groen, E. G. M. Kuijpers, P. H. Berben, “Saturation of Fatty Acids in Fixed Bed Reactors”, *Annual AOCS Convention in Anaheim*, CA, USA, May 1993.
157. Lurgi AG, Frankfurt, Germany, company brochure “Fatty Acid Technology”.
158. R. Hourticolon *et al.*, Cognis Deutschland GmbH, US 6,683,224 B1, 2004.
159. M. Greger, B. Gutsche, L. Jeromin, *Chem. Ing. Tech.* **64** (1992) no. 3, 253–259.
160. C. Litchfield, *J. Am. Oil Chem. Soc.* **40** (1963) 553–557.
161. Tallow Research Inc., US 3 065 248, 1960.
162. Henkel & Cie., DE 1 211 158, 1963.
163. J. Kristott, *Oils Fats Int.* 2004, no. 3, 20–22.
164. N. Hsu, L.L. Diosady, L. J. Rubin, *J. Am. Oil Chem. Soc.* **66** (1989) 232–236. Links
165. W.T. Koetsier, *Fat Sci. Technol.* **96** (1994) 505–507.
166. K. Furlong, *Oils Fats Int.* 2004, no. 6, 30–31.
167. G. Magnus, A. Beers, *Oils Fats Int.* 2004, no. 7, 33–35.
168. D. Packet, “Making Oleochemicals in a More Efficient Way”, *presentation at World Oleochemical Conference*, Barcelona, 14–17 April 2002, AOCS.
169. K.U. Leuven Research & Development, WO 98/54275, 1999.
170. S. Gnädig, R. Rickert, J. L. Sébédio, H. Steinhart, *Eur. J. Lipid Sci. Technol.* **103** (2001), 56–61.
171. T. E. Bradley, D. Richardson, *Ind. Eng. Chem.* **34** (1942) 237–242.
172. J. D. v. Mikusch, *Farben, Lacke, Anstrichm.* **4** (1950) 149–159
173. A. Seabo, *Adv. Conj. Lipid Res.* **2** (2003), 71–81.
174. M.P. Yurawecz *et al.*, *Advances in Conjugated Linoleic Acid Research*, Vol. 1, AOCS Press, 1999.
175. O. Berdeaux *et al.*, *J. Am. Oil Chem. Soc.*, **74** (1997) 1011–1015.
176. L. Yang *et al.*, *Chem. Phys. Lipids* **119** (2002) 23–31. Links
177. J. Schreiber, DE 512 822, 1928; DE 512 895 1930; DE 833 644, 1950.
178. GB 121 777, 1918 (J. Craven).
179. J. Scheiber, Farbe und Lack (central blatt der dentrar Farbe-und Lack-Industrie, Haumver), Nr. US (4. Des. 1929) 585–587

180. Emery Industries, Inc US 2 483 761, 1949 (C. G. Goebel).
181. Emery Industries, Inc US 2 664 429, 1953 (C. G. Goebel).
182. E. C. Leonard (ed.): *The Dimer Acids*, Humko Sheffield Chemical, Memphis, TN, 1975.
183. D. H. Wheeler, *Offic. Dig. Fed. Paint Varnish Prod. Clubs*, **322** (1951) 661.
184. Petrolite Corp. US 2 426 489, 1947 (M. De Groote).
185. Emery Industries, Inc., US 2 793 219, 1957 (F. O. Barrett, C. G. Goebel, R. M. Peters).
186. R. W. Johnson, Clay Catalyzed Dimerizatrion of Fatty Acids, AOCS Meeting, Chicago, 1991.
187. A. K. Sen Gupta, *Fette Seifen Anstrichm.* **71** (1969) 873.
188. R. Brütting, G. Spiteller, *Fat Sci. Technol.* **96** (1994) 445.
189. S. C. Cermak, T. A. Isbell, *J. Am. Oil Chem. Soc.*, **78** (2001) 557–565.
190. M. K. Schwitzer, *Castor Oil, World Conference on Oleochemicals, Into the 21th Century*, Kuala Lumpur, Malaysia, Oct. 8–12, 1990, T.H. Applewhite (ed.), AOCS, Champaign, Illinois, 1991, pp. 111–118.
191. S. M. Erhan, R. Kleimann, T. A. Isbell, *J. Am. Oil Chem. Soc.* **70** (1993) 461–465.
192. US 2 389 191, 1945 (J. D. Fitzpatrick, L. D. Meyers). US 2 450 858, 1948 (J. D. Fitzpatrick, L. D. Meyers).
193. R. W. Johnson *et al.*: Dicarboxylic Acids, in *Kirk-Othmer*, 4th edition, Vol 8, 1994.
194. Emery Ind., DE 1 040 016, 1958 (Ch. G. Goebel, A. C. Brown *et al.*).
195. H. Josten, "Development of a New Separation Process for Azelaic Acid", *presentation at World Oleochemical Conference*, Barcelona, 14–17 April 2002.
196. Abstract of Azelaic Acid Use–Patents and Journal References, Henkel Corporation, Emery Group, Technical Bulletin 171, Cincinnati 1992.
197. M. Genas, *Angew. Chem.* **74** (1962) 535–540. Links
198. M.K. Schwitzer, "Castor Oil", *World Conference on Oleochemicals, Into the 21th Century*, Kuala Lumpur, Oct. 8–12, 1990, T. H. Applewhite (ed.), AOCS, Champaign, IL, 1991.
199. K. Kroha, *inform* **15** (2004) 568–571.
200. V. B. Urlacher, S. Lutz-Wahl, R. D. Schmid, *Appl. Microbiol. Biotechnol.* **64** (2004) 317–325. Links
201. E. Dudley, D. L. Craft, R. W. Frayer, C. R. Wilson, "Biooxidation—A Platform to Produce Specialty Chemicals", lecture at the Society for Industrial Microbiology, 2003.
202. Cognis Corporation, US 6569670, 2003.
203. Cognis Corporation, WO 2004039992 A2, 2004.
204. Cognis Deutschland GmbH, WO 2003020859 A2, 2003.
205. M. E. Durchholz, K. W. Anderson, Company report, Cognis Corporation, Cincinnati, OH, Oct. 2004.
206. L. Gianfreda, M. R. Scarfi, "Enzyme Stabilization: State of Art", *Mol. Cell. Biochem.* **100**(2) 1997 97–128 Links
207. E. Wehtje, P. Adlercreutz, B. Matthisson, "Stabilization of Adsorbed Enzymes Used as Biocatalysts in Organic Solvents", *Prog. Biotechnol.* **8** (1992) 377–382.
208. M. Arroyo, J. M. Sanchez-Montero, J- V. Sinisterra, "Thermal Stabilization of Immobilized Lipase B from Candida antarctica on Different Supports: Effect of Water Activity on Enzymatic Activity in Organic Media", *Enzyme Microbial Technol.* **24** (1998) 3–12.
209. M. Arroyo, J. M. Sanchez-Montero, J. V. Sinisterra, "Stabilization of Lipase B from Candida antarctica by Immobilization on Different Supports", *Prog. Biotechnol.* **15** (1998) 565–570.
210. A. Tanaka, T. Kawamoto, M. Kawase, T. Nanko, K. Sonomoto, "Immobilized Lipases in Organic Solvents", in *Proceedings of World Conference on Biotechnology for the Fats and Oil Ind.*, Hamburg, Germany, Sept. 27–Oct. 2002, 1987, T. H. Applewhite (ed.), AOCS, 1988, pp. 123–130.
211. P. Eigtved, T.T. Hansen, C.A. Miller, "Ester Synthesis with Immobilized Lipase", in *Proceedings of World Conference on Biotechnology for the Fats and Oil Ind.*, Hamburg, Germany, Sept. 27–Oct. 02, 1987, T. H. Applewhite (ed.), AOCS, 1988, pp. 134–137.
212. X. Xu, "Engineering of Enzymatic Reactions and Reactors for Lipid Modification and Synthesis", *Eur. J. Lipid Sci. Technol.* **105** (2003) 289–304.
213. J. Piao, T. Kobayashi, S. Adachi, K. Nakanishi, R. Matsuno, "Continuous Synthesis of Lauroyl or Oleoyl Erythritol by Packed-Bed Reactor with an Immobilized Lipase", *Process Biochem.* **39** (2004) 681–686. Links
214. A. Kovac, H. Scheib, J. Pleiss, R. D. Schmid, F. Paltauf, "Molecular Basis of Lipase Stereoselectivity", *Eur. J. Lipid Sci. Technol.* **99** (2000) 61–77.
215. N. Heinsman, P. van der Weide, A. van der Padt, M. Franssen, R. Boom, K. van't Riet, "Reactor Design for the Novozym 435-Catalyzed Enantioselective Esterification of 4-Methyloctanoic Acid", *Biocatal. Biotransform.* **20** (2002) 429–436.
216. M. Soumanou, U. T. Bornscheuer, "Lipase-Catalysed Alcoholysis of Vegetable Oils", *Eur. J. Lipid Sci. Technol.* **105** (2003) 656–660.
217. H. Fineberg, *J. Am. Oil Chem. Soc.* **56 A** (1979) 805–809.
218. P. L. Layman, *Chem. Eng. News* 1983, Oct. 10, 15–16.
219. *Ullmann*, 3rd ed, **7**, 552–563.
220. N. M. Emanuel, E. T. Denisow, Z. K. Maizus: *Liquid Phase Oxidation of Hydrocarbons*, Plenum Press, New York 1967.
221. D. Osterroth, "Die Gewinnung von synthetischen Fettsäuren durch Oxidation von Paraffinen, CLB Chemie" *Labor Biotech.* **43** (1992) no. 4, pp. 177.
222. J. Falbe: *Carbon Monoxide in Organic Synthesis*, Springer Verlag, Berlin 1970.
223. Celanese Corp., US 4 138 420, 1979 (J. D. Unruh, W. J. Wells).
224. www.celanesechemicals.com.
225. H. Koch, *Brennst. Chem.* **36** (1955) 321–328.
226. Henkel & Cie, DE 2 320 641, 1974 (H. Schütt).
227. G. Mueller *et al.*, DE 4302463.7, 1993.
228. Akzo N. V., 014 910, 1977 (W. J. de Klein).
229. *Chem. Ind.* **37** (1985), 307.
230. P. Hofmann, K. Kosswig, W. Schaefer, *Ind. Eng. Chem. Pract. Res. Dev.* **19** (1980) 330–334.
231. P. Hofmann, *Chem. Ing. Tech.* **54** (1982) 694–695.
232. International Organization for Standardization (ISO), rue de Varembé 1, Case postale 56, CH-1211 Geneva 20, Switzerland.
233. *Official Methods and Recommended Practices of the American Oil Chemists' Society*, 4th Ed., AOCS Press, Champaign, IL, 1989.
234. DGF, German Standard Methods for the Analysis of Fats and Other Lipids, Wissenschaftliche Verlagsgesellschaft, Stuttgart 1984.
235. H. J. Skellon, D. M. Wharry, *Chem. Ind.* **23** (1963) 925–932.
236. M. Loury, G. Lechartier, M. Forney, *Rev. Fr. Corps Gras* **12** (1965) 253–262.
237. H. Thaler, H. J. Kleinau, *Fette Seifen Anstrichm.* **70** (1968) 465–471; **71** (1969) 92–98, 261–264.
238. W. O. Lundberg: *Autoxidation and Antioxidants*, vols. I, II, Wiley-Interscience, New York 1961, 1962.
239. *DGF-Einheitsmethoden*, Wissenschaftl. Verlags GmbH, Stuttgart 1984.
240. C. Gertz, S. Klostermann, S. Parkash Kochhar, *Eur. J. Lipid Sci. Technol.* **102** (2000) 543–551.
241. J. Lee, S.M. Lee, I.H. Kim, J.H. Jeong, C. Rhee, K.W. Lee, *J. Am. Oil Chem. Soc.* **80** (2003) 807–810.
242. K.-H. Wagner, I. Elmadfa, *Eur. J. Lipid Sci. Technol.* **102** (2000) 624–629.
243. European Community Council Directive 67/548 EEC of 1967 and later updates (until 2003: 9 amendments and 28 adaptations to technical progress).
244. Report of APAG Working Group to APAG Technical Committee, Jan. 28, 1986.
245. T. Bendt, B. Pehl, A. Werner-Busse, *Vom Wasser* **99** (2002) 53–62.

246. G. Schwedt, H. Jöckel, *CLB Chem. Labor Betr.* **35** (1984) no. 3, 115–119.
247. European Community Council Directive 91/325 EEC of 1991, 12th adaptation of technical progess of EEC 67/548.
248. G. Dieckelmann *et al.*, *Fette Seifen Anstrichm.* **85** (1983), special issue no. 1, 559–562.
249. Registry of Toxic Effects of Chemical Substances, NIOSH, 1981–1982.
250. N. I. Sax: *Dangerous Properties of Industrial Materials*, 6th ed, Van Nostrand Reinhold, New York 1984.
251. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, Wiley-Interscience, New York 1982.
252. J. J. Kabara, *J. Am. Oil Chem. Soc.* **56** (1979) 760 A–767 A.
253. G. A. Nixon, C. A. Tyson, W. C. Wertz, *Toxicol. Appl. Pharmacol.* **31** (1975) 481–490. Links
254. Y.H. Hui: *Bailey's Industrial Oil and Fat Products*, 5th ed, Wiley, New York 1996, Vol. 5, Industrial and Consumer Nonedible Products from Oils and Fats.
255. S. Warwel, "Fettsäuren als Chemie-Grundstoffe", *Fat.Sci. Technol.* **94** (1992) 512–523.
256. P. Hövelmann, B. Brackmann, *SöFW-J.* **123** (1997) 697–707.
257. A. J. Kaufmann, R. J. Ruebusch, "Oleochemicals: A World Overview", *World Conference on Oleochemicals, Into the 21th Century*, Kuala Lumpur, Malaysia, Oct. 8–12, 1990, T. H. Applewhite (ed.), AOCS, Champaign, IL, 1991, pp. 10–25.
258. B. Brackmann, *SöFW-J.* **125** (1999) 40–46.
259. Cognis Company, own estimation.
260. F. D. Gunstone, R. J. Hamilton: *Oleochemical Manufacture and Applications*, Sheffield Academic Press, Sheffield 2001.
261. C. Leonhard, "Marketing and Economics of Oleochemicals as Plastics Additives", *World Conference on Oleochemicals, Into the 21th Century*, Kuala Lumpur, Malaysia, Oct. 8–12, 1990, T.H. Applewhite (ed.), AOCS, Champaign, IL, 1991, pp. 39–44.
262. N. O. V. Sonntag, *Inform* **2** (1991) 449–463.
263. G. J. Piazza, T. A. Foglia, *Eur. J. Lipid Sci. Technol.* **103** (2001) 450–454.
264. Henkel KGaA, EP 0 902 006 A2, 1999.
265. B. Dahlke, S. Hellbardt, M. Peatow, W. H. Zech, *J. Am. Oil Chem. Soc.* **72** (1995) 349–353.
266. H. Niewiadomski, H. Szczepanska: *By-Products and Waste Materials in Fat Technology*, Ellis Horwood and Polish Scientific Publishers PWN, Warszawa 1995.
267. The Use of Fatty Acids And its Co-products as Ingredients in Compound Feedingstuffs, www.apag.org.
268. Verfütterungsverbotsgesetz–VerfVerbG, German Federal Ministry of Agriculture, update of April 2001, www.dvtiernahrung.de.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Fatty Alcohols

Standard Article

Klaus Noweck[1] and Wolfgang Grafahrend[1]
[1]Lurgi AG, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a10_277.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Saturated Fatty Alcohols
2.1. Physical Properties
2.2. Chemical Properties
2.3. Production from Natural Sources
2.3.1. Hydrolysis of Wax Esters
2.3.2. Reduction of Wax Esters with Sodium
2.3.3. Hydrogenation of Natural Raw Materials
2.3.3.1. Raw Materials and Pretreatment
2.3.3.2. Hydrogenation Processes
2.4. Synthesis from Petrochemical Feedstocks
2.4.1. Ziegler Alcohol Processes
2.4.2. Oxo Process
2.4.3. Hydrogenation of Fatty Acids produced by Oxidation of Paraffinic Hydrocarbons
2.4.4. Bashkirov Oxidation
2.4.5. Other Processes
2.5. Uses
3. Unsaturated Fatty Alcohols
4. Guerbet Alcohols
5. Bifunctional Fatty Alcohols
6. Quality Specifications
7. Storage and Transportation
8. Economic Aspects
9. Ecotoxicology and Toxicology
9.1. Ecotoxicology and Environmental Aspects
9.2. Toxicology

[Top of Page]

1. Introduction

Fatty alcohols are aliphatic alcohols with chain lengths between C_6 and C_{22}:

$$CH_3(CH_2)_nCH_2OH(n = 4 - 20)$$

They are predominantly linear and monohydric, and can be saturated or have one or more double bonds. Alcohols with a carbon chain length above C_{22} are referred to as wax alcohols. Diols whose chain length exceeds C_8 are regarded as substituted fatty alcohols. The character of the fatty alcohols (primary or secondary, linear or branched-chain, saturated or unsaturated) is determined by the manufacturing process and the raw materials used. Natural products, such as fats, oils, and waxes, and the Ziegler alcohol process provide linear, primary, and even-numbered alcohols; those obtained from

natural sources may be unsaturated. In contrast, the oxo process yields 20 – 60 % branched fatty alcohols, and also some odd-numbered ones. Guerbet dimerization results in α-branched, primary alcohols, whereas Bashkirov oxidation yields secondary alcohols.

Depending on the raw materials used, fatty alcohols are classified as natural or synthetic. Natural fatty alcohols are based on renewable resources such as fats, oils, and waxes of plant or animal origin, whereas synthetic fatty alcohols are produced from petrochemicals such as olefins and paraffins. Up to 1930, when catalytic high-pressure hydrogenation was developed ADKINS and FOLKER [7], NORMAN [8], SCHRAUTH [9], and SCHMIDT [10], the manufacture of fatty alcohols was based almost exclusively on the splitting of sperm oil. By 1962, the world production capacity from natural raw materials had grown to ca. 200 000 t/a. New processes utilizing petrochemical raw materials, e.g., the Ziegler alcohol process, the SHOP process, the oxo process, and the construction of additional plants for high-pressure hydrogenation of natural raw materials, allowed a further increase. In 2005, the world nameplate production capacity of fatty alcohols was estimated to be 2.5×10^6 t/a, being nearly equally based on natural and petrochemical feedstocks. This ratio will change because 600 000 – 700 000 t/a of additional production capacity based on natural feedstocks will come on stream by 2008. Production and consumption were estimated to amount to only 80 % or less of capacity. Fatty alcohols and their derivatives are used in polymers, surfactants, oil additives, and cosmetics and have many specialty uses.

[Top of Page]

2. Saturated Fatty Alcohols

2.1. Physical Properties

Saturated fatty alcohols up to dodecanol are clear, colorless liquids. The next higher homologues are soft materials; tetradecanol and higher alcohols have a waxy consistency. The saturated alcohols crystallize in a nearly orthorhombic lattice [11] and all have a lower specific density than water. The lower members of the series have characteristic odors; the higher fatty alcohols are odorless, except for traces of impurities such as carbonyl compounds and hydrocarbons, which are usually present.

Physical properties of linear, primary fatty alcohols are summarized in Table 1 (for additional data on pure alcohols and commercially important blends, see [12-15]).

Table 1. Physical and chemical properties of fatty alcohols

IUPAC name	Common name	CAS registry number	Molecular formula	M_r	Hydroxyl number	*mp*, °C	*bp*, °C (*p*, kPa)	Density, g/cm³ (*t*, °C)	Refractive index (*t*, °C)
1-Hexanol	caproic alcohol	[111-27-3]	$C_6H_{14}O$	102.2	548	–52	157	0.819 (20)	1.4181 (20)
1-Heptanol	enanthic alcohol	[111-70-6]	$C_7H_{16}O$	116.2	482	–30	176	0.822 (20)	1.4242 (20)
1-Octanol	caprylic alcohol	[111-87-5]	$C_8H_{18}O$	130.2	430	–16	195	0.825 (20)	1.4296 (20)
1-Nonanol	pelargonic alcohol	[143-08-8]	$C_9H_{20}O$	144.3	388	– 4	213	0.828 (20)	1.4338 (20)
1-Decanol	capric alcohol	[112-30-1]	$C_{10}H_{22}O$	158.3	354	7	230	0.829 (20)	1.4371 (20)
1-Undecanol		[112-42-5]	$C_{11}H_{24}O$	172.3	326	16	245	0.830 (20)	1.4402 (20)
1-Dodecanol	lauryl alcohol	[112-53-8]	$C_{12}H_{26}O$	186.3	300	23	260	0.822 (40)	1.4428 (20)
1-Tridecanol		[112-70-9]	$C_{13}H_{28}O$	200.4	280	30	276		
1-Tetradecanol	myristyl alcohol	[112-72-1]	$C_{14}H_{30}O$	214.4	261	38	172 (2.67)	0.823 (40)	1.4358 (50)
1-Pentadecanol		[629-76-5]	$C_{15}H_{32}O$	228.4	245	44			1.4408 (50)
1-Hexadecanol	cetyl alcohol	[36653-82-4]	$C_{16}H_{34}O$	242.5	230	49	194 (2.67)	0.812 (60)	1.4392 (60)
1-Heptadecanol	margaryl alcohol	[1454-85-9]	$C_{17}H_{36}O$	256.5	218	54			
1-Octadecanol	stearyl alcohol	[112-92-5]	$C_{18}H_{38}O$	270.5	207	58	214 (2.67)	0.815 (60)	1.4388 (60)
1-Nonadecanol		[1454-84-8]	$C_{19}H_{40}O$	284.5	196	62			1.4328 (70)
1-Eicosanol	arachidyl alcohol	[629-96-9]	$C_{20}H_{42}O$	298.6	187	64	215 (1.33)	0.806 (70)	
1-Heneicosanol		[		312.6	179	68			

		15594-90-8]	$C_{21}H_{44}O$					
1-Docosanol	behenyl alcohol	[661-19-8]	$C_{22}H_{46}O$	326.6	171	71	241 (1.33)	0.807 (80)
1-Tricosanol		[3133-01-5]	$C_{23}H_{48}O$	340.6	164	74		
1-Tetracosanol	lignoceryl alcohol	[506-51-4]	$C_{24}H_{50}O$	354.7	158	77		
1-Pentacosanol		[26040-98-2]	$C_{25}H_{52}O$	368.7	152	78		
1-Hexacosanol	ceryl alcohol	[506-52-5]	$C_{26}H_{54}O$	382.7	146	81		
1-Heptacosanol		[2004-39-9]	$C_{27}H_{56}O$	396.8	141	82		
1-Octacosanol	montanyl alcohol	[557-61-9]	$C_{28}H_{58}O$	410.8	136	84		
1-Nonacosanol		[6624-76-6]	$C_{29}H_{60}O$	424.8	132	85		
1-Triacontanol	myricyl alcohol	[593-50-0]	$C_{30}H_{62}O$	438.8	128	87		
1-Hentriacontanol	melissyl alcohol	[544-86-5]	$C_{31}H_{64}O$	452.9	124	87		
1-Dotriacontanol	lacceryl alcohol	[6624-79-9]	$C_{32}H_{66}O$	466.9	120	89		
1-Tritriacontanol		[71353-61-2]	$C_{33}H_{68}O$	480.9	116			
1-Tetratriacontanol	geddyl alcohol	[28484-70-0]	$C_{34}H_{70}O$	494.9	113	92		

Boiling points and melting points increase uniformly with chain length. Both are significantly higher than those of the hydrocarbons with the same number of carbon atoms. The influence of the polarizing hydroxyl function diminishes with increasing chain length. Thus hexanol and even octanol show some water solubility, but decanol and the higher fatty alcohols can be considered as immiscible with water. However, a slight hygroscopicity is observed even with octadecanol and higher fatty alcohols, which can absorb water vapor from air during storage. Common organic solvents such as petroleum ether, lower alcohols, and diethyl ether are suitable solvents for fatty alcohols.

2.2. Chemical Properties

The industrial importance of the fatty alcohols is due to the large number of reactions that the hydroxyl group may undergo. Figure 1 shows some typical examples (see also ⟶ Alcohols, Aliphatic); many of the resulting derivatives are intermediates of commercial importance (see section Uses). Ethoxylation with ethylene oxide yields fatty alcohol polyglycol ethers, which are of major importance for the detergent industry (⟶ Laundry Detergents; ⟶ Surfactants):

$$RCH_2OH + nH_2C\underset{O}{\diagdown\diagup}CH_2 \longrightarrow RCH_2(O-CH_2CH_2)_nOH$$

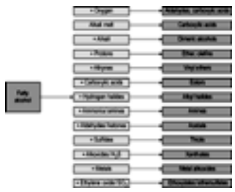

Figure 1. Typical reactions of fatty alcohols

Under normal conditions, fatty alcohols are resistant to oxidation. However, they can be converted into aldehydes or carboxylic acids using strong oxidants or by catalytic oxidation with air or oxygen [16-20]. This reaction is important for the synthesis of C_6 – C_{10} aldehydes if these are not readily available from natural sources [21] and are therefore produced from synthetic alcohols.

2.3. Production from Natural Sources

Two groups of natural raw materials are used for the production of fatty alcohols: (1) fats and oils of plant or animal origin, which contain fatty acids in the form of triglycerides that can be hydrogenated after suitable pretreatment (⟶ Fats and Fatty Oils – Refining) to yield fatty alcohols; and (2) wax esters from whale oil (sperm oil), from which the fatty alcohols are obtained by simple hydrolysis or reduction with sodium. The commercial exploitation of sperm oil has led to depletion of whale populations and is banned in some countries. Attention has therefore turned to the jojoba plant whose oil also consists of wax esters. Successful attempts to cultivate this desert shrub and develop it as a source of raw material have been made. But this wax is mainly used in cosmetic applications in a small amount.

Most fatty chemicals obtained from natural sources have chain lengths of C_{16} – C_{18}. The limited availability of compounds with 12 – 14 carbon atoms, which are important in surfactants, was one of the driving forces behind the development of

petrochemical processes as well as the intensification of plantation and the breeding of new crops for the production of fatty alcohols (see Section Synthesis from Petrochemical Feedstocks). Higher alcohols, such as C_{20} – C_{22} alcohols, can be produced from rapeseed oils rich in erucic acid and fish oils. Unsaturated fatty alcohols may be manufactured in the presence of selective catalysts. Hydroxyacids for production of diols are available from castor oil. For detailed reviews of large-scale industrial processes for the production of fatty alcohols, see [1], [22], [23].

2.3.1. *Hydrolysis of Wax Esters*

The hydrolysis of wax esters is of minor importance today. It is carried out by heating sperm oil with concentrated sodium hydroxide at ca. 300 °C and distilling the alcohol from the sodium soap.

$$R^1-\overset{\overset{\displaystyle O}{\|}}{C}-OR^2 + NaOH \longrightarrow R^1-\overset{\overset{\displaystyle O}{\|}}{C}-ONa + R^2OH$$

The distillate consists of partially unsaturated C_{16} – C_{20} alcohols, which are hardened by catalytic hydrogenation to prevent autoxidation. Since sperm oil contains only ca. 70 % wax esters, the alcohol yield is about 35 %.

2.3.2. *Reduction of Wax Esters with Sodium*

The reduction of esters with sodium was first described in 1902 by BOUVEAULT and BLANC (for a review, see [24]). Large-scale application of this process was achieved in 1928 (Dehydag).

$$\begin{aligned} &R^1-\overset{\overset{\displaystyle O}{\|}}{C}-OR^2 + 4\,Na + 2\,R^3OH \longrightarrow \\ &R^1CH_2ONa + R^2ONa + 2\,R^3ONa \xrightarrow[-4\,NaOH]{+4\,H_2O} \\ &R^1CH_2OH + R^2OH + 2\,R^3OH \end{aligned}$$

Molten sodium is dispersed in an inert solvent and the carefully dried ester and alcohol are added. When the reaction is complete, the alkoxides are split by stirring in water, and the alcohols are washed and distilled [25].

The added alcohol R^3OH, preferably a secondary alcohol, acts as a hydrogen donor. Because of side reactions, the consumption of sodium can be as much as 20 % above the stoichiometric requirement.

The reduction proceeds selectively without production of hydrocarbons and isomerization or hydrogenation of double bonds. Extensive safety measures are required due to the large quantity of metallic sodium used. The process was used until the 1950s to produce unsaturated fatty alcohols, especially oleyl alcohol from sperm oil. These alcohols can now be produced by selective catalytic hydrogenation processes using cheap raw materials, and the sodium reduction process is of interest only in special cases. A fully continuous plant with a production capacity of 3600 t/a was built in Japan in 1973 for the reduction of sperm oil [26].

2.3.3. *Hydrogenation of Natural Raw Materials*

2.3.3.1. *Raw Materials and Pretreatment*

For the production of C_{12} – C_{14} alcohols, preferably coconut oil and palm kernel oil are used. Rape seed oil, palm oil, soybean oil, and tallow are the main sources for C_{16} – C_{18} alcohols. Rapeseed oil rich in erucic acid yields fatty alcohols with 20 or 22 carbon atoms. Bifunctional fatty alcohols can be obtained from castor oil and other special oils (see Chap. Bifunctional Fatty Alcohols). More than 90 % of the 125 × 106 t/a of vegetable oils are used in food applications.

Before hydrogenation, contaminants such as phosphatides, sterols, or oxidation products and impurities such as seed particles, dirt, and water are removed in a cleaning stage, which includes refining by treatment with phosphoric acid, centrifugation, and adsorption, e.g., on charcoal or bentonite (⟶ Fats and Fatty Oils – Refining).

The refined triglycerides are then hydrolyzed to yield fatty acids (⟶ Fatty Acids) or transesterified with lower alcohols to yield fatty acid esters.

Refined free fatty acids, wax esters, and fatty acid methyl esters are used for hydrogenation. Direct hydrogenation of triglycerides is also possible; however, under the reaction conditions, glycerol is reduced to propylene glycol and propanol and therefore makes no commercial contribution as a byproduct. More hydrogen is needed and catalyst costs increase. Therefore, triglyceride hydrogenation is not used industrially.

Fatty acid esters are produced either by esterification of free fatty acids or by transesterification of triglycerides (see also ⟶ Esters, Organic).

Esterification of Fatty Acids. Esterification is an equilibrium reaction:

$$R^1-CH_2-\overset{\overset{\displaystyle O}{\|}}{C}-OH + R^2OH \rightleftharpoons R^1-CH_2-\overset{\overset{\displaystyle O}{\|}}{C}-OR^2 + H_2O$$

Excess alcohol or removal of water shifts the equilibrium toward ester formation. Industrial esterification is carried out in a column at 200 – 250 °C under slight pressure and with excess methanol. Some processes use a catalyst like an ion

exchanger [27]. Distilled fatty acids, which no longer have the composition of the original natural product, are predominantly used. Methanol reacts with the fatty acid in a countercurrent. The acid number (milligrams of KOH needed to neutralize one gram of substance) is used as a quality characteristic and for process control. Batch and continuous processes give esters of similar quality, but the continuous process uses less methanol and the residence time is reduced. The methyl ester is subsequently distilled for purification. Another uncatalyzed process esterifies the distilled fatty acid with an excess of fatty alcohols to wax esters at about 200 – 250 °C [28].

Transesterification of Triglycerides. This reaction is carried out continuously with alkaline catalysts. Like esterification, transesterification is an equilibrium reaction and is shifted toward the desired ester by excess methanol or removal of glycerol:

$$\begin{array}{l} CH_2OCOR \\ | \\ CHOCOR \\ | \\ CH_2OCOR \end{array} + 3\,CH_3OH \rightleftharpoons \begin{array}{l} CH_2OH \\ | \\ CHOH \\ | \\ CH_2OH \end{array} + 3\,RCOOCH_3$$

If the reaction is carried out under mild conditions (50 – 70 °C, atmospheric pressure, excess methanol), free fatty acids present in the oils must first be removed or pre-esterified, e.g., with glycerol. Under more severe conditions (i.e., at 9 – 10 MPa and 220 – 250 °C), pre-esterification is unnecessary, and less pure, cheaper raw materials can be used. Disadvantages of the latter process are the need for high-pressure equipment, a greater excess of methanol, and the energy-intensive further processing of the aqueous methanol. The methyl esters are purified by distillation.

Figure 2 demonstrates the production tree of physical and chemical processing of fats and oils.

Figure 2. Physical and chemical processing of fats and oils

2.3.3.2. *Hydrogenation Processes*

Three large-scale hydrogenation processes are used commercially:

1. Gas-phase hydrogenation
2. Trickle-bed hydrogenation
3. Suspension hydrogenation

The first two processes use a fixed-bed catalyst. In contrast to gas-phase and trickle-bed hydrogenation the suspension process uses a powdered catalyst with a specific particle size distribution. The selection of a process depends on the integration into existing plants and the choice of raw materials. In all cases, hydrogenation is carried out with copper-containing, mixed-oxide catalysts at 170 – 270 °C and 4 – 30 MPa. By using proprietary fixed bed catalysts methyl esters can be hydrogenated at 170 – 250 °C and 10 – 40 MPa in a trickle bed [31], [30] or in the gas phase (Fig. 12) [27]. The wax ester process operates at 170 – 270 °C and from less than 10 to 30 MPa (Fig. 13) [28].

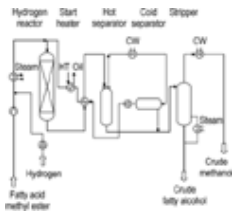

Figure 12. Production of fatty alcohols by hydrogenation of fatty acid methyl esters

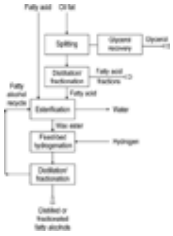

Figure 13. Wax ester route to fatty alcohols

Suspension Hydrogenation. This process is applicable to fatty acid methyl esters as well as to fatty acids.

Hydrogen (ca. 50 mol per mole of ester) and the heated methyl ester are fed separately into the bottom of a narrow reactor. The reaction is carried out at ca. 25 MPa and 250 – 300 °C in the presence of a fine powdery copper catalyst; the LHSV (liquid hourly space velocity) is approximately 1. The excess hydrogen serves to circulate the reaction mixture. The product mixture is separated into a gas phase, which is recycled to the reactor, and a liquid phase, from which the methanol is stripped. The crude product, which still has a saponification number of 6 – 10 (milligrams of KOH required to saponify one gram of substance), is distilled off after removal of the catalyst. If a stainless steel reactor is used, this process can be applied to the direct hydrogenation of fatty acids. In this case, an acid-resistant catalyst is required, and catalyst consumption is increased.

A variant that is particularly suitable for the hydrogenation of fatty acids has been developed by Lurgi (Fig. 3) [31], [32] [34] and is used by several manufacturers. This process employs a large excess of fatty alcohol in the hydrogenation reactor. Hydrogen, fatty alcohol – catalyst slurry, and fatty acids are fed separately into the reactor; the ester forms almost instantaneously [34] and is then hydrogenated in the same reactor in a slower, second reaction step. Hydrogenolysis is carried out at ca. 30 MPa and 260 – 300 °C. Catalyst consumption is 3 – 7 kg/t fatty acid depending on the acid and the concentration of catalyst-poisoning impurities like sulfur, chloride, and phoshorus. The catalyst is separated by centrifugation, and the crude fatty alcohol, which has an acid number of < 0.1 and a saponification number of 2 – 5, is purified by distillation. The selectivity of the process is > 99 %.

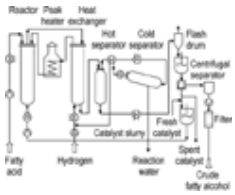

Figure 3. Suspension hydrogenation of fatty acids (Lurgi process)

By continuously replacing part of the spent catalyst, the activity of the copper chromite contact can be held constant.

Gas-Phase Hydrogenation. This process requires a vaporized substrate and is therefore particularly suitable for methyl esters, preferably with a chain length of 12 – 18 C-atoms [34]. Characteristics of the process are a large excess of recycle gas (ca. 600 mol of H_2 per mole of ester), high gas velocities and the addition of methanol to aid evaporation [35], [36]. Decomposition of methanol creates significant quantities of carbon monoxide, water, and dimethyl ether. Admixture of an inert gas to the hydrogen is claimed to make the addition of methanol superfluous and to reduce the excess of recycle gas [37]. Catalysts like copper–zinc or copper–chromium mixed oxides are used in a fixed bed. The conditions required are 4 – 10 MPa pressure and 170 – 230 °C, with an LHSV of about 0.3. Alcohol yields of > 99 % are achieved. Catalyst consumption is < 0.8 kg/t feed. The product mixture is separated into a gas and a liquid phase, the hydrogen recycled and the methanol stripped from the fatty alcohol.

Trickle-Bed Hydrogenation. In this process, the products to be reduced are used in their liquid form. The process is therefore also suitable for methyl esters [31] and non-vaporizable substrates such as wax esters and [28]. Corrosive effects of acids can be neutralized by hydrogenation in the presence of amines [38]. If a considerably lower excess of recycle gas (ca. 100 mol of H_2 per mole of ester) is used, a different plant design is necessary. The reaction is carried out at <10 – 30 MPa and about 170 – 230 °C, with an LHSV of 0.5 – 0.8. Normally, catalysts based on copper, chromium- or copper-zinc mixed oxides are employed. More rarely, supported catalysts such as copper chromite on silica are used. Catalyst consumption is < 0.8 kg/t feed. Further treatment of the product is identical to that described for gas-phase hydrogenation.

Comparison of Hydrogenation Processes. In the case of the fixed-bed processes (gas-phase and trickle-bed hydrogenation), the catalyst need not be separated from the crude fatty alcohol. However, a gradual decrease in the hydrogenating activity due to catalyst poisons such as sulfur, phosphorus, or chlorine is observed, whereas continuous replacement of the catalyst in the suspension process ensures constant activity. If methyl esters are employed, separation and further processing of the methanol is necessary.

Catalysts that contain noble metals, especially rhenium, may allow hydrogenation at lower pressures, which would reduce capital and operating costs [39-41].

2.4. Synthesis from Petrochemical Feedstocks

2.4.1. *Ziegler Alcohol Processes*

Two processes for the production of synthetic fatty alcohols are based on the work of ZIEGLER on organic aluminum compounds: the Alfol process, developed by Conoco, and Ethyl Corporation's Epal process [42], [43]. Fatty alcohols synthesized by these processes are structurally similar to natural fatty alcohols and are thus ideal substitutes for natural products.

Conoco started the first Alfol plant in the United States in 1962. This plant is now operated by Sasol Northe America. In 1964 Condea Chemie (now Sasol) installed a similar plant in Brunsbüttel, Federal Republic of Germany. Ethyl Corporation (today BP/Amoco) developed its own process (Epal process) and began operations in 1964. Additional Alfol alcohol plants were built in Ufa/ Russia in 1981 and in Jilin/ China in 1998.

Alfol process. Figure 4 shows a simplified diagram of the Alfol process [44]. A hydrocarbon is used as solvent. The process involves five steps: hydrogenation, ethylation, growth reaction, oxidation, and hydrolysis.

Figure 4. Alfol alcohol process

1. Hydrogenation

$$2\ Al(CH_2CH_3)_3 + Al + 1\tfrac{1}{2}\ H_2 \longrightarrow 3\ HAl(CH_2CH_3)_2$$

2. Ethylation

$$3\ HAl(CH_2CH_3)_2 + 3\ H_2C{=}CH_2 \longrightarrow 3\ Al(CH_2CH_3)_3$$

3. Growth Reaction

$$3\ Al(CH_2CH_3)_3 + (x+y+z)\ H_2C{=}CH_2 \longrightarrow Al\begin{cases}(CH_2CH_2)_{x+1}H\\(CH_2CH_2)_{y+1}H\\(CH_2CH_2)_{z+1}H\end{cases}$$

4. Oxidation

$$Al\begin{cases}R^1\\R^2\\R^3\end{cases} + 1\tfrac{1}{2}\ O_2 \longrightarrow Al\begin{cases}OR^1\\OR^2\\OR^3\end{cases}$$

5. Hydrolysis

$$Al\begin{cases}OR^1\\OR^2\\OR^3\end{cases} + 2\ H_2O \longrightarrow AlO(OH) + R^1OH + R^2OH + R^3OH$$

Two-thirds of the triethylaluminum produced in the ethylation reaction (b) are recycled to the hydrogenation stage (a), and one-third enters the growth reaction (c). Insertion of the ethylene molecule into the aluminum – carbon bonds occurs as a statistical process and leads to a broad distribution (Poisson distribution [45], [46]) of chain lengths, ranging from C_2 to beyond C_{26} [44]. An optimal yield of the C_{12} – C_{14} alcohols, which are important in the surfactant sector, requires addition of about four molecules of ethylene per aluminum – carbon bond (see also Table 2). A small percentage of olefins are formed as byproducts.

Table 2. Typical alcohol distributions in Alfol and Epal processes

	Distribution, %	
Alcohol	**Alfol process**	**Epal process**
Ethanol	0.5	traces
Butanol	3.4	0.1
Hexanol	9.5	1.5
Octanol	16.1	3.5
Decanol	19.5	8.0
Dodecanol	18.4	34.0
Tetradecanol	14.1	26.0
Hexadecanol	9.1	16.0
Octadecanol	5.1	8.8
Eicosanol	2.5	1.9
Docosanol	1.1	0.2

Because of the varying reactivity of partially oxidized trialkylaluminum compounds, oxidation is carried out stepwise by

passing through carefully dried air. Cooling is necessary, especially at the start of the reaction. Alkanes and oxygen-containing compounds are formed as byproducts [42].

Prior to hydrolysis, the solvent is removed by distillation. Hydrolysis with water gives high-purity hydrated alumina as a coproduct, which has many industrial applications, e.g., in catalytic processes (⟶ Aluminum Oxide – Catalytic Applications) and in ceramics. In the 1960s, hydrolysis was carried out with hot sulfuric acid at Conoco's and Ethyl Corporation's plants. Conoco changed to neutral hydrolysis, but the sulfuric acid method was used in the Epal process until its shutdown and leads to high-purity aluminum sulfate as a coproduct. The crude alcohols are finally fractionated into marketable blends and single cuts.

Epal Process (see Fig. 5). Many attempts have been made to achieve a narrower distribution of chain lengths in the growth reaction [43]. The only process that has been used on an industrial scale is the Epal process developed by the Ethyl Corporation (today BP/Amoco). The reaction steps resemble those of the Alfol process, but the growth reaction is not carried as far. The product of the growth reaction is subjected to transalkylation (290 °C, 3.5 MPa) with $C_4 – C_{10}$ olefins. The chain lengths of the resulting trialkylaluminum compounds are predominantly $C_4 – C_{10}$. Excess olefins are removed in a stripping column and then fractionated. The trialkylaluminum compound is subjected to a second growth reaction and then transalkylated (200 °C, 35 kPa) with $C_{12} – C_{18}$ olefins. Again the olefins are separated in a stripper and fractionated. At this stage, the trialkylaluminum compound consists largely of alkyl chains with 12 – 18 carbon atoms.

Figure 5. Flow diagram of the Epal process

The alcohol distributions achieved in the Epal and Alfol processes are shown in Table 2 [47]. The Epal process offers greater flexibility than the Alfol process because both the alcohols and the α-olefins that are formed as intermediates can be marketed [48]. Disadvantages of the Epal process are the higher capital and operating costs, a considerably more complicated process control, and an increased proportion of branched-chain olefins and alcohols.

2.4.2. *Oxo Process*

The oxo process (hydroformylation) consists of the reaction of olefins with an H_2– CO gas mixture in the presence of suitable catalysts. This reaction was discovered in 1938 by O. ROELEN at Ruhrchemie during work on the Fisher – Tropsch synthesis. Although the first production plant was constructed before 1945 [49], the oxo process achieved industrial importance only in the 1950s, with the increasing demand for plasticizers and detergents. The process is based on the following reaction (for details on the reaction mechanism, see [50] and ⟶ Oxo Synthesis):

$$2R-CH=CH_2 + 2CO + 2H_2 \xrightarrow{HCo(CO)_4} R-CH_2CH_2-CHO + R-\overset{\overset{CH_3}{|}}{CH}-CHO$$

α-Olefins yield approximately equal amounts of straight-chain and branched aldehydes (see also Fig. 10). Internal and branched alkenes can also be used in this reaction. Internal olefins give a product containing some primary aldehyde because the catalyst effects double-bond isomerization.

For a long time paraffin-based processes were predominant as a source of olefins, especially for detergents [48], [57]. With the development of the SHOP process (Shell's Higher Olefin Process), ethylene has become the preferred raw material [58]. The principal steps in the SHOP process are ethylene oligomerization, isomerization, and metathesis (⟶ Hydrocarbons). The products are $C_{12} – C_{18}$ α-olefins and $C_{11} – C_{14}$ internal olefins, which are all important in the area of surfactants [59], [60].

Heterogeneous hydrogenation of the oxo aldehydes at 5 – 20 MPa and 150 – 250 °C in the presence of catalysts based on nickel, molybdenum, copper, or cobalt yields the corresponding alcohols. Alternatively, the aldehyde can be subjected to an aldol reaction as in the production of 2-ethylhexanol (⟶ 2-Ethylhexanol – Industrial Production). A flow sheet of the oxo process is depicted in Figure 6. Worldwide there are three variants of the oxo process [52]: (1) the classical process using $HCo(CO)_4$ as catalyst; (2) the Shell process based on a cobalt carbonyl – phosphine complex [51]; and (3) a process using a rhodium catalyst. The key parameters of these processes are compared in Table 3.

Table 3. Typical process parameters of the oxo processes

Parameter	Oxo process		
	Classical	**Shell**	**Union Carbide**
Catalyst	cobalt carbonyl	cobalt carbonyl – phosphine complex complex	rhodium carbonyl – phosphine complex phosphine complex

Catalyst concentration, %	0.1 – 1.0	ca. 0.5	0.01 – 0.1
CO : H_2 ratio	1 : 1 – 1.2	1 : 2 – 2.5	excess hydrogen
Temperature, °C	150 – 180	170 – 210	100 – 120
Pressure, MPa	20 – 30	5 – 10	2 – 4
LHSV*	0.5 – 1.0	0.1 – 0.2	0.1 – 0.25
Primary products	aldehydes	alcohols	aldehydes
Linearity, %	40 – 50	80 – 85	ca. 90 (*n*-butanol)

* Liquid hourly space velocity

The classical, cobalt-catalyzed oxo process involves the following steps: oxo reaction, catalyst separation and regeneration, aldehyde hydrogenation, and alcohol distillation. Variants of this process differ mainly in the catalyst separation and regeneration steps [53], [54].

In the Shell process, alcohols are obtained directly because of the greater hydrogenating activity of the catalyst; the aldehyde hydrogenation step is unnecessary. Linearity is improved and the 2-methyl isomer is the main byproduct. A disadvantage is the loss of olefins due to hydrogenation to alkanes.

Hydroformylation based on a rhodium catalyst has been used by Union Carbide since the 1970s for the production of *n*-butanol and 2-ethylhexanol [55]. The higher activity of the catalyst enables operation at lower temperature and pressure. For butanol, the linearity is over 90 %. The disadvantage is the high price of rhodium.

For a review of the oxo synthesis, see [56].

Figure 6. Flow diagram of the classical oxo process

2.4.3. *Hydrogenation of Fatty Acids produced by Oxidation of Paraffinic Hydrocarbons*

The process for the oxidation of paraffinic hydrocarbons, developed in Germany before 1940, is used on an industrial scale in Russia for the manufacture of fatty acids. About 5 – 10 % of these synthetic fatty acids are converted to fatty alcohols. The products are mainly linear, primary alcohols, with 5 – 15 % branched alcohols. An overview is given in [61].

A mixture of paraffins is oxidized above 100 °C in the presence of manganese catalysts. The complex product mixture consists of aldehydes, ketones, esters, carboxylic acids, and other compounds. Since the byproducts cannot be completely removed during further processing and distillation of the carboxylic acids, the uses of fatty alcohol produced by this method are limited. As with the natural fatty acids, hydrogenation is carried out after esterification with methanol or butanol. The suspension hydrogenation process is used (see Suspension Hydrogenation.). The distillation residue contains esters of C_{10} – C_{20} alcohols, 25 % of which are secondary. These alcohols can be obtained by hydrolysis.

2.4.4. *Bashkirov Oxidation*

A variant of paraffinic hydrocarbon oxidation was developed in the 1950s in the Soviet Union by BASHKIROV [1], [3]. The paraffins are oxidized in the presence of boric acid, which scavenges the hydroperoxides that are formed as intermediates. This results in the formation of boric acid esters of secondary alcohols. These esters are relatively stable to heat and oxidation. Hydrolysis leads to a statistical distribution of secondary alcohols in which the hydroxyl function may occupy any position on the carbon chain.

Industrial oxidation is carried out at about 160 °C with a nitrogen – air mixture containing about 3.5 % oxygen [62], [63] (⟶ Alcohols, Aliphatic – Oxidation of Saturated Hydrocarbons). The conversion of paraffins is limited to a maximum of 20 % in order to minimize side reactions. The principle reaction steps are depicted in Figure 7.

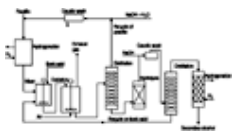

Figure 7. Bashkirov oxidation

The process is used in the former Soviet Union and in Japan (Nippon Shokubai). A plant operated by Union Carbide in the United States since 1964 is closed.

2.4.5. *Other Processes*

Fatty alcohols can also be obtained by reaction of α-olefins with hydroperoxides in the presence of transition-metal catalysts, especially molybdenum [65]:

$$R-CH=CH_2 + R-OOH \longrightarrow R-\underset{\diagdown O \diagup}{CH-CH_2} + ROH$$

If *tert*-butyl hydroperoxide is used, the coproduct isobutanol can be readily separated from the epoxide.

After separation of the low molecular mass alcohol and purification, the epoxide is hydrogenolyzed in the presence of a nickel catalyst to form the primary alcohol. About 10 – 15 % of the secondary alcohol and 2 % paraffin are obtained as byproducts [66].

Mixtures of linear, primary alcohols with average molecular masses between 400 and 700 (corresponding to a chain length of 30 – 50 carbon atoms) are marketed under the trade name Unilin [67]. The hydrocarbon content of these mixtures is about 20 %. Ethylene oligomers are oxidized to produce oxygen-containing products on an industrial scale [68].

1-Triacontanol, which does not occur in the common natural raw materials and is produced only in minute quantities in petrochemical processes such as the Alfol process, is of interest because of its ability to stimulate plant growth [69]. Several research groups investigated synthetic pathways based on cheap raw materials [70-75].

2.5. Uses

Fatty alcohols are mainly employed as intermediates. In Western Europe, only 5 % are used directly and ca. 95 % in the form of derivatives [76].

Surfactants. The amphiphilic character of fatty alcohols, which results from the combination of a nonpolar, lipophilic carbon chain with a polar, hydrophilic hydroxyl group, confers surface activity upon these compounds. Surfactants account for 70 – 75 % of fatty alcohol production [77].

Due to the their amphiphilic properties fatty alcohols orient themselves at phase interfaces and can therefore be used in emulsions and microemulsions. In cosmetic emulsions (creams, lotions) the main function of the fatty alcohols is to provide consistency, in technical emulsions they are used as cosurfactants and solution aids.

If the hydroxyl group of the fatty alcohols is replaced by other, larger hydrophilic groups, the polar character is enhanced and surfactants are obtained [4]. The most important surfactants derived from fatty alcohols are described in the following (see also ⟶ Laundry Detergents, ⟶ Surfactants).

Alkyl polyglycol ethers, fatty alcohol polyglycol ethers, fatty alcohol ethoxylates, $RCH_2(OCH_2CH_2)_nOH$, were the first nonionic surfactants produced on an industrial scale. They are synthesized in a base-catalyzed reaction. Fatty alcohol-alkylene oxide adducts, which are produced in a base-catalyzed reaction of fatty alcohols, ethylene oxide, and propylene oxide, are low-foaming surfactants.

Alkyl sulfates, fatty alcohol sulfates, belong to the group of anionic surfactants and are the longest known synthetic surfactants. In order to synthesize this class of compounds, fatty alcohols are treated with sulfur trioxide, chlorosulfuric acid, oleum, or sulfuric acid. The resulting semiesters are subsequently neutralized with alkali, mostly aqueous NaOH.

Alkyl polyglycol ether sulfates, fatty alcohol ether sulfates, also belong to the group of anionic surfactants. They are produced by reaction of fatty alcohols with ethylene oxide and the resulting adducts are then treated with sulfur trioxide or chlorosulfuric acid. The subsequent neutralization is carried out with caustic soda, ammonia, or ethanolamine.

Alkyl methylammonium chlorides belong to the group of cationic surfactants and are produced using fatty alcohols especially in the range of C_{16-18}.

Alkylpolyglycosides (APG) are prepared from fatty alcohols and sugars following various procedures. They have good skin compatibility and, if necessary, can be manufactured from renewable raw materials only. However, high production costs are usually a limiting factor for their application.

Other Uses. In addition, the polar character of the fatty alcohols allows their use as lubricants. Esters of fatty alcohols with fatty acids, the so-called wax esters, e.g., as lubricants in polymer processing and as raw materials for waxes and creams in technical applications. Esters are used also in cosmetic applications, but here usually liquid products on the basis of unsaturated fatty alcohols and/or fatty acids are preferred.

Acrylic and methacrylic acid esters of fatty alcohols are precursors of polymethacrylates, which are utilized as flow enhancer and viscosity index improvers in oils. Fumaric acid esters have similar applications. Esters of adipic, acelaic, sebacic, trimellitic, citric, and phthalic acid are used as plasticizers for PVC. The various specific applications of specialty plasticizers are based on their chemical structure, their high cold flexibility, compatibility, high thermostability, high aging stability, and low fogging-values in the finished product. Each of the mentioned esters offers special chemical and physical properties for the different application areas.

A further high volume application is in the area of fragrances and flavors. Aldehydes, esters, or alkyl groups on the basis of fatty alcohols are sometimes incorporated into more complex molecules like cinnamon aldehyde. Important for this application are high-purity single cuts which can be processed into these products without any loss in quality.

Alcohols with a chain length of C_{8-10} are used as so-called tobacco sucker control agents in the cultivation of tobacco.

Especially fatty alcohols with chain lengths $\geq C_{18}$ — either in an emulsified form or as derivatives — are suitable for protection of water reservoirs against evaporation during dry seasons. If fatty alcohols are emulsified with interface-active substances, the resulting emulsion can be applied to the water surfaces. The layer, which is formed fast and easily on the

water surface, prohibits the evaporation and therefore protects the water reserves. Field tests in Europe and in Africa have proved the effectiveness. The evaporation can be reduced to up to 50 %.

The same higher fatty alcohols can also be used as defoamer systems for papers to be produced in modern, fast operating paper machines. Products on the basis of fatty alcohols show the following advantages compared to silicon-based defoamers: high effectiveness, good biodegradability, good drainage and deaeration, water recycling, and high cost effectiveness. Each paper quality and each production facility demands different defoamer qualities, which contain emulsifiers and additives in addition to the alcohol.

[Top of Page]

3. Unsaturated Fatty Alcohols

Unsaturated fatty alcohols are special products and can only be obtained from natural sources; petrochemical processes for their manufacture do not exist. Unsaturated fatty alcohols contain at least one olefinic group in addition to the hydroxyl function. Therefore they can react both as alcohols and as olefins. The physical properties of the most important unsaturated fatty alcohols are listed in Table 4. The melting points are below those of the corresponding saturated alcohols and are influenced by the configuration of the double bond.

Table 4. Physical and chemical properties of primary unsaturated fatty alcohols

IUPAC name	Common name	CAS registry number	Molecular formula	M_r	Hydroxyl number	Iodine number	*mp*, °C	*bp*, °C (*p*, kPa)	Density, g/cm^3 (*t*, °C)	Refractive index (*t*, °C)
10-Undecen-1-ol		[112-43-6]	$C_{11}H_{22}O$	170.3	329	149	–2	133 (2.1)	0.8495 (15)	1.4509 (20)
(*Z*)-9-Octadecen-1-ol	oleyl alcohol	[143-28-2]	$C_{18}H_{36}O$	268.4	209	95	–7.5	208 – 210 (2.0)	0.8489 (20)	1.4606 (20)
(*E*)-9-Octadecen-1-ol	elaidyl alcohol	[506-42-3]	$C_{18}H_{36}O$	268.4	209	95	36 – 37	216 (2.4)	0.8388 (20)	1.4552 (40)
(*Z*,*Z*)-9,12-Octadecadien-1-ol	linoleyl alcohol	[506-43-4]	$C_{18}H_{34}O$	266.5	211	191	–5 to –2	153 – 154 (0.4)	0.8612 (20)	1.4782 (20)
(*Z*,*Z*,*Z*)-9,12,15-Octadecatrien-1-ol	linolenyl alcohol	[506-44-5]	$C_{18}H_{32}O$	264.5	212	288		133 (0.27)	0.8708 (25)	1.4775 (25)
(*Z*)-13-Docosen-1-ol	erucyl alcohol	[629-98-1]	$C_{22}H_{44}O$	324.6	173	78	34 – 35	240 – 242 (1.3)	0.8416 (33)	
(*E*)-13-Docosen-1-ol	brassidyl alcohol	[5634-26-4]	$C_{22}H_{44}O$	324.6	173	78	53 – 54	238 – 243 (1.05)		

Production. The first large-scale hydrogenation plant (Henkel) went into operation in the late 1950s. Previously, unsaturated fatty alcohols could be obtained only by hydrolysis of whale oil (Section Hydrolysis of Wax Esters) or by the Bouveault – Blanc reduction (Section Reduction of Wax Esters with Sodium).

Today, a broad range of raw materials based on animal or vegetable fats and oils are available. Both market factors and the degree of unsaturation (iodine number) required in the final product influence the selection.

For products with iodine numbers around 50, cheap beef tallow or vegetable raw materials are available. Products with iodine numbers of 80 – 100 can be produced from technical oleic acid (animal or vegetable based). Polyunsaturated products are undesirable as they tend to autoxidation. For special applications products with higher iodine numbers (> 110) based on sunflower or soybean oils are produced. Unsaturated alcohols with iodine numbers > 150 are produced, e.g., from linseed oil.

The hydrogenation processes described in Section Hydrogenation Processes are suitable for the large-scale production of unsaturated fatty alcohols. The fixed-bed processes are preferred because of the mild reaction conditions. In suspension hydrogenation, the prolonged contact between fatty alcohol and catalyst results in side reactions such as saturation of the double bond and formation of trans isomers, which leads to a higher solidification point and, hence, loss of quality. With polyunsaturated fatty acids, the formation of conjugated double bonds cannot be completely prevented.

Hydrogenation is generally carried out at 250 – 280 °C and a pressure of 20 – 25 MPa. Catalysts include zinc oxide in conjunction with aluminum oxide, chromium oxide, or iron oxide, and possibly other promoters [84-89]; copper chromite whose activity has been reduced by the addition of cadmium compounds; and cadmium oxide on an alumina carrier [90].

Selective hydrogenation can also be carried out in a homogeneous phase with metallic soaps as catalysts. An overview of early catalyst developments is given in [78]; further references can be found in [79-83]. Patent applications of the 1980s indicate great interest in selective catalysts [91-96].

Unsaturated fatty alcohols are produced in the Federal Republic of Germany (Henkel, Ecogreen) and in Japan (New Japan Chemical).

Uses. Unsaturated fatty alcohols are used in detergents, in cosmetic ointments and creams, as plasticizers and defoamers and in textile and leather processing [23], [97-99]. Oleyl alcohol is also used as an additive in petroleum and lubricating oils.

[Top of Page]

4. Guerbet Alcohols

Condensation of primary alcohols at 180 – 300 °C in the presence of alkaline condensation agents leads to primary, α-branched dimeric alcohols. The difference in structure between Guerbet and other branched fatty alcohols is shown in Figure 8. The physical and chemical properties of Guerbet alcohols are listed in Table 5.

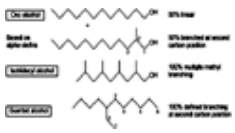

Figure 8. Examples of different types of branched fatty alcohols

Table 5. Physical and chemical properties of Guerbet alcohols

IUPAC name	CAS registry number	Molecular formula	M_r	Hydroxyl number	*mp*, °C	*bp*, °C (*p*, kPa)	Density, g/cm^3 (*t*, °C)	Refractive index (*t*, °C)
2-Methyl-1-pentanol	[105-30-6]	$C_6H_{14}O$	102.18	549		147.9 (101.3)	0.8263 (20)	1.4182 (20)
2-Ethyl-1-hexanol	[104-76-7]	$C_8H_{18}O$	130.23	431	<–76	118 (10.7)	0.8328 (20)	1.4328 (20)
2-Propyl-1-heptanol	[10042-59-8]	$C_{10}H_{22}O$	158.29	354		117 (2.7)		
2-Butyl-1-octanol	[3913-02-8]	$C_{12}H_{26}O$	186.34	301		126 – 128 (1.5)		1.4457 (20)
2-Pentyl-1-nonanol	[5333-48-2]	$C_{14}H_{30}O$	214.39	262		154 (1.7)	0.8352 (24)	1.4460 (24)
2-Hexyl-1-decanol	[2425-77-6]	$C_{16}H_{34}O$	242.45	231	–30 to –26	175 (1.5)	0.8380 (20)	1.4476 (25)
2-Heptyl-1-undecanol	[5333-44-8]	$C_{18}H_{38}O$	270.50	207	–26	198 (2.0)	0.8446 (15)	1.4550 (15)
2-Octyl-1-dodecanol	[5333-42-6]	$C_{20}H_{42}O$	298.56	188	–20	135 – 137 (0.007)	0.8329 (21)	1.4545 (19)
2-Nonyl-1-tridecanol	[54439-52-0]	$C_{22}H_{46}O$	326.61	172		164 – 167 (0.013)	0.8476 (17.5)	1.4582 (17)
2-Decyl-1-tetradecanol	[58670-89-6]	$C_{24}H_{50}O$	354.67	158		173 – 175 (0.007)	0.8413 (17)	1.4606 (17)
2-Undecyl-1-pentadecanol	[79864-02-1]	$C_{26}H_{54}O$	382.72	147				
2-Dodecyl-1-hexadecanol	[72388-18-2]	$C_{28}H_{58}O$	410.77	137		203 – 207 (0.007)		
2-Tridecyl-1-heptadecanol		$C_{30}H_{62}O$	438.83	128				
2-Tetradecyl-1-octadecanol	[32582-32-4]	$C_{32}H_{66}O$	466.88	120	38 – 39	308 – 310 (2.0)		
2-Pentadecyl-1-nonadecanol		$C_{34}H_{70}O$	494.94	113				

2-Hexadecyl-1-eicosanol	[17658-63-8]	$C_{36}H_{74}O$	522.99	107	43 – 45	270 – 280 (0.013)
2-Heptadecyl-1-heneicosanol		$C_{38}H_{78}O$	551.05	102		
2-Octadecyl-1-docosanol		$C_{40}H_{82}O$	579.11	98		
2-Nonadecyl-1-tricosanol		$C_{42}H_{86}O$	607.16	95		
2-Eicosyl-1-tetracosanol	[73761-81-6]	$C_{44}H_{90}O$	635.22	93		

The condensation reaction was first observed by M. GUERBET [100]. Its mechanism and selection of catalysts has been thoroughly studied [3], [5]. Figure 9 illustrates the reaction mechanism in simplified terms.

Figure 9. Mechanism of the Guerbet reaction

Today this reaction is performed on a technical scale with a worldwide production capacity of several thousand tons per year. The alcohol which is to be dimerized is mixed with an alkaline condensation agent and a hydrogenation/dehydrogenation catalyst and then heated to 180 – 300 °C. The water formed in the reaction is distilled off continuously together with the monomeric alcohol. Since the reaction proceeds faster at higher temperatures, especially the reaction of short-chain monomeric alcohols (C_6 – C_{10} alcohols) is performed under pressure to perform it in a suitable temperature range. However, low-molecular, α-branched dimeric alcohols, e.g., 2-ethyl-1-hexanol, can be produced more economically by other methods (⟶ 2-Ethylhexanol).

Physical Properties. The physicochemical properties of Guerbet alcohols can be summarized as follows [101]:

- Melting point or pour point are considerably lower than those of linear alcohols with the same molecular mass.
- Volatility and vapor pressure are significantly lower than those of linear alcohols with comparable consistency (but lower molecular mass).
- As compared with other liquid unsaturated alcohols the Guerbet alcohols are stable against oxidation and autoxidation.

Uses. Based on these properties Guerbet alcohols in the range of C_{12-36} can be used in the following application areas:

- Cosmetic-pharmaceutical oil components with high stability against autoxidation (rancidity)
- Plasticizers for synthetic resins, e.g., nitrocellulose lacquers (for the production of permanent templates)
- Solvents or solution aids for printing inks and specialty inks
- Lubricant components, e.g., for metal processing oils and fiber preparations
- Chemical intermediates, e.g., for production of branched carboxylic acids
- Starting materials for textile auxiliaries, e.g., esters, sulfates, phosphates, ethersulfates

The Guerbet alcohols C_{32-36}, which can be manufactured from C_{16-18}-alcohols, are used as:

- Esterification components for the manufacture of specialty waxes
- Raw materials for cosmetic stick preparations

[Top of Page]

5. Bifunctional Fatty Alcohols

Some natural oils which contain double bonds, hydroxyl groups, or other functional groups can be converted into long-chain diols. These are usually α, ω-diols or diols whose hydroxyl groups lie far apart. Table 6 lists physical properties of bifunctional fatty alcohols.

Table 6. Physical and chemical properties of bifunctional fatty alcohols

IUPAC name	CAS registry number	Molecular formula	M_r	Hydroxyl number	*mp*, °C	*bp*, °C (*p*, kPa)	Density, g/cm^3 (*t*, °C)
1,6-Hexanediol	[629-11-8]	$C_6H_{14}O_2$	118.17	949	42	134 (1.3)	0.9580

						(45)
1,7-Heptanediol	[629-30-1]	$C_7H_{16}O_2$	132.20 849	18	151 (1.9)	0.9570 (25)
1,8-Octanediol	[629-41-4]	$C_8H_{18}O_2$	146.23 767	61	167 – 168 (2.4)	0.9200 (60)
1,9-Nonanediol	[3937-56-2]	$C_9H_{20}O_2$	160.26 700	45	173.2 (1.9)	0.9160 (50)
1,10-Decanediol	[112-47-0]	$C_{10}H_{22}O_2$	174.29 644	73	175 – 176 (1.9)	0.8960 (75)
1,11-Undecanediol	[765-04-8]	$C_{11}H_{24}O_2$	188.31 596	63	178 (1.6)	
1,12-Dodecanediol	[5675-51-4]	$C_{12}H_{26}O_2$	202.34 555	81	183 – 184 (1.25)	
1,13-Tridecanediol	[13362-52-2]	$C_{13}H_{28}O_2$	216.37 519	75 – 76	195 – 197 (1.3)	
1,14-Tetradecanediol	[19812-64-7]	$C_{14}H_{30}O_2$	230.39 487	85	200 (1.2)	
1,15-Pentadecanediol	[14722-40-8]	$C_{15}H_{32}O_2$	244.42 459	70.6 – 71.6	205 – 207 (1.3)	
1,16-Hexadecanediol	[7735-42-4]	$C_{16}H_{34}O_2$	258.45 434	91.4	195 – 200 (0.53)	
1,17-Heptadecanediol	[66577-59-1]	$C_{17}H_{36}O_2$	272.48 412	96 – 96.5	204 – 205 (0.27)	
1,18-Octadecanediol	[3155-43-9]	$C_{18}H_{38}O_2$	286.50 392	97 – 98	210 – 211 (0.27)	
1,19-Nonadecanediol	[7268-65-7]	$C_{19}H_{40}O_2$	300.53 373	101	212 – 214 (0.2)	
1,20-Eicosanediol	[7735-43-5]	$C_{20}H_{42}O_2$	314.56 357	102.4 – 102.6	215 – 217 (0.2)	
1,21-Heneicosanediol	[95008-70-1]	$C_{21}H_{44}O_2$	328.58 342	105 – 105.5	223 – 224 (0.2)	
1,22-Docosanediol	[22513-81-1]	$C_{22}H_{46}O_2$	342.61 328	105.7 – 106.2		
1,23-Tricosanediol	[95491-58-0]	$C_{23}H_{48}O_2$	356.64 314			
1,24-Tetracosanediol	[22513-82-2]	$C_{24}H_{50}O_2$	370.67 302			
1,25-Pentacosanediol	[92238-33-0]	$C_{25}H_{52}O_2$	384.69 292	109		
(*Z*)-9-Octadecene-1,12-diol	[540-11-4]	$C_{18}H_{36}O_2$	284.49 394	89	182/0.06	

1,2-Diols can be produced by epoxidation of internal or α-olefins with subsequent hydrolytic cleavage of the epoxide ring (→ Alcohols, Polyhydric).

Dimerization of unsaturated fatty acids such as soybean, linseed, and tallow fatty acid, followed by esterification and catalytic hydrogenation, gives saturated diols [102].

α, ω-unsaturated fatty alcohols can be produced by pyrolysis of castor oil (Atochem) or by metathesis of oleic acid methyl ester [103].

Thermal or catalytic dimerization of unsaturated alcohols yields viscous dimers with complex structure and an average of two hydroxyl groups per molecule [104], [105].

1,12-Octadecanediol, 1,10-decanediol, and 9-octadecene-1,12-diol are obtained from castor oil by transesterification and hydrogenation, or by alkali splitting, esterification, and hydrogenation. α, ω-Diols with high purity can be produced from the corresponding dicarboxylate esters [106], e.g.,

Suberate → 1,8-Octanediol
Sebacate → 1,10-Decanediol
Dodecanedioate → 1,12-Dodecanediol

Sebacic acid is based on castor oil. Nature offers today new plants and new renewable raw materials for different applications by cultivation or genetic engineering.

Bifunctional fatty alcohols are mainly used in the production of polyesters and polyamides and as intermediates.

[Top of Page]

6. Quality Specifications

Analytical methods defined by DIN [107] and ASTM [108] standards and by the Deutsche Gesellschaft für Fettwissenschaft (DGF) [109] are used for the quality control of fatty alcohols (see Table 7). In addition, analytical methods from the different fatty alcohol suppliers might be used.

Composition of fatty alcohol mixtures is determined by gas chromatography, which can be combined with mass spectrometry. Alcohols from different sources can thus be identified; typical examples are given in Figure 10.

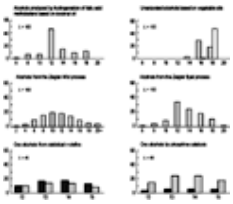

Figure 10. Typical fatty alcohol compositions obtained by different processes

(Blue = *n*-alcohols; Violet = isoalcohols; Yellow = unsaturated alcohols; *L* = linearity)

Table 7. Analytical methods for the characterization of fatty alcohols

	DGF *	DIN	ASTM
Composition		by gas chromatography	
Hydrocarbon content		by column chromatography on silica	
Color (Hazen, APHA)		53409	D 1686–81
Color (Lovibond)	C-IV 4b		D 1209–84
Refractive Index	C-IV 5	53491	D 1218–82
Density	C-IV 2b	51757	D 1298–80
			D 891–59 (1976)
Viscosity	C-IV 7	53015	D 445–83
Solidification point	M-III 4a	51570	D 87–77 (1982)
Boiling range		51751	D 1078–83
			E 133–78 (1984)
Flash point	C-IV 8	51758	D 56–82
			D 93–80
Ignition temperature		51794	D 2155–66 (1976)
Hydroxyl number	C-V 17a/b		D 1957 (1984)
Carbonyl number	C-V 18		E 411
Peroxide number	C-VI 6a		D 1022–76
Iodine number (Kaufmann)	C-V 11b		
Iodine number (Wijs)	C-V 11d		D 1959 (1984)
Acid number	C-V 2	53402	D 1613–81
Saponification number	C-V 3	53401	D 94–80
Water content	C-III 13a	51777	D 1744–83

* DGF = Deutsche Gesellschaft für Fettwissenschaft.

Linearity *L* is defined as the percentage of normal alcohols present in the mixture:

$$L = \frac{n - \text{alcohol}}{n - \text{alcohol} + \text{isoalcohol}} \times 100$$

Coconut alcohol gives very few impurity peaks. It contains < 0.1 % *n*-tridecanol and varying amounts of *n*-alkanes, depending on the hydrogenation process.

Tallow alcohol contains up to 2 % isomeric pentadecanols (isopentadecanol, anteisopentadecanol, and *n*-pentadecanol) and *n*-hexadecane, and up to 4 % isomeric heptadecanols (isoheptadecanol, anteisoheptadecanol, and *n*-heptadecanol) and *n*-octadecane.

Ziegler alcohols (Alfol and Epal alcohols) are primary, straight-chain alcohols with an even carbon number. Gas chromatography shows up to 1 % impurities, consisting of numerous even-numbered, isomeric fatty alcohols.

Oxo alcohols from statistical, cracked olefins or from paraffin dehydrogenation by the Pacol process or SHOP olefins contain

a large number of isomers that cannot be completely separated, even in the most efficient capillary columns. Both odd- and even-numbered isoalcohols occur. Oxo alcohol mixtures can be identified by spectroscopic determination of the degree of branching, i.e., the ratio of *n*-alcohols to isoalcohols, which depends on the production process and the raw materials used.

Fatty acids obtained by *oxidation of paraffinic hydrocarbons* yield even- and odd-numbered primary alcohols with a low degree of branching. Branched alkanes present in the original paraffins are lost by oxidative degradation.

The boric acid-catalyzed process (Bashkirov oxidation) yields secondary alcohols with a statistical distribution of isomers.

Guerbet alcohols are primary, α-branched fatty alcohols with two straight-chain alkyl groups of approximately equal length.

[Top of Page]

7. Storage and Transportation

During production, processing, and storage, fatty alcohols are handled as liquids. Fatty alcohols starting from $C_{12}OH$ solidify at ambient temperatures. Products with a melting point above 40 °C can therefore be transferred to, stored, and handled as flakes or pastilles.

Fatty alcohols are sensitive to oxidation and should be stored under an inert gas; the temperature should not be higher than 20 °C above the melting point. Low-pressure steam or warm water are recommended for heating. Water increases stability to autoxidation; as little as 0.1 % has a stabilizing effect. The use of antioxidants depends on further processing and quality requirements.

Stainless steel or an Al – Mg – Mn alloy (DIN 1725/1745) is used as container material; ordinary steel should have a zinc silicate coating, e.g., Dimetcote. Pumps, valves, and pipes exposed to the product should be made of stainless steel.

Fatty alcohols are considered as flammable materials and are classified according to properties such as flash point and boiling point. Their transport is governed by national and international regulations dealing with volatile and combustible materials.

[Top of Page]

8. Economic Aspects

The economic significance of fatty alcohols is reflected by the worldwide installed production capacities of ca. 2.4×10^6 t/a in 2004, which will increase to ca. 2.8×10^6 t/a by the year 2007. Table 8 shows these production capacities categorized by volume, location and raw materials. In summary, ca. 50 % of the fatty alcohols are produced from natural raw materials, but the split per geographical region is different from continent to continent and will shift to 60 % in favor of natural-based fatty alcohols by 2008. Worldwide production and consumption are estimated at $1.8 – 1.9 \times 10^6$ t/a for 2004 with a yearly increase of 2 – 3 %. In Europe only ca. 5 % of this volume is used directly as fatty alcohols. A share of 70 – 75 % is used in surfactants exclusively [76], [77], [110-114].

Table 8. Estimated production capacities for fatty alcohols in 2005, 10^3 t

Continent	Natural alcohols, ≥ C_{10}	Synthetic alcohols, ≥ C_{11}	Total
USA	190	560	750
Europe/South Africa	395	529	924
Asia	651	142	793
Total	1236	1231	2467

In the United States the synthetic fatty alcohol capacities predominate with ca. 70 % share of the total capacity. The largest producers of synthetic alcohols are Shell (SHOP/oxo process), and Sasol North America (Ziegler/Alfol process). Natural fatty alcohols are produced by Procter & Gamble and Cognis (formerly Henkel).

In Europe/South Africa, ca. 40 % of the production capacity is based on natural raw materials. The largest manufacturer is the Sasol Group, exploiting all three technologies, i.e., Ziegler, oxo and natural production processes. Other European manufacturers, such as BASF, ICI, Exxon, and Shell, predominantly use the oxo process. As the second largest Western European manufacturer for fatty alcohols, Cognis, produces exclusively from natural raw materials using high pressure hydrogenation processes.

In the Far East the share of production capacities from natural raw materials is ca. 80 %. This share will further increase by planned additional capacities of about 600 000 t by 2008. The currently largest manufacturers of natural fatty alcohols are Ecogreen, Sawit Mas, and Kao Soap Corp. Synthetic fatty alcohols are manufactured by the oxo process, e.g., by Mitsubishi Chemical, Mitsubishi Petrochemical and Fushun, or by the Ziegler process, used in Jilin.

In Eastern Europe fatty alcohols are manufactured predominantly on the basis of the paraffin oxidation process.

Considering the existing production capacities and based on the announced capacity increases, assuming an estimated yearly increase of the fatty alcohol demand of 2 % to 3 %, ca. 80 % of the capacities will be utilized through the year 2005 [76], [77], [116], [118]. The planned new capacities will be based on fats and oils and on petrochemical or Fischer – Tropsch raw materials, respectively.

The capacity utilization of each production unit depends on the overall economic development, the product mix and the prices for the used raw materials. Figure 11 shows the differences in price for raw materials based on crude oil, fats, and oils.

Table 9. Fatty alcohol production capacities (C_{10} and higher) by company, country, location, process and raw material 2005/2006

Country	Location	Company	Raw material	Process	Capacity, 10^3 t
USA	Cincinnati	Cognis	oleochemical	methyl ester hydrogenation	45
USA	Sacramento	P & G	oleochemical	methyl ester hydrogenation	145
USA	Lake Charles	SASOL	ethylene	Ziegler	60
USA	Geismar	Shell	ethylene	modified oxo	500
Total America					750
China	Dalian	Dalian Oil+Fats	oleochemical	methyl ester hydrogenation	15
China	Guandong	Jiangman	oleochemical	methyl ester hydrogenation	15
China	Fushun	Fushun	olefins	oxo	50
China	Jilin	Jilin Chem.	ethylene	Ziegler	47
India	Jalgaon	Aegis Chem.	oleochemical	fatty acid hydrogenation (Lurgi)	4
India	Anklesshawar	Godrey	oleochemical	fatty acid hydrogenation (Lurgi)	65
India	Bombay	VVF	oleochemical	wax ester hydrogenation(Lurgi)	60
Japan	Tokushima	NewJapanChem	oleochemical	methyl ester hydrogenation	7
Japan	Mizushima	Mitsubishi	olefins	oxo	10
Japan	Yokkaichi	Mitsubishi	olefins	modified oxo	35
Indonesia	North Sumatra	Ecogreen	oleochemical	fatty acid hydrogenation (Lurgi)	35
Indonesia	Batam	Ecogreen	oleochemical	methyl ester hydrogenation	80
Malaysia	Kuantan	P&G/Felda	oleochemical	methyl ester hydrogenation	75
Malaysia	Selangor	Cognis	oleochemical	methyl ester hydrogenation	30
Malaysia	Penang	Kao/Palmco	oleochemical	methyl ester hydrogenation	165
Philippines	Lucena	Colgate	oleochemical	methyl ester hydrogenation	5
Philippines	Jasaan	Kao	oleochemical	methyl ester hydrogenation	30
Philippines		Primofina / Pencentury	oleochemical	methyl ester hydrogenation (Davey)	30
Philippines	Bauan	United Coconut	oleochemical	fatty acid hydrogenation (Lurgi)	35
Total Asia					793
Romania		Government	paraffins	oxidation (idle capacity)	10
Russia		Government	paraffins	oxidation (idle capacity)	60
Russia	Ufa	Government	ethylene	Ziegler	20
Belgium	Ertevelde	Oleofina	oleochemical	methyl ester	35

				hydrogenation	
France	Boussens	Sidobre/Cognis	oleochemical	methyl ester hydrogenation	40
Germany	Brunsbüttel	SASOL	oleochemical	fatty acid hydrogenation (Lurgi)	35
Germany	Rodleben	Ecogreen	oleochemical	methyl ester hydrogenation	10
Germany	Düsseldorf	Cognisl	oleochemical	methyl ester hydrogenation	180
Germany	Marl	SASOL	oleochemical	methyl ester hydrogenation	95
Germany	Ludwigshafen	BASF	olefins	oxo	50
Germany	Brunsbüttel	SASOl	ethylene	Ziegler	42
UK	Stanlow	Shell	ethylene	modified oxo	107
Italy	Augusta	SASOL	olefins	oxo	120
S. Africa	Secunda	SASOL	FT olefins	oxo	120
Total Europe/Africa					924

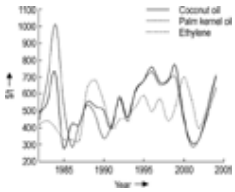

Figure 11. Price development of raw materials used for the production of fatty alcohols

While until 1973 crude oil based, synthetic fatty alcohol production units had considerable advantages, this picture changed significantly due to the price increases for crude oil in 1973 – 1974 and 1979 – 1980 and subsequently for secondary products such as naphtha, paraffins, and ethylene. On the long run the price increase rate for fats and oils was considered to be lower than for crude oil [76], [77], [115], [116], [120-127], leading to an expansion of the natural production capacities. In the beginning of 1999 the price for crude oil was again nearly as low as the price in 1974. Therefore, fatty alcohols on the basis of petrochemical raw materials had a cost advantage over natural-based fatty alcohols. This changed again in the second half of 1999, whereas in 2005 the natural-based products enjoy the advantage of high crude oil prices again. The period 1995 – 2005 was a decade of change [119]. Some units were shut down or reinstalled at other places, and larger plants were build to take advantage of economy of scale. Pants with a size of 100 000 t/a are now regarded as standard.

The long-term estimate was based on an increase in crop yields for fats and oils, new hybrids of oil plants, a wider geographical distribution of their cultivation areas, and the renewable character of these raw materials in contrast to crude oil. All of this guaranteed a reliable supply. Therefore, the ratio of the production capacities of synthetic to natural fatty alcohols changed in favor of natural products (2005). In spite of production increases for coconut oils, increased utilization of palm kernel and babassu oils, and the new hybrids of C_{12}-rich oils like cuphea varieties, natural products are mainly limited to the C_{16} – C_{18} range (90 % of all products) [128]. Synthetic alcohols, however, show their strength especially in the C_{12} – C_{15} range, which plays an important role in surfactant applications. Therefore, the C_{16} – C_{18} C-chain range is predominantly based on natural, the C_{12} – C_{15} range on petrochemical based raw materials. However, due to logistical, economical and political reasons, overlapping and reversals can occur [76], [115], [116].

[Top of Page]

9. Ecotoxicology and Toxicology

The following toxicological characterization covers alcohols with C-chain lengths from C_6 to C_{24}. Besides data regarding the acute toxicity (after single intake of the substance) and regarding local effects (skin- and mucous membrane irritation, data on sensitization effects), especially data regarding the toxicity after repeated intake of small amounts of a substance need to be taken into account for the toxicological characterization. This is especially valid for substances which have virtually no acute toxicity as e.g. fatty alcohols. Questions regarding a possible carcinogenic effect or regarding possible reproduction damage or teratogenic effects are the main focus point today when the toxicological properties of chemical substances are assessed.

For the assessment of the properties of fatty alcohols Condea data and data from scientific publications were used. Here especially the IUCLID-data [129] were consulted, which were compiled by the manufacturers according to the ECAltstoffverordnung (EEC 793/93).

9.1. Ecotoxicology and Environmental Aspects

Linear fatty alcohols are in general easily biodegradable. The biological degradation of secondary fatty alcohols proceeds slightly slower, however, the decomposition rates in standardized lab tests are comparatively high. For the assessment of the

data it has to be considered that the solubility of fatty alcohols in water decreases with increasing C-chain length. Therefore, the alcohols are present only in an extremely low concentration in aquatic environments. Effects, which are detected in various test systems, often have to be regarded as secondary effects caused by undissolved, finely dispersed substance particles, and therefore have to be distinguished from “real” substance effects.

For aquatic organisms fatty alcohols possess only moderate *acute toxicity*. In general, in the concentration range of their water solubility no toxic effects are observed. However, a number of studies was performed with concentrations which are considerably above the water solubility. The observed toxic effects can also be attributed to the aforementioned secondary effects. The available data for the *chronic toxicity* neither indicate a special toxicological potential of fatty alcohols. Dodecanol shows a more distinct toxicity in the studies. However, this toxicity cannot unmistakably be attributed to a direct substance effect, but can possibly be induced by physical effects as well. A complete separation of the aqueous solution from undissolved but finely dispersed dodecanol is virtually impossible. Therefore, a rest of undissolved dodecanol always remains in the water.

9.2. Toxicology

Linear fatty alcohols are predominantly metabolized by oxidation. The resulting fatty acids can be subsequently integrated into the fatty acid metabolism. A further metabolic reaction is conjugation of the fatty alcohols with glucuronic acid and subsequent excretion.

Fatty alcohols possess virtually no *acute toxicity* after oral intake or after exposure to the skin. Especially fatty alcohols with a shorter C-chain length irritate the skin in higher concentrations, in animal tests eye irritations were observed as well. The irritating effect decreases with increasing C-chain length. In studies with volunteers some fatty alcohols show a considerably lower skin irritating effect. This has, at least in part, to be attributed to the different test conditions of studies with volunteers as compared with animal experiments. The available data and the chemical structure of fatty alcohols do not indicate a sensitizing potential of fatty alcohols. From today's point of view even after long-term intake of the typical linear fatty alcohols, substance induced damage is not to be expected. The available data show that fatty alcohols do not possess *genotoxicity*, which is not to be expected anyway on grounds of their chemical structure. Valid data regarding a *carcinogenic effect* are not available. But, based on structure-effect considerations, it can be assumed that fatty alcohols do not have a carcinogenic potential.

For some of the fatty alcohols at least guidance data on *reproduction toxicity* are available. These data show that under the study conditions fatty alcohols neither reduced the fertility nor had an embryo or fetotoxic effect.

In Table 10 data on the toxicity of fatty acids after short-term exposure are listed. Table 11 summarizes data on the acute toxicity to aquatic organisms and degradability data [129-134].

Table 10. Acute toxicity (oral and dermal), skin Irritation, eye Irritation, and sensitization of fatty alcohols (data from animal experiments)

Substance	LD_{50} (oral)	LD_{50} (dermal)	Skin Irritation	Eye Irritation	Sensitization
1-Hexanol	3100 – 5000 mg/kg (rat)	1500 – >5000 mg/kg (rabbit)	moderately irritant	irritant	no sensitization
1-Octanol	> 3200 – >5000 mg/kg (rat)	> 5000 mg/kg (rabbit)	moderately irritant	irritant	no sensitization
Decanol	5000 mg/kg (rat)	> 3500 mg/kg (rabbit)	irritant	irritant	-
Dodecanol	> 5000 mg/kg (rat)	> 8300 mg/kg (guinea pig)	moderately irritant	moderately irritant	-
Tetradecanol	5000 – 20 000 mg/kg (rat)	> 5000 mg/kg (rabbit)	slightly irritant	slightly irritant	-
Hexadecanol	> 5000 – > 7500 mg/kg (rat)	> 5000 mg/kg (rabbit)	-	-	no sensitization
Octadecanol	> 5000 – > 8000 mg/kg (rat)	-	-	-	-
Eicosanol	>10 000 mg/kg (rat)	-	not irritant	slightly irritant	-
Docosanol	> 10 000 mg/kg (rat)	-	not irritant	slightly irritant	-
2-Butyloctanol	27 – 33 g/kg (rat)	-	not irritant	not irritant	no sensitization (human data)
2-Hexyl-1-decanol	> 33 000 mg/kg (rat)	-	slightly irritant	not irritant	no sensitization (human data)
C_{16} – C_{20}-	> 39 000 mg/kg	-	not irritant	not irritant	no sensitization

Branched Alcohols	(rat)		(rabbit)	(guinea pig)	(guinea pig, human data)
2-Octyldodecanol	> 5000 – > 42 000 mg/kg (rat)	-	slightly irritant (guinea pig)	slightly irritant (guinea pig)	no sensitization (guinea pig, human data)
2-Decyltetra-decanol	> 39 000 mg/kg (rat)	-	not irritant	not irritant	no sensitization (human data)

Table 11. Ecotoxicity of fatty alcohols

Substance	Bio-Degradability	Acute Fish Toxicity	Acute Daphnia Toxicity	Algae Toxicity	Bacteria Toxicity
1-Hexanol	biodegradable	> 100 mg/L	> 100 mg/L	10 – 100 mg/L	3000 – 10 000 mg/L
1-Octanol	biodegradable	10 – 100 mg/L	10 – 100 mg/L	10 – 100 mg/L	350 – > 10 000 mg/L
Decanol	easily biodegradable	1 – 10 mg/L	1 – 10 mg/L	-	> 10 000 mg/L
Dodecanol	easily biodegradable	< 1 mg/L	< 1 mg/L	< 1 mg/L	> 10 000 mg/L
Tetradecanol	biodegradable	-	-	-	-
Hexadecanol	biodegradable	-	-	> 100 mg/L	-
Octadecanol	biodegradable	> 100 mg/L	> 100 mg/L	> 100 mg/L	> 10 000 mg/L
Eicosanol	-	-	-	-	-
Docosanol	-	-	-	-	-
2-Butyloctanol	biodegradable	-	-	-	-
2-Hexyl-1-decanol	easily biodegradable	> 100 mg/L	-	-	-
C_{16} – C_{20}-Branched Alcohols	-	-	-	-	-
2-Octyldodecanol	biodegradable	> 100 mg/L	-	100 mg/L	> 10 000 mg/L
2-Decyltetradecanol	biodegradable	-	-	-	-

[Top of Page]

References

General References

1. *Ullmann*, 4th ed., **11**, 427 – 445.
2. *ACS Symp. Ser.* **159** (1981).
3. F. Korte (ed.): *Methodicum Chimicum*, vol. 5, Stuttgart 1975.
4. H. Stache, *Tensid Taschenbuch*, Carl Hanser Verlag, München 1981.
5. *Houben-Weyl*, **VI/1a/b**, 1 ff.
6. *Fettalkohole*, 2nd ed., Henkel, Düsseldorf 1982.

Specific References

7. H. Adkins, K. Folkers, *J. Am. Chem. Soc.* **53** (1931) 1095 – 1097. Links
8. W. Normann, *Angew. Chem.* **44** (1931) 714 – 717. DE 617 542, 1930 (W. Normann, H. Prückner); Böhme Fettchemie GmbH, DE 639 527, 1930 (W. Normann).
9. W. Schrauth, O. Schenk, K. Stickdorn, *Ber. Dtsch. Chem. Ges.* **64** (1931) 1314 – 1318.
10. O. Schmidt, *Ber. Dtsch. Chem. Ges.* **64** (1931) 2051 – 2053.
11. D. Precht, *Fette, Seifen, Anstrichm.* **78** (1976) no. 4, 145.
12. *Alfol-Alkohole, Typische Analysendaten*, Condea Chemie, Brunsbüttel.
13. *VDI-Wärmeatlas*, Düsseldorf 1984.
14. *D'Ans Lax – Taschenbuch für Chemiker und Physiker*, vol. **II**, Springer Verlag, Berlin 1964.
15. *CRC Handbook*, 55th ed., Cleveland 1974 /75.
16. *Houben-Weyl*, **E3**, 265 – 300.
17. *Houben-Weyl*, **E5/1**, 202 – 212.
18. K. Heyns, L. Blasejewicz, *Tetrahedron* **9** (1960) 67. Links

19. E. S. Gore, *Platinum Met. Rev.* **27** (1983) no. 3, 111.
20. Atlantic Richfield Co., US 3 997 578, 1976 (Ming Nan Sheng).
21. *Kirk-Othmer*, **9**, 795 ff.
22. U. R. Kreutzer, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 343. Links
23. H.-D. Komp, M. P. Kubersky in: *Fettalkohole,* 2nd ed., Henkel, Düsseldorf 1982, p. 51 ff.
24. E. F. Hill, G. R. Wilson, E. C. Steinle, Jr., *Ind. Eng. Chem.* **46** (1954) 1917.
25. M. L. Karstens, H. Peddicord, *Ind. Eng. Chem.* **41** (1949) 438.
26. H. Igo, *CEER Chem. Econ. Eng. Rev.* **8** (1976) no. 3, 31.
27. WO 90/08121, 1990.
28. DE 4343320 A1, 1993.
29. Oleofina, DE 3425758 C2, 1984/94.
30. DE 3425758 A1, 1984.
31. H. Buchold, *Chem. Eng. (N.Y.)* **90** (1983) no. 4, 42.
32. Th. Voeste, H. Buchold, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 350. Links
33. *Chem. Ind. (Düsseldorf)* **32** (1980) 739.
34. Metallgesellschaft AG, DE 2 853 990, 1978 (Th. Voeste, H.-J. Schmidt, F. Marschner).
35. Dehydag GmbH, DE 1 005 497, 1954 (W. Rittmeister).
36. Dehydag GmbH, US 3 193 586, 1965 (W. Rittmeister).
37. Henkel KGaA, DE-AS 2 613 226, 1976 (G. Demmering)
38. VEB Deutsches Hydrierwerk Rodleben, DD 213 430, 1983 (H. Aring, K. Busch, P. Franke, G. Konetzke *et al.*)
39. Institut Français du Pétrole, DE-OS 3 217 429, 1982 (R. Snappe, J.-P. Bournonville).
40. B. C. Triverdi, D. Grote, Th. O. Mason, *J. Am. Oil Chem. Soc.* **58** (1981) no. 1, 17.
41. Ashland Oil, Inc. US 4 104 478, 1978 (B. C. Triverdi).
42. *Houben-Weyl*, **13/4**, 1 ff.
43. T. Mole, E. A. Jeffery: *Organoaluminium Compounds*, Elsevier, Amsterdam 1972.
44. A. Lundeen, R. Poe: “Alpha-Alcohols,” in J. J. Mc Ketta, W. A. Cunningham (eds.): *Encyclopedia of Chemical Processing and Design*, **vol. 2**, Marcel Dekker, New York 1977, p. 465.
45. K. Ziegler, H.-G. Gellert, K. Zosel, E. Holzkamp *et al.*, *Justus Liebigs Ann. Chem.* **629** (1960) 121.
46. H. Wesslau, *Justus Liebigs Ann. Chem.* **629** (1960) 198.
47. P. H. Washecheck, *ACS Symp. Ser.* **159** (1981) 87.
48. K. L. Lindsay: “Alpha-Olefins,” in J. J. Mc Ketta, W. A. Cunningham (eds.): *Encyclopedia of Chemical Processing and Design*, vol. 2, Marcel Dekker, New York 1977, p. 482.
49. *Produkte aus der Oxosynthese*, Ruhrchemie AG, Frankfurt 1969.
50. J. Falbe: *Synthesen mit Kohlenmonoxid*, Berlin 1967.
51. R. E. Vincent, *ACS Symp. Ser.* **159** (1981) 159.
52. B. Cornils, *Compend. Dtsch. Ges. Mineralölwiss. Kohlechem.* **78/79** (1978) no. 1, 463.
53. B. Cornils, “New Syntheses with Carbon Monoxide,” *React. Struct. Concepts Org. Chem.* **11** (1980) .
54. H. Lemke, *Hydrocarbon Process.* **45** (1966) no. 2, 148.
55. C. E. O'Rourke, P. R. Kavasmaneck, R. E. Uhl, *ACS Symp. Ser.* **159** (1981) 71.
56. B. A. Murrer, J. H. Russel, *Catalysis* **6** (1983) 169.
57. A. H. Turner, *J. Am. Oil Chem. Soc.* **60** (1983) no. 3, 623. Links
58. R. L. Banks, *Appl. Ind. Catal.* 1984, no. 3, 215.
59. E. R. Freitas, C. R. Gum, *Chem. Eng. Prog.* **75** (1979) 73.
60. E. L. T. M. Spitzer, *Seifen, Oele, Fette, Wachse* **107** (1981) no. 6, 141.
61. H. Stage, *Seifen, Oele, Fette, Wachse* **99** (1973) no. 6/7, 143; no. 8, 185; no. 9, 217; no. 11, 299.
62. N. Kurata, K. Koshida, *Hydrocarbon Process.* **57** (1978) 145.
63. N. J. Stevens, J. R. Livingstone, Jr., *Chem. Eng. Prog.* **64** (1968) no. 7, 61. Links
64. N. Kurata, K. Koshida, H. Yokoyama, T. Goto, *ACS Symp. Ser.* **159** (1981) 113.
65. L. Marko, *J. Organomet. Chem.* **283** (1985) 221. Links
66. B. N. Bobylev, M. J. Farberov, S. A. Kesarev, *Khim. Promst. (Moscow)* 1979 no. 3, 138.
67. *Unilin Alcohols*, Petrolite Corp., St. Louis 1985.
68. Petrolite Corp., US 4 426 329, 1984 (J. H. Woods, C. E. Laughlin, T. R. Graves).
69. S. K. Ries, V. Wert, C. S. Sweely, R. A. Leavitt, *Science* **195** (1977) 1339. Links
70. J. Penninger, M. Biermann, H.-J. Krause, *Fette, Seifen, Anstrichm.* **85** (1983) no. 6, 239.
71. T. Gibson, L. Tulich, *J. Org. Chem.* **46** (1981) no. 9, 1821. Links
72. A. V. Rama Rao, J. S. Yadav, G. S. Annapurna, *Synth. Commun.* **13** (1983) no. 4, 331.
73. S. M. Kulkarni, V. R. Mamdapur, M. S. Chadha, *Indian J. Chem., Sect. B* **22** (1983) no. 7, 683.
74. U. T. Bhalerao, S. J. Rao, B. D. Tilak, *Tetrahedron Lett.* **25** (1984) no. 47, 5439. Links
75. G. Kirchner, H. Weidmann, *Liebigs Ann. Chem.* 1985, no. 1, 214.
76. H. J. Richtler, J. Knaut, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 160 – 175.

77. J. Knaut, H. J. Richtler, *J. Am. Oil Chem. Soc.* **62** (1985) no. 2, 317. Links
78. H. Bertsch, H. Reinheckel, K. Haage, *Fette, Seifen, Anstrichm.* **66** (1964) 763.
79. H. Bertsch, H. Reinheckel, K. Haage, *Fette, Seifen, Anstrichm.* **71** (1969) 357, 785.
80. H. Bertsch, K. Haage, H. Reinheckel, *Fette, Seifen, Anstrichm.* **71** (1969) 851.
81. K. Lindner: *Tenside, Textilhilfsmittel, Waschrohstoffe*, vol. **I, 2nd ed.**, Wissenschaftl. Verlags GmbH, Stuttgart 1964, p. 144.
82. J. Richter: Dissertation, TH Delft, 1968.
83. J. W. E. Coenen, *Fette, Seifen, Anstrichm.* **77** (1975) 431.
84. Dehydag, US 3 193 586, 1965 (W. Rittmeister).
85. Henkel, US 3 729 520, 1973 (H. Rutzen, W. Rittmeister).
86. Henkel, DE-AS 2 513 377, 1975 (G. Demmering, H. Schütt, H. Rutzen).
87. J. Cieslar, W. Bulanda, *Przem. Chem.* **63** (1984) 375; Links *Chem. Abstr.* **101** (1984) 153 810 t.
88. Institute Blachownia and Nitrogen Works Kedzierzyn, PL 118 880, 1978 (E. Fabisz, K. Chmielewski, A. Jakubowicz, A. Mankowski *et al.*).
89. Institute Blachownia and Nitrogen Works Kedzierzyn, PL 118 979, 1978 (E. Fabisz, K. Chmielewski, A. Jakubowicz, A. Mankowski *et al.*).
90. New Japan Chemical Ltd., GB 1 335 173, 1972.
91. VEB Deutsches Hydrierwerk Rodleben, DD 213 429, 1983 (H. Aring, K. Busch, P. Franke, G. Konetzke *et al.*).
92. UOP Inc., US 4 340 546, 1982 (G. Qualeatti, D. Germanas).
93. UOP Inc., US 4 446 073, 1984 (G. Qualeatti, D. Germanas).
94. Kao Soap, JP 58 210 035, 1983 (K. Kokubo, K. Tsukada, Y. Miyabata, Y. Kazama).
95. Kao Soap, JP 5 995 227, 1984 (K. Tsukada, Y. Miyabata).
96. Kao Soap, JP 59 106 431, 1984 (K. Tsukada, Y. Miyabata, K. Fukuoka).
97. H. Schütt in W. Foerst, H. Buchholz-Meisenheimer (eds.): *Neue Verfahren, Neue Produkte, Wirtschaftliche Entwicklung*, Urban und Schwarzenberg, München 1970.
98. U. Ploog, *Seifen, Oele, Fette, Wachse* **109** (1983) no. 8, 225.
99. R. R. Egan, G. W. Earl, J. Ackermann, *J. Am. Oil Chem. Soc* **61** (1984) no. 2, 324. Links
100. M. Guerbet, *C. R. Hebd. Séances Acad. Sci.* **128** (1898) 511; M. Guerbet, *Bull. Soc. Chim. Fr.* **21** (1899) no. 3, 487.
101. J. Glasl in: *Fettalkohole*, 2nd ed., Henkel, Düsseldorf 1982, p. 169.
102. Henkel & Cie GmbH, DE-OS 1 768 313, 1971 (H. Rutzen).
103. PCT/WO96/19287; PCT/DE/95/01846,
104. Henkel & Cie GmbH, DE 1 198 348, 1961 (W. Stein, M. Walther).
105. Henkel & Cie GmbH, DE 1 207 371, 1963 (H. Hennig, W. Stein, M. Walther).
106. Hüls AG, DE 38 43956,
107. “Mineral- und Brennstoffnormen,” *DIN-Taschenbuch*, Beuth-Vertrieb, Berlin-Köln-Frankfurt 1984.
108. *Annual Book of ASTM Standards*, Sections 5, 6, 15, 1985.
109. *Deutsche Einheitsmethoden zur Untersuchung von Fetten, Fettprodukten und verwandten Stoffen*, Wissenschaftl. Verlags GmbH, Stuttgart 1950 – 1984.
110. H. J. Richtler, J. Knaut, *Chem. Ind. (Düsseldorf)* **36** (1984) 131 – 134.
111. H. J. Richtler, J. Knaut, *Chem. Ind. (Düsseldorf)* **36** (1984) 199 – 201.
112. L. Marcou, *Rev. Fr. Corps. Gras.* **30** (1983) no. 1, 3 – 6.
113. *J. Am. Oil Chem. Soc.* **58** (1981) no. 11, 873 – 874A.
114. A. L. de Jong, *J. Am. Oil Chem. Soc.* **60** (1983) no. 3, 634 – 639.
115. E. C. Leonard, *J. Am. Oil Chem. Soc.* **60** (1983) no. 6, 1160 – 1161. Links
116. C. A. Houston, *J. Am. Oil Chem. Soc.* **61** (1984) no. 2, 179 – 184.
117. *J. Am. Oil Chem. Soc.* **62** (1985) no. 4, 668.
118. J. C. Dean, *Chem. Week* 1985, May 15, 3 – 34.
119. P. Healey: “The Alcohol Market, A Decade of Change”, ICIS Lor Conference, Amsterdam, Oct. 2004.
120. J. W. E. Coenen, *J. Am. Oil Chem. Soc.* **53** (1976) no. 6, 382 – 389.
121. H. Klimmek, *Chem. Ind. (Düsseldorf)* **33** (1981) 136 – 141.
122. H. Fochem, *Fette, Seifen, Anstrichm.* **87** (1985) no. 2, 47 – 52.
123. *Chem. Week* 1982, April 28, 40 – 44.
124. *Chem. Ind. (Düsseldorf)* **35** (1983) 251 – 255.
125. *J. Am. Oil Chem. Soc.* **61** (1984) no. 8, 1298 – 1314.
126. H. Lindemann, *Chem. Ind. (Düsseldorf)* **34** (1982) 713 – 716.
127. S. Field, *Hydrocarbon Process.* 1984, Oct., 34 G – N.
128. W. Stein, *Fette, Seifen, Anstrichm.* **84** (1982) no. 2, 45 – 54.
129. Public data on high volume chemicals: IUCLID CD-ROM, Year 2000 edition: European Commission, Joint Research Centre, Institute for Health and Consumer Protection, European Chemicals Bureau.
130. D. L. J. Opdyke, *Fd. Cosmet. Toxicol.* **11** (1973) 95 – 115.

131. *Food Cosmet. Toxicol.* **13** (1975) 699 – 700.
132. “Final Report on the Safety Assessment of Cetearyl Alcohol, Cetyl Alcohol, Isostearyl Alcohol, Myrsityl Alcohol, and Behenyl Alcohol”; *J. Am. Coll. Toxicol.* **7** (1988) no. 3, 359 – 411.
133. D. L. J. Opdyke, Fragrance Raw Materials Monographs “Cetyl Alcohol”; *Food Cosmet. Toxicol.* **16** (1978) 683 – 686.
134. “Final Report on the Safety Assessment of Stearyl Alcohol, Oleyl Alcohol, and Octyl Dodecanol”, *J. Am. Coll. Toxicol.* **4** (1985) no. 5, 1 – 29.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Felts

Nonwoven Fabrics and Wool are separate keywords.

Friedrich Dornedden, Filzfabrik Fulda GmbH & Co., Fulda, Federal Republic of Germany

1.	**Introduction**	1
2.	**Felting Process**	2
3.	**Pressed Felts**	2
3.1.	**Raw Materials**	2
3.2.	**Production**	2
3.2.1.	Batt Production	2
3.2.2.	Piece Felts	3
3.2.3.	Formed Felts	6
3.3.	**Properties**	6
3.4.	**Uses**	6
3.5.	**Testing**	7
4.	**Needled Felts**	7
4.1.	**Production**	7
4.2.	**Properties and Uses**	8
5.	**References**	9

1. Introduction

Felt (nonwoven) is defined as a material made entirely from mechanically consolidated fibers (DIN 61 205).

Pressed felts (wool and hair felts) are materials, composed either of fibers that can be felted or of a mixture of these and other fibers. These nonwoven fabrics are consolidated by the utilization of moisture, heat, pressure, and mechanical work (felting and, if necessary, pressing).

Needled felts are materials made from fibers that are consolidated by using a large number of needles. The needles alternately penetrate and are withdrawn from the fibers in a process called needling.

Nonwovens (→ Nonwoven Fabrics) are distinguished in DIN 61 210 from needled felts because they are consolidated by the following different methods:

1) by use of a binding agent
2) by partially dissolving the fibers
3) by melting the fibers
4) by shrinking the fibers under the influence of heat and/or a swelling agent
5) by the action of liquid jets
6) by needling and subsequent consolidation, using methods 1 to 4 [1]

The felting of wool and animal hair was discovered in ancient times. Felt caps, 3500 years old, were found in the megalithic graves of Jutland. The Chinese referred to felt as far back as 2300 B.C. [2].

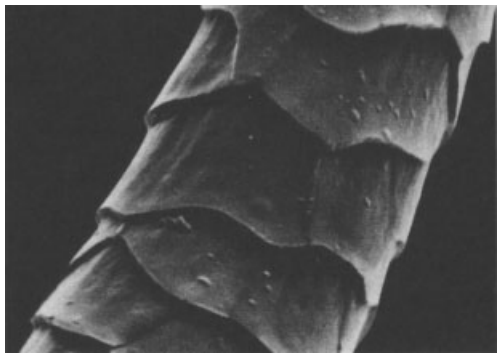

Figure 1. Scanning electron micrograph (× 3500) of a Merino wool fiber

2. Felting Process

The felting process ist still not completely understood [3], [4]. The felting ability of wool fibers depends primarily on the scales on their surface (Fig. 1) and on their elastic properties (see also → Wool). The scales of any one fiber all point in the same direction, thereby creating a directional frictional effect that allows the fibers to move more easily in the direction of their roots when force is applied. Only fibers with scales can be felted. During felting and fulling of fibrous webs (batt), the individual fibers are pressed together, forced to move against one another, and become mutually intertwined. Ends or parts of fibers that are not interlocked escape into the interspaces in the felt. On removing the applied compressing force (stress relief), the fibers try to spring back

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_297

into their original stress-free positions. However, friction prevents complete stress relief. The original position and form of each fiber is permanently changed at the end of felting and fulling, and the fibers are firmly interlocked.

3. Pressed Felts

3.1. Raw Materials

Wool and/or other animal hair are the basic raw materials for the production of pressed felts (wool felts, hair felts, and hair – wool felts). Only these fibers possess the natural felting power that enables the fibrous web made from the carded fibers to be consolidated by using moist heat pressure, and mechanical work. Other fibers that cannot be felted (cellulose fibers, synthetic fibers, cotton, etc.) can be added to obtain special effects, to produce felts with specific properties, or to reduce costs.

Hair – wool felts are produced from a mixture of wool and animal hair (preferentially hair from cattle, calves, or goats). Since the fulling power of hair is lower than that of wool, the felted, semifinished product is soaked in hot, dilute sulfuric acid before fulling.

Bating. Hat felts contain wool, rabbit fur, or hare fur (velour hats). The animal fur first has to be softened to improve its fulling power by a process known as bating. The fur is usually exposed to a mineral acid – hydrogen peroxide solution (bate). Poisonous mercury bates are employed only in the production of velour from hare fur.

In bating machines, only the ends of the hairs are exposed to the bate making them softer, more pliable, and more resilient. As a result, they can intertwine more easily when the untreated hair shafts move in the direction of their roots during felting and fulling. Bating also increases the frictional coefficient of the animal hairs, and thus improves their mutual adhesion [5].

3.2. Production

The production of pressed felts is summarized schematically in Figure 2 (cf. [6]).

3.2.1. Batt Production

The raw fiber material is fed into one or more carding willows either in pneumatic mixing plants or multiple mixers. The fiber flocks are loosened (opened), and uniformly blended during repeated passage through the machine. Addition of carding oil makes the fibers pliable for further processing. The loosened fibers are then fed in batches into a carding machine via a weigh feeder which may be electronically controlled or via a volumetrically controlled hopper. The carding machine separates the fiber flocks into their individual fibers. Different card sets are available, depending on the raw material, and the required quality and properties of the resulting batt. The fibers leaving the carding machine are longitudinally oriented in the form of a web which has a mass of $10-20\,g/m^2$. The carded web is laid in several layers on the 60 – 70-m long, continuous cloth belt of a batt frame (Fig. 3) to produce a predetermined thickness or mass of batt. The resulting batt is subsequently rolled up.

In the continuous production of "cross-laid" or "cross-lapped" batt, carding is followed by the use of a cross lapper instead of a batt frame. The cross lapper picks up the carded web and deposits it at alternating angles of about 45° (i.e., in a zig-zag pattern) across a conveyor belt (→ Nonwoven Fabrics) so that the fibers are aligned crossways. The cross-lapped batt can be laid onto another oriented web and its mass can be changed by stretching. Ideally, a multidirectional web is obtained whose tensile strengths in the lateral and longitudinal directions are the same [7].

Both methods yield batt in roll form; this is the starting material for the production of piece felts and formed felts.

3.2.2. Piece Felts

Felting. The batts used to produce piece felts are laid singly or on top of one another, depending on the thickness or mass desired in the finished product. The layers of batt are then conveyed between damp cotton cloths and fed into a plate hardener or roller hardening machine. Before entering the felting aggregate (vibrating plates or felting rollers), the batts pass over a

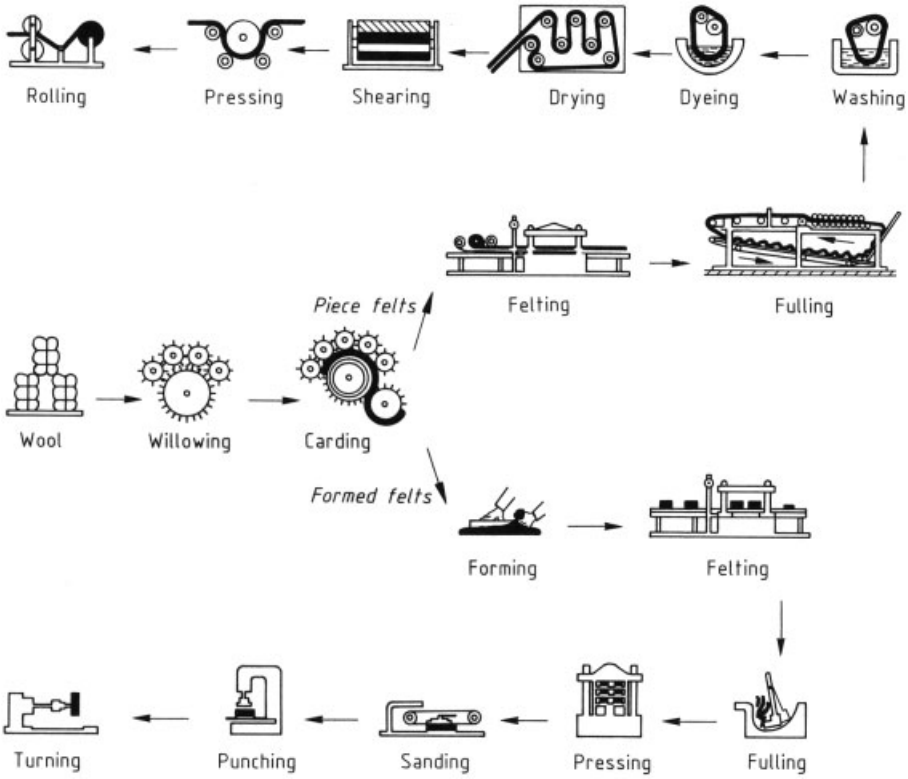

Figure 2. Production of piece and formed felts

steaming plate, where they are moistened and heated by the steam that passes through them.

The circular vibrator motion of the felting plates in the discontinuously operating *plate hardener* (Fig. 4) coupled with the action of moisture, heat, and pressure consolidate and felt the fibers.

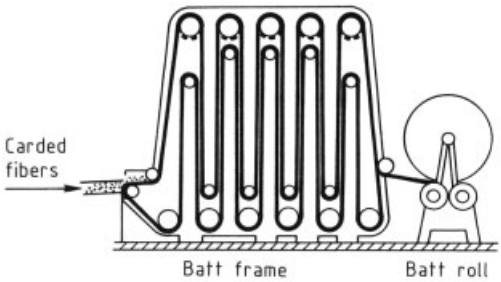

Figure 3. Schematic of a batt frame

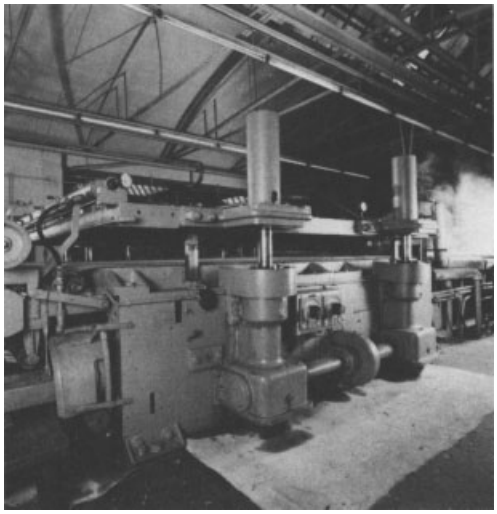

Figure 4. Plate hardener
The hardening plates are shown lifted for transporting the felt, they are lowered during felting.

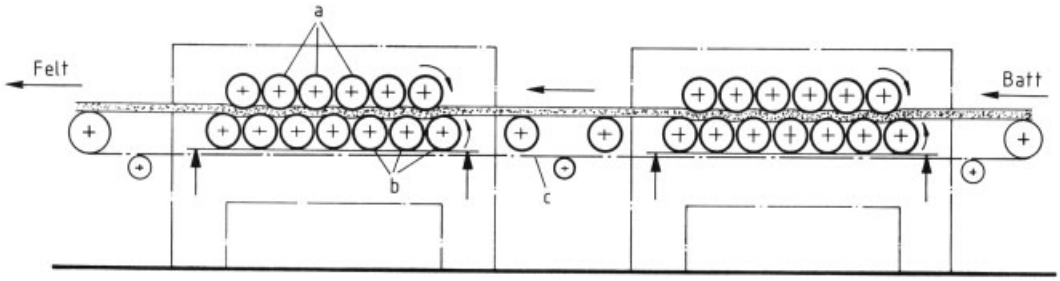

Figure 5. Roller hardening machine
a) Axially vibrating rollers: b) Heated rollers; c) Conveyor belt; d) Cotton cloth

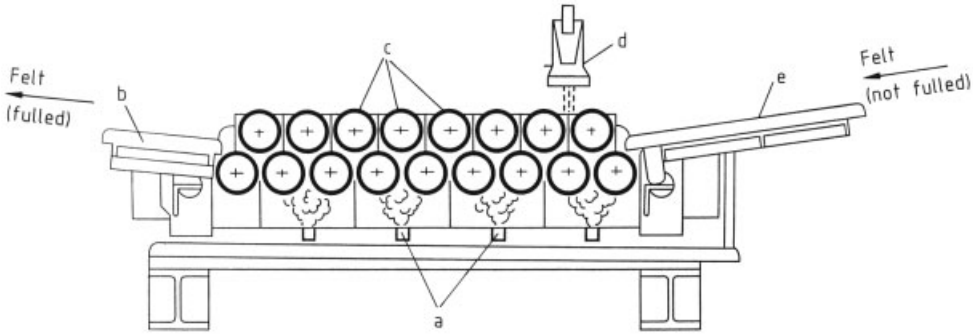

Figure 6. Pilgrim Step roller fulling machine
a) Steam heating; b) Outlet table; c) Fulling rollers; d) Sprinkling device; e) Inlet table

In the continuous *roller hardening machine*, the cotton cloths and the batt pass through two horizontal rows of rollers, one row being above the other (Fig. 5). The lower rollers are made of stainless steel and are heated, while the upper rollers are coated with plastic and vibrate axially. The batt is pressed between each pair of rollers and subjected to their vibrating action. The pressure is released between two consecutive pairs of rollers. The batt is thus consolidated and felted [8].

Fulling. The production of felts with higher densities involves the use of a procedure called fulling. In this process, the material is further felted by kneading, hammering, and pounding, often with the addition of fulling media. Some felts need no further consolidation by fulling, however. Felts used for upholstery and quilts, for example, are merely dried and, if necessary, finished after felting. Lighter piece felts and industrial felts, which may not need to be processed in a kicker mill, are fulled on a roller fulling machine or a Pilgrim Step roller fulling machine.

In a roller fulling machine (shown schematically in Fig. 2, fulling of piece felts), the felt passes between two horizontal rows of 15, 21, or 24 rollers, one row being above the other. The rollers are made of steel and are covered with rubber. For light felts, lower rollers with a diameter of 90 mm are used, but for processing heavier felts, the diameter of the lower rollers must be increased to 120 mm. Only the lower rollers are powered and their speed decreases from roller to roller, which causes the material to "pile up" into folds in the direction of transport. The diameter of the upper rollers is ca. 10 mm less than that of the lower rollers mounted exactly below them. The upper rollers have spring bearings and turn on contact with the folds of material produced by the lower rollers. At high running speed, the upper rollers hammer and beat the material as it passes through the machine [9].

In the *Pilgrim Step roller fulling machine* (Fig. 6), the felt is again transported on a conveyor belt between two horizontal rows of rollers. The rollers work in Pilgrim Step, i.e., they swing forward and backward, the backward

movement being shorter than the forward movement. Thick folded layers of felt are thus formed, which are intensively kneaded and processed as they pass through the machine.

Kicker mills (also known as hammer fulling machines) are used for fulling heavier piece felts with a thickness of several millimeters; they are indispensable for the processing of formed felts, such as disks, sheets, or blocks. They are equipped with one to four hammers (Fig. 7).

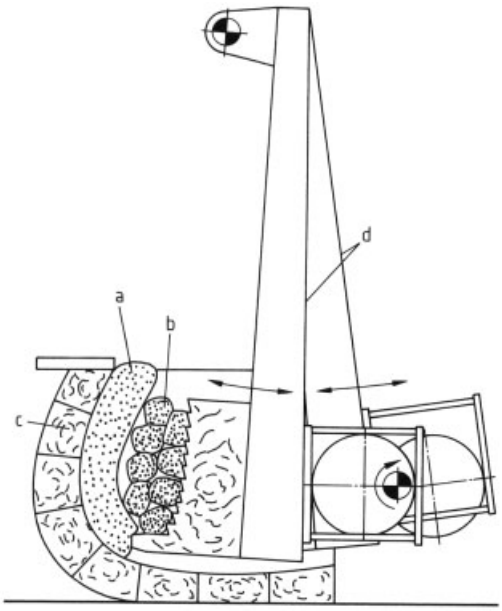

Figure 7. Kicker mill
a) Felt sheet; b) Filling material; c) Trough; d) Hammers (only two of the three hammers are shown)

The type of fulling aid employed depends on the felt quality, mass, bulk density, and the fulling method (machine). Steam or warm water is often sufficient for fulling light felts with a roller fulling machine or a Pilgrim Step roller fulling machine. Fulling with a kicker mill requires the addition of fulling aids such as sulfuric acid, soaps, anionic fatty acid condensation products, or nonionic compounds based on high molecular mass ethoxylated alcohols.

Final Treatment. The wool used for felt production often contains marking ink (pitch marks), which is manually treated with pitch solvents before washing. These solvents should then be washed off along with the lubricants, fulling media, wool grease, and dirt. Continuous cycle washing machines are used, primarily because they are less expensive and save water.

The washed, natural-colored felt can be dried and finished. However, lighter felts are generally dyed by piece dyeing in high-temperature beam dyeing machines, where the felt is wound on perforated drums, or on winch becks in the form of ropes.

Squeeze foulards or centrifuges are employed to remove water. The felt is then stretched and dried with warm air on drying frames (single layer or multilayer tenters). Finishing foulards can also be used prior to the drying tenters. The felt is held in place by means of chains studded with needles or tenter clips as it passes through the frame. Advancing or lagging results in shrinkage or stretching and, thus, determines the dimensions of the felt. Felts used for clothing, decorative purposes, and toys are subjected to finishing, which involves shearing, steaming, decatizing, and/or pressing. Felts may be given other special finishing treatments, depending on the purpose for which they are to be used; they may, for example, be stiffened, softened, flameproofed, or mothproofed.

3.2.3. Formed Felts

In the production of formed felts (summarized in Fig. 2), the carded fibrous web is roughly formed into a defined shape (round, square, or rectangular) and is then built up by crosslaying successive layers. The preformed web is conveyed between cloths over a steaming plate and steamed before being felted in a plate hardener. Felting in the hardener may take several minutes, depending on the thickness of the felt. The formed felts are then treated with dilute sulfuric acid and are either partially fulled on a Pilgrim Step roller fulling machine or completely fulled to yield a specific density of 0.40 – 0.44 g/cm^3. A kicker mill must be used if a high degree of hardness is required (greater than F 4 as specified in DIN 61 200). During fulling the felt must be constantly kept moist and warm with hot water or with fulling liquor, i.e., hot dilute sulfuric acid. The felt is subsequently deacidified in neutralization tubs and, if necessary, leveled in a Pilgrim Step process. It is then centrifuged to reduce its moisture content to ca. 60 % and dried in a drying room until the residual moisture content is

5 – 10 %. The felt disks, sheets, and blocks are then pressed and their surfaces are sanded. This material may be punched, cut, drilled, molded, or machined to provide a large variety of items.

Hat felt must be very carefully fulled. Only hard, compact felts should be used for making hats; these felts are produced as described in Section 3.2.1 by blending, opening, and cleaning the wool or bated hair. The resulting material is subsequently formed on a cone, and is then felted and fulled to a specific size on machines specially constructed for hat production. The form thus obtained is finally dyed, stiffened, formed, prepared, and decorated.

3.3. Properties

The properties of felt are determined primarily by the choice of raw materials and by the manufacturing method used. Felts can be as soft and pliable as fabrics, flexible/resilient, or even hard. Bulk densities of up to 0.80 g/cm^3 have been attained under optimal conditions. Top-quality products with a density of 1.0 g/cm^3 can be made in special cases by adding thermoplastic fibers and further pressing.

Felt displays the following special properties: elasticity, porosity (high pore volume), tensile strength combined with extensibility; high, density-dependent compressibility and recovery; thermal insulation (heat retention) and shock (vibration) and sound absorption; formability; permeability to air and moisture ("breathing," moisture exchange); absorptive capacity; and uniform dyeing. The properties of wool felt are not significantly altered by long-term exposure to lubricants and organic solvents. A low concentration of an inorganic acid, such as sulfuric or hydrochloric acid, does not have an adverse effect on wool felts, but a strong alkaline solution does. Unlike some other industrial materials, wool felts are affected very little by atmospheric factors, such as moisture, heat, cold, sun, and ozone. Felt is unchanged by temperatures from -50 to 100 °C. Two of the most important features of wool felt as a mechanical and industrial material are that it is durable and relatively maintenance-free.

3.4. Uses

Felts are sold as piece felts in lengths (rolled or doubled) or as formed felts (blocks, sheets, disks) and processed as such into felt goods. They are used industrially in the manufacture of machines, equipment, and automotive vehicles. The natural properties of felt are exploited for a wide range of applications, e.g., for dust and oil seals, oil wicks, roller coverings, air and liquid filters, polishing and buffing wheels, self-adhesive insulating tapes, prepunched insulating components, sound proofing, shock absorption linings, pads, piano felts, and orthopedic and surgical felts [10].

3.5. Testing

The most imortant tests performed on industrial felts are the determination of their *hardness* characterized by the *bulk density* ϱ_{RX} in g/cm^3 and their *thickness*. Felt hardness classes, as defined in DIN 61 200, are given in Table 1. Felt hardness is directly related to the pore volume, compressibility, and elasticity. The bulk density, according to DIN 53 855 and ASTM D461-59, is calculated from the mass and the volume of the felt [11].

Table 1. Hardness classification of felts (DIN 61200)

Bulk density ϱ_{Rx}, g/cm^3	Felt hardness
0.10 – 0.16	soft
0.18 – 0.30	intermediate
0.32 – 0.48	compact
0.52 – 0.68	hard

Standard procedures for testing felt thickness are described in DIN 53 855 and largely correspond to those stipulated in ISO 4603 and ISO 5084. The *standard thickness* of wool felt is measured on a circular testing area of 10 cm^2. A testing pressure of 5 cN/cm^2 must be applied to felt with a bulk density < 0.14 g/cm^3, and a pressure of 25 cN/cm^2 to felts with a bulk density > 0.14 g/cm^3. For smaller testing areas ($<$10 cm^2), the testing pressure must be reduced proportionately.

The *relative compressibility* ZR_{x} is defined in DIN 53 855 as:

$$ZR_{\mathrm{x}} = \frac{a_{\mathrm{x}} - a_{10\mathrm{x}}}{a_{\mathrm{x}}} 100\,\%$$

where a_x is the standard thickness of the felt determined by using the normal testing pressure (average of more than three determinations) and a_{10x} is the thickness when a 10-fold higher pressure is applied.

The *pore content* V_{px} is calculated from the density of the fibers, ϱ_N, and the bulk density, ϱ_{Rx}:

$$V_{px} = \frac{\varrho_N - \varrho_{Rx}}{\varrho_N} 100\,\%$$

In the case of mixed felts, the density of the fibers, ϱ_N, is obtained from the densities of the components of the mixture.

Further tests can be conducted for special purposes, some being performed in both longitudinal and lateral directions. These tests are described in DIN 61 206 and ASTM D461-77 and include the determination of the following properties: tensile strength and elongation at break, rupture strength, washing or ironing shrinkage, wool content, length and width, basic mass (mass related to area), air permeability, abrasion loss, pH value, residual fat content, longitudinal resistance to fluid flow, degree of sound absorption, thermal conductivity, and fiber damage based on the measurement of the alkali solubility of wool. Other properties are oil absorption, intrinsic vibration frequency, vibration damping, and heat resistance. The vibration test can be used to determine the felting capacity of wool [12]. Alkali and acid solubility tests can also be employed in the case of bated rabbit and hare fur [13].

4. Needled Felts

The use of needled felts for industrial purposes dates from the 1870s. The technique of needling creates feltlike structures from fibers that are otherwise incapable of being felted by means of numerous barbed needles that vertically penetrate and compress the fibers.

The appearance in the 1950s of new synthetic fibers with novel properties stimulated further development of the needling technique. New, efficient needling machines, in combination with improved felting needles, led to the development of products with new properties that could not be obtained by other methods [14].

4.1. Production

The choice of raw materials depends on the demands made on the needled felts and, thus, on the manifold uses of the finished product. Synthetic fibers consisting of natural and synthetic polymers and inorganic fibers are preferred, but carbon fibers can also be employed. Fibers based on aromatic polyamides (e.g., Nomex, Kevlar) and on polytetrafluoroethylene (e.g., Teflon) are of considerable industrial importance. Needled floor coverings are made chiefly from polyamide and polypropylene fibers. Products used for insulation and upholstery are made from low-grade fibrous material or from recycled natural fibers, cellulose, or synthetic fibers.

The raw material is first loosened (opened) and blended in a willower. A lubricant or antistatic agent can be added at this stage. The fibers are then aerodynamically built up as a multidirectional web. Alternatively, a crosslapped web can be produced on a carding machine and may subsequently be stretched; a multidirectional web can then be obtained with additional equipment.

In the needling machine (Fig. 8), the web fibers pass between two parallel perforated plates and are penetrated by numerous barbed needles. The barbs force the fibers down and, as needling progresses, they become increasingly intertwined thus consolidating the web. The fibers remain locked in their new positions when the needles are withdrawn because the barbs all face in one direction, i.e., toward the needle point. To attain greater compaction and consolidation, several needling machines are usually employed in series. These machines often differ from each other in their needle fittings or in the direction of penetration. The web can be subjected to needling from one side either above or below, or from both sides, either alternately or simultaneously.

In felts required for some industrial uses, the web is needled into a base material made of woven or spunbonded threads, or foil to produce high tensile strength, low stretching, and high resistance to deformation [15].

Mechanical and thermal treatments as well as chemical finishing (singeing, shrinkage, thermofixation, spraying, coating, etc.) produce special effects and influence the dimensions and

physical properties of the felt, such as thickness, porosity, elasticity, and volume. The surface of a needled felt is generally smooth, but the texture can be manipulated to give a velour, looped, or patterned surface [16].

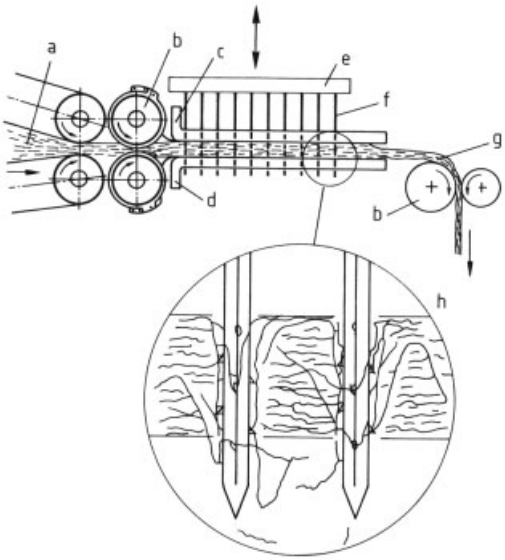

Figure 8. Needling machine
a) Incoming fibrous web; b) Rollers; c) Holddown plate; d) Perforated plate; e) Needle board; f) Felting needle; g) Needled felt; h) Magnified view of needles showing barbs and attached fibers

4.2. Properties and Uses

The properties and uses of needled felt products are directly related to the fibrous raw material, to the method of batt formation, and to the needling technique employed. The uniform distribution of fibers combined with an almost uniform density throughout the felt makes needled felts more suitable for filtration than woven or roughened sheet structures. Therefore, needled felt filters have proved to be an efficient medium for commercial air filtration (dust separation, product recovery, and air purification) and for the separation of solids from liquids.

In the paper industry, needled felts that are more than 15 m wide and are made from synthetic fibers have almost completely replaced the conventional woven felts made from wool or asbestos (woven in tubular form as round felts) [17]. Durable floor coverings for living areas are available in a wide variety of colors, surface structures, and patterns. Needled felts are also employed as floor coverings for indoor tennis courts. Other uses include interior decorations, soundproofing in cars, wall coverings, mattress covers, thermal insulation, and shock and sound absorption. A more recent application of needled felts is their use a geotextiles which are modern construction materials used in water and dam construction, land recovery, and other sectors of civil and hydraulic engineering [17].

5. References

1. H. Dietrich, *Verfahrenstechnik (Mainz)* **2** (1971) 68; *Staub Reinhalt. Luft* **5** (1974) 176; *Wasser Luft Betr.* **17** (1973) 248.
2. W. Voegelin, W. A. Vetterli, *Ciba-Rundsch.* **12** (1958) no. 139, 22 – 23.
3. G. Blankenburg: *Chemische und physikalische Eigenschaften von unveränderter und veränderter Wolle in Beziehung zum Filzvermögen,* **vol. 36,** Deutsches Wollforschungsinstitut an der T. H. Aachen, 1963, pp. 9 – 23. P. A. Koch, G. Satlow, in *Großes Textil-Lexikon,* "Filze," Deutsche Verlagsanstalt, Stuttgart 1965, pp. 427 – 440.
4. A. J. P. Martin, *J. Soc. Dyers Colour* **60** (1944) 325. G. Blankenburg, Th. Laakmann, *Melliand Textilber.* **4** (1971) 467 – 470. *Wool Sci. Rev.* **24** (1964) 34 – 48; *Text. Prax.* **2** (1965) 166. R. Walther, *Melliand Textilber.* **39** (1958) 1240 – 1242
5. R. D. B. Fraser, T. A. Pressly, *Text. Res. J.* **6** 478 – 485, and *Text. Prax.* **12** (1958) 1291.
6. I. Holme: "Wool pressed felt manufacture," *Wool Sci. Rev.* **61** (1985) 36 – 44. H. Dietrich, *Melliand Textilber.* **39** (1958) 977 – 982.
7. H. Dietrich, *Melliand Textilber.* **40** (1959) 591 – 596.
8. H. Dietrich, *Text. Prax.* **12** (1960) 1238 – 1242.
9. H. Dietrich, *Melliand Textilber.* **55** (1974) 714 – 717.
10. L. R. Mizell: "The properties, functions and uses of pressed wool felts," *Wool Sci. Rev.* **61** (1985) 3 – 35.
11. Cl. Sustmann, M. Greul, *Textilingenieurschule Aachen, Erfahrung u. Forschung* **9** (1957) 71, and *Melliand Textilber.* **44** (1963) 443, 560, 674. Broschüre des Verbandes d. Wollfilzindustrie: Filze 1964.

12. G. Blankenburg, H. Zahn, *Text. Praxis* **16** (1961) 228. G. Blankenburg, H. Philippen, *Melliand Textilber.* **51** (1970) 1345.
13. H. G. Fröhlich, *Z. Gesamte Textilind.* **58** (1956) 947.
14. J. Lünenschloß, H. Külter, *TPJ Text. Prax. Int.* **10** (1977) 1227 – 1229.
15. J. Lünenschloß, V. P. Gupta, K. Berns, *Textilbetrieb (Würzburg)* **95** (1977) Okt., 32 – 34; Nov., 23 – 28; Dez. , 47 – 54.
16. J. Ph. Dilo: "Herstellung strukturierter Nadelvliesstoffe," *Chemiefaser Text. Ind.* **31/83** (1981) 674 – 679, E 73 – 76.
17. Ch. Schaerer, *Textilveredlung* **21** (1986) 81.

Fermentation → Biotechnology
Fermium → Transuranium Elements
Ferrites → Magnetic Materials

Ferroalloys

RUDOLF FICHTE, Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

Ferroalloys are master alloys containing elements that are more or less soluble in molten iron and that improve the properties of iron and steel. These alloys usually contain a significant amount of iron [*7439-89-6*]. Ferroalloys have been used for the last 100 years, principally in the production of cast iron and steel [1]. As additives, they give iron and steel improved properties, especially increased tensile strength, wear resistance, and corrosion resistance. These effects come about through one or more of the following: (1) a change in the chemical composition of the steel; (2) the removal or the tying up of harmful impurities such as oxygen, nitrogen, sulfur, or hydrogen; and (3) a change in the nature of the solidification, for example, upon inoculation [2].

Ferroalloys are also used as starting materials in the preparation of chemicals and pure metal; as reducing agents (e.g., the use of ferrosilicon to reduce rich slags); as alloying elements in nonferrous alloys; as starting materials for special products such as amorphous metals. The individual ferroalloys are treated in detail along with the alloys of the non-iron component, e.g., ferroboron is described under → Boron and Boron Alloys. Table 1 lists some commercial ferroalloys. , Chap. 2., Chap. 3.

Over the years, the binary ferroalloys have been modified by the addition of further components. For example, magnesium is regularly added to ferrosilicon, which, when used in cast iron, produces a spheroidal graphite; and manganese or zirconium can be added to calcium–silicon.

Production. Generally, ferroalloys are prepared by direct reduction of oxidic ores or concentrates with carbon (carbothermic), silicon (silicothermic), or aluminum (aluminothermic, closely related to the Goldschmidt reaction; → Aluminothermic Processes). However, roasted ores (e.g., for ferromolybdenum) and pure technical-grade oxides (for ferromolybdenum and ferroboron) are also used as starting materials.

The large-volume ferroalloys – *ferrosilicon, ferromanganese, ferrochromium* – are prepared continuously by carbothermic reduction in large, submerged-arc furnaces. The carbon is provided by coke, and the power supply ranges from 15 to 100 MVA. Ferromanganese is also produced in blast furnaces. Both ferromanganese and ferrochromium, because of their affinity for carbon, contain high concentrations of carbon (7 – 8 wt %). In contrast, ferrosilicon contains little carbon; in fact, the carbon concentration decreases as the silicon concentration is increased, e.g., ferrosilicon containing 25 wt % Si contains 1 wt % C, while that containing 75 wt % Si contains only 0.1 wt % C.

Low-carbon ferromanganese (0.1 – 2 wt % C) or ferrochromium (0.02 – 2 wt % C) is produced in the following way: a silicon-containing alloy, ferrosilicomanganese or ferrosilicochromium, is made by carbothermic reduction of an appropriate ore and quartzite in a submerged-arc furnace. Then this silicon alloy is used for the silicothermic reduction of an appropriate ore in an electric-arc refining furnace. Oxygen is seldom used to reduce the carbon content.

Ferrophosphorus is a byproduct of the carbothermic production of → Phosphorus. *Ferrotungsten* is produced carbothermically in small, high-power electric furnaces. *Ferroboron* is produced from carbon, iron, and boron oxide or boric acid in one- and three-phase electric furnaces. *Zirconium-containing ferrosilicon* is made carbothermically in electric furnaces from zircon ($ZrSiO_4$) or baddeleyite (ZrO_2); ferrosilicon, with either 75 or 90 wt % Si, is added.

Silicothermic reduction, with the addition of some aluminum, is used to prepare *ferromolybdenum* in refractory-lined reaction vessels. The energy released in the reaction is adequate to melt both metal and slag, and the solidified block of ferromolybdenum can be recovered easily. *Ferronickel* is also mainly produced by silicothermic reduction. Aluminothermic reduction is the method of choice for the production of *ferrovanadium.* The starting materials are vanadium(V) oxide, aluminum, and iron turnings,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_305

Table 1. Ferroalloys

Alloy	Composition, wt %	Keyword in Ullmann's
Calcium – silicon [*37323-90-3*]	28 – 35 Ca	Calcium and Calcium Alloys
	60 – 65 Si	
	6 Fe	
Calcium – silicon – aluminum	15 – 25 Ca	
	10 – 40 Al	
	35 – 50 Si	
Calcium – silicon – barium	15 – 20 Ca	
	14 – 18 Ba	
	55 – 60 Si	
Calcium – silicon – magnesium	25 – 30 Ca	
	10 – 15 Mg	
	50 – 55 Si	
Calcium – manganese – silicon	16 – 20 Ca	
	14 – 18 Mn	
	58 – 59 Si	
Calcium – silicon – zirconium	15 – 20 Ca	
	15 – 20 Zr	
	50 – 55 Si	
Ferroboron [*11108-67-1*]	12 – 24 B	Boron and Boron Alloys
Ferrochromium [*11114-46-8*]	45 – 95 Cr	Chromium and Chromium Alloys
	0.01 – 10 C	
Ferrosilicochromium [*11114-46-8*]	40 – 65 Cr	
	45 – 20 Si	
Ferromanganese [*12604-53-4*]	75 – 92 Mn	Manganese and Manganese Alloys
	0.05 – 8.0 C	
Ferrosilicomanganese [*53200-80-9*]	58 – 75 Mn	
	35 – 15 Si	
Ferromolybdenum [*11121-95-2*]	62 – 70 Mo	Molybdenum and Molybdenum Compounds
Ferronickel [*11133-76-9*]	20 – 60 Ni	Nickel Alloys
Ferroniobium [*11108-69-3*]	55 – 70 Nb	Niobium and Niobium Compounds
Ferroniobiumtantalum [*11108-69-3*]	55 – 70 Nb	
	2 – 8 Ta	
Ferrophosphorus [*8049-19-2*]	ca. 25 P	Phosphorus
Ferroselenium [*11108-70-6*]	ca. 50 Se	Selenium and Selenium Compounds
Ferrosilicon [*8049-17-0*]	8 – 95 Si	Silicon
Ferrotitanium [*60591-60-8*]	20 – 75 Ti	Titanium, Titanium Alloys, and Titanium Compounds
Ferrotungsten [*12604-57-8*]	70 – 85 W	Tungsten, Tungsten Alloys, and Tungsten Compounds
Ferrovanadium [*12604-58-9*]	35 – 80 V	Vanadium and Vanadium Compounds
Ferrozirconium [*11108-72-8*]	ca. 85 Zr	Zirconium and Zirconium Compounds
Ferrosilicozirconium [*51683-55-7*]	35 – 42 Zr	
	ca. 50 Si	

chips, stampings, or nail bits. The reaction is carried out in refractory-lined reaction vessels. The reaction produces enough heat to melt both metal and slag. In fact, inert materials must be added to keep the peak temperature lower. Ferrotitanium and ferroboron are also produced aluminothermically.

The production and use of titanium as a large-volume construction material yields enough scrap for *ferrotitanium* containing ca. 70 wt % Ti to be made by melting iron and titanium scrap together. *Ferrozirconium* containing 80 – 85 wt % Zr is also made by melting metal scrap and iron together. *Ferroselenium* is produced by the exothermic reaction of iron and selenium powders.

The production of ferroalloys is of great economic importance, especially in countries that have the raw materials and the energy supply. Alloy steels constitute an ever growing fraction of total steel production because of their superior properties. As an example of the importance of ferroalloys, an estimated € 0.9×10^9 (ca. \$ 0.8×10^9) of ferroalloys were consumed in 1986 by the steel industry in the Federal Republic of Germany.

Environmental Aspects. During the production of ferroalloys, the emission of undesirable substances into the environment must be kept under control. There are guidelines for this, e.g., see [3]. ""

References

1. R. Fichte, H.-J. Retelsdorf: "Stahlveredler" in *Winnacker-Küchler,* **vol. 4,** pp. 198 – 234. Durrer – Volkert: *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin-Heidelberg-New York 1972. A. G. E. Robiette: *Electric Smelting Processes,* Griffin, London 1973. F. Leitner, E. Plöckinger: *Edelstahlerzeugung,* 2nd ed., Springer Verlag, Wien-New York 1965.
2. P. D. Deeley, K.-J. A. Kundig, H. R. Spendelow, Jr.: *Ferroalloys and Alloying Additives Handbook,* Shieldalloy Corp., Newfield, N. J., Metallurg Alloy Corp., New York 1981.
3. VDI-Richtlinien VDI 2576 (May 1983): *Emissionsminderung – Elektrothermische und metallothermische Erzeugung von Ferrolegierungen (Emission Control – Electrothermic and Metallothermic Ferroalloy Production)*, vol. 2 of VDI-Handbuch Reinhaltung der Luft, Beuth Verlag, Berlin-Köln 1983.

Ferroelectrics

HANS-GÜNTHER UNRUH, Universität des Saarlandes, Fachbereich Physik, Saarbrücken, Federal Republic of Germany

1.	**Fundamental Concepts and History**	1
1.1.	**Fundamental Concepts**	1
1.2.	**Early Development**	2
1.3.	**Development of Lattice-Dynamic Aspects**	3
2.	**Properties of Specific Families**	5
2.1.	**KH_2PO_4 (KDP) Family**	5
2.2.	**Triglycine Sulfate (TGS) Family**	6
2.3.	**$(NH_4)_2SO_4$ (AS) Family**	7
2.4.	**Oxidic Ferroelectrics**	8
2.4.1.	Barium Titanate	8
2.4.2.	Other Oxides of the Perovskite Type	10
2.4.3.	$LiNbO_3$ Family	10
2.5.	**Ferroelectric Ceramics**	11
2.6.	**Ferroelectric Polymers and Liquid Crystals**	12
3.	**References**	13

1. Fundamental Concepts and History

1.1. Fundamental Concepts

Ferroelectrics are materials exhibiting a spontaneous dielectric polarization $\boldsymbol{P}_s$, the direction of which can be reversibly changed by a suitable transient field. The electric dipole moment $\boldsymbol{m} = \boldsymbol{P}_s V$ associated with $\boldsymbol{P}_s$ in a volume V results either from a displacement of the centers of charge relative to each other or from an alignment of existing dipoles, or from a combination of these effects. The two limiting forms of ferroelectrics are classified as displacive type and order – disorder type, respectively. The reversible change in direction of $\boldsymbol{P}_s$ in the case of the displacive type requires the existence of at least two distinct, physically equivalent arrangements of atoms or ions, while in the case of the order – disorder type it requires at least two equivalent orientations of molecular dipoles.

The ferroelectric state is, as a rule, observed only below a temperature T_c, the so-called Curie point, which is characteristic of the particular material. The transition from the paraelectric high-temperature phase, with $\boldsymbol{P}_s = 0$, to the ferroelectric low-temperature phase is invariably associated with a structural phase transition at T_c, in which $\boldsymbol{P}_s$ arises as a result of cooperative displacements in the structural units of the paraelectric phase. If the displacements are viewed as a result of local fields, then the molecular field theory permits calculation of the strength of these fields from the motions and interactions of the structural units.

Since any structural transition that is strictly or approximately second order can be characterized by a suitable chosen order parameter Q, where $-Q- = 0$ for $T > T_c$ and $-Q- > 0$ for $T < T_c$, it is natural in the present case to identify Q as the spontaneous polarization. Ferroelectrics for which this choice proves correct are called proper ferroelectrics. However, there exist cases for which the physical quantity that constitutes a suitable order parameter does not directly result in a polarized state at $T < T_c$. That a ferroelectric phase can nonetheless arise is most readily deduced from the existence of thermodynamic relations of the form $P \propto Q^n$,

where n is called a faintness index. Such ferroelectrics are called improper ferroelectrics. For proper ferroelectrics the spontaneous polarization $\boldsymbol{P}_s$ is the primary order parameter, while for improper ferroelectrics it is a secondary one. Examples are given in Section 1.3.

As an extensive order parameter, Q can be associated with a conjugate intensive variable, a field h, and thus also a susceptibility, $\partial Q/\partial h$.

In the case of proper ferroelectrics these entities are the electric field $\boldsymbol{E}$ and the electric susceptibility $\chi = \partial P/\partial E$,

tensor notation being avoided here in the interest of simplicity. Since stability of a thermodynamic phase requires the reciprocal susceptibilities to be positive definite, a susceptibility can be expected to exhibit singular behavior at a phase transition. Indeed, in the vicinity of the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_309

Curie point the static electric susceptibility χ_s is found as a rule to have a temperature dependence in good agreement with the Curie – Weiss law in the form

$$\chi_s - \chi_\infty = \varepsilon_0 C / (T - T_0) \quad (1)$$

where $C \lesseqgtr 0$ for $T \lesseqgtr T_0$ and ε_0 is the permittivity of vacuum.

The Curie temperature T_0 generally is only slightly below the transition temperature T_c. The term χ_∞ represents the practically temperature-independent contribution of those degrees of freedom of the system that do not belong to the order parameter.

A homogeneous ferroelectric possesses a characteristic direction determined by spontaneous polarization. It is therefore polar, noncentrosymmetric, and consequently piezoelectric. Since $\boldsymbol{P}_s$ is in general temperature dependent, ferroelectrics are also pyroelectric. If the temperature of a ferroelectric is lowered from $T > T_c$ to $T < T_c$, as a rule no homogeneous state is produced, even in the case of single crystals. Rather, the crystals exhibit a partitioning into macroscopic domains of homogeneous polarization, analogous to the Weiss domains of ferromagnetics. The form and organization of these domains vary and depend not only on properties of the material but also on lattice defects of the particular sample. The linear dimensions of domains typically lie in the µm to mm range.

Structural phase transitions are in general referred to as ferrodistortive if no change in the periodicity volume occurs during the transition, and specifically are called ferromagnetic, ferroelectric, or ferroelastic if the (primary) order parameter is the magnetization, the polarization, or an elastic deformation, respectively. In addition, however, the transition of a high-temperature to a low-temperature phase of a crystalline substance can also be accompanied by a change in the translational symmetry owing to the increase of the periodicity volume by an integer factor. In particular, structures occur in which the dipole moments in neighboring unit cells are oriented antiparallel to each other. This results in a doubling of the unit cell. Such phases are called antiferrodistortive and, in the electrical case, antiferroelectric.

The nomenclature emphasizes several phenomenological analogies between ferroelectrics and ferromagnetics, which, however, was not a fortunate choice. Occasionally, especially in the Russian literature, the term "Seignette electric" is used. This term, which was in general use until about 1940, relates the phenomenon to the first known ferroelectric, Seignette salt, another name for Rochelle salt. The salt was first prepared by E. SEIGNETTE in La Rochelle, France.

1.2. Early Development

Ferroelectric phenomena were first observed in 1920 by VALASEK [21] in Rochelle salt [*304-59-6*], $KNaC_4H_4O_6 \cdot 4\,H_2O$. In his description, he called attention to the analogy with ferromagnetic phenomena. Besides the dielectric hysteresis in the function $\boldsymbol{P}(\boldsymbol{E})$accompanying reversal of the electric field $\boldsymbol{E}$, VALASEK also noted the extraordinarily high values of components of certain characteristic tensors, such as the dielectric function $e = c + \mathbf{1}$

and the piezoelectric modulus $\mathbf{d}$, within and near the ferroelectric temperature interval, which for Rochelle salt is 255 – 297 K. The existence of a second Curie point, below which the crystal is once again paraelectric, is not common. The second structural transformation, occurring at 255 K, gives rise to a nonpolar low-temperature phase. A satisfactory microscopic theory of the ferroelectric behavior of Rochelle salt has not yet been formulated. The first thermodynamic analysis of ferroelectric transformations was that of MUELLER [22], using Rochelle salt as an example. The central theme of this analysis is a discussion of a suitable thermodynamic potential, which, for a given reference state, is expanded in powers of independent state variables such as dielectric polarization and elastic deformation. Among the most significant results are the relationships between dielectric, elastic, and piezoelectric constants of the material. These relationships allowed the observed anomalies of the latter material constants to be interpreted as consequences of the singular behavior of the dielectric function.

Initially, the ferroelectric effect was viewed as a rare curiosity, so it did not trigger any systematic studies. Not until 1935 was a second substance, potassium dihydrogen phosphate, KH_2PO_4, discovered to be ferroelectric by BUSCH and SCHERRER [23]. This discovery

was important for the further study of ferroelectrics in several respects. One consequence was the prompt discovery that a series of alkali-metal dihydrogen phosphates and arsenates isomorphic with KH_2PO_4 also exhibit ferroelectric behavior [24], thus pointing to a more general significance of these phenomena. Moreover, KH_2PO_4 has a considerably simpler lattice structure than Rochelle salt, so that in 1941 SLATER [25] was able to present a first successful microscopic model for the origin of the ferroelectric effect. Of central importance in this model are ordering processes in the system of hydrogen bridges, which result in the development of dipole moments as a consequence of a relative displacement of the anion and cation lattices.

Since KURCHATOV [26] had assumed in 1933 that orientational processes of water dipoles are the cause of the ferroelectric property of Rochelle salt, a causal relationship between the presence of hydrogen bridges and the occurrence of the ferroelectric effect was now suspected. However, this hypothesis had to be abandoned when, during the years 1945 – 1948, the dielectric anomalies of $BaTiO_3$ were independently discovered in the United States, Russia, and Japan, and WUL and GOLDMANN [27] then observed the ferroelectric transformations of this substance. The extended phenomenological treatment of ferroelectricity of GINZBURG [28] was applied in 1949 – 1954 by DEVONSHIRE [29] to the description of the three successive ferroelectric transitions of $BaTiO_3$. DEVONSHIRE expanded the Gibbs free energy of the cubic high-temperature phase as a polynomial in polarization and deformation and showed that all three ferroelectric phases correspond to minima of the free energy when changes of crystal symmetry were taken into account.

As a structure-sensitive property, ferroelectric behavior was suspected to occur also in related oxides, and indeed in the next years MATTHIAS and collaborators [30], [31] discovered it in $LiTaO_3$, $LiNbO_3$, $KTaO_3$, and $KNbO_3$. These were followed by other titanates and, in 1951, by the first known antiferroelectric, $PbZrO_3$ [32]. In that same year KITTEL [33] published a phenomenological theory of antiferroelectricity.

During the 1950s a succession of other families of ferroelectrics were discovered, confirming the possibility of spontaneous polarization in chemically and crystallographically distinct systems. The following are examples: guanidinium aluminum sulfate hexahydrate [*10199-21-0*] (GASH), $C(NH_2)_3Al(SO_4)_2 \cdot 6\,H_2O$ [34], methylammonium aluminum alum [*15243-18-2*], $CH_3NH_3Al(SO_4)_2 \cdot 12\,H_2O$ [35], ammonium sulfate (AS), $(NH_4)_2SO_4$ [36], triglycine sulfate (TGS), $(NH_2CH_2CO_2H)_3 \cdot H_2SO_4$ [37], and sodium nitrite [*7632-00-0*], $NaNO_2$ [38].

The unusual dielectric and piezoelectric properties of ferroelectric materials soon led to their utilization as ingredients of capacitors and of diverse electroacoustic transducers, such as crystal phonograph pickups and microphones. Unlike the water-soluble and moisture-sensitive crystals, such as potassium dihydrogen phosphate and Rochelle salt, barium titanate is chemically resistant. Moreover, this material is easily manufactured as a polycrystalline ceramic in almost any desired shape. These considerations, as well as the possibility of varying the properties of the material over a wide range by suitable doping, were fundamental to the extensive use of barium titanate [11], [13].

1.3. Development of Lattice-Dynamic Aspects

ANDERSON [39] and COCHRAN [40] in 1960 attributed ferroelectric transformations to the instability of the crystal lattices with respect to certain lattice vibrations, so-called soft modes. Much earlier, FRÖHLICH [41] had associated the frequency ω_T of a transverse optical lattice vibration, which appears in the Lyddane – Sachs – Teller equation

$$\omega_T^2/\omega_L^2 = \varepsilon_\infty/\varepsilon_s$$

and anomalously high values of the static dielectric constant ε_s. However, the occurrence of a particularly soft TO mode and the consequent lattice instability if $\omega_T \to 0$ as $T \to T_c^+$ was clearly explained for the first time in 1960. In particular, COCHRAN attributed the vanishing of the force constants proportional to ω_T^2 at the transition temperature T_c of ionic crystals to an equilibrium between short-range repulsive forces and long-range attractive Coulomb forces.

The equations, which were presented in a form applicable only to certain cubic crystals, can be generalized to other lattices, and motivated a rapidly growing number of spectroscopic studies, utilizing infrared absorption, Raman scattering, and inelastic neutron scattering, which confirmed the theoretical predictions [10]. CHAVES and PORTO [42] used a frequency-dependent representation of the dielectric function $\varepsilon(\omega)$ to generalize the Lyddane – Sachs – Teller equation, including the case that the temperature-dependent low-frequency excitation responsible for the transformation has the character of a relaxation:

$$\frac{\varepsilon(\omega)}{\varepsilon_\infty} = \frac{i\omega_{\mathrm{LO}}+\omega}{i\omega_{\mathrm{TO}}+\omega} \prod_j \frac{\omega_{\mathrm{L}j}^2 + i\omega\gamma_{\mathrm{L}j} - \omega^2}{\omega_{\mathrm{T}j}^2 + i\omega\gamma_{\mathrm{T}j} - \omega^2} \tag{2}$$

Here ω_{TO} and ω_{LO} represent the transverse and longitudinal relaxation frequencies, respectively, and $\omega_{\mathrm{T}j}$ and $\omega_{\mathrm{L}j}$ the characteristic frequencies of the transverse and longitudinal oscillatory excitations, respectively. In ordered phases the structural changes can, to a first approximation, be identified as those atomic displacements that are described by the eigenvector of the soft mode. Conversely, the softening of a TO mode as a function of temperature became evidence for a ferroelectric transformation.

In proper ferroelectrics the soft mode has a wave vector $\boldsymbol{q} = 0$ (Γ point of the Brillouin zone), since the order parameter is homogeneous polarization. In improper ferroelectrics the corresponding wave vector $\boldsymbol{q}$ lies within, or on the boundary of, the first Brillouin zone. An example is gadolinium molybdate [*14985-14-9*], $Gd_2(MoO_4)_3$ [43]. The wave vector of the mode belonging to the order parameter lies at the boundary of the Brillouin zone; accordingly, below the transition temperature the unit cell is doubled. In this case the crystal symmetry permits a polarization proportional to the magnitude of the two-component order parameter, as well as a deformation. The resulting ferroelectric and ferroelastic properties of the ordered phase were correctly interpreted for the first time as effects of secondary order parameters by PYTTE [44].

Finally, the case that the instability involves a soft mode for which the wave vector $\boldsymbol{q}_{\mathrm{i}}$ lies in the interior of the first Brillouin zone is exemplified by potassium selenate [*7790-59-2*], K_2SeO_4. By means of inelastic neutron scattering, IIZUMI et al. [45] in 1977 found $\boldsymbol{q}_{\mathrm{i}} = (1-\delta(T))\,\boldsymbol{c}^*/3$, where $\boldsymbol{c}*$ is a reciprocal lattice vector. Below the transformation at T_{i}, the modulation imposed on the lattice, owing to the condition $\delta(T) \ll 1$, locally causes approximately a tripling of the unit cell. The temperature-dependent modulation is incommensurable with the lattice of the unordered phase and in the case of potassium selenate gives rise to a structure locally polar, but globally centrosymmetric and thus nonpolar. Potassium selenate becomes ferroelectric at a temperature $T_{\mathrm{c}} < T_{\mathrm{i}}$ as

$$\delta(T) \longrightarrow 0$$

and a commensurable phase that has a tripled unit cell is formed. The homogeneous polarization can be viewed as a secondary order parameter that drives this lock-in process.

For an understanding of ferroelectric transformations, structure determinations of the utmost precision have always been essential [2], [13]. Although considerable effort has been invested in X-ray and neutron diffraction studies, some qualitative and quantitative aspects still remain unelucidated, even for classical ferroelectrics such as barium titanate, because the shifts in atomic coordinates are frequently extremely small. If the symmetry of the disordered and ordered phases is known, a phenomenological description of a ferroelectric can often be accomplished within the experimental accuracy over a restricted temperature range by recourse to thermodynamic methods [7], but a general microscopic theory does not exist. The treatment of special models by the methods of the theory of phase transitions has been quite successful [8], [15].

2. Properties of Specific Families

Ferroelectric substances are customarily classified into families according to chemical and structural considerations [19], [20], although this approach distinguishes neither the type (order-disorder – displacive) nor the nature (ferroelectric – antiferroelectric, proper – improper) of the phase transformation. At present, the known pure compounds with ferroelectric properties number about two hundred, which are classified into about forty families [20]. In addition, there are numerous mixed systems, including many materials of commercial importance. In the following the properties of some

representative substances are presented. A short survey of characteristic data is given in Table 1.

Table 1. Curie points and spontaneous polarizations (at a specific absolute temperature) for various ferroelectric compounds

Substance	Curie point T_c, K	Spontaneous polarization P_s, μC/cm^2	
$PbTiO_3$	763	75	(295 K)
$LiNbO_3$	1483	71	(295 K)
$Ba_2NaNb_5O_{15}$	833	40	(295 K)
$Pb_5Ge_3O_{11}$	450	4.7	(295 K)
$Gd_2(MoO_4)_3$	432	0.29	(295 K)
SbSI	295	20	(273 K)
HCl	98.4	3.7	(95 K)
KH_2PO_4	123	4.7	(110 K)
$SC(NH_2)_2$	169	3	(140 K)
$Ca_2Sr(CH_3CH_2COO)_6$	283	0.9	(200 K)
$(NH_2CH_2COOH)_3 \cdot H_2SO_4$	323	4.5	(100 K)
$NaKC_4H_4O_6 \cdot 4H_2O$	297	0.24	(275 K)
K_2SeO_4	93	0.12	(40 K)

2.1. KH_2PO_4 (KDP) Family

The ferroelectric transition of potassium dihydrogen phosphate [*7778-77-0*] (KPD), KH_2PO_4, occurs at $T_c = 123$ K and is weakly first order. Potassium dihydrogen phosphate is tetragonal (point group $\bar{4}2\,m$) in the paraelectric phase and orthorhombic ($mm\,2$) in the ferroelectric phase, and thus is piezoelectric both above and below T_c. The static electromechanical properties of the material are represented by the coefficients of a series expansion of the energy density $U(D_j\,,\,\in_\mu)$ in the components of the dielectric displacement $\boldsymbol{D}$ and the elastic deformation $\in$. If the experimental boundary conditions require other independent variables, a suitable thermodynamic potential can be obtained by a Legendre transformation of U. In the case of KDP, such a Landau formulation yields the following fundamental linear piezoelectric equations

$$E_3 = \left(1/\varepsilon_{33}^{\in}\right) D_3 - h_{36} \in_6$$
$$\sigma_6 = -h_{36} D_3 + \left(1/s_{66}^{D}\right) \in_6$$

whereby the components in the polar c direction are taken into account. The only property of the material with an anomalous temperature dependence is the permittivity at constant deformation, $\varepsilon_{33}^{\in}$, which obeys a Curie – Weiss law (Eq. 1). For the usual experimental boundary conditions of constant field $\boldsymbol{E}$ and constant mechanical stress $\boldsymbol{\sigma}$,

$$1/\varepsilon_{33}^{\sigma} = 1/\varepsilon_{33}^{\in} - h_{36}^2 s_{66}^{D}$$
$$1/s_{66}^{E} = 1/s_{66}^{D} - h_{36}^2 \varepsilon_{33}^{\in} \qquad (3)$$

Then the "free" permittivity $\varepsilon_{33}^{\sigma}$ also obeys a Curie – Weiss law, but with a higher Curie temperature than the "clamped" permittivity $\varepsilon_{33}^{\in}$

$$T_{c,\sigma} = T_{c,\in} + h_{36} s_{66}^{D} \varepsilon_0 C$$

The elastic constant s_{66}^{E}, unlike s_{66}^{D}, is also temperature dependent. The effects described by Equations (3) are the result of symmetry-allowed coupling terms between the order parameter and other state variables in the expression for the energy density, and are characteristic of ferroelectrics [7].

Microscopic theories of the transformation of KDP begin with the structural units $H_2PO_4^-$ and K^+, which are pictured in Figure 1. The hydrogen bridges oriented at right angles to the tetragonal c axis connect adjacent phosphate tetrahedra, and for $T < T_c$ cooperatively order themselves in such a way that preferentially the two upper *or* the two lower protons lie near a phosphate tetrahedron. The resulting atomic displacements lead to spontaneous polarization. The pseudo-spin formalism often employed in connection with order – disorder transformations, together with the assumption of a tunnel mechanism of the protons along the hydrogen bridges, permits interpretation of many experimental observations, including a notable isotope effect: KD_2PO_4 [*13761-79-0*] has a transition temperature 90 K higher than that of KH_2PO_4 [10], [46].

Properties resembling those of KH_2PO_4 are found for RbH_2PO_4 [*13774-16-8*] ($T_c = 147$ K), KH_2AsO_4 [*7784-41-0*] (97 K), and RbH_2AsO_4 [*13463-57-8*] (110 K). The homologous ammonium compounds are antiferroelectric.

Preparation and Applications. KH_2PO_4 and its homologues can be prepared in the form of transparent single crystals from aqueous solutions at room temperature. Controlled growth is achieved by lowering the temperature or removing solvent from the saturated solutions in the presence of seed crystals. Occasionally a two-chamber process, with circulating solution, is employed. In one chamber polycrystalline

substance is taken up by a slightly unsaturated solution, in the other chamber single crystals grow at a somewhat lower temperature.

Members of the KDP family find use primarily by virtue of their good piezoelectric, electro-optical, and nonlinear optical properties. They are employed in the construction of Kerr cells, high-frequency light modulators, and optical frequency doublers.

2.2. Triglycine Sulfate (TGS) Family

Triglycine sulfate [*513-29-1*] (TGS), $(NH_2CH_2COOH)_3 \cdot H_2SO_4$, is ferroelectric at room temperature (point group 2) and at ca. 323 K is transformed into a centrosymmetric, paraelectric phase with the point group 2/*m*. The transition is second order; the static permittivity parallel to the ferroelectric direction obeys the Curie – Weiss law, at T_c exceeding $3\times10^5\,\varepsilon_0$. In the vicinity of T_c, the temperature dependence of the spontaneous polarization is given by $\boldsymbol{P}_s \propto (T_c - T)^{1/2}$,

in agreement with thermodynamic theory [7], [8]. Curves for $\boldsymbol{P}_s(T)$ and $\varepsilon_0/\varepsilon(T)$, which were determined by experiment, are shown in Figure 2.

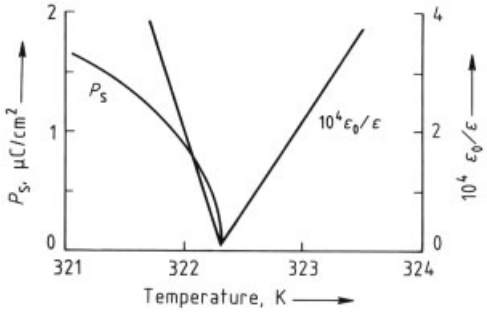

Figure 2. Temperature dependence of the reciprocal relative permittivity $\varepsilon_0/\varepsilon$ and of the spontaneous polarization P_s of triglycine sulfate near T_c

TGS is considered an excellent example of a uniaxial ferroelectric of the order – disorder type. Essentially, the transformation is attributed to the dynamics of a nonplanar glycine group, which in the paraelectric phase possesses equivalent positions in a potential with a double minimum, thus giving rise to a crystallographic mirror plane of the average structure. The other two glycine groups of the unit formula are planar and are mirror images [47]. The nonequivalence of the glycine groups can be brought out by writing the compound as

$(NH_2CH_2COOH) \cdot (NH_3CH_2COOH)_2^+ SO_4^{2-}$

The spectrum of polarization fluctuations associated with the motions of the molecular dipoles can be measured in the form of the frequency dependence of the complex permittivity

$$\varepsilon(\omega) = \varepsilon'(\omega) - i\varepsilon''(\omega)$$

In the case of TGS, the dispersion of $\varepsilon(\omega)$ is well described [48] by the Debye form for a relaxation,

$$\varepsilon(\omega) = \varepsilon_\infty + (\varepsilon_s - \varepsilon_\infty)/(1 + i\omega\tau) \tag{4}$$

The frequency $f_\varepsilon = 1/(2\pi\tau)$

likewise obeys a Curie – Weiss law and assumes the role of the soft transverse relaxation frequency in Equation (2). In the complex ε plane, the graphs for $\varepsilon(\omega)$ given by Equation (4) are semicircles. Figure 3 shows the curves for three temperatures.

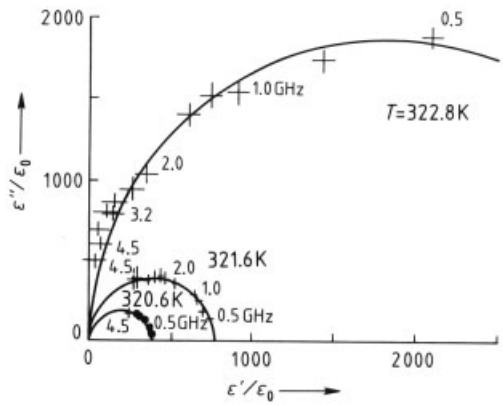

Figure 3. Diagram of the relative complex permittivity $\varepsilon(\omega)/\varepsilon_0$ of triglycine sulfate at three temperatures [48]

Numerous other glycine compounds, such as triglycine selenate [*920-25-2*],

$(NH_2CH_2COOH)_3 \cdot H_2SeO_4$

and triglycine fluoroberyllate [*2396-72-7*],

$(NH_2CH_2COOH)_3 \cdot H_2BeF_4$

have properties similar to those of TGS. In each case the deuterated compounds are ferroelectric. The resulting increase in Curie temperature is small, 11 K in the case of TGS, so that the hydrogen bridges in these structures are considered to play no special role in the transformation.

Preparation and Applications. The water-soluble crystals are prepared by the same procedures as described for KDP (see page 6). If the linear growth rate is ca. 0.5 mm/day, growth of a 100-g single crystal takes ten weeks.

By virtue of its superb pyroelectric properties, TGS is used primarily in the construction of heat sensors, infrared detectors, and thermographic transducers. If the quantity $p/(c_p\varepsilon)$ is taken as the figure of merit of a pyroelectric detector, TGS has a favorable value at room temperature owing to its high pyroelectric coefficient $p = \partial P/\partial T$

and low permittivity ε [49].

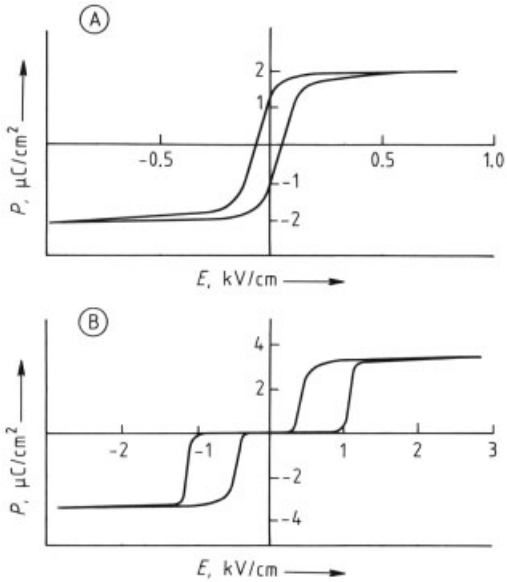

Figure 4. Hysteresis loop of ferroelectric betaine arsenate (A) [50] and of antiferroelectric betaine deuteroarsenate (B) [51]

Betaine Compounds. Ferroelectric phases also exist for compounds of the amino acid betaine with various inorganic acids. Betaine arsenate, $(CH_3)_3NCH_2COO \cdot H_3AsO_4$, is ferroelectric below $T_c = 119$ K [50]. Betaine deuteroarsenate, $(CH_3)_3NCH_2COO \cdot D_3AsO_4$, has a transition temperature 53 K higher, which suggests a significant role of hydrogen bonds in the transformation. It is remarkable that, in contrast with betaine arsenate, betaine deuteroarsenate becomes antiferroelectric at 172 K [51]. The hysteresis loops in Figure 4 show the remanent polarization at $E = 0$, namely, P_r in the case of a ferroelectric, zero in the case of an antiferroelectric. Figure 4 also shows that sufficiently high fields convert the antiferroelectric state into a ferroelectric one. The height of the subloops indicates the twofold polarization of a sublattice.

2.3. $(NH_4)_2SO_4$ (AS) Family

The ferroelectrics included in the $(NH_4)_2SO_4$ (AS) family belong to the extensive group of substances with the empirical formula A_2BX_4 that in their paraelectric high-temperature phase have the β-K_2SO_4 structure and exhibit ferroelectric properties in various low-temperature phases. Examples are $(NH_4)_2SO_4$ [*7783-20-2*], K_2SeO_4, Rb_2ZnCl_4 [*33724-11-7*], and some homologous compounds. For these compounds the polarization is, as a rule, not the primary order parameter, and in many cases a transition into an incommensurably modulated structure occurs above the ferroelectric transition (see Section 1.3).

A special case is that of $(NH_4)_2SO_4$, which at 223 K undergoes a transition from the paraelectric to a spontaneously polarized low-temperature phase. The transition is first order. The temperature dependence of the spontaneous polarization is shown in Figure 5: P_s reaches zero near 85 K, and its sign changes on further cooling [52]. In this case, distinct sublattices with opposite but no equal polarization give rise to ferrielectric behavior resembling ferrimagnetism.

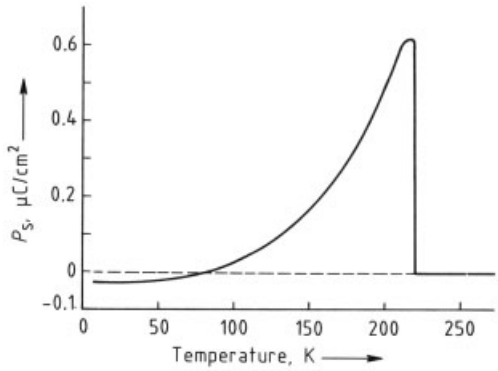

Figure 5. Temperature dependence of the spontaneous polarization of ferrielectric ammonium sulfate [52]

2.4. Oxidic Ferroelectrics

The most numerous, and commercially the most important, group of ferroelectric substances are oxides. Among these, the perovskites have been most intensively studied; they also find the most extensive industrial application in the form of polycrystalline ceramics.

2.4.1. Barium Titanate

Of all ferroelectrics, the one most thoroughly studied is barium titanate [*12047-27-7*], $BaTiO_3$. The paraelectric perovskite structure (Fig. 6) is cubic and dielectrically isotropic. Below the 403 K Curie point, essentially as the result of a relative displacement of the oxygen and titanium sublattices, there arises a dipole moment parallel to one of the cubic axes and a tetragonal distortion of the unit cell. On further reduction of the temperature, corresponding displacements and dipole moments also arise parallel to the cubic axes of the original phase that were initially not affected. As a consequence, at 278 K the crystal assumes an orthorhombic, and finally at 183 K, a rhombohedral structure. For the orthorhombic structure, $\boldsymbol{P}_s$ is parallel to a face diagonal of the original cubic cell. For the rhombohedral case $\boldsymbol{P}_s$ is parallel to a body diagonal of the original cubic cell. The permittivity obeys the Curie – Weiss law in the paraelectric phase and below the first transition experiences an especially sharp drop parallel to the polar direction (here the *c* axis, Fig. 7). An analogous temperature effect is observed for the lower-temperature transitions. In its simplest form, the thermodynamic description by Devonshire [29] of the ferroelectric transformations of $BaTiO_3$ begins with an expansion of the free elastic enthalpy density *G* in powers of the components of the polarization $\boldsymbol{P}$, while the external mechanical stress is taken to be zero:

$$\begin{aligned} G = G_0 &+ \frac{1}{2}\alpha\left(P_a^2+P_b^2+P_c^2\right) \\ &+ \frac{1}{4}\beta\left(P_a^4+P_b^4+P_c^4\right) \\ &+ \frac{1}{2}\gamma\left(P_a^2P_b^2+P_b^2P_c^2+P_c^2P_a^2\right) \\ &+ \frac{1}{6}\delta\left(P_a^6+P_b^6+P_c^6\right) \end{aligned} \tag{5}$$

where G_0 is the enthalpy density of the cubic reference phase. The individual terms in the expansion are found as the expressions invariant with respect to the symmetries of the reference phase. The equilibrium conditions

$$\partial G/\partial P_i = E_i = 0$$

lead to the desired solutions:

$P_a = P_b = P_c = 0$	for the paraelectric, cubic phase
$P_a = P_b = 0 \quad P_c \neq 0$	for the tetragonal phase
$P_a = 0, P_b = P_c \neq 0$	for the orthorhombic phase
$P_a = P_b = P_c \neq 0$	for the rhombohedral phase

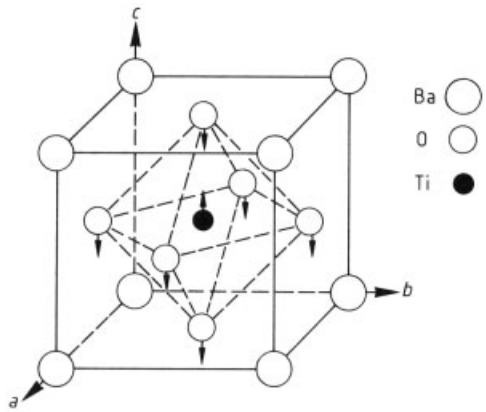

Figure 6. Unit cell of cubic barium titanate. The arrows indicate the atomic displacements that result in the tetragonal structure of the ferroelectric phase.

The temperature is explicitly introduced into the function *G* by the assumption that

$$\alpha = \alpha_0\,(T - T_0)$$

This form of the coefficient α can be justified on phenomenological [54] and latticedynamic grounds [40]. Through the relation $1/\chi = \partial^2 G/\partial P^2$,

it leads to the Curie – Weiss law. The coefficients β, γ, and δ of the expansion are, to a first approximation, viewed as temperature independent. For the spontaneous polarization of the tetragonal phase subject to equilibrium constraints, Equation (5) gives the result:

$$P_c^2 = -\,(\beta/2\delta) \pm \left[\beta^2/\left(4\delta^2\right) - \alpha_0\,(T - T_0)\right]^{1/2}$$

In the case of barium titanate, $\beta < 0$, and the transition is first order [5], [29]. The maximum thermal hysteresis for the transition at 403 K is 10 K.

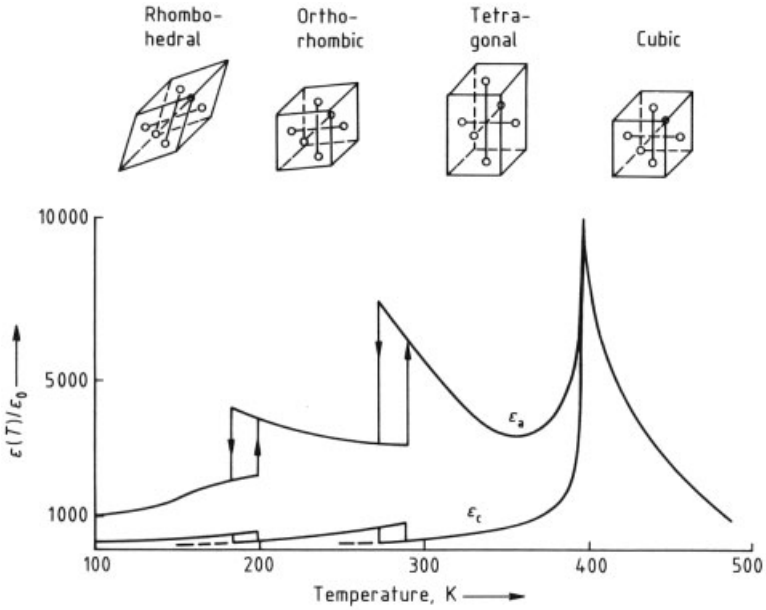

Figure 7. Lattice distortions and temperature dependence of the relative permittivitiy of $BaTiO_3$
The numerical values in all phases are for the orientations parallel (ε_c) and perpendicular (ε_a) to the tetragonal axis. In the two low-temperature phases they are only crude estimates, since in those cases the crystal has different principal axes of the dielectric tensor, and in general there is no single domain state [53].

Barium titanate was long viewed as an example of a displacive-type ferroelectric. However, structural investigations have shown that the thermal displacements of the atoms exhibit considerable anharmonicity, so order – disorder mechanisms also had to be taken into consideration. Results of infrared measurements [55] and hyper-Raman scattering [56] for the soft mode can be described within experimental error by either an oscillator model or a relaxator model. Oscillator parameters for the soft mode obtained by hyper-Raman scattering are plotted in Figure 8.

Preparation and Applications. Single crystals of $BaTiO_3$ can be obtained in sizes up to $10\,cm^3$ by lowering the temperature of a variety of fluxes, which include KF, $BaCl_2$, or excess TiO_2. Especially good results are obtained with the top-seeded solution growth technique and a TiO_2-rich melt at ca. 1630 K [16].

The standard technique of pulling single crystals out of the pure melt cannot be used for cubic barium titanate, for the melt (*mp* 1891 K) produces a hexagonal phase, which at 1733 K

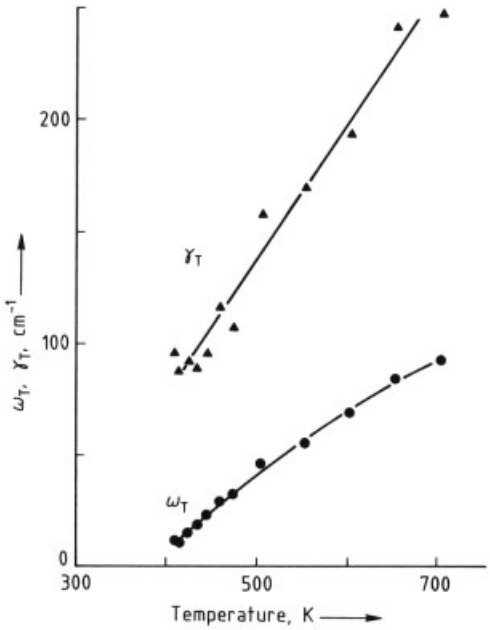

Figure 8. Temperature dependence of frequency ω_T (●) and damping coefficient γ_T (▲) of the soft mode of barium titanate obtained from measurements of the hyper-Raman effect [56]

converts into the cubic form, with consequent destruction of the crystals.

The technical uses of barium titanate single crystals are limited. As a material with a strong linear electro-optic effect, it is above all a candidate for use in light modulators, as well as for applications in dynamic holography, optical image intensification, and phase conjugation optics [57].

2.4.2. Other Oxides of the Perovskite Type

Other examples of oxidic perovskites include the ferroelectrics $KNbO_3$ [*12030-85-2*] and $PbTiO_3$ [*12060-00-3*], which wholly or partly duplicate the succession of phases exhibited by $BaTiO_3$, as well as the antiferroelectrics $PbZrO_3$ [*12060-01-4*] and $NaNbO_3$ [*12034-09-2*]. Other perovskites, for example $KTaO_3$ [*57126-93-9*], show a dielectric behavior conforming to the Curie – Weiss law, but no ferroelectric transition is observed down to low temperatures (incipient ferroelectrics). At temperatures below ca. 2 K quantum-mechanical zero-point vibrations can suppress the expected ferroelectric ordering. The ranks of ferroelectric perovskites are extended by a multitude of complex compounds as well as solid solutions of simple or complex perovskites [19], [20].

2.4.3. $LiNbO_3$ Family

The trigonal lattice structure of $LiNbO_3$ [*12031-63-9*] contains oxygen octahedra that are connected through their faces and that contain Li, Nb, or no atom at all at their center. (The oxygen octahedra in perovskite, in contrast, are connected through their corners.) The ferroelectric transition occurs at $T_c = 1483$ K, thus only slightly below the melting point of 1526 K. A member of the same family, $LiTaO_3$ [*12031-66-2*], has a substantially lower Curie point, $T_c = 893$ K.

Preparation and Applications. Lithium niobate and lithium tantalate can be produced as single crystals of centimeter dimensions and good optical quality from their melts by the Czochralski technique. The acentricity at room temperature associated with the high spontaneous polarization gives rise to strong piezoelectric, electro-optic, and nonlinear optical effects. Because of these properties, the materials find use in electro-acoustic filters, light modulators, optical amplifiers, and frequency multipliers [13], [57].

Materials with high electro-optical coefficients and high photoconductivity exhibit light-induced changes in refractive index. This *photorefractive effect* is a consequence of the presence of fields associated with space charges that are formed under conditions of nonhomogeneous illumination. The mechanism is that charge carriers freed by light from defects are captured by traps in locations where the light intensity is low. Good photorefractive properties, of use in the storage of phase holograms and for optical image processing, are found in $LiNbO_3$ and $KNbO_3$, but also in $BaTiO_3$ and in $Ba_2NaNb_5O_{15}$ [*12323-03-4*], a compound of the tungsten bronze type. Photoconductivity is often achieved by suitable doping, e.g., with iron.

2.5. Ferroelectric Ceramics

Perovskite-based ceramics are commercially produced in large quantities and find use principally in capacitors and positive temperature coefficient (PTC) thermistors and as piezoceramics.

Barium titanate ceramic is employed as a dielectric in capacitors. The dependence of the permittivity on temperature and field strength, which as a rule is undesirable, can be minimized by partial substitution of Sr or Pb for Ba as well as Zr or Sn for Ti. At the same time, the transition, which in any case is diffuse in heavily doped barium titanate, can be lowered to room temperature, thus achieving high permittivities [11]. Attention must also be paid to the grain size, because of the associated depolarization effects.

Light doping, e.g., a few tenths of mole percent of Sb, Nb, or rare earth, makes barium titanate semiconducting. If the ceramic also contains traces of Cu, Fe, or Mn, accumulation of these elements at the grain boundaries causes electrically insulating layers to form, since these impurities, as acceptors, capture free electrons.

If the acceptor concentration is sufficiently high, the ceramic constitutes a blocking layer capacitor with high effective permittivity. Lighter doping (< 0.01 %) produces so-called PTC thermistors; their resistivity can increase by more than four orders of magnitude in a 100-K interval above the Curie temperature (Fig. 9) [58]. Below the Curie point the potential barriers at the grain boundaries are nullified by polarization charges of the ferroelectric crystallites, and the electric resistance of the ceramic is determined by that of the semiconducting grains.

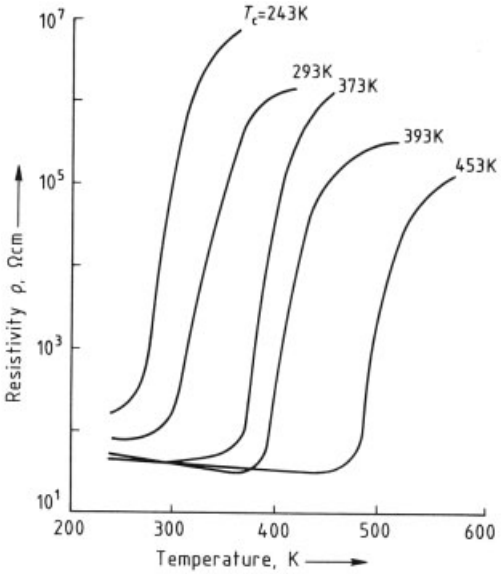

Figure 9. Temperature dependence of the resistivity of barium titanate ceramic PTC thermistors with various Curie temperatures

For piezoelectric applications the macroscopically isotropic ferroelectric ceramic is polarized in electric fields. The optimal orientation of the spontaneous polarization in the randomly arranged grains is facilitated by elevated temperatures. Ceramics of lead titanate – lead zirconate solid solutions (PZT) [*12626-81-2*] are particularly well suited for piezoceramic applications by virtue of their high Curie temperatures and high spontaneous polarization as well as their crystallographic properties. The phase diagram is shown in Figure 10. Materials with a composition close to that of the morphotropic phase boundary between the zirconate-rich rhombohedral phase and the titanate-rich tetragonal crystal structure exhibit coupling factors up to $k = 0.7$, where k^2 is the fraction of the electric energy supplied that is stored as mechanical energy [6], [11].

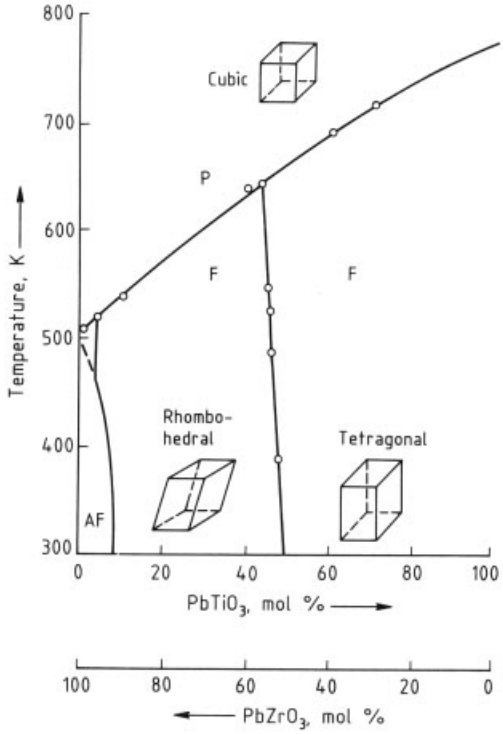

Figure 10. Phase diagram of the $PbTiO_3 - PbZrO_3$ (PZT) solid solution
AF = antiferroelectric; F = ferroelectric; P = paraelectric

The good electro-optical properties of oxidic ferroelectrics and the advantages of ceramic materials are incorporated in *opto-ceramics*. High transparency is achieved by selection of phases of minimal optical anisotropy and by special sintering procedures as in the case of lanthanum-doped lead titanate – lead zirconate (PLZT) [13, pp. 540 – 544]. Materials with the composition $Pb_{1-1.5x}La_x \square_{0.5x}(Zr_{0.65}Ti_{0.35})O_3$ with $0.07 < x < 0.1$ are used for electro-optical components such as readouts or optical shutters.

Preparation. Ferroelectric ceramics are prepared by sintering the oxide mix for several hours at temperatures between ca. 1450 and 1650 K. Starting materials are the milled oxides or, for especially fine-grained products, compounds prepared by wet-chemical methods. The low-porosity ceramics required for optical appli-

cations are obtained by hot pressing or vacuum sintering [11].

2.6. Ferroelectric Polymers and Liquid Crystals

Mechanically stretched and electrically polarized foils of the semicrystalline polymer poly-(vinylidene fluoride) [*24937-79-9*] (PVDF),

exhibit strong piezoelectric and pyroelectric effects. The further observation that PVDF shows dielectric hysteresis led to the assumption that it is ferroelectric, although no phase transition occurs up to the melting point (~ 440 K) of the crystallites of the orthorhombic β phase. Models proposed by KEPLER [59] and by BROADHURST et al. [60] assume cooperative motions of the polar chain molecules and predict a maximum spontaneous polarization of 22 μC/cm^2. A study of copolymers of PVDF with polytrifluoroethylene [61] found evidence for a transformation of a nonpolar high-temperature phase into a ferroelectric low-temperature phase.

Ferroelectric polymers find applications principally as piezoelectric transducers in the form of thin foils.

Ferroelectric properties of liquid crystal systems were first predicted in 1975 by MEYER and collaborators [62], [63]. The required symmetry constraints are satisfied by certain mesogenic phases found in the smectic group of liquid crystals: in the so-called C$*$ phase, rod-shaped, chiral molecules are arranged in layers. The long axes of the molecules are parallel to each other and inclined at an equal angle to the plane of the layer; thus, the macroscopic symmetry belongs to the monoclinic class 2. The polar axis lies parallel to the twofold rotational axis in the plane of the layer. However, the chiral character of the molecules causes the molecular axes in adjacent layers to assume a helical arrangement with respect to the normal to the layer. Consequently, a macroscopic polarization does not arise in the field-free state. In a sufficiently strong electric field directed parallel to the planes of the layers, the direction of the spontaneous polarization can be altered as in the case of a regular ferroelectric. An example of such a liquid crystal ferroelectric is *p*-decyloxybenzylidene-*p*′-amino-2-methyl-butyl-cinnamate [*97335-57-4*]:

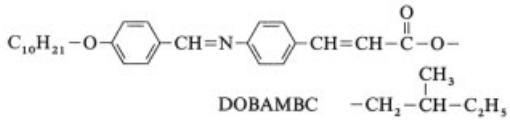

The phase transformations of this system have been studied in detail [64], [65]. As for incommensurable systems, the dielectric polarization is not the primary order parameter for this system.

3. References

1. W. Känzig: *Ferroelectrics and Antiferroelectrics,* Academic Press, New York 1957.
2. H. D. Megaw: *Ferroelectricity in Crystals,* Methuen & Co., London 1957.
3. F. Jona, G. Shirane: *Ferroelectric Crystals,* MacMillan, New York 1962.
4. H. J. Martin: *Die Ferroelektrika,* Akademische Verlagsgesellschaft, Leipzig 1964.
5. E. Fatuzzo, W. J. Merz: *Ferroelectricity,* North-Holland, Amsterdam 1967.
6. B. Jaffe, W. R. Cook, H. Jaffe: *Piezoelectric Ceramics,* Academic Press, London 1971.
7. J. Grindlay: *An Introduction to the Phenomenological Theory of Ferroelectricity,* Pergamon Press, New York 1970.
8. H. E. Stanley: *Introduction to Phase Transitions and Critical Phenomena,* Clarendon Press, Oxford 1971.
9. G. A. Smolenskij, N. N. Krainik: *Ferroelektrika und Antiferroelektrika,* Teubner, Leipzig 1972.
10. R. Blinc, B. Zeks: *Soft Modes in Ferroelectrics and Antiferroelectrics,* Elsevier, New York 1974.
11. A. Bauer et al.: *Technologie und Anwendungen von Ferroelektrika,* Akademische Verlagsgesellschaft, Leipzig 1976.
12. T. Mitsui, I. Tatsuzaki, E. Nakamura: *An Introduction to the Physics of Ferroelectrics,* Gordon and Breach, New York 1976.
13. M. E. Lines, A. M. Glass: *Principles and Applications of Ferroelectrics and Related Materials,* Clarendon Press, Oxford 1977.

14. V. M. Fridkin: *Photoferroelectrics,* Springer Verlag, Berlin 1979.
15. W. Gebhardt, U. Krey: *Phasenübergänge und kritische Phänomene,* Vieweg, Braunschweig 1980.
16. D. Elwell, H. J. Scheel: *Crystal Growth from High-Temperature Solutions,* Academic Press, London 1975.
17. R. Bechmann et al.: "Elastic, Piezoelectric, Piezooptic and Electrooptic Constants of Crystals" in *Landolt-Börnstein* **III/1** (1966); *Landolt-Börnstein* **III/2** (1969).
18. M. M. Choy et al.: "Elastic, Piezoelectric, Pyroelectric, Piezooptic, Electrooptic Constants, and Nonlinear Dielectric Susceptibilities of Crystals" in *Landolt-Börnstein* **III/11**(1979).
19. T. Mitsui et al.: "Ferro- and Antiferroelectric Substances" in *Landolt-Börnstein* **III/3** (1969); *Landolt-Börnstein* **III/9** (1975).
20. T. Mitsui et al.: "Ferroelectric and Related Substances" in *Landolt-Börnstein* **III/16 a** (1981); *Landolt-Börnstein* **III/16 b** (1982).

Specific References

21. J. Valasek, *Phys. Rev.* **15** (1920) 537 – 538; **17** (1921) 475 – 481.
22. H. Mueller, *Phys. Rev.* **57** (1940) 829 – 839; **58** (1940) 565 – 573; 805 – 811.
23. G. Busch, P. Scherrer, *Naturwissenschaften* **23** (1935) 737.
24. G. Busch, *Helv. Phys. Acta* **11** (1938) 269 – 298.
25. J. C. Slater, *J. Chem. Phys.* **9** (1941) 16.
26. I. V. Kurchatov, *Segnetoelektriki,* Moscow 1933.
27. B. Wul, I. M. Goldmann, *Dokl. Akad. Nauk SSSR* **46** (1945) 139; **49** (1945) 177; **51** (1946) 21.
28. V. L. Ginzburg, *Zh. Eksp. Teor. Fiz.* **15** (1945) 739; **19** (1949) 36.
29. A. F. Devonshire, *Philos. Mag.* **40** (1949) 1040 – 1063; **42** (1951) 1065 – 1079; *Adv. Phys.* **3** (1954) 85 – 130.
30. B. T. Matthias, *Phys. Rev.* **75** (1949) 1771.
31. B. T. Matthias, J. P. Remeika, *Phys. Rev.* **76** (1949) 1886 – 1887.
32. E. Sawaguchi, H. Maniwa, S. Hoshino, *Phys. Rev.* **83** (1951) 1078.
33. C. Kittel, *Phys. Rev.* **82** (1951) 729 – 732.
34. A. N. Holden, B. T. Matthias, W. J. Merz, J. P. Remeika, *Phys. Rev.* **98** (1955) 546.
35. R. Pepinsky, F. Jona, G. Shirane, *Phys. Rev.* **102** (1956) 1181 – 1182.
36. B. T. Matthias, J. P. Remeika, *Phys. Rev.* **103** (1956) 262.
37. B. T. Matthias, C. E. Miller, J. P. Remeika, *Phys. Rev.* **104** (1956) 849 – 850.
38. S. Sawada, S. Nomura, S. Fujii, I. Yoshida, *Phys. Rev. Lett.* **1** (1958) 320 – 324.
39. P. W. Anderson in G. I. Skanavi (ed.): *Fizika Dielektrikov,* Akad. Nauk SSSR, Moscow 1960.
40. W. Cochran, *Adv. Phys.* **9** (1960) 387 – 423.
41. H. Fröhlich: *Theory of Dielectrics,* Clarendon Press, Oxford 1949, p. 159.
42. A. S. Chaves, S. P. S. Porto, *Solid State Commun.* **13** (1973) 865 – 868.
43. H. J. Borchardt, P. E. Bierstedt, *J. Appl. Phys.* **38** (1967) 2057 – 2060.
44. E. Pytte, *Solid State Commun.* **8** (1970) 2101 – 2104.
45. M. Iizumi, J. D. Axe, G. Shirane, K. Shimaoka, *Phys. Rev. B: Solid State* **15** (1977) 4392 – 4411.
46. R. Blinc, *J. Phys. Chem. Solids* **13** (1960) 204 – 211.
47. K. Itoh, T. Mitsui, *Ferroelectrics* **5** (1973) 235 – 251.
48. G. Luther, H. E. Müser, *Z. Naturforsch. A: Astrophys. Phys. Phys. Chem.* **24 A** (1969) 389 – 392; *Z. Angew. Phys.* **29** (1970) 237 – 240.
49. S. T. Liu, D. Long, *Proc. IEEE* **66** (1978) 14 – 26.
50. A. Klöpperpieper, H. J. Rother, J. Albers, K. H. Ehses, *Ferroelectr. Lett. Sect.* **44** (1982) 115 – 120.
51. J. Albers, A. Klöpperpieper, H. E. Müser, H. J. Rother, *Ferroelectrics* **54** (1984) 45 – 48.
52. H.-G. Unruh, *Solid State Commun.* **8** (1970) 1951 –1954.
53. W. J. Merz, *Phys. Rev.* **76** (1949) 1221.
54. L. D. Landau, E. M. Lifschitz: *Statistische Physik,* Akademie Verlag, Berlin 1966, p. 469.
55. J. L. Servoin, Y. Luspin, F. Gervais, *Ferroelectrics* **37** (1981) 523 – 526.
56. H. Vogt, *Jpn. J. Appl. Phys. Part 2* **24** (1985) 112 – 116.
57. P. Guenter, *Ferroelectrics* **24** (1980) 35 – 42; **49** (1983) 39 – 46.
58. W. Heywang, *J. Am. Ceram. Soc.* **47** (1964) 484; *J. Mater. Sci.* **6** (1971) 1214 – 1226.
59. R. G. Kepler, *Annu. Rev. Phys. Chem.* **29** (1978) 497 – 518.
60. M. G. Broadhurst, G. T. Davis, J. E. McKinney, R. E. Collins, *J. Appl. Phys.* **49** (1978) 4992 – 4997.
61. A. J. Lovinger, T. Furukawa, G. T. Davis, M. G. Broadhurst, *Polymer* **24** (1983) 1225 – 1239.

62. R. B. Meyer, L. Liebert, L. Strzelecki, P. Keller, *J. Phys. Lett.* **36** (1975) 69 – 71.
63. R. B. Meyer, *Mol. Cryst. Liq. Cryst.* **40** (1977) 33.
64. G. Durand, Ph. Martinot-Lagarde, *Ferroelectrics* **24** (1980) 89 – 97.
65. B. Zeks, *Mol. Cryst. Liq. Cryst.* **114** (1984) 259.

Ferromanganese → Manganese and Manganese Alloys
Ferrotitanium → Titanium and Titanium Alloys
Ferrotungsten → Tungsten, Tungsten Alloys, and Tungsten Compounds
Ferrovanadium → Vanadium and Vanadium Compounds

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Fertilizers

Standard Article

Heinrich W. Scherer[1], Konrad Mengel[2], Heinrich Dittmar[3], Manfred Drach[4], Ralf Vosskamp[4], Martin E. Trenkel[5], Reinhold Gutser[6], Günter Steffens[7], Vilmos Czikkely[3], Titus Niedermaier[8], Reinhardt Hähndel[4], Hans Prün[9], Karl-Heinz Ullrich[4], Hermann Mühlfeld[10], Wilfried Werner[11], Günter Kluge[12], Friedrich Kuhlmann[13], Hugo Steinhauser[14], Walter Brändlein[3], Karl-Friedrich Kummer[4]

[1]Agrikulturchemisches Institut, Universität Bonn, Bonn, Federal Republic of Germany
[2]Institute for Plant Nutrition, Justus-Liebig-Universität Giessen, Giessen, Federal Republic of Germany
[3]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[4]BASF Aktiengesellschaft, Limburgerhof, Federal Republic of Germany
[5]Eusserthal, Federal Republic of Germany
[6]Lehrstuhl für Pflanzenernährung, Technische Universität München-Weihenstephan, Freising, Federal Republic of Germany
[7]Landwirtschaftliche Untersuchungs- und Forschungsanstalt, Oldenburg, Federal Republic of Germany
[8]formerly BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[9]formerly BASF Aktiengesellschaft, Limburgerhof, Federal Republic of Germany
[10]formerly Chemische Fabrik Kalk GmbH, Köln, Federal Republic of Germany
[11]Agrikulturchemisches Institut der Universität Bonn, Bonn, Federal Republic of Germany
[12]Bundesministerium für Ernährung, Landwirtschaft und Forsten, Bonn, Federal Republic of Germany
[13]Institut für Betriebslehre der Agrar- und Ernährungswirtschaft der Justus-Liebig-Universität Giessen, Giessen, Federal Republic of Germany
[14]formerly Lehrstuhl für Wirtschaftslehre des Landbaues, Technische Universität München, Freising, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a10_323.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Plant Nutrition and Soil Science
2.1. Plant Nutrients
2.1.1. Definition and Classification
2.1.2. Function of Plant Nutrients
2.2. Soil Science
2.2.1. Soil Classes, Soil Types, and Parent Material
2.2.2. Nutrient Retention in Soils
2.2.3. Soil pH, Buffer Power, and Liming
2.2.4. Soil Water–Plant Relationships
2.2.5. Organic Matter of Soils and Nitrogen Turnover
2.3. Nutrient Availability
2.3.1. Factors and Processes
2.3.2. Determination of Available Plant Nutrients in Soils
2.4. Physiology of Plant Nutrition
2.4.1. Nutrient Uptake and Long-Distance Transport in Plants
2.4.2. Effect of Nutrition on Growth, Yield, and Quality
2.5. Nutrient Balance
2.5.1. Gains and Losses of Plant Nutrients
2.5.2. Alternative Plant Nutrition
3. Standard Fertilizers
3.1. Solid Fertilizers
3.1.1. Straight Fertilizers
3.1.2. Multinutrient Fertilizers
3.1.3. Lime Fertilizers
3.1.4. Magnesium Fertilizers
3.2. Liquid Fertilizers
3.2.1. Nitrogen Liquids
3.2.2. Multinutrient Liquids
3.2.2.1. NP Liquids
3.2.2.2. NPK liquids
3.2.2.3. UAS Liquids
3.2.3. Suspensions
4. Special Fertilizers
4.1. Water-Soluble Nutrient Salts
4.2. Foliar Fertilizers
4.2.1. Production
4.2.2. Application
4.2.3. Combination with Agricultural Pesticides

4.3. Micronutrients
4.3.1. Micronutrient Forms
4.3.2. Production
4.3.3. Commercial Fertilizers
4.3.4. Use
4.4. Slow- and Controlled-Release Fertilizers
4.4.1. Introduction
4.4.2. Urea–Aldehyde Slow-elease Fertilizers
4.4.2.1. Urea–Formaldehyde Condensation Products
4.4.2.2. Other Urea– Aldehyde Condensation Products
4.4.2.3. Further Processing of Urea–Aldehyde Condensates
4.4.3. Other Organic Chemicals
4.4.4. Inorganic Compounds
4.4.5. Coated and Encapsulated Controlled-Release Fertilizers
4.4.5.1. Sulfur-Coated Controlled-Release Fertilizers
4.4.5.2. Sulfur-Coated, Polymer-Encapsulated Controlled-Release Fertilizers
4.4.5.3. Polymer-Encapsulated Controlled-Release Fertilizers
4.4.6. Anti-Float Materials
4.4.7. Controlled-Release Fertilizers on Carriers
4.4.8. Supergranules
4.4.9. Legislation
4.5. Nitrification and Urease Inhibitors
4.5.1. Introduction
4.5.2. Types of Nitrification and Urease Inhibitors
4.5.3. Pyridines
4.5.3.1. Nitrapyrin
4.5.3.2. Other pyridines
4.5.4. Dicyandiamide
4.5.5. Pyrazoles
4.5.5.1. 1-Carbamoyl-3-methylpyrazole
4.5.5.2. Outlook
4.5.6. Neem/Neem-Coated Urea
4.5.7. Urease Inhibitors
4.5.8. Environmental Aspects
4.5.9. Legal Requirements
4.6. Organic Fertilizers (Secondary Raw Material Fertilizers)
4.6.1. Fertilizers Based on Peat or Materials of Similar Stability
4.6.2. Fertilizers Based on Waste Materials of Animal Origin
4.6.3. Fertilizers Based on Wastes of Plant Origin
4.6.4. Fertilizers Based on Municipal Waste
4.7. Manure
4.7.1. Composition
4.7.2. Manure Nutrient Efficiency
4.7.3. Environmental Aspects
5. Fertilizer Granulation
5.1. Introduction
5.2. Granulator Feedstocks
5.3. Granulation Equipment
5.3.1. Pug Mill
5.3.2. Drum Granulator
5.3.3. Pan Granulator
5.3.4. The Granulator–Mixer
5.3.5. Roll Presses
5.4. Costs of Agglomeration
5.5. Bulk Blending
5.6. Quality Inspection
5.7. Fertilizer Conditioning
5.8. Environmental Aspects
6. Analysis
6.1. Sampling and Sample Preparation
6.2. Determination of Nitrogen
6.3. Determination of Phosphate
6.4. Determination of Potassium
6.5. Analysis of Calcium, Magnesium, and Trace Elements
7. Synthetic Soil Conditioners
7.1. Foams
7.1.1. Closed-Cell Expandable Polystyrene Foam
7.1.2. Primarily Open-Cell Urea–Formaldehyde Resin Foams
7.2. Colloidal Silicates
7.3. Polymer Dispersions and Polymer Emulsions
7.4. Tensides
8. Storage, Transportation, and Application
8.1. General Storage Requirements
8.2. Application

9. Environmental Aspects of Fertilizer Application
9.1. Nitrogen
9.1.1. Ground Water
9.1.2. Surface Waters
9.1.3. Atmosphere
9.1.4. Biosphere
9.1.5. Pedosphere (Soil)
9.1.6. Countermeasures
9.2. Phosphorus
9.2.1. Eutrophication
9.2.2. Heavy Metals Buildup
10. Legal Aspects
11. Economic Aspects
11.1. Economics of Fertilization
11.1.1. Input–Output Relationships: The Yield Function
11.1.2. Factors Controlling the Optimal Nitrogen Fertilization Level
11.1.3. Factors Influencing the Optimal Nitrogen Fertilization Level
11.1.4. Environmental Aspects of Fertilization
11.2. World Consumption, Production, and Trade
11.3. Future Outlook
11.3.1. Food Situation
11.3.2. Development of Fertilizer Consumption

Fertilizers are products that improve the levels of available plant nutrients and/or the chemical and physical properties of soil. An overview is given over the chemical and physical aspects of plant nutrition uptake and soil properties. The different categories of fertilizers are discussed, and special interest is given on production processes and analyses, including storage and transportation as well as environmental, legal, and economic aspects.

[Top of Page]

1. Introduction

Heinrich W. Scherer and Konrad Mengel

Fertilizers in the broadest sense are products that improve the levels of available plant nutrients and/or the chemical and physical properties of soil, thereby directly or indirectly enhancing plant growth, yield, and quality.

Fertilizers are classified as follows in terms of their chemical composition:

1. Mineral fertilizers consist of inorganic or synthetically produced organic compounds.
2. Organic fertilizers are waste products from animal husbandry (stable manure, slurry manure), plant decomposition products (compost, peat), or products from waste treatment (composted garbage, sewage sludge).
3. Synthetic soil conditioners are products whose main function is to improve the physical properties of soils, for example, friability and water and air transport.

The following categories are distinguished with respect to nutrient content:

1. Straight fertilizers generally contain only one primary nutrient.
2. Compound (complex or multinutrient) fertilizers contain several primary nutrients and sometimes micronutrients as well.
3. Micronutrient fertilizers contain nutrients required in small quantities by plants, as opposed to macronutrients; quantities range from 1 to 500 g $ha^{-1}a^{-1}$.

Finally, fertilizers can be classified as solid or liquid fertilizers and as soil or foliar fertilizers, the latter being applied exclusively by spraying on an existing plant population.

History. Fertilizing substances were applied even in antiquity. Their use can be attributed to the observation in nature that plants developed especially well in locations where human or animal excreta, ash residues, river mud, or dying plants were left. For example, the Egyptians knew about the fertility of the Nile mud, and the Babylonians recognized the value of stable manure; for example, HOMER mentions manure in the *Odyssey*. PLINY reports that the Ubians north of Mainz used “white earth,” a calcareous marl, to fertilize their fields. The Romans acknowledged the advantages of green manuring, cultivating legumes and plowing them under. At the end of the first millenium, wood ash was much used as fertilizer in Central Europe. Not until the beginning of the 19th century did guano, at the suggestion of ALEXANDER VON HUMBOLDT (1800), and Chilean caliche, on the recommendation of HAENKES (1810), come into use as fertilizers. Up to that time, however, it was still believed that the organic matter of soil, humus, was the true source of plant nutrition.

Around 1800, the nutrition problem entered a critical phase in Europe. In 1798, MALTHUS set forth his pessimistic theses, saying that the quantity of food could increase only in arithmetic progression while the population grew geometrically. Combining results obtained by others (SPRENGEL, BOUSSINGAULT) with his own pathbreaking studies, J. VON LIEBIG set forth the theoretical principles of plant nutrition and plant production in *Chemistry in Its Application to Agriculture and Physiology* (1840). He took the view, now considered obvious, that plants require nitrogen, phosphate, and potassium salts as essential nutrients and extract them from the soil. LIEBIG's *mineral theory* was well supported by experimental data of J. B. BOUSSINGAULT (1802 – 1887) in France. He and also J. B. LAWES (1814 – 1900) and J. H. GILBERT (1827 – 1901) in England showed that plants benefit from inorganic N fertilizers. LIEBIG thus became the founder of the theory of mineral fertilizers, and his doctrines led to an increasing demand for them. A number of companies were subsequently founded in Europe to produce phosphate and potash fertilizers. Superphosphate was manufactured for the first time in 1846, in England.

In Germany, this industrial development started in 1855. The importation of saltpeter on a large scale began in the area of the German Federation (56 000 t in 1878). Peruvian guano soon came into heavy use (520 000 t in 1870). Ammonium sulfate, a coke-oven byproduct, was later recognized as a valuable fertilizer, and the mining of water-soluble potassium minerals was undertaken in the 1860s [1].

The demand for nitrogen that developed at the end of the 19th century soon outstripped the availability of natural fertilizers. A crucial breakthrough came about with the discovery and large-scale implementation of ammonia synthesis by HABER (1909) and its industrial realization by BOSCH (1913).

Around the turn of the century, the technique of hydroponics led to the discovery of other essential plant nutrients. Research showed that plants in general require ten primary nutrients: carbon, hydrogen, oxygen. nitrogen, phosphorus, potassium, calcium, magnesium, sulfur, and iron. JAVILLIER and MAZE (1908) pointed out for zinc and AGULHON (1910) pointed out for boron the nutritional effects on plants. WARINGTON (1923) first described the symptoms of boron deficiency, and BRANDENBURG (1931) clearly recognized dry rot in the sugar beet as boron deficiency. Generally micronutrients were made available to the plant as liquid foliar fertilizer, a method first suggested for iron by GRIS in 1844. By 1950, this list of micronutrients had been expanded to include manganese, copper, and molybdenum.

Almost 70 years ago, serious research began on the best nutrient forms for individual plant species under various soil and climatic conditions. Besides the classical fertilizers, for example, controlled-release fertilizers, improved foliar fertilizers, nutrient chelates, and nitrification inhibitors have been developed in

recent decades. This development of new nutrient forms is still in full swing in the special fertilizers sector.

In the developed market economies of Western Europe, the United States, and Japan, however, the level of mineral fertilizer use has not been increasing since the beginning of the 1980s. In some countries, genuine agricultural overproduction has occurred recently. Since better delivery of plant nutrients has led to increasing self-reliance even in the Third World economies (e.g., China, India, Brazil), these countries are not so important as purchasers of nutrients on the world market, so that surpluses cannot be exported without limit. The production of fertilizers is also on the increase in these countries. Thus overproduction plus regional environmental problems (nitrates entering the ground water) are actually leading to a decrease in mineral fertilizer use in some areas. This decline will be limited by diminishing soil fertility in localities where fertility has been enhanced by decades of proper fertilization.

[Top of Page]

2. Plant Nutrition and Soil Science

Heinrich W. Scherer and Konrad Mengel

The science of plant nutrition is situated between soil science and plant physiology. It comprises the definition of the elements nutritive for plants; the uptake of plant nutrients and their distribution in the plant; the function of the nutritive elements in plant metabolism; their effect on plant growth; yield formation and quality parameters in crops; soil nutrient exploitation by plant roots; factors and processes that control the plant nutrient availability in soils; toxic elements in soils and their impact on plant growth; the application of plant nutrient carriers (fertilizers) and their turnover in soils; nutrient balance; and the maintenance of soil fertility.

Plant nutrition is considered mainly from two aspects, an agronomic one and an ecological one. The former is focused on the question of fertilizing soil as an efficient means to increase crop yield and to maintain or even improve soil fertility. The latter, the ecological aspect of plant nutrition, is concerned with the nutritive condition of a soil and a location and with its effect on plant growth and plant communities. Since fertilizers are the topic of this article, the agronomic aspects of plant nutrition are treated with greater depth.

The science of plant nutrition is closely associated with the science of soils. The latter comprises a broad field of scientific activity and thus cannot be considered here in all its facets. In this article only those problems of soil science relevant to understanding plant nutrition are treated.

2.1. Plant Nutrients

2.1.1. *Definition and Classification*

From a scientific point of view, the term plant nutrient is not especially precise. More appropriate is to distinguish between nutritive elements of plants and nutritive carriers. Essential *nutritive elements* for plants are the chemical elements that are required for a normal life cycle and that satisfy the following criteria:

1. A deficiency of the element makes it impossible for the plant to complete its life cycle.
2. The deficiency is specific for the element in question.
3. The element is directly involved in the nutrition of plants because of either its chemical or its physical properties.

According to this definition, the following chemical elements are nutritive elements for plants: C, H, O, N, P, S, K, Ca, Mg, Fe, Mn, Cu, Zn, Mo, B. Further elements, such as Na, Cl, and Si, may affect plant growth positively, and there are particular plant species for which these elements are of great importance. Nevertheless, they are not essential nutritive elements for plants in the strict sense of the definition. Cobalt is required by some bacteria, e.g., by dinitrogen-fixing bacteria and thus may also benefit plant growth indirectly.

Generally it is not the element itself that is provided to and taken up by the plant, but an ion or a molecule in which the nutritive element is present, e.g., C is present in CO_2, P in $H_2PO_4^-$, N in NO_3^- or NH_4^+, and B in H_3BO_3. The particular molecule or ion in which the nutritive element is present is termed the *nutrient carrier*. In the case of metals, the corresponding ion or salts of ion species in question, e.g., K^+, Ca^{2+}, Zn^{2+}, can be considered the carrier. In this sense fertilizers are nutrient carriers.

Plant nutrients may be grouped into macronutrients and micronutrients. *Macronutrients* are required in high amounts and thus are present in plant tissues in much higher concentrations than the *micronutrients*. Carbon, H, O, N, P, S, K, Ca, and Mg belong to the macronutrients. Their concentration in the dry plant matter is in the range 1 – 50 mg/g, except for C, H, and O, which have much higher concentrations (see Table 2). The concentration of the micronutrients in the dry plant matter is in the range 1 – 1000 µg/g.

From the viewpoint of fertilization, those nutrients that are required by plants in high quantities and that must be regularly supplied by fertilization are of particular interest. These nutrients are N, K, P, and to a minor degree also Ca, Mg, and S. Calcium is a *soil nutrient*, which means that it is important for an optimum soil structure. Application of micronutrients is not a common practice, but they are applied at locations where soils are low in a particular micronutrient or where soils may bind this micronutrient very strongly. This is the case for heavy metals (Fe, Mn, Cu, Zn) and B in calcareous and alkaline soils (soils with a high pH value), while Mo is strongly fixed in acid soils. Acid organic soils are known for their low available Cu content.

According to the different quantitative requirements for macronutrients and micronutrients, the former are taken up in much higher quantities than the latter. Thus a wheat stand with a yield potential of 7 t of grain per hectare requires about 100 kg K but only 100 g Cu.

From a physiological point of view, plant nutrients are grouped into four groups, as shown in Table 1. The *first group*, comprising C, H, O, N, and S, includes all major elementary constituents of organic plant matter. Their carriers are present mainly in the oxidized form, and they must be reduced during the process of incorporation. The energy required for this reduction originates directly or indirectly from photosynthetically trapped energy. Assimilation of H is basically an oxidation process, namely, the oxidation of water with the help of light energy (photolysis):

Table 1. Physiological classification of plant nutritive elements, nutrient carriers, and form in which the nutrient is taken up

Nutritive element	Nutrient carrier	Uptake
First group		
C	CO_2, HCO_3^-	CO_2 by leaves, HCO_3^- by roots
H	H_2O	H_2O by leaves, H_2O and HCO_3^- by roots
O	CO_2, HCO_3^-, O_2	O_2 and CO_2 by leaves, HCO_3^- and O_2 by roots
N	NH_4^+, NH_3, NO_3^-, NO_x	NH_4^+ and NO_3^- by roots, NH_3 and NO_x by leaves
S	SO_4^{2-}, SO_2, SO_3, H_2S	SO_4^{2-} by roots, SO_2, SO_3, and H_2S by leaves
Second group		
P	$H_2PO_4^-$, HPO_4^{2-}	$H_2PO_4^-$ and HPO_4^{2-} by roots
B	H_3BO_3, borates	H_3BO_3 and $B(OH)_4^-$ by roots
Si	silicates	$Si(OH)_4$ by roots
Third group		
K	K^+, K salts	K^+ by roots
Mg	Mg^{2+}, Mg salts	Mg^{2+} by roots

Ca	Ca^{2+}, Ca salts	Ca^{2+} by roots
Mn	Mn^{2+}, Mn salts	Mn^{2+} by roots
Fourth group		
Fe, Cu, Zn, Mo	ionic form or metal chelates, minerals containing these elements	by roots in ionic form or in the form of soluble metal chelates, Mo in the form of the molybdate

$$H_2O \xrightarrow{h\nu} 2H^+ + 2e^- + 0.5\,O_2$$

The *second group* (P, B, Si) comprises elements that are taken up as oxo complexes in the partially deprotonated (P) and protonated (B, Si) form. The oxo complex is not reduced in the plant cell, but may form esters with hydroxyl groups of carbohydrates, thus producing phosphate, borate, and silicate esters.

The *third group* comprises metals that are taken up from the soil solution in ionic form. They are only partially incorporated into the organic structure of the plant tissue: Mg in the chlorophyll molecule, Mn in the electron donor complex of photosystem II, and Ca^{2+} as countercation of indiffusible anions in cell walls and particularly in biological membranes. Potassium is virtually not incorporated into the organic plant matter. It is only weakly adsorbed by Coulombic forces. There exist, however, some organic molecules that may bind K^+ very selectively (ionophores, see Section Nutrient Uptake and Long-Distance Transport in Plants). These ionophores are likely to be involved in K^+ uptake.

The *fourth group* comprises heavy metals, of which Fe, Cu, and Zn are taken up as ions or in the form of soluble metal chelates, while Mo is taken up as molybdate. These molecules are easily incorporated into the organic structure, where they serve as essential elements of enzyme systems: Fe in the heme group and in ferredoxin, Mn in arginase [2], Cu in oxidases (polyphenol oxidase, cytochrome oxidase, ascorbate oxidase [3]), Zn in RNA polymerase [4], and Mo in nitrate reductase [5] and nitrogenase [6].

All nutritive elements of plants, therefore, are taken up in the form of inorganic complexes, mostly in oxidized form or as metal ions, i.e., in forms characterized by a low energy level. This is a unique feature of plants, and a feature in which they contrast sharply with animals and most kinds of microorganisms (bacteria and fungi). Animals and most microorganisms must take up food that is rich in chemical energy in order to meet their energy requirements. Plants, at least green plants, meet their energy requirement by converting radiation energy into chemical energy. This energy conversion process is manifest in the reduction of plant nutrient carriers (NO_3^-, SO_4^{2-}, CO_2) as already mentioned. Thus important processes of plant nutrition are closely linked with the unique function of plants in the great cycle of nature, i.e., the conversion of inorganic matter into organic form. Liebig [7] was correct in commenting on plant nutrition: “Die ersten Quellen der Nahrung liefert ausschließlich die anorganische Natur.” (The primary source of nutrition is provided exclusively by the inorganic materials in nature.)

2.1.2. *Function of Plant Nutrients*

Most plant organs and particularly plant parts that are metabolically very active, such as young leaves and roots, are rich in water (ca. 80 – 90 wt % of the total fresh matter), while their organic material is ca. 12 – 18 wt % and their mineral content is 2 – 6 wt %. As shown in Table 2, in the dry matter of plant material O and C are by far the most abundant elements, followed by H, N, and K. The elements C, O, H, and, to some extent, N are mainly structural elements in plant matter. They can, however, form chemical groups that are directly involved in metabolic processes, e.g., carboxyl groups, amino groups, hydroxyl groups.

Table 2. Mean content of chemical elements in the dry matter of green plant material

Element	Content, g/kg
O	440
C	420
H	60
N	30
K	20
P	4
All other elements	26

Since in many soils the available N is low, *nitrogen* [7727-37-9] is the most important fertilizer element, and for this reason its function in plant metabolism deserves particular interest. Nitrogen is an essential element for amino acids, proteins, nucleic acids, many coenzymes, and some phytohormones. Basic biochemical processes of meristematic growth, such as the synthesis of proteins and nucleic acids, require N. If this nutrient is not supplied in sufficient amounts, the growth rate is depressed and the synthesis of proteins affected. Nitrogen-deficient plants are characterized by low protein and high carbohydrate contents. This relationship is shown in Table 3 [8].

Table 3. Effects of N supply on yield of dry matter and the content of organic N and carbohydrates in the dry matter of young timothy plants (*Phleum pratense*) [8]

Yield and content	N supply	
	Low	Sufficient
Yield, g/pot	15.7	20.2
Content, mg/g		
Organic N,	20.5	31.5
Sucrose	46.9	22.6
Fructans *	22.2	9.2
Starch	32.8	11.7
Cellulose	169	184

* Polysaccharides of fructose.

Nitrogen is also essential for the formation of chloroplasts, especially for the synthesis of chloroplast proteins. Hence N deficiency is characterized by low chlorophyll content; the leaves, especially the older ones, are pale and yellow; the stems thin and the plants small. Nitrogen-deficient plants senescence earlier, probably because of a deficiency of the phytohormone cytokinin. Abundant N supply increases the protein content, especially the content of free amino

acids, and often also the content of NO_3^- in plants. An example of this is shown in Table 4 [9]. Excess nitrogen nutrition results in luxurious plants that frequently are susceptible to fungi attack.

Table 4. Relationship between N fertilizer rate and nitrogenous fractions in the dry matter of rye grass [9]

Nitrogen fertilizer rate, kg/ha	Nitrogenous fraction, g/kg			
	Total N	Protein N	Free amino acid N	NO_3^- and NO_2^- N
0	13.2	9.8	1.6	0.4
110	18.9	12.6	2.1	0.6
440	37.3	20.6	5.6	3.5

The ratio of N to S in plant matter is ca. 10 : 1. Hence *sulfur* [7704-34-9] is required in much lower quantities than N. Their functions are, however, similar. Sulfur is an elementary constituent of most proteins; the SH group in involved in various enzymatic processes and it is the reactive group of coenzyme A. Disulfide (S – S) bridges are essential structural elements in the tertiary structure of polypeptides and in many volatile S compounds, such as diallyl disulfide, which is the main component in garlic oil. Mustard oils occurring in many species of the Cruciferae contain a S-glycosidic bond and a sulfuryl group:

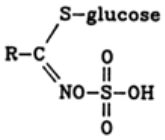

Insufficient S supply results in a decrease of growth rate with extremely low levels of SO_4^{2-} and high concentrations of free amino compounds and NO_3^- in the leaves, which is due to hampered protein synthesis. Sulfur plays an important role in the baking quality of wheat, since the concentration of S compounds in the gluten fraction is responsible for the linkages between the protein molecules [10]. Sulfur deficiency may also affect N_2 fixation of legumes by causing unfavorable conditions in the host plant or because of the relatively high S content of nitrogenase and ferredoxin [11]. Deficiency symptoms of S appear at first in the youngest leaves, which turn light green to yellow. Abundant supply with S results in an accumulation of sulfate in plant tissues.

Sulfur oxide can be taken up by the leaves and metabolized and thus can contribute to the S nutrition of plants. Too high SO_2 concentrations in the atmosphere may be toxic. The toxicity symptoms are necrotic spots in the leaves. According to SAALBACH [12], the critical SO_2 level in the atmosphere for annual plants is 120 µg/m^3. For trees and other perennials it is about half this level. The currently much discussed damage to trees in the forest of the Federal Republic of Germany (mainly spruce and silver fir) is not caused by toxic SO_2 levels.

Phosphorus [7723-14-0] is an essential element in nucleic acids and various phospholipids (phosphoglyceride and phosphosphingolipids). In both cases. phosphate is esterified with sugars (nucleic acid) or with alcohol groups of glycerol or sphingosine. Phosphate is also present in various coenzymes; the most prominent is adenosine triphosphate (ATP), which carries a kind of universal energy that is used in a number of biochemical processes. Metabolites and enzymes can be activated by phosphorylation, a transfer of the phosphoryl group from ATP to the metabolite according to the following reactions:

Activation of glucose

Glucose + ATP ⇒ Glucose-6-phosphate + ADP

Phosphorylation of an enzyme

Enzyme – OH + ATP ⇒ Enzyme–O(P) + ADP

Such reactions demonstrate the essential role of P not only in plant metabolism but also in all living organisms. Undersupply with P results in a reduced growth rate, and seed and fruit formation is affected. The leaves of P-deficient plants often show a gray dark green color; the stems may turn red. The P reserve in seeds is the Mg (Ca) salt of the inositol hexaphosphate (phytic acid):

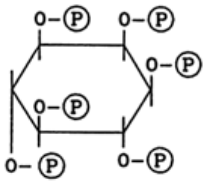

Myo–inositol hexaphosphate

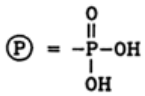

Phosphoryl group

The physiological role of boron has remained obscure until now, and therefore various hypotheses with numerous modifications exist concerning the physiological and biochemical role of boron in higher plants. Depending on the pH of the soil, boron seems to be taken up mainly as undissociated boric acid or as the borate anion. Plant species differ in their boron uptake capacity, reflecting differences in boron requirements for growth. However, there is still some controversy about boron translocation in plants. At least in higher plants, a substantial proportion of the total boron content is complexed in the cell walls in in a *cis*-diol configuration [14]. According to BIRNBAUM et al. [13], B is involved in the synthesis of uracil and thus affects UTP formation. (UTP is an essential coenzyme for the synthesis of sucrose and cell-wall components.) Also the synthesis of ribonucleic acid is hampered in the case of B deficiency. Since uracil is an integral part of ribonucleic acid (RNA), the formation of RNA may also be related to the synthesis of uracil. POLLARD et al. [15] suggest that B has a specific influence on plant membranes by the reaction of borate with polyhydroxy compounds.

Boron deficiency appears as abnormal or retarded growth of the apical growing points. The youngest leaves are misshapen and wrinkled and show a darkish blue-green color. The fact that B deficiency primarily affects the apex is in accord with the impaired synthesis of ribonucleic acids required for meristematic growth. High levels of available B in the soil may cause B toxicity in plants. This is mainly the case in arid areas; however, B toxicity can also be the consequence of industrial pollution [16]. The toxicity is characterized by yellow leaf tips followed by progressive necrosis. The leaves take on a scorched appearance and drop prematurely.

Silicon [7440-21-3] is not an essential element for plants; however, it has a beneficial effect on various plant species, mainly grasses [17]. In plants well supplied with Si, cuticular water losses are diminshed and resistance against fungal attack is improved [18]. The favorable effect of Si on rice growth is well known. Silicon-containing fertilizer is frequently applied in rice production.

Among the metal cation species, *the potassium* [7440-09-7] *ion*, K^+, is the nutrient plants take up from the nutrient medium at the highest rates. The K^+ concentration in the cytoplasm is about 100 mM and thus much higher than the concentration of other ion species [19]. Probably this high K^+ concentration has a favorable influence on the conformation of various enzyme proteins [20]. Potassium ions can easily penetrate plant membranes (see Section Nutrient Uptake and Long-Distance Transport in Plants), which often leads to a depolarization of the membranes. Membrane depolarization, it is supposed, has a favorable effect on meristematic growth, photophosphorylation, aerobic phosphorylation, and phloem loading [21]. These basic processes are important for the long-distance transport of photosynthates, the synthesis of various organic compounds, and CO_2 assimilation.

The data in Table 5 show that with an increase of K^+ in alfalfa leaves (*Medicago sativa*), the CO_2 assimilation rate increased, while the mitochondrial respiration rate decreased [22]. In the case of low K, the respiration was about 2/3 of the CO_2 assimilation, while with high K the C gained by assimilation was about 11 times higher than the C lost by respiration. This typical behavior indicates that under the conditions of K^+ deficiency much of the stored carbohydrates must be respired in order to meet the ATP demand of the plant. Plants undersupplied with K^+ have therefore a low energy status. Such plants are highly susceptible to fungal attack, water stress, and frost damage.

Table 5. Relationship between K^+ concentration in the dry matter of alfalfa leaves, CO_2 assimilation, and mitochondrial respiration [22]

Concentration of K^+, mg/g	Carbon gain and loss, mg dm^{-2} h^{-1}	
	CO_2 assimilation	Mitochondrial respiration
13	11.9	7.56
20	21.7	3.34
38	34.0	3.06

Potassium is important in determining the osmotic pressure of plant fluids, and K^+-deficient plants are characterized by inefficient water use. Sodium ions may replace some K^+ functions, e.g., the less specific osmotic functions. Important counterions of K^+ in plant tissues are Cl^-, NO_3^-, and organic anions. The frequently observed favorable effect of Na^+ and Cl^- on plant growth is related to their osmotic functions.

Plants suffering from K^+ deficiency show a decrease in turgor, and under water stress they easily become flaccid. Plant growth is affected, and the older leaves show deficiency symptoms as necrosis beginning at the margins of tips and leaves. In K^+-deficient plant tissue, toxic amines such as putrescine and agmatine accumulate.

The most spectacular function of *magnesium* [7439-95-4] is its integral part in the chlorophyll molecule. Besides this function, Mg^{2+} is required in various other processes and, the Mg fixed in the chlorophyll molecule amounts only to about 20 % of the Mg present in green plant tissues. Magnesium is an essential ion in ribosomes and in the matrix of the cell nucleus. Here Mg^{2+} is bound by phosphate groups, since the Mg^{2+} is strongly electrophilic and thus attracts oxo complexes such as phosphate [23]. The magnesium ion activates numerous enzymatic reactions in which phosphate groups are involved. The activation is assumed to be brought about by bridging the phosphate group with the enzyme or with the substrate. This is an universal function of Mg^{2+} not only relevant for plant metabolism but also for practically all kinds of organisms.

Deficiency of Mg^{2+} affects chlorophyll synthesis: leaves turn yellow or red between the veins. The symptoms begin in the older leaves. Protein synthesis and CO_2 assimilation are depressed under Mg^{2+} deficiency conditions. Recent results [24] have shown that the yellowing of spruce needles in the Black Forest is due to a Mg^{2+} deficiency and can be cured by Mg^{2+} fertilizer application.

Calcium [7440-70-2] is the element of the apoplast (cell wall and “free space”) and of biological membranes. Here it is adsorbed at the phosphate head groups of membrane lipids, thus stabilizing the membranes [25]. Most of the Ca^{2+} present in plant tissues is located in the apoplast and in the vacuole, some in the mitochondria and in the chloroplasts, while the cytoplasm is extremely low in Ca^{2+} (10^{-7} to 10^{-6}M). The maintenance of this low cytoplasmic Ca^{2+} concentration is of vital importance for the plant cell [26]. Higher cytoplasmic Ca^{2+} concentrations interfere with numerous enzymatic reactions and may even lead to a precipitation of inorganic phosphates. This low Ca^{2+} concentration suffices to form a complex with calmodulin, a polypeptide of 148 amino acids. The Ca – calmodulin complex is a universal enzyme activator. The activation is brought about by allosteric induction.

Direct Ca^{2+} deficiency in plants is rare, since most soils are relatively rich in Ca^{2+}. Physiological disorders as a consequence of an insufficient Ca^{2+} supply of particular plant parts, however, occur frequently. Calcium is mainly translocated by the transpiration stream. Hence plant parts such as fruits, which mainly feed from the phloem and less from the xylem sap, may suffer from an insufficient Ca^{2+} supply. Shear [27] cites a list of 35 such Ca^{2+}-related disorders in fruits and vegetables. Two of the most important ones involve storage tissues and result in poor crop qualities [28]: *bitter pit* in apples, characterized by small brown spots on the surface, and *blossom-end rot* in tomatoes, a cellular breakdown at the distal end of the fruit, which is then susceptible to fungal attack.

Manganese [7439-96-5] is an integral part of the superoxide dismutase and of the electron donor complex of photosystem II. Manganese may activate enzymes in the same way as Mg^{2+} by bridging the phosphate group with the enzyme or the substrate. Deficiency of Mn^{2+} leads to the breakdown of chloroplasts. Characteristic deficiency symptoms are smaller yellow spots on the leaves and interveinal chlorosis. Manganese toxicity may occur, especially on flooded soils, because of the reduction and thus solubilization of manganese oxides. Toxicity symptoms are generally characterized by brown spots of MnO_2 in the older leaves surrounded by chlorotic areas [29].

Iron [7439-89-6] is an essential element for haem and ferredoxin groups. Iron deficiency leads to chloroplast disorders; the synthesis of thylakoid membranes is disturbed and the photochemical activity affected [30]. Iron deficiency is characterized by yellow leaves. The symptoms are at first visible in the younger leaves. There is evidence that the deficiency, mainly occurring in plants growing on calcareous soils, is not induced by an insufficient Fe uptake from the soil but by a physiological disorder in leaves, affecting the Fe distribution in the leaf tissue [31].

Iron toxicity can be a problem under reducing soil conditions, which prevail in flooded soils. Under such conditions iron(III) oxides are reduced and the iron is rendered soluble. This may increase the Fe concentration in the soil solution by a factor of 10^2 to 10^3 [32] so that plants may suffer from Fe toxicity, characterized by tiny brown spots on the leaves, which later may turn uniformally brown. Iron toxicity is known as “bronzing.”

Copper [7440-50-8] is an essential element of various enzymes, such as superoxide dismutase, polyphenol oxidases, plastocyanin of the photosynthetic transport chain, and cytochrome c oxidase, the terminal oxidase in the mitochondrial electron transport chain. Deficiency in Cu leads to pollen sterility and thus affects the fruiting of plants. Copper-deficient plants often are characterized by white twisted leaf tips and a tendency to become bushy.

Zinc [7440-66-6] is an integral part of carbonic anhydrase, superoxide dismutase, RNA polymerase, and various dehydrogenases. It is closely involved in the N metabolism of plants. In Zn-deficient plants, protein synthesis is hampered and free amino acids accumulate. There is evidence that Zn is involved in the

synthesis of tryptophan, which is a precursor of indole acetic acid, an important phytohormone. Zinc deficiency is characterized by short internodes, small leaves, and chlorotic areas in the older leaves. Frequently the shoots die off and the leaves fall prematurely.

Molybdenum [7439-98-7] is present in the nitrate reductase and in the nitrogenase system that catalyzes the bacterial fixation (reduction) of dinitrogen. Deficiency of Mo frequently appears first in the middle and older leaves as a yellowish green coloration accompanied by a rolling of leaf margins. Cruciferae species are particularly susceptible to Mo deficiency. The most well-known Mo deficiency is the “whiptail” of cauliflower. For further information on the physiology of plant nutrition, see [3], [23], [33].

2.2. Soil Science

2.2.1. *Soil Classes, Soil Types, and Parent Material*

According to SCHROEDER [34], “soil is the transformation product of mineral and organic substances on the earth's surface under the influence of environmental factors operating over a very long time and having defined organisation and morphology. It is the growing medium for higher plants and basis of life for animals and mankind. As a space-time system, soil is four dimensional.”

Soils are complex, quite heterogeneous, and may differ from each other considerably. Nevertheless, all soils have some common features. They possess a mineral, an organic, a liquid, and a gaseous component. In an ideal soil, the percentage proportions of these components are 45 %, 7 %, 23 %, and 25 %, respectively. The volumes of the liquid and gaseous components may change quickly. For example, in a water-saturated soil all pores are filled with water, and in a dry soil the soil pore volume is almost completely filled with air. The mineral and organic components contain plant nutrients and adsorb plant nutrients at their surfaces, and they are therefore of importance for the storage and retention of plant nutrients. The liquid phase of the soil is the soil solution. It contains dissolved plant nutrients and is the medium for the translocation of plant nutrients from various soil sites towards the plant roots. The gaseous soil component is essential for gas exchange, especially for the supply of plant roots with oxygen and for the release of CO_2 from the soil medium into the atmosphere.

For the description, comparison, and assessment of soils, a grouping according to general criteria is indispensable. There are two main grouping systems for soils: (1) soil classes or soil texture and (2) soil types. *Textural classes* are defined according to the particle size of soils. *Soil types* relate to the parent material of soils, to the pedological genesis, and to typical properties evident in the soil profile i.e., the horizontal layers of soils, called soil horizons.

Soil Classes. *Soil particle sizes* as a main characteristic of soil classes are grouped into four major groups as shown in Table 6. The major groups (sand, silt, and clay) are subdivided into coarse, medium, fine. Designation of the soil texture (soil class) depends on the percentage proportions of the sand, silt, and clay fraction in the total fine earth, which is sand + silt + clay. Soils in which the sand fraction dominates are termed *sandy soils*, soils consisting mainly of silt and clay are *silty clays*, and soils which contain all three fractions in more or less equal amounts are called *loams*. In the German terminology, abbreviations for the fractions are used (S = sand, U = silt, T = clay, L = loam). For example, if the major fraction is silt (U) and the next sand (S), the abbreviation is sU = sandy silt. Figure 1 shows the designations of the various soil classes according to the percentage proportion of the three main particle fractions.

Table 6. Particle size of soil fractions relating to soil texture

Diameter, mm	Designation	Abbrevation
> 2	pebbles, gravels	
0.06 – 2	sand	S
0.002 – 0.06	silt	U
< 0.002	clay	T

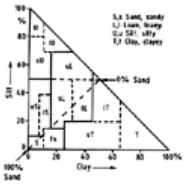

Figure 1. Diagram of soil textural classes, German system of SCHROEDER [34]

The vertical axis shows the percentage of silt, the horizontal the percentage of clay, and the dashed line the percentage of sand.

In the farmer's practice, sandy soils are called light soils, soils rich in clay heavy soils. This distinction relates to the force required to work (plough, cultivate) a soil. Soils rich in clay, but also silty soils, tend to compaction when dried and hence are heavy to work.

Although the grouping according to particle size is based on a physical factor, particle size is also associated with the chemical properties. This can be seen from Figure 2: the sand fraction consists mainly of quartz, which is a sterile material. Primary silicates (micas, feldspars) contain K, Ca, Mg, and other plant nutrients, which are released during the process of weathering. Clay minerals are less rich in plant nutrients than the primary silicates, but they possess large negatively charged surfaces that are of the utmost importance for the adsorption of plant nutrients and water.

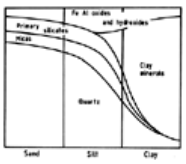

Figure 2. Mineral composition of the sand, silt, and clay fractions [34]

The various soil particles form aggregates in which organic matter is also involved. This aggregation forming fine pores and holes in the soil is of relevance for soil structure. A good soil structure is characterized by a relatively high pore volume, ca. 50 % of the total soil volume. Soil structure depends much on the Ca saturation (see Cation exchange capacity). The richer the soil is in clay, the more important a good soil structure is.

Soil Type. Soil type is related to the parent material from which a soil is developed and from the history of development, which is much influenced by climate and vegetation. Main groups of parent material are igneous rocks, sedimentary rocks, and metamorphic rocks. Also organic matter may be the main parent material. Content of plant nutrients, capacity to store plant nutrients, soil pH, and the rooting depth depend much on the parent material, but are also influenced by soil development.

In the following, a limited number of important soil types are considered according to the FAO World Soil Classification. Besides this system there are other systems, e.g., the U.S. Soil Taxonomy. The FAO classification comprises 26 classes.

A distinction can be made between young soils and old soils. The latter are generally highly weathered, their inorganic material consisting mainly of quartz and iron aluminum oxide hydroxides. Such soils are characterized by poor cation retention capacity (cation exchange capacity), low pH values, and a high phosphate fixing power. This soil type, called ferralsol, is frequent in the tropics, whereas in moderate climates highly weathered soils belong mainly to the podsols.

Young soils may be derived from the sedimentation of rivers and oceans (fluvisols) or from volcanic ash (andosols). These soils are generally rich in plant

nutrients and thus form fertile soils. The most fertile soils belong to the black earths (chernozem). They are frequent in Russia, Central and East Europe as well as North America and are derived from loess. They are characterized by a neutral pH, by a well balanced content of clay and organic matter and by a deep rooting profile. They are naturally rich in plant nutrients and possess a high nutrient storage capacity. Soils in which the parent material loess is more weathered as compared with the chernozems belong to the luvisols. This soil type is common in Germany, Austria, and France where it represents the most fertile arable land. Gleysols are soils with a high water table, rendzinas are shallow soils derived from limestone, histosols are rich in only partially decomposed organic matter.

Under arid conditions salt may accumulate in the top soil layer. Solonchaks (white alkali soils) are saline soils with a pH of ca. 8 and with neutral anions as the most important anion component. Solonetz soils (black alkali soils) possess bicarbonate and carbonate as major anion component. Their pH is in the range 8 – 10. Crop growth on saline soils is extremely poor, and in many cases only a salt flora can grow under such conditions. This is particularly true for the solonetz soils.

For further information on soil texture and soil types, see [34-37], and the Soil Taxonomy of the Soil Conservation Service of U.S. Department of Agriculture [38].

2.2.2. *Nutrient Retention in Soils*

Nutrient retention is an important characteristic of fertile soil.

Cation Exchange. Cations are retained on soil colloids having a negative charge: the cations are bound at the surface of these particles by Coulomb forces. The most important cation species are Ca^{2+}, Mg^{2+}, K^+, Na^+, Al^{3+}, $Al(OH)^{2+}$, $Al(OH)_2^+$, and H^+. This is represented in Figure 3. A distinction can be made between inorganic and organic soil colloids capable of cation adsorption. Inorganic particles belonging to the clay fraction are known as *secondary clay minerals* because they are mainly derived by weathering of primary minerals such as orthoclase, plagioclase, and particularly mica. Organic soil colloids capable of cation adsorption belong to the *humic acids*. The negative charge of the inorganic soil colloids originates from the so-called isomorphic substitution and from deprotonation. Isomorphic substitution is the replacement of Si^{4+} in the crystal lattice by Al^{3+}, Fe^{2+}, or Mg^{2+}, thus leading to a surplus of negative charge, because the anionic groups of the lattice are not completely balanced by Al^{3+}, Fe^{2+}, or Mg^{2+}. Such a negative charge is a permanent charge, in contrast to labile charges that result from deprotonation. Labile charges are typical for organic colloids (humic acids): carboxylic groups and acid hydroxylic groups of phenols may be protonated or not depending on the pH of the environment.

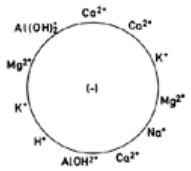

Figure 3. Schematic presentation of cations adsorbed to the negatively charged surface of a soil colloid

The secondary clay minerals are grouped into 1 : 1 clay minerals, in which a Si layer alternates with an Al layer, and 2 : 1 clay minerals, in which an Al layer is sandwiched by two Si layers. The most important representatives of the 1 : 1 clay minerals are the kaolinites. The 2 : 1 secondary clay minerals comprise the illite, transitional minerals, vermiculite, chlorite, and smectites (⟶ Clays). Most of these 2 : 1 clay minerals possess inner surfaces. They are therefore characterized by a high cation retention (= cation exchange) capacity.

These negatively charged soil colloids, often also called sorption complexes, function like a *cation exchanger*. Adsorbed cations can be replaced by other cation species. The cation exchange is stoichiometric. Adsorption and desorption depend on the concentrations of the cation species in the surrounding solution. If a soil colloid completely saturated by K^+ is exposed to increasing Ca^{2+} concentrations, for example, adsorbed K^+ is more and more replaced by Ca^{2+} until eventually the sorption complex is completely saturated by Ca^{2+} (Fig. 4). In soils, such exchange and equilibrium reactions are complex as numerous cation species and sorption complexes with differing preferences for particular cation species are involved. The principle, however, is that cations adsorbed by Coulomb forces at soil colloids equilibrate with free cations in the soil solution. Thus adding cations to a soil by fertilization, e.g., the application of a potassium salt, results in replacing adsorbed cations with the newly added cations until a new equilibrium is reached. The adsorbed cations are protected against leaching. but they are available to plant roots. The strength of cation adsorption increases with the the charge of the cation species and with the thinness of the hydration shell. Provided that there are no specific adsorption sites, the strength of cation adsorption follows Hofmeister's cation sequence:

$$Ca^{2+} > Mg^{2+} > K^+ > Na^+$$

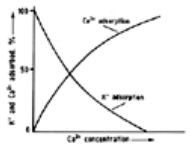

Figure 4. Ca^{2+} – K^+ exchange, K^+ desorption brought about by increasing Ca^{2+} concentration

At equilibrium, cation-exchange reactions are a helpful tool for predicting the distribution of ions between the adsorbed and solution phases of the soil as the amounts of cations present are changed. When a soil saturated with potassium is placed in a NaCl solution, the following equilibration occurs:

$$K_{soil} + NaCl \rightleftharpoons Na_{soil} + KCl$$

The exchange equation for this reaction is

$$\frac{[Na](K)}{[K](Na)} = k_1$$

Brackets refer to ions on the exchange site and parenthesis to the activity of ions in the solution. Since the proportionate strength of adsorption of the ions varies with the exchange site, values for k_1 differ for different exchange materials.

The divalent/monovalent system, which almost represents the situation in the soil, with K^+, Ca^{2+}, and Mg^{2+} as the dominant exchangeable cations, is more complex. The following equation, developed by GAPON [39], is widely used to describe monovalent/divalent exchange:

$$\frac{[K](Ca)^{1/2}}{[Ca](K)} = k_1$$

Cation exchange capacity (CEC) is defined as the quantity of cation equivalents adsorbed per unit soil or clay mineral. In Table 7 the exchange capacities of some soil classes are shown. The exchange capacity of the organic soil appears high if it is based on unit weight of soil. A more realistic picture is obtained, however, when the exchange capacity is based on soil volume, since under field conditions it is soil volume, not soil weight, that is related to a crop stand. Table 8 shows the cation exchange capacities of some important clay minerals and of humic acids in relation to the surface of these particles.

Table 7. Cation exchange capacity based on soil weight and soil volume as well as the specific weight of some soil classes

Soil class	Specific weight, kg/L	Cation exchange capacity *

		cmol/kg	cmol/L
Sandy soil	1.5	3	4.5
Loam	1.5	15	22.5
Clay soil	1.5	30	45.0
Organic soil	0.3	75	22.5

* cmol = centimole.

Table 8. Cation exchange capacity and inner and outer surfaces of some soil colloids

	Total surface, m^2/g	Inner surface, %	Cation exchange capacity, mol/kg
Kaolinite	20	0	10
Illite	100	0	30
Smectite	800	90	100
Humic acids	800	0	200

The cation exchange capacity of kaolinites and particularly of humic acids depends much on the pH of the medium. At low pH, most groups are protonated and hence the exchange capacity is low. Increasing soil pH, e.g., by liming, increases the cation exchange capacity if kaolinites and humic acids are the dominating exchange complexes.

Cation saturation of negatively charged soil colloids has some impact on soil structure, which is defined as the arrangement of soil particles into aggregates. High percentage of adsorbed Ca^{2+} favors the formation of aggregates. In well structured soils, such as in chernozems, 70 to 80 % of the total cation exchange capacity is occupied by Ca^{2+}. In acid solids, H^+ and Al cations (Al^{3+}, $Al(OH)^{2+}$, $Al(OH)_2^+$) and in saline soils Na^+ and Mg^{2+} are the dominating cation species adsorbed to soil colloids.

Anion Exchange. Soil particles may also adsorb anions. The adsorption occurs at the OH groups of aluminum and iron oxides as well as of some clay minerals. One may distinguish between a nonspecific adsorption and a specific anion adsorption. The nonspecific anion (A^-) adsorption originates from protonated hydroxylic groups.

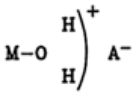

Protonation depends on soil pH and is particularly high under acid conditions. Hence nonspecific anion adsorption only plays a role in acid soils.

The specific anion adsorption is a ligand exchange. This is, for example, the case for phosphate. In step 1 $H_2PO_4^-$ replaces OH^-, resulting in a *mononuclear bond* between the phosphate and the iron oxide. In step 2, a further deprotonation of the phosphate occurs, followed by a second ligand exchange (step 3) to form a *binuclear bond* between the surface of the iron oxide and the phosphate.

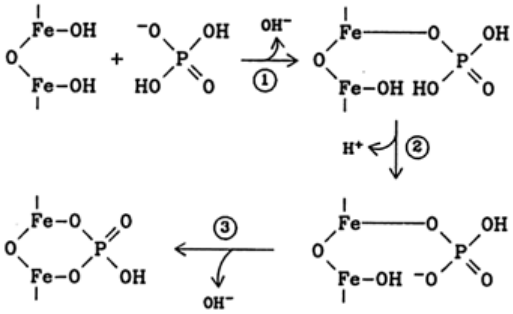

The final structure is supposed to be very stable, and the phosphate so bound is hardly available to plant roots. This reaction sequence explains why anion (phosphate) adsorption is promoted under low pH conditions. In mineral soils with pH < 7, the adsorbed phosphate represents a major phosphate fraction. Increasing the soil pH, e.g., by liming, increases phosphate availability [40]. The relationship between free and adsorbed anions can be approximately described by the Langmuir equation:

$$A = A_{max}\frac{kc}{1+kc}$$

A = surface concentration of adsorbed anions
A_{max} = maximum surface concentration
c = concentration of free anion
k = constant related to adsorption energy, the adsorption strength increasing with k

Adsorption strength depends also on anion species decreasing in the order [41]:

$$\text{phosphate} > \text{arsenate} > \text{selenite} = \text{molybdate} > \text{sulfate} = \text{fluoride} > \text{chloride} > \text{nitrate}$$

Borate and silicate may also be adsorbed, but only at high pH. Under these conditions, boric acid and silicic acid may form anions according to the following equations:

$$H_3BO_3 + H_2O \rightarrow B(OH)_4^- + H^+$$

$$H_2SiO_3 + H_2O \rightarrow H_3SiO_4^- + H^+$$

This is why in neutral to alkaline soils boron can be strongly adsorbed (fixed) by soil particles, which may lead to boron deficiency in plants. The formation of a

silicate anion can improve phosphate availability since $H_3SiO_4^-$ and phosphates compete for the same ligands at anion-adsorbing surfaces.

2.2.3. *Soil pH, Buffer Power, and Liming*

Proton concentration (pH) is of vital importance for all living organisms and also has an impact on soils and soil constituents. High H^+ concentrations (pH<4) attack soil minerals, dissolving metal cations out of the crystal lattice, and eventually lead to mineral degradation. Under low soil pH conditions (pH<5), bacterial life is suppressed while fungal life is relatively favored, which affects the decomposition of organic matter. Low soil pH (pH < 4) also affects root growth.

In many cases, however, it is not so much the H^+ but the toxic level of soluble Al species and Mn^{2+} associated with low pH conditions that considerably hamper root development and plant growth. A decrease in soil pH increases the solubility of aluminum oxide hydroxides and maganese oxides considerably. Under acid soil conditions Al^{3+}, $Al(OH)^{2+}$, $Al(OH)_2^+$ are dissolved in the soil solution or adsorbed to soil colloids. Of the three Al species, $Al(OH)^{2+}$ is considered to be the most toxic. *Aluminum toxicity* of plants is characterized by poor root growth: root tips and lateral roots become thickened and turn brown [42]. In the cell, Al interferes with the phosphate turnover and may even be adsorbed on the DNA double helix.

A distinction is made between the *actual acidity*, which is determined by the H^+ concentration in the soil solution, and the *potential acidity*, which is determined by both the H^+ of the soil solution and the adsorbed H^+. The actual acidity is measured by hydrogen-ion electrodes; the potential acidity, by titration with a base. Titration includes the Al species, which also consume OH^-

$$Al^{3+} + OH^- \rightarrow Al(OH)^{2+}$$

$$Al(OH)^{2+} + OH^- \rightarrow Al(OH)_2^+$$

$$Al(OH)_2^+ + OH^- \rightarrow Al(OH)_3 \text{ (gibbsite)}$$

Potential acidity may be high in Al-rich soils as well as in soils with a high cation exchange capacity. Such soils are characterized by a high *hydrogen-ion buffer power*: a pH change in the soil solution requires a relatively large amount of H^+ or OH^-.

In soils, different buffer power systems can be distinguished. Soils containing carbonates are buffered according to the following equation:

$$CaCO_3 + 2\,H^+ \rightarrow Ca^{2+} + H_2O + CO_2$$

Reaction may be promoted by the presence of CO_2 originating from root and microbial respiration.

$$CaCO_3 + CO_2 + H_2O \rightarrow Ca(HCO_3)_2$$

$$Ca(HCO_3)_2 + 2\,H^+ \rightarrow Ca^{2+} + 2\,H_2O + 2\,CO_2$$

At lower pH levels soils are buffered by adsorbed cations (Fig. 5) and by Al complexes, e.g.,

$$Al(OH)_3 + H^+ \rightarrow Al(OH_2)^+ + H_2O$$

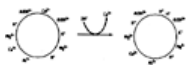

Figure 5. Principle of hydrogen-ion buffering by adsorbed cations

Soil acidification results from different processes. Plant roots extrude H^+; net release of H^+ is especially high when plants are fed with NH_4^+, while NO_3^- nutrition results in a net release of OH^-. Leguminous species living in symbiosis with *Rhizobium* extrude H^+ at particularly high rates from their roots. Microbial oxidation of organic N, organic S, and elementary S leads to the formation of strong acids, such as HNO_3 and H_2SO_4, which have a marked acidifying effect on soils. Also SO_2 and NO_x as gases as well as the acid rain formed by these oxides have an acidifying effect. In Central Europe these pollutants carry ca. 3 – 5 kmol H^+ $ha^{-1}a^{-1}$ into soils.

Under anaerobic soil conditions, e.g., after flooding, protons are consumed by the reduction of Fe^{III} and Mn^{III} or Mn^{IV} as well as by the microbial reduction of NO_3^- (denitrification):

$$2\,NO_3^- + 10\,e^- + 12\,H^+ \rightarrow N_2 + 6\,H_2O$$

Some species of *Azalea, Calluna*, *Vaccinium*, and also tea (*Camellia sinensis*) are able to grow on acid habitats. These species can mask the Al with phenols and organic acids and thus avoid Al toxicity. Rye, potatoes, oats, and lupines tolerate weakly acid soils, whereas beets (*Beta vulgaris*), barley, rape, and most leguminous crops prefer more neutral soils. Wheat takes an intermediary position with regard to soil pH. In the case of leguminous crops, it is not so much the crop itself, rather it is the *Rhizobium* bacteria living in symbiosis with the crop that are affected by low soil pH. The mulitplication of the *Rhizobium* in the soil is depressed by soil acidity.

Low pH levels can be easily overcome by *liming*, the application of alkaline materials, mainly Ca/Mg oxides, carbonates, and silicates. They react with soil acidity as follows:

$$CaO + 2\,H^+ \rightarrow Ca^{2+} + H_2O$$

$$CaCO_3 + 2\,H^+ \rightarrow H_2O + CO_2 + Ca^{2+}$$

$$CaSiO_3 + 2\,H^+ \rightarrow H_2O + SiO_2 + Ca^{2+}$$

The quantity of lime required depends on the soil pH level and the buffer power. The lower the pH and the higher the buffer power, the more lime required.

Soil acidity is mainly a problem in humid zones, where the H^+ formed in the upper soil layer replaces the adsorbed metal cations (Ca^{2+}, Mg^{2+}, K^+), which are then leached. Under arid conditions, salts may accumulate in the top soil layer. If a major part of the anions accumulated are HCO_3^- and CO_3^{2-} (solonetz soils), high soil pH levels prevail, which affect plant growth and soil structure considerably. Such soils can be meliorated by heavy applications of elementary sulfur. Under aerobic conditions the S is oxidized to H_2SO_4 by soil microorganisms. The strong acid neutralizes the HCO_3^- and CO_3^{2-}

$$H_2SO_4 + 2\,HCO_3^- \rightarrow 2\,H_2CO_3 + SO_4^{2-} \rightarrow 2\,H_2O + 2\,CO_2 + SO_4^{2-}$$

$$H_2SO_4 + CO_3^{2-} \rightarrow SO_4^{2-} + H_2CO_3 \rightarrow H_2O + CO_2 + SO_4^{2-}$$

2.2.4. *Soil Water – Plant Relationships*

Plants continuously require water that is taken up from the soil by roots and transported from them to the upper plant parts, particularly to the leaves. From here water is released into the atmosphere. This last process is called *transpiration*. Water in the plant tissues is required to maintain optimum cell turgor, which is crucial for most metabolic processes.

Plants frequently have to overcome long dry periods during which they must feed from the soil water. Very crudely, the soil can be considered as a sponge that

can store water in its pores and holes. The storage water in the soil must be retained against gravitation. The forces responsible for this retention are adsorption and capillary forces by which the water is sucked to the surface of the soil particles. This suction force can be considered as a negative pressure, and hence the strength of water binding in soils is measured in Pascals (Pa), the unit for pressure. The strength of water binding in soils is termed *water potential* (in older terminology, water tension). The higher the strength, the lower (more negative) the soil water potential. Water potentials in soils range from 0 to -1×10^6 kPa. Generally, however, soil water potentials of – 10 to – 1500 kPa prevail.

The total amount of water that can be adsorbed by a soil (all pores and holes filled with water) is called maximum water capacity. This is of minor importance; more relevant is the water quantity that can be retained against the gravitation force, the *field capacity*. Not all water of the field capacity fraction is available to plant roots. A proportion of the field capacity water is so strongly adsorbed that it can not be taken up by the roots. For most plant species this is water with a water potential < – 1500 kPa: at such a low water potential, plants wilt. Therefore, this critical water potential is also called *wilting point*. The soil water fraction not available to plants is called *dead water*. The *available soil water* thus equals the difference between the field capacity and dead water.

Soils differ much in their capacity to store water. The higher the clay content of a soil, the larger the total surface of soil particles, and the more water that can be adsorbed. The water storage capacity of soils increases with the clay content. On the other hand, the water molecules are strongly adsorbed to the surface of clays and therefore the fraction of dead water increases with the clay content. For this reason, generally medium textured soils (loamy soils), and not the clay soils, possess the highest storage capacity for available soil water. Besides soil texture, also soil structure and the rooting depth of the soil profile determine the storage capacity for available soil water.

An important criterion of available soil water is the relationship between the percent water saturation of the soil and the water potential. This is shown in Figure 6 for a sandy, a loamy, and a clay soil. The section between field capacity (– 10 kPa) and the wilting point (– 1500 kPa) is the highest for the loamy soil.

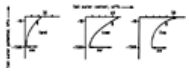

Figure 6. Relationship between the soil water content and the soil water potential and the resulting available water (AW) for three soil classes

Field capacity is –10 kPa; wilting point is – 1500 kPa.

The capability to use soil water economically differs considerably among plant species. A measure of this capability is the transpiration coefficient, the water quantity in kg (or L) required for the production of 1 kg plant dry matter:

Sorghum	277
Maize	349
Sugar beet	443
Spring wheat	491
Barley	527
Potatoes	575
Oats	583
Spring rye	634
Red clover	698
Flax	783
Alfalfa	844

Water loss under a vegetation cover results from evaporation (water release from the soil to the atmosphere) and transpiration (water release of the plant to the atmosphere). Evaporation is unproductive, transpiration productive. The relation between the two depends on plant nutrition, as can be seen in Table 9, which shows the favorable effect of N fertilizer on the productive use of soil water [43].

Table 9. Relationship between N fertilizer rate, grain yield of barley, evaporation, and transpiration [43]

Nitrogen fertilizer rate, kg/ha	Grain yield, t/ha	Transpiration,* L/m^2	Evaporation,* L/m^2
30	1.02	85	235
125	1.65	121	278
225	2.69	217	212

* Transpiration and evaporation in liters per square meter of soil surface.

2.2.5. *Organic Matter of Soils and Nitrogen Turnover*

Organic matter of soils differs considerably. Soils can be classified according to the content of organic carbon in the soil (g/kg):

Low	<5
Medium low	5 – 10
Medium	10 – 20
High	20 – 40
Rich	40 – 80
Muck	80 – 150
Peat	>150

Soils with an organic carbon content of 50 g/kg are termed organic soils, in contrast to mineral soils. Enrichment of organic matter in soils depends on location and climatic conditions. Low temperature and a lack of oxygen favor the accumulation of organic matter in soils because these conditions hamper breakdown by soil microbes. Therefore, under cold continental climate conditions (frost, long winters) and under hydromorphic soil conditions (swamps, moors, bogs), organic matter accumulates.

The fertility status of organic soils differs considerably. Moors located on the tops of mountains are generally poor in nutrients, especially in N and K, since they are fed mainly from rain. Moors in lowland fed from rivers and streams may be rich in plant nutrients, especially in N. The C : N ratio of the organic matter and soil pH are suitable indicators of the fertility status. Fertile soils possess C : N ratios of ca. 20 in their organic matter and pH values in the weak acid to weak alkaline range. Acid organic soils (highland moors) have much higher C : N ratios, ca. 50 or more in their organic matter. The C : N ratio has direct impact on the decomposition of organic N by soil microbes; the higher the ratio the lower the net release of mineral N by microbial activity.

Nitrogen turnover in soils is related not only to biological processes but also to physicochemical processes. In addition, there is a rapid exchange of N between the biosphere, the soil, and the atmosphere. The main processes of this nitrogen cycle are shown in Figure 7. Inorganic nitrogen, mainly NO_3^- and NH_4^+, including fertilizer N, can be easily assimilated by higher plants as well as by microbes (fungi, bacteria). Also dinitrogen (the N_2 of the atmosphere) can be reduced to NH_3 by some soil bacteria. The N_2 fixation capacity of the so-called free living bacteria, mainly species of *Azotobacter, Beijerinckia, Azospirillum,* and some species of the Cyanobacteria (*Anabaena, Nostoc, Rivularia*) is moderate, amounting to ca. 5 – 50 kg $ha^{-1}a^{-1}$. Symbiontic N_2-fixing bacteria, mainly species of *Rhizobium* and *Actinomyces*, have a fixation capacity about 10 times higher: for pulses (grain legumes) 50 – 100 kg/ha per growth period and for forage legumes even 200 – 500 kg $ha^{-1}a^{-1}$. They are of utmost importance in the N turnover and N availability in soils.

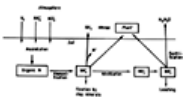

Figure 7. N cycle in nature. Transfer of N between soil, plant, and atmosphere

Inorganic nitrogen (N_2, NH_4^+, NO_3^-) assimilated by living organisms is mainly used for the synthesis of proteins, amino sugars, and nucleic acids. As soon as these organisms die, the organic N can be attacked by other microorganisms, which are able to convert the organic N into an inorganic form, a process called nitrogen mineralization. This starts with ammonification, and under aerobic conditions and favorable soil pH ammonification is followed by nitrification in the sequence:

$$\text{Organic N} \rightarrow NH_4^+ \rightarrow NO_2^- \rightarrow NO_3^-$$

Ammonification is carried out by a broad spectrum of heterotrophic organisms; nitrification only by a small number of autotrophic bacteria [44]. The microbial oxidation of NH_4^+ to NO_2^- and NO_2^- to NO_3^- requires oxygen and hence proceeds only under aerobic conditions. The oxidation of NH_4^+ to NO_2^- is brought about by species of *Nitrosomonas, Nitrosolobus,* and *Nitrospira*, oxidation of NO_2^- to NO_3^- by species of *Nitrobacter*. They all require weak acid to neutral soil conditions; in acid soils nitrification is more or less blocked.

Ammonium ions produced by microbial breakdown of organic N, including urea, as well as NH_4^+ fertilizer can also be fixed by 2 : 1 clay minerals. In this form, NH_4^+ is protected against nitrification and leaching, but, depending on the type of clay minerals, may still be available to plant roots [45]. This fixed NH_4^+ fraction is of major significance for plant nutrition in soils derived from loess.

The concentration of ammonium ion in the soil solution is governed by the equilibrium

$$NH_4^+ \rightleftharpoons NH_3 + H^+ \text{pK} = 9.25$$

At pH < 6, there is virtually no NH_3 present; with an increase in pH the deprotonation of NH_4^+ increases, and so does the risk of NH_3 loss by volatilization. In alkaline and calcareous soils considerable amounts of N can thus be lost by the soil system [46]. High losses of NH_3 may also occur from the application of slurries, which generally have an alkaline pH [47].

Ammonium as well as NO_3^- are taken up by plant roots at high rates, and vigorous crop stands can deplete the NO_3^- concentration in the soil to a great extent. Nitrate is very mobile in soils since it is virtually unadsorbed on soil colloids. It can be leached by rainfall to deeper soil layers or even into the ground water (see Section Gains and Losses of Plant Nutrients). Nitrate losses may also occur under anaerobic soil conditions, for some bacterial species are able to use the oxygen of the NO_3^- as e^- acceptor for respiration. Nitrate is thus reduced to volatile NO, N_2O, and N_2 [49]:

$$NO_3^- \rightarrow NO_2^- \rightarrow NO_{(g)} \rightarrow N_2O_{(g)} \rightarrow N_{2(g)}$$

This process, brought about mainly by species of *Pseudomonas*, *Alcalignes*, *Azospirillum*, *Rhizobium*, and *Tropionibacterium*, is called denitrification. It may cause considerable soil N losses particularly in flooded rice soils, in which anaerobic conditions prevail [50].

Loss of NO_3^- by leaching or denitrification can be reduced by blocking the NO_2^- formation by application of nitrification inhibitors such as Nitrapyrin (2-chloro-6-trichloromethylpyridine), AM (2-amino-4-chloro-6-trimethylpyrimidine), or terrazole (5-ethoxy-3-trichloromethyl-1,2,4-thiadizole). There are also some natural compounds that are nitrification inhibitors [51]. The most important is neem, which occurs in the seeds of *Azadirachta indica*, a tree common in the tropics.

Some of the organic N in the soil may be incorporated into a very stable organic form. This nitrogen, which mainly occurs in humic acids, is hardly mineralized. In most soils the humus-N fraction is by far the largest, comprising 80 – 90 % of the total soil N. It is of great importance for soil structure, but has hardly any relevance as a nutrient reserve.

The fraction of hydrolyzable soil N can be mineralized and thus may serve as a source for N absorbed by plants. The most important fraction in this respect is the N of the biomass, which comprises ca. 40 – 200 kg/ha, thus only a small fraction of total soil N, which may be 2000 – 8000 kg/ha in arable soils.

2.3. Nutrient Availability

2.3.1. *Factors and Processes*

From the total amount of N present in the soil, only a small proportion can be made available for plants (see above). This is also true for other plant nutrients. For example, in a clay-rich soil ca. 200 000 kg K^+ may be present in 1 ha within the rooting depth of 80 – 100 cm, but only 1 % may be available to plant roots.

Plant nutrient availability depends on physicochemical and biological factors. A young root pushing into a soil directly contacts only a small amount of macronutrients, which would contribute only a few percent of the total nutrient demand. By far the greatest proportion of nutrients (NO_3^-, NH_4^+, K^+, Ca^{2+}, Mg^{2+}) required by the plant must be transported towards the plant roots. This transport can be brought about by mass flow and/or diffusion. In *mass flow*, the nutrients are moved with the water flow from the soil towards the roots. Therefore mass flow depends much on water uptake and transpiration conditions. At zero transpiration (100 % relative humidity, rainy, or foggy weather) mass flow is also zero. Mass flow plays a major role for the transport of Ca^{2+}, Mg^{2+}, and also for NO_3^- in cases where the NO_3^- concentration of the soil solution is high, e.g., after fertilizer application.

Nutrients that are taken up at high rates by plant roots (K^+, NH_4^+; NO_3^-, phosphate) but that have relatively low total concentration in the soil solution are mainly transported by *diffusion*. Uptake of nutrients by roots decreases the nutrient concentration near the root surface and establishes a concentration gradient, which drives the diffusive flux of nutrients from the soil towards the plant roots. Absorbing roots thus act as a sink for plant nutrients. Typical depletion profiles are shown in Figure 8 for phosphate [52]. Phosphate and K^+ concentrations in the soil solution at the root surface may be as low as 1 μM, whereas in the bulk soil solution concentrations of 50 to 300 μM phosphate and 500 to 1000 μM K^+ may prevail. Fertilizer application increases the nutrient concentration of the soil solution and hence the concentration gradient that drives the nutrients towards the roots. Therefore, the level of the nutrient concentration in the bulk soil solution is an important factor of nutrient availability.

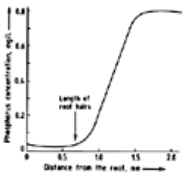

Figure 8. Phosphate depletion around a corn root: P concentration in the soil solution as a function of the distance from the root surface [52]

Diffusive flux and mass flow in the soil depend much on soil moisture. The dryer a soil, the smaller the water cross section (the size of soil pores that are still filled with water) and the more the nutrient flux is hampered. Therefore, soil moisture is another important factor of nutrient availability [53]. A third important factor of nutrient availability is the *nutrient buffer power* of a soil, a factor of particular relevance for phosphate, K^+, and NH_4^+. Here buffer power means the capability of a soil to maintain the nutrient concentration level in the soil solution (in analogy to the hydrogen-ion buffer power, see Section Soil pH, Buffer Power, and Liming). In well-buffered soils nutrients absorbed from the soil solution by roots are replenished by nutrient desorption, e.g., by cation exchange.

The most important biological factor for nutrient availability is root growth [54]. For most plant nutrients, only the soil volume around the root can be exploited by the plant. For phosphate the depletion zone extends only a few millimeter from the root surface and depends much on the length of root hairs, as can be seen from Figure 8. For K^+ and NH_4^+, the depletion zone is more extended, ca. 1 – 4 cm from the root surface. Root mass, root length, and root hairs therefore are of great importance for the portion of soil volume that can be exploited by a crop stand.

Roots excrete organic materials such as organic acids, sugars, and slimes from which bacteria feed. As a result, the bacterial colonization in the rhizosphere (the volume around the root) is much denser than in the soil apart from the roots. These bacteria are involved in the N turnover, e.g., for the N_2 fixation of free-living bacteria and for denitrifying bacteria [55].

Proton excretion of roots affects bacterial activity in the rhizosphere, the dissolution of calcium phosphates, and the cation exchange. Net proton release of plant roots is strongly affected by the type of N supply. Ammonium nutrition results in a high net release of H^+, nitrate nutrition in release of OH^- and HCO_3^-. Leguminous species living in symbiosis with *Rhizobium* are known for a high H^+ release by roots and therefore have a strong acidifying effect on soils.

Nutrient deficiency, e.g., phosphate or Fe deficiency, also increases net release of H^+, which may contribute to the dissolution of iron oxides and calcium phosphates. The release of avenic acid and mugineic acid by plant roots is of particular importance for the mobilization of Fe in the rhizosphere [56].

Plant growth may be enhanced after infection with mycorrhizal fungi, which leads to increased nutrient uptake due to increases in the effective absorptive surface of the root, mobilization of sparingly available nutrient sources, or excretion of ectoenzymes or chelating compounds. Furthermore, mycorrhizal fungi may protect roots from soil pathogens [57] and and in this way enhance root growth and nutrient acquisition of the host plant. This is particularly important when considering the nutrition of plants with immobile nutrients such as phosphorus, as fungal hyphae are known to absorb P and translocate it into the host plant [58].

2.3.2. *Determination of Available Plant Nutrients in Soils*

The level of available plant nutrients in soils can be assessed by means of plant analysis and soil analysis. In the case of *plant analysis*, the nutrient content of a particular plant organ at a certain physiological stage may reflect the nutritional status of the plant, hence also the nutrient status of the soils [59]. Such diagnostic plant analysis is particularly common for perennial crops, including fruit trees and forest trees.

In *soil analysis*, soil samples are extracted with special extractants. The quantity of nutrients extracted reflects the level of available nutrients in the soil. Numerous soil extractants have been developed. In Central Europe the DL method (double lactate method) and the CAL method (calcium acetate lactate method) are widely used for the determination of available soil K and P. In the Netherlands, available P is extracted by water (P-water method). In the United States, the Olsen method (extraction with NH_4F + HCl) is used for the determination of available soil phosphate [60]. Ion-exchange resins are also useful tools for the determination of available plant nutrients [61].

Of particular interest is the determination of available soil N early in spring before the first application of N fertilizer. Mineral N (NO_3^-, exchangeable NH_4^+) is extracted with a $CaCl_2$ or K_2SO_4 solution [62]. This technique, called the N_{min} method, provides reliable information on the level of directly available soil N. The easily mineralizable N in the soil, which frequently is the important fraction for the release of available N during the growth period, is not obtained by the N_{min} method.

Electro-ultrafiltration (EUF method) has been used for the determination of available soil nutrients [63]. This method uses an electric field to separate nutrient fractions from a soil suspension. Most plant nutrients can be extracted. The advantage of this method is that, besides inorganic nitrogen, the readily mineralizable organic N fraction is extracted [64].

In the last decade S has attracted interest as a plant nutrient. Numerous procedures have been proposed for the determination of plant-available S in soils. The procedures include extraction with water, various salts and acids, and S mineralization by incubation [65].

The relationship between soil analysis data and the response of crops to fertilizer application is not always satisfying since other factors may interfere, such as rooting depth, root morphology, soil moisture, and particularly the clay content. These factors should be taken into consideration in interpreting soil test data.

2.4. Physiology of Plant Nutrition

2.4.1. *Nutrient Uptake and Long-Distance Transport in Plants*

Oxygen and CO_2 are mainly taken up by aboveground plant parts. The process of uptake is a diffusion of CO_2 and O_2 into the plant tissue. For the entry of these gases into the plant, the stomata are of major importance. Water and other plants nutrients are mainly absorbed from the soil solution. The rate of nutrient uptake increases with the concentration of the particular nutrient in the soil solution, the rate of uptake leveling off at higher soil-solution concentrations.

Plant nutrients in the soil solution, mainly present in ionic form, diffuse into the root tissue. The outer plasma membrane of the cells (plasmalemma) is a great diffusion barrier. The transport of nutrients across this barrier is the proper process of ion uptake. This transport is not mere diffusion, but is related to specific membrane components and to metabolic processes that allow selective uptake of the plant nutrients, which is often associated with an accumulation of the nutrient in the cell. For example, the K^+ concentration in the cell (cytoplasm) may be higher by a factor of 10^2 – 10^3 than the K^+ concentration in the soil solution.

Nutrient uptake is initiated by an enzyme located in the plasma membrane called ATPase (ATP hydrolase). Its substrate is ATP. Hydrolyzation of ATP results in the splitting of H_2O into H^+ and OH^-, from which the latter remains in the cytosol of the cell while H^+ is extruded into the outer medium (Fig. 9). Thus an electrochemical potential is created between the two sides of the membrane. The proton motive force (p.m.f.) obtained in this way is described by the following equation [66]:

$$\text{p.m.f.} = -50\,\Delta\text{pH} + \Delta\varphi \quad \varphi = \text{electrical charge}$$

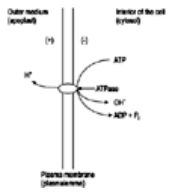

Figure 9. Scheme of plasmalemma-located ATPase, hydrolyzing ATP and pumping H^+ into the apoplast (proton pump)

The p.m.f. is the driving force for ion uptake. Cations are directly attracted by the negatively charged cell. Since the plasma membrane, however, represents a strong barrier, the entry of cation species must be mediated by particular carriers and ion channels. Little is known about these carriers and channels in plant membranes. These are assumed to be ionophores like valinomycin, nonactin, or gramicidin, which bind selectively to cation species and hence mediate a selective cation transport across the membrane. Such a type of carrier transport is shown in Figure 10. The carrier is hydrophobic and therefore quite mobile in

the membrane, which consists mainly of lipids. At the outer side of the membrane it combines selectively with a cation species, e.g., K^+. The cation carrier complex then diffuses to the inner side of the membrane, where the K^+ is released. Release and combining with K^+ depend on the electrochemical difference between the two sides of the membrane. High K^+ concentration and a positive charge favor the combining process; low K^+ concentration and a negative charge favor the release of K^+. Net transport rate becomes zero as soon as the electrochemical equilibrium is attained, which is governed by the Nernst equation:

$$E = \frac{RT}{zF} \ln \frac{a_o}{a_i}$$

E = electrical potential difference between the two sides of the membrane
F = Faraday constant
z = oxidation state of the cation
a_o = activity of the cation species in the outer solution
a_i = activity of the cation species in the inner solution

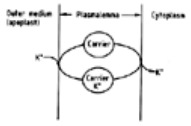

Figure 10. Scheme of K^+ carrier transport across the plasmalemma

The uptake of anions (NO_3^-, $H_2PO_4^-$) is also assumed to be driven by the plasmalemma-located ATPase. The anions presumably form protonated carriers at the outer side of the membrane and then are selectively tranported across the membrane. The protonated carrier – anion complexes are positively charged. Hence the electropotential difference between either side of the plasmamembrane represents the driving force for anion uptake.

Ion absorbed by cells of the root cortex are translocated via the symplasm in centripetal direction towards the central cylinder, where they are secreted into the xylem vessels. The actual process of this secretion is not yet understood. In the xylem the ions are translocated to the upper plant parts with the transpiration stream. They thus follow the vascular system of the plant and are distributed along the major and minor vein system of leaves from where they diffuse into the pores and intercellular spaces of cell wall (apoplast). The transport from the apoplast across the plasmamembrane into the cytoplasm of leaf cells is a process analogous to the nutrient uptake of root cells.

Some plant nutrients, such as N, P, K, and Mg, but not Fe and Ca, may also be translocated against the transpiration stream via the phloem tissue. These nutrients therefore may be transported from the tops to the roots or from older leaves to younger leaves.

2.4.2. *Effect of Nutrition on Growth, Yield, and Quality*

Meristematic growth requires plant nutrients: N and P for the synthesis of proteins and nucleic acids, K and Mg for the activation of enzymes and for membrane potentials, and all the other nutrients for various processes. The quantities required differ greatly but for practical purposes mainly N, K, and P, in some cases also Mg and Ca, limit plant growth. The growth rate is controlled by the nutrient with the lowest availability (LIEBIG's law of minimum). Application of this particular nutrient results in a growth response. This response is not linear, but rather follows a saturation type of curve (Fig. 11). Also crop yields as a function of increasing rates of fertilizer applicaton reflect this curve, which is also called the *Mitscherlich curve* since MITSCHERLICH [67] investigated these relationships extensively. The curve is described by the following equation:

$$\log (A - y) = \log (A - cx)$$

A = maximum yield
y = obtained yield
x = growth factor, e.g., fertilizer rate
c = constant

The term ($A - y$) is the increment that is required to attain the maximum yield. This increment becomes smaller as the variable x (growth factor) becomes greater. Such growth factors include not only plant nutrients, including CO_2, but also light intensity and temperature. The growth response obtained by these factors also follows the saturation type of curve shown in Figure 11.

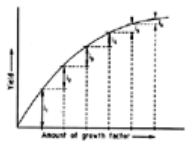

Figure 11. Mitscherlich curve, response curve of diminishing increments (i) with each further unit of growth factor

Growth and metabolic processes in plant tissues depend not only on the rates of plant nutrients supplied but also on the ratio in which the nutrients are provided. If the N supply is relatively high as compared with the supply of other nutrients, the synthesis of N-containing compounds, such as amino acids and proteins, is promoted. This may have a favorable effect on the protein synthesis in grains of cereals and hence improve their baking quality and their nutritional value. In grains for malting purposes and in particular barley, however, low protein content and high starch content are required. In this case, relatively high N rates have a detrimental effect on grain quality. An analogous case is the sugar beet, which should be high in sugar and low in N-containing compounds, especially amino acids. Relatively high N supply favors the synthesis of vitamins of the vitamin B group and the synthesis of carotenes in green plant tissue but has a negative effect on the content in vitamin C. High N rates may also increase the sensitivity of leaves and culms to fungi attack.

Phosphate and especially K^+ have a favorable impact on the energy status of plants. A relatively high supply of both nutrients promotes the synthesis of carbohydrates and the development of cell wall material, which increases the resistance against fungi attack.

2.5. Nutrient Balance

2.5.1. *Gains and Losses of Plant Nutrients*

In order to maintain the level of available plant nutrients in soils, the quantity of nutrients lost from the soil must be replenished. Nutrients may be lost by harvesting of plant material, leaching, transition into a non-available form in the soil, and volatilization. Nutrient gains result from fertilizer application, soil weathering, rainfall, and microbial dinitrogen fixation (see Section Organic Matter of Soils and Nitrogen Turnover). In most soils, weathering provides enough of those nutrients that are only required in minor amounts, such as Fe, Mn, Cu, Zn, Mo, and B. Young soils being rich in primary minerals also may weather appreciable amounts of Mg, K, and phosphate. Generally, the contribution of plant nutrients released by weathering is more important if the yield level is lower. Up to the mid-1980s, sulfur supply to crops was to a large extent brought about by SO_2 in the atmosphere and by SO_2 dissolved in precipitation. However, in the last two decades increased global environmental concern has prompted reductions in SO_2 emissions, which in past years were beneficial for crop growth in many countries [68]. With the reduction in the sulfur supply from atmospheric emissions, the greater use of high-analysis S-free fertilizers, new high-yielding crop varieties, and declining use of sulfur as a fungicide, sulfur deficiency became widespread in various parts of the world.

The quantity of nutrients taken up by crops depends mainly on the yield level and on the kind of plant organs. For a rough calculation of *nutrient uptake by crops* the following figures may be used:

1 t of grain (cereals) contains	
20 – 25	kg N
5	kg K
4	kg P
1 t of dry matter of green plant material (leaves, stems) contains	
15 – 30	kg N
15 – 30	kg K
2 – 4	kg P
1 t of roots (beets) or tubers contains	
1.5 – 3	kg N
4 – 5	kg K
0.3 – 0.6	kg P

Nutrient losses only occur from plant parts removed from the field. For example, if only grain is harvested and the straw remains on the field, only the plant nutrients present in the grain are lost from the system.

High nutrient losses may occur by *leaching*. The quantities lost depend much on climate, weather conditions, and soil properties. In Central Europe, with a rainfall of about 700 mm/a, ca. 25 – 50 % of the precipitation passes through the soil profile to a depth >1 m. Plant nutrients dissolved in this drainage water are lost from the soil. Table 10 shows maximum and minimum leaching rates obtained during a period of eight years by VÖMEL [69]. Generally, leaching rates are low if the soil clay content is high. High leaching rates prevail in fallow soils, low rates in soils with a permanent plant cover, e.g., grassland. Phosphate is effectively not leached. Leaching occurs mainly with the winter or monsoon rainfall. Under such humid conditions plant nutrients present in plant residues (stubble, straw, roots, leaves) also may be leached.

Table 10. Leaching rates of plant nutrients from soils [69]

Soil	Plant nutrient, kg ha^{-1} a^{-1}			
	N	K	Ca	Mg
Sand	12 – 52	7 – 17	110 – 300	17 – 34
Sandy loam	0 – 27	0 – 14	0 – 242	0 – 37
Loam	9 – 44	3 – 8	21 – 176	9 – 61
Clay	5 – 44	3 – 8	72 – 341	10 – 54

Volatile losses only play a role for nitrogen, which may be lost in form of NH_3, N_2, and N_2O (see Section Organic Matter of Soils and Nitrogen Turnover). Nutrient loss by conversion into a non-available form is only relevant for phosphate and in some cases for K^+.

2.5.2. *Alternative Plant Nutrition*

The terms “ecological agriculture” and “biological agriculture” are used as synonyms for “organic agriculture”. One of the main approaches of organic agriculture is a mixed farm system within a more or less closed nutrient cycle. With regard to plant nutrition the production system can be characterized by the following principles:

Nearly closed cycles of nutrients and organic matter within the farm
Predominant use of farmyard manure and compost
Slowly soluble P minerals, if necessary
K fertilizers not in the form of chloride

Besides nitrogen, which may be imported with manure, symbiotically bound nitrogen is the main N source. For this reason N_2 fixation by legumes is of great importance. Synthetic fertilizers are renounced. In the last few years major efforts were undertaken to quantify the cycles of nutrients as well as the organic matter cycles.

Organic agriculture has to deal with limited amounts of nutrients. Nutrient management, defined as systematic target-oriented organization of nutrient flow, is therefore considered as the optional combination of resources that are restricted or have to be released. Strategies must be adopted that make nutrients in the system internally available by achieving optimized utilization or which keep nutrients potentially available in the long term. The main nutrient flows in organic farm are fixed for the long term by organizing and optimizing the site-adapted crop rotation [70].

[Top of Page]

3. Standard Fertilizers

Heinrich Dittmar

Standard fertilizers include the products used in large quantities worldwide. They are applied to agricultural and large-scale garden crops. They can be classified as solid and liquid fertilizers.

3.1. Solid Fertilizers

Solid fertilizers are the most important group of fertilizers. Worldwide, nearly 90 % of all the nitrogen applied in fertilizers is in solid form (season or fertilizer year 1996/97). In Germany about 83.5 % of total nitrogen is applied in the solid form, and the corresponding figure in the United States is 44 %. Phosphate and potash are applied mainly in solid form. Solid fertilizers include granular, prilled, and compacted products (see Chap. Fertilizer Granulation).

A summary of the most important fertilizers can be found in [71].

3.1.1. *Straight Fertilizers*

Straight fertilizers contain only one nutrient, for instance, urea (N) or triple superphosphate (P_2O_5). Straight solid fertilizers are listed in Table 11.

Table 11. Straight fertilizers

Type	Production; main constituents	Analysis, wt %	Remarks	Ullmann keyword
Nitrogen fertilizers				
Calcium nitrate [13477-34-4]	synthetic	15.5 % N ($Ca(NO_3)_2$), NH_4 N, max. 1.5 %	fast acting; chiefly for special crops and refertilizing	→ Nitrates and Nitrites
Sodium nitrate [7631-99-4]	synthetic	16 % N ($NaNO_3$)	fast acting; chiefly for special crops and refertilizing	→ Nitrates and Nitrites
Chile saltpeter	from caliche; sodium nitrate	≥ 15 % N	fast acting; chiefly for special crops and refertilizing	
Ammonium nitrate [6484-52-2]	synthetic; may contain additives such as ground limestone, calcium sulfate, dolomite, magnesium sulfate, kieserite	34 % N(NH_4NO_3)	special storage provisions (see Chap. Storage, Transportation, and Application)	→ Ammonium Compounds – Ammonium Nitrate
Calcium ammonium nitrate (CAN)	synthetic; may contain additives such as ground limestone, calcium sulfate dolomite, magnesium sulfate, kieserite	26 – 28 % N, ≥ 2 % $CaCO_3$	versatile fertilizer, both fast and slow acting	→ Ammonium Compounds
Ammonium sulfate nitrate (ASN)	synthetic	added as limestone or dolomite 26 % N ($NH_4NO_3^+$ and $(NH_4)_2SO_4$), ≥ 5 % NO_3 N	versatile fertilizer with both fast and slow action	→ Ammonium Compounds
Magnesium ammonium nitrate	synthetic; ammonium nitrate and magnesium compounds (dolomite, magnesium carbonate, or magnesium sulfate)	20 % N (NH_4NO_3)	for soils and crops requiring magnesium	
Ammonium sulfate [7783-20-2]	synthetic, often as byproduct	≥ 6 % NO_3 N, 5 % MgO 21 % N ($(NH_4)_2SO_4$)	slow acting, for soils containing lime and/or high-precipitation areas	→ Ammonium Compounds – Ammonium Sulfate
Ammonium chloride [12125-02-9]	synthetic	≥ 25 % N (NH_4Cl)	rice fertilization, chiefly in Japan and India	→ Ammonium Compounds – Ammonium Chloride
Urea [57-13-6]	synthetic	46 % N ($CO(NH_2)_2$) biuret ≤ 1.2 %	optimal only when soil lime, heat, and moisture are appropriate; also suitable for foliar application; volatile losses are possible	→ Urea
Calcium cyanamide [156-62-7]	synthetic	20 – 22 % N; ≥ 75 % of total N from cyanamide; ca.40 % CaO 66 % N	slow acting, special requirements, also for weed control nitrification	→ Cyanamides – Calcium Cyanamide
Dicyandiamide [461-58-5]	synthetic		nitrification inhibitor, added pure to liquid manure; also as additive to straight and compound fertilizers containing ammonium	→ Cyanamides – Dicyandiamide
Phosphate fertilizers				
Basic slag (Thomas phosphate)	byproduct of steel production from ores containing phosphate; calcium silicophosphate	≥ 12 % P_2O_5 soluble in mineral acid, ≥ 75 % soluble in 2 % citric acid	phosphate containing up to 45 % CaO, versatile phosphate fertilizer	→ Phosphate Fertilizers
Superphosphate	digestion of ground phosphate rock with sulfuric acid; monocalcium phosphate with calcium sulfate	16 – 22 % P_2O_5 soluble in neutral ammonium citrate solution, 93 % water soluble	for all soils	→ Phosphate Fertilizers
Triple superphosphate	digestion of ground phosphate rock with phosphoric acid; monocalcium phosphate	45 % P_2O_5; soluble in neutral ammonium citrate solution, 93 % water soluble	for all soils	→ Phosphate Fertilizers
Partly digested phosphate rock	partial digestion of ground phosphate rock with sulfuric or phosphoric acid; monocalcium phosphate, tricalcium phosphate, calcium sulfate	≥ 20 % P_2O_5; soluble in mineral acid, 40 % water soluble	for all acidic soils (pH < 7)	→ Phosphate Fertilizers
Soft phosphate rock	grinding of soft phosphate rock; tricalcium phosphate and calcium carbonate	≥ 25 % P_2O_5 soluble in mineral acid, 55 % soluble in 2 % formic acid	for acidic soils and low-precipitation areas	→ Phosphate Fertilizers
Dicalcium phosphate [7789-77-7]	neutralization of phosphoric acid with calcium hydroxide dicalcium phosphate dihydrate	≥ 38 % P_2O_5 soluble in alkalized ammonium citrate; not water-soluble	for all soils	→ Phosphoric Acid and Phosphates
Thermal (fused)	thermal digestion of ground	≥ 25 % P_2O_5 soluble	for all soils and locations	

phosphate	phosphate rock with alkali-metal compounds and silicic acid; alkali-metal calcium phosphate and calcium silicate	in alkalized ammonium citrate solution and in citric acid		
Aluminum calcium phosphate	thermal digestion and grinding; amorphous aluminum and calcium phosphates	29 – 33 % P_2O_5 soluble in mineral acid, 75 % soluble in alkalized ammonium citrate solution	for all soils and locations	
Potassium fertilizers				
Potash ore	run-of-mine potash	≥ 10 % K_2O, ≥ 5 % MgO, contains sodium	for soils and crops where Na and Mg are important, especially cattle fodder	→ Potassium Compounds
Beneficiated potash ore	beneficiation of run-of-mine potash and mixing with potassium chloride	≥ 18 % K_2O	for all soils	→ Potassium Compounds
Potassium chloride [7447-40-7]	beneficiation of run-of-mine potash	≥ 37 % K_2O (up to 60 %)	for all crops not sensitive to chloride	→ Potassium Compounds
Potassium chloride with magnesium	beneficiation of run-of-mine potash and addition of Mg salts	≥ 37 % K_2O, ≥ 5 % MgO	on Mg-deficient soils for all crops not sensitive to chloride	→ Potassium Compounds
Potassium sulfate [7778-80-5]	synthetic	≥ 47 % K_2O, ≤ 3 % Cl	for all crops sensitive to chloride	→ Potassium Compounds
Potassium sulfate with magnesium	synthetic, from potassium sulfate, with addition of Mg salts; potassium sulfate and magnesium sulfate	≥ 22 % K_20, ≥ 8 % MgO; ≤ 3 % Cl	deficient soils for all crops sensitive to chloride	→ Potassium Compounds

3.1.2. *Multinutrient Fertilizers*

Multinutrient or compound fertilizers contain two or more nutrients. The term complex fertilizer refers to a compound fertilizer formed by mixing ingredients that react chemically. In bulk-blend or blended fertilizers (see Chap. Fertilizer Granulation), two or more granular fertilizers of similar size are mixed to form a compound fertilizer.

There are several routes for manufacturing fertilizers. An important route is the Odda process, in which phosphate rock is digested with nitric acid. Calcium nitrate is crystallized by cooling and removed, and the mother liquor is neutralized by addition of gaseous ammonia (see → Phosphate Fertilizers – Compound Fertilizers by the Nitro Route). Other important processes are based on digestion of phosphate rock with sulfuric acid and the so-called mixed-acid process [355]. For the production of NPK fertilizers, potassium compounds are added in the desired amount to the corresponding slurries. Magnesium (as kieserite or dolomite) and micronutrients may also be added.

NPK Fertilizers. According to the EEC Guidelines, NPK fertilizers must contain at least 3 % N plus 5 % P_2O_5 plus 5 % K_2O and at least 20 % total nutrients. The most commonly used grades (N-P_2O_5-K_2O, each in wt %) are

Nutrient ratio 1 : 1 : 1
15–15–15, 16–16–16, 17–17–17
Nutrient ratios 1 : 2 : 3 and 1 : 1.5 : 2
5–10–15, 6–12–18, 10–15–20
Nutrient ratio 1 : 1 : 1.5 – 1.7
13–13–21, 14–14–20, 12–12–17
Nutrient ratios 3 : 1 : 1 and 2 : 1 : 1
24–8-8, 20–10–10
Low-phosphate grades
15–5-20, 15–9-15

If additional numbers are given in a fertilizer grade, the fourth is the wt % MgO and the fifth is wt % S. Micronutrient contents may also be stated. In some countries the grade is expressed in terms of the elements rather than oxides.

The NPK fertilizers have the important advantage of simplified application, since all the important nutrients can be distributed in one operation. Each grain of fertilizer has the same content of nutrients. Serious errors are prevented by the harmonic nutrient ratio, provided the metering rate is correct. The nitrogen is usually present as nitrate and ammonium N in roughly equal parts. The phosphate is 30 – 90 % water-soluble, the rest being soluble in ammonium citrate solution. Most of the potassium is present as the chloride; the sulfate is used for chloride-sensitive crops.

In bulk-blended products (see Section Bulk Blending), the individual fertilizers are combined in the desired nutrient ratio. The operational advantage is the same as that of synthetic NPK fertilizers. Because, however, the individual granular products may differ in grain-size spectrum, granule surface characteristics, and density, segregation may occur during handling, storage, packaging and even during application. These products are therefore suitable for practical use only when the individual components have similar physical qualities.

NP Fertilizers. The minimum analysis for NP fertilizers under the EEC Guidelines is 3 % N and 5 % P_2O_5 and at least 18 % total nutrients. Common grades are 20–20, 22–22, 26–14, 11–52, 16–48, and 18–46. These products are appropriate for potassium-rich soils or where potash is supplied as a separate fertilizer.

NK Fertilizers. The minimum analysis for NK fertilizers under the EEC Guidelines is 3 % N and 5 % K_2O and at least 18 % total nutrients. These products are suitable for phosphate-rich soils or where phosphate is distributed separately.

PK Fertilizers. In the group of PK fertilizers, all combinations of the straight phosphate and potassium components listed in Table 11 are possible. In general, the materials are first milled and then mixed and granulated, so that a fairly homogeneous mixture is obtained. Some products are also made by bulk blending. The EEC Guidelines set forth a minimum analysis of 5 % P_2O_5, 5 % K_2O, and at least 18 % nutrients. Magnesium and micronutrient boron can be added.

3.1.3. *Lime Fertilizers*

Solid fertilizers also include lime fertilizers. The main purpose of using lime is to optimize soil pH; a secondary purpose is to supply calcium as a plant nutrient. The use of large amounts of lime to increase the pH is referred to as soil-improvement liming. Use to maintain the present pH is called maintenance liming.

The starting materials for lime fertilizers are limestone [1317-65-3] and dolomite [17069-72-6]. These are marketed in various forms. National fertilizer regulations (e.g., the type list in the Federal Republic of Germany's fertilizer law, see Chap. Legal Aspects) govern the fineness of grinding. Calcination at

900 – 1400 °C yields quicklime CaO; hydration with water gives slaked lime $Ca(OH)_2$. The most immediately effective form is finely divided quicklime with a particle size of ca. 0.15 mm or less. If a lime grade contains at least 15 % MgO or $MgCO_3$, it can be called magnesium lime. The content of MgO depends on the starting material (proportion of dolomite) and may be up to 40 %.

Calcium carbonates [471-34-1], $CaCO_3$, act slowly. They are recommended above all for lighter soils. Unslaked and slaked limes are faster acting. They are suitable mainly for loamy and clayey soils.

Besides these lime products from natural rock, there are also limes from industrial processes. They include smelting lime and converter lime, which are steel-industry byproducts. Converter lime with phosphate must contain 40 % CaO + MgO and also at least 5 % P_2O_5. These products must also satisfy minimum requirements on fineness to ensure sufficiently fast action.

Finally, the refining of beet sugar yields a product containing lime. According to the fertilizer law in the Federal Republic of Germany, this lime form must contain at least 30 % CaO and 5 % MgO.

3.1.4. *Magnesium Fertilizers*

The following magnesium fertilizers are offered in commerce: kieserite [14567-64-7] (27 % MgO) for soil application, and Epsom salts [7487-88-9] (17 % MgO) for foliar application in liquid form.

3.2. Liquid Fertilizers

Only solid fertilizers were produced and used up to ca. 1950: lower production costs, higher nutrient concentrations, and the ease of making compound fertilizers were for a long time the reasons that development was limited to the solid fertilizers. Since then, liquid fertilizers have also been developed and supplied to agriculture. Particularly in the United States, liquid fertilizers have come into heavy use. Modern production methods have lowered production costs, even for liquid fertilizers with high nutrient concentrations.

Homogeneous liquid fertilizers, in contrast to solid fertilizers, present no special problems during application by the farmer. Furthermore the storage of liquid fertilizers is less difficult than that of solid ones. Small amounts of herbicides and insecticides can be mixed with liquid fertilizers far more easily. Production plants for solid fertilizers are much more costly than those for liquid fertilizers. Liquid fertilizers have the drawbacks that they usually have lower nutrient content and are sensitive to impurities, as well as to precipitation and crystallization, especially caused by magnesium and fluorine.

Liquid fertilizers are classified as (1) anhydrous ammonia [7664-41-7] with 82 wt % N at high pressure, (2) aqueous ammonia with up to 24 wt % N, (3) ammoniates (ammoniated ammonium nitrate and/or urea solutions) with up to 50 wt % N at moderate to atmospheric pressure, and (4) nonpressure urea – ammonium nitrate solutions with up to 32 wt % N [72]. Compound liquid fertilizers include both clear liquids and suspensions.

The growing interest in liquid fertilizers results in large part from the fact that field spraying can be used for specific, exactly meterable, inexpensive, clean application with an injector. Even low nitrogen rates of 15 – 30 kg/ha can be exactly and uniformly metered out in liquid form.

Liquid-fertilizer plants are classified according to type of operation as hot mix or cold mix. Hot-mix plants use phosphoric acid and ammonia, whereby hot mix refers to the heat of reaction. Cold-mix plants mix ammonium phosphate solution with other ingredients; no heat of reaction is evolved. Cold-mix plants are popular in the United States, because they are simple and inexpensive. They are essentially a blending and mixing operation, the liquid counterpart of the bulk-blend plants (see Section Bulk Blending).

Table 12 lists the consumption of nitrogen in liquid fertilizers and their share of total nitrogen consumption by countries. In Germany consumption of nitrogen in liquids has risen since the season 1985/86, but total nitrogen consumption is decreasing [72]. The use of liquid ammonia plays no role in Germany and has strongly declined in Denmark. In the United States more nitrogen is consumed in liquid than in solid form.

Table 12. Consumption of liquid fertilizers and their share of total nitrogen consumption

Country and fertilizer year *	Liquid ammonia direct application		Nitrogen solutions		Total liquids		Total N
	10^3 t N	%	10^3 t N	%	10^3 t N	%	10^3 t N
World							
1985/86	5022.2	7.13	3811.1	5.41	8833.3	12.54	70 461.6
1993/94	5052.4	6.97	3801.9	5.24	8854.3	12.21	72 497.7
1994/95	4122.8	5.69	3794.3	5.24	7917.1	10.93	72 454.8
1995/96	4649.2	5.92	4013.6	5.11	8662.8	11.03	78 592.5
1996/97 **	4688.1	5.67	4097.0	4.96	8785.1	10.63	82 645.6
United States							
1985/86	3400.5	35.96	1887.6	19.96	5288.1	55.92	9457.0
1993/94	4116.3	35.89	2543.0	22.17	6659.3	58.06	11 469.4
1994/95	3308.7	31.12	2514.0	23.65	5822.7	54.77	10 632.1
1995/96	3693.6	33.09	2642.8	23.68	6336.4	56.77	11 161.5
1996/97 **	3651.6	32.65	2651.3	23.71	6302.9	56.36	11 184.5
Germany							
1985/86	15.0	0.66	65.0	2.84	80.0	3.50	2285.7
1993/94	0.0	0.00	205.0	12.72	205.0	12.72	1612.0
1994/95	0.0	0.00	226.0	12.65	226.0	12.65	1787.0
1995/96	0.0	0.00	282.0	15.93	282.0	15.93	1770.0
1996/97 **	0.0	0.00	290.0	16.53	290.0	16.53	1754.0
France							
1985/86	31.0	1.29	482.0	20.02	513.0	21.31	2408.0
1993/94	37.0	1.67	560.0	25.20	597.0	26.87	2222.0
1994/95	35.7	1.55	566.4	24.54	602.1	26.09	2308.3
1995/96	42.7	1.79	577.0	24.13	619.7	25.92	2391.7
1996/97 **	44.8	1.78	643.2	25.48	688.0	27.26	2523.9
Denmark							
1985/86	102.0	26.69	0.0	0.00	102.0	26.69	382.1
1993/94	0.0	0.00	4.0	1.23	4.0	1.23	325.0
1994/95	0.0	0.00	6.0	1.90	6.0	1.90	316.0
1995/96	18.0	6.19	7.0	2.41	25.0	8.60	291.0
1996/97 **	17.0	5.94	7.0	2.45	24.0	8.39	286.0

* The fertilizer year (season) runs from July 1 to 30 next year.
** Preliminary.

Liquid fertilizers can be easily mixed with a variety of micronutrients and agricultural pesticides. The dissolving of micronutrients in clear liquid fertilizers is promoted by gelatinizing agents or by complexing with polyphosphates. The pulverized additions such as copper sulfate, sodium molybdate, sodium borate, zinc oxide, or manganese oxide, which are mixed with suspensions [73], can be suspended in water by stirring.

Nonpressure liquid fertilizers are generally stored in concrete or mild steel tanks [308, p. 129]. A corrosion inhibitor is required in the case of nonpressure urea – ammonium nitrate solution. Concrete tanks have the shape of vertical cylinders. The tanks should be lined with a film of a durable plastic such as poly (vinyl chloride). Before lining, the concrete tank must be protectively coated [74]. For pressurized liquid fertilizers, double-walled vessels with leak warning systems should be employed. When single-walled tanks are used, a retaining basin must be erected.

3.2.1. *Nitrogen Liquids*

Liquid Ammonia. Since 1950, liquid ammonia has found increasing use as a direct-application fertilizer, especially in the United States, since it can be produced in large amounts at low cost. Because of its high vapor pressure (6 bar at 10 °C, 9 bar at 20 °C, and 12 bar at 30 °C), anhydrous ammonia must be stored and transported in pressure vessels and applied with special equipment.

Ammonia is injected into the soil to a depth of roughly 15 cm with injection prongs. In general, the equipment suitable for this purpose deposits the fertilizer in a pipe whose diameter depends on the soil type and soil moisture content. Figures 12 and 13 show the ammonia loss as a function of the soil type, the soil water content, and the depth of application [75]. If the soil contains 15 % water, the NH_3 loss is virtually independent of depth and, at about the 1 % level, can be neglected. Despite some advantages, the application of liquid ammonia is too expensive for most farmers in the EC. A publication for farmers on the safe use of liquid ammonia is available [76]. For storage and transportation, see ⟶ Ammonia and [308, p. 73]

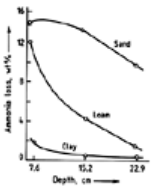

Figure 12. Ammonia losses as a function of application depth for three types of air-dry soils

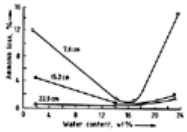

Figure 13. Ammonia loss as a function of soil water content for three depths of application

Pressurized Nitrogen Liquids. Together with ammonia, ammonium nitrate and/or urea can be converted into pressurized aqueous nitrogen fertilizers. The following categories are distinguished: low-pressure solutions (up to 1.3 bar) for direct application in agriculture, medium-pressure solutions (up to 7 bar), mostly used for ammoniating superphosphates or phosphoric acid, and high-pressure solutions (over 7 bar), used only for ammoniation. Table 13 lists some nitrogen solutions with their saturation temperatures and vapor pressures at 40 °C (313 K) [77].

Table 13. Nomenclature and physical properties of some nitrogen solutions

% N × 10	Composition *	Vapor pressure at 40 °C, bar **	Saturation temperature, °C
410	22–65–0	0.7	– 6
410	26–56–0	1.2	–32
444	25–55–10	1.5	–29
453	31–0-43	3.4	+ 8
454	37–0-33	4.0	– 9
490	33–45–13	3.6	–27
490	34–60–0	3.4	–47
530	49–36–0	7.3	–73
580	50–50–0	10.1	–34

* HN_3–NH_4NO_3–$(NH_2)_2CO$, wt %.
** Multiply by 0.1 to convert to MPa.

Nonpressure Nitrogen Liquids. Nitrogen solutions that contain little or no ammonia have no significant vapor pressure at ambient temperatures. The most common nonpressure liquid nitrogen fertilizers comprise mixtures of ammonium nitrate, urea, and water (UAN), which are the most popular nitrogen fertilizers in the United States. A commercial solution might have the following composition:

Ammonium nitrate	39.5 % = 14 % N
Urea	30.5 % = 14 % N
Water	30.0 %
UAN	100.0 % = 28 % N

The density at 15 °C is ca. 1.28 g/cm^3, corresponding to 36 kg of nitrogen in 100 L of the product.

Figure 14 gives solubility isotherms and phase boundaries for the ammonium nitrate – urea – water system. In order to optimize the amount of UAN as a function of temperature during transportation and in the field, the UAN composition is adjusted to suit the conditions [77], [79]. This is especially important for winter application. Among almost-nonpressure nitrogen solutions, mixtures containing added NH_3 (see Table 14) should also be mentioned [80].

Table 14. Composition, nitrogen concentration (N), and crystallization temperature (T_c) of nitrogen solutions [80]

Composition, wt %				N,	T_c,	Pressure at 10 °C,
NH_3	NH_4NO_3	Urea	H_2O	wt %	°C	bar *
24	56	10	10	44	–26	0.3
20	45	15	20	39.2	–32	0.2
19	58	11	12	41	–14	0.2

* Multiply by 0.1 to convert to MPa.

Figure 14. The ammonium nitrate – urea – water system [77], [78]

Solid lines are solubility isotherms; dotted lines show phase boundaries; dashed lines show constant nitrogen content.

A urea – ammonium nitrate solution can be produced by dissolving solid urea in an ammonium nitrate solution. If the raw materials are solid urea and ammonium nitrate, UAN can be produced in a slightly modified dissolver. Figure 15 gives a simplified diagram of such a production unit.

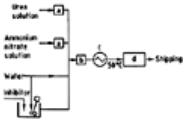

Figure 15. Production of urea – ammonium nitrate solution [112]

a) Ratio controller; b) Mixer; c) Cooler; d) Storage tank

The desired ammonium nitrate : urea ratio is obtained by mixing a 75 – 80 % urea solution at 120 °C with an 80 – 85 % ammonium nitrate solution at 40 °C, the quantities being controlled. After the addition of water, the liquid fertilizer is transferred to a storage tank after passing via a cooler. The mixer and cooler are made of stainless steel, and the equipment downstream of the cooler are made of carbon steel [94]

Since urea – ammonium nitrate solutions attack ordinary steels, steel storage tanks and tank cars should be coated with Derakane-470, a poly(vinyl ester) [82], or lined with polypropylene or polyethylene [83]. This practice also prevents stress-corrosion cracking of the steels, especially in the presence of $(NH_4)_2CO_3$ [84]. Poly(vinylidene fluoride) provides excellent corrosion protection up to 120 – 130 °C [83]. The most common inhibitor for these solutions is anhydrous ammonia, which is used to adjust the pH to 7. Ammonium thiocyanate and ammonium phosphate are also effective [308, p. 129]

In the production and use of UAN, such solutions can explode if they are evaporated to <5 % water and simultaneously heated, especially when the two components have become segregated [85].

3.2.2. *Multinutrient Liquids*

Liquid mixed fertilizers consist of aqueous solutions and dispersions of the nutrients nitrogen, phosphate, and potash. These liquid mixed fertilizers are produced mainly from phosphoric acid, anhydrous ammonia or aqueous ammonia, ammonium nitrate, urea, and potassium salts, chiefly KCl but also K_2SO_4 (hot mix). In most cases, fine-grained potassium chloride, with a somewhat higher potash content (62 % K_2O) than the normal fertilizer salt, is employed because it dissolves better. If only solid raw materials are used, the N-P-K grade of the liquid mixed fertilizer can be varied only over a narrow range. The following solid nitrogenous components are mainly employed: ammonium nitrate, urea, monoammonium phosphate, and diammonium phosphate (cold mix).

For intensive field and greenhouse cultivation of vegetables, etc., the nutrients can be introduced into the watering system in the form of a low-percentage nutrient solution. Because this solution is often delivered directly to the roots through hoses, it should have the lowest possible level of impurities.

3.2.2.1. *NP Liquids*

In the early 1950s the commercial production of mixed liquid fertilizers was started in the United States. Pure electric-furnace orthophosphoric acid was used as the phosphate source. The acid was ammoniated to give an 8–24–0 grade liquid which does not salt out at – 8 °C. At mixing plants, this base solution was mixed with urea – ammonium nitrate solution to give a liquid containing ca. 28 – 32 % nitrogen. An additional potassium content was obtained by dissolving potassium chloride in the liquid. Maximum-mix grades attainable in this way included 13–13–0 and 7–7-7. A problem was the low concentrations of the 8–24–0 base solution and the final liquid mixes. Moreover, use of the cheaper wet-process orthophosphoric acid was not possible, because on ammoniation of such acids, iron, aluminum, magnesium and other impurities form voluminous, gelatinous, and crystalline precipitates that are difficult to handle.

In 1957 TVA introduced superphosphoric acid and base solutions made by ammoniation thereof [86]. The superphosphoric acid, initially produced by the electric-furnace process, contained ca. 76 % P_2O_5 as compared to 54 % P_2O_5 in commercial wet-process acid. At this higher P_2O_5 concentration, ca. 50 % of the P_2O_5 is in the form of pyrophosphoric, tripolyphosphoric, and tetrapolyphosphoric acids [88]. Because of higher solubility of the ammonium salts of these polyphosphoric acids, ammoniation of superphosphoric acid allowed production of higher grade base solution (10–34–0 instead of 8–24–0) and higher grade mixed fertilizer solutions. A further advantage was that the polyphosphates are effective sequestering agents that prevent precipitation of many metal ions. The kinetics of formation of ammonium polyphosphates are reported in [87].

Although the introduction of superphosphoric acid was an effective solution to the problem of low-grade liquids, there were still the high costs of the electric-furnace acid process. Studies were undertaken to produce usable superphosphoric acid by concentration of the cheaper (54 % P_2O_5) wet-process phosphoric acid to about 72 % P_2O_5 and a polyphosphate content of 45 – 50 %. Problems of the wet-process superphosphoric acid of this concentration were its high viscosity and the poor storage properties of the 10–34–0 base solution derived from it. Although the polyphosphate content of the liquid initially keeps impurities in solution in complexed form, precipitates may still form during storage if the acid used in ammoniation had too high an impurity level and the polyphosphate concentration is too low. A frequent problem with NP liquids is the precipitation of insoluble magnesium compounds. If 0.4 wt % magnesium (calculated as oxide) is present, the storage time before appearance of cloudy precipitates is only about one month. For magnesium oxide concentrations of 0.2 % and 0.1 %, the storage life is about 1 – 2 years [89].

The usual raw material for phosphoric acid productions, apatite, with its impurities and its variable composition, is not a simple raw material. In the production of phosphoric acid, however, most of the impurities (e.g., aluminum, iron, and magnesium compounds) remain behind in the acid slurry.

The removal of the impurities from the wet-process phosphoric acid (Fe, Al, Mg, Ca, SO_3, F) is necessary for storage and application of the liquid fertilizer. Two typical processes reduce the concentrations of the most troublesome impurities significantly (Table 15) [95]. Process B, which is used in Europe, reduces the impurity concentrations to lower levels than the U.S. Process A, while not reducing P_2O_5 concentration as much; however, Process B is more expensive.

Table 15. Purification of wet-process phosphoric acid

	Raw acid	Purified acid Process A	Purified acid Process B
P_2O_2	52.5	40.4	48.9
Fe	0.6	0.2	0.07
Al	0.9	0.3	0.01
Mg	0.3	0.03	0.001
Fe	0.8	0.3	0.06

The cleaning of the wet-process phosphoric acid by means of continuous extraction with ammonia and acetone has been described for the “green acid” from North Carolina phosphate rock and the “black acid” from Florida phosphate rock [90]. An extraction process is described in [91]. For the use of chelating agents, see [92, p. 58].

Addition of ammonium fluoride to aqueous aluminum phosphate-containing slurries precipitates aluminum as $(NH_4)_3AlF_6$ and forms ammonium and diammonium phosphate:

$$6\,NH_4F + AlPO_4 \rightarrow (NH_4)_3AlF_6 + (NH_4)_2HPO_4 + NH_3$$

resulting in a filtrate that can be subsequently worked up to give a liquid fertilizer.

Figure 16 shows how the polyphosphate content and the N : P_2O_5 ratio affect the solubility of ammonium phosphates at 0 °C. The bottom curve represents the solubility of the orthophosphates. The middle and top curves demonstrate the increase in solubility as the polyphosphate content increases [93].

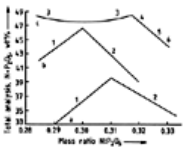

Figure 16. Effect of polyphosphate content and N : P_2O_5 ratio on solubility of ammoniated phosphoric acids at 0 °C

Polyphosphate contents as percent of total P_2O_5: a) 0; b) 45; c) 70

Crystallizing phases: 1) $(NH_4)H_2PO_4$; 2) $(NH_4)_2HPO_4$; 3) $(NH_4)_3HP_2O_7 \cdot H_2O$; 4) $(NH_4)_5P_3O_{10} \cdot 2\,H_2O$; 5) $(NH_4)_2HPO_4 \cdot 2\,H_2O$; 6) $(NH_4)_4P_2O_7 \cdot H_2O$

Table 16 characterizes NP base solutions made from superphosphoric acid. The two solutions are 10 – 34 from wet-process superacid and 11 – 37 from furnace-grade superacid. The 10 – 34 NP solution is used predominantly [77].

Table 16. Composition of NP base solutions [77]

Grade	Phosphoric acid, wt % P_2O_5	Percentage composition * Ortho-phosphate	Pyro-phosphate	Tripoly-phosphate	Tetrapoly-phosphate	Higher phosphates
10–34–0	76	49	42	8	1	0
11–37–0	79 – 80	29	42	21	5	3
	80 – 81	20	37	23	10	10

* Distribution of P_2O_5 sources: of the total P_2O_5 present (34 or 37 wt %), the percentages given in the table are present in the form of orthophosphate, polyphosphate, etc.

A large number of experimental data points were subjected to regression analysis. With the resulting equation, the total nutrient concentration in NP solutions made from ammonium phosphates and urea at 0 – 50 °C can be calculated with a relative error of 1 % or else determined with a nomograph [80].

Figure 17 shows schematically a plant for producing NP base solutions (tank reactor process).

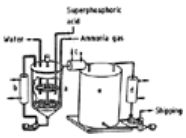

Figure 17. Plant for producing NP solution from superphosphoric acid [77]

a) Reactor; b, c, and d) Coolers; e) Storage tank

Superphosphoric acid, usually containing 40 to 50 % of the P_2O_5 as polyphosphates, ammonia, and water are continuously fed to a stirred-tank reactor. A heat exchanger operating in closed cycle holds the temperature in the reactor at roughly 70 °C. The solution concentration is controlled through density measurement, and the N : P_2O_5 ratio is controlled by means of the pH. The NP solution outlet from the reactor is cooled to 35 °C in another heat exchanger (c) before being transferred to a storage tank. To minimize hydrolysis of polyphosphate to orthophosphate, the product is further cooled to ca. 20 °C and the storage tank is protected from sunlight. The polyphosphate content of the product is always slightly lower than that of the feed acid. For example, a base solution containing about 50 % polyphosphate, which is about the minimum polyphosphate content of the product with acceptable storage properties, requires a feed acid containing 55 % polyphosphate.

Because of the expensive heat-exchange problems — heat must be supplied in the production of superphosphoric acid and removed in the subsequent ammoniation step — other techniques have been developed for producing ammonium polyphosphate solution from orthophosphoric acid in a *direct process*, thereby largely eliminating the expensive heat exchange. Two examples are the Swift process and the Tennessee Valley Authority (TVA) process.

In the Swift process (Swift Agricultural Chemical Corporation), ammonia gas reacts in a jet reactor (Fig. 18) with preheated wet-process phosphoric acid to produce ammonium polyphosphate. The phosphate melt inlet to the mixing tank is converted to liquid fertilizer by the addition of water; the grade produced is usually 10–34–0. The extent of polymerization depends on the temperature that prevails in the process reactor. At 300 °C, it is about 60 %. In addition to clear 10–34–0 liquid fertilizer, a 12–40–0 suspension can also be obtained with this process [94].

The TVA process uses a T reactor (Fig. 19) to produce ammonium polyphosphate. Wet-process phosphoric acid (52 – 54 wt % P_2O_5) is preheated and then

partly neutralized by ammonia in a simple spray washer (spray tower). The solution is then fed into the T reactor, which is also supplied with heated anhydrous ammonia from the other side. At the adjusted temperature of 232 – 243 °C, the reaction with ammonia goes to completion while the residual water is simultaneously vaporized and a foamy polyphosphate melt is formed [95]. After the product passes through a rotary separator (disengager), excess ammonia and steam are removed; the ammonia is scrubbed out, and the steam is discharged. From the melt, with a polyphosphate content of roughly 50 %, either clear liquid fertilizers such as 10–34–0 or suspensions such as 10–37–0 can be produced [95], [96].

The T reactor process has enabled reasonable equipment costs, process simplicity, and efficiency in the production of high-quality liquid fertilizers. The unique feature of the process is the conversion of a low-polyphosphate (20 – 35 %) superphosphoric acid to an high-polyphosphate (ca. 80 %) 10–34–0 or 11–37–0 base solution. The high polyphosphate content substantially improves the stability, storage life, and handling properties of the base solutions and NPK liquids. In contrast to the tank-reactor process the reaction is accomplished in a limited volume of the pipe reactor with minimal heat losses [97].

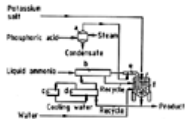

Figure 18. Swift process for producing NP (or NPK) solution from orthophosphoric acid [94]

a) Acid heater; b) Ammonia vaporizer; c) Cooling tower; d) Cooler; e) Pipe reactor; f) Mixing tank

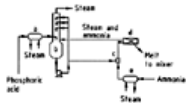

Figure 19. TVA process for producing NP solution from orthophosphoric acid [94]

a) Acid heater; b) Spray tower; c) T reactor; d) Rotating disengager; e) Ammonia preheater

In the Ugine Kuhlmann process [98], clear NP liquid fertilizers with good storage properties are produced from *unpurified wet-process phosphoric acid* [90].

The crude acid is first treated with urea to yield urea phosphate, which is crystallized and separated. The mineral impurities remaining in the mother liquor can be used in the production of solid fertilizers. Thus there is no waste disposal problem. In the second reaction step, the urea phosphate is reacted with ammonia, yielding a urea – ammonium polyphosphate melt, which is dissolved in water to obtain a 16–30–0 liquid fertilizer.

3.2.2.2. *NPK liquids*

Liquid NPK fertilizers are produced by adding a potassium component, usually potassium chloride, to UAN and NP solutions. The amount of water needed to dissolve 100 g of a mixture of UAN, NP solution, and potassium chloride at a specific temperature, as well as the solid phase, can be shown conveniently in the form of three-component phase diagrams [80]. Other triangular diagrams show the solubility relationships in mixed liquid fertilizer systems [97]. While the clear NP solutions with polyphosphate have compositions corresponding to those of solid fertilizers, the contents must be kept lower in clear ternary NPK solutions because potassium nitrate, which is formed from ammonium nitrate and potassium chloride, has a low solubility.

Table 17 presents several three-component liquid fertilizers obtained by mixing a 10–34–0 base solution with potassium chloride powder. The added component lowers the nutrient content from almost 39 % to ca. 29 – 30 % [99].

Table 17. Three-component liquid fertilizers produced by mixing 10–34–0 base solution with KCl [99]

Fertilizer type	Salt composition, wt %						Total nutrient analysis, wt % *	Freezing point, °C
	Ammonium orthophosphate	Ammonium polyphosphate	Urea	NH_4NO_3	KCl	H_2O		
1 : 1 : 1	7.93	10.11	7.8	16.3	15.6	42.2	28.85	–21
1 : 2 : 2	9.20	12.4	5.0	10.4	18.7	44.4	29.54	–18.5
1 : 1.5 : 1	10.70	13.65	7.3	15.3	14.1	38.4	30.57	–23.5
2 : 1 : 1	5.75	7.33	11.9	24.8	10.7	39.5	27.89	–24.5
1 : 1 : 0	16.10	20.5	15.9	33.2	0	14.3	38.60	–33

* $N + P_2O_5 + K_2O$.

Production takes place in either cold-mix or hot-mix plants [100], [312, p. 456]. Because cold-mix plants offer low investment costs and are easy to operate, they are in more widespread use. In a stirred vessel, the appropriate quantities of potassium chloride are added to the base solutions (e.g., 28–0-0 and 10–34–0) and mixed. Hot-mix plants, especially small-capacity ones, require relatively large investments and, because of their troublesome operation, must have well-trained personnel.

3.2.2.3. *UAS Liquids*

The use of high-analysis fertilizers containing little or no sulfur has led to soil sulfur deficiencies in areas far from industries. The critical level for sulfur in soil is 0.1 – 0.3 wt %. Sulfur deficiencies in the soil of Kuwait led to the development of a sulfur-containing liquid fertilizer: 21.2 wt % $(NH_4)_2SO_4$ + 37.6 wt % urea + 41.2 wt % water [101]. The total nitrogen content is 22 wt %, and the saturation temperature is 4 °C.

The reaction of urea with sulfuric acid

$$(NH_2)_2CO + H_2SO_4 \rightarrow (NH_2)_2CO \cdot H_2SO_4$$

yields liquid fertilizers with pH = 1. Because of corrosion, the plant must have polyethylene linings or else use expensive chrome – nickel – molybdenum steel. The following are typical UAS solutions [102]:

Sulfuric acid, wt %	N, wt %	S, wt %
28	9	27
17	16	49
15	17	52
10	18	55

Liquid fertilizers can be applied even in winter if ammonium thiosulfate is added in order to lower the salting-out temperature to – 18 °C. From the standpoint of nutrient content, this compound is better than water. Ammonium thiosulfate itself is a clear liquid fertilizer (12–0-0 + 26S) with a salting-out temperature of – 5 °

C [102].

3.2.3. *Suspensions*

Suspension fertilizers contain solid nutrient salts, especially potassium chloride, suspended in a concentrated ammonium phosphate solution. Their development has made it possible to attain nutrient contents corresponding to those of solid NPK fertilizers:

Solutions	Suspensions
8–8-8	15–15–15
3–9-9	7–21–21
2–16–12	5–15–30
10–5-5	20–10–10

What is more, suspension fertilizers have the advantage that low-priced phosphates that are difficult to dissolve in water can also be employed.

The storage and application of suspension fertilizers, however, is more difficult than that of clear liquid fertilizers, since the former tend to undergo sedimentation. If an easily handled suspension is desired, the product must be stirred vigorously before use. Swellable clays such as palygorskite (attapulgite), a magnesium silicate-containing clay (⟶ Clays), have proved suitable as suspending agents. Suspensions thus make special demands on technique. The methods of storage, transportation, and spraying cannot be transferred directly from clear liquid fertilizer solutions. As a rule, substantial changes in equipment are needed, and these require additional investments. The quality of spray distribution does not match that for clear fertilizer solutions either. For all these reasons, suspensions have scarcely become established in Europe, despite efforts to introduce them. In the United States, on the other hand, a larger scale of farming, together with a well-developed distribution service sector, has allowed the successful introduction of suspensions.

A suspension fertilizer can be produced from *solid monoammonium phosphate* by intensive stirring in water and simultaneous adding of ammonia. The addition of swellable clay (1 – 2 %) ensures a fairly stable suspension. By dilution, a product of grade 9–27–0 is obtained. Applied in a comparable way, it behaves much like the commercial 10–30–0 liquid fertilizer [103].

Anhydrous ammonia is used to improve the solubility and mixing of monoammonium phosphate. Ammonium orthophosphate displays the highest solubility at a molar ratio $NH_3 : H_3PO_4$ of 1.5. A turbine mixer is suitable for homogenizing the suspension obtained by the addition of KCl and 1 or 2 wt % palygorskite [104].

A stable suspension is obtained by ammoniating *wet-process orthophosphoric acid* (54 wt % P_2O_5) at 71 °C and a pH between 7.5 and 8.5. The product is then adjusted to the desired pH with a small amount of acid. Because of the fineness of the solid particles, the suspension (8–24–0) remains stable. After cooling, it can be blended with urea – ammonium nitrate solutions (32 % N) and potassium chloride to produce various compositions of liquid fertilizers [105].

A two-step process developed by TVA yields low-priced, high-analysis suspension fertilizers in which the phosphate component is obtained by ammoniating orthophosphoric acid (Fig. 20). In the first reaction step, the acid is partly ammoniated with ammonia, and the solution is held for about 30 min at 106 °C. In the second step, ammoniation is completed. Then the product is cooled to 60 °C, and the suspension of fine diammonium phosphate crystals is stabilized with 1.5 wt % clay. A fluid NPK fertilizer of 15–15–15 grade can be produced from the 12–40–0 suspension [106]. Raising the acid concentration from 54 wt % to 70 – 73 wt % P_2O_5 (wet-process superphosphoric acid) gives useful suspension fertilizers. In general, however, the Mg content is so high that prolonged storage is excluded because of the precipitation of struvite, $MgNH_4PO_4 \cdot 6\ H_2O$.

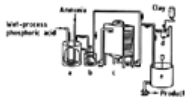

Figure 20. Production of orthophosphate suspensions

a) First reaction stage; b) Second reaction stage; c) Cooler; d) Slurry vessel for suspending agent; e) Product tank

This difficulty is avoided if *triple superphosphate* is employed as the sole phosphate source. Palygorskite is slurried with water in a mixing vessel, aqueous ammonia and urea – ammonium nitrate solution are added, and then finely-ground triple superphosphate and potassium chloride are incorporated. The suspensions made by this process have a typical grade of 12–12–12 [107]. The addition of palygorskite can be omitted if part of the triple superphosphate is first treated with aqueous ammonia to yield a voluminous precipitate, which then acts as a suspending agent for the other solid particles [108].

Suspensions from the Nitrophosphate Process. A low-cost process for producing a suspension fertilizer involves digesting crude phosphate with nitric acid. The slurry of phosphoric acid and calcium nitrate is inlet to a hot-mix plant and ammoniated to form an NP suspension fertilizer. Such processes have been developed in the United States, the USSR, and the Federal Republic of Germany [81], [109], [110]. Aside from the cheap raw material, the fixed costs of the fertilizer are relatively high. Since nitrophosphate processes yield suspension fertilizers with large amounts of calcium, the dilution effect of the calcium holds the nutrient content relatively low (e.g., 9–9-9 or 6–12–12).

High-Nitrogen Suspension Fertilizer. TVA has devised a promising method for producing a high-nitrogen suspension fertilizer (36 wt % N) [102]. This product is a mixture of microspray-crystallized urea and ammonium nitrate solution (76 wt % AN) with palygorskite. The composition is 57.5 wt % urea + 31.5 wt % ammonium nitrate + 10.0 wt % water + 1 wt % palygorskite [105], [108], [111].

Sulfur-Containing Suspensions. In regions where the effects of sulfur deficiency in the soil are pronounced in certain regions, sulfur in fertilizers not only increases yield but also can improve product quality. TVA has manufactured a 29–0–0–5S urea ammonium sulfate (UAS) suspension fertilizer in a pilot plant. The feedstocks are urea (70 % aqueous solution), sulfuric acid (93 %), anhydrous ammonia, and water. Clay (ca. 2 %) is added to the UAS solution to give the final product [102].

[Top of Page]

4. Special Fertilizers

Manfred Drach, Ralf Vosskamp, Martin E. Trenkel, Reinhold Gutser, Günter Steffens

The special fertilizers include water-soluble nutrient salts, foliar fertilizers, micronutrient fertilizers, slow- and controlled-release fertilizers, nitrification inhibitors, and organic fertilizers, including sewage sludge, compost and manure.

4.1. Water-Soluble Nutrient Salts

Especially for fertilization in commercial horticulture, but also for specialty field crops, water-soluble nutrient salts (e.g., ammonium nitrate, potassium nitrate, ammonium phosphate, magnesium sulfate) are preferably used with water-saving irrigation systems (drip irrigation/fertigation) and with soilless cultivation systems. Various crop-specific fertilizer recipes can be formulated with these salts. A convenient form for the consumer are multinutrient fertilizers made by mixing various water-soluble salts, which generally also include micronutrients. Such salt mixtures are available worldwide from most major fertilizer manufacturers.

From these salts and salt mixtures, the grower prepares highly concentrated master solutions, which for application are further diluted to render them plant-compatible and then fed into the irrigation system. By using appropriate automatic fertilizer systems it can be assured that irrigation intervals and concentration of the nutrient solution are adjusted to the crop's needs.

A distiction is made between urea-containing and urea-free nutrient salt mixtures. Since the application rate of nutrient salts is frequently monitored by means of conductivity, products containing urea cause problems as urea has no ionic conductivity and is therefore not detected. Therefore, products containing urea are mainly used for foliar fertilization (see Section Foliar Fertilizers) where these calibration problems do not exist and because urea nitrogen is the N form with the best uptake rate via the leaves.

A prerequisite for all nutrient salts is rapid dissolution in water without any residues. This is achieved by producing the salt mixtures from appropriate ingredients and by properly grinding them. Moreover, for special fields of application in horticulture (flood irrigation and soilless cultivation systems) they are rated by their nitrogen forms. For these fields of application, nitrate-based nutrient salts are preferred.

In principle, if appropriately diluted, all nutrient salts are suitable both for soil fertilization and for foliar fertilization.

4.2. Foliar Fertilizers

For over 100 years it has been known that nutrients can be taken up through the leaves. However, with some exceptions, plants cannot be provided with nutrients through the leaves alone, since even leafy plants such as potatoes cannot absorb sufficient nutrients through the leaves. Foliar fertilization, therefore, is usually a supplement to soil fertilization. However, it is often the case that proper foliar fertilization puts the plants in the position to utilize soil-derived nutrients better. Recent decades have seen foliar fertilizers come into use not only for special crops or in certain localities, but throughout agriculture.

Foliar fertilizers are substances that contain primary nutrients and/or micronutrients, are applied to the leaves, and are absorbed into the leaves. The most important foliar fertilizer is urea (46 % N) which is highly soluble in water and rapidly absorbed by plants via the leaves. Therefore, urea is frequently used as a component of fertilizer suspensions and solutions. In addition to nitrogen, other macro- and micronutrients can be added, frequently in a nutrient ratio tailored to the demand of specific target crops. Special raw materials and formulations ensure a good plant compatibility and optimum foliar nutrient uptake. Also used for foliar fertilization are solutions with organic ingredients (e.g., amino acids), readily soluble salts (e.g., potassium nitrate or micronutrients, mostly based on sulfate) and salt mixtures containing macronutrients and special micronutrient mixtures.

4.2.1. *Production*

Solid foliar fertilizers are produced by mixing salts that are readily soluble in water. Before application, the salts are dissolved for spraying. Liquid foliar fertilizers, which are made from readily soluble salts or solutions and adjusted to a given analysis, are also on the market. The production of suspensions for foliar application involves the use of dispersing agents, and in some cases ionic surfactants, to stabilize the spray.

4.2.2. *Application*

In order to achieve optimal utilization of nutrient uptake through the leaves, the nutrient salts must remain as long as possible in the dissolved liquid state on the leaf surface. For this reason, spraying in overcast weather or toward evening (“evening dew”) gives better results than spraying on dry days in bright sunlight. Depending on the sensitivity of the leaves and the concentration of the spray liquid, application in hot, sunny weather may cause burning. Table 18 gives the time required for 50 % absorption of nutrients applied to the leaves of bean plants [113].

Table 18. Nutrient absorption rates [113]

Nutrient	Time for 50 % uptake
Nitrogen	1 – 6 h
Phosphate	2.5 – 6 days
Potash	1 – 4 days
Calcium	1 – 4 days
Sulfur	5 – 10 days
Magnesium	1 h (20 % uptake)
Iron	24 h (8 % uptake)
Manganese	1 – 2 days
Zinc	1 day
Molybdenum	10 – 20 days

Nutrients placed on the leaf surface follow three paths to the leaf cells [114-116]. After the cuticle has been wet by the nutrient spray, the cutin swells, which increases the distances between the wax plates in the cuticle. As a result, the nutrients are able to diffuse through the cuticle and into the cell wall (Fig. 21). The nutrients can then diffuse into the space around the epidermal cell and to the cell membranes, where they can be absorbed, or they can diffuse further along the cell wall of the epidermal cell and adjoining cells, deeper into the leaf, where they can be absorbed by a parenchymal cell. Nutrients that have been absorbed by an epidermal cell can be transferred to other cells by cytoplasmic strands connecting the cells, the plasmodesmata, a second path. If surfactants are included in the nutrient spray, a third path becomes available: the nutrients can pass through the stomata into the air spaces of the leaf.

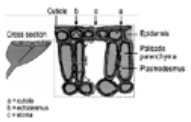

Figure 21. Access and transport paths for nutrients in leaves

Reasons for Using Foliar Fertilizers. Foliar fertilizers are applied to remedy obvious nutrient deficiencies. If a deficiency is recognized, the missing nutrient is supplied by spraying. If the deficiency is not well defined, complete foliar fertilizers or a mixed micronutrient fertilizer can be employed.

Another reason for foliar application is to remedy hidden (latent) nutrient deficiencies, which can seriously impair crop yield and quality.

Foliar fertilizers are also used to protect yield and quality, to meet peak nutrient demands (i.e., remedy latent deficiencies and stimulate the plants to absorb more nutrients from the soil), and to improve quality. The last point is especially important for special crops (wine, fruits, vegetables).

Advantages and Drawbacks of Foliar Fertilizers. Foliar application has a number of advantages over soil application: Nutrients act rapidly, nutrient utilization is high, nutrients are placed where they are needed, the risk of leaching losses and groundwater contamination is reduced. micronutrients are supplied to meet much of the demand for them, micronutrients are not fixed in the soil, application does not entail extra costs if fertilizers are combined with agricultural pesticides, and some nutritional problems can be cleared up only through application to the leaves or fruit (bitter pit in apples, stem necrosis in grapes).

The main disadvantages of foliar fertilization are that the action time is often short, the quantity of primary nutrients delivered is relatively small, and there are questions of compatibility in relation to concentration and prevailing weather conditions.

4.2.3. *Combination with Agricultural Pesticides*

Foliar fertilizers are usually applied along with agricultural pesticides for the following reasons [117]: to lower the cost of application, since pesticide spraying is required anyway; to reduce the stress due to the application of pesticide (e.g., certain herbicides); to improve the quality of pesticide spray liquids (lower the pH or increase the electrolyte content) and thus stabilize them; to reduce evaporation at low air humidity, so that most of the fine droplets reach the plant in liquid form; and to lower the surface tension of the water.

4.3. Micronutrients

The most important micronutrients (trace elements) are iron, manganese, boron, zinc, copper, and molybdenum. The delivery of micronutrients to plants is just as indispensable today as that of macronutrients or primary nutrients. Diseases resulting from an undersupply of micronutrients are seen more and more often as agriculture and gardening become more intensive, with a corresponding increase in plant nutrition [118]. For this reason, better and more specific micronutrient fertilizers have now been developed.

The term micronutrient also signifies that these substances are needed only in small quantities, often just 1 – 100 g per hectare and year. These small amounts are, however, absolutely necessary. But if the optimal dose is much exceeded, serious crop damage and loss of yield may occur. For boron, in particular, the difference between an adequate and an excessive amount is slight.

To work in the proper way, the micronutrients must be absorbed by the plants. Uptake is mainly in the form of ions: cations or metal – chelate ions of iron, manganese, copper, and zinc and anions in the case of boron and molybdenum (borates, molybdates). Absorption is through the leaves (foliar fertilizer) or the roots (soil solution).

4.3.1. *Micronutrient Forms*

Micronutrients are used in the following inorganic forms:

1. Various metal salts (sulfates, chlorides, nitrates)
2. Milled oxides
3. Metal flours and metal slags

Table 19 lists only those micronutrient compounds that are in common use. The table also states the nutrient analyses, the solubilities in water, and the various types of fertilizers in which the compounds appear.

Table 19. Micronutrient fertilizers

Substances and products	Formula	Analysis, wt %	Solubility, g/100 g H_2O	Uses
Salts (*technical*)				
Iron(II) sulfate heptahydrate	$FeSO_4 \cdot 7H_2O$	19.5	62.3	liquid fertilizers,
Manganese(II) sulfate monohydrate	$MnSO_4 \cdot H_2O$	32	63.0	nutrient-salt
Copper(II) sulfate pentahydrate	$CuSO_4 \cdot 5H_2O$	25	36.0	mixtures, solid
Zinc(II) sulfate monohydrate	$ZnSO_4 \cdot H_2O$	35.5	63.7	straight and
Zinc(II) sulfate heptahydrate	$ZnSO_4 \cdot 7H_2O$	22	165.5	multinutrient
Boric acid	H_3BO_3	17		fertilizers,
Colemanite (boron mineral)	$Ca_2B_6O_{11} \cdot 5H_2O$	11		micronutrient and
Disodium octaborate tetrahydrate *	$Na_2B_8O_{13} \cdot 4H_2O$	20.8	9.5	multimicronutrient
Hexammonium heptamolybdate tetrahydrate	$(NH_4)_6Mo_7O_{24} \cdot 4H_2O$	54		fertilizers
Oxides				
Manganese dioxide (mineral)	MnO_2	45		solid straight and
Zinc oxide (residual zinc)	ZnO	69		multinutirent
Molybdenum trioxide	MoO_3	60		fertilizers
Slags				
Copper slag flour		2.5 Cu, 0.05 Co		micronutrient fertilizers
Metal flours				
Manganese flour		20 Mn, 10 Fe, 0.5 Zn, 0.25 Cu		micronutrient
Copper flour		2.6 – 25 Cu		fertilizers

* Disodium octaborate is not a homogeneous substance. It has a $Na_2O : B_2O_3$ ratio of 1 : 4 and can be produced as a concentrated solution. It is sold by US Borax under the trade name Solubor.

In addition to these, the nitrates and chlorides of iron, manganese, copper, and zinc are also used, especially in concentrated liquid fertilizers. The trace-element nitrates and chlorides have high solubilities (e.g., 268 g per 100 g H_2O for $Fe(NO_3)_2 \cdot 6\,H_2O$ or 400 g per 100 g H_2O for $FeCl_3 \cdot 6\,H_2O$; some of the nitrates and chlorides of manganese, copper, and zinc have still higher values). Because of the price and the handling qualities in the solid state, however, the sulfates are preferred. Many crops are sensitive to trace-element chlorides in large amounts. At the levels usual in foliar application, however, this is certainly not a major problem. Trace-element sulfates and nitrates provide two nutrient elements each.

Because soluble iron compounds are fixed as phosphates or oxides in calcareous soils, organic complexes were first developed for this element. One successful complexing agent is ethylenediamine di(*o*-hydroxyphenylacetic acid) (EDDHA):

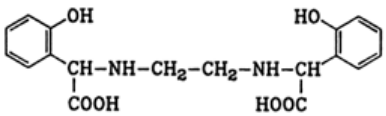

Other compounds of trace metals with *N*-carboxylalkylamino acids are less effective in the soil, especially with soil pH values > 7, but if applied on the leaves are superior to the EDDHA compounds and to the simple salt solutions.

The *N*-carboxylalkylamino acids have also come into wide use today as chelating agents for manganese, copper, and zinc:

EDTA. ethylenediaminetetraacetic acid
HEDTA. hydroxyethylethylenediaminetriacetic acid
DTPA. diethylenetriaminepentaacetic acid
EDDHA. ethylenediamine di(o-hydroxyphenylacetic acid)
EDDHMA. ethylenediamine di(*o*-hydroxy-*p*-methylphenylacetic acid)

For the formulas of the *N*-carboxylalkylamino acids, see ⟶ Ethylenediaminetetraacetic Acid and Related Chelating Agents, and the stability constants, see ⟶ Ethylenediaminetetraacetic Acid and Related Chelating Agents

Other chelating agents such as lignosulfonates and citric acid have so far been used only to a lesser extent.

4.3.2. *Production*

The list of fertilizers containing micronutrients comprises straight micronutrient fertilizers, combination micronutrient fertilizers, and ordinary fertilizers with micronutrients. All these types can be manufactured in either liquid or solid form. Among the solid micronutrient fertilizers, the powdered and fine-crystalline ones predominate, since nearly all of them can be dissolved in water and applied through the leaves. They generally consist of the several substances in Table 1 as well as the *N*-carboxyalkylamino acid chelates cited.

Solid micronutrient fertilizers for soil application include metal flours, metal slags, and micronutrient frits. Metal flours come from fabrication waste in the manganese production and processing industry and from specially prepared alloys. The source material is ground and granulated to between 0.25 and 1.5 mm. Metal slags such as copper slag flour are also won from waste products. The metal slags are quenched in water and comminuted to fine, powdery products.

Frits are made by melting glasses and incorporating the desired micronutrients (single-element and multi-element frits). The glasses are then finely milled [119].

Solid straight micronutrient fertilizers and multi-micronutrient fertilizers with complexing can be produced in various ways. One technique is to grind a metal sulfate with chelating agents, or several metal sulfates with chelating agents (in many cases borates, molybdates, and magnesium sulfate are added) and blending [120]. Chelation takes place when the material is dissolved in water before use. A second approach is to stir and thereby dissolve chelating acids and metal oxides (including magnesium oxide) in water at elevated temperatures; additives that cannot be chelated, such as borates and molybdates, are added; the product is further stirred, filtered, and spray-dried [121] or crystallized.

Liquid micronutrient fertilizers and combination micronutrient fertilizers can be prepared by the same technique but without the spray-drying step.

In the production of fertilizers with micronutrients, the micronutrients are usually added before granulation; in the case of liquid fertilizers, micronutrient compounds or chelates are dissolved along with the primary nutrient forms.

4.3.3. *Commercial Fertilizers*

Several firms in Western Europe manufacture pure micronutrients or combination micronutrient fertilizers, most of them in chelate form. Table 20 lists selected firms and examples of their products.

Table 20. Commercial fertilizers

Company	Products
ABM Chemical	Nevanaid Fe, finely powdered solid product with 9 % Fe (HEDTA)
(United Kingdom)	Nevanaid Fe, finely powdered solid product with 7 % Fe (DTPA)
	Nevanaid Mn, finely powdered solid product with 10 % Mn (EDTA)
Allied Colloids	Librel Fe-Dp, finely powdered solid product with 7 % Fe (DTPA)
(United Kingdom)	Librel Mn, finely powdered solid product with 13 % Mn (EDTA)
BASF	Fetrilon 13 %, spray-dried solid fertilizer with 13 % Fe (EDTA)
(Germany)	Fetrilon-Combi, spray-dried solid multimicronutrient fertilizer with 9 % MgO, 4 % Fe, 4 % Mn, 1.5 % Cu, 1.5 % Zn, 0.5 % B, 0.1 % Mo (all heavy metals as EDTA, boron as borate, Mo as molybdate)
	Nutribor, crystalline powder with 8 % B, 1 % Mn, 0.1 % Zn, 0.04 % Mo, 6 % N, 5 % MgO, 12 % S (EDTA)
	Solubor DF, microgranular with 17.5 % B
Ciba Geigy	Sequestrene 138, solid fertilizer with 6 % Fe (EDDHA), soil application
(Switzerland)	Ferrogan 330, solid fertilizer with 10 % Fe (DTPA)
Grace Rexoline	Rexene 224 Fe, powdered or granular fertilizer with 6 – 7 % Fe (EDDHMA), soil application
(Sweden)	Rexenol Cu, powdered solid fertilizer with 9 % Cu (HEDTA)
	Rexene Zn, liquid fertilizer with 6.5 % Zn (EDTA)

Urania (Germany) uses metal slags as raw material to produce granular copper fertilizers containing 2.5 % and 5 % Cu, as well as a copper – kieserite granular fertilizer with 2.5 % Cu and 21 % MgO. Similar micronutrient products and the same chelating agents are made in the United States and marketed as commercial fertilizers [119, pp. B68 to B83].

There are also fertilizers with micronutrient supplements, namely: granulated straight and multinutrient fertilizers, water-soluble nutrient salts, suspensions and solutions with micronutrients, which are chosen and added according to crop demand.

4.3.4. *Use*

In order to combat micronutrient deficiencies, either fertilization far exceeding actual depletion or directed foliar fertilization must be carried out [117, p. 128]. The annual depletion of micronutrients can be partly offset by the use of macronutrient fertilizers with micronutrients. On the basis of average micronutrient contents and average crops (rotation 70 % cereals, 20 % root crops, 10 % fodder), the following are withdrawn from the topsoil per hectare and year in central Europe [117, p. 122]: 400 – 700 g manganese, 260 – 400 g zinc, 150 – 200 g boron, 80 – 120 g copper. The amounts for grassland (per hectare and year) are 800 g manganese, 300 g zinc, 80 g copper, 50 g boron, and 0.5 g cobalt. In addition, leaching removes an average of 250 g manganese, 250 g boron, 100 g zinc, and 30 g copper per hectare and year [117, p. 122].

Normally, fertilization must more than make up for this consumption. Per hectare, boron-deficient soils must receive 1 – 2 kg of boron, and manganese-deficient soils must receive as much as 12 – 24 kg of manganese.

In order to prevent the development of latent deficiencies, micronutrients are often used in small quantities, especially for foliar application. This applies in particular to intensive cereal farming, where — even on well-supplied soils — the freely available micronutrients are soon consumed or else the nutrient flux at peak demand times cannot be covered from soil reserves. If relatively small amounts of a multi-micronutrient fertilizer are used in specific treatments during tillering and in the shoot and ear phases, latent deficiencies are generally eliminated and the yield potential is fully utilized [122], [123].

4.4. Slow- and Controlled-Release Fertilizers

4.4.1. *Introduction*

In plant nutrition, soil and plants are two antagonistic systems that compete for the nutrients available in or applied to the soil. This competition is the main reason why only a portion of nutrients is taken up and used by the plants and crops grown, while another portion is (temporarily) immobilized in the soil or lost by denitrification/volatilization and leaching (particularly of nitrogen) [124, p. 124]. The fertilizer industry has developed special types of fertilizers and fertilizer

modifications which avoid or at least reduce such losses, such as:

- Foliar fertilizers
- Slow and controlled-release fertilizers
- Nitrification and urease inhibitors; stabilized fertilizers

The utilization rate of nutrients is improved considerably by leaf application. However, in practice it is impossible to supply all the necessary nutrients via plant leaves [124, pp. 125 – 127], [125, pp. 14 – 16]. A more practical route is the use of nitrogen fertilizers which release the nutrients according to the plants' requirements, that is, slow- and controlled-release fertilizers.

Terminology. The Association of American Plant Food Control Officials (AAPFCO) gives the following definition: “Delay of initial availability or extended time of continued availability may occur by a variety of mechanisms. These include controlled water solubility of the material (by semipermeable coatings, occlusion, or by inherent water insolubility of polymers, natural nitrogenous organics, protein materials, or other chemical forms), by slow hydrolysis of water-soluble low molecular weight compounds, or by other unknown means” [126].

There is no official differentiation between slow-release and controlled-release fertilizers. However, the microbially degradable N products, such as urea – formaldehydes (UFs and other urea – aldehyde compositions), are commonly referred to in the trade as slow-release fertilizers, and coated or encapsulated products as controlled-release fertilizers [127, p. 12].

Advantages and Disadvantages. Slow- and controlled-release fertilizers reduce toxicity, especially in use with seedlings. The toxicity of conventional soluble fertilizers is caused by the high ionic concentrations resulting from quick dissolution. Consequently, these slow and controlled-release fertilizers permit the application of substantially larger fertilizer dressings (depot fertilization). This results in significant savings in labor, time, and energy. They also allow the full nutrient requirements of crops grown under plastic cover to be met (protected crop cultivation). These fertilizers significantly reduce possible losses of nutrients due to the gradual nutrient release (particularly losses of nitrate nitrogen). They also reduce evaporative losses of ammonia. They further contribute to a reduction in environmentally relevant gas emissions (N_2O) [127, pp. 15 – 16].

There are no standardized methods for reliable determination of the nutrient release pattern available as yet due to the lack of correlation between laboratory tests and field conditions. With urea – formaldehyde fertilizers, a proportion of the nitrogen content may be released extremely slowly or not at all. With sulfur-coated controlled-release fertilizers the initial nutrient release may be too rapid and cause damage to turf or the crop. Repeated use of sulfur-coated urea may also increase the acidity of the soil. Polymer-coated or encapsulated controlled-release fertilizers can cause an environmental problem since undesirable residues of the coating material may accumulate in the fields. However, the main disadvantage is that the cost of manufacturing slow- and controlled-release fertilizers is still considerably higher than that of conventional mineral fertilizers. At present their cost/benefit ratio prevents their wider use in general agriculture. Consequently, the vast majority is applied in nonagricultural sectors such as nurseries and greenhouses, golf courses, professional lawn care, as well as by consumers (home and garden) and landscape gardeners [127, pp. 17 – 18]. Total world consumption of slow- and controlled-release fertilizers is estimated at 562 000 t (1995/96), amounting to only 0.15 % of world total consumption of N + P_2O_5 + K_2O in the form of fertilizer material (ca. 380 × 10^6 t) [128].

Types of Slow- and Controlled-Release Fertilizers. The two most important groups are:

1. Condensation products of urea such as urea – formaldehydes (slow-release)
2. Coated or encapsulated fertilizers (controlled-release)

Of lesser importance are other organic chemicals, ion-exchange materials, and supergranules.

4.4.2. *Urea – Aldehyde Slow-Release Fertilizers*

Three types of urea – aldehyde condensation products (see also ⟶ Urea – Reaction of Urea and Its Derivatives with Aldehydes) have gained practical importance:

1. Urea – formaldehyde (UF)
2. Urea – isobutyraldehyde (IBDU/Isodur)
3. Urea – acetaldehyde/crotonaldehyde (CDU/Crotodur)

At pH values below 2, crotonaldehyde and acetaldehyde can form cyclic condensation products [131]. For production of urea – aldehyde condensates, see [129, pp. 3 – 87], [130, pp. 1 – 137], [132, pp. 153 – 156], [133], [134, pp. 247 – 279].

The urea – formaldehyde products have the largest share of the slow- and controlled-release fertilizer market (40 % of world consumption in 1995/96); IBDU- and CDU-based products are less widely used (15 % in 1995/96), since their manufacturing costs are even higher than that of urea – formaldehydes [127, pp. 61 – 63].

4.4.2.1. *Urea – Formaldehyde Condensation Products.*

Ureaform, as defined by the American Association of Plant Food Control Officials (AAPFCO) is the oldest type of urea – formaldehyde condensate. As early as 1924, Badische Anilin- & Soda-Fabrik (now BASF) registered the first patent (DRP 431 585) on urea – formaldehyde condensation fertilizers [135]. In the United States they were patented for use as fertilizers in 1947; commercial production began in 1955.

Ureaforms are a mixture of methylene – urea oligomers of various molecular masses, polymer chain lengths, and hence varying water solubilities, such as methylene diurea (MDU), and dimethylene triurea (DMTU). They also contain a certain amount of unchanged urea.

Manufacture. The manufacture of urea – formaldehyde products is a two step process [136, p.13], [137], [138], [139], [140], [141], [142], [143]:

$$NH_2CONH_2 + CH_2O \rightleftharpoons HOCH_2NHCONH_2$$

$$CH_2O + HOCH_2NHCONH_2 \rightleftharpoons HOCH_2NHCONHCH_2HO$$

First, urea and formaldehyde are combined to give the intermediates monomethylol- [1000-82-4] and dimethylolurea [140-95-4]. Under acidic conditions these methylolureas react with further urea to give various oligomers of methylene urea:

$$NH_2CONH_2 + HOCH_2NHCONH_2 \rightarrow NH_2CONHCH_2NHCONH_2 + H_2O$$

In the production of granular urea – formaldehyde products, water must be removed by evaporation. The main problem in the manufacture of urea – formaldehyde slow-release fertilizers is the production of condensation oligomers in the desired proportions. A number of processes can be used to meet this target (e.g., dilute- and concentrated-solution processes). The procedure commonly applied is to use suspensions of methylene urea or solutions of urea and methylol urea as well as solid product. In this way the condensation reactions take place in the granulator itself (in situ process) [144].

Properties. The urea – formaldehyde products are separated into the following three fractions:

1. Fraction I: cold water (25 °C) soluble (CWS), containing residual urea, methylene diurea (MDU), dimethylene triureas (DMTU), and other soluble reaction products. Depending on soil temperature the availability of Fraction I nitrogen is slow.
2. Fraction II: hot water (100 °C) soluble (HWS), containing methylene ureas of intermediate chain length: slow-acting nitrogen.
3. Fraction III: hot water insoluble (HWI) containing methylene ureas of very long chain length, insoluble in both cold and hot water; extremely slow-acting or ineffective in plant nutrition.

How the proportion of the different methylene ureas affects the release of nitrogen and the nitrogen efficiency are expressed by the activity index (AI). The AI is calculated from the solubility fractions of the fertilizer under various conditions [127, p. 22]. In the past urea – formaldehydes had an AI of about 40 – 50; more recent formulations have AI values of 55 – 65.

In general, the nitrogen content of urea – formaldehyde condensation products ranges from 35 to 42 % N. The American Association of Plant Food Control Officials (AAPFCO) specifies a minimum AI of 40, with at least 60 % of the nitrogen as cold water insoluble nitrogen (CWI N), a total N content of at least 35 % N, and an unreacted urea nitrogen content of less than 15 % of total nitrogen. In the United States, Western Europe, the Former Soviet Union, and Israel research has been carried out to reduce the fraction of HWI nitrogen [148]. In the 1980s research resulted in the development of MDU/DMTU compositions which consist of shorter chain polymers with at least 60 % CWS polymer nitrogen. Although they have higher contents of CWS nitrogen, they still have safer agronomic and environmental properties than conventional nitrogen fertilizers. Commercial products are white, colorless powders or granules. In wet granulation, the pH value and temperature must be controlled to avoid hydrolysis and thus losses of formaldehyde. Under normal conditions the finished products are stable in handling and storage. Typical properties of urea – formaldehyde products are given in Table 21.

Table 21. Properties of typical ureaforms

Property	Nitroform*	Azolon**
Total N, wt %	37.4	38.0
Insoluble in cold water, wt %	26.9	26.0
Soluble in hot water, wt %	15.4	10.4
Activity index	43	60
Bulk density, t/m^3		0.75
Granule size, mm	0.5 – 2.0	1.0 – 4

* Trademark of Nor-AM; BASF product analysis 1980.
** Trademark of Aglukon; analysis from [149].

Application. The release of plant-available nitrogen from urea – formaldehyde products mainly involves decomposition through microbial activity and dissolution by hydrolysis. Consequently, factors affecting microbial activity, such as higher temperature, moisture, pH value and oxygen availability, also affect the release of nitrogen. These products are therefore widely used in warmer climates (in the Mediterranean region of Europe and in the southern and southwestern United States).

4.4.2.2. *Other Urea – Aldehyde Condensation Products*

Due to the higher costs involved in the combination of urea with higher aldehydes, only two products have gained commercial importance. These are IBDU or Isodur (urea + isobutyraldehyde) and CDU or Crotodur (urea + acetaldehyde or crotonaldehyde).

IBDU (Isodur). Products derived through the combination of urea and isobutyraldehyde consist of mainly isobutylidene diurea [6104-30-9] (with small quantities of slow-acting by-products).

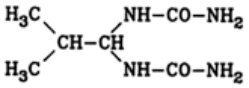

Isobutylidene diurea

Manufacture. Isobutylidene diurea is manufactured by condensation of liquid isobutyraldehyde with urea (either in solution or solid form). In contrast to urea – formaldehyde, the reaction of urea with isobutyraldehyde results in a single oligomer. To obtain an optimal proportion of IBDU it is important to stop the reaction by neutralization when the IBDU yield is at a maximum.

Isobutylidene diurea is produced in Japan and Germany [130], [150], [151], [152] (Mitsubishi, BASF). In the BASF process urea reacts with isobutyraldehyde in an aqueous solution to give a high proportion of slow-release N [153]. A Mitsubishi-developed process is operated by IB Chemicals in Alabama.

Properties and Application. IBDU is a white crystalline solid with a theoretical N content of 32.18 wt % N. The official definition (AAPFCO) requires 30 wt % N, of which 90 % must be cold water insoluble prior to grinding. Since it is almost completely soluble in hot water, the N content is therefore nearly all slow-release N. It has a calculated AI of 90 – 99 (ureaform: 55 – 65). Nitrogen is released from IBDU by hydrolysis, which is affected by soil moisture and temperature. Both urea molecules of IBDU can be liberated. The rate of N release is mainly a function of particle size: the finer the particles, the more rapid the rate of N release. IBDU is unstable in an acid media, whereby it decomposes into the starting materials. Therefore, it tends to release its nitrogen more rapidly in strongly acid soils. This can also be counteracted by using larger, well-compacted granules [158], [159].

The safety margin and agronomic response from IBDU is good with turf; occasional phytotoxity has been observed in greenhouse use. Since it is independent of microbial activity, IBDU is particularly suited to low-temperature application. Properties of IBDU are given in Table 22.

Table 22. Properties of CDU, IBDU, and Isodur

Property	CDU	IBDU	Isodur
mp, °C	245[a]	207 – 208	203 – 204[b]
Bulk density, kg/m^3	ca. 600	600 – 700	500 – 600
Solubility in water at 25 °C, g/L			
fertilizer grade		0.3 – 3.0	2.7
pure substance	1.3 (pH 2); 0.9 (pH 7)		1
Total N, wt %	31	31	31
pH (10 wt % suspension)	4 – 6	5 – 8	6 – 8
LD_{50} (rat, oral), g/kg[b]	10.0		10.0

[a] 259 – 260 °C for crotonylidene diurea (BASF).
[b] 236 °C for isobutylidene diurea (BASF).
[c] Toxicological data from BASF.

CDU (Crotodur). Crotonylidene diurea [1129-42-6] is a mixture of 75 – 80 wt % crotonylidene diurea (2 mol urea + 2 mol acetaldehyde) and 15 – 20 wt % 5-oxyethylcrotonylidene diurea [23048-84-2] (2-oxo-4-methyl-5-oxyethyl-6-ureidohexahydropyridine), 2 mol urea + 3 mol acetaldehyde), [145] and 5 – 7 wt % K_2SO_4.

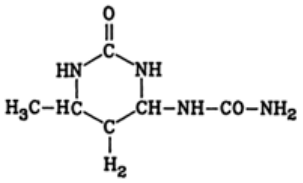

Crotonylidene diurea

Manufacture. Crotodur, which was patented as a slow-release fertilizer in 1959 [153], is produced by acid-catalyzed reaction of urea with acetaldehyde (Chisso Corp., Japan) or crotonaldehyde (BASF, Germany) [154], [155]. The mother liquor is neutralized, and the CDU isolated as a white powder by spray drying or filtration.

Properties (see Table 22). CDU or Crotodur contains 85 wt % to > 90 wt % pure crotonylidene diurea. It is almost completely insoluble in cold water, but soluble in hot water with a calculated AI of 90 – 99. In contrast to isobutylidene diurea, N release from crotonylidene diurea depends on hydrolysis and microbial activity. Only the urea molecule from the side chain (6-position) can be liberated by hydrolysis. The urea which forms part of the ring can only be released by microbial action [146]. The particle size also influences the N release; with large particles, release is strongly delayed.

Application. In Japan and Europe Crotodur is mainly used on turf and in speciality agriculture, either as a straight N fertilizer or in granulated NPK fertilizers.

4.4.2.3. *Further Processing of Urea – Aldehyde Condensates*

All commercial urea – aldehyde condensates such as ureaform, IBDU, Isodur, CDU, and Crotodur can be further processed by compaction or moist granulation with other conventional fertilizers with rapidly available plant nutrients, thus producing straight and compound fertilizers with both slow-release and rapidly available nitrogen. However, it is necessary to keep the pH between 5.4 and 6.2 [147], or preferably between 5.0 and 5.5 [156], the temperature below 90 °C, and the dwell time as well as the quantity of recycled material as low as possible [130, pp. 1 – 15], [156] to prevent hydrolysis and, particularly in the case of ureaform, further condensation to higher oligomers.

Commercial Products and Trade Names. There are a large number of various slow-release fertilizers based on ureaform 38–0-0, methylene ureas 40–0-0, IBDU 31–0-0, and CDU 31–0-0, which are formulated as straight fertilizers and in combination with P, K, and secondary and trace elements in solid form, as well as in solutions or suspensions: Scotts Granuform, Scotts MU-40, ProGrow, ProTurf (The Scotts Company); Nitroform, Nutralene (Omnicology, Inc./AgrEvo); Plantosan, Nitroform, Nutralene, Azolon (Aglukon Spezialdünger). Hydroform, Hydrolene (Hydroagri US). Folocron, CoRoN (CoRoN Corp.); Isodur (Floranid), Crotodur (Triabon) (BASF); Azorit (EniChem); Urea – formaldehyde (Mitsui Toatsu Fertilizers); IBDU (Mitsubishi Kasei); CDU (Chisso).

Analyses of some typical fertilizers containing Isodur or Crotodur are listed in Table 23.

Table 23. Analysis and physical properties of typical fertilizers containing Isodur or Crotodur *

Property	Floranid permanent 15 + 9 + 15 (+ 2)**	Floranid N 32**	Triabon 16 + 8 + 12 (+ 4)**
Nitrogen content, wt %			
Total	15	32	16
IBDU	5	29	
CDU			11.0
Ammonium	5.8		4.0
Nitrate	4.2		
Carbamide		3.0	1.0
Phosphate (P_2O_5)	9		8
Potassium (K_2O)	15		12
Magnesium (MgO)	2		4
Granule size***	0.7 – 2.8	0.6 – 2.6	1.0 – 3.5
Bulk density, kg/m^3	960	600	820

*Floranid and Triabon are trademarks of BASF.
**Floranid Permanent and Triabon also contain trace elements.
***90 % within given range [157].

4.4.3. *Other Organic Chemicals*

Some other organic compounds such as oxamide, triazones, and melamine have also been used as slow-release fertilizers but have not obtained the commercial importance of urea – aldehyde products.

Oxamide [471-46-5], the diamide of oxalic acid, N content 31.8 wt %, M_r 88.08, *mp* 419 °C, ϱ^{20} 1.667 g/cm^3, is a nonhygroscopic, colorless compound that forms needle-shaped crystals.

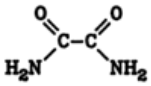

Its solubility in water is only 0.4 g/L [41, p. 320]. By hydrolysis in the soil, it is transformed first into oxamic acid with liberation of ammonia, and then into oxalic acid. Oxalic acid is toxic to plants if it is not further converted into carbon dioxide by microbial activity. Because the nitrogen is released by hydrolysis, the slow-release effect is primarily a function of particle size. Oxamide is manufactured by oxidation of hydrogen cyanide with hydrogen peroxide to give cyanogen, which is then partially hydrolyzed [161]. It can also be produced directly in one step from hydrogen cyanide, oxygen, and water, with copper nitrate as the catalyst [162], [163], [164].

In Japan, production of oxamide as a slow-release fertilizer amounts to about 1000 t/a (1995, Ube Industries) [165].

Symmetrical Triazones. Some symmetrical triazones are used as slow-release nitrogen fertilizers [160], particularly in urea – triazone solutions with 28 % N for application to soil. Triazones are synthesized by condensation reactions of formaldehyde or other aldehydes with urea, organic amines, or ammonia [131, p. 251].

Commercial Products and Trade Names. Hickson Kerley, United States: N-Sure 28–0-0, N-Sure-Lite 30–0-0, Trisert 13–3-4.

Melamine, triaminotriazine [108-78-1], is a high-nitrogen (66 wt % N) crystalline powder. It is produced through heating urea under pressure in the presence of a catalyst. Melamine slow-release fertilizer material is available as a powder (Nitrazine 66 % N). Currently only small amounts are used in the fabrication of slow-release spikes and stakes for shrubs and trees.

4.4.4. *Inorganic Compounds*

Some sparingly soluble inorganic compounds such as metal ammonium phosphates and metal potassium phosphates [130, pp. 256 – 284] are also used as slow-release fertilizers. Their solubility in water at 25 °C is [166], [167]:

$MgNH_4PO_4 \cdot H_2O$	0.014 g/100 mL
$MgNH_4PO_4 \cdot 6H_2O$	0.018 g/100 mL
$MgKPO_4 \cdot H_2O$	0.21 g/100 mL
$MgKPO_4 \cdot 6H_2O$	0.23 g/100 mL

The release of nitrogen, particularly from the two ammonium compounds, again depends on particle size (slow release only with larger particles or granules).

The manufacturing processes for magnesium ammonium and magnesium potassium phosphates are described in [130, pp. 256 – 284], [168], [169].

Other sparingly soluble inorganic compounds which have been used as slow-release fertilizers are based on crystalline ammonium potassium polyphosphates [170], and on glassy melts of ammonium dihydrogenphosphate, potassium hydrogenphosphate, and dicalcium phosphate [171].

Commercial Products and Trade Names. MagAmp (magnesium ammonium phosphate; Grace Sierra Horticultural Products); EnMag (magnesium ammonium phosphate + potassium sulfate; ICI).

4.4.5. *Coated and Encapsulated Controlled-Release Fertilizers*

Coated fertilizers are conventional soluble fertilizer materials whose plant nutrients are rapidly available and which after granulation, prilling, or crystallization are given a protective coating to control water penetration and hence the rate of dissolution and nutrient release in the soil. The AAPFCO definition is: “A product containing sources of water soluble nutrients, release of which in the soil is controlled by a coating applied to the fertilizer”.

Coated fertilizers are the fastest growing group of slow- and controlled-release fertilizers in the United States and in Japan. They accounted for 24 % of total world consumption of slow- and controlled-release fertilizers in 1995/96 [127, pp. 61 – 63]. The growth is due to improved economics in production, the possibility of controlling the release of nutrients other than nitrogen, and the greater flexibility in determining the nutrient release pattern.

Many condensation polymers, drying oils, waxes, and bitumen were tested for their suitability as coating materials [129, pp. 102 – 263], [172], [173]. However, only three categories of coated/encapsulated controlled-release fertilizers have gained commercial importance:

1. Sulfur coatings
2. Polymer coatings (e.g., PVDC copolymers, polyolefins, polyurethanes, urea – formaldehyde resins, polyethylene, polyesters, alkyd resins)
3. Sulfur – polymer coatings (hybrid products with a multilayer coating of sulfur and polymer) [174-186]

4.4.5.1. *Sulfur-Coated Controlled-Release Fertilizers*

The only inorganic coating material which has achieved any importance is sulfur [129, pp. 102 – 263], [132, pp. 151 – 169], [193-196]. The sulfur coating can be regarded as an impermeable membrane which slowly degrades in the soil through microbial, chemical, and physical processes. Nutrient release depends on the thickness of the coating in relation to the size of the granule or prill and the quality of the urea substrate. The total N content of sulfur coated ureas varies with the amount of coating applied; products currently available contain 30 – 42 % N, 6 – 30 % S, and various sealants and conditioners. Commercial production started in 1972 when ICI commissioned a pilot plant in the United Kingdom.

Manufacture. Most of the modern industrial processes are still based more or less on the technology developed in the 1960s and 1970s by TVA [193], [194], [197] (now National Fertilizer and Environmental Research Center). Preheated (71 – 82 °C) urea granules (1.7 – 2.9 mm) are introduced into a horizontal rotating cylindrical drum. Molten sulfur (143 °C) is sprayed onto the urea granules and quickly solidifies on contact. The average target thickness is 40 µm, but there are various random proportions of granules having thin (< 30 µm), medium (30 – 50 µm) and thick (> 50 µm) sulfur coatings. Any pores and cracks are closed in a second step by addition of a polymeric hydrocarbon/petroleum-based wax or a high-viscosity polymeric paraffin oil with a polyethylene sealant (2 – 3 % of total weight). A flow conditioner such as diatomaceous earth, talc, clay, or silica (2 – 3 % of total weight) is added to give a dust-free, free-flowing product with good handling and storage properties.

Other straight and compound fertilizers can also be coated with sulfur. However, ammonium nitrates and fertilizers with high contents of nitrate nitrogen are excluded due to the risk of explosion.

Agronomic Properties. Generally, sulfur-coated urea products have good slow-release properties. However, resistance of the coating to impact and abrasion is low. The quality of sulfur-coated urea (SCU) is characterized by the rate of N release into the soil solution within seven days (TVA method). SCU-30 indicates a product that releases 30 % of its nitrogen within seven days, resulting in a rather rapid initial effect. If coated too thickly they may exhibit lock-off, i.e., no effective nutrient release. These disadvantages of sulfur-coated conventional fertilizers were the reason for the development of sulfur-coated, polymer encapsulated fertilizers.

Commercial Fertilizers and Trade Names. Sulfur-coated fertilizers are mainly distributed as straight nitrogen grades, e.g., Enspan 39 % N (Hydro Agri, North America), sulfur-coated urea, and sulfur-coated potassium (Nu-Gro Canada).

4.4.5.2. *Sulfur-Coated, Polymer-Encapsulated Controlled-Release Fertilizers*

Sulfur – polymer hybrid coatings combine the controlled-release performance of polymer-coated fertilizers with the lower cost of sulfur-coated fertilizers. Figure 22 [198] shows the flow diagram of the process developed by RLC Technologies. The process yields a sulfur-coated controlled-release granular urea fertilizer with a uniform, durable polymer coating over the sulfur coating. The hot-melt polymer liquid sealant of the conventional process is replaced by specific liquid monomers. When applied sequentially onto the surface of the hot sulfur-coated urea granules, they copolymerize to form a firm, tack-free, water-insoluble polymer coating sealant. The liquid monomers used in this process are diisocyanates, such as MDI (4,4-diphenylmethane diisocyanate), and a polyol mixture of DEG (diethylene glycol) and TEA (triethanolamine); the TEA acts both as a reactive polyol and a catalyst. The resulting product has improved impact resistance. The RLC sealant provides a substantial improvement over polymeric wax and solvent-dispersed sealants [198].

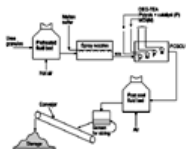

Figure 22. Flow diagram of the RLC Technologies process

Commercial Products and Trade Names. The commercial products (in the United States) generally contain 38.5 to 42 % N, 11 to 15 % S and ca. 2 % polymer sealant: TriKote PCSCU 39–42N (Pursell Technologies), Poly-S PCSCU 38.5–40N (Scotts), and POLY PLUS PCSCU 39N (Lesco).

4.4.5.3. *Polymer-Encapsulated Controlled-Release Fertilizers*

Application of controlled-release fertilizers to high-value crops requires precise control of nutrient release geared to plant requirements. Hence, a range of polymer-coated controlled release fertilizers has been developed, in which the rate of nutrient release can be altered by means of the composition and thickness of the coating, giving longevities from one to 24 months. The release pattern of these controlled-release fertilizers is significantly more linear than that of PCSCU.

Polymer coatings can be semipermeable membranes or impermeable membranes with tiny pores. Most polymers used in coating conventional fertilizers decompose extremely slowly or not at all in the soil and leave residues of up to 50 kg ha^{-1} a^{-1}). This may be considered as an environmental disadvantage even though the accumulation of 10 years (500 kg) only represents 200 ppm in dry soil. Nevertheless, extensive research is being carried out into the development of polymer coatings that are biodegraded after application. In the case of polyethylene, polypropylene, and ethylene copolymers, incorporation of ethylene – carbon monoxide copolymer promotes photochemical degradation of the coating, and coatings that contain a poly(3-hydroxy-3-alkylpropionic acid) as active ingredient are biodegradable [178], [179], [185, p. 14].

Manufacture. In the *Sierra Process* [187] the coating material is a copolymer of dicyclopentadiene with drying or semidrying oils in an organic solvent (glycerol ester of linseed oil). The granules are coated with at least two layers in a coating drum operating at 65 – 70 °C. Maleic acid, for example, is added to improve the drying of the oil. Coating weights vary from 10 to 20 wt %; the commercial products are mainly blends of different coating weights.

The *Chisso Asahi process* [188], [189] produces a particularly smooth coating. The incorporation of finely powdered inorganic materials such as talc and silica into the coating [185, pp. 15, 16] makes it possible to produce controlled-release fertilizers in which the rate of release of nutrients in the soil varies with temperature. A hot 5 wt % solution of the coating material in a hydrocarbon or chlorinated hydrocarbon solvent is sprayed onto the warm granules of fertilizer in a fluidized bed, and the solvents are immediately evaporated with hot air (60 – 70 °C).

Polyolefins, ethylene – vinyl acetate copolymers, poly(vinylidene chloride), and mixtures thereof are used as coating material. The moisture permeability is modified by means of the ratio of ethylene – vinyl acetate (high permeability) to polyethylene (low permeability) [185, pp. 12 – 16].

The release pattern is determined by a water-leach test at 25 °C; for example, T-180 indicates that the product releases 80 % of its nutrient over 180 d at 25 °C in water. Products are produced from T-40 to T-360 [185, pp. 19 – 21].

The *Pursell Technologies Reactive Layers Coating (RLC)* process [190], [191] produces attrition resistant controlled-release fertilizers by coating a plant nutrient with a coating material that chemically bonds to it. Solvent-free polyols and polyisocyanates are pumped to nozzles, located along the entire length of a horizontal, cylindrical rotating drum. The two reactive liquids are applied sequentially in ultrathin layers onto the surface of preheated urea granules, where they polymerize directly to form a continuous polyurethane encapsulating membrane. The first layer of polyisocyanate can also react with the urea granule, chemically bonding the polymer coating to the surface of the urea granule. The thickness of the layer depends on the number of reactive layers and allows controllable release durations of up to 6 months to be achieved.

The polyols used in the process are predominantly polyester polyols. The isocyanates used are also a major factor in influencing the properties of the polyurethane encapsulation. Preferred isocyanate raw materials include polymeric diphenylmethane diisocyanates. Though the process had been primarily defined for the use of urea, most other plant nutrients/fertilizers can also be utilized.

Commercial Products and Trade Names. The main polymer-coated substrates are urea (40–0-0 and 38–0-0), potassium chloride/potassium sulfate, and NP, NK, and ammonium nitrate containing NPK fertilizers. Various grades also contain secondary and minor elements.

Trade names: The Scotts Company, US: Osmocote, Osmocote Plus (NPKs + MgO + TEs), High-N, Sierra, and Sierrablen; Prokote, Scottkote. Chisso Corp., Japan: Meister (urea, KCl and K_2SO_4) Lp and Long, Nutricote (NP, NK, NPK). Pursell Technologies Inc., US: POLYON lines (PC-U, PC-SOP, PC-NPK, PC-MAP, PC-KNO_3). Agrium, Canada: Duration and ESN (clay-coated PCU). Aglukon Spezialdünger, Germany (subsidiary of AgrEvo): Plantacote Depot, Plantacote Control (NPK), and Plantacote Mix. BASF, Germany: Basacote (NPK + TE). Haifa Chemicals, Israel: Multicote 4 (PC-NPK, PC-U, PC-KNO_3). Asahi Chemical Ind., Japan: Nutricote. Mitsubishi Chemical, Japan: M cote (PCU).

4.4.6. *Anti-Float Materials*

Special controlled-release fertilizers which sink immediately on application have been developed for application to irrigated crops. Diatomaceous earth or wetting agents such as liquid surfactants are applied to the external surface of the coated fertilizer to achieve this antifloat effect by breaking the surface tension between the water and the coated fertilizer [192].

Commercial Products and Trade Names. Pursell Technologies Inc. POLYON PCU-AF/AntiFloat, marketed in Japan by Sumitomo; Haifa Chemicals resin-coated anti-floating urea MULTICOTE.

4.4.7. *Controlled-Release Fertilizers on Carriers*

Bayer manufactures controlled-release NPK fertilizers based on water-insoluble synthetic ion-exchange resins [199], [200]. The nutrients are released from the resin granules by reaction with salts in the ground water.

Commercial Products and Trade Names. Bayer AG Lewatit HD 5 NPK 18–7-15 and Lewaterr 80 NPK (31–12–33).

4.4.8. *Supergranules*

A slow release of nutrients can also be achieved by granulation or compaction of conventional fertilizers with a relatively small surface to volume ratio (supergranules, briquettes, tablets or sticks or stakes). Some of these formulations also contain urea – formaldehyde or IBDU.

4.4.9. *Legislation*

In the United States, 50 states regulate their own agricultural policies, including fertilizers. There are some guidelines and Federal EPA regulations which can be imposed on the individual states (mainly concerning registration of pesticides under RECRA — Resource Conservation And Recovery Act — in the EPA). However, fertilizers are excluded [201]. In Western Europe there are not as yet general regulations of the EU Commission on slow- and controlled-release fertilizers, and there are no coated controlled-release fertilizers in the EU type list. There are regulations concerning definitions and classification in the individual member states. These cover fertilizer types such as coated urea, coated NPK, and partly coated NPK.

In Germany, CDU, IBDU and urea – formaldehydes (UFs) are classified as individual fertilizers. Legislation also includes the group of N, NPK, NP, and NK fertilizers containing UF, CDU or IBDU; other legislation covers coated and encapsulated fertilizers [202], [203].

To achieve European standardization, a Task Force slow-release fertilizers (TFsrf) has been formed with the aim of defining the conditions under which type of fertilizer included in the EU fertilizer type list can be newly categorized as a slow- or controlled-release fertilizer [127, p. 11]: A fertilizer may be described as slow-release if the nutrients declared as slow- (controlled-) release meet, under defined conditions including that of a temperature of 25°C, each of the following three criteria:

1. No more than 15 % released in 24 h
2. No more than 75 % released in 28 d
3. At least about 75 % released in the stated release time

This European Task Force has close contacts to the Controlled Release Task Force formed in the United States by AAPFCO (Association of American Plant Food Control Officials) and TFI (The Fertilizer Institute). In Japan registration of slow- or controlled-release fertilizers requires a dissolution test in water under well defined conditions [204], [127, p. 105].

4.5. Nitrification and Urease Inhibitors

4.5.1. *Introduction*

Ammonium ions in the soil — whether from decomposition of organic material or from the application of ammonia-containing mineral fertilizers — are oxidized to nitrite and nitrate. Bacteria of the species *Nitrosomonas* spp. are responsible for the transformation into nitrite. The nitrite is relatively rapidly oxidized to nitrate by *Nitrobacter* and *Nitrosolobus* spp., so that there is normally no toxic accumulation of nitrite in the soil [205, pp. 156, 157], [206, pp. 287, 289]. The process is known as nitrification. The formation of the environmentally relevant gases N_2O and NO may be regarded as a side-reaction of the nitrification process [207].

Nitrate is readily soluble in water and in the aqueous soil medium, and so is completely mobile in the soil, in contrast to ammonia, which is strongly adsorbed in soil colloids and base-exchange complexes. Therefore, it can be leached readily from the soil [208-212]. Under unfavorable (anaerobic) conditions nitrate can be reduced by denitrification to N_2 [206, p. 289]. This can result in further considerable losses of nitrogen [213, pp. 90 – 95].

Addition of a nitrification inhibitor to ammonia containing fertilizers or urea retards nitrification and minimizes leaching of nitrogen as nitrate and losses of nitrogen due to denitrification. Furthermore, nitrification inhibitors also suppress methane emissions and lower nitrous oxide emissions [207], [212], [214-218].

On application to the soil, amide nitrogen, as in urea, UAN (urea – ammonium nitrate solution), and some NPK fertilizers, is transformed by the enzyme urease via the unstable ammonium carbamate ($H_2NCOONH_4$) to ammonia, and CO_2 [205, p. 156].

This transformation has two major drawbacks:

1. It results in sometimes very high volatilization losses of ammonia when urea is applied to the surface [219], [220] or under flooded conditions.
2. It can produce severe seedling damage by ammonia and nitrite [221].

Urease inhibitors inhibit or reduce the formation of urease, which is ubiquitous in surface soils. This slows down the rate of urea hydrolysis in the soil and prevents or at least depresses the transformation of amide nitrogen into ammonia.

Definitions. *Nitrification inhibitors* are compounds that delay bacterial oxidation of the ammonium ion by depressing the activity of *Nitrosomonas* bacteria in the soil over a certain period of time. Thus, they control leaching of nitrate by keeping nitrogen in the ammonium form longer, and preventing denitrification of nitrate [222, p. 12]. *Urease inhibitors* prevent or depress transformation of the amide nitrogen of urea into ammonium hydroxide and ammonia over a certain period of time by inhibiting hydrolytic action on urea by urease; thus they avoid or reduce volatilization losses of ammonia.

There is considerable confusion concerning the terms nitrogen stabilizers, nitrification inhibitors, urease inhibitors, and stabilized fertilizers. The terms nitrogen stabilizers and nitrification inhibitors have been used interchangeably. Strictly speaking, stabilized fertilizers refers only to those which are modified during production with a nitrification inhibitor, such as ALZON and BASAMMON. In all other cases, fertilizers and nitrification and urease inhibitors are sold separately [222, p. 12].

Advantages and Disadvantages of Nitrification and Urease Inhibitors. Nitrification inhibitors significantly reduce leaching losses of nitrate by stabilization of ammonia [219], [220] and reduce emissions of the environmentally relevant gases N_2O and NO [207], [214-216], [223]. Nitrification inhibitors indirectly improve the mobilization and the uptake of phosphate in the rizosphere [224]. Urease inhibitors reduce ammonia volatilization losses, particularly from top-dressed agricultural fields and under reduced tillage conditions [225, p. 9, 10], [226-228]. Urease inhibitors furthermore reduce seedling damage where seed-placed levels of urea-containing fertilizers are too high [219], [221].

Possible disadvantages include the fact that fertilizers containing ammonia and a nitrification inhibitor may result in increased ammonia volatilization if they are not incorporated into the soil immediately after application. Depending on the type of nitrification inhibitor, the activity of soil bacteria may not only be interrupted for a certain time period, but the soil bacteria may actually be killed. This can be regarded an undesirable interference in a natural soil process [213, p. 219], [229, pp. 37 – 44].

4.5.2. *Types of Nitrification and Urease Inhibitors*

Extensive research on nitrification and urease inhibitors has been carried out mainly in Europe, Japan, Russia, and the United States. Various chemical fumigants and pesticides have also been tested to establish their possible effectiveness in inhibiting nitrification (nematicides; soil-insecticides and herbicides) [230, pp. 547 – 554].

Until the late 1960s research was carried out in the United States and Japan on N-Serve (2-chloro-6-trichloromethylpyridine; Dow Chemical Company), AM (2-amino-4-chloro-6-methylpyrimidine; Mitsui Toatsu Chemicals), Terrazole (5-ethylene oxide-3-trichloromethyl-1,2,4-thiodiazole; Olin Mathieson); ASU (1-amide-2-thiourea; Nitto Chemical Industry), and ATC (4-amino-1,2,4-triazole hydrochloride); substituted phenyl compounds (DCS), and compounds of the *s*-triazine line (MAST) [231, pp. 64 – 82]. In Eastern Europe and the former Soviet Union, CMP (1-carbamoyle-3-methylpyrazole) and its main metabolite MP (3-methylpyrazole) were tested extensively. However, only products based on pyridines, dicyandiamide, pyrazoles have gained practical agronomic importance as nitrification inhibitors. Terrazole, AM, and ASU (thiourea) had some regional importance, particularly in the United States and in Japan.

Reseach on urease inhibitors has concentrated on phosphoric triamides. Limited research has been carried out with PPD/PPDA (phenyl phosphorodiamidate) and ATS (ammonium thiosulfate) [222, p. 32].

There are no reliable statistics publicly available on the use of nitrification inhibitors due to the unique production structure. Estimates of the acreage treated with fertilizers containing nitrification inhibitors for the United States are 1860×10^6 ha (1995/96), 1660×10^6 ha thereof with nitrapyrin and 200 000 ha with dicyandiamide. For Western Europe a very rough estimate is 200 000 ha of arable cropland treated with fertilizers containing dicyandiamide [222, p. 63, 64].

4.5.3. *Pyridines*

4.5.3.1. *Nitrapyrin*

Nitrapyrin consists of 2-chloro-6-trichloromethylpyridine [1929-82-4] and related chlorinated pyridines, such as 4,6-dichloro-2-trichloromethylpyridine [1129-19-47].

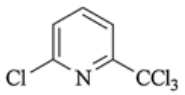

Manufacture. The (trichloromethyl)pyridine compounds are manufactured by photochlorination of methyl-substituted pyridines. The desired amount of chlorine gas is passed through the appropriate methyl-substituted pyridine, generally in the presence of its hydrochloride. The product is recovered by conventional procedures such as filtration or distillation [232]. A flow diagram for the production of chlorinated picolines from readily available raw materials such as α-picoline is given in [233].

Nitrapyrin is produced exclusively by DowElanco in the United States and distributed under the trade name N-Serve (NS) [234-236].

Properties. Nitrapyrin is a white crystalline solid with a mild sweetish odor, *mp* 62 – 63 °C; *bp* 101 °C at 133 Pa. For use in agriculture nitrapyrin is formulated as a liquid product.

N-Serve 24 nitrogen stabilizer consists of 22.2 wt % 2-chloro-6-trichloromethylpyridine, 2.5 wt % related chlorinated pyridines including 4,6-dichloro-2-trichloromethylpyridine, and 75.32 wt % Xylene-range aromatic solvent [64742-96-6].

N-Serve 24E nitrogen stabilizer consists of 21.9 wt % 2-chloro-6-trichloromethylpyridine, 2.4 wt % related chlorinated pyridines, 4,6-dichloro-2-trichloromethylpyridine, and 75.7 wt % Xylene-range aromatic solvent.

Both formulations contain 2 lb of active ingredients per gallon (≈ 240 g/L); NS24 is recommended for use with anhydrous ammonia (82 % N) and impregnation onto urea; NS24E for use with liquid fertilizers (aqueous ammonia, solutions) and with manure (slurry).

Nitrapyrin has a very selective effect on *Nitrosomonas* bacteria. However, this effect is not only bacteriostatic but also bactericidal, so that part of the population in treated soil is killed [213, p. 219], [236], [237].

Toxicity. The single dose toxicity of nitrapyrin is low; the technical material has a LD_{50} of ca. 1000 mg/kg of body weight in laboratory animals. It is slightly irritating to the eyes and skin and has a low vapor toxicity. The equivalent LD_{50} (oral, female rat) for the two formulations is 2140 mg/kg (N-Serve 24) and 3300 mg/kg (N-Serve 24E). In soil and in plants, nitrapyrin is chemically and biologically rapidly degraded into 6-chloropicolinic acid, the only significant chemical residue from its use, and further to N_2, Cl^-, CO_2, and H_2O.

Application. The recommended application rate is 1.4 – 5.6 L/ha. In warm soils the nitrification inhibiting period is normally 6 – 8 weeks; it can be 30 weeks or more in cool soils.

However, in agronomic use this long-term standard nitrification inhibitor has two drawbacks:

1. Loss by volatilization from treated fertilizers during storage
2. The bactericidal effect, which may be regarded as an undesirable interference in a natural soil process

The loss by volatilization restricts it to simultaneous application with anhydrous or aqueous ammonia or fertilizer solutions which are injected directly into the soil at a depth of at least 5 – 10 cm. This limits acceptance in regions where nitrogen fertilizer is not commonly injected. Therefore, N-Serve is available commercially only in the United States.

4.5.3.2. *Other pyridines*

3,6-dichloro-2-trichloromethylpyridine is a proven nitrification inhibitor [238]. However, it is not readily obtained by ring chlorination of a 2-substituted pyridine because such chlorinations are not sufficiently selective to produce a preponderance of the desired isomer. Hence, 3,6-dichloro-2-methylpyridine is prepared by the addition reaction 1,1-dichloro-2-propanone and acrylonitrile to give 4,4-dichloro-5-oxohexanenitrile. Subsequent cyclization in the presence of hydrogen chloride gives 3,6-dichloro-2-methylpyridine, which is chlorinated to obtain 3,6-dichloro-2-trichloromethylpyridine [239].

Commercial Products and Trade Names. N-Serve 24 Nitrogen Stabilizer, N-Serve 24E Nitrogen Stabilizer, DowElanco.

4.5.4. *Dicyandiamide*

Dicyandiamide [461-58-5] exists in two tautomeric forms (⟶ Cyanamides – Dicyandiamide). Its nitrification-inhibiting property was first reported in 1959 [240].

Manufacture. All large-scale production processes are based on calcium cyanamide. In the first step cyanamide is liberated from calcium cyanamide by carbonation in aqueous solution at pH 7 – 8. The cyanamide is then dimerized to dicyandiamide at pH 9 – 10.

$$CaNCN + H_2O + CO_2 \rightarrow H_2NCN + CaCO_3$$

$$2\,H_2NCN \rightarrow H_2N-C=NCN$$

The dicyandiamide is isolated by filtration or centrifugal crystallization.

The disadvantage of this production process is the high energy input for the production of calcium carbide, the raw material for the production of $CaCN_2$; this is reflected in the relatively high price for the product as a nitrification inhibitor. There are only three major producers worldwide: SKW Trostberg and SKW Stickstoffwerke Piesteritz in Germany; Odda Smelteverk in Norway; and Nippon Carbide in Japan. The former Canadian producer, Cyanamid Canada, has ceased production.

Properties. Dicyandiamide is generally supplied as white or colorless crystals in paper or jute bags. Under dry conditions it can be stored for an unlimited period. It has low water solubility (3.2 g/100 g water at 20 °C) and contains at least 65 % nitrogen. When applied to the soil it is decomposed (partly abiotically and partly biotically by specific enzymes) and converted via guanylurea and guanidine to urea, a conventional fertilizer [241-244].

Dicyandiamide has a bacteriostatic effect on the *Nitrosomonas* bacteria. Depending on the amount of nitrogen applied, soil moisture, and temperature, the nitrification-inhibiting effect of dicyandiamide lasts 6 – 8 weeks.

Toxicity. With an LD_{50} of 10 000 mg/kg oral (female rat), dicyandiamide is practically nontoxic. The Ames test did not reveal any mutagenic activity. Furthermore, long-term studies have shown that dicyandiamide has no cancerogenity. Therefore, any risks for human health can be excluded. This also refers to its residues [237], [245].

Processing. In Western Europe the majority of ammonium-containing fertilizers are applied in solid form. However, a satisfactory nitrification-inhibiting effect can not be obtained by simple mechanical mixing of dicyandiamide with the solid fertilizer. Dicyandiamide added to the fertilizer prior to granulation, it decomposes at the usual granulation temperatures of 100 – 140 °C. When dicyandiamide is spread onto the finished fertilizer granules it does not adhere firmly to the granule surface. The use of vegetable, animal, or mineral oils as adhesion promoters has drawbacks with regard to storage and spreading. BASF has developed a process for applying dicyandiamide to ammonium- and sulfate-containing fertilizers without decomposition and with good adherence to the surface of the granules [246]. The fertilizer granules at 60 – 130 °C are mixed in a rotary drum with 2 – 4 wt % of finely ground dicyandiamide with simultaneous addition of a 40 – 50 wt % aqueous solution of $Ca(NO_2)$. The surface of the fertilizer material is rapidly cooled, thus preventing a decomposition of the applied dicyandiamide, which adheres well to the fertilizer granules.

Freeport-McMoRan Resource Partners in the United States has developed two processes for incorporating dicyandiamide into urea in combination with other nitrification- or urease-inhibiting compounds. In the first, dicyandiamide is incorporated into urea in combination with ammonium thiosulfate and a phosphate. Controlled release of N is claimed [247]. In the second, dicyandiamide is incorporated into urea in combination with *N*-(*n*-butyl)thiophosphoric triamide (NBPT) [248].

Application. Dicyandiamide is applied directly with N or NPK fertilizers and fertilizer solution, in which it is incorporated in correct proportion for the ammonia content. These stabilized fertilizers are recommended for all agricultural crops, particularly when grown on light-textured soils where heavy precipitation is expected within 6 – 8 weeks after application or where crops have a relatively slow growth rate during the early growing stages (e.g., potatoes).

Commercial Products and Trade Names. Ensan (BASF technical product); Basammon stabil (27 % total N, of which 1.6 % is dicyandiamide N (Ensan) and 13 % S); Nitrophoska stabil 12–8-17 (12 % total N, of which 1.1 % is dicyandiamide N (Ensan), plus 2 % MgO and 7 % S). Didin (SKW technical product); Alzon 27 (27 % total of N, of which 1.6 % is dicyandiamide N (Didin) and 31 % S); Alzon 47 [47 % total N, of which 3 % is dicyandiamide N (Didin)]; Piadin (SKW Stickstoffwerke Piesteritz) liquid mixture of DCD and 3MP (3-methylpyrazole, the main metabolite of CMP) in a proportion of 15 : 1, contained in Piasin 28/Alzon-flüssig (a urea – ammonium nitrate solution).

4.5.5. *Pyrazoles*

The group of pyrazole compounds has the largest number of compounds showing a nitrification-inhibiting effect [249], [250]. They include 3-methylpyrazole; 3,4-dimethylpyrazole; 4-chloro-3-methylpyrazole; 3-methylpyrazole phosphate; 3,4-dimethylpyrazole phosphate; 4-chloro-3-methylpyrazole phosphate; compounds resulting from the addition of polyacrylic acid to 3,4-dimethylpyrazole, 4-chloro-3-methylpyrazole, 3,4-dimethylpyrazole phosphate, 4-chloro-3-methylpyrazole phosphate; and from the addition of polyphosphoric acid to 3,4-dimethylpyrazole (1:20) and (1:1). In the 1980s and 1990s various compositions were patented [324], [251-256].

Properties. The nitrification-inhibiting effect of pyrazole compounds is better than that of dicyandiamide. However, they have one or more of the following disadvantages: they are liable to hydrolysis, which lowers the stability in storage and the activity period in the soil. They are too toxic and they are highly volatile. 3-Methylpyrazole is so volatile that significant amounts are lost when it is applied onto the surface of fertilizer granules or on fertilizer storages. Metal salts and metal complexes of 3-methylpyrazole [257] are also relatively volatile.

The most extensively tested pyrazole is 1-carbamoyl-3-methylpyrazole (CMP; in Russia: KMP). This compound is not as volatile as 3-methylpyrazole, but its manufacture is difficult and high evaporation losses may still occur when it is applied to fertilizers and on storage.

4.5.5.1. *1-Carbamoyl-3-methylpyrazole*

1-Carbamoyl-3-methylpyrazole (and its metabolite 3-methylpyrazole, MP) has the molecular formula $C_5H_7N_3O$.

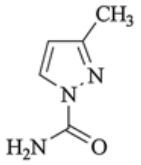

CMP was developed by VEB Agrochemie Piesteritz (now SKW Stickstoffwerke Piesteritz, Germany). Further research and practical use was mainly carried out in the former German Democratic Republic, in Central Eastern Europe, and in the former Soviet Union.

Manufacture. Carbamoyl-3-methylpyrazole (CMP) is produced from 3(5)-methylpyrazole (MP) by carbamoylation with sodium cyanate in aqueous HNO_3. After 1 h, the resulting CMP is removed by filtration, washed with water, and dried in a vacuum dryer. The yield is 90 %.

Properties. Bulk density 630 kg/m^3, solubility in water 0.56 g/100 g, *mp* 123 – 125 °C, evaporation pressure 4.41×10^{-2} Pa (20 °C). Therefore, the technical solid product has to be transported and stored in closed containers. When mixed with solid or liquid fertilizers, these have to be applied and incorporated into the soil immediately after preparation of the mixture. CMP has a bacteriostatic but not bactericidal effect on *Nitrosomonas* bacteria.

Toxicity. CMP has an LD_{50} of 1580 mg/kg (rat, oral) and its metabolite 3MP and LD_{50} of 1312 mg/kg (rat, oral).

Application. For use in field testing and agriculture CMP was formulated as a 50 % CMP-preparation for mixing with solid ammonium-containing fertilizers or solutions. However, because of its liability to hydrolysis and to prevent evaporation losses the CMP formulation could only be added at the time of applying the fertilizer and, like nitrapyrin, had to be incorporated into the soil immediately.

Commercial Products and Trade Names. At present (1998) there is no registration for 1-carbamoyl-3-methylpyrazole or any other pyrazole compound as a nitrification inhibitor in Europe. In accordance with German fertilizer law, Stickstoffwerke Piesteritz has registered a 15 : 1 mixture of dicyandiamide with 3MP under the name Piadin, recommended for use with UAN fertilizer solutions.

4.5.5.2. *Outlook*

In recent years research has concentrated on overcoming the high evaporation losses on adding pyrazole compounds to fertilizer granules and solutions, on storing such modified fertilizers, and on applying them onto the soil. BASF [258] has reported a process in which fertilizer granules are covered with salts of 3-methylpyrazole and made resistant against abraision and evaporation by addition of an inorganic or organic polyacid [259]. This treatment significantly lowers the volatility of the nitrification inhibitor and allows smaller amounts to be used. The storage stability of fertilizers treated by the BASF process is also improved. In a test for 4 weeks at 30 °C, 40 – 50 % R. H., and 1.2 m/s air velocity on 3,4-dimethylpyrazole/polyphosphoric acid 1/20 and 1/1 (DMPP), losses of the applied nitrification inhibitor were 0 and 12 %, respectively. Thus, DMPP is of future interest as a nitrification inhibitor.

4.5.6. ***Neem/Neem-Coated Urea***

The extract from the press cake of the seed of the neem tree (*Azadirachta indica*) exerts a nitrification-inhibiting effect on *Nitrosomonas* spp. This effect is reversible.

The active compound in this biological nitrification inhibitor consists of various terpenes/triterpenes (epinimbin, nimbin, deacetylnimbin, salanin, azadirachtin, deacetylsalanin).

Toxicity. The formulated product is nontoxic.

Application. For practical use, a product containing 5.0 % active ingredient is to be mixed with urea in a ration of 1 : 100 before application.

Commercial Products and Trade Names. Nimin (5 – 5.5 % active ingredient) produced by Godrej Agrovet, Pirojshanagar (India).

4.5.7. ***Urease Inhibitors***

Although it has long been known that substantial evaporation losses in form of ammonia from urea may occur, it was only in the 1980s that the first chemical substances with effective urease-inhibiting properties in soils were disclosed, namely the phosphoric triamide compounds. Losses of between 3 and 40 % and more of applied nitrogen are possible when urea is not incorporated immediately after application and stays on dry soil (particularly on soil with a high pH and with high temperature). Also if it is applied with conservative tillage, particularly no-till, and on crops which are not tilled, such as bananas, sugar cane, oil palms, and rubber, as well as on flooded rice [206, p. 290], [213, p. 211], [218], [260-263], [264, pp. 52, 59]. Reference [265] lists 146 triamide compounds, of which *N*-(*n*-butyl)thiophosphoric triamide (NBPT) showed the best inhibition values.

N-(n-butyl)thiophosphoric triamide [94317-64-3], NBPT urease inhibitor is the best developed of the *N*-alkyl thiophosphoric triamide for commercial applications.

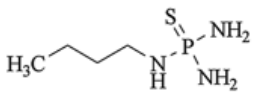

Manufacture. NBPT is prepared by a two-step synthesis in THF:

$$PSCl_3 + n\text{-}BuNH_2 + NEt_3 \rightarrow n\text{-}BuNHPSCl_2 + HNEt_3Cl$$

$$n\text{-}BuNHPSCl_2 + 4\,NH_3 \rightarrow n\text{-}BuNHPS(NH_2)_2 + 2\,NH_4Cl$$

The approximate product composition from this manufacturing process is as follows [337]:

- *N*-*n*-butylthiophosphoric triamide (NBPT), min. 85 wt %
- *N*,*N*-di-*n*-butylthiophosphoric triamide (DNBPT) 0 – 3 wt %
- *N*,*N*,*N*-tri-(*n*-butyl)thiophosphoric triamide (TNBPT) 0 – 3 wt %
- Thiophosphoric triamide (TPT) 0 – 3 wt %
- Others 0 – 10 wt %

– THF 0 – 2 wt %
– Triethylamine 0 – 2 wt %

Properties. Pure NBPT is a white crystalline solid, industrial grade NBPT is a waxy, sticky, heat- and water-sensitive material, which render this material difficult to handle. This material is susceptible to decomposition during storage and distribution. The vapor pressure is 1.1 kPa at 40 °C. The compound is an excellent urease inhibitor that inhibits the activity of urease for 12 – 14 d on dry soil. When incorporated into the soil in combination with urea or urea-containing fertilizers, it significantly reduces seedling damage when seed-placed levels of such fertilizers are too high [219], [221]. For satisfactory use on flooded crops, further research is needed. *N*-methyl-pyrrolidone (NMP) is a good carrier for NBPT [268]. However, cheaper formulations with better long-term stability are obtained with glycols and glycols with a liquid amine cosolvent [269].

Toxicity. The acute oral LD_{50} of NBPT is 1000 – 4000 mg/kg. The Ames tests were negative [266, p. 23]. The compound poses a very low acute toxicity hazard to workers and has received EPA approval. Because it inhibits free urease in the soil without affecting bacterial growth, it is not classified as a pesticide and hence not regulated under TSCA. In the soil, the product degrades into the fertilizer elements nitrogen, phosphorus and sulfur.

Application. For practical use in agriculture NBPT (Agrotain) is formulated as a green clear liquid containing [267] 25 % *N*-(*n*-butyl)thiophosphoric triamide, as active ingredient, 10 % N-methylpyrrolidone, and 60 – 65 % other nonhazardous ingredients. The recommended rate of application depends exclusively on the quantity of amide-nitrogen applied as urea, UAN, or in the form of NPK-fertilizers (1.4 kg per tonne of urea).

Agrotain is primarily recommended for pre-planting surface application of urea and urea-containing fertilizers, but may be used in pre-emergence, side-dress, top-dress, or other post-planting applications. It is not recommended for use if rain is imminent [222, pp. 38, 39].

Commercial Products and Trade Names. The only commercial product for use in agriculture is Agrotain (IMC-Agrico).

4.5.8. *Environmental Aspects*

In assessing the value of nitrification and of urease inhibitors not only must the better utilization of the applied nitrogen be taken into account, but also the possibility of maintaining clean ground water, as well as reduced emissions of ammonia and other environmentally relevant gases [222]. In Germany, where, in water catchment areas with restrictions or for other reasons, a reduction in applied nitrogen is required, the recommendation is that nitrogen application can be reduced by approximately 20 kg/ha without loss of yield.

Of equal importance are the positive environmental properties of nitrification inhibitors in significantly reducing emissions of climatically relevant gases such as N_2O and methane [207].

The application of urea or UAN-solution modified with a urease inhibitor such as NBPT would permit a substantial reduction in nitrogen losses to the atmosphere, and consequently also in the application rates, without affecting growth and yield of fertilized crops.

4.5.9. *Legal Requirements*

In Western Europe, there is no uniform regulation for DCD, the leading nitrification inhibitor (urease inhibitors are not yet in use). As in the case of slow- and controlled release fertilizers, individual countries [270], [271] have established national classification and legislation [272]. The pyrazole compound CMP (in combination with DCD) also comes under fertilizer legislation.

In the United States only fertilizers, but not additives to fertilizers, are excluded from RECRA (Resource Conservation And Recovery Act) regulations of the EPA. Therefore nitrapyrin — the active ingredient in N-Serve, the leading nitrification inhibitor in the United States — is classified as a pesticide in the EPA registration. In 1996 it was decided that all nitrification inhibitors have to be EPA registered as pesticides in the United States [222]. However, in spite of the new regulation, DCD will not require registration as a pesticide. NBPT, which does not affect soil organisms, is not classified as a biocide, but is regulated under the TSCA.

4.6. Organic Fertilizers (Secondary Raw Material Fertilizers)

At present an EC catalogue of commercial organic fertilizers is not available. Since these fertilizers mainly contain exploitable plant and animal wastes and in future will contain more waste of organic origin (with or without pretreatment such as composting, anaerobic fermentation, etc.), comprehensive legal provisions are in preparation. In Germany these organic fertilizers are designated secondary raw material fertilizers (*Sekundärrohstoffdünger*). According to the German Fertilizer Law their main components are organic waste materials for agricultural and horticultural utilization (secondary raw materials). These fertilizers are also subject to the German Waste Law (analysis of pollutant content, description of subsequent utilization) and the German Fertilizer Law (control of nutrient content) [273-275]. Depending on the quality standard, monitoring systems varying in analytic sophistication are specified. The main argument for the application of these fertilizers is the preservation of natural (and therefore limited) nutrient resources (e.g., P) by means of systematic recycling. The applied amount of these fertilizers is mainly limited by their nutrient content. In the case of low quality, the pollutant load can also determine the applied quantity. In Germany the permitted pollutant load conforms to the threshold values fixed in the decrees for the utilization ofsewage sludge (AbfKlärV), which are likely to be lowered substantially for reasons of soil protection [276]. According to the pollutant content or the ratio of nutrients to pollutants, organic wastes can be ranked as to their applicability as secondary raw material fertilizers (Fig. 23). For fertilizers containing several organic waste materials, each component must meet the legal requirements for secondary raw material fertilizers. Blending of unapproved material with compounds of low pollutant content is prohibited.

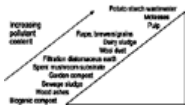

Figure 23. Applicability of wastes for agricultural use [277]

Another group of organic – mineral fertilizers are mixtures of fertilizer salts (N, P, K) with peat, composted bark, lignite dust, or, occasionally, dried slurry. The difference between these fertilizers and secondary raw material fertilizers is that the primary product of organic origin is not classified as waste. Nevertheless, threshold values for heavy metal concentrations have been set [e.g., 3 mg Cd/kg dry weight (dw); 750 mg Zn/kg dw]. In the case of peat, different combinations of fertilizer salts (N, NP, NPK, PK) are added. In the case of composted bark, lignite, or slurry, only mixtures with NPK fertilizers are supplied [274].

Organic fertilizers are also classified according to their effects:

– Fertilizers with an improving effect on soil condition have a stable organic substance as well as a slow effect on N supply (low concentration of readily available nitrogen)
– Fertilizers with short-term effects on nutrient supply, especially of nitrogen, have high contents of mineral nitrogen and/or readily available organic N compounds

Examples for the first category are composts (i.e. biogenic composts, garden composts) and fertilizers based on peat. The second category includes sewage sludge (high NH_4 concentration), blood meal, and potato starch wastewater (readily degradable organic N compounds).

Organic fertilizers with or without the addition of mineral fertilizer salts are classified as follows:

1. Fertilizers based on peat (or materials of similar stability): peat, composted bark, lignite
2. Fertilizers based on waste materials of animal origin: horns, bone meal, blood meal, hide meal, feather meal, guano
3. Fertilizers based on waste materials of plant origin (selected examples of waste materials from the food- and feedstuff industries): castor cake, cacao waste, brewer's grains, rape (marc), vinasse, spent mash, potato starch wastewater, filtration diatomaceus earth
4. Fertilizers based on municipal waste: sewage sludge, biogenic and garden composts

4.6.1. *Fertilizers Based on Peat or Materials of Similar Stability*

Peat (⟶ Peat) is an organic material of stable structure and low nutrient content that has no effect on nutrient supply without the addition of mineral fertilizers, but improves soil condition (e.g., aeration, water-retention capacity). The raw material for peat fertilizers is mainly highly decomposed upland moor peat (black peat) or acidic low moor peat (pH < 6). These peats (pH 3 – 5, volume weight 50 – 200 g/L, salt content 50 – 1500 mg/L, ash 2 – 15 %, organic matter 85 – 98 %) have high sorption capacity (cation sorption capacity 300 – 600 mval/L).

Depending on the kind of application, the crop, and the nutrient status of the soil, several mixtures of peat with mineral fertilizers (N, NP, NPK, and PK salts) are available. The following minimum requirements for nutrient content are specified: 1 % N, 0.5 % P, 0.8 % K, 30 % organic substance [274].

Similar mixtures of organic and mineral fertilizers are based on composted bark or lignite (partial substitute for peat) and other organic materials such as dried slurry or spent mushroom substrate (SMS). The minimum fraction of organic substance is set at 15 %. The organic substance of composted bark is highly stable against biological degradation and resembles the organic substance of peat. After complete decomposition, composted bark shows no N immobilization and is therefore suitable for replacing peat in mixtures with other fertilizers (pH 5 – 7, volume weight 150 – 300 g/L, salt content 100 – 1500 mg/L, ash 12 – 45 %, organic substance 55 – 88 %). Threshold values for heavy metal concentrations are specified for these fertilizers [274], [278].

4.6.2. *Fertilizers Based on Waste Materials of Animal Origin*

Only waste materials that pose no health risk may be converted into fertilizers. The raw materials arise as byproducts at butchers' shops, slaughterhouses, and carcass-disposal plants. Horn, bone meal, and blood meal are the most widely used in agriculture. In some cases, the primary products, containing mainly N and P, are upgraded by the addition of K salts [279], [280], [281], [282].

Horns. Horn materials (e.g., horns, hoofs, claws) consist mainly of the filament protein keratin. Since keratin decomposes slowly, horns represent a slow N source (22 % of the protein is in the form of cystein). The rate of N release increases with the extent of grinding (chips < grit < meal). Nutrient concentrations are listed in Table 24.

Table 24. Nutrient concentrations in fertilizers based on animal wastes (% in dry matter)

Fertilizer	N	P	K	Ca	Mg	C/N
Horn	10 – 14	0.4 – 4.0	0.2 – 0.8	1.5 – 7.5	0.5 – 1.0	3 – 4
Blood meal	9 – 13	0.2 – 1.0	0.2 – 1.5	1.5 – 3.0	0.4	2.4
Feather meal	12	0.3	0.2			4
Bone meal	4 – 6 (0.8)	7 – 12	0.2	18 – 25	0.6	4.5
Carcass meal	6 – 11	2.4 – 7.0	0.3 – 0.5	4 – 10	0.2	3.5
Guano	8 – 16	2 – 7	1 – 3	18	3 – 5	3 – 4

Bone Meal. Bones are ground, partly degreased, and cleaned. The main constituent of bone protein is the filament protein collagen (90 % of bone N is bound in collagen). By treatment with hot water and steam, collagen is converted into glutine and is removed. Therefore, steamed bone meal contains only 0.8 % N (untreated bones 4 – 6 % N). Due to the high P content (7 – 12 % P), bone meals are mainly applied as P fertilizers (Table 24). Occasionally, tricalcium phosphate is converted into monocalcium phosphate by acid treatment.

Blood Meal. Besides blood, blood meals often contain other slaughterhouse wastes such as intestine contents. These mixtures are dried and ground (N content 9 – 13 %; Table 24). Sometimes, mixtures of fresh blood (3 % N, 0.1 – 0.2 % P, 0.1 % K, 80 % water content) with solid wastes are used. Like meat meal, blood meal has a short-term effect on N supply.

Hide Meal. The main constituents of hide meal, a waste material from leather production, are skins and hair. The effect on N supply is quite slow (8 – 11 % N dw, C/N 5). In Germany, the threshold value for Cr^{3+} is set at 0.3 % in dry matter (for agricultural utilization, Cr^{6+} is prohibited in hide meal).

Feather Meals. These fertilizers show slow N release, comparable to that of horns. The N content of 13 – 14 % (Table 24) is mainly bound as keratin.

Meat and Carcass Meal. Occasionally, meat meal and carcass meal are converted into organic fertilizers. These slaughterhouse wastes have a high proportion of protein and hence a short-term effect on N supply (meat meal: 11 % N, 2.4 % P; meat-and-bone meal: 10 % N, 2.5 – 7 % P; carcass meal 6 – 10 % N; Table 24).

Guano consists of partially mineralized excrements of seabirds and can also contain feathers and carcasses. It is obtained from deposits in arid coastal regions of South America. Guano is converted to guano fertilizers by acid treatment. Guano fertilizers consist mainly of inorganic substances and contain 8 – 16 % N, 2 – 7 % P (ammonium and calcium phosphates) and 1 – 3 % K (Table 24). Since up to 50 % is in the form of ammonium, guano has a short-term effect on N supply.

4.6.3. *Fertilizers Based on Wastes of Plant Origin*

Provided the waste materials do not pose any health risk, conversion into fertilizers can be performed without special pretreatment. Both aerobic (composting) and anaerobic methods (fermentation, generation of biogas) are suitable for the treatment of plant wastes. The pretreatment and conditioning of some wastes are regulated. All substances described have low pollutant contents. Therefore, the applied quantities are only limited by nutrient contents [274], [281], [282]. The nutrient concentrations in these fertilizers are summarized in Table 25.

Table 25. Nutrient concentrations (% dw) in fertilizers based on wastes of plant origin

Fertilizer	N	P	K	Ca	Mg	org. S	C/N
Castor cake	5.5	0.8	0.8 – 1.6	0.4	0.3	80	8
Cacao waste	2.3 – 3	0.4 – 0.5	2.5 – 3.0	0.6	0.5	90	17 – 24
Brewers grains	4	0.4 – 0.7	0.5 – 1.0	0.3	0.2	65 – 75	10 – 12
Fruit pulp	1.0	0.3	1.3	0.8	0.1	85 – 95	40 – 50
Rape	1.5 – 2.5	0.4 – 0.8	3 – 3.5	1.0 – 1.8	0.2	75	20 – 30
Vinasse	3 – 4	0.15	6 – 7	0.6 – 1.2	0.3	50	8
Pulp	5 – 6	0.7 – 1.0	0.8 – 6.0			75 – 85	8 – 10
Potato starch wastewater	4 – 8 *	0.7 – 0.9 *	10 * – 12				5 * – 7
Filtration diatomaceus earth	0.7	< 0.1	< 0.1	0.2	< 0.1	6	4 – 8

* In fresh potato starch wastewater.

Castor cake is the residue of castor beans after oil has been pressed. Owing to its high protein concentration, castor cake contains 5 – 6 % N, which is ammonified in soil quite rapidly and thus becomes plant-available. The fertilizers are only traded in sealed packages due to their allergenic properties. Furthermore, the toxic substances ricin and ricinin must be destroyed by steam treatment at 120 °C for several hours, and dust has to be absorbed by treating

the castor cake with linseed oil.

Cacao waste is a residue of cocoa production and has a lower N and P content and higher K content than castor cake.

Brewers' grains (water content 75 %) and *rape* (water content 40 – 60 %) result as waste from brewing and from production of fruit juices and wine. They are used as fertilizers in fresh and composted form and have only a minor short-term effect on N supply.

Vinasse. Sugar-containing molasses is a waste material of sugar production from sugar beet. It is used for alcohol, yeast, and glutamate-production. The remaining sugar-free molasses is concentrated to 60 – 70 % and is then used as sugar-beet vinasse, an organic N and especially K fertilizer. Vinasse has a medium-term effect on N supply.

Pulp remains after alcohol production (distillery) from corn or potatoes (5 – 7 % dry matter). Potato pulp contains more K than corn pulp. Pulp has a medium-term effect on N supply, because potato and corn protein must first be decomposed by microorganisms. Ten cubic meters of pulp is equivalent to 30 – 40 kg N, 4 – 6 kg P, 5 – 40 kg K, and 200 – 300 kg C. Pulp is applied for fertilization as fresh material. Storage without addition of preservatives leads to unpleasant odors [283].

Potato Starch Wastewater. The production of potato starch generates large quantities of potato starch wastewater, which is used as a fertilizer. Depending on the production technique, 3 – 50 % of the total N is present in the ammonium form. Furthermore, the protein N is mineralized rapidly. Therefore, potato starch wastewater has a short-term effect on N supply. It also has a high K concentration. Potato starch wastewater is classified as follows:

- Fresh potato starch wastewater (ca. 6 % dry matter; 1 L is equivalent to 60 g dry matter, 5 g N, 025 g P, 6 g K)
- Treated potato starch wastewater (e.g., protein precipitation, evaporation, partly mixed with wastewater), 1 – 5 % dry matter (in the case of storable syrup, up to 55 % dry matter [284], [285])

Filtration diatomaceous earth is the filtration residue in breweries (75 % water content). Yeast proteins are the main N-containing substances. Owing to the narrow C/N ratio, this waste material has a short-term effect on N supply.

4.6.4. *Fertilizers Based on Municipal Waste*

This group is represented by biogenic and garden composts as well as by sewage sludge. In Germany the produced nutrient quantity of secondary raw material fertilizers is estimated at 270 000 t N per annum and 90 000 t P per annum, of which sewage sludge accounted for 60 % of N and 75 % of P. Composts accounted for 20 – 25 % of N and P. This demonstrates the importance of nutrient recycling by utilizing these waste materials as fertilizers. The precondition for recycling these wastes in agriculture and horticulture is a high quality standard, which strongly depends on the content of pollutants, especially heavy metals, dioxines, and furans. In Germany legal provisions for the utilization of sewage sludge and biocomposts have already been set or are in preparation (Waste Law, Fertilizer Law). These rules are likely to reduce environmental risks to a tolerable minimum and to promote the acceptance of the secondary raw material fertilizers [275], [281], [282]. Lately, sewage sludge and biocomposts have been classified as organic NPK fertilizers in the group of secondary raw material fertilizers [274].

Sewage sludge is mainly produced in municipal purification plants. Sewage sludge is classified into primary (residue after mechanical purification, deposits in settling tanks) and into secondary sludge (residue after biological purification by biomass and chemical purification by precipitation). Generally, the two kinds of sewage sludge are mixed and are subjected to various treatments. A precondition for the utilization of sewage sludge as fertilizer is an acceptable health standard. In Germany pollutant concentrations of sewage sludge have been considerably reduced over the last 20 years (e.g., Cd, Figure 24). At present, the average Cd concentration is 80 % lower than the threshold value specified in the decrees (AbfKlärV) for the utilization of sewage sludge (10 mg Cd/ kg dry matter) [276]. In 1993 dioxin and furan concentrations averaged 27 ng/kg, well below the threshold value of 100 ng/kg dry matter. Adjustment of the threshold values to the present quality standard is in preparation. In 1994 the average concentrations of heavy metals in sewage sludge applied in agriculture (in mg/kg dry matter) were Pb 84, Cd 1.8, Cr 56, Cu 251, Ni 32, Hg 1.6, Zn 977, all well below the permitted maximum loads set in the decrees for the utilization of sewage sludge (AbfKlärV) [276], [286].

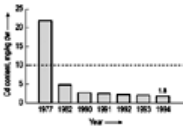

Figure 24. Cadmium concentration in sewage sludge from 1977 to 1994 [286] (the dashed line marks the threshold value for agricultural use [276])

Depending on the pretreatment (dewatering; precipitation of P with Ca, Fe, Al salts; lime addition) nutrient concentrations in sewage sludge vary considerably (Table 26). Therefore, the nutrient contents of these fertilizers must be routinely monitored and adjusted if necessary. The short-term effect on N supply depends on the NH_4 content (fraction of total N). The effect on P supply depends mainly on prior precipitation and pretreatment (lime addition, drying). The applied amount of these fertilizers is determined by nutrient concentrations. For example, the input of P should correspond to the P removed by harvest. Thus the applied quantity remains under the maximum amount of 5 t dry weight of sewage sludge per hectare in three years.

Table 26. Dry matter and nutrient contents of sewage sludge (in wt %) [275]

	Dry matter	N	P	K	Mg	CaO
Range	6 – 23	2 – 6	1.5 – 7.0	0.2 – 0.5	0.4 – 0.7	0 – 15
Maximum	75	25	15			40
Average	12	3.8	1.6	0.3	0.6	

Biogenic and Garden Composts. Biogenic compost originates from biologically degradable waste, separated and collected in households. The portion of biologically usable waste in total domestic waste is 30 – 40 wt %. Garden compost consists of waste materials of plant origin such as litter or remains from pruning, collected in gardens, parks, and embankments and in National Parks. Composts represent the final product of a controlled rotting process in piles or bioreactors. According to the extent of rotting, composts are characterized as fresh or mature composts. Biogenic wastes are subjected to biogas production and subsequent composting.

As a result of the separate collection of the organic wastes, biogenic composts have lower pollutant concentrations than composts based on total domestic wastes (Table 27). Heavy metal concentrations of biogenic and especially garden composts remain under the threshold values set in the decrees for the utilization of biogenic composts (BioAbfV, in preparation). Average concentrations of organic pollutants (e.g., dioxins, furans) of 10 – 15 ng TE/kg dry weight lie within the range of the unavoidable natural background load from the atmosphere. Nutrient concentrations of composts vary less than those of sewage sludge but still have to be routinely monitored (Table 28). Garden composts have lower nutrient contents than biogenic composts. Compared to sewage sludge (high N and P levels) composts have high K contents and mainly a soil-meliorating effect. The short-term effect on N supply is quite low (only 10 % of total N is represented by mineral N). An effect on N supply is barely detectable after applications for several years. The maximum quantity of compost application is likely to be restricted to 20 t (30 t in case of high quality , i.e., minor contents of heavy metals) dry weight over three years per hectare [i.e., 100 (150) kg ha^{-1} a^{-1} of N) [287], [289], [275].

Table 27. Heavy metal concentrations in composts originated from domestic, biogenic and garden wastes (mg/kg dry matter) [287], [288]

Element	Domestic waste	Biogenic waste	Garden waste	Maximum concentration (BioabfV *)
Pb	286	55	59	100 (150) **
Cd	3.9	0.4	0.4	1 (1.5)
Cr	60	27	28	70 (100)
Cu	261	50	36	70 (100)
Ni	40	15	14	35 (50)
Hg	2.6	0.3	0.2	0.7 (1)
Zn	1020	201	133	300 (400)

* BioabfV: Bioabfallverordnung (German Biowaste Regulations), August 1998.
** Maximum quantity of compost application: 30 (20) t dry weight over three years per hectare.

Table 28. Nutrient concentrations in composts (% dry matter) [275]

Compost from	N	P	K	Mg	CaO	Salts (g/L)
Biogenic wastes	1 – 1.8	0.2 – 0.5	0.5 – 1.3	0.3 – 1.6	1.7 – 9.5	3 – 9
Garden wastes	1.0	0.2	0.6	0.5	8.5	(1 – 6)

4.7. Manure

There are four types of animal manures: solid manure, liquid manure, slurry, and dry manure from poultry. Solid manure consists of feces, urine, and bedding. Usually, *solid manure* contains 100 % of the feces; the fraction of the urine retained depends on amount and kind of bedding material, animal type, and the way of housing. The most widely used bedding material is straw, but it can also be peat litter or sawdust. *Liquid manure* contains the urine of the animal together with some rain water and cleaning water and small amounts of feces and bedding material. *Slurry* is a mixture of feces and urine with some additional water and some bedding and feed material. *Layer dry manure* consists of the excrement layers from battery farming.

Solid manure production is the most frequent animal manure handling system for raising young layers, broilers, turkeys, and cattle, the latter especially on smaller farms. For fattening pigs and cattle on larger farms, slurry systems have been favored in the last 30 years because production costs, especially labor costs, are lower.

4.7.1. *Composition [290-296]*

A major fraction of the nutrients consumed by animals is excreted with the feces and urine. For nitrogen, this is ca. 70 – 80 % of the intake, for phosphorus ca. 80 %, and for potassium 90 – 95 %. Table 29 lists the nutrient amounts produced per year by various animals.

Table 29. Nutrient amounts excreted per year from various animals

Animal	Nutrient amount, kg		
	N *	P_2O_5	K_2O
Milking cow (6000 kg/a milk yield)	110	38	140
Fattening cattle (125 – 600 kg live weight)	42	18	44
Sow with piglets	36	19	16
Fattening pig	13	6	6
Laying hens (100)	74	41	33
Fryers (100)	29	16	16
Turkeys (100)	164	81	71

*Excluding gaseous losses from storage or application.

The nutrient contents of manures are affected by type of feed, kind and amount of bedding material, amount of water supply, and kind and length of storage. Even for the same type of manure, the nutrient content can therefore show a wide range; deviations of more than ± 50 % from the average contents are possible. Table 30 lists the average content of organic matter, nitrogen, phosphorus, and potassium in various types of manures.

Table 30. Contents of organic matter and nutrients in various manures

Manure type and amount	**Organic matter, t**	**Nutrient amount, kg**		
		N	**P_2O_5**	**K_2O**
Solid manure, cattle, 10 t	2	54	32	70
Solid manure, pig, 10 t	2	80	80	60
Solid manure, fryer, 1 t	0.2	28	21	23
Solid manure, turkey, 1 t	0.2	23	17	16
Dry poultry manure, 1 t (50 wt % solids)	0.4	28	21	15
Slurry, cattle, 10 m^3 (10 wt % solids)	0.7	47	19	62
Slurry, pig, 10 m^3 (6 wt % solids)	0.5	56	31	30
Slurry, poultry, 10 m^3 (14 wt % solids)	0.9	98	83	48

4.7.2. *Manure Nutrient Efficiency[291], [294], [297-303]*

Availability of Nutrients. In animal manures, phosphorus, potassium, magnesium, calcium, and the micronutrients are predominantly present as inorganic compounds. Their plant availability is comparable to that of mineral fertilizers. The fraction of inorganic nitrogen depends on the type of manure (Table 31).

The organic nitrogen becomes partly available to the crop after mineralization in the soil. In the long run and under optimal conditions, liquid manure reaches a nitrogen fertilizer equivalent of up to 100 %, slurry 70 – 90 %, poultry excrement 60 – 70 %, and solid manure 50 – 60 %.

Table 31. Percentage of ammonium N and organic N in various manures

Type of animal	Type of manure	Percentage nitrogen as	
		Ammonium	Organic
Various animals	solid manure	15	85
Poultry	dried feces	30	70
Poultry	fresh feces	45	55
Cattle	slurry	55	45
Poultry	slurry	60	40
Pigs	slurry	70	30

As for all fertilizers, especially with regard to nitrogen, the nutrient efficiency of manures is highly dependent on the application conditions. Especially for slurry and liquid manure, spring application usually results in a much better nitrogen efficiency than autumn or winter application, especially in areas where nitrate leaching occurs over winter. Ammonia losses can be kept low if slurry is incorporated into the soil as soon as possible after application.

Humus Effect. As shown in Table 30, animal manures contain an considerable amount of organic matter. This organic matter may increase the organic matter content of the soil, depending on the application rate and the cropping system. Application of solid organic manures usually leads to a larger increase in organic matter content than slurry due to the bedding material, which is low in N and hence less readily decomposable. The application of organic matter improves soil quality in terms of water-retention capacity, turnover nutrient availability, and nutrient pore volume, and resistance to soil erosion.

4.7.3. *Environmental Aspects [294], [295], [298], [304-307]*

Environmental problems may result from inappropriate application of animal manures, especially at times when the crop does not require nutrients, at rates which exceed the nutrient demand, or if animal manures are not incorporated into the soil immediately after application. The consequences can be:

- Nitrate losses due to leaching after nitrification of ammonium N or mineralization of organic N
- Ammonia losses from the soil or crop surface
- Potassium losses due to leaching on light soils
- Phosphorous accumulation in the top soil and in the subsoil; phosphorus leaching from peat soils

With the aim of decreasing nitrate leaching into ground water, some countries have established legal regulations that restrict the amount of manure and/or timing of application (see Chap. Environmental Aspects of Fertilizer Application). A regulation to reduce ammonia losses from manure application and storage has been implemented in the Netherlands.

[Top of Page]

5. Fertilizer Granulation

Heinrich Dittmar

5.1. Introduction

The granulation of fertilizers was one of the most significant advances in fertilizer technology, affording considerable advantages to both manufacturer and user. Today, a well-defined grain size distribution is specified just as nutrient contents and good application properties are. Although the first granular fertilizers came on the market between 1920 and 1930, a stronger trend toward granulation developed — especially in the United States — only after the end of World War II.

In 1976, both granular fertilizers and bulk blends enjoyed shares of somewhat more than 40 % in the U.S. mixed-fertilizer market. Granular fertilizers were losing ground against bulk-blend products and liquid fertilizers [313]. In 1990, the corresponding figures are about 63 % for bulk blends, 22 % for liquid fertilizers and 15 % for granular fertilizers. In Europe, Africa, and Asia, granular fertilizers are the most frequently used form, far ahead of bulk blends and fluids.

Advantages of Granular Fertilizers. Forming and subsequent conditioning are indispensable for the production of fertilizers suitable for use. It was recognized at an early stage that fertilizers in powdered or finely divided form readily cake during storage. This is less of a problem with low-surface-area granules. Only free-flowing materials allow mechanized handling and distribution. Granules often require less storage space because of their greater bulk density: they are stored and transported more economically. A further advantage of granular fertilizers over powdered and crystalline products is that they tend to produce less dust, so that product losses are reduced. A granular product with a definite grain-size spectrum is a prerequisite for uniform mechanical application with field equipment (see Section Application): granules with diameters between 1 and 5 mm are most suitable. At the same time, losses caused by the wind, and the accompanying environmental problems, are dramatically reduced. Moreover, granules produced from various feedstocks (solids, slurries, melts) by granulation do not segregate, in contrast to bulk-blended products (Section Bulk Blending).

The use of granular instead of powdered fertilizers delays nutrient delivery to the plant until the granules have disintegrated completely (controlled delivery to the plant, diminished leaching losses). In the case of some controlled-release fertilizers, larger granules release nitrogen more slowly (see Section Slow- and Controlled-Release Fertilizers). Field studies in Swedish soils have shown that granular superphosphate with a grain diameter of 1 – 3.5 mm was twice as effective as finely-divided fertilizers [314], since the granular form retards phosphate fixation in the soil [315]. This reported effect varies with the soil type, the pH, the proportion of water-soluble P_2O_5 and the type of plant [316]. In the case of mineral fertilizers not containing P_2O_5 (N, NK, and NMg fertilizers), however, the grain size has only a slight effect.

Definitions [312, Chap. 1].

- Straight fertilizer: a fertilizer containing only one nutrient.
- Compound fertilizer: a fertilizer containing two or more nutrients.
- Complex fertilizer: a compound fertilizer formed by mixing ingredients that react chemically.
- Granular fertilizer: a fertilizer in the form of particles between two screen sizes usually within the range of 1 – 4 mm.
- Prilled fertilizer: a granular fertilizer of near-spherical form made by solidification of free-falling droplets in air or other fluid medium (e.g., oil).
- Coated fertilizer: a granular fertilizer that has been coated with a thin layer of some substance to prevent caking or to control dissolution rate.
- Conditioned fertilizer: a fertilizer treated with an additive to improve physical condition or prevent caking. The conditioning agent may be applied as a coating or incorporated in the granule.
- Bulk-blend fertilizer: two or more granular fertilizers of similar size mixed together to form a compound fertilizer.

Granulation Loop Granulation may be coupled with a production step, such as the manufacture of ammoniated triple superphosphate, or on the other hand it may be only a forming step in a production process, for example, granulation in the nitrophosphate process [317]. But other production operations also come

under the heading of granulation: the preparation of feed materials and, after forming, the steps of drying (Fig. 26), cooling, screening, comminution of material with too large a grain diameter (oversize), recycling of this comminuted material and of material with too small a grain diameter (undersize) to the granulator, and finally conditioning of the particles with the desired grain size (product fraction). The processing steps, linked into a loop by the recycle, are called the granulation loop (Fig. 25).

Recycling is carried out for the following reasons:

1. The size distribution leaving the granulator differs from the required distribution
2. The ratio of liquids to solids in the available feed is in excess of the requirements for the desired size enlargement
3. Granulated material is recycled to provide nuclei for the granulation process

Processes which correspond to (1) may be described as granulation efficiency limited; an example is the agglomeration of low-solubility fertilizers. Condition (2) is generally encountered where readily soluble or high water content feeds are agglomerated and can be described as liquid phase balance limited. Liquid phase balance frequently leads to high recycle rates with consequently high processing costs (Chap. Costs of Agglomeration). For the control of granulation see [318, p. 280]

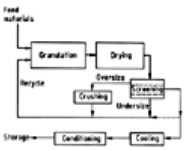

Figure 25. Granulation loop

Figure 26. Drying drum, showing granulated product

Courtesy of BASF Aktiengesellschaft

The *granulation efficiency* often is defined as the mass fraction of particulate material that leaves the granulator as finished product, that is, with grain sizes in the desired range (assuming 100 % sieve efficiency) [317]. It is also possible, however, to state the granulation efficiency as the mass fraction of finished product at the dryer outlet [319]. This definition allows for some regranulation in the dryer. The mass ratio of material not withdrawn (recycled material) to product is often referred to as the recycle ratio. For example, a 20 % granulation efficiency implies a recycle ratio of 4 : 1 if other losses are disregarded.

The *recycle ratio* is important to the process of granulation. Recycle is necessary because, after the product has passed once through the granulator, a certain quantity of particles lies outside the desired region of the grain-size spectrum (off-size material) and must be run through again. For a given mixture and a given temperature, optimal granulation takes place only within a narrow range of the solid-to-liquid ratio. The quantity of recycled fines depends not only on the chemical properties of the materials but also on the water content of the slurry and on the granulation device [320]. Recycle is also needed to generate nuclei for agglomeration and to stabilize the granulation conditions in the granulator.

The quality of the granules is influenced by the following factors:

- Type and fineness of the feedstock
- Moisture content of the granules
- Surface tension of the wetting liquid and wettability of the particles
- Mode of motion in the granulator
- Inclination and speed of the granulator
- Type and properties of the binder

Granulation Processes. Granulation processes can be classified by the nature of the feed materials to be granulated (i.e., granulation of solids, slurries, melts [308, pp. 250 – 260], [313]) and by the type of granulation equipment used. The most important types of equipment for granulating fertilizers are shown schematically in Figure 27. For the most important fertilizer materials, both straight and multinutrient, Table 32 offers an overview of the main commercial granulation processes, along with further possibilities.

Table 32. Equipment for granulation of fertilizer materials * [317]

Fertilizer material	Granulation equipment: main commercial techniques are in boldface, while possibilities are in normal typeface
Calcium nitrate	prilling, flaking, pan granulator [322], drum granulator, compaction, pugmill/blunger [320]
Ammonium nitrate	prilling [308, p. 103], [323], cold spherodizer [323], [324], pan granulator [322], [323], [325], [326], drum granulator, fluidized-bed granulation [327], [328], spouted-bed granulation, TVA falling-curtain drum granulation
Calcium ammonium nitrate	prilling [309, p. 195], pugmill/blunger [308, p. 104], [309, p. 196], drum granulator, pan granulator [322], [329], hot spherodizer, cold spherodizer, fluidized-bed granulation [327], [330], spouted-bed granulation, TVA falling-curtain drum granulation
Ammonium sulfate nitrate	pugmill/blunger [309, p. 205], drum granulator [331], prilling [309, p. 205], pan granulator [332]
Ammonium sulfate	crystallization, hot spherodizer, compaction, pipe reactor-drum [333]
Urea	prilling [323], [328], [334] cold spherodizer [323], [324], pan granulator [322], [323], [325], [326], [335], crystallization, drum granulator [336], [337], compaction [338], fluidized-bed granulation [328], spouted-bed granulation [339], TVA falling-curtain drum granulation
Urea with ammonium sulfate	prilling [341, pp. 71 – 73], cold spherodizer, pan granulator [342], fluidized-bed granulation [340], spouted-bed granulation, TVA falling-curtain drum granulation
Superphosphate	drum granulator [343], pan granulator [309, p. 234], pugmill/blunger
Triple superphosphate	drum granulator [308, p. 191], [344], pan granulator [309, p. 348], pugmill/blunger [308, p. 191], [344], compaction
Monoammonium phosphate	drum/ammoniator-granulator [311, pp. 6 – 8], [345], [346], pugmill/blunger [311, p. 30], [347], prilling [311, p. 8], [347], crystallization, compaction
Diammonium	ammoniator-granulator [311, pp. 6 – 8], [345], [346], [348], crystallization, pugmill/blunger [346],

phosphate	[347], [348], compaction
Ammonium polyphosphate	ammoniator-granulator[311, p. 217], pugmill/blunger [311, p. 217]
Nitrophosphate (NP)	hot spherodizer, prilling [349], [350], pugmill/blunger [320], [351], pan granulator [332], [351], drum granulator [351], [352], fluidized-bed granulation
Potash	compaction [353], crystallization
PK	drum granulator [343], pan granulator, pugmill/blunger, fluidized-bed granulation [354]
Nitrophosphate (NPK)	hot spherodizer [324], [355], [356], prilling [349], [350], pugmill/blunger [320], [346], [355], [356], pan granulator [357], drum granulator [345], [346], [351], [355], [356], fluidized-bed granulation, compaction [317]
Compounds on ammoniacal base	drum/ammoniator-granulator, pugmill/blunger, hot spherodizer, prilling
Compounds with ammonium nitrate	ammoniator-granulator, pugmill/blunger, hot spherodizer, drum granulator, prilling
Compounds with urea	drum/ammoniator-granulator, pugmill/blunger, hot spherodizer, compaction, prilling
Compounds with micronutrients	ammoniator-granulator, pugmill/blunger, hot spherodizer, drum granulator, compaction

* Compaction is illustrated in Figure 34. The equipment for the other processes is illustrated in Figure 27.

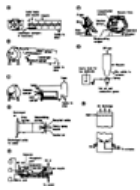

Figure 27. Granulation equipment [321]

A) Pugmill (blunger); B) Rotary drum; C) TVA ammoniator – granulator drum; D) Spherodizer process; E) SAI-R drum granulator; F) Inclined pan granulator; G) Fluidized-bed granulator/drier; H) Air-cooled prilling tower

5.2. Granulator Feedstocks

Granulation of Solids with Water or Aqueous Solutions. A solid phase (dry mixed nongranular or powdered material) and a liquid phase or steam (granulation aid) are required. Steam is discharged under the bed of material at the feed end, and water is sprayed on the bed through spray nozzles. For each mixture there is a percentage of liquid phase at which granulator efficiency is optimum. The higher the temperature, the less water and hence less drying is required [308, p. 251]. The system is granulation-efficiency controlled. Granulation takes place by agglomeration of the particles. The granulation efficiency is high and the recycle ratio is low (roughly 1 : 1). Examples are the granulation of superphosphate (with and without simultaneous ammoniation), the granulation of superphosphate in the presence of $(NH_4)_2SO_4$ and K_2SO_4, and the granulation of monoammonium phosphate together with other nutrients [313] to yield high-analysis formulations. While the granulation of solids has proved to be a flexible and economical process, it has the drawbacks of diminished quality as to physical properties and appearance. What is more, the P_2O_5 component (monoammonium phosphate or superphosphate) has to be prepared in a separate plant [313]. For an example of granulation with a solid P_2O_5 component, see [358], [359]; for granulation of NPK fertilizers containing urea, see [360].

Slurry Granulation. The materials to be granulated are in the form of a slurry, usually derived from reaction of sulfuric, nitric, or phosphoric acid with ammonia, phosphate rock, or a combination thereof. In some process modifications, solid materials may be added to the slurry before or during granulation (Fig. 28). Slurry granulation is liquid-phase controlled. Usually a thin film of a slurry having the fertilizer composition is sprayed onto small solid particles. The granules are built up in layers (layering process). The process is mainly controlled through the recycle and the slurry water content (the recycle ratio may be 5 : 1 or more [313]). Granulation is aided by various impurities (Al/Fe compounds, organic substances); see [320]. Drying can be combined with granulation into one processing step. A modification of the slurry process is the Spherodizer process developed by C & I Girdler (Section Drum Granulator). Slurry granulation is widely practiced in Europe for the production of N, NP, and NPK fertilizers. In the United States, the process has been modified so that acids, phosphoric and/or sulfuric, or partly neutralized acids are completely ammoniated in the granulator (ammoniator – granulator, Section Drum Granulator) [361]. For example, $(NH_4)_2SO_4$ can be granulated in a drum by this method [414].

Figure 28. Slurry granulation process [317, p. 23]

Granulation with solutions or slurries includes fluid-bed spray granulation (mechanism of agglomeration [362], [363]) and spray drying. In the continuous fluid-bed spray granulation process, solutions, suspensions, or melts can be converted to a granular product in a single step [364]. In contrast to spray (flash) drying, this process can be made to yield granules with a particle size of up to 5 mm [365]. The liquid for granulation is sprayed through nozzles located in or above the fluid bed onto the particles, which comprise comminuted oversize or undersize from the cyclone separator. Warm air in the fluid bed promotes the drying of the particles, and the sprayed particles increase in size. If melts are sprayed into the bed, cold fluidizing air carries off the heat of solidification. Fluidization is accomplished by blowing air through a plenum with a grid. Agglomeration with urea, NH_4NO_3, and K_3PO_4 has been reported [362].

Spray or flash drying represents a direct path from the liquid product to granules. The end product ranges from a powder to a fine grit. The feed liquid is atomized hydraulically, through feed nozzles, or pneumatically, with two-fluid nozzles or atomizer disks. The solution is sprayed into a tower-like vessel with a hot air stream and thus solidified into the fine granules. The dry product is removed pneumatically and collected in a cyclone system, or it can be removed with a bucket wheel at the bottom of the tower [366]. A few special fertilizer products are made by spray drying.

Fluidized-bed methods include the NSM process (Fig. 29) [327], [328], [370, pp. 277 – 288]. The granulator is a rectangular vessel with a perforated plate at the bottom to provide a uniform distribution of air. The fluid bed, which is initially made up of fines, has a height of 0.5 – 2 m and an area of several square meters. It is subdivided into chambers to obtain a narrow gradation in the end product. The granulation liquid or melt is sprayed into the fluid bed in each section by air. Heavier particles, which remain in the bottom portion of the fluid bed, can pass into the next section or to the outlet. In this way, the granular product migrates through the fluid bed from the first to the last section. The NSM plant has a capacity of 800 t/d for urea, and the production costs are less than those for prilling. For the properties of slurry granulation processes, see [367], [368]; for studies on the layering process, [369]. For fluid-bed granulation, see [363].

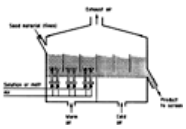

Figure 29. NSM fluidized-bed granulator [327, p. 7]

A modification combining granulation and drying in a variant of the fluid-bed process is the “spouting-bed granulator” [313], [371], [372]. The conical vessel stands on end, with a Venturi tube at the bottom, the narrowest section; into this section, either hot air propels a hot saturated solution or cold air propels a melt. The fluid bed is set up in the cone. The fast-moving particles require no perforated distributor plate. Drying or cooling takes place rapidly, and the material builds up in onion fashion. The dust collected in the cyclone is recycled (2 : 1 recycle ratio) [313]. Granulation tests on ammonium sulfate in the spouted bed have been reported [373]. Despite successful pilot tests, no full-scale unit has yet been built [371].

Melt Granulation. Spherical agglomerates produced from the melt (e.g., urea and ammonium nitrate) are called prills. These are usually obtained by spraying a salt melt or a highly concentrated solution into the top of a tower. The melt should have a very low viscosity (< 5 cP) but a high surface tension at temperatures just a few degrees above the melting point [374]. The liquid jet breaks up into droplets in the free air space. As they fall in contact with counterflowing cool air, the droplets solidify. The tower height (and thus the falling distance) and the velocity of the cool air are adjusted so that the prills are sufficiently hard when they strike the bottom of the tower [329]; tower heights are typically 45 – 55 m for ammonium nitrate [396]. The prills can be removed with scrapers or belt conveyers, or they can be cooled in a fluid-bed cooler located at the bottom end of the prilling tower [375]. Alternatively, the heat of crystallization can be carried away by spraying the droplets of melt into an oil bath. This is done, for example, with calcium nitrate [376], which is subsequently centrifuged and screened [377].

The recycle ratio in prilling is ca. 0.1 – 0.2 [315]. The prill, with a diameter of 1 – 3 mm, is usually smaller and rounder than the particle obtained by granulation. Because of the high air throughputs in the prilling space and the resulting off-gas problems, and also because of the smaller particle size, prilling has lost importance [315]. For small capacities, such as 250 t/d, granulation is economically superior to prilling; for high capacities, from 1000 t/d up, conditions determine which process is more economical [315], [378]. SCHOEMAKER and SMIT present a comparison between granulation and prilling in the manufacture of fertilizers [379]. For the prilling of NPK mixtures consisting of NH_4NO_3, $NH_4H_2PO_4$, and KCl, see [374].

For tests of oil prilling of a urea – ammonium sulfate mixture (34–0–0–9S), see [341, pp. 71 – 73], [342]; for tests of oil prilling of urea – ammonium polyphosphate mixtures, see [382]. Monoammonium phosphate can be obtained in melt form with a pipe reactor and sprayed into a prilling tower (Swift process) [347].

Multinutrient fertilizers mostly have high melting points and are very viscous [313]. One exception is a mixture of monoammonium phosphate and ammonium nitrate, which melts at a low temperature and has a low viscosity. The melt is granulated in a drum (recycle ratio 1 : 1). Depending on whether KCl is added in the granulator, formulations such as 24–24–0 and 17–17–17 are obtained. In a TVA process [380], phosphoric acid and NH_3 are reacted in a T reactor to yield an anhydrous melt; this can be granulated by itself to an ammonium phosphate/ammonium polyphosphate mixture (from 11–55–0 to 12–57–0) [313], or urea can be added to obtain a 35–17–0 or 28–28–0 granular product. If KCl is added, 19–19–19 can be produced [313], [381]. Granulation takes place in a drum or a pug mill (cf. Sections Pug Mill and Drum Granulator).

Melt-granulation processes have the advantage that a dryer, which is otherwise the largest and most expensive unit in granulation plant, can be dispensed with [308, p. 256]

5.3. Granulation Equipment

The condition for granulation is that a bed of solid particles moves, with simultaneous intensive mixing, in the presence of a liquid phase. This motion provides the particle collisions and bonding needed for granule growth. There are many types and models of granulating equipment, most of which use one of three basic intensive mixing mechanisms [321]:

1. Rotation of one or more shafts carrying staggered paddles in a fixed trough (pug mill, blunger).
2. Rotation of the whole device, such as drum or pan.
3. Movement of particles by a third phase, as by blown air in a fluid-bed granulator. In slurry granulation the third phase is usually hot air or hot combustion off-gases, which can serve as a drying medium at the same time. In this way, two processing steps in the granulation loop can be carried out in a single apparatus [365].

In order to improve pelleting conditions or pellet qualities, binding agents can be added along with the granulation liquid. The binding agents may be solid or liquid, may form films, crusts, or crystals, and may harden at standard temperature or at higher temperature [383].

Because granules can also be obtained by dry compaction, the compactor should be considered as a granulator here.

Various authors have reported data on granulators [309], [317], [321], [352], [384-388]. RIES has attempted to classify granulating equipment and processes [366], [389-391].

The granulating devices used most often in the fertilizer industry are drums, pans, and pug mills. While fluid-bed granulation has come into use in the fertilizer field, mixer – granulators and compactors are more frequently employed to form fertilizer granules. Spray drying and extrusion processes are used only for special fertilizer products.

5.3.1. *Pug Mill*

A pug mill (Fig. 27 A) consists of a U-shaped trough and, inside it, one or two shafts bearing strong paddles staggered in a screw-thread fashion. In frequency of use, two-shaft pug mills are dominant [317]. The shafts rotate at equal speeds in opposite directions in the horizontal or slightly inclined trough. The solid particles (fresh feed plus recycle) are fed in at one end of the trough and are thrown up in the middle of the trough, where they are wetted with the granulation liquid. In the trough, the paddles move, knead, and transport the moistened particles toward the discharge end. The particles can be given a better-rounded external shape either in a downstream tumbling drum or in the feed zone of the drying drum. Placement of a perforated NH_3 inlet pipe (sparger) at the bottom of the trough makes it possible to ammoniate and agglomerate the fertilizer at the same time. The pug mill is sturdy and can adapt to variable service conditions; it produces hard granules of uniform composition [321]. If the angle between the paddles and the shafts is optimized, the energy consumption can be reduced. The paddles are usually provided with a wear-resistant coating to prevent abrasion [320]. Processes have been described for granulating in a pug mill an ammonium polyphosphate melt (12–57–0), and the same melt with urea (28–28–0) [380], and the same with KCl (19–19–19) [341]. For tests on 35–17–0, see [392]. The combination of a pipe reactor with a pug mill for the granulation of NPK has been reported [393].

5.3.2. *Drum Granulator*

The drum granulator (Fig. 27 B and Fig. 30), which is the type of granulator in widest use for fertilizers, is an inclined rotating cylinder. The rotation speed is usually adjustable. For a given drum and a given granular product, there is an optimal peripheral speed that gives the highest yield of granules.

Figure 30. Drum granulator

Courtesy of BASF Aktiengesellschaft

An inclination of up to 10° from the horizontal ensures adequate movement of product toward the discharge end. Because, however, this inclination is not enough to effect classification, the discharged product has a fairly broad grain-size distribution, in contrast to the pan-granulator product (Fig. 31A).

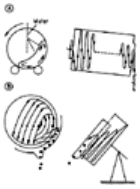

Figure 31. Schematic representation of granule development in the drum granulator (A) and pan granulator (B) [394]

Drums in which the lengthwise axis is inclined upward from the feed end to the discharge [395] give a narrower particle-size spectrum. Such an upward inclination also increases the drum fillage.

In drums currently used in the fertilizer industry, the length-to-diameter ratio is ≥1 and may reach 6 : 1. The feed end may have empirically designed distributing elements on the inside wall to spread the feed material. In the adjacent part of the drum, where the granulation liquid is fed in, a good tumbling motion should be ensured. This can be achieved with light lifting flights, but they must not lift the granulate too high. In the remainder of the drum, the pregranulated material should be tumbled to a round shape and further compacted. This is also achieved with light flights on the wall and an appropriate fillage. The fillage in the spray and tumbling areas can be controlled by means of internal ring dams. The cylinder may be either open ended or fitted with ring dams at the ends [396] to prevent overflowing at the feed end and to control the bed depth and thus residence time. Fixed or movable scrapers inside the drum or hammers or other rapping devices outside on the drum can be used to remove or reduce excessive product caking inside the drum. Some material buildup on the drum wall may promote granulation [397]. Cylindrical drums are used for continuous granulation with and without internals.

As in the case of the pug mill, recycled product (undersize) generates a moving bed of material in the drum; a slurry containing, say, 3 – 8 % water can be sprayed onto the bed [309].

Powdered feed materials (mixed and wetted in an upstream mixer if necessary to provide granule nuclei) can be granulated in the drum through spraying with water, solutions, suspensions, and highly concentrated slurries, or through blowing with steam. The bed volume should be 20 – 30 % of the cylinder capacity [398]. The recycle ratios for drum granulation are generally between 1 : 3 and 1 : 6. Optimization of these plant parameters for each product class is done by trial and error.

The drum granules are better rounded but less dense than the pug-mill granules [321]. Drums 4.5 m in diameter and 16 m long are in use in the fertilizer industry.

An important modification of drum granulation is the TVA ammoniator – granulator (Fig. 27C and Fig. 32) [308, pp. 250 –260], [399-401]. This is a drum roughly equal in length and diameter, with ring dams at the ends but no lifting flights. Ammonia reacts with phosphoric and sulfuric acids below the surface of the tumbling bed of fresh feed and recycle. The reaction generates heat, which vaporizes the water at the same time that granulation takes place. The heat is removed by injected air. The ammonia and the acids are supplied to the bed through perforated distribution pipes mounted parallel to the drum axis. The requisite bed depth is ensured by the ring dam at the drum discharge. In a modern process, a mixture of phosphoric and sulfuric acids and ammonia is neutralized in a pipe-cross reactor situated upstream of the granulating drum (Fig. 33). The slurry is then fed to the drum along with recycle. While more phosphoric acid is sprayed onto the tumbling bed, ammonia is fed into the bed [313]. In this way, NPK fertilizers can also be produced [335, pp. 44 – 48]. In the SA CROS process for monoammonium phosphate production, phosphoric acid and ammonia are mixed and reacted in a pipe reactor. The slurry is distributed over the tumbling bed together with the steam generated; no subsequent ammoniation takes place in the bed [402]. For granule improvement with an interstage pan, see [403]. The use of the pipe reactor in combination with the granulating drum for the manufacture of granular ammonium phosphates was introduced by TVA in 1973 [311, p. 45] and was later incorporated in many plants [404]. A possible improvement in the drum granulator is represented by the double-pipe granulator, which is especially well-suited to fertilizer mixtures with a high proportion of recycle (Scottish Agricultural Industries system, Fig. 27 E) [309], [405]. For example, by virtue of the high recycle ratio with corresponding residence times, a hard ammonium nitrate – ammonium phosphate mixture can be granulated.

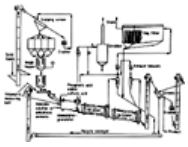

Figure 32. Ammoniator – granulator plant for NPK mixtures [317, p. 23]

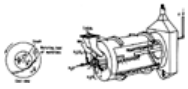

Figure 33. Location of pipe-cross reactor and spargers in ammoniator – granulator [313, p. 87]

a) Ammonia sparger; b) Phosphoric acid sparger; c) Pipe-cross reactor; d) Scrubber liquor

Reprinted by permission of John Wiley & Sons, Inc.

Another modification of the drum granulation process described is the spray-drum process (Spherodizer, Fig. 27D). In a rotating drum, preneutralized slurry is sprayed onto a dense curtain of granules cascading from baffles inside the drum. The water content of the slurry must be, say, 12 – 18 % to allow good spray dispersion [309]. During granulation, hot combustion gases flow through the drum in cocurrent [406], so that drying takes place at the same time. The dried particles are then sprayed again, redried, and so forth. The grains grow in shell fashion with an onion structure and are very hard. Spherodizer units are built in capacities of up to 650 t/d. Such an apparatus has a diameter of 4.5 m and a length of 12 m [309]. The Spherodizer, developed by C & I Girdler [407], [408], was first used on an industrial scale in 1959. The *cold* and *hot* used for the versions of the Spherodizer describe the condition of the air that flows through the drum. The cold version is used with melt feeds, especially ammonium nitrate and urea, while the hot version serves for granulation and spraying with slurries (NPK fertilizers, nitrophosphates, ammonium phosphate – nitrate, urea – ammonium phosphate) [317]. Granulation and drying thus take place in the same device. Under optimal service conditions, the recycle ratio is approximately 1 : 1.

A combination of drum granulation and fluid-bed technology is embodied in the Kaltenbach-Thuring Fluidized Drum Granulation (FDG) process [325, p. 39], [409], [410]. The technique is suitable for both melt and slurry granulation (e.g., size enlargement for urea and ammonium nitrate prills).

In a drum, the best granulation takes place at 25 – 45 % of the critical rotation speed [361], which is the rotation speed at which the weight of the granules and the centrifugal forces are in balance [386], [318, p. 204]:

$$N_{\mathrm{crit}} = 42.3\sqrt{\frac{\sin\beta}{D}}$$

D = drum diameter, m
β = drum inclination
N_{crit} = critical rotation speed, rpm

5.3.3. *Pan Granulator*

The tumbling motion of granules during agglomeration can also take place on a rotating inclined pan (Figs. 27 and 31 B).

For a given pan size, if the inclination of the pan axis to the horizontal is increased, the granules roll upward less steeply but have a longer residence time in

the pan. The granulation nuclei and the small granules initially move in the vicinity of the pan bottom. During granulation, the rotation of the pan and the force of gravity cause them to take up a spiral path. The particles grow and eventually reach the bed surface. The spiral diameter decreases continuously until the granules, finally becoming large enough, run over the rim of the pan (Fig. 31 B). Melts or slurries are often sprayed onto the bed, but water or solutions can also be used as granulation aids, and steam can be injected into the bed. If water is employed, it should be applied in the region of the largest spiral diameter [320]. Experience has shown that the spray liquid must be dispersed more finely, the finer the solids for granulation [413]. Because the overflow product has a rather uniform grain size, downstream classification can often be dispensed with. By means of an advancing and retreating scraper blade, the pan bottom can be kept fairly clean and the formation of crusts can be avoided. Here, as in the drum, some material coating the pan prevents wear and promotes the correct tumbling action [396]. The pan can also be made in the shape of a truncated cone or can have at its periphery a tumbling ring, onto which the granules fall from the pan rim; surface coating agents can be applied. Pan granulators are manufactured with diameters of up to 7.5 m [317]. Typically, the height of the rim is one-fifth of the diameter.

Concentrated salt melts of urea, ammonium nitrate, or calcium nitrate can be processed in the pan granulator; the products are easily applied fertilizer pellets, and an alternative to prills [322], [323], [383], [415].

For the production of granular triple superphosphate, phosphoric acid is added to digest finely milled crude phosphate in a granulator – mixer; this step yields a moist, crumbly product, which is directly processed in a subsequent pan to the required pellet size with the injection of steam and the addition of hot phosphoric acid [383]. For the pan granulation of urea – ammonium sulfate mixtures, see [342].

The relationship between the critical rotation speed and the pan diameter and inclination is the same as for the drum granulator [396].

In general, pan rotation speeds are $n \approx 0.6 – 0.75 n_{crit}$, where n_{crit} is the critical rotation speed; the pan axis is usually inclined at 45 – 55 ° to the horizontal [411]. According to the TVA [412], the optimal angle is ca. 65 °. The throughput of a pan granulator can be calculated roughly as follows [386]:

$$\dot{m} = k \cdot 1.5\, D^3$$

$\dot{m}$ = throughput, t/h
k = factor ca. 0.95 – 1.1 for mixed fertilizers
D = pan diameter, m

5.3.4. *The Granulator – Mixer*

While the pan granulator must be fed with powdered or pre-pelleted material, the granulator – mixer can accept friable, plastic, pasty, or crumbly feeds. If the mixing elements move at the proper speed, the material is comminuted to the desired grain sizes [383]. The disintegration of hard agglomerates is made possible by cutter rotors mounted at the sides of the mixing space [416]. Granulator – mixers are often used in batch operation, while pan granulators are run continuously. Process engineers in the fertilizer industry have also combined the two kinds of apparatus with the mixer upstream to improve product quality. The mixing vessel itself either has a fixed position or may rotate, while the movable mixing elements (e.g., mixing stars, spirals, shafts with attached vanes, plowshare mixing elements) effect intimate mixing and thus granulation by virtue of their rotation. The shape and rotation speed of the mixing elements are usually variable and can be adapted to a range of mixing and granulation jobs. Wear of the mixing elements must be expected. The mixing vessel proper can have a variety of shapes: smooth pipe, zigzag rotary pipe, pan, cone, Y, tub, and so forth. The mixer is often provided with external auxiliary heating. The liquid used for granulation can be fed to the mixing space and distributed by means of a hollow shaft [385], but suitable openings and feed pipes on the vessel can also feed in the liquid. For granulation – mixing of fertilizers, the device has a specific energy consumption of 2 – 6 kW per 100 kg of product [417]. The granulation time can be taken as 5 – 10 min. Grain sizes between 0.1 and 5 mm are achievable. The capacity is up to 30 000 kg/h per mixer [366]. RIES compares the grain-size distribution curves of fertilizer granules from granulator – mixers with those of the starting fineness [418].

5.3.5. *Roll Presses*

The size enlargement of a finely dispersed charge material by external compression is implemented in press agglomeration (Fig. 34). The charge is gripped by two counterrotating rolls, nipped in the gap, and compressed. As the void volume decreases, the material generally undergoes a two- to threefold compaction. While a charge hopper is adequate for a free-flowing material, a material that is not sufficiently free-flowing can be transported to the nip by screw feeders, with some precompaction. If the rolls are smooth, the material exiting from the nip ("shell") has a smooth surface. If the rolls have mating depressions, briquetts are produced. The shells are next reduced to the desired grain size (in a crusher or mill) and screened. The fines and oversize are recycled (Fig. 35). Rolls are manufactured in diameters up to 1.4 m and widths up to 1.5 m [390]. They may be mounted side-by-side or over-and-under.

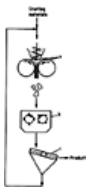

Figure 34. Press agglomeration with smooth rolls [386, p. 215]

a) Rolls; b) Crusher; c) Screen

Figure 35. Compacting of a multinutrient fertilizer [419, p. 34]

For the specific compressive forces for urea, KCl, and $(NH_4)_2SO_4$, see [353]; for data on the compaction of special fertilizers, [420], [421]; of K_2SO_4, [320]; of calcium cyanamide, [422]. For a general description of fertilizer compaction, see [423], [424]; for a monograph on roll pressing, [425]; for the principles of pressure agglomeration, [419].

5.4. Costs of Agglomeration

The costs incurred for granulation depend not just on the agglomeration properties, but — for equal or nearly equal agglomeration properties — on the size and type of equipment used in the process. For equipment and investment costs for pan granulators, drums, mixers, and roll-compaction equipment, and on hourly operating costs versus equipment size, see [426]. With regard to personnel, mixers and roll presses are considered to require one-half man (year-round, rotating shifts), while pans and drums are figured at one man each. In contrast to dry compaction (roll presses), drying costs have to be added in for mixer granulation. Up to a certain moisture content, mixer granulation with drying is quite competitive with dry compaction. For processes and costs of agglomeration, see also [427]. In comparing the granulation of solids and slurries, the investment costs are one-third higher for slurry granulation, and the operating and utility costs are likewise greater [313].

For production costs with various granulating equipment, see [428], [429]. For a cost and process comparison between prilling and granulating, see [378], [430]; for the costs of granulating monoammonium phosphate and diammonium phosphate, [311]; for the costs of NPK granulation in the Norsk Hydro nitrophosphate process, [349]; for the costs of fertilizer compaction, [431]. For economic aspects and comparative estimates of manufacturing costs, see [308, pp. 138, 266].

5.5. Bulk Blending

The mechanical mixing of single components in granular form, called bulk blending, is a special way of producing multinutrient fertilizers. Bulk blending was introduced in the USA at the beginning of the fertilizer industry [447]; it is not nearly so widespread in Europe. In the process, several of the usual starting

components, such as superphosphate, triple superphosphate, monoammonium phosphate, diammonium phosphate, urea, and potassium chloride, are combined in an uncomplicated device such as a rotating drum. The nutrient ratio can be adjusted as desired. The components are briefly mixed (up to 15 t/h) and made available to the farmer in batches that are usually loaded directly into the distributor.

The precondition for this process is that the components in the mixture be physically and chemically compatible [317], [448]. For example, urea and ammonium nitrate must not be present together, since a mixture of these is very hygroscopic and tends to deliquesce. Further, stoichiometric mixtures of urea and ammonium nitrate are sensitive to impact, and even a solution of the two can form an explosive mixture [449]. Mixtures of urea or diammonium phosphate with normal or triple superphosphate have only limited compatibility. If an aqueous salt mixture has a somewhat elevated pH and simultaneously contains NH_4^+ ion, NH_3 may be liberated.

Bulk blending has the drawback that segregation can occur during silo filling, packaging, transportation, and application. A uniform grain size or grain-size distribution is essential for reducing segregation, even if the particles differ in density. Design measures at the silo inlets and inside the silos can prevent segregation [335, pp. 37 – 39]. Drum mixing generates dust, which may necessitate cleanup measures depending on the amount of dust and the size of the mixing equipment.

The process has the advantages that the N : P_2O_5 : K_2O ratio desired by the farmer can easily be obtained; micronutrients, insecticides, and herbicides can easily be metered in, and the dealer requires less storage space.

For the production of bulk-blended urea with an appropriate gradation, see [370, pp. 277 – 288]. For the bulk-blending system in the United States, see [447], [450-452]; for the use of mechanical fertilizer mixing in Germany, see [453].

5.6. Quality Inspection

For successful handling and application, certain ranges of physical properties must be specified for the fertilizer particles. Quality control, which includes chemical analysis as well, is performed by the fertilizer manufacturer. For the determination of the physical properties of mineral fertilizers, see [308], [432-434].

Grain Shape. Fertilizer particles should have the least possible surface area, since irregular shapes lead to increased abrasion and a tendency to cake.

Grain-Size Spectrum. The diameter of ordinary commercial fertilizer grains is in the range of 0.5 – 6 mm. U.S. products generally have a somewhat finer size spectrum (primarily 1 – 3.35 mm) [317] than European products (primarily 2 – 5 mm). In exceptional cases, the product grains may be coarser, as in the case of a woodland fertilizer applied from the air (6 – 12 mm), or finer, as in the case of ammonium sulfate and crystalline mixed-salt fertilizers (< 2 mm) and other special fertilizers. The grain-size distribution is important for the intended application, for example to ensure uniform distribution of fertilizer nutrients by field equipment. The grain-size spectrum is determined by screen analysis (ISO 8397 standard screening). For the grading curves of granular fertilizers in comparison to the starting fineness, see [418].

Settled Density. The settled density is important for the sizing of packaging equipment and storage areas. For a given fertilizer grade, it should fluctuate as little as possible. The settled densities of granular fertilizers can be determined in accordance with ISO 3944; those of finely divided fertilizers (with a high content of <0.5 mm), in accordance with ISO 7837.

Compacted Density. The compacted density is generally as much as 10 % higher than the settled density. It represents the maximum bulk density that can be achieved through vigorous shaking. The compacted density can be measured in accordance with ISO 5311.

Dumping Angle. The dumping angle (angle of slope) of a fertilizer is important for the design of storage areas and for transportation. The dumping angle should be as large as possible and can be measured in accordance with ISO 8398.

Grain Hardness. The grain hardness is a measure of the fracture strength of fertilizer grains and their mechanical stability in storage. As a rule it is measured by placing grains of a definite size between two parallel plates and compressing [435].

Abrasion Resistance. The abrasion resistance is a measure of the mechanical stability of fertilizer grains moving against one another and of their stability in free fall (wear due to tumbling and dropping). Abrasion causes dusting of the fertilizers during storage operations, transportation, and application. The abrasion resistance should be as great as possible and can be determined with, for example, the TVA method [308].

Caking Tendency. One of the most important properties of a fertilizer is its storability, which can be determined through measurement of its tendency to cake. By careful drying and effective surface treatment (see Section Fertilizer Conditioning), the caking tendency can be significantly reduced and thus the storage qualities improved. This property is measured by, for example, a shear test on caked fertilizer.

Hygroscopicity. The hygroscopicity of a fertilizer characterizes its sensitivity to atmospheric humidity. Grains of a highly hygroscopic fertilizer exposed to sufficiently high air humidity take up moisture, which impairs their initial grain hardness and abrasion resistance. The hygroscopicity of a fertilizer is assessed from the water-vapor adsorption isotherm. The rate of water uptake and the critical relative humidity of the salt system can also be determined [434], [436]. Critical humidities are listed in [437]. Corresponding to the critical relative humidity is the partial pressure of water vapor over a saturated salt solution forming a very thin skin of liquid over the salt surface. If the humidity of the ambient air is less than critical, the liquid skin gives up water; if greater, the product gains moisture.

5.7. Fertilizer Conditioning

A conditioner is a material added to a fertilizer to promote the maintenance of good physical condition (flowability) during storage and handling. The use of conditioners is essential with many products, but is not required with all fertilizers. It is preferable to use other means, such as good drying, to avoid caking. Even if the fertilizer grains are dried adequately from an economic standpoint, caking may occur and impede storage, transportation, and field application.

Hardening and caking result from the crystallization of water-soluble salts and the formation of bridges between the grain surfaces during storage. The surfaces also suffer plastic deformation under pressure, and the reduction of the water vapor pressure in the joint between the new contact surfaces causes the particles to adhere to one another [438]. The reaction

$$NH_4NO_3 + KCl \rightarrow NH_4Cl + KNO_3$$

during storage may also be indirectly important in poor storage qualities [315, p. 370], [439].

The internal conditioning of fertilizers means incorporating additives in the granules before or during granulation to improve the physical properties and the anticaking qualities. Internal conditioners usually act as hardeners or crystal modifiers, for example to improve the storage properties of ammonium nitrate fertilizers. Internal conditioners inhibit or modify the effects of crystal phase inversions due to temperature changes during storage. The inversion at 32 °C can cause uninhibited ammonium nitrate granules and prills to shatter and cake. In the case of urea prills and granules, 0.2 – 0.5 % of formaldehyde or urea – formaldehyde is added to the urea melt as a hardener and anticaking additive [308, p. 301]. The addition of 1.8 % Mg $(NO_3)_2$ protects ammonium nitrate from caking. The destructure effect of the phase change at about 32 °C is avoided [440, p. 200].

External conditioning, also called coating and surface treatment, means applying to the granule surface a thin layer of powders or surfactants to reduce the caking tendency. The addition of wax and/or oil enhances the action by suppressing dust formation. This process step is carried out in a rotating drum or a fluid bed. Although coating with a fine, inert powder (kieselguhr, talcum, lime, kaolin) has long been practiced as an external form of inorganic conditioning, surface treatment with nonionic organic sealants (polyethylene waxes, paraffins, urea – aldehyde resins) and coating with surfactants to make the grain surface hydrophobic came into use later. The surfactants employed are, above all, fatty amines and sulfonates. These are also used [315] in combination with powders and waxes and/or oils [370, pp. 289 – 303]. For the use of special oils as anticaking agents, see [441]; for an example of a surfactant, [442]. For special coatings to prevent caking, see [443]. Because of the many anticaking agents and fertilizers in production, no overall recommendation can be made as to special additives for general use [438].

Intentional aging of fertilizer in a storage pile prior to bagging or bulk shipment is referred to as curing. In products that benefit from curing, chemical reactions that cause caking bonds apparently proceed to near completion during the curing period. The heat of reaction retained in the curing pile speeds the completion of the reactions. After curing there is reduced tendency for additional bonds to develop [308, p. 302]. In the manufacture of superphosphates, pile curing for about 30 d frequently is employed to improve physical properties.

5.8. Environmental Aspects

In fertilizer plants, the gaseous effluents from all equipment handling solid materials, including screens, have to be cleaned owing to their content of fine dust (and harmful gases). Although the most serious dusting occurs during the drying of granular fertilizers, dust is also formed in the granulators. These units are therefore operated under a partial vacuum. The dust in the off-gas is usually collected in cyclones and recycled. When dry dust collection is inadequate, wet separation in scrubbers, which also absorb gases like NH_3, is employed. Recycling of the scrub waters is implemented in the AZF process [444]. For NPK fertilizers, experience has shown that a soft granular product has a stronger dusting tendency than a hard one. For certain fertilizer formulations, less dust is produced from the drum than from the pug mill [445].

Gaseous effluents like ammonia, nitrogen oxides, and fluorine compounds are evolved in the production of NPK fertilizers and feedstocks. Normally these exhaust gases are scrubbed, and the resulting scrubber liquors are recycled to the process.

Liquid or aqueous effluents from fertilizer plants are usually of smaller volume compared to those vented to the atmosphere [308, p. 322]. They generally result from scrubbing equipment and can be concentrated and recycled. Spills and washings are usually collected in floor sumps and also concentrated and recycled.

For the removal of emissions from fluid-bed granulators, see [370, pp. 277 – 288]; for fluorine emission in triple superphosphate production, [446]; for a summary of environmental practices in the fertilizer industry, [308, pp. 319 – 328].

[Top of Page]

6. Analysis

Vilmos Czikkely and Titus Niedermaier

The point and purpose of fertilizer analysis is to check the declared fertilizer grade and confirm that the contents of grade-determining components, nutrients, nutrient forms, nutrient solubilities, and incidental components are as printed on the bag. In the Western countries, all fertilizers are subject to official inspection, and the laws permit only small deviations within given limits.

In nearly every country, only products with established compositions and analyses are allowed in trade; therefore, inspections must be performed during production, and factory laboratories must use testing methods in compliance with the established official techniques of analysis (Table 33)

Table 33. National and international standards for fertilizer analysis

Method	European Economic Community [454]	Federal Republic of Germany [455]	International Organization for Standardization [456]	United States (AOAC) [457]
Sampling and sample	No. L 213 on pp. 2 – 4	No. 1.1, 1.2, 1.3	ISO 3963; ISO 5306;	No. 2.1.01 – 2.1.05
preparation (see Section Sampling and Sample Preparation)	EN 1482	DIN/EN 1482	ISO 7410; ISO/DIS 7742	
Nitrogen analysis (see Section Determination of Nitrogen)				
Total nitrogen	No. 2.3; § 7.1 in No. 2.6.1; § 7.2 in No. 2.6.2	No. 3.5 (3.5.1 – 3.5.5)	ISO 5315	No. 2.4.02 – 2.4.06
Ammonia and nitrate N	2.2 (2.2.1 – 2.2.3)	No. 3.3 (3.3.2 – 3.3.4)	ISO 11 742	No. 2.4.09 and 2.4.10
Ammonia N	No. 2.1; § 7.2.5 in No. 2.6.1 and § 7.5 in No. 2.6.2	No. 3.2.1 – 3.2.6	ISO 5314, ISO 7408	No. 2.4.07 and 2.4.08
Nitrate N	§ 7.2.4 in No. 2.6.1 and § 7.4 in No. 2.6.2	No. 3.4.1 and 3.4.2	ISO 4176	No. 2.4.11 and 2.4.12
Urea N	§ 7.2.6 in No. 2.6.1, § 7.6 in No. 2.6.2	No. 3.8 (3.8.1 – 3.8.5)	ISO 8603	No. 2.4.20
Controlled-release N	EN 13 266	No. 3.10	in preparation	
Phosphate analysis (see Section Determination of Phosphate)				
Extraction	No. 3.1.1 – 3.1.6	No. 4.1.1 – 4.1.7	ISO 5316, ISO/DIS 7497	No. 2.3.01, 2.3.06, 2.3.11, 2.3.14,
Dissolved phosphate	No. 3.2	No. 4.2, 4.3	ISO/DIS 6598	No. 2.3.02 – 2.3.05, 2.3.07 – 2.3.09, 2.3.12 – 2.3.14
Potassium analyses (see Section Determination of Potassium)	No. 4.1	No. 5.1, 5.2	ISO 5317, ISO 7407, ISO 5310	No. 2.5.01 – 2.5.08
Other analyses (see Section Analysis of Calcium, Magnesium, and Trace Elements)				
Calcium		No. 6.1 and 6.2	ISO/CD 10 151	No. 2.6.01, 2.6.05 – 2.6.07
Magnesium	No. 8.1, 8.3, 8.6, 8.7	No. 7.1 and 7.2	ISO/CD 10 152	No. 2.6.01, 2.6.17 – 2.6.21
Sulfur	No. 8.1, 8.2, 8.4, 8.5, 8.9		ISO 10 084	No. 2.6.28
Micronutrients	No. 9.1 – 9.11	Chap. 8 (8.1.1 – 8.9.2)		No. 2.6.01 – 2.6.04, 2.6.10 – 2.6.16, 2.6.22 – 2.6.25, 2.6.29 – 2.6.32

6.1. Sampling and Sample Preparation

Sampling must produce a sample that corresponds to an exact, representative average of a large product batch. The most appropriate time for fertilizer sampling is during bagging or when the bulk product is being loaded into conveyances. Only when sampling at this time is impossible should samples be taken from closed bags or from a bulk product stockpile. As a rule, the sample taken should make up at least 0.1 % by weight of a batch for delivery. Fertilizer regulations also contain provisions for sampling methods. A minimum requirement is given in EN 1482.

A suitable sample divider is next used to reduce the sample quantity to some 1 – 2 kg. The final sample must be prepared for analysis by comminution, screening, and homogenizing. The smallest portion called for by the analytical technique must still be representative of the whole final sample. At the same time, this preparation must not alter the gradation of the fertilizer in such a way as to have a marked effect on typical properties such as the solubility in extractants.

6.2. Determination of Nitrogen

The most important nitrogen compounds are nitrates, ammonium salts, urea, urea – aldehyde condensates, and cyanamide. Farm-produced fertilizers (e.g., liquid manure) have most of their nitrogen in organic form (protein).

Total Nitrogen. The analytical method for determining total nitrogen is dictated by the components present in the fertilizer. For straight N fertilizers that contain only ammonia nitrogen, e.g., ammonium sulfate, it is sufficient to alkalize the sample solution and distill the liberated ammonia into an acid of known concentration in the receiving flask. For samples containing nitrate, on the other hand, the nitrate content must be reduced first. Examples of ways to reduce the nitrate are with iron or chromium powder in an acid solution, with ARND alloy in a neutral solution, or with DEVARDA alloy in an alkaline solution. Nascent hydrogen reacts with the nitrate to give ammonia in quantitative yield. Distillation from the alkalized reaction solution gives the sum of ammonia N and nitrate N.

For fertilizers that contain other forms of N besides ammonia and nitrate, the reduction step must be followed by the conversion of the other N forms (urea, cyanamide, protein) to ammonia by Kjeldahl digestion.

The final determination of total N in every case is by distillation of ammonia from the alkalized pretreated solution into a receiving flask with an acid standard solution and backtitration of the residual acid excess.

Ammonia Nitrogen Apart from Other N Compounds. In theory, the ammonia component is easily isolated from commercial mixed-salt fertilizers by distillation from sample solutions alkalized with NaOH or KOH. Urea and other organic N forms readily give additional ammonia in alkaline solution. If such compounds are present, the distillation conditions must be kept as mild as possible. The solutions are alkalized with soda, calcined magnesium oxide, or freshly precipitated calcium carbonate, and the ammonia liberated is distilled under reduced pressure at room temperature or with the help of a strong air purge.

For pure ammonium salts of strong acids, the formaldehyde method offers a quick procedure for ammonia determination. The reaction of ammonium ion with formaldehyde

$$4\,NH_4^+ + 6\,CH_2O \rightarrow (CH_2)_6N_4 + 6\,H_2O + 4\,H^+$$

liberates protons equivalent to the ammonium ions present; these can be directly titrated with sodium hydroxide standard solution.

Nitrate Nitrogen Apart from Other N Compounds. In many cases, nitrate N is determined as total N minus ammonia and urea N. The most versatile direct method is precipitation of the nitrate with nitron and final determination by gravimetry. Another sensitive method, recommended for specially formulated mixed fertilizers with low nitrate contents, is a colorimetric technique based on the nitration reactions with 2,4- or 3,4-xylenol. The nitration products give intensely colored compounds in alkaline solution.

Urea Apart from Other N Compounds. Three different methods are available for the selective determination of urea:

1. Separation of urea by precipitation with xanthydrol(9-hydroxyxanthene), yielding dixanthylurea; gravimetric final determination
2. Enzymatic hydrolysis of urea, controlled by ureases; distillation of produced ammonia into a receiving flask with acid standard solution, where it is absorbed
3. Colorimetric method (for low urea contents) based on the reaction of urea with *p*-dimethylaminobenzaldehyde

Nitrogen Fertilizers with Controlled-Release Urea – Aldehyde Condensates. Ureaform, Isodur, and Crotodur are nitrogen compounds that are only slightly soluble in cold water but are nonetheless effective as fertilizers. The delayed water solubility provides a basis for the analytical characterization of these compounds: extraction with cold water isolates the readily available N components (ammonium, nitrate, free urea). Only when the extraction residue is treated with hot water do isobutylidene diurea, crotonylidene diurea, and the short-chain components of the urea – formaldehyde condensates go into solution.

In an acid medium, especially at high temperature, condensates of urea with formaldehyde and with isobutyraldehyde tend to hydrolyze to urea and the aldehyde. In mixed fertilizers with an inherent acid tendency, there is thus a danger of altering the solubility characteristics in the course of the analysis. For this reason, the separations are carried out in buffered solutions of pH 7.5 [455].

To calculate the percent N contents, either the total N is determined in the extraction solutions or else the AOAC procedures for urea – formaldehyde condensates are carried out and a nitrogen activity index is determined.

$$\frac{\%\,\mathrm{CWIN} - \%\,\mathrm{HWIN}}{\%\mathrm{CWIN}} \times 100 = \mathrm{AI}\ \text{in}\ \%\ ,$$

where CWIN is cold-water-insoluble nitrogen and HWIN is hot-water-insoluble nitrogen.

The nitrogen activity index for urea – formaldehyde condensates is around 50; crotonylidene diurea and isobutylidene diurea have values of more than 90.

6.3. Determination of Phosphate

Because phosphorus fertilizers are rated by their contents of phosphate available to plants, various extraction methods have been devised.

Extraction with Water. Most countries define *water-soluble phosphates* as those components that go into solution at room temperature when an aqueous suspension is made with 1 g of sample to 50 mL of suspension volume (in the USA, 1 g of sample to 250 mL).

Extraction with Citrate Solution. The technique of citrate extraction determines phosphates, such as $CaHPO_4$, that are insoluble in water but soluble in complex-forming salt solutions.

The most common methods in Europe use ammonium citrate solutions with a prescribed citrate concentration. The officially approved extraction processes differ in the choice of pH. Extraction is performed in a strong alkaline medium with an ammoniacal ammonium citrate solution, or in a neutral ammonium citrate solution. Most often, extraction is carried on for 30 min at room temperature and is followed by a treatment at 40 °C; for the EEC process with neutral ammonium citrate solution, 65 °C is prescribed.

Extraction with Citric Acid Solution. Originally, the citric acid solubility was stated only for Thomas phosphate, but the lack of agreement between the plant availability of P_2O_5 and the solubility in ammoniacal citrate solution led to other straight and mixed phosphate fertilizers being rated by citric acid solubility. Treatment with 2 wt % citric acid solution (5 g of sample is agitated in 500 mL of citric acid solution for 30 min at 20 °C) dissolves phosphorus – silicon

heteropolyacids, which are hardly soluble in complex-forming salt solutions.

Extraction with Formic Acid Solution. Soft and hard phosphate rock can be distinguished by their behavior in 2 % formic acid. A 5-g ground rock sample is shaken at 20 °C for 30 min with 500 mL of the formic acid solution. In the case of soft phosphate rock, 50 – 70 % of the total phosphate dissolves; in the case of hard phosphate rock, only ca. 25 % dissolves.

Extraction for Total Phosphate Determination. A technique applicable to all fertilizers as well as crude phosphates is digestion with sulfuric acid, nitric acid, and copper sulfate. Mixtures of hydrochloric and nitric acids or nitric and perchloric acids have also been proposed for quantitative extraction.

Determination of Phosphate. After hydrolysis of polyphosphates in acid solution, if necessary, phosphate is precipitated as quinolinium molybdatophosphate and weighed. Possible substitutes for the gravimetric final determination are methods in which the precipitate, washed acid-free, is dissolved with an excess of sodium hydroxide standard solution and back-titrated with nitric acid standard solution. Reaction of the dissolved phosphate with a molybdate – vanadate solution yields a soluble, colored heteropolyacid complex, which can be determined by colorimetry.

6.4. Determination of Potassium

For almost all fertilizers that contain potassium, this component is dissolved in water. In complete fertilizers, it is advisable to perform the extraction in acid solution so that, if gypsum is present, none of the potassium will take part in double salt formation and remain undissolved as syngenite. For the determination, the potassium is precipitated with sodium tetraphenylborate from a weakly alkaline or acidic solution; the product, slightly soluble potassium tetraphenylborate, is weighed. Another proposal, mainly in the USA, is to precipitate the potassium as the chloroplatinate (AOAC). Flame photometry is also recognized as an official test method.

6.5. Analysis of Calcium, Magnesium, and Trace Elements

Complexometric methods with EDTA have come into widespread use for determining calcium and magnesium. Photometric techniques for the determination of other metal ions effective in growth (Cu, Fe, Mn, Zn, Co, Mo) have been described. To an increasing extent, however, atomic absorption spectrophotometry has gained in importance. Emission spectral analysis with plasma excitation is also used in fertilizer analysis. But this procedure is new, and as yet no official analytical specifications exist for it.

Boron must be isolated by distillation in the form of methyl borate before it can be determined. The boric acid is reacted with 1,1-dianthrimide or azomethine H, and the colored product is determined by photometry. For boron contents of 0.1 % and higher, boric acid and polyhydric alcohol (usually mannitol) are reacted to give didiolboric acid, which is determined volumetrically.

[Top of Page]

7. Synthetic Soil Conditioners

Reinhardt Hähndel and Hans Prün

Soil conditioners are substances with which the soil and substrate properties, seldom ideal for the growth of plants, can be optimized and stabilized. Their purpose is to exert biotic, chemical, or physical influences on soils in such way as to improve the soil structure and water regime [462]. Synthetic soil conditioners have well-defined composition, stable quality, and properties suited to the requirements they must meet. They may, for example, supplement or replace natural substances that are unsuitable from an environmental and ecological standpoint, are in scant supply, or must be conserved.

7.1. Foams

Foamed polystyrene and foamed urea – formaldehyde resin are used as soil conditioners; foamed phenolic and polyurethane resins are used as florists' mounting media or plant growth media [458].

7.1.1. *Closed-Cell Expandable Polystyrene Foam*

Expandable polystyrene (PS) foams consist of polystyrene pre-expanded by blowing agents at temperatures of about 100 °C and, after holding, expanded with steam. The forms used for soil conditioners are flakes, beads, or raspings [458-460], [463], [464].

Physical and Chemical Properties. Expandable PS foams have closed cells. They contain 98 % air by volume and can hold 1 – 2 % water on their surface [465], [466]. They do not, however, absorb water. Air exchange does take place [467]. Styromull foam flakes are 4 – 16 mm in diameter with a loose density of 12 – 20 kg/m^3 [459]. The permeability to water between expandable PS flakes without compression is roughly the same as that of fine gravel; under 25 % compression, it matches that of coarse sand; at 75 % compression, that of fine sand.

Expandable PS foams are largely resistant to attack by acids and alkalies; they are odorless, chemically neutral, and unobjectionable to plants. These foams have been considered resistant to degradation by soil bacteria. Physical degradation takes place under the action of ultraviolet radiation, and biodegradation, especially by soil fungi, yields carbon dioxide and water [467]. This process is slow, however, so that the foams are effective soil conditioners with a relatively long lifetime.

Raw Materials. ⟶ Styrene.

Use. Expandable PS foams in comminuted form for soil application have been described [468], [469]. Uses range from incorporation into soils and garden substrates to the conditioning of weak-structured peat [470].

Incorporated, they have soil-loosening, aerating, and draining actions. Foams are also employed as soil conditioners in crop farming and landscaping, and they have proved useful as drainage aids in slit drainage and as filter materials for covering drain pipes [458], [471], [472].

For soil aeration and structural stabilization around the roots of urban trees, expandable PS foam beads (1 – 2 mm diameter) are recommended for use with soil aeration equipment (e.g., Terralift) [473], [474].

Trade Names. Hygropor, Styromull (BASF, Ludwigshafen, Germany); Styrofoam (Dow Chemical, Midland, Mich.).

7.1.2. *Primarily Open-Cell Urea – Formaldehyde Resin Foams*

Urea – formaldehyde foams consist of condensates of urea with formaldehyde in a well-defined molar ratio. They are solid foams with rigid walls and little flexibility. The materials used as soil conditioners are mainly urea – formaldehyde foams modified for plant compatibility.

Physical Properties. Urea – formaldehyde foams can be produced with densities of 2 – 50 kg/m^3 (as water-free uncomminuted material); foams with a density of 22 kg/m^3 are used for mixing into soils and substrates. Packed moist from processing, urea – formaldehyde flakes (trade name Hygromull) weigh approximately 35 – 40 kg/m^3.

The cell walls are partly open. They are thicker, and the cell volumes smaller, the higher the density. Cell diameters lie between 100 and 300 µm [460], [475], [476]. The capillarity is low (1 – 2.5 %); the capillary elevation is 300 – 400 mm [460]. The capillary elevation for a garden substrate treated with foam is described as advantageous [477].

The grades used as soil conditioners are made water-absorbing by modification of the condensate resin and comminution to flakes measuring 4 – 20 mm. Under vacuum, the water capacity is more than 90 vol % [478]; for Hygromull flakes having a density of 22 kg/m^3, this corresponds to 4100 wt %. The water capacity can be determined appropriately by the method of DIN 11542 (1967). At atmospheric pressure, the initial water uptake is slow but rewetting is very fast [478].

Water release is uniform and goes to completion without losses due to evaporation, so that the water stored in the foam flakes can be economically utilized by

plants. The water is held in place by suction forces of pF 1.2 – 2.54 (cf. Section Soil Water – Plant Relationships; see ⟶ Ceramics, General Survey, for definition of pF). Urea – formaldehyde foams in a loose bed have the permeability of moderately fine sand; under 25 – 35 % compression, that of fine sand [479].

Chemical Properties. As the condensation product of urea and formaldehyde, the foams are synthetic organic substances. Incorporated into moist soils and substrates in flake form, they are known to be biodegradable. At pH 6 – 7, the process goes at 3 – 5 % per year [480]; at pH 3.9 – 4.1, degradation is speeded up (15 – 20 %), especially at temperatures over 40 °C. The stability falls off rapidly above 80 °C. In horticultural substrates, steam sterilization is tolerated up to a maximum of 1 h.

The components of the urea – formaldehyde foams are readily biodegradable: The urea component mineralizes to ammonium and is then converted microbially to nitrate. The formaldehyde component is liberated upon mineralization; it undergoes microbial degradation in 48 – 72 h in soil or water, even more quickly in air. Bioaccumulation does not occur [481].

A theoretical effective lifetime of some 20 – 30 years is desirable for soil conditioning [462], [482].

Raw Materials and Manufacture. A process, part of which has been patented, is used to produce foams from modified urea – formaldehyde condensation products with compressed air, foaming agents, and hardeners in fixed or mobile equipment [483].

In Germany, the maximum allowable workplace concentration (MAK) in production is 1 ppm formaldehyde; this level may be exceeded only for short periods and within limits. In the United States, the absolute MAK is 3 ppm.

Use. The use of urea – formaldehyde foams as soil conditioners has been described by DOEHLER [484]. BAUMANN advanced their use (1953 – 1968) and developed the Plastopanik technique [459]. Foams were brought into reproducible use through extensive research [458], [460], [462], [470], [485], [486].

For reasons of cost, applications cluster in the areas of garden crops, such as pot and container plants, cut flowers (e.g., carnations, chrysanthemums), flower bulbs, fruits (including strawberries), vegetables, and eucalyptus, and in landscaping, where foams are used in newly seeded lawns and woods, in the transplanting of large trees, and in sports areas and golf courses.

In terms of soil physics, urea – formaldehyde foams optimize the air and water regimes in soils and substrates. They enhance the pore volumes, the maximum, minimum, and plant-available water capacity, and the aeration, and they lower the soil density. Plants respond with improved shoot and root growth, better early growth and development rates, and increased crop yields and qualities [462].

Trade Names. Hygromull, Hygropor (mixtures of Hygromull with Styromull) (BASF Aktiengesellschaft, Ludwigshafen); Agricon (Intrinco, Vaduz, Liechtenstein).

Official licenses have been issued for Hygromull and Hygropor in Austria, Belgium, France, Luxembourg, the Netherlands, Norway, Spain, and Switzerland and also for Agricon in Saudi Arabia and other Middle Eastern countries.

7.2. Colloidal Silicates

Colloidal silicates include compounds of polysilicic acid, produced and stabilized mainly by synthetic means, with a high content of reversibly soluble silicic acid and added flocculating electrolytes [460], [462], [487-489].

Agrosil colloidal silicate consists of (1) partly dehydrated sodium silicate, precipitated (neutralized) with acids, (2) electrolytes (phosphate, sulfate), and (3) an organic additive to retard aging. Official licenses have been issued for it in Austria, Belgium, Denmark, France, Luxembourg, the Netherlands, Norway, Spain, Sweden, Switzerland, and some Middle Eastern countries.

Physical and Chemical Properties. Colloidal silicates are solid, fine-grained, and thus easily distributed substances that can be dispersed with water to form silica gels and silica sols. Silica gels, examined under the electron microscope, are described as porous; they exhibit surface activity for the addition and inclusion of water and nutrients [490], [491]. Incorporated into soils, they fill the voids between soil particles [492] and bind these together with organic and inorganic complexing agents to form water-stable crumbs [462], [493-496]. Pore redistribution takes place in this process [490], [497]. Sorptive salt buffering has been described in soil mixtures [488], [491], [498].

The chemical activity is ascribed in particular to the low-molecular-mass silica sols. In the soil, they enable the holding of phosphate in solution or its activation by desorption [490]; depending on the pH, they can irreversibly fix heavy metals [461], [462], [499-504].

Raw Materials and Manufacture. Raw materials for the production of Agrosil colloidal silicate are spray-dried and liquid sodium silicate, phosphoric and sulfuric acids, and an organic aging retardant such as urea, humic substances, pectins, or proteins. The manufacturing process is patented [505], [506].

Use. Colloidal silicates are conceived as soil supplements or soil conditioners [462], [488], [493]. To a certain extent, they make it possible to return soils affected by salts and heavy metals to agricultural use [462].

Their action as soil conditioners depends on their incorporation in the soil and their ability to take up water (precipitation, irrigation). Quantities used per are (= 10^2 m^2) are 7 – 20 kg, preferably 10 – 15 kg. In gardening and landscaping, they have an effective lifetime of 3 – 5 years; aftereffects in unworked soils have been seen for 10 and 14 years.

Direct effects due to soil conditioning include improved water and sorption capacities, the activation of plant nutrients in the soil, physical improvement of soil structure through pore redistribution and crumb formation; and the fixing of heavy metals (such as Cd, Cu, Pb, and Zn) in the presence of alkaline earths. Indirect effects include improved soil life (respiratory activity and nodule bacteria), the release of excess water (a consequence of crumb formation), and enhanced phosphate mobility. The growth of new shoots and roots is encouraged, both organic and inorganic substances are accumulated in the vegetation in greater quantities, less water is needed, wilting of grass is reduced, and fungus and bacteria resistance is increased [462], [507].

In the Swiss classification, the product is not classed as toxic.

Trade Names. Agrosil (Guano-Werke AG, BASF).

7.3. Polymer Dispersions and Polymer Emulsions

Poly(vinyl acetate), poly(vinyl propionate), butadiene – styrene copolymer, *cis*-butadiene, and various acrylic acid polymers have been described and employed as soil conditioners with actions in and on the soil [460].

Physical and Chemical Properties. The products are mainly applied to the soil surface in aqueous solutions. They cross-link the particles of the uppermost soil layer and, depending on the concentration of the active agent, form a closed film or networklike coatings. These are permeable to precipitation but diminish evaporation. Incorporated into the soil, they promote crumb formation [494].

As organic substances with a relatively high dilution, polymer dispersions are degraded relatively quickly by UV radiation when on the surface of the soil and by microbial action when in the soil. Their action is therefore limited in time, and their stability depends on ambient conditions such as cold and heat.

Raw Materials. ⟶ Polyacrylamides and Poly(Acrylic Acids); ⟶ Polyacrylates; ⟶ Poly(Vinyl Esters); ⟶ Rubber, 3. Synthetic

Use. Polymer dispersions are employed chiefly for seed protection in landscaping, and also in vegetable and flower-bulb growing. They are applied at planting time by spraying, usually along with fertilizers and soil conditioners (e.g., Agrosil) or mulches (cellulose, straw). They can also be applied after planting by area spraying or, in the case of vegetables planted in rows, by stripe spraying [508].

The quantities used depend on the product and vary between 10 and 50 g/m^2 ; the dilution with water depends on the purpose and varies between 1 : 1 and

1 : 60 (product : water).

The duration of the structure-stabilizing or protective action depends on the quantity used [509] and on environmental conditions such as weathering and insolation.

Polyacrylates and polyacrylamides are recommended for use alone or with, e.g., starch to promote water storage in soils [460], [510-515]. The waterholding effect is, however, greatly diminished by pH values higher or lower than 7.0, water hardness, and dissolved substances such as urine or soil nutrients. No satisfactory practical solution has been found for these limitations [516], but more recent generations of these so called superabsorbers are claimed to be more tolerant towards dissolved salts. Various applications to gardening substrates and flower-growing soils have been described [517].

Polymer formulations are used for protection against erosion, especially in conjunction with the sprouting of seeds. They perform an environmental function by offering limited protection to the seeds or small plants against wind and rain erosion until the plants are large enough to protect themselves. They are of increasing importance in furrow irrigation systems in North America for preventing soil erosion and increasing water infiltration into the soil. A concentration of 10 ppm in the irrigation water is necessary to achieve the desired effects [518]. Biodegradability is a prerequisite for the soil and plant application of selected polymers. Biodegradability is not assured in the case of some polyacrylates.

Trade Names. Curasol for poly(vinyl acetate) (Hoechst Aktiengesellschaft, Frankfurt); Hüls 801 for *cis*-butadiene (Chemische Werke Hüls, Marl); Aqua-Gel for polyacrylate plus starch (Miller Chemical and Fertilizer Corp.); Hydrogel Viterra for poly(ethylene oxide) (Union Carbide Corp., New York); SGP Absorbent for sodium polyacrylate plus starch (General Mills Chemicals, Inc., Minneapolis, Minn.); Soiltex G1 for polyacrylamides (Allied Colloids).

7.4. Tensides

Many tensides have been tested as soil conditioners, including ammonium laureth sulfate, ethoxylated alkyl phenols, polyoxyethylene esters of alkylated phenols, polyoxyalkylene glycols and their polymers, polyoxyethylene, polypropoxypropanol glycol butylesters, alkyl polyglycosides, sulfosuccinates, and poly (propylene oxide)s. The hydrophobic ends of the molecules of the these so-called wetting agents are adsorbed by water-repellent organic matter, and the hydrophil ends link the soil particles with water.

Use. In turfgrass culture thatch accumulation may occur due to reduced decay of the organic matter of the grass sward. Under dry conditions, this material becomes hydrophobic, and irrigation results in water losses because of poor infiltration rates and run-off [519], [520]. Spraying of wetting agents onto the affected areas at a rate of 10 – 20 L/ha increases water absorption and efficiency. Solid formulations are also available.

Trade Names. Primer (Aquatrols, USA); wetx'tra (Rhône Poulenc, France); Turf Ex (Service Chemicals, UK), Saturaid (Debco, Australia).

[Top of Page]

8. Storage, Transportation, and Application

Karl-Heinz Ullrich and Hermann Mühlfeld

Fertilizers are produced continuously, year-round, in large capital-intensive plants, but they are sold only in a few months at the beginning of and during the vegetation period. Large quantities must therefore be stored for long times. The logistic of delivering the product to the user at the proper time leads to a subdivision of storage into plant storage, dealer storage, and user storage. In this way, loading and transport are spread evenly through the year, with additional control by seasonally varied sales rebates.

The construction and operation of storage facilities depend on the size, the location, the type of fertilizer, and the danger level, all of which are dictated by the product.

8.1. General Storage Requirements

Fertilizers are classified as follows by danger level:

- Group A. Explosive fertilizers, such as those with a high ammonium nitrate content or with a low ammonium nitrate content and > 0.4 wt % organic matter.
- Group B. Fertilizers in which self-sustaining, progressive thermal decomposition (low-temperature decomposition) is possible. In the past this group comprised mainly multinutrient fertilizers which contain ammonium nitrate. Today, the ammonium nitrate based multinutrient fertilizers which are marketed are almost exclusively rated Group C.
- Groups C and D. Fertilizers that do not explode and are not susceptible to self-sustaining, progressive thermal decomposition.

The precise classification is set forth by law [521]. With its technical regulations, the German regulation concerning hazardous substances (*Gefahrstoffverordnung*) prescribes in detail how to construct and to operate storage facilities for ammonium nitrate based fertilizers. How many regulations have to be observed depends on the classification of the fertilizer. The strictest standards apply for Group A, followed by Group B. For nonflammable fertilizers (Groups C and D), the normal storage conditions for bulk materials apply.

Most fertilizers come in granular form and readily absorb moisture. As a result, they experience caking or granule disintegration; either effect makes application more difficult. The storage areas must therefore be dry. Not only the roof and walls must be tight, but the floor must be safe against rising moisture. Care must also be taken that moisture does not get into the storage building because of too much air circulation. Doors and windows must be kept closed when the relative humidity is over 65 %. The surface : volume ratio in storage should be as small as possible.

Fertilizers containing ammonium nitrate, especially calcium ammonium nitrate, must be protected from direct sunlight and from repeated temperature changes at 32 °C, because otherwise repeated recrystallization leads to granule disintegration.

When ammonium nitrate fertilizers are heated above 130 °C, low-temperature decomposition sets in, generating gases that may be toxic, depending on concentration. Practices must therefore be employed that safely prevent the action of fire or external heating. Such practices include general fire-protection measures such as no smoking, no use of fire or open light, and no unauthorized entry. Ammonium nitrate fertilizers must not be mixed with flammable materials (sawdust, coal dust, sulfur, petroleum, etc.) or with acidic or basic substances (lime, basic slag phosphate, acid salts, etc.). Welding and burning work may be carried out only with the most stringent safety precautions. Electrical equipment must be located outside the fertilizer pile. Cables and wiring must be located at least 0.5 m above the highest possible pile level. Special care should be taken that the pile does not cover up lights, especially portable lights connected by cable to the power supply.

With regard to moving mechanical parts (belt, screw, and bucket conveyors, etc.), care must be taken that these devices cannot run hot or experience fertilizer buildup at critical points. Diesel and gasoline trucks that go directly into the storage area represent a danger due to hot exhaust gases and hot mufflers. Hardened fertilizers must be loosened mechanically, not with explosives.

Low-temperature decomposition can be detected by white or brown smoke or a piercing smell. If this process develops and is recognized promptly, the sintered reaction zone can be separated from the unreacted fertilizer. Good access is needed, and precautions are needed to keep the sizable amounts of gas evolved from being a health hazard. If safe access is not possible, the decomposition must be brought to a stop by cooling with water. If any low-temperature decomposition takes place, an alarm must immediately be sent to the fire-fighting service.

Fertilizers affect most of the conventional building materials, especially when wet. Steel, aluminum, and zinc suffer more rapid corrosion and must be protected by several coats of paint (total thickness > 180 µm). Fertilizer dust can cause brittle fracture at welds in steel. Subsequent welding must be carried out with special care.

Mineral fertilizers attack concrete more or less severely, causing scaling and corrosion of the steel reinforcement. For this reason, concrete floors contacted by fertilizer should be protected by filled coal-tar pitch or epoxy-based plastic.

Wood, especially in the new form of glued lamellae, is a suitable material. Care should be taken to use corrosion-resistant fasteners.

Plant Storage. Storage at the fertilizer plant is almost always in bulk form. Large warehouses, with capacities of up to and beyond 50 000 t for one or more grades, are employed. The total storage capacity is usually several months' production. The size, shape, and outfitting of the warehouses depend on the frequency of shipping, the number of grades, the required loading capacity, and the area available. In order to reduce the risk, some grades are limited as to amount stored in a single warehouse, which may be accomplished by dividing the building. Large quantities are stored in elevators (tower silos) only under some restrictions, because of the risk of caking and the resulting high reclaiming cost.

If enough space is available and shipments are not frequent, a flat-floor warehouse in which material can be piled more or less high against the walls is the most economical design. Trough-shaped or basin-shaped storage structures with appropriate reclaiming equipment are used when space is limited and intervals between shipments are short.

The material used for the foundation and walls is usually reinforced concrete; the roof is most often made of wood. Glued lamellar wood construction makes it possible to span wide buildings without posts. Fertilizer can be piled against the walls to a height of 10 m or more. However, if the pile is too high, the material can segregate during placement. Slides during reclaiming are also a problem if material is piled too deep. The plant warehouses generally need not be heated, since the fertilizer leaves the production facility at a temperature of 35 – 45 °C (not over 50 °C). On the other hand, overhead conveyors and loading apparatus spanning moderate distances should be heated to keep them dry.

Piling is done mainly with overhead belts at the center or side, with transverse distribution by belts, chutes, or discharge pipes. Newer facilities have a distribution control room with closed-circuit television units.

Reclaiming capacity must be several times piling capacity. Because requirements vary, several methods have been devised. For high capacity and short shipment intervals, track-borne scrapers capable of automation are used. Bridge cranes are also employed. Many wheeled tractor loaders are in service; these can be shifted from one location to another, easily replaced if damaged, and reinforced if necessary. All warehouse equipment moves material by way of transfer hoppers to conveyor belts, troughs, or bucket conveyors, which carry the material to the loading point. Crucial factors in the choice of equipment are the loading capacity required, the storage qualities of the fertilizer, the cost of structures and equipment, the safety of the operation, and the expected operating costs.

Warehouses are not usually operated on the first in, first out principle. For practical reasons, however, the warehouse should be emptied completely from time to time, since varying raw materials may result in slight color variations, while going to zero aids precise inventorying, and for other reasons.

Shipping from the Plant. Fertilizers at present are usually shipped in bulk, in self-discharging railroad cars, trucks, silo cars with pneumatic loading of elevators, and canal boats and barges.

Before shipping, many plants screen the fertilizer again to remove dust, lumps, and impurities.

Shipment of fertilizers in bags is declining. The product is usually put in 50-kg open bags (pillow bags), which are then sewed or welded shut, or in valve bags. Valve bags are better suited to stacking on pallets because of their boxlike shape. Most bags are made of plastic — extruded polyethylene tubing or woven polypropylene strips coated or lined with polyethylene. Paper and jute bags are now rarely used. The machinery industry has developed largely automated equipment for bagging and loading or palleting, and much work is still under way to improve these devices and adapt them to various requirements. Because of the associated high investment costs, bagging takes place mostly at the fertilizer plant. Palleting, which simplifies subsequent transfer of the bagged product, has gained in importance. In many countries, fertilizers are shipped in large 500 – 1000-kg bags, likewise made of woven polyethylene or polypropylene.

Dealer Storage. Fertilizer is stored chiefly in buildings, which are usually divided boxwise. Because of the caking tendency and the resulting problems with reclaiming, especially in the case of ammonium nitrate fertilizers, elevators find use only for free-flowing grades, for a limited span of time, and for quantities up to 30 t.

In building storage, special care should be taken to protect against moisture, since the fertilizer no longer has its heat from production, and it often sits for several months under cool, moist conditions. Before receiving material, the warehouse must be completely dry. Fertilizer should be placed continuously, without long breaks. The pile height should not exceed 5 m. Immediately after placement, the fertilizer surface must be carefully covered with film not less than 25 µm thick, because of the mechanical stress. Joints in the film must have an ample overlap, usually 1 m. A tight seal should be made at the walls (e.g., with wood lath). Air circulation should be minimized (closed windows and doors, airtight roof). If product is reclaimed in stages, the cover must be restored each time.

A belt conveyor is suitable for placement of fertilizer. Mechanical abrasion restricts the use of screw conveyors (not over 5 – 6 m, 100 rpm).

User Storage. The interim storage of mineral fertilizers by the consumer is declining to the same extent that the dealers become more willing to undertake this kind of storage. Bulk fertilizer is kept mainly in flat storage boxes, just as in the trade, and the boxes are similar to those used for dealer storage. Elevator storage is practiced on a very limited basis.

Bagged product can be stored by the user without major precautions. To simplify transfer operations, between 20 and 30 50-kg bags are assembled on a pallet, with or without shrink-wrap; in some countries, large bags (0.5 – 1.0 t) are also employed. Bagged fertilizer can be stored outdoors for a limited time under certain conditions (no temperatures over 32 °C, white UV-stabilized covers).

Transportation. Group B fertilizers can generally be transported without special precautions. Only in exceptional cases are they considered hazardous goods as defined in the transportation regulations. Specific provisions for Group B, as well as Group A fertilizers, are set forth in international regulations for various kinds of carriers. The most important rules are recognized by nearly every industrial nation and have been incorporated into the national regulations.

Sea transport: IMDG (International Maritime Dangerous Goods Code)
Rail transport: CIM (International Convention on the Transport of Merchandise by Rail)
Road: ADR (European agreement concerning the international transportation of dangerous goods by road)
Air: IATA-RAR (IATA Restricted Articles Regulations)

The CFR contains rules for all modes of transportation in the USA.

8.2. Application

Solid mineral fertilizers are generally applied to agricultural sites in the form of granules, coarse or fine crystals, and powders. Fine crystals and powders are becoming less and less important, with the exception of lime which is used almost exclusively in powdered form.

To achieve a good nutrient efficiency and to avoid environmental pollution, the equipment used for the application of mineral fertilizers has to meet the following requirements:

1. Exact calibration of the amount of fertilizer applied and uniform distribution, both in the direction of travel and in the transverse direction, and independent of traveling speed and application rate.
2. In order to attain a high area of application, the effective band must be wide and the distributor design must permit a high rate of operation. The rate must be variable between 50 and 1000 kg/ha. Special distributors for liming must offer rates of up to 5000 kg/ha.

The uniformity of fertilizer application can be determined exactly on a test stand. Boxes with areas of 0.25 m^2 catch the fertilizer over the whole application width, and the product is then weighed. The mean is calculated from the individual values. The smaller the mean deviation from the average value, the more uniform the fertilizer distribution. Internationally, the test method is largely unified, so that the results are also easily comparable from country to country.

Application Equipment. Mineral fertilizers in the form of granules or coarse crystals are generally applied by broadcasting-type fertilizer distributors. Spreader rigs with pneumatic transverse distribution are also utilized, but their use is of declining importance. Powdered fertilizers such as lime are preferably applied by fixed-width fertilizer spreaders with mechanical transverse distribution to avoid dust formation during operation.

Broadcasting fertilizer spreaders, also called spinner spreaders, operate with working widths of 6 – 36 m. They are characterized by simple construction and easy handling, together with a high area performance. They constitute the absolute majority of all fertilizer distributors. Broadcasting fertilizer spreaders distribute the fertilizer over a semicircular to semi-elliptical area around the spreading device. The spreading width is greater than the effective working width. As the amount of fertilizer which is applied per unit area decreases towards the borders of the spreading swath, overlapping of adjacent swaths is necessary for uniform transverse distribution of the fertilizer (Fig. 36). Modern broadcasting spreaders have broadcasting patterns with extremely flat sides, so that the risk of distribution errors in the overlap zone is minimized.

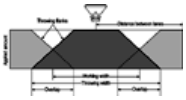

Figure 36. Broadcasting pattern for spinner spreaders

Three types of broadcasting spreaders with different spreading devices exist: single-disk, twin-disk, and oscillating spout (pendulum) spreaders (Fig. 37). The twin-disk spreader is gaining importance over the single-disk and oscillating-spout spreaders, since its basic concept is better suited for further technical development. Consequently, the spreading quality of twin-disk spreaders has been raised to a level almost matching that of pneumatic spreaders.

Figure 37. Broadcasting spreaders: single-disk centrifugal distributor (A), twin-disk centrifugal distributors with side-by-side (B) and over-and-under spinners (C), oscillating-spout distributor (D)

The spreading device in oscillating-spout spreaders is a pendulum tube that moves to and fro. Since this spreading technique allows working widths of only 6 – 15 m, whereas the large fields of modern agriculture require increasingly greater swath widths, the importance of oscillating spout spreaders is also diminishing.

Fixed-Width Fertilizer Spreaders. With fixed-width spreaders the transverse distribution of the fertilizer is achieved by means of booms, either pneumatically or mechanically. Unlike spinner spreaders, only one working width is possible.

In pneumatic fertilizer spreaders, the fertilizer is metered at the hopper by cam wheels, fed into distributor tubes, and transported by blast air to discharge outlets along the spreader booms, which are provided with deflectors (Fig. 38). Depending on the boom length, working widths of up to 24 m are possible. When working with a pneumatic spreader, overlapping of adjacent swaths is required, as with a spinner spreader. However, due to the steep flanks of the broadcasting pattern, the required overlap is much smaller (Fig. 39). Pneumatic spreaders can be used for the application of granulated and crystalline mineral fertilizers, if possible free of dust.

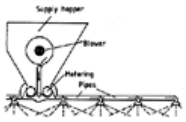

Figure 38. Schematic of a pneumatic fertilizer spreader

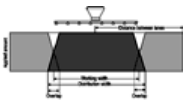

Figure 39. Broadcasting pattern for pneumatic spreaders

Spreaders with mechanical transverse distribution are generally equipped with a screw or auger (auger-type spreaders) Fertilizer flowing from the hopper is conveyed to the driven auger by various means and is then transported to the spreader tubes with adjustable discharge spouts (Fig. 40). Depending on the boom length, working widths of 6 – 8 m are possible. Auger-type spreaders are preferably used for the application of powdered fertilizers. Boom length and working width are identical. For uniform fertilizer distribution, exact driving in parallel runs is indispensable (Fig. 41).

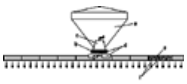

Figure 40. Schematic of a spreader with mechanical transverse distribution

a) Supply hopper; b) Elevator belt; c) Slide valve; d) Feed auger; r) Outlet

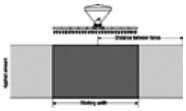

Figure 41. Broadcasting pattern of a spreader with mechanical transverse distribution

Electronically Controlled Spreading Equipment. Electronic control units are increasingly being used to monitor and to improve the metering and distribution of mineral fertilizers. Thus, the correct amount of fertilizer can be automatically discharged when the traveling speed of the spreader is changed, and the amount of applied fertilizer can be displayed, based on permanent monitoring of the fertilizer discharge.

Site-Specific Fertilizer Application Guided by GPS. Until now farmers' fields generally received a uniform fertilizer rate. With the steadily increasing size of individual fields this practice is bound to change, as increasing field size also implies increasing heterogeneity of the soil and thus of the crop yield. For economical and environmental reasons, a future development towards the site-specific management of large-sized fields is imminent. Under this aspect, the global positioning system (GPS) is currently being tested under practical conditions for its suitability for crop management, including fertilizer application. GPS allows the exact position in the field to be determined and relocated, regardless of weather, location, and time. On this basis, it should be possible to apply fertilizers site-specifically in accordance with the existing soil and crop variations within a field. The complete technology, from yield recording by the harvester to nutrient application with the fertilizer spreader, has yet to be further developed. This will probably take another 5 to 10 years.

[Top of Page]

9. Environmental Aspects of Fertilizer Application

Wilfried Werner

Fertilizers are used to enhance soil productivity or fertility (amelioration fertilization) or to maintain it at an economically and ecologically acceptable level (maintenance fertilization). The task of fertilizers is to replace the nutrients lost through the harvest and by other causes: irreversible fixation in the soil,

leaching, escape in gas form, etc. Along with the positive effect of fertilizer application, there are also negative effects, especially with improper use.

Up to the beginning and middle of the 1970s, environmental problems triggered by fertilization practices were subject to local discussion only; examples were the buildup of heavy metals in soils and crops when sewage sludge was applied, the buildup of nitrates in the groundwater where vegetables were raised in the catchment areas, and the eutrophication of surface waters due to phosphates. Only later was the partly regional character of local environmental problems acknowledged.

From the early 1990s onwards, increasing emphasis was also placed on global aspects of the impact of fertilization on the environment, i.e., the influence of agricultural production in general and fertilization in particular on the increased emission of greenhouse gases. The awareness of the mutual relationship between surplus production of agricultural products and environmental problems within the EU has continuously grown over recent decades. In particular the interest in mineral emissions from agriculture has increased substantially, both in policy making and research and in practical farming.

9.1. Nitrogen

Of all plant nutrients, nitrogen is the most effective in economic terms, but in ecological terms the most problematic. The plant absorbs nitrogen mainly as nitrate (NO_3^-) but partly as ammonium $\left(NH_4^+\right)$. The plant's ability to assimilate larger molecules with organically bound nitrogen is limited.

Problems with nitrogen arise from a local or regional excess of nitrogen, regardless of its origin. Several cases, which are mostly of a regional character, are described first, then solutions are listed. Figure 42 summarizes the environmental aspects of nitrogen.

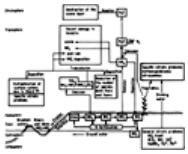

Figure 42. Agricultural nitrogen as an environmental factor

9.1.1. *Ground Water*

Various sectors of the economy — agriculture, forestry, water resources and wastewater management, transport, energy, and industry — are involved in ground water and surface water problems having to do with nitrogen. With regard to ground water, the quality of drinking water is the prime concern. A reduction in the maximum allowable nitrate content in drinking water within the EU, from 90 mg/L to 50 mg/L, has made the nitrogen problem more acute.

Until 1970 – 1975, virtually the only negative aspects considered with respect to nitrate were methemoglobinemia and the nitrate – nitrite – nitrosamine problem. Later, a number of other reactions triggered by nitrate and sulfate in the aquifer or during transport to it have come under discussion. Depending on the presence of aerobic or anaerobic conditions, the intensity of NO_3^- (and SO_4^{2-}) application, and the reserves of microbially available carbon sources, the reactions listed in Table 34 can yield the following compounds:

HCO_3^- or $Ca(HCO_3)_2$	water hardness
N_2O	stratospheric ozone break-down
NH_4^+;	undesirable, toxic, or
H_2S;	foul-smelling
Fe^{2+}, Mn^{2+}	
$Fe(OH)_3$	ferric incrustation of wells

If available carbon sources are not present, or exhausted, nitrate (or sulfate) can break through into the ground water.

From around 1980 up to now increasing emphasis was placed on ground water as prime source of diffuse nitrogen pollution of surface waters [525], [526].

Table 34. Sequence of environmental problems in the presence of excessive nitrate and sulfate in the pedosphere and hydrosphere

Oxygen and glucose in soil and ground water	Reduction processes	Energy released per mole of glucose, kJ	Problematic parameter: Input	Problematic parameter: Output *
Decreasing O_2 supply; Decreasing supply of available organic carbon	*Catabolic reduction of O_2, NO_3^-, and SO_4^{2-}*			
	Respiration			
	$C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O$	2876		
	Denitrification			
	$5C_6H_{12}O_6 + 24NO_3^- + 24H^+ \longrightarrow 30CO_2 + 42H_2O + 12N_2$	2712		
	$C_6H_{12}O_6 + 6NO_3^- + 6H^+ \longrightarrow 6CO_2 + 9H_2O + 3N_2O$	2578	NO_3^-	N_2O
	Nitrate respiration			
	$C_6H_{12}O_6 + 12NO_3^- \longrightarrow 6CO_2 + 6H_2O + 12NO_2^-$, $NO_2^- \xrightarrow{H^+} NO_x$	1946	NO_3^-	NO_x
	Ammonification			
	$C_6H_{12}O_6 + 3NO_3^- + 6H^+ \longrightarrow 6CO_2 + 3H_2O + 3NH_4^+$	1817	NO_3^-	NH_4^+
	Sulfate respiration (in the absence of NO_3^-)			
	$C_6H_{12}O_6 + 3SO_4^{2-} + 6H^+ \longrightarrow 6CO_2 + 6H_2O + 3H_2S$	486	SO_4^{2-}	H_2S
Supply of available organic carbon exhausted	*Inorganic reduction of NO_3^-*			
	Dentrification with pyrite			
	$5FeS_2 + 14NO_3^- + 4H^+ \longrightarrow 7N_2 + 10SO_4^{2-} + 5Fe^{2+} + 2H_2O$		NO_3^-	SO_4^{2-}, Fe^{2+}
	Dentrification with soluble iron			
	$10Fe^{2+} + 2NO_3^- + 24H_2O \longrightarrow N_2 + 10Fe(OH)_3 + 18H^+ + N_2$		NO_3^-	$Fe(OH)_3$
	Dentrification with elemental or organic sulfur			
	$5S + 6NO_3^- + 4H^+ \longrightarrow 3N_2 + 5SO_4^{2-} + 4CO_2 + 2H_2O$		NO_3^-	SO_4^{2-}

* Any reaction producing CO_2 is able to increase the concentration of $CaHCO_3$.

Causes of Nitrate (and Sulfate) Damage. The following cases have been described in the Federal Republic of Germany:

1. Ground water lowering, plowing of pastureland, intensive cropping with N fertilization [527]
2. Intensive cropping with vegetables and/or intensive stockbreeding with appropriate N fertilization (Fig. 44) [528-532]
3. Intensive agricultural land use — commercial gardening, animal husbandry — with appropriate N fertilization [533]
4. Commercial gardening, intensive white asparagus cultivation, overuse of mineral fertilizers (Fig. 45) [534], [535]
5. In the Netherlands, the ground water in regions with intensive stockbreeding (the provinces of Geldern and Brabant) displays not only elevated nitrate levels but often high ammonium values at depths between 10 and 25 m [536]

Recent surveys on the nitrate concentration of the ground water in Germany resulted in quite different frequency distributions for a representative overall set of measurement points and a special set of measurement points restricted to agricultural “loading areas” according to EU Directive 91/676. In the representative survey, only 10 % of the samples exceeded the EU threshold concentration for drinking water, whereas in the near-surface groundwater of so-called agricultural-loading areas this was the case in more than 60 % of the samples (Fig. 43).

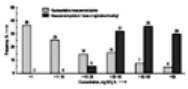

Figure 43. Frequency distribution of NO_3^- concentration in ground water in Germany [537], [538]

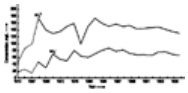

Figure 44. Nitrate and sulfate in the raw drinking water of Mussum, North Rhine-Westphalia, Federal Republic of Germany, 1910 – 1996 [532], [539].

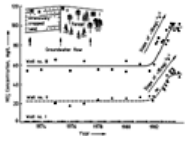

Figure 45. Nitrate concentration in water from three wells of Bruchsal Waterworks, 1973 – 1983 [540]

The insert shows the location of wells no. I – III.

A distinction should be made between acute and chronic nitrate problems. Extreme overfertilization or pastureland plowing can cause sudden, acute nitrate damage in soils threatened by leaching (sandy soils, high water table, low or declining denitrification capacity). For example, years of overapplication of N (mainly for asparagus, with more then ten times the N depletion by the crop) in the drinking-water catchment area of the Bruchsal (Federal Republic of Germany) waterworks overloaded the denitrification capacity and caused a jump in nitrate in well waters (Table 35, Fig. 45) [535], [540].

Table 35. Denitrification potential and nitrate nitrogen concentration 2 – 3 m and 3 – 4 m under the surface of forest (pine or alder) and farmed land (cereals or asparagus) at Bruchsal (Federal Republic of Germany) in February, 1986

	Forest near well number I		Intensively cropped land near well number III	
	Pine	Alder	Cereals	Asparagus
2 – 3 m under the surface				

Denitrification potential *, kg/ha	427	583	67	21
Nitrate nitrogen, kg/ha	6	18	23	56
3 – 4 m under the surface				
Denitrification potential *, kg/ha	238	258	50	17
Nitrate nitrogen, kg/ha	4	3	50	60

* The denitrification potential is the concentration of water-soluble carbon, as glucose, multiplied by 10/27, which corresponds to the stoichiometry 5 $C_6H_{12}O_6$ + 24 NO_3^- (see Table 34).

Over a long period, an acute nitrate problem may become chronic. Even under economically optimal fertilization matched to base outputs, elevated ground water nitrate levels can come about [535], [541], [542]. The base outputs arise through leaching of water-soluble nitrogen-bearing substances. The unavoidable nitrogen base outputs, primarily in winter, are between 20 and 40 kg ha^{-1} a^{-1} [543-545]. A chronic nitrate problem is much more severe than an acute one with respect to detection and propagation, as well as the cost and the success of rehabilitation measures.

The causes of damage can be classified as follows:

1. Plowing of pastureland, forced in part by prior ground water lowering. In three to four years, this process robs light soils of 6 – 7 t of organic nitrogen through mineralization and leaching out of the root zone [546], [547]. In heavy soils, the process takes much longer. The nitrate formed by mineralization is first denitrified by microorganisms, with the simultaneous degradation of the microbially available carbohydrates. The denitrification of 6 – 7 t of nitrate N requires 16 – 19 t of such carbohydrates. In this way, the available organic matter soon becomes the limiting factor, and the denitrification capacity of the soil is exhausted. In soils with shallow foundation, there is thus a danger that the N output of a nearby intensive farming operation will appear almost quantitatively in the ground water. Depending on how close the wells are, the raw water may show elevated nitrate levels for decades [548].
2. Inappropriate crop rotation. Examples of poor crop rotation include high fractions of fallow or semifallow land (e.g., corn [maize] or sunflowers), summer grains without or with inadequate (winter) intercrops, and large fractions of legumes, especially if they are plowed in too soon [545], [549]. These practices usually result in high leaching losses, chiefly in winter.
3. Corn rotation with excessive liquid manure. In regions of large-scale stockbreeding, for example, in parts of the Netherlands, Belgium, Germany, and France, excessive amounts of liquid manure were formerly — and in many forms are still — applied.
4. Incorrect nitrogen fertilization. The problem is usually excessive fertilization, poor timing, and unsuitable fertilizer form, usually not an inappropriate application method.

The danger of overfertilization with mineral fertilizers is especially great in commercial gardening and in single crops with low N depletion (orchards, vineyards). Vegetables are usually harvested in the principal growth period of the vegetative phase. Leafy vegetables, above all, offer good market quality only if a good nitrogen supply is available at the time of harvesting. Thus high residual levels of readily soluble nitrogen remain in the soil. Here again, it often happens that too much fertilizer is applied. The same holds for vinegrowing, especially over permeable limestone. In sugar beet farming, fertilizer nitrogen was formerly applied far in excess of the depletion. However, from 1980 until 1994 there was already a continuous decrease from 250 to 125 kg N per hectare in German sugar beet growing, on average [550].

In general, the danger of overapplication is less for mineral fertilizers than for organic fertilizers, since the nitrogen content in the mineral fertilizer, unlike that in manure, is known and constant. What is more, the mineral fertilizer can be applied more precisely with respect to both quantity and schedule. Finally, cost is an important factor in determining the rate of mineral fertilization.

Overapplication of organic fertilizers is a problem chiefly in heavily stock-oriented farming with liquid manure fertilization and in pasturage with grazing. Despite overapplication of organic fertilizers by a factor of one or two, compared to the residual nitrogen loss, mineral fertilizer N is often supplied in addition to improve yields. This is demonstrated by the total nitrogen balances of different farming types in Germany. The mean nitrogen surpluses were estimated for the financial year 1995/96 at 19 kg N per hectare for arable farms, 107 kg N per hectare for cattle farms, and 166 kg N per hectare for special pig and poultry farms [551].

9.1.2. *Surface Waters*

In common with the groundwater situation, many Western European rivers have shown a strong tendency towards increasing levels of dissolved nitrogen. As an example, in the period between 1954 and 1995 the nitrogen load, mainly as nitrate, from diffuse sources of the Rhine at the Lobith control station increased from 95 000 to 185 000 t N per hectare per annum [552].

The present share of diffuse sources in the total nitrogen input into surface waters in Germany is estimated at about 60 % [526], [553], [554], because the N input from point sources was drastically reduced since the mid-1980s due to the introduction of nitrification and denitrification treatments in sewage plants. Of the diffuse sources, the pathway via groundwater is highly dominant, accounting for roughly two-thirds. Drain water was responsible for an average input share of about 10 % of the diffuse nitrogen pollution of surface water in Germany. Hence, about three-quarters of total diffuse nitrogen input was due to leaching of nitrogen from upper soil layers, and nitrogen fertilization with both mineral and farm-produced fertilizers has an important impact on the level of leaching losses.

In connection with the nitrate N (and phosphorus) loading of rivers and wetlands, special attention should be paid to the drainage system in the agricultural area of the northern European lowland plains, with their intensive stockbreeding and liquid manure disposal. The danger of eutrophication not only by increasing phosphate inputs but also by rising nitrogen inputs into estuaries, coastal zones, and especially the tidal mud flats of the North and Baltic Sea coasts has come under discussion [555-557]. Sediments in these areas since about 1950 – 1960 display increasing contents of both P and N [558], [559].

9.1.3. *Atmosphere*

Agriculture is one source of problematic N compounds in the atmosphere: N_2O from nitrification and denitrification processes in the soil and, chiefly in stockbreeding areas, ammonia (Fig. 42).

N_2O Emissions. N_2O [10024-97-2], with an atmospheric residence time of 100 – 200 years, serves as a source of NO [10102-43-9]. In the upper stratosphere NO, along with HO radicals and halogenated hydrocarbons, contributes to the breakdown of ozone, which is associated with the danger of increased short-wavelength radiation on the earth and a correspondingly increased susceptibility to skin cancer in humans [560].

$$\begin{array}{l} O_3 \xrightarrow{h\nu} O + O_2 \\ O + N_2O \longrightarrow 2NO \\ 2NO + 2O_3 \longrightarrow 2NO_2 + 2O_2 \\ \hline 3O_3 + N_2O \longrightarrow 3O_2 + 2NO_2 \end{array}$$

Soil management is the most important agricultural N_2O source [561]; denitrification and nitrification are the responsible microbial processes [562], [563]. Intensive studies on N_2O emissions from agricultural soils carried out from the beginning of the 1990s gave average emission rates of 2 – 5 kg N_2O per hectare per annum for temporate zones [564], [565], [566], but much higher for tropical soils [567].

The total N_2O emission in Germany in 1990 was estimated by the Federal Environment Agency (UBA) at 143 × 10^3 t N per annum, with an agricultural contribution of 34 % [553].

The quantity of N_2O emitted from soils is determined by the content of available nitrogen, that is, by the level of N fertilization. Nitrogen fertilization was

responsible for 50 – 80 % of the emissions, and between 1 and 3 % of the fertilized N was lost as N_2O [568-571].

A worldwide estimate of direct and indirect emissions of N_2O from nitrogen fertilization and nitrogen fixation is given in Table 36. The direct emissions of N_2O are estimated at 1.25 ± 1 % of the applied N. The indirect emissions are ascribed to N from past years' fertilization, crop residues, from subsurface aquifers, and from recent atmospheric depositions [571]. An additional 0.75 % of the applied N will eventually be evolved from these indirect sources [572]. With increasing nitrate levels in the soil, the denitrification ratio N_2O/N_2 shifts in favor of N_2O [573]. The average share of N_2O in the total denitrification loss was assessed at 8.5 % [574].

Table 36. Estimates of direct and indirect global emissions of N_2O from application of fertilizer N (synthetic or animal waste) to agricultural soils and from soils growing biological N-fixing crops (10^6 t/a N_2O N) [577]

Region	Mineral N	Animal waste N	fixation	Total	Range
Africa	0.04	0.21	0.05	0.30	0.15 – 0.45
North and Central America	0.26	0.11	0.11	0.48	0.24 – 0.72
South America	0.03	0.22	0.09	0.34	0.17 – 0.51
Asia	0.75	0.52	0.19	1.46	0.73 – 2.19
Europe	0.27	0.22	0.02	0.51	0.26 – 0.77
Oceania	0.01	0.03	0.01	0.05	0.03 – 0.08
Former Soviet Union	0.17	0.18	0.03	0.30	0.19 – 0.57
Total	1.53	1.49	0.50	3.50	1.80 – 5.30

Ammonium-based fertilizers lead to higher N_2O emissions than nitrate-based fertilizers [569], [575]. Farm manures generally produce higher N_2O emissions, mainly due to the simultaneous application of quickly decomposable organic matter [576].

N_2O emissions due to N fertilization are mitigated by all measures to improve the efficiency of N use. The mitigation potential is estimated at up to 20 % [577].

NO_x ($NO + NO_2$) Emissions. NO_x, which has an average atmospheric residence time of 1.5 d, contributes to ozone synthesis in the troposphere:

$$NO_2 + O_2 \xrightarrow{h\nu} NO + O_3$$

The reverse reaction does not take place in the presence of carbon monoxide or reactive hydrocarbons, including methane, which is produced mainly by ruminants. Ozone can combine with unsaturated hydrocarbons from automobile exhausts to form ozonides, which react with NO_2 to yield peroxyalkyl nitrates

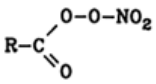

Ozone and the peroxy compounds are strong oxidizing agents with high phytotoxicity. They have been blamed for damage to vegetation in both forestry and agriculture.

The volume of NO_x emissions due to fertilization (from N reactions in the soil, nitrification more than denitrification [578]) and of plant NO_x emissions [579-581] is hard to assess. The contribution of agriculture to global NO_x emissions was estimated at 22 % (ca. 11 Tg NO_x N per annum), which consists of biomass burning (14 %) and the influence of mineral fertilization and manuring (each ca. 4 %) [582]. When the photooxidation of emitted NH_3 in the troposphere is taken into account, the figure rises to 27 %. However, NO_x is reabsorbed and metabolized by the plant, so that gross and net processes must be differentiated [583-587].

NH_3 Emissions. Ammonia emitted into the atmosphere reacts fairly rapidly (residence time < 9 d) to NH_4^+ and, after reaction with, say, SO_4^{2-}, is precipitated as ammonium sulfate, $(NH_4)_2SO_4$. Atmospheric NH_3 or NH_4^+ promotes the long distance transport of SO_2 or SO_4^{2-}. Ammonia and ammonium nitrogen forms do have fertilizing action, but in Western Europe the negative effects outweigh the beneficial ones in close-to-natural (forests, heath) and natural ecosystems (nature preserves, surface waters). Aside from the phytotoxic action of NH_3 in the area close to the emission source, deposition of NH_4^+ causes long-range damage:

1. Leaching of nutrients (e.g., K^+, Ca^{2+}, Mg^{2+}) out of the phyllosphere (leaves, needles) and the pedosphere (soil) followed by acidulation of the soil and waters [588-592]
2. Buildup of nitrogen in the pedosphere and nutrient imbalances in vegetation [589]
3. Injuries to mycorrhizae, the roots in symbiotic association with fungi [593]
4. Changes in the flora and fauna of terrestrial and aquatic ecosystems due to the promotion of nitrophilic species and the inhibition of nitrophobic species [589], [594]
5. Increased loss of N with leach water, especially in damaged coniferous forests [595-597]
6. Increased N inputs to surface waters [598], corresponding N buildups in marine sediments, especially in the North Sea and Baltic Sea coastal regions [558], [559]

Around the turn of the century (in remote regions even today) the mean annual depositions of atmospheric NO_3 N and NH_4^+ N were each ca. 1.3 kg/ha. Today, 10 – 15 kg/ha are deposited for each nitrogen form. On the basis of the filter capacity of the forest, the atmospheric N inputs may be increased by a factor of 1.2 – 2.0 (deciduous forest) or 3.0 (coniferous forest). In addition, a coniferous forest has a specific filtration capacity for NH_4^+ N but not for NH_3^- N [596], [599].

Recent measurements of the atmospheric nitrogen deposition into the soil – plant system with a new integral measurement system for wet, dry, and gaseous depositions even show annual net depositions of between 65 and 73 kg N per hectare at two locations in central Germany [600]. Clearly the earlier standard measurements underestimated the true magnitude of nitrogen deposition. Such nitrogen depositions exert a high pressure on natural or close-to-natural ecosystems because they greatly exceed the critical loads of these ecosystems.

The estimate of NH_3 emissions by agriculture is based on statistical figures of animal husbandry and fertilizer production and application, which all are multiplied by source-specific emission coefficients. Because of the different values of these specific coefficients given in the literature, all estimates of NH_3 emissions have a high range of uncertainty. An estimate for Germany (1992) resulted in the range from about 350 to 840 with a mean value of 577 Gg NH_3 N [601]. There are very large differences in the emission densities between and even within the federal states.

The estimates for Western Europe from different authors are in the range of 3.1 – 4.0 Tg/a NH_3 N [602]. As an example, Table 37 contains estimates of the anthropogenic NH_3 emissions from agriculture, industry, and other sources for 16 European countries. 90 % of the total emissions are due to agricultural activities, 74 % alone originates from animal husbandry (stabling, grazing, application of manure), and 12 % is attributed to the use of mineral fertilizers. The

total emission is 4.0 Tg/a NH_3 N ± 30 %, that is, 2.2 – 5.2 Tg/a NH_3 N.

Worldwide, including natural emissions, mineral fertilizers account for 2 – 5 % and stockbreeding 12.5 % of the total emissions of about 120 Tg/a NH_3 N [560]. The high proportion due to stockbreeding in Western Europe is accounted for by the high population density of animals and by the changeover from solid manure to slurry manure, which is less laborious but higher in NH_3 emissions. In comparison with stockbreeding, NH_3 emissions are lower even when solid mineral fertilizers high in NH_4 or amide N (ammonium sulfate, diammonium phosphate, urea) are used. This is especially true for soils where the pH and the free lime content are high or where directly incorporating the fertilizer into the soil is not possible. Liquid fertilizers with NH_4^+, because they infiltrate the soil, emit less NH_3 than solid NH_4^+ fertilizers.

Based on the specific NH_3 emission coefficients of the various N fertilizers and their application, NH_3 emission rates from mineral fertilizers in Germany are calculated by different authors to be from 2.1 to 8.1 kg per hectare per annum NH_3 N [603].

Table 37. Anthropogenic ammonia emissions from 16 European countries [602]

Country	Animal husbandry	Mineral fertilizers	Cultivated plants	Industry	Other sources	Total
Belgium/Luxembourg	99	4.2	2.3	1.1	9.2	115
Germany	504	78.4	27.1	1.5	53.1	664
Denmark	97	7.5	4.2	0.4	9.5	119
Finland	36	2.7	3.9	0.7	3.8	47
France	564	127.6	46.1	2.0	64.3	804
Greece	57	18.6	13.8	0.5	7.8	98
Great Britain	366	55.9	26.9	1.4	39.1	489
Ireland	124	12.5	8.5	0.8	12.6	158
Italy	390	64.0	21.0	3.3	41.6	520
Netherlands	200	8.5	3.0	3.6	18.7	234
Norway	32	1.1	10.5	0.9	3.1	39
Austria	62	1.6	5.3	0.4	6.1	76
Portugal	59	6.2	6.8	0.3	6.3	78
Sweden	54	2.2	5.1	0.3	5.4	67
Switzerland	48	3.4	3.0	0.0	4.7	59
Spain	270	96.1	45.8	1.8	36.0	449
Total (abs.)	2961	490	224	19	321	4016
Total (rel.)	74	12	6	0.5	8	100

Apart from the impact of nitrogen fertilization on greenhouse gases, there is growing concern about its influence on the sink capacity of soils for methane, the most important greenhouse gas after CO_2. Oxidation by methanotropic bacteria in the soil is the only biological sink for methane. An adverse effect of N fertilization on methane oxidation was first reported in 1989 [604]. Later it was found that ammonium strongly inhibits methane oxidation, whereas nitrate has only a minor effect if at all [605-607]. This is due to the competition between NH_4^+ and CH_4 for the enzyme methane monooxygenase [608].

9.1.4. *Biosphere*

Nitrogen fertilizer may produce undesirable N forms in the crop and changes in the composition of flora and fauna in nearby close-to-natural or natural ecosystems.

Undesirable N Forms in Crop. Because of the potential formation of nitrosamines, crops should not contain too much nitrate. However, the nitrosation of substances such as vitamin C [50-81-7] and α-tocopherol [59-02-9] suppresses nitrosamine formation.

Fertilization with nitrogen alone increases the content of stored N forms (such as NO_3^- N, basic amino acids, amines, and amides, and lowers the content of vitamin C in vegetative plant parts. This is more true for NH_4^+ fertilization than for NO_3^- fertilization. The content of α-tocopherol is virtually unaffected [609]. A balanced NPK application, on the other hand, can raise the vitamin C content [610], [611]. This is important above all to vegetable farming, especially with regard to the NO_3^- content arising from fertilizer, since the nitrate intake of human beings is derived about 70 % from vegetables consumed if the NO_3^- content in drinking water is low (<10 mg/L). As a result, countries in Western Europe have adopted guidelines or maximum levels for nitrate in vegetables. Limits are imposed, for example, in the Netherlands, Austria, and Switzerland. The Federal Republic of Germany has established guidelines, e.g., for fresh lettuce and spinach, 3000 and 2000 mg/kg, respectively.

Biodiversity. The diversity of both faunal and floral communities is influenced by agriculture and fertilizer use. On a site-specific local scale, fertilizer use in general and nitrogen fertilization in particular promote the growth of crops more than that of accompanying flora. In grassland, heavy use of nitrogen inhibits the growth of many herbs and legumes.

High amounts of fertilizer salts or anhydrous ammonia may have adverse effects on earthworm population by direct contact. However, the detrimental effect on the total population is very low. The greater supply of fresh organic material that becomes available when soils are raised low to high fertility by fertilization has positive effects on both earthworms and soil microorganisms.

On a global scale the use of fertilizers has positive effects on biodiversity in reducing population pressure for cultivating unsuitable fragile soils, felling of rainforests, and overgrazing, that is, by reducing or preventing soil erosion and degradation.

9.1.5. *Pedosphere (Soil)*

Nitrogen in cultivated soil (root zone, 0 – 120 cm) is an asset for agriculture, because nitrogen reserves and availability in this region are an important criterion for soil fertility. However, in close-to-natural or largely natural ecosystems, atmospheric N inputs in the forest, heath, nature preserves, lakes, and rivers can lead, over a long period, to a buildup of nitrogen in the soils; the harmless removal of this N (possible only through denitrification) is of concern to ecologists.

9.1.6. *Countermeasures*

Measures aimed at preventing or overcoming nitrogen problems due to N fertilization in all environmental zones can be divided into direct measures and indirect measures. The aim of both is to minimize the residual nitrogen in the root zone at the end of the vegetation period and hence the leaching potential.

Direct Measures in Fertilization. For arable crops such as sugar beet [550], significant progress has been made in reducing the fertilizer-related nitrogen leaching potential. To provide an economically and an ecologically accurate nitrogen recommendation, the nitrogen supply from soil reserves is assessed by methods such as electroultrafiltration (EUF) or, if not measurable, by empirical estimation. Nitrogen fertilizer is added only to account for the difference between the demand of the crop and the supply from the soil. The guideline for the plant N demand over time is the nitrogen depletion with the harvest. Because of inevitable N losses (denitrification, leaching, etc.), roughly 20 – 30 % extra nitrogen must be supplied over and above the depletion. Optimal distributing systems apply the fertilizer uniformly. Most farmers follow these principles.

The only way to accurately find the actual nitrogen demand is a nitrogen balance for a single farm or field. The ecologically acceptable N level may be higher than, equal to, or lower than the optimal fertilization level: this fact represents a potential incompatibility between economics and ecology. It is very important in this concern, that the new German “Fertilizer Utilization Decree” contains a commitment of farmers to bookkeeping, thus calculating the balances of mineral nutrients at form level.

Farm-Produced Fertilizers. The most urgent nitrogen problem in Western Europe is that the livestock population in large regions has long been above acceptable levels. To comply with clean-air and clean-water requirements, the former liquid manure regulations of some states in Germany, that still permitted using liquid manure of three live-animal units per hectare of agricultural land, an amount still far beyond what is economically beneficial and compatible with the environment, were replaced in 1996 by a federal Fertilizer Utilization Decree with some special requirements for the application of farm-produced manures, including the following:

- Limiting gaseous NH_3 losses by appropriate technical measures, consideration of weather conditions, and the incorporation of liquid manure into unplanted soils immediately after application.
- Using liquid manure in autumn after the harvest of main crops only for subsequent catch or in combination with straw incorporation and only in limited amounts (equivalent to 40 kg of NH_4 N or 80 kg of total N per hectare).
- No liquid manure application in winter from November 15 to January 15.
- Limitation of farm manure application to 170 and 210 kg N per hectare per annum on arable land and grassland, respectively, including nitrogen in animal droppings on pastures. Up to 20 % of total N in manure may be subtracted to compensate unavoidable NH_3 losses during spreading.

Mineral Fertilizers. The matching of N fertilizer application to economic and ecological requirements is much easier with mineral fertilizers than with farm fertilizers, since the mineral products — in contrast to the farm-produced ones — are defined within narrow limits as to N content, N form, and availability.

The nitrogen form is of secondary importance, since NH_4^+ and amide N are usually converted in a few days, perhaps 1 – 2 weeks, to nitrate nitrogen in agricultural soils. Nitrification inhibitors, added to NH_4^+ fertilizers to retard the conversion of NH_4^+ toNO_3^-, may afford an extra measure of safety where there is no acute ground water nitrate problem; however, they are only a partial solution to the problem [612]. (Nitrification inhibitors may promote the liberation of NH_3 from NH_4^+ fertilizers, especially when the fertilizers are not incorporated in the soil [613].)

Controlled-release forms of N may represent an improvement from the ecological point of view when they make it possible to satisfy the crop N demand over time better than can be done with conventional nitrogen fertilizers, and when no substantial amount of nitrogen remains in the soil after the vegetation period.

Indirect Measures. All practices that lead to gains in yield without additional N alleviate environmental nitrogen problems due to fertilization.

Plant Cultivation. The selection of species and varieties and the planning of the entire rotation are crucial factors in reducing economically and ecologically undesired nitrogen losses. The *evergreen system,* that is, keeping fields covered with crop plants year-round, is a desirable goal; whenever possible, the intercrops should moderate N fertilization (nitrogen harvesting). This practice also protects against erosion. Fallow land should be avoided; semifallow area, such as under corn, sunflowers, or summer grains, should be minimized.

An extraordinary reduction of nitrate loss by leaching is achieved by overhead crop irrigation, and, the extreme case, by moderate spraying to avoid frost damage and possibly the loss of a crop in an early stage.

Plant Protection. Plant protection measures, whether mechanical (especially to combat weeds and grass), biological, or chemical, generally increase yield and thus better utilize the nitrogen supply. In this way they indirectly reduce ecologically undesirable nitrogen losses.

Plant Breeding. Nitrogen-efficient plant species and varieties produce more dry matter per unit of nitrogen delivered than do inefficient species.

Soil Tillage. Tillage also has a decisive effect on ecologically undesirable N losses. Tilling as deep as possible, while still preserving soil structure, aids the availability of the nitrogen and reduces the leaching loss by creating the optimal pore volume in combination with an ordered humus regime, thus optimizing the water regime and the supply of water to the crop.

9.2. Phosphorus

Environmental problems connected with phosphate fertilizers are eutrophication and the buildup of heavy metals.

9.2.1. *Eutrophication*

The pollution of nonflowing or low-circulation surface waters (lakes, estuaries, and coastal zones, especially tidal mud flats) with plant nutrients is to be avoided on account of eutrophication. This term refers to the excessive growth of algae and aquatic plants due to too great a supply of nutrients. Because the putrefaction of these organisms requires a great deal of oxygen, the water becomes depleted in oxygen. Fish die, and the biological purification of the water ceases to function. The lake has *turned over.*

The primary limiting factor for eutrophication is phosphate. Surface waters should contain ≤ 50 µg P per liter. Nitrogen can also become a factor for eutrophication when increased biomass growth takes place.

Figure 46 compares the amounts and main sources of phosphates in surface waters of Germany from 1975 to 1995 [554]. The P input by point sources was markedly decreased by improved sewage treatment practices and by the transition to phosphate-free detergents. Consequently, the share of diffuse sources in the total P input increased from nearly 20 % to about 50 %, although the absolute diffuse input also began to fall from ca. 1990 onwards. Soil erosion and surface runoff from agricultural fields are the most important diffuse sources (Fig. 47).

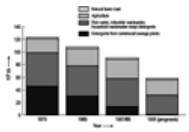

Figure 46. Changes in phosphorus input into surface waters in Germany from 1975 to 1995 [525], [526], [552], [554], [614], [615]

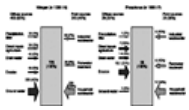

Figure 47. Nitrogen and phosphorus inputs into running waters by diffuse and point sources in Germany: prognosis for 1995 [553]

In contrast to nitrogen, phosphorus pollution via groundwater and drain water is not yet important, although vertical transport of this nutrient may create a pollution problem where the P sorption capacity of deeper soil layers is already highly saturated. This has been reported for areas with high livestock densities in the Netherlands [616] and in north-western Germany [617].

The agricultural contribution to phosphorus in surface waters on a regional scale makes the implementation of efficient countermeasures in agriculture necessary [618], [619]:

- Prevention of erosion by field arrangement (e.g., reducing the slope length), crop management (mulched seeding, underseeding, changes in crop

rotation), and soil management (tilling across the slope, improvement of soil structure and infiltration rate).
- Prevention of surface runoff from slopes by omitting the application of mineral or organic fertilizers in autumn and winter.
- Prevention of direct inputs of eroded soil and runoff water by planted margins along the banks of surface waters.
- Consequent adjustment of phosphorus fertilization to plant requirements: supply and removal must be kept in balance on soils with optimum P supply status. On P-accumulated soils, fertilization can temporarily be reduced or even omitted.

In Germany, the implementation of countermeasures is accelerated by the German “Fertilizer Utilization Decree” of 1995.

Besides economic reasons, environmental concerns were also responsible for changes in mineral fertilizer consumption since 1985. As far as phosphorus is concerned, the balance surplus showed a mean annual decrease of 7 % for the period of 1990 – 1995 [551], in particular due to reduced application of mineral P fertilizer (mean application 11.0 kg P/ha in 1995 compared to 25.5 kg P/ha in 1985).

9.2.2. *Heavy Metals Buildup*

With regard to the impact of fertilizer use on heavy metals buildup in soils, the element cadmium is of primary concern. Fertilizer phosphates produced from phosphate rock contain cadmium [7440-43-9]. The amount depends on the origin of the rock and the digestion process employed. The cadmium content of phosphate rock varies from almost zero to about 700 mg/kg P. The acidification of phosphate rock partitions the Cd between the fertilizer and byproducts of phosphoric acid production such as phosphogypsum.

The Cd input into agricultural land by phosphate fertilizers in Germany was significantly decreased during the last two decades, due to both the preference of low-cadmium phosphate rock for fertilizer production and the dramatic decrease in mean phosphorus application rates. The mean Cd content is phosphorus fertilizers is now reported with about 100 mg/kg P [620]. The mean application rate in 1995/1996 was 11 kg P/ha. Based on these figures, the average annual Cd input is 1.1 g/ha, which compares to about 3.0 g/ha in 1980.

The quantities removed by the crops and by leaching are calculated at about 2.0 g Cd/ha. The background Cd contents in agricultural fields therefore, could not be augmented by mineral fertilizers alone, even if current application rates were doubled [621-623]. However, additional cadmium sources must be taken into account. Atmospheric input from industrial air pollution is the most important source in all industrial countries (3 – 6 g Cd/ha). Manures, sewage sludges, and biowaste composts may be locally or regionally important heavy metal sources [624]. Legal regulations in Germany allow annual maximum heavy metal inputs into soils that are much higher than the realistic inputs via mineral fertilizers. For cadmium these threshold values are 18 g/ha by atmospheric depositions, 10 g/ha by biowaste composts or 16.7 g/ha by sewage sludges.

[Top of Page]

10. Legal Aspects

Günter Kluge

Since the beginning of the mineral fertilizer industry, many countries have issued regulations to protect fertilizer users. These rules deal, in particular, with nutrient contents and with safety for plants, human beings, animals, and the environment.

Fertilizer regulations are based on special fertilizer legislation or else form part of a more comprehensive regulatory scheme that governs areas such as animal feeds or pesticides. As a rule, the government department of agriculture has jurisdiction over fertilizers.

The rules on fertilizers extend only to commerce in fertilizers (offering, selling, trading); in only a few cases do they cover other areas (Table 38).

Table 38. National and international regulations concerning fertilizers

Federal Republic of Germany	Düngemittelgesetz vom 15. November 1977
	(*Bundesgesetzblatt I*, p. 2134)
	Düngemittelverordnung vom 19. November 1977
	(*Bundesgesetzblatt I,* p. 2845)
France	*Journal officiel* du 29 juin 1980 (Loi No. 79 – 595 du 13 juillet 1979 relative à l'organisation du contrôle des matières fertilisantes et des supports de culture ; *Journal officiel* du 14 juillet 1979)
Great Britain	Statuory Instruments 1977 No. 1489
	The Fertilisers Regulations
Italy	Legge 19 ottobre 1984, n. 748. Nouve norme per la disciplina dei fertilizzanti
	Gazetta Ufficiale 6. Novembre 1984
European Economic Community	Council Directive of 18 December 1975 on the approximation of the laws of the member states relating to fertilizers (76/116 EEC)
	Council Directive of 22 June 1977 on the approximation of the laws of the member states relating to sampling and methods of analysis of fertilizers (77/535 EEC)
Austria	Bundesgesetz vom 7. November 1985 über den Verkehr mit Düngemitteln, Bodenhilfsstoffen, Kultursubstraten und Pflanzenhilfsmitteln (Düngemittelgesetz – DMG)
Japan	Fertilizer Control Law of Japan. MAFF Ordinance No. 87, Juli 5, 1978
United States	Association of Official Analytical Chemists (AOAC), Washington, DC
International	International Organization for Standardization (ISO), Genève, Switzerland

Definition. Laws on fertilizers either contain a definition or simply refer to a list of fertilizers. A definition may be scientifically or pragmatically oriented. While mineral fertilizers present fewer difficulties, the diverse agents used for soil amendment are harder to define.

Examples:

Material, the main function of which is to provide plant food. (ISO)

The term “fertilizers” as used within this Law shall be defined as being any substance applied to the soil for the purpose of supplying nutrients to plants; for producing a chemical change in the soil which will contribute to the cultivation of plants; or which, when applied to the plant, will supply plant nutrients. (Japan)

Fertilizers are substances that are intended to be supplied directly or indirectly to crop plants in order to promote their growth, increase their yield, or improve their quality; excepted are . . . (Federal Republic of Germany)

In the regulations and annexes, fertilizers are systematically classified, for example, into straight and multinutrient fertilizers. The physical form, for example, in liquid fertilizers, may also figure in the classification.

Approval, Registration, Type Lists, Standards. In principle there are two ways of regulating the trade in fertilizers: type approval and individual approval of fertilizers. The European Economic Community (EEC) has chosen the following approach in the interest of harmonization of laws: A type list describes the fertilizers covered (type designation, minimum nutrient contents, expression of nutrients, method of production, essential ingredients, nutrient forms, and

nutrient solubilities). Fertilizers that meet these requirements may be marketed provided the labeling rules in all member states are complied with.

Such a system exists in similar form in many of the developed countries. Allowance is made for the fact that many fertilizers are products well known in international trade, such as urea, triple superphosphate, and NPK fertilizers. The administrative cost is low. The type list must be supplemented from time to time to take account of technical development. As a rule, effectiveness must first be proved (plant tests), as must safety for human beings, animals, the soil, and the environment.

Another approach is individual licensing for each fertilizer offered by a manufacturer or importer. An administrative action must precede the sale of any product. The government thus has a complete picture of the registered fertilizers and of the suppliers. Often, approval is granted only for a certain time. Provisional approvals are usually possible.

Because both systems have advantages and drawbacks, many countries prefer a mixed approach: type approval and registration. Standards are defined for fertilizers. Manufacturers and importers must seek registration, which requires stated tests. The certificates granted often have only a limited term.

The choice of approach is dictated, above all, by the system of laws and government in each country, in particular by the degree of supervision desired by the state.

Labeling, Terminology, Packaging. Next to approval, labeling of the product on the package or accompanying documents is the most important rule. The prescribed manner of labeling varies greatly from country to country. Before doing business, it is therefore vital to have an exact knowledge of these regulations. In general, the following information is required as a minimum:

1. type designation or name of the fertilizer
2. guaranteed content of each nutrient
3. name and address of the person responsible for sale
4. weight of product

Special rules apply to trademarks and trade names. There are special rules for packaging and sealing; most of these are restricted to certain fertilizers.

Official Inspections, Sampling, Analysis. As part of the supervision of fertilizer marketing, the sampling and analysis methods to be employed in official inspections are set forth. The variety of methods and their continual development make a selection for official tests necessary. The official checking of the guaranteed nutrient contents must be governed by rules that state whether and to what extent, unavoidable deviations are tolerated. For example, the EEC has adopted definite unified tolerances for each type of fertilizer.

Fertilizer legislation also includes the action to be taken when violations occur (fines, seizure).

Finally, besides the special laws pertaining to fertilizers there are also provisions in other areas of legislation that must be observed in the fertilizer market. The regulations on transportation and storage and the rules aimed at preventing epidemics, important for organic fertilizers, should be kept in mind.

[Top of Page]

11. Economic Aspects

Friedrich Kuhlmann, Hugo Steinhauser, Walter Brändlein, Karl-Friedrich Kummer

Consumption, production, and international trade of fertilizers are determined by technical, economic and legal factors, i.e., by output – input efficiencies, factor and product prices, and legal constraints. This applies to the single plot of an individual farm up to the agricultural area of entire nations and continents. While in developing countries consumption of fertilizers is still growing, in developed countries environmental concerns are making themselves felt, leading to legal and economic disincentives against unduly high fertilization rates. However, because of sustained world population growth, world consumption of mineral fertilizers will further increase within the next decades.

11.1. Economics of Fertilization

11.1.1. *Input – Output Relationships: The Yield Function*

Assessing economic benefits of fertilization is a relatively complex task, since the outputs of crop production processes are determined by numerous input factors, affecting the output level and being interdependent among each other. The precondition for economic considerations, therefore, is quantitative knowledge about (crop) yield functions. Although at present this knowledge is still not complete, it can safely be stated that plant growth obeys basic laws of nature, especially chemical laws. The relevant yield function for a crop — as JUSTUS LIEBIG first showed with his “law of minimum” — can thus be written as:

$$y = \min\{b_1x_1;\ \ldots\ b_jx_j;\ \ldots\ b_mx_m\} \tag{1}$$

y = amount of crop output
x_j = supply of necessary inputs
b_j = partial output – input coefficients

The x_j comprise inputs which are controllable by the farmer (seed, fertilizers, pesticides, etc.), as well as those which are not controllable (solar energy, genetic yield potential, water in rainfed agriculture, etc.). The output level y is determined by the input whose supply is minimal compared to all other inputs. Given certain supplies of all but one particular input, the output increases proportionally with increasing supply of this input, until some other input becomes the minimum factor. Further augmentation of the variable input does not lead to any further output; it would be wasted. When a fertilizer is the variable input, the relationship is called a linear response and plateau function (LRP function; Fig. 48) [627].

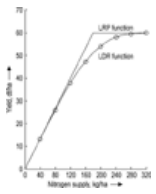

Figure 48. Types of yield function

The LRP function, however, is not compatible with results of fertilizer field experiments, investigated by means of regression analysis [628-630], [640]. Usually, these empirical investigations lead to curves shaped according to the “law of diminishing returns” (LDR function; Fig. 48). The difference between the two types of response function can be explained by the fact that available supplies of one or more of the other inputs vary from plant to plant in a plot (e.g., genetic potential, water) [641-643].

To simplify further discussion, it shall be assumed that the output (e.g., wheat) is only restricted by two inputs: the variable input x_v (e.g., nitrogen) and a given input x_g, whose amount varies randomly from plant to plant (e.g., genetic potential). Equation (1) then simplifies to:

$$y = \min\{b_gx_g;\ b_vx_v\} \tag{2}$$

Assuming a discrete distribution for the amounts of x_g, Equation (2) can be expanded to:

$$y = \sum_{i=1}^{n} p_i \min\{b_g x_{gi};\ b_v x_v\} \text{ with } \sum_{i=1}^{n} p_i = 1 \quad (3)$$

y = wheat output in dt/ha
i = number of classes in the discrete distribution of the genetic potential
p_i = probabilities of the different classes of the discrete distribution
x_{gi} = genetic potentials of the plants in the different classes of the discrete distribution, measured in dt/ha
x_v = variable supply of nitrogen, measured in kg/ha
b_g, b_v = partial input – output coefficients

If, e.g., i = 5, p_i = [0.1; 0.2; 0.4; 0.,2; 0.1], x_{gi} = [20; 40; 60; 80; 100], b_g = 1 and b_v = 0.33, increased amounts of nitrogen (in steps of 40 kg/ha) result in the plot yields shown as dots in Figure 48. Applying regression analysis leads to the LDR function also depicted. Although each plant grows according to a LRP function, the analysis of this small “experiment” suggests a law of diminishing returns.

The shape of the LDR function depends upon the variance of the spatial distribution of the input x_g within a plot. A smaller variance produces LDR curves that more closely resemble the LRP function. When the variance approaches 0, the LDR function transforms into a LRP function. This relation is especially important because modern plant breeding has reduced genetic variance of varieties (sometimes to zero), and modern land cultivation has homogenized soil conditions.

11.1.2. *Factors Controlling the Optimal Nitrogen Fertilization Level*

The above relation is of essential importance in determining the economically optimal supply of the variable input nitrogen. Assuming the LDR function of Figure 48, a wheat price of 11 €/dt, a nitrogen price of 0,6 €/kg, and a nitrogen delivery from the soil of 80 kg/ha, the optimal nitrogen fertilization rate can be derived as shown in Figure 49. The farmer tries to maximize the difference between the monetary return (LDR_1) and the nitrogen costs (NC_1). This maximum is obtained at the nitrogen input level (Nop_1) for which the slope of the monetary return curve equals the (constant) slope of the nitrogen cost line, that is, when marginal monetary returns equal marginal nitrogen costs [631].

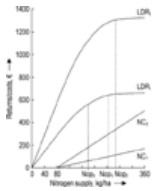

Figure 49. Economically optimal nitrogen supply in crop production, assuming the law of diminishing returns (LDR function)

Changes of output and/or input prices lead to different optimal nitrogen input levels and profitabilities. For example, doubling the wheat price to 22 €/dt results in the monetary return curve LDR_2. Tripling the nitrogen price to 1,8 €/kg results in the cost line NC_2. The first variation leads to the increased optimal nitrogen supply level Nop_2, the second variation to the decreased optimal nitrogen supply Nop_3. Generally speaking, increasing crop prices and/or decreasing factor prices result in higher fertilizer consumption, and vice versa.

If, however, the crop yield responds to variable nitrogen input levels according to an LRP function, then the farmer who tries to maximize the profitability would always provide the same nitrogen input level, regardless of price situations. Figure 50 shows monetary return functions for wheat prices of 11 €/dt (LRP_1), and 22 €/dt (LRP_2). In addition, the nitrogen cost lines, repeated from Figure 49, are depicted. Clearly, for all price situations the farmer would try to ensure the same optimal nitrogen input level Nop_1.

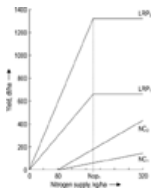

Figure 50. Economically optimal nitrogen supply in crop production, assuming linear response and plateau functions (LRP functions)

In modern agriculture, the yield functions are approaching LRP functions, so that price variations have less and less impact on levels of fertilization.

11.1.3. *Factors Influencing the Optimal Nitrogen Fertilization Level*

As can easily be derived from Equation (1), other factors, especially the levels of noncontrollable inputs for plant growth, exert a substantial influence on the optimal nitrogen input level. Firstly, the soil and climate conditions of a particular piece of land play a major role. Increased availabilities of solar energy and/or plant-available water allow for more productive fertilizer consumption, returning higher yields. As an example, Figure 51 shows the nitrogen response functions LRP_1 and LRP_2 for two parcels of land with different water conditions. The higher optimal yield y_2 of the “better” parcel can be obtained only if the nitrogen supply is increased from Nop_1 to Nop_2. Water and nitrogen are complementary inputs with respect to crop yield.

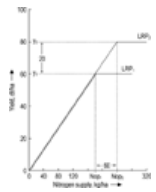

Figure 51. Effects of increased yield potentials on nitrogen requirement

Secondly, modern plant breeding successively produces varieties with higher yield potentials, enabling the plants to consume water and/or solar energy more effectively. This, of course, leads to higher yields, provided the fertilizer input level is increased appropriately. Figure 51 may be also interpreted in terms two varieties of different productivity.

Thirdly, modern crop-protection agents keep the plants healthy during the vegetation period, enabling them to consume more solar energy and water. This again leads to higher yields, provided the fertilizer input level is adjusted accordingly. The situation with and without the application of modern pestizides can also be represented by the LRP functions of Figure 51.

Under a wide range of price situations, the farmer would make full use of favorable soil and climatic conditions, as well as of the advantages of improved seeds

and crop-protection agents. As a rule, the additional monetary returns of the increased yields are much higher than the additional costs for fertilizers, seeds, and pesticides.

11.1.4. *Environmental Aspects of Fertilization*

Ecological concerns with respect to fertilization are becoming more and more important. They arise from the fact that farmers apply more fertilizers than may be consumed productively by the crops. Depending on soil conditions, some part of the surplus will not be stored in the soil to be used by the following crop, but will be leached into the underground, eventually contaminating the ground water.

However, even state-of-the-art fertilization causes some nitrogen surplus. This is due mainly to two facts: First, as discussed above, the levels of certain noncontrollable plant growth inputs are distributed randomly over a plot of land. Since the farmer does not know the exact locations, he will fertilize the whole field evenly or, in case of precision agriculture methods, at least larger parts of it. He will try to provide a nitrogen level such that the average marginal costs for the plot are equal to the average marginal monetary returns. Unavoidably, there will always be some low-yielding spots which receive too much fertilizer.

Second, under most climatic conditions for rainfed agriculture, plant-usable water supply varies from year to year. Since the farmer does not know the water supply of a particular year in advance, he will always “shoot” for a “good” year. If, for example, in a good year the higher yield y_2 in Figure 51 is achievable, and in a “dry” year only the lower yield y_1, the farmer will nevertheless always provide for the higher nitrogen level Nop_2 instead of the lower level Nop_1. The reason is simple: the expected profitability value of strategy Nop_2 is under realistic price and weather conditions much higher than that of the strategy Nop_1.

If the farmer provided only for the low fertilization level Nop_1, in a good year he would lose the monetary return of 20 dt/ha (see Fig. 51), because the insufficient nitrogen supply would limit the yield. At a wheat price of 11 €/dt, this corresponds to a loss of 220 €/ha. At the higher nitrogen level Nop_2, in a dry year 60 kg/ha nitrogen would be wasted (Fig. 51). This corresponds to a loss of 36 €/ha.

These economically induced fertilizer surpluses can be reduced by further improvements of inputs, production techniques, and prognostic information. Seed varieties with less genetic variance and homogenization of soil conditions lead to less variance of the noncontrollable growth factors within a field. Splitting nitrogen fertilization into several applications as the vegetation period develops and new information about the water supply becomes available also reduces waste of nutrients. Finally, improved long-term weather forecasts would be most valuable in reducing the uncertainty involved in fertilizing decisions.

Strict legal constraints for nitrogen application rates, however, would not be the strategy of choice. It would lead to a waste of valuable yield potential and in addition to economic losses for farmers.

11.2. World Consumption, Production, and Trade

Consumption. World consumption of the three primary nutrients — nitrogen, phosphate and potash — reached a total of 135.0 × 10^6 t in 1996/97. The previous record level of 145.6 × 10^6 t in 1988/89 was once again not reached. The consumption figures broke down as follows [625] in 10^6 t:

Nitrogen	82.9 N
Phosphate	31.1 P_2O_5
Potash	21.0 K_2O

Over the past 10 years, agricultural consumption of fertilizers experienced a deep recession, mainly due to developments in Central and Eastern Europe, as well as in the former Soviet Union. World consumption fell by 17 % from 1988/89 to 1993/94 and then recovered by 12 % up to 1996/97. Over the ten year period, nitrogen increased by 16 %, phosphate fell by 10 %, and potash fell by 19 %.

The consumption ratio N : P_2O_5 : K_2O was 1 : 0.6 : 0.5 in 1976/77; it changed in favor of nitrogen to 1 : 0.5 : 0.4 in 1986/87, and in 1996/97 it reached 1 : 0.4 : 0.25. While in North America and Western Europe the ratio is 1 : 0.4 : 0.4, it is less well balanced in other regions of the world (Table 39). Fertilizer consumption and production are given for fertilizer years, running from 1 July to 30 June. For example, 1996/97 is the period from 1 July 1996 to 30 June 1997.

Table 39. World nitrogen, phosphate and potassium consumption in millions of tonnes over the period 1976/77 – 1996/97 broken down by region/country [625] [*]

Region/country	Fertilizer season					
	1976/77		1986/87		1996/97	
	10^6 t	%	10^6 t	%	10^6 t	%
Nitrogen (N) fertilizer consumption						
Western Europe	8.8	19.0	11.4	15.9	10.1	12.2
Central/Eastern Europe	3.7	8.0	4.4	6.2	2.4	2.9
Former Soviet Union	7.3	15.7	11.5	16.1	2.8	3.4
North America	10.3	22.2	10.4	14.5	12.9	15.6
Latin America	2.2	4.7	3.8	5.3	4.1	4.9
Asia [**]	12.4	26.7	27.7	38.7	47.4	57.2
Africa	1.5	3.2	1.9	2.7	2.3	2.8
Other	0.2	0.4	0.4	0.6	0.9	1.1
World	46.4	100.0	71.5	100.0	82.9	100.0
Phosphorus (P_2O_5) fertilizer consumption						
Western Europe	5.8	21.6	5.3	15.3	3.7	11.9
Central/Eastern Europe	2.7	10.1	2.0	5.8	0.7	2.3
Former Soviet Union	4.9	18.3	9.1	26.2	0.9	2.9
North America	5.6	20.9	4.3	12.4	4.8	15.4
Latin America	2.0	7.5	2.8	8.1	2.9	9.3
Asia [**]	3.8	14.2	9.2	26.5	15.7	50.5

Africa	0.9	3.4	1.0	2.9	1.0	3.2
Other	1.1	4.1	1.0	2.9	1.4	4.5
World	26.8	100.0	34.7	100.0	31.1	100.0
Potassium (K_2O) fertilizer consumption						
Western Europe	5.5	5.7	5.9	22.6	4.3	20.5
Central/Eastern Europe	2.9	3.0	2.8	10.7	0.7	3.3
Former Soviet Union	5.6	5.8	6.7	25.7	1.2	5.7
North America	5.5	5.7	4.8	18.4	5.2	24.8
Latin America	1.2	1.2	1.9	7.3	2.6	12.4
Asia [**]	1.8	1.9	3.4	13.0	6.2	29.5
Africa	0.3	0.3	0.4	1.5	0.4	1.9
Other	0.2	0.2	0.2	0.8	0.4	1.9
World	23.0	23.9	26.1	100.0	21.0	100.0
Total fertilizer consumption ($N + P_2O_5 + K_2O$)						
World	96.2		132.3		135.0	

[*] *FAO Fertilizer Yearbook* 1980, 1990 and 1997.
[**] Excluding Asian republics of former Soviet Union.

In case only calendar year figures are available, e.g., 1996 is summarized under 1996/97. A regional analysis of fertilizer consumption shows that over 60 % is used in developing countries (in Asia, Latin America, Africa). The Asian markets are particularly important. About 50 % of all fertilizers consumed are applied in Asia. North America takes second place with 17 %, followed by Western Europe. Three decades ago, Western Europe was the world's largest fertilizer consumer; after 10 years, the Eastern Bloc had overtaken Western Europe; and another 10 years later, the developing countries have taken the lead — having increased their consumption tenfold in 30 years. This change corresponds to a mean annual growth rate of 8 %.

The intensity of fertilizer use, however, is still greatest in Western Europe. In terms of nutrient per hectare of agricultural area, Western Europe averaged 121 kg applied in 1996/97 (Table 40). The corresponding figure for the developing countries is 24 kg. In these countries, however, virtually no fertilizer is applied to pastureland. If the figures are referred to agricultural area minus pastureland, the rate in the developing countries was 88 kg of nutrients per hectare, 56 kg of this being N. Fertilizer application is extremely low in the former Soviet Union (8 kg/ha).

Table 40. Nutrient use per hectare of agricultural area, [625], [632]

Region	Nutrient application, kg/ha			
	N	P_2O_5	K_2O	Total
Western Europe	67.6	24.7	28.8	121.2
Central/Eastern Europe	36.1	10.6	9.9	56.6
Former Soviet Union	4.8	1.5	2.1	8.4
North America	26.0	9.8	10.5	46.4
Other	15.2	5.8	2.7	23.6
World average	16.9	6.3	4.3	27.5

[a]Consumption: FAO Fertilizer Yearbook, Agricultural Area: FAOSTAT Database results

Three countries account for over half of world fertilizer consumption: China with 36.6 × 10^6 t, the United States with 21.2 × 10^6 t, and India with 14.3 × 10^6 t. Next on the list are India with 14.3 × 10^6 t, France with 5.1 × 10^6 and Brazil with 4.8 × 10^6.

In years to come, a further worldwide increase is expected. This growth will take place chiefly in Asia, Latin America and Africa. The recovery of fertilizer use in Central/Eastern Europe and the former Soviet Union is anticipated but will take time.

Production. A survey of production in 1996/97 showed that Asia ranks first as world fertilizer producer, mainly for nitrogen, but also for phosphate fertilizers, followed by North America, the leading potash producer (Table 41).

Table 41. World fertilizer production by nutrient and region 1996/97 [625]

Region	Nutrient production, 10^6 t							
	N		P_2O_5		K_2O		Total	
	1986/87	1996/97	1986/87	1996/97	1986/87	1996/97	1986/87	1996/97
Western Europe	11.8	9.5	4.4	2.9	8.3	5.4	24.5	17.8
Central/Eastern Europe	6.1	4.8	2.8	0.9	0.0	0.0	8.9	5.6
Former Soviet Union	15.0	8.7	8.5	2.5	10.2	5.5	33.7	16.7
North America	13.6	19.3	9.4	11.3	8.2	8.9	31.2	39.5
Latin America	3.0	3.2	1.9	1.8	0.1	0.4	5.0	5.4
Asia [*]	26.0	42.6	7.7	11.5	2.0	3.2	35.7	57.4
Africa	1.7	2.7	1.9	2.5	0.0	0.0	3.6	5.2
Oceania	0.2	0.3	0.8	0.7	0.0	0.0	1.0	1.0
World	77.4	91.1	37.4	34.1	28.8	23.4	143.6	148.6
World supply [**]	72.6	85.8	34.9	32.2	26.1	21.8	133.6	139.7

* Excluding Asian republics of former Soviet Union.
** Available world supply was arrived at by deducting from production estimated amounts for technical uses, further processing, transport, storage, and handling losses.

Over the past ten years the production pattern by region shows a considerable decrease in Western and Central/Eastern Europe (– 27 % and – 37 %, respectively) but above all in the former Soviet Union (– 50 %). In North America (+ 27 %) and especially Asia (+ 61 %) production increased markedly. While overall world production remained nearly stagnant (+ 3.5 %), nitrogen production increased by 18 %, but phosphate and potash dropped by 8 % and 16 %, respectively.

The leaders in production of fertilizers are China, the United States, and India. They are followed by Canada, the world's largest potash producer, and Russia (Table 42).

Table 42. Fertilizer production by country, 1996/97 [625]

Country	Fertilizer production, 10^6 t			
	N	P_2O_5	K_2O	Total
China	20.1	5.8	0.2	26.1
United States	12.5 *	10.9	0.8	24.2
India	8.6	2.6	0.0	11.2
Canada	2.7 *	0.4	8.1	11.2
Russia	4.9	1.6	2.6	9.1

* Author's estimate (available supply)

The rise in fertilizer production and consumption in recent decades could not have taken place without an enormous technical advance. Significant cost savings in ammonia production, the intermediate for nitrogen fertilizer production, resulted from the changeover from multi-train to single-train plants and the conversion to more economical feedstocks, mainly natural gas.

Due to improved product quality and well-balanced particle size spectrum, rational bulk transportation and easier application become possible. Higher analysis fertilizer grades brought savings in shipping and storage per unit of nutrient.

The fertilizer processes were optimized to decrease energy consumption. Emission reduction with modern techniques improved the environmental performance of the processes and at the same time increased the nutrient yield from the raw materials. These improvements are of great importance: the 149 × 10^6 t of nutrients produced worldwide in 1996/97 meant 430 × 10^6 t of fertilizers that had to be stored, transported, and applied (authors estimate).

World Trade. Recent decades have seen a sharper rise in world trade than in consumption. Because of the limited number of potash deposits, the ratio of trade to consumption is highest for potash, with 91 %. Some 40 % of phosphate and 29 % of nitrogen are sold internationally (Table 43).

Table 43. World fertilizer trade in millions of tonnes (exports) (10^6 t) [625] *.

Nutrient	Fertilizer season		
	1976/77	1986/87	1996/97
N	8.5	17.4	24.2
P_2O_5	4.4	9.5	12.4
K_2O	13.7	17.6	19.1
Total	26.6	44.5	55.7

* *FAO Fertilizer Yearbooks*

In nitrogen, Russia and the USA have the largest exports, but Canada, Ukraine, and the Netherlands are also important in the world market. The most important importing regions are China and the USA.

American exporters dominate the phosphate market, with sales in all regions, but primarily in the developing countries in Asia, Latin America and Africa. The largest exporters of potash are Canada and Germany, followed by Belarus and Russia.

The largest exporters and importers in 1996/97 are listed in Table 44.

Table 44. Major exporting and importing countries for nitrogen, phosphate, and potash fertilizers, 1996/97 [625]

Country	Exports, 10^6 t	Country	Imports, 10^6 t
Nitrogen (N)			
Russia	3.6	China	4.7
United States	3.0	United States *	2.0
Ukraine	1.2	France	1.3
Netherlands	1.2	India	1.2
Canada *	1.3	Germany	1.2
Phosphate (P_2O_5)			
United States	5.7	China	2.8
Russia	1.2	India	0.7
Morocco	0.8	Australia	0.7

Tunisia	0.7	France	0.6
Netherlands	0.4	Thailand	0.5
Potassium (K_2O)			
Canada	7.9	United States	5.2
Germany	2.6	Brazil	1.8
Belarus	2.0	France	1.1
Russia	1.9	India	0.7
Israel	1.2	Malaysia	0.6

* Author's estimates.

Types of Fertilizers. As the quantities produced and applied have increased, the importance of fertilizer grades has changed. Low-analysis grades have decreased, while highly concentrated grades such as urea, triple superphosphate, and NPK fertilizers have gained (Fig. 52). Sulfur-containing fertilizers have gained importance to cope with sulfur deficiencies.

Figure 52. World fertilizer consumption by type, 1996/97 preliminary [632]

A) Nitrogen fertilizer; B) Phosphate fertilizer; C) Potassium fertilizer

Urea has become the preferred nitrogen fertilizer, especially in the developing countries. In Western Europe, for example, ca. 80 % of nitrogen consumption is in the form of ammonium nitrate, calcium ammonium nitrate, and ammonium nitrate-based multinutrient fertilizers. Phosphate consumption is dominated by multinutrient fertilizers. In the case of potash, straight potassium chloride is the most important grade, but potassium sulfate is of growing importance due to its sulfur contents.

11.3. Future Outlook

The decisive factors for the development of fertilizer demand are population growth and changes in the available income per capita. A growing world population needs more food and fiber. In addition, a wealthier population has a higher daily energy intake per capita and it generally consumes a higher share of animal products (Table 45). Compared to direct human consumption, three to four times more plant products such as cereals are necessary to produce the same amount of energy in the form of animal products like milk, eggs, and meat.

Table 45. Dietary energy supply per capita by economic group 1969 – 1971 and 1990 – 1992 [633]

Country Group	Total kcal/d		% animal	
	1969 – 71	1990 – 92	1969 – 71	1990 – 92
Industrialized	3120	3410	30.4	29.8
Transition	3330	3230	24.4	27.6
Low-income	2060	2430	6.1	9.3
Least developed	2060	2040	6.6	6.0
World	2440	2720	15.6	15.7

11.3.1. *Food Situation*

Since the 1950s world food production per capita has continuously increased, although world population has more than doubled (from 2.5 to almost 6.0 billion). New crop varieties and improved production techniques enabled a substantial growth of average yields. A significant advance was the broader use of effective cereal fungicides in industrialized countries and the introduction of high-yielding varieties of rice, wheat, and maize in several developing countries, known in the 1970s as the “green revolution”. These varieties, bred for tropical and subtropical regions, already offer a higher yield potential under traditional growing conditions. But their full potential can only be exploited when crop husbandry and plant nutrition are adapted to their needs, and agrochemicals provide protection against insects, diseases, and competition from weeds. In many cases they rendered the use of fertilizers and agrochemicals economic for the first time. The average yields of cereals, the most important staple food of humans, increased substantially (Table 46).

Table 46. World production and average yields of cereals 1972 – 1974 and 1996/97

	Production, 10^6 t		Yield, t/ha	
	1972 – 74	1996 – 97	1972 – 74	1996 – 97
Wheat	359	581	1.66	2.53
Maize	300	562	2.75	4.02
Barley	164	155	1.88	2.33
Rye, Oats	81	54	1.68	1.85
Rice	325	558	3.23	3.76
Total	1229	1910	2.19	3.11

According to the FAO, on average each kilogram of fertilizer nitrogen applied to the soil produces about 12 kg of cereal units (1 kg of cereal unit is equal to the nutritional value of 1 kg of barley); therefore, it can be calculated that the food for half of the world's population can only be provided due to the use of (75 – 80) × 10^6 t nitrogen in mineral fertilizers annually. Nevertheless, 840 million individuals still suffer from undernutrition and an estimated 20 million die annually of hunger and undernutrition-related diseases. Civil wars, natural disasters, and poverty, often related to unemployment, are the main reasons for starvation.

Growth in Food Demand. The 1996 update of the UN World Population Prospect shows that the rate of world population growth has declined to 1.48 % per

annum, but this still means an additional 80 million humans to be fed every year. By the year 2020 world population is predicted to reach 7.67 billions. Since economic growth is expected to continue, a larger part of the population will be able to spend more money on food. Thus an increased demand per capita for plant products to be used as animal feed must be taken into account. The demand for cereals is expected to increase by 41 % from 1993 to 2020, that for meat by 63 %, and the consumption of tubers and root is estimated to rise by 40 % [634].

Agricultural Land Use. More than 70 % of the world's surface is covered by water. Of the 13.4 billion hectares of dry land, 4.7 billion hectares are used by agriculture. The major portion is covered by extensively used grasslands, often in low-rainfall areas. Most of the food is supplied by the 1.4 billion hectares of arable land — a mere 10 % of the dry land. Much of the remaining area is not suitable for intensive agricultural production since it is either too dry, too steep, too cold, or infertile. The majority of soils usable for arable crops are covered by biotopes such as rain forests. There is no agreement to which extent they should be used for agriculture and how much must be preserved. FAO [635] calculated that Sub-Saharan Africa and Latin America have the most land reserves with crop-production potential.

At the same time 7 million hectares agricultural land are lost annually due to erosion and salination. The agricultural area available per capita declined from 0.44 ha (1960) to 0.27 ha (1990) and is expected to shrink further to 0.17 ha by 2020. Therefore FAO estimates that although newly cleared land may supply 21 % of the necessary increase in food production, and more frequent use of existing arable land 13 %, the most important contribution (66 %) must come from higher yields on soils already under cultivation. Although neither the quality of soils nor the agroclimatic conditions will allow the same high yield to be achieved everywhere, the wide variation in actual yields offers a vast potential for further improvement (Figure 53).

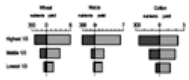

Figure 53. Average nutrient application rates (kg N + P_2O_5 + K_2O per hectare) and average yields (t/ha) of countries with high, medium, and low fertilizer application rates in 1995 [636]

Intensity of Agricultural Production. The intensity of agricultural production is highest in industrialized and transition economies, where the purchasing power for food is high. Attractive markets create an efficient production system, whereas the subsistence farmer cannot invest in yield-raising inputs. Annual nutrient application rates therefore vary widely (Figure 53). Western European countries, Egypt, Saudi Arabia, and Mexico for example apply on average more than 200 kg N + P_2O_5 + K_2O per hectare wheat and harvest 4 – 7 t of grain, whilst countries using less than 50 kg plant nutrients obtain barely more than 2 t per hectare [636]. This of course can also be the result of unfavorable growing conditions, primarily insufficient supply of water, but for most countries it reflects the socio-economic situation.

Wheat is the crop which receives most (22 %) of the fertilizer nutrients, followed by rice and maize (17 % each). Cereals in total count for 65 % of the nutrient consumption, while 9 % are used on oil crops, 5 % each on vegetables and sugar beet/cane). Thus, fertilizers are predominately used for food production. Fiber crops, tobacco, and stimulants receive 4 and 2 %, respectively [636].

In the past countries with economic and population growth have increased their fertilizer consumption (Table 47), whereas those with traditionally high consumption only show moderate growth (e.g., United States, Canada, Spain) or a decrease in consumption. This is the case for most EU countries, where nutrient levels in the soil have mostly been built up to the desirable level, environmental concern has led to strict regulations, and changes in agricultural policy, such as leaving arable land fallow, have been introduced. Changes of the economic system in the former centrally planned economies resulted in a dramatic decline in agricultural production and consequently in fertilizer consumption.

Table 47. Development of fertilizer use in the 20 countries with the highest actual nutrient consumption 1985/86 – 1995/96 in 10^6 t N + P_2O_5 + K_2O [637]

	1985/86	1995/96
China	16 852	35 527
USA	17 831	20 113
India	8504	13 876
France	5695	4915
Brazil	3197	4309
Germany	4823	2818
Indonesia	1972	2512
Pakistan	1511	2443
Canada	2325	2436
UK	2524	2264
Italy	2162	1883
Spain	1734	1817
Australia	1155	1735
Russian Fed.		1700
Turkey	1427	1700
Japan	2034	1642
Poland	3413	1511
Vietnam	386	1448
Thailand	434	1443
Malaysia	611	1247
World	128 613	130 865

11.3.2. *Development of Fertilizer Consumption*

World fertilizer consumption will continue to rise by more than 2 % per annum from 1996/97 to 2002/03 [638], [639]. The rate of growth will be highest in the densely populated countries of South Asia (+ 5.2 %), in Latin America (+ 5.0 %), Africa (+ 4.9 %) and Central Europe (+ 4.4 %) (Figure 54). The estimated increase in Central Europe signals a recovery from economic depression, while the growth rate in Africa simply reflects the low base. For Western Europe, a further decline in fertilizer consumption is forecast since the agricultural area in this densely populated, industrialized region will continue to decline, and more efficient use of organic manures (farmyard manure, sewage sludge, etc.) will be necessary to achieve the goals set for the protection of the environment.

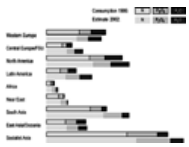

Figure 54. Fertilizer consumption forecast by region 1995 – 2002 [638]

[Top of Page]

References

1. K. Scharrer in, W. Ruhland (ed.): *Handbuch der Pflanzenphysiologie*, vol **IV**, Springer Verlag, Berlin-Göttingen-Heidelberg 1958, pp. 851 – 866.

Specific References

2. J. N. Burnell in R. D. Graham *et al.* (eds.): *Manganese in Soil and Plants*, Academic Press, London 1988, pp. 125 – 137.
3. H. Marschner: *The Mineral Nutrition of Higher Plants*, Academic Press, London 1993.
4. K. H. Falchuk, L. Ulpino, B. Mazus, B. L. Valee, *Biochem. Biophys. Res. Commun.* **74** (1977) 1206 – 1212. Links
5. M. G. Guerrero, J. M. Vega, M. Losada, *Ann. Rev. Plant Physiol.* **32** (1981) 169 – 204.
6. S. J. Brodrick, K. E. Giller, *J. Exp. Bot.* **42** (1991) 1339 – 1343. Links
7. J. v. Liebig: *Die organische Chemie in ihrer Anwendung auf Agrikultur and Physiologie*, Vieweg, Braunschweig 1841.
8. G. Hehl, K. Mengel, *Landwirtsch. Forsch. Sonderh.* **27** (1972) no. 2, 117 – 129.
9. A. K. Goswami, J. S. Willcox, *J. Sci. Food Agric.* **20** (1969) 592 – 595. Links
10. S. Haneklaus, E. Evans, E. Schnug, *Sulphur in Agriculture* **16** (1992) 31 – 34.
11. H. W. Scherer, A. Lange, *Biol Fertil Soils* **23** (1996) 449 – 453. Links
12. E. Saalbach, *Angew. Bot.* **58** (1984) 147 – 156.
13. E. H. Birnbaum, W. M. Dugger, B. C. A. Beasley, *Plant Physiol.* **59** (1977) 1034 – 1038. Links
14. H. E. Goldbach, *J. Trace Microprobe Techn.* **15** (1997) 51 – 91.
15. W. D. Loomis, R. W. Durst, *Biofactors* **3** (1992) 229 – 239. Links
16. G. K. Judel, *Landwirtsch. Forsch. Sonderh.* **34** (1977) no. 2, 103 – 108.
17. D. Werner, R. Roth, *Inorganic Plant Nutrition*. Encycl. Plant Physiol., New Ser. 15 B (1983) 682 – 694.
18. Y. Miyake, E. Takahashi, *Soil Sci. Plant Nutr. (Tokyo)* **29** (1983) 463 – 471.
19. A. D. M. Glass, M. Y. Siddiqi in P. B. Tinker, A. Läuchli (eds): *Advances in Plant Nutrition*, Praeger, New York 1984, vol. 1, pp. 103 – 147.
20. R. G. Wyn Jones, A. Pollard, *Inorganic Plant Nutrition*. Encycl. Plant Physiol. New Ser. 15 B, (1983) 528 – 562.
21. K. Mengel in R. D. Munson (ed): *Potassium in Agriculture*, American Society of Agronomy, Madison, Wis. 1985, pp. 397 – 411.
22. T. R. Peoples, D. W. Koch, *Plant Physiol*, **63** (1979) 878 – 881. Links
23. D. T. Clarkson, J. B. Hanson, *Annu. Rev. Plant Physiol.* **31** (1980) 239 – 298.
24. R. F. Hüttl, H. W. Zöttl, *Forest Ecol. Management* **61** (1993) 325 – 338.
25. C. R. Caldwell, A. Haug, *Physiol. Plant.* **54** (1982) 112 – 118. Links
26. J. B. Hanson in P. B. Tinker, A. Läuchli (eds): *Advances in Plant Nutrition*, Praeger, New York 1984, vol. 1, pp. 149 – 208.
27. C. B. Shear, *Hort. Science* **10** (1975) 361 – 365.
28. H. Marschner, *Neth. J. Agric. Sci.* **22** (1974) 275 – 282.
29. L. C. Campell, R. O. Nable in R. D. Graham, R. J. Hannan, N. C. Uren (eds.): *Manganese in Soils and Plants*, Kluwer Academic Press, Dordrecht 1988.
30. S. Spiller, N. Terry, *Plant Physiol.* **65** (1980) 121 – 125. Links
31. K. Mengel, G. Geurtzen, *J. Plant Nutr.* **9** (1986) 161 – 173. Links
32. J. C. G. Ottow, G. Benckiser, I. Watanabe, S. Santiago, *Trop. Agric. (Trinidad)* **60** (1983) 102 – 106.
33. K. Mengel, E. A. Kirkby: *Principles of Plant Nutrition*, 4th ed., International Potash Institute, Berne 1987.
34. D. Schroeder: *Soils – facts and concepts*, International Potash Institute, Berne 1984.
35. N. C. Brady: *The Nature and Properties of Soils*, Macmillan Publishing Co., New York 1984.
36. F. Scheffer, P. Schachtschabel: *Lehrbuch der Bodenkunde*, 14th ed., Enke Verlag, Stuttgart 1998.
37. R. Dudal in *SEFMIA Proc. Intern. Seminar on Soil Environment and Fertility Management in Intensive Agriculture*, Society of Science of Soil and Manure of Japan, Tokyo 1977, pp. 78 – 88.
38. Soil Taxonomy of the Soil Conservation Service, U.S. Department of Agriculture, Handbook no. 436.
39. E. N. Gapon, *J. Gen. Chem. USSR* **3** (1933) 144 – 152.
40. R. J. Haynes, *Adv. Agron.* **37** (1984) 249 – 315. Links
41. R. L. Parfitt, *Adv. Agron.* **30** (1978) 1 – 50. Links
42. C. D. Foy, R. L. Chaney, M. C. White, *Annu. Rev. Plant Physiol.* **29** (1978) 511 – 566.
43. C. E. Kallsen, T. W. Sammis, E. J. Gregory, *Agron. J.* **76** (1984) 59 – 64.
44. R. L Tate, *Soil Microbiology*, Wiley & Sons, New York 1995.
45. K. Mengel, *Adv. Soil Sci.* **2** (1985) 65 – 131.
46. G. L. Terman, *Adv. Agron.* **31** (1979) 189 – 223. Links
47. R. Vandré, J. Clemens, *Nutrient Cycling in Agrosystems.* **47** (1997) 157 – 165.
48. R. J. Buresh, S. K. DeDatta, *Adv. Argon.* **45** (1991) 1 – 59.
49. S. K. DeDatta, R. J. Buresh, *Adv. Soil Sci.* **10** (1989) 143 – 169.
50. N. K. Savant, S. K. DeDatta, *Adv. Agron.* **35** (1982) 241 – 302. Links
51. J. H. G. Slangen, P. Kerkhoff, *Fert. Res.* **5** (1984) 1 – 76.
52. L. Hendriks, N. Claassen, A. Jungk, *Z. Pflanzenernaehr. Bodenkd.* **144** (1981) 486 – 499.
53. P. H. Nye, *Adv. Agron.* **31** (1979) 225 – 272. Links
54. M. Silberbush, S. A. Barber, *Agron. J.* **75** (1983) 851 – 854.
55. C. A. Neyra, J. Döbereiner, *Adv. Agron.* **29** (1977) 1 – 38. Links
56. V. Römheld, H. Marschner, *Adv*, Plant Nutr. 2 (1986) 155 – 204.
57. R. Perrin, *Soil Use Manag.* **6** (1990) 189 – 195.
58. H. Marschner, B. Dell in A. D. Robson *et al.* (eds.): *Management of Mycorrhizas in Agriculture, Horticulture and Forestry*, 1994 , 89 – 102.
59. A. Finck: *Dünger und Düngung - Grundlagen und Anleitung zur Düngung der Kulturpflanzen*, VCH Verlagsgemeinschaft, Weinheim, Germany 1992.
60. A. L. Page, R. H. Miller, D. R. Keeney: *Methods of Soil Analysis - Chemical and microbiological properties*, 2nd ed., **part 2**, American Society of Agronomy, Inc., Madison, WI 1982.
61. E. Sibbesen, *J. Sci. Food Agric.* **34** (1983) 1368 – 1374. Links
62. G. Hoffmann, *Methodenbuch*, vol. **1, Untersuchung von Böden**, VDLUFA-Verlag, Darmstadt 1991/1997.
63. K. Németh, *Plant and Soil* **83** (1985) 1 – 19.
64. K. Németh, J. Maier, K. Mengel, *Z. Pflanzenernaehr. Bodenk.* **150** (1987) 369 – 374.
65. A. Link, Ph. D. Thesis, University of Hannover, Grauer Verlag Stuttgart 1997.
66. R. J. Poole, *Annu. Rev. Plant Physiol.* **29** (1978) 437 – 460.
67. E. A. Mitscherlich: *Bodenkunde für Landwirte, Forstwirte und Gärtner*, 7th ed., Verlag Paul Parey, Berlin – Hamburg 1954, p. 168.
68. S. P. Ceccotti, D. L. Messick, *Agrofood Industry hi-tech* **5** (1994) 9 – 14.
69. A. Vömel, *Z. Acker Pflanzenbau* **123** (1965/1966) 155 – 188.
70. U. Köpke: “Optimized Rotation and Nutrient Management in Organic Agriculture,” in H. vanKeulen, E. A. Lantinga, H. H. van Laar (eds.): *Proc. Workshop Mixed Farming Systems in Europe*, Dronten, The Netherlands 1998.

General References

71. *Guidelines of the Council of the European Communities for the Harmonization of Legal Provisions in the Member States Concerned with Fertilizers*, 1975.

Specific References

72. J. Norden, *Raiffeisen* 1979, 50 – 55.
73. F. P. Achorn, W. C. Scott, J. A. Willbanks, *Fert. Solution* **14** (1970) 40.
74. V. Buhlmann, *Vereinigte Landwesenkaufleute* 1980, no. 3, 20, 22, 24.
75. J. Norden, *DIZ* (1978) 256 – 260, 512 – 515.
76. *Farm Chem.* **137** (1974) no. 10, 10.
77. T. P. Hignett, *Chem. Technol.* **2** (1972) 627 – 637.
78. T. P. Hignett (ed.): *Fertilizer Manual*, Martinus Nijhoff, Dr. W. Junk Publishers, Dordrecht – Boston – Lancaster 1985, p. 129.
79. D. M. Sloan, R. W. Veales, *J. AOAC* **60** (1977) 876.
80. R. Kümmel, *Chem. Techn. (Leipzig)* **25** (1973) 732.
81. W. R. Grace, NL-A 6 414 487, 1964.
82. *Derakane-Vinylesterharze*, DOW Chemical Europe (1977) Nov., p. 6.
83. *Handbuch der chemischen Beständigkeit der Dynamit Nobel AG*, 1982, p. 60.
84. W. von Beackmann, D. Funk, *3R, Rohre Rohrleitungsbau Rohrleitungstransp.* **17** (1978) 443.
85. L. G. Croysdale, W. E. Samuels, J. A. Wagner: "Urea – Ammonium Nitrate: Are They Safe?," *Chem. Eng. Prog.* **61** (1965) no. 1, 72 – 77. Links J. F. Anderson: "Explosions and Safety in Handling Urea – Ammonium Nitrate Solutions, *Saf. Air Ammonia Plants* **9** (1967) 70 – 71. Van Dolah *et al.*: "Explosion Hazards of Ammonium Nitrate under Fire Exposure," *Rep. Invest. U.S. Bur. Mines* no. 6773 (1966). Th. M. Groothiuzen: *Danger Aspects of Liquid Ammonium Nitrate. Part I. Detonation Properties*, report no. M3038–1 (1979). Th. M. Groothuizen: *Thermische stabiliteit van een ureum-ammonium-nitraatoplossing bij 60 °C*, report no. 484/485/70, Opdracht NSM-NV no. 9139 (1970).
86. J. M. Potts (ed.): "Fluid Fertilizers," Bulletin Y 185, Tennessee Valley Industry, Muscle thoals, Alabama 1984.
87. R. P. Rilo, *Zh. Prikl. Khim. (Leningrad) USSR* **50** (1977) 1676; *J. Appl. Chem. Russ. (Engl. Transl.)* **50** (1977) 1610.
88. A. L. Huhti, P. A. Gartaganis, *Can. J. Chem.* **34** (1956) 785. Links
89. W. P. Moore, "Liquid Fertilizers Have Long Storage Life," *Chem. Eng. News* **49** (1971) no. 44, 37.
90. E. U. Huffman, *New Developments in Fertilizer Technology*, Tennessee Valley Authority, Muscle Shoals, Ala. 1974.
91. A. Alon: "Recent Developments in Phosphoric Acid Manufacture," *ISMA Technical Conference*, Edinburgh, 14th to 16th Sept. 1965, Chap. V.
92. M. S. Casper, *Liquid Fertilizers*, Noyes Data Corp., Park Ridge, N.J., 1973.
93. A. V. Slack: *Fertilizer Developments and Trends*, Noyes Development Corp., Park Ridge, N. J., 1968, p. 344.
94. *Nitrogen* 1971, no. 4, 30.
95. R. D. Young: "Advances in Fertilizer Technology That Will Have Significant Impact," *Agric. Chem.* **26** (1971) 12 – 16, 29 – 31.
96. R. S. Meline, C. H. Davis, R. G. Lee, *Farm Chem.* **133** (1970) no. 11, 26, 28, 30.
97. National Fertilizer Solutions Association: Fluid Fertilizer Manual, Peoria, IL 1985.
98. Ugine Kuhlmann, DE-OS 2 100 413, 1971.
99. G. V. Kostyukhova, N. N. Malakhova, V. M. Lembrikoc, L. M. Soleveva, *Zh. Prikl. Khim. (Leningrad)* **51** (1978) 3; *J. Appl. Chem. USSR (Engl. Transl.)* **51** (1978) 1.
100. J. Silverberg, *Assoc. Am. Fert. Control Off.*, Off. Publ., no. 21 168.
101. M. F. Abd El-Hameed, "Experimentation on the Use of Urea – Ammonium Sulphate (UAS) Liquid Fertilizer in Kuwait." Preprint J. F. Techn. Conf. (1984).
102. *Nitrogen*, 1985, no. 153, 20 – 31.
103. "Phosphoric Acid Sludge: An Inept Phosphatic Material?" *Phosphorus Potassium* 1985, no. 135, 26.
104. "Suspension Fertilizers without Polyphosphates," *Phosphorus Potassium* 1977, no. 89, 31.
105. J. A. Wilbanks: "Suspensions and Slurries," *Agric. Chem.* **22** (1967) 37.
106. "New Developments in Fertilizer Technology, 9th Demonstration," Tennessee Valley Authority, Muscle Shoals, Ala., 1972, pp. 53 – 55.
107. H. E. Mills, *Fert. Solutions* **6** (1962) no. 6, 38.
108. A. V. Slag, *Farm Chem.* **128** (1965) no. 5, 24.
109. M. N. Nabiev, A. F. Glagoleva, *Uzb. Khim, Zh.* 1960 no. 4, 3.
110. BASF, DE-AS 1 592 567, 1965.
111. W. C. Scott, J. A. Wilbanks, L. C. Faulkner: "Fluid Fertilizer," *Proc. Annu. Meet. Fert. Ind. Round Table* **25** (1975) 87.
112. "Urea-Ammonium Nitrate Solutions," *Nitrogen*, 1971, no. 69, 30 – 31.
113. S. H. Wittwer, F. G. Teubner, *Annu. Rev. Plant Physiol.* **10** (1959) 13 – 32.
114. W. Franke, *Planta* **55** (1960) 533 – 541; *Planta* **61** (1964) 1 – 16; Links *Z. Pflanzenkr. (Pflanzenpath.) Pflanzenschutz, Sonderh.* **8** (1977).
115. H.-P. Pissarek, *Dtsch. Landw. Presse* **22** (1972) no. 10.
116. G. Jürgens, H. Ramstetter, *Dtsch. Weinbau* **35** (1980) no. 19, 715 – 717.
117. A. Buchner, H. Sturm, *Gezielter düngen*, DLG-Verlag, Frankfurt 1985, pp. 148 – 155.
118. W. Bergmann, A. Hennig, *Chem. Tech. (Leipzig)* **25** (1973) 391 ff.
119. *Farm Chemicals Handbook '86*, Meister Publishing Co., Willoughby, Ohio, 1986.
120. BASF, DE 2 313 921, 1973 (J. Jung, B. Leutner, H. Sturm). C. A. Whitcomb, EP 0 023 364, 1980, prior. US 61 842, 1979.
121. BASF, EP 0 053 246, 1981 (B. Leutner *et al.*), DE 3 044 903, prior., 1980; EP-A 0 173 089 =DE-OS 3 427 980 1984 (A. Jungbauer, B. Leutner, G. F. Jürgens)
122. G. Jürgens: "Spurennährstoffe in die ähre," *Bauernblatt für Schleswig-Holstein* **33** (1979) no. 129, 24.
123. G. Jürgens, *Chemie und Technik in der Landwirtschaft* 1983, no. 4, 128 – 129.
124. A. Amberger: *Pflanzenernährung*, 4t ed., "UTB 846", Eugen Ulmer Stuttgart, 1996.
125. A. Finck: "Fertilizers and their efficient use," in D. J. Halliday, M. E. Trenkel, W. Wichmann (eds.): *IFA (1992): World Fertilizer Use Manual*, International Fertilizer Industry Association, Paris 1992.
126. Association of American Plant Food Control Officials (AAPFCO) (ed.): *Official Publication No. 48*, Association of American Plant Food Control Officials, Inc., West Lafayette IN, 1995.
127. M. E. Trenkel: *Controlled-Release and Stabilized Fertilizers in Agriculture*, (FAO) IFA, Paris 1997.
128. IFA (ed.): *World Fertilizer Consumption*, International Fertilizer Industry Association, Paris 1996.
129. R. Powell: *Controlled Release Fertilizer*, Noyes Development Corp., Park Ridge, 1968.
130. J. Araten: *New Fertilizer Materials*, Noyes Development Corp., Park Ridge, N. J., 1968.
131. H. Petersen: "Synthesis of Cyclic Ureas by Ureidoalkylation," *Synthesis* 1973, 243 – 244, 262, 264.
132. A. V. Slack: *Fertilizer Development and Trends*, Noyes Development Corp., Park Ridge, N. J., 1968, p. 153 – 156.
133. H. Petersen: "Grundzüge der Aminoplastchemie," *Kunststoff-Jahrbuch*, 10th ed., Wilhelm Parnsegau Verlag, Berlin, 1968, pp. 30 – 107.
134. R. J. Church: "Chemistry and Processing of Urea and Ureaform," in V. Sauchelli (ed.): *Fertilizer Nitrogen Its Chemistry and Technology,* Reinhold Publ., New York 1964.

135. BASF (ed.): *Die Landwirtschaftliche Versuchsstation Limburgerhof 1914 – 1964. 50 Jahre landwirtschaftliche Forschung in der BASF*, BASF Aktiengesellschaft, Limburgerhof 1965.
136. R. Hähndel, "Langsamwirkende Stickstoffdünger — ihre Eigenschaften und Vorteile," *Mitt Landbau* (1986) 4.
137. DuPont, US 2 766 283, 1951 (E. T. Darden) = DE 1 002 765, 1952.
138. DuPont, CA 550 618, 1954 (H. M. Kvalnes, E. T. Darden), US prior., 1953/1954.
139. Allied Chemical, US 2 644 806, 1950 (M. A. Kise), US 2 845 401, 1955 (C. E. Waters).
140. J. M. O'Donnell, US 2 916 371, 1956.
141. Hercules Powder, US 3 227 543, 1965 (J. M. O'Donnell).
142. Borden Co., US 2 810 710, 1955 (D. R. Long).
143. Fisons, GB 789 075, 1954 (A. L. Whynes).
144. O. M. Scott & Sons, US 3 076 700 ,1957 (V. A. Renner); DE-OS 1 592 751, 1963 (V. A. Renner, R. H. Czurak).
145. Unpublished results of the BASF Ammonia Laboratory, 1962.
146. T. Ushioda, *Jpn. Chem. Q* **5** (1969) no. 4, 27.
147. E. Rother: "Depotdünger," 1975, and "Arbeiten zur Entwicklung eines Herstellverfahrens für Mehrnährstoffdünger mit Langzeitstickstoff," 1985, both internal BASF reports.
148. H. Schneider, L. Veegens: "Practical Experience with Ureaform Slow Release Nitrogen Fertilizer During the Past 20 Years and Outlook for the Future', *Proc. Fert. Soc.* **180** (1979) 4 – 39.
149. Aglukon company broschure: *"Nutralene, Nitroform, AZOLON — Biologically Controlled Nitrogen Release,"* Aglukon Spezialdünger GmbH, Düsseldorf 1993.
150. J. Jung, J. Dressel, *Landwirtsch. Forsch. Sonderh.* **26/I** (1971) 131.
151. Mitsubishi, DE 1 146 080, 1961 (M. Hamamoto, J. Sakaki). Mitsubishi, US 3 322 528, 1961 (M. Hamamoto).
152. Mitsubishi, DE 1 543 201, 1965 (S. Kamo, N. Yanai, N. Osako), JA prior (1964); GB 1 099 643, 1965, JA prior (1964). Hoechst, DE 1 303 018, 1963 (H. Schäfer, P. Krause); DE 1 244 207, 1962 (U. Schwenk, F. Kalk, H. Schäfer). BASF, DE-OS 1 618 128, 1967 (G. Rössler); DE 2 355 212, 1973 (O. Grabowsky *et al.*).
153. BASF, DE 1 081 482, 1959 (J. Jung *et al.*)
154. BASF, DE 1 176 39, 1963 (H. Kindler *et al.*); DE 1 244 ;207, 1962 (H. Brandeis, H. Petersen, A. Fikentscher).
155. Chisso Corp., JP 492 348, 1967; DE-OS 1 670 385, 1966 (K. Fukatsu *et al.*); US 3 488 351, 1966. (K. Fukatsu *et al.*).
156. C. J. Pratt, R. Noyes: *Ureaform in Nitrogen Fertilizer Chemical Process*, Noyes Development Corp., Pearl River, N. Y., 1965, pp. 232 – 234.
157. BASF: *Produktinformation für Erwerbsgartenbau, Landschaftsbau und Landwirtschaft,* COMPO GmbH, Münster 1995.
158. J. Jung, *Z. Pflanzenernaehr. Dueng. Bodenkd.* **94** (1961) 39; J. Jung, C. Pfaff, *Die Landwirtschaftliche Versuchsstation Limburgerhof 1914 bis 1964*, BASF, Ludwigshafen 1964.
159. A. J. Patel, G. C. Sharma, *J. Am. Soc. Hortic, Sci.* **102** (1977) no. 3, 364 – 367.
160. BASF, DE-OS 1 467 377, 1982 (J. Jung *et al.*).
161. *Chem. Ing. Tech.* **46** (1974) A 100; *Chem. Age (London)* **108** (1975) no. 2852, 17 – 18.
162. Hoechst, DE-OS 2 308 941, 1973.
163. W. Riemenschneider: "A New Simple Process for Manufacturing Oxamide," in *Prod. Tech. Plant Nutr. Effic. Br. Sulphur Corp. Int. Conf. Fert., 2nd, 1978*, (1978 – 1979) 1 – 17.
164. *Eur. Chem. News* **24** (1973) no. 604; **24** (1973) no. 609. T. Okada, *J. Ammonium Industry* **1** (1975) no. 2, 28; *Nitrogen* 1979, no. 117, 43.
165. T. Fujita, personal communication, 1996.
166. G. L. Bridger *et al.*, *J. Agric. Food Chem.* **10** (1962) 181. Links
167. M. L. Salutzky, R. P. Steiger, *J. Agric. Food Chem.* **12** (1962) 486 – 491.
168. Mitsui Toatsu Chem., JP-Kokai 067 089/82, 1980. Nugata Sulphuric Acid Co, JP 7 035 210, (Prior) 1966; *Chem. Abstr.* **75** (1971) 34 643.
169. Wintershall AG, DE-OS 1 792 725, 1963 (W. Jahn-Held); DE 1 252 210, 1962 (W. Jahn-Held); DE 1 265 726, 1964 (W. Jahn-Held), DE 1 592 810, 1963 (W. Jahn-Held), DE 1 592 811, 1963 (W. Jahn-Held). F. Müller, DE 1 924 284, 1969. Kali & Salz, DE 3 204 238, 1982; DE 3 216 973, 1982. Knapsack, DE 1 260 450, 1966 (W. Kern, J. Cremer, H. Harnisch).
170. Monsanto, GB 1 226 256, 1968.
171. BFG Glass Group, BE 878 884, 1978.
172. J. Jung. *Z. Pflanzenernähr. Düng Bodenkd.* **91** (1960) 122 – 130. Archer Daniels Midland, US 3 223 518, 1961 (L. J. Hansen), DE 1 254 162, 1960 (L. J. Hansen).
173. M. J. Brown, R. E. Luebs, P. F. Pratt, *Agron. J.* **58** (1966) no. 2, 175 – 178. J. J. Oertli, O. R. Lunt, *Soil Sci. Soc. Am. Proc.* **26** (1962) 579 – 587.
174. Aglukon company brochure: *"Plantacote — The System of Crop-Specific Plant Nutrition,"* Aglukon Spezialdünger GmbH, Düsseldorf, 1992.
175. T. Fujita, "Invention of fertilizer coating technology using polyolefin resin and manufacturing of polyolefin coated urea," *Jap. J. Soil Sci. Plant Nutr.* **67** (1996) 3.
176. T. Fujita *et al.*: *MEISTER with controlled availability — Properties and efficient utilization,* Chisso-Asahi Fertilizer Co., Ltd., Tokyo.
177. T. Fujita: *"Technical Development, Properties and Availability of Polyolefin Coated Fertilizers,"* Proceedings Dahlia Greidinger Memorial International Workshop on Controlled/Slow Release Fertilizers, Technion – Israel Institute of Technology, Haifa, 7 – 12 March 1993.
178. Chisso Corp., CA 1 295 849 ,1992 (T. Fujita, S. Yoshida, K. Yamahira).
179. Chisso Corp., CA 1 295 848 ,1992 (T. Fujita, Y. Yamashita, S. Yoshida, K. Yamahira).
180. T. Fujita, S. Maeda, M. Shibata, C. Takahashi: *"Research & development of coated fertilizer,"* Proceedings Fertilizer, Present and Future, Japan. Society of Soil Science and Plant Nutrition, Tokyo 1989, pp. 78 – 100.
181. T. Fujita, S. Maeda, M. Shibata, C. Takahashi: *"Research & development of coated fertilizers,"* Japanese Society of Soil Science & Plant Nutrition, Proceedings of the Symposium on Fertilizer, Present and Future, Sept. 25 – 26, Tokyo 1989.
182. H. M. Goertz: *"Technology Developments in Coated Fertilizers,"* Proceedings Dahlia Greidinger Memorial International Workshop on Controlled/Slow Release Fertilizers, Technion – Israel Institute of Technology, Haifa, 7 – 12 March 1993.
183. S. P. Landels: *Controlled-Release Fertilizers: Supply and Demand Trends in U.S. Nonfarm Markets*, SRI International, Menlo Park, CA 1994.
184. W. P. Moore: *"Reacted Layer Technology for Controlled Release Fertilizers,"* Proceedings Dahlia Greidinger Memorial International Workshop on Controlled/Slow Release Fertilizers, Technion – Israel Institute of Technology, Haifa, 7 – 12 March 1993.
185. S. Shoji, A. T. Gandeza: *Controlled Release Fertilizers with Polyolefin Resin Coating*, Kanno Printing Co. Ltd., Sendai 1992.
186. The O.M. Scott & Sons Co., US 5 089 041 ,1992 (H. E. Thompson, R. A. Kelch).
187. Archer Daniels Midland, DE-AS 1 242 573, 1964 (Le Roy B. Sahtin), US prior., (1963). Archer Daniels Midland, US 3 264 088, 1965 (L. J. Hansen). Sierra Chemical Europe, EP-A 184 869, 1985; NL prior., 1984 (J. M. H. Lembie).
188. Chisso Asahi Fertilizer Co., JP 634 150, 1972, 650 860, 1972, DE-AS 2 461 668, 1974 (T. Fujita *et al.*), JP prior., 1973; US 4 019 890, 1974 (T. Fujita *et al.*), JP prior., 1973.
189. Chisso Asahi Fertilizer Co., DE 2 834 513, 1978 (T. Fujita *et al.*), JP prior., 1978.
190. RLC Technologies, L.L.C., US 5 599 374, 1997 (J. H. Detrick); US 5 547 486, 1996 (J. H. Detrick, F. T. Carney, Jr. US 5 374 292, 1993 (J. H. Detrick, F. T. Carney, Jr. US 4 969 947, 1990 (W. P. Moore); US 4 804 403, 1989 (W. P. Moore); US 4 711 659, 1987 (W. P. Moore).
191. Pursell Inc. company brochure: *"POLYON Polymer Coatings and the RLC Process,"* Pursell Industries, Inc., Sylacauga, Alabama 1995.
192. H. J. Detrick, PURSELL Technologies Inc., personal communication, 1997.
193. D. W. Rindt, G. M Bloin, J. G. Getsinger, *J. Agric. Food Chem.* **16** (1968) 773 – 778. Links

194. *Ag Chem Commer. Fert.* **28** (1973) no. 2, 14 – 15; *Chem. Eng. (N. Y.)* **81** (1974) Dec. 23, 32 – 33. *Sulphur Inst. J.* **10** (1974) no. 3/4, 7.
195. S. F. Allen *et al.*, *Crops and Soils* **21** (1968) no. 3; *Agron. J.* **63** (1971) 529.
196. D. W. Rindt, G. M. Bloin, O. E. Moore, *J. Agric. Food. Chem.* **19** (1971) no. 5, 801 – 808.
197. Tennessee Valley Authority, US 3 295 950, 1965 (G. M. Bloin, D. W. Rindt); US 3 342 577, 1966 (G. M. Bloin, D. W. Rindt); US 3 903 333, 1972; (A. R. Shirley, R. S. Meline). O. M. Scott & Sons, US 4 042 366, 1976 (K. E. Fersch, W. E. Stearus).
198. RLC Technologies, L.L.C., US 5 599 374 ,1997 (J. H. Detrick).
199. Bayer, DE-AS 2 240 047, 1972; DE-OS 3 020 422, 1980; EP 46 896, 1981; EP-A 95 624, 1983 (H. Heller, D. Schäpel).
200. Bayer, EP 90 992, 1983, DE 3 212 537, 1982 (H. Brunn *et al.*).
201. B. Crawford: *Florida Commercial Fertilizer Law. Rules and Regulations*, State of Florida, Department of Agriculture and Consumer Services, Tallahassee FL, 1995.
202. Bundesminister für Ernährung, Landwirtschaft und Forsten: *Verordnung zur Änderung düngemittelrechtlicher Vorschriften*, 1995, BGBl. I.
203. G. Kluge, G. Embert: *Das Düngemittelrecht mit fachlichen Erläuterungen*. Bonn 1996.
204. T. Fujita, personal communication, 1997.
205. A. Amberger: *Pflanzenernährung*, 4th ed., **Taschenbücher 846**, Verlag Eugen Ulmer, Stuttgart 1996, pp. 156, 157.
206. K. Mengel: *Ernährung und Stoffwechsel der Pflanze*, VEB Gustav Fischer Verlag, Jena 1968, pp. 287, 289.
207. Bundesrat, Drucksache 239/96 *Stellungnahme der Bundesregierung zu der Entschließung des Bundesrates zur Verordnung zur Änderung düngemittelrechtlicher Vorschriften,* Bundesanzeiger Verlagsgesellschaft mbH, Bonn 1996.
208. A. Amberger: "Potential of the Nitrification Inhibitor Dicandiamide to Control Nitrogen Management and Environmental Pollution," *Proceedings, Dahlia Greidinger Memorial International Workshop on Controlled/Slow Release Fertilizers, Technion,* Israel Institute of Technology Haifa, March 7 – 12 1993.
209. B. Scheffer, "Application of nitrogen fertilizers with nitrification inhibitors in water drainage areas," *GWF, Gas Wasserfach: Wasser/Abwasser* **135** (1994) 15 – 19.
210. B. Scheffer: "Nitratgehalte im Dränwasser bei Einsatz stabilisierter Stickstoffdünger," Stabilisierte Stickstoffdünger — ein Beitrag zur Verminderung des Nitratproblems, Proccedings: 15./16. Oktober 1991, Würzburg, Publisher; BASF Aktiengesellschaft, Limburgerhof/SKW Trostberg AG, 1991, pp. 83 – 94.
211. P. Schweiger: "Wege zur Minimierung des Nitratproblems," Proceedings: Stabilisierte Stickstoffdünger — ein Beitrag zur Verminderung des Nitratproblems. Fachtagung: 15./16. Oktober 1991, Würzburg. Publisher; BASF Aktiengesellschaft, Limburgerhof/SKW Trostberg AG, 1991, pp. 59 – 69.
212. W. Zerulla: "N_{min} — Gehalte im Boden nach der Düngung bzw. nach der Ernte," Proceedings: Stabilisierte Stickstoffdünger — ein Beitrag zur Verminderung des Nitratproblems. Fachtagung: 15./16. Oktober 1991, Würzburg. Publisher; BASF Aktiengesellschaft, Limburgerhof/SKW Trostberg AG, 1991, pp. 111 – 119.
213. H. Sturm, A. Buchner, W. Zerulla: *Gezielter düngen. Integriert, wirtschaftlich, umweltgerecht*, VerlagsUnion Agrar, DLG-Verlags-GmbH, Frankfurt a.M. 1994, pp. 90 – 95.
214. K. F. Bronson, A. R. Mosier, "Suppression of methane oxidation in aerobic soil by nitrogen fertilizers, nitrification inhibitors, and urease inhibitors," *Biol. Fertil. Soils* **17** (1994) 4, 263 – 280. Links
215. K. F. Bronson, A. R. Mosier: "Effect of Nitrogen Fertilizer and Nitrification Inhibitors on Methane and Nitrous Oxide in Irrigated Corn," *Biochemistry of Global Change. Radiatively Active Trace Gases*, Chapman & Hall, New York – London 1993.
216. K. F. Bronson, A. R. Mosier, S. R. Bishnoi, "Nitrous Oxide Emissions in Irrigated Corn as Affected by Nitrification Inhibitors," *Soil Sci. Soc. Am. J.* **56** (1992) 1, 161 – 165. Links
217. H.-J. Klasse: "Versuchsergebnisse zur Wirkung stabilisierter Stickstoffdünger auf die Nitratverlagerung bzw. -auswaschung. N_{min} — Gehalte im Boden nach der Düngung bzw. nach der Ernte," Proceedings: Stabilisierte Stickstoffdünger — ein Beitrag zur Verminderung des Nitratproblems. Fachtagung: 15./16. Oktober 1991, Würzburg. Publisher; BASF Aktiengesellschaft, Limburgerhof/SKW Trostberg AG, 1991, pp. 103 – 110.
218. M. Koshino: *The Environmental Protection Framework Concerning Fertilizer Use in Japan*, National Institute of Agro-Environmental Sciences, Department of Farm Chemicals, Tsukuba 1993.
219. C. A. Grant, S. Jia, K. R. Brown, L. D. Baily: "Volatile losses of NH_3 surface-applied urea and urea ammonium nitrate with and without the urease inhibitors NBPT or ammonium thiosulphate," *Canadian Journal of Soil Science* **76** (1996) 3. Links
220. C. J. Watson *et al.*, "Soil properties and the ability of the urease inhibitor N-(n-butyl) thiophosphoric triamide (nBTPT) to reduce ammonia volatilization from surface-applied urea," *Soil Biol. Biochem.* **26** (1994) 9, 1165 – 1171. Links
221. C. A. Grant, R. Ferguson, R. Lamond, A. Schlegel, W. Thomas: "Use of Urease Inhibitors in the Great Plains," *Proceedings Great Plains Soil Fertility Conference,* Denver, Co 1996.
222. M. E. Trenkel: *Controlled-Release and Stabilized Fertilizers in Agriculture*, (FAO) IFA, Paris 1997, p. 12.
223. J. A. Delgado, A. R. Mosier: "Mitigation alternatives to decrease nitrous oxides emissions and urea – nitrogen loss and their effect on methane flux," *J. Environ. Qual.* **25** (1996) 5. Links
224. A. Amberger: "Mobilization of rock phosphates in soil," *2nd African Soil Science Society Conference,* Cairo 1991.
225. F. Bayrakli, S. Gezgin: "Controlling ammonia volatilization from urea surface applied to sugar beet on a calcareous soil," *Commun. Soil. Sci. Plant Anal.* **27** (1996) 9, 10.
226. Z. P. Wang, O. van Cleemput, L. Baert: "Movement of urea and its hydrolysis products as influenced by moisture content and urease inhibitors," *Biol. Fertil. Soils* **22**, (1994) 1/2.
227. Z. P. Wang, O. van Cleemput, P. Demeyer, L. Baert: "Effect of urease on urea hydrolysis and ammonia volatilization," *Biol. Fertil. Soils* **11** (1991) 1, 43 – 47. Links
228. Z. P. Wang, L. T. Li, O. van Cleemput, L. Baert: "Effect of urease inhibitors on denitrification in soil," *Soil Use Manage.* **7** (1991) 4, 230 – 233.
229. B. Zacherl, A. Amberger: "Effect of the nitrification inhibitors dicyandiamide, nitrapyrin and thiourea on Nitrosomonas europea," *Fert. res.* **22** (1990) 37 – 44.
230. J. K. R. Gasser: "Nitrification Inhibitors — Their Occurence, Production and Effects of their Use on Crop Yields and Composition," *Soils Fert.* **33** (1970) 547 – 554.
231. K. Sommer, *Nitrifizide,* Landw. Forschung Sond. 27/II, 1972, pp. 64 – 82.
232. US 3 135 594, 1964 (Cleve A. I. Goring), Serial No. 653, 065 1957 expired in 1981; continuation-in-part of copending application.
233. US 3 424 754, 1969 (William H. Taplin III.), expired in 1986. US 3 420 833, 1969 (William H. Taplin), expired in 1983.
234. DowElanco(ed.): *Quick Guide to Nitrogen Management,* DowElanco, Champaign, IL.
235. DowElanco(ed.): *References on N-Serve® Nitrogen Stabilizer — Bibliography,* DowElanco, Champaign, IL 1989.
236. J. Huffman, DowElanco, personal communication, 1997.
237. W. Zerulla, BASF Aktiengesellschaft, personal communication, 1996.
238. US 3 135 594, 1964 (Cl. A. I. Goring).
239. US 5 106 984, 1992 (M. E. Halpern, J. A. Orvik, J. Dietsche); EP 0 306 547 B1, 1992 (M. E. Halpern, J. Dietsche, J. A. Orvik, B. J. Barron).
240. FR 1 232 366, 1960 (L. Soubies, R. Gadet).
241. S. Hallinger: "*Untersuchungen zur biologischen Metabolisierung von Dicyandiamid,* Technische Universität München," Isnstitut für Bodenkunde, Pflanzenernährung und Phytopathologie, Lehrstuhl für Pflanzenernährung, Dissertation, 1992.
242. M. Hauser, K. Haselwandter: "Degradation of Dicyandiamide by Soil Bacteria," *Soil Biol. Biochem.*, **22** (1990) 113 – 114. Links
243. K. Vilsmeier: "Fate of ammonium-N in pot studies as affected by DCD addition," *Fert. res.* **29** (1991) 3, 187 – 189.
244. K. Vilsmeier: "Turnover of ^{15}N-Ammonium Sulfate with Dicyandiamide under Aerobic and Anaerobic Soil Conditions," *Fert. Res.* **29** (1991) 191 – 196.
245. R. Roll: *"Zur Toxikologie von Dicyandiamid,"* Proceedings: Stabilisierte Stickstoffdünger — ein Beitrag zur Verminderung des Nitratproblems.

Fachtagung: 15./16. Oktober 1991, Würzburg. BASF Aktiengesellschaft, Limburgerhof/SKW Trostberg AG, 1991, pp. 77 – 82.

246. US 4 764 200, 1988 (O. Meiss, R. E. Nitzschmann); DE 4 128 828 A1, 1993 (K. Engelhardt, H. Dittmar, S. Luther).
247. US 4 994 100, 1991 (A. R. Sutton, Ch. W. Weston, R. L. Balser).
248. US 5 352 265, 1994 (Ch. W. Weston, L. A. Peacock, W. L. Thornsberry, Jr., A. L. Sutton).
249. DD 222 471, 1985 (H. J. Herbrich *et al.*).
250. DD 247 894, 1987 (K. Gunther *et al.*).
251. US 4 975 107, 1990 (K. E. Arndt *et al.*), US 4 969 946, 1990 (K. E. Arndt *et al.*), US 4 545 803, 1985 (K. E. Arndt *et al.*).
252. DE 4 446 194, 1996 (J. Dressel *et al.*), DE 4 137 011, 1993 (R. Kästner *et al.*), DE 3 820 739, 1990 (U. Baus *et al.*), DE 3 820 738, 1990 (U. Baus *et al.*), DE 3 409 317, 1985 (H. Bohm *et al.*).
253. DE 4 405 393, 1995 (M. Grabarse *et al.*), DE 4 405 392, 1995 (M. Grabarse *et al.*), DE 4 237 688, 1994 (I. Berg *et al.*), DE 4 237 687, 1994 (M. Grabarse *et al.*), DE 4 219 661, 1994 (M. Grabarse *et al.*), DE 4 218 580, 1993 (H. J. Hartbrich *et al.*), DE 4 211 808, 1993 (H. Böhland *et al.*), DE 4 018 395, 1991 (M. Grabarse *et al.*), DD 273 829, 1990 (H. J. Hartbrich *et al.*), DD 279 470, 1990 (H. J. Halbenz *et al.*),
254. DD 260 486, 1989 (H. Carlson *et al.*),
255. RU 2 046 116, 1996 (L. A. Gorelik *et al.*).
256. SU 1 680 683, 1992 (O. V. Kuznetsova *et al.*).
257. US 4 523 940, 1985 (K. E. Arndt *et al.*), US 4 522 642, 1985 (K. E. Arndt *et al.*).
258. DE 4 128 828 A1, 1993 (H. Dittmar *et al.*).
259. WO 98/05607, 1998 (Th Barth *et al.*)
260. S. Kincheloe: “The manufacture, agronomics and marketing of Agrotain,” IFA Agro-Economics Committee Conference: 'Plant Nutrition in 2000', Tours, June 1997.
261. S. Kincheloe, A. R. Sutton: “The Manufacture, Agronomics and Marketing of agrotain,” Book of Abstracts, 212th ACS National Meeting, American Chemical Society, Washington, D.C. 1996.
262. F. J. Dentener, P. C. Crutzen: “A Three-Dimensional Model of the Global Ammonia Cycle,” *Atmos. Chem.* **19** (1994) 331 – 369.
263. W. H. Schlesinger, A. E. Hartley: “A Global Budget for NH_3,” *Biogeochem.* **15** (1992) 191 – 211.
264. A. Finck: *Dünger und Düngung*, Verlag Chemie, Weinheim – New York 1979, pp. 52, 59.
265. US 4 530 714, 1985 (J. F. Kolc, M. D. Swerdloff, M. M. Rogic, L. L. Hendrickson, M. van der Puy).
266. Ch. F. Wilkinson: *N-(n-butyl) thiophosphoric triamide (NBPT, Agrotain®): A Summary of its Chemistry, Toxicology, Ecotoxicology and Environmental Fate,* IMC-Agrico Company, Mundelein, IL 1996, p. 6.
267. IMC Agrico (ed.): *Material Safety Data Sheet , AGROTAIN® Urease Inhibitor,* IMC-Agrico Company, Mundelein, IL 1995.
268. US 5 352 265, 1994 (C. W. Weston *et al.*), US 5 364 438, 1994 (C. W. Weston *et al.*).
269. US 5 698 003, 1997 (B. A. Omilinsky *et al.*).
270. Ministére de L'Environment de la République Français: “Bonne pratiques d'épandage et de stockage des fertlisants,” Journal Officiel de La Republique Française, Jan. 5, 1994.
271. Bundesminister für Ernährung, Landwirtschaft und Forsten: *Verordnung zur Änderung düngemitelrechtlicher Vorschriften,* Bundesgesetzblatt Jahrgang 1995, Teil I., Bonn 1995.
272. G. Kluge, G. Embert: *Das Düngemittelrecht mit fachlichen Erläuterungen 1992*, Landwirtschaftsverlag GmbH, Münster – Hiltrup 1992.
273. KrW/AbfG: Kreislaufwirtschafts- und Abfallgesetz, Gesetz zur Vermeidung, Verwertung und Beseitigung von Abfällen, BGBl. I (1994) p. 2705.
274. G. Kluge, G. Embert: *Das Düngemittelrecht mit fachlichen Erläterungen, (inkl. 1. Supplement 1997)*, Landwirtschaftsverlag GmbH, Münster – Hiltrup 1996.
275. R. Gutser, *VDLUFA Schriftenreihe* **44** (1996), 29 – 43.
276. AbfKlärV: Klärschlammverordnung, BGBl. I (1992)
277. K. Severin, *VDLUFA Schriftenreihe* **44** (1996) 85 – 101.
278. R. Gutser, K. Teicher, P. Fischer, *Acta Horticulturae* **150** (1984) 175 – 184.
279. A. Fink: *Dünger und Düngung*, VCH Verlagsgesellschaft, Weinheim, Germany 1992, pp. 164 – 174.
280. E. Knickmann: “Handelshumusdüngemittel”, in K. Scharrer, Linser (eds.): *Handbuch der Pflanzenernährung und Düngung*, Springer-Verlag, Wien 1968, pp. 1400 – 1431.
281. “Abfallstoffe als Dünger”, *VDLUFA-Schriftenreihe* **23** (1988).
282. “Sekundärrohstoffe im Stoffkreislauf der Landwirtschaft”, *VDLUFA-Schriftenreihe* **44** (1996).
283. R. Gutser, *Z. Branntweinwirtschaft* **138** (1998) 2 – 6.
284. U. Bayer, Diss. Techn. Univ. München, 1987.
285. T. v. Tucher, Diss. Techn. Univ. München, 1990.
286. Bundesministerium für Umwelt, Naturschutz- und Reaktorsicherheit: Entwicklung der Schwermetallgehalte von Klärschlämmen in der BRD von 1974 – 1994, Informationsblatt 1996.
287. A. Gronauer *et al.*: Bioabfallkompostierung — Verfahren und Verwertung, Bayerisches Landesamt für Umweltschutz, Schriftenreihe Heft 139 (1997).
288. P. Fischer, M. Jauch, *VDLUFA Schriftenreihe* **33** (1991) 751 – 756.
289. M. Bertoldi, P. Sequi, B. Lemmes, T. Papi: “The Science of Composting,” *European Comm. Int. Symp.*, Udine 1995, Blackie Academic & Professional, an imprint of Chapman – Hall, 1996.
290. J. Köhnlein, H. Vetter: “Die Stalldüngerrotte bei steigender Stroheinstreu,” *Z. Pflanzenernähr. Düng. Bodenkd.* **63** (1953) no. 108, 119 – 141.
291. G. J. Kohlenbrander, L. C. N. de la Lande Cremer: *Stahlmest en Gier*, H. Veenman & Zonen, Wageningen 1967.
292. H. Koriath und Kollektiv: *Güllewirtschaft, Gülledüngung*, VEB Deutscher Landwirtschaftsverlag, Berlin 1975.
293. K. Rauhe: “Wirtschaftseigene Düngemittel” in *Handbuch der Pflanzenernährung und Düngung*, vol. **2**, Springerverlag, Wien-New York, 1968 pp. 907 – 943.
294. D. Strauch, W. Baader, C. Tietjen: *Abfälle aus der Tierhaltung*, Verlag Eugen Ulmer, Stuttgart 1977.
295. H. Vetter, G. Steffens: *Wirtschaftseigene Düngung*, DLG-Verlag, Frankfurt 1986.
296. *Faustzahlen für die Landwirtschaft*, 10th ed., Landwirschaftsverlag, Münster-Hiltrup 1983.
297. A. Amberger: “Gülle – ein schlechtgenutzter Dünger,” *DLG-Mitteilungen* **2**, 1982 pp. 78 – 80.
298. J. C. Brogan (ed): *Nitrogen Losses and Surface Run-off from Landspreading of Manures*, Martinus Nijhoff/Dr. W. Junk Publishers, Den Hague-Boston-London 1981.
299. K. Früchtenicht, H. Vetter: “Phosphatzufuhr und Phosphatwirkung aus organischer Düngung,” in *Phosphatdüngung aus heutiger Sicht*, Arbeitskreis Phosphat der deutschen Düngerindustrie, Hamburg 1983.
300. J. K. R. Gasser: *Effluents from Livestock*, Applied Science Publishers, London 1980.
301. T. W. G. Hucker, G. Catroux: *Phosphorus in Sewage Sludge and Animal Waste Slurriers*, Reichel, Dordrecht 1981.
302. G. Schechtner: “Nährstoffwirkung und Sonderwirkungen der Gülle auf dem Grünland,” in *Bericht über die 7. Arbeitstagung “Fragen der Güllerei”*, Bundesversuchsanstalt für alpenländische Landwirtschaft, 1981, pp. 135 – 196.
303. J. H. Voorburg: *Utilization of Manure by Land Spreading*, Janssen Services, London 1977.
304. L. Diekmann: “Schadgase aus Flüssigmist” in *Rationalisierungskuratorium für Landwirtschaft* (*RKL*), 1983, pp. 839 – 885.
305. P. Förster: “Einfluß hoher Güllegaben und üblicher Mineraldünger auf die Stoffbelastung (NO_3, NH_4, P und SO_4) im Boden und Grundwasser in

Sandböden Nordwestdeutschlands," *Z. Acker Pflanzenbau*, **137** (1973) 270 – 286.
306. W. R. Kelly: *Animal and Human Health Hazards Associated with the Utilization of Animal Effluents*, Commission of the European Community, Luxembourg 1978.
307. H. Kuntze: "Düngung und Gewässergüte," *Landwirtsch. Forsch. Sonderh.* **35** (1979) 14 – 24.

General References

308. *Fertilizer Manual, Development and Transfer of Technology*, Series no. 113, UNIDO, United Nations, New York 1980.
309. A. Schmidt: *Chemie und Technologie der Düngemittelherstellung*, Hüthig-Verlag, Heidelberg 1972.
310. A. J. More (ed.): *Granular Fertilizers and Their Production*, The British Sulphur Corp., London 1977.
311. *Fertilizers: Ammonium Phosphates*, **Process Economics Program Report no. 127**, SRI International, Menlo Park, Ca., 1979.
312. F. T. Nielsson: *Manual of Fertilizer Processing*, vol. **5 of the Fertilizer Science and Technology Series**, 1987.

Specific References

313. *Kirk-Othmer*, **10** (1980) 31 – 125.
314. T. P. Hignett, *Farm Chem.* 1963, no. 1, 34 – 35.
315. *Winnacker-Küchler*, 4th ed., **2** pp. 334 – 378.
316. J. Jung, *Proc. Int. Congr. Phosphorus Compd., 3rd*, 1983, 463 – 495.
317. "Granular Fertilizers," supplement to *Phosphorus Potassium* **1985**.
318. V. Sauchelli, *Chemistry and Technology of Fertilizers, ACS Monograph Series*, Reinhold Publishing Corporation, New York; Chapman & Hall Ltd., London 1960, p. 280.
319. A. T. Brook, "Developments in Granulation Techniques," paper read before the Fertiliser Society of London, 1957.
320. O. Bognati, L. Buriani, I. Innamorati, *Proc. ISMA Tech./Agric. Conf.*, 1967.
321. T. E. Powell, *Proc. Technol. Intern.* **18** (1973) 271 –278.
322. *Nitrogen* 1981, no 131, 39 – 41.
323. *Nitrogen* 1975, no. 95, 31 – 36.
324. *World Guide to Fertilizer Process and Constructors*, 6th ed., The British Sulphur Corp., London 1979, pp. 138 – 139.
325. *Nitrogen* **175** (1988) 32 – 40.
326. O. E. Skauli: "Pan Granulation of Ammonium Nitrate and Urea," (Symposium on New Fertilizer Technology) ACS National Meeting, 168th, 1974.
327. J. P. Bruynseels: "Fluid Bed Granulation of Ammonium Nitrate and Calcium Ammonium Nitrate," *Proc. of the Fertiliser Society* 1985, no. 235.
328. J. P. Bruynseels, *Hydrocarbon Process.* **60** (1981) no. 9, 203 – 208.
329. *H. B. Ries, Aufbereit. Tech.* **18** (1977) 633 – 640.
330. *Nitrogen* **193** (1990) 22 – 26.
331. Private communication.
332. R. S. Meline, I. W. McCamy, J. L. Graham, T. S. Sloan *J. Agr. Food Chem.* **16** (1968) 235 – 240.
333. *Nitrogen* **177** (1989) 21 – 24.
334. J. W. McCamy, M. M. Norton, *Farm Chem.* **140** (1977) no. 2, 61 – 68, 70.
335. *New Developments in Fertilizer Technology, 11th Demonstration*, Tennessee Valley Authority, Muscle Shoals, Ala., 1976, pp. 33 – 35.
336. *Nitrogen* 1975 no. 98, 37 – 39.
337. L. M. Nunnely, F. T. Carney: "Melt Granulation of Urea by the Falling-Curtain Process," ACS National Meeting, 182th, 1981.
338. W. Pietsch, *Aufbereit. Tech.* **12** (1971) 684 – 690.
339. C. Debayeux in [340] , p. 8.1 – 8.5.
340. V. Bizzotto, IFA Technical Conference 5 – 8 Nov. 1984, Paris, p. 7.1 – 7.14 (Reprints).
341. *New Developments in Fertilizer Technology, 8th Demonstration*, Tennessee Valley Authority, Muscle Shoals, Ala., 1970, pp. 41 – 47.
342. G. C. Hicks, J. M. Stinson, *Ind. Eng. Chem. Proc. Des. Dev.* **14** (1975) 269 – 276.
343. A. S. Maartensdijk: "Direct Production of Granulated Superphosphates and PK-Compounds," *New Dev. Phosphate Fert. Technol. Proc. Tech. Conf. ISMA Ltd., 1976* (1977).
344. G. H. Wesenberg: "Concentrated Superphosphate: Manufacturing Process" in [312]
345. F. P. Achorn, D. G. Salladay: "Granulation Using the Pipe-Cross Reactor" in [312]
346. J. A. Benes, A. Hemm: "Modern Process Design in Compound Fertilizer Plants", *Seminar of Trends and Developments in the Fertiliser Industry,* 1985.
347. G. C. Hicks: "Review of the Production of Monoammonium Phosphate" in [312]
348. G. H. Wesenberg: "Diammonium Phosphate Plants and Process" in [312]
349. J. Steen, E. A. Aasum, T. Heggeboe: "The Norsk Hydro Nitrophosphate Process" in [312]
350. *Phosphorus Potassium* 1975, no. 76, 48 – 54.
351. I. W. McCamy, G. C. Hicks, *Proc. Annu. Meet. Fert. Ind. Round Table*, 26th, 1976, 168 – 176.
352. G. C. Hicks, I. W. McCamy, B. R. Parker, M. M. Norton; "Basis for Selection of Granulators at TVA," paper presented at the *Proc. Annu. Meet. Fert. Ind. Round Table*, 28th, 1978.
353. H. Stahl, *Aufbereit. Tech.* **21** (1980) 525 – 533.
354. G. Heiseler, D. Baier, D. Juling, W. Kretschmer, *Chem. Tech. (Leipzig)* **25** (1973) 410 – 414.
355. L. Diehl, K. F. Kummer, H. Oertel: "Nitrophosphates with Variable Water Solubility: Preparation and Properties," paper read before the Fertiliser Society of London, 1986.
356. M. Autti, M. Loikkanen, P. Suppanen: "Economic and Technical Aspects in some Differential NPK-Processes," *New Dev. Phosphate Fert. Technol. Proc. Tech. Conf. ISMA Ltd*, 1976 (1977).
357. H. B. Ries, *Aufbereit. Tech.* **16** (1975) 17 – 26.
358. I. S. E. Martin, F. J. Harris in [310], pp. 253 – 270.
359. J. Kotlarevsky, J. Cariou in [310], pp. 245 – 252.
360. M. Kuwabara, S. Hayamizu, A. Hatakeyama in [310] pp. 125 – 147.
361. J. L. Medbery, F. T. Nielsson, *Ind. Eng. Chem. Prod. Res. Dev.* **22** (1983) 291 – 296.
362. B. Dencs, Z. Ormós, *Int. Symp. Agglomeration, 3rd, 1981*, G 2 – G 15.
363. A. Maroglou: "Fluidized Bed Granulation," Ph. D. Thesis, University of Birmingham 1985.
364. S. Mortensen, S. Hoomand, *Chem. Eng. Prog.* **79** (1983) no. 4, pp. 37 – 42.
365. S. Mortensen, A. Kristiansen, *Int. Symp. Agglomeration, 3rd, 1981*, F 77 – F 90.
366. H. B. Ries, *Aufbereit., Tech.* **11** (1970) 262 – 280.
367. I. S. Mangat in [310] pp. 148 – 161.
368. J. L Hawksley in [310] pp. 110 – 124.
369. P. J. Sherrington, *Can. J. Chem. Eng.* **47** (1969) June, 308 – 316. Links
370. A. I. More (ed): *Fert. Nitrogen: Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol., 4th, 1981*.
371. Y F. Berguin in [310] pp. 296 – 303.
372. P. J. Sherrington, R. Oliver: *Granulation*, Hlyden&Sons Ltd., London p. 172.

373. O. Uemaki, K. B. Mathur, *Ind. Eng. Chem. Process Des. Dev.* **15** (1976) 504 – 508.
374. F. E. Steenwinkel, J. W. Hoogendonk: "The Prilling of Compound Fertilisers," paper read before The Fertiliser Society (1969).
375. *Nitrogen* 1980, no. 127, 45.
376. P. J. van den Berg, G. Hallie: "New Developments in Granulation Techniques," *Proc. Fert. Soc.* 1960, no. 59.
377. *"Supra Today*," 1977, p. 23.
378. R. P. Ruskan, *Chem. Eng. (N. Y.)* **83** (1976) June 7, 114 – 118.
379. I. R. Schoemaker, I. A. C. M. Smit, "Criteria for a Choice between various Granulation and Prilling Technologies for Fertilisers," paper read before the Fertiliser Society of London (1981).
380. R. G. Lee, M. M. Norton, H. G. Graham, *Proc. Annu. Meet. Fert. Ind. Round Table*, 24th, 1974, 79.
381. R. G. Lee, R. S. Meline, R. D. Young, *Ind. Eng. Chem. Process. Des. Dev.* **11** (1972) 90 – 94.
382. C. J. Pratt, *Br. Chem. Eng.* **14** (1969) 185 – 188.
383. H. B. Ries, *Int. Symp. Agglomeration, 3rd*, 1981, F 34 – F 59.
384. H. Rumpf, *Chem. Ing. Tech.* **30** (1958) 329 – 336.
385. G. Nöltner, *Int. Symp. Agglomeration, 3rd, 1981*, F 16 – F 33xs.
386. H. Schubert, E. Heidenreich, F. Liepe, Th. Neeße, *Mechanische Verfahrenstechnik II*, VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1979, p. 204.
387. G. Heinze: "Verfahrenstechniken des Agglomerierens," Seminar of the VDI-Bildungswerk, Stuttgart-Vaihingen 1984.
388. J. Eimers: "Granuliertechniken bei der Düngemittelherstellung," paper presented at the congress AGRICHEM 81, Bratislava 1981.
389. H. B. Ries, *Aufbereit. Tech.* **11** (1970) 147 – 153.
390. H. B. Ries, *Aufbereit. Tech.* **11** (1970) 615 – 621.
391. H. B. Ries, *Aufbereit. Tech.* **11** (1970) 744 – 753.
392. *New Developments in Fertilizer Technology, 13th Demonstration*, Tennessee Valley Autority, Muscle Shoals, Ala., 1980, pp. 75 – 79.
393. R. J. Milborne, D. W. Philip, "Adapting a Pipe Reactor to a Blunger for NPK Production," paper read before the Fertiliser Society of London, 1986.
394. W. Pietsch, *Aufbereit. Tech.* **7** (1966) no. 4, 177 – 191.
395. G. Heinze, *Chem. Tech. Heidelberg* **15** (1986) no. 6, 16, 18, 21.
396. R. H. Perry, C. H. Chilton: *Chemical Engineers' Handbook*, 5th ed., McGraw-Hill, New York 1973 pp. 57 – 65.
397. J. O. Hardesty, *Chem. Eng. Prog.* **51** (1955) 291 – 295. Links
398. F. T. Nielsson: "Granulation" in [312]
399. Ch. Davis, R. S. Meline, H. G. Graham, *Chem. Eng. Prog.* **64** (1968) no. 5, 75 – 82. Links
400. *Farm Chemicals Handbook*, 1986, p. B 35.
401. R. D. Young, G. C. Hicks, C. H. Davis, *J. Agric. Food Chem.* **10** (1962) , no. 1, 68 – 72.
402. F. G. Membrillera, J. L. Toral, F. Codina in [310], pp. 162 – 173.
403. "The Cros Fertilizer Granulation Process", *Phosphorus Potassium* **1977,** no. 87, 33 – 36.
404. *Phosphorus Potassium* 1986, no. 144, 27 – 33.
405. J. W. Baynham: "The SAI-R Process for the Manufacture of Compound Fertilisers containing Ammonium Nitrate and Diammonium Phosphate", *Proc. ISMA Tech. Conf. 1965*.
406. *Nitrogen* 1970, no. 64, 25 – 26.
407. E. Pelitti, J. C. Reynolds in [310], pp. 95 – 109.
408. R. M. Reed, J. C. Reynolds, *Chem. Eng. Prog.* **69** (1973) no. 2, 62 – 66. Links
409. E. Vogel, *Nitrogen*, 1986, no. 161, 28 – 31.
410. *Nitrogen* 1987, no. 166, 39.
411. K. Sommer, W. Herrmann, *Chem. Ing. Tech.* **50** (1978) 518 – 524.
412. F. D. Young, T. W. McCamy, *Can. J. Chem. Eng.* **45** (1967) 50 – 56. Links
413. H. Klatt, *Zem. Kalk Gips* **11** (1958) no. 4, 144 – 154.
414. Tennessee Corp., US 769 058, 1985.
415. J. W. McCamy, M. M. Norton in [310], pp. 68 – 94.
416. St. Ruberg, *Aufbereit.-Tech.* **28** (1987) 75 – 81.
417. H. B. Ries, *Tech. Mitt.* **77** (1984) 583 – 589.
418. H. B. Ries, *Maschinenmarkt* **92** (1986) 25 – 28.
419. H. Stahl: "Press-Agglomeration," seminar of the VDI-Bildungswerk, Stuttgart-Vaihingen 1984.
420. S. Maier in [310], pp. 283 – 295.
421. A. Seixas, J. D. Ribeiro in [340] p. 11.1 – 11.21.
422. H. Rieschel, *Aufber. Tech.* **11** (1970) 133 – 146.
423. A. Tassin, Chr. Fayard, *Inform. Chem.* **269** (1986) no. 1, 127 – 131.
424. P. Heroien, C. Fayard, *Proc. Inst. Briquet. Agglom. Bienn. Conf.* **22** (1991) 173 – 185.
425. W. Herrmann: *Das Verdichten von Pulvern zwischen zwei Walzen*, Verlag Chemie, Weinheim, Germany 1973.
426. W. Herrmann, K. Sommer, *Int. Symp. Agglomeration, 3rd*, 1981, F2 – F15.
427. W. Herrmann, *Chem. Ing. Tech.* **51** (1979) 277 – 282.
428. G. J. Thorne, J. D. C. Hemsley, H. Hudson, *Chem. Process. (London)* **19** (1973) no. 1, 17 – 18.
429. I. Podilchuck, W. T. Charlton, M. D. Pask: *Modern Approach to the Design of Granular Fertilizer Plants*, Fisons Ltd., Fertilizer Division, 1975.
430. H. Vernède, *Nitrogen* 1969, no. 60, 29 – 33.
431. Sahut-Conreur, *Verfahrenstechn.* 1987, no. 5, 8, 10, 12.
432. G. Hoffmeister: *"Physical Properities of Fertilizers & Methods for Measuring Them"*, TVA Bulletin Y 147, 1979.
433. L. Jäger, P. Hegner, *Fert. Technol.* **1** (1985) 191 – 221.
434. K. Niendorf, W. Teige, *Chem. Tech. (Leipzig)* **26** (1974) 771 – 774.
435. M. Brübach, A. Göhlich, *Verfahrenstechnik* **7** (1973) 268 – 272.
436. P. Runge, *Chem. Tech. (Leipzig)* **28** (1976) 163 – 165.
437. J. Silverberg, *Proc. Ann. Meet. Fert. Ind. Round Table 16th* 1966, 81.
438. D. C. Thompson: "Fertiliser Caking and its Prevention," *Proc. Fert. Soc.* 1972, no. 125.
439. O. Kjöhl, A. Munthe-Kaas: "Study of the Mechanism of Caking in some Nitrophosphate Fertilizers," *Proc. ISMA Tech. Conf.*, 1970.
440. G. C. Lowrison: *Fertilizer Technology*, Ellis Horwood Limited, Chichester, 1989, p. 406.
441. P. A. Mackay, K. S. Sharples: "The Use of Special Oils and Coatings to Prevent Caking of Fertilizers," *Proc. Fert. Soc.* 1985, no. 239.
442. *Nitrogen* 1986, no. 164, 33 – 34.
443. *Nitrogen* 1990, no. 187, 20 – 27.
444. *Nitrogen* 1987, no. 166, 19.

445. E. Alto, P. Suppanen: “Dust Control in NPK Production,” *Proc*, ISMA Tech. Conf. 1974.
446. J. Baretincic: “Environmental Regulations,” in [312]
447. J. E. Seymour, in [310], pp. 174 – 180.
448. *Bulk Blend Quality Control Manual*, The Fertilizer Institute, Washington, D.C., 1975.
449. L. G. Croysdale, W. E. Samuels, J. A. Wagner, *Chem. Eng. Prog.* **61** (1965) 72 – 77. Links
450. G. Hoffmeister, “Bulk Blending,” in [312]
451. L. Taylor, “Latest Technology and Methods Used for Bulk Blending and Fertilizer Storage in the United States,” paper read before the Fertiliser Society of London, 1987.
452. H. W. Frizen: “Verfahrensketten für feste und flüssige Mehrnährstoffdünger in den USA,” Dissertation, Universität Bonn, 1974.
453. K. H. Ullrich, *Vereinigte Landwarenkaufleute* 1979, no. 3.
454. Official Bulletin of the European Community, 20th ed., No. L213 from 22 August 1977, 77/535/EEC: Guidelines of the commission from 22 June 1977 on the approximation of the laws of the member states concerning sampling and methods of analysis of fertilizers. The Guidelines are extended and modified in: Bulletin L265 from 12.09.89, 89/519/EEC for Ca., Mg, Na and S; and Bulletin L113 from 07.05.93, 93/1/EEC for trace elements B, Co, Cu, Fe, Mn, Mo, Zn.
455. Methodenbuch, Band II: Die Untersuchung von Düngemitteln, [Book of Methods. Vol. II: The Analysis of Fertilizers] Verband Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (LUFA) [Association of German Agricultural Research Institutes] LUFA-Verlag, Darmstadt, 1995.
456. ISO (International Organization for Standardization).
457. Official Methods of Analysis of AOAC International, 16th ed, 3rd rev., AOAC International, Gaithersburg, Maryland, USA, 1997.

General References

458. B. Wermingshausen: *Kunststoffe im Landbau*, Kohlhammer-Verlag, Stuttgart-Berlin-Köln-Mainz 1985.
459. H. Baumann: *Plastoponik, Schaumkunststoffe in der Agrarwirtschaft*, Dr. Alfred Hüthig Verlag, Heidelberg 1967.
460. A. Kullmann: *Synthetische Bodenverbesserungsmittel*, VEB Deutscher Landwirtschaftsverlag, Berlin 1972.
461. M. G. Voronkov, G. I. Zelchan, E. Lukevitz in K. Ruhlmann (ed.): *Silizium und Leben, Biochemie, Toxikologie und Pharmakologie des Si*, Akademie-Verlag, Berlin 1975.

Specific References

462. H. Prün, *Rasen Turf Gazon* **17** (1986) no. 1, 120 – 127.
463. B. Werminghausen, *Chemie und Technik in der Landwirtschaft* 1962, 359 – 360.
464. H. Knobloch, *Wasser und Boden* **12** (1964) 403 – 405.
465. F. Penningsfeld, *Das Gartenamt* **28** (1979) 1 – 4.
466. J. Günther, *Gb + Gw*, Gärtnerbörse und Gartenwelt, 81 (1981) no. 31, 714 – 716.
467. B. Werminghausen, private communication, 1987.
468. B. Werminghausen, *Mitt. Landbau* 1964, no. 4, 1 – 8.
469. BASF, DE 1 245 205, 1962 (F. Stastny, J. Jung)
470. H. Will, paper presented at the meeting of the Gesellschaft für industriellen Pflanzenbau, Wien 1975.
471. H. Knobloch, *Wasser Boden* **18** (1960) 415 – 418.
472. H. Knobloch, *Z. Kulturtech. Flurbereinig.* **8** (1967) 54 – 59.
473. E. Zinck, R. Krötz, *Umweltmagazin* 1980, no. 3, 38 – 40.
474. E. Zinck, DE 2 742 606, 1977.
475. H. Rasp, *Z. Pflanzenernähr. Bodenk.* **133** (1972) no. 1/2, 111 – 123.
476. H. Rasp, *Mitteilungen Deutsche Bodenkundl. Gesellschaft* **15** (1972) 181 – 183.
477. W. Tepe, private communication based on an internal report (1976).
478. S. Maier, *Landwirtsch. Forsch. Sonderh.* **18** (1970) 14 – 20.
479. B. Werminghausen, *Acta Hortic.* **26** (1972) 159 – 164.
480. A. Buchner, J. Jung, P. Weisser, H. Will, *Landwirtsch. Forsch.* **22** (1969) no. 2, 94 – 99.
481. J. F. Kitchens, R. E. Casner, G. S. Edwards, W. E. Harward III, B. J. Macri, EPA 560/2–76–009, ARC 49–5681, Springfield, Va., 1976, pp. 1 – 200.
482. W. Skirde, *Z. Vegetationstechnik* **5** (1982) 1 – 6.
483. L. Unterstenhöfer, *Kunststoffe* **57** (1967) no. 11, 850 – 855.
484. F. Doehler, DE 839 944, 1948.
485. P. di Dio, paper presented at the Congress Chemistry in Agriculture, 2nd., Pressburg/Bratislawa 1972.
486. *Guide to Evaluating Sands and Amendments Used for High Trafficked Turfgrass*, University of California, Agricultural Extension, AXT no. 113 1970.
487. W. Ganssmann, *Phosphorsäure* **22** (1962) 223 – 241.
488. E. Seifert, Mitteilungen Leichtweiss-Institut Technische Universität Braunschweig, **1970,** nos. 25 and 27.
489. T. Niedermaier, *Landwirtsch. Forsch.* **32** (1979) no. 1/2, 129 – 137.
490. H. Gebhardt, *Mitteilungen Deutsche Bodenkundl. Gesellschaft* **15** (1972) 225 – 245.
491. W. Büring, *Das Gartenamt* **23** (1974) no. 5, 258 – 281.
492. K. Wiede, Dissertation, Bonn 1976.
493. W. Büring, *Rasen und Rasengräser.* 1969, no. 6, 78 – 83.
494. B. Walter, E. Petermann, *Weinberg und Keller* **21** (1974) no. 12, 513 – 532.
495. A. H. Sayegh, A. M. Osman, B. Hattar, M. Barasiri, M. Chaudri, M. Nimeh, *Agrochimica*, **26** (1982) no. 1, 22 – 30. Links
496. H. H. Becher, R. Woytek, *Z. Kulturtech. Flurbereinig.* **25** (1984) 46 – 53.
497. E. Schulze, unpublished reports of the Institut für Pflanzenbau, Universität Bonn, 1970 – 1977.
498. W. Büring, *Das Gartenamt* **31** (1982) 659 – 667.
499. A. Okuda, E. Takahashi, *Nippon Dojo Hiryogaku Zasshi* **32** (1961) 533.
500. L. H. P. Jones, K. A. Handreck, *Adv. Agron.* **19** (1967) no. 19, 107 – 149. Links
501. J. Vlamis, D. E. Williams, *Plant Soil* **27** (1967) 131. Links
502. H. Rasp, *Landwirtsch. Forsch. Sonderh.* **35** (1978) 394 – 411.
503. W. Friedmeann, W. Pesch, private communication (1979).
504. H. O. Leh, *Gesunde Pflanze* **38** (1986) no. 6, 246 – 256.
505. Henkel, DE 2 208 917, 1972 (E. Seifert, H. von Freyhold).
506. Guano-Werke, DE 2 612 888, 1976 (J. Bochem, G. Eisenhauer).
507. J. Cronjaeger, A. Hardt, H. Bernard, *Gesunde Pflanze* **45** (1993) 7.
508. M. Kretschmer, K. Schaller, K. Wiegand, *Landwirtsch. Forsch.* **35** (1982) , no. 1 – 2, 66 – 76.
509. H. Prün, Landwirtsch. Forsch. *Sonderh.* **30** (1973) 288 – 296.
510. B. Meyer, *Mitteilungen Deutsche Bodenkundl. Gesellschaft* **15** (1972) 253 – 258.
511. F. Callebaut, D. Gabriels, M. D. Boodt, *Med. Fac. Landbouwwet., Rijksuniv. Gent* **49** (1984) no. 1, 53 – 61.

512. E. McGuire, R. N. Carrow, J. Troll, *Agron. J.* **70** (1978) 317 – 348.
513. M. J. Carrol, A. M. Petrovic, *Proceed. Int. Turfgrass Research Conference, 5th*, Avignon 1985, pp. 381 –389.
514. L. M. Rivière, S. Charpentier, B. Valat, *Rev. Horic.* **267** (1986) 43 – 49.
515. H. Prün, paper presented before the Deutsche Rasengesellschaft, Heidelberg 1984.
516. F. Masuda, *CEER Chem. Econ. Eng. Rev.* **15** (1983) no. 11, 19 – 22.
517. R. L. Flannery, W. J. Busscher, *Comm. Soil Sci. Plant Anal.* **13** (1982) no. 2, 103 – 111.
518. *Kirk-Othmer* 4th ed., vol. **22.**
519. C. A. York, N. A. Baldwin, *J. Sports Turf Res. Inst.* **68.**
520. C. A. York, N. A. Baldwin, *International Turfgrass Society Research Journal* **7** (1993).
521. Gefahrstoffverordnung vom 26.08.1986 (letzte Anpassung 06.98) ergänzt durch technische Regeln (TRGS511), Carl Heymann-Verlag, Köln.
522. M. Brübach: "Der Einfluß der Korngröße, der Granulatfestigkeit und der Reibung auf die Verteilung von Dünger- und Pflanzenschutzgranulaten," Dissertation, Berlin 1973.
523. A. Buchner, H. Sturm: *Gezielter düngen*, DLG-Verlag, Frankfurt 1985.
524. K. Rühle, "Die Verteilgenauigkeit pneumatischer Mineraldüngerstreuer," *KTBL-Schrift* no. 198, Darmstadt-Kranichstein 1975.
525. W. Werner, H. W. Olfs, K. Auerswald, K. Isermann: "Stickstoff- und Phosphoreintrag in Oberflächengewässer über diffuse Quellen," in A. Hamm (ed.): *Studie über die Wirkungen und Qualitätsziele von Nährstoffen in Fließgewässern*, Academia Verl., St. Augustin 1991, pp. 665 – 764.
526. W. Werner, H. P. Wodsack: "Stickstoff und Phosphoreintrag in die Fließgewässer Deutschlands unter besonderer Berücksichtigung des Eintragsgeschehens im Lockergesteinsbereich der ehemaligen DDR," *Agrarspektrum*, vol. **22**, DLG-Verl., Frankfurt 1994.
527. O. Strebel: "Untersuchungen zur Grundwasserbelastung bei land- und forstwirtschaftlicher Bodennutzung von Sandböden im Raum Hannover," *Z. Kulturtech. Flurbereinig.* **18** (1977) 310 – 312.
528. P. Obermann: "Hydrochemische/hydromechanische Untersuchungen zum Stoffgehalt von Grundwasser bei landwirtschaftlicher Nutzung," *Bot. Mitt. Dtsch. Gewässerkd.* **42** (1981) 1 – 217.
529. P. Obermann: "Die Belastung des Grundwassers aus landwirtschaftlicher Nutzung nach heutigem Kenntnisstand," in H. Nieder (ed.): *Nitrat im Grundwasser. Herkunft, Wirkungen, Vermeidung,* Verlag Chemie, Weinheim 1985, pp. 53 – 64.
530. P. Obermann, G. Einars, H. Zakosek, G. Schollmayer: *Hydrogeologische/bodenkundliche Untersuchungen im Bereich des Schüttensteiner Waldes bei Bocholt/Westfalen, 1st interim report to the Bundesministerium für Ernährung, Landwirtschaft und Forsten (Ministry of Nutrition, Agriculture, and Forestry): Beitrag der Landwirtschaft zur Verminderung der Nitratbelastung des Grundwassers*, **parts B2 und E**, 1985.
531. P. Obermann, W. Leuchs, G. Einars, H. Zakosek, G. Schollmayer, S. Schulte-Kellinghaus, H. Zepp: *Hydrogeologische/bodenkundliche Untersuchungen im Bereich des Schüttensteiner Waldes bei Bocholt/Westfalen*, final report, 1986.
532. P. Obermann, W. Leuchs: *Chemische Genese eines pleistozären Porengrundwasserleiters bei Bocholt/Westfalen, final report*, 1987.
533. *LWA Nordrhein-Westfalen: Grundwasserbericht 85,* Landesamt für Wasser- und Abfallwirtschaft Nordrhein-Westfalen, Düsseldorf 1985.
534. U. Rohmann: "Landwirtschaftliche und aufbereitungstechnische Maßnahmen zur Lösung des Nitratproblems," DVGW-Schriftenr. no. 46 (1985) 105 – 123.
535. K. Isermann: "Vertikale Verteilung von Nitrat- und Denitrifikationsparametern bis in 10 m Bodentiefe bei unterschiedlicher Landbewirtschaftung," *Herbsttagung Limburgerhof 1986*. K. Isermann: "Vertical Distribution of NO_3-N and Denitrification Parameters in the Soil Down to Greater Depths (about 10 m) under Various Soil Management/Nitrogen Fertilization," International Symposium of CIEC Agricultural Waste Management and Environmental Protection, 4th, Braunschweig 1987.
536. I. F. I. Gast, J. Taat, W. van Duijvenbooden: "Landelijk meetnet — grondwaterkwaliteit," Rijksinstituut voor volksgezondheit en milieuhygiene 2, DEEL Rapport no. 840 382002, April 1985.
537. LAWA – Länderarbeitsgemeinschaft Wasser: "Nitrat," *Bericht zur Grundwasserbeschaffenheit*, Umweltministerium Baden-Württemberg, Stuttgart 1995. LAWA – Länderarbeitsgemeinschaft Wasser: Ergebnisse der Überwachung der Gewässer gemäß Artikel 5 Abs. 6 Satz 2 der Richtlinie 91/676/EWG. Anlage 3 der Mitteilung der Bundesrepublik Deutschland (BMU) vom 25. Oktober 1996 über die Umsetzung der Richtlinie 91/676/EWG.
538. EU Nitrate Directive 19/676.
539. Bocholter Energie und Wasserversorgung GmbH, personal communication, August 17, 1998.
540. U. Rohmann, H. Sontheimer: *Nitrat im Grundwasser. Ursachen, Bedeutung, Lösungswege*, DVGW Forschungsstelle am Engler-Bunte-Institut, Karlsruhe 1985.
541. Der Rat von Sachverständigen für Umweltfragen: *Umweltprobleme der Landwirtschaft, Sondergutachten März 1985*, Verlag W. Kohlhammer, Stuttgart-Mainz 1985.
542. D. Sauerbeck: *Funktionen, Güte und Belastbarkeit des Bodens aus agrikulturchemischer Sicht. Materialien zur Umweltforschung*, published under the auspices of the Rat von Sachverständigen für Umweltfragen, Verlag W. Kohlhammer, Stuttgart-Mainz, 1985.
543. J. Jung, J. Dressel: "Zum Stand des Nitratproblems," *Herbsttagung Limburgerhof 1981*, 1 – 7.
544. A. Vömel: "Stickstoffdüngung und Stickstoffverluste auf ackerbaulich genutzten Böden," *Landwirtsch*. no. 197 Sonderh. (1981) 205 – 222.
545. O. J. Furrer, 1986: "Einfluß von Fruchtfolge und Düngung auf den Nitrateintrag ins Grundwasser," vol 10 C, Pro Aqua-Pro Vita, Internationale Fachmesse und Fachtagungen für Umweltschutztechnik und ökologie, 10th, Basel 1986.
546. W. Burghardt, cited in H. Kuntze: "Möglichkeiten und Grenzen des Bodenschutzes durch den Landwirt — aus ackerbaulich-bodenkundlicher Sicht," *Bodenschutz — mit der Landwirtschaft*, Arbeiten der DLG **185** (1986) 33 – 35.
547. O. Strebel, J. Böttcher, W. H. M. Duynisfeld: "Einfluß von Standortbedingungen und Bodennutzung auf Nitratauswaschung und Nitratkonzentration des Grundwassers," *Landwirtsch. Forsch., Sonderh.* **37** (1984) 34 – 44.
548. O. Strebel, J. Böttcher: "Ermittlung von Ursachen und Prognosen der mittleren Nitratkonzentration im Grundwasser von Wassereinzugsgebieten," *Mitt. Dtsch. Bodenkd. Ges.* **53** (1987) 305 – 307.
549. F. Hess: "Acker- und pflanzenbauliche Strategien zum verlustfreien Stickstofftransfer beim Anbau im Organischen Landbau," *Mitt. Ges. Pflanzenbauwiss.* **3** (1990) 241 – 244.
550. I. Günther, Farm Economic Data Working Group, Assoc. of Sugar Beet Growers, personal communication, 1995.
551. M. Bach, H. G. Frede, G. Lang: "Entwicklung der Stickstoff-, Phosphor- und Kaliumbilanz der Landwirtschaft in der Bundesrepublik Deutschland," Studie der Gesellschaft f. Boden- und Gewässerschutz e.V., Wettenberg 1997.
552. H. Behrendt: "Detection of anthropogenic Trends in Time series of river rhine lead using Windows of Discharge and long-term Means," Report of the ICES/OSPAR-Workshop "Identification of statistical methods for temporal trends" Kopenhagen, 25 – 26. Sept. 1997, pp. 20 – 24.
553. Umweltbundesamt Berlin (UBA): Forschungsbericht 1994, Berlin, 1995.
554. H. G. Frede, S. Dabbert (eds.): *Handbuch zum Gewässerschutz*, Landwirtschaft. Verlagsges. ECOMED, Landberg. 1998
555. H. Carlson: "Quality Status of the North Sea," *Dtsch. Hydrogr. Z.* (1986) suppl. vol. B, no. 16.
556. E. Schwedhelm, G. Irion: "Schwermetalle und Nährelemente in den Sedimenten der deutschen Nordseewatten," CFS Cour. Forschungsinst. Senkkenberg no. 73 (1985).
557. K. G. Malle: "Wie schmutzig ist die Nordsee?" *Chem. Unserer Zeit* **21** (1987) no. 1, 9 – 16. Links
558. G. Müller, J. Dominik, R. Reuther, R. Malisch, E. Schulte, L. Acker: "Sedimentary Record of Environmental Pollution in the Western Baltic Sea," *Naturwissenschaften* **67** (1980) 595 – 600.
559. G. Müller: "Zur Chronologie des Schadstoffeintrages in Gewässer," *Geowiss. Unserer Zeit* **1** (1983) 2 – 11.
560. P. J. Crutzen: "Atmospheric Interactions — Homogeneous Gas Reactions of C, N, and S Containing Compounds," in B. Bolin, R. B. Cook (eds.): *The Major Biochemical Cycles and their Interactions,* J. Wiley & Sons, Chichester, England 1983, pp. 65 –114.
561. T. Granli, O. C. Bøckman: "Nitrous Oxide from Agriculture," *Norwegian J. Agric. Sci.* (1994) Suppl. 12.
562. R. F. Weiss: "The Temporal and Spatial Distribution of Tropospheric Nitrous Oxide," *J. Geophys. Res. C* **86** (1981) 7185 – 7195.

563. R. J. Stevens, R. J. Laughlin, L. C. Burns, J. R. M. Arah, R. C. Hood: “Measuring the Contributions of Nitrification and Denitrification to the Flux of Nitrous Oxide from Soil,” *Soil. Biol. Biochem.* **29** (1996) 139 – 151. Links
564. E. A. Kaiser, O. Heinemeyer: “Temporal Changes in N_2O-Losses from two arable Soils,” *Plant and Soil* **181** (1996) 57 – 63. Links
565. F. C. Thornton, R. J. Valente: “Soil Emissions of Nitric Oxide and Nitrous Oxide from no-till Corn,” *Soil Sci. Soc. Am. J.* **60** (1996) 1127 – 1133.
566. E. A. Kaiser *et al.*, “What predicts Nitrous Oxide Emission and Denitrification N-Loss from European Soils?” *Z. Pflanzenernähr. Bodenk.* **159** (1996) 541 – 547.
567. E. A. Davidson, P. A. Matson, P. D. Brooks: “Nitrous Oxide Emissions controls and inorganic Nitrogen Dynamics in fertilized tropical agricultural Soils,” *Soil Sci. Soc. Am. J.* **60** (1996) 1145 – 1152. Links
568. K. A. Smith, I. P. McTaggart, H. Tsuruta, K. Smith: “Emissions of N_2O and NO associated with Nitrogen Fertilization in intensive Agriculture, and the Potential for Mitigation,” *Soil Use Manage.* **13** (1997) 296 – 304.
569. M. J. Eichner: “Nitrous Oxide Emissions from fertilized Soils: Summary of available Data,” *J. Environ. Qual.* **19** (1990) 272 – 280.
570. A. F. Bouwman (ed.): *Soils and the Greenhouse Effect*, Wiley and Sons, New York 1990.
571. C. Henault, X. Devis, J. L. Lucas, J. C. Germon: “Influence of different agriculture Practices (type of Crop, form of N-Fertilizer) on Soil Nitrous Oxide Emission,” *Biol. Fertil. Soils* **27** (1998) 299 – 306. Links
572. K. Isermann: “Agriculture's share in the emission of trace gases affecting the climate and some source-oriented proposals for sufficiently reducing this share”, *Environm. Pollution* **83** (1994) 95 – 111.
573. K. L. Weier, J. W. Doran, J. F. Power, D. T. Walters: “Denitrification and the Dinitrogen/Nitrous Oxide Ratio as affected by Soil Water, available Carbon, and Nitrate,” *Soil Sci. Soc. Am. J.* **57** (1993) 66 – 72. Links
574. J. C. G. Ottow, G. Benckiser, M. Kapp, G. Schwarz: “Denitrifikation — die unbekannte Größe,” *DLG-Mitteilungen* **105** (1990) 8 – 10.
575. H. Flessa, W. Pfau, P. Dörsch, F. Beese: “The influence of Nitrate and Ammonium Fertilization on N_2O Release and CH_4 Uptake of a well-drained Topsoil demonstrated by a Soil Microcosm Experiment,” *Z. Pflanzenernähr. Bodenk.* **159** (1996) 499 – 503.
576. H. Clayton, I. P. McTaggart, J. Parker, L. Swan: “Nitrous Oxide Emission from fertilized Grassland: a 2-Year Study of the Effects of N Fertilizer Form and environmental Conditions,” *Biol. Fertil. Soil* **25** (1997) 252 – 260.
577. V. Cole *et al.*: “Agricultural Options for Mitigation of Greenhouse Gas Emissions,” in R. T. Watson, M. C. Zinyowera, R. H. Moss (eds.): *Climate Change 1995*, "Chap. 23", Cambridge University Press, Cambridge 1996, pp. 745 – 771.
578. R. Lipschultz, O. C. Zafiriou, S. C. Wofsy, M. B. McElroy, F. W. Valois, S. W. Watson: “Production of NO and N_2O by Soil Nitrifying Bacteria,” *Nature* **294** (1981) 641 – 643. Links
579. C. A. Stutte, R. T. Weiland: “Gaseous Nitrogen Loss and Transpiration of Several Crops and Weed Species,” *Crop. Sci.* **18** (1978) 887 – 889. Links
580. C. A. Stutte, R. T. Weiland, A. R. Blem: “Gaseous Nitrogen Loss from Soybean Foliage,” *Agron. J.* **71** (1979) 95 – 97.
581. J. V. Dean, J. E. Harper: “Nitric Oxide and Nitrous Oxide Production by Soybean and Winged Bean During the In Vivo Nitrate Reductase Assay,” *Plant Physiol.* **82** (1986) 718 – 723. Links
582. I. E. Galbally: “Factors controlling NO_x emissions from soils”, *Exchange of trace gases between terrestrial ecosystems and the atmosphere*, John Wiley and Sons, New York 1989 pp. 23 – 27.
583. T. Yoneyama, A. Hashimoto, T. Totsuka: “Absorption of Atmospheric NO_2 by Plants and Soils. 4. Two Routes of Nitrogen Uptake by Plants from Atmospheric NO_2: Direct Incorporation into Aerial Plant Parts and Uptake by Roots after Absorption into the Soil,” *Soil Sci. Plant Nutr. (Tokyo)* **26** (1980) no. 1, 1 – 7.
584. R. Lensi, A. Chalamet: “Absorption de l'oxyde nitreux par les parties aeriennes du mais,” *Plant Soil* **59** (1981) 91 – 98.
585. J. C. Ryden: “N_2O Exchange between a Grassland Soil and the Atmosphere,” *Nature* **292** (1981) 235 –237. Links
586. H. H. Rogers, V. P. Aneja: “Uptake of Atmospheric Ammonia by Selected Plant Species,” *Environ. Exp. Bot.* **20** (1980) 251 – 257. Links
587. M. J. Koziol, R. R. Whatley: *Gaseous Air Pollutants and Plant Metabolism*, Butterworths, Borough Green, Kent, England, 1984.
588. F. G. M. Roelofs, A. J. Kempers, A. L. F. M. Houdijk, J. Jansen: “The Effect of Air Borne Ammonium Sulphate on Pinus Nigra var. maritima in the Netherlands,” *Plant Soil* **84** (1985) 45 – 56. Links
589. J. G. M. Roelofs, A. W. Boxman, H. E. G. Van Dijk: “Effects of Airbone Ammonium on Natural Vegetation and Forests,” *Proceedings of the EEC Symposium*, Grenoble 1987.
590. K. Isermann: “Bodenversauerung beim natürlichen Wald- und Grasland-ökosystem von Rothamsted (England) unter dem Einfluß atmosphärischer und pedogener Protonen-Anlieferung von 1880 bis heute,” a paper presented at the Kongreß der Internationalen Bodenkundlichen Gesellschaft, 13th, Hamburg 1986 and published in *AFZ, Allg. Forstzeitschr.* **1987,** pp. 379 – 381.
591. K. Isermann: “Die Rolle des Stickstoffs bei den “neuartigen” Waldschäden,” paper presented at the Symposium Waldschäden am Biologischen Institut II der Universität Freiburg, 3rd, 1986.
592. L. De Temmermann, A. Ronse, K. Van den Cruys, K. Meeus-Verdinne: “Ammonia and Pine Tree Dieback in Belgium,” *Proceedings of the EEC Symposium*, Grenoble 1987.
593. W. M. J. Den Boer: “Ammonia, Not only a Nutrient but also a Cause of Forest Damages,” Wiss. Symposium zum Thema Waldschäden, Neue Ursachenhypothesen, Text 19, Umweltbundesamt, Berlin 1986.
594. H. Ellenberg: “Veränderungen der Flora unter dem Einfluß von Düngung und Immissionen,” *Schweiz. Z. Forstwes.* **136** (1985) 19 – 39.
595. G. Büttner, N. Lamersdorf, R. Schultz, B: Ulrich: *Deposition und Verteilung chemischer Elemente in küstennahen Waldstandorten — Fallstudie Wingst; Abschlußbericht, report of the Forschungszentrum Waldökosysteme/Waldsterben, series B, vol. 1*, 1986.
596. K. Kreutzer, J. Bittersohl: “Stoffauswaschung aus Fichtenkronen (Picea abies (L.) Karst.) durch saure Beregnung,” *Forstwiss. Centralbl.* **105** (1986) no. 4, 357 – 363.
597. M. Hauhs: “Die regionale Bedeutung des Nitrataustrages unter Wald für die Gewässerversauerung,” DFG-Workshop Modellierung des bodeninternen Stickstoffumsatzes, Hohenheim 1987.
598. G. L. Hutchinson, F. G. Viets: “Nitrogen Enrichment of Surface Water by Adsorption of Ammonia Volatilized from Cattle Feedlots,” *Science (Washington, D.C.)* **166** (1969) 514 – 515.
599. J. Gehrmann: “Derzeitiger Stand der Belastung von Waldökosystemen in Nordrhein-Westfalen durch Deposition von Luftverunreinigungen,” *Forst-Holzwirt.* **42** (1987) no. 6, 141 – 145.
600. S. Mehlert: “Untersuchungen zur atmogenen Stickstoffdeposition und zur Nitratverlagerung”, UFZ-Bericht 22, UFZ-Umweltforschungszentrum Leipzig-Halle GmbH, 1996.
601. L. Grünhage, H. J. Jäger, K. Freitag, K. Hanewald: “Emissionskataster Hessen — Landesweite Abschätzung der Emissionen aus biogenen und nicht gefaßten Quellen,” Umweltplanung, Arbeits- und Umweltschutz, no. 184, Hessische Landesanstalt für Umwelt, Wiesbaden 1996.
602. ECETOC (ed.): “Ammonia Emissions to Air in Western Europe” Tech. Rep. No. 62, European Centre for Ecotoxicology and Toxicology of Chemicals, Brussels 1996.
603. “Gasförmige Stickstoffverbindungen aus der Landwirtschaft”, Bericht für die DFG-Senatskommission zur Beurteilung von Stoffen in der Landwirtschaft, Schlußentwurf, Bonn April 1996.
604. P. A. Steudler, R. D. Bowden, J. M. Melillo, J. D. Aber: “Influence of nitrogen fertilization on methane uptake in temperate forest soils”, *Nature* **341** (1989) 314 – 316.
605. S. Hansen, J. E. Maehlum, L. R. Bakken: “N_2O and CH_4 fluxes influenced by fertilization and tractor traffic”, *Soil Biol. Biochem.* **259** (1993) 621 – 630.
606. B. W. Hütsch, C. P. Webster, D. S. Powlson: “Long-term effects of nitrogen fertilization on methane oxidation in soil of the Broadbalk Wheat Experiment”, *Soil Biol. Biochem.* **25** (1993) 1307 – 1315.
607. K. F. Bronson, A. R. Mosier: “Suppression of methane oxidation in aerobic soil nitrogen inhibitors and urease inhibitors”, *Biol. Fertil. Soils* **17** (1994) 68 – 84.

608. C. Bédard, R. Knowles: "Physiology, biochemistry and specific inhibitors of CH_4, NH^+_4 and CO oxidation by methanotrophs and nitrifier," *Microbiol. Rev.* **53** (1989) 68 – 94. Links
609. K. Mengel, E. A. Kirkby: *Principles of Plant Nutrition*, International Potash Institute, Worblaufen-Bern 1982.
610. G. Pfützer, C. Pfaff: "Einfluß der Düngung und Belichtung auf die Vitaminbildung in der Pflanze," *Umschau (1897)* **39** (1935) 917.
611. C. Pfaff, G. Pfützer: "über den Einfluß der Ernährung auf den Carotin- und Ascorbinsäuregehalt verschiedener Futterpflanzen," *Z. Angew. Chem.* **50** (1937) 179.
612. G. A. Rodgers, A. Penny, F. V. Widdowson, M. V. Hewitt: "Tests of Nitrification and of Urease Inhibitors, when Applied with Either Solid or Aqueous Urea, on Grass Grown on a Light Sandy Soil," *J. Agric. Sci.* **108** (1987) 109 – 117.
613. R. Prakasa, K. Puttanna: "Nitrification and Ammonia Volatilisation Losses from Urea and Dicyandiamide-treated Urea in a Sandy Loam Soil," *Plant Soil* **97** (1987) 201 – 206.
614. H. Bernhard: *Phosphor: Wege und Verbleib in der Bundesrepublik Deutschland. Probleme des Umweltschutzes und der Rohstoffversorgung*, Verlag Chemie, Weinheim-New York 1978.
615. A. Hamm: "Wie und woher kommen die Nährstoffe in die Flüsse?" in J. L. Lozan, H. Kausch (eds.): *Warnsignale aus Flüssen und Ästuaren, Wissenschaftliche Fakten*, Verl. Paul Parey, Berlin – Hamburg 1996 pp. 105 – 110.
616. A. Breeuwsma, J. G. A. Reijerink, O. F. Schoumans: Fosfaatverzadigde gronden in het oostelijk, central en zuidelijk zandgebied, Report 68, Staring Centrum, Wageningen 1990.
617. U. Pihl, W. Werner: "Erhebungsuntersuchungen zu Phosphatgehalten, Phosphatsorptionskapazität und relativer Phosphatsättigung der Böden in veredlungsstarken Gebieten Nordrhein-Westfalens als Prognosekriterien des potentiellen Phosphataustrags in Drain- und Grundwasser," Forschungsber. 64, Lehr- und Forschungsschwerpunkt "Umweltverträgliche und standortgerechte Landwirtschaft", Landwirtschaftliche Fakultät, Universität Bonn, Bonn 1998.
618. Th. Diez, K. Bucksteeg: "P-fluxes in a Watershead Area with Intensive Agriculture and Eutrophication of a Lake," paper presented at the Int. Symposium of CIEC Agricultural Waste Management and Environmental Protection, 4th, Braunschweig-Voelkenrode 1987.
619. W. Werner: "Implementation and efficiency of countermeasures against diffuse nitrogen and phosphorus inputs into surface waters from agriculture", in E. Romstad, J. Simonsen, A. Vatn (eds.): *Controlling mineral emissions in European agriculture*, CAB International, Oxon – New York 1997 , pp. 73 – 88.
620. Industrieverband Agrar e.V., personal communication, 1997.
621. K. Isermann: "Einfluß der Phosphatdüngung auf den Cadmiumgehalt des Bodens, ermittelt anhand zahlreicher Dauerversuche in Westeuropa," *Landwirtsch. Forsch. Sonderh.* **39** (1982) 283 – 301.
622. B. Machelett, W. Podlesak, J. Garz: "Zur Wirkung des Cadmiumgehalts im Düngerphosphat auf die Cadmiumkonzentration in Boden und Pflanze in einem langjährigen Phosphatdüngungsversuch," *Arch. Acker-Pflanzenbau Bodenkd.* **28** (1984) no. 4, 247 – 251.
623. J. J. Mortvedt: "Cadmium Levels in Soils and Plant Tissues from Long Term Soil Fertility Experiments in the United States," *Congr. Int. Soc. Soil Science, 13th*, vol. **3**, Hamburg 1986.
624. W. Werner, C. Brenk: "Entwicklung eines integrierten Nährstoffkonzeptes als Basis eines umweltverträglichen, flächendeckenden Recyclings kommunaler Abfälle in Nordrhein-Westfalen und regionalisierte Bilanzierung der Schwermetallflüsse", Forschungsber. 48, Landwirtschaftliche Fakultät Universität Bonn, Bonn 1997.

General References

625. *FAO Fertilizer Yearbook* , Food and Agricultural Organization of the United Nations, Rome.
626. *FAO Production Yearbook* , Food and Agricultural Organization of the United Nations, Rome.

Specific References

627. Q. Paris: "The von Liebig-Hypothesis," *Am. J. Agric. Econ*, **74** (1992) 4, 1019 – 1028.
628. E. von Boguslawski: *Ackerbau, Grundlagen der Pflanzenproduktion*, Frankfurt/M. 1981.
629. K. Baeumer: "Düngung", in R. Diercks, R. Heitefuss (eds.): *Integrierter Landbau*, 2nd ed., München 1994 pp. 88 – 110.
630. E. O. Heady, J. L. Dillon: *Agricultural Production Functions*, Ames, Iowa 1961.
631. F. Kuhlmann: *Betriebslehre der Agrar- und Ernährungswirtschaft*, 2nd ed., Frankfurt/M. 2003.
632. *World Fertilizer Consumption Statistics April 1998* IFA International Fertilizer Industry Association, 1998. FAOSTAT Database results Internet http://www.FAO.org
633. FAO: *The Sixth World Food Survey*, Food and Agricultural Organisation of the United Nations, Rome 1996.
634. P. Prinstrup-Anderson *et al.*: *The World Food Situation: Recent Developments*, Emerging Issues and Long Term Prospects, IFPRI, Washington 1997.
635. FAO: *Dimension of Need, An Atlas of Food and Agriculture*, Food and Agricultural Organisation of the United Nations, Rome 1995.
636. G. Harris: *"An Analysis of Global Fertilizer Application Rates for Major Crops*," International Fertilizer Development Centre (IFDC), Muscle Shoals 1998.
637. FAO: *Fertilizer Yearbook*, vol. **46**, Food and Agricultural Organisation of the United Nations, Rome 1996.
638. IFA: *"Fertilizer Demand Forecast*," International Fertilizer Industry Association, Paris 1998.
639. FAO: *Fertilizer Outlook*, Food and Agricultural Organisation of the United Nations, Rome 1996
640. J. Debruck, G. Fischbeck, W. Kampe: *Getreidebau aktuell*, 7th ed., Frankfurt/M. 1983.
641. F. Kuhlmann: "Zum 50. Todestag von Friedrich Aereboe: Einige Gedanken zu seiner Intensitätslehre," *Agrarwirtschaft* **41** (1992) 222 – 230.
642. F. Kuhlmann, K. F. Frick: "Produktionstheoretische Betrachtungen über das "Gesetz des Minimums", und das "Gesetz vom abnehmenden Ertragszuwachs"," *Berichte über Landwirtschaft* **73** (1995) 591 – 623.
643. K. F. Frick: "Analyse und Entwicklung von Standardproduktionsfunktionstypen in der pflanzlichen Produktion," *Sonderheft der Agrarwirtschaft*, vol. **153**, Holm 1997.

[Top of Page]

Related Articles

Individual fertilizers are also described under the separate keywords ⟶ Ammonia, ⟶ Ammonium Compounds, ⟶ Nitrates and Nitrites, ⟶ Phosphate Fertilizers, ⟶ Potassium Compounds, and ⟶ Urea

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Fiber Optics

Standard Article

Tsuneo Nakahara[1], Shigeru Tanaka[2], Masayuki Nishimura[2]
[1]Sumitomo Electric Industries, Ltd., Osaka, Japan
[2]Sumitomo Electric Industries, Ltd., Yokohama, Japan

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a10_433.pub2
Article Online Posting Date: December 15, 2006

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Structure of Optical Fibers
2.1 Basic Structure
2.2 Multimode Fibers
2.3 Single-mode Fibers
2.4 Dispersion-modified Fibers
2.4.1 Dispersion-shifted Fibers
2.4.2 Nonzero Dispersion-shifted Fiber
2.4.3 Dispersion-flattened Fibers
2.5 Functional Fibers
2.5.1 Dispersion-compensating Fibers
2.5.2 Rare-earth-doped Fibers
2.5.3 Polarization-maintaining Fibers
3. Raw Materials and Production
3.1 Silica Fibers
3.1.1 Preform Production
3.1.1.1 Modified Chemical Vapor Deposition (MCVD) Method
3.1.1.2 Outside Vapor Deposition (OVD) Method
3.1.1.3 Vapor-Phase Axial Deposition (VAD) Method
3.1.1.4 Plasma Chemical Vapor Deposition (PCVD) Method
3.1.2 Fiber Drawing
3.2. Non-silica Fibers
3.2.1 Multicomponent Glass Fibers
3.2.2 Plastic Cladding Fibers
3.2.3 All-plastic Fibers
3.2.4 Infrared Fibers
4. Characterisitics of Optical Fibers
4.1 Transmission Loss
4.2 Dispersion
4.3 Multimode Dispersion
4.3.1 Chromatic Dispersion
4.3.2 Polarization Mode Dispersion
4.4 Nonlinearity
5. Reliability of Optical Fibers
5.1 Mechanical Strength
5.2 Hydrogen-induced Loss Increase
5.3 Radiation-induced Loss Increase
6. Applications of Optical Fibers

6.1 Optical Fiber Cables
6.2 Fiber Connection Technologies
6.3 Optical Fiber Sensors
7. Future Prospects
7.1 New Fibers
7.1.1 Photonic Crystal Fiber (PCF)
7.1.2 Holey Fiber (HF)
7.1.3 Photonic Band-gap Fiber (PBGF)
7.2 New Systems
7.2.1 Soliton System
7.2.2 Code Division Multiplex (CDM)

The optical fiber is one of the most important key technologies that support the current information technology (IT) era. Optical fibers are now widely used not only as an excellent transmission medium for optical communication but also as optical functional devices, such as optical amplifiers. Both optical fibers and their manufacturing technologies have evolved dramatically toward higher performance and improved productivity. Their applications have also been expanded from optical communication to a variety of fields, such as optical sensors. This article reviews various aspects of optical fibers: structures & materials, fabrication techniques, optical characteristics & reliability, and applications.

[Top of Page]

1. Introduction

The oldest scientific record concerning light transmission in a confined path is attributed to the British physicist J. TYNDALL who, in 1870, studied the transmission of light in a curved stream of water. In 1930, the German physicist H. LAMM conducted experiments on image transmission through a crude assembly of quartz fibers. In 1936, G. C. SOUTHWORTH, in the United States, carried out theoretical and experimental studies on the behavior of electromagnetic waves in cylindrical nonconductive materials; in the late 1950s, T. NAKAHARA made a theoretical and experimental study on a loosely bounded electromagnetic wave along variously shaped dielectric lines using quasi-millimeter waves [1]. The phenomena they observed are exactly the same as those encountered in the transmission of laser light through optical fibers.

In the 1930s, many physicists studied the transmissivity of materials (e.g., F. URBACH, K. A. STACEY, and B. SZIGETY). They observed peaks in light absorption in the infrared region, which suggested the existence of a band with low absorption and thus a low transmission loss between the ultraviolet and infrared regions.

An optical communication method using glass as a transmission medium had been proposed in Japan in 1936 [2]. In the early 1960s, optical waveguides using lenses or lenslike media for optical signal transmission were studied by Bell Laboratories in the United States. KAO and HOCKHAM [3] of Standard Telephone Laboratories (STL) in the United Kingdom demonstrated in 1966 that removal of impurities from glass fibers reduced the loss (attenuation) of optical transmission to a level acceptable for use in long-distance communications; they recommended quartz as a suitable transmission material. In 1968, KAWAKAMI and NISHIZAWA [4] proposed that an optical fiber with a refractive index distribution similar to that of a lens could be used effectively as a high-capacity transmission medium.

Methods for fabricating transparent, pure fused silica glass have been in use since 1910. In 1935, a method of producing synthetic silica glass was developed, based on the decomposition of volatile silicon compounds such as silicon tetrachloride [10026-04-7]. Furthermore, by the early 1960s, the discovery had been made that the refractive index of pure fused silica glass could be increased easily with various additives; the structure, materials, and characteristics of a suitable optical fiber for signal transmission had also been described on a theoretical basis.

In 1969, Nippon Sheet Glass and the Nippon Electric Corporation (NEC) jointly produced a graded-index multicomponent glass fiber, which they called Selfoc, using an ion-exchange process [5]. In the first trial production, the light attenuation of the fiber was 80 to 200 dB km^{-1}. In 1970, Corning Glass Works in the United States announced a silica fiber with an attenuation of 20 dB km^{-1} [6], which was manufactured by depositing a particulate TiO_2–SiO_2 glass layer inside a silica tube by flame hydrolysis. In 1974, Bell Laboratories reported the production of a low-loss fiber with an attenuation of 1.1 to 2 dB km^{-1} by modified chemical vapor deposition (MCVD) [7]. The method of outside vapor deposition (OVD) was announced by Corning Glass Works in 1973 [8]. Sumitomo Electric filed a continuous vapor-phase axial deposition process patent that covered an arbitrary refractive index profile in 1974. Philips reported the plasma chemical vapor deposition (PCVD) method in 1975 [9]. The Nippon Telegraph & Telephone Corporation (NTT) announced the development of the vapor-phase axial deposition (VAD) process for step-index profile in 1977 [10]. The VAD process was rapidly improved, leading in 1981 to the successful manufacture of long, uniform optical fibers whose light attenuation was 0.2 dB km^{-1}.

The theoretical transmission loss of a silica fiber at a wavelength of 1.55 µm is about 0.13 dB km^{-1} and is limited by Rayleigh scattering and the vibrational absorption of the Si—O bonds. After the development of fluorine-stuffed pure silica fiber announced in 1972, the transmission loss was decreased. Champion data of transmission loss 0.154 dB km^{-1} at 1.55-µm wavelength and 0.1484 dB km^{-1} have been recorded in 1986 and in 2002, respectively [11].

As for mechanical strength, ordinary manufacturing equipment could not be used owing to the extreme brittleness of silica fiber. Sufficient mechanical strength for practical use could not be attained by applying polyamide (nylon) extrusion to drawn

fiber. However, the development in 1974 of the method of tandem resin coating immediately after drawing dramatically improved the mechanical strength [12].

[Top of Page]

2. Structure of Optical Fibers

2.1. Basic Structure

Optical fibers consist of a central core and a cladding layer that surrounds the core. The core has a slightly higher refractive index than that of the cladding. The transmission of light in the optical fiber is illustrated in Figure 1.

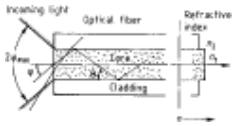

Figure 1. Fundamental structure of an optical fiber

n_1 = refractive index of core; n_2 = refractive index of cladding; θ = angle of reflection at core–cladding interface; φ = launching angle

When light is launched into the core at the endface of the optical fiber, the portion of the light whose angle is within a certain range will be trapped inside the core by total internal reflection at the core–cladding interface. The critical angle for the total internal reflection depends on the refractive index difference between the core and the cladding. It can be derived from Snell's law. As a parameter determining the light-capturing efficiency, the numerical aperture, *NA*, is defined by the following equation:

$$NA = \sin \varphi_{max} = n_1 \sin \theta_c = (n_1^2 - n_2^2)^{1/2} \quad (1)$$

where n_1 and n_2 are the refractive indexes of the core and the cladding, respectively, φ_{max} is the maximum launching angle of the light, and θ_c is the critical angle for total internal reflection.

Optical fibers are generally classified into two types: multimode fibers and single-mode fibers (SMFs).

2.2. Multimode Fibers

Multimode fibers have a relatively large refractive index difference (typically 1 to 2 %) between the core and the cladding and a large core diameter, which is significantly larger than the wavelength of light and is typically several tens of micrometers or larger. Because of these structural parameters, the multimode fibers have many light paths (or modes) traveling through the core with slightly different angles.

The multimode fibers are further classified into two types depending on the refractive index profiles of the core: step-index fibers and graded-index fibers. Figure 2 illustrates those two types of multimode fibers. The core of the step-index fiber has a constant refractive index higher than that of the cladding, while the graded-index fiber has the refractive index gradually changing as a function of the radial position.

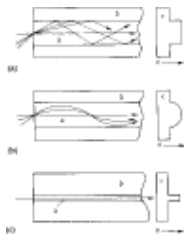

Figure 2. Classification of optical fibers according to propagation modes

(a) step-index fiber; (b) graded-index fiber; and (c) single-mode fiber. a) Core; b) Cladding; c) Refractive index, *n*

2.3. Single-mode Fibers

When both the core–cladding refractive index difference and the core diameter are reduced, the number of the transmitting modes (light paths) is decreased. Only one mode can be transmitted through the core, when these two parameters are sufficiently small. This is the case of the SMF as illustrated in Figure 2.

The basic type of silica-based SMF has a typical refractive index difference of 0.3 to 0.4 % and a core diameter of several µm. This standard type SMF is currently being used very widely in telecommunications. In optical communication systems using the standard SMFs, the optical signal wavelengths mostly used are 1.3 µm, at which the fibers have small or zero chromatic dispersion, and 1.55 µm, at which the fibers exhibit the lowest transmission loss, as described in detail in Section Characterisitics of Optical Fibers.

2.4. Dispersion-modified Fibers

Although the standard SMFs are superior transmission media for optical communication, even higher performance optical fibers of various types have been developed within the SMF category. The key parameter here is the chromatic dispersion. Since all silica-based fibers exhibit the lowest transmission loss around the 1.55-µm wavelength, this wavelength band is naturally suited for long-distance transmission. The chromatic dispersion of the standard SMF is, however, relatively large in the 1.55-µm band. This may limit the transmission capacity or the transmission distance. Various types of SMFs whose chromatic dispersion, especially in the 1.55-µm band, is modified or tailored to improve the transmission performance have been developed. These dispersion-modified fibers usually have somewhat complicated refractive index profiles as illustrated in Figure 3.

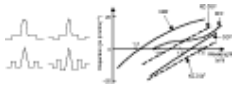

Figure 3. Refractive index profiles of dispersion-modified fibers and chromatic dispersion characteristics of various fibers

2.4.1. *Dispersion-shifted Fibers*

The dispersion-modified fibers of the first generation are dispersion-shifted fibers (DSFs) [13, 14]. The DSFs have their zero dispersion wavelengths (the wavelengths at which the chromatic dispersion is zero) shifted from the 1.3-µm band to the 1.55-µm band, where the transmission loss is minimized. The DSFs are well suited to long-distance transmission of high bit-rate optical signals.

2.4.2. *Nonzero Dispersion-shifted Fiber*

As the optical transmission technologies have recently been evolved with the advent of the optical fiber amplifier and wavelength-division-multiplexed (WDM) transmission, further modification of the chromatic dispersion characteristics has been applied to the DSFs. In long-distance WDM transmission with narrow wavelength spacing using optical amplifiers, the signal quality tends to deteriorate because of nonlinear effects in the optical fiber. Such detrimental effects become significant especially when the chromatic dispersion of the optical fiber is close to zero in the signal wavelength band.

The modified versions of DSFs whose chromatic dispersion is small but not zero in the 1.55-µm band are called nonzero dispersion-shifted fibers (NZ-DSFs) [15, 16]. The NZ-DSFs are being widely used in long-distance and large-capacity WDM transmission systems.

2.4.3. *Dispersion-flattened Fibers*

In general, the chromatic dispersion of the optical fiber has some dependence upon the wavelength. The wavelength dependence of the dispersion (the differential of the dispersion with respect to the wavelength) is called *dispersion slope*. Since both the DSFs and the NZ-DSFs have some dispersion slopes, the chromatic dispersion varies to some extent within the signal wavelength range. This dispersion variation is increased and may become undesirable when the signal wavelength band to be used is extended.

The SMFs whose dispersion slope is nearly zero in the wavelength range of concern are called dispersion-flattened fibers (DFFs) [17, 18]. Because precise control of such dispersion characteristics is generally difficult in view of the fiber design and fabrication, DFFs are still in the developmental stage and not yet fully commercialized.

2.5. Functional Fibers

Although optical fibers were originally developed as transmission media for optical signals, they have also found a number of applications in which their optical functionality, other than optical transmission, is utilized. Among these functional fibers are dispersion-compensating fibers (DCF), rare-earth-doped fibers, and polarization-maintaining fibers (PMFs).

2.5.1. *Dispersion-compensating Fibers*

When high bit-rate optical signals in the 1.55-µm band are transmitted through long standard SMFs, the signal waveforms are deformed because of the relatively large positive chromatic dispersion. This signal deterioration can be suppressed by inserting some SMFs that have large negative chromatic dispersions, which compensate for the positive dispersions of the transmission fibers [19]. These DCFs are typically coiled and packaged in small boxes. Such DCF modules are placed in the transmission equipment and used as optical functional devices.

2.5.2. *Rare-earth-doped Fibers*

Some of rare earth elements are known to be luminescent. Optical fibers whose cores are doped with specific rare earth elements function as optical amplifiers. One of the most important rare earth elements is erbium, which is capable of amplifying optical signals around the wavelength of 1.55 µm [20, 21]. The 1.55-µm band is one of the major wavelength bands used in optical fiber communication, because silica fibers exhibit the lowest transmission loss. Praseodymium and thulium are also known as additives for optical amplification in the 1.3-µm and 1.45-µm bands, respectively [21, 22].

Figure 4 illustrates the basic configuration of the erbium-doped fiber amplifier (EDFA). The pumping light at 0.98 µm or 1.48 µm launched into the fiber excites the erbium ions in the core, and the 1.55-µm optical signals traveling through the fiber are amplified through stimulated emission from the erbium ions. The EDFAs are not only highly efficient in terms of the signal gain or the output power but are also capable of simultaneously amplifying multiple optical signals that are WDM in the 1.55-µm band. This feature has strongly supported the recent dramatic progress of WDM transmission.

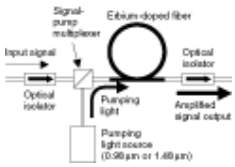

Figure 4. Basic configuration of an erbium-doped fiber amplifier

2.5.3. *Polarization-maintaining Fibers*

Single-mode optical fibers that maintain the linear polarization state of the light transmitted through the fibers are called polarization-maintaining fibers (PMFs). A typical cross section of the PMF is illustrated in Figure 5 [23]. The polarization of light is maintained by the birefringence that is induced by two stress-applying parts composed of SiO_2–B_2O_3 placed beside

the core. The PMFs are often used in optical fiber sensors, such as fiber-optic gyroscopes, or for interconnection between optical devices that have polarization dependence, such as laser diodes, optical modulators, and so on.

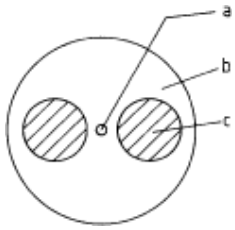

Figure 5. An example of a polarization-maintaining fiber

a) Core; b) Cladding; c) Stress-applying part

[Top of Page]

3. Raw Materials and Production

Optical fibers are basically made of optically transparent materials. Most of the optical fibers currently used in optical communication are silica-based fibers because of their low transmission loss. There are also several materials other than silica that are used for optical fibers, such as multicomponent glass, plastic, fluoride glass and so on.

3.1. Silica Fibers

Silica fibers are composed of SiO_2 and some additives that modify the refractive index of silica. The additives are essential for forming the basic fiber structure consisting of the core and the cladding, which have slightly different refractive indexes. The refractive index of silica can be decreased by adding fluorine or B_2O_3 and increased by adding other metal oxides or anions. The most popular additive or dopant for increasing the refractive index is GeO_2. P_2O_3 and Al_2O_3 are also known to increase the refractive index. Figure 6 shows the effects of various materials on the refractive index of silica.

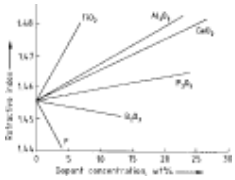

Figure 6. Relationship between dopant concentration and refractive index of silica

3.1.1. *Preform Production*

In the production of silica fibers, silica-based glass rods, called preforms, are prepared first, and then the preforms are drawn into thin fibers. A vapor-phase method is generally used to form silica fiber preforms. There are four major preform fabrication techniques: modified chemical vapor deposition (MCVD), outside vapor deposition (OVD), vapor-phase axial deposition (VAD), and plasma chemical vapor deposition (PCVD).

In preform production by the MCVD and the PCVD methods, the following oxidation reaction is utilized to synthesize SiO_2:

$$SiCl_4 + O_2 \rightarrow SiO_2 + 2\,Cl_2$$

The following hydrolysis reaction is employed in the OVD and the VAD methods:

$$SiCl_4 + 2\,H_2O \rightarrow SiO_2 + 4\,HCl$$

The typical raw materials are volatile liquids (e.g., $SiCl_4$, $GeCl_4$, BBr_3, $POCl_3$) and gases (e.g., BCl_3, CCl_2F_2, SiF_4). Since the vapor pressures of transition-metal halides are much lower than those of the raw materials, the fibers fabricated using these techniques are very pure in terms of transition-metal impurities.

3.1.1.1. *Modified Chemical Vapor Deposition (MCVD) Method*

The MCVD technique, which was developed by Bell Laboratories [24], achieves glass deposition by letting the raw material vapors and oxygen flow through a rotating silica tube, as illustrated in Figure 7. The tube is heated by using a traversing O_2–H_2 burner. To obtain the desired refractive index profile, a number of glass layers of different compositions are deposited on the inner wall of the tube. In some cases, P_2O_5 is added to decrease the deposition temperature. After the deposition, the temperature is raised to 1800 to 2000 °C to collapse the tube into a solid preform.

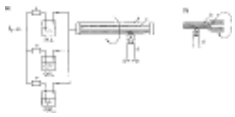

Figure 7. Modified chemical vapor deposition (MCVD) method

(a) Deposition and (b) collapsing. a) Mass flow controller; b) Rotating silica tube; c) Synthetic glass layer; d) Traversing burner

3.1.1.2. *Outside Vapor Deposition (OVD) Method*

The OVD technique was developed by Corning Glass Works [8]. Figure 8 illustrates the principle of the OVD technique. Fine

glass particles (soot) are produced by flame hydrolysis of the raw materials with an O_2–H_2 burner and deposited on a rotating rod. A number of soot layers are deposited by moving the burner many times along the rotating rod. The vapor composition is changed layer by layer to produce a desired refractive index profile. After the deposition, the original rod is removed from the soot rod, which is then dehydrated by heating in chlorine gas and sintered to form a solid preform.

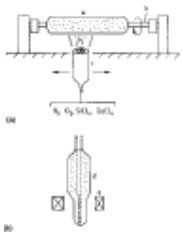

Figure 8. Outside vapor deposition (OVD) method

(a) Soot deposition and (b) consolidation. a) deposited soot; b) rotating rod; c) traversing burner; d) soot rod; e) heater

3.1.1.3. *Vapor-Phase Axial Deposition (VAD) Method*

The VAD technique was developed by several Japanese companies [25]. Soot generated by flame hydrolysis is deposited on the end of a rotating seed rod. A soot preform then grows in the axial direction, as illustrated in Figure 9. In the case of graded-index fiber preform production, the refractive index profile is determined by controlling the temperature distribution on the growing soot surface. Multiple O_2–H_2 burners may also be used to simultaneously deposit both the core and the cladding for step-index fiber preforms. The soot preform is then dehydrated by heating in chlorine gas, and finally sintered to form a solid preform.

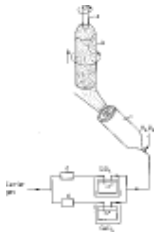

Figure 9. Vapor-phase axial deposition (VAD) method

a) rotating seed rod; b) soot preform; c) O_2–H_2 burner; d) mass flow controller

Since the soot grows continuously in the axial direction in the VAD method, there is no fundamental limit to the soot length that can be fabricated. The VAD technique enables fabrication of very large preforms with high productivity.

3.1.1.4. *Plasma Chemical Vapor Deposition (PCVD) Method*

The PCVD technique was developed by Philips [26]. Its concept is illustrated in Figure 10. A silica tube is evacuated to reduce its internal pressure and raw materials and oxygen are fed in. Oxidation occurs in nonisothermal plasma produced by a microwave cavity. The tube is typically placed in a furnace whose temperature is raised to 1200 °C, for example. After the deposition, the tube is collapsed at a high temperature.

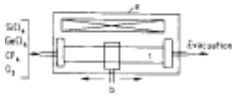

Figure 10. Plasma chemical vapor deposition (PCVD) method

a) Furnace; b) Traversing microwave cavity; c) Silica tube

3.1.2. ***Fiber Drawing***

Silica fibers are fabricated by heating the end of a preform in a furnace and drawing it, as shown in Figure 11. The drawing temperature is typically 2100 °C. There are two types of drawing furnaces: a carbon resistance furnace and a radio-frequency induction furnace. The fiber diameter (typically 125 µm) is precisely controlled during the drawing process by monitoring the diameter and adjusting the drawing speed.

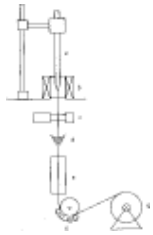

Figure 11. Fiber drawing process

a) Preform; b) Furnace; c) Fiber diameter monitor; d) Coating applicator; e) Curing furnace; f) Capstan; g) Take-up bobbin

In order to maintain the mechanical strength of silica fibers, it is essential to immediately coat the fiber with some sort of resin before it comes in contact with rollers or take-up bobbins in the drawing process. This prevents the generation of microscopic cracks on the fiber surface. Figure 12 shows a comparison of the tensile strength between uncoated fibers and coated fibers. Typical coating materials are urethane acrylate and silicone. The former is cured by ultraviolet light, and the latter is usually cured by heat. A standard coating diameter is 250 µm.

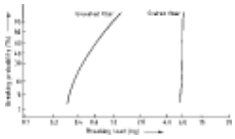

Figure 12. Tensile strength of coated and noncoated fibers

3.2. Non-silica Fibers

3.2.1. *Multicomponent Glass Fibers*

Multicomponent glass fibers are fabricated from ultrapure powders such as SiO_2, Na_2CO_3, $CaCO_3$, and KNO_3. These powders are mixed and melted in a crucible to form homogeneous, bubble-free glass. The core and cladding glasses are melted separately, and then processed by the so-called double-crucible method, which employs two crucibles that have concentric orifices at the bottom. The core glass is placed in the center crucible and the cladding glass in the outer crucible. The crucibles are heated and the core and cladding glasses are drawn into a fiber having the core–cladding structure.

Although the multicomponent glass fibers are relatively easy to fabricate because of their low melting temperatures, their transmission losses are higher than silica fibers. The transmission loss is typically 5 dB km^{-1} at 0.85 µm [27].

3.2.2. *Plastic Cladding Fibers*

While the standard silica fibers have the core and the cladding both of which are composed of silica, the cladding can be replaced by some kind of polymer whose refractive index is lower than that of the silica core. This type of optical fiber is called plastic cladding fiber.

The plastic cladding fibers can be fabricated by using techniques very similar to those developed for silica fibers. The fiber is drawn from a silica rod and coated immediately with a low refractive index resin, such as silicone, as the cladding. The core can be either pure silica with a constant refractive index (the step-index fiber) or germanium-doped silica with a graded-index profile (the graded-index fiber).

3.2.3. *All-plastic Fibers*

All-plastic fibers consist of a plastic core and a plastic cladding. Polystyrene or polymethyl methacrylate is used as the core material and fluoropolymer as the cladding material, for example. The all-plastic fibers feature excellent flexibility and large dimensions (large core and cladding diameters), which enable easy connection and handling. Their transmission losses are, however, relatively high at wavelengths exceeding 0.6 µm because of C–H vibration absorption. Lower transmission losses at 1.3 µm have recently been achieved by using fluoropolymer as the core material [28]. Recent advances in fabrication techniques have also enabled the formation of graded-index profiles [29]. Graded-index all-plastic fibers featuring large bandwidths are believed to be useful for short-distance large-capacity transmission in buildings.

3.2.4. *Infrared Fibers*

As described in detail in Chapter Characterisitics of Optical Fibers, silica fibers have the lowest transmission loss around 1.55 µm, because of Rayleigh scattering, which is proportional to λ^{-4} and therefore decreases as the wavelength is increased, and the infrared vibration absorption by the Si —O bond in silica. It is anticipated that alternative materials having vibration absorption in much longer wavelength bands may exhibit a transmission loss lower than that of silica at wavelengths longer than 1.55 µm. Such non-silica materials that have been investigated are germanium dioxide, fluoride glasses, chalcogenide glasses, and crystalline halides [30, 31].

Figure 13 shows examples of the transmission loss spectra of those infrared fibers. It is clear that the infrared fibers realize transmission losses lower than those of silica fibers in wavelengths longer than 2 µm. Attempts to achieve ultralow loss (lower than the silica fiber's minimum loss around 1.55 µm) in the infrared region, however, have never been successful. The infrared fibers are used in short-distance applications with signal wavelengths longer than 2 µm, such as optical sensors.

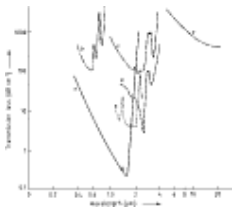

Figure 13. Typical loss spectra of infrared fibers compared with those of silica and plastic fibers

a) Silica fiber; b) Plastic fiber; c) Chalcogenide glass fiber; d) Fluoride glass fiber; e) GeO_2 fiber; f) Polycrystalline (AgBr) fiber

[Top of Page]

4. Characterisitics of Optical Fibers

4.1. Transmission Loss

The transmission loss of the optical fiber represents a decrease in optical power per unit length. It is directly related to the maximum distance over which optical signals can be transmitted in optical communication systems, and is one of the most important parameters of the optical fiber. The transmission loss of the optical fiber is generally given by the following equation:

$$\alpha(\lambda) = \frac{A}{\lambda^4} + B + C(\lambda) \quad (2)$$

where

λ = wavelength of light (µm)

A = Rayleigh scattering coefficient (dB km^{-1} µm^{-4})

B = loss induced by structural imperfection (dB km^{-1})

$C(\lambda)$ = wavelength-dependent loss induced by impurities and intrinsic absorption (dB km^{-1})

A typical transmission loss spectrum of silica fiber is shown in Figure 14. The first term in Eq. (2) is Rayleigh scattering, which is caused by microscopic fluctuation of the density and/or the composition in glass, and it is one of the fundamental losses of glass. The Rayleigh scattering loss is proportional to λ^{-4}. The Rayleigh scattering coefficient, A, is approximately 0.7 dB km^{-1} µm^{-4} for pure silica, and it increases as the concentration of index-changing additives, such as GeO_2, is increased.

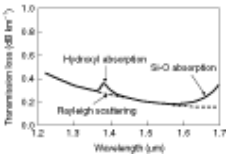

Figure 14. Typical transmission loss spectrum of silica fiber

The second term, the structural-imperfection-induced loss, is related to longitudinal fluctuation of the core – cladding interface, for example, and it is generally independent of the wavelength. Improvement of the fabrication techniques has reduced the structural-imperfection-induced loss to negligibly small levels.

One of the major contributions included in the third term of Eq. (2) is the intrinsic absorption induced by vibration of the Si —O bond in silica. A large absorption having a peak at 9.1 µm affects the transmission loss at wavelengths longer than 1.55 µm. Other important causes of absorption are hydroxyl bonds and transition-metal impurities, such as iron, copper, cobalt, or nickel. In general, the concentration of the transition metal is so small that it does not affect the transmission loss of silica fibers manufactured properly. On the other hand, the hydroxyl bonds exhibit a peaky absorption loss spectrum around 1.39 µm. However, it is also possible to reduce the hydroxyl content to sufficiently small levels by using improved dehydration techniques in preform production [32]. Optical fibers with no hydroxyl loss around 1.39 µm are now commercially available.

In the end, the fundamental loss causes in silica fibers are only Rayleigh scattering and the Si —O bond absorption. Since the Rayleigh scattering coefficient depends on the dopant concentration in the core, the lowest transmission loss can be achieved by a pure-silica-core fiber, which consists of a pure silica core with no additives and a cladding containing fluorine that lowers the refractive index. Sumitomo Electric has reported the lowest transmission loss record of 0.1484 dB km^{-1} [11], which is significantly lower than the standard germanium-doped-core fibers' loss of 0.2 dB km^{-1}.

4.2. Dispersion

The dispersion represents the variation in the group velocity of light transmitted through the optical fiber. In general, the dispersion determines the deformation of the optical signal waveforms, thus the transmission bandwidth of the optical fiber. The dispersion is classified into three categories: multimode dispersion, chromatic dispersion, and polarization mode dispersion.

4.2.1. *Multimode Dispersion*

The multimode dispersion is the variation in the group velocity among many different transmission modes, or many light rays traveling with slightly different angles through multimode fibers. The multimode dispersion is applicable only to multimode fibers. The concept of the multimode dispersion is shown in Figure 2.

The refractive index profiles of the graded-index fibers are designed so that almost all light paths have the same group velocity. Note that the apparently shortest light path traveling straight along the core axis does not necessarily have the highest velocity, since the velocity of light depends on the refractive index of the material (the higher the refractive index, the lower the velocity). The optimum refractive index profiles of the graded-index fibers, which minimize the group velocity variation among all transmission modes (light paths), are known to be nearly parabolic. The bandwidths of the step-index fibers are significantly smaller than those of the graded-index fibers because of the multimode dispersion.

4.2.2. *Chromatic Dispersion*

There is no multimode dispersion in SMFs, since they have only one transmission mode. Therefore, SMFs, in general, exhibit much larger transmission bandwidths than multimode fibers. The dispersion characteristics of SMFs are determined by the chromatic dispersion and the polarization mode dispersion.

Figure 15 schematically shows the chromatic dispersion. The chromatic dispersion is the differential of the group delay of light transmitted through the optical fiber with respect to the wavelength. The chromatic dispersion can be divided into two effects: material dispersion and waveguide dispersion. The material dispersion is the characteristic intrinsic to the material, silica in this case, and it varies depending upon the additives, such as germanium, but not very much. On the other hand, the waveguide dispersion is the dispersion induced by the fiber structure, and it strongly depends on the refractive index profile. By properly designing the refractive index profile, it is possible to some degree to tailor the waveguide dispersion, and thus the chromatic dispersion. This is the base of the various dispersion-modified fibers described in Chapter Structure of Optical

Fibers.

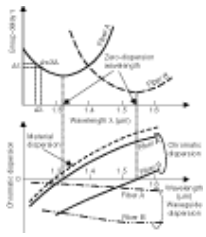

Figure 15. Group delay and chromatic dispersion characteristics of optical fibers

4.2.3. *Polarization Mode Dispersion*

Actually, SMF has two transmission modes, both of which are basically the same but whose polarization axes are perpendicular to each other. The polarization mode dispersion (PMD) is the group delay difference between those two orthogonal polarization modes. If the optical fiber is perfectly circular, the polarization modes have the same group velocity, but slight noncircularity of the fiber structure and/or asymmetric stress distribution applied to the fiber induce some birefringence, thus the PMD. Because optical power transfer between the two modes randomly occurs during transmission through the fiber, behavior of the PMD has a somewhat statistical nature. The PMD is generally proportional to the square root of the fiber length.

4.3. Nonlinearity

In general, phenomena whereby the material properties of a given medium change when high-intensity light is transmitted through it are called *nonlinear effects* (see ⟶ Nonlinear Optics and [33]). The nonlinearity of silica is significantly small compared with various other glasses. It has, however, become important recently because optical fiber amplifiers increase the optical signal power and the number of multiplexed optical signals in the optical fiber, resulting in noticeable levels of nonlinear effects even in silica fibers.

The nonlinear effects that are of most concern for optical fibers are nonlinear refractive index (optical Kerr effect) and stimulated scattering. The nonlinear refractive index is a small variation of the refractive index dependent upon the light intensity. This may cause various detrimental effects in the optical fiber communication, such as self-phase modulation, cross-phase modulation, and four-wave mixing. Among the stimulated scattering phenomena are stimulated Raman scattering and stimulated Brillouin scattering. While nonlinear effects are generally harmful in optical fiber communication, they are also useful in a variety of applications, such as fiber Raman amplifiers.

[Top of Page]

5. Reliability of Optical Fibers

5.1. Mechanical Strength

The mechanical lifetime of the optical fiber depends on the stress applied to the fiber and the environmental conditions [34]. Figure 16 shows the relationship between the average time to fiber breakage and the applied stress in air and water. The time to fiber breakage or the time to failure, t_f, is given by

$$\ln t_f = -n \ln \sigma_a + \ln k_s \tag{3}$$

where σ_a is the applied stress, and n and k_s are constants that depend on the glass material and the environmental conditions. For silica fibers in air, n is approximately 25 and k_s about 2.2×10^{12}. In production of the optical fiber, a certain level of stress is applied to the fiber after drawing in order to eliminate any weak portion; this is called *proof testing*. The expected lifetime of the fiber, t_{min}, is given by

$$\ln t_{min} = n \ln \sigma_a + \ln \left(B\sigma_p^{n-2}\right) \tag{4}$$

where σ_p is the stress applied during the proof testing and σ_a is the stress applied to the fiber in use, for example, the fiber in a cable. The constant B depends on the glass material and the environmental conditions. In general, high humidity and high temperature shorten the fiber lifetime. However, proper proof testing in the fiber production and appropriate cable designs ensure long-term reliability in standard applications.

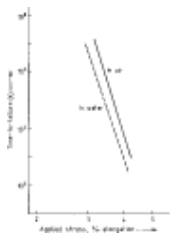

Figure 16. Relationship between time to failure and applied stress for silica fibers in air and water

5.2. Hydrogen-induced Loss Increase

Hydrogen may be produced in optical fiber cables as a result of thermal dissociation of plastic, corrosion of metal by water, or

electrolysis of water when two different metals are used in a cable. Two types of loss increase due to hydrogen are known: absorption by hydrogen gas permeated into the glass and absorption by hydroxyl bonds formed as a result of reactions between the glass and the hydrogen gas [35, 36]. The former loss increase is reversible, while the latter is not. Figure 17 shows the loss spectra of an optical fiber before and after 36-week exposure to hydrogen gas at room temperature. The permeated hydrogen causes absorption of 8 dB km^{-1} at 1.24 µm.

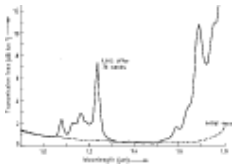

Figure 17. Transmission loss spectra of a silica fiber before and after 36-week exposure to hydrogen at room temperature

Figure 18 shows the loss spectra of two SMFs before and after exposure to hydrogen at 200 °C. At high temperatures, the hydroxyl absorption having a peak at 1.39 µm increases in the standard germanium-doped-core fiber, while the pure-silica-core fiber exhibits no loss increase. This is because GeO_2 reacts with hydrogen to form hydroxyl bonds but pure silica does not. The pure-silica-core fiber is superior in terms of the transmission loss stability against hydrogen.

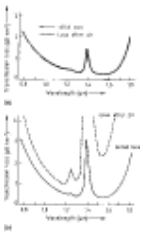

Figure 18. Stability of optical fibers after heating in hydrogen at 200 °C

(a) Single-mode fiber with a pure silica core; and (b) single-mode fiber with a GeO_2-doped silica core

5.3. Radiation-induced Loss Increase

The transmission loss is also increased by irradiation [37]. The loss increase depends upon the type of radiation, its intensity (dose rate), the composition of the fiber, the fiber fabrication process, and the fiber temperature during irradiation. The effects of ϒ-ray, X ray, ultraviolet light, or neutron irradiation have been examined.

Among various silica-based fibers, the pure-silica-core fiber exhibits the smallest loss increase after irradiation. Addition of GeO_2 or P_2O_5 to silica results in a high sensitivity to all forms of irradiation, while B_2O_3 specifically increases the sensitivity to neutron irradiation. The sensitivity is reduced by adding hydroxide ions or hydrogen and is increased by adding chlorine.

For a given total radiation dosage, the transmission loss increase is much larger for higher dose rates. In addition, irradiation at a higher temperature results in a smaller loss increase and faster recovery of the radiation-induced loss. Visible or infrared light irradiation also helps to suppress the loss increase and accelerate the loss recovery; this phenomenon is called *photobleaching* [38].

[Top of Page]

6. Applications of Optical Fibers

6.1. Optical Fiber Cables

Optical fibers are used most typically in the form of optical fiber cables. Optical fiber cables are generally thinner and lighter than conventional communication cables, such as copper wire cables or coaxial cables. Optical fiber cables are free from any influence from electromagnetic induction; this is another advantage over conventional copper cables. Optical fiber cables have a wide variety of applications, such as long haul or metropolitan telecommunication networks, submarine cable networks, local area networks, broadband access (fiber to the home) networks, and so on. There are also various composite type cables, which contain optical fibers for addition of the communication function. Examples of such composite cables are electric power/optical fiber composite cables and optical ground wires (OPGW).

Figure 19 illustrates the cross sections of three major optical fiber cable types used in telecommunications: the loose tube type cable, the slotted-core type cable, and the single tube type cable. Optical fiber cables are designed so that the optical fibers inside are free from excessive strain or bending during and after installation. In loose tube type cables, optical fibers are loosely encased in plastic tubes, and the tubes are stranded around a central strength member. In the case of slotted-core type cables, a number of individual fibers or fiber ribbons (fiber arrays formed with some resin) are placed in each slot or groove, which is helically formed on the surface of a central rod having a strength member. Single tube type cables contain individual fibers or fiber ribbons in the central tube. In any case, proper design of the cable structure ensures stable performance of optical fibers.

Figure 19. Examples of optical fiber cables

(a) Loose tube type cable; (b) Slotted-core type cable; and (c) Single tube type cable

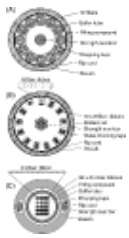

6.2. Fiber Connection Technologies

It is highly important to connect optical fibers in various situations. The loss of the optical signal at the connection should be as small as possible. Since the core of the optical fiber is very small, precise alignment of the ends of two optical fibers to be connected is key. There are three types of fiber connecting technologies: fusion splicing, optical connectors, and mechanical splicing.

Figure 20 illustrates the principle of fusion splicing. Fiber ends are melted by means of a high-frequency electric discharge and permanently joined. Prior to fusion splicing, perfectly flat fiber ends are prepared by utilizing fiber-cleaving tools. In the case of the optical connector, the ends of the optical fibers are inserted and fixed in high-precision ferrules, which are then mated by using an alignment sleeve, as illustrated in Figure 21. In fabrication of the optical connectors, the end faces must be polished flat. The optical connectors can be connected and disconnected repeatedly. Multifiber connectors, which simultaneously connect multiple fibers or fiber ribbons with the fiber count of 4, 8, or more, have also been developed and put to practical use.

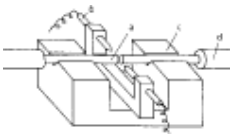

Figure 20. Principle of fusion splicing

a) Optical fiber; b) Electrode; c) V-groove; d) Fiber coating

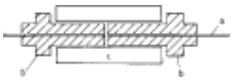

Figure 21. Principle of an optical fiber connector

a) Optical fiber; b) Ferrule; c) Alignment sleeve

Mechanical splicing is a technique for semipermanent fiber connection that utilizes relatively simple tools, unlike fusion-splicing machines. The optical fiber ends are prepared by using the fiber-cleaving tool and then connected by utilizing high-precision fiber aligning parts. In general, the connection loss is lowest for fusion splicing, and relatively high with optical connectors and mechanical splicing.

6.3. Optical Fiber Sensors

A variety of sensors that utilize optical fibers have been developed, and some of them have already been brought into practical use. Examples of such optical sensors are fiber-optic gyroscopes and fiber-optic temperature distribution sensors.

The fiber-optic gyroscopes utilize a phenomenon called the *Sagnac effect*, in which light traveling clockwise and light traveling counterclockwise in the same optical fiber coil experience slightly different phase shifts when the optical fiber coil is rotating. Since the phase difference depends on the angular velocity of the coil, the rotation can be detected by observing the phase difference. Ultrahigh sensitivity can be realized by properly designed fiber-optic gyroscopes.

Fiber-optic temperature distribution sensors utilize the Raman scattering effect, which is one of the nonlinear effects in optical fibers. Since the amount of light generated by the Raman scattering effects in the optical fiber depends upon the temperature of the fiber, it is possible to determine the fiber temperature by observing the Raman scattering–induced light signals. By further examining the signals as a function of the location along the fiber, the temperature distribution along the fiber can be observed. Fiber-optic temperature distribution sensors are actually being used, for example, for monitoring the temperature distribution along electric power cables (which contain optical fibers as sensors).

[Top of Page]

7. Future Prospects

In the ubiquitous communications era, the transmission and exchange of huge amounts of data will be indispensable. As described in the previous chapters, various technologies have been developed to increase data transmission rates, such as Time Division Multiplexing (TDM) and Wavelength Division Multiplexing (WDM).

7.1. New Fibers

At present, dispersion-controlled silica fibers are developed for long-distance communication with TDM and/or WDM systems to achieve maximum transmission speed. Also, several types of transmission along new types of optical fiber have been proposed. These fibers are briefly described below.

7.1.1. *Photonic Crystal Fiber (PCF)*

In 1995, BIRKS et al. reported on Photonic Crystal Fiber (PCF) and Holey Fiber (HF), a type of PCF, and thereafter these types of fibers have been actively investigated [39]. PCF is manufactured by stacking an array of silica capillary rods to form a macroscopic preform and then by drawing it into fiber. Typically, PCF is made from hexagonal lattice of circular air columns in which the periodicity is relatively uniform.

Photonic Band Gap (PBG) is defined as the range of frequency and propagation constant of light in which light is prohibited to propagate. It appears by arranging air holes in silica glass adequately.

7.1.2. *Holey Fiber (HF)*

HF has a periodic lattice of air holes distributed in the oblique section so as to form a core-clad microstructure that enables total reflection.

Owing to the large refractive index contrast that exists between glass and air, HF has unique properties, such as very high nonlinearity, a small effective core area, an enormous amount of waveguide dispersion that can be used for compensating the large material dispersion of silica, and so on [40].

7.1.3. *Photonic Band-gap Fiber (PBGF)*

CREGAN et al. demonstrated in 1999 that light can be strictly guided without leakage along a hollow core [41]. Light in certain well-defined wavelength bands is trapped in air by a full two-dimensional photonic band gap. PBGF is expected to greatly reduce Rayleigh scattering and nonlinearity. It also may hold the promise of yet undiscovered properties that are not attainable in standard fibers.

Recently, ultrashort pulse lasers such as femtosecond and picosecond pulse lasers have been developed. With ultrashort pulses, the wavelength spectrum can be very broadly spread. However, there are no fibers that can cover such a wide spectrum. Therefore, new types of fiber that use new materials or an air containing structure with a wider wavelength region will be required.

7.2. New Systems

As far as using ordinary optical fiber bit rate and distance of repeater spacing are inevitably limited owing to dispersion of optical fiber, that is, wavelength dependency of group velocity. In order to cope with these limitations, optical soliton transmission utilizing the nonlinear optics effect and Code Division Multiplex (CDM) have been proposed as new breakthrough technologies.

7.2.1. *Soliton System*

In 1973, HASEGAWA and TAPPERT proposed that soliton waves could be applied to optical fiber communications [42]. The pulse, a soliton wave, is generated when the group velocity dispersion of glass fiber and the tertiary nonlinear optical effect called self-phase modulation effect are balanced.

The self-phase modulation effect appears when the power level of optical pulse is increased and the nonlinear optical effect accordingly becomes larger. An equilibrium condition occurs and very long-distance repeater spacing is possible thanks to the absence of distortion.

This proposition has been verified through laboratory testing. However, no practical soliton system has yet been deployed, because power level control has not yet reached a sufficient level of sophistication.

7.2.2. *Code Division Multiplex (CDM)*

With very high-speed digital transmission, all-photonic switching will be required because electronic switching cannot function at such high speed. Therefore, new components including fiber for all-photonic switching, such as two-dimensional space modulation, will be required.

The extension of today's TDM and WDM technologies will have limitations because of nonlinearity, polarization, and the temperature dependency of fiber. Therefore, new modulation systems, such as CDM using a concept like spread spectrum [43], will be required in the future.

[Top of Page]

References

1. M. Sugi, T. Nakahara, *I.R.E. Trans. M.T.T.* **7** (1958) no. 3), 366.
2. JP 125 946, 1936 (M. Seki, H. Negishi).
3. K. C. Kao, G. A. Hockham, *Proc. IEE* **133** (1966) no. 7, 1151 – 1158.
4. S. Kawakami, H. Nishizawa, *IEEE Trans.* **MTT-16** (1968) no. 10, 814 – 818.
5. T. Uchida, M. Furukawa, I. Kitano, K. Koizumi, H. Matsumura, *IEEE J. Quantum Electron.* **QE-6** (1970) no. 10, 606 – 612.
6. F. P. Kapron, D. B. Keck, R. D. Maurer *Proceedings of Conference on Trunk Telecommunications by Guided Waves* (1970), London, Sept.–Oct., pp. 148–153.
7. J. B. MacChesney, P. B. O'Conner, F. V. DiMarcello, J. R. Simpson, P. D. Lazay, *Proceedings of 10th International Congress on Glass* (1974), pp. 6/40–6/45.
8. Corning Glass Works, US 3 737 292, 1973 (D. B. Keck, B. Flats, P. C. Schultz).
9. D. Küppers, J. Koenings, *Proceedings of ECOC '76* (1976) pp. 49–54.

10. T. Izawa, S. Kobayashi, S. Sudo, F. Hanawa, *Proceedings of IOOC '77*, Tokyo/Osaka (1977), pp. 375–378.
11. K. Nagayama, M. Kakui, M. Matsui, T. Saitoh, Y. Chigusa, *Electron. Lett.* **38** (2002) 1168.
12. T. Nakahara, A. Tsukamoto, H. Kumamaru, M. Hoshikawa, *Proc. IEEE* **123** (1976) no. 6, 603–608.
13. N. Kuwaki, M. Ohashi, C. Tanaka, N. Uesugi, *Electron. Lett.* **21** (1985) 1186.
14. D. T. Croft, J. E. Ritter, *Technical Digest of OFC'85*(1985) Paper WD2.
15. D. W. Peckham, A. F. Judy, R. B. Kummer, *Proceedings of ECOC '98* (1998) Vol. 1, pp. 139, 140.
16. T. Ooishi, Y. Kubo, Y. Yokoyama, T. Kato, M. Yoshida, M. Onishi, Y. Makio, *Proceedings of OECC 2000* (2000), Paper 12B3-4, pp. 136, 137.
17. Y. Liu, W. B. Mattingly, D. K. Smith, C. E. Lacy, J. A. Cline, E. M. DeLiso, *Proceedings of ECOC '98* (1998) Vol. 1, pp. 37, 38.
18. Y.Yokoyama, T. Kato, M. Hirano, M. Onishi, E. Sasaoka, Y. Makio, M. Nishimura, *Proceedings of ECOC '98* (1998) Vol. 1, pp. 131, 132.
19. J. M. Dugan, A. J. Price, M. Ramadan, D. L. Wofl, E. F. Murphy, A. J. Antos, D. K. Smith, D. W. Hall, *Technical Digest of OFC '92* (1992) Paper PD14, pp. 367–370.
20. E. Desurvire: *Erbium-Doped Fiber Amplifiers*, John Wiley & Sons, New York 1994.
21. S. Sudo: *Optical Fiber Amplifiers: Materials, Devices, and Applications*, Artech House, Norwood 1997.
22. Y. Ohishi, T. Kanamori, T. Kitagawa, S. Takahashi, E. Snitzer, G. H. Sigel, *Technical Digest of OFC '91* (1991) Paper PD2.
23. T. Hosaka, K. Okamoto, T. Miya, Y. Sasaki, T. Edahiro, *Electron. Lett.* **17** (1981) 530, 531.
24. J. B. MacChesney, P. B. O'Conner, F. V. DiMarcello, J. R. Simpson, P. D. Lazay, *Proceedings of 10th International Congress on Glass* (1974) pp. 6/40–6/45.
25. T. Izawa, S. Kobayashi, S. Sudo, F. Hanawa, *Proceedings of IOOC '77* (1977) pp. 375–378.
26. D. Küppers, J. Koenings *Proceedings of ECOC '76* (1976) pp. 49–54.
27. K. J. Beeles, P. W. France, S. F. Carter, S. Partington, *Proceedings of ECOC '82* (1982) pp. 601–605.
28. Y. Koike, *Proceedings of ECOC '96* (1996) pp. 41–48.
29. T. Ishigure, Y. Koike, *Technical Digest of OFC 2001* (2001) Paper ThC7.
30. T. Kanamori, S. Takahashi, *Jpn. J. Appl. Phys.* **24** (1985) L785–L760.
31. S. Takahashi, I. Sugimoto, *J. Lightwave Technol.* **LT-2** (1984) 272–275.
32. F. Hanawa, S. Sudo, M. Kawachi, M. Nakahara, *Electron. Lett.* **16** (1981) 669, 700.
33. G.P. Agrawal: *Nonlinear Fiber Optics*, Academic Press, New York 1989.
34. Y. Mitsunaga, Y. Katsuyama, H. Kobayashi, Y. Ishida, *J. Appl. Phys.* **53** (1982) 4847–4853. Links
35. K. Mochizuki, Y. Namihira, H. Yamamoto, *Electron. Lett.* **19** (1983) 743–745. Links
36. N. Uchida, N. Uesugi, Y. Murakami, M. Nakahara, *Proceedings of ECOC '83* (1983) PD Paper.
37. G. H. Sigel, B. D. Evans, *Appl. Phys. Lett.* **24** (1974) 410–412. Links
38. E.J. Friebele, M.E. Gingerich, *J. Non-Cryst. Solids* **38** (1980) 245–250.
39. T. A. Birks, P. St. J. Russell, D. M. Atkin, T. J. Shepherd, *Electron. Lett.* **31** (1995) no. 22, 1941. Links
40. J. C. Knight, T. A. Birks, P. St. J. Russell, D. M. Atkin, *Opt. Lett.* **21** (1996) no. 19, 1547. Links
41. R. F. Cregan, B.J . Mangan, J. C. Knight, T. A. Birks, P. St. J. Russell, P. J. Roberts, D. C. Allan, *Science* **285** (1999) 1537–1539. Links
42. A. Hasegawa, F. Tappert, *Appl. Phys. Lett.* **23** (1973) 142–144. Links
43. T.Nakahara, *Proc. International Symposium on Microwave and Optical Technology (ISMOT 2001)* (2001).

Further Reading

Cancellieri, G. (1993), "Single-Mode Optical Fiber Measurement: Characterization and Sensing". Boston: Artech House.

Izawa, T., Sudo, S. (1987), *Optical Fibers: Materials and Fabrication*. Tokyo: KTK Scientific Publishers.

Li, T. (1985), *Optical Fiber Communications*. New York: Academic Press.

Miller, S.E., Chynoweth, A.G. (1979), *Optical Fiber Telecommunications*. San Diego: Academic Press.

[Top of Page]

Glossary

MCVD

Modified chemical vapor deposition, one of the silica optical fiber fabrication methods developed by Bell Laboratories. The deposition of the core and the cladding layer is achieved by passing the oxygen and the raw material vapors, that is, $SiCl_4$ and $GeCl_4$, through a rotating heated silica tube. After deposition, the temperature is increased to collapse the tube into a solid preform. The preform is drawn into fiber by heating it in a drawing furnace.

Optical Communication System

The combination of a light source that converts electrical signals to light signals, an optical fiber that is a transmission medium for the optical signals, and a detector that reconverts the optical signals to electrical signals. It is a low-cost system because of the long repeater spacing and high transmission capacity compared to a conventional telecommunication system.

Optical Fiber

A transmission medium for optical communications systems made with transparent materials, i.e, glass or plastic. It consists of a core surrounded by a cladding layer that has a lower refractive index than the core material. An optical

signal is transmitted in the core by total internal reflection.

OVD

Outside vapor deposition, one of the silica optical fiber fabrication methods developed by Corning Glass Works. Fine glass particles produced by flame hydrolysis of the raw materials, i.e, $SiCl_4$ and $GeCl_4$, with a $C_2H_6O_2$ burner are deposited on the lateral surface of a rotating rod. After deposition, the rod is removed from the soot boule, which is then dehydrated by heating in chlorine gas and sintered to a transparent preform. The preform is drawn into fiber by heating it in a drawing furnace.

PCVD

Plasma chemical vapor deposition, one of the optical fiber fabrication methods developed by Philips. The oxygen and the raw material vapors, that is, $SiCl_4$ and $GeCl_4$, pass through a rotating heated silica tube, which is evacuated. In the silica tube, oxidation of the raw materials occurs in a plasma, and core and cladding glass layers are deposited on the inner surface of the tube. After deposition, the temperature is increased to collapse the tube into a solid preform. The preform is drawn into fiber by heating it in a drawing furnace.

VAD

Vapor-phase axial deposition, one of the silica optical fiber fabrication methods developed by Japanese companies. Fine glass particles produced by flame hydrolysis of the raw materials, that is, $SiCl_4$ and $GeCl_4$, with an O_2–H_2 burner are deposited on one end of a rotating seed rod. A soot preform then grows along the axis of the rod. The core and the cladding soots are deposited simultaneously. After that, it is dehydrated by heating in chlorine gas and sintered to transparent preform. The preform is drawn into fiber by heating it in a drawing furnace.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Fibers, 1. Survey

Fritz Schultze-Gebhardt, Düsseldorf, Federal Republic of Germany (Chaps. 2 – 7)

Karl-Heinz Herlinger, Institut für Chemiefasern der Deutschen Institute für Textil- und Faserforschung Stuttgart, Denkendorf, Federal Republic of Germany (Chaps. 2 – 8)

1. Introduction 1
2. History 2
3. Characteristics of Fibers 3
3.1. Fineness 3
3.2. Tenacity and Modulus of Elasticity . . 3
3.3. Elongation 3
4. Spinning 3
4.1. Wet Spinning 4
4.2. Dry Spinning 4
4.3. Melt Spinning 4
5. Prerequisites for Fiber Formation . . 5
5.1. Molecular Mass and Fiber Formation 5
5.2. Molecular Structure and Fiber Properties 5
5.3. Property Requirements for the Formation of Fiber Structures 6
5.4. Crystallization 7
5.5. Organization of Structural Elements 8
5.6. Structural Models 8
5.7. Molecular Symmetry and Physical Properties 9
5.8. Changes in Properties Caused by Symmetry Defects 9
6. Fiber Properties Required by Textiles 10
6.1. Requirements to Be Met by Textiles . 11
6.2. Modification of Fiber Properties . . . 11
6.3. Comfort Properties of Textiles 12
7. Economic Aspects 13
8. Tabular Survey of Fibers 14
9. References 35

1. Introduction

The term fibers refers collectively to a wide variety of forms of fibrous materials. Standards have been established to introduce order into the field. The most frequently employed terms are defined here. Natural fibers that can be spun into yarn are called *staple fibers.* The primary spinning of man-made fibers results in the production of *continuous filaments* (endless fibers). Indeed, both *monofilaments* (threads; spinneret has one hole) and *multifilaments* (spinneret has many holes) can be produced. *A filament yarn* consists of a large number of filaments that can be given texture by twisting, crimping and/or heat setting.

The word *tow* refers to a *fiber tape* made from thousands of filaments. If a tow is cut or torn, it gives rise to staple fibers. Fiber tapes obtained by cutting or tearing parallel to the fibers are known as *fiber bands.*

Conventional standards divide fibers into (1) *natural fibers* and (2) *man-made fibers* (also called chemical fibers).

Natural fibers are subdivided into plant and animal fibers. Man-made fibers are subdivided into natural materials brought into fiber form by a chemical reaction (regenerate fibers) and fibers made from synthetic polymers (synthetic fibers). The following classification includes standardized abbreviations that are used occasionally throughout this article:

1. Natural fibers
1.1. Plant fibers, e.g.,
cotton
flax
hemp
jute
1.2. Animal fibers, e.g.,
wool
camel hair
angora
silk
2. Man-made fibers
2.1. Fibers based on natural polymers, e.g.,
viscose rayon (CV)
Lyocell (CLY)
cellulose acetate (CA)
elastodiene (LA)
2.2. Fibers based on synthetic polymers (synthetic fibers), e.g.,
polypropylene (PP)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_451

polyacrylonitrile (PAN)
poly(vinyl chloride) (PVC),
polyamide 6 (PA 6, nylon 6)
polyamide 66 (PA 6,6; nylon 66)
poly(ethylene terephthalate) (PET)
poly(ether ether ketone) (PEEK)
polyurethane (PUR; spandex)

2. History

Natural Fibers. Natural fibers have been used by humans for thousands of years. Animal hair and plant fibers were spun into yarn and woven into textiles. Indeed, the textile industry today is still based on this ancient technology.

Man-Made Fibers. The idea of making "artificial" fibers and threads is over 300 years old. In fact, the start of the chemical fiber industry dates from about 1884. The process of regenerating fiber-forming cellulose was based on the discovery of solvents such as copper oxide – ammonia by E. Schweizer in 1857. Later, Count Chardonnet succeeded in forming derivatives of cellulose which could be solubilized and spun into threads; "Chardonnet silk" was first made in 1884. A detailed account of the history of man-made fibers is not given in this article. The subject was dealt with competently by H. Klare in 1985 [19].

The most important developments in the production of man-made fibers based on cellulose were:

1) the cuprammonium process for solubilization and spinning of cellulose;
2) the formation of spinning solutions of cellulose derivatives, such as cellulose nitrate and cellulose acetate;
3) the intermediate derivatization of cellulose to cellulose xanthate which in turn is spun into cellulose fibers;
4) the development of new solvent systems for cellulose, such as *N*-methylmorpholine oxide (NMMO) or dimethylacetamide – lithium chloride.

In 1927, H. Staudinger used polyoxymethylene as a model of cellulose to demonstrate that fiber-forming polymers were linear polymeric molecules.

The production of synthetic fibers was the result of pioneering work on the formation of synthetic polymers and the development of extrusion techniques known as wet, dry, and melt spinning.

Nylon 66 was first synthesized by W. H. Carothers in 1935. This was closely followed by the discovery of nylon 6 by P. Schlack in 1938. The work of J. R. Whinfield and I. T. Dickson led to the development of polyester fibers in 1941. Today, polyamides and polyesters are the largest-volume polymers capable of being melt-spun. The production of synthetic fibers based on familiar polymers such as polyacrylonitrile, which cannot be melt-spun, was made possible by the introduction of new solvents, e.g., dimethylformamide (H. Rein, 1942). Discovery of the stereospecific polymerization of propylene (G. Natta, 1954) led to the introduction of polyolefins into the fiber industry. The development of technology and the simultaneous elucidation of structure and properties resulted in "tailor-made" polymers and fibers. Specific properties such as rubber elasticity (O. Bayer, 1947) or extreme tenacity and stiffness were realized (high-modulus fibers, P. W. Morgan and S. L. Kwolek, 1968). In addition, high-temperature fibers were made from polyheterocyclic compounds.

Apart from organic polymers, inorganic substances such as glass, carbon, BN, and SiC were also spun into fibers.

Future developments are likely to be directed toward classical mass production, particularly towards the attaining of optimal processing characteristics and clothing comfort. New types of application – above all in the field of industrial fibers and in medical technology – will stimulate the development of special fibers with very specific properties.

3. Characteristics of Fibers

3.1. Fineness

The thickness of fibers and filaments ranges from 1 to 100 μm. Measuring thickness by means of conventional instruments is very difficult. Indeed, even small variations in fiber uniformity, thickness, and cross-sectional area hinder microscopic measurement. To make a sta-

tistically valid statement about the fineness of a fiber (fiber density), length and mass are computed instead of thickness. The unit *tex* is used to express fiber fineness. According to the ISO,

$$1\,\mathrm{tex} = 1\,\mathrm{g\,per\,1000\,m}$$

If 1000 m of fiber weigh 1 g, the fineness is 1 tex. Another unit commonly used to express fineness is dtex. If the mass per length is given in dtex, the numerical values will be comparable to those in the formerly used unit *denier* [46]:

$$1\,\mathrm{denier} = 1\,\mathrm{g\,per\,9000\,m}$$

3.2. Tenacity and Modulus of Elasticity

The cross-sectional area of fibers is generally nonuniform and cannot be determined easily. Hence, measurements based on area have limited applicability. For this reason, tenacity is not measured in gigapascal (GPa) units (ISO) but is expressed relative to the fineness. Usually, the maximal tensile force (at break) is employed, expressed in centinewtons per tex (cN/tex); other units used are cN/dtex (1 cN/dtex = 10 cN/tex) or N/tex (1 N/tex = 100 cN/tex). Textile and industrial fibers have tenacities ranging from 10 to more than 300 cN/tex (see Table 4). The modulus of elasticity corresponds to the tangent to the stress – strain curve at the origin (→ Fibers, 6. Testing and Analysis, Chap. 3.). It is also expressed relative to the fineness, in cN/tex. Typical moduli range from 50 to 1000 cN/tex for textile fibers and from 1000 to 40 000 cN/tex for industrial fibers. (See Table 5.)

3.3. Elongation

The highest possible degree of elongation at break (→ Fibers, 6. Testing and Analysis, Chap. 3.) varies greatly with the type of fiber. Industrial fibers, e.g., carbon fibers and high-modulus aramids, have values of 0.1 – 2 %; textile fibers and yarns have values of 5 – 70 %; and values of 300 – 700 % are obtained for spandex fibers (see Table 4).

In practical applications, fibers are elongated only to a small extent. Maximal elongation is never attained because otherwise fibers undergo partially irreversible stretching.

Temperature and moisture (weather) affect both the tenacity and the elongation of fibers to varying extents, depending on the type of polymer. The same applies to dimensional stability and wrinkling tendency, properties that are also influenced by the type of fiber and by the textile structure.

4. Spinning

Both natural and synthetic fibers are consisting of linear polymers. These polymers are converted into fibrous form by growth (animal hair, plant fibers) or extrusion (spider, silkworm, spinning technology) and are specifically oriented to the fiber axis. The many mechanisms found in nature for the formation of filamentous structures are, by no means completely understood (see e.g. [87]). Only a few processes are available for large-scale production of man-made fibers. Specific production techniques are based on the deformability of fiber polymers. In principle, a distinction is made between spinning methods involving solutions and those involving molten polymer. For a detailed description of spinning, see → Fibers, 3. General Production Technology.

4.1. Wet Spinning

Polymer solutions are converted into fibers by diluting a highly concentrated polymer solution in a coagulating bath. The extremely viscous polymer solution, e.g., 3.5 – 10 Pa · s (still higher values are obtained with viscose rayon) can be extruded, i.e., spun, by forcing it through a spinneret to form threadlike structures. In general, polymers are soluble only in specific solvents. The viscous polymer solution is extruded into a coagulating bath where the solvent is diluted, which results in precipitation of polymer threads. For instance, polyacrylonitrile is soluble in dimethylformamide (DMF). In the coagulating bath, DMF is washed out of the viscous polymer threads with water, and polyacrylonitrile fibers precipitate (Fig. 1).

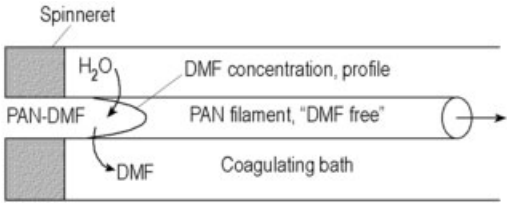

Figure 1. Fiber formation by solvent extraction in the spinning bath
PAN = polyacrylonitrile: DMF = dimethylformamide

The solvent concentration in the fiber decreases along the length of the coagulating bath. In wet spinning, several physicochemical processes occur simultaneously, e.g., solvent diffusion, polymer precipitation, and formation of membrane-like interfaces and system-specific morphological structures. One of the main purposes of developing wet-spun polymers is to produce specific fiber structures in the coagulating bath. This is achieved, in principle, by modification in the coagulating bath, which is described in more detail under → Fibers, 4. Synthetic Organic, Chap. 5.3.3., wet spinning of polyacrylonitrile. In a spinning process of this type, the polymer can also undergo a chemical reaction; e.g., the viscose process involves acid hydrolysis of cellulose xanthate in the coagulating bath (→ Cellulose, Chap. 3.1.3.).

To make the process economical, the polymer concentration should be kept as high as possible during spinning. In the case of soluble polymers, the viscosity of the spinning solution, and thus its basic spinnability, are directly related to the concentration and solvation of the polymers. However, complicated processes, such as exchange, diffusion, and drying, that occur during wet and dry spinning result in polymer- and solvent-dependent limitations on the polymer concentration.

4.2. Dry Spinning

In dry spinning, the polymer solution is also forced through a spinneret. Solvent is then evaporated in a warm current of air to produce almost solvent-free filaments. Cellulose acetate, polyacrylonitrile, and polyurethane are spun by this technique (→ Cellulose Esters, → Fibers, 4. Synthetic Organic).

4.3. Melt Spinning

In melt spinning, the polymer is melted by heating and then passed through a spinneret via a spinning pump. Only polymers that are thermally stable under melt conditions can be subjected to this extrusion (i.e., spinning) process. The fluid threads emerging from the spinneret solidify on cooling (usually air cooling) to form filaments. This technique is applied in spinning, e.g., nylon 6, nylon 66, or poly(ethylene terephthalate) (→ Fibers, 4. Synthetic Organic).

The difference between melting and spinning temperatures depends on the viscosity of the melt. Usually, the spinning temperature is about 30 K higher than the melting temperature T_m. However, if the viscosity is still too high, the spinning temperature must be increased even further. For example, T_m of polypropylene is 170 °C but the polymer is spun at 260 °C. The melt index, not the viscosity, is generally used to characterize the melting properties of polymers. In fact, the tedious measurement of melt viscosity has now been replaced by simple mass determination (→ Fibers, 6. Testing and Analysis).

Hence, melt spinning requires polymers that are thermally stable and, as far as possible, resistant to thermal oxidation at temperatures approximately 30 K higher than their melting temperatures (PA-6, PA-66, and PET).

Production conditions must be optimally coordinated with the properties of the product to achieve the desired melt viscosity, on the one hand, and the required fiber tenacity, on the other. Rapid advances in process technology, however, have made possible not only the processing of very highly viscous melts, but also the extrusion of solids.

5. Prerequisites for Fiber Formation

5.1. Molecular Mass and Fiber Formation

The spinnability of polymer melts or solutions depends on their viscosity at technically feasible spinning conditions. The viscosity η of linear polymers depends, in turn, on the molecular mass M_r:

Table 1. Polymer type and molecular mass suitable for fiber formation

Polymer	Repeating unit	Fiber formation at M_r	Degree of polymerization
Polyethylene	$-CH_2-CH_2-$	> 1 000 000	> 62 500
Polypropylene	$-CH(CH_3)-CH_2-$	> 1 000 000	> 45 455
Polyacrylonitrile	$-CH(C\equiv N)-CH_2-$	> 30 000 – 70 000	> 1070 – 2500
Aliphatic polyamides	$-CH_2-NH-CO-CH_2-$	15 000 – 30 000	380 – 770
Aliphatic polyesters	$-CH_2-O-CO-CH_2-$	50 000	1300
Aliphatic–aromatic polyesters	$-O-CH_2-CH_2-O-CO-C_6H_4-CO-$	22 000 – 26 000	220 – 260
Aromatic polyamides	$-NH-C_6H_4-NH-CO-C_6H_4-CO-$	60 000	550

$$\eta = K \cdot M_r^{\alpha}$$

where K is a constant that is known for most common polymers.

The fiber-forming tendency of a polymer increases with increasing molecular mass. The varying forces of interaction between individual structural groups of polymers are directly related to the chemical nature of these linear molecules. For this reason, optimal conditions for fiber formation are very polymer-specific. In principle, the weaker the interaction between structural segments, the higher is the molecular mass required (Table 1) [47].

5.2. Molecular Structure and Fiber Properties

The thermochemical and photochemical stability of a fiber polymer is controlled by its chemical structure. The possibility of chemically modifying textile properties is determined by the chemical reactivity and polarity of the individual structural units (Table 2). The type of chemical bonds and the stereochemistry are important for molecular geometry, interaction of structural units, and thermal conformational stability. Low conformational mobility is generally related to high mechanical tenacity, modulus, and torsion modulus. Finally, the molecular organization is influenced by the symmetry of the structural units and of the entire molecular chain. The interaction energy (cohesion energy) controls the solubility and compatibility of structural units with each other. These factors finally lead to specific morphological domains, which frequently occur in block structures of varying chemical composition along the polymer chain, e.g., in polyurethane (spandex) fibers. The general relationship between the structure and properties of polymers suitable for fiber production is discussed in the literature [47]. In this section the structural elements of technically important fibers are compared.

Table 2. Structural characteristics of polymers and their importance for fiber properties

Structural characteristics	Importance for fiber properties
Chemical structure (primary structure)	thermochemical and photochemical stability, textile chemical processes (dyeing, textile finishing)
Molecular organization and symmetry	general properties, melting point, shrinkage behavior
Type of binding and stereochemistry (secondary structure)	thermophysical and thermomechanical properties (moduli, elasticity)
Conformation and conformational mobility	
Morphological structures, superstructures	diffusion (dyeing), breaking mechanism

Crystallization. The interaction between structural units is of considerable importance for crystallization. The chemical structure, stereochemistry, and symmetry of polymers determine, to a large extent, the structural conforma-

tion of the polymer under prevailing conditions such as temperature (melt or solid), condition of solution (solvation of molecules), coil structure, and aggregation. Higher orientation or parallelism of the chain molecules increases orientation of the structure elements. Their interaction also depends on the possible molecular conformations.

During spinning, a preliminary orientation of the molecules occurs, which is enhanced during drawing, and stabilized by partial crystallization. The individual segments usually crystallize independently of each other; hence amorphous intercrystalline areas exist unavoidably. The course of polymer crystallization, especially of fibers, is quite different from the crystallization characteristics of low molecular mass compounds.

5.3. Property Requirements for the Formation of Fiber Structures

Fiber-forming polymers require certain properties that make possible the formation of typical fiber structures; these are linearity, intermolecular forces, and the possibility of crystallization whereby, in general, supermolecular structures result.

1) *Linearity.* Only linear polymers can be expected to have optimal fiber-forming properties. Linearity is specifically a prerequisite for the spinnability of polymer solutions and melts.
2) *Intermolecular Forces.* After orientation of linear polymers during spinning and drawing, dispersion forces, e.g., dipole-dipole interactions or hydrogen bonds between the polymer segments contribute to the fixation of the polymer chains.
3) *Crystallization.* The well-developed organization of the polymer chains usually results in crystallization. Polymer-specific crystallization results in crystalline domains and amorphous intercrystalline areas.
4) *Supermolecular Structure.* The crystallites are oriented so that they lie parallel to the fiber axis, which results in supermolecular structures, e.g., fibrils and fibril bundles (→ Fibers, 2. Structure).

These requirements were fulfilled by STAUDINGER's model of cellulose as polymeric formaldehyde [48]. Even this simple compound met all the conditions for fiber formation. It is a meltable linear polymer, capable of crystallization with formation of fiber fibrils.

The simplest model for intermolecular dispersion forces (van der Waals forces) is shown in a polyolefin molecule such as polyethylene or polypropylene:

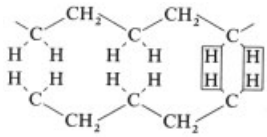

In poly(ethylene terephthalate), interaction between the aromatic π-electron systems of the benzene rings and of the carbonyl groups makes dipole – dipole interaction possible:

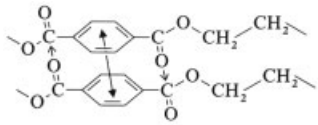

Hydrogen bonds act on amide or urea groups in polyamides, such as nylon 66 or nylon 6, or in polyurethane, respectively:

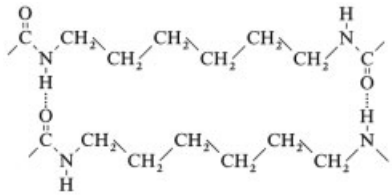

The OH groups of poly(vinyl alcohol) and of cellulose (viscose rayon) are also able to form hydrogen bonds. Very strong dipole – dipole interaction occurs between the nitrile groups of polyacrylonitrile:

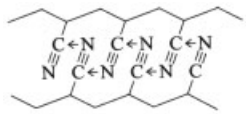

However, this all-trans conformation of polyacrylonitrile can change into lower energy helical structures [49]:

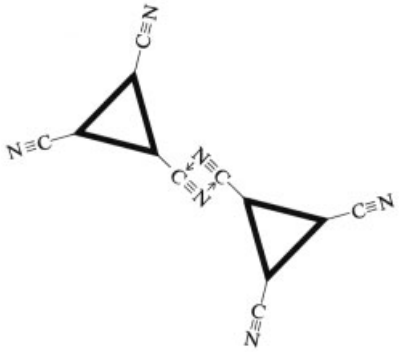

The strong dipole–dipole interaction of polyacrylonitrile is one of the reasons for the specific solubility behavior of that polymer.

Weak dispersion forces are effective only in a highly symmetrical conformation. This high degree of symmetry, in turn, is possible only with a highly symmetrical arrangement of substituents, e.g., along an aliphatic carbon chain. Such a high degree of symmetry in linear polymers can be achieved only by coordination polymerization, in the presence of organometallic catalysts (Ziegler – Natta), to yield isotactic polymers. If this symmetry (e.g., in polypropylene) is disturbed by the presence of other structural units (e.g., carboxylic acid ester groups; acrylic acid comonomers), both polymer and fiber properties change drastically.

5.4. Crystallization

The flexible linear polymer takes on a variety of conformations during fiber formation, seeking the conformation with the lowest free energy (i.e., highest molecular organization). However, spinning conditions do not allow polymers to attain this state completely, and fiber polymers usually undergo only partial crystallization. The ratio of crystalline to amorphous regions varies from almost totally amorphous to highly crystalline, single-phase systems (Fibers, 2. Structure).

Linear polymers exist in solution or as molten polymer in a more or less coiled form (Fig. 2 A). In the shearing field during spinning, the molecules are partially uncoiled and oriented in the direction of flow (Fig. 2 B). A higher degree of orientation is achieved during the subsequent drawing operation (Fig. 2 C). In this step, spun fibers are stretched by more than 200 %, resulting in partial crystallization of polymer molecules (Fig. 2 D). In high-speed spinning, the extremely high windup speeds (3000 – 6000 m/min; 50 – 100 m/s) induce crystallization during fiber formation. The resulting preoriented yarns (POY) have to be stretched by only 40 – 100 % in subsequent steps (texturing). Additional polymer-specific increases in windup speeds lead to fully oriented yarns, which require almost no further stretching.

Wu et al. [77] succeeded to orientate spinning PET in a threadline modification process. The one-step orientation (without separate drawing) could be realized by the installation of a liquid bath in the spinline. High-performance filaments, however, have to be drawn after spinning to reach maximum fiber strength.

Liquid Crystals. Some linear polymers, in solution or molten form, tend to undergo self-organization by forming liquid – crystalline structures. Soluble linear polymers (lyotropic systems) form liquid crystals as a result of their rigid conformation and the formation of polyelectrolytes, e.g., the system poly(*p*-phenyleneterephthalamide) – H_2SO_4. Liquid crystals can also be formed in polymer melts (thermotropic systems). Examples of suitable systems are the aromatic polyesters of the poly(4-hydroxybenzoic acid) type or aromatic blocks built up in segments and separated by flexible spacers. Figure 2 E shows the resulting structures of some high-modulus aramids; pronounced amorphous intercrystalline areas are not visible.

5.5. Organization of Structural Elements

The arrangement of fibrils and fibril bundles in man-made fibers is controlled by the production procedure, i.e., by physical and physicochemical processes. Therefore, resulting structures depend on production parameters, and despite the different types of polymers employed, the tertiary structures of all man-made fibers are very similar.

In contrast, natural fibers have organized, evolved structures which differ considerably from each other, depending on the organism. Structural variations are reflected in different geometric positions of the fibrils with respect to the

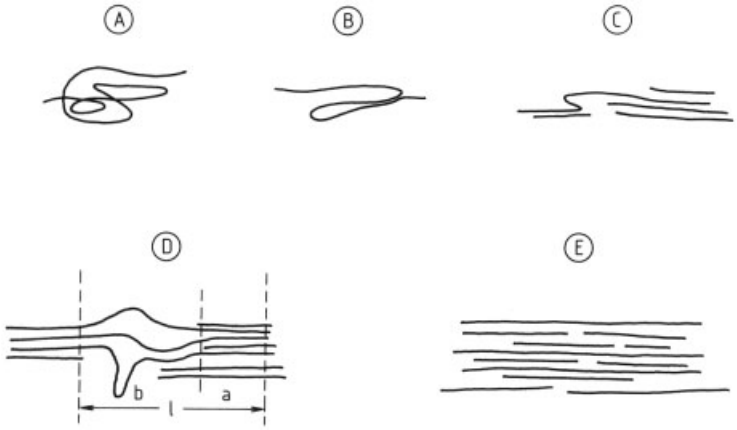

Figure 2. Various conformations of fiber polymers
A: Unstirred solution (coil of polymer molecules)
B: Molecular orientation in the shearing field during spinning
C: Molecular orientation during drawing or liquid crystal formation from the melt or solution (polyelectrolyte molecules)
D: Formation of structure on crystallization of flexible polymer chains. Examples are polyethylene, polypropylene, polyamide, and polyester.
a) Crystallite; b) Intercrystalline (amorphous) region; l) Long period (10 – 20 nm)
E: Formation of structure of lyotropic and thermotropic systems, such as crystallization of rigid polymer chains. Examples are high-modulus fibers of the aramid type, polyesters, and carbon fibers.

fiber axis (e.g., in cotton fibers) or in different chemical structures along the fiber cross section (e.g., in animal hair, such as wool). Fibers could thus be classified into two principal structural categories, namely, organized and unorganized, if the structure of each fiber were known.

5.6. Structural Models

The chemical structure of linear polymers determines the degree of conformational mobility of the chain segments of these molecules. Both chain mobility and molecular structure, in turn, greatly influence fiber structure. Flexible chain molecules may crystallize partially to form structures that conform with the classical crystalline – amorphous model (Fig. 3 A). Spandex fibers, on the other hand, have a special domain structure, because the chemical structure of these polymers changes considerably in the various segments along the polymer chain (Fig. 3 B).

The relatively planar structure of aramids and polyheterocyclic compounds, their conformational rigidity, and their tendency to form lyotropic structures in the spinning solution are consistent with a structural model in which the intercrystalline segments are bridged (Fig. 3 C).

Even more planar molecules form graphite-like structures (Fig. 3 D). Structural defects lead to cavities and weakly bent planes of the aromatic systems; for more details on carbon fibers, see → Fibers, 5. Synthetic Inorganic.

5.7. Molecular Symmetry and Physical Properties

The importance of molecular symmetry for thermophysical properties (melting point, glass transition temperature) can be demonstrated for polyamides. Figure 4 shows the fluctuation of melting points in the series of polyamides PA-3 to PA-13 (nylon 3 to nylon 13). Apparently, polyamides with an even number of carbon atoms and those with an odd number of carbon atoms belong to different structural series [47]. This can be explained by the different overall symmetry and, consequently, the different possibilities for hydrogen-bond formation (Fig. 5). In PA-7, more hydrogen bonds can be formed – at least in the all-trans conformation – than in PA-6. Analogously, all nylon types with an even number of carbon atoms have lower melting points than those with an odd number of carbon atoms (Fig. 5).

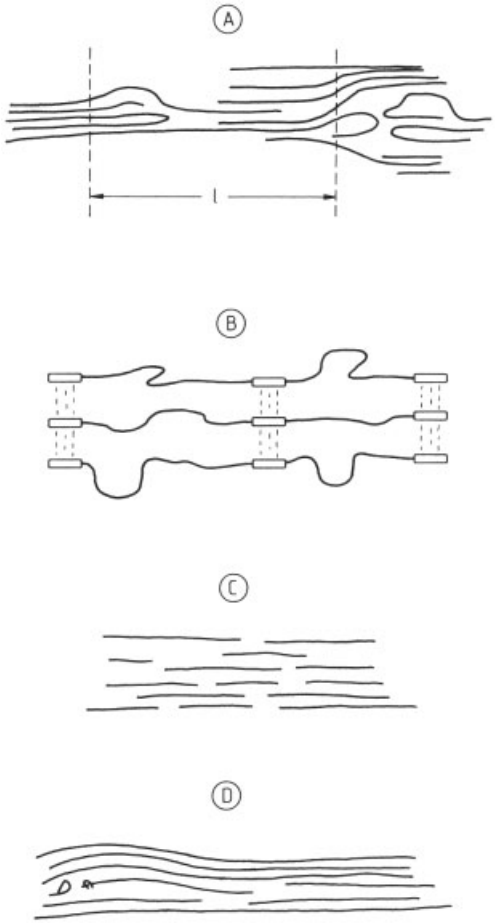

Figure 3. Various structural models of different polymer types
A: Classical structural model of amorphous – crystalline fiber polymers. Examples are polyamide, polyester, and viscose rayon. Long period (l) 10 – 30 nm.
B: Structural model of spandex fibers (polyurethane). Length of hard segment 2.5 nm; length of soft segment 15.0 nm
C: Structural model of aramid fibers (p-structures). No amorphous phases present; stretched molecules due to lyotropic structures during fiber formation
D: Structural model of carbon fibers; graphite structure with defects (cavities, bent layers; D = position of defect).

5.8. Changes in Properties Caused by Symmetry Defects

The incorporation of comonomers affects not only chemical structure, but also symmetry and, therefore, other fiber properties. For instance, the tacticity of polypropylene is changed by the insertion of comonomers, such as vinyl compounds. The resulting copolymer has altered properties and is no longer suitable for textile production. In addition, polar comonomers interfere with the isotactic course of the polymerization process itself.

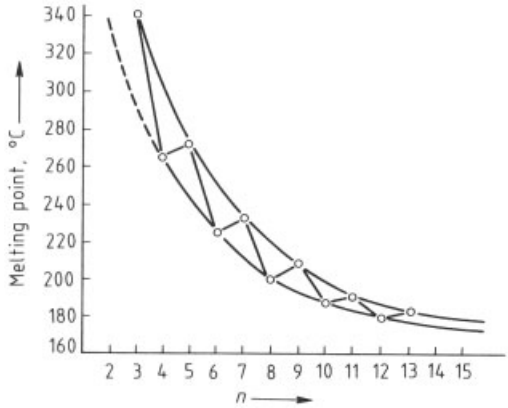

Figure 4. Melting points of nylons as a function of chain length n.

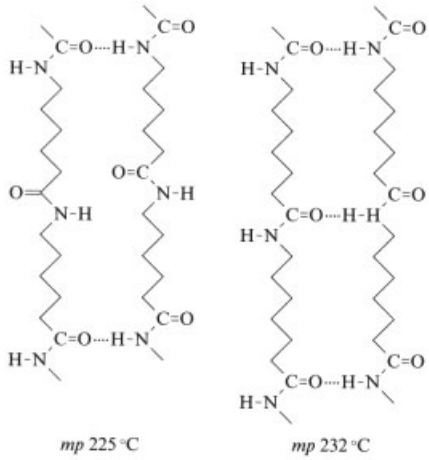

Figure 5. Hydrogen bonding in nylon 6 (A; *mp* 225 °C) and nylon 7 (B; *mp* 232 °C)

Polyacrylonitrile is rarely used for textile fibers without comonomers being incorporated into the polymer, because the desired fiber properties (e.g., deformability, dyeability) cannot be achieved with polyacrylonitrile alone. This, of course, changes the solubility of the polymer.

Polyesters of similar chemical structure, but quite different molecular symmetry, also have very different melting points. For example, the

highly symmetrical poly(ethylene terephthalate) has a melting point of 265 °C, whereas the melting point of poly(*m*-phenyleneisopthalamide) is only 102 °C [47].

The same situation arises with aramids. Here, too, the highly symmetrical para-linked polyamides have higher melting points (in this case, decomposition temperatures) than the less symmetrical meta-linked polymers (Fig. 6). The different dipole orientation, oppositely directed or in the same direction (arrows in Fig. 6), and the total symmetry cause higher fiber tenacity and modulus.

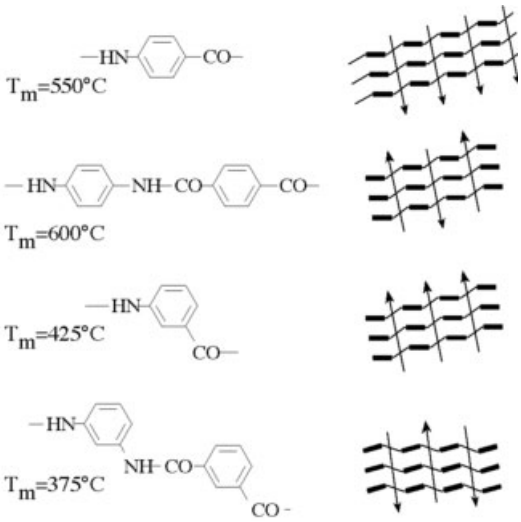

Figure 6. Effect of dipole orientation on the melting point (decomposition temperature) of aramids

6. Fiber Properties Required by Textiles

Despite the fact that a large number of polymers possess the basic properties required for fiber formation, they are not all suitable for large-scale production of fibers. In textile production and in commercial applications, polymers must fulfill a variety of requirements:

1) *Chemical Stability.* Polymers must be stable under the influence of heat, light, air, water, and the chemicals commonly used in textile finishing and care.
2) *Thermomechanical Stability.* Polymers must possess the ability to be solution- or melt-spun and, at the same time, retain their dimensional stability during processing, use and care.

The requirements to be met by fibers and their parent polymers are, therefore, determined by the demands on the final products.

6.1. Requirements to Be Met by Textiles

The need for a variety of textile fabrics having specific properties should be clear from the diversity of textile uses. The priorities and order of importance of properties depend on the application in question.

For *clothing,* textile properties like appearance, aesthetics, optics, drape, formability (ironing, pleating), comfort, heat and moisture transport, handle, ease of care, and fastness of all types are of considerable importance. In addition, certain chemical requirements, such as dyeability, crease-resistant finishing, and dimensional stability, must be met in the production of textile materials. Finally, processibility to the finished textile product in partially automated production lines has become increasingly important.

Industrial textiles must also possess specific properties, depending on their application. For example, high tenacity is required for ropes; high modulus, for fiber reinforcement; heat stability, for protective clothing and insulation; high water absorption, for articles of hygiene; resistance to chemicals and flue gas, for filter materials; and long-term stability, for geotextiles [45].

In the past, fiber production and the utilization of specific properties were optimized primarily. Today, fibers with desired properties are made specifically to fulfill the requirements of the final product. The correlation between requirements and properties of the finished product and corresponding fiber (polymer) properties is given in Table 3.

6.2. Modification of Fiber Properties

The specific properties of the finished product depend directly on its end use. Since the fiber itself is not the final product, changes in fiber structure result in changes in the property profile

Table 3. Correlation of textile properties with fiber properties

Properties of the final product	Corresponding fiber properties
Optical properties	luster, fiber surface
Aesthetics	profile, fiber cross section
Mechanical properties	modulus of elasticity, tenacity, elongation
Comfort	
Physiological properties of clothing	moisture absorption, moisture transport
Antistatics	electrical resistance
Thermal insulation	heat capacity, porosity, heat conduction ∗
Hand	textile structure, bending modulus
Feel	roughness, modulus of elasticity, fineness
Ease of care and washability	wetting, moisture absorption, glass transition temperature (wet and dry) [78]
Dry cleaning	polymer (in)solubility, swelling
Soiling characteristics	zeta potential, adsorption and dissolution of soil components
Fastness	
Mechanical stability	tenacity, elongation, moduli, abrasion resistance
Dimensional stability	melting point, glass transition point (wet and dry)
Forming (e.g., pleating)	thermoplasticity, glass transition temperature (T_g)
Lightfastness	chemical structure, sensitizers, stabilizers
Lightfastness of dyeing	polymer – dye interaction, radical lifetime
Specific properties	
Flame resistance	chemical structure, combustion mechanism
Impermeability to water	moisture absorption, wetting properties
Water vapor permeability	surface properties, morphology, yarn structure
Dyeability	glass transition point during dyeing
Mechanical properties (ropes, material, tire cord)	modulus of elasticity, tenacity, elongation, dynamic modulus
Rubber elasticity	chemical structure, glass transition temperature, morphological structure, domain structure

∗ The thermal insulation of a textile is largely due to the air inside the pores, which is of low thermal conductivity.

of the finished product. Matching the requirement profile with the property profile is not always possible. the purpose of numerous modification procedures (e.g., of the fiber polymer, production method, processing, and finishing) is to confer special functional properties on the end product. some of these procedures are described in more detail in the following sections.

copolymers. Fibers are used predominantly as raw materials in textile manufacture. specific fabric properties, such as dimensional stability, brightness, color, pattern, or luster, are imparted to textiles during the physical and physicochemical processes of textile finishing. These characteristics should remain unchanged, particularly during use and cleaning. Chemical modification of the fiber polymer is generally employed to achieve this goal. Chemical treatment involves, above all, the introduction of special structural groups into the basic polymer, groups that are capable of taking part in polar and homopolar interactions with the chemicals used in textile finishing. Some examples are

1) introduction of a sulfonic acid group into poly(ethylene terephthalate) to bind cationic dyes (sulfoisophthalic acid as comonomer);
2) introduction of a sulfonic acid group into polyacrylonitrile with the help of vinylsulfonic or styrenesulfonic acid or a similar comonomer to make possible the use of cationic dyes;
3) introduction of flexible chain segments into poly(ethylene terephthalate) to change its dyeing characteristics. Polymer modification [copolycondensation with poly(ethylene glycol)] in the production of carrier-free dyeable polyester makes use of this principle.

Polymer Mixtures. Polymers are generally incompatible. If two different polymers are spun together, fibril structures are formed (matrix fibril fibers) that are known as *biconstituent fibers.* If two polymer streams are combined in the spinneret without blending, *bicomponent fibers* are obtained (→ Fibers, 3. General Production Technology).

Additives. Many specific fiber characteristics can be achieved by use of additives in the spinning solution or the melt. For example, the following effects can be obtained:

1) Delustering: use of titanium dioxide as additive in spinning
2) Spin dyeing: use of pigments as additives in spinning
3) Antistatic properties: use of electrically conducting additives in spinning
4) Flame resistance: use of flame retardants as additives in spinning

The use of additives in spinning produces a very broad spectrum of properties. For economic reasons, this type of fiber modification is the method of choice, if it produces the required effect. By comparison, the development of special polymers is a much more expensive procedure.

6.3. Comfort Properties of Textiles

Fibers used to produce clothing must possess a variety of properties. The term comfort refers to those properties responsible for the wash and wear performance of garments. Physiological studies have shown that apart from fiber substance, i.e., the chemical structure of the fiber polymer, the most important factors that influence textile wearability are fiber, yarn, and textile construction [50]. Along with moisture absorption, moisture transport across the fiber surface or through textile spaces is of particular significance for the moisture balance of textiles. Textiles must retain their appearance and form during washing and dry cleaning. In this connection, certain inherent fiber properties are of vital importance. The thermomechanical stability of textiles is controlled by the glass transition temperature (T_g) of the fiber polymers. However, this temperature is not a structural property but, rather, is directly related to the fiber system, e.g., the type of polymer and polymer – water interaction. Deformability generally increases with increasing moisture absorption because the chain-segment mobility of polymers is increased. The glass transition temperature of fiber polymers with a high moisture-absorbing capacity decreases dramatically with increasing moisture content. The resulting structural changes in the fiber lead to fiber shrinkage or stretching, depending on the degree of thermosetting. In other words, the textiles lose their shape.

Even in the dry state, dimensional stability decreases with increasing temperature, fibers can undergo permanent deformation at dry heat. Suitable fixation and textile finishing, e.g., thermofixation (heat setting) of polyester, felting or shrink-resistant finishing of wool, and crease-resistant finishing of cotton, can enhance dimensional stability and ease of care. Textiles that are still not dimensionally stable enough to be washed with water must be dry-cleaned. Here, soil release is achieved by using organic solvents such as perchloroethylene or fluorochloro-hydrocarbons (→ Dry Cleaning); the temperature must be maintained below the glass transition points of the fiber – solvent system throughout the process.

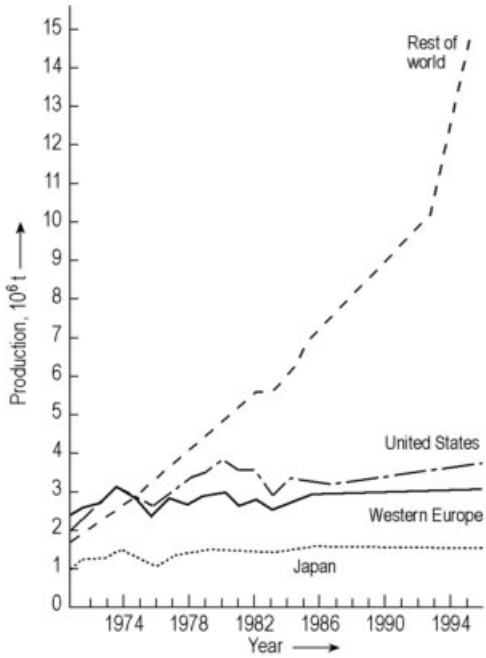

Figure 7. Total production of man-made fibers [2]

7. Economic Aspects

The production of natural and man-made fibers is influenced by a number of factors, such as demand (state of the economy), availability (production and trade), prices, and availability of raw materials. The per capita consumption varies

considerably from country to country and depends on climate, supply, domestic production versus imports, standard of living, fashion, state of the economy, and degree of industrialization.

The balance between the consumption of natural and of man-made fibers is also controlled by these factors, although a certain "awareness of nature" currently tips the scales in favor of natural fibers. The production of natural fibers, e.g., cotton or wool, is determined in terms of the profit made per area cultivated. A dramatic increase in production cannot be expected at present. However, new developments in genetic engineering could rapidly improve the yields obtained per given area.

The economic importance of individual fiber types can be deduced from their annual worldwide consumption [51], and Figure 7 shows the worldwide production of man-made fibers.

Both the consumption and the production of fibers also depend directly on the level of development and on per capita income [52]. The consumption per person in individual countries differs considerably for a variety of reasons. Some countries have already reached a high level of consumption, and despite their high standard of living, the population spends only a fixed percentage of their available income on textiles. The growth of fiber production for industrial use, on the other hand, is likely to continue. In 1984, 7 kg of fiber was consumed per capita; for the year 2000, a total fiber production of 43×10^6 t is expected if per capita consumption remains at 7 kg, or 48.6×10^6 t if it increases to 8 kg. The growth rate of 2.8 % achieved in the period 1968 – 1984 is not expected to continue. Up to 1995 the average growth rate was only 2.3 %/a. Nevertheless, a world fiber volume of about 50×10^6 t is expected in the year 2000 [52], [79].

The limited means of production in Eastern Europe also contributed to reduced consumption. In the developing countries, lack of spending power, capital, and means of production all combine to reduce growth of consumption. Regional differences in population growth will lead to an increase in fiber consumption in the developing countries.

8. Tabular Survey of Fibers [6], [53], [54]

A quantitative survey of the most important properties of fibers based on polycondensation and polymerization, as well as some inorganic fibers, is shown in Tables 4, 5, 6, 7, 8, 9, 10. Properties of these fibers, are compared with those of natural and semisynthetic fibers. The properties of all textile and industrial fibers are determined by variables related to production and aftertreatment. Hence, in general, only a range of values characteristic of individual fiber classes is given in this survey. In fact, only the order of magnitude is indicated for electric resistance of fibers because of considerable variations in measured values. No entries have been made when reliable information was unavailable.

The numbers in brackets are plausible estimated values. Figures for tenacity and modulus are based on both fiber fineness ($1\,\mathrm{tex} = 1\,\mathrm{g/km}$) and cross-sectional area of the sample. The following relationship is applicable:

$$\text{Fineness (tex)} = 10^5 \times \text{Density}\left(\mathrm{g\,cm^{-3}}\right) \times \text{Crosssection}\left(\mathrm{cm^2}\right)$$

In general, the characteristic stress – strain curve $\sigma(\varepsilon)$ of fibers does not show a constant gradient, even at low elongation. Hence, the modulus of elasticity is usually defined as the differential quotient $\mathrm{d}\sigma(\varepsilon)/\mathrm{d}\varepsilon$. Experimental determination at low fiber elongation ε is difficult but can be conducted successfully by extrapolation of the function $\mathrm{d}\sigma(e)/\mathrm{d}e$ for $e = 0$ [55].

The elastic recovery of a fiber (in percent) can be measured by standard methods [56] and is shown in Table 5 as

$$\left(1 - \frac{e_\mathrm{r}}{e}\right) 100$$

where ε_r = residual elongation, ε = total elongation.

Table 4. Mechanical properties of fibers (data at break)

Fiber name/polymer	Trade names	Fiber type [a]	Density (ϱ), g/cm^3	Elongation ε_H (21 °C) at 65 % R.H. [b], %	Elongation ε_H (21 °C) "wet/dry," %	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H/ϱ, cN/tex	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H GPa	Tenacity (σ_H) at 21 °C Wet/dry (relative), %	Relative loop tenacity [62], %	Relative knot tenacity [63], %
Natural fibers										
Cotton		st	1.50 – 1.54	6 – 15	100 – 110	25 – 50	0.35 – 0.7	100 – 110	65 – 75	60 – 100
Wool		st	1.32	25 – 50	110 – 130	10 – 20	0.15 – 0.25	70 – 90	75 – 85	80 – 85
Silk		fi	1.25	10 – 30	120 – 200	25 – 50	0.3 – 0.6	75 – 95	60 – 80	80 – 85
Hemp		st	1.48	2 – 5	≈ 100	36 – 75	0.5 – 1.1	100 – 106		
Flax		st	1.438 – 1.456	1 – 4	110 – 125	30 – 55	0.43 – 0.8	105 – 120	20 – 40	
Jute		st	1.436	1.3	100	20 – 40	0.3 – 0.57	99 – 104		
Regenerate cellulose										
Viscose rayon, Viscose Modal	Evlan, Fibro, Sarille,...	st/fi	1.52 – 1.54	10 – 30	100 – 130	16 – 45	0.25 – 0.7	40 – 80	25 – 65	25 – 60
Cuprammonium rayon	Asahi Bemberg, Bemberg	st/fi	1.52	10 – 40	170 – 200	14 – 21	0.2 – 0.32	60 – 70	60 – 75	60 – 85
Cellulose acetate	Arnel, Celco, Dicel	st/fi	1.29 – 1.32	20 – 45	120 – 150	10 – 15	0.13 – 0.2	50 – 80	70 – 95	80 – 90
Cellulose triacetate	Estron, Silene, Tricel	st/fi	1.29 – 1.32	20 – 45	120 – 150	10 – 15	0.13 – 0.2	50 – 80	70 – 95	80 – 90
NMMO fibers [88]	Lyocell, Tencel,...	st	1.5	10 – 16	110 – 115	32 – 48	0.5 – 0.7	55 – 70	45 – 50	≈ 75
Protein fibers		st	1.315	25 – 60	60 – 110	11 – 12	0.14 – 0.16	40 – 65		
Alginate fiber (Ca-Alginate)		fi	1.78	5		11 – 18	0.2 – 0.3	27 – 30		
Polycondensate fibers										
Nylon 6 (PA 6)	Perlon, Acelon, Amilan, Anso, Caprolan, Grilon,...	st	1.14	30 – 70	105 – 125	30 – 40	0.35 – 0.45	80 – 90	65 – 85;	
		fi	1.14	20 – 45	105 – 125	40 – 60	0.45 – 0.7	85 – 90	70 – 95	80 – 90;
		hd	1.14	15 – 20	105 – 125	60 – 90	0.7 – 1	85 – 90	70 – 90	60 – 70
Nylon 66 (PA 66)	Nylon, Antron, Cantrece, Meryl, Timbrelle, Ultron...	st	1.14	30 – 60	105 – 125	35 – 40	0.4 – 0.45	see PA 6	see PA 6	see PA 6
		fi	1.14	20 – 40	105 – 125	40 – 60	0.45 – 0.7			
		hd	1.14	15 – 20	105 – 125	60 – 90	0.7 – 1			

Table 4. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Density (ϱ), g/cm^3	Elongation ε_H (21 °C) at 65 % R.H. [b], %	Elongation ε_H (21 °C) "wet/dry," %	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H/ϱ, cN/tex	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H GPa	Tenacity (σ_H) at 21 °C Wet/dry (relative), %	Relative loop tenacity [62], %	Relative knot tenacity [63], %
Aramids							Compressive strength, axial in GPa, [80], [81], [82]			
Poly(*m*-phenyleneisophthalamide)	Nomex, Conex	fi	1.38	15 – 30	60 – 80				≈95	80 – 85
	Kevlar 29, Twaron	fi	1.44	3 – 4		185 – 195	2.7 – 2.8	0.35	≈45	≈50
Poly(*p*-phenyleneterephthalamide) (PPTA)	Kevlar 49	fi/an	1.45	2 – 2.5		185 – 195	2.7 – 2.8	0.365	50 – 78	35 – 50
	Kevlar 149	fi/an	1.47	1.2 – 1.9		150 – 160	2.2 – 2.4			
	Kevlar 981, Twaron	fi/hd		2.8 – 3.5		230 – 260	3.5 – 3.8			
	Kevlar Hp	fi	1.44	3.6		140 – 150	2.05			
Poly(ethylene terephthalate) (PET)	Dacron, Diolen,	st	1.36 – 1.41	25 – 50	100 – 105	30 – 55	0.4 – 0.75	95 – 100	75 – 95	
	Fortrel, Grilene,	st/ap	1.36 – 1.41	30 – 55	100 – 105	25 – 40	0.35 – 0.55	95 – 100	75 – 85	
	Serene, Terylene,	fi	1.36 – 1.41	20 – 30	100 – 105	40 – 60	0.55 – 0.85	95 – 100	70 – 98	70 – 80
	Trevira,....	fi/hd	1.36 – 1.41	8 – 20	100 – 105	60 – 100	0.85 – 1.5	95 – 100	60 – 90	40 – 70
Poly(butylene terephthalate) (PBT)	Trevira 810, Trevira 813	st/fi	1.25	35 – 40		40 – 45	0.5 – 0.56			
Poly(1,4-dimethylenecyclohexane terephthalate) (PDCT) [59]	Kodel	st	1.22 – 1.23	15 – 35		22 – 36	0.37 – 0.45	≈ 100	30 – 95	
Polyarylate	Ekonol [60], Vectran	fi	1.40 – 1.41	2 – 5		200 – 280	2.8 – 3.9	Compressive strength, axial [80], [81], [82] GPa 0.04 – 0.075		23 – 34
Poly(etherether ketone) (PEEK)	Zyex	fi	1.27 – 1.298	15 – 60		20 – 100	0.25 – 1.3		55 – 80	65 – 68
Polyimide (PI)	P 84	st/fi	1.41	for st 30 – 38 and for fi 19 – 21		27 – 37	0.38 – 0.53		60 – 65	70 – 75
Novolak	Kynol, Philene, Novoloid	st	1.2 – 1.3	20 – 60		13 – 22	0.17 – 0.27		70	80
Poly(amide imide) (PAI)	Kermel	st		10 – 25		25 – 60				
Poly(ether imide) (PEI)		st/fi		16 – 38		15 – 27	0.2 – 0.35			
Polyurethane elastomers (spandex)	Dorlastan, Lycra, . . .	fi	1.1 – 1.3	400 – 700	≈ 100	5 – 12; 30 – 70 = $\sigma_H (1 + \varepsilon_H)/\varrho$ [c]	0.06 – 0.15	75 – 100		

Table 4. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Density (ϱ), g/cm^3	Elongation ε_H (21 °C) at 65 % R.H. [b], %	Elongation ε_H (21 °C) "wet/dry," %	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H/ϱ, cN/tex	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H GPa	Tenacity (σ_H) at 21 °C Wet/dry (relative), %	Relative loop tenacity [62], %	Relative knot tenacity [63], %
Polymerizate fibers										
Polyethylene (HD – PE) (HM – HD – PE: $\bar{M}_r > 10^6$)	Hiralon, Vegon,...	fi	0.95 – 0.96	10 – 45	100	30 – 70	0.3 – 0.65	100	60 – 90	70 – 90
	Spectra 1000...	hd	0.97	3 – 6	100	270 – 370	2.6 – 3.6	100		≈ 50
Polypropylene (PP)	Herculon, Meraklon...	st/fi	0.90 – 0.92	15 – 50	100	25 – 60	0.22 – 0.55	100	85 – 95	70 – 90
Polyacrylonitrile (PAN)	Dralon T,...	st/fi	1.17 – 1.19	25 – 40	≈ 100	35 – 58	0.42 – 0.69	80 – 100	≈ 60	≈ 70
Copolymerizates with PAN ≥ 85 %	Dolan, Dralon,	st	1.17 – 1.19	20 – 60	100 – 120	18 – 32	0.2 – 0.37	75 – 95	30 – 80	75 – 80
	Euroacril, Leacril,...	fi	1.17 – 1.19	15 – 40	100 – 120	35 – 45	0.42 – 0.53	80 – 100	30 – 80	≈ 70
Modacrylics (50 – 84 % PAN)	Kanekalon, SEF,...	st	1.3 – 1.4	25 – 50	100 – 110	15 – 25	0.2 – 0.35	80 – 100	50 – 70	≈ 80
Poly(vinyl chloride) (PVC)										
atactic	Clevil, Rhovil,....	st/fi	1.38 – 1.40	10 – 40	100	20 – 24	0.28 – 0.34	100	35 – 70	
syndiotactic [61]		st	1.40	35 – 60	100	25 – 30	0.35 – 0.42	100	60 – 90	65 – 85
Poly(vinyl alcohol) (PVA)	Kuralon, Mewlon,	st	1.26 – 1.31	13 – 26	120 – 140	20 – 55	0.25 – 0.7	65 – 85	35 – 40	55 – 65
	Solvron, Vilon,...	fi	1.26 – 1.31	9 – 22	120 – 140	55 – 77	0.7 – 1	65 – 85	35 – 40	55 – 65
(HM: $\bar{M}_r > 10^6$)		hd	1.31	3 – 4		150 – 230	2 – 3			
Polytetrafluoro-ethylene (PTFE)	Gore-Tex Fibers,...	st/fi	2.1 – 2.3	20 – 40	100	8 – 18	0.16 – 0.38	100	60 – 90	75 – 90
Poly(phenylene sulfide) (PPS)	Procon, Rylon	st/fi	1.37 – 1.38	25 – 40		28 – 47	0.4 – 0.65		78	60 – 80
Melamine resin fiber	Basofil		1.4 – 1.5	15 – 25		15 – 40	0.2 – 0.6			

Table 4. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Density (ϱ), g/cm^3	Elongation ε_H (21 °C) at 65 % R.H. [b], %	Elongation ε_H (21 °C) "wet/dry," %	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H/ϱ, cN/tex	Tenacity (σ_H) at 21 °C at 65 % R.H. σ_H GPa	Tenacity (σ_H) at 21 °C Wet/dry (relative), %	Relative loop tenacity [62], %	Relative knot tenacity [63], %
Inorganic fibers		Fiber diameter, µm						Compressive strength, axial, GPa		
Textile glass	Gevetex, Fiberglas		2.45 – 2.6	2 – 5	100	70 – 120	1.7 – 2.7		15 – 30	
E glass	E-Fiber, Enkafort	5 – 15	2.52 – 2.54	2 – 3.5	100	80 – 140	2 – 3.5	0.5		
Aluminum silicate (e.g., $3\,Al_2O_3 \cdot 2\,SiO_2 \cdot B_2O_3$)	Nextel 312	17	2.63 – 2.7			65	1.7			
	Nextel 440		3.1							
	Nextel 480		3.05	1.1		75	2.3			
β-Silicon carbide	Nicalon (Nippon Carbon)	5 – 20	2.6 – 3.1	1		≈ 100	3			
	FP (DuPont)	20		0.4			1.4 – 1.5			
	Saffil, Saffimax (ICI)	3 – 3.2	2.8 – 3.3	0.7		30 – 60	1 – 2			
Silica (SiO_2)	Silica (Akzo)	9 – 10	1.8 – 2.0	1.4 – 2.0		15 – 40	0.25 – 0.8			
Steel (20 % Cr, 7 % Ni)	Bekinox	fi	7.9	1 – 2	100	22 – 29	1.75 – 2.25		65 – 75	60 – 70
Carbon (pitch based)										
High-modulus types (HM)	Tornel, Grafil,	6 – 9	2.0 – 2.06	0.4 (-2)		≈ 100	2 (-4)	0.4 – 0.8	0	
High-strength types (HS)	Fiber G	9 – 12	2.02	0.5 – 2		≈ 250	≈ 5	0.5 – 1.1	0	
Carbon (PAN based)										
High-modulus types (HM)	Magnamite, Pyrofil,	6 – 7	1.8 – 1.96	0.4 – 1.2		100 – 180	1.9 – 3.5	1.4 – 1.5	0	
High-strength types (HS)	Torayca,...	7 – 9	1.74 – 1.82	1.5 – 2.4		170 – 380	3 – 7	1.44 – 4	0	

[a] st = staple fiber, fi = filament, hd = highly drawn, an = annealed, ap = antipilling; [b] R.H. = relative humidity; [c] effective breaking strength

Table 5. Elastomechanical properties of fibers

Fiber name/polymer	Trade names	Fiber type [a]	Elastic recovery, %: $(1-\varepsilon_r/\varepsilon)$ 100 for $\varepsilon = 2\,\%$	$(5\,\%)$	Initial Modulus (elongation $\varepsilon \to 0$) E/ϱ, N/tex	E, GPa	Torsion modulus [64] G/ϱ, N/tex	G, GPa	Torsion brittleness [66], degrees
Natural fibers									
Cotton		st	75	45	3 – 6	4.5 – 9	1.6	2.4	53 – 56
Wool		st	95 – 99	60 – 70	1.5 – 3	2 – 4	0.8 – 1	≈ 1	48 – 52
Silk		fi	95	70	7 – 10	9 – 12.5	1.5 – 2.3	1.9 – 2.9	51
Hemp		st			60	90			
Flax		st	≈ 60		8 – 25	12 – 36	0.95	1.3 – 1.4	68
Jute		st			21 – 22	30 – 31			
Regenerate cellulose									
Viscose rayon, Viscose Modal	Evlan, Fibro, Sarille,..	st/fi	70 – 100	40 – 60	2 – 3 (6.5 – 7.5)	3 – 4.5 (4.2 – 4.9)	0.5 – 1	≈ 1	51 – 55
Cuprammonium rayon	Asahi Bemberg, Bemberg	st/fi	70 – 95	40 – 60	2 – 3	3 – 4.5			48 – 57
Cellulose acetate	Arnel, Celco, Dicel,	st/fi	90 – 95	(40 – 60)	2 – 3.5	2.5 – 4.5	0.6 – 0.8	≈ 1	44 – 50
Cellulose triacetate	Estron, Silene, Tricel,...	st/fi	90 – 95	55 – 70	2 – 3.5	2.5 – 4.5	0.6 – 0.8	≈ 1	44 – 50
NMMO Fiber	Lyocell, Tencel,..	st/fi			4	6			
Protein fiber		st		26	2	2 – 3	0.9	≈ 1	30 – 40
Alginate fiber (Ca-Alginat)		fi			7	12 – 13	1.2	≈ 2	
Polycondensate fibers									
	Perlon, Acelon,	st							27 – 41
Nylon 6 (PA 6)	Amilan, Anso,	fi	95 – 100	95 – 100	0.5 – 3	0.6 – 3.5	≈ 0.3	≈ 0.3	33 – 42
	Caprolan, Grilon,...	hd	95 – 100	90 – 95	4 – 5	4.5 – 5.5	0.7	0.8	
	Nylon, Antron,	st							27 – 41
Nylon 66 (PA 66)	Cantrece, Meryl,	fi	95 – 100	95 – 100	0.5 – 3	0.6 – 3.5	0.45	0.5	40 – 41
	Timbrelle, Ultron,...	hd	95 – 100	90 – 95	4 – 5	4.5 – 6			

Table 5. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Elastic recovery, %: $(1-\varepsilon_r/\varepsilon)$ 100 for $\varepsilon = 2\,\%$ (5 %)		Initial Modulus (elongation $\varepsilon \to 0$) E/ϱ, N/tex	E, GPa	Torsion modulus [64] G/ϱ, N/tex	G, GPa	Torsion brittleness [66], degrees
Aramids									
Poly(*m*-phenyleneisophthalamide)	Nomex, Conex,	fi			8 – 14	11 – 20			≈ 50
	Kevlar 29, Twaron,	fi	95 – 100		41 – 53	59 – 77			
Poly(*p*-phenyleneterephthalamide) (PPTA)	Kevlar 49, Twaron,	fi/an	95 – 100		85	124			
	Kevlar 149,	fi/an	96		103	150			
	Kevlar 981, Twaron,	fi/hd			70 – 83	102 – 120			
	Kevlar Hp	fi			50	72			
Poly(ethylene terephthalate) (PET)	Dacron, Diolen,	st			2.5 – 4	3.4 – 5.5			30 – 49
	Terylene, Trevira,	ap			2.5 – 4	3.4 – 5.5			≈ 48
	Fortrel, Grilene,	fi	90 – 98	70 – 90	7 – 18	10 – 25	0.65	0.9	41 – 48
	Serene,...	hd			7 – 18	10 – 25	1.1	1.5	47 – 48
Poly(butylene terephtalate) (PBT)	Trevira 810...813	fi	100	99 – 100	4 – 5	5 – 6			
Poly(1,4-dimethylene cyclohexane terephthalate) (PDCT) [59]	Kodel	st	85 – 95	50 – 60	2.5 – 4	3 – 5			49
Polyarylate	Vectran, Ekonol [61]	fi			60 – 128	80 – 180			
Poly(ether ether ketone) (PEEK)	Zyex	fi			4 – 12	5 – 15			
Polyimide (PI)	P84	st/fi			3.5 – 5	5 – 7			
Novolak	Kynol, Philene, Novoloid	st			2.6 – 3.6	3.3 – 4.6			
Poly(amide imide) (PAI)	Kermel	st			3.5 – 7.5				
Poly(ether imide) (PEI)		st			2.7 – 4.2	3.5 – 5.1			
Polyurethane elastomers (spandex)	Dorlastan, Lycra,...	fi	93 – 98 for $\varepsilon = 300\,\%$		≈ 0.005 – 0.01	0.006 – 0.012	0.004	0.005	

Table 5. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Elastic recovery, %: $(1-\varepsilon_r/\varepsilon)$ 100 for $\varepsilon = 2\,\%$	$\varepsilon = 5\,\%$	Initial Modulus (elongation $\varepsilon \to 0$) E/ϱ, N/tex	E, GPa	Torsion modulus [64] G/ϱ, N/tex	G, GPa	Torsion brittleness [66], degrees
Polymerizate fibers									
Polyethylene (HD – PE)	Hiralon, Vegon,	fi	95 – 100	90 – 95	0.2 – 5	0.2 – 5	0.05	0.05	
(HM – HD – PE: $\bar{M}_r > 10^6$)	Spectra 1000,...	hd	100		51 – 165	50 – 160	1.5 [65]	1.5 [65]	
Polypropylene (PP)	Herculon, Meraklon,...	st/fi	95 – 100	85 – 95	0.5 – 5	0.5 – 5			
Polyacrylonitrile (PAN)	Dralon T,...	st/fi	90 – 95	50 – 90	9.5 – 14	11 – 17	1.5 – 1.7	1.8 – 2	
Copolymerizates with PAN $\geq 85\,\%$	Dolan,Dralon	st	90 – 95	50 – 90	3 – 5	3.5 – 6			45 – 52
	Euroacril, Leacril,...	fi	90 – 95	50 – 90	9 – 10	10 – 12	1	1.2	55 – 60
Modacrylics (50 – 84 % PAN)	Kanekalon, SEF,..	st	95 – 99	85 – 98	1 – 2.5	1.3 – 3.5			
Poly(vinyl chloride) (PVC)									
atactic	Clevyl, Rhovyl,...	st/fi	70 – 90	55 – 65	2 – 4	3 – 5.5	0.6 – 0.7	0.8 – 1	27 – 50
syndiotactic [61]		st	90 – 95	75 – 80	2.5 – 3	3.5 – 4			
Poly(vinyl alcohol) (PVA)	Kuralon, Mewlon	st		60 – 80 40 – 60	3 – 4	4 – 5	0.9	1.2	
	Solvron, Vilon,...	fi		60 – 80 40 – 60	3 – 4.5	4 – 6	1.2 – 1.5	1.6 – 2	
(HM: $\bar{M}_r > 10^6$)		hd			30 – 60	40 – 80			
Polytetrafluoroethylene (PTFE)	Gore-Tex Fibers	st/fi			0.35 – 2	0.7 – 4	0.2 – 0.3	0.4 – 0.7	
Poly(pheneylene sulfide) (PPS)	Procon, Ryton,...	st/fi			2.7 – 5	3.7 – 6.8			

Table 5. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Elastic recovery, %: $(1-\varepsilon_r/\varepsilon)$ 100 for $\varepsilon = 2\,\%$ (5 %)	Initial Modulus (elongation $\varepsilon \to 0$) E/ϱ, N/tex	E, GPa	Torsion modulus [64] G/ϱ, N/tex	G, GPa	Torsion brittleness [66], degrees
Inorganic fibers		Fiber diameter, µm						
Textile glass	Gevetex, Fiberglas,...		100	28 – 34	70 – 90	$\lesssim 16$	$\lesssim 40$	85 – 88
E glass	E-Fiber, Enkafort,...	5 – 15	100	28 – 34	70 – 90	$\lesssim 16$	$\lesssim 40$	85 – 88
Aluminum silicate (e.g. $3\,Al_2O_3 \cdot 2\,SiO_2 \cdot B_2O_3$)	Nextel 312,	17		5.5 – 52	15 – 140			
	Nextel 440,			65	200			
	Nextel 480 [89]			73	224			
β-Silicon carbide	Nicalon (Nippon Carbon)	5 – 20		67 – 100	200 – 300			
	FP (DuPont)	20		120 – 125	350 – 385			
	Saffil, Saffimax (ICI)	3 – 3.2		33 – 83	100 – 250			
Silica (SiO_2)	Silica (Akzo)	9 – 10		6 – 30	12 – 56			
Steel (20 % Cr, 7 % Ni)	Bekinox	fi	100 for $\varepsilon = 1\,\%$	19 – 25	150 – 200	7.9 – 9.5	60 – 70	
Carbon (pitch based)								
High-modulus types (HM)	Thornel, Grafil,	6 – 9		250 – 400	500 – 800			
High-strength types (HS)	Fiber G	9 – 12		124 – 200	250 – 400			
Carbon (PAN based)	Magnamite							
High-modulus types (HM)	Pyrofil,	6 – 7		160 – 260	300 – 500			
High-strength types (HS)	Torayca,...	7 – 9		110 – 165	200 – 300	2.2	4	

[a] st = stable fiber, fi = filament, hd = highly drawn, an = annealed, ap = antipilling.

Table 6. Thermal properties of fibers

Fiber name/polymer	Trade names	Fiber type [a]	Fiber shrinkage in water of 95 °C, %		Specific heat capa-	Thermal conductivity,	Glass transition temperature,	Melting temperature	Heat resistance in air up to *t*, °C
			drawn	thermoset	city [67], $J\,g^{-1}\,K^{-1}$	$W\,m^{-1}\,K^{-1}$	°C	(decomposition temperature), °C	
Natural fibers									
Cotton		st			1.3	0.3 – 0.5		(400)	120
Wool		st			1.3 – 1.6	0.2 – 0.4	45 [90]		120
Silk		fi			1.4	0.2 – 0.4	40 – 50	(170 – 180)	120
Flax		st			1.4	0.3			
Regenerate cellulose									
Viscose rayon, Viscose, Modal	Evlan, Fibro, Sarille,...	st/fi	0.5 – 10 [b]		1.3 – 1.5	0.3 – 0.6		(175 – 190)	120
Cuprammonium rayon	Asahi Bemberg, Bemberg	st/fi	0.5 – 10 [b]		1.4			(175 – 205)	120
Cellulose acetate	Arnell, Celco, Dicel,	st/fi	5 – 20		1.3 – 1.5	0.3		255	120
Cellulose triacetate	Estron, Silene, Tricel,...	st/fi	5 – 20	0.7	1.5	0.3	170 – 180 [68]	300	130
Polycondensate fibers									
	Perlon, Acelon,	st	8 – 15	0.5 – 1.5	1.5 – 1.8		80 – 85 (dry)	215 – 220	120
Nylon 6 (PA 6)	Amilan, Anso,	fi	8 – 15	1 – 5	1.5 – 1.8	0.2 – 0.3	≈ 20 (3.7 % water)	215 – 220	120
	Caprolan, Grilon,...	hd	8 – 15		1.5 – 1.8	0.4	90 – 95 (dry)	215 – 220	120 – 150
	Nylon, Antron,	st	8 – 15	0.5 – 1.5	1.5		90 – 95 (dry)	255 – 260	120
Nylon 66 (PA 66)	Cantrece, Meryl,	fi	8 – 15	1 – 5	1.5	0.2 – 0.3	≈ 30 (3.7 % water)	255 – 260	120
	Timbrelle, Ultron,...	hd	8 – 15		1.5	0.4		255 – 260	120 – 150

Table 6. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Fiber shrinkage in water of 95 °C, %		Specific heat capa-	Thermal conductivity,	Glass transition temperature,	Melting temperature	Heat resistance in air up to *t*, °C
			drawn	thermoset	city [67], $J\,g^{-1}\,K^{-1}$	$W\,m^{-1}\,K^{-1}$	°C	(decomposition temperature), °C	
Aramids									
Poly(*m*-phenyleneisophthalamide)	Nomex, Conex	fi		1.5	1.2	0.13	280–290	(≈ 370)	180–200
	Kevlar 29, Twaron,	fi			1.4	0.05	340–360	(≈ 550)	180–200
Poly(*p*-phenyleneterephthalamide) (PPTA)	Kevlar 49, Kevlar 149,Twaron,	fi/an				0.05	340–360	(≈ 550)	180–200
	Kevlar 981, Twaron,	fi/hd					340–360		
	Kevlar Hp	fi					340–360		
	Dacron, Diolen,	st		0.5–1			80–110	250–260	120–150
Poly(ethylene terephthalate) (PET)	Fortrel, Grilene,	ap		0.5–1			80–110	250–260	120–150
	Serene, Terylene,	fi	5–10	0.5–1.5	0.4–1.9 [70]	0.2–0.3	80–110	250–260	120–150
	Trevira, . . .	hd	7–8				80–110	250–260	150–160
Poly(butylene terephtalate) (PBT)	Trevira 810, Trevira 813	st/fi					68–75	224	120
Poly(1,4-dimethylene cyclohexane terephthalate) (PDCT) [59]	Kodel	st		0.1–0.5			≈ 100	285–290	≈ 200
Polyarylate	Vectran, Ekonol [60]	fi						320–290	180–260
Poly(ether ether ketone) (PEEK)	Zyex	fi		0.4–0.5			139–153	334–345	200–250
Polyimide (PI)	P84	st/fi					315		260
Novolak	Kynol, Philene, Novoloid	st		0					
Poly(amide imide) (PAI)	Kermel	st						(> 380)	250
Poly(ether imide) (PEI)		st/fi	4–8	0.7			215–225		160–170
Polyurethane elastomers (spandex)	Dorlastan, Lycra,...	fi	≈ 3–12			0.15	Polyether: ≈ − 60 [70] ; Polyester − 40 to − 20	230–290	120

Table 6. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Fiber shrinkage in water of 95 °C, %		Specific heat capa-	Thermal conductivity,	Glass transition temperature,	Melting temperature	Heat resistance in air up to *t*, °C
			drawn	thermoset	city [67], $J g^{-1} K^{-1}$	$W m^{-1} K^{-1}$	°C	(decomposition temperature), °C	
Polymerizate fibers									
Polyethylene (HD-PE)	Hiralon, Vegon,	fi	5 – 10		1.4 – 2.0	0.2 – 0.4	− 20 to − 30	124 – 138	70 – 90
(HM – HD – PE: $\bar{M}_r > 10^6$)	Spectra 1000	hd				10 – 38 [84], [85]	− 20 to − 30	124 – 138	
Polypropylene (PP)	Herculon, Meraklon	st/fi		0 – 5	1.6	0.2 – 0.3		160 – 175	≈ 120
Polyacrylonitrile (PAN)	Dralon T	fi/st	14 – 16	≈ 1		0.2	≈ 95 (dry)	≈ 320	140
Copolymerizates with PAN ≥ 85 %	Dolan, Dralon	st	(20 – 40 [c])	0.5 – 5	1.2 – 2.5	0.2	85 – 95 (dry); 50 – 60 (wet)	(>250)	140
	Euroacril, Leacril,	fi	16 – 22	≈ 1	1.2 – 1.5	0.2	85 – 95 (dry); 50 – 60 (wet)	(>250)	140
Modacrylics (50 – 84 % PAN)	Kanekalon, SEF,...	st		0.2 – 5			85 – 95	(≳ 170)	≈ 120
Poly(vinyl chloride) (PVC)									
atactic	Clevyl, Rhovyl,	st/fi	20 – 30		0.8 – 0.9	0.16 – 0.17	70 – 90	(160 – 200)	< 65
syndiotactic [61]		st		0			90 – 100	(160 – 200)	
Poly(vinyl alcohol) (PVA)	Kuralon, Mewlon,	st		2 – 3			75 – 90		140
	Solvron, Vilon,	fi					75 – 130	(240 – 260)	(water vapor: 120)
Polytetrafluoroethylene (PTFE)	Gore-Tex Fibers	st/fi		2	1.0 – 1.1	0.23	127 – 130	327 – 342 [72]	180
Poly(phenylene sulfide) (PPS)	Procon, Ryton,...	st/fi	≳ 16				85 – 95	270 – 290	190 – 200
Melamine resin fiber	Basofil	st							200 – 220

[a] st = stable fiber, fi = filament, hd = highly drawn, an = annealed, ap = antipilling;
[b] Depending on fiber type;
[c] High-shrinkage (HS) fiber.

Table 7. Thermal and electrical properties of inorganic fibers

Fiber name	Trade name	Fiber diameter, µm	Specific heat capacity [67], $Jg^{-1}K^{-1}$	Thermal conductivity, $Wm^{-}K^{-1}$	Heat resistance in air up to , °C	Fire Limiting Oxygen Index (LOI), %	Melting temperature (Glass transition T_g), °C	Specific electrical resistance, Ω cm
Textile glass	Gevetex, Fiberglas	5 – 15	≈ 0.75	≈ 0.8	300 – 400		$(600 \lesssim T_g \lesssim 700)$	$10^{12} - 10^{15}$
E Glass	E-Fiber, Enkafort	5 – 15	≈ 0.75	≈ 0.8	300-400		$(600 \lesssim T_g \lesssim 700)$	$10^{12} - 10^{15}$
Aluminum silicate	Nextel 312	17			1200 – 1300		1800	
(e.g., $3\,Al_2O_3 \cdot 2\,SiO_2 \cdot B_2O_3$)	Nextel 440				1200 – 1300		1800	
	Nextel 480				1200			
	Nicalon (Nippon Carbon)	5 – 20		12	> 800		2700	
β-Silicon carbide								
	FP (DuPont)	20			900			
	Saffil, Saffimax (ICI)	3 – 3.2			1000			
Silica (SiO_2)	Silica (Akzo)	9-10			1000 – 1100		1750	
Steel (20 % Cr, 7 % Ni)	Bekinox	filaments	0.46	15			1400 – 1450	
Carbon (pitch based)								0.7×10^{-4}
High-modulus types (HM)	Thornel, Grafil,	6 – 9			300 (-500)			
High-strength types (HS)	Fiber G,...	9 – 12			300 (-500)			$> 10^{-3}$
Carbon (PAN based)								
High-modulus types (HM)	Magnamite, Pyrofil,	6 – 7		60 – 115 [73]	300 (-500)	> 60	3600	$\approx 10^{-3}$
High-strength types (HS)	Torayca,...	7 – 9	0.7	15 – 20 [73]	300 (-500)	> 60	$\approx 1.5 \times 10^{-3}$	

Table 8. Electrical resistance, ironing temperature, water absorption and solubility of fibers

Fiber name/polymer	Trade names	Fiber type [a]	Specific electrical resistance, [b] Ω cm	Ironing temperature, °C	Water absorption at 21 °C, 65 % R.H., [d] %	Water retention [74], %	Solubility in selected solvents [c]
Natural fibers							
Cotton		st	$10^6 - 10^8$	180 – 220	7 – 11	40 – 50	concentrated H_2SO_4, Cuoxam-solution
Wool		st	$10^8 - 10^{11}$	160 – 170	15 – 17	40 – 45	conc. inorg. acids, conc. KOH
Silk		fi	$10^9 - 10^{10}$	140 – 165	9 – 11	40 – 45	conc. inorg. acids, conc. KOH, conc. HCO_2H
Flax		st		215 – 240	8 – 10	50 – 55	concentrated H_2SO_4
Regenerate cellulose							
Viscose rayon	Evlan, Fibro, Sarille,...	st/fi	$10^6 - 10^7$	150 – 180	12 – 14	85 – 120	conc. inorg. acids
Cuprammonium rayon	Asahi Bemberg, Bemberg,...	st/fi	$\approx 10^7 - 10^8$	150 – 180	11 – 12	100 – 125	conc. inorg. acids
Cellulose acetate	Arnel, Celco, Dicel	st/fi	$10^9 - 10^{12}$	180	6 – 7	20 – 28	conc. inorg. acids, acetone, dioxane, phenols
Cellulose triacetate	Estron, Silene, Tricel,...	st/fi	$\approx 10^{14}$	220 – 250	2 – 5	10 – 18	conc. inorg. acids, acetone, dioxane, phenols
NMMO Fiber	Lyocell, Tencel,...	st/fi		150		65 – 70	*N*-methylmorpholine oxide (NMMO)
Polycondensate fibers							
	Perlon, Acelon,	st		150	3.5 – 4.5	10 – 15	conc. inorg. acids, phenols
Nylon 6 (PA 6)	Amilan, Anso,	fi	$10^9 - 10^{11}$	150	3.5 – 4.5	10 – 15	conc. inorg. acids, phenols
	Caprolan, Grilon,...	hd		150	3.5 – 4.5	9 – 11	conc. inorg. acids, phenols
	Nylon, Antron,	st		180 – 200	3.5 – 4.5	10 – 15	conc. inorg. acids, phenols
Nylon 66 (PA 66)	Cantrece, Meryl,	fi	$10^9 - 10^{11}$	180 – 200	3.5 – 4.5	10 – 15	conc. inorg. acids, phenols
	Timbrelle, Ultron,...	hd		180 – 200	3 – 4	9 – 11	conc. inorg. acids, phenols
Aramids							
Poly(*m*-phenyleneisophthalamide)	Nomex, Conex	fi			4.5 – 5	12 – 17	polar org. solvents and solutes (LiCl, $CaCl_2$), conc. H_2SO_4
	Kevlar 29, Twaron,	fi			2 – 3	7	conc. H_2SO_4
Poly(*p*-phenyleneterephthalamide) (PPTA)	Kevlar 49, Kevlar 149, Twaron	fi/an	10^{15}		2 – 3	7	conc. H_2SO_4
	Kevlar 981, Twaron	fi/hd			2 – 3	7	conc. H_2SO_4
	Kevlar Hp	fi			2 – 3	7	conc. H_2SO_4
	Dacron, Diolen,	st		150 – 200	0.3 – 0.4	3 – 5	conc. H_2SO_4, conc. KOH, phenols, tetrachloroethane
Poly(ethylene terephthalate) (PET)	Fortrel, Grilene,	fi	$10^{11} - 10^{14}$	150 – 200	0.2 – 0.5	3 – 5	conc. H_2SO_4, conc. KOH, phenols, tetrachloroethane
	Terylene, Trevira,...	hd		150 – 200	0.2 – 0.4	3 – 5	conc. H_2SO_4, conc. KOH, phenols, tetrachloroethane
Poly(1,4-dimethylene cyclohexane terephthalate) (PDCT) [59]	Kodel	st		≈ 200	0.2		conc. H_2SO_4, conc. KOH, phenols, tetrachloroethane

Table 8. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Specific electrical resistance, [b] Ω cm	Ironing temperature, °C	Water absorption at 21 °C, 65 % R.H., [d] %	Water retention [74], %	Solubility in selected solvents [c]
Polyarylate	Ekonol [60], Vectran	fi			0		NMP + $CaCl_2$
Poly(ether ether ketone) (PEEK)	Zyex	fi					conc. H_2SO_4
Novolak	Kynol, Philene, Novoloid,...				6 – 8		
Poly(amide imide) (PAI)	Kermel				3		polar org. solvents (DMA, NMP)
Poly(ether imide) (PEI)		st/fi			0.25 – 1.25		dichloromethane
Polyurethane elastomers (spandex)	Dorlastan, Lycra, ...	fi		150 – 180	0.5 – 1.5	7 – 11	noncross-linked fibers: DMA, DMF, HMPA, conc. inorg. acids (decomp.)
Polymerizate fibers							
Polyethylene (HD-PE)	Hiralon, Vegon, Spectra,...	fi	$10^{13} - 10^{17}$		0	0	conc. H_2SO_4, benzene, chlorinated hydrocarbons
Polypropylene (PP)	Herculon, Meraklon,...	st/fi	$> 10^{13}$	130	0	0	conc. H_2SO_4, toluene, chlorinated hydrocarbons
Polyacrylonitrile (PAN)	e.g., Dralon T	fi/st	$\approx 10^{14}$	150 – 180	≈ 1	4 – 6	conc. inorg. acids, DMA, DMF, DMS, ethylene carbonate, conc. $ZnCl_2$ or NaSCN solutions
Copolymerizates with PAN $\geq$ 85 %	Dolan, Dralon,...	st/fi	$10^{8} - 10^{14}$	150 – 180	1 – 1.5	5 – 12	conc. inorg. acids, DMA, DMF, DMS, ethylene carbonate, conc. $ZnCl_2$ or NaSCN solutions
Modacrylics (50 – 84 % PAN)	Kanekalon, SEF,...	st	$10^{12} - 10^{13}$		0.4 – 3	10 – 20	conc. H_2SO_4, DMF, acetone, phenol, cyclohexanone
Poly(vinyl chloride) (PVC)							
atactic	Clevyl, Rhovyl	st/fi	$10^{12} - 10^{14}$		0 – 0.2	4 – 6	conc. H_2SO_4, chlorinated hydrocarbons, dioxane, cyclohexanone, DMF
syndiotactic [61]		st	$10^{12} - 10^{14}$		0	4 – 6	conc. H_2SO_4, chlorinated hydrocarbons, dioxane, cyclohexanone, DMF

Table 8. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Specific electrical resistance, [b] Ω cm	Ironing temperature, °C	Water absorption at 21 °C, 65 % R.H., [d] %	Water retention [74], %	Solubility in selected solvents [c]
Poly(vinyl alcohol) (PVA)	Kuralon, Mewlon,...	st/fi			3.5 – 5	25 – 35	conc. inorg. acids, phenols, DMF
Poly(tetrafluoroethylene) (PTFE)	Gore-Tex Fibers,...	st/fi	$> 10^{14}$		0		perfluorinated solvents > 300 °C
Poly(phenylene sulfide) (PPS)	Procon, Ryton,...	st/fi			≈ 0.03 – 0.25 (< 0.6)		< 200 °C no solvent
Melamine resin fiber	Basofil	st			9		
Inorganic fibers							
Textile glass	Gevetex, Fiberglas	st/fi	$10^{12} - 10^{15}$		0.1	0	hydrofluoric acid
E glass	E-Fiber, Enkafort	fi	$10^{12} - 10^{15}$		0.1	0	hydrofluoric acid
Aluminum silicate	Nextel 312	st			≈ 0	0	hydrofluoric acid
(e.g., $3\,Al_2O_3 \cdot 2\,SiO_2 \cdot B_2O_3$)	Nextel 440	st			≈ 0		hydrofluoric acid
Steel (20 % Cr, 7 % Ni)	Bekinox	fi	0.7×10^{-4}		0		(attacked by halogenated solvents)

[a] st = staple fiber, fi = filament, hd = highly drawn, an = annealed, ap = antipilling;
[b] of fibers without additives or special finishes;
[c] DMA = dimethylacetamide, DMF = dimethylformamide, DMS = dimethyl sulfoxide, HMPA = hexamethylphosphoramide, NMP = *N*-methyl pyrrolidone;
[d] R.H. = relative humidity.

Table 9. Resistance of fibers [c]

Fiber name/polymer	Trade names	Fiber type [a]	Heat (air)	Light/weather [b]	Biological influence [c]		Chemicals[c]		Fire
			up to t, °C	Residual	Micro-	Insects	Residual tenacity, %		limiting
				tenacity, %	organisms [57], [58], [75], [76]		After 1000 h at 20 °C/10 h at 100 °C dilute acid	dilute alkali	oxygen index (LOI), %
Natural fibers									
Cotton		st	120	20 – 30/0 – 20	–	(+)	60 – 80/0 – 20	80 – 100/80 – 100	19 – 20
Wool		st	120	0 – 20/0 – 20	–	–	90 – 100/selective	selective	25 – 28
Silk		fi	120	0 – 20/0 – 20	(+)	(+)	lower than wool	selective	
Flax		st		0 – 20/0 – 20	bleached (+)		more resistant than cotton	lower than cotton	
Regenerate cellulose									
Viscose rayon, Viscose Modal	Evlan, Fibro, Sarille,...	st/fi	120	0 – 30/0	–	–	similar to cotton		19 – 20
Cuprammonium rayon	Asahi Bemberg, Bemberg	st/fi	120	0 – 30/0	–	–	similar to cotton		19 – 20
Cellulose acetate (CA)	Arnell, Celco, Dicel,	st/fi	120	20 – 45/0 – 25	(+)	–	60 – 100/selective	0 – 20/20 – 60 NaOH	18 – 19
Cellulose triacetate	Estron, Silene, Tricel,...	st/fi	130	20 – 45/0 – 25	(+)	–	60 – 100/selective	better than CA	18 – 19
NMMO Fiber	Lyocell, Tencel	st/fi	150		similar to cotton		similar to cotton		similar to cotton
Polycondensate fibers									
	Perlon, Acelon,	st	120	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5
Nylon 6 (PA 6)	Amilan, Anso,	fi	120	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5
	Caprolan, Grilon,...	hd	120 – 150	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5
	Nylon, Antron,	st	120	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5
Nylon 66 (PA 66)	Cantrece, Meryl,	fi	120	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5
	Timbrelle, Ultron,...	hd	120 – 150	20 – 30/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 21.5

Table 9. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Heat (air) up to t, °C	Light/weather [b] Residual tenacity, %	Biological influence [c] Micro-organisms [57], [58], [75], [76]	Insects	Chemicals[c] Residual tenacity, % After 1000 h at 20 °C/10 h at 100 °C dilute acid	dilute alkali	Fire limiting oxygen index (LOI), %
Aramids									
Poly(*m*-phenyleneisophthalamide)	Nomex, Conex	fi	180 – 200	/50		+	80 – 100/80 – 100	90 – 100/90 – 100	26 – 30
Poly(*p*-phenyleneterephthalamide) (PPTA)	Kevlar 29, Twaron,	fi	180 – 200	65 – 80/ after 16 weeks	(+)	+	+	+	
	Kevlar 49, Twaron, Kevlar 149,	fi/an	180 – 200	65 – 80/ after 16 weeks	(+)	+	+	+	29 – 31
	Kevlar 981, Twaron,	fi/hd	180 – 200	65 – 80/ after 16 weeks	(+)	+	+	+	29 – 31
	Kevlar Hp	fi	180 – 200	65 – 80/ after 16 weeks	(+)	+	+	+	29 – 31
Poly(ethylene terephthalate) (PET)	Dacron, Diolen,	st	120 – 150	60 – 80/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 22
	Fortrel, Grilene,	ap	120 – 150	60 – 80/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 22
	Terylene, Serene,	fi	120 – 150	60 – 80/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 22
	Trevira,...	hd	150 – 160	60 – 80/5 – 15	(+)	+	90 – 100/90 – 100	90 – 100/90 – 100	20 – 22
Poly(butylene terephtalate) (PBT)	Trewira 810,...813	st/fi	120						
Poly(1,4-dimethylene cyclohexane terephthalate) (PDCT) [59]	Kodel	st	≈ 200			+	70 – 80 HCl	50 – 70 Na_2CO_3 solution	
Polyarylate	Vectran, Ekonol [60]	fi	180 – 260				+/96	-	36
Poly(ether ether ketone) (PEEK)	Zyex	fi	200 – 250	-			99/23-74 selective	/95	30 – 35
Polyimide (PI)	P 84	st/fi	260				$\gtrsim$ 80	(+)	36 – 38
Novolak	Kynol, Philene, Novoloid	st	200				+	-	30 – 39

Table 9. (Continued)

Fiber name/polymer	Trade names	Fiber type [a]	Heat (air) up to t, °C	Light/weather [b] Residual tenacity, %	Biological influence [c] Micro-organisms [57], [58], [75], [76]	Insects	Chemicals [c] Residual tenacity, % After 1000 h at 20 °C/10 h at 100 °C dilute acid	dilute alkali	Fire limiting oxygen index (LOI), %
Poly(amide imide) (PAI)	Kermel	st	250	-			+	(+)	31 – 32
Poly(ether imide) (PEI)		st/fi	160 – 170	+			95 – 100/	85 – 100/	33
Polyurethane elastomers (spandex)	Dorlastan, Lycra,...	fi	120	0/0	+	+	selective; polyether more resistant than polyester		
Polymerizate fibers									
Polyethylene (HD)-PE	Hiralon, Vegon,,...	fi	70 – 90	+	–	(+)	+	+	
(HM – HD – PE: $\bar{M}_r > 10^6$)	Spectra 1000	hd	80 – 90	+	–	(+)	+	+	
Polypropylene (PP)	Herculon, Meraklon,...	st/fi	≈ 120	0/0	–	(+)	+	+	19 – 20
Polyacrylonitrile (PAN)	e.g., Dralon T	fi/st	140	60 – 80/50 – 60	(+)	+	90 – 100/80 – 100	90 – 100/60 – 100	18 – 20
Copolymerizates with PAN ≥ 85 %	Dolan, Dralon, Euroacril, Leacril,...	st/fi	140	60 – 80/50 – 60	(+)	+	90 – 100/80 – 100	90 – 100/60 – 100	18 – 20
Modacrylics (50 – 84 % PAN)	Kanekalon, SEF,...	st	120	lower than PAN	+	+	+	+	25 – 30
Poly(vinyl chloride)									
atactic	Clevyl, Rhovyl,...	st/fi	< 65	60 – 90/	+	+	+	+	37 – 46
syndiotactic [61]		st			+	+	+	+	37 – 46
Poly(vinyl alcohol) (PVA)	Kuralon, Mewlon,	st	140 (water vapor: 120)				better than nylon		20
	Solvron, Vilon,...	fi	140 (water vapor: 120)	+/(+)	(+)	(+)	better than nylon	+	20
(HM: $\bar{M}_r > 10^6$)		hd	140 (water vapor: 120)				better than nylon	+	20
Poly(tetrafluoroethylene) (PTFE)	Gore-Tex Fibers,...	st/fi	180		+	+	+	+	≳ 40
Poly(phenylene sulfide) (PPS)	Procon, Ryton,...	st/fi	190 – 200	-	+	+	+/100	+/100	34 – 35
Melamine resin fiber	Basofil	st	200 – 220				0	≈ 70/	30

[a] st = staple fiber, fi = filament, hd = highly drawn, an = annealed ap = antipilling;
[b] no UV stabilizers added, one-year exposure in Florida;
[c] + resistant, (+) moderately resistant, – not resistant.

Table 10. Dyeing behavior of fibers

Fiber name/polymer	Trade name	Dyes
Natural fibers		
Cotton		substantive
Wool		anionic
Silk		anionic, cationic, substantive, reactive, vat
Regenerate cellulose		
Viscose rayon	Evlan, Fibro,...	substantive
Cuprammonium rayon	Asahi Bemberg,...	substantive
Cellulose acetate	Arnel, Celco, Dicel,	substantive, disperse
Cellulose triacetate	Estron, Silene, Tricel,...	substantive, disperse
NMMO fiber	Lyocell, Tencel,...	substantive, reactive
Polycondensate fibers		
Nylon 6	Perlon, Capron, Grilon,...	anionic, metal-complex disperse
Nylon 66	Nylon, Antron, Ultron,...	anionic, metal-complex, disperse
Poly(*m*-phenyleneisophtalamide) (aramid)	Nomex	cationic plus carrier, high-temperature conditions
Poly(ethylene therephthalate)	Dacron, Diolen,...	disperse plus carrier, high-temperature conditions
Poly(1,4–dimethylenecyclohexane)	Kodel	disperse plus carrier, high-temperature conditions
Poly(butylene terephthalate)	Trevira 810,..813	disperse
Novolak	Kynol,...	disperse
Poly(ether imide)		disperse
Polyurethane elastomers	Dorlastan, Lycra	anionic
Polymerizate fibers		
Polypropylene	Herculon, Meraklon,...	disperse
Polyacrylonitrile	Dralon, Dolan,...	cationic, disperse
Modacrylics	Kanekalon, SEF,...	cationic, disperse
Poly(vinyl chloride)	Clevyl, Rhovyl,...	disperse
Poly(phenylene sulfide)	Procon, Ryton	disperse

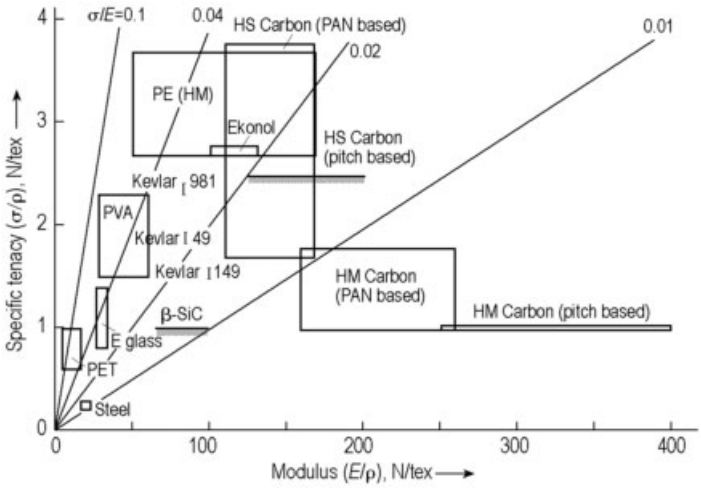

Figure 8. Specific tenacity (σ/ϱ) and modulus (E/ϱ) of high-performance fibers (ϱ = density)

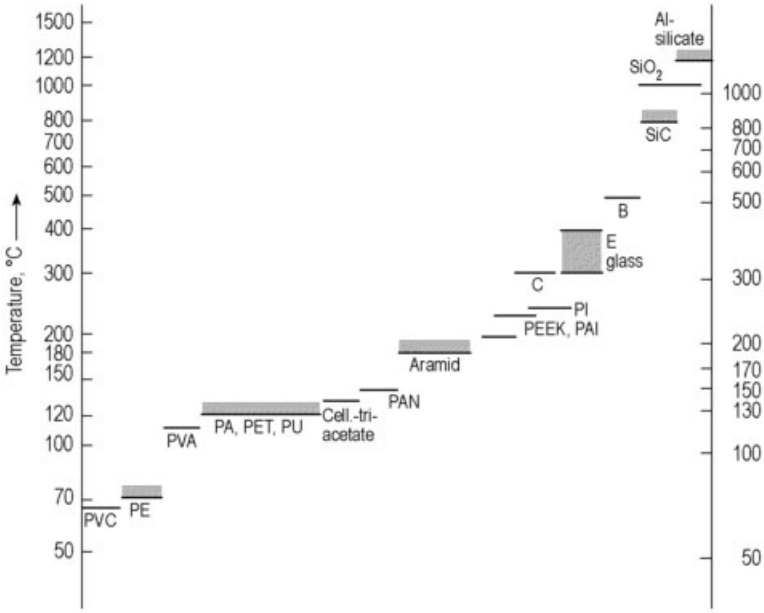

Figure 9. Highest temperatures for fiber application without significant loss in tenacity

Values for fiber elongation, fiber shrinkage, and relative properties such as wet tenacity, loop tenacity and other parameters are expressed as percent. If properties such as fiber resistance to light, weather, chemicals, or organisms cannot be expressed numerically, they are characterized in a simplified qualitative manner in Table 9. In individual cases, the respective conditions and the type of attack (e.g., by certain kinds of pests [57], [58]) are such that widely differing results can be expected.

Some promising fibers that are still being developed have also been included, in particular, high-performance fibers made from high molecular mass poly(vinyl alcohol), the thermoresistant PEEK, PAI, PI (see Fig. 9, and the polyarylate " Ekonol," an example of a thermotropic polymer [59]. Among the inorganic fibers, apart from carbon and glass, the properties of silicon carbide, and ceramic fibers have been included. Figure 8 shows the ranges of Young's moduli and tenacities of presently known high-performance fibers used mainly for reinforcing organic resins.

Although fibers are highly anisotropic compared with other materials and therefore especially strong in the axial direction their Young's modulus is always inferior to the theoretical (crystal) modulus. The ratio of fiber to crystal modulus attains about 0.1 for textile commodity fibers, 0.3 for high-performance fibers (e.g., from HM-HDPE), especially those with extended chain structure, and as high as 0.8 for the LC polymer Kevlar 149 with unfolded, oriented PPTA molecules. The deficit of the fiber modulus is a consequence of incompletely uncoiled and disentangled macromolecules in as-spun and drawn fibers. The mechanical response of fibers is controlled by the fraction of taut tie molecules connecting neighboring crystallites. This fraction can be estimated from measurements of the elastic modulus by using a simple rheological model or other methods to be about 0.05 – 0.1 for drawn fibers [83].

Compressibility and lateral compliance increase with increasing fiber anisotropy (on drawing) and this is undesirable for reinforcing fibers (e.g., Kevlar or carbon fibers) in composites. In some cases thermal aftertreatment of such high-performance fibers allows a compromise to be found between the desired compressibility and lower tenacity (see also compressive strength of fibers in Table 4).

The applicability of industrial fibers at higher temperatures (see Tables 6 and 7) is of current

interest. Figure 9 reviews the thermal long-term stability in air.

This compilation of important data is by no means complete. For further information, see → Fibers, 4. Synthetic Organic and → Fibers, 5. Synthetic Inorganic.

9. References

General References

1. H. Batzer (ed.): *Polymere Werkstoffe,* Thieme Verlag, Stuttgart 1986.
2. R. Bauer, H. J. Koslowski: *Chemiefaser-Lexikon,* 10th ed. Deutscher Fachverlag, Frankfurt 1993.
3. J. Brandrup, (ed.): *Polymer Handbook* , 2nd ed., Interscience, New York 1976.
4. R. M. Brown, (ed.): *Cellulose and Other Natural Polymer Systems,* Plenum Publishing, New York 1982.
5. M. E. Carter: *Essential Fiber Chemistry,* Dekker, New York 1971.
6. J. G. Cook: *Handbook of Textile Fibres,* Merrow, Watford 1968.
7. H. Driesch: *Welche Chemiefaser ist das,* Franckh, Stuttgart 1962.
8. A. A Dembeck: *Guidebook to Man-made Textile Fibres and Textured Yarns of the World,* 3rd ed. The United Piece Dyed Works, New York 1969.
9. H. Doehner (ed.): *Wollkunde,* Parey, Berlin 1964.
10. H. F. Mark (ed.): *Encyclopedia of Polymer Science and Technology,* "Plastics, resins, rubbers, fibres," Wiley-Interscience, New York 1967.
11. B. Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981.
12. F. Fourné (ed.): "Herstellung und Verarbeitung," *Synthetische Fasern,* Wissenschaftl. Verlags GmbH, Stuttgart 1964.
13. B. C. Gaswami, J. G. Martindale, F. L. Scardino: *Textile Yarns,* Wiley-Interscience, New York 1977.
14. M. Grayson (ed.): *Encyclopedia of Textiles, Fibers, and Nonwoven Fabrics,* Wiley-Interscience, New York 1984.
15. F. Happey: *Applied Fibre Science,* Academic Press, London 1979.
16. H. W. Haudek, E. Viti: *Textilfasern,* Bondi, Wien 1980.
17. *High polymers,* a Series of Monographs on the Chemistry, Physics and Technology of High Polymeric Substances, Interscience, New York.
18. Internat. Baumwoll-Inst. (ed.): *Handbuch der Baumwollstoffe,* Frankfurt 1983.
19. H. Klare: *Geschichte der Chemiefaserforschung,* Akademie Verlag, Berlin 1985.
20. P.-A. Koch, (ed.): *Großes Textil-Lexikon,* Deutsche Verlagsanstalt, Stuttgart 1966.
21. H. A. Krässig, J. Lenz, H. F. Mark: *Fiber Technology,* Dekker, New York 1984.
22. R. Vieweg, G. W. Becker (ed.): *Kunststoff-Handbuch,* Hanser Verlag, München 1965.
23. P. Lennox-Kerr: *The World Fibres Book,* The Textile Trade Press, Mancheser, 1972.
24. M. Lewin, S. B. Sello (ed.): *Handbook of Fiber Science and Technology,* Dekker, New York 1983.
25. W. Loy: *Chemiefaserstoffe,* Schiele und Schön, Berlin 1978.
26. J. A. Maclaren, B. Milligan: *Wool Science,* Science Press, Marrickville 1981.
27. H. F. Mark (ed.): *Man-made Fibers, Science and Technology,* Interscience, New York 1968.
28. R. Meredith: *Elastomeric Fibres,* Merrow, Watford 1971.
29. B. E. Messerli (ed.): *Seide,* Textilwerkstattverlag Hannover, 1986.
30. L. Miles: *Cotton,* Wayland, Have 1980.
31. R. Moncrieff: *Man-made-Fibres,* 4th ed., Heywood, London 1966.
32. J. S. Robinson (ed.): *Fiber-forming Polymers,* Noyes Data Corp., Park Ridge 1980.
33. J. S. Robinson (ed.): *Spinning, Extruding and Processing of Fibers,* Noyes Data Corp., Park Ridge 1980.
34. J. S. Robinson (ed.): *Manufactue of Yarns and Fabrics from Synthetic Fibers,* Noyes Data Corp., Park Ridge 1980.
35. W. J. Roff: *Fibres, Films, Plastics and Rubbers,* Butterworths, London 1971.
36. Z. A. Rogovin: *Chemiefasern,* Thieme Verlag, Stuttgart 1982.
37. H. E. Schiecke: *Wolle als textiler Rohstoff,* Schiele und Schön, Berlin 1979.
38. K. A. Schmidt: *Technologie textiler Glasfasern,* Zechner, Speyer 1964.
39. Textil-Fakten: *Markt- und Strukturdaten der Textil- und Bekleidungswirtschaft,* Deutscher Fachverlag, Frankfurt 1983.
40. *Textile Faserstoffe,* Fachbuchverlag, Leipzig 1967.
41. C. A. Tisdell: *Economics of Fibre Markets,* Univ. New Castle, New Castle, Austr. 1977.

42. E. Wagner: *Die textilen Rohstoffe,* 6th ed., Spohr, Wuppertal 1981.
43. A. Ziabicki: *Fundamentals of Fibre Formation,* Wiley Interscience, New York 1976.
44. A. Ziabicki: *High Speed Fibre Spinning,* Wiley Interscience, New York 1985.
45. F. Schultze-Gebhardt, "Survey of the Most Important Properties of High-Performance Fibers and their Technical Use," *Chemiefasern/Textilind.* **43** (1993) T 194 – 196, E 135.

Specific References

46. P.-A. Koch (ed.): *Großes Textil-Lexikon,* Deutsche Verlagsanstalt, Stuttgart 1966.
47. H. Batzer, (ed.): *Polymere Werkstoffe,* Thieme Verlag, Stuttgart 1986.
48. H. Staudinger et al., *Phys. Chem.* **126** (1927) 3.
49. G. Henrici-Olivé, S. Olivé, *Adv. Polym. Sci.* **51** (1983) 1.
50. K.-H. Umbach, *Chemiefasern/Textilind.* **33** (1983) 85, 136, 204.
51. *Chemiefasern/Textilind.* **37** (1987) 182.
52. J. D. Geerdes: *World Fibre Production, Trends and Outlok,* Internat. Conference on Man Made Fibres, Beijing, Nov. 1985.
53. P.-A. Koch: *Faserstoff-Tabellen,* Konradin-Verlag Kohlhammer GmbH, Stuttgart 1968, 1977, 1979.
54. E. Kleinhansl, J. Mavely: *Denkendorfer Fasertafeln 1986,* Textilpraxis int., Leinfelden-Echterdingen 1986.
55. F. Schultze-Gebhard: "Mechanische Eigenschaften," in B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981, p. 64.
56. DIN 53 835 (1975).
57. W. Kerner-Gang, H. Kühne in G. Schreyer (ed.): *Konstruieren mit Kunststoffen,* Hanser Verlag, München 1972.
58. H. Kühne, *TPI Text. Prax. Int.* **29** (1974) 57; **30** (1975) 598,718.
59. E. V. Martin, H. Busch, *Angew. Chem.* **74** (1962) 624.
60. Sumitomo Chem. Com., US 4 503 005, 1983 (K. Ueno, H. Sugimoto, K. Hayatsu).
61. C. Mazzolini: *The Development of a New Fibre from Syndiotactic Polyvinylchloride,* 2nd Shirley International Seminar, Manchester 1970.
62. DIN 53 843 (1976).
63. DIN 53 842 (1976).
64. R. Meredith, *J. Text. Inst., Trans.* **45** (1954) T 489. N. Adams, *J. Text. Inst., Trans.* **47** (1956) T 530. I. D. Owen, *J. Text. Inst., Trans.* **56** (1965) T 329.
65. C. L. Choy, W. P. Leung, *J. Polym. Sci. Polym. Phys. Ed.* **23** (1985) 1759 – 1780.
66. P. A. Koch, G. Feier, B. Hoffmann: "Untersuchungen über die Querspröd igkeit neuer Synthesefasern," *Forschungsber. Landes Nordrhein-Westf.* no. 2299, Westdeutscher Verlag, Opladen 1973.
67. W. Götze, F. Winkler, *Faserforsch. Textiltech.* **18** (1967) 119, 385.
68. K. E. Perepelkin, *Faserforsch. Textiltech.* **25** (1974) 251.
69. G. W. Urbanczyk, G. Michalak, *J. Appl. Polym. Sci.* **32** (1986) 3841 – 3846.
70. H. Hespe, E. Meisert, U. Eisele, L. Morbitzer, W. Goyert, *Kolloid Z. Z. Polym.* **250** (1972) 797. D. J. Hourston, R. Meredith, *J. Appl. Polym. Sci.* **17** (1973) 3259.
71. G. Hinrichsen, *Angew. Makromol. Chem.* **20** (1971) 121.
72. R. L. McGee, J. R. Collier, *Polym. Eng. Sci.* **26** (1986) 239 – 242.
73. H. Böder, D. Gölden, P. Rose, H. Würmseher: "Kohlenstoffasern – Herstellung, Eigenschaften, Verwendung," *Z. Werkstofftech.* **11** (1980) 275.
74. DIN 53 814 (1974).
75. I. A. Ermilova, L. N. Alekseeva, I. I. Šamolina, V. A. Chochlova, *Sowj. Beiträge Faserforsch. Textiltech.* **19** (1982) 274 – 276.
76. Y. L. Hsieh, D. A. Timm, J. Merry, *Textile Res. J.* **57** (1987) 20 – 28.
77. G. Wu, J.-D. Jiang, P. A. Tucker, J. A. Cuculo, *J. Polym. Sci.: B: Polym. Phys.* **34** (1996) 2035 – 2047.
78. F. Schultze-Gebhardt, *Chemiefasern/Textilind.* **43** (1993) 432 – 433, E 66, E 68.
79. *Chem. Fibers Int* **46** (1996) 230. Econom. Fiber Bureau, Washington: Fiber Organon, Washington 1996.
80. T. F. N. Johnson, *Chem. Fibers Int.* **46** (1996) 280, 282 – 284, 286.
81. S. R. Allen, *J. Mater. Sci.* **28** (1993) 853 – 859.
82. G. J. Hayes, D. D. Edle, J. M. Kennedy, *J. Mater. Sci.* **28** (1993) 3347 – 3257.
83. M. Miwa et al., *J. Mater. Sci.* **31** (1996) 499 – 506.
84. F. Schultze-Gebhardt, *Acta Polym.* **41** (1990) 512 – 513, *Chemiefasern/Textilind.* **40** (1990) T 56, T 58, E 49 – E 50.
85. B. Poulaert et al., *Polym. Commun.* **31** (1990) 48 – 51.
86. C. L. Choy, Y. Fei, T. G. Xi, *J. Appl. Polym. Sci.* **31** (1993) 365 – 370.

87. D. T. Grubb, L. W. Jelinski, *Macromolecules* **30** (1997) 2860 – 2867.
88. W. Albrecht, M. Reintjes, B. Wulfhorst: "Lyocell-Fasern, Faserstoff-Tabellen nach P.-A. Koch, 1. Ausg. 1997," *Melliand Textilber.* **78** (1997) 575 – 581.
89. H. Blumberg, *Chem. Fibers Int.* **47** (1997) 36 – 41.
90. J. M. Kure et al., *Textile Res. J.* **67** (1997) 18 – 22.

Fibers, 2. Structure

JOHN W. S. HEARLE, The University of Manchester Institute of Science and Technology, Manchester, United Kingdom

1. Introduction . . . 1
1.1. Fibers as Industrial Chemical Structures . . . 1
1.2. Structural Types . . . 3
1.3. Development of Knowledge of Fiber Structure . . . 4
2. General Features of Major Textile Fibers . . . 4
2.1. Chemical Forms . . . 4
2.2. Polymer Characterization . . . 5
2.3. Other Chemical Constituents . . . 6
2.4. Fine Structure . . . 7
2.4.1. Characterization . . . 7
2.4.2. Models . . . 11
2.4.3. Unified Parametric Approach . . . 14
2.4.4. Relation to Fiber Properties . . . 18
2.4.5. Relation to Fiber Formation . . . 19
2.4.6. Structural Transitions . . . 20
2.5. Intermediate Structures . . . 21
2.6. Gross Inhomogeneities . . . 21
2.7. Fiber Shape . . . 22
2.7.1. Cross-sectional Shape . . . 22
2.7.2. Axial Shape . . . 23
3. Particular Textile Fibers . . . 24
3.1. Cellulosic Fibers . . . 24
3.1.1. Chemistry . . . 24
3.1.2. Cotton . . . 25
3.1.3. Other Plant Fibers . . . 26
3.1.4. Rayon . . . 26
3.1.5. Acetate Fibers . . . 27
3.2. Protein Fibers . . . 27
3.2.1. Wool and Hair . . . 29
3.2.2. Silk . . . 29
3.2.3. Regenerated Protein Fibers . . . 29
3.3. Major Melt-Spun Fibers . . . 29
3.3.1. Chemistry . . . 29
3.3.2. Fine Structure . . . 30
3.3.3. PLA Fibers . . . 31
3.4. Acrylic Fibers . . . 31
3.5. Other Polymers . . . 32
4. High-Modulus, High-Tenacity Fibers . . . 32
4.1. Glass . . . 32
4.2. Highly Oriented Linear Polymers . . . 32
4.2.1. Aramid . . . 33
4.2.2. Polyethylene . . . 33
4.2.3. Other Polymers . . . 33
4.3. Carbon Fibers . . . 34
4.4. Ceramic and Other Inorganic Fibers 35
4.5. Structure and Properties . . . 35
5. Other Fibers . . . 36
5.1. High-Toughness Fibers . . . 36
5.2. Elastomeric Fibers . . . 36
5.3. Thermally and Chemically Resistant Fibers . . . 37
5.4. Special-Purpose Fibers . . . 38
5.5. Nanofibers and Carbon Nanotubes 39
5.6. Smart Fibers . . . 39
6. References . . . 40

1. Introduction

1.1. Fibers as Industrial Chemical Structures

In two important senses, fibers are structures. First, the definition of fibers as a particular form of matter depends on their gross structural size and shape. Second, they usually have a rich complexity of internal structure at several levels, which is determined by the biological or industrial circumstances of the formation of any particular fiber type and, in turn, determines the physical and chemical properties of the fiber as well as its utility.

The Textile Institute [1] defines fibers as units of matter characterized by flexibility, fineness, and high ratio of length to width. The geometrical aspects of this definition are paramount because even an inherently stiff and brittle ceramic material is flexible in the fiber structural form and, at the other extreme, a flexible rubber rod is

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_477.pub2

not a fiber. To provide a convenient quantitative marker in units commonly used in the man-made fiber industry, which can be converted to strict SI units, a typical fiber may be taken to have the following dimensional parameters:

Linear density 1 decitex (dtex)=10^{-7} kg/m
Length 1 cm
Density 1 g/cm^3

The International Organization for Standardization (ISO) has accepted tex, defined as grams per kilometer, as a unit of linear density within the SI system. However, an older unit denier, defined as grams per 9000 m, was widely used in the silk and man-made fiber industries. This has led to the adoption of decitex (dtex) as a common industrial unit.

Other derived parameters, if the fiber is assumed to be a solid with a circular cross section, are

Mass	1 µg
Diameter	11.2 µm
Length-to-diameter ratio	1120
Surface area	0.351 mm^2
Specific surface area	351 m^2/kg

The fineness of the structural division can be demonstrated by noting that a small piece (1 m^2) of a lightweight textile fabric (100 g/m^2) would contain a hundred million such fibers. The other structural aspect of the definition, the high ratio of length to diameter, intensifies the situation since fibers can interact, mechanically or otherwise, within a structure wherever they cross, as shown in Figure 1. Thus the essential fiber structural element should be taken as having a length equal to its diameter. The above parameters give the element a length of 10 µm and a mass of 1 ng, or 10^{12} fiber elements per kilogram.

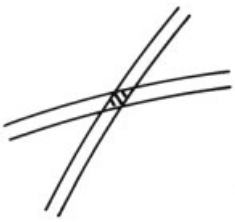

Figure 1. Zone of interaction between two fibers over a distance approximately equal to the fiber diameter

The industrial situation is more complex than indicated by the "three ones" (1 dtex, 1 cm, 1 g/cm^3) of the example. The use of the word fiber in other contexts may cover other sizes. Even for fibers used in the textile industry, there are no precisely defined limits for the dimensions of what may be described without qualification as a fiber. For the natural fibers, which have been used for thousands of years, and the first generation of manufactured fibers, which together account for the major usage of textile fibers, the "three ones" are at the lower limit of fineness, length, and density. For these commercial fibers, the practical limits may be roughly given as follows.

Fineness ranges from about 1 dtex (10^{-7} kg/m) to 30 dtex (3×10^{-6} kg/m), giving diameters up to about 50 μm. Anything much coarser than the above limit of 30 dtex would be referred to as a thick *monofil* or *bristle*. From the last quarter of the 20th century onward, there has been considerable interest in finer fibers. Polyester *microfibers* are directly spun down to 0.1 dtex (10^{-8} kg/m) and finer still by using special techniques, such as the *islands-in-a-sea* procedure, in which fine component fibers are extruded within a removable, soluble matrix [2, 3]. *Nanofibers*, down to 10^{-7} dtex (10^{-14} kg/m) can be made by electrospinning, in which polymer melts or solutions are extruded in a strong electric field [4]. Carbon nanotubes are now being directly spun and collected as yarns [5, 6]. Other terms for finer fibers are *ultrafiber, superfine fiber*, *fibril*, and, for single crystals, *whisker*.

The shortest fibers used in the regular textile industry are about 1 cm in length, but shorter fibers are used in papermaking and in discontinuous-fiber composite reinforcement. The longest individual or *staple fibers* in commercial use are about 10 cm in length; however, man-made *continuous filaments*, occasionally produced as *monofilament yarns* but usually as *multifilament yarns* or *tows*, are effectively infinite in length (they can always be made longer by winding another turn on the package). The natural "continuous filament," namely silk, is about 800 m in usable length.

Density depends on the material, but the lightest common fiber material is polypropylene with a density of about 0.91 g/cm^3, depending on crystallinity. Most common textile fibers have densities between 1 and 1.5 g/cm^3, but the newer fibers such as carbon, glass, ceramic, and metal

fibers, used in composites and other engineering applications, have higher densities.

The structural form, emphasized in the above discussion, is the basis for the industrial exploitation of fibers. The consequences of the structure manifest themselves in several different ways. Traditionally, the fiber form provides the combination of flexibility and strength needed in textile materials for clothing and for a range of household and industrial uses. In modern high-performance composites, the crack-stopping characteristic provides the major advantage over bulk material. A second important feature in clothing is the large volume of space that can be trapped within a fiber assembly to provide softness and warmth; this is also significant in filtration and other industrial uses. Related to this feature is the high accessible surface area which is utilized in other applications. Continuity over long lengths is exploited by nature in nerve fibers and industrially in optical fibers. Networking deserves more exploration.

Finally, three fundamental, academic aspects are also of great industrial relevance:

1) Large structural defects, which are a major source of weakness in bulk materials, cannot occur in fibers. Any defect must be appreciably smaller than the fiber diameter.
2) Because fibers are so fine, their internal structure can be controlled in ways that cannot be used for bulk materials. Even a polymeric fiber, with relatively poor thermal conductivity, can be heated and cooled at rates faster than 1000 °C/s. Patterns of crystallization, which would be impossible with the slow cooling of large volumes, can thus give forms of fine structure that can only be found in fibers (or thin films). Metal fibers can be quenched at rates that result in crystalline (amorphous) glassy metals. Similar arguments apply to chemical changes, with fast diffusion through the fiber, and mechanical stresses. Thus, the chemical, thermal, and mechanical conditions that determine the form of fine structure can, in principle and to a considerable degree in practice, be controlled to a precise sequence, with minimum gradients disturbing the structure formation.
3) The third aspect relates to fibers in systems. In use, the fiber is a component in an assembly. Typically, a collection of fibers will make up a yarn, and yarns will be interlaced to make a fabric, which is the "material" used in the construction of the final product. Both within the fiber and in the arrangement of fibers in the assembly, the multiplicity of levels of structure, illustrated in Figure 2, can be exploited by the designer to achieve the required function. In simpler materials, many of these choices disappear, and some will be missing in various fiber systems.

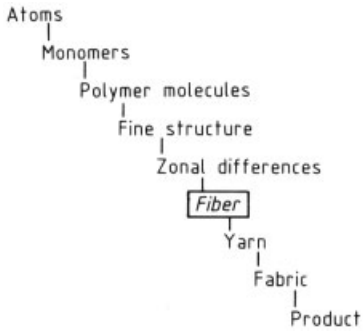

Figure 2. Levels of structure within a fiber and in an assembly of fibers as a useful product

Questions of yarn, fabric, or industrial composite structure are beyond the scope of this article. The subject of fiber structure includes the external form but concentrates within fibers on large-scale inhomogeneities (zonal differences), fine structure (molecular packing), and chemical structure.

1.2. Structural Types

Within the increasing diversity of commercial fiber types, the most useful structural classification is based on the use of the fiber. The following categories are defined and treated separately in the following sections of this article: general-purpose textile fibers, high-stiffness fibers, and special-purpose fibers.

General-Purpose Textile Fibers For ordinary textile usage, a material is needed which can be produced in fiber form; has reasonable mechanical, thermal, and chemical stability, and can be colored by dyeing. However, the most restrictive feature in terms of structural possibilities is that the fiber must have intermediate elas-

tic extensibility, with a breaking extension between 5 and 50 % (usually between 10 and 30 %), as well as appreciable recovery from extensions near break. This rules out as too inextensible the common brittle materials such as crystals and glasses, and ductile materials such as metals; elastomeric rubbers are ruled out as too extensible. The only class of material structure that meets this requirement consists of partially oriented, partially crystalline, linear polymers.

High-Stiffness Fibers The new high-performance fibers used in advanced composites have more diversity of structure. The forms are

1) very highly oriented, highly crystalline linear polymers;
2) the graphitic, irregular, two-dimensional crystalline form of carbon fibers;
3) glass and ceramic structures; and
4) crystalline forms.

Special-Purpose Fibers Fibers for special purposes have an even greater diversity of structure, because a variety of materials can be made in fiber form when the incentive to exploit particular physical or chemical properties is present. Some of the more important materials are mentioned in the Section 5.4.

1.3. Development of Knowledge of Fiber Structure

Textile fibers were used for thousands of years before their structure was studied. During the first half of the 20th century, many studies by optical microscopy provided information on the gross structural features. Much information on a wide range of fibers is contained in [7] and [8].

Only in the 1920s was the polymer (macromolecular) hypothesis established; it was also recognized at that time that the natural plant fibers and regenerated rayons were composed of the linear macromolecule, cellulose and that natural animal fibers were composed of macromolecular proteins. This development coincided with the first X-ray diffraction studies of fibers, which demonstrated a combination of ordered and disordered material within the fiber. This was first interpreted as a dispersion of separate crystalline micelles within an amorphous matrix. However, in the early 1930s the molecular length was recognized as being much greater than the estimated size of the micelles; this led to the fringed micelle model in which the linear polymer molecules run continuously between crystalline micelles and the disordered regions. In the 1950s, the simple view was challenged in three ways: (1) by electron microscope studies which showed the presence of fibrils as a basic structural unit in many types of fiber; (2) by new ideas coming from fundamental studies of polymer crystallization, in particular, chainfolding in polymer single crystals and spherulitic growth in bulk crystallization; (3) by new thinking about the meaning of order and disorder. Some unifying ideas have been proposed, but possibly the most important point is that different fibers have different structures, so one can be specific only in relation to particular cases.

The early development of ideas of fine structure of fibers was reviewed in [9, Chapter 6], which also contains more detailed information on many fiber types; this reference was brought up to date in 1970 in [10]. The understanding and control of fiber structure is reviewed in [11], which covers both natural and manufactured fibers and introduces ideas from the science of complexity, including polymer dynamics, fractals, kinetics of nonhomogeneous processes, nonlinear systems, chaos, and the possibility that fiber formation involves quantum effects. Other books and reviews concerned with various aspects of fiber science and technology also include information on fiber structure [12 – 31].

2. General Features of Major Textile Fibers

2.1. Chemical Forms

This section is particularly concerned with the major textile fibers, which are discussed individually in Chapter 3, although the investigative techniques and some of the general comments are also applicable to the specialist fibers covered in later sections. Of the many linear polymers which have been synthesized in the laboratory or occur in nature, these major textile fibers of the 20th century consisted of only six basic chemical types. These fibers are listed in

Table 1, together with some less important ones. Others polymers have been tried for general textile use, and some may still be used in small quantities. In 2002, the Federal Trade Commission recognized a new generic fiber type: PLA, which is manufactured from poly(lactic acid) or poly(lactate) derived from naturally occurring sugars or starch. The commercial fiber has the trade name NatureWorks.

Table 1. Polymers in textile fibers

Class	Large-volume production	Small-scale or experimental production
1 Cellulosic	cellulose	
	secondary cellulose acetate	cellulose triacetate
2 Protein	keratin in wool	fibroin in silk
3 Polyamide	nylon 6 nylon 66	nylon 4, nylon 7, and others
4 Polyester	poly(ethylene terephthalate)	poly(butylene terephthalate) poly(trimethylene terephthalate), poly(ethylene naphthalate), and others
5 Vinyl/Acrylic	polyacrylonitrile	poly(vinyl alcohol), poly(vinyl chloride), modacrylic and others
6 Polyolefin	polypropylene	polyethylene

The desirable features of the polymer molecule are (1) it should be regular in order to form a crystal lattice, and (2) the stiffness of the chains and the interaction between them should be high enough to give stability to the crystal, but not so stiff as to cause difficulty in formation or excessive rigidity in the fiber. The principles of selection are illustrated in Table 2 by the various combinations of four simple groups in repeating polymer units. Chains composed only of methylene ($-CH_2-$) and ester ($-COO-$) groups are too inert and flexible to give a high enough melting point. However, inclusion of the amide ($-CONH-$) group, which interacts by hydrogen bonding, or of the rigid and electronically interactive benzene ring results in commercially successful polyamide and polyester materials. Nevertheless, these chains still contain sequences of $-CH_2-$ and $-COO-$, which provide enough flexibility. Experiments with other polyamides and polyesters suggest that six is about the right number of main-chain atoms to form such a sequence. If the flexible sequences are eliminated as in the aromatic polyesters and aramids, then the melting point becomes too high, with decomposition occurring first, and the materials are too stiff and intractable for general textile use. However, the ability of these polymers to form liquid crystals can be utilized to make high-modulus, high-tenacity (HM–HT) fibers.

The other synthetic route to obtain useful fibers by modifying the polyethylene molecule is to add side groups, which either are interactive, such as the polar $-C{\equiv}N$ groups in polyacrylonitrile and the hydrogen-bonding $-OH$ groups in poly(vinyl alcohol), or act like the $-CH_3$ groups in polypropylene and cause the chain to form a coil with greater flexural rigidity. In nature, more complicated forms occur, such as cellulose which contains a stiff ring and hydrogen-bonding $-OH$ groups, and proteins which have a variety of possible interactions.

Although the chemical structure of the synthetic polymer fibers thus seems simple, complexity can be introduced through copolymerization. End groups must, of course, be present, but other groups can be introduced to improve or modify dye affinity or to give other special effects. This is done in specialty polyamide and polyester fibers to give deep-dyeing effects or a selective attraction for particular classes of dyes. Copolymerization is always carried out on acrylic fibers, since pure polyacrylonitrile cannot be dyed; thus, considerable diversity exists in the chemical composition of acrylic fibers from different manufacturers, which have different trade names, or of different types having the same main trade name. The modacrylics are a more extreme example of copolymerization, containing less than 85 % acrylonitrile and adapted to give improved flame resistance or other special properties.

2.2. Polymer Characterization

In addition to their chemical composition, the physical form of polymer molecules must be characterized. With the exception of keratin, which is a complex protein system, the chains are normally linear with little branching or cross-linking, although cellulose fibers may be cross-linked by resin treatment of textile fabric to impart crease resistance.

Table 2. Various polymer forms

Type	Typical repeating unit	Comment
Aliphatic polyester	$-CO(CH_2)_2COO(CH_2)_2O-$	$mp < 100\,°C$, no use
Polyethylene	$-CH_2-$	$mp \approx 135\,°C$, little used as fiber
Polyamide	$-CO(CH_2)_4CONH(CH_2)_6NH-$	$mp \approx 260\,°C$, textile fiber
Polyester	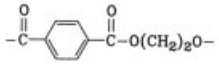	$mp \approx 260\,°C$, textile fiber
Aromatic polyester	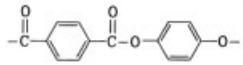	high *mp*, not used
Aramid	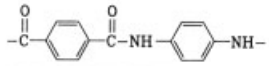	does not melt, high-modulus, high-tenacity fiber

The most important parameters are thus the degree of polymerization (chain length) and its distribution. Generally, the longer the chains, the greater is the fiber strength. However, the maximum degree of polymerization is limited by the conditions of formation. In the natural plant fibers, biosynthesis within the cell enables a very high degree of polymerization to be obtained. However, in the production of rayon fibers the mechanical and chemical attack necessary to get cellulose into solution, inevitably reduces the degree of polymerization. Considerable improvement in properties can be achieved by attention to detail, so that a high degree of polymerization is obtained. The elimination of low molecular mass fractions from the distribution is particularly advantageous. With synthetic polymers, the attainable degree of polymerization depends on the polymerization chemistry: addition polymers are generally able to reach high values, but condensation polymers are limited by the reaction equilibrium.

Although, for ultimate properties, a high degree of polymerization is desirable, this cannot be carried too far because the viscosity of the melt or solution will be too high for extrusion in spinning.

The average degree of polymerization may be found by standard techniques with measurement of viscosity or chemical end-group analysis being the most common. In particular cases, characterization of tacticity, or the form of copolymerization, whether random, block, or graft, may also be useful.

2.3. Other Chemical Constituents

Impurities. In addition to the polymer content, fibers usually contain within their structure appreciable amounts of other substances. In cotton, these include salts and other minor impurities, but in some plant fibers hemicelluloses and lignin may also be present in significant amounts. Wool contains various biological residues, and man-made fibers may contain catalytic residues, low molecular mass analogues such as dimers and trimers, and other manufacturing contaminants. All of these substances may be classed as impurities.

Pigments and Antistatic Agents. Some other constituents are deliberately introduced. Titanium dioxide particles are commonly added to man-made fibers as a delusterant, which gives a more matt appearance to fabric and increases opacity. Carbon black or colored pigments may be added to give a range of "spun-dyed" fibers. Such components are finely distributed within the fiber, but continuous strands of carbon or a colorless conducting polymer may be introduced to give antistatic fibers.

Water. Most of the polymers used in fibers absorb water, which is dispersed within the structure. At an intermediate humidity of 65 %, this gives a moisture content of about 13 % in wool and rayon, 7 % in cotton, 6 % in secondary acetate, 4 % in triacetate and nylon, 1.5 % in acrylic, and 0.5 % in polyester fibers; at saturation, without liquid water being present, some

fibers may absorb their own mass of water, giving a moisture regain of 100 % or a moisture content of 50 % based on total mass. The water will be absorbed predominantly in the noncrystalline regions of the fiber structure or on the surfaces of crystallites. The absorption may be divided between those water molecules that are located directly on active moisture-absorbing groups in the polymer (particularly −OH and −CONH− groups) and those that form secondary layers on top of the directly absorbed molecules. Because of its effects on dimensional swelling, heat of absorption, mechanical properties, diffusion within the fiber, thermal transitions, electrical conductivity, and other properties, moisture is an important element in the structure of many fibers.

Waxes, Lubricants, and Surface Finishes. Finally, other substances are commonly present on the surface of fibers. Natural fibers, as formed, contain oils and waxes, which are usually removed at some stage of chemical processing. However, in the production of man-made fibers and at various stages in the textile processing of all types of fiber, surface finishes are normally added to change the frictional properties of the fibers, usually to lubricate but sometimes to increase friction; frequently, to control static electricity; or for special purposes, e.g., to influence other properties such as wettability. The constitution of these surface components, which typically add 1 – 3 % to the total mass, is very complicated; they may contain a variety of natural oils, synthetic lubricants, industrial surface-active agents, and other substances. Little information has been published on this subject, but a review presents some information from patents [43].

2.4. Fine Structure

2.4.1. Characterization

The level immediately above the molecules is the most specialized and significant aspect of fiber structure in the context of industrial chemistry. Polymerization chemistry and the use of polymers in other applications are discussed elsewhere. The larger scale features are more a matter of engineering. However, the fine structure is characteristic of fibers, is rarely similar to structures found in other forms of the material, and results from the interaction of physical chemistry with the biological or industrial circumstances of fiber formation. This discussion is concerned with how the polymer molecules pack together and involves the range from 1 to 100 nm (10 to 1000 Å).

The linear polymer molecules in all general-purpose textile fibers are oriented preferentially parallel to the fiber axis to a degree which gives high anisotropy but is well short of perfect orientation. The molecules are ordered to about midway between the perfect order of a crystal lattice and the extreme disorder of an amorphous linear polymer, e.g., rubber or glass. The problem of characterizing fiber structure then has two aspects: (1) a qualitative description of the structural form which, while still being partially oriented and partially crystalline, can include many different models; and (2) establishment of the numerical values of relevant parameters, which may be general and relate to many structural models or may be specific to a particular structural concept.

Although various techniques provide information relevant to an understanding of fiber structure, they are all limited in what they provide and subject to artefacts, experimental error, and uncertainty of interpretation. Substantial, unambiguous information about fiber structure cannot be obtained from any single technique. One special danger is that if some particular form of model is assumed, experimental measurement may lead to a precise calculation of some parameter of that model. This will then reinforce faith in the model; yet if another model has been used as a starting-point, it too could be fitted to the experimental results with equal precision. The approach to understanding is thus a matter of integrating a diverse set of data from a variety of sources into a coherent unity. For some natural fibers, this has been achieved with confidence in the results; but for man-made fibers, a clear picture still does not exist.

The analytical techniques that provide almost all the information on fiber fine structure have been used for at least 40 years. Recent advances have occurred through some variation of methods, more powerful instrumentation, and — most important — the use of digital methods to collect data and computers to analyze data.

Degree of Crystallinity. If the fine structure is assumed to be a mixture of crystalline and amorphous material, the degree of crystallinity can be determined in various ways. If the density of the crystalline material is ϱ_c, of amorphous material ϱ_a, and of the fiber ϱ, then the degree of crystallinity δ_{cv} is the volume fraction of crystalline material given by

$$\delta_{cv} = (\varrho - \varrho_a) / (\varrho_c - \varrho_a) \quad (1)$$

The fiber density can be determined most easily in a density gradient tube containing a mixture of two liquids, graded so that the proportion of dense liquid increases from top to bottom of the tube, and calibrated by particles of known density. Experimental errors may be associated with such factors as absorbed water in the fibers, absorption of the test liquids, entrapment of air, and the presence of voids within the fiber. Usually, the dimensions of the crystal unit cell of the polymer are known accurately from X-ray diffraction studies, not necessarily on fibers, and the mass in the unit cell is known from the chemical composition, so that ϱ_c is known, although even here values are being challenged and revised. If the polymer can be quenched to an amorphous glassy state, then a direct measurement of ϱ_a can be made; otherwise, a value must be estimated by some form of extrapolation.

At best, there is some uncertainty about the true values of ϱ, ϱ_c, and ϱ_a; but the reality is worse. The best estimate of ϱ_c is for a unit cell within a large, perfect, unstrained crystal. Within the fiber, the crystals are very small, may contain imperfections, and are strained by the pull of connecting chains; each of these factors affects the density. Defining the true value of ϱ_a is even more difficult, because highly disordered structures may be packed with different degrees of efficiency. Aging of an amorphous structure enables the thermal vibrations to shake the structure down into tighter packing in a more favorable energy state. Furthermore, when a direct measure of amorphous density is available, (e.g., from a rapidly cooled polyester fiber), it is on a volume of material which is large in comparison to molecular dimensions and involves the dubious assumption that small regions of disordered chains, which connect crystallites by lengths of about ten repeating units, would have the same density as the bulk amorphous material. Also, because the scale of fine structure is so small, the interface regions between crystal and amorphous material constitute a substantial part of the whole, and the densities in these regions are likely to be different.

Finally, this whole discussion of degree of crystallinity is based on a tacit assumption that the fiber can be regarded as a two-phase mixture of a crystalline component and an amorphous component, typically represented by a model of crystalline micelles embedded in an amorphous matrix. However, the reality may be a more complicated distribution of intermediate degrees of order or, at the other extreme, a completely uniform, partially ordered, structure.

The problems arising from the application of Equation (1), which appears to be a very simple statement about fine structure, are discussed here at length because they also illustrate the nature of the uncertainties that arise in other techniques for investigating structure, which will be dealt with more briefly. The difficulties of interpretation in relation to models apply universally, but many of the more specialized techniques, while potentially yielding more detailed structural information, are subject to greater errors through artifacts of specimen preparation and problems of interpreting data.

An alternative definition of the degree of crystallinity exists, namely, the mass fraction of crystalline material:

$$\varrho_{cm} = (V_a - V) / (V_a - V_c) \quad (2)$$

where V, V_a, and V_c are the specific volumes, which are reciprocals of the respective density. Since the difference in densities is not large, the two quantities will not differ greatly.

Although one must be very cautious about a precise and absolute interpretation of the degree of crystallinity, the fractions defined in Equations (1) and (2) provide a scale of degree of order in fibers, from poorly ordered to highly ordered, even if the scale is arbitrary and empirical. This interpretation would only be upset in highly exceptional circumstances (e.g., in ice and water) where the ordered crystal lattice had a lower density than the disordered arrangement.

Other physical and chemical measurements can be used to define the degree of order (degree of crystallinity) by equations analogous to (1) and (2). These include *refractive index*, which is proportional to density; *moisture take-up*, which occurs within amorphous regions but usually not

within crystalline regions, although absorption will occur on crystal surfaces; other *absorption techniques*; and *latent heat of melting*, although here the fraction would equal the ratio of fiber latent heat to crystal latent heat (L/L_c), because of the assumption that the amorphous material does not contribute. Other techniques exist in which the responses X_c and X_a from the crystalline and amorphous materials can be separated, so that the relevant fraction becomes $X_c/(X_c + X_a)$. Such techniques include X-ray diffraction, electron diffraction, infrared absorption, NMR and Raman spectrometry. However, the problem here is that the two responses overlap, so the separation of X_c and X_a is uncertain.

Provided any errors of experimentation or interpretation directly associated with the techniques have been eliminated, all the above methods give the same value of the degree of crystallinity for a simple model of large crystals dispersed in amorphous material (like a mixture of crystals and liquid in an organic compound at the melting point). However, when the real structure is not defined so simply, the results will differ because each technique will be defining a different scale of the degree of order.

X-Ray Diffraction. One of the most useful techniques for studying fiber structure is wide-angle X-ray diffraction, in which a beam of X-rays is passed through a bundle of parallel fibers. Specialists should be consulted about the minimization of errors associated with the X-ray source and its collimation, specimen preparation, the method and geometry of detection (photographic film or solid-state detection), and the extraction and interpretation of quantitative data. Because of the inverse form of Bragg's law, $\sin\theta = n\lambda/2d$ (where θ is the angle of a strong reflection from a crystal lattice, n is an integer defining the order of the diffraction, λ is the X-ray wavelength, and d is the crystal lattice spacing), wide-angle scattering (WAXS) gives information at the level of atomic spacing, whereas small-angle scattering (SAXS), which requires a different instrument, gives information on larger scale structural features.

A perfectly crystalline, perfectly oriented fiber would give a wide-angle X-ray diffraction pattern of the form shown in Figure 3A (a). The symmetry of crystals oriented parallel to a single axial direction, but with random orientation around this axis, leads to the reflections from particular lattice spacings appearing on layer lines, with the separate orders at increasing distance from the meridian. Real textile fibers give patterns of the form indicated in Figure 3B (a). The first-order reflections are easily seen, but higher orders are usually very faint. The sharp spots are diffused, and more general scattering occurs in the form of a halo from disordered material.

The relation between various features of the X-ray diffraction pattern and structure is as follows:

1) Location of spots: size and shape of unit cell; arrangement of atoms in crystal lattice
2) Radial broadening of spots: increasing with reduced size and perfection of crystals
3) Circumferential spread into arcs: decreasing orientation of crystals
4) Ratio of sharp spots to diffuse scattering: degree of crystallinity

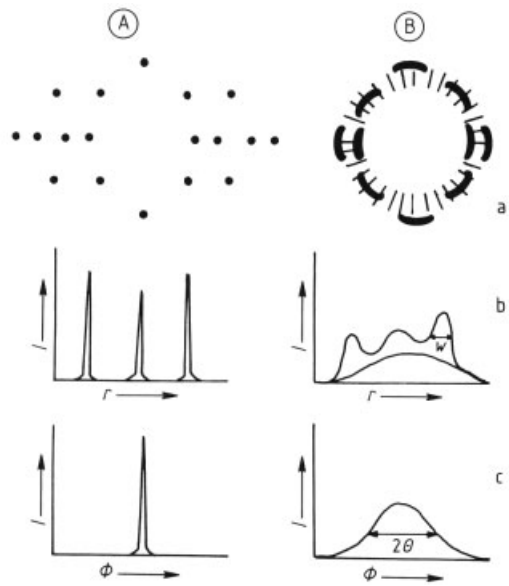

Figure 3. A: Perfectly crystalline fiber; B: Typical fiber a) Simplified representations of wide-angle X-ray diffraction patterns; b) Radial scans of intensity I against an interval of radius r, showing peaks broadened to a half-width w, superimposed on an amorphous background A; c) Scans along a circumferential arc of intensity I against polar angle ϕ, showing broadening 2 θ resulting from range of orientation

In Figure 3 (b) and (c) is illustrated how quantitative parameters can be determined roughly,

but much more detailed analysis is needed to get precise measures. In principle, the complete X-ray scattering contains information from the complete pattern of packing within the fibers, including the size, shape, and perfection of the crystals and statistical features of the amorphous material. The problem is to devise analytical schemes to extract this information without misinterpretation.

The small-angle X-ray scattering pattern consists of diffuse spots from which an indication of long spacings within the fine structure can be obtained.

The techniques continue to advance. Digital collection of data and computational analysis increase the useful information which can be generated. The use of intense synchrotron sources of X-rays enables information to be obtained in much shorter times and allows dynamic studies of structure development in a threadline to be performed.

Electron Microscopy. Transmission electron microscopy is another source of information, which has proved particularly useful where there are well-defined structural units. Thus the cellulose in cell walls can be shown to be in the form of fine fibrils, either by disintegration or by taking surface replicas (→ Cellulose); and in wool, transverse and longitudinal sections can be cut and stained, and clearly show a fibril – matrix composite structure, with dimensions that can be measured (→ Wool). However, these simple statements hide the great skill and care needed to extract useful and reliable information. In man-made fibers, where the structure is less well-defined, information is more difficult to extract, but new techniques of high-resolution electron microscopy on thin sections show promise. A particularly promising technique is the use of high-voltage electron microscopy on thick sections. By tipping the specimen, information of the three-dimensional structure can be obtained and shown by tomography. This method has been used by Bryson et al. [44] to determine the helical angles in macrofibrils in wool.

Electron Diffraction. Electron diffraction generates the same sort of information as X-ray diffraction, but because it is carried out with a minute beam on a thin specimen, it can be used to distinguish different areas of a cross section and different directions in the fiber.

Other Microscopy Techniques. Since its introduction in the 1960s, scanning electron microscopy (SEM) has been used to study the overall form of fibers, and, particularly, the details of fiber fractures [45]. Recent advances in high-resolution imaging enable finer structural details down to 1 nm or less to be examined [46].

Atomic force microscopy is being used to investigate coarser features of fiber structure, such as the cell structure in wool, and to measure the mechanical properties of individual structural components. Refinements of the technique are likely to improve resolution.

Due to limits on resolution and contrast, it is not possible to obtain much information about the internal structure of fibers by conventional optical microscopy. However, the development of confocal light microscopy now allows coarser features, such as the cells in wool, to be observed. In this technique a small spot of light is scanned across the specimen at different depths. The resultant reflected or fluorescent light is picked up on a detector and gives images of successive cross sections of a fiber.

Infrared Spectrometry. Infrared spectrometry can be used to detect the presence of particular chemical groups, but its relevance to fine structure results from the fact that the absorption frequency changes with the environment of the group. Degree of order can be determined by separating the bonds associated with crystalline and amorphous materials. A difference also exists between *cis* and *trans* configurations, and crystallographic chain folds may give a characteristic shift. Other information can be inferred from the change of frequency with the state of strain of a fiber. The use of polarized infrared radiation can yield data on orientation of absorbing groups in the material. Infrared absorption, particularly with high-speed Fourier transform techniques, has been very useful in polymer studies, but problems of suitable arrangement have limited its value in studies of fibers.

Raman spectrometry. In its simplest form, Raman spectroscopy gives information similar to infra-red spectroscopy. However its value has been transformed by the introduction of the Renishaw Raman microscope. Now, in addition to

detecting local differences in molecular environments, the shift in bonds can be calibrated to determine molecular strains.

Nuclear Magnetic Resonance Spectrometry. Another major analytical technique used to a small extent to obtain information on fibers is nuclear magnetic resonance. STATTON [47] pioneered the use of wide-line NMR, which relates to mobility of molecular segments within a fiber structure, but the method has not been used much since his work.

Refractometry. Use of the refractive index as a measure of density has been mentioned, but the most important use of an optical technique is the determination of birefringence as a measure of orientation. This is one of the most straightforward methods since, apart from any errors resulting from form birefringence, the difference between the refractive index n_1 for light polarized parallel to the fiber axis and n_2 for light polarized perpendicular to the axis is directly related to the orientation of groups contributing to the dipole moment. The Herman's orientation function f, defined below in terms of n_1 and n_2, is related to the average angle of inclination of molecular segments, ϕ:

$$f = (n_1 - n_2)/(n_1' - n_2') = 1 - 3\sin^2\phi/2 \quad (3)$$

where n_1' and n_2' refer to a perfectly oriented fiber.

The orientation determined in this way relates to the whole fiber, whereas the crystal orientation is given by X-ray diffraction. The average amorphous orientation can be calculated from the other two values.

Dichroism. Dichroism, including fluorescence methods, is another optical technique for determining orientation. Other structural information can be inferred from light scattering.

Thermal Analysis. The use of latent heat as a measure of crystallinity has been mentioned, but the complete differential scanning calorimetry (DSC) or differential thermal analysis (DTA) response contains more information. The presence of multiple endotherms indicates the existence of different structures, with the more stable forms, such as larger crystals, melting at the higher temperatures. Generally, any endotherm or exotherm is an indication of some structural change.

Other Sources of Information. In addition to the analytical techniques, two other important sources can be used to understand fiber fine structure. The first is a consideration of the way in which the molecules would be expected to behave under the conditions of fiber formation. This involves an application of knowledge of the movement and deformation of polymer molecules and of polymer crystallization, and varies with the chemical constitution of the polymer.

The second is the relation between structure and various physical and chemical properties. For example, a difference in appearance after dyeing is perhaps the most sensitive way of detecting a structural change, but this cannot be interpreted easily. The range of properties related to structure includes diffusion and accessibility; mechanical and thermomechanical responses, particularly dynamic modulus studies; electrical properties; and others. Attempts to explain these properties lead circumstantially to views about fine structure. Most certainly, if a particular structural model gives good predictions of properties without the use of adjustable parameters, this is strong evidence in support of the model.

2.4.2. Models

The first widely used model of textile fiber fine structure was the *fringed micelle* model, illustrated in Figure 4. This model reconciled the indications of the presence of both crystalline and amorphous regions with the great length of the chain molecules by postulating that a given molecule fringed in and out of the crystalline micelles. In fibers, as distinct from unoriented polymer forms, the orientation was assumed to be preferentially parallel to the fiber axis. This model was plausible because it might be expected to occur if crystallization started at many places within the fiber. Furthermore, it gave reasonable explanations of many fiber properties.

When chain folding in polymer crystals was discovered, a *modified fringed micelle* structure was suggested (Fig. 5). This model shows a mixture of chain folding and fringing at the edge of each micelle, and recognizes that relatively high degrees of crystallinity, such as 50 %, can only be accommodated if the micelles are regularly

stacked in a pseudofibrillar array. If required by the crystal lattice, the ends of the micelle will be angled. Models of this sort, proposed by various scientists, remain the best working models for some types of fiber.

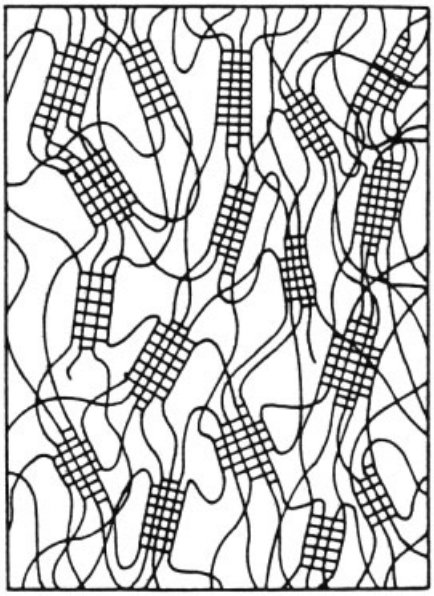

Figure 4. Fringed micelle model, as drawn by Hearle [48]

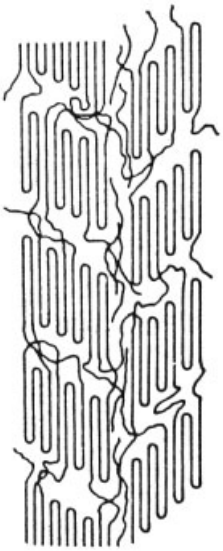

Figure 5. Modified fringed micelle model, as drawn by HEARLE [49]

Characterizations such as Figures 4 and 5 are not explicit theories of structure. They are inadequate pictorial representations of poorly defined views of structure. They do not attempt to be quantitative in terms of size, shape, and spacing of crystalline regions or of molecular paths in amorphous regions. Differences in representation by different authors reflect artistic skill and temperament more than scientific intention, as well as the limitations imposed by size and a two-dimensional format.

An attempt by MURTHY et al [50] to give a more quantitative picture of the structure of a nylon 6 fiber on the basis of SAXS and WAXD data is shown in Figure 6. The main features of Figure 5 can be seen, but with more irregularity in the structure.

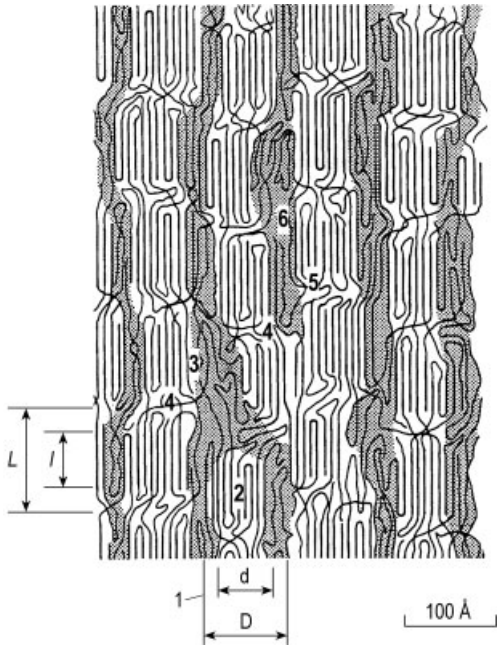

Figure 6. The structure of a nylon 6 fiber, as envisaged by MURTHY et al [50] on the basis of SAXS and WAXD data. 1) Fibrils; 2) Lamellae; 3) Partially extended chains in the interfibrillar regions; 4) Tie molecules in the interlamellar regions; 5) Free chain ends; 6) Amorphous segments with large free volume, which may give rise to voids; 7) Fusion of adjacent microfibrils. Shaded areas represent the interfibrillar amorphous regions.

In some fibers, evidence suggested that the crystals might be in the form of long fibrils, and this led to the suggestion of a *fringed fibril struc-*

ture (Fig. 7). *Fringed lamellar forms*, (Fig. 8) are also possible, but they are more likely to occur in bulk polymers than in fibers.

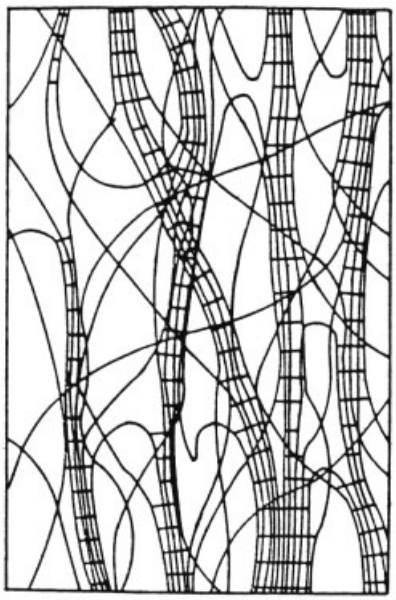

Figure 7. Fringed fibril structure, as drawn by HEARLE [51]

All these models are two-phase structures in which crystalline and amorphous regions are clearly differentiated. A radically different approach is to propose a *uniform one-phase structure* of intermediate form. HOSEMANN [53] investigated the paracrystalline structure in which the disorder results from a statistical deviation from the true lattice spacings: this maintains an approximation to short-range order, but means that there is no long-range order. KARGIN [54], from the other direction, suggested that no real crystallinity exists, at least in some types of fiber, and that the structure is amorphous, with a degree of order introduced by some correlation between the positions of neighboring molecular segments.

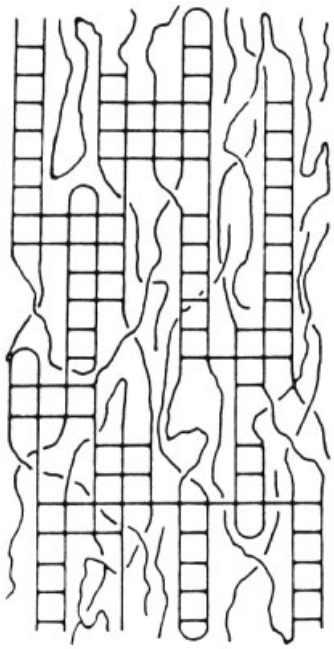

Figure 9. Dynamic crystalline gel model in oriented fiber, as drawn by HEARLE [10]

Adopting a somewhat more localized view, RENEKER [55] has explored the explanation of disorder as the result of particular forms of crystal defect, somewhat analogous to dislocations in metals; this is probably a useful representation for polyethylene and other chains with short repeating units. On the other hand,

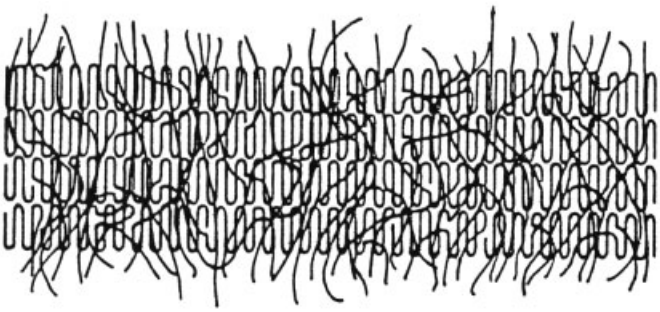

Figure 8. Fringed lamellar structure, as drawn by HEARLE [52]

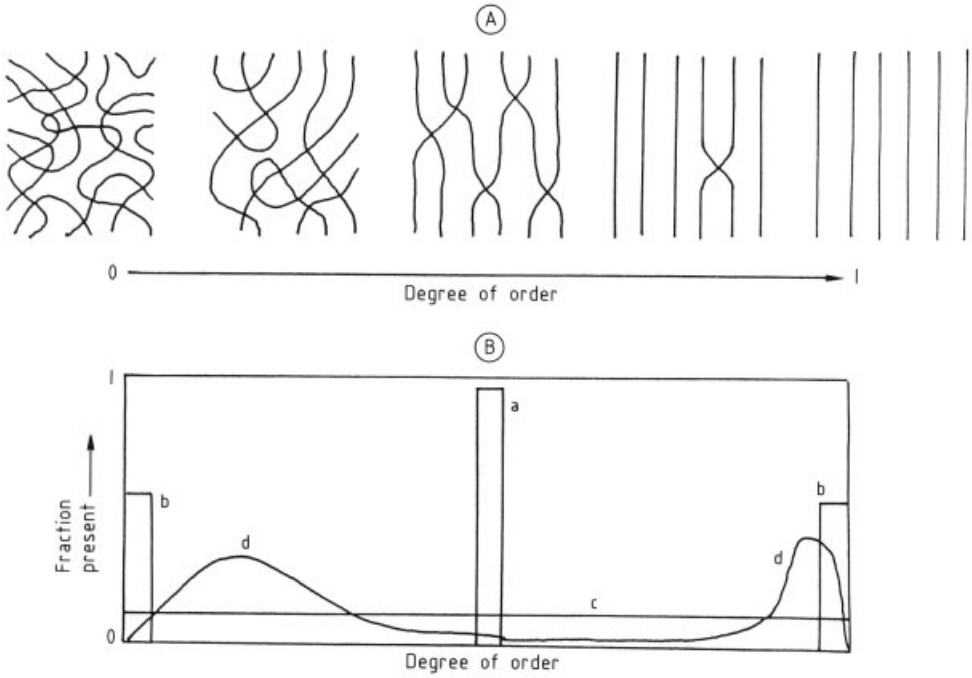

Figure 10. A: Schematic representation of examples of degree of order from amorphous (zero) to the perfect crystal (one); B: Distributions with different degrees of localization of order
a) Uniform at degree of order of 0.5 (zero localization); b) Perfect crystal plus amorphous (maximum localization); c) Hypothetical distribution with all degrees equally represented; d) Mixture of high and low order (moderate localization)

Hearle [56], following a thermodynamic consideration of multiple melting in nylon and polyester fibers, has suggested a *dynamic crystalline gel* in which small molecular segments are in register throughout the material, as illustrated in Figure 9.

According to another view, amorphous material as such might not be present, but the structure may be an assembly of fine crystalline fibrils with disorder coming from imperfections in the packing of the fibrils. The density, for instance, would be reduced by the space between fibrils.

Considerable conflict and diversity of view about fiber fine structure appear to exist, but two important points must be remembered: (1) different fibers have different structures, so different choices of model will give the best representation for particular fibers; (2) real structures are likely to be much more complicated than the models, which are usually simplified to an extreme in an attempt to clarify understanding. Real fibers are likely to show features from several models. All the structures can be represented as a distribution of paracrystals in a matrix of partially correlated disorder, with the possibility of localized defects in the crystal or localized segments in register in the amorphous part, although such a description is too general to add much understanding.

2.4.3. Unified Parametric Approach

An alternative way of dealing with the problem is to try to define a set of parameters that are generally applicable and describe important aspects of structure.

A perfectly crystalline material can be described by a limited number of coordinates in a crystal lattice, and a highly amorphous material by a limited number of statistical parameters. Structures of intermediate order can only be truly described by specifying the location of every molecular segment in a region large enough to be typical of the whole. This would lead to a set of parameters approaching Avogadro's number, so that one would have to be highly selective in choosing a reduced set!

The most useful parameters appear to be the following:

1) Degree of order
2) Degree of localization of order
3) Aspect (length-to-width) ratio of crystalline units
4) Degree of orientation
5) Size of crystalline units
6) Molecular extent

The idea of order is discussed above in relation to measurement of "crystallinity." Theoretically, a *degree of order* could be defined as a parameter based on statistical correlation between vectors representing neighboring molecular segments. In a perfect crystal, the vectors would be perfectly correlated in location and direction; at maximum disorder, there would be no correlation. Practically, a scale of degree of order can be defined by density or some other measurable quantity. In a uniform structure, the degree of order can vary continuously between the two extremes as illustrated in Figure 10A.

The idea of *localization of order* is then illustrated by Figure 10B. An average degree of order at some intermediate value, say 0.5, can be given by many different distributions. At one extreme (a), the whole material can be uniform, with zero localization of order. At the other extreme (b), the structure may be completely localized with perfect crystals embedded in a wholly amorphous matrix. Another form (c), at an intermediate degree of localization, would have equal regions of all degrees of order. These simple idealized forms are unlikely to occur in real fibers; the most common form is likely to be some variant of (d), namely, imperfect crystals mixed with varied disordered material, but certain fibers do tend more to the extremes of very high localization or complete uniformity.

The *aspect ratio*, or ratio of length to width, is a measure of the shape of the crystalline regions (with the assumption that the localized units are the highly ordered parts dispersed within a disordered matrix). This can range from infinity to one, through fibrils of great length, limited length, and elongated micelles, and then from one to zero through short, wide micelles, lamellae of limited width, and very wide lamellae, as illustrated in Figure 11.

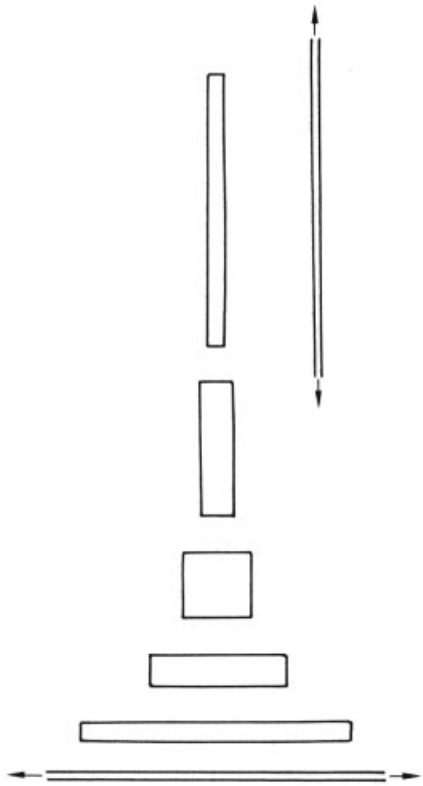

Figure 11. Shapes of crystalline regions of varying aspect ratio, decreasing from top to bottom

The different models of structure described in Section 2.4.2 can then be located on a three-dimensional plot, as shown in Figure 12. However, different models, with subtle differences in molecular packing, can occupy the same place in the diagram, and a continuous range of structures exist, which merge imperceptibly from one form to another. The named models are idealizations which lead to simple description. Figure 13 indicates where different types of fibers lie on such a plot. The justification for the location, and the uncertainty, will become apparent in later sections dealing with particular fibers.

Fortunately, the three parameters plotted are enough to describe the essential character of a structure. The other features can be regarded as superimposed on the basic form.

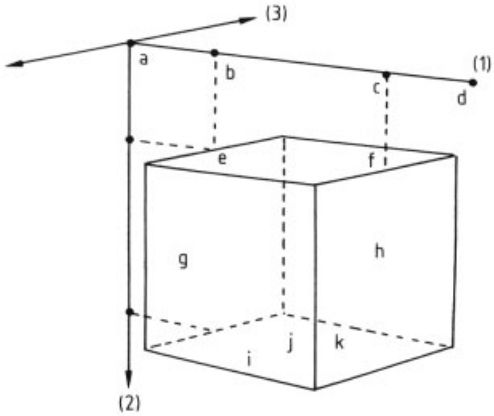

Figure 12. Representation of models of structure on a three-dimensional plot with axes (1) degree of order, (2) degree of localization of order, (3) aspect ratio of localized units
Forms are: a) Wholly amorphous; b) Amorphous with correlation; c) Paracrystalline; d) Perfect crystal; e) Amorphous with varying correlation; f) Crystalline with distributed defects; g) Networks of regions of high and low order; h) Crystalline with concentrations of defects; i) Fringed fibril; j) Fringed micelle; k) Fringed lamellar

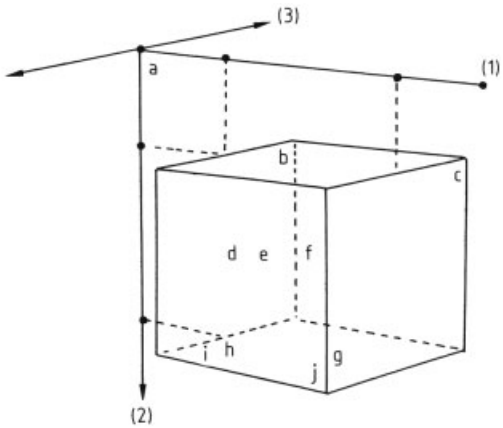

Figure 13. Estimated location of fiber types on the plot of Figure 12
a) Rubber, glassy polymer; b) Acrylic; c) Melt-spun polyethylene; d) High wet modulus rayon; e) Rayon; f) Nylon, polyester; g) Aramid, high molecular mass polyethylene; h) Polyurethane (spandex); i) Wool; j) Cotton

Any of the suggested fiber structures can occur at varying *degrees of orientation*. As discussed with relation to birefringence in Section 2.4.1, the degree of orientation is a quantity which is easily defined as some average of the orientation of individual molecular segments, and easily measured. Crystallite orientation is simply related to the distribution of crystal axis angles, such as those shown in Figure 4, relative to the fiber axis. In natural fibers, as discussed below, genetically controlled mechanisms lead to helical forms of orientation. For melt-spun fibers, such as nylon and polyester, it is not accidental that Figures 5 and 6 show perfect crystallite orientation. Orientation is induced by drawing with a typical draw ratio of 4:1. The combination of high draw and a crystallinity of around 50%, which implies that in each 3D direction the spacing between the micellar crystals is, on average, one sixth of the crystallite dimensions, means that a very high degree of crystallite orientation is inevitable. Amorphous orientation is given by the distribution of angles of molecular segments. Earlier drawings, such as Figures 4 and 5, tend to show an almost random distribution of angles, but the more specific model of Figure 6 shows a fairly high degree of amorphous orientation. In wholly amorphous polymers, such as rubbers, orientation of molecular segments occurs during stretching and can be set by cooling below the glass transition temperature, only to be released on heating. In the melt-spun fibers the orientation is induced by the draw tension and the driving force that pulls molecular segments into crystallites. On release of the draw tension, the geometrical constraints of the structure limit the extent to which de-orientation of molecular segments in the amorphous regions is possible.

The *size of the crystalline regions* in any of the structures can range from small to large. The structure may have a fine texture, with small crystallites close together, or a coarse texture with large crystallites farther apart.

The first five parameters are all obvious characteristics of fine structure. The concepts are clear, and qualitative comments and rankings can certainly be made, although experimental difficulties in measurement can arise and debate may occur about the best mathematical forms of definition and about problems in interpreting measurements. Any diagram of fine structure includes implicitly the information needed to calculate values of each parameter, even if the diagrams are so simplified that the values are not realistic.

The sixth parameter, *molecular extent*, is a less obvious structural feature. It has never been measured directly and could not be estimated from a typical diagram because the relevant information is contained in molecular paths over distances greater than those shown in the diagrams. However, molecular extent almost certainly ranks equally with the other five parameters in determining important fiber properties.

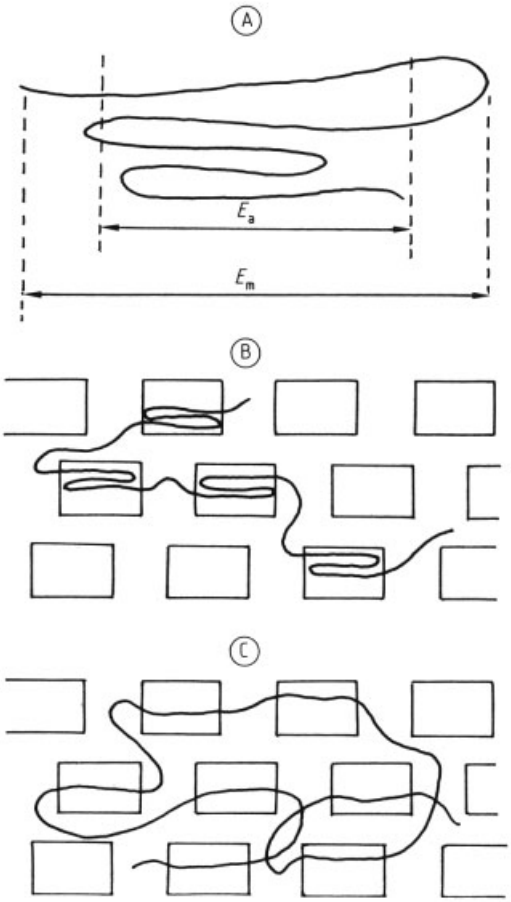

Figure 14. A: Definition of molecular extent, either as an extreme value E_m or an average value, E_a
B: Molecular extent with crystallographic folding
C: Molecular extent without crystallographic folding

Figure 14 A illustrates the basic definition, with Figures 14 B and C giving examples in micellar structures with and without crystallographic folding. Molecular extent is a measure of the distance over which the molecule extends. Particular definitions might relate either to the extreme value E_m or to some average of length of the folds E_a. In principle, molecular extent could be measured in different directions, but in fibers the molecular extent parallel to the fiber axis is important. As obvious from Figures 14 B and C molecular extent is closely related to two frequently discussed structural features: the number of chain folds and the number of tie molecules between crystals. In an amorphous isotropic polymer, it would be related to radius of gyration.

Reference to the analogous quantity in textile yarns spun from short fibers like cotton or wool might be helpful. As the material comes from carding, the fibers are folded backward and forward, and thus have a limited fiber extent. Various drafting operations then pull the fibers out into more extended forms until, in the limit, the fiber may be fully extended along the yarn with no loops or folds. This increases yarn strength because any fold is as much a source of weakness as a fiber end, since both interrupt continuity in the transmission of stress. The fiber extent can be measured experimentally by incorporating colored tracer fibers, but an analogous procedure is not possible for the molecules within the fiber. Neutron scattering of deuterated molecules might be a way of estimating molecular extent, but the experiments would be very expensive.

The three different parameters — degree of order, degree of orientation, and molecular extent — must not be confused. The problem is that at the extremes, which never occur in practice, they do become identical, as shown in Figure 15 A and B. A tendency exists to assume that they go together through the intermediate states. However, this is not so, and Figures 15 C – F give various schematic examples of structures in which one parameter is high and another low. Analogous situations will occur in the real structure of textile fibers.

As stated at the start of this section, specification of six parameters still leaves open an infinite range of possible structures, with a large number of forms showing significant differences. Other features which can be discussed in the context of the structure of particular fibers or which must be considered in theoretical analysis of the relation between structure and properties include (1) more detailed specification of the size and shape of crystalline regions; (2) the nature of imperfections within crystals; (3) the form of molecular

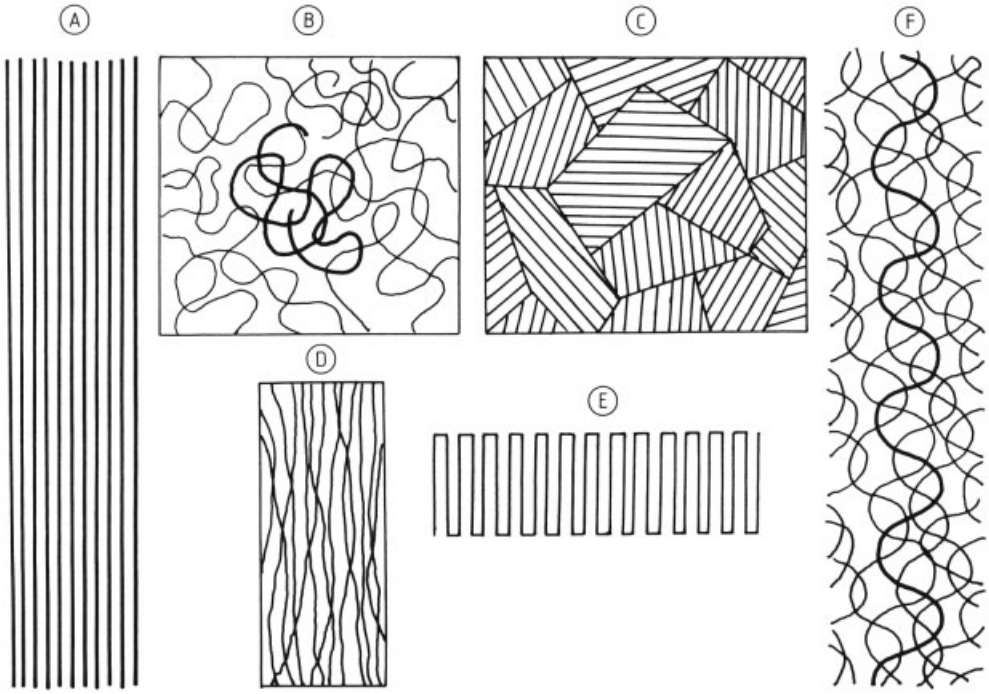

Figure 15. Schematic illustration of independence of degree of order, degree of orientation, and molecular extent, except at the extreme values
A: Perfect order gives perfect orientation and maximum molecular extent
B: Maximum disorder gives zero orientation and a low extent
C: High order and zero orientation in a polycrystalline material
D: Low order in a structure with all segments oriented at a small angle to a preferred axis
E: High orientation with low molecular extent
F: High molecular extent with zero preferred orientation of segments

packing in amorphous regions; (4) the relative positions of crystallites on a regular or irregular superlattice; and (5) details of the interconnections between crystallites.

2.4.4. Relation to Fiber Properties

Although the relation between fine structure and fiber properties can be discussed in detail only in the context of particular fiber types, some general trends can be mentioned.

Crystallinity tightens the structure and gives it greater stability. Consequently, increasing degree of order usually yields increased strength (unless bad stress concentrations are present) and increased stiffness, reduced moisture absorption and accessibility to chemicals such as dyes, and increased dimensional stability under changes of temperature and humidity.

The *degree of localization of order* seems to have little influence on most mechanical and other physical properties. A low degree of localization gives a uniform structure: if this reflects the coherence of ordered material, diffusion within the fiber will be resisted; but if, as is more likely particularly in hot or wet conditions, this represents the mobility of disordered material, then the movement of substances such as dyes will be relatively easy. Generally, low degrees of localization of order will be thermodynamically unstable, so that such materials can be set in a more stable form with an increase in the localization.

Crystallite aspect ratio has a major effect on mechanical properties, with a high ratio promoting greater stiffness (initial modulus), which is usually associated with reduced extensibility and greater strength.

Degree of orientation has a major influence on mechanical properties, as illustrated by the family of stress – strain curves in Figure 16. As indicated below, this is also a parameter that can be controlled easily. Increasing orientation trades extensibility for strength, with the breaking points lying on a locus of increased tenacity T and reduced breaking extension E, which approximates to a constant value of $T\,E^{1/2}$. The initial modulus also increases. For clothing and household uses, the softer, lower modulus fibers are preferred, and the lower strength is adequate. For industrial uses such as tire cords and ropes, the more oriented high-tenacity fibers are preferred [57].

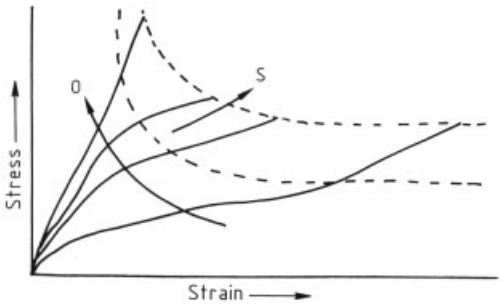

Figure 16. Family of stress – strain curves, showing effect of increasing orientation (→ O) or improving structure (→ S)

The *size of localized units* does not have much influence on the initial part of the stress – strain curve, but it does influence the strength. A reduction in value is one way of improving structure, which moves the locus of failure in Figure 16 further out. This is because a coarse structure of larger units tends to give greater stress concentrations and less effective connections between crystallites, with more free ends and fewer tie molecules. An increase in size also makes the spaces between crystallites greater, and this increases the mobility of dye molecules. Treatments that change structure in this way are thus prone to cause unsightly dye streaks in fabrics if they are not well controlled.

Increasing *molecular extent* is another way of increasing the strength of a structure, usually associated with some increase in stiffness. This will also push the locus of failure in Figure 16 further out, since a high extent means that the structure is more strongly connected along its length. The mechanics can be analyzed in terms of slippage, which occurs at free ends of molecules or at folds, and leads to a lack of contribution to tension.

2.4.5. Relation to Fiber Formation

The development of fine structure can also be discussed in detail in relation to particular materials and manufacturing methods, but once again, a few general comments can be made. An important distinction exists between natural fibers and man-made fibers. The former, except for silk, are laid down very slowly as living cells over a period of days, weeks, or months, under the genetic control of mechanisms of biosynthesis of the polymer in living organisms. This leads to highly specialized structures. In contrast, man-made fibers are formed from a melt or solution of the polymer by high-speed manufacturing operations. Typical spinning speeds will range from 500 to 5000 m/min (ca. 10 to 100 m/s), so that the solidification which establishes the structure, occurs in a fraction of a second. The manufacturer is thus left with what happens in the transition from disorder to partial order, and this is likely to be a rather poorly defined structure. Thermodynamically, the structure will be metastable, so that in the melt-spun thermoplastic polymer fibers, it can be changed by subsequent heat treatment at temperatures approaching the melting point. The structure of fibers, like cellulose or the acrylics which can only be spun from solution because they decompose before they melt, will be less easily changed, since this would require the action of plasticizing or swelling agents, such as caustic soda which changes the structure of cotton in mercerization.

A particular example of the difference between natural and man-made fibers is shown by the *degree of order* in cellulose fibers. Natural plant fibers, like cotton, have a crystallinity of about 66 %, but rayon only reaches about 33 %. Rapid crystallization from a disordered solution or melt leads to crystallinities which are characteristic of the polymer stiffness and complexity. Therefore, nylon and polyester will give values of about 50 %, but linear polyethylene can give higher values. Usually, different treatments will not change the value by more than about

± 5 %. However, when a different manufacturing method is available, such as the liquid crystal systems used in aramids or gel spinning for high-modulus polyethylene, much higher values can be attained.

The *degree of localization of order* is also an almost inevitable consequence of the manufacturing system, although it may be increased by heat treatment of thermoplastic fibers.

The *aspect ratio* of the crystallites depends on the circumstances of crystallization, which are not usually amenable to control in manufacturing, with rapid crystallization from a disordered state tending to give micellar crystals. The contrast between natural and regenerated cellulose fibers is again instructive: the former are fibrillar, but the latter are usually micellar. By changing the chemistry of the regeneration of viscose rayon, crystallization can be delayed until the fiber is solid and oriented, and this does lead to a more fibrillar form of rayon with a higher modulus.

Degree of orientation is easy to control in manufacturing operations, so that the family of products represented by Figure 16 can be generated. Orientation may be increased by increasing the stretch imposed either between extrusion and solidification, provided the spinning speed is high enough to limit relaxation, or in a subsequent drawing operation. In thermoplastic fibers, orientation can be increased further by hot stretching of yarns or reduced by relaxation and shrinkage. Thus, the fiber manufacturer has a number of ways at his disposal to adjust orientation. Technically, no lower limit exists, but the market only requires products with orientations above a certain level. An upper limit to what can be achieved without breaking fibers does exist, however, usually at around 10 % breaking extension. Other mechanisms are involved in producing the high orientation of the high-modulus, high-tenacity fibers.

Crystallite size is not controlled easily, except that it tends to be increased by heat treatment of thermoplastic fibers. Major changes in manufacturing methods also cause differences. Thus a change in the chemistry of regeneration of viscose rayon fibers, with the sodium cellulose xanthate going through an intermediate zinc cellulose xanthate before being converted to cellulose, leads to a finer structure than is achieved by direct generation. High-speed spinning of polyester fibers, in which all the orientation is induced in the initial spinning operation, seems to give a structure with larger crystallites than a process of solidification followed by drawing.

Molecular extent is also not easy to control, except by the use of special techniques such as gelspinning, or by biological control, which gives a very high value in natural cellulose fibers. In manufacturing from solution or melt, the molecular extent will depend on the natural configuration of the molecule in the disordered state, represented by a radius of gyration, transformed by the elongation during fiber formation.

2.4.6. Structural Transitions

Fiber structure is often pictured in static terms, like an instantaneous snapshot of some geometric pattern. However, the dynamics, particularly the transitions between different degrees of mobility, also have to be considered.

The major transitions such as melting and chemical decomposition lead to a loss of fiber identity. In principle, any linear polymer is thermoplastic and will melt, but in practice, a distinction exists between those that melt and those that decompose and char before a melting temperature is reached. The chief thermoplastic fibers are nylon, polyester, polypropylene, and triacetate. Close to the melting point, the structure develops a degree of mobility, particularly if it has not been annealed for a long time to develop large stable structures. Consequently, these fibers can have their structure permanently changed, or heat-set, by high-temperature processing. The chief decomposing or charring fibers are cellulose and polyacrylonitrile. Wool and secondary acetate tend to liquefy and decompose simultaneously.

Below the melting point, glass transitions occur which have some of the features of thermodynamic second-order transitions. They occur in amorphous regions and involve a change in some feature of the structure from rigidity to mobility. Rigidity, which causes the material to be glassy, can come either from stiff chains or from strong interactions between chains; either of these will prevent relative movement. In a simple polymer, a single transition will occur from a rigid glass to a flexible, rubbery form; however, in some important fibers, the total change is divided

between multiple transitions. These transitions cause a major reduction in fiber stiffness, and cooling through the transition gives a temporary set.

Mobility can also be induced by plasticization; this is particularly important in moisture-absorbing fibers, where the transition temperature becomes much lower in the wet state.

Another form of minor but first-order transition, which occurs in some fibers, is between different crystal lattices.

2.5. Intermediate Structures

Some textile fibers contain structural entities at a scale immediately above the fine structure. Thus in cotton, which is a unicellular fiber, the fibrils are arranged in lamellar layers; and in wool, the microfibrils of the basic structure are grouped in macrofibrils within the individual cells which make up the whole fiber. Similar macrofibrillar associations may be present in man-made fibers, but they are not well defined and their reality has not been established conclusively. The fibrillar strands, which are often seen when fibers are pulled apart, may merely reflect lines of weakness within the fine structure.

Spherulites. The slow crystallization of polymers in bulk from the melt usually gives a spherulitic structure as a result of repeated branching from widely spaced nuclei, but the rapid cooling of fibers normally leads to a high concentration of nuclei which leaves no room for spherulites to develop. However, certain cooling patterns do allow time for some spherulite growth to start on a few heterogeneous nuclei that are stable at a higher temperature. These spherulites will grow to a certain size, but as the temperature falls, the spherulite growth will slow down and many more nuclei will be available for the final slow crystallization into a typical micellar fine structure. The spherulites will be left as occlusions within the fiber and, on drawing, will be transformed from a spherical form to a cigar-shaped ellipsoid. Generally, the conditions of fiber formation are adjusted to avoid this, but it does sometimes occur and can even be beneficial in some applications as a result of particular forms of light scattering.

Voids. In spinning from solution, coagulation commonly occurs before the solvent has been removed completely from the fiber. The material precipitates in a spongy form containing voids filled with solvent. In subsequent processing, these voids collapse as solvent dries off, and they become elongated as the fiber is stretched. The void surfaces seem to disappear to give a solid fiber, but a vestigial effect will remain with the fiber divided into fibrillar zones separated by weaker divisions. This can show up in ease of fibrillation and in a granular surface in tensile fracture.

Cells. Some natural fibres, such as wool and plant fibres like flax, but not cotton, which is a single plant cell, are multicellular. They are formed by the growth of separate cells, which are held together by chemical or physical interactions. Between the cells of wool, there is a cell membrane complex containing lipids and proteins.

2.6. Gross Inhomogeneities

Some man-made fibers have essentially the same structure over their whole cross section, but many textile fibers contain major sub-divisions. A clear example is wool, illustrated schematically in Figure 17, in which the core is bisected into two components having different chemical structure and physical fine structure, with a third form of structure present in the surface layer. Such effects in natural fibers derive from genetic control.

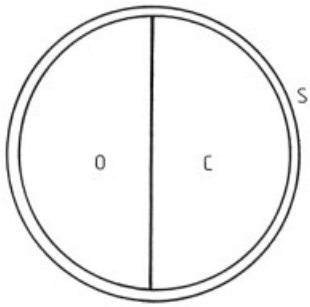

Figure 17. Schematic view of subdivisions of wool fiber: ortho-cortex (O), para-cortex (C), and surface scales (S)

In the less deterministic and faster industrial production of fibers, differences between skin

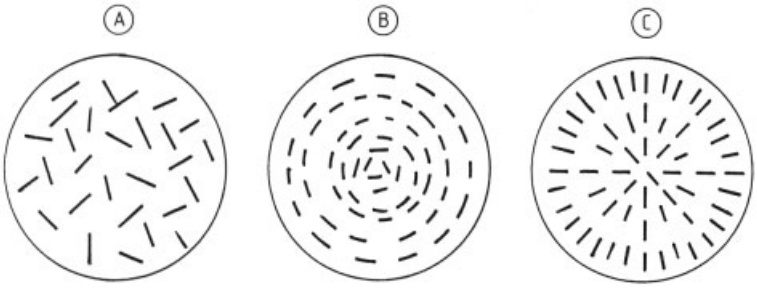

Figure 18. Orientation of ribbonlike molecules in directions perpendicular to the molecular axis
A) Random; B) Radial; C) Circumferential

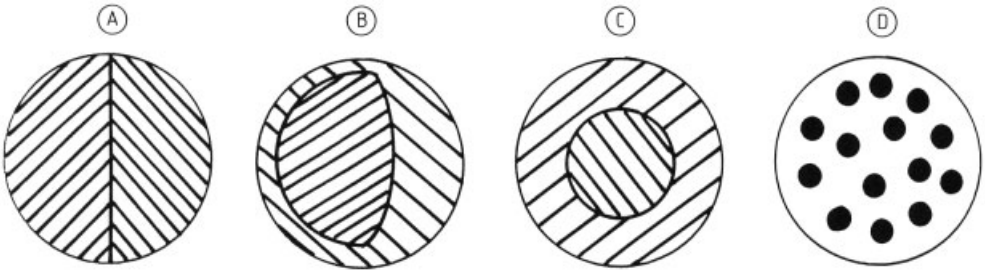

Figure 19. Bicomponent fibers
A), B) Side-by-side; C) Sheath-core; D) Islands-in-a-sea

and core are hard to avoid, since they experience differences in cooling, in evaporation or coagulation (if spun from solution), in nucleation, and in stresses. The patterns of crystallization are consequently different. Where molecules or crystal lattices are asymmetric in cross section, with a flat ribbonlike form which may be due to cyclic groups, the orientation within the cross section can be random, radial, or circumferential as indicated in Figure 18.

Man-made fibers can be deliberately produced as multicomponent, usually bicomponent, forms. Side-by-side fibers (Figs. 19 A and B) will develop crimp as a result of differential shrinkage. Sheath-core fibers (Fig. 19 C) are used when some special feature, such as bonding on heating, is required in the surface layer. The islands-in-a-sea fibers (Fig. 19 D) are used to make very fine fibers by removal of the matrix.

2.7. Fiber Shape

2.7.1. Cross-sectional Shape

The largest scale structural feature of a fiber is its shape. In natural fibers, this can be specialized as a result of the genetic control of growth. However, if a fiber is melt-spun by extrusion through a circular orifice, it will have a simple circular cross section. In spinning from solution, the volume reduction usually modifies the cross section. First, solidification occurs on the outside of the fiber; then the skin has to collapse as solvent is lost from the interior. Various degrees of modification are illustrated in Fig. 20. The circular form occurs when there is no loss of volume or when the chemistry of the solidification process is controlled so that it occurs uniformly throughout the fiber. When a skin forms first and a subsequent small reduction in volume occurs, a dog-bone cross section will result; but when there is a large volume reduction, highly serrated forms are produced.

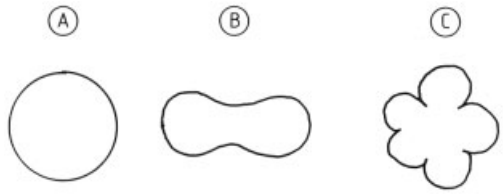

Figure 20. Increasing modification of fiber shape
A) Circular (unmodified); B) Dog bone; C) Serrated

Modification of fiber shape is often desirable to change the luster or other properties of a fiber. This can be achieved in melt spinning by extruding through noncircular orifices and ensuring that solidification occurs before the fiber reverts to the equilibrium circular form. For example, extrusion through a triangular hole will lead to a rounded, three-sided shape, as suggested in Figure 21 A; but extrusion through a Y-shaped orifice will give a higher degree of modification of shape, as in Figure 21 B. The shape of the fiber can be controlled, within limits, by altering the shape of the hole and by adjusting the balance during the cooling of the molten fiber between surface tension, which favors a circular form, and viscosity, which resists change of shape. Octalobal, star-shaped fibers are another common form produced in this way.

Greater bulk can be achieved in fibers by making them hollow.

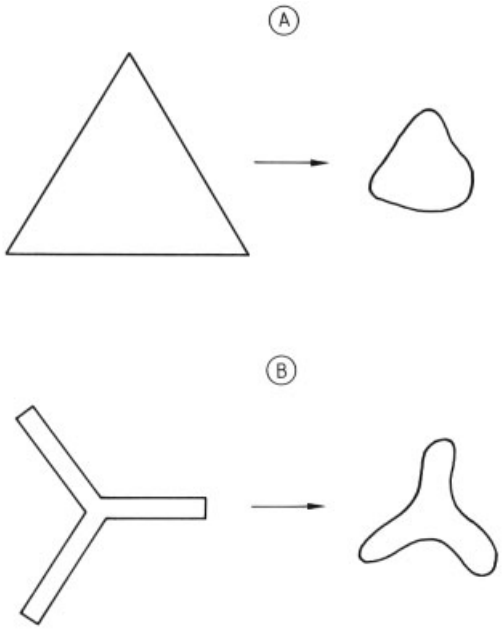

Figure 21. Fiber shape modification
A) Moderate shape modification with a triangular hole; B) Greater modification with a Y-shaped hole

2.7.2. Axial Shape

Most natural fibers have a natural *crimp*, namely, a waviness in axial shape, which is beneficial in textile processing and gives bulk, warmth, and softness to the final product. In cotton, the crimp is at a low level and somewhat irregular, but wool has a pronounced helical crimp like a spring. Simple industrial production of fibers will give straight fibers without crimp, which has to be introduced for many applications.

For some markets, the need for high crimp justifies the expense of *bicomponent fiber* production. Differential shrinkage then causes the fiber to develop curvature, as indicated in Figure 22 A. The preferred lowest energy form for the fiber would be a tight coil, shown in Figure 22 B; but axial restraint would cause this to open into an extended coil (Fig. 22 C). However, torsional restraint is also present. In order to coil up, as in Figures 22 B and 22 C, the end of the fiber would have to rotate. Since such twisting is not possible in a long continuous fiber length, the system accommodates by going into alternating left-handed and right-handed helices, as shown in Figure 22 D. The typical form of filaments in a bicomponent crimped product consists of short sections of helical crimp in opposite directions joined by reversals.

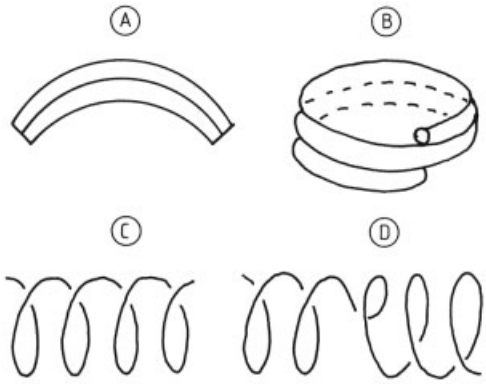

Figure 22. Crimp development resulting from differential shrinkage in a bicomponent fiber
A) Preferred crimp curvature; B) Minimum energy coiled form; C) Extended coil; D) The real form, namely, alternating helices

Texturing. Simple multifilament or monofilament yarns can be used in some applications, as dense, lustrous, silky fabrics, with the filaments remaining straight and without crimp. However, a substantial part of the textile industry is devoted to changing the axial shape

by texturing [58]. In one technique applicable to any type of multifilament yarn, the inherent fiber shape is not changed, but *projecting loops* are formed and trapped within a twisted or interlaced yarn structure. All other techniques involve heat-setting thermoplastic filaments, such as nylon and polyester, in contorted shapes. The dominant techniques are *false-twist texturing* for fine textile yarns and *jet-screen bulking* for coarse carpet yarns. *Stuffer-box, knit – deknit*, and *edge crimping* are among the other methods that have been tried.

In *twist texturing*, the yarn is set in a highly twisted state (Fig. 23 A) and then untwisted (Fig. 23 B). The filaments thus have a natural helical shape and are also under torsional stresses. When the untwisted yarn is allowed to contract by 10 – 20 % (Fig. 23 C), the internal stresses are relieved by the filaments' forming alternating helices, which can be stabilized by a second heat-setting step if a yarn with high bulk but low stretch is needed. A single-heater process gives high-stretch yarn in which the filaments contract to highly snarled forms in order to relieve torque (Fig. 23 D).

In the *jet-screen process*, the yarn is fed into a turbulent steam jet and then impacted on a screen where it is heated and cooled in order to set the filaments in complex buckled forms.

In the manufacture of man-made fibers intended for staple yarns, the tows are normally crimped before being cut into short fibers. One method involved passing the hot tow between intermeshing gear wheels, so that the filaments were set in an axial shape by the serrations. In stuffer-box crimping in which the tow is forced into a confined space that causes the filaments to buckle into a tight zigzag form; this leads to a planar sawtooth crimp.

3. Particular Textile Fibers

Although other articles in this encyclopedia deal in detail with different types of industrial fibers (→ Fibers, 4. Synthetic Organic; → Fibers, 5. Synthetic Inorganic; → Cellulose; → Wool), the preceding sections clearly demonstrate the great diversity of structure even within the limited group of general-purpose textile fibers. Consequently, to avoid leaving the subject of fiber structure in a state of uncertainty, the structure of specific materials should be summarized briefly.

3.1. Cellulosic Fibers

3.1.1. Chemistry

The repeat unit of the cellulose molecule is shown below, although the possibility that occasional groups might have a different form (i.e., be derived from a different monosaccharide) cannot be eliminated.

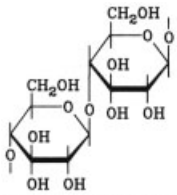

Cellulose [9004–34–6]

The molecule is ribbon shaped but, because of the oxygen (−O−) links, has some flexibility, at least for bending in one plane and twisting. The hydroxyl (−OH) groups readily form hydrogen bonds, which give a relatively strong attraction between neighboring chains and also lead to water absorption. In natural cellulose, the degree of polymerization is very high, but it is much less in regenerated cellulose.

A point of specialized interest is that the chain is directional, as a result of the asymmetry of the ring structure. Crystallization can occur in different forms, and controversy has existed about the precise crystal lattices. The belief now is that cellulose I, which is the usual form in natural cellulose, has all the chains aligned in the same direction (i.e., parallel), whereas cellulose II, which is formed on mercerization with caustic soda and is also the usual form in regenerated cellulose, has neighboring chains in opposite directions (i.e., antiparallel). Cellulose III and IV occur after special treatments.

The relatively stiff chain and the strong hydrogen bonding mean that dry cellulose has a high second-order transition temperature. Indeed, this temperature is above 200 °C, where measurement is not possible because the fiber chars. However, the absorption of water results in mobility, and in the wet state, the transition temperature is below room temperature. Wetting

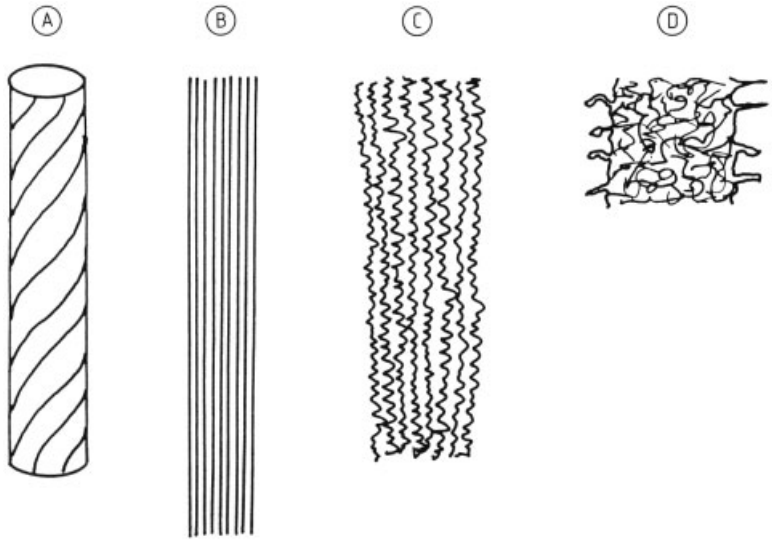

Figure 23. Structural forms in twist texturing
A) Yarn as set; B) Untwisted yarn held to length under tension; C) Yarn contracted by 15 %; D) Yarn fully contracted

and drying thus constitute a mechanism for setting.

In cellulose triacetate, the $-OH$ groups are replaced by acetate ($-OCOCH_3$) groups. For the more common secondary cellulose acetate, about one-sixth of the acetate groups are more or less randomly hydrolyzed back to hydroxyl groups.

3.1.2. Cotton

Cotton [39] is a unicellular fiber in which an outer primary wall first grows out from the seed coat to the final dimensions of the cotton fiber and then is filled in with a secondary wall. Cellulose is synthesized within the cell plasma by the addition of glucose molecules to growing chain molecules via an enzyme complex. About 30 molecules grow together and naturally form crystalline fibrils with chains in the parallel array of cellulose I. At the level of fine structure, the best model is an assembly of microfibrils, with the disorder coming from a lack of register in packing the fibrils together, and the density reduced by some space between fibrils. The equivalent crystallinity is about two-thirds. One unanswered question is what causes the fibrils to retain a clear identity instead of coalescing into larger crystals. Beyond the fine structure is a rich complexity of structural detail.

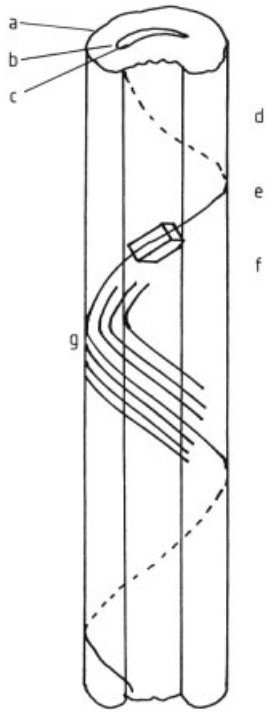

Figure 24. Schematic of major features of cotton structure
a) Primary wall; b) Secondary wall; c) Lumen; d) Molecule; e) Crystal lattice; f) Fibrils; g) Reversal

The thin primary wall forms with the fibrils wound in a specific array, but the structure is dominated by the thick secondary wall. The development of structure is shown schematically in Figure 24. The fibrils form lamellar sheets that wind around the inside of the cylinder at an angle which changes slightly through the thickness, but is about 21 °. At intervals, the helices reverse from right-handed (*Z*) to left-handed (*S*). Daily growth rings are apparent.

The structure at all levels is important in determining the properties of cotton fibers. Thus, moisture absorption, swelling, and accessibility to dyes occur on and between the fibril surfaces. The stress-strain curve can be explained as a deformation of the crystal lattice, with the modulus reduced by a factor approximating $\cos^4 \theta$, resulting from the helix angle θ, and additional extension occurring because of rotation at the reversals and the straightening out of gross convolutions [15].

3.1.3. Other Plant Fibers

Other plant fibers, such as flax, jute, and ramie, which stiffen stems, and sisal, which stiffens leaves, are multicellular and may contain more non-cellulosic impurities [37]. The fibrillar model of fine structure is applicable, but the geometry of arrangement in the cells is different. Usually, the helix angle is smaller, around 10 °, so that the fiber modulus is higher than in cotton.

3.1.4. Rayon

Cellulose fibers can be regenerated through several solvent systems, and these do produce structural differences [33]. The most important process is the viscose process in which a solution of sodium cellulose xanthate ($-$OH replaced by $-OCSS^-Na^+$) in caustic soda is extruded into an acid bath. Viscose rayon is of considerable structural interest because control of the chemistry of regeneration enables fibers with different structure, properties, and uses to be made. Figure 25 gives a schematic indication of the differences.

The degree of order in regenerated cellulose is about half that in natural plant fibers, with a crystallinity of about one-third. In most forms, the fine structure is assumed to be fringed micellar. The addition of zinc ions to the acid bath gives improved properties. Two regeneration routes then occur. On the surface of the fiber, the zinc salt forms first, followed by reaction with the acid to give cellulose. This yields a fine micellar texture, with small crystallites well-linked by tie molecules, which has good strength. The zinc ions do not reach the core fast enough to prevent the direct regeneration of cellulose from the sodium salt. The resulting structure is coarser in texture with larger crystals further apart, which results in a larger proportion of free ends to links, and greater stress concentrations; consequently, it is generally weaker. The collapse of the skin leads to a highly serrated cross section, with the ribbonlike molecules and crystals forced into a transverse orientation that is locally perpendicular to the surface. This is the common form of *textile viscose rayon.*

When the zinc concentration in the bath is increased and additives are included in the viscose solution that slow down the direct regeneration reaction, the whole fiber can be forced into the strong, finer textured structure. This, together with some attention to degree of polymerization and the elimination of a low molecular mass fraction, is the basis for the *high-tenacity tire-cord rayons.* The more uniform coagulation also replaces the serrated cross section by a roughly circular cross section.

Selection of other additives causes the fiber to burst open after forming a skin. This gives a bicomponent fiber with a thicker skin on one side than on the other. Such fibers naturally crimp on drying. The best *carpet fibers* combine some aspects of the high-tenacity route with the crimped route.

In all these processes, the coagulation results from the regeneration of cellulose, which crystallizes into an unoriented fiber. This is then stretched to promote orientation, to a greater extent in industrial fibers than in fibers for clothing.

When the viscose solution is extruded into a weak acid bath, such as sodium dihydrogen phosphate, the sequence changes by first coagulating a fiber of the sodium cellulose xanthate, then orienting it by stretching, and finally regenerating the cellulose. This gives the so-called *polynosic* or *high wet-modulus rayons*, which are more cottonlike in character. Crystallization

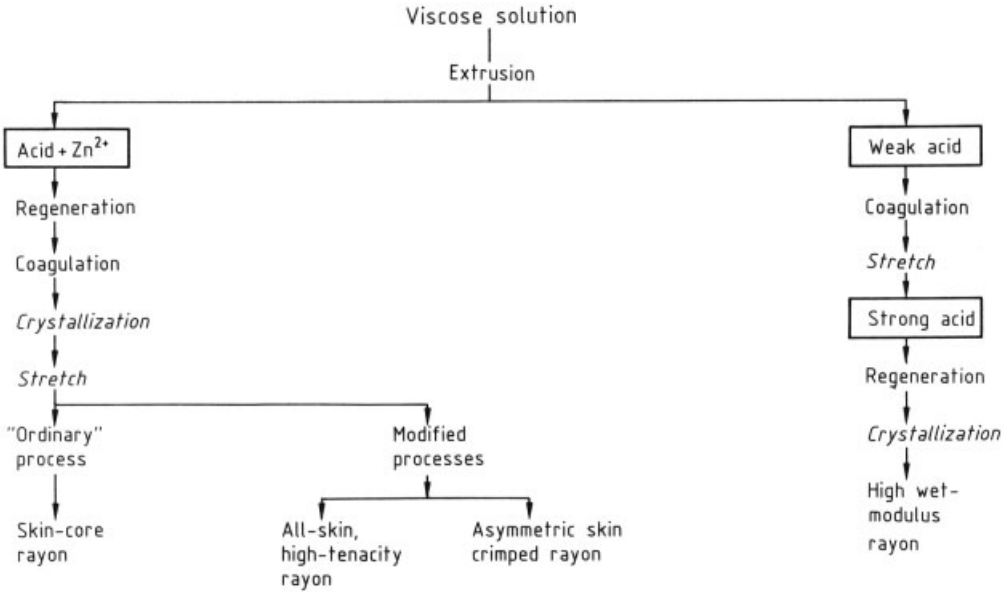

Figure 25. Schematic of routes to different types of viscose rayon

occurs in what is already a solid, oriented fiber, and this results in a fibrillar structure. The fibrils are not as well defined as in natural fibers; they consist of long crystalline units, with molecules fringing off into amorphous regions.

The late 20th century led to another means of spinning regenerated cellulose fibers, which is more environmentally acceptable. *Lyocell* fibers are high-DP cellulose fibers spun from solution in an amine oxide. They also have a fibrillar fine structure (→ Cellulose, Chap. 3.2).

Table 3 gives a comparative summary of the various cellulose fibers in terms of important structural parameters.

3.1.5. Acetate Fibers

Because of their irregular chemical structure, secondary acetate fibers are very poorly crystalline. This shows up in a low softening temperature, low resistance to extension, poor recovery, and low strength. Indeed, this oriented but almost amorphous structure gives properties which would now be inadequate to launch a new fiber, although acetate remains in markets established earlier.

Triacetate is more crystalline and is somewhat similar to fibers like nylon and polyester, although it has a lower melting point and is weaker. The structure is assumed to be micellar.

3.2. Protein Fibers

The repeat unit in proteins is an amino acid residue:

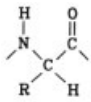

Protein

Diversity results from the selection of a particular sequence of side groups (R) from the usual 20 different types, including two which depart from the form shown in the formula. Proline, a ring of three $-CH_2-$ groups joined to the nitrogen atom in place of hydrogen. Cystine, $-CH_2SSCH_2-$, which is a bifunctional group linking two protein chains together, is produced by the cross-linking of cysteine, in which R is $-CH_2-SH$.

Keratin — a generic name for the material in wool, hair, and other appendages — is a complex mixture of different proteins with significant amounts of all types of side groups. Current understanding identifies a limited number

Table 3. Estimated relative values of structural parameters in cellulose fibers

	Degree of order	Localization of order	Aspect ratio	Orientation	Size	Molecular extent
Cotton	high	high	high	medium	fine	high
Flax	high	high	high	high	fine	high
Ordinary rayon						
core	low	medium	low	medium	coarse	low
skin	low	medium	low	medium	fine	low
High-tenacity rayon	low	medium	low	high	fine	medium
High-modulus rayon	low	medium	high	medium	fine	medium
Lyocell	low	medium	high	medium	fine	medium

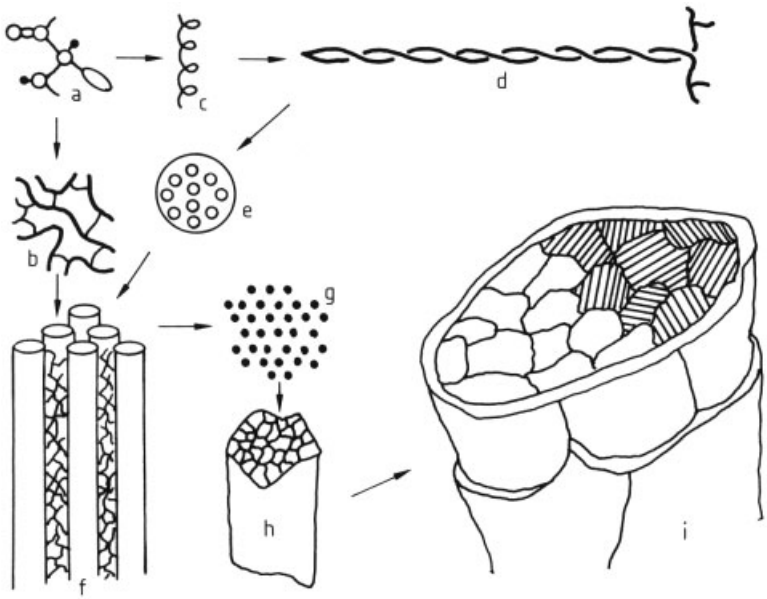

Figure 26. Schematic representation of levels of structure in wool
a) Chemical repeat unit; b) High-cystine molecular segments; c) Low-cystine molecular segment; d) Protofibril with link to matrix; e) Microfibril; f) Fine structure of microfibril and matrix; g) Macrofibril; h) Cell; i) Whole fiber showing ortho- and para-cortex and scale cells

of proteins making up keratin intermediate filaments (microfibrils) and a large number of keratin-associated proteins in the matrix [36, p. 72].

Silk fibroin, which contains no cystine, is a simpler protein and a natural block copolymer with sections limited to a regular sequence of residues with the side groups $-H$, $-CH_3$, and $-CH_2OH$, and other sections with more diversity, including substantial amounts of tyrosine, $-CH_2-4\text{-}C_6H_4OH$.

Proteins of the type found in fibers exhibit two different forms of crystal lattice. The α-form, a perturbation of Pauling's simple α-helix, is a coiled form with about 3.6 amino acid residues per turn. The lattice is stabilized by hydrogen bonds and other interactions between coils in the same chain. The β-form is an extended chain lattice, with hydrogen bonds between neighboring molecules.

The matrix of wool and hair is formed from globular proteins with two main families. One family of high-sulfur proteins is rich in cysteine, which is cross-linked to cystine during the keratinization stage in the growth of hair fibers. The other family consists of high tyrosine/glycine proteins.

3.2.1. Wool and Hair

Wool [36] and other hair fibers [35] show a series of structural features at various levels up to the full-scale fiber, as indicated in Figure 26.

The fine structure consists of crystalline microfibrils, composed of low-sulfur protein (without cystine), parallel to the fiber axis. This is embedded in a rubbery matrix, which is cross-linked by cystine and connected to the fibrils by tails of the low-sulfur molecules. Figure 27 shows the tensile stress – strain curve of a uniform wool fiber in the wet state. The crystal lattice is normally a coiled chain α-helix; under tension it can transform with 80 % extension to an extended chain β-lattice. The interaction with the matrix in the composite system leads to the flat yield region in the stress – strain curve, the stiffer post-yield region, and the characteristic recovery curve [59, 60].

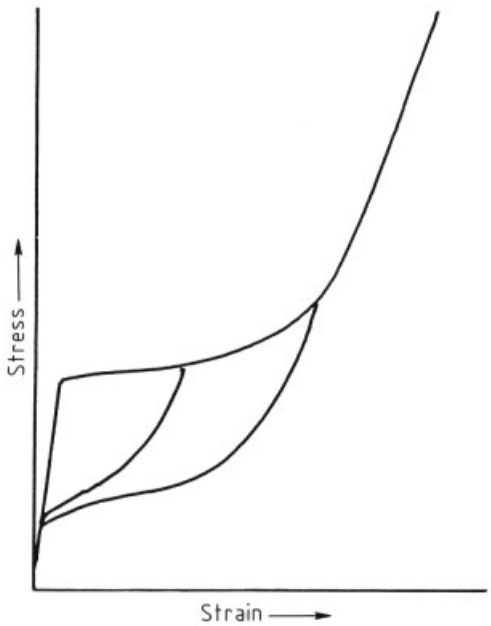

Figure 27. Stress – strain curve of wool

The larger scale macrofibril and cell structures do not have much influence on the properties, but there may be slippage due to debonding at the lipid-rich cell membrane complex between individual cells. However, the division of the main part of the fiber, called the cortex, into an ortho-cortex, with less cystine, and a para-cortex, with more cystine, gives a bicomponent fiber. The microfibrils in the macrofibrils of the para-cortex are parallel to the fiber axis, but in the ortho-cortex the macrofibrils consist of helical assemblies of microfibrils [44]. On drying, the ortho-cortex macrofibrils tend to increase in length, due to the decrease in helix angle as the macrofibril diameter reduces, but the para-cortex macrofibrils tend to remain constant in length. The differential length change leads to a helical crimp, if the components are side-by-side, but a straight fiber if they are symmetrically distributed [61, 62]. A total model proposed for the structural mechanics of wool and hair fibers covers eight levels between molecules and the whole fiber [63].

The scale cells on the surface, with an escarpment facing the tip of the fiber, cause directional friction.

3.2.2. Silk

Silk [35], which does not form as a living cell but is an excretion, presumably evolved into a tough fiber; this would result from a crystallization of the regular sections linked in a rubbery matrix by the less regular sections. The chemical constitution would control the fine structure in a fiber which coagulates after extrusion, to form two triangular filaments held together by gum (→ Silk).

3.2.3. Regenerated Protein Fibers

Fibers have been made from the proteins in milk (casein), corn (zein), and other natural products. Generally, these fibers have a poorly ordered structure with a weakness that has led to a lack of commercial success.

3.3. Major Melt-Spun Fibers

3.3.1. Chemistry

The most important melt-spun fibers [38], dating from the middle of the 20th century, are the polyamides *nylon 66* and *nylon 6*, and the polyester *poly(ethylene terephthalate)* (PET),

with the following repeating units (see also → Fibers, 4. Synthetic Organic, Chap. 1):

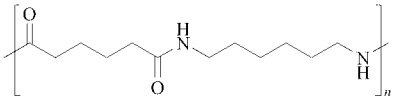

Nylon 66 [*25038-54-4*]

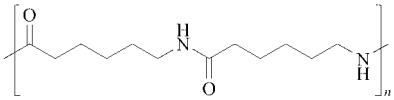

Nylon 6 (two repeats) [*32131-17-2*]

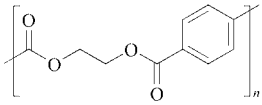

Poly (ethylene terephthalate) (PET) [*25038-59-9*]

More recently, the commercially available syntheses of poly(trimethyleneterephthalate) (3GT) and poly(ethylenenaphthalate) (PEN) have led to the production of these polymers as textile fibers.

Little difference exists between the two polyamides: in contrast to nylon 66, nylon 6 has five and five $-CH_2-$ groups, instead of four and six; has the middle $-CONH-$ group reversed to $-NHCO-$; and consequently, has only half the length of repeat unit. Perhaps the most important feature is that the nylon 6 molecule is directional, so neighboring chains in a crystal must be in proper relation to each other. This might be the reason nylon 6 melts at a temperature ca. 50 °C lower than nylon 66 and PET.

All three polymers have a very long repeating unit, which gives fewer choices of location in crystals than in simpler polymers such as polyethylene, and thus leads to lower maximum crystallinities. The other common characteristic is the presence of a flexible aliphatic sequence with weak interactions. In between these sequences, the polyamides have $-CONH-$ groups which interact by hydrogen bonding, and polyester has benzene rings which are both stiff and interactive.

The crystal lattices of these polymers are simple extended chain forms and are known. In nylon 6, some complications result from the existence of two alternative forms.

The glass-to-rubber transition in nylon and polyester occurs in two stages, which is a valuable feature. A transition below room temperature gives a measure of extensibility and toughness to the fiber, while sufficient stiffness is retained. The remaining inevitable transition to a fully rubbery, amorphous form between crystallites occurs at a higher temperature. The first transition is associated with the onset of mobility in rotation around $-CH_2-$ and similar bonds in the aliphatic sequences. In nylon, the second transition is associated with the hydrogen bonds moving into a state of dynamic equilibrium, so that the chain segments in amorphous regions become completely free. This transition temperature is much lower in wet fibers. In polyester, analogous interactions of the benzene rings occur, but the transition temperature is somewhat higher and is little affected by water. The benzene rings lead to a higher initial modulus in PET than in nylon. 3GT with more aliphatic groups is more like nylon, but the double ring in PEN leads to a stiffer fiber with a higher melting-point.

Another major melt-spun fiber is *isotactic polypropylene* [*9003-07-0*], $\text{+}CH_2CH(CH_3)\text{+}_n$. Because of the bulky side group, the molecule forms a threefold helical coil, which gives it a greater stiffness in the melt or in amorphous regions, but a lower axial modulus in the crystal. Polypropylene has a single broad glass transition near room temperature.

3.3.2. Fine Structure

Despite their major importance as industrial products for about 50 years, surprisingly little is known about the fine structure of melt-spun fibers. The degree of order corresponds to a crystallinity of about 50 %, not much changed by differences in processing. The degree of orientation can be controlled within fairly wide limits by the speed of windup, subsequent drawing, and hot-stretching or relaxing treatment.

As a working model, the structure is assumed to consist of crystalline micelles, partly folded and partly fringed at the ends, and linked by tie molecules. The crystallites are assumed to be stacked in pseudofibrils, as shown in Figure 5. Some authors have suggested significant differences between amorphous material joining crystallites within a pseudofibril and amor-

phous material separating pseudofibrils. However, good quantitative information is lacking about the size and shape of crystallites, their arrangement in some form of superlattice, the relative amount of folds and tie molecules, and the detailed pattern of interconnections between crystallites. Evidence exists that annealing treatment leads to the development of larger crystallites spaced farther apart. If the fiber is free to contract, chain folding may develop.

Consideration of the change in endotherms observed in differential scanning calorimetry indicates that rapid "crystallization" leads to a form I. This has been suggested [56] to be a more uniform structure (see Fig. 9), with small neighboring segments in crystallographic register, but in dynamic equilibrium just below the almost constant melting point observed for form I. Annealing then leads to the formation of small separate crystallites, initially with a lower melting point but a lower free energy at the annealing temperature. Prolonged annealing increases the size and perfection of the crystallites, and raises the melting point.

Despite many similarities, nylon and polyester have significant differences. Cooling of an unoriented fiber, being wound up at a low speed such as 1000 m/min (ca. 20 m/s), leads to crystallization in nylon but not in polyester. Only the subsequent drawing causes the polyester to crystallize. Textile polyester fibers have a higher initial modulus than nylon and split axially at an angle closer to the fiber axis. These differences are presumably associated with the stiffness of the benzene ring, but the structural implications are not understood in detail.

When polyester fibers are wound at a higher speed, around 3000 m/min (ca. 50 m/s), they develop an incipient crystallinity that makes them more stable and easier to process. These are the so-called partially oriented yarns (POY) used as draw-texturing feed yarns. At still higher speeds, sufficient orientation and crystallinity can be induced to give fibers suitable for some applications without drawing. However, the degree of orientation that can be achieved this way is lower than in the two-stage route. The fine structure is also different and seems to contain larger crystallites.

3.3.3. PLA Fibers

The newest melt-spun fiber is PLA [64], which is produced with the tradename NatureWorks. There is less structural information available on this type of fiber. It can be assumed that it is a poly(lactic acid) fiber, which has an ester linkage between the −OH and −COOH groups at opposite ends of the lactic acid molecule. The basic chemical formula is:

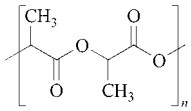

The US Federal Trade Commission definition issued in 2002 states that there must be at least 85 % of lactic acid ester units in the chain, but includes poly(lactate)s, which could contain other units. For example, there could be copolymerisation with ethylene terephthalate or other groups.

3.4. Acrylic Fibers

The remaining major synthetic fiber type has acrylonitrile, $-CH_2CH(CN)-$, as its principal repeating unit [38] (see → Fibers, 4. Synthetic Organic, Chap. 5). However, for dyeability, up to 15 % of a comonomer must be included. Suggestions abound in the patent literature, along with a diversity of practice by manufacturers, about both the type of comonomer and the form of copolymerization. The degree of tacticity of the polyacrylonitrile [*25014-41-9*] is also not well documented but probably atactic. Because of these features, acrylic fibers are rather different in structure from most other types.

The molecule appears to act as a cylindrical rod with strong, but not very localized, attractive forces coming from the dipole in the $-C{\equiv}N$ group. The rods pack regularly side by side, like chalk in a box, but are not in register along their length. The fine structure is thus assumed to consist of such ordered pseudocrystalline regions mixed with less ordered regions. A big second-order transition occurs around 100 °C, when the segments in the less ordered regions overcome the interactions and can slide over one another easily; other transitions then take place,

at higher temperatures, until the structure decomposes while still solid.

3.5. Other Polymers

Other linear polymers have been used to a limited extent in textile fibers. Their fine structures have usually not been studied in detail but are similar to the most nearly corresponding major fibers. Among melt-spun fibers there has been some production of other variant *polyamides* and *polyesters*. Some of these, such as nylon 4 and poly(butylene terephthalate), differ only in the length of $-CH_2-$ sequences. Qiana (Du Pont), marketed as a luxury fashion fiber but no longer made, was a more complicated polyamide.

Polyethylene [*9002-88-4*], $\text{+}CH_2-CH_2\text{+}_n$, fibers are spun from the melt or, in the spunbonded material Tyvek (Du Pont), are flash-spun from solution to give a fibrous network. The structure is highly crystalline, probably with the lamellar form characteristic of crystallization in this polymer; disorder would then arise from the localized defects possible with such a small repeat unit (see → Fibers, 4. Synthetic Organic, and Section 4.2.2). High-modulus polyethylene is dealt with in Chapter 4.

Poly(vinyl chloride) [*9002-86-2*], PVC, $[CH_2CH(Cl)]_n$, is similar to acrylic fibers, as are other vinyl and vinylidene copolymers with chlorine, acetate, or other side groups (see → Fibers, 4. Synthetic Organic, Chap. 6).

Poly(vinyl alcohol) [*9002-89-5*], PVA, $[CH_2CH(OH)]_n$, has special features associated with hydrogen bonding of $-OH$ groups. In one form, the attraction for water leads to solution on washing, but this can be eliminated by appropriate treatment. Full information on PVA fibers is given in [65] (see also → Fibers, 4. Synthetic Organic, Chap. 7).

4. High-Modulus, High-Tenacity Fibers

An important, new generation of fibers has appeared, intended for engineering use either in flexible systems (e.g., ropes and other tension members), in special tires or in rigid composites, and characterized by high stiffness and high strength. Most of these fibers also show high temperature and chemical resistance, and some fibers selected mainly for thermal stability have high stiffness even if not high strength. The choice of category is thus sometimes rather arbitrary, but this section includes fibers selected mainly for mechanical performance, while Section 5.4 includes those selected mainly for thermal and chemical stability. The strong fibers are reviewed in [34, 66].

The advantages of using a material in fiber form are discussed in Section 1.1, but in some cases the structure of the fiber material differs little from other forms of the material. In these cases, only brief comments will be given in this article, and reference should be made to other articles in this encyclopedia for more general and detailed information on structure (→ High-Performance Fibers).

4.1. Glass

The structure of glass fibers is essentially the same as the structure of bulk glass [34, pp. 191–238]. Solidification from the molten state gives an isotropic three-dimensional network based on a tetrahedron of four oxygen atoms around a silicon atom, but made irregular and amorphous by metal ions. A typical composition of E-glass fiber intended for electrical insulation is 52 – 56 % silicon dioxide, 16 – 25 % calcium oxide, 12 – 16 % aluminum oxide, 8 – 13 % boron oxide, 0 – 1 % sodium and potassium oxides, and 0 – 6 % magnesium oxide. Different compositions are used in the newer glass fibers optimized for other markets; for example, the high-strength S-2 glass fiber is low in alkali and high in magnesium. Glass fibers are reviewed in [67] and [68]; see → Fibers, 5. Synthetic Inorganic.

4.2. Highly Oriented Linear Polymers

High-tenacity, high-modulus polymer fibers have a structure that satisfies the three requirements of high order, high orientation and — most important — high molecular extent. In practice, such a structure can be reached only by specialized routes, and this has not been achieved with such common polymers as nylon and polyester (PET). There are two main production methods. The first route requires very stiff

and strongly interactive molecular chains, which inhibit chain folding. In most cases, this group of fibers depends on orientation during extrusion of a liquid crystal solution or melt. The second route involves a "combing out" of the chain molecules by some means such as gel spinning or super-drawing and requires very flexible and weakly interactive molecules.

4.2.1. Aramid

The first high-tenacity aramid fiber was introduced by Du Pont with the name *Kevlar*; other manufacturers now also produce aramid fibers [34, pp. 23–61]. It is composed of poly(*p*-phenylene terephthalamide) (PPTA)

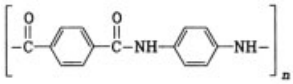

a stiff, rodlike molecule with strong hydrogen-bonding interactions at the −CONH− groups and electronic interaction between benzene rings. In solution, the molecule forms liquid crystals which are easily oriented on extrusion. The resulting fiber has a structure that is very highly oriented and highly crystalline. The molecules are fully extended. The crystal perfection persists only for limited distances, with a packing pattern derived from the rodlike liquid crystals. Suggestions have been made of the occurrence of pleated sheets arranged radially in the structure. However, specialized articles should be consulted for such details [69, 70], (→ High-Performance Fibers).

Teijin produces another aramid fiber called *Technora* [71], but this is a copolyterephthalamide of *p*-phenylenediamine and 4,4′-diaminodiphenyl ether, and is a 50 mol % copolymer of

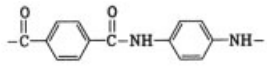

and

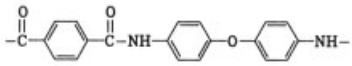

Evidence exists for the presence of the crystal structures of both homopolymers. In contrast to PPTA fibers, the highly-oriented, chain-extended structure is not formed from a liquid-crystal solution, but results from drawing and heat-treating of fiber coagulated from an extruded isotropic solution.

Other aramid and related aromatic fibers with similar structures have been produced in Russia [34, pp. 115–132].

4.2.2. Polyethylene

Polyethylene, which has the simple $-CH_2-$ repeat unit, can be produced as a high-strength, high-modulus fiber that is highly crystalline and highly oriented, and has the molecules highly extended with a minimum of chain folding. The polymer will always be linear, but the degree of polymerization depends on the production method.

The highest strength and stiffness is produced by gel spinning from a fairly concentrated solution of ultrahigh molecular mass polymer, which allows for a high degree of subsequent elongation [34, pp. 62–92]. Melt-spinning and super-drawing is a cheaper process but requires lower molecular mass, so that the resulting properties are not as good. Solid-state extrusion of a crystal mat is another method of achieving the required structure [34, pp. 132–144]. Three reviews give many details [72 – 74]; see also → Fibers, 4. Synthetic Organic.

4.2.3. Other Polymers

Several other fiber types are produced by the liquid-crystal route.

Vectran, which is melt-spun and then heat-treated to increase molecular mass, is a fully aromatic co-polyester with the formula:

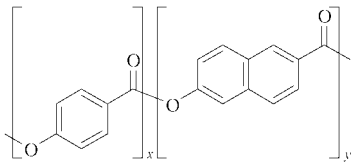

Other research, aimed at superior properties rather than lower cost, is utilizing some newly synthesized materials such as stiff-chain aromatic heterocyclic polymers (with the general designation PBZ) based on benzothiazole (PBS)

or benzoxazole (PBO) ring structures. Typical repeat units are

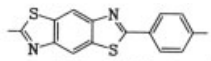

PBS [95-16-9]

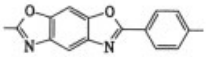

PBO [273-53-0]

PBO is now commercialized as *Zylon* [34, pp. 101–108]. Although the molecules of these fibers are highly oriented and packed side by side in a regular array, they are not necessarily in crystallographic register along their length.

Aramid and PBO fibers have relatively weak intermolecular bonding and consequently have low shear resistance and fail in compressing by internal buckling. Sikkema [34, pp. 108–115], [75] has synthesized the new polymer M5, which is made from tetraaminopyridine and dihydroxyterephthalic acid, and prepared experimental fibers. Projecting OH groups lead to hydrogen bonding in both directions, as distinct from the planar hydrogen bonding in the para-aramid fiber, with a resulting improvement in shear and compression resistance in fibers. The chemical formula is:

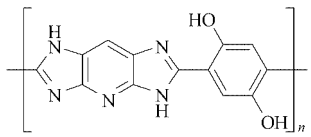

There have been reports on studies of poly(vinyl alcohol) fibers produced by gel spinning or superdrawing, and other materials are being investigated.

A third special route has been applied to polydiacetylenes, which have the following general formula:

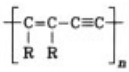

where R represents a large number of possible side groups.

Single-whisker crystals of the precursor diacetylene can be grown, which is then polymerized in the solid state. The resulting fiber is essentially a perfect polymer single crystal.

Another route that has been applied to polyethylene, is virgin polymerization in which the orientation, crystallinity, and molecular extent are established at a point of polymerization of chains.

4.3. Carbon Fibers

When a suitable organic precursor fiber is subjected to a sequence of heat treatments, ending with a very high temperature under tension, a high-modulus carbon fiber is produced with a quasi-graphitic structure [34, pp. 156–190]. Depending on the starting material, usually polyacryonitrile fibers or pitch fibers, and the details of the heat treatments, considerable variations in structure occur, which lead to a range of combinations of strength and stiffness.

Pure graphite crystals consist of a stack of planar sheets of carbon atoms in a hexagonal array. Most carbon fibers are turbostratic, with neighboring parallel sheets not in crystallographic register. However, any assembly of parallel sheets would be useless in fibers due to lack of lateral cohesion, which leads to the use of graphite as a lubricant. Some disorder, which links sheets, is needed. Carbon fibers are a disturbed and distorted assembly of graphitic crystalline elements, for which many different pictures have been drawn. The sheets are arranged in separate layers that are bent and folded within the structure. It seems likely that the conversion of the original linear polymers into two-dimensional sheets may have developed in different regions, which leads to occasional cross-linking C−C bonds between neighboring portions of sheets to give a sparse three-dimensional network. An important feature for high modulus is that the planes should all be aligned closely parallel to the fiber axis. This still leaves two degrees of freedom of local orientation of a plane. Rotation about axes perpendicular to the plane allows neighboring planes to be at different angles in a turbostratic structure. Rotation of the planes around a line parallel to the fiber axis, which must occur cooperatively between neighboring sheets to maintain close packing, allows for a variety of transverse orientations within carbon fibers. Radial, circumferential and other patterns occur.

Carbon fibers are dealt with in several chapters in [66], as well as in [76] and [77]; see also → Fibers, 5. Synthetic Inorganic.

4.4. Ceramic and Other Inorganic Fibers

Quartz fibers have a long history in scientific instrumentation, but the production of fine ceramic fibers with diameters of the order of 10 μm for use in high-temperature composites was first commercialised in the 1980s. Ceramic fibers, are covered in several chapters in [26, 51, 80]. The account below is based on the more recent publication [34, pp. 239–258] (see also → Fibers, 5. Synthetic Inorganic, Chap. 4).

Since the 1980s, there have been considerable development and the market has stabilized with silicon carbide based and alumina based fibers as the main types. The fibers have a structure which is typical of polycrystalline ceramics, but it has been shown that small differences in the fabrication processes lead to considerable differences in microstructure.

One group of silicon carbide based fibers (see also → Fibers, 5. Synthetic Inorganic, Chap. 3) are made by treatments of organosilicon polymer fibers. A typical *Nicalon* fiber from Nippon Carbide contains 56 % Si, 32 % C and 12 % O. β-SiC is present in grains of about 2 nm size, with the remaining 45% of the fiber composed of a metastable Si-C-O phase, which impedes grain growth and allows creep to occur, and free carbon particles. In *Hi-Nicalon* the oxygen content is reduced to 0.5 % or less but the fiber contains more free carbon. The SiC grains are 5 to 20 nm in size, surrounded by less crystalline SiC and turbostratic carbon. *Tyraano Lox* fibers from Ube Industries include about 2 % of Ti with 55 % of Si and varying amounts of C and O and have microstructures similar to the high-oxygen Nicalon *fibers*. Near-stoichometric SiC fibers, which contain more than 95 % SiC, are made by sintering processes. Grain sizes are of the order of 200 nm with pockets of free carbon between the grains.

Alumina fibers are derived from the extrusion of viscous solutions of aluminum salts, which are processed to give over 99 % Al_2O_3 fibers. Up to around 1000 °C grain sizes range from 10 to 100 nm, but above 1100 °C rapid growth of micron-sized α-alumina grains occurs and the fibers become brittle. Small grain size is maintained in alumina – silica fibers. For example, *Altex* from Sumitomo contains 15 % of SiO_2 and consists of 25 nm γ-alumina grains surrounded by amorphous silica. *Nextel* fibers from 3M contain B_2O_3 and have the composition of mullite. With 14 % B_2O_3, *Nextel* 312 is described as quasi-amorphous, but *Nextel* 440 with 2 % B_2O_3 consists of γ-alumina in silica. The presence of silica leads to more flexible fibers, which creep and lose strength above 900 °C. *Nextel* 650 is an alumina – zirconia fiber, which contains 10 % of ZrO_2 and 1 % of Y_2O_3.

Large diameter ceramic "fibers" can be made by vapor deposition on a substrate such as tungsten or carbon filament. Another type of ceramic fiber consists of monocrystalline inorganic crystal whiskers. Ceramic fibers, are covered in several chapters in [26, 66, 77]; see also → Fibers, 5. Synthetic Inorganic.

4.5. Structure and Properties

By definition, all the fibers included in this section have a *high modulus* and most have a *high strength*. Although each fiber type should be examined separately, a broad generalization is that the linear polymers give very high strength but somewhat lower modulus; that the carbon fibers can have high strength and the highest moduli; that the ceramic fibers have lower strength but high modulus; and that the amorphous structure of glass leads to lower values of both strength and modulus.

However, the structural differences cause big differences in some other properties. The highly oriented linear polymer fibers inevitably have a low *compressive yield stress*, since only weak forces exist between molecules and the chains can buckle within the fibers. They are prone to splitting into fibrils and usually break by long axial splits. By contrast, glass and ceramic fibers, which have a three-dimensionally bonded structure, and carbon, which is also coherent in all directions, show similar behavior in compression and tension. They usually break cleanly across the fiber.

Although the three-dimensional fiber structures are good in compression, the same effect causes the fibers to be poor in *flexibility* because

they snap easily. On bending, the outside is extended and the inside is compressed; the smallest radius of curvature R that can be achieved is given by $R = r/e$, where r is the fiber radius and e the fiber breaking strain. Therefore, the tightest achievable curvature is more than 20 times the fiber radius. In contrast to this, the linear polymer fibers can be bent back on themselves to nominally zero values of curvature without breaking, because yielding on the compression side allows the neutral plane to move out so that the extension on the outer surface does not exceed the fiber breaking extension. Another benefit, arising partly from this and partly from the axial splitting, is that linear polymer fibers are more effective in absorbing energy. They are the best materials for ballistic impact resistance fabrics, in which glass or carbon is useless, and generally show greater damage tolerance in structures.

The other important difference is in *thermal resistance*, which depends on chemical composition. Polyethylene melts at a low temperature, and shows excessive creep and loss of mechanical properties even at 100 °C. The aramids and the newer rigid rod polymers do not melt but decompose at around 500 °C. Glass begins to soften at similar temperatures. The ceramic and carbon fibers withstand very high temperatures.

5. Other Fibers

The two major groups of fibers are the general-purpose textile fibers, whose structure is described in Chapters 2 and 3, and the high performance engineering fibers, described in Chapter 4. However, a miscellaneous group of fibers is also used for a variety of purposes to be mentioned in this section. Generally, the main structural interest is in the chemical composition, and other structural details are either little known or similar to the most common forms of the particular material, which are discussed in other articles in this encyclopedia.

5.1. High-Toughness Fibers

Some engineering uses of fibers require high toughness, shown by a high work of rupture (energy to break, or area under stress – strain curve), usually accompanied by a fairly high modulus. The structure of these fibers is described in Section 3.3; they include the melt-spun polyamides, polyester, and polypropylene. However, the industrial yarns are processed so as to maximize the degree of orientation.

5.2. Elastomeric Fibers

The elastomeric fibers are at the opposite extreme to the high-modulus fibers. Elastomeric fibers are intended for applications in which high stretch is required; the breaking extension may be as high as 500 %, and the modulus is inevitably low.

Natural rubber, either in the form of thread cut from sheet or as extruded latex, was the first material to be used. It consisted of thick monofilaments or bundles of monofilaments. Their structure is the usual cross-linked amorphous network formed by vulcanization.

However, modern elastomeric fibers are mostly spandex fibers composed of segmented polyurethanes and able to be made in typical textile filament diameters. The polymer is an alternating block copolymer, in which "soft" segments provide the high extension by changing from coiled to extended forms, and the "hard" segments interact and form small crystalline or quasicrystalline domains to hold the structure together. The fine structure is thus an unoriented, fringed micelle form with a low degree of crystallinity (high fraction of a rubbery amorphous phase). Under low or moderate tension, the mechanism of rubber elasticity will allow high stretch by the uncoiling of soft segments, but under high tension the alignment of soft segments may lead to some crystallization that reinforces and strengthens the structure.

In contrast to natural rubber, in which chemical cross-links join chains in pairs, the clustering of connections in the hard segment domains is an important structural feature for good mechanical properties.

In principle, many possible chemical forms exist for soft and hard segments, and even with a particular choice of components the proportions may be varied. Couper [78] reports a typical monomer for a soft-segment block to be

$$\mathrm{OCN{-}[A]{-}NHCO}\left(\mathrm{{-}O{-}[B]{-}CONH{-}[A]{-}NHCO}\right)_n\mathrm{{-}O{-}[B]{-}OCONH{-}[A]{-}NCO}$$

where OCN−[A]−NCO is an aromatic diisocyanate, O−[B]−OCO is the polyether or polyester chain of a macroglycol, and n is a small number varying from molecule to molecule.

This monomer is then linked by bifunctional chain extenders, usually diamines, to the hard segments. A typical hard segment is reported to be

$$-\!\!\left(\begin{array}{l}\text{CONH}-[\text{A}]-\text{NHCO}-\text{NH}-[\text{C}]-\text{NH}\\ -\text{CONH}-[\text{A}]-\text{NHCO}\end{array}\right)\!\!-_m$$

where [C] is the aliphatic sequence in a diamine and m is an integer.

5.3. Thermally and Chemically Resistant Fibers

Apart from polyethylene, all the high-modulus, high-tenacity fibers described in Chapter 4 have good thermal and chemical resistance. This is particularly so for ceramic and carbon fibers. However, other fibers, some of which have only moderate or poor mechanical properties are used for insulation protection, filtration, chemical engineering functions, or reduction of fire hazards [34, pp. 259–324].

Asbestos, which is very brittle and now recognized to be medically dangerous, was once widely used. It is an inorganic mineral of structural interest because its fibrous form derives from the fact that the unit cell of the crystal is naturally curved in a small radius. The crystals grow by forming cylinders of many layers of rolled-up sheets, limited by the extent of allowable distortion as the inner and outer radii depart from the preferred value. The space in the center of the cylinder and between cylinders is filled with secondary crystallization, in which the cylinders are incomplete (→ Asbestos).

Rock wools formed from a melt dispersed into fiber form have a polycrystalline structure.

Several linear organic polymers are used to make resistant fibers with reasonable textile properties. One route is to choose rather stiff chains with strong interactions, which have a high melting point and a chemical structure selected to avoid flammability and prevent decomposition at too low a temperature. One such polymer is the m-aramid, poly(m-phenylene isophthalamide) [*9016-75-5*], marketed as Nomex (Du Pont):

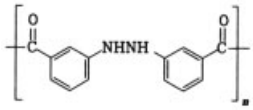

The shape of the chain makes it easier to form into fibers than the p-aramid, and leads to a more extensible fiber with good textile properties.

Various arimid fibers have been produced experimentally. The one that is commercially available is *P84*, which was developed by Lenzing but is now produced by Inspec Fibers. Its composition is:

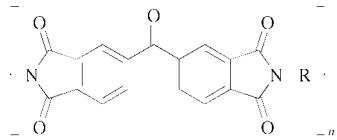

where R is $C_6H_4CH_2$.

Kermel from Rhône-Poulenc is a poly(aramide-imide) fiber. The chemical structure is reported as:

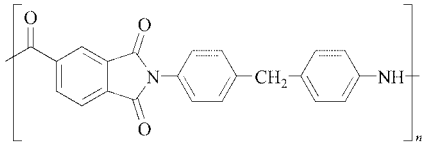

A fiber with higher temperature stability and high moisture absorption, is poly[2,2′-(m-phenylene)-5,5′-bibenzimidazole] or PBI [79], with the following repeat unit:

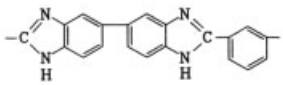

This is stabilized by forming a salt with sulfuric acid.

The fine structures of these fibers are similar to those of other crystallizable polymers spun from solution.

If the route to carbon fibers is stopped after the first stage of treatment of acrylic fibers, thermally resistant, partially carbonized fibers are produced. The disorganized structure of *Panox* fibers leads to comparatively low strength, but textile properties which give reasonably comfortable flame-proof protection. Further heat treatment gives improved thermal resistance but greater brittleness, as shown by Curlon fibers.

If good textile properties are sacrificed, thermoset resins with a characteristic amorphous, three-dimensional, polymer network structure provide thermally resistant fibers. The novoloid fiber *Kynol* (formerly made by Carborundum now by Kynol Corporation) is a thermoset phenolic resin, with the structure being a three-dimensionally cross-linked network, typified by forms such as:

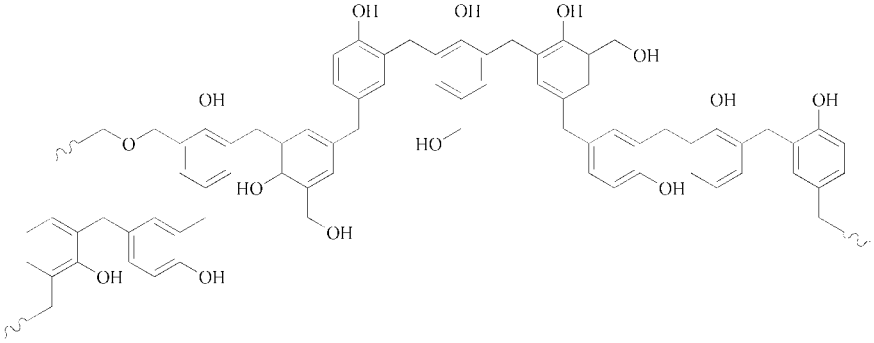

Basofil from BASF is a melamine-formaldehyde resin.

Polytetrafluoroethylene [*9002-84-0*], PTFE, is a very inert and stable polymer with the simple repeat unit $-(CF_2)-_n$. It has good chemical resistance at high temperatures. The coiling of the molecules in shallow turns stiffens the chains and leads to a high glass transition temperature and stable crystals, which show some lattice transitions. The polymer decomposes at a high temperature before it melts and is insoluble. Consequently, it must be spun from a dispersion and then consolidated by sintering, which is reflected in the structure (→ Fibers, 4. Synthetic Organic, Chap. 8).

Various other polymers with good chemical resistance can be melt-spun and would be expected to have structures similar to other melt-spun fibers; copolymers would have lower crystallinity. Poly(vinylidene chloride) (PVDC, $-CCl_2CCl_2-$) can only be spun when copolymerized with other vinyl or acrylic monomers. Fluorinated polymers, poly(vinyl fluoride), poly(vinylidene fluoride) and fluorinated ethylene polymers and copolymers are another group of materials.

PEEK, poly(etheretherketone) is melt-spun at high temperature by Zyex in a range of cross-sections and sizes. The chemical formula is:

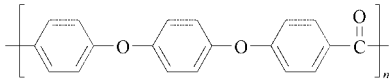

Other polyetherketones have not proved commercially successful.

Poly(phenylene sulfide) [29] fibers are produced by a number of companies. The chemical formula is:

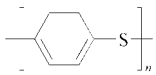

Poly(etherimide) fiber is another chemically resistant fiber that can be melt-spun. However, its complicated chain form brings about an amorphous structure. The chemical formula is:

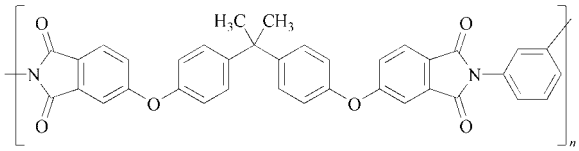

5.4. Special-Purpose Fibers

Many fibers are made for a variety of special purposes, but only some brief indications of chemical and physical structure can be given. A number of these fibers are described in [27].

Fibers that dissolve in water, possibly hot or with soap or other common chemicals present, have some uses. Such fibers can be made from poly(vinyl alcohol). $\text{+CH}_2\text{CH(OH)+}_n$, or from calcium alginate derived from seaweed. A number of fibers for biomedical applications are made from linear polymers selected for their beneficial interaction with the body [81].

Many fibers are naturally absorbent, and this can be enhanced by modifying the chemical or physical structure. A particular example is the conversion to *activated carbon fibers* with many accessible microvoids. Unlike the high performance carbon fibers, those fibers have not been subjected to high-temperature graphitization.

Particular structural features may also occur on a large scale. *Hollow fibers*, made of a suitable material, can be used for chemical separation processes or the controlled release of reagents.

Metal fibers may be extremely fine-drawn wires used for special purposes, such as tungsten filaments in light bulbs or fibers for leakage of static electricity in carpets. A decoratively used fiber is *Lurex* (Sildorex), which consists of aluminum anodized to give color, cut into fine strips from foil, and laminated with plastic protection. These materials have the usual fine structure of metals. However, a special structure, which can only be found in fibers or thin films, is the amorphous structure of metallic glasses formed by rapid chilling [82]. The blocking of defect mobility leads to a very high yield stress.

Materials with special electrical or optical properties can be made in fiber form. For example, fibers with the structure of electrets are useful in filtration. The most important application in this category consists of optical fibers for communication. These are usually made of specially constituted inorganic glasses, but polymeric fibers have also been tried (→ Fiber Optics).

5.5. Nanofibers and Carbon Nanotubes

Electrospinning [9], which can be carried out on small-scale equipment, has made it possible to produce fibers with diameters down to a few nanometers. It was once thought that the fineness resulted from a splitting of the extruded fiber, but it is now known that it results from a massive attenuation of the fiber in a whirling mode. The web that is collected is a single fiber of enormous length. Spinning can be from solution or melt, so that a wide range of polymer nanofibers can be produced. The chemistry and fine structure would be expected to be similar to that of larger fibers made from the same material, though the fact that all the material is as close to the surface as the outer layer of a coarse fiber may lead to some differences.

The other type of nanofiber now in experimental production consists of carbon nanotubes. These derive from the dsicovery of C_{60}, in which carbon atoms are covalently bonded into a spherical network known as a buckyball. It was then found that hemispherical ends could be linked by a tube with a similar network. The basic structure is thus a two-dimensional, covalently bonded sheet of carbon atoms, which wraps around to form a tube. Short lengths of multiwalled tubes with diameters of 4 to 50 nm were reported in 1991 [83]. Long singlewalled carbon nanotubes with diameters of about 1.5 nm can be produced in carbon arcs [84, 85] or by laser ablation [86] and can be collected in rope form [87]. However, these methods are only suitable for producing minute amounts for scientific study or specialist uses.

The production of coarse multiwalled carbon fibers by vapor deposition had been known for over 100 years. 21st century research [5, 6] has shown that is possible to grow finer carbon nanotubes by vapor deposition and collect them as long lengths of yarn. Terminology [5] is confusing because this "yarn" is referred to as a "fiber", and not as a yarn composed of nanotube fibers. The nanotubes may be singlewalled with 1.6–3.5 nm in bundles with a diameter of 30 nm or multiwalled with diameters of 30 nm and aspect ratios of $\gg$1000:1 [5].

5.6. Smart Fibers

A major development of the 21st century is the production of smart fibers [41, 42]. Typically they are electrically, optically, mechanically, thermally, or chemically active and hence respond to external stimuli. It is beyond the scope of this chapter to go into detail on smart fibers. From a structural view point, it can be said that some fibers may be made of a single active

material, but more commonly the special properties result from inclusions or surface coatings. Two examples are the incorporation of globules of a phase-change material, which reduces the effect of changes of temperature, and coating fibers with a piezoelectric polymer, which enables fabrics to act as electric switches.

6. References

1. M. J. Denton, P. N. Daniels (eds.): *Textile Terms and Definitions*, 11th ed., Textile Institute, Manchester 2002.
2. M. Okamoto, in D. Bunnschweiler, J. W. S. Hearle (eds.): *Polyester: 50 Years of Achievement*, Textile Institute, Manchester 1993, pp. 108 – 111.
3. R. Binachi, R. Maglione, in D. Bunnschweiler, J. W. S. Hearle (eds.): *Polyester: 50 Years of Achievement*, Textile Institute, Manchester 1993, pp. 196 – 198.
4. H. Fong, D. H. Reneker, Electrospinning and the formation of nanofibers, in D.R. Salem (ed.): *Structure Formation in Polymeric Fibers*, Hanser, Munich 2000, pp. 223–246.
5. Y.-L. Li, A. Kinloch, A. H. Windle, *Science* **304** (2004) 276–278.
6. M. Zhang, K. R. Atkinson, R. H. Baughman, *Science* **306** (2004) 1358–1361.
7. H. R. Mauersberger (ed.): *Matthews' Textile Fibers*, 6th ed., Wiley-Interscience, New York 1954.
8. C. A. Farnfield, D. R. Perry (ed.): *Identification of Textile Materials*, 7th ed., Textile Institute, Manchester 1975.
9. J. W. S. Hearle, R. H. Peters (ed.): *Fibre Structure*, Textile Institute, Manchester 1963.
10. J. W. S. Hearle, R. Greer: *Text. Prog.* **2** (1970) no. 4, 1 – 203.
11. J. W. S. Hearle, *J. Appl. Polymer Sci.: Appl. Polymer Symp.* **47** (1991) 1 – 31.
12. R. Hill (ed.): *Fibres from Synthetic Polymers*, Elsevier, Amsterdam 1953.
13. P. H. Hermans: *Physics and Chemistry of Cellulose Fibres*, Elsevier, Amsterdam 1949.
14. P. Alexander, R. F. Hudson (revised by C. Earland): *Wool its Chemistry and Physics*, 2nd ed., Chapman and Hall, London 1963.
15. E. Lord: *The Characteristics of Raw Cotton*, Textile Institute, Manchester 1961.
16. R. Meredith, J. W. S. Hearle (ed.): *Physical Methods of Investigating Textiles*, Textile Book Publishers, New York 1959.
17. H. F. Mark, S. M. Atlas, E. Cernia (ed.): *Man-made Fibers, Science and Technology*, vols. **1, 2, 3**, Interscience, New York 1968.
18. J. G. Cook: *Handbook of Textile Fibers*, 4th ed., vols. **1, 2**, Merrow, Watford 1968.
19. W. E. Morton, J. W. S. Hearle: *Physical Properties of Textile Fibres*, 3nd ed., Textile Institute, Manchester 1993.
20. M. Lewin, J. Preston (ed.): *High Technology Fibers, part A*, Dekker, New York 1985.
21. R. W. Moncrieff: *Man-made Fibres*, John Wiley, New York 1975.
22. A. Ziabicki, H. Kawai (ed.): *High-speed Fiber Spinning*, Wiley-Interscience, New York 1985.
23. M. Lewin, E. M. Pearce (ed.): *Fiber Chemistry*, Marcel Dekker, New York 1985.
24. E. M. Hicks, R. A. Craig, E. L. Wittbecker, J. G. Lavin et al., *Text. Prog.* **3** (1971) no. 1, 1 – 127.
25. A. J. Hughes, J. E. McIntyre (ed.), *Text. Prog.* **8** (1976) no. 1, 1 – 177.
26. A. R. Bunsell, M.-H. Berger (eds.): *Fine Ceramic Fibers*, Marcel Dekker, New York 1999.
27. T. Hongu, G. O. Phillips: *New Fibers*, 2nd ed., Woodhead Publ., Cambridge 1997.
28. T. Nakajiama (ed.): *Advanced Fiber Spinning Technology*, Woodhead Publ., Cambridge 1994.
29. S. K. Mukhopadhyay (ed.): *Advances in Fiber Science*, Textile Institute, Manchester 1992.
30. H. H. Yang: *Kevlar Aramid Fiber*, John Wiley & Sons, New York 1993.
31. D. Bunnschweiler, J. W. S. Hearle (eds.): *Polyester: 50 Years of Achievement*, Textile Institute, Manchester 1993.
32. D. R. Salem (ed.): *Structure Formation in Polymeric Fibers*, Hanser, Munich 2000.
33. C. Woodings (ed.): *Regenerated Cellulose Fibres*, Woodhead Publishing, Cambridge 2001.
34. J. W. S. Hearle (ed.): *High-performance Fibres*, Woodhead Publishing, Cambridge 2001.
35. R. R. Frank (ed.): *Silk, Mohair, Cashmere and Other Luxury Fibres*, Woodhead Publishing, Cambridge 2001.
36. W. S. Simpson, G. H. Crawshaw (eds.): *Wool: Science and Technology*, Woodhead Publishing, Cambridge 2002.
37. R. R. Frank (ed.): *Bast and Other Plant Fibres*, Woodhead Publishing, Cambridge 2005.
38. J. E. McIntyre (ed.): *Synthetic Fibres: Nylon, Polyester, Acrylic, Polyolefin*, Woodhead Publishing, Cambridge 2005.

39. S. Gordon, Y.-L. Hsieh (eds.): *Cotton: Science and Technology*, Woodhead Publishing, Cambridge, in preparation.
40. T. Hongu, G. O. Phillips, M. Taigami: *New Millenium Fibers*, Woodhead Publishing, Cambridge 2004.
41. X. M. Tao (ed.): *Smart Fibres, Fabrics and Clothing*, Woodhead Publishing, Cambridge 2001.
42. X. Tao (ed.): *Wearable Electronics and Photonics*, Woodhead Publishing, Cambridge 2005.
43. W. Postman, *Text. Res. J.* **50** (1980) 444 – 453.
44. W. G. Bryson, D. N. Mastonarde, J. P. Caldwell, J. L. Woods, Proc. 10th Int. Wool Textile Res. Conf. Aachen, CD Publication ST-3, 1–10.
45. J. W. S. Hearle, B. Lomas, W. D. Cooke: *Atlas for Fibre Fracture and Damage to Textiles*, Woodhead Publ., Cambridge 1998.
46. E. Boyce, *ICEM 13*, Paris (1994) 51 – 54.
47. W. O. Statton, *J. Polym. Sci. Part C Polym. Symp.* no. 3 (1963) 3 – 8.
48. J. W. S. Hearle, *J. Polym. Sci. Part C Polym. Symp.* no. 20 (1967) 215 – 251.
49. J. W. S. Hearle, R. Greer, *J. Text. Inst.* **61** (1970) 240 – 244.
50. N. S. Murthy, A. C. Reimschussel, V. J. Kramer, *J. Appl. Polymer Sci.* **40** (1990) 249 – 265
51. J. W. S. Hearle, *J. Polym. Sci.* **28** (1958) 432 – 435.
52. J. W. S. Hearle: *Polymers and their Properties*, Ellis Horwood, Chichester 1982.
53. R. Hosemann, *Polym.* **3** (1962) 349 – 392.
54. V. A. Kargin, *J. Polym. Sci.* **30** (1958) 247 – 258.
55. D. H. Reneker, *J. Polym. Sci.* **59** (1962) S39 – S42.
56. J. W. S. Hearle, *Appl. Polym. Symp.* **31** (1977) 137 – 161.
57. H. A. McKenna, J. W. S. Hearle, N. O'Hear: *Handbook of fibre rope technology*, Woodhead Publishing, Cambridge 2004.
58. J. W. S. Hearle, L. Hollick, D. K. Wilson: *Yarn Texturing Technology*, Woodhead Publishing, Cambridge 2001.
59. B. M. Chapman, *Textile Res. J.* **39** (1969) 1102–1109.
60. J. W. S. Hearle, *Int. J. Biol. Macromol.* **27** (2000) 123–138.
61. W. A. Munro, G. A. Carnaby, *J. Textile Inst.* **90** (1999) 123–126.
62. W. A. Munro, *J. Textile Inst.* **92** (2001) 213–221.
63. J. W. S. Hearle, *Wool Tech. & Sheep Breeding* **51** (2003) 95–117.
64. J. Lunt, *Int. Fiber J.* 2000 (June) 48–52.
65. I. Sakurada: *Polyvinyl Alcohol Fibers*, Marcel Dekker, New York 1985.
66. W. Watt, B. Perov: "Strong Fibers," in A. Kelly, Y. N. Rabotnov (eds.): *Handbook of Composites*, vol. **1**, North-Holland, Amsterdam 1985.
67. M. S. Aslanova: [66], 3 – 60.
68. A. J. Majumdar: [66], 61 – 86.
69. M. G. Dobb: [66], 673 – 704.
70. M. Jaffe, R. S. Jones: [66], 349 – 392.
71. S. Ozawa, *Polymer. J.* **19** (1987) 119 – 125.
72. I. M. Ward: [66], 705 – 740.
73. P. J. Barham, A. Keller, *J. Mater. Sci.* **20** (1985) 2281 – 2301.
74. P. J. Lemstra, R. Kirschbaum, *Polymer.* **26** (1985) 1372 – 1384.
75. D. J. Sikkema, *Textiles Engineering for Performance*, UMIST Conference, Manchester 22 – 24 Apr. 1998.
76. M. Ehrburger, J. Donnet: [20], 169 – 220.
77. P. Bracke, H. Schurmans, J. Verhoerst: *Inorganic Fibers and Composite Materials*, Pergamon Press Oxford 1984.
78. M. Couper: [20], 51 – 86.
79. A. B. Conciatori, A. Buckley, D. E. Stuetz: [20], 221 – 268.
80. J. G. Scruggs, J. O. Reed: [20], 335 – 348.
81. S. W. Shalaby: [20], 87 – 126.
82. J. J. Gilman, *Phys. Today* **28** (1975) 46 – 53.
83. S. Ijima, *Nature* **354** (1991) 56.
84. S. Ijima, *Nature* **363** (1993) 56.
85. D. S. Bethune, C. H. Khiang, M. S. de Vries, G. Gorman, R. Savoy, J. Vasquez, *Nature* **363** (1993) 605.
86. T. Guo, P. Nikolaev, A. Thess, D. T. Colbert, R. E. Smalley, *Chem. Phys. Lett.* **243** (1995) 483–487.
87. A. Thess et al., *Science* **273** (1996) 483 ff.

Fibers, 3. General Production Technology

Werner Stibal, EMS Inventa-Fischer AG, Domat/Ems, Switzerland (Chaps. 1, 3.2, 5, and 7)

Raimund Schwarz, EMS Inventa-Fischer AG, Domat/Ems, Switzerland (Sections 2.1, 2.2, and 2.4)

Ulrich Kemp, EMS Inventa-Fischer AG, Domat/Ems, Switzerland (Section 3.1)

Klaus Bender, EMS Inventa-Fischer AG, Domat/Ems, Switzerland (Chap. 4)

Friedrich Weger, EMS Inventa-Fischer GmbH, Berlin, Germany (Chap. 6)

Manfred Stein, EMS Inventa-Fischer AG, Domat/Ems, Switzerland (Section 2.3, Section 3.3)

1. Introduction 3
2. Theory and Mechanisms of Fiber Production 3
2.1. Spinning 3
2.1.1. General Description 3
2.1.2. Theory and Mechanism of Spinning 4
2.1.2.1. Spinnability of Fluids 4
2.1.2.2. Flow Processes in Spinning 5
2.1.3. Melt Spinning 10
2.1.4. Solution Spinning 17
2.1.4.1. Dry Spinning 17
2.1.4.2. Wet Spinning 19
2.2. Drawing 21
2.2.1. Elongation Properties of As-Spun Yarns 21
2.2.2. Mechanism of Cold Drawing 22
2.2.3. The Drawing Process 23
2.2.4. Draw Ratio and Tensile Strength . . 24
2.2.5. Ultrahigh-Strength Synthetic Yarns 25
2.3. Crimping 25
2.3.1. General Description 25
2.3.2. Crimping Processes 25
2.3.2.1. Spontaneous Crimping 26
2.3.2.2. Deformation Crimping 26
2.3.2.3. Bicomponent Crimping 27
2.3.3. Crimping of Staple Fibers and Filament Yarns 27
2.4. Heat-Setting 27
2.4.1. Purpose 27
2.4.2. Mechanism 28
2.4.3. Shape Retention and Molecular Memory 30
3. Manufacture of Filament Yarns . 31
3.1. Melt Spinning of Filament Yarns 31
3.1.1. General Aspects 31
3.1.2. Conventional Spinning 31
3.1.3. High-Speed Spinning 34
3.1.4. Superhigh-Speed Spinning 35
3.1.5. Draw Twisting and Draw Winding . 36
3.1.6. Draw Warping 37
3.1.7. Horizontal Drawing 38
3.1.8. Spin Drawing 39
3.2. Solution Spinning of Filament Yarns 39
3.2.1. Dry Spinning 40
3.2.2. Wet Spinning 40
3.3. Texturing of Filament Yarns . . . 41
3.3.1. Development of Texturing Processes 41
3.3.2. False-Twist Texturing 41
3.3.3. Air-Jet Texturing 44
3.3.4. Stuffer Box Texturing 47
3.3.5. Bistructural Texturing 48
4. Manufacture of Tow and Staple Fibers 49
4.1. Fiber Spinning 49
4.1.1. Melt Spinning 49
4.1.1.1. Conventional Process 49
4.1.1.2. Compact Process 50
4.1.1.3. Continuous One-Step Process 53
4.1.2. Wet Spinning 54
4.1.3. Dry Spinning 54
4.2. Posttreatment of Tow and Staple Fibers 54
4.2.1. Drawing of Tow 56
4.2.2. Washing and Finishing 57
4.2.3. Drying and Heat-Setting 58
4.2.4. Crimping 59
4.2.5. Cutting 60
4.2.6. Packaging 62
5. Manufacture of Monofilaments . 62
6. Process Technologies for Spunbonded Nonwovens 64
7. Economic Aspects 64
8. References 66

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_511

Acronyms and Symbols

BCF	bulked continuous filament
DSC	differential scanning calorimetry
FDY	fully drawn yarn
HE	highly elastic
LOY	low orientation yarn
NVC	nip-controlled vector
PES	polyester
PETP	poly(ethylene terephthalate)
POY	partially oriented yarn
RFR	ribbon-free random wind
F	degree of setting, force
G	free energy, polymer throughput per spinneret hole
K	energy of cohesion
L	maximum thread length
$m*$	mass flow rate
$\bar{M}_{\mathrm{w}}$	weight-average molecular mass
N	normal stress difference
S	entropy, shrinkage
T	titer (fineness), temperature
U	internal energy
v	velocity
v_{L}	take-up velocity
v_0	extrusion velocity
V	volume
V_{z}	spin-draw ratio
x	distance
α	countersink angle
β	angle of transition
ε	elongation
$\dot{\varepsilon}$	elongation rate
ε_{H}	Hencky measure of elongation
$\dot{\gamma}$	shear rate
θ	temperature, normal stress coefficient
η	shear (dynamic) viscosity
η_{D}	elongational viscosity
σ	tensile stress, tenacity at break (based on titer)
σ_0	tenacity based on initial fineness
σ_{s}	surface tension
σ_{t}	tenacity based on cross-sectional area
τ	shear stress, relaxation time

1. Introduction

Man-made fibers are manufactured by converting high-molecular polymer materials into continuous fibers (monofilaments, filaments, tow, or staple fibers). Monofilament is a single continous fiber, filaments are composed of a certain number of continous spun and collected single fibers, tow consists of a collection of thousands of continous filaments, and staple fibers are tow that has been cut into discrete lengths. For a general description of fibers and their properties, see → Fibers, 1. Survey.

Key process steps in fiber production are

1) Spinning
2) Drawing
3) Posttreatment

In the first of these steps, *spinning*, polymers are forced by dosing units through small capillaries, either as a melt or in solution and then solidified into fibers by cooling, precipitation, or solvent evaporation. During spinning, the macromolecular chains are only partially oriented in a preferred direction. (Spinning may also be defined to include the process in which fibers of finite length, i.e., staple fibers, are combined to form yarn; this is, however, part of a different technology and is not treated in this article.)

In the second step, *drawing*, the as-spun fibers are often stretched to many times their original length. Drawing increases the molecular orientation along the fiber axis and is frequently accompanied by an increase in crystallinity and fiber strength.

In the third step, *posttreatment*, the drawn fibers are heated under tension or under defined conditions of relaxation (heat-setting or thermosetting). If necessary, filaments, which after spinning and drawing normally have a flat characteristic, can also be textured, and tow can be crimped. Heat-setting changes the fiber morphology and allows the properties of the filament yarn, staple fibers, or tow to be adapted to end-use requirements. A finishing oil must be applied to the fibers to improve their processing properties. The proper formulation of a finishing oil requires a balanced combination of lubricating, cohesive, and antistatic properties.

The classical production of synthetic fibers involves several separate steps. New technologies have been developed in the past few years with the goal of integrating as many production steps as possible into a single continous process; this has improved both the economics of synthetic fiber manufacture and the quality of the products.

2. Theory and Mechanisms of Fiber Production

2.1. Spinning

This section deals solely with melt and solution spinning. Other spinning methods such as emulsion and suspension spinning, phase-separation spinning, reaction spinning, and interfacial, and surface spinning are described in [1], [2].

2.1.1. General Description

In all spinning processes, a spin pump continuously forces a fluid spinning mass, sometimes under considerable pressure, through a spinneret orifice (the shaping element) into a suitable liquid or gaseous medium. The resulting filament is solidified by adjusting the temperature or the concentration of the medium. The spinning mass is converted into a fluid state by melting or by dissolution, depending on the properties (thermal stability) of the raw material. If the material is neither fusible nor soluble, special spinning processes are required. As the filament leaves the spinneret, it is narrowed, while still in the fluid state, by drawing off at a constant take-up velocity. The polymer molecules become weakly oriented along the longitudinal axis of the fiber as solidification progresses. Changes in morphology may occur depending on the process used.

Characteristic Variables. An important characteristic of the spinning process is the *spin-draw ratio* V_Z which represents the ratio of the take-up velocity v_L to the extrusion velocity v_0, i.e., the speed at which the spinning mass leaves the spinneret:

$$V_Z = v_L / v_0$$

The extrusion velocity v_0 is calculated from the ratio of the volume throughput (cubic centimeters per second) to the free cross-sectional area of the spinneret holes (square centimeters); both parameters are based on either the entire spinneret plate or on only a single filament (one hole). The spin-draw ratio varies with the spinning process; in melt spinning, it may be as high as several hundred, but in solution spinning the extrusion velocity for some polymers exceeds the take-up velocity (i.e., draw ratio < 1).

The most important characteristic of the as-spun material is the titer T which is a measure of its fineness and is also called the linear density. The titer is defined as the mass per unit length (grams per meter) and is expressed in tex:

$$1\,\text{tex} = 1\,\text{g per } 1000\,\text{m}$$
$$1\,\text{dtex} = 1\,\text{g per } 10000\,\text{m}$$

By differentiating fiber mass and length with respect to time, the equation defining the fineness can be rewritten as:

$$\text{Titer} = \frac{\text{Mass flow rate (g/s)}}{\text{Take-up velocity (m/s)}}$$
$$\text{T (dtex)} = \frac{m^* \text{ (g/s)}}{v_L \text{ (m/s)}} \times 10000$$

This equation is frequently used in spinning technology. The yarn number is obtained by assuming $m*$ to be the throughput of the entire spinneret plate. To obtain the fineness of an individual filament, the yarn number is divided by the number of spinneret holes. If the spinning mass is diluted, $m*$ must be corrected appropriately to calculate the fineness. The equation shows that the as-spun fineness depends only on the spinneret throughput and the take-up velocity. The diameter of the spinneret hole (and the associated spin-draw ratio) does not affect fineness but does influence the process of fiber formation occurring between the spinneret and the take-up unit.

The *fiber diameter* d, in millimeters, can be calculated from the titer with the aid of the volume density ϱ. For a fiber with a circular cross section:

$$d = 0.02 \times \sqrt{\frac{T}{\pi \cdot \varrho}}$$

where T is expressed in decitex and ϱ in grams per cubic centimeter.

2.1.2. Theory and Mechanism of Spinning

2.1.2.1. Spinnability of Fluids

Spinnability, i.e., the ability to form threads, is exhibited by many materials which differ widely in structure and composition; it is by no means

a unique property of macromolecules. Chain-like macromolecules are, however, a prerequisite for the production of useful fibers with good mechanical properties. According to Ziabicki [1], the maximum thread length attainable on application of tensile stress is a measure of spinnability. The maximum thread length L depends not only on the material, but also on the spinning conditions. For a given substance, L depends primarily on the product $v \cdot \eta$, where v is the pulling velocity in meters per second and η is the shear viscosity. As shown in Figure 1, L passes through a maximum at intermediate $v \cdot \eta$ values where it is affected to some extent by the ratio of surface tension to viscosity.

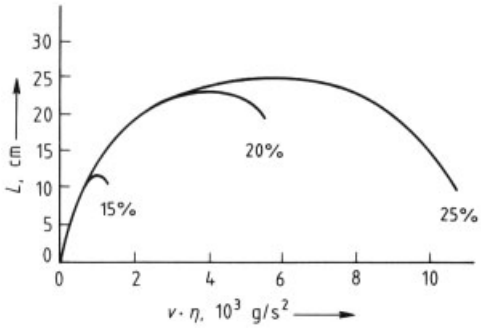

Figure 1. Maximum length L of a fluid thread as a function of the product of pulling velocity v and shear viscosity η. A thread was pulled by a rod from solutions containing the indicated concentrations of cellulose acetate in weight percent; intrinsic viscosity $[\eta] = 1.92$ dL/g in an 85 : 15 (vol : vol) acetone – water mixture [3]

A fluid, as-spun filament is thought to break by two mechanisms [4]. The first of these is *cohesive failure* in which fracture occurs when the tensile stress exceeds the tensile strength of the material at some point along the path of filament formation. Tensile strength depends on the energy of cohesion (critical elastic energy) of the material. Tensile stress can build up if the fluid, in contrast to ideally viscous Newtonian media, is able to store energy of deformation, i.e., if it is viscoelastic. The second failure mechanism is termed *capillary wave failure* and depends on the ratio of surface tension to viscosity. The larger the ratio, the more rapid is the increase in the amplitude of initially minute waves on the filament surface. After a certain distance, the capillary filament is divided into droplets and breaks (a fine stream of water also exhibits this type of behavior).

Depending on prevailing conditions, the filament will fail by the mechanism that predicts failure at the shorter filament length. By plotting theoretical curves for both models, a region of spinnability becomes apparent which coincides with that observed in practice (see Fig. 2).

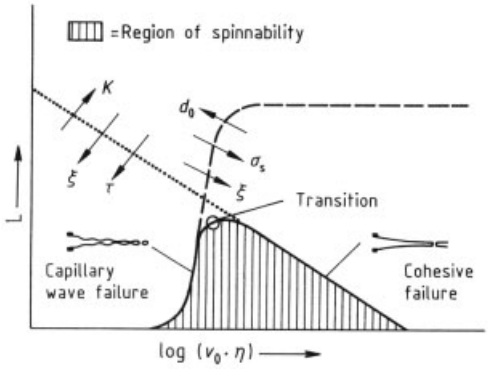

Figure 2. Superposition of theoretical curves of capillary wave failure (— — —) and cohesive failure (. . . .).
The maximum length L of a fluid filament is plotted as a function of the product of the extrusion velocity v_0 and the shear viscosity η; d_0 = initial diameter; $\xi = \mathrm{d}\ln(v/v_0)/\mathrm{d}x$ = axial deformation gradient; σ_s = surface tension; K = energy of cohesion; τ = relaxation time, time for the stress (stored energy) created by a sudden deformation to decay to the (1/e)th part [4].
The arrows show the direction in which the theoretical curves shift when the indicated parameter increases

2.1.2.2. Flow Processes in Spinning

Internal and external flow processes occur during fiber spinning. The flow of the spinning mass toward the spinneret and its passage through the spinneret holes are termed *internal processes*. The *external processes* include the behavior of the jet of spun material on leaving the spinneret hole, as well as drawing and solidification of the fiber. The flow processes occurring inside and on leaving the spinneret are governed by the shear flow properties of the spinning mass under the prevailing extrusion conditions (i.e., shear viscosity η and shear rate $\dot{\gamma}$). Subsequent flow processes are characterized by drawing properties and conditions (extensional or elongational viscosity η_D and elongation rate $\dot{\varepsilon}$). These two types of flow are shown schematically in Figure 3.

Flow Properties of Fluids. In contrast to solids, liquids begin to flow continuously when

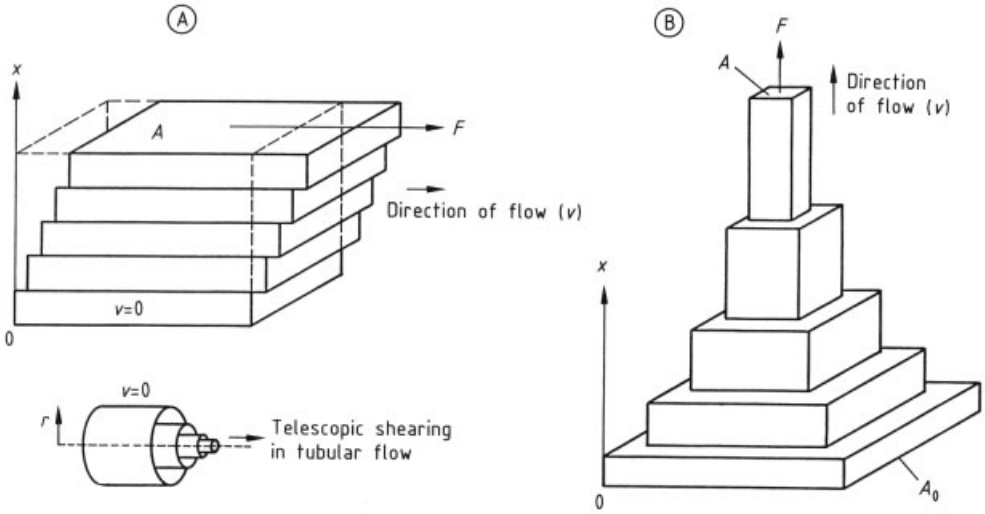

Figure 3. Schematic of flow processes in fiber spinning
A) Laminar shear flow shown as simple shear between horizontal layers and as tubular flow with telescopic shearing; B) Monoaxial, steady-state elongational flow
F = force; A = area; x = coordinate axis; r = radius; v = velocity

a force is applied. If the force is applied tangentially to a given plane, an equal and opposite force must act on the opposing interface, and shear flow occurs. Infinitely thin laminar layers are displaced in parallel and exhibit a velocity profile; the energy brought into play is converted to frictional heat. In tubular flow, the layers are concentric and the driving force is applied to the circular face. The opposing force is created by friction against the tube wall. The layers shown in Figure 3A must be imagined as being infinitely long in the direction of flow. In the steady-state, the fluid leaving the element under consideration is constantly replaced. The following terms are defined in laminar, linear shear flow:

Shear rate	$\dot{\gamma} = \mathrm{d}v\,(x)\,/\mathrm{d}x$ for a planar surface, or $dv\,(r)\,/\mathrm{d}r$ for a tube
Shear stress	$\tau = F/A$
Shear viscosity (also called Newtonian viscosity or dynamic viscosity)	$\eta = \tau/\dot{\gamma}$

where x = coordinate axis, v = velocity, r = tube radius, F = applied force, and A = area.

Thus, to achieve a given average flow rate, a force must be applied which is proportional to the shear viscosity of the material used.

In Newtonian fluids, η is constant over the entire range of shear rate (i.e., η is a material constant). However, polymer melts and polymer solutions used in fiber production generally exhibit a more complex behavior. They are pseodoplastic (η decreases as the shear rate increases, i.e., non-Newtonian) or viscoelastic (viscosity is time-dependent), and also display rubber-like elastic deformation. GLEISSLE discovered that all three phenomena can be attributed to the elastic behavior of polymers [5]. Thus, the steady-state viscosity function $\eta\,(\dot{\gamma})$ can be used to deduce the time-dependent behavior of a viscoelastic fluid under stress (mirror relations). At the molecular level, the high elasticity of polymers, in contrast to solids, cannot be explained by interatomic forces (with correspondingly low elongation). In macromolecular chains, the restoring force is instead produced by entropy effects which are based on the relatively unhindered rotation of the main single valence bonds and the nonextended bond angles. The most entropically favorable (i.e., most probable) conformation is that of a ball with randomly coiled chains [6]. In the absence of an external elongational force, the macromolecular polymer chains will adopt this conformation.

Elongational flow occurs when a force acts perpendicular to an area with a finite cross sec-

tion (see Fig. 3B). Because the volume of an incompressible fluid must remain constant, elongation is accompanied by a simultaneous transverse contraction. Hence during steady elongational flow, an infinitely small volume of fluid with an initial cross section A_0 accelerates in the direction of flow under the action of the elongating force. In the steady state, the outflow is continuously replaced by inflow through the area A_0. The following terms can be defined and are analogous to those already given for shear flow:

Elongation rate	$\dot{\varepsilon} = \mathrm{d}v(x)/\mathrm{d}x$
Tensile stress (also called elongational or normal stress)	$\sigma = F/A(x)$
Elongational viscosity (also called extensional or Trouton viscosity)	$\eta_\mathrm{D} = \sigma/\dot{\varepsilon}$

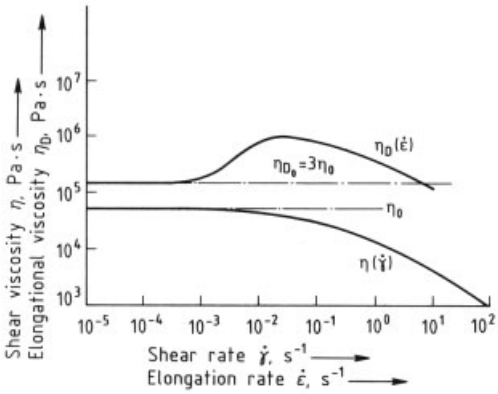

Figure 4. Steady-state shear and elongational viscosity as a function of the rate of deformation
Measurements were performed on branched polyethylene at 150 °C [7] (η = viscosity; η_D = elongational viscosity; η_0 = limiting viscosity at zero shear)

For an incompressible Newtonian material the following relation holds: $\eta_\mathrm{D} = 3\,\eta$ (analogous to the material parameters during the deformation of a solid). For non-Newtonian materials this relation is valid only at low rates of deformation (see Fig. 4); under these apparently Newtonian conditions, the steady-state viscosity does not change (i.e., $\eta = \eta_0$, where η_0 is the limiting viscosity at zero shear). As shown in Figure 4, further elongation is at first accompanied by consolidation, i.e., η_D increases. The shear viscosity η simultaneously decreases, however, because the shearing force increasingly orients the randomly coiled, ball-shaped molecules into elongated threads; consequently, the flow resistance drops.

Flow Processes in the Spinneret and Filament Formation. The most obvious evidence for the elastic behavior of polymers and their solutions during spinning is the fact that the extrudate leaving the spinneret has a larger diameter than the spinneret hole (see Fig. 5A). This effect is called die swell, onion formation, or the Barus effect. The ratio of the maximum diameter of the extrudate to the diameter of the spinneret, d_max: d_0, is called the die swell ratio or swelling factor.

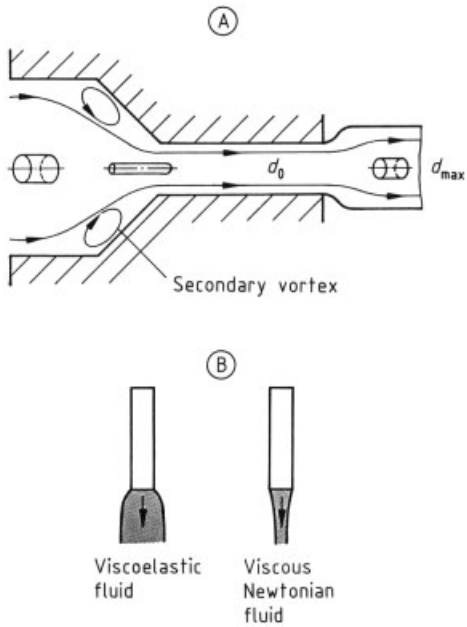

Figure 5. Spinneret flow and die swell ∗
A) Deformation of a cylindrical fluid element during extrusion through a spinneret; B) Flow of fluids out of a tube
d_0 = capillary diameter; d_max = maximum diameter of the extrudate
∗ Reproduced with permission from [8].

There are two reasons for die swell:

1) The inflowing mass in front of the spinneret capillary is subjected to an elongating deformation in addition to shear. The resultant elastic stress partially relaxes during flow through the spinneret capillary, the residual elasticity increases with decreasing residence time (i.e., lower ratio of the capillary length l to the capillary diameter d_0). Once the mass has left the spinneret, its diameter is

not limited by the capillary walls; the residual elasticity recovers, and this leads to expansion. The $l:d_0$ ratio usually varies from 1 to 3. Values for d_0 range from 0.1 to 0.4 mm with residence times of 10^{-4} to 10^{-2} s; polymer relaxation times can be as high 1 s.

2) Even polymer melts or solutions which emerge from a long tube without an orifice exhibit an increase in diameter (Fig. 5B). As mentioned previously, this is because shear flow produces entropy – elastic molecular orientations, as a result of the radial velocity field; a persistent tensile stress is thus created which is a function of the shear velocity (dynamic orientation).

For non-Newtonian fluids, the sum of these two elastic deformations causes the emerging filament to contract. The contraction in length is proportional to the square of the simultaneous swelling (die swell) factor. Under normal conditions, the filament diameter expands to 1.1 – 2.5 times the capillary diameter. For a viscous Newtonian fluid, however, the diameter of the emerging strand is approximately 14 % smaller than that of the capillary (see Fig. 5B).

For a complete description of the deformation of non-Newtonian materials during shear flow (such as die swell), normal stresses occurring along the three major coordinates must also be considered [8]. In the limiting case of incompressible fluids, rheology is not affected by absolute hydrostatic pressure; only the differences between the normal stresses are important. These are defined as follows (see also Fig. 6):

$$N_1 = \tau_1 - \tau_2 \text{ (first normal stress difference)}$$

$$N_2 = \tau_2 - \tau_3 \text{ (second normal stress difference)}$$

where τ_1, τ_2, and τ_3 are the normal stresses acting on the areas perpendicular to the three space coordinates.

In the case of simple, linear shear, $N_2 = 0$ and $N_1 = \theta \cdot \dot{\gamma}^2$, where $\dot{\gamma}$ is the shear rate and θ is the normal stress coefficient for the steady state. The parameter N_1 has the same effect on a fluid stream as an externally applied tensile stress. The resultant phenomena are called normal stress effects or the Weissenberg effect.

The mirror relations derived by Gleissle [5] can be used to calculate θ directly from the viscosity function $\eta \cdot (\dot{\gamma})$. The parameter θ is the greater, the more the shear viscosity has been reduced by the shear rate $\dot{\gamma}$. The importance of normal stresses is illustrated by the fact that, for the polymers used in the manufacture of man-made fibers, N_1 is greater than τ, even at low shear rates $\dot{\gamma} =$ ca. $10^{-1} - 10^{-2}\,\mathrm{s}^{-1}$).

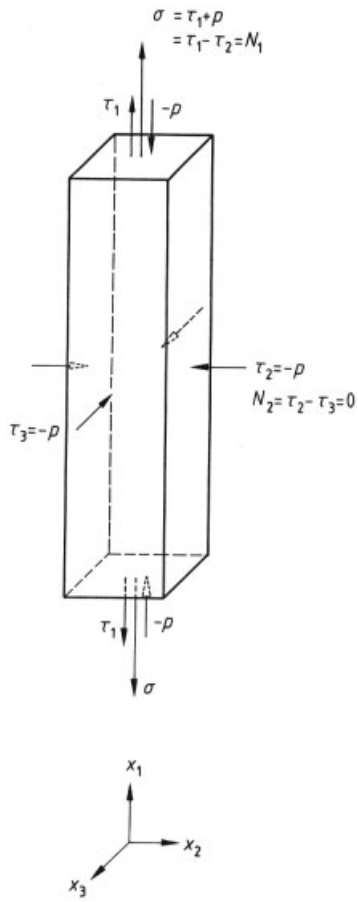

Figure 6. State of normal stress in a fluid element
p = hydrostatic pressure acting on all sides; τ_1, τ_2 and τ_3 = normal stresses acting on the areas perpendicular to the coordinates x_1, x_2, and x_3, respectively

Shear rates at the walls of spinneret capillaries are usually $10^2 - 10^5\,\mathrm{s}^{-1}$. If the ratio $N_1 : \tau$ becomes too large (i.e., if the shear rates are too high), melt fracture occurs. Melt fracture is a form of unstable flow that is produced by oscillating elastic deformations. This instability is not related to the Reynolds criterion for turbulence used in hydrodynamics. Melt fracture can be recognized by the fact that the filament emerging from the spinneret is not smooth: its diameter is irregular and varies more or less periodically

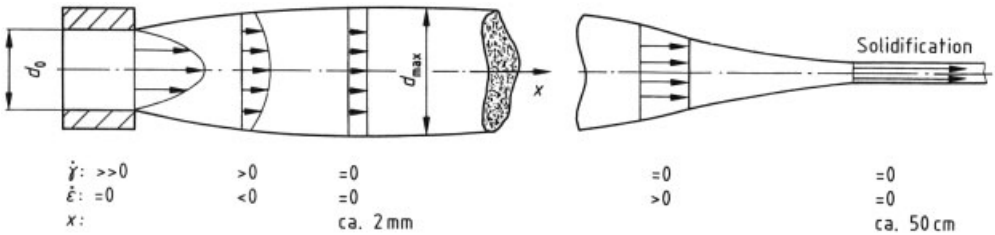

Figure 7. Velocity profile of an extruded, tapering filament
d_0 = capillary diameter; $d_{\max}$ = maximum filament diameter; $\dot{\gamma}$ = shear rate; $\dot{\varepsilon}$ = elongation rate; x = distance from the spinneret orifice

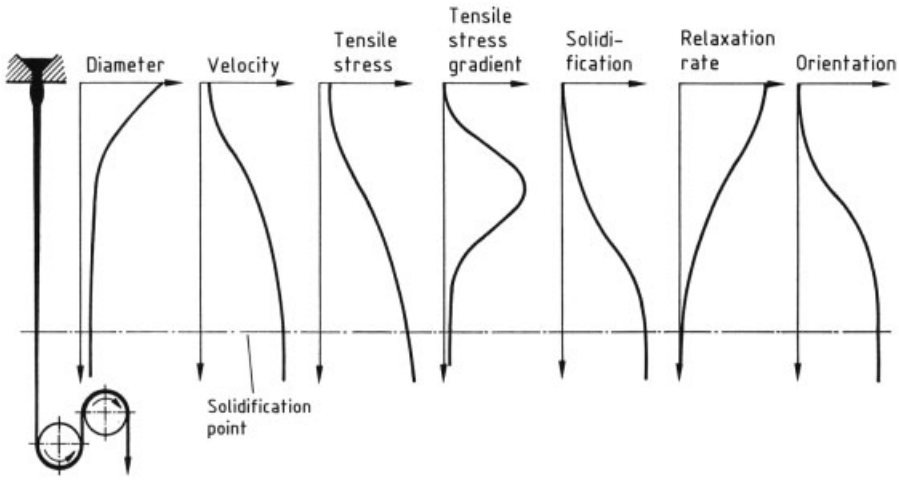

Figure 8. Schematic of the spinning process [12]

[9], [10]. GLEISSLE [11] discovered a universal critical value of the ratio $N_1 : \tau$ for different polymers, above which melt fracture occurs:

$$(N_1/\tau)_{\text{crit}} = 4.6 \pm 0.2$$

The spinneret hole diameter is the major factor affecting the shear rate and therefore the first normal stress difference N_1. The diameter must therefore have a suitable minimum value; its upper limit is determined by the fact that the spin-draw ratio and elongation rate must not be too great if excessive preorientation and cohesive failure are to be avoided.

Under steady-state conditions, the extruded filament exhibits the velocity profile shown in Figure 7. A parabola-shaped, shear flow profile develops in the spinneret capillary. After the material leaves the spinneret, the profile flattens out because frictional force is no longer exerted by the capillary walls and deceleration results simultaneously due to swelling. Once the maximum filament diameter is reached, normally within a distance of 1 – 3 mm, the tensile force begins to act on the fluid thread. The radial velocity over the whole cross section is constant; hence, the jet of spun fluid is accelerated and begins to taper. Under normal spinning conditions, the axial velocity gradient is ca. 10 – 70 s^{-1}, the largest deformation occurring within the first 20 – 30 cm. Spinnability is stabilized by the onset of coagulation or solidification and, at least in the initial region, by the increase in elongational viscosity shown in Figure 4.

The typical behavior of several fiber parameters during spinning is depicted schematically in Figure 8. Consolidation of the filaments, in the sense of solidification, is achieved by cooling (quenching) or coagulation in a spinning medium and is associated with an increase in the limiting viscosity at zero shear η_0. Since relaxation is slower in a viscous mass, the buildup of tension resulting from the elongating forces and increased viscosity is accompanied by an

increase in molecular orientation along the filament axis. This orientation can be measured by birefringence and remains in the as-spun filament after it has solidified at its glass transition temperature. The glass transition temperature corresponds to the softening point of the amorphous polymer regions; the melting point of a polymer, however, is defined as the melting point of its crystallites, above which the melt is thermoformable. Once the solidification point has been reached and deformation is complete, the stress continues to increase because of external friction between the thread and the spinning medium.

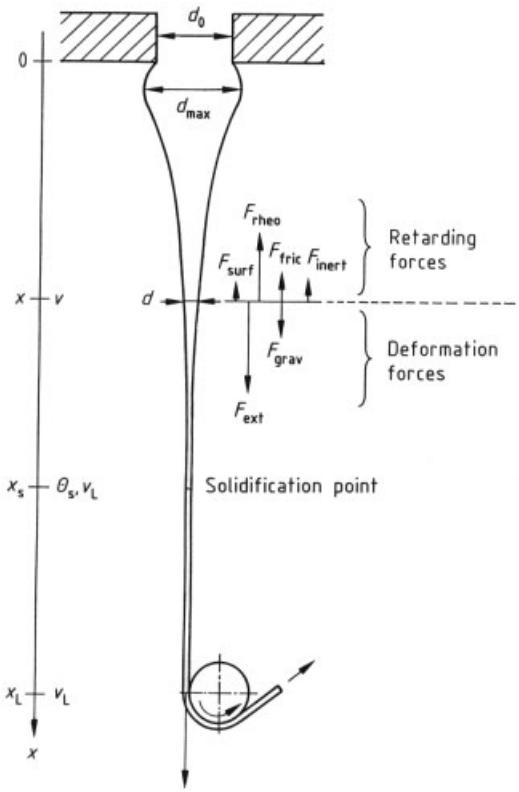

Figure 9. Balance of forces in a spun filament (direction of spinning is downward)
d = filament diameter; d_0 = capillary diameter; d_{max} = maximum filament diameter; x = distance from spinneret; v = filament velocity; θ_s = solidification temperature; subscripts S and L indicate solidification point and end of spinning path, respectively (see text for definition of forces)

Partially crystalline regions may be formed during spinning, depending on how rapidly the polymer tends to crystallize. In general, however, crystallization is difficult to predict because the interaction of polymer properties, spinning conditions, orientation, and crystallization are complicated.

During spinning, a filament is subject to the balance of physical forces shown in Figure 9. The first detailed investigation of the forces operating during spinning was carried out by Ziabicki [13]. At any given distance x from the spinneret, the retarding forces are balanced by the deformation forces:

$$F_{rheo}(x) = F_{ext} + F_{grav}(x) - F_{inert}(x) - F_{fric}(x) - F_{surf}(x)$$

where

F_{rheo} = rheological (viscous) force, tensile force
F_{ext} = external take-up force (constant along the entire threadline)
F_{grav} = force of gravity (weight of filament suspended from point x) – if the fiber is spun vertically upward F_{grav} is negative; if it is spun horizontally, F_{grav} is zero
F_{inert} = force of inertia (due to acceleration to v_L, the velocity at the end of the spinning path)
F_{fric} = frictional force between the surface of the filament and the spinning medium
F_{surf} = force due to surface tension (tends to minimize surface area of the filament by shrinkage)

At the solidification point ($x = x_s$) and thereafter, F_{inert} is zero; in addition, the surface and interfacial tension are often negligible. At the end of the spinning path ($x = x_L$), F_{grav} (in downward spinning) and F_{fric} also equal zero; at this point, the tensile force applied to the filament, F_{rheo}, is therefore equal to the external take-up force. Complicated mathematical procedures are, however, required to predict the course of filament formation (i.e., decrease in diameter) from the elongational viscosity and the calculated force components on the right-hand side of the above equation.

2.1.3. Melt Spinning

Melt spinning is the simplest spinning process. It is applied to polymers that can be melted and which do not decompose or change substantially

when maintained in the molten state for several minutes at a temperature up to 300 °C (e.g., polyesters, polyamides, polyolefins). The highly viscous melt is forced through a spinneret at a high pressure (ca. 10 – 20 MPa), therefore the spinneret plate has to be quite thick (ca. 10 mm). Figure 10 shows the geometry of a spinneret hole used in melt spinning.

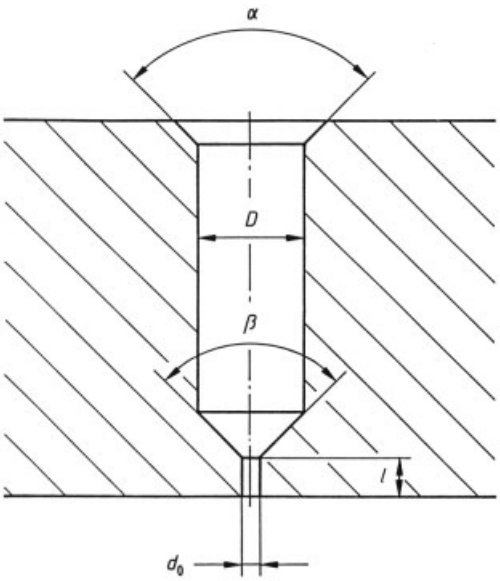

Figure 10. Vertical section through a spinneret hole used for melt spinning
α = countersink angel; β = angle of transition; d_0 = capillary diameter (ca. 0.2 – 0.4 mm, permissible variation = ±0.002 mm); D = prechannel diameter; l = capillary length (ca. 2 d_0)

After leaving the spinneret, molten filaments are cooled (quenched) in air as a spinning medium; cooling is usually accelerated by blowing the air across the filaments. In special cases, a liquid is used as a spinning medium to accelerate cooling even further (see also production of monofilaments, Chap. 5). This is not to be confused with wet spinning, which is dealt with in Section 2.1.4.2.

The key characteristic of melt spinning is that the filaments are extruded from a melt and solidified by heat exchange with a spinning medium. Since no exchange of material occurs, under steady-state conditions the equation of continuity (the simplest form of material balance) applies at every distance x along the filament:

$$\varrho \cdot A(x) \cdot v(x) = G = \text{constant}$$

where

ϱ = polymer density
A = cross-sectional area of filament
v = filament velocity
G = polymer throughput per spinneret hole (e.g., grams per minute)

If the filament cross section is circular and $A = \pi d^2/4$
(d = filament diameter), it follows that

$$v(x) = \frac{4G}{\pi \varrho d(x)^2} \text{ or } d(x) = 2 \times \sqrt{\frac{G}{\pi \varrho v(x)}}$$

Hence, the decreasing diameter of the as-spun filament is a direct measure of the increasing speed during spin elongation. Because frictional forces are relatively low when spinning is performed in air, take-up speeds of 8 – 25 m/s are obtained in classical melt spinning; speeds up to 100 m/s or more are possible in high-speed spinning [14].

Mathematical Models. Melt spinning is the spinning process most easily described by theoretical models and mathematics. Terms have been derived for the force and heat balance required to quantitatively describe filament formation during melt spinning [13], [15]. The heat balance is necessary to calculate the temperature-dependent elongational viscosity. Hamana [16] derived a coupled system of three first-order differential equations containing the filament variables velocity $v(x)$, tensile force $F_{\text{rheo}}(x)$, and temperature $\theta(x)$. The equations and both the initial and boundary conditions also take into consideration the physical properties of the polymer and air, as well as the spinning parameters (diameter of the spinneret hole, extrusion temperature, throughput per hole, take-up speed, and velocity of blown air). The mathematical solution of this system of equations is difficult because it requires numerical integration and knowledge of the initial value of the filament tensile force at the spinneret, F_0. The parameter F_0 is defined indirectly by the boundary condition that the filament speed must equal the take-up velocity when the filament reaches the take-up unit or is at its glass transition temperature. Hence, the solution requires several iterative integrations in which an estimated initial

value of F_0 is fitted until it meets the boundary condition.

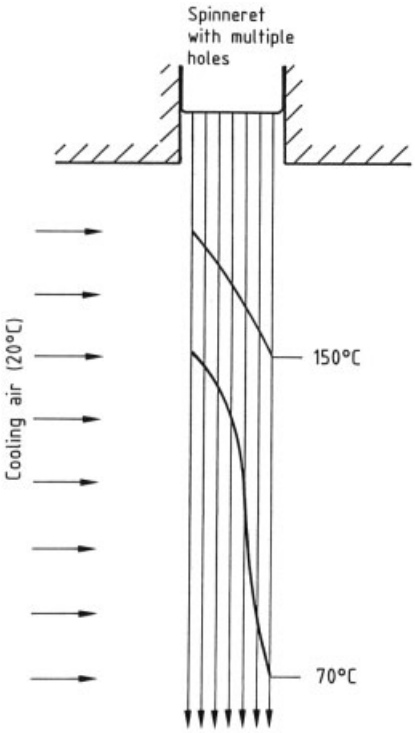

Figure 11. Temperature variation along a bundle of filaments during melt spinning of poly(ethylene terephthalate) (PETP) [17]
Spinning path, ca. 1 m long; 70 °C isotherm = solidification curve at the glass transition temperature (T_g)

Computer Simulation. Hamana [16] was the first to write a computer program for simulating the melt spinning process. His measurements on the variation of temperature and diameter of poly(ethylene terephthalate) PETP and nylon 6 filaments confirmed his theoretical calculations. Poly(ethylene terephthalate) has a high glass transition temperature and does not develop any significant crystallinity during spinning up to medium speed. The birefringence, determined as a measure of molecular orientation, is proportional to the tensile stress calculated by the program. Furthermore, the birefringence of the quenched, as-spun filament corresponds to the tensile stress at the glass transition temperature. Mathematical simulation can thus be used to predict the properties of as-spun PETP filaments as a function of spinning conditions. The effect of spinning parameters on filament formation can be simulated easily, but this does not make spinning experiments superfluous.

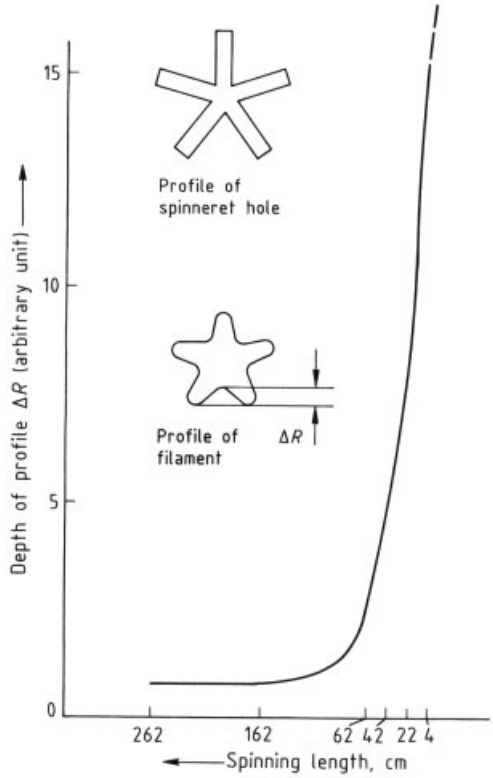

Figure 12. Change in the cross section of nylon filaments spun from a spinneret with star-shaped holes [18].
Spinning length indicates distance between spinneret and liquid cooling bath

The program used by Hamana [16] has the drawback that it can only be used to simulate the formation of individual, isolated filaments under defined quenching conditions. In practice, fibers are spun in bundles which contain up to 3000 individual filaments, and quenching conditions varying considerably from filament to filament. Since air is blown from one side, cooling conditions can be very poor for filaments furthest from the air source for two reasons: (1) air reaching the rear filaments is already preheated and (2) the velocity of the transverse stream of air decreases because the front filaments entrain part of the blown air in their aerodynamic boundary layer [1]. Therefore, the program for individual filaments has been extended to include the heat and mass balance of air so that calculations

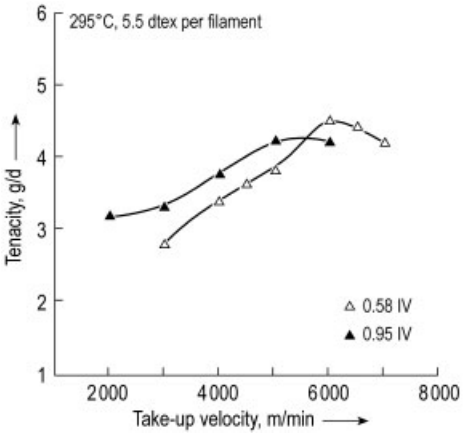

Figure 13. Tenacity versus take-up speed for PETP filaments of two different intrinsic viscosities [20], 1 cN/dtex $\triangleq$ 0.8839 g/den

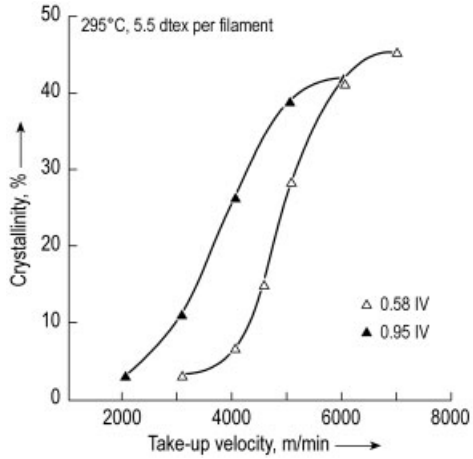

Figure 14. Birefringence versus take-up speed for PETP filaments of two different intrinsic viscosities [20]

can be applied to the entire filament bundle. Figure 11 shows the isotherms obtained from such a simulation. The cooling path in the first row of filaments is quite short and corresponds to that obtained for individual filaments. However, filaments in the fifth and subsequent rows take about twice as long to solidify. In practice, this can lead to undesirable sticking when the filaments come into contact with one another. Fiber quality data and the filament-yarn bundle simulation program have been used to show how air blowing and quenching can be improved and standardized without having to reduce the number of filaments [17].

Filament Shape. Noncircular spinneret holes may be used to shape the filament profile and thus modify the properties of the product (e.g., luster, touch, coloration). The cross section of the spun filaments generally differs from that of the spinneret holes because surface tension tends to minimize the surface area of the filament. For any given length, a minimal surface area is achieved when the cross section is a solid circle. This change, however, does not occur instantly because it is retarded by viscosity. An example of the change in cross section along the spinning path is illustrated in Figure 12.

A well-developed profile is obtained by spinning at a low temperature, with rapid quenching; the viscosity then increases rapidly and the profile is "frozen" when the glass transition temperature is reached. The spinning temperature must not be too close to the polymer melting point because the swelling factor increases significantly just above the melting point, which makes the emerging filament rounder. Cooling conditions are most difficult to prescribe for hollow fibers spun from spinnerets that usually consist of a ring slit that is divided into one or more segments by fillets (see Fig. 67); individual segments must then flow together before solidification occurs. The external pressure is higher than that inside the hollow filament cavity and thus reduces the size of the cavity. As a result of this pressure difference, the air which later fills the hollow filament flows into or is sucked into the filament immediately below the spinneret; it enters at the fillets between the circular segments, before the filament closes up. Expensive spinneret designs have been patented in which this problem is overcome by spinning the hollow filament from a closed ring and blowing an inert gas into the center of the hollow internal cavity [19].

Filament Formation in Melt-Spinning Processes. In the 1970s a new fiber manufacturing process was developed that combined drawing and texturing. With this new process, undrawn, partially oriented yarn (POY) is still produced today. In the case of poly(ethylene terephthalate), PETP, the spinning speeds of this

process is between 2700 and 3500 m/min, depending on the desired product. However, higher spinning speeds are disadvantageous for the producer because of the increase in orientation and crystallinity, which deteriorate the required properties of this pre-yarn. At spinning speeds above 3500 m/min the textile physical properties of the textured yarn are negatively influenced. Especially the draw ratio in the texturing process is much lower than that of conventional POY, so that the productivity would be reduced. Because the draw ratio is lower and the final titer (fiber diameter) is fixed, the POY titer must also be reduced. Hence, if a conventional PETP is used, typical limits exist which are attributable to the inherent properties of the spin polymer or the corresponding polymer physics.

Which factors influence the fiber formation process? For two different intrinsic viscosities (IVs), Figures 13, 14, and 15 show the tenacity, crystallinity, and birefringence as a function of spinning speed [20]. For an intrinsic viscosity of 0.95 dL/g, which corresponds to high molecular mass PETP for the production of tire cord, molecular orientation (birefringence), crystallinity, and tenacity already increase at relatively low spinning speeds.

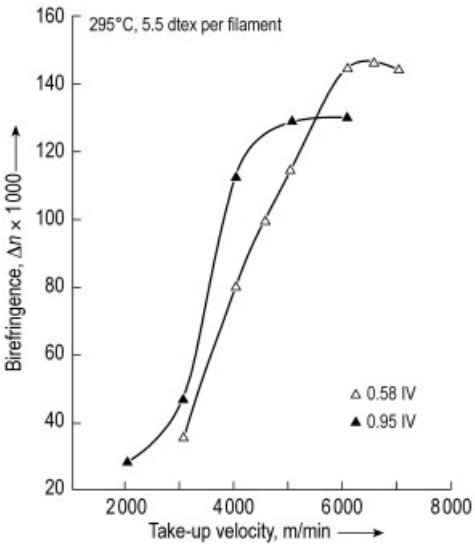

Figure 15. Crystallinity versus take-up speed for PETP filaments of two different intrinsic viscosities [20]

For both IVs a maximum exists in all three diagrams, but to a first approximation the curves for the higher molecular mass, which corresponds to higher intrinsic viscosity, are shifted to lower spinning speeds, and the maximum heights of the curves are different. For a higher intrinsic viscosity (0.95 dL/g) the onset of the strong increase in molecular orientation starts at lower velocities, but the maximum attained by the low-viscosity PETP is higher (Figs. 13, 14, 15). Therefore, by varying the intrinsic viscosity, it is possible to shift the molecular orientation at a given spinning speed to lower or higher values. But this adjustment is limited. Not every molecular mass is suitable for all kinds of end product. However, polymer modifications can also be used to influence the POY textile physical properties (see page 16).

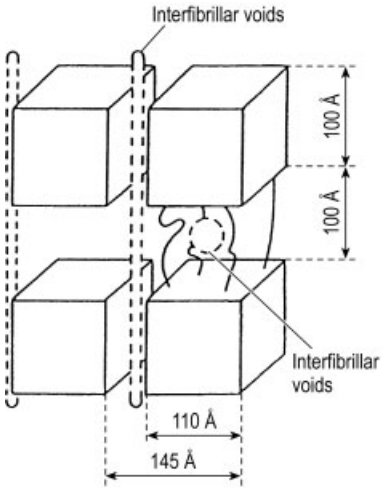

Figure 16. Fine-structure model for a spinning speed of 10 000 m/min [24]

Morphology of the As-Spun Fiber. Figure 16 shows the morphology of a semicrystalline PETP fiber. Between the crystallites, which are always more or less highly oriented, are the noncrystalline regions, which begin to orient themselves at a certain spinning stress. When a critical molecular orientation is attained, the intermolecular forces are so high that crystallites are spontaneously formed (Fig. 17) [21]. In high-speed spinning, the structure-forming process is so fast that inter- and intrafibrillar voids are generated. This can be detected by the positron-annihilation method, whereby the lifetime of the positron is used to detect voids in the polyester [22]. Positrons and electrons can form a positronium atom, the annihilation

characteristic of which can be used to measure the microvoids in the polyester [23].

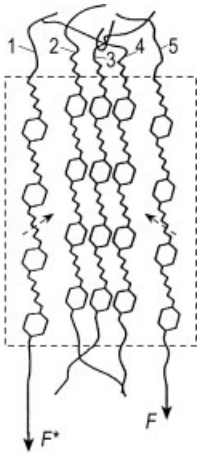

Figure 17. Model for PETP crystal growth under high tensile stress [21]
F and $F*$ are forces ($F* > F$)

Because of the different fiber morphology that results from high-speed spinning as compared to conventional low-speed processes, dye pickup is drastically improved.

The molecular chain alignment at which the fiber is still noncrystalline is called the amorphous orientation or oriented meso phase. The meso phase is intermediate between a totally coiled amorphous polymer chain arrangement and a highly ordered structure which launches the crystallization process above a critical degree of orientation. Figure 18 shows the development of the molecular orientation, beginning with the free-fall yarn directly under the spinneret to the oriented meso phase, the necklike deformation, and ending with the semicrystalline end state of this spin yarn [25], [26].

The amorphous orientation can be measured by different methods. Beside the investigation by the birefringence method [27], the amorphous orientation of PETP can be measured directly by detecting the natural intrinsic fluorescence of its benzene rings [28]. Crystallite sizes, intrafibrillar distances, and crystallite orientation are detectable by X-ray investigations [29].

Radial Distribution of Molecular Orientation. The higher the spinning velocity at a given fiber diameter, the more inhomogeneous is the radial distribution of the molecular orientation in the as-spun fiber [30], [31]. A synonym for this morphology is skin – core structure.

This characteristic morphology can be clearly observed in yarns which were spun at very high velocities, e.g., speeds higher than 7000 m/min. This molecular structure can be measured by the birefringence method as a function of the fiber diameter (Fig. 19) [32–35]).

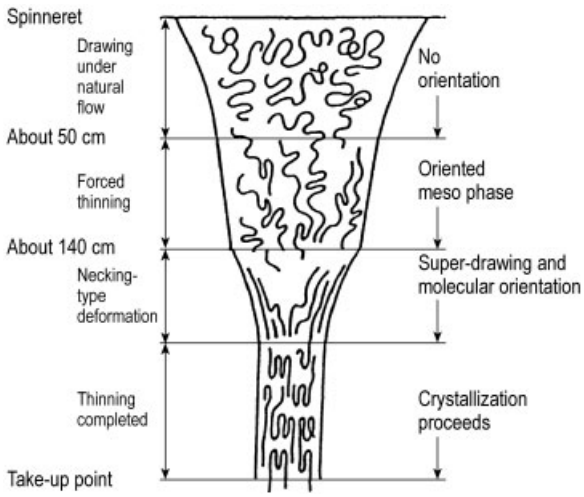

Figure 18. Changes in fiber structure during high-speed spinning [25], [26]

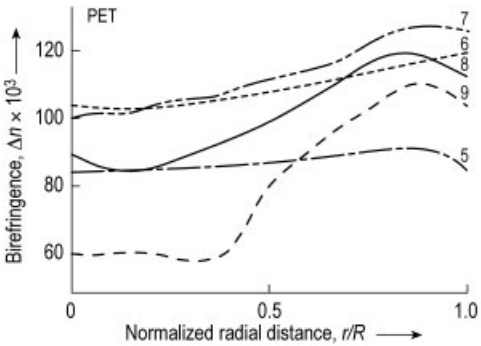

Figure 19. Radial distribution of birefringence at different take-up speeds (Take-up velocity in km/min) [32]

Figure 20 illustrates the radial distribution of the polymer (fiber) velocity gradient, the temperature in the filament, the corresponding spinning stress, and the resulting distribution of the molecular orientation (birefringence) [36].

At extreme spinning velocities with very fast cooling of the filament, the orientation process can only take place at the outer parts of the fiber [32], [24]. So at ultrahigh spinning speeds, e.g., 10 000 m/min, the spinning stress is concentrated in the boundary layer. However, a certain minimum spinning stress and enough orientation time are required for development of high molecular alignment and crystallinity. Hence, the molecules in the skin part of the fiber, which are oriented first, are frozen before the spinning stress can propagate through the whole thread. The result is a very inhomogeneous radial orientation distribution. The skin part is much more highly oriented than the molecules in the middle of the fiber. This characteristic form is generally called skin/core structure.

Further investigations of the development of the skin/core structure in fiber spinning can be found in [37–39], [40].

Polymer Modification. Polymer additives can be favorably used to influence the development of the fiber morphology. Chemical modification is performed by several producers. But these chemical modifiers have limits because the high concentrations which would have to be used to strongly influence the fiber properties deteriorate the textile physical properties of the products. So mostly these are not suitable for the conversion to end products.

Beside chemical polymer modifications [41], also polymeric additives are used, for example, to influence the orientation of the molecular chains in the fiber. The resulting polymer blends can be classified as two-phase, one-phase, or intermediate systems. Where phase separation occurs, the components are incompatible.

A polymer blend consisting of polystyrene/polypropylene and polystyrene/high-density polyethylene yields a decrease in melt viscosity [42], [43], and a decrease in melt viscosity was observed for polyethyleneterephthalate/PA66 blends [44]. Interesting fiber properties were obtained by adding a liquid-crystalline copolymer to polyethyleneterephthalate as a modifier. At a given take up velocity the molecular orientation could be drastically reduced so that the elongation at break (extension) was substantially increased (Fig. 21) [45], [46].

The addition of liquid-crystalline polymers is very effective with regard to the elongation increase, but the advantages in production cost limited because of the high price of such additives.

The elongation increase on adding a further polymer can be attained also in the case of nylon and other matrix polymers. Poly(methyl methacrylate) was also effective, but concentrations of more than 2 % are required to give a sufficient increase in elongation increase at higher velocities, for example, at spinning speeds of 5000 m/min [46]. Additionally yarn breaks were observed at high velocities.

An additive consisting of 50 – 90 % imidized polymethacrylic acid alkyl ester [48] can be used concentrations of less than 1 wt % for the above applications. The additives are mostly used for the production of yarns which can be directly used (fully drawn yarn). The advantage of this kind of modifier is its high effectiveness; the disadvantage is the high price of such copolymers. A further copolymeric additive consisting of two of the monomers A, B, and C, where A is polystyrene or methyl-substituted polystyrene; B is imidized maleic anhydride; and C is maleic anhydride or maleic acid is described in [49].

For POY processes which run at high speeds of, for example, 5000 m/min Stibal [50] describes an investment cost advantage of 25 % and the corresponding production cost saving for

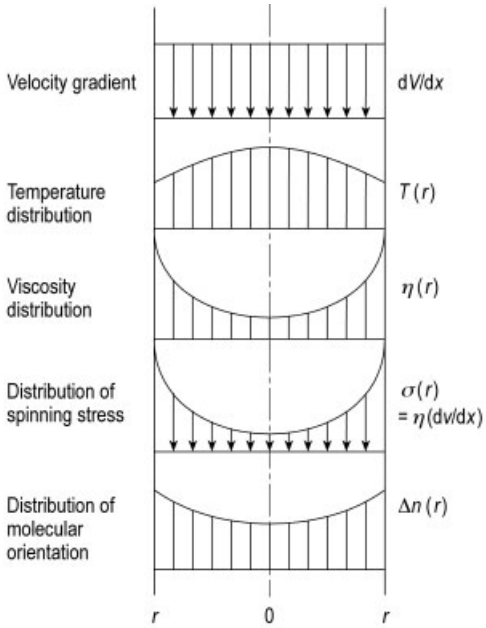

Figure 20. Radial distribution of take-up speed, viscosity, spinning stress, and molecular orientation [36]

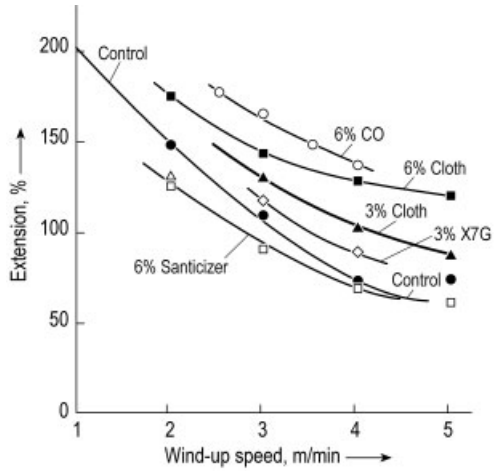

Figure 21. Fiber elongation with different modifiers (Santicizer, X7G, Cloth, Central) [45], [46]

a melt modification process in continous polymerisation plants. The fiber forming process is in general an elongation process. So when incompatible polymeric additives are added to a polymer melt, additional shear forces occur between the matrix and the additive [45], and rotational components are developed during the fiber forming process (Fig 22), [51]. At the same time the morphology is frozen by fast cooling, so that the orientation of the resulting fiber is reduced. The orientation decrease can be measured by an increased elongation at break and/or by a reduction in birefringence (Figs. 23 and 24), [52].

2.1.4. Solution Spinning

Polymers that do not meet the requirements for melt spinning may be spun if they can be dissolved in a suitable solvent. The polymer swells when dissolved and may form a completely homogeneous solution, depending on prevailing bonding forces and interactions. Polymers prepared by solution polymerization (e.g., polyacrylonitrile) can be spun directly without intermediate processing. Much lower spinning pressures (ca. 0.5 – 4 MPa) are used than in melt spinning because the dissolved, diluted polymer flows more freely than a polymer melt. Polymers with a very high molecular mass (generally up to a weight-average molecular mass $\bar{M}_w$ of 250 000) can also be spun. This is not possible in melt spinning because the limiting viscosity at zero shear η_0, and thus the spinning pressure, increase in proportion to $\bar{M}_w^{3.4}$. In solution spinning, the effect of molecular mass on viscosity can be compensated for by appropriate dilution [1], [53]. However, the concentration should not be too low since this reduces polymer throughput and increases the cost of solvent recovery.

2.1.4.1. Dry Spinning

In dry spinning, the polymer solution is spun into a hot gas whose temperature is higher than the normal boiling point of the solvent. The solvent thus evaporates, most of the heat of evaporation being supplied by the drying gas which also removes solvent vapor. Evaporation of the solvent along the spinning path is determined mainly by its rate of diffusion through the filament, which

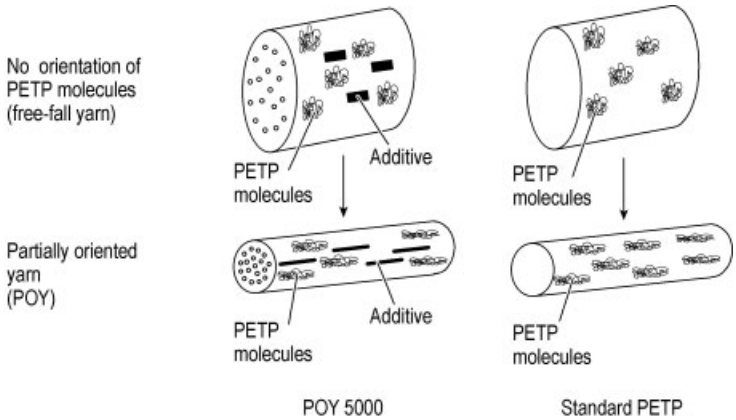

Figure 22. Structure-formation process related to the reduction of molecular orientation resulting from a polymer additive [51]

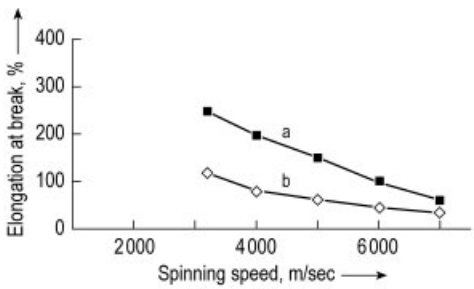

Figure 23. Fiber elongation versus spinning speed for physically modified (HP5000 PETP) and unmodified PETP polymer [52]

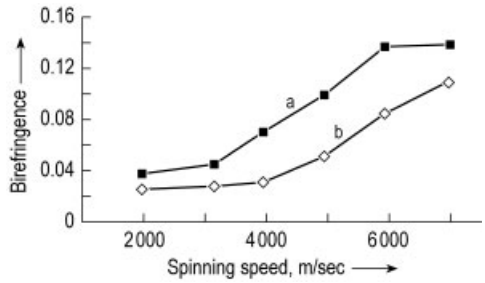

Figure 24. Birefringence versus spinning speed for physically modified (HP5000 PETP) and unmodified PETP polymer [52]

decreases as solidification progresses. Solidification occurs as a result of the decrease in solvent concentration and the associated increase in viscosity; the filament passes through a gel form into the solid phase. Since the heat supplied to the filament by the drying gas is largely dissipated by solvent evaporation, the temperature of the filament does not change significantly. The temperature increases only toward the end of the spinning path. Solidification by means of material transport (i.e., diffusion of solvent from the interior of the filament to the surface and subsequent evaporation) leads to pronounced radial inhomogeneities, particularly in spinning solutions with low polymer concentrations. Since solidification proceeds more rapidly at the surface of the filament than inside, a typical sheath–core structure is formed. Diffusion of the solvent which remains in the interior is thus hindered by the already solidifying skin. If evaporation from the surface is very rapid, the skin becomes so solid that its circumference remains practically constant. Further diffusion of solvent from the core results in a loss of mass from the interior, with formation of cavities and negative pressure; the skin then collapses and folds inward. The resulting filaments display cross sections that have the appearance of wrinkled lobes, almost like a peanut.

In dry spinning, the filament always has a residual solvent content of 5–25 wt %, which exerts a plasticizing action. This is desirable because it facilitates subsequent drawing of the filament; therefore the residual solvent is removed later in the process.

The spin-draw ratio based on the extrusion rate of the spinning solution ranges from 1 to 10 and is thus substantially lower than that found in melt spinning which may be as high as several hundred. The ratio is lowered by a factor of $1/c_p$ due to dilution alone (c_p is the polymer concentration).

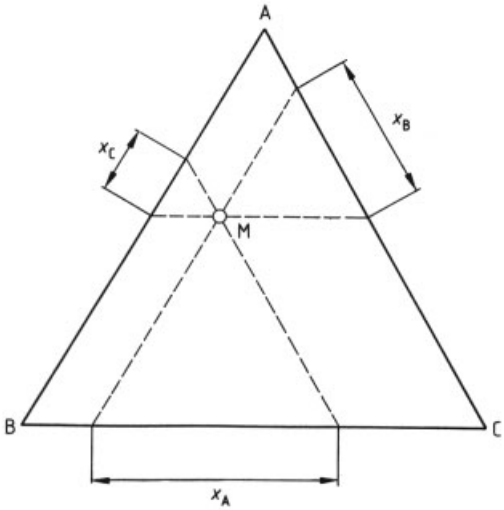

Figure 25. A triangular diagram representing a ternary mixture.
Points A, B, and C represent the pure components; point M represents a mixture with composition $x_A + x_B + x_C$; the dashed lines pass through M and are parallel to the sides of the triangle

Although general physical laws and models of material transport are applicable to dry spinning, a mathematical description is even more difficult to derive than for melt spinning [1], [54]. A simple model for the upper part of the filament (where diffusion is not limiting) is based on the assumption that the ratio of the material transfer coefficient to the heat transfer coefficient is constant. For an overview of the dependence of fiber properties on dry-spinning parameters, see [55].

2.1.4.2. Wet Spinning

Wet spinning is used if the polymer is dissolved in a nonvolatile solvent or if the polymer dissolves only after chemical transformation (which necessitates a subsequent reverse reaction). In wet spinning, the polymer solution is spun into a liquid coagulating bath. Heat exchange with the spinning medium is not responsible for solidification of the filament. Instead, solidification occurs as a result of coagulation caused by phase separation. Phase separation is induced by a component of the spin bath which is incompatible with the polymer (i.e., a nonsolvent). The polymer, solvent, and nonsolvent constitute a ternary system, and processes occurring during wet spinning can best be explained by using a triangular diagram as shown in Figure 25. The pure components A, B, and C are represented by the corners of an equilateral triangle. Point M then represents a mixture with the composition $x_A + x_B + x_C = 100\,\%$. If M lies on an edge of the triangle, the mixture is binary.

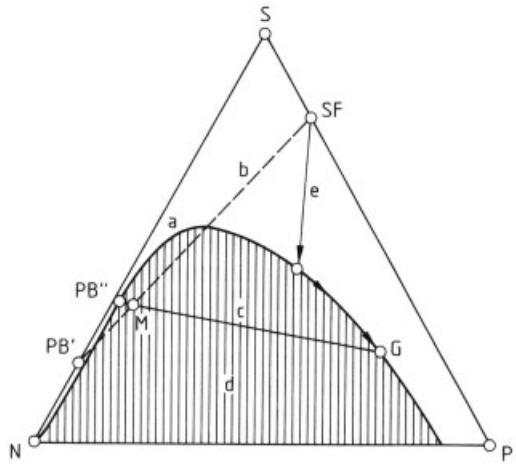

Figure 26. Phase diagram of a ternary system during wet spinning
Points: P = polymer; S = solvent; N = nonsolvent; SF = spinning solution; PB′ = inflow into coagulating bath; M = composition of mixture (on mixture line according to the law of leverage); G = gel-like filament after phase separation (in thermodynamic equilibrium); PB″ = outflow from coagulating bath (located opposite G on the binodal curve)
a) Binodal curve (limit of miscibility gap); b) Mixture line; c) Line of phase equilibrium (passing through M); d) Two-phase region (miscibility gap, shaded); e) Path of change in composition of the spun filament in the coagulating bath (diffusion-controlled)

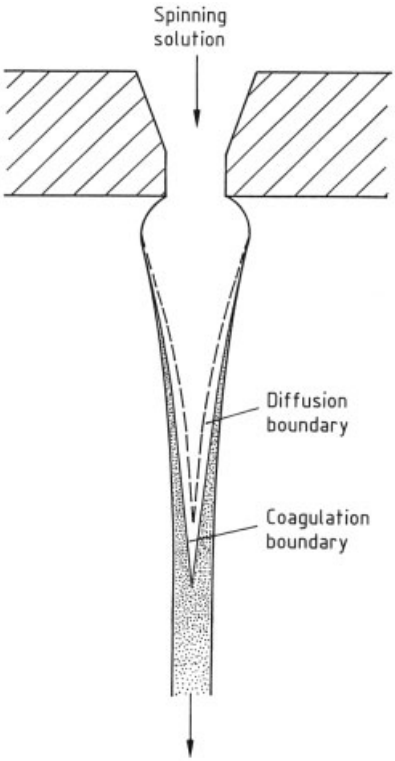

Figure 27. Filament formation during wet spinning [2]
The stippled area indicates coagulation

The situation in the coagulating bath is depicted in the form of a triangular phase diagram in Figure 26. The most important feature is a rather broad (shaded) immiscible region under the so-called binodal curve. Any mixture which, based on its average theoretical composition, lies in this region separates into two immiscible, liquid phases. The composition of these two phases is given by the points at which the equilibrium line passing through M intersects the binodal curve. (Any two opposing points on the binodal curve are connected by such a line.)

The nonsolvent N (frequently water) must be chosen so that it is miscible with the solvent S in all proportions. When the spinning solution SF enters the coagulating bath with an initial composition PB′, the total mixture can be represented by the point M; the position of M on the line connecting SF and PB′ is such that the ratio of the distance SF–M to the distance PB′–M is the same as the ratio of the amount of PB′ to the amount of SF (law of leverage). By using point M as an aid in graphic construction, points PB″ (final composition of the coagulating bath) and G are obtained which correspond to thermodynamic equilibrium. Point G represents the final state of a swollen gel fiber; its surface corresponds to the interfacial boundary with the spin bath. The route by which G is attained, if at all, primarily affects the structure of the gel

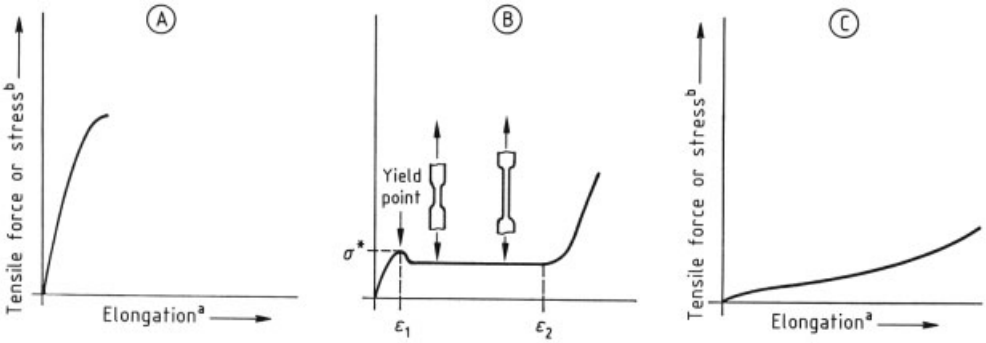

Figure 28. Representative stress – strain curves of spun filaments
A) $T_g \gg T$, crystalline – oriented; B) $T_g = T$, low orientation (cold drawing) – movement of the draw shoulders or neck is illustrated; C) $T_g \ll T$, plastic
T = temperature of measurement which increases from left to right if material used in A, B, and C is identical; T_g = glass transition temperature; $\sigma* =$ critical stress
[a] Elongation $\varepsilon = \Delta l / l_0$ where Δl is the increase in length of a filament with an original length l_0; [b] Tensile stress based on the initial fineness (titer)

fiber being formed, it depends on the kinetics of solidification and on the residence time.

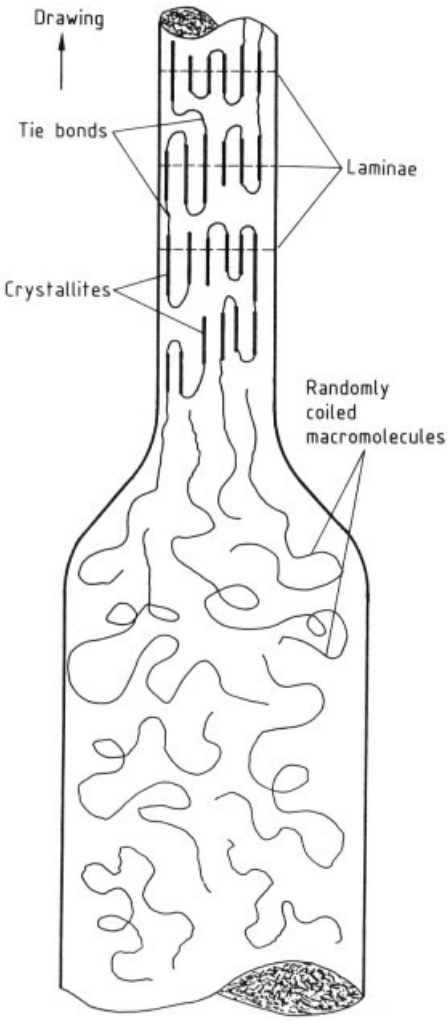

Figure 29. Draw point (bottleneck) and molecular orientation obtained during cold drawing

Two types of material transport occur during wet spinning, their rates are determined by diffusion constants. First, a flow of solvent S∗ occurs from the spinning jet into the spin bath. Second, a stream of nonsolvent N∗ diffuses into the spinning jet. The ratio S∗ : N∗ determines the change in composition of the original spinning solution SF [1]. If, for example, S∗ : N∗ = 1, the polymer concentration in the spinning jet does not change, i.e., a change in composition would proceed along a straight line starting at SF and run parallel to the side S – N of the triangle. Figure 26 shows a general route (e) which corresponds to S∗ : N∗ > 1. Phase separation begins at the point of intersection with the binodal curve; the filament coagulates and becomes gel-like. Its composition continues to change along the binodal curve in the direction of G, and the concentration of P increases.

As in dry spinning, well-developed sheath – core structures and cavities are also formed in wet spinning due to radial diffusion. In wet-spun polyacrylonitrile fibers, for example, this leads to filaments with kidney-shaped sections. Figure 27 represents the formation of a wet-spun filament. The composition of the spinning solution within the diffusion boundary has not yet changed. The coagulation boundary indicates the position of the binodal line. The spinneret holes have a characteristic shape and are about 0.1 mm in diameter. The thickness of the spinneret is only a few tenths of a millimeter.

In wet spinning, the spin-draw ratio is usually less than 3, and occasionally even less than 1. Spin elongation still occurs, however, due to the increased concentration of polymer in the spinning jet. High frictional forces in the fluid limit the spinning speed to about 0.8 m/s, which can be compensated for by a large number of spinneret holes. A survey of common solvents and coagulants used in wet spinning is given in [2].

A mathematical formulation for wet spinning is even more difficult than for dry spinning. White and Hancock [56] were the first to carry out precise analysis and model calculations.

2.2. Drawing

2.2.1. Elongation Properties of As-Spun Yarns

As-spun polymer filaments normally exhibit a low degree of molecular preorientation and crystallization; they can therefore still undergo considerable elongation in the cold state which may be as high as five times their initial length. Depending on the properties of the material and the temperature used for testing, elongation or draw-

ing produces one of the characteristic stress–strain curves shown in Figure 28.

If the glass transition temperature far exceeds the temperature of testing or if the filament is already crystalline, the tensile force (stress) increases rapidly but the filament breaks at low elongation (strain) (Fig. 28A). Conversely, the filament can exhibit a high plastic elongation with a slow increase in force if the testing temperature far exceeds the glass transition temperature (Fig. 28C). Figure 28B represents an intermediate stress–strain curve, where the testing temperature is close to the glass transition temperature and which is typical of *cold drawing*; this term was coined in the early days of synthetic fiber development to signify that drawing occurs far below the melting point. Cold drawing proceeds as follows: in the initial stage of low elongation, deformation is elastic and reversible, i.e., the molecules retain their mutual points of cohesion. At a certain critical stress $\sigma*$ (elongation $= \varepsilon_1$), the yield point is reached where the tensile force passes through a maximum. A constriction (bottleneck) then begins to form at a random point along the drawn filament. As elongation proceeds, this constriction wanders along the filament in both directions at constant yield force, resulting in the formation of a draw shoulder (telescope effect). In this region, drawing is heterogeneous. At an elongation ε_2, the draw shoulder occupies the whole length of the filament. Since a criterion of mathematical instability is met [1], filament regions of varying diameter will always exist between elongation values ε_1 and ε_2. The constricted region exhibits a reduced diameter, corresponding to a local elongation ε_2; the remaining sites are still not drawn, however, because they lie in the elastic region below the yield point. These undrawn sites disappear only when the filament is elongated to ε_2. The remarkable drawing factor $1 + \varepsilon_2$ is called the natural draw ratio and equals the ratio of the fineness of the undrawn regions to that of the constricted regions. Once ε_2 is exceeded, deformation again becomes homogeneous with increasing force (up to the break point). If the material has a higher preorientation, ε_2 is lower and the yield force is higher.

2.2.2. Mechanism of Cold Drawing

Cold drawing can be explained at a molecular level (see Fig. 29). In the undrawn filament the molecular chains are largely disordered and coiled but are immobile just below the softening point. Since the filament is extended beyond the region of reversible elongation, the irreversible part of the mechanical work of extension is converted to heat [1], which is concentrated in the zone of flow where deformation takes place. This local heating can lead to melting if the rate of elongation is too high; it provides the molecules with the mobility required for rearrangement and sliding during extension. Once the flow process has begun, the stress–strain curve in the zone of flow begins to change due to heating, so that a maximum loading at the stationary point of draw no longer has to be overcome.

The heat liberated during extension consists of two parts: (1) internal frictional heat and (2) heat produced by thermodynamic changes resulting from orientation of the molecules under stress. The thermodynamic changes can be explained as follows. Extension results in parallelization of the chain segments and a decrease in entropy. Neighboring molecules can nevertheless still interact with one another because of their temporary mobility. They thus assume a more energetically favorable crystal structure in which atoms of parallel molecular chains occupy the lattice points of a highly ordered, compact crystal lattice. The associated release of enthalpy thus results in an overall more favorable thermodynamic state than that of the frozen, amorphous, as-spun filament. Since the filament cools or coagulates during spinning, it does not have enough time to assume the most stable thermodynamic state permitted by the temperature of the surroundings, i.e., that of unordered crystallites, in which small crystals, also called micelles, are embedded in an amorphous matrix. A partially crystalline, as-spun filament would, however, have a lower residual drawability because the crystallites stabilize its morphological structure. Only those crystallites whose molecules are aligned at right angles to the fiber axis can be broken up and rearranged during drawing [57]. The polymer macromolecules are too long to be present in crystallites in extended form; they are usually folded several times and

arranged in lamellar planes which are perpendicular to the molecular axes (i.e., at right angles to the filament axis in the oriented state). Some molecules are anchored in two or more lamellae and are called tie molecules. They thus connect different crystalline regions in the longitudinal direction of the filament, which are separated by amorphous regions. Apart from the tie molecules, the amorphous regions contain only free chain ends and loosely folded loops. Together with chain loops, the tie molecules thus provide for force transfer within such microfibrils [58].

2.2.3. The Drawing Process

If filaments (e.g., PETP) with a uniform diameter are to be obtained by a continuous cold-drawing process, a draw ratio greater than the natural draw ratio must be used. This is accomplished with two sets of rollers running at different peripheral speeds; the ratio of these two speeds is the machine draw ratio. The number of yarn loops on each roll must be sufficient to prevent slippage according to the formula for rope friction, taking into account the drawing tension. Since a small proportion of the drawing deformation is elastic, the effective draw is slightly less than the machine draw ratio. When a draw point (bottleneck) appears in cold drawing, it must be stabilized locally within the field of drawing by frictional or thermal measures to maintain steady-state drawing [this is not necessary for plastic drawing (see Fig. 28C) at an approximately uniform rate of elongation].

The dynamics of drawing can be simulated mathematically [1]. As in melt spinning, the equation of continuity applies, but simulation is more difficult because of thermal and viscoelastic peculiarities. The drawing process can, however, be investigated on a laboratory scale by carrying out creep experiments in which the time-dependent change in length is measured under constant tensile force [59], [60]. The length of the drawing field (i.e., the distance between the sets of rollers) required at a given feed velocity can be calculated from these experiments. Various approaches can be used to reduce the drawing force and/or improve drawability; they all have the same objective – to soften the as-spun material by lowering its glass transition temperature T_g (with plasticizers, swelling agents, oligomers, or residual solvent) or by increasing the drawing temperature. Softening also leads to disappearance of the telescope effect. For example, PETP ($T_g = 70\,°C$) no longer forms a neck when drawn at 75 °C. If it is drawn in water, the plasticizing action of water begins to dominate at 50 °C, so that the drawing tension approaches zero above 90 °C [61]. Water is a good conductor of heat and thus prevents overheating of the flow zone which would otherwise increase the drawing force. When drawing takes place above the glass transition temperature, however, the relaxation time of the molecules diminishes rapidly, which means that stresses decay quickly, especially during slow drawing, and the original internal orientation is then reestablished. Thus only part of the filament extension has an orienting effect; the remainder produces flow. This fact can be exploited in extreme cases to produce superfine filaments. The as-spun filaments are first drawn without orientation in a hot-water bath or some other swelling agent (just below 100 °C in the case of PETP); orientation is then increased in a subsequent draw. This method permits very high elongations, which may be as much as 20 times the original length [62].

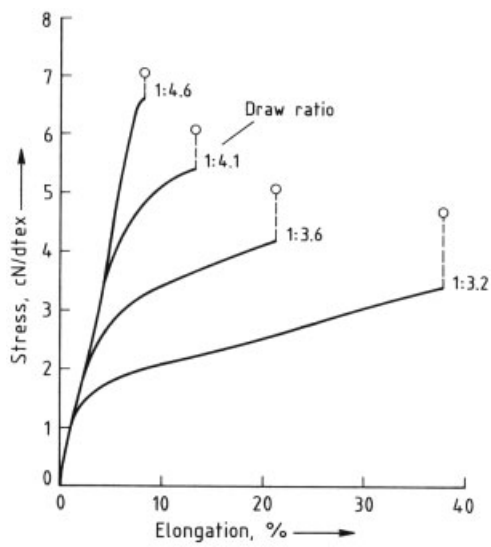

Figure 30. Stress – strain curves of poly(ethylene terephthalate) (PETP) filaments drawn at different draw ratios [12] Stress shown in the curves (σ_0) is based on the initial fineness; open circles indicate σ, the effective tenacity at break based on the actual fineness

2.2.4. Draw Ratio and Tensile Strength

Any material being spun can be subjected to varying degrees of draw. The draw ratio selected depends on the textile properties desired and the existing preorientation. As illustrated in Figure 30, the higher the draw ratio, the greater is the tenacity at break. At a higher draw ratio, the modulus of elasticity and the crystallinity increase, whereas the elongation at break is reduced.

The tenacity at break is equal to the force applied at break divided by the fineness. Since the actual fineness of the fiber at break is lower than its original fineness, the effective tenacity at break based on the actual fineness (σ) will be higher than the tenacity based on the initial fineness (σ_0). The two tenacities are related as follows:

$$\sigma = \sigma_0 (1+\varepsilon)$$

where ε is the elongation and is equal to $\Delta l/l_0$, Δl being the increase in fiber length and l_0 the initial length. In the case of low draw ratios close to the natural draw ratio, the effective tenacity is almost constant (see Fig. 30). A high draw ratio is required to obtain a large increase in tenacity.

The maximum fiber strength is obtained when the macromolecules are largely straightened out during drawing and are aligned along the longitudinal axis of the filaments so that they can crystallize. Since a molecular chain which is embedded and laterally anchored in the crystal lattice almost totally resists being pulled away along its axis, any subsequent tensile loading will put the greatest stress on the tie molecules which bridge the amorphous regions (cf. Section 2.2.2). If the load per tie molecule becomes too great, the filament breaks due to the rupture of the main bonds (C – C) of these force-transmitting chain segments. Because only a small fraction of the molecular chains (shown in the cross section of the drawn part of the filament in Fig. 29) act as force transmitters, the resultant strength is a fraction of that which is theoretically possible. Appropriate processing and use of a higher draw ratio increase the number of tie molecules and thus the tenacity at break up to a certain point by improving molecular uncoiling and alignment and preventing chain refolding. Two-step drawing may be important in this context [63]. The upper limit of drawing is reached when the tie molecules are fully extended. When the filament leaves the drawing zone, the chains in the amorphous regions recoil to some extent.

As explained previously, the value of the draw ratio used depends on the preorientation of the as-spun material which, in melt spinning, is largely governed by the take-up speed. Figure 31 illustrates the decrease in residual draw ratio with increasing take-up (spinning) speed, corresponding to a higher spinning orientation. However, this decrease is less pronounced than the concomitant increase in spinning speed. As a result, the spinneret throughput, and thus the productivity of a given spinning position, increase with increasing spinning speed, provided that the final fineness remains constant. (Only for polyamide does no advantage result from increasing the spinning speed in the range from 33 – 50 m/s.)

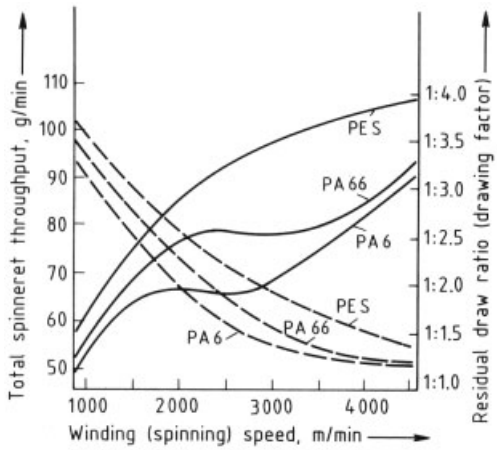

Figure 31. Total spinneret throughput and residual draw ratio as a function of winding (spinning) speed in melt spinning [64]
PA = polyamide; PES = polyester [poly(ethylene terephthalate)] ; —- Throughput; – – – – Residual draw ratio
Final yarn fineness = 167 dtex; Elongation at break = 25 %

High-speed spinning has become an important processing variant of melt spinning because only a low postdraw is then required, which is usually integrated into one of the other processing steps. High-speed spinning also reduces shrinkage due to formation of oriented crystallites [65].

2.2.5. Ultrahigh-Strength Synthetic Yarns

Maximum fiber strength would be obtained if very long (theoretically, infinitely long), completely extended polymer chains were to form an ideal, flawless crystal because they would then be in their most compact form and all be able to transfer force. In such a fiber, the macroscopic modulus of elasticity would correspond to the modulus of individual molecular chains (i.e., the strength of the C – C bonds). Theoretical tenacities of various polymers have been calculated [66]. Polyethylene is theoretically the most suitable polymer for obtaining fibers with very high strength because its chains have the smallest cross-sectional area. High-performance fibers with a very high molecular mass ($\bar{M}_w > 10^6$ g/mol) have been obtained from polyethylene by wet spinning (cf. Section 3.2.2.) [67]. The fibers have a tenacity of 3.5 GPa (35 cN/dtex), which far exceeds that of steel, but this still corresponds to only about 10 % of the maximum theoretical value. Ingenious drawing processes have been developed in an attempt to approach maximum theoretical values more closely [68], [69]. Very high tenacities have been achieved with liquid-crystalline, anisotropic polymer melts although their theoretical strength is distinctly lower than that of polyethylene.

When different materials are being compared, conversion of the textile tenacity σ in centinewtons per decitex to the tenacity σ_T in gigapascals (1 GPA = 10^9 N/m^2) may be useful:

$$\sigma_T = 0.1 \cdot \varrho \cdot \sigma$$

where ϱ is the fiber density (in g/cm^3). The term σ_T is commonly used for construction and reinforcement applications.

2.3. Crimping

2.3.1. General Description

Natural fibers, unlike synthetic fibers, have a so-called texture, which means that they are not straight but are crimped in three dimensions along their axes. This basic geometric property has shaped the design of fiber-processing machinery and largely determines the characteristics of yarns and fabric made from fibers.

The process used to impart a suitable texture to synthetic fibers which are processed into staple fiber or filament yarn is thus an important part of fiber manufacture.

The term texturing means the production of a permanent crimp in an initially straight fiber or filament yarn. Crimping is the result of local changes in the intermolecular interactions occurring in the polymers from which the fiber is made. The crimp is maintained to a greater or lesser degree during deformation or is stored as a type of memory [55], [70].

Crimp Rating. Many proposals have been made for rating crimp [71–74]; these include optical, mechanical [75], and thermomechanical [76] methods like those prescribed in DIN 53 840. Some important parameters for stuffer box crimped fiber are illustrated in Figure 32, others are:

$$\mathrm{KK} = (l_g - l_z)/l_g = \text{crimp contraction}$$
$$\mathrm{KG} = l_g/l_z = \text{crimp degree}$$
$$\mathrm{KB} = (l_g - l_b)/(l_g - l_z) = \text{crimp stability}$$

where

l_g = length of elongated (noncrimped) fiber
l_z = length of crimped fiber
l_b = length of crimped fiber after stress application

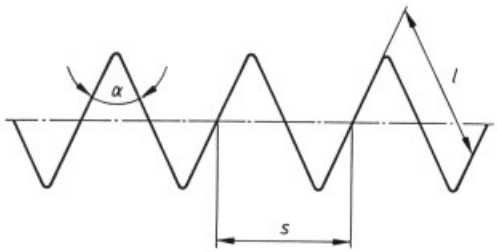

Figure 32. Structure of a crimped fiber (stuffer box)
α = buckling angle; l = buckling length; s = length of double arc

2.3.2. Crimping Processes

Various types of crimp and crimp stabilities can be obtained depending on the crimping process. Three main types are used: spontaneous, deformation, and bicomponent crimping.

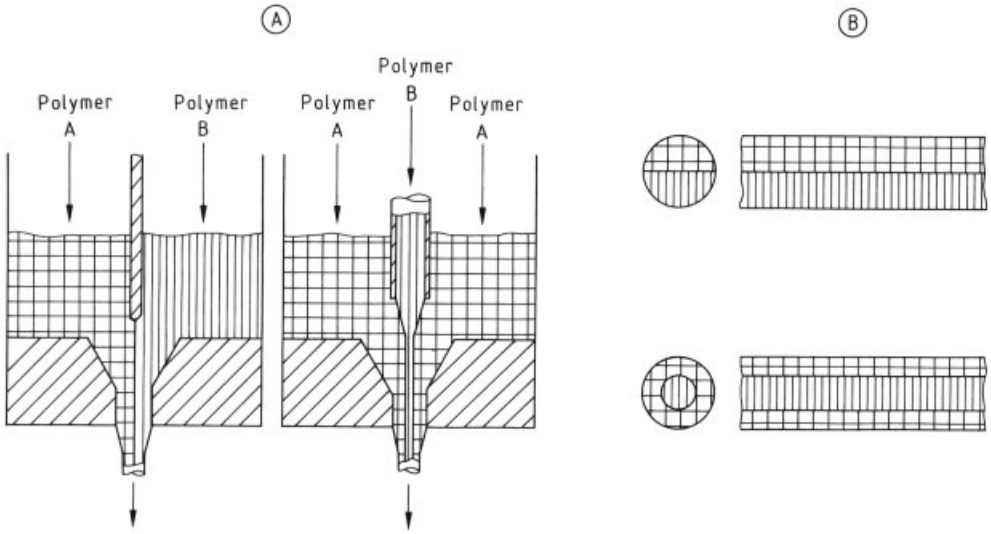

Figure 33. A) Bicomponent spinneret for side-by-side (left) and sheath – core (right) filaments; B) Transverse and longitudinal sections through a side-by-side (upper) and sheath – core (lower) filament

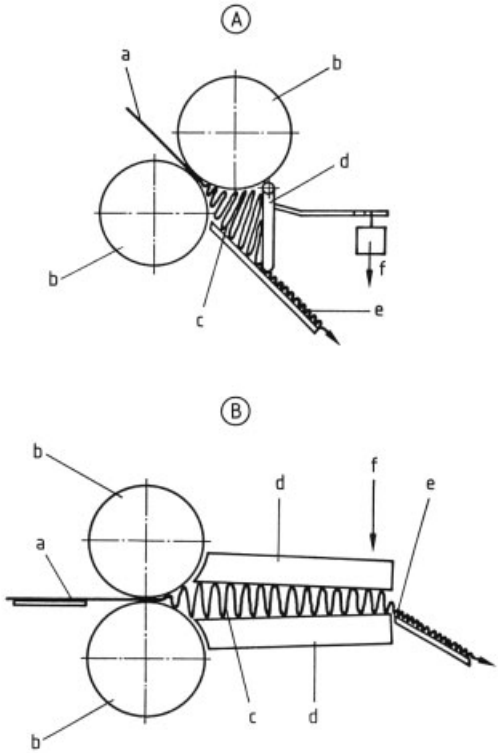

Figure 34. A) Crimper with weighted swivel flap; B) Crimper with adjustable stuffer box
a) Incoming tow; b) Crimping rollers; c) Plug of material; d) Stuffer box; e) Crimped tow; f) Pressure applied to stuffer box

2.3.2.1. Spontaneous Crimping

Spontaneous crimping is caused by the anisotropy of various fiber properties such as density, molecular orientation, or crystallinity both in its cross-sectional plane and along its axis. Anisotropy can be obtained or reinforced by applying appropriate measures during spinning and drawing. In melt spinning, for example, the filament can be cooled on one side just below the spinneret by pulling it past a cooled body [77]. Alternatively, a swelling agent may be applied unilaterally in the spinning and/or drawing region. The resulting anisotropy causes differential stresses to be "frozen" into the filament which, when liberated (e.g., in hot water without tension), produce limited differences in shrinkage and, thus, irregular crimping of the material.

2.3.2.2. Deformation Crimping

In deformation crimping, crimp is produced by application of an external mechanical force. Deformation crimping is currently the most important method for texturing both staple fibers and filament yarns. Various techniques have been developed. The fibers can be crimped in a stuffer box (cf. Section 2.3.3) by knit-de-knit technique, or by feeding them through a pair

of gears (gear crimping). In another technique called edge crimping, the heated filaments are continuously drawn over a sharp edge. Mechanical force can also be applied by means of torsion (false-twist texturing cf. Section 3.3.2) or by allowing the filaments to pass through an air jet where the turbulence blows them together into a random structure (cf. Section 3.3.3).

2.3.2.3. Bicomponent Crimping

In bicomponent crimping, two polymers which usually differ in their chemical or physical properties are coextruded from a single spinneret capillary (Fig. 33A).

By using a suitable spinneret, the polymer components may be arranged side by side or in the form of a sheath – core structure. Figure 33B shows the structures of the resulting filaments. A stable, three-dimensional crimp is formed as a result of the different shrinkage and viscoelasticity of the polymers; it depends on the distribution of the polymer mass (e.g., 1 : 1 in the side-by-side arrangement or 3 : 7 in the sheath – core structure with an eccentric core).

2.3.3. Crimping of Staple Fibers and Filament Yarns

The stuffer box principle dominates in the crimping of staple fibers. Tow of up to 6×10^6 dtex is fed through two feed rollers at speeds of up to 8 m/s into a closed chamber called a stuffer box. This box may be equipped with a swivel flap to which a load can be applied (Fig. 34A), or entire upper and lower cover flaps can be closed pneumatically about a hingepoint (Fig. 34B).

Two delivery rollers push the tow into a chamber, and a folded plug builds up due to the forces and friction in the chamber. The fibers kink between the plug and the pinchpoint of the tow in the delivery rollers; a two-dimensional, sawtooth crimp is thus produced (Fig. 35). Crimp formation is determined by the roller and chamber geometry, the chamber pressure, and friction. Crimp stability is also affected by the temperature control of the feed material, the delivery rollers, and the chamber.

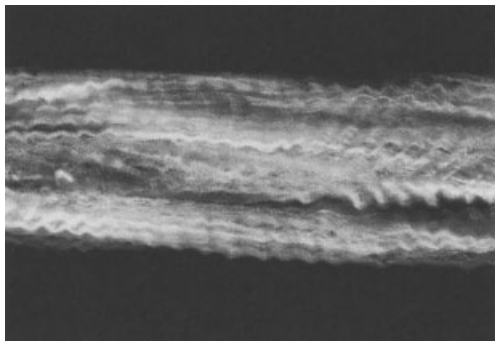

Figure 35. Tow with two-dimensional, sawtooth crimp

Many processes have been developed for texturing filament yarns, which include filament deformation, setting, and restoring of deformation in a continuous cycle. Different processes are used depending on the type of polymer and the filament fineness; they are described in Section 3.3.

2.4. Heat-Setting

2.4.1. Purpose

Drawn polymeric filaments have a high strength and elasticity, but when heated they begin to contract (i.e., shrink) at a certain temperature. This effect is enhanced in the presence of swelling agents such as water. These filaments are not generally suitable for use or further processing due to their lack of dimensional stability. Fabrics made from such filaments display, for example, poor crease resistance and poor shape retention during dyeing or washing. Therefore, the drawn filaments must be heat-set with or without a swelling agent and with or without tension. Shrinkage occurs during heat-setting and is accompanied by a simultaneous change in the fine structure mechanical properties of the filament. Since internal stresses are relieved, the fiber will not shrink during subsequent processing as long as the conditions are milder than those used for setting. However, a fiber should not be "heat-set to death" because while it is being processed into yarn, fabric, and finally, clothing, more extreme setting conditions must be used so that the article retains each new geometric shape imposed on it.

Definitions. If a filament or fiber of initial length l_0 is shortened to a length l by heat-setting or by the conditions of use, its shrinkage S and degree of setting F expressed in percent are defined as follows:

$$S = [(l_0 - l)/l_0] \times 100$$
$$F = (l/l_0) \times 100 = 100 - S$$

The degree of setting characterizes the dimensional stability of the fiber product under certain conditions of treatment and is, of course, greater if more presetting has occurred.

2.4.2. Mechanism

The dimensional stabilization of a synthetic fiber by heat-setting is made possible by the partial crystallinity of the polymers and the rearrangement of some of the crystalline regions in the amorphous matrix. The crystallites act as physical cross-links and are able to rearrange when the mobility of the molecular chains increases sufficiently during setting. Permanent, irreversible setting could otherwise only be produced by cross-linking the chains with covalent bonds as is the case, for example, in the vulcanization of rubber.

After being drawn, the fibers are generally not in thermodynamic equilibrium since stresses and thermally less stable crystallites are "frozen" in a colloidal structure. When this colloidal structure is heated to a temperature above the glass transition temperature (i.e., has a certain degree of thermal mobility), its thermodynamic equilibrium is characterized by a relative minimum in the free energy G. According to the laws of thermodynamics, G consists of the variables H (enthalpy) and S (entropy) [79]:

$$G(T,p,\sigma) = H(T,p,\sigma) - T \cdot S(T,p,\sigma)$$

where T, p and σ are the independent variables temperature, pressure, and (for a stretched fiber) tensile stress, respectively. The enthalpy H, in turn, consists of the internal energy U and the work content of the system:

$$H(T,p,\sigma) = U(T) + pV - \sigma V \ln(l/l_0)$$

where

$p\,V =$ pressure energy

$\sigma\,V \ln(l/l_0) = \sigma\,V\varepsilon_{\mathrm{H}} =$ stress energy

(V = volume and ε_{H} = Hencky measure of elongation related to the stress-free length l_0).

Stress energy is negative with respect to pressure energy because external energy is supplied to the system when the fiber length is extended ($l > l_0$), whereas the system must do compression work on the surroundings if expansion occurs in all directions.

When changes in state occur at constant pressure, the amount of heat exchanged with the surroundings is represented by ΔH; according to convention, ΔH is negative if heat passes from the system to the surroundings. For example, in the isochoric extension of a rubber band, the work of extension is liberated as heat and the rubber is temporarily warmed.

Hence, when a drawn polymeric fiber is heated above its glass transition temperature, it tends toward a more stable state by reducing its free enthalpy. This can be accomplished either by an increase in entropy (recoiling of chain segments) or by a decrease in enthalpy (by liberating heat due to crystallization, whereby U and pV decrease). Contrary to [80], however, the work of shrinkage performed in response to a tensile stress does not contribute to a decrease in enthalpy; on the contrary, in so doing the fiber would take up an equivalent amount of heat from the surroundings (Gough – Joule effect).

The amorphous matrix between the larger, more stable crystallites contains many small crystallites with some defects in ordering. Since stabilization can be achieved in two opposing ways, heat-setting entails a disproportionation of the colloidal structure (see Fig. 36) [80], and the amorphous matrix separates into two structural regions. On the one hand, the amorphous regions loosen due to relaxation of blocked stresses in the molecular chains (recoiling and displacement of segments); this is connected with the breakup of the least thermally stable, stressed crystallites and a certain amount of shrinkage. On the other hand, the more stable crystallites grow to form larger, more ordered regions which prevent further fiber shrinkage. An increase in the short-range order and/or degree of crystallinity can be measured by following the increase in average fiber density, which represents an interpolation between the density of the amorphous and crystalline regions. This two-phase

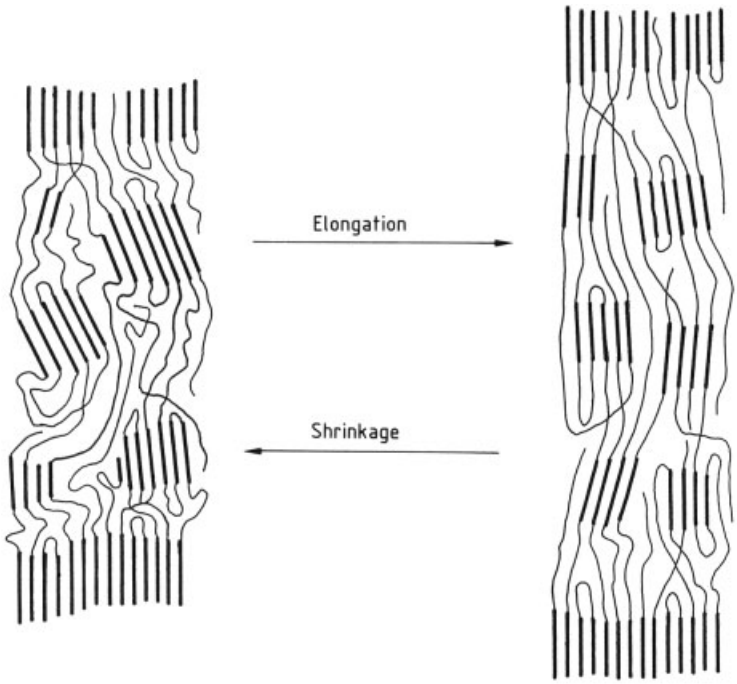

Figure 36. Colloidal structure of a polymeric filament
Crystalline regions indicated by thick lines

model is not, however, completely accurate because the packing density of the two phases can vary.

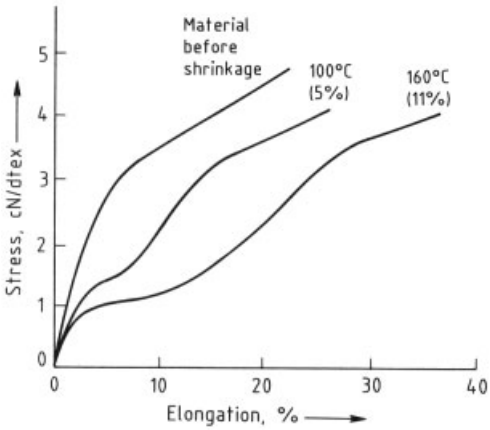

Figure 37. Stress – strain curves of poly(ethylene terephthalate) (PETP) filaments shrunk freely at different temperatures [12]
Filaments were heat-set without tension, shrinkage obtained during setting is indicated in parentheses

Effect of Stress on Heat-Setting. The minimum free energy, and hence optimal thermal stability, at the heat-setting temperature is obtained if the fiber can yield to the retractive force $(\partial G/\partial l)_T$ and is able to shrink freely [80]. The disadvantage of this alternative, however, is that the loosening of the amorphous regions flattens the stress – strain curve of the fiber (see Fig. 37). Tenacity at break decreases and elongation at break increases by approximately the amount of free shrinkage. In addition, the height of the elastic region is reduced by the formation of a flow saddle.

This deterioration of mechanical properties can be avoided by heat-setting the fiber at constant length under a force that opposes the shrinkage force (e.g., by setting on hot rollers immediately after continuous drawing). The orientation (i.e., entropy) and the stress – strain curve then remain practically unchanged [81]. Even without external shrinkage, the fiber can relax by means of internal creep due to its viscoelastic properties [82]. Crystallization then remains unaffected [81] or may even be enhanced if the applied stress is low because the

enthalpy – elastic portion of the restoring force is negative in the initial phase of extension [80]. However, because stresses applied at constant length do not completely relax to zero, a small amount of shrinkage takes place on unloading [80]. Because the heat-setting equilibrium attained under stress has a higher free enthalpy than that of the stress-free system, a certain residual shrinkage occurs when fibers are used or tested under stress-free conditions. The total shrinkage obtained after heat-setting under tension and heat treatment without tension is, however, much smaller than that obtained after heat-setting without applied tension [62], because the crystallites created under stress evidently stabilize the anisotropic composite structure more effectively. However, since the structure is only slightly loosened, properties such as dye uptake are also reduced.

Other Important Variables. If setting takes place in a medium which also acts as a swelling agent, the glass transition temperature is lowered by ΔT_g. This means that the temperature T at which setting is performed corresponds to an effective heat-setting temperature T_{eff} in an inert medium:

$$T_{eff} = T + \Delta T_g$$

Thus, for polyamides saturated with water, ΔT_g is approximately 80 °C [80]. This means that polyamide fibers must be heat-set in hot air without tension at a minimum temperature of 180 °C to be free of shrinkage when boiled in water without applied tension. Conversely, heat-setting in water at 100 °C provides resistance to shrinkage in hot air up to 180 °C. (For PETP, ΔT_g in water is ca. 30 °C [61].) Hence, drying can have a simultaneous heat-setting effect.

Other variables that affect the degree of thermal stabilization are the rate of heating (e.g., greater with saturated steam) and the residence time. If the rate of heating is too high, the setting shrinkage increases because disintegration of the unstable regions is more rapid than consolidation of the structure by postcrystallization. At higher setting temperatures, equilibrium is attained more quickly.

Effective Temperature and Effective Stress. The effective heat-setting temperature T_{eff} corresponds to the melting temperature of the crystallites formed in the matrix, which are relatively unstable compared to the main micelles. As discussed above, T_{eff} does not necessarily correspond to the actual temperature of treatment. At thermodynamic equilibrium of the crystalline – amorphous phase change ($\Delta G = 0$), the melting temperature T_m equals $\Delta H_m / \Delta S_m$ (ΔH_m and ΔS_m being the enthalpy and entropy of fusion, respectively). The enthalpy of fusion and hence the melting temperature are lower than for the overall melting of the fiber because the crystallites which are influenced by heat-setting are under tension; the internal energy of the matrix system is therefore not reduced fully during crystallization. Since melting of the crystallites is always associated with shrinkage, T_{eff} of a material can be determined in two ways: (1) by determining the onset of shrinkage on a shrinkage – temperature curve, and (2) by using differential scanning calorimetry (DSC). In addition to a major peak, the DSC diagram exhibits a smaller melting peak which appears at the effective temperature [83]. The heat of melting measured by DSC is also a measure of average crystallinity. Furthermore, stresses locked into the fiber and hence the effective stress of heat-setting, σ_{eff}, can also be determined by measuring the equilibrium shrinkage force [78], [83].

2.4.3. Shape Retention and Molecular Memory

Every solid with elastic properties is characterized by a stress-free reference state which is in thermodynamic equilibrium. Perturbation of this equilibrium by a mechanical strain ε at a temperature lower than the effective heat-setting temperature T_{eff} creates a restoring force which strives to reestablish the equilibrium. The stabilized structure, which represents a molecular memory of the equilibrium shape, has been called a *setting system*; the nonequilibrated shape with the perturbed molecular interaction has a been termed a *stressed system* because it generates stresses which strive toward the original shape [80]. Shape is not restricted merely to fiber length but includes all three dimensions,

e.g., crimp in the form of bends. If a fiber is bent forcibly, the outer, convex side is extended locally, whereas the inner, concave region is compressed. Because the fiber "remembers" its straight form (setting system 1, F 1) it normally straightens out when the force is removed. If deformation is conducted at a temperature higher than T_{eff}, the stressed system (stressed system 2, S 2) can be converted to molecular setting system 2 (F 2) by heat-setting (described in Section 2.4.2). At a higher temperature, sufficient energy of activation is available to obtain a minimum in the free energy; this transformation is thus irreversible and the memory of F 1 is erased and replaced by F 2. When the crimp is stressed mechanically, F 2 is transformed into a stressed system 3 (S 3). As long as deformation is reversible, crimp will later be restored. Even crimp which has been extended severely at a temperature below the glass transition temperature recovers when heated to a temperature higher than T_g but lower than T_{eff} because of its imprinted memory [80].

3. Manufacture of Filament Yarns

3.1. Melt Spinning of Filament Yarns

3.1.1. General Aspects

Multifilaments of polymeric materials were first produced by melt spinning in 1940. Subsequent developments have led to much higher operating speeds, throughputs, and product quality [84].

Spinning processes are categorized according to the processing speed and type of end product (Table 1). Speciality products may require speeds other than those listed here.

Depending on their production speed, the yarns have different properties, as can be seen in the strain – stress diagram (Fig. 38).

According to end use requirements, the luster of filaments is changed by modifying the cross section or by varying amount of dulling agent in the polymer. Spin dyeing by adding a suitable dye to the polymer prior to spinning is another possibility.

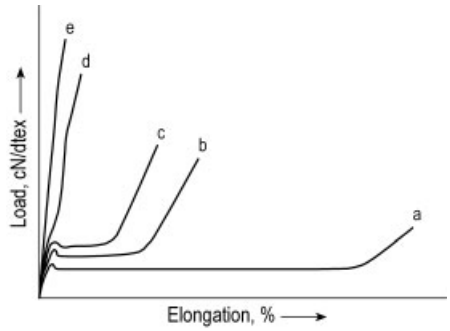

Figure 38. Stress – strain diagram for PETP produced at different production (winding) speeds
a) Undrawn yarn; b) Medium-oriented yarn; c) Partially oriented yarn; d) Highly oriented yarn; e) Fully drawn yarn

According to the application, yarns are produced in different counts and with different numbers of single filaments (Table 2).

3.1.2. Conventional Spinning

Spinning systems are generally arranged to give a vertical downward process flow over several floors, as shown in Figure 39 [61]. In the *chips spinning process*, polymer chips are brought to the required state of dryness by batchwise or continuous drying. They are then fed pneumatically to the chips silo (a) from where they fall by gravity into the feeding opening of the extruder (b). The chips are transported by the rotating extruder screw (feeding zone) and melted between the electrically heated barrel and the screw (melting zone). Adequate pressure and uniform output of the polymer is obtained by means of a progressive increase in the core diameter of the screw (metering zone). The speed of the screw is controlled to ensure that the melt is fed to the distribution line (d) at constant pressure even if the consumption of the spinning machine varies.

In the direct spinning process, the liquid polymer is fed from the polycondensation stage directly to the distribution line (d) via a pressure pump; the solid-phase stages of chips cutting and drying are omitted.

Polymer of insufficient purity is prefiltered by means of a main melt filter (c). The filter inserts can be changed semi- or fully automatically when their service life is over without disturbing the spinning process. Metal filters, sintered metals, and powdered metal, or silicic materials

Table 1. Classification of spinning processes

Yarn type	Description	Processing speed, m/s	Processing speed, m/min	Product
UDY	undrawn yarn	8 – 30	500 – 1800	intermediate
MOY	medium-oriented yarn	28 – 43	1700 – 2600	intermediate
POY	partially oriented yarn	43 – 83	2600 – 5000	intermediate
HOY	highly oriented yarn	83 – 100	5500 – 7500	final
FDY	fully drawn yarn	75 – 83	4500 – 6500	final

Table 2. Classification of yarns

Class of filament	Titer per single filament, dtex	Number of filaments	Total filament count, dtex
Monofil (textile)	7 – 20	1 – 3	10 – 22
Regular	2 – 6	7 – 72	22 – 167
High-count	1 – 2	48 – 120	55 – 167
Microfilament	0,5 – 0.99	48 – 192	55 – 167
Superfine filaments	< 0.5	120 – 384	55 – 167

with a defined pore size are used as filter media; a combination of woven and nonwoven metal filters is usually preferred.

The melt distribution line (d) carries the polymer with equal residence times to the spin heads in the spin beam (e). Each spin head contains all the elements for the preparation of the melt and formation of the filaments. Individual spin heads can be shut down by means of freezing or mechanical valves if servicing of the spin pump is required. The spinning system is vapor-heated (g) using a mixture of diphenyl – diphenyl oxide ($bp = 256\,°C$). For different temperatures, a corresponding over- or underpressure is maintained. The vapor is generated by an electrically heated boiler which may be separate or integrated into the lower section of the spin beam.

In the spin beam, the melt stream is carried to multiple gear pumps (f) where it is metered and conveyed to the spin pack (h) at a spinning pressure of 5 – 40 MPa. In the spin pack the melt is filtered, homogenized, and transported to the spinneret which contains the desired number of capillaries (3 – 350) with a circular or shaped cross section. Depending on the polymer, filament fineness, and spinning speed, the diameter of the capillary holes is 0.15 – 0.8 mm and the capillary length one to five times their diameter. Filtration is performed with the same filter media as described above for the main melt filter, a combination of different filter media may also be employed. As a rule, at a single spinning position several yarn ends (2 – 16) are spun, quenched, finished, and wound up either together or in two groups. Several spinning positions producing the same product are combined into a spinning machine (cf. Fig. 40). The spin packs have a limited service life and can be changed after stopping the relevant spin pump.

The liquid filaments emerge vertically from the spinneret and are cooled to room temperature by using transverse or radially blown air with a temperature of 18 – 26 °C and a relative humidity of 50 – 70 % (i). Depending on the process and the production schedule, the length of the blow-air zone is 25 – 180 cm and the air velocity ranges from 0.25 to 1.2 m/s. The cooled yarn ends are guided vertically to the take-up by the conditioner (j). The take-up (Fig. 40) contains all the elements needed for the posttreatment and winding of the yarn.

A film of finish is applied either after the yarn is cooled in the quenching duct [Fig. 39 (i)] or when it leaves the conditioner [Fig. 39 (j)] at the take-up [Fig. 39 (k)]. The finish consists of a mixture of mineral and/or ester oils and emulsifiers which improve the cohesion of the filaments, as well as their electrostatic and frictional properties. The yarn is covered with a film of finish by running it tangentially over a ceramic roller which rotates at a constant speed in a trough containing the finishing oil [Fig. 39 (k)]. Several yarns can be finished simultaneously with one roller. A more precise method of applying finish involves metering the finish for each yarn by means of a gear pump and a ceramic applicator pin. The amount of finish applied is determined by the speed of the rollers

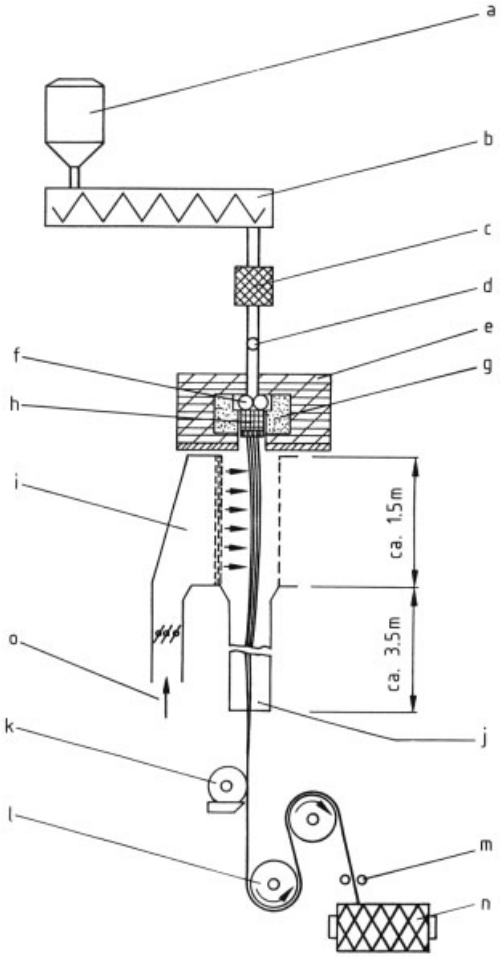

Figure 39. Schematic of melt spinning
a) Chips silo; b) Extruder; c) Melt filter; d) Melt distribution line; e) Spin beam; f) Spin pump; g) Vapor-heated jacket; h) Spin pack; i) Quenching duct; j) Conditioner; k) Finish applicator; l) Godets; m) Traversing guide; n) Spin bobbin; o) Blow air

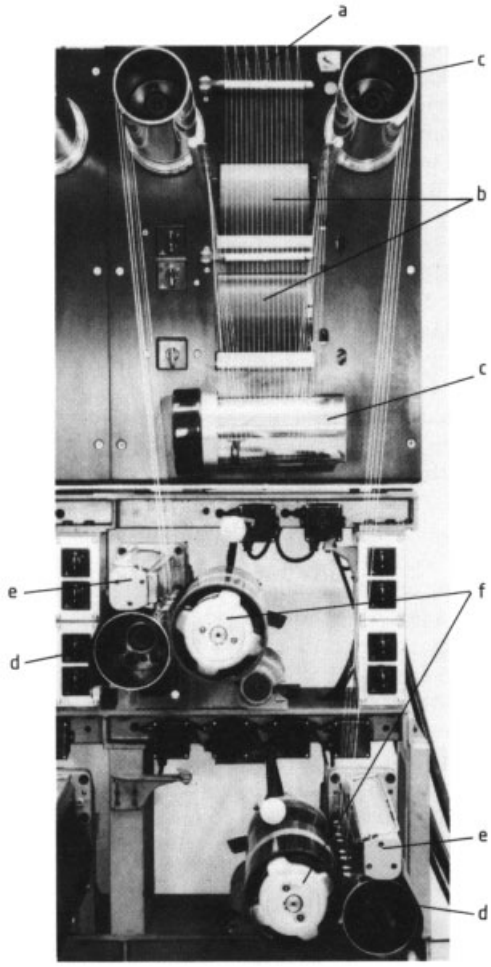

Figure 40. Take-up unit for undrawn yarn (UDY) with 16 ends per spinning position (only one position is shown)
a) Filament yarn; b) Finishing-oil roller; c) Godets; d) Friction roller; e) Traversing unit; f) Chuck with eight yarn packages
Courtesy of Barmag, Remscheid, Germany

and pumps. The finishing oil can be emulsified in water, dissolved, or used neat.

As shown in Figure 39, the finished yarn ends pass over two godets (l), which determine the spinning speed, and are then led to the winder. The yarn is wound by means of a traversing guide (m) onto cylindrical cardboard or resin tubes to form spinning bobbins (n). A friction roller ensures that the speed of the tube with its wound yarn remains constant. The increase in diameter of the bobbin is compensated by backward motion of the winding head with the friction roller or of the chuck bearing the wound yarn. Traverse motion is usually performed by a simple cam-shaft traversing systems that moves a ceramic guide along the bobbins stroke. The traverse speed is modulated at frequent intervals to prevent slippage of yarn layers when package patterns are formed due to an even-number ratio between the rates of bobbin rotation and traverse. The yarn is threaded into the take-up

by means of a movable injector gun powered by compressed air.

Undrawn yarn (UDY) is wound at 8 – 30 m/s and has a low orientation (LOY). It can only be stored for short periods and must be drawn three to four times in a subsequent drawing process to obtain a product with properties that are suitable for its end use [2], [61], [84]. To attain high yarn strength for industrial uses, the yarn is passed through a heated tube with a length of ca. 30 cm (hood heater) after the spinneret.

Auxilliary plants are necessary to provide auxilliaries, raw materials etc. in the required quality neccesary for production of the filaments:

- Chips drying to adjust water content (if extruder spinning is used)
- Spin-pack, spin-pump, and melt-filter cleaning and servicing
- Spin-finish make-up
- Winder maintenance
- Quality control for raw material and final product

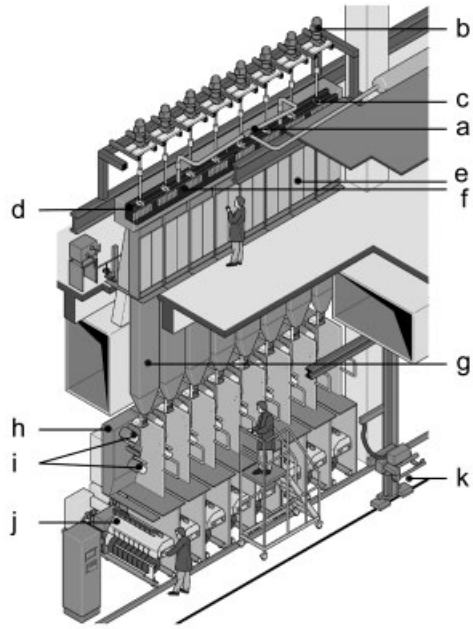

Figure 41. Spinning machine for partially oriented yarn (POY) with S-wrap-around godets (8 ends per spinning position)
a) Melt distribution line; b) Motor for spin drive pump; c) Spin pump; d) Spin head; e) Blow duct; f) Spin pack; g) Conditioner; h) Take-up unit; i) Godets j) Winder; k) Automatic doffing system
Courtesy of Barmag, Remscheid, Germany

3.1.3. High-Speed Spinning

Since 1970, spinning at conventional winding speeds of 8 – 30 m/s has been replaced in many areas by high-speed spinning with winding speeds of 43 – 75 m/s. High-speed spinning yields a partially oriented yarn (POY) which subsequently has to be drawn 1.2- to 2.2-fold for further use. Since POY has a higher degree of orientation than UDY, it has a longer storage life and further external processing (e.g., texturing) is thus possible. In general, a higher productivity per spinneret is achieved (see Fig. 31). The higher initial modulus makes POY less delicate than UDY; higher operating speeds and yields are therefore possible in subsequent processing.

High-speed spinning proceeds basically as described in Section 3.1.2; a typical spinning unit is shown in Figure 41. As a result of the high operating speed, higher demands are made on the polymer itself, its handling, its filtration, and the equipment. The increased take-off speed results in a comparatively high yarn winding tension so that godets may be required to control winding tension especially for high count and low-dtex single yarns.

After quenching, a finish is normally applied at a variable distance (50 – 200 cm) below the spinneret to control yarn tension. Finish is usually applied by means of applicator pins and metering pumps.

The yarn is frequently interlaced immediately before winding to improve yarn take-off from the package. The individual filaments are entangled by directing the yarn past a nozzle made of sintered ceramic which emits a jet of compressed air at a pressure of 0.3 – 0.6 MPa [2], [86].

The yarn is finally wound on specially developed high-speed winders which nowadays usually are equipped with automatic doffing (changing) devices (Fig. 42). Automatic winders operate practically without waste and are labor-saving. Often they are combined with an automatic doffing system which removes the full

Figure 42. Automatic doffing system for unloading the full bobbins from the winder
Coutesy of Toray Engineering Co., Ltd.

bobbins from the winder and replace them with empty tubes.

The bobbin chuck is usually driven directly by a motor (axle drive). The peripheral speed of the bobbin is kept constant by controlling either the number of revolutions of a friction-driven sensor roll or by means of the yarn winding tension (godet machines only). The yarn is distributed over the traverse stroke by means of a shuttlelike moving traversing guide onto cylindrical cardboard or for higher speeds with the aid of two or more opposing rotors over the width of the traverse (Fig. 43)[88]. One winder is capable of simultaneously producing two to ten packages on one chuck with a weight of 8 – 40 kg, depending on the number of ends.

To ensure good package formation and unwinding conditions, traverse modulation (see Section 3.1.2) can be supplemented by a step precision winding unit which changes the traverse speed by means of a microprocessor [89]. Typical winding speed is 50 m/s for PETP POY, with modified polymer 83 m/s, and for Nylon POY 70 m/s.

3.1.4. Superhigh-Speed Spinning

Since about 1980, yarn with increased orientation has been obtained by increasing the winding speed above 83 m/s to 133 m/s; 117 m/s is the maximum speed used commercially. The product is a flat yarn which usually requires no further drawing. The yarn displays a lower strength, lower initial modulus, and higher elongation in comparison to mechanically drawn yarn. In the case of PETP the boil-off-shrinkage of the yarn is much lower, too. The modified yarn properties result in a limited application window for this product.

Production is basically the same as that described in Sections 3.1.2 and 3.1.3, but greater demands are made on the equipment and the polymer because of the extremely high operating speeds. The yarn is spun without mecanical drawing, and final elongation is determined by take-up speed and the titer of single filaments. Tension control with godets is necessary if a spinning machine of conventional height is used. In compact spinning [85], omission of the conditioner between the quenching zone and the winder reduces the height of the spinning unit from ca. 12 m (conventional height) to ca. 6 m. This decrease, mainly in the thread-pass length, reduces yarn tension by lowering the air drag, so that godets can be dispensed with.

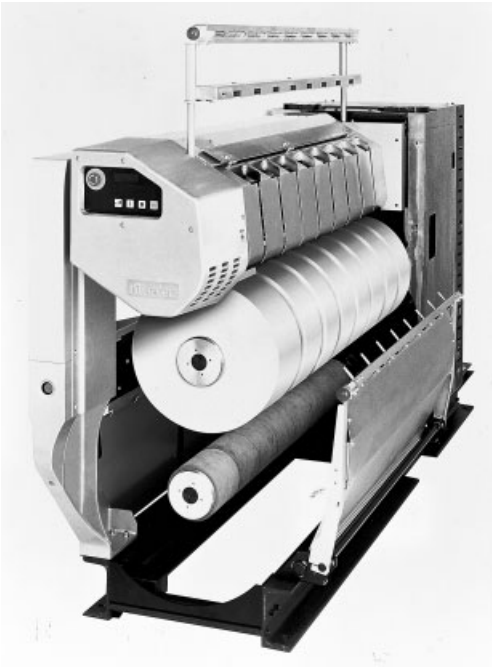

Figure 43. Eight-end winder with an automatic doffing system for 2500 – 6000 m/min with rotating traverse system Courtesy of Murata Machinery Ltd., Osaka, Japan

The yarn may also be reheated in a heating tube 1 – 3 m in length immediately after quenching to obtain a further increase in orientation [87]. Two open, looped godets are used in this case to control yarn tension. Due to the very high yarn speed, only winders with a rotating traverse system are in use for this process (cf. Fig. 43) [88].

To ensure good unwinding conditions, traverse modulation (see Section 3.1.3) can be supplemented by a step precision winding unit which changes the traverse speed by means of a microprocessor [89]. To improve cohesion and take-off properties, the yarn is commonly interlaced as described in Section 3.1.2. Because of the high speeds involved, manual string up is performed at lower speeds by means of an injector, after which operating speed is reached automatically, and an automatic doff is used to segregate yarn with unsuitable operating speed. Further doffs are carried out automatically at full operating speed [2].

3.1.5. Draw Twisting and Draw Winding

The final properties of low-oriented yarns, produced as described in Section 3.1.2 and 3.1.3, are established by subsequent drawing. Drawing results in higher strength, lower elongation at break, and in some cases, lower boil-off and hot-air shrinkage (drawing is discussed in detail in Section 2.2). The final product is a flat yarn with suitable properties for weaving, knitting, and industrial applications. In *draw twisting,* the yarn is simultaneously drawn and twisted, whereas in *draw winding* it is simply drawn and wound.

In many areas drawing machines have been replaced by integrated, high-speed processes due to their lower labor costs and their higher yarn quality and outputs, as described in Section 3.1.8.

The path of the yarn in the drawing machine can differ depending on the polymer used. Figure 44 shows the path for PETP, for which the hot plate (g) and the second godet (e) are heated, while the other godets are cold. For nylon 6 and nylon 66, which are drawn cold [84], [90], the hot plate and second godet are omitted. A draw ratio of 1/1.2 to 1/5 is applied between rollers (a), (c), and (e) for PETP, and between (a) and (c) for nylon. The yarn is directed onto the cop (i) by a ring (j) that moves up and down and holds a traveler (k). The traveler is pulled along the ring by the yarn-winding tension. The difference between the speeds of the traveler and the cop twists the yarn while it is being wound (draw twisting). Yarn buildup on the cop (i) is programmed, the most common patterns being

differential, cop, and parallel winding. The take-up speed is 12 – 24 m/s for textile yarn [2].

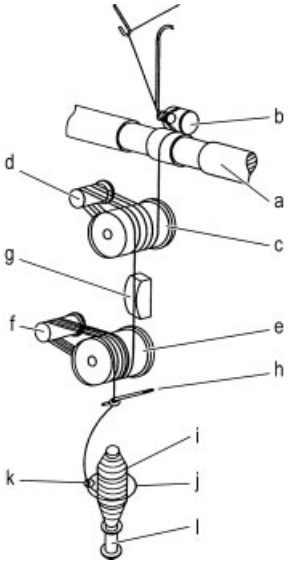

Figure 44. Draw twisting of PETP yarn
a) Feed roller; b) Pressure roller; c) First draw roller (godet); d) Idler; e) Second draw roller (godet); f) Idler; g) Hot plate; h) Pigtail guide; i) Cop; j) Ring; k) Traveler; l) Spindle

Figure 45. Draw winding machine with two godets and a hot plate, two ends per position
Courtesy of Zinser Synthetic Systems, Germany

Draw winding is carried out similarly, but instead of being twisted the yarn is wound onto cylindrical packages rather than biconical cops by using a winder at each drawing position. The wound yarn is practically free of twist. Take-up speed is 20 – 28 m/s; on some machines up to 33 m/s is possible. The temperature of the draw rollers and heaters is monitored by scanners. Interlacing of the yarn before winding (cf. Section 3.1.3) is often used to improve cohesion, especially in draw winding.

Industrial yarns are processed on draw winding machines of reinforced design, similar to that illustrated in Figure 45. Depending on the end use of the yarn, the machine may have up to three drawing zones, three hot draw rollers, three drawing pins, and two hot plates in the drawing fields. The total draw ratio can be as high as 1/7. Take up of the yarn and interlacing is as described above. The take-up speed is 3 – 13 m/s [2].

Most industrial yarn nowadays is produced by one-step processes; speciality products are still produced by draw winding or draw twisting.

3.1.6. Draw Warping

Undrawn or partially drawn textile filament yarn, produced as described in Sections 3.1.2 and 3.1.3, is subjected to integrated drawing and warping to produce section beams for use in knitting and weaving. The drawing process described in Section 3.1.5 is then omitted. The feed yarn must be of above-average quality to minimize machine stops and any resultant dyeing defects.

Draw warping is illustrated schematically in Figure 46. The yarn is taken off 600 – 1400 bobbins arranged in a creel (a) and fed into the drawing unit as a warp via yarn brakes and a reed with eyelets (f). A series of seven to ten sets of rollers with S loops are arranged in three groups (b) – (d) to form a drawing zone (draw ratio 1 : 1.2 to 1 : 5) and a relaxation zone. Drawing can also be assisted by incorporating a heated water bath, a heated draw pin, or heated feed rollers in the drawing zone. The yarn warp is then heat-set by guiding it over a hot plate (e). Yarn defects (slubs) are located by means of optical detectors (g); the unit is stopped automatically so that defects can be removed. To minimize damage to the yarn during machine stoppage, the hot plate is swung out of the way, the heated rollers are cooled, and the draw pins are rotated so that the yarn is in contact with the unheated side. Yarn cohesion is improved by applying a finishing oil by means of a wetted roller rotating in a trough (h). Yarn cohesion is further improved by interlacing the yarns with an entangling reed (f) be-

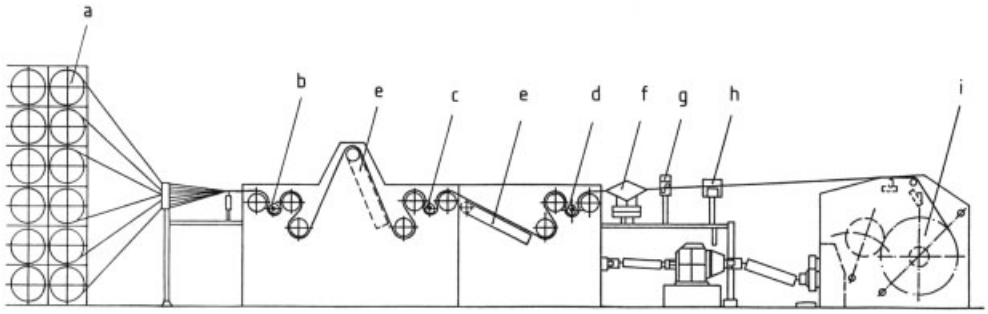

Figure 46. Schematic of a draw-warping unit
a) Creel; b) Feed rollers; c) Draw rollers; d) Relaxing rollers; e) Hot plate; f) Tangle reed; g) Slub detector; h) Finish applicator; i) Beam
Courtesy of Karl Mayer, Obertshausen, Germany

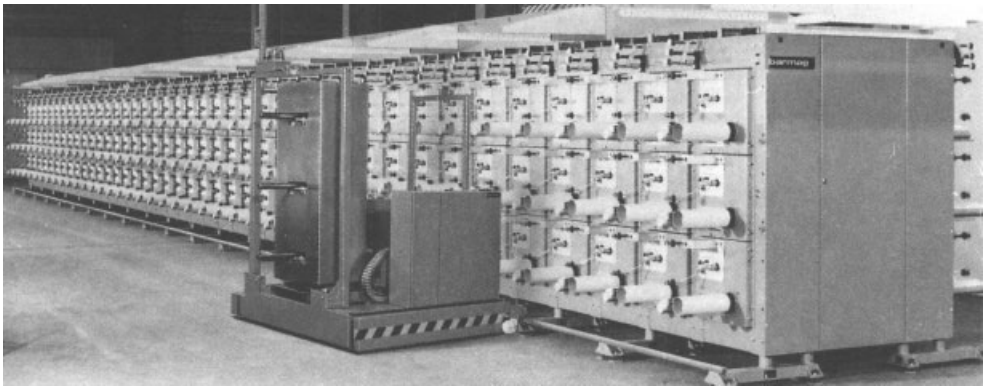

Figure 47. Yarn take-up unit with 216 tension-controlled winders and an automatic doffer
Courtesy of Barmag, Remscheid, Germany

fore the oil is applied; the filaments of one yarn end are interlaced as they pass a jet supplied with compressed air at a pressure of 0.3 – 0.6 MPa. Some machines are equiped with a sizing process instead to improve yarn cohesion further for delicate yarns that are to be woven. The finished yarn is wound under constant winding tension at speeds up to 10 m/s to form a section beam (i). Sized yarn is wound at speeds of up to 3.3 m/s [93–96].

3.1.7. Horizontal Drawing

Undrawn or partially drawn yarn [92], produced as described in Sections 3.1.2 and 3.1.3, can also acquire its final properties in the form of a warp on a multi-end drawing machine rather than a draw twister. Horizontal drawing is used primarily to process industrial yarn. Since a relatively low drawing speed is used (2 – 8 m/s), yarns with very low shrinkage are obtained due to the comparatively high residence time in the machine's heater.

The yarn ends are drawn off a large number of bobbins placed in a creel and are then fed into the drawing unit as a warp. The drawing unit consists of five to nine open, heated feed rollers and a subsequent, similarly constructed set of draw rollers with 1.2 – 7-fold greater surface speed. The warp is looped around the two sets of rollers and drawn as a result of their different speeds. Hot-air shrinkage of the yarn is reduced to the desired value by heating it in a channel between the draw rollers. The yarn is then looped around a third, unheated set of draw rollers with three to seven open, looped feed rollers; these rollers run at a slower speed than the previous set and thus permit a defined relaxation of the yarn. The yarn is cross-wound on cylindrical bobbins in a bobbin creel. The speed of the winders is con-

trolled by yarn tension. A yarn take-up unit is shown in Figure 47. If necessary, yarn cohesion can be improved by interlacing individual yarn filaments or by applying a finishing oil as described in Section 3.1.6.

3.1.8. Spin Drawing

In spin drawing, the spinning and drawing processes described in Sections 3.1.2, 3.1.3, and 3.1.5 are combined into a one-step process which produces a flat, ready-to-use, fully drawn yarn (FDY) for textile or industrial purposes. Spin drawing has largely replaced the conventional two-step process because it is less labor-intensive, quality control costs are lower, and a better quality yarn is produced. A spin-drawing machine is shown in Figure 48.

Figure 48. Spin-drawing machine for industrial yarns, with four ends per spinning position
Courtesy of Rieter, Winterthur, Switzerland

For textile application spinning is carried out as described in Sections 3.1.2 and 3.1.3. In the 1970s conventional spinning speeds of up to 17 m/s and winding speeds of ca. 67 m/s were used. Today, high-speed spinning at 37 – 75 m/s with winding speeds of 70 – 83 m/s is preferred; this results in higher throughput and more uniform quality. After finish has been applied, the yarn is looped several times around a feed godet with an idler roll or around a pair of godets. The feed godet may be heated inductively if necessary (e.g., in the case of PETP). A second, similarly constructed drawing unit runs 1.2 – 4 times faster (the draw ratio), at a temperature of 130 – 180 °C. This unit draws the yarn, relaxes it, and allows thermal shrinkage to be adjusted to the desired value.

If processing speed is sufficiently high, the godets do not necessarily have to be heated, relaxation of the yarn prior to winding (e.g., by means of steam) is sufficient [97].

In the production of industrial yarn, the spinning speed is usually 7 – 12 m/s. Three to four sets of godets are needed so that a total draw ratio of 1/5 to 1/7 is obtained in two drawing steps. The temperature of the godets is between 80 and 220 °C. A final drawing zone with a draw ratio between 1/0.85 and 1/1.05, serves as a relaxation zone. The yarns are interlaced as described in Section 3.1.4 to improve their cohesion for further processing [2]. For the production of industrial yarn according to the HMLS (high-modus, low-shrinkage) process, spinning speed is between 30 and 38 m/s. Three to four sets of godets are used for a total draw ratio between 1.18 and 1.22. Winding speed is between 93 and 108 m/s.

3.2. Solution Spinning of Filament Yarns

The basic principles of solution spinning are described in Section 2.1.4. Examples of polymers which are spun commercially into filament yarn from solutions are polyacrylonitrile and its copolymers [98], poly(vinyl alcohol) [99], and aramids [2]. Filament yarns with high heat resistance, a high modulus of elasticity, and/or a high strength (tenacity) have also been introduced. These yarns are made from polybenzimidazole, poly(amide – imide), polyoxydiazole, polytriazole, and other special polymers [2] and can be produced only by solution spinning.

Economic preferences for wet or dry spinning cannot be generally defined because process economics depend strongly on use of the optimal solvent. The required properties of the filament yarn may also influence the choice of process. In the case of poly(vinyl alcohol), for example, silklike yarn is obtained by dry spinning [100], whereas wet spinning produces high-strength, high-modulus yarns [101].

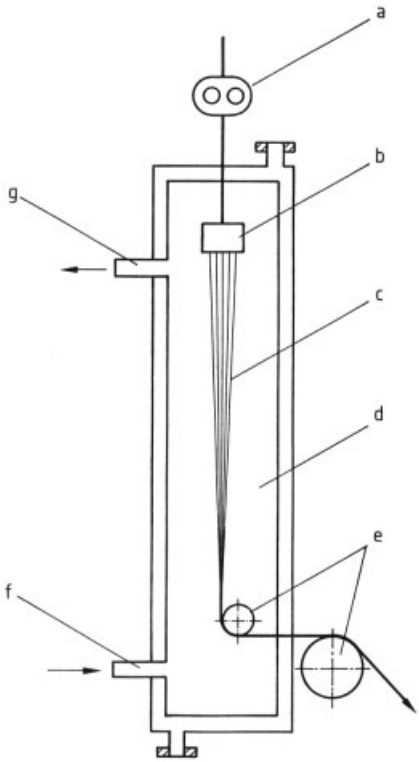

Figure 49. Schematic of a dry-spinning shaft
a) Spin pump; b) Spinneret; c) Filament yarn; d) Drying chamber; e) Take-up roller; f) Inlet for drying gas; g) Outlet for drying gas and solvent vapor

3.2.1. Dry Spinning

Strict requirements are placed on the solvents used in dry spinning. In addition to being excellent solvents for the polymer in question, they must be nontoxic, and must have a low heat of evaporation, and a low risk of explosion. The spinning solutions can be prepared continuously or batchwise in commercial solubilizers, kneaders, or screws. Solutions should be as concentrated as possible and may contain up to 45 wt % of solid material, depending on the polymer and its degree of polymerization. The highly viscous spinning solution is filtered and degassed; it is then conveyed, if necessary in heated pipes, to the spin pumps and the spinnerets by means of extruders or gear pumps. The spin pumps and spinnerets are located in the upper part of a spinning shaft, which is shown schematically in Figure 49. The shaft is 4 – 8 m long and consists of a vertical, double-jacketed, heated duct fitted with an inlet and an outlet for the drying gas. The drying gas (air or an inert gas) is heated to a temperature corresponding to the boiling point of the solvent; gas flow is either cocurrent or countercurrent with respect to the spun filaments. To avoid fluctuations in quality of the filament yarn, all spinning parameters must be kept constant within a narrow range. Optimal recovery of solvent is an extremely important economic factor in fiber production. The filaments emerging from the dry-spinning shaft are rerouted by take-off rolls, coated with finish, and wound at speeds of up to 17 m/s. Further processing of the spun material is analogous to that described in Section 3.1.5. As a rule, 12 to 64 dry-spinning shafts are arranged side by side in series and constitute a unit known as a spinning machine.

3.2.2. Wet Spinning

The principles of wet spinning are described in Section 2.1.4.2. Preparation and purification of the highly concentrated spinning solution are similar to dry spinning. The viscosity and therefore the concentration of the polymer are selected to suit the requirements of the spinning units and to give a pressure at the spin pump of ca. 1 – 2 MPa. The spinning solution is filtered and conveyed to the spinnerets; a number of spinnerets generally spin into a common coagulating bath. The coagulation conditions are an extremely important factor in determining the solidification of the swollen, gel-like filaments and therefore the filament quality. Immense know-how is required to maintain a constant concentration and temperature in the coagulating bath during spinning. The residence time of the yarn in the coagulating bath is affected by a number of factors, such as speed of injection and drawoff, and must be optimized to ensure uniform, flawless coagulation of the yarn.

In older units, yarn from the coagulating bath is subjected to drawing and posttreatment in separate processes, but modern integrated units operate continuously in a one-step mode, as shown in Figure 50. Filament yarns from many neighboring spinnerets are drawn from the coagulating bath and stretched by warp drawing. They are then posttreated and wound at speeds of 2.5 – 15 m/s on cylindrical bobbins on individual winders combined in creels. Posttreatment

such as drying, drawing, and heat-setting differs depending on the polymer being processed and the desired properties of the filament yarn. The drawing unit for warp drawing shown in Figure 50 is designed for the production of high-strength, high-modulus poly(vinyl alcohol) yarn [102].

Gel Spinning. Gel spinning is a type of solution spinning that has been defined as as dry spinning at a very high polymer concentration (up to 80 %), that is, a cross between dry and melt spinning. However, the term is also used because of the gel-like appearance of the resulting filaments [67], [103].

This spinning technology was developed at the end of the 1970s [104] for the production of high-performance polyethylene fibers and filament yarns which are distinguished by high strength and high modulus of elasticity. These special properties are achieved because gel spinning orients the macromolecular polymer chains along the fiber axis with a high degree of parallelism. Polyethylene with an extremely high molecular mass ($\bar{M}_w > 10^6$ g/mol) is used and must be spun from solutions at a relatively low concentration. Under these conditions, the molecular chains uncoil and can orient themselves along the fiber axis before crystallization sets in. The polymer solution (ca. 20 wt %) is prepared in a hydrocarbon solvent such as Decalin (decahydronaphthalene [*91-17-8*]) and then forced through the spinnerets at 120–130 °C and spun in a dry-spinning shaft. Spinning is followed by hot drawing at 70–150 °C.

High-performance polyethylene fibers produced by gel spinning are being introduced commercially under the trade name DYNEEMA [103]. Other polymers, such as polyacrylonitrile, are also assumed to be accessible to gel spinning and the manufacture of high-performance fibers.

3.3. Texturing of Filament Yarns

3.3.1. Development of Texturing Processes

In many applications, the original smooth multifilament yarns must be textured (crimped) in two or three dimensions so that they acquire the optimum properties required for their end use (cf. Section 2.3). Texturing increases the bulk and improves the elasticity of the yarn. Only a few of the many texturing processes developed during the 1950s and 1960s have survived on a commercial scale [2], [105]. For textile titers in the range 22–330 dtex, *false-twist texturing* is now used for more than 90 % of all polyesters and polyamides; twist texturing, air texturing, and bicomponent technology are only of minor importance. During the past ten years, textured polyester filament yarns have been primarily responsible for the growing popularity of *friction false-twist texturing*. Worldwide production of textured polyester yarn is anticipated to be 3.0×10^6 t/a by 1993, which is much more than the total annual worldwide consumption of wool. Further strong increases up to the year 2000 are only expected in Asia. In 1997 about 79 % of 179 176 friction false twist texturing spindles were shipped to Asia [20].

Until the early 1970s, three separate process steps were required to produce textured filament yarn. The yarn was melt-spun conventionally at a draw-off speed of 13–25 m/s; it was then drawn on draw twisters or draw winders and, finally, textured on special machines. Since 1972, however, the combined one-step draw-texturing process has become increasingly important worldwide. Compared to the conventional three-step process, this technology results not only in remarkable improvements in the quality of the textured yarn but also in substantial financial savings. The pneumatic stuffer box has increasingly replaced the mechanical stuffer box in the production of textured yarns with a fineness of 1000–≥4000 dtex; these so-called bulked continuous-filament (BCF) yarns are used for carpets and for decorative or upholstery fabric.

3.3.2. False-Twist Texturing

In false-twist texturing, thermoplastic multifilaments are twisted by torsion, the twist is heat-set, and the yarns are subsequently detwisted in a continuous process. Textured yarns are obtained which exhibit good bulk but are almost completely free of twist. Also since 1972, false-twist texturing on spindles has largely been replaced by texturing in which twist is imparted by friction. Friction texturing allows considerably greater productivity because drawoff speeds of

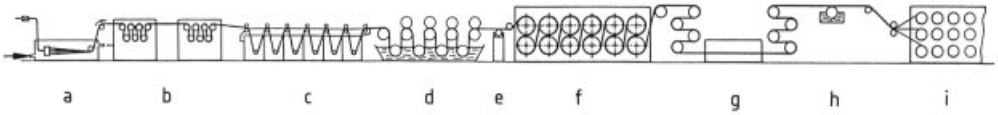

Figure 50. Schematic of a wet-spinning – warp-drawing machine
a) Coagulating bath; b) Wet – cold drawing; c) Neutralization bath; d) Washing bath; e) Lubrication; f) Drying; g) Hot drawing; h) Final finishing; i) Winding of filament yarn

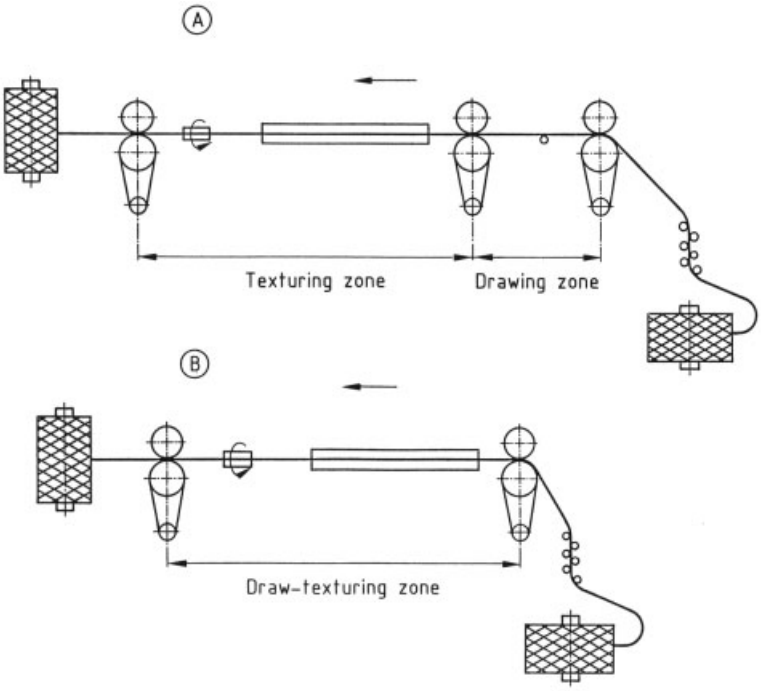

Figure 51. Draw-texturing process
A) Sequential draw texturing; B) Simultaneous draw texturing

10 – 20 m/s can be obtained; texturing on spindles only allowed a maximum draw-off speed of 5 m/s. The rapid development of friction texturing would not have been possible without the commercial introduction of high-speed spinning and draw texturing [108].

Draw texturing may be either sequential or simultaneous (see Fig. 51). In sequential draw texturing, the yarn is drawn in a drawing zone which precedes the texturing zone; in simultaneous draw texturing, drawing and texturing are integrated in a single zone.

High-speed spinning of partially oriented yarn (POY) in combination with simultaneous draw texturing has made its mark worldwide in the manufacture of polyester yarns. Depending on whether the crimp of textured yarns is reduced by thermal treatment, the yarns are divided into non-heat-set, highly elastic (HE) yarns with a crimp of 35 – 50 % and heat-set yarns with a 20 – 25 % crimp. Since worldwide capacity for conventionally spun polyamide yarns with a low orientation (LOY) is still existing, these polyamide yarns are usually crimped by sequential draw texturing. New installations for the manufacture of polyamide yarn employ high-speed spinning so that the resultant POY is also subjected to simultaneous draw texturing.

Commercial frictional draw-texturing machines are based on a similar design. They all contain the following essential elements: a supply creel, heater, cooler, friction-texturing unit, and winding unit. A typical draw-texturing machine is depicted in Figure 52.

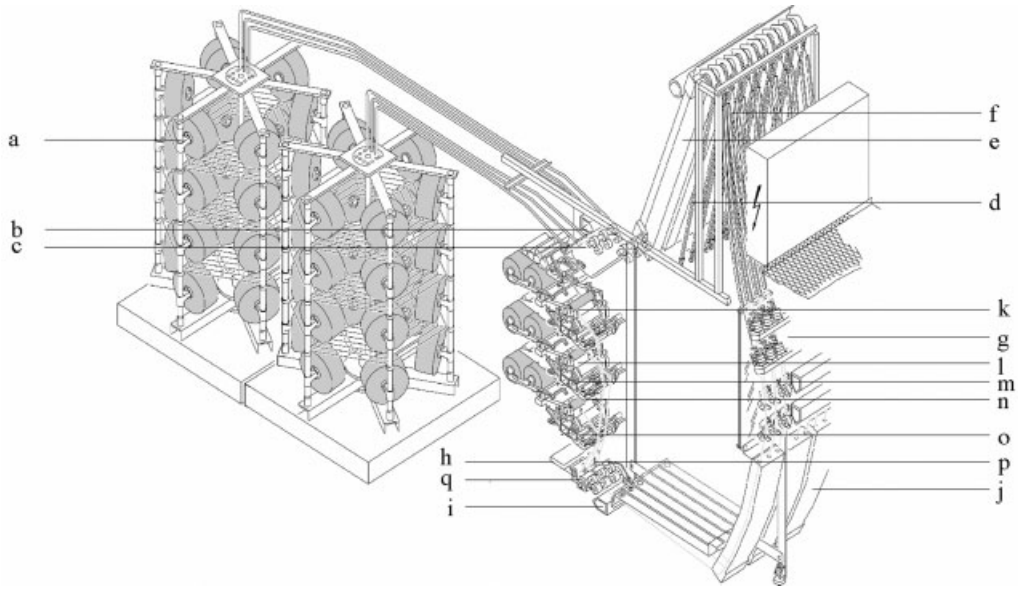

Figure 52. Double-heater draw-texturing machine
a) Rotary creel; b) Yarn cutter; c) Feed system; d) Threading device; e) Primary heating; f) Cooling plate.; g) Ringtex texturing unit; h) Yarn break sensor; i) Yarn aspirator; j) Secondary heating; k) Take-up unit; l) Tube holder; m) Reversing shaft; n) Drive rollers; o) Traversing yarn guide; p) Yarn guide elements; q) Yarn oiler
Courtesy of Barmag/Saurer Group, Remscheid, Germany

Yarn runs from the POY bobbins mounted on the supply creel (a) to feed rollers (c), over the primary heater (e) and cooling plate (f) to the frictional texturing unit (g) and a second set of feed rollers (c). In the case of heat-set yarn, the drawn and textured filaments pass over a secondary heater (j) before being wound in the take-up unit (k). To attain favorable properties for further processing, the cohesion of the textured yarn must be improved. For economic reasons, textured yarns are interlaced rather than twisted immediately after draw texturing.

Table 3. Advantages and disadvantages of disk materials

	Ceramic disks	Polyurethane disks
Tenacity of textured yarn	lower	higher
Sensitivity to spin finish	lower	higher
Snow generation	higher	lower
Crimp	lower	higher
Lifetime	long	shorter
Costs	lower	higher

Disk-friction units with three axes, (see Fig. 53) are commonly used to impart twist, the disks can be made of pure ceramic or polyurethane, as well as steel or aluminium, coated with different materials. These different disk materials have different advantages and disadvantages which are summerized in Table 3.

A draw-texturing machine is shown in Figure 54.

The number of disks on a draw texturing unit depends on production speed, yarn thickness, disk surface and spin-finish content. Usually this number varries from four to eight. Sometimes also polished disks are used at the top and the end of a disk arrangement.

Two alternative twisters based on the principle of nip-controlled vector (NCV) drive have been developed by Murata Machinery (Japan) and Barmag (Germany). In these machines, twist is imposed on the yarn by two elastic surfaces running in opposite directions. Only Murata Machinery has commercialized the texturizing unit with crossed belts and one large disk with two belts (one on each side of the disk).

The Murata system uses small intersecting elastic belts (Fig. 55), and the Barmag Ringtex system textures the yarn between two intersecting, ring-shaped friction surfaces (Fig. 56). The NVC false-twist systems are said to per-

mit higher texturing speed, reduce abrasion loss, improve yarn strength, and be generally gentler. However, because of its better process stability the frictional draw-texturing process has more widely established in the textile industry.

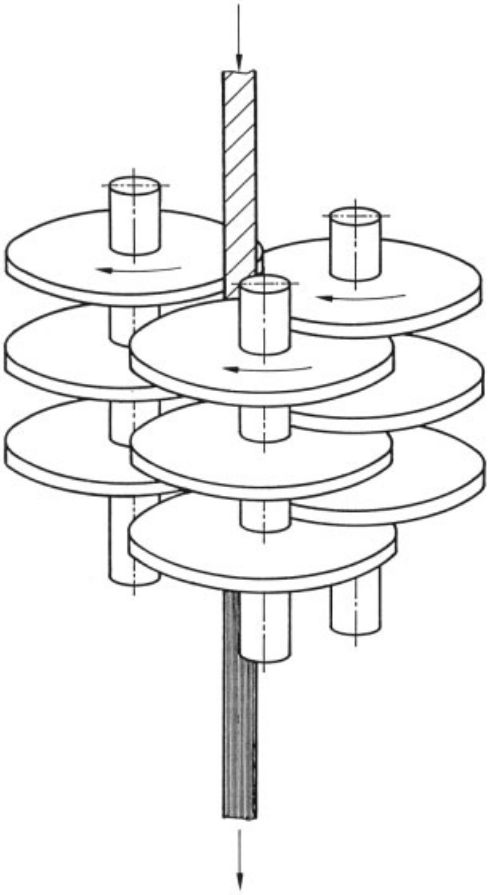

Figure 53. Disk-friction false twister with three axes

Typical lengths of heating zones are about 2 – 2.5 m each. To minimize the machine height and to improve the operation of the machine, different yarn paths have been realized. Figure 52 shows a so-called collapsed or folded yarn running path. However, rerouting the yarn can be detrimental to yarn quality, especially in the case of profiled or very fine filaments. To improve the texturing performance, different machine designs are available, as shown in Figure 57.

Attempts are being made to obtain even higher texturing speeds for economic reasons. Therefore, noncontact high-temperature heaters were developed. Such heaters operated at about 400 °C as opposed to the temperature of standard contact heater of about 200 °C. The high-temperature noncontact heaters are shorter and so the yarn tension caused by friction is much lower, especially at high speeds. With such heaters special care must be taken with the spin finish, since small deposits of heat-cracked spin finish in the heater can limit the operation drastically.

To guarantee a uniform and failure-free quality of the textured yarn, new machines are equipped with an online quality inspection system. Such systems monitor the yarn tension of each yarn path and classify each bobbin into different levels of quality. In combination with an automatic doffing system, such systems reduce costs by decreasing labor requirements and allowing night shifts without personnel.

Figure 54. Draw-texturing machine (Mach Crimper 33H) Courtesy of Murata Machinery, Japan

Attempts are also being made to produce POY-bobbins with better runoff properties. Barmag has, for example, introduced the ribbon-free random wind (RFR) system [109].

3.3.3. Air-Jet Texturing

Widespread use of air-jet texturing was hindered by its high consumption of air and the relatively low production speed (ca. 2 m/s). In the 1970s, however, production speed was successfully increased (7 – 10 m/s) and air consumption reduced. These developments, together with the trend to yarn with a character more like that of staple fibers, have led to increased use of air-jet texturing.

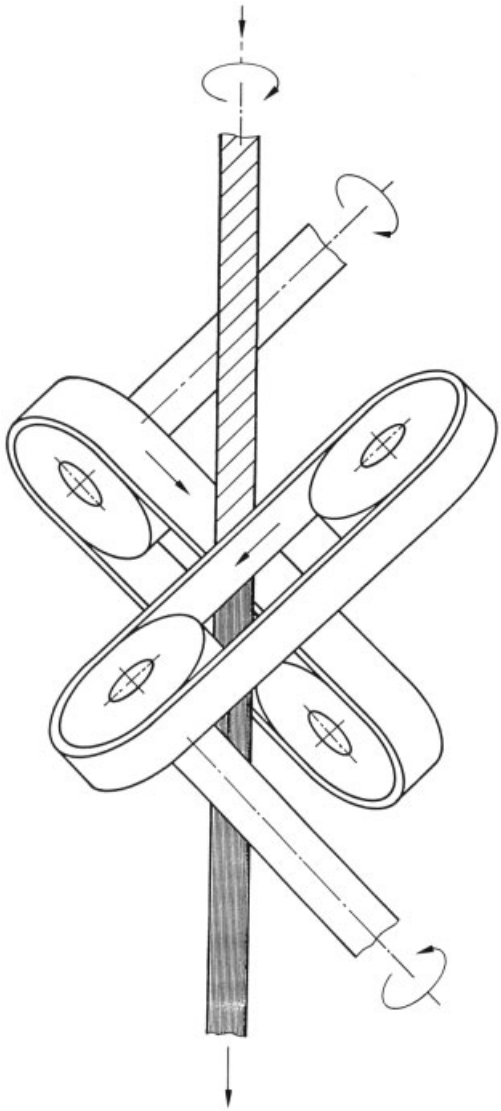

Figure 55. Double-belt false twister

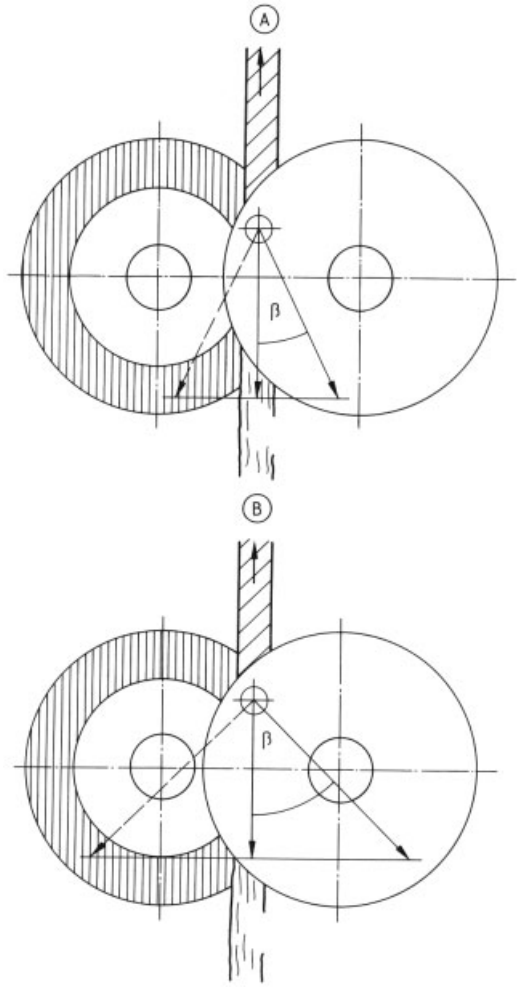

Figure 56. Disk-friction false twister (Ringtex system)
A) Small twist angle, low degree of twisting; B) Large twist angle, high degree of twisting

In air-jet texturing, one or more multifilament yarns are fed into the texturing jet where they are swirled in a turbulent stream of air and form loops and curls. This process is illustrated schematically in Figure 58. The basic mechanism of air-jet texturing involves mechanical deformation of individual filaments [110]; it can therefore be used to texture both thermoplastic and nonthermoplastic filaments.

The construction of air-jet texturing machines is similar to that of false-twist texturing machines (see Section 3.3.2), but an air-texturing jet replaces the false twisting unit. Since air-jet texturing machines can be fed with both drawn and undrawn (POY or LOY) filament yarns, they are equipped with one or two drawing zones that precede the texturing zone [111–113].

The crimp character of an air jet textured yarn is much different to that of a friction false twisted textured yarn. The two types of textured yarn are

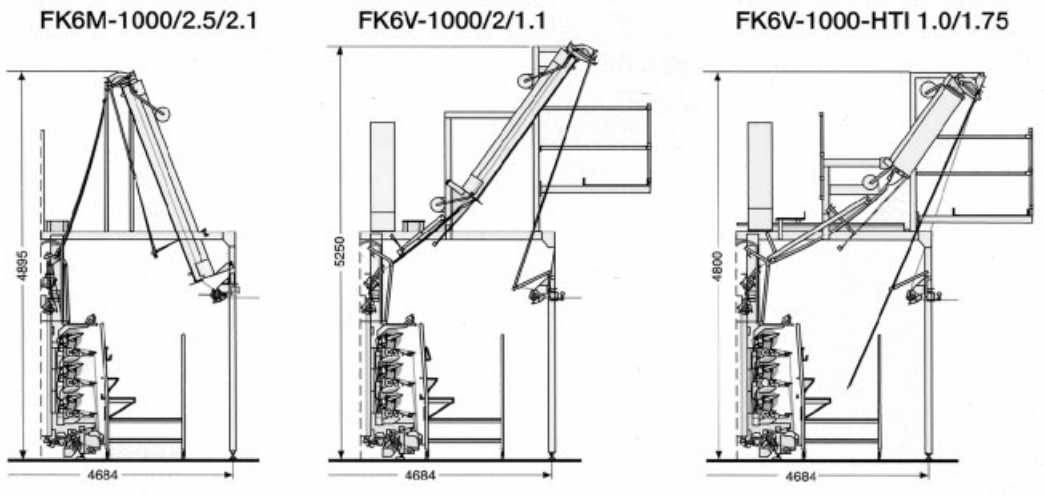

Figure 57. Different designs of friction false twisting machines
Courtesy of Barmag/Saurer Group, Germany

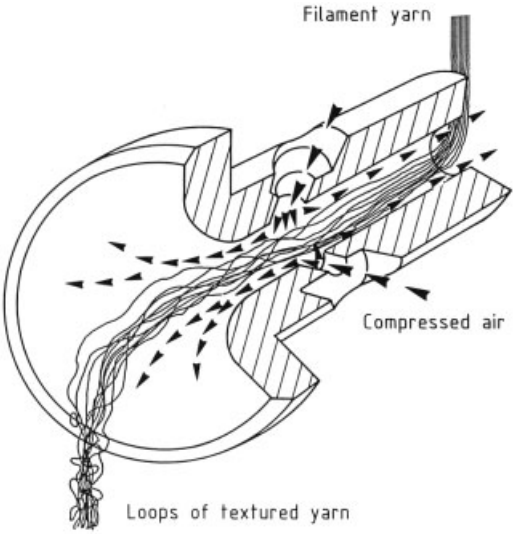

Figure 58. Schematic representation of loop formation in an air jet
Courtesy of Heberlein, Wattwil, Switzerland

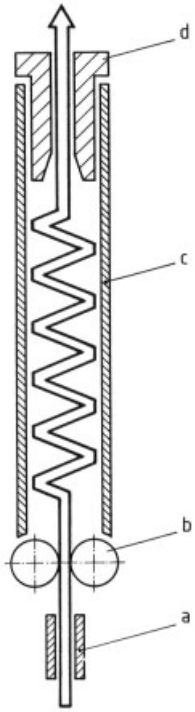

Figure 59. Mechanical stuffer-box crimping
a) Incoming yarn; b) Feed rollers; c) Heated stuffer box; d) Closing device with changeable weight

used in different fields of application in the textile industry.

3.3.4. Stuffer Box Texturing

In the mechanical stuffer box process (shown schematically in Fig. 59), texturing is achieved by using a set of rollers to feed the filament yarn into a heated, pressure-controlled stuffer box. The yarn undergoes buckling deformation, and a sawtooth, two-dimensional crimp is formed [114]. Stuffer-box crimped yarns up to 330 dtex are of minor commercial importance for textile applications. The pneumatic stuffer box has almost completely replaced the mechanical stuffer box for texturing coarser yarns with a fineness of ca. 500 – 3300 dtex, which are primarily made of polyamides or polypropylene. The major reasons for this are probably the substantially higher texturing speed of up to 67 m/s and the three-dimensional crimp which is far superior in bulk and volume to the sawtooth, two dimensional crimp. The basic principle of the pneumatic stuffer box was described in a patent application in 1957 [115]. The system is shown schematically in Figure 60.

Two different process variants are available for the production of textured BCF yarn with a fineness of 1000 – $\geq$ 4000 dtex for the carpet and upholstery industries: discontinuous draw texturing and continuous spin-draw texturing [116–118].

Discontinuous draw texturing consists of two steps. The undrawn filaments are first spun on conventional spinning units and wound to give cylindrical bobbins with weights of ca. 20 – 50 kg. In a second step, the spun bobbins are then placed in supply creels, and the individual filament yarns are conveyed to the individual positions of the draw-texturing machine via yarn guides and brakes. A typical draw-texturing machine is shown in Figure 61.

The draw unit of the machine consists of a system of heated godets formed either by pairs of rollers or by feed rollers with idler rolls. After being drawn, the filament yarn is textured in a pneumatic stuffer box, deposited on cooling drums, and subsequently wound on cylindrical bobbins. As a rule, the textured yarn is also interlaced prior to winding to improve its processing characteristics. One to four yarn ends can be processed simultaneously at each machine position, i.e., in each draw-texturing unit. One- and two-end operations are carried out commercially at up to 67 m/s, but the speed must be reduced to 42 – 50 m/s in four-end operations.

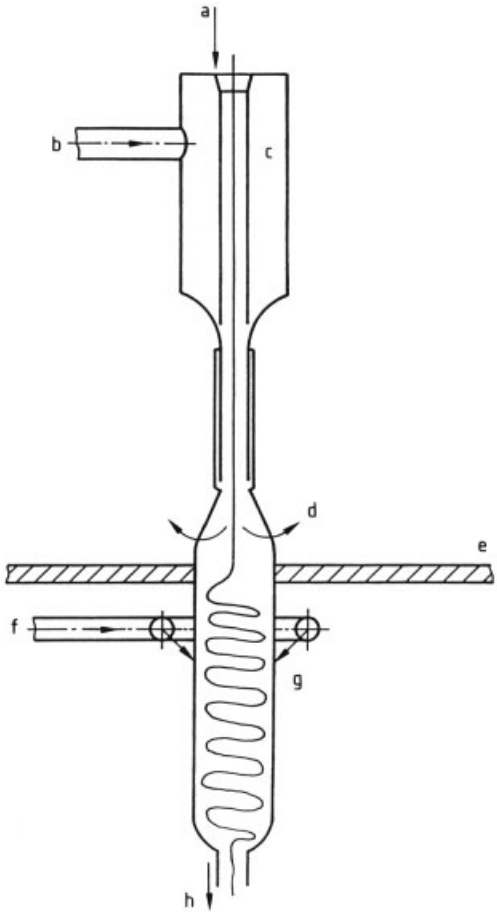

Figure 60. Pneumatic stuffer box crimping [115]
a) Yarn; b) Heated air; c) Injection tube; d) Stuffer box; e) Partition for separating air flow; f) Cold gas; g) Setting zone with air-permeable walls; h) Crimped yarn

The discontinuous, two-step process is particularly suited to the manufacture of speciality BCF yarns. Various spun-dyed filament yarns of different fineness or different dyeing properties can be plied and processed together in a single texturing chamber to give special BCF yarns.

The high texturing speed attainable with the pneumatic stuffer box system allows the spinning of the filament yarn and draw texturing to be combined in a highly economical, single-step, *continuous spin-draw texturing process.* A spin-

draw texturing machine is shown schematically in Figure 62; the draw-texturing unit is located directly below the conditioner of the spinning machine. The spun filament yarns are drawn immediately, textured, interlaced, and usually wound at a rate of 50 – 67 m/s by automatic winders.

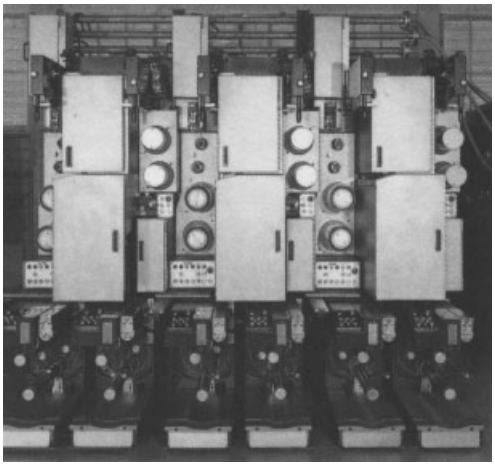

Figure 61. Draw-texturing machine with four ends for bulked continuous-filament (BCF) yarn
Courtesy of Rieter, Winterthur, Switzerland

3.3.5. Bistructural Texturing

In bistructural texturing, also called spin texturing, the latent crimp present in the spun yarn is activated by a process known as development [119], [120] (see also Section 2.3.2.1). Texture can be developed by means of thermal or hydrothermal treatment either during or after separate drawing [121]. Alternatively, texture development can occur during dyeing or heat-setting after the smooth yarn has been processed into textiles [122]. Spinning, drawing, texture development, and winding can also be integrated into a one-step process [123]. In Section 2.3.2.1 a number of processes are described which can be used to produce anisotropic filament structures that lead to spin-textured yarns. These processes are all based on the inhomogeneity of the yarn material in the cross-sectional plane of the individual filaments. Differences in the shrinkage of the individual filaments, caused by chemical or physical inhomogenity, are primarily responsible for the formation of three-dimensional crimp. This method of texturing allows the two-step process technology of high-speed spinning and subsequent draw texturing (described in Section 3.3.2) to be developed into a totally integrated, one-step process.

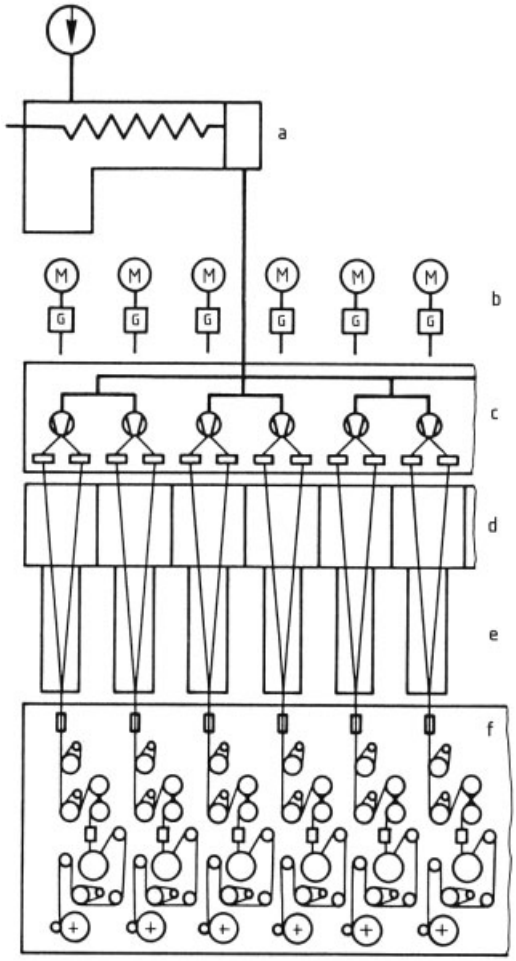

Figure 62. One-step, spin-draw texturing machine for nylon 6, bulked continuous-filament (BCF) yarn
a) Extruder; b) Motor and gears ; c) Spin beam; d) Quenching system; e) Conditioner; f) Drawing, texturing, and winding systems

However, spin texturing is not widely used in industry (contrary to earlier expectations) [119], [124]. The reason for this could be that bicomponent technology has been predominantly used to produce bistructurally textured multifilament yarns. This technique involves the simultaneous spinning of two polymers and is much more demanding and also more expen-

sive than spinning a single polymer. In addition, bistructurally textured multifilament yarns differ in hand (touch) and appearance (crimp) from the friction-textured yarns that currently dominate the market and have therefore had difficulty in being accepted. Especially for polyesters (which represent about 80 – 90 % of world textured yarns) the poor crimp stability of bistructural textured yarns has led only to a few special textile applications.

4. Manufacture of Tow and Staple Fibers

4.1. Fiber Spinning

Spinning processes used in the production of tow and staple fibers can be divided into three major groups which are listed below in order of their commercial importance:

1) Melt spinning (conventional, compact, and continuous one-step processes)
2) Wet spinning
3) Dry spinning

The spun filaments are combined to form a tow, which is drawn, washed, heat-set, dried, crimped, and packaged. Staple fibers are obtained by cutting the tow into defined lengths.

4.1.1. Melt Spinning

4.1.1.1. Conventional Process

The basic principles of melt spinning are described in Section 2.1.3. Polyesters, polyamides, and polyolefins are currently the most important melt-spinnable polymers.

The molten polymer can be pumped directly from a polycondensation unit to the spinning unit (direct spinning). Alternatively, the polymer granulate is dried and/or conditioned, melted in a suitable screw extruder, and conveyed to the spinning unit [125]. Since a spinning unit generally consists of several identical spinning positions, the polymer is transported to the spin beam along paths of equal length by means of a distributor. The spin beam (Fig. 63) is connected to the spin pump (Fig. 64) and also accepts the spin packs (Fig. 65) by either top or bottom loading.

To achieve a uniform temperature throughout the entire spinning system, it should be heated with condensing diphyl (a eutectic mixture of diphenyl and diphenyl oxide, boiling range 256 – 258 °C at standard pressure) as heat-transfer medium.

The spin pump conveys a precisely measured volume of melt to the spin pack where it is distributed and filtered. Pressures may reach 20 MPa. The melt stream finally reaches the spinneret plate and is extruded through a multiplicity of holes [126] (Fig. 66). The hole capillary may display various profiles which shape the emerging melt stream and the resulting filament (Fig. 67).

After emerging from the capillary, the melt stream is cooled by heat exchange with the surroundings and solidifies to form a filament. Conditioned air is usually used as a cooling medium and is blown through the bundle of filaments. The air impinges on the filaments with a laminar flow at a defined velocity [127]. If circular or rectangular spinnerets are used, the necessary amount of air is blown across the filament bundle by means of a quenching shaft. In the case of annularly arranged holes or annular spinnerets, central quenching is used, in which air is blown either outward or inward from a tubular distributor and thus flows radially past the filaments (Fig. 68). Central quenching systems can handle a much larger number of holes than air shafts and, hence, increase polymer productivity per spinneret. They are thus more economical and, in addition, produce yarn of excellent quality [17]. Spinneret plates with a hole density of ca. 25 holes per square centimeter and a total of ca. 7000 holes can be used. Throughputs per spinning position of 240 kg/h and machine capacities up to 180 t/d can thus be attained. In contrast, the maximum possible number of holes in rectangular spinnerets is ca. 3000, and in circular spinnerets with cross-flow quenching systems ca. 700.

The extruded, solidifying filaments are drawn off by an external force and, depending on the polymer and the fineness, attain their final shape in a zone 3 – 200 cm below the spinneret. The bundle of filaments is then gathered together, redirected, and combined with bundles from other spinning positions to form a tow.

Spin finish is applied to the filaments prior to gathering by means of either a static system

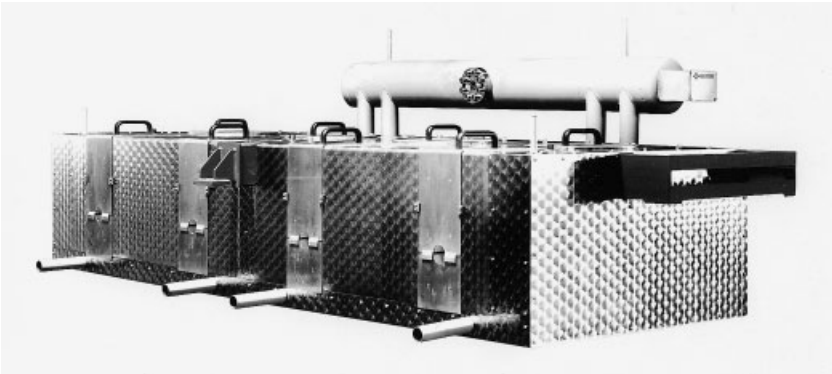

Figure 63. Spin beam for fiber production
Courtesy of Inventa-Fischer AG, Domat/EMS, Switzerland,

(e.g., slot finishing) or a dynamic system (e.g., roller finishing). The spin finish normally consists of a mixture containing demineralized water and 0.1 – 5 % of a lubricant. In both cases the spin finish is applied by the finish applicator to obtain good tow cohesion, satisfactory electrostatic behavior, and the desired degree of fiber-to-fiber and/or fiber-to-metal friction. The tow then reaches the take-off unit which usually comprises four to six electrically driven godets (Fig. 69). The tow, which may be as large as 20 – 100 ktex ($\triangleq 2 \times 10^5 - 1 \times 10^6$ dtex) and moves at a speed of 8 – 33 m/s, finally reaches the delivery unit (also shown in Fig. 69). Here, the tow is usually picked up by so-called sunflower wheels, slightly undulated, and deposited in a container known as a spin can. The spin cans traverse below the delivery unit to ensure a uniform tow distribution (Fig. 70. They normally hold 500 – 8000 kg of tow.

Attempts have been made to replace this system by one in which the relatively high transverse tow velocity is reduced by forcing a rotary motion on it [129], [130]. However, these attempts failed due to economic reasons or damage to the tow.

Melt-spinning plants similar to the one described above usually occupy several floors (Fig. 71).

4.1.1.2. Compact Process

A compact, one-step process has been developed for the production of polyolefin fibers in which spinning is followed immediately by drawing [131]. In some cases this technology can also be applied to polyamides. The reason for this development is that polyolefins are best dyed with pigments in a melt. Since a broad range of colors is available, relatively small units are required with rapid changeover facilities.

Compact units have been designed which consist of a melt extruder, mixing devices for colored pigments, and a spin beam with 4 – 16 spinning positions. Spinnerets may have up to 80 000 holes. Extruded filaments are cooled by intense air blowing immediately below the spinneret. Throughputs of ca. 100 kg/h per position are attained at take-off speeds of 0.2 – 0.5 m/s. After being finished, the filament bundles are gathered, redirected, and fed immediately into a drawing zone. The speed of drawing is ca. 2.5 m/s. After drawing the tow is annealed, crimped, and, if necessary, heat-set. It is then cut into defined lengths and packed in a baling press. A 16-position machine operating in a fineness range of 1 – 3 dtex produces tow at a rate of ca. 15 t/d, the rate is higher for coarser fibers. Such a machine has a maximum height of 5.5 m and occupies a floor area of 35 × 6 m (Fig 72).

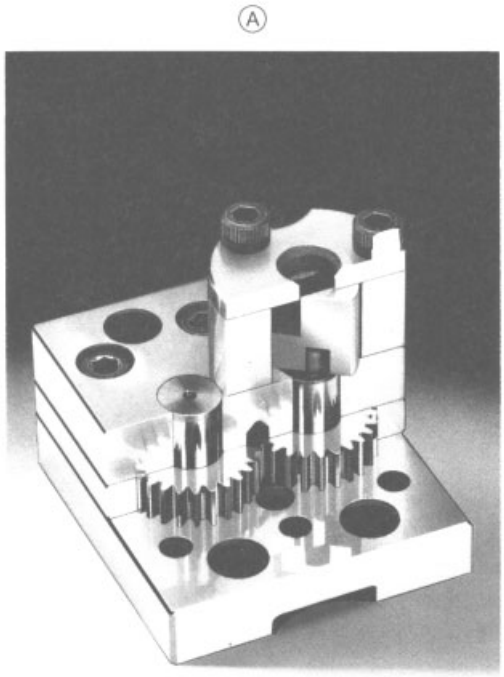

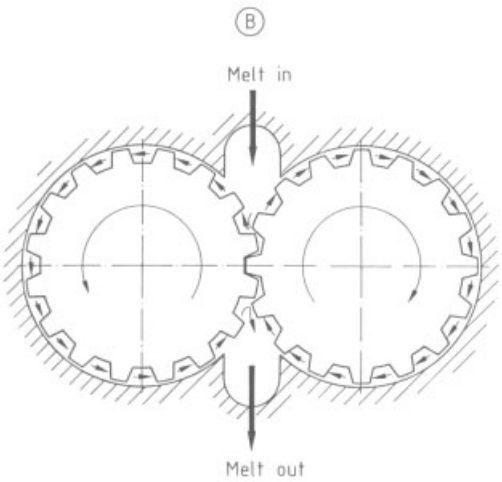

Figure 64. Gear spin pump
A) Cut-away model (courtesy of Feinpruf, Göttingen, Germany); B) Mechanism of operation

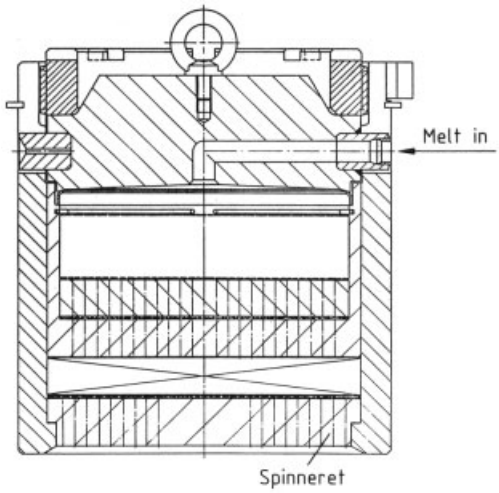

Figure 65. Spinneret pack

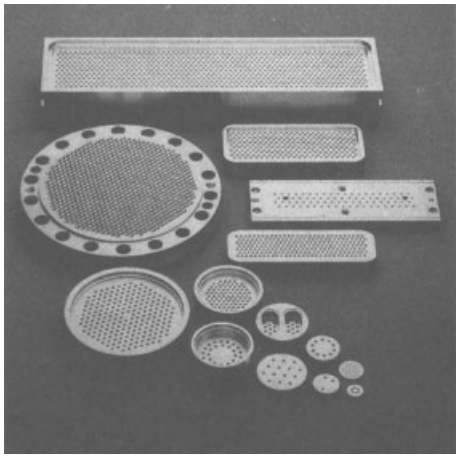

Figure 66. Spinnerets made of high-grade steel used for melt spinning of staple fibers
Courtesy of Enka tecnica, Heinsberg, Germany

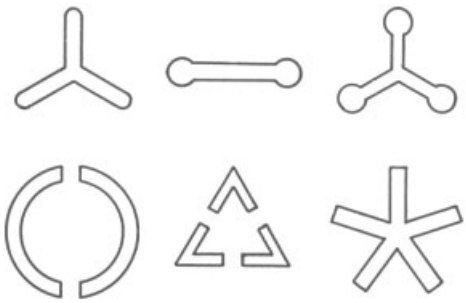

Figure 67. Cross sections of spinneret capillaries

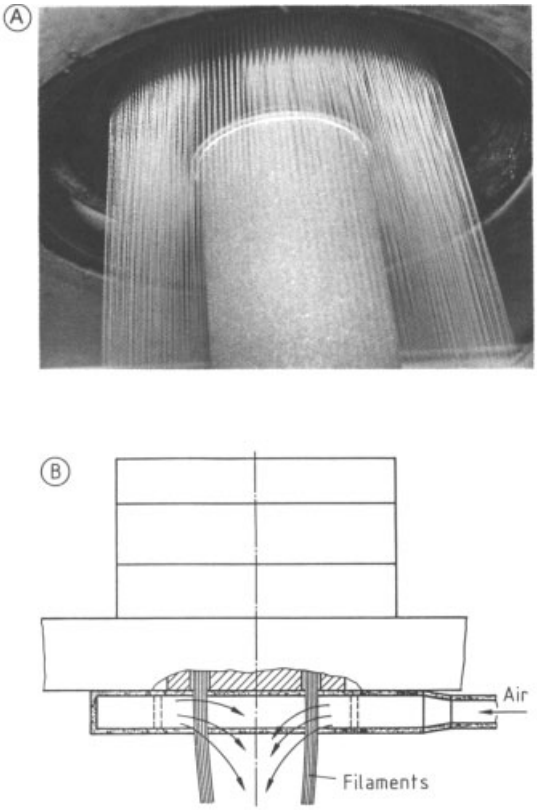

Figure 68. Central quenching of spun filaments
A) Outward quenching (courtesy of Inventa-Fischer AG, Domat/Ems, Switzerland); B) Schematic of inward quenching [128]

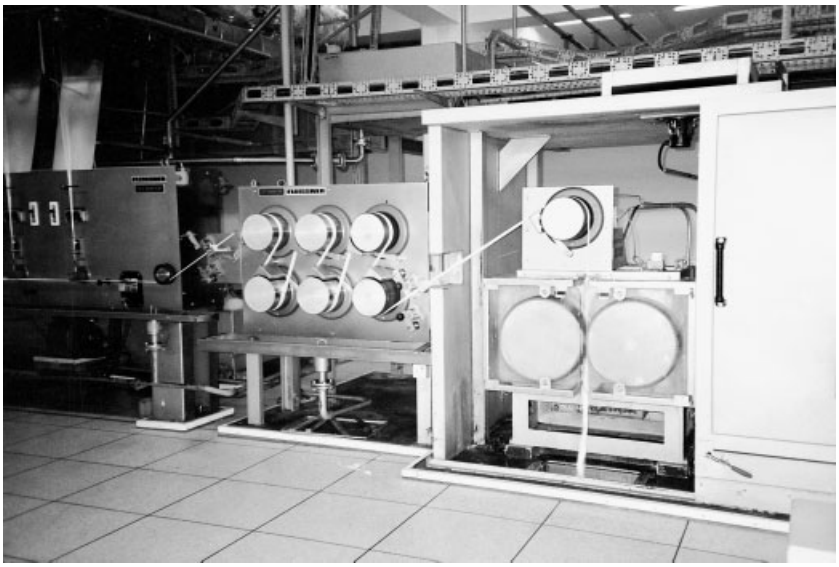

Figure 69. Spinning machine with take-off godets and delivery unit
Courtesy of Fleissner, Egelsbach, Germany

Figure 70. Traversing spin can
Courtesy of Fleissner, Egelsbach, Germany

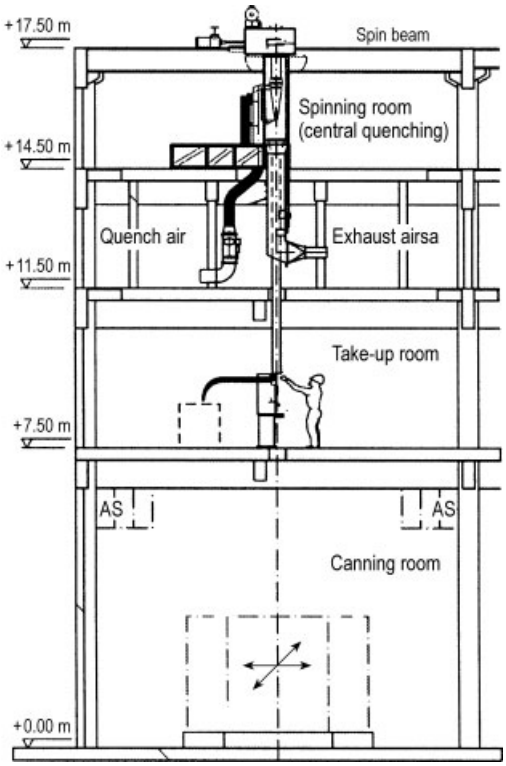

Figure 71. Conventional spinning plant for polyester staple fiber production

4.1.1.3. Continuous One-Step Process

The continuous one-step process [132], [133] was developed as a result of the lower speeds (< 16 m/s) used in conventional spinning and integrates drawing, annealing, crimping, heat-setting, and cutting. This process is primarily used for manufacturing polyolefin fibers but in special cases can also be applied to other melt-spinnable polymers such as polyester and polyamide.

Rectangular spinnerets are used for spinning, and four spinning positions arecombined into a draw-texturing unit. The delivery speed of the drawing and stabilizing step is 33 – 42 m/s at a fineness of 30 000 – 40 000 dtex. The multifilament tow is then crimped by a hot air-jet texturing system (cf. Section 3.3.3) and subsequently deposited on a heat-setting conveyor belt. A specially developed cutter, based on the principle of rapid cutting (see Section 4.2.5) divides the textured filament tow into staple fibers of the desired length.

An installation of this type has a height of ca. 5.4 m and is ca. 4 m wide and 18 – 30 m long, depending on the length of the heat setting oven; it can produce staple fibers at a rate of 8 – 10 t/d (Fig 73).

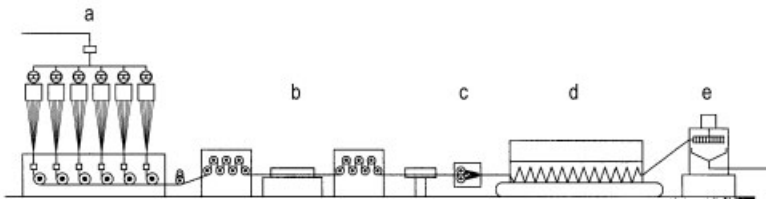

Figure 72. Flow diagram of a compact one-step staple fiber spinning process
a) Spinning; b) Drawing; c) Crimping; d) Heat setting e) Cutting

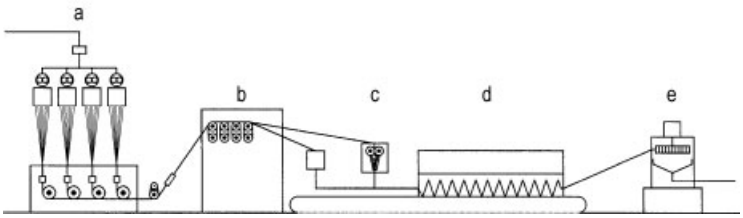

Figure 73. Flow diagram of a high-speed one-step process for continuous staple fiber production
a) Spinning; b) Drawing; c) Crimping; d) Heat setting; e) Cutting

4.1.2. Wet Spinning

The basic principles of wet spinning are described in Section 2.1.4.2, and the production of continuous filaments by this method is decribed in Section 3.2.2. If staple fibers are desired, the as-spun filaments produced by compact or combination spinnerets (Figs. 74 and 75) are gathered into a tow in, or on leaving, the spin bath. A typical spinning machine is shown in Figure 76. Spinning positions may be arranged on one or both sides. Spun filaments are drawn off at relatively low speed, but this is offset by using a very large number of spinneret holes.

In the continuous processes predominantly used today, the tow is guided directly to a drawing unit for further processing (cf. Section 4.2). Wet-spinning lines produce up to 90 t of polyacrylonitrile (PAC) tow per day (total fineness = 770 dtex) at a rate of 1.3 m/s. However, PAC fibers are also produced by the more economical melt-spinning process [134].

4.1.3. Dry Spinning

The basic principles of dry spinning are described in Section 2.1.4.1 and the production of continuous filaments by this method is described in Section 3.2.1. In the production of staple fibers, the tow leaving the spinning shaft is first guided over a finish applicator, then redirected and combined with the tow from other spinning positions. It is then taken off by driven rollers and deposited in spin cans. A dry-spinning unit for the production of staple fibers is illustrated schematically in Figure 77. Production rates range from 3 – 8 m/s.

4.2. Posttreatment of Tow and Staple Fibers

The properties of spun fibers do not meet the requirements for their final application. Strength is low, elongation and shrinkage are too high, and further processibility and blendability with natural fibers are not assured. As described for the one-step processes (Sections 4.1.1.2 and 4.1.1.3) and the wet-spinning process (Section 4.1.2),

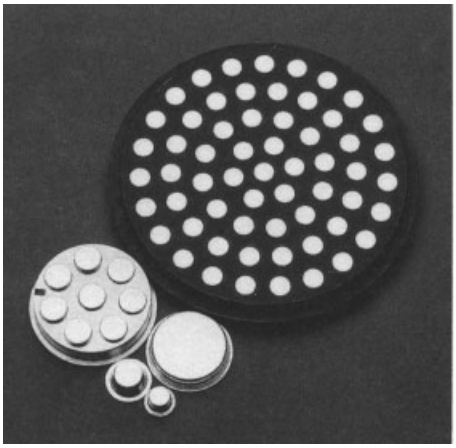

Figure 74. Spinnerets used in the wet spinning of staple fibers
Courtesy of Enka tecnica, Heinsberg, Germany

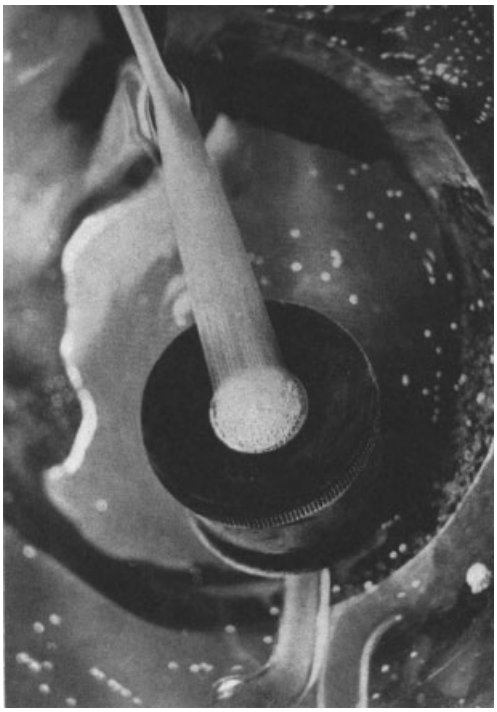

Figure 75. Filaments being spun through a spinneret in the production of staple fibers by wet spinning
Courtesy of Enka, Wuppertal, Germany

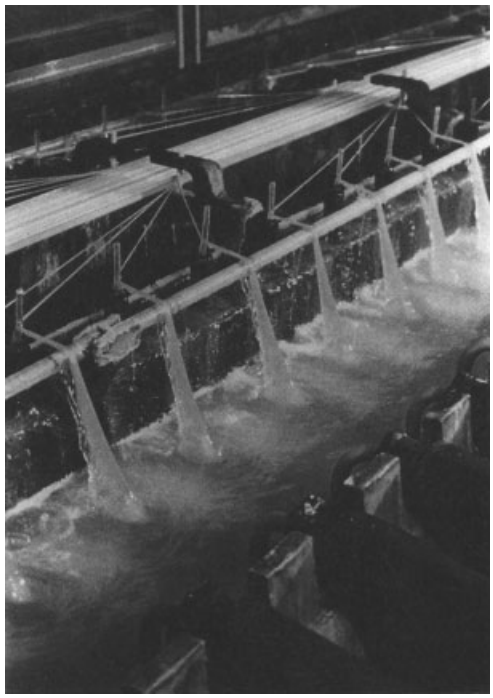

Figure 76. Wet-spinning machine for the production of staple fibers
Courtesy of Enka, Wuppertal, Germany

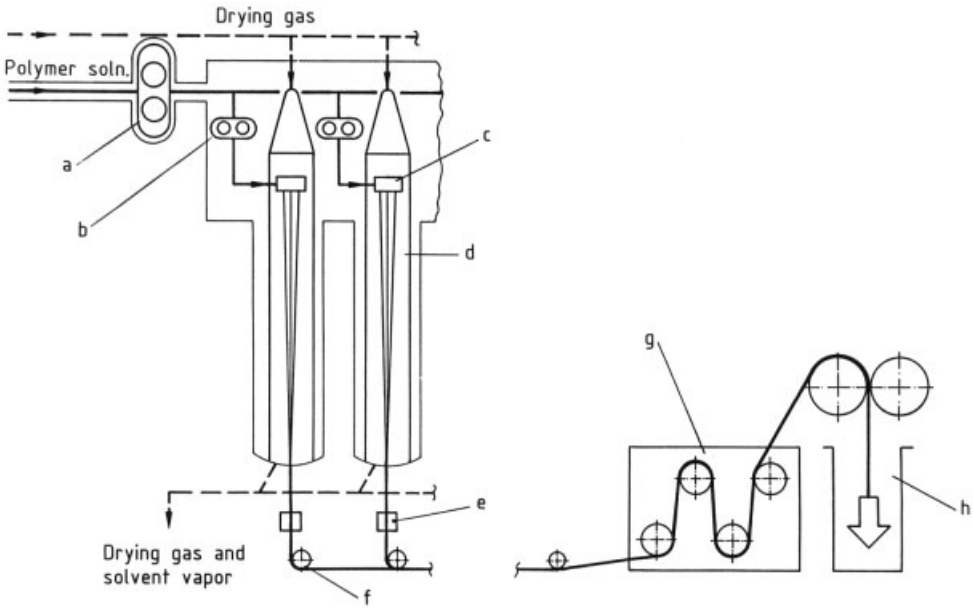

Figure 77. Dry-spinning unit for production of staple fibers
a) Pump; b) Spin pump; c) Spinneret; d) Spin shaft; e) Finish applicator; f) Redirecting godet, g) Take-off unit; h) Spin can

staple fibers must be converted to useful fibers by a suitable sequence of processing steps which include drawing, washing, heat-setting, drying, crimping, cutting, and packaging. These processes are carried out on fiber band production lines which can handle a large number of tows and may be up to 150 m long. The multiple tows are fed into the line and are processed simultaneously. Four widely used examples of such lines are illustrated in Figure 78. Lines produce 200 t of staple fibers per day and have a tractive power of 18 t, a production speed of 5 m/s and can handle a total drawn tow of 6×10^6 dtex [135].

4.2.1. Drawing of Tow

In the case of fibers produced by melt or dry spinning, the spin cans containing the as-spun fiber tow are placed in a can creel. As shown in Figure 79, the tow from these cans is plied and conveyed as a band to the drawing unit of the production line.

Prior to being drawn, the fiber band passes through an immersion bath filled with fluid to adjust its frictional properties and obtain a uniform temperature (Fig. 80A). The fluid consists of demineralized water and a lubricant. Drawing is performed between two or three sets of rollers and is referred to as one- or two-step drawing, respectively [136]. Drawing itself is always divided into two fields: the first comprising the range of draw up to the natural draw ratio and the second comprising the region beyond this ratio. The most common types of drawing used commercially are summarized in Figure 80.

Temperature control, the position of the draw point, and the type of draw used depend on the polymer and the production process and are proprietary know-how of the manufacturer. A suitable draw ratio must also be chosen to obtain a desired type of fiber from a specific polymer. Particular care must be taken not to leave any undrawn sections in the filament, which would result in uneven dyeing. Important factors in this respect are the density (decitex per centimeter) of the tow on the rolls, the uniformity of temperature control in the rollers and baths, modification of the frictional behavior of fibers by finish application, the drawing speed, the position of the draw point, and last but not least, maintainenance of constant operating parameters [1], [61], [84]. Today high-capacity polyester fiber lines (100 – 200 t/d) are equiped with a two-step pro-

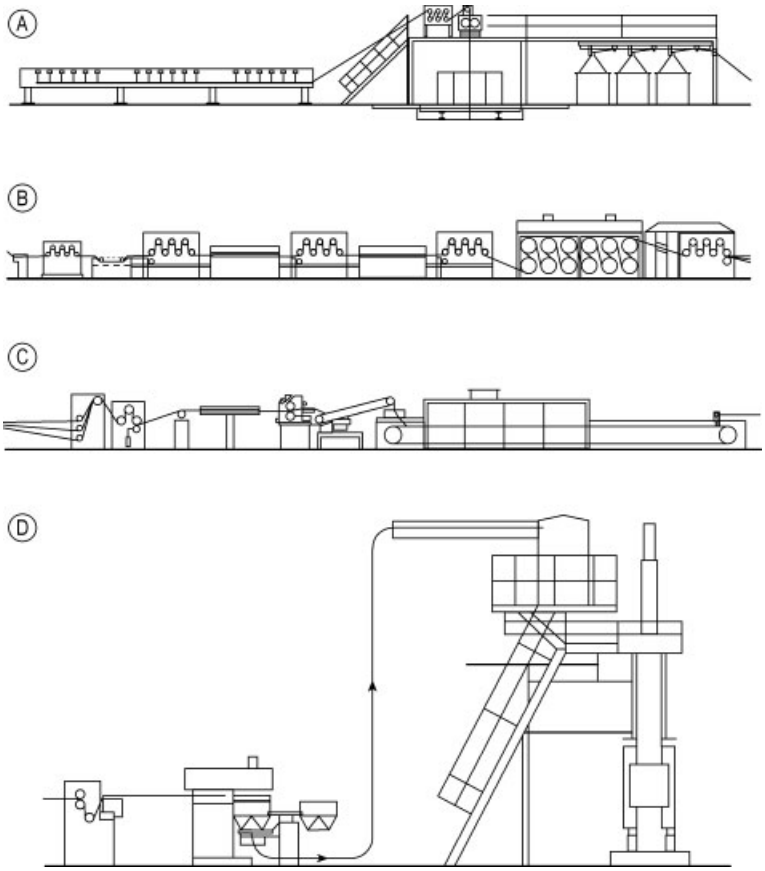

Figure 78. Production line for polyester staple fibers
A) Can creel; B) Drawing and heat setting; C) Crimping and drying; D) Cutting and baling

cess with a immersion draw chest as the first stage.

The draw rollers are water- or vapor-heated and have a diameter up to 45 cm. The draw roller units are cantilevered and up to 1.8 m long, they can process more than 200 t of fiber per day. Draw ratios range from 1 : 2 to 1 : 6 and draw fineness is up to 600 ktex. The draw fibers leave the drawing zone at speeds up to 6 m/s. The drawing units must be able to withstand drawing forces up to 210 kN. A typical drawing unit is shown in Figure 81.

In the case of fibers produced by wet spinning, the multifilaments extruded from a large number of spinnerets are gathered to form a partial tow and washed prior to drawing. Multistep or continuous drawing is usually carried out in hot water which contains a small amount of solvent. The total draw ratio can be as high as 1 : 10.

4.2.2. Washing and Finishing

Washing. Fibers contain various components which must be removed by washing. Wet-spun fibers, for example, contain reaction products, and melt-spun fibers of nylon 6 may contain monomeric caprolactam. Washing can take place at various points along the fiber band production line: during or after drawing, after heat-setting, crimping (tow), or after cutting (flock). The following devices can be used to wash the tow or flock: dip tanks (tow), screen drums (tow/flock), and screen belts (flock).

Washing fluid (water) is circulated in the bath, and its volume, temperature, chemical composition, and concentration are controlled.

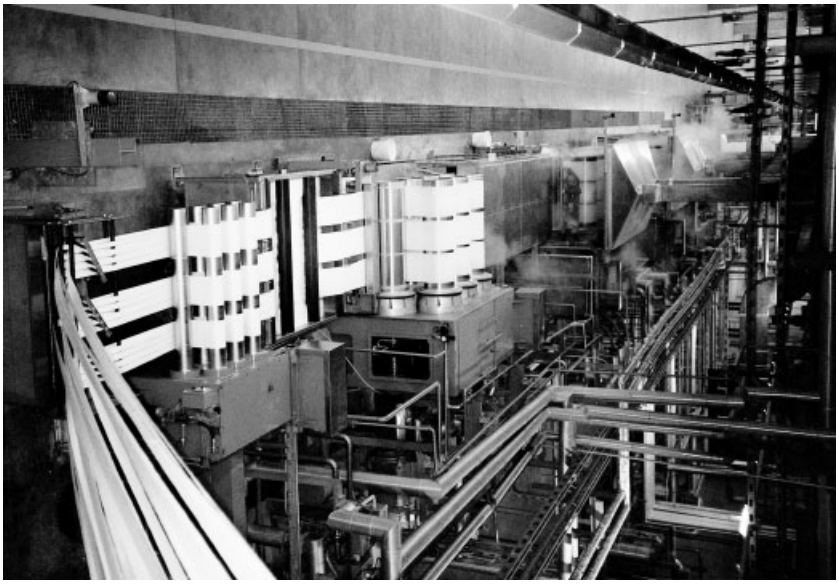

Figure 79. Production line showing the fiber band
Courtesy of Fleissner, Egelsbach, Germany

Finishing. The fiber surface must be provided with an appropriate finish to ensure suitable running properties during production and subsequent processing. In a fiber production line, the finish is applied by immersing the tow in an immersion bath at the head of the line which is filled with finish consisting of special lubrication oils (Fig. 82).

Alternatively, finish can be applied before the crimper using kiss rollers (Fig. 82D), into the crimper or after the crimper, or before the cutter by using a spraying unit (Fig. 82B and C). Finishes are applied with minimum loss, which means that little or no rinsing is required between process steps; this eliminates recovery costs and minimizes pollution of the environment.

4.2.3. Drying and Heat-Setting

Fibers are dried primarily to remove water and to obtain a commercially acceptable weight. However, heat-setting begins if drying is performed at high temperature; this considerably improves the fine structure of fibers (especially that of cellulose or wet-spun fibers). On the other hand, polyester fibers with stress-elongation character similar to that of cotton are dried at very low temperature ($< 90\,^{\circ}C$) to prevent changes in fiber structure; this represents pure drying. If the drying temperature were increased in this case, the fibers would be heat-set and thus lose their typical cotton-like properties. Drying and heat-setting can be carried out at constant length or with free or limited shrinkage.

After being drawn, cotton-like polyester fibers are heat-set at constant length on calenders (Fig. 83).

Calender rollers are heated with a suitable heat transfer oil and their surface temperature is brought up to $230\,^{\circ}C$. The tow is run over the requisite number of rollers, with appropriate contact times.

Heat-setting with free shrinkage is usually carried out in a plate belt dryer (Fig. 84A) or a screen drum dryer (Fig. 84B). Here, the material is deposited in meandering loops or in the form of flocks at the entrance to the dryer – setter and is transported through the system. The material is dried and heat-set by circulating a stream of heated gas, usually air. In addition to transfer of heat to the fiber material by means of contact or convection, radiation may also be used.

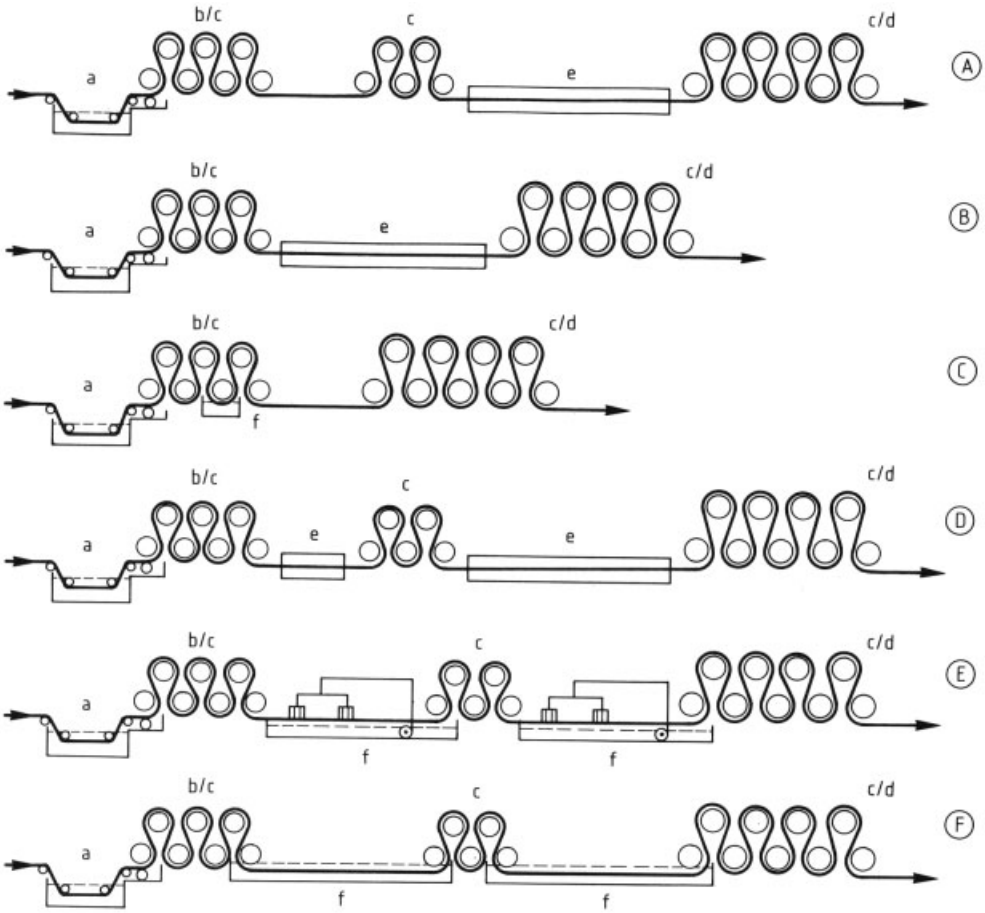

Figure 80. Principal types of drawing used commercially
A) Two-field, two-step drawing; B) One-field, one-step drawing; C) One-field, two-step drawing; D) Two-field, two-step drawing; E) Two-field, two-step drawing; F) Two-field, two-step drawing
a) Immersion bath; b) Feed rollers; c) Draw rollers; d) Heat-setting rollers; e) Steam channel; f) Hot-water bath

4.2.4. Crimping

Texturing in staple fiber production lines is mainly performed with stuffer box crimpers. They are constructed in widths of 20 – 660 mm, can process 3 – 10 ktex of tow per centimeter of width, and operate at a speed of up to 7 m/s. The machines must be built with high precision and must be extremely sturdy. A crimping unit used in staple fiber production lines is shown schematically in Figure 85.

For defect-free processing of tow the material must be fed into the crimper homogeneously. To reduce the working widths of the tow bands from the preceding stage (drawing/heat setting) to the width of the crimper, they are folded with the aid of concave rolls or tow plyers. A dancing roll or a measurement unit is used to obtain uniform, constant tow band tension. Uniform temperature throughout the tow is provided by a steam or vapor zone which precedes the crimper and also affects the modulus of elasticity of the fiber and thus crimp formation.

In the crimper, the material is transported over temperature-controlled rollers, pushed against the plug of material that is forming, and thus forced to buckle (cf. Section 2.3.2.2). Crimp formation is determined by material properties (modulus of elasticity, diameter, friction), the geometry of the stuffer box, and the pressure selected in the box. Steam is often injected into

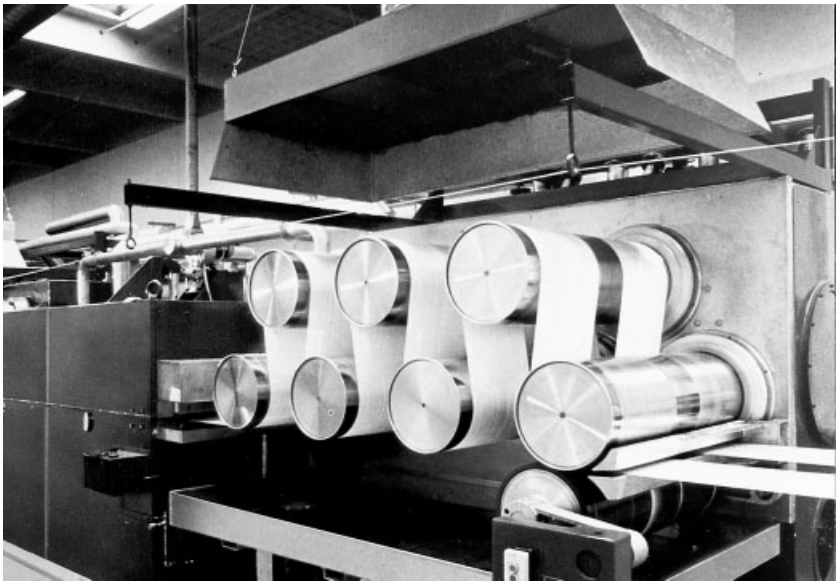

Figure 81. Drawing unit
Courtesy of Fleissner, Egelsbach, Germany

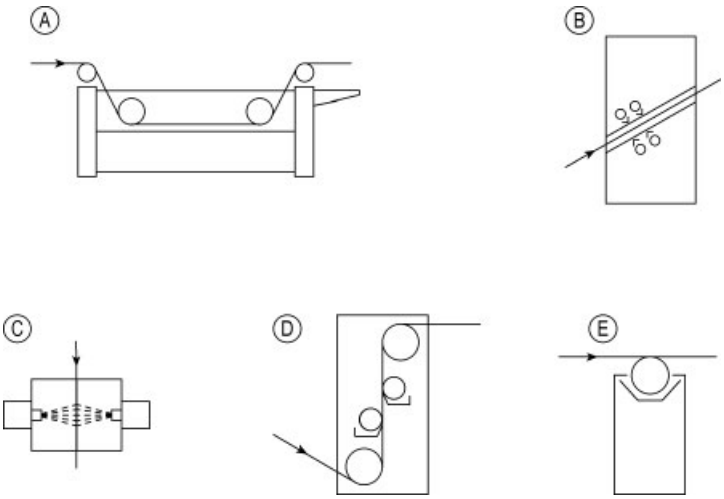

Figure 82. Finishing units for staple fibers
A) Immersion bath; B,C) Spraying units; D,E) Finishing units with kiss rollers

the box to set the texture. A crimper is shown in Figure 86.

4.2.5. Cutting

The continuous tow produced on the staple fiber production line is finally cut by mechanical cutters. For short fibers, the tow is cut into lengths of 6 – 18 mm. Staple lengths of 30 – 45 mm are suitable for processing in equipment designed to handle cotton; lengths of 60 – 150 mm can be processed in equipment designed to handle wool or wool blends.

In older, obsolete cutters, the tow moved over vertically running, slotted wheels and was cut between the slots by rotating sets of blades. Different types of cutters are now used, however.

Figure 83. Heat-setting calender
Courtesy of Fleissner, Egelsbach, Germany

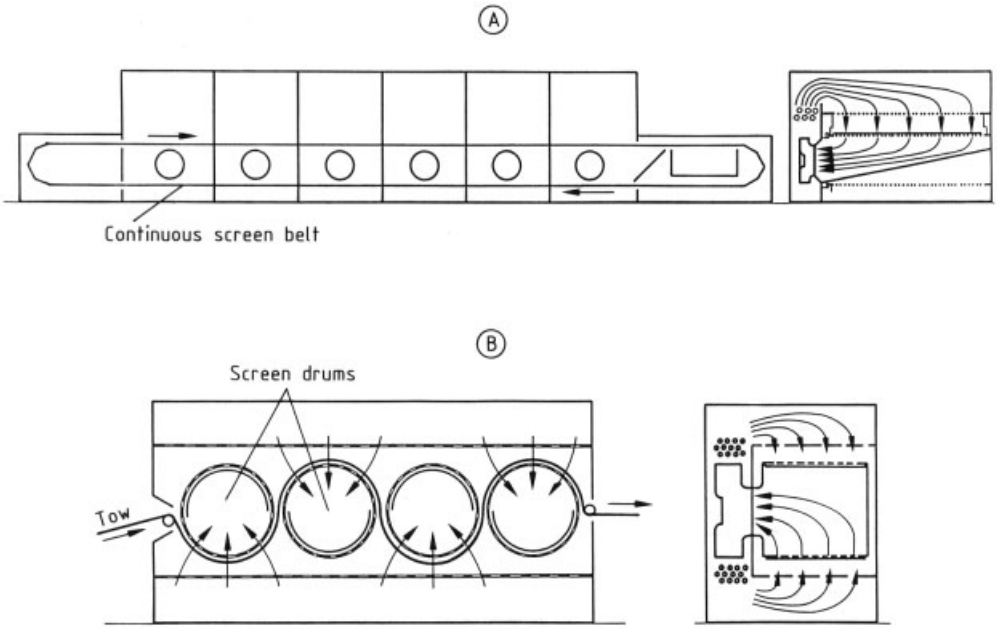

Figure 84. Dryers used for drying and heat-setting staple fibers
A) Plate belt dryer; B) Screen drum dryer
Cross sections of the dryers are shown on the right.

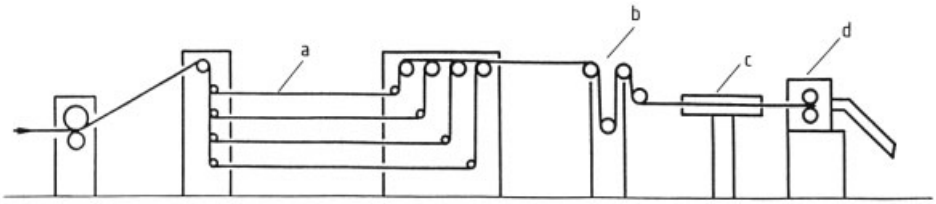

Figure 85. Crimping unit for staple fibers
a) Tow plyer; b) Dancing roller; c) Steam channel; d) Stuffer box

The LUMMUS system is capable of cutting tow at up to 600 ktex at a rate of 5 m/s. The material is wrapped around a rotating, horizontal ring of blades and pressed against the cutting edge of the blades by a pressure roll (Fig. 87).

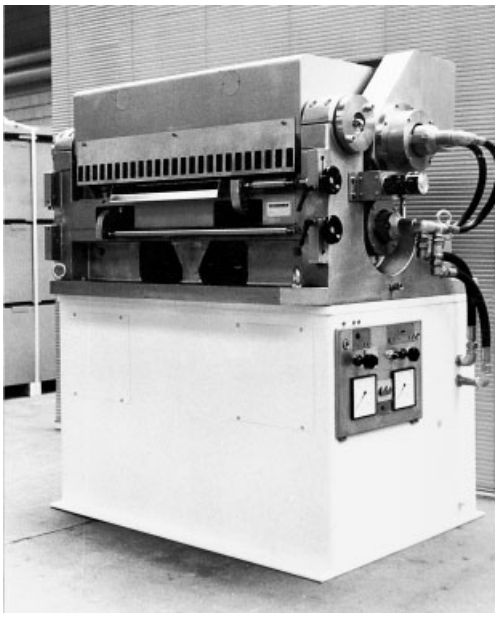

Figure 86. Jumbo Crimper
Courtesy of Fleissner, Egelsbach, Germany

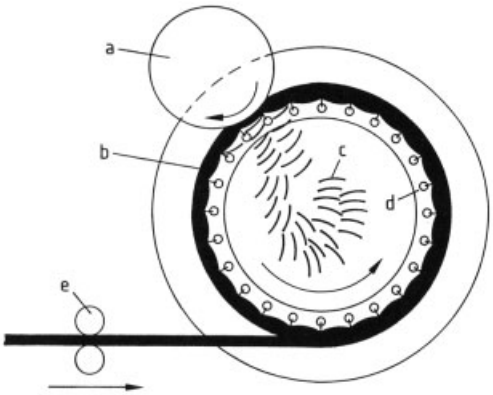

Figure 87. Cutter operating by the LUMMUS principle
a) Pressure disk; b) Tow; c) Cut staple fiber; d) Knife blade; e) Feed rollers

In a high-performance cutting system, the tow is drawn between a rotating cutter head and an inclined, free-running pressure disk positioned above it. The narrowing gap between the pressure disk and the cutter head builds up sufficient pressure in the tow to push the material through the blades (Fig. 88). This cutter can operate at high cutting speeds, and 10-ktex tows can be cut at 83 m/s.

In a third type of cutter, a continuous tow is cut by blades which are arranged in a pattern similar to the thread of a screw. The parallel arrangement of the fibers is thereby retained and the cut fibers therefore form a band. In contrast to the other cutting methods, natural fibers in the form of slivers can be blended with the staple fibers in a second step (converter process).

4.2.6. Packaging

Staple fibers are compressed into rectangular bales weighing up to 500 kg with material densities of > 0.5 g/cm^3. Usually, the fibers are transported pneumatically from the cutters to a prepress for preliminary compacting and are then brought to balesize in a main press. In the press, the material is wrapped in plastic sheeting and tied with metal or plastic straps.

Continuous tow is folded, deposited in boxes, and compacted with press rams. It can also be packed in bales if the tow is first folded and compacted in suitable containers.

5. Manufacture of Monofilaments

Monofilaments with diameters of ca. 0.06 – 2 mm have a wide range of applications, particularly in industry [137]. Monofilaments can only be produced from thermoplastic, melt-spinnable polymers such as polyamides [138], polyesters [137], polypropylene, polyethylene, and poly(vinyl chloride). Polymers that can only be spun by wet spinning (e.g., polyacrylonitrile) are not suitable for the manufacture of monofilaments. Strict requirements are placed on the properties of monofilaments, especially on the constancy of diameter, shrinkage, and strength. Specially optimized units have been developed which provide the required narrow tolerances and, at the same time, the most economical method of monofilament production [139]. Modern, industrial production of monofilaments is carried out in one step in which spinning, drawing, posttreatment, and winding are combined into a continuous process. A monofilament production plant is shown schematically

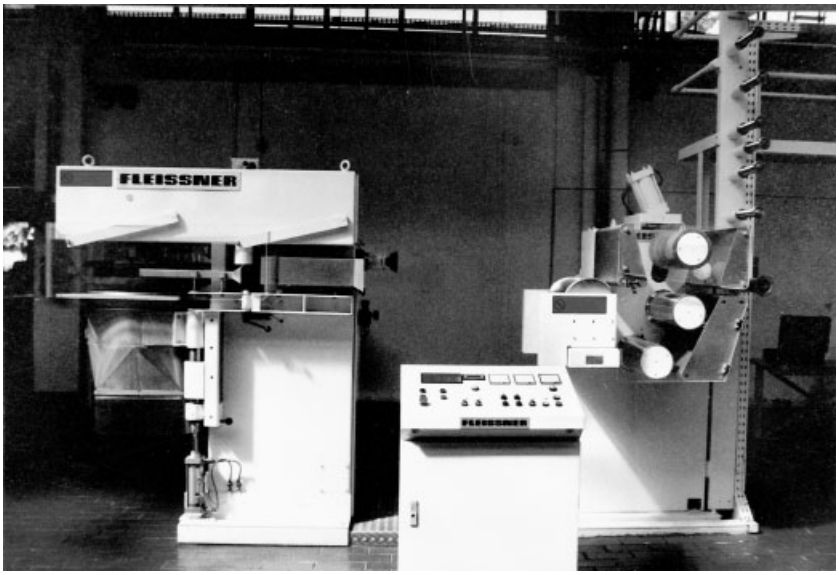

Figure 88. Tension stand with cutter
Courtesy of Fleissner, Egelsbach, Germany

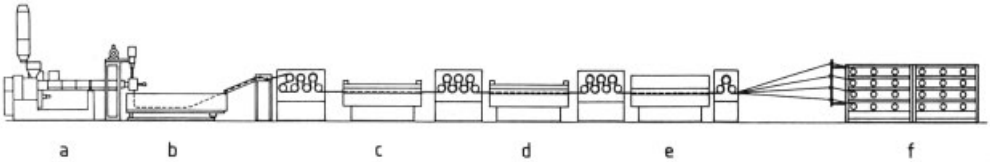

Figure 89. Monofilament production plant
a) Granulate extrusion; b) Water bath; c) First drawing zone; d) Second drawing zone; e) Heat-setting; f) Winding

in Figure 89. Figure 90 shows a monofilament winding machine.

The polymer granulate is melted in an extruder; the melt is then filtered, fed into the spin nozzle and pressed through spinnerets. Additives such as lubricants or stabilizers can be added to the granulate or fed as a side stream into the extruder or into the melt line leaving the extruder. The additives are then distributed homogeneously in the melt with the aid of mixing devices.

Depending on the polymer and the diameter of the monofilament, up to 200 – 300 monofilaments can be spun simultaneously at each spinning position and posttreated in the form of a yarn warp. The melt pressure and the temperature of the monofilament spin heads must be kept constant within narrow limits to ensure that the simultaneously spun monofilaments are of consistent quality. The fluid melt monofilaments are spun into a water bath and drawn off continuously as a yarn warp over guide rolls. They are fed into drawing units equipped with three to seven draw rolls and hot steam or hot air channels.

Figure 90. Monofilament winding machine

Drawing temperature and draw ratio depend on the polymer used to produce the monofilaments and on the desired properties of the end product. Drawing can be followed by heat-setting and/or relaxing before the monofilaments are individually wound on flanged, conical, or biconical bobbins at a speed of 2 – 3.3 m/s. If necessary, a finish oil can be applied prior to winding. Computer-controlled winders with loss-free, fully automatic bobbin changers and up to 64 winding positions are available.

6. Process Technologies for Spunbonded Nonwovens

Nonwovens are classically produced by carding staple fibers (→ Nonwoven Fabrics). Spunbonded nonwovens, however, are made from continuous filaments in a one-step procedure; investment and production costs are thus lower. Thermoplastic polymers such as polypropylene, polyethylene, polyamides, or polyesters are melt spun through special spinnerets to form continuous filaments which are drawn by means of air streams and directly deposited as a random web on a moving collection screen belt. Finally, the filaments are bonded together by heat, chemical, or other entangling techniques, e.g., needle punching or hydroentanglement.

Several different process technologies are available for the industrial production of spunbonded nonwovens. They differ as to whether the filament bundles are spun from circular spinnerets or from rectangular spinnerets like a filament curtain. They also differ with regard to the production of air streams (over- or underpressure), their guidance and generally in depositing the filaments on the collecting belt (web formation).

A typical spunbond process [106] is illustrated in Figure 91. Spinnable thermoplastic polymer chips from a silo are melted in an extruder, into which colored pigments and other additives can be injected. The polymer is prefiltered through a melt filter (continuous type) and fed to the spinning beam. The spinning system is vapor-heated with a mixture of diphenyl and diphenyl oxide and ensures a constant temperature profile across the full width of spinneret. A spin pump forces the melt through the melt distribution line to the spinneret which contains several thousand capillary holes. Hole number, size (0.3 – 0.8 mm), and l/d ratio (1 – 6) mainly depend on the polymer properties, filament fineness, and required capacity.

The liquid filaments emerge from the spinneret, are cooled to room temperature with conditioned air, and enter into the aerodynamic draw-off system. To achieve a high tensile strength it is important that the drawing force caused by friction between air and filament surface is applied near the glass transition point. The drawing forces needed for high filament speeds are generated by the high velocities of the air (up to 16000 m/min) that is sucked in through the drawing slot. Depending on the polymer, filament speeds of, e.g., up to 6000 m/min for polyester and up to 4000 m/min for polypropylene are reached. The air – filament stream is directed downward into the web formation section, and the filaments are deposited on the moving screen belt. The filaments are distributed in a random and entangled manner by controlled turbulence and turbulent vortices in the air stream. As the web-formation chamber is directly connected to the drawing part and this space is fully protected, no disturbance by ambient air or other surrounding effects can take place. This feature minimizes edge trimming requirements and results in economical use of polymer. Underneath the conveyor belt the air is sucked off and recycled to the spinning section. Under predetermined spinning conditions, the fabric weight can be adjusted by varying the speed of the moving screen belt.

For boosting the productivity additionally one or more spinning beams can be added or a melt-blown system can be installed between two beams. This permits the production of spunbond/meltblown/spunbond (SMS). The power consumption of this suction process is relatively low compared with the process where the drawoff system operates with compressed air.

The process permits the production of webs with fabric weight of 8 to 300 g/m^2, widths between 0.6 and 4.6 m, and filament fineness ranging from 0.6 to 6 dtex.

7. Economic Aspects [145–147]

Since the mid 1970s, excess production capacities have resulted in fierce international compe-

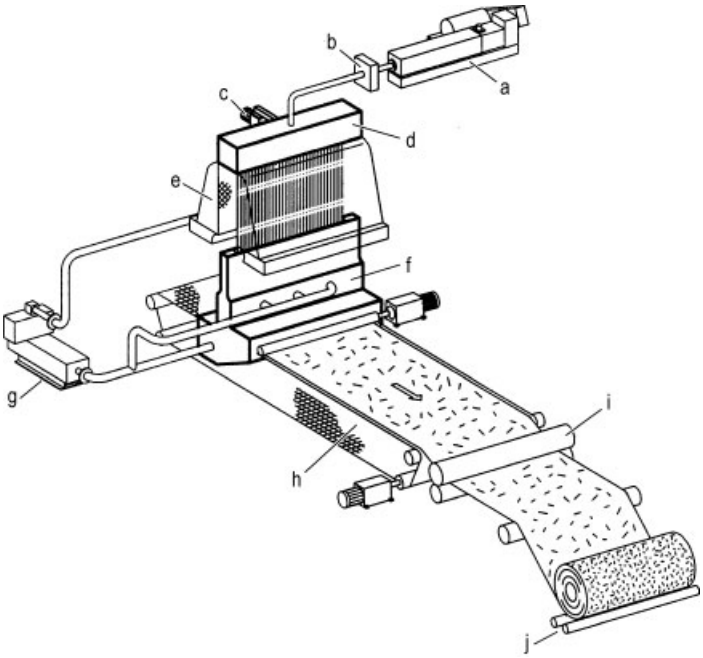

Figure 91. Schematic of the production of spunbonded nonwovens
a) Extruder; b) Melt filter; c) Spin pump; d) Spinning beam;e) Quenching chamber; f) Web-formation chamber; g) Suction blower; h) Conveyor screen; i) Calender; j) Winder

tition among the producers of man-made fibers, particularly in industrialized Western countries. Later the most strongly growing market for man-made fiber was in Asian Countries. The economics and thus the competitiveness of a man-made fiber enterprise are affected to a high degree by the total capacity of its fiber and filament plants [2], by the size of the individual production lines, and by the technologies used.

Until the early 1970s, for example, the polymer required for the production of polyester fiber was made primarily in batchwise or continuous polycondensation plants with capacities of 30 – 60 t/d; dried granulate, as a rule, was thus the starting material used for spinning staple fibers and filament yarns. The profitability of a polyester plant increases considerably with increasing output capacity of a production line [17], [85], [142]. A particularly dramatic reduction in manufacturing costs can be achieved in continuous installations by omitting the granulate step and employing direct spinning in which the fibers are spun directly from the melt. Investment costs for buildings and equipment are lower, and energy consumption is reduced because granulate no longer has to be produced, dried, and remelted. As a result, almost all polyester fiber plants built today are continuous production lines which employ direct spinning; production capacities are 100 – 200 t/d in one line.

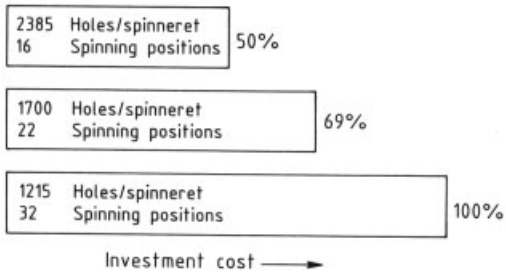

Figure 92. Dependence of investment costs on number of spinning positions and number of holes per spinneret
Figures are based on a melt-spinning unit with central quenching and a production capacity of 45 t/d

Investment and manufacturing costs of polyester staple fibers are affected to a high degree by the output of the spinning positions and the capacity of the drawing units. Figure 92 shows that investment costs of a staple fiber

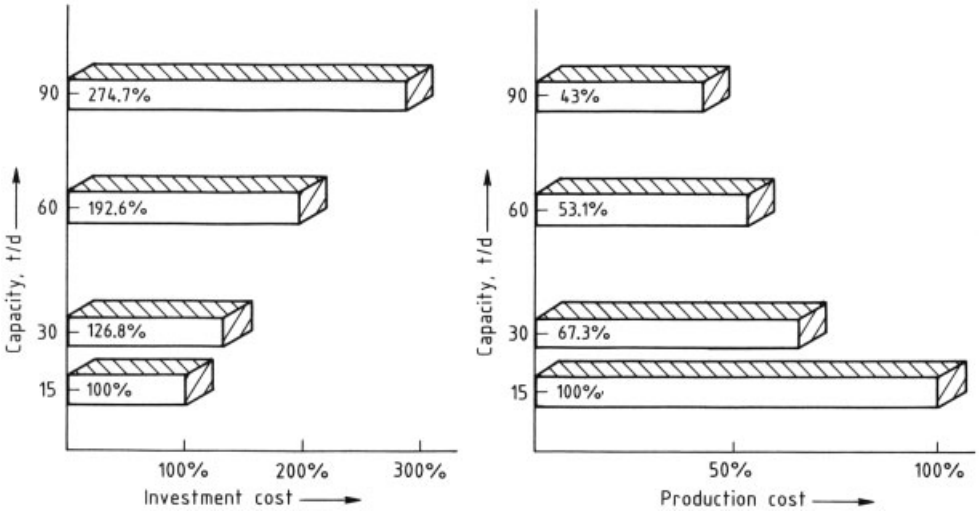

Figure 93. Comparison of investment and production costs for fiber production lines in relation to their capacity

spinning plant can be halved by doubling the 1215 spinneret holes per spinning position and halving the number of spinning positions. Figure 93 demonstrates that manufacturing costs in the drawing unit can be more than halved if the installation has a capacity of 90 t/d instead of 30 t/d.

Efforts to improve competitiveness in the area of filament yarns have focused on reducing investment costs by combining multistep, discontinuous processes into a single, continuous process. One-step processing also reduces manufacturing costs by cutting down personnel requirements. However, modern one-step processes not only provide cost advantages, but generally result in an improvement in the quality of filament yarns.

Studies of the most important factors affecting the manufacturing costs involved in draw texturing show that an integrated, one-step, spin-draw texturing process is more economical than conventional draw texturing only if the texturing speed can be increased above 67 m/s [143].

In processes used to produce filament yarns (e.g., spinning, high-speed spinning, spin drawing, draw texturing), the number of yarn ends per spinning or drawing position has a considerable influence on the manufacturing cost of the end product. In draw texturing used to produce BCF yarns, cost advantages are obtained if finenesses up to ca. 1000 dtex are processed on a four-end machine instead of a two-end machine [144]. The break-even point is calculated to be a fineness of 1200 dtex.

Fierce international competition in the production of synthetic fibers and the construction of synthetic fiber plants is forcing the man-made fiber industry and engineering companies to develop new, lower cost, better quality technologies. Only the replacement of out-of-date, unprofitable production facilities by new, highproductivity technologies can assure the long-term profitability and competitiveness of an enterprise.

8. References

1. A. Ziabicki: *Fundamentals of Fibre Formation,* Wiley-Interscience, New York 1976.
2. B. von Falkai: *Synthesefasern,* Verlag Chemie, Weinheim 1981.
3. Y. Oshima, H. Maeda, T. Kawai, *J. Chem. Soc. Kogyo Kagaku Zarshi* **60** (1957) 311.
4. A. Ziabicki, R. Taksermann-Krozer, *Kolloid Z.* **198** (1964) 60.
5. W. Gleissle, Vortrag AIChE annual meeting, New Orleans 1981.
6. P. J. Flory: *Statistical Mechanics of Chain Molecules,* Wiley-Interscience, New York 1969.
7. H. M. Laun, H. Münstedt, *Rheol. Acta* **17** (1978) 415; *Rheol. Acta* **18** (1978) 492.

8. M. H. Pahl: *Praktische Rheologie der Kunststoffschmelzen und Lösungen,* VDI-Verlag, Düsseldorf 1983.
9. J. Meissner, *Kunststoffe* **61** (1971) 576, 688.
10. C. J. S. Petrie, M. M. Denn, *AIChE J.* **22** (1976) 209.
11. W. Gleissle, *Rheol. Acta* **21** (1982) 484.
12. *Winnacker-Küchler,* **6,** 626.
13. A. Ziabicki, *Kolloid Z.* **175** (1961) 14.
14. A. Ziabicki, H. Kawai: *High-Speed Fiber Spinning,* Wiley-Interscience, New York 1985.
15. S. Kase, T. Matsuo, *J. Polym. Sci. Part A,* **3** (1965) 2541.
16. I. Hamana: "Vortragsschrift zur VII. Internationalen Chemiefasertagung in Dornbirn (Oesterreich) Juni 1968," *Lenzinger Ber.* **26** (1968) 118.
17. H. Lückert, W. Stibal, *Chemiefasern Textilind.* **36** (1986) 22 – 29.
18. V. Gröbe, H. Versäumer, *Faserforsch. Textiltech.* **14** (1963) 249.
19. Du Pont, US 3 686 377, 1971 (R. Y. Hayes).
20. C. Chen, J. A. Cuculo, P. A. Tucker, *J. Appl. Polym. Sci.* **44** (1992) 447.
21. J. A. Cuculo, P. A. Tucker, G. Chen, *J. Appl. Polym. Sci. Appl. Polym. Symposium* **47** (1991) .
22. ICI, WO 97/34758, 1996 (S. Jenkins).
23. L. Xie, G. B. DeMaggio, W. E. Friese, J. DeVries, D. W. Gidley, H. A. Histrov, A. F. Yee, *Phys. Rev. Lett.* **74** (1995) 4947.
24. S. Ishizaki, K. Iohara, K. Fujimoto, *Sen-i Gakkaishi* **45** (1989) 234.
25. J. Shimizu, T. Kikutani, A. Takaku, N. Okui, *Sen-i Gakkaishi* **37** (1981) T-135.
26. J. Shimizu, *Sen-i Gakkaishi* **38** (1985) 243.
27. R. J. Samuels, *J. Polym. Sci.* **10** 1972) 781.
28. B. Clauss, D. R. Salem, *Polymer* **33** (1992) 3193.
29. H. Yasuda: "Structure Development and Physical Properties of High-Speed-Spun Pet Fibers", in A. Ziabicki, H. Kawai (eds.): *High Speed Fiber Spinning,* Wiley, New York 1985, pp. 363 – 381.
30. K. Fujimoto, Ph. D. Thesis, Institute of Technology, Tokyo 1989.
31. J. Shimizu, N. Okui, T. Kikutani, *Sen-i Gakkaishi* **37** (1981) T-135.
32. K. Iohara, S. Ohawaki, Y. Murase, Preprints of Gordon Conference, 1985.
33. G. Kühnle, E. Schollmeyer, H. Herlinger, *Makromol. Chem.* **178** (1977) 2723.
34. G. Kühnle, E. Schollmeyer, H. Herlinger, *Makromol. Chem.* **179** (1978) 661.
35. G. Kühnle, E. Schollmeyer, H. Herlinger, *Makromol. Chem.* **180** (1979) 472.
36. J. Shimizu, N. Okui, T. Kikutani, A. Takaku, Text. Ind. Conf. , Raleigh, NC, 1981, p. 156.
37. C. Chen, J. A. Cuculo, P. A. Tucker, *J. Appl. Polym. Sci.* **44** (1992) 447.
38. J. F. Hotter, J. A. Cuculo, P. A. Tucker, *J. Appl. Polym. Sci.* **43** (1991) 1511.
39. K. Fujimoto, K. Iohara, S. Ohawaki, Y. Murase, *J. Appl. Polym. Sci.* **42** (1991) 1509.
40. J. S. Raghavan, J. A. Cuculo, *J. Appl. Polym. Sci. B Polym. Phys.* **37** (1999) 1565.
41. Akzo, WO 98/47936, 1997 (H. Hüther, C. De Jong, D. Schilo, U. Seidel, Ch. Vieth, H. Zengel).
42. C. D. Han et al., *J. Appl. Polym. Sci.* **15** (1971) 1163.
43. C. D. Han et al., *J. Appl. Polym. Sci.* **18** (1974) 2589.
44. L. A. Utracki, A. M. Catani, G. L. Bata, M. R. Kamal, V. Tan, *J. Appl. Polym. Sci.* **17** (1982) 1913.
45. H. J. Brody, *Appl. Polym. Sci.* **31** (1986) 2753.
46. ICI, EP-0 041 327, 1980 (H. Brody).
47. Teijin, EP 0 047 464, 1980 (K. Iohara, M. Matsui).
48. Akzo, EP 0 631 638 B1, 1992 (B. Böhringer, D. Schilo).
49. Lurgi-Zimmer, DE 19 747 867, 1997 (A. Klein, J. Czoillek, H. Schurmann, U. Thiele, D. Wandel).
50. W. Stibal, Lecture at the ″World Polyester Conference″, Zürich 1997, EMS-Inventa AG.
51. Internal publication, INVENTA-FISCHER AG 1999.
52. W. Stibal, ″HP5 – A Novel Polymer for the Fiber Producer″, Lecture at the ″World Polyester Conference″ Zürich 1998, EMS-INVENTA AG.
53. Y. Oyanagi, M. Matsumoto, *J. Colloid Sci.* **17** (1962) 426.
54. Y. Ohzawa, Y. Nagano, T. Matsuo, *Proc. 5th Int. Congress on Rheology,* Kyoto 1968, vol. 4, (1970) 393.
55. J. Corbière in H. F. Mark, S. M. Atlas, E. Cernia (eds.): *Man-Made Fibers, Science and Technology,* **vol. 1,** Wiley-Interscience, New York 1967.
56. J. L. White, T. A. Hancock, *J. Appl. Polym. Sci.* **26** (1981) 3157.
57. A. Peterlin in H. F. Mark, S. M. Atlas, E. Cernia (eds.): *Man-Made Fibers, Science and Technology,* **vol. 1,** Wiley-Interscience, New York 1967.
58. H. Zahn, *TPI Text. Prax. Int.* **41** (1986) 1293.

59. H. Dittmann, *Acta Polym.* **32** (1981) 260; *Acta Polym.* **35** (1984) 652.
60. F. Schultze-Gebhardt, *Acta Polym.* **37** (1986) 247.
61. H. Ludewig: *Polyesterfasern,* Akademie-Verlag, Berlin 1975.
62. Akzo, CH 631 311, 1979 (H. Stapp); DE 2 824 500, 1978 (H. Stapp).
63. R. H. Butler, D. C. Prevorsek, Y. D. Kwon, *Polym. Eng. Sci.* **22** (1982) no. 6, 329.
64. J. Lünenschloss, *Chemiefasern Text. Anwendungstech. Text. Ind.* **23** (1973) 1067.
65. I. Jacob, H. R. Schröder, *Chemiefasern Textilind.* **28** (1980) 114, 228.
66. T. He, *Polymer* **27** (1986) no. 2, 253.
67. *Chemiefasern Textilind.* **35** (1985) 1, 33.
68. M. P. Nosov, V. A. Smirnova, *Acta Polym.* **34** (1983) no. 7, 434.
69. G. Hinrichsen, *Tech. Rundsch.* **78** (1986) 38, 20.
70. *Ullmann,* 4th ed. **11,** 282 – 284.
71. K. Riggert, *Melliand Textilber.* **4** (1977) 274 – 279.
72. P. Ehrler, J. Mavely, *Melliand Textilber.* **2** (1983) 99 – 106.
73. P. M. Latzke, R. Hesse: *Textilien Prüfung/Untersuchung/Auswertung,* Verlag Schiele & Schön, Berlin 1974, pp. 198 – 217.
74. P. Ehrler, W. Sprenkmann, *Chemiefasern Textilind.* **31** (1981) 656 – 664.
75. Zweigle: Company brochure and handbook "Kräuselwaage", p. 160.
76. DIN 53 840, Nov. 1983.
77. ICI, US 4 038 357, 1977 (B. Boyes, A. Jones).
78. H. J. Berndt, *TPI Text. Prax. Int.* **38** (1983) 1241.
79. K. Denbigh: *The Principles of Chemical Equilibrium,* Cambridge University Press, London 1973.
80. F. Schultze-Gebhardt, *Angew. Makromol. Chem.* **98** (1981) 63.
81. L. Vollrath, *Textiltechnik (Leipzig)* **31** (1981) 11, 703.
82. G. Valk, G. Heidemann, *Chemiefasern Textilind.* **31** (1981) 854.
83. G. Heidemann, H. J. Berndt, *Chemiefasern Textilind.* **31** (1981) 866.
84. H. Klare, E. Fritsche, V. Gröbe: *Synthetische Fasern aus Polyamiden,* Akademieverlag, Berlin 1963, pp. 239.
85. H. Lückert, M. Busch, *Chemiefasern Textilind.* **33** (1983) 29 – 38.
86. M. Mayer, *Chemiefasern Textilind.* **35** (1985) 24 – 25.
87. ICI, DE-OS 2 836 514, 1979 (P. Carr).
88. A. Ziabicki, H. Kawai: *High Speed Fiber Spinning,* J. Wiley & Sons, New York 1985, pp. 487 – 505.
89. F. Maag, *Melliand Textilber.* 1986, 98.
90. F. Fourné: *Synthetische Fasern,* West Verlag, Stuttgart 1964, p. 96.
91. F. Fourné: *Synthetische Fasern,* West Verlag, Stuttgart 1964, p. 135.
92. Hoechst, EP 0 173 221 A2, 1985 (H. Thaler).
93. F. Maag, *Chemiefasern Textilind.* **35** (1985) 306.
94. R. C. Mears, *Chemiefasern Textilind.* **35** (1985) 413.
95. F. Maag, *TPI Text. Prax. Int.* **40** (1985) 231.
96. R. Th. Maier, *Chemiefasern Textilind.* **36** (1986) 517.
97. M. Busch, U. Kemp, *Fiber Prod.* **9** (1981) 26, 28, 30, 32, 60.
98. H. F. Mark, S. M. Atlas, E. Cernia: *Man-Made Fibers,* **vol. 3,** Wiley-Interscience, New York 1968, pp. 173 – 243.
99. H. F. Mark, S. M. Atlas, E. Cernia: *Man-Made Fibers,* **vol. 3,** Wiley-Interscience, New York 1968, pp. 264 – 302.
100. T. Okuda, *Chemiefasern Text. Anwendungstech.* **20** (1970) 857.
101. Kurashiki Rayon Co., DE-OS 1 937 985, 1968 (A. Tadao, K. Susumu).
102. EMS-Inventa, CH 658 076, 1982 (H. Dalla Torre); EMS-Inventa, DE 2 850 337, 1978 (P. Schäfer, M. Capaul, W. Griehl, P. Meier)
103. R. Kirschbaum, H. Yasuda, E. H. M. Van Gorp, *Chemiefasern Textilind.* **36** (1986) T135 – T139.
104. Stamicarbon, US 4 422 993, 1980, (P. Smith, P. J. Lemstra). P. Smith, P. J. Lemstra, *Polymer* **21** (1980) 1341 – 1343.
105. B. Wulfhorst: *Stand und Zukunft der Texturierung,* Deutscher Fachverlag, Frankfurt am Main 1975, pp. 10 – 12.
106. World Markets for Textured Yarns, Textile Intelligence Ltd., Romford 1995.
107. H. Rabenstein: *Vortrag Reutlinger Texturier-Colloquium,* Oct. 1984.
108. L. Riehl, *Chemiefasern Textilind.* **35** (1985) 27 – 31.
109. BARMAG, Information Service Nr. 25, 20 – 22.
110. G. Bock, J. Lünenschloss, *Chemiefasern Textilind.* **34** (1984) 492 – 496.
111. F. J. van Aken, *Chemiefasern Textilind.* **29** (1979) 108 – 116.
112. E. Krenzer, *Chemiefasern Textilind.* **35** (1985) 674 – 678.

113. J. Kessler, *Chemiefasern Textilind.* **36** (1986) 295 – 296.
114. C. Iyer: *Dissertationsarbeit,* TH Stuttgart 1974.
115. H. Brüttig, DD 17 786, 1957 (H. Bruettig).
116. H. Schellenberg, *Chemiefasern Textilind.* **34** (1984) 902 – 908.
117. D. Ahrendt, H. Dietrichs, *Chemiefasern Textilind.* **31** (1981) 207 – 209.
118. H. Vieth, E. Göbel, *Textiltechnik (Leipzig)* **28** (1978) 618 – 621; *Textiltechnik (Leipzig)* **29** (1979) 469 – 473.
119. P. Piller: *Bulked Yarns,* SNTL-Publishers of Technical Literature, Praque, in coedition with the Textile Trade Press, 11 Albert Sq., Manchester M2 5 HD 1973, pp. 430 – 470.
120. R. Jeffries: *Biocomponent Fibers,* Merrow Publishing Co. Ltd., 276 Hemstead Road, Watford, Herts, England 1971.
121. ICI, GB 809 273, 1955 (R. H. Speakman); ICI, US 3 629 386, 1969 (P. E. Knapp).
122. Monsanto, US 4 600 644, 1982 (J. E. Bromley, J. M. Chamberlin, J. Yu).
123. Monsanto, US 4 202 854, 1977 (P. T. Howse, A. E. Wilkie).
124. F. Fourné, *Chemiefasern Textiltechnik* **18** (1968) 353 – 359.
125. L. Riehl, *Chemiefasern Textilind.* **37** (1987) 32 – 42.
126. M. E. Schwab, *Chemiefasern Textilind.* **36** (1986) 499 – 502.
127. F. Fourné, DE 2 930 553, 1979 (F. Fourné).
128. Du Pont, US 3 050 821, 1962 (J. J. Kilian).
129. Neumag, DE 2 553 866, 1975 (E. Bauch).
130. IWEKA, DE 2 747 706, 1977 (H. Enneking, G. Schubert, L. Hanger).
131. A. Schweitzer, *Chemiefasern Textilind.* **28** (1978) 41 –42.
132. BARMAG, DE 3 523 321 A1, 1985 (M. Bussmann, F. Hensen, W. Runkel).
133. D. Ahrendt, *Textiltechnik (Leipzig)* **36** (1986) 242 –246.
134. Amer. Cyanamid, US 4 163 770, 1979 (H. Porosoff).
135. F. Schmitt, *Melliand Textilber.* **63** (1982) 830 – 835.
136. K. Riggert, *Chemiefasern Textilind.* **31** (1981) 638 – 648.
137. G. Beyer, *Chemiefasern Textilind.* **32** (1982) 452.
138. R. Vieweg, A. Müller, *Kunststoff-Handbuch,* **vol. VI,** Hauser Verlag, München 1966, pp. 336 – 340.
139. *Chemiefasern Textilind.* **32** (1982) 458.
140. L. Hartmann, F. Fourné, *Chemiefasern Textilind.* **32** (1972) 231 – 236.
141. R. Hoffmeister, *Chemiefasern Textilind.* **32** (1982) 724.
142. L. Riehl, H. Reiter, *Chemiefasern Textilind.* **33** (1983) 22 – 29.
143. F. Maag, *Chemiefasern Textilind.* **31** (1981) 118 –124.
144. H. Schellenberg, *Chemiefasern Textilind.* **34** (1984) 902 – 908.
145. World Markets for Textured Yarns, Textile Intelligence Ltd., Romford/UK, 1995.
146. Chemical Fibers International, Volume 48 (September 1998) 278.
147. BARMAG AG, Technical data, FK6 - 1000, Tex 100/2 (9/1998).

Fibers, 4. Synthetic Organic

Leland L. Estes, (E. I. Du Pont de Nemours & Co.), Nashville, Tennessee 37215, United States (Chap. 1)

Helmut Sattler, Königstein-Schneidhain, Hoechst Aktiengesellschaft, Federal Republic of Germany (Chap. 2)

Hans Berg, Königstein-Schneidhain, Hoechst Aktiengesellschaft, Federal Republic of Germany (Chap. 2)

Karl-Heinz Wolf, Bayer AG, Dormagen, Federal Republic of Germany (Chap. 3)

Michael Kausch, Bayer AG, Dormagen, Federal Republic of Germany (Chap. 3)

Hans Schröer, Bayer AG, Dormagen, Federal Republic of Germany (Chap. 3)

Antonio Pellegrini, Moplefan, Milano, Italy (Section 4.1)

Paolo Olivieri, Moplefan, Milano, Italy (Section 4.1)

Werner Schoene, BASF AG, Ludwigshafen, Federal Republic of Germany (Sections 4.2–4.4)

Alfred Nogaj, Bayer AG, Dormagen, Federal Republic of Germany (Chap. 5)

Carlhans Süling, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

Jacques Menault, (Rhovyl), St-Foy-les-Lyons, France (Chap. 6)

Tetsuro Osugi, Nippon Chemtec Consulting Inc., Osaka, Japan (Chap. 7)

Osamu Morimoto, Nippon Chemtec Consulting Inc., Osaka, Japan (Chap. 7)

Peter E. Frankenburg, E. I. Du Pont de Nemours, Wilmington, Delaware 19898, United States (Chap. 8)

1. Polyamide Fibers 4
1.1. Properties 4
1.2. Polyamides for Synthetic Fibers . 5
1.2.1. Aminocarboxylic Acid Type Polyamides 6
1.2.2. Diamine – Dicarboxylic Acid Type Polyamides 9
1.2.3. Copolyamides 10
1.3. Special Requirements for Polyamides 11
1.4. Modifying Additions 12
1.5. Technology of Melt Spinning . . . 13
1.5.1. Melt Spinning of Nylon 13
1.5.2. Special Techniques and Products . . 15
1.6. Uses 16
1.7. Economic Aspects, Manufacturers, and Trade Names 16
2. Polyester Fibers 17
2.1. Production 17
2.1.1. Starting Materials and Polymer . . . 17
2.1.2. Spinning 19
2.1.3. Drawing 21
2.1.4. Aftertreatment 22
2.1.5. Dyeing of Poly(ethylene terephthalate) Fiber Products 23
2.2. Structure and Properties 25
2.2.1. Polymer 25
2.2.2. As-Spun Yarns 27
2.2.3. Drawn Materials 28
2.2.4. Dyed Poly(ethylene terephthalate) Material 28
2.3. Poly(ethylene terephthalate) Fibers 29
2.3.1. Staple Fibers 29
2.3.2. Continuous-Filament Yarns for Apparel and Home Furnishings 32
2.3.3. Continuous-Filament Yarns for Industrial Applications 35
2.3.4. Spunbonds 38
2.3.5. Monofilaments 39
2.4. Other Polyester Fibers 40
2.4.1. Poly(tetramethylene terephthalate) . 40
2.4.2. Poly(1,4-dimethylenecyclohexane terephthalate) 41
2.5. Uses 41
2.6. Economic Aspects 44
2.7. Polyester: Summary and Future Prospects 45
3. Polyurethane Fibers 48
3.1. Physical Properties 48
3.2. Chemical Properties 50
3.3. Raw Materials 50
3.4. Production of Polymer 52
3.5. Spinning Processes 52

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a10_567

3.6. Uses 53
4. Polyolefin Fibers 54
4.1. Polypropylene Fiber 55
4.1.1. Chemical and Physical Properties . 55
4.1.2. Production 57
4.1.2.1. Raw Materials 57
4.1.2.2. Staple Fiber Production Process . . 57
4.1.2.3. Continuous Filament Yarn Production Process 59
4.1.2.4. Stabilization 59
4.1.2.5. Pigmentation 60
4.1.3. Uses 61
4.1.4. Economic Aspects 63
4.2. Polyolefin Film Yarns and Split Yarns 63
4.2.1. Raw Materials 63
4.2.2. Production of Tapes and Split Yarns 64
4.2.3. Uses 66
4.3. Monofilaments 67
4.3.1. Production 67
4.3.2. Uses 68
4.4. High-Modulus Polyolefin Yarns . 68
4.4.1. Production 69
4.4.2. Properties and Uses 69
5. Polyacrylonitrile Fibers 69
5.1. Physical Properties 70
5.2. Chemical Properties 70
5.3. Production 71
5.3.1. Polymerization 71
5.3.2. Dissolving 73
5.3.3. Spinning 74
5.3.4. Aftertreatment 75
5.3.5. Special Processes 77
5.4. Analysis 78
5.5. Types of Fibers 80
5.6. Economic Aspects 83
6. Poly(vinyl chloride) and Poly(vinylidene chloride) Fibers . 83
6.1. Fibers from Pure Poly(vinyl chloride) Homopolymer 84
6.1.1. Fibers from Pure Atactic Homopolymer 84
6.1.2. Fibers from Blends of Poly(vinyl chloride) Homopolymer 85
6.1.3. Products, Properties, and Uses . . . 86
6.2. Fibers from Postchlorinated Poly(vinyl chloride) (CPVC) . . . 87
6.3. Fibers from Vinyl Chloride Copolymers 87
6.4. Fibers from Poly(vinylidene chloride) and Copolymers 88
7. Poly(vinyl alcohol) Fibers 88
7.1. Physical and Chemical Properties 89
7.2. Production 89
7.3. Poly(vinyl alcohol) for Fiber . . . 89
7.3.1. Wet Spinning 89
7.3.2. Dry Spinning 91
7.3.3. Mixed Spinning 91
7.4. Uses 92
7.5. Economic Aspects 92
8. Polytetrafluoroethylene Fibers . . 92
8.1. Physical and Chemical Properties 92
8.2. Production 93
8.3. Uses 93
8.4. Toxicology and Occupational Health 94
9. References 94

The following is a list of specific abbreviations used in this article.

BCF	bulked continuous filament
CPVC	postchlorinated poly(vinyl chloride)
DMA	dimethylacetamide
DMF	dimethylformamide
DMSO	dimethyl sulfoxide
DMT	dimethyl terephthalate
FOY	fully oriented yarn
HALS	hindered amine light stabilizer
HDPE	high-density polyethylene
HE	highly elastic
HOY	highly oriented yarn
HT	high temperature
LDPE	low-density polyethylene
LOI	limiting oxygen index
LOY	low-oriented yarn
MDI	methyl diisocyanate
MFI	melt flow index
M_rD	molecular mass distribution
PAC	polyacrylonitrile
PACM	bis(4-aminocyclohexyl)methane
PBT	poly(butylene terephthalate) = poly(tetramethylene terephthalate)
PDCT	poly(1,4-dimethylenecyclohexane terephthalate)
PETP	poly(ethylene terephthalate)
POY	partially oriented yarn
PP	polypropylene
PTFE	poly(tetrafluoroethylene)
PUR	polyurethane
PVA	poly(vinyl alcohol)

PVC	poly(vinyl chloride)
PVDC	poly(vinylidene chloride)
TDI	tolyl, tolylene, or toluene diisocyanate
TMA	thermomechanical analysis
TPA	terephthalic acid
UHMM	ultrahigh molecular mass
UHMMPE	ultrahigh molecular mass polyethylene

1. Polyamide Fibers

This chapter deals predominantly with linear polyamide fibers derived from aliphatic monomers, all of the generic class called *nylon.* Linear, aliphatic polyamides are conveniently divided into two groups: those made from aminocarboxylic acids and those made from diamine – dicarboxylic acids. These are characterized by the number of carbon atoms in the monomeric starting material. For example, nylon 11 is the polycondensate of 11-aminoundecanoic acid [$H_2N(CH_2)_{10}COOH$]. Nylon 610 is the polycondensate of 1,6-diaminohexane [$H_2N(CH_2)_6NH_2$] and sebacic acid [$HOOC(CH_2)_8$ COOH]. Fibers of aromatic polyamides in which the amide linkages attach directly to two aromatic rings are called aramids, and are covered under → High-Performance Fibers. World production of nylon fiber reached 3.4×10^6 t/a in 1985 and is expected to grow slowly. The 45-year history of the beginning and growth of nylon has been reviewed [7].

1.1. Properties

Nylon fibers are superior to all natural and regenerated fibers with respect to high tenacity and elongation, abrasion resistance, and insensitivity to rotting and moth attack. By heat treatment, fabrics of nylon fiber can be set into stable shapes through formation of additional hydrogen bonds. The highly polar nature of carbonamide groups also results in swelling of nylon fibers in water or in polar solvents, and permits simple dyeing processes with disperse, acid, and metallized dyes. Nylon fibers absorb moisture and swell longitudinally. This limits "outdoor" use because of poor dimensional stability.

Other general properties of nylon are as follows:

Shrinkage: dependent on conditions; 1 – 15 % when setting in water or hot air

Abrasion resistance: outstanding; especially suitable for textile floor coverings

Electrostatic charge: high at low relative humidity; reduced significantly in antistatic fibers (addition of carbon black or polyglycol ether compounds)

Solubility: at room temperature in concentrated inorganic acids, *m*-cresol, *o*-chlorophenol, and phenol – alcohol mixtures; at the boil in benzyl alcohol and phenylethyl alcohol

Dyeability: with acid, metal-complex, disperse, and reactive dyes; special types with basic dyes

Stability:

Light stability: highly dependent on pigmentation and presence or absence of stabilizers

Thermal oxidation: yellowing and molecular degradation reduced by stabilizers (Cu^+ or antioxidants)

Weathering: relatively poor outdoor life, especially with fine deniers

Dilute acid: adequate resistance

Dilute alkali: good resistance

Biological stability: good to very good resistance against decay-producing bacteria, mold fungi, and moth larvae

Shape stability: good crease resistance with dry materials. On absorbing water, nylon fibers swell in the direction of the fiber axis. The lengthening must be considered in the manufacture of clothing, especially with nylon 6 and nylon 66. With felts for floor coverings, dimensional stability is improved by blending with nonswelling fibers such as polypropylene.

Physical properties of four nylon fibers are given in Table 1.

Table 1. Physical properties of some nylon fibers

Property	Nylon 6	Nylon 66	Nylon 11	PACM-12∗
mp, °C	214 – 220	255 – 260	190	275
ϱ, g/cm^3	1.14	1.14	1.04	1.03
Water absorption, %				
65 % R.H., 21 °C	3.5 – 4.5	3.5 – 4.5	1.3	1.5 – 2
Water retention	9 – 15	9 – 15	2.9	ca. 3
Tenacity, cN/dtex	3 – 9	3 – 9	4 – 7	2.5 – 3
Elongation at break, %	80 – 15	80 – 15	40 – 15	30 – 25
Wet – dry strength, %	80 – 90	80 – 90	100	NA∗∗
Loop – straight strength, %	70 – 90	70 – 90	NA∗∗	NA∗∗
Elastic modulus, cN/dtex				
Textile fibers	5 – 30	5 – 30	45	20 – 30
Industrial fibers	60 – 90	60 – 90	NA∗∗	NA∗∗
Setting temperature, °C	185 – 195	200 – 230	NA∗∗	190

∗ Nylon from bis(4-aminocyclohexyl)methane and dodecanedicarboxylic acid.
∗∗ Not applicable.

1.2. Polyamides for Synthetic Fibers

The high molecular mass, linear polyamides that are useful for synthetic fibers contain mainly methylene groups between the carbonamide groups. They comprise those compositions that can be melted and spun into filaments without decomposing.

The introduction of cyclic segments into an aliphatic polyamide chain generally stiffens the chain and raises the melting point (Table 2). The effect is most pronounced when the ring structure is aromatic and symmetrical, and when the amide group is connected directly to an aromatic ring. Connection of two aromatic rings to the amide group defines the aramid class of polyamides, which usually degrade without melting and cannot be melt-spun. Melt temperature is also strongly affected by ring symmetry, as seen in the 90 °C difference in melting points of the meta and para isomers of the two xylene diamines.

Introduction of C-alkyl side chains in nylons disrupts molecular order and leads to both a reduction in melt temperature and increased solubility in organic solvents. The influence of N-alkyl substituents is even greater, because they eliminate the possibility of hydrogen bonding between polyamide molecules. The polyamide of piperazine and adipic acid, which melts at 185 °C, is of scientific interest because the effect of the missing hydrogen bonds is largely compensated for by chain stiffening through the piperazine radical.

The effect of C-methyl and N-methyl groups on the melting temperature of nylons made from adipic acid and a diamine is as follows:

$H_2N-(CH_2)_4-NH_2$	251 °C
$H_2N-(CH_2)_2-CH(CH_3)-(CH_2)_3-NH_2$	180 °C
$HN(CH_3)-(CH_2)_6-NH_2$	145 °C
$HN(CH_3)-(CH_2)_6-NH(CH_3)$	− 75 °C

Substitution of oxygen or sulfur for a methylene group in aliphatic polyamides also results in a lower melting point. In the low molecular mass region, nylon melt temperatures increase significantly with increasing molecular mass. This effect, however, can be neglected at the high molecular masses necessary for fiber production. The strength of nylon fibers also increases with increased molecular mass, but melt spinning has technological limits because of the difficulty of handling very high melt viscosities.

The melting temperature of nylon decreases with increased ratio of methylene groups to carbonamide groups, and approaches the melting temperature of polyethylene (*mp* 130 – 150 °C, depending on molecular mass). Thus, nylon 22 melts at 145 °C. However, melting temperatures of the two types of nylon do not decrease uniformly, because nylons in the homologous series with an even number of methylene groups between carbonamide groups melt at higher temperatures than the adjacent nylons with an odd number of methylene groups (Tables 3 and 5). This is because the geometrical symmetry of the even-numbered configuration permits tighter chain packing and better hydrogen bonding. Wa-

Table 2. Comparison of melting points for various polyamides

Diamine	Diacid	CAS registry no.∗	*mp*, °C
m-Xylylene-α,α'-diamine	adipic	[*25805-74-7*]	243
1,6-Diaminohexane	adipic	[*9011-55-6*]	265
p-Xylylene-α,α'-diamine	adipic	[*24938-72-5*]	333
1,6-Diaminohexane	terephthalic	[*24938-03-2*]	370
p-Xylylene-α,α'-diamine	terephthalic	[*24938-64-5*]	> 400

∗ For the polyamide from the diamine and diacid.

Table 3. Physical properties of some aminocarboxylic acid type polyamides

Polyamide	*mp*, °C	Water absorption, %		Wet strength as a	ϱ, g/cm^3
		65 % R. H.	100 % R. H.	% of dry strength	
4	265	9.1	28.0	70 – 80	
6	223	4.3	9.5	85 – 90	1.14
7	233	2.8	5.0		1.10
8	200	1.7	4.0	93 – 97	1.08
9	209	1.5	3.3		1.06
10	188	1.4	3.0		1.04
11	190	1.3	2.8		1.04
12	179	1.3	2.7	100	

ter absorption and nylon density also decrease with increased number of methylene groups, i.e., as the molecule becomes more alkane-like (Table 3).

1.2.1. Aminocarboxylic Acid Type Polyamides

Nylons of the aminocarboxylic type can be prepared by the following methods:

1) polycondensation of ω-aminocarboxylic acids having more than five carbon atoms
2) hydrolytic polymerization of lactams having more than six ring members, and
3) anionic polymerization of lactams with exclusion of water.

Tables 1 and 3 list the most important *physical properties* of this type. Fibers are characterized by a high tenacity of 6 – 7 g/dtex and a break elongation of 20 – 30 %. An exception is nylon 4, with a tenacity of ca. 4 g/dtex.

Nylon 1 [*32010-01-8*]. Nylon 1 falls into this class by virtue of the linear connection of carbonamide groups; it is a polyamide based on its chemical structure [8]:
−CO−NH−CO−NH−CO−NH−.
However, derivatives prepared from monoisocyanates have no polyamide properties.

Nylon 2 [*25734-27-4*]. The polyamide from aminoacetic acid is really the first member of the series defined by the general formula $H[HN-(CH_2)_n-CO]_xOH$ ($n = 1$). So far, preparing nylon 2 with sufficiently high molecular mass for the production of fibers has not been possible. A high order of crystallization would be expected, and spinning would only be possible by a solvent spinning process. Filaments of nylon 2 could resemble natural silk very closely in physical properties.

High molecular mass nylon 2 derivatives alkylated on the α-carbon atom may be also viewed as polypeptides of α-aminocarboxylic acids and can be prepared by ionic polymerization of *N*-carboxylic acid anhydrides:

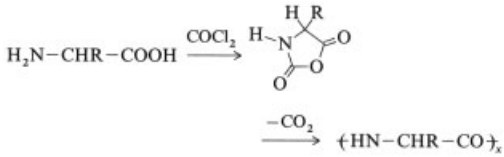

Polymerization can be initiated with a variety of catalysts such as alkali hydroxides and carbonates, phenols, and amines. Spinnable polymers have been obtained only from pure optical isomers, i.e., L- or D-aminocarboxylic acid

derivatives. The Japanese have produced semi-works quantities of polypeptide fibers based on L-glutamic ω-ester, which are reported to have unusual properties.

Nylon 3 [*25513-34-2*]. Nylon 3 would be expected to have good fiber properties, but no economically attractive route to its preparation has been found. It melts at 330 °C and should be spinnable to fibers from superheated aqueous solutions [9].

C-Alkyl derivatives of nylon 3 are of interest for fibers because of new syntheses of corresponding β-lactams [10].

Chlorocyanogen and sulfur trioxide react to form the easily accessible *N*-carbonylsulfamidic acid chloride. Under mild conditions, this reacts with olefins to form the β-lactam-*N*-sulfonyl chloride, which can be converted to the β-lactam through hydrolysis in a weakly acidic medium:

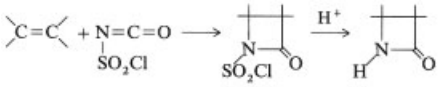

Because both the β-lactams and the corresponding polyamides generally decompose above 250 °C, these nylons are prepared preferably through anionic polymerization in solution, at temperatures of − 15 to 30 °C. For copolymerization, the β-lactams must have reasonably similar polymerization rates. Polymerizations are anhydrous with alkaline catalysts, e.g., alkali lactams or N-alkyl lactams. In contrast to melt-spun polyamides in which molecular masses of 15 000 to 25 000 are adequate for good fiber properties, molecular masses of > 200 000 are required for β-lactams, where fibers are made by solution spinning. Fibers with good textile properties and remarkedly good oxidative stability can be prepared from poly(β,β-dimethyl-β-propriolactam), spun from a ca. 15 % solution. Typical solvents are trifluoroethanol, phenol, and dimethyl sulfoxide.

Nylon 4 [*24938-56-5*]. The polycondensate of 4-aminobutyric acid is prepared by anionic polymerization of γ-butyrolactam in the presence of alkaline catalysts at < 60 °C.

The reaction is carried out in an inert hydrocarbon solvent, and conversion is ca. 80–85 %. Precipitated polymer must be washed thoroughly to remove residual monomer, traces of which accelerate decomposition of the polymer at the melting point. Nylon 4 can be melt-spun at 279–280 °C, but residence time in the melt must be short to avoid decomposition or reversion to γ-butyrolactam. Nylon 4 can also be spun from a 45 % formic acid solution, but lower tenacity is obtained. Because of its relatively high water absorption (9 % at 65 % R.H.), nylon 4 can be dyed easily, but it has relatively poor crease resistance. The fibers have not been significant economically, in part because of their poor temperature sensitivity in spinning [11].

Nylon 6 [*25038-54-4*]. While not ideal for optimum fiber properties, nylon 6 represents about half of the commercial production of nylon fibers and occupies a transitional position in this series. Whereas 4-aminobutyric and 5-aminovaleric acid proceed smoothly to the corresponding lactams on heating above melt temperature, 6-aminocaproic acid forms a temperature-dependent equilibrium condensate. At 250 °C, this consists of ca. 89 % linear polyamide, 8.5 % caprolactam, and 2.5 % larger ring amides. This equilibrium condensate is identical with the equilibrium polymer that results from hydrolytic polymerization of ε-caprolactam.

Problem-free continuous polymerization of caprolactam (→ Caprolactam) can be carried out in perpendicular flow-through tubes, with small additions of water, at 240–270 °C and atmospheric pressure. Batch polymerization in pressure autoclaves can also be carried out, which is a preferred technique for small production units or for polymer with special additives. Anhydrous caprolactam can also be polymerized with alkaline catalysts or cocatalysts, but this "fast polymerization" is not used in the manufacture of fibers. Excess caprolactam and higher ring amides must be removed from the equilibrium mixture before further processing to fibers. The equilibrium melt is extruded, solidified, cut to granules or flake, washed repeatedly with boiling water, and then vacuum dried. This can be done batchwise or continuously. Careful exclusion of oxygen is essential in all processing steps. Nylon 6 can also be polymerized and spun directly by passing the equilibrium polymer through a continuous, thin-film evaporator to remove impurities and then pumping it

directly to the spinning machine [11, pp. 121–134].

Because of its relatively low melting point and high water absorption, and the lack of a technically simple way to convert monomer to a continuously spinnable melt, nylon 6 is not the best aminocarboxylic acid type for producing synthetic fibers. The reason for its great importance and large production volume is economic. The high-purity caprolactam required for fiber production is made from favorably priced raw materials (benzene, cyclohexane, or phenol).

Nylon 7 [*25035-01-2*]. Nylon 7 is generally considered to be the optimum aminocarboxylic acid type polyamide for fibers. In both the condensation of 7-aminoheptanoic acid and the hydrolytic polymerization of enantholactam, yields of linear polymer exceed 99 %. This means that nylon 7 can be produced continuously in a technically simple polymerization-spinning process. Textile properties of nylon 7 fibers are superior to those of nylon 6 and, to a degree, those of nylon 66. Of particular interest, compared to nylon 6, are the 10 °C higher melting temperature and reduced water absorption. These two properties lead to important advantages, i.e., a considerably higher wet modulus and improved crease recovery (Table 4).

To date, no economically competitive route for production of the monomer has been reported. Small quantities of nylon 7 are produced in the former States of USSR. For this, 7-aminoheptanoic acid is reportedly prepared by telomerization of carbon tetrachloride and ethylene, via the intermediate 7-chloroheptenoic acid.

Nylon 8 [*25035-02-3*]. Nylon 8, which melts at 200 °C, is prepared by hydrolytic polymerization of caprolactam. The equilibrium polymerizate contains only traces of ring amide, so that fibers can be produced by using a continuous polymerization-spinning process. Because of the reduced number of carbonamide groups, water absorption is low, which provides relatively good wet strength and wet elasticity. However, raw materials for the synthesis of caprolactam are too expensive for this polymer to be attractive economically.

Nylon 11 [*25035-04-5*]. The synthesis of nylon 11 is based on a five-step process from castor oil to 11-aminoundecanoic acid. A 30–40 % aqueous dispersion of the acid can be polycondensed in three steps, and the nylon 11 formed can be isolated and melt-spun. Addition of phosphoric or phosphorus acid facilitates spinning by raising the melt viscosity and limiting side-chain growth. As expected from the longer hydrocarbon chain, nylon 11 shows hardly any sensitivity or dimensional change in the presence of moisture. The fibers are too expensive for general use and are restricted to a few special applications in luggage, ropes, transmission belts, and bristles.

Nylon 12 [*24937-16-4*]. Nylon 12 is prepared by hydrolytic polymerization of lauryl lactam, made by a four-step process from butadiene. The reaction to give polymer is essentially complete, so spun products can be prepared by using a continuous polymerization-spinning process. Because nylon 12 has good thermal stability, polymerization temperatures can be raised to 310–340 °C [13]. Except for a 10 °C lower melting point, fiber properties of nylon 12 closely correspond to those of nylon 11. These fibers are not made in large volume because of cost, and their use is restricted mainly to industrial specialty areas.

1.2.2. Diamine–Dicarboxylic Acid Type Polyamides

Nylons of this group are prepared for fiber use primarily by thermal polycondensation of equimolar amounts of diamines and dicarboxylic acids or their salts. Monomers that can undergo intramolecular condensation, like glutaric and succinic acids, are not suitable for thermal polycondensation. In commercial practice, aqueous solutions or suspensions of the salt are processed in batchwise or continuous operation. Polycondensation is induced initially by treatment of the concentrated salt under pressure to avoid loss of diamine, and then under vacuum to increase molecular mass to the degree required for fiber production.

This type of nylon can also be prepared from diamines and dicarboxylic acid dichlorides through interfacial polymerization. Some nylons, with melting temperatures in a range suit-

Table 4. Physical and textile properties of high-strength filaments from nylons 6, 7, and 66 [12]

Property	Nylon 6	Nylon 7	Nylon 66
Tenacity, cN/dtex	8.0 – 8.5	8.1 – 8.5	8.2 – 8.7
Wet strength, % of dry	81	92	85
Elongation at break, %	24	14 – 16	18
Elastic modulus, cN/dtex∗	41	46 – 49	54
Wet modulus, % of dry∗∗	27	62	31
Boil-off shrinkage, %	12	8 – 10	10
ϱ, g/cm^3	1.14	1.10	1.14
mp, °C	214 – 220	225 – 233	255 – 260
Water absorption, %			
65 % R.H.	4.3	2.8	4 – 4.5
100 % R.H.	9 – 10	5	9 – 10

∗ At 1 % elongation.
∗∗ In 25 °C water.

Table 5. Melting temperature of nylons

Diamine	Dicarboxylic acid	Nylon	*mp*, °C
1,4-Diaminobutane	adipic	46	278/295
	pimelic	47	233
	suberic	48	250
	azelaic	49	233
	sebacic	410	239
1,5-Diaminopentane	glutaric	55	198
	adipic	56	223
	pimelic	57	183
	suberic	58	202
	azelaic	59	179
	sebacic	510	185 – 195
1,6-Diaminohexane	adipic	66	250 – 265
	pimelic	67	202 – 228
	suberic	68	220 – 232
	azelaic	69	185 – 226
	sebacic	610	209 – 221
1,7-Diaminoheptane	adipic	76	226 – 250
	pimelic	77	196 – 214
	sebacic	710	187 – 208
1,8-Diaminooctane	adipic	86	235 – 250
	suberic	88	205 – 225
	sebacic	810	197 – 210
1,9-Diaminononane	adipic	96	205
	azelaic	99	165
	sebacic	910	179
1,10 -Diaminodecane	adipic	106	230 – 236
	suberic	108	208 – 217
	sebacic	1010	194 – 203

able for melt spinning, 180 – 280 °C (Table 5), are obtained by combination of aliphatic diamines and dicarboxylic acids. However, only nylon 66 has any economic significance.

Nylon 46 [*24936-71-8*]. The fiber from 1,4-diaminobutane and adipic acid is claimed to have better high-temperature properties than nylon 66 [14]. Good strength retention and low shrinkage suggest industrial applications.

Nylon 66 [*32131-17-2*]. Because raw materials are favorably priced for producing both adipic acid (cyclohexane) and 1,6-diaminohexane (acrylonitrile, butadiene, and adipic acid), nylon 66 has outstanding economic importance in the production of synthetic fibers (→ Adipic Acid; → Cyclohexane).

For the production of fiber-grade polymer, a 50 – 60 % aqueous nylon salt solution is heated in a pressure autoclave, with careful exclusion of oxygen. The pressure is released slowly, and the

resultant low molecular mass polymer is heated under vacuum in the same vessel to increase molecular mass and melt viscosity to the level required for fiber spinning. Polycondensation can also be carried out continuously [15] and the resultant melt pumped directly to the spinning machine.

Nylon 610 [*9008-66-6*]. Nylon 610 is prepared under conditions similar to those for nylon 66. Because of low water absorption and high elasticity, nylon 610 is especially suitable for production of bristles. The cost of sebacic acid, produced from castor oil, is fairly high and limits the use of this polymer.

Nylons Containing Aliphatic or Aromatic Rings. There is currently no significant commercial production of linear polyamides containing aliphatic or aromatic ring structures. Many have been investigated for use in tires for improved flat spotting, for better carpet aesthetics, and for improved wrinkle resistance and tactile aesthetics in nylon apparel fabrics. Systematic studies have been reported on fibers or films using phenylenediamine, $C_6H_4(NH_2)_2$, and C_4 to C_{12} diacid chlorides [16], and on the relative effects of cyclohexylene and phenylene linkages in polyamide fibers prepared from C_6 to C_{12} aliphatic dicarboxylic acids [17].

One ring-containing nylon fiber, T-472 [*26403-85-0*], was marketed commercially in the early 1970s by Du Pont. It was based on bis(4-aminocyclohexyl)methane and dodecanoic acid.

The diamine is obtained by catalytic hydrogenation of 4,4′-diaminodiphenylmethane (made from aniline and formaldehyde), resulting in a mixture of cis, trans isomers. The mixture must be fractionally crystallized to yield the preferred 70 % trans, trans product [18]. This isomeric mixture is polycondensed with dodecanoic acid to yield a melt-spinnable polymer (*mp* 275 °C) and fibers with high tenacity and resilience, and low water sensitivity.

The luxury, silklike fabric made from T-472 fiber was marketed under the registered trade name Qiana. The product was withdrawn from the market in the early 1980s for economic reasons.

1.2.3. Copolyamides

Nylon copolymers are formed easily by polycondensation of monomer mixtures. Compared to the corresponding pure polyamides, they have lower melt temperatures and higher solubilities. In the case of copolyamides of 6-aminocaproic acid and 1,5-hexanediammonium adipate, the melt temperature minimum is ca. 160 °C at 60 wt % of the amino acid. The melting temperature of the copolyamides can be reduced still further through ternary monomer mixtures. Low-melting copolyamides are suitable, for example, for thermal bonding of needle felts and nonwoven fabrics. Copolyamides of ε-caprolactam and 11-aminoundecanoic acid are especially suitable for fishing lines because of good flexibility and increased transparency.

1.3. Special Requirements for Polyamides

The *purity* of all nylon starting materials is especially important because reactive impurities result in uncontrolled chain growth and can reduce the quality of the spun product, e.g., by increased yellowing tendency. Mechanical impurities must be removed carefully by filtration before polymerization. The polymer melt is also filtered to remove mechanical impurities and pigment agglomerates, by passing it through a bed of sand or powdered metal and fine screens, just before it is extruded in filament form. This is especially important in the production of fine-denier filaments in which coarse impurities can reduce quality and yield through filament breaks.

Because nylons are particularly sensitive to oxidation at elevated temperature, *oxygen* must be scrupulously excluded in both the polymerization and the spinning steps. The oxygen content of the inert gases used, primarily carbon dioxide and nitrogen, should be maintained below 0.001 %. In both the hydrolytic polymerization of lactams and the polycondensation of aminocarboxylic acids, an equilibrium exists whose position is influenced substantially by the water content of the polyamide melt:

$$\mathrm{H[HN{-}(CH_2)_n{-}CO]_aOH + H[HN{-}(CH_2)_n{-}CO]_bOH} \rightleftharpoons \mathrm{H[HN{-}(CH_2)_n{-}CO]_{a+b}OH + H_2O}$$

Water removal leads to an increase in average molecular mass; water addition reduces average molecular mass. The same is true for nylons of the diamine – dicarboxylic acid type. *Viscosity stabilizers* are added to the monomeric starting materials to control the degree of polymerization and prevent an undue increase in viscosity in remelt. Both monofunctional and bifunctional compounds are used to limit chain growth. Especially suitable for this purpose are primary and secondary aliphatic amines or diamines, aliphatic and aromatic carboxylic acids, and dicarboxylic acids. Examples are cyclohexylamine, 1,6-diaminohexane, acetic acid, benzoic acid, and adipic acid. Trifunctional compounds are undesirable because they lead to molecular branching.

Good *analytical control* of polymer quality is essential to good spinning continuity and high fiber quality for all synthetic polymers. A narrow molecular mass distribution is desirable. Broad distributions can result from improper design or improper operation of a continuous polymerizer for nylon, and this frequently leads to inferior fiber properties.

Molecular mass distribution

can be determined by using the ultracentrifuge [19], by turbidity titration, by fractional precipitation or partition between two liquids, or by chromatography. From the molecular mass distribution, molecular mass can be calculated as the number average (M_n) or the mass average (M_m). The ratio M_m/M_n is always greater than 1 and is a numerical measure of molecular nonuniformity. The absolute value of M_n can be determined cryoscopically, osmotically, or through end-group analysis; M_m can be determined by light scattering [20] or by the hypsometric distribution of macromolecules in ultracentrifugal sedimentation.

The simplest and most commonly used method for obtaining relative values of mass-average molecular mass (M_m) is measurement of solution viscosity [19, pp. 49 – 67] in a capillary viscometer at constant temperature, where

$\eta_{relative}$ = flow time of the solution divided by flow time of the solvent

and

$\eta_{specific} = 1/\eta_{relative}$

Suitable solvents are concentrated sulfuric acid, formic acid, and phenol. The specific solution viscosity, $\eta_{specfic}/c$, referred to the polyamide concentration in grams per 100 mL, is concentration dependent. Plotting viscosity as a function of decreasing concentration and extrapolating to zero yield the intrinsic viscosity, which is directly proportional to M_m. In the production of nylon fibers, sufficient control of the process can be achieved by routinely measuring and controlling the viscosity of a standard solution, e.g., 1 g/100 mL. The melt viscosity, of practical importance for melt spinning, can be determined in rotary, torsional viscometers and capillary viscometers, or by using the falling ball method. In measuring melt viscosity, oxygen and moisture must be rigorously excluded and temperature held constant. On melting, polyamides can change by amide ring formation, molecular mass increase, and degradation, so melt viscosity is likely to vary according to method and measuring conditions. In general, the melt viscosity of polyamides for melt spinning is ca. 50 – 500 Pa · s.

The *free amino and carboxyl end groups* of polyamides are determined by potentiometric or conductometric titration with dilute perchloric acid and alkali hydroxide solution, respectively. A solvent mixture of phenol, ethanol, and water is suitable for amino-group titration. A supersaturated polyamide solution, in benzyl alcohol or α-phenylethyl alcohol, is used for carboxyl end-group determination. This analytical procedure cannot be used for all polyamides.

Water analysis is carried out by heating the flake or fiber, and released water vapor is carried with dry nitrogen into Karl Fischer reagent. In this method, both the physically bound water and that formed by chemical reaction are determined.

Extractable materials in nylon 6, caprolactam, and higher ring amides can be determined by extracting the finely divided polymer for several hours in water or lower alcohols. However, in this relatively simple method, the high ring amides are not extracted completely. For quantitative determination and separation of ring amides, the polymer is dissolved in formic acid and the linear polyamide molecules are then precipitated by addition of methanol. Ring amides can be determined quantitatively in a methanol solution by using gel chromatography.

1.4. Modifying Additions

Nylons that absorb water and swell, such as nylons 6 and 66, can be dyed relatively easily with *disperse, acid, or metallized dyes*. The dye molecules migrate readily into the fiber structure. The dyeing of nylon fibers with acid dyes results from reaction with the amino end groups. The number of such reactive sites is influenced by M_n but can be modified by additives such as viscosity stabilizers. For example, carboxylic stabilizers will block amino groups and reduce affinity for acid dyes. The use of primary aliphatic amines or diamines provides additional basic sites and increases acid dyeability. Introduction of tertiary amine groups into the chain strongly increases absorption of acid dyes.

Nylons can be modified to accept *basic dyes* by introduction of strong acidic groups such as sulfonic acids, e.g., 5-sulfoisophthalic acid, in equimolar mixture with 1,6-diaminohexane [21]. Polycolor effects (graduated color depth, variable color and shades) can be achieved by dyeing fabrics of mixed nylon fibers with different dyeabilities in a single dye bath containing acid, basic, and disperse dyes.

Delustering is the practice of altering the optical properties of a fiber, usually by incorporation of a white pigment. This is most often done for textile uses. The most common delusterant is titanium dioxide [*13463-67-7*], added in amounts of 0.03 – 2 %, depending on the effect desired. Of the two forms of titanium dioxide available, the anatase variety is preferred for continuous-filament yarns because the harder rutile modification causes excessive abrasion and wear on guide surfaces and machine parts over which the yarn runs. The rutile form can be used with staple fibers in which guide wear potential is not so serious. Other pigments, such as clay, can also be used. Delustering can also be accomplished by surface roughening or by introducing voids into the filament.

Size distribution [22] of titantium dioxide pigments and pigment aggregates is generally 0.1 – 1.0 µm, with a frequency maximum in the region of optimum turbidity at 0.2 – 0.4 µm. However, commercial pigments also contain larger agglomerates, which must be broken down through wet grinding or separated from the aqueous dispersion by fractional sedimentation. Pigment agglomeration must be avoided in pigment addition to the monomer or to the polymer melt, because the particles result in filter plugging as well as weak sections in the filament. For continuous production of nylon flake or fiber, adding a pigment – polymer concentrate of 20 – 40 % titanium dioxide to the melt just before spinning is advantageous. This permits quick conversion to different degrees of delustering, but also imposes strict requirements on control and mixing.

The magnitude of pigment addition can be checked by ash determination or photometric titanium analysis. The uniformity of pigment distribution, as well as the number and size distribution of pigment agglomerates, can be analyzed by microscopic examination of thin fiber cross sections.

Aging and light protection must be provided for some nylons by addition of special additives [23]. Manganese(II) compounds, in amounts as low as 5×10^{-4} %, are especially effective for nylons 6 and 66. Titanium dioxide pigments, especially of the anatase type, reduce light resistance, and one variety of anatase TiO_2 is coated with manganese to offset this effect. Many manganese compounds, such as the acetate, sulfate, phosphate, polyphosphate, and glycerophosphate salts have been proposed. In selecting the type and amount of additive, undesirable side effects, such as yellowing of the filaments and extraction of the manganese(II) during dyeing or wet finishing, must be considered in addition to the light-protective ability desired. Use of a manganese salt with a reducing anion, such as manganese(II) hypophosphite, improves the whiteness of fibers. Addition of sodium phenylphosphinate, in combination with manganese(II) hypophosphite, substantially improves both whiteness and quality [24].

The *heat-aging resistance* of nylons must be improved for uses such as the tire cord. The thermal – oxidative degradation of nylon can be reduced considerably by adding antioxidants (0.1 – 1.0 %), such as substituted aromatic diamines, and their action can be increased by combining them with trialkyl phenyl phosphites. Textile nylon, for which yellowing cannot be tolerated, requires special phenolic antioxidants substituted with alkyl radicals in the positions ortho to the – OH group. Examples are the amine salts of 3,5-dialkyl-4-hydroxyphenylcarboxylic acids. Copper(I) salts, in amounts as

low as 5×10^{-4} % copper, are also effective. Many synergistically acting stabilizer mixtures are described in the literature, e.g., complex compounds of copper(I) iodide and hydrocarbon phosphines or phosphites are said to be especially effective.

Spin Dyeing and Optical Brightening. Inorganic and organic pigments for spin coloration of nylons must be stable under melt processing conditions and must give high colorfastness [25]. Pigment particle size should be less than 1 – 2 µm to avoid impairment of fiber strength, and agglomerates must be carefully milled or separated by sedimentation. Colored pigments are generally added to the starting monomer as aqueous dispersions. They may, however, be added to the melt as pigment – polymer concentrates [26]. This technique is desirable for short commercial runs but requires precise metering and good mixing.

Suitable additives for *spun-dyed nylon* are the yellow to red cadmium pigments, anhydrous iron oxide pigments, ultramarine, copper phthalocyanine, perylene, and quinacridone pigments. Because nylon melts generally have a strong reducing action, perylene pigments can be reduced on the surface with a significant change in color. However, reoxidation often occurs in air shortly after extrusion of the filament. Some azo and anthraquinone dyes are also soluble in nylon melts and are sufficiently stable for spin dyeing.

Dyeing of nylon flake in aqueous baths, before spinning, is an economical way to produce a large range of colors from one flake type. However, the resulting colors do not have the outstanding fastness of pigment-spun fibers.

Black nylon fibers are obtained by using carbon black pigments with an average particle size of 10 – 30 µm. Carbon black pigment in too high a concentration can block carboxyl groups, reduce the degree of polymerization, and impair flow behavior of the melt.

Optical brighteners are frequently used to improve aesthetics of textile fibers, especially to hide the effects of soiling. The more stable brighteners are added to the monomeric starting material, the less stable ones, such as the coumarin type, are injected into the melt.

1.5. Technology of Melt Spinning

1.5.1. Melt Spinning of Nylon

All commercially significant nylons are melt-spun at temperatures up to ca. 300 °C [27]. Processes may be continuous from monomer to spun fiber, or they can be staged. When a variety of small-volume products is required, producing nylon flake and remelting it for the spinning process are economically advantageous. This is especially true for nylon 6, which must be washed to extract polycondensation byproducts.

In modern plants, flake is usually stored in large hoppers and then transported pneumatically or in smaller hoppers to the spinning area. Prior to spinning, the flake must be conditioned to proper moisture content because this can have a strong effect on melt viscosity. Great care is taken to store and transport flake with exclusion of moisture. Also, before being delivered to the melt unit, fines and excessively large cuts, as well as contaminating metal particles, are removed.

Nylon flake is commonly melted by using screw extruders or, in the older process, heated grids or grates. In either case, oxygen must be rigorously excluded. Spinning temperature depends mainly on the thermal stability and melt viscosity of the nylon. With nylon 6 and nylon 66, temperatures are usually kept at 280 – 290 °C. After leaving the melt unit, polymer is delivered to a metering pump and then to a spin pack. The pack contains filter media, passages for distributing polymer within the pack, and a spinneret. The filter section serves to remove mechanical impurities and to provide high shear for the melt. In a typical filter section, the molten polymer passes through layers of either powdered metal or fine sand, then through layers of very fine mesh screen, and finally through a photoetched plate before reaching the spinneret.

The spinneret is a metal plate, usually 6 – 10 mm thick, filled with holes, each hole forming one filament. Filaments from round spinneret holes have circular cross sections. The cross section can be altered to almost any desired filament shape to achieve unusual luster effects, as well as altered physical and tactile properties. Because the surface tension of the melt leads to a rounded filament, melt temperature, melt viscosity, and

cooling rate must all be adjusted to achieve the desired filament cross section.

Spinneret holes of 200 – 400 μm can be used to produce nylon filament and staple yarns. The molten polymer extruded from the hole is pulled away by downstream rolls, reducing the diameter 5- to 50-fold. The diameter, or the mass per unit length, of the spun filament is determined by the rate at which polymer is metered through the spinneret hole and the rate at which it is pulled away. Small particles that pass through the filter leave weak spots in the filament, which may break in subsequent processing.

Nylons up to ca. 30 dtex/filament can be spun into air, but higher diameters require spinning into water or other liquids. Except for a few specialty uses, such large diameter filaments are generally made for nontextile applications. Turbid monofilaments are usually obtained when spinning into water, because of crystallization and surface corrosion. Clear filaments can be prepared by spinning nylon melts into cooled organic liquids such as carbon tetrachloride.

Unusual and useful filament structures can be prepared with specially constructed spinnerets [28]. Two melts are fed simultaneously to a spinneret hole, which results in a bicomponent filament that crimps spontaneously after drawing and finishing. Suitable melt pairs for this are nylon 66 and a 66 – copolyamide [29] or a copolyamide 612 [30]. Spinnerets with two concentric holes yield sheath-core filaments from two melts. Injection of an inert gas into the center hole causes a void to develop and results in a hollow filament cross section.

Grid spinning. Grid spinning was the original process for melt spinning of nylon.

Flake is gravity-fed from an overhead hopper and melted on a spiral coil or grid heated with diphenyl. An atmosphere of superheated steam is generally provided to control melt equilibration. In starting up a grid-spinning head, melted polymer dropping from the melting grid collects in a conical sump, flows into a pressure pump, and is fed to a metering pump where it is metered through the filter pack and spinneret. Because pump demand must always be smaller than the maximum melt capacity, the sump fills until the melt grid is immersed in the melt, thus automatically regulating melting via the free-grid surface. When the process is interrupted, the grid heat must be shut off, otherwise molten polymer will rise into the cooler zone and solidify to form a bridge that blocks the flake feed. The melting capacity of a grid-spinning unit with a grid diameter of 160 – 200 mm is ca. 30 – 60 g/min. Higher capacity can be achieved with larger grids. One melting grid can feed four or more spinnerets.

Extrusion Spinning. More modern spinning equipment uses melt extruders instead of grid units.

Extruders provide a shorter dwell time in the melt, better mixing, and better viscosity control. Their main advantage, however, is the large melt capacity, to ca. 1500 kg/d for a 150-mm-diameter screw.

Screws with length – diameter ratios of 20 – 30 are preferred to assure a more uniform melt temperature through the longer compression zone. An additional mixing zone connected to the compression zone improves homogeneity of the melt and permits increased throughput [31].

The melt is fed at constant pressure from the extruder through a jacket-heated transfer line to the spinning machine. One melt extruder can supply 30 – 40 individual spinning positions, each of which might spin eight or more individual thread lines. Use of a central filter for the melt, located between the extruder and the spinning block, is considered advantageous for some processes. This type of filtration is practical for nylon 6 but not as effective for nylon 66 because of the tendency of the latter to degrade and form gels in, and after, the filter.

Melted polymer must be fed to the individual spinnerets with the same average dwell time. Distribution piping is designed in cascade form to provide the same melt travel distance to each spinneret. In addition, dwell time differences arising from parabolic velocity distribution of melt inside the pipe must be balanced with static mixers ahead of each branch. Local or chronological nonuniformity of the melt can usually be detected first as dyeing differences in filament yarns.

Direct Spinning. Economic and quality advantages result by feeding continuously polymerized melt directly to the spinning machine. This is most easily accomplished with nylon 66, which produces only small amounts of low molecular mass cyclic amides. Direct spinning

becomes most attractive when a few products can be made in high volume and such products can be made with little change in overall throughput. Large transitional changes are to be avoided. In such systems, introducing additives into the melt downstream of the polymerization unit is also desirable. This requires additional expenditures for the preparation of melt concentrates, as well as for the melting, metering, and homogenization equipment required for injection.

As noted earlier, nylon 6 fibers can also be made by direct spinning, but the equilibrium mixture, containing substantial amounts of caprolactam and higher cyclic amides, must be dealt with. This is accomplished through *thin-film evacuation.*

The melt, prepared continuously in a polymerization tube, is passed to an evacuating assembly at 270 °C [32]. It is distributed through an annular slit onto a cylindrical wall, as a coherent film of melt. As the melt flows down the wall, volatile components (caprolactam and water) evolve at a pressure of ca. 10 Pa. At the bottom of the assembly, the evacuated nylon 6 melt is homogenized through a mixing stirrer and conveyed by a compression screw through the discharge and to the spinning machine.

1.5.2. Special Techniques and Products

Fine-Denier Nylon Filament Yarns. Most fine-denier filament yarns are used in wearing apparel and household upholstery. Products are in the region of 10 – 400 dtex total, with filament counts from 1 to > 100. Single filaments run from 1.5 to ca. 6 dtex (9 – 36 µm).

Spinnerets for fine-denier textile filaments are usually made of stainless steel, 40 – 70 mm in diameter and 10 – 15 mm thick. Because of the high viscosity of the melt, the spinneret hole initially has a large diameter (2 – 3 mm), called the counterbore, which tapers in a conical section to a smaller exit hole (200 – 400 µm) from which the melt is extruded.

With a windup speed of 600 – 1200 m/min, solidification in the quench (cooling) chimney is completed within ca. 50 cm of the spinneret. Cool air is blown at right angles across the filament path, and careful attention must be given to avoiding pulsation or turbulence in air flow.

Filaments leaving the quench chimney are moistened with a spinning finish (oils emulsified in water) to provide static protection and lubrication for further processing. Many filaments are combined after finish application to form the final, multifilament bundle. If the yarn is to be drawn in a separate step, it is normally passed through a steam tube to promote development of the proper crystalline structure. For coupled spinning and drawing, yarn can be forwarded from the feed roll, which pulls the filaments away from the spinneret, directly to a set of faster moving draw rolls and then to a package-winding unit. Textile yarns are commonly drawn three to four times the original spun length to develop the required tensile – elongation properties. Depending on the process and the end use, packages of ca. 1 – 20 kg are produced.

Economics of the spinning operation have been improved over the years by winding multiple packages on one winding unit, by winding larger packages for each thread line, and by operating the entire process at higher speed. Filament yarns are now prepared at spinning speeds of 5000 m/min, and useful fiber properties are now achieved in some cases without a separate drawing step [33]. Spinning, drawing, and texturing may also be combined in one continuous operation.

Coarse-Denier Nylon Filament Yarns. Coarse nylon yarns range from 400 to > 35 000 dtex total and from 6 to ca. 20 dtex per filament.

Rectangular spinnerets are commonly used, spinning several hundred holes per spinneret. In the production of high-strength nylon yarns (e.g., tire cord), draw ratios of 5 – 6 are used and drawing is commonly done in two stages, with heating, to develop maximum tenacity and low elongation. Tenacities of 8.0 cN/dtex and 15 – 17 % elongation are common for nylon 66 tire yarns. Package weights of 40 kg are common, and ends from several hundreds of these packages are combined and wound on a larger spool (beam) for shipment to the customer.

Nylon filament yarns for carpets and other floor coverings are usually sold in the range of 1000 – 3000 dtex total and 15 dtex per filament. Nylon for netting and ropes reach 35 000 dtex total, with 20 dtex per filament. These yarns are drawn four to six times their original length with

heat assists and are deposited in rope form in a can. Some of these are then stretched further in a hot water bath to develop maximum tenacity.

Wires and Bristles of Nylon. Coarser nylon filaments (0.05 – 5 mm) are generally made with a higher degree of polymerization. Copolyamides are often more suitable for fishing lines.

Several processing steps are frequently coupled in one continuous operation. Flake is melted in a single-screw extruder, and the melt is extruded through spinneret holes (1 – 10 mm in diameter). The coarse filaments are solidified in a water bath and drawn between rollers in air or hot water at a speed of 200 – 400 m/min. Thicker wires may be drawn through dies to set the draw point and round the cross section. To improve dimensional stability and stiffness, coarse filaments for bristles are heat-set in boiling water, either continuously or on reels or bobbins.

1.6. Uses

Fine-denier filament yarns, from 1 to 6 dtex per filament and up to about 400 dtex total, are used broadly in wearing apparel, such as ladies underwear, lingerie, stockings, pantyhose, and outerwear; in shirting, elastic outer clothing, and recreational clothing; and in home upholstery and automotive decorative coverings. Heavy-denier filament yarns, 6 – 20 dtex per filament and up to 3000 dtex total, are used in carpets and other floor coverings, belt fabrics, tire cord, netting, and ropes. Nylon staple fibers, ranging from 1.0 to 15 dtex per filament, are used directly or blended with cotton, felt, and other synthetics for apparel, floor covering, and industrial uses. Monofilament nylons, 20 to > 1000 dtex per filament, are used in items such as fishing lines, luggage, paintbrush bristles, and cordage.

1.7. Economic Aspects, Manufacturers, and Trade Names

Some 3.9×10^6 t of nylon fiber was produced worldwide in 1988 (see Table 6), with a value of over \$ 11×10^9 [302]. Over 98 % of the total is either nylon 6 or nylon 66. Whereas the fraction of nylon 6 was only 9 % in 1953, vs. 91 % for nylon 66, worldwide shipments of the two are now about equal. Installed worldwide capacity, approaches 4×10^6 t. In 1985, the geographic distribution of production was 36 % North America, 23 % Western Europe, 17 % Asia (including Japan and Taiwan), and 24 % all other; sales by end use were 38 % for apparel and home furnishings, 34 % for carpets (mostly face fibers), and 28 % for industrial uses (truck tires, conveyor belting, etc.); an additional 1.0×10^6 t of nylon resin was produced for nonfiber applications. Major producers of nylon fibers are listed in Table 7 along with types and representative trade names.

Table 6. World production of nylon fibers∗

Year	Production, t	Production, %∗∗
1940	5 400	95
1950	62 000	90
1960	410 000	58
1970	1 908 000	38
1980	3 151 000	30
1985	3 396 000	27
1988	3 900 000	

∗ Data compiled from *Textile Organon.*
∗∗ Percentage of world production of synthetic fibers.

Table 7. Major producers and trade names of nylon fibers∗

Manufacturer	Nylon	Trade names
Du Pont (United States)	66	Antron, Cordura, Cantrece
BASF (FRG)	6	Zeftron
Rhône-Poulenc (France)	66	Nylon
Allied Signal (United States)	6	Caprolan, Anso, Captiva
Monsanto (United States)	66	Nailon R, Cadon, Cumuloft
	6	Helion, Polycord
ICI (United Kingdom)	66	Bri-Nylon
BASF	6	Zeftron
Bayer (FRG)	6	Bayer-Perlon, Dorix
Courtaulds (France)	6	Celon
Enka-Glanstoff (FRG)	6	Enka-Perlon, Enkalon, Enkatron
Hoechst (FRG)	6	Perlon Hoechst
Snia Viscosa (Italy)	6	Lilion
Teijin (Japan)	66	Teijin-Nylon
Toray (Japan)	6	Amilan
	66	Toray-Nylon
Visco-Suisse (Switzerland)	6	Flixor
	66	Edlon

∗ The first six are listed in order of sales volume; the remainder are alphabetical.

2. Polyester Fibers

History. The first spinnable high molecular mass polyesters were prepared by W. H. Carothers and co-workers in the laboratories of E. I. du Pont de Nemours & Co. around 1930 [34]. Dicarboxylic acids in use then were sebacic acid, hexadecanedicarboxylic acid, and others; some diols used were ethylene glycol and propylene glycol. The discovery that hardened spun strands could be drawn to several times their original length and, thereby strengthened, was made with these polyesters [35]. However, their melting points were ca. 75 °C, too low for practical application. For this reason, Carothers turned his attention to the more promising polyamides [36].

In the late 1930s, investigations showed that terephthalic acid and other aromatic acids could yield useful polyamides. Therefore, polyesters containing terephthalic acid as the dicarboxylic acid were studied at several locations [37]. Poly(ethylene terephthalate) was discovered in the laboratories of Calico Printers Association. Beginning in 1940, Whinfield and Dickson devoted themselves mainly to poly(ethylene terephthalate) [38], [39], while Schlack was working on poly(tetramethylene terephthalate) [40].

In 1947, ICI and Du Pont acquired patent rights for Great Britain and the United States, respectively, and initiated the commercial development of polyester fibers [37]. In the early 1960s, ICI licensed its technology worldwide to some large fiber producers whose own developments led to today's mass production. Thus, poly(ethylene terephthalate) has gradually become the most important fiber polymer [41].

According to DIN definition 60 001, fibers are designated polyester if the polymer chain contains at least 85 % of an ester of a diol and terephthalic acid. Poly(ethylene terephthalate) [*25038-59-9*] (PETP), poly(oxy-1,2-ethanediyloxycarbonyl-1,4-phenylenecarbonyl), is a polyester with terephthalic acid [*100-21-0*] as the dicarboxylic acid and ethylene glycol as the diol ($n = 80 - 150$):

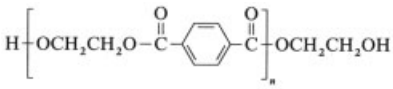

For a detailed description of the chemistry, raw materials, production, and analysis of polyesters, see → Polyesters. The following discussion contains information pertinent to an understanding of the production and properties of these fibers. Section 2.1 covers aspects of production common to all types of PETP fibers; specific information can be found under the individual fibers (Sections 2.2 and 2.3).

2.1. Production

2.1.1. Starting Materials and Polymer

Ethylene Glycol [*107-21-1*]. Ethylene glycol is produced commercially by the reaction of ethylene oxide with water, which in turn is obtained by the oxidation of ethylene (→ Ethylene Glycol; → Ethylene). Prerequisites for smooth polycondensation to poly (ethylene terephthalate) and for the desired polymer properties are extreme purity of ethylene glycol and other raw materials and the absence of water.

Dimethyl Terephthalate [*120-61-6*] (**DMT**). Until the end of the 1960s, economical production of terephthalic acid in the purity required for fibers was difficult. For that reason, the more easily available dimethyl ester of terephthalic acid was generally chosen as the starting component for PETP (→ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid).

This material is produced commercially primarily by the *Witten process* (identical to the Imhausen or Katzschmann process [42]).

p-Xylene [*106-42-3*] is subjected to a two-step, liquid-phase oxidation in air. The *p*-toluic acid so produced is esterified with methanol to the methyl ester of *p*-toluic acid, which is then esterified with methanol to DMT.

Terephthalic Acid [*100-21-0*] (**TPA**) (→ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid).

Since the early 1970s, extra pure terephthalic acid of fiber quality has been produced commercially primarily by the *Amoco process* [43].

Inhibition of the oxidation of the second methyl group of *p*-xylene is suppressed with the aid of added bromine-containing promoters as

cocatalysts. As with DMT, the oxidation takes place in air and produces raw terephthalic acid, which is dissolved at high temperature under pressure in water, hydrated, and thus purified.

Polymer. The polymer is prepared in two steps:

1) Transesterification of dimethyl terephthalate (**1**) with ethylene glycol to bisglycol terephthalate (**2**) (continuous or batchwise) or direct esterification of terephthalic acid.
2) Polycondensation (continuous or batchwise) to give poly(ethylene terephthalate) (**3**).

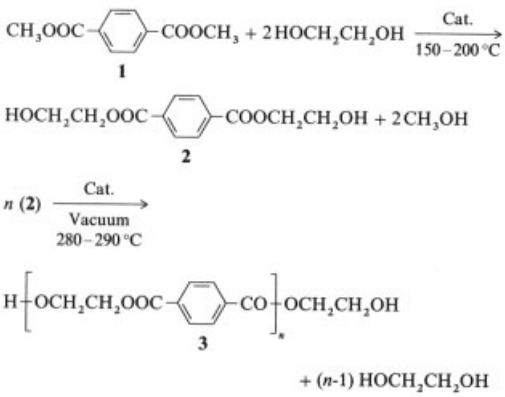

The batch process is preferred for special polymer types, particularly for small amounts of polymer. The continuous process is used for standard polymer types in large volume, especially if spinning immediately follows polycondensation.

Poly(ethylene terephthalate) from Dimethyl Terephthalate. Commercial transesterification of dimethyl terephthalate with subsequent polycondensation was initially carried out batchwise.

Batchwise Process. First, DMT is esterified as quantitatively as sible with an excess of ethylene glycol (molar ratio range of DMT to ethylene glycol 1 : 2.5 to 1 : 4.5) between 150 and 180 °C. Catalysts are primarily zinc, calcium, and manganese acetates. Bisglycol terephthalate, as well as higher homologues, are formed. Excess glycol is distilled at temperatures up to 250 °C.

Polycondensation is carried out at higher temperatures, up to ca. 285 °C, under vacuum. In this step, the catalysts are primarily antimony trioxide or oxides of similar amphoteric metals. Phosphorus compounds are added as stabilizers to inactivate the transesterification catalyst.

Other additives, e.g., titanium dioxide for delustering or carbon black for black coloration, can be added to the melt prior to polycondensation. Once polycondensation is complete, the condensate is forced out of the vessel with nitrogen and quenched in water. The melt, which is thus solidified largely in the amorphous state, can then be formed into chips.

Continuous Process. Continuous transesterification and polycondensation occur in several, mostly horizontal, reactors in a manner similar to the batchwise process.

As the polyester melt is formed during condensation, care is taken to keep its layer thickness small and its surface correspondingly large. This is necessary because diffusion of excess glycol determines the rate of polycondensation. Discharge from the last reactor occurs by means of an extruder or a pump. The polyester melt is then either passed to the spinning units and spun immediately or chilled in water and cut into chips. Continuous processing units are generally quite large, with throughputs of up to 100 t/d and average holdup times of the material in the unit of 5 – 10 h.

Poly(ethylene terephthalate) from Terephthalic Acid. Since highly purified, fiber-grade terephthalic acid has become available at low cost, it is increasingly being esterified directly with ethylene glycol (molar ratio of terephthalic acid to ethylene glycol 1 : 1.1 to 1 : 1.5).

The reaction occurs at about 260 °C under pressure and generally requires no catalyst. Lower linear oligomers are formed. After the glycol has been distilled, polycondensation follows. Its course is similar to that with DMT as the starting product. A polycondensation catalyst is also required.

The economic advantages of this route are that circulation of methanol is eliminated, substantially less excess glycol is necessary, and above all, the amount (in grams) of starting material required is less by ca. 14 % because of the lower molecular mass of TPA vs. DMT. However, direct esterification results in somewhat more diethylene glycol byproduct. This process is used in continuously operating modern plants.

Poly(ethylene terephthalate) with Very High Molecular Mass. Industrial end uses require very high molecular masses (corresponding to very high average lengths of molecules) to attain the necessary tenacity and modulus. Because removal of excess glycol in polycondensation is rate determining, reducing the thickness of the melt layer as the reaction progresses would be appropriate. However, due to the melt viscosity increase during polycondensation, this becomes more and more difficult. Holdup time at high melt temperature is thus increased, which produces thermal degradation in parallel with polymerization. Polymer whiteness thus deteriorates, and COOH end groups are formed in undesirable concentrations in addition to the OH end groups. To get around these problems, PETP of very high molecular mass is now generally produced by *postcondensation in the solid phase* (solid condensation).

Granulate PETP, produced by melt condensation, is dried and exposed to heat either batchwise in vacuum or continuously under an anhydrous gaseous protective blanket free of oxygen (generally nitrogen) at temperatures up to ca. 250 °C (i.e., somewhat below the melting point). A lamellar structure with noncrystalline regions and voids forms (→ Fibers, 2. Structure). Polycondensation occurs in parallel with recrystallization and additional crystallization. The glycol – water mixture produced diffuses through the noncrystalline regions and mainly through the interlamellar voids that are being formed. The result, in addition to acceptable polymer color, is primarily the desired low concentration of COOH end groups. Once the desired average molecular mass has been attained, the granulate is chilled and packaged. Alternatively, it can be melted directly with the exclusion of water and sent to the spinning units.

2.1.2. Spinning

The melting point of poly(ethylene terephthalate) (265 °C) is clearly below the decomposition temperature of the melt; therefore, PETP can be melt-spun (see also → Fibers, 3. General Production Technology, Chap. 3.1.). During melt spinning, a molecular orientation is created in the filaments by the drawdown tension; this orientation is frozen during cooling. Here, PETP exhibits a special characteristic which distinguishes it from other commercial melt-spinnable polymers: because of its relatively high glass transition temperature (70 °C), the spun material solidifies completely in the spinning shaft to form a stable, supercooled melt with molecular orientation more or less in the yarn direction [37]. This orientation, measured by birefringence, is proportional to the tensile stress along the path of the yarn at the time and place of solidification, [44], [45].

As filaments solidify, they also attain their final speed (winding speed) and thus their final linear density (as-spun fineness). Where the solidification point lies along the path of the as-spun yarn, depends on spinning conditions; distances between 20 and 150 cm from the spinneret have been reported [45].

Because of the relatively low tendency of PETP to crystallize at room temperature, the as-spun yarn is amorphous as long as drawdown tensions are not too great. Stress-induced crystallization occurs only at higher tensile stress as a result of spinning speeds above ca. 3000 – 4000 m/min or of very intense air quench; this crystallinity also is frozen into the yarn. Under these processing conditions, the as-spun yarn is partially crystalline [46], [47].

At very high spinning speeds (at about 8000 m/min), the crystallinity of the as-spun yarn can be as high as 50 % [48], [49], and the distribution of directions of the crystallites becomes even more narrow. At the same time, the layered structure of crystalline order, which resembles that of PETP spun slowly and then drawn, is slowly transformed into a more fibrillar structure of larger ordered regions. Despite the high orientation of the crystalline fractions, the total orientation again decreases somewhat; therefore, orientation of the amorphous material between the fibrils is assumed to decrease [50].

At spinning speeds > 4000 m/min, an additional sheath-core orientation occurs in the filaments [48]. This effect can cause the orientation in the sheath, measured by birefringence, to be almost twice as great as in the filament core [51], [52]. Correspondingly, the sheath of the filaments is also more highly crystalline. These differences remain during further processing of the as-spun yarn and can lead to nonuniformities in dyeing if they are not controlled in the process [53].

As mentioned, the orientation and even the crystallinity of filaments depend on spinning conditions and increase with increasing spinning speed, stronger air quench, and higher molecular mass. Both characteristics decrease with increasing spinning temperature, with higher mass transport, and thus with greater as-spun linear density. If the spinning speed is increased and the fineness is held constant by a proportional increase in mass transport, the effect of spinning speed predominates. The orientation and even the crystallinity increase [45][46][47]. The diameter of the spinneret holes, hence the spin–draw ratio, has relatively little effect on the final spin orientation and crystallinity except when spinning speeds are $\geq$5000 m/min [50].

Orientation and crystallinity of the filaments are characteristics of the as-spun yarn that must be considered in further processing because of their effect on the properties of the end product. Both factors are held as constant as possible with respect to time and between the spinning units.

Spinning Process. As described in Section 2.1, PETP is produced either continuously or batchwise. In the continuous process, the spinning and polymerization units can be connected. Thus, freshly condensed melt flows directly to the spinning units. Because the economic advantages of direct spinning are obvious, its use for high-volume end products, e.g., for staple fibers and standard filament yarns, which are always made by the same process, is increasing constantly [84].

The PETP granulate that is delivered to a spinning plant is first stored in silos. The granulate must be dried immediately before melting because even the smallest amount of water in the polymer causes hydrolytic degradation of the melt. Residual water content < 0.01 % is desirable.

Drying is carried out mostly in continuous shaft dryers. The granulate moves slowly downward, while a hot, often predried, countercurrent air stream blows over it. Since amorphous PETP granulate tends to become sticky during heating, a crystallizer in which the chips are slightly crystallized, at least on the surface, often precedes the dryer [55]. Many dryer designs are known, with throughputs amounting to several tons per hour. For some products, drying the PETP granulate in a vacuum is better, for example, in batches in large tumble dryers.

From the dryer, the granulate reaches the melt extruder, which transports the melt to one or more spinning machines. Because of thermal degradation of the melt, maintaining equal holdup times in the hot lines between the extruder and the attached spinning units is important; the temperature of these lines must also be the same. Only then can the uniformity of the as-spun yarn and the end products be assured.

Spinning units differ considerably in construction and processing data, depending on the end product. The same is true for the method of cooling, finishing, drawing down, and collecting the as-spun yarn [37]. Common operational winding speed ranges from ca. 500 m/min for industrial yarns to 6500 m/min for filament yarns spun at the highest speed. Thick filaments, e.g., monofilaments, are spun at a much lower speed. Common to all is the spinning temperature, which is 270–310 °C. For details, see Section 2.3.

2.1.3. Drawing

The glass transition temperature of PETP is ca. 70 °C, which is somewhat above room temperature. Therefore, when PETP yarn spun at relatively low speed is elongated slowly, cold drawing occurs with a draw point [54]. From the linear densities on both sides of the draw point or from the length of the plateau of the stress–strain diagram (see Section 2.2.2), a "natural draw ratio" can be specified. This ratio depends on the orientation of the as-spun yarn, which already represents a part of the total orientation [56]. The product of the orienting spin–draw and the natural draw ratio is constant to a first approximation and is equal to ca. 6–7 [57].

In commercial practice, as-spun PETP yarn is always drawn at a temperature $\geq$ 80 °C. Because of the added work of drawing, the real yarn temperature in the draw zone is frequently higher than the external temperature, depending on the drawing arrangement.

During drawing [44], the molecular network of the as-spun yarn is oriented further. The increasing parallelism of the molecular chains initiates PETP crystallization. With further drawing, the crystalline regions and the amorphous

areas between them become more and more oriented [58]. Orientation, crystallinity, and strength of the drawn yarn increase with increasing draw ratio, whereas break elongation decreases [59]. This is illustrated by the stress–strain diagrams of yarns drawn differently (Fig. 1).

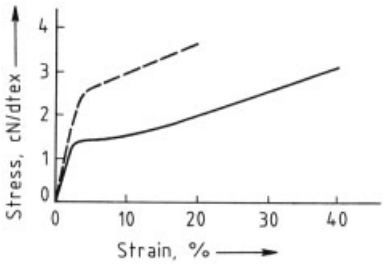

Figure 1. Stress – strain curves of two drawn PETP yarns
—— Draw ratio 1 : 1.52; – – – – Draw ratio 1 : 1.83

In addition, the product of orienting spin-draw and draw ratio is approximately constant for the maximum possible draw ratio and for drawing to a given break elongation [57]. Thus, PETP yarn spun at higher speeds needs to be drawn less to attain a given break elongation. However, this affects the structure of the drawn material [53], [57], [60]. Because less energy is added during drawing, the crystallinity is somewhat lower, and both the total orientation, as measured by birefringence, and the tenacity of the yarn are reduced [61]. These relations begin to change at spinning speeds of ca. 5000 m/min. Then the as-spun yarn is so highly crystalline that the ordered regions are no longer loosened during drawing but orient themselves as far as possible as a whole along the fiber axis. The crystallinity again increases [48].

The more the drawing temperature exceeds the glass transition temperature the more easily can PETP be drawn [44]. Drawing tension drops if drawing temperature is increased and also if the glass transition temperature is decreased, either by water absorption by the as-spun yarn [37] or by modifications of the polymer. However, drawing temperature must not be set too high, for then the purely plastic and nonorienting elongation becomes too large a percentage of the total draw.

The draw ratio would have to be set very high to achieve high strength. The yarn in the draw zone would then no longer be able to absorb stresses and would break. As with other polymers, two-stage drawing is necessary; the predrawn yarn enters a second draw zone where it is then capable of withstanding the drawing tension.

Drawing Process. In industrial practice, PETP as-spun yarn is almost always drawn continuously between two godets or sets of rolls, whose speeds determine the draw ratio [54]. In so doing, the required draw temperature is imparted to the yarn by heaters or heated rolls [37]. Keeping the draw point stationary is important; otherwise the drawn product will contain differentially drawn areas, which produce defects, particularly nonuniform dye uptake. Details of each drawing arrangement are dictated by the product to be drawn (see also → Fibers, 3. General Production Technology, Chap. 2.2.).

2.1.4. Aftertreatment

Thermal Aftertreatment. At room temperature, oriented and noncrystallized segments of the PETP molecular chains in a drawn thread-like structure are frozen and immobile; the entire structure is not in equilibrium. When the yarn is heated beyond the glass transition temperature, the amorphous chain segments begin to move so as to achieve an entropically more favorable, less oriented position, to the extent permitted by their restraints in the crystalline regions. These regions also lose orientation and, as an overall result, the yarn shrinks. The extent of shrinkage depends on the orientation and crystallinity of the drawn structure; it increases with increasing temperature and, initially, also with the duration of heat treatment [44], [59], [62]. Any plasticizers, e.g., water, that have diffused into the structure favor shrinkage.

The thermal instability of drawn PETP yarn structures would be quite troublesome during further processing and in textile end products. Therefore, it is often reduced during fiber and yarn manufacture by a thermal aftertreatment called *heat-setting* (see also → Fibers, 3. General Production Technology, Chap. 2.4.) [37], [63].

Structures that are allowed to shrink without restraint between 110 and 180 °C are thermally stable below the setting temperature. In this process, orientation decreases markedly, which

manifests itself on the stress – strain curve as a saddle with a yield stress of ca. 1 cN/dtex (Fig. 2). If the product, which has been heat-set without restraint, is loaded to this yield stress saddle, it is permanently deformed. In general, normal textile applications rarely reach this amount of loading.

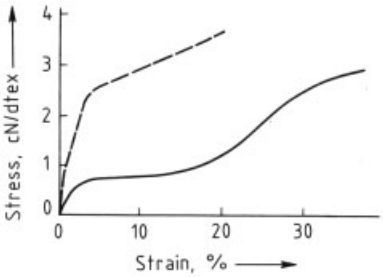

Figure 2. Stress – strain curves of drawn PETP yarn, subsequently allowed to shrink freely
– – – – Drawn yarn; —— Drawn and allowed to shrink freely for 5 min in steam

During unrestrained heat-setting, crystallinity increases considerably, primarily favoring the ordered regions already present. Degrees of crystallinity up to 50 % have been reported. The amorphous fraction is correspondingly reduced, but the chain segments are disoriented in response to shrinkage. With increasing crystallinity, maximum dye uptake (i.e., the saturation concentration) decreases but dye diffusion becomes much easier as the amorphous chain segments disorient. In this way, dyeability reaches a minimum at heat-setting temperatures of 150 – 180 °C [59], [62], [64].

If heat-setting takes place at constant length without any shrinkage, the effects are intensified with temperature and duration of heat-setting. Crystallinity also increases but less so than during unrestrained heat-setting. The PETP products are not completely stable thermally but will retain some residual shrinkage [59], [63]. Therefore, the temperature at which this type of heat-setting is carried out is frequently very high, e.g., 220 °C. Cooling the material under tension after it leaves the heat-setting zone is important to prevent any shrinkage. Heat-setting at constant length decreases extensibility and increases strength.

Hot postdrawing of a previously drawn yarn represents heat-setting with negative shrinkage. Crystallinity and orientation increase markedly, and the initial slope of the stress – strain curve of the yarn and the tenacity increase. The extent of changes in shrinkability depends on processing conditions.

In practice, a middle course is often adopted, i.e., heat-setting while constant shrinkage is allowed [63]. This increases crystallinity and somewhat reduces orientation. The increase in extensibility of the material approximately corresponds to the shrinkage permitted; at the same time, a slight saddle is formed on the stress – strain curve. The residual shrinkage of material heat-set in this way is smaller by about the amount of shrinkage permitted during setting. The many processing parameters, such as draw ratio, heat-setting temperature, and shrinkage allowed during setting, permit matching the properties of the PETP product quite closely to the requirements of each end use.

Heat-setting does more than prevent thermal instability. The structural rearrangements it induces make prior changes in shape impressed on the yarn permanent [37]. This feature is used widely in producing and processing fiber products, e.g., in ironing or pleating when creases are to be set hot. However, in producing the fibrous structure and in further heat-setting, care must be taken to chose conditions which will assure that any subsequent heat-setting will still be effective. "Deadening" a fiber (or a textile structure formed from it), so that it can no longer be heat-set, should not be done.

During setting, primarily the orientation of the amorphous parts changes; however, growing crystalline regions also disorient if any shrinkage has been permitted. This leads to a corresponding change in the superstructure of the PETP fiber. In general, the layered structure becomes more pronounced. The orientation angle of the crystalline layers with respect to the fiber axis increases [59].

Structural rearrangements during heat-setting and the resultant changes in textile properties depend in a complicated way on the structure and prior history of the PETP yarn and on the conditions of the heat-setting process. Therefore, formulating general, quantitative relationships is impossible. These depend on the given heat-setting process, which differs from product to product. All setting processes are included in the manufacturing process immediately after drawing or deformation.

Deformation. Melt-spun fibers including PETP are generally flat and, in this condition, unsuitable for many applications [37]. Hence, all staple fibers and tows must be matched to natural fibers by *crimping*.(→ Fibers, 3. General Production Technology, Chap. 2.3.) PETP filament yarns for apparel are *textured* (→ Fibers, 3. General Production Technology, Chap. 3.3.) on a large scale. The processes and installations used in both cases are quite similar for all manmade fibers. The deformations produced are frequently heat-set, either without tension (staple fibers heat-set without restraint) or with tension (texturing), depending on the product. The features specific to PETP are product-related and, where necessary, are described in Chapter 3.

2.1.5. Dyeing of Poly(ethylene terephthalate) Fiber Products

Dyeing of textiles and dyeing technology in general are discussed under → Textile Dyeing.

As a result of polycondensation and subsequent degradation, PETP contains OH and COOH end groups. Neither is suitable for bonding ionic dyes. Therefore, PETP is dyed with disperse dyes unless it has been modified to contain ionic end groups (→ Disperse Dyes). These dyes are available in powder or liquid form. They are only slightly soluble in water and are, therefore, used as aqueous dispersions in the dye bath or even as foam. Dye preparations necessarily contain dispersants.

Migration of dyes into the fiber is controlled by the laws of diffusion [65]. The parameter determining the rate of migration is thus the diffusion constant D, which depends on the size and structure of the dye molecule. The saturation concentration c_∞, reached at very long dyeing times, is the second parameter required to describe the kinetics of dyeing.

Investigations have shown that D is affected substantially by the orientation of the molecular segments in the noncrystalline region of the fiber. With increasing orientation, molecular mobility decreases, and the glass transition temperature T_g, a measure of mobility, increases. Beginning at ca. 68 °C in an isotropic, amorphous PETP, T_g can reach 100 °C in a highly oriented, partially crystalline fiber [66], [67].

The diffusion constant is also affected by the ordering of the molecules, though to a much lesser degree. This becomes evident if the temperature of heat treatment prior to dyeing exceeds 180 °C. Above this temperature, the noncrystalline regions (layers between crystalline layers) are loosened, which restores the mobility of the molecular segments in these regions; parts of molecules that up to then had been extended can coil up or even refold, which leads to a growth in crystalline regions transverse to the molecular axis and to an "emptying" of the noncrystalline regions. Material that has been spun at high speed and then drawn behaves similarly (see Section 2.3.2) [53]. However, such ordered states in the noncrystalline regions of the fiber are first of all key to controlling the saturation concentration c_∞.

The fiber can take up dye to any significant degree only at temperatures above T_g. At these temperatures, the polymer is present as a highly viscous liquid in which the mobility of molecular segments in the noncrystalline regions is restricted. The saturation concentration evidently depends on the accessible free volume in the noncrystalline regions and in voids. It is also certainly affected by the structure of the interfaces of these volumes. A quantitative description is not yet possible.

Thus, with reference to fiber structure, the diffusion constant D is the most important orientation-dependent kinetic parameter of dye migration through the "liquid" polymer regions. The saturation concentration is determined by the states of order of the supercrystalline structure (small-angle X-ray structure). However, practical dyeing behavior is essentially determined by the diffusion constant.

Poly(ethylene terephthalate) fibers can be dyed to deep shades without using auxiliary chemicals only at temperatures ca. 50 °C above the glass transition temperature. Such dyeing is carried out in aqueous media under pressure in a closed vessel and is called high-temperature (HT) dyeing. High-temperature dyeing of textiles made of 100 % PETP or blends of PETP with cotton or other cellulosic fibers presents no problems. Any wool in the blend, however, is damaged.

Carriers are used for open dyeing in boiling water, especially of blends of PETP and wool. These chemicals are swelling agents, i.e.,

poor solvents, for the polyester fiber. They temporarily loosen intermolecular bonds in the fiber and thus increase the mobility of molecular segments. In other words, they lower the glass transition temperature. The problems connected with cleanup of wastewater are a disadvantage.

Therefore, PETP fibers have been developed that can be dyed to deep shades in boiling water without carrier (see Polymer, Section 2.2.1) [68]. These fibers contain block comonomers, generally polyglycols, and meet all requirements of the apparel and home furnishings trades. However, minor losses in light- and washfastness of the dyeings compared to unmodified PETP are unavoidable.

Ionic dyeing for differential dyeing requires appropriate ionic end groups in the molecule. Sodium salts of 5-sulfoisophthalic acid dimethyl ester [*138-25-0*] and 5-sulfoisophthalic acid [*22326-31-4*] have been used; they can be added during transesterification of DMT or during direct esterification of TPA [69]. Fibers can be dyed with cationic dyes without difficulty, but the dyeings are much less lightfast.

Dyeing with acid dyes requires basic end groups. Thus far, building such end groups into PETP, e.g., via amines, has not been possible without yellowing or high aminolytic degradation, resulting from the high polycondensation and spinning temperatures.

Dyeing Processes. The most important dyeing processes are the exhaust process and the continuous process; aftertreatment of PETP dyeings is very important for the quality of the dyeing.

The dyeing process most widely used with PETP fiber products is the *exhaust process.* Depending on the situation in each dye plant (wastewater) and the chemical composition of the dyed goods (presence of wool), dyeing can be carried out under pressure at HT conditions (60 – 90 min at 125 – 140 °C), or it may be carried out open or under low pressure, respectively, with carrier (60 – 90 min at 95 – 110 °C). Combining both processes is rare and used only if highly oriented PETP products are to be dyed to deep shades.

Continuous processes are also used. With the aid of saturated steam, superheated steam (pad steam process, 2 – 8 min holdup time at 160 – 180 °C), or hot air (pad thermosol process, 15 – 90 s holdup time at 180 – 220 °C), high throughputs, which improve the economics of dyeing, can be achieved. Blends of PETP with cotton and other cellulosic fibers are preferably dyed by these processes.

Aftertreatment of dyed goods is essential for the quality of the dyeing. This is especially true with regard to their fastness, but also holds for other properties of dyed goods (fibers, yarns, and textile fabrics) such as processibility, appearance, hand, and mechanical properties. In deep dyeing, particularly, residual dye remains on the fiber surface after drying and heat-setting with deleterious effects, particularly on fastness. A reducing, alkaline, postdyeing, cleaning treatment removes these dyes, as well as any oligomers that are diffused from the interior to the surface of the fibers. Temperature and pH of the postcleaning bath must be carefully controlled because of the danger of hydrolytic degradation.

2.2. Structure and Properties

2.2.1. Polymer

Chemical Structure.

Homopolymer. In addition to the OH end groups of the macromolecules, PETP fibers contain 10 – 40 milliequivalents of carboxyl (COOH) end groups per kilogram of fiber. These are produced in the melt at elevated temperature (especially beginning at 280 °C) by breaking of the ester bonds. Because COOH end groups catalyze hydrolytic degradation of the polyester, they are occasionally protected (e.g., with epoxide compounds or carbodiimides), especially in products destined for industrial use [70].

Diethylene glycol, which is present in the PETP chain in concentrations up to ca. 1 mol %, affects softening behavior of the polyester, dyeability of the fibers, and to a certain extent, lightfastness. For this reason, diglycol formation is suppressed as much as possible by adjusting the conditions of the reaction.

In addition to macromolecules, the polymer contains ca. 2 % oligomers, primarily cyclic trimers, but also linear oligomers. The former can adversely affect washing and dyeing.

Additional impurities in the polyester are acetaldehyde and degradation products. At

higher thermal loading of the melt during further processing, these can lead to formation of gel particles and discoloration.

In the solid state, PETP absorbs ca. 0.4 % of water from the atmosphere at 20 °C and 65 % relative humidity. Because even minute amounts of water can cause, or at least substantially favor, hydrolytic degradation, the polyester must be carefully dried to a moisture content below 0.01 % prior to spinning.

Copolymer. The principle of copolymerization of PETP with suitable monomers is used to improve certain fiber properties.

The relatively high glass transition temperature of the fibers limits dyeability in boiling water to light and medium shades (see Section 2.1.5). Deep shades can be obtained by using higher dyeing temperature (HT dyeing at ca. 135 °C under pressure) or carriers. When these procedures are impractical, e.g., because of the presence of wool or the contamination of wastewater, fibers of block copolymers (PETP with polyglycols) can be used [68], [69]. By loosening the noncrystalline regions, the glass transition temperature of these fibers is lowered far enough to attain both the required diffusion and the desired saturation concentration of disperse dyes in the fiber.

In differential dyeing (single bath dyeing of textile fabrics to various shades), fibers are needed that can be dyed with both cationic and disperse dyes. The sodium salt of 5-sulfoisophthalic acid [*51876-65-4*] is almost exclusively built into the PETP molecule for this purpose [69].

The low pilling property of textile fabrics, necessary for practical use, requires low-pilling PETP staple fibers (see Section 2.3.1). Low molecular mass PETP is used for these fibers. In addition, cross-links sensitive to hydrolysis are also built into the polyester with the aid of, for example, organosilicon compounds [71].

A particular requirement for home furnishings is low flammability. For years PETP fibers have been used in which flame retardants in the form of suitable phosphorus compounds are co-condensed with PETP (→ Flame Retardants) [72], [73]. In this way, flame protection becomes permanent; it cannot be lost in washing or dry cleaning.

Additives. In many cases, the polymer includes additives intended to change the normally transparent and brilliant white appearance of fiber products.

The most frequently used additive is titanium dioxide [*13463-67-7*], mostly anatase but occasionally the harder rutile, which serve to dull the fiber. Particles with average grain sizes between 0.3 and 0.4 µm are added to the transesterification product in concentrations of 0.1 – 3 %.

Dyed fiber products are made by adding soluble dyes or pigments, if extreme lightfastness of dyed goods is required, to the melt. To produce black fibers, carbon black is usually added.

Many attempts have been made to achieve permanently antistatic polyester fibers by admixture of suitable compounds. Surface resistances of 10^{10} Ω have been attained [74]. However, not much practical use has as yet been made of these possibilities. In general, the desired electrical surface conductivity is obtained by finishes applied to the fiber.

Physical – Chemical Structure. The most important structural variable of fiber-forming polymers is the *average molecular mass*. This describes the mass of the average macromolecule and is the product of the molecular mass of the monomer and the average degree of polymerization (→ Polymerization Processes). A measure of the molecular mass and, hence, the length of molecules is the specific viscosity, η_{spec}, measured in dilute solution:

$$\eta_{\mathrm{spec}} = (\eta - \eta_0)/\eta_0$$

where η is the viscosity of the solution and η_0 the viscosity of solvent. For poly(ethylene terephthalate), $\eta^{\mathrm{c}}_{\mathrm{spe}}$ is between 0.5 and 1.5, depending on the application of the fiber products.

Generally, molecular mass is described by the intrinsic viscosity [η], often designated as *IV*

$$[\eta] = \lim_{c \to 0} \eta_{\mathrm{spec}}/c$$

where c is the concentration of polymer in the solution being measured. Again, depending on the area of application, [η] for PETP ranges from 0.45 dL/g for low-pilling fibers to 1.05 dL/g for tire cord, which corresponds to a span of 50 to over 200 for the average degree of polymerization.

The *distribution of molecular mass* corresponds to that given by Flory for condensation

polymers [75]. Changes in this distribution function by degradation can be reversed by melting. The distribution function of molecular mass is rarely measured and is not used for product testing.

Melt viscosity is measured with capillary or rotary viscometers. In polymers intended for standard fibers, it is ca. 250 Pa · s at 285 °C. Melt viscosity increases rapidly with increasing $[\eta]$ and decreasing temperature. It is one of the basic variables of spinnability. The orientation of the macromolecules during filament formation, is particularly dependent on melt viscosity.

The *transition temperatures,* generally measured by differential thermal analysis, are the final determinants of the orientation, crystallization, and melting behavior of PETP. The glass transition temperature is measured as ca. 68 °C on the isotropic and amorphous polymer. It depends on the orientation and crystallinity of the polymer and thus can be as high as 100 °C in highly oriented, semicrystalline, industrial, continuous-filament yarns. In the isotropic and amorphous polymer, the crystallization temperature is ca. 100 °C; the maximum rate of crystallization is observed between 140 and 180 °C. The melting point of PETP, which depends on crystallinity and crystal size, lies between 255 and 265 °C.

Physical Structure. Poly(ethylene terephthalate) quenched from the melt is isotropic and almost amorphous. The density of amorphous PETP is 1.335 g/cm^3. The polymer can be crystallized primarily by the action of heat but also by the presence of agents such as dichloromethane or dioxane. The crystal structure is triclinic; the density of the crystalline regions is 1.515 g/cm^3 [76].

The configuration and conformation of monomeric terephthalic acid and ethylene glycol in the macromolecule are also known. Infrared spectroscopy shows that the terephthalic acid component is present in the trans configuration in the crystalline part, but in the cis configuration in the noncrystalline part. The IR method shows that the glycol component exists in the trans conformation in the undisturbed crystalline region and in both the gauche and the trans conformation in the noncrystalline region. The mass ratio of the glycol components in the gauche and trans conformations changes not only during crystallization but also during orientation and disorientation. This ratio correlates to the mechanical properties of shaped PETP articles [77].

When PETP granulate is melted and the melt is forced through spinneret holes, as-spun filaments are formed. The orientation of the macromolecules in these filaments, characterized by the birefringence Δn of the filaments, depends to a large extent on windup speed. Thus, Δn at a windup speed of 1500 m/min is only ca. 0.010, but at 3500 m/min it is as much as 0.055. Density, too, is subject to this dependence. At 1500 m/min it is ca. 1.34 g/cm^3, and at 3500 m/min it is 1.35 g/cm^3. The degrees of crystallinity behave similarly [78].

During drawing of as-spun filaments, the macromolecules are oriented further, and the birefringence of filament yarns for industrial uses increases to ca. 0.240. The ability of PETP to crystallize is exploited in many ways during drawing and also in subsequent aftertreatment. Thus, the fiber density ranges from 1.36 g/cm^3 for drawn high-shrinkage staple fibers up to 1.41 g/cm^3 for highly drawn low-shrinkage filament yarns.

2.2.2. As-Spun Yarns

Nonuniformities of as-spun yarn along a filament or filament bundle, but also from filament to filament, from bundle to bundle, or even from plant to plant, are reflected in the structure of the final product. There, they appear as differences in textile properties, primarily in dye uptake. For this reason, yarn is carefully monitored during spinning. The glass transition temperature of PETP (70 °C) causes the as-spun yarn to be quite stable at room temperature and makes measurement of orientation and even of crystallinity relatively easy. Intervening in the manufacturing process in case of deviation is then possible.

Many physical measurement techniques are available to investigate the structure of PETP as-spun yarn. The total orientation of a filament is measured by *optical birefringence.* Except for highly crystalline as-spun yarn, the birefringence is proportional to the orientation factor [44]. If suitably measured, it provides indications of any possible sheath-core structure.

Another measure of orientation is the velocity of sound and with it the *sonic modulus* [79], [80]. They change little at lower orientation and increase noticeably only at spinning speeds ≥ 2500 m/min [47].

The degree of crystallinity of as-spun yarn can be calculated easily from its *density*; care is required, however, because the densities of completely amorphous and completely crystalline PETP, required in these calculations, are not entirely independent of orientation [47]. In addition, true separation of the amorphous and crystalline phases is presumed, which is sometimes questionable.

Much information can be gained from *X-ray* investigations. Wide-angle patters or goniometry provide data on the size, quantity, and orientation of the crystallites, and the approximate degree of crystallization. With their aid, total orientation can be divided into orientation of the amorphous and of the crystalline fractions of the filament. Small angle X-ray scattering permits conclusions to be drawn concerning the superstructure, i.e., the mutual arrangement of crystalline and amorphous regions and also the state of order of the latter.

Thermal Measurements. If as-spun yarn or tow is heated at constant length beyond the softening point to ca. 100 °C, a *shrinkage stress* that corresponds to the orienting stress during spinning is created [57]. This stress is, therefore, a measure of orientation and is largely proportional to the optical birefringence. These two quantities diverge only at high spinning speeds.

If as-spun yarn is heated without restraint, it shrinks, sometimes very much. In so doing, the orienting elongation D, frozen in during spinning, is recovered; hence, D can be calculated from the *spun yarn shrinkage* [81]. As expected from theory, the quantity $D^2 - 1/D$ is approximately proportional to birefringence and shrinkage stress at constant length, which are measures of orientation [81], [82]. However, this is true only for spinning speeds up to 1500 m/min with normal quench. Beyond that, shrinkage is impeded by crystalline regions developed in the sample during spinning or by the regions created during measurement. Thus, spun yarn shrinkage goes through a maximum of 60 – 70 % at spinning speeds between 2000 and 4000 m/min and then drops rapidly with increasing speed [46], [53], [57]. In this spinning range, spun yarn shrinkage is again recommended as a measure of crystallinity, even of dyeability of the final product [53], [61].

Differential Thermal Analysis (DTA). The glass transition temperature, the temperature of the maximum rate of crystallization, the melting point of crystallites, and the final melting point can all be determined from DTA curves [47].

Mechanical Tests. If a PETP as-spun yarn is elongated at room temperature and its tensile stress is measured at the same time, the stress – strain curves shown in Figure 3 are obtained. For as-spun yarn of lower orientation, a saddle appears after the initial rise in stress, while a draw point crawls along the clamped yarn [54]. The percentage of elongation reached at the end of this saddle is called the natural draw ratio R_n of the as-spun yarn. With increasing spinning orientation, shown in the figure by higher spinning speeds, R_n decreases; this is because a greater amount of extensibility has been anticipated during spinning [57]. If spinning orientation becomes too large, R_n can no longer be specified [53], [60]. Then either the elongation of the as-spun yarn at a given load or the force necessary to produce a given elongation is used [57]. In practice, determining the operational draw ratio that imparts a definite property, such as break elongation or dye uptake to the final product might be useful. This, however, assumes that no changes occur in further processing.

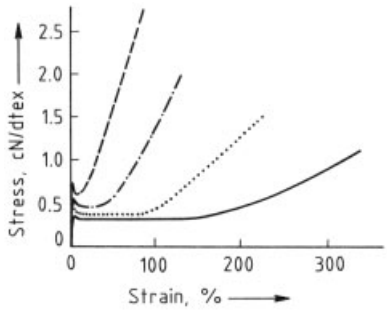

Figure 3. Stress – strain curves of PETP as-spun yarns spun at four spinning speeds
Spinning speeds:
——— 1200 m/min
· · · · · · 2000 m/min
— · — · 3000 m/min
– – – – 4000 m/min

To monitor the spinning process and possibly pinpoint defects in commercial practice, tests should be limited to those that can be most read-

ily and accurately carried out and that provide averages over a sufficient quantity of as-spun yarn. Each kind of test, taken separately, is useful only for monitoring, not for control. If a test quantity exceeds the tolerance range, use of another kind of test is recommended to readjust the process conditions.

2.2.3. Drawn Materials

During drawing, PETP crystallizes in the triclinic configuration mentioned in Section 2.2.1, "Poly(ethylene terephthalate) with Very High Molecular Mass." The supermolecular structure is assumed to consist of somewhat parallel layers, inclined to the fiber axis. With increasing draw, the distance between the layers, measured in a direction normal to the layer surface, decreases, and the layers rotate more in the direction of the fiber axis [59], [83].

In practice, material that has been drawn but not processed further is almost never available as an intermediate product. A posttreatment almost always follows immediately, as described in Section 2.1.4.

2.2.4. Dyed Poly(ethylene terephthalate) Material

Fastness Properties. The lightfastness of dyed textiles for apparel and home furnishings is important in preserving their coloration. A lightfastness rating of 6 can be achieved readily in articles of unmodified PETP fibers with most commercially available disperse dyes. The lightfastness of PETP fibers, modified to be dyeable to deep shades at the boil without carrier, may be half a rating lower. Dyeings of PETP fibers, modified to be cationic dyeable, generally do not rate higher than 4 – 5 in lightfastness.

Heat-setting fastness, rubbing fastness, and washfastness of dyeings of PETP fibers meet all practical requirements. The only exception is the slightly lower washfastness of modified PETP fiber, dyeable without carrier. A wash temperature of 40 °C should not be exceeded.

Degradation and Mechanical Properties. In discussing alkaline postdyeing cleaning treatment, the danger of hydrolytic damage to PETP fibers is mentioned (see Section 2.1.5, Dyeing Processes). This damage consists of molecular degradation, which can extend to different degrees from the fiber skin to the core. The consequence, in general, is a reduction in tensile strength and breaking elongation of the fibers caused by the embrittlement connected with degradation. At the same time, the hand of the textile fabric can occasionally be changed. However, sometimes such changes of the hand are desired (e.g., by alkalization).

Another change in the properties of PETP fibers or textiles is sometimes produced by shrinkage during dyeing, which may be accompanied by crystallization processes. Carefully controlling the dimensions of, and stresses in, textile fabrics as well as the drying and setting temperatures is required.

2.3. Poly(ethylene terephthalate) Fibers

2.3.1. Staple Fibers

Special Features of the Manufacturing Process. In the manufacture of PETP staple fibers, polymers with intrinsic viscosities $[\eta]$ of 0.65 – 0.75 are generally used; for special types, polymers of much lower viscosity with $[\eta]$ ca. 0.5 are occasionally used. The equipment on which PETP staple fibers are produced is usually large, with daily production up to 50 t and more. Therefore, feeding the spinning units directly from a condensation unit makes economic sense (see also → Fibers, 3. General Production Technology) [84].

A spinning unit for PETP staple fibers consists of 8 – 32 spinning positions including their accompanying spinning shafts, aligned side by side. Each position is supplied with melt, either simultaneously or in groups, by an extruder or directly from a condensation unit [37].

Depending on the type of air quench, spinnerets may be circular with uniformly arranged spinneret holes, they may be plates in which the holes are arranged rectangularly, and finally they may be circular with an annular arrangement of holes [84]. The number of holes is based on the as-spun fineness and the feed of the spinning position. The spinneret hole shape is not always round. For profiled fibers with nonround cross sections, such as triangles, five-pointed stars, or

even hollow fibers, the holes must have profiles that exaggerate those of the final product because the profile impressed by the hole tends to blur somewhat during spinning.

The output of a spinning position has been increased from 500 g/min to 1500 g/min and more by increasing the number of spinneret holes, by forceful quenching, and by increasing the spinning speed.

For example, a filament bundle with a linear density of 8 dtex per filament is spun from an 1125-hole spinneret at a speed of 1500 m/min. The output per spinning position is 1350 g/min. A 16-position spinning machine then provides about 31 t/d of as-spun yarn in the form of a tow with total length of 2160 km and a linear density of 144 000 dtex, consisting of 18 000 individual filaments.

The length of the spinning shafts is determined by the need not only to ensure that the filaments are cooled below 70 °C but also to prevent the moving individual filaments from sticking together. The shafts are generally 5 m long but, with a suitable quench, they may be shorter [84].

Finish is applied to the filament bundle at the lower end of the shaft to ensure trouble-free packaging of the as-spun material. The spin finish is also used to impart surface characteristics to the fibers needed downstream during manufacturing and textile processing [85], [86], [87]. This is more economical and protects the environment better, but some compromises with respect to processibility during tow drawing and textile processing are unavoidable.

After the finish application, strands coming down from the spinning units are turned horizontally and combined into a tow, which is drawn off and deposited in a spinning can. Depending on the installation, these cans hold 0.5 – 2 t of tow, all of which should have the same length.

The fibrous as-spun product is preoriented; in newer spinning units with very effective quench, for example, it is characterized by a birefringence of $(8-14)\times10^{-3}$ or by a natural draw ratio of ca. 1 : 2.5.

High-speed spinning at ca. 3000 m/min, which is advantageous for textile filament yarns, has been attempted for production of PETP staple fibers [88]. Problems still exist in developing a reliable process for depositing the tow in the spinning can [89]. A sheath-core effect in the filaments also presents difficulties.

As customary in the production of staple fibers and tows from melt- or dry-spun synthetic polymers, the as-spun tows are combined and further processed continuously as a tow band. The tow band (up to ca. 1.20 m wide) is first immersed in a warm bath where the stresses in the individual strands are equalized; the tow is thoroughly wetted, and if necessary, additional finish is added. From the immersion bath, the tow enters the draw zone between two sets of rolls [85].

During drawing, the draw points or draw zones of all filaments in the band must be as close together as possible and must not wander [85]. A number of drawing processes are available for this [90]. If the tow is cold when it enters the draw zone, the yield stress must be lowered by sudden heating, e.g., via a steam jet, so that drawing can occur. In this process, the draw ratio is limited, and if necessary, a second drawing step must be provided.

If the tow is heated on hot feed rolls, drawing begins on the last or next-to-last feed roll and can even be facilitated, if necessary, by an immersion bath under the third roll from the end. It is also usual to pass the tow through an additional hot zone in the draw stand. With this process, reaching higher draws with one drawing is possible.

Drawing in water is a compromise in which only the last feed roll is heated and dipped into the hot water draw bath. Drawing begins at that point.

If postdrawing is to be accomplished in a separate draw zone, the draw rolls of the first stage are heated because they represent the feed rolls of the second draw, which is then carried out in a heating device.

The stress – strain behavior of the end product can be preselected by the draw ratio but is more critically affected by the type and place of heat-setting. Thus, fibers that are to exhibit a steep stress – strain curve with high initial modulus, similar to cotton, are relatively highly drawn, e.g., 1 : 4.5 to 1 : 6, depending on the as-spun yarn. After drawing, the tow is heat-set under tension, usually at constant length. The heat-setting temperature is between 160 and 230 °C, depending on the shrinkage desired. After setting, the tow traverses a cooling zone under

tension and then enters the crimper. To facilitate crimping, it is first slightly moistened or steamed. The tow must not become too hot in the crimper; otherwise it relaxes and textile properties deteriorate. After crimping, finish can again be applied if necessary, in which case the tow must again be carefully dried [90].

Fibers that are to have a flat stress – strain curve similar to wool do not need to be so highly drawn; a draw ratio range of ca. 1 : 2.5 to 1 : 3 will suffice. After drawing, the tow is sent immediately to the crimper via a short steam zone. The crimp is then heat-set without tension, often on traveling grate machines. The heat-setting temperature, ca. 120 – 180 °C, is selected according to the amount of shrinkage desired in the end product. The flat stress – strain curve is formed by the tensionless setting, and the crimp becomes highly stable.

Finally, the crimped and, if necessary, shrinkage-stabilized tow is packaged. Either it is compressed into a bale and sold as endless tow, or it passes through a cutter and is cut into staple fibers, which are then baled. The usual mass of a bale can range up to 500 kg.

Staple fiber draw lines are up to 80 m long and produce from 20 to 250 t/d, the latter, of course only for high-volume products. Delivery speeds range from 150 to 300 m/min, with tow sizes up to 600 ktex and more [91]. Investment and manufacturing costs per ton per day are reduced considerably with increasing productivity of these draw lines [89].

Completely continuous installations with capacities of 12 – 15 t/d produce smaller lots. They are supposed to be superior to the usual tow band installations in terms of investment and manufacturing costs, but they produce fibers with three-dimensional crimp. Whether such fibers can be spun in blends with natural fibers and whether fabrics produced from such yarns will be acceptable in the marketplace remain open questions [89].

Types of Fibers, Properties, and Uses. Polyester staple fibers encompass the broadest product spectrum of any fiber. They can be found in all areas of application [37, 92–94]. They are used in apparel for overcoats, jackets, leisure- and sportswear, ladies' outerwear, men's suiting, work and protective clothing, children's wear, underwear, and sleepwear. In home furnishings, they range from drapery and curtain fabrics, through furniture coverings, pillows and pillow stuffing, table and bed linen, to wall and floor coverings. They have a whole host of industrial end uses along with filament yarns, spunbonded products, and monofilaments. A few examples of applications are in filtration, automobile interiors, geotextiles, and material for the sewing trade.

Yarns made from PETP fibers can be processed without reservation in weaving and knitting as well as in tufting. The fibers are completely suitable for processing as unblended yarns (100 % PETP), as well as in blends with cotton, wool, silk, cellulosic man-made fibers, acrylic fibers, and many more.

Of all polyester fibers, *cotton types* have the largest market share. These are processed into blended yarns, and these in turn into blended articles. These contain cotton, viscose, or modal fibers in what are today considered the "classical" blend ratios of 65 : 35 or 1 : 1 (PETP to cotton) and in other ratios.

The stress – strain curve of cotton types is shown in Figure 4. Its initial slope corresponds approximately to that of cotton. Their finenesses of 0.4, 0.9, 1.3, and 2.4 dtex, and the major one of 1.7 dtex, at cut lengths of 30 mm and 38 mm are matched to cotton.

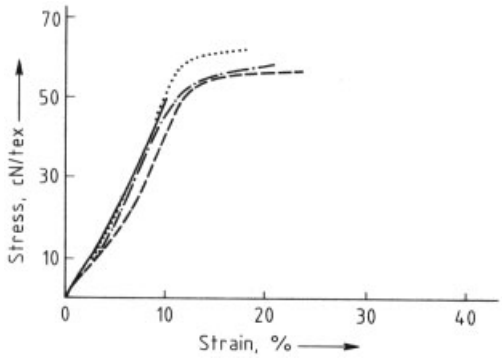

Figure 4. Stress – strain curves of PETP cotton type staple fibers
———— Cotton
– – – – – PETP cotton type (normal shrinkage)
— · — · — PETP cotton type (low shrinkage)
· · · · · · · · PETP sewing thread type

Fibers with normal high shrinkage and others with reduced shrinkage are available and used to vary the hand and appearance of textile fabrics over a wide range. With the additional pos-

sibilities offered by the choice of fineness and weave construction, the textile processor can exert a major influence on the permeability of, for example, tent, overcoat, shirt, and blouse fabrics. In comparison to wool types of polyester fiber, dyeing of highly oriented cotton types is somewhat slower but, like the former, presents no problems.

Special cotton-type fibers, which are also offered as continuous tow for processing on stretch – breaking converters, are intended for sewing. These fibers are especially highly drawn and, therefore, have the highest moduli (resistance to elongation) and tenacities.

Wool types are intended primarily for processing in blends with wool. The major blend ratio is 55 : 45 (PETP to wool). The stress – strain curves of these fibers are matched to that of wool (see Fig. 5) by tensionless heat-setting during fiber manufacture, as described in Section 2.1.4. Shrinkage, therefore, is low, and dyeability is optimum. The spectrum of fineness and cut length ranges from 1.7 dtex (38 mm) to 6.7 dtex (120 mm). Wool types are also available as continuous tow for processing on cutter – converters.

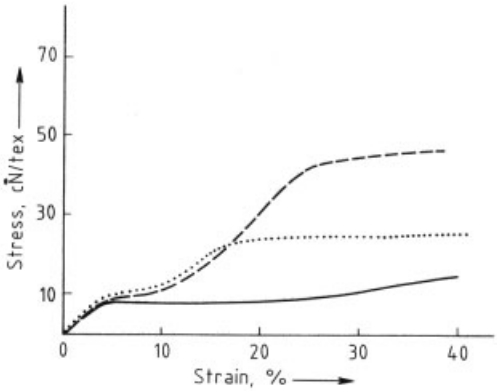

Figure 5. Stress – strain curves of PETP wool types
—— Wool
– – – – PETP wool type (normal)
· · · · · PETP wool type (low pilling)

The so-called pilling phenomenon, which occurs in blends with wool, has been responsible for considerable developmental cost. Small balls or pills are formed on the surface of textile fabrics during use. Suitable fibers have been developed to solve this problem [59], [93]. Their special structure eventually causes the fibers anchoring the pill to break. In this manner, the pill disappears soon after it is formed.

A third group consists of the *fiberfill types* commercially available as staple fibers but also as continuous tow. These fibers are used both in apparel (primarily for heat insulation) and in home furnishings (primarily in pillows, cushions, and bedding). Their crimp is two- or three-dimensional. To improve the hand of fiberfill, such fibers are often provided with a special finish, based primarily on silicone. The range of linear densities of these fibers is between 1.5 and 14 dtex at a cut length of 30 – 80 mm (most common length, 60 mm). This group also contains high-shrinkage fibers with a shrinkage at the boil of 50 %.

In home furnishings and automobile interiors, *polyester carpet types* are used. These also include poly(tetramethylene terephthalate) fibers, see Section 2.4.1 (→ Polyesters, → Floor Coverings). Carpet fibers based on PETP are used as both normal and high-shrinkage types in home furnishings and in needle-punched felt. Such fibers are also made with various cross sections; their finenesses and cut lengths range from 6.7 dtex (135 mm) to 20 dtex (150 mm). An essential characteristic of polyester floor coverings is that their mechanical properties are unaffected by moisture.

Special types of PETP staple fibers have secured a place in the market. Some can be dyed in an open vessel without carrier to deep shades [68]. Such staple fibers and tows meet special requirements of environmental protection where wool blends have to be used (carriers) and HT dyeing is, therefore, excluded. These fibers exist as standard types and as pill-resistant tows. Cationic dyeable staple fibers are used in relatively small amounts in the wool area and in polyester carpets.

Since the mid 1970s,flame-retardant PETP fibers have rounded out the product spectrum [73]. These fibers are used in apparel and, especially, in home furnishings. They meet legal requirements and even exceed them in some cases. For these, PETP copolymers with phosphorus compounds have proved particularly useful. Fibers of this type combine the comfort expected of unmodified PETP fibers with modern safety requirements. They are distinguished by deep dyeability with disperse dyes. Because of their diverse applications, these fiber types

are available in finenesses and cut lengths from 1.7 dtex (38 mm) to 13 dtex (150 mm).

2.3.2. Continuous-Filament Yarns for Apparel and Home Furnishings

Special Features of the Manufacturing Process. Poly(ethylene terephthalate) that is to be spun into continuous-filament yarn for apparel generally exhibits an intrinsic viscosity [η] of 0.70 – 0.75 [37]. It must be highly purified, especially when used in high-speed spinning, which is often the case. The polymer must also be dried carefully to $< 0.01\,\%$ H_2O.

As in staple fiber units, the spinning positions of a spinning machine are aligned in a row. Quenching is mostly done by a transverse air flow with several spinnerets in a single shaft so that all filament bundles are affected uniformly. At the lower end of the spinning shafts, the filament bundles are individually combined into yarns while finish is applied to them. Finishing with a perforated yarn guide has been generally adopted. Below the spinning shafts, the yarns are usually drawn off by godets, the circumference and revolutions of which determine the spinning speed; the yarns are then wound individually on bobbins [37].

Originally, spinning speeds were 500 – 1500 m/min. The as-spun yarn was amorphous and characterized by an orientation birefringence Δn of $(3-10)\times 10^{-3}$, a so-called *low orientation yarn* (*LOY*) [53], [95]. The as-spun yarn was drawn by passing it, with friction, over heated pins in the draw zone and then additionally, with surface contact, over a hot plate. This process was replaced by drawing on heated rolls with several wraps [37]. For heat-setting, the delivery roll of the set of draw rolls was also heated but at considerably higher temperature. Final drawoff was by a third pair of rolls. The drawn yarn was wound on a twister (draw twisting) or on a winder (draw winding); frequently it was then twisted, heat-set, and rewound. Such units, which involve slow spinning and separate drawing, are still operating but are probably no longer being built.

Improvements in winding frames for drawn yarn led to the next development, spin drawing . In this process, a set of draw rolls is mounted below each spinning shaft. The as-spun yarn is drawn down at ca. 700 m/min and finally wound at ca. 3000 m/min. Winders developed for high-speed spinning in 1983 permit drawing down at speeds up to ca. 4000 m/min with final windup speeds up to 6000 m/min [87]. This process is highly cost-effective, and product quality is good.

In a variant of the spin-drawing process, the draw rolls are mounted near the spinneret at the point where the filaments have just attained their final fineness and are still warm [96]. This process, too, is rated highly for cost and quality of product.

Spinning speeds have gradually been raised to 2800 – 4500 m/min to increase productivity of the spinning units. The result has been production of *partially oriented yarn (POY)* [53], [95]. These yarns are less sensitive because of their inherent slight crystallinity and require less drawing. Drawing can now be combined with texturing into a single process, with considerable savings in investment and operating costs [97].

Drawoff rolls or godets are not required in high-speed spinning, and drawing-off can be left to the winder; however, special measures must be taken to match the spinning tension to the optimum (mostly lower) windup tension [87].

Yarn spun at higher speed and then drawn takes up dye more readily; because it is more highly oriented, postdrawing is lower and entails less additional crystallinity. Material spun at very high speed is, moreover, said to have an especially loose structure, with fewer large crystalline regions, even after drawing [53].

Commercial high-speed spinning has become possible only through the development of suitable winders, which can also be used for as-spun yarn. They can go as high as 6000 m/min and, for experimental purposes, up to 10 000 m/min [98].

In the speed range of 3000 – 4000 m/min, productivity no longer increases substantially with spinning speed. Hence, yarn spun at 2800 – 4000 m/min has been adopted as the starting material for draw texturing [95]. Therefore, additional investment in increased spinning speed is probably not worthwhile [99]. Efforts have been directed at reducing spin orientation and thus increasing draw ratio and productivity. For example, a heating tube below the spinneret reduces spin orientation [100], as do minor additions to the polymer of low-viscosity polymers

[101] or components with slight cross-linking action [102].

The drawing process is combined with texturing in draw texturing to produce textured yarns from material spun at high speed. In this process, the yarn is drawn in the texturing zone on the first heater of the texturing unit, i.e., between the delivery rolls and the twister. This places stringent requirements on the as-spun yarn. Its structural properties, such as orientation and crystallinity, must be uniform to ensure uniform dyeability of the end product. In addition, the as-spun yarn must not have any defects such as broken filaments. In most cases, the yarn is additionally heat-set on a second heater in the texturing unit with a 7 – 14 % contraction in length. This setting process reduces the crimp elasticity of the yarn, which is then dimensionally stable.

The major problem in the manufacture of textured PETP yarns is uniform dyeability. The many process steps and, hence, variables require special attention during manufacture. Total testing costs are high.

A smaller fraction of PETP filament yarns for apparel is processed without texturing. Flat yarns are obtained by the old process of separate spinning and drawing mentioned previously, as well as by spin drawing. More recently, flat yarns have also been spun at high speed [48], [53]. Several processes, such as high-speed spin drawing and warp drawing [103], [104], can accomplish this economically and with high product quality.

Because PETP yarns spun at 6000 m/min still have flat stress – strain curves, they cannot be used as is [95]. Thermal treatment in the spinning shaft, e.g., a heating zone in the shaft, helps increase their orientation and crystallinity [105], [106]. The stress – strain behavior of this *highly oriented yarn (HOY)* then suffices for several special applications. At still higher spinning speed, the sheath – core effect of orientation and crystallinity can exert a negative influence [53], [107].

Types of Continuous-Filament Yarn, Properties, and Uses. Continuous-filament yarns are products of the man-made fiber industry that can be converted directly to textile fabrics, i.e., without intermediate processing. Whereas staple fibers are produced by first cutting crimped filaments which then must be spun into a yarn, possibly in mixtures with other fibers, in a relatively expensive process, filament yarn is available immediately.

Filament yarn is offered flat, i.e., with silk-like, or variously crimped textures; thus, it provides opportunities to change the hand, bulk, and optical appearance of downstream goods and to accomodate styling requirements in apparel and home furnishings, which far exceed the range attainable with natural fibers [37], [92][93][94].

Partially oriented yarns (POY) have, since the introduction of high-speed spinning technology (see Section 2.1.2), led to successful intermediates for internal fiber production as well as to salable products. These filament yarns are spun and wound at speeds of ca. 3000 – 4500 m/min. They are precursors for pirn drawing (cops drawing), warp drawing (→ Fibers, 3. General Production Technology), as well as warp drawing – sizing or draw texturing. Most POY products with a fineness of 300 dtex and 32, 34, or 36 filaments per yarn (dtex 300 f 32, etc.) are produced for draw texturing, which is the largest end use.

Flat Yarns. For direct textile processing in weaving and knitting, suitable flat *fully oriented yarns (FOY)* are available on pirns, bobbins, or as warp beams. They are sold with round, trilobal, triangular, or even octalobal cross sections with an individual filament fineness of 2 – 5 dtex and a yarn fineness of 22 – 167 dtex.

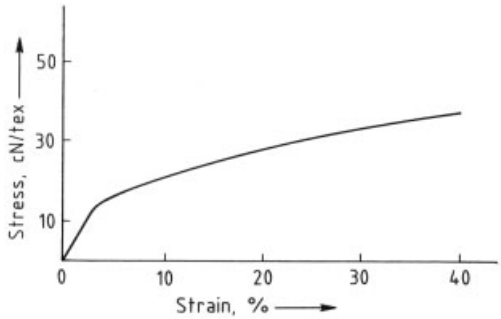

Figure 6. Stress – strain curve of highly oriented filament yarns (HOY)

Pirn (cops) material intended for weaving and warp knitting is produced "classically" with twist from LOY or from POY. Material in bobbins, on the other hand, is used exclusively in weaving and comes primarily from one-step processes such as spin drawing or high-speed spinning (4000 – 6000 m/min) (see Section 2.1.2).

The fineness of individual filaments made by this latter process ranges from 0.6 to 3 dtex. These HOY yarns are distinguished by uniform dyeability. A stress – strain curve, which, in contrast to standard yarn, exhibits higher tensile elongation, is shown in Figure 6. Warp beams, finally, are produced by the warp-drawing process from POY exclusively for warp knitting. For weaving, the warp draw – sizing process is used.

Textured Yarns. The major application for textured yarn, primarily false-twist textured, is in knitting and weaving for apparel. Air-textured yarn, on the other hand, is used predominantly in home furnishings and is processed by weaving.

False-twist textured yarns are draw-textured, starting with POY, mostly by the friction process, but in some cases by the spindle process. Their linear density is 50 – 200 dtex (singles yarn), with an individual filament fineness of 1.25 – 5 dtex. The round cross section of POY filaments is distorted by draw texturing. However, special yarn types are produced with a triangular profile, which is preserved during draw texturing. This type of individual filament cross section imparts a special luster, almost a glitter, to textile fabrics. The linear density of most false-twist textured yarns now used in knitting and weaving is ca. 167 dtex. Fine yarn down to 50 dtex/40 filaments is used in ladies' blouses, for example.

Air-textured yarn can be made from either POY or FOY. With POY, a drawing step precedes air texturing on the texturing machine, whereas FOY is textured directly. In this process the filament cross section, usually round, is preserved. The linear density of such yarn is 200 – 3000 dtex for single yarn, which is appropriate for its primary use in home furnishings. The fineness of individual filaments is 2 – 3 dtex.

Special Types. The extensive product line of continuous-filament yarns includes many special types. Examples of interesting new developments are filament yarns with staple yarn character, i.e., filament yarns with the less regular and bulkier structure of staple yarns, which is achieved by the type and arrangement of individual filaments. In some cases, free fiber ends even protrude from the yarn.

A large variety of properties is possible today. Among the most important are yarns with high bulk character, which is achieved by differential shrinkage of individual filaments. The properties of such yarns, however, are closer to flat yarns. Other yarns are formed by cotexturing different types of individual filaments. These approach staple yarns, even with regard to the comfort of clothing made from them.

Extremely low fineness filaments with individual counts down to 0.1 dtex (produced by splitting thicker filaments) or 0.6 dtex (spun directly) are used in textiles with the appearance and feel of buckskin.

In differential dyeing cationic dyeable yarns are used in addition to normal disperse dyeable filament yarns. When yarns are made from blends of disperse dyable individual filaments and cationic dyeable ones, they may be dyed in different colors.

Flame retardancy in home furnishings, especially in drapes and curtain fabrics, is achieved with filament yarns that consist of modified PETP. By inserting a phosphorus compound into the fiber molecule, the flame retardancy of such yarns is permanent and cannot be removed in washing or dry cleaning.

Finally, filament yarns with permanent antistatic properties are available. In general, however, the surface conductivity required for processing and for some end uses is achieved by applying finishes.

2.3.3. Continuous-Filament Yarns for Industrial Applications

Special Features of the Manufacturing Process. Filament yarns for industrial use are generally spun from higher molecular mass PETP, which is the only way to achieve high strength. For low-fineness, high-tenacity yarn (final fineness < 1000 dtex) PETP with an intrinsic viscosity $[\eta]$ of 0.75 – 0.80 is used; for coarse, high-tenacity yarns (final fineness ≥ 1000 dtex), the intrinsic viscosity $[\eta]$ of PETP is 0.95 – 1.05. Such molecular masses are achieved either by continuous polycondensation with direct spinning or, preferably, by postcondensation in the solid state [37].

During condensation in the melt, COOH end groups, which catalyze hydrolytic degradation of PETP in the melt, are formed in the polymer and can disarrange the structure in the final product. These end groups can be protected with carbodiimides [70] or similarly acting sub-

stances, which are added to the polymer in small amounts prior to spinning. No COOH groups are formed if solid-state condensation is carried out in a vacuum or under a protective gas and at temperatures below the melting point. Care must be taken, however, that the polymer does not absorb water during storage or during transportation from the condensation to the melting equipment.

Polymer with a high viscosity requires a rather high spinning temperature, e.g., ca. 290 °C for fine yarn and ca. 310 °C for coarse yarn [37]. To minimize thermal degradation of the melt, the lines carrying melt from extruder to spinneret must be as short as possible. Also, the melt in the distribution lines after the extruder is kept relatively "cold" and warmed to the higher spinning temperature only in the spinning pack [108].

Fine, high-tenacity yarn has a linear density similar to that of textile yarn. For this reason, it is spun on a similar machine. The yarn is either spun and drawn in separate process steps or spin-drawn continuously. The spinning speed for fine high-tenacity yarn in the two-step process is approximately 1000 m/min and 500 m/min and higher during spin drawing. The draw ratio is ca. 5; the yarn is usually allowed to shrink ca. 10 % during heat-setting at a hot roll temperature of ca. 240 °C. The delivery speed of the drawing machines is ca. 700 m/min, but during spin drawing, it is substantially higher (ca. 2500 – 3000 m/min).

A special, high-volume variant of fine, high-tenacity yarn is sewing yarn. By plying, twisting, and renewed heat-setting, throwsters turn this yarn into sewing thread, whose most important characteristic is the sewing length achievable. The goal is sewing yarn that needs no further plying and twisting to make it comparable in processability to sewing thread of staple fiber.

During spinning of particularly viscous polymers for coarse industrial yarn, very high drawing forces develop in the filaments. These forces can be relieved by delaying filament cooling with the aid of a heating tube directly below the spinneret, which reduces spin orientation and equalizes the structure of the filaments [100], [109]. In this way, noticeably higher final strength can be attained [110].

Coarse high-tenacity PETP yarn is produced either in one step by spin drawing or in two steps by spinning and further treatment as a warp. During spin drawing, the spinning speed is ca. 700 m/min; when spun on bobbins for warp processing, it is ca. 700 – 1000 m/min. Yarn has been spun at much higher speed, in which case the disturbing influences mentioned previously are compensated in other ways [111].

The spin-draw machines for coarse high-tenacity yarn resemble those for fine high-tenacity yarn except that some parts are even more massive [94]. Because of the high operating speed (up to 4000 m/min) the residence time on heat setting rolls and beyond is short. Therefore, preparing low-shrinkage material by spin drawing is difficult; however, the strength that can be achieved is considerable.

In the draw-line process, a warp of yarn is drawn in two steps and then heat-set. After the last set of rolls, which determines the setting shrinkage, the yarn is wound on individual bobbins. The delivery speed of such draw lines is 120 – 300 m/min, with a productivity of 4 – 10 t/d.

The ability of the end product to shrink and its shrinkage force can be reduced, both in spin drawing and in warp processing, by permitting shrinkage in the heat-setting step. This increases elongation of the yarn at loads important in practical applications. The much higher residence time of the yarn in postdrawing and setting in the warp process leads to lower residual shrinkage at equal relative elongation of the end product (relative elongation being the elongation at a given load). Inversely, at equal residual shrinkage, the relative elongation of the product from the warp process is lower. These relationships are influenced further by the polymer used, the spinning process, and the drawing conditions.

In certain end uses, coarse industrial PETP yarn must exhibit good adhesion to rubber. For this purpose, an appropriate adhesive finish is applied during spinning or in warp processing prior to heat-setting. This finish may contain epoxy compounds (Epikote) or similarly acting material.

Industrial Yarns, Properties, and Uses. Of the three product forms used industrially, i.e., high-tenacity filament yarns, spunbonded products, and monofilaments, the filament yarns have by far the broadest application [37], [92][93][94]. This ranges from the building in-

Table 8. Filament yarns for industrial end uses

Types	Shrinkage in hot air at 200 °C, %	Fineness, dtex	Number of filaments	Filament fineness, dtex	Tenacity, cN/tex	Elongation, %
Yarns for sewing threads, normal shrinkage	17 – 18	74 – 455	24 – 96	3.1 – 4.7	68 – 70	12 – 13
Yarns for sewing threads, low shrinkage	7	49 – 940	16 – 200	2.5 – 4.7	68 – 70	16 – 17
Fine yarns, normal shrinkage	18 – 19	280 – 940	48 – 200	3.3 – 5.8	65 – 70	12 – 14.5
Fine yarns, normal shrinkage (spun-dyed black)	19 – 22	550 – 940	50 – 100	9.4 – 11	66 – 69	19 – 22
Fine yarns, low shrinkage	4 – 14	140 – 550	24 – 96	5.8	60	21
Coarse yarns, normal shrinkage	15.5 – 22	1100 – 1670	100 – 200	5.5 – 11	72 – 73	13 – 15.5
Coarse yarns, normal shrinkage (spun-dyed black)	15.5 – 22	1100 – 1670	100 – 200	8.4 – 11	66 – 72	12.5 – 13.5
Coarse yarns, low shrinkage	3.5 – 5.0 (7.5 – 8.5)∗	1100 – 1670	200	5.5 – 8.4	72 (73 – 75)∗	16.5 – 17.5 (11.5 – 12.5)∗
Coarse yarns for tire cord	10 – 18	1100 – 1400	200	5.5 – 7	75	12
Coarse yarns (plied fineness)	22	2200 – 11 000	400 – 2000	5.5	other values depend on single yarn	

∗ With adhesive finish.

dustry through machine and motor vehicle construction, chemistry, mining, agriculture and forestry, transportation, sporting goods, protection of the workplace and the environment, to apparel and home furnishings. Such yarns are offered on the market in three major categories: sewing yarn, fine yarn, and coarse yarn. Depending on end use, these yarns are employed either as normal-shrinkage or low-shrinkage types. In addition to fine and coarse white yarns, spun-dyed black types are also available. The extraordinarily broad product line is rounded out by plied fineness of coarse filament yarns. For an overview see Table 8.

Sewing Yarns. The value of sewn textiles, particularly their appearance and durability, depends to a large extent on seam performance. To best complement textile design and construction, the seam must meet many requirements, the most important of which are neat appearance, high tenacity and abrasion resistance, high lightfastness (especially to ultraviolet radiation), resistance to chemicals, and washfastness. In addition, good sewability is significant for the sewing yarns; this is defined as achieving a high sewing length at high sewing speeds without yarn breakage. All these requirements are met or exceeded by fine denier, high-tenacity PETP filament yarns. They have performed well alongside PETP staple yarns and PETP stretch-broken yarns in 100 % PETP sewing yarns as well as in core-spun yarns.

Sewing Yarns with Normal Shrinkage. Filament yarns with normal shrinkage are suitable for 100 % PETP sewing yarns, which must be heatset (Table 8). Sewing threads prepared from them are used especially in high-performance areas, e.g., safety belts, cushions, leather shoes, luggage, and handbags.

Sewing Yarns with Low Shrinkage. Low-shrinkage filament yarns are offered in a broad range of finenesses (Table 8). The largest end use is in core-spun yarns. By spinning a sheath of cotton, viscose, or other polymeric staple fiber around a PETP filament core, the properties of both fiber types are combined. The sheath assures sewability; the core, mechanical efficiency. Such yarns meet high requirements of aesthetics, dyeability, and temperature resistance.

Fine Filament Yarns. Both normal- and low-shrinkage fine filament yarns are available for

general use. Filament yarns with normal shrinkage are offered as white or spun-dyed black types. Both types are preferred in lashing straps, tapes, belts, ropes, nets, and fire hoses.

Low-shrinkage fine filament yarns are offered only in white (Table 8). They have a wide range of applications, extending from base fabrics for PVC or rubber coating for tarpaulins, membranes, containers, awnings, and tent roofs, through conveyor belts, V-belts, hoses, filter fabrics, and drying screens for paper making to sailcloth. Its use in sailcloth demonstrates the ability of PETP to meet the requirements of low mass, high tenacity, and uniformity, as well as resistance to UV light, chemicals, water, and mildew.

Coarse Filament Yarns. Coarse yarns, like fine yarns, are offered with either normal or low shrinkage (Table 8). Yarns with normal shrinkage are available in white or spun-dyed black versions. White yarns, analogous to the corresponding fine yarns, are intended for lashing straps, belts, ropes, nets, hoses (especially pressure hoses), and safety belts. The situation is quite similar with respect to the black types, which are used primarily in safety belts and lashing straps.

Low-shrinkage coarse filament yarns, which are offered exclusively in white, have equally wide-ranging applications, such as base fabrics for PVC or rubber coating for tarpaulins, membranes, containers, awnings and tent roofs, hoses, filter fabrics, belts, conveyor belts, V-belts, and drying.

An especially interesting end use for low-shrinkage, high-tenacity, adhesively finished filament yarns is in automobile tires. The properties of such yarns make them highly suitable for the carcass of radial tires. Tires with these yarns no longer exhibit the disturbing phenomenon of flat spotting, observed in nylon-reinforced tires. (Flat spotting is the formation of an elliptical footprint in the tire during standing, resulting from molecular deformation of nylon because of its low glass transition temperature. It causes a bumpy ride until the tire warms up.) Since the early 1960s, much effort has been expended to solve the problems of PETP tire-cord material. In addition to rubber adhesion, these included the hydrolytic stability of PETP. High-tenacity filament yarns for tires are rather low in shrinkage. They have high moduli, exhibit high flexural strength, and thus produce little heat during mechanical loading. They, therefore, have all the prerequisites to utilize fully the high chemical stability (resistance to hydrolysis) provided by extremely low concentrations of COOH groups. Solid-phase polycondensation combined with chemical capping of the majority of any still-existing COOH groups (see Section 2.2.1) makes possible the high stability of the tire cord.

The high-tenacity, coarse-filament single yarns discussed so far are complemented for many end uses by a series of plied yarns. These include some that are adhesively finished for use with rubber, others that have low-shrinkage like the single yarns, and still others that have normal shrinkage. Their finenesses extend up to 11 000 dtex. These yarns are intended for use in the coating sector as well as in straps, belts, ropes, nets, and hoses.

2.3.4. Spunbonds

Special Features of the Manufacturing Process. Spunbonded products are consolidated random layers of 2- to 10-dtex continuous filaments [112]. They are produced as continuous sheets up to 5 m wide in a continuous process from polymers of average intrinsic viscosity ($[\eta]$ ca. 0.7). Their unit weights range from 20 to 500 g/cm^2 (→ Fibers, 3. General Production Technology).

A spunbond production line usually consists of several rows of spinnerets with mutually displaced spinning packs. Each filament bundle is accelerated by a draw tube under the spinneret to a final speed of $\geq$ 5000 m/min and thus is drawn [113]. Spreading devices open the filament bundle and lay the filaments down on a moving belt in overlapping random patterns. Various designs are known for these spreading devices; i.e., mechanical, pneumatic, and electrostatic [112].

Properties of the filaments are the same as those of filament yarns spun at high speed: tenacity ca. 3.5 cN/dtex, break elongation ca. 70 %, and shrinkage at 200 °C ca. 3 %. The orientation is characterized by a birefringence Δn of ca. 100×10^{-3}, and the degree of crystallinity is ca. 40 %.

The loose filament web, laid down as uniformly as possible, is then consolidated by using various processes singly or in combination, such

as needle punching, calendering, and bonding. For certain applications, the web must be made thermally stable by additional heat-setting.

Spunbonded Products, Properties, and Uses. Polyester spunbonds are available in unit weights of 20 – 500 g/m^2 and in widths of 1 – 5 m. The consolidation discussed in the preceding section via needle punching, bonding, or calendering requires, in addition to PETP and suitable binders, the use of polyesters of lower melting point or even of PETP fibers with a sheath of a different polymer. Polyester spunbonds are offered commercially under various *trade names:*

Asahi	Asahi Kasei
Colbond BV	Coltron
EniChemFibre SpA	Terbond-Spun
Hoechst	Trevira-spunbond
Lutravil	Lutradur
Reemay, Inc.	Reemay
Rhône-Poulenc	Bidim
Teijin	Unisel
Toray	Axter
Toyobo	Toyobo Volans
Unitika	Marix

Lightweight webs (20 – 50 g/m^2) are available with individual filament counts of 2 – 3 dtex. They are used as cover webs in sanitary products and in horticulture.

Nonwoven products (50 – 130 g/m^2) have filaments with individual finenesses of 4 – 5 dtex. Their end uses are in construction and in shoe caps.

Materials with unit weights of 130 – 500 g/m^2 and individual filament finenesses up to 10 dtex are intended especially for use in construction (e.g., carriers for asphalt sheets for roofing), needle felts, embossed PVC coatings, underground construction, hydraulic engineering, and road building.

In addition to their large use in construction, polyester spunbonds are already being used successfully in automobiles (e.g., trunk linings) as well as in protective clothing and for environmental protection (covering of storage basins).

Web properties depend predominantly on the often different technologies for web production (laydown and bonding). No systematic data exist up to now that relate web structure to various end uses. However, in general, needle-punched webs are bulky, whereas thermally bonded webs are thin. Web strengths are ca. 2 – 4 N g^{-1} m^{-2} throughout (measured on strips 5 cm wide). Break elongations are 60 – 100 % for needle-punched webs, 20 – 60 % for bonded and needled webs, and generally < 30 % for thermally bonded webs.

2.3.5. Monofilaments

Special Features of the Manufacturing Process. In the production of PETP monofilaments, polymers with intrinsic viscosities $[\eta]$ of 0.7 – 0.9 are used. Polymers of higher viscosity are often brought to their required value by solid-phase condensation. Because these polymers then have a lower content of COOH groups, they have advantages for use in hydrolysis-resistant monofilaments with increased hot – wet durability. Occasionally, the COOH end groups of the polymers are also protected chemically.

Monofilaments with final thicknesses of 0.03 – 2 mm are spun in air, cooled immediately in water, then drawn, heat-set, and packaged on a directly connected draw line [114]. Each extruder usually has its own spinning head and, accordingly, its own draw line. Various designs are recommended for spinning heads and spinnerets, such as circular spinnerets and spinnerets with straight rows of holes [115]. Feed for each spinneret hole should be the same. A few centimeters below the spinneret, the filaments enter a water bath, whose temperature is above the glass transition temperature of PETP (70 °C). The filaments are cooled but remain soft. In the bath, they change direction, are brought out of the bath as a warp sheet, and are drawn off by a set of rolls at a rate of less than 100 m/min. The filaments have no spin orientation.

The first draw almost always takes place in water at 90 – 95 °C; the second draw, also takes place in a water bath, a higher boiling liquid, or a heater tunnel. Die plates are sometimes used for more accurate adjustment of the thickness. Heat-setting occurs in a heater tunnel, usually at 130 – 180 °C. Depending on the type desired, more or less shrinkage is permitted and the total draw ratio (ca. 1 : 4 to 1 : 6) is selected. At the end of the roll train, the monofilaments are taken up on individual bobbins on a creel.

In almost all applications of PETP monofilaments a uniform diameter is the most important mark of quality. Tolerance of ± 3 % is re-

quired; if the process is carefully controlled, it will often be less. Additional quality requirements are smooth surface and, if necessary, uniform cross section. Textile properties of a given type must always remain constant, especially shrinkage tension and shrinkage. In monofilaments with improved resistance to hydrolysis, the residual strength after prolonged use should be preserved uniformly.

Types of Monofilaments, Properties, and Uses. Monofilaments are produced with diameters of 0.034 – 2.00 mm These are thick filaments (the finenesses corresponding to these diameters are 12.5 – 43 000 dtex) and are, therefore, often called wires. A broad spectrum of types of PETP monofilaments is found commercially in woven fabrics and zippers. These end uses place strict requirements on uniformity of diameter and properties, and especially on stress – strain behavior and shrinkage. The mechanical properties of typical PETP monofilaments are evident from the stress – strain curves in Figure 7.

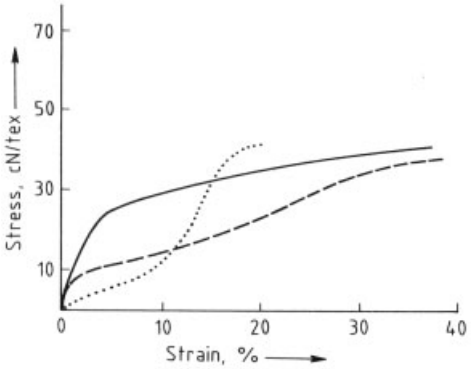

Figure 7. Stress – strain curves of typical PETP monofilaments
——— Finest monofilament (0.055 mm diameter)
– – – Monofilament for weaving (weft type)
····· Monofilament for zippers

Monofilaments for weaving are used in the building industry in interior construction and interior installation. Other, often large, areas of application are paper machine sieves, safety belts, filter and screening fabrics, as well as conveyor belts. Fish nets, cordage, fishing lines, and medical applications all require the special properties of monofilaments for weaving.

The finest monofilaments, with diameters of 0.034 – 0.055 mm, have high elongation (see Fig. 7) and, at the same time, high shrinkage. Coarser wires range from 0.080 – 2.00. In addition, some wires have a rectangular cross section with diameters from 0.25 and 0.50 mm to 0.57 and 0.88 mm.

Monofilaments for weaving, especially for paper machine sieves are available as normal, hydrolysis-resistant, and chemically stabilized types. The most important representatives have high elongation (wires for the weft) or low elongation (wires for the warp) but low shrinkage.

Another important group comprises *zipper monofilaments*, with a diameter of 0.40 – 1.16 mm. This competes with polyamide wire. A major advantage with PETP monofilaments is that their stress – strain properties and shrinkage behavior can be varied to a great extent. Thus, zipper wires are available over a whole gamut of combinations from low, elastic extensibility with high shrinkage to high, plastic extensibility with low shrinkage. The majority of these types exhibit low, elastic extensibility and medium shrinkage at the boil.

In general, the combination of modulus and shrinkage properties of zipper wire and wires for weaving is determined by the special manufacturing process. It is chosen according to the load – time program and the thermal exposure of monofilaments during use. The ability to vary their properties so widely makes possible the adaptation of PETP monofilaments to any kind of processing and to any end use.

2.4. Other Polyester Fibers

2.4.1. Poly(tetramethylene terephthalate)

The polyester of terephthalic acid and 1,4-butanediol [*110-63-4*] is a comparatively old fiber polymer [40].

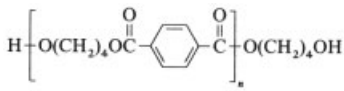

Because 1,4-butanediol has always been more expensive than ethylene glycol, poly(tetramethylene terephthalate) [*30965-26-5*], also called poly(butylene terephthalate) or PBT, remained more or less on the sidelines. Not until the middle of the 1960s

were special PBT products developed by several producers; the advantages of these products compared to PETP permitted use of the more expensive 1,4-butanediol [116]. Today, these PBT products have become well-entrenched in the marketplace.

Production of PBT is similar to that of PETP. Dimethyl terephthalate is transesterified and subsequently polycondensed, either batchwise or continuously. However, the catalysts used with PETP promote ring closure of 1,4-butanediol, leading to the potentially explosive tetrahydrofuran. Therefore, tetraisopropyl titanate [*546-68-9*], for example, is used as catalyst in both process steps.

Because the aliphatic sections of the molecular chains are longer and thus more mobile, the transition temperatures of PBT, *mp* ca. 230 °C and glass transition temperature ca. 25 °C, are lower than those of PETP. In addition, PBT crystallizes more readily than PETP. The PBT unit cell exists in two modifications. Of these, one is more extended, ca. 11 % longer, and reverts to the shorter one when unloaded. This is responsible for the better recovery of PBT products from stretching and bending.

Like PETP, PBT is spun on standard machines from the melt. Because of the lower glass transition temperature, formation of spin orientation not only occurs along the spinning path but continues into the drawoff and the windup. In addition, clearly recognizable crystallization of PBT occurs during spinning, which in turn also affects orientation. Properties of the as-spun yarn thus depend in a complicated manner on spinning conditions.

After the usual drawing and heat-setting, PBT has a flatter stress – strain behavior than PETP (Fig. 8). The permanent extension of a PBT product after straining, however, is much less (Fig. 9) [116]. For example, after a strain of 15 %, the permanent extension of PETP is 8.5 %; that of PBT, only 2.5 %. For this reason, PBT is particularly suited for carpet fibers, which must exhibit good recovery after loading. Unlike PETP, PBT can be dyed easily to deep shades at the boil without carrier because of its lower glass transition temperature.

In addition to its elastic behavior and easy dyeability, PBT has other favorable properties. For example, it tends to swell less than PETP. All these factors, taken together, make PBT an interesting raw material for textile zippers. Finally, because of their resistance to alkali, good abrasion resistance, and elastic behavior, PBT monofilaments are advantageously used in screen fabrics, which are constantly exposed to hot water or slightly alkaline media. In this way, fibers for carpets and monofilaments for zippers and screens have become products in which the more expensive polymer (PBT) pays for itself.

2.4.2. Poly(1,4-dimethylenecyclohexane terephthalate)

Fibers of poly(1,4-dimethylenecyclohexane terephthalate) [*25135-20-0*], PDCT, have been on the market since the late 1950s under the trade name Kodel II (Tennessee Eastman Company, United States) (see below).

The intermediates are dimethyl terephthalate (DMT) (Section 2.1.1) and 1,4-bis(hydroxymethyl)cyclohexane [*105-08-8*] commonly known as 1,4-cyclohexanedimethanol. This diol is produced by the two-step hydrogenation of DMT at 160 – 180 °C and high pressure in the presence of catalysts (→ Alcohols, Polyhydric, Chap. 2.7.). The diol can exist as either the trans or the cis isomer. By appropriate choice of reaction conditions, the trans – cis ratio can be adjusted to 7 : 3. To prepare the polymer, DMT is transesterified with the diol and after excess diol is distilled, polycondensation takes place in a manner analogous to PETP.

The polymer is transparent and amorphous. It crystallizes in the triclinic system. The crystalline density is 1.265 g/cm^3 for polymer with trans isomer and 1.303 g/cm^3 for polymer with cis isomer [117], [118]. Wide-angle X-ray diagrams of crystalline samples show a continuous transition of the trans crystal lattice to the cis crystal lattice with increasing cis – trans ra-

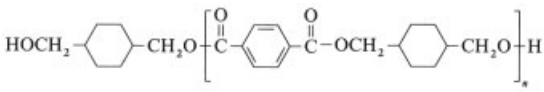

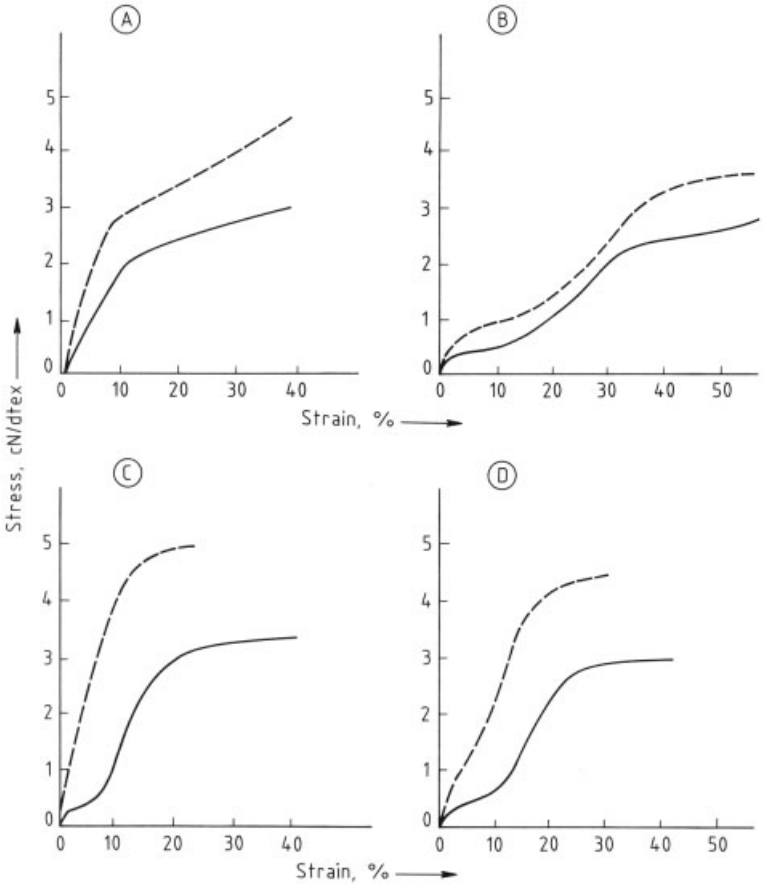

Figure 8. Stress – strain curves of drawn and heat-set PETP and PBT yarns [116]
A) Drawn; B) Drawn and heat-set without clamping; C) Drawn and heat-set with clamping; D) Drawn, heat-set with clamping, plus heat-set without clamping
– – – – PETP; —— PBT

tio. The fiber density is 1.22 – 1.23 g/cm^3. Small-angle X-ray investigations show a fibrillar structure with lattice layers at an angle to the fiber axis.

The glass transition temperature of the polymer is 76 °C; in drawn and crystalline fibers, it can be as high as 100 °C. The melting point is 285 – 295 °C [92].

Compared to PETP, the tenacity of PDCT staple fibers is ca. 30 % less, with the same elongation at break. The abrasion resistance behaves similarly. The tensile or bending recovery of the fibers from equal deformation at room temperature, on the other hand, is greater for PDCT because of the higher glass transition temperature of the polymer [118]. The result is greater crimp stability of PDCT fibers and, hence, greater bulk of nonwovens and textile fabrics. Because of the higher glass transition temperature, however, PDCT fibers are more difficult to dye than PETP fibers. Therefore, the dyeing temperature for PDCT fibers must be increased or suitable carriers and dyes chosen.

Applications of PDCT fibers are in areas such as fiberfill and carpets that require especially high crimp stability and generally high recovery forces following deformation at room temperature. These fibers are also used in blankets and synthetic furs [118].

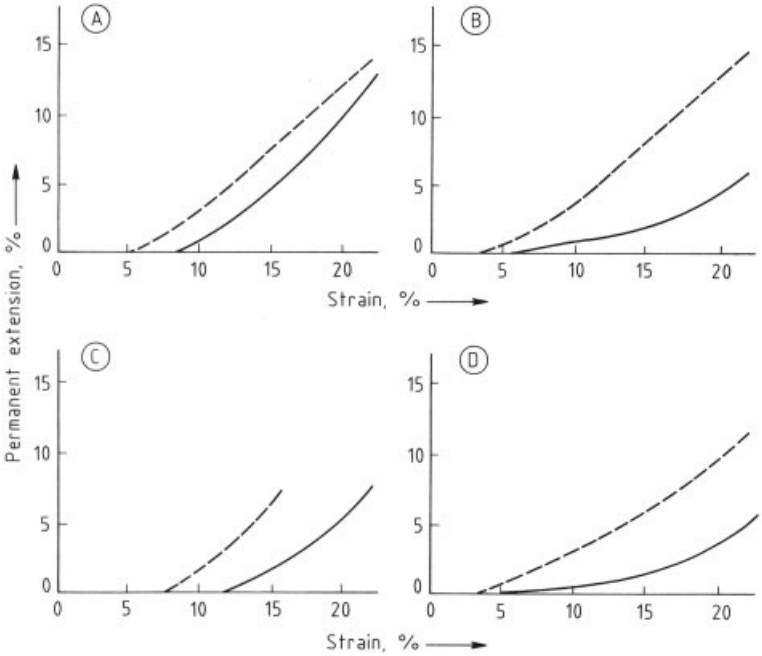

Figure 9. Permanent extension of PETP and PBT yarns after strain [116]
A) Drawn; B) Drawn and heat-set without clamping; C) Drawn and heat-set with clamping; D) Drawn, heat-set with clamping, plus heat-set without clamping
– – – – PETP; —— PBT

2.5. Uses

Staple fibers, continuous-filament yarns, monofilaments, and spunbonds of polyester are found in all major textile markets. Among the polyesters, poly(ethylene terephthalate) is the most important polymer. Apparel, such as outerwear or sportswear, as well as work clothing and protective clothing, could not be imagined without PETP fibers, either 100 % or in blends with other man-made or natural fibers. Similarly, PETP fibers are widely used in home furnishings, such as drapes and curtains, furniture covers, table and bed linens, cushions, pillows, and carpets.

Polyester products have been particularly successful in industrial markets. Here, they are encountered everywhere fibers can be used. In many cases, because of their properties, especially their relatively low density (ca. 1.38 g/cm^3), they have provided elegant and, at the same time, economical solutions to problems. Examples are sewing thread, tarpaulins and awnings, safety belts, V-belts, tire cord, zippers, woven screens, geotextiles, roofing sheets, or book covers.

The extremely high performance of PETP is reflected in the breadth of end uses [37], [59], [69], [92, 93], [119], [121]. This performance is based primarily on two properties of the fiber products: the dimensional stability of the textiles made from PETP under mechanical and thermal stress, and their ease of care. Both properties are primarily determined by the glass transition temperature which, depending on the chemical structure of the polymer and the degree of orientation and order, lies between ca. 70 and 100 °C (see Section 2.2.1, → Fibers, 2. Structure). The fibers can be crystallized at any stage of their manufacture and during further processing. This permits production of any desired shape.

Polyester fibers are available in all possible forms in which natural and man-made fibers exist and can be processed on equipment used to manufacture yarns, woven fabrics, knit goods, and even nonwovens. Grouped broadly, the forms of PETP are the following:

1) Staple fibers with a fineness of 0.4 – 20 dtex; short cut fibers, 6 mm length, fibers of 30 – 60 mm up to 150 mm; round or modified cross sections; crimped (except for flat cut fibers).
2) Textile continuous-filament yarns with a fineness of 20 – 200 dtex; individual filament fineness of 0.5 dtex (special types 0.1 dtex) to 5 dtex; flat, textured, or with staple yarn character; round or modified cross sections. The fineness of air-textured, continuous-filament yarns is 200 – 3000 dtex.
3) Fine high-tenacity continuous-filament yarns and yarns for sewing threads with a fineness of 50 – 940 dtex, with individual filaments of 3 – 10 dtex; flat with round cross sections.
4) Coarse high-tenacity continuous-filament yarns with a fineness of 1100 – 1670 dtex (plied fineness up to 11 000 dtex); individual filaments of 5 – 11 dtex; flat with round cross sections.
5) Spunbonds in widths of 1 – 5 m; unit weights of 20 – 500 g/m^2, with an individual filament fineness of 2 – 10 dtex.
6) Monofilaments with a diameter of 0.034 – ca. 2 mm; flat with round or modified cross sections.

2.6. Economic Aspects

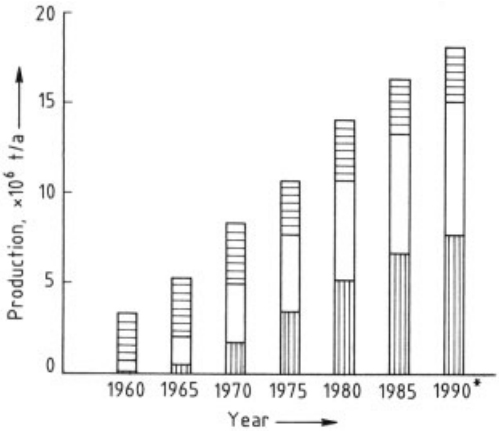

Figure 10. World production of man-made fibers
Vertically hatched area = polyester fibers; Plain area = other synthetic fibers; Horizontally hatched area = cellulosic fibers
∗ Estimated.

World production of fibers increased from ca. 14.9×10^6 t/a in 1960 to 34.8×10^6 t/a in 1985. Natural fibers, essentially cotton and wool, contributed ca. 7.6×10^6 t/a of this increase, whereas man-made fibers contributed ca. 12.2×10^6 t/a [41], [120]. Production of man-made fibers has increased fivefold from 1960 to 1985. Because the fraction contributed by cellulosic fibers decreased, the increase in production of man-made fibers can be ascribed solely to synthetic fibers (Fig. 10).

From small beginnings prior to 1960, polyester fibers, essentially poly(ethylene terephthalate), attained a volume of 6.5×10^6 t/a by 1985. Their percentage of the production and consumption of all synthetic fibers tripled between 1960 and 1985 (Fig. 11) and is now almost 50 %.

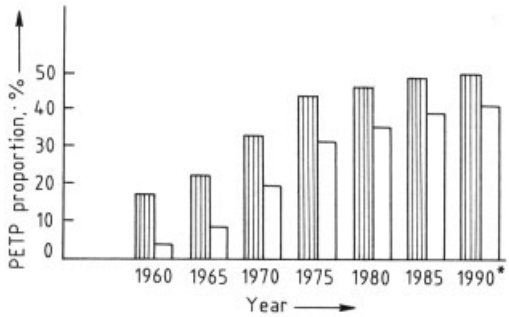

Figure 11. Production of polyester fiber compared to all chemical fibers and to synthetic fibers
Plain area = PETP proportion of all chemical fiber production; Vertically hatched area = PETP proportion of synthetic fiber production
∗ Estimated.

The two most important fiber forms are staple fibers and filament yarns. For *staple fibers,* including tow and nonwovens (Fig. 12), the production of cellulosic fibers remained approximately constant between 1965 and 1985. During this time, production of polyamide and polyacrylonitrile fibers, the latter predominating, increased sixfold. The increase in poly(ethylene terephthalate) is about twice as great, i.e., twelvefold. As far as end uses are concerned, PETP fibers compete with other synthetic fibers only in a few limited areas. They are used mainly in areas previously reserved for natural and cellulosic fibers.

In *continuous-filament yarns,* including industrial yarns and monofilaments (Fig. 13), pro-

duction of cellulosic fibers has decreased. Even before the rise of PETP, polyamide yarns had a secure niche in the market; between 1960 and 1985, their production increased approximately threefold. The production of polyester yarns increased twentyfold to equal that of polyamide yarns. Little competition exists between the two; if any, it is in the industrial area. Here, too, polyester continuous-filament yarns have primarily penetrated areas previously dominated by cellulosic continuous-filament yarns or by spun yarns of natural fibers.

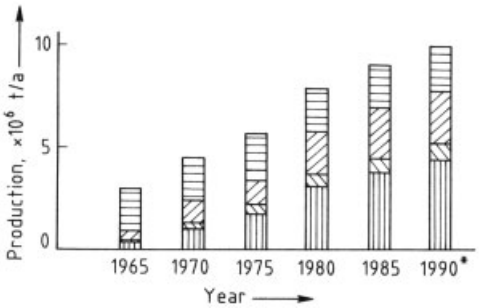

Figure 12. World production of staple fibers (key fiber materials)
Vertically hatched area = polyester; Diagonally hatched, inclined to the left = polyamides; Diagonally hatched, inclined to the right = polyacrylonitrile; Horizontally hatched area = cellulose
∗ Estimated.

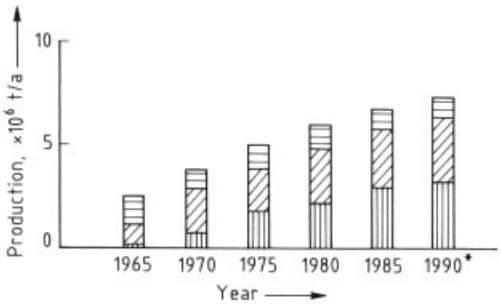

Figure 13. World production of continuous-filament yarns (key fiber materials)
Vertically hatched area = polyester; Diagonally hatched, inclined to the right = polyamides; Horizontally hatched area = cellulose
∗ Estimated.

Understandably, a vast amount of industrial development work and capital was required to build such enormous PETP fiber production. Sample figures are available for the early years of the PETP fiber, Terylene [38]. Large capital reserves were needed to maintain production during occasional wide swings in the market. For this reason, only large chemical companies were and are in a position to operate PETP fiber plants. A selection of trade names and producers follows [119]:

Allied Polyester (Allied, United States)
Dacron (Du Pont, United States)
Diolen (Enka, Federal Republic of Germany)
Fortrel (Celanese, United States)
Grilene (Emser Werke, Switzerland)
Grisuten (German Democratic Republic)
Kodel (Tennesee Eastman Co., United States)
Tergal (Rhône-Poulenc, France)
Terital (Montefibre, Italy)
Terlenka (Enka, The Netherlands)
Tersuisse (Viscosuisse, Switzerland)
Terylene (ICI, United Kingdom)
Teteron (Teijin, Toray, Japan)
Trevira (Hoechst, Federal Republic of Germany, United States)
Wistel (Snia Viscosa, Italy)

2.7. Polyester: Summary and Future Prospects

Properties. The following properties are responsible for polyester fibers being used in all areas where fibers are applied:

1) the extremely broad range of mechanical properties of staple fibers and filament yarns,
2) their variable and easily adjustable shrinkage properties,
3) their ease of dyeing combined with high fastness properties,
4) their high chemical, solvent, and UV resistance,
5) the low flammability of even the standard polymer, which can be further improved by modification without significantly impairing other valuable fiber properties,
6) the deformability of the fibers and the fabrics made from them,
7) the ability to make these deformations (e.g., crimp, creases, and pleats) permanent,
8) the low moisture uptake and, hence, rapid drying,
9) the fact that most soil remains on the surface where it is easily removed with water or solvents.

These properties are largely insensitive to the presence of water at moderate temperature. Together, they are responsible for the dimensional stability and easy care of textiles made from polyester fibers.

Comparison with Other Man-made Fibers. Polyester fiber products, especially those from PETP, are used commercially in large amounts in all common fiber forms: staple fibers, filament yarns for textile and industrial applications, spunbonded products, and monofilaments. The plethora of products and broad spectrum of applications are not nearly the rule for other major synthetic fibers, e.g., nylon 6 and 66 polyamides or polyacrylonitrile.

Both nylon 6 and nylon 66 filament yarns are used worldwide, especially in textured yarns and as tire cord, as are nylon spunbonds and monofilaments. The same is true for staple fibers in carpets. However, polyamide fibers are not easy to heat-set, which limits their application in apparel mainly to woven and knit goods.

With polyacrylonitrile, the product with the highest volume is staple fiber. They are used in apparel, especially in knitting, and in home furnishings in weaving, as well as in carpets. The strong dependence of mechanical properties on the presence of moisture at elevated temperature (poor dimensional stability) makes these fibers unattractive for woven fabrics in apparel. Filament yarns and fibers for industrial uses are still specialty products. To some extent, however, they are very successful, for example, as asbestos replacement in fiber-reinforced concrete. Polyacrylonitrile monofilaments or spunbonds are not produced because of the problems presented by solution spinning (coagulation and residual solvent).

Polypropylene, like polyester, is available in all types of fiber and nonwoven forms. However, the relatively low temperature resistance and lightfastness, as well as shortcomings in dimensional stability under permanent load, limit the use of this polymer to special areas of apparel, home furnishings, and industrial applications. Dyeability problems are forcing the use of spin dyeing in the apparel and home furnishings sectors.

Future Process Development. Polyester fiber polymer for staple fibers, filament yarns, spunbonds, or monofilaments will continue to be made by condensation. It will, likewise, continue to be melt-spun through spinnerets; the as-spun material will be drawn in some manner; and the drawn material will be posttreated according to end use. Processes, that differ in principle from this scheme have not shown up. However, individual efforts are underway in many locations to further improve the quality of intermediate and end products, to make processes more economical, and to adapt the products to the frequently changing and sometimes even entirely new needs of downstream processors and end users [94].

For example, attempts are underway to improve *polymer uniformity* from plant to plant and over long periods of time by careful process control, e.g., via automation and self-regulation. Equipment changes and variations in the catalyst system as well as in the temperature – vacuum program should improve the economics of polymer production. The key to this is increasingly combining continuous polycondensation with spinning, i.e., direct spinning. This combination of process steps, not only here but in the entire production of polyester products, will do much to improve the quality and manufacturing cost of the end product. Finally, various polymer additives will impart novel properties to the end products.

The uniformity and purity of *staple fibers* are being improved by changes in plants, e.g., in the manner of quenching during spinning and in the machinery of the draw line. This work will also improve the economics of the entire process. Additionally, staple fibers and tows must be adapted to the constantly evolving processes and plants of downstream processors. These adaptations involve mainly crimp and finish. The changing market, which often differs from country to country, requires manufacturers to alter the balance of properties of their products accordingly. This will necessitate variations in machine settings and in the equipment of the draw lines. A distant goal of staple fiber process development is to spin at high speed and to draw and posttreat in one step.

Purity and especially uniformity of dyeing also head the list of quality requirements for *filament yarns for apparel and home furnishings.* They are being improved constantly by careful process control and automation. Pro-

ductivity should be increased by even higher spinning speeds, although considerable physical fiber problems and equipment problems remain to be solved. Work is underway in several locations.

With *filament yarns for industrial use*, the question is to adapt strength and shrinkage behavior permanently to the requirements of each downstream processor and end user, and to explore additional end uses. This work involves the polymer as well as the plants and processing conditions. It entails improving the economics of polymer production, spinning, and processing of as-spun yarn.

In *spunbonded products,* uniformity of textile properties and uniformity of thickness are important. They can be improved by careful monitoring and automation of the process. Constructional work on spinning, drawing, and laydown equipment should result in both better and lower cost nonwovens. The greatest effort, however, is to open even broader and different end use areas to spunbonds by varying such web properties as filament fineness, web thickness, unit mass, and type of bonding. Because spunbonds represent a relatively new type of product, this work is likely to be successful.

Monofilaments require as much uniformity as possible with respect to textile properties and thickness in all applications. Work is, therefore, underway continuously to improve the process by equipment changes and monitoring. Furthermore, much work is being done on varying the textile properties and the monofilament polymer to open up new end uses.

Future Applications. Based on the success of polyester *staple fibers* as a component of blends with wool, cotton, and other cellulosic and man-made fibers, their use especially in blends and also in 100 % articles is expected to increase. This expectation is supported by the outstanding wear comfort of clothing made from these mixtures [121]. Polyester fibers cause moisture given up by the skin to be transported away, through the clothing to the outside. New types of fibers with voids will reinforce this transport mechanism in the clothing and fiberfill sector and will permit further improvements in comfort. New developments, especially in dry-laid nonwovens, will open up further end uses for polyester fibers. In so doing, consolidation by binders will be replaced more and more by thermal bonding. The use of polyester carpet fibers is expected to broaden constantly because of the development of new fiber types and carpet constructions.

Development work will focus on new types of *filament yarns for apparel and home furnishings*, particularly further improvement in aesthetics and wear comfort. In silklike textiles, the trend to ever finer yarns and individual filaments, combined with variation in filament cross section, will continue. In heavier fabrics, however, filament yarns with the character of staple yarns will continue to find new markets.

High-tenacity polyester *filament yarns for industrial end uses* with optimum all-around balance of properties and relatively low cost will continue to be available between standard textile yarns and special high-modulus fibers for constantly growing, as well as new, areas of application.

Spunbonded products based on polyester will probably be used in the future in composite systems, to an even greater extent than now. Contributing to this will be further exploration of the possibility of widely varying their structure and their properties.

A considerable expansion of end use areas for *monofilaments,* especially into the industrial sector, is expected, based on the further replacement of other fiber and natural products, especially metals.

Future Economic Aspects. Figures 10, 11, 12, 13 present estimates of world production of man-made fibers in 1990. Overall, an increase is expected, which, in absolute terms, is only slightly less than that between 1980 and 1985 (Fig. 10). The fraction of polyester fibers as a percentage of synthetic fibers and of all man-made fibers will continue to grow, but no longer as rapidly (Fig. 11). Nevertheless, the fraction of polyester fibers is expected to reach 50 % of all synthetic fibers.

Among the staple fibers, polyester will increase somewhat more than products of other polymers (Fig. 12), while filament yarns of polyester will remain about even with those of polyamides (Fig. 13).

Table 9 shows the average yearly increases in consumption of all polyester fibers from 1985 to 1990 and beyond, as a whole and divided into in-

dividual fiber types. Staple fibers and monofilaments are expected to develop like the average of all polyester fibers, while filament yarns for apparel seem to be approaching saturation. In contrast, well above-average increases are expected for industrial filament yarns and especially for spunbonds. Whereas these two products represent only 5.4 and 1.7 %, respectively, of the volume of all polyester fibers, they are, nevertheless, highly interesting from an economic point of view.

Table 9. Estimated average annual growth in the consumption of polyester fiber and other types of fiber

Fiber	Growth, % per year	
	1985 – 1990	1990 – 1993
Staple fibers and tow	3.2	2.7
Textile filament yarns	2.2	1.5
Industrial filament yarns	4.4	3.4
Spunbonds	8.4	5.3
Monofilaments	3.3	2.4
All polyester fibers	3.0	2.4
Other synthetic fibers	1.6	0.8
All synthetic fibers	2.3	1.6
Cellulosic fibers	0.2	−3.8
All man-made fibers	1.9	0.7
Natural fibers	1.3	1.1
All fibers	1.6	0.9

For comparison, Table 9 also shows the average yearly growth rates of other fiber groups. At least until 1990, the consumption of man-made fibers will increase somewhat more than world consumption of all fibers. Cellulosic fibers are initially expected to remain about the same and then to decrease markedly, possibly because of displacement by competing polyester fibers. Synthetic fibers as a whole will grow much more than all fibers. Natural fibers cannot grow appreciably because little room can be found for expansion of land under cultivation. The increasing worldwide demand for fibers will have to be met to a greater extent by synthetic fibers [119].

Because the remaining synthetic fibers are expected to grow no more than all fibers, the most significant growth compared to other types of fibers will be in polyester fibers. The reason for this is their extreme versatility. By varying the polymer and the processing conditions, polyester fibers can be adapted to the manifold requirements of the marketplace.

Understandably, these estimates are uncertain. The global growth rates will undoubtedly vary from country to country and from sector to sector. They will also be subject to fluctuations in world economy, especially to those affecting the cost of polymer intermediates and the price of cotton. The effects of unpredictable changes in consumer preferences and, hence, in fashions also contribute to the uncertainties of these estimates.

Overall, however, estimates of the average yearly growth in consumption of various fiber types suggest that the producers of polyester fibers will have to continue to improve the quality of their existing products and to open up new areas of application by developing new products.

3. Polyurethane Fibers

Following the discovery in 1937 of the diisocyanate addition polymerization process by O. Bayer and co-workers [122], this principle was applied to the production of polyurethane (PUR) fibers. Hard PUR fibers, first developed to compete with nylon 66, are no longer important. On the other hand, the same principle of synthesis was successfully applied to the manufacture of highly elastic yarns [123], [124].

Synthetic high-polymer fibers containing at least 85 wt % of segmented polyurethane are called *elasthane* [125]; the term *spandex* is used in the United States and in this chapter.

Spandex fibers are characterized by high elongation and elasticity. Like rubber yarns (elastodiene fibers), they belong to the class of elastomeric fibers. Spandex, produced as continuous-filament yarn, is used almost exclusively to manufacture elastic textiles and is always employed in combination with other fibers. Although spandex has gained considerable technical and economic importance, worldwide production, which amounted to 23 000 t in 1985, is relatively small compared to other textile fibers. Commercial products (see Table 10) have distinctive raw materials bases, chemical structures, and manufacturing and spinning processes, which, in the final analysis, are responsible for the specific balance of properties [126][127][128][129][130].

3.1. Physical Properties

Spandex is highly extensible by application of an external force. When the load is removed,

Table 10. Commercial spandex products as of 1985

Trade name	Producer	Country	Composition*	Spinning process
Acelan	Tae Kwang	Korea	polyether/MDI/diamine	
Cleerspan (Glospan)	Globe Manufacturing	United States	polyester/TDI/diamine polyether/MDI/diamine	reaction dry
Dorlastan	Bayer AG	Federal Republic of Germany	polyester/MDI/diamine	dry
Elaston	Chemitex Cellviskoza	Poland		
Espa	Toyobo	Japan	polyether/MDI/diamine	dry
Fujibo Spandex	Fuji Spinning	Japan	polyester/polyether/MDI/diamine	wet
Loobell	Kanebo	Japan	polyether/MDI/diol	from the melt
Lycra	Du Pont	United States Brazil The Netherlands Canada Great Britain Mexico Argentina	polyether/MDI/diamine polyester/MDI/diamine	dry dry
Lynel	Fillatice	Italy	polyester/MDI/diamine	reaction
Mobilon	Nisshinbo	Japan		from the melt
Spinning				
Opelon	Toyo Products Co.	Japan	polyether/MDI/diamine	dry
Roica	Asahi Chem. Ind	Japan	polyether/MDI/diamine polyester/MDI/diamine	dry
Spandaven	Gomelast	Venezuela	polyether/MDI/diamine	wet
Vispan	Elastofibre Pirelli	Italy	polyesterurethane	reaction

* TDI, toluene diisocyanate; MDI, 4,4′-methylenebis(phenyl isocyanate).

the material rapidly and completely reverts to its original state. This is a result of the two-phase structure of the hard and soft segments. The soft segments, which are highly mobile at the use temperature, are coiled in the unloaded state and stretched in the loaded state. The restoring force results from the increase in entropy during reversion to the coiled state (entropy elasticity). The crystalline hard segments of symmetrical aromatic urea groups form a network structure via hydrogen bonds and thus provide form stability [131–144].

Characteristic stress – strain curves of a commercial spandex are shown in Figure 14 in a cyclical loading – unloading experiment (stress – strain cycle). The mechanical properties of spandex filament yarns for the usual range of types and finenesses are listed in Table 11 (see → Fibers, 1. Survey, Chap. 5.).

Table 11. Mechanical properties of continuous filament spandex yarns

Property	Range of values
Fineness	20 – 5000 dtex
Elongation at maximum tensile load	
Standard-elongation types	420 – 570 %
High-elongation types	600 – 800 %
Tenacity	
Referred to original fineness	0.45 – 1.2 cN/dtex
Referred to fineness at break [a]	3 – 7 cN/dtex
Cyclic loading to 300 % elongation	
Modulus, 150 % [b]	0.04[c] – 0.1 [d] cN/dtex
Modulus, 300 % [b]	0.1[c] – 0.25 [d] cN/dtex
Residual elongation	10 – 30 %

[a] Fineness at maximum textile elongation.
[b] Referred to original fineness.
[c] High-elongation types.
[d] Standard-elongation types.

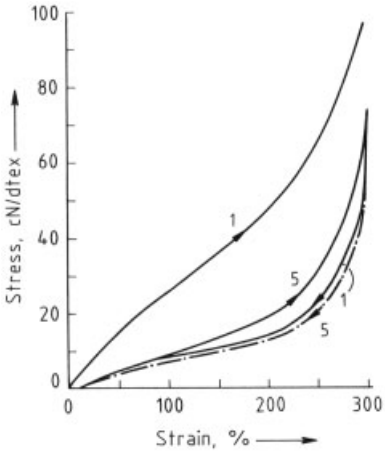

Figure 14. Stress – strain cycle of a 480-dtex spandex yarn (Dorlastan type V 400) ∗
∗ 1 is first curve; 5 is fifth curve; –▷– is elongation; –◁– is recovery.

Below the glass transition temperature of the soft segments, spandex loses its elasticity. In commercial products, this temperature of 210 – 250 K is far below normal environmental temperature range. Critically important for textile processors is the behavior at high temperature (160 – 170 °C, which is the temperature during textile treatment and shaping processes).

3.2. Chemical Properties

Compared to rubber yarns, spandex is substantially more stable to chemical influences during textile processing and subsequent use. Any possible damage manifests itself primarily in a loss of strength and an onset of yellowing. As expected, polyether spandex is more affected by oxidative processes and polyester spandex by hydrolysis. The stability of spandex subjected to various chemical treatments is given in Table 12.

Spandex-containing textiles can be cleaned chemically without problem. Sensitivity to photo oxidative damage is increased by unsaturated fatty acids and cosmetic oils [146]. Nitrogen oxides lead to yellowing and, hence, are detrimental to white textiles that contain spandex. Compared to rubber yarns, spandex has a greater affinity for textile dyes.

3.3. Raw Materials

Segmented polyurethane, the raw material for spandex fibers, consists of long-chain dihydroxy compounds (macrodiols), aromatic diisocyanate compounds, and short-chain diamino or dihydroxy compounds (chain extenders). Because of their chain mobility, the long-chain diols (soft segments) are responsible for the extensibility. The urethane and urea groups, formed by aromatic diisocyanates and short-chain diols (hard segments), function as physical cross-links and are the source of yarn strength.

Macrodiols. Macrodiols suitable for the manufacture of spandex have molar masses of 1 – 3 kg/mol and melting points < 50 °C. Only polyether and polyester diol are important. The most frequently used polyether diols is poly(tetramethylene ether glycol) [*25190-06-1*], polymerized from tetrahydrofuran via ring opening (→ Polyoxyalkylenes, → Polyoxymethylenes). Spandex fibers with polyether soft segments are highly resistant to hydrolysis and alkali, and have excellent low-temperature properties. They are, however, sensitive to light, atmospheric oxygen, and the chlorinated water of swimming pools and must be specially protected from them.

Among the polyester diols used, adipic acid polyesters have gained great importance, whereas polycaprolactone diols have not. For polyadipates, mixtures of $C_2 - C_6$ diols, e.g., 1,2-ethanediol – 2,4-butanediol and 1,6-hexanediol – 2,2-dimethyl-1,3-propanediol are used most frequently. Polyester spandex is more resistant to chlorinated water than polyether types and is insensitive to oxidative damage. Its sensitivity to hydrolysis can be reduced by the use of diols with a greater number of carbon atoms. Polyesters of mixed diols give spandex with a low tendency to crystallize during extension and with improved elastic properties.

Trade names. Polyethers: Terathane (Du Pont), Polymeg (Quaker Oats), Poly-THF (BASF), PTMG (Mitsubishi).

Polyesters: Niax (Union Carbide, Desmophen (Bayer).

Diisocyanates. Commercial spandex is almost exclusively produced from 4,4′-methy-

Table 12. Chemical resistance of spandex

Chemical treatment	Polyether spandex[a]	Polyester spandex[a]	Remarks
HCl 10 %, 20 °C, 24 h	low	high	
H_2SO_4 10 %, 20 °C, 24 h	low	high	
CH_3COOH 10 %, 20 °C, 24 h	none	none	
Cl_2-H_2O 20 mg/L, 20 °C, 24 h	high[b]	low	some yellowing
NaOH, followed by rinsing, suctioning, rinsing	none	none	
Cleaning in gasoline	low	low	
Xeno-test[c], 50 h	low	low	some yellowing
Xeno-test[c], 100 h	medium	low	little to moderate yellowing

[a] In the classification used, the following strength loss is to be expected: none, 0–5 %; low, 5–20 %; medium, 20–40 %; high, 40–90 %; total, 90–100 %.
[b] Can largely be prevented by use of ZnO as spinning additive [145].
[c] Light source: Hanau 450 (UV-emitting xenon gas lamp).

lenebis(phenyl isocyanate) [*101-68-8*] (MDI; → Isocyanates, Organic). Toluene diisocyanate [*26471-62-5*] (TDI), which was used in the early years of spandex production, is rarely employed today. Although aliphatic diisocyanates lead to polyurethanes with increased light durability, they are of little use because of increased difficulties in processing the spinning solutions. However, 1,6-diisocyanatohexane can be employed as a modifier [147].

Chain Extenders. Most spandex producers use short-chain aliphatic diamines, such as 1,2-diaminoethane or 1,2-diaminopropane, as chain extenders because they give fibers with the best thermal and hydrothermal properties. Hydrazine hydrate, carbonic acid, dicarboxylic acid dihydrazides, aminocarboxylic acid hydrazides, and semicarbazidocarboxylic acid hydrazides [148], [149] were used earlier as chain extenders, but are no longer of any significance. Occasionally, small amounts of codiamines, such as 1,3-diaminocyclohexane [150] or *N*-methylbis (3-aminopropyl)amine [151], are used in combination with ethylenediamine. This favorably affects certain properties, e.g., solubility and storage behavior. 1,4-Butanediol, frequently used in thermoplastic polyurethanes, results in spandex with unsatisfactory properties. 1,4-Bis(2-hydroxyethoxy)benzene, on the other hand, produces a melt-spun spandex with improved properties [152].

Solvents. Spandex fibers based on polyurethane–polyurea decompose during melting; therefore, they are spun from solution. Only dipolar aprotic solvents, usually dimethylacetamide and dimethylformamide, can be used.

Stabiliziers and Adjuvants. Because of the use of aromatic diisocyanates, light tends to discolor spandex [153]. Spandex in which polyethers are used as soft segment is, in addition, attacked photooxidatively. The following are used as stabilizers for these processes: phenolic antioxidants, sterically hindered amines, light-protective agents (e.g., benzotriazoles), and phosphites. They must meet special requirements such as low thermal volatility and good laundering resistance.

Spandex yellows under the influence of nitrogen oxides, which are contained primarily in combustion gases. Tertiary amine compounds have proved to be effective protectors against these; in addition, the patent literature describes a whole series of protective agents, e.g., carboxylic acid hydrazides [154], esters [155], [156], and phosphites [157], [158].

Polyether spandex, which is sensitive to chlorinated water, can be protected by tertiary amine compounds; polymers or polyurethanes with basic groups are used, such as poly(*N,N*-dialkylamino)ethyl acrylates and methacrylates [150], [159]. Protective action is also obtained by inclusion of zinc oxide during spinning [145].

The dyeability of mixed articles (fabrics containing a mixture of yarns of polyamide and spandex) with acid dyes can be improved by adding basic compounds to the spandex. Alternatively, basic substituents, e.g., *N*-alkyldialkanolamines or *N*-alkyldiaminoalkylamines, can be built into the polymer

molecule [160]. Previously, spandex was produced mainly as delustered yarns. Titanium dioxide was used as a delusterant, usually in concentrations of 2 – 5 %, to obtain as good a degree of whiteness as possible. Additional improvement in whiteness was obtained by adding nuancing dyes or pigments. However, nondelustered spandex types have gained greater prominence because they make possible more brilliant colors in dyed polyamide – spandex fabrics [161].

Freshly spun spandex yarn is slightly tacky. After being wound on the bobbin, cohesion between yarn layers develops with storage time. To counteract this and to improve the unwinding and processing behavior, separating and lubricating agents are used, e.g., polydimethylsiloxane and metal soaps [162].

3.4. Production of Polymer

The fiber raw material is produced by the two-stage prepolymer – chain extender process, frequently used in polyurethane chemistry.

Production of Isocyanate Prepolymer. In this process step, an α, ω-OH functional polymer (macrodiol) is reacted with an aromatic diisocyanate (called a polymer-analogous reaction) to yield an α, ω-NCO functional polymer (macrodiisocyanate). The reaction is carried out in the melt or in a highly polar solvent (e.g., dimethylacetamide) at 50 – 80 °C in a batchwise or a continuous process. The molecular mass of the prepolymer, which subsequently plays an important role in determining the elastic properties of the spandex, can be adjusted via the molecular mass of the macrodiol and the molar ratio of NCO : OH [163]. The prepolymers have number-average molar masses of 4 – 8 kg/mol, as well as NCO contents of 2 – 4 wt %, and are generally processed immediately without intermediate storage.

Chain Extension (Polymerization). In chain extension, the macrodiisocyanate is reacted with a diamine in a highly polar solvent (e.g., dimethyl acetamide) to give a high molecular mass polyurethane – polyurea. The most frequently used short-chain aliphatic diamines react rapidly. To obtain a good polymer structure, rapid and intense mixing is, therefore, required during chain extension. Reactivity can be reduced by formation of a carbamate from the diamine and carbon dioxide prior to chain extension [164].

Both batchwise and continuous processes are used during chain extension. Spandex solutions with solids contents of ca. 20 – 35 % and viscosities of ca. 50 – 300 Pa · s (25 °C) are obtained. Molecular mass is controlled during chain extension by using monoamines, e.g., diethylamine, as chain terminators. The polyurethane – ureas produced from diamines are linear polymers and, hence, soluble in the reaction media used.

With trifunctional starting components, covalently cross-linked polyurethanes are obtained, which are simultaneously chain extended and formed into a yarn [149] in the reaction-spinning process (see Section 3.5).

3.5. Spinning Processes

In contrast to elastic yarns based on rubber, spandex can be spun by using technologies generally known in the synthetic fiber industry. Because spandex based on polyurethane cannot be processed thermoplastically, solution-spinning processes are used to produce yarns.

Solution Spinning. Of the solution-spinning processes (dry, wet, and reaction spinning), dry spinning has become the most important because it lends itself more readily than the others to production of the fine spandex yarns that are increasingly in demand today.

Dry-Spinning Process (see Fig. 15). The high-viscosity elastomer solution is spun through multihole spinnerets (d) into heated spinning cells (e). During the process, the solvent evaporates, is absorbed by the entrained hot spin gases (air or inert gas), and is carried away. The solvent is separated by condensation and possibly scrubbing of the off-gas, distilled, and recycled. The individual filaments are cemented with the aid of a false-twisting device (j) to form a multifilament [165], [166]. Finish is applied to the yarn, which is wound up at a rate of 200 – 800 m/min. The drawoff speed is greater than the spinning speed (spin draw). The orientation thus produced affects yarn properties. To improve productivity, several multifilaments are spun from a single

spinning cell at the same time. Figure 16 shows a cross section of a dry-spun spandex yarn.

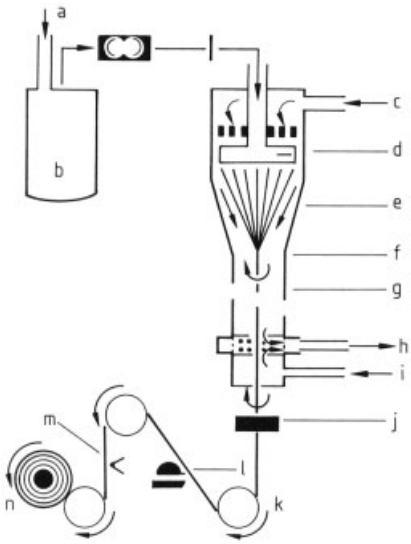

Figure 15. Dry-spinning process for spandex filament yarns [149]
a) Metering pump; b) Spinning solution; c) Hot spinning air; d) Spinneret; e) Heated spinning cell wall; f) Twisting point; g) Cell length, ca. 4 – 8 m; h) Spinning gas exhaust; i) Fresh gas; j) False-twisting device; k) Godet roll; l) Finish; m) Traverse; n) Windup device

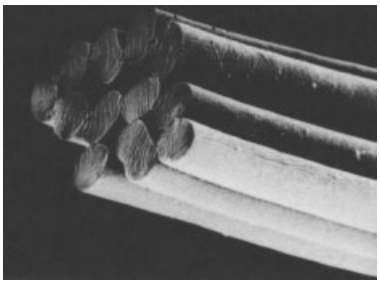

Figure 16. Dorlastan fiber cross sections, 160 dtex (12 individual capillaries) ∗
∗ ca. 50 µm.

Wet-Spinning Process (see Fig. 17). In the wet-spinning process, the elastomer solution is spun into aqueous precipitation baths (d). In the process, the water-insoluble polyurethane coagulates, and the spinning solvent is carried off by the water. As in dry spinning, the individual filaments are cemented to a multifilament. After the filament has passed through wash baths and drying cabinets, finish is applied and the filament is wound. Windup speeds are significantly lower than in dry spinning, but equipment and energy costs are lower.

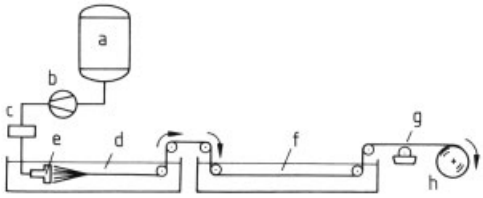

Figure 17. Schematic of a spandex wet- or reaction-spinning machine [126]
a) Spinning solution or prepolymer; b) Spinning pump; c) Filter; d) Precipitating or reaction spin bath; e) Spinneret; f) Wash or subsequent reaction bath; g) Finish; h) Windup

Reaction-Spinning Process. In this process, chain extension of the isocyanate prepolymer by the amine and formation of the yarn occur simultaneously in the precipitation bath [167]. During spinning of the prepolymer into the reaction spin bath, a solid polymer skin forms immediately on the yarn surface and provides the stability necessary to draw off the yarn. Complete solidification is achieved in a second step. With this process, producing chemically cross-linked spandex by using trifunctional starting materials is also possible.

Melt-Spinning Process. Only elastic yarns that are exclusively based on pure polyurethane can be prepared by the high-productivity melt-spinning process. These materials do not have the properties of commercial spandex based on polyurea – polyurethane and are, therefore, of little technical significance

3.6. Uses

Production of Elastic Yarns and Fabrics. Spandex fibers are processed and used exclusively in combination with hard fibers to impart high elasticity to textiles. In most cases, the spandex is used in bare form (i.e., unwrapped). Because of its rubberlike hand and poor frictional properties, spandex intended for certain end uses, such as woven elastic clothing fabrics [168], support stockings, and socks, is wrapped or corespun with hard fibers. Of these elastic combination yarns [168], wrapped yarns are the most important. Wrapping can be carried out with one or two textile yarns. Corespun

yarns are produced on special secondary spinning machines on which staple fibers are twisted around the stretched spandex yarn. Corespun yarns are distinguished by a remarkably soft hand. To improve the processibility of the combination yarns, they can be changed by a combined stretch – steam process so that they (temporarily) lose their elastic extensibility. After being processed into fabric, the good elastic properties are restored by dyeing at the boil in the relaxed state.

The most important types of processing used to produce articles containing spandex are listed in Table 13. As can be seen, both knit (processing on circular or warp knitting machines) and woven fabrics can be produced. Articles with the highest extensibility can be prepared by the circular knitting process [169]. The elastic yarn is processed directly from the bobbin, the yarn being taken off over the end of the spool.

To prepare warp knitwear, the elastic yarns must first be assembled [170]. To do so, up to 1500 yarns are wound parallel, side by side, under constant extension onto a warp beam. Here, processing is exclusively in the bare form. By varying the fineness, the hard fiber yarn, and the pattern notation technique, the properties and appearance of the articles can be influenced in many ways with this process and can be adapted to both functional requirements and the demands of fashion [171], [172].

Processing of elastic yarns in woven fabrics has increased substantially. End uses are suit-coat and pants fabrics for articles with increased wear comfort.

The major end uses for coarser spandex yarns and elastic wrapping yarns are tapes for various textile and engineering applications. In Western Europe, 45 % of spandex production is used in girdle and underwear manufacture, 20 % each in swim wear and hosiery, and ca. 15 % in other areas.

Textile Properties of Elastic Fabrics. The most important property of elastic fabrics is the combination of extensibility and elastic recovery. Spandex can elongate until the hard fiber construction obstructs further elongation. Extensibilities vary from ca. 25 % (woven fabrics) to > 200 % (knitwear). The elastic behavior of a warp-knitted fabric is shown in Figure 18.

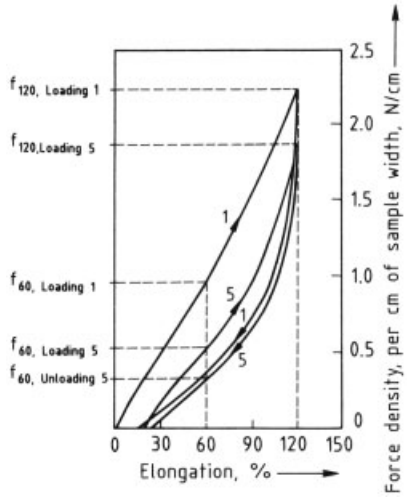

Figure 18. Curves of cyclic elongation of a tricot fabric (longitudinal direction) containing 18.5 % 45-dtex Dorlastan, unit mass 220 g/m^2, longitudinal extensibility 210 % ∗ [126]
∗ 1 is first curve; 5 is fifth curve; –▷– is elongation; –◁– is recovery.

By using spandex, the textile industry can fabricate elastic textiles in great variety, having given shaping and support forces, and also meeting all requirements with respect to end-use behavior, comfort, and fashion.

4. Polyolefin Fibers

Polyolefin fibers, yarns, and monofilaments are usually produced by melt spinning. Film yarns and split yarns, obtained from films by uniaxial stretching, are also produced in large quantities. Polypropylene (PP) is the predominant raw material for fibers and yarns. High-modulus yarns of ultrahigh molecular mass high-density polyethylene (HDPE), commercialized in the mid 1980s, are an exception.

Both PP and HDPE are used to make film yarns and monofilaments; however, HDPE is used to a much lesser extent. Areas of application for PP fibers include carpet yarns, home furnishing fabrics, industrial fabrics, and nonwovens. Polyolefin tapes and monofilaments are used in bags, tarpaulin, industrial fabrics, carpet backing, twine, rope, and bristles.

The production of textile products from polyolefins in 1985 was estimated to be ca. 1.7×10^6 t, about 91 % (1.55×10^6 t) of which was PP and

Table 13. End uses of spandex continuous-filament yarns

End use	Manufacturing technology	Amount of spandex, wt %	Yarn type	Fineness, density, dtex
Lingerie and lingerie bands	circular knit	5 – 12	bare, wrapped, corespun	22 – 400
	warp knit		bare, wrapped	22 – 80
Girdles	warp knit	10 – 40	bare, wrapped	22 – 800
Swim wear	warp knit	10 – 32	bare,	22 – 80
	circular knit	6 – 12	bare, wrapped	44 – 160
Hosiery bands, sock borders	circular knit	10 – 20	bare, wrapped	135 – 400
Support stockings	circular knit	10 – 60	bare, wrapped, corespun	22 – 480
Tapes	tape weaving	10 – 50	bare, wrapped	320 – 5000
Outerwear	weaving	2 – 10	rubber threads, wrapped yarns	44 – 160

about 9 % (150 000 t) HDPE. About 30 % of all polypropylene produced is processed into tapes, fibers, and monofilaments. In 1985, this amounted to ca. 630 000 t in Western Europe, 47 % of which went into tapes, split fibers, and monofilaments; 32 %, into staple fibers; 13 % into filament yarns, and 8 % into spunbonds. The growth of polypropylene yarns, fibers, and nonwovens should continue to be above the growth of polypropylene consumption in general as well as that of all other synthetic fibers.

4.1. Polypropylene Fiber

Polypropylene fiber was developed in Italy by Montecatini during the second half of the 1950s [173]. Several other companies initially played an active role in its development, including Hercules, Reeves Brothers, Vectra, and Phillips in the United States; ICI in Europe; and Toray, Mitsubishi Rayon, and Toyobo in Japan. Some of them, like ICI, disappeared from the scene after the first few years; others improved their line of products and became extremely important in the market. There are many producers; in Western Europe they increased in number during the late 1970s, when the Montecatini patents expired.

4.1.1. Chemical and Physical Properties

Chemical Structure. Polypropylene is a hydrocarbon made up of $-CH_2CH(CH_3)-$ units. These propylene units form a stereoregular helix, with the methyl groups at an angular distance of 120° (isotactic polypropylene). A small percentage (ca. 2.5 – 5 %) is sterically random (atactic polypropylene).

Because its structure contains no polar groups, PP is resistant to most chemical agents, in particular to acids and alkalies; insensitive to water; and not dyeable with common types of dyes. Moreover, it has a low resistance to thermooxidative and photooxidative degradation. The problem of nondyeability has been solved by mass dyeing techniques (Section 4.1.2.5) and the problem of degradation by the use of suitable additives (Section 4.1.2.4). Because of the absence of polar groups, only van der Waals forces exist between macromolecules; these forces are much weaker than the hydrogen bonds found, for example, in polyamide. Because mechanical characteristics depend on these forces and on molecular entanglement, very high molecular masses (M_r from ca. 170 000 to 300 000, compared to ca. 20 000 for polyamide) must be used to obtain satisfactory strength (fiber tenacity).

Rheological Aspects [174]. Polypropylene fiber is produced by melt spinning, i.e., by extruding the molten polymer through suitable spinnerets. Therefore, the rheological behavior of the polymer melt must be considered in selecting processing conditions or even in choosing a certain polymer for specific spinning conditions. The first aspect to be considered is that isotactic PP in the molten state is highly non-Newtonian and exhibits a strong pseudoplastic character; the apparent viscosity of the melt is not constant with varying shear stress but it decreases when shear stress increases. The deviation from Newtonian behavior is markedly dependent on the relative molecular mass (M_r) and polydispersity (molecular mass distribution M_rD); for polymers of comparable polydispersity, the melt

viscosity increases with increasing viscosity–average molecular mass. An extremely useful, practical indication of M_r is the melt flow index (MFI), which is a measure of flow under standard conditions (ASTM D 1238: grams extruded in 10 min through a standard orifice under a pressure of 2.16 kg/cm^2 at 230 °C); MFI is inversely proportional to melt viscosity. The MFI is measured at a shear rate close to those commonly met in usual spinning processes.

The initial MFI of PP polymers used in fiber production ranges from 2 to 35, depending on the process involved and the type of fiber produced. In addition to pseudoplasticity, PP melt exhibits appreciable signs of viscoelasticity. In a viscoelastic material the deformations undergone by the melt have not only a viscous but also an elastic component, which can therefore be recovered when the applied stress ceases, i.e., when the extruded polymer emerges from the spinneret capillary. The residual elastic energy, whose magnitude depends on the residence time in the capillary, can cause, among other phenomena, enlargement of the extrudate to a diameter larger than that of the capillary (die swell). The polymer swelling increases as the shear stress increases and as the ratio of length to capillary radius decreases.

Die swell also depends on the molecular mass distribution (M_rD): at comparable melt viscosities, die swell is larger for a polymer with broad M_rD than for a polymer with a narrow M_rD. Swelling increases as temperature decreases. All these rheological characteristics are taken into account when spinneret hole geometry, polymer type (MFI, M_rD), and extrusion conditions are being selected.

Structural Characteristics [174, pp. 327–337], [175]. The stereoregular structure of isotactic PP macromolecules makes an ordered three-dimensional configuration (i.e., a crystalline structure) possible when the molten polymer is cooled to the solid state. Under normal conditions, isotactic PP crystallizes in the monoclinic system; if cooled rapidly to room temperature, it exhibits a paracrystalline form, called smectic. In practice, PP fiber is a partially crystalline material. The crystalline structure of the fiber depends on the conditions of extrusion and quenching, the take-up speed in the spinning step, and the draw ratio and annealing conditions in the finishing step. Normally, the crystallinity of the finished fiber does not exceed 70 %; values of 40–50 % are usual. Many physical properties are influenced by crystallinity. Mechanical properties, however, are affected less by the percentage of monoclinic material than by the molecular orientation imparted during drawing; orientation is evaluated by measurement of birefringence.

Mechanical Properties. Depending on spinning conditions (extrusion temperature, quenching, spinning speed, and rate of deformaton in the molten state), the mechanical properties of as-spun PP fibers can vary widely, as shown schematically in Figure 19. The mechanical properties of drawn fibers are influenced substantially by the applied draw ratio and, consequently, also vary widely (Fig. 20).

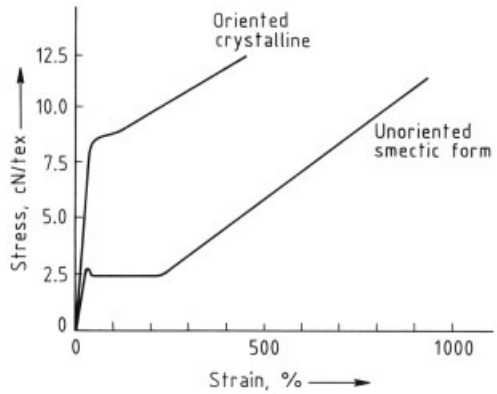

Figure 19. Stress–strain curves of as-spun polypropylene fibers

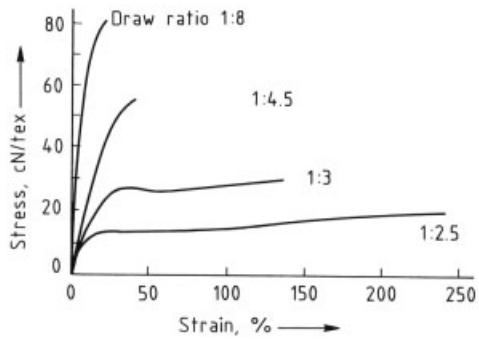

Figure 20. Stress–strain curves of drawn polypropylene fibers

The relationship between tenacity and draw ratio depends on the polymer used and, in particular, its polydispersity. As polydispersity decreases, the tenacity obtained at a given draw ratio increases, but the drawability of the spun fiber decreases at the same time. Therefore, polymers with narrow M_rD are not suitable for production of high-tenacity fibers; they do, however, have advantage in terms of spinning continuity at high spinning speeds.

4.1.2. Production

4.1.2.1. Raw Materials

Polypropylene fibers are obtained from PP homopolymer. They may also be based on PP and a propylene – ethylene random copolymer or a blend of polypropylene and polyethylene homopolymers. In the latter case, the definition "polyolefin fiber" or "olefin fiber" is more correct (see Section 4.2). However, polyolefin is also frequently used for PP homopolymer fiber.

Polypropylene is used in the production of side-by-side bicomponent polypropylene – polyethylene fibers. The Customs Cooperation Council of the European Economic Community (EEC) adopted the following definition of textile polypropylene:

Fibers composed of aliphatic saturated hydrocarbon linear macromolecules having in the chain at least 85 % by weight of units with one carbon atom in two carrying a methyl side chain in an isotactic position, and without further substitution.

Availability. The availability of raw materials is not a problem, although the supply of propylene may be influenced by the demand for ethylene because ethylene and propylene are coproducts in the cracking of virgin naphtha. Other sources, of lesser importance at the moment, are liquefied refinery gases containing propylene and propane gases.

4.1.2.2. Staple Fiber Production Process

Polypropylene can be produced by two different processes: (1) *conventional spinning* is based on high spinning speeds and spinnerets with a rather limited number of holes; (2) *short spinning* is based on low spinning speeds and spinnerets with a large number of holes. In the first process, because of the high spinning speeds, postspinning operations are performed in a separate step; in the second process, because of the low spinning speeds, spinning and postspinning are carried out continuously. Postspinning operations are basically the same for both processes.

Conventional Spinning Process. The polymer, to which stabilizers and pigments have been added, is introduced into the extruder (a) (Fig. 21). The molten material is fed to the spinnerets (e) by metering pumps (c), which control the throughput of each spinneret. The extruder is heated electrically, while the melt distribution manifold is heated by means of a diathermic fluid (e.g., Dowtherm and Diphil). Filaments (f) emerging from the spinneret are quenched with cold air (air speed 1 – 2 m/s) and, when solidified, collected in cans (i). An antistatic and lubricating agent is usually applied to the filaments before they come in contact with any solid surface (g).

The filament fineness depends on the throughput of the metering pump, the number of holes in the spinneret, and the take-up speed, according to the following relationship:

$$d = \frac{1000}{3.6}\,\frac{Q}{n \cdot v}$$

where d is the fineness, decitex; Q is the pump throughput, kilograms per hour; n is the number of holes; and v is the take-up speed, meters per second. The spinning speed of PP fiber is usually 600 – 1200 m/min, and the temperature of the molten polymer in the spinneret is 220 – 300 °C, depending on molecular mass of the polymer, spinning speed, and desired characteristics of the fiber being produced. Spinnerets used for staple production are 150 – 250 mm in diameter and have between 200 and 2000 holes, depending on the fineness of the fiber. In some cases, rectangular spinnerets are used to achieve higher productivity. The number of holes is limited to avoid having the filaments stick together (filament marriage). Polymers with a fairly low M_r (MFI 5 – 20) are used commonly in the production of standard fibers; higher M_r (MFI ca. 2) polymers are used only for high-tenacity fibers. To obtain satisfactory spinning continuity, the

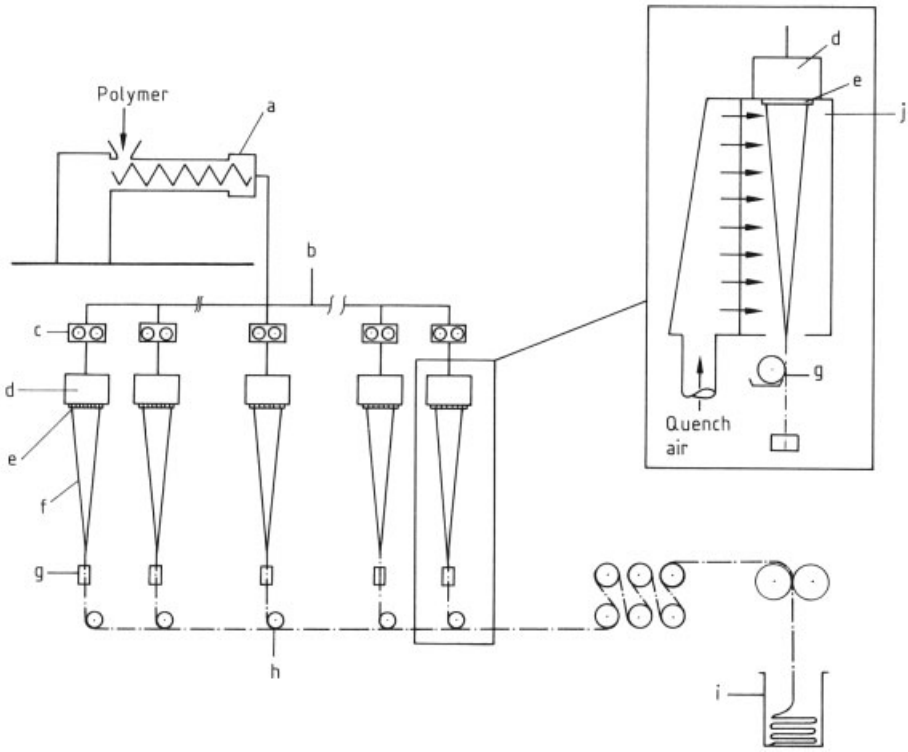

Figure 21. Conventional spinning plant
a) Extruder; b) Melt distribution; c) Metering pump; d) Spinning pack; e) Spinneret; f) Filament; g) Oiling device; h) Take-up; i) Can coiler; j) Quench box

polymer must be filtered before reaching the spinneret; both sand filters and screen filters are used.

Postspinning Operations. Drawing, crimping, heat setting, cutting, and baling are the steps that follow spinning in the production process.

Several tows coming from the spinning section are processed together to reach a total fineness of 200 – 300 ktex (0.2 – 0.3 kg/m). They are stretched between two sets of rolls; the draw ratio depends on the ratio of the roll speeds of these two sets. The draw ratio for standard fibers is between 1 : 3 and 1 : 5. To obtain uniformly drawn fibers, the tow is heated in a steam chamber located between the two sets of rolls. Alternatively, heated rolls can be used in the first set. If high-tenacity fibers are desired, the drawing operation is sometimes divided into two steps by means of a third set of rolls. Crimping is intended to impart bulk and cohesion to the fiber for subsequent conversion to final products. The fiber is compressed in the crimper (stuffer box) and subsequently dried and heat-set in a hot-air oven (100 – 130 °C) to stabilize crimp and reduce thermal shrinkage after drawing. A finish is usually applied to the fiber during the postspinning operations. Staple cutting and baling operations are the final stages of production. The speed of these operations is limited by the crimping speed and is generally 150 – 250 m/min.

Short Spinning Process. The rheological behavior of PP is such that, when spinning speeds are fairly low ($\leq$ 120 m/min), and the stream of quenching air blows through the filaments close to the spinneret, the filaments do not stick together even with close hole spacing (ca. 1 mm). Then, the speed of consecutive operations can be matched, and a continuous,

uninterrupted, one-step process can be achieved. A consequence of the high hole density and the reduction in height of the quenching section is that plants can be made more compact for a lower cost. For this reason and reduced manpower requirements, the short spinning process has become very popular.

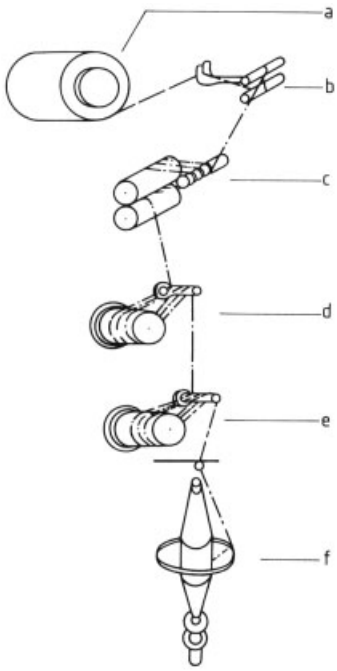

Figure 22. Draw-twisting machine
a) Spun yarn; b) Pre-tension guide; c) Pre-tension rolls and guides; d) Feed roll; e) Draw roll; f) Twisting take-up

4.1.2.3. Continuous Filament Yarn Production Process

In principle, spinning continuous filament yarn is similar to spinning staple fiber; the main difference is that in the former, each yarn is collected on an individual bobbin. Drawing of the yarn (Fig. 22) is usually performed by two pairs of rolls (f), which are generally heated electrically. For some applications, continuous filament is used as it comes from the drawing machine (flat yarn), but in most cases, filament bulk must be increased by means of a texturing system. A mechanical texturing system is preferred for the very fine yarns used in the textile industry. In this case, the yarn is highly twisted, heat-set, and detwisted in a continuous false-twisting process to give each individual filament the shape of a spring. A subsequent heat-setting operation can be introduced in the detwisting zone to reduce yarn elasticity. For coarser continuous filament yarns (carpet yarns), systems based on fluid bulking jets are used. Consistent with the continuing trend toward reduced production costs, machines have been developed in which the different operations are performed continuously in a single step.

4.1.2.4. Stabilization

Polypropylene, although relatively stable to heat and light in the absence of oxygen, has poor resistance to thermooxidative and photooxidative degradation [175, pp. 38 – 93]. Thermooxidative degradation occurs during conversion of the polymer to a fiber. Photooxidative degradation occurs when the end product, which contains PP fiber, is exposed to direct or diffuse solar radiation in the presence of oxygen. As a hydrocarbon, PP is transparent to UV light but it is sensitive to it because of the presence of traces of oxygen-containing groups, such as hydroperoxides and ketones, which absorb at 310 – 330 nm. Depending on the temperature during exposure, thermooxidative degradation and photooxidative degradation often occur simultaneously. Both processes generate polymer radicals, which catalyze further molecular cleavage.

Therefore, suitable antioxidants and light stabilizers are added to protect the polymer melt at high temperature and to impart sufficient aging resistance to the finished products. (→ Antioxidants, Chap. 5.4.). These stabilizers can be classified according to the mechanism by which they protect the polymer: (1) chain terminators or free-radical scavengers, which terminate the chain reaction by giving rise to nonreactive radicals; (2) hydroperoxide deactivators, which decompose hydroperoxides without formation of radicals; (3) UV absorbers, which screen UV radiation; and (4) quenchers, which quench the excited states of ketone groups; and (5) free-radical scavengers that trap the radicals and regenerate themselves (HALS: hindered amine light stabilizers). The following are some of the chemicals used at present:

Chain terminators	phenolic antioxidants
Peroxide decomposers	disulfides, trialkyl phosphites
UV absorbers	derivatives of hydroxybenzophenone and benzotriazole
Energy quenchers	nickel chelate stabilizers
Free-radical scavengers	hindered amines [176]

Some producers of stabilizers are:

Ciba-Geigy	B. F. Goodrich Co.
Borg Warner Chemicals	Enichem
Uniroyal	Ethyl Co.

The effectiveness of the stabilizer formulation depends on the choice of individual components. This choice must also take into account the pigments used because they can affect heat and light stability significantly and can even interfere with stabilizers. In addition, attention must be paid to possible concomitant phenomena, such as gas fading, which is the color change caused by nitrogen oxides present in air. These oxides can react with some components of the stabilizing system, resulting in nitration of aromatic rings or addition of nitrogen oxides to double bonds. Nitrogen oxides can also react with certain finishes and cause color changes via some of the reactions mentioned previously.

The stabilizer formulation also includes antacids, typically calcium stearate, which react with the HCl formed by reaction of catalyst residue with substances containing active hydrogen. Stabilizer systems that comply with international regulations for chemicals in food contact applications are also available.

Stabilizer Evaluation. The following test methods have been developed to select the most suitable stabilizers and predict actual fiber behavior.

Oxidation Resistance. Fiber bundles, hung on a rotating frame under a small weight, are exposed in a forced-draft oven at 110 °C. Resistance to oven aging is expressed as days to failure. For operating conditions, see ASTM D 3045-74 (reapproved 1979).

Light Resistance. Accelerated tests with instruments based on a xenon lamp are preferred to those using a carbon-arc apparatus because the xenon lamp gives a spectrum similar to that of summer sunlight, whereas the carbon-arc apparatus is too high in UV radiation. Exposure conditions are selected so that the temperature is comparable to that experienced in use because the rate of photooxidation is a function of temperature. For indoor uses, such as carpets and wall coverings, fiber bundles are exposed in a Xenotest or Weatherometer according to ASTM G 26 – 84, DIN 54 004, ISO B 02. The resistance to artificial light is expressed as residual tenacity after exposure. Fiber exposed to combined heat and light, e.g., on rear window shelves in cars, is tested in the Xenotest at high temperature.

Gas-Fading Resistance. The fiber specimen is exposed to nitrogen oxide containing gases derived from the combustion of butane. Color variation is assessed by comparing the exposed and original samples after four exposure cycles of controlled duration.

4.1.2.5. Pigmentation

Polypropylene cannot be dyed by conventional methods because it lacks polar groups that can provide suitable sites for dyes. The techniques proposed thus far for making PP dyeable by traditional acid or disperse dye systems do not have near- or medium-term commercial potential, either because of the relatively high cost of additives involved or because of the as yet insufficient colorfastness. Only nickel-modified fibers are used commercially for space dyeing or printing.

For these reasons, PP fiber is colored by blending the polymer with suitable pigments prior to extrusion; this system is known as "melt dyeing" or "mass dyeing." Because of various requirements, such as thermal stability at high temperature, the range of pigments that can be used is limited; however, mass-dyed PP fibers are characterized by high colorfastness to light, heat, rubbing, and washing. Colorfastness may, of course, be affected by stabilizer – pigment interaction or by additives that facilitate extraction of the pigments. The main classes of pigments used are the following:

Inorganic:	titanium dioxide (rutile or anatase)
	iron oxides
Organic:	insoluble azo pigments
	copper phthalocyanines
	vat pigments
	carbazole dioxyazine
	carbon black

Some pigment producers are:

BASF	Bayer
Sandoz	Farbwerke Hoechst
Ciba-Geigy	Cabot

4.1.3. Uses

Types of Fibers Produced. Polypropylene fiber is produced in a wide range of finenesses, both as staple and continuous filament. Staple fibers range from ca. 1 to 250 dtex per filament. The fiber cross section is primarily circular; the fineness range of continuous-filament yarn is 70 – 12 000 dtex. The filament cross section is predominantly circular in the low to medium fineness for apparel, upholstery, and industrial uses, and triangular or trilobal for the fibers used in floor covering. For the latter application, the basic fineness range is 1300 – 4000 dtex; higher fineness up to 12 000 dtex is obtained by comingling.

Commercial Mass. The International Bureau for the Standardization of Man-Made Fibers (BISFA) has agreed on the following definition of commercial mass: "the commercial mass shall be obtained by adding to the oven-dry mass, i.e., after removal of finish, etc., the mass corresponding to the conventional allowance" where "the conventional allowance for polypropylene staple fiber, sliver or top and tow is 2.0 %."

Properties and Applications. Polypropylene fiber has physical, chemical, and mechanical properties that make it suitable for a wide range of applications. These properties can be divided into the following groups:

Good
Relative density
Resistance to acid and alkali
Abrasion resistance
Moisture absorption
Mechanical properties under dry and wet conditons
Dimensional stability
Stain and soil removal
Easy washing
Quick drying
Static buildup in carpets
Resistance to mildew, microorganisms, and insects
Thermal bondability
Colorfastness of suitably pigmented fibers
Light resistance of suitably stabilized fibers

Fair
Flammability
Frictional characteristics
Oil-based soil removal

Poor
Dyeability with common dyeing methods (need for pigmentation)
Heat and light resistance (need for stabilization)
Resistance to dry-cleaning solvents
Behavior during ironing

Some properties are inherent in the component material; others are imparted or enhanced by the production process or by suitable additives. A brief survey of the major properties follows, with an indication of their effect on fiber application.

Relative density (0.91) is the lowest of all commercial textile fibers; this is an advantage in all applications because it means lightness and high cover (especially in nonwoven construction).

Chemical resistance, especially to acid and alkali, is outstanding. Polypropylene fiber was treated at 20 °C with the following chemicals; the residual tenacity after 96 h is given (in percent):

Nitric acid (66 %)	80 – 87
Sulfuric acid (95 %)	95 – 100
Formic acid (75 %)	100
Sodium hydroxide (40 %)	100
Concentrated aqueous ammonia	95 – 100
Trichloroethylene	80
Perchloroethylene	80
Xylene	80
Toluene	90
Benzene	90
Acetone	100
Sodium hypochlorite (5 % active chlorine)	85
Hydrogen peroxide (12 parts)	90

Hot halogenated solvents swell PP fiber, with subsequent shrinkage on drying (trichloroethylene), or even dissolve it (boiling perchloroethylene). Moreover, some boiling aromatic solvents dissolve the fiber (e.g., xylene, decahydronaphthalene), and strong oxidizing agents attack it. In the last case, however, the fiber can be improved

considerably by suitable stabilizer formulations. The high chemical resistance is exploited in industrial filtration.

Abrasion resistance is comparable to polyamide. It increases with increasing crystallinity and molecular mass and is influenced by frictional properties. The high abrasion resistance is exploited in floor covering, upholstery, hosiery, and industrial applications.

Moisture regain is virtually nil (0.05 % at 65 % R.H. and 21 °C). The insensitivity of PP fiber to water brings about the following advantages: dimensional stability, fabric dryness in contact with the skin, quick drying, and no variation in mechanical properties under wet conditions.

Temperature Behavior. Polypropylene fiber does not distort when exposed to temperatures of 120 – 130 °C during conversion processes; however, it is sensitive to pressure at elevated temperature, so ironing can cause the fabric to stiffen. The softening point is 150 °C, and melting occurs at 168 °C. The relatively low melting point is exploited in thermal bonding, which has become important in hygienic and medical applications (calender bonding is generally used, but flow-through bonding with hot air can also be applied). High-temperature tentering treatment is used to produce dense nonwovens suitable as coating substrates.

Dyeability. Because PP fiber is not dyeable by common dyeing methods, pigmentation is used. Although a nickel-modified fiber is available for printing or space dyeing with chelatable dyes, its use is limited. The absence of polar groups in the PP molecule makes for easy stain removal; on the other hand, PP is more easily soiled by oil-based stains than other synthetic fibers.

Resistance to Mildew, Microorganisms, and Insects. Polypropylene fiber is neither attacked by mildew and molds nor damaged by insects.

Static Buildup. Limited charges are separated when a person walks on a PP carpet and, consequently, a low body voltage is generated, lower than that at which uncomfortable shocks can be felt.

Flammability. Application of a flame causes PP to shrink, melt, and burn if contact with the flame is sufficiently prolonged. Polypropylene burns slowly and drips like a candle. Burning can continue even if the flame is withdrawn because the limiting oxygen index of PP is 18.5. The self-ignition temperature is 570 °C.

Frictional Characteristics. The coefficient of friction of PP fiber, both fiber-to-fiber and fiber-to-metal, is fairly high. Suitable lubricants are applied to keep it low so as to allow conversion by textile processing. Cotton spinning is one of the most critical processes.

Electrical Properties. Polypropylene is a good insulator and exhibits an extremely low power loss, even at high frequencies. The dielectric constant is 2.1 at 60 Hz and 2.2 at 1 MHz; the power loss factor is 0.0002 at 60 Hz and 0.0003 at 1 MHz; volume resistivity is 4.9×10^{14} Ω cm.

Thermal Properties. The thermal conductivity of PP fiber is the lowest among commercial fibers (0.138 $W\,m^{-1}\,K^{-1}$). This contributes to high thermal insulation in textiles, although the major contribution comes from air trapped in the structure.

Allergic Phenomena and Nontoxicity. Polypropylene itself does not cause skin irritation or sensitization and is not toxic; surface finishes must be selected correctly to retain these advantages.

Applications. The three major fields of application are in industrial, home textile, and apparel end uses.

Industrial applications include rope, twine, conveyor belts, industrial sewing thread, filter cloth, paper reinforcement, geotextiles, coating substrates, and carpet backing. Diaper facing is an example of a hygienic – medical product in which PP has made rapid inroads.

Home textile applications include floor covering (tufted, woven, and needle punched), upholstery fabric, wall covering, fiberfill, blankets, warp yarns in woven carpets, and yarns for secondary backing for tufted carpets. Floor coverings constitute the most important outlet for PP fiber, in both bulked continuous filament (BCF) and staple.

The *apparel uses* of PP are limited by lack of dyeability, difficulty in ironing, and inability to be dry-cleaned. However, the fiber is established in knitwear, especially sportswear. Staple and textured continuous filament yarns are used. Other applications include fleece fabrics, hosiery, and hand knitting yarns.

4.1.4. Economic Aspects

Cost. Although the price of PP in fiber production may be influenced by the demand for polymer in other areas of application, the price is generally lower than that of other fiber-forming polymers because of the shorter "chemical path" from virgin naphtha from which they are all derived. Another positive aspect is the improvement of bulk polymerization processes based on high-yield catalysts, which reduce production costs considerably.

Production. In 1985, world production of PP fiber was about 650 000 t [177], 250 000 t of which were produced in Western Europe (180 000 t of staple fiber and 70 000 t of continuous-filament yarn). Some producers and *trade names* are listed in Table 14 [178], [179].

Table 14. Trade names of polypropylene fibers

Country	Producer	Trade name *
Austria	Chemie Linz	Asota (ST, Y)
former Czechoslovakia	Chemosvit	Prolen (Y)
Denmark	Danaklon	Danaklon (ST)
		Danaklon ES (ST)
Federal Republic of Germany	Faserwerke Bottrop	Vegon (ST)
	Steen	Polysteen (ST)
Italy	Filatura di Delebio	Delebion (Y)
	Moplefan	Meraklon (ST)
		Meraklon CF (Y)
		Meraklon BCF (Y)
Japan	Chisso	Chisso Polypro
		Chisso ES (ST, Y)
	Daiwabo	Daiwabo Polypro (ST)
	Mitsubishi Rayon	Pylen (Y)
United Kingdom	F. Drake	Gymlene (ST)
	Plasticizers	Charisma, Duron (ST, Y)
United States	Amoco	Marquesa Lana, Propex III (ST, Y)
	Hercules	Herculon
		Herculon Nouvelle (ST, Y)
	Phillips	Marvess (ST, Y)

* ST = staple fiber; Y = continuous filament yarn.

The amount of fiber produced is not commensurate with other synthetic fibers, such as polyester, polyamide, and acrylic fibers, but is, nevertheless, significant and continues to increase steadily. The reason for this steady, though not explosive, increase is not only the interesting performance of the product but also the favorable cost of the raw material and the development of production systems with low investment costs and reduced labor requirements. Another significant factor is the solution of problems regarding stabilization as well as the availability of economical and effective pigmentation methods. Pigmentation has eliminated the basic drawback of the fiber, i.e., its nondyeability.

4.2. Polyolefin Film Yarns and Split Yarns

The production of polyolefin film yarns and split yarns is a relatively recent route to the manufacture of textile products, compared to customary spinning processes that have been in use for a long time. Basically, the process involves production of a primary film, which is subsequently cut, stretched, and possibly fibrillated mechanically. This method had been developed and patented by the end of the 1930s [180] but was not exploited industrially on a broader basis until the 1960s.

In principle, other thermoplastics can also be produced by this means. However, it is important only for polypropylene and for high-density polyethylene (HDPE) (→ Polyolefins).

The usual size of individual PP film ribbons is between 300 and 2500 dtex; split yarns often have sizes up to 10 000 dtex. The smallest size attainable by special processes is ca. 10 dtex.

The literature on PP film yarns and split yarns has been reviewed [181–185].

4.2.1. Raw Materials

The choice of raw materials depends not only on what is required of the finished product but also on the processing properties. Basically, raw materials with higher molecular masses (lower melt flow indices) are used to produce film yarns and monofilaments than to produce melt-spun multifilaments. The choice between PP and HDPE is governed by the following properties:

1) Polypropylene has higher strength, i.e., greater tenacity, than HDPE (see Figure 23) [186], [187]. In addition, PP exhibits less "cold flow"; i.e., products of PP have lower elongation under continuous mechanical loading than those of HDPE.

2) Tapes and monofilaments of HDPE are more flexible than those of PP and give finished products with a softer hand.
3) Polypropylene can withstand a higher thermal load because of its higher melting temperature (≅160 °C vs. ≅130 °C for HDPE).
4) On the other hand, HDPE has better resistance to cold. Although stretching substantially improves the resistance of PP to cold, HDPE is at present preferred for applications below freezing temperature because it offers a greater margin of safety.
5) Tapes and monofilaments of HDPE have lower coefficients of sliding friction than those of PP. Consequently, they are better suited for processes in which good sliding ability is needed (for example, processing on knitting machines).
6) The tendency of HDPE tapes to split is much lower than that of PP. Depending on end use, this can be an advantage or a disadvantage.
7) The UV stability of HDPE is inherently better than that of PP. Nevertheless, very good weathering resistance has now been achieved with PP tapes and monofilaments by using high-efficiency UV stabilizers of the hindered amine light stabilizer (HALS) type.

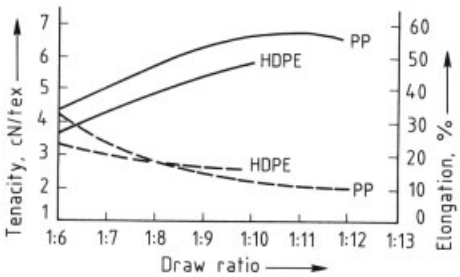

Figure 23. Tenacity (——) and break elongation (– – – –) of film yarn of polypropylene (PP) and high-density polyethylene (HDPE) as a function of draw ratio [186]

The HDPE types used have densities between 0.945 and 0.960 g/cm^3 and melt flow index values (MFI 190/2.16, i.e., MFI measured at 190 °C under a load of 2.16 kg/cm^2) between 0.3 and 0.7 g/10 min; a broad molecular mass distribution is preferable.

Polypropylene tapes are produced almost exclusively from homopolymers. Depending on the process employed and end uses envisioned, melt flow indices (MFI 230/2.16) between 1.5 and 6 g/10 min, and preferably between 2.5 and 4 g/10 min, are common. If the process involves cooling via a water bath, PP types with low water pickup (carry over) are required.

Coloration is achieved by pigment compounds; to minimize splitting and assist in weaving PP ribbons, inorganic fillers (e.g., chalk) are used.

In special cases, copolymers (i.e., blends with other polymers) or coextrudates can be used to achieve certain properties. Coextruded films permit the production of crimped tapes and yarns [188][189][190].

4.2.2. Production of Tapes and Split Yarns

The stretching process in which the film strips are drawn to many times their original length (in most cases, six to ten times) is the central processing step. It leads to high molecular orientation in the longitudinal direction and to substantial increase in strength with simultaneous decrease in break elongation. However, strength in the transverse direction clearly decreases. At high draw ratios, this effect can cause the tape to "split," i.e., to fibrillate even at low mechanical stress. This phenomenon, which is more pronounced for PP than for HDPE tapes, is undesirable in some cases, e.g., for bagging and packaging fabrics. In other cases, it is exploited and reinforced intentionally (e.g., by profiling the ribbons or by using fibrillators) to arrive at products with lower fineness and more textile-like properties (split yarns and fibers).

Figure 24 is a generalized scheme for the production of film ribbons and split yarns. The processing steps follow.

Production of Film. In general, extruders (a) with screw diameters of 90 or 120 mm and screw lengths $\geq$ 25 times the diameter are used. Screws with a compression ratio of ca. 1 : 3, with a shear and mixing section, have proved useful. Cooled groove boxes in the feed section lead to more uniform granulate feed and higher throughput. In this case, flat cut screws with combined high-efficiency mixing and shearing sections can be used. The type of film production as well as the conditions of cooling and extrusion affect processing and properties of the end product [186], [191], [192]. Tapes of HDPE are usually produced from tubular films (b_1), and

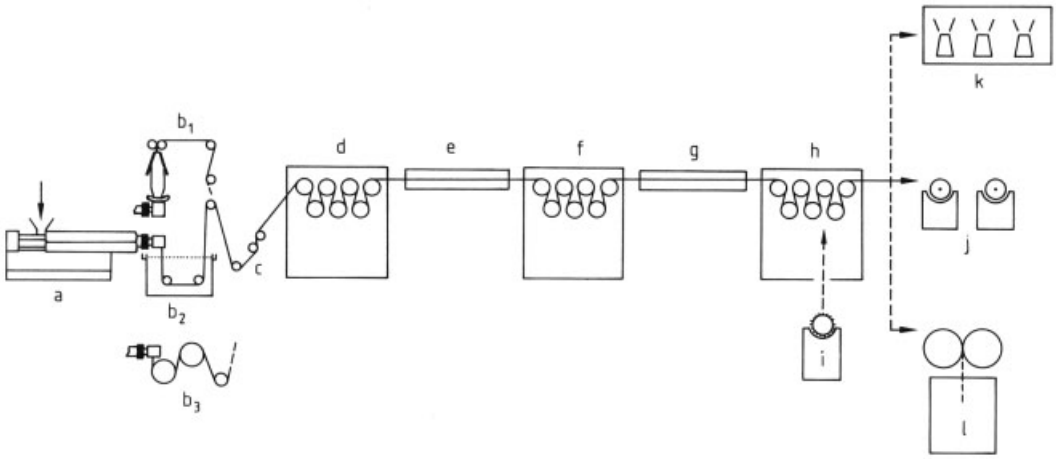

Figure 24. Production of film yarns and split fibers
a) Extruder; b_1) Tubular film (air cooled); b_2) Cast film (water quenched); b_3) Cast film (chill roll quenched); c) Knife beam; d) First godet stand; e) Stretching oven; f) Second godet stand; g) Setting oven; h) Third godet stand; i) Fibrillator (optional, only used for split fiber production); j) Windup; k) Twister; l) Staple cutter

PP tapes preferably from cast films. The cast film can be cooled in a water bath (b_2) at 15 – 40 °C. This type of cooling is particularly effective because it permits high operating speed, leads to the formation of small crystallites, and is advantageous with respect to stretching behavior and strength of the tapes obtained. Raw materials with low water carry over must be chosen, and wiping, suctioning, and squeezing devices must be provided so that no water is entrained in the stretching section or the edge strip return. Film production by the chill roll process (b_3) is used only when exceptionally close thickness tolerance and high tape uniformity are required. Throughput is normally lower in tubular film installations than in cast film units because of lower cooling intensity; in addition, larger variations in film thickness are to be expected.

Cutting. Normally the film is cut prior to stretching. This is accomplished with blades arranged at the desired distance on a knife beam (c). In principle, it is also possible to stretch the whole film and cut it only after stretching and setting. The high splitting tendency of PP makes it difficult to obtain a cut completely free of nicks, so only HDPE is processed this way [193], [194].

Stretching. Stretching is accomplished by guiding the film strips over mechanically driven cantilevered feed rolls. The stretching unit is made up of two sets of godets (d) and (f) usually consisting of seven godets each. The second set runs at a higher speed than the first; the draw ratio is determined by the difference in speeds. Draw ratios from 1 : 6 to 1 : 8 are used to produce tape for weaving and warp knitting; for rope, cable, and twine, draw ratios are ca. 1 : 9 to 1 : 10. Operating speeds are ≤ 300 m/min.

Stretching in the short gap between rolls occurs mainly through reduction in thickness without appreciable loss in width. This leads to some orientation in the transverse direction and, hence, a reduction in splitting tendency. Stretching is generally carried out in hot air ovens (e) or on heated plates.

The overall stretching process is exothermic. However, heat must be supplied to initiate it and to assure uniformity at the high speeds required. The required temperature depends on the raw material, the operating speed, and the desired tape properties; for the most part, it is between 130 and 180 °C. Whereas the temperature of the tape itself cannot be measured, the uniformity of stretching conditions can be checked by measuring yarn tension [195].

Setting. After stretching, setting (g), i.e., controlled shrinkage, is accomplished between the second and third godet stand by again supplying heat (in a hot-air oven or with the aid

of heated rolls). In this annealing process, the orientation stresses introduced during stretching are relieved. In so doing, subsequent undesired and uncontrolled shrinkage is reduced or prevented.

The setting temperature is either equal to the stretching temperature or 10 – 20 °C below it. A higher setting temperature leads to less shrinkage but adversely affects tape strength. Setting shrinkage results from the difference in drawoff speeds between the second and third godet stand. For most applications, adjusting the setting shrinkage to ca. 5 % is sufficient. For end uses in which particularly high dimensional stability is required (e.g., carpet backing), a higher setting shrinkage (10 – 15 %) must be selected.

Splitting and Fibrillating. Polypropylene tape has a relatively high tendency to split, and the propensity for fibrillation increases with increasing degree of stretching. The spontaneous splitting tendency is occasionally exploited in the production of twine and packaging yarns. The tapes split even during twisting (k) and give a fibrous, textilelike end product. The splitting tendency can be reinforced by adding other polymers [196].

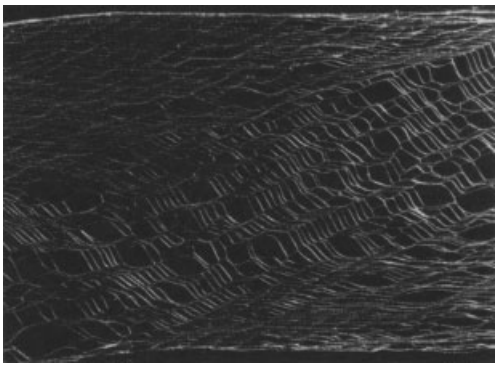

Figure 25. Networklike fibrillated yarn of propylene

Fibrillated yarns or fibers with still lower individual fineness and greater softness are obtained by special fibrillation processes. Stretched film strips are generally guided over a needle roll revolving at controlled speed (i); the direction of rotation of the roll coincides with the running direction of the tapes. The tapes are split into a networklike multiplicity of fibrils (Fig. 25). Depending on the number, fineness, and arrangement of the needles, as well as the speed of rotation of the roll, various fibrillar structures can be obtained [196–200].

Strongly fibrillated tape yarn can also be obtained without needle roll from profiled primary film. Profiling is produced either by a comblike jet lip (Barfilex process [201]) during extrusion or by roll embossing [202], [203]. In both cases, the impressed longitudinal grooves act as preset break points, which tear at the high draw ratios used (ca. 1 : 10). However, these processes are of minor practical importance compared to needle roll fibrillation.

Split fibers are obtained by cutting the fibrillated tapes in staple cutters (l). This process, however, is used rarely because of the development of short spinning technology to produce PP spun fibers of appropriate deniers at a reasonable cost.

End use for fibrillated tapes (split yarns) include twine, rope, reinforcement and weft yarn in woven carpets, and coarse yarn for geotextiles.

Coloration and Finishing. Tapes and split yarns are almost exclusively bulk-dyed with the aid of pigment concentrates, which are added prior to extrusion. The problem of coloring polyolefins by bath dyeing has not yet been solved satisfactorily.

Inorganic compounds such as chalk are added to reduce splitting and facilitate weaving. Ultraviolet stabilizers, required for many areas, are often provided by the manufacturer of the raw material, but the processor can also blend in concentrates. In general, UV stabilizers of the HALS type are used. They also impart good weathering resistance to PP ribbons [204]. However, HALS stabilizers may interact with certain pigments.

Antistatic agents and incorporated lubricants are generally not effective for ribbons and monofilaments. In addition, they can cause problems during film and ribbon production (e.g., plate out and water carry over). Subsequent external application (finishing) gives better results.

4.2.3. Uses

Tapes and split yarns of polyolefins are used in textile products that require high strength; low weight; chemical, moisture, and abrasion resistance; as well as other properties. Because of

the ease of processing and the low volume cost compared to other fiber raw materials, their use is especially economical in many cases.

To a large extent, PP and HDPE tapes have replaced jute, hemp, sisal, and other natural fibers. Film yarns and split yarns (tenacities commonly 4 – 6 cN/dtex) are used particularly in the manufacture of bags, flexible bulk containers and other packaging fabrics and knits; industrial fabric (e.g., tarpaulin, geotextiles, filter fabrics); carpet backing; netting (protection against hail, birds, solar radiation, etc.); awnings; camouflage shields; twine; rope and cable.

Packaging and Industrial Fabrics. Wide width fabrics are usually produced on flat looms [205]. Circular looms are useful for bagging, because in addition to simpler preparation for weaving (no beaming; weaving directly from the reel), elimination of the sideseam gives better bag strength and saves labor as well as material [206].

By subsequent extrusion coating, mostly with LDPE or PP, water- and dustproof fabrics are obtained. In addition, the coating prevents ribbon slippage and loosening of the fabric.

Carpet Backing. Because of the relatively high thermal stress during carpet manufacture (dispersion coating and drying), only PP ribbon can be used. The major application is as primary backing in the manufacture of tufted carpets; to a lesser extent, PP is used as secondary backing and as binding warp yarns in conventional woven carpets. The splitting behavior must be precisely controlled, especially in primary backings, and shrinkage must be low ($<1\,\%$ at 132 °C, 20 s, European Association for Textile Polyolefins method) [192]. Too little splitting leads to deflection of the tufting needles and, hence, defects in the carpet as well as breakage of individual tapes. Too much splitting also affects fabric strength adversely. Tenter setting, previously used to minimize shrinkage of the entire fabric, is employed only occasionally.

Knit Goods. Cloth for bagging or for shade, camouflage, and hail protection can be made on warp knitting machines as well as on looms [193], [207]. In most cases, netlike, coarse-stitch products, e.g., bags and pouches for fruit, vegetables, and other crops, are made.

The raw material is almost exclusively HDPE, frequently via the split-knitting process [193], [194]. Polypropylene tapes are difficult to fabricate on knitting machines because of poor sliding properties and tendency to split and break.

Twine, String, and Rope. Twine and string for packaging are obtained from highly stretched, spontaneously splitting, or fibrillated tapes of PP, by twisting and plying. For high-strength ropes and cables, fibrillated ribbons, PP monofilaments, and long staple fibers are used [208].

Floor Covering. Split fibers can, like other fibers, be processed into needle felts, which are used as floor or wall coverings. The use of split fibers in these areas has, however, decreased considerably and given way to spun fibers.

4.3. Monofilaments

4.3.1. Production

Monofilaments of PP or HDPE usually have a round or elliptical cross section, with a diameter of 0.1 – 3 mm (mostly 0.1 – 0.5 mm).

The steps in monofilament production are, for the most part, identical to those in tape manufacture. Monofilament machines consist of extruder, spinning unit (spinning head with spinning pump and spinneret), cooling bath, roll stretching units (usually two sets of seven rolls and one set of three rolls), hot-air ovens or a hot-water bath (for HDPE), and windups [184], [209]. Raw materials are generally propylene homopolymers with melt flow indices between ca. 2 and 6 g/10 min (MFI 230/2.16). If improved abrasion and rubbing resistance as well as better cold toughness are required, propylene copolymers are also used. Among the HDPE types, those with densities of 0.950 – 0.960 g/cm^3, MFI 190/2.16 values between 0.4 and 1.0 g/10 min, and narrow molecular mass distribution have been found to perform best.

Extruder – Spinneret. Extruders with a screw diameter of 45 – 90 mm are used for the production of monofilaments. Spinneret holes are arranged in concentric circles on the plate.

Depending on the size of the unit and the diameter of the monofilaments, ca. 50 – 100 monofilaments are extruded through one spinneret. In general, the spinneret hole diameter for polyolefins is three to four times the final diameter.

Cooling Bath. Monofilaments are produced exclusively with water bath cooling; the water temperature is 20 – 60 °C, depending on working material and thickness of the monofilaments.

Stretching. Stretching of PP monofilaments occurs in hot-air ovens between two stretching units with seven rolls each. The usual draw ratio is 1 : 8 to 1 : 12, and the temperature is 120 – 160 °C. Because of their lower softening range, HDPE monofilaments can be stretched in water baths at, or slightly below, 100 °C. Compared to stretching in hot-air ovens, this method has the advantage of better and more uniform heat transfer. Stretch ratios are generally between 1 : 8 and 1 : 11.

Setting. As in tape production, setting occurs in hot air after stretching. It is also done under tension, which is controlled by suitable adjustment of drawoff roll speed. Setting temperature is in the range of stretching temperature; setting shrinkage is usually adjusted to 4 – 8 %.

Windup. in contrast to film yarns (which are wound on packages with traverse laydown), monofilaments are usually wound in parallel on flanged disk spools. Windup speed is ca. 100 – 180 m/min. Monofilaments for bristles are wound on large-diameter reels or wheels to avoid too high a curvature.

Coloration and Finish. Coloring and finishing are similar to the process steps for tapes. However, addition of mineral fillers to reduce splitting is not required for monofilaments. To avoid fibrillation of the outer layers, surface finishes are applied (e.g., during twisting) to improve lubricity.

4.3.2. Uses

Monofilaments of PP and HDPE are processed into fabrics, knit goods, rope, and cable on the same machines used for film ribbons and split yarns.

Areas of application include fishing nets (HDPE); nets for protection from hail, birds, etc.; packaging nets; sieves and filter fabrics; rope and cable; bristles; reinforcing grids for construction sheeting; and tarpaulin.

Bristles for sweepers, road brooms, and the like are sometimes mechanically split at the ends. Because sufficient weathering resistance can be obtained with little or no UV stabilizer, HDPE is often used for protective netting.

4.4. High-Modulus Polyolefin Yarns

Conventional processes for producing polyolefin tapes, monofilaments, and fibers give a structure that consists of more or less oriented crystals as well as amorphous regions. Considerably higher values of the modulus of elasticity and tensile strength are to be expected if chain folding can be avoided, and as many molecular chains as possible can be oriented in the direction in which the load is applied. This ideal case can be approximated most closely with HDPE. Based on its relatively high packing density in the crystal lattice, HDPE should give the highest theoretical stiffness and strength value [210]. On the other hand, HDPE can be relatively easily prepared with very high molecular mass, so the number of chain ends and defects in the crystal can be minimized. The theoretically attainable stiffness and strength values of polypropylene are lower because of its helical structure. However, PP is also of interest because of its higher heat resistance.

The necessary "ultrahigh" orientation of the molecular chains can be achieved essentially in two ways with high molecular mass polyolefins: (1) drawing in the solid state (at a temperature clearly below the melting range) [211–214], and (2) spinning of highly viscous solutions (gel spinning) and subsequent one- or two-stage drawing [215], [216].

Higher modulus and strength are attainable by drawing in solution than by drawing in the solid state because products with higher molecular mass can be used, and a higher degree of orientation can be achieved.

4.4.1. Production

Two technologies are available; both are still in the process of commercialization. Drawing in the solid state is accomplished in draw jets at ca. 100 °C [217]. At draw ratios up to 1 : 30, moduli of ca. 50 GPa are attained for HDPE and ca. 20 GPa for PP. This process is being commercialized by the British Petroleum (BP) and Celanese companies.

In gel spinning (Fig. 26), a highly viscous solution in Decalin, paraffin oil, or paraffin wax is first prepared in a kneading or stirring unit [218]. It is then extruded through a spinneret (c) into a prepicitating bath (d). After entering the precipitating and cooling bath, the filaments are in a gel state in which the molecules remain largely disentangled. For orientation in the fiber direction and subsequent crystallization with final removal of solvent, a single- or multiple-step drawing process follows (e). Drawing occurs under heat and with a stretch factor $\geq$30. Preferably, HDPE with a molecular mass of ca. 2×10^6 is used (UHMMPE). The yarns can subsequently be cross-linked using radiation.

4.4.2. Properties and Uses

By gel spinning, HDPE yarns with moduli of elasticity of 50 – 100 GPa, and tenacities of $\leq$ 35 cN/dtex can be produced [219]. Because of their low density (0.97 g/cm^3), these yarns have an extremely high specific strength and are very abrasion, light, chemical, and moisture resistant. In addition, they exhibit high energy absorption and knot strength as well as good resistance to cold. Limitations in use include a relatively high tendency to creep and a relatively low melting temperature (ca. 150 °C). Radiation cross-linking increases heat resistance and reduces the tendency to creep.

End-use possibilities are extremely high-strength rope and cable, fabric for sails, industrial filters, bulletproof vests, and reinforcing layers for composites.

Trade names: Deyneema (DSM – Toyobo, The Netherlands – Japan); Spectra (Allied, United States); Tekmilon (Mitsui, Japan)

5. Polyacrylonitrile Fibers

"Polyacrylonitrile fibers" is a collective name for all fibers that contain at least 85 % polymerized acrylonitrile [*107-13-1*]. These fibers are also referred to as "acrylic fibers." Most polyacrylonitrile [*25014-41-9*] (PAC) fibers consist of ternary copolymers with 89 – 95 % acrylonitrile, 4 – 10 % of a non-ionogenic comonomer, and 0.5 – 1 % of an iogenic comonomer containing a sulfo ($-SO_3H$) or sulfonate ($-OSO_3H$) group. Fibers of 100 % acrylonitrile are also produced for industrial use [220]. Because polyacrylonitrile decomposes before melting, it cannot be spun from the melt but only from solution.

History. In the early 1930s, H. Rein at I.G. Farbenindustrie carried out experiments to produce fibers from polyacrylonitrile. Because of its insolubility in organic solvents, its nonmelting properties, and its high softening point, advantageous fiber properties were expected. Patents for the use of dimethylformamide (DMF) as a solvent for acrylonitrile polymerization were filed in Germany in 1942 by I.G. Farben and in the United States in 1944 by Du Pont. It was the first industrially useful solvent for the production of polyacrylonitrile fibers. Soon thereafter, dimethylacetamide (DMA) was also found to be a solvent. Technology for dissolving the polymer in aqueous ions of certain organic salts was also developed. The first acrylic fiber was commercialized by Du Pont in 1948. In the Federal Republic of Germany, production of acrylic fibers began in 1954 [221].

The composition of the copolymer (cf. Section 5.3.1) determines the dyeing behavior of the fibers. The non-ionogenic comonomer serves as an internal plasticizer and increases the rate of dye uptake. The sulfonate group of the ionogenic comonomer makes possible chemically bonding ionogenic dyes, whose color-imparting ion is a cation (basic dyes). The amount of iogenic comonomer determines the saturation value, i.e., the maximum amount of dye that can be chemically bonded by the fibers. Acidic ionogenic comonomers are replaced by basic comonomers if use of acid dyes is desirable.

Basic-dyeable acrylic fibers are the major type of fiber consumed; those made of 100 %

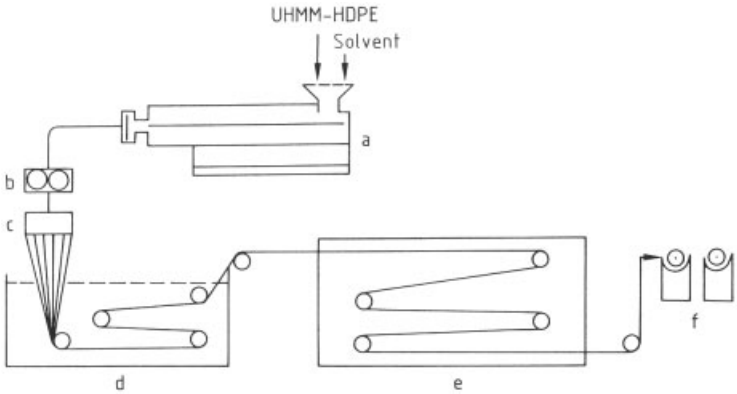

Figure 26. Gel-spinning process for the production of high-modulus UHMM HDPE ∗ yarns
a) Extruder; b) Spinning pump; c) Spinneret; d) Precipitating bath – extraction; e) Heating – drawing; f) Windup
∗ Ultrahigh molecular mass high-density polyethylene.

acrylonitrile and those dyeable with acid dyes lag far behind.

The term "modacrylic" designates all fibers that contain 50 – 85 % of bound polyacrylonitrile. Fibers with 40 – 60 % copolymerized vinyl chloride and fibers with 30 – 45 % copolymerized vinylidene chloride are important because of their flame-retardant properties.

Polyacrylonitrile fibers are wool-like in nature. An important step in their development was the improvement in dyeability by modification with suitable comonomers. Their wool-like nature and good dyeability have established a secure place for acrylic fibers in the textile market along with polyester and polyamide fibers.

World production of acrylic fiber increased from 1 449 000 t/a in 1974 to 2 334 000 t/a in 1984 and to 2 383 000 t/a in 1992. In 1984, polyacrylonitrile fibers represented 14.9 % of all synthetic fibers produced and 6.5 % of total fiber production (synthetics, cotton, and wool); world production of polyacrylonitrile fibers exceeded wool production (1 672 000 t) by 662 000 t [222].

5.1. Physical Properties

Textile Characterization. Articles of acrylic fibers have a wool-like hand. Lower fineness fibers produce a softer hand; higher fineness make it harsher. The high bulk and low density (compared to natural fibers) make possible light, airy textiles with good recovery, warmth retention, and covering power. Because acrylic fibers do not swell, articles made from them remain permeable to air even when wet. The tensile strength of acrylic fibers and the abrasion resistance of articles made from them are greater than those of natural fibers.

Commercial fibers have different dye affinities and dye uptake rates. For light shades, dispersion and basic dyes are used; for dark shades, only basic dyes. Dyeing temperature is generally ca. 100 °C. Dyeings are distinguished by brilliance and fastnesses. The degree of luster is determined by the type, brilliant or delustered.

Articles of acrylic fibers fall into the easy-care category. They absorb little moisture, can be washed at moderate temperature, and dry rapidly. They are resistant to many organic solvents, and stains can be removed easily. They do not felt. Weathering and lightfastness are excellent and unsurpassed by any natural or synthetic fiber. This makes the fiber especially suitable for outdoor uses such as awnings. The fiber is not attacked by bacteria, mildew, or insects, and does not rot.

5.2. Chemical Properties [223]

Good chemical resistance of the homopolymer to hydrolysis permits industrial applications up to 140 °C in a humid atmosphere, e.g., for filter fabrics. The resistance of acrylic fibers to

industrial gases, smoke, and soot is very good. Acrylic fibers are resistant to moderate concentration of mineral acids and to oxidizing agents. They are soluble in concentrated sulfuric acid and concentrated nitric acid. Resistance to weak alkali is sufficient. Acrylic fibers are insoluble in most organic solvents, e.g., benzene, mineral oil, carbon tetrachloride, aliphatic hydrocarbons, aliphatic alcohols (e.g., ethanol), aliphatic esters, and aliphatic ketones (e.g., acetone).

In the burn test, the fibers melt, ignite, and burn with a yellow, sooty flame; a dark, hard, and brittle residue remains. In contrast, modacrylic fibers melt in the flame to dark, hard lumps; burn partially; but extinguish again if the flame is removed.

For a summary of the properties of acrylic and modacrylic fibers, see → Fibers, 1. Survey, Chap. 5.

5.3. Production

Production of polyacrylonitrile fibers includes the following processing steps (Fig. 27) [224]:

1) Polymerization (Section 5.3.1)
2) Dissolving (Section 5.3.2)
3) Spinning (Section 5.3.3)
4) Aftertreatment (Section 5.3.4)

5.3.1. Polymerization

The fiber raw materials for acrylic or modacrylic fibers are produced by polymerization or copolymerization of acrylonitrile (→ Acrylonitrile). For further details concerning reactivity and material constants see [225].

Polyacrylonitrile can be made by anionic or radical polymerization. Although polymer solutions can be produced in high space–time yields by anionic polymerization, radical-initiated polymerization or copolymerization of acrylonitrile is generally the method of choice in the manufacture of raw materials for acrylic or modacrylic fibers. Either solution polymerization or precipitation polymerization is used.

Like all vinyl monomers, acrylonitrile can polymerize with the formation of isotactic as well as syndiotactic sequences. The processes used to produce acrylonitrile polymers give about equal amounts of isotactic and syndiotactic sequences [226].

For commercially useful acrylic fibers or yarns, homo- or copolymers of acrylonitrile are used as raw materials. The raw materials for acrylic fibers in *textile applications* are acrylonitrile terpolymers.

The same initiator systems and similar reaction conditions can be used for the industrial unipolymerization of acrylonitrile as are used to produce acrylonitrile copolymers, the raw materials for the usual acrylic fibers in textile applications.

Interest has developed in very high molecular mass acrylonitrile copolymers that contain small amounts of comonomers. These materials are used to produce high-strength acrylic yarns [227].

To obtain the desired properties in acrylic fibers for textile applications, acrylonitrile terpolymers are used as raw materials. These polymers contain 4 – 10 % comonomers, are free of ionic groups, and are particularly suited for statistical copolymerization with acrylonitrile. Examples are the methyl esters of acrylic or methacrylic acid, vinyl acetate, or acrylamide. Depending on the need for raw materials for cationically or anionically dyeable acrylic fibers, an acidic or basic comonomer is used as an additional component (cf. Section 5.5). Monomers containing sulfonic acid groups have proved particularly usefull as acidic or anionic comonomers, e.g., methallyl sulfonic acid, styrene sulfonic acid, and 2-methyl-2-sulfopropylacrylamide, or soluble salts of these sulfonic acids, respectively. Examples of basic comonomers are dialkyl aminoalkyl methacrylates or vinylpyridines.

Although the processing technology of the terpolymerization reaction is largely independent of monomer composition, special attention must be given to the structure of the dyeing comonomer when selecting the initiator or initiator system.

Precipitation Polymerization. In the precipitation polymerization of acrylonitrile in water to produce acrylic fiber raw materials, use is made of the relatively good water solubility of the monomer. To introduce end groups into the fiber raw material, which can be utilized to adjust dyeability properties initiator systems are

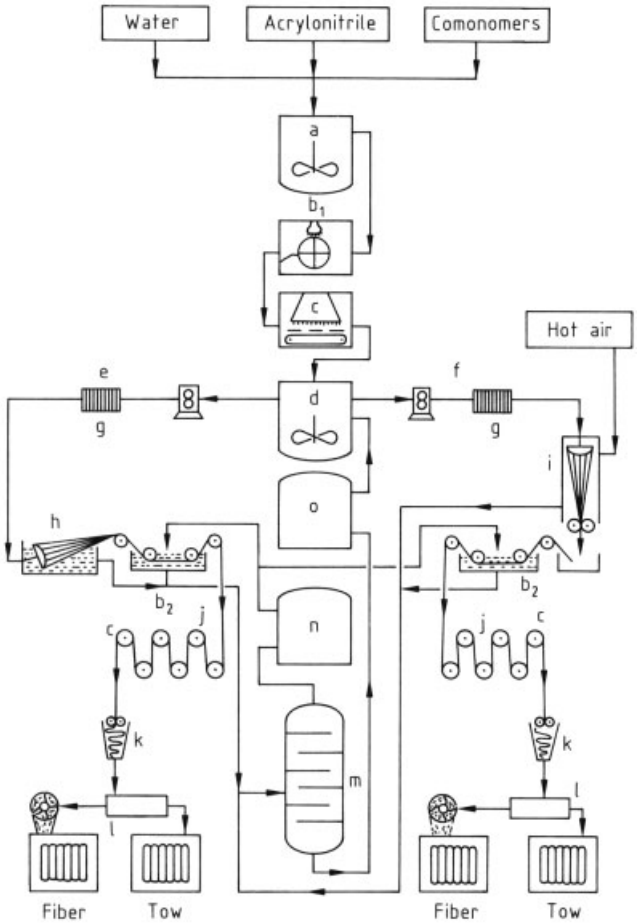

Figure 27. Manufacture of acrylic fiber ∗ [224]
a) Polymerization; b_1) Washing and filtering of polymer; b_2) Washing of filaments; c) Drying; d) Dissolving; e) Wet-spinning process; f) Dry-spinning process; g) Filter; h) Spinning bath; i) Spinning shaft; j) Drawing, drying; k) Crimping; l) Steaming; m) Distillation; n) Water; o) Solvent
∗ Wet spinning – lower left; dry spinning – lower right.

used. The polymerization is initiated with water-soluble peroxy disulfate redox systems, such as potassium, ammonium, or sodium peroxydisulfate, combined with water-soluble salts of sulfur (IV) or derivatives of sulfur (II). Especially useful in the production of raw materials for cationically dyeable acrylic fibers is the initiator system peroxydisulfate – hydrogen sulfite. The decomposition reaction of this redox system has been thoroughly studied [228].

The relationship between dyeability and the number of different end groups in the copolymers is known [229]. The type and concentration of polymer end groups of the fiber raw material have a major influence not only on the coloring properties but also on the resistance to thermally induced yellowing and the degree of whiteness of the resulting acrylic fibers.

To obtain cationic dyeable acrylic fibers, precipitation polymerization is carried out at 40 – 70 °C and at a pH of <6. Peroxydisulfate – hydrogen sulfite is used as initiator of polymerization with a severalfold molar excess of hydrogen sulfite relative to peroxydisulfate. To purify and demonomerize the polymer completely and

gently, a compact polymer granulate, which can be dried after washing and filtering, is produced.

In the synthesis of fiber raw materials for acrylic fibers that are to be dyed with anionic dyes, redox systems, e.g., H_2O_2 – thioglycerine [230], that do not result in anionic end groups are used.

A detailed description of redox polymerization of acrylonitrile and a comprehensive review of the use of redox systems can be found in handbooks [231]. The constancy of distribution of molecular mass, chemical composition, and degree of whiteness is particularly significant to the final properties of acrylic fibers [220, p. 200].

Solution Polymerization. The copolymerization of acrylonitrile in solution results in polymer solutions that can be spun without additional process steps after the elimination of residual monomers. The polymerization reaction conditions must be optimized so that monomer conversion is almost complete; low molecular mass fractions that could adversely affect the mechanical and thermal properties of the resulting acrylic fibers must be avoided as much as possible. Examples of suitable solvents are dimethylformamide, dimethyl sulfoxide, nitric acid, and concentrated aqueous solutions of ammonium thiocyanate or zinc chloride. In radical polymerization, these solvents – but especially the polymer solutions – are effective chain-transfer agents. For that reason, specially selected initiator systems and special reactors are used for the large-scale commercial solution polymerization of acrylonitrile [232].

The choice of precipitation polymerization or solution polymerization often depends on the prior history of the acrylic fiber producer. Both processes have advantages and disadvantages, and economics – size of installation, type of spinning process in use and raw material situation – is often the deciding factor.

5.3.2. Dissolving

The solvents used in dry and wet spinning are listed in Table 15.

Commercial process. A dough, which is as free of clumps as possible in order to avoid long dissolution times, is prepared from the powdery polymer and the solvent by growing a compact polymer granulate; intimate mixing of polymer and solvent stream, e.g., in screw mixers, turbines, or paddle mixers; and, providing low solvent temperature, which initially produces a suspension with good distribution of polymer in the solvent.

A two-step process is used for dissolution in aqueous salt solutions. The polymer is first suspended in a salt solution that does not dissolve the polymer; then the salt concentration is increased to the amount necessary for dissolution. This process takes place in agitating pans, either batchwise or continuously. The batchwise process requires several hours to produce dimethylformamide solutions. The dissolving temperature is between 60 and 100 °C. In the continuous process, the holdup time in the agitating pan is shorter (e.g., 5 – 30 min), but the temperature of dissolution is correspondingly higher (e.g., up to 120 °C) and the intensity of agitation is greater [233]. The solution can then be treated thermally in heat exchangers or exposed to shearing forces in screws or turbines.

The concentration of the spinning solutions is higher for dry spinning than for wet spinning. When dimethylformamide is the solvent, 25 – 30 % solutions are normally used for dry spinning and 20 – 25 % solutions for wet spinning. Concentration and viscosity of the spinning solution are monitored constantly.

To keep the spinning solutions free of gel particles or other impurities, which is a prerequisite for long spinneret life, they are filtered in filter presses. Heat, amines, oxygen, and ions of heavy metals (especially iron) damage initial color (degree of whiteness). Therefore, an inert atmosphere, acid, and reducing and complexing stabilizers are needed [234], [235]. Additives have also been developed to protect against salt-like clouding of the solutions (in dimethylformamide) because of the polymerization catalyst [236], [237].

Stable viscosity is desirable so that spinning remains a steady-state process; it is obtained by optimizing the temperature of dissolution (ca. 70 – 100 °C) or by using suitable additives [238], [239].

Gelation occurs at low temperature via secondary valence forces and at high temperature via cross-linking reactions. The higher the viscosity of the solution and the molecular mass of

Table 15. Distribution of installed capacity by solvents, 1985

Solvent	CAS registry number	Formula	Spinning process	Percent of total capacity
N,N-Dimethylformamide (DMF)	[*68-12-2*]	$HCON(CH_3)_2$	dry wet	15 17
Sodium thiocyanate*	[*540-72-7*]	NaSCN	wet	24
N,N-Dimethylacetamide (DMA)	[*127-19-5*]	$CH_3CON(CH_3)_2$	wet	24
Nitric acid	[*7697-37-2*]	HNO_3	wet	12
Zinc chloride*	[*7646-85-7*]	$ZnCl_2$	wet	3
Dimethyl sulfoxide (DMSO)	[*67-68-5*]	$(CH_3)_2SO$	wet	2
Ethylene carbonate	[*96-49-1*]		wet	3

* Aqueous solution.

the polymer, the greater is the tendency to gel; more uniform polymers have less tendency to gel [240].

Aqueous zinc chloride solutions and solutions in dimethyl sulfoxide and ethylene carbonate also tend to gel. Because partial hydrolysis of the nitrile groups occurs when nitric acid is used as the solvent, these spinning solutions should not be heated above 5 °C.

5.3.3. Spinning

The polymer contained in the spinning solution is shaped into a yarn by either wet or dry spinning. In both processes, the spinning solution is pressed through small holes in the spinneret, thus giving the polymer its threadlike shape. The processes differ in the state of aggregation of the medium that the spinning solution enters on leaving the spinneret. In wet spinning, a precipitating bath of water and the solvent is involved; in dry spinning, hot air or a hot inert gas.

Wet Spinning. The solvents used industrially for wet spinning are listed in Table 15. Examples from the patent literature of spinning conditions for various solvents are given in Table 16.

The 10 – 25 % polymer solutions are deaerated, filtered, and pumped through the 10 000 – 80 000-hole spinnerets into the precipitating bath, where the solution thread line is coagulated to a highly swollen gel yarn by the exchange of solvent in the yarn for water. Coagulation conditions determine the strength of the sheath – core structure of the yarn, as well as its porosity and cross-sectional shape. Wet-spun yarns are aftertreated immediately without intermediate storage.

Dry-Jet Process.

In a variation of the wet-spinning process, the solution thread line leaving the spinneret is first passed through air before entering the precipitating bath. Fibers with higher strength and elongation, smoother surface, and greater uniformity can be made by this process [241].

Dry Spinning. The only solvent used in dry spinning of polyacrylonitrile fibers is dimethylformamide having the following properties:

bp (101.2 kPa)	153 °C
Heat of vaporization	571.24 kJ/kg
Water	<0.3 %
Amines	<0.01 % dimethylamine
Iron	<0.05 ppm

The pH of a 20 % aqueous solution at 25 °C should be 6.5 – 9.

The degassed and filtered spinning solution is heated to 100 – 150 °C and spun into vertical spinning shafts through spinnerets (200 – 2000 spinneret holes; hole diameter, 0.1 – 0.5 mm). Several spinning shafts make up a spinning machine. The walls of the shafts are heated to 150 – 220 °C; hot air or a hot inert gas (200 – 240 °C),

Table 16. Solvents and conditions for wet spinning

Solvent [a] (wt %)	Concentration [b], wt %	Composition [c], wt %	Temperature,[c] °C
DMF (100)	17 – 25	DMF (40 – 60)	
			5 – 25
		H_2O (60 – 40)	
DMA (100)	20	DMA (40 – 65)	20 – 30
		H_2O (60 – 35)	
DMSO (100)	20	DMSO (50)	10 – 40
		H_2O (50)	
NaSCN (50)	10 – 15	NaSCN (10 – 15)	
			0 – 20
H_2O (50)		H_2O (90 – 85)	
$ZnCl_2$ (54)	10	$ZnCl_2$ (14)	25
NaCl (4)		NaCl (1)	
H_2O (42)		H_2O (85)	
HNO_3 (70)	10 – 13	HNO_3 (30)	3
H_2O (30)		H_2O (70)	
Ethylene carbonate (90 – 85)	15 – 20	ethylene carbonate (20 – 40)	40 – 90
H_2O (10 – 15)		H_2O (80 – 60)	

[a] DMF: *N,N*-dimethylformamide; DMA: *N,N*-dimethylacetamide; DMSO: dimethyl sulfoxide.
[b] Of the spinning solution.
[c] Of the precipitation bath.

passes through the shaft in the same direction as the yarn. The oxygen and water content of the spinning gas should be as low as possible to avoid discoloration of the spun yarn and decomposition of dimethylformamide. Dimethylformamide evaporates in the spinning shaft and is exhausted together with the spinning gas at the bottom of the shaft. The performance of a spinning shaft is, in the final analysis, determined by the geometry of the engineering apparatus and by the amount of heat transferred to the spinning threads via the hot gas and radiation from the heated walls of the shaft. Of special significance for the quality and performance of the spun material is the air guidance in the immediate vicinity of the spinneret [242–244].

Solidified, dried spun yarns are drawn off at a rate of 200 – 500 m/min. Spun yarns from the shafts of one spinning machine are collected and deposited in a can. (In the manufacture of continuous filaments, spinnerets with a maximum of 200 holes are used. In this case, an oily finish is applied to the spun material below the spinning shaft, and it is wound up on bobbins.)

The dimethylformamide that has evaporated is condensed, distilled, and recycled. The spun material still contains 7 – 30 % DMF, which is leached out during aftertreatment and recovered by distillation.

The spinning conditions (temperature of solution, temperature of shaft, temperature and amount of spinning gas, and drawoff speed) greatly affect both further processing of the spun yarns (e.g., drawability) and properties of the end product (e.g., strength and elongation).

Melt spinning. Mixtures of polyacrylonitrile and water (ca. 13 – 33 % relative to the polymer), in contrast to anhydrous polyacrylonitrile, can form a molten, single-phase system without decomposition of the polymer. Fibers from a polyacrylonitrile – water mixture were prepared as early as 1948 and again around 1970 – 1971 [220, p. 260 ff]. The advantage of such a process could be the absence of solvent and lower costs. However, more than 15 years of research and development have not yet resulted in a commercial process. Possible reasons for this are the high viscosities of the melts and the difficulty of preventing the melt from becoming two-phase at any point in the process.

5.3.4. Aftertreatment

The aftertreatment steps are as follows:

1) Washing
2) Drawing
3) Finish application
4) Crimping
5) Drying
6) Cutting
7) Steaming
8) Packaging

Cutting is omitted in the production of tow.

During aftertreatment, the spun material acquires the necessary physical properties, such as strength, modulus, elongation, shrinkage, crimp, and crimp stability. Large-scale aftertreatment units are capable of processing tow up to 770 ktex (770 g/m) at rates up to 120 m/min in dry spinning and up to 80 m/min in wet spinning. The daily production of a large unit is 110 – 120 t; smaller units (30 t/d) are also used [245].

Washing. Washing removes residual solvent and salts by leaching with hot water (e.g., 80 °C) in multilayer or drum washing machines. A closed construction is used to protect the environment. The washing machine consists of many individual zones. The wash water travels in a direction opposite to the direction of travel of the yarn tow. The wash water is distilled or otherwise processed to recover the solvent. Washing usually precedes drawing because of the longer holdup time in the washing baths resulting from the relatively low tow speeds prior to drawing, which makes the washing process correspondingly more intense.

Drawing. Drawing imparts the required strength to the yarn and occurs in two or more steps in drawing tanks between drawing units. The preferred drawing medium is water just below the boil (e.g., 99 °C). The draw ratio is between 1 : 2 and 1 : 10. Wet-spun yarns must be drawn more (because of the absence of preorientation) than dry-spun yarns to attain the same strength. The rule "the higher the draw, the higher are the strength and modulus and the lower is the break elongation" applies to both processes.

After being drawn, the yarns have a boil-off shrinkage, which, depending on the draw ratio, is between 15 and 45 %. Shrinkage increases as draw temperature decreases and passes through a maximum with decreasing draw ratio. This behavior is applied to production of shrinkable fibers. Fibers already shrunk can be made shrinkable by a postdrawing treatment.

In addition to water drawing, steam drawing or drawing on feed wheels is also possible.

Finishing. Finish application occurs in an immersion bath. Most commonly, active anionic compounds are applied to the fibers from an aqueous medium. Examples are ethoxylated fatty acids, salts of ethoxylated sulfonic acids, or phosphoric acid esters of long-chain aliphatic alcohols. Finish application prevents the buildup of static charges during further processing of the fibers and imparts the proper degree of adhesion and smoothness to the surface. The finish content of the finished fiber is 0.2 – 0.4 %. It can be determined by extracting with ether or a mixture of methanol – benzene, evaporating the solvent, and weighing.

Crimping. In general, crimping is carried out in a stuffer box after drying (stuffer box crimping). A steaming zone, in which saturated steam condenses on the tow, heats the tow and makes it more formable. Two crimping rolls, kept at constant temperature by circulating cooling water, pull tow bands of up to 3×10^6 dtex into the crimping chamber, stuff them, and push them continuously from the chamber at a predetermined pressure. In certain crimping units, steam can be introduced into the stuffer box to shrink the tow and increase the crimp stability (setting).

The type of crimp is important in processing fibers because of its effect on cohesion and on the hand of the yarns produced from the fibers. Crimp is determined by the conditions of crimping (temperature and stuffing box pressure), the percentage of comonomer in the polymer, and other factors that affect the thermal plasticity or friction, e.g., moisture content or type of finish.

Drying. Screen drum dryers, screen belt dryers, or calenders are used for drying. Drying temperatures of 120 – 170 °C are common.

During tensionless drying, i.e., drying loosely laid tow bands or cut fibers, the shrinkage built into the fiber in the drawing process is released and removed. This process can be enhanced by steaming. When dried under tension, the fibers

must subsequently be relaxed, usually by steaming. Drying reduces the moisture content of the fibers to 1 – 2 % and increases their density.

Although polyacrylonitrile does not crystallize like polyamide or polyester, an improvement or increase in regions of near-order by drying or annealing can, nevertheless, be demonstrated by X-ray analysis. Because of the consolidating effect of drying, uniform drying is especially important for uniform dyeability of the fibers.

Steaming. Steaming generally occurs in screen belt steamers or in autoclaves. Steaming with saturated steam shrinks the fibers and increases the crimp stability. To ensure that the shrinking process is complete, a high and uniform steam atmosphere (>95 % steam) is required.

Steaming increases the elongation of the fibers and slightly reduces their strength. Steaming under pressure is especially effective in increasing the rate of dye uptake.

Cutting and Packaging. In the production of endless tow for processing on stretch – break machines, the crimped and dried tow bands are packaged in cartons or bales. The mass of such tows ranges from 50 to 110 g/m. The bundle mass is 100 kg per carton and up to 800 kg per bale.

In the production of staple fibers, the crimped and dried tow bands are cut into staple fibers with slotted wheels or rotary cutters. The usual cut lengths are 40, 60, 80, 100, or 150 mm, depending on end use. Cut fibers are compressed and packaged into bales of ca. 200 – 300 kg in packing presses.

The aftertreatment of *filament yarns* of polyacrylonitrile involves processes similar to those for tow and fibers: drawing and setting. Equipment is tailored to the particular filament count of the product. In addition, the filaments are twisted to tighten the yarn and packaged on cones, which are shipped at a mass of 1.5 – 2.0 kg. Commercial counts of industrial filaments are 220 dtex/96 filaments, 440 dtex/192 filaments, or 890 dtex/384 filaments.

5.3.5. Special Processes

Continuous Dry Spinning Process. Modern continuous dry spinning and aftertreatment processes have eliminated storage following the dry-spinning process [246].

Because of the relatively high spinning speed (compared with wet spinning), the tow speed after drawing must be 500 – 2000 m/min. A vibrating trough is used to transport the spun tow during washing. Drawing is carried out in steam; crimping, in an aerodynamic crimper with a gaseous medium (at 1.5 – 1.6 MPa and 50 – 210 °C). Finish application can occur before, during, or after crimping. The steaming process is carried out in a vibrating trough in a steamer; drying, in a belt dryer.

In this process, products of uniform quality are obtained; solvent emission and waste are reduced; and manufacturing costs are lower [246].

In one variant of this process, washing is also eliminated [247].

Producer Coloring. The manufacture of colored fibers is of economic interest if large amounts of one color are to be produced. Fiber producers can make colored fibers in the following ways:

1) Spin dyeing,
2) Gel dyeing,
3) Dyeing of spun material

Spin Dyeing [248]. Dyes, either cationic or pigment dyes, are added to the spinning solution prior to spinning.

The dyes are free of adulterants. They are soluble in the solvent used and should be insoluble in water (the subsequent washing medium). Dye solutions also contain polymer to adjust their viscosity to that of the spinning solution. In general, combinations of dye solutions are metered into the polymer solution to achieve a given color; they are homogenized by means of statically or mechanically operating mixers.

Ionic bonding of the color-imparting cation to the sulfite or sulfate groups of the polyacrylonitrile polymer is rapid and quantitative. Uniformly dyed tows with brilliant color and good textile fastness are obtained by both the dry-spinning and the wet-spinning processes.

Organic pigments are used when especially high lightfastness is required, e.g., in textiles used outdoors. The pigments are dispersed in a dilute solution of polymer. In so doing, stabilizers are added to prevent coagulation of the

dispersed dyes. This is followed by metering into the spinning solution as with cationic dyes.

Gel Dyeing. Dyeing during aftertreatment generally occurs while yarns are in the gel state after wet spinning. Dyeing is possible because cationic dyes rapidly diffuse into the highly swollen yarn. It can occur in the precipitating bath, the drawing tank, or a separate bath. It must be carried out prior to drying, which collapses the structure of the swollen yarn.

Dyeing by both the extraction method and the padding method has been patented. Important variables are temperature, concentration of dye, and time in the dye bath; these determine the amount of dye required and the wash- and light-fastness of the dyed material [249]. The choice of dyes is also important in the quality of dyeing [250].

Gel dyeing has the following advantages [251]:

1) Good dye uniformity within the batch and from batch to batch
2) Same tow quality as with undyed fibers because all physical perturbations are avoided
3) Economical batch size ($\geq$5 t) because hardly any material is wasted during dye change

Dyeing of Spun Material. Dry-spun yarns can also be dyed during aftertreatment, but the rate of dye uptake must be increased by suitable measures [252].

Uses. Fibers dyed in the course of fiber manufacture are used in clothing, home furnishings, and outdoor textiles. Yarns, pigment-colored during spinning, are recommended for outdoor textiles (awnings, sunshades, camping and terrace furniture, camping tents, and boat covers).

5.4. Analysis

Cross Section (Fig. 28). The cross-sectional structure of polyacrylonitrile fibers and modacrylic fibers is determined by the composition of the polymer, the spinning process, and spinning conditions. The shape of the cross section determines the luster, moisture regain, and mechanical properties of the fiber [253].

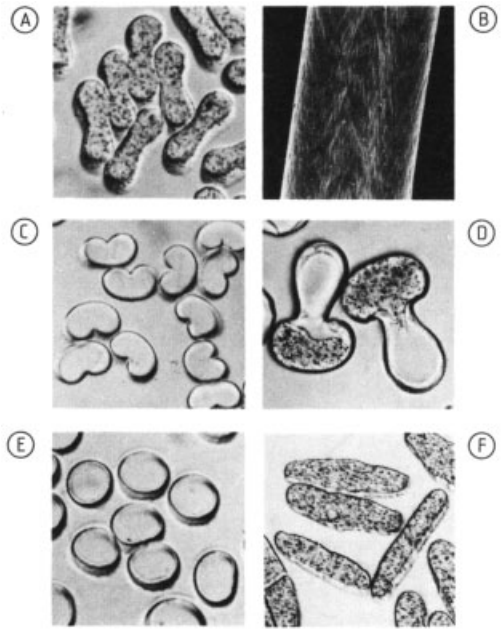

Figure 28. Acrylic fibers, cross sections and surfaces
A) Dry spun, delustered; B) Surface of A; C) Wet spun; D) Bifilar, dry spun, single component, delustered; E) Modacrylic, dry spun; F) Modacrylic, wet spun, delustered

Stress – Strain [220, pp. 73, 208]. The stress – strain curve of acrylic fibers (Fig. 29) is characterized below their glass transition temperature (50 – 95 °C with decreasing moisture content) by a steep initial region A, followed by zone B, where small changes in stress result in large changes in elongation. Above the glass transition temperature, A no longer exists; the fiber is more extensible, with slowly increasing resistance until a yield stress σ_c, at which structurally determined resistance to plastic sliding processes are overcome, is attained.

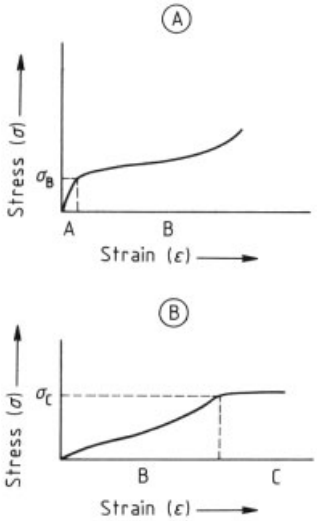

Figure 29. Stress – strain diagram of polyacrylonitrile (PAC) fibers [220]
A) Yield points below glass transition temperature $T < T_e$;B) Yield points above glass transition temperature $T > T_e T_e$ = second order transition temperature; σ_B and σ_C = points at which small increases in stress result in large changes in elongation

The stress – strain curve is affected by the type and amount of comonomer, the molecular mass and molecular mass distribution and the conditions of spinning and aftertreatment. Figure 30 shows the effect of the draw ratio.

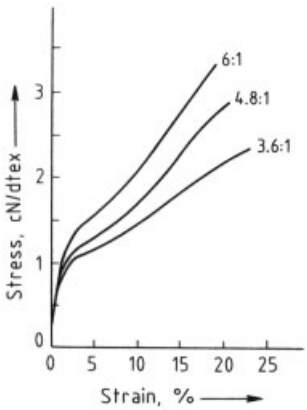

Figure 30. Effect of three different draw ratios on elongation and stress

Thermomechanical Analysis [220, pp. 77, 209]. In thermomechanical analysis (TMA), the change in length of a fiber under a constant preload is observed at a constant rate of heating. From the changes in length of fibers under low load, conclusions can be drawn concerning their degree of heat-setting. In Figure 31, brands 1 and 2 were more intensively heat-set during manufacture than brand 3, in which extension of the yarns above 110 °C is prevented, principally by release of shrinkage tension.

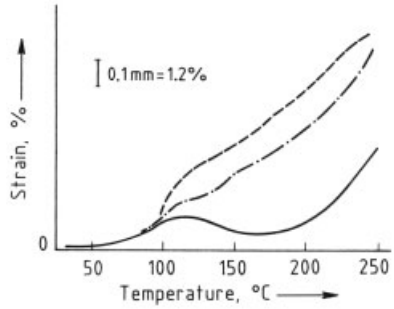

Figure 31. Thermomechanical deformation of acrylic fiber tow of three different brands ∗ [220]
– – – – Brand 1; — · — Brand 2; —— Brand 3
∗ Pretension 0.015 cN/dtex; rate of heating 10 °C/min.

Analytical Procedure. The *cross section* of the fiber material provides a good indication that an acrylic fiber is present. Dumbbell-shaped cross sections indicate a dry-spun acrylic fiber; beanshaped cross sections, a wet spun fiber (cf. Figure 28).

Type reactions can be used to recognize and distinguish an acrylic fiber from other synthetic fibers. Each of these reactions, must be carried out on a freshly prepared sample; acrylic fibers should be [223]:

1) soluble in concentrated nitric acid at room temperature,
2) insoluble in concentrated formic acid at room temperature, and
3) insoluble in boiling cyclohexanone.

Analysis of *polymer composition* is carried out by IR spectroscopy [254]. A first indication of the composition of the copolymer is obtained from analysis of the acrylonitrile content. To determine the brand, the exact composition of the polymer is required. The major neutral comonomers are vinyl acetate, methyl methacrylate, and methyl acrylate. Major ionogenic comonomers are styrene sulfonate, methallyl sulfonate, and itaconate, or in acid-dyeable types, cyclic or aliphatic amines. In addition, the following are present in flame resis-

tant types (modacrylic fibers): vinyl bromide, vinyl chloride, and vinylidene chloride [255].

The number of ionogenic groups can be determined titrimetrically. Acid groups can be determined potentiometrically in DMF solution with 0.1 M NaOH [256]; basic groups, in nitromethane with perchloric acid in dioxane solution [257]. The quantitative determination of sulfur-containing polymer building blocks can also be accomplished by ignition in an oxygen flask and determination of the sulfate formed or by determining the sulfur with X-ray fluorescence analysis [258]. Indications of the "physical history" of an acrylic fiber are obtained by physical methods, e.g., thermomechanical analysis (previously discussed) or by determining the degree of relaxation with a solubility test [259].

The percentage of acrylic or modacrylic fibers in fiber blends can be determined by selective dissolution in DMF or DMA [260]. For comprehensive information on analysis, see [220, pp. 379 – 423].

5.5. Types of Fibers [261]

Classification. The multiplicity of acrylic fibers can be classified by various systems. Chemical composition determines certain properties that are used in classification; fibers can be

1) basic dyeable,
2) dark dyeable,
3) acid dyeable,
4) self-crimping, or
5) flame retardant.

Characteristic descriptions, such as brilliant, delustered, bright, and spun-dyed, are also used.

Subdivision into spun fiber and tow accounts for various processing possibilities: (1) spun fiber for woolen, cotton, worsted, and open-end (rotor) spinning processes; and (2) tow for processing on stretch – break machines followed by spinning. Classification can be by fineness (denier) range (in decitex):

1) Finest: 0.4, 0.6, 0.9
2) Fine: 1.6, 2.2, 2.6
3) Medium: 3.3, 5.0, 6.7
4) Coarse: 11.0, 15.0, 17.0

Cut lengths are 28, 40, 50, 60, 100, 120, and 150 mm.

Physical modifications important to the quality of the goods are also used in classification. The following distinctions are made with respect to fiber shrinkage:

N-fiber	0 – 2 % shrinkage
S-fiber	ca. 20 % shrinkage
HS-fiber	ca. 40 % shrinkage

The type of crimp is another criterion that affects the quality of the goods. Qualitative data on crimp and crimp stability (low, medium, or high) serve as characterization parameters.

The range of finenesses for tow for stretch-breaking is 1.5 – 17 dtex. Tow can be bright, delustered, or spun-dyed.

The number of combinations of various characteristics for basic dyeable fibers alone runs into several hundred. Some inkling of the extraordinary diversity of possible applications of the most important synthetic staple fiber can be obtained by considering that the processor has available more than 20 different brands with varying properties, that polyacrylonitrile fiber types can be blended with each other or with other synthetic or natural fibers to achieve certain properties, and that the quality of the goods is substantially affected by the many possible yarn and fabric constructions.

Industrial Fibers. Fibers for industrial use consist mostly of 100 % or almost 100 % acrylonitrile. Molecular mass is made as high as processibility will permit because the higher the molecular mass, the higher the strength. The fibers absorb very little moisture and, therefore, do not swell. They are resistant to hydrolysis, especially in acid medium, and are heat resistant up to 140 °C.

Uses. Fibers are used in filter hoses and filter pockets in dry filtration, e.g., in the separation of fly ash in coal-fired power stations. Another end use is in dryer felts in paper machines, which absorb water that could not be squeezed out of the paper sheets. Fibers of 100 % acrylonitrile are also used as asbestos replacement. For this end use, particularly high modulus and strength are required [262]. Nonmelting fibers that are nonflammable in air are obtained by

heat-treating acrylic fibers in an oxidizing atmosphere at 300 °C. These fibers can also replace asbestos [263]. Filaments of 100 % polymerized acrylonitrile are good starting materials for high-strength carbon fibers [264].

Dyeing Modifications. Dyeability is generally achieved in polyacrylonitrile fibers by inclusion of a neutral, plasticizing comonomer (5 – 10 %) and an ionogenic comonomer (e.g., methallyl sulfonic, 0.5 – 1.0 %) that has an affinity for dyes. By increasing the amount of comonomer with dye affinity to 3 – 4 %, fibers are obtained with a four- to fivefold increase in rate of dye uptake. Such fibers can be dyed at a temperature as low as 85 °C rather than at the boil. Dye migration is especially good in such fibers, and the time required to fix dyes is markedly reduced. Use of a rapidly absorbing fiber is based on the differential dyeing effect, which occurs when the fiber is dyed with a conventional fiber. Special-effect yarns can be produced from homogeneous or inhomogeneous blends of fibers or rovings, e.g., yarns of mixed colors, slub yarns, and flake twist yarns. Combinations of yarns result in twisted yarns with multicolor effects.

Another dye modification is represented by acid-dyeable fibers that contain 5 – 15 % of a basic comonomer with affinity for acid dyes. Possible comonomers are the vinylpyridines, e.g., 2-methyl-5-vinylpyridine [*29994-42-1*], or even aliphatic amines, e.g., 2-(dimethylamino)ethyl methacrylate [*2867-47-2*]. Yarns of acid-dyeable acrylic fibers are processed and dyed together with yarns of basic-dyeable acrylic fibers. Dyeing is carried out in the same dye bath containing both acid and basic dyes. In this way, interesting color effects can be achieved such as two-color effects or tone-on-tone effects. If an undyed – colored combination is desired, only acid dyes are used, to which a conventional acrylic yarn is relatively insensitive.

Modacrylic Fibers [265]. Modacrylic fibers contain 50 – 80 % bound acrylonitrile. The comonomer that makes up the difference is essentially a halogen-containing monomer. Vinyl chloride [*75-01-4*] and vinylidene chloride [*75-35-4*] are the most frequently used halogen-containing comonomers. The main advantage of modacrylic fibers is their reduced flammability. Whereas polyacrylonitrile fibers melt in the flame, ignite, and continue to burn after removal of the flame, modacrylic fibers are self-extinguishing once the flame is removed. In addition, they do not exhibit an undesirable dripping of molten fiber. The burnt residue is solid and black. In case of fire a disadvantage is the formation of hydrogen chloride. Fibers containing vinyl chloride exhibit even less flammability than vinylidene chloride fibers.

Bromine-containing monomers, e.g., vinyl bromide or vinylidene bromide, are also used. The development of polymerizable phosphorus compounds has become known as well.

In addition to acrylonitrile and halogen-containing comonomers, modacrylic fibers generally contain other comonomers that are intended to improve the dyeability of the fibers. These are mostly monomers with acid groups, e.g., sodium styrene sulfonate [*2695-37-6*] or sodium methallyl sulfonate [*149-44-0*]. Neutral comonomers, e.g., alkyl acrylamides, which improve dyeability by loosening fiber structure, can also be included. The manufacture of modacrylic fibers in essence follows the same process steps as the manufacture of polyacrylonitrile fibers.

Because of the type and high percentage of comonomer, modacrylic fibers exhibit higher plasticity, which is expressed as lower sticking temperature or higher shrinkage (for other properties, see → Fibers, 1. Survey, Chap. 5.). Attempts were made to eliminate this disadvantage by reducing the halogen content of the polymer and including additives in the spinning solution that act synergistically to reduce flammability. Such additives are antimony trioxide, organic antimony compounds, or halogen compounds of antimony or tin. Halogen compounds can form the oxide and can be dispersed in the fiber so finely that they do not act as delusterants. Even compounds of metals, e.g., barium titanate, or phosphorus compounds, e.g., derivatives of the acids of phosphorus, serve as flame retardants.

Modacrylic fibers tend to lose their luster under hot – wet treatment, even during dyeing. Salts are added to the dye bath or dyeing is done at lower temperature to prevent this. However, the best route is to reluster the dyed fibers with dry heat. Luster stability has been improved by changes in spinning and aftertreatment technol-

ogy, by specific changes in polymer composition, and by the addition of luster stabilizers.

Uses. Reduced flammability favors the use of modacrylic fibers wherever there are fire hazards or where the danger in case of fire is particularly severe. This includes such uses as children's sleepwear; home furnishings, especially drapes; wall coverings in public buildings; and automobile upholstery. A large area of application is in pile fabrics, carpets, and imitation furs. Modacrylics are also used in wigs.

Modacrylic fibers are processed by themselves or in blends with other fibers, e.g., acrylic fibers, polyester fibers, or poly(vinyl chloride) fibers. By blending and proper construction of the finished article, special processing requirements, as well as the requirements of flammability tests, can be met.

Absorbent Fibers [220]. To produce fibers with greater wear comfort, special attention has been paid to developing moisture-absorbent fibers. Such fibers must be able to absorb moisture from the air and from water, either by virtue of their ability to swell or by means of a system of pores in the interior of each fiber. With acrylic fibers, the formation of a pore system is especially significant.

The system of pores must be protected by a sheath of proper thickness to ensure that the fiber can be processed without problem. The sheath must have a multiplicity of fine channels that can conduct water into the porous interior of the fiber. The properties of such a fiber include high adsorptivity, no swelling, high threshold of clammy feel, rapid transport of water, rapid evaporation, and low density [266].

The manufacture of absorptive fibers by both dry- and wet-spinning processes has been described. In dry spinning, an appropriate precipitant can be added to the spinning solution. This results in the formation of a sheath around the filament and an internal system of pores during spinning [267]. In wet spinning, highly swollen yarns with many pores are first formed during coagulation in the precipitating bath [268]. By skillful tuning of processing conditions, these pores can be retained in the final product. Water-soluble substances can also be allowed to diffuse into the swollen filaments, which are then dried. By subsequent leaching of water-soluble compounds, a system of pores is created.

Fibers with high swellability can be produced by the inclusion of comonomers with hydrophilic groups. Examples of such comonomers are acrylic acid, methacrylic acid, or 2-(dimethylamino)ethyl methacrylate. The saponification of comonomers on the fiber produces hydrophilic carboxyl groups. A popular route to the production of water-absorbent fibers is the spinning of mixtures of polyacrylonitrile and other polymers. Addition of cellulose acetate favors pore formation. Addition of hydrophilic polymers, e.g., saponified acrylonitrile, favors water absorption.

Bicomponent Fibers. The side-by-side type of bicomponent fibers in which both components are placed eccentrically in the cross section and run along the entire length of the fiber, is commercially important (see Fig. 28D). Differential shrinkage of the components forming the fiber produces three-dimensional crimp when the shrinkage is released. The differential shrinkage of the two components is achieved by a difference in the amount of comonomers. If the concentration of hydrophilic, usually ionic, groups differs in the two components, reversible crimp results. This refers to a crimp that can be removed by washing and regenerated by drying. If the concentration of hydrophilic groups of the two components is essentially equal, the crimp is irreversible, i.e., permanent. The shrinkage of the fiber, and hence the crimp, can be released during fiber manufacture by drying or steaming or during subsequent processing, e.g., dyeing of the fiber or yarn.

Bicomponent fibers are produced both as staple fibers and as tow. Knit fabrics of these yarns are distinguished by good volume, pleasant, wool-like hand, and good stitch definition. Bicomponent fibers are also used in carpets because of covering power [269], [270]. The market for these fibers is primarily in Japan and the United States.

Shrinkable Fibers. Shrinkable fibers can be produced from tow by processing on stretch-break machines. In general, the fiber shrinkage produced on these machines is 15 – 24 %. If suitable measures are taken during tow production and stretch-breaking, fiber shrinkages of 24 – 40 % are attainable.

Shrinkable staple fibers are generally offered at two levels of shrinkage: (1) shrinkable fibers with ca. 20 % shrinkage and (2) high-shrink fibers with 35 – 40 % shrinkage. The desired level of shrinkage can be obtained most easily during the manufacture of shrinkable fibers by suitable choice of draw ratio and draw temperature with subsequent mild drying. Care must be taken, however, to produce a shrinkable fiber free of voids so that no electrostatic problems occur in further processing and no changes in coloration or luster occur in the processed fiber. Good technical properties are also desirable. Criteria for a void-free structure are negative birefringence and density $>1.180\ g/cm^3$ [271].

Shrinkable fibers are processed in blends with nonshrinkable fibers to give high-bulk yarns, which are distinguished by their high volume. High-shrinkage fibers are also used in imitation furs as bottom hair, in the production of multistep plushes, and in felts. Manufacturing methods are described in detail in many patents [220]. Important measures are consolidation of the yarn by removal of solvent, drying, and heat treatment in water or steam, partly prior to drawing and partly between two drawing steps. The rate of drawing is also important.

5.6. Economic Aspects

With 2 334 000 t (Western Europe 784 000 t), acrylic fibers amounted to 19 % of synthetic fiber (polyester, polyamide, and polyacrylic) production in 1984. Acrylic fibers are produced almost exclusively as staple fiber and tow. Acrylic fibers represented exactly one-third of the synthetic staple fibers in 1984 (6 977 000 t) [222].

Table 17. Consumption of acrylic fibers in the EEC (1983)

Use	Consumption, t/a	%
From yarns		
Fabrics	93 600	22.4
Knit goods	205 000	49.1
Carpets	26 300	6.3
Other	43 700	10.5
From fibers		
Fiberfill	1 300	0.3
Nonwovens	5 300	1.2
Other	42 600	10.2

In the European Economic Community (EEC) (Belgium, Federal Republic of Germany, France, Italy, Luxembourg, The Netherlands, and Great Britain), the consumption of acrylic fibers in 1983 is shown in Table 17.

Acrylic fibers are used both by themselves and in blends with others fibers. The major end use of acrylic fibers in clothing is in sweaters; in home furnishings, they are used in furniture covers and blankets [222].

Acrylic fibers represent the largest percentage (33 %) of fibers processed on knitting machines (Western Europe, 1983), ahead of cotton with 22.8 %. Other areas in which acrylic fibers represent the leading raw material, ahead of all other fibers, are furniture pile fabrics (80.0 %), knitted and woven furs (79.8 %), and awnings (85.0 %) [224].

The United States used ca. 207 000 t of acrylic fibers in 1985. Preferred areas of application were sporting goods (ca. 24 %), sweaters (ca. 17 %), hosiery (ca. 13 %), and knitting yarns (13 %).

World capacity of acrylic fibers is ca. 2.8×10^6 t/a. About one-third of this capacity is in Western Europe (major countries: Federal Republic of Germany, Great Britain, Italy, and Spain) and another third in Asia (major countries: Japan, South Korea, Taiwan, and Turkey). Important countries in North America are the United States and Mexico, and in the Eastern Block, the former Soviet Union. The major producers are shown in Table 18.

6. Poly(vinyl chloride) and Poly(vinylidene chloride) Fibers

According to standard ISO terminology, the generic name for poly(vinyl chloride) [*9002-86-2*] (PVC) and poly(vinylidene chloride) [*9002-85-1*] (PVDC) fiber is *chlorofiber*; this designates products containing at least 50 % vinyl chloride or vinylidene chloride units. In the United States, the Textile Fiber Products Identification Act defines any fiber containing at least 85 wt % of vinyl chloride units as *vinyon* and any fiber composed of at least 80 wt % of vinylidene chloride as *saran*.

Compared to acrylic or polyester fibers, the 1987 world production and consumption of chlorofibers is small, ca. 20 000 t/a.

Table 18. Major producers of acrylic fibers

Trade Name	Producer	Country	Process
Acrilan	Monsanto	USA	wet
Aksa	Acrilik Kimya Sanayii	Turkey	wet
Beslon	Toho Beslon Co.	Japan	wet
Bulana	Kombinat "Dimitar Domov"	Bulgaria	wet
Cashmilon	Asahi Chemical Industries	Japan, Argentina, Ireland, India	wet
Courtelle	Courtaulds Fibres	Great Britain, France	wet
Creslan	Amer. Cyanamid	USA	wet
Crilenka	Courtaulds Fibers	Spain	wet
Crysel	Celulosa y Derivados	Mexico	wet
Dolan	Hoechst	FRG	dry (wet)
Dralon	Bayer	FRG, Peru	dry (wet)
Euroacryl	Anicfibre	Italy	wet
Exlan	Japan Exlan Co.	Japan	wet
Fisisa	Fibras Sinteticas	Mexico	wet
Hanilon	Hanil Synthetic Fiber Industries	South Korea	wet
Kanebo Acryl	Kanebo Synthetic Fibres	Japan	wet (dry)
Kanekaron (Modacryl)	Kanegafuchi Chemical Industries Co.	Japan	wet
Leacril	Montefibre Societa Italiana Prodotti Acrylic	Spain Great Britain, Italy	wet
Malon	Organsko Hemijska Industrija Skopje	Former Yugoslavia	wet
Melana	Uzina de Fibre Sintetice Savinesti	Romania	wet
Nitron	Chemiefaserkombinat; Navoi Polotsk, Saratow	former States of USSR	wet
Orlon	Du Pont	USA	dry
Tairylan	Formosa Plastic	Taiwan	wet
Toraylon	Toray	Japan	wet
Townflower	Tong Hwa Synthetic Fiber	Taiwan	wet
Velicren	Snia Fibre	Italy	wet
Vonnel	Mitsubishi Rayon	Japan	wet
Wolpryla	VEB Chemiefaserwerk "Friedrich Engels"	GDR	wet

In terms of comfort, flame retardancy, chemical resistance, and shrinkability, these fibers may be considered specialty items, essentially having the properties of the homopolymer, e.g., chemical composition, low crystallinity, high dipole moment, and low moisture regain. The obstacles to commercial development of these fibers result mainly from the low softening point, which leads to poor dimensional stability at elevated temperature and poor downstream processability compared to other synthetic fibers.

6.1. Fibers from Pure Poly(vinyl chloride) Homopolymer

6.1.1. Fibers from Pure Atactic Homopolymer

The poly(vinyl chloride) generally employed for fiber production is simply a variant of the standard polymer used in plastic compounds. Bulk polymerization at 60 – 70 °C is generally preferred because contamination from polymerization additives is low; this results in an atactic polymer with a molecular mass of 70 000 – 85 000. The melting point of pure homopolymer is close to the decomposition point. Dry or wet spinning from solution is, therefore, the only industrial route available to produce commercial fibers.

Dry Spinning. To avoid thermal decomposition with dehydrochlorination, the spinning temperature must be kept relatively low; therefore, the boiling point of the solvent also must be low.

Only a few common solvents are suitable for the homopolymer. Blends of equal volumes of carbon disulfide and acetone (*bp* 38 °C) and benzene and acetone (*bp* 64.5 °C) were patented and are presently used in production [272]. The spin-

ning composition consists of polymer, solvent, and additives. Polymer content varies with the molecular mass of the polymer and ranges from 25 to 30 wt %. The additives have definite functions, e.g., improving heat stability, lightfastness, whiteness, and luster [273]. Certain producers of homopolymer chlorofiber overcome the difficulties in conventional dyeing caused by the low shrinkage temperature of the fiber by adding dye to the spinning dope.

Early swelling of the complex polymer grain and high dope viscosity make dope preparation difficult. The three possible ways to overcome this problem are (1) flash mixing of polymer and solvent with a short residence time in the shearing system [274], (2) premixing the polymer with a nonsolvent [275], and (3) kneading under high shearing force with long residence time.

When the solution is processed at ambient temperature, the microcrystalline regions of the polymer are not dissolved and the solution is colloidal. Dissolution at a temperature >100 °C and processing the solution above the boiling point of the solvent result in a more homogeneous solution and a large decrease in viscosity, with concomitant advantages in further processing [276]. After the dope is dissolved, it is filtered in one or two steps.

The solution is heated or cooled to the spinning temperature and extruded through a 1000- to 2000-hole spinneret in a long tube with hot air moving concurrently with the filaments. The emerging filaments are cooled, wetted by an aqueous lubricant, and wound up or collected at 300 – 700 m/min. The spinning parameters are important because they determine characteristics of the as-spun filament, such as cross-sectional shape and specific surface.

Staple Route. After spinning, the tow is preheated at a temperature (80 °C) lower than the glass transition temperature and drawn 400 – 800 % at a temperature (85 °C) higher than the glass transition temperature. The drawing medium may be hot water or low-pressure steam.

After drawing, the fiber is thermally unstable and has a strong tendency to return to its original length; the shrinkage and shrinkage stress may be as great as 70 % and 0.6 cN/dtex, respectively. For this reason, the tow is annealed under tension in water or steam [277]. The residual shrinkage is removed by relaxing the fiber tension-free at the boil or slightly above.

The next steps in the production of standard fiber are oiling, drying, crimping, and cutting, by conventional technology. Some commercial interest has been shown in high-shrinkage (30 – 65 %) chlorofibers with their attendant high shrinkage stress. Shrinkable fibers are obtained by varying the draw ratio and avoiding the annealing and relaxation steps.

Wet Spinning. At present, one industrial plant (Kustanai, former States of USSR) uses the wet process to produce commercial chlorofiber. Suitable common solvents are *N,N*-dimethylacetamide (DMA) and *N,N*-dimethylformamide (DMF). Special additives are required to limit discoloration of the solution. With these solvents, the rate of coagulation is high and voids appear in the coagulated gel, permanently impairing fiber performance. To exert positive control over the coagulation rate, water is added to the solvent before the solution is prepared; the solvent is then heated to improve solubility [278]. High solvent content (up to 85 %) and low temperature (5 – 15 °C) are suitable for the coagulating bath.

After the gel has coagulated, it is drawn at a low draw ratio and then washed on a multiple-stage countercurrent machine to recover the solvent as a concentrated aqueous solution. A second draw of 400 – 600 % is carried out at boiling or intermediate temperature; the tow is then dried under tension. Annealing can be carried out during or after drying. The tow is then relaxed tension-free, oiled, cut into staple, and baled.

In wet processing, obtaining a homogeneous structure and avoiding discoloration and residual solvent are inherently difficult. The properties of the fibers are generally inferior to those produced by the dry process.

6.1.2. Fibers from Blends of Poly(vinyl chloride) Homopolymer

To avoid the inherent limitations of homopolymer fibers without loss of specific and outstanding properties, spinning blends of homopolymer and minor amounts of a polymer with fibrillar

structure to act as reinforcement has been proposed. Postchlorinated PVC (CPVC) or cellulose acetate [279] can be used without modification of the process. Other blends require special processes. Dry-spun fibers, which contain ca. 80 – 85 % PVC homopolymer and 15 – 20 % CPVC, exhibit the structure of a matrix reinforced with fibrils and have greatly improved heat dimensional stability, mechanical properties, and creep [280]. Such fibers are produced on a commercial scale by Rhône-Poulenc. If cellulose acetate or cellulose is used to replace CPVC, the level of heat stability increases [279], [281]. Researchers in the former Soviet Union proposed wet-spinning a blend of homopolymer with 4-phenylbenzimidazoterephthalamide in dimethylacetamide [282]. In all these cases, the PVC homopolymer is the matrix.

A Japanese firm (Kohjin Co.) produces a fiber (Cordelan) in which PVC and PVC – PVA [poly(vinyl alcohol)] graft-polymer particles are dispersed in a PVA matrix. This fiber gains heat dimensional stability, with concomitant reduction in chemical resistance, thermal insulation, and flame retardancy.

6.1.3. Products, Properties, and Uses

Products. Three major grades of products are available: (1) shrinkable products with different levels of shrinkage, (2) stabilized products for tow dyeing for apparel, and (3) highly stabilized products for piece dyeing and home furnishings (matrix-fibrils). Table 19 lists the physical characteristics of these grades. The stabilized grades are available in a wide range of fineness from 2.4 to 15 dtex and of cut lengths required for cotton, woolen, or worsted spinning systems.

Chlorofiber can be spun on cotton, woolen, or worsted systems as 100 % highly stabilized fiber or as blends of shrinkable and stabilized fibers. The fibers may be dyed at the boiling temperature or slightly below by use of selected disperse dyes. Only highly stabilized fibers can be yarn- or piece-dyed.

Properties and Related End Uses. The main end uses for chlorofibers are related to comfort, flame retardancy, chemical resistance, and surface processing.

Comfort. Chlorofibers exhibit high thermal and tactile comfort [283], [284]:

- high thermal resistance (0.23 – 0.25 m^2 °C $W^{-1}cm^{-1}$)
- constant thermal resistance with humidity transfer
- water transfer
- high rate of evaporation of perspiration
- feeling of warmth
- no lag time between sweat production and evaporative heat loss

Flame Retardancy. Chlorofibers have a very high level of flame retardancy as shown by the limiting oxygen index (LOI) of 46 % at 50 °C and 49 % at 150 °C. They do not ignite and spread flame, nor do they melt and form droplets in contrast to many other synthetic fibers. They may be blended with other flammable, staple fibers, such as wool and polyester, and retain their flame retardancy over a specific range of blend ratios. They can also be blended with thermally resistant fibers, such as aramid and preoxidized acrylic fiber, and exhibit outstanding heat resistance in terms of ignition temperature (640 – 680 °C) and ignition flux (6.4 W/cm^2).

Chemical Resistance. Homopolymer fibers withstand weathering and are resistant to most chemicals with the exception of polar organic compounds (chlorinated hydrocarbons, aromatic hydrocarbons, ketones, and a few others).

Surface processing. Woven or nonwoven fabrics of chlorofibers may be processed by thermal molding. Under the combined action of temperature and pressure, any shape can be achieved with a woven or knitted fabric composed of chlorofibers. These fibers may be heated and welded by high-frequency heating because of their low dielectric loss factor (0.013 at 50 Hz).

Trade Name, Producer (Plant Location): *Retractyl-Fivravyl, Thermovyl L9,* and *ZC,* Rhovyl (Tronville-en-barrois, France); *Teviron,* Teijin (Iwakuni, Japan); (Kustanai, former States of USSR); (Canton, Chunking, Foochow, Shanghai, Sengli, and Tientsin, all People's Republic of China).

Table 19. Physical properties of stabilized grades of PVC products

Property	Shrinkable fibers	Stabilized fibers	Highly stabilized fibers
Density, g/cm^3	1.4	1.38	1.38
Break strength, cN/dtex	2.5 – 3.0	1.6 – 2.2	2.1 – 2.5
Break elongation, %	15 – 20	85 – 105	45 – 55
Initial modulus, cN/dtex	30 – 40	10 – 11	12.5 – 13.5
Shrinkage in air, %			
at 100 °C	40 – 55	0.2	0
at 110 °C	45 – 65	15 – 25	5 – 10
at 115 °C	50 – 70	25 – 30	10 – 15
Temperature of maximum shrinkage stress, °C	80 – 100	135 – 140	135 – 140
Shrinkage stress, cN/dtex	0.04 – 0.06	0.02 – 0.03	0.015 – 0.02
Moisture regain (at 65 % R.H., 25 °C), %	0.05 – 0.1	0.05 – 0.1	0.05 – 0.1

6.2. Fibers from Postchlorinated Poly(vinyl chloride) (CPVC)

Poly(vinyl chloride) homopolymer (56.7 % chlorine) can be postchlorinated to a 73.2 % maximum chlorine content by free-radical chlorination. Solubility in common solvents, especially in acetone or blends containing acetone, depends on the chlorine content of different molecular chains and the distribution of blocks of poly(vinyl chloride) and 1,2-dichloroethane in each chain.

The *solution process* is believed to be the most suitable method of obtaining concentrated spinning solutions, with maximum solubility at a chlorine content of 63 %. Soviet (Chlorin) and German Democratic Republic (Piviacid) fiber producers claim a solution process using either tetrachloroethane or dichlorobenzene as solvent. The processing steps include solution preparation, chlorination, precipitation of polymer with nonsolvent, polymer washing, and solvent recovery [278].

The *quality specifications* of spinnable polymer are:

Chlorine content	63 – 65 wt %
Specific viscosity at 0.2 % in cyclohexane	0.2
Degree of polymerization	900 – 1000
Ash content	0.03 – 0.08 wt %
Iron content	1×10^{-3} wt %

Filament or staple fiber based on CPVC is spun by the dry or wet process in acetone. In wet spinning, the coagulation bath consists of water and a small amount of acetone; the spun gel is drawn 300 – 700 % in one or two steps.

The properties of the filament are as follows:

Linear mass	3.6 – 4.2 – 5.6 dtex
Break strength	>1.6 cN/dtex
Break elongation	>40 %
Density	1.44 g/cm^3
Shrinkage in boiling water	72 %
Limiting oxygen index	51.5 %

Fibers of CPVC have properties similar to homofibers, with slightly higher shrinkage and flame retardancy and lower chemical resistance and ease of dehydrochlorination [285].

Some attempts have been made to improve certain limiting characteristics: (1) grafting methyl methacrylate to increase moisture regain, (2) grafting acrylonitrile to improve thermal stability, and (3) blending CPVC with cellulose acetate at a ratio of 85 : 15 to increase the shrinkage temperature [278].

6.3. Fibers from Vinyl Chloride Copolymers

Copolymers of PVC with a 13 – 15 % vinyl acetate content and a vicosity index of 57 – 58 exhibit a lower glass transition temperature (73 °C) than homopolymers, and the bondability improves at temperatures lower than the melting point.

The process used is *dry spinning*. A solution of copolymer (≤50 wt %) in acetone, containing heat stabilizers and additives, is prepared at ambient temperature, filtered, and extruded at a temperature dependent on the polymer content (50 – 80 °C). A hot aqueous treatment without

tension, crimping, or oiling, and low-temperature drying follow if the fiber is to be used in the dry-laid nonwoven process; for the wet-laid nonwoven process, crimping is omitted.

The relevant key *properties* are:

Chlorine content	46 %
Break strength	0.6 – 0.7 cN/dtex
Break elongation	150 – 200 %
Softening point	69 °C
mp	155 – 160 °C
Decomposition temperature	230 °C
Shrinkage in water	18 – 28 % at 80 °C
Shrinkage tension	0.003 cN/dtex
Shrinkage temperature	60 °C

Bonding may be carried out between 80 and 150 °C or between 150 and 170 °C with transformation of the fiber into melt droplets. Thermal bonding is used to produce nonwovens, blended with other natural, regenerated, or synthetic fibers at a ca. 30 % maximum content.

Trade Name, Producer (Country): *Wacker MP-Faser,* Wacker Chemie (France); *Vinyon HH,* Celanese Co. (United States).

6.4. Fibers from Poly(vinylidene chloride) and Copolymers [278], [286], [287]

The homopolymer of PVDC has a softening point of 185 – 200 °C. Its two outstanding properties, high crystallinity and insolubility, make it difficult to process. Therefore, to ease processing conditions, copolymers with small quantities of vinyl chloride are used.

The usual copolymers contain 85 – 90 wt % vinylidene chloride (saran). Copolymers are processed by melt spinning or extrusion. Because they are heat sensitive and have a limited thermal life at the extrusion temperature, specially designed extruders and alloys are required. Depending on end use, plasticizers, heat and light stabilizers, and colored pigments may be added to the copolymer. Monofilaments or multifilaments are extruded at 160 – 175 °C, quenched in water at 10 – 15 °C with a short distance between the spinneret and the level of the quenching bath, and then drawn 500 – 600 %.

Products, Properties, and Uses. The range of *products* includes monofilaments (70 – 120 dtex) and multifilaments (10 – 78 dtex). Typical properties are:

Density	1.65 – 1.75 g/cm^3
Break strength	1.5 – 2.0 cN/dtex
Break elongation	15 – 25 %
Initial modulus	6 – 10 cN/dtex
Moisture regain	>0.1 %
Softening point	115 – 135 °C
mp	170 – 177 °C
Shrinkage temperature	70 °C
Dielectric constant at 60×10^6 Hz	3 – 5
Dielectric loss	tan δ 0.05

Poly(vinylidene) chloride has excellent resistance to acid, alkali (except ammonium hydroxide), and weathering but poor heat resistance. The products cannot be dyed easily; disperse dyes may be used but fastness is poor.

7. Poly(vinyl alcohol) Fibers

W. O. Hermann and W. Haenel applied for the first patent on poly(vinyl alcohol) [*9002-89-5*] (PVA) fiber in 1931. They reported that the fiber could be made from PVA by the well-known wet- and dry-spinning methods, and that the water resistance of the fiber could be improved by physical and chemical aftertreatment. However, they did not investigate further the production of a water-resistant textile fiber because their interest was limited mainly to the solubility of the fiber in water. Water-soluble PVA monofilament, made by dry spinning and called "Synthofil," was manufactured for a short time in Germany.

In 1940, I. Sakurada and his colleagues improved the hot-water resistance of wet-spun PVA fiber by heat-treating it in hot air and acetalizing it with formaldehyde; they developed the technology for production of a water-resistant textile fiber [288].

Wet-spun staple fiber based on this technology was first available commercially in Japan in 1950; the fiber was called "Vinylon." Initially, Vinylon was distinguished from other synthetic fibers by its hydrophilic properties and it was marketed as a substitute for cotton in both clothing and industrial fibers. Improved production technology led to remarkable improvement in the mechanical properties of the fiber. In 1959, high-tenacity filament produced by dry spinning, and later by wet spinning, became available. The use of Vinylon has become almost en-

tirely industrial. In addition to the water-resistant fiber, a water-sensitive variety has been produced for water-soluble fiber and binder fiber for making paper.

The staple fiber has been also produced commercially in the Democratic People's Republic of Korea since 1961 and the People's Republic of China since 1965.

7.1. Physical and Chemical Properties

The mechanical properties of commercial PVA fibers (Vinylon) produced in Japan are shown in Table 20 [289] (see → Fibers, 1. Survey, Chap. 5.). The tenacity and modulus are the highest among commercial general-purpose synthetic fibers. Fiber having a tenacity of 12 cN/dtex and a modulus of 274 cN/dtex has been produced for cement reinforcement. Resistance to abrasion and fatigue is also excellent. The fiber absorbs slightly more water than polyamide fiber; and its mechanical properties in the wet state are somewhat lower. Formalization is effective in improving water resistance of the fiber but may lower the mechanical properties. The elastic recovery is slightly lower than that of acrylic fiber but better than cotton or rayon. Acetalization with benzaldehyde is an effective technique for improving the elastic recovery of PVA fibers used in clothing. The resistance of PVA fiber to various chemicals is generally superior to that of polyamide and polyester fibers. In particular, its excellent alkali resistance is important in the reinforcement of cement products [290]. Because of its high polarity, PVA fiber adheres well to matrix materials in fiber-reinforced composites for resin, rubber, and cement. Its weather resistance is the highest of the various synthetic fibers.

The fiber is softened by heating above 230–240 °C, but it has no definite melting point. Commercial fibers with sufficient hot-water resistance, and water-soluble fibers that dissolve in 20–90 °C water, are produced by applying the appropriate heat or acetalization treatments. For use in wearing apparel, the formalized staple fiber, mainly blended with cotton, is dyed with direct, vat, sulfur, or metallized dyes.

7.2. Production

7.3. Poly(vinyl alcohol) for Fiber

Poly(vinyl alcohol) for fiber is produced by polymerization of vinyl acetate in methanol and subsequent alkaline hydrolysis of poly(vinyl acetate). The material must have a degree of polymerization of 1200–2500 and a narrow molecular mass distribution. These properties are attained by adjusting polymerization conditions, and especially by conducting the polymerization at a lower temperature and keeping the conversion of vinyl acetate in methanol below 70 %.

For the production of industrial fiber, PVA with a degree of polymerization of ca. 1700 is generally used. The amount of residual acetyl groups in PVA should be <0.1 mol % for heat treatment to produce a sufficiently water-resistant fiber.

7.3.1. Wet Spinning

Sodium Sulfate Coagulation Bath Method. Staple fiber and tow are produced by wet spinning.

In ordinary wet spinning, PVA is dissolved in water to a concentration of 14–16 % by heating. After filtration and deaeration, the spinning solution is kept above 70 °C to avoid gelation. It is extruded through spinnerets into a coagulation bath containing a saturated solution of sodium sulfate at 40–50 °C. The coagulation speed is much slower than that of viscose, and a longer residence time in the bath is necessary. For the coagulation operation to run smoothly, and to minimize floor space requirements, a vertical spinning machine is preferred [291]. The spun fiber tow is usually drawn while passing through a second hot coagulation bath and then dried.

The wet-spun and dried fiber is soluble in hot water. It is further drawn and heat-treated in hot air or by hot rollers at 210–240 °C. The temperature of the heat-treatment process must be higher than that of the hot drawing process. In heat treatment, a small amount of shrinkage is usually allowed in order to stabilize fiber structure. This treatment orients and crystallizes the polymer chains in the fiber, reducing the fiber solubility in hot water.

Table 20. Mechanical properties of Vinylon fibers [289]

Property	Staple and tow		Filament	
	Regular tenacity	High tenacity	Regular tenacity	High tenacity
Tenacity, cN/dtex				
Standard	3.5 – 5.7	6.0 – 9.3	2.6 – 3.5	5.3 – 10.6
Wet	2.8 – 4.6	4.7 – 7.9	1.9 – 2.8	4.4 – 9.3
Wet – standard ratio, %	72 – 85	78 – 85	70 – 80	75 – 90
Loop tenacity, cN/dtex	2.8 – 4.6	4.4 – 5.1	4.0 – 5.3	6.2 – 11.5
Knot tenacity, cN/dtex	2.1 – 3.5	4.0 – 4.6	1.9 – 2.6	2.4 – 4.4
Elongation, %				
Standard	12 – 26	9 – 17	17 – 22	6 – 22
Wet	12 – 26	9 – 17	17 – 25	8 – 26
Young's modulus, cN/dtex	22 – 62	62 – 221	53 – 79	62 – 221
Resilience at 3 % elongation, %	70 – 85	72 – 85	70 – 90	70 – 90
Relative density	1.26 – 1.30	1.26 – 1.30	1.26 – 1.30	1.26 – 1.30
Commercial regain, %	5.0	5.0	5.0	5.0
Water absorbency, %				
20 °C, 20 % R.H.	1.2 – 1.8	1.0 – 1.5	1.2 – 1.8	1.0 – 1.5
20 °C, 65 % R.H.	4.5 – 5.0	3.5 – 4.5	3.5 – 4.5	2.5 – 4.5
20 °C, 95 % R.H.	10.0 – 12.0	8.0 – 10.0	10.0 – 12.0	8.0 – 10.0

The heat-treated fiber is further acetalized in an aqueous solution containing 1 – 5 % formaldehyde, 10 –20 % sulfuric acid, and 5 – 20 % sodium sulfate at 50 – 80 °C for 10 – 60 min to reach a degree of formalization of 25 – 35 mol %. Acetalization is affected not only by reaction conditions but also by conditions of the drawing and heat-treatment processes applied to the fiber. Acetalization greatly improves the boiling-water resistance of the fiber and also changes its mechanical and chemical properties to some extent.

In the production of staple fiber, the heat-treated tow is usually crimped mechanically and cut, the cut fiber is formalized. For industrial application, the formalized tow is subjected directly to a stretch-break spinning system (Perlock or Converter system) to convert it to spun yarn. This yields spun yarn with high tenacity and high toughness.

Short-cut fiber, destined to be water-swellable binder fiber or water-resistant substrate fiber for making paper, is produced by cutting wet-spun and dried tow, or tow that has been heat-treated and sometimes formalized, to the proper length of 3 – 12 mm.

Alkaline Coagulation Bath Method. Fiber that has been wet-spun in a sodium sulfate coagulation bath has a cocoon-shaped cross section and skin – core of inhomogeneous structure because coagulation proceeds with rapid dehydration from the outside of the extruded jet of spinning dope. This makes it difficult to obtain a clear fiber with high tensile strength.

When PVA dope is spun into an aqueous solution of concentrated alkali, fiber with a round cross section and a homogeneous, compact structure can be obtained [292]. The fiber is transparent and can be highly drawn to give improved tensile strength and modulus, and better water resistance. In this case, the spinning dope extruded in the coagulation bath is gelated instantaneously by contact with the alkaline coagulant, so dehydration proceeds homogeneously. This wet-spinning method is used in Japan. In commercial practice, the aqueous coagulation bath contains sodium hydroxide at a concentration >200 g/L and at a temperature of 40 –50 °C.

The alkaline coagulation method was modified to give more highly drawn fiber with a higher modulus by using aqueous acidic spinning dope containing a small amount of boric acid [293]. Because of the presence of boron, cross-linking appears to occur between the polymer chains of PVA during fiber formation. This suppresses entangling of the polymer chains and makes coagulation more homogeneous.

The spinning dope, containing boric acid of 0.5 – 2.0 wt % (relative to PVA) at pH 3 – 5.5, is spun into an aqueous alkaline coagulation bath containing 20 – 100 g/L of sodium hydroxide and 100 – 300 g/L of sodium sulfate at pH 13.5 – 14.0. After removal from the coagula-

tion bath, the spun fibers are treated with acid, washed with water, dried, and drawn to a draw ratio >10 to give high tenacity (>9 cN/dtex) and high modulus (>177 cN/dtex).

This technology was developed initially for production of high-tenacity, high-modulus filament yarn for use in belt cord for radial tires. The yarn is modified further to produce high-modulus fiber, which can be substituted for asbestos in cement reinforcement; this fiber has a tenacity of 12 cN/dtex and a modulus of 274 cN/dtex [294].

Wet Spinning with Organic Solvent. Polyvinyl alcohol has a high crystalline modulus comparable to that of polyethylene. This high modulus has not been exploited extensively to date in the preparation of fiber. Various investigations have been carried out into the preparation of PVA fiber with a much higher modulus. In many cases, this has involved modified wet-spinning methods that use organic solvents.

When a hot, dilute solution of PVA in polyol is cooled slowly, highly crystallized PVA is deposited [295]. This phenomenon has been connected with gel-spinning – ultrahigh-drawing procedures. When the gel fiber from a solution of PVA in glycerin, having an ultra-high degree of polymerization of 38 000, was hot-drawn at high temperature, the resulting fiber had a tenacity of 17 cN/dtex and a modulus of 555 cN/dtex [296].

7.3.2. Dry Spinning

Dry spinning of PVA is used in the production of filament yarns. The two classifications of dry spinning are low-draft spinning [297] and high-draft spinning [298].

Low-Draft Spinning. Aqueous spinning dope, which has a concentration of 41 – 45 % and an extremely high viscosity, is used in low-draft spinning. It is prepared by dissolving PVA granules, conditioned with water, by heating them in an extruder under pressure.

The high-vicosity spinning dope at a temperature of 130 – 160 °C, is spun in air at ca. 50 °C. The cooled and solidified filament is wound in a low-draft ratio of 0.3 – 1.0 based on the extrusion rate of the spinning dope from the spinneret (i.e., draft ratio = winding rate/extrusion rate), and gradually dried while the drying temperature is raised. The dried filaments have a circular cross section, a homogeneous structure, and low strain; they can be drawn to the highest draw ratio and heat-treated near the melting point. The resulting filament has adequate hot-water resistance and is used commercially without formalization.

This method is suitable for producing heavier filament for industrial applications. It was originally developed for the production of heavier monofilaments of 110 – 1100 dtex. It is now also applied to the production of multifilaments, such as 1330 dtex/200 filaments.

High-Draft Spinning. In high-draft spinning, aqueous spinning dope containing PVA at a concentration of 28 – 41 % is spun at 90 – 95 °C into a spinning tube consisting of a drafting zone at 30 – 80 °C and 55 – 95 % R.H., and a drying zone at >80 °C and <50 % R.H. While passing through the spinning tube, the filament is highly drafted and dried so as to be taken off at a winding speed of 200 – 500 m/min. The dried filaments are hot-drawn and heat-treated, and sometimes further acetalized.

This method can give finer filaments (<5 dtex) for textile applications. Some filament for apparel, made from a mixture of PVA and aminoacetalized PVA is produced commercially and can be dyed with acid dyes. This method is now used primarily to produce water-soluble filament. Filament soluble in water at 20 – 90 °C is produced commercially by using PVA with low degrees of polymerization and saponification, and by adjusting heat-treatment conditions [299].

7.3.3. Mixed Spinning

The technology for spinning bicomponent fiber from PVA and poly(vinyl chloride) (PVC) in an emulsion was developed in 1961 [300].

Vinyl chloride is emulsion-polymerized in an aqueous solution of PVA, and the emulsion is mixed with PVA to give a spinning dope containing PVC and PVA in ca. 1 : 1 mass ratio. The fiber is produced in essentially the same way as regular PVA fiber, by using the sodium sulfate

coagulation bath method. The fiber is characterized by its flame retardance and is produced commercially in Japan under the name Polychlal.

7.4. Uses

In Japan, about 90 % of Vinylon fiber production, ca. 30 000 t/a, is consumed in industrial applications that require such advantages as high tenacity, excellent weather resistance, and effective reinforcing effect in matrices. The main industrial uses are fishing nets, ropes, belts, canvasses, and sewing threads. Perlock-type spun yarns from 100 % Vinylon tows and filament yarns are used predominantly. The water-soluble filaments are used to produce chemical lace fabrics.

The development of special grades has increased the demand for Vinylon remarkably. The water-swellable, short-cut fibers are popular as binders in the production of paper and wet-laid nonwoven fabrics. Another short-cut fiber, modified to have the highest tenacity and modulus as well as a special surface structure that enhances its adhesion to matrices, is used as a substitute for asbestos to reinforce cement products. These special grades are exported to the world market.

Only 10 % of Vinylon production is used domestically in Japan for apparel, mainly by blending Vinylon with cotton and other fibers. The main apparel application is work clothing that requires durability and chemical resistance.

In Korea and China, significant quantities of PVA fiber are blended with cotton and used for apparel. In these countries too, fiber applications in the future will shift gradually to industrial uses.

7.5. Economic Aspects

Poly(vinyl alcohol) fibers are produced in Japan, the Democratic People's Republic of Korea, and the People's Republic of China. In Japan, production levels reached a maximum of 76 000 t in 1971 and decreased to 36 000 t by 1984, while production capacity remained constant at 80 000 t/a [289]. Korea and China are presumed to have production capacities of more than 50 000 and 150 000 t/a, respectively.

The producers of Vinylon and Polychlal fibers in Japan, and their brand names, are as follows:

Vinylon:	Kuraray Co. (Kuraray Vinylon, Cremona,Kuralon), Unitika Chemical Co. (Unitika Vinylon, Mewlon), Nitivy Co. (Vilon, Solvron; water-soluble)
Polychlal:	Kohjin Co. (Cordelan)

8. Polytetrafluoroethylene Fibers

Polymerization technology for polytetrafluoroethylene [*9002-84-0*] (PTFE) has been known since 1941 and has become well established, leading to utilization of the polymer not only as a plastic but also in fiber form [301]. This is undoubtedly because it combines interesting chemical (inertness) and physical (low friction, temperature resistance, and triboelectric characteristics) properties.

Tetrafluoroethylene is polymerized under pressure in the presence of persulfate or peroxide in water; oxygen is excluded. The reaction is exothermic and increases in velocity with increased pressure. The viscosity of the polymer, as well as its molecular mass cannot be measured readily; however, indications are that the polymer chain is linear. At the normal polymerization level, molecular masses are estimated to be between a few hundred thousand and several million.

8.1. Physical and Chemical Properties

Some of the physical properties of polytetrafluoroethylene are as follows:

mp	327 °C
Glass transition temperature	ca. 120 °C
ϱ	2.1 g/cm^3
Melt viscosity (380 °C)	10 GPa

These fibers are similar to the polymer (i.e., highly crystalline 93 – 98 %) and, because of the high energy of the C–F bond, display appreciable resistance to heat and chemicals (see → Fibers, 1. Survey, Chap. 5.; → Fluoropolymers, Organic). As a result, the

polymer is difficult to burn. The fibers also are distinctive because of their good light and weathering durability, and their utility at temperatures up to 290 °C. Phase changes that occur below 30 °C are responsible for the cold flow characteristics of the material, although the fibers are reported to be more dimensionally stable than bulk polymer because of their molecular orientation. Decomposition products are generated above 290 °C at a rate of 0.0002 %/h. Copolymers, containing minor amounts of other fluorinated monomers, are also available. These are melt extrudable but, because of their high melt viscosities, only into monofils of ≥100 dtex. A 50 – 50 copolymer with ethylene (Tefzel, Du Pont) has a lower melt viscosity and can be extruded at commercially attractive speeds; however, its temperature–chemical characteristics are altered by the presence of the nonfluorinated comonomer.

8.2. Production

Polytetrafluoroethylene can be sintered above its melting point into a tough, viscous melt. However, melt spinning is not feasible with the homopolymer because the temperature required to overcome the viscosity leads to polymer degradation. Because economically attractive solvents for the polymer have not been found (a few fluorinated hydrocarbons dissolve several percent of polymer above 300 °C), either paste extrusion below the melting point of the polymer or dispersion spinning to form fibers is used. To promote agglomeration during extrusion of plasticized (with 20 wt % kerosene) homopolymer, extrusion is done under conditions favorable to production of particulates in ribbon form, so that at least 30 % have a width of 0.1 nm, an aspect ratio of at least 5 : 1, and a cross section of $<0.001\ nm^2$. Sintering, followed by drawing and heat-setting results in solid or expanded films (depending on process conditions), which are then split into filaments of ca. ≥100 dtex.

An alternate process is used to produce finer filaments. A 60 % solid dispersion of polytetrafluoroethylene polymer, stabilized with aryl or alkyl polyglycol ethers, is blended with viscose solution (5 %). The latter acts as a matrix during early phases of the process. The mixture is filtered and then extruded at room temperature through a multihole spinneret (hole diameter 0.15 mm) into a coagulating bath containing 6 % sulfuric acid, 16 % sodium sulfate, and 0.3 % zinc sulfate. The spinning velocity is about 12 m/min.

After coagulation, the yarn is washed with water at 90 °C, dried at 190 °C, and then raised to 340 – 360 °C, the temperature at which polymer particles are sintered and the viscose residue degrades and is partially burned. After a brief contact time at 360 – 390 °C, the yarn is stretched four to eight times its original length.

The resultant continuous-filament yarn has a tenacity of 1.8 cN/dtex and an elongation of 20 – 30 %. The filament mass is 6 dtex. The fibers are brown and can be bleached in boiling sulfuric acid containing nitric acid or by exposure for several days in an air oven. The latter must be done cautiously by raising the temperature stepwise to 350 °C, to avoid exotherms because of the sudden oxidation of the viscose residues.

8.3. Uses

Polytetrafluoroethylene fibers are utilized in woven and nonwoven forms as filter media for aggressive liquids and gases (i.e., strong acids, bases, and oxidizing agents), including air pollution control fabrics. Fibers are also used in gaskets and pump seals because of their chemical resistance and low-friction characteristics. In electrical systems, they are utilized as insulators for wires and components, and as cleaners – wipers in xerographic devices, where their high-temperature and triboelectric characteristics are important. Because of their great chemical stability, these fibers are used as the binder component in asbestos chlor-alkali cell diaphragms. The low-friction characteristics and high length – diameter ratio of the fibers are important in their use as solid lubricants in engineering plastics. Monofils have found application in demisters in the production of mineral acids.

In addition, dispersion of fibers on woven or nonwoven substrates imparts chemical stability to the surface and results in layered composites with excellent dielectric properties and corrosion resistance for utilization in xerography.

Trade Names. Teflon [*98846-34-5*], Teflon FEP [*58516-09-9*], Teflon TFE, Teflon PFA

[*71767-39-0*], Tefzel [*25038-71-5*], Tefaire (Du Pont, United States); Goretex (Gore, United States); Polyflon (Daikin Kogyo, Japan); Toyoflon (Toyo Rayon, Japan); Fluon (ICI, Great Britain).

8.4. Toxicology and Occupational Health

Adequate ventilation should be provided during processing of PTFE fibers. At elevated processing temperatures, PTFE fluoropolymer fibers liberate vapors that may be harmful. Generally, adequate ventilation and care to prevent such vapors from being inhaled will provide ample safety. Care should be exercised not to contaminate smoking tobacco with PTFE fibers. A mixture of tobacco and PTFE polymer would be especially hazardous if the mixture were combusted and inhaled in smoking. This can be avoided by not carrying tobacco into areas where it may become contaminated.

9. References

General References

1. A. Ziabicki, H. Kawai: *High Speed Spinning-Science and Engineering Aspects,* J. Wiley & Sons, New York 1985.
2. J. F. McKellar, J. S. Allen: *Photochemistry of Man-Made Polymers,* Applied Science Publishers, London 1979.
3. M. Lewin, E. M. Pearce (eds.): *Handbook of Fiber Science and Technology,* **vol. 4,** Marcel Dekker, New York 1985, pp. 73 – 169.
4. *Ullmann,* 4th ed., **11,** p. 336.
5. T. Osugi: "PVA Fibers," in H. Mark et al.: *Man-Made Fibers,* Science & Technology, 3rd ed., Interscience Publ., New York 1968, p. 268.
6. I. Sakurada: *Polyvinyl Alcohol Fibers,* Marcel Dekker, New York 1985.
7. O. L. Shealy, *Fiber Prod. Int.* **12** (1984) no. 2, 32 – 41.
8. P. Schlack, *Chemiefasern* **13** (1963) 560 – 562.
9. Kurashiki Rayon Co., JP 406 628, 1965 (K. Masubayashi, T. Kawaguchi).
10. R. Graf et al., *Angew. Chem.* **74** (1962) 523 – 530.
11. J. S. Robinson (ed.): *Fiber Forming Polymers-Recent Advances,* Noyes Data Corp., Park Ridge, N. J., 1980, pp. 135 – 165.
12. C. F. Horn et al., *Angew. Chem.* **74** (1962) 531 – 540.
13. Monsanto, FR 1 419 132, 1964 (J. E. Tate). Monsanto, US 3 402 152, 1968 (E. T. Erignac, R. T. Wright).
14. D. O'Sullivan, *Chem. Eng. News* **62** 1984, no. 21, 21 – 22.
15. BASF, US 4 060 517, 1977 (F. Mertes et al.).
16. P. W. Morgan, S. W. Kwolek, *Macromolecules* **8** (1975) no. 2, 104 – 111.
17. J. S. Ridgeway, *J. Polym. Sci., Polym. Chem. Ed.* **12** (1974) no. 9, 2005 – 2016.
18. Du Pont, US 3 502 624, 1970 (C. F. Black).
19. F. Happy (ed.): *Applied Fiber Science,* **vol. 2,** Academic Press, New York 1981, pp. 1 – 44.
20. F. Happy (ed.): *Applied Fiber Science,* **vol. 1,** Academic Press, New York 1978, pp. 381 – 416.
21. Du Pont, US 3 184 436, 1965 (E. E. Magat).
22. I. M. Kolthoff, P. J. Elving, F. H. Stross (eds.): *Treatise on Analytical Chemistry,* **vol. 4,** J. Wiley & Sons, New York 1977, part 3, pp. 273 – 313.
23. N. S. Allen, J. F. McKellar, *J. Polym. Sci., Macromol. Rev.* **13** (1978) 241 – 281.
24. W. Schnabel: *Polymer Degradation – Principles and Practical Applications,* Macmillan Publ., New York 1981, pp. 40 – 44.
25. BASF, DE-OS 1 570 261, 1965 (C. Dorfner et al.).
26. Glanzstoff, DE-OS 1 494 765, 1965 (E. Neyer, H. Linhartt).
27. M. Sittig: *Polyamide Fiber Manufacture, Chemical Process Review no. 61,* Noyes Data Corp., Park Ridge, N. J., 1972, pp. 183 – 254.
28. C. Placek: *Multicomponent Fibers,* Noyes Data Corp., Park Ridge, N. J., 1971, pp. 40 – 52.
29. Du Pont, US 3 779 853, 1973 (E. H. Olsen).
30. ICI, US 3 675 408, 1972 (D. M. Pacini et al.).
31. K. Riggert et al., *Chemiefasern Text. Anwendungstech./Text.-Ind.* **23** (1973) 1183 – 1186.
32. BASF, DE-OS 1 495 177, 1964 (P. Maahs et al.).
33. R. A. Vaidya, *Indian Text. J.* **94** (1984) no. 10, 53 – 58.
34. H. Mark, G. S. Whitby (eds.): *Collected Papers of W. H. Carothers on High Polymeric Substances,* Wiley-Interscience, New York 1940.
35. W. H. Carothers, J. W. Hill, *J. Am. Chem. Soc.* **54** (1932) 1579 – 1587.

36. H. Klare: *Geschichte der Chemiefaserforschung,* Akademie-Verlag, Berlin 1985.
37. H. Ludewig: *Polyesterfasern,* 2nd ed., Akademie-Verlag, Berlin 1975.
38. J. R. Whinfield, *Text. Res. J.* **23** (1953) 289 – 293.
39. Calico Printers' Assn., DE 973 553, 1940 – 1941 (J. R. Whinfield, J. T. Dickson).
40. P. Schlack, *Textil-Prax.* **8** (1953) 1055 – 1062.
41. International Rayon and Synthetic Fibres Committee (ed.): *Information on Man-Made Fibres*, vol. 22, Paris 1985.
42. E. Katzschmann, *Chem.-Ing.-Techn.* **38** (1966) 1 – 10.
43. Midcentury Corp., DE-AS 1 081 445, 1960 (S. Barker, A. Saffer).
44. A. Ziabicki: *Fundamentals of Fibre Formation,* Wiley-Interscience, London 1976.
45. I. Hamana, M. Matsui, S. Kato, *Melliand Textilber.* **4** (1969) 382 – 388, 499 – 503.
46. E. Liska, *Kolloid-Z. Z. Polym.* **251** (1973) 1028 – 1037.
47. H. M. Heuvel, R. Huisman, *J. Appl. Polym. Sci.* **22** (1978) 2229 – 2243.
48. A. Ziabicki, H. Kawai (ed.): *High-Speed Fiber Spinning,* Wiley-Intersience, New York 1985.
49. H. H. George, H. Matsui, H. Yasuda, "Literature and Patent Survey," in A. Ziabicki, H. Kawai (ed.): *High-Speed Fiber Spinning,* Wiley-Intersience, New York 1985, pp. 553 – 565.
50. W. Dietrich, G. Reichelt, H. Renkert, *Chemiefasern/Text.-Ind.* **32/84** (1982) 612 – 625.
51. J. Shimizu, N. Okui, T. Kikutani, *Sen'i Gakkaishi* **37** (1981) T135 – T142.
52. M. Jambrich, A Diačiková, O. Ďurčová, A. Chvála, *Acta Polym.* **37** (1986) 118 – 121.
53. G. W. Davis, A. E. Everage, J. R. Talbot, *Fiber Prod.* **12** (1984) no. 2, 22 – 28.
54. I. Marshall, A. B. Thompson, *J. Appl. Chem.* **4** (1954) 145 – 153.
55. K. Scharley, *Bühler Nachr.* **222** Bühler-Miag, Braunschweig 1983.
56. H. H. George, A. Holt, A. Buckley, *Polym. Eng. Sci.* **23** (1983) 95 – 99.
57. I. Jacob, H.-R. Schröder, *Chemiefasern/Text.-Ind.* **30/82** (1980) 114 – 119, 228 – 232.
58. J. H. Dumbleton, *J. Polym. Sci., Part A* **2** (1968) 795 – 800.
59. H. Berg, *Chemiefasern/Text. Ind.* **22/74** (1972) 215 –222.
60. H. Brody, *J. Macromol. Sci., [B] Phys.* **22** (1983) no. 1, 19 – 41.
61. R. W. Miller, J. H. Southern, R. L. Ballman, *Text. Res. J.* **53** (1983) 670 – 677.
62. R. Huisman, H. M. Heuvel, *J. Appl. Polym. Sci.* **22** (1978) 943 – 965.
63. K. Riggert, *Chemiefasern* **19** (1969) 816 – 823.
64. J. H. Dumbleton, J. P. Bell, T. Murayama, *J. Appl. Polym. Sci.* **12** (1968) 2491 – 2508.
65. H. Güllemann, *Melliand Textilber.* **53** (1972) 910 – 914.
66. J. H. Dumbleton, T. Murayama, F. P. Bell, *Kolloid-Z. Z. Polym.* **228** (1968) 54 – 58.
67. V. B. Gupta, C. Ramesh, A. K. Gupta, *J. Appl. Polym. Sci.* **29** (1984) 3727 – 3739.
68. H. Zimmermann, *Melliand Textilber.* **57** (1976) 828 – 832.
69. F. Jakob, *Chemiefasern/Text.-Ind.* **74** (1972) 388 –396.
70. A. Williams, I. T. Ibrahim, *Chem. Rev.* **81** (1981) 589 – 636.
71. S. Hoffrichter et al., *Dtsch. Textiltech.* **20** (1970) 774 – 782.
72. V. Freudenberger, F. Jakob, *Angew. Makromol. Chem.* **105** (1982) 203 – 215.
73. H. Zimmermann, *Chemiefasern/Text.-Ind.* **80** (1978) 1054 – 1060.
74. Teijin, EPS 0 038 429, 1984 (K. Ozaki, M. Matsui).
75. P. Flory: *J. Amer. Chem. Soc.* **58** (1936) 1877 – 1885.
76. S. Fakirov, E. W. Fischer, G. F. Schmidt, *Makromol. Chem.* **176** (1975) 2459 – 2465.
77. H. M. Heuvel, R. Huisman, *J. Appl. Polym. Sci.* **30** (1985) 3069 – 3093.
78. B. von Falkai, W. Giessler, F. Schultze-Gebhardt, G. Spilgies, *Angew. Makromol. Chem.* **108** (1982) 9 – 39.
79. W. H. Charch, W. W. Moseley, *Text. Res. J.* **29** (1959) 525 – 535.
80. W. W. Moseley, *J. Appl. Polym. Sci.* **3** (1960) 266 – 276.
81. H. Sattler, *Kolloid-Z. Z. Polym.* **187** (1963) 12 – 18.
82. E. Liska, *Angew. Makromol. Chem.* **65** (1977) 147 – 168.
83. R. Bonart, *Kolloid-Z. Z. Polym.* **199** (1964) 136 – 144.
84. H. Lückert, W. Stibal, *Chemiefasern/Text.-Ind.* **36/88** (1986) 24 – 29.
85. L. Riehl, *Chemiefasern/Text.-Ind.* **28/80** (1978) 1039 – 1046; **29/81** (1979) 24 – 30.
86. N. H. Bezaz, *Chemiefasern/Text.-Ind.* **32/84** (1982) 262 – 265.

87. H. Lückert, M. Busch, *Chemiefasern/Text.-Ind.* **33/85** (1983) 29 – 38.
88. P. Dammann, *Chemiefasern/Text.-Ind.* **26/78** (1976) 521 – 522.
89. H. Lückert, *Chemiefasern/Text.-Ind.* **29/81** (1979) 1019 – 1022.
90. K. Riggert, *Chemiefasern/Text.-Ind.* **31/83** (1981) 638 – 648.
91. F. Schmitt, *Melliand Textilber.* **63** (1982) 830 – 835.
92. B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981.
93. J. Thimm, *Chemiefasern/Text.-Ind.* **75** (1973) 811 –817.
94. G. W. Davis, A. E. Everage, J. R. Talbot, *Fiber Producer* **12** (1984) no. 6, 45 – 62.
95. H. Treptow, *Chemiefasern/Text.-Ind.* **35/87** (1985) 411 – 412.
96. Rhône-Poulenc-Textile, FR 1 551 878, 1967 (G. Barbe, P. Curtillat, C. Lequay).
97. Du Pont, DE-AS 2 241 718, 1971 (D. G. Petrille, , M. J. Piazza, C. E. Reese).
98. *Fiber World* 1985, 22 – 51.
99. J. Lünenschloss, *Textiltechnik* **32** (1982) 232 – 235.
100. Du Pont, GB 900 009, 1960 (L. Cenzato).
101. Teijin Ltd., EP 159 875, 1984 (H. Yamada, N. Sayama, M. Kuno).
102. Monsanto, US 4 092 299, 1976 (D. L. G. R. MacLean, R. T. C. Estes).
103. F. Maag, *Chemiefasern/Text.-Ind.* **34/86** (1984) 173 – 178.
104. R. C. Mears, *Chemiefasern/Text.-Ind.* **35/87** (1985) 413 – 414.
105. Imperial Chem. Ind., DE 2 836 514, 1977 (P. L. I. Carr).
106. Imperial Chem. Ind., EP 42 664, 1980 (P. L. I. Carr).
107. Du Pont, US 4 134 882, 1977 (H. R. E. Frankfort, B. W. Knox).
108. K. Riggert, *Chemiefasern* **21** (1971) 379 – 384.
109. Hoechst AG, DE 2 115 312, 1971 (R. Johne, M. Bechter).
110. DuPont, US 3 216 187, 1962 (W. A. Chantry, A. F. Molini).
111. Hoechst AG, DE-OS 3 431 831, 1984 (H. Thaler).
112. L. Hartmann, *Chemiefasern/Text-Ind.* **22/74** (1972) 231 – 236, 324 – 328.
113. L. Gerking, *Verfahrenstechnik (Mainz)* **10** (1976) 779 – 784.
114. F. Hensen, S. Braun, *Chemiefasern/Text.-Ind.* **29/81** (1979) 844 – 848.
115. H. Reinbold, *Chemiefasern/Text.-Ind.* **35/87** (1985) 420 – 422.
116. H. Sattler, *Textil-Prax.* **33** (1978) 1175 – 1178.
117. E. V. Martin, H. Busch, *Angew. Chem.* **74** (1962) 624 – 628.
118. B. von Falkai, G. Spilgies, F. Schultze-Gebhardt, H. Wilsing, *Acta Polym.* **34** (1983) 86 – 98.
119. H. Sattler: "Chemiefasern," in Winnacker-Küchler (ed.): *Chemische Technologie,* 4th ed., **vol. 6,** Hauser Verlag, München 1982, pp. 611 – 734.
120. *Text. Organon* **41** (1970) 98; **45** (1974) 71; **51** (1980) 75; **57** (1986) 114.
121. K.-H. Umbach, *Chemiefasern/Text.-Ind.* **85** (1983) 136 – 139, 204 – 208.
122. Bayer, DE 728 981, 1937 (O. Bayer, H. Rinke, W. Siefken, L. Orthner et al.).
123. Bayer, DE 888 766, 1951 (W. Brenschede).
124. Du Pont, US 2 957 852, 1955 (P. E. Frankenburg, A. H. Frazer).
125. DIN 60 001, 1970.
126. H. Gall, K. H. Wolf in E. Becker, D. Braun (eds.): *Kunststoff-Handbuch,* **vol. 7,** Carl Hanser Verlag München-Wien 1983, pp. 611 – 627.
127. M. Couper in M. Lewin, J. Preston (eds.): *Handbook of Fiber Science and Technology, High Technology Fibers,* **vol. 3,** part A, Marcel Dekker, New York 1985, pp. 51 – 85.
128. H. Oertel in B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981, pp. 179 – 190.
129. P. A. Koch, *Chemiefasern/Text.-Ind.* **29/81** (1979) 95.
130. K. H. Wolf, *Text.-Prax. Int.* **36** (1981) 839.
131. D. J. Hourston, R. Meredith, *J. Appl. Polym. Sci.* **17** (1973) 3259.
132. R. Bonart, L. Morbitzer, *Kolloid-Z. Z. Polym.* **241** (1970) 909.
133. H. Hespe, E. Meisert, U. Eisele, L. Morbitzer et al., *Kolloid-Z. Z. Polym.* **250** (1972) 250.
134. R. Bonart, E. H. Müller, *J. Macromol. Sci. [B] Phys.* **10** (1974) 177.
135. R. Bonart, E. H. Müller, *J. Macromol. Sci. [B] Phys.* **10** (1974) 345.
136. R. Bonart, L. Morbitzer, E. H. Müller, *J. Macromol. Sci. [B] Phys.* **9** (1974) 447.
137. P. Hirt, H. Herlinger, *Angew. Makromol. Chem.* **40/41** (1974) 71.
138. H. Ishihara, J. Kimura, K. Saito, H. Ono, *J. Macromol. Sci. [B] Phys.* **10** (1974) 491.
139. J. Blackwell, K. H. Gardner, *Polymer* **20** (1979) 13.
140. L. L. Harrell, *Macromolecules* **2** (1969) 607.

141. H. Herlinger, P. Hirt, W. Aldinger, *Melliand Textilber.* 1979, 525.
142. L. Born, H. Hespe, J. Crone, K. H. Wolf, *Colloid Polym. Sci.* **260** (1982) 819.
143. L. Born, J. Crone, H. Hespe, E. H. Müller et al., *J. Polym. Sci., Polym. Phys. Ed.* **22** (1984) 163.
144. L. Born, H. Hespe, *Colloid Polym. Sci.* **263** (1985) 335.
145. Du Pont, DE 3 124 197, 1981 (E. M. Kenneth).
146. B. Küster, H. Herlinger, *Textil.-Prax. Int.* **36** (1981) 15.
147. Bayer, US 3 377 308, 1968 (H. Oertel, H. Rinke, W. Thoma).
148. H. Oertel, *Lenzinger Ber.* **45** (1978) 38 – 52.
149. H. Oertel, *Chemiefasern/Text.-Ind.* **27** (1979) no. 79, 1090 – 1096; **28** (1978) no. 80, 44 – 49.
150. Du Pont, DE 1 918 066, 1969 (Ch. G. Bottomley, O. L. Larry).
151. Amer. Cyanamid, US 3 577 388, 1969 (J. S. Megna).
152. Kanebo, JP 56 110 712, 1981.
153. T. Timm, *Kautsch. Gummi, Kunstst.* **37** (1984) 1021.
154. Bayer DE-AS 1 184 947, 1963 (F. Rosendahl, H. Rinke, H. Oertel).
155. Du Pont, EP 137 407, 1985 (C. C. Ketterer).
156. Teijin, JP 4 917 449, 1974 (H. Moriga).
157. Du Pont, EP 137 408, 1985 (W. Lewis).
158. Amer. Cyanamid, US 3 573 251, 1968 (J S. Megna, F. A. V. Sullivan).
159. Du Pont, DE 1 669 511, 1967 (O. L. Hunt).
160. Bayer, DE 1 495 830, 1964 (H. Oertel, H. Rinke, F. Moosmüller, F.-K. Rosendahl).
161. H. Gall, *Chemiefasern/Text.-Ind.* **32** (1982) no. 84, 510.
162. Du Pont, EP 0 046 073, 1980 (R. S. Hanzel, P. J. Sauer).
163. J. H. Saunders, K. C. Frisch: *Polyurethanes Chemistry and Technology in High Polymers,* **vol. 16 – 1,** Interscience Publishers, New York 1962, p. 129.
164. Bayer, DE 1 223 154, 1963 (W. Thoma, H. Rinke, H. Oertel).
165. Bayer, DE OS 1 288 235,1963 (H. Marzolph, H. Lenz, K. Bernklau, H. Scherzberg).
166. Du Pont, US 3 094 374, 1961 (P. M. Smith).
167. H. A. Pohl, *Text. Res. J.* **28** (1958) 473.
168. Bayer AG, Geschäftsbereich Fasern, Vertrieb: Produktinformation Dorlastan "Empfehlung zur Herstellung gewebter elastischer Bekleidungsstoffe mit Bayer Textilfaser Dorlastan," (available in English, "Suggestions on the Production of Woven Elastic Apparel Fabrics with Bayer Textile Fiber Dorlastan"), Bayer AG, Leverkusen, 1982.
169. Bayer AG, Geschäftsbereich Fasern, Vertrieb: Produktinformation Dorlastan "Bayer Textilfaser Dorlastan in der Rundstrickerei," (available in English, "Bayer Textile Fiber Dorlastan in Circular Knitting"), Bayer AG, Leverkusen, 1983.
170. K. Murenbeeld, *Textil.-Praxis Int.* **24** (1969) 335.
171. W. Müller, *Chemiefasern/Text.-Ind.* **20** (1970) 45.
172. Bayer AG, Geschäftsbereich Fasern, Vertrieb: Schrift "Bayer Textilfaser Dorlastan in der Kettenwirkerei," (available in English, "Suggestions on the Processing of Bayer Textile Fiber Dorlastan in the Warp-Knitting Industry"), Bayer AG, Leverkusen, 1977.
173. Montecatini, IT 535 712, 1954 (G. Natta, P. Pino, G. Mazzanti). Montecatini, IT 537 425, 1954 (G. Natta, G. Mazzanti, P. Pino). Montecatini, BE 538 782, 1954 – 1955 (G. Natta, P. Pino, G. Mazzanti, K. Ziegler).
174. M. Compostella in A. D. Ketley (ed.): *The Manufacture and Commercial Applications of Stereoregular Polymers,* 1st ed., **vol. 1,** Marcel Dekker, New York 1967, pp. 322 – 337.
175. M. Ahmed: *Polypropylene Fibers – Science and Technology,* Elsevier Scientific Publ. Co., Amsterdam-Oxford-New York 1982, pp. 181 – 184.
176. A. Tozzi, G. Cantatore, F. Masina, *Text. Res. J.* **48** (1978, Aug.) 433 – 436.
177. J. Helberg, *Chemiefasern/Text.-Ind.* **36/88** (1986, Jun.) 493 – 499
178. *Text. Organon* **57** (1986) no. 6, 116 – 140.
179. *Chemiefasern/Text.-Ind.* **33/85** (1983, Nov.) 670.
180. I. G. Farbenindustrie, DE 667 234, 1936 (H. Jacqué); DE 746 593, 1939.
181. H. Krässig, *J. Polym. Sci. Macromol. Rev.* **12** (1977) 321 – 410.
182. F. Hensen, S. Braun, *Kunststoffe* **68** (1978) 221 – 229.
183. F. Hensen, *Kunststoffe* **71** (1981) 643 – 652.
184. F. Hensen, M. Noe in VDI-Gesellschaft Kunststofftechnik (ed.): *Folien Gewebe Vliesstoffe aus Polypropylen,* VDI-Verlag, Düsseldorf 1979, pp. 107 –149, 151 – 174.
185. The Plastics & Rubber Inst. (ed.): *Prepr. Conference on Polypropylene Fibres and Textiles,* The Plastics & Rubber Inst., London Sept. 30 – Oct. 1, 1975; Sept 26 – 29, 1979; Oct. 4 – 6, 1983.
186. L. H. Gouw, W. H, Skoroszewski, *Plastica* **22** (1969) no. 10, 438 – 447.

187. W. H. Skoroszewski, *Plast. Polym.* **40** (1972) no. 147, 142 – 152.
188. Hercules, US 3 470 285, 1969 (C. W. Kim, S. D. Samluk).
189. A. R. Freedman, *Mod. Plast.* **47** (1970) no. 6, 108 –110.
190. J. Harms, H. Krässig, F. Sasshofer, *Chemiefasern Text.-Anwendungstech./Text.-Ind.* **23** (1973) 845 –849, 979 – 981.
191. D. E. Hanson, F. H. Reed, *Plast. Rubber Process.* **1** (1976) no. 1, 28 – 32.
192. P. E. Nott, *Kunststoffe* **67** (1977) 490 – 497.
193. "Das Verarbeiten von Polyolefin-Bändchen auf Raschelmaschinen," *Kettenwirkpraxis* **4** (1970) 9 – 18.
194. A. Balaz, *Melliand Textilber.* **6** (1971) 664 – 665.
195. G. Kautz, *Kunststoffe* **63** (1973) 682 – 686.
196. D. Moorwessel, G. Pilz, *Kunststoffe* **59** (1969) 205 – 210, 539 – 544.
197. Phillips Petroleum Co., US 3 302 501, 1967 (M. E. Greene).
198. Shell International Research M., NL 6 511 455, 1967.
199. Chevron Research Co., GB 1 128 274, 1967.
200. J. H. L. Choufoer: *Conference on Polypropylene Fibers in Textiles,* York Sept. 30 – Oct. 1, 1975.
201. Barmag, DE-AS 1 660 230, 1971 (H. Schippers, R. Hessenbruch); DE-OS 2 127 792, 1972 (E. Lenk, R. Jung, G. Albrecht).
202. Shell International Research M., NL 6 905 427, 1969; FR 2 046 406, 1971.
203. G. J. Nichols, *Plast. Polym.* **40** (1972) no. 146, 84 – 88.
204. F. Guggumus in R. Gächter, H. Müller (eds.): *Plastics Additives Handbook,* 2nd ed., Hanser-Verlag, Munich-Vienna-New York 1984, pp. 152, 158.
205. K. H. Kessels, *Text. Prax. Int.* **31** (1976) no. 3, 250 – 260; no. 4, 355 – 358.
206. H. Kirchenberger, *Melliand Textilber.* **7** (1974) 599 –603.
207. K. D. Darlington, *Knitting Times* 1975, (Aug.) 12 – 17.
208. M. R. Parsey: *Conference on Polypropylene Fibres and Textiles,* York Oct. 4 – 6, 1983.
209. F. Hensen, S. Braun, *Kunststoffe* **64** (1974) 228 – 233.
210. G. W. Ehrenstein, C. Maertin, B. Pornnimit, D. Scherz, *Plastverarbeiter* **37** (1986) no. 10, 206 – 219.
211. G. Capaccio, I. M. Ward, *Polymer* **15** (1974) 233 – 238.
212. A. G. Gibson, I. M. Ward, *J. Mater. Sci.* **15** (1980) 979 – 986.
213. P. D. Coates, I. M. Ward, *Polymer* **20** (1979) 1553 – 1560.
214. W. G. Perkins, R. S. Porter, *J. Mater. Sci.* **17** (1982) 1700 – 1712.
215. P. Smith, P. J. Lemstra, *J. Mater. Sci.* **15** (1980) 505 – 514.
216. Stamicarbon, BE 881 587, 1979; US 4 422 993, 1983 (P. Smith, P. J. Lemstra).
217. A. G. Gibson, I. M. Ward, *Polymer Eng. Sci.* **20** (1980) no. 18, 1229 – 1235.
218. *Chemiefasern/Text.-Ind.* **35/87** (1985) no. 1, 33 – 34.
219. R. Kirschbaum, H. Yasuda, E. H. M. van Gorp, *International Man-Made Fibers Congress,* Dornbirn, Sept. 24 – 26, 1986.
220. B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim, 1981 p. 198.
221. H. Klare: *Geschichte der Chemiefaserforschung,* Akademie-Verlag, Berlin 1985, pp. 180 – 186.
222. Comité International de la Rayonne et des Fibres Synthétiques (CIRFS) (ed.): *Zahlen über Chemiefasern,* CIRFS, Paris Jan. 22, 1985.
223. P. A. Koch, "Faserstoff-Tabellen," *Chemiefasern/Text.-Ind.* **27/29** (1977, Jun.) 513 – 524.
224. *Textil-Praxis Int.* **5** (1985) 549 – 552.
225. J. F. Kroschwitz (ed.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., J. Wiley & Sons, New York-Chicester-Brisbane-Toronto-Singapore 1985, p. 426.
226. G. Svegliado, G. Talamini, G. Vidotto, *J. Polym. Sci., Part. A* **5** (1967) 2875 – 2881.
227. Stamicarbon, EPA 0 144 983, 1985 (R. M. A. M. Schellekens, P. J. Lemstra, G. M. L. M. Leherte et al.).
228. P. Fritsche, J. Ulbricht, *Faserforsch. Textiltech.* **14** (1963) 320 – 325; **14** (1963) 517 – 521; **15** (1964) 93 – 100.
229. L. H. Peebles, Jr., R. B. Thompons, Jr., J. R. Kirby, M. E. Gibson, *J. of Appl. Polym. Sci.* **17** 1973 113. J. Tsuda, *J. of Appl. Polym. Sci.* **13** (1961) 104.
230. Du Pont, DE-AS 1 174 070, 1964 (G. N. Milford, W. K Wilkinson).
231. *Houben-Weyl,* **14/1,** 297 ff and p. 973 ff.
232. M. Sittig: *Chem. Process Review No. 62: Acrylic and Vinyl Fibers,* Noyes Data Corp., Park Ridge, N. J., 1972.
233. Hoechst AG, DE 3 048 059, 1982 (E. Höroldt, H. Vollmüller, H. D. Berndhard, H. Strobel).

234. J. Runge, *Acta Polym.* **33** (1982) no. 12, 708 – 713.
235. J. Runge, *Acta Polym.* **34** (1983) no. 10, 631 – 636.
236. VEB Friedrich Engels, DD 215 342, 1984 (J. Aurich, H. Ebeling, E. Seidel, P. Leppin).
237. Bayer AG, DE 2 832 212, 1980 (K. H. David, A. Nogaj, H. Rinkler).
238. Bayer AG, DE 3 333 145, 1985 (H. Engelhard, A. Nogaj, M. Reichardt, P. Kruchem).
239. Bayer AG, DE 3 333 146, 1985 (J. König, S. Korte, H. Engelhard).
240. K. Jost, *Rheol. Acta* **1** (1958) 303 – 315
241. G. C. East, J. E. Mc Intyre, G. C. Ratel, *J. Text. Inst.* **75** (1984) no. 3, 196 – 200.
242. Bayer AG, DE 3 424 343, 1986 (Ch. Pieper, H. K. Burghartz, R. Hirsch, N. Rink).
243. Bayer AG, DE 3 141 490, 1983 (H. J. Behrens, H. Herold, E. Muschelknautz, R. Vogelsgesang).
244. Du Pont, DE 3 339 501, 1984 (Berry, Jr., W. Cunningham).
245. F. Schmitt, *Melliand Textilber.* **12** (1982) 830 – 835.
246. Bayer AG, DE 3 308 657, 1984 (M. Bueb, D. Paulini, E. Muschelknautz, W. Wagner et al.).
247. Bayer AG, DE 3 225 266, 1984 (M. Reinehr, H. Uhlemann).
248. W. Teige, *Chemiefasern/Text.-Ind.* **33/85** (1983, Sept.) 636 – 642.
249. R. Detscheva, R. Fltscheva, St. Stojanov, K. Dimov, *Textilveredlung* **9** (1974) 312 – 316.
250. Hoechst AG, DE 3 141 082, 1983 (M. Hähnke, W. Teige).
251. D. E. Titheridge, *Text.-Betr.* **100** (1982, Jul./Aug.) 45, 46, 58.
252. Bayer AG, DE 2 317 132, 1974 (M. Reinehr, A. Nogaj, G. Hölzing).
253. J. G. Cook: "Man-made Fibers," *Handbook of Textile Fibers,* vol. 2, Merrow Publ. Co. Watford, England, 1968, p. 442.
254. T. Müller, *Chemiefasern/Text.-Ind.* **35/87** (1985, Jun.) 390 – 392.
255. H. Engelhard, *Chemiefasern/Text.-Ind.* **34/86** (1984, Jun.) 400 – 401.
256. O. Glenz, W. Beckmann, *Melliand Textilber.* **38** (1957) 296 – 300, 783 – 787.
257. C. A. Streuli, *Fresenius Z. Anal. Chem.* **153** (1956) 71.
258. H. Nestler, J. Mai, V. Stepputat, *Acta Polym.* **31** (1980) no. 4, 267 – 271.
259. J. Glacén, J. Maillo, J. T. Baixauli, *Melliand Textilber.* **66** (1985) 127 – 132.
260. R. V. Flor, M. J. Prager, *Text. Res. J.* **53** (1983, Jan.) 15 – 18.
261. "Spezialitäten-Lieferprogramm der Chemiefaser-Hersteller," *Chemiefasern/Text.-Ind.* **33/85,** (1983, Nov.) 782 – 786.
262. H. Haehne, *Chemiefasern/Text.-Ind.* **33/85** (1983, Dec.) 839, 842 – 844, 846.
263. J. Vogelsgesang, H. D. Gölden, *Chemiefasern/Text.-Ind.* **32/84** (1982, Jun.) 422 – 432.
264. F. Fourné, *Chemiefasern/Text.-Ind.* **32/84** (1982, Jun.) 433–436.
265. J. Atkinson in P. J. Alvey (ed.): *The Production of Man-made Fibers, Textile Progress,* **vol. 8,** no. 1, Textile Institute, Manchester 1976, pp. 61 – 69.
266. W. Körner, G. Blankenstein, P. Dorsch, U. Reinehr, *Chemiefasern/Textil.-Ind.* **29/81** (1979, Jun.) 452 –462.
267. Bayer AG, DE 2 554 124, 1977 (U. Reinehr, T. Herbertz, H. Jungverdorben, P. Kleinschmidt et al.).
268. Mitsubishi, GB 1 377 200, 1974.
269. W. E. Fitzgerald, J. P. Knudsen, *Text. Res. J.* **37** (1967) no. 6, 447 – 453.
270. O. Heuberger, A. J. Ultee, *Lenzinger Ber.* **38** (1975) 154 – 163.
271. Du Pont, DE 1 435 611, 1970 (M. L. Davis, L. Merton, S. C. Camden).
272. Rhodiaceta, FR 913 919, 1942.
273. Ch. Heiberger, *Encyclopedia of PVC,* Dekker, New York 1986, pp. 416 – 429, 417 – 428, 632 – 636.
274. Rhône-Poulenc, FR 2 475 418, 1980 (A. Berchoux).
275. Teijin, JP 352 668, 1968.
276. Rhovyl, FR 2 524 475, 1982 (G. Achard, G. Anouilh, J. Menault).
277. Rhône-Poulenc, FR 2 495 645, 1980 (G. Achard, J. Menault, C. Thouvenot).
278. A. Rogovin, W. Albrecht: *Chemie Fasern,* G. Thieme Verlag, Stuttgart 1981, pp. 314 – 322.
279. Rhovyl, FR 2 524 498, 1982 (G. Achard, P. Chion, J. Menault).
280. B. Catoire, G. Gastadi, R. Hagege, *Bull. Sci. Inst. Text. Fr.* **11** (1974) 205 – 217.
281. Rhône-Poulenc, FR 2 483 966, 1980 (P. Chion, J. Menault, A. Rodier, J. P. Sacre).
282. E. A. Rassolova et al., *Fibre Chem.* **15** (1983) no. 4, 8.
283. J. Menault, *Ind. Text. (Paris)* **1158** (1985) 829 – 831.
284. Shirley Institute, Report DS. 1045, Dec. 12, 1983, unpublished.
285. D. Werner, *Textiltechnik* **35** (1985) 8.

286. B. von Falkai, *Synthesefasern,* Verlag Chemie, Weinheim 1981, pp. 216 – 218.
287. *Ullmann,* 4th ed., **7,** pp. 296 – 308.
288. Nippon Kagaku Sen'i Kenkyusho, JP 147 958, 1939 (I. Sakurada, S. Lee, M. Kawakami); JP 157 076, 1940 (S. Lee, I. Sakurada, H. Kawakami, K. Hirabayashi et al.).
289. Japan Chemical Fibers Assn. (ed.): *Kasen Handbook (Handbook of Chemical Fibers),* Japan Chemical Fibers Assn., Tokyo 1985.
290. A. Mizobe, *Sen'i Gakkaishi* **41** (1985) 180.
291. Kurashiki Rayon Co., US 2 642 333, 1950 (T. Tomonari, T. Akaboshi, M. Nagai, T. Osugi).
292. Noguchi Laboratory, JP 162 708, 1942 (M. Nakanishi, M. Kubota).
293. Kurashiki Rayon Co., US 3 660 556, 1969 (T. Ashikaga, T. Kosaka).
294. Kuraray Co., JP-Kokai 56-125 264, 1984 (S. Oka, M. Mizobe, J. Hikasa, M. Okazaki et al.).
295. Kurashiki Rayon Co., US 3 427 298, 1969 (K. Tsuboi, T. Mochizuki).
296. Allied, US 4 440 771, 1982 (Y. D. Kwon, S. Kavesh, D. C. Prevosek).
297. Kurashiki Rayon Co., JP 317 932, 1958 (S. Nakajo, E. Morita).
298. Kurashiki Rayon Co., JP 232 163, 1956 (K. Kawai, S. Miyazaki, K. Tanabe, T. Osugi).
299. M. Uzumaki, *Sen'i Gakkaishi* **33** (1977) 55.
300. Toyo Chemical Co., JP 417 267, 1961 (S. Okamura, H. Asakura).
301. Du Pont, US 2 559 750, 1951 (K. L. Berry).
302. *Kirk–Othmer,* **vol 19,**1992 p. 519

Fibers, 5. Synthetic Inorganic

Erich Fitzer, Institut für Chemische Technik der Universität Karlsruhe, Karlsruhe, Federal Republic of Germany (Chaps.1 and 5)

Rudolf Kleinholz, Gevetex Textilglas-GmbH, Herzogenrath, Federal Republic of Germany (Section 2.1)

Hartmut Tiesler, Grünzweig + Hartmann und Glasfaser AG, Ladenburg, Federal Republic ofGermany (Section 2.2)

Martyn Hugh Stacey, ICI Chemicals & Polymers Ltd., Runcorn, United Kingdom (Chaps. 3 and 6)

Roger De Bruyne, N.V. Bekaert S.A., Zwevegem, Belgium (Chap. 4)

Ignace Lefever, N.V. Bekaert S.A., Zwevegem, Belgium (Chap. 4)

Michael Heine, Institut für Chemische Technik der Universität Karlsruhe, Karlsruhe, Federal Republic of Germany (Chap. 5)

1. Survey . . . 3
1.1. General Description of Inorganic Fibers . . . 3
1.1.1. Commercial Forms and Terminology . . . 3
1.2. Crystalline Structure and Mechanical Properties . . . 4
1.3. Testing Methods for Inorganic Fibers . . . 6
1.4. Inorganic Fibers for Thermal Insulation . . . 7
1.5. Inorganic Fibers for Reinforcement Purposes . . . 8
1.5.1. Principles of the Fiber Reinforcement . . . 8
1.5.2. Commercial Continuous Inorganic Fibers . . . 9
1.5.3. Surface Chemistry, Surface Treatment, and Surface Properties . . . 11
1.5.4. High-Temperature Behavior of Inorganic Fibers . . . 12
2. Glass Fibers . . . 13
2.1. Textile Glass Fibers . . . 13
2.1.1. Chemical Composition and Properties . . . 14
2.1.2. Manufacture . . . 16
2.1.2.1. Glass Filaments . . . 16
2.1.2.2. Glass Staple Fibers . . . 18
2.1.3. Textile Glass Products: Processing and Application . . . 18
2.1.3.1. Product Designation . . . 18
2.1.3.2. Correlation Between Processing and Application . . . 18
2.1.3.3. Description of Textile Glass Products . . . 19
2.1.3.4. Applications . . . 20
2.1.4. Product Standards . . . 21
2.1.5. Toxicology and Occupational Health 21
2.1.6. Economic Aspects . . . 22
2.1.7. Producers . . . 22
2.2. Glass Fibers for Insulation . . . 22
2.2.1. Physical Properties . . . 22
2.2.2. Chemical Properties . . . 23
2.2.3. Raw Materials and Chemical Composition . . . 23
2.2.4. Production . . . 24
2.2.5. Environmental Protection . . . 27
2.2.6. Quality and Analysis . . . 28
2.2.7. Applications . . . 28
2.2.8. Toxicology and Occupational Health 29
3. Refractory Fibers . . . 29
3.1. Properties and Raw Materials . . . 30
3.1.1. Physical Properties . . . 30
3.1.2. Chemical Properties . . . 33
3.1.3. Resources, Raw Materials . . . 33
3.2. Production . . . 34
3.2.1. Boron Monofilament . . . 34
3.2.2. Silicon Carbide Fibers . . . 35
3.2.3. High Alumina Fibers . . . 36
3.2.4. Zirconia Fibers . . . 37
3.3. Environmental Protection . . . 38
3.4. Uses . . . 38
3.4.1. High-Temperature Insulation and Filtration . . . 38
3.5. Miscellaneous Uses . . . 39
3.6. Economic Aspects . . . 39
3.7. Toxicology and Occupational Health . . . 40
4. Metal Fibers . . . 40
4.1. Production . . . 40
4.1.1. Bundle Drawing . . . 40
4.1.2. Melt Spinning . . . 41
4.1.3. Other Processes . . . 42
4.2. Bundle-Drawn Fibers . . . 43

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_001

4.2.1. Form and Dimensions 43
4.2.2. Composition 43
4.3. Uses 43
4.4. Manufacturers 45
4.5. Melt Spun Fibers 45
4.5.1. Properties 45
4.6. Uses 45
5. Carbon Fibers 45
5.1. Properties 47
5.1.1. Mechanical Properties 47
5.1.2. Chemical Composition 47
5.1.3. Structure 48
5.1.4. Physical Properties 51
5.1.5. Chemical Properties 51
5.2. History 54
5.2.1. Rayon-Based Carbon Fibers 54
5.2.2. PAN-Based Carbon Fibers 55
5.2.3. Pitch-Based Carbon Fibers 55
5.2.4. Vapor-Grown Carbon Fibers 56
5.3. Production 56
5.3.1. PAN Precursor Fiber 57
5.3.2. Stabilization 58
5.3.3. Process Parameters 59
5.4. The Pitch Fiber Process 59
5.4.1. Isotropic and Anisotropic Pitch Fiber Precursors 59
5.4.2. The Mesophase Formation 60
5.4.3. Process Steps 60
5.5. Comparison of the Carbon Fibers-Based on PAN and on Pitch 63
5.6. Commercial Availability and Economic Aspects 64
5.6.1. Companies and Trade Names 64
5.6.2. World Capacity 64
5.7. Future Importance 65
6. Refractory Ceramic Fibers 67
7. References 67

1. Survey

1.1. General Description of Inorganic Fibers

The inorganic fibers are treated in five chapters: Glass Fibers, Refractory Fibers, Metal Fibers, Carbon Fibers and Refractory Ceramic Fibers.

In this introduction, the similarities with organic synthetic fibers, as well as special characteristics, are treated generally.

Unlike organic fibers, there are no natural sources for inorganic fibers, with one exception, asbestos fibers, which are a separate keyword (→Asbestos). As synthetic products, inorganic fibers have obvious similarities with the synthetic organic fibers, first of all in the morphology and terminology of the commercially available forms and in their field of technical application. Synthetic organic (Chardonnet silk, a regenerated cellulose fiber (see Fibers, 1. Survey) and inorganic fibers were first fabricated a hundred years ago. The critical difference between inorganic and organic fibers is seen in the differing properties, which correspond in general to those of organic polymers on the one hand and to those of inorganic and metallic materials on the other hand.

Nonmetallic inorganic fibers are stronger, stiffer, higher melting, and thus more heat resistant, and in all cases nonflammable, but, except for metal fibers, they are also characterized by brittleness. Based on these general properties, they are less important as textiles per se; however, they are superb for reinforcement of other materials, in the so-called composites (→Composite Materials).

They are excellent as thermal insulation at high temperatures and in corrosive environments.

1.1.1. Commercial Forms and Terminology

The term *fiber* can be used for a formed material that has a length-to-diameter ratio of at least 10 : 1, and with a cross-sectional area below 0.005 mm^2 and a thickness below 0.25 mm [1]. This standardized nomenclature also includes *whiskers*, which are mainly single-crystal materials, and are treated under a separate keyword (→Whiskers). Nonmetallic fibers include both continuous (endless) and short fibers. Diameters range between 5 and 15 µm, but some are much thicker, for instance, the fibers made by chemical vapor deposition on a substrate fiber such as B or SiC fibers which have diameters of ca. 100 – 150 µm. The diameter is of critical importance in regard to the tensile strength, because all materials showing brittle fracture increase in strength as the diameter decreases (Fig. 1).

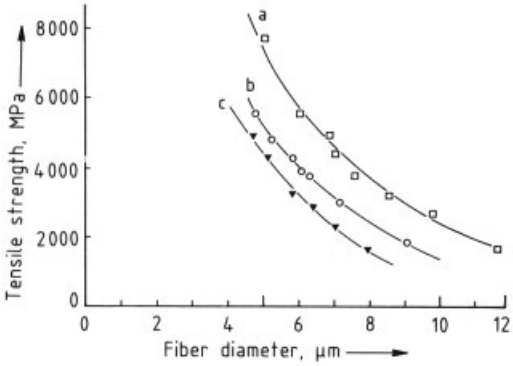

Figure 1. Tensile strength of monofilaments (carbon fibers) as a function of fiber diameter. a) Hercules AS4 (Hercules); b) Torayca T300 (Toray); c) Torayca M40 (Toray)

Short fibers can be fabricated by a melt spinning process directly to staple fibers (see pp. 16 and 23 – 24) or by chopping (cutting) of continuous fibers. Commercial short-fiber products include wool, mats, and felts. The spinning of short fibers to continuous yarn, as with natural organic fibers, is generally not practiced for synthetic inorganic fibers.

Continuous fibers are fabricated by melt spinning, if the fiber materials are meltable, such as is the case for glass or mineral fibers. The advanced inorganic fibers, such as carbon fibers and also those of refractory ceramics, are made from organic precursor fibers via thermal degradation. For the spinning of the precursors the methods for fabrication of synthetic organic fibers are used. This precursor technique can produce both monofilaments and multifilaments, as described in "Organic Fibers" (Fibers, 4. Synthetic Organic). Only filament yarns which consist of up to 300 000 filaments are commercially available (see "multifilament tow" in Chap. 5.6.2). For reinforcement in advanced composites, tows of 3000 to 12 000 monofilaments (3 – 12 K tows) are used.

Meltspun glass fibers are also fabricated as filament tow and applied in rowings of some thousands of monofilaments. The terms *rovings* and *tows* are synonymous for filament yarn. Fabrics from filament yarns are commercially available. They are fabricated by weaving, braising, or similar techniques without binder to textiles, tubes, filter sieves, cables, etc. The chemical vapor deposition (CVD) method of fabrication, a method developed to produce boron fibers and still used for B and SiC fibers, can only be used to produce monofilament fibers. Such monofilament fibers are used to reinforce polymers, although of decreasing importance. However, such fibers may be useful in the reinforcement of light metals. The CVD technique can be used with success to fabricate yarns of high-melting substrate fibers with a coating on the individual monofilament, e.g., SiC or B_4C on boron or TiN, SiC, etc. on carbon.

1.2. Crystalline Structure and Mechanical Properties

Glass and mineral fibers are characterized by their amorphous structural state, while metal and some oxide fibers such as the DuPont A1203 fiber are characterized by their polycrystallinity.

In the case of most refractory ceramics produced from organic precursors, and also in the case of carbon fibers, the structure is not fully amorphous, but the crystals in the fibers are so imperfect that the term polycrystallinity can be used only with reservation. As a consequence, postheat-treatment of such fibers above 1000 °C easily initiates crystallization processes as demonstrated on quite pure Si_3N_4 fiber (experimental-type Tonen in 1988, [2]) in Figure 2. The room temperature strength of the fabricated fiber with final pyrolysis temperatures of 1000 °C in nitrogen is 2450 MPa, an impressive value, but after heat treatment at

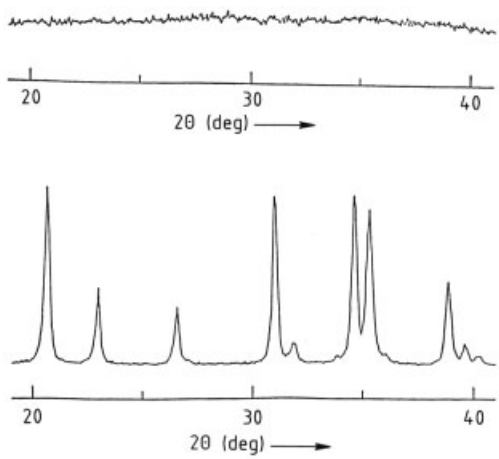

Figure 2. Composition and X-ray diffraction of an experimental-type silicon nitride fiber with the composition in mass fraction: 59.2 % Si, 39.1 % N, 0.5 % C, 1.1 % O, 0.1 % H; θ = Bragg angle [2]

1600 °C, a perfect β-Si_3N_4 crystal structure develops, and room-temperature strength drops below 1000 MPa, a considerable lowering of the mechanical properties. Figure 3 shows further examples where the decrease in strength due to thermally activated crystallization is even more extreme.

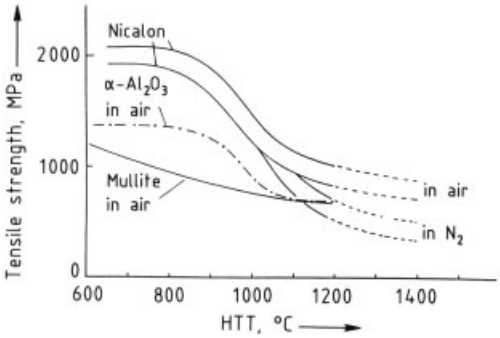

Figure 3. Decrease of tensile strength of various inorganic fibers made by the precursor method (room temperature tensile strength) HTT = heat treatment temperature

The well-known strength decrease of glass fibers at elevated temperatures is caused by facilitated flow processes because of lowered viscosity (Fig. 4).

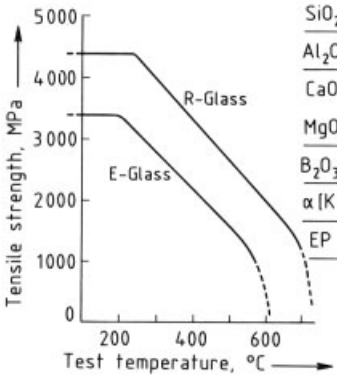

	E-Glass	R-Glass
SiO_2	53-54%	60%
Al_2O_3	14-15%	25%
CaO	20-24%	6%
MgO		9%
B_2O_3	6-9%	
α [K^{-1}]	5×10^{-6}	4×10^{-6}
EP	840°C	990°C

Figure 4. Decrease in tensile strength and composition of glass fibers at elevated test temperatures

The drop in strength by postheat-treatment is also observed for carbon fibers and explained by thermal healing of structural defects within the polyaromatic layers, thus breaking cross-linkages between them. This healing process enormously increases the tensile modulus of the fibers because of the extreme anisotropy of the graphite-like carbon structure. In no case can three-dimensional crystallinity, as in the graphite structure, be achieved by postheat-treatment of carbon fibers, not even above 2700 °C. Therefore, carbon fibers never consist of polycrystalline graphite and the term "graphite fibers," as used commercially in the United States, is scientifically unjustified and misleading.

A characteristic of inorganic fibers, whether amorphous or polycrystalline is that they all exhibit brittle fracture without plastic deformation. Figure 5 shows some examples. Only metallic fibers behave differently.

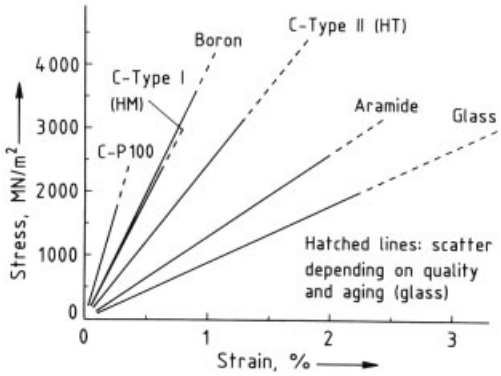

Figure 5. Stress – strain diagram of some inorganic fibers

The strain-to-failure (elongation at fracture) for the brittle fibers follows precisely the ratio strength to tensile modulus. This explains why high-modulus inorganic fibers show the lowest values of strain to failure, values below 0.5 %. For reinforcement of polymers, as in advanced composites, an elongation up to 2 % is required by aircraft designers. This requirement is met by the low-modulus glass fibers but not by the advanced carbon and ceramic fibers. Glass fibers, therefore, are privileged in engineering application if fracture toughness and impact resistance are required without the need for extreme stiffness. If high tensile moduli have to be maintained, the elongation can be improved only by increasing the fiber strength (Fig. 5).

It has been well known since GRIFFITH that the fracture in brittle materials is initiated by tensile stresses far below the intrinsic strength of the material due to notches of a minimum – the so-called critical length [3]:

$$\sigma = \sqrt{\frac{E \cdot \gamma}{C}}$$

The probability for the presence of such a notch decreases with decreasing volume of the tested sample. For fibers having the same diameter, the measured values for the tensile strength increase with shorter gauge length and if the same testing length is used, with smaller diameter of the same fiber type. Figure 1 shows the tensile strengths of monofilaments of commercial fibers as a function of the fiber diameter.

As a consequence, the efforts to develop new high strength inorganic fibers for reinforcement are directed to smaller diameter filaments, for instance 5 µm. However, this way to improve strength and elongation is limited by the danger of carcinogenicity of all fiber materials thinner than 3 µm that have length to diameter ratio larger than 3 ($l/d > 3$)

if the fiber material is persistent *in vivo*. A hypothesis concerning cancerogenic potency on geometric parameters was published in 1978 (Fig. 6) [4].

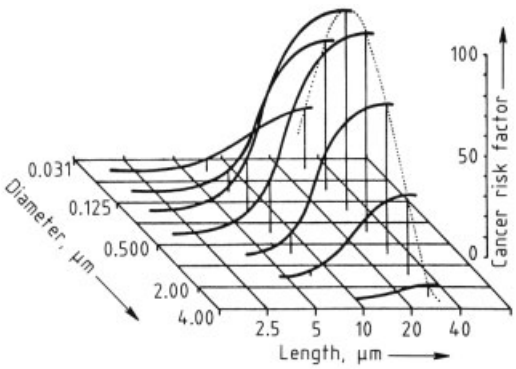

Figure 6. Carcinogenicity of thin inorganic fibers [4]

As a further consequence, the testing method, especially the gauge length of the fiber to be tested for tensile strength, has to be strictly standardized.

1.3. Testing Methods for Inorganic Fibers

Inorganic fibers are characterized mainly by tensile strength. Longfibers, at least 30 mm in length, are tested as monofilaments in a paper frame as demonstrated in Figure 7. Each commercial testing machine can be used for this purpose. A gauge length of 25 mm is the standard [5]. The critical problem is the determination of the monofilament cross section. In the case of circular morphology, the measurement of the monofilament diameter by a scanning microscope gives reproducible results. The diameters of the monofilaments in a bundle can vary ± 10 % [6]. If the fibers have irregular morphology, only integration of the cross sectional areas of the monofilaments in samples cut perpendicularly to the fiber axis using optical or electronic microscopy comes into question. Figure 8

The only commercial method is the strand method [7].

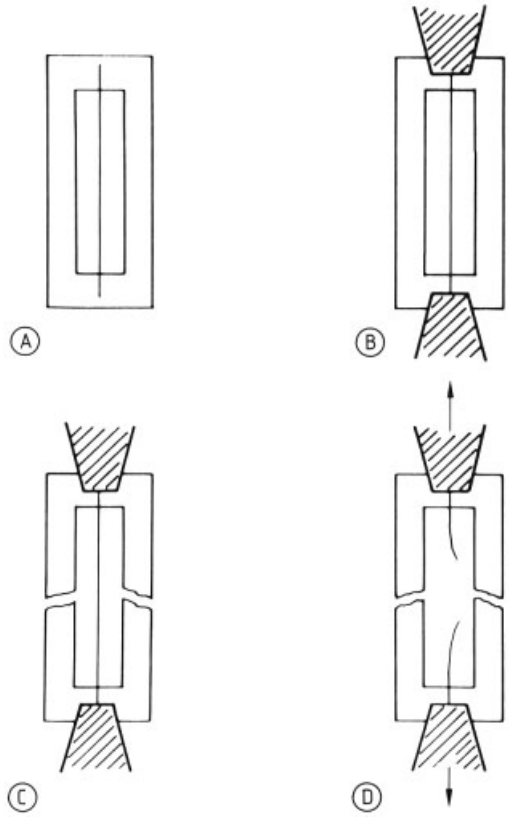

Figure 7. Paper frame method for testing room temperature tensile strength of thin monofilaments
A) Paper frame with the monofilament test sample fixed by an adhesive on the frame; B) The paper frame in the holding equipment of the testing machine; C) The paper frame after cutting; D) The test specimen after measurement

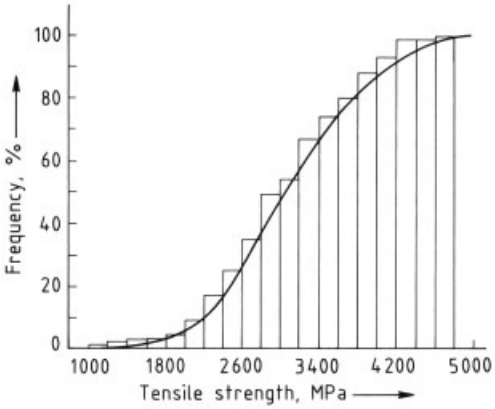

Figure 8. Distribution of tensile strength of monofilaments of a carbon fiber tow [6]

The cross section of the bundle may also be determined by the mass per length of the tow and by the pycnometer density or alternatively by the He density of the fiber material. The precondition for this cross section determination is a circular morphology.

The term "dtex" (1 tex = 1 g per 1000 m) is used to characterize the mass of a filament fiber; this is the same as for synthetic organic fibers (→ Fibers, 4. Synthetic Organic).

A special test is the *loop test*, which is applicable for thicker monofilaments (see Fig. 9). This test can be modified to the *knot test* for fast qualitative characterization of the flexibility of filament fibers.

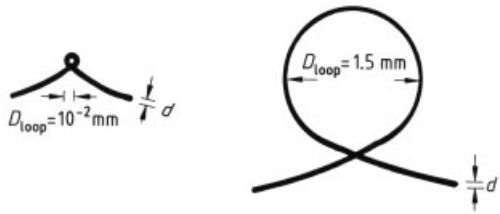

Figure 9. Loop test for low modulus A) and high modulus B) monofilaments indicating flexibility of inorganic fibers

$$D_{\text{loop}} \sim \left(\frac{d}{\varepsilon}\right)_{\text{fiber}} = \left(\frac{d \cdot E}{\sigma}\right)_{\text{fiber}}$$

D = loop diameter, d = fiber diameter, ε = fiber elongation, E = fiber modulus, σ = fiber strength

The tensile modulus is measured during the tensile strength test or nondestructively by the *ultrasonic method*, which gives good comparative results.

The compressive strength of the various inorganic fibers is a further characteristic property, but can only be measured if the fibers are inbedded in a unidirectional composite with a resin as matrix and a fiber volume fraction of 60 %. Isotropic fibers have high compressive strength, but the fibers with anisotropic structure and high preferred orientation parallel with the fiber axis show lower compressive strengths. The worst compressive strengths appear for the ultra-high-modulus types; here, the mesophase pitch-based carbon fibers are surpassed in a negative sense by the high-modulus organic fibers.

The determination of the linear thermal expansion of fibers is straight-forward; however, the determination in the perpendicular direction is extremely difficult. For carbon fibers this is important because of the high degree of anisotropy and the variety of degrees in preferred orientation. Data from direct optical measurements are compiled in Table 1. An indirect method is the comparative dilatation measurement using the pure resin and the fiber-reinforced composite.

Table 1. Thermal expansion of inorganic fibers

Fiber type	Thermal expansion, K^{-1} *	
	$\alpha_{\parallel} \cdot 10^{-6}$	$\alpha_{\perp} \cdot 10^{-6}$
Glass fiber		
E glass	5.4	5.4
S glass	1.6	1.6
Carbon fiber	(−0.5) – 0.2	10 – 15
CVD Monofilaments		
SiC fibers	3.1 – 4.5	3.1 – 5.2
Oxide fibers	7.0 – 8.6	7.0 – 8.6

* parallel (||) or perpendicular (⊥) to the fiber axis

1.4. Inorganic Fibers for Thermal Insulation

The use of inorganic fibers for insulation is based not only on the low thermal conductivity of the material itself (one to two orders of magnitude less than that of metals) but also on the large volume between the fibers. These two aspects are discussed in detail in the chapter 2. Application of glass fibers is limited as shown in Table 13, but refractory oxide fibers allow use at much higher temperatures. Saffil fibers Al_2O_3 and ZrO_2 wool from ICI are examples (see Section 3.2.4). Upper application temperature is ca. 1650 °C. Even carbon felt can be used as thermal insulation; in

Table 2. Thermal conductivity of insulation layers from inorganic fibers

	Measured at °C	in W/mK
Al_2O_3-Blanket (Saffil)	200	0.1
	1000	0.26
	1400	0.5
Zirkonia-Blanket (Saffil)	200	0.05
	1000	0.22
	1400	0.42
Silica-Wool	200	0.04
(Quarz and Silica)	800	0.14
Mullite-Texture (Nextel)	RT	0.09
	200	0.12
	1000	0.87
Zirkonia-Insulation-Plate		
Type ZYZ3 (Zircar)	400	0.09
	800	0.11
	1100	0.14
	1400	0.19
	1650	0.23
Carbon-Felt (Sigri)		
HTT = 1000 °C	20	0.06
HTT = 2000 °C	20	0.08

For comparison thermal conductivity of various materials, measured at room-temperature in W/mK

Cu	Al	Steel	C-Fiber	Glass	Polyethylene
370	175	46	12 – 23	0.9	0.23

Effect of insulation layers

Type	Thickness	Hotface	Coldface
Mullite	1 layer (ca. 3 mm)	1000 °C	500 °C
(in air)	3 layers (ca. 9 mm)	1000 °C	300 °C
Carbon Felt	40 mm	1600 °C	360 °C
(in argon)	40 mm	1200 °C	260 °C

spite of the high thermal conductivity characteristic of all carbon materials, the upper application temperature in a non-oxidizing environment is 2000 °C (see Table 2).

Mostly short fibers, such as wools, felts, textiles, and similar fabrics, are used. Mechanical properties and variation in individual fiber thickness is of little importance.

1.5. Inorganic Fibers for Reinforcement Purposes

1.5.1. Principles of the Fiber Reinforcement

Controlled inclusion of fibers into a homogeneous matrix produces a fiber-reinforced composite. The properties of the composite can be precalculated from the properties of the tow components, their volume fractions, and the geometrical arrangement of the fibers within the isotropic matrix. The simplest case is the unidirectional (UD) reinforcement with continuous fibers, where the rule of mixtures can be applied (Fig. 10A). Comparison of the precalculated and actual composite properties reveals the efficiency of the translation of the fiber properties into the composite. With good adhesion between fibers and matrix, efficiencies of 80 – 100 % can be achieved, especially as far as the moduli are concerned. Exact precalculations are described in [8] and elsewhere.

Figure 10 (see next page). Principles of reinforcement in unidirectional composites
A) Rule of mixture for strength data in fiber direction E_c = Young's modulus of composite, $E_{F,M}$ = Young's modulus of fiber and matrix, $x_{F,M}$ = Volume fraction of fiber and matrix, η_E = Yield of fiber modulus in the composite, a) Fiber; b) Composite; c) Matrix
B) Angle dependence of strength a) in unidirectional long fiber reinforced composites and b) in isotropic short fiber composites [9]
C) Stiffening effect by short carbon fibers in aluminum and thermoplastics a) C-Fiber/Thermoplastics; b) C-Fiber/Aluminum

Estimation of the strengths by precalculation, even in unidirectional reinforced samples, is more difficult, because the fracture processes overlap with the reinforcement effect. If we consider a polymer matrix with strain-to-failure exceeding that of the fibers, which is generally the case, the crack starts in the fibers. The degree of adhesion between fibers and matrix and the nature of the interphase contribute to the extent of the reinforcement effect. In addition, chemical and physical compatibility must be considered. Nevertheless, the simple rule of mixtures is still useful and reasonably valid.

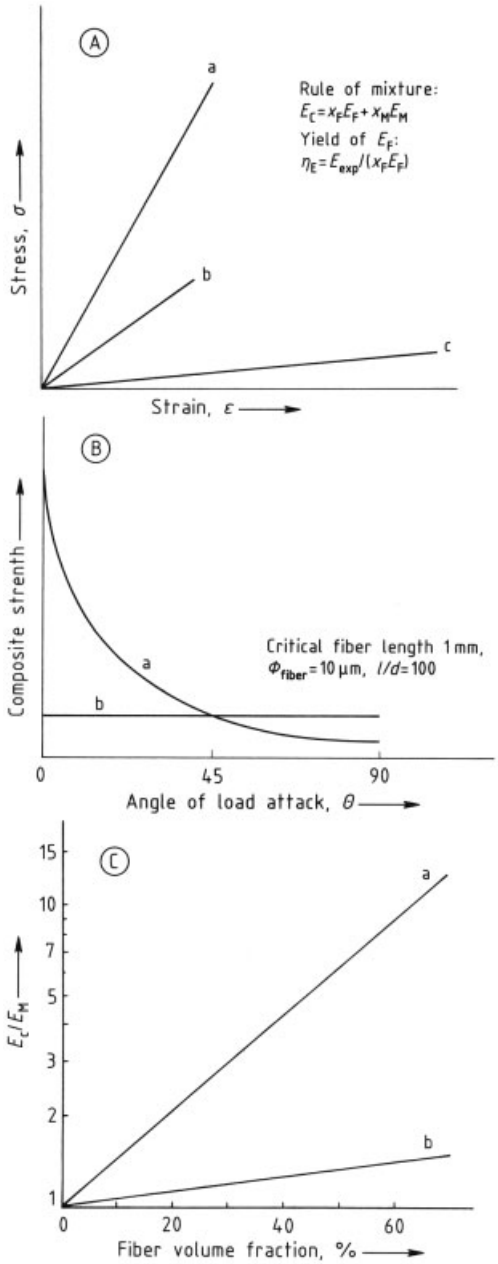

The precalculated reinforcement for UD composites is achieved only in the fiber direction. For other directions, reinforcement is less effective. The strength of a UD composite across the fiber axis is that of the pure polymer (Fig. 10B) [9].

Short fibers arranged statistically in all directions, produce reinforcement of 1/6 of that precalculated for UD composites by the rule of mixtures, but this strengthening is isotropic. However, it should be demonstrated in each case whether or not reinforcement is worthwhile for a specific material combination. For example, as shown in Figure 10C, addition of short carbon fibers to an aluminum matrix is not effective [10]. Precise stiffness and strength information is a prerequisite for the successful design of fiber-reinforced matrices.

1.5.2. Commercial Continuous Inorganic Fibers

Mechanical Properties For a better understanding of the character, properties, and the importance of today's inorganic fibers and future trends, a brief history is helpful. The fiber reinforcement of polymers started during the 1930s but was not commercially successful until the 1940s, only after the interface problem had been solved by means of coupling agents. In the 1950s the U.S. Air Force and later the NASA supported development of reinforcing fibers to achieve light weight and high stiffness for future airplanes. Boron, a light but high-melting element, was selected as the most promising candidate. A new production technology was developed: *chemical vapor deposition* on thin tungsten wires. The boron fibers fabricated in this way are monofilaments with diameters between 120 and 150 µm and have more the character of thin wires than that of fibers. Therefore chemists started to develop filament fibers from the neighboring element carbon, which is also light and has a high melting point, and from which fibers have been demonstrated 100 years earlier as fila-

Table 3. Mechanical and physical properties of the representative inorganic fibers

Fiber type	Room temperature properties		
	Density, g/cm^3	Tensile strength, MPa	Tensile modulus, GPa
Glass fiber	2.50 – 2.62	3400 – 4500	70
Carbon fiber	1.76 – 2.10	2000 – 7000	240 – 700
CVD ∗ Monofilaments			
SiC fibers	2.55 – 3.50	2000 – 3700	200 – 420
Oxide fibers (whisker)	3.90	1200 – 1400	340 – 400

∗ CVD = chemical vapor deposition

ments for electric light bulbs. The high stiffness was achieved by hot working, similar to the processing of metal wires. Real technical success was achieved, however, only after this high-temperature forming process became superfluous upon introduction of the preferred orientation in the process step before thermal decomposition of the polymer fiber to the carbon fiber. These carbon fibers, which can be handled like glass fibers, have completely replaced boron fibers in their function of stiffening polymer composites. They are widely used today in combination with glass fibers in so-called advanced hybrid composites. Boron fibers on wire substrates have also been used to reinforce light metals. But in this field, too, a new competitor has appeared, namely, CVD-SiC fibers, thick monofilaments, much like thin wires. Silicon carbide is oxidation resistant in air up to 1400 °C, but has a density of 3.5. It is heavier than the light metals or boron. It is a superior fiber component in SiC and Si_3N_4 ceramics, especially for toughening. The troublesome CVD technique and the sticking of the monofilaments, both dangerous for the workers, have led to the development of a new competitor, SiC and Si_3N_4 filament yarns, which are fabricated much like carbon fibers by thermal degradation of polymeric precursor fibers, carbosilanes, and silazanes. For oxide composites, refractory oxide fibers are best. Fabrication techniques starting from organometallic precursor fibers as shown above, have so far produced only low-strength fibers, but improvement can be expected. Fabrication from ultrafine crystalline powder produces higher strength fibers [11].

Commercially available today are glass fibers, carbon fibers (HT and HM types), CVD boron, SiC monofilaments, SiC and Si_3N_4 fibers, and Al_2O_3 fibers. The physical and mechanical properties are compiled in Table 3. The specific strengths and tensile moduli are presented in a more schematic way in Fig. 11 [12]. All mechanical properties are determined at room temperature and under standard conditions. Composites with these stiffnesses are brittle, however. The impact resistance of the various fibers is therefore an important criterion for their application. Carbon fibers behave worst in this respect in spite of the other overwhelmingly excellent properties, as shown in Table 4 (see next page). One way to overcome this brittleness is to mix the carbon fibers with glass fibers (Fig. 12). Organic polyaramide fibers (Kevlar from DuPont or Twaron from ENKA) are included in Table 4 and Figure 12 to allow comparison.

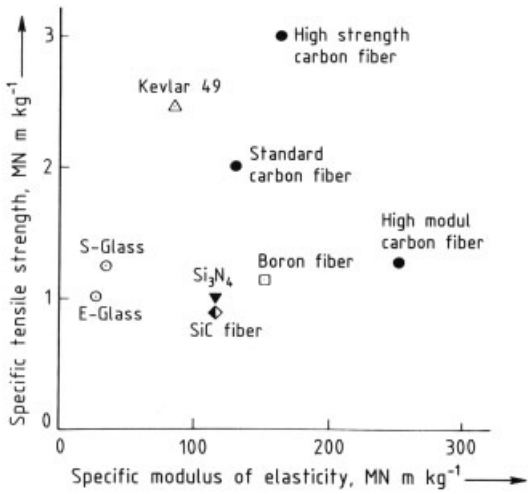

Figure 11. Mechanical room temperature properties of inorganic fibers in a stress – modulus diagram [12]

Table 4. Qualitative comparison of mechanical properties of glass, polyaramid, and carbon fibers

Static load						Dynamic load				
Strength			Modulus			Impact	Fatigue	Ilss	Temperature resistivity	Density
Tensile	Flexural	Compressive	Tensile	Flexural	Compressive					
Glass										
+	○	+	—	—	—	+	—	+	○	—
Aramide										
○	—	—	○	○	○	○	○	○	—	+
Carbon										
+	+	○	+	+	+	—	+	+	+	○

+ = better than average; ○ = average; — = less than average
Ilss = interlaminar shear strength

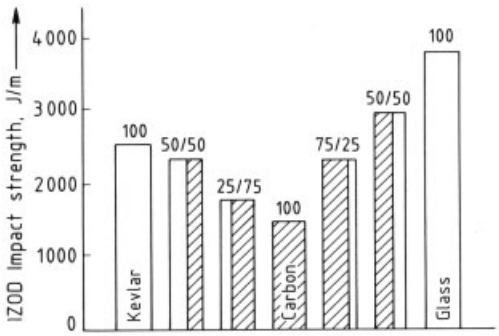

Figure 12. Improvement of impact strength of carbon fiber-reinforced composites by polyaramide (Kevlar) and glass hybridization, respectively

1.5.3. Surface Chemistry, Surface Treatment, and Surface Properties

All commercial inorganic fibers are sized, i.e., coated with a thin organic, usually polymeric, layer to protect the fiber surface and to improve the handling of the brittle fiber material. The most common sizing materials are poly(vinyl alcohol) and uncured epoxy or polyimide resins. Glass fibers are coated by coupling agents, usually silanes, to improve adhesion in the matrix resins, which are usually unsaturated polyesters and, in advanced composites, epoxy resins. Glass fibers have silanol groups on the surface, which can react easily with functional groups in the resin. As such, interfaces within the composites are very sensitive to attack by water, and the glass fibers must be protected by linear organic molecules that have a hydrophobic group on one end of the molecule. The mechanism of protection is indicated in Figure 13 [13].

Carbon fibers are thoroughly wetted by the various polymers if they possess such functional surface groups as acid and basic surface oxides [14]. Also, the adhesion in the composite is explained by chemical interchange with these sur-

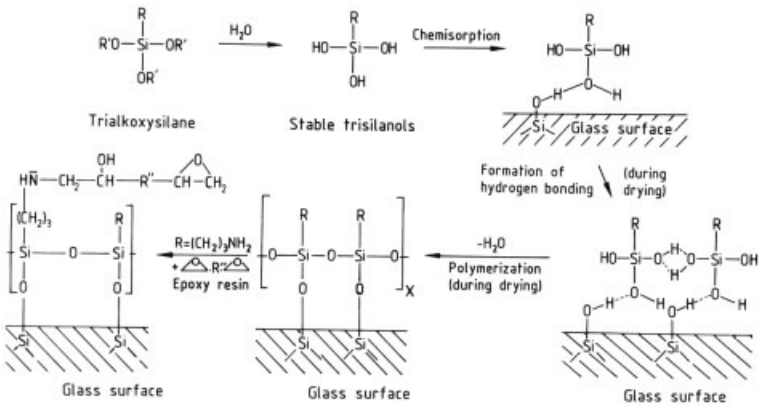

Figure 13. The effect of coupling agents for glass fibers in polymers (schematic) [13]

face groups. The surface oxides are formed on the carbon fibers as the last step during fabrication. All commercial carbon fibers are surface-treated and sized in this way. The process parameters of the surface treatment strongly influence the degree of surface oxidation, and thus the character of the interfaces within the composite. Surface treatment is, however, usually a proprietary process kept secret by the supplier. The suitability of the commercially surface treated and sized fibers must be tested and confirmed by the user.

Adhesion between fiber and matrix plays an important role in respect to the impact resistance of the composite. Dependant on the strain-to-failure of the fiber, impact resistance can be improved or worsened by a high degree of adhesion, which in turn is determined by the degree of surface oxidation [15]; Figure 14 shows some correlations.

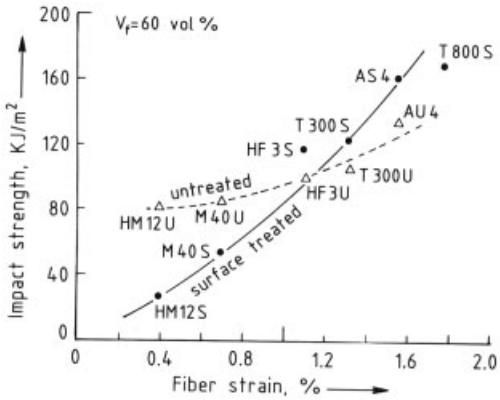

Figure 14. Impact strength of carbon fiber-reinforced unidirectional composites as a function of fiber strain for the untreated and surface-treated state
U: Untreated; S: Surface treated

The chemical compatibility with the matrix is another criteria for inorganic reinforcement fibers. Glass fibers are compatible with polymers, with most oxides, but not with concrete. Carbon fibers are also compatible with polymers, and in addition they are compatible with concrete. They are not compatible with most metals, with the exception of lead and tin. They are incompatible with oxides above 400 °C. Surface coatings as diffusion barriers improve the chemical compatibility. Boron fibers are compatible with polymers, but react with metals. Even in the case of reinforcement of aluminum, boron fibers must be coated by diffusion barriers, thin layers of SiC or B_4C deposited from the vapor as the last step of the fabrication process (Fig. 15) [16, 17]. SiC fibers are compatible with polymers, and generally with most inorganic materials, if 1000 °C is not exceeded. Refractory oxide fibers would be ideally suitable for combination with light metals if properties and economic considerations justify it.

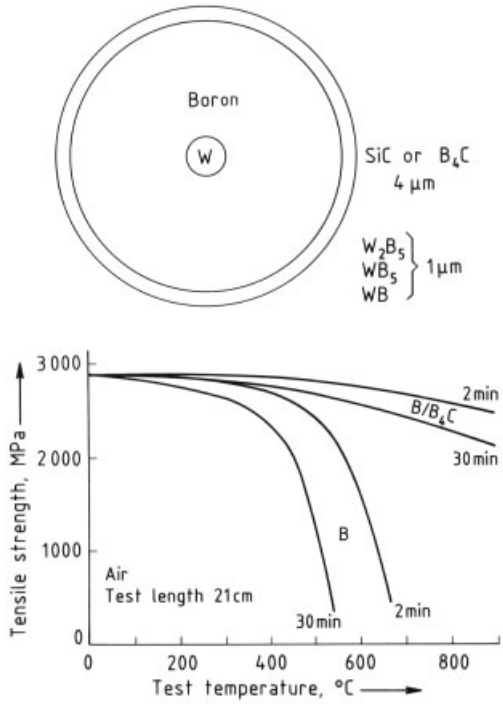

Figure 15. Boron fiber fabrication on tungsten substrate by the CVD process (chemical vapor deposition) [16]

1.5.4. High-Temperature Behavior of Inorganic Fibers

Glass fibers are limited to temperatures below ca. 300 °C. Carbon fibers are not resistant to oxidation above 400 °C, but are thermally stable up to 2000 °C in non-oxidative environments. They can be applied in all high-temperature polymers without restriction. For the highest temperatures, the performance of PAN-based carbon HM-fiber must be considered as exemplary. They can be used in composites with

carbon as matrix without limitation (carbon–carbon composites [18]), but diffusion barriers are needed in the cases of matrices of oxides, metals, and non-oxidic refractory compounds [19]. Some examples are given in Figures 16, 17, and 18 [20]. All other refractory fibers lose their room-temperature strength values if heat-treated above 1200 °C because of crystallization processes (see Figs. 3 and 4). Inorganic fibers resistant to high temperatures are a fertile field for further development.

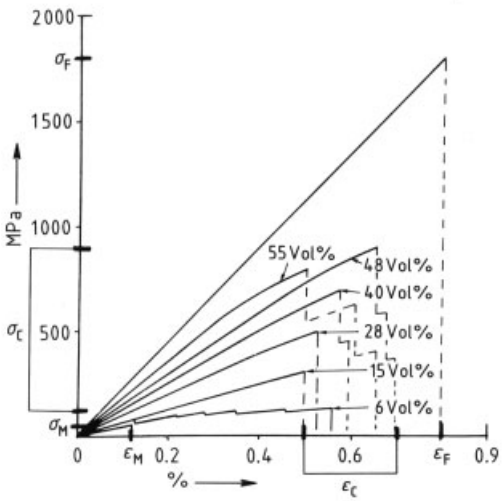

Figure 16. Reinforcement effect by carbon fibers with silicon carbide coatings in quartz glass [20]
σ_F, σ_C, σ_M = stress of fibre, composite, and matrix respectively; ε_F, ε_C, ε_M = strain of fibre, composite, and matrix respectively

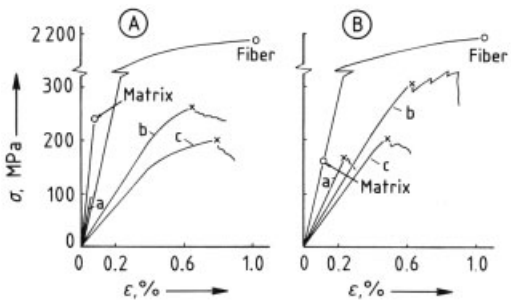

Figure 17. Strengthening and toughening of mullite by silicon carbide-coated carbon fiber
A) Al_2O_3 Matrix; a) 10 %, b) 54 % and c) 32 % coated
B) Mullite – Matrix; a) 28 %; b) 41 % and c) 34 % coated

2. Glass Fibers

2.1. Textile Glass Fibers

Textile glass fibers have gained in importance in spite of the availability of other, newer reinforcing fibers, such as carbon or aramid fibers [21–25]. A large part of this growth is a result of glass fiber-reinforced plastics (GRP). The market for GRP in the Federal Republic of Germany has grown from ca. 10 000 t in 1960 to ca. 200 000 t in 1985.

Glass fibers that can be processed by conventional textile techniques such as weaving are called textile glass fibers, to contrast them with insulation glass fibers. Textile glass fiber is a collective term for fine fibers possessing an approximately uniform, mostly circular cross section made from molten glass. The terminology of textile glass and processing aids is standardized in national and international standards.

Depending on the manufacturing process, textile glass fibers are subdivided into (1) glass filaments and (2) glass staple fibers, which are defined as follows in ISO 6355:

1) A glass filament is a textile glass fiber of practically unlimited length of a given diameter drawn from molten glass.
2) A glass staple fiber is a textile fiber of limited length (spun fiber) and of a given diameter.

Textile glass is used both for textile purposes, e.g., decoration, insulation, filtration, as well as for reinforcement of a matrix to form a composite. Matrices can be plastics (both thermosetting and thermoplastic), bitumen, rubber, cement, gypsum, or other materials. Of all glass filaments ca. 90 % are used to reinforce a matrix.

Textile glass fibers can also be categorized by the type of products and the diameter of the filaments:

1) Textile products, such as textile glass yarns and plied yarns with filament diameters of 5 – 13 µm
2) Plastic products such as mats, rovings, chopped strands, and milled fibers with filament diameters of 9 – 24 µm

The textile products are mostly woven and are then available as fabrics for reinforcing plastics

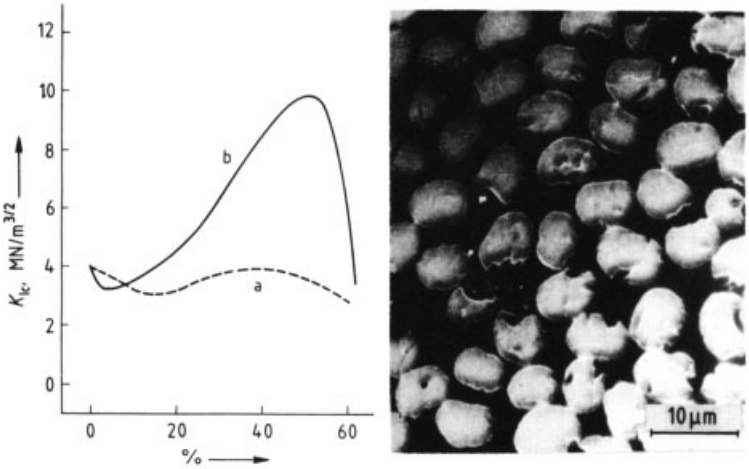

Figure 18. Toughening of alumina by silicon carbide coated carbon fibers a) Uncoated C-Fibers; b) SiC-Coated C-Fibers

and other applications such as thermal insulation and decoration.

Plastic products can be manufactured to parts by hand lay-up, winding, molding, etc. (Chapter 4).

History. The technique of ornamenting glass vessels with glass yarns was used in ancient Egypt around 1600 B.C. and developed further by the Romans. In the Middle Ages, this technique–based on drawing heated rods into fine filaments–was further refined in Venice and Murano.

The first indications of the use of glass yarns in fabrics are found in the work of the physicist FERCHAULT DE REAUMUR in 1713. The first known factory for the manufacture of glass yarns was founded in 1866 in Vienna by JULES DE BRUNFAUT. The inventor of the bushing, however, was LOUIS SCHWABE, who first produced glass yarns by drawing the melt through fine holes. In 1908, a rather coarse, endless glass fiber was demonstrated in Hamburg by W. V. PACZINSKY. True industrial production of textile glass fibers was realized in the United States in 1930 and in Germany in 1939.

2.1.1. Chemical Composition and Properties

Chemical Composition. The chemical composition of glass fibers essentially determines the properties and the applications. Table 5 shows the known types of textile glass fibers and their compositions.

Table 5. Chemical composition of textile glass (nominal values, wt %)

Component	Types of glass								
	Multi-purpose	Acid resistant			Alkali resistant	High strength		Good dielectric properties	
	E	A	C	E-CR	[31]	R	S	D	Quartz
SiO_2	53–55	70–72	60–65	58	60	60	60–65	73–74	100
Al_2O_3	14–15.5	0–2.5	2–6	12–13	0.7	25	20–25		
CaO	20–25	5–9	14	21	5	6		0.5–0.6	
MgO	20–25	1–4	1–3	4.5		9	10		
B_2O_3	6.5–9	0–0.5	2–7	<0.1			0–1.2	22–23	
F	0–0.7			<0.15					
Na_2O	≤1	12–15	8–10	0.6	14		0–1.1	1.3	
ZrO_2	≤1				18				
K_2O	≤1	1		0.4				1.5	
Fe_2O_3	≤1			0.3	0.1				
TiO_2	≤1			2.1					

The following types are recognized:

E glass. An alumino-borosilicate glass with a mass fraction of alkali ≤ 1 %, which is generally used in glass-reinforced plastics and in electrical applications

A glass. An alkali-lime glass with little or no addition of boron trioxide and a mass fraction of alkali ≥ 1 %, for special applications

E-CR glass. An alumino-lime silicate glass with a mass fraction of alkali ≤ 1 %, especially suited for the reinforcement of plastics because of acid resistance

C glass. An alkali-lime glass with high boron trioxide content, with good chemical resistance for special applications

D glass. A borosilicate glass, a specialty glass with high dielectric properties

R glass. An aluminosilicate glass without added calcium and magnesium oxides, glass that meets high mechanical requirements

S glass. An aluminosilicate glass without added calcium oxide but with a mass fraction of magnesium oxide of ca. 10 %

The most widely used are E-glass fibers, which are encountered in all areas of industry and everyday life, especially in combination with plastics (Chapter 4). In glass staple fibers, C glass is used primarily [29].

More recently, acid-resistant E-CR-glass fibers have been used in special projects such as pipes for flue-gas desulfurization plants. Alkali-resistant glass fibers are primarily intended for cement reinforcement.

High-strength textile glass fibers such as R- and S-glass fibers are preferred in applications where high tensile strength is needed, e.g., in containers under high internal pressure.

The raw materials for the components are largely natural minerals. They are subject to stringent quality control when received, since the raw materials must be within narrow, well-defined limits to produce textile glass fibers having uniform properties.

Properties. Textile glass fibers are distinguished by particularly high tensile strengths. Measurements on filaments drawn directly from the bushing yield values of 3400 MPa for E glass and 4400 MPa for R glass (Table 6).

The modulus of elasticity, which is 1/3 that of steel and quite adequate for many applications, and the density of 2.6 g/cm^3, which is high compared to plastics, are not sufficient for large-surface, light-weight parts (airplane construction). For such parts, fibers of lower density and higher modulus of elasticity (aramid and carbon fibers) were developed.

Textile glass fibers have a linear stress – strain curve up to the break, so that the break elongations ε_B can be calculated from the modulus of elasticity E and the tensile strengths σ_B (Fig. 19, see next page). Good resistance to weathering and heat, nonflammability, good dielectric properties, low thermal expansion, and, depending on the type of textile glass, good resistance to corrosion, are additional characteristics of textile glass fibers. The thermal conductivity of textile glass products is determined by the density of fiber packing. The properties of textile glass fibers to a great extent determine the properties of the composites (→Composite Materials).

Table 6. Properties of textile glass (nominal values)

Property	Types of glass								
	Multi-purpose	Acid resistant			Alkali resistant	High strength		Good dielectric properties	
	E	A	C	E-CR	[31]	R	S	D	Quartz
Density, g/cm^3									
Bulk glass	2.61	2.46	2.46			2.55		2.16	2.2
Fiber glass	2.59		2.45	2.70	2.68	2.53	2.48	2.14	
Modulus of elasticity, MPa	73 000	74 000	71 000	72 000	76 000	86 000	85 500	55 000	6200 – 7200
Tensile strength, MPa									
Freshly drawn fiber	3 400	3 100	3 100	3 330	3 700	4 400	4 590	2 500	400 – 500
Fiber in composite	2 500		2 400			3 750		1 550	
Dielectric constant at 1 MHz	5.8 – 6.4		6.8	7		5.6 – 6.2	4.9 – 5.3	3.85	3.78

2.1.2. Manufacture

2.1.2.1. Glass Filaments

Glass filaments are produced continuously by the bushing fiberizing process. There are two commercial processes for the manufacture of glass filaments: the marble melting process and the direct melt process.

In the older indirect *marble melting process,* the production of glass and the glass spinning process are separate. E glass is initially produced in the form of glass marbles (15 – 20 mm diameter), which are then re-melted in a second phase in the glass forming process (two-stage process). This process is still in use today for making fine filament yarns.

In the *direct melt process,* molten glass flows directly from the melting furnace into a system of heated feeders (channels) in which the bushings are placed. In this one-step process, there is no separation between glass manufacture and glass forming (Fig. 20). Temperatures in the pro-

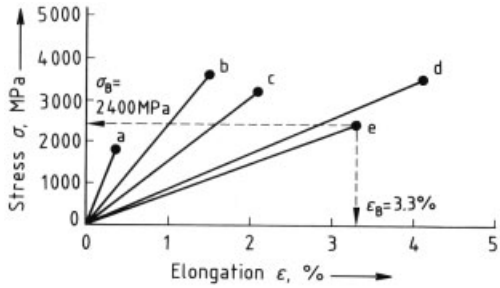

Figure 19. Stress – strain curves of five fibers
a) Carbon fiber, high modulus; b) Carbon fiber, high strength; c) Aramid fiber, high modulus; d) R-glass fiber; e) E-glass fiber σ_B = tensile strength; ε_B = break elongation

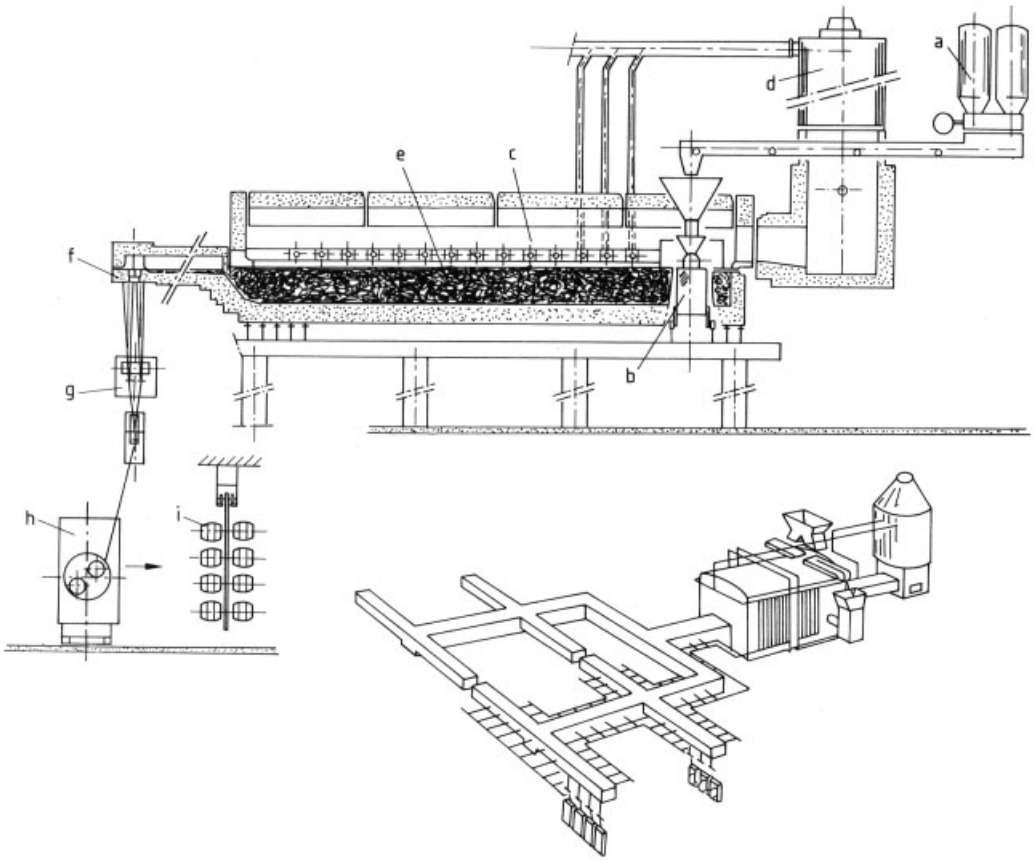

Figure 20. Direct melting
a) Mixing silos; b) Furnace feed; c) Furnace; d) Heat exchanger; e) Melt; f) Bushing; g) Sizing applicator; h) Cake winder; i) Spun cake

duction of E glass are 1310 – 1390 °C in the furnace and 1225 – 1260 °C in the feeders [24].

In principle, the fiber forming process involves attenuating glass filaments continuously from a Pt – Rh bushing by applying a mechanic drawing force at offtake speeds of up to 80 m/s (Fig. 21). Bushings may have 400, 800, 1200, 1600, or more holes of 1 – 2.5 mm diameter. A single filament is drawn from each hole. The filaments are collected into a bundle (strand) and wound as a spun cake or directly processed into rovings, continuous strand mats, or chopped strands. The deformation during drawing is high: for example, with bushing-tip diameters of 2 mm and a filament diameter of 10 µm. A typical ratio of cross sections is 40 000 : 1 (Fig. 22).

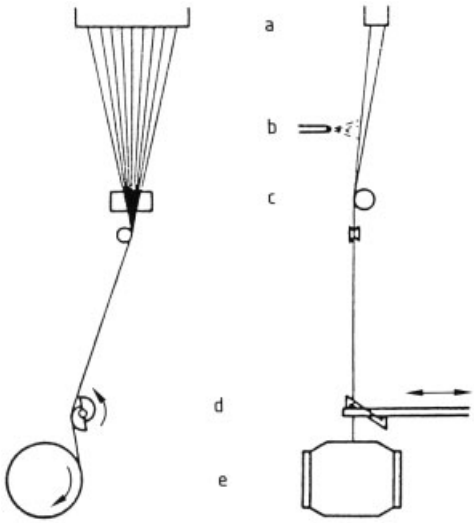

Figure 21. Spin-draw process (schematic)
a) Bushing; b) Water spray; c) Sizing application; d) Strand guide; e) Spun cake

Formation of the filaments involves two basic processes:

1) efflux of molten glass from the bushing tips under the condition that

$$F_p + F_g > F_\eta + F_\sigma$$

2) drawing of the meniscus at the bushing tip to a continuous filament by a force under the condition that

$$F_d + F_g > F_a + F_\sigma$$

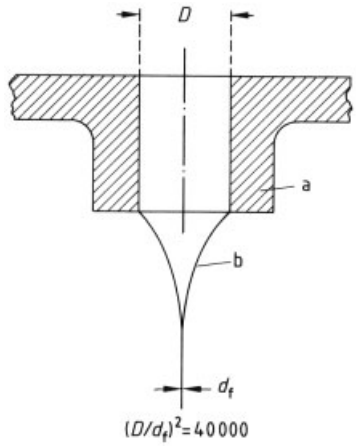

Figure 22. Bushing
a) Bushing nipple; b) Drawing bulge D = hole diameter; d_f = diameter of glass filament

where
F_p = force due to hydrostatic pressure
F_g = force of gravity
F_σ = force due to surface tension
F_η = force due to viscosity
F_d = drawing force
F_a = sum of frictional forces

The viscosity of the melt must be increased sharply by cooling with air and water to enable a filament to be formed in the shortest possible time after the melt leaves the bushing tip.

An aqueous solution, called size or sizing, is applied to the cooled filaments; this solution consists basically of film formers, lubricants, and, if necessary, coupling agents. After the forming packages are dry, the sizing (0.5 – 1.5 wt %) has two functions:

1) Reducing the glass – glass coefficient of friction from ca. 1.0 to 0.2 to make further processing of the yarn possible. Due to the high coefficient of friction, textile glass strands without size cannot be processed further since they cannot be pulled off the cakes.
2) Imparting properties that are prerequisite for further application of the textile glass products as plastic reinforcement (plastic products) or as textile products with optimum weaving and desizing properties.

2.1.2.2. Glass Staple Fibers

The principles discussed with regard to the drawing of glass filaments also apply to glass staple fibers, but the manufacturing processes themselves are different.

In the *bushing blowing process,* the molten effluent is fibrillated by streams of gas. Adhesive sizing is added, and the fibers are collected on a rotating perforated drum or on a suction belt (nonwovens) under reduced pressure.

In the *drum drawing process,* fibers are drawn either from bushings or from rods by a rotating drum and scraped off. During the drawing process, sizing is applied to the fibers for later cementing. The fibers broken by scraping are swirled in a funnel.

In both processes the fibers are finally consolidated in a finishing tube and drawn off as a sliver. The sliver can be converted directly into fabrics, yarns, or plied yarns.

2.1.3. Textile Glass Products: Processing and Application

2.1.3.1. Product Designation

Glass filament and glass staple yarns have a standard form of designation.

Glass Filament Yarns (ISO 2078 and ISO 3598) EC9-34 Z28 – sizing x

E = E glass
C = continuous
9 = filament diameter in µm
34 = tex ∗
Z28 = Z-twist, 28 turns per meter

Glass Staple Yarns(ISO 2078 and ISO 3598) CD11-200 Z140

C = C glass
D = discontinuous
11 = nominal fiber diameter in µm
200 = tex ∗
Z140 = Z-twist, 140 turns per meter

∗ One tex is equivalent to 1 g/1000 m. For glass filament products the relationship between the number of holes in the bushing N, the filament diameter d_f, and the tex number is:

$$\text{tex} = 0.002 d_f^2 \times N$$

where d_f is in µm. The density of glass ϱ_G is taken to be 2.5 g/cm^3.

The other textile glass products are designated in a similar manner.

Typical properties of products used in textiles and plastics are summarized in Tables 7 and 8.

2.1.3.2. Correlation Between Processing and Application

Textile glass fibers can be processed as follows:

1) Textile processes such as weaving, knitting, and braiding, suitable for textile glass yarns, plied yarns, and rovings
2) Plastic processes for reinforcing thermosetting and thermoplastic resins, suitable for textile glass fabrics, layers, knits, mats, rovings, milled fibers, tissues, and chopped strands
3) Processes for reinforcing other matrix materials, e.g., bitumen, paper, rubber, gypsum, and plaster, suitable for textile glass fabrics, yarns, surface mats, rovings, and chopped strands

The processes used in plastic reinforcement are summarized in Table 9 (see next page).

Table 7. The tex number of textile glass yarns as a function of filament diameter and number of filaments in a strand

Filament diameter			Number of filaments per strand					
µm	Nominal µm	U.S. System	50	100	200	400	800	1600
5.2	5	D	2.8	5.5	11			
6.4	6	DE			(17)	34	68	136
7.4	7	E		11	22			
9.2	9	G			34	68	136	272
11.3	11	J			51	102	204	408
13.0	13	K				136	272	544

Table 8. The tex number of glass yarns for plastic reinforcement as a function of filament diameter and number of filaments in a strand

µm	Nominal µm	U.S. System	Number of filaments per strand							
			50 ∗	100 ∗	200 ∗	400	800	1600	2000	4000
10.0	10	H	10	20	40	80	160	320		
11.3	11	J	12.5	25	50					
14.2	14	L				160	320	640		
17.3	17	N					480	960	1200	2400
19.4	19	OP					600	1200		
24.5	24	ST							2400	4800

∗ Strands separated during spinning

Table 9. Processing of glass-reinforced plastics (GRP)

Process	Textile glass products
	Glass-reinforced thermosetting resins:
Manual processes	
Hand lay-up	chopped strand mat, woven fabric, surface mat
Fiber/resin spray-up	roving
Intermediate processes	
Injection process	surface mat, continuous mat
Vacuum injection process	surface mat, continuous mat
Wet molding	surface mat, continuous mat, woven fabric
Industrial processes	
Processing of semi-finished products	(compression molding, injection molding)
Prepregs	woven fabric
SMC ∗	roving
Molding compounds	chopped strands
Rotation process	roving, woven fabric
Reinforcement of polyurethane	milled fiber, continuous mat
Filament winding	roving, woven fabric, chopped strand mat
Continuous impregnation	continuous mat, chopped strand mat, woven fabric
Pultrusion	roving, woven fabric, continuous mat, chopped strand mat
	Glass-reinforced thermoplastic resins:
Injection molding of glass fiber-reinforced granulate	milled fiber, roving, chopped strands
Compression molding of glass mat-reinforced thermoplastic plates	continuous mat, roving

∗ SMC = sheet molding compounds

2.1.3.3. Description of Textile Glass Products

There are separate standards for special products:

textile glass rovings	ISO 2797
textile glass mats	ISO 2559
textile glass fabrics	ISO 2113
glass filament yarns	ISO 2078 and ISO 3598
glass staple fiber yarns and rovings	ISO 2078 and ISO 3598

The entire line of products is presented in Figure 23 (see next page).

Textile glass yarns usually have a protective twist and are sold with textile or plastic sizes. Textile sizes based on starch/fats are especially suitable for high-speed looms, e.g., air looms. For certain applications, e.g., circuit boards for printed circuits, the textile finish applied to the fiber must be removed and replaced by an adhesive finish (coupling agent) via a special treatment given to the fabric.

Textile glass fabrics come in different weaves, depending on the application.

Textile glass rovings with untwisted, approximately parallel filaments or strands are offered with inside or outside take-off. Textile glass rovings are either assembled into a larger strand comprising several strands or produced by a direct process below the bushing. In the second process roving strands of various tex numbers are obtained by changing the number of holes in the bushing or by changing the filament diameters. Spun rovings of textile glass are suitable for the pultrusion process because of their good mechanical properties in the transverse direction.

Textile glass mats are made from chopped strands (chopped strand mats) or of strands laid up directly below the bushing (continuous strand mats). Consolidation in both cases is accomplished by chemical binders. Chopped strand mats are used primarily for hand lamination; continuous strand mats for all types of injection molding, compression molding, and continuous impregnation (Table 5).

Chopped strands with lengths of 3, 4.5, 6, or 12 mm have been found useful in the production of molding compounds and as reinforcement of thermoplastics. As shown in Figure 5, chopped strands are either cut directly below the bushing

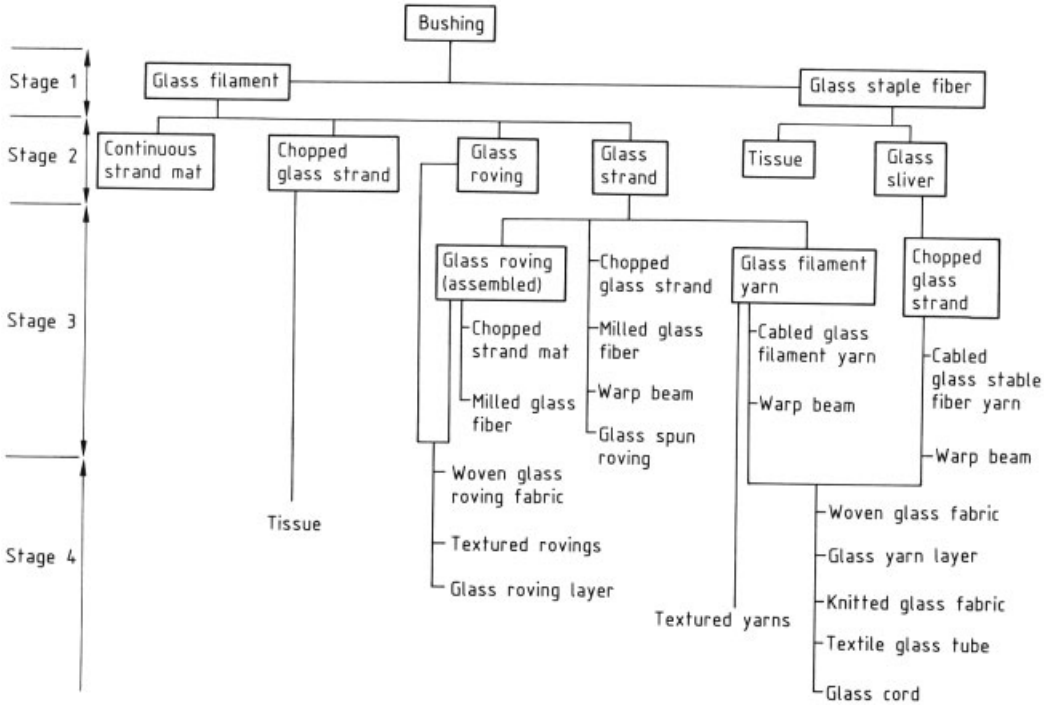

Figure 23. Textile glass products
Stages 1 – 3 are in the hands of the textile glass producers, while Stage 4 is in the hands of the processors, e.g., weavers

or produced from intermediate products, e.g., textile glass roving or glass strands.

Milled textile glass fibers with fiber lengths of 0.2 mm are suitable as reinforcement of thermoplastics and polyurethane foams. In addition, they are used in the production of primers, fire-protection coatings, and the like.

Tissues of glass filaments or glass staple fibers are produced both by dry and wet processing. In both cases the fibers are chemically bonded.

Because of their bulky nature, *textured rovings or yarns* have turned out to be good replacement products for asbestos, e.g., in the manufacture of clutch facings. Textured yarns are also used in decorative fabrics.

*Layers*are planar structures of bonded, nonwoven textile glass yarns with crisscross or parallel yarns. Layers are primarily used for reinforcing paper or bitumen or in the preparation of prepregs (preimpregnated sheets).

*Textile glass knit goods*are more elastic than woven fabrics of textile glass. One area of application is medical bandages of pre-impregnated textile glass knits, which take the place of plaster casts.

Warp beams are used in the preparation of warps for weaving and consist of a multiplicity of yarns wound in parallel. Beams of strands with plastic sizing are used for unidirectional reinforcement of plastics for the bumper systems of motor vehicles.

2.1.3.4. Applications

As seen in Table 10 (on next page) there is a distinct trend toward the use of GRP in motor vehicles, both passenger cars and trucks. Glass fiber reinforced thermosetting resins are used in passenger cars primarily for body parts (hoods, tail gates) and for structural elements (bumpers, bumper supports). Glass fiber reinforced thermoplastics are used for operating parts within the engine compartment, noise suppressors below the engine, hub caps, etc.

Textile glass fabrics for *epoxy resin prepregs,* the basic material of circuit boards, is not considered to be GRP but rather is considered among

Table 10. Percentage of textile glass fibers used to reinforce resins by area of application in the Federal Republic of Germany, 1973 and 1985 [32]

Area	Thermosetting resins (except circuit boards)		Thermoplastic resins	
	1973	1985	1973	1985
Construction	24	16	0	1
Industry and agriculture	32	24	23	12
Transport	9	25	25	55
Consumer goods	6	4	14	6
Sport and leisure articles	6	3	0	0
Electronic industry	17	25	36	23
Other	6	3	2	3

other applications. However, it is the most important such other application, representing one-third of the amount used for other applications.

The following other applications are of importance:

1) Thermal insulation and heat protection (an asbestos replacement) with woven glass fabrics of glass filament and glass staple fiber yarns for primary insulation jackets, heat shielding curtains, fire blankets, pipe insulation, etc.
2) Wall coverings of staple fiber yarns
3) Plaster reinforcement of textile glass mesh fabrics and textile glass nonwovens
4) Reinforcement of grinding wheels with woven glass fabrics
5) Filter materials for high-temperature filtration with woven glass fabrics and nonwovens
6) Bitumen reinforcement with blend fabrics of glass staple fiber yarns, glass filament yarns, and textile glass nonwovens
7) Gaskets with glass cord and glass twine (an asbestos replacement)
8) Decorative fabrics of textured glass filament yarns and glass staple fiber yarns
9) Expansion joints of impregnated woven glass fabrics
10) Cable insulation and reinforcement with textile glass yarns
11) Reinforcement of adhesive tapes with textile glass yarns

2.1.4. Product Standards

The textile glass standards described in Section 4 contain supplementary basis sheets with specifications. Beyond that, there are customer acceptance standards, which list minimum requirements. In order for the values determined to be comparable and reproducible, standardized test procedures are used in which all testing conditions are specified precisely.

Processors of textile glass products should store the products in dry rooms at temperatures below ca. 25 °C at a relative humidity level between 55 and 65 %. Exposure to direct sunlight is to be avoided. Prior to processing, the material should be equilibrated under the temperature-humidity conditions present in processing.

2.1.5. Toxicology and Occupational Health

Textile glass products do not represent a health or environmental hazard, mainly because of the following properties:

1) Textile glass fibers do not contain poisonous and/or hazardous components.
2) Fiber structure, because of the chemical composition, is affected by body fluids in such a way as to reduce the carcinogenic potential.
3) Textile glass fibers do not split longitudinally, which largely prevents their penetration into the alveoli of the lungs.

Extensive systematic examinations of many thousands of people working for decades in the production and processing of textile glass products have shown that continuing exposure to textile glass fibers does not produce increased morbidity or mortality rates. Likewise, innumerable animal experiments (inhalation and ingestion of textile glass fibers as well as skin contact) have revealed no fibrinogenic or carcinogenic changes in test animals, a contrast to asbestos.

2.1.6. Economic Aspects

The growth of textile glass fiber production is illustrated by comparing the total figure of 1 135 000 for 1985 (Table 11) with that for 1973, 865 000 t. A 56 % increase was achieved.

Table 11. Production of textile glass fibers by final use and geographic region, 1985 [33]

Area	Production, t	
	Uses in glass fiber reinforced resins	Other uses, including fabrics for circuit boards
United States, Mexico, Canada	520 000	80 000
Western Europe	240 000	65 000
Japan and Korea	165 000	35 000
Eastern Europe, Former States of USSR, China	155 000	40 000
Others	50 000	
Total	*1 130 00*	*220 000*

2.1.7. Producers

The principal producers of textile glass fibers are listed in Table 12.

Table 12. Companies producing textile glass fibers and the location of their plants

Company	Location of plants
Asahi Fiberglass	Japan
Certain Teed Corp.	United States
Fiberglass Ltd.	Great Britain
Glaswerk Schuller *	Federal Republic of Germany
Manville Corp.	United States
Nippon Electric	Japan
Nippon Glassfiber	Japan
Nitto Boseki	Japan
Owens Corning Fiberglass	United States, Belgium, France
Pittsburgh Plate Glass (PPG)	United States
Scandinavian Glassfiber	Sweden
Silenka	The Netherlands
Skloplast	Former Czechoslovakia
VEB Glasseidenwerk Oschatz	Former German Democratic Republic
Vetrotex	Federal Republic of Germany ** (Gevetex), France, Italy, Spain
Vitrofil	Italy

* Glaswerk Schuller specializes in staple fibers.
** The Vetrotex subsidiary Gevetex operates the plant in Germany.

2.2. Glass Fibers for Insulation

Glass fibers for insulation are fibers of vitreous structure produced from silicate melts by mechanical or pneumatic processes. For insulation, nontextile fibers are used exclusively. Such fibers are also called *glass wool,* slag wool, *rock wool,* or mineral wool.

While the manufacture of glass filaments goes back to antiquity, it was not until 1840 that the production of nontextile man-made mineral fibers (MMMF) was demonstrated in Wales [34]. This was an effort to utilize blast furnace slag and at the same time to find a replacement for natural asbestos. In the United States, rock wool production is documented to have taken place in 1893 [35]. An early patent for glass wadding by the centrifugal process was granted in 1912, but industrial exploitation of the process did not occur until 1930 [36]; the blowing process for glass wool, still in use today, was patented in 1933 [37].

The hope to replace asbestos could not be realized. Instead, these products found their own areas of application. Synthetic resin binders were developed around 1950, which made possible the production of dimensionally stable products and assured a uniform quality for insulation.

2.2.1. Physical Properties

Fibrous insulating materials of man-made mineral fibers are a highly disperse system of finely divided fibrous glass with a specific surface of $0.1-0.2\,m^2/g$. The fibers are oriented predominantly in the horizontal direction. For this reason, the material is isotropic in two dimensions but not in the third. It is an open-pore system with little resistance to flow. This is the basis for its thermal and acoustic properties.

The fiber diameters are distributed over a broad spectrum. The logarithm of the fiber diameter approximately follows a normal distribution (Fig. 24, see next page). The minimum fiber diameter is process dependent and is between 0.18 and 0.30 µm [38]. However, the properties of the individual fibers are less important than those of the final insulating material, i.e., the fibers and binder as a whole.

Table 13 (see page after next) lists the limiting temperature in use which is affected by the chemical composition, the fiber diameter, the amount and type of binder as well as the bulk density and the actual conditions of installation.

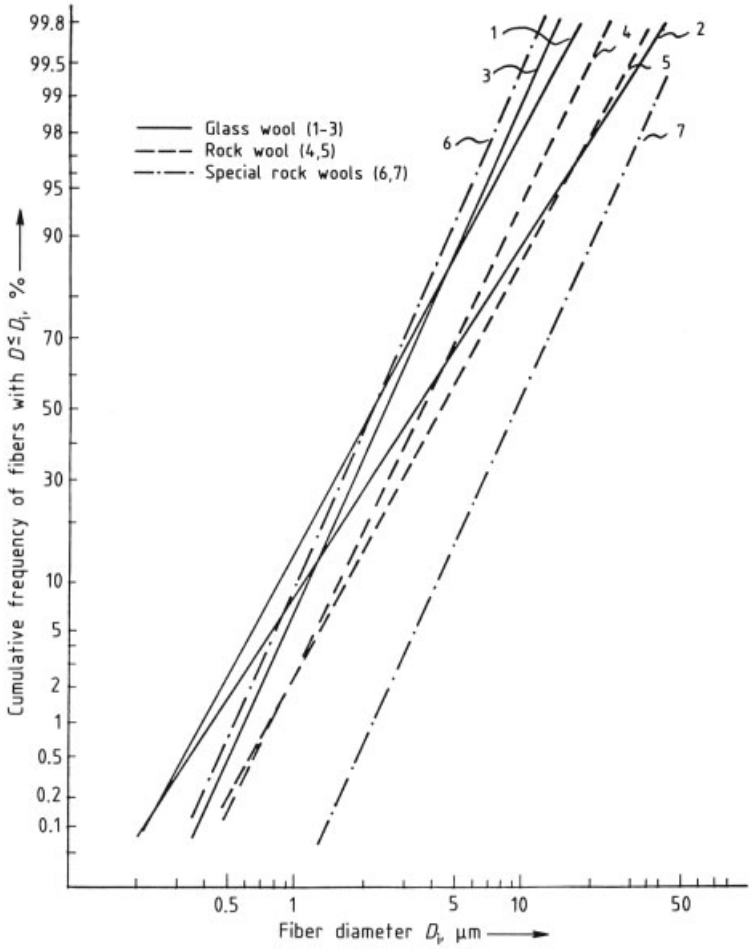

Figure 24. Distribution of fiber diameters in glass and rock wools

The most important property of insulation is the *thermal conductivity,* which is shown in Figure 25 for 14 mineral wools. The thermal conductivity of a particular fibrous material is determined by its bulk density ϱ_B and its average fiber diameter $\overline{D}$. The mechanical properties are primarily determined by bulk density and binder content; in any case, however, fibrous insulating materials cannot be used as load-bearing structural units.

2.2.2. Chemical Properties

Mineral wool insulation is inert to organic solvents, even at elevated temperatures. Extractables are mineral oils, emulsifiers, silicones, and low molecular mass components of binders.

Water and aqueous solutions attack the glass surface, the attack magnified by the large fiber surface. Attack by water also loosens the fiber–resin bond. Calciferous products (slag wool, rock wool) are soluble in acids and are attacked by acid flue gases.

2.2.3. Raw Materials and Chemical Composition

Glass wool (wadding, fibers) are produced from mixtures of minerals and industrial chemicals that produce the required final composition. In

Table 13. Chemical composition, fiberization process, and limiting temperature in use of glass fibers for insulation

	Glass wool			Rock wool				Slag wool	
	1	2	3	4	5	6	7	8	9
Composition, wt %									
SiO_2	57.4	64.6	53.9	52.7	46.3	47.6	46	46.8	40.2
Al_2O_3	3.6	2.9	3.6	8.0	13.3	10.2	15	10.2	15.6
Fe_2O_3	0.35	0.35	0.35	1.8	5.6				
							6.5 *		0.9 *
FeO	0.10	0.10	0.09	1.4	6.3	7.3		1.0	
MnO		0.05			0.2			0.3	0.3
TiO_2		0.1		0.6	2.9	1.9	1.5		0.7
ZrO_2	4.0								
MgO	4.0	3.1	4.6	4.0	9.7	12.0	12	2.3	4.35
CaO	8.6	6.6	23.4	20.9	10.9	14.8	16	37.7	34.9
BaO	1.8	2.4							
Na_2O	11.2	14.0	9.4	10.3	2.6	1.6	2.5	0.9	
K_2O	1.7	1.1	1.0	0.6	1.3	1.6		0.6	1.5
B_2O_3	5.0	5.9	3.0	0					
F	2.3	0	0	0					
Fiberization process **	C	C	A	A	A	B	B	D	D
Limiting temperature in use, °C	570	550	700	700	750	750	750	730	700

* Fe_2O_3 and FeO.
** A) Owens process; B) Rock wool process; C) Rotary process; D) Centrifugal blowing process. These four processes are illustrated in parts A, B, C, and D, respectively, of Figure 26.

addition to quartz sand, the minerals used include limestone, dolomite, phonolite, feldspar, nepheline-syenite, kernite, colemanite, ulexite, and heavy spar (barite). The industrial chemicals include sodium carbonate, sodium sulfate, and barium carbonate. The use of recycling glass has gained considerable importance.

In contrast, rock wool and slag wool are, as a rule, prepared from a single mineral raw material of suitable composition, such as basalt, rhyolite, or diabase. To improve meltability, limestone is added as required. Table 1 shows examples of chemical compositions.

The mineral fiber is the predominant component of the insulating materials:

90 – 99.8 %	vitreous mineral fibers, average fiber diameter 3 – 6 µm
10 % max.	phenol – formaldehyde resins, frequently modified with urea or melamine
1 % max.	aliphatic mineral oil
0.2 % max.	emulsifiers, preferably nonionic polyethoxylenes
0.5 % max.	polymethylsiloxanol

2.2.4. Production

Melting. Two types of melting units are in use for the production of silicate melt: glass melting tank furnaces and cupola furnaces.

The *glass melting tank furnaces* are similar in construction to those used in other areas of glass manufacture [39]. Preferred are cross-fired recuperative furnaces and U-shaped flame furnaces, but all-electric melters are also used if energy costs are right. Maximum productivity is 300 t/d, and up to 3.5 $tm^{-2}\ d^{-1}$ can be achieved. *Cupola furnaces*are used to produce molten rock. They require coarse feed and yield a less uniform melt. The traditional fuel is coke but oil- or gas-fired units are also used. Furnace productivity rarely exceeds 150 t/d.

Fiber Production. Numerous processes and process variants have been developed for fiberizing silicate melt. Of these, only four have withstood the test of time (Fig. 26).

Owens Process. In the Owens blowing process (Fig. 26A), the melt, in the form of a thick primary strand, emerges from holes at the bottom of a platinum crucible (the bushing) and is subsequently entrained by a downward-directed stream of gas from the blow-jet. The blowing medium may be steam or compressed air. The primary strand is drawn to form fibers with an average diameter of 3 – 6 µm (Fig. 27). The length

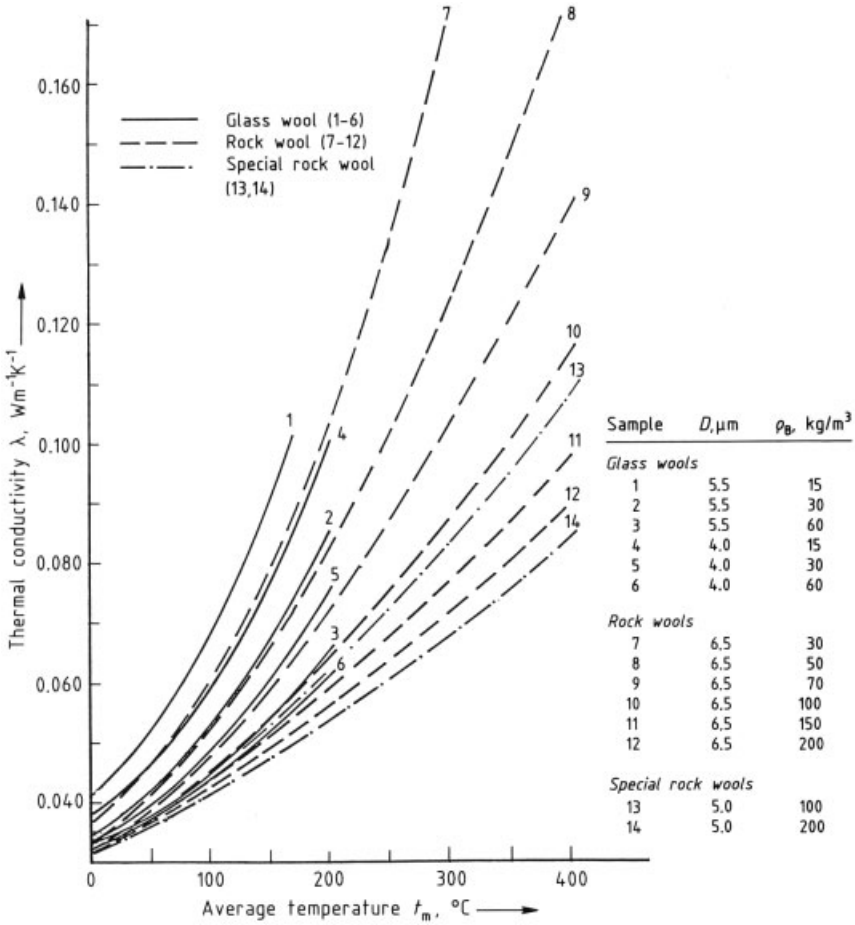

Figure 25. Thermal conductivity of glass and rock wools as a function of temperature

of the fibers is about 3 – 10 cm. The product contains slugs. Glass and rock melts are processed in this way.

Rock Wool Process. This centrifugal cascading process (Fig. 26B) is also a one-step fiberizing process. A single stream of melt hits the face of a horizontal rotor which distributes it to two or three additional rotors; from which it is thrown in the form of fine fibers by centrifugal force. The fiber length is in the range 3 – 10 cm. The product contains slugs. Rock melts are used in this process.

Rotary Process. The rotary process (Fig. 26C) is a two-step process. The melt, in the form of a thick stream, passes through the hollow shaft of a rotating centrifuge into a rotating throw-basket of platinum or special steel, which distributes the melt evenly on the inside of a rotating spinner. Primary fibers issuing from holes in the face of this spinner, are entrained by the hot combustion gases of a circular burner and drawn to long fibers with diameters of 3 – 6 µm. The product is free of slugs (Fig. 28), but the process can be used only for "soft, long glasses," not for rock melts. The process is highly variable with respect to process parameters but is easy to control.

Other Processes. The centrifugal blowing process, shown in Figure 26D, is of little significance in the production of slag wool. The rod drawing process and a variant of the blowing process serve to produce textile glass fibers for glass mats. Special glass microfibers are made by the flameattenuation process. However, these

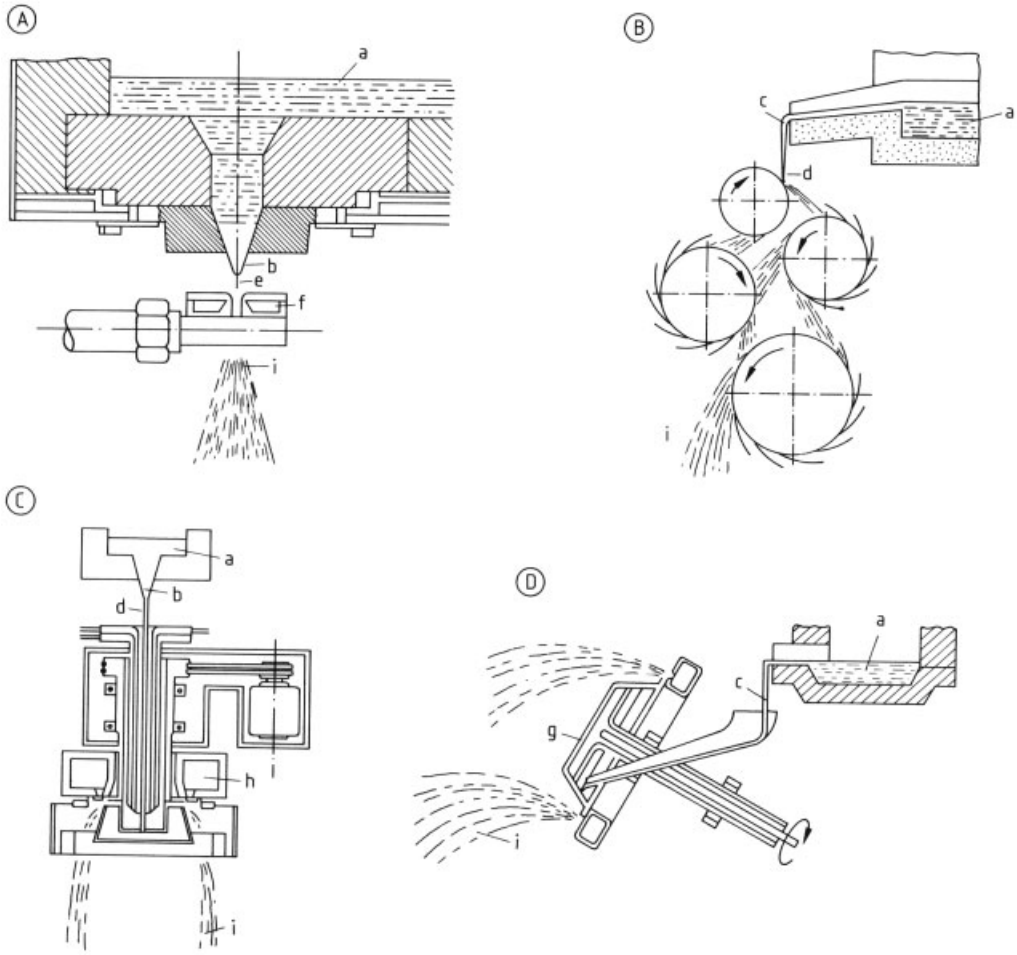

Figure 26. Fiberization processes
A) Owens process [40]; B) Rock wool process [41]; C) Rotary process [41]; D) Centrifugal blowing process [41]
a) Melt; b) Bushing; c) Spout; d) Melt stream; e) Primary fiber; f) Jet; g) Centrifugal spinner; h) Burner; i) Fiber stream

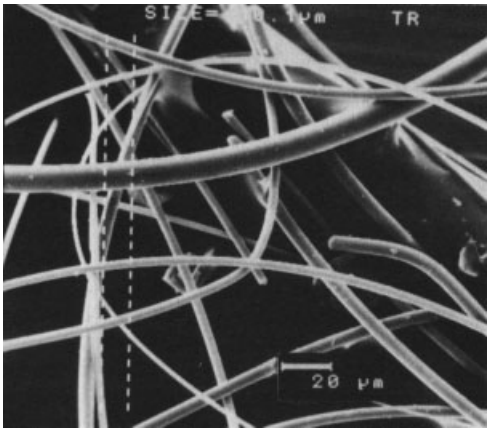

Figure 27. Scanning electron microscopy (SEM) photograph of rock wool produced by the Owens process

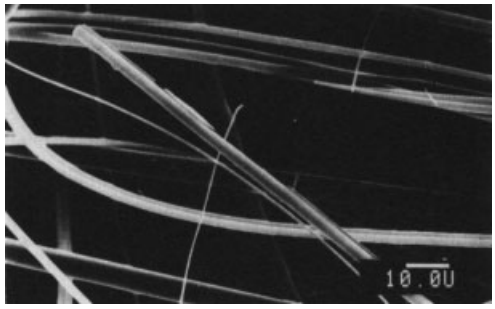

Figure 28. SEM photograph of glass wool produced by the rotary (TEL) process

processes have limited productivity and consume large amounts of energy for fiberization,

and therefore they are not used to produce fibrous insulating materials.

Product Forming. The fibers are formed into a continuous sheet by pneumatic processes, while an aqueous solution of binder is sprayed onto the freely falling individual fibers. The binders are slightly pre-condensed phenol–formaldehyde resins. They are often modified by aminoplastic resins (urea, melamine, melamine ether, dicyanodiamide) to suppress the tendency of phenolics to exothermally autoxidize. In the industry these aminoplastic resins are called antipunking agents.

Figure 29 shows schematically the overall manufacturing process. The binder resins are cured at temperatures around 180–250 °C in a continuous flow dryer heated by circulated air. For shaped articles or pipe insulation, for example, special ovens that operate on a similar principle are used. On the other hand, presses with purely convective heat transfer are rarely used since they require long cycle times.

Mineral-fiber insulating materials can be cut to any desired dimension with knives, saws, shears, and milling cutters.

2.2.5. Environmental Protection

Characteristic emissions arise in the various stages of producing insulating material from mineral fibers. Emissions from the melting units are similar to those of other glass melting installations. Electric precipitators have proved useful in dust removal but greatly strain the fluegas conditioning system. They can be combined with the usual wet and semi-wet installations for sorption of fluorine and desulfurization.

The binder resins should contain a minimum of monomers. Lubricants, mineral oils, hydrophobic agents, and other organic additives must be selected with an eye to their volatility. Scrubbing systems with wet electric precipitators or deep-bed aerosol separators are suitable for minimizing emissions.

Wastewater from the manufacturing plant is biodegradable. Treatment with municipal wastewater is always preferable. The high water requirements of the off-gas conditioning system favors installation of closed circulating systems with the result that manufacturing plants can be operated free of wastewater.

Local ordinances must be observed in removing process waste. The insulating material itself, like other building materials, is inert and does not have the potential for endangering water. Both the dust from the filter units and the process waste is suitable for internal recycling. The fiberization units are serious noise emitters; therefore, the manufacturing plant must be housed in massive buildings.

The overall effect of mineral-fiber fibrous insulating material on the environment is positive.

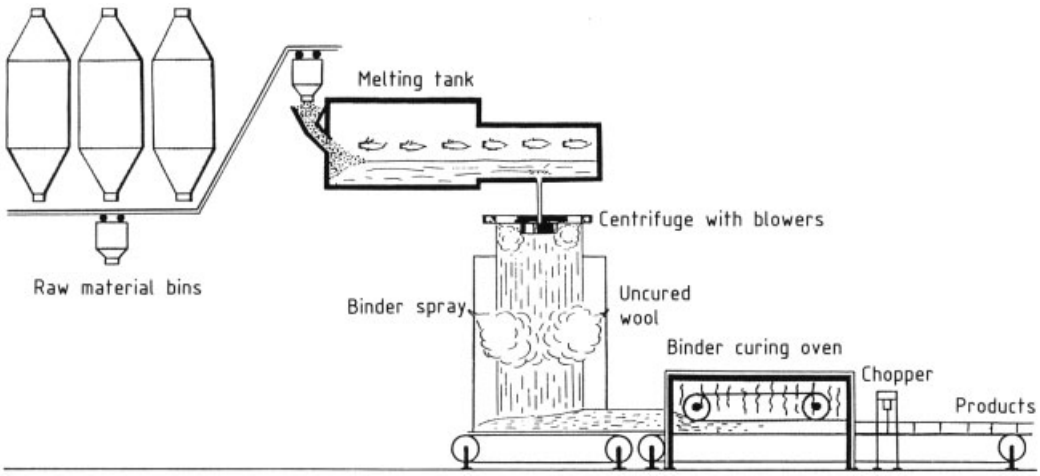

Figure 29. Mineral-fiber insulation, from raw material to product [42]
The rotary process is used to form fibers.

There is no other building material that offers a similar potential for minimizing the consumption of primary energy for the same raw material and energy requirements. Average energy amortization times of 0.3 years and ecological amortization times of less than 0.1 years have been calculated.

2.2.6. Quality and Analysis

Each country has very detailed standards for fibrous insulating materials. International standardization is in progress, e.g., in the areas of thermal insulation in high-rise construction, fire protection, and industrial acoustics [43].

Figure 30 shows a scheme of analysis for mineral-fiber insulating material. Glass analysis alone is, as a rule, not sufficient to identify the producer. However, analysis of characteristic trace elements generally allows identification of the producer. Certain finishing specialities of the final product may permit identification.

From the ecological point of view, an eluate analysis is essential. The eluate is obtained by shaking the insulating material at room temperature with a tenfold amount of water for 24 hours.

For the determination of formaldehyde, the perforator method DIN-EN 120 and the gas analysis method of DIN 52693 are suitable. Analogous methods are specified in other countries.

2.2.7. Applications

Mineral-fiber insulating materials are finished products for use in protection from heat, cold,

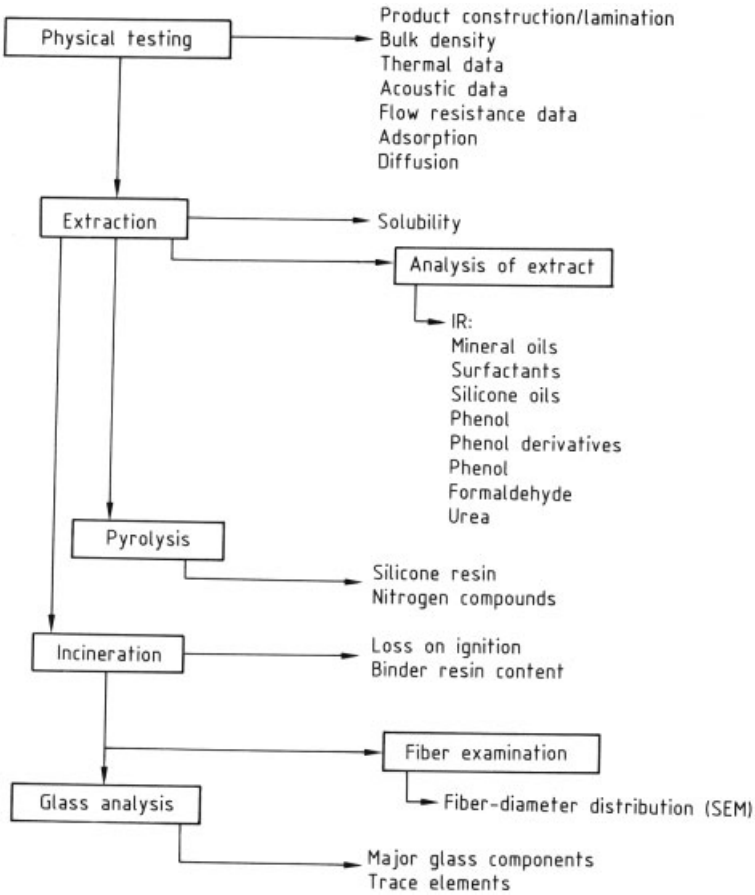

Figure 30. Scheme of analysis for mineral-fiber insulating materials

fire, and noise. The major end use area is high-rise construction. Considerable quantities are used in industrial and machine construction for thermal insulation. Fibrous insulating materials are also important as noise absorbers in buildings and elsewhere.

Economic Significance. World production of mineral-fiber insulating material reached a maximum of ca. 6×10^6 t in 1980 and has stagnated since (Table 14). About 75 % of the total is produced in the industrialized countries with temperate and subarctic climate. Consumption is closely tied to investment and building activity. About 60 % of the demand for insulation in industrialized countries is met by mineral-fiber insulating material.

Table 14. Production of mineral-fiber insulating materials in various geographic areas, 1973 and 1983.

Area	Production				
	1973		1983		Increase, %
	t	%	t	%	
North America	1 600 000	43	2 350 000	42	46.9
Western Europe	1 200 000	32	1 750 000	31	45.8
Eastern Europe and the Former States of USSR	600 000	16	800 000	14	33.3
Japan	200 000	5	400 000	7	100
Australia	} 120 000	} 3	50 000	1	} 150
Other			250 000	5	
World total	3 750 000	100	5 600 000	100	49.3

2.2.8. Toxicology and Occupational Health

Synthetic mineral fibers are hard, abrasive materials. Constant handling can damage the epidermis, especially the finger tips and other exposed areas of the body. Therefore, work gloves should be worn, and the skin should be kept smooth by applying oily, protective skin lotions.

The fibrous dust formed during mechanical loading, cutting, sawing, or rubbing may irritate the skin or the mucous membranes. Especially the coarse fibers, with diameters greater than about 12 µm, and the nonfibrous components are able to penetrate the epidermis, thus causing injuries and facilitating the entry of germs. The irritations are slight and rapidly heal with normal body hygiene.

Ordinances pertaining to protection of the workplace prescribe a general permissible limit of dust; in most countries, this limit is 5–10 mg/m^3. In the Federal Republic of Germany the limit is 6 mg/m^3. Epidemiological investigations suggest that there is no health risk associated with the handling of mineral-fiber insulating material if these limits are observed.

Suspicions of a *carcinogenic potential* were derived from animal experiments with extremely fine glass fibers carried out in 1972–1978. Up to now these suspicions have not been confirmed with fibers from insulating materials. Animal experiments involving dust inhalation gave no indication of carcinogenicity. Extensive epidemiological investigations could neither confirm nor deny these suspicions. Further research is needed.

Fibrous insulation materials are free of preservatives, fungicides, or biocides. They do not contain asbestos or silicogenic material. Formaldehyde is usually absent or present only in trace amounts (<10 mg/kg).

3. Refractory Fibers

Refractory fibers are used at temperatures in excess of 1000 °C. They are distinguishable from mineral and synthetic aluminosilicate fibers by their greater durability at high temperature and by the fact that they have polycrystalline microstructures rather than being amorphous. They are relatively new products, having only been developed during the last 20 years, and additional varieties continue to be announced at a significant rate. The majority of fibers of this type have been developed for thermal insulation at high temperatures or for advanced composites having the special physical properties required in aerospace applications. These materials are expensive since they are still manufactured on a small scale.

Historical Background. Until the 1960s, refractory fibers were either naturally occurring fibrous minerals of the asbestos type or melt-spun glass fibers such as E glass, rockwool, or kaolin-derived fibers. Their maximum use temperatures were limited by their melting points, kaolin fibers having the highest use temperature of 1250 °C. Since then more refractory grades

of melt-spun kaolin fibers, capable of use up to 1400 °C, have appeared.

The development of several types of novel oxide fibers since 1970 has had several motivations. At the time, an awareness was developing that a safe, noncarcinogenic, cheap asbestos replacement might be possible. Asbestos was widely used in medium-temperature insulation, in textiles, and in friction materials so that replacement in these uses appeared to be an attractive technical target. About the same time the worldwide rise in oil prices increased the demand for energy-saving insulation, especially in high-temperature furnaces operating above 1250 °C. Consequently the need for improved fiber refractoriness and freedom from nonfibrous impurities (combined with an awareness that very fine fibers like asbestos might be carcinogenic) lead to the development of new high-alumina and zirconia staple fibers. Simultaneously, continuous textile-grade multifilament fibers have also been developed to satisfy the need for high-temperature resistant textile applications such as furnace curtains, fire barriers, hot gas filters, electrical insulation, and protective fire-fighting suits. In the 1980s renewed interest in the potential of metal-matrix composites is also sustaining the further development of novel high-alumina fibers.

During the 1960s the needs of the U.S. aerospace program for high-modulus, low-density reinforcing fibers led to the development of entirely new non-oxide fibers; initially to CVD boron fiber and later to carbon fibers. To a large extent the development of cheap, high-quality, carbon fibers in the 1970s has overtaken that of boron fiber since the price of the latter has not been reduced significantly. A major disadvantage of carbon and boron fibers is their ready oxidation at moderate temperatures, and so during the late 1960s and 1970s various kinds of oxidation-resistant silicon carbide fibers have been evolved. For example, the makers of CVD boron have turned to CVD silicon carbide as a potentially cheaper and better product. Another driving force in the development of continuous textile-grade silicon carbide fibers (such as Nicalon) has been the desire to have good high-temperature properties combined with resistance to oxidation and to attack by molten metals. At present, therefore, advanced oxide fibers and silicon carbide variants represent the two main categories of refractory fibers.

The chief materials of interest are boron [*7440-42-8*], silicon carbide [*409-21-2*], high-alumina [*1344-28-1*] (0 – 30 % SiO_2/Al_2O_3), and zirconia-based [*1314-23-4*] products. For each of these materials there is a choice of several grades from a range of suppliers. Fibers are available in a variety of primary forms such as continuous multifiber yarn, short staple random fibers, or as whiskers (→Whiskers). Typically, products have diameters within the range 1 – 150 µm. Some of these primary products may be converted easily to secondary forms such as woven cloths, ropes, papers, boards, prepregs, etc. All products on the market have complex compositions and microstructure so that each can only be referred to uniquely by its trade name. These are listed in Table 15.

Other materials that are or have been available in fiber form are boron carbide (B_4C), silicon nitride [*12033-89-5*] (Si_3N_4, whiskers), and boron nitride (BN). For these materials the supply situation varies since their applications tend to be highly speculative.

3.1. Properties and Raw Materials

3.1.1. Physical Properties

The behavior of fibrous materials is strongly dependent on the shape and diameter of the fibers. Typically the finer fibers are in diameter, the more flexible they become and the more difficult it is to pack them into high volume fractions. Fine random staple and whisker fibers are usually low bulk density materials, while thick monofilament or multifilament continuous fibers, which are both more rigid and less flexible, are easily packed into high volume fractions (60 % for example).

These characteristics determine utility and the methods of handling. Fine staple fibers are easily used as low density thermal insulation and for the manufacture of low volume fraction composites, but may present dust problems in handling. Large diameter continuous fibers are more readily woven into textile forms for high-temperature use and are also more readily made into unidirectional high-volume fraction composites. On the other hand, the main handling

Table 15. Fiber properties *

Manufacturer	Fiber name	Chemical type	Type (fiber count)	Dimensions			Mechanical properties		Compressive strength, MPa	Fiber density kg/m^3,	Thermal properties			Crystal Phases
				Particulate	Diameter, µm	Length	Modulus, GPa	Tensile strength, MPa			Coefficient of thermal expansion, 10^6 K^{-1}	Melting point or liquidus, °C	Maximum use temperature, °C	
Avco	Boron	boron (tungsten)	filament		140	continuous	390	3000–5000	6600	2400	4.7–7.0	2200	500	Amorphous
Avco	SCS2	silicon carbide	filament		140	continuous	420	3400–4500	4000	3000		2830	1600	β-SiC
Nippon Carbon	Nicalon	silicon carbide	yarn (500)		15	continuous	176–196	2450–2950	>3000	2600	3.1		ca. 1400	β-SiC/graphite
Ube Industries	Tyranno	silicon—Ti—C—O	yarn		10–13	continuous	220	2800–3000		2300–2500	3.1		>1300	Amorphous
Arco Metals-Silag	Silar SCS9	silicon carbide	whisker	10–20 %	0.6	10–80 µm	700	7000		3200		2700		α-SiC
Tokai Carbon	Tokomax	silicon carbide	whisker	<1 %	0.1–0.5	50–200 µm	400–700	3000–14000		3190		2700	1600	β-SiC
Tateho Chemical	SCW#1	silicon carbide	whisker		0.1–0.5	10–40 µm	480	20700		3210				β-SiC
Los Alamos		silicon carbide	whisker		3–10	10–25 mm	520–650	2000–20000						β-SiC
Tateho Chemical	SNW#1	silicon nitride	whisker		0.5–2	50–300 µm	380	14000		3180		1900		
3M	Nextel 312	24 % Silica/ 14 % Boria/ Alumina	yarn (600, 900, 1200)		11	continuous	152	1720		2700	3.5	1800	1200	"inverse mullite"
3M	Nextel 440	28 % Silica/	yarn (600,		11	continuous	207–240	1720		3100	5.0	1890	1430	η- or γ-Al_2O_3

Table 15. (Continued)

Manufacturer	Fiber name	Chemical type	Type (fiber count)	Dimensions			Mechanical properties		Compressive strength, MPa	Fiber density kg/m^3,	Thermal properties			Crystal Phases
				Particulate	Diameter, µm	Length	Modulus, GPa	Tensile strength, MPa			Coefficient of thermal expansion, 10^6 K^{-1}	Melting point or liquidus, °C	Maximum use temperature, °C	
		2 % Boria/ Alumina	900, 1200)											
Carborundum	Fibermax	27 % Silica/ Alumina	staple		2 – 3.5		150	850				1890	1650	
Denka	Alcen	20 % Silica/ Alumina	staple; yarn	2 – 3; 10		170	1700		3300		1890	1600	mullite/α-Al_2O_3	
Sumitomo		15 % Silica/ Alumina	yarn (1000)		17	continuous	200	1500	2800 – 4400	3200	8.8	1890	1250	γ-Al_2O_3
ICI	Catalytic Saffil	5 % Silica/ Alumina	staple	<1 %	1 – 5	5 cm	100	1000		1800		1980		η-Al_2O_3
ICI	Saffil RF	5 % Silica/ Alumina	staple	<1 %	1 – 5	5 cm	300	2000		3300		1980		δ/θ-Al_2O_3
ICI	Saffil HA	5 % Silica/	staple	<1 %	1 – 5	5 cm	>300	1500		3400		1980	1600	α-Al_2O_3/mullite
Du Pont	Fiber FP	>99 % Alumina	yarn (200)		20	continuous	380	>1400	6900	3900		2040	1320	α-Al_2O_3
3 M	Nextel Z11	32 % Silica/ Alumina	yarn		14	continuous	76	1310		3700		2000	1000	tetragonal ZrO_2
Zircar Products	ZYBF-2	8 % Yttria/ Zirconia	yarn/staple		3 – 10	continuous/ 1.6 mm				6000		2580	2200	tetragonal ZrO_2

∗ In air, fiber strength severely affected.

risk is skin irritation caused by penetration of the rigid fibers.

The physical properties of most importance are fiber diameter, length, shape, density, modulus, tensile and compressive strength, expansion coefficient, thermal conductivity, and maximum working temperature. These are summarized from available manufacturers brochures in Table 15.

Only data on fibers with diameters greater than 1 μm are reliable. Data for finer whiskers are implicit from data on composites. Nevertheless, single crystal whiskers of silicon carbide have the best mechanical properties (modulus around 600 GPa, strengths up to 20 000 MPa). Polycrystalline Avco silicon carbide and boron and Du Pont α-alumina are of comparable stiffness, but Du Pont alumina is weaker than the others (modulus about 400 GPa and strengths 1500 – 5000 MPa). Finally polycrystalline transition aluminas, mullites, and Nicalon have moduli around 200 GPa and strengths around 2000 MPa. In composite applications specific modulus and strength are more important, and here boron possesses the greatest advantage since its density is the lowest. Silicon carbide and transition aluminas follow at ca. 3000 kg/m^3, while α-alumina and zirconia-based fibers follow with densities of ca. 4000 kg/m^3. Figure 31 shows SEM microphotographs of three different refractory fibers.

3.1.2. Chemical Properties

Refractory fibers as a class of materials are of low reactivity. The chief area of concern is their resistance to oxidation at high temperatures and, if used in metal – matrix composites, their inertness to corrosion by metals during fabrication and use. Reported thermal data are also given in Table 1, including claimed maximum use temperatures. Boron fibers are the most reactive towards oxidative and metallic attack. Several attempts have been made to coat boron fibers with inert layers, which has led to the BorSic product (SiC-coated boron). The chemical vapor deposition process for making boron fibers lends itself to laying down additional coatings. In France boron carbide-coated boron has also been tested [44].

The properties of silicon carbide fibers and whiskers in composites have been thoroughly reviewed by Andersson and Warren [45,46]. Their resistance to metal attack exceeds that of boron by 200 – 300 °C. High alumina fibers are inert in air and light alloys, but are slowly attacked by hot strong acids or alkalis. Zirconia fibers have exceptional thermal stability and are also resistant to hot alkali.

3.1.3. Resources, Raw Materials

The raw materials for refractory fibers are usually specially selected intermediates. Since the elements needed are among the most abundant,

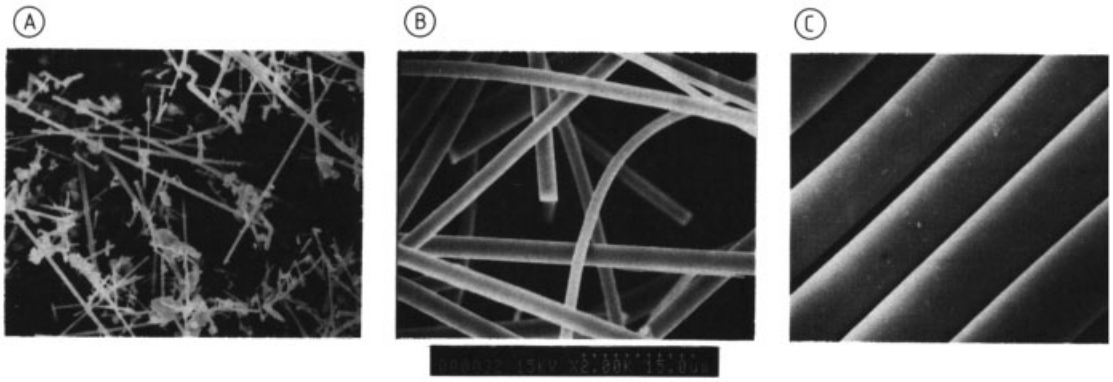

Figure 31. SEM microphotographs of several refractory fibers; scale: —⟶—10 μm
A) Silicon carbide whiskers; B) Aluminum oxide staple fiber (Saffil, ICI); C) Titanium-modified silicon carbide fiber (Tyranno, Ube Industries)

there is no shortage of primary sources. The intermediates, however, are only readily available in industrialized countries. In the case of silicon carbide whiskers, a process has been commercialized which uses rice hulls as raw material [47].

3.2. Production

Because of their relative novelty, the processes for making refractory fibers are almost all proprietory, and full details have never been published. A summary of the patent literature [48] since 1970, relating to both the production of refractory fibers and their uses in composites, has appeared. There is a special section on refractory fibers in BELA VON FALKAI's book [49]. For further useful information refer to [50, 51].

3.2.1. Boron Monofilament

CARLSSON [52] has reviewed general methods of making boron fibers. In the 1960s, Avco (USA) developed a chemical vapor deposition (CVD) process to make 140-µm boron monofilaments having an 8-µm tungsten core [52, 53] on a scale of 15 t/a. According to the company's latest brochure, the present capacity is only slightly greater. More recently, a cheaper variant was made available, in which the core is a 30-µm carbon fiber, but this has now been withdrawn. This development was initially financed by the U.S. Government to develop advanced composites for use in the space shuttle and in aerospace applications [53].

The process deposits amorphous boron with a nodular morphology from a BCl_3–H_2 mixture onto the electrically heated tungsten core at ca. 1300 °C.

$$BCl_3 + 3/2\,H_2 \rightarrow B + 3\,HCl$$

The operation is carried out in single-stage or two-stage tube reactors (Fig. 32), the filament passing at 3 – 4 m/min through mercury seals, which also act as electrical contacts. Control of deposition temperature is critical because too high values lead to crystallization of the boron, which reduces the strength of the filaments. Auxiliary heating techniquescan increase quality

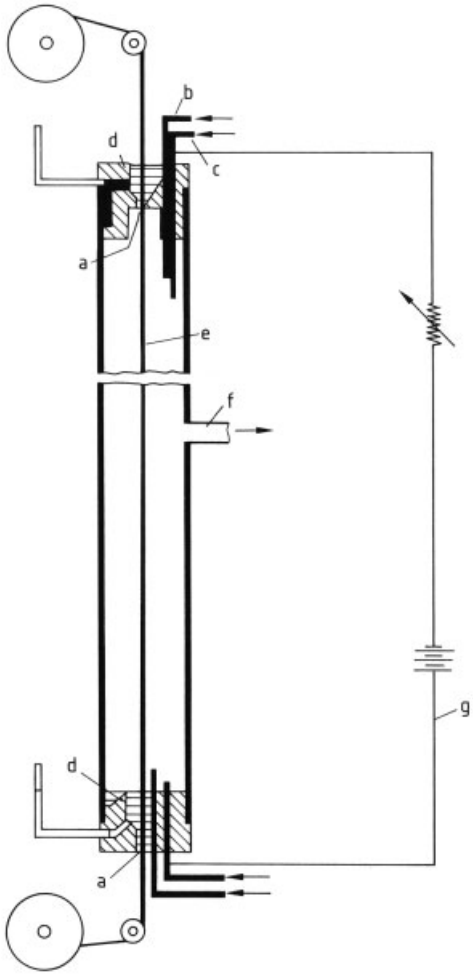

Figure 32. Apparatus with mercury seals for the preparation of boron fibers by two-stage chemical vapor deposition [76].
a) Apertures (25 – 254 µm diameter); b) hydrogen feed tube; c) BCl_3 feed tube; d) Mercury pool; e) Tungsten wire substrate; f) Gas effluent; g) Power supply (d.c.)

and output. A range of products with various diameters has been produced. Initially, the tungsten core made up a large proportion of raw materials costs, but later the price of tungsten dropped, and larger diameter fibers were developed so that this cost factor decreased in importance. The introduction of carbon-cored B filament has lead to further cost reduction [53]. Various surface treatments, such as acid polishing, have been investigated to increase the strength from 2 GPa to 5 – 7 GPa. Also surface coatings such as silicon or boron carbide are applied by

the CVD technique [54] to reduce the reactivity with aluminum alloys and oxidation at temperatures above 300 °C.

In the commercial plant, hundreds of reactors are fed in parallel with the gaseous reactants. Since the conversion of gas per pass is low, the unconverted gases are recovered and recycled, while the byproduct HCl is sold.

More recently, more rapid boron deposition methods have been reported: these methods operate at high temperature and high pressure and do not require resistive heating of the substrate. Pulse discharges have been used for boron coating of filaments [55]. The excited gas species are formed by inductive or capacitive coupling of pulsed high electric power to a boron-containing gas mixture at a pressure of 0.1 – 0.3 MPa (Fig. 33). The gases admitted to the deposition zone are hydrogen, mixed with boron halides, boranes, or alkyl borates. The deposition process is performed at 20 – 350 °C on filaments suspended in the deposition zone: tungsten filaments, graphite yarn, and polyacrylonitrile yarns were coated with boron to a thickness of about 50 µm. The residence time of the substrate within the deposition zone was about 5 min.

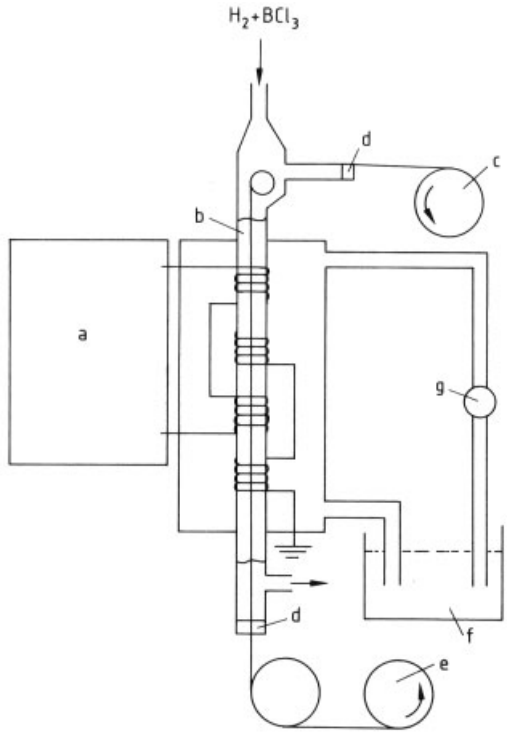

Figure 33. Apparatus for the production of boron fibers by capacitive coupling of pulsed high electric power [55]. a) Power supply (pulsed high electric power); b) Oil-cooled reaction tube; c) Let-off spool; d) Seal; e) Take-up spool; f) Oil reservoir; g) Pump

3.2.2. Silicon Carbide Fibers

Because of the highly attractive physical properties of fully crystalline silicon carbide and the low cost of raw materials, more effort has been devoted to generating silicon carbide in fibrous forms than for any other material [45, 46, 56].

Silicon Carbide Monofilament. This material is available in composites from Avco and is probably made by a CVD process [57] similar to that used for boron filament. The reagents are an alkylchlorosilane or a mixture of alkane and chlorosilane condensed onto a tungsten or carbon core. The product has a columnar grain structure with the growth direction normal to the fiber axis. In order to increase the strength of the fiber it is usually coated with a thin carbon layer (CVD process), which fills in any grain boundary faults at the filament surface.

Silicon Carbide Continuous Yarn. Silicon carbide continuous yarn was invented by S. Yajima in the 1970 – 1975 period and has since been commercialized by Nippon Carbon Co. as Nicalon [58]. A second generation carbide fiber, namely Tyranno fiber from Ube Industries, contains titanium in addition to silicon and is reported to have superior strength and chemical resistance.

The Nicalon fiber process begins with the conversion of dimethyldichlorosilane (an intermediate in silicone manufacture) to poly(dimethylsilane) by dechlorination with sodium. The poly(dimethylsilane) is heated under Ar in an autoclave at 470 °C for 8 – 14 h to convert it to a poly(carbosilane) of complex structure (see [58]). After vacuum distillation to 280 °C to remove low molecular mass fractions, the poly (carbosilane) (M_r 1500) is then melt-spun at 350 °C as a multifilament tow. The raw fibers are then cross-linked by oxidation in air for 30 min up to 190 °C and heat-treated to 1200 °C in inert atmosphere to convert them to the fin-

ished fiber having maximum strength. The fibers have a complex constitution consisting of a fine-grained mixture of graphite, β-SiC, and glassy silica. The fiber stability is limited above 1200 °C, and the modulus (which is a function of fiber diameter) is only 50 % of the value of CVD polycrystalline silicon carbide.

Silicon Carbide Whiskers (→Whiskers, Chap. 4.1.). These have traditionally been made by high-temperature synthesis between SiO_2 and a carbon source under reducing conditions. A more recent development uses as raw material rice hulls, which contain Si and are processed at 1500 – 1660 °C [47, 59]. They are usually single crystals, although there is some confusion in the literature as to the direction of the growth axis. Very high strength β-silicon carbide whiskers (6 µm diameter) are grown by a vapor – liquid – solid (VLS) method at 1400 °C from methane and silicon monoxide using powdered iron catalyst [60, 61]. Properties are given in Table 1.

3.2.3. High Alumina Fibers

High alumina fibers containing less than 30 wt % SiO_2 cannot be readily spun from melts because the temperatures are too high. Fibers in this class are either made from water-soluble or dispersible precursors (sol – gel products) or from an organometallic precursor (only Sumitomo's alumina). Fibers that are available range from fine diameter (3 µm) random staple (ICI Saffil alumina fiber) through to relatively coarse (20 µm) continuous fibers (Du Pont FP alumina). This class includes mullite [*1302-93-8*] fibers (27 % SiO_2) and 3 M Nextel alumina – boria – silica fibers. Single-crystal whisker products are also available (→Whiskers, Chap. 4.3.).

Continuous Yarn Fibers. Du Pont, 3 M, and Sumitomo have supplied products of this type for several years. ICI, Denka Kagaku, and Mitsui mining announced new alumina-based continuous or semicontinuous fibers.

In each case a different precursor is used, but the processes are otherwise broadly similar (Fig. 34, see next page). First a viscous precursor solution is prepared and then converted to green (or gel) fibers by extrusion through multi-hole spinnerets. The gel fibers are converted to refractory fibers by heat treatment, usually in multistage furnace treatments. Major features of these processes are the ability to make any composition and the large shrinkage experienced during calcination.

According to patents [62, 63], Du Pont's FP fibers may be made from colloidal α-alumina (e.g., Alcoa XA16 fractionated to 90 % <0.5 µm diameter) dispersed by HCl treatment and blended with an aluminum salt, such as the chlorohydrate (commercially available basic aluminum chloride [*1327-41-9*], $Al_2(OH)_5Cl \cdot 2-3\,H_2O$), acetate, or nitrate. The viscous mixture is extruded into air to give gel fibers, which are then subjected to two- or three-stage calcination to give the final fiber (either flame-polished in an oxypropane flame or at 1400 – 1500 °C in a furnace). The fibers have a diameter of 20 µm and are polycrystalline α-alumina with a grain size <0.5 µm. Their strength can be improved by a surface coating ca. 5 nm thick of silica derived from a silicone [64].

According to a patent [65] 3 M's Nextel fibers are made from a silica hydrosol blended with aluminum formoacetate and boric acid together with organic improvers such as polyvinylpyrollidone (PVP), poly(vinyl alcohol) (PVA), lactic acid, or glucose. After concentration to a viscosity between 45 and 500 Pa · s, the solution is extruded through a multi-hole spinneret into air to give gel fibers. These gel fibers are calcined in air to at least 1000 °C (polycrystalline η- or γ-alumina <60 nm grain size) or to 1400 °C (polycrystalline mullite >50 nm grain size) if they are to be used up to 1400 °C. The finished fibers have an oval cross section with a mean diameter of 11 µm and are suitable for weaving cloths.

Sumitomo has described their process [66, 67]: first triethylaluminum or triethoxyaluminum is hydrolyzed with a controlled amount of water to a polyalkyl(oxy)aluminoxane (degree of polymerization of 10 – 200), which is then dissolved in an organic solvent. An alkyl silicate is added (to give 15 wt % SiO_2 in the final product) to prepare the viscous solution (0.1 – 500 Pa · s) for spinning. The fibers are dry-spun into air. These gel fibers are then calcined in air to 700 °C to decompose them to oxide and further fired at 1000 °C to crystallize them to a dense

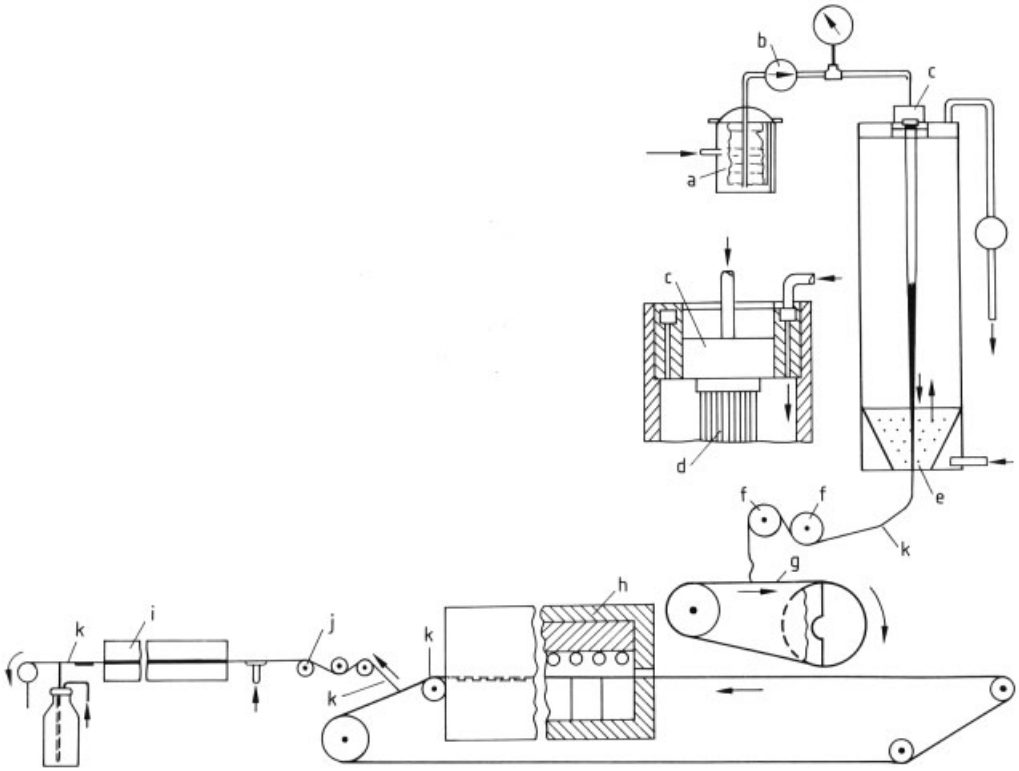

Figure 34. Flow chart of continuous refractory oxide fiber production [77].
a) Precursor solution; b) Metering pump; c) Multi-hole spinneret; d) Fiber bundle; e) Hot air distributor; f) Draw rolls; g) Transfer conveyor; h) Pre-fire resistance furnace; i) High-temperature tube furnace; j) Transfer rolls; k) Multi-filament fiber tow

polycrystalline fiber having a γ-alumina and/or a mullite structure with a grain size of several nanometers). Production fibers are made with a diameter of 17 µm.

So far no details are available for methods of producing the newest continuous fibers.

Staple Fibers. Staple fiber products include ICI Saffil alumina, Carborundum's Fibermax (and the equivalent Toshiba Monofrax), and several products available from Japan such as Denka's Alcen.

In sol–gel processes [68], an aqueous spinning solution is formulated from basic aluminum chloride (aluminum chlorohydrate, $Al_2(OH)_5Cl \cdot x\,H_2O$), a water-dispersible siloxane, or a silica sol, and a viscosity modifier such as PVP, PVA, or poly(ethylene oxide). Fibers are either blow-spun through multi-hole spinnerets or centrifugally spun into warm dry air and collected as a random mat on a conveyor belt. They are decomposed at 200–500 °C and calcined in one or two stages to recrystallize and sinter the initially amorphous material. Various products are available, with different finishing temperatures and containing different crystalline-transition alumina phases (Table 1). Centrifugally spun products contain more nonfibrous material than blow-spun products.

3.2.4. Zirconia Fibers

Zircar fibers (yttria stabilized) are made by the relic process, where a continuous cellulose fiber is impregnated with suitable salts and then calcined in air to decompose the salts to oxides and burn away the cellulose. Yttria is added to maintain the zirconia in the tetragonal phase, preventing the destructive conversion to monoclinic phase on cooling. Alternative phase stabilizers

that have been investigated are calcium oxide and magnesia. An alternative process, which has been used by ICI to make staple fibers, is the sol – gel route. This is described by others in the literature [69] and generally uses zirconyl acetate as the raw material. Other details are similar to those used for alumina fibers. Nextel Z-11 continuous zirconia – silica [*14940-68-2*] fibers from 3 M are made much like the alumina – silica – boria products [70].

3.3. Environmental Protection

Since these products have only been made in small quantities so far and are physically and chemically inert under ambient conditions, they pose no significant threat to the environment. Ultimate disposal is likely to be by burial. The byproducts of fiber manufacture are either acids, which are scrubbed and sold or neutralized before disposal, or carbon oxides and small amounts of volatile organic compounds.

3.4. Uses

The principal uses of refractory fibers are in high-temperature insulation and composites.

3.4.1. High-Temperature Insulation and Filtration

By reason of their refractoriness, those fibers which are not attacked by oxygen or which form passive oxide layers (oxides, SiC) can be used in furnaces as high-temperature insulation or structural elements. High-alumina fibers dominate these applications. In general, the staple and whisker materials that have diameters below 6 μm are best for low-density insulation having a low thermal conductivity. Conductivities at 1400 °C as low as $0.3 - 0.4\,Wm^{-1}\,K^{-1}$ can be reached. Use of such low thermal mass insulation allows faster cycle times in batch furnaces, giving energy savings of 20 – 40 %.

In furnace applications, the short fiber products can be used in a variety of forms. When resilience is required, blanket or loose wool can be employed in several ways. In one method, strips of blanket are glued to a rigid backing (stack-bonding) to give a surface with high gas erosion resistance. This technique can be applied to existing brick-lined gas-fired furnaces (veneering) to give substantial energy savings. Staple fibers can also be converted by wet-lay vacuum-forming processes to rigid forms. For such products the refractory fibers are often blended with amorphous alumino-silicate fibers so as to uprate the maximum working temperature of the amorphous fibers, e.g., from 1250 to 1400 or 1500 °C. In these forms, sacrificial organic binders, such as starch, are often combined with a refractory binder such as a silica sol. It is also possible to make papers by blending small proportions of fine glass fibers into the slurry.

Thermal insulation for nuclear reactors presents a requirement of high resilience under high acoustic vibration loadings at modest temperatures. Very low shot level staple fibers are superior for this application.

The yarn and monofilament materials are best used as woven cloths [71], which can be employed as flame barriers and radiation shields or as bag filters for hot-gas filtration. Where density is not a prime consideration, Nextel Z11 is claimed to have superior firewall properties. The long fiber products can be converted to string, woven cloths, ropes, tapes, thread, braiding for electrical wiring, etc. Such yarns, ropes, etc. can be employed as expansion joint packings, door seals, and the like. Both staple and continuous fibers can be used for space vehicle insulation systems, e.g., tiles as on the space shuttle.

Both short and long fiber products can be used in advanced composites. Large volumes of data are available for composite properties using either resin matrices, metal matrices, or ceramic matrices. These present a range of opportunities.

The refractory properties of the fibers are most effectively exploited in metal – matrix or ceramic – matrix composites because good properties are then available above 200 °C in many such systems. Both types of material are presently undergoing intensive development to evaluate their potential for replacement of conventional metal alloys.

Resin-Matrix Composites. In resins, refractory fibers are in competition with carbon fibers and Kevlar fibers. In situations where the composite is required to have high compressive

strength, the use of boron, silicon carbide, or alumina fibers is preferable. This has found utility in parts of sporting goods such as golf clubs, tennis rackets, and fishing rods. Aerospace applications have also been explored [53]. The U.S. F14 fighter had boron – epoxy horizontal-stabilizer skins. Where electrical insulating properties as well as mechanical properties are important, then alumina or silicon carbide fibers can be more effective than glass or Kevlar fibers.

Metal – Matrix Composites. Metal – matrix composites (MMCs) where the matrix is an aluminum, magnesium, titanium, or nickel alloy, should be able to replace parts currently made of steel, titanium, or nickel alloys. Fabrication and shaping of MMCs presents formidable problems. Corrosion of fibers by the metal during fabrication and use is the chief problem. Aluminas are superior in this respect to silicon carbide and boron, which, for example, is only stable to 200 °C in 2024 aluminum alloy. However, boron fibers are usable when SiC-coated. Because of high costs, nearnet-shape forming methods such as squeeze-casting, extrusion, and forging are preferred. Aluminum-based MMCs have reached commercialization in a few applications. Boron filament – aluminum composites have been utilized for cargo bay spars in the space shuttle. In the future, other fibers, such as silicon carbide (continuous and whiskers) [72, 73] and alumina [74], will find uses in aerospace and automotive engineering applications where low density must be combined with rigidity, strength, heat, and/or wear resistance.

The use of small amounts (<30 vol %) of short alumina fibers in diesel engine pistons has been commercialized by Toyota. Here the effect of fibers is to increase the wear resistance of the piston alloy in the region of the ring grooves so that high nickel alloy inserts are not required. In more advanced applications such fibers in piston crowns can increase resistance to fatigue cracking so that more efficient engines become possible. High volume fractions of unidirectional long alumina fibers are currently being evaluated in connecting rods for internal combustion engines to decrease the mass of moving parts, which should lead to increased engine efficiency.

BorSic and silicon carbide monofilaments have been successfully incorporated into Ti-6Al-4V alloy [44] with the aim of increasing the maximum operating temperature of this alloy, but this matrix is generally too aggressive to be compatible with other types of fibers.

Ceramic – Matrix Composites. When materials are required to exceed the maximum working temperature of nickel-based superalloys, ceramic – matrix composites are forecast to fill the gap. If the more severe processing and property requirements can be met, they will be utilized in advanced gas-turbine blades and in advanced internal combustion engine components such as pistons and exhaust turbo-charger rotors.

Currently much effort is devoted to improving the fracture toughness of monolithic engineering ceramics such as glasses, glass ceramics, silicon carbide, silicon nitride, alumina, and partially-stabilized zirconia by the incorporation of fiber reinforcement. This is the most severe application of refractory fibers since they are required to resist corrosion by the matrix during fabrication of the composite at temperatures often exceeding 1200 °C. Some encouraging results have been reported in the literature [57], but a real application has not yet been successful.

3.5. Miscellaneous Uses

Miscellaneous applications of refractory fibers have developed, such as the use in nuclear technology. Here, boron fibers find uses in fabricating flexible neutron shields.

Zirconia fibers find utility because of their great resistance to alkali attack and are recommended for use as battery separators. Nextel Z-11 has outstanding flame resistance and is preferred for fire-wall uses.

Porous fibers, such as ICI catalytic grade Saffil alumina, have been found to be effective as catalyst supports in domestic flameless heaters.

3.6. Economic Aspects

Refractory fibers are specialities made on a relatively small scale where raw materials costs make little contribution to the selling price. The major costs are related to processing and capital costs.

Of all the fibers considered, short staple alumina fibers are the product available in the greatest quantity and at lowest price. World capacity probably exceeds several hundred tonnes per year. Prices range from \$ 30 to \$ 60 per kg (ca. 30 – 60 €/kg).

Prices for continuous fibers or whiskers vary from \$ 80 – 900 per kg. The cheapest is Nextel 312 at \$ 80 per kg with Nextel 440 at \$ 260 per kg. SiC whiskers, Nicalon, Sumitomo alumina, and DuPont FP fibers all cost ca. \$ 450 – 550 per kg, while Avco boron fiber (tungsten core) is \$ 825 – 900 per kg. These products are all probably made on a 10 – 20-t/a scale. Du Pont has projected that at 500 t/a FP alumina fiber would cost \$ 88 per kg.

Secondary forms of the above materials–such as vacuum-formed shapes for staple fiber or woven cloths for continuous fibers–can cost double the basic fiber cost. Some products are available as epoxy prepregs, and there is an appropriate markup in price for such intermediates.

3.7. Toxicology and Occupational Health

Fibers with diameters exceeding 3 µm are not respirable and, like glass fibers, are difficult to handle only because of skin irritation. For such materials, operators handling the fibers are recommended to wear loose-fitting clothing.

Fine fibers with diameters below 3 µm are respirable. Because of the known ability of asbestos dust to cause both asbestosis and mesothelioma, considerable effort has been devoted to assessing other fibrous mineral and ceramic dusts for similar effects. In general, suspicion has now focussed on those materials that are durable in the body and have diameters below 0.5 µm and lengths greater than 8 µm [75]. Products that have diameters significantly above this limit have been found to be inert biologically. According to a proposal, such materials should be classified as nuisance dusts, for which an MAK of 6 mg/m^3 is appropriate. In the UK and the United States the MAK is 10 mg/m^3. The finer materials, especially whiskers, remain under suspicion. Similar products have been withdrawn from the market because animal tests have proved them to have positive carcinogenic potential. For such materials an MAK equivalent to that of asbestos has been recommended until more definitive information is available.

4. Metal Fibers

The definition for a metal fiber is not straightforward, especially when metal fibers are classified according to nature, structure, properties, or manufacturing processes. What cannot be considered a metal fiber is easier to state:

A single drawn wire used separately; anything much thicker than 100 µm; and steel wool. Metal fibers can usually be handled in the same way as textiles, paper, or polymers, and have small, equivalent diameters.

In this chapter, emphasis is placed upon bundle-drawn and melt-spun fibers. For metal whiskers see →Whiskers. Other methods of manufacturing fibers are mentioned briefly.

4.1. Production

4.1.1. Bundle Drawing

Very fine wires can be produced by wire drawing. Production costs are extremely high, however, and increase exponentially as the diameter decreases. To overcome this problem bundle drawing was developed as early as 1936 [78]. Instead of drawing wires singly, a number (in some cases several thousand) are bundled and drawn simultaneously [79] (Fig. 35, see next page). The sole difficulty remaining is to separate the individual fibers with a suitable material prior to bundling so that the final fibers can be individualized again. Quite often used is another metal that has sufficiently similar properties to the wire metal to be made into fibers, to withstand the drawing and annealing operations, but a lower chemical resistance to allow the fibers to be freed from this matrix in a leaching operation. Whereas this method greatly reduces drawing and annealing costs, it only produces bundles of fibers and not reels of single wires such as is needed to produce fine tungsten wire to serve as core material for boron fiber manufacturing.

4.1.2. Melt Spinning [80–82]

The basic idea behind all the from-the-melt production methods is the possible economic advantage of single-step processes over conventional time- and energy-consuming, labor-intensive fiber-drawing processes. These from-themelt processes can be roughly divided into two classes: melt extrusion and melt extraction. Dating back to the 1930s, from-the-melt fibers developed rapidly in the 1960s after Duwez developed the splat cooling technique: a revolution in metallurgy, resulting in what are now generally known as the RST (Rapid Solidification Techniques), sparked intense interest in glassy or amorphous metals. The Battelle Development Corporation has pioneered all these processes in one way or another, and is still doing so.

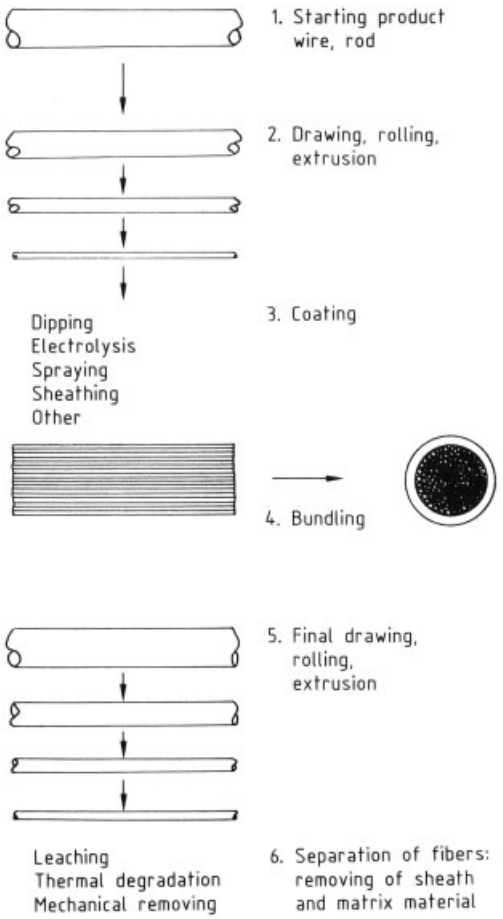

Figure 35. Manufacturing alternatives of bundle-drawn metal fibers

Melt Extrusion. Melt extrusion (melt spinning) involves the solidification of a liquid jet of metal extruded through an orifice (Fig. 36). The stream of molten metal is inherently unstable, mainly due either to its high surface tension and low viscosity or to a violent drop in its viscosity at its spinning temperature.

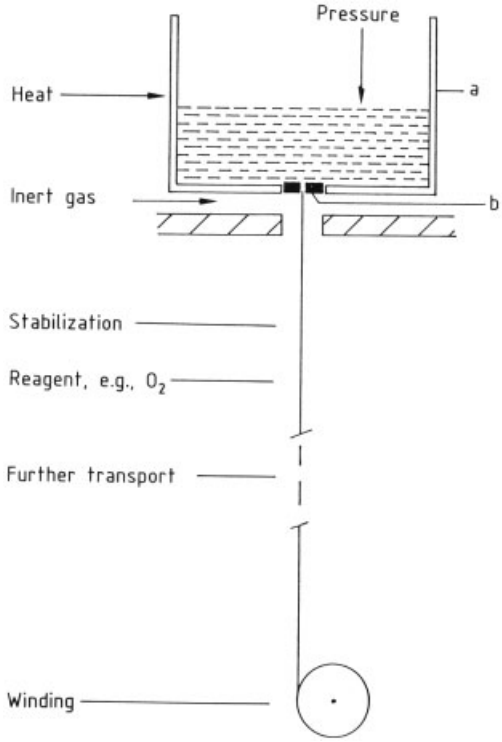

Figure 36. Melt spinning of metal fibers.
a) Crucible; b) Nozzle

The stream must be stabilized. There are a number of variants, which can be grouped in four divisions:

1) Jet stabilization in a controlled atmosphere, triggering off a surface reaction and oxide film formation (e.g., Monsanto, Michelin, Bekaert, Brunswick).
2) Jet stabilization by sheathing in glass and spinning the resulting composite. This so-called Taylor process (Fig. 37) has been successful on a semi-industrial scale (Glass Development, U.K.), but has virtually been abandoned in spite of the excellent fiber quality, mainly because of low production capac-

ity, hence high cost. Fibers down to 1 μm are possible.

3) Stabilization by heat transfer to a liquid (in-rotating-liquid spinning) (e.g., Allied Signal, Michelin, Unitika).
4) Solidification by direct contact with a moving-solid, heat-extracting surface. Although round fibers cannot be made by this technique, it was and still is the only one that has developed into a full-blown industrial process (Allied Signal, 1976). In this respect, it is the process responsible for the commercialization of amorphous metals (ribbons and sheets).

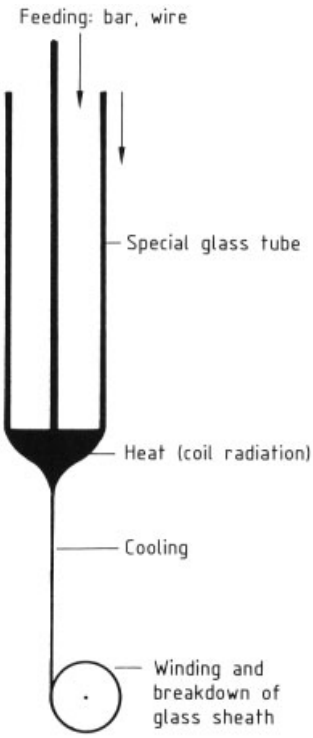

Figure 37. Taylor process for the production of metal fibers

Most of the processes failed to become fully industrial because they could not be upgraded from workable single-fiber processes to multi-fiber operations. The most promising method is the in-rotating liquid spinning, established in 1981 and the only method that enables fibers with circular cross sections to be manufactured in continuous lengths.

Improvements remain to be realized with respect to production costs (development of the multi-hole process), raw-material prices (up to 10 % boron), and some vital mechanical properties, especially torsion characteristics and fatigue behavior.

Melt Extraction. The melt extraction processes avoid the need for orifices because the molten metal is extracted from a melt by a wedge-shaped edge of a rotating disk: CME (crucible melt extraction) or PDME (pendant drop melt extraction), see Figure 38. The process was invented and developed by Battelle and exploited on an industrial scale by Fiber Technologies (U.K.).

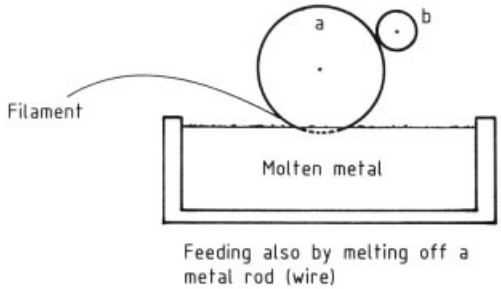

Figure 38. Crucible melt extraction (CME) for the production of metal fibers.
a) V-edged wheel; b) Wiping wheel

4.1.3. Other Processes

Metal fibers have also been made by chemical vapor deposition, mechanical methods, and electrolytic procedures. Other less important methods include slurry drying, decomposition, and reduction; in-situ decomposition of inorganic and organic compounds; and physical vapor deposition.

Chemical Vapor Deposition (Whiskers). Metallic whiskers are generally made from the gas phase by decomposing a compound of the metal (e.g., pentacarbonyl iron). Whiskers of iron, copper, and tin have been manufactured.

Mechanical Methods. Mechanical shaving methods, similar to those for steelwool production can be used, starting from either wire or stacked thin foils. The fibers so produced have irregular cross sections, either triangular or rectangular with a surface roughness and poor mechanical properties. The shaving methods were im-

proved by the introduction of superimposed vibrations, which result in more homogeneous cross sections and properties.

Electrolytical Procedures. After carbonization, organic fibers can be coated by electrolysis to produce metal sheaths. The organic and carbon residues can also be decomposed to leave hollow fibers. Metal fibers of Ni, Cr, or Cu as well as their alloys, can be formed. Another process consists of depositing continuous fibers on a suitably grooved mandrel.

4.2. Bundle-Drawn Fibers

4.2.1. Form and Dimensions

Bundle-drawn fibers can be produced in the form of continuous bundles, broken bundles (slivers), cut fibers, spun yarns, threads, cables, web, sintered web, needle felt, filter media, etc. All these morphologies are obtained by textile industry-like operations combined with metallurgical techniques, processes, and methods. The fiber diameters depend to a large extent on the metallurgical and structural purity of the starting material (wire rod), itself a function of the alloying elements and the manufacturing and purification procedures during wire-rod production.

4.2.2. Composition

The following are typical alloys lending themselves to the bundle-drawn fiber manufacturing:

Stainless AISI 316L
Stainless AISI 302
Carpenter 20Cb3
Inconel 601
FeCrAlloy
Hastelloy X
Nichrome
titanium

Standard diameters are 4, 8, 12, and 22 µm (Fig. 39). Of course, larger diameters are possible, such as 40, 50, and even 100 µm. For some alloys 12 µm is the lower limit. For very pure alloys, fibers 2 µm and even finer are possible (Fig. 40).

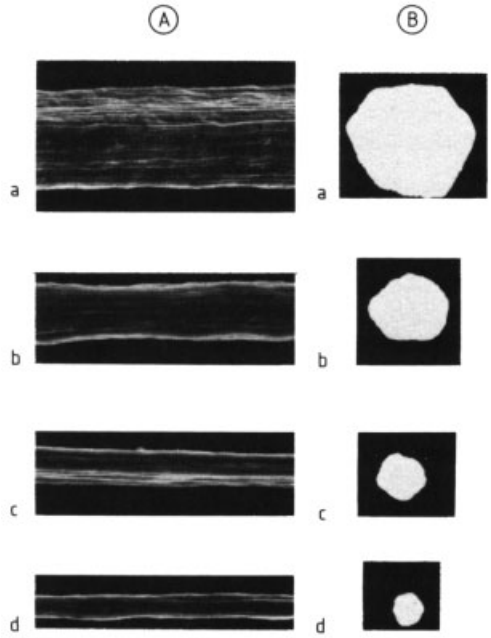

Figure 39. Microphotographs of bundle-drawn metal fibers A) Scanning electron microscope (×1250) B) Optical microscope (×1100)
a) 22 µm; b) 12 µm; c) 8 µm; d) 4 µm

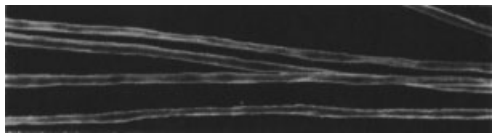

Figure 40. Very thin bundle-drawn metal fibers, diameter 2 µm (×625)

4.3. Uses

Anti-Static Applications. Since metal fibers are good electrical conductors, they were first used in anti-static carpeting to dissipate the static charges. Very soon the same principle was applied to clothing for medical use as well as for work with explosives or sensitive electronic components of instruments, followed by anti-static filterbags for use with explosive dusts [83]. Cast floors, e.g., as in operation rooms, are made sufficiently conductive to eliminate the possibility of sparks and electric discharges. The same also applies to bed linen for use with oxygen tents, etc.

Filter Media [84]. Filters are one of the most promising areas for the use of metal fibers, since the existing products, such as woven wire mesh and sintered metal powders, no longer fulfill all criteria:

1) Fine pore size from 20 µm to a few micrometers
2) Large permeability (4 – 10 times higher than at present)
3) High dirt holding capacity, therefore longer operational life (fourfold or more than at present)
4) Acceptable and competitive price in the application niches

Only sintered metal fiber filter media can satisfy these requirements successfully. The principal areas of metal fiber filter media applications include the following:

polymer fiber and foil manufacturing
hot gas filtration
aircraft fuel de-icing and filtration
filtration in the fields of hydraulics and robotics
fine chemicals and pharmaceuticals
crossflow filtration
generation of nuclear energy: gas cooled reactors, cooling
water condensate, elimination of radioactive aerosols from waste gas flows during emergencies

Sound Absorption. Sintered metal fiber sheets can be deployed as highly effective sound barriers under circumstances where other absorbers are inefficient, e.g., in high-temperature environments and at high-energy sound levels. This is particularly the case for aircraft and turbine noise reduction.

Gas Burner Panels and Tubes [85]. Highly porous sintered FeCrAlloy sheets and tubes produce an efficient radiant surface for burning natural gas. The advantages over the usual ceramic burners are improved life cycle, lower NO_x, better malleability, and sturdiness.

Thermal Insulation and Flame Arresting. The heat conductivity of porous metal fiber structures is comparable to that of an average quality glass fiber mat, and such products have solved high-temperature insulation problems such as those involving the exhaust gases of the space shuttle thrust rockets. These sintered metal fiber structures also make highly permeable, extremely efficient, versatile flame arrestors.

Medical Applications. Titanium fibers are bio-compatible and can consequently be applied as porous ingrowth materials for implants to evoke bone development in the pores, guaranteeing perfect fixing and load transmission. Stainless steel fibers can be used to reinforce dental bioglass.

Electromagnetic Interference (EMI) [86]. With the proliferation of electronic appliances and instruments in our daily life, electromagnetic interference must be controlled. More and more appliances are shielded against outside interference, as well as against creating interference themselves. Present-day housings for electronic devices are made from composite materials of polymer matrix and metal fibers. The presence of a low concentration of conductive metal fibers of the correct alloy, structure, diameter, and length give these casings a total attenuation of more than 80 dB, this without altering the other properties of the polymer matrix.

Radar Camouflage. As a special case of EMI, an amount of the correct metal fiber in radar camouflage netting that matches as closely as possible the absorption and reflection characteristics of radar waves for the environment has the effect of concealing an object. Structures containing metal fiber can be adjusted to vary from zero to almost 100 % reflection or absorption.

Microwave Detection. Every metallic fiber structure has a unique microwave signature, or fingerprint, that can be detected, registered, personalized, stored, and recognized. This feature can serve for the identification and verification of valuables, security papers, passports, credit cards, identification cards, etc.

High-Voltage Protection. The effect of the strong electromagnetic field on persons working on live high-tension lines can be harmful to their health. A protective suit made from cloth containing up to 25 % metal fibers acts as an efficient shield.

Capillary Uses. Since bundle-drawn metal fibers are finely grooved on their surfaces and form very fine tubes in between themselves, a fiber wick exhibits large capillary action. This effect is used in heat pipes, space technology, etc.

Other Uses. Many other uses are under consideration or are being developed. A few important ones should be mentioned: catalysts and catalyst carriers, heat-exchanger media, diffusion barriers, coalescing and demisting media, gas–liquid separator media, high-temperature seals, and the heating of textiles (e.g., curtains).

4.4. Manufacturers

There are presently three major metal fiber producers: Brunswick Corporation (U.S.A.), Nippon Seisen (Japan), and N.V. Bekaert (Belgium). Not all of them produce all the various qualities and shapes available; some are integrated into the endproducts. N.V. Bekaert S.A. makes the largest variety of alloys, diameters, half-products, and shapes, as company policy is not to be integrated.

4.5. Melt Spun Fibers

4.5.1. Properties

The quality and the properties of fibers produced by most melt extrusion processes are quite satisfactory, and this is particularly the case for the amorphous or microcrystalline fibers (Fig. 41). In addition to the somewhat unusual magnetic and physical properties of the fibers, tensile strengths up to 7000 N/mm^2 have been obtained. However, poor dynamic behavior (e.g., fatigue) is still a weakness. Equivalent diameters (generally the fibers are not round) range from 25 to 300 µm, depending on the process and the melt composition.

Producing continuous lengths by melt extraction has proved to be extremely difficult, but by notching the edge, short fibers and flakes with an equivalent diameter of 20 – 200 µm can be produced. Filaments with circular cross sections are not available commercially.

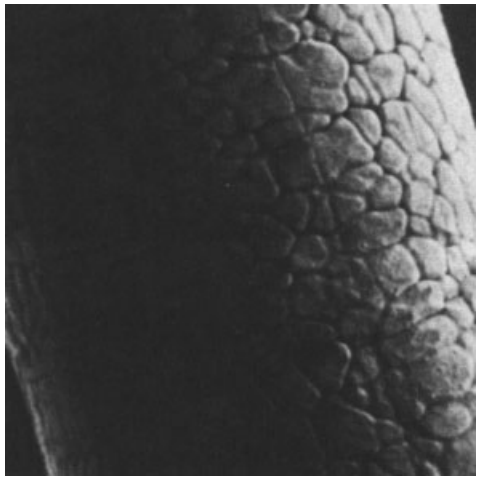

Figure 41. Surface of melt-extruded metal fiber, diameter ca. 0.36 µm (magnification ×250)

4.6. Uses

Melt Extrusion Fibers. If amorphous or microcrystalline fibers with circular cross section become commercially available, there would be an opening in those areas that already are served by amorphous ribbons: transformers and other electromagnetic equipment, shielding, sensors, etc. However, by far the biggest market would be for structural uses, especially in tire reinforcement, with a potential of $0.5 - 1.0 \times 10^6$ t/a.

Melt Extraction Fibers. Melt extraction fibers are mainly used for the reinforcement of concrete and, in particular, of refractory cements where they compete with conventionally drawn carbon and stainless steel wires. The smaller flakes also have some potential as fillers in polymers for electromagnetic interference applications, where they must compete with the bundle-drawn metal fibers.

Economic Aspects. Metal fiber prices are tending downwards as more and more high volume applications are being implemented.

5. Carbon Fibers

Carbon fibers are black fibers used as yarns, felt, or powderlike short monofilaments with diameters smaller than 10 µm. They are mainly

applied to reinforce polymers, much like glass fibers have been used for decades in fiber glass, with the principal difference that composites with carbon fibers are so-called *advanced composites* or *high-performance composite materials.* The progress achieved with carbon fibers, as compared with glass reinforcement fibers, is based on the superior stiffness of carbon fibers, combined with high strength and low density. As can be seen from Figure 42, the density of polyaramide fibers, introduced into the market nearly simultaneously with carbon fibers, is even lower, but strength and stiffness of these organic fibers do not reach near the top values for carbon fibers.

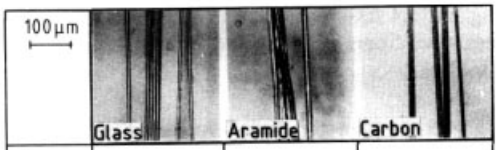

	Glass	Aramide	Carbon
ρ g/cm³	2.5	1.45	1.8 (2.1)
σ MPa	2000–4000	3000	3000–5000 (2000)
E ∥ GPa	50–80	70–130	200–500 (700)
ε %	4–5	2–4	0.5–1.9 (0.3–0.5)

Figure 42. Comparison of the reinforcing fibers for advanced composites with polymer matrix: glass, polyaramides, and carbon

Figure 43 shows the superior specific stiffness (Young's modulus measured in fiber direction divided by the density) of various types of carbon fibers in comparison with those of polyaramide and glass fibers. The extremely low values for isotropic conventional metals and the theoretical maximum for the anisotropic graphite single crystal are also included. This figure demonstrates not only the superiority of carbon fibers but also the variability of stiffness among the various carbon fiber types. In the best case, up to 80 % of the theoretical modules of a graphite layer can be realized in a commercial material.

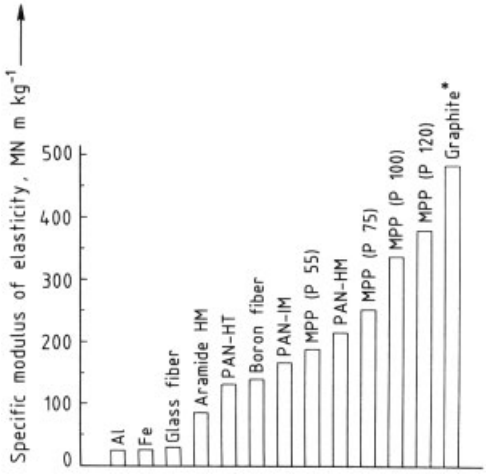

Figure 43. Specific tensile modulus (Young's modulus divided by density) of the various carbon fiber types (PAN and MPP) in comparison to other structural materials
∗ = single crystal (a-direction)

Nomenclature. Terms for description of new materials should be based on scientific characterization. Trivial terms, introduced locally, and still used, should be avoided. In UK and France the term *carbon fiber* is used correctly. In USA, however, the scientifically unjustified term "graphite fiber" is preferred in industry, and in German, as well as in Eastern European languages, the tradition of describing carbon materials as coal has been carried over to the new material carbon fiber (German: *Kohlefaser*). The IUPAC has therefore established a sub-committee for description and characterization of carbon and graphite materials in general and of carbon fibers in specific. Some tentative recommendations for the terminology [87], partly still under discussion, are the following:

Carbon fibers are fibers (filaments, yarns, rowings) consisting of at least 92 wt % carbon, regularly in the nongraphitic stage, that is, with two-dimensional long-range order of the carbon atoms in planar hexagonal networks, but without any measurable crystallographic order in the third direction (c direction) apart from more or less parallel stacking.

HM-type carbon fibers are carbon fibers with a value of the modulus of elasticity in the fiber direction larger than 30 % of the elastic number c_{11} of a graphite single crystal, but with a ratio of tensile strength to tensile modulus below 1 %.

HT-type carbon fibers are carbon fibers with a value of the tensile strength above 3000 MPa and a ratio of strength to stiffness of ca. 1.5 – 2.0 %.

IM-type carbon fibers belong to the HT type carbon fibers because of comparable values of tensile strength but are characterized by improved stiffness (Young's modulus up to 30 % of the theoretical c_{11} value with remaining values of the ratio strength to modulus above 1 %).

In addition, the precursor material of the carbon fibers is important because the combination of various fiber properties and the behavior of the carbon fiber mechanically, physically, and chemically depend strongly on the raw material.

PAN-based carbon fibers are carbon fibers obtained from polyacrylonitrile (PAN) precursor fibers by stabilization treatment, carbonization, and possibly final heat treatment at even higher temperature.

Isotropic pitch-based carbon fibers are carbon fibers obtained from isotropic pitch fibers by stabilization treatment and carbonization.

Anisotropic pitch-based carbon fibers (MPP-based carbon fibers) are carbon fibers obtained from mesogenic pitch after this has been transformed into mesophase pitch (MPP) during the process of spinning, by stabilization of the spun mesophase pitch fibers, carbonization, and final heat treatment at even higher temperature.

Rayon-based carbon fibers are carbon fibers obtained from rayon precursor fibers by chemical pretreatment and carbonization. They are isotropic carbon fibers and can be transformed into anisotropic carbon fibers with high strength and stiffness by hot stretching above 2800 K.

Gas-phase grown carbon fibers are carbon fibers grown in a hydrocarbon atmosphere with the aid of fine particulated solid catalysts such as iron or another transition metals. They consist of graphitizable carbon and can be transformed into graphite fibers by heat treatment above 2800 K.

5.1. Properties

5.1.1. Mechanical Properties

The strength values differ dramatically among the various carbon fiber types. Because carbon fibers exhibit brittle fracture behavior, the strength: stiffness ratio strictly controls the fracture elongation (strain to failure) of the fibers. In Figure 44 these relationships are demonstrated graphically in three coordinates. The values of all commercially available carbon fibers are grouped around two areas, which describe opposing combinations of properties: the high modulus domains (HM-type fiber) and the high-strength, high-tenacity domains (HT-type fiber).

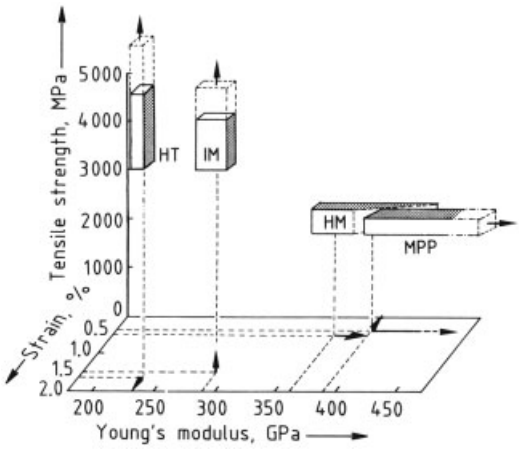

Figure 44. Characterization of the various commercial carbon fiber types by tensile strength and Young's modulus
HT = high tensile; IM = intermediate modulus; HM = high modulus; MPP = mesophase pitch-based high-modulus fibers

Nearly all carbon fibers available commercially have a near-circular cross section. As shown in Figure 45 (see next page) the cross section of the carbon fibers mirrors the fiber precursor. Currently (1986) wetspun PAN is usually the raw material. Gasphase-grown carbon fibers exhibit the best graphitic order and preferred orientation, but have no importance for reinforcement.

5.1.2. Chemical Composition

Carbon fibers consist of elemental carbon. As all synthetic carbon materials, they are fabricated from carbon compounds by thermal degradation (pyrolysis), and the final (i.e., maximum) heat-treatment temperature controls the nature and amount of remaining ligand atoms in the residual carbon. The HM-type fibers consist

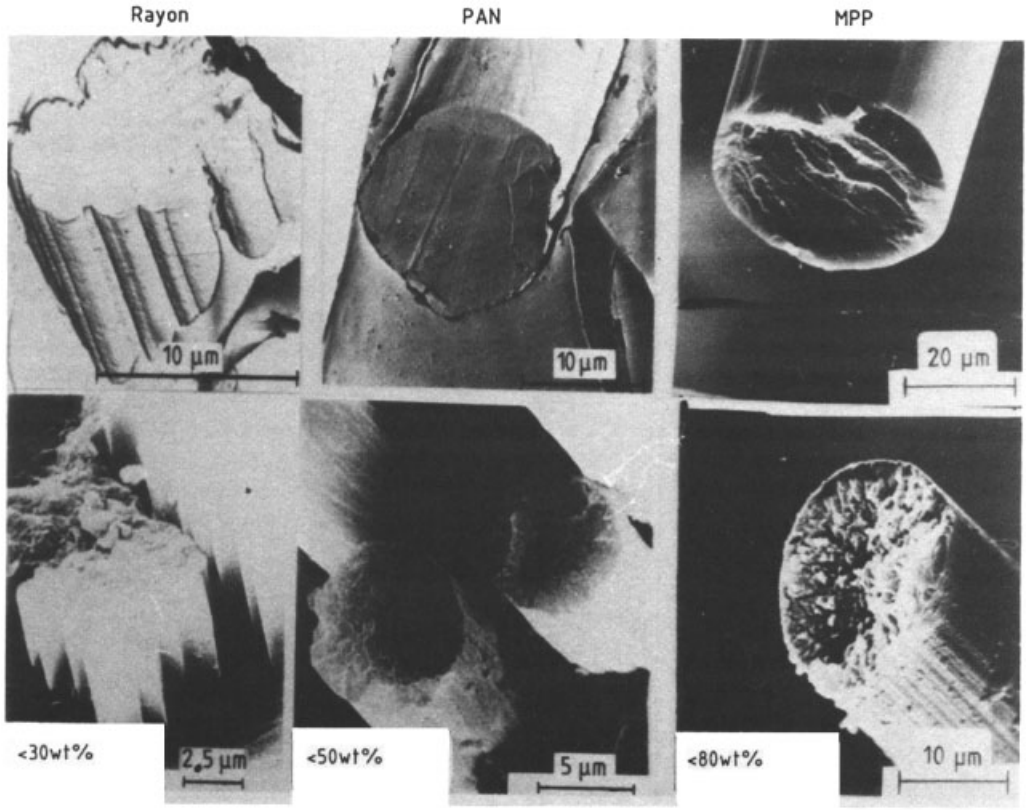

Figure 45. Fracture surface of the precursor fibers (rayon, wetspun PAN, mesophase pitch) and of the resulting carbon fibers (REM)
Carbon yield given in wt %

of nearly pure carbon; however, the HT-types contain some nitrogen (<4 wt %) and oxygen (<1 wt %). Hydrogen contents can be neglected [88, 89].

Surface-treated carbon fibers contain, in addition, solid oxygen complexes on the surface and organic coatings (sizing), which protect the carbon fibers during handling and transport.

5.1.3. Structure

Solid elemental carbon has two equilibrium modifications: diamond with isotropic homopolar bonds between the carbon atoms and sp^3 hybridization of the electron levels and the hexagonal graphite with sp^2 hybridization, thus anisotropic structure. Rhombohedral graphite is a result of mechanical deformation of the hexagonal form. The broad variety of disordered carbons belongs to the graphitic modification mainly with sp^2 hybridization and thus a tendency to form planar polyaromatic molecules but with only two-dimensional order and no hexagonal ABAB-shaped arrangement of these polyaromatics in the third dimension. In this sense carbon fibers can be described as disordered carbon with more or less ordered polyaromatic layers and a hypermolecular anisotropic arrangement of these polyaromatics parallel to the fiber axis. Perfectness, size, and arrangement of these polyaromatics control the physical and especially the mechanical properties of the fibers. Partial sp^3 hybridization was confirmed in all artificial carbon, and especially in carbon fibers, and is explained by the carbon atoms on the edges and in disclinations of the polyaromatics [90, 91].

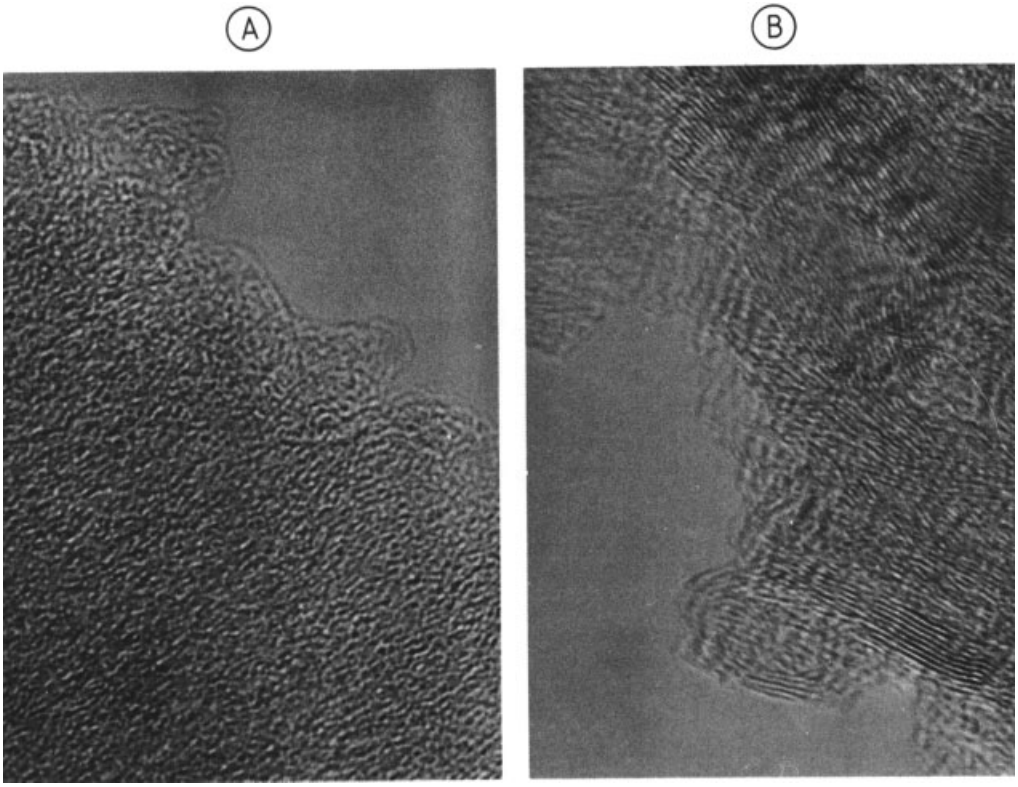

Figure 46. TEM bright field images (cut perpendicular to the fiber axis) [90, 92]
A) Typical commercial HT type (T 300) B) HM type (SIGRAFIL HM) carbon fiber

Transmission electron microscopy (TEM) produces bright field images of thin-layer cuttings of HT- and HM-type carbon and reveals the different degree of defect structure (Fig. 46) [90, 92].

The preferred orientation in the direction of the fiber axis as indicated by the Ruland parameter q (0 corresponds to a statistic arrangement and −1.0 means completely preferred orientation) is shown in Figure 47 [93]. Some experimental fiber types based on mesophase pitch reach nearly 100 % of the theoretical Young's modulus, i.e., that of a single crystal. Such a preferred orientation is combined, however, with the disadvantage of low shear resistivity.

The layer distance, which is easily measured, is a good indication of the degree of defect structure of a carbon fiber. Figure 48 shows the values for commercial fibers as received and the reduction of the $\bar{c}/2$ value by further heat treatment, which heals structural defects [94].

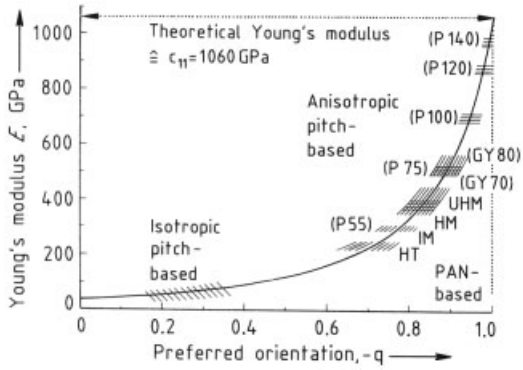

Figure 47. Young's modulus of the various types of carbon fibers correlated with the degree of preferred orientation measured by X-ray diffraction and expressed by the Ruland parameter [93]
0 = statistical distribution, 1.0 = orientation of graphite single crystal

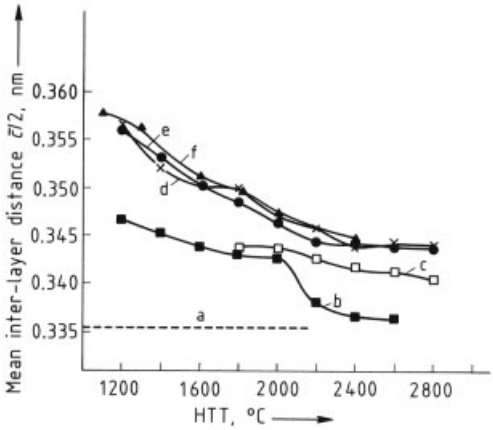

Figure 48. Main interlayer distances $\bar{c}/2$ of various commercial carbon fibers and their decrease due to subsequent heat treatment. The $\bar{c}/2$ values of graphitizable petroleum coke are included for comparison [92, 94, 95]
a) Ideal graphite; b) Petroleum coke; c) P55, MPP-based; d) AS4, PAN-based; e) T300, PAN-based; f) T800, PAN-based

This healing process results in higher density and increased Young's modulus, which is shown in Figure 49B. The strength passes a maximum at heat-treatment temperatures around 1500 °C (Fig. 49A). The reduction of strength, with the minimum around 1800 °C, is caused by the escape of the remaining ligands, which causes severe structural disturbance. Only minor improvement takes place during further annealing. The variety of curves for PAN-based fibers is caused by differing oxidative pretreatment during the stabilization before carbonization [96, 137].

Figure 50 (see next page) shows the effects on the ultrastructure of heat treatment. The ordering process at the highest temperatures is easy to recognize for a commercial HT fiber by the enlargement of domains with parallel layer stacking (Figs. A and B). Such improvement of strength by annealing is not found in the case of pitch-based fibers with superior preferred orientation because of increased shear sensibility by annealing (Figs. 50 C and D). The much higher tendency of mesophase pitch-based fiber towards structural rearrangement can also be seen in Figure 50.

Based on electron microscopic studies, several models for structure in carbon fibers have been proposed. Four of these are shown in Figure 51 [97–99]. In all cases, parallel stackings of polyaromatic layers are arranged in the fiber direction in forms such as ribbons,

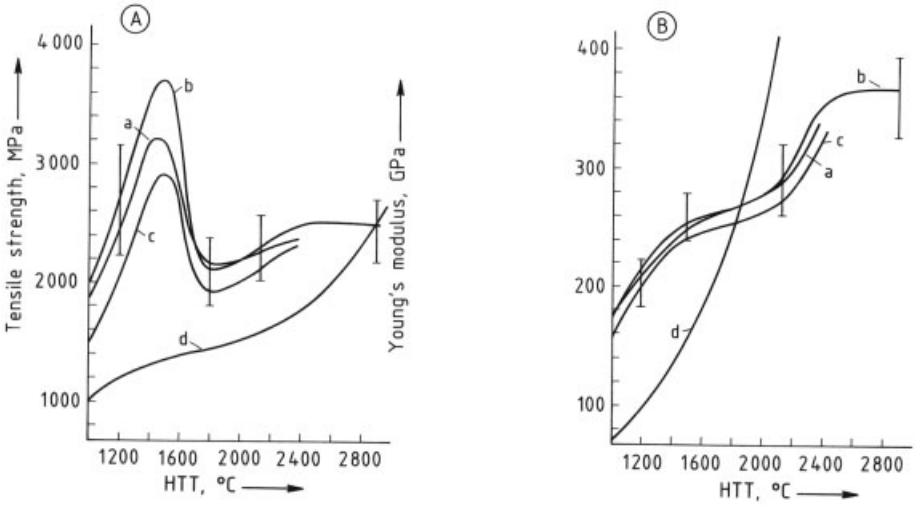

Figure 49. The effects of heat posttreatment of carbon fibers [96, 137]
A) Tensile strength
SAF, stabilization treatment a) 230 – 260 °C/40 min; b) 230 – 270 °C/40 min; c) 230 – 290 °C/40 min; d) MPP
B) Young's modulusSAF, stabilization treatment a) 230 – 260 °C/40 min; b) 230 – 270 °C/40 min; c) 230 – 290 °C/40 min; d) MPP
The effect of varied heat pretreatment during stabilization is included. For all PAN-based fibers, SAF precursor was used [137]. For comparison [96], the behavior of MPP carbon fibers is included.

Table 16. Physical properties of typical commercial PAN-based carbon fibers.

	Density, g/cm^3	Electrical resistance, 10^{-3} Ω m	TEP, μV/K	Susceptibility χ ∗, 10^{-6} emu/g		Specific heat Cp, J/gK	Thermal conduct. λ, W/mK
				⊥	‖		
PAN-based HT	1.76 – 1.81	1.5 – 2.4	−1.6 to −2.4	−0.3	−0.05	0.71	8 – 20
PAN-based HM	1.82 – 2.00	0.6 – 1.0	17.2	−6.6	−0.63	0.71	60 – 100
Anisotropic pitch-based	2.00 – 2.15	0.4 – 1.1	22.6	−9 to −10.4	−0.70	0.93	110 – 375
Isotropic pitch-based	1.60	5 – 10		−0.74 to −5.32	−0.64 to −4.26		15

∗ Perpendicular (⊥) or parallel (‖) to fiber axis

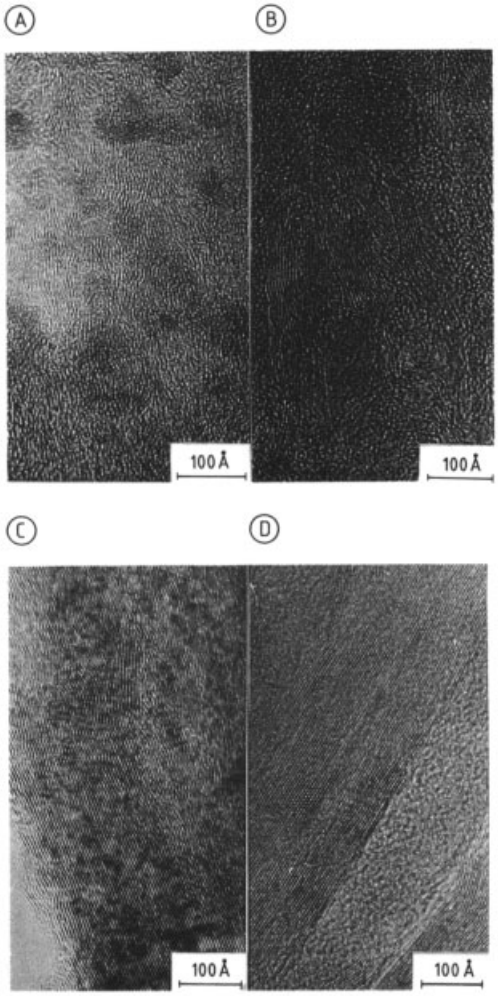

Figure 50. Effect of high-temperature heat treatment on the ultra structure and especially on the enlargement of the domains with parallel stacking of layers (TEM bright field of cuts parallel to the fiber axis)
A) PAN-based HT fibers (AS4W) as received; B) The same after-treatment at 2400 °C; C) MPP-based fiber (P55) as received; D) The same after treatment at 2440 °C

plates, hoses, tubes, or combinations of the various structural elements.

5.1.4. Physical Properties

Table 16 shows the densities, electrical resistivities, thermoelectric powers (TEP), magnetic susceptibilities, heat capacities, and thermal conductivities of the various carbon fiber types.

The reversible thermal expansion is highly anisotropic (Table 17). In PAN-based HT-type carbon fibers high anisotropy is also found, as is the case for graphite single crystals. In contrast, the reported anisotropy in a mesophase pitch-based HM-type fiber is much smaller [100]. Experimental measurement for the direction perpendicular to the fiber axis, however, is very difficult, and a comprehensive study is still not available.

Table 17. Reversible thermal expansion of some commercially available carbon fibers.

	CTE $\alpha_{\parallel}$, 10^{-6} K^{-1}	CTE $\alpha_{\perp}$, 10^{-6} K^{-1}	$-\alpha_{\perp}/\alpha_{\parallel}-$
Graphite, single crystal	0.5 – 1.0	29	29 – 58
PAN-based HT	0.6	5 – 30	8.3 – 50
PAN-based HM	1.0	15	15
Anisotropic, pitch-based	0.9 – 1.4	6	4.3 – 6.7
Isotropic, pitch-based	3	3	1

Typical mechanical properties of carbon fibers are listed as a part of Figure 44.

5.1.5. Chemical Properties

The chemical resistivity of carbon and graphite materials against aqueous solutions and most

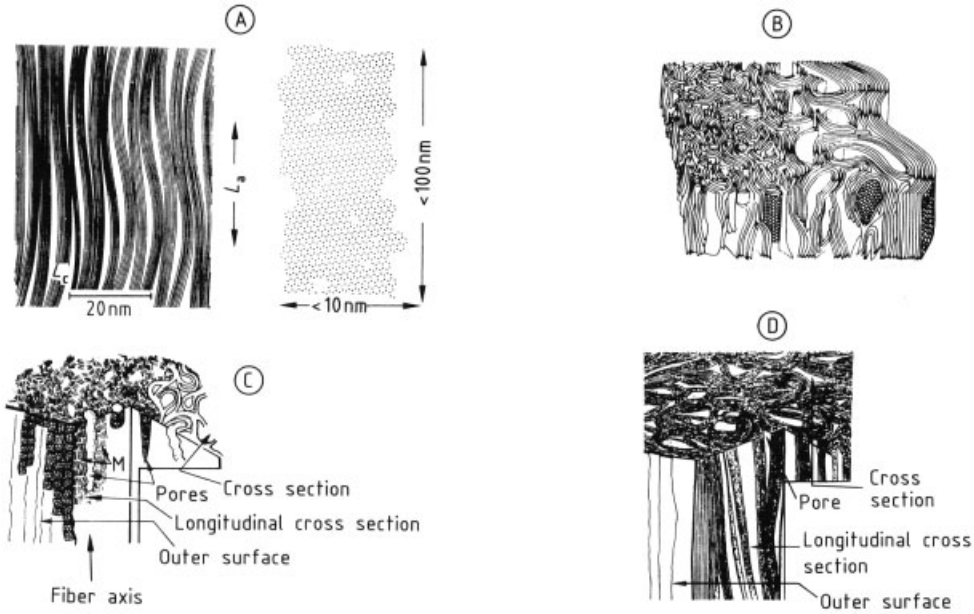

Figure 51. Structural models of carbon fibers
A) Two-dimensional model proposed by RULAND [97] (ribbon model); stretched graphitized rayon-based fibers B) Three-dimensional model proposed by JOHNSON [98] (plate/sheet model); PAN-based fibers C) Ring/tube model proposed by OBERLIN [98] for all carbons resulting from thermal degradation of carbonizable solids [99] D) Hose-shaped layer stacking model proposed for highly heat-treated carbon fibers [99]

chemicals at ambient temperatures is well-known and has been utilized in chemical apparatus. The behavior of carbon fibers is similar. Only fluorine acids as well as strong oxidants attack carbon fibers strongly. The tendency of graphite to form intercalation compounds with halogens, salts, oxidizing acids, and some other substances is also well-known. The majority of carbon fibers, which consist of nongraphitizable carbon is less susceptible to intercalation attack. Only gasphase grown fibers can be intercalated easily.

Most destructive for carbon fibers are their reactivity with metals, above all aluminum at temperatures below 1000 K, and their reactivity with solid oxides at temperatures above 1000 K. This reactivity limits the use of uncoated carbon fibers for reinforcement of metals and oxides. However, if reaction and diffusion barriers are built in – silicon carbide, titanium nitride, and similar coatings have proved to be successful barriers (Fig. 52) –carbon fibers can be used to reinforce metals and oxides [101, 102].

Figure 52. SiC-coated carbon fibers in a SiO_2 glass matrix (parallel to the fibers) [101]

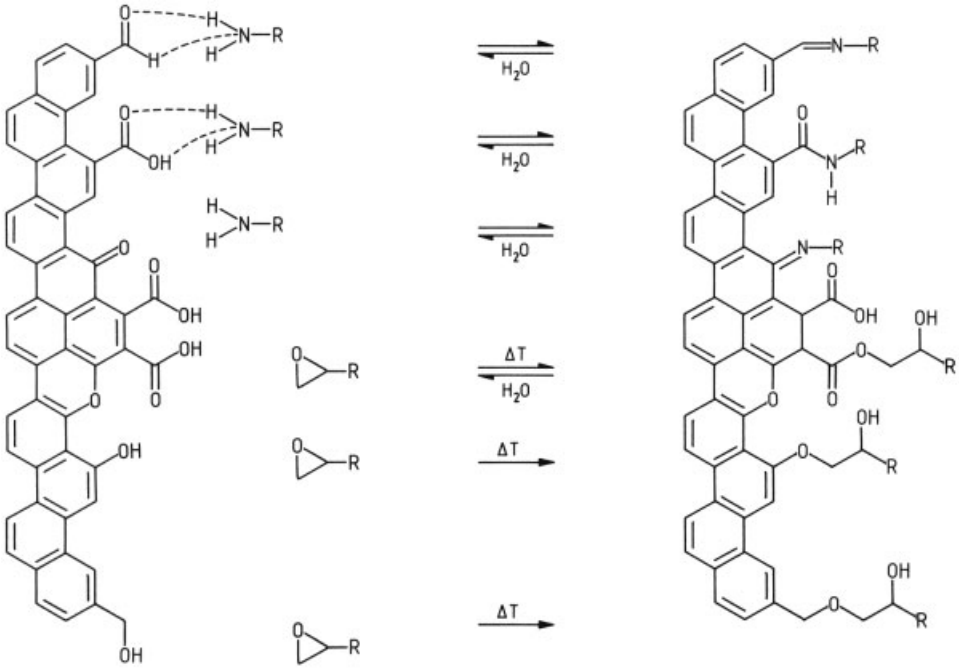

Figure 53. Oxide complexes on surface-treated carbon fibers and their affinity to polar groups in epoxy resins and curing agents [103]

Solid surface oxides form as an intermediate stage before formation of volatile carbon oxides. Such solid surface oxide complexes improve the adhesion of carbon fibers in polymers and matrix in a composite [105, 106]. The thermal stability of surface oxides is limited to temperatures below 800 K for carboxylic groups and below 1000 K for nonacidic surface oxides. Figure 53 [103] shows the various types of surface oxides and their affinity to polar groups in the resins, which initiates at least chemisorptive adhesion and perhaps even formation of extensive chemical bonding between fiber and resin.

The degree of adhesion is measured by interlaminar shear strength (Ilss). Figure 54 shows the effect of stepwise desorption of the surface

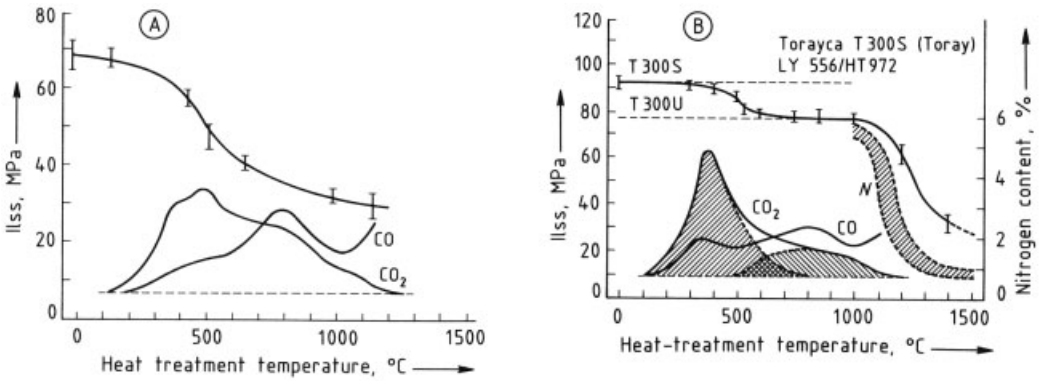

Figure 54. Gases (CO_2, CO) liberated during thermal desorption of surface-treated carbon fibers [89]
A) HM type, Sigrafil HM12S (Sign), 120 h oxidized in 68 % HNO_3
B) HT type, showing loss of nitrogen, content of PAN-based carbon fibers above 1000 °C
a) T300S; b) T300U; S: surface treated; U: untreatedIlss = interlaminar shear strength

oxides in surface-treated carbon fibers by heat treatment in argon and the resulting effects of such stepwise desorption on the remaining adhesion in an expoxy matrix. The strong improvement of adhesion in surface-treated HM fibers is associated with the carboxylic groups. The improvement of adhesion in HT-type fibers by surface treatment is much smaller, a maximum of 30 % of the total adhesion. The remaining adhesion in HT-type fibers is strongly correlated with the nitrogen content in the fiber. The adhesion of the fibers decreases abruptly if they are heat-treated above 1200 °C, which drives off the nitrogen remaining from the PAN precursor [88].

The controlled formation of surface oxides can be attained by three routes:

1) Thermal oxidation at temperatures above 700 K
2) Wet oxidation in boiling nitric acid or in a similar oxidative liquid
3) Anodic oxidation in aqueous solution

Anodic oxidation is the only method used industrially for surface oxidation of carbon fibers. This process is the easiest to control and forms the greatest amount of surface oxides (Fig. 55).

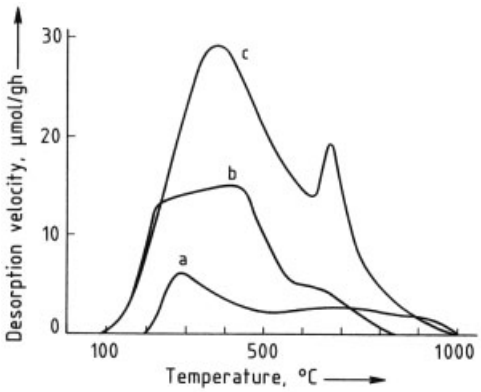

Figure 55. Amount of surface oxides formed by thermal, wet, and anodic surface treatment of HM fibers as measured by weight loss during thermal desorption in argon [104]
a) Thermal oxidation; b) in HNO_3 oxidized; c) anodic oxidation

Thermal oxidation, although simple to do, is dangerous because of the rapid reaction rate and the danger of fiber damage due to total oxidation. Figure 56 shows in an Arrhenius plot the experimental weight loss on oxidation of carbon fibers in air [104]. The HT-type fiber, which is given a maximum heat treatment at 1450 °C during fabrication, is much more sensitive to total oxidation than the HM-type fibers, which are heat treated above 2500 °C during fabrication. Total oxidation can be catalyzed even by alkali-metal compounds, as is the case for carbon and graphite materials. Heavy metal oxides are effective catalysts. Experimental results on the effect of copper oxide and vanadium oxide on the relatively resistant HM-type fibers are included in Figure 56. As a consequence, metallic impurities in carbon fibers have to be strictly avoided during industrial production.

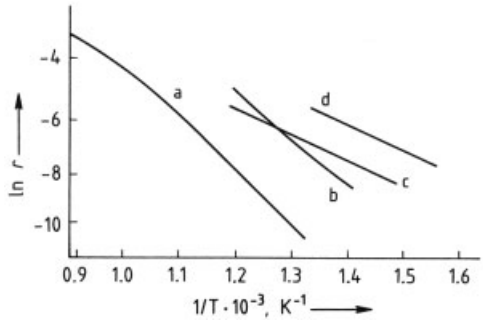

Figure 56. Thermal oxidation rates of HM- and HT-type carbon fibers (Arrhenius plots), also showing the catalytic effects of heavy metal oxides on the oxidation of HM fibers [104]; r = rate constant
a) HM (Sigrafil); b) HT (Toho); c) HM + CuO; d) HM + V_2O_5

5.2. History

Public disclosures of practical processes for converting textile fibers into useful carbon fibers have been confined largely to the patent literature.

5.2.1. Rayon-Based Carbon Fibers

For rayon as the precursor material, a comprehensive survey was published by R. Bacon [107]. One hundred years ago, Edison [108] described an incandescent-lamp filament made by carbonizing a filament of natural and regenerated cellulose material. In the 1950s Abbott [109] developed a process for converting

rayon into a fibrous carbon material for insulation, filtration, and absorption. This material was carbonized at temperatures up to 1000 °C and possessed tensile strengths of only 280 MPa. In 1959, Union Carbide [110–112] started commercial production of similar fibers, and other companies in the U.S.A. followed.

Various fiber types heat-treated at final temperatures around 1200, 1800, and even above 2500 °C have become commercially available, e.g., the carbonized VIB type and the WYB type heat-treated at graphitization temperatures, both from UCC (Union Carbide). The tensile strength of such fibers ranged from 300 to 700 MPa, while their moduli were between 30 and 50 GPa. The most important application was the reinforcement of phenolic resins to form ablative materials for rocket and missile components. This early period of carbon fiber production and application is reviewed by Schmidt and Hawkins [113].

In these first years of the carbon fiber industry, several companies introduced a pretreatment process for the precursor fibers to increase carbon yields and speed up the processing rates, especially in the early heat treatment stage. The Pluton fiber of 3 M [114] was introduced in 1961. A decisive improvement of the mechanical properties of the carbon fibers was the plastic deformation of the fibers at temperatures above 2500 °C. This UCC process [115] led to the commercial production of the Thornell type, which had in 1965 a tensile strength of 1260 MPa and a modulus of 175 GPa (Thornell 25).

This was the period when the US Air Force Material Laboratory in Dayton, Ohio, strongly supported further development in the field of high-modulus carbon fibers [116, 117]. The goal was to replace boron fibers as soon as it proved possible to fabricate high-performance carbon fibers with a minimum modulus of 350 GPa. In 1970, UCC reached the goal with the commercially available Thornell 75 HM-type carbon fiber, which also had a Young's modulus of 500 GPa, a value corresponding to 50 % of the theoretical c_{11} value. Hitco [118] entered the market with yarns produced by similar methods. Although these carbon fibers were satisfactory in applications, the final decision of the market was to restrict use to the PAN-based carbon fibers.

5.2.2. PAN-Based Carbon Fibers

The history of PAN-based carbon fibers starts in 1950 with the observation of R. C. Houtz [119] that polyacrylonitrile fibers are converted into black, flameproof materials by heat treatment at 300 °C in air. However, products produced by this method were not used before 1959 [120]. At that time the discovery was made that flexible, flameproof textiles with the original fiber structure can be obtained only if the exothermic reaction between 150 and 300 °C is strictly controlled and if the high shrinkage during this thermal treatment is minimized by mechanical means. In 1960 the process to produce a flameproof fiber, starting with Dupont's AF Orlon fiber and commercialized as Black Orlon, was described in detail by Vosburgh [121]. Heat treatment up to 15 h at 220 °C was needed to produce a well stabilized fiber with a strength of 40 % of the value of the original PAN fiber. The oxygen content of the final product was 10 – 20 wt %.

Surprisingly, this oxidizing-blackening treatment of the PAN fibers was not complemented by carbonization. Such a combination was first claimed by Shindo [120] in 1959. A maximum tensile strength of 100 MPa was obtained. In 1965, Johnson, Philip, Watts [122] added to this combined process the need to restrain shrinkage or alternatively to stretch the PAN fiber during stabilization treatment. They obtained carbon fibers with Young's moduli up to 420 GPa. This process is now the standard process for fabrication of PAN-based carbon fibers, although patent protection turned out to be limited and is completely gone today. A more detailed history is given by Watts [123], and a complete literature survey is given by D. J. Mueller [124].

5.2.3. Pitch-Based Carbon Fibers

The processes for carbon fibers from pitches are based on the observation of Otani [125] in 1963 that some pitches are spinnable to pitch fibers, which can be converted into carbon fibers. Pitch from poly(vinyl chloride) turned out to be the best, but a variety of other pitches can be used after specified pretreatment as summarized by Sugio Otani and Asao Oya [126]. The mechanical properties of the resulting carbon fibers were

in the range 800 – 1800 MPa for tensile strength and 20 – 50 GPa for Young's modulus. These pitch-based carbon fibers were commercialized in 1970 by Kureha Chemical Industries and are still the only ones commercially available in the form of continuous yarns [127]. In 1985, FREEMAN [128] reported that the Ashland Company started production of a variety of carbon fiber types using petroleum pitch as the raw material. Coal tar pitch is recommended as an alternative precursor material by ROMMEY [129]. All these pitch-based carbon fibers belong to the isotropic, low-modulus type.

Fabrication of anisotropic pitch-based carbon fibers with greatly improved mechanical properties was achieved by LEWIS and SINGER et al. [130] in 1976. This process is based on the mesophase, a liquid crystal formed within the pitch before spinning that enables the formation of a preferred orientation in the pitch fiber, thus also in the final carbon fiber. Values of tensile strength up to 3000 MPa and an extremely high Young's modulus, up to 600 GPa, can be achieved.

Laboratory developments have continued in many countries of the Western World. Commercial products are available from AMOCO (previously Union Carbide) [131].

5.2.4. Vapor-Grown Carbon Fibers

Whiskers and needles of carbon deposited from the gas phase were described in 1966 by FITZER and SCHLESINGER [132]. FITZER and RHEE [133] continued these studies and reported in 1970 growth rates and mechanical properties of the whiskers and needles. Tensile strengths up to 1500 MPa and stiffness values up to 300 GPa were reported. The decisive progress was reported by ENDO [134] in 1974 when he recognized that the whisker growth follows a catalytic mechanism starting from fine, dispersed solid metal particles in the substrate. Whisker lengths up to 250 mm and tensile strength values up to 26 000 MPa have been reported so far by ENDO [134]. Commercialization has not been announced, however, probably because of the completely different structure of vapor-grown carbon fibers. These consist of soft carbon, that is, the graphitizable less-disordered variety. Therefore, high mechanical properties cannot be expected. They are most suitable for intercalation (high electrical conductivity).

5.3. Production

All commercial production processes for carbon fibers are based on carbonization of polymer fiber precursors. The preconditions for useful fiber formation compiled in Table 18 must be fulfilled. Thus, the number of suitable polymer fibers is limited. The fiber form is always preserved during carbonization in the case of nonmeltable polymers, such as thermosetting resins, which can be thermally degradated directly. Thermoplastic precursor fibers have to be converted into a nonmeltable form before carbonization. This is called stabilization treatment and occurs on conversion of the linear polymers into ladder polymers and on formation of additional cross-linkages. In the case of pitch precursor fibers, the melting of the mixture of polyaromatic compounds must be hindered by oxidation. Three precursor materials are satisfactory: (1) rayon fibers, (2) PAN fibers, and (3) pitch fibers. Rayon fibers do not require stabilization, only preimpregnation is necessary to improve the carbon yield. PAN and pitch fibers need sta-

Table 18. Restrictions on the precursor fiber for carbon fiber fabrication

Preconditions for carbon fiber manufacture	Parameters	
	Precursor	Process
Preservation of the fiber form of the polymer	thermosetting or thermoplast fibers	direct thermal degradation or stabilization treatment of meltable precursor before thermal degradation
High carbon yield	selection of the polymers	optimization of the stabilization and carbonization parameters
Preferred orientation of the polycarbon layers with highest degree of defects in the ultra structure	highly cross-linked, unmeltable polymers	stretching of the precursor before or during stabilization and/or hot-stretching of the carbon fiber

bilizing pretreatment. The carbon yield is lowest for rayon fibers and highest for stabilized pitch fibers:

Rayon fibers	25 %
Oxidized pitch fibers	80 %
Polyacrylonitrile	50 %
Oxidized pitch	up to 80 %
Phenolic and furan resins	up to 85 %

The most important parameter in producing high-quality carbon fibers is the preferred orientation of the polyaromatic layers parallel to the carbon fiber axis. Only then can the high strength of the bonds in the molecular structure of the graphite lattice (c_{11}= 1060 GPa, c_{33}= 33 GPa) be transferred into stiffness and strength of the carbon fiber. The low shear modulus in the graphite crystal structure (c_{44}= 4.6 GPa) must be avoided by having the highest degree of defects in the ultrastructure of the carbon fiber. In the case of rayon precursor fibers the needed high degree of structural imperfectness can be realized, however the preferred orientation is realizable only by hot working of the carbon fiber at 3000 K. This hot stretching process is difficult to carry out on an industrial scale. Therefore, carbon fibers are no longer produced from rayon.

In PAN- and pitch-based anisotropic carbon fibers, the preferred orientation is introduced into the precursor before carbonization. Polyacrylonitrile is the most important precursor material. More than 90 % of all commercially produced carbon fibers that are used for reinforcement of polymer material are based on this polymer. Only a minor part is produced from pitch.

5.3.1. PAN Precursor Fiber

Textile PAN fibers usually consist of ternary mixtures of copolymers with 89 – 95 wt % acrylonitrile (AN), 4 – 10 wt % of a non-ionogenic comonomer plasticizer, and 1 wt % of additive to chemically bond dyes [135]. For the precursors for carbon fibers, a lower content of comonomers, usually below 3 wt %, is used.

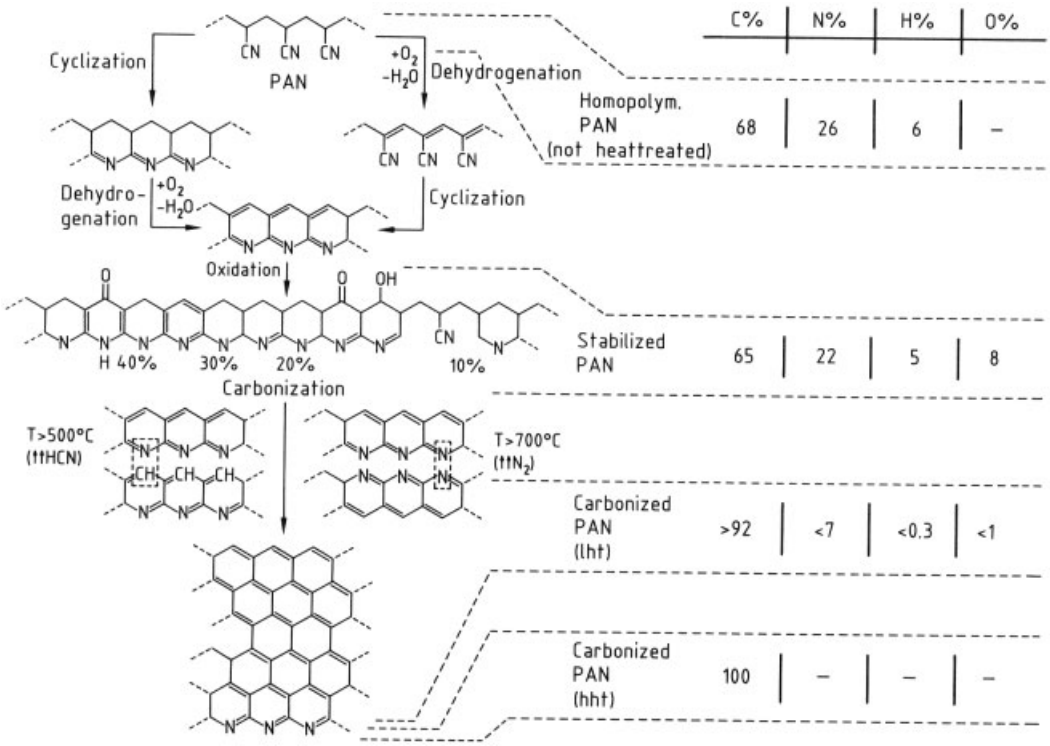

Figure 57. The chemistry of stabilization and carbonization: fabrication from homopolymer PAN and stabilized PAN [137, 138, 147]

Polymer fibers can be spun by melt, wet, and dry spinning processes. In the case of polyacrylonitrile as a carbon fiber precursor, the wet spinning process is the most important. In 1986 a new melt spinning process was presented which would offer economic advantages [136].

The solvents dimethylformamide (DMF) and dimethylacetamide (DMA) are usually preferred for wet spinning today. For carbon fiber precursors, however, the inorganic solvents nitric acid, stannum chloride and thiocyanate are also used. In any case, copolymer composition of the precursor, the spinning technology, and especially the solvent are almost always strict commercial secrets. Only Courtaulds (United Kingdom) describes the use of dimethyl sulfoxide for fiber precursor fabrication.

The main difference between textile PAN fibers and PAN precursor fibers is the different degree of stretching: the stretching ratio for textile fibers is 1 : 2 to 1 : 10, that for PAN precursor fiber 1 : 12 or more. There is the danger of overstretching, which introduces defects into the polymer fiber that initiate defects in the carbon fibers, thus causing low strength.

The extremely high-strength and high-tenacity carbon fibers are mainly produced from precursors more carefully spun and stretched to a higher degree. Dry-jet wet spinning would be a good method for this.

5.3.2. Stabilization

Polyacrylonitrile is a thermoplastic polymer that decomposes above 370 °C. A suitable stabilization process is needed before carbonization to obtain an infusible intermediate that preserves the orginal fiber structure. This intermediate is known as inflammable textile (e.g., PAN-Ox).

The stabilization step is the most important one from the viewpoint of carbon fiber quality and process economy. The reaction is diffusion controlled, and a residence time of ca. 1 h at temperatures between 200 and 280 °C is needed. Fiber shrinkage during stabilization must be avoided (see Section5.2.2).

The chemistry of the stabilization process is illustrated schematically in Figure 57 (see previous page). The thermoplastic chain is cyclized to a naphthyridine ring system and simultaneously dehydrated in the backbone to form conjugated double bonds. The resulting aromatic polymer has ladder units of about five, corresponding to the average number of isotactic sequences in the nitrile polymer. Cyclization can occur with or without simultaneous dehydrogenation as shown in Figure 58 [137]. In both cases strong exothermic reactions are observed. In industrial processes, cyclization is always carried out in an oxidizing atmosphere because of the much higher carbon yield during the subsequent carbonization [139].

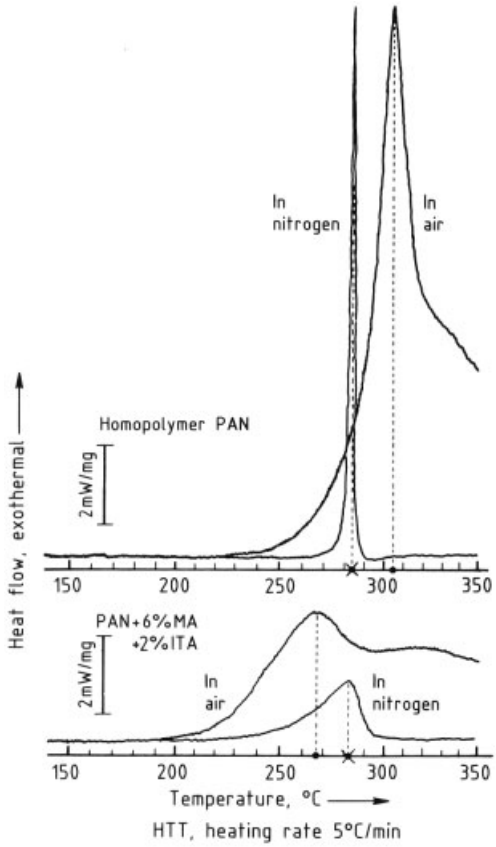

Figure 58. The exothermic stabilization in nitrogen (cyclization) and in air (cyclization and dehydrogenation) as revealed by DSC (differential scanning calorimetry) [137]

There is a direct correlation between the amount of oxygen absorbed and the density of the stabilized fiber. The best mechanical properties result for a precursor density of 1.4 – 1.42 g/cm^3, which corresponds to oxygen contents of 10 – 14 wt % (Fig. 59).

The stabilization causes the color to change from the white of polyacrylonitrile to yellow

and brown and finally to the black of the intermediate, the oxidized, inflammable polyacrylonitrile. The stabilization chemistry has been treated comprehensively in numerous papers, e.g., [137, 138, 140–147].

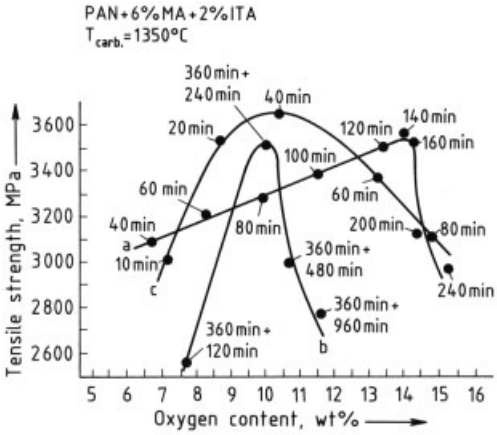

Figure 59. The effect of oxygen uptake during stabilization on the strength of the final carbon fiber
a) 230 °C isotherm/air; b) 230 °C isotherm, 1st stage N_2, 2nd stage air; c) 230 → 270 °C nonisotherm/air

5.3.3. Process Parameters

The carbonization step can be carried out much faster than the stabilization step; however, the evolution of the volatile byproducts, mostly H_2O, HCN, CO_2, and N_2, must be taken into account [148].

Practical temperature – time profiles have been published [137, 142, 149–151]. The quality of the final carbon fiber product is controlled by the precursor fiber, the stabilization treatment and the oxygen content, and finally by the carbonization schedule.

An example of a continuous stabilization and carbonization for fabrication of carbon cloth from PAN from patent literature is shown in Figure 60 [152].

5.4. The Pitch Fiber Process

5.4.1. Isotropic and Anisotropic Pitch Fiber Precursors

Carbon fibers made from pitch precursor are available with isotropic or anisotropic structure, the former made from isotropic pitch, the later from mesophase pitch. The isotropic pitch-based carbon fibers have low mechanical properties. They have been produced industrially from coal tar pitch by Kureha (Japan), for more than 15 years. The petroleum pitch producer Ashland (United States) manufactures isotropic carbon fibers, which are used mostly for thermal and mechanical insulation but also as fillers.

Anisotropic carbon fibers, which can be used for reinforcement because of their excellent mechanical properties, are produced in a process that takes advantage of the larger planar polyaromatic molecules in pitch and their tendency to form liquid crystals with parallel polyaromatic layers. The mesophase pitch is melt spun and stretched in the plastic deformable stage to produce pitch fibers with preferred orientation.

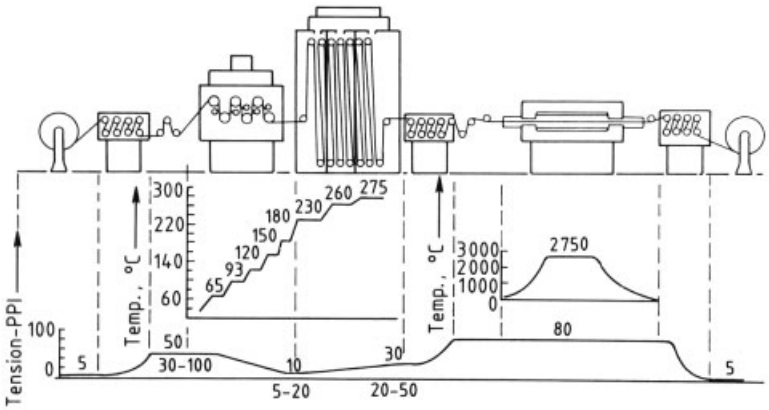

Figure 60. Hitco process for production of carbon cloth from PAN cloth showing temperature and tension profiles [152]

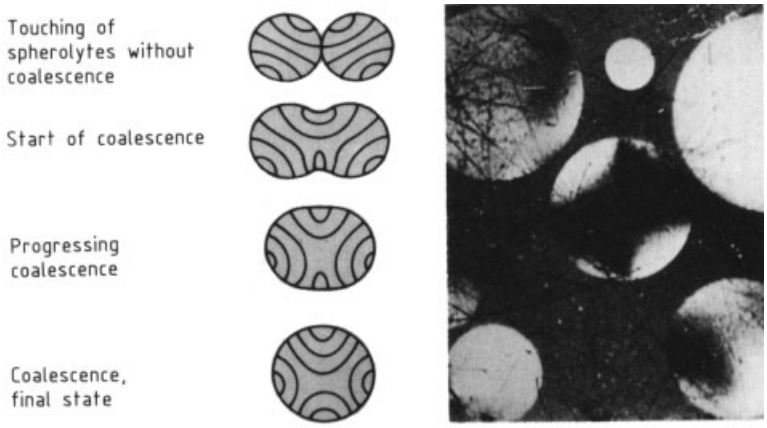

Figure 61. The coalescence of discotic liquid crystals in pitch

All pitch fibers must be stabilized before carbonization, but pitch fibers with preferred orientation consisting to a large extent of mesophase not only hinder melting during carbonization but also cross-link the polyaromatics upon oxidation thus hindering formation of graphitized carbon fibers.

5.4.2. The Mesophase Formation

Pitch is the residue from distillation of coal tar or crude oil or from cracking hydrocarbons. All pitches consist of thousands of aromatic hydrocarbons along with more or less cyclic aliphatics and aliphatic side chains. The molecular mass of the polyaromatic compounds ranges between 200 and 800. All the compounds in a single pitch have never been identified. Pitches are characterized by fractionation: their solubility in a sequence of standard solvents (Table 19). The model molecules shown in the table are suggested by comprehensive analytical investigations [153]. The melting points of the various pure compounds cover a range of hundreds of degrees. Pitches, however, form multi-eutectic systems with softening temperatures between 50 and 300 °C, much lower than the melting temperatures of the pure aromatic components.

The mesophase is formed during thermal degradation in the temperature range 400–450 °C and is controlled by the hydrogen content remaining in the aromatics during this heat treatment. Aromatics with an average molecular mass around 800 (coal tar pitches) and aromatics with aliphatic side chains and a molecular mass around 1000 (petroleum pitches) are able to form nuclei for discotic liquid crystals [153].

The liquid crystals in pitch are spherical with the molecular orientation shown in Figure 61. The coalescence of these liquid crystals with disklike molecules is similar to that of the liquid crystals with rodlike molecules.

From the viewpoint of the phenomenology, the formation of liquid crystals is similar to that of crystallization, starting from observable nuclei with diameters larger than 1 μm. The growth and coalescence proceed until all isotropic pitch matrix is consumed. Insoluble impurities in the molten pitch matrix (primarily quinoline insolubles, i.e., the pitch residue) inhibit coalescence and therefore also the formation of large volumes of anisotropy.

5.4.3. Process Steps

The viscosity of pitch, a precondition for satisfactory melt-spinning behavior, is determined by pitch composition. Figure 62 shows the apparent viscosities of a typical petroleum pitch and a typical coal tar pitch as functions of heat treatment [154]. The behavior of toluene-soluble (TS) and toluene-insoluble (TI) fractions of the coal tar pitch and of two mixtures of TS and TI fractions are included. Softening of the pitches upon heating is indicated by viscosity decrease; solidification due to mesophase formation, by viscosity increase. Spinning is difficult for

Table 19. Characterization of typical coal tar and petroleum pitches by solvent fraction and model molecular structures for the fractions [153]

wt %	Coal tar pitch		wt %	Petroleum pitch	
52.2	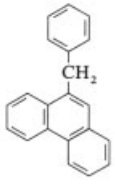	MW = 268 C/H = 1.31 $\frac{H_{arom}}{H_{aliph}} = 7$ $\frac{C_{arom}}{C_{aliph}} = 20$	67.1	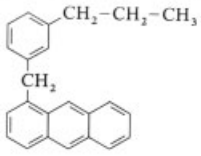	MW = 306 C/H = 1.08 $\frac{H_{arom}}{H_{aliph}} = 0.73$ $\frac{C_{arom}}{C_{aliph}} = 25$
7.8	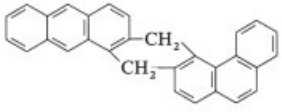	MW = 380 C/H = 1.5 $\frac{H_{arom}}{H_{aliph}} = 4$ $\frac{C_{arom}}{C_{aliph}} = 14$	11.0	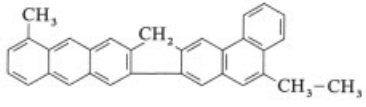	MW = 408 C/H = 1.33 $\frac{H_{arom}}{H_{aliph}} = 1.4$ $\frac{C_{arom}}{C_{aliph}} = 7$
1.6	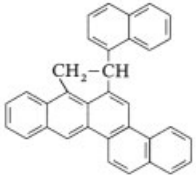	MW = 429 C/H = 1.6 $\frac{H_{arom}}{H_{aliph}} = 6.3$ $\frac{C_{arom}}{C_{aliph}} = 16$	9.5	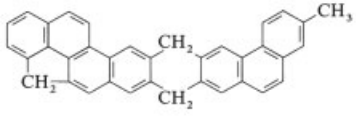	MW = 456 C/H = 1.5 $\frac{H_{arom}}{H_{aliph}} = 1.66$ $\frac{C_{arom}}{C_{aliph}} = 8$
5.8	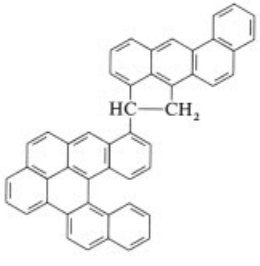	MW = 602 C/H = 1.48 $\frac{H_{arom}}{H_{aliph}} = 7.6$ $\frac{C_{arom}}{C_{aliph}} = 23$	2.0	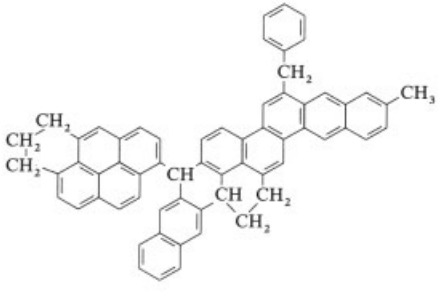	MW = 812 C/H = 1.45 $\frac{H_{arom}}{H_{aliph}} = 1.59$ $\frac{C_{arom}}{C_{aliph}} = 6$

Table 19. Continued.

wt%	Coal tar pitch	wt%	Petroleum pitch
3.4	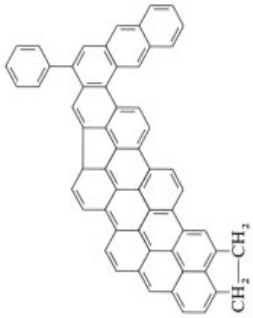 MW = 737 C/H = 2.03 $\frac{H_{arom}}{H_{aliph}} = 6.3$ $\frac{C_{arom}}{C_{aliph}} = 29$	2.6	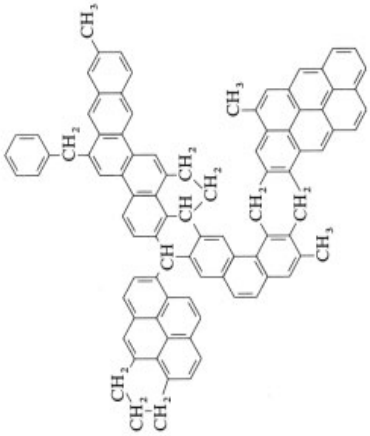 MW = 1203 C/H = 151 $\frac{H_{arom}}{H_{aliph}} = 1.10$ $\frac{C_{arom}}{C_{aliph}} = 5.7$
19.7	C/H = 2.14	5.1	C/H = 1.64
9.5	C/H = 2.90	2.7	C/H = 1.95

high mesophase contents (which produce the HM fibers), however carbon yield increases with viscosity (Fig. 63) [154]. These contradicting conditions cause technological problems.

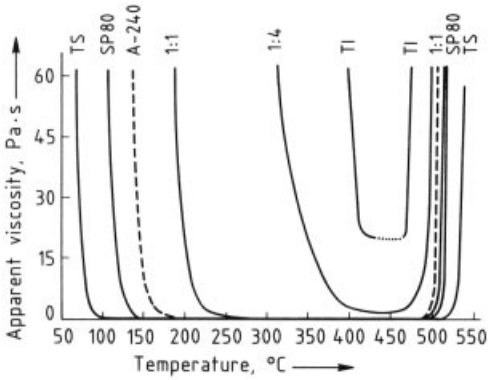

Figure 62. Apparent viscosity (rotation viscosimeter) of the petroleum pitch A240, the purified coal tar pitch SP80, fractions resulting from toluene extraction of SP80 (TS = toluene soluble, TI = toluene insoluble), as well as two mixtures of TS and TI [154]

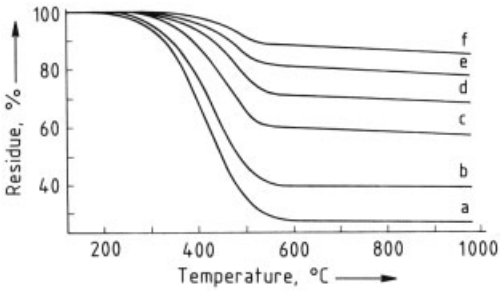

Figure 63. Carbonization residues from the pitch samples described in Figure 62 [154]
a) Toluene soluble; b) SP80 and 4 : 1 mixture; c) 1 : 1 mixture; d) 1 : 4 mixture; e) 1 : 9 mixture; f) toluene insoluble

Stabilization is carried out between 200 and 300 °C [155], much like in the PAN process. The stabilization times are dramatically shorter, however, between 10 and 30 min. The patent literature indicates that commercial pitch-based carbon fibers are given a final heat treatment around 1800 °C (P55 type). Fiber types with a relatively high Young's modulus, in the area of 400 GPa, can be achieved (cf. Fig. 44). A flow sheet for the entire fabrication process is given in Figure 64 [156].

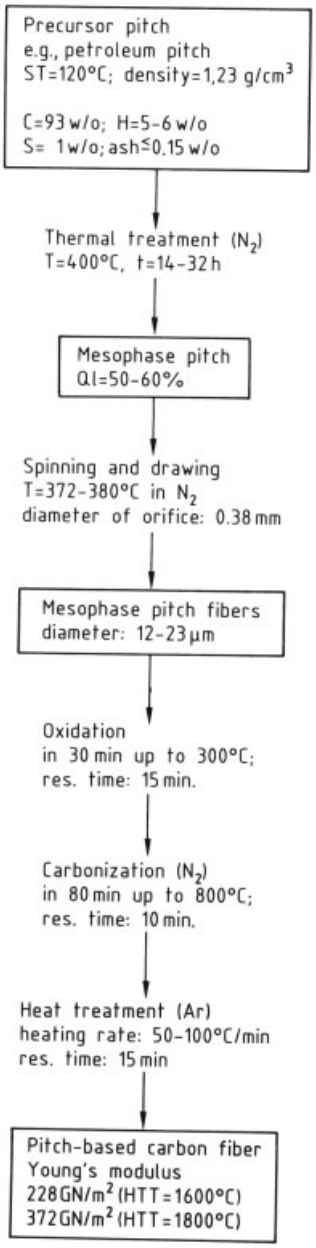

Figure 64. Flow sheet for manufacture of carbon fibers from pitch (Union Carbide Corp. (UCC) process) [156]

5.5. Comparison of the Carbon FibersBased on PAN and on Pitch

Figure 65 shows the stress – strain behavior of PAN-based carbon fibers, as commercially available, of the intermediate modulus fibers in development, and of isotropic and anisotropic pitch-based fibers.

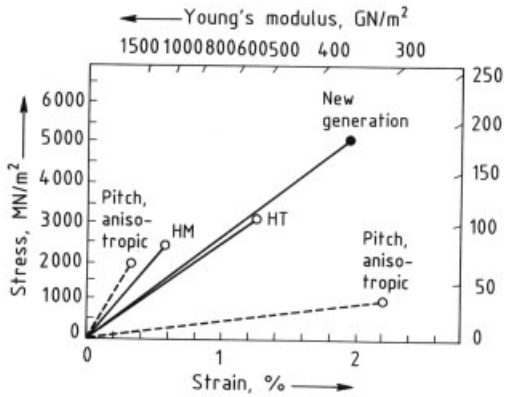

Figure 65. Stress – strain behavior of commercially available PAN- and pitch-based carbon fibers

The pitch process for carbon fibers potentially offers the capacity to produce carbon fibers covering the whole mechanical property range of the carbon fiber family. However, these expectations for pitch-based fibers have not yet been realized in spite of worldwide, intensive research.

The overwhelming part of commercial carbon fibers is the PAN-based HT type. High strain-to-failure of the fibers is seen as one of the key properties for the fracture toughness needed for broader application of advanced composites. The best strain values, between 1.5 and 2 %, are possible on industrial scale only with PAN-based HT-type fibers. Further, manufacturing problems are reduced if fabrication starts from a fully synthetic precursor material, with strictly controlled quality, instead of from a semi-natural raw material, such as petroleum or coal tar pitches, with strongly varying compositions depending on crude or coal type and the thermal history.

The applications of ultrahigh modulus carbon fibers from anisotropic pitch will most likely remain restricted to space structures and to carbon fiber-reinforced carbon fiber material (carbon – carbon composites). The challenge to increase the applications of anisotropic pitchbased carbon fibers can be satisfied only if production of medium-quality fiber types can be coupled with further drastic price reductions.

5.6. Commercial Availability and Economic Aspects

5.6.1. Companies and Trade Names

In the 1960s, the industrial production of carbon fibers started in UK, Germany, Japan, and USA with a total volume below 1000 t/a. Very soon, however, Japan had the advantage because of the availability of PAN-fiber plants using the wet process, whereas in USA and Germany PAN was produced mostly by the older dry process. Two Japanese companies specializing in the production of synthetic fibers, namely Toray and Toho Beslon, have succeeded in greatly expanding carbon fiber production. Today, these two companies dominate the world's carbon fiber production by licenses and joint ventures.

Hercules (USA) is independent from the two Japanese companies, but works jointly with Sumitomo, another Japanese PAN-precursor producer. Courtaulds fabricates precursors and is thus independent from Japanese imports. It is also the only company that sells PAN-precursor fibers suitable for a carbon fiber manufacture the world over without the limitations of licenses or joint ventures. In Germany, Sigri is independent, but fabricates no precursors. The producers, their production, their trade names, and some commercially available carbon fiber types from PAN precursors are compiled in Table 20.

5.6.2. World Capacity

The world capacity to produce carbon fibers is estimated to exceed 5000 t/a, and is increasing rapidly. The development of world production capacity during the last 15 years is shown in Figure 66. The small quantities of anisotropic and

Table 20. Carbon fiber producing companies, trade names, and commercial availability in the world market.

Producer (country)	Production capacity, t/a state 1988	Trade name	Types			Licensed by
			HT	IM	HM	
Toho (Japan)	1380	Besfight	HTA	IM 400	HM 35	
			ST 3	IM 500	HM 40 HM 45	
Toray (Japan)	1500	Torayca	T 300 T 700	T 800 T 1000	M 30 M 40 M 50	Asahi Nippon (Japan) 300
Amoco (USA)	360	Torayca	T 300 T 700	T 800 T 1000	M 30 M 40 M 50	Toray
Hercules (USA)	1360	Magnamite	As 4 As 6	IM-6 IM-7 IM-8	HMS HMU	Toray
BASF (USA)	450	Celion	G 30 – 400 G 30 – 500 G 30 – 600	G 40 – 600 G 40 – 700	G 50 – 300 Gy 70 Gy 80	Toho
Hysol (UK + USA)	700	Grafil	XA-S	IM-S	HM-S	
Sigri (FRG)	100 (PANOX 400)	Sigrafil	C	–	+	
AKZO (FRG + USA)	860	Tenax	HTH ST 3			Toho
Soficar (F)	300	Filkar	T 300 T 400	T 800	M 30 M 40 M 50	Toray
Amoco (USA)	230 ∗	Thornel		P 55	P 75 P 100 P 125	
Mitsubishi Chem. Ind.	500 ∗	Dialead				
Ashland (USA)	140 ∗	Carboflex		isotropic types		
Kureha (Japan)	900 ∗			isotropic types		

∗ pitch-based carbon fiber types

isotropic pitch-based carbon fibers are included. There is disequilibrium between capacity, production, and real consumption, but more exact information is not available. It is estimated that capacity will grow up to 10 000 t/a and increase between 30 and 50 % per year.

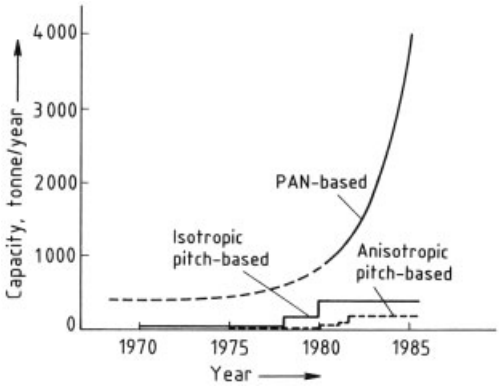

Figure 66. World production capacity for carbon fibers

Average prices of carbon fibers are shown in Figure 67. The price for PAN-based carbon fibers strongly depends on the number of filaments in one tow. The so-called low-filament tows, with 3000 monofilaments, have been used for years. Currently (mid-1980s) most prepregs use tows with 6000 monofilaments. The tendency is to try to increase the filament number up to 12 000.

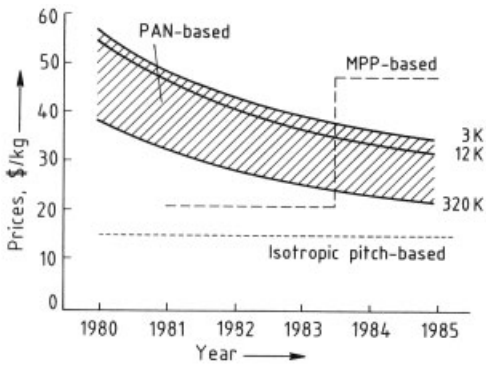

Figure 67. The development of carbon fiber prices

PAN-multifilament tows are the cheapest form of PAN-based carbon fibers: continuous 320 000 monofilament tows as used in textile application are difficult to handle in high-performance composites. They are mainly used for cutting into chopped fibers. However, multifilament tows with only 40 000 monofilaments have been introduced. The technological gap between the low filament tows (12 K and increasing) and the multifilament tows (40 K and decreasing) is diminishing. This should result in lower prices.

The price situation for anisotropic pitch-based carbon fibers is complicated. Around 1981 Union Carbide started to produce high-quality carbon fibers and sell them at $ 20 per kg, much below the price of PAN-based fibers at that time. As shown in Figure 67, the market situation and production difficulties forced the producer in 1983 to more than double the price. Anisotropic pitch-based high-modulus fibers are now much more expensive than PAN-based HT fibers.

The price situation is stable for isotropic pitch-based carbon fibers as delivered by Kureha for decades. Ashland, the Major petroleum pitch producer, has entered the market.

5.7. Future Importance

As has become evident over the last decade, polymers reinforced with carbon fibers present the opportunity not only to replace but also to displace structural metals [157]. The advanced composites (carbon fiber-reinforced plastics, CFRPs) are superior to conventional metals in respect to strength, stiffness, lightness, thermal expansion, fatigue, and corrosion resistance (Fig. 68).

Comparable Costs. The comparable costs of strength and stiffness of a unit volume of an advanced composite reinforced with PAN-based HT or HM carbon fibers, polyaramide fibers, glass fibers, former boron, or silicon carbide fibers have been estimated and compiled in Table 21.

From the viewpoint of strength, the glass fiber reinforcement is cheapest, followed by polyaramide and PAN-based HT fibers. PAN-based HM fibers are much more expensive. The CVD boron and silicon carbide fibers cannot be considered on an economical basis. From the viewpoint of stiffness, the carbon fiber-reinforced composites, although attractive on a price basis, are limited to the low modulus range.

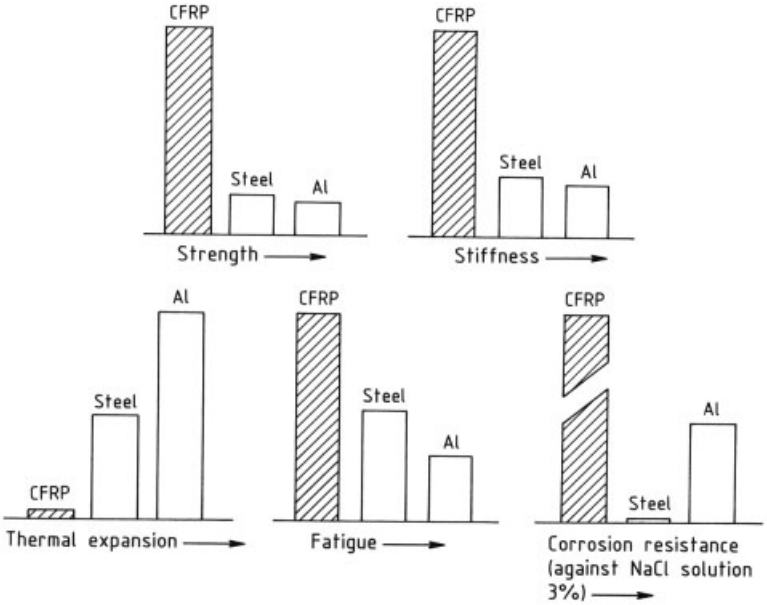

Figure 68. The technical advantages of CFRPs over conventional metals

Table 21. The price related to specific tensile strength and specific stiffness.

Fiber type	Price, DM/kg	Density, 10^3 kg/m^3	Price divided by specific	
			Tensile strength, DM/ 10^4 kgm	Young's modulus, DM/ 10^4 kgm
E glass	10	2.50	1.5	0.07
Kevlar	40	1.45	2	0.05
C_{PAN} (HT)	70	1.75	3	0.05
C_{MPP}	200	1.99	17.5	0.19
C_{PHN} (HM)	450	1.85	27	0.30
B/c	1000	2.50	67	0.50
SiC/c	4000	3.50	400	3

Technological Fields. The aerospace field can expect the highest technical and economic benefit from application of these new materials. In the past such risky innovations were the only chance to realize some ambitious projects. The U.S. space shuttle is the best example. Although NASA designed this shuttle in the early 1970s essentially in conventional materials, it was not realizable without carbon fiber-reinforced doors for the loading room. In this way, carbon fiber-reinforced epoxies were tested in practice and accepted as performing structural materials for primary structures in manned flying vehicles. Today, military aircraft contain up to 50 wt % carbon-reinforced resins instead of light metals like aluminum. Commercial aircrafts are following step by step to replace aluminum by carbon fiber-reinforced composites.

The cost-benefit schedule of typical composite applications is:

space market	5 000 €/kg
air craft	500 €/kg
racing car	50 €/kg
automotives	up to 5 €/kg

The present calculation gives load savings of 15 – 30 % in aircraft and component costs of 77 – 150 €/kg. In automotives, load savings of 20 – 50 % but component costs up to 2.6 €/kg. Although the cost-benefit ratio is lowest in automotive applications, this field is the most promising for the future. To produce large cars with extremely light weight using the design rules of the aerospace industry presents a real challenge to the auto industry. The question becomes more and more a question of economy of scale for the fiber production on the one hand, and of rationalized processing for final products on the other. Plans for future production lines have already been published [158].

Another argument supporting a shift from metals to carbon fiber-reinforced composites is the energy requirements shown in Table 22

[159]. The numbers in the table include both the energy needed to process the materials and to process the raw material. The composites appear advantageous in terms of volumes and even more so in terms of mechanical properties per unit volume.

Table 22. Energy requirement in kWh for the production of Al, Fe, CFRP's and PVC [159].

Material	Density, $\frac{g}{cm^3}$	Energy requirement for production		
		$\frac{kWh}{kg}$	$\frac{kWh}{L}$	$\frac{kWh}{L} / \frac{N}{mm^2}$
Al	2.8	37.3	104.5	232.2
Fe	7.9	6.8	53.6	26.8
CFRP (UD)	1.5	3.7	5.6	3.5
CFRP (2D)	1.5	3.7	5.6	6.9
PVC	1.4	7.9	11.1	1110.0

6. Refractory Ceramic Fibers

High-purity aluminosilicate wool fibers, essentially free from group 1 and group 2 oxides, are melt-spun from calcined kaolin or pure alumina and silica mixtures in a fashion similar to that for glass wool, rock wool, and mineral wool fibers (see Chap. 2). Five compositions are in common use: 50/50 wt % alumina/silica (maximum use temperature of 1250 °C), 60/40 wt % alumina/silica (maximum use temperature of 1400 °C), and chromia-, boria-, or zirconia-modified 50/50 wt % alumina/silica (also for use at 1400 °C). The physical properties of the fibers are mainly determined by their diameters. most products have a weight mean diameter in the range of 2 – 3 µm.

Further details concerning production technology and composition can be found in [160]. The principle use of such fibers is for hot-face insulation of pottery kilns and metal-working furnaces. Production in Europe is of the order of 16 000 t/a. Correspondingly large amounts are manufactured in the USA and in Japan. Worldwide, the major manufacturers are the Morganite Group (USA, Japan, UK, FRG), the Carborundum Group (USA and FRG), and the Société Européenne des Produits Réfractaires (SEPR) in France [161].

The fibers are manufactured as a low density blanket (64 –96 kg/m^3), which is flexible and can be used as such. Alternatively, rigid boards and shapes can be created by vacuum-forming wet slurries of raw fibers and a refractory binder (usually a silica sol). Products can be designed to be dimensionally-stable and to have excellent low thermal conductivity up to 1500 °C when combined with high-alumina products such as ICI "Saffil" alumina (see Section 3.4.1) Rigid boards lend themselves to modular furnace designs with rapid heating and cooling cycles or where high gas velocities would lead to erosion of blankets. In all cases, the use of low-density fiber aggregates leads to significant energy savings in batch firing regimes compared to furnace designs employing dense refractories.

7. References

1. ASTM D3878.
2. Tonen, 1988, Data sheet.
3. A. A. Griffith, *Philos. Trans. R. Soc. London* **221 A** (1920) 163.
4. F. Pott *Staub.-Reinhalt. Luft* **38** (1978) 181.
5. ASTM D3379.
6. H. Böder, D. Gölden, G. Popp, J. Vogelsang, *SAMPE* (1983), paper 12.
7. ASTM D4018-81.
8. S. W. Tsai, *NASA Tech. Rep.* CR-224 (1965).
9. E. Fitzer, W. Hüttner, Bericht über das 93. Wehrtechn. Symp., Verbundwerkstoffe in der Wehrtechn. part I, Bundesakademie für Wehrverwaltung und Wehrtechnik, p 93.04.01. 22 – 24 May 1978, Mannheim.
10. E. Fitzer, M. Heym, *Chemistry and Industry,* 21 August 1976, 663 – 676.
11. A. K. Dhingra, *Philos. Trans. R. Soc. London* **A 294** (1980) 411 – 417.
12. R. W. Lang, H. Stutz, M. Heym, D. Nissen, Lecture held on the symposium organized by the Fachgruppe "Makromolekulare Chemie" der Gesellschaft Deutscher Chemiker on "Neue Polymere – Spezielle Eigenschaften – Moderne Technologien," Bad Nauheim 14. – 15. April 1986.
13. R. Weiss, Ph. D. Dissertation University of Karlsruhe 1984.
14. J. B. Donnet, P. Ehrburger, *Carbon* **15** (1977) 143.
15. E. Fitzer: *Fibre Reinforcement for Composite Materials,* A. R. Bunsell, Amsterdam.
16. E. Fitzer, J. Schlichting, 18. Int. Chemiefasertagung Dornbirn (1979), in: Anorganische Fasern – Herstellung, Eigenschaften, Verwendung – 2. Werkstofftechnik, **11** (1980) 330 – 341.

17. G. Jacobsen, "Influence of the Interface on Transverse Strength, Creep Behaviour and Impact Behaviour of Boron- and Silicon Carbide-Fibre Reinforced Aluminum," ICCM-IV, Tokyo, 25. – 28. 10. 1982, *Prog. Sci. Eng. Compos.,* 1323 – 1330.
18. E. Fitzer, *Carbon* **25** (1987) 163 – 190.
19. E. Fitzer, *Pure Appl. Chem.* **60** (1988) no. 3, 287 – 302.
20. E. Fitzer, P. Schubert, *Rev. Int. Hautes Temp. Refract.* **16** (1979) 147 – 155.

General References
21. Tagungshandbücher SPI, Conference reports, (Society of the Plastic Industry) (SPI), New York.
22. *Tagungshandbücher AVK (conference reports)*, Arbeitsgemeinschaft Verstärkte Kunststoffe, Frankfurt.
23. K. A. F. Schmidt: *Textilglas für die Kunststoffverstärkung,* Zechner und Hüthig, Speyer 1972.
24. K. L. Loewenstein: *The Manufacturing Technology of Continuous Glass Fibers,* Elsevier, Amsterdam-Oxford-New York 1983.
25. H. S. Katz, J. V. Milewski: *Handbook of Filers and Reinforcements for Plastics,* van Nostrand Reinhold, New York 1978.
26. W. Bobeth, W. Böhme, J. Techel: *Anorganische Textilfaserstoffe,* VEB Verlag Technik, Berlin 1955.
27. J. G. Mohr, W. P. Rowe: *Fiber Glass,* Van Nostrand Reinhold Co., New York 1978.

Specific References
28. R. Kleinholz, J. Molinier: "Verstärkungswerkstoffe nach Maß," AVK-Tagung, 1985.
29. R. Kleinholz: "Neue Erkenntnisse bei Textilglasfasern zum Verstärken von Kunststoffen," Internationale Chemiefasertagung 22. Dornbirn, Austria, 1983.
30. J. Eule: "Glasstapelfaser-Vorgarne für die Herstellung textiler Erzeugnisse," *Textilbetrieb Würzburg,* **100** (1982).
31. J. Bijen: "Durability of Some Glass Fiber Reinforced Cement Composites," *J. Am. Concr. Inst.* **80** (1983) no. 7, 8.
32. Vetrotex report 2/86, Gevetex, Aachen 1986.
33. Gevetex Textilglas GmbH, Aachen 1986.
34. C. Z. Carroll-Porczinski: *Inorganic Fibers,* National Trade Press, London 1958.
35. P. A. Koch, *Glastech. Ber.* **23** (1950) 204 – 212.
36. G. von Pazciczky, DE 272 441, 1912. F. Rosengarth, DE 539 738, 1930.
37. Owens Illinois Glass, US 2 133 236, 1933 (G. Slayter, J. H. Thomas).
38. E. Poeschel, R. König, R. Haag, U. Roth et al.: *Umweltrelevanz künstlicher Fasern als Substitut für Asbest,* Erich Schmidt Verlag, Berlin 1978.
39. W. Trier: *Glasschmelzöfen,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1984.
40. W. Giegerich, W. Trier: *Glasmaschinen,* Springer Verlag, Berlin-Göttingen-Heidelberg 1964.
41. J. A. Škol'nikov, B. M. Polik, E. P. Kočarov, E. R. Nigin: *Steklannoe štapel'noe volokno,* Isd. Khimija, Moscova 1969.
42. F. V. Tooley: *The Handbook of Glass Manufacture,* 3rd ed., vol. 2, Ashlee Publishing, New York 1984.
43. ISO/DIS 8142, ISO/DIS 8144, ISO/DIS 8145.
44. J. Thebault, R. Pailler, G. Bontemps-Moley, M. Bourdeau et al., *J. Less Common Met.* **47** (1976) 221 – 223.
45. C.-H. Andersson, R. Warren, *Composites* **15** (1984) 16 – 24.
46. C.-H. Andersson, R. Warren, *Composites* **15** (1984) 101 – 111.
47. J. G. Lee, I. B. Cutler, *Am. Ceram. Soc. Bull.* **54** (1975) 195 – 198.
48. P. Bracke, H. Schurmans, J. Verhoest: *Inorganic Fibres and Composite Materials,* 1st ed., Pergamon Press, Oxford 1984.
49. B. von Falkai (ed.): *Synthesefasern,* Section 6.2.4, Verlag Chemie, Weinheim 1981, pp. 268 – 274.
50. W. C. Miller, *Kirk-Othmer,* **20** (1982) 65 – 77.
51. G. Winter, *Angew. Chem.* (1972); *Angew. Chem. Int. Ed. Engl.* **11** (1972) 751 – 760.
52. J.-O. Carlsson, *J. Mater. Sci.* **14** (1979) 255 – 264.
53. V. Krukonis in H. S. Katz, J. V. Milewski (eds.): *Handbook of Fillers and Reinforcements for Plastics,* Chap. 28, Van Nostrand, New York 1978, pp. 545 – 561.
54. A. M. Tsirlin in W. Watt, B. V. Perov (eds.): *Handbook of Composites,* "Strong Fibres," vol 1, Chap. 5, Elsevier, Amsterdam 1985, 156 – 194.
55. K. C. Hou, US 3 677 799, 1972.
56. J. Schlichting, *Ber. Dtsch. Keram. Ges.* **57** (1980) 289 – 298.
57. T. F. Foltz, *Ceram. Eng. Sci. Proc.* **6** (1985) 1206 – 1220.

58. S. Yajima in W. Watt, B. V. Perov (eds.): *Handbook of Composites,* "Strong Fibres," vol 1, Chap. 6, Elsevier, Amsterdam 1985, pp. 201 – 237.
59. Tokai Carbon Co., GB 2 129 025, 1983 (Akiyama, Yamada, Takashata).
60. J. V. Milewski, F. D. Gac, J. J. Petrovic, S. R. Skaggs, *J. Mater. Sci.* **20** (1985) 1160 – 1166.
61. J. J. Petrovic, J. V. Milewski, D. L. Rohr, F. D. Gac, *J. Mater. Sci.* **20** (1985) 1167 – 1177.
62. Du Pont, US 3 808 015, 1968; GB 1 264 973, 1969 (L. E. Seufert).
63. A. K. Dinghra, *Philos. Trans. R. Soc. London. Ser. A* **294** (1980) 411 – 417.
64. Du Pont, GB 1 417 791, 1973 (J. R. Green, R. F. Tietz).
65. 3 M, US 4 047 965, 1976 (K. A. Karst, H. G. Sowman).
66. Y. Abe, S. Horikiwi, K. Fujimura, E. Ichiki in T. Hayashi, K. Kawata, S. Umekawa (eds.): *Prog. Sci. Eng. Compos.,* ICCM-IV, Tokyo 1982, pp. 1427 – 1434.
67. Sumitomo, GB 1 457 801, 1973; US 4 495 121, 1982.
68. J. D. Birchall, J. A. A. Bradbury, J. Dinwoodie in W. Watt, B. V. Perov (eds.): *Handbook of Composites,* "Strong Fibres," vol 1, Chap. 4, Elsevier, Amsterdam 1985, pp. 115 – 154.
69. E. Leroy, C. Robin-Brosse, J. P. Torre in L. L. Hench, D. R. Ulrich, (eds.): *Ultrastructure Processing of Ceramics, Glasses and Composites,* Wiley-Interscience, New York 1984, pp. 219 – 231.
70. 3 M, US 3 793 041, 1971, (H. G. Sowman).
71. H. G. Sowman, D. D. Johnson, *Ceram. Eng. Sci. Proc.* **6** (1985) 1221 – 1230.
72. H. Teranishi, H. Ichikawa, T. Ishikawa, *New Mater. New Processes* **2** (1983) 379 – 385.
73. S. V. Nair, J. K. Tien, R. C. Bates, *Int. Met. Rev.* **30** (1985) 275 – 290.
74. A. K. Dinghra, *CHEMTECH* 1981, 600 – 608.
75. F. Pott, *Staub Reinhalt. Luft* **38** (1978) 486 – 490.
76. United Aircraft Corporation, BE 1 177 854 (1967); US 3 549 424, 1971, 3 574 649, 1971.
77. US 3 760 049, 1973.
78. S. J. Everett, US 2 050 298, 1936.
79. J. A. Roberts in H. S. Katz, J. V. Milewski (eds.): *Handbook of Fillers and Reinforcements for Plastics,* "Metal Filaments," Chapter 31, Van Nostrand Reinhold, New York, 1978, pp. 597 – 616.
80. S. Kavesh: *Spinning Wire from Molten Metal,* vol. 74, AIChE no. 180, 1978.
81. H. Jones: "Experimental Methods in Rapid Quenching from the Melt," *Treatise Mater. Sci. Technol.* **20** (1981).
82. E. A. Feest: "Rapid Solidification for the Wire Industry," *Wire Ind.* June 1985.
83. P. Vansteenkiste: "Prevention of Static Charges in Dust Filter Fabrics with Stainless Steel Fibres," *Filtr. Sep.* **18** (1981) March-April.
84. R. De Bruyne: "Multipass Test Performance of Metal Fibre Filter Media," *Proceedings World Filtration Congress III,* vol. 2, Downingtown, 1982, pp. 400 – 408.
85. L. C. Shirvill, I. Lefever: "Metallic Fibre Surface Combustion Radiant Gas Burners," *Proceedings 1986 International Gas Research Conference,* Toronto, Canada, 1986.
86. J. Samyn: "Bekiscan, a Fast Nondestructive Method to Evaluate Emi-Shielding Properties of Metal Fibre-Filled Poylmers," *Antec '86 Proceedings,* Boston 1986.Chapter 5
87. J. B. Donnet, E. Fitzer, K.-H. Köchling, *Proceedings of 4th International Carbon Conference (Carbon '86),* Baden-Baden 30. 6. – 4. 7. 1986, pp. 701 – 705.
88. J. Simitzis, Ph. D. Thesis, University of Karlsruhe, FRG, 1975.
89. H. Jäger, Ph. D. Thesis, University of Karlsruhe, FRG, 1986.
90. E. Fitzer, F. Rozploch, *High Temp. – High Pressures,* 1988 in print.
91. E. Fitzer, E. Gantner, F. Rozploch, D. Steiner, *High Temp. – High Pressures,* 1988, in print.
92. E. Fitzer, W. Frohs, M. Heine, F. Rozploch, *Proceedings of 4th International Carbon Conference (Carbon '86),* Baden-Baden 30. 6. – 4. 7. 1986, pp. 850 – 852.
93. W. Ruland: *J. Polym. Sci. Part C* **28** (1969) 5.
94. E. Fitzer, W. Frohs, F. Rozploch, *High Temp. – High Pressures* **19** (1987) 311 – 322.
95. E. Fitzer, P. Laudenklos, G. Schwarz, *Ber. Dtsch. Keram. Ges.* **53** (1976) no. 6, 169 – 174.
96. A. A. Bright, L. S. Singer, *Carbon* **17** (1979) 59 – 69.
97. W. Ruland, *J. Appl. Phys.* **38** (1967) 3585.
98. S. C. Bennet, D. J. Johnson: *Proc. London Int. Carbon Graphite Conf. 5th* **1** (1978) 337.
99. M. Guigon, A. Oberlin, G. Desarmot, *Fibre Sci. Technol.* **20** (1984) 117 – 198.
100. W. Marciniak, F. Rozploch, *High Temp. – High Pressures* **11** (1979) 709 – 710.
101. P. Schubert, Ph. D. Thesis, Univ. of Karlsruhe, FRG, 1977.
102. K. Brennfleck, E. Fitzer, G. Schoch, 1984 in Extended Abstracts, Carbone 84, Bordeaux, France (Talence: Centre de Recherche Paul Pascal) pp. 162 – 163.

103. H.-P. Boehm, *Catalysis* **16** (1966) 176.
104. J. Cziollek, Ph. D. Thesis, University of Karlsruhe, FRG, 1983.
105. E. Fitzer, *Carbon Fibres and Their Composites,* Springer Verlag 1985.
106. E. Fitzer, R. Weiss, *Carbon* **25** (1987) no. 4, 455 – 467.
107. R. Bacon, in P. J. Walker, jr. (ed.): *Chemistry and Physics of Carbon,* **vol. 9,** Marcel Dekker, New York 1973, p. 7.
108. T. A. Edison, US 223 898/470 925, 1879/1892.
109. W. F. Abbott, US 3 053 775, 1962.
110. *J. Electrochem. Soc.* **106** (1959) 147 C.
111. G. E. Cranch: *in Proceedings of the Fifth Conference on Carbon,* **vol. 1,** Pergamon Press, New York 1962, p. 589.
112. C. E. Ford, C. V. Mitchel, US 3 107 152, 1963.
113. D. L. Schmidt, H. T. Hawkins, in Reinforced Plastics '65, *Society of Plastics Engineers Regional Conference,* Seattle, 1965, p. 111.
114. E. M. Peters, US 3 235 353, 1966.
115. R. Bacon, G. E. Cranch, R. O. Moyer, jr., W. H. Watts, US 3 305 315, 1967.
116. R. Bacon, W. A. Schalamon, *Eighth Biennial Conference on Carbon,* Buffalo, N. Y., June 1967, Paper MI-58.
117. W. J. Spry, GB 1 093 084, 1967; US 3 454 362, 1969.
118. D. W. Gibson, G. B. Langlois, *ACS Polymer Preprints* **9** (1968) 1376.
119. R. C. Houtz, *Tex. Res. J.* **20** (1950) 786.
120. A. Shindo, JP 28 287, 1959.
121. W. G. Vosburgh, *Tex. Res. J.* **30** (1960) 882 – 896.
122. W. Johnson, H. Phillips, W. Watt, GB 1 110 791, 1965.
123. W. Watt, in A. Kelly et al. (eds.): *Handbook of Composites,* "Strong fibres," vol. 1, Elsevier, Amsterdam 1985.
124. D. J. Müller, Ph. D. Thesis, University of Karlsruhe, *"Zur Herstellung von Kohlenstoffaser aus Polyacrylnitril über die katalytische Cyclisierung des Polymeren,"* 1973.
125. S. Otani, *Mol. Cryst. Liq. Cryst.* **63** (1981) 249.
126. S. Otani, A. Oya, *Kogyo Kagahu Zasshi* **73** (1970) 493.
127. S. Otani, A. Oya, in *Petroleum Derived Carbon,* Am. Chem. Soc. Symp. Ser. 303, Washington, D.C., 1986, p. 323.
128. W. P. Hettinger, J. W. Newmann, R. P. Krook, D. C. Boyer, *Earth Moving Industry Conference,* Peoria, Illinois, USA, April 8 – 10, 1986 in SAE (Society of Automotive Engineers) Technical Paper series 860 767.
129. I. Romey, "Preprint," *2. Int. Kohlenstofftagung,* Baden-Baden 1976, p. 470.
130. J. B. Barr, S. Chwastiak, R. Didchenko, R. T. Lewis, J. C. Lewis, L. Singer, *Appl. Polym. Symp.* **29** (1976) 261.
131. Union Carbide, US 4 005 183, 1977.
132. E. Fitzer, H. Schlesinger, *Ber. Dtsch. Keram. Ges.* **43** (1966) no. 3, 209 – 218.
133. E. Fitzer, W. Fritz, B. Rhee, *High Temp. – High Pressures* **4** (1972) 219 – 224.
134. T. Koyama, M. Endo, Y. Hishiyama, *Jpn. J. Appl. Phys.* **13** (1974) no. 12, 1933 – 1939.
135. A. Nogaj, in B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981, pp. 197 ff.
136. G. P. Daumit, Y. S. Ko, in K. Brunsch et al. (eds.): in *High Tech – the Way into the Nineties,* Elsevier, Amsterdam, 1986, pp. 201 – 213.
137. E. Fitzer, W. Frohs, M. Heine, *Carbon* **24** (1986) 387 – 395.
138. K. Morita et al., *Ext. Abstracts Int. Symp. on Carbon,* Toyohashi 1982, p. 321.
139. E. Fitzer, M. Heine, G. Jacobsen, *Int. Symp. on Carbon,* Toyohashi, Japan 1982.
140. E. Fitzer, D. Müller, A. K. Fiedler, *Int. Conference on Carbon Fibres, Their Composites and Application,* no. 2, London 1971.
141. E. Fitzer, A. K. Fiedler, D. J. Müller, *Circ. Inf. Tech. Cent. Doc. Sider.* **43** (1971) 923 ff, 1022 ff.
142. E. Fitzer, A. K. Fiedler, *Preprints Carbon 72,* Baden-Baden, Juni 1972, paper 21.
143. E. Fitzer, D. J. Müller, *Chem. Z.* **96** (1972) 20 – 26.
144. E. Fitzer, M. Heym, *Chem. Ind.* **16** (1976) 663–676.
145. E. Fitzer, T. Müller, *Proc. of Carbon '82 Ext. Abstracts,* London 1982, pp. 249 – 252.
146. N. Grassie, R. McGuchan, *Eur. Polym. J.* **7** (1971) 1357.
147. E. Fitzer, D. J. Müller, *Carbon* **13** (1975) 63.
148. E. Fitzer, M. Heine, W. Metzler, *Proc. of Carbon '86,* Baden-Baden, DKG 1986, pp. 853 – 855.
149. Toho, Beslon Co., Ltd. DE-OS 31 32 784, 1981.
150. Toray Industries, JP 48 228, 1975.
151. E. Fitzer, A. Gkogkidis, G. Jacobsen, *Extended Abstracts Int. Symp. on Carbon,* Toyohashi 1982, p. 284, Carbon Society of Japan, Tokyo.
152. D. W. McGuffin, R. J. Larsen, R. W. Mitchell, B. E. Caruso, US 3 803 672, 1974 assigned to HITCO.

153. T. Trendelenburg, Ph. D. Thesis, University of Karlsruhe, FRG, 1984.
154. D. Kompalik, Ph. D. Thesis, University of Karlsruhe, FRG, 1986.
155. E. Fitzer, D. Kompalik, *Sprechsaal* **119** (1986) no. 10, 911 – 916.
156. Union Carbide, US 4 005 183, 1977.
157. *Engineering,* Sept. 1982, p. 611.
158. H. Böder, SIGRI GmbH Meitingen/Augsburg, Vortrag anläßlich Carbon '86 in Baden-Baden.
159. H. Fleckenstein, *VDI-Z.* **125** (1983) nr. 4, 123 – 131.
160. W. C. Miller, *Kirk-Othmer* **20** (1982) 65 – 77.
161. J. Griffiths, "Synthetic Mineral Fibres," *Ind. Miner. (London),* Sept. (1986) 20 – 43.

Fibers, 6. Testing and Analysis

Ernst Kleinhansl, Institut für Textil- und Verfahrenstechnik, Denkendorf, Federal Republic of Germany

Mary C. Sawyer, E. I. Du Pont de Nemours and Company, Wilmington, DE 19898, United States

1. Introduction 2
1.1. History and Current State of the Art 2
1.2. Definitions and Basic Terms 2
1.3. Test Procedures, Rules, Standards 3
1.4. Climate 3
1.5. Sampling 3
2. Tests of Form and Mass 4
2.1. Length 4
2.1.1. Single-Fiber Length 4
2.1.2. Comb Sorter Method 4
2.1.3. Electronic Method – Capacitive Scanning 5
2.1.4. Electronic Method – Optical Scanning 6
2.2. Width, Thickness 7
2.3. Fineness 7
2.3.1. Gravimetric Method 7
2.3.2. Projection Microscope Method (Fiber Diameter) 7
2.3.3. Diameter Measurement with Laser Beam 8
2.3.4. Projection Microscope Method with Image Analysis Techniques 8
2.3.5. Vibration Method 8
2.3.6. Airflow Method 9
2.3.7. Microscopic Area Measurement . . . 9
2.4. Microscopic Investigation of Cross Section 9
2.5. Microscopic Investigation of Fiber Surface 10
2.6. Crimp 11
2.6.1. Crimp Frequency 11
2.6.2. Crimp Ratio, Crimp Retention 11
2.7. Density 12
2.7.1. Pycnometer Method 12
2.7.2. Density Balance (Archimedes' Principle) 12
2.7.3. Gradient Method 12
2.8. Chemical Identification of Fibers by Solubility 12
2.9. Other Identification Methods for Fibers 12
2.10. Identification of Wool and Hair . . 13
2.11. Analysis of Fiber Impurities and Extractable Matter 13
2.11.1. Trash and Dust Content of Cotton (Nonlint Content) 13
2.11.2. Honeydew on Cotton 13
2.11.3. Grease Content of Wool – Extractable Matter Content 14
2.11.4. Preparation and Finishing Agents for Man-Made Fiber 14
2.12. Maturity of Cotton 14
3. Tensile Properties 15
3.1. Tensile Test 15
3.2. Elastic Behavior 16
3.3. Modulus 16
3.4. Loop Tensile Test 16
3.5. Bundle Stress Test 16
4. Behavior toward Water 17
4.1. Moisture Adsorption 17
4.1.1. Oven Drying Test 17
4.1.2. Other Tests for Moisture Content of Fibers 17
4.2. Water Absorption 18
5. Thermal Behavior 18
5.1. Melting Point 18
5.2. Burning Test 18
6. Dyeing Properties 18
6.1. Laboratory Dyeing 18
6.2. Identification of Fiber Type by Dyeing (Neocarmine) 19
7. Processing 19
7.1. General 19
7.2. Spinning Devices for Trials 19
7.3. Electrostatic Charging 19
7.4. Rotor Ring Method 19
7.5. Combination of Micro Dust and Trash Analyzer with the Rotor Ring Method and a Spinning Device for Trials 20
8. Automated Test Units 20
9. References 20

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_067

1. Introduction

1.1. History and Current State of the Art

For as long as textile fibers have been used, people have tested them in one way or another usually aided by nothing more than their eyes, their fingers, and occasionally, their noses. Visual methods, in refined form, are still in use today, generally because objective methods of measurement are lacking or are very expensive. Visual test methods are subjective and require great experience on the part of the person making the test.

Over the past decades, various methods of measurement and appropriate test apparatus have been developed for measurement of individual fibers. Such measurements presume great care, intuition, and skill on the part of the individual; nevertheless, the tester usually influences the results. Many individual values must be obtained because of the relatively high dispersion of fiber characteristics and properties; hence, measurements are tedious and costly.

Current developments tend toward so-called collective testing procedures which, together with automation and the use of computers, provide information on fiber properties very rapidly and minimize the influence of the individual tester. Frequently, the required properties themselves are not measured, but others that correlate with them. By coupling several testing devices, so-called test beds (high-volume testing, high-volume instruments) are created. This technology is still evolving.

1.2. Definitions and Basic Terms

Fibers are linear products that can be processed on textile machinery. Their common characteristic is a relatively high ratio of length to cross section, sufficient strength, and flexibility. Depending on their origin or composition, they are divided into the following groups:

1) Natural fibers (vegetable, animal, and mineral)
2) Man-made fibers from natural or synthetic polymers
3) Industrially produced fibers based on inorganic substances

Depending on their form, fibers are designated as

1) Staple fibers (fibers of limited length)
2) Filaments (fibers of practically unlimited length)

Test methods can sometimes be used with all kinds of fibers and sometimes only with certain kinds. In addition, methods are available for measuring characteristics specific to fiber type.

1.3. Test Procedures, Rules, Standards

To assure comparable and reproducible results, various organizations issue test procedures, rules, and standards. The following are some of the most important ones.

International standards and rules:

ISO International Organization for Standardization
EN European Standards
BISFA International Bureau for the Standardization of Man-Made Fibers
IWTO International Wool Textile Organization

Some national standards:

ASTM United States
AATCC United States
DIN Germany
BS United Kingdom
NF France

Standards or rules do not exist for all test methods. In that case, measurements are carried out according to methods developed by companies, institutes, associations, or instrument manufacturers.

Any existing standards or rules promulgated by ISO, CEN, ASTM, BISFA, IWTO, and DIN are listed with each of the test methods described below.

1.4. Climate

Fiber properties, especially those of natural fibers and of man-made fibers from natural polymers, depend on the testing climate. The relative

humidity is of primary importance here. In addition, fibers exhibit hysteresis when water is absorbed. Fibers take up different amounts of water, depending on whether they have been conditioned by adsorption or desorption. Therefore, fibers are tested in a standard climate:

Temperature $(20 \pm 2)\,°C$
Relative humidity $(65 \pm 2)\,\%$

Samples must always be conditioned to the standard climate in the adsorptive phase. When necessary, they must be dried first if, prior to testing, they have been exposed to a climate with more than 65 % R.H. Conditioning of samples is prescribed by standards ISO 139; ASTM D 1776; and DIN 53 802.

Currently, the European standardization commitees are striving for a change in the standard climate, with the aim of harmonization with testing laboratories in other fields such as plastics, paper, leather, and biology. The standard climate would then be:

Temperature $(23 \pm 2)\,°C$
Relative humidity $(50 \pm 5)\,\%$

1.5. Sampling

Fiber samples must be tested so that from the values obtained for a given property, the value of that property for the source material can be deduced. The sample must, therefore, be representative of the source material. Details of sampling are governed by the following standards and rules: ISO 1130; BISFA Rules; ASTM D 1060, D 1441, D 1771, D 2525, D 3333; DIN 53 803 (ASTM D 1771 has been discontinued without replacement, although copies are available from ASTM headquarters). A new standard is in preparation, prEN 12751:1997, which replaces all European national standards for sampling and later later the international standard ISO 1130.

2. Tests of Form and Mass

2.1. Length

Staple fiber length is an important fiber characteristic which greatly affects the spinning process as well as the quality of the textile end product. The lengths of natural fibers vary widely; therefore, many individual measurements are needed. The large number of fibers in a test lot and the inhomogeneities (e.g., from bale to bale or from batch to batch) also necessitate many individual measurements. Man-made fibers, with few exceptions, are cut to a certain length; substantially less variation occurs, but the length can be reduced by stresses on the fiber while it is being processed into yarn. This increases the deviation.

2.1.1. Single-Fiber Length

With the aid of two pairs of tweezers, a fiber is picked up at both ends, the crimp is pulled out by applying a slight tension, and the length is read off on a ruler. Crimp can also be removed by using a pair of tweezers to draw the fiber over a glass plate covered with paraffin oil or vaseline. The length is then measured with a ruler.

The measuring process can be simplified by mechanical devices. One of them is shown in Figure 1. With a pair of tweezers, the fiber is drawn under a loading weight and the crimp is thus removed. The results are presented in the form of a cumulative frequency diagram (Fig. 2); average fiber length and coefficient of variation are calculated.

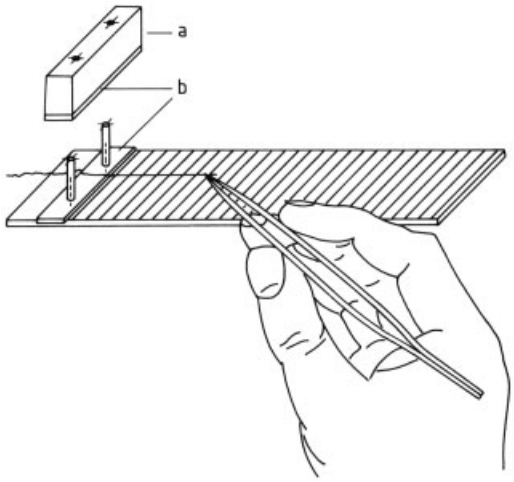

Figure 1. Length measurement of single fibers
a) Loading weight; b) Velvet strips

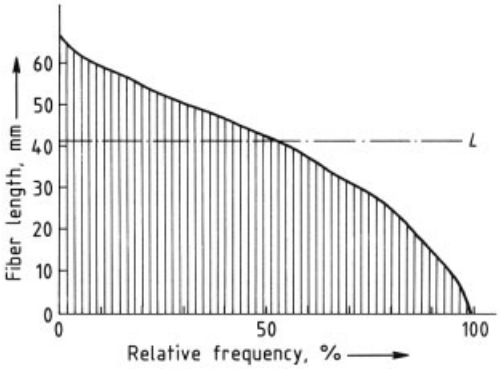

Figure 2. Cumulative distribution of fiber length
L = average fiber length

Specific standards for length measurement are ISO 6989; ASTM D 3661; IWTO-5-60; DIN 53 808 T 1. For future reference, ASTM is developing another method (57-507) which is assigned "D".

2.1.2. Comb Sorter Method

A prepared fiber sample containing at least 10 000 fibers is first end-aligned with the aid of two banks of combs and special forceps (Fig. 3). The distance between pins depends on the type of fiber being tested: e.g., 3 mm for cotton, 10 mm for longer wool fibers. This distance is also the class interval for length classes. Once ordered, fibers of individual length classes are drawn from the bank and weighed. The weight is used to calculate the number of fibers of each class and, thus, to determine the average fiber length and the coefficient of variation. The result is presented as shown in Figure 2.

Mechanized devices that perform the ordering operation have been developed to test wool. Manual preparation, particularly end alignment, is difficult, time-consuming, and dependent on the care and skill of the tester. The procedure is applicable to fibers from flock (raw material), slivers, and rovings.

Standards for comb sorting are ISO 920; IWTO-1-66; ASTM D 519, D 1440, D 1575, D 3660; DIN 53 806.

2.1.3. Electronic Method – Capacitive Scanning

Two steps have been taken to simplify measurement of fiber length. (1) Parallelizing has been automated. This is more readily accomplished with sliver than with flock. Long fibers such as wool or long staple man-made fibers can be ordered more easily and accurately. (2) The

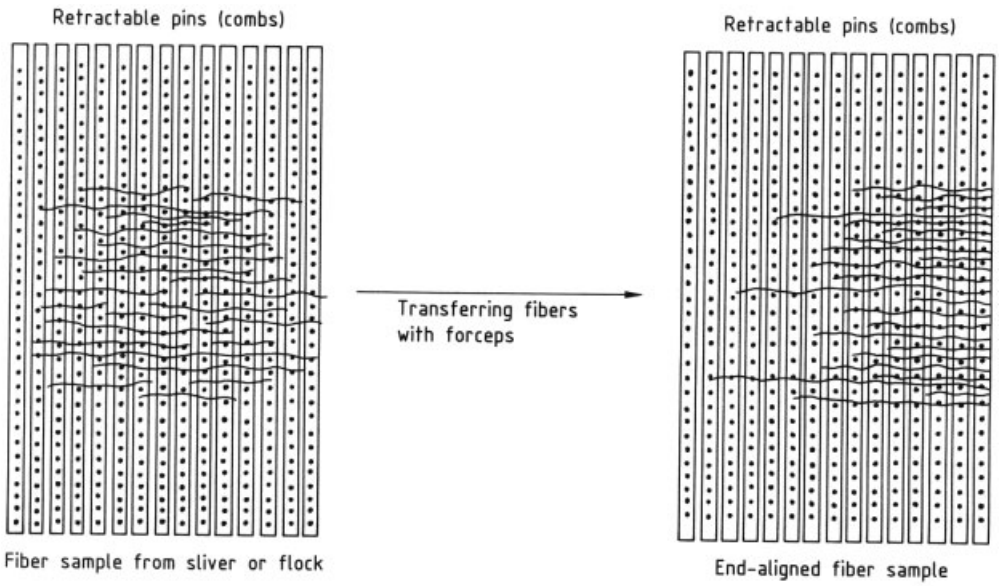

Figure 3. Comb sorting principle

frequency of individual length classes is determined by means other than drawing and weighing.

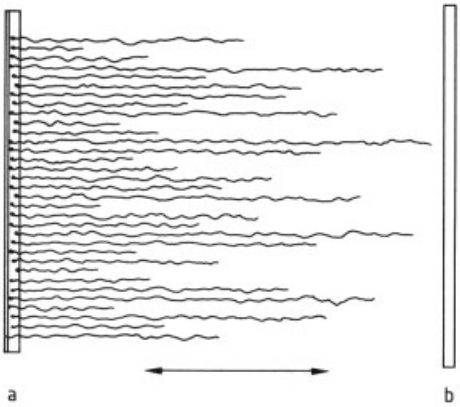

Figure 4. End-aligned beard with capacitive measuring principle
a) Clamp; b) Measuring condenser

WIRA Fiber Diagram Machine. The end-aligned beard is prepared by mechanical means. The ends are thermally bonded with plastic film. The fiber beard, which is now easy to handle, is drawn through the electrodes of a condenser, and the fiber mass for each length zone is measured. Standards are ISO 2646 and IWTO-16-67.

In the latest version of this machine, all characteristics of fiber length measurement, including the fiber length diagram, are printed out. The procedure is suitable for wool and long staple man-made fibers in sliver form.

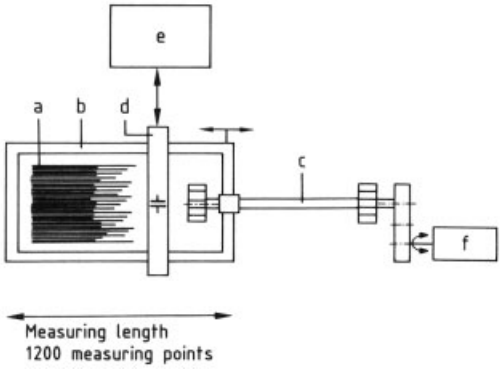

Figure 5. Sketch showing principle of Almeter
a) End-aligned fiber sample; b) Measuring sled; c) Spindle; d) Measuring condenser; e) Detector; f) Stepping motor

Almeter. Fibers are end-aligned by a fully automatic sample preparation machine; the sample is then measured in the electric field of a condenser (Figs. 4 and 5).

This method is suitable for all fibers in flock or sliver form; the standards are ISO 2648 and IWTO-17-85.

2.1.4. Electronic Method – Optical Scanning

Optical Scanning of a Beard. The Fibrograph was developed specifically for measuring the length of cotton fibers. A beard is formed in a special comb by using the Fibrosampler. The beard is not end-aligned. In the measuring zone, the optical density of the beard is measured as a function of fiber length (Fig. 6).

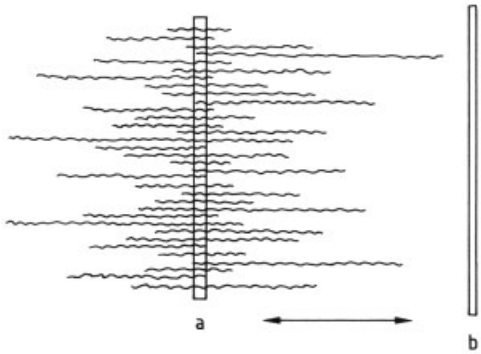

Figure 6. Beard, not end-aligned, with optical scanning
a) Clamp; b) Optical sensor

Results are given in terms of so-called span length, defined as a function of cumulative frequency. The frequency diagram is called a fibrogram (Fig. 7). It has become customary to report the span length at 2.5 and 50 % frequency, abbreviated as the 2.5 % SL value and the 50 % SL value. In addition, the uniformity index is given, which correlates with the deviation of fiber length.

The Fibrograph was developed by SpinLab (Knoxville, USA) and is now available from Zellweger (Uster, Switzerland) under the trade name Uster Fibrograph 730. Many measurements can be made in a short time with this machine, which is especially important in testing raw cotton. Standards are ISO 4913; ASTM D 1447.

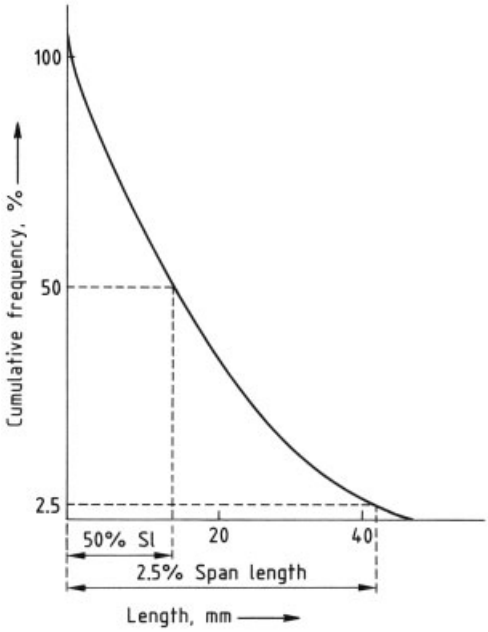

Figure 7. Fiber length measurement with Fibrograph Fibrogram shows 2.5 % and 50 % span lengths

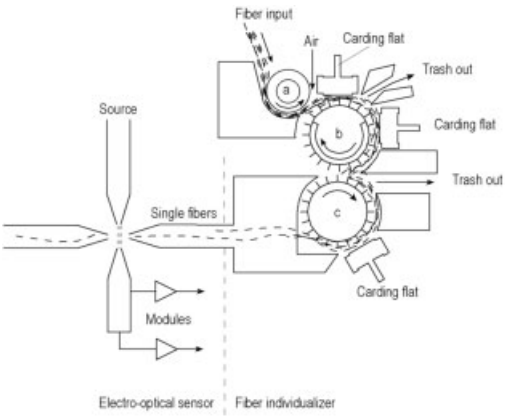

Figure 8. Schematic of the AFIS system

Optical Scanning of Single Fibers. In the AFIS tester (AFIS = advanced fiber information system; Figure 8) flocks or slivers of fibers are individualized by a special device. The individual fibers are aligned longitudinally and stretched in an air stream. They then pass through a tapered channel into an optical measuring system. The data are evaluated by various modules, the most important of which is the L & D module, which measures fiber length and diameter (or width for nonround fibers). The fiber diameter or width is a measure of the fineness. Other modules measure fiber neps (Module N) and impurities (Module T).

2.2. Width, Thickness

Further geometric characterization of a fiber should include width or thickness and, for circular cross section, the diameter. Because of the multiplicity of fiber cross sections (see Section 2.4), a precise definition of one or more width measurements is not possible. Such measurements are, therefore, made only in exceptional cases. Fiber diameter is used as a characteristic of fineness only for wool (see Sections 2.3.2, 2.3.3, 2.3.4).

2.3. Fineness

Fiber fineness is defined as mass referred to a length (length-related mass). In the internationally accepted tex system (ISO 1144), the mass in grams is referred to 1 km:

1 tex = 1 g/km

From this, the decitex unit (dtex), frequently used for fibers, is derived:

1 dtex = 1 dg/km = 1 g/10 km

In certain countries, the millitex unit (mtex) is preferred:

1 mtex = 1 mg/km

2.3.1. Gravimetric Method

In gravimetric determination of fineness, the length and weight of the fibers are measured. The following practical method is used: a carefully parallelized, combed fiber bundle is cut at both ends when as much crimp as possible has been pulled out by suitable tensioning. The length should be as great as possible. Groups of 50 fibers – generally ten groups – are taken from this bundle. The total length is calculated from the cut length and the number of fibers. The mass of the individual groups of fibers is determined by weighing, and the fineness is calculated from the mass and the length. Standards are ISO 1973, EN 21 973; BISFA Rules; ASTM D 1577; DIN 53 812 T 1.

2.3.2. Projection Microscope Method (Fiber Diameter)

Because the fineness of individual wool fibers varies widely, the fineness distribution of individual fibers, in addition to the average fineness, must be determined. As a rule, at least 300 fibers must be measured. Figure 9 shows the results of a wool fineness measurement.

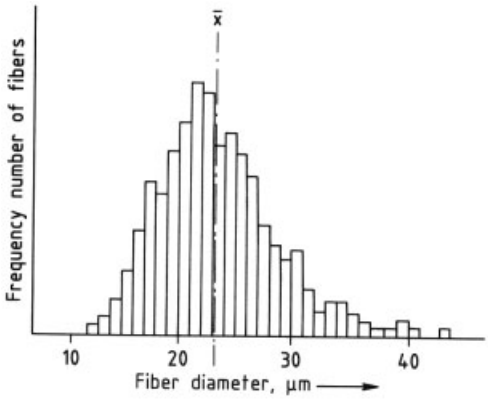

Figure 9. Example of wool fineness measurement
$\bar{x}$ = average fiber diameter

To calculate the fineness in decitex, the following formula is used:

$$T_d = \frac{\pi d^2 \cdot \varrho}{400}$$

where T_d = fineness, decitex; d = diameter, micrometers; ϱ = density, grams per cubic centimeter.

Appropriate standards are ISO 137; IWTO-8-89; ASTM D 2130; DIN 53 811.

2.3.3. Diameter Measurement with Laser Beam

Because measuring the fiber diameter of wool is very costly and time-consuming, simpler test methods are being sought. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) of Ryde, New South Wales, Australia, has developed the so-called fiber distribution analyzer (FDA) [5]. Wool fibers are first cut to a length of ca. 2 mm and then finely dispersed in a nonswelling liquid. The liquid with the pieces of fiber then flows through a measuring chamber where it is measured by means of a laser beam and an electronic optical measuring device. About 30 – 50 fibers per second can be measured. The measurements are analyzed by computer, and the results are printed out.

This instrument is still used in many laboratories, but has since been replaced by a better and more effective measurement system (see Section 2.3.4) and is therefore no longer manufactured.

2.3.4. Projection Microscope Method with Image Analysis Techniques

With the aid of an electronic image analyzer, microscopic pictures of wool fibers can be suitably analyzed; thickness measurements can also be carried out. A more detailed description of this measuring procedure is contained in ASTM standard D 3510 (discontinued without replacement, copies are available from ASTM headquarters).

A new method for rapid and automatic measurement of fiber diameter has been developed by Australian Wool Testing Authority (AWTA) Ltd., Kensington, Victoria, Australia. It is based upon image analysis techniques and is called FIDAM (Fiber Image Display and Measurement). The measurement section consists essentially of a high-quality CCD (Capacitive Coupled Diode) television camera attached to a low-power microscope. The camera output feeds signals to a flash analog-to-digital converter (ADC) and then to a desktop personal computer which is the heart of the system. Sample preparation is also largely automatic; it involves microcoring the sample and spreading the resulting snippets, which are air-mounted, onto a large microscope slide. The calibration, which is linear, is based on the Interwoollabs' Calibration Tops and is very stable [6].

A further development is the OFDA (optical fiber diameter analyzer) instrument, marketed by Zellweger, Uster, Switzerland, in which computer image processing is performed through a projection microscope. In 50 s, 2000 – 5000 fiber sections can be measured (standard: IWTO-22-92).

2.3.5. Vibration Method

The vibroscope is based on the vibrating string principle. Linear density, or mass per unit length, can be calculated from the fundamental resonant frequency of transverse vibration of a fiber measured under known conditions of length and tension. This device is often combined with a fiber tensile tester. The measured fiber fineness is combined directly with the breaking force, and the breaking tenacity is calculated from it. The applicable standards are ISO 1973; EN 21 973; ASTM D 1577; DIN 53 812 T 2.

2.3.6. Airflow Method

The resistance of a plug of fibers to airflow depends on fiber fineness. For cotton, a characteristic called "Micronaire value" has been defined which can be read off the Micronaire instrument, designed for this purpose, as a function of resistance to airflow ("Micronaire reading"). The instrument is calibrated with fixed cotton standards. Resistance to airflow also depends on maturity. For wool, the scale is calibrated in micrometers. Figure 10 shows a sketch of the principle of this device. This instrument is now available from Zellweger (Uster, Switzerland) under the trade name Uster Micronaire 775.

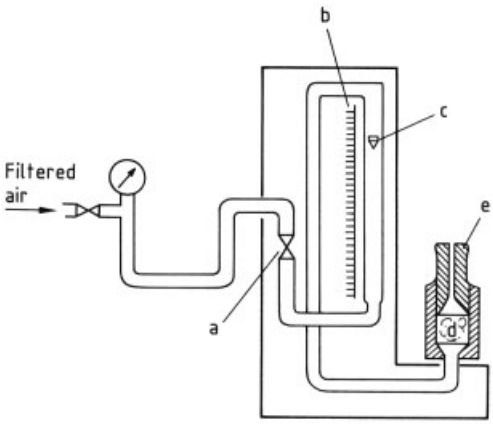

Figure 10. Micronaire tester Sheffield 60 600
a) Pressure regulator; b) Scale; c) Rotameter plummet; d) Sample holder with fiber plug; e) Fiber compression plunger

Further details can be found in the following standards: ISO 1136, ISO 2403; ASTM D 1282, D 1448; IWTO-6-90, IWTO-28-89; DIN 53 941.

2.3.7. Microscopic Area Measurement

The area can be measured microscopically when few fibers are available for measurement and when these fibers are not circular in cross section. In this case, the above-mentioned methods cannot be used. Fiber cross sections for microscopic examination are prepared with a microtome, and drawings or photographs of these cross sections are made. The cross-sectional area can be determined by planimetry or by electronic analysis of the pictures; fineness is calculated from the area.

2.4. Microscopic Investigation of Cross Section

An investigation of fiber cross section may be necessary for several reasons:

1) Identification of fibers
2) Estimation of blend ratios
3) Determination of fineness (see Section 2.3.7)
4) Determination of maturity of cotton (Section 2.12)
5) Determination of depth of penetration of dyes or optical brighteners, sheath – core zones of fibers
6) Measurement of cross-sectional parameters, e.g., modification ratio (the "deviation" from a circle) and angle of lobes

Modification ratio is the ratio of the diameter of an inscribed circle, touching all the valleys in a lobed cross section, to the diameter of a circumscribed circle touching all the peaks in that cross section.

In addition, Special Instrument Laboratory (SpinLab) (Knoxville, TN 37901) has developed the "Fibrofine" Specific Surface Tester Model 820, used for measuring fiber fineness of cotton, to correlate with modification ratio measurements from cross-sections.

The *identification* of specific fiber types is possible by viewing their cross sections (Fig. 11). Man-made fibers, which are normally extruded from spinneret holes with circular cross section, can also be produced with cross sections

of different shapes (Fig. 12). Test Method 20 (AATCC) shows photographs of common textile fibers.

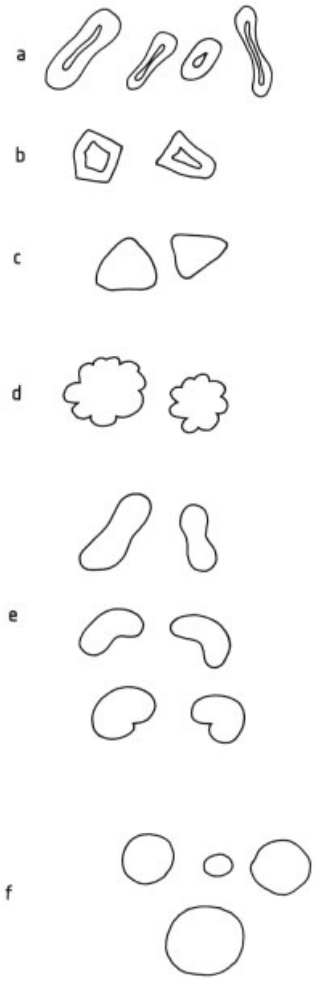

Figure 11. Cross sections of various fiber types
a) Cotton; b) Flax; c) Silk; d) Viscose rayon (normal); e) Most acrylic fibers (including modacrylic); f) Wool, cuprammonium rayon, polyester, polyamide, polypropylene, some acrylic fibers

Mixing Ratios. When the components of a fiber blend have different cross sections, an approximate blend ratio can be established through counting, by using photomicrographs of the cross sections. Mixtures of fiber of the same type but with different fiber fineness can also be examined in this way provided the difference in fineness is large enough (e.g., 3 – 6 dtex). Test Method 20 A (AATCC) "Fiber Analysis: Quantitative," has sections on microscopic analysis and procedures which cover this subject.

Dye Penetration. The distribution of dyes across the fiber cross section can be determined if the colorants are still recognizable in the relatively thin cross sections. The distribution of optical brighteners is determined easily under the microscope in UV light. Fibers with sheath – core structure can be made visible by using special dyes, e.g., Solophenyl Blue-Green BL (Ciba-Geigy). Some so-called bicomponent fibers can also be detected microscopically.

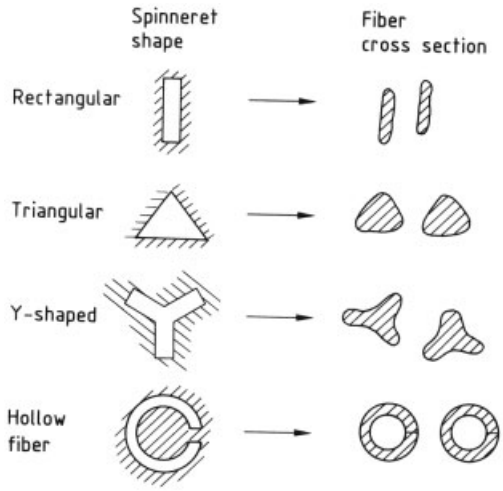

Figure 12. Some special cross sections in man-made fiber production

2.5. Microscopic Investigation of Fiber Surface

In certain cases, the fiber type can be identified by microscopic examination of a longitudinal view, i.e., of the fiber surface (Fig. 13). Thus, for example, wool can be recognized by its scale structure. When an unknown fiber must be identified, microscopic examination begins with the surface, because preparation of a specimen for longitudinal viewing is considerably faster and simpler than preparing one for cross-sectional viewing. Wool, cotton, flax, and viscose fibers can be identified unmistakably in this way. Test Method 20 (AATCC), also has information on longitudinal characteristics of natural fibers which is important to their identification.

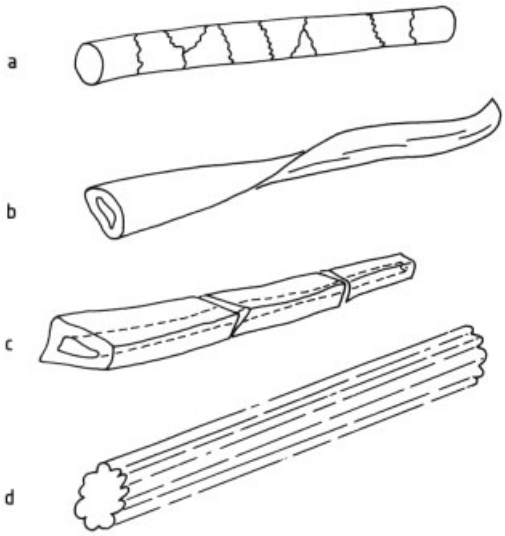

Figure 13. Longitudinal view of some fiber types
a) Wool (characteristic: scales); b) Cotton (characteristic: longitudinal torsion); c) Flax (characteristic: lumen, transverse cracks); d) Viscose rayon (characteristic: longitudinal stripes)

For man-made fibers, the longitudinal view also indicates if the fiber contains delusterants or if it is of the bright type.

Microscopic examination of the longitudinal view can also serve many other purposes, such as

1) Determination of the surface structure of fibers, pores, vacuoles, cracks
2) Detection of foreign matter on the fiber
3) Inspection of crack formation at the fiber ends, an indication of overload
4) Investigation of thermal damage due to excessive friction
5) Checking wool fiber ends for hair roots, and differentiating between shorn wool and lime wool
6) Detection of insect damage

These investigations are best carried out with scanning electron microscopes (SEM) which provide high magnification and good depth of field.

2.6. Crimp

Crimp is very important for the processibility of fibers and the properties of textiles made from them. The crimp of wool is relatively uniform and characterized by true waves. Man-made fibers, on the other hand, may exhibit irregular crimp, and the shapes of the waves can differ widely (Fig. 14).

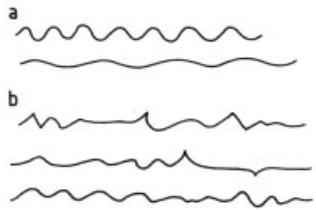

Figure 14. Different forms of crimped fibers
a) Wool; b) Synthetic fiber with stuffer box crimping

Several characteristics can be used to describe crimp; two of these are discussed in detail below.

2.6.1. Crimp Frequency

The number of crimp waves per centimeter is determined. The waves are counted either directly on the fiber or on enlargements, for example, with the aid of a projector. Care must be taken to count either the peaks or the valleys, not both (Fig. 15).

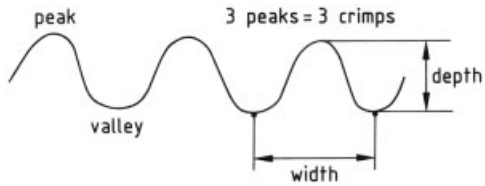

Figure 15. Definitions of crimping

In addition, the depth and width of the crimp waves can be measured with the aid of enlargements (ASTM D 2491; discontinued without replacement but available from headquarters).

2.6.2. Crimp Ratio, Crimp Retention

The crimp ratio E is defined by two parameters, the crimped length and the extended length of a fiber (Fig. 16).

Fibers should be straightened so that the crimp is removed but the material is not stretched. Crimp retention B is another characteristic that can be measured along with crimp.

The fiber which has been straightened to measure the crimp remains under this load for 5 min. After unloading, the fiber recovers partially; the extent depends on crimp retention. The length of this partially crimped fiber is then measured. The testing device used is the so-called crimp balance (Fa. Zweigle, Reutlingen, Federal Republic of Germany) which greatly facilitates the entire operation.

The change in length from the crimped to the straightened fiber can also be measured on a tensile testing machine. The Vibrotex machine (Lenzing AG, Lenzing, Austria) works by this principle.

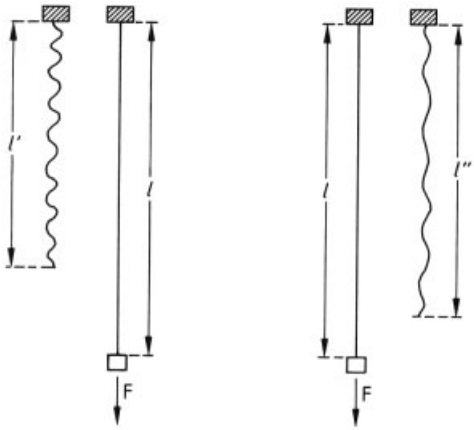

Figure 16. Definition of crimp ratio and crimp retention
Crimp ratio:
$E = \frac{l - l'}{l}$
or $E\ (\text{in}\,\%) = \frac{l - l'}{l} \cdot 100\%$
Crimp retention:
$B\ (\text{in}\,\%) = \frac{l - l''}{l - l'} \cdot 100\%$

2.7. Density

2.7.1. Pycnometer Method

The pycnometer is a container of known volume, usually 50 cm^3. The mass P of the empty pycnometer is known. It is filled with a measuring liquid, preferably toluene, and weighed $(P + F_0)$. The container is then emptied, filled with the fiber whose mass is to be measured $(P + G)$, and filled to the mark with measuring liquid $(P + G + F)$. From these four weight determinations and the known density of the measuring liquid, the density of the fiber can be calculated.

2.7.2. Density Balance (Archimedes' Principle)

A small specimen is weighed in air and then immersed in a liquid of lower density than the fiber. The difference in weight is equal to the weight of the liquid displaced, i.e., the volume of the fiber specimen times the density of the fiber specimen times the density of the liquid.

Displacement liquids used are *n*-heptane ($\varrho = 0.648$ g/cm^3) or methanol ($\varrho = 0.791$ g/cm^3). The standard applied is ASTM D 276.

2.7.3. Gradient Method

A liquid column with a density gradient is prepared by stacking layers of two liquids of different densities (e.g., xylene with $\varrho = 0.9$ g/cm^3 and pentachloroethane with $\varrho = 1.7$ g/cm^3), one above the other. The fiber sample to be measured remains suspended at the height at which the densities of liquid and fiber coincide.

2.8. Chemical Identification of Fibers by Solubility

Specific fiber types are also identified by a combination of microscopic and solubility tests. To save space, a complete solubility test scheme is not presented here (see ASTM 276). Some solvents for fibers are

Wool: potassium hydroxide
Polyamide: hydrochloric acid (6 mol/L)
Polyacrylonitrile: dimethylformamide
Polyester: 1,2-dichlorobenzene
Cellulose acetate: acetone

2.9. Other Identification Methods for Fibers

Other characteristics can be used to identify fibers, provided suitable measuring devices are available:

Fiber density (see Section 2.7)

Melting point (see Section 5.1)
Infrared spectroscopy
Refractive index and birefringence

The methods, including all necessary specifications and characteristic IR spectra, are described in ASTM D 276.

2.10. Identification of Wool and Hair

The wool and hair of different animals consist of the same basic substance; to differentiate among them under a standard microscope is often difficult or impossible. Other characteristics, such as fineness, fiber length, and crimp, are not sufficiently different to assure unambiguous identification. The scanning electron microscope has proved suitable for identifying various wools and hairs. Good information is available in Test Method 20 of AATCC.

Distinguishing characteristics are the scales of the fibers and their thickness. Institutes that have a scanning electron microscope and a staff experienced with animal fibers are capable of carrying out such investigations [7], [8].

2.11. Analysis of Fiber Impurities and Extractable Matter

Natural fibers contain various impurities. In cotton, these impurities originate predominantly from the cotton plant. During ginning, cotton fiber is contaminated further with seed-coat fragments. Scoured wool also contains vegetable impurities which cannot be removed by scouring. In addition, the surface is covered with wax, grease, or oil, either of natural origin or applied subsequently. Those substances which are applied later are called preparations or finish agents and, collectively, extractable matter (i.e., substances that can be removed by organic solvents).

2.11.1. Trash and Dust Content of Cotton (Nonlint Content)

Raw cotton contains various impurities such as stems, leaves, sand, seed-coat fragments, and dust which, collectively, are called trash and dust. Slivers from carding and drawing may also be more or less contaminated, depending on the intensity of cleaning. The various testing machines for determining trash and dust operate on the mechanical – pneumatic principle. Two of these are mentioned below.

Shirley Analyzer. A known weight of raw cotton, partially processed cotton, or waste is fed into the machine. The machine separates the foreign matter from the cotton and discharges the non-lint cotton into a separate chamber. The amounts of lint and nonlint recovered are calculated as a percentage of the original specimen mass (ASTM D 2812).

Micro Dust and Trash Tester. The Institute of Textile Technology and Process Engineering Denkendorf (Germany) has developed a testing machine based on the opening device of a rotor spinning machine, which separates trash and dust from the fiber material. Trash and micro dust are picked up separately by separation and suction, respectively, and weighed [9]. This instrument has been adopted by Zellweger (Uster, Switzerland), modified, and sold under the trade name MDTA 3 (micro dust and trash analyzer).

2.11.2. Honeydew on Cotton

The secretion of the white fly, a sticky sugar-containing substance called honeydew, is deposited on the cotton shrub. Honeydew is very disruptive during spinning. Various procedures have been developed in recent years for determining honeydew, none of which are satisfactory. The highly nonuniform distribution of honeydew deposit creates difficulties. Some procedures utilize chemical test solutions which discolor appropriately in the presence of sugar [10–12], e.g.,

Benedict test
Perkins test
Bremer honeydew test

By means of extensive tests, the "Honeydew Working Group" of the International Cotton Test Conference has found that these tests are not fully applicable to industrial conditions. Various measurement techniques in which the stickiness of fiber mats between heated metal plates is used

as a measure of the honeydew content are being evaluated, and corresponding standards are in preparation.

2.11.3. Grease Content of Wool – Extractable Matter Content

Various solvents can be used to dissolve grease or oil. Because other substances can be picked up at the same time, depending on the solvent, the term "extractable matter" is more accurate.

The following solvents are used for wool:

Ethanol ASTM D 584
Trichlorotrifluoroethane ASTM D 1574
Dichloromethane ISO 3074; IWTO-10-62

Fiber samples are treated with the appropriate solvent in a Soxhlet extractor. The amount of material extracted is referred to the dry or commercial weight of the wool.

2.11.4. Preparation and Finishing Agents for Man-Made Fiber

A Soxhlet extractor is also used for man-made fibers; however, a different solvent is employed. No uniform international standard exists. The following solvents are specified in DIN 54 278:

Methanol: polyester, polyacrylonitrile, and poly(vinyl chloride) fibers
Petroleum ether: polyamide and cellulose acetate fibers

Another reference is ASTM D 2257, Standard Test Method for "Extractable Matter in Yarns," which can be applied to staple fiber. It uses Soxhlet extractions with trichlorotrifluoroethane and then ethanol.

2.12. Maturity of Cotton

Cotton contains both fully matured and immature fibers. The characteristic of immature cotton fibers is a wall width less than the lumen width (Fig. 17). The degree of maturity can be determined by various procedures.

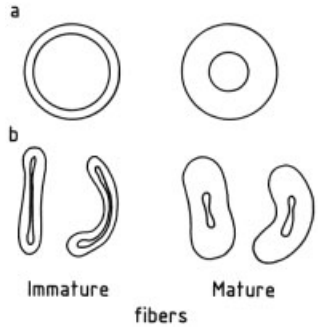

Figure 17. Mature and immature cotton fibers
a) Before husk is opened; b) After husk is opened

Microscope. Fibers are swollen by addition of an 18 % solution of sodium hydroxide. Because of their smaller width, immature fibers can be spotted and counted under a microscope. In polarized light, immature fibers can be recognized even without swelling; the same is true for the microscopic examination of fiber cross sections. Standards are ISO 4912; ASTM D 1442; DIN 53 943 T 2 and T 3.

Fiber Fineness. Fineness involves a determination of the maturity index and is used to estimate the linear density of loose cotton fibers by measuring, with an airflow instrument, the resistance of a plug of cotton fibers to airflow before and after treatment with a sodium hydroxide solution (Causticaire method). The standards are ASTM D 2480; DIN 54 943 T 4.

A new procedure measures the linear density, by using an airflow instrument, at two different compression ratios of the cotton plug. The two readings are substituted in a formula, and the degree of maturity is calculated. The standards are ISO 10 306, EN 30 306, ASTM D 3818.

Dyeing. Immature fibers acquire very different coloration, especially with certain test dyes. The red – green test is based on this property (red coloration = mature; green = immature). The degree of maturity is estimated from a comparison with standard samples.

Fibrograph. Another method for measuring cotton maturity is ASTM D 3817, Standard Test Method for "Maturity Index of Cotton Fibers by Fibrograph" which uses the Uster Fibrograph 730 (see page 6).

3. Tensile Properties

3.1. Tensile Test

To determine its deformation properties, a fiber is clamped at each end in the jaws of a tensile testing machine and loaded in the longitudinal direction until it breaks. According to standards in 1989, only tensile testing machines that operate by constant rate of extension principles may be used. The relevant terms are defined as follows:

Gauge Length. The gauge length is the distance between the clamp jaws as traversed by the mounted specimen (which is not always a straight line between the jaws, as is the case with capstan clamps). The gauge length should not be less than 10 mm; 20 mm or more is preferred. The method ASTM D 2101, which is applied on filaments, uses a 25-mm gauge length.

Breaking Load. The maximum force applied to a specimen in a tension test is called breaking load. The force is expressed in newtons, centinewtons, or millinewtons. In ASTM standards, it is also expressed in gram-force (gf).

Tenacity. The tensile stress expressed as force per unit linear density of the unstrained specimen is called tenacity. Tenacity is expressed in centinewtons or millinewtons per tex; ASTM also uses gram-force per tex (gf/tex). The tenacity corresponding to the breaking load is called *breaking tenacity.*

Elongation. Elongation is the increase in length, expressed in units of length or as a percentage of the gauge length (Fig. 18). The elongation corresponding to the maximum load is called *elongation at break.*

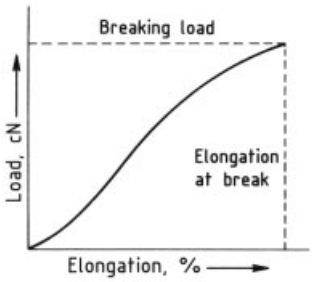

Figure 18. Load – elongation curve

Tensile testing machines must have the precision necessary to measure fibers with sufficient accuracy at a gauge length of 10 mm and at relatively low tensile forces. The rates of elongation are determined as a function of the expected elongation (Table 1). In addition to the numerical indication of force and elongation, the testing machine must be equipped with a device on which the load – elongation curve can be recorded. Elongation rate (percent of initial specimen length in mm) at the corresponding elongation at break

Most testing machines on the market today are combined with equipment for measuring fineness. Fineness is first measured by the vibroscope method on each individual fiber, and the result is fed into a computer. The same fiber is then clamped in the tensile testing machine and tested. The measured force is also fed into the computer, which then calculates the tenacity at break. In addition, the computer calculates an average of all the individual values and all other statistical parameters; as a rule, at least 50 individual values are required per test.

For wet testing, which is common for viscose fiber, for example, the tensile testing machine must be equipped with a suitable moistening device. The fiber clamped in the jaws is wetted directly.

Standards are ISO 5079; EN 25 079; BISFA Rules; ASTM D 3822. The method ASTM D 2101 is used for longer fibers (25-mm gauge length).

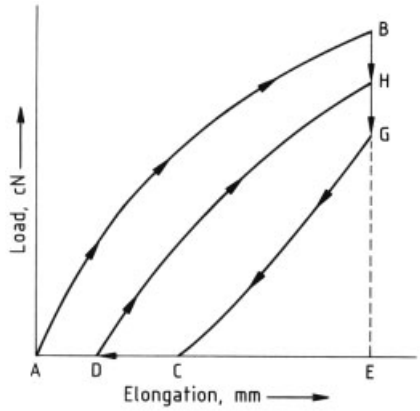

Figure 19. Determination of elastic properties of fibers
A to B: load to required elongation
B to G: wait (simultaneous relaxing of stress)
G to C to A: back to zero elongation; wait
A to D to H: load to same elongation
Permanent deformation, % $\frac{\overline{AD}}{\overline{AE}} \cdot 100$
Tensile strain recovery, % $\frac{\overline{DE}}{\overline{AE}} \cdot 100$

Table 1. Elongation rate (percent of initial specimen length in mm) at the corresponding elongation at break

Estimated elongation at break, %	Elongation rate (ISO, EN)	Estimated elongation at break, %	Elongation rate (ASTM)
< 8	50	< 8	10
≥ 8	100	8 – 100	60
		> 100	240

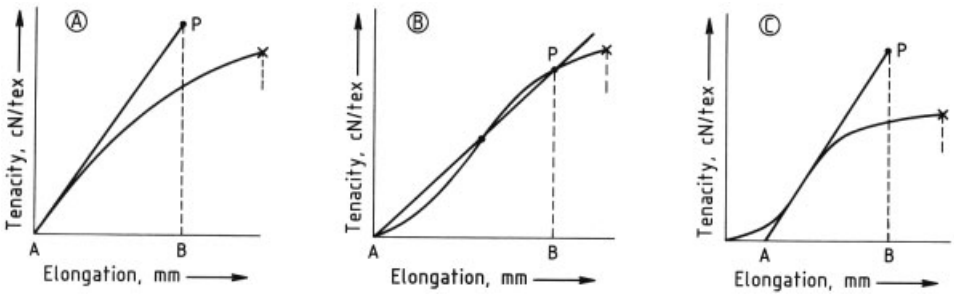

Figure 20. Determination of modulus
A) Initial modulus; B) Secant modulus; C) Tangent modulus
$\left(\frac{\overline{PB}}{\overline{AB}/l_0}\right)$
where l_0 = specimen length (gauge length)

3.2. Elastic Behavior

The elastic properties of fibers are measured by determining the amount of recoverable energy exhibited by a fiber when it is stretched to a predetermined extension at a specific elongation rate, held at this extension for a specified length of time, and finally relaxed at the same rate at which it was extended. The fiber can also be loaded in multiple cycles, and the loading and unloading time in each can be varied.

Permanent deformation and tensile strain recovery can be determined from the load – elongation curves (Fig. 19). These values can be ascertained at various degrees of elongation, e.g., 2, 5, and 10 %. Standards are ASTM D 1774; DIN 53 835.

3.3. Modulus

The modulus is the ratio of the change in tenacity, expressed in centinewtons per tex, to the change in elongation, expressed as a fraction of the original length. These values can be obtained in various ways from the tenacity – elongation curve (Fig. 20). The standard used is ASTM D 3822; ASTM D 2101 is used for longer fibers (25-mm gauge length).

3.4. Loop Tensile Test

Loop breaking tenacity is a fundamental property used to establish limitations on fiber processing and on end use. The specimen, consisting of two single fibers looped in a configuration shown in Figure 21, is broken on a testing machine with a constant rate of extension.

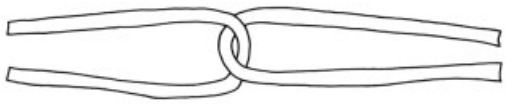

Figure 21. Loop specimen configuration

3.5. Bundle Stress Test

Tensile testing of single fibers is very time-consuming and requires careful handling. The relatively short fiber length of cotton presents additional problems. For this and other reasons, collective test methods have been developed in which a bundle of parallelized fibers is mounted

in suitable clamps and broken. For some machines, zero gauge length is recommended; for others, the gauge length is 3.2 mm (1/8 inch).

Pressley Tester. For the Pressley tester, the standard gauge length is zero, but it can be adjusted to 3.2 mm. A combed, parallelized bundle of cotton is secured in two special clamps and loaded by releasing a weight that slides down an inclined track. When the bundle breaks, the weight stops on the track and the breaking force can be read off the scale. The fiber bundle remaining in the clamps is then weighed so that the breaking force can be related to fineness. The machine is calibrated with standard cotton (test cotton), prescribed by the U.S. Department of Agriculture (USDA). The result is calculated as the Pressley Index (DIN) or as breaking tenacity in gram-force per tex (ASTM). Standards are ISO 300; ASTM D 1445; DIN 53 942.

Stelometer. In a stelometer, force is applied by the pendulum principle. The gauge length is 3.2 mm. In addition to the breaking force, the elongation at break can also be read. By means of fineness (weight and length) of the fiber bundle, tenacity at break is calculated in units of centinewtons per tex and, in the United States, also in gram-force per tex. This instrument is available under the trade name Uster Stelometer 654.

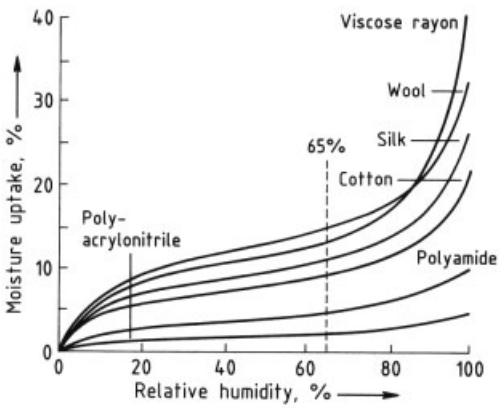

Figure 22. Moisture uptake of different fibers at 25 °C [1]

Test cotton or, if wool is to be tested, test wool can be used for calibration. Both are prescribed by the USDA. Standards are ISO 3060; IWTO-32-82; ASTM D 1445, D 2524. In principle, any tensile testing machine with suitable force range can be used to perform the bundle test, provided appropriate clamps are available (ASTM D 1294).

4. Behavior toward Water

4.1. Moisture Adsorption

Fibers take up more or less water from the atmosphere, depending on hygroscopicity and relative humidity (Fig. 22).

Moisture uptake is subject to hysteresis (see Section 1.4); i.e., fibers take up different amounts of moisture, depending on whether they are brought into the standard climate from the dry or from the wet state. Moisture affects various physical properties and is important in calculating commercial weight.

4.1.1. Oven Drying Test

The fibers are weighed, dried in an oven with hot air (105 °C), and reweighed. The difference between the original mass and the oven-dry mass is the moisture content, expressed in percent. Standards are ISO 6741; IWTO-33-88, IWTO-34-85; BISFA Rules; ASTM D 629, D 1576, D 2495, D 2654; DIN 53 800. For calculating commercial weight, various organizations specify commercial moisture regain, but no uniform international standard existed in 1989 (see also BISFA Rules; ASTM D 1909; DIN 54 201).

4.1.2. Other Tests for Moisture Content of Fibers

Fiber moisture also affects its electrical properties, e.g., electrical resistance or capacitive parameters in a condenser field. On this basis, instruments have been developed with which the moisture content of fibers can be measured. They are usually not accurate enough to determine commercial weights, but because they can be read directly and are easy to use, they have been adopted for control purposes.

Measurement of Electrical Resistance. Two needle-shaped electrodes are inserted in the fiber material (e.g., in the bale) and moisture is measured via resistance.

Capacitive Measurement. Fibers, even whole bales, are measured in the electrical field of a condenser. These instruments must be calibrated by oven drying. This calibration should be carried out at two different moisture levels for each fiber type.

4.2. Water Absorption

Water is taken up from the liquid phase in different ways. Part of the water enters the interior of the fiber, part of it is bound to the fiber surface, and part remains between fibers as a result of capillary action. The last is especially difficult to measure because it is greatly affected by the position of the fibers and the packing density. Therefore, water retention is used as a measure of water absorption because the conditions of measurement are defined more precisely and are easier to observe.

Determination of Water Retention. A fiber specimen is placed in a basket, immersed in water, and subsequently centrifuged. The conditions for centrifuging are defined by the standards

ASTM D 2402: 1 h at an acceleration of 1000 g (9800 m/s^2)
DIN 53 814: 20 min at an acceleration of 800 – 1000 g (7800 – 9800 m/s^2)

In this method, water retained by fiber masses includes (1) water absorbed from the prevailing atmosphere, (2) water imbibed after immersion, and (3) moisture adhering to fiber surfaces after being subjected to acceleration.

5. Thermal Behavior

Changes produced by thermal effects are almost exclusively determined on the yarns or textile fabrics made from fibers rather than on the fibers themselves. Only a few characteristic temperatures can be measured on the fibers and be used to identify fiber types:

Glass transition temperature
Softening range
Melting point
Decomposition temperature
Autoignition point
Char temperature

Two determinations used more frequently with fibers – melting point and burning test – are discussed below.

5.1. Melting Point

The melting point can be determined for fibers by means of a microscope equipped with a special heating device. This is a way of distinguishing nylon 6 from nylon 66, for example (standard ASTM D 276).

5.2. Burning Test

The burning test serves exclusively to identify fibers. Flammability, combustion behavior, odor, and residue are characteristic of individual fiber types (see also Test Method 20 of AATCC). Some examples are

Cellulose Fibers (cotton, viscose rayon)	easily flammable; burns rapidly with bright flame; white-gray ash
Wool	difficultly flammable; odor of burnt hair; coal-like ash
Polyamide	moderately flammable; melts while burning, drips; sphere-like residue

6. Dyeing Properties

6.1. Laboratory Dyeing

Dyeing tests are carried out preponderantly on textile fabrics and yarns. If these are not available, test dyeings are carried out on fiber flock in laboratory dyeing apparatus. In flock dye houses, such test dyeings serve as control samples. The test dyeing itself can take place in simple dye beakers or in a dyeing apparatus with stirrers and programmed heating cycles. For high-temperature dyeing (above the boiling point), a special apparatus is required.

6.2. Identification of Fiber Type by Dyeing (Neocarmine)

Neocarmine dyes (e.g., Neocarmine W [*51394-11-7*], obtainable from FESAGO, Heidelberg, Federal Republic of Germany) are used to identify fiber types by comparison with a color scale. This is an especially useful method if a microscope or other expensive equipment is not available.

7. Processing

7.1. General

The processibility of fibers into yarns or nonwovens is important to spinners and nonwoven producers. This property is highly complex and contains many subproperties and characteristics which make laboratory testing difficult. Starting with a processing test on production machinery, a scaled-down, semiworks machine is constructed in which the number of production units (e.g., the number of spindles on a spinning machine) is reduced. The amount of fiber to be processed is thus kept to a minimum, but a certain amount of information on processibility is lost.

Occasionally, attempts are made to predict processibility with the aid of specially developed testing devices. This is only partially successful. With the present state of the art, only a few processibility characteristics can be estimated from such devices and from spinning tests on semiworks spinning units.

7.2. Spinning Devices for Trials

A mini carding machine is used to open the fibers. Next, the fiber web is parallelized on small draw rolls and spun on a spinning tester. The spun yarn can be checked for various properties. Rovings can also be fed to spinning positions to determine the behavior of fibers in the spinning process or to investigate certain spinning parameters.

Only limited conclusions concerning processibility on production machines can be drawn from such tests, especially if high-production machines are involved. For this reason, a test run under actual operating conditions on the production machine itself is often preferred.

7.3. Electrostatic Charging

Fibers that absorb little moisture, which include all synthetic fibers, tend to be electrostatically charged. They are, therefore, provided with antistatic finishing agents. The antistatic effect depends on the relative humidity of the atmosphere.

The tendency to be charged electrostatically can be measured with the aid of so-called field-intensity meters at the card or mini card or on the drawing frame. The measured field intensity in volts per meter, together with the relative humidity, indicates the running properties that can be expected and possible processing problems.

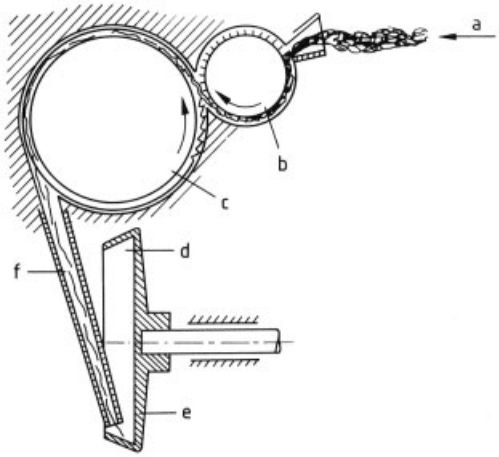

Figure 23. Principle of rotor ring method
a) Feeding of fibers; b) Feeder; c) Opener; d) Rotor ring (fleece of parallelized fibers); e) Rotor; f) Feeding tube

7.4. Rotor Ring Method

The rotor ring device developed at the Institute for Textile Technology and Process Engineering Denkendorf (Rotor Ring Device 580) consists, in essence, of a rotor spinning unit. By measuring the width of the fiber ring taken from the rotor and the electrical output of the opening roll, important conclusions can be drawn concerning the processibility of fibers, crimp, finish

properties, fiber – fiber friction, and fiber – metal friction (Fig. 23) [13].

7.5. Combination of Micro Dust and Trash Analyzer with the Rotor Ring Method and a Spinning Device for Trials

At the Institute for Textile Technology and Process Engineering Denkendorf, a combination of the following test methods has been developed and is marketed by Zellweger under the trade name Uster Quickspin System:

- Micro Dust and Trash Analyzer, MDTA 3
- Rotor Ring Method, Rotorring 3 (see Section 7.4) with a large rotor (diameter ca. 320 mm)
- Quickspin rotor spinning device for procucing a rotor yarn

A modified rotor ring device (see Fig. 23) separates micro dust and trash by filtration and suction and weighs them. The cleaned fibers pass through the feeding tube (f) into the rotor (e), whose diameter of ca. 320 mm is such that the fiber ring removed from the rotor has a length of ca. 1 m. The fiber ring is used for the determination of some of the properties described in Section 7.4. The fiber ring is then cut to give a ca. 1 m long sliver, which is spun into a yarn ca. 200 m in length. Physical investigations on the yarn then allow conclusions to be made about the properties of the fibers.

The Quickspin System is mainly intended for rapid evaluation of raw cotton. It enables cotton dealers and processors to quickly evaluate the quality of a cotton delivery on the basis of small samples.

8. Automated Test Units

Test lines have been developed for testing cotton where several fiber properties can be determined sequentially, largely in an automated or mechanized mode. Control of the operation, calibration, recording, and analysis of results are performed with the aid of computers. The essential properties measured are

- Fiber length (fibrograph principle)
- Fiber strength and elongation (bundle test)
- Fineness (Micronaire)
- Trash content (scanning)
- Color

Other combinations, including measurement of maturity or trash and dust content, are possible. Two manufacturers have produced such units, several hundred of which are in use worldwide:

1) High Volume Testing (HVT) from Spinlab, Inc., Knoxville, Tennessee, and
2) High Volume Instruments (HVI) from Motion Control, Inc., Dallas, Texas

The units are calibrated with test cotton whose properties are determined and approved by a central test laboratory. This can be considered something of a disadvantage. Calibration by means of a standard independent of the material to be tested would be preferable. The advantage of these test lines is that complete measurement of several properties can be performed in a very short time (a few minutes) by one person. With a complete knowledge of raw materials data, directed use in spinning becomes possible.

9. References

General References

1. H. Sommer, F. Winkler: *Handbuch der Werkstoffprüfung,* "Die Prüfung der Textilien," Springer-Verlag, Berlin 1960.
2. P. M. Latzke, R. Hesse: *Textilien, Prüfen – Untersuchen – Auswerten,* Fachverlag Schiele & Schön GmbH, Berlin 1974.
3. E. Wagner: *Mechanisch-technologische Textilprüfungen,* Dr. Spohr-Verlag, Stuttgart 1966.
4. A. Agster: *Färberei- und textilchemische Untersuchungen,* Springer-Verlag, Berlin 1967.

Specific References

5. L. J. Lynch, N. A. Michie, *Text. Res. J.* **46** (1976) 653.
6. A. R. Edmunds, H. McNally, IWTO Technical Committee Report No. 7, June 1987.
7. P. Kusch, G. Stephani, *Schriftenr. Dtsch. Wollforschungsinst. Tech. Hochsch. Aachen* **96** (1984).

8. P. Kusch, W. Arns, *Melliand Textilber.* **64** (1983) 427 – 429.
9. E. Kirschner, E. Bay, *TPI Text. Prax. Int.* **39** (1984) 985 – 990.
10. F. Hadwich, *Melliand Textilber.* **42** (1961) 487.
11. H. H. Perkins, jun., *Text. Bull.* **97** (1971) no. 8, 21.
12. S. Sismann, A. Schenek, *Melliand Textilber.* **65** (1984) 642 – 644.
13. P. Ehrler, O. Schreiber, *TPI Text. Prax. Int.* **39** (1984) 347 – 352.

Fibrinogen → Blood
Fibrinolytics → Blood
Fillers → Paints and Coatings
Fillers → Paper and Pulp
Fillers → Plastics, Additives
Fillers → Rubber, 4. Chemicals
Film Deposition → Thin Films

Films

PETER SCHMITZ, Hoechst AG, Werk Kalle, Wiesbaden, Federal Republic of Germany

SIEGFRIED JANOCHA, Hoechst AG, Werk Kalle, Wiesbaden, Federal Republic of Germany

1. Introduction 2
2. Production Technology 3
2.1. Solution Casting 3
2.2. Thermoplastic Film Forming . . . 5
2.2.1. Calendering 5
2.2.2. Extrusion 5
2.3. Stretching 7
2.3.1. Monoaxial Stretching 7
2.3.2. Biaxial Stretching 7
2.4. Film Modification and Converting 9
2.4.1. Corona Treatment 9
2.4.2. Coating 10
2.4.3. Lamination 10
3. Additives 10
4. Delivery Forms 11
5. Test Methods 11
6. Film Products 13
6.1. Cellulose Derivatives 13
6.1.1. Cellulose Hydrate 13
6.1.2. Cellulose Esters 13
6.2. Hydrocarbon Polymers 18
6.2.1. Low-Density Polyethylene (LDPE) 18
6.2.2. High-Density Polyethylene (HDPE) 18
6.2.3. Linear Low-Density Polyethylene (LLDPE) 18
6.2.4. Polypropylene (PP) 18
6.2.5. Poly(4-methyl-1-pentene) (PMP) . 19
6.2.6. Ionomers 19
6.2.7. Polystyrene (PS) 19
6.3. Halogenated Hydrocarbon Polymers 20
6.3.1. Poly(Vinyl Chloride) (PVC) 20
6.3.1.1. Rigid PVC 20
6.3.1.2. Plasticized PVC 20
6.3.2. Poly(Vinylidene Chloride) (PVDC) 20
6.3.3. Fluoropolymers 21
6.3.3.1. Poly(Vinyl Fluoride) (PVF) 21
6.3.3.2. Poly(Vinylidene Fluoride) (PVDF) 21
6.3.3.3. Polytetrafluoroethylene (PTFE) . . 21
6.3.3.4. Polychlorotrifluoroethylene (PCTFE) 22
6.4. Poly(Methyl Methacrylate) (PMMA) 22
6.5. Poly(Vinyl Alcohol) (PVA) 22
6.6. Ethylene–Vinyl Alcohol Copolymer (EVOH) 22
6.7. Polyamide (PA) 23
6.8. Polyester 23
6.9. Polycarbonate (PC) 24
6.10. Polysulfone (PSU) 24
6.11. Polyimide (PI) 24
6.12. Other Films 24
6.12.1. Polyetherimide (PEI) 24
6.12.2. Polyethersulfone (PES) 25
6.12.3. Poly(ether ketone) (PEK) 25
6.12.4. Poly(Phenylene Sulfide) (PPS) . . . 25
6.13. Composite Films 26
7. Summary of Uses 26
8. Environmental Aspects 28
9. References 28

Abbreviations used in this article:

BOPP biaxially oriented polypropylene
CA cellulose acetate
CAB cellulose acetate butyrate
CAP cellulose acetate propionate
ETFE ethylene–tetrafluoroethylene copolymer
EVM ethylene–vinyl acetate copolymer
EVOH ethylene–vinyl alcohol copolymer
FEP tetrafluoroethylene–hexafluoropropylene copolymer
HDPE high-density polyethylene
LDPE low-density polyethylene
LLDPE linear low-density polyethylene
MD machine direction
PA polyamide
PBT poly(butylene terephthalate)
PC polycarbonate
PCTFE polychlorotrifluoroethylene
PEI polyetherimide
PEK polyetherketone
PEN poly(ethylene-2,6-naphthalene dicarboxylate)
PES polyethersulfone

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_085

PETP poly(ethylene terephthalate)
PFA perfluoroalkoxy copolymer
PI polyimide
PMMA poly(methyl methacrylate)
PMP poly(4-methyl-1-pentene)
PP polypropylene
PPS poly(phenylene sulfide)
PS polystyrene
PSU polysulfone
PTFE polytetrafluoroethylene
PVA poly(vinyl alcohol)
PVC poly(vinyl chloride)
PVDC poly(vinylidene chloride)
PVDF poly(vinylidene fluoride)
PVF poly(vinyl fluoride)
TD transverse direction

1. Introduction

The term *film* is applied to thin materials present in a continuous planar form, which are at the same time self-supporting and flexible. A film may consist of one or more layers, usually of differing composition. The term generally encompasses both metallic and polymeric substances, which may be either homogeneous or made heterogeneous by the presence of incompatible additives or blending components. In this article, a film is assumed to be a polymeric material which can be wound into a roll. Film thicknesses range from 0.5 µm to ca. 1 mm.

The characteristics of a film are determined primarily by the raw materials employed, but they are additionally affected by processing methods, modification, and converting (e.g., stretching, coating and lamination).

Film materials are of considerable economic importance. In 1986, the annual worldwide consumption of film for packaging and technical applications was ca. 11×10^6 t. Figures for the worldwide consumption of the more important film materials are listed in Table 1.

Films are used as flexible packaging materials, as insulation for electrical conductors, as protective layers on sensitive surfaces, and as windable base materials for subsequent application of magnetic coatings, electrically conducting, and light-sensitive layers as well as printing inks. Multitudes of film and sheet types have been developed to meet the increasing demands of these applications. These products provide a broad spectrum of mechanical, optical, and electrical properties, as well as various degrees of permeability and stability to temperature and aggressive media.

Table 1. Worldwide consumption of some of the most important types of film in 1986*

Material	Consumption, 10^6 t
Low-density polyethylene (LDPE)	8.00
Polypropylene, biaxially oriented (BOPP)	0.55
Poly(vinyl chloride) (PVC), rigid and plasticized	1.42
Poly(ethylene terephthalate) (PETP), biaxially oriented	0.55
Cellophane	0.26

* Hoechst marketing research, Kalle Works, Federal Republic of Germany.

Historical Aspects [1]. Polymeric films were first developed at the end of the 19th century, when cellulose nitrate was introduced as a base material for photographic emulsions. Other cellulose-based films followed, the most important of which were derived from cellulose acetate and cellulose hydrate. Cellulose triacetate films quickly replaced the highly flammable cellulose nitrate films in the photographic industry; cellulose hydrate films (cellophanes) became important packaging materials, maintaining their dominance until polyolefin films were introduced in the 1950s. Cellophane coated with cellulose nitrate or poly(vinylidene chloride) also played a key role in this respect due to its low permeability to water vapor and oxygen, coupled with desirable sealing properties.

All cellulose-based films are prepared from polymer solutions, however the subsequent introduction of totally synthetic polymers has led to new film production methods, as well as new types of films with unique properties. These developments have been encouraged by the close working relationship between fiber and film producers, particularly in terms of polymer compositions and processing technology.

Important milestones in the development of films included the introduction of

1) poly(vinyl chloride) films in the late 1930s
2) low-density polyethylene films in the late 1940s
3) biaxially oriented polyester and polypropylene films in the 1950s

Parallel and subsequent to these developments, a series of specialty films has emerged. These films are distinguished either by high mechanical or thermal stability, or by special properties such as impermeability to oxygen, advantageous electrical characteristics, or resistance to chemical attack.

2. Production Technology [2]

Generally speaking, the manufacture of films from both natural and synthetic polymers relies on two basic techniques:

1) solution casting
2) thermoplastic film forming

The method chosen is generally determined by the behavior of the starting materials with respect to solvents and temperature. However, in certain cases the desired product characteristics are decisive.

For *solution casting,* the raw materials must first be dissolved in an appropriate solvent. The solution is then pressed through a slot die either onto a moving belt or into a precipitation bath.

*Thermoplastic film forming*can be divided into two categories: (1) calendering (film formation in a rolling process) and (2) film formation from a melt (an extrusion process).

2.1. Solution Casting

Solution casting is applicable to soluble polymers that are difficult or impossible to thermoform. It also has advantages in the preparation of films for optical applications. Cellulose acetate, polycarbonate, and polyimide films are prepared in this way.

Solution casting consists of three stages:

1) preparation of the casting solution
2) film formation on a moving surface (drum or endless belt) or in a precipitation bath
3) removal of residual solvent

Figure 1 is a schematic illustration of a belt casting machine that is suitable for the production of cellulose acetate film.

The casting solution is prepared in a large agitator vessel, if possible using a low-boiling solvent. The solution is normally very viscous (ca. 20 Pa · s) and is passed through a filter press equipped with a cotton filter. After the viscosity has been checked, the solution finally enters the casting reservoir. Entrapped air is removed either by evacuation or by heating.

The solution is fed from the reservoir under pressure through the die and is cast onto a rotating drum or a moving metal belt. The die gap is adjusted according to the desired thickness of the film.

The metal belt may be made of copper, nickel, or steel, and its surface finish is of critical importance in determining the characteristics of the resulting film. For optical-grade films, the surface must be polished and filtration must be highly effective. Belt dimensions range from 25 to 55 m in length and 1 to 2 m in width; belt speed varies from 1 to 60 m/min, depending on the desired thickness of the film.

The film is dried on the belt by a countercurrent of warm air and is then removed, further dried, and wound. Typical thicknesses of films prepared in this way are 15 – 250 µm.

In some cases, the polymer solution emerges from a slot die directly into a liquid precipitation bath. The polymer solvent is miscible with the bath fluid, whereas the polymer is insoluble in this fluid. Consequently, the polymer coagulates in the bath to form a gel film with a microporous structure. Thus, this process is frequently employed for the preparation of membranes (→ Membranes and Membrane Separations).

Cellulose hydrate (cellophane) films [3] are produced by a modification of this process in which the gel film undergoes a chemical reaction. The casting solution is prepared by reacting cellulose with caustic soda solution and carbon disulfide to form viscose. The resulting xanthate solution is then pressed into a spinning bath containing sulfuric acid, causing decomposition of the xanthate and forming a gel web composed of regenerated cellulose (see also → Cellulose, Chap. 3.1.).

A slot die is immersed in the precipitation bath (see Fig. 2). Shortly after its formation, the gel film possesses sufficient stability to permit its transport over rollers. The film is subsequently washed and bleached, passed through plasticizer baths, dried, and wound.

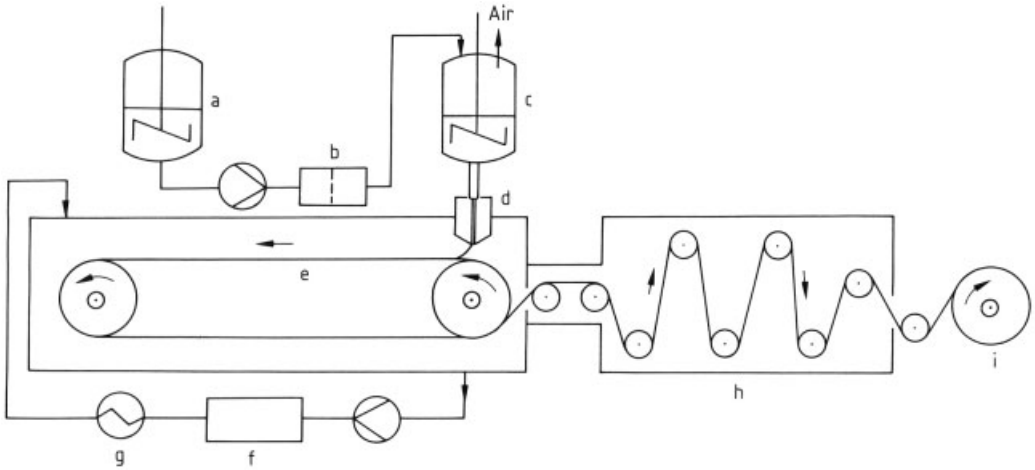

Figure 1. Film production by solution casting on an endless belt
a) Agitator vessel; b) Filter; c) Degasser/casting reservoir; d) Die/caster; e) Metal belt; f) Solvent recovery; g) Air heater; h) Final dryer; i) Winding

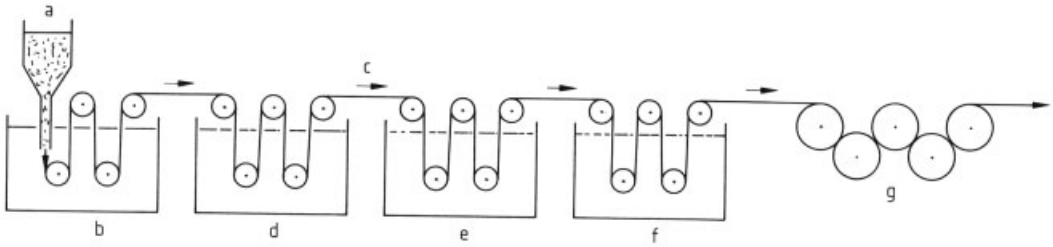

Figure 2. Production of cellulose hydrate film by casting into a precipitation bath
a) Casting solution, slot die; b) Precipitation bath; c) Regenerated film web; d) Regeneration bath; e) Wash and bleaching bath; f) Plasticizer bath; g) Drying rolls

Film thickness ranges from 12 to 45 µm with a typical width of ca. 2 m. Production rates of 80 m/min are common for films in the mid-thickness range. The normal slot die may be replaced by an annular orifice, in which case the product is a tube of regenerated cellulose [4]. Further processing is as described above for flat cellulose hydrate films.

2.2. Thermoplastic Film Forming

2.2.1. Calendering [5–7]

Only polymers with a broad range of plasticity and high viscosity can be processed by the calendering method; these criteria are met by homo- and copolymers of poly(vinyl chloride) (PVC).

The production of calendered (e.g., PVC) film is shown in Figure 3; three steps are involved:

1) preparation of a blend of the components in a mixer
2) homogenization and plastification
3) film formation in the calender

The powdered mixture is usually converted into a fused mass by continuous compounding in a kneader or planetary-gear extruder using heat, shearing forces, and pressure. The mass is then rolled into a film web, ideally with a four- or five-roll calender, in which the nip rolls are mounted in an "L"-form. These rolls are made of steel or cast iron and they generally have a smooth surface. The nip rolls are heated to 180 – 220 °C, depending on the process and the polymer mixture; heating must be uniform. The calendered

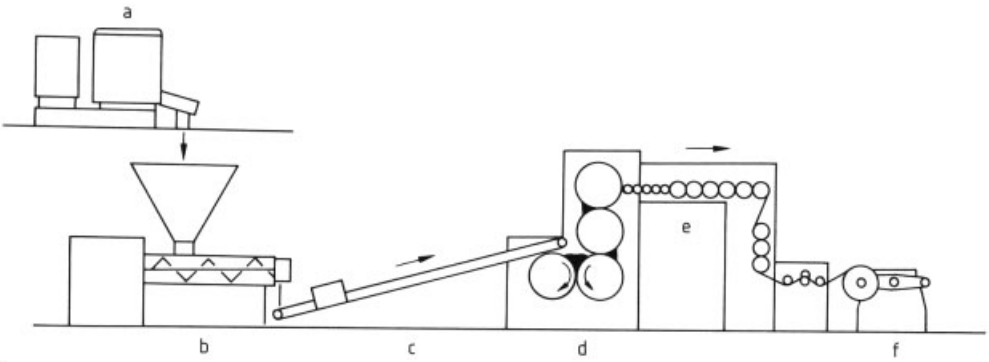

Figure 3. Film production by calendering
a) Mixer; b) Kneader; c) Feed band; d) Calender; e) Cooling and tempering rolls; f) Winder

web then passes over cooling rolls followed by an edge-trimmer, in some cases a stretching facility, and finally a winding unit. Calendering equipment operates at high pressure and must, therefore, be solidly constructed, making it very expensive. The largest production facilities for processing plasticized PVC have nip rolls 85 cm in diameter and are 280 cm wide. Calenders for rigid PVC utilize rolls up to 70 cm in diameter and are 250 cm wide. Production rates of 50 – 100 m/min are achieved depending on the desired film thickness; with stretching, line speeds approach 300 m/min. These figures correspond to an output of 500 – 2000 kg/h.

2.2.2. Extrusion [8]

Advantages of extrusion include a wide range of attainable film thickness, large-scale processing, and the fact that recovery systems for expensive solvents are not needed.

The technique utilizes an extruder to melt the polymer, a die to confer shape to the melt, and a cooling system to solidify it. The process is applicable only to raw materials for which there is a sufficiently large temperature difference between melting and the onset of decomposition (e.g., polyolefins, polyamides, and polyesters).

Chill Roll Extrusion . Figure 4 is a schematic illustration of a chill roll extrusion line. Moisture-containing raw materials (moisture content usually $\ll 1\,\%$) are subjected to continuous or discontinuous drying, preferably utilizing heated air or a vacuum, prior to melting in the extruder.

Various extruder designs can be used including single-screw and twin-screw extruders, the latter with either corotating or contrarotating screws. Extremely high outputs of 4 – 5 t/h can be achieved with two extruders combined in sequence (tandem extruder). Single-screw extruders are the most common. Screw geometry must be adapted to the raw materials being processed and has a major effect on the output and quality of the melt. The largest single-screw extruders currently in use have a screw diameter (D) of 30 cm and a screw length of 30 D.

The melt is filtered in one of several ways depending on the desired purity: wire mesh, sintered metal, and wire–felt disks are all used. The filtered raw materials are forced through a slot die. The required pressure can normally be achieved with the extruder, however in special cases a gear pump is installed before the filter to ensure constant metering of the melt. In coextrusion (see Section 2.4.2), the die combines several different melt streams. The width of the die and the die gap determine the width and thickness of the film, respectively. In large production units, the dies may be as wide as 3.5 m. Adjustment of the die gap permits the film thickness profile to be corrected during production. Initial adjustment and readjustment can be performed either manually or automatically, e.g., by thermal expansion of adjusting screws that are heated as a function of the relative film thickness.

When the melt film emerges from the die, it is cast onto a highly polished, chrome-plated chill roll. The takeoff speed depends on both film thickness and output and may be as high as

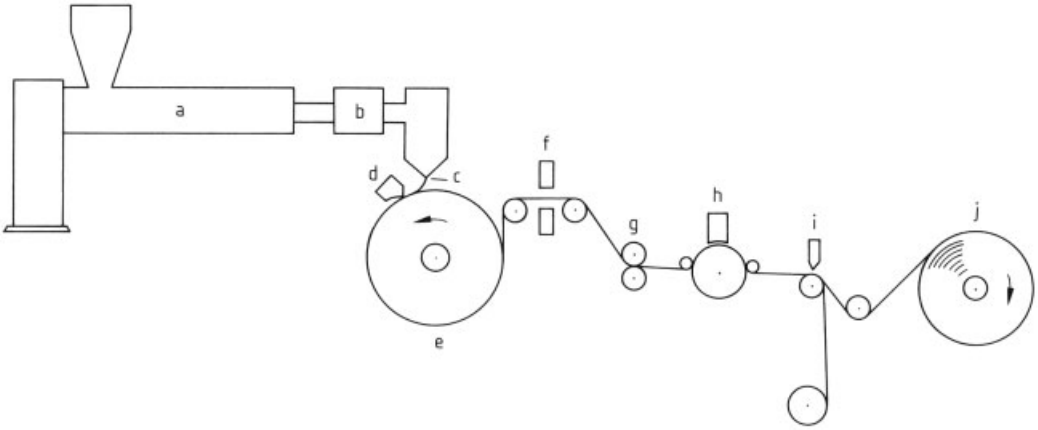

Figure 4. Film production by chill roll extrusion
a) Extruder; b) Filter; c) Slot die; d) Air knife; e) Cooling roll; f) Device for measuring film thickness; g) Tension rolls; h) Corona treatment; i) Edge trimming; j) Winding

100 m/min. Entrapment of air between the melt and the roller surface is prevented by using an air jet (air knife), electrostatic "pinning", or a vacuum device.

The thickness of the film is measured with the aid of a traversing radiation source and detector. The film is then edge-trimmed and wound to a roll. If the film is to be printed, corona pretreatment can also be carried out (see Section 2.4.1).

The film is usually cut to the designated width in a separate process–the actual width depending on the application. If monoaxial or biaxial orientation is desired, the film is fed directly into a stretching unit (see Section 2.3) without intermediate winding.

Film Blowing (Blown-Bubble Extrusion). In the film blowing process, the filtered melt is pressed through an annular die instead of a slot die; this results in the production of a tubular film (Fig. 5). The tube (or "bubble") is inflated by means of air blown through the center of the die. The film is thereby expanded at a temperature close to its melting point and is then cooled by an airstream located above the die. A number of techniques are employed for air cooling; the most common involves passing the tube through a cooling ring. Supplementary internal cooling is also usually used if a high throughput is required. If biaxial stretching is to be performed, the melt must be solidified immediately after it emerges from the die (see Section 2.3.2).

The cooled tube is collapsed by transporting it through a convergent collapsing frame and a set of nip rolls. The film can then either be wound directly as a flattened tube or slit open on both sides and directed to two separate winding stations. The largest film blowing systems utilize annular dies with a diameter of 2 m and an output of 1.5 t/h.

The film blowing process is employed mainly for polyolefin and polyamide films. Multilayer films can be produced by coextrusion of tubular films (see Section 2.4.2).

2.3. Stretching [8]

Many films acquire their desired properties only after stretching. Stretching below the softening temperature orients the polymer molecules and crystallites in the direction of stretch, leading to changes in the stress–strain characteristics of the material. Both the modulus of elasticity and the ultimate strength are increased by this process, but elongation at break and dimensional stability with respect to temperature decrease. A number of other characteristics may be altered as well, depending on the type of film involved, e.g. gas permeability, electrical properties, heat sealability, abrasion resistance, coefficient of friction, puncture resistance, density, and optical properties. Stretching can be conducted either solely in the machine (longitudinal) direction (monoaxial

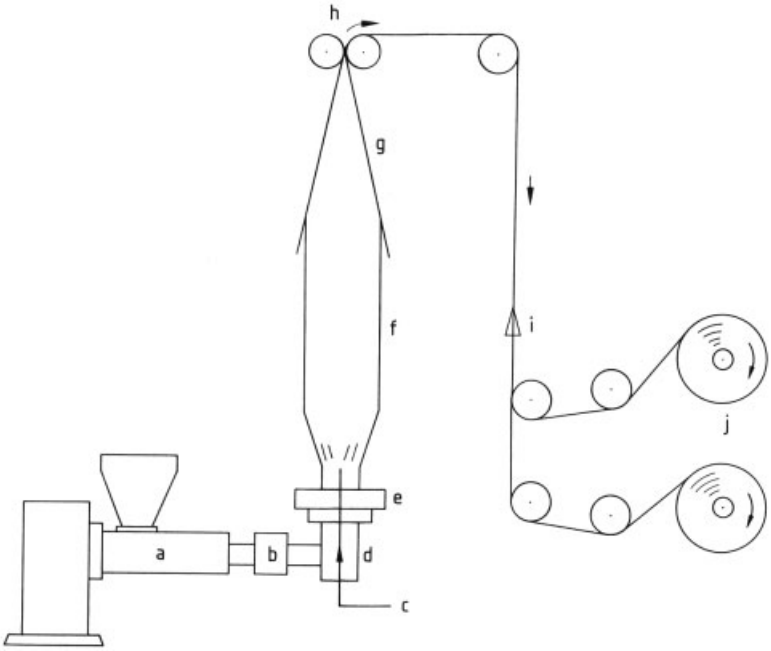

Figure 5. Film production by the blowing method
a) Extruder; b) Filter; c) Blowing air; d) Annular die; e) Air cooling ring; f) Film tube (bubble); g) Flattening; h) Nip rolls; i) Tube separation (slitting); j) Winding

stretching) or in both the machine and the transverse directions (biaxial stretching).

Monoaxial stretching (see Section 2.3.1), which is also commonly used in the production of high-strength fibers, produces films of enhanced strength in the direction of manufacture. The resulting films are thus anisotropic.

Biaxial stretching (see Section 2.3.2) can be conducted in one or two steps (simultaneous and sequential stretching, respectively). Sequential stretching may result in either anisotropic or isotropic films depending on the ratio of the final length to the initial length (the draw ratio). Simultaneous stretching tends to produce isotropic films.

The majority of films prepared from orientable polymers are amenable to either sequential or simultaneous biaxial stretching. However, films derived from polymers with a high rate of crystallization during orientation (stress crystallization) [e.g., some polyamides, poly(butylene terephthalate), and ethylene–tetrafluoroethylene copolymer] can only be oriented by simultaneous stretching.

2.3.1. Monoaxial Stretching

Monoaxial stretching is accomplished by subjecting a film to an increasing velocity gradient as it passes through one or more gaps between temperature-controlled rolls. Numerous stretching machines are available; they differ with respect to the number and arrangement of the rolls. Monoaxial stretching is of particular importance in the manufacture of high-strength tape from polyethylene and polypropylene, as well as for the orientation of poly(vinyl chloride), polycarbonate, and poly(vinylidene fluoride) films. Longitudinal stretching units are also incorporated in film production lines that employ sequential stretching technology for the manufacture of biaxially oriented films (see Section 2.3.2).

2.3.2. Biaxial Stretching

Sequential Stretching. Sequential biaxial stretching is currently the preferred orienting method for example with polyester and polypropylene films. Figure 6 illustrates a production

line that uses longitudinal–lateral stretching (the so-called machine direction–transverse direction, MD−TD sequence) for the manufacture of biaxially stretched film. The film line consists essentially of a chill roll facility coupled with forward and transverse stretching units.

The polymer is first dried, melted in an extruder, passed through a slot die, and vitrified on a chill roll. Longitudinal stretching follows using a system of rollers, whereas drawing in the transverse direction occurs in an air-heated oven (stenter) comprising several temperature-controlled zones. When the longitudinally stretched film enters the stenter, it is grasped at both edges by two series of continuously circulating clips which are linked together in a chain. In the zone where the film reaches the stretching temperature, the two series of parallel-running clips are gradually drawn apart and hence stretch the film in the transverse direction (Fig. 6). The film is then heat-set (crystallized), cooled, released from the clips, and wound to a roll. Draw ratios for polyester lie typically between 3 : 1 and 4 : 1 [9], [10].

Facilities for edge-trimming, thickness determination, and, where appropriate, corona treatment (see Section 2.4.1) are located between the stenter and the winding unit. Winding speeds range from 80 to 250 m/min. Film thickness varies from 1 to 350 μm.

Polyester films with a high modulus of elasticity in the machine direction are best prepared by a TD−MD stretching sequence, or in a three-step process with MD−TD−MD stretching.

Simultaneous Stretching. Simultaneous biaxial orientation is most easily achieved by inflation of a tubular film. In the continuous process, a tube is first created by cooling the melt in a water bath or with a cooling mandrel. This tube is subsequently heated to the stretching temperature with a ring of radiant heaters. It is then inflated, collapsed, and transported as described for blown films (see Section 2.2.2). This technique is widely used for preparing biaxially oriented poly(vinylidene chloride), polyamide, polyester, and especially polypropylene tubular films; all of these films have a high shrinking potential. For this reason, oriented polypropylene tubular films are normally slit, separated, and then heat-set, either with the aid of heated rollers or in a hot-air tunnel.

Simultaneous biaxial stretching of a flat film is technically more complicated than that of tubular materials. Normally, a special stretching (stenter) frame is used in which the clips are not linked in a chain as described for sequential stretching. The clips are individually accelerated as they pass through the transverse-stretching zone by means of a progressive transport system.

2.4. Film Modification and Converting

Films are normally intermediates from which end products are manufactured in subsequent processing steps. The terms *film modification* and *converting* refer to operations such as surface treatment, coating, or lamination, which enhance the value of a film. Commonly encountered end products derived from films include packaging materials (e.g., printed bags and adhesive tapes), capacitors, magnetic tapes, and photographic films.

Untreated films often lack the specific characteristics that are required in the final product, e.g., stability during handling, heat sealability, wettability, adhesion of printing ink or other coatings, and barrier properties. Further processing is then needed to remedy these deficiencies. Sometimes the desired results can be obtained with additives (cf. Chap. 3), which include antistatic agents and lubricants. However, surface treatment or coating is often necessary, and in some cases two or more films must be combined to form a laminate.

2.4.1. Corona Treatment

Corona treatment is the principal method of preparing a film surface for printing or coating. A continuously moving film is exposed to an ionized atmosphere, which is produced by an alternating electrical field [11]. The changes induced in the film surface enhance both its wettability and adhesion properties. Corona treaters can be integrated into the film production process (see Sections 2.2.2 and 2.3.2) and are widely utilized in the pretreatment of polyolefin films.

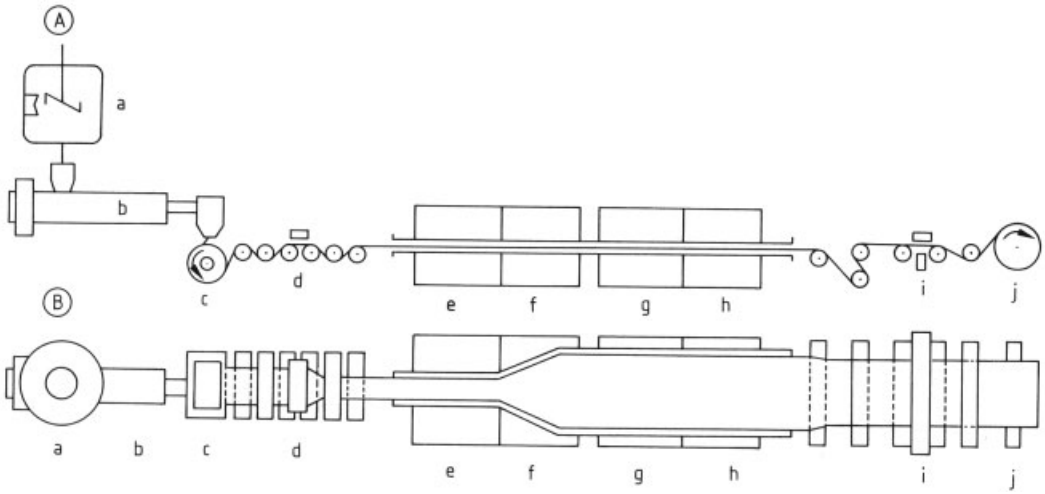

Figure 6. Production of biaxially oriented film by sequential stretching in a machine direction–transverse direction (MD–TD) sequence [production unit viewed from the side (A) and from above (B)]
a) Dryer; b) Extruder; c) Slot die and chill roll; d) Stretching in the machine direction (MD); e) Preheating zone; f) Zone for stretching in the transverse direction (TD); g) Heat-setting (crystallizing) zone; h) Cooling zone; i) Thickness measurement; j) Winding

2.4.2. Coating

Films may be coated with dissolved, dispersed, emulsified, or molten materials [8], [12]. The techniques employed differ considerably, depending on whether the material is applied as a liquid or as a melt.

Liquid materials are normally coated on one or both sides of a moving film web either by immersion or with the aid of applicator rolls. The coated film is then dried in a heated-air oven or by means of rollers.

Coating is normally performed separately from film production (*off-line coating*), although in some special cases, the two processes can be integrated (*in-line coating*) [13]. In in-line coating (e.g., of polyester films), the coating assemblies are preferentially placed between the MD and TD stretching units.

Coating materials such as nitrocellulose, polyacrylates, and poly(vinylidene chloride) are used to improve heat sealability and gas barrier properties. Adhesion is improved by coating films with polyurethanes, cross-linkable polyacrylate, copolyesters, or epoxides. Other coatings include antistatic or slip agents, scratchresistant and antiadhesive materials.

Films with a different functional coating on each surface possess a combination of properties (e.g., sealing and barrier properties) and can be produced on a large scale.

Molten materials are applied to films with the aid of a die by a process analogous to chill extrusion (see Section 2.2.2). The film itself is supported by a roller. Extrusion coating can be conducted either off- or in-line; the latter approach is employed with biaxially stretched films.

Coextrusion represents a special case of melt coating [14], [15] in which two or more polymer melt streams are combined either in the melt pipes or in a multimanifold die just before the material emerges from the die lips. In the former case, special adapters of varying design ensure that the melt streams do not mix in the die and hence leave the die gap in the form of a multilayer film analogous to that produced by a multimanifold die. This process is also adaptable to annular die systems.

Coextrusion has gained considerable importance. Sealable, biaxially oriented polypropylene film with a triple-layer structure is a particularly important product in the packaging indus-

try. The thickness of the sealable layer is typically 0.5 µm.

Coextruded, double-layer films are commonly employed for magnetic tape and capacitors; they have a smooth side which can be coated, and a rough side, which ensures a low coefficient of friction for adequate processability.

2.4.3. Lamination [8], [12]

Two or more different films can be bonded to form a composite or multilayer film. Their individual properties are thus combined, resulting in a variety of high-quality materials for both the packaging industry and the technical sector. For example, a heat-sealable film may be laminated to a film that acts as a barrier to gases and/or to a high-strength film. The potential number of property combinations is very large.

Laminated composite films are commonly prepared by the *dry lamination process,* in which a dissolved or dispersed adhesive is coated on the surface of the film. After evaporation of the solvent in a dryer, the coated film is bonded to a second film web in a laminating unit consisting of a pair of nip rolls, one made of rubber and the other of steel.

The simplest laminating system is composed of two unwinding stations, an adhesive applicator, a dryer, a laminating unit, and a windup station. However, many different systems are in use, depending on the desired number of laminate layers and the choice of adhesive. In some cases, adhesion can be achieved without using solvents, e.g., by using hot melts, reactive adhesives, or systems that can be hardened by irradiation.

3. Additives

Nearly all films contain additives which perform two main functions: (1) stabilization of the polymers and (2) improvement of processability and the potential uses of the film [16], [17] (see also → Plastics, Additives, → Plasticizers).

Additives can be introduced when the polymer is being prepared or processed, or during film production. An additive must withstand all stages of processing without any loss of activity. Additives used in films that come into contact with food must also comply with legal health and safety requirements.

Processing Stabilizers. The thermal decomposition of some film raw materials must be suppressed by addition of processing stabilizers so that the melt can be thermoformed. The choice of stabilizer depends on the decomposition mechanism involved. Thus, sterically hindered phenols effectively stabilize polypropylene, while poly(vinyl chloride) responds to treatment with metal salts (in particular organic tin salts) or *N,N′*-diphenylthiourea [*102-08-9*]. Additive concentrations range from <1 to 3 wt %.

Plasticizers. Plasticizers are particularly useful for adjusting the flexibility of poly(vinyl chloride), cellulose acetate, and cellulose hydrate films. The plasticizer concentration may be as high as 40 % (plasticized PVC); typical examples include esters of phthalic and phosphoric acids.

Slip or Antiblocking Agents. Films with smooth surfaces possess a high coefficient of friction and therefore tend to block during winding and other processing operations. To improve the frictional properties of these films, their surface is finely structured by adding small amounts of finely dispersed pigments. The surface of biaxially oriented polyester films can be improved by adding inert inorganic pigments (e.g., calcium carbonate, silicon dioxide, or calcium phosphate) or particles composed of crosslinked polyacrylates. The particles are < 1 µm in diameter, and are used at a concentration of < 0.5 wt %. Lubricants such as fatty amides or polysiloxanes may also be introduced.

Fillers and Dyes. The optical properties of films can be altered by the incorporation of fillers and dyes. A wide range of concentrations is employed. Many fillers are inert inorganic substances, e.g., titanium dioxide, barium sulfate, kaolin, and carbon black. Colored films are normally produced from raw materials that have been dyed in bulk. However, in special cases the films themselves may be dyed either by a diffusion process or by printing the entire surface.

Antistatic Agents. Films are rendered antistatic by mixing the virgin polymer with a substance bearing hydrophilic groups. The additive migrates to the surface layer. Quaternary ammonium compounds, ethyl sulfonates and, especially with polyolefin films, ethoxylated alkyl amines are used as antistatic agents.

Weathering Stabilizers. Most films for outdoor applications are prone to photooxidative degradation and are stabilized by adding UV-absorbing substances. Piperidyl derivatives, for example, are effective stabilizers of polypropylene.

4. Delivery Forms

Commercial quantities of films are available in various thicknesses (ca. 1 – 1000 µm) and are supplied in the form of rolls with cores of a standard diameter (e.g., 6″).

Standardized thicknesses are demanded for many applications. Attempts are being made to establish a standardized system, since this also simplifies production planning and stock control. Films are offered in widths ranging from a few millimeters to several meters. Standard widths have only been established in a few areas of application.

The maximum length of a film depends on the intended application as well as on the film thickness. For example, with poly(ethylene terephthalate) (PETP), films with a thickness of 5 µm are supplied in rolls with a typical length of 30 km.

Films that are precut to size are also marketed, although only to a very limited extent.

5. Test Methods [18–24]

Many of the test methods employed for process control in film manufacture have been standardized. However, the results obtained are not always compatible because the test conditions specified in different standards often vary.

The *chemical composition* of a film is normally apparent from its IR spectrum. Trace elements are determined by atomic absorption spectroscopy, mass spectrometry, or neutron activation analysis. Relative molecular masses can be measured from the viscosity of a film solution. The identity of a film can often be determined more simply with the aid of systematic analytical schemes (e.g., [20], Appendix B).

The *average thickness* of a film is established according to DIN 53 370 or ASTM E 252. No standard method for measuring thickness in both the machine and transverse directions exists as yet. During manufacture, film thickness can be determined by measuring the absorption of either β- or IR radiation or by a dielectric method. The uniformity of the thickness of a film sample can be determined mechanically, electrically, or optically. The accuracy of measurement in all cases is 0.1 µm.

Methods for determining the *mechanical properties* of films are described in ASTM D 882, D 1822, D 1894, D 1922; ISO 1184 (1983). Individual methods are as follows: tear strength, DIN 53 455; elongation at break, DIN 53 457; modulus of elasticity, DIN 53 448; impact strength, DIN 53 374; tear resistance, DIN 53 375; and tear propagation resistance, DIN 40 634 and DIN 53 363.

The *coefficient of friction* is of considerable importance in the converting of films, and conventional methods for its determination are described in DIN 53 375 and ASTM D 1894. Special techniques have been developed for some applications (e.g., for magnetic tapes).

Film density is determined according to ASTM 1505 IC.

The *electrical properties* of films are tested according to VDE 0345/1.65 and 0303/7/1.65; ASTM D 149, D 150, D 229, and D 252; and DIN/sec 15 C. Parameters that are measured include the dielectric strength, dielectric constant, dissipation factor, resistivity, surface resistance, electrolytic corrosion, and glow discharge ignition voltage.

The melting point, glass transition temperature, crystallization and *softening behavior,* and *decomposition temperature* are easily determined by differential thermal analysis. The *long-term thermal stability* of films designed for use as electrical insulators is normally determined according to VDE 0530 or IEC 216; the films can then be classified according to the temperature at which they may be used, i.e., Class T (90 °C), A (105 °C), E (120 °C), B (130 °C), F (155 °C), H (180 °C), and C (> 180 °C).

The coefficient of expansion, dimensional stability, and *thermal conductivity* can be determined by reference to the following standards, which are not, however, exclusively designed for films: DIN 1341, 40 634, 52 328; ASTM D 229, D 1204; VDE 0530/1/1.66.

The following standards apply to the measurement of *optical properties* (transparency, haze, gloss, refractive index, and color): DIN 5033, 53 491; ASTM 1003, D 307, D 1925, D 2457. The measurements are easy to perform, but interpretation of the results can be difficult.

Determination of *permeability* with respect to liquids, vapors, and gases is relatively complicated, but this parameter is especially valuable, since films are widely used as packaging materials. Appropriate procedures are described in DIN 53 122 and 53 386, and ASTM D 1434 and E 96 (E).

Chemical resistance is most easily determined by examining the changes induced in ultimate strength and elongation at break as a result of exposure to the substances concerned. However, the following standards are also applicable: DIN 53 476; ASTM D 543, D 756, D 1435, E 96; ISO/R 175 and R 511.

Water absorption is determined according to DIN 53 472 and ASTM 0570.

The most useful test of *surface structure* is carried out with a profilometer. Since the 1980s this technique has proved itself to be particularly useful for examining film substrates used in the manufacture of magnetic tapes. The best known procedures are those described in DIN 4760 – 4774, the Japanese Industry Standard JIS 0601-70, and ISO/DIS 4267/1. In special circumstances, additional information can be obtained by using microscopy (interference methods and scanning electron microscopy).

The *roll hardness* of film rolls can be evaluated by using the Shore hardness as a criterion (DIN 53 505, ISO 12 868). A better procedure makes use of a falling sphere [24].

The *flammability* of films is evaluated according to DIN 4102, ASTM D 1433, ISO R 1715, and Underwriters Laboratory Standard UL 94 VTM.

Measurement of the *extraction of low molecular mass components* of films with solvents has not been standardized. Most studies are based on special procedures developed by individual manufacturers and end users. Interest centers on films designed for the packaging industry. Information regarding the extractable components of a film often serves as the basic for its legal approval (e.g., FDA approval) [25].

6. Film Products [20], [26], [27]

A remarkable variety of film materials has been developed as a result of several factors: Film manufacturers have a wide choice of potential raw materials and can also use special processing steps such as stretching to modify certain properties of a film during its manufacture. Further variants are made possible by the introduction of additives, by the application of functional coatings, or by lamination of two or more types of film. The following discussion is limited to representative examples from characteristic film groups. Important physical properties of these materials are compiled in Tables 2, 3, 4, 5. (The lists of trade names supplied are by no means exhaustive.)

6.1. Cellulose Derivatives

6.1.1. Cellulose Hydrate (→ Cellulose, Chap. 3.1.)

Cellophane [*9005-81-6*] is a trade name applied to films based on cellulose hydrate (regenerated cellulose).

Cellophane is prepared from cellulose via the sodium xanthate derivative in a solution casting process (see Section 2.1) and is available in thicknesses ranging from 12 to 45 µm. Untreated cellophane film is crystal clear, glossy, and highly permeable to water vapor; it is unstable with respect to water, but stable to fats and oils. Cellophane is not heat-sealable, but can be bonded; it is odorless, tasteless, and authorized for use as a food packaging material by the FDA and similar institutions. It possesses desirable mechanical characteristics at the normal water content of 7 – 8 %. Properties are listed in Table 2.

Cellophane is increasingly being replaced by oriented polypropylene films; 70 – 80 % of its original applications (mainly in packaging) have already been affected.

Table 2. Properties of selected films [a]

	Cellophane (uncoated)	Cellulose acetate (CA) [b]	Polyethylene (PE)		Poly(4-methyl-1-pentene) (PMP)	Polypropylene (BOPP)
			low-density (LDPE)	high-density (HDPE)		
Mechanical properties						
Thicknesses, µm	12 – 45	12 – 350	25 – 200	50 – 1000	75 – 300	4 – 80
Density, g/cm^3	1.45	1.3	0.92	0.95	0.83	0.91
Ultimate strength						
longitudinal, N/mm^2	100	100	25 [g]	23 [g]	24	140
lateral, N/mm^2	60	100	20 [g]	23 [g]	24	280
Elongation at break						
longitudinal, %	20	30	400 [g]	400	10	140
lateral, %	60	30	700 [g]	400	10	50
Modulus of elasticity						
longitudinal, N/mm^2	5 300	1500	170	900	1100	2 000
lateral, N/mm^2	2 800	1500	170	900	1100	4 000
Electrical properties						
Dielectric strength [c], kV/mm	100	150	200	200	65	340
Dielectric constant at 1 kHz	3.2	4.5	2.3	2.2	2.1	2.2
Dissipation factor	0.015	0.02	0.0002	0.0004	0.003	0.0002
Volume resistivity, Ω cm	10^{11}	10^{14}	10^{17}	10^{17}	10^{16}	5×10^{17}
Thermal properties						
mp, °C			110	130	235	165
Application temperature range [d], °C	− 15 to + 150	− 15 to + 120	− 60 to + 95	− 50 to + 100	−20 to +130	− 20 to + 120
Optical properties n_D^{20}	1.6	1.4 – 1.5	1.50	1.50	1.46	1.50
Dimensional stability						
Coefficient of thermal expansion, K^{-1}	6×10^{-5}	6×10^{-5}				
Shrinkage at 100 °C (30 min)						
longitudinal, %	0.7 – 3.0	0.4	1.0 (80 °C)	0.5	1.6 (160 °C)	3
lateral, %	0.7 – 3.0	0.2	0	0	0	1
Physicochemical properties						
Water absorption in 24 h, wt %	45 – 115	5	< 0.1	< 0.1	< 0.1	< 0.1
Stability toward						
daylight	good	good	limited	limited	poor	limited
organic solvents	moderate	unstable	good	good	good	good
acid	unstable	unstable	good	good	very good	good
alkali	unstable	unstable	good	good	very good	good
salt solutions	unstable	stable	good	good	very good	good
Permeability						
Water vapor, g m^{-2} d^{-1}	very high	350	2.5	1.0	0.6	1.5
Oxygen [e], cm^3 m^{-2} d^{-1} bar^{-1}	10	1 500	4 000	1 600	38 000	600
Carbon dioxide [e], cm^3 m^{-2} d^{-1} bar^{-1}	100	10 000	16 000	7 000	110 000	1 800
Nitrogen [f], cm^3 m^{-2} d^{-1} bar^{-1}	12	300	1 300	400	9 400	140

[a] Standard values determined as described in Chapter 5.
[b] Cellulose acetate film (2 1/2-acetate) containing plasticizer.
[c] At 50 Hz, film thickness = 40 µm, plate/plate.
[d] Temperature range over which no significant loss in strength or stiffness occurs.
[e] Film thickness = 40 µm, 23 °C.
[f] Film thickness = 200 µm.
[g] Film thickness = 50 µm.
[h] Decomposition temperature.
[i] For biaxially stretched (balanced) film.
[j] 20 °C, 0 % relative humidity.
[k] Film thickness = 25 µm.
[l] 40 °C, 90 % relative humidity.
[m] Film thickness = 100 µm.

Table 3. Properties of selected films[a]

	Polystyrene (PS)	Poly(vinyl chloride) (PVC)		Poly(vinylidene chloride) (PVDC)	Polytetrafluoroethylene (PTFE)	Poly(vinyl fluoride) (PVF)
		rigid	plasticized			
Mechanical properties						
Thicknesses, µm	4 – 500	30 – 700	10 – 1 000	25	12 – 1 000	12 – 50
Density, g/cm^3	1.06	1.38	1.24 – 1.3	1.66 – 1.75	2.1 – 2.2	1.38 – 1.57
Ultimate strength						
longitudinal, N/mm^2	60	50	20 – 35	30 – 50	10 – 30	60 – 150
lateral, N/mm^2	60	50	20 – 35	30 – 50	10 – 30	60 – 150
Elongation at break						
longitudinal, %	10	70	250	150 – 250	100 – 350	100 – 200
lateral, %	10	40	250	150 – 250	100 – 350	100 – 200
Modulus of elasticity						
longitudinal, N/mm^2	3 000	2 500	300 – 1 000	230	400	1 400
lateral, N/mm^2	3 000	2 500	300 – 1 000	230	400	1 400
Electrical properties						
Dielectric strength [c], kV/mm	200	60 [f]	50 – 100 [f]		28 [m]	140
Dielectric constant at 1 kHz	2.5	3.5	3.3 – 5.3	4.5 – 6.0	2.0 – 2.1	8.5
Dissipation factor	0.0005	0.02	0.03 – 0.1	0.06 – 0.07	0.0002	1.6
Volume resistivity, Ω cm	10^{16}	10^{15}	$10^{11} - 10^{14}$	$10^{13} - 10^{14}$	10^{18}	10^{13}
Thermal properties						
mp, °C	240	160[h]		150	330	200
Application temperature range [d], °C	− 60 to + 100	− 15 to + 75	− 30 to + 50	0 to + 140	− 190 to +250	−70 to +110
Optical properties n_D^{20}	1.59	1.54		1.62		
Dimensional stability						
Shrinkage at 100 °C (30 min)						
longitudinal, %	25	3	0.5	10 – 35		1
lateral, %	25	3	0.5	10 – 35		1
Physicochemical properties						
Water absorption in 24 h, wt %	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.5
Stability toward						
daylight	limited	good	good	limited	very good	very good
organic solvents	limited	good	good	limited	very good	very good
acid	good	good	good	good	very good	very good
alkali	good	good	good	good	very good	very good
salt solutions	good	good	good	good	very good	very good
Permeability						
Water vapor, g m^{-2} d^{-1}	30	7.5	10 – 100	1.5	thickness < 400 µm	32
Oxygen [e], cm^3 m^{-2} d^{-1} bar^{-1}	3 000	80	700 – 1 000	30	contains	30
Carbon dioxide [e], cm^3 m^{-2} d^{-1} bar^{-1}	17 000	160	3 000 – 5 000	200	micropores	110
Nitrogen [f], cm^3 m^{-2} d^{-1} bar^{-1}	500	12	150 – 1 000	5		25

[a] Standard values determined as described in Chapter 5.
[b] Cellulose acetate film (2 1/2-acetate) containing plasticizer.
[c] At 50 Hz, film thickness = 40 µm, plate/plate.
[d] Temperature range over which no significant loss in strength or stiffness occurs.
[e] Film thickness = 40 µm, 23 °C.
[f] Film thickness = 200 µm.
[g] Film thickness = 50 µm.
[h] Decomposition temperature.
[i] For biaxially stretched (balanced) film.
[j] 20 °C, 0 % relative humidity.
[k] Film thickness = 25 µm.
[l] 40 °C, 90 % relative humidity.
[m] Film thickness = 100 µm.

Table 4. Properties of selected films[a]

	Poly(ethylene terephthalate) (PETP)	Polycarbonate (PC), monoaxially stretched	Nylon 6 (PA), biaxially stretched	Poly(methyl methacrylate) (PMMA)	Ethylene–vinyl alcohol copolymer (EVOH), biaxially stretched
Mechanical properties					
Thicknesses, µm	1.0–350	2–500	12–25	50–500	12–25
Density, g/cm^3	1.4	1.21	1.16	1.14	1.19
Ultimate strength					
longitudinal, N/mm^2	200[i]	250	260	40	210
lateral, N/mm^2	200[i]	8	260	40	210
Elongation at break					
longitudinal, %	100[i]	30	100–120	75	50
lateral, %	100[i]	100	100–120	75	50
Modulus of elasticity					
longitudinal, N/mm^2	4 000[i]	2 100	1 400	1 900	3 600
lateral, N/mm^2	4 000[i]	800	1 400	1 900	3 600
Electrical properties					
Dielectric strength[c], kV/mm	320	300	50	90	
Dielectric constant at 1 kHz	3.3	2.8	3.4–3.7	4.4	
Dissipation factor	0.005	0.0012	0.016	0.046	
Volume resistivity, Ω cm	10^{18}	2×10^{17}	5×10^{17}	10^{16}	
Thermal properties					
mp, °C	260	260	200[h]	160	181
Application temperature range[d], °C	−269 to +180	−140 to +140	−200 to +180	−20 to +120	
Optical properties n_D^{20}	1.60	1.60			
Dimensional stability					
Coefficient of thermal expansion, K^{-1}	2×10^{-5}	7×10^{-5}			
Shrinkage at 100 °C (30 min)					
longitudinal, %	< 1	< 1	< 1	−11	−4 (140 °C)
lateral, %	< 1	< 1	< 1	0	−0.5
Physicochemical properties					
Water absorption in 24 h, wt %	0.5	0.35	5	0.5	5.9
Stability toward					
daylight	limited	limited	good	very good	
organic solvents	moderate	moderate	good	good (aliph.) poor (arom.)	good
acid	moderate	moderate	poor	very good	
alkali	stable only at low concentration	stable only at low concentration	good	very good	
salt solutions	good	good	good	very good	
Permeability					
Water vapor, g m^{-2} d^{-1}	8	35	100	75	40
Oxygen[e], cm^3 m^{-2} d^{-1} bar^{-1}	50	8 000	26		0.5[j]
Carbon dioxide[e], cm^3 m^{-2} d^{-1} bar^{-1}	250	25 000	100		12
Nitrogen[f], cm^3 m^{-2} d^{-1} bar^{-1}	12	250	9		0.08

[a] Standard values determined as described in Chapter 5. [b] Cellulose acetate film (2 1/2-acetate) containing plasticizer.
[c] At 50 Hz, film thickness = 40 µm, plate/plate. [d] Temperature range over which no significant loss in strength or stiffness occurs.
[e] Film thickness = 40 µm, 23 °C. [f] Film thickness = 200 µm.
[g] Film thickness = 50 µm. [h] Decomposition temperature.
[i] For biaxially stretched (balanced) film. [j] 20 °C, 0 % relative humidity.
[k] Film thickness = 25 µm. [l] 40 °C, 90 % relative humidity.
[m] Film thickness = 100 µm.

Table 5. Properties of selected films[a]

	Polysulfone (PSU)	Polyimide (PI)	Polyether-imide (PEI)	Polyether-sulfone (PES)	Polyether-ketone (PEK)	Poly-(phenylene-sulfide) (PPS)
Mechanical properties						
Thicknesses, µm	25 – 1 000	7.5 – 125	75 – 500	25 – 300	75 – 250	2.0 – 125
Density, g/cm^3	1.24	1.42	1.27	1.37	1.24 – 1.30	1.35
Ultimate strength						
longitudinal, N/mm^2	80	180	100	85	120	220
lateral, N/mm^2	80	180	100	85	120	220
Elongation at break						
longitudinal, %	50	70	70	15	200	55
lateral, %	50	70	70	15	200	55
Modulus of elasticity						
longitudinal, N/mm^2	2 500	2 700	3 000	2 400	330	4 000
lateral, N/mm^2	2 500	2 700	3 000	2 400	330	4 000
Electrical properties						
Dielectric strength[c], kV/mm	200	280[k]	145	125	170	240
Dielectric constant at 1 kHz	3.14	3.5	3.56	3.90	3.5	3.0
Dissipation factor	0.001	0.003	0.0013	0.003	0.002	0.0006
Volume resistivity, Ω cm	5×10^{16}	10^{18}	3.4×10^{16}	1.7×10^{16}	10^{17}	10^{17}
Thermal properties						
mp, °C	370	406[h]	216[h]	223[h]	334	285
Application temperature range[d], °C	− 80 to + 170	− 260 to + 400	≤ 170	≤ 180	≤ 240	≤ 180
Optical properties n_D^{20}	1.63	1.78				
Dimensional stability						
Coefficient of thermal expansion, K^{-1}	5.6×10^{-5}	2×10^{-5}	4.5×10^{-5}	4.4×10^{-5}	4×10^{-5}	3×10^{-5}
Shrinkage at 100 °C (30 min)						
longitudinal, %	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
lateral, %	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1	< 0.1
Physicochemical properties						
Water absorption in 24 h, wt %	0.4	1.3 – 2.9	1.0[l]	1.4[l]	0.4[j]	0.05
Stability toward						
daylight	good	good	good			good
organic solvents	moderate	moderate	satisfactory	satisfactory	very good	very good
acid	good	good	good	good	very good	very good
alkali	good	stable only at low concentration	good	good	very good	very good
salt solutions	good	good	good	good	very good	very good
Permeability						
Water vapor, g m^{-2} d^{-1}	98	84	43.5	100	28.5	5
Oxygen[e], cm^3 m^{-2} d^{-1} bar^{-1}		3.8		200	55	100
Carbon dioxide[e], cm^3 m^{-2} d^{-1} bar^{-1}		6.9				500
Nitrogen[f], cm^3 m^{-2} d^{-1} bar^{-1}		0.9				50

[a] Standard values determined as described in Chapter 5.
[b] Cellulose acetate film (2 1/2-acetate) containing plasticizer.
[c] At 50 Hz, film thickness = 40 µm, plate/plate.
[d] Temperature range over which no significant loss in strength or stiffness occurs.
[e] Film thickness = 40 µm, 23 °C.
[f] Film thickness = 200 µm.
[g] Film thickness = 50 µm.
[h] Decomposition temperature.
[i] For biaxially stretched (balanced) film.
[j] 20 °C, 0 % relative humidity.
[k] Film thickness = 25 µm.
[l] 40 °C, 90 % relative humidity.
[m] Film thickness = 100 µm.

Manufacturers. Wolff Walsrode, Federal Republic of Germany; Chemiefaser Lenzing, Austria; British Cellophane, United Kingdom; Flexel, Atlanta, United States; Futamura, Japan.

6.1.2. Cellulose Esters (→ Cellulose Esters)

Films derived from cellulose acetate [*9004-35-7*] (CA) (2-acetate, 2 1/2-acetate, and 3-acetate), cellulose acetate propionate [*9004-39-1*] (CAP), and cellulose acetate butyrate [*9004-36-8*] (CAB) are available; they all contain plasticizers–phthalate esters, aliphatic alcohols, triphenyl phosphate. Solution casting is the usual production method. Methylene chloride [*75-09-2*] is used as a solvent [9]. The casting solution has a cellulose acetate concentration of 18 – 26 % and a viscosity of 15 – 30 Pa · s.

Cellulose acetate films are the most important of the cellulose ester films. The properties of cellulose acetate vary as a function of the number of acetate groups per cellulose molecule. The properties of cellulose acetate film (2 1/2 acetate) are listed in Table 2. Films with the highest acetate concentration, corresponding to the minimum number of hydroxyl groups, show reduced sensitivity to moisture as well as minimal absorption of water and low water-vapor permeability. These films are crystal clear in their untreated state; they are dimensionally stable at ambient temperature and humidity and possess good electrical properties. Special surface characteristics (e.g., gloss, roughness) can be conferred during manufacture by casting on an appropriate type of metal belt.

Cellulose acetate films are used as supports for photographic coatings, as transparent windows, release films, decorative materials, and as electrical insulators.

Trade Names. Acetatfolie (Hoechst, Federal Republic of Germany); Triafol (Bayer, Federal Republic of Germany); Ultraphan (Lonza, Federal Republic of Germany); Bexoid (Bexford, United Kingdom); Cellittor (British Cellophane, United Kingdom).

6.2. Hydrocarbon Polymers

6.2.1. Low-Density Polyethylene (LDPE) (→ Polyolefins)

Low-density polyethylene [*9002-88-4*] (LDPE) has a molecular mass of > 50 000 and is prepared by the high-pressure process. It is converted into film mainly by the film blowing process at a melt temperature of 150 – 210 °C. For properties of the film, see Table 2.

Film thickness of LDPE ranges between 25 and 200 µm. The films are transparent to crystal clear, depending on the choice of relative molecular mass and process technology. They are soft, flexible, and resistant to creases and show high tear and tear propagation resistance. Mechanical properties are stable at low temperature and chemical resistance is very high. Films made from LDPE display low permeability to water vapor but high permeability to gases. They are nontoxic and contain neither plasticizers nor odorous materials. Low-density polyethylene films are easily welded and thermoformed. Specific shrinking characteristics can be imparted to the films by stretching at low temperature.

Films made from LDPE are used for packaging and as protective coverings. They are also employed in the production of composite films. Low-density polyethylene has become commercially important. Worldwide annual production is estimated at more than 8×10^6 t.

Trade Names. Plastotrans (Folienfabrik Forchheim, Federal Republic of Germany); Waloplast (Wolff Walsrode, Federal Republic of Germany); Helioflex (Feldmühle, Federal Republic of Germany); Bexthene (Bexford, United Kingdom); Cryovac L (Cryovac Div., W. R. Grace, United States).

Film derived from a copolymer of ethylene with vinyl acetate (EVM film) is used similarly as a packaging material. Its exceptional adhesive properties make it particularly suitable as a cling and stretch-wrap film.

Trade Name. Montothene (American Can, United States).

6.2.2. High-Density Polyethylene (HDPE) (→ Polyolefins)

High-density polyethylene [*9002-88-4*] (HDPE) prepared by the low-pressure method is also gaining importance as a raw material for films. Films made from HDPE are produced both by chill roll extrusion and by film blowing. Characteristics of the films are given in Table 2.

Films derived from HDPE have a higher chemical resistance than LDPE films and can be employed over a somewhat wider temperature range; HDPE films are used in construction and civil engineering as insulators. High-density polyethylene film can be welded and bonded, and is readily thermoformable.

Trade Names. Plastin (Folienfabrik Forchheim, Federal Republic of Germany); Fortiflex (Crown Zellerbach, United States).

Polyolefins (mainly HDPE) are the basis of several types of "synthetic paper," which are also prepared by film-forming techniques. The market for this material in Great Britain in 1985 was estimated to be 40 000 t/a.

6.2.3. Linear Low-Density Polyethylene (LLDPE) (→ Polyolefins)

The 1980s witnessed the development of films derived from linear low-density polyethylene (LLDPE), which are produced in the same way as LDPE films. Linear low-density polyethylene films have better mechanical properties (greater strength and higher ultimate strength) than ordinary LDPE films. As a result, LLDPE films have found applications as sacks for transporting heavy materials. Since LLDPE also has a higher melting point (118 °C) than LDPE (110 °C), these sacks can be filled with hot materials (e.g., freshly produced cement). Stretch-wrap packaging represents another expanding market. Like other polyethylene films, LLDPE films can be combined with additional materials to form composite films.

6.2.4. Polypropylene (PP) (→ Polyolefins)

Polypropylene [*9003-07-0*] (PP) film is available in an unoriented form as well as in both mono- and biaxially oriented versions.

Unstretched PP film is prepared either by chill roll casting or by blowing. Its properties are comparable to those of LDPE film, but with the advantages of greater transparency and a wider application temperature range. Unstretched PP film is used exclusively for packaging purposes. West European consumption in 1986 was 80 000 t.

Monoaxially stretched PP film exhibits increased strength in the direction of orientation and is used for three major applications: (1) cable insulation, (2) tapes and industrial fabrics, and (3) yarns and ropes.

Biaxially oriented polypropylene (*BOPP*) *film* is the most important type of polypropylene film; its properties are listed in Table 2. Worldwide consumption in 1986 was 550 000 t of which 460 000 t could be attributed to the packaging industry and 60 000 t to the technical sector. An additional 30 000 t was processed into adhesive tapes. Films made of BOPP currently occupy ca. 80 % of the packaging market once held by cellophane. Factors contributing to this development include superior properties, lower cost, and an ecologically preferable manufacturing technology. Likewise, glassine (a thick semitransparent greaseproof paper) is being replaced by opaque BOPP films that are heat- and cold-sealable. These films were introduced in the late 1970s by Hoechst [28].

Biaxial orientation can be achieved by any of the standard methods (see Section 2.3.2), but sequential stretching is preferred because this method permits a wide range of thicknesses. Most BOPP packaging films are provided with a separate sealing layer, formed either by coextrusion or in an off-line coating process. Coextruded sealing coatings normally consist of polyolefin copolymers, whereas polyacrylate or poly(vinylidene chloride) is used for off-line coating. Poly(vinylidene chloride) has the added advantage that it simultaneously improves the oxygen barrier properties of the film.

Biaxially oriented polypropylene films range in thickness from 4 to 80 µm. They are used industrially in laminates and as release and capacitor films.

Trade Names. Trespaphan (Hoechst, Federal Republic of Germany); Propafilm (ICI, United Kingdom); Shorco (British Cellophane, United Kingdom); Walothen (Wolff Walsrode, Federal

Republic of Germany); Bicor (Mobil, Belgium); Moplefan (Montedison, Italy).

6.2.5. Poly(4-methyl-1-pentene) (PMP) [29] (→ Polyolefins)

Poly(4-methyl-1-pentene) [*9016-80-2*] (PMP) films are produced by extrusion and possess the lowest density of all available films (0.83 g/cm^3). They have a higher softening point (*mp* 240 °C) than PP films (*mp* 165 °C). Film thickness ranges from 50 to 300 µm. For properties, see Table 2.

PMP films have been used only in combination with other materials, mainly with cardboard for the packaging of oven-ready meals.

6.2.6. Ionomers (→ Specialty Plastics)

The term *ionomer* describes a family of polymers incorporating both covalent and ionic bonds. Ionomers [*25608-26-8*] are based on polyethylene and have carboxyl groups located along their polymer chains. These groups provide ionic cross-links between chains [20]. The corresponding films are prepared by extrusion.

Many of the properties of ionomer films are comparable to those of LDPE films. However, ionomer films are more resistant at room temperature to fats and oils than LDPE films, and they possess exceptionally good sealing properties. They are also extremely tough and are ideally suited for packaging.

Trade Name. Surlyn A (Du Pont, United States).

6.2.7. Polystyrene (PS) (→ Polystyrene and Styrene Copolymers)

Polystyrene [*9003-53-6*] (PS) films are available in unoriented as well as in mono- and biaxially stretched form. They are produced by chill roll extrusion or by blowing. For properties, see Table 3. For foamed PS films see →Foamed Plastics, Chap. 4.2.2.

Oriented films are crystal clear with a high gloss but are extremely brittle. Their most important application is for envelope windows. The polymer must be modified with synthetic rubber to increase its strength. However, the resulting increase in impact strength is obtained at the expense of reduced transparency. Film thicknesses are between 4 and 500 µm.

The favorable dielectric characteristics of PS film result in its use in high-voltage capacitors and as an insulator for cables.

Trade Names. Styroflex (Norddeutsche Seekabelwerke, Federal Republic of Germany); Trycite (Dow Chemical, United States).

6.3. Halogenated Hydrocarbon Polymers

6.3.1. Poly(Vinyl Chloride) (PVC) [6] [→ Poly(Vinyl Chloride)]

Films formed from poly(vinyl chloride) [*9002-86-2*] (PVC) may be either rigid or flexible. Rigid films consist solely of PVC; flexible films are produced by addition of plasticizers or by copolymerization with acrylate esters or vinyl acetate. Both rigid and plasticized PVC films are produced by calendering processes, although chill roll casting and film blowing are also possible. Some PVC films are also subjected to stretching.

Rigid and plasticized PVC films are of great economic importance. Worldwide production is estimated at 1.4×10^6 t/a.

6.3.1.1. Rigid PVC

Rigid PVC film is available in a wide range of thicknesses (30 – 700 µm); important properties are listed in Table 3. The material varies from crystal clear to transparent and may be colored by the addition of pigments. Rigid PVC is unaffected by water and is virtually impermeable to odors. It possesses outstanding chemical resistance and is also highly flame resistant. The film can be welded, heat-sealed, embossed, thermoformed, stamped, bonded, imprinted, and (under high vacuum) metallized. These remarkable properties combine to produce a broad spectrum of applications for rigid PVC film both in packaging and the industrial sector. Adhesive tapes account for a particularly large share of the market.

6.3.1.2. Plasticized PVC

Plasticized PVC films have an even wider range of applications than rigid PVC because the properties of plasticized PVC can be modified by varying the type and amount of plasticizer added. Plasticized PVC films are available with a thickness of 10 – 1000 µm; for important properties, see Table 3. The material can be strained and shows little tendency to wrinkle; its consistency varies from soft to hard. Plasticized PVC film is resistant to both tear and tear propagation, and it can be embossed, welded, bonded, and stitched. It is resistant to most inorganic and organic chemicals, as well as to water and light. It also shows good aging and weathering properties. Permeability to water vapor and gases is minimal.

Plasticized PVC film is used mainly in packaging and the building industry, as a protective covering, in laminated films, and for the modification of wood surfaces. It is also processed into commercial adhesive tapes.

Trade Names. Genotherm, Guttagena, Genopak, Suprotherm (Hoechst, Federal Republic of Germany); Mipolam (Dynamit, Federal Republic of Germany); Folan, Polytherm (Folienfabrik Forchheim (Federal Republic of Germany); Pentafilm (Pentaplast, United States).

6.3.2. Poly(Vinylidene Chloride) (PVDC)

[27], [30] [→ Poly(Vinylidene Chloride)]

Poly(vinylidene chloride) [*9002-85-1*] (PVDC) films are made from a vinylidene chloride–vinyl chloride–vinyl acetate copolymer [*9003-20-7*]. The comonomers render the product sufficiently stable for extrusion. Plasticizers are also added to improve processing properties. The films are manufactured by the tubular drawing process. Since the stability of the melt is low, the film tube must be cooled as it emerges from the die prior to orientation by blowing (see Section 2.3.2). For properties, see Table 3.

Poly(vinylidene chloride) film has been on the market under the trade name Saran since the late 1940s and is still widely used because of its outstanding water vapor and oxygen barrier properties. Its principal application is in the packaging of perishable foodstuffs.

The material is supplied as a shrinkable monofilm or coextruded film; PVDC laminates are also available. Coextruded products are prepared in both tubular and flat form. The PVDC generally is sandwiched between other polymeric layers which impart strength and toughness. Adhesion problems are overcome by simultaneously extruding an adhesive tie layer.

Trade Names. Saran, Saranex (Dow Chemical, United States); Cryovac (Grace, United States); Krehalon (Kureha, Japan); Supralon (Asahi, Japan).

6.3.3. Fluoropolymers [26]

(→ Fluoropolymers , Organic)

Incorporation of fluorine into hydrocarbon chains results in polymers that display antiadhesive characteristics, thermal stability, flame resistance, insulating properties, and, most important of all, extreme stability with respect to weathering and aggressive media. As a result of these properties, which depend on the degree of substitution, these films are used mainly under extreme conditions.

From a chemical standpoint, the fluoropolymers used for film production fall into three categories:

1) perfluorinated polymers,
2) partially fluorinated polymers,
3) polymers containing chlorine as well as fluorine

Polytetrafluoroethylene (PTFE) is a perfluorinated polymer that cannot be processed from a melt by conventional techniques. Polytetrafluoroethylene films are prepared by isostatically sintering the powdered material, followed by "skiving" (paring) of the sintered block. All other fluoropolymers–including copolymers of PTFE with perfluorinated vinyl components–are extrudable and can, therefore, be made into films either by the chill roll process or by film blowing.

The market for fluoropolymer films in the United States is dominated by films made from poly(vinyl fluoride) (PVF), PTFE and polychlorotrifluoroethylene (PCTFE). Extrudable copolymer films made from tetrafluoroethylene–hexafluoropropylene copolymer (FEP),

perfluoroalkoxy polymer (PFA), and ethylene–tetrafluoroethylene copolymer (ETFE) have also become somewhat important, or at least have a considerable potential for development. Film made of poly(vinylidene fluoride) (PVDF) is also worth mentioning because of its piezoelectric properties.

6.3.3.1. Poly(Vinyl Fluoride) (PVF)

Poly(vinyl fluoride) [*24981-14-4*] (PVF) film is commercially available in both transparent and pigmented form in thicknesses of 12 – 50 µm. For properties, see Table 3. Its antiadhesive characteristics, together with its resistance to mechanical, chemical, and weathering effects, make PVF film suitable for use as a protective lamination on both interior and exterior paneling. It also serves as a release film in the pressing of synthetic resin.
Trade Name. Tedlar (Du Pont, United States).

6.3.3.2. Poly(Vinylidene Fluoride) (PVDF) [31]

Poly(vinylidene fluoride) [*24937-79-9*] (PVDF) films are solvent cast or extruded with a thickness of 50 –500 µm. They possess excellent resistance to weathering and aggressive media.

Poly(vinylidene fluoride) film has an exceptionally high dielectric constant and a relatively high dissipation factor. Because of its piezoelectric properties, the monoaxially stretched film is often employed as a mechanical–electrical transducer in membranes. The consumption of film for piezoelectric applications is small (ca. 1 t); its price is extremely high.
Trade Name. Fluorex (Rexham, United States); Kynar (Pennwald, United States).

6.3.3.3. Polytetrafluoroethylene (PTFE)

Polytetrafluoroethylene [*9002-84-0*] (PTFE) films have the highest resistance to aggressive agents of any polymer film and are thermally stable up to 250 °C. Since PTFE films are produced via sintering and skiving, they are somewhat porous; some properties are listed in Table 3. Most of the PTFE film that is produced is used for sealing pipe joints and flanges.

Two related copolymers are extrudable: FEP [*25036-53-7*], derived from tetrafluoroethylene and hexafluoropropylene, and PFA [*2645-79-6*], a copolymer of tetrafluoroethylene and a perfluorovinyl ether. Films produced from these materials are not porous, but otherwise display essentially the same characteristics as PTFE itself. The mechanical and optical properties of the extruded films are superior to those of skived PTFE films. Commercial products are available in thicknesses from 12 –1000 µm. Their principal uses are as release films, in the electrical industry, and especially as container linings.
Trade Name. Hostaflon (Hoechst, Federal Republic of Germany); Teflon (Du Pont, United States); Fluon (ICI, United Kingdom).

6.3.3.4. Polychlorotrifluoroethylene (PCTFE)

Films made from polychlorotrifluoroethylene [*9002-83-9*] (PCTFE) are used widely in the packaging of pharmaceuticals and medical supplies; this is because the films possess excellent moisture barrier properties and chemical resistance, coupled with good processability, sterilizability, and UV stability. These films are available in thicknesses ranging from 12 to 250 µm.
Trade Name. Aclar (Allied, United States).

6.4. Poly(Methyl Methacrylate) (PMMA) [32] (→ Polyacrylates)

Poly(methyl methacrylate) [*9011-14-7*] (PMMA) films are prepared either by chill roll extrusion or by film blowing; thicknesses range from 50 to 500 µm. They show high resistance to heat and weathering, but their ultimate strength and elongation at break are relatively low. For properties, see Table 4.

Poly(methyl methacrylate) film is frequently used as a protective layer on metal and plastic, thereby competing with PVF film.
Trade Name. Korad (Polymer Extruded Products, United States); Plexiglas-Folie (Röhm, Federal Republic of Germany); Shinkolite Film (Mitsubishi Rayon, Japan).

6.5. Poly(Vinyl Alcohol) (PVA) [33]

(→ Polyvinyl Compounds, Other)

Poly(vinyl alcohol) [*9002-89-5*] (PVA) is produced by hydrolyzing poly(vinyl acetate). Both the fully and partially hydrolyzed polymers are utilized for film manufacture; the fully hydrolyzed polymer is water-soluble.

Poly(vinyl alcohol) film can be prepared by casting an aqueous PVA solution or by extrusion. Prior to extrusion, the poly(vinyl alcohol) is compounded into a paste with water and a plasticizer. Film thicknesses are in the range of 20 – 75 µm.

The characteristic features of PVA film include its high transparency and gloss, toughness, absence of static charge, and, in the absence of moisture, desirable oxygen barrier properties. Water solubility varies, depending on the vinyl acetate content.

Poly(vinyl alcohol) is employed as a monofilm, but can also be coated with PVDC or laminated with polyolefins. Biaxially oriented films can be produced by simultaneous stretching [34].

Poly(vinyl alcohol) films are used as packaging materials for food, textiles, and as release films in the manufacture of synthetic resin products. They are also employed in the manufacture of water-soluble sachets for disposable packaging of hygienic products and as a basis for polarized films.

Trade Names. Vinylon (Kuraray, Japan); Emblar-OV (Unitika, Japan); Bovlon (Nippon Gohsei, Japan).

6.6. Ethylene–Vinyl Alcohol Copolymer (EVOH)

A partially crystalline copolymer of ethylene and vinyl alcohol, EVOH [*25067-34-9*] is prepared by hydrolysis of an ethylene–vinyl acetate copolymer. This material first appeared on the market in the 1970s [35]; it can be processed into a film by extrusion.

Films derived from EVOH are excellent gas barriers. Biaxial orientation leads to increased strength and further reduction of gas permeability. Since the gas barrier properties of EVOH films are adversely affected by moisture, they are primarily used in the form of laminates or coextrudates. Some properties of oriented EVOH film are listed in Table 4.

The principal application of these materials is in food packaging.

Trade Names. Eval (Kuraray, Japan); Exceed (Okura, Japan).

6.7. Polyamide (PA) [20] (→ Polyamides)

Polyamide (PA) films (thickness 12 – 25 µm) are made mainly from nylon 6 [*25038-54-4*] and nylon 66 [*32131-17-2*] by chill roll extrusion or film blowing. The films display high tensile and tear strengths, as well as high resistance to puncture, abrasion, oils, and fats. Nylon films have a low permeability to gases, depending on the atmospheric humidity. Nylon 11 [*25035-04-5*] and nylon 12 [*24937-16-4*] are used for films with lower water-vapor permeability.

Orientation improves the properties of nylon films. Biaxially oriented films are produced from nylon 6 by simultaneous stretching. Properties of these films are listed in Table 4.

Polyamide films are used as monofilms but are employed mainly in composite structures produced by lamination, extrusion coating, or coextrusion with sealing or barrier resins.

Trade Names. Capran (Allied, United States); Dartek (Du Pont, United States); Bonyl (Kohjin, Japan); Emblem (Unitika, Japan); Supronyl (Hoechst, Federal Republic of Germany).

6.8. Polyester [10] (→ Polyesters)

The term *polyester film* generally refers to films prepared from poly(ethylene terephthalate) [*25038-59-9*] (PETP). However, other polyesters are also processed into film, e.g., poly(butylene terephthalate) [*24968-12-5*] (PBT) and poly(ethylene-2,6-naphthalene dicarboxylate) [*24968-11-4*] (PEN) (trade name: Q-Film, Teijin/Japan) [36].

Polyester films are manufactured by extrusion; the majority are biaxially oriented and produced in flat form; only a small amount of tubular film is made.

Biaxially oriented PETP films are prepared in thicknesses ranging from 1.0 to 350 µm and vary in appearance from clear to transparent, depending on their end use. Pigments can also be added for coloration.

Polyester films have favorable mechanical and electrical properties, are resistant to solvents, and display long-term thermal stability up to 130 °C (VDE 0530). They can be welded and bonded by special techniques. Furthermore, these films can be printed by flexographic, gravure, and screen printing techniques; they are also suitable for high-vacuum metallization without the need for special pretreatment. Some polyester films can be subjected to further deforming subsequent to their manufacture, for example, by thermoforming or thermally induced shrinking. Properties of PETP film are listed in Table 4; its permeability to gases and water vapor can be significantly decreased by coating with PVDC.

Polyester films are used for more applications than any other type of film. They are employed as packaging materials, dielectrics, and electrical insulators, as base films for photographic and magnetic coatings, as release films in synthetic resin moldings, and as color and montage film in the graphics industry. These films also serve as a base for flexible printed circuits, as drafting films, and as decorative film. Polyester films have gained considerable importance in the manufacture of videotapes. The demand for polyester film as a base for thermal transfer tapes is also increasing.

Polyester films can be combined with other materials (PE film, paper webs, and aluminum foil) to produce excellent composites.

The economic importance of polyester film is formidable: 550 000 t was sold worldwide in 1986, most of this being used in the manufacture of electrical insulation, magnetic tape, photographic film, and packaging film.

Trade Names. Hostaphan (Hoechst, Federal Republic of Germany); Mylar (Du Pont, United States); Melinex (ICI, United Kingdom); Lumirror (Toray, Japan); Tetoron (Teijin, Japan); Diafoil (Diafoil, Japan); Terphane (Rhône Poulenc, France); Espet (Toyobo, Japan).

6.9. Polycarbonate (PC) [20]

(→ Polycarbonates)

Polycarbonate [*25037-45-0*] (PC) film first appeared on the market in 1957 and is prepared from bisphenol A and phosgene. The principal manufacturers of both polycarbonate and polycarbonate film are General Electric and Bayer.

Polycarbonate films are produced either by extrusion (chill roll) or by solution casting and are available with a thickness of 2 – 500 µm. The thinner grades are subjected to monoaxial stretching. The properties of monoaxially stretched PC film are listed in Table 4.

The special characteristics of these films include their sterilizability, dimensional stability, thermoformability, and long-term thermal stability up to 130 °C (VDE 0530). They are transparent to crystal clear in appearance, and possess good electrical properties. Applications have been found in packaging, the electrical industry, and above all in the field of graphics. Consumption of PC film in the United States is ca. 3000 t/a.

Increased usage is anticipated, partly as a result of new developments in the raw material sector.

Trade Names. Lexan (General Electric, United States); Makrofol (Bayer, Federal Republic of Germany); Pokalon (Lonza, Federal Republic of Germany).

6.10. Polysulfone (PSU) (→ Specialty Plastics)

The first commercially available polysulfone (PSU) was Udel, which was introduced by Union Carbide in 1965 (see also Section 6.12.2). Udel polysulfone [*25135-51-7*] is produced by reacting bisphenol A and 4,4′-dichlorodiphenyl sulfone. The material is amorphous, with a glass transition temperature of 190 °C, and can be converted into film either by extrusion or by solution casting.

Polysulfone films are available in thicknesses ranging from 25 to 1000 µm. They are highly transparent, possess high heat and dimensional stabilities, and can be thermoformed. Their electrical properties are comparable to those of biaxially oriented polyester films. Polysulfone films are suitable for long-term use at temperatures up

to 150 °C, making them superior in this respect to polyester films. Polysulfone films are also stable to alkali, acid, oils, and detergents, although they are attacked by ketones and chlorinated or aromatic hydrocarbons. The properties of Udel PSU film are listed in Table 5.

Principal applications are in packaging, the electrical industry, adhesive tape manufacture, and display technology [37].

Trade Names. Sumilite FS 1200 (Sumitomo Bakelite, Japan); Europlex-Folie (Röhm, Federal Republic of Germany).

6.11. Polyimide (PI) (→ Polyimides)

The most important polyimide [*25036-53-7*] (PI) for film production is derived from pyromellitic acid and 4,4′-diaminodiphenyl ether [38]. Films made from this material do not melt, and they have the highest thermal stability of any synthetic film. They are suitable for long-term applications at temperatures up to 240 °C and short exposure to peak temperatures of up to 400 °C. For properties, see Table 5.

Polyimide films are prepared by dissolving the appropriate monomers in a suitable solvent (e.g., *N,N*-dimethylacetamide), in which they react to form a soluble intermediate (polyamic acid). The film is cast from this solution and is subsequently converted into the insoluble polyimide by intramolecular condensation at 300 °C.

The combination of high thermal stability and flame resistance, along with stability to radiation and favorable electrical properties, makes this film attractive for applications in the aerospace industry (e.g., as an electrical insulator), and the manufacture of electronic components [39]. Commercial grades range in thickness from 7.5 to 125 µm.

Worldwide capacity in 1986 was 1400 t/a with an expected annual growth rate of 10 %.

Trade Names. Kapton (Du Pont, United States; Toray/Du Pont, Japan); Upilex (Ube Industries, Japan); Novax (Mitsubishi Chemical, Japan).

6.12. Other Films

6.12.1. Polyetherimide (PEI)

The backbone of polyetherimides (PEI) contains both ether and amide linkages. The PEI used for film production was first marketed by General Electric in 1982 under the trade name Ultem and is an amorphous thermoplastic polymer with the structure:

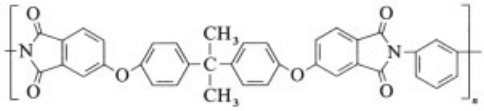

[*61128-24-3*]

Poly(etherimide) film is prepared by the chill roll process at an extrusion temperature of 350 °C and is available with a thickness of 75 – 500 µm.

This film is especially suited for electrical applications due to its favorable mechanical and electrical properties. The maximum temperature for long-term use is 170 °C; thus, thermal stability is substantially less than that of conventional polyimides. Nonetheless, stability is adequate for many purposes and significantly higher than that of polyester film [37]. The film is used mainly for electrical insulation. For properties, see Table 5.

Trade Names. Ultem Films (Westlake, Dixon, United States); Sumilite FS-1400 (Sumitomo Bakelite, Japan); Europlex PEI (Röhm, Federal Republic of Germany).

6.12.2. Polyethersulfone (PES) (→ Specialty Plastics)

The poly(ethersulfone) (PES) used in film manufacture is an amorphous, thermoplastic material manufactured by ICI under the trade name Victrex. It consists of aromatic rings linked alternately by ether and sulfone groups:

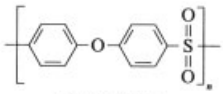

[*25667-42-9*]

Poly(ethersulfone) film is normally manufactured by the chill roll process, although it can also be cast from solution. Thicknesses vary from 25 to 300 µm. Its range of applications is determined by its electrical and optical properties combined with its thermostability and flame resistant behavior [37], [40], [41].

Poly(ethersulfone) films are used in electrical insulation, interior aircraft fittings, and liquid crystal displays.
Trade Names. Stabar (ICI, United Kingdom); Victrex PES-5200 G (Westlake, United States); Talpa 1000 (Mitsui Toatsu, Japan); Sumilite FS 1300 (Sumitomo Bakelite, Japan); Europlex-Folie (Röhm, Federal Republic of Germany).

6.12.3. Poly(ether ketone) (PEK)

Poly(ether ketones) [*2738–27–4*] can be extruded into either fibers or films in the temperature range 350–420 °C.

Films made from PEK display high thermal stability (up to 240 °C), solvent and chemical resistance, good dimensional stability, flame resistance, and stability to both radiation and hydrolysis. Properties are listed in Table 5. Since PEK film also possesses favorable electrical properties and is a low-density material, it is used in the aerospace industry.

Both crystalline and amorphous unoriented films are commercially available in thicknesses from 75 to 250 µm [37], [41], [42]. Oriented films are still being developed.
Trade Names. Stabar K and XK (ICI, United Kingdom); Sumilite FS 1100 (Sumitomo Bakelite, Japan); Talpa 2000 (Mitsui Toatsu, Japan).

6.12.4. Poly(Phenylene Sulfide) (PPS)
(→ Specialty Plastics)

Poly(phenylene sulfide) [*9016-75-5*] (PPS) was marketed as early as 1973 by the Phillips Petroleum Company under the trade name Ryton. However, it was not until 1985 that Phillips, in collaboration with Toray, succeeded in producing the material in film quality.

Poly(phenylene sulfide) is a partially crystalline polymer prepared from sodium sulfide and 1,4-dichlorobenzene. Poly(phenylene sulfide) film is manufactured by extrusion at 300 °C and is similar to PETP film in that it can be biaxially oriented and heat-set. Properties of PPS film are given in Table 5. Thicknesses are in the range 2 – 125 µm.

Poly(phenylene sulfide) film combines favorable electrical characteristics with high thermal stability [43]. Thus, it is mainly used in the electrical industry, although it also has potential as a magnetic storage medium.
Trade Names. Torelina (Toray, Japan).

6.13. Composite Films [44]

The most important production methods for composite films are adhesive lamination, extrusion coating (see Section 2.4.3), and, more recently, the simultaneous extrusion (coextrusion) of several raw materials (see Section 2.4.2). The latter approach is unique in that it can produce very thin multilayer films (ca. 12 µm) that still display combined properties.

The incorporation of PETP, BOPP, PA, and PVC into composite films permits optimization of ultimate strength, elongation at break, and puncture resistance.

Composite films are especially important as packaging materials, particularly for bulk materials regardless of their state, i.e., gases, liquids, and solids.

Composite films that contain polyethylene can be welded by either heat impulse or thermal sealing methods. Thin polyethylene layers (ca. 30 µm) may suffice for packaging lightweight goods, provided that no need exists for evacuation or a controlled atmosphere. Moderately thick layers (≥ ca. 40 µm) are required for vacuum and inert atmosphere packings. Thicker PE film (≥ ca. 60 µm) is employed if the material to be packaged is a liquid or a paste.

The combination of plastic films with aluminum foil produces semirigid composites with exceptionally low permeability to gases, water vapor, and odors. Special heat-sealable, polyolefin-based layers result in composite films that can be separated by peeling. Composite films incorporating ionomers are especially suitable for packaging oily materials.

7. Summary of Uses [2], [10], [26], [27]

The most important uses of individual films have been described in Chapter 6 and are summarized in Table 6. They can be divided into three main categories: packaging, electrical insulation, and use as bases for other active materials.

Table 6. Uses of films

Application	Film material*																							
	Cellophane	CA	PE	BOPP	PMP	Ionomer	PS	PVC	PVDC	PVF	PTFE	PCTFE	PMMA	PVA	EVOH	PA	PETP	PC	PSU	PI	PEI	PES	PEK	PPS
Packaging																								
Food (incl. snack and convenience foods)	x		x	x	x	x		x	x					x	x	x	x	x	x					
Nonfood items	x	x	x	x		x	x	x				x		x		x	x	x					x	
Electrical Insulation																								
Capacitors				x			x										x	x	x					x
Insulation		x		x	x		x				x						x		x	x	x	x	x	x
Printed circuits																	x	x		x	x	x	x	x
Base Film for Functional Coatings																								
Adhesive tapes	x	x		x				x									x		x		x			
Magnetic tapes																	x							
Floppy disks																	x							x
Photographic film		x															x							
Reprography		x															x	x		x				
Decoration	x	x		x				x									x							
Transfer metallization				x													x							
Graphics		x															x	x						
Miscellaneous																								
Office supplies		x	x	x			x	x									x							
Heat mirrors																	x							
Protective lamination			x	x	x			x	x	x	x		x				x				x	x		
Solar collectors																	x			x				
Release film	x	x		x	x					x	x			x			x							
Protective covers			x	x				x																
Display items														x			x		x					

* For explanation of abbreviations, see abbreviations list.

Packaging. Packaging is the most important application of films; virtually every type of film finds some application in the packaging field. Plain and printed films are commonly used in packaging a wide range of foods and nonfood articles. Legal requirements for food packaging are described in [25] (see also → Foods, 4. Food Packaging, Chap. 8.4.).

Tubular film prepared either directly by blowing or by the subsequent sealing of flat film is used for the manufacture of a wide variety of bags, sacks, and blister or cushion packs. Thicker films are thermoformed into bottles, containers, cups, trays, and bowls. Shrinkable film is employed in palette packaging. Film is also the basis of skin packings, packing tapes, and even tear-off bands. Decorative iridescent packaging films are available which consist of layers of two or more different polymers with different refractive indices. The iridescent color is produced by light interference [45]. Special characteristics such as impermeability to gases and water vapor can be obtained by applying coatings or using laminated structures.

Electrical Insulation. Films with a high dielectric constant and a low dielectric loss factor are particularly useful in a.c. and d.c. capacitors. Such films may be as thin as 1.0 μm. The films themselves are dielectrics, but they can be converted to electrodes either by vacuum metallizing or by combining them with aluminum foil. Polyester and PI films, either alone or in combination with paper or synthetic nonwoven materials serve as insulating material in electrical equipment. Polymer films are also employed in cable sheathing; thin strips of film are often used as electrical insulation or labels; they also create plasticizer barriers, provide thermal and mechanical insulation, and, in combination with metallic film, confer protection against static charging. Polyester (PETP), PC, and PI films are used in the production of flexible printed circuits because of their exceptional dimensional and thermal stability. Film switches represent a new area of application, especially for PETP film.

Base Films for Functional Coatings. Polyester film in particular has acquired major significance as a base material for magnetic coatings. Special polyester films combining extra strength with defined surface roughness have been developed for this purpose. Magnetic tapes based on these films are marketed internationally for audio cassette players, video recorders, computers, and other commercial and industrial recording equipment (→ Information Storage Materials).

Films with a high dimensional stability (e.g. PETP) are used as substrates for drafting and photocopying. Films are also used in the printing industry as carriers for transferring color to paper or plastics (stamping foil). Film coated with deformable materials are used as embossing films. Iridescent films and films coated with firmly bonded layers of metal and/or lacquers are frequently used for decorative or display purposes. Films are also the basis for most adhesive tapes. The introduction of a wide range of film-based typewriter ribbons are another application.

Miscellaneous. The furniture industry employs films as veneers for the surface treatment of sheet materials. Thicker films (sheets) are thermoformed into ceiling and wall panels, as well as decorative fixtures and reproductions.

The use of films in heat mirrors and solar collectors (→ Solar Technology) is a more recent development. Heat mirrors employ coated films (mostly PETP) that transmit visible light and reflect infrared light. In solar collectors, solar energy is absorbed and passed to a flowing heat-transfer medium (e.g., water). Solar collectors may be constructed from UV-stable film.

Film is frequently used as a protective covering for equipment and materials. Many modern greenhouses are constructed almost entirely from transparent film material. The building industry generally makes frequent use of films on account of their waterproofing properties. Small amounts of films are used for rainwear and clothing.

8. Environmental Aspects

Most of the scrap rolls and waste generated in film production as a result of edge trimming and manufacturing problems is collected and recycled. Special facilities are required for shredding the film and compressing it into granules suitable for reprocessing.

Film constitutes 1 – 2 % of normal household waste. Separate collection and recycling of waste film is in theory possible, but is expensive and requires considerable organization. All films can be regarded as environmentally harmless, provided that they are disposed of properly. Combustion of film waste presents no particular technological problems. Even PVC film can be safely incinerated, although public controversy suggests otherwise; modern combustion facilities are capable of trapping any hydrogen chloride vapor that is formed.

Films can create litter problems; a number of film additives and other chemical modifications have been proposed for accelerating the rate at which films undergo photooxidative decomposition. However, apart from one exception these suggestions have not been implemented, primarily because the required precision in formulation for a given application is difficult to achieve. In addition they conflict with the present ideas of litter prevention. The sole exception is the *mulch film* employed in agriculture for stimulating seed germination and growth; this product is designed to have a lifetime of only one growing season.

9. References

1. J. H. Stickelmeyer: "History of Plastic Films," in W. R. R. Park (ed.): *Plastics Film Technology,* Van Nostrand Reinhold Company, New York 1969, pp. 3 – 9.
2. O. J. Sweeting: *The Science and Technology of Polymer Films,* **vol. I,** II, Interscience, New York 1968, 1970.
3. J. Voss: "Aus Lösungen hergestellte Hydratcellulose-Kunststoffe" in R. Vieweg, E. Becker (eds.): *Kunststoffhandbuch,* vol. III, Hanser Verlag, München 1965, pp. 84 – 165.
4. E. Karmas: *Sausage Casing Technology,* Noyes Data Corp., New Jersey 1974, pp. 146 – 177.
5. L. I. Nass: *Encyclopedia of PVC,* vol. 3, Marcel Dekker, New York – Basel 1977.
6. G. W. Becker, D. Braun (eds.): *Kunststoffhandbuch, Polyvinylchlorid 2/2,* 2nd ed., Hanser Verlag, München – Wien 1985; M. Kreutzer, "Extrudieren von Folien und Platten," pp. 957 – 977; M. Kreutzer, C. F.-J. Holthausen, F. Pilz, "Kalandrieren von Folien," pp. 1001 – 1029.
7. R. A. Elden: "Calandering of Film and Sheeting" in A. Whelan, J. L. Craft (eds.): *Developments in PVC Production and Processing – 1,* Applied Science Publ., London 1977, pp. 205 – 221.
8. F. Hensen, W. Knappe, H. Potente (eds.): *Handbuch der Kunststoff-Extrusionstechnik,* "Extrusionsanlagen," vol. II, Hanser Verlag, München – Wien 1986; G. Winkler: "Extrusion von Blasfolien," pp. 76 – 106; R. Hessenbruch: "Coextrusion von Blasfolien," pp. 107 – 124; H. Bongaerts: "Extrusion von Flachfolien nach dem Kühlwalzenverfahren," pp. 125 – 186; F. Hensen: "Extrusion von gereckten Folien," pp. 243 – 270; E. Schöllkopf, K. Roesch: "Beschichten und Kaschieren durch Extrusion," pp. 320 – 353.
9. S. Janocha: "Herstellung von Polyethylenterephthalat Folien," in *Winnacker-Küchler,* **7**, 520 – 523.
10. E. Werner, S. Janocha, M. J. Hopper, K. J. Mackenzie: "Polyesters, Films" in *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 12, J. Wiley & Sons, New York 1988, pp. 193 – 216.
11. R. J. Ashley, D. Briggs, K. S. Ford, R. S. A. Kelley: "Adhesion Problems in the Packaging Industry" in D. M. Brewis and D. Briggs (eds.): *Industrial Adhesion Problems*, Orbital Press, Oxford 1985, pp. 213 – 218.
12. H. L. Weiss: *Coating and Laminating Machines,* Converting Technology Company, Milwaukee 1983.
13. ICI, GB 1 264 338, 1969 (G. A. E. Pears).
14. S. Levy: *Plastics Extrusion Technology Handbook,* Industrial Press Inc., New York 1981.
15. R. H. Cramm, W. R. Sibbach: *Coextrusion Coating and Film Fabrication,* Tappi Press 1983.
16. R. Gächter, H. Müller: *Plastics Additives Handbook,* Hanser Verlag, Munich–Vienna–New York 1985.
17. J. Siepek, H. Daoust: *Additives for Plastics,* Springer Verlag, New York–Heidelberg–Berlin 1983.
18. B. Carlowitz: *übersicht über die Prüfung von Kunststoffen,* Kunststoff-Verlag, München 1983.
19. R. Hummel: *Atlas der Polymer- und Kunststoffanalyse,* Verlag Chemie, Weinheim 1978 – 1983.
20. J. H. Briston, L. L. Katan: *Plastics Films,* 2nd ed. Longman Scientific & Technical, Harlow, United Kingdom, 1983.

21. ASTM Standards,Parts 26, 27, 1967.
22. DIN-Taschenbuch 18, 21, Beuth-Verlag, Berlin 1986.
23. J. V. Schmitz, W. E. Brown: *Testing of Polymers,* **vol. I – IV,** Interscience, New York–London–Sydney 1965 – 1969.
24. Beloit, US 3 425 267, 1969 (J. D. Pfeiffer).
25. Office of Federal Register, *Code of Federal Regulations, 21 CFR,* National Archives and Records Administration, Washington 1986.
26. C. R. Oswin: *Plastic Films and Packaging,* Applied Science Publ., London 1975.
27. C. J. Benning: *Plastic Films for Packaging,* Technomic Publishing Co., New York 1983.
28. Hoechst, EP 004 633, 1979 (D. Gebhardt, G. Crass, S. Janocha).
29. Data Sheet 1982 A, Mitsui Petrochemical Ind., Tokyo.
30. R. A. Wessling: *Polyvinylidene Chloride,* Gordon and Breach Science Publishers, New York–London–Paris 1977, pp. 176 – 180.
31. Data Sheet, Kynar Film, Westlake Plastics Company.
32. Data Sheet, Korad A, Georgia Pacific Corp. Data Sheet, Shinkolite-Film, Mitsubishi Rayon, Tokyo.
33. C. A. Finch: *Polyvinyl Alcohol,* J. Wiley & Sons, London–New York–Sydney–Toronto 1973, pp. 378 –387.
34. Data Sheet, Emblar-OV, Biaxially oriented PVA-Film, Unitika Ltd., Osaka. Data Sheet, Bovlon, PVA, Biaxially Oriented Film, Nippon Gohsei, Osaka.
35. H. Iwasaki, K. Sato, K. Akao, K. Watanabe: "Development and Commercial Production of EVAL-Resin," *CEER Chem. Econ. & Eng. Rev.* **9** (1977) no. 10, 32 – 37. Data Sheet, EVAL EF, Biaxial Oriented EVAL-Film, Kuraray, Osaka.
36. Teijin, DE-AS 2 337 815, 1973 (S. Shimotsuma, T. Asai, M. Hosoi, K. Sagamihara et al.).
37. Data Sheet, *Highly Functional Engineering Plastics Films,* Sumilite FS-1000 Series, Sumitomo Bakelite Co., Tokyo.
38. C. E. Sroog, *Macromol. Rev.* **11** (1976) 161 – 208.
39. Data Sheet, E 36 484-2, Kapton, Du Pont, Genf, Switzerland 1983.
40. Data Sheet SR TD 10, Stabar S, Polyethersulfone Film, ICI, Welwyn Garden City, England 1984.
41. Data Sheet, *High Performance Film Talpa-1000/2000,* Mitsui Toatsu Chemicals, Tokyo.
42. Data Sheets SR TD 11, SR TD 12, *Stabar K/XK, Polyetherketone Films,* ICI, Welwyn Garden City, England 1985. A. S. Wood, "Unmatched Performance, a Build-up in Ketone Resins," *Modern Plastics,* April 1987, pp. 46 – 48.
43. Toray, DE 2 916 841, 1979 (T. Asakura, Y. Noguchi, O. Hiroaki). "PPS Film Torelina," Technical Information, Toray Industries, Tokyo 1987.
44. Data Sheets Hoechst, Werk Kalle, Wiesbaden 1980 – 1987.
45. "Mearl Iridescent Films," Technical Information Sheet TIB-G, The Mearl Corporation, 1050 Lower South Street, Peekskill, New York 10566, April 1980.

Filters, Optical → Optical Materials

Fire Extinguishing Agents

Orville M. Slye, Jr., Loss Control Associates, Levittown, Pennsylvania 19056, United States

1. **Introduction** 1
2. **Water as an Extinguishing Agent** . . . 1
3. **Fire-Fighting Foams** 3
3.1. **Limitations and Uses** 4
3.2. **Foam Concentrates** 4
3.3. **Guidelines for Using Fire-Fighting Foams** 6
4. **Dry Chemical Extinguishing Agents** . 6
5. **Inerting Gas Extinguishing Agents** . . 7
6. **Halogenated Extinguishing Agents** . . 8
7. **Extinguishing Agents for Combustible Metal Fires** 9
8. **References** 10

1. Introduction

Fire extinguishing agents include water, water plus additives, "dry chemicals," inerting gases, and halogenated compounds, each specifically suited for certain types of fires. Water is the oldest means of extinguishing fire, and chemicals are used as additives to enhance its characteristics for fire control. Dry chemical agents are mixtures of powders applied from portable devices or from fixed systems. Inerting systems using nitrogen, carbon dioxide, or flue gas have long been used in fire control, especially in marine applications. Use of halogenated agents, with superior flame extinguishing properties, was increasing worldwide in the 1980s.

The type of extinguishing agent depends on the class of the combustible material. Various systems are used to classify combustibles and to determine the most effective type of extinguishing agent [1].

Table 1 illustrates three major systems for fire classification. Some of these systems categorize fire control agents according to the class or classes of fire for which they are most effective and applicable. Although separation into classes of fire is often difficult in actual fire situations, the classification systems allow an accurate description of a fire situation and identify some of the hazards associated with the fire-fighting activity.

2. Water as an Extinguishing Agent

Water is the single most useful and efficient material for controlling and halting combustion and fire. Because water is generally plentiful, its use is relatively inexpensive and effective. The efficiency of water for extinguishing fire is primarily due to the change in physical state from liquid to vapor. The heat of vaporization, which is 2260 kJ/kg, removes heat from the combustion process. As an example of the heat removal capability of water, a typical fire-fighting nozzle applying water at 380 L/min (100 US gal/min) directed into a fire removes nearly 1000 MJ/min. In comparison with the complete combustion of 1 L of gasoline, which produces >36 MJ/kg, the extinguishing efficiency of water can be fully appreciated [2].

Additional advantages of water as an extinguishing agent include the following:

1) Has the ability to dissolve and wash away many products of combustion so that the seat of the fire may be reached
2) Can be easily pumped through hoses and pipes because of the small change in viscosity with temperature in the liquid phase between 1 and 99 °C
3) Expands to steam, which has ca. 1700 times the volume of the liquid; makes confined spaces inert by displacing air and oxygen, aiding in fire extinguishment
4) Has high surface tension (0.073 N/m), which enables water to be issued in a consolidated stream or in discrete water droplets from spraying devices or "fog" nozzles
5) Has a density high enough to provide a physical mass to a nozzle stream, which assists in physical penetration

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_113

Table 1. Classification of fire-extinguishing agents

Classification	United States	United Kingdom	Federal Republic of Germany
Class A fire	normal combustibles: wood, cloth, paper, rubber, and plastics	solid materials (organic matter): fire occurs with formation of glowing embers	*
Class B fire	flammable and combustible liquids, gases, and greases	liquids or liquefiable solids	*
Class C fire	energized electrical equipment that may be harmed by the extinguishing media	gases or liquefied gases from a liquid spill or a liquid or gas leak	*
Class D fire	combustible metals, such as magnesium, titanium, zirconium, sodium, or potassium	combustible metals	*
Electrical fire	see Class C	no separate classification; de-energize and use appropriate agent for class of fire remaining	see Class E
Class E	see Class C	see Electrical fire	extinguishing agent chosen for fire of surrounding material: foam should not be used; water should be used only in the form of a fog and at a safe distance (>5 m), depending on the voltage

* Identical to United Kingdom classification.

6) Has molecular stability; does not break down or dissociate appreciably at <1650 °C, which exceeds ordinary flame temperature

Although water is the most useful and efficient heat-removing agent for fire-fighting purposes, some physical properties restrict use under certain conditions, including the following:

1) When frozen, water cannot be transported or applied easily to a fire.
2) High surface tension restricts ability to penetrate and wet below the surface area of combustibles to extinguish deeply sourced fires.
3) Water readily runs off surfaces, leaving a coating too thin for heat removal or surface protection.
4) Transportation through hoses or pipes results in loss of pressure because of friction.
5) Density and immiscibility reduce effective control of hydrocarbon fires; water sinks below the fuel even when used in a fine spray.
6) Water vapor explosion may occur if spontaneous vaporization is possible, e.g., in the case of molten salts, bitumen, or other products with a temperature >120 °C.
7) Use of water to fight fires of lower molecular mass metals may be dangerous.

Chemicals that change water's characteristics by depressing freezing point, reducing surface tension, increasing viscosity, and modifying flow properties or water density have been developed and extend the capability of water as an extinguishing agent.

Inorganic freezing point depressants prevent water from crystallizing at low temperature. Calcium chloride [*10043-52-4*], used in 94 % pelleted grade, depresses the freezing point of water to as low as −40 °C; sodium chromate [*10034-82-9*] is added (0.5 wt % of the calcium chloride used) for corrosion protection. Other antifreeze materials, such as ethylene glycol [*107-21-1*], propylene glycol [*57-55-6*], and glycerol [*56-81-5*], are also used for cold-weather protection, primarily in automatic sprinkler systems. However, care must be taken to prevent contamination of public water supplies. The use of these materials in portable extinguishers is not recommended because residual antifreeze is flammable after the fire evaporates the water.

Wetting agents lower the surface tension of water from a normal level of 0.073 N/m to ca. 0.025 N/m, which allows water to penetrate solid materials and extinguish fires below the surface. Wetting agent solutions are most useful when used on tightly consolidated Class A fire hazard materials, such as straw, mattresses, or cotton bales, where the water, with a lower surface tension, reaches the deep-seated fire by capillary action more quickly and with better efficiency than water without a wetting agent.

Viscosity-enhancing materials or "thickeners" increase the viscosity of water so that it may be applied to nonhorizontal surfaces. Two principal types of water thickeners are used: those that produce an organic gel and those that give a slurry. Gels either coat the material that is burning or prevent the material from igniting. Bentonite [*1302-78-9*], a powdered mineral, yields a water slurry, as does ammonium phosphate [*7783-28-0*] or ammonium sulfate [*7783-20-2*]. Ammonium sulfate gives a thick slurry, which is used particularly in combating forest fires. These slurry-type mixtures retard flames because of their solids content and the capability of ammonium salts to halt flaming and glowing ignition.

Flow property modifiers reduce friction resulting from water flowing at high velocity in a hose. Water pumped at high velocity through the small-diameter hose that is commonly used for fire fighting loses pressure because of friction between the water and the hose walls and because of turbulent flow within the hose. Dilute solutions of poly(ethylene oxide) [*25322-68-3*] can yield nonturbulent water flow in a fire hose. Flow increases of 55 – 70 % can be achieved in a 3.8-cm fire hose when 0.017 vol % poly(ethylene oxide) is added. Nozzle pressure at the end of the hose is doubled when the flow modifier is used. An added benefit from the water stream modifer is the ability to use small and more manageable hose sizes without loss of nozzle pressure or flow volume.

Water density modification for fire-fighting purposes is achieved by adding air to water to form a semistable air foam that is lighter than most flammable and combustible liquids or to emulsify the top layer of a burning liquid to obtain a nonflammable floating emulsion of water and fuel. As do wetting agents, synthetic detergents lower the water's surface tension. Detergents are chosen for their emulsification action. When they are sprayed or directed into flammable or combustible fuels, detergents mix rapidly to produce a suspension of the fuel in the detergent solution. Emulsification lowers the fuel's vapor pressure, which reduces the vapor concentration to below the lower flammable limit in air; thus, the fuel no longer burns. Synthetic detergent emulsifying agents are used at a concentration of 0.5 – 2.0 vol % yielding a fuel – solution – air emulsion with a lower density than that of either water or fuel. Although the emulsion may be temporary, fires involving fuels with a high flash point require only a short interruption of combustion to halt reignition.

3. Fire-Fighting Foams

Fire-fighting foam is a mass of gas-filled bubbles formed from aqueous solutions of specially formulated, concentrated liquid foaming agents. The gas normally is air, but may be an inert gas.

Because the density of the resulting foam is so low, it floats on the surface of flammable and combustible liquids, producing an air-excluding, cooling, continuous layer of vapor-sealing, water-bearing materials that halts or prevents combustion [3]. Various methods and types of equipment are available to generate and apply foam.

Formulation. Fire-fighting foams are formulated in several ways for fire extinguishing action. Foams are produced by mixing an appropriate amount of foam concentrate with water and then aerating and agitating the solution to form the bubble structure.

Some foams are thick and viscous, forming a tough heat-resistant blanket over a burning liquid or on a vertical surface; others are thinner and spread more rapidly. Foams that produce a vapor-sealing film of surface-active water solution on a liquid surface are available. Some foams are used as large volumes of wet gas cells for filling cavities and inundating surfaces.

Foams are defined by their *expansion ratio*, which is the ratio of final foam volume to original solution volume before the air is added. Foams are arbitrarily subdivided into three ranges:

1) low-expansion foam: expansion up to 20 : 1
2) medium-expansion foam: expansion range from 20 : 1 to 100 : 1
3) high-expansion foam: expansion range from 200 : 1 to 1000 : 1

3.1. Limitations and Uses

Low-expansion foam develops a cooling, coherent blanket and is used primarily to extinguish burning flammable or combustible liquid fires, such as in spills or in storage tanks. Foam extinguishes fires progressively by first providing

a cover on the surface of the liquid, such as in a storage tank, to prevent vapor transmission and then by cooling hot surfaces adjacent to the burning liquid to prevent reignition. Foams are also able to diminish or halt the generation of flammable vapors from liquids or solids. Medium expansion foam is used primarily to control spill fires and to fill cavities where toxic or flammable vapors may collect.

Use of medium- or high-expansion foam that has air in the bubbles may be dangerous in a small room. If the vapor concentration in the room is higher than the explosive limit, the air from the foam may be enough to cause an explosion. Such accidents have occurred.

When attacked by heat and flame, foam breaks down and the water content vaporizes. Foam must be applied to a burning liquid surface in sufficient volume and rate to compensate for evaporation, with additional foam required to establish a residual layer over any exposed fuel. Foam is an unstable "air – water emulsion" that is easily dissipated by mechanical or physical forces, by some chemical vapors or fluids, and by incompatible extinguishing agents. Turbulent air or a violent uplift of combustion gases can also divert light foam from the burning area.

Foam solutions are conductive and are not recommended for use on electrical fires. If foam is used near electrical equipment, precautions should be taken to minimize chances of electrical shock.

3.2. Foam Concentrates

Some foam concentrates are available for specific applications, whereas others are more suitable for extinguishing fires in all types of flammable liquids, including water-soluble and foam-destructive liquids.

Protein Foaming Agents. Concentrates of these foaming agents consist of protein fragments derived from chemical digestion and hydrolysis of natural protein solids. The material gives elasticity, mechanical strength, and water retention capability to the resulting foams.

The concentrates also contain dissolved polyvalent metallic salts, which strengthen the bubbles when the foam is exposed to heat and flame. Organic solvents are added to the concentrates to improve foam formation and uniformity, as well as to control viscosity at lower temperature.

Protein-type concentrates are available for use with either fresh water or saltwater at 3 or 6 vol % concentrations. The resulting foam is dense, viscous, and highly stabile; its resistance to high heat and burnback is superior to that of other foaming agents. These foams cling to vertical surfaces and insulate them if the proportioning rate is increased or the water content reduced.

Fluoroprotein Foaming Agents. These agents contain an additional fluorinated, surface-active compound that provides a fuel-shedding property to the resulting foam. These foams are particularly effective for fire-fighting situations in which the foam becomes coated with fuel, such as when subsurface or base-injected foam is used for fires in storage tanks or for foam streams applied from hose nozzles where falling foam may plunge into the burning fuel pool. In addition, the resulting foam has a greater compatibility with dry chemical agents than that of protein foam. As with protein foam, the concentrate is used with either fresh water or saltwater at concentrations of 3 or 6 vol %.

Aqueous Film-Forming Foaming Agents (AFFF). The AFFF agents are synthetic compounds that form air foams similar to those produced by the protein-based materials. In addition, these foaming agents are capable of forming water solution films on the surface of flammable liquids from which the term aqueous film-forming foam is derived. The AFFF concentrates are used at either 3 or 6 vol % with fresh water or seawater.

The air foams generated from AFFF solutions possess low viscosity, have fast spreading and leveling characteristics, and, like other foams, act as surface barriers to exclude air and halt fuel vaporization. The unique feature of these foams is the continuous aqueous layer that forms under the foam. This floating film on the hydrocarbon fuel surface helps suppress combustible vapors and cools the fuel substrate. The film is capable of spreading over fuel surfaces not fully covered by foam; it is self-healing when mechanically disrupted and continues to spread as long as a reservoir of nearby foam remains. Hot surfaces can reduce foam effectiveness as can aromatic

hydrocarbons. The insulating effect in protecting flammable or combustible liquids against radiant heat is lower with these foams compared to protein foams.

Concentrates of AFFF agents contain fluorinated, long-chain synthetic hydrocarbons with particular surface-active properties. Various water-soluble, high-molecular mass, synthetic polymers are added to aid in strengthening the bubble wall and in retarding breakdown.

As used, these foams are nontoxic and biodegradable in diluted form. Because of the low surface tension of solutions draining from AFFF, the foams may be useful under mixed class fire situations (Class A and B), where deep penetration of water is needed to extinguish deeply sourced fires in Class A materials. The AFFF concentrates must not be mixed with other types of foam concentrates; however, the foam itself does not break down other types of foams in fire-fighting operations. The AFFF agents are compatible with dry chemical fire-fighting compounds.

Synthetic Hydrocarbon Surfactant Foaming Agents. These hydrocarbons are synthetically produced surface-active compounds that foam copiously in water solution; however, their use for fire fighting requires some care. These compounds are also used at concentrations of 3 or 6 vol % in water.

Fire-fighting characteristics depend on the volume of foam layer on the burning surface that restricts air and controls combustible vapor production. In addition, a minor cooling effect comes from the water contained in the foam, which is a function of the rapid drain time of the water from the foam mass. This drain time leaves a bubble mass that is highly vulnerable to heat or mechanical disruption. A rate of application higher than that used with ordinary fire-fighting foams is required to gain control and extinguishment. The foams from these agents are sometimes destructive to normal fire-fighting foams.

Alcohol-Type Foaming Agents. Alcohol-compatible agents are used on fires involving fuels that are water soluble, water miscible, or are polar solvents. They can also be applied to normal flammable liquids, but are most effective when used on fires in alcohols, enamel and lacquer thinners, methyl ethyl ketone, acetone, and acrylonitrile. Relatively small amounts of these substances rapidly deteriorate ordinary fire-fighting foams.

Proprietary compositions of several alcohol-type concentrates, based on concentrate bases of protein, fluoroprotein, or AFFF foams, have been developed. The most common are usually described as polymeric alcohol-resistant AFFF concentrates, which produce foams suitable for application to either flammable hydrocarbons or water-miscible liquids. These substances exhibit AFFF characteristics when used on hydrocarbons and produce a floating gel-like mass for foam buildup on water-miscible fuels.

High-Expansion Foaming Agents. High-expansion foaming agents are used to control and extinguish Class A and Class B fires and are particularly well suited for flooding confined spaces. Applications are generally indoor because out-door use is limited by the weather.

Liquid concentrates for producing high-expansion foam consist of a synthetic hydrocarbon surfactant that will foam copiously with a small input of turbulent action. Normally 1 – 2 vol % of concentrate in water is used. The air supply for foam generation should be outside the fire area because combustion products have an adverse effect on the resulting foam volume and stability. Tests have shown that expansion ratios of ca. 500 : 1 are the most effective for controlling fires and reducing vaporization from liquefied natural gas (LNG) spills. As water drains from the foam, a thin ice layer is formed. This ice layer floats on the LNG and supports the high-expansion foam blanket, retarding vapor generation and possibly building a shield to protect the LNG surface from radiated heat.

Chemical Foaming Agents and Powders. These materials have become obsolete because of the superior economics and ease of handling of liquid foam-forming concentrates. Chemical foam is formed by the aqueous reaction of aluminum sulfate [*10043-01-3*] and sodium hydrogen carbonate [*144-55-8*]; the foam also contains proteinaceous stabilizers. The carbon dioxide [*124-38-9*] gas that results is trapped in the foam.

3.3. Guidelines for Using Fire-Fighting Foams

Foams are effective for fire control if the following general rules are applied:

1) Foam application should be gentle because agitated application increases the time to extinguish the fire and the amount of foam concentrate used.
2) Successful use of foam depends on the rate of application. Rates are expressed in terms of the amount of foam solution (by volume) reaching the total area of fuel surface each minute. Thus, an expansion rate of 1 : 8 with an application rate of 4.1 $L\,min^{-1}\,m^{-2}$ provides 32.8 $L\,min^{-1}\,m^{-2}$ of finished foam. Increasing foam application rate over the minimum generally reduces time to extinguish the fire. If application rates fall below the minimum, extinguishment is improbable. Application rates are determined by vendors or approval laboratories.
3) Foams are more stable when generated with water at an ambient temperature of 2 – 27 °C; either fresh water or seawater can be used. Foam contaminants, such as detergents, oil residues, or certain corrosion inhibitors, may adversely affect foam quality.
4) The location of foam-making device air inlets is important because combustion products can adversely affect foam quality.
5) The pressure of water in foam-making devices affects foam quality; high or low limits should not be exceeded.
6) Some fire-fighting chemicals, such as vaporizing liquid extinguishing agents and dry chemicals, can adversely reduce the effectiveness of foam.

4. Dry Chemical Extinguishing Agents

Dry chemical agents, which are mixtures of powdered chemical compounds of varying composition, have the most efficient flame-halting capability. The compounds are stored in moisture-proof containers and are pneumatically blown into a fire by a pressurized gas discharge that conveys the solid powder into the combustion zone. Dry chemicals, known as dry powders in the United Kingdom, do not extinguish fire by cooling, smothering, or emitting an inert gas, but instead provide instantaneous and momentary flame extinction [1, p. 199], [4].

Extinguishers for applying dry chemicals onto fires evolved from hand-held tubes of sodium hydrogen carbonate to a variety of portable and fixed fire extinguishing systems using dry chemical pressurized-gas propellant. Particle size range of the mixture is 5 – 75 µm, with most in the range of 20 – 25 µm.

In 1959 the U.S. Navy developed the *potassium hydrogen carbonate* [*298-14-6*] formula for dry chemicals. This material has twice the flame-quenching capacity per unit mass as the sodium compound. The material is called Purple K because it exhibits a purple flame when it encounters a flame. Purple K is compatible with AFFF when applied in twined nozzle systems and provides rapid control of burning spilled flammable liquids, such as during an airplane crash.

Ammonium dihydrogen phosphate [*7722-76-1*]has an extinguishing action on flammable liquid fires similar to that of sodium hydrogen carbonate. However, ammonium dihydrogen phosphate also decomposes when heated to yield a glassy coating of metaphosphoric acid [*37267-86-0*], HPO_3, which may halt further combustion of Class A fuels such as wood. This extinguishing agent is useful for fires in either Class A or Class B fuels. Electrical fires, Class C, can also be extinguished by dry chemicals because the powder is not electrically conductive. Extinguishers containing this agent are "multipurpose" because they can be used on Class A, B, or C fires.

A potassium-based dry chemical consisting of *potassium chloride* [*7447-40-7*] or Super K has been developed for fire fighting. This agent exhibits a flame-extinction efficiency that is slightly less than that of potassium hydrogen carbonate.

Monnex, an extinguishing agent developed by ICI (United Kingdom), consists of the addition product of urea and potassium hydrogen carbonate and is unique in its mode of reaction. Urea [*57-13-6*] can form adducts, or inclusion complexes, under certain processing conditions. These complexes are crystalline mixtures in which the molecules of one of the components

are contained within the crystal lattice framework of the other compound; the original components separate when the mixture is heated. In Monnex, urea is contained in the potassium hydrogen carbonate lattice.

A finely powdered dry chemical is a more efficient flame-quenching agent than the same powder with a coarser particle-size distribution. Fine powders cannot be projected great distances, however, which makes them impractical for fire extinguishing purposes. The urea–potassium hydrogen carbonate particles are of normal and customary size distribution and can be projected efficiently through a nozzle under gas pressure. However, when heated, the crystals shatter, yielding fine particles and resulting in more efficient flame-quenching than is normally obtained.

5. Inerting Gas Extinguishing Agents

Inerting gases extinguish or prevent fire by partial exclusion of oxygen from contact with the combustible material. Inerting gases include gases that do not react with combustible materials and can dilute oxygen or air atmospheres to the point where oxidation (or combustion) either cannot start or cannot continue. Helium [*7440-59-7*], argon [*7440-37-7*], krypton [*7439-90-9*], nitrogen [*7727-37-9*], carbon dioxide [*124-38-9*], and sometimes water vapor are used to produce inert atmospheres. Flue gases or products of combustion gases, such as those from oil burners, are also inerting gases because of their high carbon dioxide and nitrogen content and low (ca. 5 %) oxygen content. Applications for flue gas inerting include marine petroleum tankers where flue gases are used for storage tank inerting. Carbon dioxide and nitrogen are the principal inerting gases currently used.

Carbon dioxide gas has long been used as an inerting agent for fire extinguishment and prevention [5]. Carbon dioxide is noncombustible and does not react with most substances. The gas provides its own pressure for discharge from storage containers used for fire-protection applications. As a gas or as a finely divided solid (dry ice), carbon dioxide does not conduct electricity and is ideal for use on energized electrical equipment.

Carbon dioxide is the only inerting gas of which large volumes can be liquefied and maintained in that state in a small container volume at ordinary temperature and at moderate pressure. Carbon dioxide vapor is more dense than air, so the gas sinks to the base of the flames, replacing air and oxygen and maintaining a flame-smothering atmosphere. Carbon dioxide extinguishes flammable and combustible liquid and gas fires at concentrations of ca. 30 – 38 vol % carbon dioxide; however, a higher concentration may be necessary to extinguish certain gases that have wide flammable limits. Specific requirements of the vendor, regulatory standards, or fire codes should be consulted for particular applications.

Reignition is possible once the carbon dioxide dissipates and air (oxygen) contacts the combustible material before the material has cooled below its ignition temperature or fire point. Deep seated fires may continue in Class A materials after flame control has been achieved with carbon dioxide application. The gas diffuses slowly through materials that are difficult to penetrate. Therefore, cooling of these materials is necessary and is normally accomplished by maintaining an atmosphere of ca. 38 % carbon dioxide for a period of time.

Two principal types of *storage units* are used to contain carbon dioxide for fire-protection applications. In the high-pressure method, the liquid and gas phase are at equilibrium in a steel container at ca. 6 MPa pressure. The familiar portable CO_2 extinguisher is an example of the high-pressure container. The low-pressure storage unit is maintained at ca. 2 MPa and kept at −18 °C by a refrigeration unit. This method allows large amounts of CO_2, mostly in the liquid state, to be stored in light, steel-walled containers. Larger systems generally incorporate fire detectors and automatic or remote manual activation devices. An alarm, sounding 15 – 20 s before gas release, is required for confined spaces to avoid asphyxiation of any personnel that may be present.

Nitrogen is used as an oxygen diluent to halt combustion and produce an inert atmosphere [1, p. 191], [6]. However, nitrogen is not easily condensed. The gas must be stored at high pressure

(ca. 14 MPa) in heavy cylinders, which limits its application as a fire extinguishing material or inerting gas. In addition, because it is slightly lighter than air, nitrogen is difficult to direct or maintain near the base of flames or seat of the fire, where it is needed to be the most effective. Applications in liquid form have been attempted, but the extensive quantities that are needed are not currently practical.

Water vapor can be used, in special cases, as an extinguishing agent or inerting gas. However, high-pressure steam can generate static electricity, which can ignite flammable vapors, unless special precautions are taken.

6. Halogenated Extinguishing Agents

Although elemental halogens are vigorous oxidizing agents when reacted with metallic and nonmetallic elements with one or more atoms of oxygen, the resulting ionic halogens chemically combine with and reduce many compounds. Reaction of elemental halogens with highly oxidizable hydrocarbons, such as methane [*74-82-8*], ethane [*74-84-0*], or propane [*74-98-6*], using special high-temperature methods in the absence of oxygen, yields a gaseous or liquid organic compound that is capable of decomposition and is an efficient fire-extinguishing agent. Early halogenated extinguishing agents included carbon tetrachloride [*56-23-5*] and methyl bromide [*74-83-9*].

Halons. In 1946 it became obvious that of all of the halogenated compounds, the lower aliphatic hydrocarbons substituted with bromine resulted in powerful extinguishing agents. However, the materials were toxic. The fluorine bond to carbon was known to be very stable, and if a compound containing some new ratio of fluorine : bromine : carbon could be synthesized, perhaps the result would be less toxic. This theory led to the development of the halogenated fire extinguishing agents known as Halons [1, p. 192].

Nomenclature for the Halons has been simplified. Integers are used to represent the various constituents in the material: the first digit represents the number of carbon atoms in the compound molecule; the number of fluorine atoms are indicated by the second digit; chlorine atoms are represented by the third digit; the fourth digit indicates the number of bromine atoms; and the fifth indicates the number of iodine atoms, if any. Hydrogen atoms are not numbered.

Some common halogenated extinguishing agents are listed in Table 2. Most countries allow use of Halon 1211 and 1301 for fire control; Halon 2402, which is highly toxic, has limited applications where humans are not present, e.g., for fuel tank or engine compartment inerting [7].

Halons are effective flame extinguishing agents because their density causes flames to sink and stay in the region of the base of the flames. Knowledge is limited about combustion kinetics and the influence of the actions and reactions of halogenated material. Evidence has shown that Halon agents remove the active chemical species involved in the chain reactions of flame propagation. In fire situations, Halons decompose at flame temperature and the resulting products have an affinity for the highly active free radicals ($\cdot$O, $\cdot$H, and $\cdot$OH) produced by burning fuels, subsequently removing them from further flame propagation. Halogenated agent efficiency occurs because of the following reactions:

1) Capture of active atomic hydrogen by halogen, removing hydrogen from further heat-producing reactions 2) Reaction of active free-radical hydroxyl groups from oxidation of fuels, followed by a halogen hydride reaction that results in water
2) Regeneration of active atomic halogen by the preceding reaction so that the processes of capture and removal of atomic and free-radical groups may be repeated

The flame-extinguishing performance of Halon 1301 is slightly superior to that of Halon 1211. In general, ca. 5 vol % of Halon 1301 extinguishes fires in most hydrocarbon fuels, compared with, e.g., 6 vol % of Halon 1211. Individual fuels should be tested or reference made to standard tables to predict the actual extinguishing concentration. An 8 – 9 vol % concentration is required to prevent flame propagation in combustible mixtures of ordinary fuels in air. Halon 1301 is ca. 2.5 times more effective than carbon dioxide.

Solid combustibles can sometimes be extinguished by using Halons; however, soaking time

Table 2. Halogenated extinguishing agents

Chemical name	CAS registry number	Trade name	Molecular formula
Bromomethane	[*74-83-9*]	Halon 1001	CH_3Br
Bromochlorodifluoromethane	[*74-97-5*]	Halon 1211	$CBrClF_2$
Bromotrifluoromethane	[*75-63-8*]	Halon 1301	CF_3Br
1,2-Dibromo-1,1,2,2-tetrafluoroethane	[*124-73-2*]	Halon 2402	$C_2F_4Br_2$

is required to allow the Halon to reach the seat of the fire in a mass of material. Soaking time varies with the solid combustible material composition and geometry.

Toxicology and Occupational Health. When Halons are used in occupied areas, toxicity, characterized by anesthetic or narcotic effects, must be considered. High Halon concentration can lead to dizziness, impaired coordination, and reduced mental acuity. At high concentration, prolonged exposure can result in unconsciousness or even death.

Halon 1301 is the least toxic and, for most applications, can be discharged into occupied spaces. Its TLV-TWA values and MAK are 1000 ppm and 6100 mg/m^3, respectively [8], [9]. System designs exceeding 7 vol % concentration have induced light-headedness and have reduced dexterity. Above 10 vol % toxic effects are considered potentially serious and these levels should be avoided. Halon 1211 has a lower threshold limit than Halon 1301 (TLVs and MAKs have not been established), and the effects at concentrations required for fire extinguishment are comparable to those of carbon dioxide.

In general, application from portable extinguishers in typical rooms is considered safe, but fixed systems should be restricted to unoccupied areas. Concentrations up to 4 vol % have been shown to be tolerable for a 1-min maximum exposure, which compares with the same exposure time for a 10 vol % concentration of Halon 1301.

7. Extinguishing Agents for Combustible Metal Fires

Class D combustible metal fires are an important consideration when materials that burn in air, such as elemental metals and their alloys, are being used. Many of these metals require extensive heat exposure before reaching ignition temperature, whereas others may ignite spontaneously.

Metals burn at extremely high temperature and are highly reactive. No extinguishing agent is known that will halt the oxidation reactions of burning metals. Dry sand, cement, and special powders are possible materials. Combustible metal fires can be extinguished by removing the metal's contact with air or oxygen, by smothering, by inerting, or by cooling the surrounding area to lower the temperature of the mass below its ignition point [10].

One proprietary mixture consists of graphite and foundry coke, carefully sized to facilitate easy handling, packing, and application to the burning metal surface by scoop or shovel. Graphite is difficult to ignite and is a good heat conductor; i.e., it absorbs heat and radiates it from the burning area. An organic phosphate added to the mixture decomposes to yield a noncombustible smoke that penetrates the spaces between graphite particles to exclude oxygen. The mixture is a reducing agent and is moderately effective for all metal fires.

Another proprietary mixture consists largely of carefully sized sodium chloride [*7647-14-5*] with tricalcium phosphate [*7758-87-4*] powder, and zinc stearate [*557-05-1*] added to facilitate discharge from a pressurized extinguisher through a delivery nozzle. A small amount of powdered thermoplastic material is added to the mixture to melt and hold the sodium chloride particles together for efficient cooling and smothering of the burning metal. This material is suitable for all metal fires. A powder specially formulated for sodium fires contains sodium carbonate [*497-19-8*] instead of sodium chloride [1, p. 205].

Inert gas blanketing of burning metals, e.g., with argon, helium, and neon [*7440-01-9*], is an effective extinguishing method where feasible.

Ternary eutectic chloride powder contains potassium chloride [*7447-40-7*], sodium chloride, and barium chloride [*10361-37-2*]. This mixture, when exposed to burning metals, melts and can cover the burning surface, thus excluding oxygen. Water, applied judiciously from a safe distance, can be used for burning magnesium or titanium. Water dissociates to hydrogen and oxygen at the burning metal surface, thus it may increase the burning rate. Water is not safe to use on burning sodium and other alkali metals.

8. References

1. R. L. Tuve: *Principals of Fire Protection Chemistry,* National Fire Protection Association, Boston 1976, p. 143.
2. R. M. Hodnut in A. E. Cote (ed.): *Fire Protection Handbook,* 16th ed., Chapter 1, Section 17, Chapter 6, Section 21, National Fire Protection Association, Boston 1986.
3. Reference 2, N. R. Lockwood, Chapter 4, Section 19, p. 32.
4. Reference 2, W. Haessler, Chapter 3, Section 19, p. 24.
5. Reference 2, H. V. Williamson, Chapter 1, Section 19, p. 2.
6. Reference 2, P. F. Johnson, Chapter 6, Section 19, p. 55.
7. Reference 2, D. W. Moore, Chapter 2, Section 19, p. 11.
8. ACGIH (ed.): *Threshold Limit Values (TLV) and Biological Exposure Indices,* ACGIH, Cincinnati, Ohio, 1986 – 1987.
9. DFG (ed.): *Maximum Concentrations at the Workplace (MAK)* , VCH Verlagsgesellschaft, Weinheim 1986.
10. Reference 2, A. S. Prokopovitsh, Chapter 5, Section 19, p. 49.

Fire Resistance → Flame Retardants
Fireworks → Pyrotechnics
Fischer – Tropsch Synthesis → Coal Liquefaction
Fish Oils → Fats and Fatty Oils

Flame Retardants

BURTON J. SUTKER, Ethyl Corporation, Sayreville, New Jersey 08872, United States

1. Introduction 1
2. Mechanism 1
2.1. Dilution 1
2.2. Generation of Noncombustible Gas . 2
2.3. Gas-Phase, Free-Radical Inhibition . 2
2.4. Solid-Phase Char Formation 3
3. Test Methods 5
4. System Development 7
4.1. Selection Criteria of Flame Retardants 7
4.2. Formulation with Reactive Flame Retardants 8
4.3. Formulation with Additive Flame Retardants 11
4.4. Interactions between Flame Retardants and Other Formulation Ingredients 14
4.5. Formulation to Reduce Smoke 15
5. Flame Retardant Markets 15
6. Toxicology 17
7. Future Directions in Flame Retardants 18
8. References 19

1. Introduction

*Flame retardation*is a process by which the normal degradation or combustion processes of polymers have been altered by the addition of certain chemicals. The polymers exhibit resistance to decomposition by thermal stress when measured by standard tests. The degree of resistance to flammability is usually defined by a numerical value of the test as being more resistant to flame or fire than the unmodified polymer. Results of these tests should not be used exclusively to predict performance in an actual fire, although they are a guide in selecting materials. *Flame-retarded materials*, can be formulated to be more resistant to ignition than non-flame-retarded materials, or to have slower rates of flame spread in a major fire that is initiated by some other source; however, the flame-retarded article will ultimately burn.

2. Mechanism

A basic understanding of the different theories by which flame retardants function is essential for identifying and selecting the optimum material for any given situation. Many alternative materials can be used, and the cost performance often dictates the choice. Flame retardants function by their interaction or interference with one of the three required components of a fire:

1) a combustible substance or fuel
2) heat, supplied either externally or from the combustion process itself
3) an oxidizing gas, primarily oxygen.

2.1. Dilution

Reducing the total quantity of combustible matter improves overall flame retardation. For example, adding fillers, such as clays, to polymer systems (often 50 – 200 parts by weight per 100 parts of polymer) reduces flammability. However, in some cases, such as glass fiber reinforced composites, the glass fiber stiffens the polymer. On exposure to heat or a flame, the glass may prevent the polymer from melting away from the flame; thus, the material burns. In addition, the glass can act as a heat sink so that less energy input is required to ignite the polymer on a second exposure to heat.

2.2. Generation of Noncombustible Gas [1]

Some materials decompose when heated, generating noncombustible gases. These gases lower the oxygen concentration at the front of the flame and often result in flame snuffing because of lack of oxygen. This result is transitory because once the material decomposes the residual matter acts

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_123

as an incombustible diluent of the polymer matrix. An example of this type of flame retardant is *calcium carbonate* [*471-34-1*] which decomposes at 825 °C to generate the solid, calcium oxide, and the gas, carbon dioxide; these products do not support combustion. The use of calcium carbonate with brominated flame retardants is not recommended, however. Calcium carbonate and its more alkaline decomposition product, calcium oxide, absorb hydrogen bromide that is liberated from the brominated flame retardants to form inactive species, such as calcium bromide.

Some materials decompose to produce water vapor as the noncombustible gas. Heat is absorbed because of decomposition as well as vaporization of liquid water. *Aluminum oxide trihydrate* [*21645-51-2*], ($Al_2O_3 \cdot 3H_2O$) begins to decompose at 230 °C with the release of 34.5 wt % of its original mass as water vapor. *Magnesium hydroxide* [*1309-42-8*], or hydrated magnesium oxide, begins to decompose at 340 °C with the release of 31 wt % of the original mass as water. For processing polymers at elevated temperature, the higher decomposition temperature of magnesium hydroxide makes it preferable to hydrated alumina [1]. Typically, 50 – 100 parts by weight of these compounds are required per 100 parts of polymer to achieve flame retardation.

Azo compounds are used in forming foam polymer systems. These materials liberate nitrogen. Although effective in producing foams, their cost effectiveness compared to that of other compositions has eliminated their use as flame retardants.

2.3. Gas-Phase, Free-Radical Inhibition

[2], [3]

Combustion of hydrocarbons creates highly active fragments in the solid or condensed phase. These fragments vaporize, react with oxygen, and form free radicals. Free-radical formation is highly exothermic, resulting in volatilization of additional active fragments from the condensed phase. The process continues unless free-radical formation is interrupted and stable species are produced. Some of the reactions that may occur are as follows [2]:

$HO\cdot + CO \longrightarrow CO_2 + H\cdot$	Highly exothermic
$H\cdot + O_2 \longrightarrow HO\cdot + O\cdot$	Chain branching
$O\cdot + HBr \longrightarrow HO\cdot + Br\cdot$	Chain transfer
$HO\cdot + HBr \longrightarrow H_2O + Br\cdot$	Chain termination

The HBr from decomposing *brominated compounds* is very effective in deactivating free radicals in the vapor phase. The formation of new radicals is also reduced as less heat is generated, and the entire combustion process is slowed.

Chlorinated compounds function in the same manner. In practice, often twice as much chlorine-containing compound is required as bromine-containing compound. The difference approaches the atomic mass ratio for the two halogens, i.e., 79.90 : 35.45 = 2.25.

Compounds containing *iodine* are known, but are of limited use as flame retardants because of their poor thermal stability. The dark color of many organoiodine compounds is also a negative factor.

Compounds containing *fluorine* generally exist as functional polymers instead of as materials added to other polymeric systems to provide flame retardation. These polymers are very stable and decompose only at high temperature. Hydrofluoric acid, when finally liberated, is an effective free-radical deactivator.

Antimony oxide acts as a synergist with halogens, particularly chlorine and bromine [3]. It is almost totally ineffective if used without halogen. The trioxide [*1309-64-4*] is the common material used, although the pentoxide is also available. The pentoxide [*1314-60-9*], as manufactured, generally has a much finer particle size and is more efficient per unit mass added than the trioxide. Semitranslucent thermosetting polyester laminates containing the pentoxide are also possible. Because some thermoplastic polyesters are sensitive to residual activity in all forms of antimony oxide, alkaline salts of antimony oxide are used in these critical cases.

Synergism may occur through a series of reactions; the basic reaction in the case of brominated flame retardants is:

$$Sb_2O_3 + 6\,HBr \rightleftarrows 2\,SbBr_3 + 3\,H_2O$$

Antimony tribromide forms a dense white smoke that snuffs the flame by excluding oxygen from the front of the flame. Water from normal combustion causes the equilibrium to shift

to the left, forming antimony oxide, which is then available to react with fresh HBr from decomposing brominated compounds.

Some applications cannot take advantage of the synergy between halogen and antimony oxide. For example, epoxy – glass fiber printed circuit boards contain high levels of bromine instead of combinations of bromine and antimony oxide. Trace quantities of antimony oxide poison chemical treatments used to etch the copper-foil printed circuits.

The effectiveness of halogen in inhibiting radical formation is often a function of the organic structure to which it is attached. In polypropylene, a cycloaliphatic bromine compound is more efficient at low levels than a more stable aromatic bromine compound. This is because of the relative decomposition characteristics of the bromine compound and the polymer. However, with rubber-modified, high-impact polystyrene, the more thermally stable aromatic compound is more efficient.

Typically, halogenated compounds used in applications of flame retardants contain either 40 – 70 wt % chlorine or 45 – 80 wt % bromine. From 20 to 40 parts of chlorinated compound or 5 – 30 parts of brominated compound would be used per 100 parts of polymer, depending on the flame retardant requirements. Antimony oxide use is typically 25 – 50 % that of the halogenated material.

Poly(vinyl chloride) has inherently good flame-retardant characteristics because of the high chlorine content. Antimony oxide (3 – 5 wt %) is often helpful in passing the more stringent flammability tests.

2.4. Solid-Phase Char Formation

Several different types of flame retardants form insulating or minimally combustible chars on polymer surfaces exposed to external heat sources. This char reduces volatilization of active fragments and absorbs and dissipates heat. The effectiveness of the flame retardant is specific for each polymer. For example, phosphorus-based flame retardants are effective in producing a minimally combustible char in phenylene oxide – ether polymers, but are essentially ineffective in styrenic polymers.

Phosphorus and Boron Compounds. *Phosphorus* is available in the elemental state, as organic phosphates, or as inorganic acid phosphates. All forms are used as flame retardants in different polymer systems, e.g., elemental red phosphorus in polyamides, organophosphates in polyphenylene oxides or polyurethanes, and inorganic acid phosphates in cellulosics. In the presence of heat and oxygen, phosphorus compounds decompose to form water vapor and phosphorus oxides. The phosphorus oxides subsequently react with the polymer matrix and dehydrate it, reforming phosphoric acids. These acids again decompose to reform water vapor and phosphorus oxides.

Ultimately as the water available from normal combustion of hydrocarbons diminishes, the phosphorus oxides react with hydrocarbon fragments to produce a very high melting point char at the interface between the polymer and the heat source. The chars, which contain phosphorus, rapidly dissipate heat energy and lose their "glow". This antiglow property of phosphorus contributes to its effectiveness as a flame retardant.

Boron compounds can function in a manner similar to phosphorus compounds. However, elemental boron is not used commercially as a flame retardant. Organoboron compounds are less stable and typically much more expensive than organophosphorus compounds. Inorganic boron compounds are used as extenders for antimony oxide or in low-cost treatments for cellulosics.

Typically, 1 – 10 parts of contained phosphorus or boron is used per 100 parts of polymer. Care should be taken, however, in developing formulations that use both gas-phase and solid-phase flame retardant mechanisms because combinations of halogen, antimony oxide, and phosphorus can be synergistic or antagonistic.

Sulfur Compounds. Some inorganic sulfur compounds, such as ammonium sulfate and ammonium sulfamate, form anhydrous sulfur oxides on decomposition. Sulfur compounds are used almost exclusively as low-cost flame retardants for cellulosics. Sulfur compounds are not used commercially as flame retardants in thermoplastics or other polymer systems, although research is being conducted in this

area. On the basis of improved flame retardancy of poly(phenylene sulfide) nonmigrvs. poly(phenylene ether) as a neat polymer, an improvement in flame-retardant properties from those of sulfur-containing materials vs. oxygenated materials could be predicted.

Silicon Compounds [4], [5]. Organosilicon compounds can decompose to form incombustible silicon oxide structures. For example, organosilicon-containing polymers can be formulated to produce low-density silicone foams with excellent flame-retardant characteristics. Proprietary metallic catalysts, such as platinum, are quite efficient in these systems.

Halogenated phenol organosilicon compounds have been developed and are more efficient than analogous compounds without silicon [4], [5]. This efficiency is probably a result of formation of a char containing silicon oxide, and of the vapor-phase activity of the halogen. Care must be exercised in the choice and use of silicon-containing materials. Finely divided calcium silicates used as thickening agents in polymer systems can adversely affect flame retardation.

Polynuclear Aromatics. Some materials, which form highly condensed aromatic structures when heated, ultimately burn because they are hydrocarbons, but do so with difficulty. Untreated *phenolic polymers* have good flame-retardant properties. However, they can glow and radiate heat for prolonged periods of time after the original heat source is removed and, therefore, can ignite combustible matter in contact with them.

Complex aromatic polyesters resulting from the alkaline hydrolysis of poly(ethylene terephthalate) residues are effective in polyurethane, polyisocyanurate, and hybrid foam systems as char formers. *Isocyanurate foams* show better flame retardation than conventional urethanes because of the more condensed polynuclear aromatic structure. *Polyamides* based on aliphatic diamines and aliphatic diacids have a limited degree of inherent flame-retardant properties. However, polyamides that have aromatic rather than aliphatic reactant groups have been made into fibers with excellent flame retardation. In fact, Nomex fibers can be made into firefighters' uniforms. *Aromatic polyimides* are also inherently flame retardant (→ High-Performance Fibers).

The concept of polynuclear aromatic effectiveness is also used in *additive bromine compounds,* such as decabromodiphenyl oxide [*1163-19-5*] (Saytex 102 flame retardant) and ethylene bis(tetrabromophthalimide) [*32588-76-4*] (Saytex BT-93 flame retardant). In many systems BT-93 is more effective than 102 even though BT-93 has a lower bromine content. Thermal gravimetric analysis shows significant char formation on decomposition for BT-93 and almost none for decabromodiphenyl oxide.

Intumescence [6]. Systems containing a polyhydric alcohol (e.g., pentaerythritol), a nitrogen source (e.g., melamine), and a phosphorus donor (e.g., ammonium polyphosphate) can form thermally stable foams when heated. The major historical use of this intumescent chemistry has been in water-based paints and coatings. Untreated polypropylene has also been flame retarded effectively with intumescent systems, but at high loadings of 30 – 50 parts per 100 parts of polymer. These polymer formulations generate little smoke when burned.

Even under polypropylene processing conditions, care is required to avoid premature foaming of the ingredients. More thermally stable systems have been proposed for styrenic polymers [6]. These systems are reported to produce little smoke and good resistance to ultraviolet light.

3. Test Methods

Test methods have been developed that vary in complexity from simple (e.g., ASTM D 2863 and UL 94, which require specimens that are 10 cm long) to very complex (e.g., ASTM E 84, which require specimens 750 cm long) [7].

Test methods may evolve on the basis of particular requirements of a country or industrial group. In the United States, materials must meet Underwriters Laboratory standards of performance: polymer compositions used in general electrical components must satisfy the UL 94 test; wire and cable compositions, UL 44 or 62; television components, UL 1410. Products going into coal mines in the People's Republic of China must meet standards based on ASTM D 2863. Insulating foams for sale in the

Federal Republic of Germany must be rated by a DIN test.

These tests are conducted on small specimens instead of finished articles, even though configuration does affect flammability results. Many industrial and regulatory authorities have realized the shortcoming of such simplistic test procedures and have developed larger-scale tests on finished articles. *Full-scale burns* of finished electrical wires and cables, foam-upholstered chairs, and total roofing assemblies are now more common. Because of cost, only the largest companies have facilities to conduct these tests. Specialized consulting testing laboratories have emerged to do this work for smaller companies and to provide unbiased data for product acceptance or regulatory purposes for companies of all sizes.

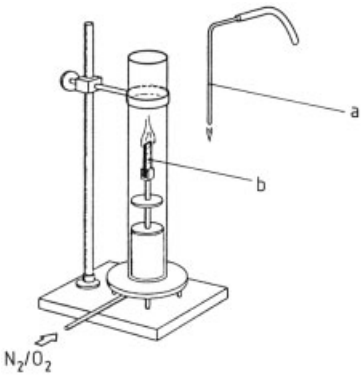

Figure 1. Oxygen index tester, ASTM D 2863 [7, p. 183] a) Pilot flame; b) Burning specimen

ASTM D 2873-77. The oxygen index or limit oxygen index test determines the oxygen concentration at which sustained combustion occurs using a vertically mounted specimen ignited at the upper end (Fig. 1). The test is performed at room temperature, but equipment is available to test for oxygen index at elevated temperature. Results are expressed as $[V_{O_2}/(V_{O_2} + V_{N_2})] \cdot 100\,\%$. Typical oxygen index values (in vol %) for polymers are as follows:

Polyolefins	17.4
Flame-retarded polyolefins	24 – 28
Polystyrene	18.1
Flame-retarded polystyrene	27 – 30
Epoxy resin	24.0
Polyurethane foam	18.0
Rigid poly(vinyl chloride)	40.0
Polytetrafluoroethylene	95.0

The oxygen index test is simple and reproducible and is used to guide development work. Conclusions about fire performance in an actual fire should not be made based on results from the oxygen index test or any other test.

UL 94 of Underwriters Laboratory. The UL 94 test rates materials on the basis of their ability to resist combustion on repeated application of a flame. The test also rates materials for their resistance to dripping with subsequent ignition of a combustible substance (cotton) from the resulting drips.

Test specimens (12.7 × 1.27 cm × various thicknesses) are mounted vertically in a chamber (Fig. 2). Ignition takes place at the lower end of the test specimen. After a 10-s ignition, the flame is removed. If the specimen extinguishes, the afterflame time is recorded and the specimen is reignited for 10 s.

Materials are classified on the basis of afterflame time and ignition of the cotton caused by flaming drips. The most widely accepted classes are 94 V-0, 94 V-1, and 94 V-2. Class 94 V-1 allows a longer afterflame time than 94 V-0, whereas class 94 V-2 has the same afterflame time as 94 V-1 but allows for cotton ignition. Variations of the test utilize different shape specimens and orientations, e.g., horizontal.

Many applications require UL 94 V-0 ratings at very thin sections. Historically, a V-0 rating at 0.32 cm was acceptable; however, more recent products are formulated to achieve a V-0 classification at 0.08 cm. Achieving the V-0 classification at thinner sections is difficult because of the tendency of thermoplastics to melt and drip. Although polymers that drip are permitted by UL 94 V-0, a delicate formulating balance is required to ensure consistent passage of the test.

Because of the worldwide importance of the United States market, companies that export electronic assemblies to the United States manufacture products to the UL 94 standard. This

standard often becomes the accepted norm for products for the individual domestic marketplace as well. Multinational corporations seeking uniformity in many countries also utilize the UL 94 standard to simplify testing.

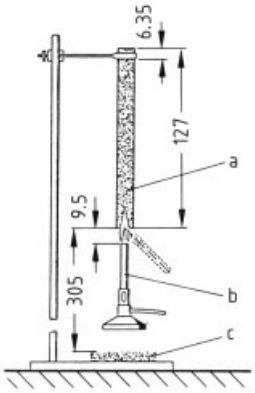

Figure 2. Underwriters Laboratory, UL 94 ∗ [7, p. 350]
a) Specimen; b) Burner; c) Surgical cotton (ca. 50×50 ×6.5 mm)
∗ Dimensions are in mm.

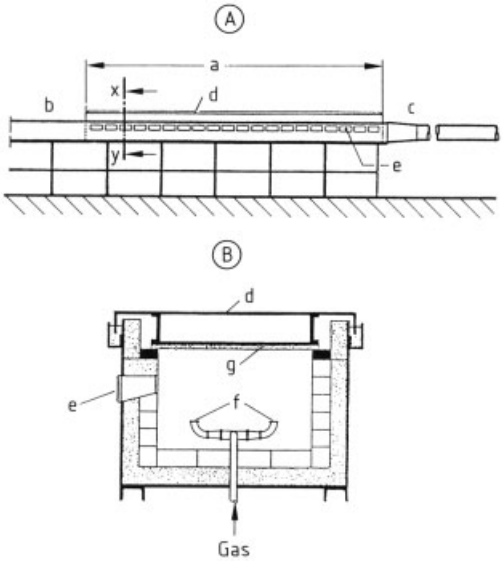

Figure 3. Steiner tunnel, ASTM E 84 [7, p. 229]
A) Longitudinal view: a) Test chamber (7.62 m); b) Fire end; c) Vent end; e) Viewing windows
B) Cross-sectional view of section x – y: d) Removable cover plate; e) Viewing windows; f) Gas burner; g) Specimen

ASTM E 84. The ASTM E 84, or Steiner tunnel test (Fig. 3), is used to rate building and construction materials in the United States for their flame spread, fuel contribution, and smoke evolution.

Test specimens (51.43 ± 1.9 cm wide by 629.9 ± 1.27 cm long) are placed in a tunnel and ignited at one end. Ratings are based on the performance of a standard redwood board (flame spread 100) and asbestos board (flame spread 0).

Typically only the flame spread and smoke values are used to rank materials. Local United States building codes often include E-84 values in their material acceptance criteria. Code values for flame spread are 25 for class I and 75 for class II; smoke density values are 450 for both classes, relative to cement – asbestos (0) and red oak (100).

DIN 4102 Part 1. The Federal Republic of Germany rates building and construction materials by this test as noncombustible (A) or combustible with low (B-1), moderate (B-2), or high (B-3) flammability characteristics.

Four specimens (19×100 cm) are mounted around a chimney and ignited from the bottom for 10 min. At the end of the test, specimens are measured and other observations made. To be rated B-1, the mean length of the specimens must still be 15 cm.

In some variations of the DIN test, surface charring is an important test criteria. Thus, the total specimen could still be B-1 by DIN 4102, but surface charring would render it a B-2 or B-3 classification. Some aliphatic bromine compounds are more effective in reducing this surface flash charring than others.

4. System Development

Flaming can be retarded by incorporation of the flame-retardant molecule into the polymer itself or by admixture of the flame retardant with the polymer. The former is the *reactive approach*, and the latter is the *additive approach*.

The major advantage of the reactive approach is the non-migratory nature of the flame retardant. This eliminates mold plate-out or blooming, which often occurs with flame retardants in plastics systems. A secondary advantage is the ability to produce optimum polymer properties by using the proper balance of reactive material and other monomers. Additive-type flame retar-

dants generally reduce some property, such as impact resistance or heat resistance.

Costs associated with reactive systems are usually higher than with additive systems because of the required development of new polymerization technology. Reactive systems may not be as effective as additive systems because of the more controlled availability of the flame retardant. The ability of flame retardants to melt and migrate to the flame front in the heat of a fire is often desirable; such movement is restricted if the flame retardant is part of the polymer backbone.

Typical examples of systems in which reactive flame retardants can be used are polyurethanes, unsaturated polyesters, epoxy resins, and polycarbonates. However, some additive systems exist for each of these polymer classes, and the choice between additive, reactive, or hybrid systems is based on other formulation parameters, such as properties and costs. For example, glass fiber reinforced bathroom showers and tubs are typically produced with high loadings of aluminum oxide trihydrate as an additive flame retardant. Alternatively, a brominated or chlorinated acid or alcohol can be used in forming the polyester resin itself.

4.1. Selection Criteria of Flame Retardants [8]

Identification of the *flammability test requirements* is the initial criterion for selection of the appropriate flame retardant. In some cases, changing part design or construction material obviates the need for flame retardants.

Establishing the desired *cost–performance parameters* is the second key criterion. For example, a UL 94 V-0 polypropylene may be produced with 10 – 100 parts of various flame retardants per 100 parts of polypropylene. The properties and costs of systems containing high loadings of flame retardant are vastly different from those at low loadings.

Environmental exposure requirements of the finished article is a third criterion. For example, unsaturated polyesters based on dibromoneopentyl glycol [*3296-90-0*] have superior resistance to discoloration by UV light compared to products based on tetrabromophthalic anhydride [*632-79-1*]. Ethylene bis(tetrabromophthalimide) has superior resistance to UV light in styrenic polymers compared with halogenated diphenyl oxides. However, television cabinets and computer components are often painted to mask color changes resulting from fluorescent or UV light exposure in preference to using flame retardants with good light resistance. Other special environmental exposure criteria are optical clarity and corrosion resistance.

Aromatic bromine-based flame retardants have resulted in the development of new products that overcame previous performance deficiencies [8]; these compounds are listed in Table 1.

Compounding (or reacting) of the flame retardant must take into account the stability of the flame retardant, the matrix in which it is suspended, and the effect of the flame retardant or its degradation products on processing or compounding equipment. Typically, thermal gravimetric analysis is used to evaluate the decomposition or volatilization of flame retardants and to aid in selection of the most cost-effective material.

4.2. Formulation with Reactive Flame Retardants [9], [10]

Reactive flame retardants must be combined with the polymer during manufacture or be capable of being grafted during subsequent processing. Each flame retardant has a unique technology that provides the desired level of flame retardation and polymer properties.

Polyurethane Foams. Flame-retardant rigid foams can be produced from polyhydroxy-terminated chlorine- and bromine-containing compounds with a nonreactive organophosphate as diluent to reduce viscosity (Solvay IXOL B-251), or from ester ethers of tetrabromophthalic anhydride (Ethyl Saytex RB-79 flame retardant) [*75790-69-1*]. Derivatives of dibromoneopentyl glycol or organophosphates that have a hydroxyl functionality, such as diethyl-*N,N*-bis(2-hydroxyethyl)aminomethyl phosphonate (Stauffer's Fyrol 6) [*2781-11-5*], can also be used. A formulation for a rigid urethane foam that meets ASTM E 84 Class I is as follows:

Table 1. Aromatic bromine-based flame retardants

Bromine compound	CAS registry no.	Bromine, %	*mp*, °C	10 wt %, loss by TgA, ∗ °C
Pentabromotoluene	[*87-83-2*]	82	280	220
Pentabromophenylallyl ether	[*3555-11-1*]	76	163	190
Pentabromoethylbenzene	[*85-22-3*]	80	136	231
Decabromodiphenyl	[*13654-09-6*]	84	375	330
Pentabromodiphenyl oxide	[*32534-81-9*]	71		260
Octabromodiphenyl oxide	[*32536-52-0*]	79	70 – 150	322
Decabromodiphenyl oxide	[*1163-19-5*]	83	300 – 310	365
Ethylene bis(tetrabromophthalimide)	[*32588-76-4*]	67	445	450
Tetradecabromodiphenoxybenzene	[*58965-66-5*]	82	370	420

∗ TgA = thermogravimetric analysis.

Mobay Multranol 4034	18.6 wt %	
Saytex RB-79 flame retardant	28.0 wt %	
Dow Corning 193 surfactant	0.7 wt %	
Abbott Labs Polycat 8	0.7 wt %	
DuPont Freon 11	13.0 wt %	
Upjohn PAPI 135	39.0 wt %	
Isocyanate index	110.0	
Bromine	14.0 wt %	
ASTM E 84 test		
Thickness	5.1 cm	10.2 cm
Flame spread index	25.0	25.0
Smoke density	135.0	65.0

Manufacturers of foam may make their own formulation or may purchase preformulated packaged systems, which come in two or three parts. One part is usually polymeric isocyanate. A second part contains polyol or a polyol mixture, water, flame retardant, silicone surfactant, catalyst, and a blowing agent. Special catalysts may be supplied as a third part.

High-speed, high-intensity mixers are used to ensure adequate mixing. Reaction takes place between hydroxyl groups and isocyanate groups:

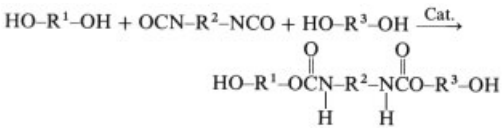

As the complex structure begins to form, heat is generated, the fluorocarbon blowing agent volatilizes, and the foam rises. Cross-linking occurs between the active hydrogens of the carbamate groups with additional isocyanate.

Polycarbonates are produced in complex systems containing both aqueous and organic phases. The reaction between bisphenol A [*80-05-7*] and phosgene [*75-44-5*] is as follows:

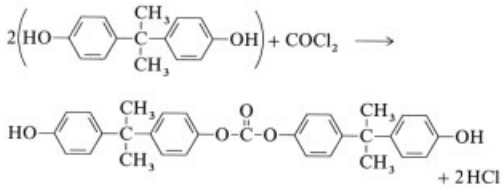

A reactive flame-retarded polycarbonate is produced by substituting tetrabromobisphenol A [*79-94-7*] or another halogenated biphenol for some of the bisphenol A.

Unsaturated Polyesters [11]. Polyesters are produced by reacting glycols with organic acids. Typical ingredients are ethylene and propylene glycols and phthalic or maleic anhydride. Acid anhydrides are preferred because they generate only half as much water of reaction as the corresponding diacid. Esterification catalysts, such as *p*-toluenesulfonic acid or titanium complexes, are often used.

Unsaturated polyesters contain maleic anhydride or tetrahydrophthalic anhydride, which results in aliphatic unsaturation of the backbone. These polyesters are capable of further cross-linking with styrene, divinylbenzene, or another vinyl monomer to produce highly rigid, complex networks. The styrene – unsaturated polyester systems are usually further reinforced with fiberglass to produce a wide range of products from bathtubs to boats.

When flame retardation is required, halogens, most often chlorine or bromine, are selected. The halogen is introduced in several ways:

Table 2. Formulation of two typical flame-retardant unsaturated polyesters containing tetrabromophthalic anhydride

Component	mol	mol
Tetrabromophthalic anhydride	0.40	0.73
Phthalic anhydride	0.60	0.27
Maleic anhydride	1.00	1.00
Ethylene glycol	2.06	
Propylene glycol		2.06
% Bromine	18.0	25.0
Oxygen index	34.3	34.0

Table 3. Formulation of a flame-retardant unsaturated polyester containing dibromoneopentyl glycol

Component	mol
Part A ∗	
Maleic anhydride	1.9
Phthalic anhydride	1.9
Dibromoneopentyl glycol	4.0
Part B ∗	
Maleic anhydride	4.5
Phthalic anhydride	4.5
Propylene glycol	10.0

∗ Part A contains 30 wt % bromine and part B contains no bromine; the two parts are mixed to achieve 21 wt % bromine.

1) Direct bromination of the unsaturated aliphatic polyesters (Koppers Dion resins)
2) Dibromoneopentyl glycol [*3296-90-0*] as a replacement for ethylene or propylene glycols (Dow Chemical FR-1138 or Dead Sea Bromine FR-522)
3) Hydroxyethylated tetrabromobisphenol A [*4162-45-2*] as a replacement for a portion of the glycols
4) Tetrabromophthalic anhydride as a replacement for a portion of the phthalic anhydride (Ethyl Saytex RB-49 flame retardant or Great Lakes Chemical PHT-4)
5) Chlorendic anhydride [*115-27-5*] (Velsicol Chemical Co.) as a replacement for a portion of the phthalic anhydride
6) Tetrachlorophthalic anhydride [*117-08-8*] (Monsanto Tetraphthal) as a replacement for a portion of the phthalic anhydride
7) Diallyl tetrabromophthalate [*49693-09-6*] to replace some or all of the cross-linking styrene
8) Halogenated styrene of divinylbenzene to replace some or all of the cross-linking styrene

Cost and product performance are the controlling variables in selecting the flame retardant. Typical formulations containing tetrabromophthalic anhydride or dibromoneopentyl glycol [11] are shown in Tables 2 and 3.

Epoxy Resins. Epoxy resins are produced from the reaction of bisphenol A and epichlorohydrin (→ Epoxy Resins). Reaction conditions are chosen carefully to ensure consistency of epoxide concentration in the resin. Of particular concern are trace impurities, such as inorganic halides, that remain in the finished product. Inorganic halides can adversely affect electrical properties in finished epoxy resin–glass fiber printed circuit boards. Organic halides that are readily hydrolyzed by mild alkali can interfere with the rate and extent of cure. Most epoxy resin manufacturers stress consistency of both hydrolyzable and ionic halides so that users need not make changes in their formulation or operation.

Table 4. Formulation of a low-bromine flame-retardant epoxy resin

Component	parts by wt
Solution A	
Shell Epon 828 resin	64.57
Saytex RB-100	35.43
Triphenyl phosphine	0.20
Acetone	25.00
Solution B ∗	
Resin solution A	125.20
Dicyandiamide	2.90
Benzyldimethylamine	0.20
Acetone or methyl cellosolve	75.00

∗ Solution B is used to make the glass laminates.

Flame-retarded epoxy resins are made by using tetrabromobisphenol A to partially or almost totally replace the bisphenol A. These resins contain either 20 or 40 wt % bromine and are referred to as low bromine and high bromine, respectively. The choice of which resin to use is based on the flame-retardant requirements of the printed circuit board or other application.

Typical FR-4 board manufacture. Printed circuit board stock is manufactured from catalyzed epoxy resins and glass cloth. The resin is applied in an impregnating bath. Partial curing or B staging occurs so that the glass cloth can be handled but is not too tacky or sticky. Multilayers of the B-staged cloth and copper foil are placed in a

curing press, and the lamination is then formed and cured (→ Epoxy Resins, Chap. 4.2.1.).

A typical low-bromine epoxy resin formulation is given in Table 4.

Printed circuit boards are also made with polyester or phenolic resins. The polyester boards are based on brominated resins with antimony oxide synergist. Phenolic boards often contain brominated diphenyl oxides and occasionally a phosphate plasticizer or diaminodiphenylsulfone.

Halogenated monomers , if available, can be used to make inherently flame-retarded polymers, such as vinyl chloride [*75-01-4*] in poly(vinyl chloride) or ethylene – vinyl chloride copolymers. Vinylidene chloride [*75-35-4*] resins are also inherently flame-retardant. Vinyl bromide [*593-60-2*] is copolymerized with acrylic monomers in flame-retardant carpet fiber applications.

Halogenated styrene monomers have been investigated to find compounds having high halogen content, adequate storage stability, and a satisfactory reaction rate. Although development quantities of both chloro- and bromostyrenes are available, none have achieved commercial success.

Acrylic or allylic monomers with pendant tribromophenol [*118-79-6*] or pentabromophenol [*608-71-9*] groups are also available. These can be either co- or homopolymerized. The homopolymer itself can be used as a flame retardant. Two examples of polymeric or oligomeric flame retardants from halogenated monomers are poly(dibromophenylene oxide) [*26023-27-8*] (Great Lakes PO-64P), made from 2,4,6-tribromophenol [*118-79-6*]; and poly(pentabromobenzyl)acrylate [*59447-57-3*] (Dead Sea FR-1025), made from pentabromobenzyl bromide [*38521-51-6*].

Glycidyl derivatives of brominated phenols have also been manufactured and have many uses.

4.3. Formulation with Additive Flame Retardants

Compounds and formulations with additive flame retardants depend on the polymer matrix and the end-use requirements [12–17]. Combinations of chain-stiffening and plasticizing flame retardants to achieve an optimum balance of properties in thermoplastics have been described [16], [17].

The major emphasis in this section is on additives used in conjunction with thermoplastics. Instead of the reactive flame retardants discussed in Section 4.2, additives may be used in the same polymer systems as follows:

1) polyurethane foams – organophosphates, brominated organics, alumina trihydrate
2) polycarbonates – brominated organic oligomers, sulfonate salts
3) unsaturated polyesters – brominated organics, alumina trihydrate
4) epoxy resins – brominated organics

Laboratory Compounding [1], [18], [19]. The equipment used for laboratory compounding of thermoplastics, such as polyolefins, styrenics, and polyesters, depends on the characteristics of the polymer, flame retardant, and other additives.

Typically the polymer is fluxed, masticated, or otherwise prepared for acceptance of the flame retardant mixture. When both halogenated flame retardant and antimony oxide are used, they are preblended and added to the fluxed polymer. In general, the procedures and equipment for thermoplastics are also used for thermoplastic or thermosetting elastomeric polymers.

Three types of laboratory equipment are employed:

The laboratory *Brabender Plasticorder* has two counter-rotating, intermeshing, shear kneading rollers that masticate the polymer. Heat is provided electrically.

A typical compounding *mill* has two counter-rotating chrome-plated steel rolls, which run at different speeds to increase shearing and mixing, and can be individually heated or cooled.

The Haake compounding twin-screw *extruder* has two intermeshing screws and a single drive. More complex equipment has independent drives with either concurrent or countercurrent rotation. Heat is varied along the barrel to increase or reduce melt viscosity and shear. Some equipment has an individual port along the barrel for addition of various ingredients or for venting off-gas.

Table 5. Thermoplastics processing temperatures

Polymer	Operating temperature, °C		
	Brabender	Mill	Extruder
Acrylonitrile – butadiene – styrene	175	170	170
Polystyrene (high-impact)	190	175	175 – 200
Polypropylene	175	210	175 – 200
Polyethylene (low-density)	120	140	
Polyethylene (high-density)	175	165	175 – 200
Polyamide	260		260 – 270
Poly(butylene terephthalate)	270		240 – 260
Poly(ethylene terephthalate)	270		240 – 260

Operating temperatures for the three types of plastics compounding equipment are shown for a variety of polymers in Table 5.

Once the plastic mass has been well compounded, finished plaques for further testing are formed by compression molding in standard heated presses or by injection molding. Molding conditions for some thermoplastics are shown in Table 6.

Table 6. Thermoplastics injection molding

	Temperature, °C	Pressure, MPa
Polyethylene	190 – 260	3.6 – 3.8
Polypropylene	230 – 275	3.6 – 3.8
Poly(vinyl chloride)	175 – 200	3.4 – 3.8
Polystyrene (crystalline)	220 – 260	4.0 – 4.1
Polystyrene (high-impact)	205 – 230	4.0 – 4.1
Acrylonitrile – butadiene – styrene	220 – 260	3.6 – 4.1
Poly(ethylene terephthalate) (glass filled)	235 – 275	3.6 – 4.1
Poly(butylene terephthalate) (glass filled)	230 – 275	3.8 – 4.1
Polyamide 6 and Polyamide 66 (natural resin only)	275 – 295	3.4 – 3.6
Polycarbonate (natural resin only)	260 – 295	3.4 – 3.6

Test specimens are evaluated for flame retardation and for other physical properties. These tests are also used in thermoplastic or thermosetting systems based on reactive flame retardants. Typical property tests are

Heat distortion temperature	ASTM D 648
Vicat softening point	ASTM D 1525
Izod impact strength	ASTM D 256
Dart drop resistance	ASTM D 1709
Tensile – elongation	ASTM D 638
Flexural strength	ASTM D 790

In almost all cases, flame-retardant additives reduce the impact strength of virgin polymers. Very high-melting or reinforcing flame retardants increase thermal deformation resistance, whereas low-melting flame retardants reduce thermal deformation resistance. Flexural strength and tensile strength are increased with high-melting flame retardants, but percent elongation is reduced. Table 7 shows the wide divergence in some physical properties attainable in an acrylonitrile – butadiene – styrene polymer with five flame retardants [8].

Chlorine and bromine flame-retardant compounds have been compared in wire and cable compositions [18]; retardants for thermoplastic polyurethanes have been described [19]. Differences in flame retardation were observed between ester and ether polyurethanes.

Commercial Compounding. Commercial equipment varies only slightly from the laboratory counterparts. Mills are larger or contain three rolls rather than two. Extruders are also larger.

In place of the Brabender, commercial facilities use Banburys with various designs of kneading devices. The Farrel continuous mixer (FCM, Farrel Manufacturing Co.) combines some of the features of both a Banbury and an extruder. The FCM has two counter-rotating mixing motors similar to a Banbury. The ends of the rotors are extended with screw flights evolving into a metering twin-screw extruder. The plastic is continuously extruded through a pelletizer die.

The producer of a television set, for example, purchases a molded cabinet from a custom molder. The molder purchases plastic pellets from a manufacturer who warrants that the plastic meets a UL 94 V-0 rating. The pellet manufacturer has the following two options:

1) Mix base resin with flame retardant and antimony oxide and make a finished compound in pellet form.

Table 7. Properties of flame-retardant acrylonitrile – butadiene – styrene polymers

	Neat resin	DBDPO[a]	OBDPO[b]	TDBDPB[c]	TBBPA[d]	HBDPE[e]
Oxygen index	18.4	27.1	28.3	31.2	33.7	25.8
UL 94 (0.32 cm[f])	burn	V-0	V-0	V-0	V-0	V-0
(0.16 cm[f])	burn	V-0	burn	V-0	V-0	burn
Tensile strength, MPa	36	40	44	39	33	40
Flexural modulus, GPa	2.2	1.7	2.0	2.3	2.7	2.0
Vicat softening point (0.32 cm[f]), °C	100	102	97	102	85	96
Izod impact strength (0.32 cm[f]), J/m	34.3	6.9	13.7	5.9	4.9	12.8

[a] Decabromodiphenyl oxide.
[b] Octabromodiphenyl oxide.
[c] Tetradecabromodiphenoxybenzene.
[d] Tetrabromobisphenol A.
[e] Bis(tribromophenoxy)ethane.
[f] Sample thickness.

2) Mix base resin with a premixed concentrate of flame retardant and antimony oxide purchased from a concentrate manufacturer. These concentrates, which can contain up to 90 wt % of active material with 10 wt % binder, are "let down" with base resin to the desired level of active ingredient in the finished compound.

Although individual manufacturers are the best source of specific formulation information, following are some typical flame-retardant systems for various thermoplastics. All are based on 100 parts of original base polymer. Unless otherwise indicated, the formulations give UL 94 V-0 compositions of 0.32 cm thickness.

High-impact polystyrene: 14 parts decabromodiphenyl oxide with 4.75 parts of antimony oxide.

Acrylonitrile – butadiene – styrene (ABS):

A total of 23 parts of octabromodiphenyl oxide [*32536-52-0*] with 5 parts of antimony oxide, or 34 parts of tetrabromobisphenol A and 8.6 parts of antimony oxide.

Polypropylene:

A total of 4 parts of ethylene bis(5,6-dibromonorbornane-2,3-dicarboximide) (Ethyl Saytex BN-451 flame retardant) [*41291-34-3*], or bis(2,3-dibromopropyl ether)tetrabromobisphenol S (Great Lakes Chemical PE68) [*21850-44-2*], or bis(2,3-dibromopropyl ether) tetrabromobisphenol S (Marubishi Non-Nen 52) [*42756-55-1*] with 2 parts of antimony oxide produces UL 94 V-2 formulations. Ten parts of Non-Nen 52 with four parts of antimony oxide produces a V-0. Intumescent formulations based on pentaerythritol, phosphorus donor, and melamines produce a V-0 rating at 24 – 40 parts per 100 of polymer; V-0 is also achievable with 40 – 80 parts of aluminum oxide trihydrate or magnesium hydroxide (hydrated magnesia).

Glass-reinforced (30 wt %) poly(buylene terephthalate) (PBT) or poly(ethylene terephthalate) (PET):

several products are used commercially such as tribrominated polystyrene (Ferro Pyrochek 68PB) [*57137-10-7*], polycarbonate oligomers of tetrabromobisphenol A (Great Lakes Chemical BC-52 [*28906-13-0*] or BC-58 [*71342-77-3*]), poly(dibromophenylene oxide) (Great Lakes Chemical PO64P), decabromodiphenyl oxide or ethylene bis(tetrabromophthalimide) (EthylSaytex BT-93 flame retardant). From 6 to 8 wt % aromatic bromine with 3 – 5 parts of antimony oxide are used. Actual levels of the flame-retardant compound itself are adjusted on the basis of its individual bromine content. The ratio of brominated compound to antimony oxide is generally 2 – 3 : 1. However, in countries such as the People's Republic of China, where antimony oxide is plentiful, economics often dictate the ratio of materials, with the result that more antimony oxide is used.

Systems not containing glassrequire a higher amount of flame retardant. Levels are increased to 10 – 14 % of contained bromine and 4 – 8 parts of antimony oxide.

Low-density polystyrene foam:

A total of 1 – 2 parts of hexabromocyclododecane [*3194-55-6*], pentabromomonochlorocyclohexane (Dow Chemical 651P) [*87-84-3*], and dibromoethyldibromocyclo-

hexane (Ethyl Saytex BCL-462 flame retardant) [*3322-93-8*] are adequate to produce foam that meets building code requirements in the United States; UL 94 values on these foams are not significant.

Poly(phenylene oxide) high-impact polystyrene blends:

A total of 15 – 20 parts of an aromatic phosphate, such as triphenylphosphate [*115-86-6*], in nominal 50 : 50 blends of the two polymers.

In addition to their use in plastics, flame retardants are also necessary in other polymer systems, such as elastomers, textiles, paints, paper, and coatings. Where processing temperatures or environmental exposure requirements are not severe, chlorinated alkanes can economically replace brominated organics. Where mass per unit volume or physical properties are not critical, aluminum oxide trihydrate or other fillers can replace more expensive flame retardants.

Commercial applications of the flame retardant are a function of end-use requirements. For example, flame retardants can be made an integral part of a polypropylene upholstery or carpet fiber through melt processing or can be applied to the finished fabric as a back coating. In the former case, the retardant must be fluid at the spinning temperature or be a fine, uniform particle size powder. Either a liquid or solid flame retardant can be used for back coating; decabromodiphenyl oxide dispersed in a latex binder system is used widely for polypropylene drapery or upholstery fabric.

Clothing has also been treated for flame retardation with a variety of halogen- and phosphorus-containing materials. Of particular concern are the effects of the chemicals on direct skin contact. Durability to washing and dry cleaning are also critical factors. Finally, the flame retardant must not affect the fabric color or feel. In most cases, proprietary formulations, which are specific for fiber content, have been developed. Testing of flame-retarded fabrics is by both horizontal and vertical procedures.

Paper is also flame-retarded. Typical nondurable systems are based on inorganic salts, such as ammonium phosphates and sodium borates. More durable flame retardants are based on chlorine-containing emulsion polymers synergized with antimony oxide. The flame retardants are most often applied to the paper in a saturation or coating operation.

4.4. Interactions between Flame Retardants and Other Formulation Ingredients

Fillers. Calcium carbonate can interfere with the flame-retardant action of halogenated flame retardants because it absorbs and neutralizes HCl or HBr. Carbon blacks, with high levels of residual oils on their surface, make flame retarding difficult in any system. Fine-particle colloidal silica used to minimize dripping in flame-retarded thermoplastics also contributes to increased flammability because of its high surface area.

When high loadings or fillers are used, dispersion is difficult and the effect on polymer physical properties is significant. Fillers with surface treatments of organotitanate or organosilicon coupling agents are useful to overcome the dispersion and physical property loss problems. In some instances flame retardance has been improved because of the addition of a coupling agent.

Some evidence even suggests that aluminum oxide trihydrate or magnesium hydroxide at low loadings can act synergistically with halogen antimony oxide systems.

Colorants. Some inorganic cadmium and zinc pigments are deleterious to flame retardation. The mechanism for this phenomenon is unknown.

Thermal Stabilizers and Processing Aids. Zinc-based processing aids can cause discoloration of a polymer matrix when used in combination with halogenated materials. Calcium stearates are generally used when this problem occurs. Depolymerization of styrenic polymers caused by some halogenated flame retardants during polymer processing can be minimized by organotin systems. These systems also usually contain complex phenolic antioxidants, such as Irganox 1010 (Ciba-Geigy) [*6683-19-8*], Ethanox 330 (Ethyl) [*1709-70-2*], and occasionally thiols.

Synergists [20]. Tin compounds are effective synergistic agents as partial or total replacements for antimony oxide [21]. A total flame retardant for polypropylene is reported to contain 32 % bromine, 1 % phosphorus, and 1.5 % tin.

Phosphites can act synergistically with certain brominated flame retardants in the absence of antimony oxide. In other systems containing antimony oxide, loss of flame retardation has been observed with phosphites.

Peroxide-producing chemicals, such as 2,3-dimethyl-2,3-diphenylbutane [*1889-67-4*], can be used in combination with aliphatic or cycloaliphatic bromine compounds in styrenic polymers. Additive levels of the bromine can sometimes be significantly reduced by as much as 50 % in polystyrene foam manufacture.

Combinations of zinc borate [*12536-65-1*] and antimony oxide mixed with decabromodiphenyl oxide or ethylene bis(tetrabromophthalimide) are used in wire and cable formulations based on thermoplastic styrene – butadiene copolymers. Higher oxygen indices are obtained with these mixtures than with antimony oxide alone.

Ultraviolet Light Stabilizers and Antioxidants (→ Antioxidants, Chap. 5.4.), [22–24]. Each polymer system with its flame retardant is unique and requires a specific stabilization package. What is acceptable in high impact polystyrene or ABS may be totally ineffective in alloys or blends of these substances with other polymers.

Consistency of color under exposure to ultraviolet or fluorescent light has become a critical issue in business machines and other plastic cabinetry. It is equally important in fabrics, paints, and coatings.

Impact polystyrene containing an aromatic bromine or an aliphatic bromine – chlorine compound has been studied with a wide range of light stabilizers in combination with antioxidants [22], [23]. In all cases, the flame-retarded impact polystyrene had poorer light stability than the polymer alone.

Benzotriazoles combined with hindered phenolic antioxidants were most effective with aromatic bromine-based flame retardants. Further color improvement was obtained by adding an epoxy novolac. The system containing an aliphatic halogen compound was improved with benzotriazoles. Hindered amines were effective when combined with benzotriazoles.

Flame retardants that are stable to UV light have been developed. However, the most demanding applications require additional UV stabilizers. The UV stability of brominated flame retardants has been rated as follows [8]:

1) brominated diphenyl oxides: poor
2) bis(tribromophenoxy)ethane (Great Lakes Chemical FF680) [*37853-59-1*]: good
3) ethylene bis(tetrabromophthalimide) (Ethyl Saytex BT-93 flame retardant): excellent

4.5. Formulation to Reduce Smoke

All burning materials generate smoke as a result of incomplete combustion. Styrenic polymers and polyurethanes, for example, produce thick black, carbonaceous, sooty smoke. Flame retardants that interfere with the normal combustion process change the rate of smoke evolution and the amount and composition of the smoke. Less smoke allows individuals more time to safely exit from confined areas.

Much of the development work on smoke suppression has been done with poly(vinyl chloride). Zinc borate reduces smoke and can be used as a partial replacement for antimony oxide in some formulations. Molybdenum compounds are also used in PVC formulations.

Unfortunately, PVC smoke reduction technology has had limited success in other polymers. Molybdenum compounds do function in unsaturated polyesters, however.

Iron compounds, such as ferrocene (Arapahoe DFR-121) [*102-54-5*], have been used in poly(vinyl chloride) and polyurethanes. Iron compounds have also been reported to reduce smoke in polyamide formulations that contain chlorinated organic flame retardants and antimony oxide.

Other metals, such as lead or bismuth, have had limited applications as have quinoline [*91-22-5*] and phthalocyanine complexes. Typically from 0.5 to 6 parts of smoke suppressant per 100 parts of polymer are used.

Smoke from halogen – antimony oxide formulations can be eliminated by using other flame

retardants. Aluminum oxide trihydrate and magnesium hydroxide do not generate smoke. Intumescing systems based on pentaerythritol, a phosphorus donor, and melamine also do not materially contribute to smoke.

Measurement of Smoke Evolution [21], [25], [26]. Smoke is measured peripherally to flammability in some tests, such as ASTM E 84, and independently in others. Smoke measurement testing by small-scale laboratory techniques and full-scale experiments has been reviewed [21].

Little correlation between test methods was observed, as each initiates combustion with different energy input and either a fixed volume or moving air stream. Some protocols include either a radiant energy source or a flaming source as the combustion initiator. Most are based on some form of visual observation; however, the Arapahoe smoke test (ASTM D 4100) rates materials by mass transfer per unit time from the specimen to a collecting filter.

Smoke measuring devices and smoke reduction techniques in olefin polymers have been studied [25].

The most widely used device in the United States is the National Bureau of Standards Chamber (ASTM E 662). The chamber is shown in Figure 4.

The Ohio State University release rate apparatus is gaining acceptance. This device monitors smoke continuously as it exits the apparatus. Since air flow rates can be varied, actual fire situations can be simulated more closely than with the fixed-volume, static ASTM E 662 procedure. Flame retardants that reduced smoke levels significantly have been found using the device [24].

5. Flame Retardant Markets [27–32]

Modern Plastics magazine publishes an annual listing of the major classes of flame retardants in the United States [27]. Included in the article are new products that have appeared in the trade press. Table 8 is the September, 1985 list of major flame-retardant types and their consumption.

Flame retardants in the United States were the subject of studies by SRI International [28]. Updates appear in SRI's tabulations of plastics additives and specialty chemicals. The growth in consumption of flame-retardant chemicals in Western Europe from 1985 to 1992 is shown in Table 9 [32]. The Japanese market for flame retardant chemicals was estimated to be 80×10^3 t in 1984. Japan is the predominant consumer in the Far East, although Taiwan, Korea, and the People's Republic of China are growing users. There are smaller users in South America and South Africa.

Table 8. Consumption of flame retardants by type * [27]

Type of flame retardant	U.S. consumption, 10^3 t	
	1984	1985
Additives		
Aluminum oxide trihydrates	91.8	93.7
Antimony oxides	15.0	15.9
Boron compounds	5.0	5.0
Bromine compounds	15.0	16.9
Chlorinated alkanes and cycloaliphatics	15.0	13.0
Phosphate esters		
Nonhalogenated	16.9	17.0
Halogenated	11.0	11.0
Others **	8.0	8.1
Total	177.7	180.6
Reactives		
Epoxy resin intermediates	11.0	10.0
Polycarbonate intermediates	3.0	3.0
Polyester intermediates	5.0	5.9
Urethane intermediates		
Flexible foam	1.0	1.0
Rigid foam	4.0	4.0
Others	4.0	4.0
Total	28.0	27.9
Grand total	205.7	208.5

* Include secondary plasticizers with flame retardant functionality.
** Include molybdenum, zinc, and metal oxides.

Electronic, electrical [29], and construction [30] applications use the bulk of flame retardants. Wire and cable applications represent a significant specialty subset of electronics and electrical [31].

The Fire Retardant Chemicals Association, with headquarters in Lancaster, Pennsylvania, was formed to allow a common forum for discussion and interchange on matters of concern to manufacturers and users of flame retardants.

Listings of suppliers of flame retardants in the United States are given in the SRI report [28], and lists of suppliers on a worldwide basis including the United States, Europe, and Japan are available [2].

Markets for flame retardants were developed after governmental regulatory action; loss of life

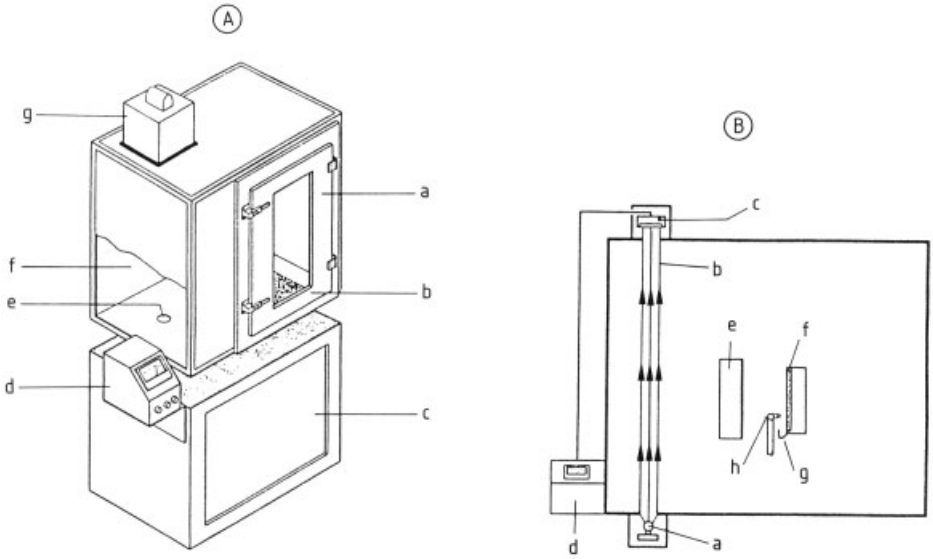

Figure 4. National Bureau of Standards (NBS) smoke chamber, ASTM E 662 [7, pp. 394, 395]
A) View of chamber: a) Door with glass window; b) Blow-out panel; c) Controls and measuring instruments; d) Photometer; e) Window for optical system; f) Test chamber; g) Photodetector housing
B) Operational schematic: a) Light source; b) Light beam; c) Photodetector; d) Photometer; e) Infrared radiator; f) Specimen; g) Trough for specimen melt; h) Six-tube burner

and property because of fire resulted in public pressure to provide "safer materials." Insurance companies also exerted pressure by increasing payment rates in unprotected environments. Pressure still exists for improved products with lower levels of smoke and with reduced toxicity of the combustion products.

Table 9. Estimated West European consumption of flame-retardant chemicals

Product group	Consumption, 10^3 t	
	1985	1992
Polystyrene	4.0 – 4.5	4.5 – 5.0
ABS	1.0 – 1.5	1.2 – 1.8
Polyesters	7.5 – 8.0	8.5 – 9.0
Epoxy resins	3.5 – 4.0	4.0 – 4.5
Polyolefins	10.0 – 12.0	11.0 – 13.0
Poly(vinyl chloride)	25.0 – 27.0	27.0 – 29.0
Polyurethanes	12.0 – 13.5	13.5 – 15.0
Engineering plastics	1.5 – 1.8	1.7 – 2.0
Paper and textiles	9.0 – 10.0	10.0 – 11.0
Rubber and elastomers	5.0 – 6.0	6.0 – 7.0
Other	11.5 – 11.7	12.6 – 12.7
Total	90.0 – 100.0	100.0 – 110.0

Electrical and Electronic Applications. Cabinets and housings for television sets and the polymeric components in the sets (e.g., connectors) must meet flammability standards. Printed circuit boards and internal wires are also processed with flame retardants.

All commercial wires and cables along with connectors, plugs, and spacers are treated with flame retardants.

Construction. Thermal insulating materials are flame retardant. Molded fiberglass units to be used as tubs and shower stalls must meet flammability standards. Institutional mattresses represent a specific use of flame retardants. Furniture cushioning of all types is under active flame retardant development.

Transportation. Polymer components inside mass transit vehicles (airplanes, buses, and subways) must meet flammability standards. Most polymeric materials in private automobiles are flame-retardant, but comply with different flammability standards than mass transit vehicles. Smoke and toxic gas generation from burning vehicle interiors are a significant concern in restricting free exit. Seating material is a major problem because of its total mass in the vehicle and accessibility to accidental or deliberately set fires.

Fabrics and Apparel. Children's sleepwear is subjected to flammability standards, as is workers' clothing in specific industries. Carpets and draperies are treated with flame retardants. Military fabrics, such as tenting, are flame-retardant.

6. Toxicology

Regulatory Influences. The development and introduction of new materials are becoming complex. Information is required by regulatory agencies on the health, safety, and disposal of the product and on its raw materials and byproducts. Often, expensive and lengthy toxicology testing is required prior to commercialization. This testing can include developmental toxicology, as well as teratology.

Workers are becoming more concerned with the long-term safety and health effects of the chemicals they handle and use. Tris(2,3-dibromopropylphosphate) [*126-72-7*] has been banned for use in children's sleepwear, but other uses have not been banned in the United States. The Federal Republic of Germany is evaluating legislation that would restrict the use of chemicals containing or generating chlorinated dibenzoparadioxins. The Japanese Ministries require detailed testing on the effects of chemicals on various fish species.

The regulatory climate and its cost will hinder development of new materials and will emphasize compounding already existing products as much as possible.

Toxic Products of Combustion [32–36]. When materials burn, they liberate both gaseous and solid combustion productions. Inhalation of these decomposition products can cause both short-term and long-term effects. Flame retardants alter the combustion process so that the decomposition products are different from non-flame-retarded materials. The entire process is dynamic, with the composition of liberated gases varying with time.

The evaluation of the toxic effects of combustion products has only recently received significant attention. Toxicity of combustion products is important with materials exposed to fire, but the situation should be viewed together with other aspects of a material's performance, not independently. Resistance to ignition, reduction in the rate or extent of flame spread, and minimization of smoke evolution are all effects produced by flame retardants, which result in the saving of life and property. Combustion toxicity should be included in an overall hazard index with proper emphasis on other protection benefits.

Currently, an accepted standard method for measuring combustion toxicology does not exist. More work is needed in evaluating laboratory tests before material-restricting regulations can be formulated on the basis of the test results [33]. Material design with total fire safety in mind is essential [34].

Time and concentration factors in combustion toxicology have been evaluated [34]. Both are important in determining short- and long-term toxicologic effects. Because fires are dynamic, both time and concentration of various toxicant gases change. Reducing this complex situation to laboratory testing of relatively pure substances is a difficult, if not impossible, task.

Some chair-seat cushioning material produces more toxic gases under smoldering ignition than others [35]. The toxicologic effects of thermoplastics have been compared at normal processing temperatures and above [36]. At some elevated temperatures, Douglas fir had a higher irritation potency than many of the plastics although the effect was more transitory. Large amounts of carbon monoxide were liberated from commercial plastics formulations with and without a flame retardant at temperatures significantly above normal processing temperatures.

Several states in the United States are reviewing toxicity of products of combustion data to determine whether or not regulatory action is required. Political activists in the Federal Republic of Germany are pressing for a ban on certain flame retardants because they produce toxic combustion products.

7. Future Directions in Flame Retardants

New Market Requirements. Flame retardants are used only when this is required by the marketplace. Market changes can increase

or decrease the usage of flame retardants. Although few regulations exist in South America, multinational companies producing subassemblies for the United States, Japanese, or European markets meet the importing country's flame retardant requirements. To ease inventory and production problems, the same companies make flame retardant assemblies for domestic consumption.

Polyurethane and polystyrene foams are efficient thermal insulators per unit thickness, but are more combustible than traditional materials, such as rock wool or fiberglass. Flame retardants are required, therefore, as these materials replace conventional insulation. Other polymer foams (isocyanurate and phenolic) have inherent flame retardance but are deficient in physical properties.

Glass fiber reinforced epoxy printed circuit boards and phenolic paper boards are flame retarded. These low-cost mass-produced articles have been used widely. Developments in flexible circuitry and molded circuits have overcome the rigidity of previous materials. Many high-performance polymers (e.g., polyimides) used in newer circuits are inherently flame retardant.

Specification Changes. Electronic cabinetry is almost exclusively rated UL 94 V-0, with some rated UL 94 V-2. This evolved from electron-tube-powered television sets, where heat evolution from the tubes could result in fire. With the advent of transistorized circuitry with little heat output, no commensurate change was made in the flammability test requirement. An effort is under way to modify the internal configuration of electronic components, shield them, and thus totally eliminate the possibility of their causing plastics ignition. If the rating does change, the need for flame retardants would be reduced or eliminated.

On the other hand, institutional bedding and mass transit vehicles are developing more stringent requirements for flame retardants. The effort to thwart damage results from purposeful and accidental ignition.

8. References

1. G. E. Bowen: *Magnesium Hydroxide: A Halogen-Free Flame and Smoke Suppressant for Polypropylene,* 2nd International Flame Retardants Symposium, London 1985, pp. 2/1.
2. Herbert Jenkner in R. Gachter, H. Muller (eds.): "Flame Retardants for Thermoplastics." *Plastics Additives Handbook,* Hanser Publishers, München p. 535.
3. Irv Touval, *Plast. Compd.* 1982, (Sept. – Oct.) 31.
4. Ethyl Corp., US 4 476 268, 1984 (H. J. Barda, S. Hussain)
5. Ethyl Corp., US 4 476 267, 1983 (H. J. Barda, S. Hussain)
6. Reference I, O. Cichetti, R. Fontanelli, A. Pagliavi: "Advances in the Field of Intumescent Halogen-Free FRs for Polymers, pp. 13/1.
7. J. Troitzch: *International Plastics Flammability Handbook,* Hanser Publishers, Munich 1983, Chapter 7.
8. Reference I, B. J. Sutker, C. J. Abbott: "Evolution of Aromatic Flame Retardants," pp. 5/1.
9. D. P. Miller, *Plast. Eng.* 1980, (Feb.) 29.
10. R. C. Nametz, *Plast. Compd.* 1984, (Sept. – Oct.) 54.
11. E. R. Larsen, W. C. Weaver: "FR-1138 Dibromoneopentyl Glycol-Based Unsaturated Polyesters: Preparation and Evaluation. *Proc. 28th Annual Technical Conference,* Reinforced Plastics/Composites Institute, Soc. of the Plastics Industry, Washington, D.C., 1973, sec. 2 – A, p. 1.
12. N. Albee, *Plast. Compd.* 1981, (Jul. – Aug.) 89.
13. N. Albee, *Plast. Compd.* 1981, (Sept. – Oct.) 95.
14. J. P. Hamilton, *Plast. Compd.* 1978, (Sept. – Oct.) 54
15. R. C. Nametz, *Plast. Compd.* 1979, (Jan./Feb.) 31.
16. E. P. Chang, *J. Appl. Polym. Sci.* **21** (1977) 937.
17. E. P. Chang, *J. Appl. Polym. Sci.* **21** (1977) 2167.
18. J. M. Lesniewski: "Comparison of the Relative Cost Effectiveness of Brominated and Chlorinated Flame Retardants in EPDM," *Proc. Symp. International Wire and Cable,* 1980, Cherry Hill, N.J., p. 90.
19. J. Sutker, S. Walter, *J. Fire Sci.* **1** 1983, (Jan./Feb.) 66.
20. "Tin Chemicals Fight Fire, Tin and its Uses" International Tin Research Institute, Middlesex, England, 1985 p. 146.
21. J. G. Quintiere, *Fire Mater.* **6** (1982) no. 3, 4, 145.

22. R. Bradley, J. Farber, L. Tests: "Light Stabilization of Flame Retardant Impact Polystyrene," National Technical Conf. of Soc. of Plastic Engineers, Louisville, Ky., Nov. 17 – 19, 1975.
23. R. Bradley, J. Farber, L. Tests, *Polym. Eng. Sci.* **7** (1977) no. 11, 782.
24. Polyplastics Co., EP 0 054 392, 1981 (M. Matsuoka, Y. Ikenaga, O. Matsuno).
25. B. J. Sutker: "Factors Affecting Smoke Production During Combustion of Polyolefin Polymers," Standard Technical Publication 816, ASTM, Philadelphia, 1983, p. 78.
26. A. L. Bridgman, G. L. Nelson, *J. of Fire Flammability* **13** (1982) 114.
27. "Chemicals and Additives, Special Report," *Mod. Plast.* 1985, (Sept.) 59.
28. A. Brief, T. Chastain, C. Cipriani, S. Epstein et al.: "Fire Retardant Plastics," a multiple client study by Skeist Laboratories, Livingston, N.J., 1985.
29. T. Chastain, S. Epstein, M. Kusumgar, G. F. Lillquist et al.: "Electronic Applications for Polymers, II," a multiple client study by Skeist Laboratories, Livingston, N.J., 1982.
30. "A Market, Technical, and Opportunity Study, Flame Retardant Plastics in the Construction Industry," TPC Business Research Group, Westport, Conn., 1981.
31. G. C. Tesoro, *Polym.-Plast. Technol. Eng.* **18** (1982) no. 2, 123.
32. H. L. Kaplan, A. F. Grand, G. E. Hartzell: "A Critical Review of the State-of-the-Art of Combustion Toxicology, final report, SwRI project no. 01-6862, The Society of the Plastics Industry, New York 1982.
33. J. O. Punderson, *Fire Mater.* **5** (1981) no. 1, 41.
34. S. C. Packam, G. E. Hartzell, *J. Test. Eval. (ASTM)* **9** (1981) 341.
35. Y. Alarie, M. F. Stock, M. Matijak-Schaper, M. M. Birky, *Fundam. Appl. Toxicol.* **3** (1983) 619.
36. G. K. Shanga, M. Matijak, Y. Alarie, *J. Am. Ind. Hyg. Assoc.* **42** (1981) 481.

Flavors and Fragrances

Karl-Georg Fahlbusch, Corporate Research Division, DRAGOCO Gerberding & Co. Aktiengesellschaft, Holzminden, Federal Republic of Germany

Franz-Josef Hammerschmidt, Corporate Research Division, DRAGOCO Gerberding & Co. Aktiengesellschaft, Holzminden, Federal Republic of Germany

Johannes Panten, Corporate Research Division, DRAGOCO Gerberding & Co. Aktiengesellschaft, Holzminden, Federal Republic of Germany

Wilhelm Pickenhagen, Corporate Research Division, DRAGOCO Gerberding & Co. Aktiengesellschaft, Holzminden, Federal Republic of Germany

Dietmar Schatkowski, Corporate Research Division, DRAGOCO Gerberding & Co. Aktiengesellschaft, Holzminden, Federal Republic of Germany

based on the article of:

Kurt Bauer, Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany

Dorothea Garbe, Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany

Horst Surburg, Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany (Chap. 3)

1. Introduction 4
1.1. The Chemical Senses 4
1.2. Definition 4
1.3. History 4
1.4. Odor Descriptors, Odor Thresholds, Odor Value 5
1.4.1. Qualitative Measurements 5
1.4.2. Quantitative Measurements 5
1.4.2.1. Odor Threshold 5
1.4.2.2. Odor Value 5
1.5. Regulations and Labeling Requirements 6
1.5.1. Flavors 6
1.5.2. Fragrances 6
2. Single Fragrance and Flavor Compounds 6
2.1. Aliphatic Compounds 7
2.1.1. Hydrocarbons 7
2.1.2. Alcohols 7
2.1.3. Aldehydes and Acetals 8
2.1.4. Ketones 11
2.1.5. Acids and Esters 11
2.1.6. Miscellaneous Compounds 14
2.2. Acyclic Terpenes 14
2.2.1. Hydrocarbons 14
2.2.2. Alcohols 14
2.2.3. Aldehydes and Acetals 21
2.2.4. Ketones 24
2.2.5. Acids and Esters 25
2.2.5.1. Geranyl and Neryl Esters 25
2.2.5.2. Linalyl and Lavandulyl Esters . . . 26
2.2.5.3. Citronellyl Esters 27
2.2.6. Miscellaneous Compounds 27
2.3. Cyclic Terpenes 28
2.3.1. Hydrocarbons 28
2.3.2. Alcohols and Ethers 30
2.3.3. Aldehydes and Ketones 34
2.3.4. Esters 40
2.3.5. Miscellaneous Compounds 42
2.4. Other Cycloaliphatic Compounds 42
2.4.1. Alcohols 43
2.4.2. Aldehydes 44
2.4.3. Ketones 45
2.4.4. Esters 48
2.5. Aromatic Compounds 50
2.5.1. Hydrocarbons 50
2.5.2. Alcohols and Ethers 50
2.5.3. Aldehydes and Acetals 53
2.5.4. Ketones 56
2.5.5. Esters of Araliphatic Alcohols and Aliphatic Acids 59
2.5.6. Aromatic Acids 60
2.5.7. Esters Derived from Aromatic and Araliphatic Acids 60
2.5.7.1. Benzoates 60
2.5.7.2. Phenyl acetates 61
2.5.7.3. Cinnamates 61
2.5.8. Miscellaneous Compounds 62
2.6. Phenols and Phenol Derivatives . 63
2.6.1. Phenols, Phenyl Esters, and Phenyl Ethers 63
2.6.2. Phenol Alcohols and their Esters . . 67

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_141

2.6.3. Phenol Aldehydes 67
2.6.4. Phenol Ketones 70
2.6.5. Phenol Carboxylates 70
2.7. O- and O, S-Heterocycles 72
2.7.1. Cyclic Ethers 72
2.7.2. Lactones 76
2.7.3. Glycidates 81
2.7.4. Miscellaneous Compounds 81
2.8. N- and N, S-Heterocycles 81
3. Natural Raw Materials in the Flavor and Fragrance Industry 83
3.1. Introduction 83
3.2. Isolation of Natural Fragrance and Flavor Concentrates 83
3.2.1. Essential Oils 84
3.2.2. Extracts 84
3.3. Survey of Natural Raw Materials 85
4. Quality Control 121
5. Economic Aspects 122
6. Toxicology and Environmental Aspects 122
7. References 123

1. Introduction

1.1. The Chemical Senses

The sensations "taste" and "odor" are part of our five senses. Whereas seeing, hearing, and feeling react to a physical input (visible light waves, sound waves, and temperature and pressure, respectively) the two so-called chemical senses are stimulated by direct interactions with defined chemicals. The chemical senses are regarded phylogenetically as the oldest senses, i.e., the earliest living species communicated by emitting and receiving chemical signals in their aqueous environment. This communication system was refined when the first species left water to live on land and had to rely on molecules that were transported in the surrounding atmosphere. These molecules had to be volatile and were detected by the odor receptors which consequently developed to a much higher sophistication than the taste receptors. The currently accepted doctrine is that only four basic taste impressions (salty, sweet, sour, and bitter [1]) exist, whereas the types of odor impressions are unlimited. Newer results, however, indicate that the limitation to only four taste impressions is too narrow and that more complex stimulations to the sense of taste do exist [2].

According to E. L. Wick, the term flavor as sensory impression results from the simultaneous stimulation of the two chemical senses, odor and taste [3], so the description of volatile materials will cover both flavor and fragrance materials. As mentioned above, the sense of odor developed to a much higher degree than the sense of taste and in some highly developed mammals (e.g., all canine species) became the most important sense of all. Even in humans, where the sense of sight has become predominant, odor influences considerably the behavior, consciously and subconsciously, it controls for example intake of food and emotions and according to Proust helps in the "search for the past".

1.2. Definition

Flavor and fragrance compounds are molecules that stimulate the human chemical senses. The physiological importance of compounds that elicit a sensory reaction in other living beings apart from humans will not be treated in the context of this paper.

In the following only compounds that stimulate the sense of smell for flavors and fragrances are discussed. These compounds, that are detected in the upper end of the nose have to have a certain volatility to be transported with the aspired air stream to the receptors. It is generally assumed that compounds with a molecular mass > 300 are not volatile enough to fulfill these conditions. Even if many compounds are used both as flavor and fragrance ingredients, some distinctions exist: in general, flavor compounds consist of molecules that occur naturally in edible materials whereas many fragrance compounds are not produced in nature (see Section 1.5).

1.3. History

The history of odor is as old as the history of mankind. Incense was already offered to the God of the Sea in Ninive, in 5000 B.C. The Gilgamesch Epos, which is considered to be the oldest written document about the genesis

of mankind, mentions the use of cedarwood as source of odorants already in 3200 B.C. Sandalwood is described as part of the gifts that the Queen of Sheba gave to Salomon. Greeks and Romans used oil extracts of herbs and flowers as therapeutic agents and for their sensory pleasures. During the middle ages, no real progress in preparing and using odorant materials was reported, apart from the discovery of making concentrated alcohol by the Arabs [4]. Real progress in preparation of odorant compounds and mixtures started nearly simultaneously with the beginning of organic chemistry which is marked by WÖHLER's synthesis of urea [5].

Already in 1830, WÖHLER and LIEBIG synthesized benzaldehyde, which was subsequently identified as the main aroma compound in bitter almonds by ROBIQUET and BOUTRON-CHARLARD in 1932. Worthy of mention is the synthesis of vanillin by REIMER and THIEMANN in 1876, which led to the foundation of the German flavor and fragrance industry.

Analysis of food flavors proved to be cumbersome due to the very low natural abundance of these molecules and the analytical methods available at that time. By 1960, about 450 molecules had been identified to occur in the volatile part of edible materials.

The real breakthrough in knowledge about the chemical nature of odorants started with the invention and application of gas chromatography by JAMES and MARTIN [6]. This analytical and also preparative separation method is the technique of choice for volatile materials such as odorants. The value of gas chromatography was quickly recognized and this technique has undergone enormous progress, e.g., through the introduction of capillary columns [7] and coupling of this separation tool to identification systems such as mass spectrometry and infrared spectroscopy. Extensive use of these and other analytical procedures have since allowed the identification of about 8000 volatile compounds in food materials [8]. Parallel to investigations into foods this method was also used to analyze many odorant extracts from materials such as flowers, leaves, and other natural sources used historically as fragrance materials. Development in synthetic procedures allowed economical syntheses of these identified compounds which can be used as such or as molecular models (lead compounds) for the synthesis of other not naturally occurring odorants.

1.4. Odor Descriptors, Odor Thresholds, Odor Value

Odors can be measured and described qualitatively and quantitatively, but contrary to the stimuli that elicit our other senses (i.e., sight, hearing, touch, temperature) no physical analytical method exist for odor measurement.

1.4.1. Qualitative Measurements

Qualitative measurements are restricted to verbal descriptions of the odor impression. First approaches to classify odors by their descriptors were initiated by HARPER, BATE SMITH, and LAND [9]. Their general conclusion is that the more complex the odor perception the more impossible becomes an exact odor description. A more detailed analysis into description of odors and thus their qualitative measurement has been initiated by KASTNER [10], who distinguishes three main descriptor categories, i.e., general descriptive terms, special descriptive terms, and precise descriptive terms, in which the first one refers to adjectives like attractive, good, bad, stinky, rough, etc.; the second category comprises explications like green, herbaceous, flowery, earthy; and the third one uses terms like sulfury, ambery, sweaty, etc.

1.4.2. Quantitative Measurements

1.4.2.1. Odor Threshold

Odor strength can be described by its odor threshold. Generally, two different thresholds are used: (1) the *detection threshold*, which is defined as the lowest detectable concentration of a compound versus the pure solvent and (2) the *recognition threshold*, which is the lowest concentration that allows an identification of the odor impression of a compound. Both values are strongly dependent on the solvent system [11].

Thresholds measured in air are generally much lower than those measured in water. A compilation of about 2000 thresholds in water

and in air document the enormous differences that can occur [12].

Thresholds in air can be measured with commercial olfactometers or by standardized smelling from a gas chromatograph as described in [13], whereas thresholds in water and other liquid solvents are determined by the so-called multiple pair test [14], in which the panel members have to decide which of the bottles contain the compound in question. The concentrations are generally presented in a geometrical line. The threshold is defined as the concentration that is correctly identified by 50 % of a panel consisting of a statistically relevant number of testers [15].

1.4.2.2. Odor Value

The quantitative contribution of a compound to the total odor impression can be described by its odor value. This term has been introduced by Rothe et al. [16] and is defined as the concentration of a compound in a given mixture divided by the odor threshold of this compound, i.e., the higher the concentration and the lower the threshold the more the compound contributes to the overall odor impression.

1.5. Regulations and Labeling Requirements

1.5.1. Flavors

Different legislative situations for definition and labeling of flavor active materials exist: The Americas base their systems on the United States Code of Federal Regulations, specifically CFR 101.22, which describes as *artificial flavor* or *artificial flavoring* any substance that imparts flavor, which is not derived from a spice, fruit or fruit juice, vegetable or vegetable juice, edible yeast, herb, bark, bud, root, leaf or similar plant material, meat, fish, poultry, eggs, dairy products, or fermentation products thereof [16]. *Natural flavor* or *natural flavoring* means the essential oil, oleoresin, essence or extractive, protein hydrolysate, distillate, or any product of roasting, heating or enzymolysis, which contains the flavoring constituents derived from a spice, fruit or fruit juice, vegetable or vegetable juice, edible yeast, herb, bark, bud, root, leaf or similar plant material, meat, seafood, poultry, eggs, dairy products, or fermentation products thereof, whose significant function in food is flavoring rather than nutritional [17]. In Europe, these regulations are defined in the "EU Flavouring Directive 88/388 EEC". The term flavoring is restricted to flavoring substances, flavoring preparations, process flavorings, smoke flavorings, or mixtures thereof. A flavoring substance is a defined chemical substance with flavoring properties which is obtained: (i) by appropriate physical processes (including distillation and solvent extraction) or enzymatic or microbiological processes from material of vegetable or animal origin either in the raw state or after processing for human consumption by traditional food-preparation processes (including drying, torrefaction, and fermentation), (ii) by chemical synthesis or isolated by chemical processes and which is chemically identical to a substance naturally present in material of vegetable or animal origin as described in (i), (iii) by chemical synthesis but which is not chemically identical to a substance naturally present in material of vegetable or animal origin. "Process flavoring" means a product which is obtained according to good manufacturing practices by heating to a temperature not exceeding 180 °C for a period not exceeding 15 min from a mixture of ingredients not necessarily themselves having flavoring properties in which at least one contains nitrogen (amino) and another is a reducing sugar. "Smoke flavoring" means a smoke extract used in traditional foodstuffs smoking processes [18].

1.5.2. Fragrances

For fragrances no differentiation is made between natural, nature-identical, and artificial substances. Many compounds used in fragrances are components of natural origin, e.g., constituents of essential oils, resins, and, to a much lesser extent, of animal secretions. In addition many chemicals are used in fragrances, whose molecular structures are not found to occur in nature but their fragrance characteristics mimic those of natural products. For use in fragrance components, these chemicals have to comply with the national legal requirements. In addition they

have to pass the so-called IFRA tests for (see Chap. 6):

Acute toxicity
Skin irritation
Skin sensitization
Phototoxicity and photosensitization

For those products (e.g., fragrance chemicals, essential oils, resinoids) for which IFRA has edited use recommendations this is mentioned by 'IFRA recommendation'. If IFRA suggests a maximum use level then one will find the remark 'limiting IFRA recommendation'. In those cases one may refer to the IFRA publications for further information.

2. Single Fragrance and Flavor Compounds

Fragrance and flavor compounds of commercial interest are arranged according to the Beilstein system of functional groups, not according to their organoleptic properties, since relationships between odor and structure are difficult to establish. However, the Beilstein system has been abandoned in a few cases for practical reasons.

In each class of parent compounds, hydrocarbons and oxygen-containing compounds are described first. Nitrogen- and sulfur-containing compounds are treated at the end of each of these sections under the heading Miscellaneous Compounds. Aliphatic compounds are discussed in Section 2.1, followed by the terpenes. The terpenes constitute a very important group of compounds and are subdivided into acyclic terpenes (Section 2.2) and cyclic terpenes (Section 2.3). Nonterpenoid cycloaliphatics are described in Section 2.4. Aromatic compounds are discussed in Section 2.5. Phenols and phenol derivatives are described under a separate heading (Section 2.6) on account of their biogenetic and odor relationships. Methylenedioxyphenyl derivatives are also described under this heading for the same reason even though, systematically, they belong to the oxygen-containing heterocycles (Section 2.7). Compounds that are only produced in small quantities, but which are important due to their high odor intensity, are mentioned but not described in detail.

2.1. Aliphatic Compounds

The acyclic terpenes are discussed separately in Section 2.2. Some of the cycloaliphatic fragrance and flavor compounds are structurally related to the cyclic terpenes and are, therefore, discussed in Section 2.4 after the cyclic terpenes.

2.1.1. Hydrocarbons

Saturated and unsaturated aliphatic hydrocarbons with straight as well as branched chains occur abundantly in natural foodstuffs, but they contribute to the odor and taste only to a limited extent. The highly unsaturated hydrocarbons 1,3-*trans*-5-*cis*-undecatriene [*51447-08-6*] and 1,3-*trans*-5-*trans*-undecatriene [*19883-29-5*], however, contribute to the odor of galbanum oil [19].

2.1.2. Alcohols (→ Alcohols, Aliphatic; → Fatty Alcohols)

Free and esterified saturated primary alcohols occur widely in nature, e.g., in fruit. Since their odor is relatively weak, their use as components in fragrance compositions is limited. Their use in aroma compositions, especially for fruit flavors, is by far more important (e.g., straight-chain $C_4 - C_{10}$ alcohols, isoamyl alcohol). Unsaturated alcohols are most important (e.g., leaf alcohol with its intensely green odor) and may impart characteristic notes to compositions.

Naturally occurring fatty alcohols used in the fragrance industry are produced principally by reduction of the methyl esters of the corresponding carboxylic acids, which are obtained by transesterification of natural fats and oils with methanol. Industrial reduction processes include catalytic hydrogenation in the presence of copper–chromium oxide catalysts (Adkins catalysts) and reduction with sodium (Bouveault–Blanc reduction); for details, see → Alcohols, Aliphatic, Chap. 2.3.3., → Fatty Alcohols, Chap. 2.3., → Fatty Alcohols, Chap. 2.4. Unsaturated alcohols can also be prepared by the latter method. Alcohols are starting materials for aldehydes and esters.

3-Octanol [*589-98-0*], $CH_3(CH_2)_4CH(OH)CH_2CH_3$, $C_8H_{18}O$, M_r 130.23, *bp* (97.6 kPa) 176 – 176.5 °C, d_4^{20} 0.8264, n_D^{20} 1.4252, may occur in its optically active form. It is a colorless liquid that has a mushroomy-earthy odor and occurs in mushrooms. 3-Octanol can be obtained by hydrogenation of 3-octanone; it is used in lavender compositions and for imparting mushroom-like odors.

2,6-Dimethyl-2-heptanol [*13254-34-7*], $C_9H_{20}O$, M_r 144.26, *bp* (101.3 kPa) 170 – 172 °C, d_4^{20} 0.8085, n_D^{20} 1.4248, which has not yet been found in nature, is a colorless liquid with a delicate, flowery odor reminiscent of fresias. It is synthesized from 2-methyl-2-hepten-6-one and methylmagnesium chloride by a Grignard reaction, followed by hydrogenation, and is used in flowery perfume compositions.

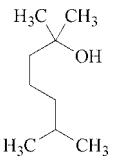

Trade Names. Dimetol (Givaudan Roure), Freesiol (Haarmann & Reimer).

***trans*-2-Hexen-1-ol** [*928-95-0*], $CH_3CH_2CH_2CH{=}CHCH_2OH$, $C_6H_{12}O$, M_r 100.16, *bp* (101.3 kPa) 155 °C, d_4^{20} 0.8459, n_D^{20} 1.4382, occurs in many fruits and has a fruity, green odor, which is sweeter than that of the isomeric *cis*-3-hexen-1-ol and is, therefore, preferred in aroma compositions.

***cis*-3-Hexen-1-ol** [*928-96-1*], leaf alcohol, $CH_3CH_2CH{=}CHCH_2CH_2OH$, $C_6H_{12}O$, M_r 100.16, *bp* (101.3 kPa) 156 – 157 °C, d_4^{20} 0.8495, n_D^{20} 1.4384, is a colorless liquid with the characteristic odor of freshly cut grass. *Robinia pseudacacia* and mulberry leaf oil contain up to 50 % leaf alcohol, and green tea up to 30 %. Small quantities occur in the green parts of nearly all plants.

A stereospecific synthesis for *cis*-3-hexen-1-ol starts with the ethylation of sodium acetylide to 1-butyne, which is reacted with ethylene oxide to give 3-hexyn-1-ol. Selective hydrogenation of the triple bond in the presence of palladium catalysts yields *cis*-3-hexen-1-ol. Leaf alcohol is used to obtain natural green top notes in perfumes and flavors. In addition, it is the starting material for the synthesis of 2-*trans*-6-*cis*-nonadien-1-ol and 2-*trans*-6-*cis*-nonadien-1-al.

1-Octen-3-ol [*3391-86-4*], $CH_3(CH_2)_4CH(OH)CH{=}CH_2$, $C_8H_{16}O$, M_r 128.21, *bp* (94.6 kPa) 175 – 175.2 °C, d_4^{20} 0.8383, n_D^{20} 1.4378, may occur in the optically active form. It is found, for example, in lavender oil and is a steam-volatile component of mushrooms. 1-Octen-3-ol is a liquid with an intense mushroom, forest-earthy odor that can be prepared by a Grignard reaction from vinylmagnesium bromide and hexanal. It is used in lavender compositions and in mushroom aromas.
Trade Name. Matsutake alcohol (Takasago).

9-Decen-1-ol [*13019-22-2*], $CH_2{=}CH(CH_2)_6CH_2OH$, $C_9H_{18}O$, M_r 142.24, *bp* (270 Pa) 85 – 86 °C, n_D^{20} 1.4480, has been identified as a trace constituent of cognac. It is a colorless liquid with a fresh, dewy, rose note that can be prepared by partial dehydration of 1,10-decanediol. It is used in rosy-floral soap perfumes.
Trade Name. Rosalva (IFF).

10-Undecen-1-ol [*112-43-6*], $CH_2{=}CH(CH_2)_8CH_2OH$, $C_{11}H_{22}O$, M_r 170.29, *bp* (2.1 kPa) 133 °C, d_4^{20} 0.8460, n_D^{20} 1.4500, has not yet been found in nature and is a colorless liquid with a fatty-green, slightly citrus-like odor. It can be synthesized from 10-undecylenic acid and is used to give flower perfumes a fresh note.

2-*trans*-6-*cis*-Nonadien-1-ol [*28069-72-9*], $CH_3CH_2CH{=}CHCH_2CH_2CH{=}CHCH_2OH$, $C_9H_{16}O$, violet leaf alcohol, M_r 140.22, *bp* (1.5 kPa) 96 – 100 °C, d_4^{10} 0.8622, n_D^{20} 1.4740, occurs, for example, in cucumber oil, violet leaf oil, and violet blossom oil. It is a colorless liquid with an intense, heavy-fatty, green odor, reminiscent of violet leaves. The starting material for the synthesis of 2-*trans*-6-*cis*-nonadien-1-ol is *cis*-3-hexen-1-ol, which is converted via its halide into the corresponding Grignard reagent. The Grignard reagent is reacted with acrolein to give 1,6-nonadien-3-ol, which is converted

into 2-*trans*-6-*cis*-nonadien-1-ol by allylic rearrangement.

Nonadienol is a powerful fragrance substance. It is used in fine fragrances to create refined violet odors and to impart interesting notes to other blossom compositions. In aroma compositions it is used for fresh-green cucumber notes.

2.1.3. Aldehydes and Acetals

(→ Aldehydes, Aliphatic and Araliphatic)

Aliphatic aldehydes are among the most important components used in perfumery. Although the lower fatty aldehydes $C_2 - C_7$ occur widely in nature, they are – with the exception of hexanal – seldom used in fragrance compositions. The lower aldehydes (e.g., acetaldehyde, isobutyraldehyde, isovaleraldehyde, and 2-methylbutyraldehyde) impart fruity and roast characters to flavor compositions. Fatty aldehydes C_8–C_{13}, however, are used, singly or in combination, in nearly all perfume types and also in aromas. Their odor becomes weaker with increasing molecular mass, so that aldehydes $> C_{13}$ are not important as perfume ingredients.

In addition to the straight-chain saturated aldehydes, a number of branched-chain and unsaturated aliphatic aldehydes are important as fragrance and flavoring materials. The double unsaturated 2-*trans*-6-*cis*-nonadienal [*557-48-2*], "violet leaf aldehyde" (the dominant component of cucumber aroma), is one of the most potent fragrance and flavoring substances; it is, therefore, only used in very small amounts.

Acetals derived from aliphatic aldehydes have odor characteristics that resemble those of the aldehydes but are less pronounced. These acetals contribute to the aroma of alcoholic beverages, but can rarely be used in flavoring compositions because they are not sufficiently stable. Since they are resistant to alkali, a number of them (e.g., heptanal dimethyl acetal and octanal dimethyl acetal) are occasionally incorporated into soap perfumes.

Fatty aldehydes are generally produced by dehydrogenation of alcohols in the presence of suitable catalysts (see → Aldehydes, Aliphatic and Araliphatic, Chap. 2.3.). The alcohols are often cheap and available in good purity. Aldehyde synthesis via the oxo process is less suitable since the resultant products are often not pure enough for flavor and perfume purposes. Specific syntheses for the branched-chain and unsaturated aldehydes that are important in perfumery and flavoring techniques are described under the individual compounds.

Hexanal [*66-25-1*], caproaldehyde, $CH_3(CH_2)_4CHO$, $C_6H_{12}O$, M_r 100.16, *bp* (101.3 kPa) 128 °C, d_4^{20} 0.8139, n_D^{20} 1.4039, occurs, for example, in apple and strawberry aromas as well as in orange and lemon oil. It is a colorless liquid with a fatty-green odor and in low concentration is reminiscent of unripe fruit.

Hexanal is used in fruit flavors and, when highly diluted, in perfumery for obtaining fruity notes.

Octanal [*124-13-0*], caprylaldehyde, $CH_3(CH_2)_6CHO$, $C_8H_{16}O$, M_r 128.21, *bp* (101.3 kPa) 171 °C, d_4^{20} 0.8211, n_D^{20} 1.4217, occurs in several citrus oils, e.g., orange oil. It is a colorless liquid with a pungent odor, which becomes citrus-like on dilution. Octanal is used in perfumery in low concentrations, in eaux de cologne, and in artificial citrus oils.

Nonanal [*124-19-6*], pelargonaldehyde, $CH_3(CH_2)_7CHO$, $C_9H_{18}O$, M_r 142.24, *bp* (101.3 kPa) 190 – 192 °C, d_4^{20} 0.8264, n_D^{20} 1.4273, occurs in citrus and rose oils. It is a colorless liquid with a fatty-roselike odor and is used in floral compositions, particularly those with rose characteristics.

Decanal [*112-31-2*], capraldehyde, caprinaldehyde, $CH_3(CH_2)_8CHO$, $C_{10}H_{20}O$, M_r 156.27, *bp* (101.3 kPa) 208 – 209 °C, d_4^{20} 0.8259, n_D^{20} 1.4287, is a component of many essential oils (e.g., neroli oil) and various citrus peel oils. It is a colorless liquid with a strong odor, reminiscent of orange peel, that changes to a fresh citrus odor when diluted. Decanal is used in low concentrations in blossom fragrances (especially to create citrus nuances) and in the production of artificial citrus oils.

Undecanal [*112-44-7*], $CH_3(CH_2)_9CHO$, $C_{11}H_{22}O$, M_r 170.29, *bp* (2.4 kPa) 117 °C, d_4^{20} 0.8250, n_D^{20} 1.4325, occurs in citrus oils. It is a colorless liquid with a flowery-waxy odor that

has aspects of freshness. Undecanal is the prototype of the perfumery aldehydes and is widely used in perfume compositions for imparting an "aldehydic note".

Dodecanal [*112-54-9*], lauraldehyde, lauric aldehyde, $CH_3(CH_2)_{10}CHO$, $C_{12}H_{24}O$, M_r 184.32, *bp* (13.3 kPa) 185 °C, d_4^{20} 0.8350, n_D^{20} 1.4350, is a colorless liquid with a waxy odor; in high dilution it is reminiscent of violets. Dodecanal occurs in several citrus oils and has been found in small amounts in essential oils obtained from several *Pinus* species. It is used in perfumery in conifer fragrances with fatty-waxy notes, but also in many other odor types. It is added to aroma compositions to obtain citrus notes.

Tridecanal [*10486-19-8*], $CH_3(CH_2)_{11}CHO$, $C_{13}H_{26}O$, M_r 198.34, *bp* (1.3 kPa) 128 °C, d_4^{20} 0.8358, n_D^{18} 1.4384, occurs in lemon oil and has been identified as a volatile constituent of cucumber. It is a colorless liquid having a fatty-waxy, slightly citrus-like odor. Addition of tridecanal to fragrance compositions imparts fresh nuances in the top note as well as in the dry out.

2-Methyldecanal [*19009-56-4*], methyloctylacetaldehyde, $CH_3(CH_2)_7CH(CH_3)CHO$, $C_{11}H_{22}O$, M_r 170.29, *bp* (98.8 kPa) 119–120 °C, d_4^{20} 0.8946, n_D^{20} 1.4205, is not reported to have been found in nature. It is a colorless liquid with an aldehydic, citrus-peel-like, waxy-green odor. 2-Methyldecanal is obtained as a byproduct in the manufacture of 2-methylundecanal by hydroformylation of 1-decene. It is used in perfumery to refresh green and citrus nuances.

2-Methylundecanal [*110-41-8*], methylnonylacetaldehyde, $CH_3(CH_2)_8CH(CH_3)CHO$, $C_{12}H_{24}O$, M_r 184.32, *bp* (1.3 kPa) 114 °C, d_4^{20} 0.8303, n_D^{20} 1.4321, has not yet been found in nature. It is a colorless liquid, with an odor markedly different from that of the isomeric dodecanal. It has a fatty odor with incense and ambergris notes.

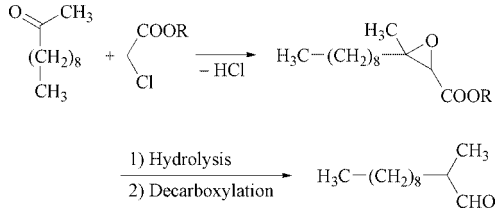

2-Methylundecanal is produced by two routes:

1) Methyl nonyl ketone is converted into its glycidate by reaction with an alkyl chloroacetate. Saponification of the glycidate, followed by decarboxylation, yields 2-methylundecanal.
2) The second synthesis is based on the conversion of undecanal into 2-methyleneundecanal by reaction with formaldehyde in the presence of catalytic amounts of amines [20]. Hydrogenation of 2-methyleneundecanal yields methylnonylacetaldehyde. A convenient process starts from 1-decene: hydroformylation gives a mixture consisting mainly of undecanal and 2-methyldecanal. Reaction of the crude product with formaldehyde in the presence of dibutylamine yields a mixture containing over 50 % 2-methyleneundecanal. After hydrogenation of the double bond, pure 2-methylundecanal is separated from byproducts by fractional distillation [21].

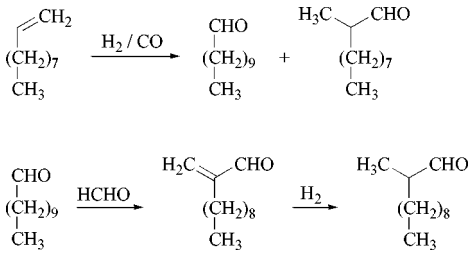

In comparison with other fatty aldehydes, 2-methylundecanal is used in perfumery in rather large amounts to impart conifer notes, particularly fir impressions, but frequently also in phantasy compositions.

***trans*-2-Hexenal** [*6728-26-3*], leaf aldehyde, $CH_3CH_2CH_2CH{=}CHCHO$, $C_6H_{10}O$, M_r 98.14, *bp* (101.3 kPa) 146–147 °C, d_4^{20} 0.8491, n_D^{20} 1.4480, is the simplest straight-chain unsaturated aldehyde of interest for perfumes and

flavors. It occurs in essential oils obtained from green leaves of many plants. *trans*-2-Hexenal is a colorless, sharp, herbal-green smelling liquid with a slight acrolein-like pungency. Upon dilution, however, it smells pleasantly green and apple-like. The aldehyde can be synthesized by reacting butanal with vinyl ethyl ether in the presence of boron trifluoride, followed by hydrolysis of the reaction product with dilute sulfuric acid [22].

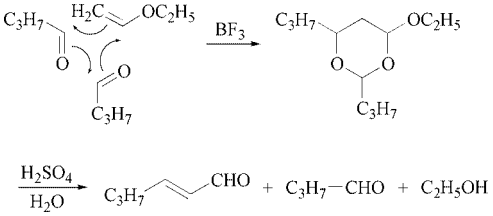

trans-2-Hexenal has an intense odor and is used in perfumes to obtain a green-leaf note, and in fruit flavors for green nuances (limiting IFRA recommendation). (See Sec. 1.5.2 Fragrances).

***cis*-4-Heptenal** [*6728-31-0*], $CH_3CH_2CH{=}CHCH_2CH_2CHO$, $C_7H_{12}O$, M_r 112.17, *bp* (1.33 kPa) 41 °C, n_D^{20} 1.4343, is a widespread volatile trace constituent of food flavors. It is a colorless, oily liquid with a powerful, fatty, somewhat fishy and, in high dilution, creamy odor. It can be prepared from 1-butyne (via lithium 1-butynide) and acrolein (which is converted into 2-bromopropionaldehyde dimethyl acetal). The resulting 4-heptynal dimethyl acetal is cleaved and the triple bond is hydrogenated catalytically to give *cis*-4-heptenal [23].

cis-4-Heptenal is used in cream, butter, and fat flavors.

10-Undecenal [*112-45-8*], $CH_2{=}CH(CH_2)_8CHO$, $C_{11}H_{20}O$, M_r 168.28, *bp* (400 Pa) 103 °C, d_4^{20} 0.8495, n_D^{20} 1.4459, has not yet been found in nature. It is a colorless liquid with a fatty-green, slightly metallic, heavy-flowery odor. The aldehyde can be synthesized from undecylenic acid, for example, by hydrogenation of the acid chloride (Rosenmund reduction) or by reaction with formic acid in the vapor phase in the presence of titanium dioxide. In perfumery, 10-undecenal is one of the aldehydes essential for creating the "aldehydic note".

2.1.4. Ketones (→ Ketones)

Aliphatic monoketones are of minor importance as fragrance and aroma substances. 2-Alkanones (C_3-C_{15}) have been found in the volatile fractions of many fruits and foods, but they do not contribute significantly to their aroma. An exception are the odd-numbered methyl ketones C_7, C_9, C_{11}, which possess a characteristic nutty note; they are used, e.g., in cheese flavor compositions. In perfumery, aliphatic ketones are used for accentuation, e.g., 3-octanone [*106-68-3*] for lavender notes. The hydroxyketone acetoin and the diketone 2,3-butanedione are commercially important aroma substances.

Acetoin [*52217-02-4*], 3-hydroxy-2-butanone, $CH_3COCH(OH)CH_3$, $C_4H_8O_2$, M_r 88.11, *bp* (101.3 kPa) 148 °C, d_4^{20} 1.0058, n_D^{20} 1.4171, has a pleasant buttery odor and both of its optical isomers occur widely in nature. It is synthesized by partial oxidation of 2,3-butanediol and is obtained as a byproduct in the fermentation of molasses. It is used for flavoring margarine.

2,3-Butanedione [*431-03-8*], diacetyl, $CH_3COCOCH_3$, $C_4H_6O_2$, M_r 86.09, *bp* 88 °C (101.3 kPa), d_4^{20} 0.9831, n_D^{20} 1.3950, is a constituent of many fruit and food aromas and well-known as a constituent of butter. Many methods are known for its manufacture, e.g., dehydrogenation of 2,3-butanediol with a copper chromite catalyst [24]. It is used mainly in aromas for butter and roast notes. Large quantities are used for flavoring margarine; small amounts are used in perfumes.

2.1.5. Acids and Esters (→ Carboxylic Acids, Aliphatic; → Esters Organic)

Straight-chain, saturated aliphatic acids are found in many essential oils and foods. These acids contribute to aromas, but are not important as fragrance substances. In flavor compositions, aliphatic acids up to C_{10} are used to accentuate certain aroma characteristics (C_3-C_8 for fruity notes; C_4, C_6-C_{12} for cheese flavors). However, straight-chain and some branched-chain

aliphatic acids are of considerable importance as starting materials in the manufacture of esters, many of which are valuable fragrance and flavor materials. Aliphatic esters contribute to the aroma of nearly all fruits and many foods. Some are responsible for a particular fruit aroma, or for the smell of a particular flower; however, many of these esters possess a nonspecific fruity odor.

Most of the esters used are acetates and ethanol is the most common alcohol component. In nature, most esters are derived from alcohols and acids with an even number of carbon atoms. In addition to straight-chain saturated compounds, branched-chain compounds such as isoamyl esters and unsaturated compounds such as hexenyl esters are important.

Although the odor of aliphatic esters with a small number of carbon atoms is strictly fruity, it changes to fatty-soapy and even metallic as the number of carbon atoms increases.

Esters are usually prepared by esterification of carboxylic acids with alcohols. Industrial procedures depend on the physical properties of the esters concerned (see → Esters, Organic).

In perfumery, acetates are the most important aliphatic esters; formates do not keep well. Animal and fatty notes become more pronounced in esters of higher fatty acids. Acetates of alcohols up to C_6 are used principally for fruity notes, whereas the acetates of C_8, C_{10}, and C_{12} alcohols are employed for blossom fragrances and for flower notes in general. Lauryl acetate in particular is also used for conifer notes.

In flavor compositions, aliphatic esters are preferred for artificial fruit aromas; as in nature, acetates and ethyl esters prevail.

Ethyl formate [*109-94-4*], $HCOOCH_2CH_3$, $C_3H_6O_2$, M_r 74.08, *bp* (101.3 kPa) 54.5 °C, d_4^{20} 0.9168, n_D^{20} 1.3598, occurs widely in fruits. It is a liquid with a slightly pungent, fruity, ethereal odor and is used in fruit flavors.

***cis*-3-Hexenyl formate** [*33467-73-1*], $HCOO(CH_2)_2CH{=}CHCH_2CH_3$, $C_7H_{12}O_2$, M_r 128.17, *bp* (101.3 kPa) 155 °C, d_4^{20} 0.9121, n_D^{20} 1.4270, has been identified in tea. It possesses a green-fruity odor and is used in perfumery and flavor compositions to impart fruity green notes.

Ethyl acetate [*141-78-6*], $CH_3COOCH_2CH_3$, $C_4H_8O_2$, M_r 88.11, *bp* (101.3 kPa) 77.1 °C, d_4^{20} 0.9003, n_D^{20} 1.3723, is a fruity smelling liquid with a brandy note and is the most common ester in fruits. It is used in fruit and brandy flavors.

Butyl acetate [*123-86-4*], $CH_3COO(CH_2)_3CH_3$, $C_6H_{12}O_2$, M_r 116.16, *bp* (101.3 kPa) 126.5 °C, d_4^{20} 0.9003, n_D^{20} 1.3942, is a liquid with a strong fruity odor. It occurs in many fruits and is a constituent of apple aromas.

Isoamyl acetate [*123-92-2*], $CH_3COO(CH_2)_2CH(CH_3)_2$, $C_7H_{14}O_2$, M_r 130.19, *bp* (101.3 kPa) 142.5 °C, d_4^{20} 0.8730, n_D^{20} 1.4006, is a strongly fruity smelling liquid and has been identified in many fruit aromas. It is the main component of banana aroma and is, therefore, also used in banana flavors.

Hexyl acetate [*142-92-7*], $CH_3COO(CH_2)_5CH_3$, $C_8H_{16}O_2$, M_r 144.21, *bp* (101.3 kPa) 171.5 °C, d_4^{20} 0.8730, n_D^{20} 1.4092, is a liquid with a sweet-fruity, pearlike odor. It is present in a number of fruits and alcoholic beverages, and is used in fruit aroma compositions.

3,5,5-Trimethylhexyl acetate [*58430-94-7*], $CH_3COO(CH_2)_2CH(CH_3)CH_2C(CH_3)_3$, isononyl acetate, $C_{11}H_{22}O_2$, M_r 186.29, does not occur in nature. Commercial isononyl acetate contains small amounts of byproducts. It is a colorless liquid with a woody-fruity odor and is prepared from diisobutene by the oxo synthesis, followed by hydrogenation to the alcohol and acetylation. It is used in household perfumery.

***trans*-2-Hexenyl acetate** [*2497-18-9*], $CH_3COOCH_2CH{=}CH(CH_2)_2CH_3$, $C_8H_{14}O_2$, M_r 142.20, *bp* (2.1 kPa) 67 – 68 °C, d_4^{20} 0.8975, n_D^{20} 1.4277, occurs in many fruits and in some essential oils, e.g., peppermint. It is a fresh-fruity, slightly green smelling liquid and is used in fruit flavors.

***cis*-3-Hexenyl acetate** [*3681-71-8*], $CH_3COO(CH_2)_2CH{=}CHCH_2CH_3$, $C_8H_{14}O_2$, M_r 142.20, *bp* (1.6 kPa) 66 °C, has been identified in many fruit aromas and green tea. It is a prototype for green odors and is often used in combination with *cis*-3-hexenol.

Ethyl propionate [*105-37-3*], $CH_3CH_2COOCH_2CH_3$, $C_5H_{10}O_2$, M_r 102.13, *bp* (101.3 kPa) 99 °C, d_4^{20} 0.8896, n_D^{20} 1.3839, is found in many fruits and alcoholic beverages. It has a fruity odor reminiscent of rum and is used in flavor compositions for creating both fruity and rum notes.

Ethyl butyrate [*105-54-4*], $CH_3(CH_2)_2COOCH_2CH_3$, $C_6H_{12}O_2$, M_r 116.16, *bp* (101.3 kPa) 121 – 122 °C, d_4^{20} 0.8785, n_D^{20} 1.4000, occurs in fruits and alcoholic beverages, but also in other foods such as cheese. It has a fruity odor, reminiscent of pineapples. Large amounts are used in perfume and in flavor compositions.

Butyl butyrate [*109-21-7*], $CH_3(CH_2)_2COOCH_2(CH_2)_2CH_3$, $C_8H_{16}O_2$, M_r 144.21, *bp* (101.3 kPa) 166 °C, d_4^{20} 0.8709, n_D^{20} 1.4075, is a liquid with a sweet-fruity odor. It is a volatile constituent of many fruits and honey and is used in fruit flavor compositions.

Isoamyl butyrate [*106-27-4*], $CH_3(CH_2)_2COO(CH_2)_2CH(CH_3)_2$, $C_9H_{18}O_2$, M_r 158.23, *bp* (101.3 kPa) 178.5 °C, d_4^{20} 0.8651, n_D^{20} 1.4106, is a liquid with strongly fruity odor that occurs, e.g., in banana. It is used mainly in fruit flavors.

Hexyl butyrate [*2639-63-6*], $CH_3(CH_2)_2COO(CH_2)_5CH_3$, $C_{10}H_{20}O_2$, M_r 172.27, *bp* (101.3 kPa) 208 °C, d_4^{20} 0.8652, n_D^{20} 1.3920, is a liquid with a powerful fruity odor. It has been identified in a number of fruits and berries and is an important constituent of fruit flavor compositions.

***cis*-3-Hexenyl isobutyrate** [*41519-23-7*], $(CH_3)_2CHCOO(CH_2)_2CH{=}CHCH_2CH_3$, $C_{10}H_{18}O_2$, M_r 170.25 is found in spearmint oil. It smells fruity-green and is used in perfumery to create freshness in blossom compositions.

Ethyl isovalerate [*108-64-5*], $(CH_3)_2CHCH_2COOCH_2CH_3$, $C_7H_{14}O_2$, M_r 130.19, *bp* (101.3 kPa) 134.7 °C, d_4^{20} 0.8656, n_D^{20} 1.3962, is a colorless liquid with a fruity odor reminiscent of blueberries. It occurs in fruits, vegetables, and alcoholic beverages. It is used in fruity aroma compositions.

Ethyl 2-methylbutyrate [*7452-79-1*], $CH_3CH_2CH(CH_3)COOCH_2CH_3$, $C_7H_{14}O_2$, M_r 130.19, *bp* (101.3 kPa) 131 – 132 °C, d_4^{25} 0.8689, n_D^{20} 1.3964, is a liquid with a green-fruity odor reminiscent of apples. It is found, for example, in citrus fruits and wild berries and is used in fruit flavor compositions.

Ethyl hexanoate [*123-66-0*], ethyl caproate, $CH_3(CH_2)_4COOCH_2CH_3$, $C_8H_{16}O_2$, M_r 144.21, *bp* (101.3 kPa) 168 °C, d_4^{20} 0.8710, n_D^{20} 1.4073, is a colorless liquid with a strong fruity odor, reminiscent of pineapples. It occurs in many fruits and is used in small amounts for flowery-fruity notes in perfume compositions and in larger quantities in fruit flavors.

2-Propenyl hexanoate [*123-68-2*], allyl caproate, $CH_3(CH_2)_4COOCH_2CH{=}CH_2$, $C_9H_{16}O_2$, M_r 156.22, *bp* (2 kPa) 75 – 76 °C, d_4^{20} 0.8869, n_D^{20} 1.4243, and also to occurs in nature. It has a typical pineapple odor and is used in, for example, artificial pineapple flavors. (Limiting IFRA recommendation) (see Sec. 1.5.2 Fragrances).

Ethyl heptanoate [*106-30-9*], ethyl enanthate, $CH_3(CH_2)_5COOCH_2CH_3$, $C_9H_{18}O_2$, M_r 158.24, *bp* (101.3 kPa) 187 – 188 °C, d_4^{20} 0.8680, n_D^{20} 1.4120, is a colorless liquid with a fruity odor reminiscent of cognac. It is found in fruits and alcoholic beverages and is used in appropriate aroma compositions.

2-Propenyl heptanoate [*142-19-8*], allyl enanthate, $CH_3(CH_2)_5COOCH_2CH{=}CH_2$, $C_{10}H_{18}O_2$, M_r 170.25, *bp* (101.3 kPa) 210 °C, d_4^{20} 0.8927, n_D^{20} 1.4290, has not yet been found in nature. It is used in perfume compositions for apple-like (pineapple) notes. (Llimiting IFRA recommendation) (see Sec. 1.5.2 Fragrances).

Ethyl octanoate [*106-32-1*], ethyl caprylate, $CH_3(CH_2)_6COOCH_2CH_3$, $C_{10}H_{20}O_2$, M_r 172.27, *bp* (101.3 kPa) 208 °C, d_4^{20} 0.8693, n_D^{20} 1.4178, is a liquid with a fruity-flowery odor. It occurs in many fruits and alcoholic beverages and is used in fruit flavors.

Ethyl 2-*trans*-4-*cis*-decadienoate [*3025-30-7*], $CH_3(CH_2)_4CH{=}CHCH{=}CHCOOCH_2CH_3$, $C_{12}H_{20}O_2$, M_r 196.29, *bp* (6 kPa) 70–72 °C, has been identified in pears and has the typical aroma of Williams pears. Synthesis of ethyl 2-*trans*-4-*cis*-decadienoate starts from *cis*-1-heptenyl bromide, which is converted into a 1-heptenyllithium cuprate complex with lithium and copper iodide. Reaction with ethyl propiolate yields a mixture of 95 % ethyl 2-*trans*-4-*cis*- and 5 % ethyl 2-*trans*-4-*trans*-decadienoate. Pure ethyl 2-*trans*-4-*cis*-decadienoate is obtained by fractional distillation [25].

Methyl 2-octynoate [*111-12-6*], $CH_3(CH_2)_4C{\equiv}CCOOCH_3$, $C_9H_{14}O_2$, M_r 154.21, *bp* (1.3 kPa) 94 °C, d_4^{20} 0.9320, n_D^{20} 1.4464. Limiting IFRA recommendation.

Methyl 2-nonynoate [*111-80-8*], $CH_3(CH_2)_5C{\equiv}CCOOCH_3$, $C_{10}H_{16}O_2$, M_r 168.24, *bp* (2.7 kPa) 121 °C, d_4^{20} 0.9162, n_D^{25} 1.4395. Both methyl 2-nonynoate and methyl 2-octynoate have a triple bond and are liquids with a fatty, violet-leaf-like odor. They are used in perfume compositions. Limiting IFRA recommendation.

2.1.6. Miscellaneous Compounds

A number of volatile aliphatic compounds that contain nitrogen or sulfur atoms are important aroma constituents. Alkyl thiols, dialkyl sulfides and disulfides, and alkyl thiocyanates belong to this group. They occur widely in foods and spices and determine the odor of, for example, onions, garlic, and mustard. Because of their potent smell, they are used in high dilution and are often produced only in small quantities. The same is true for 3-methylthiohexanol [*51755-66-9*], $CH_3(CH_2)_2CH(SCH_3)CH_2CH_2OH$, a volatile constituent of passion fruits. Allyl isothiocyanate, however, is an exception in that it is produced in large quantities.

Allyl isothiocyanate [*57-06-7*], allyl mustard oil, $CH_2{=}CHCH_2N{=}C{=}S$, C_4H_5NS, M_r 99.14, *bp* (101.3 kPa) 152 °C, d_4^{20} 1.0126, is the main component of mustard oil (> 95 %). It is a colorless oil with a typical mustard odor and can be prepared by reacting allyl chloride with alkaline-earth or alkali isothiocyanates [26].

2.2. Acyclic Terpenes

For a general description of terpenes, see → Terpenes.

2.2.1. Hydrocarbons

Acyclic terpene (C_{10}) and sesquiterpene (C_{15}) hydrocarbons find little use in flavor and fragrance compositions. They are relatively unstable and some have a slightly aggressive odor due to their highly unsaturated structure. Myrcene, ocimene, and farnesene, are present in many fruits and essential oils, but find only limited use in perfumery.

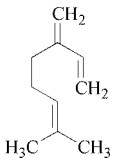

Myrcene

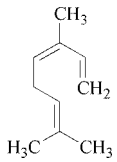

Ocimene

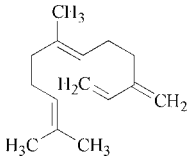

β-Farnesene

2.2.2. Alcohols

Acyclic terpene and sesquiterpene alcohols occur in many essential oils. These alcohols were formerly isolated from oils in which they are major components. Currently, large-scale synthesis of terpenoids permits production without the uncertainties associated with isolation from natural sources. However, the odor qualities of synthetic products often differ from those of compounds isolated from natural sources, since the desired natural product often cannot be separated from small amounts of compounds with similar physical properties but different odor quality.

The acyclic terpene alcohols geraniol, linalool, and citronellol are the most important

terpene alcohols used as fragrance and flavor substances. Geraniol and linalool are, in addition to nerol and lavandulol, primary products in terpene biosynthesis. The fully saturated alcohols tetrahydrogeraniol and tetrahydrolinalool are also used in large quantities in fragrance compositions. The fragrance materials myrcenol, identified in lavender oil, and its dihydro and tetrahydro derivatives, which have not yet been found in nature, belong structurally to the terpenes. The sesquiterpene alcohols farnesol and nerolidol are popular materials for perfume compositions.

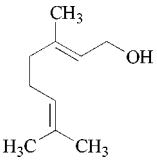
Geraniol

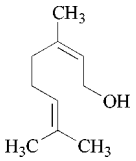
Nerol

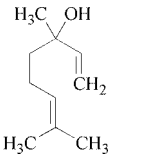
Linalool

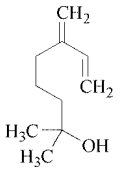
Myrcenol

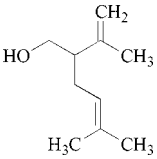
Lavandulol

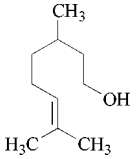
Citronellol

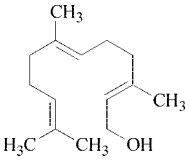
trans,trans-Farnesol

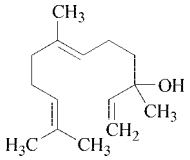
trans-Nerolidol

Geraniol and nerol are *cis–trans*-isomers. In the rarely occurring lavandulol, the isoprene units are not coupled in the normal head-to-tail manner.

The farnesols and nerolidols are sesquiterpene analogs of geraniol–nerol and linalool. These compounds are formed by extending one of the methyl groups in the 7-position of the corresponding monoterpene with an isoprene unit. Because these compounds have an extra double bond, they also have an additional possibility for *cis–trans*-isomerism. Thus, there are four stereoisomers of farnesol and two of nerolidol.

Geraniol [*106-24-1*], 3,7-dimethyl-*trans*-2,6-octadien-1-ol, $C_{10}H_{18}O$, M_r 154.25, *bp* (101.3 kPa) 230 °C, d_4^{20} 0.8890, n_D^{20} 1.4777, occurs in nearly all terpene-containing essential oils, frequently as an ester. Palmarosa oil contains 70 – 85 % geraniol; geranium oils and rose oils also contain large quantities. Geraniol is a colorless liquid, with a flowery-roselike odor.

Since geraniol is an acyclic, doubly unsaturated alcohol, it can undergo a number of reactions, such as rearrangement and cyclization. Rearrangement in the presence of copper catalysts yields citronellal. In the presence of mineral acids, it cyclizes to form monocyclic terpene hydrocarbons, cyclogeraniol being obtained if the hydroxyl function is protected. Partial hydrogenation leads to citronellol, and complete hydrogenation of the double bonds yields 3,7-dimethyloctan-1-ol (tetrahydrogeraniol). Citral is obtained from geraniol by oxidation (e.g., with chromic acid), or by catalytic dehydrogenation. Geranyl esters are prepared by esterification.

Production. Dehydrogenation of geraniol and nerol is a convenient route for synthesizing citral, which is used in large quantities as an intermediate in the synthesis of vitamin A. Large-scale processes have, therefore, been developed for producing geraniol. Currently, these are far more important than isolation from essential oils. Nevertheless, some geraniol is still isolated from essential oils for perfumery purposes.

1) *Isolation from Essential Oils.* Geraniol is isolated from citronella oils and from palmarosa oil. Fractional distillation of, for example, Java citronella oil (if necessary after saponification of the esters present) yields a fraction containing ca. 60 % geraniol, as well as citronellol and sesquiterpenes.
   A product with a higher geraniol content and slightly different odor quality for use in fine fragrances is obtained by fractionating palmarosa oil after saponification of the geranyl esters.
2) *Synthesis from β-Pinene.* Pyrolysis of β-pinene yields myrcene, which is converted into a mixture of predominantly geranyl,

neryl, and linalyl chloride by addition of hydrogen chloride in the presence of small amounts of catalyst, e.g., copper(I) chloride and an organic quaternary ammonium salt [27]. After removal of the catalyst, the mixture is reacted with sodium acetate in the presence of a nitrogen base (e.g., triethylamine) and converted to geranyl acetate, neryl acetate, and a small amount of linalyl acetate [28]. After saponification and fractional distillation of the resulting alcohols, a fraction is obtained that contains ca. 98 % geraniol.

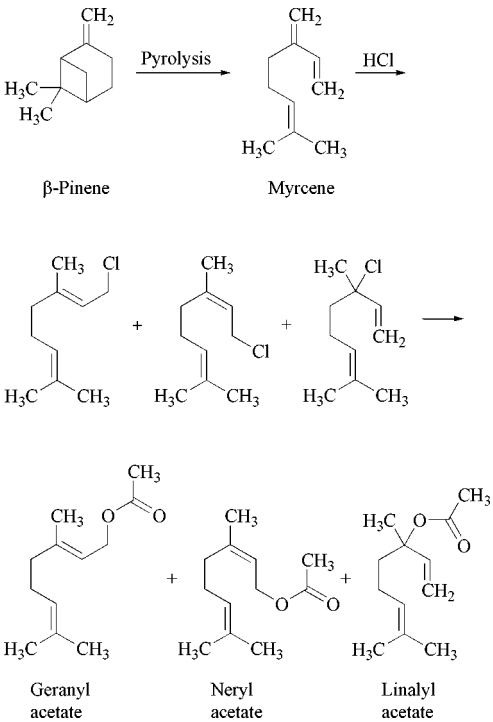

3) *Synthesis from Linalool.* Recently, a 96 % pure synthetic geraniol prepared by isomerization of linalool has become commercially available. Orthovanadates are used as catalysts, to give a > 90 % yield of a geraniol–nerol mixture [29]. Geraniol of high purity is finally obtained by fractional distillation.
A considerable portion of commercially available geraniol is produced by a modified process: linalool obtained in a purity of ca. 65 % from α-pinene is converted into linalyl borates, which rearrange in the presence of vanadates as catalysts to give geranyl and neryl borates. The alcohols are obtained by hydrolysis of the esters [30].

Uses. Geraniol is one of the most frequently used terpenoid fragrance materials. It can be used in all flowery-roselike compositions and does not discolor soaps. In flavor compositions, geraniol is used in small quantities to accentuate citrus notes. It is an important intermediate in the manufacture of geranyl esters, citronellol, and citral.

Nerol [*106-25-2*], 3,7-dimethyl-*cis*-2,6-octadien-1-ol, $C_{10}H_{18}O$, M_r 154.25, *bp* (99.3 kPa) 224 – 225 °C, d_4^{20} 0.8796, n_D^{20} 1.4744, occurs in small quantities in many essential oils where it is always accompanied by geraniol; its name originates from its occurrence in neroli oil. Nerol is a colorless liquid with a pleasant roselike odor which, unlike that of geraniol, has a fresh green note.

Nerol undergoes the same reactions as geraniol, but cyclizes more readily in the presence of acids.

Nerol is produced along with geraniol from myrcene in the process described for geraniol (see Geraniol). It can be separated from geraniol by fractional distillation.

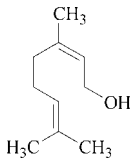

Uses. Nerol is used in perfumery not only for the same purposes as geraniol, e.g., in rose compositions, to which it lends a particular freshness, but also in other blossom compositions. In flavor work it is used for bouquetting citrus flavors. Technical-grade nerol, often in a mixture with geraniol, is used as an intermediate in the production of citronellol and citral.

Linalool [*78-70-6*], 3,7-dimethyl-1,6-octadien-3-ol, $C_{10}H_{18}O$, M_r 154.25, *bp* (101.3 kPa) 198 °C, d_4^{20} 0.8700, n_D^{20} 1.4616, occurs as one of its enantiomers in many essential oils, where it is often the main component. (−)-Linalool [*126-90-9*], for example, occurs at a concentration of 80 – 85 % in Shiu oils from *Cinnamomum camphora; Cayenne linaloe* oil contains

ca. 80 %. (+)-Linalool [*126-91-0*] makes up 60 – 70 % of coriander oil.

Properties. (±)-Linalool [*22564-99-4*] is, like the individual enantiomers, a colorless liquid with a flowery-fresh odor, reminiscent of lily of the valley. However, the enantiomers differ slightly in odor [31]. Together with its esters, linalool is one of the most frequently used fragrance substances and is produced in large quantities.

In the presence of acids, linalool isomerizes readily to geraniol, nerol, and α-terpineol. It is oxidized to citral by chromic acid. Oxidation with peracetic acid yields linalool oxides, which occur in small amounts in essential oils and are also used in perfumery. Hydrogenation of linalool gives tetrahydrolinalool, a stable fragrance compound. Its odor is not as strong as, but fresher than, that of linalool. Linalool can be converted into linalyl acetate by reaction with ketene or an excess of boiling acetic anhydride [32].

Production. In the 1950s nearly all linalool used in perfumery was isolated from essential oils, particularly from rosewood oil. Currently, this method is used only in countries where oils with a high linalool content are available and where the importation of linalool is restricted.

Since linalool is an important intermediate in the manufacture of vitamin E, several large-scale processes have been developed for its production. Preferred starting materials and/or intermediates are the pinenes and 2-methyl-2-hepten-6-one. Most perfumery-grade linalool is synthetic.

1) *Isolation from Essential Oils.* Linalool can be isolated by fractional distillation of essential oils, for example, rosewood oil, Shiu oil, and coriander oil, of which Brazilian rosewood oil is probably the most important.
2) *Synthesis from α-Pinene.* α-Pinene from turpentine oil is selectively hydrogenated to *cis*-pinane [33], which is oxidized with oxygen in the presence of a radical initiator to give a mixture of ca. 75 % *cis*- and 25 % *trans*-pinane hydroperoxide. The mixture is reduced to the corresponding pinanols either with sodium bisulfite ($NaHSO_3$) or a catalyst. The pinanols can be separated by fractional distillation and are pyrolized to linalool: (−)-α-pinene yields *cis*-pinanol and (+)-linalool, whereas (−)-linalool is obtained from *trans*-pinanol [34].

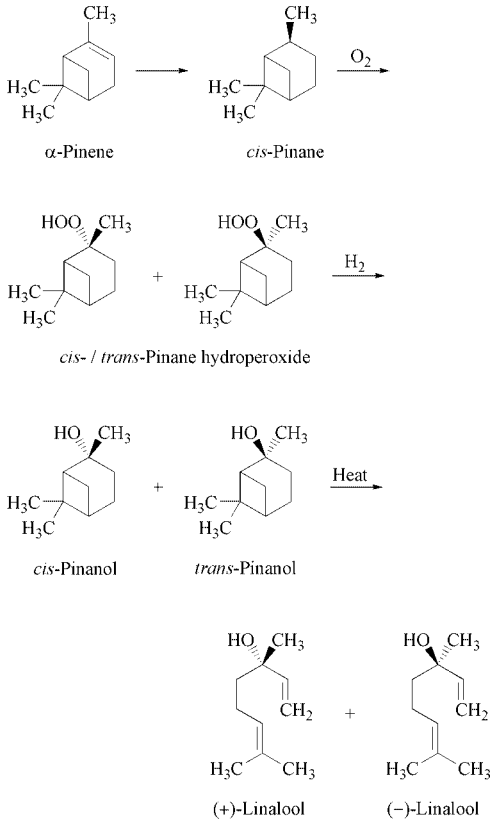

3) *Synthesis from β-Pinene.* For a description of this route, see under Geraniol. Addition of hydrogen chloride to myrcene (obtained from β-pinene) results in a mixture of geranyl, neryl, and linalyl chlorides. Reaction of this mixture with acetic acid–sodium acetate in the presence of copper(I) chloride gives linalyl acetate in 75 – 80 % yield [35]. Linalool is obtained after saponification.
4) *Synthesis from 2-Methyl-2-hepten-6-one.* The total synthesis of linalool starts with 2-methyl-2-hepten-6-one; several large-scale processes have been developed for synthesizing this compound:

a) Addition of acetylene to acetone results in the formation of 3-methyl-1-butyn-3-ol, which is hydrogenated to 3-methyl-1-buten-3-ol in the presence of a palladium catalyst. This product is converted into its acetoacetate derivative with diketene [36] or with ethyl acetoacetate [37]. The acetoacetate undergoes rearrangement when heated (Carroll reaction) to give 2-methyl-2-hepten-6-one:

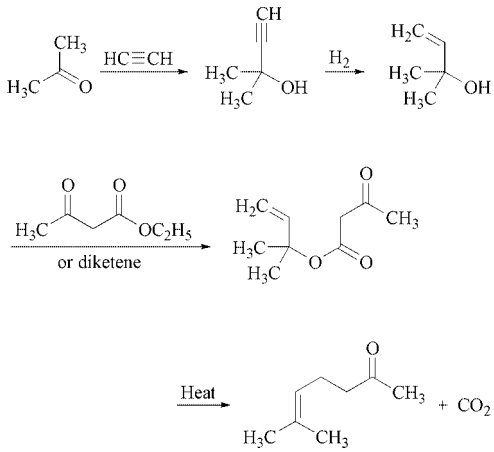

b) In another process, 2-methyl-2-hepten-6-one is obtained by reaction of 3-methyl-1-buten-3-ol with isopropenyl methyl ether followed by a Claisen rearrangement [38]:

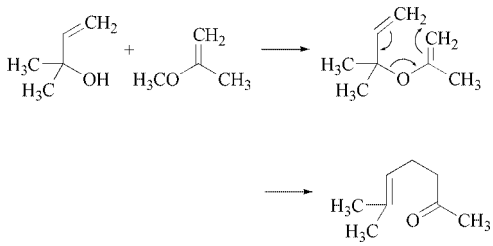

c) A third synthesis starts from isoprene, which is converted into 3-methyl-2-butenyl chloride by addition of hydrogen chloride. Reaction of the chloride with acetone in the presence of a catalytic amount of an organic base [39] leads to 2-methyl-2-hepten-6-one:

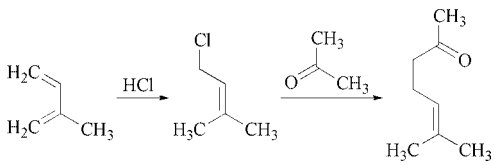

d) In another process, 2-methyl-2-hepten-6-one is obtained by isomerization of 2-methyl-1-hepten-6-one [40]. The latter can be prepared in two steps from isobutylene and formaldehyde. 3-Methyl-3-buten-1-ol is formed in the first step [41] and is converted into 2-methyl-1-hepten-6-one by reaction with acetone [42].

2-Methyl-2-hepten-6-one is converted into linalool in excellent yield by base-catalyzed ethynylation with acetylene to dehydrolinalool [43]. This is followed by selective hydrogenation of the triple bond to a double bond in the presence of a palladium–carbon catalyst.

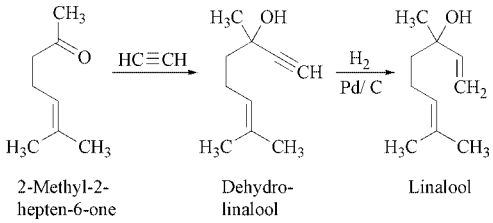

Uses. Linalool is used frequently in perfumery for fruity notes and for many flowery fragrance compositions (lily of the valley, lavender, and neroli). Because of its relatively high volatility, it imparts naturalness to top notes. Since linalool is stable in alkali, it can be used in soaps and detergents. Linalyl esters can be prepared from linalool. Most of the manufactured linalool is used in the production of vitamin E.

Myrcenol [*543-39-5*], 2-methyl-6-methylene-7-octen-2-ol, $C_{10}H_{18}O$, M_r 154.25, *bp* (6.7 kPa) 78 °C, d_4^{20} 0.8764, n_D^{20} 1.4731, is an isomer of geraniol and linalool. It has been identified in Chinese lavender oil [44] and some medicinal plants. It is a colorless liquid with a fresh-flowery, slightly limelike odor. Due to its conjugated double bonds, it tends to polymerize; polymerization can be suppressed by adding inhibitors (e.g., antioxidants such as ionol).

Myrcenol can be prepared by treating myrcene with diethylamine to give a mixture of geranyl- and neryldiethylamine. These compounds are hydrated with a dilute acid to the corresponding hydroxydiethylamines. Deamination to myrcenol is effected by using a palladium–phosphine-cation complex as a catalyst [45].

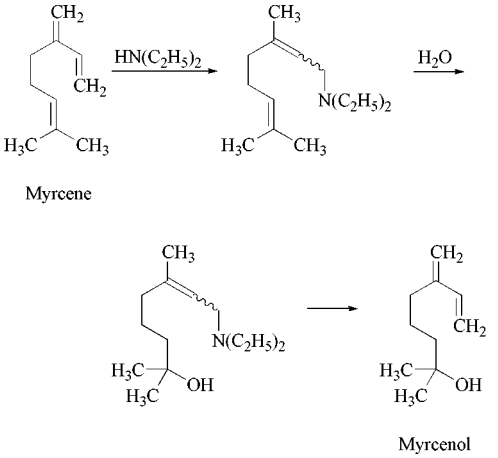

Myrcenol is used in perfumery to obtain a lifting top note in citrus and lavender compositions.

Citronellol [*26489-01-0*], 3,7-dimethyl-6-octen-1-ol, $C_{10}H_{20}O$, M_r 156.27, *bp* (101.3 kPa) 244.4 °C, d_4^{20} 0.8560, n_D^{20} 1.4558, $[\alpha]_D^{20}$ + resp. − 5 to 6 °, occurs as both (+)-citronellol [*1117-61-9*] and (−)-citronellol [*7540-51-4*] in many essential oils.

(−)-Citronellol isolated from natural sources is often named rhodinol. At present, the name rhodinol is also used for the isopropenyl isomer, α-citronellol; therefore, exclusive use of the systematic name is better.

In many natural products citronellol occurs as a mixture of its two enantiomers; the pure (+) or (−) form is seldom found. (+)-Citronellol dominates in oils from *Boronia citriodora* (total citronellol content ca. 80 %) and *Eucalyptus citriodora* (citronellol content 15 – 20 %). (−)-Citronellol is the predominant enantiomer in geranium and rose oils, both of which may contain up to 50 % citronellols.

Citronellol is a colorless liquid with a sweet roselike odor. The odor of (−)-citronellol is more delicate than that of (+)-citronellol.

Citronellol undergoes the typical reactions of primary alcohols (→ Alcohols, Aliphatic, Chap. 2.2.). Compared with geraniol, which contains one more double bond, citronellol is relatively stable. Citronellol is converted into citronellal by dehydrogenation or oxidation; hydrogenation yields 3,7-dimethyloctan-1-ol. Citronellyl esters are easily prepared by esterification with acid anhydrides.

Production. (−)-Citronellol is still obtained mainly from geranium oil by saponification followed by fractional distillation. Although of high odor quality, it does not possess the true (−)-citronellol odor due to impurities. Much larger quantities of (+)- and (±)-citronellol are used and are prepared by partial or total synthesis.

1) *Synthesis of (+)- and (±)-Citronellol from the Citronellal Fraction of Essential Oils.* (+)-Citronellal is obtained by distillation of Java citronella oil and is hydrogenated to (+)-citronellol in the presence of a catalyst (e.g., Raney nickel). Similarly, (±)-citronellol is prepared from the (±)-citronellal fraction of *Eucalyptus citriodora* oil.
2) *Synthesis of (±)- or Slightly Dextrorotatory Citronellol from Geraniol Fractions of Essential Oils.* This citronellol is produced by catalytic hydrogenation of saponified geraniol fractions (also containing (+)-citronellol) obtained from Java citronella oil, followed by fractional distillation. Selective hydrogenation of the double bond in the 2-position of geraniol in geraniol–citronellol mixtures isolated from essential oils can be achieved by using Raney cobalt as a catalyst; overhydrogenation to 3,7-dimethyloctan-1-ol can be largely avoided by this method [46].

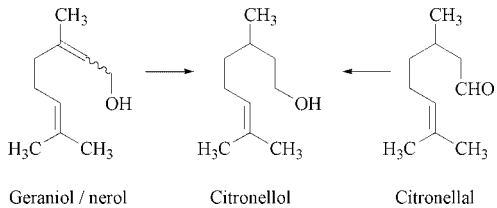

3) *Synthesis of (±)-Citronellol from Synthetic Geraniol– Nerol or Citral.* A considerable amount of commercial synthetic (±)-citronellol is produced by partial hydrogenation of synthetic geraniol and/or nerol. Another starting material is citral, which can be hydrogenated, e.g., in the presence of a catalyst system consisting of palladium, ruthenium, and trimethylamine [47].
4) *Preparation of (−)-Citronellol from Optically Active Pinenes.* (+)- *cis*-Pinane is readily synthesized by hydrogenation of (+)-α-pinene or (+)-β-pinene, and is then pyrolyzed to give (+)-3,7-dimethyl-1,6-octadiene. This

compound is converted into (−)-citronellol (97 % purity) by reaction with triisobutylaluminum or diisobutylaluminum hydride, followed by air oxidation and hydrolysis of the resulting aluminum alcoholate [48].

Uses. Citronellol is one of the most widely used fragrance materials, particularly for rose notes and for floral compositions in general. As flavor material, citronellol is added for bouquetting purposes to citrus compositions. It is the starting material for numerous citronellyl esters and for hydroxydihydrocitronellol, an intermediate in the production of hydroxydihydrocitronellal.

Dihydromyrcenol [*18479-58-8*], 2,6-dimethyl-7-octen-2-ol, $C_{10}H_{20}O$, M_r 156.27, *bp* (1.3 kPa) 77 – 79 °C, d_4^{20} 0.841, is a colorless liquid with a fresh citrus-like odor and a lavender note. It is prepared from 2,6-dimethyl-2,7-octadiene, the pyrolysis product of *cis*-pinane [49], by addition of hydrogen chloride and hydrolysis of the resulting 2,6-dimethyl-2-chloro-7-octene [50].

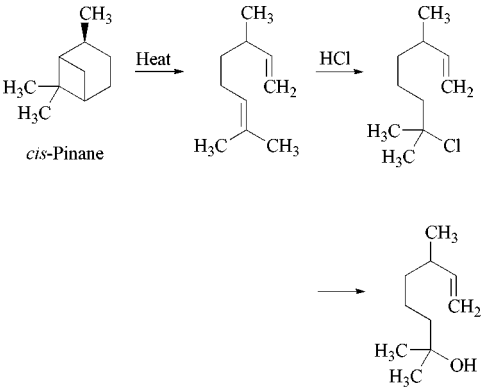

cis-Pinane

Dihydromyrcenol

Dihydromyrcenol is used in soap perfumes for lime and blossom notes.

Tetrahydrogeraniol [*106-21-8*], 3,7-dimethyloctan-1-ol, $C_{10}H_{22}O$, M_r 158.28, *bp* (101.3 kPa) 212 – 213 °C, d_4^{20} 0.8285, n_D^{20} 1.4355, has been identified in citrus oils and is a colorless liquid with a waxy, rose-petal-like odor. It is prepared by hydrogenation of geraniol or citronellol in the presence of a nickel catalyst and is a byproduct in the synthesis of citronellol from geraniol or nerol. Because of its stability, it is often used to perfume household products.

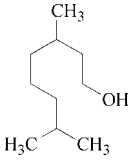

Tetrahydrolinalool [*78-69-3*], 3,7-dimethyloctan-3-ol, $C_{10}H_{22}O$, M_r 158.28, *bp* (1.3 kPa) 78 – 79 °C, d_4^{20} 0.8294, n_D^{20} 1.4335, is a constituent of honey aroma. It is a colorless liquid with a linalool-like odor that is slightly fresher but distinctly weaker than that of linalool. Tetrahydrolinalool is prepared by catalytic hydrogenation of linalool and is used as a substitute for the less stable linalool in perfuming aggressive media.

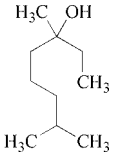

Farnesol [*4602-84-0*], 3,7,11-trimethyl-2,6,10-dodecatrien-1-ol, $C_{15}H_{26}O$, M_r 222.37, *bp* (1.6 kPa) 156 °C, d_4^{20} 0.8846, n_D^{20} 1.4890, is a component of many blossom oils. It is a colorless liquid with a linden blossom odor, which becomes more intense when evaporated, possibly due to oxidation.

Of the four possible isomers (due to the double bonds in the 2- and 6-positions), the *trans*–*trans*-isomer is the most common in nature and occurs, for example, in ambrette seed oil. 2-*cis*-6-*trans*-Farnesol has been identified in petitgrain oil Bigarade.

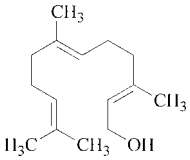

trans-trans-Farnesol

Since the odors of the isomers differ very little, natural farnesol in compositions can be replaced by synthetic farnesol, which is a mixture of isomers obtained by isomerization of nerolidol.

Farnesol is particularly suited for use in flower compositions and is valued for its fixative properties. IFRA recommends to use only qualities containing at least 96 % of the farnesol isomers.

Nerolidol [*7212-44-4*], 3,7,11-trimethyl-1,6,10-dodecatrien-3-ol, $C_{15}H_{26}O$, M_r 222.37, *bp* (1.6 kPa) 145 °C, d_4^{20} 0.8778, n_D^{20} 1.4898, is the sesquiterpene analogue of linalool. Because of the double bond at the 6-position, it exists as *cis*- and *trans*-isomers. Each of these isomers can exist as an enantiomeric pair, since the carbon atom in the 3-position is chiral.

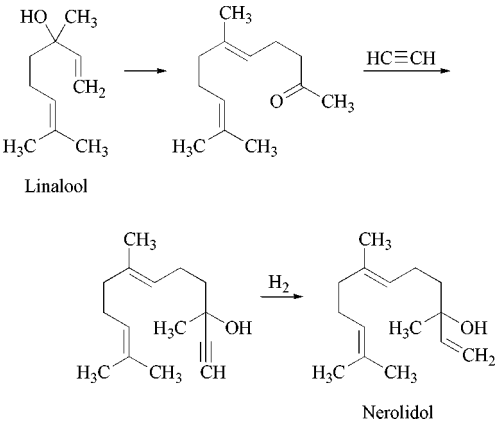

Nerolidol is a component of many essential oils. (+)-*trans*-Nerolidol occurs in cabreuva oil; (−)-nerolidol has been isolated from *Dalbergia parviflora* wood oils.

Synthetic nerolidol consists of a mixture of (±)-*cis*- and (±)-*trans*-nerolidol and is a colorless liquid with a long-lasting, mild flowery odor.

Industrial synthesis of nerolidol starts with linalool, which is converted into geranylacetone by using diketene, ethyl acetoacetate, or isopropenyl methyl ether, analogous to the synthesis of 2-methyl-2-hepten-6-one from 3-methyl-1-buten-3-ol (see page 16). Addition of acetylene and partial hydrogenation of the resultant dehydronerolidol produces a mixture of *cis*- and *trans*-nerolidol racemates.

Nerolidol is used as a base note in many delicate flowery odor complexes. It is also an intermediate in the production of vitamins E and K_1.

2.2.3. Aldehydes and Acetals

Among the acyclic terpene aldehydes, citral and citronellal hold key positions as fragrance and flavor chemicals, as well as starting materials for the synthesis of other terpenoids. Hydroxydihydrocitronellal is one of the most important fragrance compounds. Derivatives of these aldehydes, particularly the lower acetals, are also used as fragrance compounds. Acyclic sesquiterpene aldehydes are not very important as such, but they contribute to the characteristic fragrance and aroma of essential oils, for example, in the case of α- and β-sinensal and sweet orange oil (seepage 94).

Citral [*5392-40-5*], 3,7-dimethyl-2,6-octadien-1-al, $C_{10}H_{16}O$, M_r 152.24, occurs as *cis*- and *trans*-isomers (citral a and b, respectively) analogous to the corresponding alcohols, geraniol and nerol: citral a [*141-27-5*] (geranial), *bp* (2.7 kPa) 118 – 119 °C, d_4^{20} 0.8888, n_D^{20} 1.4898; citral b [*106-26-3*] (neral), *bp* (2.7 kPa) 120 °C, d_4^{20} 0.8869, n_D^{20} 1.4869.

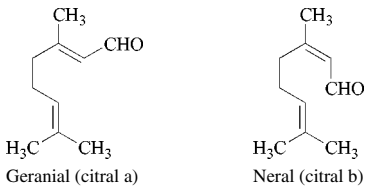

Natural citral is nearly always a mixture of the two isomers. It occurs in lemongrass oil (up to 85 %), in *Litsea cubeba* oil (up to 75 %), and in small amounts in many other essential oils. The citrals are colorless to slightly yellowish liquids, with an odor reminiscent of lemon.

Since citral is an α,β-unsaturated aldehyde with an additional double bond, it is highly reactive and may undergo reactions such as cyclization and polymerization. Geraniol, citronellol, and 3,7-dimethyloctan-1-ol can be obtained from citral by stepwise hydrogenation. Citral can be converted into a number of addition compounds; the *cis*- and *trans*-isomers can be separated via the hydrogen sulfite addition compounds. The condensation of citral with active methylene groups is used on an industrial scale in the synthesis of pseudoionones, which are starting materials for ionones and vitamins.

Production. Since citral is used in bulk as a starting material for the synthesis of vitamin A, it is produced industrially on a large scale. Smaller quantities are also isolated from essential oils.

1) *Isolation from Essential Oils.* Citral is isolated by distillation from lemongrass oil (see p. page 98) and from *Litsea cubeba* oil (see p. page 106). It is the main component of these oils.
2) *Synthesis from Geraniol.* Currently, the most important synthetic procedures are vapor-phase dehydrogenation and oxidation of geraniol or geraniol–nerol mixtures. Catalytic dehydrogenation under reduced pressure using copper catalysts is preferred [51].
3) *Synthesis from Dehydrolinalool.* Dehydrolinalool is produced on a large scale from 2-methyl-2-hepten-6-one and acetylene (see page 16) and can be isomerized to citral in high yield by a number of catalysts. Preferred catalysts include organic orthovanadates [52], organic trisilyl oxyvanadates [53], and vanadium catalysts with silanols added to the reaction system [54].

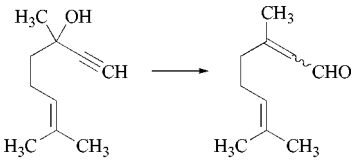

Dehydro-linalool — Citral

4) *Synthesis from Isobutene and Formaldehyde.* 3-Methyl-3-buten-1-ol, obtained from isobutene and formaldehyde [41], isomerizes to form 3-methyl-2-buten-1-ol [55]. However, it is also converted into 3-methyl-2-butenal by dehydrogenation and subsequent isomerization [56,57]. Under azeotropic conditions in the presence of nitric acid, 3-methyl-2-buten-1-ol and 3-methyl-2-butenal form an acetal (shown below) [58], which eliminates one molecule of 3-methyl-2-buten-1-ol at higher temperatures. The intermediate enol ether undergoes Claisen rearrangement followed by Cope rearrangement to give citral in excellent yield [59]:

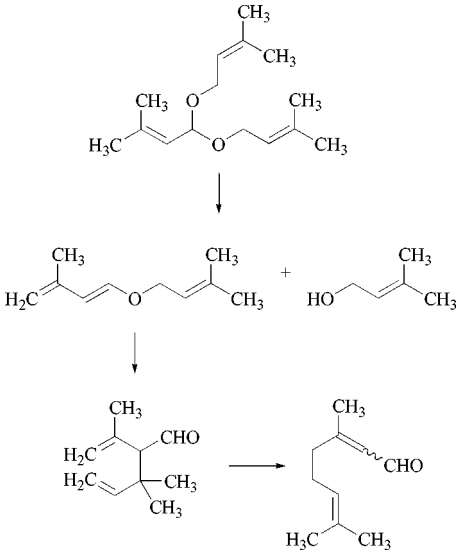

Uses. Because of its strong lemon odor, citral is very important for aroma compositions such as citrus flavors. In perfumery it can be used only in neutral media due to its tendency to undergo discoloration, oxidation, and polymerization. Limiting IFRA recommendation. It is used as a starting material in the synthesis of ionones and methylionones, particularly β-ionone, which is an intermediate in vitamin A synthesis.

Citral diethyl acetal [*7492-66-2*], 3,7-dimethyl-2,6-octadien-1-al diethyl acetal, $C_{14}H_{26}O_2$, M_r 226.36, *bp* (2 kPa) 140 – 142 °C, d_4^{20} 0.8730, n_D^{20} 1.4503, is a colorless liquid with a flowery, warm-woody citrus odor. It is relatively stable in alkali and can, therefore, be used in soap.

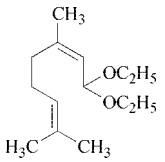

Citronellal [*106-23-0*], 3,7-dimethyl-6-octen-1-al, $C_{10}H_{18}O$, M_r 154.25, *bp* (101.3 kPa) 207 – 208 °C, d_4^{20} 0.851, n_D^{20} 1.4477, $[\alpha]_D^{20}$ + resp. − 15 °, $[\alpha]_D^{20}$ − 13.1 °, occurs in essential oils in its (+) and (−) forms, often together with the racemate. (+)-Citronellal [*2385-77-5*] occurs in citronella oil at a concentration of up

to 45 %; *Backhousia citriodora* oil contains up to 80 % (−)-citronellal [*5949-05-3*]. Racemic citronellal [*26489-02-1*] occurs in a number of *Eucalyptus citriodora* oils at a concentration of up to 85 %.

Pure citronellal is a colorless liquid with a refreshing odor, reminiscent of balm mint. Upon catalytic hydrogenation, citronellal yields dihydrocitronellal, citronellol, or dihydrocitronellol, depending on the reaction conditions. Protection of the aldehyde group, followed by addition of water to the double bond in the presence of mineral acids or ion-exchange resins results in formation of 3,7-dimethyl-7-hydroxyoctan-1-al (hydroxydihydrocitronellal, see page 24). Acid-catalyzed cyclization to isopulegol is an important step in the synthesis of (−)-menthol (seepage 30).

Production. Citronellal is still isolated from essential oils in considerable quantities; it is also produced synthetically.

1) *Isolation from Essential Oils.* (+)-Citronellal is obtained from citronella oils by fractional distillation. (±)-Citronellal is isolated from *Eucalyptus citriodora* oil (see page 100); when necessary, it is purified by using an addition compound, e.g., the bisulfite derivative.
2) *Synthesis from Geraniol or Nerol.* (±)-Citronellal can be obtained by vapor-phase rearrangement of geraniol or nerol (see page 16) in the presence of, e.g., a barium-containing copper–chromium oxide catalyst [60].

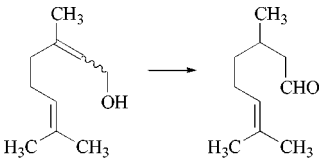

Geraniol / nerol Citronellal

3) *Synthesis from Citronellol.* (±)-Citronellal can also be obtained by dehydrogenation of citronellol (see page 19) under reduced pressure with a copper chromite catalyst [61].
4) *Synthesis from Citral.* Selective hydrogenation of citral to citronellal can be accomplished in the presence of a palladium catalyst in an alkaline alcoholic reaction medium [62].

Uses. Citronellal is used to a limited extent for perfuming soaps and detergents. Its main use is as a starting material for the production of isopulegol, citronellol, and hydroxydihydrocitronellal.

7-Hydroxydihydrocitronellal [*107-75-5*], hydroxycitronellal, 3,7-dimethyl-7-hydroxyoctan-1-al, $C_{10}H_{20}O_2$, M_r 172.27, *bp* (130 Pa) 85 – 87 °C, d_4^{20} 0.9220, n_D^{20} 1.4488, has been reported to occur in essential oils [63]. It is a colorless, slightly viscous liquid with a flowery odor reminiscent of linden blossom and lily of the valley. Commercially available hydroxycitronellal is either optically active or racemic, depending on the starting material used. Hydroxycitronellal prepared from (+)-citronellal, for example, has a specific rotation $[\alpha]_D^{20}$ + 9 to + 10 °.

Hydroxycitronellal is relatively unstable toward acid and alkali and is, therefore, sometimes converted into more alkali-resistant acetals, particularly its dimethyl acetal.

Production. The most important synthetic routes to hydroxydihydrocitronellal are listed below.

1) *Synthesis from Citronellal.* One of the oldest routes to hydroxycitronellal is the hydration of the citronellal bisulfite adduct (obtained at low temperature) with sulfuric acid, followed by decomposition with sodium carbonate. A more recent development is hydration of citronellal enamines or imines, followed by hydrolysis [64].
2) *Synthesis from Citronellol.* Citronellol (see page 19) is hydrated to 3,7-dimethyloctan-1,7-diol, for example, by reaction with 60 % sulfuric acid. The diol is dehydrogenated catalytically in the vapor phase at low pressure to highly pure hydroxydihydrocitronellal in excellent yield. The process is carried out in the presence of, for example, a copper–zinc catalyst [65]; at atmospheric pressure noble metal catalysts can also be used [66].
3) *Synthesis from 7-Hydroxygeranyl–Neryl Dialkylamine.* The starting material can be obtained by treatment of myrcene with a dialkylamine in the presence of an alkali dialkylamide, followed by hydration with sulfuric acid. The 7-hydroxygeranyl–neryl dialkylamine isomerizes to the corresponding 7-hy-

droxyaldehyde enamine in the presence of a palladium(II)–phosphine complex as catalyst. Hydrolysis of the enamine gives 7-hydroxydihydrocitronellal [67].

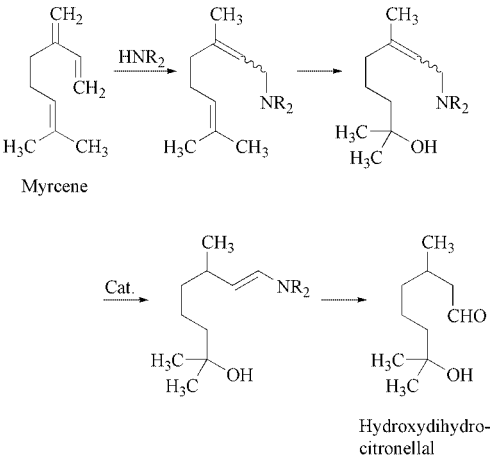

Uses. Because of its fine, flowery odor, hydroxydihydrocitronellal is used in large quantities in many perfume compositions for creating linden blossom and lily of the valley notes. It is also used in other blossom fragrances such as honeysuckle, lily, and cyclamen. Limiting IFRA recommendation.

Methoxydihydrocitronellal [*3613-30-7*], 3,7-dimethyl-7-methoxyoctan-1-al, $C_{11}H_{22}O_2$, M_r 186.29, *bp* (60 Pa) 60 °C, n_D^{20} 1.4375, is a colorless liquid with a fresh, green, blossom odor and is used in perfumery in floral compositions for fresh-green nuances.

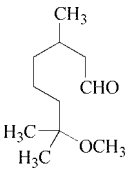

Hydroxydihydrocitronellal dimethyl acetal [*141-92-4*], 3,7-dimethyl-7-hydroxyoctan-1-al dimethyl acetal, $C_{12}H_{26}O_3$, M_r 218.34, *bp* (1.6 kPa) 131 °C, d_4^{20} 0.931, n_D^{20} 1.4419, is a colorless liquid with a weak, flowery odor. Since the acetal is stable to alkali, it is used occasionally in soap perfumes.

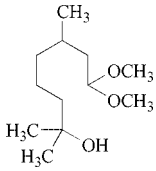

Citronellyloxyacetaldehyde [*7492-67-3*], $C_{12}H_{22}O_2$, M_r 198.30, *bp* (1.6 kPa) 128–130 °C, is a colorless, slightly viscous liquid with a delicate-flowery odor reminiscent of rose and lily of the valley. The aldehyde is obtained by hydrolysis (with dilute acetic acid) of its acetals, which are prepared from sodium citronellate and chloroacetaldehyde dialkyl acetals [68]. It is used in perfumery to obtain heavy-flowery effects.

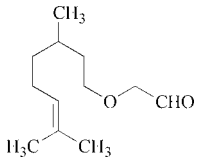

2,6,10-Trimethyl-9-undecenal [*141-13-9*], $C_{14}H_{26}O$, M_r 210.36, *bp* (1.2 kPa) 133–135 °C, d_4^{20} 0.8483, n_D^{20} 1.447–1.453, is a colorless to slightly yellow liquid with an intense aldehyde-waxy, slightly flowery odor. It is synthesized from hydrogenated pseudoionone (primarily the tetrahydro compound) and an alkyl chloroacetate by means of a glycidic ester condensation; this is followed by hydrolysis and decarboxylation.

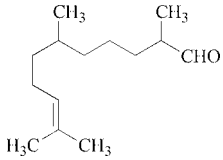

2,6,10-Trimethyl-9-undecenal is a richly fragrant compound that is used in flower compositions to obtain an aldehydic note.
Trade Names. Adoxal (Givaudan Roure), Farenal (Haarmann & Reimer).

2.2.4. Ketones

Unlike the terpene alcohols, aldehydes, and esters, acyclic terpene ketones are not particularly

important as fragrance or flavor substances; thus, they are not discussed here in detail. 2-Methyl-2-hepten-6-one (see page 16) is an important intermediate in the synthesis of terpenoids. Its odor properties are not impressive. It occurs in nature as a decomposition product of terpenes. Tagetone [*6752-80-3*] is a major component of tagetes oil. Solanone [*1937-54-8*] and pseudoionone [*141-10-6*] are acyclic C_{13} ketones with a terpenoid skeleton. Solanone is one of the flavor-determining constituents of tobacco, pseudoionone is an intermediate in the synthesis of ionones (see page 36).

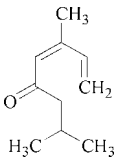
Tagetone

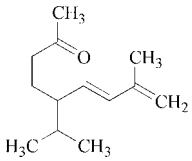
Solanone

Geranylacetone [*689-67-8*], 6,10-dimethyl-5,9-undecadien-2-one, $C_{13}H_{22}O$, M_r 194.32, *bp* (1.3 kPa) 124 °C, d_4^{20} 0.8729, n_D^{20} 1.4674, occurs in *cis* as well as *trans* form and has been identified in fruits and in essential oils. It is a colorless liquid with a fresh-green, slightly penetrating, roselike odor.

Geranylacetone is an intermediate in nerolidol synthesis (see page 21). It is used in perfumery in rose compositions, for example, in soap perfumes.

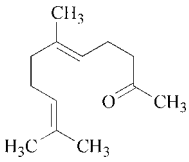
trans-Geranylacetone

2.2.5. Acids and Esters

Although a small amount of acyclic terpene acids such as geranic acid and citronellic acid occurs in many essential oils, often as esters, they are rarely used in perfume and flavor compositions.

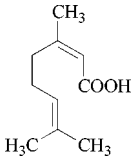
cis-Geranic acid

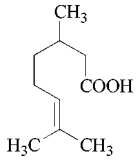
Citronellic acid

Methyl geranate is an intermediate in α-damascone synthesis and is sometimes needed in the reconstitution of essential oils. However, the lower fatty acid esters (particularly the acetates) of the acyclic terpene alcohols geraniol, linalool, and citronellol are extremely important both as fragrance and as flavor substances. The acetates occur in many essential oils, sometimes in rather high amounts. Formates, propionates, and butyrates occur less frequently. As a result of the development of large-scale production processes for terpenes, the esters of acyclic terpene alcohols are nearly always made synthetically. All acyclic terpene esters that are used as fragrance and flavor materials can be prepared by direct esterification of the appropriate alcohols. However, special precautions are required for the esterification of linalool.

Because the lower fatty acid esters of geraniol, linalool, and citronellol are important contributors to the odor of many essential oils, these esters are widely used in the reconstitution of such oils, as well as in perfume and flavor compositions. The acetates, particularly linalyl acetate, are most widely used. The use of formates is limited by their relative instability. Higher esters are not important in terms of quantity, but are indispensable for creating specific nuances.

In aroma compositions, fatty acid esters of the acyclic terpene alcohols are used for obtaining citrus notes and for rounding off other flavor types.

The most important and most frequently used acyclic terpene esters are described below.

2.2.5.1. Geranyl and Neryl Esters

Geraniol is described on page 15, nerol on page 16.

Geranyl formate [*105-86-2*], $C_{11}H_{18}O_2$, M_r 182.26, *bp* (101.3 kPa) 229 °C (decomp.), d_4^{20} 0.9209, n_D^{20} 1.4659, is a liquid with a fresh, crisp-herbal-fruity rose odor. It is used as a mod-

ifier of, among others, rose, geranium, and neroli compositions.

Geranyl acetate [*105-87-3*], $C_{12}H_{20}O_2$, M_r 196.29, *bp* (1.5 kPa) 98 °C, d_4^{20} 0.9136, n_D^{20} 1.4624, occurs in varying amounts in many essential oils: up to 60 % in oils from *Callitris* and *Eucalyptus* species, and up to 14 % in palmarosa oil. A smaller amount occurs in, for example, geranium, citronella, petitgrain, and lavender oils. Geranyl acetate is a liquid with a fruity, rose note reminiscent of pear and slightly of lavender. It is used frequently in perfumery to create not only flowery-fruity nuances (e.g., rose), but also for citrus and lavender notes. A small amount is added to fruit aromas for shading.

Geranyl propionate [*105-90-8*], $C_{13}H_{22}O_2$, M_r 210.32, *bp* (101.3 kPa) 253 °C, d_4^{20} 0.9207, n_D^{20} 1.459, has a fruity rose odor and is used in perfumery in heavy blossom fragrances with a secondary fruity note.

Geranyl isobutyrate [*2345-26-8*], $C_{14}H_{24}O_2$, M_r 224.34, *bp* (101.3 kPa) 265 °C, d_4^{20} 0.9209, n_D^{20} 1.4576, is a liquid with a fruity rose odor. It is used in floral perfume compositions and in fruit flavors.

Geranyl isovalerate [*109-20-6*], $C_{15}H_{26}O_2$, M_r 238.37, *bp* (101.3 kPa) 279 °C, d_4^{20} 0.8920, n_D^{20} 1.4640, is a liquid with a strongly fruity rose odor. It is used in perfume and flavor compositions.

Neryl acetate [*141-12-8*], $C_{12}H_{20}O_2$, M_r 196.29, *bp* (3.4 kPa) 134 °C, d_4^{20} 0.9110, n_D^{20} 1.4624, is the *cis*-isomer of geranyl acetate. It is present in helichrysum oil and has also been identified in, among others, neroli oil and petitgrain oil Bigarade. It is a colorless, flowery-sweet-smelling liquid and is used in perfumery for blossom compositions (e.g., orange blossom and jasmin).

2.2.5.2. Linalyl and Lavandulyl Esters

Among the linalyl esters (linalool, see page 16), the acetate is by far the most important fragrance and flavor substance. The formate, propionate, and butyrates are used in small amounts.

Linalyl formate [*115-99-1*], $C_{11}H_{18}O_2$, M_r 182.26, *bp* (1.3 kPa) 100 – 103 °C, d_4^{20} 0.9134, n_D^{20} 1.4530, is a liquid with a fruity odor, reminiscent of bergamot. Linalyl formate is moderately stable and is used in lavender fragrances and eaux de cologne.

Linalyl acetate [*115-95-7*], $C_{12}H_{20}O_2$, M_r 196.29, *bp* (101.3 kPa) 220 °C, d_4^{20} 0.8951, n_D^{20} 1.4500, occurs as its (−)-isomer [*16509-46-9*] as the main component of lavender oil (30 – 60 %, depending on the origin of the oil), of lavandin oil (25 – 50 %, depending on the species), and of bergamot oil (30 – 45 %). It has also been found in clary sage oil (up to 75 %) and in a small amount in many other essential oils. (±)-Linalyl acetate [*40135-38-4*] is a colorless liquid with a distinct bergamot-lavender odor.

Production. Linalyl acetate is synthesized by two methods:

1) Esterification of linalool requires special reaction conditions since it tends to undergo dehydration and cyclization because it is an unsaturated tertiary alcohol. These reactions can be avoided as follows: Esterification with ketene in the presence of an acidic esterification catalyst below 30 °C results in formation of linalyl acetate without any byproducts [69]. Esterification can be achieved in good yield, with boiling acetic anhydride, whereby the acetic acid is distilled off as it is formed; a large excess of acetic anhydride must be maintained by continuous addition of anhydride to the still vessel [32]. Highly pure linalyl acetate can be obtained by transesterification of *tert*-butyl acetate with linalool in the presence of sodium methylate and by continuous removal of the *tert*-butanol formed in the process [70].
2) Dehydrolinalool, obtained by ethynylation of 2-methyl-2-hepten-6-one, can be converted into dehydrolinalyl acetate with acetic anhydride in the presence of an acidic esterification catalyst. Partial hydrogenation of the triple bond to linalyl acetate can be carried out with, for example, palladium catalysts deactivated with lead [71].

Uses. Linalyl acetate is used extensively in perfumery. It is an excellent fragrance material for, among others, bergamot, lilac, lavender,

linden, neroli, ylang-ylang, and phantasy notes (particularly chypre). Smaller amounts are used in other citrus products. Since linalyl acetate is fairly stable toward alkali, it can also be employed in soaps and detergents.

Linalyl propionate [*144-39-8*], $C_{13}H_{22}O_2$, M_r 210.31, *bp* (101.3 kPa) 226 °C, d_4^{20} 0.9009, n_D^{20} 1.4505, is a liquid with a fresh bergamot note, reminiscent of lily of the valley. It is used in perfumery in, for example, bergamot, lavender, and lily of the valley compositions.

Linalyl butyrate [*78-36-4*], $C_{14}H_{24}O_2$, M_r 224.34, *bp* (101.3 kPa) 232 °C, d_4^{20} 0.8987, n_D^{20} 1.4523, is a liquid with a fruity bergamot note and a subdued animalic tone. It is used in lavender perfumes and in many blossom compositions.

Linalyl isobutyrate [*78-35-3*], $C_{14}H_{24}O_2$, M_r 224.34, *bp* (20 Pa) 63 – 65 °C, d_4^{20} 0.8991, n_D^{20} 1.4513, has a fresh-fruity lavender odor, which is more refined than that of the butyrate. It is used in lavender compositions and in several flowery notes.

Lavandulyl acetate [*25905-14-0*] (for formula of lavandulol, see Section 2.2.2), $C_{12}H_{20}O_2$, M_r 196.29, *bp* (1.7 kPa) 106 – 107 °C, d_4^{20} 0.9129, n_D^{20} 1.4561, occurs in its (−)-form at a concentration of ca. 1 % in French lavender oil and lavandin oil. It is a liquid with a fresh-herbal rose odor and is used in perfumery for lavender- and lavandin-oil reconstitutions.

One synthetic route to lavandulyl acetate starts with prenyl acetate, which dimerizes in the presence of a Friedel–Crafts catalyst, such as boron trifluoride–diacetic acid [72].

2.2.5.3. Citronellyl Esters

The following citronellyl esters (citronellol, see page 19) are used in relatively large amounts as fragrance and flavor materials:

Citronellyl formate [*105-85-1*], $C_{11}H_{20}O_2$, M_r 184.28, *bp* (2 kPa) 97 – 98 °C, d_4^{20} 0.8972, n_D^{20} 1.4556, is a liquid with a strongly fruity, roselike odor, which is suitable for fresh top notes in rose and lily of the valley fragrances.

Citronellyl acetate [*67650-82-2*], $C_{12}H_{22}O_2$, M_r 198.30, *bp* (101.3 kPa) 240 °C, d_4^{20} 0.8889, n_D^{20} 1.4515, occurs in many essential oils either as one of its optical isomers or as the racemate. The odor of racemic citronellyl acetate differs little from that of the optical isomers. (±)-Citronellyl acetate is a liquid with a fresh-fruity rose odor. It is often used as a fragrance, for example, for rose, lavender, and geranium notes as well as for eaux de cologne with citrus nuances. Since it is relatively stable to alkali, it can be used in soaps and detergents. Citrus flavors acquire specific character through the addition of citronellyl acetate; it is also used to round off other fruit flavors.

Citronellyl propionate [*141-14-0*], $C_{13}H_{24}O_2$, M_r 212.33, *bp* (2 kPa) 120 – 124 °C, d_4^{20} 0.8851, n_D^{20} 1.4452, is a fresh-fruity, roselike smelling liquid. It is used in perfume and flavor compositions in the same way as the acetate.

Citronellyl isobutyrate [*97-89-2*], $C_{14}H_{26}O_2$, M_r 226.36, *bp* (1.6 kPa) 131 – 132 °C, d_4^{20} 0.8760, n_D^{20} 1.4418, is a liquid with a sweet-fruity note and is used in perfumery for fruity-floral nuances.

Citronellyl isovalerate [*68922-10-1*], $C_{15}H_{28}O_2$, M_r 240.39, *bp* (4 kPa) 194 – 196 °C, d_4^{20} 0.8790, n_D^{20} 1.4429, has a heavy, rosy-herbal odor and is used in oriental perfume compositions among others.

Citronellyl tiglate [*24717-85-9*] (with *trans*-$CH_3CH{=}C(CH_3)COOH$ as the acid component), $C_{15}H_{26}O_2$, M_r 238.37, *bp* (900 Pa) 144 – 145 °C, d_4^{20} 0.9043, is a liquid with a flowery rosy, fruity, mushroom-like odor. It is used in geranium oil reconstitutions.

2.2.6. Miscellaneous Compounds

The number of nitrogen- and sulfur-containing derivatives of acyclic terpenoids that are known to be important fragrance and flavor substances is even smaller than in the non-terpenoid aliphatic series discussed in Section 2.1.6. However, a few nitriles are used in rather large amounts in soap perfumes because of their relatively high stability toward alkali.

Geranyl nitrile [*5146-66-7*], geranonitrile, $C_{10}H_{15}N$, M_r 149.24, *bp* (1.3 kPa) 110 °C, d_4^{20} 0.8642, n_D^{20} 1.4759, occurs as a mixture of its *cis*- and *trans*-isomers. It is a liquid with a crisp-fresh, lemon-like, green odor. The nitrile can be prepared from citral by reaction with hydroxylamine and subsequent dehydration with acetic anhydride.
Trade Name. Citralva (IFF).

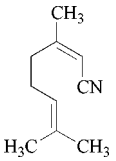

Citronellyl nitrile [*51566-62-2*], $C_{10}H_{17}N$, M_r 151.25, *bp* (2 kPa) 110 – 111 °C, d_4^{20} 0.845 – 0.846, n_D^{20} 1.4485 – 1.4500, is a colorless liquid with a strong, lemon-like odor. The nitrile can be prepared from citronellal oxime in the same way as geranic acid nitrile.
Trade Name. Agrunitril (DRAGOCO).

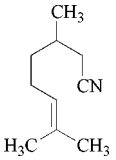

2.3. Cyclic Terpenes

2.3.1. Hydrocarbons

Cyclic terpene hydrocarbons occur in essential oils, sometimes in large amounts. They often serve as starting materials for the synthesis of fragrance and flavor compounds. By themselves they generally contribute relatively little to fragrance and aroma. They are used mainly in household perfumery and for reconstitution of essential oils.

Of the various types of monocyclic terpene hydrocarbons, those with the *p*-menthadiene structure are the most important. Examples are as follows:

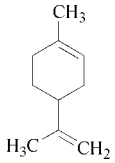
Limonene

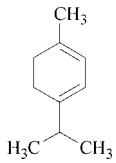
α-Terpinene

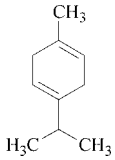
γ-Terpinene

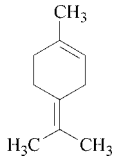
Terpinolene

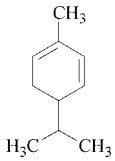
α-Phellandrene

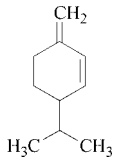
β-Phellandrene

Of the bicyclic terpene hydrocarbons, the pinenes are by far the most important industrially. Camphene and 3-carene are used as starting materials for fragrance compounds.

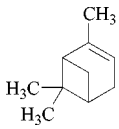
α-Pinene

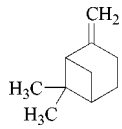
β-Pinene

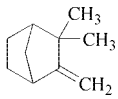
Camphene

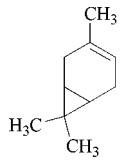
3-Carene

Many cyclic sesquiterpenes of various structural types have been isolated from essential oils. As in the case of the cyclic monoterpene hydrocarbons, a number of the cyclic sesquiterpenes are used as starting materials in the synthesis of fragrance and flavor compounds or for the reconstitution of essential oils. Typical examples are as follows:

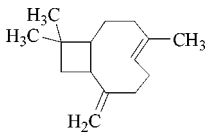
Caryophyllene

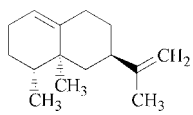
(+)-Valencene

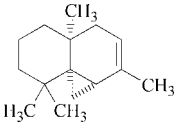
Thujopsene

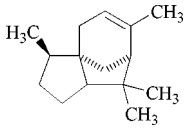
α-Cedrene

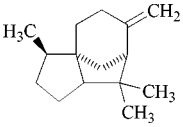
β-Cedrene

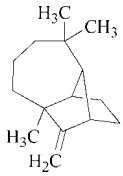
Longifolene

components of artificial essential oils. (IFRA recommendation)

Limonene [*138-86-3*], 1,8-*p*-menthadiene, $C_{10}H_{16}$, M_r 136.24, *bp* (101.3 kPa) 178 °C, d_4^{20} 0.8411, n_D^{20} 1.4726, $[\alpha]_D^{20}$ + or − 126.3 °; (+)-limonene [*5989-27-5*] and (−)-limonene [*5989-54-8*] as well as the racemate (dipentene) [*7705-14-8*] occur abundantly in many essential oils. The (+)-isomer is present in citrus peel oils at a concentration of over 90 %; a low concentration of the (−)-isomer is found in oils from the *Mentha* species and conifers.

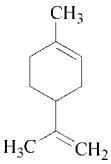

Limonene is a liquid with lemon-like odor. It is a reactive compound; oxidation often yields more than one product. Dehydrogenation leads to *p*-cymene. Limonene can be converted into cyclic terpene alcohols by hydrohalogenation, followed by hydrolysis. Nitrosyl chloride adds selectively to the endocyclic double bond; this reaction is utilized in the manufacture of (−)-carvone from (+)-limonene (see page 35).

(+)-Limonene is obtained in large amounts as a byproduct in the production of orange juice; (−)-limonene is isolated in relatively small quantities from essential oils. Racemic limonenes, which are commercially available under the name dipentene, are formed as (by)products in many acid-catalyzed isomerizations of α- and β-pinene. Distillation of the so-called dipentene fraction yields limonenes in varying degrees of purity.

The limonenes are used as fragrance materials for perfuming household products and as

γ-Terpinene [*99-85-4*], 1,4-*p*-menthadiene, $C_{10}H_{16}$, M_r 136.24, *bp* (101.3 kPa) 183 °C, d_4^{20} 0.8493, n_D^{20} 1.4747, is a colorless liquid with an herbaceous citrus odor and can be prepared by isomerization of limonene.

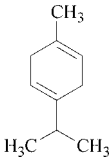

(−)-α-Phellandrene [*4221-98-1*], 1,5-*p*-menthadiene, $C_{10}H_{16}$, M_r 136.24, *bp* (99 kPa) 172 °C, d_4^{20} 0.8410, n_D^{20} 1.4708, $[\alpha]_D^{20}$ − 183 °, is a colorless liquid with a citrus odor and a slight peppery note. It is isolated, for example, from *Eucalyptus dives* oil.

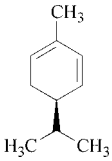

Pinenes are widespread, naturally occurring terpene hydrocarbons. The α- and β-forms occur in varying ratios in essential oils.

α-Pinene [*80-56-8*], 2-pinene, $C_{10}H_{16}$, M_r 136.24, *bp* (101.3 kPa) 156 °C, d_4^{20} 0.8553, n_D^{20} 1.4662, $[\alpha]_D^{20}$ + or − 51.9 °, is the most widespread pinene isomer. (+)-α-Pinene [*7785-70-8*] occurs, for example, in oil from *Pinus palustris*Mill. at a concentration up to 65 %; oil from *Pinus pinaster* Soland. and American oil from *Pinus caribaea* contain 70 % and 70 – 80 %, respectively of the (−)-isomer [*7785-26-4*].

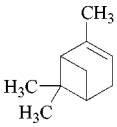
α-Pinene

α-Pinene undergoes many reactions, of which the following are used in the fragrance industry: Upon hydrogenation α-pinene is converted to pinane, which has become an important starting material in the industrial processes used

in the fragrance and flavor industry. α-Pinene can be isomerized to β-pinene with high selectivity for β-pinene formation [73]. Hydration with simultaneous ring opening yields terpineol and *cis*-terpin hydrate. Pyrolysis of α-pinene yields a mixture of ocimene and alloocimene.

Pure α-pinene is obtained by distillation of turpentine oils (see page 118). As a fragrance substance it is used to improve the odor of industrial products. However, it is far more important as a starting material in industrial syntheses, for example, terpineols (see page 33), borneol (see page 34), and camphor (see page 36).

β-Pinene [*127-91-3*], 2(10)-pinene, $C_{10}H_{16}$, M_r 136.24, *bp* (101.3 kPa) 164 °C, d_4^{20} 0.8712, n_D^{20} 1.4763, $[\alpha]_D^{20}$ + or − 22.6 °, occurs in many essential oils. Optically active and racemic β-pinenes are present in turpentine oils, although in smaller quantities than α-pinene.

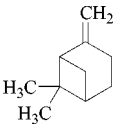

β-Pinene

β-Pinene is similar to α-pinene in its reactions. Pyrolytic cleavage of β-pinene to myrcene, the starting material for acyclic terpenes, is used on an industrial scale. Addition of formaldehyde results in the formation of nopol; nopyl acetate is used as a fragrance material. β-Pinene is produced in large quantities by distillation of turpentine oils. It is used as a fragrance material in household perfumery. However, most β-pinene is used in the production of myrcene.

2.3.2. Alcohols and Ethers

Although cyclic terpene alcohols occur widely in nature, few have the physiological properties that make them important fragrance or flavor compounds. Exceptions are α-terpineol and (−)-menthol, the latter because of its cooling/refreshing effect. Of the bicyclic monoterpene alcohols, borneol deserves mention.

Many cyclic sesquiterpene alcohols are key odor components in essential oils, for example, cedrol in cedarwood oil, the vetiverols in vetiver oil, and the santalols in sandalwood oil. Since these alcohols have not yet been synthesized on an industrial scale, they are described under the oil in which they occur (Chap. 3). Some of their derivatives, however, are discussed in this section.

Menthol, *p*-menthan-3-ol, $C_{10}H_{20}O$, M_r 156.27, has three asymmetric carbon atoms in its cyclohexane ring and, therefore, occurs as four pairs of optical isomers. The configuration of four of these isomers is given below; the other four are their mirror images.

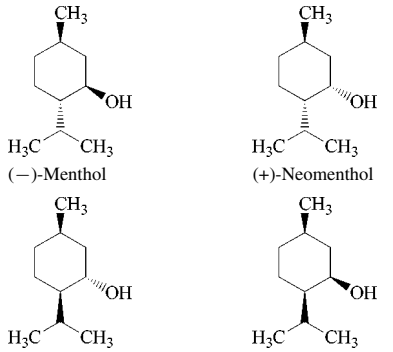

(−)-Menthol is the isomer that occurs most widely in nature. It is the main component of peppermint and cornmint oils obtained from the *Mentha piperita* and *Mentha arvensis* species. Esterified menthol also occurs in these oils (e.g., as the acetate and isovalerate). Other menthol stereoisomers may be present in these oils as well.

Physical Properties. The eight optically active menthols differ in their organoleptic properties [74]. (−)-Menthol has a characteristic peppermint odor and also exerts a cooling effect. The other isomers do not possess this cooling effect and are, therefore, not considered to be "refreshing". (±)-Menthol occupies an intermediate position; the cooling effect of the (−)-menthol present is distinctly perceptible.

The enantiomeric menthols have identical physical properties (apart from their specific rotation), but the racemates differ from the optically active forms in, for example, their melting points. Although the differences between the boiling points of the stereoisomers are small, the four racemates can be separated from each other

by fractional distillation. Boiling points (in °C at 101.3 kPa) are as follows:

Neomenthol	211.7
Neoisomenthol	214.6
Menthol	216.5
Isomenthol	218.6

Other physical constants of commercially available levorotatory and racemic menthols are: (−)-menthol [*2216-51-5*], *mp* 43 °C, n_D^{20} 1.4600, $[\alpha]_D^{20}$ −50 °; (±)-menthol [*15356-70-4*], *mp* 38 °C, n_D^{20} 1.4615.

Chemical Properties. Hydrogenation of menthols yields *p*-menthane; oxidation with chromic acid or catalytic dehydrogenation yields menthones. Dehydration under mild conditions yields 3-*p*-menthene as the main product. Reaction with carboxylic acids or their derivatives yields menthyl esters, which are used mainly as aroma substances and in pharmaceutical preparations and formulations. The esterification of menthols with benzoic acid is used on an industrial scale in the resolution of racemic menthol.

Production. Many industrial processes exist for the production of menthols. For (−)-menthol, isolation from peppermint oil (see Mint Oils, page 107) competes with partial and total syntheses. When an optically active compound is used as a starting material, optical activity must be retained throughout the synthesis, which generally consists of several steps. Total syntheses or syntheses starting from optically inactive materials require either resolution of racemic mixtures or asymmetric synthesis of an intermediate. Recently used processes are the following:

1) (−)-*Menthol from Cornmint Oil. Mentha arvensis* oils, which may contain 70–80 % free (−)-menthol, are cooled and the crystals separated by centrifugation. Since the crystalline product contains traces of cornmint oil, this menthol has a slightly herbaceous-minty note. Pure (−)-menthol is obtained by recrystallization from solvents with low boiling points.
2) (−)-*Menthol from Dementholized Cornmint Oil.* Dementholized cornmint oil, from which (−)-menthol has been removed by crystallization and which still contains 40–50 % free menthol, can be reused for producing (−)-menthol. The fairly large quantity of (−)-menthone in the oil (30–50 %) is hydrogenated to form a mixture of mainly (−)-menthol and (+)- neomenthol; the (−)-menthyl esters present (chiefly (−)-menthyl acetate) are saponified. Additional (−)-menthol is then separated from other components by crystallization, distillation, or via the boric acid esters.
3) (−)-*Menthol from* (+)-*Citronellal.* This process uses the readily occurring cyclization of citronellal to isopulegol. (+)-Citronellal can be isolated with an optical purity of ca. 80 % from citronella oil (see page 98). Alternatively, it can be synthesized with a purity of 98 % from dialkylgeranylamine (obtained from myrcene and a dialkylamine) by enantioselective isomerization to (+)-citronellaldialkylenamine followed by hydrolytic cleavage to (+)-citronellal. Isomerization is effected in the presence of a chiral rhodium–phosphine complex as a catalyst [75]. (+)-Citronellal is cyclized in the presence of acidic catalysts (e.g., silica gel) to give a mixture of optically active isopulegol isomers containing ca. 20 % of the corresponding racemates:

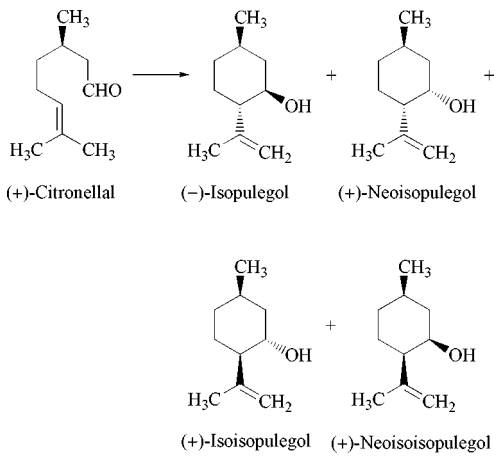

(−)-Isopulegol can be isolated from this mixture and hydrogenated to (−)-menthol. The remaining isopulegol stereoisomers can be partly reconverted into (+)-citronellal by pyrolytic cleavage and reused in the cyclization procedure [76].

The isopulegol mixture can also be hydrogenated to produce a mixture of menthols; the individual stereoisomers are then sep-

arated by distillation. To obtain optically pure (−)-menthol, a resolution step involving a suitable crystalline derivative (such as the benzoate) is required. The undesired stereoisomeric menthols, mainly (+)-neomenthol and (+)-isomenthol, are epimerized to an equilibrium mixture (e.g., by heating in the presence of sodium menthylate). (−)-Menthol is then again separated from the mixture.

4) (−)-*Menthol from* (−)-*Piperitone or Piperitol.* (−)-Menthol can also be prepared from (−)-piperitone, the main component of *Eucalyptus dives* Typus oils. Hydrogenation in the presence of Raney nickel yields a mixture of menthols, from which (−)-menthol can be separated by crystallization and saponification of its chloroacetate.
Analogously, (+)-*trans*-piperitol (obtained from α- or β-phellandrene via piperityl chloride [77]) can be hydrogenated to give a mixture of 97 % (+)-isomenthol and 3 % (+)-menthol. Pure (+)-isomenthol is obtained by crystallization and undergoes rearrangement to give an equilibrium mixture of (+)-neomenthol and (−)-menthol; the latter is separated by distillation.
5) (−)-*Menthol from* (+)-*3-Carene.* An Indian manufacturing process for (−)-menthol starts from 3-carene, the major component of Indian turpentine oil (55 – 65 %). (+)-3-Carene isomerizes to (+)-2-carene, which can be pyrolyzed to (+)-*trans*-2,8-*p*-menthadiene. Isomerization of the latter yields (+)-isoterpinolene, which is hydrogenated to give > 50 % (+)-3-*p*-menthene. Epoxidation and subsequent rearrangement lead to a menthone–isomenthone mixture, which gives a mixture of menthols when it is catalytically hydrogenated. Fractional distillation and crystallization yield commercially acceptable (−)-menthol [78].
6) (−)-*Menthol from* (±)-Menthol. (±)-Menthol can be prepared via several routes and subsequently resolved into the optical isomers:
   a) Racemic menthol can be synthesized by hydrogenation of thymol. This yields a mixture containing the four stereoisomeric menthols in various proportions. (±)-Menthol is separated from the other isomers by distillation.

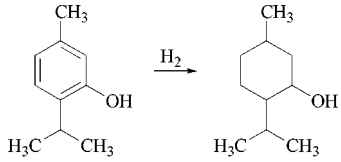

Thymol Menthol

The remaining isomeric menthols, neomenthol, isomenthol, and a trace of neoisomenthol, can be epimerized, under the conditions used for the thymol hydrogenation, to give ca. 6 : 3 : 1 equilibrium mixture of (±)-menthol, (±)-neomenthol, and (±)-isomenthol, respectively. (±)-Menthol can, again, be distilled from the equilibrium mixture.
   b) (±)-Menthol can be resolved into its optical antipodes by several routes. A large-scale industrial process utilizes selective crystallization of either (+)- or (−)-menthyl benzoate by seeding saturated solutions or supercooled melts of (±)-menthyl benzoates with crystals of (+)- or (−)-menthyl benzoate. Pure (+)- or (−)-menthol is obtained following hydrolysis of the esters [79]. The undesired (+)-menthol can be reconverted into the racemate. Biochemical resolution methods have also been developed [80].

Uses. Because of its cooling and refreshing effect, (−)-menthol is used in large quantities in cigarettes, cosmetics, toothpastes, chewing gum, sweets, and medicines. (±)-Menthol can be used in medicines and liniments.

Isopulegol, 8-*p*-menthen-3-ol, $C_{10}H_{18}O$, M_r 154.25, pure (−)-isopulegol [*89-79-2*], *bp* (1 kPa) 74 °C, d_4^{26} 0.9062, n_D^{26} 1.4690, $[\alpha]_D^{26}$ − 23.6 °. Like menthol, isopulegol has three asymmetric carbon atoms and, therefore, four stereoisomers, each occurring as a pair of optically active antipodes.

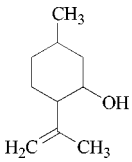

The isopulegols occur in a large number of essential oils, often in optically active or part-

ly racemic form. Since citronellal readily cyclizes to isopulegol, the latter occurs frequently in citronellal-containing essential oils, in which it is formed during the recovery of the oil.

Isopulegol produced industrially from (+)-citronellal is a mixture of isomers containing a high percentage of (−)-isopulegol. The isopulegols are colorless liquids with a minty-herbaceous odor. They are converted into the corresponding menthols by means of hydrogenation. Cyclization of citronellal, in the presence of acidic catalysts, yields a mixture of isomeric isopulegols; (+)-citronellal obtained synthetically or from citronella oil is most frequently used as the starting material.

Isopulegol is used in perfumery in various blossom compositions, as well as for geranium notes. It is an important intermediate in (−)-menthol production.

Terpineols are unsaturated monocyclic terpene alcohols and are formed by acid-catalyzed hydration of pinenes; α-, β-, γ-, and δ-isomers exist:

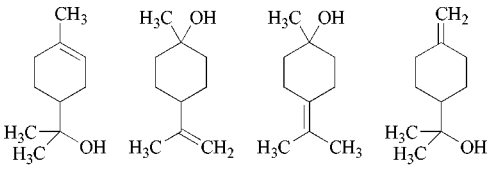

α- and β-Terpineol occur in optically active forms and as racemates. α-Terpineol is an important commercial product. It occurs in a large number of essential oils, primarily as (−)-α-terpineol (for example, in conifer and lavandin oils). Small quantities of (+)- and (±)-α-terpineol are found in many other essential oils; β-, γ-, and δ-terpineol do not occur widely in nature.

α-Terpineol [*98-55-5*], 1-*p*-menthen-8-ol, $C_{10}H_{18}O$, M_r 154.25, *mp* (pure enantiomers) 40–41 °C, *mp* (racemate) 35 °C, *bp* (101.3 kPa) 218–219 °C, d_4^{20} 0.9357, n_D^{20} 1.479, $[\alpha]_D$ +106.4° (solution in ether, 4 %) is a colorless, crystalline solid, smelling of lilac. The most important commercial grade of terpineol consists of a liquid mixture of isomers, that contains mainly α-terpineol and a considerable amount of γ-terpineol. This mixture has a stronger lilac odor than does pure crystalline α-terpineol.

Hydrogenation of α-terpineol yields *p*-menthan-8-ol. Terpineol is readily dehydrated by acids, yielding a mixture of unsaturated cyclic terpene hydrocarbons. Under mildly acidic conditions, terpin hydrate is formed. The most important reaction for the fragrance industry is esterification, particularly acetylation to terpinyl acetate.

Production. Although α-terpineol occurs in many essential oils, only small quantities are isolated, e.g., by fractional distillation of pine oils.

A common industrial method of α-terpineol synthesis consists of the hydration of α-pinene or turpentine oil with aqueous mineral acids to give crystalline *cis*-terpin hydrate (*mp* 117 °C), followed by partial dehydration to α-terpineol. Suitable catalysts are weak acids or acid-activated silica gel [81].

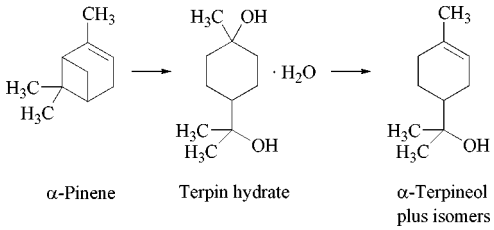

Selective conversion of pinene, 3-carene, and limonene or dipentene to terpineol, without terpin hydrate formation, is also used. Addition of organic acids (weak acids require catalytic amounts of mineral acids) produces terpinyl esters, which are subsequently hydrolyzed to terpineol, sometimes in situ.

Uses. Terpineol with its typical lilac odor is one of the most frequently used fragrance compounds. It is stable and inexpensive, and is used in soaps and cosmetics.

1-Terpinen-4-ol [*562-74-3*], 1-*p*-menthen-4-ol, $C_{10}H_{18}O$, M_r 154.25, *bp* (101.3 kPa) 212 °C, *bp* (500 Pa) 73.5 °C, d^{20} 0.9315, n_D^{20} 1.4799, occurs as (+)-, (−)-, and racemic 1-terpinen-4-ol in many essential oils (e.g., from *Pinus* and *Eucalyptus* species) and in lavender oil. It is a colorless liquid with a spicy, nutmeg-like, woody-earthy, and also lilac-like odor.

1-Terpinen-4-ol is a byproduct in the synthesis of terpineol from terpin hydrate, and occurs in commercial terpineol. Pure 1-terpinen-4-ol can be prepared from terpinolene by photosensitized oxidation, reduction of the resulting 1-methyl-4-isopropenyl-1-cyclohexene-4-hydroperoxide, and selective hydrogenation of the corresponding alcohol [82].

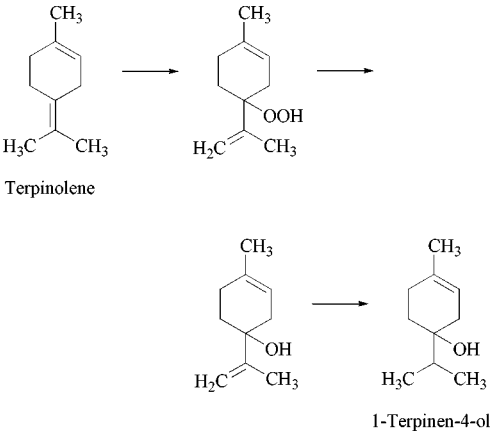

It is used, for example, in artificial geranium and pepper oils and in perfumery for creating herbaceous and lavender notes.

Borneol [*507-70-0*], 2-bornanol, $C_{10}H_{18}O$, M_r 154.25, *mp* (pure enantiomers) 208 °C, *mp* (racemate) 210.5 °C, d_4^{20} 1.011, $[\alpha]_D^{20}$ + or − 37.7 °, is a bicyclic terpene alcohol. Borneol is an *endo* isomer; the corresponding *exo* isomer is isoborneol [*124-76-5*]:

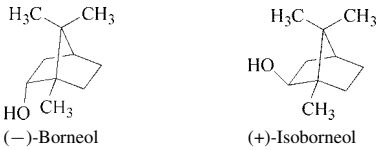

Borneol occurs abundantly in nature as a single enantiomer or, less frequently, as the racemate. (−)-Borneol [*464-45-9*] occurs particularly in oils from Pinaceae species and in citronella oil. (+)-Borneol [*464-43-7*] is found, for example, in camphor oil (Hon-Sho oil), in rosemary, lavender, and olibanum oils.

Borneol is a colorless, crystalline solid. (+)-Borneol has a camphoraceous odor, with a slightly sharp, earthy-peppery note, which is less evident in (−)-borneol. Commercial borneol is often levorotatory ($[\alpha]_D^{20}$ − 18 to − 28 ° in ethanol), and contains (−)-borneol and up to 40 % isoborneol.

Borneol is oxidized to camphor with chromic or nitric acid; dehydration with dilute acids yields camphene. Borneol is readily esterified with acids, but on an industrial scale bornyl esters are prepared by other routes. For example, levorotatory borneol is synthesized industrially from levorotatory pinenes by Wagner-Meerwein rearrangement with dilute acid, followed by hydrolysis of the resulting esters [83].

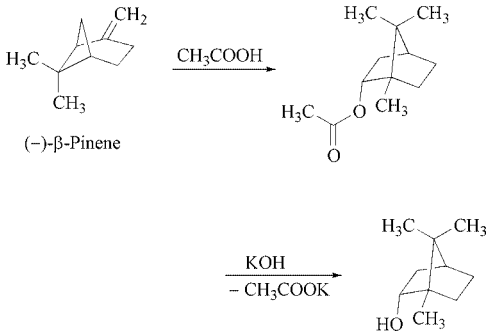

Borneol is used in the reconstitution of the essential oils in which it occurs naturally.

Cedryl methyl ether [*19870-74-7*], 1*H*-3a,7-methanoazulene, octahydro-6-methoxy-3,6,8,8-tetramethyl-(6*β*), and [*67874-81-1*], 1*H*-3a,7-methanoazulene, octahydro-6-methoxy-3,6,8,8-tetramethyl-(6*α*), $C_{16}H_{18}O$, M_r 236.40, d_{25}^{25} 0.974 – 0.979, n_D^{20} 1.494 – 1.498, is a colorless liquid with a fine cedarwood odor and a distinct amber nuance. It is prepared by methylation of cedrol and is used in perfumes, soaps, and cosmetics.
Trade Name. Cedramber (IFF).

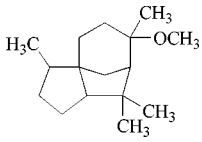

2.3.3. Aldehydes and Ketones

Cyclic terpene aldehydes occur in essential oils only in low concentration. These aldehydes are seldom used as single fragrance compounds.

A few of the cyclic terpene ketones are commercially important as fragrance and flavor compounds, for example, menthone and carvone, which have the *p*-menthane skeleton, and the ionones, which have a (trimethylcyclohexenyl)alkenone skeleton. The ionones and their methyl-substituted homologues are some of the most valuable fragrance materials. Some cyclic terpene ketones are the main components of essential oils (e.g., camphor in camphor oil); others, although not main components, may be essential for a fragrance (e.g., β-damascenone, which is an important component of Bulgarian rose oil; see Rose Oils, page 114). The cyclic sesquiterpene ketone, nootkatone, is one of the characteristic components of grapefruit aroma.

Menthone, *p*-menthan-3-one, $C_{10}H_{18}O$, M_r 154.25, exists as two stereoisomers, menthone and isomenthone, each of which occurs as a pair of enantiomers, due to the two asymmetric centers present in the molecule.

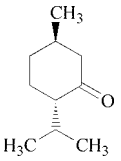

(−)-Menthone

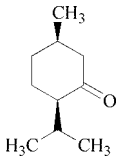

(+)-Isomenthone

Both stereoisomers occur in many essential oils, often as a single enantiomer species. A particularly high concentration (sometimes $>50\,\%$) is found in oils from *Mentha* species. The menthones are colorless liquids that possess a typically minty odor; the odor of isomenthone is slightly musty. They have a strong tendency to interconvert and are, therefore, difficult to obtain in high purity. Industrial products are mixtures of varying composition. Physical constants of industrially important menthone isomers are listed in Table 1.

The menthones are converted into the corresponding menthols by hydrogenation; for example, (−)-menthone yields (+)-neomenthol and (−)-menthol. (−)-Menthone can be obtained by distillation of dementholized cornmint oil or by oxidation of (−)-menthol (e.g., with chromic acid). Dehydrogenation of (−)-menthol (e.g., with copper chromite) yields a mixture of (−)-menthone and (+)-isomenthone. (±)-Menthone is prepared analogously from (±)-menthol. However, it can also be synthesized by hydrogenation of thymol in the presence of palladium–carbon catalysts [84].

Menthone and isomenthone are used for synthetic peppermint oils and bases.

Carvone, 1,8-*p*-menthadien-6-one, $C_{10}H_{14}O$, M_r 150.22, *bp* (101.3 kPa) 230 °C, d_4^{20} 0.960, n_D^{20} 1.499, $[\alpha]_D^{18}$ (+)-carvone +64.3 °, $[\alpha]_D^{20}$ (−)-carvone −62.5 °, occurs as (+)-carvone [*2244-16-8*], (−)-carvone [*6485-40-1*], or racemic carvone [*22327-39-5*]. The optical isomers differ considerably in their organoleptic properties. They occur in high percentages in a number of essential oils. (+)-Carvone is the main component of caraway oil (ca. 60 %) and dill oil; (−)-carvone occurs in spearmint oil at a concentration of 70 – 80 %.

Properties. The carvones are colorless to slightly yellow liquids. (+)-Carvone has an herbaceous odor reminiscent of caraway and dill seeds, whereas (−)-carvone has a herbaceous odor reminiscent of spearmint. Depending on the reaction conditions, hydrogenation of carvone yields either carveols or dihydrocarvone, which are also used as flavor compounds. When treated with strong acids, carvone isomerizes to carvacrol.

Production. In the past, (+)- and (−)-carvones were isolated by fractional distillation of caraway oil and spearmint oil, respectively. However, these carvones are now prepared syn-

Table 1. Physical properties of industrially important menthone isomers

Compound	CAS registry number	*bp* (101.3 kPa), °C	d_t^{20}	n_D^{20}	$[\alpha]_D$
(−)-Menthone	[*14073-97-3*]	210	0.896∗	1.450	−28.5 °
(±)-Menthone	[*1074-95-9*]	210	0.896∗	1.450	
(+)-Isomenthone	[*1196-31-2*]	212	0.900∗∗	1.453	+95.0 °
(±)-Isomenthone	[*36977-92-1*]	212	0.900∗∗	1.453	

$*t = 20\,°C$; $**t = 4\,°C$

thetically, the preferred starting material being (+)- and (−)- limonenes, which are converted into the corresponding optically active carvones. Since optical rotation is reversed in the process, (+)-limonene is the starting material for (−)-carvone.

The preferred industrial method of carvone synthesis utilizes the selective addition of nitrosyl chloride to the endocyclic double bond of limonene. If a lower aliphatic alcohol is used as solvent, limonene nitrosochloride is obtained in high yield. It is converted into carvone oxime by elimination of hydrogen chloride in the presence of a weak base. Acid hydrolysis in the presence of a hydroxylamine acceptor, such as acetone, yields carvone [85].

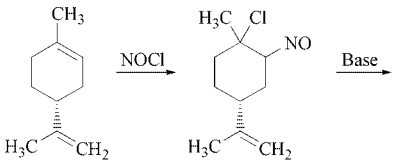

(+)-Limonene (+)-Limonene nitrosochloride

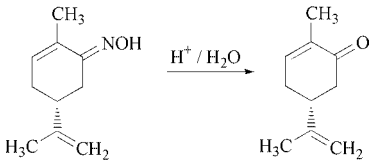

(−)-Carvone oxime (−)-Carvone

Uses. Both (+)- and (−)-carvone are used to flavor a number of foods and beverages. (−)-Carvone is produced in much larger quantities and is also used in oral hygiene products.

Camphor [*21368-68-3*], 1,7,7-trimethylbicyclo[2.2.1]heptan-2-one, $C_{10}H_{16}O$, M_r 152.24, (+)-camphor: *bp* (101.3 kPa) 204 °C, *mp* 178.8 °C, $[\alpha]_D^{20}$ +44.3 °; both optical isomers are found widely in nature, (+)-camphor [*464-49-3*] being the more abundant. It is, for example, the main component of oils obtained from the camphor tree *Cinnamomum camphora.*

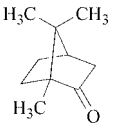

(+)-Camphor

Camphor is produced by fractional distillation and crystallization of camphor oil (see page 90) or, synthetically, by dehydrogenation of isoborneol (from isobornyl acetate, see page 41) over a copper catalyst.

Due to its characteristic penetrating, slightly minty odor, camphor is only used in perfuming industrial products. It is far more important as a plasticizer.

Fenchone [*1195-79-5*], 1,3,3-trimethylbicyclo[2.2.1]heptan-2-one, $C_{10}H_{16}O$, M_r 152.24, *bp* (101.3 kPa) 193 °C, d_4^{20} 0.9484, n_D^{20} 1.4628, $[\alpha]_D^{20}$ + or − 66.8 °, occurs as its (−)-isomer in a number of fennel oils. It is a colorless, slightly viscous liquid with a camphoraceous odor.

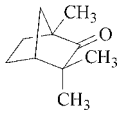

(+)-Fenchone

(+)-Fenchone [*7787-20-4*] containing a small amount of the (−)-isomer [*4695-62-9*] is prepared by dehydrogenation of (−)-fenchol. (−)-Fenchyl esters are obtained, along with other compounds, by addition of carboxylic acids to α-pinene. Hydrolysis of the esters yields (−)-fenchol.

Fenchone is used to prepare artificial fennel oils and to perfume household products.

Ionones and Homologous Compounds. The C_{13} ketones α- and β-ionone are cyclic terpenoids that occur in many essential oils. However, being metabolites of the corresponding carotenoids [86], they occur in only small amounts. A third isomer, γ-ionone, has not yet been observed in nature.

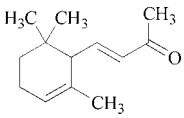

α-Ionone

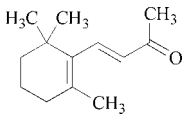

β-Ionone

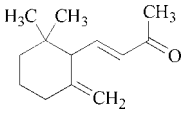

γ-Ionone

Both optical isomers of α-ionone are found in nature. Generally, ionones have a *trans* config-

uration. *trans*-α-Ionone can be converted into the *cis*-isomer by exposure to ultraviolet light. Under the same conditions, *trans*-β-ionone rearranges to the retro compound.

The *irones* are ionone homologues that have an additional methyl group adjacent to the twin methyl groups in the cyclohexane ring. The number of possible irone isomers is larger than that of the ionones due to the additional methyl group on the ring. Some of these irone isomers occur in essential oils from the roots of *Orris* species (see Orris Root Oil, page 111).

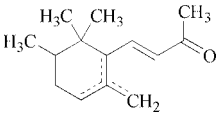

Irone, α-, β-, γ-

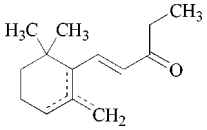

n-Methylionone, α-, β-, γ-

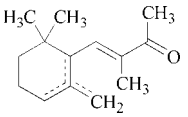

Isomethylionone, α-, β-, γ-

Other ionone homologues are the *methylionones,* in which the oxoalkenyl group carries an extra methyl substituent. The methylionones also exist as α-, β-, and γ-isomers, each of which can occur in the *cis* or *trans* form; the isomers may also be optically active. Their natural occurrence is debated [87].

All ionones, irones, and methylionones, as well as the corresponding pseudocompounds (their synthetic acyclic precursors) are slightly viscous yellowish liquids. Commercial irones and methylionones are mixtures of isomers that are named according to their main component. Their composition varies with the method used to prepare and cyclize the pseudocompound and fluctuates considerably between different manufacturers.

Physical and Chemical Properties. Physical and odor properties of the best known ionones are listed in Table 2.

β-Ionone is converted into intermediates for vitamin A synthesis (→ Vitamins, Chap. 2.7.2.). The hydrogenation of ionones and methylionones is of some importance. Dihydro or tetrahydro derivatives or ionols can be obtained depending on reaction conditions.

Table 2. Physical properties of ionones

Name	Formula	CAS registry number	M_r	*bp* °C (*p*, kPa)	d_4^{20}	n_D^{20}	Odor
α-Ionone	$C_{13}H_{20}O$	[*127-41-3*]	192.30	121 – 122 (1.3)	0.9319	1.4982	sweet-floral, reminiscent of violet
β-Ionone	$C_{13}H_{20}O$	[*79-77-6*]	192.30	121.5 (0.93)	0.9461	1.5202	reminiscent of cedarwood, violet-like upon dilution
γ-Ionone	$C_{13}H_{20}O$	[*79-76-5*]	192.30	82 (0.16)	0.9317	1.4985	violet-like with woody-resinous tonality (intermediate in the synthesis of γ-dihydroionone, a component of ambergris)
α-Irone	$C_{14}H_{22}O$	[*79-69-6*]	206.33	109 (0.36)	0.9340	1.4998	responsible for the fragrance of natural orris oil
β-Irone	$C_{14}H_{22}O$	[*79-70-9*]	206.33	108 – 109 (0.21)	0.9465	1.5183	reminiscent of β-ionone, but slightly more intense
α-*n*-Methylionone	$C_{14}H_{22}O$	[*127-42-4*]	206.33	97 (0.35)	0.9210 [a]	1.4938 [b]	reminiscent of α-ionone, but milder and more delicate
β-*n*-Methylionone	$C_{14}H_{22}O$	[*127-43-5*]	206.33	102 (0.35)	0.9370	1.5155	β-ionone-like, but with a distinct leather note
α-Isomethylionone	$C_{14}H_{22}O$	[*127-51-5*]	206.33	130 – 131 (1.3)	0.9345	1.5019	reminiscent of orris and violets, possesses the finest odor of all ionones
β-Isomethylionone	$C_{14}H_{22}O$	[*79-89-0*]	206.33	94 (0.4)	0.9376	1.5053	interesting, powdery, orris-like odor with slightly woody aspects
Allylionone [c]	$C_{16}H_{24}O$	[*79-78-7*]	232.35	102 – 104 (0.02)	0.9289	1.5040	floral violet odor with a fruity pineapple note and high tenacity

[a] d_4^{25}
[b] n_D^{25}
[c] Obtained by cyclization of allylpseudoionone (from citral and allylacetone)

With Raney nickel–copper chromite catalysts, methylionones are converted into tetrahydromethylionols, which are also used as fragrance materials [88].

Production. Ionones, irones, and methylionones, as well as allylionone, are all produced by analogous routes. Special procedures must be used to obtain a particular isomer, either pure or as the main component. These are described where appropriate.

In all processes an acyclic precursor, called a pseudoionone, pseudoirone, etc., is prepared by base-catalyzed condensation of citral or 6-methylcitral with acetone, methyl ethyl ketone, or allylacetone, as appropriate.

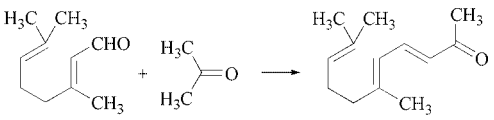

In the synthesis of vitamin A, the dependence on natural sources as well as steadily increasing production via β-ionone as an intermediate have led to the development of a method for synthesizing citral from dehydrolinalool (see page 21). More recent routes employ dehydrolinalool as the starting material for pseudoionone. Dehydrolinalool is converted into pseudoionone by using either diketene [89] or a suitably substituted acetoacetate (Carroll reaction) [90]:

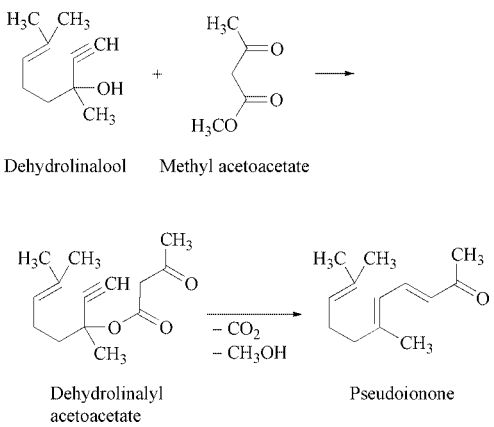

A milder reaction for synthesizing pseudoionone from dehydrolinalool is transetherification with an alkoxyalkene [91]:

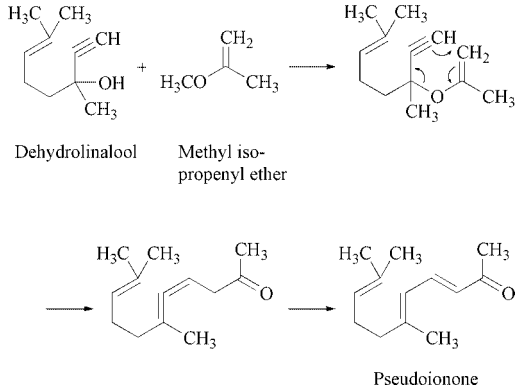

In the methylionone synthesis, condensation of citral with methyl ethyl ketone results in a mixture of *n*-methyl- and isomethylpseudoionone, each of which may occur as one of four possible *cis–trans*-isomers.

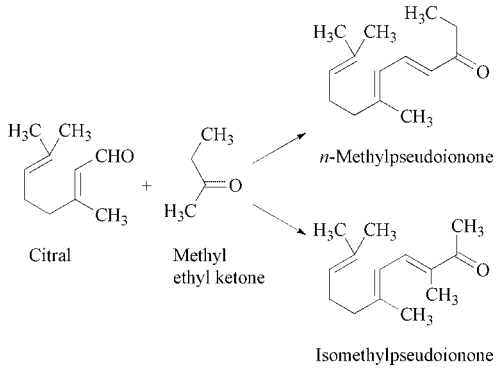

The ratio of the major isomers in the mixture depends on the condensation catalyst and the reaction conditions. In the presence of common alkaline catalysts (e.g., sodium hydroxide), straight-chain isomers are formed preferentially. Strongly alkaline catalysts, such as quaternary ammonium bases, favor the formation of isomethylpseudoionone [92]. This compound is a precursor for the highly valued fragrance substance α-isomethylionone and can be obtained as the main component by reacting dehydrolinalool with the enol ether of methyl ethyl ketone and methanol [93]. Acidic and Lewis catalysts are employed in the cyclization of the pseudocompounds to the cyclic ketones. The primary cyclization products are α-ionone or its homologues, which are isomerized to the β-compounds by strong acids. Concentrated sul-

furic acid converts pseudoionone almost exclusively into β-ionone; 85 % phosphoric acid yields α-ionone in ca. 80 % purity. γ-Ionone can be obtained together with a small amount of α- and β-ionone when boron trifluoride etherate is used as the catalyst and dimethylformamide as the solvent [94]; γ-ionone is of little commercial importance. Since α- and β-ionone can be separated on an industrial scale by fractional distillation in high-performance columns, other methods of separation are seldom used.

Uses. The volume of the production of β-ionone, which serves as an intermediate in vitamin A synthesis, shows that it is by far the most important. The ionones and their homologues are components of blossom and phantasy perfume compositions. The ionones and irones are used in aroma compositions as well, although on a much smaller scale. α-Ionone is a highly valued fragrance material. The methylionones are among the most important fragrance substances, α-isomethylionone being the most important. The irones, isomers of the methylionones, are produced in limited quantities, mainly due to their high cost. Likewise, allylionone is manufactured in small amounts.

Damascones, 1-(2,6,6-trimethylcyclohexenyl)-2-buten-1-ones, $C_{13}H_{20}O$, M_r 192.30, are ionone isomers. Damascone exists in α-, β-, γ-, δ-, and ε-forms, depending on the position of the double bond in the cyclohexane ring. Commercially important are α-damascone [*43052-87-5*], β-damascone ([*23726-91-2*] for (E) and δ-damascone [*57378-68-4*] as mixtures of their *cis*- and *trans*-isomers.

The α- and β-damascones have been identified as components of tea aroma. They are liquids with a fruity, roselike odor. One synthetic route to the damascones starts with an appropriate cyclogeranic acid derivative (halide, ester, etc.). This is reacted with an allyl magnesium halide to give 2,6,6-trimethylcyclohexenyl diallyl carbinol, which on pyrolysis yields the desired 1-(2,6,6-trimethylcyclohexenyl)-3-buten-1-one. Damascone is otained by rearrangement of the double bond in the side-chain [95].

The damascones are used in perfume compositions, especially rose perfumes, and in flavor compositions, to which they impart naturalness and body. (IFRA recommendation)

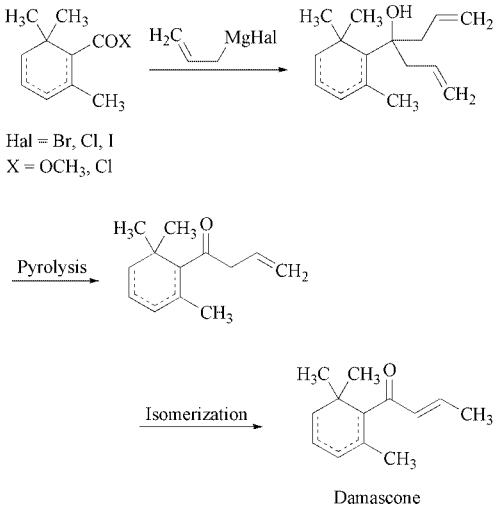

***β*-Damascenone** [*23726-93-4*], 1-(2,6,6-trimethyl-1,3-cyclohexadienyl)-2-buten-1-one, $C_{13}H_{18}O$, M_r 190.28, is a constituent of Bulgarian rose oil, which has a powerful fragrance although it is only present at a concentration of 0.05 %.

It is used in small quantities in flavor and perfume compositions to impart naturalness and brilliance. (IFRA recommendation)

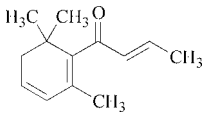

Nootkatone [*4674-50-4*], 5,6-dimethyl-8-isopropenylbicyclo[4.4.0]-1-decen-3-one, $C_{15}H_{22}O$, M_r 218.33, *mp* 35 °C, has been isolated from grapefruit peel and juice and identified in other citrus oils as well. The commercially available product is a colorless to yellowish liquid with a typical grapefruit odor.

Nootkatone can be prepared by oxidation of valencene, a sesquiterpene hydrocarbon isolated from orange oils. (IFRA recommendation)

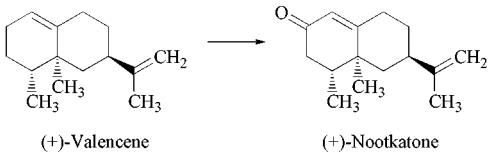

Nootkatone is used for flavoring beverages.

Cedryl methyl ketone, $C_{17}H_{26}O$, M_r 246.39, is a long-lasting wood fragrance which is prepared by acetylation of cedarwood oil fractions that contain sesquiterpene hydrocarbons, mainly α-cedrene and thujopsene (for structures, see Section 2.3.1). Acetylation is carried out in the presence of an acidic catalyst (e.g., polyphosphoric acid). Commercially available cedryl methyl ketone is a multicomponent mixture. The main component is 1-(2,6,6,8-tetramethyltricyclo[5.3.1.0^{1,5}]undec-8-en-9-yl)ethan-1-one [*32388-55-9*]. One of its odor-determining compounds is 1-(2,2,8-trimethyltricyclo[6.2.2.0^{1,6}]dodec-5-en-5-yl)ethan-1-one [*32388-56-0*]:

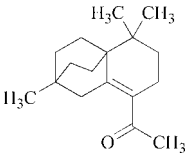

Trade Names. Lignofix (Dragoco), Lixetone (Quest), Vertofix (IFF).

2.3.4. Esters

Esters derived from cyclic terpene alcohols, especially the acetates, are common fragrance and flavor components. Menthanyl, menthenyl, bicyclic bornyl acetates, and a few acetates of sesquiterpene alcohols are extensively used in perfume and aroma compositions.

(−)-Menthyl acetate [*2623-23-6*], (−)-*p*-menthan-3-yl acetate, $C_{12}H_{22}O_2$, M_r 198.30, *bp* (3 kPa) 116 °C, d_4^{20} 0.9253, n_D^{20} 1.4467 – 1.4468, $[\alpha]_D^{20}$ − 81.1 °, occurs in peppermint oils. It is a colorless liquid with a fresh-fruity, peppermint odor.

(−)-Menthyl acetate is prepared by acetylation of (−)-menthol (e.g., with acetic anhydride). It is used mainly in peppermint flavors and reconstituted peppermint oils, but also to a small extent in perfumery.

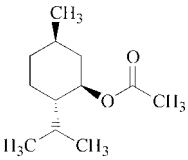

(±)-Menthyl acetate [*29066-34-0*] occurs in essential oils. It is synthesized by esterification of racemic menthol. Its odor is crisper and less fruity than that of (−)-menthyl acetate. It is used for essential oil compositions and occasionally in household perfumery.

***p*-Menthanyl acetate,** $C_{12}H_{22}O_2$, M_r 198.30, *bp* (200 Pa) 67 – 70 °C, d_{25}^{25} 0.931 – 0.937, n_D^{20} 1.446 – 1.451, is commercially available as a *cis–trans* mixture of *p*-menthan-1-yl acetate [*26252-11-9*] (**1**) and *p*-menthan-8-yl acetate [*80-25-1*] (**2**). It is a colorless liquid with a citrus-fresh, pine-needle odor and a secondary, slightly herbaceous note. Menthanyl acetate is produced by hydrogenation of terpinyl acetates (mixtures of isomers) (e.g., in the presence of Raney nickel [96]) or by esterification of a mixture of isomeric *p*-menthanols.

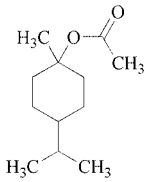

1

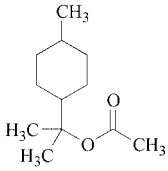

2

The ester mixture is highly stable and is, therefore, used in perfumery for detergents and other household products.

α-Terpinyl acetate [*80-26-2*], 1-*p*-menthen-8-yl acetate, $C_{12}H_{20}O_2$, M_r 196.29, *bp* (5.3 kPa) 140 °C, d_4^{20} 0.9659, n_D^{21} 1.4689, $[\alpha]_D^{20}$ of the enantiomers + or − 79 °; the enantiomers and the racemate occur in many essential oils (e.g., Siberian pine needle oil and cypress oil), but generally not as the main component. Pure α-terpinyl acetates are colorless liquids with a fresh bergamot-lavender odor. Commercially available terpinyl acetate consists mainly of α-terpinyl acetate, but also contains a number of other isomeric compounds. It can be prepared by acetylating the terpineol mixture obtained from terpin hydrate, using a customary procedure for tertiary alcohols.

Because of its odor properties, stability, and low price, large quantities of terpinyl acetate are used in perfumery for lavender and bergamot types, as well as in essential oil reconstitutions.

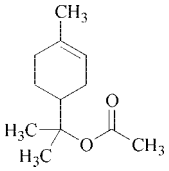

Nopyl acetate [*35836-72-7*], (−)-2-(6,6-dimethylbicyclo[3.1.1]hept-2-en-2-yl)ethyl acetate, $C_{13}H_{20}O_2$, M_r 208.30, *bp* (1.5 kPa) 122 °C, d_4^{20} 0.9811, n_D^{20} 1.4733, $[\alpha]_D$ −30.9° (CAS registry no. of racemate [*128-51-8*], is not found in nature and has a fresh-fruity-woody odor. (−)-Nopyl acetate is prepared by acetylation of (−)-nopol with acetic anhydride. (−)-Nopol is obtained from (−)-β-pinene and paraformaldehyde in a Prins reaction. Nopyl acetate is used in perfumes for soap and household products.

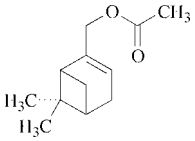

Bornyl acetate, 2-*endo*-bornanyl acetate, $C_{12}H_{20}O_2$, M_r 196.29, *bp* (101.3 kPa) 223–224 °C, d_4^{20} 0.9838, n_D^{20} 1.4630, $[\alpha]_D^{20}$ + or −44.4°, *mp* (+)- and (−)-form 29 °C, occurs in its optically active forms and as a racemate in many essential oils.

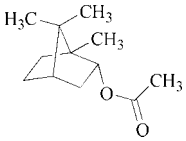

(−)-Bornyl acetate

(−)-Bornyl acetate [*5655-61-8*] is a characteristic component of most conifer oils. It has a camphoraceous, pine-needle-like odor. Both (+)-bornyl acetate [*20347-65-3*] and (−)-bornyl acetate form colorless crystals; the racemate [*36386-52-4*] is a colorless liquid. Bornyl acetate is prepared by esterification of borneol with acetic anhydride or via the process described under borneol (see page 34).

Due to its characteristic pine needle odor, bornyl acetate is frequently used in conifer needle compositions, soap, bath products, room sprays, and pharmaceutical products.

Isobornyl acetate [*125-12-2*], 2-*exo*-bornanyl acetate, $C_{12}H_{20}O_2$, M_r 196.29, *bp* (1.6 – 1.7 kPa) 102 – 103 °C, d_4^{20} 0.9841, n_D^{20} 1.4640, has been identified in a number of essential oils. It is a colorless liquid with a pleasant, pine-needle odor. Isobornyl acetate is prepared from camphene and acetic acid in the presence of acidic catalysts (e.g., sulfuric acid) [97], or ion-exchange resins (e.g., sulfonated styrene–divinylbenzene polymers) [98].

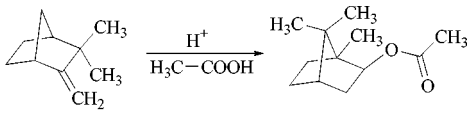

(−)-Camphene (+)-Isobornyl acetate

Isobornyl acetate is used in large amounts for perfuming soap, bath products, and air fresheners. However, the major use of isobornyl acetate is as an intermediate in the production of camphor.

Guaiyl acetate, $C_{17}H_{28}O_2$, M_r 264.41, *bp* (300 Pa) 118 – 123 °C, d_{25}^{25} 0.965 – 0.990, n_D^{20} 1.489 – 1.495, is obtained by esterification of guaiac wood oil with acetic anhydride and consists of the acetates of the natural sesquiterpene alcohols guaiol and bulnesol (see Guaiac Wood Oil, page 103). Guaiyl acetate is a yellow to amber-yellow liquid with a weak but lasting, woody odor. It is used in perfumery for tea-rose and wood nuances.

Cedryl acetate [*77-54-3*], $C_{17}H_{28}O_2$, M_r 264.41, *bp* (400 Pa) 146 – 150 °C, d_{25}^{25} 0.966 – 1.012, n_D^{20} 1.495 – 1.506, occurs in cedarwood oils. The pure compound is crystalline (*mp* 80 °C). Commercial cedryl acetate is a colorless to amber-yellow liquid, with a cedarwood-like odor. It is prepared by esterification of the cedrol-rich fraction from cedarwood oil and is used in perfumery for wood and leather notes, and as a fixative.

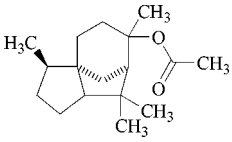

Vetiveryl acetate [*62563-80-8*], *bp* (300 Pa) 125 – 128 °C, d_{25}^{25} 0.979 – 1.015, n_D^{20} 1.5050 – 1.5180, is not a single compound; its main component is khusimyl acetate [*61474-33-7*]. Vetiveryl acetate is prepared by esterification of the sesquiterpene alcohols isolated from vetiver oils.

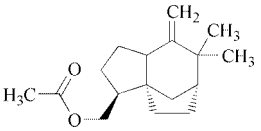

Khusimyl acetate

Vetiveryl acetate is a light yellow liquid with a dry, fresh-woody odor. It is a popular fragrance mixture that is frequently used in luxury perfumery; it is also used as a fixative in many fine fragrances. (Limiting IFRA recommendation)

2.3.5. Miscellaneous Compounds

Of the few known terpene compounds that contain heteroatoms such as nitrogen or sulfur, the thiol 8-mercapto-*p*-menthan-3-one described below has qualitatively important applications as a fragrance and flavor substance. The second thiol, 1-*p*-menthene-8-thiol, is described because its odor threshold value is far lower than that of most other fragrance and flavor compounds.

8-Mercapto-*p*-menthan-3-one [*38462-22-5*], $C_{10}H_{18}OS$, M_r 186.31, *bp* (13 Pa) 57 °C, is an essential odoriferous constituent of buchu leaf oil. The commercial product is synthesized from pulegone and is a liquid mixture of *cis*–*trans*-isomers with a typical blackcurrant odor. 8-Mercapto-*p*-menthan-3-one is prepared by reacting pulegone with hydrogen sulfide in the presence of a base (e.g., triethylamine) [99]:

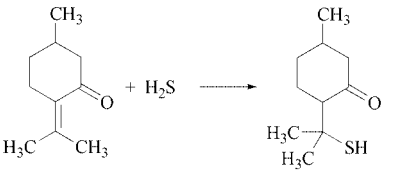

Pulegone 8-Mercapto-*p*-menthan-3-one

This powerful fragrance and flavor substance is used in perfume and aroma compositions.

1-*p*-Menthene-8-thiol [*71159-90-5*], $C_{10}H_{18}S$, M_r 170.31, *bp* (45 Pa) 40 °C, d_4^{20} 0.948, n_D^{20} 1.503, has been identified in grapefruit juice. It is a liquid with an extremely powerful, obnoxious odor; when diluted it has the typical aroma of fresh grapefruit juice. Its odor threshold value is extremely low: 2×10^{-5} µg/kg for the (+)-R and 8×10^{-5} µg/kg for the (−)-S isomer [100].

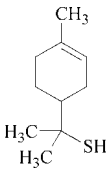

2.4. Other Cycloaliphatic Compounds

In addition to cyclic terpenoids, several other cycloaliphatic compounds have above-average importance as fragrance materials; some of them are structurally related to the terpenes. Ketones are most widely represented and include cyclopentanone derivatives, such as the jasmin fragrance compounds, and cyclic ketones with 15 – 17-membered carbon rings, such as muscone [*541-91-3*] and civetone [*542-46-1*], which are constituents of the extremely expensive animal products, musk and civet.

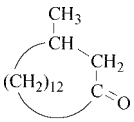

Muscone

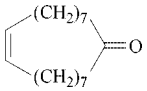

Civetone

Many syntheses have been devised for these macrocyclic ketones. However, due to the many steps involved and poor yields, they are not used widely on an industrial scale. Cyclopentadecanone, a natural musk fragrance, and the unsaturated 5-cyclohexadecen-1-one, which has not yet been found in nature, are more easily synthesized. Their odor characteristics are similar to those of muscone and civetone and are, therefore, used instead of them.

A few alicyclic aldehydes are valuable perfume materials and are obtained by Diels – Alder

reactions using terpenes and acrolein. Esters derived from hydrogenated aromatic compounds, such as *tert*-butylcyclohexyl and decahydro-β-naphthyl acetates, are also used in large amounts as fragrance materials.

2.4.1. Alcohols

Some alicyclic alcohols are important as synthetic sandalwood fragrances. The annual world production of natural sandalwood oil is about 100 t with a current price of about $ 400 per kilogram. Due to the limitations of the harvest and this highly fluctuating price, chemists have been searching for cheap synthetic substitutes for this important perfumery ingredient. Since the 1970s a new class of sandalwood products has been developed, which all originate from campholenic aldehyde, an easily available product of the rearrangement of α-pinene epoxide, catalyzed by zinc chloride or zinc bromide.

5-(2,2,3-Trimethyl-3-cyclopenten-1-yl)-3-methylpentan-2-ol [*65113-99-7*] (**3**), $C_{14}H_{26}O$, M_r 210.36, d_4^{20} 0.896 – 0.904, n_D^{20} 1.470 – 1.476, is a fragrance ingredient that does not occur in nature. It is prepared by condensation of campholenic aldehyde with methyl ethyl ketone, followed by selective hydrogenation of the resulting unsaturated ketone [101].

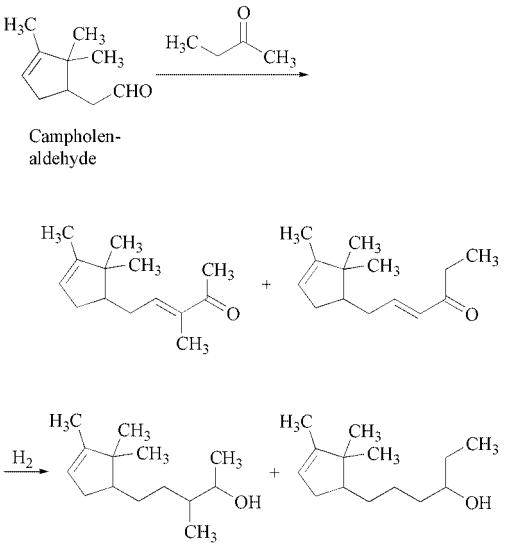

Compound **3** can be used either in a pure state or as a mixture with its byproduct **4** in perfume compositions and soap perfumes.
Trade Name. Sandalore (Givaudan Roure).

2-Ethyl-4-(2,2,3-trimethyl-3-cyclopenten-1-yl)-2-buten-1-ol [*8219-61-6*], $C_{14}H_{26}O$, M_r 208.35, *bp* (3 kPa) 114 – 116 °C, d_4^{20} 0.9160 – 0.9168, n_D^{20} 1.4870 – 1.4880, is another sandelwood-like fragrance also derived from campholenic aldehyde. Aldol condensation of this aldehyde with butyric aldehyde gives an α,β-unsaturated carbonyl compound that can be reduced to an unsaturated alcohol.

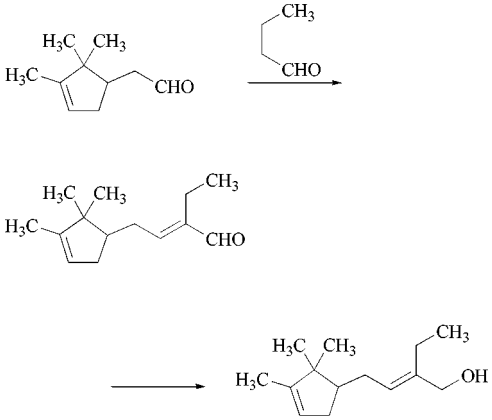

Trade Names. Bacdanol (IFF), Bangalol (Quest), Sandolene (Haarmann & Reimer), Sandranol (DRAGOCO).

In addition some other materials of this type are on the market:

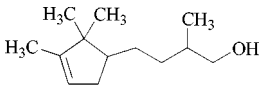

Trade Name. Brahmanol (DRAGOCO).

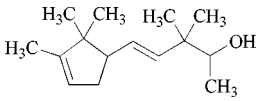

Trade Name. Polysantol (Firmenich).

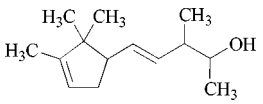

Trade Name. Ebanol (Givaudan Roure).

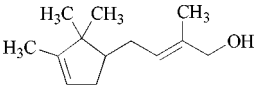

Trade Names. Madrol (DRAGOCO), Sandelice (Cognis).

3-*trans*-Isocamphylcyclohexanol [*3407-42-9*], $C_{16}H_{28}O$, M_r 236.40, does not occur in nature. A synthetic mixture of terpenylcyclohexanol isomers has a sandalwood odor. A commercially available mixture containing 3-*trans*-isocamphylcyclohexanol is prepared by reacting camphene and guaiacol in the presence of an acidic catalyst (e.g., boron trifluoride), followed by catalytic hydrogenation of the resulting terpenylguaiacols. In the alkylation reaction, camphene rearranges to the isobornyl, isofenchyl, and isocamphyl skeletons. These substituents may be introduced in guaiacol at four positions. In the subsequent hydrogenation with simultaneous elimination of the methoxy group, additional possibilities for isomerism arise because the hydroxyl group may be either axial or equatorial to the terpenyl moiety. Therefore, the actual content of the desired isomer, 3-*trans*-isocamphylcyclohexanol, is low in most products. The other isomers are either weak in odor or odorless.

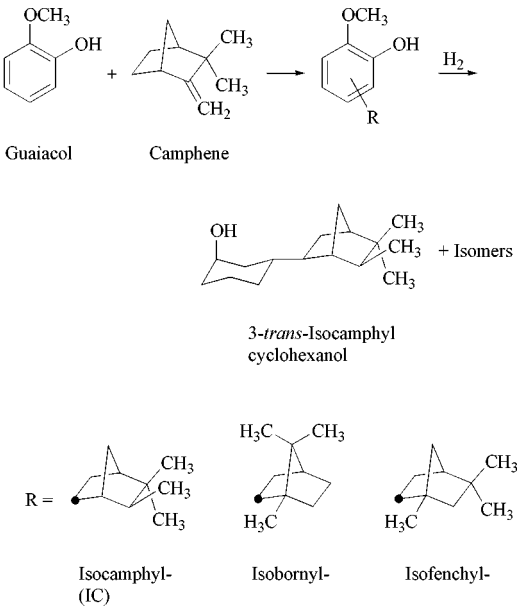

A process starting from catechol, instead of guaiacol, yields a mixture with a higher content of 3-*trans*-isocamphylcyclohexanol [102]. Moreover, the process starting from guaiacol has been improved by converting the main component formed in the first step, *p*-isocamphylguaiacol (**5**), into a mixed ether (**6**) by reaction with diethyl sulfate. Nucleophilic cleavage of the ether with alcoholates or Grignard reagents results in a high yield of *m*-isocamphylguethol (**7**), which is hydrogenated catalytically, with concomitant loss of the ethoxy group, to give a high yield of 3-*trans*-isocamphylcyclohexanol (**8**) [103].

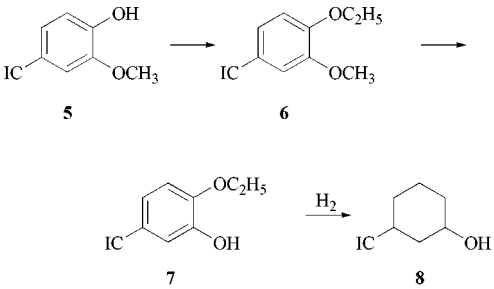

IC = Isocamphyl

The mixture is used as such in large amounts as a replacement for sandalwood oil in soaps, cosmetics, and perfume compositions.

Trade Names. Sandela (Givaudan Roure), Sandel H+R (Haarmann & Reimer), Sandiff (IFF), Rhodanthial (Rhodia), Santalex (Takasago).

2.4.2. Aldehydes

2,4-Dimethyl-3-cyclohexene carboxaldehyde [*68039-49-6*], $C_9H_{14}O$, M_r 138.21, *bp* (4 kPa) 94–96 °C, n_D^{25} 1.4696, is prepared as a mixture of its *cis*- and *trans*-isomers by a Diels–Alder reaction of 2-methyl-1,3-pentadiene and acrolein. It is a liquid with a strongly green, slightly herbaceous, citrus note. It is used for perfuming cosmetic preparations as well as household products.

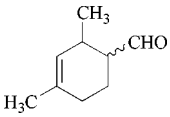

Trade Names. CyclalC (Givaudan Roure), Ligustral (Quest), Triplal (IFF), Vertocitral (Haarmann & Reimer).

4-(4-Methyl-3-penten-1-yl)-3-cyclohexene carboxaldehyde [*37677-14-8*], $C_{13}H_{20}O$, M_r 192.30, is prepared, together with its 3-isomer, by a Diels – Alder reaction of myrcene and acrolein. The mixture has a fresh-fruity, slightly citrus-like odor and is used to perfume household products.

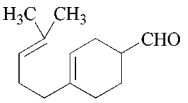

Trade Names. Empetal (Quest), Myrac aldehyde (IFF), Vertomugal (Haarmann & Reimer).

4-(4-Hydroxy-4-methylpentyl)-3-cyclohexene carboxaldehyde [*31906-04-4*] (**9**), $C_{13}H_{22}O_2$, M_r 210.32, *bp* (130 Pa) 120 – 122 °C, d_4^{20} 0.9941, n_D^{20} 1.4915, is a fragrance substance that does not occur in nature. It is a colorless, viscous liquid with a sweet odor reminiscent of lily of the valley. The aldehyde can be prepared by a Diels – Alder reaction of myrcenol and acrolein in the presence of a Lewis catalyst (e.g., zinc chloride) [104]:

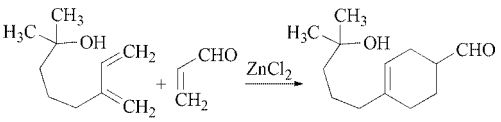

Myrcenol Acrolein **9**

Reaction of myrcenol with acrolein at elevated temperatures, without a catalyst, yields a 70 : 30 mixture of the 4- and 3-substituted cyclohexene carboxaldehydes [105]. This mixture is a commercial product. Compound **9** has excellent fixative properties and is used especially in soap and cosmetics perfumery.
Trade Names. Lyral (IFF), Kovanol (Takasago).

2.4.3. Ketones

2-Heptylcyclopentanone [*137-03-1*], $C_{12}H_{22}O$, M_r 182.31, *bp* (1.3 kPa) 130 °C, d^{20} 0.890, n_D^{20} 1.4530, is a fragrance material with a fruity, slightly herbaceous, jasmin odor; it has not yet been found in nature. It is a colorless, viscous liquid prepared by condensation of cyclopentanone with heptanal to give 2-heptylidenecyclopentanone, with subsequent hydrogenation of the double bond.

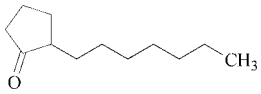

2-Heptylcyclopentanone and its lower mass homologues 2-pentyl- and 2-hexylcyclopentanone are used in, for example, jasmin, honeysuckle, and lavender compositions.
Trade Names. Fleuramone (IFF), Projasmon P (Haarmann & Reimer).

Dihydrojasmone [*1128-08-1*], 3-methyl-2-pentyl-2-cyclopenten-1-one, $C_{11}H_{18}O$, M_r 166.26, *bp* (2.7 kPa) 87 – 88 °C, d^{25} 0.9157, n_D^{25} 1.4771, is a colorless, slightly viscous liquid with a typical jasmin odor, resembling that of the naturally occurring *cis*-jasmone. Dihydrojasmone can be synthesized by various routes. A preferred method is intramolecular aldol condensation of 2,5-undecanedione, which can be prepared from heptanal and 3-buten-2-one in the presence of a thiazolium salt, such as 5-(2-hydroxyethyl)-4-methyl-3-benzylthiazolium chloride [106]:

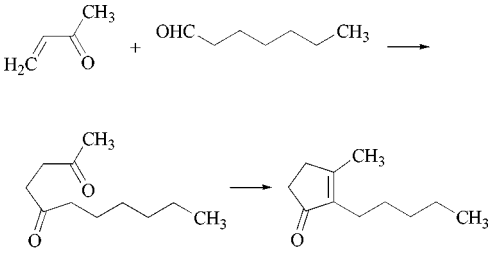

Dihydrojasmone

Dihydrojasmone is used in perfumery in jasmin bases and, more generally, in blossomy and fruity fragrances.

***cis*-Jasmone** [*488-10-8*], 3-methyl-2-(2-*cis*-penten-1-yl)-2-cyclopenten-1-one, $C_{11}H_{16}O$, M_r 164.25, *bp* (1.6 kPa) 78 – 79 °C, d^{20} 0.9423, n_D^{20} 1.4989, occurs in jasmin absolute and contributes to its typical jasmin odor. It is a pale yellow, viscous liquid with a strong jasmin odor. Various stereospecific syntheses for *cis*-jasmone have been reported. A patented method involves alkylation of 3-methyl-2-cyclopenten-1-one with *cis*-2-pentenyl chloride in an alkaline

medium in the presence of a phase-transfer catalyst (e.g., tricaprylmethylammonium chloride) [107]:

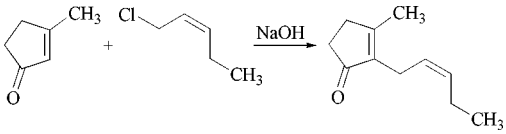

cis-Jasmone

cis-Jasmone is used in perfumery in fine jasmin bases and floral compositions.

1-(5,5-Dimethylcyclohex-1-en-1-yl)pent-4-en-1-one [*56973-85-4*], and 1-(3,3-dimethylcyclohex-1-en-1-yl)pent-4-en-1-one [*56973-84-3*], $C_{13}H_{20}O$, M_r 192.30, d_4^{20} 1.490, n_{20}^{20} 0.930. The isomeric mixture is a colorless to pale yellow liquid with an extremely powerful and diffusive odor reminiscent of galbanum, with a pineapple and hyacinth character.

The cyclic ketone can be formed by dehydratization of dehydrolinalool (**10**), subsequent alkylation of **11** with allyl chloride to form **12**, acid-catalyzed cyclization of (**12**) and hydrolysis of the product **13** to give a mixture of **14** and **15** [108].

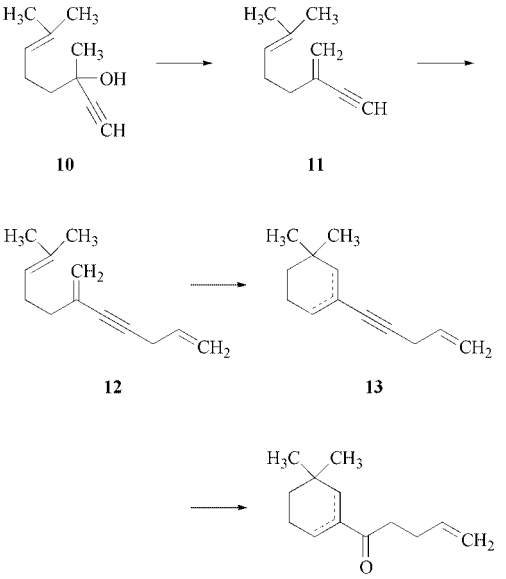

Trade Name. Dynascone (Firmenich).

1,2,3,5,6,7-Hexahydro-1,1,2,3,3-pentamethyl-4*H*-inden-4-one [*33704-61-9*], $C_{14}H_{22}O$, M_r 206.32, d_4^{20} 0.962, n_D^{20} 1.500, has not yet been found in nature. It is colorless to pale yellow solid to semi-solid mass with a diffusive, spicy, musk-like odor with strong floral reinforcement and powdery, velvet nuances.

The indenone **18** can be prepared by hydrogenation of pentamethylindan (**16**) and subsequent oxidation of the intermediate **17** with air over copper acetate to yield the corresponding ketone **18** [109, 110].

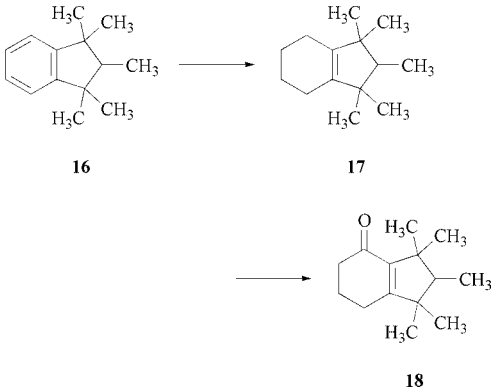

Trade Name. Cashmeran (IFF).

Cyclopentadecanone [*502-72-7*], $C_{15}H_{28}O$, M_r 224.39, *mp* 65 – 67 °C, *bp* (7 Pa) 85 °C, is a musk fragrance found in the scent gland of the male civet cat.

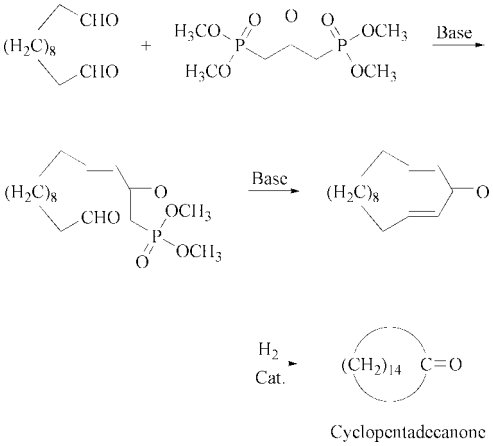

A number of syntheses have been developed for its manufacture. In a more recent method,

1,12-dodecanedial is reacted with 1,3-bis(dimethylphosphono)propan-2-one in the presence of a base, preferably in two steps via the intermediates 15-(dimethylphosphono)pentadec-12-en-14-on-1-al and 2,14-cyclopentadecadien-1-one. Subsequent hydrogenation yields cyclopentadecanone [111].

Cyclopentadecanone is used in fine fragrances.

Trade Name. Exalton (Firmenich).

3-Methylcyclopentadec-4/5-en-1-one, (*E*/*Z*) [*82356-51-2*], $C_{16}H_{28}O$, M_r 236.40, d_{20}^{20} 0.925 – 0.935, has a powerful musk odor, reminiscent of musk ketone.

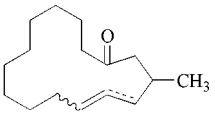

The synthesis of muscenone starts with a condensation of cyclododecanone and dimethyl methylsuccinate, followed by cyclization with polyphosphoric acid. Hydrogenation and elimination yields methylbicyclo[10.3.0]-pentadec-1(12)-ene. Ozonolyses results in a diketone, which is converted via hydrogenation and elimination to the unsaturated muscone [112].

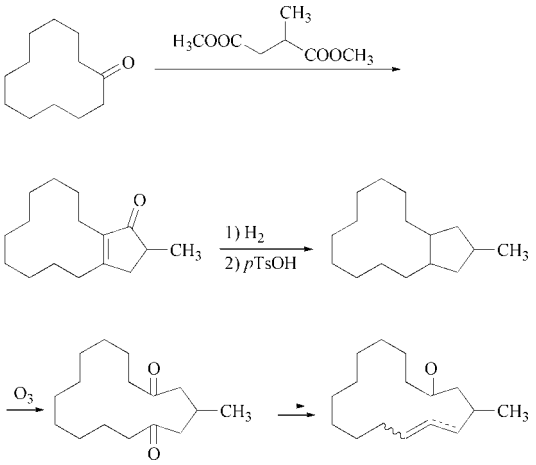

3-Methylcyclopentadec-4/5-en-1-one is used as nitromusk replacer, at low dosage (0.05 – 0.5 %) as a fragrance enhancer, at higher dosage it is "used for its musky character in addition to its boosty effect".

Trade Name. Muscenone (Firmenich).

5-Cyclohexadecen-1-one [*37609-25-9*], $C_{16}H_{28}O$, M_r 236.40, *bp* (10 Pa) 121 °C, n_D^{25} 1.4865, is commercially available as a 40 : 60 mixture of its *cis*- and *trans*-isomers. It is a colorless liquid with an intense musk odor.

A three-step synthesis starts from cyclododecanone (**19**). Reaction with chlorine gives 2-chlorocyclododecanone (**20**) which is reacted with 2 mol of vinylmagnesium chloride to give 1,2-divinylcyclododecan-1-ol (**21**). This is finally converted into 5-cyclohexadecen-1-one by an oxy-Cope rearrangement [113]:

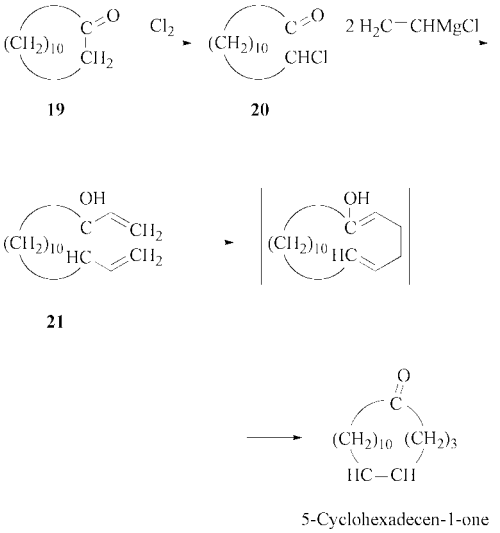

5-Cyclohexadecen-1-one can be added to perfume compositions as a substitute for the natural macrocyclic ketone musks.

Trade Name. TM II SP (Soda Aromatic), Ambretone (Takasago).

2,3,8,8-Tetramethyl-1,2,3,4,5,6,7,8-octahydro-2-naphthalenyl methyl ketone [*54464-57-2*] (**22**), $C_{16}H_{26}O$, M_r 234.38, *bp* (370 Pa) 134 – 135 °C, n_D^{20} 1.4975 – 1.500, is a synthetic amber fragrance. It is prepared by a Diels – Alder reaction of myrcene and 3-methyl-3-penten-2-one in the presence of aluminum chloride and cyclization of the substituted cyclohexenyl methyl ketone intermediate with phosphoric acid. Some 3-acetyl isomer is also formed [114].

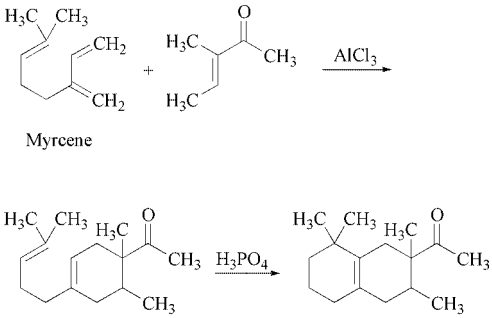

22

The ketone is used in perfume bases for soaps, eaux de cologne, and detergent compositions.
Trade Name. Iso E Super (IFF).

3-Methyl-2-cyclopenten-2-ol-1-one [*80-71-7*], $C_6H_8O_2$, M_r 112.13, *mp* (monohydrate) 106 °C, occurs in beechwood tar and has a caramel-like odor. It has been identified as a flavor component in food. Crystals of the compound usually contain 1 mol of water. Synthetic routes of production are of limited importance in comparison with isolation from beechwood tar.

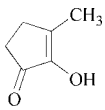

The compound is frequently used in flavor compositions for its caramel note, e.g., in beverages and in confectionery. It is rarely used in perfumery, and then mainly as an intensifier.

2.4.4. Esters

2-*tert*-Butylcyclohexyl acetate [*88-41-5*], $C_{12}H_{22}O_2$, M_r 198.30, *mp* (pure *cis*-isomer) 34.5 – 35.4 °C; commercial product: d_{25}^{25} 0.938 – 0.944, n_D^{20} 1.4500 – 1.4560; this compound does not occur in nature and exists in *cis* and *trans* forms. Pure 2-*cis*-*tert*-butylcyclohexyl acetate is a crystalline solid, with a fruity, agrumen-like odor. The commercial product is a colorless liquid and consists of a mixture of *cis* and *trans*-isomers, which contains 60 – 95 % of the *cis* ester. With an increasing percentage of the *trans*-isomer, the odor becomes more woody-camphory.

The acetate is prepared by esterification of 2-*tert*-butylcyclohexanol, which is obtained from 2-*tert*-butylphenol. It is highly stable and is used for perfuming soap as well as bath and household products.

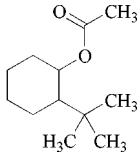

Trade Names. Agrumex (Haarmann & Reimer), Verdox (IFF).

4-*tert*-Butylcyclohexyl acetate [*32210-23-4*], $C_{12}H_{22}O_2$, M_r 198.30, *bp* (1 kPa) 89 – 95 °C, d_{25}^{25} 0.933 – 0.939, n_D^{20} 1.450 – 1.454, does not occur in nature and exists in *cis* and *trans* forms. The *trans*-isomer has a rich, woody odor, while the odor of the *cis*-isomer is more intense and more flowery. Considerable variations in *cis*–*trans* ratios in commercial mixtures have little effect on the physical constants. Therefore, the composition of mixtures should be determined by gas chromatography.

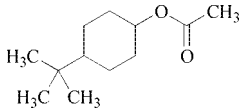

The ester is prepared by catalytic hydrogenation of 4-*tert*-butylphenol followed by acetylation of the resulting 4-*tert*-butylcyclohexanol [115]. If Raney nickel is used as the catalyst, a high percentage of the *trans*-isomer is obtained. A rhodium–carbon catalyst yields a high percentage of the *cis*-isomer. The *trans* alcohol can be isomerized by alkaline catalysts; the lower-boiling *cis* alcohol is then removed continuously from the mixture by distillation [116].

4-*tert*-Butylcyclohexyl acetate is used particularly in soap perfumes.
Trade Names. 4-*tert*-Butylcyclohexylacetate (Degussa), Oryclon extra (Haarmann & Reimer), Vertenex (IFF).

Decahydro-β-naphthyl acetate [*10519-11-6*], $C_{12}H_{20}O_2$, M_r 196.29, *bp* (650 Pa) 120 °C, d^{25} 1.005 – 1.015, n_D^{25} 1.475 – 1.482, is a fragrance substance that does not occur in nature. It consists of a mixture of several stereoisomers and is a colorless liquid with a sweet-fruity-herbaceous odor and a slight jasmin note.

Decahydro-β-naphthyl acetate is prepared by esterification of technical-grade decahydro-β-naphthol (e.g., with acetic anhydride). The acetate is used for perfuming household products.

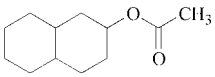

4,7-Methano-3a,4,5,6,7,7a-hexahydro-5(or 6)-indenyl acetate [*2500-83-6*] (or [*5413-60-5*]), $C_{12}H_{16}O_2$, M_r 192.26, *bp* (1 kPa) 119 – 121 °C, d_4^{25} 1.0714, n_D^{25} 1.4935, is a colorless liquid with a herbal, fresh-woody odor. It consists of a mixture of isomers that is obtained by addition of acetic acid to dicyclopentadiene in the presence of an acid catalyst. It is used for perfuming soaps, detergents, and air fresheners.

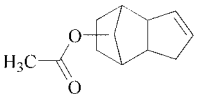

Trade Names. Cyclacet (IFF), Herbaflorat (Haarmann & Reimer), Verdylacetate (Givaudan Roure).

Allyl 3-cyclohexylpropionate [*2705-87-5*], $C_{12}H_{20}O_2$, M_r 196.29, *bp* (130 Pa) 91 °C, has not yet been found in nature. It is a colorless liquid with a sweet-fruity odor, reminiscent of pineapples. The ester is prepared by esterification of 3-cyclohexylpropionic acid (obtained by hydrogenation of cinnamic acid) with allyl alcohol. It is used in perfumery to obtain fruity top notes as well as pineapple and chamomile nuances. (Limiting IFRA recommendation)

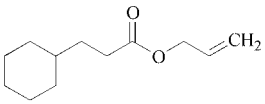

Methyl dihydrojasmonate [*24851-98-7*], methyl (3-oxo-2-pentylcyclopentyl)acetate, $C_{13}H_{22}O_3$, M_r 226.32, *bp* (1 Pa) 85 – 90 °C, $d_4^{21.6}$ 1.003, $n_D^{20.2}$ 1.4589, is a jasmin fragrance that is closely related to methyl jasmonate, which occurs in jasmin oil. Methyl dihydrojasmonate has been identified in tea. It is a liquid with a typical fruity, jasmin-like blossom odor. Methyl dihydrojasmonate is prepared as *cis*/*trans*-isomeric mixture by Michael addition of malonic esters to 2-pentyl-2-cyclopenten-1-one (**23**), followed by hydrolysis and decarboxylation of the resulting 2-pentyl-3-oxocyclopentyl malonate (**24**), and esterification of the (2-pentyl-3-oxocyclopentyl)acetic acid (**25**) [117]. Dealkoxycarbonylation of **24** can also be accomplished directly with water at elevated temperature [118].

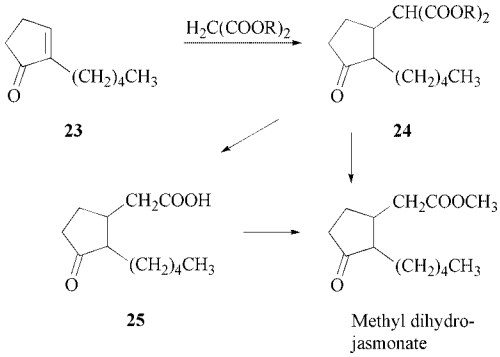

Methyl dihydrojasmonate is used in perfumery for blossom fragrances, particularly in jasmin types.
Trade Names. Hedione (Firmenich).

***cis*-Methyldihydrojasmonate** [*24851-98-7*], methyl (3-oxo-2-pentylcyclopentyl)acetate, $C_{13}H_{22}O_3$, M_r 226.32, *bp* (7 Pa) 180 °C, d_{20}^{20} 1.001 – 1.009, has an extremely powerful hedionic, jasminic odor.

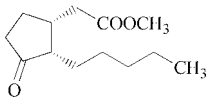

Although *cis*-methyldihydrojasmonate is the minor isomer of hedione (see page 49), it is the much more intensive one. Its odor threshold is about 20 times lower than that of the *trans*-isomer. Several attempts have been made to synthesize the pure *cis*-isomer. One approach is heating of the *trans*-isomer with sodium carbonate followed by fractional distillation [119, 120]:

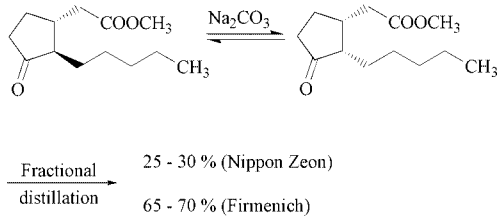

A *cis* : *trans* ratio of 96 : 4 was obtained by catalytic asymmetric hydrogenation of the acid in the presence of the chiral ruthenium catalyst DuPhos (**26**):

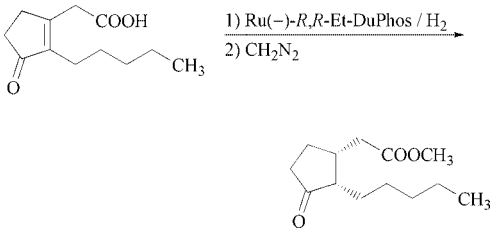

Due to its thermodynamic instability *cis*-methyldihydrojasmonate is only stable in media having a pH value of 5.5 to 6.5.
Trade Names. Hedione HC (Firmenich), Cepionate (Nippon Zeon).

Methyl 2-hexyl-3-oxocyclopentanecarboxylate [*37172-53-5*] (**27**), $C_{13}H_{22}O_3$, M_r 226.32, *bp* (8 Pa) 85 °C, n_D^{20} 1.4562, is a colorless liquid with a long-lasting, floral, jasmin-like odor, that has only little of the fatty aspect characteristic of many jasmin fragrances. The product has not yet been found in nature.

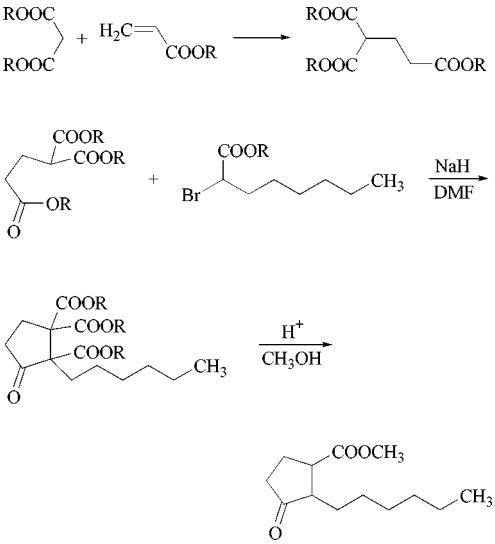

27

Compound **27** can be prepared by condensing an alkyl α-bromocaprylate with a trialkyl propane-1,1,3-tricarboxylate to give a substituted cyclopentanone. Hydrolysis, decarboxylation, and esterification of the resulting monocarboxylic acid with methanol yields the desired ester [121]. Trialkyl propane-1,1,3-tricarboxylates can be prepared by Michael addition of dialkyl malonates to alkyl acrylates.

Compound **27** is used in perfumery in floral compositions.
Trade Name. Jasmopol (Haarmann & Reimer).

2.5. Aromatic Compounds

2.5.1. Hydrocarbons

A few alkyl- and aralkyl-substituted aromatic hydrocarbons find limited use in perfumery. Examples include *p*-cymene [*99-87-6*], which is a component of many essential oils and when pure has a weak, citrus odor, as well as diphenylmethane [*101-81-5*] which has an odor like geranium:

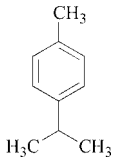

p-Cymene

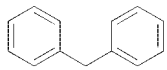

Diphenylmethane

2.5.2. Alcohols and Ethers

Phenethyl alcohol is qualitatively and quantitatively one of the most important fragrance substances that belongs to the class of araliphatic alcohols. Its lower homologue (benzyl alcohol) and higher homologue (dihydrocinnamic alcohol) also have characteristic odor properties, but are more frequently used in the form of their esters. Cinnamic alcohol, the most important unsaturated araliphatic alcohol, is valuable for both fragrances and flavors.

The araliphatic alcohols mentioned above occur in many natural fragrances and flavors, but are generally not the main components. These alcohols are nearly always prepared synthetically for use in compositions. The branched-chain homologues of phenethyl and dihydrocinnamic alcohols (dimethyl benzyl carbinol and dimethyl phenethyl carbinol, respectively) are used in fairly large amounts as fragrance materials, but have not been found in nature.

Benzyl alcohol [*100-51-6*], C_7H_8O, M_r 108.14, *bp* (101.3 kPa) 205.4 °C, d_4^{20} 1.0419, n_D^{20} 1.5396, occurs in many essential oils and foods. It is a colorless liquid with a weak, slightly sweet odor. For further details on the chemical properties and synthesis of benzyl alcohol, see→ Benzyl Alcohol, Chap. 3.

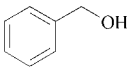

Because of its relatively weak odor, benzyl alcohol is used in fragrance and flavor compositions mainly as a solvent and for dilution. It is the starting material for a large number of benzyl esters, which are important fragrance and flavor substances.

Phenethyl alcohol [*60-12-8*], 2-phenylethyl alcohol, $C_8H_{10}O$, M_r 122.17, *bp* (101.3 kPa) 219.8 °C, d_4^{20} 1.0202, n_D^{20} 1.5325, is the main component of rose oils obtained from rose blossoms. It occurs in smaller quantities in neroli oil, ylang-ylang oil, carnation oil, and geranium oils. Since the alcohol is rather soluble in water, losses occur when essential oils are produced by steam distillation.

Properties. Phenethyl alcohol is a colorless liquid with a mild rose odor. It can be dehydrogenated catalytically to phenylacetaldehyde and oxidized to phenylacetic acid (e.g., with chromic acid). Its lower molecular mass fatty acid esters as well as some alkyl ethers, are valuable fragrance and flavor substances.

Production. Many synthetic methods are known for preparing phenethyl alcohol; the following are currently of industrial importance:

Friedel–Crafts Reaction of Benzene and Ethylene Oxide.
In the presence of molar quantities of aluminum chloride, ethylene oxide reacts with benzene to give an addition product, which is hydrolyzed to phenethyl alcohol:

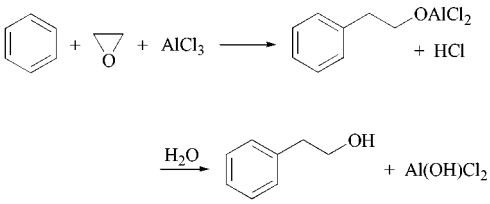

Phenethylalcohol

Formation of byproducts, such as 1,2-diphenylethane, is largely avoided by using an excess of benzene and low temperature. Special purification procedures are required to obtain a pure product that is free of chlorine and suitable for use in perfumery.

Hydrogenation of Styrene Oxide.
According to a patent, excellent yields of phenethyl alcohol are obtained when styrene oxide is hydrogenated at low temperature, using Raney nickel as a catalyst and a small amount of sodium hydroxide [122].

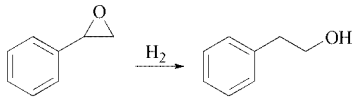

Uses. Phenethyl alcohol is used frequently and in large amounts as a fragrance compound. It is a popular component in rose type compositions, but it is also used in other blossom notes. It is stable to alkali and, therefore, ideally suited for use in soap perfumes.

Phenethyl methyl ether [*3558-60-9*], $C_9H_{12}O$, M_r 136.19, *bp* (94.6 kPa) 185–186 °C, d^{27} 0.9417, n_D^{24} 1.4970, is a colorless liquid with a sharp, rosy-green odor. It is used in oriental type perfumes as well as in artificial keora oil.

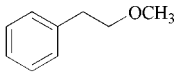

Phenethyl isoamyl ether [*56011-02-0*], $C_{13}H_{20}O$, M_r 192.30, d_{25}^{25} 0.901–0.904, n_D^{20} 1.481–1.484, is a colorless liquid with a green, sweet-flowery odor of chamomile blossoms and a secondary, soapy note; it is used in perfumes.

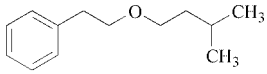

Trade Name. Anther (Quest).

1-Phenylethyl alcohol [*98-85-1*], styrallyl alcohol, $C_8H_{10}O$, M_r 122.17, *mp* 20 °C, *bp* (101.3 kPa) 203 °C, d_4^{20} 1.0135, n_D^{20} 1.5275, has been identified as a volatile component of food (e.g., in tea aroma and mushrooms). The alcohol is a colorless liquid with a dry, roselike odor,

slightly reminiscent of hawthorn. It can be prepared by catalytic hydrogenation of acetophenone. 1-Phenylethyl alcohol is used in small quantities in perfumery and in larger amounts for the production of its esters, which are more important as fragrance compounds.

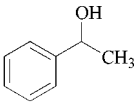

Dihydrocinnamic alcohol [*122-97-4*], 3-phenylpropanol, hydrocinnamic alcohol, $C_9H_{12}O$, M_r 136.19, *bp* (100 kPa) 237.5 °C, d_4^{20} 1.008, n_D^{20} 1.5278, occurs both in free and esterified form in resins and balsams (e.g., benzoe resin and Peru balsam). It has been identified in fruit and cinnamon.

Hydrocinnamic alcohol is a slightly viscous, colorless liquid with a blossomy-balsamic odor, slightly reminiscent of hyacinths. Esterification with aliphatic carboxylic acids is important because it leads to additional fragrance and flavor compounds.

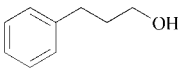

Hydrocinnamic alcohol is prepared by hydrogenation of cinnamaldehyde. A mixture of hydrocinnamic alcohol and the isomeric 2-phenylpropanol can be obtained from styrene by a modified oxo synthesis. The two isomers can be separated by distillation [123].

Hydrocinnamic alcohol is used in blossom compositions for balsamic and oriental notes.

α,α-Dimethylphenethyl alcohol [*100-86-7*], 1-phenyl-2-methyl-2-propanol, α,α-dimethyl benzyl carbinol, DMBC, $C_{10}H_{14}O$, M_r 150.22, *mp* 24 °C, *bp* (101.3 kPa) 214 – 216 °C, d_4^{20} 0.9840, n_D^{20} 1.5170, has not yet been found in nature. The alcohol has a floral-herbaceous odor, reminiscent of lilac, and is prepared by a Grignard reaction of benzylmagnesium chloride and acetone. It is used in perfumery for various flower notes (e.g., lilac, hyacinth, mimosa). The alcohol is stable to alkali and, thus, is suited for soap perfumes. It is used to prepare a number of esters, which are also used as fragrance compounds.

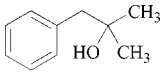

4-Phenyl-2-methyl-2-butanol [*103-05-9*], α,α-dimethylphenethyl carbinol, $C_{11}H_{16}O$, M_r 164.25, *bp* (1.9 kPa) 124 – 125 °C, $d_4^{20.7}$ 0.9626, $n_D^{20.7}$ 1.5077, is a colorless liquid with a dry-flowery, lily-like odor. It has been identified in cocoa aroma and is prepared by a Grignard reaction of benzylacetone and methylmagnesium chloride. It is used in blossom compositions.

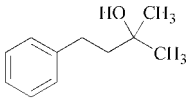

1-Phenyl-3-methyl-3-pentanol [*10415-87-9*], phenethyl methyl ethyl carbinol, $C_{12}H_{18}O$, M_r 178.28, *bp* (1.7 kPa) 129 – 130 °C, d_4^{25} 0.9582, n_D^{20} 1.509 – 1.513, has not yet been found in nature. It is a colorless liquid with a delicate peony, slightly fruity odor. Phenethyl methyl ethyl carbinol can be prepared from benzylacetone and ethylmagnesium chloride by a Grignard reaction. It is used to perfume soap, cosmetics, and detergents.

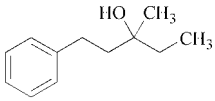

Cinnamic alcohol [*104-54-1*], 3-phenyl-2-propen-1-ol, $C_9H_{10}O$, M_r 134.18; *trans*-isomer [*4407-36-7*]: *mp* 34 °C, *bp* (101.3 kPa) 257.5 °C, d_4^{20} 1.0440, n_D^{20} 1.5819; this alcohol can exist in *cis* and *trans* forms. Although both isomers occur in nature, the *trans*-isomer is far more abundant and is present, for example, in styrax oil. *trans*-Cinnamic alcohol is a colorless, crystalline solid with a hyacinth-like balsamic odor.

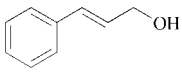

trans-Cinnamic alcohol

Cinnamic alcohol can be dehydrogenated to give cinnamaldehyde and oxidized to give cinnamic acid. Hydrogenation yields 3-phenylpropanol and/or 3-cyclohexylpropanol. Reaction with carboxylic acids or carboxylic

acid derivatives results in the formation of cinnamyl esters, some of which are used as fragrance compounds.

Production. Cinnamic alcohol is prepared on an industrial scale by reduction of cinnamaldehyde. Three methods are particularly useful:

1) In the *Meerwein–Ponndorf reduction,* cinnamaldehyde is reduced to cinnamic alcohol (yield, ca. 85 %) with isopropyl or benzyl alcohol in the presence of the corresponding aluminum alcoholate.
2) A 95 % yield of cinnamic alcohol is obtained by selective hydrogenation of the carbonyl group in cinnamaldehyde with, for example, an *osmium–carbon* catalyst [124].
3) High yields of cinnamic alcohol can be obtained by reduction of cinnamaldehyde with *alkali borohydrides.* Formation of dihydrocinnamic alcohol is thus avoided [125].

Uses. Cinnamic alcohol is valuable in perfumery for its odor and fixative properties. It is a component of many flower compositions (lilac, hyacinth, and lily of the valley) and is a starting material for cinnamyl esters, several of which are valuable fragrance compounds. In aromas, the alcohol is used for cinnamon notes and for rounding off fruit aromas. It is used as an intermediate in many syntheses (e.g., for pharmaceuticals such as the antibiotic chloromycetin). (Limiting IFRA recommendation)

2.5.3. Aldehydes and Acetals

Many of the aldehydes discussed in this section are treated in detail elsewhere (→ Aldehydes, Aliphatic and Araliphatic).

Several araliphatic aldehydes are of special commercial importance as fragrance and flavor materials. These include cinnamaldehyde and its homologues in which the side-chain carries an alkyl substituent; α-amyl- and α-hexylcinnamaldehyde are particularly important. Other important members of this group are the substituted phenylpropanals, 4-isopropyl- and 4-*tert*-butyl-α-methyldihydrocinnamaldehyde. Arylacetaldehydes and arylpropionaldehydes are, in comparison, seldom used in compositions. The corresponding acetals are more stable and are used as well, although their odor is slightly different and significantly weaker.

The simplest araliphatic aldehyde, benzaldehyde and its 4-isopropyl homologue, cuminaldehyde, are used to a limited extent as fragrance and flavor compounds. However, both compounds are used in large quantity for the production of cinnamic and dihydrocinnamic aldehydes, respectively.

Benzaldehyde [*100-52-7*], C_7H_6O, M_r 106.12, *bp* (101.3 kPa) 178.1 °C, d_4^{15} 1.0415, n_D^{20} 1.5463, is the main, characteristic component of bitter almond oil. It occurs in many other essential oils and is a colorless liquid with a bitter almond odor. For further details on the chemical properties and synthesis of benzaldehyde, see → Benzaldehyde, Chap. 3.

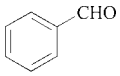

Benzaldehyde is used in aroma compositions for its bitter almond odor. It is the starting material for a large number of araliphatic fragrance and flavor compounds.

Phenylacetaldehyde [*122-78-1*], C_8H_8O, M_r 120.15, *bp* (101.3 kPa) 195 °C, d_4^{20} 1.0272, n_D^{20} 1.5255, has been identified in many essential oils and as a volatile constituent of foods. It is a colorless liquid with a sweet-green odor, reminiscent of hyacinth. Since it readily undergoes oxidation and polymerizes, it must be stabilized by addition of antioxidants and by dilution with, for example, diethyl phthalate before use in compositions.

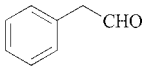

Phenylacetaldehyde can be obtained in high yield by vapor-phase isomerization of styrene oxide, for example, with alkali-treated silica–alumina [126]. Another process starts from phenylethane-1,2-diol, which can be converted into phenylacetaldehyde in high yield. The reaction is performed in the vapor phase in the presence of an acidic silica–alumina catalyst [127].

Phenylacetaldehyde is used in perfume compositions, in particular for hyacinth and rose notes. (IFRA recommendation)

Phenylacetaldehyde dimethyl acetal [*101-48-4*], $C_{10}H_{14}O_2$, M_r 166.22, *bp* (100.2 kPa) 219 – 221 °C, d^{18} 1.004, is a colorless liquid with a strong, rose-petal odor. The dimethyl acetal is more stable than phenylacetaldehyde itself. It imparts a herbal green note to many flower compositions.

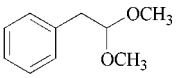

Trade Names. Lilas Vert (Firmenich), Viridine (Givaudan Roure).

Dihydrocinnamaldehyde [*104-53-0*], 3-phenylpropanal, $C_9H_{10}O$, M_r 134.18, *bp* (2.7 kPa) 112 °C, d_4^{20} 1.019, n_D^{20} 1.5266, occurs in Sri Lanka cinnamon oil, among others. The aldehyde is a colorless liquid with a strong, flowery, slightly balsamic, heavy hyacinth-like odor. It tends to undergo self-condensation.

Dihydrocinnamaldehyde can be obtained with scarcely any byproducts by selective hydrogenation of cinnamaldehyde. It is used in perfumery for hyacinth and lilac compositions.

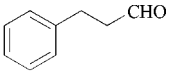

Hydratropaldehyde [*93-53-8*], 2-phenylpropanal, $C_9H_{10}O$, M_r 134.18, *bp* (1.5 kPa) 92 – 94 °C, d_4^{20} 1.0089, n_D^{20} 1.5176, is a colorless liquid with a green hyacinth odor. Hydratropaldehyde can be hydrogenated to hydratropic alcohol, which is also used to a limited extent as a fragrance compound. Hydratropaldehyde is obtained from styrene by oxo synthesis; small quantities of the isomeric dihydrocinnamaldehyde are formed as a byproduct. Hydratropaldehyde is used in perfumery in blossom compositions.

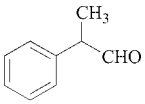

Hydratropaldehyde dimethyl acetal [*90-87-9*], $C_{11}H_{16}O_2$, M_r 180.25, n_D^{20} 1.4938, is a liquid with a mushroom-like, earthy odor. It is used for green nuances in flower compositions.

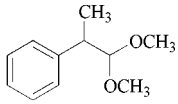

4-Methylphenylacetaldehyde [*104-09-6*], $C_9H_{10}O$, M_r 134.18, *bp* (101.3 kPa) 221 – 222 °C, d_4^{20} 1.0052, n_D^{20} 1.5255, which has been identified in corn oil, is a colorless liquid with a strong green odor. It can be prepared by reaction of 4-methylbenzaldehyde with chloroacetates, followed by hydrolysis of the resulting glycidates and decarboxylation. The aldehyde is used in flower compositions for green notes.

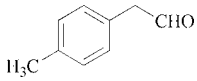

Cyclamenaldehyde [*103-95-7*], 2-methyl-3-(4-isopropylphenyl)propanal, $C_{13}H_{18}O$, M_r 190.28, *bp* (300 Pa) 108 – 108.5 °C, d_4^{20} 0.9502, n_D^{20} 1.5068, has been reported in nutmeg [128]. The commercially available racemate is a colorless to yellowish liquid with an intense flowery odor reminiscent of *Cyclamen europaeum* (cyclamen, sowbread).

Production. Two main processes are used for the industrial synthesis of cyclamenaldehyde:

1) Alkaline condensation of 4-isopropylbenzaldehyde and propanal results, via the aldol **28**, in the formation of 2-methyl-3-(4-isopropylphenyl)-2-propenal (**29**). The unsaturated

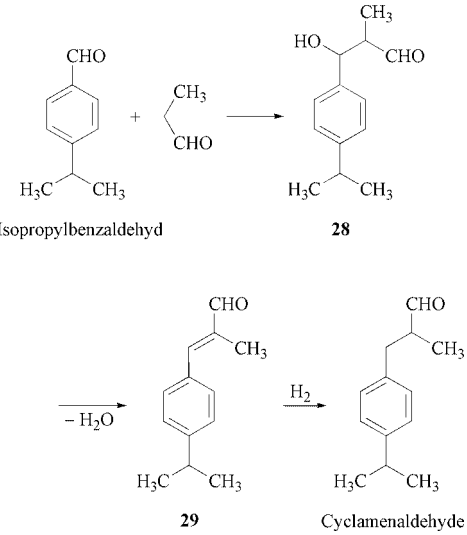

aldehyde is hydrogenated selectively to the saturated aldehyde in the presence of potassium acetate and a suitable catalyst, such as palladium–alumina [129]: see previous page.

2) Friedel–Crafts reaction of isopropylbenzene and 2-methylpropenal diacetate (methacrolein diacetate) in the presence of titanium tetrachloride/boron trifluoride etherate gives cyclamenaldehyde enolacetate, which is hydrolyzed to the aldehyde [130]:

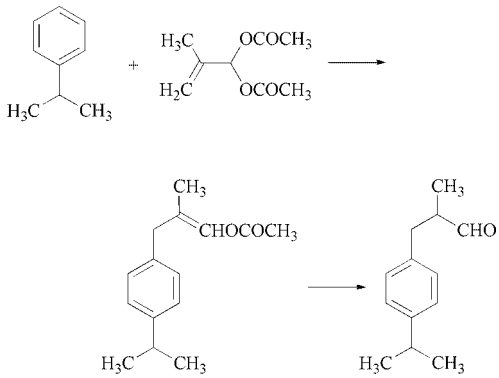

Cyclamenaldehyde

Uses. Cyclamenaldehyde is an important component for obtaining special blossom notes in perfume compositions, particularly the cyclamen type. Because of its fresh-flowery aspect, it is also used as the top note in many other blossom fragrances.

4-*tert*-Butyl-α-methyldihydrocinnamaldehyde [*80-54-6*], 2-methyl-3-(4-*tert*-butylphenyl)propanal, $C_{14}H_{20}O$, M_r 204.31, *bp* (800 Pa) 126 – 127 °C, d_4^{20} 0.9390, n_D^{20} 1.5050, is a homologue of cyclamenaldehyde, but is not found in nature. The racemic compound is a colorless to slightly yellow liquid with a mild-flowery odor, reminiscent of cyclamen and lily of the valley.

The aldehyde is prepared by the same routes as cyclamenaldehyde (see under Cyclamenaldehyde, page 54).

A patent describes the following preparation from α-methylcinnamaldehyde. α-Methylcinnamaldehyde (from benzaldehyde and propionaldehyde) is hydrogenated to α-methyldihydrocinnamic alcohol. The alcohol is alkylated with *tert*-butyl chloride or isobutene to 4-*tert*-butyl-α-methyldihydrocinnamic alcohol, which is subsequently dehydrogenated to the desired aldehyde [131].

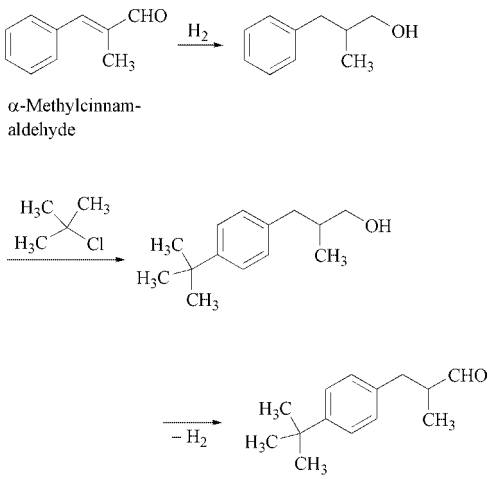

α-Methylcinnamaldehyde

4-*tert*-Butyl-α-methyldihydrocinnamaldehyde

The compound is more stable than cyclamenaldehyde and is a popular component of flower compositions, particularly lily of the valley and linden types, because of its mild, pleasant, blossom fragrance. Large quantities are used in soap and cosmetic perfumes.

Trade Names. Lilestral (IFF), Lilial (Givaudan Roure).

Cinnamaldehyde [*14371-10-9*], 3-phenyl-2-propenal, C_9H_8O, M_r 132.16, *trans*-isomer *bp* (101.3 kPa) 253 °C, d_4^{20} 1.0497, n_D^{20} 1.6195, is the main component of cassia oil (ca. 90 %) and Sri Lanka cinnamon bark oil (ca. 75 %). Smaller quantities are found in many other essential oils. In nature, the *trans*-isomer is predominant.

trans-Cinnamaldehyde is a yellowish liquid with a characteristic spicy odor, strongly reminiscent of cinnamon. Being an α,β-unsaturated aldehyde, it undergoes many reactions, of which hydrogenation to cinnamic alcohol, dihydrocinnamaldehyde, and dihydrocinnamic alcohol are important. Cinnamic acid is formed by autoxidation.

On an industrial scale, cinnamaldehyde is prepared almost exclusively by alkaline condensation of benzaldehyde and acetaldehyde. Self-condensation of acetaldehyde can be avoided by using an excess of benzaldehyde and by slowly adding acetaldehyde [132].

Cinnamaldehyde is used in many compositions for creating spicy and oriental notes (e.g., soap perfumes). It is the main component of artificial cinnamon oil. In addition, it is an important intermediate in the synthesis of cinnamic alcohol and dihydrocinnamic alcohol.

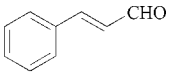

α-Amylcinnamaldehyde [*122-40-7*], 2-pentyl-3-phenyl-2-propenal, $C_{14}H_{18}O$, M_r 202.30, *bp* (0.7 kPa) 140 °C, d_4^{20} 0.9710, n_D^{20} 1.5381. Its natural occurrence as a genuine product is debatable. It is a light yellow liquid with a flowery, slightly fatty odor, which becomes reminiscent of jasmin when diluted. The aldehyde is relatively unstable and must be stabilized by antioxidants.

It is prepared from benzaldehyde and heptanal in the same way as cinnamaldehyde.

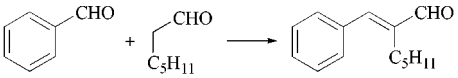

α-Amylcinnamaldehyde

α-Amylcinnamaldehyde is a very popular fragrance substance for creating jasmin notes. It is stable to alkali and long-lasting; large quantities are used, particularly in soap perfumes.

α-Hexylcinnamaldehyde [*101-86-0*], 2-hexyl-3-phenyl-2-propenal, $C_{15}H_{20}O$, M_r 216.32, *bp* (2 kPa) 174 – 176 °C, d^{24} 0.9500, n_D^{25} 1.5268, has not yet been identified in nature. It is a yellow liquid with a mild, slightly fatty, flowery, somewhat herbal odor and a distinct jasmin note. Like the α-amyl homologue, α-hexylcinnamaldehyde must be protected against oxidation by the addition of stabilizers. It is prepared in a manner similar to α-amylcinnamaldehyde by alkaline condensation of excess benzaldehyde with octanal (instead of heptanal). α-Hexylcinnamaldehyde is widely used in flower compositions (e.g., jasmin and gardenia) and, because of its stability to alkali, in soap perfumes.

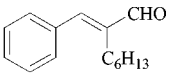

2.5.4. Ketones (→ Ketones)

The aromatic ketones that occur or are used as fragrance and flavor materials are predominantly aryl methyl ketones, which include acetophenones and β-naphthyl methyl ketone. Several acetylpolymethylindanes and -tetralins are commercially important as musk fragrances. Musk ketone, a substituted nitroacetophenone, is discussed in page 62.

Acetophenone [*98-86-2*], methyl phenyl ketone, C_8H_8O, M_r 120.15, *mp* 20.5 °C, *bp* (101.3 kPa) 202.0 °C, d_4^{20} 1.0281, n_D^{20} 1.5372, is a naturally occurring component of a large number of foods and essential oils. It is a colorless liquid with a penetrating sweet odor, reminiscent of orange blossom.

Acetophenone can be hydrogenated catalytically to 1-phenylethanol. It is obtained as a byproduct in the Hock phenol synthesis and is purified from the high-boiling residue by distillation (→ Ketones, Chap. 7.1.). The quantitites obtained from this source satisfy the present demand.

Acetophenone is used for perfuming detergents and industrial products and is an intermediate in the synthesis of other fragrance compounds.

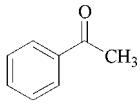

4-Methylacetophenone [*122-00-9*], *p*-tolyl methyl ketone, $C_9H_{10}O$, M_r 134.18, *mp* 28 °C, *bp* (101.3 kPa) 226 °C, d_4^{20} 1.0051, n_D^{20} 1.5335, has been identified in Brazilian rosewood oil and in pepper. It occurs as colorless crystals with a flowery-sweet odor that is milder than that of acetophenone. 4-Methylacetophenone is prepared from toluene and acetic anhydride or acetyl chloride by a Friedel–Crafts reaction. It is used for blossom notes in mimosa and hawthorn type perfumes, especially in soap perfumes.

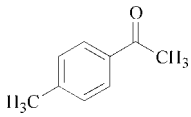

Benzylacetone [*2550-26-7*], 1-phenyl-3-butanone, $C_{10}H_{12}O$, M_r 148.20, *bp* (101.3 kPa) 233 – 234 °C, d_4^{22} 0.9849, n_D^{22} 1.5110, has been identified as a volatile component of cocoa. Benzylacetone is a sweet-flowery smelling liquid, which can be prepared by selective hydrogenation of benzylidene acetone (from benzaldehyde and acetone). It is used in soap perfumes.

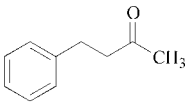

Methyl β-naphthyl ketone [*93-08-3*], $C_{12}H_{10}O$, M_r 170.21, *bp* (1.7 kPa) 171 – 173 °C, *mp* 56 °C, d_4^{20} 1.171, n_D^{20} 1.6752, has been identified in some essential oils. It smells like orange blossom and is a colorless crystalline solid. It is usually prepared by Friedel–Crafts acetylation of naphthalene (with acetylchloride, acetic anhydride, etc.) in the presence of aluminum chloride. In polar solvents (e.g., nitrobenzene), the percentage of the simultaneously formed α-isomer is lower. Methyl β-naphthyl ketone is used in eaux de cologne, soap perfumes, and detergents. It is a good fixative.

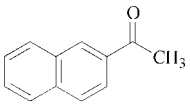

Benzophenone [*119-61-9*], diphenyl ketone, $C_{13}H_{10}O$, M_r 182.22, *bp* (6.3 kPa) 200.5 °C, *mp* 48.1 °C, d^{50} 1.976, has been identified as a flavor component of grapes. It is a colorless crystalline solid with a rosy, slightly geranium-like odor. It can be prepared in several ways (→ Ketones, Chap. 7.3.), for example, by Friedel–Crafts reaction of benzene and benzoyl chloride with aluminum chloride, or of benzene and carbon tetrachloride, and subsequent hydrolysis of the resulting α,α-dichlorodiphenylmethane. Benzophenone can also be prepared by oxidation of diphenylmethane. It is used in flower compositions and as a fixative.

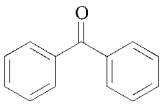

6-Acetyl-1,1,2,3,3,5-hexamethylindane [*15323-35-0*], $C_{17}H_{24}O$, M_r 244.38, *mp* 35 °C, is a synthetic musk fragrance. It is prepared by Friedel–Crafts acetylation of 1,1,2,3,3,5-hexamethylindane, which can be obtained as a 70 : 30 mixture with 1,1,3,5-tetramethyl-3-ethylindane by reacting α,*p*-dimethylstyrene with amylenes or 2-methyl-2-butanol in a mixture of acetic acid and concentrated sulfuric acid [133]:

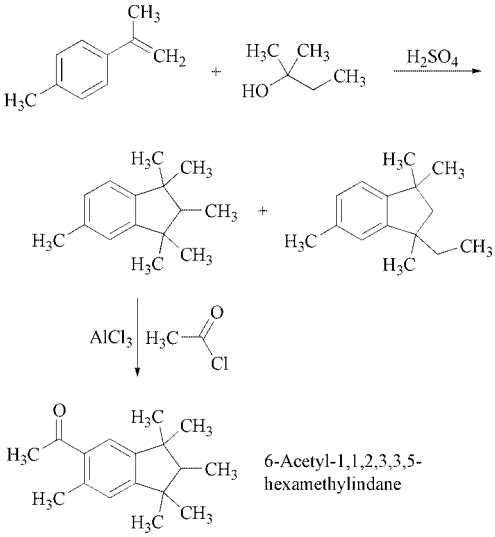

The indane ketone is a musk fragrance that is stable to light and in soap. It is used in perfumes and cosmetics for its fixative properties as well as its fragrance.
Trade Name. Phantolide (Haarmann & Reimer).

4-Acetyl-1,1-dimethyl-6-*tert*-butylindane [*13171-00-1*], $C_{17}H_{24}O$, M_r 244.38, *mp* 76.7 – 77.2 °C, is a musk fragrance that does not occur in nature. It is prepared by reacting *tert*-

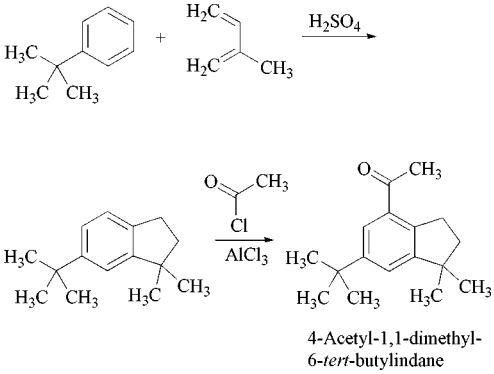

butylbenzene with isoprene in the presence of sulfuric acid, followed by acetylation of the resulting 1,1-dimethyl-6-*tert*-butylindane [134]: see previous page.

The indane is light-stable and is mainly used for perfuming soaps and cosmetics.
Trade Names. Celestolide (IFF), Crysolide (Givaudan Roure).

5-Acetyl-1,1,2,6-tetramethyl-3-isopropylindane [*68140-48-7*], $C_{18}H_{26}O$, M_r 258.40, *bp* (130 Pa) 144 – 146 °C, n_D^{20} 1.5301, is also a musk fragrance that does not occur in nature. It is prepared from toluene and isobutyryl chloride by a Friedel–Crafts reaction that yields *p*-tolyl isopropyl ketone; the ketone is reduced to the corresponding alcohol. Chlorination and treatment with 2-methyl-2-butene yield 1-isopropyl-2,3,3,5-tetramethylindane, which by a Friedel–Crafts reaction with acetyl chloride gives the title compound [135]:

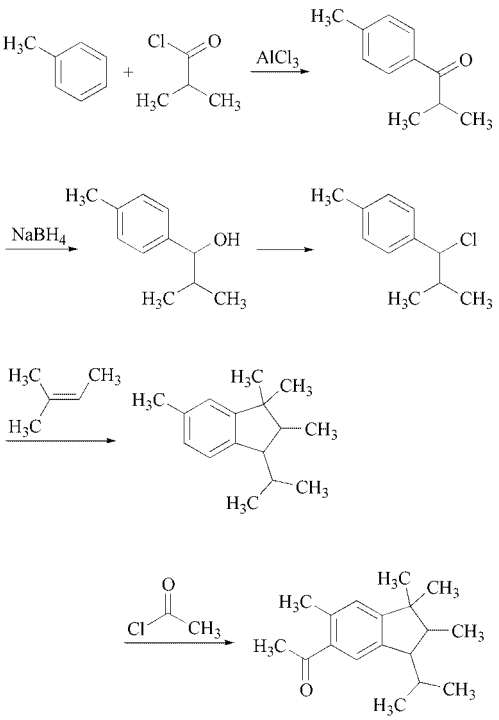

5-Acetyl-1,1,2,6-tetramethyl-3-isopropylindane

It is used in perfume compositions for soaps and detergents.
Trade Name. Traseolide (Quest).

6-Acetyl-1,1,2,4,4,7-hexamethyltetralin [*1506-02-1*], $C_{18}H_{26}O$, M_r 258.40, *mp* 55.5 °C, *bp* (250 Pa) 119 °C, is a synthetic musk fragrance. It is prepared from 1,1,2,4,4,7-hexamethyltetralin, which is obtained by one of the following routes:

1) Reaction of α,*p*-dimethylstyrene with tetramethylethene [136] or 2,3-dimethylbutan-2-ol [133] in an acetic acid–sulfuric acid mixture yields the desired hexamethyltetralin, in addition to polymerized starting materials:

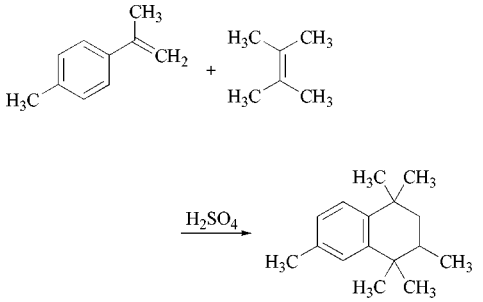

1,1,2,4,4,7-Hexamethyltetralin

2) Reaction of *p*-cymene with 2,2-dimethyl-3-butene (neohexene) and a tertiary alkylhalide (as a hydrogen scavenger) in the presence of catalytic amounts of anhydrous aluminum halide in inert solvents produces a high yield of the hexamethyltetralin [137].

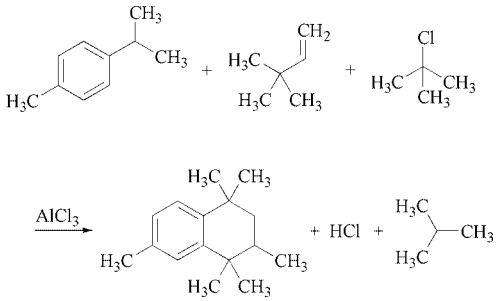

1,1,2,4,4,7-Hexamethyltetralin is subsequently acetylated to 6-acetyl-1,1,2,4,4,7-hexamethyltetralin, e.g., with acetyl chloride and aluminum chloride (see next page).

The product is a light-stable, versatile musk fragrance that is used in soap and cosmetics.
Trade Names. Fixolide (Givaudan Roure), Tonalide (Haarmann & Reimer).

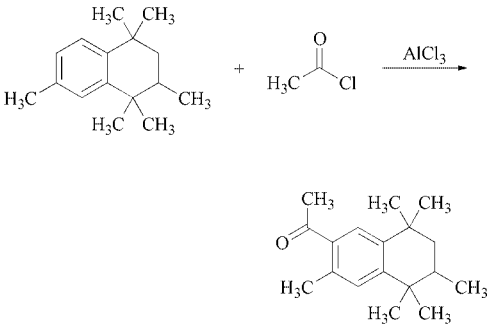

6-Acetyl-1,1,2,4,4,7-hexamethyltetralin

2.5.5. Esters of Araliphatic Alcohols and Aliphatic Acids

Esters of araliphatic alcohols and aliphatic acids are interesting as flavors and fragrances because of their characteristic odor properties. Acetates are the most popular esters. Benzyl acetate is particularly important commercially and occupies a prominent position in the fragrance and flavor industry.

Esters of other fatty acids are used to a lesser extent. In addition to benzyl esters and phenethyl esters, isomeric homologues with substituted side-chains are used in fairly large amounts in perfume compositions because of their special blossom odors. Not all have yet been found in nature. The esters are prepared from the corresponding alcohols via the customary routes (→ Esters, Organic).

Benzyl esters of lower molecular mass fatty acids occur widely in nature. The following are important fragrance and flavor materials:

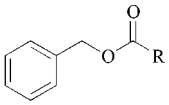

Benzyl acetate [*140-11-4*], R = CH_3, $C_9H_{10}O_2$, M_r 150.18, *bp* (101.3 kPa) 215 °C, d_4^{20} 1.0550, n_D^{20} 1.5232, is the main component of jasmin and gardenia oils. It occurs as a minor component in a large number of other essential oils and extracts. It is a colorless liquid with a strong, fruity, jasmin odor. Benzyl acetate is prepared by esterification of benzyl alcohol with acetic anhydride (e.g., with sodium acetate as a catalyst) or by reaction of benzyl chloride with sodium acetate (→ Benzyl Alcohol, Chap. 8.2.). In terms of volume, benzyl acetate is one of the most important fragrance and flavor chemicals.

Benzyl propionate [*122-63-4*], R = CH_2CH_3, $C_{10}H_{12}O_2$, M_r 164.20, *bp* (101.3 kPa) 219–220 °C, d_4^{20} 1.0335, n_D^{20} 1.4996, is a liquid with a sweet-fruity odor, which is used in perfumery for floral-fruity notes and in fruit flavor compositions.

Benzyl isovalerate [*103-38-8*], R = $CH_2CH(CH_3)_2$, $C_{12}H_{16}O_2$, M_r 192.26, *bp* (101.3 kPa) 245 °C, d_4^{20} 0.9900, n_D^{20} 1.4878, is a liquid with a heavy, flowery odor, which is used in perfumery for oriental and heavy blossom odors.

Phenethyl Esters. The most commonly used *phenethyl esters* are the following:

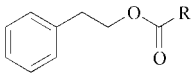

Phenethyl acetate [*103-45-7*], R = CH_3, $C_{10}H_{12}O_2$, M_r 164.20, *bp* (101.3 kPa) 232.6 °C, d_4^{20} 1.0883, n_D^{20} 1.5171, occurs in a number of essential oils and is a volatile aroma component of many fruits and alcoholic beverages. Phenethyl acetate is a colorless liquid with a fine rose scent and a secondary, sweet, honey note. It is used in perfumery as a modifier of phenethyl alcohol, e.g., in rose and lilac compositions. In addition, it is used in a large number of aromas, in keeping with its natural occurrence.

Phenethyl isobutyrate [*103-48-0*], R = $CH(CH_3)_2$, $C_{12}H_{16}O_2$, M_r 192.26, *bp* (2 kPa) 122–124 °C, d^{15} 0.9950, n_D^{20} 1.4871, occurs in peppermint oils. It has a heavy, fruity, blossom odor and is used accordingly in perfume and flavor compositions.

Phenethyl isovalerate [*140-26-1*], R = $CH_2CH(CH_3)_2$, $C_{13}H_{18}O_2$, M_r 206.28, *bp* (7.2 kPa) 141–145 °C, d^{15} 0.9845, n_D^{20} 1.4855, has been identified as a volatile aroma component of peppermint oils. The fruitiness of its odor is even more pronounced than that of the isobutyrate. It is used in small quantities for the same purposes as phenethyl isobutyrate.

1-Phenylethyl acetate [*93-92-5*], styrallyl acetate, $C_{10}H_{12}O_2$, M_r 164.20, *bp* (1.3 kPa) 92.5 °C, d_4^{20} 1.0277, n_D^{20} 1.4954, is a liquid with a dry, fruity-green, blossom odor, reminiscent

of gardenia. It can occur in the form of optically active enantiomers, but only the racemate is used in perfumery. Styrallyl acetate is a key ingredient in gardenia fragrances and is added to many other blossom compositions, particularly for dry top notes.

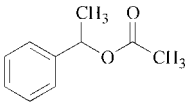

α,α-Dimethylphenethyl Esters. The following α,α-dimethylphenethyl esters are commercially important:

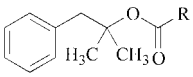

α,α-Dimethylphenethyl acetate [*151-05-3*], R = CH_3, $C_{12}H_{16}O_2$, M_r 192.26, *mp* ca. 30 °C, *bp* (400 Pa) 90 °C, d_{25}^{25} 0.998 – 1.000, n_D^{20} 1.4923, is a colorless liquid with a flowery-woody odor. The ester is used in blossom compositions, e.g., lily of the valley, rose, and geranium.

α,α-Dimethylphenethyl butyrate [*10094-34-5*], R = $(CH_2)_2CH_3$, $C_{14}H_{20}O_2$, M_r 220.31, *bp* (400 Pa) 96 °C, d_{25}^{25} 0.971 – 0.974, n_D^{20} 1.4860 – 1.4900, is a colorless liquid with a slightly herbal, strongly fruity odor, reminiscent of prune and apricot. It is used in perfumery as a modifier of the alcohol and for oriental notes.

Cinnamyl acetate [*103-54-8*], $C_{11}H_{12}O_2$, M_r 176.21, *bp* (1.3 kPa) 139 – 140 °C, d^{22} 1.0520, n_D^{20} 1.5420, is the only ester of cinnamic alcohol of any importance. *trans*-Cinnamyl acetate occurs in cassia oil and is a colorless liquid with a sweet-flowery-fruity, slightly balsamic odor. It is a good fixative and is used in blossom compositions (e.g., lilac and jasmin) and for oriental notes. In aroma compositions, it is used for cinnamon-fruity effects.

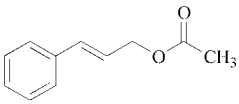

2.5.6. Aromatic Acids

Aromatic acids (e.g., benzoic acid) and araliphatic acids (e.g., phenylacetic, cinnamic, and dihydrocinnamic acids) occur in numerous essential oils and have also been identified in the aromas of many foods. However, phenylacetic acid is the only acid that is used in significant quantities as a fragrance and flavor substance.

Phenylacetic acid [*103-82-2*], $C_8H_8O_2$, M_r 136.15, *bp* (101.3 kPa) 265.5 °C, *mp* 78 °C, $d_4^{79.8}$ 1.0809, occurs in Japanese peppermint oil, in neroli oil, and in traces in rose oils. It is a volatile aroma constituent of many foods (e.g., honey). It forms colorless crystals that have a honey odor. For further details on chemical properties and synthesis, see → Phenylacetic Acid.

Because of its intense odor, phenylacetic acid is added to perfumes in small quantities for rounding off blossom odors. Addition to fruit aromas imparts a sweet honey note.

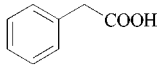

2.5.7. Esters Derived from Aromatic and Araliphatic Acids

The acid moiety generally determines the odor of esters derived from aromatic or araliphatic acids. Unless stated otherwise, the esters are prepared from the corresponding acids or acid derivatives and alcohols by the customary methods (→ Esters , Organic).

2.5.7.1. Benzoates

The following benzoates are used in fairly large quantities as perfumery materials.

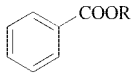

Methyl benzoate [*93-58-3*], R = CH_3, $C_8H_8O_2$, M_r 136.15, *bp* (101.3 kPa) 199.6 °C, d_4^{20} 1.0888, n_D^{20} 1.5164, has been found in essential oils (e.g., ylang-ylang and tuberose blossom

oil). It is a colorless liquid with a strong, dry-fruity, slightly phenolic odor. For further properties and synthesis, seeBenzoic Acid and Derivatives, Chap. 8.5. .

Methyl benzoate is used in perfume bases, such as ylang-ylang and tuberose types.

Hexyl benzoate [*6789-88-4*], R=$(CH_2)_5CH_3$, $C_{13}H_{18}O_2$, M_r 206.28, *bp* (102.6 kPa) 272 °C, is a liquid with a balsamic-green, melon-like odor. It is used in perfumery.

Benzyl benzoate [*120-51-4*], R = $CH_2C_6H_5$, $C_{14}H_{12}O_2$, M_r 212.25, *bp* (2.0 kPa) 170–171 °C, *mp* 21–22 °C, d_4^{25} 1.1121, n_D^{20} 1.5680, is the main component of Peru balsam. It occurs in fairly large amounts in a number of blossom concretes and absolutes (e.g., tuberose and hyacinth). It forms either a viscous liquid or solid flakes and has a weak, sweet-balsamic odor. For further properties and synthesis, see Benzoic Acid and Derivatives, Chap. 8.5.

Benzyl benzoate is used in perfumery as a fixative and as a modifier in heavy blossom fragrances.

2.5.7.2. Phenyl acetates

Of the phenyl acetates, the following are particularly important fragrance and flavor substances:

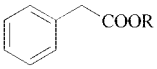

Ethyl phenyl acetate [*101-97-3*], R = CH_2CH_3, $C_{10}H_{12}O_2$, M_r 164.20, *bp* (101.3 kPa) 227 °C, d_4^{20} 1.0333, n_D^{20} 1.4980, is a volatile aroma component of fruit and honey. It is a colorless liquid with a strong, sweet odor reminiscent of honey. Small amounts are used in flower perfumes and in fruit flavors.

Geranyl phenyl acetate [*102-22-7*], R =$CH_2CH{=}C(CH_3)CH_2CH_2CH{=}C(CH_3)_2$, $C_{18}H_{24}O_2$, M_r 272.39, has not been found in nature. It is a yellow liquid with a mild rose odor and a secondary honey note. It is used as a fixative in rose compositions and heavy perfumes.

Phenethyl phenyl acetate [*102-20-5*], R = $CH_2CH_2C_6H_5$, $C_{16}H_{16}O_2$, M_r 240.30, *bp* (600 Pa) 177–178 °C, *mp* 26.5 °C, d_{25}^{25} 0.880, n_D^{20} 1.5496–1.5504, has not yet been identified in natural sources. It occurs as a colorless liquid or crystals, which have a heavy, sweet, rose or hyacinth odor and a distinct honey note. The ester is used particularly in flowery fragrance compositions and as a fixative.

2.5.7.3. Cinnamates

Three cinnamates are of some importance in perfumery:

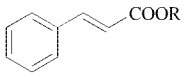

Methyl cinnamate [*103-26-4*], R = CH_3, $C_{10}H_{10}O_2$, M_r 162.19, *bp* (101.3 kPa) 261.9 °C, *mp* 36.5 °C, d_4^{20} 1.0911, n_D^{21} 1.5766, occurs in essential oils, mostly as the *trans*-isomer. It is the main component of oils isolated from *Alpinia* species (content up to 80 %) and *Ocimum canum* varieties (> 50 %). It has also been identified as a volatile aroma component of cinnamon and strawberries.

Methyl cinnamate is a colorless crystalline solid with a fruity, sweet-balsamic odor. In addition to the common esterification methods, it can be prepared by Claisen condensation of benzaldehyde and methyl acetate in the presence of sodium. Methyl cinnamate is used in soap perfumes, as well as in blossom and oriental perfumes, and is sometimes added to aromas.

Benzyl cinnamate [*103-41-3*], R = $CH_2C_6H_5$, $C_{16}H_{14}O_2$, M_r 238.29, *mp* 35–36 °C, occurs in balsams and balsam oils. It forms white, sweet-balsamic-smelling crystals. Benzyl cinnamate is used as a fixative in perfumes and as a component of heavy, oriental perfumes.

Phenethyl cinnamate [*103-53-7*], R = $CH_2CH_2C_6H_5$, $C_{17}H_{16}O_2$, M_r 252.31, *mp* 65–68 °C, occurs in extracts from *Populus balsamifera* buds. It is a crystalline solid with a heavy, rosy-balsamic odor. It is used as a fixative in blossom fragrances.

2.5.8. Miscellaneous Compounds

Of the few aromatic, nitrogen-containing fragrance substances, the nitro musks are by far the most important. They are still considerably important commercially as substitutes for the natural macrocyclic musks that are difficult to synthesize, but nitro musks are being increasingly replaced by lightfast musk fragrances. Methyl anthranilate and its *N*-methyl derivative are also aromatic, nitrogen-containing compounds that are used as fragrances and flavors in fairly large amounts. A number of aromatic nitriles have been introduced; they are stable to alkali and, therefore, are used in soap perfumes.

Musk xylol [*81-15-2*], 2,4,6-trinitro-1,3-dimethyl-5-*tert*-butylbenzene, $C_{12}H_{15}N_3O_6$, M_r 297.27, *mp* 114 °C, does not occur in nature. It forms yellow crystals with a persistent musk odor. Musk xylol is prepared by alkylation of *m*-xylene with isobutene and subsequent nitration with a sulfuric acid–nitric acid mixture.

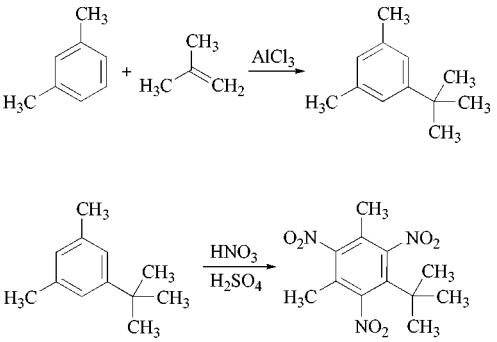

Musk xylol

Musk xylol is used in large quantities in inexpensive perfumes for soap and household products.

Musk ketone [*81-14-1*], 3,5-dinitro-2,6-dimethyl-4-*tert*-butylacetophenone, $C_{14}H_{18}N_2O_5$, M_r 294.31, *mp* 137 °C, does not occur in nature. It forms yellowish crystals with a sweet, very persistent, slightly animal musk odor. Musk ketone is prepared by Friedel–Crafts acetylation of 1,3-dimethyl-5-*tert*-butylbenzene, and nitration of the resulting 2,6-dimethyl-4-*tert*-butylacetophenone with nitric acid.

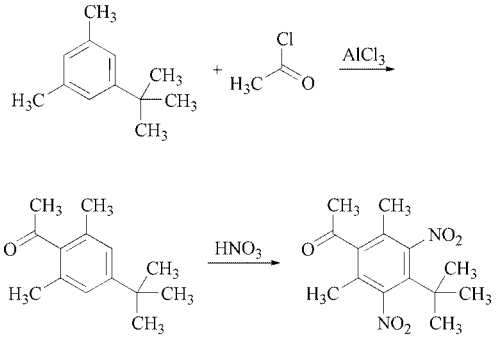

Musk ketone

Musk ketone is widely used as a fixative in blossom and phantasy compositions.

Cinnamonitrile [*4360-47-8*], C_9H_7N, M_r 129.16, *bp* (1.7 kPa) 135 – 135.5 °C, *mp* 23.5 – 24 °C, d^{25} 1.0244, n_D^{25} 1.6001; *trans*-cinnamonitrile is a colorless crystalline solid or a colorless viscous liquid with a spicy, slightly flowery odor.

Cinnamonitrile can be prepared by one of the common routes to nitriles, e.g., by dehydration of cinnamaldoxime. It is stable to alkali and is used for perfuming soap and detergents.

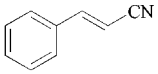

Trade Name. Cinnamalva (IFF).

5-Phenyl-3-methyl-2-pentenonitrile [*93893-89-1*], $C_{12}H_{13}N$, M_r 171.24, *bp* (20 Pa) 82 – 88 °C, d^{25} 0.979, n_D^{20} 1.5340, is a colorless liquid with a citrus-fruity, slightly balsamic odor. The commercial product is a 2 : 3 mixture of the *cis* and *trans*-isomers.

The nitrile is prepared by condensation of benzylacetone with cyanoacetic acid in the presence of pyridine and by elimination of carbon dioxide. The mixture is used as such in soap and detergent perfumes [140].

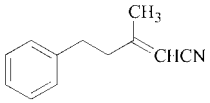

Trade Name. Citronitril (Haarmann & Reimer).

Methyl anthranilate [*134-20-3*], $C_8H_9NO_2$, M_r 151.16, *bp* (2 kPa) 135.5 °C,

mp 24 – 25 °C, d_4^{19} 1.1682, n_D^{20} 1.5815, occurs in a large number of blossom essential oils (e.g., neroli, ylang-ylang, and jasmin oils), grapes, and citrus oils. It occurs as white crystals or a yellowish liquid, that show blue fluorescence and have an orange blossom odor. Methyl anthranilate is prepared by esterification of anthranilic acid with methanol or by reaction of isatoic anhydride with methanol [141].

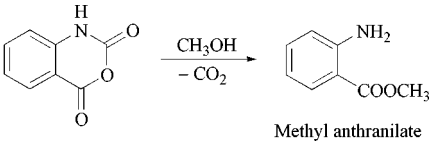

Methyl anthranilate

It is used in a large number of blossom fragrances. However, its use in perfumes for soaps and cosmetics is limited because it causes discoloration. It is used in aroma compositions (e.g., in grape and citrus flavors).

Methyl *N*-methylanthranilate [*85-91-6*], $C_9H_{11}NO_2$, M_r 165.19, *bp* (1.6 kPa) 130 – 131 °C, d_4^{20} 1.1295, n_D^{20} 1.5796, is the main component of petitgrain oil from mandarin leaves and is also found in mandarin peel oil. It is a pale yellow, fluorescent liquid with a delicate mandarin odor. The ester can be prepared by methylation of methyl anthranilate. It is used in soap and cosmetic perfumes as well as in aromas, particularly for mandarin flavors. (IFRA recommendation)

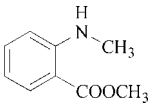

Some Schiff's bases of methyl anthranilate are interesting fragrance materials, e.g., those formed with hydroxydihydrocitronellal [*89-43-0*] and 2,4-dimethyl-3-cyclohexene carboxaldehyde [*68845-02-3*]. Both have a heavy blossom odor and a high tenacity.

2.6. Phenols and Phenol Derivatives

For a detailed description of phenols and their derivatives, see → Phenol, → Phenol Derivatives.

2.6.1. Phenols, Phenyl Esters, and Phenyl Ethers

Of the phenols and phenyl ethers used as fragrance and flavor compounds, 4-allyl- (**30**, R = H) and 4-propenylphenols (**31** and **32**, R = H) and their methyl ethers (**30** – **32**, R = CH_3) occur particularly frequently in essential oils.

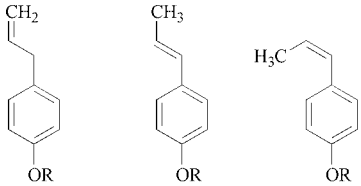

30 **31** **32**

A second hydroxyl or methoxyl substituent is often present; 2-methoxy-4-allylphenol and 2-methoxy-4-propenylphenol are the most important compounds belonging to this category.

Diphenyl ether [*101-84-8*], $C_{12}H_{10}O$, M_r 170.21, *bp* (1.34 kPa) 121 °C, *mp* 26.8 °C, d^{20} 1.0748, has not been observed in nature. It is a colorless liquid or a crystalline solid with an odor reminiscent of geranium leaves. Diphenyl ether is obtained as a byproduct in the production of phenol by high-pressure hydrolysis of chlorobenzene. For further details of properties and synthesis, see → Phenol Derivatives, Chap. 6.3. Because of its stability and low price, diphenyl ether is used in large quantities in soap perfumes. However, its main application is as a heat-transfer medium (eutectic mixture with diphenyl).

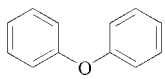

Thymol [*89-83-8*], 2-isopropyl-5-methylphenol, $C_{10}H_{14}O$, M_r 150.22, *bp* (101.3 kPa) 232.5 °C, *mp* 51.5 °C, d_4^{20} 0.9756, n_D^{20} 1.5227, is the main constituent of thyme and some origanum oils; it also occurs in many other essential oils. It forms colorless crystals with a spicy-herbal, slightly medicinal odor reminiscent of thyme. For further details of properties and synthesis, see → Phenol Derivatives, Chap. 1.5.3.

Thymol is used as a dry top note in lavender compositions, in men's fragrances, and as a disinfectant in oral hygiene products. It is also important as a starting material for the production of racemic menthol (see page 30).

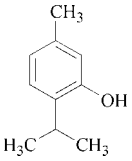

Anethole [*104-46-1*], 1-methoxy-4-propenylbenzene, $C_{10}H_{12}O$, M_r 148.20, *trans*-isomer: *bp* °C, d_4^{20} 0.9883, n_D^{20} 1.5615, occurs both as its *cis*- and *trans*-isomers in nature; however, *trans*-anethole is always the main isomer. Anethole occurs in anise oil (80–90 %), star anise oil (> 90 %), and fennel oil (80 %).

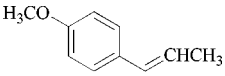

trans-Anethole [*4180-23-8*] forms colorless crystals (*mp* 21.5 °C) with an anise-like odor and a sweet taste. Anethole is oxidized to anisaldehyde (e.g., with chromic acid); when hydrogenated it is converted into 1-methoxy-4-propylbenzene.

Production. Isolation of anethole from anethole-rich essential oils has lost its significance in favor of isolation from sulfate turpentine oils (see page 118).

1) Anethole can be crystallized from oils in which it occurs as a major component (anise, star anise, and fennel oils).
2) A fraction of American sulfate turpentine oil (0.5 % of the total) consists mainly of an azeotropic mixture of anethole and caryophyllene. *trans*-Anethole can be isolated from this mixture by crystallization.
3) Another fraction of American sulfate turpentine oil (1 % of the total) consists essentially of an azeotropic mixture of estragole (1-methoxy-4-allylbenzene, *bp* (101.3 kPa) 216 °C) and α-terpineol. Treatment with potassium hydroxide yields a mixture of anethole isomers and α-terpineol, which can be separated by fractional distillation.

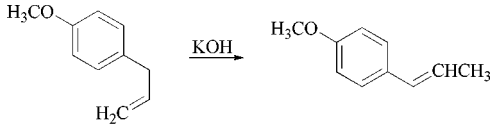

Estragole Anethole

Since isolation from oils does not always satisfy demand, synthetic processes starting from anisole have been developed.

4) *Synthesis from Anisole and Propionic Acid Derivatives.* Anisole is converted into 4-methoxypropiophenone by Friedel–Crafts acylation with propionyl chloride or propionic anhydride. The ketone is hydrogenated to the corresponding alcohol with a copper chromite catalyst. The alcohol is dehydrated in the presence of acidic catalysts to a mixture of *cis*- and *trans*-anethole [142].

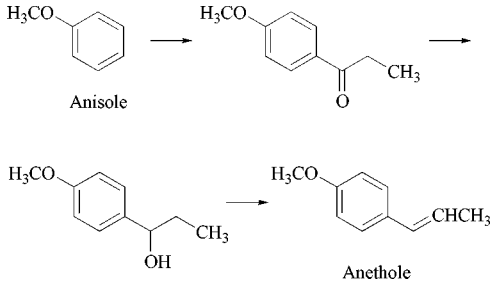

Anisole Anethole

5) *Synthesis from Anisole and Propanal.* Anisole is converted into a mixture of 1,1-bis(4-methoxyphenyl)propane and small amounts of isomeric compounds by condensation with propionaldehyde in the presence of catalytic amounts of acid. 1,1-Bis(4-methoxyphenyl) propane is cleaved when heated to give anethole and anisole [143].

Uses. Anethole is used in large quantities in the alcoholic beverage industry (Pernod, Ouzo) and in oral hygiene products. Some crude anethole is converted into anisaldehyde.

2-Phenoxyethyl isobutyrate [*103-60-6*], $C_{12}H_{16}O_3$, M_r 208.26, *bp* (530 Pa) 125–127 °C, d_{25}^{25} 1.044–1.050, n_D^{20} 1.492–1.496, is a fragrance compound that does not occur in nature. It is a colorless liquid with a sweet, flowery-fruity odor.

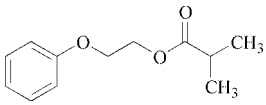

The ester is prepared by esterification of 2-phenoxyethanol with isobutyric acid and is used as a fixative in perfumes (rose and lavender types) as well as for fruity notes.
Trade Name. Phenirat (Haarmann & Reimer).

β-Naphthyl Alkyl Ethers. The β-naphthyl alkyl ethers described below are used in perfumery, especially in soap perfumes. The ethers are prepared by O-alkylation of β-naphthol. They have not been observed in nature.

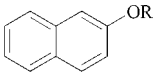

β-Naphthyl methyl ether [*93-04-9*], R = CH_3, $C_{11}H_{10}O$, M_r 158.20, *bp* (1.3 kPa) 138 °C, *mp* 73 – 74 °C, forms white crystals with an intense orange blossom odor.

β-Naphthyl ethyl ether [*93-18-5*], R = CH_2CH_3, $C_{12}H_{12}O$, M_r 172.23, *bp* (1.3 kPa) 148 °C, *mp* 37 – 38 °C, forms white crystals with a mild, long-lasting, orange blossom fragrance.

β-Naphthyl isobutyl ether [*2173-57-1*], R = $CH_2CH(CH_3)_2$, $C_{14}H_{16}O$, M_r 200.28, *mp* 33 – 33.5 °C, forms white crystals with a fruity, orange blossom odor.

Hydroquinone dimethyl ether [*150-78-7*], 1,4-dimethoxybenzene, $C_8H_{10}O_2$, M_r 138.17, *bp* (2.7 kPa) 109 °C, *mp* 57 – 58 °C, occurs in hyacinth oil and has also been identified in tea. It is a white crystalline solid with an intensely sweet, somewhat herbal, nutlike odor.

Hydroquinone dimethyl ether is prepared by etherification of hydroquinone and is used in soap perfumes.

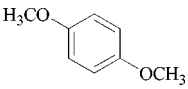

Isoeugenol [*97-54-1*], 2-methoxy-4-(1-propenyl)phenol, $C_{10}H_{12}O_2$, M_r 164.22; *cis*-isomer [*5912-86-7*]: *bp* (1.7 kPa) 134 – 135 °C, d_4^{20} 1.0837, n_D^{20} 1.5726; *trans*-isomer [*5932-68-3*]: *mp* 33 – 34 °C, *bp* (1.7 kPa) 141 – 142 °C, d_4^{20} 1.0852, n_D^{20} 1.5784. Isoeugenol occurs in many essential oils, mostly with eugenol, but not as the main component. Commercial isoeugenol is a mixture of *cis*- and *trans*-isomers, in which the *trans*-isomer dominates because it is thermodynamically more stable. Isoeugenol is a yellowish, viscous liquid with a fine clove odor, that of the crystalline *trans*-isomer being the more delicate.

Isoeugenol can be hydrogenated catalytically to form dihydroeugenol. Vanillin was formerly prepared by oxidation of isoeugenol. Additional fragrance compounds are prepared by esterification or etherification of the hydroxyl group.

Production. Starting materials for the synthesis of isoeugenol are eugenol and guaiacol.

1) *Synthesis from Eugenol.* The sodium or potassium salt of eugenol is isomerized to isoeugenol by heating. Isomerization can also be carried out catalytically in the presence of ruthenium [144] or rhodium [145] compounds.

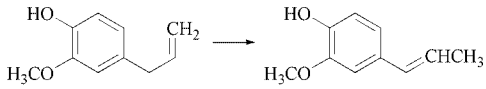

Eugenol Isoeugenol

2) *Synthesis from Guaiacol.* Guaiacol is esterified with propionic acid, and the resulting guaiacyl propionate rearranges in the presence of aluminum chloride to give 4-hydroxy-3-methoxypropiophenone. Reduction of the ketone to the corresponding secondary alcohol and dehydration finally yield isoeugenol [146–148].

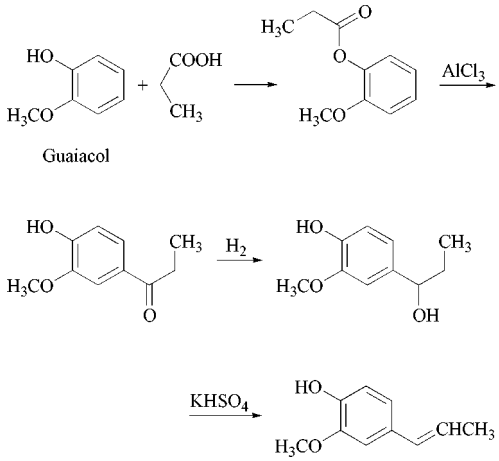

Guaiacol

Uses. Isoeugenol is used in perfumery in a large number of blossom compositions, mostly for clove and carnation types, but also in oriental perfumes. (Limiting IFRA recommendation)

Small amounts are employed in aromas and in reconstituted essential oils.

Isoeugenol methyl ether [*93-16-3*], $C_{11}H_{14}O_2$, M_r 178.23, *cis*-isomer [*6380-24-1*]: *bp* (900 Pa) 137 – 137.5 °C, d_4^{20} 1.0530, n_D^{20} 1.5628; *trans*-isomer [*6379-72-2*]: *bp* (700 Pa) 126 °C, d_4^{20} 1.0556, n_D^{20} 1.5699, occurs in small quantities in several essential oils. It is a colorless to pale yellow liquid with a mild clove odor.

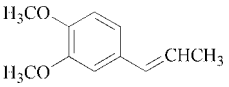

Isoeugenol methyl ether is used in perfumery in clove and carnation bases and as a fixative in spicy-floral compositions.

Eugenol [*97-53-0*], 2-methoxy-4-allylphenol, $C_{10}H_{12}O_2$, M_r 164.20, *bp* (1.3 kPa) 121 °C, d_4^{20} 1.0652, n_D^{20} 1.5409, is the main component of several essential oils; clove leaf oil and cinnamon leaf oil may contain > 90 %. Eugenol occurs in small amounts in many other essential oils. It is a colorless to slightly yellow liquid with a spicy, clove odor.

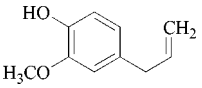

Catalytic hydrogenation (e.g., in the presence of noble-metal catalysts) yields dihydroeugenol. Isoeugenol is obtained from eugenol by shifting the double bond (see Isoeugenol, page 65). Esterification and etherification of the hydroxyl group of eugenol yield valuable fragrance and flavor materials (e.g., eugenyl acetate and eugenyl methyl ether).

Production. Since sufficient eugenol can be isolated from cheap essential oils, synthesis is not industrially important. Eugenol is still preferentially isolated from clove leaf and cinnamon leaf oil (e.g., by extraction with sodium hydroxide solution). Nonphenolic materials are then removed by steam distillation. After the alkaline solution is acidified at low temperature, pure eugenol is obtained by distillation.

Uses. Eugenol is used in perfumery in clove and carnation compositions as well as for oriental and spicy notes. It is a common component of clove and other aroma compositions. In dentistry, it is used as an antiseptic.

Eugenol methyl ether [*93-15-2*], $C_{11}H_{14}O_2$, M_r 178.23, *bp* (1.5 kPa) 127 – 129 °C, d_4^{20} 1.0396, n_D^{20} 1.5340, occurs in numerous essential oils, sometimes at a very high concentration; leaf and wood oil from *Dacrydium franklinii* Hook. (Huon pine oil) contain more than 90 %. The ether is an almost colorless liquid with a mild-spicy, slightly herbal odor. It is prepared by methylation of eugenol and is used in perfumery (e.g., in carnation and lilac compositions) and in flavor compositions.

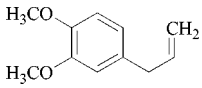

Eugenyl acetate [*93-28-7*], $C_{12}H_{14}O_3$, M_r 206.24, *bp* (400 Pa) 120 – 121 °C, *mp* 29 °C, d_4^{20} 1.0806, n_D^{20} 1.5205, occurs in clove oil, together with eugenol. It is a crystalline solid or yellowish liquid with a slightly fruity, clove odor. Eugenyl acetate is prepared by acetylation of eugenol with acetic anhydride and is used in clove compositions to accentuate flowery character.

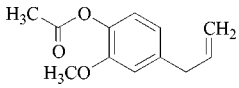

Propenylguethol [*94-86-0*], 2-ethoxy-5-(1-propenyl)phenol, $C_{11}H_{14}O_2$, M_r 178.23, exists in *cis* (*mp* 35 – 36 °C) and in *trans* (*mp* 86 °C) forms. The *trans*-isomer has a sweet vanilla-like odor. Propenylguethol can be prepared from isosafrole by reaction with methylmagnesium chloride or by ethylation of isoeugenol followed by selective demethylation with alkali [149].

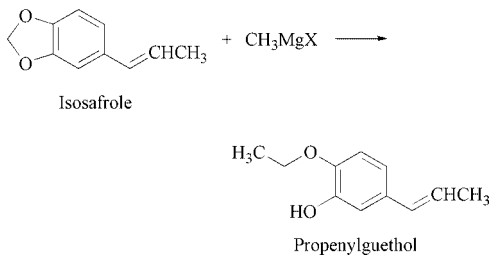

It is used in perfumery, for example, in soaps and cosmetics, to create or enhance vanilla notes.

***p*-Cresyl phenylacetate** [*101-94-0*], $C_{15}H_{14}O_2$, M_r 226.27, is prepared by esterification of *p*-cresol with phenylacetic acid. It forms crystals (*mp* 75 – 76 °C) with a narcissus odor and a honey note. It is used in blossom compositions with a slight animal note.

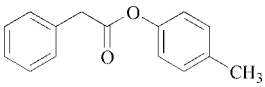

2.6.2. Phenol Alcohols and their Esters

In comparison with the araliphatic alcohols discussed in Section 2.5.5, very few phenol alcohols are used as fragrance and flavor materials. None of the alcohol derivatives of vanillin, ethylvanillin, and heliotropin or their esters have special organoleptic properties. Anise alcohol and its acetate are the only products that are used to some extent in perfume and aroma compositions.

Anisyl alcohol [*105-13-5*], 4-methoxybenzyl alcohol, $C_8H_{10}O_2$, M_r 138.17, *bp* (1.3 kPa) 136 °C, d_4^{20} 1.1140, n_D^{25} 1.5420, occurs in vanilla pods and in anise seeds. It is a colorless liquid with a sweet-flowery, slightly balsamic odor.

Pure anise alcohol for perfumery and flavor purposes is prepared by hydrogenation of anisaldehyde. It is used in perfumery in blossom compositions (e.g., lilac and gardenia types) and in flavors for confectionery and beverages.

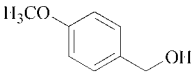

Anisyl acetate [*104-21-2*], $C_{10}H_{12}O_3$, M_r 180.20, *bp* (1.5 kPa) 133 °C, d_4^{20} 1.1084, has been found in several types of berries. It is a colorless liquid with a fruity, slightly balsamic blossom odor and is used occasionally in sweet-flowery compositions, but more frequently in flavor compositions for fruity notes.

2.6.3. Phenol Aldehydes

Phenol aldehydes are generally pleasant-smelling products. Some of them are particularly important as fragrance and flavor compounds. Anisaldehyde and certain derivatives of protocatechu aldehyde (3,4-dihydroxybenzaldehyde) are well-known representatives. The monomethyl ether of protocatechu aldehyde, vanillin, is perhaps the most widely used flavor compound. Other important derivatives of this aldehyde are veratraldehyde (dimethyl ether) and piperonal (formaldehyde acetal derivative); they are not only used as fragrance and flavor substances, but also are intermediates in many industrial processes.

***p*-Anisaldehyde** [*123-11-5*], 4-methoxybenzaldehyde, $C_8H_8O_2$, M_r 136.15, *bp* (1.85 kPa) 132 °C, d_4^{25} 1.1192, n_D^{25} 1.5703, occurs in many essential oils, often together with anethole. It is a colorless to slightly yellowish liquid with a sweet, mimosa, hawthorn odor. *p*-Anisaldehyde can be hydrogenated to anise alcohol and readily oxidizes to anisic acid when exposed to air. Synthetic routes to anisaldehyde usually involve the oxidation of *p*-cresyl methyl ether. Manganese dioxide and sulfuric acid are usually used for oxidation. In a Russian process, *p*-cresyl methyl ether is oxidized with alkali peroxysulfates in the presence of silver salts [150].

p-Anisaldehyde

Other industrial processes are the liquid-phase oxidation in the presence of cobalt catalysts [151] and the electrochemical oxidation in the presence of lower aliphatic alcohols via the corresponding anisaldehyde dialkyl acetal [152].

p-Anisaldehyde is frequently used in sweet blossom compositions (e.g., in lilac and hawthorn types) as well as in flavor compositions for confectioneries and beverages. *p*-Anisaldehyde is an intermediate in many industrial processes. Its hydrogensulfite derivative is used as a brightener for metals in galvanic baths.

Vanillin [*121-33-5*], 4-hydroxy-3-methoxybenzaldehyde, $C_8H_8O_3$, M_r 152.15, *bp* (1.3 kPa) 155 °C, *mp* 82 – 83 °C, d_4^{20} 1.056, is found in many essential oils and foods, but is often not essential for their odor or aroma. However, it does determine the odor of essen-

tial oils and extracts from *Vanilla planifolia* and *V. tahitensis* pods, in which it is formed during ripening by enzymatic cleavage of glycosides.

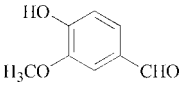

Annually about 12 000 t vanillin are produced synthetically and about 20 t originate from natural sources [153].

Properties. Vanillin is a colorless crystalline solid with a typical vanilla odor. Because it possesses aldehyde and hydroxyl substituents, it undergoes many reactions. Additional reactions are possible due to the reactivity of the aromatic nucleus. Vanillyl alcohol and 2-methoxy-4-methylphenol are obtained by catalytic hydrogenation; vanillic acid derivatives are formed after oxidation and protection of the phenolic hydroxyl group. Since vanillin is a phenol aldehyde, it is stable to autoxidation and does not undergo the Cannizzaro reaction. Numerous derivatives can be prepared by etherification or esterification of the hydroxyl group and by aldol condensation at the aldehyde group. For example, several of these derivatives are intermediates in the synthesis of pharmaceuticals.

Production. Most commercial vanillin is obtained by processing waste sulfite liquors; the remainder is synthesized from guaiacol. Preparation by oxidation of isoeugenol is of historical interest only.

Preparation from Waste Sulfite Liquors:
The starting material for vanillin production is the lignin present in sulfite wastes from the cellulose industry (→ Lignin). The concentrated mother liquors are treated with alkali at elevated temperature and pressure in the presence of oxidants. The vanillin formed is separated from the byproducts, particularly acetovanillone (4-hydroxy-3-methoxyacetophenone), by extraction, distillation, and crystallization.
A large number of patents describe various procedures for the (mainly) continuous hydrolysis and oxidation processes, as well as for the purification steps required to obtain high-grade vanillin [154]. Lignin is degraded either with sodium hydroxide or with calcium hydroxide solution and simultaneously oxidized in air in the presence of catalysts. When the reaction is completed, the solid wastes are removed. Vanillin is extracted from the acidified solution with a solvent (e.g., butanol or benzene) and reextracted with sodium hydrogen sulfite solution. Reacidification with sulfuric acid followed by vacuum distillation yields technical-grade vanillin, which must be recrystallized several times to obtain food-grade vanillin. Water, to which some ethanol may be added, is used as the solvent in the last crystallization step.

Preparation from Guaiacol and Glyoxylic Acid.
Several methods can be used to introduce an aldehyde group into an aromatic ring. Condensation of guaiacol with glyoxylic acid followed by oxidation of the resulting mandelic acid to the corresponding phenylglyoxylic acid and, finally, decarboxylation continues to be a competitive industrial process for vanillin synthesis.
Currently, guaiacol is synthesized from catechol, which is prepared by acid-catalyzed hydroxylation of phenol with hydrogen peroxide. Glyoxylic acid is obtained as a byproduct in the synthesis of glyoxal from acetaldehyde and can also be produced by oxidation of glyoxal with nitric acid. Condensation of guaiacol with glyoxylic acid proceeds smoothly at room temperature and in weakly alkaline media. A slight excess of guaiacol is maintained to avoid formation of disubstituted products; excess guaiacol is recovered. The alkaline solution containing 4-hydroxy-3-methoxymandelic acid is then oxidized in air in the presence of a catalyst until the calculated amount of oxygen is consumed [155]. Crude vanillin is obtained by acidification and simultaneous decarboxylation of the (4-hydroxy-3-methoxyphenyl) glyoxylic acid solution. Commercial grades are obtained by vacuum distillation and subsequent recrystallization as described under method (1).
This process has the advantage that, under the reaction conditions, the glyoxyl radical enters the aromatic guaiacol ring almost exclusively para to the phenolic hydroxyl group. Tedious separation procedures are thus avoided.

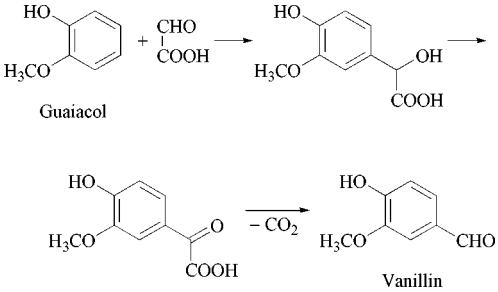

Uses. The main application of vanillin is the flavoring of foods (e.g., ice cream, chocolate, bakery products, and confectioneries). Small quantities are used in perfumery to round and fix sweet, balsamic fragrances. Vanillin is also used as a brightener in galvanotechnical processes and is an important intermediate in, for example, the production of pharmaceuticals such as L-3,4-dihydroxyphenylalanine (L-DOPA) and methyldopa.

Veratraldehyde [*120-14-9*], 3,4-dimethoxybenzaldehyde, $C_9H_{10}O_3$, M_r 166.18, occurs in a few essential oils and is a crystalline solid (*mp* 44.5 – 45 °C) with a woody, vanilla-like odor.

Veratraldehyde can be prepared by methylation of vanillin. It is used in oriental and warm-woody fragrances, as well as in flavor compositions for vanilla notes. It is an intermediate in, for example, the synthesis of pharmaceuticals.

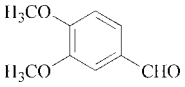

Ethylvanillin [*121-32-4*], 3-ethoxy-4-hydroxybenzaldehyde, $C_9H_{10}O_3$, M_r 166.18, *mp* 77 – 78 °C, does not occur in nature. Its odor resembles that of vanillin but is approximately three times as strong. Ethylvanillin can be prepared by method (2) as described for vanillin, using guethol instead of guaiacol as the starting material.

Ethylvanillin is used extensively in the chocolate industry. It gives a sweet, balsamic note to flowery and fruity perfume compositions.

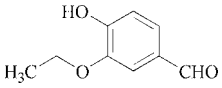

Piperonal [*120-57-0*], heliotropin, 3,4-methylenedioxybenzaldehyde, $C_8H_6O_3$, M_r 150.13, *bp* (1.6 kPa) 139.4 °C, *mp* 37 °C, $d_4^{43.2}$ 1.2792, occurs in a number of essential oils, but never as the main component. It forms white crystals with a sweet-flowery, slightly spicy, heliotrope-like odor.

Production. Piperonal is produced by two main routes:

From Isosafrole.
For many years, oxidative cleavage of isosafrole was the only route applicable on an industrial scale. Examples of oxidants that give good yields of piperonal are chromium(VI) salts, oxygen, and ozone.

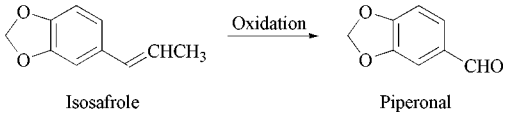

This method is still used currently because safrole (the starting material for isosafrole) can be isolated from essential oils relatively inexpensively and in sufficient quantity.

From Catechol. Several routes have been developed for the synthesis of piperonal from catechol. In one such route, catechol is converted into 3,4-dihydroxymandelic acid with glyoxylic acid in an alkaline medium in the presence of aluminum oxide. 3,4-Dihydroxymandelic acid is oxidized to the corresponding keto acid (e.g., with copper(II) oxide), which is decarboxylated to 3,4-dihydroxybenzaldehyde [156]. The latter product is converted into piperonal, for

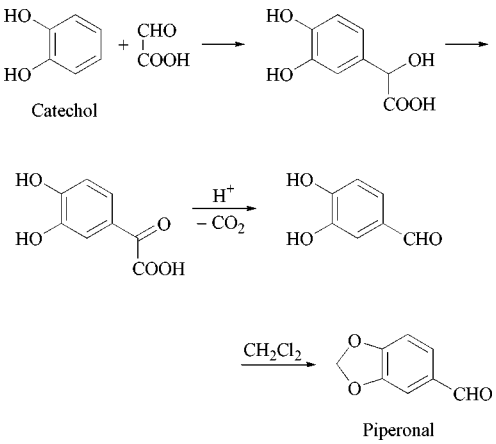

example, by reaction with methylene chloride in the presence of quaternary ammonium salts [157].

In another route, catechol is first reacted with methylene chloride and converted into 1,2-methylenedioxybenzene [157]. Reaction with glyoxylic acid in strongly acidic media yields 3,4-methylenedioxymandelic acid [158]. Subsequent oxidation and decarboxylation with nitric acid affords piperonal.

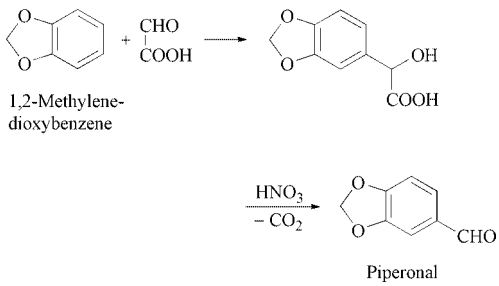

Uses. Piperonal is used in many flowery-spicy fine fragrances and is also an important ingredient of flavor compositions.

2.6.4. Phenol Ketones

Few of the phenol derivatives that have a keto substituent in their side-chain are of interest as fragrance or flavor substances. A number of phenols and phenyl ethers acetylated in the benzene ring have been identified as volatile components of foods. 4-Methoxyacetophenone is of some interest as a fragrance compound. 4-Hydroxybenzylacetone, a higher mass phenol ketone, has a characteristic raspberry aroma.

4-Methoxyacetophenone [*100-06-1*], acetanisole, $C_9H_{10}O_2$, M_r 150.18, d^{41} 1.0818, n_D^{41} 1.5470, occurs in anise oil. It forms white crystals (*mp* 38 °C) with a sweet odor, reminiscent of hawthorn. 4-Methoxyacetophenone is prepared by Friedel–Crafts acetylation of anisole and is used in soap perfumes.

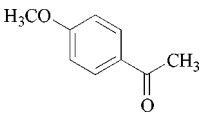

1-(4-Hydroxyphenyl)-3-butanone [*5471-51-2*], raspberry ketone, $C_{10}H_{12}O_2$, M_r 164.20, *mp* 82 – 83 °C, is a highly characteristic component of raspberry aroma. It forms colorless crystals with a sweet-fruity odor strongly reminiscent of raspberries.

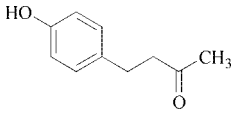

Raspberry ketone is prepared by alkali-catalyzed condensation of the alkali salt of 4-hydroxybenzaldehyde and acetone, followed by selective hydrogenation of the double bond in the resulting 4-hydroxybenzalacetone. Other syntheses start from phenol which is converted into 1-(4-hydroxyphenyl)-3-butanone with methyl vinyl ketone (e.g., in the presence of phosphoric acid) [159] or with 3-oxo-1-butanol in the presence of concentrated sulfuric acid [160].

The ketone is used in fruit flavors, particularly in raspberry compositions.

Trade Names. Frambinon (DRAGOCO), Oxyphenylon (IFF).

2.6.5. Phenol Carboxylates

Alkyl and aralkyl salicylates, as well as methyl 3-methylorsellinate (an oakmoss fragrance) are important phenol carboxylates that are used in flavors and fragrances. The following salicylates are used in perfume and flavor compositions and can be prepared by esterification of salicylic acid.

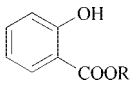

Methyl salicylate [*119-36-8*], R = CH_3, $C_8H_8O_3$, M_r 152.15, *bp* (1.6 kPa) 98 °C, d_4^{25} 1.1782, n_D^{25} 1.5350, is the main component of wintergreen oil and occurs in small quantities in other essential oils and fruit. It is a colorless liquid with a sweet, phenolic odor. Methyl salicylate is used in perfumery as a modifier in blossom fragrances and as a mild antiseptic in oral hygiene products.

Isoamyl salicylate [*87-20-7*], R = $CH_2CH_2CH(CH_3)_2$, $C_{12}H_{16}O_3$, M_r 208.26, *bp* (2 kPa) 151 – 152 °C, d_4^{20} 1.0535, n_D^{20} 1.5065, has been found in a number of fruit aromas. It is a colorless liquid with a sweet, clover-like odor and is used in perfumery for floral and herbal notes, particularly in soap perfumes.

Hexyl salicylate [*6259-76-3*], R = $(CH_2)_5CH_3$, $C_{13}H_{18}O_3$, M_r 222.28, *bp* (1.6 kPa) 167 – 168 °C, d_{25}^{25} 1.035, n_D^{25} 1.5049, has been reported in carnation flower absolute [161]. It is a colorless liquid with a green, flowery-spicy odor, reminiscent of azaleas. It is used for blossom and herbal notes in perfumes, e.g., in soap, personal hygiene products, and detergents.

***cis*-3-Hexenyl salicylate** [*65405-77-8*], R = *cis*-$(CH_2)_2CH{=}CHCH_2CH_3$, $C_{13}H_{16}O_3$, M_r 220.27, *bp* (150 Pa) 125 °C, d_{25}^{25} 1.0589, n_D^{20} 1.5210, has been identified in carnation flower absolute. It is a colorless liquid with a long-lasting, sweet, green balsamic odor. It is used in fine fragrances and for scenting soaps, cosmetics, and detergents.

Benzyl salicylate [*118-58-1*], R = $CH_2C_6H_5$, $C_{14}H_{12}O_3$, M_r 228.25, *bp* (1.3 kPa) 186 – 188 °C, d_4^{20} 1.1799, n_D^{20} 1.5805, which occurs in several essential oils, is a colorless, viscous liquid with a weak, sweet, slightly balsamic odor. Benzyl salicylate is used as a fixative in flowery-spicy perfume compositions and in flavors.

Phenethyl salicylate [*87-22-9*], R = $CH_2CH_2C_6H_5$, $C_{15}H_{14}O_3$, M_r 242.27, *mp* 44 °C, which has not been reported to occur in nature, is a crystalline solid with a weak, long-lasting, balsamic, blossom odor, reminiscent of rose and hyacinth. It is used in perfumery for spicy and balsamic blossom compositions.

Methyl 3-methylorsellinate [*4707-47-5*], methyl 2,4-dihydroxy-3,6-dimethylbenzoate, $C_{10}H_{12}O_4$, M_r 196.20, *mp* 145 °C, is an odor-determining constituent of oakmoss extract and forms colorless crystals with a mossy-earthy odor. It can be prepared from acyclic compounds. In an industrial process, dimethyl malonate is condensed with 4-hexen-3-one (or a mixture of this ketone and 5-chlorohexan-3-one) to give a substituted 3-hydroxy-2-cyclohexenone (**33**). Aromatization, in good yield, is achieved by reaction of the hydroxycyclohexenone with a suitable *N*-haloamide. The intermediate 3-hydroxy-2-cyclohexenone can also be obtained by condensation of methyl 3-oxopentanoate with methyl crotonate [162].

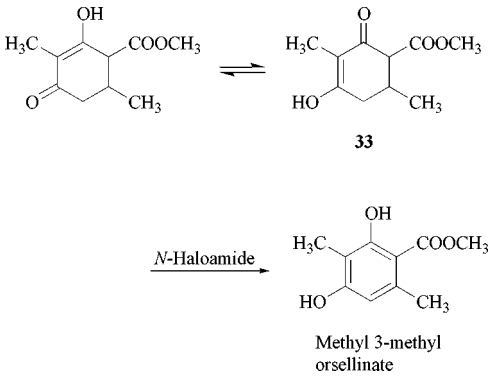

Another route to methyl 3-methylorsellinate starts from β-orcinol, which is converted into a carboxylic acid derivative by reaction with carbon dioxide (Kolbe synthesis). The acid is esterified (e.g., with dimethyl sulfate) [163].

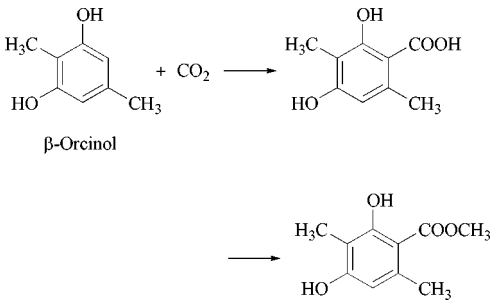

Methyl 3-methylorsellinate is used as a substitute for oakmoss extract in fine fragrances, soaps, and cosmetics.

Trade Name. Evernyl (Givaudan Roure), Veramoss (IFF).

2.7. O- and O, S-Heterocycles

2.7.1. Cyclic Ethers

Cyclic ethers used as fragrances include a number of terpenoid compounds. Some of them, such as 1,4-cineole [*470-67-7*] and 1,8-cineole, occur in essential oils in significant quantities. Others are only minor components; examples are rose oxide, nerol oxide [*1786-08-9*], and rose furan [*15186-51-3*], which contribute to the specific fragrance of rose oil, and caryophyllene oxide [*1139-30-6*], which has a woody, slightly ambergris-like odor. α-Cedrene oxide [*11000-57-0*] is another wood-fragrance compound, that can be easily prepared by epoxidation of cedarwood oil hydrocarbons.

Except for some of the above-mentioned compounds, only a few other cyclic ethers are important, for instance, 4,6,6,7,8,8-hexamethyl-1,3,4,6,7,8-hexahydropenta[*g*]benzopyran, a musk fragrance that is used in large amounts.

Numerous furan and pyran derivatives, many of which originate from heat treatment of carbohydrates, largely determine the odor of processed foods. Of this group, 2,5-dimethyl-4-hydroxy-2*H*-furan-3-one and maltol are used in fairly large quantities in flavors. The following compounds are used in relatively small amounts in flavor compositions:

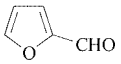

2-Furaldehyde [*98-01-1*], freshyl baked bread odor

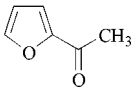

2-Acetylfuran [*1192-62-7*], sweet balsamic odor

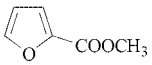

Methyl 2-furoate [*611-13-2*], fruity, mushroom-like odor

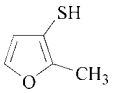

2-Methylfuran-3-thiol [*28588-74-1*], roast beef aroma

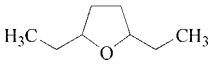

2,5-Diethyltetrahydrofuran [*41239-48-9*], fruity, herbal-minty note

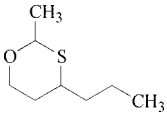

2-Methyl-4-propyl-1,3-oxathiane [*67715-80-4*], typical sulfury note of tropical fruit

1,8-Cineole [*470-82-6*], 1,8-epoxy-*p*-menthane, eucalyptol, $C_{10}H_{18}O$, M_r 154.25, *bp* (101.8 kPa) 176 – 177 °C, *fp* 1 °C, d_4^{20} 0.9267, n_D^{20} 1.4586, occurs in many terpene-containing essential oils, sometimes as the main component. For example, eucalyptus oils contain up to 85 % 1,8-cineole and laurel leaf oil contains up to 70 %. It is a colorless liquid with a characteristic odor, slightly reminiscent of camphor.

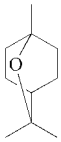

1,8-Cineole is one of the few fragrance materials that is obtained exclusively by isolation from essential oils, especially eucalyptus oils. Technical-grade 1,8-cineole with a purity of 99.6 – 99.8 % is produced in Spain in large quantities by fractional distillation of *Eucalyptus globulus* Labillardière. A product essentially free from other products can be obtained by crystallization of cineole-rich eucalyptus oil fractions.

1,8-Cineole has a fresh odor and is used in large quantities in fragrances as well as in flavors (e.g., in oral hygiene products).

4-Methyl-2-(2-methyl-1-propenyl)tetrahydropyran [*16409-43-1*], rose oxide, $C_{10}H_{18}O$, M_r 154.25, *bp* (1.6 kPa) 70 °C, d_4^{20} 0.875, n_D^{20} 1.4570, $[\alpha]_D$ for the optically pure (−)-*cis* form − 58.1 °, occurs in small quantities, mainly the levorotatory *cis* form, in essential oils (e.g., Bulgarian rose oil and geranium oil). Commercial synthetic products are either optically active or inactive mixtures of the *cis*- and

trans-isomers. Their physical constants, particularly the optical rotation, depend on the starting material and the method of synthesis. They are colorless liquids with a strong odor reminiscent of geranium oil and carrot leaves. Rose oxide is usually prepared from citronellol (see page 19), which can be converted into a mixture of two allyl hydroperoxides (e.g., by photosensitized oxidation with oxygen). Reduction of the hydroperoxides with sodium sulfite yields the corresponding diols [164]. Treatment with dilute sulfuric acid results in allylic rearrangement and spontaneous cyclization of one of the isomers; a mixture of diastereoisomeric rose oxides is thus formed. The unreacted diol isomer is separated by distillation. (−)-Citronellol as the starting material yields approximately a 1 : 1 mixture of (−)-*cis*- and (−)-*trans*-rose oxide.

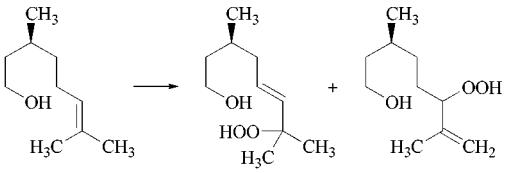

(−)-Citronellol

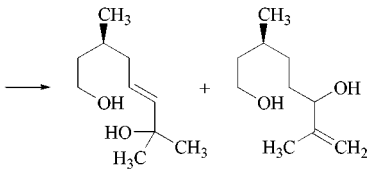

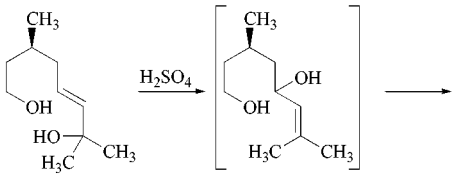

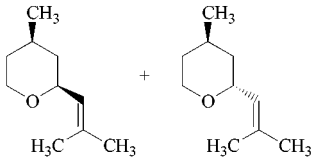

(−)-*cis*-
Rose oxide

(−)-*trans*-
Rose oxide

Rose oxide is used in rose and geranium perfumes.

Trade Name. Rose oxide (DRAGOCO).

Menthofuran [*494-90-6*], 3,6-dimethyl-4,5,6,7-tetrahydrobenzofuran, $C_{10}H_{14}O$, M_r 150.22, *bp* (1.3 kPa) 78 – 79 °C, d_4^{20} 0.9676, n_D^{20} 1.4855, $[\alpha]_D^{20}$ +94.6 °, occurs mainly as the (+)-isomer, in numerous essential oils (e.g., Mentha oils). It is a colorless liquid with a minty odor.

(+)-Menthofuran [*17957-94-7*] is isolated from Mentha oils or is prepared synthetically, for example, by treatment of (+)-pulegone with fuming sulfuric acid in acetic anhydride and pyrolysis of the resulting sultone.

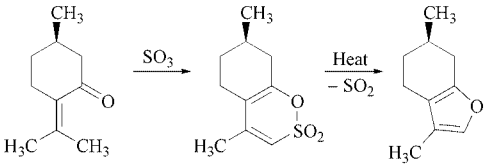

(+)-Pulegone

Menthofuran

Menthofuran is used mainly in peppermint oil reconstitutions.

Linalool oxide, 2-methyl-2-vinyl-5-(α-hydroxyisopropyl)tetrahydrofuran, $C_{10}H_{18}O_2$, M_r 170.25, *bp* (101.3 kPa) 188 °C, d_4^{20} 0.939 – 0.944, n_D^{20} 1.451 – 1.455, has been identified in essential oils and in fruit aromas. Commercial linalool oxide is a mixture of the *cis* and *trans* forms, [*5989-33-3*] and [*34995-77-2*], respectively. It is a liquid with an earthy-flowery, slightly bergamot-like odor.

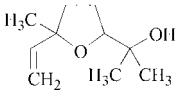

Linalool oxide

Linalool oxide is prepared by oxidation of linalool (see page 16). The isomeric compound 2,2,6-trimethyl-6-vinyltetrahydro-2*H*-pyran-3-ol [*14049-11-7*], which also occurs in nature, is formed as a byproduct:

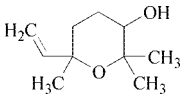

Linalool oxide is used in perfumery (e.g., for lavender notes) and for reconstitution of essential oils.

A dehydrated linalool oxide, 2-methyl-2-vinyl-5-isopropenyltetrahydrofuran [*13679-86-2*], occurs naturally; it has a minty eucalyptol odor and is used in perfumery.

3a,6,6,9a-Tetramethyldodecahydronaphtho-[2,1-*b*]furan [*6790-58-5*] (**34**), $C_{16}H_{28}O$, M_r 236.40, *mp* 75 – 76 °C, is a crystalline autoxidation product of ambrein (see Ambergris, page 87) with a typical ambergris odor. It is prepared from sclareol, a diterpene alcohol obtained from plant waste in the production of clary sage oil (see Sage Oils, page 115). Oxidative degradation to a lactone, hydrogenation of the latter to the corresponding diol and dehydration yield (**34**).

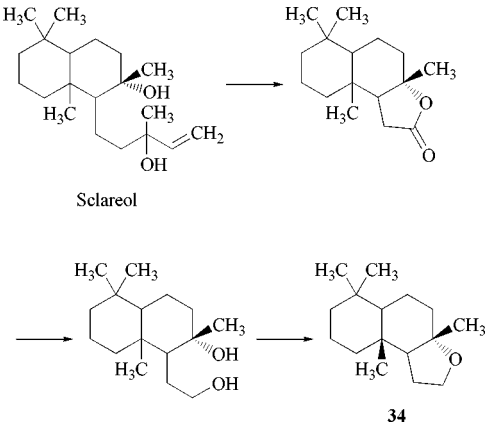

The product is used in perfumery for creating ambergris notes.

Trade Names. Ambrox (Firmenich), Ambroxan (Cognis).

2,5-Dimethyl-4-hydroxy-2*H*-furan-3-one [*3658-77-3*], $C_6H_8O_3$, M_r 128.13, is a constituent of pineapple aroma and is also found in other foods. It forms colorless crystals (*mp* 77 – 79 °C) with a relatively weak, nonspecific odor. Dilute solutions develop a pineapple, strawberry-like odor. It can be prepared by cyclization of hexane-2,5-diol-3,4-dione in the presence of an acidic catalyst [165]. The dione is the ozonization product of 2,5-hexynediol, which is obtained by ethynylation of acetaldehyde.

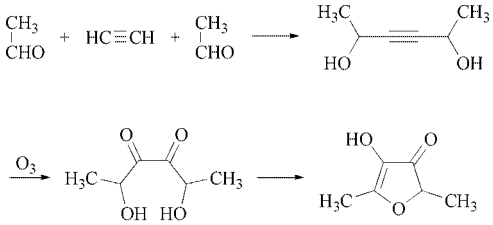

2,5-Dimethyl-4-hydroxy-2*H*-furan-3-one

In another process, a dialkyl α-methyldiglycolate (formed from an alkyl lactate and an alkyl monochloroacetate) is reacted with dialkyl oxalate in the presence of a sodium alkoxide and dimethylformamide. The reaction product is cyclized, alkylated, hydrolyzed, and decarboxylated [166].

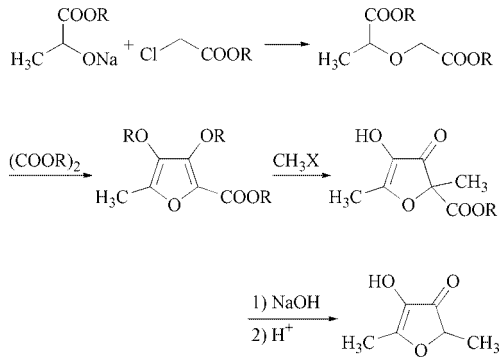

The compound is used in the flavoring of foods.

Trade Name. Furaneol (Firmenich).

Maltol [*118-71-8*], 3-hydroxy-2-methyl-4*H*-pyran-4-one, $C_6H_6O_3$, M_r 126.11, *mp* 162 – 164 °C, occurs in pine needles and the bark of young larch trees. It is produced when cellulose or starch are heated and is a constituent of wood tar oils. It forms crystals with a caramel-like odor, reminiscent of freshly baked cakes.

Although many routes are known for its synthesis, maltol is still isolated mainly from beechwood tar. It is used in aroma compositions with a caramel note and as a taste intensifier, for example, in fruit flavors (particularly in strawberry flavor compositions).

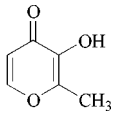

3-Pentyltetrahydro-2*H*-pyran-4-ol acetate [*18871-14-2*], $C_{12}H_{22}O_3$, M_r 214.30, *bp* (0.13 kPa) 102 – 103 °C, d_{25}^{25} 0.974 – 0.978, n_D^{20} 1.448 – 1.451, is a colorless to pale yellow liquid with a sweet-floral, fruity, slightly woody, jasmin-like odor. It is prepared by a Prins reaction of 1-octene with formaldehyde and acetic acid and is used in perfumes for various flower types, especially jasmin.

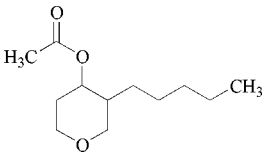

Trade Names. Jasmophyll (Haarmann & Reimer), Jasmopyran (Quest), Jasmal (IFF).

4,6,6,7,8,8-Hexamethyl-1,3,4,6,7,8-hexahydrocyclopenta[*g*]benzopyran [*1222-05-5*] (**35**), $C_{18}H_{26}O$, M_r 258.40, *bp* (1.1 kPa) 129 °C, d_4^{20} 1.0054, n_D^{20} 1.5342, is a viscous liquid with a musklike odor. It is one of the most frequently used synthetic, artificial musk fragrances. The starting material for its synthesis is 1,1,2,3,3-pentamethylindane, which is prepared by cycloaddition of *tert*-amylene to α-methylstyrene. The pentamethylindane is hydroxyalkylated with propylene oxide in a Friedel–Crafts reaction using aluminum chloride as a catalyst (analogous to the synthesis of 2-phenethyl alcohol from benzene and ethylene oxide, see page 51). Ring closure of the resulting 1,1,2,3,3-pentamethyl-5-(β-hydroxyisopropyl)indane to (**35**) is accomplished with paraformaldehyde and a lower aliphatic alcohol via the acetal [167] or with paraformaldehyde and a carboxylic acid anhydride via the acylate [168].

The commercial product is diluted with solvents (e.g., diethyl phthalate, isopropyl myristate, benzyl benzoate) to make it less viscous. It is alkali-stable and does not discolor in light. Therefore, it is a popular ingredient of perfume compositions for soaps, detergents, and cosmetics and is used in large amounts.
Trade Names. Galaxolide (IFF).

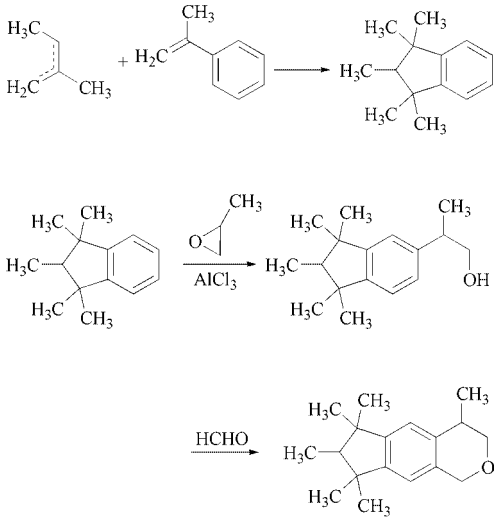

35

4,4*a*,5,9*b*-Tetrahydroindeno[1,2-*d*]-1,3-dioxane [*18096-62-3*], $C_{11}H_{12}O_2$, M_r 176.21, *bp* (1.3 kPa) 134 °C, *mp* 35 – 36 °C, $d_{15.5}^{15.5}$ 1.159, n_D^{20} 1.559, forms crystals with an indole-like odor. It is prepared by a Prins reaction from indene and formaldehyde in the presence of dilute sulfuric acid [169]. It is used in perfumes for soaps and detergents.

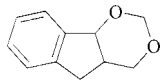

Trade Names. Indoflor (Haarmann & Reimer), Indolarom (IFF).

The 2,4-dimethyl homologue [*27606-09-3*] is prepared by using acetaldehyde instead of formaldehyde. Its odor is reminiscent of geranium and magnolia [171].

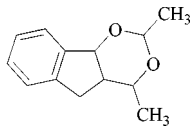

Trade Name. Magnolan (Haarmann & Reimer).

2,4,6-Trimethyl-4-phenyl-1,3-dioxane [*5182-36-5*], $C_{13}H_{18}O_2$, M_r 206.28, *bp* (300 Pa) 100 °C, d_4^{20} 1.159, n_D^{20} 1.504, is a colorless to pale yellow liquid. This cyclic acetal has a smell of grapefruit with herbal and fresh aspects.

2,4,6-Trimethyl-4-phenyl-1,3-dioxane can be produced by a Prins reaction of methylstyrene with acetaldehyde.

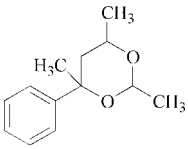

Trade Name. Vertacetal (DRAGOCO), Floropal (Haarmann & Reimer).

Isolongifolanone ethylene glycol acetal [*154171-77-4*], $C_{17}H_{28}O2$, M_r 264.41, *bp* (3 kPa) 142 – 143 °C, d_4^{20} 1.0510, n_D^{20}1.5051 is commercially available as a colorless liquid with strong woody odor with flowery-fresh effects and with a velvety moss/ambra accent. It has an extremely long lasting odor and acts as fixative. Isolongifolanone ethylene glycol acetal is frequently used in soaps, cosmetics, and perfumery compositions.

For the production of isolongifolanone acetal (**40**), longifolene (**36**), which is a main component of Indian turpentine oil, is first isomerized to isolongifolene (**37**). The epoxide (**38**), obtained by epoxidation of isolongifolene, is subsequently transformed to isolongifolanone (**39**), which is reacted with ethylene glycol to the cyclic isolongifolanone ethyleneglycolacetal (**40**) [170].

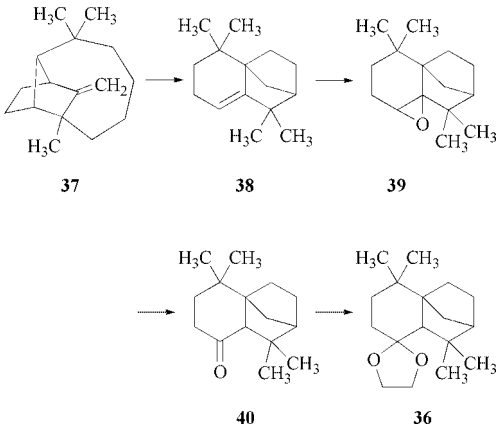

Trade Name. Ysamber-K (DRAGOCO).

5,5,7,9,9,13-Hexamethyl-4,6-dioxatetracyclo[6.5.1.^{1,10}0^{3,7}]tetradecane [*211299-54-6*], cedrene acetonide, $C_{18}H_{30}O_2$, M_r 278.44, *mp* 157 – 158 °C, is a fragrance compound with a powdery-woody, long-lasting amber note. Its industrial synthesis starts with the epoxidation of cedarwood oil fractions that contain sesquiterpene hydrocarbons, mainly α-cedrene. Acid-catalyzed ring opening and acetalization leads to cedrene acetonide [172].

The pure compound is crystalline and is frequently used in soaps, cosmetics and perfumery compositions as a fixative with high substantivity.

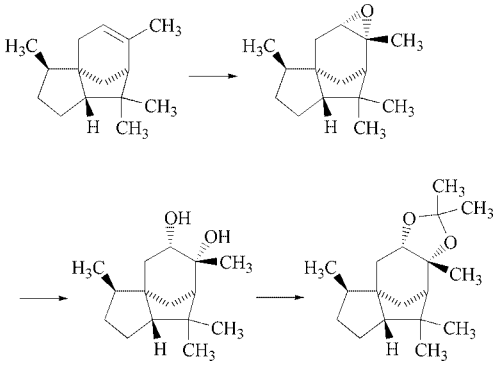

Trade Name. Ambrocenide 10 (DRAGOCO).

2.7.2. Lactones

Naturally occurring organoleptically important lactones are mainly saturated and unsaturated γ- and δ-lactones, and to a lesser extent macrocyclic lactones. The occurrence of these types of lactones reflects their ready formation from natural acyclic precursors.

The γ- and δ-lactones are the intramolecular esters of the corresponding hydroxy fatty acids. They contribute to the aroma of butter and various fruits. 3,4-Dimethyl-5-pentylidenedihydro-2(5*H*)-furanone [*774-64-1*] is an unsaturated γ-lactone with a distinct nutty flavor. 15-Pentadecanolide is responsible for the musklike odor of angelica root oil. Of the naturally occurring bicyclic lactones, phthalides are responsible for the odor of celery root oil, and coumarin for woodruff.

The macrocyclic esters hold a special position among the industrially produced lactone fra-

grance compounds. Like the well-known macrocyclic ketones (see Section 2.4), they have outstanding odor properties as musks. However, the lactones can be prepared more easily than the ketones, for example, by depolymerization of the corresponding linear polyesters. Since replacement of a methylene unit by oxygen affects the odor of these compounds very little, oxalactones with 15 – 17-membered rings are commercially produced in addition to 15-pentadecanolide. Several cyclic diesters prepared from long-chain α,ω-dicarboxylic acids and glycols are also valuable musk fragrances.

The γ-lactones described below can be prepared in good yield in a one-step process by radical addition of primary fatty alcohols to acrylic acid, using di-*tert*-butyl peroxide as a catalyst. A patent claims a high yield when the reaction is carried out in the presence of alkali phosphates or alkali sulfates [173].

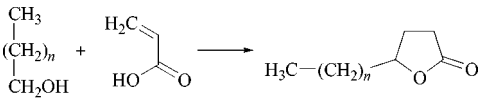

γ-Lactones

γ-Octalactone [*104-50-7*], ($n = 3$), $C_8H_{14}O_2$, M_r 142.20, *bp* (1.3 kPa) 116 – 117 °C, d_4^{20} 0.977, n_D^{25} 1.4420, occurs as an aroma constituent in many processed and unprocessed foods. It is a pale yellow liquid with a fruity-coconut-like odor and is used both in aroma compositions and in heavy blossom perfumes.

γ-Nonalactone [*104-61-0*], ($n = 4$), "so-called aldehyde C_{18}" $C_9H_{16}O_2$, M_r 156.22, *bp* (1.7 kPa) 136 °C, d_4^{20} 0.9676, n_D^{20} 1.446, occurs in many foods and is a pale yellow liquid with a coconut-like aroma. It has numerous applications, similar to those of γ-octalactone, in aroma compositions and perfumery.
Trade Names. Abricolin (Haarmann & Reimer), Prunolide (Givaudan Roure).

γ-Decalactone [*706-14-9*], ($n = 5$), $C_{10}H_{18}O_2$, M_r 170.25, *bp* (2.3 kPa) 156 °C, d_4^{21} 0.952, $n_D^{19.5}$ 1.4508, is present in a wide variety of foods and is an almost colorless liquid with an intensely fruity odor, reminiscent of peaches. It is used in perfumery for heavy, fruity flower odors and in aroma compositions, particularly peach flavors.

γ-Undecalactone [*104-67-6*], ($n = 6$), "so-called aldehyde C_{14}" $C_{11}H_{20}O_2$, M_r 184.28, *bp* (2.0 kPa) 167 – 169 °C, d_4^{20} 0.944, n_D^{20} 1.4514, occurs in foods and is an almost colorless liquid with a peachlike odor. In addition to preparation by radical addition of 1-octanol to acrylic acid, γ-undecalactone is also prepared by intramolecular cyclization of 10-undecylenic acid with 70 – 80 % sulfuric acid with migration of the double bond.

γ-Undecalactone has many applications in perfume and aroma compositions, similar to those of γ-decalactone.
Trade Name. Pêche pure (Givaudan Roure).

δ-Decalactone [*705-86-2*], $C_{10}H_{18}O_2$, M_r 170.25, *bp* (3 Pa) 117 – 120 °C, $d_4^{27.5}$ 0.9540, n_D^{26} 1.4537, is a flavor constituent of many types of fruit, cheese, and other dairy products. It is a colorless, viscous liquid with a creamy-coconut, peachlike aroma.

δ-Decalactone can be prepared by peracid oxidation of 2-pentylcyclopentanone. It is used in perfumes and for cream and butter flavorings.

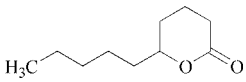

15-Pentadecanolide [*106-02-5*], ω-hydroxypentadecanoic acid lactone, $C_{15}H_{28}O_2$, M_r 240.39, *bp* (1.3 – 1.4 kPa) 169 °C, *mp* 37 – 38 °C, d_4^{40} 0.940, occurs in small quantities in, for example, angelica root oil. It forms colorless crystals with a delicate, musklike odor.

Production. The main industrial syntheses start from cyclododecatriene derivatives (→ Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene). The 16-membered ring is formed either by expansion of the 12-membered ring or by depolymerization of polyesters of ω-hydroxypentadecanoic acid.

Preparation by Ring Expansion of Cyclododecanone. Radical addition of allyl alcohol to cyclododecanone, for example, with di-*tert*-butyl peroxide as a radical initiator, yields 2-(γ-hydroxypropyl)cyclododecanone (**41**). This is converted into 13-oxabicyclo-[10.4.0]hexadec-1(12)-ene (**42**)

by acid-catalyzed dehydration [174]. Addition of hydrogen peroxide, in the presence of sulfuric acid, gives 12-hydroperoxy-13-oxabicyclo[10.4.0]hexadecane (**43**). Cleavage of the peroxide by heating in xylene gives 15-pentadecanolide as well as a small amount of 15-pentadec-11(and 12)-enolide and 11 12-hydroxy-15-pentadecanolide [175].

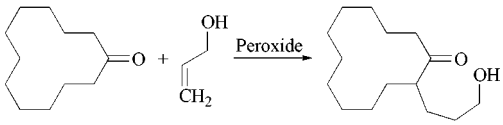

Cyclododecanone **41**

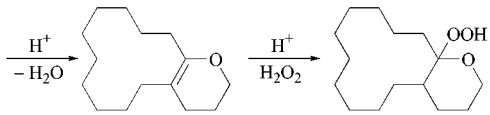

42 **43**

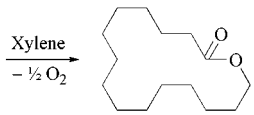

15-Pentadecanolide

Preparation from Polyesters of 15-Hydroxypentadecanoic Acid.
In a Japanese process, the required ω-hydroxy acid is prepared from 1,12-dodecanediol in several steps. The diol is added to methyl acrylate in a radical reaction, using di-*tert*-butyl peroxide as a catalyst. The free hydroxyl group in the resulting ω-hydroxy-γ-pentadecalactone is acetylated with acetic anhydride, and the resulting ω-acetoxy-γ-pentadecalactone is converted into 15-hydroxypentadecanoic acid by hydrogenolysis and hydrolysis [176] (see upper one of following figures).
Another route to 15-hydroxypentadecanoic acid starts from 13-oxabicyclo[10.4.0]hexadec-(1,12)-ene (**42**), see (1). Nitrosation to 12-oximino-15-pentadecanolide, followed by a Wolff–Kishner reduction yields 15-hydroxypentadecanoic acid [177]. The reaction can be carried out without isolation of the intermediate oxime.

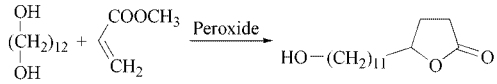

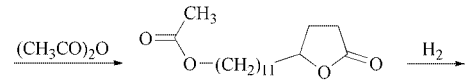

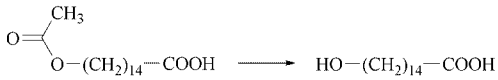

15-Hydroxypentadecanoic acid

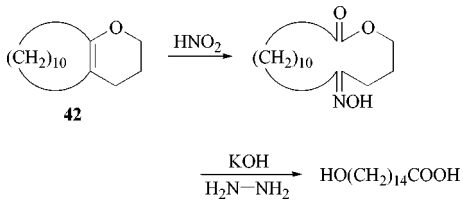

42

The polyester of 15-hydroxypentadecanoic acid is prepared by customary methods and is cleaved under high vacuum in the presence of transesterification catalysts.

Uses. 15-Pentadecanolide is a highly valuable fragrance material that is used in fairly large amounts in fine fragrances as a fixative with a delicate musk odor.
Trade Names. Cyclopentadecanolid (Haarmann & Reimer), Exaltolid (Firmenich), Pentalide (Soda Aromatic).

Oxacyclohexadec-12/13-en-2-one, (*E/Z*) [*34902-57-3*], $C_{15}H_{26}O_2$, M_r 238, *bp* (40 Pa) 88 – 93 °C , d_{20}^{20} 0.961 – 0.971, possesses a musklike, ambrette, and animalic odor.

Production (see production of 15-pentadecanolide page 77). Cleavage of 12-hydroperoxy-13-oxabicyclo[10.4.0]-hexadecane (**43**) in the presence of $Cu(OAc)_2$ and $Fe(SO_4)_2$ gives 43 % of *trans*-pentadec-11-enolide, 26 % *trans*-pentadec-12-enolide, 18 % *cis*-pentadec-11-enolide, 5 % *cis*-pentadec-12-enolide, and 5 % pentadecanolide [178].

Another synthesis uses a pyrolysis of an acetoxy derivative of the lactone [179].

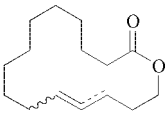

Uses. 15-Pentadec-11/12-enolide is a very powerful musk and combines the strength and tenacity of the aromatic musks with the pleasantness and elegance of the macrocyclic musks.
Trade Names. Habanolide (Firmenich), Globalide (Haarmann & Reimer).

12-Oxa-16-hexadecanolide [*6707-60-4*], 16-hydroxy-12-oxahexadecanoic acid lactone, $C_{15}H_{28}O_3$, M_r 256.38, does not occur in nature. Its odor is comparable to that of 15-pentadecanolide, but less intense. It is prepared by reacting methyl ω-bromoundecanoate with the monosodium salt of 1,4-butanediol. The resulting methyl 16-hydroxy-12-oxapalmitate is condensed to the corresponding polyester, which is subsequently depolymerized.

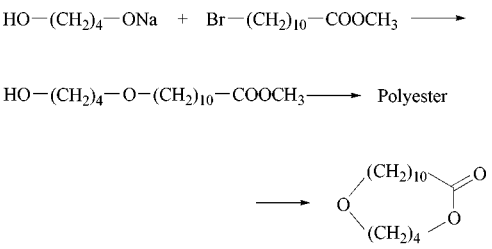

12-Oxa-16-hexadecanolide

12-Oxa-16-hexadecanolide as well as the stronger smelling 11-oxa isomer [*3391-83-1*], which is obtained in the same way from the corresponding hydroxyoxa acid, are used as substitutes for 15-pentadecanolide in fine fragrances.

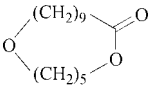

11-Oxa-16-hexadecanolide

Trade Names. 12-Oxa-16-hexadecanolide = Cervolide (Quest), 11-Oxa-16-hexadecanolide = Musk R 1 (Quest).

α,ω-Dodecanedioic acid ethylene ester [*54982-83-1*], (**44**, $n = 10$), $C_{14}H_{24}O_4$, M_r 256.34, *mp* 18 °C, *bp* (2.7 kPa) 139 – 141 °C, d_4^{60} 1.0303, n_D^{40} 1.4588, is a synthetic musk that is prepared by thermal depolymerization of the polyester obtained from 1,12-dodecanedioic acid and ethylene glycol in the presence of a catalyst (e.g., stannous salts of aliphatic monocarboxylic acids) [180].

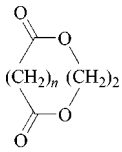

44

The compound is used in perfumery as a musk fragrance, but is not as long-lasting as the following homologous compound, ethylene brassylate.
Trade Name. Arova 16 (Degussa).

α,ω-Tridecanedioic acid ethylene ester [*105-95-3*], ethylene brassylate (**44** ,$n = 11$), $C_{15}H_{26}O_4$, M_r 270.37, *bp* (1.3 kPa) 140 °C, d_4^{60} 1.0180, n_D^{20} 1.4702, is an artificial fragrance compound, with a sweetish, slightly fatty, musk odor. Like α,ω-dodecanedioic acid ethylene ester, the ester is obtained by depolymerization of the corresponding polyester. Brassylic acid (α,ω-tridecanedioic acid) is prepared by ozonolysis of erucic acid [181]:

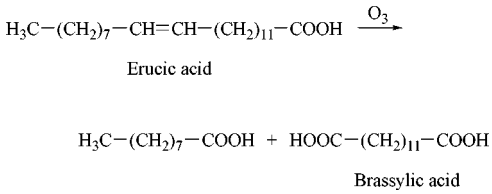

Ethylene brassylate is used in large amounts in perfumery as a fixative and for rounding off sweet-flowery odor notes.
Trade Name. Musk T (Takasago), Ethylene Brassylate (Degussa).

Coumarin [*91-64-5*], 2*H*-1-benzopyran-2-one, $C_9H_6O_2$, M_r 146.15, *bp* (1.33 kPa) 153.9 °C, *mp* 70.6 °C, d_4^{20} 0.935, occurs widely in nature and determines, for example, the odor of woodruff. It forms white crystals with a hay-like, spicy odor. When treated with dilute alkali, coumarin is hydrolyzed to the corresponding coumarinic acid salt (*cis*-2-hydroxycinnamic

acid). Heating with concentrated alkali or with sodium ethanolate in ethanol results in the formation of *o*-coumaric acid salts (*trans*-2-hydroxycinnamic acid). 3,4-Dihydrocoumarin is obtained by catalytic hydrogenation, for example, with Raney nickel as a catalyst; octahydrocoumarin is obtained if hydrogenation is carried out at high temperature (200 – 250 °C).

Production. Coumarin is currently produced by Perkin and Raschig syntheses.

Synthesis from Salicylaldehyde by the Perkin Reaction.
In the presence of sodium acetate, salicylaldehyde reacts with acetic anhydride to produce coumarin and acetic acid. The reaction is carried out in the liquid phase at elevated temperature.

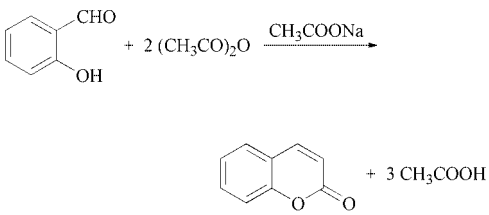

Coumarin

In a special process, the sodium acetate catalyst is retained in the reactor by a built-in filter and is reused [182].
Synthesis from o-Cresol by the Raschig Process.

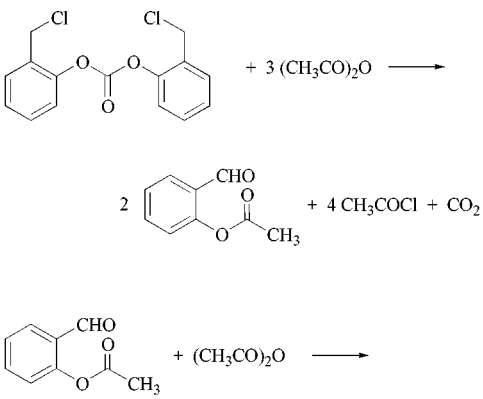

Coumarin

The hydroxyl group of *o*-cresol is protected by a carbonate or phosphate group, and the ester can be converted into the corresponding benzal chloride derivative by chlorination of the methyl group. Reaction of this intermediate with acetic anhydride yields *O*-acetylsalicylaldehyde and acetyl chloride. Ring closure of the aldehyde with acetic anhydride gives coumarin.
Since the odor of coumarin is relatively weak, strong-smelling byproducts (e.g., vinylphenol) must be removed. Many purification methods have been reported and patented.

Uses. Coumarin is one of the most widely used fragrance compounds. It is used in fine fragrances as well as in soap perfumes for spicy green notes. It is also used in galvanization as a brightener.

Dihydrocoumarin [*119-84-6*], 3,4-dihydro-2*H*-benzopyran-2-one, $C_9H_8O_2$, M_r 148.16, *bp* (2.3 kPa) 145 °C, *mp* 24 °C, n_D^{25} 1.5528, occurs in a few essential oils and forms colorless crystals with a sweet-herbal odor. Dihydrocoumarin is prepared by hydrogenation of coumarin, for example, in the presence of a Raney nickel catalyst. Another process employs the vapor-phase dehydrogenation of hexahydrocoumarin in the presence of Pd or Pt–Al_2O_3 catalysts [183]. Hexahydrocoumarin is prepared by cyanoethylation of cyclohexanone and hydrolysis of the nitrile group, followed by ring closure to the lactone [184].

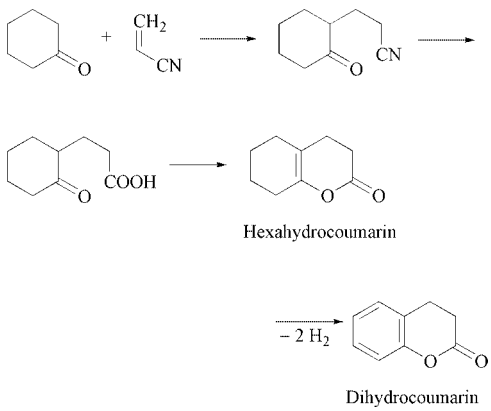

Dihydrocoumarin is used in woodruff-type flavor compositions. (Limiting IFRA recommendation)

2.7.3. Glycidates

A number of glycidates are important intermediates in the synthesis of fragrance compounds. They are prepared either by epoxidation of the corresponding acrylates or by condensation of aldehydes or ketones with α-chloro substituted fatty acid esters (Darzens reaction).

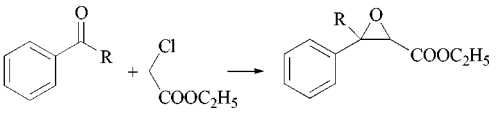

Ethyl 3-phenylglycidate [*121-39-1*], "so-called special aldehyde C_{16}", $C_{11}H_{12}O_3$, M_r 192.21, *bp* (40 Pa) 104 °C, d_4^{20} 1.1023, n_D^{30} 1.5095, is a colorless liquid with a strawberry-like odor; it is not known to occur in nature.

It is prepared by treating ethyl cinnamate with peracetic acid [185] or by condensation of benzaldehyde with ethyl chloroacetate (in the above Darzens reaction, R = H). The glycidate is used as a long-lasting fragrance compound for creating harmonic, fruity notes in household and fine fragrances.

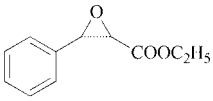

Ethyl 3-methyl-3-phenylglycidate [*77-83-8*], "so-called aldehyde C_{16}", $C_{12}H_{14}O_3$, M_r 206.24, *bp* (2.4 kPa) 153–155 °C, d_{25}^{25} 1.506–1.513, occurs in two optically active pairs of *cis*- and *trans*-isomers; each isomer has a characteristic odor [186]. The commercial product is a racemic mixture of all four isomers and has a strong, sweetish, strawberry odor. The *cis* : *trans* ratio obtained in the Darzens condensation of acetophenone ($R = CH_3$) and ethyl chloroacetate depends on the base used in the reaction.

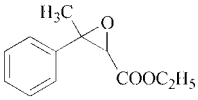

The glycidate is used in household perfumery for fruity notes.
Trade Name. Fraise pure (Givaudan Roure).

2.7.4. Miscellaneous Compounds

2-Furylmethanethiol [*98-02-2*], furfuryl mercaptan, C_5H_6OS, M_r 114.16, *bp* (101.3 kPa) 160 °C, d_4^{20} 1.1319, n_D^{20} 1.5329, is an important constituent of the aroma of roasted coffee. It is a liquid with an unpleasant odor, which becomes like coffee when diluted. Furfuryl mercaptan is prepared from furfuryl alcohol, thiourea, and hydrogen chloride. The resulting *S*-furfurylisothiouronium chloride is cleaved with sodium hydroxide to give furfuryl mercaptan.

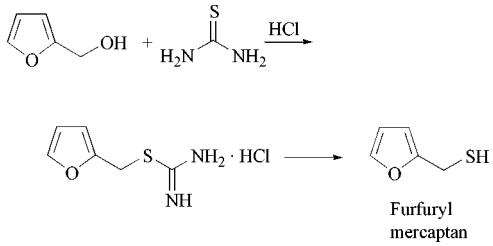

The thiol is used in coffee aromas.

2.8. N- and N, S-Heterocycles

Many nitrogen- and sulfur-containing heterocycles have been identified in the aroma fractions of foods [187]. In roasted products (e.g., coffee) and heat-treated foods (e.g., baked bread or fried meat), these heterocycles are formed from reducing sugars and simple or sulfur-containing amino acids by means of Maillard reactions [188, 189]. Their odor threshold values are often extremely low and even minute amounts may significantly contribute to the aroma quality of many products [190, 191]. Therefore, N- and N, S-heterocyclic fragrance and flavor substances are produced in far smaller quantities than most of the products previously described.

Pyrroles, indoles, pyridines, quinolines, and pyrazines are examples of N-heterocycles that are produced as fragrance and flavor compounds. Thiazoles and dithiazines are examples of nitrogen- and sulfur-containing heterocycles. These heterocyclic compounds are mainly used in aroma compositions, exceptions are indoles and quinolines, which are important fragrance substances.

Representatives of the above-mentioned classes are as follows:

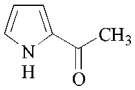

2-Acetylpyrrole [*1072-83-9*], roast odor

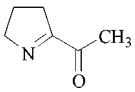

2-Acetyl-3,4-dihydro-5*H*-pyrrole [*85213-22-5*], characteristic odor of white bread crust

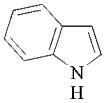

Indole [*120-72-9*], fecal odor, floral in high dissolution

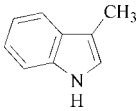

3-Methylindole, skatole [*83-34-1*], indole-like odor

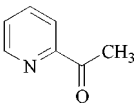

2-Acetylpyridine [*1122-62-9*], roast odor

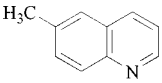

6-Methylquinoline [*91-62-3*], blossom odor, sweet-animalic upon dissolution

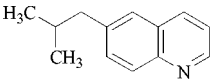

6-Isobutylquinoline [*68141-26-4*], mossy-earthy odor

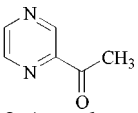

2-Acetylpyrazine [*22047-25-2*], popcorn-like odor

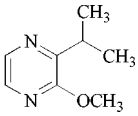

2-Methoxy-3-isopropylpyrazine [*25773-40-4*], green pea odor

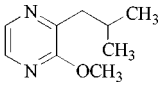

2-Methoxy-3-isobutylpyrazine [*24683-00-9*], green-pepper odor

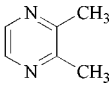

2,3-Dimethylpyrazine [*5910-89-4*], and its 2,5- [*123-32-0*] and 2,6- [*108-50-9*] isomers, roast odor, reminiscent of nuts

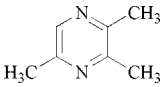

Trimethylpyrazine [*14667-55-1*], roast odor, reminiscent of coffee and cocoa

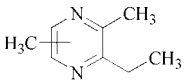

Mixture of 3-ethyl-2,5-dimethylpyrazine [*13360-65-1*] and 2-ethyl-3,5-dimethylpyrazine [*13925-07-0*], roast odor, reminiscent of nuts

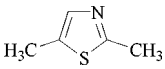

2,5-Dimethylthiazole [*4175-66-0*], meat-like odor

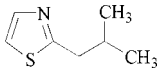

2-Isobutylthiazole [*18640-74-9*], tomato odor

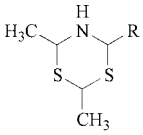

and

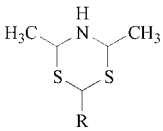

R = Methyl, isopropyl, isobutyl, 2-methylbutyl

Alkyldimethyl-1,3,5-dithiazines, roast odor

3. Natural Raw Materials in the Flavor and Fragrance Industry

3.1. Introduction

Although synthetic flavor and fragrance materials are produced on an industrial scale, naturally occurring raw materials continue to be essential, important ingredients in the manufacture of flavor and fragrance compositions for several reasons. First, the composition and organoleptic nature of natural products are often too complex to be reproduced by a combination of synthetic fragrance substances. Second, the characteristic flavor and fragrance substances of a particular product often cannot be synthesized at a competitive price. Third, the use of natural materials in the production of certain flavor compositions is compulsory. Demand for perfumes based on natural materials has increased.

Raw materials for the flavor and fragrance industry are obtained from more than 250 different plant species, but only a handful of products originate from animals.

Raw materials are isolated from various parts of plants, e.g., blossoms, buds, fruit, peel, seeds, leaves, bark, wood, roots, or from resinous exudates. Different parts of the same plant may yield products with different compositions. For instance, steam distillation of the bark of the cinnamon tree gives cinnamon bark oil, which contains mainly cinnamaldehyde, whereas cinnamon leaf oil obtained from the leaves of the tree contains eugenol as its major constituent.

The quality of natural products depends considerably on their geographic origin, even if they are isolated from the same plant species. This may be partly due to variations in cultivation conditions, such as soil structure and climate, but also results from the fact that different varieties of the same plant species are cultivated in different areas. Thus, more than 500 natural raw materials are available for the creation of perfumes and flavors.

The flavor and fragrance industry has expanded so much that the plants required to supply the raw materials are now grown on a very large scale. Examples are the peppermint and spearmint plantations in the United States and the lavandin plantations in southern France.

The economic importance of the cultivation of aromatic plants has led to the systematic breeding of new varieties. New species are obtained either by alternation of generations or by vegetative means in an attempt to improve yield, oil quality, and resistance to disease and insects.

The production of some essential oils has decreased to low levels or even been discontinued due to competition from synthetic products. Nevertheless, the worldwide production of flavor and fragrance materials of natural origin has increased recently due to breeding successes, but their total market share has decreased. Annual worldwide sales currently amount to $\$600 \times 10^6$ [192].

Raw materials derived from intensive agricultural cultivation are usually relatively inexpensive. However, the prices of some natural materials may exceed \$1000 per kilogram because cultivation and harvesting of these plants are tedious and product yields are very low. Examples of extremely valuable ingredients of fragrance and flavor creations include rose oil, jasmine absolute, tuberose absolute, orris root oil, ambrette seed oil, angelica root oil, and orange flower oil.

3.2. Isolation of Natural Fragrance and Flavor Concentrates

Three main methods are used to concentrate plant flavor and fragrance substances:

1) distillation
2) mechanical separation ("pressing")
3) solvent extraction

The qualitative and quantitative composition and, thus, the organoleptic properties of the product depend on the isolation procedure. For example, an extract contains large amounts of nonvolatile components that are not found in essential oils obtained by distillation. Since these components markedly influence odor development (complexing and fixing), the two products may have completely different organoleptic properties, even though the compositions of their odorous volatile constituents are comparable.

In addition, the distillation of essential oils at elevated temperature results in the transformation of thermolabile substances, and some typical components are only released from their precursors in the plants under distillation conditions.

Solvent extraction is generally applied in the separation of heat-labile plant materials or if an essential oil can be obtained only in very low yield (e.g., from blossoms). It is also used if the nonvolatile components are desired for their fixative properties (e.g., in the preparation of resinoids from exudates).

3.2.1. Essential Oils

Production. Essential oils are obtained from plant materials by distillation with water or steam. After condensation of the vapor phase, the oil separates from the aqueous phase and is removed. The yield of essential oil, based on the starting plant material, generally ranges from a few tenths of 1 % to a few percent. The apparatus used in the production of natural fragrance concentrates is described in [193].

Essential oils consist of volatile, lipophilic substances that are mainly hydrocarbons or monofunctional compounds derived from the metabolism of mono- and sesquiterpenes, phenylpropanoids, amino acids (lower mass aliphatic compounds), and fatty acids (long-chain aliphatic compounds). Unlike fatty oils, essential oils do not leave a grease stain when dabbed on filter paper.

Essential oils are to be distinguished from the so-called *distillates,* which are ethanol-containing products that are obtained from plant materials by distillation with ethanol or with ethanol–water mixtures. *Essence oils* are defined as essential oils that separate from the aqueous phase in the distillation receiver during the distillative concentration of fruit juices (usually citrus juices). *Citrus peel oils* are a special type of essential oil. They are isolated by pressing the peel to release the volatile substances stored in the pericarp in small oil glands. The resulting products are termed essential oils because they consist largely of highly volatile terpene hydrocarbons. However, they also contain small amounts of nonvolatile compounds, such as dyes, waxes, and furocoumarins.

Uses. Most essential oils are used directly as starting materials in the production of flavor and fragrance compositions. However, some essential oils are fractionated or concentrated by distillation, partitioning, or adsorption. Substances that are important for the desired characteristic odor and taste are thus concentrated, and other components, which possess either an unpleasant or very faint odor or are unsuitable for the application in question, are removed.

Individual compounds can be isolated from essential oils containing one or only a few major components by distillation or crystallization. Examples are eugenol from clove oil, menthol from cornmint oil, citronellal from *Eucalyptus citriodora* oil and citral from *Litsea cubeba* oil. These compounds are used as such or serve as starting materials for the synthesis of derivatives, which are also used as flavor and fragrance substances. However, the importance of some of these oils has decreased substantially because of the development of selective synthetic processes for their components.

Although essential oils or their fractions are mixtures of many substances, these oils are occasionally converted as a whole into derivatives. Examples of such derivatives are vetiveryl acetate from vetiver oil, guaiyl acetate from guaiac wood oil, and acetyl cedrene from cedarwood terpenes. These products are also employed as fragrance substances.

3.2.2. Extracts

Extracts of fragrance and flavor substances obtained from plants are termed pomades, concretes, absolutes, resinoids, or tinctures according to their method of preparation.

Pomades consist of fats that contain fragrance substances and are produced by the hot or cold enfleurage of flowers. *Hot enfleurage* is the oldest known procedure for preserving plant fragrance compounds. In this method, flowers (or other parts of a plant) are directly immersed in liquid or molten wax.

In *cold enfleurage,* the volatile components released by flowers into their sourroundings are absorbed with fats over a longer period of time. This industrial procedure was developed in southern France in the 19th century for the production of high-grade flower concentrates. It involves the application of fresh flowers to a fat layer, consisting of a mixture of specially refined lard and beef tallow, which is spread on a glass

plate in a closed container. This method, however, has been almost totally replaced by the less tedious technique of solvent extraction.

Concretes are prepared by extracting fresh plant material with nonpolar solvents (e.g., benzene, toluene, hexane, petroleum ether). On evaporation, the resulting residue contains not only volatile fragrance materials, but also a large proportion of nonvolatile substances including waxy compounds. For this reason, concretes (like pomades) are not completely soluble in alcohol and, thus, find limited use as perfume ingredients. However, they can be employed in the scenting of soaps.

Concretes, which are actually intermediate products (see below), are prepared mainly from flowers (rose, jasmine, tuberose, jonquil, ylang-ylang, mimosa, boronia, etc.), but also from other plant materials (lavender, lavandin, geranium, clary sage, violet leaves, oak moss, etc.). A yield of ca. 0.3 % based on the starting flower material, is obtained in the production of jasmine concrete.

Absolutes are prepared by taking up concretes in ethanol. Compounds that precipitate on cooling are then removed by filtration. After evaporation of the ethanol, a wax-free residue called an absolute is left behind. Absolutes are completely soluble in ethanol and, therefore, can be freely used as perfume ingredients. They are usually formed in a yield of ca. 50 %, based on the concrete as starting material.

In rare cases, absolutes can be obtained directly by extracting the plant material with alcohol (e.g., tonka absolute).

Resinoids are prepared by extracting plant exudates (balsams, oleo gum resins, natural oleo resins, and resinous products) with solvents such as methanol, ethanol, or toluene. Yields range from 50 to 95 %. The products are usually highly viscous and are sometimes diluted (e.g., with phthalates or benzyl benzoate) to improve their flow and processing properties.

Resinoids mainly consist of nonvolatile, resinous compounds and are primarily used for their excellent fixative properties.

The resinoids described above should be distinguished from prepared *oleoresins* (e.g., pepper, ginger, and vanilla oleoresins), which are concentrates prepared from spices by solvent extraction. The solvent that is used depends on the spice; currently, these products are often obtained by extraction with supercritical carbon dioxide. Pepper and ginger oleoresins contain not only volatile aroma compounds, but also substances responsible for pungency.

Tinctures are alcoholic solutions that are prepared by treating natural raw materials with ethanol or ethanol–water mixtures. They can also be obtained by dissolving other extracts in these solvents. Tinctures are sometimes called *infusions.*

3.3. Survey of Natural Raw Materials

The following survey of the most important, well-known raw materials used in the flavor and fragrance industry is by no means complete; the materials are listed in alphabetical order. Physical standards for essential oils are described as specified by the ISO or the Essential Oil Association of the United States (EOA). The physical standards cited according to ISO are taken from the most recent draft, because many standards are already very old but under discussion. The ISO standard number remains the same; so the reader can also check for the older data. EOA specifications were only cited in those cases where no ISO specifications are available. These become more and more the most widely distributed standards. Gas chromatograms are widely used for analysis and quality control, but have not been included due to lack of space. Further details are given in the literature, e.g., [194], and in ISO specifications, which now include gas chromatograms.

The entire components of the individual products have not been listed for the same reason; further information is available in the literature, e.g., [187, 195]. Physical data for extracts or concentrates consisting largely of nonvolatile material are not given because the composition of these products varies widely according to the isolation and manufacturing procedure used.

The botanical names of plants are cited in accordance with the International Code of Botanical Nomenclature (ICBN) as described, for example, in [196].

Allium oils are obtained from garlic and onion (Liliaceae). Their quality is assessed on the basis of their odor and aroma rather than their physical and chemical properties. The EOA specifications given below are, therefore, of limited value only.

1) *Garlic oil* is obtained by steam distillation of crushed bulbs of the common garlic, *Allium sativum* L.; it is a clear, reddish-orange liquid, with a strong, pungent, characteristic garlic odor.
   d_{25}^{25} 1.040 – 1.090, n_D^{20} 1.5590 – 1.5790 [197]. Diallyl disulfide [*2179-57-9*] is an essential odor component of garlic oil:
   $CH_2{=}CHCH_2SSCH_2CH{=}CH_2$
2) *Onion oil* is obtained by steam distillation of the crushed bulbs of the common onion, *Allium cepa* L. It is an amber-yellow to amber liquid with a strongly pungent, lasting, characteristic onion odor.
   d_{25}^{25} 1.050 – 1.135; n_D^{20} 1.5495 – 1.5695 [198].
   Aliphatic sulfur compounds, in particular disulfides such as methyl propyl disulfide, dipropyl disulfide, and especially *cis*- and *trans*-propenyl propyl disulfide are mainly responsible for the typical odor of onion oil. Garlic and onion oil are used in seasoning mixtures for the food industry.

Allspice oil, see Pimento Oils, page 113.

Ambergris (ambra), see Animal Secretions, page 87.

Ambrette seed oil is obtained by steam distillation of the dried, crushed seeds of *Abelmoschus moschatus* Medik. (*Hibiscus abelmoschus* L., Malvaceae), a flowering shrub growing in tropical areas. Due to its content of long-chain fatty acids, the crude product is a waxy mass and, therefore, also called "Ambrette beurre." Removal of the fatty acids with alkali gives a clear yellow to amber-yellow liquid with the strong, musky odor of ambrettolide.

d_{25}^{25} 0.898 – 0.920; n_D^{20} 1.4680 – 1.4850; $[\alpha]_D$ $-2°30'$ to $+3'$; acid value: max. 3; saponification value: 140 – 200 [199].

The constituents responsible for the musk odor of the oil are ambrettolide, (*Z*)-7-hexadecen-16-olide [*123-69-3*] and (*Z*)-5-tetradecen-14-olide [*63958-52-1*]:

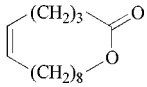
5-Tetradecen-14-olide

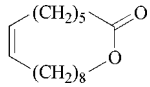
Ambrettolide

Other components are acyclic aliphatic esters and terpenes, such as farnesol and farnesyl acetate. Ambrette seed oil is one of the most expensive essential oils and, thus, is used mainly in fine fragrances.

Amyris oil is obtained by steam distillation of the wood from the tree *Amyris balsamifera* L. (Rutaceae), which grows in the Caribbean area and around the Gulf of Mexico. Annual production worldwide is ca. 40 t. It is a pale yellow to amber-yellow, slightly viscous liquid with a mild wood odor.

d_{20}^{20} 0.946 – 0.978; n_D^{20} 1.505 – 1.510; $[\alpha]_D^{20}$ + 10 ° to + 60 °; solubility: 1 vol in 1 vol of 90 % ethanol at 20 °C; solutions sometimes become opalescent on dilution; acid value: max. 3.0; ester value (after acetylation): 180 – 198 [200].

The oil is sometimes incorrectly called West Indian sandalwood oil. However, its composition and odor are different from those of the oils obtained from sandalwood species. The major components of amyris oil are sesquiterpenoids such as valerianol [*20489-45-6*], elemol [*639-99-6*], β-eudesmol [*473-15-4*], and *epi*-γ-eudesmol [*15051-81-7*] [201].

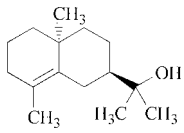
epi-γ-Eudesmol

Amyris oil is used in perfume compositions, mainly as a fixative.

Angelica oil is prepared from *Angelica* roots or seeds.

1) *Angelica root oil* is obtained by steam distillation of the dried roots of *Angelica archangelica* L. Apiaceae [*Archangelica officinalis* (Moench) Hoffm.], a plant occurring predominantly in Europe. The oil is a pale yellow to deep amber-yellow liquid with a green, herbaceous, peppery, musk-like odor and a bitter-sweet taste.

d_{25}^{25} 0.850 – 0.880, occasionally up to 0.930 for oils from stored roots; n_D^{20} 1.4735 – 1.4870; $[\alpha]_D$ 0° to +46°; acid value: max. 7; ester value: 10 – 65; solubility: 1 vol in at least 1 vol of 90 % ethanol, often with turbidity [202]. (Limiting IFRA recommendation)

2) *Angelica seed oil* is similarly obtained from fresh seeds of the plant. It is a light yellow liquid with an odor that is sweeter and more delicate than that of the root oil.
d_{25}^{25} 0.853 – 0.876; n_D^{20} 1.4800 – 1.4880; $[\alpha]_D$ +4° to +16°; acid value: max. 3; ester value: 14 – 32; solubility: 1 vol in at least 4 vol 90 % ethanol, often with considerable turbidity [203].
Angelica root oil contains ca. 90 % terpenoids and sesquiterpenoids. α-Pinene, 3-carene, limonene, and β-phellandrene are the major components (together ca. 60 %). In addition, the oil contains a large number of oxygen-containing compounds, of which the macrolides 15-pentadecanolide and 13-tridecanolide are essential odor components. The two angelica oils are used mainly in the alcoholic beverage industry.

Animal secretions are of minor commercial importance because some of the animal species from which they are obtained are virtually extinct, and the killing quota for others has been sharply reduced. Few odoriferous secretions of mammals have actually been shown to possess pheromone-like properties linked with reproduction. However, the odor of animal secretions is known to be important for communication and behavior of a particular species. Most of the products described below contain strong-smelling compounds with relatively high molecular masses. Therefore, they are used as longlasting fragrance complexes. Most of their odoriferous constituents are now produced synthetically and are used for the same purposes.

1) *Ambergris* (*ambra*) is a secretion of the sperm whale *Physeter catodon* (*P. macrocephalus* L.), that possibly results from a pathological condition. Ambergris has a lower density than water and washes ashore along the ocean coasts. The major quantity is obtained from killed animals, but only a low percentage contain ambergris in their intestines.

Fresh ambergris is almost black, but it turns light gray and develops a pleasant odor when exposed to light and seawater over a period of time. The major components of ambergris are epicoprosterol (up to 85 %) and the odorless triterpene alcohol, ambrein [*473-03-0*].

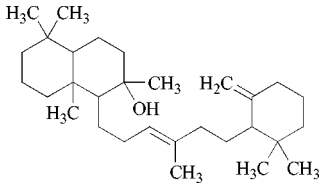

Ambrein

Ambrein is the likely precursor of a number of strongly odoriferous mono-, bi-, and tricyclic compounds that are formed by autoxidation or photooxidation [204]. Examples are as follows:

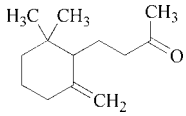

γ-Dihydroionone [*13720-12-2*] (odor: tobacco)

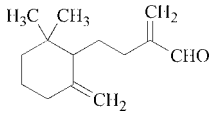

2-Methylene-4-(2,2-dimethyl-6-methylenecyclohexyl)butanal [*72892-63-8*] (odor: seawater)

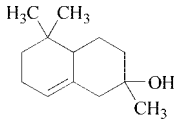

α-Ambrinol [*41199-13-3*] (odor: moldy, animal, fecal)

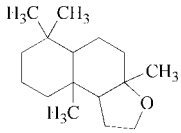

3a,6,6,9a-Tetramethyldodecahydronaphtho[2,1-b]furan [*6790-58-5*] (odor: ambergris)

Together, these compounds largely represent the odor of ambergris. Ambergris is applied as a 3 % tincture in 90 % ethanol, which is matured by standing over a period of time

with occasional shaking. It is used in fine fragrances.

2) *Beeswax absolute* is obtained by alcohol extraction of beeswax. The yield is generally less than 1 %. The yellowish-brown viscous product has a mild, honey-like odor and high tenacity; it is used almost exclusively in fine fragrances.

3) *Castoreum* is an unpleasantly sharp-smelling, oily substance secreted by special glands of beavers, *Castor fiber* L. (Castoridae), living in Canada, Alaska, and Siberia. Both sexes secrete the substance, which accumulates in an abdominal pouch, also called castoreum. Dilute castoreum (e.g., as a tincture in ethanol) smells pleasantly of birch tar and musk and is slightly fruity.
   Castoreum is a byproduct of the fur industry. The beaver pouches are dried in the air or over a wood fire; the color of their contents then changes from yellow to dark brown, and the consistency from a butter-like to resinous character.
   In addition to alcoholic tinctures, castoreum is available in the form of resinoids, which are prepared by extracting dried, comminuted pouches with acetone, benzene, or ethanol; yields are 30 – 50 %, 20 – 25 %, and 45 – 80 %, respectively. The resinoids have the following specifications [205]

Acid value	15 – 30	30 – 35	15 – 30
Ester value	145 – 200	145 – 200	145 – 200
Carbonyl value	55 – 70	40 – 60	10 – 20

   The intense, for the Siberian beaver leathery, odor of castoreum is caused largely by phenolic compounds (e.g., 4-alkylphenols and catechol derivatives), which beavers take in with their food and excrete into their abdominal pouch. Castoreum is used mainly in fine fragrances for its characteristic, long-lasting odor, particularly for delicate leather nuances.

4) *Civet* is a glandular secretion produced by both sexes of the civet cat (Viverridae). Two species are known: *Civetticitis civetta* that inhabits Ethiopia, and *Viverra zibetha* that is found in India, Malaya, Indochina, and Indonesia.
   The animals are kept in cages, and the fresh secretion is taken from the pouches at regular (about one-week) intervals. Civet is almost liquid with a light yellow color. It darkens when exposed to light and takes on a consistency like salve. In dilutions (e.g., as an alcoholic tincture), civet has a pleasant, sweetish odor. A resinoid prepared by extraction with acetone is a dark brown-red mass; *mp* 20 – 25 °C; acid value: 100 – 140; ester value: 20 – 40; carbonyl value: 15 – 25 [206]. Civetone (for formula, see Section 2.4) is the main odoriferous constituent of civet (2.5 – 3.4 %). Civet contains other macrocyclic ketones such as cyclohexa- and cycloheptadecanone and 6-*cis*-cycloheptadecenone. Traces of indole and skatole contribute to the animal note.
   Civet has a distinctly different odor from musk and was formerly a versatile ingredient of fine fragrances.

5) *Musk* is secreted exclusively by the male animals of *Moschus moschiferus,* a wild deer living in the mountains of Nepal, Tibet, and Mongolia. The light yellow secretion with a salve consistency accumulates in an abdominal pouch and probably serves to attract females. When the pouch is dried, the secretion solidifies to form a brittle, brown mass with a characteristic odor. Since several *Moschus* species occur, large variations exist in the quality and specifications of musk. Hunting of the animals has been prohibited; therefore, only small quantities of musk are available at extremely high prices.
   (−)-Muscone and related macrocyclic ketones are responsible for the odor of musk [207]. Like other animal secretions, musk is preferentially used as an alcoholic tincture in fine fragrances.

Anise oil is obtained by steam distillation of the fruits of *Pimpinella anisum* L. (Apiaceae). It is a colorless to pale yellow liquid or crystalline mass with a powerful, sweet odor, characteristic of anethole.

d_{20}^{20} 0.980 – 0.990; n_D^{20} 1.552 – 1.561; $[\alpha]_D^{20}$ − 2 ° to + 2 °; solubility: 1 vol in 3 vol of 90 % ethanol at 20 °C; *fp* 15 – 19.5 °C [208].

The main component of anise oil is *trans*-anethole, which may be present at a concentration up to 94 % and which determines the melting point of the oil.

The oil was formerly produced in many countries, mainly in eastern Europe, but has now been

replaced to a large extent by the less expensive star anise oil (see page 117), which also contains a high percentage of anethole. Annual worldwide production of both oils is ca. 500 t.

Anise oil is used for flavoring foods, beverages, and oral hygiene products.

Armoise oil is obtained by steam distillation of the herb *Artemisia herba-alba* Asso (Asteraceae) that grows in Morocco. It is a light yellow to yellow liquid with a light herbaceous odor characteristic of thujone.

d_{25}^{25} 0.917 – 0.935; n_D^{20} 1.4600 – 1.4720; $[\alpha]_D^{20}$ – 25 ° to – 8 °.

The major components of armoise oil (Marrakesh-type) are the ketones camphor (40 %) and α- and β-thujone (35 % and 5 % respectively). Since *Artemisia herba-alba* exists as various chemotypes, the composition of the oil may vary widely. For example, each of the above-mentioned ketones may occur in a concentration over 50 % or less than 10 %.

Armoise oil is used in fairly large amounts in fine fragrances (e.g., for chypre notes).

Basil oil is available in several types that differ in their major components; the two most important are described.

1) *Réunion basil oil* is obtained by steam distillation of the flowering tops or whole plants of *Ocimum basilicum* L. (Lamiaceae). This oil is produced mainly in Réunion, the Comores, Madagascar, and the Seychelles. It is a pale yellow to amber-yellow liquid with a fresh, green, spicy odor characteristic of estragole. d_{20}^{20} 0.948 – 0.970; n_D^{20} 1.5100 – 1.5200; $[\alpha]_D^{20}$ – 1 ° to + 2 °; solubility: 1 vol in 7 vol of 80 % ethanol [209].
   The oil typically has a high content of methylchavicol (min. 75 %, max. 85 %).

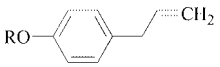

R = H_3C, methylchavicol (estragole) [*140-67-0*]
R = H, chavicol [*501-92-8*]

   Annual production (ca. 200 to 300 t) is used predominantly for seasoning foods.
2) *European basil oil* is produced mainly in France, Italy, the Balkan countries, Egypt, and South Africa. The major components of this oil are linalool (35 – 50 %) and methylchavicol (15 – 25 %). It is used for food flavoring and in perfumery.

Bay oil is obtained by steam distillation of the leaves of *Pimenta racemosa* (Mill.) J. W. Moore (Myrtaceae). It is a dark brown liquid with a strong, spicy, clovelike odor.

d_{20}^{20} 0.943 – 0.984; n_D^{20} 1.505 – 1.517; $[\alpha]_D^{20}$ close to 0 °; phenols content: min. 50 % [210].

Evergreen bay trees or bay-rum trees, which are up to 12 m high, grow wild and are also cultivated in northern South America and in the West Indies. The main cultivation area is the island of Dominica.

The major components of the oil are myrcene, eugenol, and chavicol. The phenol content is determined largely by the latter two compounds.

Total production of bay oil is ca. 50 t/a. It has antiseptic properties because of its high phenol content and is, therefore, used for perfuming after-shave lotions.

Benzoe Siam resinoid is obtained by solvent extraction of the resin from *Styrax tonkinensis* (Pierre) Craib ex Hartwich trees (Styracaceae). The wild growing *Styrax* tree is widespread in Thailand, Laos, Cambodia, and Vietnam.

Benzoe Siam resinoid is a reddish to light brown, viscous liquid with a long-lasting, chocolate-like, sweet, balsamic odor. It is used in perfumery for balsamic nuances and as a fixative. (IFRA recommendation)

Benzoe Sumatra resinoid is obtained by solvent extraction of the resin from *Styrax benzoin* Dryand, a tree growing predominantly on the island of Sumatra. Benzoe Sumatra resinoid is a dark brown viscous liquid with a warm, powdery, sweet-balsamic odor. Its main volatile, odor-determining components are derivatives of benzoic and cinnamic acids and vanillin [211].

Benzoe Sumatra resinoid is used in perfumery, mainly as a fixative with a warm, balsamic note. (IFRA recommendation)

Bergamot oil see Citrus Oils, page 94.

Bitter almond oil (free from hydrogen cyanide) contains benzaldehyde as its main component. Benzaldehyde does not occur as such in the plant, but is formed, together with hydrogen cyanide, by the hydrolytic cleavage of the

glycoside amygdalin. Amygdalin is present in bitter almonds, the seeds of *Prunus amygdalus* Batsch var. *amara* (DC.) Focke, and ripe apricot kernels, *Prunus armeniaca* L. (Rosaceae). The press cake, which remains after removal of the fatty oils, is macerated with water and left to stand for several hours, after which the "essential oil" is separated by steam distillation. The crude oil contains 2–4 % hydrogen cyanide, which is removed by washing with alkaline solutions of iron(II) salts. Subsequent redistillation yields an oil free from hydrogen cyanide. It is a colorless to slightly yellow liquid with an intense, almond-like, cherry aroma and a slightly astringent, mild taste.

d_{25}^{25} 1.040–1.050; n_D^{20} 1.5410–1.5460; $[\alpha]_D$ 0° to +0°25′; acid value: max. 8; aldehyde content (calculated as benzaldehyde): min. 95 %; solubility: 1 vol in 2 vol of 70 % ethanol [212].

Bitter almond oil is used almost exclusively in natural aroma compositions.

Black currant absolute (Bourgeons de cassis absolute) is obtained by solvent extraction via the concrete obtained from the dormant buds of the black currant bush *Ribes nigrum* L. (Saxifragaceae). The yield is ca. 3 %. Black currant absolute is a dark green paste with the characteristic, powerful, penetrating odor of black currants.

It has been used as a major ingredient in some fine fragrances.

Bois de rose oil, see Rosewood Oil, page 115.

Buchu leaf oils are obtained by steam distillation of *Agathosma betulina* (Bergius) Pillans (syn. *Barosma betulina*) and *A. crenulata (L.)* Pillans (syn. *Barosma crenulata*) (Rutaceae) leaves. The oils are dark yellow to brown liquids with a characteristic minty-fruity odor, reminiscent of black currant.

d_{25}^{25} 0.912–0.956, n_D^{20} 1.474–1.488; $[\alpha]_D$ −36° to −8°; acid value: 1–5; ester value: 20–85.

The bushes grow wild and are cultivated in South Africa. The major components of the oils are (+)-limonene (ca. 10 %) and other cyclic terpenoids that are structurally related to menthone. However, the constituents responsible for the characteristic black currant odor are *p*-menthane-8-thiol-3-one [*38462-22-5*] and its S-acetate derivative [*57074-34-7*] which are two of the small number of naturally occurring sulfur-containing terpenoids known to date [213, 214].

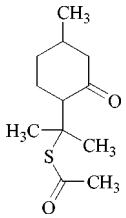

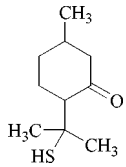

Buchu oil is used as a flavor ingredient (e.g., in fruit aromas) and in perfumery in chypre bases and in certain types of eaux de cologne; only very small amounts are employed because of its intensity.

Calamus oil is obtained by steam distillation of fresh or unpeeled, dried roots of *Acorus calamus* L. (Araceae). It is a yellow to medium brown, moderately viscous liquid with a pleasant, spicy, aromatic odor. (Limiting IFRA recommendation)

The plant occurs in polyploid varieties and the corresponding essential oils differ predominantly in their content of β-asarone (*cis*-isoasarone) [*5273-86-9*]:

diploid (American)	0 %
triploid (European)	0–10 %
tetraploid (Indian)	up to 96 %

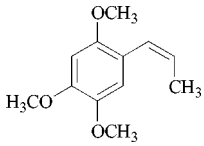

β-Asarone

The following data are typical for European and Indian oils, respectively:

d_{25}^{25} 0.940–0.980 and 1.060–1.080; n_D^{20} 1.5010–1.5160 and 1.5500–1.5525; $[\alpha]_D$ +5° to +35° and −2° to +6.5°; acid value: max. 4; ester value: 3–20; solubility: 1 vol in 5 vol of 90 % ethanol; solutions may be turbid [215].

Calamus oil is used in perfumery for spicy-herbaceous notes; small quantities are also employed in the alcoholic beverage industry. However, use is legally restricted because of the potential toxicity of β-asarone.

Camphor oil is obtained by steam distillation of the wood of the camphor tree *Cinnamomum camphora* Sieb. (Lauraceae) growing in China, Taiwan, and Japan. The main constituent of the crude oil is camphor (ca. 50 %), which can be separated by cooling and centrifugation. Fractionation of the mother liquor gives two oils:

1) *White camphor oil* is the first distillation fraction (ca. 20 % of the crude camphor oil). It is a colorless or almost colorless liquid with a cineole-like odor.
   d_{25}^{25} 0.855 – 0.875; n_D^{20} 1.4670 – 1.4720; $[\alpha]_D$ +16° to +28°; solubility: 1 vol in 1 vol of 95 % ethanol; solutions usually become cloudy on further dilution [216].
   In addition to monoterpene hydrocarbons, this oil contains up to 35 % cineole.
   Camphor oils with a higher cineole content can be obtained by further fractionation or by steam distillation of the leaves of other varieties of the camphor tree. They are marketed under the trade name eucalyptus oil 70 – 80 % by Chinese producers, but are not true eucalyptus oils (see Eucalyptus Globulus Oil, page 100).
2) *Brown camphor oil* is a fraction with a boiling point higher than that of camphor (ca. 20 %). It is a pale yellow to brown liquid with the characteristic odor of sassafras oil.
   d_{25}^{25} 1.064 – 1.075; n_D^{20} 1.5100 – 1.5500; $[\alpha]_D$ 0° to +3°; *fp* 6 °C; solubility: 1 vol in 2 vol of 90 % ethanol [217].

The oil contains more than 80 % safrole and, like Brasilian sassafras oil, is therefore used as a raw material for the production of piperonal via isosafrole (see page 69). (IFRA recommendation)

Camphor oils with a high safrole content can also be obtained by steam distillation of the wood of *Cinnamomum parthenoxylon* Nees.

The production of natural camphor and camphor oils was formerly several thousand of tons per year. Nowadays it has declined to 500 t/a as a result of the production of synthetic camphor. The same is true for the distillation of linalool-containing camphor oils (Ho oil, Ho leaf oil), which are derived from other varieties of the camphor tree.

Cananga oil, see Ylang-Ylang and Cananga Oils, page 120.

Caraway oil is obtained by steam distillation of crushed caraway seeds from *Carvum carvi* L. (Apiaceae). It is an almost colorless to amber liquid with a characteristic caraway odor and a mild-spicy taste.

d_{20}^{20} 0.905 – 0.920; n_D^{20} 1.4840 – 1.4890; $[\alpha]_D^{20}$ +67° to +80°; ketone content (calculated as carvone): 50 – 60 %; solubility: 1 vol in 5 vol of 80 % ethanol; acid value: max. 1 [218].

The major constituents of common caraway oil are (+)-limonene and (+)-carvone, which together may make up more than 95 % of the oil. (+)-Carvone is the essential odor component.

Caraway oil is produced mainly in Poland and Hungary. It is used primarily for flavoring foods and alcoholic beverages, but also for the production of (+)-carvone.

Cardamom oil is obtained by steam distillation of the seeds of *Elettaria cardamomum* (L.) Maton var. minuscula Burkill (Zingiberaceae). It is a colorless or very pale yellow liquid with an aromatic, penetrating, slightly camphoraceous odor and a persistent, pungent, strongly aromatic taste.

d_{20}^{20} 0.919 – 0.936; n_D^{20} 1.4620 – 1.4680; $[\alpha]_D^{20}$ +22° to +41°; solubility: 1 vol in 2 to 5 vol of 70 % ethanol, solutions sometimes slightly opalescent [219].

The major components of cardamom oil are 1,8-cineole and α-terpinyl acetate (ca. 30 % each). It is produced from cultivated or wild plants in the mountainous regions of southern India, Sri Lanka, Indonesia, Thailand, southern China, and Guatemala. Cardamom oil is used primarily for seasoning foods and occasionally in perfumery.

Carrot seed oil is obtained by steam distillation of the crushed seeds of *Daucus carota* L. (Apiaceae). It is a light yellow to amber-yellow liquid with a pleasant, aromatic odor.

d_{20}^{20} 0.900 – 0.945; n_D^{20} 1.4800 – 1.4930; $[\alpha]_D^{20}$ – 30° to – 4°; acid value: min. 5.0; ester value: 9 – 58; ester value (after acetylation): 50 – 95; solubility: 1 vol in 2 vol of 80 % ethanol [220].

The main constituent of carrot seed oil is carotol [*465-28-1*], which may be present in over 50 % concentration and is responsible for the characteristic odor.

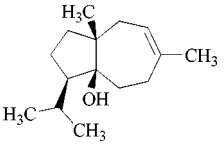

Carotol

Carrot seed oil is used in the alcoholic beverage industry, in food flavors, and in perfume compositions.

Castoreum, see Animal Secretions, page 87.

Cedar Oils. Several different conifer species are called cedars and the corresponding oils vary considerably in composition. Worldwide annual production is about 800 t. The following cedar oils are commercially important:

1) *Atlas cedarwood oil* is obtained by steam distillation of wood from *Cedrus atlantica* (Endl.) Manetti ex Carr. (Pinaceae). It is a yellowish to brownish, slightly viscous liquid with a sweet aromatic odor and a lasting, woody undertone.
d_{25}^{25} 0.925 – 0.940; n_D^{20} 1.5060 – 1.5160; $[\alpha]_D$ + 50 ° to + 77 °; ketone content (calculated as atlantone): min. 20 %; solubility: 1 vol in at least 5 vol of 95 % ethanol [221].
Atlas cedarwood oil and Himalaya cedarwood oil [from *Cedrus deodara* (D. Don) G. Don] both contain α-atlantone [*32207-08-2*] as their main odoriferous component. They are used mainly in soap perfumes.

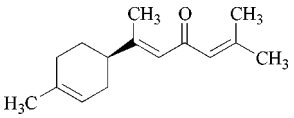

E-(+)-α-Atlantone

2) *Cedar leaf oil* (Thuja oil) is produced by steam distillation of fresh leaves and branch ends of the tree *Thuja occidentalis* L. (Cupressaceae). It is a colorless to yellow liquid with a powerful, herbaceous odor, characteristic of thujone [*546-80-5*].
d_{25}^{25} 0.910 – 0.920; n_D^{20} 1.4560 – 1.4590; $[\alpha]_D$ – 14 ° to – 10 °; ketone content (calculated as thujone): min. 60 %; solubility: 1 vol in 3 vol of 70 % ethanol at 25 °C [222].

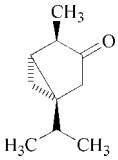

(–)-Thujone

The oil is produced in the northern states of the United States and in Canada. It is used in perfumery for dry nuances in citrus and woody compositions.

3) *Chinese cedarwood oil* is similar in composition to Texas cedarwood oil (see below). Chinese cedarwood oil is obtained by steam distillation of the wood of *Chamaecyparis funebris* Endl. = *Cupressus funebris* Endl. (Cupressaceae), which is a weeping cypress, indigenous to China [223]. Annual production lies between 850 and 900 t. Chinese cedarwood oil is an almost colorless to light yellow oil with a woody, bitter odor. d_{20}^{20} 0.938 – 0.960; n_D^{20} 1.5000 – 1.5080; $[\alpha]_D^{20}$ – 35 ° to – 20 °; cedrol content: min. 10 %, max. 16 % (by GLC); solubility: 1 vol in 5 vol of 95 % ethanol at 20 °C [224].
4) *Texas cedarwood oil* is produced by steam distillation of chopped wood of the Texas cedar, *Juniperus mexicana* Schiede (Cupressaceae). It is a brown to reddish-brown, viscous liquid that may partially solidify at room temperature. It has a characteristic cedarwood odor.
d_{20}^{20} 0.950 – 0.966; n_D^{20} 1.5020 – 1.5100; $[\alpha]_D^{20}$ – 52 ° to – 30 °; total alcohol content (calculated as cedrol): 35 – 48 %; cedrol content: min 20 % (by GLC), solubility: 1 vol in 5 vol of 90 % ethanol at 20 °C [225].
For uses, see Virginia cedarwood oil.
5) *Virginia cedarwood oil* is produced by steam distillation of sawdust, finely chipped waste wood from the manufacture of cedarwood products, or from stumps and logs of the red cedar; *Juniperus virginiana* L. (Cupressaceae). It is an almost colorless to pale yellow, viscous liquid with a characteristic cedarwood odor. The oil sometimes solidifies at room temperature.
d_{20}^{20} 0.941 – 0.965; n_D^{20} 1.5020 – 1.5100; $[\alpha]_D^{20}$ – 36 ° to – 16 °; solubility: 1 vol in 5 vol of 95 % ethanol; sum of cedrol and widdrol: min. 15 %, max. 30 % (by GLC) [226, 227].

While the Texas cedar grows in Mexico and other Central American countries, the Virginia cedar grows exclusively in the Southeast of the United States. Both cedar oils (production around 1600 t/a) are used mainly for perfuming soaps and other products, as well as a starting material for the isolation of cedrol [*77-53-2*]. Further fragrance substances such as cedryl acetate (see p. page 41) and cedryl methyl ether (see p. page 34) are produced from cedrol. The oils contain other sesquiterpenes, particularly α-cedrene and thujopsene, which can be converted to valuable fragrance materials. By acetylation of the mixture cedryl methyl ketone (see page 39) is produced; for production of ambrocenide pure α-cedrene is used (see page 76).

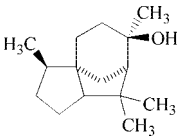

(+)-Cedrol

Celery seed oil is obtained by steam distillation of the crushed, ripe seeds of celery, *Apium graveolens* L. (Apiaceae). It is an almost colorless to brownish-yellow liquid with a characteristic, pervasive, spicy, very persistent odor.

d_{20}^{20} 0.867 – 0.908; n_D^{20} 1.4780 – 1.4880; $[\alpha]_D^{20}$ +65° to +78°; solubility: 1 vol in no more than 6 vol of 90 % ethanol at 20 °C; acid value: max. 6; ester value: 20 – 70 [228].

Major mono- and sesquiterpene hydrocarbons in the oil are (+)-limonene (58 to 79 % according to ISO) and β-selinene [*17066-67-0*]. Its typical, long-lasting odor is caused primarily by two lactones, 3-butylphthalide [*6066-49-5*] and sedanenolide [*62006-39-7*] which each occur at a concentration of ca. 1 %.

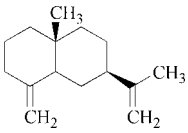

(+)-β-Selinene

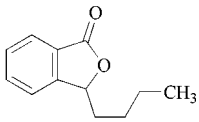

3-Butylphthalide

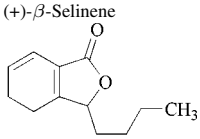

Sedanenolide (3-butyl-4,5-dihydrophthalide)

Celery for the production of seed oil is grown mainly in southern France, India, and California. The seed oil is used chiefly for flavoring foods, although small quantities are also used in perfumery.

Chamomile oils are available in three types:

1) *Blue chamomile oil* (German chamomile oil) is obtained by steam distillation of the flowers and stalks of *Chamomilla recutita* (L.) Rauschert (*Matricaria recutita.* L., *M. chamomilla* L. p.p.-Asteraceae), which is grown mainly in central and eastern Europe, Egypt, and Argentina. Annual production is ca. 10 t. Blue chamomile oil is a deep blue or bluish-green liquid that turns green and, finally, brown when exposed to light and air. The oil has a strong, characteristic odor and a bitter-aromatic taste. Chamazulene [*529-05-5*] is responsible for its blue color. Chamazulene and (−)-α-bisabolol [*23089-26-1*] contribute to the anti-inflammatory properties of blue chamomile oil.

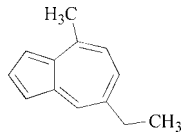

Chamazulene

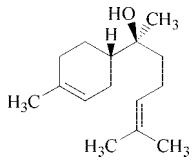

(−)-α-Bisabolol

d_{25}^{25} 0.910 – 0.950; acid value: 5 – 50; ester value: 0 – 40; ester value (after acetylation): 65 – 155; solubility: solutions in 95 % ethanol usually remain turbid [229].

2) *Roman chamomile oil* (English chamomile oil) is produced from the dried flowers of *Chamaemelum nobile* (L.) All. (*Anthemis nobilis* L.), Asteraceae, which are cultivated primarily in Belgium, but also in England, France, and Hungary. The oil is distilled in France and England. It is a light blue or light greenish-blue liquid with a strong aromatic odor characteristic of the flowers.
d_{20}^{20} 0.900 – 0.920; n_D^{20} 1.4380 – 1.4460; $[\alpha]_D^{20}$ −2° to +3°, acid value: max. 8; ester value: 250 – 340; solubility: 1 vol in 0.6 vol of 90 % ethanol (solutions may contain a slight precipitate) [230].

The constituents of Roman chamomile oil include esters of angelic and tiglic acids.

3) *Moroccan chamomile oil* (wild chamomile oil) is produced from *Ormenis* species (e.g., *O. multicaulis, O. mixta*) that grow wild in Morocco. It is a light yellow liquid with a sweet, herbaceous, woody balsamic odor. Its main constituents are irregular monoterpenes, such as santolina alcohol [*35671-15-9*] and artemisia alcohol [*29887-38-5*].

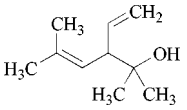

Santolina alcohol

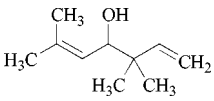

Artemisia alcohol

Cinnamon oils

1) *Cassia oil* (Chinese cinnamon oil) is obtained by steam distillation of the leaves, leaf stalks, and young twigs, of *Cinnamomum aromaticum* Nees (*C. cassia* Bl., Lauraceae). It is a yellow to reddish-brown liquid with a sweet-spicy, cinnamon-like odor, annual production around 300 t.
d_{20}^{20} 1.052 – 1.070; n_D^{20} 1.6000 – 1.6140; solubility: 1 vol in 3 vol of 70 % ethanol at 20 °C; acid value: max. 15; content of carbonyl compounds (calculated as cinnamaldehyde): min. 80 % [231].
In contrast to cinnamon bark oil, cassia oil contains a considerable amount of 2-methoxycinnamaldehyde (5 to 10 %) in addition to its main constituent, cinnamaldehyde. Cassia oil is used predominantly in flavor compositions. (Limiting IFRA recommendation)

2) *Cinnamon leaf oil* is produced by steam distillation of the leaves of the cinnamon tree, *Cinnamomum zeylanicum* Blume. The main countries in which the oil is produced are Sri Lanka, the Seychelles, southern India, Madagascar, and the Comoro Islands. It is a reddish-brown to dark brown liquid with a characteristic spicy odor, reminiscent of clove buds.
Specifications of cinnamon leaf oil from Sri Lanka are d_{20}^{20} 1.037 – 1.053; n_D^{20} 1.5270 – 1.5400; $[\alpha]_D^{20}$ – 2.5 ° to + 2 °; solubility: 1 vol in 2 vol of 70 % aqueous ethanol at 20 °C; phenol content: 75 – 85 % [232].
The main component of cinnamon leaf oil is eugenol. The oil is used as such or as a source of high-grade eugenol (see page 66).

3) *Sri Lanka cinnamon bark oil* is obtained by steam distillation of the dried bark of the cinnamon tree. It is a yellow liquid with the odor and burned-spicy taste of cinnamon.
d_{25}^{25} 1.010 – 1.030; n_D^{20} 1.5730 – 1.5910; $[\alpha]_D$ – 2 ° to 0 °; aldehyde content (calculated as cinnamaldehyde): 55 – 78 %; solubility: 1 vol in at least 3 vol of 70 % ethanol [233].
The oil is used predominantly in flavor compositions. (Limiting IFRA recommendation)

Cistus oil, see Labdanum Oil, page 104.

Citronella oil, see Cymbopogon Oils, page 98.

Citrus oils comprise both essential oils obtained from the peels of citrus fruits and essence oils obtained by concentrating citrus juice (see Section 3.2.1). The essential oils and the essence oils have similar compositions.

Oils isolated from other parts of citrus plants (blossoms and leaves) are not classified as citrus oils because the former show marked differences in composition and organoleptic properties, they are described in other sections (see Neroli Oil, page 109 and Petitgrain Oils, page 112).

Production of Citrus Peel Oils. Apart from distilled lime oil, citrus peel oils are produced by pressing. Pressing of the peels for oil is often combined with juice production.

In the first industrial production process, the citrus fruit was cut into halves and the juice was then pressed out. Further pressing of the peel liberated the desired oil. This mechanical procedure, which is still used, is known as *cold pressing.* Currently, most citrus peel oils are produced by other methods. For instance, the peel of the whole fruit is rasped or punctured before juice extraction to release the oil. The oil is then rinsed off with water and is subsequently separated from the resulting emulsion by centrifugation.

Production of Citrus Essence Oils. Distillative concentration of citrus juices yields essence oils, which separate from the aqueous phase in the receiver when the distillate condenses. The composition of essence oils is similar to that of

peel oils, but the essence oils usually contain a larger amount of aliphatic ethyl esters (e.g., ethyl butyrate in orange essence oil). Thus, their aroma resembles that of a particular juice more than that of peel oils.

Citrus oils contain up to 95 % monoterpene hydrocarbons (usually limonene, but others as well, e.g., lemon oil also contains α-terpinene and β-pinene). The important aroma-determining components of citrus oils are functionalized terpenes and aliphatic compounds (predominantly carbonyl compounds and esters), present only in very low concentrations. Thus, several methods are employed to concentrate citrus oils on an industrial scale. The monoterpene hydrocarbon content is decreased by distillation, liquid–liquid partitioning between two immiscible solvents, or absorption on a carrier such as silica gel. The deterpenized concentrates are marketed under the name "Citrus oil *x*-fold", depending on the concentration factor.

1) *Bergamot oil, Italian* is obtained by cold-pressing peels from the unripe fruits of *Citrus aurantium* L. subsp. *bergamia* (Risso et Poit.) Wight et Arn. (Rutaceae). It is a green to greenish-yellow liquid, which sometimes contains a deposit. The oil has a pleasant, fresh, sweet, fruity odor.
d_{20}^{20} 0.876 – 0.883; n_D^{20} 1.465 – 1.470; $[\alpha]_D^{20}$ + 15 ° to + 32 °; evaporation residue 4.5 – 6.4 %; solubility: 1 vol in 1 vol of 85 % ethanol at 20 °C; acid value: max. 2.0; ester value: 86 – 129, corresponding to an ester content of 30 – 45 % (calculated as linalyl acetate) [234].
Bergamot is still grown almost exclusively in the Italian province of Calabria, where the soil and climatic conditions are very favorable for its cultivation. The bergamot is grafted onto bitter orange plants. The quality of the oil is determined by its ester content and varies with the species. Annual production is ca. 90 t.
Linalyl acetate, linalool (10 %), and citral (1 %) are important components of bergamot oil. Its terpene content (25 – 50 %) is relatively low for a citrus oil.
Use of the untreated oil in cosmetics is limited by the photosensitizing properties of bergaptene (a furocoumarin) present in the oil, thus an IFRA recommendation exists. Bergamot oil is used in many perfumes and is an important component of eaux de cologne.
2) *Grapefruit oil* is obtained by cold pressing of the outer peels of the fruits of *Citrus paradisi* Macfad. (Rutaceae). It is a yellow to reddish orange liquid, with fresh citrus-like odor characteristic of grapefruit peel.
d_{20}^{20} 0.852 – 0.860; n_D^{20} 1.4740 – 1.4790; $[\alpha]_D^{20}$ + 91 ° to + 96 °; carbonyl value: min. 3.25; evaporation residue: min. 5.0 %, max. 10.0 % [235].
Grapefruit oil is produced mainly in the United States and its composition varies with the species. White grapefruit oil obtained from the varieties Marsh seedless and Duncan that are grown in Florida, contains ca. 90 % terpene hydrocarbons (mainly (+)-limonene), 0.5 % alcohols (mainly linalool), 1.8 % aldehydes (mainly octanal and decanal), 0.5 % ketones (mainly nootkatone), 0.3 % esters (mainly octyl and decyl acetate), and 7.5 % nonvolatile components [236, 237].
The sesquiterpene ketone nootkatone (for structure, see page 39) is primarily responsible for the characteristic aroma of grapefruit oil.
Grapefruit oil is used mainly for flavoring fruit beverages. Annual production is 200 to 220 t. (Limiting IFRA recommendation)
3) *Lemon oil* is obtained by pressing peel from the fruits of *Citrus limon* (L.) Burm. f., (Rutaceae). It is a pale yellow to pale greenish-yellow liquid with a characteristic lemon peel odor. ISO differentiates between American, Mediterranean, and Equatorial type. The American type is subdivided into 'coastal' and 'desert' type. In the following the physical properties of the 'coastal' type are given.
d_{20}^{20} 0.852 – 0.858; n_D^{20} 1.4730 – 1.4760; $[\alpha]_D^{20}$ + 57 ° to + 65.6 °; evaporation residue: 1.75 – 3.6 % [238]; acid value: max. 2 %; content of carbonyl compounds (calculated as citral): 8.0 – 14.0 %.
The composition of lemon oils depends on the variety of lemon and the country of origin. Their main components are terpenes – (+)-limonene (ca. 65 %), β-pinene, and γ-terpinene (8 – 10 % each) being the most important. The characteristic odor of lemon oil differs from that of other citrus oils and is

largely due to neral and geranial. The content of these compounds is in the coastal type between 1.6% and 2.9 %, in Italian lemon oil between 1.4 and 3.2 % [238].

Annual worldwide production is about 3.7×10^3 t, most of which originates from the United States, Italy, and Argentina.

Lemon oil is used in many food flavors. Because of its fresh odor, relatively large quantities are also employed in eaux de cologne and other perfumery products (see IFRA recommendation).

4) *Lime oil* may be either cold-pressed or distilled, but the distilled oil is produced on a much larger scale.

Two varieties of limes are of importance in the commercial production of lime oils: the West Indian (Mexican or Key) lime, *Citrus aurantiifolia* (Christm. et Panz.) Swingle and the Persian (Tahiti) lime, *C. latifolia* Tanaka (Rutaceae). The former has small fruits with many seeds, and the latter bears larger, seedless fruits. West Indian limes are grown primarily in Mexico, the West Indies and Peru; Persian limes are cultivated in Florida and Brazil.

Persian lime oils contain lower concentrations of the typical organoleptic components than the West Indian lime oils and are thus comparatively mild and flat [239]. Consequently, West Indian lime oils are more popular.

*Pressed lime oils*are obtained by rasping and puncturing (type B). Other methods are also employed for West Indian limes: the whole fruits may be chopped and the oil–juice emulsion subsequently centrifuged (type A) [240,241].

Pressed lime oil (Mexican type, A) is a yellowish green liquid with a strong, characteristic odor, reminiscent of lemon. Limiting IFRA recommendation.

Type A: d_{20}^{20} 0.875 – 0.884, n_D^{20} 1.482 – 1.486, $[\alpha]_D^{20}$ + 35 ° to + 42 °; carbonyl value: min. 16, max. 31, corresponding to 4.5 to 8.5 % of carbonyl compounds expressed as citral; evaporation residue: 10.0 – 14.5 % [242].

Cold-pressed lime oils are produced in a quantity of about 300 t/a. The composition and uses of pressed lime oil are similar to those of pressed lemon oil.

Distilled lime oils are produced by steam distillation of an oil–juice emulsion that is obtained by chopping the whole fruit. The acid present in the juice acts on the oil released from the peel and changes its characteristics. The original components are modified to form a series of new compounds.

Distilled lime oil is a colorless to pale yellow liquid, with a characteristic odor, which differs from that of the fresh fruit and the cold-pressed oil.

d_{20}^{20} 0.856 – 0.865; n_D^{20} 1.474 – 1.478; $[\alpha]_D^{20}$ + 34 ° to + 45 °; evaporation residue: max. 2.5 %; content of carbonyl compounds (calculated as citral): max. 1.5 % [243].

Acid-catalyzed cyclization and dehydration of citral and linalool give rise to several compounds that occur at comparatively high concentrations and contribute to the typical aroma of distilled lime oil (e.g., 1,4-cineole [*470-67-7*], 1,8-cineole [*470-82-6*], 2,2,6-trimethyl-6-vinyltetrahydropyran [*7392-19-0*], and 2-(2-buten-2-yl)-5,5-dimethyltetrahydrofuran [*7416-35-5*]).

Annual worldwide production of distilled West Indian lime oil is about 750 t. The oil is used primarily in soft drinks.

5) *Mandarin oil* is obtained by cold pressing the peel of mandarin oranges, the fruits of *Citrus reticulata* Blanco (Rutaceae). The oil is a greenish-yellow to reddish-orange liquid, depending on the degree of ripeness of the fruit, with a pale blue fluorescence and a characteristic odor, reminiscent of mandarin peel.

d_{20}^{20} 0.848 – 0.855; n_D^{20} 1.474 – 1.478, $[\alpha]_D^{20}$ + 64 ° to + 75 °; evaporation residue: 1.6 – 4.0 %; content of carbonyl compounds (calculated as decanal): 0.4 – 1.2 %; acid value: max. 2.0 [244].

The main components are limonene (ca. 70 %) and γ-terpinene (ca. 20 %). The characteristic feature of mandarin oil is its content of α-sinensal, methyl *N*-methylanthranilate (which is responsible for fluorescence), and long-chain unsaturated aliphatic aldehydes.

Annual production of mandarin oil in Italy is ca. 100 t; smaller quantities are produced in Spain, Brazil, and Argentina.

Mandarin oil is used to enrich the bouquet of flavor compositions containing sweet orange

oils as the main component. It is also used in liqueurs and perfumery.

6) *Orange oil bitter* is obtained by pressing fresh peel from the fruits of *Citrus aurantium* L. subsp. *aurantium* (Rutaceae). It is produced mainly in Mediterranean countries and is a pale yellow to yellowish-brown liquid with a slightly mandarin-like odor and a somewhat bitter aroma.
d_{20}^{20} 0.8480 – 0.8600; n_D^{20} 1.4720 – 1.4760; $[\alpha]_D^{20}$ +88° to +98°; evaporation residue: 2 – 5 %; aldehyde content (calculated as decanal): min. 0.5 %, max. 2.9 % [245].
The composition of bitter orange oil differs from that of sweet orange oil; i.e., its aldehyde content is lower and its ester content is higher.
Worldwide production of bitter orange oil (ca. 10 t/a) is lower than that of other pressed peel oils. Bitter orange oil is used in eaux de cologne and liqueurs. Limiting IFRA recommendation.

7) *Orange oil sweet* is obtained from the peel of the fruits of *Citrus sinensis* (L.) Osbeck. It is a yellow to reddish-yellow liquid with the characteristic odor of orange peel and may become cloudy when chilled. Sweet orange oil is often produced in combination with orange juice (e.g., in the United States, Brazil, Israel, and Italy). Specifications for the oils depend on the cultivated varieties and the country of origin [246]
The ISO specification of Orange oil sweet is as follows:
d_{20}^{20} 0.842 – 0.850; n_D^{20} 1.4700 – 1.4760; $[\alpha]_D^{20}$ +94° to +99°; evaporation residue: 1.0 – 5.0 %; content of carbonyl compounds (calculated as decanal): 0.9 – 3.1 %.
The oils have a high terpene hydrocarbon content (> 90 %, mainly (+)-limonene), but their content of oxygen-containing compounds differs and affects their quality. Important for aroma are aldehydes, mainly decanal and citral, and aliphatic and terpenoid esters. The sesquiterpene aldehydes α-sinensal [*17909-77-2*] and β-sinensal [*6066-88-8*], which contribute particularly to the special sweet orange aroma, also occur in other citrus oils, although in lower concentration.
Worldwide production of cold-pressed orange oil is over 27 000 t/a. Its main uses are the flavoring of beverages and confectioneries and perfuming eaux de cologne, soap, and household products.

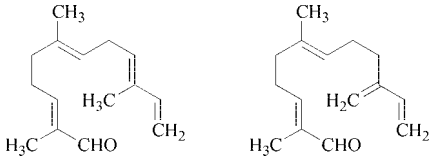

α-Sinensal β-Sinensal

Civet, see Animal Secretions, page 87.

Clary sage oil, see Sage Oils, page 115.

Clove oils are produced from the 15 – 20-m high clove tree *Syzygium aromaticum* (L.). Merr. et L. M. Perry, syn: *Eugenia caryophyllus* (Spreng.) Bullock ex S. G. Harrison (Myrtaceae).

1) *Clove bud oil* is obtained in 15 – 20 % yield by steam distillation of the dried flower buds. Clove bud oil, like the leaf oil, is a yellow to brown, sometimes slightly viscous liquid. It turns dark purple-brown on contact with iron. The oil has the spicy odor characteristic of eugenol.
d_{20}^{20} 1.042 – 1.063; n_D^{20} 1.5280 – 1.5380; $[\alpha]_D^{20}$ −1.5° to 0°; solubility: 1 vol in 2 vol of 70 % ethanol at 20 °C; phenol content: min. 85 %, max. 93 % [247].

2) *Clove leaf oil* is obtained in 2 – 3 % yield by steam distillation of the leaves of the abovementioned.
d_{20}^{20} 1.039 – 1.049; n_D^{20} 1.5280 – 1.5350; solubility: 1 vol in 2 vol of 70 % ethanol at 20 °C; phenol content: minimum 82 % [248].
The main component of all clove oils is eugenol (draft for a GLC specification for leaf oils: 80 to 92 %), which is responsible for their odor and antiseptic properties. Other major constituents are eugenyl acetate and caryophyllene. Clove bud oil has a higher acetate content and a more delicate odor than leaf oil, but is much more expensive than the leaf oil. Leaf oil is produced and used in the largest quantities.
The most important countries that produce clove oils are Madagascar, Tanzania, and Indonesia. Smaller quantities are produced in

other tropical areas (e.g., Malaysia and Sri Lanka). Worldwide production of clove oils is 2000 – 2200 t/a, of which Indonesia produces about half.

The oils are used in many perfume and aroma compositions, because of their spicy clove odor. A small amount is used as an antiseptic, mainly in dentistry. The leaf oil, in particular, is also used as a raw material for the production of eugenol, which is the starting material for further commercially important fragrance compounds, such as isoeugenol and eugenyl esters (see page 65).

Copaiba (balsam) oils are obtained by steam distillation of the exudate (balsam) from the trunk of several species of *Copaifera* L. (Fabaceae), a genus of trees growing in the Amazon basin. They are colorless to light yellow liquids with the characteristic odor of the corresponding balsams and an aromatic, slightly bitter, pungent taste.

d_{25}^{25} 0.880 – 0.907, n_D^{20} 1.4930 – 1.5000; $[\alpha]_D$ – 33 ° to – 7 °; solubility: 1 vol in 5 – 10 vol of 95 % ethanol [249].

The oils consist primarily of sesquiterpene hydrocarbons; their main component is caryophyllene (> 50 %, for formula, see Section 2.3).

Copaiba balsam oils and balsams are used mainly as fixatives in soap perfumes.

Coriander oil is obtained by steam distillation of ripe fruits of *Coriandrum sativum* L. (Apiaceae). It is an almost colorless to pale yellow liquid with a characteristic spicy odor, reminiscent of linalool.

d_{20}^{20} 0.862 – 0.878; n_D^{20} 1.4620 – 1.4700; $[\alpha]_D^{20}$ + 7 ° to + 13 °; solubility: 1 vol in 8 vol of 65 % ethanol at 20 °C; acid value: max. 3.0; linalool content by chromatography: min. 68 % [250].

The main component of coriander oil is (+)-linalool (60 – 80 %). Mono- and polyunsaturated fatty aldehydes, although minor components, contribute to the characteristic aroma of the oil because of their powerful odor.

Coriander is mainly cultivated in the CIS, Poland, and Hungary. Worldwide oil production is 50 – 100 t/a. Coriander oil is no longer important as a raw material for the production of linalool and its derivatives. However, it is still used extensively in seasoning mixtures and in perfume compositions.

Cornmint oil, see Mint Oils, page 107.

Cumin oil is obtained by steam distillation of ripe fruits of *Cuminum cyminum* L. (Apiaceae). It is a light yellow to brown liquid with a strong, somewhat fatty and green odor.

d_{20}^{20} 0.905 – 0.930; n_D^{20} 1.490 – 1.5060; $[\alpha]_D^{20}$ + 1 ° to + 8 °; solubility: 1 vol in 8 vol of 80 % ethanol at 20 °C; aldehyde content: 45 – 58 % (calculated as cuminaldehyde) [251].

Cymbopogon oils are produced from several aromatic grasses that belong to the genus *Cymbopogon* Spreng. (Poaceae). The oils are obtained by steam distillation of the parts of the plants that grow above ground. The following oils are of commercial interest:

1) *Citronella oil* is available in two types:
   a) *Ceylon or Sri Lanka citronella oil* is produced by steam distillation of fresh or partly dried leaves and stems of the grass species *Cymbopogon nardus* (L.) W. Watson–"lenabattu"–cultivated in Sri Lanka. It is a pale yellow to brownish liquid with a fresh, grassy, camphoraceous odor.
   d_{20}^{20} 0.891 – 0.910; n_D^{20} 1.479 – 1.490; $[\alpha]_D^{20}$ – 25 ° to – 12 °; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C; ester value (after acetylation): 157 – 200; carbonyl value: 18 – 55, corresponding to 5 – 15 % carbonyl compounds (calculated as citronellal) [252].
   Ceylon oil is less valuable than Java oil and is used almost exclusively for perfuming toilet soaps, washing powders, and household products.
   b) *Java citronella oil* is obtained by steam distillation of fresh or partially dried stems and leaves of *Cymbopogon winterianus* Jowitt syn. *Cymbopogon nardus* W. Watson var. mahapengiri Winter, which is grown in Southeast Asia, India, and Indonesia, as well as in Central and South America. It is a pale yellow to pale yellowish brown liquid with a slightly sweet, flowery, roselike odor with the strong citrus note of citronellal.

d_{20}^{20} 0.880 – 0.893; n_D^{20} 1.467 – 1.473; $[\alpha]_D^{20}$ $-5°$ to $0°$; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C, opalescence is sometimes observed when ethanol is continuously added; ester value (after acetylation): min. 250, corresponding to 85 % acetylizable compounds (calculated as geraniol, this percentage includes citronellal, since it is converted quantitatively into isopulegyl acetate under the acetylation conditions); carbonyl value: min. 127, corresponding to 35 % carbonyl compounds (calculated as citronellal) [253].

Java citronella oil may contain up to 97 % acetylizable compounds and up to 45 % carbonyl compounds, depending on the time of harvesting. Is is used extensively not only in perfumery, but also as one of the most important raw materials for the production of citronellal. In addition, a fraction with a high geraniol content is obtained from the oil. Both citronellal (see page 22) and the geraniol fraction (see page 15) are starting materials for the synthesis of a large number of other fragrance compounds. The oil produced in Taiwan and in Java contains, in addition to the major components citronellal and geraniol, 11 – 15 % citronellol, 3 – 8 % geranyl acetate, 2 – 4 % citronellyl acetate, and many other minor components.

Annual worldwide production was reported to be > 5000 t in 1971 and is now less than 2000 t. Main producers are Taiwan, China (province of Hainan), and Java [254].

2) *Lemongrass oil* is available in two types, which are produced by steam distillation.
   a) *West Indian or Guatemala lemongrass oil* is obtained from *Cymbopogon citratus* (DC.) Stapf in Central and South America, as well as in a number of African and East Asian countries. It is a pale yellow to orange-yellow liquid with a lemon-like odor, characteristic of citral.
   d_{20}^{20} 0.872 – 0.900; n_D^{20} 1.483 – 1.489; $[\alpha]_D^{20}$ $-6°$ to $0°$; content of carbonyl compounds (calculated as citral): min. 75 %; solubility: freshly distilled oil is soluble in 70 % ethanol at 20 °C, but solubility diminishes on storage and the oil may become insoluble in 90 % ethanol. Residues remain after vacuum distillation of oils stored for longer times due to the high molecular mass products formed by polymerization of myrcene. The oil contains up to 20 % myrcene [255].
   b) *Indian lemongrass oil* is obtained from the so-called Indian variety of lemongrass, *Cymbopogon flexuosus* (Nees ex Steud.) Stapf. The oil is produced mainly in India. Its physical constants differ slightly from those of the West Indian type, the main difference being in the density, which is d_{20}^{20} 0.885 – 0.905 [256].

   The two oils were formerly the main source of natural citral, obtained as a ca. 4 : 1 mixture of geranial and neral by distillation (see page 21). However, lemongrass oil has declined in commercial importance due to the competitive synthesis of citral (see page 21) and isolation of natural citral from Litsea cubeba oil (see page 106). Nevertheless, between 300 and 500 t/a are still produced. In addition to being processed into citral, it is used to some extent for perfuming soap and household products.

3) *Palmarosa oil* is obtained by steam distillation of wild or cultivated *Cymbopogon martinii* (Roxb.) W. Watson. var. *motia,* collected when in blossom. It is a pale yellow liquid with a characteristic roselike odor and a grassy note.

   d_{20}^{20} 0.880 – 0.894; n_D^{20} 1.4710 – 1.4780; $[\alpha]_D^{20}$ $-1.4°$ to $+3°$; solubility: 1 vol in 2 vol of 70 % ethanol at 20 °C; acid value: max. 1.0; the following data are valid for oil from India, oils of other origin possess slightly different data, ester value: min. 7, max. 36; ester value (after acetylation): 260 – 280, corresponding to a total alcohol content of 88 – 95 %, free alcohol content (calculated as geraniol): 74 – 94 % [257]. High-grade palmarosa oil may contain up to 95 % geraniol and its esters; it is produced in smaller quantities than other oils obtained from aromatic grasses. Annual worldwide production is ca. 150 t. Palmarosa oil is the starting material for geraniol and geranyl esters of high odor quality, but it is also used for perfuming soaps and cosmetics.

4) *Gingergrass oil* is produced in India from the *sofia* variety of *Cymbogon martinii* and is less important than palmarosa oil.

Cypress oil is produced by steam distillation of terminal branches of *Cupressus sempervirens* L. (Cupressaceae). It is a liquid with a woodlike odor that has an ambergris note.

d_{20}^{20} 0.863 – 0.885; n_D^{20} 1.468 – 1.478; $[\alpha]_D^{20}$ + 15 ° to + 30 °; solubility: 1 vol in 8 vol of 90 % ethanol.

This oil is produced exclusively in southern France and Algeria (ca. 10 t/a); its major components are α-pinene and 3-carene. Degradation products of higher terpenoids are responsible for the typical ambergris note [258, 259].

Davana oil is obtained by steam distillation of the herb *Artemisia pallens* Wall. (Asteraceae), grown in south India. It is an orange-brown liquid with a sweet tealike odor reminiscent of dried fruits. The composition of the oil is very complex; its main components are furanoid sesquiterpenes. It is used predominantly for aroma compositions.

Dill oil is obtained from the dill plant, *Anethum graveolens* L. (Apiaceae), in two different forms:

1) *Dill weed oil* is prepared from flowering dill by steam distillation. Its main components are α-phellandrene (up to 60 %) and the so-called (+)-dill ether [*74410-10-9*] (up to 30 %). The latter is responsible for the typical organoleptic properties of the dill plant and, thus, of dill weed oil.

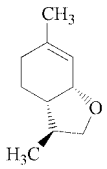

(+)-Dill ether

2) *Dill seed oil* is prepared by steam distillation of the crushed ripe fruits of the dill plant. Its main components are limonene (up to 40 %) and (+)-carvone (up to 60 %). This oil has a typical caraway odor and taste in contrast to the weed oil, which contains only a small amount of carvone.

Commercially available dill oils are mixtures of both types, because the plants are harvested not according to the criteria for the special oils, but when the state of ripeness guarantees a maximum yield. They generally contain α-phellandrene (10 – 20 %), limonene (30 – 40 %), dill ether (3 – 10 %), and carvone (30 – 40 %).

The main growing areas are the Balkan states, Hungary, and the United States. Dill oils are used primarily for seasonings in the pickling and canning industries.

Elemi oil, Elemi resinoid are obtained from exuded gum resin of *Canarium luzonicum* (Miq.) (Burseraceae), a tree growing in the Phillipines. The resin is extracted with a solvent to form the resinoid, which is a yellow to orange mass of high viscosity. The oil is produced by steam distillation of the gum oleoresin and is a colorless to light yellow liquid.

d_{20}^{20} 0.850 – 0.910; n_D^{20} 1.4720 – 1.4900; $[\alpha]_D^{20}$ + 44 ° to + 85 ° [260].

The major components of elemi oil are limonene, α-phellandrene, and the sesquiterpene alcohol elemol. Both the resinoid and the oil have a fresh, citrus-like, peppery odor and are used predominantly in soap perfumes.

Estragon oil, see Tarragon Oil, page 117.

Eucalyptus oils are produced from plants belonging to the genus *Eucalyptus* (Myrtaceae), which includes ca. 500 species in Australia, the country of origin, alone. Correct botanical classification was possible only by determining the chemical composition of the essential oils obtained from the leaves. At present, few of these oils are commercially important.

1) *Cineole-rich Eucalyptus oils*
   a) *Australian eucalyptus oil* is obtained by steam distillation of the foliage of certain Eucalyptus species indigenous to Australia, e.g., from *Eucalyptus fruticetorum* F.v. Muell. (*E. polybractea* R. T. Bak.), *E. smithii* R. T. Bak., and *E. leucoxylon* F.v. Muell.
   d_{20}^{20} 0.918 – 0.928; n_D^{20} 1.458 – 1.465; $[\alpha]_D^{20}$ – 2 ° to + 2 °; solubility: 1 vol in 3 vol of 70 % ethanol at 20 °C; 1,8-cineole content: 80 – 85 % [261].
   The minor components of this oil differ from those of *E. globulus* oil.

b) *Eucalyptus globulus oil* is produced by steam distillation of the leaves and twigs, recently collected of *Eucalyptus globulus* Labillardière. It is an almost colorless to pale yellow liquid with a fresh odor, characteristic of 1,8-cineole. The crude oil contains ca. 65 % cineole and more than 15 % α-pinene. Commercially available qualities are rectified, and have a cineole content of 70 – 75 % or 80 – 85 %. The respective specifications of these two types are as follows:
d_{20}^{20} 0.904 – 0.920 and 0.906 – 0.920; n_D^{20} 1.460 – 1.468 and 1.458 – 1.465; $[\alpha]_D^{20}$ 0 ° to + 10 ° and + 2 ° to + 10 °; solubility 1 vol in 5 vol of 70 % ethanol and 1 vol in 5 vol of 80 % ethanol at 20 °C [262].
The oil is produced mainly in Spain and Portugal, where the wood is used in the cellulose industry, and in China. Worldwide production is between 3500 and 4000 t/a. Eucalyptus oils with a high cineole content are used for cineole production (see page 72). The oils and cineole itself are used primarily in pharmaceutical preparations. Fairly large quantities are also used in perfumery, e.g., to imitate the odor of cineole-containing essential oils. Distillation fractions of camphor oil (see page 90) are also marketed as eucalyptus oils with a cineole content of 70 – 80 %.

Eucalyptus citriodora oil is obtained by steam distillation of leaves and twigs of *Eucalyptus citriodora* Hook. It is an almost colorless, pale yellow, or greenish-yellow liquid with a citronellal-like odor.
d_{20}^{20} 0.860 – 0.870; n_D^{20} 1.450 – 1.456; $[\alpha]_D^{20}$ − 1 ° to + 3 °; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C; carbonyl value (calculated as citronellal): min. 70 % [263].
In addition to the main component, citronellal, the oil contains citronellol and isopulegol (5 – 10 % each).
Young *E. citriodora* trees that are grown exclusively for essential oil production are cut back to a height of 1 – 1.50 m and develop into shrubs. The leaves can be harvested throughout the year; more than 200 kg of oil can be obtained per hectare.
The major producer is Brazil with over 500 t/a, but considerable quantities are also produced in other countries (e.g., South Africa and India). The annual production worldwide is about 1300 t. *Eucalyptus citriodora* oil is a starting material for the manufacture of citronellal and products derived from it. It is also used in perfumery for the same purposes as citronellal (see page 22).

3) *Eucalyptus dives oil* is obtained by steam distillation of fresh leaves of *Eucalyptus dives* var. *Schau, Typus,* grown in New South Wales.
d_{15} 0.889 – 0.910; n_D^{20} 1.477 – 1.481; $[\alpha]_D^{20}$ − 75 ° to − 50 °; content of carbonyl compounds (calculated as piperitone): 40 – 50 % [264].
In addition to (−)-piperitone (for formula see page 30), the oil contains 15 – 25 % α- and β-phellandrene. The oil was previously used as a starting material in the manufacture of (−)-menthol, but has lost much of its significance. Annual worldwide production in 2002 has dropped to 50 t.

4) *Eucalyptus staigeriana oil* is produced by steam distillation of the leaves of *Eucalyptus staigeriana* F. v. Muell. It has a fresh, lemon-like odor and contains ca. 30 % citral. Further constituents are geraniol and nerol and their acetates, as well as (−)-limonene. It is used mainly in perfumery [265].

Fennel oil is commercially available in sweet and bitter types that are obtained from two varieties of a subspecies of common fennel, *Foeniculum vulgare* Mill., subsp. *capillaceum* Gilib. (Apiaceae). The sweet oil is obtained from the *dulce* variety and the bitter oil from the variety *vulgare* (Mill.) Thell.

The oils are produced both in the Mediterranean countries and in other continents and are obtained by steam distillation of crushed fennel seeds. They are colorless to pale yellow liquids with a typical fennel odor and a camphoraceous, initially sweet taste that later becomes bitter.

d_{20}^{20} 0.961 – 0.977; n_D^{20} 1.5280 – 1.5430; $[\alpha]_D^{20}$ + 11 ° to + 24 °; *fp* 3 – 11.5 °C; solubility: 1 vol in 5 vol of 80 % ethanol.

Both oils are similar in composition; their major components are *trans*-anethole (> 80 % in sweet fennel oil) and fenchone (up to ca. 25 % in bitter fennel oil). Sweet fennel oil is more important because of its higher anethole content. While both oils are used in flavors, small quantities are also employed in perfumery.

Fir needle oils, see Pinaceae Oils, page 113.

Galbanum oil and galbanum resinoid are produced from the gumlike exudate of *Ferula galbaniflua* Boiss. and Buhse (Apiaceae) growing in northern Iran and *F. rubricaulis* Boiss. growing in southern Iran.

The gum is collected from a cut in the upper part of the uncovered roots. The annual yield of gum is ca. 100 t.

The oil is produced by steam distillation and is a colorless to light yellow liquid with a green, slightly spicy odor.

d_{20}^{20} 0.8670 – 0.8900; n_D^{20} 1.4780 – 1.4850; $[\alpha]_D^{20}$ +7° to +17°; acid value: max. 2; solubility: 1 vol in at least 6 vol of 90 % ethanol [266].

In addition to 75 % monoterpene hydrocarbons and ca. 10 % sesquiterpene hydrocarbons, galbanum oil contains a fairly large number of terpene and sesquiterpene alcohols and their acetates. Minor components, with entirely different structures and low odor threshold values, contribute strongly to the chacteristic odor. Examples are as follows:

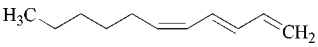

(*E,Z*)-1,3,5-Undecatriene [*51447-08-6*]

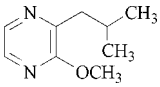

2-Methoxy-3-isobutylpyrazine [*24683-00-9*]

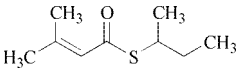

S-sec-Butyl 3-methyl-2-butenethioate [*34322-09-3*]

Galbanum oil is used for creating green top notes in perfume compositions.

Galbanum resinoid is produced by extraction of the gum with a nonpolar solvent. It is used for the same purposes as the oil and has excellent fixative properties.

Geranium oil is obtained by steam distillation of the flowering herb *Pelargonium graveolens* l'Heritier ex Aiton, *P. roseum* Willdenow, and other nondefined hybrids that have developed into different ecotypes in different geographical regions. The oil is an amber to greenish-yellow liquid with the characteristic roselike odor of the plant.

The main cultivation areas are Réunion and Madagascar (Bourbon type), Morocco and Egypt (North African type), and China. The Bourbon quality is more valuable and thus more expensive. Annual worldwide production is ca. 300 t. Specifications of geranium oils [267] are given in Table 3.

The composition of the Bourbon oil differs quantitatively as well as qualitatively from that of North African oil. However, they both contain an unusually high percentage of (−)-citronellol, isomenthone, formates, and tiglates, which are rarely found in essential oils. The two types of oil can be differentiated by two characteristic minor constituents: the Bourbon type contains (−)-6,9-guaiadiene [*36577-33-0*] and the African type contains 10-*epi*-γ-eudesmol [*15051-81-7*] (for structure see page 86). The Chinese type also contains (−)-6,9-guaiadiene, however a smaller amount of menthone/isomenthone and a higher amount of (−)-citronellol than the Bourbon type.

Table 3. Specifications of geranium oils

Parameter	Type of geranium oil			
	Bourbon	Egyptian	Moroccan	Chinese
d_{20}^{20}	0.884 – 0.892	0.887 – 0.892	0.883 – 0.900	0.882 – 0.8
n_D^{20}	1.4620 – 1.4680	1.4660 – 1.4700	1.4640 – 1.4720	1.4600 – 1
$[\alpha]_D^{20}$	−14° to −8°	−12° to −8°	−13° to −8°	−14° to −8°
Solubility in 70 % ethanol (20 °C), vol : vol	1 : 3	1 : 2.2	1 : 3	1 : 5
Acid value (max.)	10	6	10	10
Ester value	52 – 78	42 – 58	35 – 80	55 – 75
Ester value (after acetylation)	205 – 230	210 – 222	192 – 230	215 – 232
Content of carbonyl compounds (as isomenthone), %	max. 16	max. 16	max. 16	max. 10
Alcohol content (as citronellol), %	42 – 55	40 – 58	35 – 58	

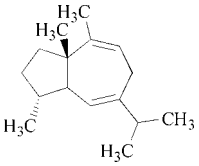

(−)-6,9-Guaiadiene

Geranium oil is one of the most important natural raw materials in the fragrance industry. It is used for creating rosy notes, especially in soap.

Ginger oil and ginger oleoresin are produced from the ginger plant *Zingiber officinale* (L.) Rosc. (Zingiberaceae).

Ginger oil is produced by steam distillation of dried, ground rhizomes. It is a light yellow to yellow liquid with the aromatic, persistent odor of ginger, but lacking the pungency usually associated with ginger. The citrus note of ginger oil is due to traces of citral.

d_{25}^{25} 0.871 – 0.882; n_D^{20} 1.4880 – 1.4940; $[\alpha]_D$ − 45 ° to − 28 °; saponification value: max. 20; soluble in ethanol, solutions are usually turbid [268].

The major components of the oil are β-sesquiphellandrene [*20307-83-9*] (about 15 %) and zingiberene [*495-60-3*] (30 to 40 %).

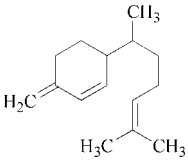

β-Sesquiphellandrene

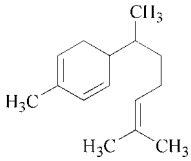

Zingiberene

Ginger oleoresin is prepared by extracting ginger rhizomes with acetone or alcohol. The product contains the essential oil along with the substances responsible for the pungency of ginger. These compounds are substituted phenols of the following structure:

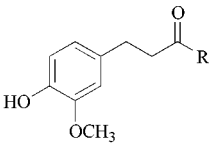

Gingerols: $R = -CH_2-\underset{\underset{OH}{|}}{CH}-(CH_2)_n-CH_3$; $n = 1–4, 8, 10$

Shogaols: $R = -CH=CH-(CH_2)_n-CH_3$; $n = 2, 4, 6, 8$

Main cultivation areas for ginger are India, Malaysia, Taiwan, and the Fiji Islands. Ginger concentrates are used in large amounts in beverages (e.g., ginger ale), as well as in baked goods and confectioneries.

Grapefruitoil, see Citrus Oils, page 94.

Green cognac oil, see Lie de Vin Oil, page 106.

Guaiac wood oil is obtained by steam distillation of ground wood and sawdust from the tree *Bulnesia sarmienti* Lorentz (Zygophyllaceae), which is up to 20 m high and grows wild in the Gran Chaco region of Paraguay and in Argentina. The annual production is ca. 150 t. The oil is a dark yellowish, viscous liquid with a mild, pleasant odor reminiscent of that of tea roses and faintly of violets. The oil solidifies at room temperature to a yellow-white to light amber colored mass (*mp* 40 – 50 °C).

d_{25}^{25} 0.960 – 0.975; n_D^{20} 1.5020 – 1.5070; $[\alpha]_D$ − 12 ° to − 3 °; solubility: 1 vol in at least 7 vol of 70 % ethanol; solutions are sometimes slightly turbid or opalescent; total alcohol content (calculated as guaiol): min. 85 % [269].

The main constituents of the oil are the sesquiterpene alcohols guaiol [*489-86-1*] and bulnesol [*22451-73-6*]. The oil may be used as a starting material for the synthesis of guaiazulene, which has anti-inflammatory properties.

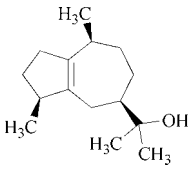

Guaiol

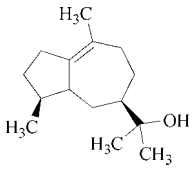

Bulnesol

Guaiac wood oil is used extensively in perfume compositions for its excellent fixative properties. Also the acetates of the sesquiterpene alcohols of this oil are used in perfumery.

Gurjun balsam oil is produced by steam distillation of balsams obtained from several *Dipterocarpus* species (Dipterocarpaceae); the trees grow in South and East Asia. The oil is a yellow, slightly viscous liquid with a weak, woodlike odor.

d_{15} 0.918 – 0.930; n_D^{20} 1.5010 – 1.5050; $[\alpha]_D$ – 130° to – 35°; solubility: 1 vol in 10 vol of ethanol.

It consists almost entirely of sesquiterpene hydrocarbons, and its main component (> 60 %) is α-gurjunene [*489-40-7*].

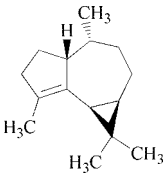

α-Gurjunene

Gurjun balsam oil and gurjun balsams are used for their good fixative properties, e.g., in soap perfumes and serve also as a starting material for the production of guaiazulene.

Jasmin absolute is obtained by solvent extraction, via the concrete, from the flowers of *Jasminum grandiflorum* L. (Oleaceae), cultivated in Egypt, Italy, Morocco, and India, and of *J. sambac* (L.) Ait. from China and India. The concrete is usually a brown to dark brown waxy mass, with a characteristic jasmin odor.

mp 48 – 51 °C; ester value: 70 – 125.

The absolute is generally a reddish-brown liquid with a delicate jasmin odor; the color deepens on storage.

n_D^{20} 1.4780 – 1.4920; acid value: 8 – 14; ester value: 120 – 220.

One ton of jasmin blossoms yields ca. 2.5 – 3 kg of concrete when it is extracted two or three times with hexane or benzene. Extraction of the concrete with ethanol gives a ca. 60 % yield of the absolute.

The main volatile component of jasmin oil is benzyl acetate. However, minor components such as indole [*120-72-9*], *cis*-jasmone [*488-10-8*], and methyl jasmonate [*1211-29-6*] contribute strongly to the typical jasmin fragrance.

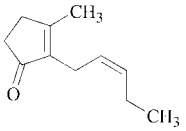

cis-Jasmone

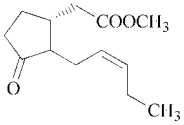

Methyl jasmonate

The absolute is one of the most valuable blossom fragrances used in fine fragrances.

Juniper berry oil is obtained by steam distillation of berries of *Juniperus communis* L. var. *erecta* Pursh. (Cupressaceae). It is a colorless, pale green or yellowish liquid with a characteristic, conifer-like odor, and an aromatic-bitter taste.

d_{20}^{20} 0.857 – 0.872; n_D^{20} 1.4710 – 1.4830; $[\alpha]_D^{20}$ – 15° to 0°; solubility: 1 vol in 10 vol of 95 % ethanol, solutions usually turbid; acid value: max. 2; ester value: min. 12 [270].

The slightly turpentine-like odor and the relatively low solubility of the oil are caused by its high content of α-pinene and other monoterpene hydrocarbons. The main oxygen-containing component is 1-terpinen-4-ol.

Most juniper berries are obtained from bushes growing in European countries, such as Italy, Czech Republic, Slovakia, Hungary, and former Yugoslavia.

The oil is used to a limited extent in perfumery for creating fresh, dry effects and as an aroma ingredient in alcoholic beverages of the gin type.

Labdanum absolute and Labdanum oil are obtained from labdanum gum, which is exuded when twigs of *Cistus ladaniferus* L. (Cistaceae) are boiled in water.

Labdanum absolute is produced by extraction of the gum with alcohol (e.g., methanol or ethanol). So-called colorless absolutes are obtained from absolutes by extraction with, for example, hexane.

Steam distillation of the gum yields *labdanum oil* which is a golden yellow, viscous liquid that quickly turns dark brown on standing:

d_{25}^{25} 0.905 – 0.993; n_D^{20} 1.4920 – 1.5070; $[\alpha]_D$ + 0° 15′ to + 7°, often difficult to determine due to its dark color; solubility: 1 vol in 0.5 vol of 90 % ethanol, solutions frequently opalescent to turbid, paraffins may separate upon further dilution; acid value: 18 – 86; ester value: 31 – 86 [271].

Other odoriferous materials are derived from the leaves and young twigs of *Cistus ladaniferus.* Cistus oil is obtained by steam distillation; solvent extraction yields cistus concrete. Cistus oil, in contrast to labdanum oil, consists mainly of monoterpene hydrocarbons.

Cistus and labdanum products come from Mediterranean countries, mainly Spain, where the *Cistus* shrub grows abundantly. They are used widely in perfumery, giving perfume compositions a warm, balsamic tonality with a touch of ambergris. The resinoids and absolutes are excellent natural fixatives.

Laurel leaf oil is obtained by steam distillation of leaves from *Laurus nobilis* L. (Lauraceae), an evergreen tree cultivated primarily in Mediterranean countries. The oil is a light yellow to yellow liquid with an aromatic, spicy odor.

d_{25}^{25} 0.905 – 0.929; n_D^{20} 1.4650 – 1.4700; $[\alpha]_D$ $-19°$ to $-10°$; solubility: 1 vol in at least 1 vol of 80 % ethanol; acid value: max. 3; saponification value: 15 – 45; saponification value (after acetylation): 36 – 85 [272].

The main component of the oil is 1,8-cineole (30 – 70 %); other important components are linalool (ca. 10 %) and eugenol.

Laurel leaf oil is used extensively in the food industry, e.g., for seasoning meat products and soups.

Lavandula products comprise the following oils and extracts:

1) *Lavender oil* is produced by steam distillation of freshly cut, flowering tops of *Lavandula angustifolia* Mill. (Lamiaceae). It is a pale yellow liquid with a fresh, sweet, floral, herbaceous odor on a woody balsamic base. The following physical constants are valid for spontaneous lavender from France.
d_{20}^{20} 0.880 – 0.890; n_D^{20} 1.458 – 1.464; $[\alpha]_D^{20}$ $-11.5°$ to $-7°$; solubility: 1 vol in 2 vol of 75 % ethanol at 20 °C; acid value: max. 1.0; ester value: 102.5 – 165, corresponding to an ester content of 35.8 – 58 % (calculated as linalyl acetate) [273].
True French lavender grows in the Haute Provence at an altitude of 600 – 1500 m. The plants are grown from seeds of the wild lavender ("spontaneous" or "population" lavender). Lavender oil is produced in a yield of 10 – 25 kg/ha; total production is ca. 100 t/a. It has the following typical composition (%): *cis*-β-ocimene (4 – 10), *trans*-β-ocimene (1.5 – 6), cineole (< 1), camphor (< 0.5), linalool (25 – 38), linalyl acetate (25 – 45), 1-terpinen-4-ol (2 – 6), and lavandulyl acetate (> 2).
Cloned varieties of lavender (e.g., *Maillette* and *Matherone*) yield more oil per hectare and can be grown at lower altitudes; however, they produce a poorer quality oil. Bulgarian lavender oil is similar in composition to that obtained from French population lavender, but is of poorer quality: about 50 to 70 t are produced per year.
Lavender oils of a special quality are also produced in several other areas throughout the world (e.g., Tasmania).
2) *Spanish spike lavender oil* is produced by steam distillation of flowering tops of spike, *Lavandula latifolia* Medik. (Lamiaceae). It is an almost colorless to light orange liquid with a characteristic, rough odor slightly like cineole and camphor.
d_{20}^{20} 0.894 – 0.907; n_D^{20} 1.4610 – 1.4680; $[\alpha]_D^{20}$ $-7°$ to $+2°$; solubility: 1 vol in 3 vol of 70 % ethanol; solutions may become opalescent on dilution; acid value: max. 1.0; ester value: 3 – 14; ester value (after acetylation): 130 – 200 [274].
The main components of Spanish spike lavender oil are linalool (34 – 50 %), cineole (16 – 39 %), and camphor (8 – 16 %).
Spike plants grow wild in the entire Mediterranean area and prefer warmer, lower-lying regions than lavender and lavandin. Oil is primarily produced from plants cultivated in Spain. At present, annual worldwide production is ca. 150 – 200 t.
3) *Lavandin oil abrialis* is obtained by steam distillation of freshly cut flowering tops of lavandin, which is a hybrid of lavender and spike (*Lavandula angustifolia* Mill. x *Lavandula latifolia* Medik.). It is a pale yellow liquid with a lavender-like odor and a slightly camphoraceous note.
Data for lavandin oil abrialis: d_{20}^{20} 0.887 – 0.897; n_D^{20} 1.4560 – 1.4660; $[\alpha]_D^{20}$ $-5°$ to $-2°$; solubility: 1 vol in 4 vol of 70 % ethanol at 20 °C; acid value: max. 1.0; ester value: 77 – 108, corresponding to an ester content of 27 – 38 % (calculated as linalyl acetate) [276].
Data for *lavandin oil grosso* (a specific clone of lavandin, cultivated in the South of France: d_{20}^{20} 0.891 – 0.899; n_D^{20} 1.4580 – 1.4620; $[\alpha]_D^{20}$ $-7°$ to $-3.5°$; solubility:

1 vol in 3 vol of 70 % ethanol at 20 °C; acid value: max. 1; ester value: 100 – 137, corresponding to an ester content of 35 – 48 % (calculated as linalyl acetate) [262].

Lavandin plants are sterile and can be propagated only by using cuttings. The oils from the most important varieties, *abrialis* and *grosso,* contain linalool and linalyl acetate as major constituents as well as 6 – 11 % cineole and 7 – 11 % camphor for abrialis, and 4 – 7 % cineole and 6 – 8 % camphor for grosso. A third variety is called *super* because its oil contains a high concentration of linalyl acetate, and, thus, resembles lavender oil most closely.

Although lavender oil is more valuable than lavandin oil as a fragrance raw material, lavandin plants are more commonly cultivated because they give a higher yield of oil (ca. 50 – 100 kg/ha) and are hardier than lavender plants.

Cultivation in southern France is no longer restricted to the traditional lavender regions, but also now includes the Languedoc. Former Yugoslavia is another lavandin-producing country. Worldwide annual production is up to 1100 t.

All three oils are used primarily in soap perfumes; considerable quantities are also employed in eaux de cologne and in bath products.

4) *Lavender and lavandin extracts* are also commercially important and are produced in southern France by solvent extraction of flowering lavender and lavandin herbs. Production of lavandin concrete is higher than that of lavender. Extraction of the pastelike concretes with ethanol, followed by evaporation, yields absolutes. These extracts differ from the essential oils in being more soluble and in having a dark green color and a longer-lasting odor with a haylike, spicy note. They are also used in eaux de cologne and fine fragrances, sometimes after decoloration (removal of chlorophyll with activated charcoal).

Lemon oil, see Citrus Oils, page 94.

Lemongrass oil, see Cymbopogon Oils, page 98.

Lie de vin oil (green cognac oil or wine lees oil) is obtained by steam distillation of the yeast and other sediments (lees) formed in wine. It is a green to bluish-green liquid with a characteristic cognac aroma.

d_{25}^{25} 0.864 – 0.870; n_D^{20} 1.4275 – 1.4295; $[\alpha]_D$ – 1 ° to + 2 °; acid value: 32 – 70; ester value: 200 – 245; solubility: 1 vol in at least 2 vol of 80 % ethanol [277].

Lie de vin oil consists mainly of the ethyl and isoamyl esters of fatty acids, formed during fermentation. It is used mostly in aroma compositions; only very small amounts are employed in perfume compositions.

Lime oil, see Citrus Oils, page 94.

Litsea cubeba oil is produced by steam distillation of the fruits of *Litsea cubeba* C. H. Persoon (Lauraceae) growing in East Asia. It is a pale yellow to dark yellow liquid with a fresh odor, reminiscent of citral.

d_{20}^{20} 0.880 – 0.892; n_D^{20} 1.4800 – 1.4870; $[\alpha]_D^{20}$ + 3 ° to + 12 °; carbonyl content (calculated as citral): min. 74 %; solubility: 1 vol in 3 vol of 70 % ethanol at 20 °C [278].

The oil is used mainly for the production of citral; smaller quantities are employed for perfuming household products. The main producer is China. The worldwide annual production is about 1000 t.

Lovage oils are obtained by steam distillation of either the leaves (leaf oil) or the fresh roots (root oil) of the perennial plant *Levisticum officinale* Koch (Apiaceae). Both oils are yellow-greenish-brown to dark brown liquids with a strong, characteristic, aromatic odor reminiscent of celery and angelica.

d_{25}^{25} 0.960 – 0.985 (leaf oil); d_{20}^{20} 1.010 – 1.090 (root oil); n_D^{20} 1.5350 – 1.5580; $[\alpha]_D^{20}$ – 1 ° to + 5 °; acid value: 2 – 16; ester value: 170 – 260; solubility: 1 vol in 1.5 vol of 85 % ethanol at 20 °C [279].

The main constituent of the leaf oil is α-terpinyl acetate, while the main constituent and odor-determining component of the root oil is ligustilide [*4431-01-0*]. A few percent ligustilide are also present in the leaf oil.

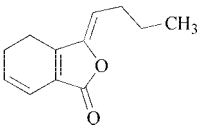

Ligustilide

The oils are very expensive. They are mainly used in the alcoholic beverage industry and for flavoring of tobacco; small amounts are also used in perfumery.

Mace oil, see Nutmeg Oil, page 109.

Mandarin oil, see Citrus Oils, page 94.

Marjoram oil (sweet marjoram oil) is obtained by steam distillation of the herb *Origanum majorana* L. (*Majorana hortensis* Moench, Lamiaceae). The plants are cultivated in France, Hungary, Egypt, and Tunisia. The oil is a yellow to greenish-yellow liquid with a characteristic earthy-spicy odor.

d_{25}^{25} 0.890 – 0.906; n_D^{20} 1.4700 – 1.4750; $[\alpha]_D$ + 14 ° to + 24 °; acid value: max. 2.5; saponification value: 23 – 40; saponification value (after acetylation): 68 – 86; solubility: 1 vol in 2 vol of 80 % ethanol [280].

The main constituent of the oil is 1-terpinen-4-ol (> 20 %), which with (+)-*cis*-sabinene hydrate [*15537-55-0*] (3 – 18 %) is responsible for the characteristic flavor and fragrance of marjoram oil.

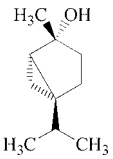

(+)-*cis*-Sabinene hydrate

Sweet marjoram oil is used mainly for flavoring foods, but also in smaller amounts in perfumery for spicy shadings in men's fragrances.

Spanish marjoram oil is obtained by steam distillation of the flowering tops of *Thymus mastichina* L. (Lamiaceae), a wild plant growing in Spain. The oil is a slightly yellow liquid with a fresh camphoraceous, spicy odor.

d_{20}^{20} 0.890 – 0.920; n_D^{20} 1.460 – 1.470; $[\alpha]_D^{20}$ − 4 to + 6 °; 1,8-cineole content: min. 30 %, max. 68 %; solubility: 1 vol in 3 vol of 70 % ethanol [281].

Mentha citrata oil, see Mint Oils.

Mimosa absolute is obtained from the blossoms of the mimosa tree *Acacia dealbata* Link (Mimosaceae), which grows in the forests of southern France and in Morocco.

The absolute is a yellowish brown liquid with a slightly green, flowery odor and extremely high tenacity.

Annual production in France (Grasse) is ca. 3 – 6 t [282]. Mimosa absolute is used mainly in fine fragrances as a flowery fixative.

Mint oils are listed in order of the quantities produced.

1) *Cornmint oil* is produced by steam distillation of the flowering herb *Mentha arvensis* L. var. *piperascens* Malinv. and var. *glabrata* Holmes (Lamiaceae). The crude oil, containing ca. 70 % (−)-menthol, is then partially dementholized by freezing and centrifuging. The dementholized oil is an almost colorless to amber-yellow liquid with a characteristic minty odor; typical specifications for oils of Chinese and Indian origin are as follows:
   d_{20}^{20} 0.890 – 0.910; n_D^{20} 1.4570 – 1.4650; $[\alpha]_D^{20}$ − 24 ° to − 13 °; ester content (calculated as menthyl acetate): 3 – 9 %; total alcohol content (calculated as menthol): 40 – 60 %; ketone content (calculated as menthone): 25 – 45 %; solubility: 1 vol in 4 vol of 70 % ethanol [283].
   Dementholized oil still contains ca. 45 % (−)-menthol; (−)-menthone and (−)-menthyl acetate are other major components. Dementholized *Mentha arvensis* oil was formerly produced primarily in Brazil. Currently, cornmint oil is produced mainly in China; smaller amounts are produced in India and Paraguay. The world production of the oil is estimated to be 17 000 to 18 000 t/a. Dementholized cornmint oil tastes more bitter and stringent than peppermint oil. Thus, the former is used as a cheaper substitute for the latter and for the production of (−)-menthol (see page 30).
2) *Peppermint oil* is produced by steam distillation of the aerial plant *Mentha* x *piperita* L. var. *piperita*. It is an almost colorless to pale

greenish-yellow liquid with a characteristic peppermint odor.
d_{20}^{20} 0.898 – 0.918; n_D^{20} 1.4590 – 1.4650; $[\alpha]_D^{20}$ – 30° to – 14°; solubility: 1 vol in 5 vol of 70 % ethanol at 20 °C; ester value: 12 – 30; ester value (after acetylation): 135 – 200; carbonyl value: 54 – 115 [284].
As in cornmint oil, the main component of peppermint oil is (–)-menthol; it also contains (–)-menthone (ca. 20 %) and (–)-menthyl acetate (up to 8 %). However, peppermint oil, unlike cornmint oil, has a high content of (+)-menthofuran [*17957-94-7*] (1 – 8 %, according to GLC).

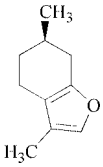

(+)-Menthofuran

The leading producer of peppermint oil is the United States, where annual production from the variety *M. piperita* L. var. *vulgaris* Sole (Black Mitcham) is ca. 3600 t. Cultivation areas are located in the Midwest states, Idaho, Oregon, and Washington. The oils differ in quality and are named according to their geographic origin (e.g., Midwest, Idaho, Madras, Willamette, Yakima). Production of European oil has decreased significantly, despite its high quality. Peppermint oil is used mainly for flavoring toothpaste, other oral hygiene products, and chewing gum. Smaller quantities are used for flavoring confectioneries. Due to its high price, peppermint oil is not used for the production of menthol.

3) *Spearmint oil* is obtained by steam distillation of the flowering tops of *Mentha spicata* L. (native spearmint) and *Mentha* x *cardiaca* Ger. (Scotch spearmint). It is an almost colorless to yellow-green liquid with a fresh, caraway-minty odor.
d_{20}^{20} 0.920 – 0.937; n_D^{20} 1.485 – 1.491; $[\alpha]_D^{20}$ – 60° to – 45°; solubility: 1 vol in 1 vol of 80 % ethanol at 20 °C; carbonyl value: min. 200, corresponding to a carvone content of 55 % [285].
The compositions of Scotch and native spearmint oils differ as shown in Table 4 [286]. The main producer of spearmint oil is the United States, primarily the state of Washington. Smaller amounts are also produced in some Midwest states. Total annual production is ca. 1750 t, 55 % being of the native type and 45 % of the Scotch type.

Table 4. Characteristic components of Scotch and native spearmint oils

Compound	Content of compound in spearmint oil, %	
	Scotch	Native
Limonene	15	9
3-Octanol	2	1
Menthone	1	< 0.1
Dihydrocarvone	1	2.5
Sabinene hydrate	0.1	1.5
Dihydrocarvyl acetate	< 0.1	0.6
cis-Carvyl acetate	0.1	0.6

By far the most spearmint oil is used for flavoring toothpaste and chewing gum. Smaller quantities are used in other oral hygiene products and in pharmaceutical preparations.

4) *Mentha citrata oil* is distilled from the herb *Mentha citrata* Ehrh. It has a lavender-like odor and contains linalool and linalyl acetate as major components (together ca. 70 %). Recently, hybrids of the plant have been developed in the United States and other countries that yield oils with a total linalool content of over 80 % [287]. Currently, the main cultivation area is Eastern Asia. *Mentha citrata* oils are used mainly for perfuming soap and cosmetics.

5) *Pennyroyal oil* is obtained by steam distillation of the flowering tops of *Mentha pulegium* L. It is an almost colorless to amber-yellow liquid with a herbaceous, minty odor.
d_{20}^{20} 0.930 – 0.944; n_D^{20} 1.4800 – 1.4900; $[\alpha]_D^{20}$ + 15° to + 24°; solubility: 1 vol in 3 vol of 70 % ethanol; content of carbonyl compounds (calculated as pulegone) min. 80 % [288].
The main component of pennyroyal oil is (+)-pulegone [*89-82-7*] (ca. 80 %). About 50 t of the oil are produced per year, mainly in southern Spain and Tunisia.

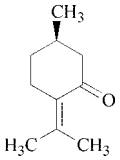

(+)-Pulegone

An oil of similar composition, also called pennyroyal oil, is obtained in North America from *Hedeoma pulegioides* L. Pennyroyal oil has lost its significance as a starting material for the production of (−)-menthol. However, it is still used in perfume compositions for soap and oral hygiene products.

Musk, see Animal Secretions, page 87.

Myrrh oil and myrrh resinoids are produced from the air-dried gum of *Commiphora myrrha* (Nees) Engl. var. *molmol* Engl., *C. abyssinica* (Berg) Engl. and *C. schimperi* (Berg) Engl. (Burseraceae), shrubs that grow in Northeast Africa and Arabia.

Myrrh resinoids are obtained in 20 – 45 % yield by extracting the gum with suitable solvents (e.g., benzene, hexane). They are waxlike, brown-yellow to red-brown masses with a balsamic odor and an aromatic bitter taste.

Acid value: 20 – 35; ester value: 165 – 200; carbonyl value: 20 – 75 [289].

Myrrh oil is obtained from the gum by steam distillation; it is a light brown or green liquid with the characteristic odor of the gum.

d_{25}^{25} 0.995 – 1.014; n_D^{20} 1.5190 – 1.5275; $[\alpha]_D$ − 83 ° to − 60 °; acid value: 2 – 13; saponification value: 9 – 35; solubility: 1 vol in 7 – 10 vol of 90 % ethanol; solutions are occasionally opalescent or turbid [290].

Typical aroma-determining compounds of the myrrh plant are furanosesquiterpenoids such as lindestrene [*2221-88-7*] [291].

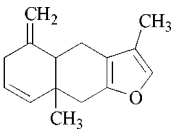

Lindestrene

Neroli oil and orange flower absolute are obtained from the blossoms of the bitter orange tree, *Citrus aurantium* L. subsp. *aurantium,* syn. *Citrus aurantium* L. subsp. *amara var. pumilia*, (Rutaceae), which is grown in France, Italy, and North Africa. *Neroli oil* is produced by steam distillation (about 4 t/a) and is a pale yellow to amber-colored liquid, with a slight blue fluorescence and a characteristic sweet, spicy-bitter odor of orange blossom.

d_{20}^{20} 0.864 – 0.876; n_D^{20} 1.460 – 1.474; $[\alpha]_D^{20}$ + 2 ° to + 11 °; solubility: 1 vol in 3.5 vol of 85 % ethanol at 20 °C; acid value: max. 2.0; ester value: 26 – 60 [292].

After separation of neroli oil, the aqueous layer of the steam distillate, known as orange blossom water, is extracted with suitable solvents (e.g., petroleum ether). Evaporation gives *orange flower water absolute* (absolue de l'eau de fleurs d'oranger), which is a dark brown-red liquid. It contains less terpene hydrocarbons and correspondingly more polar compounds than neroli oil.

Orange flower absolute is obtained from the blossoms by solvent extraction via the concrete. It is a dark brown liquid with a warm, spicy-bitter odor.

The main volatile constituent of all three products is linalool. Their typical flavor is created by a number of nitrogen-containing trace constituents, such as indole and derivatives of anthranilic acid.

Neroli oil and the related products are some of the most expensive natural raw materials and are produced only in small quantities (a few tons per year). They are used in fine fragrances; neroli oil, for example, is one of the classical components of eaux de cologne.

Nutmeg (mace) oil is obtained by steam distillation of the dried and crushed nuts and/or the coverings (mace) of nutmeg, which are the fruits of *Myristica fragrans* Houtt. (Myristicaceae). The tree grows in Indonesia and in the West Indies and becomes 15 – 20 m high. Nutmeg oil is a colorless to pale yellow liquid with a pleasant spicy odor. Its physical constants and odor vary with the place of origin. Specifications of Indonesian and (in parentheses) West Indian nutmeg oil are as follows:

d_{20}^{20} 0.885 – 0.907 (0.862 – 0.882); n_D^{20} 1.4750 – 1.4850 (1.4720 – 1.4760); $[\alpha]_D^{20}$ + 6 ° to + 18 ° (+ 25 ° to + 40 °); solubility (both oils): 1 vol in 5 vol of 90 % ethanol at 20 °C (solutions

sometimes opalescent), for freshly distilled oils, 1 vol in 3 – 4 vol [293].

Indonesian oils contain a higher percentage of higher boiling components than do West Indian oils (produced mainly in Grenada).

The oils contain ca. 90 % terpene hydrocarbons, mainly sabinene and α- and β-pinene. Major oxygen-containing constituents are 1-terpinen-4-ol and phenol ether derivatives (mainly myristicin).

Nutmeg oil is used mainly in food flavorings and to a lesser extent in perfumery. Worldwide annual production is 400 to 450 t, more than 120 t/a is produced in Indonesia alone.

Oakmoss absolute and tree moss absolute are obtained from tree lichens. *Oakmoss absolute* is derived from *Evernia prunastri* (L.) Arch. (Usneaceae), a lichen growing on oak trees. The lichen is first extracted with nonpolar solvents to give a concrete. The waxes are then removed by precipitation with ethanol, leaving an absolute.

The concretes are green to brown waxy pastes; the absolutes are more or less viscous liquids with an earthy, mossy, woody odor and a slight phenolic, leather note.

Resorcinol derivatives, e.g., orcinol [*504-15-4*], β-orcinol [*488-87-9*], their monomethyl ethers, and methyl 3-methylorsellinate [*4707-47-5*] are mainly responsible for the characteristic earthy-moss-like odor of the oakmoss products.

Tree moss concretes and absolutes are prepared from *Pseudevernia furfuracea*, a lichen growing on conifer bark. Their odors are different from those of the corresponding oakmoss products.

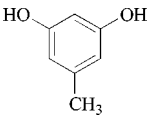

Orcinol

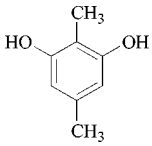

β-Orcinol

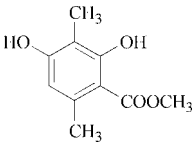

Methyl 3-methylorsellinate

Large quantities of oakmoss (mousse de chêne) and tree moss (mousse d'arbre) are collected annually in former Yugoslavia, France, and Morocco [294]. In 1997 1900 t of Pseudevernia furfuracea and 700 t of Evernia prunastri were processed, mainly in Grasse [295].

The extracts and absolutes are used in perfumery for nuances and as a fixative to give compositions a dry, sweet base note, e.g, in fougère and chypre perfumes. (IFRA recommendation).

Olibanum oil and olibanum resinoid are obtained from frankincense, which is a gum resin collected from the bark of the tree *Boswellia carteri* Birdw. (syn. *B. sacra* Flueck.) or *B. frereana* Birdw. (Burseraceae) growing in Yemen and Somalia. The resinoid is produced by solvent extraction, and steam distillation gives the oil, which is a pale yellow, slightly viscous liquid with a balsamic odor and a faint lemon note.

d_{25}^{25} 0.862 – 0.889; n_D^{20} 1.4650 – 1.4820; $[\alpha]_D$ $-15°$ to $+35°$; solubility: 1 vol in 6 vol of 90 % ethanol, solutions occasionally opalescent; acid value: max. 4.0; ester value: 4 – 40 [296].

Various qualities are commercially available. Their compositions may vary considerably because they are prepared from the resins of different *Boswellia* species [297].

Olibanum oil and resinoid are used in oriental type perfumes for their odor and their good fixative properties.

Opopanax oil and opopanax resinoid (also: Opoponax) are obtained from the resin of *Commiphora erythraea* Engl. var. *glabrescens* Engl., a tree growing in Somalia (Burseraceae). The resinoid is prepared by solvent extraction, and steam distillation of the resin gives the essential oil, which is a yellow to greenish-yellow liquid with a warm, sweet, balsamic odor.

d_{25}^{25} 0.865 – 0.932; n_D^{20} 1.488 – 1.504; $[\alpha]_D$ $-32°$ to $-9°$; acid value: max. 4; saponification value: 4 – 20; solubility: 1 vol in 10 vol of 90 % ethanol; solutions are occasionally turbid [298].

Opopanax oil and resinoid are used in perfume compositions with oriental characteristics. (IFRA recommendation).

Orange flower absolute, see Neroli Oil, page 109.

Orange oils, see Citrus Oils, page 94.

Origanum oils are produced from several species of the flowering herb *Origanum* (Lamiaceae). Spanish origanum oil is derived from *Coridothymus capitatus* Rchb. and is often called white thyme oil. The oils differ mainly in their content of carvacrol and thymol, which are major constituents. Oil with a high carvacrol content [*499-75-2*] (for formula, see page 116) is a yellowish-red to dark brown liquid with a spicy, herbaceous odor, reminiscent of thyme. The color quickly turns to black when in contact with iron.

d_{25}^{25} 0.935 – 0.960; n_D^{20} 1.5020 – 1.5080; $[\alpha]_D$ – 2 ° to + 3 °; phenol content: 60 – 75 %; solubility: 1 vol in 2 vol of 70 % ethanol; solutions may become turbid when further diluted [299].

According to ISO the Spanish origanum oil is derived from *Thymbra capitata* (L.) Cav. It is a yellowish to dark brown, almost black oil with an aromatic, phenolic, and spicy odor; d_{20}^{20} 0.930 – 0.955; n_D^{20}1.5000 – 1.5130; $[\alpha]_D^{20}$ – 5 ° to + 2 °; solubility: 1 vol in 4 vol of 70 % ethanol at 20 °C; carvacrol content by GLC: 60 to 75 % [300]. (Both botanical names – *Coridothymus* and *Thymbra* – are not mentioned in the Zander dictionary [196].)

Origanum oils are used in perfume compositions and in seasoning mixtures.

Orris root oil is obtained by steam distillation of the rhizomes of the two species of iris, *Iris pallida* Lam. or *Iris germancia* L. (Iridaceae). Prior to distillation the rhizomes are stored for about two years in order to develop the aroma and are then ground. The steam distillate is a light yellow to brown-yellow solid mass with a violet-like odor. The solid liquefies to a yellow to yellow-brown liquid at 38 – 50 °C.

Acid value: 175 – 235, corresponding to an acid content of 71 – 95 % (calculated as myristic acid); ester value: 4 – 35; ketone content (calculated as irone): 9 – 20 %; soluble with ethanol in all proportions at 50 °C [301].

Because of its high content of myristic and other fatty acids, the steam distillate is a waxy mass that resembles a concrete and is sold under this name or as *Beurre d'Iris* (Orris butter). In addition, a neutral product, obtained after removal of the acids with alkali, is marketed under the name *orris oil absolute* or as *orris oil 10-fold.*

The main constituents of the oil are (–)-*cis*-γ-irone [*89888-04-0*] and (–)-*cis*-α-irone [*89888-03-9*], which are responsible for its odor [302].

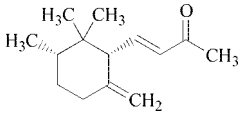

(–)-*cis*-γ-Irone

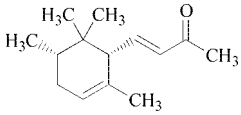

(–)-*cis*-α-Irone

The oil, respectively the butter, is very expensive and is used in perfumery and in flavor compositions.

Palmarosa oil, see Cymbopogon Oils, page 98.

Parsley oils are produced from the plant *Petroselinum crispum* (Mill.) Nyman. ex A.W. Hill (P. *sativum* Hoffm.), (Apiaceae):

1) *Parsley leaf oil* is produced by steam distillation of the aerial parts of the plant, including the immature seeds. It is a yellow to light-brown liquid with the typical odor of the herb. d_{25}^{25} 0.908 – 0.940; n_D^{20} 1.5030 – 1.5300; $[\alpha]_D$ – 9 ° to + 1 °; acid value: max. 2; soluble in 95 % ethanol; solutions may be opalescent [303].
   The oil consists mainly of monoterpene hydrocarbons. Its main constituent is 1,3,8-menthatriene [*18368-95-1*], which – together with two pyrazines and (*Z*)-dec-6-enal – is important for the aroma of parsley leaves [304].
2) *Parsley seed oil* is obtained by steam distillation of ripe fruits of parsley. It is an almost colorless to amber-yellow liquid whose dry odor is characteristic of the crushed fruit, but different from that of the green parts of the plant.
   d_{20}^{20} 1.043 – 1.083; n_D^{20} 1.5130 – 1.5220; $[\alpha]_D^{20}$ – 10 ° to – 4 °; acid value: max. 6; ester value: min. 1, max. 10; solubility: 1 vol in 6 vol of 85 % ethanol, solutions are sometimes slightly turbid [306].
   Characteristic components of parsley seed oil are myristicin [*607-91-0*], apiol [*523-80-8*], and 2,3,4,5-tetramethoxyallylbenzene [*15361-99-6*] [305].

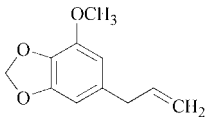
Myristicin

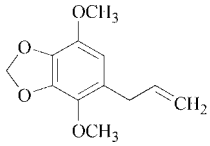
Apiol

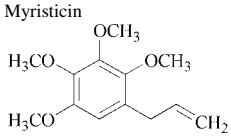
2,3,4,5-Tetramethoxyallylbenzene

Parsley oils are used in the food industry in seasonings, for example, for meat and sauces.

Patchouli oil is produced by steam distillation of the dried leaves of *Pogostemon cablin* (Blanco) Benth. (Lamiaceae). It is a yellow to reddish-brown, more or less viscous liquid with a characteristic, slightly camphoraceous, woody balsamic odor.

d_{20}^{20} 0.952 – 0.975; n_D^{20} 1.5050 – 1.5150; $[\alpha]_D^{20}$ – 60 ° to – 40 ° (in exceptional cases to – 32 °); solubility: 1 vol in 10 vol of 90 % ethanol at 20 °C; acid value: max. 4.0; ester value: max. 10 [307].

The patchouli shrub is cultivated primarily in Indonesia. It grows to 1 m, but during harvesting is cut back to 10 – 15 cm. The oil yield from preferably sun-dried leaves is ca. 2 %.

Although the sesquiterpene alcohol (–)-patchoulol [*5986-55-0*] is the main component of patchouli oil (27 – 35 %), this compound contributes less to the characteristic odor of the oil than norpatchoulenol [*41429-52-1*] present only at a concentration of 0.2 – 1 % [308].

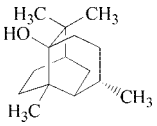
Patchoulol

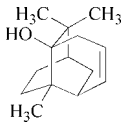
Norpatchoulenol

Worldwide production lies between 900 and 1000 t/a. Patchouli oil is very tenacious and is used in perfumery for oriental and fougère notes.

Pennyroyal oil, see Mint Oils, page 107.

Pepper oil and pepper oleoresin are obtained from the black pepper *Piper nigrum* L. (Piperaceae). Pepper oil is produced by steam distillation of whole or broken unripe fruits. The oil is an almost colorless to bluish-green liquid with a characteristic odor, reminiscent of pepper, but without the pungency of the spice.

d_{20}^{20} 0.870 – 0.890; n_D^{20} 1.480 – 1.492; $[\alpha]_D^{20}$ – 16 ° to + 4 °; solubility: 1 vol in 3 vol of 95 % ethanol; ester value: max. 11 [309].

The oil contains mainly monoterpene hydrocarbons (ca. 80 %), its main constituent being sabinene [*3387-41-5*] (20 – 25 %).

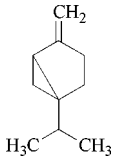
Sabinene

Pepper oleoresin is produced by solvent extraction and, unlike the oil, contains the principal pungent compound, piperine [*94-62-2*], along with some essential oil.

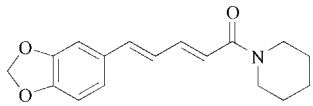
Piperine

Pepper products are used for flavoring foods; pepper oil is also employed to a small extent in perfume compositions.

Peppermint oil, see Mint Oils, page 107.

Peru balsam oil is obtained by distillation of the balsam of *Myroxylon pereirae* (Royle) Klotzsch (Fabaceae). It is a yellow to pale brown, slightly viscous liquid, which may form crystals with a rather sweet, balsamic odor.

d_{25}^{25} 1.095 – 1.110; n_D^{20} 1.5670 – 1.5790; $[\alpha]_D$ – 1 ° to + 2 °; acid value: 30 – 60; ester value: 200 – 225; solubility: 1 vol in at least 0.5 vol of 90 % ethanol [310].

Peru balsam is produced almost exclusively in El Salvador. The oil yield from the balsam is ca. 50 %. Major components are benzyl benzoate and benzyl cinnamate. Nerolidol and vanillin contribute to the odor.

The oil is used for its excellent fixative properties in perfumes for soap and cosmetics. Use of Peru balsam itself is now banned because it is allergenic. (IFRA recommendation)

Petitgrain oils are obtained by steam distillation of the leaves of citrus trees (Rutaceae family). The oils derived from the bitter orange tree are the most important. Other petitgrain oils (mandarinier, citronnier, and bergamotier) are less important. Petitgrain oil mandarinier is a source of natural methyl *N*-methylanthranilate, which is present at a concentration of more than 50 %. Petitgrain oils are essential constituents of eaux de cologne.

Petitgrain oil Paraguay is obtained from an acclimatized variety of the bitter orange tree, *Citrus aurantium* L. subsp. *aurantium,* which is grown in Paraguay. Leaves and twigs are steam distilled. Petitgrain oil Paraguay is a colorless to pale yellow liquid with a characteristic, strong, pleasant odor, reminiscent of linalool and linalyl acetate.

d_{20}^{20} 0.882 – 0.893; n_D^{20} 1.455 – 1.463; $[\alpha]_D^{20}$ $-6°$ to $-1°$; solubility: 1 vol in 4 vol of 70 % ethanol at 20 °C; acid value: max. 1; ester value: 127 – 184, corresponding to an ester content of 45 – 65 % (calculated as linalyl acetate) [311].

Annual production is ca. 250 – 280 t.

Petitgrain oil bigarade is obtained by steam distillation of the leaves, twigs, and small green fruits of the bitter orange tree *Citrus aurantium* L. subsp. *aurantium,* grown in France, Italy, and North Africa. It is a pale yellow to amber yellow liquid, with a slight blue fluorescence and with a characteristic ethereal and pleasant odor.

d_{20}^{20} 0.886 – 0.898; n_D^{20} 1.4560 – 1.4720; $[\alpha]_D^{20}$ $-6°$ to $+1°$; solubility: 1 vol in 5 vol of 70 % ethanol at 20 °C; acid value: max. 2; ester value: 140 – 217, corresponding to an ester content of 50 – 77 % (calculated as linalyl acetate) [312].

Annual production is ca. 3 to 5 t.

Pimento oils (allspice oil) are derived from pimento fruits and leaves. Pimento berry oil is obtained by steam distillation of the dried, fully grown, unripe, berry-like fruits of the pimento shrub, *Pimenta dioica* (L.) Merrill. (Myrtaceae), growing in Jamaica. It is a pale yellow to brown liquid with a spicy odor, reminiscent of eugenol.

d_{20}^{20} 1.027 – 1.048; n_D^{20} 1.5250 – 1.5400; $[\alpha]_D^{20}$ $-5°$ to $0°$; solubility: 1 vol in 2 vol of 70 % ethanol at 20 °C; phenol content: min. 65 % [313].

The major components of pimento berry oil are eugenol (up to 75 %), 1,8-cineole, and caryophyllene. Distillation of the leaves gives an oil that has an even higher content of eugenol (80 – 90 %). Annual production of leaf oil is ca. 50 t, which exceeds that of the berry oil.

Pimento oils, like the spice itself, are used mainly in the food industry, as well as in perfume compositions for creating spicy, clovelike notes.

Pinaceae needle oils from Pinaceae species contain (−)-bornyl acetate as their main odoriferous component. Other main constituents are monoterpene hydrocarbons such as α- and β-pinene, limonene, 3-carene, and α- and β-phellandrene. The oils are used in perfumes for soap, bath products, and air fresheners and in pharmaceutical preparations.

1) *European silver fir oil* is produced in central Europe by steam distillation of needle-bearing twigs of *Abies alba* Mill. It is a colorless to pale yellow liquid with a pleasant odor of freshly cut fir needles.
   d_{25}^{25} 0.867 – 0.878; n_D^{20} 1.4700 – 1.4750; $[\alpha]_D$ $-67°$ to $-34°$; solubility: 1 vol in 7 vol of 90 % ethanol; ester content (calculated as bornyl acetate): 4 – 10 % [314].
   Silver fir cone oil (templin oil) obtained from the cones of *Abies alba* Mill. has similar properties.
2) *Siberian fir needle oil* is obtained by steam distillation of needles and twigs of *Abies sibirica* Ledeb., the Siberian silver fir. It is an almost colorless or pale yellow liquid with a characteristic fir odor.
   d^{25} 0.898 – 0.912; n_D^{20} 1.4685 – 1.4730; $[\alpha]_D$ $-45°$ to $-33°$; solubility: 1 vol in 1 vol of 90 % ethanol, solutions may become hazy when further diluted; ester content (calculated as bornyl acetate): 32 – 44 % [315].
3) *Canadian fir needle oil* (balsam fir oil) is produced in Canada and in several northern states of the United States from needles and twigs of *Abies balsamea* (L.) Mill. It is an almost colorless to pale yellow liquid with a pleasant balsamic odor.
   d_{25}^{25} 0.872 – 0.878; n_D^{20} 1.4730 – 1.4760; $[\alpha]_D$ $-24°$ to $-19°$; solubility: 1 vol in 4 vol of 90 % ethanol (turbidity may occur), ester content (calculated as bornyl acetate): 8 – 16 % [316].
4) *Pine needle oil* is produced primarily in Austria (Tyrol), former Yugoslavia, and the CIS by steam distillation of the needles of the

Norwegian or Scotch pine, *Pinus sylvestris* L. It is a colorless or yellowish liquid with an aromatic, turpentine-like odor.
d_{25}^{25} 0.857 – 0.885; n_D^{20} 1.4730 – 1.4785; $[\alpha]_D$ $-4°$ to $+10°$; solubility: 1 vol in 6 vol of 90 % ethanol, sometimes slightly opalescent; ester content (calculated as bornyl acetate): 1.5 – 5 % [317].

5) *Dwarf pine needle oil* is obtained in Austria (Tyrol) and other mountainous areas of central and southeast Europe by steam distillation of fresh needles and twig tips of *Pinus mugo* Turra subsp. *mugo* Zenari and subsp. *pumilio* (Haenke) Franco. It is a colorless liquid with a pleasant, balsamic odor.
d_{25}^{25} 0.853 – 0.871; n_D^{20} 1.475 – 1.480; $[\alpha]_D^{25}$ $-16°$ to $-3°$; solubility: 1 vol in 10 vol of 90 % ethanol; bornyl acetate content is similar to that of pine needle oil.
6) *Spruce and hemlock oils* are produced in Canada and the Northeast of the United States by steam distillation of needles and twigs from *Picea mariana* (Mill.) B.S.P. (black spruce), *Picea glauca* (Moench) Voss (white spruce), *Tsuga canadensis* (L.) Carrière (Eastern Hemlock), and related species. They are very pale to light yellow liquids with a pleasant odor reminiscent of pine needles.
d_{25}^{25} 0.900 – 0.915; n_D^{20} 1.4670 – 1.4720; $[\alpha]_D$ $-25°$ to $-10°$; solubility: 1 vol in at least 1 vol of 90 % ethanol; solutions may become hazy when > 2 vol of 90 % ethanol are added; ester content (calculated as bornyl acetate): 37 – 45 % [318].

Rose oil and rose absolute are used mainly in fine fragrances. Rose oil is also used in small amounts for flavoring purposes.

1) *Rose oil* is obtained by steam distillation of blossoms of *Rosa* x *damascena* Mill. (Rosaceae) that is mainly cultivated in Turkey, Bulgaria (Kazanlik rose), and Morocco. Since a reasonable amount of rose oil is still dissolved in the aqueous phase after steam distillation, the distillation water (*rose water*) is redistilled or extracted.
Rose oil is a yellow to greenish-yellow liquid with the characteristic odor of rose blossoms and distinct tea and honey notes.
d_{20}^{20} 0.848 – 0.880; n_D^{25} 1.4520 – 1.4700; $[\alpha]_D^{20}$ $-5°$ to $-1.8°$; *fp* between 16 and 23.5 °C; ester value: min. 7, max. 24. Due to the presence of paraffin hydrocarbons (stearoptenes) the determination of the physical constants at 20 °C can be difficult or impossible because of a low freezing-point [319].
The major constituents of rose oil are (−)-citronellol, geraniol, and nerol. In contrast to the absolute (see below), the oil contains only a minor amount of phenethyl alcohol, which is extremely soluble in water.
In addition to the major components mentioned above, Bulgarian rose oil also contains a number of components which, although present in low concentrations, contribute to the characteristic fragrance. Among these are β-damascenone (see page 39) and rose oxide (see page 72) [320].
Since ca. 3 t of blossoms are required to prepare ca. 1 kg of oil, rose oil is one of the most expensive essential oils. Production is limited and is only about 30 t per year.
2) *Rose absolute* is prepared from the concrete extracted from *Rosa* x *damascena* in the countries mentioned above, as well as from *Rosa centifolia* L. (rose de mai) in Morocco and the south of France.
The absolute is a reddish liquid with a typical rose odor. The phenethyl alcohol content of its volatile fraction is 60 – 75 % [321].

Rosemary oil is obtained by steam distillation of the twigs and flowering tops of *Rosmarinus officinalis* L. (Lamiaceae). It is an almost colorless to pale yellow or yellow-greenish liquid with a characteristic, refreshing, pleasant odor.

For oils type *Spain*: d_{20}^{20} 0.892 – 0.910; n_D^{20} 1.4640 – 1.4760; $[\alpha]_D^{20}$ $-5°$ to $+8°$; ester content (calculated as bornyl acetate): 2 – 5 %; acid value: 1.0; solubility: 1 vol in 3 vol of 90 % ethanol.

For oils type *Tunisia and Morocco*: d_{20}^{20} 0.907 – 0.920; n_D^{20} 1.4640 – 1.470; $[\alpha]_D^{20}$ $-2°$ to $+5°$; ester content (calculated as bornyl acetate): 2 – 5 %; acid value: 1.0; solubility: 1 vol in 2 vol of 80 % ethanol [322]

Major components are 1,8-cineole, α-pinene, and camphor, which comprise up to 55 %, 30 % and 22 %, respectively of commercial oils. The

composition of the oil varies with its origin. North African oils (e.g., from Tunisia) are comparatively high in cineole. Spanish oils have a lower cineole content, but a higher camphor (up to 22 %) and monoterpene hydrocarbon content (30 – 40 %). Verbenone [*18309-32-5*] is also important for the odor; it may occur in traces or as a major component; its content in the Spanish oils is higher than in the North African ones [323].

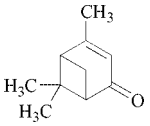

Verbone

The main producers of rosemary oil are in Tunisia and Spain. About 100 to 150 t are produced annually. Rosemary is used widely in perfumery and in large amounts for perfuming bath foams, shampoos, and hair tonics.

Rosewood oil, Brazilian (Bois de rose oil) is obtained by steam distillation of wood from *Aniba rosaeodora* Ducke var. *amazonica* Ducke and/or *A. parviflora* (Meissner) Mez (Lauraceae). It is an almost colorless to pale yellow liquid with a characteristic, sweet odor, reminiscent of linalool.

d_{20}^{20} 0.870 – 0.887; n_D^{20} 1.4620 – 1.4690; $[\alpha]_D^{20}$ $-2°50'$ to $+4°$; solubility: 1 vol in 9 vol of 60 % ethanol at 20 °C; acid value: max. 1; ester value: max. 5; ester value (after acetylation): 247 – 280, corresponding to an alcohol content of 82 – 96 % (calculated as linalool) [324].

Brazilian rosewood oil is no longer competitive as a raw material for linalool. Increasing production costs, as well as the development of large-scale processes for fully synthetic linalool (for production of vitamin A), have led to a sharp decline in production (annual production ca. 20 – 30 t). Currently, rosewood oil is only used in small amounts in perfumes, e.g., for soaps.

Sage oils are of three main types:

1) *Clary sage oil* is obtained by steam distillation of flowering tops and foliage of cultivated *Salvia sclarea* L. (Lamiaceae). The quality *Broyée en vert* is obtained after cutting the fresh material and immediate distillation. Another quality named *Traditionelle* exists, which is obtained by steam distillation after drying of the tops and foliage on the field. Both qualities are pale yellow to yellow liquids with a herbaceous odor and a winelike bouquet.
Physical constants for the quality *Broyée en vert*: d_{20}^{20} 0.890 – 0.902; n_D^{20} 1.4560 – 1.4660; $[\alpha]_D^{20}$ $-20°$ to $-10°$; solubility: 1 vol in 3 vol of 80 % ethanol; solutions become opalescent on dilution; acid value: max. 1; ester content (calculated as linalyl acetate): 63 – 75 % [325].
In addition to linalyl acetate, the oil contains linalool and other terpene alcohols, as well as their acetates. When the volatile components are evaporated, a distinct ambergris note develops that is attributed to oxidative degradation products of sclareol [*515-03-7*]. Sclareol is the main component in the concrete, obtained by solvent extraction of *S. sclarea* L. leaves.

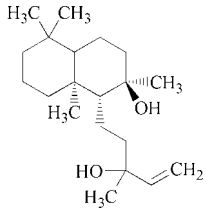

Sclareol

Sclareol is used as a starting material for a number of ambergris fragrances. Major cultivation areas for *S. sclarea* L. are the CIS, the Mediterranean and Balkan countries, and more recently the United States. The annual production is between 50 and 150 t.
Clary sage oil is used extensively in fine fragrances.

2) *Dalmatian sage oil* is obtained by steam distillation of the partially dried leaves of *Salvia officinalis* L. (Lamiaceae). It is a colorless to yellow liquid with a warm camphoraceous, thujone-like odor and sharp and bitter taste.
d_{20}^{20} 0.910 – 0.930; n_D^{20} 1.4580 – 1.4740; $[\alpha]_D$ $+2°$ to $+30°$; saponification value: 6 – 20; ketone content (calculated as thujone): 28 – 78 %; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C [326].
The major components of *S. officinalis* oil are camphor and α-thujone [*546-80-5*]. β-

thujone [*471-15-8*] is present in ca. 5 % concentration.

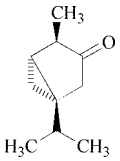

α-Thujone [(−)-thujone]

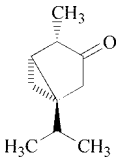

β-Thujone [(+)-isothujone]

The oil is used in pharmaceutical preparations and in perfumery to create dry, spicy-herbaceous notes.

3) *Spanish sage oil* is produced by steam distillation of leaves and twigs of *Salvia lavandulifolia* Vahl (Lamiaceae) growing in Spain in the provinces Granada, Murcia, Almeria, and Jaén. The oil is an almost colorless to pale yellow liquid with the characteristic camphoraceous odor of the leaves.
d_{20}^{20} 0.907 – 0.930; n_D^{20} 1.465 – 1.472; $[\alpha]_D^{20}$ − 12 ° to + 17 °; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C; acid value: max. 2.0; ester value: min. 15, max. 55 [327].
Unlike Dalmatian sage oil, Spanish sage oil contains no or only traces of α- or β-thujone; camphor and 1,8-cineole are the major components and are responsible for the odor of the oil, which is used primarily in pharmaceutical preparations.
Altogether about 50 t of Spanish and Dalmatian oils are produced annually.

Sandalwood oil, East Indian is obtained by steam distillation of the heartwood of *Santalum album* L. (Santalaceae). It is a slightly viscous, almost colorless to golden yellow liquid with a characteristic, sweet, woody, long-lasting odor.

d_{20}^{20} 0.968 – 0.983; n_D^{20} 1.5030 – 1.5080; $[\alpha]_D^{20}$ − 21 ° to − 12 °; solubility: 1 vol in 5 vol of 70 % ethanol at 20 °C; ester value: min. 10; free alcohol content (calculated as santalol): min. 90 %.

East Indian sandalwood oil consists almost exclusively of closely related sesquiterpenoids; the alcohols *cis*-α-santalol [*115-71-9*] (45 – 55 %) and *cis*-β-santalol [*77-42-9*] (18 – 24 %) are mainly responsible for its odor [328].

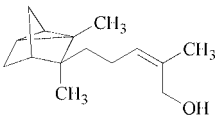

α-Santalol

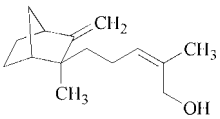

β-Santalol

The trees cultivated for oil production must be at least 30 years old to make oil production profitable. The oil yield, calculated on the amount of wood used for steam distillation, is 4 – 6.5 %. Indian production is ca. 60 t/a. Smaller quantities are produced in neighboring countries.

East Indian sandalwood oils are used extensively in perfumery as very valuable, stable fixatives.

Sassafras oil, Brazilian is obtained by steam distillation of the roots, trunks, and branches of *Ocotea pretiosa* (Nees) Mez (Lauraceae). The oil was formerly called *Ocotea cymbarum* oil due to incorrect botanical naming. It is a yellow to brownish liquid with the characteristic odor of safrole.

d_{20}^{20} 1.082 – 1.094; n_D^{20} 1.5330 – 1.5370; $[\alpha]_D^{20}$ − 0.5 ° to + 3.0 °; solubility: 1 vol in 2 vol of 90 % ethanol; *fp* + 7.5 to + 9.5 °C [329].

The main component of the oil is safrole, which may make up more than 90 % of the oil and determines its freezing point. (IFRA recommendation)

Production from trees that grow wild in Brazil, Paraguay, and Columbia exceeds that of North American sassafras oil, which is similar in composition. The main use of the oil is for the production of safrole and products derived from it (e.g., isosafrole and piperonal, see page 69).

Chinese sassafras oils (annual production about 950 t) are fractions, rich in safrole, that are obtained from oils of different species of the camphor tree (see Camphor Oils, page 90).

Savory oil is obtained by steam distillation of the whole dried herb *Satureja hortensis* L. (Summer savory) or *S. montana* L. (Lamiaceae) (Winter savory). It is a light yellow to dark brown liquid with a spicy odor, reminiscent of thyme or origanum.

d_{25}^{25} 0.875 – 0.954; n_D^{20} 1.4860 – 1.5050; $[\alpha]_D$ − 5 ° to + 4 °; phenol content: 20 – 59 %; saponification value: max. 6; solubility: 1 vol in 2 vol of 80 % ethanol at 25 °C; solutions in 10 vol of ethanol may be slightly hazy [330].

Several qualities of savory oil exist, depending on the *Satureja*species used. The main cultivation areas are France, Spain, some of the Balkan countries, and some midwestern states of the United States. Characteristic of savory oil is its high content of carvacrol [*499-75-2*].

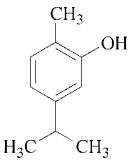

Carvacrol

Savory oil is used mainly in the food industry, e.g., for flavoring sauces and soups.

Spearmint oil, see Mint Oils, page 107.

Spike oil, see Lavandula Products, page 105

Star anise oil is obtained by steam distillation of the star-shaped fruits of *Illicium verum* Hook. f. (Illiciaceae), an evergreen tree growing in Vietnam and China. It is a colorless to pale yellow liquid, which solidifies on cooling.

d_{20}^{20} 0.979 – 0.985; n_D^{20} 1.5530 – 1.5560; $[\alpha]_D^{20}$ $-2°$ to $+2°$; solubility: 1 vol in 3 vol of 90 % ethanol; *fp* min. 15 °C [331].

The main component of star anise oil, as in anise oil, is *trans*-anethole (86 – 93 %, determined by GLC). Pure anethole can be obtained by rectification (see page 64). Star anise oil has replaced true anise oil derived from *Pimpinella anisum* in the production of natural anethole.

Star anise oil and its product, anethole, are used primarily in the alcoholic beverage industry (anise liquors), but also for flavoring food and toothpaste.

Styrax oil and styrax resinoid are obtained from styrax balsam, which is collected from the artificially injured trees, *Liquidamber orientalis* Mill. (Asia Minor) and *L. styraciflua* L. (Central America) (Hamamelidaceae).

Styrax resinoid is a light grey to dark grey-brown viscous liquid that is prepared by solvent extraction. It has a sweet balsamic, slightly grasslike odor and is used in perfumery as a fixative.

Steam distillation of the balsam gives *styrax oil*, a light yellow liquid that contains a relatively large amount of cinnamic acid, which partially crystallizes. Styrax oil has a sweet-balsamic odor with a styrene-like top note. It is used in perfumery in flowery compositions. (IFRA recommendation).

Tagetes oil (Marigold oil) is produced by steam distillation of the flowering plants *Tagetes minuta* L. (*T. glandulifera* Schrank., Asteraceae). The main producers are in France, South Africa, and India. The oil is a yellow to dark orange liquid with a strong, aromatic-fruity odor. Its main components are *cis*-ocimene, dihydrotagetone, tagetone, and *cis*- and *trans*-ocimenone.

It is used in accentuated flowery fragrances and in aroma compositions to achieve fruity effects. (IFRA recommendation)

Tarragon oil (estragon oil) is produced by steam distillation of leaves, stems, and flowers of *Artemisia dracunculus* L. (Asteraceae). It is a colorless to pale yellow liquid with a characteristic, spicy, delicate estragon odor reminiscent of liquorice and sweet basil.

d_{20}^{20} 0.918 – 0.943; n_D^{20} 1.5080 – 1.5180; $[\alpha]_D^{20}$ $+2°$ to $+6°$; acid value: max. 1; ester value: max. 18; solubility: 1 vol in 4 vol of 90 % ethanol [332].

The main component of tarragon oil is estragole (methylchavicol, see page 89), which may make up > 60 % of the oil.

Worldwide production of tarragon oil only amounts to a few tons per year. It is used mainly in aroma compositions, smaller quantities are employed in perfumery.

Tea tree oil is obtained by steam distillation of leaves and twig tips of *Melaleuca alternifolia* Cheel, *M. linariifolia* Sm. and *M. dissitiflora* F. Mueller, and other *Melaleuca* species, which are trees growing in Australia (Myrtaceae). It is a colorless to pale yellow liquid with an earthy, spicy odor.

d_{20}^{20} 0.888 – 0.906; n_D^{20} 1.4750 – 1.4800; $[\alpha]_D^{20}$ $+5°$ to $+15°$; cineole content: max. 15 % (by GLC); terpinen-4-ol content: min. 30 % (by GLC); solubility: 1 vol in 2 vol of 85 % ethanol [333].

Tea tree oil is used in perfumery for creating certain nuances and for earthy notes. In addition there is an increasing interest for using this oil in aroma therapy. Annual production is ca. 150 t.

Thuja oil, see Cedar Leaf Oil, page 92.

Thyme oil is obtained by steam distillation of flowering plants of *Thymus zygis* L. var. *gracilis* Boissier (Lamiaceae). Thyme oil is not to be confused with white thyme oil; see under Origanum Oil. Thyme oil is a red or reddish liquid with a strong, characteristic, spicy-phenolic odor and a sharp, lasting taste.

d_{20}^{20} 0.910 – 0.937; n_D^{20} 1.4940 – 1.5040; $[\alpha]_D^{20}$ $-6°$ to $-1°$; total phenol content: min. 38 %, max. 56 %; solubility: 1 vol in 3 vol of 80 % ethanol [334].

The main constituents of thyme oil are thymol and carvacrol (up to 60 %); for structures of these compounds, see p. page 63, page 116. Other chemotypes of *T. vulgaris* are limited to specific areas and yield, e.g., oils that contain geraniol, linalool, α-terpineol, and cineole; these oils are of minor importance.

Thyme shrubs grow in France, Spain, Algeria, and Morocco. Spain is the main producer of the oil. Thyme oil is used mainly for flavoring foods and oral hygiene products, but is also used in perfumery to create spicy, leathery notes. Because of its high phenol content, thyme oil has germicidal and antiseptic properties.

Tolu balsam resinoid is produced by extraction of the balsam of *Myroxylon balsamum* (L.) Harms (Fabaceae). It is a dark orange brown mass with a sweet, resinous, long-lasting odor, reminiscent of hyacinth. An essential oil is also distilled from the balsam.

Tolu balsam resinoid contains a fairly large amount of benzyl and cinnamyl esters of benzoic and cinnamic acid.

Both the resinoid and the balsam are used in perfumery, mainly for their fixative properties.

Tonka bean absolute is prepared by solvent extraction either directly from so-called Tonka beans or via the concrete. Tonka beans are the seeds of fruits from *Dipteryx odorata* (Aubl.) Willd. (Fabaceae).

Tonka bean absolute is a solid or crystalline mass with a sweet, caramel-like odor, reminiscent of coumarin. In perfumery tonka bean absolute is used as a fixative and produces a dry sweetness in men's fragrances.

Tree moss absolute, see Oakmoss Absolute, page 110.

Tuberose absolute is obtained by solvent extraction, via the concrete, from the blossoms of *Polianthes tuberosa* L. (Agavaceae). The plants are cultivated in Morocco, Egypt, and India. Tuberose absolute was formerly produced in southern France by enfleurage. It is an orange to brown liquid with a sweet-narcotic blossom odor and is used in modern, flowery perfume compositions.

Turpentine oils (→ Turpentines) are used in large quantities by the fragrance industry as starting materials for the manufacture of single fragrance and flavor compounds. Turpentine oils are obtained either from balsams or from the wood of various Pinaceae species. They are less valuable as odor materials than the oils obtained from needles, young twigs, and cones (see Pinaceae Needle Oils, page 113). Nevertheless, they are used either as such or indirectly as fragrant solvents for perfuming household products. IFRA recommends that these oils should only be used when the level of peroxides is kept to the lowest practical level.

Balsam turpentine oil is obtained from the resins of living trees of suitable *Pinus* species by distillation at atmospheric pressure and temperatures up to 180 °C, or by other fractionation methods, which do not change the terpene composition of the resins. *Wood turpentine oils,* on the other hand, are generally obtained by steam distillation of chopped tree trunks, dead wood, or of resin extracted from this wood. *Sulfate turpentine oil* is produced as waste in the manufacture of cellulose by the sulfate process and is also a wood turpentine. Pine oil is another wood turpentine oil that is obtained by dry distillation of suitable pine and fir trees, followed by fractionation.

Balsam and wood turpentine oils are colorless liquids with a mild, characteristic odor. Oils obtained by dry distillation often also have a phenolic note. The specifications of turpentine oils are listed in Table 5 [335].

Worldwide production of turpentine oils is ca. 300 000 t/a, of which approximately one third is produced in the United States and Canada as sulfate turpentine oil. Other major producers are the CIS, Scandinavia, and China.

The main components of all turpentine oils are terpene hydrocarbons. The balsam and sulfate turpentine oils produced in the South-

Table 5. Specifications of turpentine oils

Parameter	Type of turpentine oil			
	Balsam	Wood, steam distilled	Sulfate	Wood, dry distilled
d_4^{20}	0.855 – 0.870	0.855 – 0.870	0.860 – 0.870	0.850 – 0.865
n_D^{20}	1.465 – 1.478	1.465 – 1.478	1.465 – 1.478	1.463 – 1.483
Distillate up to *bp* 170 °C, %	90	90	90	60
Evaporation residue, %	2.5	2.5	2.5	2.5
Acid value	1	1	1	1

east of the United States contain mainly α- and β-pinene (60 – 70 % and 20 – 25 %, respectively); *Pinus palustris* Mill. (long leaf pine) gives the (+)-enantiomers and *Pinus caribaea* Morelet (slash pine) gives the (−)-enantiomers. The remaining constituents are *p*-menthadienes, oxygen-containing terpenoids, and anethole. In contrast with the Scandinavian and Russian turpentine oils, the American oils contain very little 3-carene or camphene. Large amounts of the American oils are separated by fractional distillation into their components, which are used as starting materials in the manufacture of fragrance and flavor compounds. European turpentine oils originate mainly from Portugal, where distillation of *Pinus pinaster* Aiton resin or wood gives a turpentine oil that consists of 70 – 85 % (−)-α-pinene and 10 – 20 % (−)-β-pinene [336].

Valerian oil is obtained by steam distillation of well-dried ground roots of *Valeriana officinalis* L. (Valerianaceae). It is a yellow-green to yellow-brown liquid with an extremely strong, characteristic, penetrating odor. The oil becomes dark and viscous on aging or on exposure to air.

d_{25}^{25} 0.942 – 0.984; n_D^{20} 1.4860 – 1.5025; $[\alpha]_D$ − 28 ° to − 2 °; acid value: 5 – 50; saponification value: 30 – 107; solubility: 1 vol in 0.5 – 2.5 vol of 90 % ethanol at 25 °C, solutions are clear to opalescent when up to 10 vol of ethanol is added [337].

The main component of the oil is (−)-bornyl acetate, but it also contains other bornyl esters (e.g., bornyl isovalerate), terpene and sesquiterpene hydrocarbons, as well as free isovaleric acid, which contributes strongly to the odor of the oil.

Valerian is cultivated in Belgium, northern France, east Germany, Poland, Hungary, and southern Russia. The so-called Kesso roots obtained from *Valeriana officinalis* L. var. *angustifolia* Miq., cultivated in Japan, give a much larger yield of oil.

Valerian oil is produced in limited quantities and is used in flavor and fragrance compositions and, to a small extent, in pharmaceutical preparations.

Vanilla extract (an oleoresin) is produced by extraction of the pods of *Vanilla planifolia* Andrews or *V. tahitensis* J. W. Moore (Orchidaceae) with a polar solvent (e.g., methanol, ethanol, or acetone, which may also contain water). The composition of the extract depends on the type and amount of solvent used. Generally, the percentage of vanillin in the extract (yield 25 – 30 %) is 3 – 4 times higher than that in the pods.

Vanillin and phenol derivatives are primarily responsible for its aroma. The main producers of *V. planifolia* pods are Madagascar, the Comoro Islands, and Réunion (Bourbon); *V. tahitensis* pods are grown mainly in Tahiti.

Vanilla extracts are used extensively in chocolate and baked products, but even more so in ice cream.

Verbena oil is produced by steam distillation of the leaves of *Aloysia triphylla* (L'Hér.) Britton [*Lippia citriodora*(Lam) Humb., Bonpl. et Kunth, *L. triphylla* (L'Hér.) Kuntze, (Verbenaceae)]. It is a light yellow liquid with a delicate lemon odor.

d_{20}^{20} 0.890 – 0.920; n_D^{20} 1.4820 – 1.4880; $[\alpha]_D$ − 18 ° to − 10 °; content of carbonyl compounds (calculated as citral): 32 – 38 %; solubility: 1 vol in 1 vol of 90 % ethanol.

The major components of "true" verbena oil are citral (ca. 40 %), geraniol, and nerol (ca. 5 % each).

The verbena plant, originates in South America, but is cultivated mainly in southern France. The oil is used in fine fragrances. IFRA recommendation.

Vetiver oil is produced by steam distillation of the roots of the grass *Vetiveria zizanioides* (L.) Nash (Poaceae), which grows wild or is cultivated in many tropical and subtropical countries. The oil is a brown to reddish-brown, viscous liquid with a characteristic precious-wood- and rootlike, long-lasting odor. The following specifications are for oil produced in Haiti:

d_{20}^{20} 0.980 – 1.001; n_D^{20} 1.5160 – 1.5270; $[\alpha]_D^{20}$ + 22 ° to + 46 °; acid value: min. 1, max. 14; ester value: 5 – 35; carbonyl value: 23 – 59; solubility: 1 vol in 2 vol of 80 % ethanol at 20 °C.

The tenacity of the highly complex vetiver oil is attributed to its high sesquiterpene content. The ketones α-vetivone [*15764-04-2*] and β-vetivone [*18444-79-6*], which form up to 10 % of the oil, as well as khusimol [*16223-63-5*] (up to 15 %) and its derivatives, contribute significantly to its odor [338].

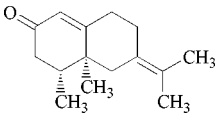

α-Vetivone

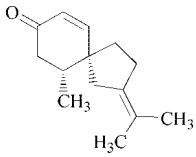

β-Vetivone

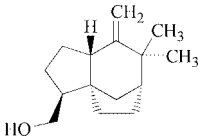

Khusimol

The oil contains a considerable number of bi- and tricyclic primary, secondary, and tertiary sesquiterpene alcohols, called vetiverols or vetivenols. These alcohols, as well as their acetates, are valuable fragrance materials. Since several varieties of vetiver grass exist and since fresh as well as air-dried roots are distilled under conditions that vary with the producer, the quality of the commercial oils differs considerably. Oil yields are up to 3 %. The main producers are Indonesia, China, and Haiti (ca. 300 t/a).

Vetiver oil is used in fine fragrances for long-lasting precious-wood notes. It is also used as starting material for vetiveryl acetate (see page 41).

Violet leaf absolute is obtained by solvent extraction, via the concrete, from the leaves of *Viola odorata* L. (Violaceae), which is grown predominantly in southern France.

The absolute is a dark green to brown liquid with a strong, green odor. The main constituent of the volatile fraction is 2-*trans*-6-*cis*-nonadienal [*557-48-2*] (violet leaf aldehyde).

Violet leaf absolute is used frequently in perfume compositions, but only in very low concentration because of its intense odor.

Ylang-ylang and cananga oils are essential oils that are obtained from two subspecies of the cananga tree.

1) *Ylang-ylang oils* are obtained by steam distillation of freshly picked blossoms of *Cananga odorata* (Lam.) Hook. f. et Thomson (Annonaceae). These cananga trees normally grow to a height of 20 m but are pruned to a height of 2.10 – 3.00 m and flower throughout the year. The oil is produced mainly in Madagascar and the Comoro Islands. Four fractions are collected at progressively longer distillation times and are known as "Extra", "I", "II", and "III". They are all pale to dark yellow liquids with a characteristic floral, spicy, balsamic odor. The first fractions are the most valuable; they have a higher density and a higher saponification value. Specifications of fractions obtained from Comoro Island oils are given in Table 6 [339, 340].
   The compositions of the various oil fractions depend on the duration of distillation. The first fraction, ylang-ylang oil Extra, has the highest content of strongly odoriferous constituents such as *p*-cresyl methyl ether (about 8 %), methyl benzoate (about 5 %), (−)-linalool (about 11 %), benzyl acetate (about 30 %), and geranyl acetate (about 4 %). The other fractions contain increasing amounts of sesquiterpene hydrocarbons (mainly caryophyllene and germacrene D), which can reach 70 % in ylang-ylang III. Components such as *p*-cresol, eugenol, and isoeugenol are important for the odor, although they are present only in low concentration.
   Ylang-ylang Extra and I are used mostly in fine fragrances; ylang-ylang II and III are employed in soap perfumes.
2) *Cananga oil* is produced by steam distillation of the flowers of *Cananga odorata* (DC) Hook. f. et Thomson subsp. *macro-*

Table 6. Specification of ylang-ylang oils from the Comoro Islands

Parameter	Ylang-ylang fractions			
	Extra	I	II	III
d_{20}^{20}	0.955 – 0.976	0.938 – 0.960	0.925 – 0.945	0.906 – 0.925
n_{D}^{20}	1.498 – 1.506	1.501 – 1.509	1.502 – 1.511	1.503 – 1.513
$[\alpha]_{D}^{20}$	$-40°$ to $-20°$	$-46°$ to $-25°$	$-60°$ to $-35°$	$-72°$ to $-45°$
Acid value	<2	<2	<2	<2
Ester value	140 – 185	100 – 160	75 – 115	45 – 75

phylla (Annonaceae). The yield, based on the weight of the flowers, is ca. 1 %: The oil is a light to dark yellow liquid with a characteristic, floral, slightly woody odor.
d_{20}^{20} 0.906 – 0.923; n_{D}^{20} 1.4950 – 1.5030; $[\alpha]_{D}^{20}$ $-30°$ to $-15°$; solubility: 1 vol in 1 vol of 95 % ethanol at 20 °C (solution may be turbid); acid value: max. 2; ester value: 13 – 35 [341].
The qualitative composition of cananga oil resembles that of ylang-ylang III oil, but is distinguished by its higher caryophyllene content (30 – 40 %).
Cananga oil originates almost exclusively in Java, where the flowers are collected throughout the year; annual production is 30 – 50 t. The oil is used mainly in soap perfumes.

4. Quality Control

Quality control of fragrance and flavor substances, as well as the products derived from them, comprises the comparison of sensory, analytical, and if necessary, microbiological data with standards and specifications. To a large extent these have been established in official specification collections (Pharmacopoeias, ISO, Association Française de Normalisation (AFNOR), Essential Oil Association, Merck Index).

In the past few decades, a precise analytical methodology has been developed for sensory evaluation and has proved to give reliable results [342]. Analytical determination of identity and purity aids greatly in establishing the acceptability of fragrance and flavor materials.

Single fragrance and flavor compounds are characterized by generally accepted physical constants, such as density, refractive index, optical rotation, and melting point. The advantages of these parameters are their speed of measurement and the possibility of comparison with other laboratories; therefore, they will continue to be indispensable. In addition, these measurements can be done more and more by robots controlled by computers.

Determinations of content by chemical parameters, such as ester and carbonyl values, are being increasingly supplemented by chromatographic procedures such as GC and HPLC and by spectroscopic techniques (UV, IR, NIR, and NMR).

Standardization of specifications for complex fragrance and flavor materials, such as essential oils and animal secretions, is far more difficult than for single compounds. In addition to organoleptic and physical properties, the content of certain typical components is determined. Problems concerning the natural, botanical, and geographical origins of these products are solved by using modern chromatographic and spectroscopic analytical techniques (and combinations such as GC – MS). Furthermore the determination of chiral properties of the molecules by enantioselective capillary GC has become a routine method. Isotope ratio mass spectrometry (IRMS) reveals the stable isotope ratios of the bioelements C, H, O, N, and S and the positional isotope analysis can be performed by Site-specific Natural Isotope Fractionation – NMR (SNIF – NMR) [343]. These methods have become very effective tools for authenticity control.

The analysis of trace components (halogens, heavy metals, and pesticides) in flavors and fragrances that are used in foods and cosmetics is becoming increasingly important. Radioactivity values in materials of natural origin are also monitored.

5. Economic Aspects

The flavor and fragrance industry has grown from an artisanal development into an industrial business. The total market for these product categories is estimated to about $ 16×10^9, Europe and North America being the leading markets with shares of about $ 5.4 and 4.1×10^9, respectively. Total annual growth rates till 2005 are forecasted to reach about 4.5 %, growth in Europe and North America will be below 3 %. In general it is predicted that sales of flavors will grow by about 5.5 % and fragrances on a whole by about 3 %.

The flavor and fragrance industry has undergone a great degree of consolidation since the mid-1990s. The top ten flavor and fragrance producers have combined sales in excess of $ 9×10^9, corresponding to about 55 % of the total value. The remaining 45 % of the market are shared by more than 1000 other companies, many of them serving only one to three customers. Out of the top ten companies, four have annual sales in excess of $ 1×10^9, namely Givaudan and Firmenich of Switzerland, IFF of the USA and Quest, which is affiliated with ICI. Takasago of Japan also reports sales in excess of $ 1×10^9, however its sales in the flavors and fragrances field are less than that. The remaining top ten companies, in addition to the aforementioned, are Dragoco and Haarmann and Reimer, both in Germany, Universal Flavors in the USA, and Hasegawa in Japan. The last one of this group, BBA of the USA, has recently been taken over by IFF. In conclusion, it can be foreseen that further consolidation will increase the market shares of these top ten companies over the next five years.

6. Toxicology and Environmental Aspects

The addition of flavor and fragrance compounds to foods, cosmetics and toiletries, and other consumer products is intended to improve the quality of those products. So these chemicals must not have adverse effects on health or well-being. To avoid any problems caused by perfumed consumer products, the fragrance industry decided to build up a self-regulatory system.

Already in 1966 in the USA the Research Institute for Fragrance Materials, Inc. (RIFM, Two University Plaza, Suite 406, Hackensack, N.J. 07601) was established for the sole purpose of fragrance safety. As an independent nonprofit organization, RIFM has a continuing program, testing fragrance materials of good commercial quality. All judgements in matters pertaining to the evaluation of safety are made exclusively, and independently by an international expert panel of toxicologists, pharmacologists, and dermatologists drawn from the academic world. Those results are published as 'RIFM monographs' in the journal 'Food and Chemical Toxicology' [344] and in the Internet under www.rifm.org/data/start.cfm. These monographs contain specifications and data on biological activity (e.g., acute, short-term, and long-term toxicity, toxicokinetics, and metabolism).

In 1973 national associations of fragrance manufacturers have formed the International Fragrance Association (IFRA, 8, Rue Charles Humbert, CH-1205 Genève, Switzerland), an international organization with scientific aims concerned with all aspects of safety evaluation and regulation of the fragrance industry. IFRA represents over 100 fragrance manufacturers in 15 countries and edits guidelines advising against the use of fragrance materials under conditions that might provoke irritation and sensitization reactions or phototoxic effects (see Section 1.5.2). The activities of RIFM and IFRA are harmonized by a Joint Advisory Committee. Similarly to IFRA, the International Organization of the Flavor Industry (IOFI, same adress as IFRA) publishes lists of naturally occurring and nature-identical flavor components, which may be used up to given concentrations, and of artificial flavor substances that are considered to be nontoxic.

For more than 40 years the Flavor and Extract Manufacturers' Association's (FEMA) Expert Panel Food Additives conducts a comprehensive program of safety evaluation of flavoring substances under the authority of Section 201(s) of the Federal Food, Drug, and Cosmetic Act. FEMA publishes compendiums with flavor components commonly known as GRAS.

In the 1980s and especially in the 1990s a growing awareness developed that flavor and fragrance chemicals can also affect the environment. So the Material Safety Data Sheets of fla-

vors and fragrances (MSDS, see EEC Directive 91/155/EEC) now contain in addition to the toxicological information also ecological information, mainly: biodegradability and aquatic toxicity. In most countries, the use of flavor and fragrance materials in consumer products is extensively regulated.

Recent Discussions. In 2002 *nitromusks* were identified in surface water, fish, and human fat and milk samples. Due to their low biodegradability these substances – having been used for 100 years – accumulate in the environment. In addition a relatively high log (partition coefficient octanol – water) (log $P_{ow} > 5$) shows that these substances are lipophilic and therefore accumulate also in fat tissue. RIFM has initiated the execution of various studies: acute and subchronic toxicity, irritation, sensitization and photosensitization potential, genotoxicity, carcinogenicity bioassay; studies on dermal absorption and pharmacokinetic.

Reports of *polycyclic musks* in the environment have also prompted a review of the safety of these important fragrance ingredients and similar studies as for nitromusks were conducted. For both groups the following summary is valid: Margins of safety based on the no-adverse-effect levels from subchronic studies are sufficient to conclude that direct exposure from use of these materials in fragrances or indirect exposure from their presence in the environment pose no significant risks to human health. Their environmental risk assessment is the subject of separate reviews.

The discussions on the nitro and polycyclic musks have given increased evidence of the importance of knowing as much as possible about properties of fragrance materials such as biodegradability and partition coefficient octanol – water.

7. References

The authors would like to thank B. M. Lawrence and K. D. Protzen for providing information about production quantities of essential oils.

The abbreviation EOA refers to the Essential Oil Association of the United States.

1. E. von Skramlik, *Z. Psychol. Physiol. Sinnesorg.* **53** (1922) 36.
2. W. Pickenhagen in W. Pickenhagen, C.-T. Ho, A. M. Spanier (eds.): *The Contribution of Low- and Nonvolatile Materials to the Flavor of Foods,* Allured Publishing Corporation, Carol Stream, Ill. 1996, p. 3.
3. E. Wick, *Food Technol.* **19** (1965) 145.
4. G. Ohloff: *Scent and Fragrances,* Springer Verlag, Berlin – Heidelberg – New York 1994.
5. F. Wöhler, *Ann. Phys. Chem.* **12** (1828) 253.
6. A T. James. A. I. P. Martin, *Analyst* **77** (1952) 915.
7. E. I. Golay, Gas Chromatog. Proc. Symposium, Amsterdam 1958, 36 – 38.
8. L. M. Nijssen et al. (eds.): *Volatile Compounds in Food: Qualitative and Quantitative Data,* TNO Nutrition and Food Research, Zeist 1963 – 1999.
9. R. Harper, E. C. Bate Smith, D. G. Land: *Odour Description and Odour Classification,* J.+A. Churchill Ltd., London 1968.
10. D. Kastner, *Parfümerie Kosmetik* **54** (1973) 97.
11. J. S. Jellinek, *Ann. NY Acad. Sci.* **116** (1963) 725.
12. L. J. van Great, A. H. Nettenbretter: "Compilation of odour threshold values in air and water", published by The National Institute for Water Supply, Voorburg and the Netherlands Central Institute for Nutrition and Food Research, TNO, Zeist 1977.
13. F. Drawert, N. Christoph: "Significance of the sniffing-technique for the determination of odour thresholds and detection of aroma impacts on trace volatiles", in P. Schreier (ed.): *Analysis of Volatiles, Proceedings International Workshop Würzburg, Sept. 1983,* Walter de Gruyter, Berlin 1984.
14. D. G. Guadagni, R. G. Buttery, *J. Food Sci.* **43** (1978) 1346.
15. Manual on Sensory Testing Methods, ASTM Spec. Tech. Publ. 434, American Society for Testing Materials, Philadelphia 1968.
16. M. Rothe, B. Thomas, *Z. Lebensm. Unters. Forsch.* **119** (1963) 302.
17. CFR Ch. I,4-1-00 Edition, § 101.22 a, (1).
18. CFR Ch. I, 4-1-00 Edition, § 101.22 a, (3).
19. Firmenich SA, CH 581 446, 1973 (F. Naef, G. Ohloff, A. Eschenmoser).
20. International Flavors & Fragrances Inc., FR 1 508 854, 1968 (H. J. Blumenthal).
21. Ruhrchemie AG, EP 7 609, 1978 (H. Bahrmann, B. Cornils, G. Diekhaus, W. Kascha et al.).

22. Union Carbide, US 2 628 257, 1953 (R. I. Hoaglin, D. H. Hirsh).
23. Naarden International N.V., NL 76 08 333, 1976 (P. C. Traas, H. Boelens, H. J. Wille).
24. Monsanto Chemical Corp., US 2 455 631, 1945 (O. J. Weinkauff).
25. Firmenich SA, CH 544 803, 1970 (F. Naef).
26. IG Farbenindustrie AG, DE 705 651, 1941 (W. Flemming, H. Buchholz).
27. Union Camp Corp., EP-A 132 544, 1984 (P. W. Mitchell, L. T. McElligott, D. E. Sasser).
28. The Glidden Company, US 3 031 442, 1958 (R. L. Webb).
29. Kuraray Co., JP 75 58 004, 1973 (O. Yoshiaki, Y. Ninagawa, Y. Jujita, T. Hosogai et al.).
30. SCM Corp., US 4 254 291, 1978 (B. J. Kane).
31. G. Ohloff, E. Klein, *Tetrahedron* **18** (1962) 37 – 42.
32. Rhône-Poulenc Ind., SA, DE-AS 2 025 727, 1970 (P. S. Gradeff, B. Finer).
33. SCM Corp., US 4 018 842, 1976 (L. A. Canova).
34. Studiengesellschaft Kohle, US 3 240 821, 1966 (G. Ohloff, E. Klein, G. Schade).
35. Glidden Co., US 3 076 839, 1958 (R. L. Webb).
36. Hoffmann-La Roche, CH 342 947, 1959 (W. Kimel, W. N. Sax).
37. BASF, GB 848 931, 1960 (H. Pasedach, M. Seefelder).
38. Hoffmann-La Roche, BE 634 738, 1964 (R. Marbet, G. Saucy).
39. Kuraray, DE 2 356 866, 1974 (Y. Tamai, T. Nishida, F. Mori, Y. Omura et al.).
40. BASF, DE 1 267 682, 1967 (H. Müller, H. Köhl, H. Pommer).
41. BASF, DE-AS 1 618 098, 1967 (H. Müller, H. Overwien, H. Pommer).
42. BASF, DE-AS 1 286 020, 1967 (H. Pommer, H. Müller, H. Overwien).
43. BASF, DE-AS 1 643 710, 1967 (H. Pasedach, W. Hoffmann, W. Himmele).
44. H. Dorn, N. Tian, H. Qiao, X. Lu, Sepu **1985,** 2(2) 75 – 79; *Chem. Abstr.* **103** (1985) 119978k.
45. Takasago Perfumery, EP-A 112 727, 1984 (S. Mitsuhashi, H. Kumobayashi, S. Akutagawa).
46. Universal Oil Products, US 3 275 696, 1961 (E. Goldstein).
47. BASF, DE-OS 2 934 250, 1979 (M. Horner, M. Irrgang, A. Nissen).
48. K. Ziegler, DE-AS 1 118 775, 1960.
49. Hercules Powder, US 2 388 084, 1945 (A. L. Rummelsburg).
50. Glidden, US 2 902 510, 1959 (R. L. Webb).
51. K. Kogami, J. Kumanotani, *Bull. Chem. Soc. Japan* **41** (1968) 2508 – 2514; *Chem. Abstr.* **70** (1969) 20230y.
52. Rhône-Poulenc, DE-AS 1 811 517, 1968 (P. Charbardes).
53. Hoffmann-La Roche, DE 2 353 145, 1973 (N. Hindley, D. A. Andrews).
54. Kuraray Co., JA 51 048 608, 1974 (T. Hosogai, T. Nishida, K. Itoi); *Chem. Abstr.* **85** (1976) 142638t.
55. BASF, DE-AS 1 901 709, 1969 (H. Overwien, H. Müller).
56. BASF, DE-AS 2 715 209, 1977 (C. Dudeck, H. Diem, F. Brunnmüller, B. Meissner et al.).
57. BASF, DE-AS 2 715 208, 1977 (B. Meissner, W. Fliege, O. Woerz, C. Dudeck et al.).
58. BASF, DE-AS 2 625 074, 1976 (A. Nissen, G. Kaibel).
59. BASF, EP 21 074, 1980 (A. Nissen, W. Rebafka, W. Aquila).
60. V. N. Kraseva, A. A. Bag, L. L. Malkina, O. M. Khol'mer et al., SU 118 498, 1958; *Chem. Abstr.* **53** (1959) 22067h.
61. Glidden, US 3 028 431, 1962 (R. L. Webb).
62. Rhodia Inc., US 3 971 830, 1976 (P. S. Gradeff).
63. D.-J. Wang, *K'o Hsueh Fa Chan Yuch K'an* **7** (1979) no. 10, 1036 – 1048; *Chem. Abstr.* **92** (1980) 124929d.
64. L'Air liquide, DE-AS 2 045 888, 1970 (M. Vilkas, G. Senechal).
65. T. Kumano, JP 70 03 366, 1965; *Chem. Abstr.* **72** (1970) 111650t.
66. Givaudan, DE-AS 2 211 421, 1972 (L. M. Polinski, B. Venugopal, E. Eschinasi, J. Dorsky).
67. Takasago Perfumery, DE 3 044 760, 1980 (S. Akutagawa, T. Taketomi).
68. Soda Aromatic Co., JP 73 22 411, 1971 (T. Yasukawa, A. Abe, R. Eto, M. Ishitani et al.); *Chem. Abstr.* **79** (1973) 18085x.
69. Distillers Company, GB 878 680, 1960 (P. Nayler).
70. BASF, DE 1 768 980, 1968 (H. Pasedach).
71. Hoffmann-La Roche, GB 774 621, 1957 (J. A. Birbiglia, G. O. Chase, J. Galender).
72. Haarmann & Reimer GmbH, BE 615 962, 1962 (K. Bauer, J. Pelz).
73. SCM Corp., US 4 000 207, 1976 (G. L. Kaiser).
74. R. Emberger, R. Hopp in R. G. Berger, S. Nitz, P. Schreier (ed.): *Topics in Flavour Research,* H. Eichhorn Marzling-Hangenham 1985, p. 201 – 218.

75. Takasago Perfumery, EP 68 506, 1982 (S. Otsuka, K. Tani, T. Yamagata, S. Akutagawa, H. Kumobayashi, M. Yagi).
76. Bush Boake Allen, FR 1 374 732, 1964 (B. T. D. Sully, P. L. Williams).
77. Glidden, US 3 031 442, 1958 (R. L. Webb).
78. Sukh Dev: "New Industrial Processes Based on Carene", *IUPAC Int. Symp. Chem. Nat. Prod. 11th* 1978, 4 (Part 1) 433 – 447 (N. Marekov, I. Ognyanov, A. Arahovates ed.).
79. Haarmann & Reimer, DE 2 109 456, 1971 (J. Fleischer, K. Bauer, R. Hopp).
80. Takasago Perfumery, DE-AS 1 815 845, 1968 (T. Moroe, S. Hattori, A. Komatsu, Y. Yamaguchi).
81. Hoechst AG, DE 955 499, 1956 (H. Hoyer, G. Keicher, H. Schubert).
82. Dragoco, DE 1 235 306, 1965 (E. Klein).
83. B. Masumoto, T. Funahashi, JP 5 368, 1952; *Chem. Abstr.* **48** (1954) 8263g.
84. Dow Chemical Co., US 3 124 614, 1958 (L. J. Dankert, D. A. Pernoda).
85. Glidden, US 3 293 301, 1964 (J. M. Derfer, B. J. Kane, D. G. Young).
86. Y.-R. Naves, *J. Soc. Cosmet. Chem.* **22** (1971) 439 – 456.
87. Y.-R. Naves, *Riv. Ital. Essenze* **58** (1976) no. 10, 505 – 514.
88. Haarmann & Reimer, DE 2 455 761, 1974 (G. Blume, R. Hopp, W. Sturm).
89. BASF, DE-AS 1 286 019, 1967 (H. Pommer, W. Reit, W. Pasedach, W. Hoffmann).
90. BASF, DE-AS 1 286 018, 1967 (H. Pommer, W. Reif, H. Pasedach, W. Hoffmann).
91. Hoffmann-La Roche, DE 1 109 677, 1959 (R. Marbet, G. Saucy).
92. Polak & Schwarz, GB 812 727, 1959 (M. G. J. Beets, H. van Essen).
93. Hoffmann-La Roche, GB 865 478, 1961.
94. Studiengesellschaft Kohle, FR 1 355 944, 1962 (G. Ohloff, G. Schade).
95. Firmenich, CH 563 951, 1972 (K. H. Schulte-Elte).
96. J. Kulesza, J. Podlejski, A. Falkowski, PL 57 707, 1966; *Chem. Abstr.* **72** (1970) 12914a.
97. Japan Terpene Chemical Co., JP 68 27 107, 1966 (Y. Matsubara, K. Takei); *Chem. Abstr.* **70** (1969) 58071k.
98. W. Klose, K. Bochow, DD 69 586, 1967.
99. Firmenich, DE 2 008 254, 1970 (E. Sundt, G. Ohloff).
100. E. Demole, P. Enggist, G. Ohloff, *Helv. Chim. Acta* **65** (1982) no. 6, 1785 – 1794.
101. Givaudan, CH 629 461, 1977 (W. M. Easter, R. E. Naipawer).
102. International Flavors & Fragrances, US 4 014 944, 1976 (J. B. Hall, W. J. Wiegers).
103. Haarmann & Reimer, DE-OS 2 921 139, 1979 (G. K. Lange, K. Bauer).
104. International Flavors & Fragrances, US 4 007 137, 1975 (J. M. Sanders, W. L. Schreiber, J. B. Hall).
105. International Flavors & Fragrances, US 2 947 780, 1960 (R. W. Teegarden, L. Steinbach).
106. Bayer AG, DE-OS 2 437 219, 1974 (H. Stetter, H. Kuhlmann).
107. International Flavors & Fragrances, US 4 045 489, 1976 (W. J. Wiegers, J. B. Hall).
108. Firmenich, DE 2 502 767, 1975 (K.-H. Schulte-Elte).
109. IFF, DE 2 036 936, 1971 (J. B. Hall, E. T. Theimer).
110. IFF, DE 2 220 820, 1973 (J. B. Hall, L. K. Lala).
111. Firmenich, US 4 302 607, 1981 (G. H. Büchi, H. Wüest).
112. Firmenich, FR 1 558 413, 1969.
113. Toray Industries, DE 2 141 309, 1971 (A. Miyake, M. Nishino, H. Kondo).
114. International Flavors & Fragrances, US 3 929 677, 1975 (J. B. Hall, J. M. Sanders).
115. Bayer, US 2 582 743, 1952 (M. Bollmann, E. Kroning).
116. Van Ameringen-Haebler, US 2 927 127, 1960 (W. T. Somerville, E. T. Theimer).
117. Firmenich, GB 907 431, 1961 (A. Firmenich, R. Firmenich, G. Firmenich, R. E. Firmenich).
118. Roure Bertrand Dupont S.A., DE 2 732 107, 1977 (P. J. Teisseire, M. Plattier, E. Giraudi).
119. Nippon Zeon, EP 0 399 788 A2, 1990 (T. Yamada, H. Fujisawa, H. Tanaka).
120. Firmenich, WO 96/00 206, 1996, (V. Rautenstrauch, J.-J. Riedhauser).
121. Polak's Frutal Works, NL 71 15 065, 1972 (A. M. Cohen).
122. Toyo Soda, JP 78 31 635, 1976 (K. Kondo, K. Komiya, H. Iwamoto); *Chem. Abstr.* **89** (1978) 75313a.
123. Ruhrchemie, FR 2 015 475, 1970.
124. Engelhard Minerals and Chemicals Corp., US 3 655 777, 1968 (P. N. Rylander, D. R. Steele).
125. H. I. Schlesinger, H. C. Brown, US 2 683 721, 1954.
126. Ajinomoto, JP 72 42 278, 1971 (Y. Matsuzawa, T. Yamashita, S. Ninagawa); *Chem. Abstr.* **78** (1973) 135878y.
127. Mitsubishi Petrochemical Co., JP 74 25 654, 1970 (T. Hashimoto, M. Tanaka); *Chem. Abstr.* **82** (1975) 170375m.

128. Lee et al., *Gas Chromatography,* Academic Press, New York 1962, p. 475.
129. Universal Oil Products, US 3 520 934, 1968 (W. Dunkel, D. J. Eckardt, A. Stern).
130. Rhône-Poulenc, DE-AS 1 145 161, 1958 (I. Scriabine).
131. BASF, EP 32 576, 1980 (N. Goetz, W. Hoffmann).
132. United States Rubber Co., US 2 529 186, 1950 (H. R. Richmond).
133. Polak's Frutal Works, FR 1 392 804, 1963 (J. Stofberg, B. van der Wal, K. N. Nieuwland).
134. Givaudan, US 3 078 319, 1960 (T. F. Wood).
135. Naarden Int. N.V., NL-A 78 02 038, 1977.
136. Polak's Frutal Works, DE-AS 1 045 399, 1954 (E. H. Polak).
137. Givaudan, US 3 856 875, 1973 (T. F. Wood, E. Heilweil).
138. Recherche et Industrie Therapeutiques, GB 1 032 879, 1965 (P. Crooy).
139. Warszawa Fabryka Syntetykow Zapachowych, PL 50 349, 1963 (T. Zardecki); *Chem. Abstr.* **66** (1967) 55210m.
140. Haarmann & Reimer, DE-AS 2 256 483, 1972 (K. Bauer, U. Harder, W. Sturm).
141. Council of Scientific and Industrial Research, IN 146 359, 1977 (Y. R. Ramachandra, S. N. Mahapatra); *Chem. Abstr.* **92** (1980) 22281u.
142. All-Union Scientific Research Institute of Synthetic and Natural Perfumes, SU 261 380, 1970 (G. E. Svadkowskaya, L. A. Kheifits, A. I. Platova, K. Denisenkova); *Chem. Abstr.* **73** (1970) 14450r.
143. Haarmann & Reimer, DE-AS 2 418 974, 1974 (K. Bauer, R. Mölleken).
144. Rhône-Poulenc, DE-AS 1 936 727, 1969 (P. Gandilhon).
145. Givaudan, DE-OS 2 508 347, 1975 (A. J. Chalk).
146. S. G. Polyakova, L. L. Malkina, O. M. Khol'mer, I. M. Lebedev et al., SU 114 197, 1958; *Chem. Abstr.* **53** (1959) 14140e.
147. O. M. Khol'mer, S. G. Polyakova, SU 126 879, 1960; *Chem. Abstr.* **54** (1960) 19596c.
148. All-Union Scientific Research Institute of Synthetic and Natural Perfumes, SU 407 872, 1973 (G. E. Svadkovskaya, S. G. Polyakova, N. A. Daev, N. N. Zelenetskii et al.); *Chem. Abstr.* **80** (1974) 108715c.
149. Collins Chemical Co., US 3 104 265, 1963 (A. Fiecchi).
150. All-Union Scientific Research Institute of Synthetic and Natural Perfumes, DE-OS 2 221 116, 1972 (L. K. Andreeva, I. A. Kheifits, T. P. Cherkasova, E. A. Berger et al.); *Chem. Abstr.* **78** (1973) 58046w.
151. J. Imamura, *Yuko Gosei Kagaku Kyokaishi* **37** (1979) 667 – 677; *Chem. Abstr.* **92** (1980) 6158d.
152. BASF, DE 2 848 397, 1978 (D. Degner, M. Barl, H. Siegel).
153. U. Krings, D. Ganßer, R. G. Berger, *Lebensmittelchem.* **52** (1998) 123.
154. D. W. Goheen in K. V. Sarkanen (ed.): *Lignins,* Interscience, New York 1971, pp. 797 – 831.
155. Ube Industries, JP 76 128 934, 1975 (S. Nagai, T. Enomiya, T. Nakamura); *Chem. Abstr.* **86** (1977) 189514k.
156. Ube Industries, DE 2 804 063, 1978 (S. Umemura, N. Takamitsu, T. Enomiya, H. Shiraishi et al.).
157. Brichima, DE-OS 2 703 640, 1977 (P. Maggioni).
158. Haarmann & Reimer, DE-OS 2 754 490, 1977 (K. Bauer, R. Mölleken).
159. Dragoco Gerberding & Co., GB 876 684, 1958 (F. Porsch).
160. BASF, DE-AS 2 145 308, 1971 (R. Fischer, W. Körnig).
161. P. Buil, J. Garnero, D. Joulain, *Parfums, Cosmet., Aromes* **52** (1983) 45 – 49.
162. PFW Beheer, DE 2 359 233, 1974 (A. M. Cohen).
163. Takasago Perfumery Co., JP 77 122 336, 1976 (T. Hagihara, H. Tsuruta, T. Yoshida); *Chem. Abstr.* **88** (1978) 136324h.
164. Studiengesellschaft Kohle, DE 1 137 730, 1961 (G. O. Schenck, G. Ohloff, E. Klein).
165. Firmenich, CH 474 500, 1967 (L. Re, G. Ohloff).
166. Polak's Frutal Works, DE-OS 2 812 713, 1978 (A. M. Cohen).
167. International Flavors & Fragrances, US 3 910 964, 1975 (J. M. Sanders, L. H. Michael).
168. Haarmann & Reimer, EP-A 120 257, 1984 (H. Finkelmeier, R. Hopp).
169. NV Polak & Schwarz's Essencefabrieken, GB 714 645, 1954.
170. DRAGOCO, EP 0 543 470 B1, 1992 (E. J. Brunke, D. Schatkowski).
171. May & Baker, FR 1 577 817, 1967.
172. DRAGOCO, DE 197 044 84 A1, 1997 (W. Pickenhagen, D. Schatkowski).
173. Ube Industries, JP 76 95 058, 1975 (T. Yamasaki, M. Nakai, Y. Kuroki, M. Oda, Y. Kawaguchi, M. Kobayashi); *Chem. Abstr.* **85** (1976) 192191n.
174. Haarmann & Reimer, DE 2 136 496, 1971 (R. Hopp, K. Bauer).

175. Firmenich, DE 2 026 056, 1970 (J. J. Becker).
176. Soda Koryo, DE 2 818 126, 1978 (K. Suzuki, T. Eto, T. Otsuka, S. Abe).
177. Haarmann & Reimer, DE 2 731 543, 1977 (K. Bauer, A. Körber).
178. Firmenich, EP 0 424 787 A2, 1989 (P. Fankhauser, P. Fantini).
179. Haarmann & Reimer, EP 0 512 348, 1992(R. Hopp et al.).
180. Chemische Werke Hüls, DE 2 547 267, 1975 (J. Ritter, K. Burzin, K. H. Magosch, R. Feinauer).
181. K. D. Carlson, V. E. Sohns, R. B. Perkins, E. L. Huffmann, *Ind. Eng. Chem., Prod. Res. Dev.* **16** (1977) no. 1, 95 – 101.
182. Dow Chemical, US 4 036 854, 1975 (K. Y. Chang).
183. Stamicarbon, NL 72 02 539, 1972 (J. A. Thoma, J. M. Deumens, E. J. Verheijen).
184. Stamicarbon, DE 2 309 536, 1973 (S. Schaafsma, J. Deumens).
185. Union Carbide, GB 863 446, 1961 (B. Phillips, P. Staacher, D. L. Mac'Peek).
186. M. Ishihara, K. Yonetani, T. Shibai, Y. Tanaka, *7th International Congress of Essential Oils,* Kyoto 1977, pp. 266 – 269; *Chem. Abstr.* **92** (1980) 75722v.
187. H. Maarse, E. A. Visscher, *Volatile compounds in Foods, Qualitative Data,* TNO-Division for nutrition and food research TNO-CIVO Food Analysis Institute, Zeist (Netherlands) 1983 – 1987.
188. R. Tressl, T. Kossa, M. Holzer in F. Drawert (ed.): *Geruch- und Geschmackstoffe, ,* ‚Bildung flüchtiger Aromastoffe durch Maillard-Reaktionen,“ Verlag Hans Carl, Nürnberg1975, pp. 33 – 47.
189. G. Vernin, “Mechanism of Formation of Heterocyclic Compounds in Maillard and Pyrolysis Reactions”, *Chem. Heterocycl. Compd. Flavours Aromas* 1982, 151 – 207.
190. G. Ohloff, I. Flament, “The Role of Heteroaromatic Substances in the Aroma Compounds of Foodstuffs” in *Progress in the Chemistry of Organic Natural Products,* vol. 36, Springer Verlag, Wien 1979, pp. 231 – 283.
191. G. Ohloff, I. Flament, W. Pickenhagen, “Flavor Chemistry”, *Food Reviews International* **1** (1985) no. 1, 99 – 148.
192. K. D. Protzen, in R. Carle (ed.): *Ätherische Öle - Anspruch und Wirklichkeit,* Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart 1993, p. 23.
193. B. Meyer-Warnod, *Perfum. Flavor.* **9** (1984) no. 04/05, 93.
194. K. Formazek, K. H. Kubeczka, *Essential Oil Analysis by Capillary Chromatography and Carbon-13 NMR Spectroscopy,* J. Wiley & Sons, New York 1982.
195. a) B. Lawrence, *Progress in Essential Oils,* bimonthly column in Perfum. and Flavor. b) Miltitzer Berichte, collection of brief reports on essential oils, flavors, and fragrances, published annually, VEB Chemisches Werk Miltitz, Miltitz (German Democratic Republic).
196. D. Zander in Erhard, Götz, Bödecker, Seybold (eds.): *Handwörterbuch der Pflanzennamen,* 16th ed., Verlag E. Ulmer, Stuttgart 2000.
197. EOA No. 180.
198. EOA No. 183.
199. EOA No. 147.
200. ISO 3525- 1979.
201. T. A.van Beek, R. Kleis, M. A. Posthumus, A. van Veldhuizen, *Phytochemistry* **28** (1989) no. 7, 1909 – 1911.
202. EOA No. 96.
203. EOA No. 97.
204. B. D. Mookherjee, R. R. Patel, *VIIth International Congress of Essential Oils,* Kyoto 1977, pp. 479 – 482; Chem. Abstr. 92 (1980) 203412h.
205. EOA No. 194.
206. EOA No. 195.
207. B. D. Mookherjee, R. W. Trenkle in *Fedarom, Technical Data of the VIIIth Intern. Congress of Essential Oils,* Cannes 1980, published 1982 (Grasse), p. 580.
208. ISO/CD 3475- 2000.
209. ISO/DIS 11043– 1996.
210. ISO/WD 3045- 2000.
211. H. M. Boelens et al., *Perfum. Flavor.* **6** (1981/1982) no. 6, 15.
212. EOA No. 215.
213. R. Kaiser, D. Lamparsky, P. Schudel, *J. Agric. Food Chem.* **23** (1975) no. 5, 943 – 950.
214. B. M. Lawrence, *Perfum. Flavor* **23** (1998) no. 4, 47 – 50.
215. EOA No. 101.
216. EOA No. 98.
217. EOA No. 69.
218. ISO/DIS 8896– 1985.
219. ISO 4733- 1981.
220. AFNOR NF T 75-352 1989.
221. EOA No. 229.
222. EOA No. 86.
223. O. Motl, S. K. Paknikar, *Collect. Czech. Chem. Commun.* **33** (1968) 1939 – 1942; *Chem. Abstr.* **69** (1968) 25059p.

224. ISO/DIS 9843– 2000.
225. ISO/CD 4725- 2000.
226. EOA No. 36 B.
227. ISO/WD 4724- 1997.
228. ISO/DIS 3760- 2000.
229. EOA No. 156.
230. AFNOR T 75-253 1992.
231. ISO/DIS 3216- 1994.
232. ISO/DIS 3524- 2000.
233. EOA No. 87
234. ISO/FDIS 3520- 1998.
235. ISO/WD 3053- 1975.
236. R. J. Braddock, J. W. Kesterson, *J. Food Sci.* **41** (1976) no. 5, 1007 – 1010.
237. M. G. Moshonas, E. Shaw, *J. Agric. Food Chem.* **22** (1974) 282 – 284.
238. ISO/CD 855- 2001.
239. L. Haro, W. E. Faas, *Perfum. Flavor.* **10** (1985) no. 5, 67.
240. L. Haro-Guzman in *Fedarom, Technical Data of the VIIIth International Congress of Essential Oils Cannes 1980,* published 1982 (Grasse), p. 181.
241. D. McHale in *Fedarom, Technical Data of the VIIIth International Congress of Essential Oils Cannes 1980,* published 1982 (Grasse), p. 177.
242. ISO/WD 3809- 1997.
243. ISO/DIS 3519- 1994.
244. ISO/DIS 3528- 1994.
245. ISO 9844- 1991.
246. ISO/DIS 3140- 1989.
247. ISO/DIS 3142- 1994.
248. ISO/DIS 3141- 1994.
249. EOA No. 10.
250. ISO/DIS 3516- 1994.
251. ISO/DIS 9301- 2000.
252. ISO/CD 3849- 2000.
253. ISO/DIS 3848- 2000.
254. O. P. Virmani, S. C. Datta, *Flavour Ind.* **2** (1971) no. 10, 595 – 602.
255. ISO/WD 3217- 2000.
256. ISO/CD 4718- 2000.
257. ISO 4727- 1982.
258. J. Garnero, P. Buil, D. Joulain, R. Tabacchi, *Parfums, Cosmet. Aromes* **20** (1978) 33 – 36.
259. AFNOR NF T 75-254 1992.
260. ISO/FDIS 10624- 1997.
261. ISO 3065- 1974.
262. ISO/CD 7704- 1997.
263. ISO/DIS 3044- 1994.
264. E. Guenther, *Perfum. Essent. Oil Rec.* **59** (1968) 634 – 641.
265. F. Porsch, H. Farnow, H. Winkler, *Dragoco Report 1965,* 47 – 49.
266. ISO/FDIS 14716- 1998.
267. ISO 4731- 1978.
268. EOA No. 13.
269. EOA No. 63.
270. ISO 8897- 1991.
271. EOA No. 181.
272. EOA No. 119.
273. ISO/DIS 3515- 2000.
274. ISO/FDIS 4719- 1999.
275. ISO/DIS 3054- 2000.
276. ISO/FDIS 8902- 1999.
277. EOA No. 157.
278. ISO/DIS 3214- 1998.
279. ISO/FDIS 11019- 1997.
280. EOA No. 76.
281. ISO/CD 4728- 1999.
282. C. Ehret, *Neue Inhaltsstoffe aus dem Mimosablütenöl,* lecture given at the meeting of the Vereinigung von Seifen-, Parfümerie- und Waschmittelfachleuten, Bad Dürkheim,Oct. 1986.
283. ISO/FDIS 9776- 1998.
284. ISO/CD 856- 2000.
285. ISO/DIS 3033- 1986.
286. B. M. Lawrence, *Perfum. Flavor.* **5** (1980) no. 4, 6 – 17.
287. W. A. Todd, M. J. Murray, *Perfum. Essent. Oil Rec.* **59** (1968) 97 – 102.
288. ISO 3714- 1980.
289. EOA No. 212.
290. EOA No. 141.
291. R. A. Wilson, B. D. Mookherjee, *IXth Intern. Congress of Essential Oils,* Singapore 1983, Technical Paper, vol. 4, p. 1.
292. ISO/DIS 3517- 2000.
293. ISO/FDIS 3215- 1998.
294. T. H. Moxham in E. Brunke (ed.): *Progress in Essential Oil Research,* De Gruyter, Berlin 1986, p. 491.
295. S. Huneck in W. Herz et al (ed.): *Progress in the Chemistry of Organic Natural Products* **vol. 81**, Springer-Verlag, New York 2001, p. 230.
296. EOA No. 68.
297. B. M. Lawrence, *Perfum. Flavor.* **17** (1992) no. 3, 61 – 64.
298. EOA No. 67.
299. EOA No. 142.
300. ISO/FDIS 14717- 1999.
301. EOA No. 116.
302. B. M. Lawrence, *Perfum. Flavor.* **20** (1995) no. 6, 35 – 42.
303. EOA No. 293.
304. C. Masanetz, W. Grosch, *Flavor Frag. J.* **13** (1998) 115 – 124.
305. ISO/DIS 3527- 1998.

306. ISO 3527- 1975.
307. ISO/CD 3757- 1999.
308. P. Teisseire, *VIIth International Congress of Essential Oils,* Kyoto 1977, pp. 315 – 319; Chem. Abstr. 92 (1980) 129084j.
309. ISO 3061- 1979.
310. EOA No. 65.
311. ISO/DIS 3064- 1998.
312. ISO/CD 8901- 1999.
313. ISO 3043- 1975.
314. EOA No. 126.
315. EOA No. 50.
316. EOA No. 112.
317. EOA No. 133.
318. EOA No 32.
319. ISO/CD 9842- 1999.
320. G. Ohloff, E. Demole, *J. Chromatogr.* **406** (1987) 181.
321. N. Nicolov, P. Dragostinov, L. Luba, A. Tzoutzoulova, F. Portarska, *Recherches sur l'Absolue de Rose Bulgare V. Congress international de Oleos Essenciais Brasil 1971,* vol. 44 (Suplemento), Anais da Academia Brasileira de Ciencias, 1972.
322. ISO/DIS 1342- 1999.
323. R. Granger, J. Passet, G. Arbousset, *Parfums, Cosmet., Savons Fr.* **3** (1973) 307 – 312.
324. ISO/WD 3761- 2000.
325. AFNOR NF T 75-255 1992.
326. ISO/DIS 9909- 1991.
327. ISO 3526- 1991.
328. ISO/DIS 3518- 2000.
329. ISO 590- 1981.
330. EOA No. 197.
331. ISO/FDIS 11016- 1999.
332. ISO/DIS 10115- 1989.
333. ISO/DIS 4730- 1992.
334. ISO/FDIS 14715- 1999.
335. ISO/R 412- 1965.
336. ISO/FDIS 11020- 1997.
337. EOA No. 165.
338. ISO/DIS 4716- 2000.
339. ISO/CD 3063- 2000.
340. E. M. Gaydou, R. P. Randriamiharisoa, J.-P-Bianchini, J.-R. Llinas, *J. Agric. Food Chem.* **36** (1988) 574 – 579.
341. ISO/DIS 3523- 2000.
342. G. Jellinek, *Sensory Evaluation of Food,* VCH Weinheim/Ellis Horwood, Ltd., Chichester 1985.
343. A. Mosandl in R. Carle (ed.): *Ätherische Öle – Anspruch und Wirklichkeit,* Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart 1993, p. 103 – 133.
344. *Monographs on Fragrance Raw Materials, Special Issue IX, Food and Chemical Toxicology, 38, Supplement 3,* 2000, p.1 – 236.

Flocculants

Hans Burkert, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Jürgen Hartmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Classification of Flocculants 2
2.1. Inorganic Flocculants 2
2.2. Natural Organic Flocculants 2
2.3. Synthetic Organic Flocculants . . . 2
2.3.1. Chemical Structure 2
2.3.2. Production Requirements 3
3. Flocculation Mechanisms 5
3.1. Double Layer Compression 5
3.2. Specific Ion Adsorption 5
3.3. Polymer Charge Patch 5
3.4. Polymer Bridging 6
4. Uses 6
5. Choice of Optimum Flocculant . . . 6
6. Legal Aspects 8
7. Economic Aspects 9
8. References 9

1. Introduction

Flocculants are used to optimize the separation of the solid phase from the liquid phase in aqueous suspensions. These suspensions usually consist of organic or inorganic particles, which are of finely divided to colloidal consistency and are distributed in water as a dispersion medium. With smaller particles, the suspension is more stable, i.e., the lower is the tendency for the particles to agglomerate and sediment. However, addition of a flocculant drastically increases the rate of particle sedimentation so that a clear, supernatant liquid is obtained. The settled sludge can then be filtered rapidly or centrifuged, resulting in a more efficient separation.

In this article, a distinction is made between organic and inorganic flocculants in accordance with the international convention.

Inorganic flocculants are mostly aluminum or iron salts that form insoluble hydroxide precipitates in water. Impurities in the water are adsorbed onto the sedimenting precipitate flocs and are thus removed from the water. A combination of inorganic and organic flocculants is frequently employed in the treatment of water.

Organic flocculants are natural or synthetic water-soluble polymers. They may also be called sedimentation, clarification, filtration, or centrifugation aids, according to their final use. The ability of synthetic polymers to increase the sedimentation rate of suspensions was first observed in 1939 [4]. The flocculation effect exerted by synthetic water-soluble polymers was discovered at the beginning of the 1950s at BASF, and at about the same time, Cyanamid filed a patent application for polymeric flocculants in Canada [5]. Flocculants produced by BASF were first used successfully in 1955 for the clarification of washing water in German collieries [6]. Shortly afterwards, Dow Chemical introduced their flocculants into the United States market. Since then, organic flocculants have steadily gained in importance and are now used worldwide in many branches of industry and in water-treatment plants.

2. Classification of Flocculants

2.1. Inorganic Flocculants

The most important inorganic flocculants are the trivalent salts of aluminum and iron, as well as activated silica, and are listed in Table 1.

2.2. Natural Organic Flocculants

The group of natural organic flocculants includes water-soluble starch [*9005-25-8*] [7], guar gum, alginates [*9005-38-3*], and products based on chitin [8], glue, and gelatin [9]. Owing to their low relative molecular masses (10 000 – 100 000), these flocculants are only moderately efficient. Despite the very low cost of some of these natural products, the cost-to-performance

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_251

Table 1. Inorganic flocculants

Type	Formula	CAS registry number	Comments
Alum	$Al_2(SO_4)_3 \cdot 14\,H_2O$	[*7784-31-8*]	used as a powder (17 wt % Al_2O_3) or 50 wt % solution (8.5 wt % Al_2O_3)
Poly(aluminum chloride)	$Al(OH)_{1.5}(SO_4)_{0.125}Cl_{1.25}$	[*7446-70-0*]	obtained by partially hydrolyzing $AlCl_3$ [*7446-70-0*], used as a solution (10 wt % Al_2O_3)
Sodium aluminate	$NaAlO_2$	[*1302-42-7*]	used as a powder and solution
Iron(III) chloride	$FeCl_3$	[*7705-08-0*]	used as a powder (100 wt % $FeCl_3$) or solution (35 – 45 wt %)
Iron(III) sulfate	$Fe_2(SO_4)_3$	[*10028-22-5*]	used as a granulate
Iron(III) sulfate chloride	$FeClSO_4$	[*12410-14-9*]	obtained by oxidation of $FeSO_4$ with chlorine in solution, used as a solution (30 – 40 wt %)
Iron(II) sulfate	$FeSO_4 \cdot 7\,H_2O$	[*7782-63-0*]	used as a crystalline powder
Sodium silicate (activated silica)	Na_2SiO_3	[*1344-09-8*]	used as a solution

ratio of synthetic flocculants is considerably more favorable due to their very high relative molecular masses. Consequently, natural organic flocculants have steadily declined in importance. Only soluble starch was able to maintain its market position in certain fields for any length of time, e.g., in the preparation of bauxite (red mud flocculation [10], → Aluminum Oxide, Chap. 3.1.4.) and the purification of water (→ Water). Starch was chosen for the latter application because, unlike synthetic organic flocculants, its use did not require legal permission. Flocculants based on natural organic polymers are now of little importance and are not, therefore, further discussed.

2.3. Synthetic Organic Flocculants

2.3.1. Chemical Structure

The chemical structure of synthetic organic flocculants determines their flocculating properties. They are water-soluble polymers which can be classified into nonionic polymers [11], and ionic polymers (polyelectrolytes); the latter can be further subdivided into anionic, cationic, and amphoteric types [12]. Examples of flocculants from each of these groups are given in Table 2.

Nonionic Polymers. Other nonionic homopolymers not cited in Table 2 can be produced from 1-vinyl-2-pyrrolidone [*88-12-0*], *N*-vinylformamide [*13162-05-5*], methoxyethylene [*107-25-5*], etc. Poly(vinyl alcohol) [*9002-89-5*] is synthesized by hydrolysis of poly(vinyl acetate). Very high molecular mass poly(ethylene oxide) [*25322-68-3*] is obtained by polymerizing ethylene oxide over special catalysts, e.g., alkaline-earth carbonates or aluminum alcoholates (C_1–C_4) [13], [14]. Nonionic homopolymers are, however, of little commercial importance as flocculants.

Ionic Polymers. The most important ionic polymeric flocculants are copolymers derived from acrylamide [*79-06-1*] and a charged comonomer [→ Polyacrylamides and Poly(Acrylic Acids)].

Anionic Polymers. Almost all anionic polymeric flocculants are synthesized as homopolymers or acrylamide copolymers of the alkali-metal or ammonium salts of acrylic acid [*79-10-7*]. Methacrylic acid [*79-41-4*], maleic acid [*6915-18-0*], ethylenesulfonic acid [*1184-84-4*], 4-styrenesulfonic acid [*98-70-4*], and 2-methyl-2-[(1-oxo-2-propenyl)amino]-1-propanesulfonic acid [*15214-89-8*] are also used as comonomers in acrylamide copolymers, but are of little practical importance.

Cationic Polymers. Other water-soluble comonomers are used for the production of cationic polymers in addition to those listed in Table 2, the most important being the following: substituted acrylamide and methacrylamide salts [16–19]; *N*-vinylformamide [*13162-05-5*] and *N*-vinylacetamide [*5202-78-8*], which are polymerized and hydrolyzed in alkaline or acidic media to "vinylamine" copolymers [*593-67-9*] [20–23]; and salts of *N*-vinylimidazole [*1072-63-5*],

2-vinylpyridine [*100-69-6*] or 4-vinylpyridine [*100-43-6*].

Table 2. Synthetic organic flocculants

Type	Formula	CAS registry number	Comments
Polyacrylamide	$\left[CH_2CH(CONH_2)\right]_n$	[*9003-05-8*]	main nonionic organic flocculant
Poly(ethylene oxide)	$\left[CH_2CH_2O\right]_n$	[*25322-68-3*]	nonionic; can be used, like polyacrylamide, with other flocculants
Poly(sodium acrylate)	$\left[CH_2CH(COO^-Na^+)\right]_n$	[*9003-04-7*]	important anionic polyelectrolyte, used mostly as a copolymer with acrylamide
Poly[2-(*N*,*N*,*N*-trimethylamino)-ethyl acrylate] (chloride salt)	$\left[CH_2CH(COOCH_2CH_2\overset{+}{N}(CH_3)_3)\right]_n\ Cl^-$	[*54076-97-0*]	important cationic polyelectrolyte, used mostly as a copolymer with acrylamide
Polyethylenimine	$\left[N(CH_2CH_2NHCH_2CH_2)(CH_2CH_2NHCH_2CH_2NH_2)\right]_n$	[*26913-06-4*]	cationic polyelectrolyte with a high charge density in its neutralized form
Poly[*N*-(dimethylamino-methyl)acrylamide]	$\left[CH_2CH(CONHCH_2N(CH_3)_2)\right]_n$	[*25765-48-4*]	cationic polyelectrolyte formed from polyacrylamide, formaldehyde, and a secondary amine by the Mannich reaction [a], mostly neutralized or quaternized

Water-insoluble or sparingly water-soluble comonomers can also be incorporated, provided the resulting copolymers are still soluble in water.

Cationic homopolymers and copolymers of dialkyldiallylammonium chlorides, e.g., diallyldimethylammonium chloride [*7398-69-8*], can be produced by free-radical polymerization; these polymers are composed of chains bearing cyclic structures [24–28]. The polyamines are another important class of cationic flocculants and are obtained by polycondensation. The starting materials are alkylene dichlorides or epichlorohydrin and ammonia, low molecular mass alkylene polyamines, or polyaminoamides [29–33]. Polycondensation of ethyleneimine [*151-56-4*] yields polyethylenimines.

Amphoteric polymers that bear both anionic and cationic charges are of little practical importance as flocculants.

2.3.2. Production Requirements

The raw materials used to synthesize organic flocculants are described in Section 2.3.1.

High molecular mass polymers are required for efficient flocculation (cf. Section 3.4). Thus, very strict requirements are imposed on the purity of the starting materials, especially the monomers. Contamination by inhibitors or by chain-transfer or cross-linking agents must be avoided. Water is the preferred solvent because other solvents either yield polymers of lower molecular mass due to transfer reactions or pose an ecological risk.

A second important prerequisite for obtaining a high molecular mass is a polymerization temperature that is as low as possible. The polymerization temperature depends largely on the choice of initiator. Conventional free-radical initiators, such as alkali-metal persulfates and peroxides, are being superseded by azo initiators, which decompose at lower temperatures, e.g., 2,2′-azobis(2-amidinopropane)hydrochloride, 2,2′-azobis(2,4-dimethylvaleronitrile). Alternative initiators are redox catalysts (e.g., ascorbic acid) and UV and γ-radiation [34–39]. Attempts are also being made to achieve homogeneous temperature distribution by removing the heat produced during polymerization as efficiently as possible from the reaction vessel, e.g., by using evaporative cooling, special monomer feed techniques in thin layers [36], [40], [41], or two-phase systems. A method for trapping the heat of polymerization with endothermic reactions has also been described [42], [43].

Synthetic organic flocculants are marketed as liquids (either solutions or emulsions) and solids.

Solutions. Polycondensation in an aqueous solution is an important technique for producing polyamine flocculant solutions. A polymer content of ca. 50 wt % can be attained.

When acrylamide is copolymerized in solution, the high molecular masses that are desired result in highly viscous, gel-like solutions, which are difficult to handle when the polymer content exceeds 10 wt %. Special metering and solubilizing equipment is required [44]. These solutions are important starting materials for the manufacture of solid flocculants.

Emulsions. Emulsion flocculants are more important than solution flocculants and are made by the polymerization of water-in-oil emulsions [45–50] (for details of emulsion polymerization, → Polymerization Processes). A concentrated aqueous monomer solution is emulsified with emulsifiers in a dispersion phase that usually consists of a high-boiling mixture of aliphatic hydrocarbons. The particle size of the emulsions is 0.1 – 10 μm and their viscosity is low, usually less than 1500 mPa · s. Polymerization is then started by adding oil- or water-soluble initiators; azo initiators are preferred, but the use of UV irradiation has also been described [51]. Wetting agents are added to the polymerized water-in-oil emulsions prior to use; the emulsions are then inverted in water [52]. Variations on this simple batch method dissipate the heat of polymerization more efficiently either by evaporative cooling [53], [54] or by special monomer feed techniques [55], [56]. These methods yield emulsions with a polymer content of 25 – 45 wt %. An even higher polymer content can be obtained by azeotropic distillation of the internal aqueous phase and some of the hydrocarbon mixtures [57].

Water-in-oil emulsions are easy to handle despite their high polymer content and require no drying with the attendant risk of product damage. However, the starting materials are more expensive and the stability of the emulsions is limited.

Solids. Liquid products can be dried to give powders. Drying methods include precipitation of the polymers with low molecular mass alcohols or acetone in a heavy-duty mixer, drying on rollers, spray-drying, azeotropic distillation with inert organic solvents, and irradiation with microwaves. Water-in-oil emulsions are not often dried because of the relatively high initial cost of the raw materials. Solid flocculants are usually obtained from aqueous polymer solutions. High polymer concentrations are used to reduce drying costs. The solutions form firm gels, which are cut into small pieces and dried. The particle size of the final powder is adjusted by milling and sieving.

An industrially important alternative for obtaining solid flocculants is based on polymerization of a water-in-oil suspension [58–61] (for details of suspension polymerization, → Polymerization Processes). The concentrated aqueous monomer solution is emulsified in the presence of special protective colloids in a hydrophobic phase. The hydrophobic phase may be a low-boiling aliphatic compound, an aromatic hydrocarbon, or a halogen derivative of the latter. Polymerization is started with water-soluble initiators and proceeds by evaporative cooling to produce polymer beads that contain water and have a particle size of 0.1 – 0.5 mm. The water is then removed by azeotropic distillation, the hydrophobic phase serving as an entraining agent. The solid product is obtained after separation of the organic solvent, drying, and sieving.

Solid flocculants with high molecular masses can also be prepared by precipitation polymerization in, for example, *tert*-butanol or its mixtures with water [62], [63], but this method is less important.

The advantages of solid products are their high flocculant concentration, which results in lower marketing costs and increased stability. Their disadvantages are the necessity for a drying stage, with the associated risk of product damage, and their slow rate of dissolution. The latter problem has been largely solved by the development of modern dosing devices.

3. Flocculation Mechanisms

The stability of a suspension depends on the number, size, density, and surface properties of the solid particles (internal or dis-

persed phase) and on the density of the external phase or dispersion medium. The stability and destabilization of colloids, suspensions, and emulsions are discussed in detail elsewhere (→ Colloids, Chap. 5.,→ Colloids, Chap. 6.,→ Colloids, Chap. 7.; → Disperse Systems and Dispersants, Chap. 3., → Disperse Systems and Dispersants, Chap. 4.; → Emulsions).

In aqueous suspensions, the particle surface has an electrical (usually negative) charge, which, for example, in mineral solids originates from the dissociation of protons from surface hydroxyl groups belonging to silicate groups [64]. If counterions (e.g., Ca^{2+} or Mg^{2+}) are present in the surrounding water, they accumulate on the surface of the suspended particles, forming an ionic double layer [65]. The excess negative charge at the shear surface of the double layer (the zeta potential) can be measured with a zeta meter. As the zeta potential increases, the coulombic repulsion between the particles becomes stronger and the suspension is more stable.

Many authors distinguish between "coagulation" and "flocculation" [66]. *Flocculation* is usually taken to mean the process in which particles are interlinked by molecular chains with no change in the zeta potential (anionic and nonionic organic flocculants). *Coagulation,* however, is taken to mean the electrical neutralization of suspended particles, as brought about by iron and aluminum hydroxide.

Four models are used to explain how flocculants aid particle agglomeration. However, more than one of these mechanisms probably act simultaneously.

3.1. Double Layer Compression

As the cation concentration in the water increases, the thickness of the ionic double layer decreases (i.e., the double layer is compressed), thereby reducing the number of collisions between the particles and rendering the suspension unstable [67]. This mechanism accounts to some extent for the destabilization of suspensions by inorganic salts, especially by monovalent ions.

3.2. Specific Ion Adsorption

The mechanism of specific ion adsorption always applies when cations are bound chemically to the particle surface, e.g., by forming complexes with the surface atoms. Thus, the particle charge may be partially or completely neutralized [68], [69]. In the latter case, the particles no longer repel each other and the conditions for agglomeration are optimal.

Trivalent aluminum and iron salts are examples of flocculants that function in this way. In water they are partially or totally hydrolyzed to form positively charged dibasic, tribasic, or polymeric hydroxides that adsorb onto the particle surface and neutralize its charge according to their degree of hydrolysis.

3.3. Polymer Charge Patch

When cationic polymers are added to a suspension, they are adsorbed to the surfaces of particles with negative charge as described in Section 3.2. They neutralize the negative charge, but their charge density is higher than that of the particle surfaces. Thus, positively charged polymer centers (polymer charge patches) are formed on the particle surfaces and, in turn, come into contact with the negatively charged surfaces of other particles and also destabilize the suspension [70], [71]. Flocculation involving relatively low molecular mass cationic polymers (e.g., polyethylenimine) is thought to proceed by this mechanism.

3.4. Polymer Bridging

In polymer bridging, high molecular mass polymer chains are adsorbed onto the particle surface and form bridges between adjacent particles [72–75]. Adsorption occurs by electrostatic forces, van der Waals forces, hydrogen bonding, and chemical bonding. This model is used to account for the flocculation produced by high molecular mass organic flocculants. The bridging mechanisms of anionic, cationic, and nonionic organic flocculants are different.

Mineral (anionic) suspensions are generally flocculated with anionic polyelectrolytes for

economic reasons (e.g., in the coal mining industry). This is only possible beause the anionic particles are not suspended in deionized water, but in hard water. Water hardness varies and is primarily caused by divalent calcium and magnesium ions. These positively charged ions are adsorbed onto the negatively charged particle surface to form an ionic double layer. The positively charged ions face outward and allow the anionic flocculant to approach the particle closely enough to be adsorbed to its surface [72]. In this case, bridging is the result of a genuine chemical bond in which the divalent cations link the solid particles with the active groups on the polymer chain. However, hydrogen bonding may also play a role. Obviously, mineral suspensions cannot be flocculated in very soft water due to the absence of divalent ions. Therefore, the suspensions must be hardened with anhydrous lime prior to the addition of the polymer. Alternatively, cationic flocculants can be used.

Obviously, the probability that a polymer can be adsorbed simultaneously onto different particles and act as a bridge between them increases with an increase in its molecular mass. Accordingly, the most effective anionic flocculants have molecular masses of ca. $(10-20)\times10^6$. Thus, sedimentation is quicker and filtration is easier when large stable agglomerates or flocs are formed [76].

Since cationic flocculants are positively charged, they partially or completely neutralize negative charge on particle surfaces, and the polymer chains cross-link adjacent particles.

The only bridging mechanism that is available to nonionic synthetic flocculants, e.g., polyacrylamide and poly(ethylene oxide), is hydrogen bonding.

4. Uses

The fact that suspended solid particles can be made to agglomerate to form large, rapidly settling, easily filterable flocs is exploited to clarify suspensions and dewater the resulting sludges. Flocculants are added to accelerate settling in basins and to dewater sludge in vacuum filters, sand filters, belt filters, filter presses, and centrifuges. However, water purification is not the sole application of flocculants. Each branch of industry has its own priorities, which may be maximum recovery of a solid matter and/or the separation of a clear aqueous phase. Combinations of organic and inorganic flocculants can, to some extent, remove dyes and oily impurities from effluents.

If the density of the flocs is similar to that of the aqueous phase, they may float and not settle. Thus, flocculants may promote flotation, but this may be troublesome if solids of different composition are to be separated (coal and ore flotation) [77].

The most important uses of flocculants are summarized in Table 3.

5. Choice of Optimum Flocculant

Flocculant manufacturers now offer comprehensive ranges of anionic, cationic, and nonionic flocculants that differ from one another in the chemistry of their monomeric building blocks, the nature and size of their ionic charge, and their molecular mass [137]. It is not always easy to predict which flocculant is most suitable to obtain the desired sedimentation rate for a specific application. Broadly speaking, anionic and nonionic organic flocculants are best for neutral suspensions that consist primarily of inorganic solids, whereas cationic flocculants are best for neutral suspensions containing predominantly organic solids.

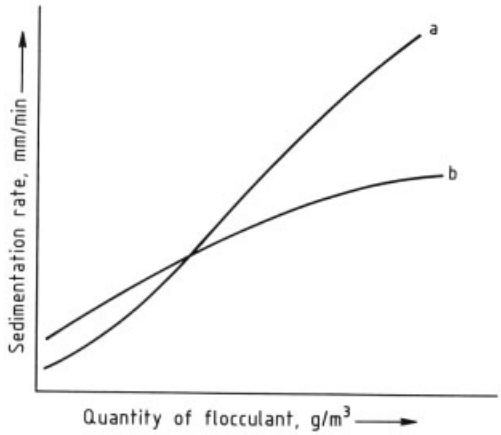

Figure 1. Relationship between the sedimentation rate of a suspension and the quantity of flocculant added
a) Weakly anionic flocculant; b) Strongly anionic flocculant

Inorganic flocculants are used mainly to clarify suspensions containing a low concentration

Table 3. Uses of Flocculants

Field of application	Use	Flocculant type∗	Reference
Mining industries			
Coal	clarification of washwater and flotation tailings,	a-PAA	[78], [79]
	filtration of concentrates,	a-PAA	[80], [81]
	dewatering of tailings	a-PAA, PEI, c-PAA	[82], [83]
Ores	clarification of flotation tailings,	a-PAA	[84–87]
	thickening and filtration of concentrates	PAA, a-PAA	
Potash	clarification of potash solutions	a-PAA	[88], [89]
	(hot solution and flotation processes),		
	dewatering of tailings	PAA, PEO	
Fluorite and barytes	clarification of flotation tailings	a-PAA	[90], [91]
Kaolin	thickening of kaolin slurries	a-PAA	[92], [93]
Salt works	clarification of raw brine, sludge dewatering	PAS, a-PAA	[94]
Phosphate	water recovery in mines,	PAA, a-PAA	
	filtration of gypsum,	a-PAA	[95–97]
	clarification of phosphoric acid	PAA	
Uranium processing	filtration of acidic or alkaline solutions	PAA, PEO	[98], [99]
Foundries and metalworking	clarification of water for scrubbing gas from blast furnaces, clarification of effluents from		
	sintering and pickling plants,	a-PAA	[100–102]
	purification of electrolytic solutions,	PAA, a-PAA	[103–105]
	clarification of electroplating effluents	a-PAA	[101]
Bauxite benification	thickening of red muds (Bayer process)	PAS	Aluminum Oxide, [106]
TiO_2 production	clarification of black liquor,	c-PAA, CCP	[107]
	filtration of TiO_2	PAA	
Paper industry	retention of kaolin and	PEI, c-PAA, PA	[108–114]
	dewatering of slurries on paper machines,	PEI, c-PAA, PA	
	pulp recovery	PEI, c-PAA, PA	
	effluent clarification and dewatering	a-PAA, c-PAA	
Communal and industrial water treatment plants	effluent purification,	IF, a-PAA	[115–117]
	thickening of activated sewage sludge,	c-PAA	[118]
	flotation of activated sewage sludge,	c-PAA	[119]
	mechanical dewatering of sewage sludge,	c-PAA	[120–124]
	phosphate precipitation	IF, a-PAA	[125]
Clarification of potable and non-potable water	flocculation of surface water,	IF, a-PAA	[126–128]
	clarification of wash water used to flush filters,	a-PAA	
	filter conditioning	a-PAA, PEI, PA	[129], [130]
Sugar processing	clarification of sugar juice, dewatering of sludge	a-PAA	[131–133]
Biotechnology	dewatering of biological sludges	PEI, PA, c-PAA	[134–136]

∗ CCP = cationic carbamoyl polymers; IF = inorganic flocculants; PA = poly(alkylene polyamines); PAA = polyacrylamide; a-PAA = copolymers of acrylamide and sodium acrylate; c-PAA = copolymers of acrylamide and cationic substituted (meth)acrylates, neutralized and quaternized; PAS = poly(sodium acrylate); PEO = poly(ethylene oxide); PEI = polyethylenimine.

of particulate material, e.g., in the purification of surface water to potable water. The voluminous hydroxide flocs adsorb both solid and dissolved contaminants. If the flocs sediment too slowly, the sedimentation rate can be accelerated by adding an organic flocculant.

pH Value. The pH is important in selection of a flocculant because it determines the degree of dissociation of anionic and cationic polyelectrolytes. Complete dissociation of the flocculant is desired, and anionic and cationic flocculants are, therefore, preferred for alkaline and acidic suspensions, respectively. Nonionic polymers are affected very little by the pH. In the case of highly acidic suspensions, nonionic polymers are often preferred to cationic flocculants for economic reasons.

Flocculant Addition. At a given flocculant addition rate, the suspension settles more slowly if its solid content increases or if its average particle diameter decreases [6]. However, the amount of flocculant that is required also depends on the strength of its ionic charge. Figure 1 illustrates the difference in behavior between a strong and a weak anionic flocculant in a neutral suspension.

Powder flocculants tend to clump when they are simply added to water, and a great deal of energy is needed to incorporate emulsions into water. Manufacturers supply instructions with their products and also offer simple devices for

overcoming the above-mentioned difficulties as well as problems encountered in continuous processes. Integrated semiautomatic and automatic equipment is commercially available. Polymeric flocculants are generally used as 0.05 – 0.1 wt % solutions. Inorganic flocculants are supplied in crystalline form or as aqueous solutions and are used as 10 wt % solutions.

Normally 0.1 – 0.5 g of flocculant is added per cubic meter of suspension in the thickener or settling basin. When sludge is dewatered with the aid of filters and centrifuges, 2 – 8 kg of flocculant per ton of dry sludge is necessary.

Laboratory Tests. When a flocculant is chosen, it is always helpful to carry out laboratory trials with a series of different flocculants [138]. Fresh suspensions must be used. Representative samples should be taken if the composition of the suspension varies considerably. The manufacturer's literature should be consulted for further details. The following tests can be used, depending on the intended application:

1) Sedimentation tests in a shaker provide information about the sedimentation rate and clarity of the supernatant [139]. Laboratory results can be helpful in calculating flocculant doses required in continuously operating thickeners, thus eliminating long trial periods in the initial period of use [140].
2) Filter plate tests (vacuum filtration).
3) Büchner funnel tests (filtration with and without vacuum) [141], [142].
4) Pressure filter tests [143].
5) Determination of filtration rate (capillary suction time) [144], [145].Tests 2 – 5 provide information about the dewatering of sludges, especially with regard to the thickness and moisture content of the filter cake and the filtration efficiency.
6) Determination of floc stability [143]; this test provides information about the dewatering of sludges in centrifuges.
7) Jar test [146]: In this test the flocculant is added to a suspension in a jar (beaker). Slow stirring induces the formation of voluminous hydroxide flocs that surround the slurry. The sedimentation and clarity of the test samples are compared after the stirrer has been switched off. This test is performed on slurries containing very little sedimentable material for which organic flocculants are generally unsuitable.
8) Dynamic flocculation measurement with the aid of a laser beam [147]. This method is also used for optimizing the flocculant metering rate in continuously operating systems.

6. Legal Aspects

The high molecular mass polymeric flocculants are not toxic in themselves. Certain precautions should be taken, however, because they all contain small proportions of residual toxic monomers [148], especially acrylamide. Legal controls apply when the flocculants are used in the manufacture of food, in the purification of potable water, or for the clarification of sugar juice.

A list of the flocculants that are approved by the EPA for water purification is printed annually in the United States [149]. Many other countries have their own regulations, but these regulations differ very little from each other. Generally, the residual monomeric acrylamide content must be less than 0.05 wt % of the polymer content; the amount of flocculant added to untreated water should not exceed 1 g/m^3 [150].

The United States regulations on the clarification of sugar juice [151] are generally observed worldwide; they stipulate that the residual acrylamide monomer content must account for less than 0.05 wt % of the polymer content, and the quantity of added flocculant must not exceed 5 g/m^3. Methods for determining the residual monomeric acrylamide content of polymers and the residual flocculant content of treated water have been described [150], [152].

7. Economic Aspects

In 1984, the use of synthetic organic flocculants (calculated as dry mass) in Western Europe was ca. 15 000 t and in the United States ca. 40 000 t, half of this being accounted for by polyacrylamides. Large markets also exist in Canada, South America, Africa, Australia, and Japan.

Of the inorganic flocculants, alum ($Al_2(SO_4)_3 \cdot 18\,H_2O$) enjoys the largest market share. In 1978, ca. 400 000 t was used in the

United States and ca. 150 000 t in Western Europe, in both cases almost exclusively for water treatment.

Representative worldwide producers of inorganic and synthetic organic flocculants are given in Tables 4 and 5.

Table 4. Representative worldwide producers of inorganic flocculants

Flocculant	Producers
Alum	Allied (USA)
	Amer. Cyanamid (USA)
	Boliden (Sweden)
	Giulini Chemie (FRG)
	Stauffer (USA)
	Nikkei Kako (Japan)
	Scintoma Aluminium Smelting (Japan)
Poly(aluminum chloride)	Giulini Chemie (FRG)
	Sachtleben Chemie (FRG)
Sodium aluminate	Giulini Chemie (FRG)
	Nalco Chemical (USA)
	Sachtleben Chemie (FRG)
Iron(III) chloride	BASF (FRG)
	Dow Chemical (USA)
	Pennwalt (USA)
	Solvay (Switzerland)
Iron(III) sulfate	Giulini Chemie (FRG)
Iron(III) sulfate chloride	Kronos Titan (FRG)
Iron(II) sulfate	Kronos Titan (FRG)
	Pfizer (USA)
	Société des Fabriques de Produits Chimiques de Thann et Mulhouse (France)
Sodium silicate	Allied (USA)
	Solvay (Belgium)
	Wöllner-Werke (FRG)

Table 5. Primary worldwide producers of synthetic organic flocculants

Producer	Flocculant trade name
Allied Colloids (UK)	Magnafloc, Percol, Zetag
Amer. Cyanamid (USA)	Superfloc, Accurac
BASF (FRG)	Sedipur, Polymin
Betz Laboratories (USA)	Polyfloc, Betz
Calgon (USA)	Calgon, Hydraid
Chemische Fabrik Stockhausen (FRG)	Praestol
Dow Chemical (USA)	Separan, Purifloc
Hercules (USA)	Hercofloc, Reten
Kurita Kogyo (Japan)	Kurifloc
Mitsubishi Chemical (Japan)	Diaclear
Nalco Chemical (USA)	Nalco
Röhm (FRG)	Rohafloc
Sankyo Chemical (Japan)	Sanpoly
Sanyo Chemical (Japan)	Sanfloc
Société Nationale de Floculant (France)	Floerger

8. References

General References

1. H. G. Flock, E. G. Rausch in N. M. Bikales (ed.): *Encyclopedia of Polymer Sci. and Techn.,* "Water-Soluble Polymers," vol. 2, Plenum Press, New York 1973, pp. 21 – 73.
2. J. Gregory in C. A. Finch (ed.): *Chemistry and Technology of Water-Soluble Polymers,* Plenum Press, New York 1983, pp. 307 – 320.
3. Philip Molyneux: *Water-Soluble Synthetic Polymers,* "Properties and Behavior," vol. 2, CRC Press, Boca Raton 1984, pp 190 – 217.

Specific References

4. O. Schweitzer, *Kautschuk* **15** (1939) 169.
5. Amer. Cyanamid, CA 734 540, 1952 (R. B. Booth, J. M. Dobson).
6. F. Drexler, *Glückauf* **92** (1956) 1023; B. Franke, *Glückauf* **92** (1956) 1028.
7. W. Petersen, *Glückauf* **70** (1934) 125.
8. R. A. A. Muzzarelli: *Chitin,* Pergamon Press, New York 1977.
9. W. Petersen, F. Gregor, *Glückauf* **68** (1932) 621.
10. Corn Products Co., BE 626 712, 1963 (E. F. Paschall, D. D. Wolfe).
11. P. Molyneux, *Water-Soluble Synthetic Polymers,* "Properties and Behavior," vol. I, CRC Press, Boca Raton 1984.
12. R. W. Armstrong, U. P. Strauss in N. M. Bikales (ed.): *Encyclopedia of Polymer Science and Technology,* **vol. 10,** Interscience, New York 1969, pp. 781 – 861 M. F. Hoover, *J. Macromol. Sci. Chem.* **A4** (1970) no. 6, 1327 – 1418 P. Molyneux, *Water-Soluble Synthetic Polymers,* "Properties and Behavior," vol. II, CRC Press, Boca Raton 1984, pp. 1 – 62.
13. Union Carbide, US 2 987 489, 1961 (F. E. Bailey, F. N. Hill).
14. Union Carbide, US 3 271 328, 1966 (F. N. Hill).
15. A. M. Schiller, T. J. Suen, *Ind. Eng. Chem.* **48** (1956) 2132 – 2137.
16. Amer. Cyanamid, US 2 595 907, 1952 (W. M. Thomas, J. A. Price).
17. Union Carbide, US 3 014 896, 1961 (C. E. Colwell, R. C. Miller).
18. Lubrizol Corp., US 3 666 810, 1972 (D. I. Hoke).
19. Jefferson Chemical Comp., US 3 878 247, 1975 (P. H. Moss, R. M. Gipson).
20. M. Nagata, O. Yoshikawa, M. Sakai, Y. Sakakibara et al., *Nippon Kagaku, Kaishi* **50** (1975) 1819 – 1822.

21. D. J. Dawson, R. D. Gless, R. E. Wingard, *J. Am. Chem. Soc.* **98** (1976) 5996 – 6000.
22. Dynapol, US 4 217 214, 1980 (P. L. Dubin).
23. BASF, EP-A 71 050, 1983 (M. Kroener, R. Schneider, H. Burkert, F. Linhart et al.).
24. Amer. Cyanamid, US 3 147 218, 1964 (R. B. Booth, W. F. Linke).
25. Calgon Corp., US 3 412 019, 1968 (M. F. Hoover, B. Park, R. J. Schaper, J. E. Boothe).
26. J. E. Boothe, H. G. Flock, M. F. Hoover, *J. Macromol. Sci. Chem.* **A4** (1970) no. 6, 1419 – 1430.
27. J. E. Lancaster, L. Baccei, H. P. Panzer, *J. Polym. Sci. Polym. Lett. Ed.* **14** (1976) 549 – 554.
28. C. Wandrey, W. Jaeger, *Acta Polym.* **36** (1985) 100 – 102.
29. Nalco Chem. Comp., US 3 174 928, 1965 (G. T. Kekish).
30. Dow Chemical, US 3 275 588, 1966 (D. C. Garms).
31. Calgon Corp., US 3 391 090, 1968 (D. L. Schiegg, M. Run).
32. A. T. Coscia in N. M. Bikales (ed.): *Encyclopedia of Polymer Science and Technology,* **vol. 10,** Interscience, New York 1969, pp. 616 – 622.
33. Amer. Cyanamid, US 3 567 659, 1971 (D. E. Nagy).
34. R. Schulz, G. Renner, A. Henglein, W. Kern, *Makromol. Chem.* **12** (1954) 20 – 34.
35. Progil, FR 2 064 729, 1969 (J. P. Communal, J. Fritz, B. Papillon).
36. Rhone-Poulenc, FR 2 348 227, 1976 (J. Boutin, J. Neel).
37. Nautschno-Issledowatelskij Fisikochimitschesky institut imeni S. J. Karpowa, DE-AS 1 645 065, 1966 (A. Abkin, A. P. Scheinker, P. M. Chomikowsky, M. K. Yakowlewa et al.).
38. Hercules Inc., US 3 926 756, 1975 (A. J. Restaino).
39. Japan Atomic Energy Research Inst., US 4 066 522, 1978 (S. Machi, T. Wada, H. Sekiya).
40. BASF, DE-AS 1 032 922, 1956 (H. Fikentscher, H. Burkert, H.-J. Krause).
41. Sanyo Chem. Ind. Ltd., JP 5 4155-296, 1978 (K. Takada, M. Suzuki).
42. Dow Chemical, FR 1 253 602, 1960 (A. C. Proffitt).
43. Allied Colloids Ltd., EP-A 123 486, 1983 (P. Flesher, D. Farrar, I. M. Johnson).
44. Hercules Inc., US 4 113 688, 1978 (S. C. Pearson).
45. Dow Chemical, US 3 284 393, 1966 (J. W. Vanderhoff, R. M. Wiley).
46. Nalco Chemical Comp., US 3 826 771, 1971 (D. R. Anderson, A. J. Frisque).
47. J. W. Vanderhoff, E. B. Bradford, H. L. Tarkowski, J. B. Shaffer et al.: "Polymerisation and Polycondensation Processes," *Adv. Chem. Ser.* **34** (1962) 32 – 51.
48. W. Baade, K. H. Reichert, *Eur. Polym. J.* **20** (1984) 505 – 512.
49. J. W. Vanderhoff, F. V. DiStefano, M. S. El-Aasser, R. O'Leary, *J. Dispersion Sci. Technol.* **5** (1984) 323 – 363.
50. C. Pichot, C. Graillat, V. Glukhikh, *Makromol. Chem. Suppl.* **10/11** (1985) 199 – 214.
51. Bayer, DE-AS 2 354 006, 1973 (W. Keller, F. Müller).
52. Nalco Chemical Comp., US 3 624 019, 1970 (D. R. Anderson, A. J. Frisque).
53. Hercules Inc., US 3 095 404, 1959 (D. C. Lincoln, W. P. Shyluk).
54. Petrolite Corp., US 4 312 969, 1982 (H. L. Becker, R. S. Buriks, J. G. Dolan).
55. Allied Colloids, GB 1 482 515, 1973 (D. B. Anderson, J. Forder, I. M. Johnson).
56. Röhm GmbH, DE-OS 2 840 894, 1978 (J. Masanek, H. Pennewiß, H. Trabitzsch).
57. C. S. Scanley, US 4 052 353, US 4 090 992, 1972.
58. Dow Chemical, US 2 982 749, 1961 (R. E. Friedrich, R. M. Wiley, W. L. Garrett).
59. Röhm GmbH, DE 2 009 218, 1970 (H. Pennewiß, H. Knoll, J. Masanek).
60. BASF, DE 2 710 372, 1977 (R. Hoene, D. Jung, H. U. Schenck, H. Spoor).
61. Mitsubishi Chem. Ind., Kyoritsu Yuki, DE-OS 2 800 520, 1978 (O. Kamada, K. Watanabe, K. Sagamikara, S. Sawayama).
62. Cassella Farbwerke Mainkur AG, DE-AS 1 301 532, 1966 (H. v. Brachel, F. Engelhardt).
63. Hercules Inc., US 3 336 269, 1967 (D. J. Monagle, W. P. Shyluk).
64. T. W. Healy, *J. Macromol. Sci. Chem.* **A8** (1974) 603.
65. R. Aveyard, D. A. Haydon: *An Introduction to the Principles of Surface Chemistry,* Cambridge University Press, New York 1973.
66. D. H. Everett, *Pure Appl. Chem.* **31** (1972) 579.
67. P. Sennett, J. P. Olivier, *Ind. Eng. Chem.* **57** (1965) no. 8, 32.
68. R. O. James, G. R. Wiese, T. W. Healy, *J. Colloid Interface Sci.* **59** (1977) 381.
69. E. Matijevic, *J. Colloid Interface Sci.* **43** (1973) 217.

70. J. Gregory, *J. Colloid Interface Sci.* **42** (1973) 448.
71. A. Rembaum in T. F. Yen (ed.): *Recycling and Disposal of Solid Wastes,* "Flocculation of Model Waste Particles by Means of Ionene Polymers," Ann Arbor Science, Ann Arbor, Mich., 1974.
72. A. S. Michaels, *Ind. Eng. Chem.* **46** (1954) 1485 – 1490.
73. J. A. Kitchener, *Br. Polym. J.* **4** (1972) 217.
74. F. M. Fowkes, M. A. Mostafa, *Ind. Eng. Chem. Prod. Res. Dev.* **17** (1978) no. 1, 3.
75. Y. A. I. Attia, J. A. Kitchener: *Proc. Eleventh Internat. Mineral Proc. Congress,* Cagliari, Italy, Apr. 1975, pp. 1233 – 1248.
76. H. E. Ries, B. L. Meyers, *J. Appl. Polym. Sci.* **15** (1971) 2023 – 2034.
77. W. Hey, K.-H. Kubitza, D. Leininger, *Glückauf-Forschungshefte* **44** (1983) 226.
78. D. A. Dahlstrom, R. P. Klepper: *VIII. Intern. Coal Preparation Congress Donezk D3,* 1979.
79. H. A. Hamza, *CIM Bull.* **72** (1979) 116.
80. W. Blankmeister, *Aufbereit. Tech.* **21** (1980) 171.
81. K. Legner, J. Schwerdtfeger, *Glückauf* **119** (1983) 237.
82. H. Hennig, S. Ehlert, E. Beekere, *Aufbereit. Tech.* **25** (1984) 681.
83. D. Leininger, P. Wilczynski et al., *Glückauf* **115** (1979) 467.
84. J. Lange, *Erzmetall* **35** (1982) 117.
85. J. Bialas, J. Ampulski, *Aufbereit. Tech.* **26** (1985) 240.
86. M. J. Pearse, J. Barnett: *Second World Filtration Congress,* London 1979, pp. 333 – 354.
87. F. N. Kemmer, J. A. Beardsley, *Eng. Min. J.* **172** (1971) 92.
88. W. Schubert, H. Schubert, *Neue Bergbautech.* **7** (1977) no. 1, 64 – 71.
89. V. A. Arsentiev, I. Leja, *CIM Bull.* **70** (1977) 154.
90. R. Rüffer, H. Teichmann, *Gas- und Wasserfach* **102** (1961) 603.
91. B. Heide, *Aufbereit. Tech.* **24** (1983) 15.
92. E. Gmey, *Montan. Rundsch.* 1958, 333.
93. Calgon Corp., EP 47 118, 1982 (J. A. Aufman).
94. Mitsubishi Chemical, J 49 012 048, 1974 (Miyaji Kazuo, Ueda Hideo).
95. I. Dombalov, *Khim. Ind. (Sofia),* **51** (1979) 117.
96. D. W. Leyshon, R. T. Schneider, *Phosphorus Potassium* **104** (1979) 32.
97. E. Jemaa, AFCFP-Quarterly J. Arab. Federation of Chemical Fertilizer Producers (1979) 44.
98. R. Bodu, L. Clappier et al., *Ind. Miner. Tech.* 1981, 417.
99. P. Young, R. Gatzweiler, W. Flöter, *Erzmetall* **35** (1982) 489.
100. J. M. Reuter, *Umwelt* 1981, no. 1, 27.
101. R. Weiner: *Die Abwässer in der Metallindustrie,* 3. ed., Eugen G. Lenze Verlag, Saulgau 1965, pp. 279 – 289.
102. H. Schramek, *WLB Wasser, Luft, Betr.* **21** (1977) 450.
103. A. v. Röpenack, U. Kleinert, *Erzmetall* **24** (1971) 421.
104. R. R. Knobler, T. I. Moore, R. L. Capps, *Erzmetall* **32** (1979) 109.
105. P. Mouchet, *Ind. Miner. Mineralurgie* 1977, 148 – 159.
106. Amer. Cyanamid, DE 3 245 806, 1983 (N. J. Brownrigg).
107. Allied Colloids Ltd., BE 848 271, 1977 (D. A. Frost, D. Farrar).
108. *Ullmann,* 4th ed., **17,** 577 – 635.
109. R. Nicke, *Zellst. Pap. (Leipzig)* **31** (1982) no. 1, 19.
110. W. Auhorn, *Rev. A T I P* **37** (1983) 1.
111. U. Beck, F. Müller, *Wochenbl. Papierfabr.* **107** (1979) 411.
112. F. Müller, U. Beck, *Papier (Darmstadt)* **33** (1979) 89 – 95.
113. D. Horn, Zellst. Pap. (Leipzig) **28** (1979) no. 3, 129.
114. H. Heel, *Umweltmagazin,* 1980, no. 1, 44.
115. H. Engelhardt, M. Neuwirth, W. Weisbrodt, *Korresp. Abwasser* **32** (1985) 940.
116. J. C. Ginocchio, H. Bischofsberger, A. Gmünder, *Chem. Ing. Tech.* **56** (1984) 310.
117. H. Predel, *Erdöl und Kohle-Erdgas-Petrochem.* **37** (1984) 502.
118. A. Bauer, W. Böwe, *Korresp. Abwasser* **31** (1984) 39.
119. R. Schrieber: *Berichte der ATV/Jahreshauptversammlung Neumünster* 1982, p. 351.
120. B. Sander, 311. Dechema-Kolloquium, Frankfurt (10. 1. 1985).
121. R. Römer, H. Lauer, *Chem. Ing. Tech.* **55** (1983) 568.
122. H. Perchthaler, *Abwassertech. Abfalltech.* **36** (1985) 32.
123. D. Reimann: *3. Europäisches Abwasser- und Abfallsymposium,* München, EA 75 (1975).
124. G. D. Abraham, *Umwelt* **5** (1984) 419.
125. R. Klute, V. Mertsch, H. H. Hahn, *Korresp. Abwasser* **31** (1984) 809.

126. R. Rehwoldt, *Environ. Sci. Technol.* **16** (1982) 616 A.
127. H. Bernhardt, *GWF Gas Wasserfach: Wasser Abwasser* **125** (1984) 312.
128. J. C. Ginocchio, *Tech. Rundsch. Sulzer* **65** (1983) no. 3, 12.
129. V. Mayer, *GWF Gas Wasserfach: Wasser Abwasser* **125** (1984) 373.
130. M. Jekel, R. Böckle, *Vom Wasser* **59** (1982) 195.
131. A. G. Noble: *Proc. Queensl. Soc. Sugar Technol., 33th Conf.,* 1971, p. 251.
132. O. L. Crees, E. Whayman, A. L. Willersdorf: *Proc. Queensl. Soc. Sugar Technol., 44th Conf.,* 1977, p. 225.
133. V. Agius, R. Attard et al.: *Proc. Queensl. Soc. Sugar Technol., 44th Conf.,* 1977, p. 277.
134. Miles Laboratories Inc., GB 53 764, 1982 (D. F. Hershberger, M. M. Sternberg).
135. Amer. Cyanamid, GB 1 438 484, 1976 (P. F. Wilde).
136. E. Wottgen, A. N. Grigorian, D. Beck, B. A. Tschernucha et al., DD 104 554, 1974.
137. J. Vostrcil, F. Juracka: *Commercial Organic Flocculants,* Noyes Data Corp., Park Ridge, N. J., 1976.
138. R. F. Packham, *Br. Polym. J.* **4** (1972) 305 – 315.
139. Flockungsmittel für Betriebswässer, DIN 23 007 (March 1983).
140. J. Sievers, *Mitteilung für den Bergbau aus der Anwendungstechn. Abt. der BASF AG,* 1965, p. 345 d.
141. B. Sander, *Chem. Ing. Tech.* **51** (1979) 684 – 685.
142. B. A. Heide, R. Kampf, H_2O **11** (1978) 339.
143. B. Sander, *Chem. Ing. Techn.* **52** (1980) 282 – 283.
144. R. C. Baskerville, R. S. Gale, *J. Water Poll. Control. Federation* **67** (1968) 233.
145. U. Loll, *Korresp. Abwasser* **24** (1977) 295.
146. A. P. Black, A. M. Buswell et al., *J. Am. Water Works Assoc.* **49** (1957) 1414 – 1424.
147. J. Eisenlauer, D. Horn, *Z. Wasser Abwasser Forsch.* **16** (1983) 9.
148. D. D. McCollister et al., *Toxicol. Appl. Pharmacol.* **7** (1965) no. 5, 639.
149. Report on Coagulant Aids for Water Treatment,Environmental Protection Agency, Office of Drinking Water, Cincinnati, Ohio, Feb. 1978.
150. Polyacrylamide zur Wasseraufbereitung, DIN 19 622 (Dec. 1977).
151. Federal Register, §121.1092 (Nov. 30th, 1963) 12 747.
152. H. Burkert, *GWF Gas Wasserfach; Wasser Abwasser* **111** (1970) 282.

Floor Coverings

Gudrun Heisterberg-Moutsis, Du Pont de Nemours (Deutschland) GmbH, Östringen, Federal Republic of Germany (Chap. 1)

Rainer Heinz, Du Pont de Nemours (Deutschland) GmbH, Östringen, Federal Republic of Germany (Chap. 1)

Thomas F. Wolf, Du Pont de Nemours (Deutschland) GmbH, Östringen, Federal Republic of Germany (Chap. 1)

Dominic J. Harper, Du Pont de Nemours (Deutschland) GmbH, Östringen, Federal Republic of Germany (Chap. 1)

David James, Du Pont de Nemours (Deutschland) GmbH, Östringen, Federal Republic of Germany (Chap. 1)

Richard P. Mazzur, Armstrong World Industries, Lancaster, Pennsylvania 17604, United States (Chap. 2)

Volker Kettler, Witex AG, Augustdorf, Federal Republic of Germany (Chap. 3)

Hansgert Soiné, Witex AG, Augustdorf, Federal Republic of Germany (Chap. 3)

1. Textile Fiber Floor Covering 2
1.1. Fiber Production 2
1.1.1. Staple Fiber 2
1.1.2. Bulked Continuous Filament 2
1.1.3. Polymer Selection 3
1.1.4. Fiber Design 3
1.2. Yarn Preparation 4
1.2.1. Staple Yarn Processing 4
1.2.2. Twisting, Heat-Setting, and Intermingling 4
1.3. Production of Textile Floor Coverings 6
1.3.1. Flooring Types and Manufacturing Methods 6
1.3.2. Carpet Constructions 7
1.3.3. Coloration and Finishing 8
1.4. Carpet Properties and Performance 9
1.4.1. Carpet Esthetics 9
1.4.2. Carpet Performance 9
1.4.3. Test Methods 11
1.5. Disposal and Recycling 11
2. Resilient Floor Covering 12
2.1. Introduction 12
2.2. Linoleum 13
2.3. Vinyl Flooring 13
2.3.1. Raw Materials 14
2.3.2. Sheet Vinyl Flooring 15
2.3.3. Tile Flooring 18
2.4. Rubber Tile and Sheet 20
2.5. Specialty Flooring Products 21
3. Laminate Flooring 21
3.1. Development 21
3.2. Construction 22
3.3. Materials 22
3.4. Production and Installation 22
3.5. Testing 25
3.6. Environmental Aspects 26
3.7. Economic Aspects 26
4. References 26

1. Textile Fiber Floor Covering

Textile floor coverings have been highly appreciated in the human environment ever since textile manufacturing techniques were developed. They deliver an individually designed artificial ground which can be soft, warm, sound-insulating, and colored/patterned.

Global carpet production grew to 2.5×10^9 m^2 in 1999, of which 52 % was produced in the United States, 27 % in Western Europe, 9 % in the Asia/Pacific region, and 12 % in other regions of the world. World carpet fiber production amounted to 2.1×10^6 t in 1999. On the basis of area, the main fiber materials used worldwide are nylon (55 %), polypropylene (34 %), polyester (5 %), and wool (3 %). The majority of carpeting is used for residential fitted carpets and rugs (68 %); the remaining 32 % is split between commercial installations (e.g., offices, hotels, banks, schools, and retail stores), automotive and transportation, and outdoor carpeting [1].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_263

For the production of carpets, either bulked continuous filament (BCF) or staple fiber can be used. Whilst BCF yarn experienced strong growth in the mass market during the 1980s and 1990s, staple fiber retains a very strong position in more upmarket segments. BCF yarns continue to gradually substitute staple fiber yarns, the substitution being driven by superior process economics along the value chain.

1.1. Fiber Production

The production of synthetic fibers for use in carpet manufacture starts with polymers, which are made from monomers (or intermediates) derived from crude oil. For example, Nylon 66, the most prevalent fiber material, is made from hexamethylenediamine and adipic acid. These react with one another to form a monomer (in this case known as nylon salt), which is then polymerized. The polymer is fed in the form of solid pellets or chips into the fiber-forming melt spinning machine. Alternatively, polymer production and fiber spinning can be coupled in a continuous melt-spinning operation. Melt spinning is used for the majority of synthetic fiber carpet yarns, including Nylon 66 and 6, polyester, and polypropylene. For a detailed discussion of fiber production and specific fibers, see → Fibers, 3. General Production Technology; → Fibers, 4. Synthetic Organic.

1.1.1. Staple Fiber

The molten polymer is extruded through shaped holes to form filaments, which are then quenched (cooled) so that they solidify. The filaments are collected in a large bundle, known as a tow or rope, consisting of hundreds of thousands of filaments. The rope is stored for further processing.

Natural fiber such as wool and, to a lesser extent, cotton, sisal, and silk, have an inherent crimp, which gives the fiber desirable value and esthetic properties, such as bulk (i.e., volume or space filling), and an ability to recover from deformation. Furthermore, most natural fibers are not continuous but instead have relatively short lengths. In a second processing operation, therefore, the synthetic fiber tow is drawn, crimped while in a relaxed state and heat-set. The tow is then cut into the desired staple fiber lengths and baled.

1.1.2. Bulked Continuous Filament

Bulked continuous filament yarn is so named because the fibers are used in continuous form (not cut, as with staple from synthetic and natural fibers), and the fiber is crimped (or bulked) during manufacture.

As with staple, filaments are extruded from the melt and quenched. They are then collected into yarn bundles comprising tens of filaments. The yarn is then drawn, crimped (bulked), and heat-set. Bulk is usually achieved by compression of the heated yarn into a small hot chamber whose dimensions determine the character of the bulk/crimp. After a residence time in the chamber sufficient to ensure heat-setting, the yarn is cooled, straightened under tension, and wound onto packages. The advantage of heat-setting is that, despite being straightened, the fibers retain a "memory" of the crimp and recover a major part of it when subjected to a combination of heat and moisture in subsequent processing at low tension. Generally, the character of crimp achieved in BCF yarns is not the same as that of staple yarns; a wider range of bulk in the finished carpet can be achieved with BCF. Yarn properties also depend on other production parameters, such as draw ratio.

1.1.3. Polymer Selection

Polymer selection is one of the most important determining factors in the balance of properties achievable in the carpet, as well as the cost and ease of fiber and subsequent carpet manufacture.

Nylon (polyamide 6 and 66) is the major carpet fiber polymer in use today, followed by polypropylene, polyester [poly(ethylene terephthalate)], and wool. Nylon is widely used because its molecular structure imparts outstanding abrasion resistance, excellent recovery from high strains, and excellent dyeability, lending the polymer great versatility, performance, and durability. Polyamide 66 is made from hexamethylene diamine and adipic acid, whereas polyamide 6 is produced from caprolactam. While the two polymers are similar

in many ways, the less symmetrical molecular structure of nylon 6 results in a more open structure and lower crystallization potential.

Polypropylene (PP) has gained market shares over the last few decades because of lower raw material cost and the low technological entry barrier for fiber extrusion. Several carpet mills have started to produce polypropylene fibers in house (vertically or backward integrated carpet manufacture). (The same is true to some extent of Nylon 6 and polyester, but Nylon 66 yarns remain more difficult to produce, and their manufacture is still exclusively the preserve of specialized yarn manufacturers.) Polypropylene carpets have lower appearance retention than Nylon carpets, but are appropriate for less demanding applications. PP is usually produced as pigmented (spin-colored, solution-dyed) staple or BCF, since an economical route to a dyeable PP fiber has not been established.

Polyester fibers are available mainly in staple form; polyester is less versatile in dyeing and less resilient in carpet than Nylon.

Wool, which has traditionally been regarded as a premium fiber, tends (due to lower abrasion resistance) not to perform as well as Nylon in like-for-like carpets, though carpets made from wool tend to be more expensive. Wool quality can vary widely.

1.1.4. Fiber Design

Fiber design is at least as critical to carpet properties as polymer selection. The esthetic properties and performance characteristics of a carpet are determined by a variety of fiber-related factors. These include:

- Yarn thickness or size, which is measured for BCF yarns in *decitex* (dtex = g/10 km) or *denier* (g/9 km), and varies typically from 600 to 2500 dtex (as singles yarns). Staple yarn thickness is measured in Nm (metric count, m/g) or c.c. (cotton count, 840 yards/lb).
- The thickness of individual filaments within a yarn bundle, often referred to as dpf or decitex per filament.
- Fiber cross section (Fig. 1), dictated by the shape of the extrusion hole and various polymer and production parameters. All melt-spun fibers can be shaped at extrusion, but this is best exploited by the polymer types most able to withstand abrasion.
- Luster or brightness, most commonly controlled by the addition of a delustrant (usually TiO_2) to the polymer, but also affected by fiber cross section and by voids, striations, and spherulites, which may be induced in spinning, and by crimping.
- Character and degree of bulk or crimp

Fiber design and its impact on carpet properties is further discussed in Section 1.2.

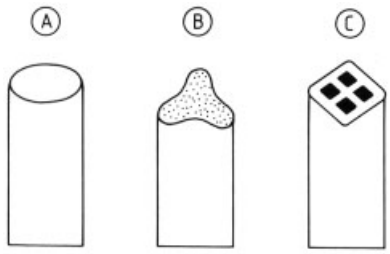

Figure 1. Carpet fiber cross sections
A) Round; B) Trilobal; C) Square hollow filament

1.2. Yarn Preparation

The short, crimped staple fibers must be spun together to form a staple yarn before further use. Both staple and BCF yarns can then be tufted directly as singles yarns into carpet from producer packages. Alternatively, a number of further processing steps can first be employed, depending on the esthetic and performance characteristics desired in the carpet. These include twisting and heat-setting, and intermingling.

1.2.1. Staple Yarn Processing

Staple yarn production involves opening of the fiber bales, possibly blending with other staple fibers, carding (to align the individual fibers), drawing (drafting), and spinning or twisting (to lock the fibers into the final yarn).

The use of staple allows the blending of fibers of different cross section, luster, or dyeability. This relatively intimate blending at the fiber level provides different styling options from those achieved by plying or twisting different yarns, as described below. Even different materials are blended to obtain desired carpet properties. For

example, an 80/20 wool/nylon blend provides markedly improved wear performance versus 100 % wool. Blending, of course, can also have a beneficial effect on the uniformity of the yarn produced and therefore the final carpet.

1.2.2. Twisting, Heat-Setting, and Intermingling

Depending on the desired carpet esthetic and performance characteristics, yarns of the same or different lusters, decitexes and/or dyeabilities may be combined by using a twisting and heat-setting process or an air entanglement/intermingling process.

Twisting. Two (or sometimes more) singles yarn packages are plied and twisted around each other to form a twisted yarn ready for heat-setting. *Direct cabling* (Volkmann, Verdol: see Fig. 2) is the most common twisting technique.

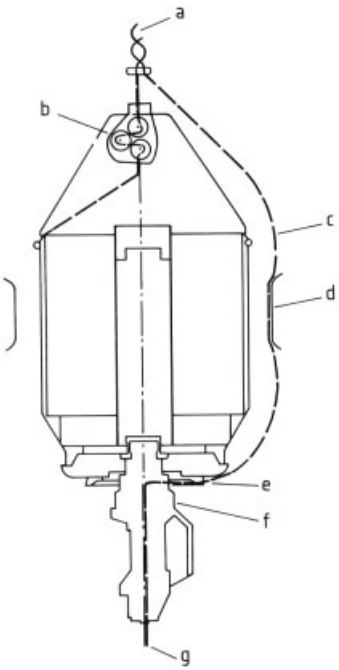

Figure 2. Operating diagram of a Verdol cable twister
a) T_3, plied yarn; b) Bucket yarn tensioner; c) Creel yarn balloon; d) Balloon restraining ring; e) Storage disk; f) Spindle; g) T_2, creel yarn; Tension (T) is adjusted via weighted disks or other devices at the package

Twisting has several purposes. While simple plying achieves a desired total yarn decitex, twisting provides the tuft integrity (tuft definition, texture) which is characteristic of saxony and hard twist (friezé) carpet styles. The amount of twist put into the yarn influences not only the initial appearance or texture of the carpet, but also the retention of this texture, which degrades in use according to the severity of traffic exposure. Increased twist level, however, increases yarn cost and results in loss of carpet bulk because of compression of the yarn crimp. The twist level used for a given style is therefore a balance of many factors. The yarns twisted together are usually identical, but dissimilar yarns may be paired to achieve additional styling flexibility.

Heat-Setting. The twisted yarn structure is sufficiently stable to allow winding onto packages, but it must be heat-set to lock in the twist and withstand the stresses of carpet dyeing and finishing and the traffic received in use. Also, yarn ″leanness″ can occur because the crimp has been pulled out in the various winding and packaging operations. If the yarn is heated while relaxed, much of this crimp, and hence yarn bulk, can be recovered. Therefore, both bulk recovery and setting of the structure are necessary so that the yarn remains stable in its intended end-use area.

For Nylon 66 or 6, setting can be achieved in a hot dry atmosphere at 180 – 220 °C or, alternatively, in saturated steam at 120 – 140 °C. Machines made by Suessen and Superba embody these two continuous heat-setting processes, which have now largely displaced the earlier autoclave batch process. BCF is primarily Superba steam set, while Suessen hot air setting is the preferred method for staple in North America.

Intermingling is another technique used to increase styling options in carpet design. The effect achieved is different to that obtained by twisting and heat-setting. Two or more singles yarns are unwound and passed together through a small chamber to which compressed air is applied. The yarns become intermingled or comingled with each other and are then rewound as a single yarn. They do not require subsequent heat-setting. Intermingling is carried out predominantly with BCF yarns, and more often to

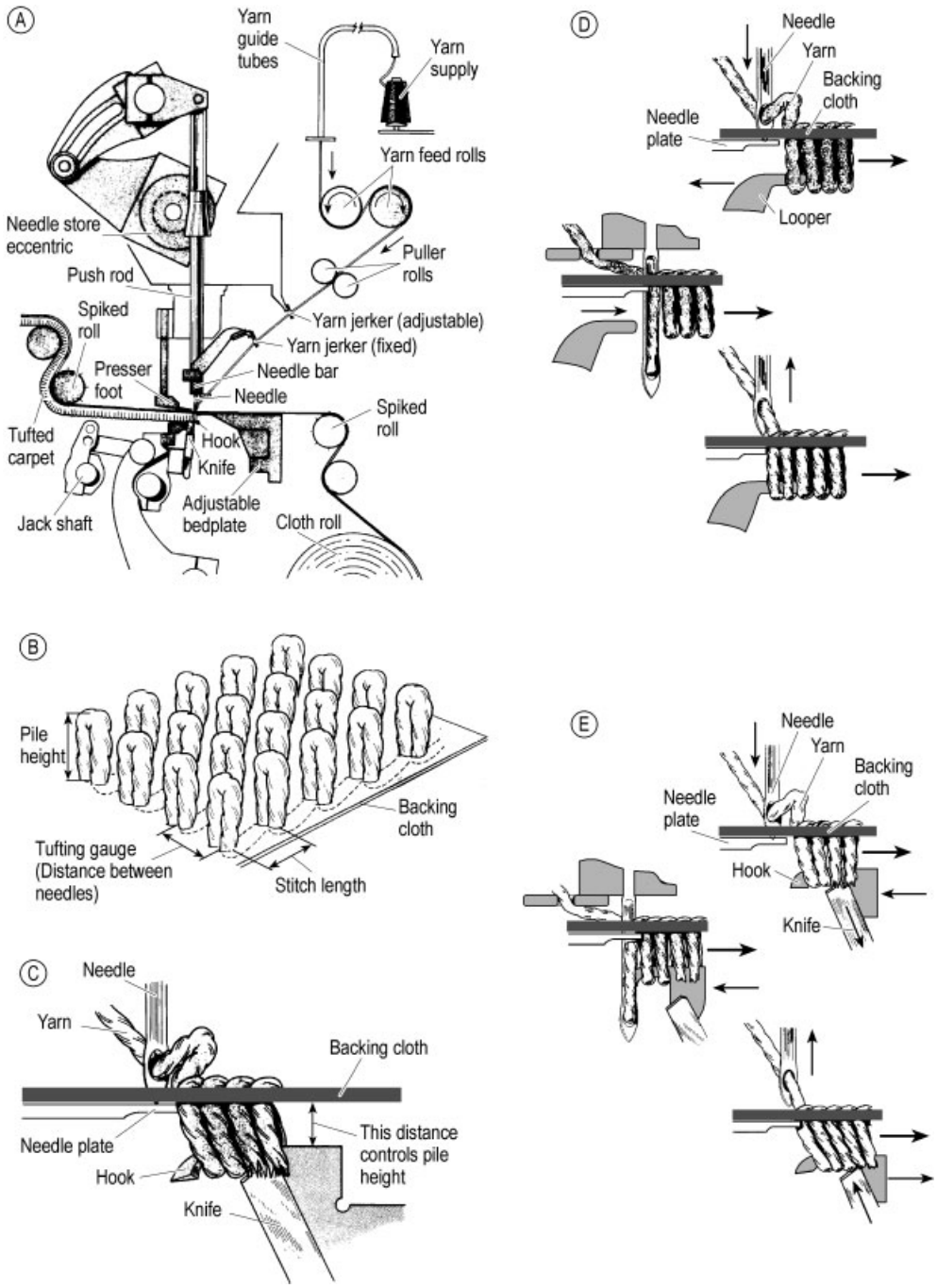

Figure 3. Production of tufted carpet
A) Cross section through a tufting machine, illustrating machine components and yarn path; B) The number of stitches per unit length in the tufting direction, the distance between needles, and the yarn thickness determine carpet density; C) The pile height is determined by the distance between the loopers and the backing cloth; D) Loop pile production; E) Cut pile production
Reproduced with permission of Cobble

combine dissimilar rather than identical yarns, for example yarns with different dyeabilities.

1.3. Production of Textile Floor Coverings

1.3.1. Flooring Types and Manufacturing Methods

The major types of textile floor coverings are tufted, woven, and nonwoven. Further manufacturing techniques include fusion bonding, flocking, and knitting.

Tufting is currently the fastest, most economical, and most automated method for producing carpets (Fig. 3). Needles held on a needle bar are each threaded with a pile yarn and pushed through a primary backing fabric. Under the backing, a hook or looper catches and holds the yarn to the desired pile height (Fig. 3 A – C). For cut-pile yarn, a knife cuts the loops as they accumulate on the looper; for loop-pile yarn, the hook withdraws and leaves a loop (Fig. 3 D and E). Carpet density is determined by the number of stitches per unit length in the tufting direction, by the spacing between needles, and by yarn thickness.

Thicker yarns require coarser needles with larger spacing. Common spacings between needles (gauges) range from 0.16 cm for European fine velours up to 0.8 cm (1/16, 5/64, 1/10, 1/8, 5/32, 3/16, 5/16 inches). Pile heights vary between 3 and 50 mm. Pile weights consequently range from 300 to 2500 g/m^2. The average tufted pile weight in Europe is 580 g/m^2 versus ca. 1000 g/m^2 in North America (worldwide average 840 g/m^2). The needle bar of a tufting machine can be moved sideways, and this enables the creation of tuft patterns if the needles are fed with yarns of different types. A wider range of patterning options is available on machines with double sliding needle bars, full repeat scrolls, and other modifications (e.g., Graphics, Cobble Colortec, Tuftco).

The application of a latex lock coat to the underside of the primary backing is the preferred method of locking the filaments together and into the primary backing. A secondary backing is then usually applied to enhance dimensional stability, strength, and lay-flat stiffness and to protect the underside of the carpet. It provides the contact for fixing the textile floor covering to the ground on installation. In the past, residential carpets in continental Europe (low pile weights) usually had a foam secondary backing. Recent landfill regulations and restrictions, as well as efforts to simplify carpet construction for environmental reasons, have led to the introduction of textile layers (woven fabrics; nonwoven needled fleeces, often made from polyolefin or polyester with a content of recycled material), which provide similar underfoot comfort and thermal and acoustic insulation. These textile backings have largely displaced foam. In countries where piles are typically higher, where underlays are used (UK, North America), and in commercial installations, a very thin fabric provides the interface between the latex and the underlay or floor.

Woven carpets are produced by the Wilton or the Axminster method; the main differences between them are the design of the loom and the weave used. In Wilton production, the pile yarn is carried through the backing, whereas in Axminster, the pile yarn is inserted in the backing.

Flocked carpets and fusion-bonded floor coverings with an upright pile use melt or normal adhesives to anchor fibers or yarns to a backing (Fig. 4). *Needled floor coverings* (fleeces, nonwovens) are produced by needle punching with optional subsequent fiber bonding. They can be homogeneous or consist of more than one layer (heterogeneous).

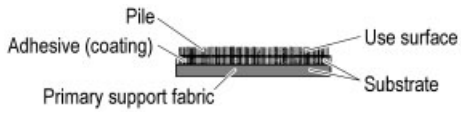

Figure 4. Flocked-pile carpet
Reproduced by courtesy of ISO

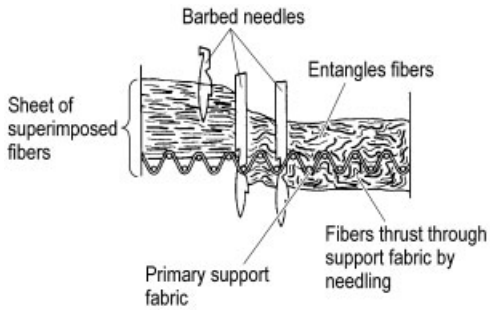

Figure 5. Manufacturing principle of needled textile floor covering
Reproduced by courtesy of ISO

In needle punching, layers of loose fibers are passed under a bank of barbed needles (Fig. 5). The needles catch filaments and entangle the layers of fibers with each other. Special techniques have been developed which allow the creation of needled pile floor coverings (e.g., Dilour).

In thermobonding, contact points between fibers are melted together by heat treatment. This requires fibers with low melting point, core-sheath two-component fibers with low melting polymer sheaths, or melt adhesives.

Bonding compounds (e.g,. latex) act as adhesives. Needled floor coverings can be fully impregnated or back coated.

A backing layer can be used to add volume, enhance stability, allow for easier installation, etc.

Fleeces are used as inexpensive commercial floor coverings in Europe and feature in special applications such as outdoor and automotive carpeting.

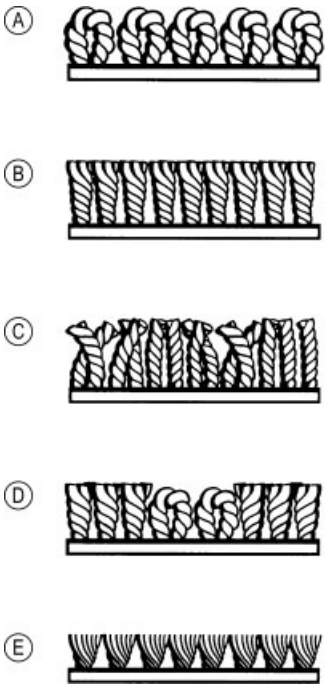

Figure 6. Types of carpet construction
A) Loop pile; B) Saxony; C) Friezé (twist); D) Cut and loop; E) Velour

1.3.2. Carpet Constructions

Carpets can be grouped into two major construction classes: loop pile and cut pile (Fig. 6).

Loop pile carpets (Fig. 6 A) can have level pile surfaces, patterned loops, or loops with two or three different pile heights to give a textured effect. Styling beyond pile height and carpet density is achieved by coloration of the carpet or yarn. Loop pile carpets, which achieve good durability by virtue of the loop construction and therefore do not require twisted, heat set yarns, are usually produced by direct tufting of BCF singles or intermingled yarns. However, many new loop styles now use twisted and often heat set yarns.

Cut pile carpets can be further subdivided into a variety of different constructions. These include saxony, hard twist (friezé), cut-and-loop, shag, velour (velvet), and COC. *Saxony carpets* (Fig. 6 B) use twisted, heat-set yarns having 1.4 – 2.4 turns per centimeter (3.5 – 6 turns per inch). This results in regular pinpoint tufts. *Friezé* (*hard twist*) carpets (Fig. 6 C) contain more highly twisted yarns, which are kinked before the heat setting process. This leads to irregular tuft-end formation on yarn cutting (during tufting) and variable light reflection from either bent or upright fibers at the pile surface. In *Cut-and-loop* carpets (Fig. 6 D), some of the tufts are cut and others remain whole loops, forming a pattern. *Shags* are high pile, very low density saxony carpets. The pile height exceeds 2 cm, but the individual tufts lie flat rather than standing upright. Carpets of *velour* (velvet) construction (Fig. 6 E) use staple yarns with little or no twist. The yarn stands upright in the carpet, and adjacent tufts are not visible as single entities, so that the surface appears uniform. Velours form an important segment of the residential market in continental Europe. *COC* is an acronym for cross-over-cut and refers to constructions in which Nylon singles yarns of different dyeabilities or yarns of different color are directly tufted with a sliding needle bar to form a pattern. They are found primarily in the continental European commercial sector. Velours and COCs are examples of the direct tufting of staple and BCF singles yarns, respectively. In the United Kingdom and North America, by contrast, staple yarns are twisted and heat set and used to make the other carpet structures described above. The majority

of BCF yarns used for cut-pile are also twisted and heat set before tufting. The vast majority of staple fibers are used in cut pile carpets.

1.3.3. Coloration and Finishing

Of all manufacturing operations in a carpet mill, coloration and finishing have perhaps the most significant effect on the visual and tactile properties of the end product. *Coloration* can refer to dyeing or to printing of white (typically Nylon) carpet material. Both continuous and batch dyeing processes are in commercial operation. Depending on the fiber selection, plain colors or multicolored effects can be achieved from one dye liquor. In a printing process, multicolored designs/patterns are produced by screen or spray printing. *Finishing* can refer to the application of a chemical aftertreatment, such as a protective stainblocking or anti-soil treatment. Shearing of the pile to improve the final surface uniformity of the carpet is an example of a mechanical finishing process. The last finishing step for tufted carpets is the application of a latex lock coat and then a secondary backing.

Dyeing. Tufted carpets made from white Nylon yarns are being increasingly dyed on continuous dyeing machines. Quick-change color applicators (Küsters, Fleissner) are used to speed process changes. In the continuous process, dye liquor is poured evenly over the carpet running below. After color application, the carpet runs into a steamer with vertical loops (festoons), and is heated with saturated steam for 2 – 6 min, during which the dyes are fixed onto the fiber. Water rinsing, hydro-extraction, and drying follow. In the less productive batch method, carpets are dyed in winches (becks). Here, a section of carpet runs in a loop through a full dyebath with temperatures up to boiling point for 40 – 60 min.

Acid leveling dyes are used for most Nylon carpets in residential and commercial end uses. This class of dyes allows the best color uniformity with good fastness properties in later end use. For combinations of standard and chemically modified Nylon (differential-dye yarns), which allow multicolor effects from a single dyebath, acid leveling, acid contrast, and cationic dyes are used. Metal-complex dyes are mainly used for automotive carpets, which require the highest standards of dye lightfastness, only achievable with this class of dyes.

Standard polyester and polypropylene are not commercially dyeable on current piece-dyeing equipment at atmospheric pressure. To avoid high-temperature dyeing ($> 100\,°C$ at > 0.1 MPa), carrier-free dyeable polyester, which is dyeable with disperse dyes, has been developed, and offers at least a limited styling flexibility. Most PP is colored by addition of pigments to the polymer melt during yarn manufacture.

Printing. White tufted Nylon carpets can be printed on flatbed screen printing machines or rotary screen printing machines, which allow the highest production speeds (30 – 50 m/min). The spray printing machines, Chromojet (Zimmer) or Millitron (Milliken), similar in principle to an ink-jet printer, run at lower speeds but are more flexible in terms of design possibilities and design change. Because the risk of poor color uniformity is much lower in printing, many kinds of acid and metal-complex dyestuffs are in use for standard prints. For the widely used resist and displacement printing techniques, specially selected groups of acid and metal-complex dyes are used, in combination with appropriate resist and blocking chemicals.

After color application the prints are run through a horizontal steamer for 3 – 6 min in saturated steam for color fixation, followed by water rinsing, hydro-extraction, and drying.

1.4. Carpet Properties and Performance

1.4.1. Carpet Esthetics

Color, pattern, and structure are the most obvious elements of the *visual esthetics* or appearance of a carpet, and are dictated by polymer/fiber choice, carpet construction, yarn choice (BCF/staple, singles/twisted, heat-set/intermingled), and coloration. Further important visual factors can be described in terms of luster, sparkle, glitter, sheen, clarity, brightness, dullness, chalkiness, etc. These are influenced by fiber design (Section 1.1.4), the shape of the fiber cross section (Fig. 1), mattness/brightness, crimp, yarn, and filament decitex.

The *tactile esthetics* of a carpet are influenced by many of the same factors; filament and yarn decitex (thickness/size), twist level of the plied yarn, crimp level of the filaments, and carpet construction, including pile height and stitch density.

1.4.2. Carpet Performance

The visual and tactile esthetics listed in Section 1.4.1 describe a new carpet. The performance characteristics of a carpet in use are also very important, and requirements vary from end use to end use.

Appearance Retention. Carpet appearance retention or "newness" retention in use has several aspects. It includes the loss of texture due to traffic (leading to damaged and opened tufts and consequent deterioration of surface texture), fiber loss, and recovery from compression after a weight (e.g., furniture) is removed. Appearance retention also covers changes to colors and patterns, whether due to fading (poor dye lightfastness) or blurring (due to deterioration of surface texture of the carpet). Further end use specific requirements include, for example, the ability to withstand the high downward and rotational forces exerted by castor chair wheels (office environments), and bleach resistance (hospitals).

Appearance retention depends on the nature of the pile fiber (basic molecular structure as well as subsequent processing), on carpet construction, and on carpet installation. Fiber material plays an important role in all mechanical aspects, whereby Nylon 66 has the best performance and is hence the fiber of choice for high-traffic commercial installations and premium residential carpets. The fiber cross section, delustrant content, bulk level, yarn decitex, and decitex per filament are significant factors. For example, fiber cross section and filament decitex are critical to wear performance. Carpet construction and the appropriate choice of yarn are also important. The floor performance of velour and loop-pile carpets is dependent on the density of the construction, determined by pile yarn mass and height. Using a bulkier yarn also increases the apparent density. The tightness of the yarn structure is a significant factor as well. Performance can be improved by increasing the twist or entanglement of the yarn. In saxony carpets, the most important factor is the twist level of the yarn. A low twist level is ca. 1.4 turns per centimetre (3.5 turns per inch) in singles yarn and in plied yarn. A minimum of two turns above primary backing is recommended for good appearance retention, which for cut-pile carpets like saxonies is most apparent in the tip definition of the tufts.

Pattern and color contrast of a textile floor covering also influence appearance retention by hiding physical changes in the surface structure.

The topic of appearance retention is complicated by the fact that some of the factors listed directly influence the amount of physical change that occurs in the carpet in use, while others affect the degree to which the change is visible, and some affect both.

Static electricity can be generated by the separation of the shoe sole and carpet pile material when walking on a carpet in a dry room climate. The increase in body voltage per step depends on the charge generated between the separated materials (rate of charge vs. rate of dissipation) and on the relative humidity, which has a marked effect on carpet resistance. This phenomenon is not limited to synthetic carpet fibers, but occurs with wool as well. Significant body voltages can be created in several ways, including by rising from furniture. Dissipation of charges (decay rate) is most effectively achieved by the addition of antistatic (conductive) filaments to the pile fiber in combination with a conductive lock coating. Electric charge can then be dissipated from the body through the carpet, even under dry conditions, preventing build up of voltages likely to be perceived as electric shocks. An alternative, but less durable solution for carpet installations in low-traffic areas, is the use of antistatic topical treatments. Computer rooms require particularly good antistatic protection.

Soiling refers to dirt on the carpet (rather than staining) and is basically a surface-transfer phenomenon in which dirt from shoe soles is transferred to the fiber surface of the carpet and vice versa. Airborne dirt usually plays a minor role. Dirt consists mostly of a blend of small inert particles, such as silicates, phosphates, oxides, and carbonates, held together by various amounts of

fatty organic materials and moisture (agglomerate size ca. 2 μm). It is first deposited at the top of a carpet but gradually migrates into the pile as traffic exposure continues.

Any fiber property that affects the size and nature of the carpet's exposed fiber surface will have a direct effect on soiling. This is the reason why coarse fibers with smooth surfaces and low surface area, in shapes that cannot trap soil, are best suited to reduce the soiling propensity. The chemical affinity of dirt particles for carpets is successfully reduced by surface application of fluorochemical coatings (e.g., polyfluoroalkylurethanes).

In addition to reducing the amount of soil deposited per traffic exposure, the visibility of soil can be reduced by proper carpet color selection, and by making fibers less transparent by addition of delustrants to increase light scattering. A particularly effective way to hide soil through light scattering is to incorporate voids into the fibers. Hollow fibers with low surface area and three to six voids (e.g., Fig. 1 C) are therefore preferred for commercial carpet installations where high traffic levels are expected.

Staining. Another feature of carpet fibers is their *staining* propensity. Nylon, which by virtue of its chemical nature is the most flexible fiber for all coloration methods – an advantage in offering consumer choice and at the production stage – is more prone to staining by, for example, food acids and beverages than some other substrates. Stain-resist or stain-blocking systems are applied which elevate Nylon's stain resistance to the level of polypropylene and polyester. Even coffee stains can be avoided with suitable systems [2].

Flammability of carpets is largely determined by fiber material, carpet construction (the denser the better), latex fillers, and secondary backing. Of all pile materials, wool can most easily be treated with flame retardants. Among the synthetic fibers, Nylon 66 performs best due to its high ignition temperature, whereas polypropylene burns most easily. This is a further reason for the limited use of polypropylene fibers in commercial installations. Carpets in transportation (particularly aircraft) end uses have very stringent flammability requirements.

Color Fastness. In addition to dye lightfastness, wash and rubbing fastness, resistance to fading in the presence of high NO_x and ozone levels is important for some end uses or geographical regions. Resistance to high sunlight and temperature exposure is critical in automobiles.

Health and hygiene issues carry increasing weight in modern society. Textile floorings have variously been blamed for poor indoor air quality, the propagation of dust mites (to the excreta of which, some people are allergic) and the harboring of other allergens and microbes. Many recent studies, however, show carpets to have a positive influence. For example, in combination with adequate and regular cleaning, carpets act to trap dust out from the air, reducing the amount of breathable dust (and potential allergens) in the atmosphere, thus improving air quality. Due to the relatively low humidity levels in carpets, they play a very limited role as breeding grounds for dust mites, which prefer bedding and other furniture. Studies conducted on Nylon 66 have demonstrated that it is essentially nontoxic through both the expected dermal route to exposure as well as by other possible but unlikely routes (oral and inhalation). Nylon 66 will not support mildew or bacteria, nor is it eaten by moth larvae [5]. The role of carpeting in the prevention/mitigation of noise pollution, as a safe, non-slip surface, and as a thermal insulator is increasingly recognized.

1.4.3. Test Methods

A system of carpet tests and classification schemes has been developed. End use recommendations for particular carpets are derived from these.

While a global classification for textile floor coverings is still pending, a common standard (EN 1307) has been agreed upon within Europe. This involves certain basic requirements such as

- Acceptable color fastness (to light: ISO 105-B02; to wet and dry rubbing; ISO 105-X12; to water: ISO 105-E01)
- Fiber bind [fuzzing and loss of mass (fiber loss), in preparation: ISO 11856]

It also introduces Use Classes (1 – 4) and Luxury Rating Classes (1 – 5).

Use Class reflects the resistance to wear (fiber loss; Lisson test ISO 12951 and appearance retention (Vettermann drum or Hexapod drum: ISO 10361, assessment of changes in appearance: ISO 9405). The Luxury Rating depends on carpet type (tufted, woven, flocked), thickness of pile (ISO 1766), mass of pile per unit area (ISO 8543) above primary backing, and number of tufts per unit area (ISO 1763).

Further properties are listed in EN1307 for particular end uses. These include:

- Castor chair suitability (EN 985, ISO 4918 in preparation)
- Electrical properties (walking test: ISO 6356, and vertical resistance: ISO 10965, at 25 % relative humidity)
- Acoustic properties (acoustic insulation from impact noise: ISO 140-8; acoustic absorption: ISO 354)
- Thermal resistance (ISO 8302)
- Suitability for use in periodically humid conditions such as bathrooms (dimensional stability: ISO 2551; higher color fastness to wet and dry rubbing: ISO 105-X12)
- Suitability for use on stairs (ISO 13746, in preparation)
- Fraying behavior (ISO 10833, in preparation).

Globally recognized tests do exist for soiling (Kappasoil ISO 11378-1 and drum test ISO 11378-2) and are in preparation for staining (ISO 13750) and flammability (Radiant Panel ISO 9239-1). For appearance retention, there is more reliance in North America on floor or "walker" trials (in tandem with laboratory tests like the Vetterman Drum) than is the case in Europe.

Needled floor coverings have their own classification system in Europe (EN 1470 needled pile floor coverings: EN 13297).

1.5. Disposal and Recycling

The majority of used carpets are currently still disposed of in landfill, though some are now incinerated with energy recovery. Carpets are a complex and intimate mixture of many different and incompatible materials: various polymers, adhesives, natural fibers, chalk, dyes, and dirt resulting from use (silicates, metal oxides, and organic matter). Collection and identification and the nature of the product make effective recycling anything but simple, but new legislation and improving process economics are powerful forces for change.

Several companies and major industry consortia in North America and in Europe have invested considerable effort in recent years, and significant progress has been made in the areas listed below:

- Some fiber manufacturers include melt-recycled plastics in new fibers.
- Carpet construction is becoming simpler and more recovery-friendly with the substitution of foam backing by more easily handled textile fleeces (which themselves can include recycled material)
- The polymer type in carpet pile and/or backing can now be identified with a new generation of spectroscopic analysers, either online or with portable instruments.
- Mechanical separation processes have been developed which are capable of separating the different carpet materials with varying degrees of purity. These can then be recycled, usually through a remelting process and generally to lower grade end uses (car parts, synthetic lumber, etc.). There are several such recycle programs operational in both North America and Europe.
- Chemical recycling is now available at least at pilot plant level for nylon polymer types in post-consumer carpet waste [6]. Polyester can also be depolymerized, but there is as yet no process for carpet waste. PP cannot be depolymerized into its constituent building blocks.

A number of pilot collection and recycle schemes are operating in North America and Europe and include collection, sorting, separation, remelting, chemical recycling, and energy recovery. Significant advances are to be expected by 2010.

2. Resilient Floor Covering

2.1. Introduction

Resilient flooring is a family of flooring products that includes such materials as linoleum, asphalt tile, rubber and vinyl flooring. The term resilient refers to the ability of these flooring materials to deform and recover from the compressive forces caused by the dynamic action of foot fall, cart wheels and casters, and the static loads of appliances and furnishings. Resilience distinguishes this group of flooring products from carpeting, wood, ceramics, and stone.

Types of Resilient Flooring. Resilient flooring is produced in two basic formats: sheet and tile. inthe case of vinyl flooring each has, with some exceptions, a distinctly different composition and manufacturing process, whereas rubber flooring employs essentially the same process and formulation for both forms.

Economic Aspects. The market for flooring is divided into two distinct segments: residential and commercial. The total area of existing residential floor space in the United States in 1998 was about 17×10^9 m^3, of which an estimated 16.2 % was covered with resilient flooring material, the remainder being covered with carpet, wood, ceramic tile, etc. Most residential resilient flooring is used in kitchens and baths, and some in foyers and family rooms.

The total commercial and institutional U.S. floor space in 1998 was approximately 36 billion square feet (3.4×10^9 m^3, excluding warehousing, manufacturing areas, and parking garages. Resilient flooring, both vinyl and rubber, constituted about 25.5 % of this area, carpet 56.5%, and others (stone, wood, ceramic, etc.) 18 %. Market segments for resilient flooring in the commercial and institutional area include mercantile, health care, education, light industry, offices, and banks.

The total apparent hard surface flooring consumption for 1998, including U.S. manufacturer shipments and imports, but excluding exports, amounted to $\$5.4 \times 10^9$, of which resilient flooring constituted 44.2 %.

History. [7–10]. The resilient flooring produced today evolved from painted fabric or floor cloth that was laid over board flooring to provide a decorative, washable surface and to keep out drafts from the cracks between the floor boards. Several attempts, dating back as far as the 1600s, were made to establish an industry to manufacture floor cloth and oil cloth. The invention of oil cloth is recorded in a 1627 patent. A patent from 1763 describes a floor covering made by coating fabric with a mixture of resin, tar, Spanish brown, beeswax, and linseed oil.

In 1844, a floor cloth made from India rubber, sawdust, and cork, named Kamtulicon, was introduced. At about the same time, waterproof fabrics coated with rubber became quite popular and drove the price of rubber up to the point where it became uneconomical for flooring use. A patent was issued to Frederick Walton in 1860 for an oxidized linseed oil intermediate called linoxyn which made possible the production of Linoleum, a name coined by Walton to differentiate it from the rubber-based Kamtulicon. In 1864 Linoleum was commercially introduced in England by Walton, Taylor & Co. This was essentially the beginning of the modern resilient flooring industry.

Linoleum is based on a binder system of oxidized drying oils, resins, and gums, filled with ground cork, wood flour, whiting, and pigments. Although it has excellent wearing characteristics, linoleum suffers from degradation by alkaline hydrolysis and is therefore unsuited for applications where alkaline moisture is present, e.g., on concrete subfloors in contact with the ground. It may also be damaged by frequent washings and stripping of wax polishes. Because of these limitations and the versatility of PVC-based flooring, its production was discontinued in the United States in 1973.

In those applications where linoleum was not suitable because of its sensitivity to moisture and alkalinity, such as on subgrade concrete, asphalt tile was used. This product was based on asphalt binders or, for lighter colored tiles, coumorone indene or hydrocarbon resins, filled with asbestos and limestone. Asphalt tile is essentially obsolete today.

2.2. Linoleum

Linoleum is experiencing a resurgence in popularity in the North American market because

of its use of natural and renewable raw materials and its unique design capabilities. Expected sales in 1999 are about 3 % of the total U.S. resilient-flooring market [11].

Production. Linseed oil and dryers (i.e., catalysts for oxidation) are combined and partially oxidized in a reactor. At a certain stage in the process, melted resins are added to form linoleum cement. When the right consistency and degree of oxidation is achieved, the linoleum cement is extruded, cut into billets, dusted with wood flour, and further aged to continue the oxidation process. It is then blended with the fillers and pigments in a high-intensity mixer and calendered or milled into sheets. If the finished product is to be a solid color flooring material or desk top material, it may be calendered directly to a backing of jute or a glass scrim. To form variegated colors, accent sheets are sliced or diced to appropriate sizes and calendered with a base color to form a jaspé. If this step gives the finished design, the material is calendered to a backing. If a marbled pattern is desired, unbacked jaspé sheets are cut into strips and cross-sheeted parallel to the calender nip and calendered. The resulting sheet is then laminated onto a backing.

The linoleum must be cured. It is festooned in tall ovens with about 18 – 28 m in each loop and cured at elevated temperature for 2 – 3 weeks. The cured product is finish coated with acrylic or nitrocellulose lacquer, inspected, and cut to length, generally the length of one loop.

Historical processes include inlaid linoleum, in which die-cut shapes from various colored and patterned sheets were laid into matching die-cut spaces in base sheet; and stenciled linoleum with "scratch mix" granules, mottled, and placed into pattern position through thick metal stencils. The scratch mix was produced by running a pin-studded roll against the linoleum composition while it was worked on a two-roll mill.

2.3. Vinyl Flooring

Although rubber flooring enjoys a significant market share, the dominant types of resilient flooring manufactured today have a binder system based on poly(vinyl chloride), PVC. This material is well suited for this application because: (1) it can tolerate high levels of inert filler, (2) it is clear and colorless, (3) it is inherently fire resistant, (4) it can be effectively plasticized to a wide range of flexibilities, and (5) it is not adversely effected by moisture or alkali. When moderately filled and plasticized it retains a high proportion of its inherent strength and durability.

Vinyl or PVC flooring was first introduced by Dynamit Nobel in 1934 under the trade name Mipolam [10]. Today there are many types of vinyl flooring in the market, both in sheet and tile form.

The U.S. custom is to refer to these products as vinyl flooring, whereas they are commonly called PVC flooring in Europe. This has led to some confusion in the American market with imported flooring products, which are sometimes thought to be based on different polymers. These terms should be considered synonymous in the flooring trade.

2.3.1. Raw Materials

Poly(Vinyl Chloride) Resins. Both vinyl chloride homopolymers [→ Poly(Vinyl Chloride)] and vinyl chloride/vinyl acetate copolymers are used in the production of resilient flooring. As a general rule, vinyl chloride homopolymers are used in sheet flooring and solid vinyl tiles, while vinyl composition tiles usually contain copolymers of vinyl chloride and vinyl acetate.

The weight-average molecular mass M_w of the resins most useful in flooring ranges from 40 000 to 200 000. The higher molecular mass polymers are frequently used in flooring wear layers because of their greater ultimate tensile strengths and abrasion resistance, whereas the lower molecular mass polymers, because of their lower melt viscosity, are most useful in producing foams for cushioned flooring and saturants for glass scrim reinforcements.

Flooring manufacture employs suspension- and modified bulk-polymerized PVC in processes requiring intensive mixing, dry blending, or calendering, and emulsion-polymerized resin in processes involving liquid coating of plastisols.

Stabilizers. As with any compounded poly(vinyl chloride) system, the flooring composition must be protected against degradation

from the effects of both heat and ultraviolet radiation during processing and during its service life. The stabilizers most commonly used in vinyl flooring are soaps of barium, calcium, and zinc, and certain organotin compounds. Lead and cadmium soaps have been used in the past but, although effective stabilizers, they are not widely used today because of environmental concerns and their tendency to form colored sulfides, a phenomenon referred to as sulfur staining). Epoxidized oils such as soybean oil, tallate esters, and other partially unsaturated oils are frequently used as costabilizers and secondary plasticizers in compounds stabilized with metallic soaps. Organophosphites are also frequently used as chelating agents and antioxidants in conjunction with metallic soap systems.

The particular blends of stabilizers used in flooring are proprietary to the manufacturers because of their subtle effects on other properties important to the production or performance of the products. An example of interaction is the accelerating effect by zinc soaps on the decomposition of the blowing agents used to expand the foamed layer in some vinyl flooring.

Plasticizers. The vinyl binder system in flooring is plasticized to provide flexibility and resilience, as well as to facilitate processing. Di(2-ethylhexyl) phthalate has been the most frequently used plasticizer for flooring because of its relatively low cost, permanence, color, and overall balance of performance properties. However, because of current controversy regarding long-term health effects of this plasticizer, it has been largely replaced in the United States by diisononyl, diisodecyl, and diisoheptyl phthalates.

Butylbenzyl phthalate, historically used for production of vinyl composition tile, is increasingly being used in plastisol coatings and foamable base. This is a very efficient, highly solvating plasticizer, and when used in plastisol coatings, viscosity-depressant plasticizers such as TXIB (2,2,4-trimethylpentane-1,3-diol diisobutyrate) and dodecylbenzene are frequently included [12].

Fillers and Pigments. With the exception of the clear wear layer in most residential and some commercial flooring structures, the stabilized and plasticized vinyl formulation, i.e., the binder, is mixed with varying amounts of inorganic filler to provide mass and thickness at a reasonable cost. Pigments are added to provide color, opacity, and design.

Calcium carbonate (limestone) is the filler of choice in flooring formulations because of its low cost, lack of color, and low plasticizer absorption properties [13]. Talc and clay are rarely used, because of their significant stiffening effect on plasticized vinyl. Asbestos was used in the past in vinyl asbestos and asphalt tiles, and some sheet flooring, but it is no longer used in flooring made in the United States.

Fillers provide not only bulk and mass at reasonable cost but also increased dimensional stability, opacity, resistance to cigarette burns, improved flame spread ratings, reduced smoke generation, and reduced memory and elasticity.

Pigments used in flooring must be heat stable, lightfast and resistant to change in color from alkaline moisture The preferred white pigment is titanium dioxide. Most colored pigments are inorganic, but certain colors, primarily the blues and greens and some yellows, are only available as organic compounds.

Other Compounding Ingredients. In many public facilities, flooring must meet certain code requirements with regard to fire and smoke properties. Specific additives may be used to reduce the flame spread and smoke generation of the products. Agents used for this purpose include alumina trihydrate, antimony trioxide, zinc oxide, boron compounds, phosphate esters, and chlorinated hydrocarbon plasticizers.

The foam layer in cushioned flooring is chemically expanded by the thermal decomposition of azobis(formamide) blowing agents. Some formulations may also contain processing aids and lubricants depending on specific processing methods and requirements.

2.3.2. Sheet Vinyl Flooring

Cushioned Rotovinyls. Cushioned rotovinyls are the most common form of sheet flooring on the market today and are used extensively in residential applications.

The structure (Fig. 7) typically consists of a backing of felt composition, glass scrim, or an

extensible vinyl film; a foam layer for cushioning and to facilitate pattern embossing; a rotogravure printed pattern, and a transparent or translucent unfilled vinyl wear layer.

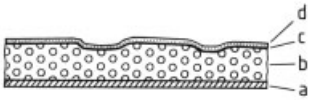

Figure 7. Typical rotogravure sheet flooring structure
a) Felt composition backing; b) Foam interlayer; c) Rotogravure print; d) Clear wear layer

Production. The production of cushioned rotovinyls depends largely upon the use of vinyl plastisols in which emulsion-polymerized PVC, blending resins, pigments, stabilizers, and other additives are dispersed in the plasticizer system. When heated to approximately 180 – 200 °C the plasticizer(s) solvate the resins to form a tough plastic film. Typical formulations employed are shown in Table 1.

Table 1. Typical formulations for some vinyl flooring structures

Ingredients	Formulation∗, parts by weight per 100 parts resin				
	A	B	C	D	E
Vinyl resin(s)	100	100	100	100	100
Plasticizer(s)	30	60	30	35	40
Epoxidized oil	5	5	4	5	
Processing aid					8
Stabilizers	2.5	2	2	3	2
Fillers and pigments		30	210	70	640
Blowing agent		5			

∗ A = clear coat for rotogravure flooring; B = foam coat for rotogravure flooring; C = filled wear layer for inlaid vinyl flooring and vinyl tile; D = homogeneous sheet flooring; E = vinyl composition tile.

Rotovinyl sheet flooring is built up of several layers of plastisol coatings applied to the carrier by reverse-roll or knife-over-roll coating. The backing felt or other carrier is coated with a foamable plastisol containing the blowing agent [typically azobis(formamide)], which decomposes at elevated temperatures to create the expanding gases. In some structures a nonfoamable base coat may first be applied. After coating, the foamable plastisol is gelled (partially fused) at a temperature below the decomposition temperature of the blowing agent to form a smooth printable substrate (Fig. 8).

The gelled foam coat is then decorated with vinyl inks by direct or offset rotogravure printing to create the desired pattern. Registered embossing is usually achieved by incorporating agents in selected inks that interfere with the expansion of the foam coat [14]. There are also sophisticated systems for mechanically embossing in register.

A transparent or translucent plastisol is coated onto the printed gel coat, and the entire structure is heated to a temperature adequate for complete fusion of the plastisols and expansion of the foam layer to produce the final product. Some processes use mechanical embossing techniques to create overall textures in this final step.

Characteristics and Uses. The most common width of cushion rotovinyls is 12 feet, (3.66 m) although 6, 9, and 15 foot (1.83, 2.74, and 4.57 m) widths are also produced. The backings, foam, and wear layer are available in various gauges depending on the intended applications. The composition and thickness of the components are descrided in ASTM F 1303-97 [15]. The pattern design potential is virtually unlimited because of the versatility of rotogravure printing.

Cushion rotovinyl is today the most popular resilient flooring product in the residential market. Most products are promoted as not requiring waxing for normal use and many have surface coatings of polyurethane, urethane acrylate, or other proprietary cross-linking materials for improved stain resistance and sustained gloss retention. These coatings may be cured by ultraviolet light or conventional heat in the final stage of production.

The foam interlayer provides underfoot comfort and warmth. It is also fairly effective in reducing the transmission of impact sound from the floor to the space below, and combined with appropriate subfloor structures, can be used to achieve certain impact-isolation class requirements in multifamily housing units.

Inlaid Sheet Vinyl Flooring with Backing. Products in this category have a felt or filled vinyl backing with an inlaid vinyl wear layer of varying thickness, depending on intended end use [15]. The heavier gauges are used primarily for commercial and institutional applications (Fig. 9).

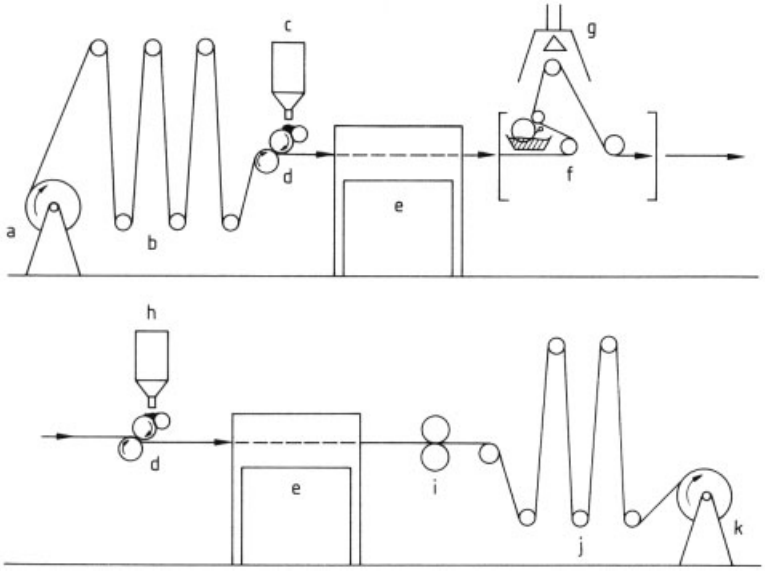

Figure 8. Rotogravure production line
a) Backing unroll; b) Infeed accumulator; c) Foamable plastisol base coat; d) Reverse roll coater; e) Gelling or fusing oven; f) Rotogravure print station (4 – 7 stations); g) Drying heater with fume hood; h) Clear coat plastisol; i) Mechanical embossing laminator (optional); j) Outfeed accumulator; k) Windup roll

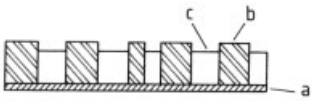

Figure 9. Typical inlaid vinyl sheet flooring structure
a) Felt composition backing; b) Filled vinyl composition; c) Unfilled vinyl matrix

Production. A typical formulation for the filled component of the wear layer is shown in Table 1, column C. The resins used are primarily suspension PVC homopolymer, but some copolymer may be used in accent colors to control graining. The primary plasticizers are generally of the phthalate type, although some products may contain phosphates and/or chlorinated hydrocarbons to improve flammability ratings. The filler is usually calcium carbonate.

The ingredients are mixed and fused in a high-intensity mixer or a reciprocating screw extruder and milled and/or calendered into sheets of plain or variegated color. These are then fractured, cut, or ground into various shapes and sizes depending on the finished design requirements. Different colors of particles are frequently combined to produce a mottled mix. To produce random inlaid design the particles are cast onto the backing and fused and consolidated by heat and pressure, with or without an unfilled vinyl matrix surround, in flat-bed or rolling presses. Deliberate imitative designs may be achieved by using a stenciling operation to position granulated color onto the carrier to produce a specific pattern.

Characteristics and Uses. Inlaid sheet vinyl flooring is normally supplied in widths of 6 – 6.5 feet (1.8 – 2 meters). The inlaid construction of these products makes them sturdy and durable. Their more massive wear layer, supported directly by the felt or filled vinyl backing, provides resistance to damage by cuts and punctures and thus makes them suitable for commercial applications.

These products have found widespread use in health-care facilities, schools, and institutions. They are used also in abusive residential applications such as dormitories and military housing.

Homogeneous Sheet Vinyl Flooring. These products may also be referred to as solid vinyl sheet flooring or PVC sheet flooring. They have been popular in Europe for many years, particularly for commercial and institutional

applications, and since the 1980s have found increased popularity in the United States.

Production. Homogeneous vinyl sheet flooring is normally manufactured by a combination of calendering, consolidating, and/or laminating processes. A generalized formulation is provided in Table 1, column D. The types of resins, plasticizers, stabilizers, and fillers are similar to those discussed under inlaid vinyl flooring.

The compounding ingredients are mixed and fused in a high-intensity, high-shear mixer, (Fig. 10 A, e) and sheeted through one or more two roll mills (f, g). Previously prepared accent colors are added to the web at one of the mill nips to produce sheets with the desired graining. The material is then calendered through a four-roll calender (i). If the resulting sheets are to be built up to achieve the required finished gauge, they may be laminated together in a continuous or intermittent pressing operation to obtain the desired thickness (Fig.10).

Characteristics and Uses. These floors are produced primarily as roll goods, but some manufacturers cut the stock into sheets, usually three feet (0.9 m) square, or into one foot square (0.3 m) tiles. Roll widths vary from four feet (1.22 m) to 6.5 feet (2 m). The most frequently specified thickness for commercial applications is 80 mil (2 mm), although 60 mil (1.5 mm) gauge is also produced by some manufacturers. Depending on the structure and process employed in production, the pattern and graining is fairly consistent throughout the thickness of the product. Some laminated structures, while being of solid vinyl construction, may have the pattern only in the top layer. Their construction and performance requirements are described in ASTM F 1913-98 [16].

This is the most durable type of vinyl flooring available today. It is formulated with fairly high molecular mass homopolymers and contains relatively little inert filler. This, combined with its thickness and through-color construction, results in excellent resistance to abrasion and other mechanical damage from punctures, cuts, and gouges. Its composition also provides good chemical resistance.

Because of these factors and the fact that this flooring is installed with heat-welded seams, it is used in commercial and institutional areas having rigorous performance requirements such as industrial, medical, and educational laboratories, and clean rooms; veterinary clinics; outpatient health care clinics; and in hospital procedure rooms, including surgeries.

Layered Composite Flooring (Heterogeneous Sheet Flooring). Versions of solid vinyl flooring becoming increasingly popular in the United States are composite structures having a clear wear layer of at least 20 mils (0.5 mm) in thickness over a printed pattern on a filled vinyl backing. The product may or may not be reinforced with glass scrim.

The compounding ingredients are similar to those mentioned for both residential rotovinyls and homogeneous vinyl sheet flooring. In those cases where the wear layers are produced by hot-melt calendering, no viscosity reducing plasticizers are necessary. The products are produced by a combination of calendering the individual layers, plastisol coating or saturating the fiber glass reinforcement, if employed, and laminating in a continuous belt laminator (Fig. 10 B).

2.3.3. Tile Flooring

Vinyl Composition Tile. (VCT) is the most widely used resilient tile product today. Because of the general concern over the use of asbestos, it has superseded vinyl asbestos tile (VAT).

Production. A typical formulation for VCT is shown in Table 1, column E. The vinyl resins normally used in VCT are vinyl chloride – vinyl acetate copolymers having a vinyl acetate content in the 6 – 13 % range. The copolymers may also be blended with PVC homopolymer to achieve certain processing characteristics. Processing aids, typically low-melting hydrocarbon resins, are used to lower the fusion temperature and to assist in the mixing process.

The plasticizer system consists of varying blends of dihexyl phthalate, diisoheptyl phthalate, diisononyl phthalate, and butylbenzyl phthalate. Epoxidized soy oil is also commonly used as a coplasticizer/stabilizer. Manufacturers carefully guard their actual formulations, particularly with regard to fillers and stabilizers because of their subtle, yet important effect on performance properties.

The ingredients are combined in a high-power, high-shear, heated mixer, (Fig. 11 a) such as a Banbury mixer to blend and fuse the in-

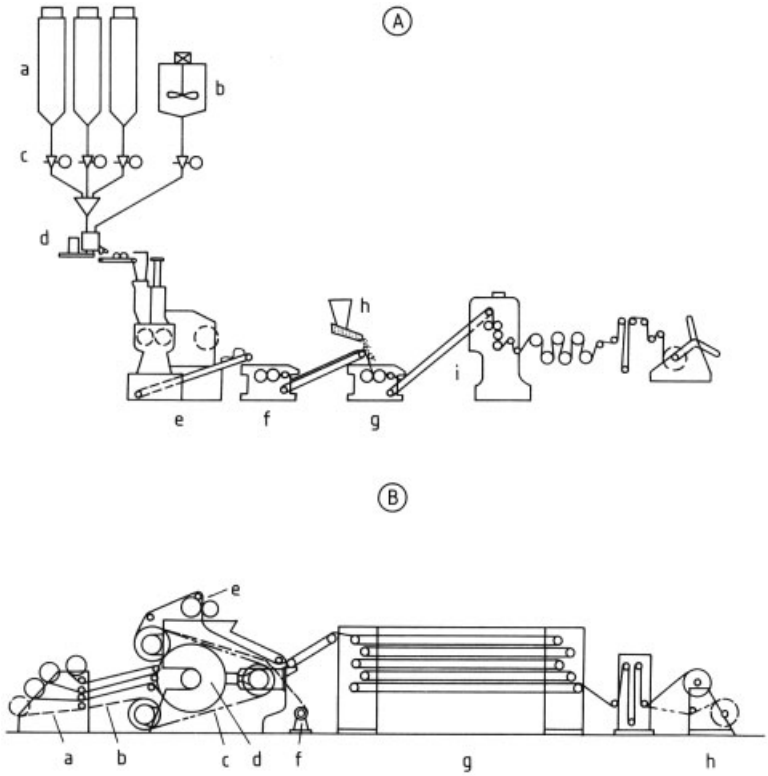

Figure 10. Production of homogeneous sheet vinyl flooring
A) Production line for calendered sheet vinyl flooring: a) Bulk storage silos; b) Plasticizers; c) Metering; d) Premixer; e) Intensive mixer; f) Two-roll mill no. 1; g) Two-roll mill no. 2; h) Accent color feed; i) Four-roll calender
B) Laminating system for calendered sheet flooring: a) Unroll; b) Carrier fabric; c) Press belt; d) Laminating drum; e) Embosser; f) Roll up for backing; g) Annealing and cooling chamber; h) Windup roll
∗ Adapted from [12, p. 27].

gredients together. A heavy doughlike mass is formed and dropped from the mixer onto a two-roll mill (c). Accent colors (b), of the same or similar composition are added to the mill nip in the manufacture of grained or jaspé sheet.

When the mixing is complete and the graining judged right, the stock is cut (doctored) off the mill roll and folded into blankets to create more uniform graining when passed through the sheeting calender (f). The continuous web exiting from the sheeting calender is passed through a second finishing calender (h) to reduce it to the final gauge and provide a smooth face. For the production of tile having a surface applied design, previously prepared accent chips are applied (g) to the face of the tile between the first and second calender. Through-grained tile may also be produced by casting pre milled and fractured color chips onto a carrier belt, heating in an oven or by radiant heater, and consolidating in the sheeting calender (f). Tile having an embossed surface texture is embossed by passage between another pair of rolls (i) with tooled embossing on the top roll.

After exiting the finishing calender or, if used, the embosser, a factory finish of wax or synthetic polish may be applied (j) to the hot sheet to provide uniform gloss, and prevent blocking of the tile in storage. The sheet is then cooled (k) and die-cut in a punch press (m) to the finished size. The cut tiles are then inspected and packed (n). Rejected tile and frame scrap from the punching operation is recycled (o) to the mixer.

Characteristics and Uses. Vinyl composition tile is produced in several gauges, sizes, and styles for differing intended end uses. For residential applications, VCT is offered in service gauge, which is approximately 1/16 of an

inch (ca 1.5 mm) thick. These tiles are frequently embossed and may contain surface applied patterns and grainings. Most residential tiles today have a printed decorative pattern with a clear vinyl wear layer laminated to a solid color vinyl composition tile backing and embossed in register. Many also have high-performance no-wax finishes, as described for sheet flooring (Section 2.3.2).

Most commercial tile is 1/8 of an inch (3.2 mm) thick, however 3/32-inch (2.4-mm) tile is also available for use in light commercial applications. Tile for commercial applications is usually of the through-grained variety rather than surface applied designs. The standard size of VCT is 12 × 12 inch (305 mm square), but other custom sizes may be available.

In the manufacture of vinyl composition tile, particular attention is given to the dimensional stability, size, squareness, and the quality of the edge cutting. These qualities are essential for successful installation of the tile floor. Structure and performance requirements of VCT are contained in ASTM F 1066-95a [17].

Vinyl composition tile is considered to be the standard or base-grade commercial finish flooring. It has the lowest relative installed cost and has established a good performance record for this use. The major outlet today is the mercantile market. It is used almost exclusively for the general floor area of grocery stores, supermarkets, and discount department stores. It is also used extensively in schools, health-care facilities, and to a lesser extent in offices, banks, and light industry.

Vinyl Tile. Vinyl tile may be referred to as "solid vinyl tile", "homogeneous vinyl tile" and even erroneously, "pure vinyl tile". The primary difference between vinyl composition tile (VCT) and vinyl tile is the binder content. Vinyl tile generally has a binder content in excess of 34 % of the formula weight [18]. There is no minimum binder content requirement for vinyl composition tile.

Production. The formulation of solid vinyl tile is very similar to that used for the filled component of inlaid sheet flooring or some homogeneous flooring. See Table 1, column C for a typical formulation.

Since there are many proprietary methods used for the production of vinyl tiles it is not possible to describe a specific manufacturing process here. In many respects the processes are similar to those used in the production of homogeneous sheet vinyl flooring (see page 17)

Characteristics and Uses. Vinyl tile is more varied in construction than is vinyl composition tile. Appliqué tiles consist of two distinct layers, a base layer in one color and a overlay cut to resemble brick or other pavers. Other styles have through-grained marble or mottle designs, while still others are solid-color pavers. The overall thickness is usually 1/8 inch and it is available in both 9 × 9 inch (0.2 m) and in 12 × 12 inch (0.3m) sizes.

Vinyl tile is considerably more flexible than vinyl composition tile and has different specification requirements, particularly with regard to recovery from indentation and flexibility. Because of its high binder content it has significantly higher resistance to abrasive wear and is more resistant to impact damage. Conversely, its higher binder content makes it less resistant to damage from accidental cigarette burns.

The initial cost of vinyl tile is significantly higher than VCT, but because of its flexibility, resilience, and high degree of abrasion resistance, this product is very suitable for areas subjected to high pedestrian and cart traffic. Many vinyl tiles are offered in imitative patterns of brick, slate, and ceramic pavers and are frequently used in mercantile settings to highlight special merchandising areas in the store.

2.4. Rubber Tile and Sheet

Rubber tile and sheet flooring was introduced in the 1920s and was widely used through the 1950s, particularly for applications unsuited to linoleum and asphalt tile. Its popularity declined through the 1960s and 1970s because of the relative success of vinyl tiles and sheet flooring. In the mid-1980s it made a comeback in certain specific applications, due partly to the popularity of the raised-disk design used for slip retardance in areas of heavy pedestrian traffic.

Production. Modern rubber flooring uses styrene – butadiene rubber (SBR) as the primary binder. Other rubbers that may be incorporated with the SBR include Neoprene and nitrile rubbers as well as ethylene propylene diene

monomer elastomers (EPDM). A typical rubber flooring formulation contains the rubber or rubber blend, a high-styrene resin as hardener, perhaps a hydrocarbon oil extender, fillers and pigments, and vulcanizing agents such as sulfur and zinc oxide. The preferred fillers for rubber flooring are clays and talcs because of their reinforcing nature. Limestone whiting may be used as an extender in some compounds.

The ingredients are blended typically in a Banbury mixer, followed by sheeting through a two-roll mill. The sheet from the mill, along with previously compounded accent colors is rolled into a "pig" that is fed to a four-roll calender; there, the accent colors produce the marbleized graining, and a relatively smooth sheet approaching the final gauge of the product is prepared.

For the manufacture of tile and nonroll goods, this sheet is cut into slabs that are press cured at elevated temperature for 10 – 12 min in multiple or stack presses. Roll goods are cured in a continuous press such as a Rotopress.

After curing, the material is allowed to age a day or two before cutting to the finished size. The back of the tile may be sanded or have embossed ribbing to provide better grab of the installation adhesive.

The wearing surface may be smooth or may have a molded embossing such as round disks, raised squares, or other configurations to furnish slip-retarding characteristics and provide special design effects.

Characteristics and Uses. Rubber flooring is very durable, with excellent resistance to abrasion, recovery from indentation, quietness under foot, and the best resistance to accidental cigarette burns of all the modern resilient flooring products. It also has better resistance to polar organic solvents, such as acetone and other ketones, esters, and cyclic ethers. However, it has poorer resistance to aliphatic and aromatic hydrocarbons and vegetable oils than do vinyl products.

Rubber flooring is frequently specified in light industry and laboratories, where its unique solvent resistance is required. It is also used in health-care facilities and schools, where its good wear resistance and resilience are needed. Raised-disk rubber flooring has found widespread acceptance in areas where wet slip resistance is required or where that particular decorative effect is desired.

Rubber flooring composition is also used extensively for molded stair tread, risers, and landings in public stair ways.

2.5. Specialty Flooring Products

Conductive flooring is used in areas where explosive or flammable vapors or dusts may exist, or where control of static electricity for sensitive electronic equipment is needed. These products obtain their conductivity by the use of conductive carbon black in the formulation which somewhat limits the color and design capabilities. Conductive flooring is manufactured to meet the requirements of the National Fire Protection Association (NFPA) for explosive or flammable environments [19].

Static-Dissapative Flooring. For the control of static electricity for electronics applications, flooring products are available that are less conductive than conductive flooring. The resistance to ground need only be in the range of 10^6 to 10^9 ohms, which can be achieved by incorporating quaternary ammonium compounds in the formulation.

Slip-retarding flooring is used on ramps or in areas where the floor will be subject to frequent wetting such as entryways, vestibules, food-preparation areas, etc. This flooring may have raised areas or deep embossing to help prevent hydroplaning when wet; alternatively, mineral particles such as carborundum, quartz, or other mineral aggregates may be incorporated in the surface to provide increased friction with shoe soling. Combinations of these two configurations may also be used.

3. Laminate Flooring

For some years high-precision manufacturing methods have been available for the production of flooring elements that can be laid – mainly as floating floors – with surfaces ready to be walked on. These can be products with wood surfaces (parquet flooring) or laminated elements.

A distinction is made between flooring elements that are laminated during manufacture and those which are combined with laminated materials [high-pressure laminate (HPL) or continuous pressed laminate (CPL)] in a separate operation. The former are produced exclusively on high-density fiberboard (HDF) substrates, whereas particleboard can also be used for the latter. Because nowadays very thin decorative laminate sheets can be produced, these are often indistinguishable from films.

3.1. Development

The first laminated floors were produced by Perstorp in Sweden in the 1970s. This major manufacturer of decorative laminates uses only materials produced in its own factory, which are applied by the short-cycle pressing technique to particleboard. For many years, Perstorp was the only commercial manufacturer. Then, as rumours grew of allergic reactions to textile floor coverings, wood-based materials, whose health credentials were unquestioned, became more important.

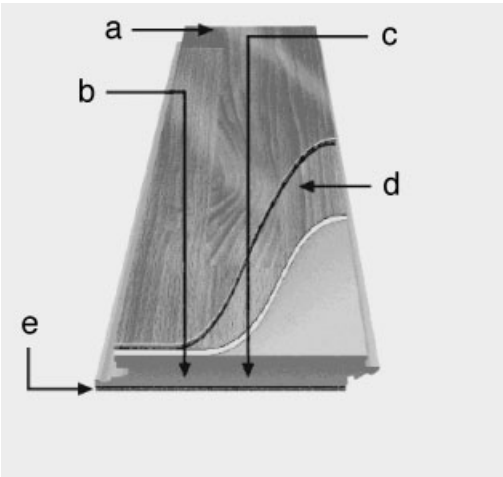

Figure 11. Construction of a laminate flooring tile (Witex) a) Very hard resin coating; b) Tongue-and-groove joint with extra impregnation to give water resistance; c) HDF substrate; d) Resinized patterned film, e.g. ,with the appearance of wood or stone, stable to UV light; e) High-strength film for extra hardness and impact resistance; f) Resinized protective film affords dimensional stability

After many years of rapid market growth and healthy profits, there is now stagnation and considerable continuing growth in production capacities. Meanwhile, laminated flooring elements are sold by large building suppliers at prices below 10 DM/m^3. Stable market development will be assured not only by favorable prices, but also by products that are suitable for do-it-yourself floor laying.

3.2. Construction

The construction of laminated flooring elements is very simple. To the bottom of the HDF substrate is applied the protective film, and to the top of the substrate are applied either the patterned film alone, the patterned film plus the overlay, or the underlay plus the patterned film plus the overlay (Fig. 11).

3.3. Materials

The substrate is the load-carrying element of the design. It must be as dense and hard as possible to resist loading pressures from furnishings. Recently, melamine-resin-bonded HDF has been used for high-quality products because of its considerably better resistance to moisture relative to urea-resin-bonded products. The underlying protective film increases the stability of the element, ensures constructional symmetry, and prevents the ingress of moisture from below. An underlay prevents the often dark background color of the substrate from showing through light-colored patterned films and reinforces the whole assembly against high-pressure loads.

The main function of the patterned film is decorative. It also has a load-bearing function and is responsible for the wear-resistance properties if it is not covered by an overlay. In the former case, it is resin-treated in two stages, first by conventional impregnation and then, after a short intermediate drying stage, with an additional melamine resin layer.

The overlay consists of carrier papers of pure α-cellulose containing micronized corundum powder, which is added during the production process. These special papers are patented and are produced in the United States. They enabled high-quality laminated flooring elements to be manufactured for the first time.

3.4. Production and Installation

Pressing. Hydraulic presses are the principal machines used in the manufacturing process, and these are also used in the production of self-supporting decorative laminated plates, for which all three types of commonly used presses are used: multiplate presses with cooling, one-plate presses without cooling, and double-band presses with or without cooling. Direct lamination is by far the most important process in the manufacture of flooring elements, mainly by short-cycle pressing. However, some double-band presses have been installed recently.

In *short-cycle pressing*, pressing times of 10 – 20 s (depending on the number of films to be applied) and pressures of 3 MPa are normally used. Up to two cycles per minute at pressing temperatures of 180 °C are currently achieved. The following operations are performed fully automatically:

- From height-adjustable holding racks, usually with three compartments, film is withdrawn by pulling devices. When the protective film is in position, the substrate is offered up to it by means of a movable arm with suction pick-up, and the film is applied from above in a second station. The film is applied as precisely as possible to minimize film overlap and reduce formation of waste by splintering. For flooring elements, especially with patterned strip flooring, the symmetry of the pattern across the width is very important. The printed areas at the joints must be in exact alignment.
- The short-cycle presses of various machine manufacturers differ only in their feed systems, whereby speed is of the greatest importance. Here, the critical dwell time, which begins with the application of heat in the compression space and ends when sufficient pressure is reached, plays an important role. It is ca. 2 s. Such short times can only be achieved with the aid of equipment that alternately feeds materials and discharges the tiles at high speed.
- The final operations are to clean the edges of the tiles, to check the surfaces visually, and in some cases to cut the tiles and stack them.

Changing the pressing plates plays a very important role in short-cycle pressing technology, as the stainless steel pressing plates also act as embossing matrices. A rapid changeover gives great flexibility in the short-cycle pressing process. The plate-changing operation is therefore also automated. The equipment for this purpose has access to a compartmented magazine that normally stores 15 different matrices. The changeover cycle requires only 10 min. It is operated by a movable arm, which uses suction to pick up the plate that has been taken out of the compression space by the discharge equipment and places it on the band conveyor which takes it to the plate magazine. After the old plate has been placed in the magazine, a new plate is taken out by reversing the process.

There are two fundamentally different *double-band pressing* systems in competition with each other. One operates with a surrounding sealed cushion of compressed air, and the other with free-flowing oil films on which the stainless steel bands slide. For coating operations with relatively simple process parameters, the much cheaper version with the air cushion can be considered, while for high-pressure coating materials the other system is recommended because of its ability to meet demanding requirements.

Double-band presses are now available in every working width demanded today. In contrast to the short-cycle pressing technique, in which the films are seldom fed from rolls, double-band presses are always supplied from rolls of film. This is one of the advantages of continuous pressing. Film storage and feeding mechanisms occupy very much less space than is required in short-cycle pressing technology. As flooring products are produced in large quantities, the rolls of film can be used continuously without interruption, and the management of left-over materials on rolls is minimized.

The process parameters are approximately the same as those in short-cycle pressing technology. The feeding rate, which determines the production capacity, is ca. 18 m/min. Assuming a heated film length of 2000 mm, this would give a pressing time of ca. 6.5 s. This extremely short pressing time can only be realized if the films are impregnated in-house with adherence to strict conditions. These include the quality and age of the resin for the impregnation process and also the storage time of the impregnated films. The film resins must have sufficient time under pressure to enable them to flow. Only then can the required fully densified surfaces be achieved.

Production capacity is an important criterion for choosing between short-cycle and double-band pressing. Thanks to the continuous improvements being made to digital servo-motors and other control systems, the speed of the short-cycle pressing technique is continually being increased, so that charging times play a continually diminishing role. Thus, it was only a question of time until the 18 m/min or even 20 m/min of the double-band presses could be matched by long short-cycle presses, albeit at the cost of increased space requirements. A major press manufacturer in Westphalia has delivered a short-cycle press of 13 m operating length to Italy. It is expected to operate at two cycles per minute. Double-band technology will respond with an extended compression area. However, as the hardening reaction takes place essentially over the first meter, this imposes strict demands.

Embossed pressing bands are much more expensive than the equivalent pressing plates. Also, changing the pressing band is much more time-consuming and difficult. However, flooring elements without embossing are of lower quality, both visually and for practical purposes. The wide range of embossing patterns that can be produced by pressing plates is not available with bands, simply because of the exorbitant costs of the latter. However, the continuous process achievable by continuous presses gives considerable potential savings in the consumption of electrical and thermal energy.

Conditioning and Cutting. After pressing, the laminated plates are allowed to stand so that internal stresses caused by unevenly distributed moisture and temperatures can even out. Short-cycle products are generally stacked for about 5 d. Attempts are being made to shorten this time by using star coolers. Double-band presses with cooling give products that are much easier to cut and which only require about half the time for stress relief.

After conditioning of the full-sized products, they are precisely cut to the required dimensions. Patterned products have a clearly visible symmetry line that was applied applied during pressing. Prior to the multibladed saws, alignment devices adjust the position of the sheet until two optical sensors indicate that the symmetry axis has been reached.

Two types of multibladed saws are used. One operates with a single motor which drives a central, replaceable drive shaft to which the saws are attached, usually by quick-release bushes; the other has a drive for each saw blade, and these can be rapidly adjusted to different widths by digital controllers. Inoperative motors are stored in parking stations. Depending on the format of the sheets, lengthwise cutting can precede several transverse cuts. The cut products are conditioned for up to two more days to relieve the stresses that are an inevitable consequence of the cutting process.

Formatting and Profiling. Variants of the above process all aim to reduce the conditioning time. These include direct linking of the cutting saw with the profiling production line. In all cases, this involves risks. Whether these outweigh the gain in rationalization is a question that must be considered. After cutting, the workpieces always lie with the patterned side upwards. On placing the tiles in the profiling line, the patterned sides are turned to face downwards so that the underside of the element is subject to the tolerances of the profiling step.

Profiling is performed by two angled double-ended profilers. The first, for the lengthways machining, can be replaced by a grooving machine. Modern production lines have high outputs. The newest plant can produce 120 elements per minute, sort them, pack them into cartons, stack the cartons on palettes, and load the palettes into containers.

The design of the milling stations is decisive for the fit of the profiles. An absolutely level surface of the fitted flooring, with gap-free joints both lengthwise and crosswise, is the criterion by which the customer will eventually judge the quality. A precisely level surface can only be achieved if the workpieces are milled to give a close fit one with another over the entire batch. With high feed rates, there can be undesired frictional heating, which can be aleviated by a carefully controlled air cushion. The profiles must not only mate cleanly, but above all must show high repeatability. From one consignment to the next, year after year, there must be no flaws of any kind.

The sharpness and life of the tools play an important role. Overlays filled with corundum powder lead to tool wear because of their very

abrasive properties. Hence, only diamond tools are used, and these are replaced at high speed with the aid of a quick-action chuck. Polycrystalline and monocrystalline diamond fragments (PCD and MCD, respectively) are under discussion for this application. It would appear that the extremely expensive MCD tools are not cost-effective, although their useful life is many times that of PCD tools.

Water-Resistant Finishing. As it is not yet standard to use melamine-bonded HDF as substrate, the profiles are impregnated with water-resistant material after both lengthwise and crosswise machining. The water-repellent material is applied evenly over the whole length of the joint without allowing it to spread over the surface; it is then allowed to soak in and dry. The water-resistant material must not affect the adherent properties of the adhesive used for the installation of the floor.

Packaging. A quantity of floor tiles equivalent to approximately 2 m^2 of flooring has become established as a suitable package size. Packaging is carried out in the profiling production line, and is therefore subject to demands for high production rates. Few manufacturers use completely automatic packing equipment. The cheapest method is to secure the packages, which usually contain eight flooring elements, with sealing tape. These, together with the product information, which is also automatically packed, are then shrink-wrapped to give protection from dirt and moisture. Packing in cartons gives greater protection against damage. The elements can either be placed, from above, in an automatically prefolded carton or be packed in a slip-lid carton. Here, too, additional foil wrapping is advisable. The product information and preferably also the patterned surface should always be visible from outside, for example, through a window.

Installation. Laminate flooring is laid by floating on sheeting that deadens impact sound and protects from moisture (Fig. 12). In accordance with the guidelines for laying PVC flooring, the base, usually concrete, must be well dried and even, as these thin flooring materials are not capable of bridging over large discontinuities.

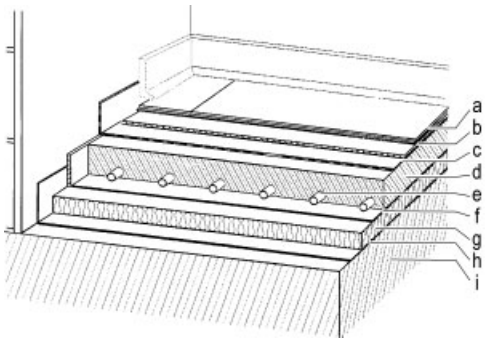

Figure 12. Laying of laminate flooring
a) Laminate flooring; b) Wool-felt board; c) Vapor barrier; d) Flooring plaster; e) Heating tube; f) Screening; g) Insulation; h) Moisture barrier; i) Concrete floor

The floated surface forms a large unit with material-corresponding shrinkage or growing, and water imbibition values. It must therefore always be kept at a distance of at least 10 mm from walls and pillars, and these gaps can be covered with matching edge strips. Both adhesive and mechanical fixing of flooring elements are used. The former still dominates the market, but will probably be replaced by dry methods. The poly(vinyl acetate) adhesive must exude upwards through the joints along their entire length to prevent water penetration. With mechanically fixed elements, this additional protection is not provided, and a water-resistant finish or water-resistant substrate is necessary. The fixing adhesives must be moisture-proof, tough, elastic, and of high strength.

3.5. Testing

Laminate flooring was defined in early 2000 in the European Standard EN 13 329 and its specifications and properties were described. Laminate flooring is described as having closed pores, and this property can be measured by testing its sensitivity to staining by certain test materials (e.g., graphite). According to EN 438, laminate flooring is Grade 3, but it should preferably be Grade 4.

Before the appearance of the laminate flooring standard EN 13 329, the abrasion resistance was determined and used as a single value in accordance with EN 438. Because of the unsuitability of the testing of highly wear-resistant

Table 2. Abrasion resistance classes for laminate flooring

	Level of use					
	Domestic			Commercial		
	Moderate	General	Heavy	Moderate	General	Heavy
Class	21	22	23	31	32	33
Abrasion resistance	AC1	AC2	AC3	AC3	AC4	AC5

laminate flooring by methods developed for furniture, values of 4000 – 20 000 revolutions were quoted as test results. In the new standard the wear resistance test has been improved, and the test for highly wear resistant overlays has been modified.

The new test results enable the products to be grouped according to their abrasion resistance, so that each type of overlay can be assigned to an abrasion-resistance class (Table 2).

The well-known hardness tests such as Brinell and ball indentation values are of only limited suitability for determining the hardness of a laminate floor. These tests therefore do not feature in the European laminate standard. The French laminate standard NFB 54 020 gives a minimum value of 4 or 5 MPa.

Laminate floors are produced with a wide range of gloss values. Two or three years ago these were in the 10 – 15 range, but now more and more producers offer gloss values of 18 – 20. These high values enable reproduction to be even more realistic.

As a result of overlay technology and the existence of melamine resins, laminate floorings are very resistant to heat and are therefore resistant to a glowing cigarette (better than Grade 4 of EN 438). As a rule, laminate floorings meet the requirements of B1 of DIN 4102 Part 14.

Taking all their properties into consideration, laminate floorings stand up very well in comparison to other floor coverings. Past criticisms with respect to moisture sensitivity and footstep noise are to be set against the developments in modern laminate floorings. Special backing coatings and nonswelling substrates are reducing footstep noise and improving edge instability.

Whereas attempts are still being made to develop a scientific test for footstep noise, testing methods for edge-swelling behavior and the measurement of dimensional stability by modified testing procedures are described in EN 13329, maximum edge-swelling values being 18 % for commercial and 20 % for domestic use.

Other test specifications of relevance for laminated floorings are listed below:

EN 309	Wood Particleboards – Definition and Classification
EN 311	Particleboards – Surface Soundness of Particleboards – Test Method
EN 316	Wood Fibreboards – Definition, Classification and Symbols
EN 318	Fibreboards – Determination of Dimensional Changes Associated with Changes in Relative Humidity
EN 438	Decorative High Pressure Laminates (HPL) – Sheets Based on Thermosetting Resins
EN 685	Resilient Floor Coverings – Classification
EN 12529: 1998	Castors and Wheels – Castors for Furniture – Castors for Swivel Chairs – Requirements

3.6. Environmental Aspects

Laminate floorings are built up of layers of various materials. This often leads to the assumption that "plastic floor coverings" are difficult to dispose of in an environmentally friendly manner. However, laminate floorings can be disposed of like furniture and with as few problems. In both cases, because of the high wood content, the most appropriate disposal route is controlled combustion, the heating values being good.

VOC emissions from the product are very low, so that a laminate floor is classified as a very low emission material. Formaldehyde emissions are in the range 0.02 – 0.04 ppm, i.e., on the order of those of natural wood.

Milling operations produce wood dust and shavings. The usual national recommendations apply to these materials. When the material is cut during the course of installation on a building site, dust emissions can vary. Saws should preferably be provided with dust-extraction equipment.

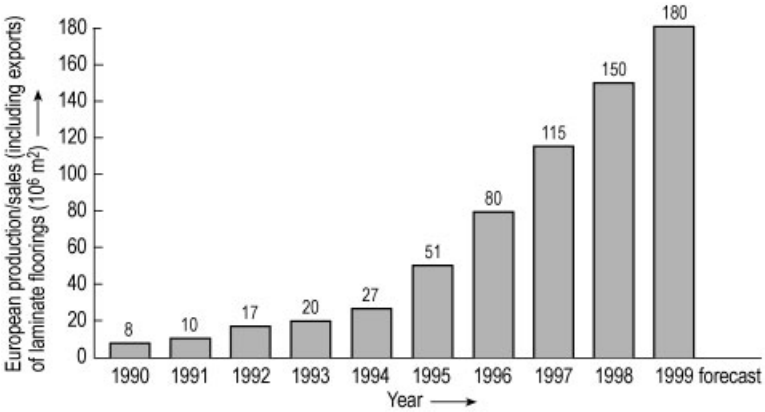

Figure 13. European sales (including exports) of laminate flooring (in 10^6 m^2)

3.7. Economic Aspects

The European market for laminate floorings continues to grow rapidly and is forecast to be 180×10^6 m^2 in the year 2000 (Fig. 13, while the German market is expected to level out at ca. 40×10^6 m^2. In western Europe, laminate flooring had a ca. 6.8 % share of the total flooring market in 1998, corresponding to ca. 100×10^6 m^2.

4. References

1. Carpet Market Overview, DuPont, 1999.
2. R. C Buck: New Nylon Stain Resist Technology. 36th International Man-Made Fibers Congress, Dornbirn, Austria, 1997.
3. M. P. W. Wilson: The Chemical and Physical Properties of Polyamides: a Review, 37th International Man-Made Fibers Congress, Dornbirn, Austria, 1998.
4. M. P. W. Wilson (1999) Designing Carpet Fibers for Esthetics and Performance. 38th International Man-Made Fibers Congress, Dornbirn, Austria.
5. P. L. Hauck: Carpets and Hygiene, Recent Key Learning's, TIFCON 1994.
6. P. L. Hauck: Sustainable Growth and Environmental Excellence are Interdependent, TIFCON 1999.
7. Frederick Walton: *The Infancy and Development of Linoleum Floorcloth,* London 1925.
8. G. Forster et al., *Linoleum* , Verlag moderne Industrie, Landsberg 1996.
9. R. F. Lonzillotti: *The Hard Surface Floor-Covering Industry,* State College of Washington Press, Pullman, Washington, 1955.
10. B. Berkley: *Floors, Selection and Maintenance,* LTP Publishers, Chicago 1968.
11. *Floor Covering Weekly* **47** (1998) no. 28.
12. A. S. Wilson, *Plasticizers, Principles and Practice,* Institute of Materials, London 1995, p. 36.
13. A. W. McKee: "Study of the Effects of Fillers in Compounding Plastic Flooring," *Soc. Plast. Eng. J.* **18** (1962, Feb.) no. 2, 186 – 190.
14. Congoleum-Nairn Inc., US 3 293 094; 3 293 108, 1966 (R. Frank Nairn et al.).
15. Standard Specification for Sheet Vinyl Floor Covering with Backing, ASTM F 1303-97.
16. Standard Specification for Sheet Vinyl Floor Covering without Backing, ASTM F 1913-98.
17. Standard Specification for Vinyl Composition Floor Tile, ASTM F 1066-95a.
18. Standard Specification for Solid Vinyl Floor Tile, ASTM F 1700-96.
19. *Standard for Health Care Facilities,* National Fire Protection Assn., NFPA 99, Quincy, Mass. 1984, p. 24.

Fluorescent Dyes

HANSRUDOLF SCHWANDER, Ciba-Geigy AG (retired), Basel, Switzerland (Chaps. 1–11)

PAUL HENDRIX, Radiant Color NV, Houthalen, Belgium (Chap. 12)

1. Introduction 1
2. Naphthalimide Dyes 2
3. Coumarin Dyes 2
4. Xanthene Dyes 3
5. Thioxanthene Dyes 5
6. Naphtholactam Dyes 6
7. Azlactone Dyes 6
8. Methine Dyes 7
9. Oxazine and Thiazine Dyes 7
10. Miscellaneous Fluorescent Dyes . . 7
11. Special Uses 9
12. Daylight Fluorescent Pigments . . 11
12.1. Production 11
12.1.1. Dyes for Daylight FluorescentPigments 11
12.1.2. Pigment Matrices 11
12.2. Quality Specifications 12
12.3. Uses 12
12.4. Toxicology 13
13. References 13

1. Introduction

Fluorescent dyes differ from normal dyes in that they produce exceptionally bright colors because they not only absorb light, but also emit it. Fluorescent dyes are used mainly as disperse or reactive dyes (→ Disperse Dyes, → Reactive Dyes) in the coloring of textile fibers and plastics, in the production of inks and printing inks and, in combination with appropriate pigments, in the production of paints and lacquers. Fluorescent dyes are also employed as laser dyes (→ Laser Dyes). Furthermore, daylight fluorescent dyes may be dissolved in a resin matrix and then pulverized to form fluorescent pigments (see Chap. 12).

Fluorescence occurs when molecules that have absorbed light and are in their lowest excited state S_1 return to their ground state S_0 and emit light (for a detailed discussion of fluorescence, see → Dyes, General Survey, Chap. 2.4.). Some of the absorbed light energy also activates nuclear vibrations and is released as heat. Further energy is lost due to the conversion of the lowest excited state S_1 to a triplet state T_1 (*intersystem crossing*) and is not available for fluorescence. The *quantum yield* is a measure of the fluorescent efficiency of a compound, i.e., the proportion of the absorbed light energy that is emitted as fluorescent light.

Fluorescent dyes absorb and emit light in the visible region of the spectrum. The fluorescent dyes treated in this article are distinguished from optical or fluorescent brighteners (→ Optical Brighteners), which absorb in the invisible UV region and emit blue to blue-violet light.

Fluorescent dyes usually have extremely rigid, extended π-systems. Rigidity is of importance because it suppresses the release of energy due to activated nuclear vibrations. Substituents such as heavy atoms (chlorine and bromine) or nitro groups are detrimental to fluorescence because they favor intersystem crossing.

Many dyes exhibit fluorescence, but to be of practical use, fluorescent dyes must satisfy certain requirements: they must produce a pure color dictated by their absorption and emission spectra, they must have a high molar extinction, and most important, they must have a high quantum yield. These requirements are met by very few dyes. A disadvantage shared by many fluorescent dyes is their poor lightfastness, but there are some exceptions.

In this article some of the most important fluorescent dyes and their applications are described. For a general discussion of the properties of organic dyes and pigments, see → Dyes, General Survey; → Pigments, Organic.

2. Naphthalimide Dyes

A very wide variety of naphthalimide dyes are known (→ Naphthalimide Dyes and Pigments).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_279

Only a few important representative examples, which show exceptionally brilliant, greenish-yellow fluorescent colors, are discussed in this section; they all show poor lightfastness. Brilliant Sulfoflavin FF (**1**),C.I. 56205 [*2391-30-2*] serves as a dye for wool and polyamides [6]. Solvent Yellow 44 (**2**),C.I. 56290 [*2478-20-8*] is used to dye plastics and is of some importance in the production of fluorescent pigments [7].

Compound **1** is prepared from *p*-toluidine and 4-amino-3-sulfo-1,8-naphthalic anhydride [*52173-68-9*] which is obtained by sulfonating 4-amino-1,8-naphthalic anhydride. The production of (**2**) involves the reaction of 4-chloro-1,8-naphthalic anhydride with *m*-xylidine, followed by the use of ammonia to substitute the chlorine atom with an amino group.

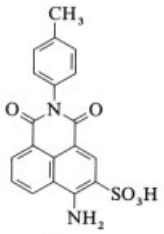

1

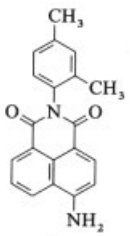

2

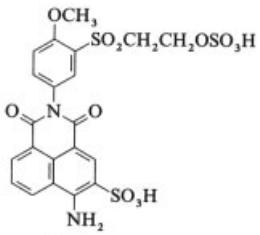

3

4-Amino-3-sulfo-1,8-naphthalimide is being increasingly used to dye wool and *N*-methyl(4-methylamino)-3-sulfo-1,8-naphthalimide [*54229-25-3*] to dye polyamides. Compound **3** [*23741-82-4*] is a reactive dye which is employed to dye cellulose fibers [8].

3. Coumarin Dyes

Dyes derived from coumarin, 2*H*-1-benzopyran-2-one [*91-64-5*], are usually intensely colored, brilliant greenish-yellow dyes with high quantum yields and high molar extinctions. They are moderately lightfast.

The most important fluorescent coumarin dyes are derivatives of 7-dialkylaminocoumarin, which has a heterocyclic residue at the 3-position (**4**).

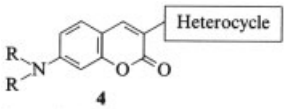

4

R = alkyl

Fluorescent coumarin dyes are used as disperse dyes for polyesters and polyamides, their sulfonic acid derivatives are employed for dyeing polyamides, and their cationic derivatives are used for dyeing polyacrylonitriles. Some of these dyes are important as laser dyes or are used in the production of fluorescent pigments (see Chap. 12).

Coumarin dyes are obtained by reacting 2-hydroxy-4-dialkylaminobenzaldehyde with the nitrile (or acetate) derivative of a heterocycle as in the following example:

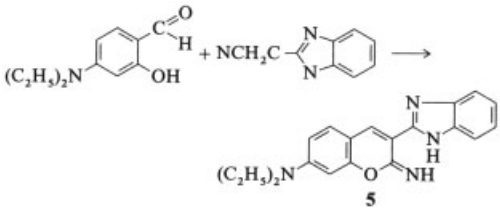

5

The use of a nitrile derivative results in the introduction of an imino group at the 2-position of the coumarin ring, as in compound **5**. Hydrolysis occurs under the conditions employed for dyeing, and the imino group is replaced by oxygen, yielding the corresponding coumarin derivative.

Some important 7-diethylcoumarin dyes (compounds **6**–**13**) are listed in Table 1. The introduction of a nitrile group at the 4-position of the coumarin residue gives the dye a red color, as in compound **14** [*70546-25-7*], which is the 4-nitrile derivative of **9** [15].

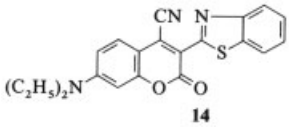

14

Table 1. 7-Diethylcoumarin dyes [9], [10], [11], [12], [13], [14]

Constitution of dye 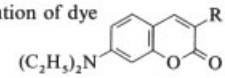 R =	CAS registry number	Color	Use	Reference
6	[*27425-55-4*]	yellow	disperse dye	[a]
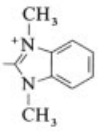 **7**	[*12221-86-2*] (C.I. Basic Yellow 40)	yellow	for dyeing polyacrylonitrile, fluorescent pigment	[a]
8	[*38215-36-0*]	yellow	disperse dye, laser dye	[a]
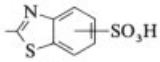 **9**		yellow	for dyeing polyamide	[b]
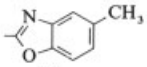 **10**	[*34564-13-1*]	yellow	disperse dye	[c]
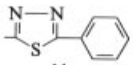 **11**	[*62143-26-4*]	yellow	disperse dye	[d]
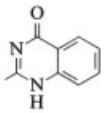 **12**	[*28754-28-1*]	yellow	disperse dye	[e]
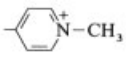 **13**	[*55470-53-6*]	orange	for dyeing polyacrylonitrile	[f]

Compound **14** is used as a disperse dye for polyesters and produces a brilliant red color of moderate lightfastness.

4. Xanthene Dyes

Fluorescein (**15**), C.I. 45350 Acid Yellow 73 [*518-47-8*] is the best known xanthene dye (→ Triarylmethane and Diarylmethane Dyes).

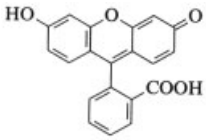

15

Even a very dilute solution of this dye exhibits an intense yellowish-green fluorescence. It is synthesized by reacting phthalic anhydride with resorcinol. The eosines (bromine compounds) and erythrosines (iodine compounds), phloxine

B, C.I. 45410 [*18472-87-2*], and Rose Bengal, C.I. 45440 [*632-68-8*] are halogenated derivatives of fluorescein.

Rhodamine dyes (aminoxanthenes) are a very important class of xanthene dyes and are of considerable commercial importance. They are intensely brilliant, blue-red, fluorescent dyes with very high molar extinctions and quantum yields and are used for coloring inks, paper, and plastics. Their sulfonated derivatives are used to dye wool and polyamides. Rhodamine dyes are also the starting materials for the production of red fluorescent pigments; some of them are employed as laser dyes. The most important rhodamine dyes are synthesized by reacting 3-dialkylaminophenols with phthalic anhydride. The carboxyl group of the product may then be esterified.

The first of the rhodamine compounds to be reported was Rhodamine B (**16**), C.I. 45170 Basic Violet 10 [*81-88-9*] [16], which is used in fluorescent pigments.

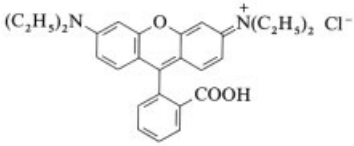

16

Rhodamine 3 B (the ethyl ester of Rhodamine B), C.I. 45175 Basic Violet II [*2390-63-8*], and Rhodamine 6 G (**17**), C.I. 45160 Basic Red 1 [*989-38-8*], have properties similar to those of **16** [17].

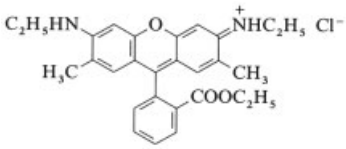

17

The product of the reaction of 3-diethylaminophenol with benzaldehyde-2,4-disulfonic acid is Sulforhodamine B (**18**), C.I. 45100 Acid Red 52 [*3520-42-1*] [18], which is used to dye wool and polyamides. It is also employed in the production of fluorescent pigments, and as a laser dye.

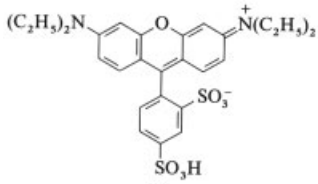

18

The alkyl groups on the nitrogen atoms of Sulforhodamine 101 (**19**) [*60311-02-6*] are less flexible than those of Sulforhodamine B (**18**) [19], and the molecule is thus more rigid. Dyes of this type have very high quantum yields, approaching 100 %, and are used as laser dyes.

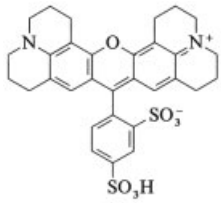

19

Compound **20** is used as a reactive dye for dyeing cellulose fibers [20].

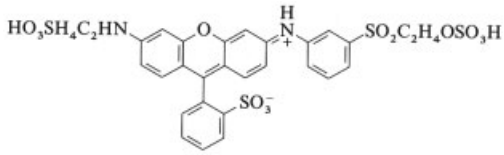

20

The reaction of a coumarin dye with dicyanomethane produces compound **21**, which can be considered as an azaxanthene dye [21].

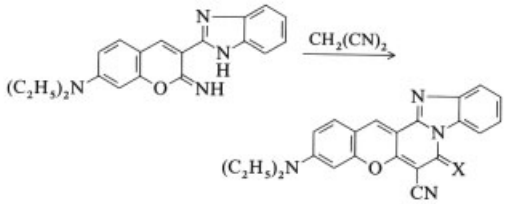

21

21a [*52372-39-1*]: X = NH
Hydrolysis of **21a** gives **21b** [*52372-36-8*]: X = O

This compound is used as a disperse dye for polyesters. It produces an intensely brilliant, moderately lightfast, red color.

5. Thioxanthene Dyes

The color of fluorescent thioxanthene dyes ranges from yellow to red. These dyes are usually less brilliant than the coumarin dyes and the rhodamines, but some of them show good lightfastness. The thioxanthene ring system is produced by means of a Pschorr ring closure, as shown for Fluorescent Red GG (**22**), C.I. Solvent Orange 63 [*16294-75-0*] [22]:

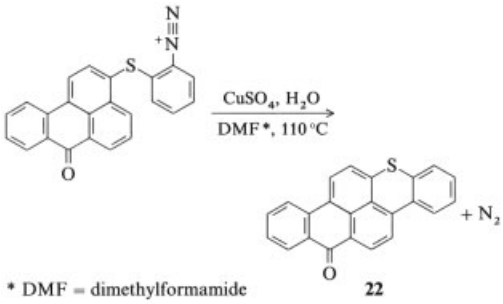

Plastics can be colored red with compound **22** which has a good lightfastness.

Samaron Brilliant Yellow H 6 GL (**23**), C.I. 56235 Disperse Yellow 105 [*14121-47-2*], is an intensely colored, lightfast, sublimate-free disperse dye. It is used to impart a yellow color to polyesters [23] and is prepared by reacting benzo[*k,l*]-thioxanthene-3,4-dicarboxylic anhydride with aminomethoxypropane. If the latter is replaced by stearylamine, a compound is formed that is particularly useful for dyeing plastics in bulk, mineral oils, fats, and waxes [24].

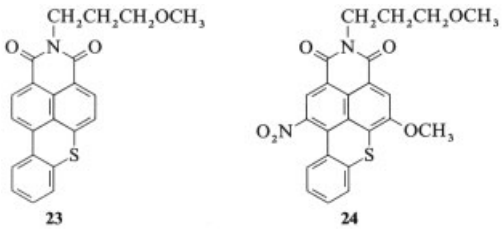

Compound **24** [*18014-08-9*] can be employed as a disperse dye, with good lightfastness, to impart a brilliant pink color to polyesters [25]. Compounds **23** and **24** are also classified as naphthalimide dyes.

Disperse dye **25**[*35763-62-3*] and its isomer **25a** give a very brilliant yellowish-red color having moderate lightfastness when applied to polyesters [26]. The naturally occurring dye Iachnanthofluorone (**26**) [*53766-48-6*] is a red fluorescent xanthene derivative that is closely related to compounds **23**–**25** [27].

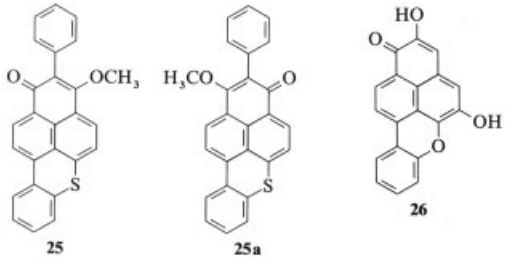

6. Naphtholactam Dyes

Naphtholactam dyes have colors ranging from yellow to red and are suitable for dyeing polyesters and, in some cases, polyamides [28]. Although they show excellent lightfastness, these dyes do not have the high molar extinctions nor the quantum yields of, for example, the coumarins. Naphthostyril [*130-00-7*] (the lactam of 1-aminonaphthalene-8-carboxylic acid) reacts with a weakly basic aromatic or heterocyclic amine or with a reactive methylene group (e.g., a barbituric acid derivative) in the presence of phosphoryl chloride in an inert solvent to yield naphtholactam dyes. Compound **27** [*67880-03-9*] is obtained if the reaction is performed with 2-cyano-4-nitroaniline [29], as shown below; this compound produces a reddish-yellow color when used as a disperse dye with polyesters.

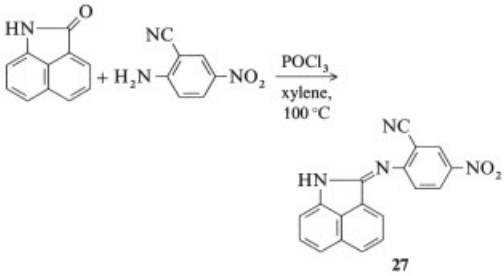

The disperse dye **28** [*58470-74-9*] gives a yellowish-red color with good lightfastness when applied to polyesters [30]; it is especially suitable for textile printing.

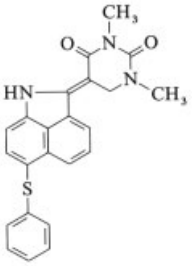

28

Compound **29** can be used to dye polyamides orange with a high degree of lightfastness [31].

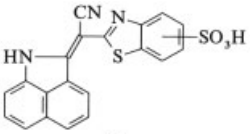

29

Polyesters can be dyed orange to red with a high degree of lightfastness by using dyes typified by compound **30** [32].

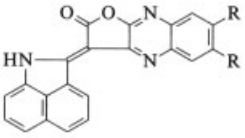

R = H, alkyl, or alkoxy

30

7. Azlactone Dyes

Azlactones are formed by condensation of benzoylglycine (hippuric acid) with an aromatic aldehyde in the presence of acetic anhydride and sodium acetate.

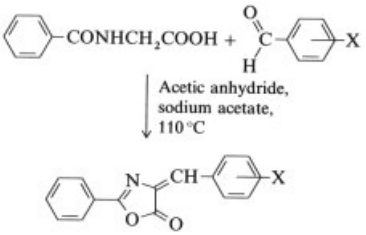

Although azlactone dyes of the donor-acceptor type (**31**) are very brilliant, they are not lightfast. However, the greenish-yellow disperse dyes (**32**) [*25744-09-6*] and (**33**) [*51202-86-9*] can be effectively used on polyesters. They possess excellent coloring power and lightfastness [33], [34].

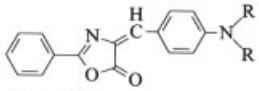

R = alkyl

31

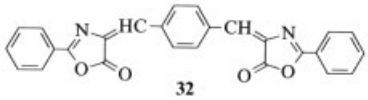

32

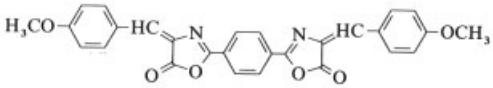

33

8. Methine Dyes

Methine dyes (→ Methine Dyes and Pigments) are cationic dyes that are used to color polyacrylonitrile and acid-modified polyester fibers. Dye **34** is a fluorescent greenish-yellow [35] and is synthesized by condensation of 1,3,3-trimethylindoline-2-ylidene acetaldehyde [*84-33-3*] (Fischer base aldehyde) with 3-phenyl-2-pyrazoline [*936-47-0*].

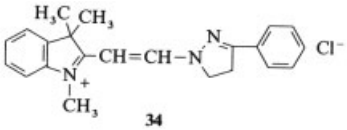

34

Dye **35** is formed by coupling diazotized *p*-toluidine to 1,3,3-trimethyl-2-(3-chloro-1-propylidene)indoline [36]. This product imparts a fluorescent, lightfast, orange color to polyacrylonitrile.

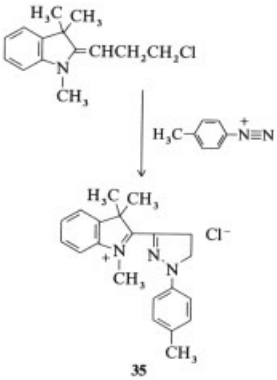

35

Another methine dye (**36**) [*23406-34-0*] colors polyacrylonitrile a brilliant red, but with only moderate lightfastness [37].

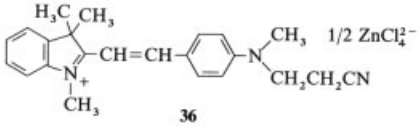

Dye **36** is formed by the condensation of 1,3,3-trimethyl-2-methyleneindoline with 4-(*N*-methyl-*N*-cyanoethylamino)benzaldehyde.

9. Oxazine and Thiazine Dyes

Oxazine dyes (→ Azine Dyes), such as Basic Blue 3 (**37**), C.I. 51004 [*63589-47-9*], are used predominantly to color polyacrylonitrile a greenish-blue [38]. Some of these dyes, including **37**, are employed as laser dyes [39].

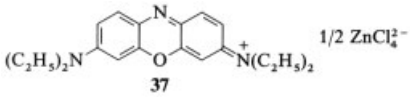

Thiazine dyes, e.g., dye **38**, exhibit fluorescence at long wavelengths from 650 to 680 nm and are also used as laser dyes.

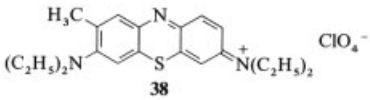

Some dioxazines are also fluorescent dyes; dye **39** imparts a fluorescent, yellowish-red color to polyamides [40].

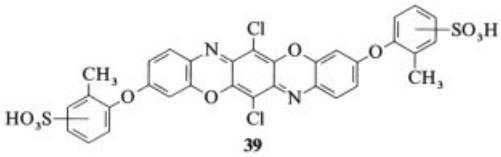

10. Miscellaneous Fluorescent Dyes

Diphenyl Brilliant Flavine 7 GFF (**40**), C.I. Direct Yellow 96 [*61725-08-4*] [41], can be regarded as an optical brightener (→ Optical Brighteners) which absorbs light in the visible region of the spectrum. Its maximum absorption is shifted to longer wavelengths in comparison with stilbene systems due to its more extensive conjugation. It is used as a direct dye for coloring cellulose fibers greenish-yellow (see bottom).

Disperse Yellow 77 (**41**), C.I. 70150 [*22281-18-1*] [42], can be prepared by acylation of 1-aminoanthraquinone with acetic anhydride, followed by condensation with anthranilic acid and zinc chloride at 100 – 110 °C. Both steps are carried out in the same reaction vessel.

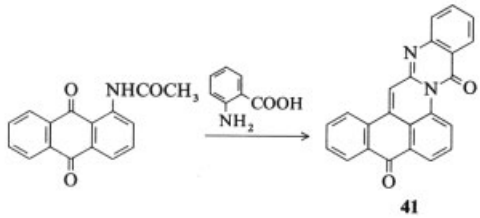

Compound **41** is a brilliant, fluorescent, disperse dye that can be used to dye polyesters; it produces a greenish-yellow, sublimation–fast color with especially good lightfastness.

Fluorol 242 (**42**), C.I. 68410 [*6871-92-7*], a derivative of anthraquinone [43], is synthesized by replacing the bromine atom of 4-bromo-6-aminoanthrapyrimidine with an octadecylamino residue, followed by acylation with butyric chloride. Compound **42** is a fluorescent, yellowish-green dye and is suitable for dyeing fats and oils [24].

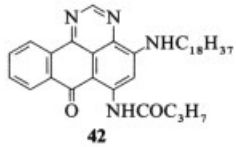

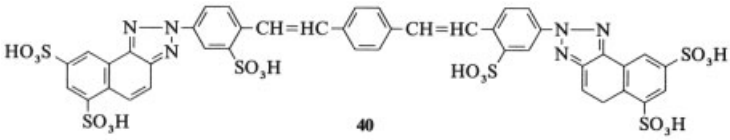

Solvent Green 5 (**43**), C.I. 59075 [*2744-50-5*], is a heat-resistant, lightfast dye [44]. It is used to color thermoplastic materials yellow in bulk and is suitable for material testing (see Chap. 11).

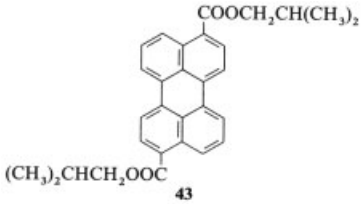

Perylenetetracarboxylic acid diarylimide derivatives typified by **44** [45] are used in the bulk dyeing of plastics (→ Naphthalimide Dyes). They produce extremely lightfast, red colors and are suitable for the production of fluorescent collectors (see Chap. 11).

The vat dye Thioindigo Red B (**45**), C.I. 73300 [*522-75-8*], is not fluorescent. However, this substance and its derivatives can be used conveniently in the bulk dyeing of plastics. They dissolve in the plastic material producing lightfast, heat-resistant fluorescent colors [48]. Asymmetrical thioindigo derivatives are suggested as especially suitable for the bulk dyeing of polyesters [47].

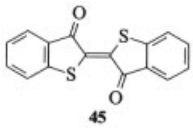

The naphthyridine dione dye Ciba Lake Red B (**46**), C.I. 73095 [*6417-51-2*] [48], has also been recommended for the bulk dyeing of plastics; it produces red fluorescent shades. This dye is synthesized by the acylation of indigo with phenylacetyl chloride and ring closure.

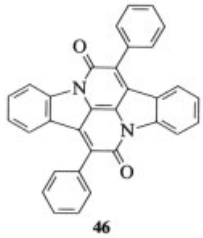

Azo coupling of 5-nitro-2-aminoindazole to 2,4,6-tris(*N*-dimethylamino)pyrimidine yields Azo Orange (**47**). Compound **47** is not itself fluorescent, but is converted to the yellow fluorescent dye **48** [*38432-19-8*] by treatment with acid [49]. Compound **48** can be used as a disperse dye to impart a lightfast, yellow color to polyesters. The formation of **48** from **47** illustrates that the stiffening of a molecular structure can convert a nonfluorescent azo dye into a fluorescent system.

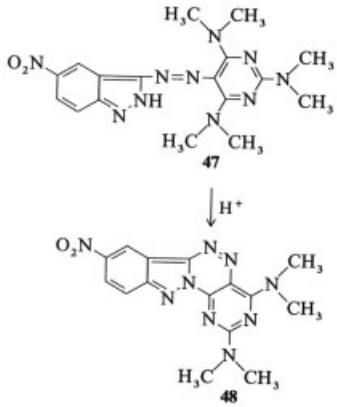

11. Special Uses

The general uses of fluorescent dyes as colorants for fibers and plastics have been described in the preceding sections. This chapter describes special applications of fluorescent dyes, but does not include their use as laser dyes [50] because this is a separate keyword (→ Laser Dyes). Fluorescent daylight pigments are treated in Chapter 12.

Photochemistry. A variety of fluorescent dyes, in particular xanthene dyes (e.g., Rose Bengal), are used as sensitizers in photochemistry [51].

Material Testing. Fluorescent dyes, e.g., Solvent Green 5 (**43**), are used in the testing of materials to detect microflaws in the surfaces of metals (→ Nondestructive Testing).

Fluorescent Collectors. Fluorescent collectors consist of highly transparent plastic sheets, e.g., poly(methyl methacrylate), which are colored with a fluorescent dye. Exposure of the sheets to diffuse sunlight produces fluorescent light, which is then concentrated and collected by total internal reflection and released from the thin surfaces at the edges of the sheets. The collected light energy can be used to produce electricity by means of solar cells (→ Solar Technology) [52]. Dyes typified by **44** have proved to be most useful for this purpose.

Indicators. Fluorescent dyes are used as indicators in analytical titrations (→ Indicator Reagents).

In precipitation titrations, the fluorescent dye is adsorbed to the surface of the precipitate and undergoes a clearly recognizable change in fluorescence at the end point. Silver ions can be titrated in this way, using fluorescein (**15**) or Rhodamine 6 G (**17**) as an indicator.

Another class of fluorescent indicators consists of water-soluble compounds whose fluorescence changes at the end point. These indicators permit the titration of turbid or colored solutions. Thiols, for example, can be titrated with organic mercury compounds using fluorescein as an indicator.

Derivatizing Reagents. Substances that are normally very difficult to detect (amines, amino acids, peptides, sugars, carboxylic acids, and lipids) can be converted into their fluorescent derivatives with a suitable reagent. The derivative is then separated by means of TLC or HPLC and its fluorescence is measured. Detection sensitivity is very high; aniline, for example, can be detected at a concentration of 2 ng/mL [53]. Examples of such reagents are given in Table 2 (compounds **49**–**53**).

Compounds **52** and **53** are particularly interesting fluorogenic compounds because they produce intensely fluorescent derivatives, although they themselves do not fluoresce.

Fluorescent Probes. Fluorescent probes permit measurement of various parameters in cellular systems. For example, the pH can be measured by treating thin tissue sections with a fluorescent pH indicator, such as 4-methyl-7-hydroxycoumarin (**54**) [*90-33-5*] ($pK_s = 7.6$) [54].

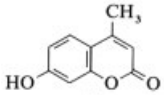

54

Other fluorescent probes are used to determine the partial pressure of oxygen, the membrane potential, the redox status, the degree of lipophilia, viscosity, etc. [55]. Interference caused by the addition of foreign fluorescent probes can often be avoided by exploiting the intrinsic fluorescence of many natural products, such as thiamine [56]. Test strips coated with fluorescent probes have attained considerable importance in medical diagnostics [57]. For example, theophylline in serum may be determined by this method. Fluorescence-based immunoassays have also been developed (→ Diagnostic Reagents, Chap. 3.3.3.)

Enzyme Substrates. Enzyme activity can be measured by allowing enzymes to act on a substrate to produce an intensely fluorescent product. Suitable substrates are provided by the carboxylic acid esters, phosphates, and sulfates of highly fluorescent phenols with pK values between 6 and 9 such as 4-methyl-7-hydroxycoumarin (**54**), fluorescein (**15**), and 2-hydroxy-3-naphthoic acid (**55**) [*92-70-6*].

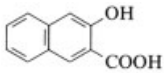

55

The glycosidic ethers of these phenols formed with a variety of sugars (e.g., glucose and galactose) are important in determining the activity of enzymes that cleave sugars.

Esters formed from compound **54** and long-chain fatty acids are commercially available and are suitable for measuring the activity of carboxyl esterases. The palmitic acid ester of compound **55** exhibits very weak fluorescence at 400 nm and is cleaved to palmitate and **55** by the

Table 2. Reagents used to synthesize fluorescent derivatives for chemical analyses

Name	CAS registry number	Structural formula	Derivatized compounds
Naphthyl isocyanate (**49**)	[*86-84-0*]	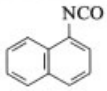	alcohols and amines
1-Dimethyl-5-naphthalene-sulfonyl chloride (dansyl chloride) (**50**)	[*605-65-2*]	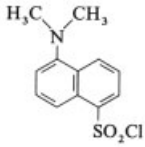	amines and phenols
Fluorescein isothiocyanate (FITC) (**51**)	[*27022-45-3*]	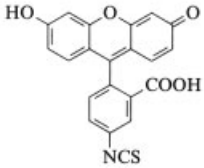	amines, amino acids, and proteins
1,2-Benzenedicarboxaldehyde (*o*-phthalic anhydride, OPTA) (**52**)	[*643-79-8*]	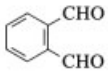	primary amines
N-4-(2-benzimidazolyl) phenylmaleimide (BIPM) (**53**)	[*27030-97-3*]	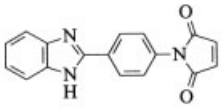	thiols

enzyme lipase. Since compound **55** fluoresces very intensely at 460 nm, the rate at which the intensity of fluorescence increases can be used to measure enzyme activity (direct continuous kinetic assay) [58].

12. Daylight Fluorescent Pigments

Pigments are particulate solids which, unlike dyes, are practically insoluble in the medium to which they are added. Most daylight fluorescent pigments consist of finely divided resin particles that contain daylight fluorescent dyes. Their luminosity and brilliance make them particularly useful if intense or long-distance visibility is needed.

Daylight fluorescent pigments absorb UV and visible light from daylight and reemit it at a higher wavelength as visible radiation. As a rule, the time interval between light absorption and emission is very short (10^{-8} s), and the fluorescence persists only in the presence of an exciting light source. These pigments should not be confused with UV fluorescent pigments, which only fluoresce under UV light, or with phosphorescent pigments, which continue to emit light in the dark.

12.1. Production

Very few pure crystalline pigments have daylight fluorescent properties; one such substance, however, is Pigment Yellow 101, C.I. 48052 [*2387-03-3*] [59]. Most daylight fluorescent pigments are composed of one or more fluorescent dyes and a solid dye-carrying substrate or matrix. Optical brighteners, UV absorbers, or ordinary pigments (e.g., phthalocyanines or titanium dioxide) are sometimes also included. The matrix has a high affinity for the dye, thereby protecting it from environmental influences. Fluorescent pigments can be prepared by adsorbing fluorescent dyes onto the surface of fine silica gel particles [60], but selected grades of polymers are usually used as matrices.

12.1.1. Dyes for Daylight FluorescentPigments

Many dyes exhibit daylight fluorescence, but very few are useful for making pigments; the most important are listed in Table 3.

Table 3. Dyes for daylight fluorescent pigments

C.I. number	CAS registry number	Color
Solvent Yellow 44 (**2**)	[*2478-20-8*]	yellow
Solvent Yellow 160 : 1	[*61902-43-0*]	greenish-yellow
Basic Yellow 40	[*12221-86-2*]	greenish-yellow
Basic Red 1	[*989-38-8*]	pink
Basic Violet 10 (**16**)	[*81-88-9*]	pink
Acid Red 52 (**18**)	[*3520-42-1*]	pink

The concentrations of the dyes in the matrix range from ca. 0.1 to 10 wt %. Red fluorescent dyes exhibit the highest emission intensity at a concentration of ca. 0.5 wt %. Yellow dyes may be added at a concentration of up to 10 wt % without causing appreciable quenching. Mixtures of fluorescent dyes very often emit light at a higher intensity than that produced by the separate dyes; this is due to sensitized fluorescence. Sensitized fluorescence is one of the basic principles used in the production of fluorescent pigments and occurs when a portion of the excitation energy of one dye is transferred to a second dye with similar electron energy levels. The second dye is thus excited (sensitized) and fluoresces. Bright fluorescent yellow, orange, red, pink, and magenta shades are obtained by mixing yellow, pink, and violet dyes.

12.1.2. Pigment Matrices

Most fluorescent pigments are based on toluenesulfonamide–melamine–formaldehyde resins that are prepared by *bulk polymerization.* The fluorescent dyes are added to the partially polymerized monomers. Further reaction yields a highly condensed, colored resin, which solidifies when cooled and is then pulverized to give a fine pigment powder [61], [62]. The pigments may also be prepared by *emulsion polymerization,* in which the colored precondensate is further reacted after being mixed at high speed into, for instance, a liquid resin [63] or a molten wax [64]. This results in the formation of spherical pigment particles.

Other matrices are based on benzoguanamine–formaldehyde resins which are prepared by emulsion polymerization in an aqueous phase [65]. For further details of bulk and emulsion polymerization, see → Polymerization Processes.

Pigments based on polyamide [66] and polyester matrices [67] are highly resistant to temperature. Nonplasticized poly(vinyl chloride) and its copolymers [68], urethane resins derived from isocyanates, and compounds containing hydroxyl groups [69] have also been used as pigment matrices.

12.2. Quality Specifications

Daylight fluorescent pigments are offered in a series of colors with carriers having the same basic structure. Properties, such as color strength, solubility, particle size, and resistance to solvents, light, and heat, differ within the various series. Special series are available for most applications to satisfy the specific requirements of processing techniques and formulations.

Fluorescent pigments are generally fine powders with a particle size of 2 – 10 µm. They are easily dispersed in commonly used binders, resin vehicles, and polymers.

Since the pigments are translucent, optimum fluorescence is obtained by applying them to a dense white surface. They are slowly degraded when exposed to sunlight but incorporation of UV absorbers ensures optimal colorfastness. Thick, highly pigmented films are generally recommended for exterior applications. The pigments are almost indefinitely stable to indoor illumination.

At specific wavelengths, daylight fluorescent pigments reflect more light than a white surface does. This is because normal reflectance is reinforced by the emission of fluorescent light. A white surface reflects a maximum of 100 % of the incident light, whereas daylight fluorescent pigments can show total reflectances of over 240 %. Examples of pigment reflectances at different wavelengths are listed in Table 4.

Figure 1 shows typical reflectance curves for an orange, daylight fluorescent surface and a conventional nonfluorescent surface. The high luminance of the fluorescent surface results from

a 3 – 4-fold higher reflectance value at the peak emission wavelength. Fluorescent pigments are a powerful source of monochromatic light.

Table 4. Spectrophotometric parameters of daylight fluorescent pigments

Color∗	Peak wavelength of emission, nm	Reflectance, %
Blue	450	65
Green	520	75
Yellow	530	170
Orange	610	240
Red	620	210
Pink	625	215

∗ Tests were conducted on pigments from the Radglo R series (Radiant Color, Belgium)

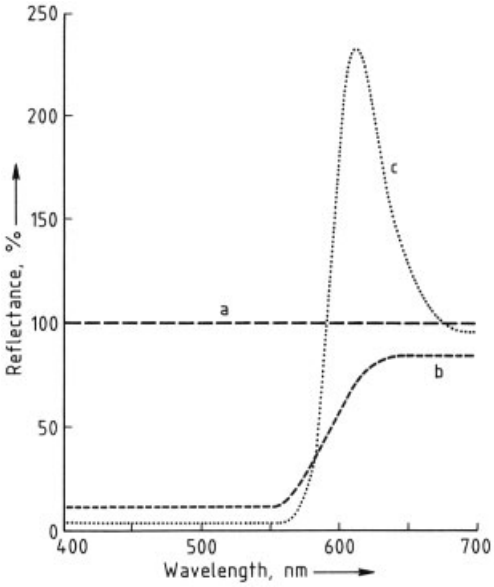

Figure 1. Reflectance spectra of colored surfaces
a) White; b) Conventional orange (Pigment Red 104, C.I. 77605 [*12656-85-8*]); c) Daylight fluorescent orange (Radglo Fluorescent Orange R 13, Radiant Color, Belgium)

12.3. Uses

Many military, civil, and industrial authorities stipulate the use of fluorescent colors where safety, high visibility, and rapid identification are important. Typical applications are the marking of danger areas, aircraft, ambulances, naval vessels and buoys, safety clothing, traffic signs, and symbols on slow-moving vehicles.

As a result of their striking visual impact, daylight fluorescent pigments are used to enhance a wide range of products (e.g., consumer goods, packagings, coated papers, and textiles), as well as for publicity and advertising. As UV-responsive colors, these pigments are employed in the coding and tracing of documents, in theater scenery, and in entertainment decorations. The wide demand for daylight fluorescent pigments has resulted in the development of special products for printing, coating, and painting applications. The use of fluorescent pigments in the plastics industry has almost unlimited possibilities.

12.4. Toxicology

Daylight fluorescent pigments are considered to be nontoxic and are not radioactive. They do not contain inorganic phosphorus, and heavy metals are only found in trace amounts. However, the pigments do not have FDA approval and therefore, should not be used in connection with food.

13. References

General References

1. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 2,** Academic Press, New York 1952, pp. 740 – 795.
2. P. Rys, H. Zollinger: *Leitfaden der Farbstoffchemie,* Verlag Chemie, Weinheim 1970, pp. 87 – 102, 113 – 123.
3. Th. Förster: *Fluoreszenz Organischer Verbindungen,* Vandenhoek und Ruprecht, Göttingen 1951.
4. Th. Förster in W. Foerst (ed.): *Optische Anregung Organischer Systeme,* "Umwandlung der Anregungsenergie," Verlag Chemie, Weinheim 1966, pp. 509 – 552.
5. E. Lippert in W. Foerst (ed.): *Optische Anregung Organischer Systeme,* "Die Medienabhängigkeit der Fluoreszenzfarbe," Verlag Chemie, Weinheim 1966, pp. 342 – 355.

Specific References

6. General Aniline, US 2 006 017, 1932 (W. Eckert, W. Gmelin).
7. Hoechst, DE-OS 2 423 546, 1974 (Th. Papenfuhs).

8. Sumitomo, JP 44/15296, 69/15296, 1966 (H. Sugiyama, H. Otsuka, M. Yamamato, T. Ikeda).
9. Geigy, US 3 014 041, 1958 (H. Häusermann, J. Voltz).
10. BASF, US 3 691 187, 1970 (G. Grau).
11. Bayer, US 3 985 763, 1970 (H. Harnisch).
12. BASF, US 4 055 568, 1975 (M. Patsch, C. Vamvakaris).
13. Sumitomo, US 3 704 302, 1969 (S. Enomoto, K. Sato, G. Suzuki). Sandoz, US 4 007 188, 1972 (W. Koch).
14. CIBA-GEIGY, US 3 872 132, 1970 (S. Koller, R. Zink, D. Reichel, J. Voltz).
15. Ciba-Geigy, US 4 547 579, 1977 (P. Möckli). P. Möckli, *Dyes Pigm.* **1** (1980) 3 – 15.
16. BASF, US 377 349, 1887; US 377 350, 1887 (M. Cérésole).
17. BASF, US 516 584, 1893 (H. A. Bernthsen).
18. Hoechst, US 1 003 738, 1906 (Emmerich).
19. Bayer, US 4 100 509, 1975 (H. Walther, W. Hartig, A. Friedrich, R. Raue).
20. Hoechst, DE-OS 1 955 849, 1969 (F. Meininger, F. Kohlhaas).
21. BASF, US 4 005 111, 1976 (W. Mach, H. Scheuermann).
22. Hoechst, US 3 829 439, 1971 (E. Spietschka, J. Landler).
23. Hoechst, US 3 367 937, 1964 (O. Fuchs, H. Troester).
24. Chi-Kand Dieu in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* "Solvent Dyes," vol. 8, Academic Press, New York 1978, pp. 115 – 120.
25. Hoechst, US 3 502 678, 1965 (O. Fuchs, H. Troester).
26. Hoechst, US 3 781 302, 1970 (O. Fuchs, E. Spietschka, J. Hohn, H. Troester).
27. J. M. Edwards, U. Weiss, *Phytochemistry* **13** (1974) no. 8, 1597 – 1602.
28. H. R. Schwander, Vortrag auf dem Farbensymposium 1976 in Freudenstadt.
29. M. Mostoslavskii, Ya. Shteinberg, C. Ivonina, I. Krutovskaya, *Zh. Prikl. Khim. (Leningrad)* **51** (1978) no. 6, 1424 – 1426.
30. BASF, US 4 009 165, 1974 (E. Schefczik). Bayer, US 3 980 651, 1973 (A. Brack).
31. Bayer, US 3 927 005, 1973 (H. Harnisch).
32. Ciba-Geigy, US 4 056 528, 1975 (H. R. Schwander, K. Burdeska, Chr. Zickendraht).
33. BASF, DE-OS 2 440 405, 1974 (N. Zimmermann, E. Daubach, H. Saukel).
34. Hoechst, DE-OS 2 225 546, 1972 (R. Schickfluss, W. Steckelberg).
35. Bayer, US 4 143 228, 1969 (A. Brack).
36. Bayer, US 3 770 727, 1976 (H. Psaar, W. Kruckenberg).
37. Ciba, DE-OS 1 936 460, 1968 (A. Kaiser, W. Lang, R. Jeanneret).
38. Bayer, DE 300 253, 1916; Friedländer 13, 360.
39. Kodak, US 3 731 222, 1972 (K. H. Drexhage).
40. Ciba-Geigy, DE-OS 2 506 098, 1974 (H. R. Schwander, A. Pugin, K. Burdeska).
41. Geigy, US 2 684 966, 1952 (E. Keller, R. Zweidler).
42. Bayer, US 3 037 836, 1959 (H. S. Bien, F. Baumann).
43. Bios **987** (1948) 174.
44. Bios **987** (1948) 187.
45. BASF, GB 1 391 543, 1971 (F. Graser). BASF, DE-OS 2 210 170, 1972 (F. Graser).
46. Hoechst, US 3 937 666, 1970 (H. Schäfer, E. Janousch).
47. Mitsui Toatsu, US 3 963 429, 1973 (M. Tsujimoto, K. Osawa, I. Okubi). Ciba-Geigy, US 4 304 919, 1979 (N. Bühler, H. Bosshard).
48. Ciba, US 1 043 682, 1912 (G. Engi). Allied Chemical, GB 1 222 314, 1968 (W. R. Demki, A. F. Lachlewski).
49. Geigy, US 3 720 671, 1968 (G. Kaupp). Hoechst, DE-OS 2 355 967, 1973 (E. Fleckenstein, R. Mohr). Hoechst, DE-OS 2 360 986, 1973 (E. Fleckenstein, R. Mohr). E. Bamberger, *Ber. Dtsch. Chem. Ges.* **32** (1899) 1797; *Justus Liebigs Ann. Chem.* **305** (1899) 354.
50. F. C. Strome: "The Dye Laser," *Eastman Org. Chem. Bull.* **46** (1974) no. 2, 1 – 4.
51. H. Meier in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* "Photochemistry of Dyes," vol. IV, Academic Press, New York 1971, pp. 494, 502 –504.
52. W. Stahl, A. Zastrow, *Phys. Unserer Zeit* **16** (1985) 167 – 179.
53. Y. Kanaoka, *Angew. Chem.* **89** (1977) 142 – 152; *Angew. Chem. Int. Ed. Engl.* **16** (1977) 137. J. Haugland in R. F. Steiner (ed.): *Excited States of Biopolymers,* "Covalent Fluorescent Probes," Plenum Press, New York 1983, pp. 29 – 58. B. Lingemann et al., *J. Liq. Chromatogr.* **8** (1985) 789 – 874. G. Schwedt, *Fluorimetrische Analyse,* Verlag Chemie, Weinheim 1981.
54. O. S. Wolfbeis et al., *Fresenius Z. Anal. Chem.* **314** (1983) 119 – 127.
55. J. Haugland, *Handbook of Fluorescent Probes Mol. Probes Inc.,* Junction City, Oregon 1986.
56. O. S. Wolfbeis in "Fluorescence of Natural Organic Products," Part I, S. G. Schulman (ed.): *Molecular Luminescence Spectroscopy,* Wiley-Interscience, New York 1985.

57. A. Zipp, E. Hornby, *Talanta* **31** (1984) 863 – 877. B. Walter, *Anal. Chem.* **55** (1983) 498 A – 514 A. B. Walter, A. C. Greenquist, W. E. Howard III, *Anal. Chem.* **55** (1983) 873 – 878. W. E. Howard III, A. C. Greenquist, B. Walter, F. Wogoman, *Anal. Chem.* **55** (1983) 878 – 881.
58. G. G. Guiltbault: *Enzymatic Methods of Analysis,* Pergamon Press, Oxford 1977. H. Bergmeyer: *Methods of Enzymatic Analysis,* **vol. 1 – vol. 12,** VCH Verlagsgesellschaft, Weinheim 1983 – 1986.
59. IG Farbenindustrie, DE 728 303, 1940 (F. Müllbauer).
60. Sherwin Williams Company, US 3 518 205, 1970 (M. S. Vukasovich).
61. Sterling Colour Ltd., FR 1 321 034, 1962.
62. Switzer Brothers, US 2 938 873, 1960 (Z. Kazenas).
63. Hercules Inc. (Radiant Color), US 4 079 026, 1978 (J. G. Mone).
64. Hercules Inc. (Radiant Color), US 3 682 854, 1972 (D. Bennahmias, H. P. Locklin, P. G. Locklin).
65. Nippon Shokubai Kagaku Kogyo Co. Ltd., DE-AS 2 634 415, 1976 (Tsubakimoto, Tsuneo, Fuzikawa, Iwao et al.).
66. Dayglo Color Corp., GB 1 383 051, 1972.
67. Hercules Inc. (Radiant Color), US 3 922 232, 1975 (A. K. Schein).
68. Hercules Inc. (Radiant Color), US 3 796 668, 1974 (R. T. Hickcox).
69. Siegle GmbH, US 3 741 907, 1973 (H. P. Beyerlin).

Fluorine

Michel Jaccaud, ATOCHEM – Centre de Recherches Rhône Alpes, Pierre-Bénite, France (Chaps. 1 - 3and 4.1 in part)

Robert Faron, COMURHEX, Paris, France (Chaps. 4–7, 4.1 in part)

Didier Devilliers, Université Pierre et Marie Curie, Paris, France (Chaps. 4–7, 4.1 in part)

René Romano, COMURHEX, Paris, France (Chaps. 4–7, 4.1 in part)

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
3. Mineral Resources 4
4. Industrial Production 4
4.1. History 5
4.2. Cell Tanks, Cooling Systems, Skirts, and Diaphragms 6
4.3. Electrodes, Polarization, and Anode Effect 7
4.4. Electrolyte 8
4.5. Overall Industrial Process 9
5. Handling, Storage, and Transport . . 11
6. Environmental Protection and Disposal of Fluorine 12
7. Uses and Economic Aspects 12
8. References 14

1. Introduction

Under normal conditions, fluorine [*7782-41-4*] is a pale yellow gas of stinging odor, which rapidly causes suffocation. Having the atomic number $Z = 9$, it occupies an extreme position in the Periodic Table: at the top of Group 17, the last element before neon, the closest rare gas. Its electronic configuration ($1s^2\ 2s^2\ 2p^5$) makes it the first halogen.

This permits prediction of the strong electronegativity of the fluorine atom, which seeks to take on the electron configuration of neon by capturing an electron. This tendency and the virtual impossibility for an electron of the atom to occupy a higher energy level make fluorine an essentially univalent element [9].

This structure also leads to the formation of the diatomic F_2 molecule, the most highly oxidizing stable substance known. This extreme oxidizing power explains the nonexistence of the F_2 molecule in nature and the impossibility of obtaining it by the action of an oxidant on the fluoride ion, F^-. The only method of preparing elemental fluorine is electrolysis in nonaqueous media.

Fluorine was prepared for the first time in 1886 by the French chemist Henri Moissan, by electrolysis of a solution of potassium fluoride in anhydrous hydrogen fluoride [*7664-39-3*] [10]. The name fluorine comes from the Latin *fluere*, meaning *to flow*, by reference to the main fluorine ore, fluorite, CaF_2, used as a metallurgical flux. Until World War II, fluorine remained a laboratory curiosity.

The discovery of nuclear fission, with all its military and civilian implications, brought about rapid development of the production of fluorine, a necessary intermediate in isotope concentration via uranium hexafluoride.

2. Properties

2.1. Physical Properties

Fluorine has a relative atomic mass A_r of 18.998403 [11]. Its electron affinity is 333 kJ/mol. The fluoride ion is the smallest anion, with an ionic radius of 136 pm; this small size promotes the formation of hydrogen bonds in compounds involving the F^- ion. Depending on the scale used, the electronegativity of fluorine is 4.10 (Allred–Rochow), 3.98 (Pauling), or 3.91 (Mulliken).

The fluorine molecule, F_2, M_r 37.997, is small. The values of the F–F distance determined by electron diffraction [12] and Raman spectroscopy [13] are in agreement, and the currently accepted value is 141 pm. The covalent

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_293

radius of the F atom is therefore 71 pm, making fluorine the smallest halogen. The value of the F–F bond energy has been the subject of numerous controversies; the most recent value is 158.3 kJ/mol [14]. This energy is low and contributes to the high reactivity of the fluorine molecule, frequently leading to highly exoenergetic reactions.

The most important physical properties of the F_2 molecule are given in Table 1; the essential thermodynamic properties are listed below:

Enthalpy of fusion ΔH_f^0	509 J/mol
Enthalpy of vaporization ΔH_v (85.0 K)	6509 J/mol
Standard entropy S^0	203 J mol^{-1} K^{-1}
Standard heat capacity C_p^0	31.3 J mol^{-1} K^{-1}

The reversible potential of the standard fluorine electrode, 2.85±0.04 V, has been calculated from the free enthalpy of formation of HF in aqueous solution [15] but is impossible to achieve experimentally. It is high as a result of the very high solvation energy of the F^- ion in water.

Table 1. Physical properties of fluorine

Melting point *mp*	53.5 K
Boiling point *bp*	85.0 K
Critical temperature τ_{crit}	144 K
Critical pressure p_{crit}	5.571 MPa
Density of the solid ϱ_s	1900 kg/m^3
Density of the liquid ϱ_l	
at 66.0 K	1633 kg/m^3
71.0 K	1608 kg/m^3
77.2 K	1562 kg/m^3
85.0 K	1512 kg/m^3
Density of the gas ϱ_g (STP)	1696 kg/m^3
Viscosity of the liquid η_l	
at 69.2 K	0.414 mPa s
85.2 K	0.257 mPa s
Viscosity of the gas η_g (STP)	0.0218 mPa s
Thermal conductivity λ of the gas (STP)	0.0248 W m^{-1} K^{-1}
Dielectric constant ε of the liquid at 71.2 K	1.54

Vapor pressure of fluorine:

T, K	50.2	59.0	65.5	70.5	80.0	85.0
p, kPa	0.1	1.3	5.3	13.3	53.3	101.3

From an industrial point of view, it is more useful to know the minimum decomposition voltage of HF, taking into account the reaction for the electrolytic production of fluorine:

$$2\,HF \rightleftharpoons H_2 + F_2$$

This voltage, calculated under the conditions achieved in industrial electrolyzers (KF · 2 HF medium, 85° C) is $U_{decomp.} = 2.901$ V [16].

This is actually the most probable value, after a series of determinations frequently lacking in homogeneity, as noted by A. T. Kuhn is his review article [17].

2.2. Chemical Properties

The chemical properties of fluorine are determined by its very powerful oxidizing nature, which frequently leads to sudden or even explosive reactions. The simplest fuel, hydrogen, combines with fluorine in a highly exoenergetic reaction without the need for external initiation.

Reaction with water forms, in addition to HF, hydrogen peroxide [*7722-84-1*], H_2O_2, and oxygen fluoride [*7783-41-7*], OF_2, which cannot be obtained by direct reaction. Compounds in aqueous solutions are generally oxidized by a complex process, with oxoanions forming peroxoanions and metal ions being brought to their highest oxidation state. For example, the chlorate ion, ClO_3^-, is oxidized to perchlorate ion, ClO_4^-. The phosphate ion, PO_4^{3-}, is converted to ortho and pyro perphosphate ions.

Fluorine combines with nonmetals to form compounds with unusual properties.

Thus, sulfur burns in fluorine to form sulfur hexafluoride [*2551–62–4*], SF_6, an extremely stable gas. This is the only hexahalogenated sulfur derivative, which is due to the low steric hindrance of the fluorine atom, permitting it to form the largest number of single covalent bonds with a given element.

Carbon and fluorine combine directly to form carbon tetrafluoride [*75-73-0*], CF_4, a particularly inert gas. The equivalent silicon derivative, SiF_4, in contrast, is very easily hydrolyzed. Controlled fluorination of carbon is possible, leading to the solids C_2F and CF.

Fluorine does not react directly with nitrogen, regardless of the temperature and pressure. However, nitrogen trifluoride [*7783-54-2*], NF_3, can be formed in an $N_2 + F_2$ plasma. This compound illustrates the particular nature of the reactions of fluorine: the low bond energy of the

F_2 molecule permits the reaction $N_2 + F_2$ to be thermodynamically favored, in contrast to the reaction $N_2 + Cl_2$:

$$N_2 + 3F_2 \longrightarrow 2NF_3 \quad \Delta H_f = -108.78\ \text{kJ/mol}$$
$$N_2 + 3Cl_2 \longrightarrow 2NCl_3 \quad \Delta H_f = +230.12\ \text{kJ/mol}$$

Two derivatives of phosphorus, PF_3 [*7783-55-3*] and PF_5 [*7647-19-0*], and their homologues, AsF_3 [*7784-35-2*], AsF_5 [*7784-36-3*], SbF_3 [*7783-56-4*], and SbF_5 [*7783-70-2*], are also known. Generally, the highest oxidation state is obtained.

Halogens react readily with fluorine, and numerous derivatives are known, such as ClF_3 [*7790-91-2*], ClF_5 [*13637-63-3*], BrF_3 [*7787-71-5*], BrF_5 [*7789-30-2*], IF_5 [*7783-66-6*], IF_7 [*16921-96-3*]. Chlorine trifluoride is a much more reactive fluorinating agent than fluorine itself, while IF_5 is a milder fluorinating agent.

Fluorine combines directly with the heaviest rare gases, i.e., those with the lowest ionization energy. The best known rare-gas fluorides are XeF_2, XeF_4, XeF_6, KrF_2, KrF_4. The above compounds are obtained by the prolonged action of fluorine at high temperature and most frequently under pressure; these compounds show structural analogies with the halogen fluorides.

Fluorine reacts with numerous oxygenated derivatives of metalloids and can lead to addition products. Chlorine dioxide, ClO_2, forms ClO_2F; nitric oxide, NO, leads to nitrosyl fluoride, NOF; and NO_2 forms nitryl fluoride, NO_2F. Fluorophosgene, COF_2, yields trifluoromethyl hypofluorite, CF_3OF. All of these compounds are gases. The action of fluorine on a chlorate produces ClO_3F, while concentrated perchloric or nitric acid produces ClO_4F and NO_3F, perchloryl and nitryl oxyfluoride, respectively.

Numerous metals react suddenly with fluorine, and these reactions have presented numerous problems for the industrial production of fluorine. The critical point in the F_2–metal reactions is the formation of a compact surface film of fluoride, opposing the diffusion of fluorine and of the metal itself. This is referred to as passivation. Under these conditions, passive iron and steel are highly resistant to fluorine, provided the temperature is below 200 °C. Above 200 °C only nickel and a nickel–copper alloy (70 wt % Ni–30 wt % Cu) show sufficiently stable passivation.

Certain non-oxidizable metals, such as the noble metals, show good resistance to fluorine at low temperature, but are rapidly destroyed above a certain temperature by the formation of a volatile fluoride, such as PtF_4.

Fluorine also reacts with metal compounds and, in particular, with halogenated compounds, from which it displaces the halogen. With a few exceptions, this reaction leads to the maximum oxidation state of the metal. The direct fluorination of chromium stops at CrF_5. Chromyl fluoride, CrO_2F_2, is obtained by fluorination of the chloride.

With manganese, the reaction does not go beyond MnF_3, even upon fluorination of the permanganate ion. On the other hand, in the presence of KF, an oxidation state of +4 is stabilized in the form of $KMnF_5$. Likewise, nickel leads directly only to the difluoride, NiF_2, but in the presence of KF, K_2NiF_6 can be prepared. Cobalt forms the trifluoride CoF_3 [*10026-18-3*], while tantalum and niobium form the pentafluoride.

Uranium leads to the hexafluoride UF_6, a gaseous compound essential in the nuclear fuel cycle, but many other metals also form a volatile hexafluoride, including rhenium, iridium, molybdenum, tungsten, osmium, neptunium, plutonium, and technetium.

The direct action of fluorine on hydrocarbons is most frequently destructive. The energy released in the reaction is such that it causes rupture of the C–C bonds, and it is therefore appropriate to dilute the fluorine to a very low concentration in an inert gas or to work in a solvent. In the liquid phase, the organic molecule is diluted in a fluorine-resistant solvent, most frequently at low temperature, and by introducing the fluorine diluted with nitrogen or argon.

Direct fluorination generally leads to a mixture of perfluorinated compounds, and it is extremely difficult to obtain a partially fluorinated product. Furthermore, partial destruction of the molecule cannot be prevented, with saturation of the double bonds. Provided that the rings of aromatic compounds are preserved, the homologous cyclic perfluorinated derivatives are obtained. Under certain conditions, fluorination preserves the hetero atoms present in the organic chain, especially O and N. For example, perfluorinated ketones and amines can be obtained from their hydrocarbon homologues.

3. Mineral Resources

While the element fluorine does not exist in nature in the form of F_2, because of the extreme reactivity of this molecule, the element is very widespread in nature in the combined form. The earth's crust contains between 0.06 and 0.07 wt % of fluorine, putting this element in thirteenth place on the scale of abundance. Numerous inorganic compounds contain fluorine, but, on the industrial level, only three families of products represent economically exploitable starting materials:

1) *Natural cryolite*, $AlF_3 \cdot 3NaF$ [*15096-52-3*]. This is the richest ore, very widely exploited since the early days of commercial aluminum production, but now practically exhausted and abandoned in favor of synthetic cryolite.
2) *Fluorite* or fluorspar, CaF_2 [*7789-75-5*]. Presently the most widely exploited ore [18].
3) *Fluorapatite*, $Ca_5(PO_4)_3F$ [*1306-05-4*]. Fluorine is a byproduct of the fertilizer industry. This is certainly the ore of the future because it represents the world's largest reserve of fluorine.

In 1984, the world capacity for CaF_2 production (from fluorite and fluorapatite) was estimated at 6.5×10^6 tonnes. There are three commercial grades of CaF_2: acid, ceramic, and metallurgical. It is the acid grade ($\geq$97 % CaF_2) that is used in the manufacture of hydrogen fluoride, the starting material for the electrochemical production of fluorine.

The presence of fluorine compounds in many ores and their leaching result in natural soft waters frequently containing 0.2 mg/L of fluoride ions. Sea water has a distinctly higher content, from 1 to 1.4 mg/L.

Finally, fluorine compounds exist in traces in numerous living organisms. In the human body, fluorine is an important trace element in the teeth and bones at the level of $1-2 \times 10^{-3}$ wt % (dry product).

4. Industrial Production

Except the laboratory-scale method described in [19], it is presently not known how to prepare fluorine commercially by a purely chemical method, i.e., starting from a fluorine derivative that has itself been obtained without the participation of elemental fluorine.

All fluorides are in equilibrium with fluorine:

$$MF_{x+y} \rightleftharpoons MF_x + y/2\ F_2$$

Complex fluorides are known which show a high fluorine vapor pressure below 900 K.

The fluorine complexes of Ni(IV) are capable of oxidizing the fluoride ion to fluorine [20]:

$$K_2NiF_6 + KF \xrightarrow{673\,K} K_3NiF_6 + \frac{1}{2}\ F_2$$

This reaction has been proposed for the preparation of extremely pure oxygen-free fluorine, but these fluorides are themselves prepared by the action of fluorine, and it is not known how to prepare them otherwise.

In view of this, the only method for the production of fluorine, both in the laboratory and in industry, is the electrolysis of anhydrous hydrogen fluoride containing dissolved potassium fluoride to achieve sufficient conductivity. It is important to work in as anhydrous a medium as possible, to prevent the formation of oxygen, or oxygen fluoride, at the anode. In all cases, hydrogen is produced at the cathode and fluorine at the anode.

4.1. History

Following the historic preparation by H. Moissan in 1886, which was carried out in a platinum U tube, with platinum electrodes, ca. 30 years went by without any significant improvement. For example, the German chemist O. Ruff, a pioneer in the chemistry of fluorine compounds, used an apparatus very similar to that of H. Moissan for his work. This process used an electrolyte rich in HF ($\geq$60 wt %) but below ambient temperature to limit the vapor pressure of HF.

At the end of World War I, W. L. Argo and his co-workers [21] were the first to use a fused electrolyte and a graphite anode. This involved a so-called high-temperature cell, with a $KF \cdot HF$ medium, which has a melting point of 512 K.

This new technology represented an important advance, because it avoided the use of platinum, which is strongly corroded in the Moissan

process. However, two major problems arose: the excessively rapid increase in the melting point with the consumption of HF, and a phenomenon of anode passivation, referred to as the *anode effect.*

Mastery of the KF · HF phase diagram goes back to G. H. Cady [22]; this diagram makes it possible to understand the difficulties in using the high-temperature cells: the bath used is located in a range in which the melting point is sensitive to the HF concentration (see Fig. 1).

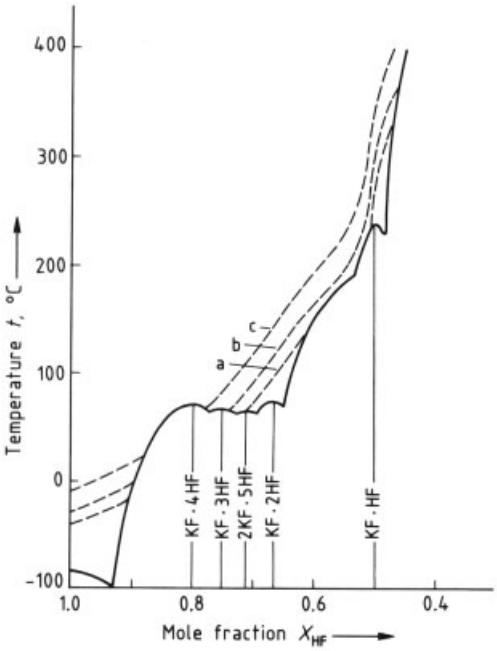

Figure 1. Phase diagram of the system KF · xHF
The dotted lines are isobars of HF vapor pressure above the liquid: a) 50 mm Hg (6.7 kPa); b) 100 mm Hg (13.3. kPa); c) 250 mm Hg (33.3 kPa).

The anode effect is a phenomenon often observed in fused-salt electrolysis. It consists of a sudden or progressive increase in voltage at the terminals of the cell for a constant imposed amperage. A considerable amount of work has been devoted to this phenomenon, which was originally attributed to the presence of an insulating gas film on the anode surface. However, it was observed that the wettability of the anode by the electrolyte decreased considerably after voltage was applied, thus promoting spreading of the gas film. It was O. Ruff who suggested in 1936 that a surface compound of the type CF_x, not wetted by the bath, could be formed.

Since that date, the presence of this compound has been demonstrated and is now an acknowledged fact [23], [24]. Apparently, electron transfer through this barrier layer of fluorinated carbon takes place by a tunnel effect. This mechanism has a high activation energy, which is the origin of the high overvoltage of the fluorine anode [25]. In normal operation, this passivating film protects the carbon against destruction. Under certain more or less well-known conditions, electron transfer is completely blocked.

Industrially, various means are attempted to prevent or retard the anode effect (see4.3).

For 25 years (1919 – 1944), 80 % of the cells described in the literature were of the high-temperature type. However, P. Lebeau and A. Damiens, starting in 1925, may be considered as forerunners in the use of a medium-temperature technology [26]. They used the salt KF · 3HF, which has a melting point of 65.8 °C (see Fig. 1). The anode was made of nickel, and it suffered severe corrosion. In 1942, G. H. Cady described a cell using the salt KF · 2HF, with a melting point of 71.7 °C [27], and an anode made of either nickel or carbon. This date marks the start of the industrial development of fluorine production linked to the exploitation of nuclear fission. This development became more intense at the end of World War II. The fluorine industry is a new industry, only about forty years old, which has grown together with the nuclear industry.

For numerous applications, such as the fluorination of organic compounds, only small quantities of fluorine are required. Since fluorine storage requires an appropriate technology, it may be advantageous to have electrolysis cells available on a laboratory scale. Cells of this type have been supplied by ICI to universities and industry for many years [32].

In all modern fluorine-production units, the electrolyte has the approximate composition KF · 2HF (40.8 wt % HF) and a temperature ranging from 80 to 110 °C [28–31]. The main characteristics of several industrial cells are listed in Table 2. As an example, the usual COMURHEX – ATOCHEM fluorine cell is shown schematically in Figure 2. The new French double-size cell (12 kA) is shown in Figure 3. Other cell designs have been reviewed in the literature [1], [28], [31].

Table 2. Main characteristics of four industrial fluorine cells

Construction feature or operating characteristic	Union Carbide (USA) [29], [32]	ICI (UK) [29]	COMURHEX (France) [33], [34]	Montecatini (Italy) [35]
Electrolyte composition				
% HF	40 – 42	40 – 41.5	40	39 – 41
Electrolyte capacity, t	1.55	1.73	2.4	1.25
Temperature, °C	88 – 104	80 – 85	85 – 100	95 – 100
Dimensions of cell body, m	2.26 × 0.96 × 1	3 × 0.7 × 0.56	2.3 × 0.9 × 0.85	2 × 1 × 0.6
Materials of construction:				
cell body	Monel	mild steel	Monel	Monel
cathodes	mild steel	mild steel	steel	iron
carbon anodes	ungraphitized	porous	ungraphitized	amorphous
diaphragm	Monel	none required	none required	Monel
skirt	Monel	Monel	Monel	Monel
Number of anodes	32	24	32	32
Operating characteristics:				
current intensity, kA	4 – 6	5 – 6.5	6	5
current density, A/dm^2	10 – 15	18	13	7
current efficiency, %	90 – 95	93	90 – 96	90 – 95
fluorine output, kg/h	2.54 at 4 kA 3.47 at 6 kA	3.3	4	3.3
operating voltage, V	8 at 4 kA 12 at 6 kA	10 at 5 kA	8.5 – 10	8.5 – 9.5

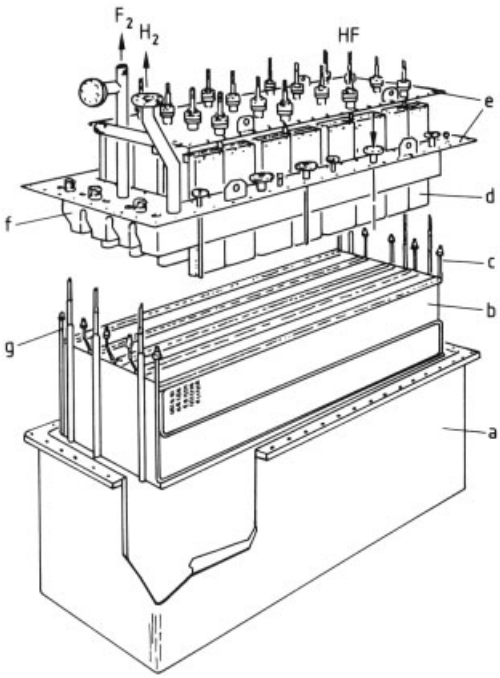

Figure 2. The COMURHEX – ATOCHEM fluorine cell
a) Cell tank; b) Cathode; c) Cooling water inlet; d) Anode; e) Cell covers; f) Gas separation skirt; g) Cooling water outlet

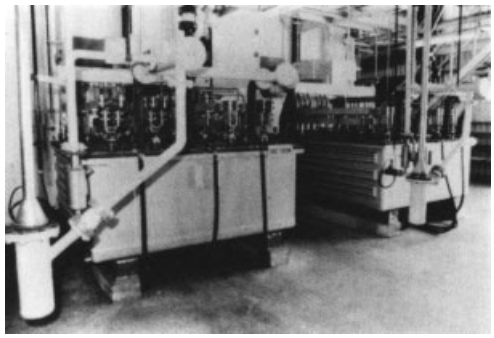

Figure 3. The new French 12-kA fluorine cell

New cell designs are described in the patent literature [36], [37] but are not fundamentally different from the classical cells.

4.2. Cell Tanks, Cooling Systems, Skirts, and Diaphragms

Monel (nickel – copper alloy) is frequently used for the cell-tank construction; mild steel tanks, sometimes operating at cathode potential to minimize corrosion, are also used [38]. The corrosion rate of metals in contact with the vapor phase is lower than that of the tank walls in contact with the electrolyte. In fact, the use of Monel

as a cell-cover material is not required: mild steel or Monel-clothed steel are adequate.

During electrolysis, a large quantity of heat resulting from the Joule effect must be removed so that the cell can operate at a constant temperature; for this purpose, the French cell is surrounded by a water jacket and additional Monel cooling tubes are suspended through the cover. When the cell is shut down, the electrolyte must be kept molten, and heat is then supplied by introducing steam through the Monel tubes. In other types of cells, circulation of the electrolyte is used to promote the removal of heat during electrolysis [39].

Monel, magnesium, or steel skirts are used to separate the hydrogen and fluorine produced by electrolysis. In some cells, a diaphragm (i.e., a skirt extending further into the electrolyte and with an immersed part, which may be perforated) helps to direct the flow of the gases as they are released. The use of a diaphragm results in a better current efficiency, but the ohmic drop is higher, and thus the operating voltage is also higher. The greatest benefit of the diaphragm is to prevent broken anodes from touching cathodes thus avoiding damage to the cell caused by short circuits. Considerable bipolar corrosion of metallic skirts is observed, because the surface of the metal exposed to the cathode becomes anodic. This problem may be overcome by positioning the cathode below the level of the skirt [40]. Small gaps between the skirt and highly permeable anodes are possible because the bubbles of fluorine escape through the interconnecting pores; larger gaps must be provided for low-permeability carbons, for the fluorine bubbles are evolved at the electrode surface.

4.3. Electrodes, Polarization, and Anode Effect

Metallic cathodes are used in all industrial cells. Copper and steel are suitable materials, but steel is preferred because of its lower cost. The corrosion rate of a metal dipped in the $KF \cdot 2HF$ electrolyte is not high if this metal acts as a cathode. The metallic cell body is no longer used as the cathode in industrial cells. In the French cells, the cathode is in the form of a box surrounding the anode assembly, with the cathode connections at the top of the cell. In the British cells, internally cooled cathodes have been used, with cathode connections being made under the cell.

The main cause of failure in medium-temperature cells is related to the anodes: polarization, the anode effect, and deterioration of the mechanical and electrical contact between the electrode and the bus bar. Graphite is not a satisfactory anode material, because of swelling and disintegration, but amorphous carbon is. Nickel anodes have been tested, but their corrosion reduces the current efficiency and insoluble nickel fluoride accumulates.

The difficulties encountered with carbon anodes in fluoride melts have been related to a wetting problem since the early years of fluorine electrochemistry [21]. The *anode effect* in fluoride melts has been reviewed by Rudge [29] and by Watanabe et al. [41]. To avoid this phenomenon, two solutions have been proposed: the first consists in using porous carbon anodes, with the bubbles escaping through the interconnecting pores. However, these electrodes are somewhat fragile and good electrical and mechanical contact between the carbon block and the bus bar is difficult to obtain. The second solution is to use low-permeability carbon anodes and to add nickel or lithium salts to the electrolyte in order to improve the wettability [42]. In the Union Carbide cells, fitted with low-permeability carbon anodes, no additives are deliberately introduced. Rudge [29] claimed that the low concentrations of nickel necessary to reduce the anode – electrolyte contact angle, originate from the corrosion of Monel skirts. In the same way, in the Allied Chemical (now Allied-Signal) cells, corrosion of magnesium alloy skirts occurs, and this introduction of magnesium fluoride has the same effect in promoting wetting of the anodes as the deliberate addition of LiF [29].

A Japanese patent describes the production of special types of carbon suitable for fluorine production [43]. To reduce polarization problems, the Harshaw anodes were impregnated with copper [44], whereas the anodes in the Du Pont cells were strengthened by impregnation with carbonaceous materials and rebaking [31].

High-permeability carbons to overcome polarization problems have been suggested by Howell et al. [45]. The same authors have built a fluorine cell in which the hollow anode is wholly immersed in the electrolyte; F_2 is led away by a copper pipe, which also provides electrical con-

nections to the carbon [46]. More recently, composite carbon electrodes having an outer part with a higher porosity than the inner part have been patented [47], [48]. Finally, impregnation of anodes with lead salts has been proposed to suppress CF_4 formation during electrolysis [49].

Deterioration of the anode contact frequently occurs because it is subject to attack by fluorine, electrolyte, and hydrogen fluoride vapor. A metal rod screwed into the top of the carbon block can be used only if the carbon is very hard; anodes may also be bolted to a copper or steel hanger bar.

Improved electrical connections have been described in the patent literature. They are formed by electrodepositing [50] or spraying [51] a layer of copper or nickel and welding the electrical conductor to the electroplated or sprayed region of the electrode. Bottom-entering anodes sealed with frozen electrolyte have been successful only in laboratory cells. Polytetrafluoroethylene, loaded with 25–50 wt % powdered calcium fluoride, has been widely used as an insulator in packing glands supporting the anode and cathode conductor posts.

4.4. Electrolyte

The specific electrical conductivity κ of the electrolyte has been measured by Schumb et al. [52] over a wide range of temperatures and HF concentrations. The authors found: $0.12<\kappa<0.21\Omega^{-1}\mathrm{cm}^{-1}$.

A high HF content in the electrolyte actually promotes corrosion, misting, and anode polarization [53], whereas low HF contents lead to a decrease in the conductance and thus to an increase in the cell operating voltage [53].

During electrolysis, the composition of the melt changes, since HF is consumed. In many cells, the electrolyte level is maintained by the intermittent addition of gaseous HF, but control of the electrolyte composition is important, and HF should be added automatically to maintain a preselected optimum concentration. Circulation of the electrolyte is not necessary to maintain a uniform composition, because the rate of dispersion of HF is high, but good circulation is necessary to promote the rapid dissipation of heat from the cell.

The electrolyte $KF \cdot 2HF$ is prepared from commercial Frémy salt $KF \cdot HF$ [*7789-29-9*] and HF. Special care is taken in its preparation, because the purity of the fluorine produced and the anode behavior depend strongly on the electrolyte composition.

The COMURHEX specification (in wt %) for $KF \cdot HF$ is $KF \cdot HF$ 99.0 % min., K_2SiF_6 0.1 % max., H_2O 0.3 % max., SO_4^{2-} 0.01 % max., Na^+ 0.3 % max., Cl^- 0.001 % max., Fe, and Pb 0.001 % max. The specification for the HF used in electrolyte preparation and feeding of COMURHEX cells is HF 99.9 % min., H_2SO_4 0.05 % max., H_2SiF_6 0.05 % max., SO_2 0.005 % max., and H_2O 0.05 % max. The preparation of $KF \cdot 2HF$ is an exothermic reaction. Gaseous or liquid HF is added to the Frémy salt in a reaction vessel jacketed to control its temperature.

If it is not sufficiently pure, the electrolyte is treated with fluorine or preelectrolyzed, using a cell with a nickel anode. In particular, the water content must be reduced to avoid OF_2 and CO_2 evolution in the anodic compartment. The sulfate content must be low to avoid attack of the carbon anode [32] and SO_2F_2 [*2699-79-8*] evolution. Chloride and SiF_6^{2-} give rise to the evolution of chlorine fluorides (or oxychlorofluorides) and silicon tetrafluoride [*7783-61-1*], SiF_4. Other impurities, such as heavy metals, contribute to a rise in the cell operating voltage.

Some compounds deliberately or accidentally introduced into the melt have been reported to have a beneficial effect on the electrolytic process. For example, cell behavior is strongly influenced by the presence of nickel salts in the electrolyte, resulting from corrosion of nickel or Monel skirts, diaphragms, or cooling tubes [29]. The addition of lithium fluoride to the electrolyte is useful because it lowers the freezing point of the melt and reduces anodic polarization problems [40], [54]; however, the LiF concentration decreases with time, because it is progressively carried down into the sludge at the bottom of the cell, and the beneficial effects finally disappear [52]. Quantities from 1 to 2 wt % are used as standard additions to the melts. The critical current densities are increased by the addition of LiF; LiF is considered to promote the decomposition of the passivating film of $(CF)_n$ formed on the anode surface [55]; the addition of LiF, NiF_2, or AlF_3 has been reported to be effective in preventing the anode effect [56] by reducing

the anode – electrolyte contact angle [29]. During operation, the carbon anode is coated with a layer containing nickel or lithium ions; this layer is wetted by the electrolyte, and fluorine bubbles break away easily [42].

Finally, numerous other additives to the electrolyte have been proposed in a Japanese patent [57], and as a method for purification of the electrolyte [58].

The electrolysis of a mixture of NH_3 and HF (liquid at room temperature) has been studied [59–61]. Patents describe cells for the industrial production of fluorine using the NH_4F – HF system [62] or a mixture of NH_3, HF, and KF [63–65], but, this process is not yet in use because of corrosion of metals and carbon in this electrolyte.

4.5. Overall Industrial Process

The flow diagram of the French fluorine production plant is shown in Figure 4.

After preparation, the KF · 2HF electrolyte is transferred to the preelectrolysis cell for removal of water. After this important step, lithium fluoride is added to the electrolyte, and the electrolyte is transferred into the fluorine production cell.

Special techniques for the elimination of polarization have been developed: a new 6000 A cell is first started up at low current ($I = 1000$ A) for several hours. The current is increased step by step until steady-state conditions at 6000 A are established. During the low-voltage electrolysis, the last traces of water are removed; the anodic gases produced (i.e., fluorine contaminated with OF_2, CO_2, O_2) are diluted with nitrogen and destroyed.

The usual operating voltage is 8.5 – 10.5 V for $I = 6000$ A (corresponding to $j = 12$ A dm^{-2}). The current efficiency is about 90 – 95 % but the energy consumption is about 14 – 17 kWh/kg F_2, i.e., about four times the theoretical value. This large energy loss is mainly due to the high anodic overvoltage.

As a result of the electrolysis of HF, hydrogen is produced in the cathode compartment, while fluorine evolution takes place in the anode compartment. The cell off-gases contain hydrogen fluorine, whose concentration depends on the composition of the electrolyte and the cell operating temperature (average 6 % [66] or 10 % [67]. In the COMURHEX cell, this concentration is 7 – 10 %. The HF is separated from hydrogen or fluorine and reprocessed.

Other impurities in the cell off-gases, such as CO_2, O_2, or OF_2, may result from the electrolysis of traces of water in an insufficiently preelectrolyzed electrolyte. In the same way, sulfuryl fluoride (SO_2F_2) and silicon tetrafluoride result from the electrolysis of sulfate and fluorosilicate impurities. As shown by Downing et al. [67], the quantity of these compounds in the anodic gases, which may rise to several percent at the start-up of the cell, decreases rapidly, provided that purified HF is used to feed the cells. In addition, very low concentrations ($\leq$30 ppm) of tetrafluoromethane CF_4 and higher fluorocarbons are always observed during electrolysis. Their presence has been attributed to the continuous thermal decomposition of the graphite fluoride passivating layer [24]:

$$n\,CF_x \longrightarrow a\,C + b\,CF_4 + c\,C_2F_6 + d\,C_yF_z + \ldots$$

The CF_4 concentration is controlled by infrared spectrophotometry. When it is greater than 50 ppm, the operating voltage is reduced, in order to avoid anode deterioration. The percentage of O_2 and N_2 in anodic gases is determined by gas chromatography.

The operation of fluorine cells is not dangerous if special precautions are rigorously followed, e.g.:

1) Avoid contact with electrolyte, HF, or F_2.
2) Protective equipment must be worn by operators working near the cells: safety goggles, face shields, protective clothing, gloves. In addition, a gas mask with canister for acid-gas protection may be worn.
3) Prevent development of explosive gas mixtures within the cell (i.e., air leaks into the cathode compartment must be eliminated).
4) Reactive contaminants, such as grease, must be kept out of the system. If they are not, highly exothermic combustions take place, and the metal in piping or valves may melt and burst under pressure.
5) Fluorine plants must have forced ventilation systems to provide fresh air for personal comfort and safety.

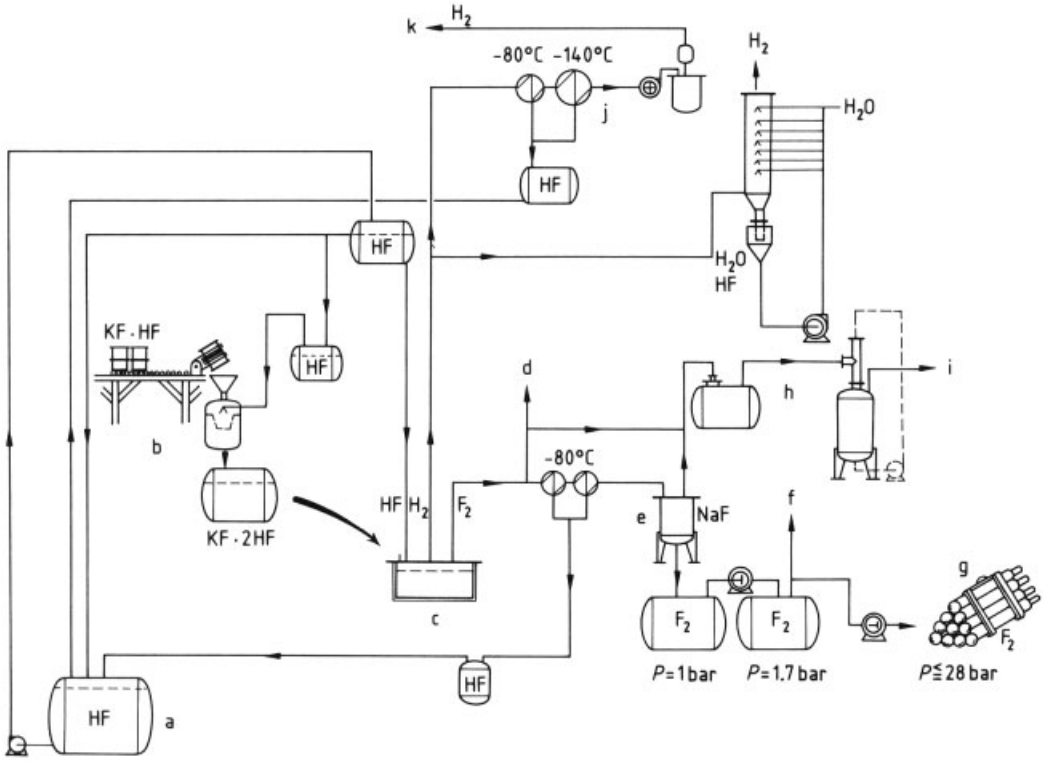

Figure 4. Simplified flow diagram of the French fluorine production plant
a) HF storage; b) Electrolyte preparation; c) Electrolysis cells; d) Utilization (UF_6 or SF_6 units); e) Fluorine purification, f) Purified fluorine; g) Fluorine storage; h) Fluorine disposal; i) Vent; j) Hydrogen purification; k) Burner

5. Handling, Storage, and Transport

Handling [28], [68], [69]. Because of the reactivity of elemental fluorine, special care must be taken in handling this gas. A passivation of the whole system (line and apparatus) with F_2 is imperative before each operation. Corrosion of passivated piping and valves is not a serious problem if moisture and reactive contaminants, such as grease, are kept out of the system. The whole system must be carefully tested for leaks; manipulations require a dry, ventilated room. The selection of materials to be in contact with fluorine depends on the expected temperature. Aluminum, copper, stainless steel, nickel, brass, Inconel, and Monel have been proposed for industrial use. Tables of corrosion data have been presented and reviewed in the literature [69]. For the handling of pure fluorine under pressure, the use of nickel or Monel pipes is preferable.

Fluorine is not difficult to handle near atmospheric pressures: in 1947, Cady claimed that the hazards of working with fluorine had previously been greatly over-rated. However, handling fluorine under high pressure is a different matter [29].

At moderate temperatures, reactions can be carried out, usually on a laboratory scale, in quartz, Pyrex, Teflon (polytetrafluoroethylene), or Kel'f (polymonochlorotrifluoroethylene). Metallic equipment is necessary for work at relatively high temperatures. Laboratory equipment is reviewed in [70].

Storage [71–73]. Storage only relates to a small percentage of the production. Electrolyzers are integrated in SF_6 or UF_6 units, and the major part of the elemental fluorine produced is directly used in the plant. In order to avoid the handling of bottled fluorine, ICI supplied 10-A and 60-A laboratory-scale cells to universities and to industry for many years [29]. At present, the use of reliable cylinders of compressed fluorine is preferred.

Compression and Liquefaction. Fluorine requires purification before liquefaction. HF must be removed to prevent icing; this is accomplished by a combination of refrigeration and NaF absorption traps [72]. Heating fluorine to 300 – 400 °C before liquefaction is useful for destroying traces of impurities such as oxygen fluorides; if they are not removed, these compounds spontaneously explode when the liquid phase is warmed for evaporation.

The compression of pure fluorine with ordinary compressors is difficult because lubricants react with the gas. Compressors of the diaphragm type have been used successfully for low-pressure storage, but for higher pressures the method used involves liquefaction of F_2 by cooling with liquid nitrogen and later reevaporation of the gas to the desired pressure in the shipping container.

Transport. The French specification for the transport of fluorine is available from the "Imprimerie Nationale, Paris":

Transport by rail and road (April 25, 1945, and June 27, 1951)

Transport by air (January 14, 1983).

The European specification for transport by road (A.D.R.) is available from the United Nations (Geneva), and the specification for transport by rail (R.I.D.) can be obtained from the S.N.C.F. (Paris).

Fluorine must be packaged under a pressure not exceeding 2.86 MPa (28 bars); the steel cylinders, previously tested at 200 bars, must be equipped with special valves and must not contain more than 5 kg of elemental fluorine. They must be labeled with the toxic-product label number 6 – 1. Cylinders containing 2.2 kg and special packages of ten cylinders are available from COMURHEX.

In the United States, fluorine is supplied by Air Products and Chemicals as a mixture of 10 % F_2 in nitrogen. Liquid F_2 was available in the 1960 s in tank trucks containing up to 2270 kg of fluorine. The inner steel shell of the tank contained the liquefied product, the intermediate stainless steel shell was filled with liquid nitrogen, and the outer shell was filled with an insulating powder.

6. Environmental Protection and Disposal of Fluorine

In laboratories or in small fluorination units, a soda – lime tower is usually sufficient for destroying waste fluorine. Suitable methods for the disposal of fluorine and hydrogen fluoride contained in gaseous effluents of fluorine-production plants have been reported [28].

Liquid effluents coming from electrolysis, compression, or purification units contain dilute hydrofluoric acid, which is generally disposed of in a lime neutralization system.

Several methods for fluorine disposal have been described in the literature. Their aim is to convert fluorine to less active and less toxic materials. The reaction of fluorine and carbon to form tetrafluoromethane has been investigated. Although several authors report that the reaction is sometimes explosive [74], others consider the disposal of fluorine by reaction with charcoal satisfactory [29], [75]. TURNBULL et al. [76] claim that the best method (Du Pont) consists of injecting fluorine into the flame of a natural-gas burner. The burner operates with propane or butane and a deficiency of air; water and caustic scrubbers are added to the system to remove the hydrogen fluoride formed by the reaction:

$$C_3H_8 + 10F_2 \longrightarrow 3\,CF_4 + 8\,HF$$

In the ICI process, waste fluorine is scrubbed in spray towers by reaction with 10 % caustic potash solution; the contact time and alkali concentration must be adjusted to avoid the formation of oxygen fluorides [29].

At present, in the COMURHEX process, waste fluorine is reacted with sulfur. Sulfur hexafluoride, oxyfluorides, and lower fluorides of sulfur are obtained; the hydrolyzable compounds are treated in an aqueous caustic scrubber; inert gases are vented to the atmosphere.

Hydrogen and fluorine produced in electrolytic cells are contaminated with a mist of fine solid particles of electrolyte; in addition, the concentration of hydrogen fluoride in the gas produced is ca. 5 – 10 vol %. The entrained electrolyte is removed by steel entrainment separators (American plants) or by an electrostatic precipitator using a nickel wire at a potential of 15 000 V (French plant [34]). Hydrogen fluoride contained in fluorine is partially removed

by condensation at −80 °C; if a low level of contamination is required (<0.02 vol %), sodium fluoride tray absorbers are used as a secondary absorption system [34]. Sodium fluoride traps are very efficient for removing HF, forming NaF · HF [*1333-83-1*]. The absorbent is regenerated by heating. In the past, HF-containing hydrogen was vented directly to the atmosphere. At the present time, HF is usually removed by condensation. After that, the gas is, if necessary, scrubbed with water or caustic solution before being vented. If may also be burned (Union Carbide Nuclear, 1962 [77], and COMURHEX, 1986).

7. Uses and Economic Aspects

Elemental fluorine is necessary to prepare the highest oxidation state fluorides, because of the combination of its oxidizing power and its fluorinating abilities. For many applications, especially in organic chemistry, milder or more specific fluorinating agents are required; these compounds are prepared from elemental fluorine. The element is also needed for specific new applications such as the synthesis of graphite fluorides, but its main application is in the field of nuclear energy [78]. All of these syntheses proceed by exothermic reactions, and temperature control is frequently necessary. Most experiments require the use of special metallic equipment (see 5). Preparative methods on the laboratory scale were reviewed [70].

Economic Aspects. The world's production of elemental fluorine, except for countries with planned economies, is concentrated in seven countries with an overall capacity of ca 17 000 t/a:

Canada Eldorado Nuclear
Federal Republic of Germany Kali Chemie
France COMURHEX
Italy Montedison
Japan Daikin Kogyo Co., Kanto Denka
Kogyo Co., Asahi Chemicals
U.K. British Nuclear Fuels, Imperial Smelting Corp.
U.S.A. Allied Chemicals, Kerr McGee, Air Products and Chemicals

Most of the elemental fluorine produced is used in two areas: (1) manufacture of 18 000 – 21 000 t/a of UF_6 [*7783-81-5*], corresponding to 6000 – 7000 t of F and (2) manufacture of 5100 – 7700 t/a of SF_6 [*2551-62-4*], corresponding to 4000 – 6000 of F. Approximately 2000 t/a are required for miscellaneous applications.

The price of elemental fluorine is estimated at $ 5 – 8 per kg, but, because of difficulties in storage and handling, the price is much higher when F_2 is supplied in cylinders. To avoid these problems, units for elemental fluorine production are integrated into SF_6 or UF_6 units.

Production of Uranium and Sulfur Hexafluorides. Elemental fluorine was used on a large scale for the first time during World War II: separation of the isotopes of natural uranium was carried out by a diffusion process involving uranium hexafluoride vapor. Fluorinated oils, greases, and polymers were needed in the associated technology. The main effort in fluorine chemistry was concentrated in the war years, but there has been continuing activity because of the peaceful application of fission processes.

Elemental fluorine is essential in the final step of processing uranium ores to produce uranium hexafluoride. UF_4 [*10049-14-6*] is first produced by the reaction of UO_2 [*1344-57-6*], and HF, and UF_6 is then prepared exclusively by the action of F_2 on UF_4:

$$UF_4 + F_2 \longrightarrow UF_6$$

Two techniques (flame reactor, fluidized beds) are used to carry out this reaction [79] (→ Uranium , Uranium Alloys, and Uranium Compounds).

Fluorine is also used to produce sulfur hexafluoride by direct combination of the elements (→ Fluorine Compounds, Inorganic, Chap. 6.1.1.).

Production of Fluorinating Agents. Chemists can now use a large variety of fluorinating compounds, some of them almost as reactive as fluorine (ClF_3, KrF_2, ...), others only mild fluorinating agents (IF_5). Only

ClF_3, BrF_3, BrF_5, and IF_5 are of industrial importance (→ Fluorine Compounds, Inorganic, Chap. 6.5.).

The great oxidizing strength of elemental fluorine is necessary to prepare high oxidation state fluorides such as CoF_3. Sometimes ClF_3 or BrF_3 are used instead of F_2 because they are easier to store and handle; however, these halogen fluorides may form stable adducts that are difficult to decompose.

Other compounds, *solid fluorine carriers*, are useful in organic syntheses. They are composed of elements with two possible oxidation states: Co(III)/Co(II), Ag(II)/Ag(I), Ce(IV)/Ce(III), Pb(IV)/Pb(II). The most widely used is CoF_3; this releases part of its fluorine in the free form when the temperature is raised to 300 – 400 °C,

$$2\,CoF_3 \longrightarrow 2\,CoF_2 + F_2$$

CoF_3 is regenerated by the reverse reaction [79].

Production of Graphite Fluoride. The name *graphite fluorides* is assigned to nonstoichiometric solid fluorocarbons with the general formula CF_x $(0 < x \leq 1.24)$. The structure of two of these intercalation compounds is known, corresponding to simple F/C ratios: CF [*11113-63-6*] or $CF_{0.5}$ [*71183-38-5*]. The general properties of these compounds have been reviewed [80–82]. Graphite fluorides are prepared by direct fluorination of various carbons at 300 – 600 °C [83]. The chemical composition and crystal structure strongly depend on the reaction temperature and the crystallinity of the starting carbons.

Minor Uses. Propulsion systems employing fluorine as a rocket-fuel oxidizer have been suggested; the abandonment of this is due to the difficulties involved in storage and handling of liquid fluorine and to pollution problems, since the reaction of F_2 with hydrogen-containing materials forms hydrogen fluoride [84].

The possible use of fluorine for the reprocessing of nuclear fuels has been investigated [85], but these studies have not yet gone beyond the pilot level [79]. This fluoride volatilization method is also called the *dry process.*

The fluorination of tungsten and rhenium forms volatile hexafluorides used in chemical vapor deposition (CVD): the thermal decomposition of WF_6 [*7783-82-6*] and ReF_6 [*10049-17-9*] form highly pure metal coatings on metallic or nonmetallic substrates [86].

Fluorine is also used to prepare NF_3 from NH_3 or $NH_4F \cdot HF$, to etch silicon [87], to produce solvent-resistant polyolefin containers [88], and to improve moisture transport and soil release of synthetic fibers (polyamide, polyester, polyolefin, polyacrylonitrile fibers) [89–91].

Elemental fluorine or fluorine diluted with nitrogen or helium can be used in preparative organic chemistry for the synthesis of highly fluorinated molecules. In this case, the lack of selectivity is not a major problem [92].

8. References

General References

1. *Gmelin,* Syst. No. 5, Fluor, 1926, 1959, 1980.
2. *Ullmann,* 4th ed., **11,** 587 – 630.
3. *Kirk Othmer,* 3rd ed., **10,** 630 – 654.
4. C. Slesser, S. R. Schram (eds.): *Preparation, Properties and Technology of Fluorine and Organic Fluoro Compounds,* McGraw-Hill, New York 1951.
5. J. H. Simons (ed.): *Fluorine Chemistry*, vol. I, Chaps. 8– 10, Academic Press, New York 1950.
6. M. C. Sneed, J. L. Maynard, R. C. Brasted: *Comprehensive Inorg. Chem.,* **vol. 3,** Van Nostrand, New York 1954, pp. 9 – 46.
7. T. A. O'Donnel in J. C. Bailar, H. J. Emeléus, R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorg. Chem.,* **vol. 2,** Pergamon Press, Oxford 1973, pp. 1009 – 1038.
8. D. Naumann: "Fluorine and Fluorine compounds," *Special Inorganic Chemistry,* vol. 2, Steinkopff, Darmstadt 1980.

Specific References

9. A. G. Sharpe, *Q. Rev. Chem. Soc.* **11** (1957) 49 – 60.
10. H. Moissan, *C. R. Hebd. Seances Acad. Sci.* **102** (1886) 1543 – 1544.
11. IUPAC, XXXI General Assembly, 1981.
12. M. T. Rogers, V. Schomaker, D. P. Stevenson, *J. Am. Chem. Soc.* **63** (1941) 2610 – 2611.
13. D. Andrychuk, *Can. J. Phys.* **29** (1951) 151 – 158.

14. K. P. Huber, G. Herzberg in: *Molecular Spectra and Molecular Structure*, 4th Constants of Diatomic Molecules, 2nd ed., Van Nostrand Reinhold, New York, N.Y., 1979.
15. W. Latimer, *J. Am. Chem. Soc.* **48** (1926) 2868 – 2869.
16. D. Devilliers, F. Lantelme, M. Chemla, *J. Chim. Phys. Phys. Chim. Biol.* **76** (1979) no. 5, 428 – 432.
17. A. T. Kuhn in A. J. Bard (ed): *Encyclopedia of Electrochem. of the Elements,* **vol. 4,** Dekker, New York 1975, pp. 43 – 86.
18. L. Pelham, *J. Fluorine Chem.* **30** (1985) 1 – 17.
19. K. O. Christe, *Inorg. Chem.* **25** (1986) 3721 – 3722.
20. U.S. Energy Res. Dev. Adm., 3 989 808, 1975 (L. B. Asprey).
21. W. L. Argo, F. C. Mathers, B. Humiston, C. O. Anderson, *Trans. Am. Electrochem. Soc.* **35** (1919) 335 – 349.
22. G. H. Cady, *J. Am. Chem. Soc.* **56** (1934) no. 7, 1431 –14 34.
23. H. Imoto, T. Nakajima, N. Watanabe, *Bull. Chem. Soc. Jpn.* **48** (1975) no. 5, 1633 – 1634.
24. D. Devilliers, M. Vogler, F. Lantelme, M. Chemla, *Anal. Chim. Acta* **153** (1983) 69 – 82.
25. D. Devilliers, Thèse, Paris 1984.
26. P. Lebeau, A. Damiens, *C. R. Hebd. Seances Acad. Sci.* **181** (1925) 917 – 919.
27. G. H. Cady, D. A. Rogers, C. A. Carlson, *Ind. Eng. Chem. Ind. Ed.* **34** (1942) no. 4, 443 – 448.
28. R. J. Ring, D. Royston, Australian Atomic Energy Commission Report AAEC/E 281, 1973; *Chem. Abstr.* **80** (1974) 140 412 n.
29. A. J. Rudge in A. T. Kuhn (ed.): *Industrial Electrochemical Processes,* Elsevier, Amsterdam 1971, pp. 1 – 69.
30. R. Papp, R. Romano, *Centernary Discovery Fluorine Int. Sym.,* Paris 1986, p. 193.
31. *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol. II, Suppl. 1, Longmans, London 1956.
32. A. J. Rudge, *Chem. Ind. (London)* 1966, 482 – 488.
33. M. Bergeret, *Energ. Nucl. (Paris)* **7** (1965) 93 – 99.
34. A. Level, *Chim. Ind. Génie Chim.* **102** (1969) 1077 – 1082.
35. D. Sianesi, *Chim. Ind. (Milan)* **46** (1964) no. 8, 883 – 894.
36. Montecatini Edison S.p.A., FR 2 095 766, 1972 (A. Tricoli, A. Battara, G. Rebua, L. Bestetti).
37. Asahi Glass Co. Ltd., JP-Kokai 59 229 487, 1984.
38. U.S. Atomic Energy Commission, US 2 540 248, 1951 (R. C. Downing).
39. British Nuclear Fuels Ltd., US 4 125 443, 1978 (A. E. Grant, S. Wainwright).
40. J. T. Pinkston Jr., *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 255 – 258.
41. N. Watanabe, Y. Kanaya, *Denki Kagaku Kogyo Butsuri Kagaku* **39** (1971) 139 – 147.
42. J.-P. Randin in A. J. Bard (ed.): *Encyclopedia of Electrochemistry of the Elements,* **vol. 7,** Dekker, New York 1976, pp. 1 – 291.
43. Central Glass Co. Ltd.; Toyo Tanso Co. Ltd., DE-OS 3 027 371, 1981 (N. Watanabe, M. Aramaki, Y. Kita).
44. U.S. Atomic Energy Commission, US 2 506 438, 1950 (G. C. Whitaker).
45. Imperial Chemical Industries Ltd., GB 655 098, 1951 (W. N. Howell, H. Hill).
46. Imperial Chemical Industries Ltd., GB 642 812, 1950 (W. N. Howell, H. Hill).
47. British Nuclear Fuels plc, GB 2 135 334, 1984 (A. Marshall).
48. British Nuclear Fuels plc, GB 2 135 335, 1984 (A. Marshall).
49. Asahi Glass Co. Ltd., JP-Kokai 5 867 877, 1983.
50. Imperial Chemical Industries Ltd., GB 651 107, 1951 (A. J. Rudge, W. N. Howell).
51. Imperial Chemical Industries Ltd., GB 685 461, 1951 (A. J. Rudge, W. N. Howell).
52. W. C. Schumb, R. C. Young, K. J. Radimer, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 244 – 248.
53. R. C. Downing in C. Slesser, S. R. Schram (eds.): *Preparation, Properties and Technology of Fluorine and Organic Fluoro Compounds,* Chap. 3 McGraw-Hill, New York 1951.
54. U.S. Office of Scientific Res. Devel., US 2 422 590, 1947 (W. C. Schumb, A. J. Stevens).
55. H. Nakayama, A. Tasaka, K. Nakanishi, N. Watanabe, *Denki Kagaku Kogyo Butsuri Kagaku* **42** (1974) no. 10, 531 – 534.
56. H. Imoto, K. Ueno, N. Watanabe, *Bull. Chem. Soc. Jpn* **. 51** (1978) 2822 – 2825.
57. Asahi Glass Co. Ltd., JP-Kokai 5 881 981, 1983.
58. Central Glass Co. Ltd., JP-Kokai 5 884 117, 1983.
59. G. Pourcelly, M. Rolin, *J. Chim. Phys. Phys. Chim. Biol.* **74** (1977) no. 4, 401 – 408.
60. N. Watanabe, A. Tasaka, K. Nakanishi, *Denki Kagaku Kogyo Butsuri Kagaku* **36** (1968) 685 – 689.

61. A. Tasaka, T. Yamamoto, K. Kawamura, T. Terada, Proc. 8th Symposium on Molten Salt Chemistry, Tokyo 1985, pp. 60 – 61.
62. Société des Usines Chimiques de Pierrelatte, FR 2 082 366, 1971 (M. Caron, P. Coste, C. Coquet, M. Rey).
63. Pechiney Ugine Kuhlmann, FR 2 304 691, 1976 (R. Faron, A. Cathala).
64. Produits Chimiques Ugine Kuhlmann, FR 2 343 821, 1977 (R. Faron, A. Cathala).
65. Allied Corp., US 4 511 440, 1985 (A. M. Saprokhin, D. J. Friedland, R. M. Baran, J. T. Kim et al.).
66. R. L. Murray, S. G. Osborne, M. S. Kircher, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 249 – 254.
67. R. C. Downing, A. F. Benning, F. B. Downing, R. C. McHarness et al., *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 259 – 262.
68. J. F. Gall, H. C. Miller, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 262 – 266.
69. R. Landau, R. Rosen, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 281 – 286.
70. J. Grannec, L. Lozano in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 17 – 76.
71. S. G. Osborne, M. M. Brandegee, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 273 – 274.
72. J. F. Froning, M. K. Richards, T. W. Stricklin, S. G. Turnbull, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 275 – 278.
73. H. F. Priest, A. V. Grosse, *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 279 – 280.
74. J. H. Simons, L. P. Block, *J. Am. Chem. Soc.* **61** (1939) 2962 – 2966.
75. H. Pulley, R. L. Harris, U.S.A.E.C. Report KY-638, 1972; *Chem. Abstr.* **77** (1972) 117 909 j.
76. S. G. Turnbull, A. F. Benning, G. W. Feldmann, A. L. Linch et al., *Ind. Eng. Chem. Ind. Ed.* **39** (1947) 286 –288.
77. W. K. Henderson, W. B. Goode, S. Bernstein, E. J. Tullos, U.S.A.E.C Report KY-399, 1962; *Chem. Abstr.* **57** (1962) 12286 c.
78. H. G. Bryce in J. H. Simons (ed.): *Fluorine Chemistry,* **vol. 5,** Academic Press, New York 1964, pp. 297 – 498.
79. B. Cochet-Muchy, J. Portier in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 565 – 605.
80. P. Kamarchik, Jr., J. L. Margrave, *Acc. Chem. Res.* **11** (1978) 296 – 300.
81. Y. Kita, N. Watanabe, Y. Fujii, *J. Am. Chem. Soc.* **101** (1979) 3832 – 3841.
82. N. Watanabe, H. Touhara, T. Nakajima, N. Bartlett et al. in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 331 – 369.
83. N. Watanabe, T. Nakajima in R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Chap. 9, Ellis Horwood, Chichester (England) 1982.
84. R. H. Langley, L. Welch, *J. Chem. Ed.* **60** (1983) 759 – 761.
85. M. Bourgeois, B. Cochet-Muchy, *Energ. Nucl. (Paris)* **10** (1968) 192 – 200.
86. COMURHEX, FR 2 574 988, 1986 (M. Bargues, R. Romano).
87. J. A. Mucha, V. M. Donnelly, D. L. Flamm, L. M. Webb, *J. Phys. Chem.* **85** (1981) 3529 – 3532.
88. Air Products and Chemicals, Inc., US 3 862 284, 1975 (D. D. Dixon, D. G. Manly, G. W. Recktenwald).
89. Air Products and Chemicals, Inc., DE-OS 2 247 615, 1973 (D. D. Dixon).
90. Air Products and Chemicals, Inc., DE-OS 2 500 598, 1975 (D. D. Dixon, L. J. Hayes).
91. Air Products and Chemicals, Inc., US 4 020 223, 1977 (D. D. Dixon, L. J. Hayes).
92. M. R. C. Gerstenberger, A. Haas, *Angew. Chem. Int. Ed. Engl.* **20** (1981) 647 – 667.

Fluorine Compounds, Inorganic

Jean Aigueperse, ATOCHEM, Usine de Pierre-Bénite, Pierre-Bénite, France (Chaps. 1 and 2)

Paul Mollard, Ste Foy Les Lyon, France (Chap. 3)

Didier Devilliers, Université Pierre et Marie Curie, Paris,France (Chaps. 4 and 6)

Marius Chemla, Université Pierre et Marie Curie, Paris, France (Chaps. 4 and 6)

Robert Faron, COMURHEX, Paris, France (Chaps. 4 and 6)

René Romano, COMURHEX, Paris, France (Chaps. 4 and 6)

Jean Pierre Cuer, ATOCHEM, Centre de Recherche Rhône-Alpes, Pierre-Bénite, France (Chap. 5)

1. Hydrofluoric Acid 3
1.1. Physical Properties 3
1.1.1. Anhydrous Hydrofluoric Acid (AHF) 3
1.1.2. Aqueous Solutions 3
1.2. Chemical Properties 4
1.3. Manufacture 5
1.3.1. Fluorite Processes 5
1.3.2. Fluosilicic Acid Processes 7
1.3.3. Corrosion-Resistant Materials . . . 8
1.3.4. Specifications and Analysis 9
1.4. Storage and Transport 9
1.5. Toxicology and Occupational Health 9
1.6. Economic Aspects and Uses 10
2. Calcium Fluoride 11
2.1. Physical and Chemical Properties 11
2.2. Production and Resources 12
2.3. Uses 13
2.4. Toxicology 15
2.5. Specifications 15
2.6. Chemical Analysis 15
3. Aluminum Fluoride and Fluoroaluminates 16
3.1. Aluminum Fluoride 16
3.1.1. Properties 16
3.1.2. Manufacture 17
3.1.2.1. Hydrofluoric Acid Processes 17
3.1.2.2. Fluosilicic Acid Processes 17
3.1.2.3. Recovery of Fluorine During Aluminum Production 19
3.2. Fluoroaluminates 19
3.2.1. Cryolite, $AlF_3 \cdot 3\,NaF$ 20
3.2.1.1. Properties 20
3.2.1.2. Manufacture 20
3.2.2. Ammonium Cryolite 23
3.3. Environment Protection 23
3.4. Specifications 23
3.5. Chemical Analysis 23
3.6. Uses 23
3.7. Economic Aspects 24
3.8. Toxicology 24
4. Other Metal Fluorides 24
4.1. Ammonium Fluorides 25
4.2. Cobalt Fluorides 26
4.3. Lithium Fluorides 26
4.4. Magnesium Fluoride 26
4.5. Nickel Fluorides 27
4.6. Potassium Fluorides 27
4.7. Sodium Fluorides 28
4.8. Tungsten Hexafluoride 28
5. Fluorine – Silicon Compounds . . 28
5.1. Silicon Tetrafluoride 29
5.1.1. Physical Properties 29
5.1.2. Chemical Properties 29
5.1.3. Manufacture 29
5.1.4. Transport 30
5.1.5. Analysis 30
5.1.6. Use 30
5.1.7. Toxicology 30
5.2. Fluosilicic Acid 31
5.2.1. Physical Properties 31
5.2.2. Chemical Properties 31
5.2.3. Manufacture 32
5.2.4. Transport 33
5.2.5. Quality and Analysis 33
5.2.6. Use 33
5.2.7. Economic Aspects 33
5.2.8. Toxicity 33
5.3. Fluosilicates 33
5.3.1. Properties 34
5.3.2. Manufacture 34
5.3.3. Analysis 34
5.3.4. Uses, Economic Aspects 34
5.3.5. Toxicity 35
6. Other Inorganic Fluorine Compounds 35
6.1. Sulfur Fluorides 35

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_307

6.1.1. Sulfur Hexafluoride 35
6.1.2. Sulfur Tetrafluoride 36
6.2. Nitrogen Fluorides 37
6.3. Oxygen – Fluorine Compounds . . 37
6.4. Phosphorus – Fluorine Compounds 38
6.5. Halogen Fluorides 39
6.6. Rare-Gas Fluorides 40
6.7. Graphite Fluorides 41
7. References 41

1. Hydrofluoric Acid

Hydrofluoric acid [*7664-39-3*], M_r = 20.01, has been known since the end of the eighteenth century as a product capable of attacking glass. The anhydrous acid was obtained by Fremy from potassium bifluoride in 1856. However, the uses of hydrofluoric acid did not become truly diversified and developed until 1930, with the manufacture of aluminum fluoride and synthetic cryolite for the aluminum industry, the first production of fluorocarbons, and the pickling of stainless steels. From 1940 on, hydrofluoric acid has been used for the production of uranium fluorides and as an alkylation catalyst in gasoline production. Industrial growth brought the world production of hydrofluoric acid to the level of 1×10^6 t at the beginning of the 1970 s. Then stagnation set in, a state that has lasted up to the present, because of the reductions in use required by environmental protection and the decrease in certain economic activities.

Hydrofluoric acid is a colorless liquid boiling at 19.5 °C at atmospheric pressure [1–3]. It is characterized by special physical properties resulting from the association of HF molecules by means of hydrogen bonds. It is also characterized by chemical reactivity and toxicity, which make it dangerous. All industrial hydrofluoric acid production starts from fluorite. Processes based on the recovery of fluorosilicic acid from phosphate rock have not yet gone beyond the development stage.

1.1. Physical Properties

1.1.1. Anhydrous Hydrofluoric Acid (AHF)

Regardless of physical state–solid, liquid, or gas–HF molecules tend to associate by means of hydrogen bonds:

– FH – FH –

In the solid state, these bonds form continuous zigzag planar chains [4]. Shorter and less rigid chains are present in the liquid [5]. The association of the HF molecules in the gaseous state has attracted the largest number of studies [1–3], [6–13].

Equilibria are established between the molecules HF, $(HF)_2$, $(HF)_3, \ldots$, $(HF)_n$ [8]. These equilibria are governed by the temperature (Fig. 1A [7]) and pressure (Fig. 1B [2], [14]). At the boiling point under atmospheric pressure, the average degree of association is 3.74 [1], corresponding to an average molecular mass M_r = 74.9. At atmospheric pressure, hydrofluoric acid vapor is completely depolymerized only above 100 °C.

This molecular association also affects certain physical properties, such as the boiling point and density, and also the variation of these properties with temperature and pressure. The enthalpy of polymerization is $\Delta H = -25 \pm 4$ kJ/mol [2], [10], [11], [15], [16]. This should be taken into consideration both in the calculation of the amounts of heat to be supplied for superheating hydrofluoric acid vapors and in the calculation of exchange surfaces for desuperheating these same vapors.

The principal physical properties of anhydrous hydrofluoric acid are listed in Table 1. The vapor pressure is shown as a function of temperature in Figure 2 [1].

1.1.2. Aqueous Solutions

Aqueous hydrofluoric acid solutions are colorless. They fume in air when their concentration is higher than 40 %. The specific heat of solution is 834 J for a 50 % solution and 160 J for an infinite dilution of one gram of liquid HF in water at 25 °C [25].

The principal physical properties–density [26], melting point [27], boiling point, and liquid – vapor composition [28]–are shown as func-

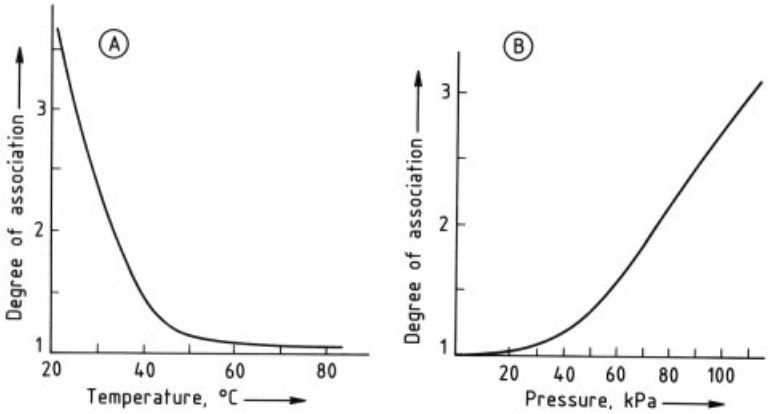

Figure 1. Degree of association of hydrogen fluoride as a function of A) Temperature at atmospheric pressure; B) Pressure at 26 °C

Table 1. Physical properties of anhydrous hydrofluoric acid

Property	Value	Reference
Melting point *mp*	-83.37 ± 0.02 °C	[17]
Heat of fusion ΔH_f	196.49 ± 0.17 kJ/kg	[17]
Boiling point *bp*	19.5 °C	[1–3]
Heat of vaporization ΔH_v at 101.3 kPa	374 kJ/kg	[2]
Density ϱ		
Liquid at 0 °C	1.015 g/cm^3	[3]
Liquid at 20 °C	0.970 g/cm^3	[3]
Saturated vapor at 20 °C	3.170 g/L	[1]
Critical temperature t_{crit}	188 ± 3 °C	[18]
Critical pressure P_{crit}	6.5 MPa	[18]
Specific heat of the liquid *c* at 20 °C	2.54 J g^{-1} K^{-1}	[19]
Heat of formation of the gas at 25 °C ΔH_f°	13.59 ± 0.025 kJ/g	[20]
Viscosity η of the liquid at 0 °C	0.256 mPas	[21]
Surface tension of the liquid σ		
at 0 °C	10.2 mN/m	[22]
at 19.5 °C	8.6 mN/m	[22]
Electrical conductivity *K* at 0 °C	$1 \times 10^{-6}\ \Omega^{-1}\text{cm}^{-1}$	[23]
Dielectric constant ε at 0 °C	83.6	[24]

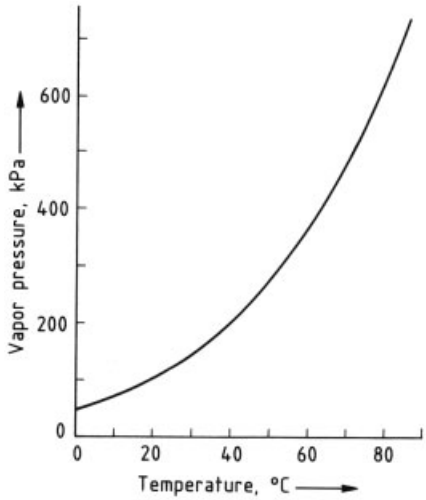

Figure 2. Vapor pressure of hydrogen fluoride as a function of temperature

tions of concentration in Figure 3. The partial pressures of the hydrofluoric acid solutions, p_{HF} and p_{H_2O}, can be calculated from the empirical relationship $\log p\ (\text{kPa}) = B - A/T$ [29],

wt % HF	HF		H_2O
	A	*B*	*AB*
30	2500	7.757	7.786 2247
50	2120	7.449	7.738 2386
70	1830	7.439	

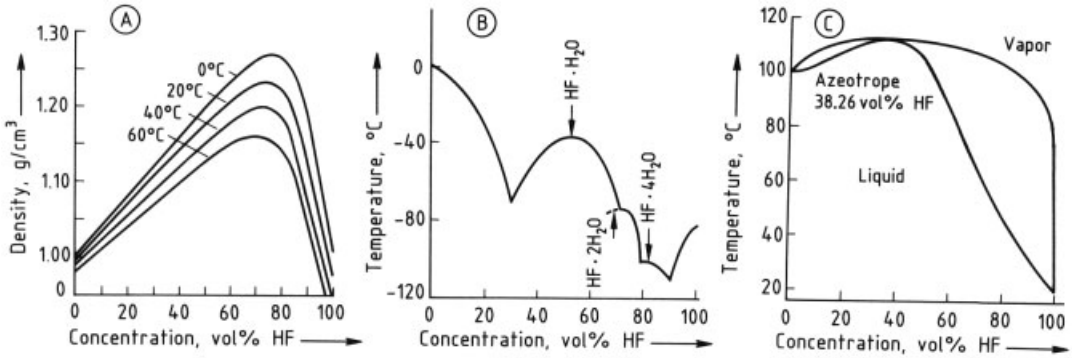

Figure 3. Physical properties of aqueous hydrofluoric acid solutions
A) Density; B) Solid-liquid phase diagram; C) Liquid – vapor phase diagram

1.2. Chemical Properties

Hydrofluoric acid is characterized both by the stability of its molecule, which, with a dissociation energy of 560 kJ, is classified among the most stable diatomic molecules, and by its reactivity.

Hydrofluoric acid has a special affinity for oxygen compounds. It reacts with boric acid to form boron trifluoride, silica and silicates to give SiF_4 and fluosilicic acid, sulfur trioxide and sulfuric acid to form fluorosulfonic acid. The last reaction demonstrates the dehydrating power of anhydrous hydrofluoric acid.

Liquid hydrofluoric acid is a much studied nonaqueous ionizing solvent [30]. It is a strong acid in the Brönsted – Lowry sense, which gives it a catalytic activity in alkylation, isomerization, and polymerization reactions. In aqueous solution, hydrofluoric acid is a weak acid, with an activity comparable to that of formic acid.

Hydrofluoric acid is used in organic chemistry as a fluorinating agent in substitution reactions with aliphatic halides and in addition reactions with unsaturated compounds.

1.3. Manufacture

The industrial processes for the manufacture of anhydrous hydrofluoric acid are all based on the action of sulfuric acid on fluorite. However, phosphate rock, which contains fluorapatite, $CaF_2 \cdot 3\ Ca_3(PO_4)_2$, constitutes the most important natural reserves of fluorine. In fact, this ore, which is used all over the world for the manufacture of phosphoric acid, contains between 3 and 4 wt % of fluorine. The reserves are five times greater than the proven reserves of fluorite.

In the treatment of phosphate rock with sulfuric acid to produce phosphoric acid, the fluorine can be recovered in the form of dilute fluosilicic acid solution. Numerous investigations have been carried out to develop processes for the production of anhydrous hydrofluoric acid from fluosilicic acid solutions. These processes, none of which has yet undergone industrial development, are still of undeniable potential interest.

1.3.1. Fluorite Processes

Fluorite is attacked by sulfuric acid at temperatures between 100 and 300 °C. The concentration of sulfuric acid must be between 98 and 99 wt %. More dilute acid leads to increased reactor corrosion, while more concentrated acid promotes the formation of fluorosulfonic acid, HSO_3F, which would have an unfavorable effect on the reaction kinetics [31].

Acid-grade fluorspar with a grain size below 150 µm is used; CaF_2 content should be above 97 %. Main impurities:

SiO_2	1 % max
S (sulfides)	0.05 % max
$CaCO_3$	1 % max
$Al_2O_3 + Fe_2O_3$	1.5 % max
H_2O	0.1 % max

Since fluorspar is generally the more expensive of the two starting materials, a slight excess

of sulfuric acid over the stoichiometric ratio is used to ensure maximum conversion of fluorite. The unreacted raw materials are entrained by the calcium sulfate.

The main reaction

$$CaF_2 + H_2SO_4 \longrightarrow 2\,HF + CaSO_4$$

is accompanied by a number of secondary reactions:

$$SiO_2 + 4\,HF \longrightarrow SiF_4 + 2\,H_2O$$
$$CaCO_3 + H_2SO_4 \longrightarrow CaSO_4 + CO_2 + H_2O$$
$$Fe_2O_3 + 3\,H_2SO_4 \longrightarrow Fe_2(SO_4)_3 + 3\,H_2O$$
$$Al_2O_3 + 3\,H_2SO_4 \longrightarrow Al_2(SO_4)_3 + 3\,H_2O$$
$$MS + H_2SO_4 \longrightarrow MSO_4 + H_2S$$

The sulfuric acid is also reduced by the steel reactor walls:

$$Fe + 2\,H_2SO_4 \longrightarrow FeSO_4 + SO_2 + 2\,H_2O$$

The action of sulfuric acid on fluorspar therefore produces:

1) a gas phase containing hydrofluoric acid, with SiF_4, H_2O, SO_2, H_2S, and CO_2 as the main impurities
2) a solid phase consisting of calcium sulfate, containing the unreacted raw materials and iron and aluminum sulfates as the main impurities.

The reaction is heterogeneous and endothermic. These two characteristics determine the technology of all industrial reactors. The nature of the reaction medium changes during the course of the reaction, and shows that an industrial reactor must necessarily include effective stirring:

Extent of reaction	Reaction medium
0 – 40 %	liquid, pasty, then sticky and corrosive
40 – 70 %	powdered
70 – 80 %	pasty, sticky, and corrosive
80 – 100 %	powdered

The heat liberated by the secondary reactions represents less than 10 % of the heat required by the endothermic main reaction, which is close to 100 kJ for 2 mol HF at 160 °C. A supply of heat is therefore necessary for good development of the reaction of sulfuric acid with fluorspar.

The main industrial reactors can be grouped into five categories.

Simple Rotary Furnace. The two raw materials are fed into one end of the rotary furnace, and calcium sulfate is removed from the other end. This is a cocurrent reactor. The gas phase is generally discharged close to the inlet. The heat supply is provided by circulation of combustion gases in a double jacket, possibly with preheating one or both raw materials. In a Du Pont process, the heat is supplied by introducing part of the sulfuric acid in the form of a mixture of sulfur trioxide and steam [32].

The reliability of this type of reactor is based on effective agitation of the reaction mixture and on good protection against corrosion. The specific devices have not been described.

Rotary Furnace with Premixer. Developed by Buss (Fig. 4), the special co-kneader premixer ensures an intimate mixing of the raw materials. At the exit of the premixer, the extent of reaction is close to 30 %. The corrosion at the entrance to the rotary furnace is significantly reduced; however, it is not eliminated, and the dangers of crust formation on the furnace walls remain.

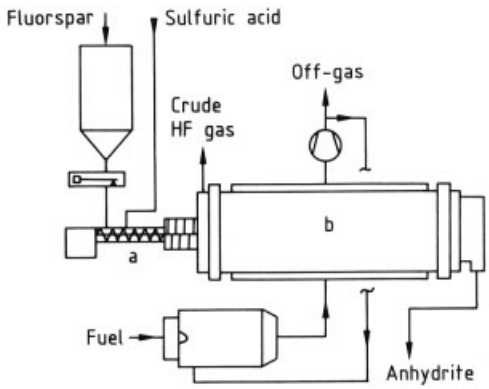

Figure 4. Premixer-rotary furnace [33]
a) Buss prereactor; b) Main reactor

Rotary Furnace with Prereactor [34]. Sufficient heat supply and residence time allow the conversion rate to reach 40 – 50 % at the exit of the prereactor, which supplies a powdered reaction mixture to the furnace. The furnace must, however, still contend with the second corrosive, crust-forming phase. Scraping devices combined with anti-corrosion linings are therefore necessary, or furnaces must be of large

size to permit the retention of a large volume of reaction mixture.

Rotary Furnace with Recycle [35]. The furnace is directly supplied with fluorspar and sulfuric acid, but an internal device removes a suitable quantity of mixture at the end of the reaction, in the vicinity of the outlet, and recycles it to the furnace head. The throughputs are selected so that the composition of the reaction mixture in the furnace head corresponds to a conversion rate of 80 %, and the dangers of corrosion and crust formation no longer exist. However, the furnace must be large to ensure sufficient residence time because of the high recycling rate.

Rotary Furnace with Recycle and Prereactor [36]. Combination of the two devices just described leads to units with especially high productivity.

Treatment of the Gases. The crude gas-phase hydrofluoric acid discharged from the reactor contains

hydrofluoric acid
entrained fluorite and calcium sulfate dusts
impurities heavier than HF (H_2O, H_2SO_4 as vapor and bubbles)
impurities lighter than HF (SiF_4, CO_2, SO_2, H_2S, air)

When the temperature of the gas drops to 60 – 70 °C, solid sulfur forms,

$$SO_2 + 2\,H_2S \longrightarrow 3\,S + 2\,H_2O$$

Initially, gas treatment consists of removing the solid particles and heavy impurities by one of two schemes. Scheme one:

Separation of the solid particles in a cyclone or coke box,
Separation of the heavy impurities in a 60 °C precondenser and recovery of 70 % hydrofluoric acid solution rich in sulfuric acid, which can be recycled to the furnace after the addition of oleum.

Scheme two:

Separation of dust and water in a washing column sprayed with sulfuric acid
Reconcentration of the wash acid by addition of oleum.

In a second stage, the gases are cooled to condense crude hydrofluoric acid, which must be purified by distillation (generally two columns). Finally, in a third stage, the noncondensable gases saturated with HF are washed before discharge to the atmosphere. This washing generally includes a column sprayed with sulfuric acid, which recovers the hydrofluoric acid, followed by a column sprayed with water, which recovers SiF_4 as fluosilicic acid solution.

Treatment of the Calcium Sulfate. The reaction of sulfuric acid with fluorite produces ca. 3.7 t of calcium sulfate per tonne of hydrofluoric acid manufactured. This byproduct is orthorhombic anhydrite, crystallographically identical to natural anhydrite. It can be used in cement to control the settling rate. It can also be converted to a material usable in building, mining, or road construction. However, these applications have been developed only in Germany and Japan. Everywhere else, economic conditions do not permit utilization of the calcium sulfate, which must be discarded.

1.3.2. Fluosilicic Acid Processes

The solutions containing 15 – 20 % of fluosilicic acid that are recovered in the reaction of phosphate rock constitute the raw material of all proposed processes.

Hydrolysis of SiF_4. High-temperature hydrolysis in a flame reactor, a process initially studied by Reed [37], consists of concentrating the fluosilicic acid solution to 30 %, vaporizing this solution according to reaction

$$H_2SiF_6 \longrightarrow SiF_4 + 2\,HF$$

and hydrolyzing SiF_4 in an air – hydrogen or air – hydrocarbon flame,

$$SiF_4 + 2\,H_2O \longrightarrow SiO_2 + 4\,HF$$

The by-product silica is usable. The hydrofluoric acid is collected as a dilute solution. Concentration with sulfuric acid yields anhydrous

hydrofluoric acid. The diluted sulfuric acid solutions are used for the treatment of phosphate rock.

Controlled hydrolysis of SiF_4 can be used to increase the concentration of the initial fluosilicic acid solution:

$$3\,SiF_4 + 2\,H_2O \longrightarrow 2\,H_2SiF_6 + SiO_2$$

After separation of the silica, the concentrated solution can be vaporized and decomposed to SiF_4 and HF. After the absorption of HF in a selective solvent, SiF_4 is recycled to the hydrolysis stage. Distillation regenerates the selective solvent and produces anhydrous hydrofluoric acid. Buss [38] recommends the use of a polyglycol as the selective solvent. Concentrated cold sulfuric acid is also recommended [39].

In another process claimed by Stauffer and Tennessee Corporation [40], [41], decomposition of the concentrated fluosilicic acid solution is achieved by the action of cold sulfuric acid. Under these conditions, only SiF_4 is released into the gas phase, and can therefore be directly recycled to the hydrolysis stage, while HF remains in solution in sulfuric acid diluted to 60 %. The HF – H_2SO_4 separation is carried out by distillation or stripping.

Alkaline Method. The treatment of H_2SiF_6 solutions with NH_3, the alkaline method, precipitates silica in a solution of ammonium fluoride:

$$H_2SiF_6 + 6\,NH_3 + (n+2)\,H_2O \longrightarrow SiO_2 \cdot n\,H_2O + 6\,NH_4F$$

After separation of the silica, the ammonium fluoride solution is concentrated by evaporation and decomposed to produce the bifluoride, NH_4HF_2, and ammonia.

Several processes have been proposed for the production of hydrofluoric acid from ammonium bifluoride. Goulding Chemicals [42] treats NH_4HF_2 with NaF to form sodium bifluoride, $NaHF_2$, which is then decomposed to NaF and HF. Du Pont [43] proposes the treatment of ammonium bifluoride with sulfuric acid at 180 °C to liberate hydrofluoric acid and form ammonium bisulfate.

Reaction of Na_2SiF_6, K_2SiF_6, Li_2SiF_6, or $CaSiF_6$ with HCl at temperatures between 350 and 600 °C is another method proposed to produce hydrogen fluoride [44].

None of these processes has been used industrially. However, economic studies show that the processes based on the hydrolysis of SiF_4 could compete with the processes for the manufacture of anhydrous hydrofluoric acid from fluorite, but with one restriction: these processes must necessarily be used at the sites of phosphoric acid production.

In addition, technological problems probably remain to be solved, because of the highly corrosive nature of the reaction media.

1.3.3. Corrosion-Resistant Materials

In industries producing or consuming hydrofluoric acid, material selection must take into account the acid concentration, temperature, and nature of the impurities present in both the acid and the material. Guides to selection are given in [45–47]. Nevertheless, material selected should be tested under the actual conditions of use: any error in the selection of the materials can affect not only the reliability but also the safety of the plant.

Metals. Carbon steel is commonly used with liquid or gaseous anhydrous hydrofluoric acid up to 100 – 150 °C. Carbon steel is passivated by the formation of a layer of iron(III) fluoride. Since erosion of this protective layer increases the corrosion rate, the flow rate of liquid acid in contact with a carbon steel wall should be limited to 1.5 m/s. The corrosion of carbon steel produces hydrogen, and may thereby have unexpected consequences:

- formation of an explosive air – hydrogen mixture in the dome of a storage tank
- appearance of blistering on the inner surface of the metal sheets used for storage and transport of hydrofluoric acid [48]
- embrittlement of the steel, which may, for example, involve the rupture of bolts.

The use of silicon-containing cast iron is prohibited. Silica-free cast iron cannot be recommended, because of its porosity and the danger of blistering.

Austenitic or nickel – chrome steels show no better resistance to anhydrous hydrofluoric acid than carbon steels. In particular, they must not be used at temperatures exceeding 150 °C. Monel

metal, nickel, and Inconel can be used in the presence of gaseous hydrofluoric acid up to 600 °C.

Pure aluminum shows good resistance to corrosion by anhydrous HF. However, aluminum should be selected only after tests carried out under the planned conditions of use.

The possibility of using metals in contact with aqueous hydrofluoric acid are limited. Only platinum is unattacked up to the boiling point of the solutions, regardless of their concentration. Carbon steel can be used at ambient temperature for solutions with a concentration ≥70 %. Copper is quite resistant to solutions with concentrations >60 %, up to their boiling points, provided that the flow rate is limited to 1.5 m/s. Copper is sometimes used in the construction of distillation columns for the production of anhydrous hydrofluoric acid.

Monel metal is more resistant than nickel and Inconel to solutions with concentrations ≥50 % at ambient temperature. Titanium and tantalum are strongly attacked by hydrofluoric acid, regardless of its concentration.

Ceramics. Materials containing silica are attacked by HF even in dilute solution and at low temperature.

Graphite and alumina show good resistance to solutions with concentrations <60 % up to their boiling points. Synthetic sapphire glasses maintain their transparency in contact with hydrofluoric acid.

Plastics. Polyethylene, polypropylene, and poly(vinyl chloride) show good resistance to dilute solutions containing up to 50 % of HF at ambient temperature. Polytetrafluoroethylene (PTFE) shows good resistance to anhydrous hydrofluoric acid and to its aqueous solutions. However, PTFE is not completely impermeable to hydrofluoric acid. If this material is used in a protective layer, sufficient thickness must be provided. Poly(vinylidene difluoride) (PVDF) can be used for solutions with concentrations <60 %, up to their boiling points.

Synthetic and natural rubbers are generally attacked by hydrofluoric acid. However, certain grades of ebonite are used at ambient temperature as a lining for solutions with concentrations <50 %.

1.3.4. Specifications and Analysis

Commercial anhydrous hydrofluoric acid is a highly pure product. The following standard analysis is representative of the average quality delivered by the principal producers:

HF	99.9 % max
H_2O	0.02 %
H_2SO_4	0.01 %
SO_2	0.005 %
H_2SiF_6	0.01 %

Hydrofluoric acid is also marketed in 70 % and 50 % aqueous solutions. The analytical methods for the anhydrous acid are standardized:

H_2O	ISO 3699	Karl Fischer method
H_2SO_4	ISO 3138	titrimetric method
SO_2	ISO 3702	iodometric method
H_2SiF_6	ISO 3701	photometric method with reduced molybdosilicate

1.4. Storage and Transport

The storage conditions for anhydrous hydrofluoric acid or its solutions are subject to regulation. In Europe, the European Technical Fluorine Committee [49], which includes the principal hydrofluoric acid producers, has established recommendations relating to storage equipment: construction materials, conditions of calculation, thermal insulation, supporting device, filling level, valves, safety valves, retaining tank, inspection, etc.

Anhydrous hydrofluoric acid and its aqueous solutions are subject to national and international regulations relating to the transport of hazardous materials, e.g., for transport by railroad (RID) and by road (ARD) in Europe, and the international regulations for transport by sea (IMDG code) and air (IATA).

These regulations specify the types of authorized packaging, the filling levels, the required safety devices, labeling, and traffic prohibitions. For example, the transport of anhydrous hydrofluoric acid by air is prohibited.

1.5. Toxicology and Occupational Health

In humans, hydrofluoric acid causes lesions of the respiratory system, through inhalation of vapor, and of the skin by direct contact [50], [51].

The inhalation of HF vapors causes a respiratory insufficiency that can range from simple irritation to acute pulmonary edema. The renal function can also be affected [51]. The presence of HF vapors in the air can be detected by smell, with the olfactory limit varying between 0.03 and 0.11 mg/m^3, depending on the individual. The permissible concentration limits for an exposure time exceeding eight hours vary between 0.5 and 2.5 mg/m^3, depending on the country.

An environment containing 50 mg of HF per m^3 of air is dangerous. In fact, while tolerable for one minute, it must be considered as an upper limit for a five-minute exposure and may have fatal results if the exposure is continued for more than 30 minutes. A conscious person is not able to remain for 30 minutes in an atmosphere so irritating to nose and eyes, but the danger is real for an unconscious person.

Because of the seriousness of the risk to skin from hydrofluoric acid [52], [53], the safety instructions furnished by all producers of hydrofluoric acid must be strictly followed in any handling of hydrofluoric acid. The safety equipment, gloves, safety glasses, boots, moisture-proof coveralls, breathing apparatus, and showers must be maintained in good condition.

An accident caused by hydrofluoric acid, even if it appears minor, must always be treated as a serious accident requiring the rapid intervention of competent medical staff. But it is primarily through the speed with which first aid is administered that the severity of the accidents can be considerably reduced. The rules are simple: victims of inhalation accidents must be evacuated from the contaminated areas as rapidly as possible, and burns caused by sprays must be washed with copious amounts of water, starting immediately and continuing for 15 minutes.

1.6. Economic Aspects and Uses

Table 2 shows the hydrofluoric acid production capacities of the Western world in 1982. No information concerning the production of hydrofluoric acid in the Eastern bloc countries has been published.

Table 2. Hydrofluoric acid production capacities (Western countries, 1982)

Region or country	Capacity (1000 t)
North America *	430
United States	232
Canada	106
Mexico	92
Western Europe *	339
France	98
Italy	80
Federal Republic of Germany	78
Great Britain	49
Spain	25.5
Netherlands	5
Greece	3.5
Japan *	63
Others (Tunisia, Australia, Republic of South Africa, Argentina)	28
Total	*860*

* Reproduced with permission from [54].

The two main uses, which represent more than 75 % of the total consumption of hydrofluoric acid, are (1) the manufacture of fluorocarbons and (2) the manufacture of aluminum fluoride and synthetic cryolite for the aluminum industry. However, these two applications have been subject to restrictions imposed by environmental protection for the last 10 to 15 years.

The most important fluorocarbons are chlorofluoromethanes 11 (CCl_3F), 12 (CCl_2F_2), and 22 ($CHClF_2$), and chlorofluoroethanes 113 ($CCl_2F-CClF_2$), 114 ($CClF_2-CClF_2$), and 115 ($CClF_2-CF_3$), all produced by the action of hydrofluoric acid on the corresponding chlorinated hydrocarbons. They are used as aerosol propellants, refrigerating fluids, industrial degreasing solvents, blowing agents for polyurethane, polystyrene, or polyolefin foams, and as raw materials for the manufacture of fluorinated polymers (polytetrafluoroethylene).

The restrictions concern chlorofluoromethanes 11 and 12 used as aerosol propellants. The measures were taken initially in the United States and then in Europe after the publication in 1974 of a hypothesis of Molina and Rowland [55] that the molecules of CFM 11 and 12 undergo photochemical reactions in the upper atmosphere to release chlorine atoms,

which would be capable of causing decomposition of the stratospheric ozone. The disappearance of this ozone layer would lead to an increase in the UV radiation reaching the earth's surface, which would have extremely harmful ecological consequences.

Despite numerous studies, this hypothesis has been neither confirmed nor refuted. The restrictions on the use of fluorocarbons aerosol propellants remain. As far as other applications are concerned, however, the major countries have not made any definitive, clear decisions [56].

Aluminum fluoride and synthetic cryolite are constituents of the fused alumina bath from which aluminum is produced by electrolysis. These compounds act only to lower the melting point of the bath, and the consumption is mostly losses to the atmosphere. Mainly environmental pressure has reduced the consumption, in terms of HF, from 35 kg/t of aluminum in 1960 to less than 25 kg/t of aluminum at the present. Over the long term, this market for hydrofluoric acid may be reduced by the development of aluminum production processes that do not involve any fluorine products, such as an Alcoa process based on the electrolysis of aluminum chloride.

The other applications of hydrofluoric acid are:

1) Pickling of stainless steel sheets. Hydrofluoric acid (aqueous 70 % solution) is used for certain grades of stainless steel at the rate of 10 kg HF/t steel.
2) Manufacture of UF_6. Hydrofluoric acid is used to produce UF_4 from U_3O_8; UF_4 is then converted to UF_6 with fluorine, which is obtained by electrolysis of anhydrous hydrofluoric acid.
3) Catalysis. Alkylation processes converting light olefins and isoparaffins to gasoline can be catalyzed by H_2SO_4 or anhydrous HF [57]. These processes have been developed primarily in the United States. The hydrofluoric acid consumption is ca. 2 kg/m^3 of alkylate.
4) Engraving, glass polishing, polishing of electronic silicon.
5) Manufacture of various fluorides, such as NaF or BF_3.

2. Calcium Fluoride

Fluorspar is a calcium fluoride [*7789-75-5*] ore, whose deposits are distributed over all five continents.

Crystals of fluorite were used in antiquity for the manufacture of ornamental cups and vases. The flux properties, to which fluorspar owes its name, have been known since the sixteenth century. The first mines were started up at the end of the nineteenth century in England, and then in the United States, to meet the requirements of the iron and steel industry (Bessemer and Martin processes). Later on, the electric furnace and, above all, the basic oxygen furnace, caused fluorite consumption to increase further.

In parallel, the chemical industry, using hydrofluoric acid, produced synthetic aluminum fluoride and cryolite for the aluminum industry and the fluorocarbons. Fluorspar is the only fluorine raw material used to manufacture hydrofluoric acid.

The other uses of fluorspar (manufacture of glasses, enamels, and welding fluxes) use only small quantities of ore.

World fluorite production has been at a ceiling level of five million tonnes per year since 1975. Investigations to develop a calcium fluoride (or hydrofluoric acid) manufacturing process starting from the fluosilicic acid (phosphoric acid byproduct) have not yet led to any industrial process.

2.1. Physical and Chemical Properties

The principal physical properties of calcium fluoride, M_r 78.08, are shown in Table 3. Calcium fluoride crystallizes in the face-centered cubic system, which leads to cubic or, less frequently, octahedral crystals. Crystals of pure calcium fluoride are transparent and colorless, while natural fluorite also has a broad variety of colors, ranging from red to blue. Some varieties are fluorescent in ultraviolet radiation.

Calcium fluoride forms eutectics with most metallic oxides, which makes it a useful flux. Fluorite is slightly hydrolyzed by heating in moist air above 1000 °C. It is only slightly soluble in dilute, cold acids but is attacked by hot sulfuric acid to produce hydrofluoric acid.

Table 3. Physical properties of calcium fluoride

Property	Value	Reference
Melting point *mp*	1414 °C	[58]
Heat of fusion ΔH_f	29.7 kJ/mol	[58]
Boiling point *bp*	2513 °C	[59]
Heat of vaporization ΔH_v at the *bp*	335 kJ/mol	[60]
Specific heat C_p at 25 °C	67 J/mol K	[61]
Heat of formation ΔH_f° at 25 °C	−1215 kJ/mol	[62]
Density ϱ	3.181 g/cm^3	[63]
Hardness (Mohs scale)	4	[64]
Solubility in water at 25 °C	1.7 mg/100 g H_2O	[65]
Refractive index n_D^{24}	1.43382	[66]

2.2. Production and Resources

High-purity calcium fluoride, which is used for the manufacture of special glasses and for the growth of single crystals, is obtained by the reaction of hydrofluoric acid and pure calcium carbonate.

Recovery. The locations of the fluorspar deposits are closely associated with continental rifts and lineaments (Table 4). Prospecting may be simple or difficult, depending on whether vegetation is sparse or abundant and on whether the deposit reaches the surface or not. The deposits are present either in the form of veins in vertical fissures, or in the form of horizontal beds. These two types of deposit can be exploited in surface mines or in pits and galleries.

Fluorite is generally associated with other minerals, such as quartz, barite, calcite, galena, siderite, sphalerite, or chalcopyrite (Table 5). In order to be economically workable, a deposit must have an average CaF_2 content above 20 % and, with rare exceptions, the extracted ores must be concentrated to be marketable. Selective separations are obtained in the gravity concentration processes, because calcium fluoride has a density that is generally higher than that of the minerals constituting the gangue. Preconcentration of the crushed ore can be obtained in a cone separator, using an aqueous suspension of finely ground ferrosilicon as a separation medium. The relative density of the suspension is maintained between 2.55 and 2.62 at the top of the cone and between 2.85 and 3.1 at the bottom of the cone. The ore is introduced at the top of the cone, from which the light particles are removed, while the heavier particles, rich in fluorite, drop to the bottom of the cone, where they are recovered. The ferrosilicon is recycled. A product containing 40 % of calcium fluoride is obtained from 20 % ore. Flotation then provides the different commercial grades of fluorite. Most flotation units are of the multistage type and permit the recovery of byproducts.

Table 4. Geologic fluorite deposits [67], [68]

a) *Gangue deposits*	Example
1. in magmatites	Massif Central (France)
in metamorphites	Oberpfalz (Federal Republic of Germany)
2. in clastic sedimentary rocks	Rottleberode/East Harz (German Democratic Republic)
3. in calcareous dolomite sedimentary rocks	Rosiclaire/Illinois (United States)
b) *Stratified deposits*	
1. displacements in limestone and dolomites	La Collada/Asturia (Spain)
2. marine sedimentary deposits in limestone and dolomites	Coahuila/Mexico, Cave-in-Rock Distr./Illinois (United States)
3. vulcanic sedimentary deposits	Pianciano/Latium (Italy)
c) *Deposits of different formation*	
1. occurrences and impregnations in magmatic and different rocks	Vergenoeg (Republic of South Africa)
2. deposits in carbonatites	Okoruso (South West Africa)

Production. Table 6 lists the world production of fluorite, with the contribution of each principal producing country for the years 1975 and 1980. Also listed are the reserves (calculated as finished product) considered to be economically workable in 1978.

World production has reached a ceiling of five million tons per year. Mexico is the leading producer, with over 20 % of world production.

Production at the current level would exhaust the reserves by ca. 2000. However, a shortage of calcium fluoride will be prevented by:

1) Recourse to fluorspar reserves that are not economically workable under present conditions
2) Continuation of prospecting and the discovery of new deposits
3) Production of calcium fluoride and hydrofluoric acid from fluosilicic acid.

Table 5. CaF_2 contents and accompanying minerals of fluorite in different types of deposits [67]

	Contents of CaF_2 (%)	Other minerals	Deposits
Gangue deposits	50 – 80	barite, uranium ores	Nabburg/Wölsendorf/Oberpfalz (Federal Republic of Germany)
	60 – 90	calcite, quartz (barite and sulfite missing)	Zaragoza and Rio Verde District/ San Luis Potosi (Mexico)
	60	barite, quartz	Massif Central etc. (France)
	80 – 95	sulfides (of minor importance)	La Charbonniere near Limoges (France)
	47 – 50	barite, PbS, ZnS, Cu, gravel, quartz	Rosiclaire/Illinois (United States)
	70 – 80	barite, PbS, quartz, etc.	Durham (Great Britain)
	50 – 70	3 – 25 % barite 1 – 2 % PbS	Eyam District/Derbyshire (Great Britain)
	35 – 50	little barite	Burin peninsula/New Foundland (Canada)
	18	40 % barite	Cape Breton Island/Nova Scotia (Canada)
	16 – 50	0.3 % Sn, 0.04 % WO_3, PbS, ZnS, Mo_2S, Cu gravel, etc. partly 0.2 % Be	Lost River/Alaska (United States)
	13 – 14	5 % Pb, 7 % Zn, 0.5 % Cu	San Francisco del Oro/Chihuahua (Mexico)
Stratified deposits	60 – 85	0.3 % Be (bertrandite), 10 – 12 % calcite, 5 – 10 % SiO_2	Aguachile Coahuila (Mexico)
	65 – 70, locally 80 – 90 [1]	calcite, quartz	Pico Etereo District/Coahuila (Mexico)
	20 – 60	PbS, ZnS, barite, marcasite, Cu gravel	Cave-in-Rock District/Illinois (United States)
	65 – 67	15 % Fe_2O_3 (goethite, limonite), 15 % SiO_2, little PbS	Escaro/Pyrenees (France)
	35	10 – 15 % barite, Pb/Zn sulfides, quartz	Morvan, Massif Central (France)
	20 – 40	20 – 25 % barite, 40 – 45 % SiO_2, calcite	La Collada/Asturia (Spain)
	36	40 % barite, 12 % celestite, 6 % SiO_2, little PbS, ZnS, pyrite, calcite	Djebel Zaghouan (Tunisia)
Deposits of different formation	25 – 55	13 % barite, apatite, etc.	Pianciano/Latium (Italy)
	variable Ø 15 – 20	quartz, feldspar	Buffalo, Waterberg District, Bushveld (Republic of South Africa)
	15 – 70	to 60 % Fe_2O_3 (hematite) Ø 55 % Fe_2O_3	Vergenoeg, Farm Kromdraal, Warmbad District (Republic of South Africa)

[1] These contents are unusually high for stratified deposits.

Several processes have been proposed for the manufacture of calcium fluoride from the fluosilicic acid recovered from wet phosphoric acid production, but none has been developed industrially.

The U.S. Bureau of Mines process [71] includes two stages: treatment of the fluosilicic acid solution with ammonia to precipitate filterable silica:

$$H_2SiF_6 + 6\,NH_3 + 2\,H_2O \longrightarrow 6\,NH_4F + SiO_2$$

followed by treatment of the ammonium fluoride solution with milk of lime to precipitate calcium fluoride:

$$6\,NH_4F + 3\,Ca(OH)_2 \longrightarrow 3\,CaF_2 + 6\,NH_3 + 6\,H_2O$$

The ammonia solution is recycled.

Bayer [72] and Kali Chemie [73] treat the fluosilicic acid solution with calcium carbonate suspension to a metastable solution of silica and a precipitate of calcium fluoride, which is separated by filtration and then dried.

2.3. Uses

Table 7 shows the consumption of calcium fluoride in the United States in 1980 by type of application. Ninety-eight percent of the tonnage

Table 6. World production and reserves of fluorite [69], [70] (in thousands of tonnes of commercial spar)

Country	Production, 1000 t		Reserves, 1000 t
	1975	1980	1978
World	*5000*	*4900*	*100 000*
North America			
Mexico	1089	1106	13 000
USA	127	84	6 000
South America			
Argentina	54	30	1 800
Brazil	109	55	2 000
Western Europe			
France	478	318	3 100
Great Britain	235	186	7 800
Spain	335	245	3 500
Italy	231	152	2 400
Federal Republic of Germany	75	87	1 300
Eastern Europe			
former USSR	475	520	5 300
former Czechoslovakia	90	96	800
former metallurGerman Democratic Republic	90	100	800
Africa			
South Africa	202	523	31 000
Tunisia	30	39	1 100
Morocco	47	64	1 300
Kenya	55	93	2 900
Asia			
China	350	400	4 200
Mongolia	302	450	4 000
Thailand	286	306	3 800
North Korea	30	40	

is consumed by two industries: the chemical industry, for the manufacture of hydrofluoric acid, and the iron and steel industry.

Table 7. Use of fluorite in the United States in 1980 (in thousands of tons) [69]

Hydrofluoric acid	587
Steelmaking	367
Basic oxygen furnaces	243
Open hearth furnaces	58
Electric furnaces	64
Iron and steel foundry	10
Glass and fiberglass	10
Welding rod coatings	1.3
Enamel, pottery	0.6
Other uses	2.1
Total	*976*

Chemical Industry. Fluorspar is the starting material for the production of hydrofluoric acid (see Chap. 1).

Iron and Steel. Fluorspar is a flux in the manufacture of steel, lowering the melting point and increasing the fluidity of the slag. Furthermore, when mixed with lime, it reduces the phosphorus and sulfur content of the metal. The use of fluorspar varies from 2 – 5 kg/t metal, depending on the type of furnace.

Fluorspar is used in the manufacture of cast iron at the rate of 6 – 8 kg/t of metal, and also in the manufacture of certain ferroalloys.

Other Uses. Fluorite constitutes 5 – 7 % of the aluminum electrolysis bath. It is used in numerous fluxes and, in particular, for welding rod coatings.

Milky or colored glasses and glass fibers can contain 10 – 20 % of fluorite. Enamels for metallic or ceramic substrates contain between 3 and 10 % of fluorite, which acts as an opacifier.

The addition of 1 % of a mixture of equal parts of gypsum and fluorspar to the raw materials for

cement reduces the clinkering temperature by 50 – 150 °C without lowering the quality of the cement [74].

2.4. Toxicology

Fluorite is not classified as a dangerous material. Nevertheless, every precaution should be taken to prevent accidental contact of calcium fluoride with an acid that would lead to the formation of hydrofluoric acid. Ingestion of fluorite dusts can be dangerous and must be avoided. For inorganic fluorides, NIOSH recommends a limiting concentration, expressed as fluorine, of 2.5 mg/m^3 air [75].

2.5. Specifications

Three commercial grades of fluorite are available:

acid-grade fluorspar (acid spar)
ceramic-grade fluorspar (ceramic spar)
metallurgical-grade fluorspar (met spar)

The specifications for acid spar are as follows:

CaF_2	97 % min
SiO_2	1 % max
S (sulfide)	0.05 % max

Users can also establish a particle size and maximum contents of $CaCO_3$, B, Be, As, and P. For ceramic spar:

CaF_2	85 – 97 %
SiO_2	2.5 – 3 % max
$CaCO_3$	1 – 1.5 % max
Fe_2O_3	0.12 % max
Pb, Zn	traces

The percentage of CaF_2 determines the grade. For met spar:

60 – 85 % CaF_2

Frequently, and particularly in the United States, the content of met spar is expressed in units of *effective CaF_2*. This content is obtained by subtracting from the CaF_2 content a value equal to 2.5 times the SiO_2 content, to take into account the fluorine-consuming reaction:

$$2\,CaF_2 + SiO_2 \longrightarrow SiF_4 + 2\,CaO.$$

Expressed as effective CaF_2, the specification is generally 70 % minimum. The content of free sulfur or sulfide is limited to 0.3 %. Maxima may also be established for lead (0.25 %) and for As, Ba, and Zn, at the level of 100 ppm. The particle size is preferably between 5 and 80 mm.

The presence of fines allows significant losses by entrainment. To avoid this, briquettes are made by agglomeration of fines with molasses acting as a binder. Numerous additives (MnO_2, Al_2O_3, $Ca(OH)_2$, feldspar, ilmenite, etc.) can be incorporated at the user's request. The shape and size of the briquettes differ: the current peach-pit type have a length of 2.5 cm. In the United States, steelmaking used 110 000 t of briquettes in 1980.

2.6. Chemical Analysis

The analytical methods for fluorite have been standardized or are in the process of standardization:

H_2O	ISO 4282	Drying at 105 °C
Fluorine	Draft ISO 5439	Willard – Winters method
SiO_2	Draft ISO 5438	Colorimetry of the reduced blue silicomolybdate complex
$BaSO_4$	Draft ISO 5437	Gravimetry
Carbonates	Draft ISO 4283	Acidimetry after release and absorption of CO_2 in sodium hydroxide solution
Sulfides	Draft ISO 4284	H_2S converted to cadmium sulfide and titrated iodometrically
Phosphorus	Draft ISO 6676	Colorimetry of the yellow phosphato-vanadomolybdate complex

Calcium can also be determined by complexometry with EDTA, boron by colorimetry with carminic acid, and arsenic by colorimetry of the red complex obtained by the absorption of AsH_3 in a solution of silver diethyldithiocarbamate in pyridine.

3. Aluminum Fluoride and Fluoroaluminates

Aluminum fluoride [*7784-18-1*] and cryolite [*15096-52-3*] are the two aluminum – fluorine compounds of industrial importance. Their main application is in baths for electrolytic aluminum production, accounting for over 95 % of their use.

For a long time, the fluorine raw material was hydrofluoric acid, but its importance has decreased considerably since 1975. Between 1975 and 1981, 20 – 25 % of the production of aluminum fluorides was based on fluosilicic acid, an effluent in the manufacture of phosphoric acid and hydrofluoric acid [76, no. 7391002 H]. This proportion should increase in the future, for the following reasons:

1) In 1985, only 40 % of the phosphoric acid units were equipped for fluosilicic acid recovery and, in all cases, 80 % of this acid was destroyed [77].
2) Fluosilicic acid recovery is easy, and the cost of the fluorine is well below that from hydrofluoric acid.
3) An estimate made in 1978 showed that the confirmed reserves of fluorine in phosphates were five times greater than those of fluorine in fluorite [78], [79].
4) The recovery of fluorine effluents of the aluminum industry will intensify under the pressure of the strict regulations for fluorine-containing wastes. Efforts between 1960 and 1981 to recycle fluorine effluents back to the electrolysis cells, reduced fluorine consumption from 35 to 24 kg/t aluminum [80]. Progress remains to be made in fluorine recovery both in phosphoric acid plants and aluminum plants.

3.1. Aluminum Fluoride

3.1.1. Properties

The only important aluminum fluoride is the trifluoride, M_r 83.98: white solid consisting of a trigonal crystal and usually forming rhombohedrous (88°49′), space group $R32D_3^7$ with 2 AlF_3 per unit cell and dimensions of $a_0 = 0.504$ nm and $\alpha = 58°31'$ [81]. Each Al atom is coordinated by six F atoms, with three Al – F distances of 0.170 nm and three of 0.189 nm. The F – F spacing is 0.253 nm. The lattice vacancies can be occupied by water molecules.

At atmospheric pressure, aluminum trifluoride does not melt but rather sublimes. The main characteristics of aluminum fluoride are shown in Table 8.

Table 8. Physical properties of aluminum fluoride

Property	Value	Reference
Sublimation point	1272 – 1278 °C	[82]
Vapor pressure P at 1000 °C	910 – 920 Pa	
Density ϱ	3.10 g/cm^3	
Heat of formation at 298 K ΔH_f°	−1489 kJ/mol	[83], [84]
Free energy of formation at 298 K ΔG_f°	−1426 to −1431 kJ/mol	[84]
Refractive index n_D	1.376	
Entropy S_{298}°	66.2 – 66.5 J/mol K	[82]

Solubility in water: the trihydrate prepared by precipitation below 25 °C has a solubility of 4.1 g/L (expressed as AlF_3). On the other hand, concentrated solutions (ca. 200 g/L) prepared and kept for 10 h at 90 – 100 °C and stirred in the presence of the solid precipitate the β form. The solution above the β form maintains a concentration close to 16 g/L, regardless of temperature.

The remarkable aspect is the stability of the supersaturated concentrated solution. When cooled abruptly from 100 to 20 °C, in the absence of solid, it can be kept for several hours at concentrations of close to 200 g/L. This stability varies with the contents of HF and AlF_3 [85].

In the presence of HF, the solubility of AlF_3 in water at 25 °C goes through a maximum of 17 wt % for an HF concentration close to 39 %, probably related to the presence of $AlF_3 \cdot 3\,HF$ (AlF_6^{3-} ion) [86]. The solubility in organic solvents is very low.

Aluminum fluoride, when heated above 250 °C in the presence of steam, undergoes partial hydrolysis:

$$2\,AlF_3 + 3\,H_2O \longrightarrow 6\,HF + Al_2O_3 \qquad (1)$$

For the dehydration of $AlF_3 \cdot 3\,H_2O$, this hydrolysis must be avoided, and the drying and calcination conditions must be appropriate.

3.1.2. Manufacture

Raw materials:

1) Partially dehydrated aluminum hydroxide, and to a lesser extent uncalcined alumina
2) Gaseous hydrofluoric acid, 100 % or diluted with air, and concentrated solutions containing 30 – 70 % of HF
3) Fluosilicic acid (20 – 30 % H_2SiF_6)

3.1.2.1. Hydrofluoric Acid Processes

Aqueous Hydrofluoric Acid Process [87]. Installations using hydrofluoric acid solution are increasingly rare. The process consists of reacting alumina trihydrate, $Al_2O_3 \cdot 3\,H_2O$, continuously or discontinuously with hydrofluoric acid solution at 80 °C until the alumina dissolves completely. The stirred supersaturated solution of AlF_3, with or without seeding, deposits $AlF_3 \cdot 3\,H_2O$, which is separated by filtration. The solid is dried and calcined in one or two stages to furnish a product containing 90 – 98 % AlF_3. The mother liquor, with a content of ca. 18 g/L of AlF_3, can be used for cryolite production.

Gaseous Hydrofluoric Acid Process. The processes using this method operate with a two- or three-stage fluidized bed [88] with gaseous hydrofluoric acid (possibly diluted) as a carrier. The alumina trihydrate reacts countercurrently, and is dehydrated and fluorinated at temperatures of 500 – 600 °C :

$$Al_2O_3 \cdot 3\,H_2O(s) + 6\,HF(g) \longrightarrow 2\,AlF_3(s) + 6\,H_2O(g)$$
$$\Delta H^0_{298} = -126\,\text{kJ/mol}\,AlF_3$$
$$Al_2O_3 \cdot H_2O(s) + 6\,HF(g) \longrightarrow 2\,AlF_3(s) + 4\,H_2O(g)$$
$$\Delta H^0_{298} = -182.7\,\text{kJ/mol}\,AlF_3$$
$$Al_2O_3(s) + 6\,HF(g) \longrightarrow 2\,AlF_3(s) + 3\,H_2O(g)$$
$$\Delta H^0_{298} = -212.1\,\text{kJ/mol}\,AlF_3$$

(The heats of reaction were obtained by using for the heat of formation of $AlF_3(s)$ the value of −1488.9 kJ reported by Kubachewski [83], which is more reliable than the values found in other references, which range from −1400 to −1510 kJ.)

The installations differ in the way the gaseous hydrofluoric acid is used. Hydrofluoric acid that is practically pure, except for small traces of SO_2 and SiF_4, is obtained by vaporization of anhydrous HF. This technique is used by numerous companies: Allied Chemical, Bayer, Kaiser, Montedison, Stauffer, and Ugine Kuhlmann before 1970.

Another method takes the HF directly at the exit of the rotary production furnace and heats it with an internal burner to furnish a mixture with one-fifth the usual HF content (approx. 200 g/m^3), which reacts directly with the fluidized alumina. This technique is more economical, because it avoids the anhydrous HF stage, but the fluidized bed reactor must be bulkier. It is used by Pechiney, the world's largest producer, with a capacity of 90×10^3 t in its plant at Salindres (France). The effluent gas, containing SiF_4 and a small amount of HF, is washed with water to produce a solution of fluosilicic acid and HF. These techniques produce a grade containing only 89 – 91 % of AlF_3. Numerous technological improvements have been made over the years [89].

3.1.2.2. Fluosilicic Acid Processes

Direct Process. For more than 25 years, aluminum fluoride has been produced throughout the world from fluosilicic acid, which originates mainly from the recovery of fluorine in the phosphate industry and to a lesser extent from the manufacture of HF from fluorite and of AlF_3 from HF. The first installation was started up in Linz, Austria, by O.S.W. (Oesterreichische Stickstoffwerke) in 1962 [90], followed by plants in Sweden, Japan, USA, former states of USSR, India, Poland, Spain, Romania, France, and the German Democratic Republic. Numerous patents have been filed, particularly by O.S.W.

The aluminum fluoride produced must meet the purity standards

$SiO_2 < 0.2\,\%$
$P_2O_5 < 0.025\,\%$

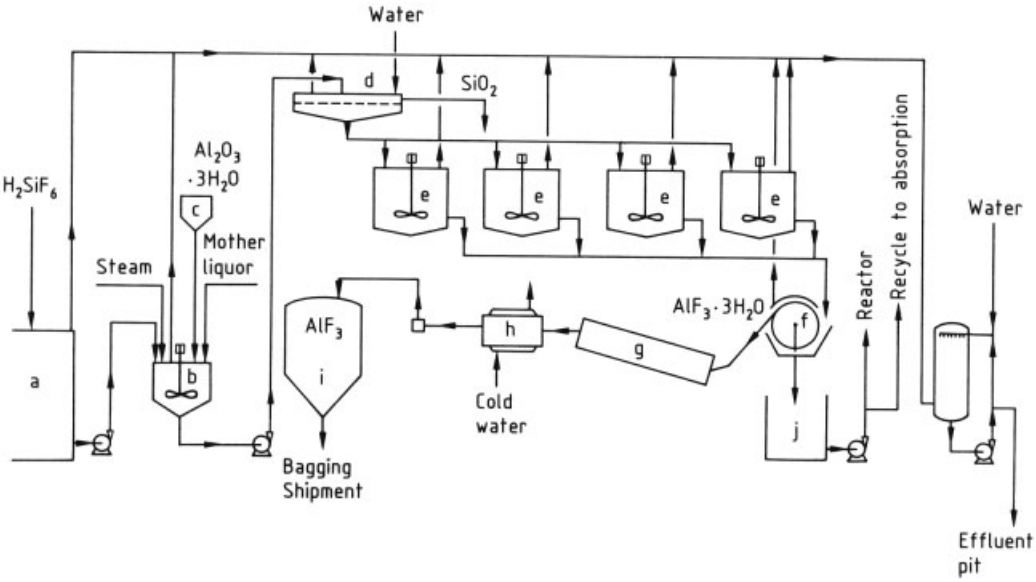

Figure 5. Simplified process flow sheet for the production of AlF_3 from H_2SiF_6 and $Al_2O_3 \cdot 3\,H_2O$
a) H_2SiF_6 storage tank; b) Reactor; c) Alumina hopper; d) Silica filter; e) Crystallizers; f) Trihydrate filter; g) Drier; h) Cooler; i) Finished product storage hopper; j) Weak mother liquor storage tank

required for aluminum electrolysis. The silica standard is easy to achieve, but the P_2O_5 standard, equivalent to a P_2O_5 : F ratio of <0.0004, is more of a challenge. The principal effort to remove P_2O_5 must be made during recovery of fluosilicic acid.

The various phases of the process for the manufacture of AlF_3 from fluosilicic acid are shown in Figure 5 [91].

Alumina trihydrate is dissolved in a batch process or less frequently in a continuous process in 15 wt % fluosilicic acid preheated to 60 °C:

$$H_2SiF_6 + Al_2O_3 \cdot 3\,H_2O \longrightarrow 2\,AlF_3 + SiO_2 + 4\,H_2O$$

The heat evolved brings the temperature of the mixture to 90 – 100 °C, and reaction is complete within 20 minutes. The precipitated silica must be easily filterable. The following parameters therefore are varied during the reaction: temperature, concentration and excess of reagents, agitation, pH. All producers have their own precise conditions. The silica is filtered off and washed on a belt filter, bucket filter, or even a centrifugal filter. The supersaturated AlF_3 solution is placed in stirred tanks maintained at 80 – 90 °C for 5 – 8 h, with or without seeding. After the crystallization of β-$AlF_3 \cdot 3\,H_2O$, this is separated on a rotary or belt filter, with or without washing of the cake.

The mother liquors contain ca. 18 g/L of AlF_3 and can be reused: (1) for dilution of the fluosilicic acid, (2) as the first washing of the silica on the filter, (3) in washing the gases of phosphoric acid and superphosphate units, and (4) in the manufacture of cryolite. The limit of recovery of these mother liquors is fixed by their P_2O_5 content, which must not exceed 100 ppm.

The moist trihydrate is dried in two stages, a method known since 1946 [92]. Since then numerous variations have been proposed, and many patent applications have been filed. In the first stage the temperature is limited to 250 °C to prevent hydrolysis (Eq. 1) and to produce a composition close to $AlF_3 \cdot 0.5\,H_2O$. This is subjected, in a second stage, to heating to 500 – 600 °C, either slowly in an indirectly heated rotary furnace or more rapidly in a fluidized bed [93]. In all cases, a calcined product containing 97 – 98 % AlF_3 is obtained, which must be protected against moist air [94].

One variation would be of interest: crystallization of a fluoride with a low degree of hydra-

tion (0.15 – 1 H_2O) with crystallization at high temperature and pressure [95].

Indirect Process. The intermediate is ammonium cryolite, $AlF_3 \cdot 3\,NH_4F$, which is manufactured by reacting aqueous ammonium fluoride

$$H_2SiF_6 + 6\,NH_4OH \longrightarrow 6\,NH_4F + SiO_2 + 4\,H_2O$$

with partially dehydrated alumina [96]:

$$12\,NH_4F + Al_2O_3 \cdot x\,H_2O \longrightarrow 2\,AlF_3 \cdot 3\,NH_4F + 6\,NH_3 + (x+3)\,H_2O \quad (2)$$

or with a solution of AlF_3:

$$3\,NH_4F + AlF_3 \longrightarrow AlF_3 \cdot 3\,NH_4F \quad (3)$$

This last reaction can be carried out in a fluidized bed [97], which produces calibrated crystals of ammonium cryolite, thus facilitating further treatment.

The manufacture of AlF_3 from ammonium cryolite can be carried out by two methods:

1) Thermal decomposition at 400 – 500 °C, liberating NH_4F:

$$AlF_3 \cdot 3\,NH_4F \longrightarrow AlF_3 + 3\,NH_4F$$

which is recycled. The difficulty is collecting sublimed NH_4F by washing.

2) Reaction with partially dehydrated alumina, liberating NH_3, which is easier to recycle:

$$2\,AlF_3 \cdot 3\,NH_4F + Al_2O_3 \cdot x\,H_2O \longrightarrow 4\,AlF_3 + 6\,NH_3 + (x+3)\,H_2O$$

These two processes are not as simple as the direct process because of the recycling of NH_4F or NH_3. The one advantage, except pointed out by the Tennessee Valley Authority (TVA) and Mekog Albatros [98]: P_2O_5 is removed during precipitation of the ammonium cryolite.

3.1.2.3. Recovery of Fluorine During Aluminum Production

The fluorine in the gases emitted by the electrolysis cells (hydrolysis of the bath or sublimation in the form AlF_3 or atmolite, $NaAlF_4$) must be recovered for economic and environmental reasons. At the present time, virtually the only recovery process being developed further involves the dry method.

Dry Method. In the dry method, classical Sandy Coarse alumina, with a BET surface area between 45 and 80 m^2/g, is used to recover the fluorine in new automated cells, cf. [99]. This alumina is injected into the circuit for collection of the cap gases [100], and 99.99 % of the hydrogen fluoride is chemisorbed within a few seconds. All particles (alumina and solid dusts) are retained in a bag or sack filter. Only neutral gases can pass through the alumina cake fixed on the bags. The alumina fluorinated in this way, which also contains some impurities, is recovered and added to the cell.

This system is presently employed in all modern cells using the Aluminum Pechiney technology: St. Jean de Maurienne, Tomago, Becancour, Karmoy, and others. Less than 1 kg HF/t of aluminum is allowed to escape.

Wet Method. In the wet method, the effluent gases of old aluminum cells are washed with ca. 20-g/L AlF_3 solution, absorbing the dilute hydrofluoric acid, to bring the solution to approximately 5 g/L of HF in one or two stages. Neutralization with alumina produces a solution containing approximately 27 g/L of AlF_3, which precipitates 11.5 g/L of $AlF_3 \cdot 3\,H_2O$. The mother liquor, containing approximately 20 g/L of AlF_3, is returned to the gas washing stage [101].

Although this acid washing technique does not retain the SO_2, its low productivity, 11.5 g of $AlF_3 \cdot 3\,H_2O$ per liter in circulation, is a disadvantage.

3.2. Fluoroaluminates

The fluoroaluminates constitute a class of products containing aluminum and fluorine that can be written in the form

$$x\,AlF_3 \cdot y\,MF$$

where $1 \leq x \leq 3$ and $1 \leq y \leq 5$; M represents an alkali or alkaline-earth metal.

Cryolite, $AlF_3 \cdot 3\,NaF$, and chiolite, $3\,AlF_3 \cdot 5\,NaF$, are the two principal minerals. However, the quantities mined do not exceed 50×10^3 t of cryolite per year.

In addition to sodium, other cations, such as Ca, Mg, Li, K, Sr, NH_4, and Li/Na combinations are known in nature. Compounds containing oxygen and OH groups in addition to Al and F also exist.

The two most important fluoroaluminates obtained by synthesis are sodium cryolite, or simply cryolite, and the intermediate, ammonium cryolite, $AlF_3 \cdot 3\,NH_4F$.

3.2.1. Cryolite, $AlF_3 \cdot 3\,NaF$

3.2.1.1. Properties

At room temperature, crystalline cryolite, M_r 209.94, is monoclinic, with a slight deviation from orthorhombic symmetry ($R = 90°\ 11'$). The space group is P21/m, with $a = 0.546$, $b = 0.561$, and $c = 0.780$ nm [102]. The physical properties are shown in Table 9.

Table 9. Physical properties of cryolite

Property	Value	Reference
Melting point *mp*	1009 – 1012 °C	
Vapor pressure *P*		
at 1000 °C	253 Pa	[103]
Density ϱ	2.97 g/cm^3	
Heat of formation ΔH_f°	−3297 kJ/mol	
Heat of fusion ΔH_f		
at 1012 °C	225 kJ/mol	
Refractive index n_D	1.339	
Entropy S_{298}°	238 J/mol K	[103], [104]
Electrical conductivity *K*		
Liquid at 1012 °C	2.82 Ω^{-1}cm^{-1}	
Solid at 400 °C	$4 \times 10^{-6}\ \Omega^{-1}$cm^{-1}	

Heating cryolite causes crystalline transformations, accompanied by changes in density and electrical conductivity [104].

Cryolite is important because it is the main constituent of the baths for aluminum production; it lowers the melting point below 1000 °C, while dissolving AlF_3, NaF, CaF_2, LiF, and the oxides Al_2O_3, MgO, TiO_2, etc. In the fused bath (65 wt % $AlF_3 \cdot 3\,NaF$) the important equilibrium expression is [105]

$$AlF_3 \cdot 3\,NaF \rightleftharpoons 2\,NaF + AlF_3 \cdot NaF$$

The properties of this bath have been often studied, especially density, viscosity, solubilities, conductivity [104], [106]. Conductivity is provided primarily by the Na^+ ions.

Cryolite has a solubility in water of 0.042 g at 25 °C and 0.135 g at 100 °C per 100 g of solution; the solubility is lower in the presence of NaF and is higher in the presence of acid [107], base, or aluminum, or iron salts [108]. If air is excluded, the melting loss at 400 °C is less than 0.5 % for natural cryolite and between 2 and 3 % for synthetic cryolites.

3.2.1.2. Manufacture

Although the specific consumption of cryolite is lower than that of aluminum fluoride in electrolytic aluminum production, and the ratio is continuously changing to AlF_3, the manufacture of cryolite continues to be the subject of new studies.

The processes developed during 1931 – 1951 have lost appeal for two reasons [109]:

1) Complexity, execution requiring several stages
2) Raw materials that are scarce or out of favor (high-silica fluorite, CO_2–H_2SO_4–$NaHCO_3$, $AlCl_3$, etc.).

Modern processes use raw materials such as HF, H_2SiF_6, $Al_2O_3 \cdot 3\,H_2O$, $Al_2O_3 \cdot x\,NaOH$, NaOH, NaCl, and Na_2SO_4. Usually, manufacture is placed close to the sources of fluorine and aluminum, to reduce costs further.

The quality of the final product has been significantly improved:

1) NaF/AlF_3 ratios close to 3, and adjustable around this value
2) high bulk densities, close to 1.5 g/cm^3
3) spherical particles, average diameter adjustable from 50 – 200 µm, which are easy to filter, wash, dry, and calcine [110], [111].

These techniques have been placed on an industrial scale by Ugine Kuhlmann.

Hydrofluoric Acid Processes. The main hydrofluoric acid processes can be summarized by the equation

$$12\,HF + Al_2O_3 \cdot 3\,H_2O \longrightarrow 2\,AlF_3 \cdot 3\,HF + 6\,H_2O$$

followed by one of the three equations

$$2\,AlF_3 \cdot 3\,HF + 6\,NaOH \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 6\,H_2O$$

$$2\,AlF_3 \cdot 3\,HF + 3\,Na_2CO_3 \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 3\,H_2O + 3\,CO_2$$

$$2\,AlF_3 \cdot 3\,HF + 6\,NaCl \longrightarrow 2\,AlF_3 \cdot 3\,NaFO + 6\,HCl$$

These operations are generally carried out by a batch process: after the alumina is dissolved in 40–60 % HF solution in one hour at 70 °C, cryolite is precipitated by addition of NaOH, Na_2CO_3, or NaCl. After the slurry is filtered, the solids are calcined at 600 °C in a rotary or platform furnace. These operations can also be carried out continuously in a fluidized bed to produce granular cryolite [111]. Ugine Kuhlmann put into industrial use between 1965 and 1970 another continuous, fluidized bed process (capacity 8000 t/a) using a solution of sodium aluminate from Bayer alumina production

$$12\,HF + Al_2O_3 \cdot 6\,NaOH \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 9\,H_2O \quad (4)$$

The aluminate solution can be taken from the concentration stage following the filtration of alumina and contains approximately 150 g/L of Al_2O_3 and 360 g/L of NaOH (→ Aluminum Oxide). After cooling, this solution is stable and can be stored and transported without solids formation. Its cost is very attractive. This process [111] can produce calibrated cryolite with an average diameter of approximately 80 µm at the exit of a fluidized bed supplied with aluminate and 40–70 wt % hydrofluoric acid. Filtration and drying are facilitated. No dust is formed.

Fluosilicic Acid Processes [112], [113]. Fluosilicic acid, an effluent from HF and AlF_3 manufacture, and above all, from phosphate and superphosphate plants, must meet very rigorous requirements with respect to its P_2O_5 content: the cryolite produced must have a P_2O_5:F ratio $\leq$0.0004 [114]. The standard for SiO_2 is less rigorous (<0.2 %) and is easier to attain if good conditions of precipitation and filtration of the silica are provided in the course of the neutralization of the H_2SiF_6 [115].

Three processes can be distinguished: (1) NH_4F intermediate, (2) AlF_3–NaF intermediate, and (3) Na_2SiF_6 intermediate.

NH_4F Intermediate . A solution is obtained after neutralization of the fluosilicic acid with aqueous ammonia and separation of the silica

$$H_2SiF_6 + 6\,NH_4OH \longrightarrow 6\,NH_4F + SiO_2 + 4\,H_2O$$

The intermediate can be reacted with sodium aluminate:

$$12\,NH_4F + Al_2O_3 \cdot 6\,NaOH + 3\,H_2O \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 12\,NH_4OH$$

This process was used as early as 1940 by I.G. Farbenindustrie at Oppau and is presently in use in two plants in former Yugoslavia [112]. The effluent ammonia solution is recycled. In a fluidized bed on a pilot-plant scale, Ugine Kuhlmann was able to verify this technique to manufacture granular cryolite [111].

The intermediate can also be reacted with alumina or with aluminum fluoride [Eqs. (2) and (3)] to produce ammonium cryolite, $AlF_3 \cdot 3\,NH_4F$, which can be transformed to sodium cryolite by reaction with sodium hydroxide:

$$AlF_3 \cdot 3\,NH_4F + 3\,NaOH \longrightarrow AlF_3 \cdot 3\,NaF + 3\,NH_4OH$$

with recycling of the ammonia solution to NH_4F manufacture. A sodium salt is more economical than sodium hydroxide:

$$AlF_3 \cdot 3\,NH_4F + 3\,NaCl \longrightarrow AlF_3 \cdot 3\,NaF + 3\,NH_4Cl$$

An ammonium cryolite stage makes is possible to remove the P_2O_5 in the NH_4F solution [98].

Manufacture from NaF and AlF_3 [116]. Two solutions are prepared from fluosilicic acid: (a) a solution of 150 g/L of AlF_3 and (b) a solution of 40 g/L of NaF obtained by neutralization of fluosilicic acid with NaOH or Na_2CO_3 and separation of the silica:

$$H_2SiF_6 + 6\,NaOH \longrightarrow 6\,NaF + SiO_2 + 4\,H_2O$$

The two solutions are mixed to obtain cryolite:

$$AlF_3 + 3\,NaF \longrightarrow AlF_3 \cdot 3\,NaF$$

The AlF_3 solution can also be the mother liquor from the filtration of $AlF_3 \cdot 3\,H_2O$ production, a means of effectively using these mother liquors, which contain 18 g/L of AlF_3 and 7 % of the fluorine yield of AlF_3 manufacture.

Na_2SiF_6 Intermediate [117]. By starting from impure fluosilicic acid, a sodium fluosilicate with a low P_2O_5 content is obtained; furthermore, Na_2SiF_6 has a higher fluorine content (61 % F) than 30 % fluosilicic acid (24 % F), which makes transport less expensive when cryolite cannot be produced where H_2SiF_6 is recovered.

A number of processes are used for the manufacture of cryolite from Na_2SiF_6 in the various countries–USA, Japan, and India. There are two principal techniques: hydrolysis with Na_2CO_3 and hydrolysis with ammonia.

The process described by Kaiser Aluminum [118] has been known since 1960. A sodium fluoride solution is prepared by treating sodium fluosilicate with Na_2CO_3 and separating silica:

$$Na_2SiF_6 + 4\,Na_2CO_3 + 2\,H_2O \longrightarrow 6\,NaF + SiO_2 + 4\,NaHCO_3$$

The resulting solution is reacted with sodium aluminate to precipitate cryolite:

$$12\,NaF + Al_2O_3 \cdot 2\,NaOH + 8\,NaHCO_3 \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 8\,Na_2CO_3 + 5\,H_2O \qquad (5)$$

The sodium carbonate is recycled.

The reaction of sodium fluorosilicate with aqueous ammonia [119], after separation of the silica, produces a solution of sodium and ammonium fluorides:

$$Na_2SiF_6 + 4\,NH_4OH \longrightarrow 2\,NaF + 4\,NH_4F + 2\,H_2O + SiO_2$$

To obtain cryolite, the two fluorides are reacted either with a solution containing an aluminum salt and a sodium salt,

$$4\,NaF + 8\,NH_4F + Al_2(SO_4)_3 + Na_2SO_4 \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 4(NH_4)_2SO_4$$

or with an aluminate solution

$$2\,NaF + 4\,NH_4F + NaAlO_2 + 4\,H_2O \longrightarrow AlF_3 \cdot 3\,NaF + 4\,NH_4OH + 2\,H_2O$$

In the second case, the ammonia solution is recycled.

In a process that has not gone beyond the pilot-plant stage, a mixture of solids Na_2SiF_6 and $AlF_3 \cdot 3\,H_2O$ is heated to 800 °C and to limit hydrolysis of AlF_3 and SiF_4 the mixture is impregnated with aqueous hydrofluoric acid:

$$3\,Na_2SiF_6 + 2\,AlF_3 \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 3\,SiF_4 \uparrow$$

The major disadvantage is the need to convert SiF_4 to H_2SiF_6, which represents a recycling of four of the six fluorine atoms contained in the starting Na_2SiF_6 [120].

Recovery of Fluorine During Aluminum Production. *Washing of Emitted Gases.* After separation of the soots and metal oxides on a bag or electrostatic filter, the gases are washed either with water or an alkaline solution (NaOH or Na_2CO_3). In the first case, a dilute solution of hydrofluoric acid (6 – 10 g/L) is obtained. When it is treated with sodium aluminate solution [Eq. (4)] and the resulting solution crystallized in a fluidized bed, a sandy cryolite 100 – 130 µm in diameter is obtained [111]. This technique has been in use since 1960 in the Société Péchiney aluminum plant at Lannemezan (France).

In the second case, the solution obtained contains a mixture of NaF and $NaHCO_3$, which is treated according to Equation (5) with an aluminate solution to precipitate cryolite. The sodium carbonate is recycled.

Leaching of Cathode Crusts. During repair of the electrolysis cells, crusts containing fluorine and essentially all of the sodium introduced both by the alumina and the cryolite are removed. Their composition is variable: sodium – fluoride – aluminate. Treatment for recovery is not simple and rarely profitable: after crushing and grinding of the crusts, it consists of extraction with water, aqueous sodium hydroxide, and finally hydrofluoric acid, producing solutions of sodium aluminate, sodium fluorides, and dilute solutions of AlF_3 and HF [121]. By reacting these solutions alone or in the presence of hydrofluoric acid, or of alkaline gas-washing solutions, cryolite is precipitated:

$$5\,NaF + Al_2O_3 \cdot 2\,NaOH + (4\,NaF + 8\,NaHCO_3) \longrightarrow 2\,AlF_3 \cdot 3\,NaF + 8\,Na_2CO_3 + 5\,H_2O$$

$$2\,NaF + Al_2O_3 + 10\,NaOH + 2\,AlF_3 + 16\,HF \longrightarrow 4\,AlF_3 \cdot 3\,NaF + 13\,H_2O$$

The carbonate can be recycled to the leaching stage.

3.2.2. Ammonium Cryolite

Ammonium cryolite, $AlF_3 \cdot 3\,NH_4F$, is an intermediate in the production of AlF_3 (Section 3.2) and cryolite (see page 21) from H_2SiF_6.

Another role of ammonium cryolite as an intermediate, in this case in the production of anhydrous HF from H_2SiF_6, may be cited. Ugine Kuhlmann has studied all of the reactions on a prepilot stage. The process can be summarized in four equations:

$$H_2SiF_6 + 6\,NH_3 + 2\,H_2O \longrightarrow SiO_2 + 6\,NH_4F$$
$$2\,NH_4F \longrightarrow NH_4F \cdot HF + NH_3$$
$$2\,NH_4F \cdot HF + AlF_3 \cdot NH_4F \longrightarrow AlF_3 \cdot 3\,NH_4F + 2\,HF\uparrow$$
$$AlF_3 \cdot 3\,NH_4F \longrightarrow AlF_3 \cdot NH_4F + 2\,NH_4F$$

The NH_4F produced in the last reaction is recycled in the second equation. The net equation is then:

$$H_2SiF_6 + 2\,H_2O \longrightarrow SiO_2 + 6\,HF\uparrow$$

3.3. Environment Protection

Aluminum fluoride and cryolite are produced from two raw materials capable of presenting storage and handling problems: hydrofluoric and fluosilicic acids. In the manufacture of the aluminum fluorides, HF and SiF_4 may be released as a result of hydrolysis occurring during drying and calcination at 500 – 600 °C. The quantities of HF may reach 3 – 10 % of the fluorine contained.

The environmental problem is solved by washing the off-gases with water. The fluorine is recovered, or the wash water is neutralized with lime.

3.4. Specifications

The principal specifications for aluminum fluoride are as follows:

$AlF_3 \geq 90\,\%$	(quality of the product obtained from $H_2SiF_6 \geq 97\,\%$)
$Al_2O_3 \leq 10\,\%$	(quality of the product obtained from $H_2SiF_6 < 3\,\%$)
$SiO_2 < 0.2\,\%$	
$Fe_2O_3 < 0.1\,\%$	
$P_2O_5 < 0.025\,\%$	
Average diameter 40 – 60 µm.	

The principal specifications for cryolite are as follows:

$AlF_3 \cdot 3\,NaF$	$\geq$95 %
F	50 – 53.5 %
Al	13 – 15 %
Free Al_2O_3	3 – 5 %
SiO_2	$\leq$0.2 %
Fe_2O_3	$\leq$0.1 %
P_2O_5	$\leq$0.02 %
NaF : AlF_3	2.8 – 3.1

Particle size distribution: average diameter 50 – 120 µm for the calibrated particles obtained by fluidized bed crystallization, or fine powder between one and ten micrometers in size.

3.5. Chemical Analysis

The 1984 ISO catalog provides the following references to analyses of the principal constituents of AlF_3 and $AlF_3 \cdot 3\,NaF$.

	AlF_3	$AlF_3 \cdot 3\,NaF$
F	2362 – 1972	1693 – 1976
Al	2367 – 1972	2367 – 2972
		2830 – 2973
Na		Oliver and Winter, *Anal. Chim. Acta* **51** (1970) 409
Fe	2368 – 1972	
SiO_2	2369 – 1972	1620 – 1976
H_2O	3392 – 1976	
	3393 – 1976	
P_2O_5	5930 – 1979	
	6374 – 1981	

3.6. Uses

Aluminum Fluoride. Almost all of the aluminum fluoride (over 95 %) is used in electrolysis cells for the production of aluminum (15 – 20 kg per tonne of aluminum) and in cells for electrolytic aluminum refining. In the second

case, the high-purity aluminum fluoride (>97 % AlF_3) obtained by the H_2SiF_6 method is preferred.

The specific consumption of AlF_3 is increasing at the expense of that of cryolite, for the following reasons:

1) The alumina introduced into the electrolysis cells contains sodium (equivalent to 5 000 ppm of Na_2O, on the average), which is converted to cryolite in the fused bath in the presence of AlF_3.
2) The gases emitted by the cell contain sodium tetrafluoroaluminate, $AlF_3 \cdot NaF$, containing a larger quantity of AlF_3 than cryolite, $AlF_3 \cdot 3\,NaF$, requiring a make-up of AlF_3.
3) Use of a more acid bath to increase the reactivity.

The other uses of aluminum fluoride are considerably less important [122]:

glass for optical purposes, lasers, and electrical conductors
ceramics
protector of baths for the refining of light metals
welding flux
constituent of catalysts for the preparation of certain halogenated hydrocarbons

Cryolite. Cryolite is the principal constituent of the electrolysis bath for aluminum production. Nevertheless, its specific consumption (4–10 kg/t aluminum) is relatively low and is decreasing, for the reasons indicated above. However, it plays a very important role, because of its ability to dissolve fluorides and oxides. In this way, the temperature, conductivity, viscosity, and density of the bath can be affected. This property is especially useful in refining baths, which must have a high density to facilitate separation of the fused aluminum. High density is attained by dissolution of barium fluoride in the cryolite melt.

The other uses for cryolite are practically the same as those of aluminum fluoride. In addition, cryolite is used as a filler for resins and in grinding wheels.

3.7. Economic Aspects

The principal use of aluminum fluoride and cryolite is in electrolytic aluminum production. The forecasts of aluminum production made in 1975 turned out to be far from accurate: for example, production in the USA in 1983 was below 4×10^6 t [123, nos. 7 021 001 G and T], as against a forecast of close to 10×10^6 t. The same is true for Japan: the production dropped from 994×10^3 t in 1981 to 446×10^3 t in 1984.

This unpredicted decrease in aluminum production explains the significant variations in the production and consumption of aluminum fluoride and cryolite recorded in the principal countries between 1975 and 1984 and, above all, the significant differences found between capacity and production. When all aluminum potlines have been equipped with the dry method's process (Section 3.1.2.3), the consumption of aluminum – fluorine compounds will decrease considerably (Tables 10 and 11).

3.8. Toxicology [126]

The low solubility of *aluminum fluoride* in water limits the toxicity; however, the following data have been noted:

Minimum oral lethal dose for guinea pigs: 600 mg/kg
Threshold limit value: 88 mg/m^3
Contact for 24 h of a dose of 500 mg with the eye of an adult rabbit produces erythemas, edemas, etc.
Subcutaneous injection to frogs: LD_{50}Lo: 1 680 mg/kg

The only danger presented otherwise by AlF_3 is caused by its hydrolysis during heating >300 °C in the presence of water vapor: in this case, it is the HF that is toxic.

The low solubility of *cryolite* in water makes it relatively nontoxic; however, on ingestion, the gastric acidity increases this solubility. LD_{50}: 200 mg/kg.

4. Other Metal Fluorides

Ammonium fluorides are included in this chapter because their behavior is very similar to that of the alkali-metal fluorides.

Table 10. Economic data for aluminum fluoride

Country or region	Year	Capacity 10^3 t/a	Production 10^3 t/a	Consumption	Prinicpal companies
Western Europe	1980		52 [123, no. 7022000Z]		Pechiney (France)
	1984		100 [123, nos. 7022000D and T]	80 [123, no. 7022000D]	and
	1985	165 (est.) [123, no. 7022000Y]			ODDA (Norway)
USA	1980		135 [123, no. 7022000U]		
	1981		129 [123, no. 7022000U]		
	1982		76 [123, no. 7022000U]		Allied-Signal Company
	1984		115 (est.) [123, no. 7021001Q]	60 – 85 [123, no. 7022000W]	Alcoa
	1985	131 [123, no. 7022000T]			Kaiser
Japan	1982	56 [123, no. 7021001T]	41 [123, nos. 7022000D and Z]		Hashimoto
	1984	77.4	48	11	and Chem Central Glass Moritakagaku Kogyo Co.
Western world	1984		263	150 – 175	
	1985	$\leq$370			
Worldwide[338]	1990	$>$400	700		

Table 11. Economic data for cryolite

Country or region	Year	Capacity, 10^3 t	Production, 10^3 t	Consumption	Principal companies
Western Europe	1981		104 [124]		Fluorosid S.p.a. (Italy)
	1985	85 (est.)	[125]		
USA	1981	48 – 55 including 30 from H_2SiF_6	[76], nos. 7391002M, D, and S, 7022000U]		Kaiser
Japan	1977 to 1980			15 – 20, including 97 % by recycling	

4.1. Ammonium Fluorides

Ammonium fluoride [*12125-01-8*] (M_r = 37.04), NH_4F, is a colorless, hygroscopic salt. It crystallizes in the hexagonal system. Its solubility in water is about 1000 g/L at 0 °C. It decomposes on heating to NH_3 and $NH_4F \cdot HF$. The latter, known as ammonium bifluoride [*1341-49-7*] (M_r = 57.04), is a white, transparent, hygroscopic salt (*mp* 126.1 °C, *bp* 239.5 °C) crystallizing in the orthorhombic system. Anhydrous ammonium fluoride is prepared by reacting anhydrous NH_3 and HF or by heating aqueous solutions of NH_4F at reduced pressure, so that a mixture of H_2O and NH_3 is distilled [127]. Very pure metal fluorides (CaF_2, MgF_2, etc.) may be prepared by reaction of NH_4–HF with the metal carbonates [128]. Direct fluorination of $NH_4F \cdot HF$ by elemental fluorine is employed for the manufacture of NF_3 (see Section 6.2).

Aqueous solutions of $NH_4F \cdot HF$ are used instead of hydrofluoric acid in many fields: decoration of glassware, glass washing, decolorization of iron stains on textiles, fungicidal treatment of wood [129], and etching of metals [129]. These surface treatments with $NH_4F \cdot HF$ solutions are used on many metals, for example: formation of a glazed finish on aluminum surfaces [130], removal of thin layers of surface oxides from thin aluminum films in integrated semiconductor circuits [131], and activation of magnesium [132] or steel [133] substrates before nickel plating or phosphatizing [133].

Solutions of $NH_4F \cdot HF$ in anhydrous HF have been proposed as electrolytes, for example, for the industrial production of elemental

fluorine [134], but this process is not yet in use because of corrosion problems (see Fluorine, Chap. 4.4.); however. these electrolytes are suitable in special small-scale electrolysis cells for the fluorination of organic molecules [135] and the deposition of conducting polyaniline coatings on electrodes [136].

4.2. Cobalt Fluorides

Cobalt difluoride [*10026-17-2*] (M_r= 96.93), CoF_2, crystallizes with the rutile structure (*mp* 1200 °C); it is obtained by the action of anhydrous HF on cobalt carbonate, whereas the great oxidizing strength of elemental fluorine or ClF_3 is needed for the preparation of cobalt trifluoride [*10026-18-3*]. CoF_3 (M_r= 115.93) is considered a *solid fluorine carrier* because it can release part of its fluorine in the free form when the temperature is raised:

$$2\,CoF_3 \longrightarrow 2\,CoF_2 + F_2$$

Glemser [137] claims that the decomposition does not really occur below 600 °C. Nevertheless, CoF_3 is a strong fluorinating agent that may be regenerated after use [130]. CoF_3 is especially useful in organic synthesis [138], e.g., for the preparation of perfluorocarbons from the hydrocarbons. Addition of fluorine to halogenated alkenes, such as trichloro trifluoro propene, by reaction with CoF_3 is also possible under carefully controlled conditions [139]. On the other hand, high temperatures and long contact times generally produce highly fluorinated compounds. CoF_3 is also capable of converting most metal oxides to high-oxidation-state metal fluorides.

Special care must be taken during handling, because CoF_3 is toxic and hygroscopic.

4.3. Lithium Fluorides

Lithium fluoride [*7789-24-4*], (M_r= 25.94), LiF, is a white salt (*mp* 846 °C; *bp* 1680 °C) that is sparingly soluble in water. It has the NaCl structure, like all alkali-metal fluorides. It is prepared by the action of hydrofluoric acid on lithium carbonate or hydrogen carbonate. The bifluoride [*12159-92-1*], LiF · HF, is unstable.

Quantities of 1 – 2 % of LiF are added to the electrolyte KF · 2 HF for the production of elemental fluorine. These additions have been reported to be effective for preventing the anode effect by improving the wettability of the carbon electrodes [140], [141]. This is probably related to the property of LiF to form highly conducting ternary compounds $C_xF_yLi_z$ with graphite fluoride [142].

LiF is also used in flux compositions for metal joining and in the ceramics industry. Extremely oxidation-resistant, it is a more efficient lubricant than previously reported soft oxides (PbO) at temperatures above 500 °C [130]. Large, pure crystals of LiF are employed in optical systems for ultraviolet, visible, or infrared light and in X-ray monochromators.

Finally, LiF is a component of many fused electrolytes. For example, the eutectic mixture of LiF, NaF, and KF, known as FLiNaK (composition in mol %: LiF 46.5, NaF 11.5, KF 42.0; *mp* 454 °C) is a suitable electrolyte for the electrodeposition of refractory metals (Zr, Ta, Nb, Cr, Mo, W, and their alloys) as structurally coherent, ductile, fine-grained deposits [143], [144]. This electrolyte is also used in high-temperature batteries.

In nuclear energy production, fused salt breeders, using a mixture of lithium, beryllium, and thorium fluorides at 500 °C, have been proposed as a replacement for the conventional pressurized-water reactors employing uranium. Molten LiF or Li_2BeF_4 have also been considered for fusion reactors in the breeding part or for cooling [145].

The world consumption of LiF is ca. 250 tonnes per year [146].

4.4. Magnesium Fluoride

Magnesium fluoride [*7783-40-6*], (M_r= 62.31), MgF_2, is a colorless salt with the rutile structure. It is formed by reaction of magnesium oxide and HF or magnesium carbonate and NH_4F · HF [128]. It is also a byproduct from the manufacture of elements such as beryllium by reduction of the corresponding fluoride by magnesium metal [147].

In Mg metallurgy and in the ceramics industry, MgF_2 is used as a flux. Single crystals of alkaline-earth fluorides, such as MgF_2, are

suitable for optical applications because of their large domain of transparency from the ultraviolet to the middle infrared region [130]. Infrared transparent windows may be prepared by hot-pressing MgF_2 powder [148].

Ternary intercalation compounds of graphite with fluorine and MgF_2 have been prepared [142]; these compounds have high electrical conductivity and would therefore have an important potential as cathodes or new electroconductive materials. Finally, the eutectic $NaF-MgF_2$ has been proposed in advanced latent-heat energy storage for solar power systems [149].

4.5. Nickel Fluorides

Nickel(II) fluoride [*10028-18-9*], (M_r= 96.71), NiF_2, is a green salt that sublimes above 1000 °C. It crystallizes with the rutile structure and is slightly soluble in water or anhydrous HF and more soluble in aqueous HF.

Nickel(II) fluoride is used in the manufacture of zinc oxide-based ceramic varistors [150] and in organic synthesis as a hydrofluorination catalyst [151]. The addition of NiF_2 to molten $KF \cdot 2\,HF$ has been reported to be effective in preventing the anode effect in fluorine manufacture [140], probably because, during electrolysis, the carbon anode is coated with a catalytic layer containing nickel ions in the higher oxidation state (+4).

Numerous complex salts of nickel(III) or (IV) have been reported in the literature [152], such as M_3NiF_6 and M_2NiF_6, where M = Na, K, Rb, or Cs. They are prepared by the action of a strong fluorinating agent on a mixture of nickel and alkali-metal halides. For example, K_2NiF_6 [*17218-47-2*] may be prepared by the action of elemental fluorine on a mixture of KCl and $NiCl_2$ at 275 °C [153]. These complex salts of Ni(III) and (IV) are fluorinating agents in anhydrous HF, and some of them can be used as sources of pure elemental fluorine [154].

Bis(tetrafluoroammonium) hexafluoronickelate [*63105-40-8*], $(NF_4)_2NiF_6$, which is a solid fluorine carrier, has been proposed as a high-energy solid-propellant oxidizer. It is prepared from Cs_2NiF_6 and NF_4SbF_6 in anhydrous HF [155].

4.6. Potassium Fluorides

Potassium fluoride [*7789-23-3*], (M_r= 58.10), KF, is a colorless to white hygroscopic salt that crystallizes with the NaCl structure (*mp* 858 °C, *bp* 1505 °C). Two hydrates are known: $KF \cdot 2\,H_2O$ (*mp* 41 °C) and $KF \cdot 4\,H_2O$ (*mp* 19 °C).

Potassium fluoride is prepared by reacting potassium carbonate (or KOH) with aqueous hydrofluoric acid. Care is necessary in handling the anhydrous salt, to prevent its hydration.

The anhydrous salt is used in preparing silver soldering fluxes. For brazing, mixtures of alkaline borates and potassium fluoride are generally preferred, but a mixture of KF and AlF_3 has been proposed for brazing aluminum and its alloys [156].

Anhydrous KF is used in organic synthesis as a catalyst for various reactions [157], [158] or to introduce fluorine into organic molecules [146], [159]. For example, fluoro compounds can be prepared by replacing labile chlorine atoms by fluorine atoms, as in the manufacture of sodium fluoroacetate, a rat poison. The nucleophilic strength of F^- and the solubility of KF in aprotic organic solvents may be improved by using crown ethers. The "naked" fluoride ion obtained is an efficient fluorinating agent [159].

Potassium fluoride is a component of an electrolyte known as FLiNaK (see Section 4.3).

In France, potassium fluoride may be added to table salt to increase resistance to dental caries: 0.76 g KF, corresponding to 0.250 g of fluorine, is added per kilogram of NaCl [160].

The tetrahydrate, $KF \cdot 4\,H_2O$, has been proposed for the storage of heat (below 100 °C) because its melting point is close to room temperature [145].

Potassium bifluoride [*7789-29-9*] (M_r= 78.11), $KF \cdot HF$, also called Frémy's salt, is a white, crystalline salt (*mp* 239 °C). It is prepared from potassium carbonate or hydroxide and aqueous hydrofluoric acid. The concentrated solution is cooled to permit crystallization. A process including precipitation in the presence of a saturated alcohol (ethanol, propanol) has been described in the patent literature [161]. $KF \cdot HF$ can also be obtained from fluosilicic acid [162]. Thermal decomposition of $KF \cdot HF$ produces HF.

Potassium bifluoride is used in fluxes [163], [164], in etching metals such as aluminum [129], [165], in frosting glass [166], and in the treatment of wood to prevent fungal attack [129], [167]. $KF \cdot 2\,HF$ [*12178-06-2*] (*mp* 71.7 °C) is prepared from commercial Frémy's salt and HF. In all modern fluorine-production units, the electrolyte has the approximate composition $KF \cdot 2\,HF$ for temperatures ranging from 80–110 °C (see Fluorine, Chap. 4.4.).

4.7. Sodium Fluorides

Sodium fluoride [*7681-49-4*] (M_r = 41.99), NaF, is a white, crystalline salt with the NaCl structure (*mp* 992 °C, *bp* 1704 °C). Its solubility in water is ca. 40–50 g/L between 0 and 100 °C. Relatively high concentrations of fluoride ions have toxic effects; the lethal dose for a 70-kg human is estimated to be 5–10 g of sodium fluoride [168]. NaF is made by reacting sodium carbonate or hydroxide with hydrofluoric acid; separation of the NaF from the solution and drying may be improved by adding a saturated alcohol (ethanol, propanol) to the solution [169]. If excess acid is added, sodium bifluoride [*1333-83-1*] (M_r = 61.99), $NaF \cdot HF$, is formed. At present, $NaF \cdot HF$ and NaF are frequently prepared from inexpensive materials such as the waste gases from superphosphate manufacture by conversion of H_2SiF_6 [170], [171]. The annual consumption of NaF is estimated at several million tonnes [146].

On heating above 160 °C, $NaF \cdot HF$ decomposes to NaF and HF. Use is made of this property in fluorine manufacture: the anodic gases, mainly composed of fluorine, contain ca. 7–10 % HF, which can be partially removed by condensation. If a low level of contamination is required, e.g., <0.02 vol %, HF is absorbed by sodium fluoride, forming $NaF \cdot HF$. The absorbent is regenerated by heating.

Sodium fluoride may be added to drinking water (0.7–1 mg F^-/L) to increase resistance to dental caries. It is also used for pickling stainless steel, in soldering and metallurgical fluxes, and in wood treatment [167].

In the reprocessing of irradiated fuels, ruthenium, niobium, and antimony can be separated by absorbing them on, or complexing them with, NaF at 400 °C [130]. NaF is a component of an electrolyte known as FLiNaK (see Section 4.3).

Sodium bifluoride is a good laundry sour: it removes iron stains and does not attack textile fibers. It is also used in etching and frosting glass [166], and in cleaning stone and brick facades. In aluminum manufacture, cryolite may be prepared by the reaction of aluminum hydroxide, $Al(OH)_3$, with solid $NaF \cdot HF$ [172].

4.8. Tungsten Hexafluoride

Tungsten hexafluoride [*7783-82-6*] (M_r = 297.84), WF_6, is a colorless gas at room temperature (*mp* 2.5 °C, *bp* 17.1 °C), which is hydrolyzed by water to WO_3. Its reactivity as an oxidant has been compared with that of hexafluorides of other transition metals [173], [174]: It is a mild oxidant and a poor fluorinating agent [174]. Its electrochemical behavior in anhydrous hydrogen fluoride was studied [175], and a new general preparative route for lower fluorides and hexafluorometallates of the transition metals ($NaWF_6$, etc.) was reported.

Tungsten hexafluoride is manufactured by the direct action of elemental fluorine on solid powdered tungsten, e.g., in a flame reactor. It is stored in steel containers (1–1000 kg). Its major use is in chemical vapor deposition (CVD), which produces tungsten coatings on both conducting and insulating substrates. Two chemical reactions are used to form such a deposit: (1) reduction of WF_6 by hydrogen at an elevated temperature or (2) thermal decomposition of WF_6. High-purity CVD coatings of tungsten, sometimes alloyed with rhenium, have found application, e.g., as targets for the production of X-rays [176] and solar energy collectors [177].

5. Fluorine–Silicon Compounds

Fluorine–silicon compounds contain two elements, silicon and fluorine, that are abundant in the earth's crust. They are raw materials used in the manufacture of broadly developed industrial products, such as aluminum and its derivatives. They are also used in the fine chemistry of electronic silicon. They provide potential access to anhydrous hydrofluoric acid. Because of the diversity of their reactions, silicon tetrafluoride

and fluosilicic acid and its salts are important intermediates.

5.1. Silicon Tetrafluoride

Under normal temperature and pressure, silicon tetrafluoride or tetrafluorosilane [*7783-61-1*], SiF_4, M_r= 104.09, is a colorless gas with a suffocating odor, which hydrolyzes rapidly in humid air. It can be manufactured by direct synthesis from the elements, but it usually arises as an industrial byproduct of the treatment of fluorine-containing minerals, such as fluorite or fluorapatites. The market for this product is limited. The potential for development lies in the production of finely divided silicas, electronic silicon, and optical fibers.

5.1.1. Physical Properties

The principal physical properties are shown in Table 12 [178], [179]. The sublimation curve is represented by the equation:

$$\log p = -1346.2T^{-1} + 12.61$$

The vaporization curve is represented by:

$$\log p = -975.0T^{-1} + 10.578$$

with p expressed in Pascal [180]. The changes in the other principal physical properties with temperature have been presented in the form of curves [181].

Table 12. Physical properties of silicon tetrafluoride

Property	Value
Boiling point *bp*	−90.2 °C (175.7 kPa) −86.8 °C (223.8 kPa)
Density ϱ	
Gas (STP)	$4.69 \cdot 10^{-3}$ g/cm^3
Liquid at −80 °C	1.598 g/cm^3
Solid at −195 °C	2.145 g/cm^3
Critical constants	
Temperature t_{crit}	−14.15 °C
Pressure P_{crit}	3.713 MPa
Thermodynamic constants	
Standard heat of formation ΔH_f°	−1614.94 kJ/mol
Standard free energy of formation ΔG_f°	−1572.65 kJ/mol
Specific heat C_p	73.64 Jmol^{-1} K^{-1}
Entropy S_{298}°	282.49 Jmol^{-1} K^{-1}

5.1.2. Chemical Properties

Tetrafluorosilane is an extremely stable product: it is not decomposed by an electric spark or by heating to 800 °C. However, its important property is its hydrolysis in the presence of water or water vapor. In the gas phase, hydrofluoric acid and finely divided silica form [182], creating abundant silica smoke,

$$SiF_4 + 2\,H_2O \rightleftharpoons SiO_2 + 4\,HF - 107.8\,kJ$$

In the aqueous phase, the final products are fluo-silicic acid and silica:

$$3\,SiF_4 + 2\,H_2O \longrightarrow 2\,H_2SiF_6 + SiO_2$$

Bases and sodium carbonate in dilute aqueous solution hydrolyze SiF_4 with formation of fluosilicate. e.g.:

$$3\,SiF_4 + 4\,NaOH \longrightarrow 2\,Na_2SiF_6 + SiO_2 + 2\,H_2O$$

Excess alkaline reagent decomposes the fluosilicate to fluoride:

$$Na_2SiF_6 + 4\,NaOH \longrightarrow 6\,NaF + SiO_2 + 2\,H_2O$$

Gaseous ammonia forms the solid adduct $SiF_4 \cdot NH_3$. Amines react similarly [183].

Metal oxides react with SiF_4 to form silica and the corresponding fluoride or fluosilicate. Metal chlorides lead to exchange reactions between Cl and F, with the formation of chloride fluorides [184].

Organic compounds form addition compounds. This is the case with alcohols [185] and especially with organomagnesium compounds (Grignard reagents), which lead to alkyl or aryl silicyl fluoride derivates [186].

5.1.3. Manufacture

Processes. The oldest manufacture of tetrafluorosilane was carried out by the action of sulfuric acid on fluorite. The presence of silica in the ore leads to the formation of SiF_4:

$$2\,CaF_2 + 2\,H_2SO_4 + SiO_2 \longrightarrow SiF_4 + 2\,CaSO_4 + 2\,H_2O$$

However, the most widely used processes start with either silica and hydrofluoric acid or fluosilicic acid or its salts.

Hydrofluoric acid is injected in gaseous form at the bottom of a quartz-lined vertical reactor. Improvements consist of injection of either gaseous or liquid hydrofluoric acid into different stages of the reactor for better control of the reaction temperature [187].

The silica may be dispersed in concentrated sulfuric acid, into which the gaseous hydrofluoric acid is injected [188]. A more advantageous continuous process uses fluosilicic acid instead of hydrofluoric acid, introduced with silica into a column whose upper part is fed with concentrated sulfuric acid [189].

Sodium fluosilicate can be treated with concentrated sulfuric acid, with stirring, in the presence of silica:

$$2\,Na_2SiF_6 + 2\,H_2SO_4 + SiO_2 \longrightarrow 3\,SiF_4(g) + 2\,Na_2SO_4 + 2\,H_2O$$

Sodium fluosilicate can also be heated to cleave the molecule:

$$Na_2SiF_6 \longrightarrow 2\,NaF + SiF_4$$

Heating the product to around 900 °C produces silicon tetrafluoride at pressures of 0.06 MPa [190]. Other fluosilicates, e.g., the barium salt can be used. These processes lead to a very pure product.

All of these processes lead to gaseous SiF_4, which is dried, generally in a column of concentrated sulfuric acid, and then compressed and stored in pressure-resistant steel containers at 6.9 MPa.

Materials. In the manufacture of silicon tetrafluoride, hydrofluoric acid is present either as a raw material or in a stage of the process. This requires material and technology related to those used in the manufacture of hydrofluoric acid. However, pure anhydrous gaseous SiF_4 is only very slightly corrosive.

Environmental Protection. The product is a gas that, upon contact with atmospheric moisture, hydrolyzes to form hydrofluoric acid. Ventilation should be provided in the work areas, in case of the accidental emission of gases. The effluents are to be treated with an alkaline solution to trap the acid and with a calcium salt, which precipitates F^- ions. The waste standards depend on the legislation of each country.

5.1.4. Transport

Silicon tetrafluoride, a toxic gas with a secondary corrosive danger, is subject to national and international transport regulations. The product is listed in the UNO international nomenclature under number 1859.

5.1.5. Analysis

A chromatographic method is used for gases. After dissolution in aqueous medium, analytical methods for fluosilicic acid can be applied (see 5.2.5).

5.1.6. Use

The industrial use of silicon tetrafluoride is limited. The treatment with SiF_4 of concretes and cement parts that must withstand corrosion has been developed during the last twenty-five years [191]. The hydrolysis of silicon tetrafluoride in a high-temperature flame is a process for the manufacture of finely divided pyrogenic silica with a large specific surface area [192]. Silicon tetrafluoride has been used in the manufacture of silane [*7803-62-5*] and of photovoltaic or electronic silicon, e.g., in the processes of Allied Chemical Company [193] and Sharp Corporation [194].

Byproduct silicon tetrafluoride is converted in the phosphoric and hydrofluoric acid plants to fluosilicic acid, which can be transported and used as a raw material.

Union Carbide (Matheson Department) is one of the oldest producers. New facilities producing under 100 t/a have been announced by Allied Chemical Company (USA) and Central Glass (Japan). Sales are small, e.g., $ 1 000 000 in the USA in 1983 [195].

5.1.7. Toxicology [196]

Silicon tetrafluoride is a nonflammable gas that is highly irritating to the skin and eyes. When

heated, it decomposes and emits toxic fluorine compounds. Threshold limit value (time-weighted average): 2.5 mg (F)/m^3.

5.2. Fluosilicic Acid

Fluosilicic acid or hydrofluosilicic acid [*16961-83-4*], H_2SiF_6, M_r = 144.09, is used mainly in the form of aqueous solutions obtained by the treatment of fluorapatites. It is a raw material for the manufacture of aluminum fluoride and synthetic cryolite. The foreseeable development lies as a raw material to replace fluorite in the manufacture of anhydrous hydrofluoric acid, when the material problems have been resolved.

5.2.1. Physical Properties

Fluosilicic acid does not exist in the anhydrous state but only in aqueous solutions, whose properties depend on the concentration of H_2SiF_6 and the temperature. At 17.5 °C the density of fluosilicic acid solutions has the following values:

% H_2SiF_6	5	10	15	20	25	30
d, g/cm^3	1.0407	1.0834	1.1281	1.1748	1.2235	1.2742

At 25 °C, a 60.79 % solution has a specific gravity of 1.4634 [197], [198].

The freezing point depression in water is 0.317 °C for 1 wt % fluosilicic acid. Mono-, di-, and tetrahydrates are present in the solid state.

Distillation of a fluosilicic acid solution produces a gas phase consisting of silicon tetrafluoride, hydrofluoric acid, and steam. A mole ratio HF: SiF_4 of 2 in the gas phase, corresponding to fluosilicic acid, is the case only over an aqueous phase containing 13.3 wt % of fluosilicic acid. For lower concentrations, the mole ratio HF : SiF_4 is greater than 2 in the gas phase, i.e., the aqueous phase is enriched in silica. For solutions of 13.6 wt % or higher, the molar ratio HF : SiF_4 decreases, becoming less than 1 for 30 wt % fluosilicic acid. Hydrofluoric acid is enriched in the aqueous phase, and silicon tetrafluoride is enriched in the gas phase [199].

In contrast, another study found the partial pressures of SiF_4 and HF in the vapor phase to fit the following equations [200]:

$$\log p_{HF} = (1808.5 - 109.89c)/T - 8.625 + 0.49c$$
$$\log p_{SiF_4} = (1472.89 + 50.02c)/T - 1.625 + 0.386c$$

where p is the partial pressure in Pascal, T the absolute temperature, and c the concentration in wt % of H_2SiF_6 in the aqueous solution.

The presence of a pseudoazeotrope in concentrated solutions is due to the molecular association $H_2SiF_6 \cdot SiF_4$ and to a desolvation of SiF_4 [201]. Distillation and concentration of fluosilicic acid, without precipitation of silica, is possible if hydrofluoric acid is present. For example, a solution containing 36 % of fluosilicic acid and 10 % of hydrofluoric acid, boiling at 116.1 °C at 101.26 kPa, can be obtained [202].

The most recent thermodynamic values [203], [204] are:

$\Delta H_f^0 = -2376.5$ kJ/mol
for H_2SiF_6 in a solution containing 41 moles of HF and 75 moles of H_2O
$\Delta H_f^0 = 211.4$ kJ/mol
for $2\,HF(g) + SiF_4(g) \longrightarrow H_2SiF_6(g)$

For the SiF_6^{2-} ion:

$\Delta H_f^0 = -2389.1$ kJ/mol
$\Delta G_f^0 = -2199.4$ kJ/mol
$S_{298} = 122.2$ J mol^{-1}K^{-1}

5.2.2. Chemical Properties

Fluosilicic acid is a strong acid in aqueous solution. The neutralization and the decomposition of the acid can be distinguished in the overall neutralization curve. Two equilibria occur in aqueous solutions:

$$SiF_6^{2} \rightleftharpoons SiF_4 + 2\,F^-$$
$K = 1 \cdot 10^{-6}$ [205]
$$SiF_6^{2} + 2\,H_2O \rightleftharpoons SiO_2 + 4\,H^+ + 6\,F^-$$
$K = 1.2 \cdot 10^{-27}$ [206]

The careful addition of strong bases under vigorous agitation leads to the formation of the fluosilicate. Sodium or ammonium fluosilicate can be produced by this reaction. On the other hand, any excess of reagent over the stoichiometric quantity gives silica and the fluoride, as

a result of a displacement of the second equilibrium. Sodium or ammonium fluoride can be manufactured by this method. In the case of the aluminum salt, the silica precipitates but the aluminum fluoride supersaturates.

In highly acid medium, particularly in concentrated sulfuric acid, the fluosilicic acid is cleaved according to the first equilibrium to give silicon tetrafluoride, which is insoluble in the medium, and hydrofluoric acid, which stays in solution.

Fluosilicic acid displaces the Cl^- ion of sodium chloride with formation of hydrochloric acid and sodium fluosilicate:

$$H_2SiF_6 + 2\,NaCl \longrightarrow Na_2SiF_6 + 2\,HCl$$

5.2.3. Manufacture

Raw Materials. Fluosilicic acid can be manufactured from silica and aqueous hydrofluoric acid solutions, but in practice only silicon tetrafluoride, a byproduct from other processes, is used.

Fluosilicic acid solutions are produced in hydrofluoric acid plants as a result of the presence of silica in the fluorite ore. Approximately 50 kg of H_2SiF_6 is produced per tonne of HF. This represents an annual potential of 50×10^3 t of fluosilicic acid on a worldwide basis (see Section 1.3.2).

Fluosilicic acid is produced during the manufacture of fertilizers or of wet-process phosphoric acid by the action of sulfuric acid on fluorapatite ores. Silicon tetrafluoride and hydrofluoric acid are emitted during the reactions and, above all, during the concentration of phosphoric acid from 30 wt % to 40 and 52 wt % P_2O_5.

Processes. The reaction of a dilute solution of hydrofluoric acid and fluosilicic acid on silica takes place in equipment permitting good contact of the reagents. A vertical column containing powdered silica and fed with liquid at approximately 50 °C is suitable [208] and produces a concentrated fluosilicic acid solution (42 – 43 wt % H_2SiF_6) containing only traces of hydrofluoric acid.

The processes for fluosilicic acid manufacture are normally integrated in the plants for manufacture of superphosphate or phosphoric acid. They consist of a recovery of the gases rich in silicon tetrafluoride and hydrofluoric acid. These gases also contain water vapor and droplets of phosphoric acid.

The recovery is carried out in several columns placed in series and fed by a fluosilicic acid solution whose content is established by the desired recovery yield, a function of the liquid : gas ratio at operating temperature. The solution obtained contains 15 – 25 wt % of H_2SiF_6, [209–212].

The effectiveness of the processes lies in limiting the entrainment of phosphoric acid. Several systems exist; the following are typical:

1) Mechanical devices, such as cyclones, and baffles operating dry or with aqueous medium and recycling of the liquid [213].
2) Packed or sprayed columns providing intimate contact between the gases and the sprayed liquid that consists of a fluosilicic acid solution, which is recycled to the phosphoric acid concentrators [214].

A very effective process consists of using as the wash liquid sulfuric acid containing 80 % H_2SO_4 which, when saturated with hydrofluoric acid, lets the SiF_4–HF mixture pass, while ensuring the retention of phosphoric acid droplets. The contaminated sulfuric acid is recycled to the phosphate ore treatment unit. This process can be applied to the gases discharged from the manufacture and concentration of phosphoric acid or from superphosphate fertilizer units [215].

The use of one of these two types of purifiers, or of a combination of the two, considerably reduces the phosphorus content of the fluosilicic acid recovered downstream.

The phosphate content, expressed as a mass ratio P_2O_5 : F, of close to 0.03 – 0.05 in the entering gases is reduced to 0.002 – 0.005 with a mechanical system. With a wash column, the ratio drops to 0.001. With the two systems placed in series and a low gas velocity, a ratio of 0.0007 can be reached [216]. This reduced phosphate content is important when the acid is to be used in the manufacture of aluminum fluoride or cryolite.

Materials. The construction materials for the columns and storage units must be resistant to an aqueous medium of fluosilicic acid and dilute hydrofluoric acid. Ebonite-coated steels are suitable. Plastics, such as polyethylene and polypropylene, show good resistance. The use of metals should be preceded by corrosion tests.

5.2.4. Transport

Fluosilicic acid, a dangerous and corrosive material, is subject to national and international regulations relating to transport by railroad (RID), road (ADR), and to international maritime transport regulations. The quantities that can be transported by air are limited. The product is listed in the UNO international nomenclature under number 1778.

5.2.5. Quality and Analysis

Fluosilicic acid is marketed in the form of aqueous solutions containing 20 – 35 wt % H_2SiF_6. It generally contains an excess of fluorine over the stoichiometric quantity of the formula. The commercial product may also contain chloride, phosphate, sulfate, and traces of sulfur dioxide and metallic impurities.

The number of standard analytical methods is small: potentiometric determination of fluorine, ISO 6677 (1983), and some standard analyses for sodium fluosilicate, other analyses have been proposed. Classical methods, such as acidimetry in alcoholic medium [217–219] and determination of $HF - H_2SiF_6$ acid mixtures [220], can be used. Impurities are determined by conventional methods after evaporation in the presence of sodium carbonate.

5.2.6. Use

The first use of fluosilicic acid was as a bactericidal and fungicidal agent (breweries, wood). Like silicon tetrafluoride, fluosilicic acid changes the setting rate and the texture of cement [191].

The aluminum industry is the largest consumer of fluosilicic acid (40 % in the USA), where it is used for the manufacture of aluminum fluoride and cryolite [221].

The manufacture of sodium, ammonium, magnesium, and potassium fluosilicates is also important (30 % of the demand in the USA). Finally, the use fluosilicic acid in water fluoridation, in competition with alkali metal fluorides, is developing in the USA.

Fluosilicic acid is also a raw material for the manufacture of hydrofluoric acid, but corrosion problems are limiting the development of this type of process.

5.2.7. Economic Aspects [222]

The USA is the country producing the largest amount of fluosilicic acid, approximately 90 000 t/a in terms of 100 % H_2SiF_6 by more than 25 companies operating phosphoric acid and superphosphate plants. Only three companies carry out production not combined with phosphate production.

In other countries, fluosilicic acid is used principally in the manufacture of aluminum fluoride. There are few commercial transactions, because the product is frequently used on site. The listed selling price was of the order of $ 100/t of 100 % material in 1982 [223].

5.2.8. Toxicity [224]

Fluosilicic acid irritates mucous membrane, eyes, and skin and emits corrosive and toxic fluoride vapors. The lowest published subcutaneous lethal dose for the frog is 140 mg/kg, while the threshold limit value (time-weighted average) is 2.5 mg (F)/m^3.

5.3. Fluosilicates

Fluosilicates are the salts produced by reaction of fluosilicic acid with bases or metal salts. The most important fluosilicate is sodium fluosilicate. Other fluosilicates, such as those of ammonium, potassium, zinc, and magnesium, have also been produced commercially.

Table 13. Solubilities of some fluosilicates in water in g/100 g of water

Fluosilicate	Temperature 20 °C	25 °C	50 °C	100 °C
$BaSiF_6$ [*17125-80-3*]		0.025		0.044
K_2SiF_6 [*16871-90-2*]		0.177		0.82
$MgSiF_6$ [*16949-65-8*]	23.00		28.54	
Na_2SiF_6 [*16893-85-9*]	0.64	0.76	1.27	2.45
$(NH_4)_2SiF_6$ [*16919-19-0*]		18.5	26.15	37.90
$ZnSiF_6$ [*16871-71-9*]	35.16		38.30	42.18

5.3.1. Properties

The solubilities of some fluosilicates in water as functions of temperature [225] are shown in Table 13. Solutions of fluosilicates are sensitive to hydrolysis in alkaline medium [226]:

$$SiF_6^{2-} + 4\,OH^- \rightleftharpoons SiO_2 + 2\,H_2O + 6\,F^-$$

Dispersions of fluosilicates by peptization of solutions are obtained by adding a salt, a polyhydric alcohol, and a weak base to the solution [227]. In the presence of acid, fluosilicate solutions release silicon tetrafluoride.

The thermal decomposition of fluosilicates takes place with release of gaseous silicon tetrafluoride and formation of the solid fluoride [228]. This decomposition, particularly in the case of the barium salt, provides SiF_4 for use in electronics [229].

5.3.2. Manufacture

The manufacture of fluosilicates, especially the sodium salt, is carried out by neutralization of fluosilicic acid. The operation is carried out under vigorous agitation, with control of the ratio of reagents in order not to simultaneously produce the fluoride by introduction of excess alkali. Sodium fluosilicate precipitates in good yield because of its low solubility. It can be purified by dissolution in aqueous sodium chloride [230]. For the more soluble ammonium fluosilicate, vacuum crystallization is necessary to increase the yield, or crystallization must be carried out at low temperature [231]. The precipitates are washed and dried.

Sodium and potassium fluosilicates are also prepared by the action of fluosilicic acid on the alkali-metal chloride. The products are filterable solids [232], [233].

Zinc, magnesium, and barium fluosilicates are obtained by the action of fluosilicic acid on a salt of the corresponding metal, such as the carbonate.

5.3.3. Analysis

There are some standard analytical methods for sodium fluosilicate:

ISO TC 47/4281 for free acidity
ISO TC 47/5442 for sulfur
ISO TC 47/54443 for iron
ISO TC 47/5915 for granulometry
ISO TC 47/6229 for free silica

5.3.4. Uses, Economic Aspects

Sodium fluosilicate, the most important fluosilicate, is used in water treatment in some countries, especially in the USA, in the treatment of effluents of industrial laundries [234], and in glass manufacture as a polishing or etching additive [235]. It is also a raw material for cryolite and for silicon tetrafluoride.

The less important ammonium fluosilicate is also used in industrial laundries. The other fluosilicates have different uses: the magnesium salt acts as a hardener for concrete, and the potassium salt is used in the manufacture of porcelain enamels and porcelains [236]. The building industry uses fluosilicates in coatings for the protection of walls and floors against certain chemical agents [237–239]. The barium salt is a luminophor in ultraviolet lamps [240].

The fungicidal and insecticidal properties of fluosilicates have been used in certain commercial products [241].

The price of sodium fluosilicate was $ 270 per tonne in 1982 [242].

5.3.5. Toxicity

For sodium fluosilicate [243]:

oral LD_{50} for the rat	125 mg/kg
published minimum	70 mg/kg
subcutaneous lethal dose for the rat	
published minimum oral lethal dose for the rabbit	125 mg/kg
published minimum	448 mg/kg
subcutaneous lethal dose for the frog	2.5 mg (F)/m^3
threshold limit value (time-weighted average)	

Other slightly soluble or soluble fluosilicates, where the cation is not toxic, have similar toxicities.

6. Other Inorganic Fluorine Compounds

6.1. Sulfur Fluorides [244], [245]

Among the sulfur–fluorine compounds S_2F_2, SF_2 [*13814-25-0*] and its dimer S_2F_4 [*27245-05-2*], SF_4 [*7783-60-0*], S_2F_{10} [*5714-22-7*], and SF_6 [*2551-62-4*], only two have important industrial applications: SF_4 and SF_6.

Physical properties of S_2F_2 have been reported since 1923 [246]. Two isomeric forms have been established [247]: pyramidal molecules of thiothionyl fluoride $S=SF_2$ [*16860-99-4*] and nonplanar chains F–S–S–F of difluorodisulfane [*13709-35-8*].

Sulfur difluoride, SF_2, and its dimer S_2F_4 are not especially stable. A review of the properties of these lower sulfur fluorides, including difluoropolysulfanes S_nF_2, has been published [248].

Disulfur decafluoride, S_2F_{10} (M_r = 254.11), is extremely toxic, causing the death of rats within 16–18 h of exposure to S_2F_{10} vapors at a concentration of 1 ppm. It is a colorless, volatile liquid (*mp* – 92 °C, *bp* 29 °C) that does not react with water at ambient temperature. The molecule is composed of two octahedral SF_5 groups linked by an S–S bond. It decomposes at ca. 200 °C:

$$S_2F_{10} \longrightarrow SF_4 + SF_6$$

a reaction used in the purification of SF_6.

The preparation, physical, and chemical properties of thionyl fluoride [*7783-42-8*], SOF_2 [249], and sulfuryl fluoride [*2699-79-8*], SO_2F_2, [250] have been reviewed. The latter has special industrial applications: fumigation of food material [251] and catalysis.

6.1.1. Sulfur Hexafluoride [252]

Preparation. Elemental fluorine is not necessary for the preparation of SF_6, which can be made by pyrolyzing SF_5Cl, the latter manufactured by the reaction of sulfur or a sulfur chloride with chlorine and an amine –HF complex [253]. The electrochemical production of SF_6 has been patented [254], [255].

However, almost all methods consist of burning sulfur in a stream of fluorine:

$$S + 3F_2 \longrightarrow SF_6 \quad \Delta H = -1096\,\text{kJ/mol}$$

Several techniques are described in the literature [256–259] that use simple sulfur melting pot or flame reactors. The heat of reaction must be removed by cooling the reactor jacket.

The crude gas produced, consisting mainly of SF_6, must be purified because it contains SOF_2, SO_2F_2, and toxic lower fluorides such as S_2F_2, SF_4, and S_2F_{10}. The gases are first washed in a KOH spray scrubber to remove the major portion of the hydrolyzable compounds (S_2F_2, SF_4, HF, F_2, SO_2F_2, etc.). The gas mixture is then passed through a metal tube at 400 °C to pyrolyze the S_2F_{10} to SF_4 and SF_6. Another washing with an alkaline solution is necessary to remove the last portions of hydrolyzable compounds. The gases next pass through a tower containing a suitable drying material, such as alumina, phosphorus pentoxide, or barium oxide, before being purified over charcoal, distilled, compressed, and transferred to steel cylinders for storage.

Properties. Discovered in 1900 [260], sulfur hexafluoride (M_r = 146.05) is a colorless, tasteless, incombustible gas (sublimation point −63.9 °C). Practically inert from a chemical and biological standpoint, it does not react with water, caustic potash, or strong acids and can be heated above 500 °C without decomposition. However, at very high temperatures, or in the presence of an electric arc, SF_6 slowly decomposes, yielding toxic byproducts including lower

fluorides of sulfur. The chemical inertness of SF_6 is due to the fluorine shielding of the sulfur atom: the molecule is octahedral, S being covalently saturated by six fluorine atoms. SF_6 has a high dielectric strength and a high recombination rate after dissociation by electric discharges. Its dielectric constant has a value of 1.002049. Finally, it is a good acoustic and thermal insulating agent.

These properties are responsible for the specific uses of the gas.

Uses. Because of the good dielectric properties of SF_6 and its ability to trap electrons during arc blowing, 90 – 95 % of the SF_6 produced is used in electromechanical equipment such as high-voltage circuit breakers, compact switching stations that are very safe and need little maintenance, and insulation of electric cables for transmission of high electric power through difficult or protected sites (the importance of this will increase if its cost can be reduced).

Minor uses of SF_6 (about 5 % of the production): insulation of nuclear particle accelerators and X-ray equipment, microwave and radar equipment, thermoacoustic insulation of windows (double-pane systems enclosing SF_6 and offering a high degree of acoustic and thermal insulation [261]), meteorology and medical uses related to its inertness, treatment with a plasma containing SF_6 for etching silicon [262] or for increasing the water repellency and wet strength of kraft paper [263], and protection of molten magnesium in the magnesium industry [264].

Specifications. The byproducts from the production of SF_6 must be completely removed from the product, since they are toxic. The absence of toxic impurities in SF_6 is verified by exposing five albino mice for 24 h to an atmosphere containing 79 % by volume of SF_6 and 21 % O_2. After 72 h, they must show no signs of intoxication [265].

The other specifications [266] are:

air	<500 ppm by weight
CF_4	<500 ppm
H_2O	<15 ppm
acidity, computed as HF	<0.3 ppm
hydrolyzable fluorides, computed as HF	<1 ppm
mineral oil content	<10 ppm

Economic Aspects. SF_6 is shipped as a liquefied gas in steel cylinders (several kg) or in large containers (several hundred kg). Bulk quantities are available in tube trailers.

In 1986, the world market was estimated to be 5000 – 7700 t/a. SF_6 is manufactured by Allied Chemical Corporation and Air Products and Chemicals in the United States; Asahi Glass, Hashimoto Chemical Industries, and Kanto Denka Kogyo in Japan; and Montedison, Kali-Chemie, and ATOCHEM-COMURHEX in Europe. Its price is ca. \$ 4 – 8 per kg and \$ 8 – 12 per kg in 1992 [339].

6.1.2. Sulfur Tetrafluoride [267]

SF_4 (M_r = 108.06) is a colorless, nonflammable gas (mp – 121 °C, bp – 40.4 °C) with a penetrating odor; it is very toxic. The pure gas was first prepared in 1955 by the reaction of fluorine and sulfur at low temperature [268], a technique still used for its manufacture [269]. SF_4 is also produced by reaction of SCl_2 with AM · nHF, where AM is an amine (aliphatic tertiary amine, pyridine, quinoline, etc.) and n = 1 – 4 [270], [271]. Other reactions have been proposed in the patent literature: reaction of SCl_4 or a mixture of SCl_2 and Cl_2 with HF in the presence of CCl_4 [272] or oxidation of sulfur with UF_6 [273].

The SF_4 molecule is a distorted trigonal bipyramid. It reacts rapidly with water:

$$SF_4 + H_2O \longrightarrow SOF_2 + 2\,HF$$

As a consequence, all reactions with SF_4 must be performed under anhydrous conditions.

The high reactivity of SF_4 is used to form selectively fluorinated compounds [274]: $-CF_2-$ groups from ketones or thiocarbonyl compounds; $-CF_3$ groups from carboxylic acids, esters, anhydrides, or amides; and inorganic fluorides MF_x from oxides or sulfides. However, several less important uses have been developed: increasing the water repellency of polymeric materials by treatment with SF_4, preparation of high-lubricity elastomers, and treatment of carbon blacks to prepare improved rubber vulcanizates.

In spite of its interesting properties, the applications of SF_4 are limited because of its *toxicity.* A concentration of 50 ppm was found to

be lethal to one of two animals exposed for four hours; all animals exposed to a concentration of 10 ppm for one hour showed irregular breathing and signs of irritation [275].

SF_4 is supplied as a liquefied gas in steel cylinders (Air Products and Chemicals). The US production is several hundred kilograms per year. The price was about $ 170 per kg in 1982.

6.2. Nitrogen Fluorides [276]

Nitrogen Trifluoride . Among the numerous N – F compounds, only NF_3 [*7783-54-2*] has achieved commercial importance: NF_3 (M_r = 71.00) is a colorless toxic gas (*mp* – 206.8 °C, *bp* – 129 °C). It does not react with common metals, water, or dilute basic or acid solutions at ordinary temperatures, although it is a strong oxidizing agent at elevated temperatures. It fluorinates organic compounds, but the reaction is frequently explosive. It reacts with many metals to produce N_2F_4 and the metal fluoride.

Nitrogen trifluoride is manufactured by direct fluorination of NH_3 or $NH_4 \cdot HF$ by F_2 [277], [278] or by electrolysis of ammonium fluoride-hydrogen fluoride melts [279], [280].

Nitrogen trifluoride is preferred to F_2 in high-power chemical lasers because of its ease of handling at ambient temperatures. Its production is expected to increase in the near future with the development of these lasers. Minor uses have been mentioned in the literature: plasma etching of semiconductor materials [281], production of perfluoroammonium salts, a proposed gas filler to increase the life and brightness of lamps [282], and as an oxidizing agent in rocketry (not yet commercialized).

Other Nitrogen – Fluorine Compounds. Ammonium fluoride is discussed in Chapter 4.

Perfluoroammonium salts, such as NF_4SbF_6 [*16871-76-4*] or NF_4AsF_6 [*16871-75-3*], have been synthesized by reaction of NF_3(g), F_2(g), and SbF_5(l) or AsF_5(l) at 100 – 200 °C under a pressure of $5 - 20 \times 10^6$ Pa (50 – 200 atm) [283]. They are convenient solid sources of F_2 and NF_3 because they are stable up to 200 °C but decompose rapidly above 300 °C, releasing NF_3 and F_2. The thermochemistry of these salts has been studied by Bougon et al. [284].

The properties of the nitrogen – fluorine compounds without commercial importance at the present time have been reviewed in the literature [276]: tetrafluorohydrazine [*10036-47-2*], N_2F_4, difluorodiazine [*10578-16-2*], N_2F_2, and fluorine azide [*14986-60-8*], N_3F. Among oxygenated N – F compounds: nitryl fluoride [*10022-50-1*], NO_2F, nitrogen fluoride oxide (nitrosyl fluoride) [*11094-71-6*], NOF, and fluorine nitrate [*7789-26-6*], NO_3F.

6.3. Oxygen – Fluorine Compounds

All oxygen fluorides have strong oxidizing power. Oxygen difluoride is the most stable.

Oxygen Difluoride [285], [286]. Oxygen difluoride [*7783-41-7*] (M_r = 54.00) is a toxic, colorless gas (*mp* – 223.8 °C, *bp* – 145.3 °C), discovered by Lebeau and Damiens in 1927. The molecule is nonlinear. It may be prepared by several methods: electrolysis of molten KHF_2 in the presence of water [287], electrolysis of aqueous hydrofluoric acid [288], or fluorination of 0.5 M aqueous NaOH [289].

$$2\,F_2 + 2\,NaOH \longrightarrow OF_2 + 2\,NaF + H_2O$$

Pure, dry OF_2 is stable up to 250 °C, but in the presence of water its handling becomes dangerous: mixtures of OF_2 and steam can explode easily because of the reaction

$$OF_2 + H_2O \longrightarrow 2\,HF + O_2$$

Oxygen difluoride reacts like elemental fluorine as a powerful oxidizing agent, but higher activation energies are required to initiate its reactions. Perhaps suitable for rocket and missile propellants, no reactions of particular interest in industrial preparative chemistry have been found, and it remains mainly a laboratory chemical.

Other Oxygen Fluorides [285], [286]. Dioxygen difluoride [*7783-44-0*], O_2F_2, is formed by an electric discharge through a gaseous $O_2 - F_2$ mixture at low pressure and liquid-air temperature. This unstable compound decomposes rapidly to O_2 and F_2 above –78 °C; it reacts explosively with most oxidizable materials, even at cryogenic temperatures.

Trioxygen difluoride, O_3F_2, and tetraoxygen difluoride, O_4F_2, are extremely unstable compounds, even at temperatures as low as −180 °C. They are two of the most powerful oxidants known.

Other Compounds. Hypofluorous acid [*14034-79-8*], HOF, formed by passing F_2 over cold water, has been suggested to be the active intermediate in the F_2–H_2O reaction. Compounds containing the –OF group have been classified as fluoroxides or oxyfluorides, but for inorganic compounds the term hypofluorites appears to be more suitable. The properties of these compounds have been reviewed in the literature [290].

6.4. Phosphorus – Fluorine Compounds [291]

Phosphorus fluorides, phosphorus oxyfluoride, the three fluorophosphoric acids, and their salts, are the most important phosphorus – fluorine compounds.

Phosphorus Fluorides [292], [293]. The most widely studied phosphorus fluorides are PF_3 [*7783-55-3*] and PF_5 [*7647-19-0*]. *Phosphorus trifluoride* (M_r = 87.97) is a colorless gas, slowly hydrolyzed by moist air (*mp* – 151.5 °C, *bp* – 101.8 °C). The molecule has a pyramidal structure and forms many adducts [e.g., $Ni(PF_3)_4$, $PtCl_2(PF_3)_2$] by donating the electron pair on the P atom. Its toxicity is related to its ability, like CO, to form a stable complex with hemoglobin.

A mixture of PF_3 with oxygen can detonate violently, but activation is necessary (electric spark).

$$PF_3 + 1/2\,O_2 \longrightarrow POF_3$$

Both F_2 and Cl_2 react rapidly, with PF_3 leading to PF_5 and PF_3Cl_2, respectively. PF_3 is usually prepared by an exchange reaction between PCl_3 and AsF_3 or SbF_3 in the presence of SbF_5 as the catalyst [294].

Phosphorus pentafluoride (or pentafluorophosphorane), PF_5 (M_r = 125.97), has been widely studied. It is a colorless, strongly fuming gas (*mp* – 93.8 °C, *bp* – 84.6 °C) that is very hygroscopic and easily hydrolyzable: a slight trace of moisture forms POF_3 and HF, and with more water, hydrolysis yields phosphoric acid, with oxyfluorophosphates as intermediates.

The molecule has a trigonal bipyramidal structure corresponding to sp^3d hybridization; PF_5 behaves as a typical electron-accepting Lewis acid. For this reason, it is an excellent Friedel-Crafts polymerization catalyst [295]. Adducts are formed with amines, ethers, and nitriles.

Exchange reactions may be used to prepare PF_5 [292], for example:

$$3\,PCl_5 + 5\,AsF_3 \longrightarrow 3\,PF_5 + 5\,AsCl_3$$

Other fluorinating agents, such as SbF_3 and SbF_5, are more convenient. PF_5 may also be obtained from PF_3, or from the thermal decomposition of sodium hexafluorophosphate between 300 and 1000 °C.

Phosphorus Oxyfluoride. Phosphorus oxyfluoride (phosphoryl fluoride) [*13478-20-1*] (M_r = 103.97), POF_3, is a gas at room temperature. It is formed from PF_5 by reaction with water; further hydrolysis forms a mixture of fluorophosphoric acids. Halide exchange methods allow preparation of POF_3 from $POCl_3$.

The tetrahedral structure of POF_3 is more stable than the 5-coordinate structures; POF_3 is considerably less active than PF_5 toward electron donors.

Fluorophosphoric Acids [296]. *Monofluorophosphoric acid* [*13537-32-1*] (M_r = 99.99), H_2PO_3F, is used as a polymerization catalyst; hydrolysis is complete in dilute solutions:

$$H_2PO_3F + H_2O \longrightarrow H_3PO_4 + HF$$

The most suitable method for the preparation of H_2PO_3F consists of the reaction of phosphorus pentoxide with concentrated hydrofluoric acid:

$$P_2O_5 + 2\,HF + H_2O \longrightarrow 2\,H_2PO_3F$$

Sodium monofluorophosphate is used in toothpastes; it probably facilitates the formation of fluoroapatite in teeth, which increases

resistance to dental caries. Passivation of metallic surfaces by K_2PO_3F (potassium monofluorophosphate) is another field of interest [297]: these coatings provide a good combination of corrosion protection and paint adherence and are superior to Cr(VI) compounds.

The esters of monofluorophosphoric acid are highly toxic nerve gases as a result of their cholinesterase inhibiting power. For example, Sarin [*107-44-8*] has an oral LD_{50} in rats of 550 µg/kg. These compounds are hydrolyzed in alkaline solutions.

Difluorophosphoric acid [*13779-41-4*], HPO_2F_2 (M_r = 101.98), is used as a catalyst. Its potassium salt has been reported to be a stabilizing agent in chloroethylene polymers [298].

Hexafluorophosphoric acid [*16940-81-1*], HPF_6 (M_r = 145.97), was first observed in the product obtained by reacting P_2O_5 with 40 % aqueous HF. The PF_6^- anion is stable and is isoelectronic with SF_6; however, heating with mineral acids results in hydrolysis. Dry alkali-metal salts heated to high temperatures decompose according to the equilibrium:

$$KPF_6 \rightleftarrows KF + PF_5$$

The lithium and potassium salts have been used as electrolytes in lithium anode batteries.

6.5. Halogen Fluorides [299–302]

Fluorine reacts exothermally with the other halogens over a wide temperature range. High oxidation states may be obtained: Br(V) and I(VII). Four compounds have commercial importance: ClF_3, BrF_3, BrF_5, and IF_5.

Physical Properties and Toxicology. Table 14 summarizes the physical properties of the well-known halogen fluorides. The structures of halogen fluorides and oxyfluorides have been reviewed [303]: ClF_3 and BrF_3 (vapor) both have a T-shaped structure, while the structure of BrF_5 and IF_5 is that of a tetragonal pyramid.

At room temperature, ClF_3 is a gas, while BrF_3, BrF_5, and IF_5 are liquids. All of these compounds are toxic; ClF_3 is considered to be the most toxic (of the order of HF) and most hazardous. It is extremely irritating to the eyes, skin, and respiratory tract. Concentrations of 50 ppm or higher are fatal in 30 minutes; a threshold limit value of 3 ppm has been suggested [304]. For IF_5, a threshold limit value of 2 mg/m^3 is given for an 8-hour workday and 40-hour workweek [305].

Preparation. Chlorine trifluoride is produced on a scale of several hundred tonnes per year by the direct reaction of fluorine with chlorine. The mixture is passed through a nickel tube at ~290 °C; temperature control is necessary because ClF_3 decomposes to F_2 and ClF at high temperatures.

BrF_3 and BrF_5 are manufactured by the gas-phase reaction:

$$Br_2 + xF_2 \longrightarrow 2\,BrF_x \quad x = 3 \text{ or } 5.$$

Several tonnes of BrF_3 are produced per year.

Many methods have been proposed to prepare IF_5. Several hundred tonnes are prepared each year. Fluorination of iodine with AgF, ClF_3, or BrF_3 is used for small-scale preparation. Industrial-scale production often involves the reaction of elemental fluorine with iodine [301]. For example, in a continuous process, 1 % iodine is dissolved in liquid IF_5 and introduced into a reactor in which the fluorination is achieved with fluorine [306].

Chemical Properties and Uses. Like fluorine, halogen fluorides are used for their fluorinating and oxidizing powers. Halogen fluorides such as ClF_3, BrF_3, or IF_5 are fluorinating agents that serve as substitutes for elemental fluorine because their reactions with particular compounds are milder or more specific and because handling is easier. Economy of operation is never claimed, because fluorine is needed for their preparation.

For example, IF_5 has become an important industrial compound because it is an easily storable liquid source of fluorine; it can be used as a solvent and has selective fluorinating properties in organic chemistry.

Metals react with halogen fluorides. In some cases, a passivating fluoride film is formed. Passivation of metallic piping with ClF_3 is used in the nuclear energy field to protect the metal against further corrosion by UF_6. Other possible uses of ClF_3 and BrF_3 in the nuclear energy area are reported: conversion of uranium to UF_6 for

Table 14. Physical properties of halogen fluorides

Compound	M_r	*mp*, °C	*bp*, °C
ClF [*7790-89-8*]	54.45	−155.6	−100.1
ClF_3 [*7790-91-2*]	92.45	− 76.3	11.7
ClF_5 [*13637-63-3*]	130.45	−103	− 13.1
BrF [*13863-59-7*]	98.9	− 33	20
BrF_3 [*7787-71-5*]	136.9	8.8	125.8
BrF_5 [*7789-30-2*]	174.9	− 60.5	41.3
IF_5 [*7783-66-6*]	221.9	9.4	104.5
IF_7 [*16921-96-3*]	259.9	4.5 (subl.)	

isotopic enrichment and recovery of uranium as volatile UF_6 from nuclear fission products, but at present UF_6 is apparently prepared exclusively by the action of F_2 on UF_4 [252].

Halogen fluorides react with many nonmetallic elements except rare gases, nitrogen and oxygen; water is violently decomposed, but dry Pyrex glass is not attacked below 100 °C.

Metallic halides (other than fluorides) react with all halogen fluorides to produce the fluorides and release free halogens. The highest oxidation state of the metal is usually obtained when ClF_3 is used. The fluorinating efficiency decreases in the order $ClF_3 > BrF_5 > IF_7 > ClF > BrF_3 > IF_5 > BrF$ [301]. Thus, only partial replacement of O by F is accomplished when oxides such as P_4O_{10} or WO_3 are dissolved in IF_5, whereas complete replacement of oxygen takes place in the reaction of WO_3 with BrF_3.

ClF_3 and the bromine fluorides react vigorously with organic compounds, leading to mixtures of products. Reactions are controlled more easily with iodine fluorides. Iodine pentafluoride is used on an industrial scale for the production of perfluoroiodoalkanes, which are intermediates in the synthesis of functional perfluoro compounds [252], [307], [308].

Other minor uses of halogen fluorides are mentioned in the literature: ClF_3 has been proposed as an oxidizer for rocket fuels such as hydrazine, but development has not been reported. Intercalation compounds of halogen trifluorides in graphite have been prepared [309]; highly conductive materials are obtained when cis-rich $(CH)_x$ polymer is exposed to IF_5 vapors as a result of doping the polyacetylene film [*25067-58-7*] [310]. During World War II, ClF_3 was used by Germany as an incendiary gas. At the present time the strong oxidizing properties of ClF_3 and BrF_3 are used in an oil-well-tubing cutter [311], [312].

Oxyfluorides of the halogens have been reviewed [313], [314].

6.6. Rare-Gas Fluorides [315–319]

The oxidizing properties of platinum hexafluoride, discovered by Bartlett et al. [320], [321], opened the door to rare-gas chemistry. The first xenon compound was formed by interaction of Xe and PtF_6 vapor at room temperature. Isolable compounds of the heavier gases (krypton, xenon, radon) are known, but work with radon has been limited by its intense radioactivity [319]. Xenon fluorides are prepared by heating mixtures of xenon and fluorine in nickel vessels; high fluorine pressures and low temperatures favor XeF_6 [*13693-09-9*] formation, whereas an excess of xenon leads to a good yield of XeF_2 [*13709-36-9*] [322]. Disproportionation of XeF_2 to XeF_4+ Xe is observed above 623 K.

The preparation of KrF_2 [*13773-81-4*] is carried out at liquid-nitrogen temperature, using a krypton – fluorine mixture irradiated with γ rays or UV radiation. An electrical discharge also gives good results. This fluoride is thermodynamically unstable and must therefore be handled with special care; it is one of the most powerful fluorinators known and serves for the preparation of the highest oxidation states of the elements. For example, it can be used to prepare gold pentafluoride.

The oxidative fluorinating strength increases along the series $XeF_2 < XeF_4 < XeF_6 < KrF_2$ and has been exploited in inorganic chemistry (for example, synthesis of tetravalent lanthanides, preparation of UF_6 from UO_2F_2, UF_4 or UO_3 [323]) and especially in organic chem-

istry [324]. The byproduct is often a noble gas. Graphite intercalation compounds of XeF_4 and XeF_6 have been tested as mild fluorinating agents for aromatic systems. Finally, noble-gas fluorides have been studied for possible uses in high-power lasers [325].

6.7. Graphite Fluorides

The name graphite fluorides is assigned to nonstoichiometric solid fluorocarbons with the general formula $(CF_x)_n$ or $CF_x (0 < x < 1.24)$. The structure of two of these intercalation compounds is known, corresponding to simple F : C ratios: poly(carbon monofluoride) CF [*11113-63-6*] and poly(dicarbon monofluoride) $CF_{0.5}$ [*71183-38-5*]. The general properties of these compounds have been reviewed [326–328].

The industrial product is a grey to white powder that is prepared by direct fluorination of carbon or graphite at 300 – 600 °C [329]. The chemical composition and crystal structure strongly depend on the reaction temperature and the crystallinity of the starting carbons. Graphite fluorides have a very low surface energy. Their conductivity is divided by 10^7 when x increases from 0.3 to 0.5. When $x > 0.5$, the compounds are insulators.

CF is used as a solid lubricant for high temperatures [330], [331], and as a cathodic material for primary Li batteries [332], [333], but carbon black or graphite must be added to the electroactive CF for improving the conductivity. A non-aqueous electrolyte is required, for example $LiBF_4$ or $LiClO_4$ in propylene carbonate or butyrolactone. These cells are now used on a large scale, especially in Japan. Their working voltage (3.2 V) and energy density (300 Wh/kg) are respectively twice and five times that of a conventional carbon/zinc cell.

CF_x/KOH/Zn primary batteries have been patented [334]. A surfactant must be added to insure the wetting of the CF powder by the aqueous electrolyte.

Ternary intercalation compounds of graphite with fluorine and a metallic fluoride such as LiF [*7789-24-4*], AlF_3 [*7784-18-1*], MgF_2 [*7783-40-6*], or CuF_2 [*7789-19-7*] were prepared [335–337]. It was found that some of these have a high electrical conductivity, comparable to that of graphite, and would therefore show a significant potential as new electroconductive materials.

7. References

1. R. L. Jarry, W. Davis, Jr., *J. Phys. Chem.* **57** (1953) 600 – 604.
2. C. E. Vanderzee, W. N. Rodenburg, *J. Chem. Thermodyn.* **2** (1970) 461 – 478.
3. I. Sheft, A. J. Perkins, H. H. Hyman, *J. Inorg. Nucl. Chem.* **35** (1973) 3677 – 3680.
4. M. Atoji, W. N. Lipscomb, *Acta Crystallogr.* **7** (1954) 173 – 175.
5. I. Sheft, A. J. Perkins, *J. Inorg. Nucl. Chem.* **38** (1976) 665 – 668.
6. J. Simons, J. H. Hildebrand, *J. Am. Chem. Soc.* **46** (1924) 2183 – 2191.
7. K. Fredenhagen, *Z. Anorg. Allg. Chem.* **210** (1933) 210 – 224.
8. G. Briegleb, *Naturwissenschaften* **28** (1941) 420 – 421.
9. W. Strohmeier, G. Briegleb, *Z. Electrochem.* **57** (1953) 662 – 667.
10. G. Briegleb, W. Strohmeier, *Z. Electrochem.* **57** (1953) 668 – 674.
11. J. N. Maclean, F. J. C. Rossotti, H. S. Rossotti, *J. Inorg. Nucl. Chem.* **24** (1962) 1549 – 1554.
12. R. L. Redington, *J. Phys. Chem.* **86** (1982) 561 – 563.
13. J. M. Beckerdite, D. R. Powel, E. T. Adams, *J. Chem. Eng. Data* **78** (1983) no. 3, 287 – 293.
14. W. Schotte, *Ind. Eng. Chem. Process Des. Dev.* **19** (1980) 432 – 439.
15. E. U. Franck, F. Meyer, *Z. Electrochem.* **63** (1959) 571 – 582.
16. D. F. Smith, *J. Mol. Spectrosc.* **3** (1959) 473 – 485.
17. J. H. Hu, D. White, H. L. Johnston, *J. Am. Chem. Soc.* **75** (1953) 1232 – 1236.
18. E. U. Franck, W. Spalthoff, *Z. Electrochem.* **61** (1957) 348 – 357.
19. R. M. Yabroff, J. C. Smith, E. H. Lightcap, *J. Chem. Eng. Data* **9** (1964) no. 2, 178 – 182.
20. H. M. Feder, W. N. Hubbard, S. S. Wise, J. L. Margrave, *J. Phys. Chem.* **67** (1963) 1148 – 1150.
21. J. H. Simons, R. D. Dresdner, *J. Am. Chem. Soc.* **66** (1944) 1070 – 1072.
22. J. H. Simons, J. W. Bouknight, *J. Am. Chem. Soc.* **54** (1932) 129 – 135.
23. J. Shamir, A. Netzer, *J. Sci. Instrum.* **1** (1968) 770 – 771.

24. K. Fredenhagen, J. Dahmlos, *Z. Anorg. Allg. Chem.* **178** (1929) 272 – 274.
25. G. K. Johnson, P. N. Smith, W. N. Hubbard, *J. Chem. Thermodyn.* **5** (1973) 793 – 809.
26. Allied Chemical Corp., *Hydrofluoric Acid* (1978).
27. G. H. Cady, J. H. Hildebrand, *J. Am. Chem. Soc.* **52** (1930) 3843 – 3846.
28. P. A. Munter, O. T. Aepli, R. A. Kossatz, *Ind. Eng. Chem.* **39** (1947) no. 3, 427 – 431.
29. P. A. Munter, O. T. Aepli, R. A. Kossatz, *Ind. Eng. Chem* **41** (1949) no. 7, 1504 – 1508.
30. *Gmelin* **3,** 181 – 238.
31. Buss, WO 82/03848, 1982 (D. Boese, R. Etter).
32. Du Pont, US 3 102 787, 1960 (D. McMillan, C. C. Quarles).
33. Buss, CH 319 931, 1957 (H. List).
34. H. List, EP 100 783, 1984.
35. Allied, US 3 607 121, 1969 (W. E. Watson, R. P. Troeger); US 3 718 736, 1970 (W. E. Watson, R. P. Troeger).
36. Atochem, FR 2 564 449, 1985 (P. Laroche, P. Thiery, Y. Verot).
37. R. S. Reed, US 4 036 938, 1977.
38. Buss, GB 1 262 571, 1969 (W. Blochl, B. Oberbacher).
39. Wellman Power Gas, US 3 758 674, 1969 (W. R. Parish).
40. Stauffer, US 3 257 167, 1963 (L. E. Tufts, J. T. Ruk-ker, T. H. Dexter).
41. Tennessee Corp. US 3 218 124-5-6-7-8-9, 1962 (T. T. Houston).
42. Goulding Cheminals Ltd., US 4 144 315, 1976 (R. E. Worthington).
43. Du Pont, US 3 914 398, 1974 (C. R. Faust).
44. Dow Chemical, US 2 832 669, 1957 (D. R. Allen).
45. *Mater. Perform.* **22** (1983) no. 11, 9 – 12.
46. K. Hauffe, *Z. Werkstofftech.* **15** (1984) 427 – 435.
47. K. Hauffe, *Z. Werkstofftech.* **16** (1985) 259 – 270.
48. R. L. Schuyler, *Mater. Perform.* **18** (1979) no. 8, 9 – 16.
49. C. T. E. F., 250 avenue Louise Boîte 71, 1050 Bruxelles.
50. Occupational Exposure to Hydrogen Fluoride, U.S. Dept. of Health, Niosh, March 1976.
51. M. A. Trevino, G. H. Hermann, W. L. Sprout, *JOM J. Occup. Med.* **25** (1983) no. 12, 861 – 863.
52. J. Braun, H. Stoss, A. Zober, *Arch. Toxicol.* **56** (1984) 50 – 54.
53. J. Camarasa, *Arch. Mal. Prof. Med. Trav. Secur. Soc.* **44** (1983) no. 6, 422 – 425.
54. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," *Chemical Economics Handbook,* SRI International, August 1982.
55. J. M. Molina, F. S. Rowland, *Nature (London)* **249** (1974) 810 – 812.
56. F. Lombard, *J. Fluorine Chem.* **18** (1981) 1 – 24.
57. L. E. Chapin, G. C. Liolios, T. M. Robertson, *Hydrocarbon Process.* **64** (1985) no. 9, 67 – 71.
58. B. F. Naylor, *J. Am. Chem. Soc.* **67** (1945) 150 – 152.
59. D. A. Schulz, A. W. Searcy, *J. Phys. Chem.* **67** (1963) 103 – 106.
60. O. Ruff, L. Leboucher, *Z. Anorg. Allg. Chem.* **219** (1934) 376 – 381.
61. S. S. Todd, *J. Am. Chem. Soc.* **71** (1949) 4115 – 4116.
62. A. Finch et al., *Can. J. Chem.* **46** (1968) 3447 – 3451.
63. H. E. Swanson, E. Tatge, *Nat. Bur. Stand. US Circ.* **539** (1953) no. 1, 69.
64. G. L. Kehl: *The Principles of Metallographic Laboratory Practice,* McGraw-Hill, New York 1949, p. 242.
65. H. S. Booth et al., *J. Am. Chem. Soc.* **72** (1950) 2567 –2575.
66. *Tech. News Bull. Nat. Bur. Stand. U.S.* **47** (1963) 91.
67. Bundesanstalt für Bodenforschung, Hannover, und Deutsches Institut für Wirtschaftsforschung, Berlin: [Federal Institute for Soil Research, Hannover, and German Institute for Economic Research, Berlin]: Untersuchung über Angebot und Nachfrage mineralischer Rohstoffe, IV Flußspat [Investigation Concerning Supply and Demand of Mineral Raw Materials, IV Fluorspar] (1974).
68. K. H. Jakob, Erzmetall **24** (1971) 186 – 191.
69. G. Clarke, *Ind. Miner. (London)* (1982) 25 – 46.
70. D. E. Morse, *Mineral Facts and Problems,* Bull 671 US Dept of Int., Bureau of Mines (1980).
71. H. E. Blake, Jr., et al., *Rep. Invest. U.S. Bur. Mines* **RI 7502** (1971).
72. Bayer, DE 2533 128, 1975 (S. Schneider).
73. Kali-Chemie, DE 2 535 658, 1975 (K. H. Hellberg).
74. S. S. Kumar, S. S. Kataria, *World Cem. Techn.* **12** (1981) no. 6, 279 – 285.

75. *Registry of toxic effects of chemical substances,* **vol. 1,** National Institute for Occupational Safety and Health, 1981 – 1982.
76. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH (Chemical Economics Handbook),* S.R.I. International, August 1982.
77. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH (Chemical Economics Handbook),* S.R.I. International, August 1982, no. 7391000 H.
78. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH (Chemical Economics Handbook),* S.R.I. International, August 1982, no. 7391000 I.
79. K. Weiss, *GDMB Schriftenr.* **35** (1980) 189 – 206.
80. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH (Chemical Economics Handbook),* S.R.I. International, August 1982, no. 7391002 G.
81. R. Hoppe, D. Kissel, *J. Fluorine Chem.* **24** (1984) no. 3, 327 – 340.
82. N. W. Chase et al.: "Janaf-Thermochemical Tables," *J. Phys. Chem. Ref. Data,* NSDRS-NBS (1971) 37.
83. O. Kubaschewski, *J. Less Common Met.* **16** (1968) 295 – 342.
84. J. D. Cox, *J. Chem. Thermodyn.* **9** (1977) no. 8, 705 – 706.
85. E. V. Khamskii, A. S. Bogatyrenko, *Zh. Prikl. Khim. (Leningrad)* **53** (1980) no. 1, 208 – 209.
86. A. Seidell: "Solubilities Inorganic and Metal Organic Compounds," vol. 1, 4th ed., Am. Chem. Soc. Washington, D.C. (1958) 179 – 180.
87. Bayer, DE 1 220 839, 1964 (G. Broja, H. Gradinger).
88. Imperial Smelting, GB 1 026 131, 1962 (J. L. Pegler).
89. Vereinigte Aluminium Werke, Kaiser Aluminium, DE-OS 3 405 452, 1984 (H. Sauer, J. N. Anderson, F. Kaempfa). V. Wolf, *Erzmetall* **37** (1984) no. 9, 4, 21 – 30.
90. F. Weinratter, *Chem. Eng. N.Y.* **71** (1964) no. 9, 32.
91. J. Y. Morlock, B. Nochet, P. Mollard, *2nd Congress on Phosphorus Compounds Proceeding, Boston, USA,* 21 May 1980, ed. Imphos, 799 – 808.
92. Alcoa, GB 647 216, 1946. OSW, FR 1 278 454, 1961. Swiss Aluminium Ltd., DE 3 042 367, 1982 (A. Skania, R. Stank, H. P. Schenkel).
93. Lentia, DE 2 837 211, 1977 (W. Tschebull, W. Kepplinger).
94. N. A. Solntseva, A. M. Zagudaev, E. A. Burakov, T. V. Stratonova, *Tsvetn. Met.* 1983, no. 1, 47 – 48.
95. M. Grobelny, PL 73 113, 1971. Bayer, DE-OS 2 307 925, 1973 (W. Becker, H. Jonas, W. Weiss).
96. D. Harder, *I.G. Farben Industrie Bitterfeld,* Document no. 0635 (1942) 1 – 17. H. E. Blake, R. K. Koch, *Bur. Mines Rep.* **6524** (1964) 12 – 15. P. M. R. Versteegh, T. J. Thoonen, *Chem. Process Eng. (London)* **53** (1972) no. 6, 11.
97. Ugine Kuhlmann, FR 2 052 118, 1969 (P. Mollard).
98. T. V. A., US 2 981 597, 1961 (G. Tarbutton, T. D. Barber-Farr, T. M. Jones, H. T. Lewis). J. B. Cotton, *Chem. Eng. Prog.* **66** (1970) no. 10, 56 – 62.
99. Bayer, DE-OS 2 064 696, 1970 (P. Langrock, E. Podschus, M. Schulze, E. Weise).
100. Elkem, DE-OS 2 313 495, 1973 (F. Steineke). Vereinigte Aluminium-Werke – Lurgi, DE-OS 2 403 282, 1974 (V. Sparwald).
101. Sumitomo Chal Cy, FR 2 305 396, 1976 (T. Abe, T. Okamoto, T. Sakakura, T. Tateno). Anaconda Aluminum Co., US 3 941 874, 1976 (T. F. Payne).
102. I. Naray-Szabo, K. Sasvari, *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **99** (1938) 27 – 31.
103. K. Papp, J. Kurty-Komlosi, *Thermochim. Acta* **92** (1985) 513 – 516.
104. M. Rolin, *Rev. Int. Hautes Temp. Refract.* **8** (1971) no. 2, 127 – 132. M. Rolin, P. Desclaux, *Rev. Int. Hautes Temp. Refract.* **8** (1971) no. 3, 221 – 225. M. Rolin, *Electrochim. Acta* **17** (1972) no. 12, 2293 – 2307. J. R. Gaspard, J. Goret, M. Ferber, *Light Met. (Warrendale, Pa.)* **1** (1975) 79 – 94. AIME Ny, L. H. Marcus: "Cryolite June 70–June 80," *NTJS from G.O.V. Rep. Annonce Index* **80** (1980) no. 26, 5649.
105. H. Kvande, *High Temp. High Pressures* **15** (1983) no. 1, 63 – 71; **15** (1983) no. 6, 685 – 691.
106. J. Gerlach, U. Hennig, H. D. Poetsch, *Erzmetall* **31** (1978) no. 7 – 8, 333 – 338.
107. N. P. Mokretzki, J. P. Antipina, *Tsvetn. Met. (Moscow)* 1984, no. 5, 51 – 54.
108. A. Seidell: "Solubilities Inorganic and Metal Organic Compounds," Am. Chem. Soc., Washington D.C., vol. 1, 4th ed. (1958) 184.
109. IG Oppau, DE 646 109, 1935 (W. Strathmeyer). BASF, DE 873 690, 1942. R. Vogel, *Chem. Ztg.* **75** (1951) 603.

110. Kakutani, JP 7 432 708, 1974 (M. Kakutani, S. Isobe).
111. S.E.C.E.M. and A.E.U, FR 1 187 352, 1957, and patent addition no. 76 352, 1961 (P. Mollard). Ugine Kuhlmann, FR 1 519 691, 1967 (R. Bachelard). Ugine Kuhlmann, FR 2 076 577, 1970 (P. Mollard).
112. W. Wolfrom, W. Källing, *Tech. Umweltschutz* **25** (1982) 149 – 161.
113. Z. Klimek, *Przem. Chem.* **63** (1984) no. 9, 476 – 478.
114. D. V. Fedorin, A. Samoilova, *Tsvetn. Met.* **1978,** no. 12, 23 – 24. J. Xavier, ID 145 960, 1979.
115. J. Lobo, V. Damodaran, *Res. Ind.* **18** (1973) no. 4, 127 – 129.
116. Derivados del Fluor, FR 2 459 204, 1979 (H. J. Gutierrez).
117. M. Plebankiewicz, Z. Klimek, A. Bernat, *Mater. Ogolnopol. Symp. Zwizki Fluorowe* 1980, 127 – 129.
118. Kaiser Aluminium and Chem. Corp., US 3 061 411, 1960 (D. C. Gernes).
119. W. Augustyn, J. Chmiel-Pela, M. Grobelny, D. Rozycka, *Przem. Chem.* **56** (1977) no. 11, 565 – 566.
120. Ugine Kuhlmann, DE-OS 1 667 720, 1968 (R. Bachelard).
121. Vereinigte Alumium-Werke, US 2 991 159, 1958 (G. Wendt); US 2 943 914, 1960 (E. Moser).
122. Montedison, DE-OS 2 415 471, 1974 (M. Vecchio, P. Andrigo).
123. B. Heydorn, R. Barrere, J. Shimosato in: *CEH (Chemical Economics Handbook)*, S.R.I. International, Sept. 1985.
124. The Economics of Fluorspar (1983) Roskill 164.
125. SRI International: *Directory of Ch*[al] *Products,* 1985, p. 1675.
126. *Registry of Toxic Effects of Chemical Substances,* NIOSH 1 (1981 – 82) 314.
127. Allied Chemical Corp., DE-OS 2 140 602, 1972 (B. M. Lichstein, C. Woolf).
128. General Electric, US 3 357 788, 1967 (J. F. Ross).
129. Kali-Chemie, DE 1 010 504, 1957 (R. Schober).
130. B. Cochet-Muchy, J. Portier in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 565 – 605.
131. Motorola, US 4 235 648, 1980 (L. Richardson).
132. Texas Instruments, US 3 767 582, 1973 (G. A. Miller).
133. R. O. Hull & Co., US 3 296 141, 1967 (W. A. Lieb, E. Billow).
134. Société des Usines Chimiques de Pierrelatte, FR 2 082 366, 1971 (M. Caron, P. Coste, C. Coquet, M. Rey).
135. S. Nagase in P. Tarrant (ed.): *Fluorine Chemistry Reviews,* **vol. 1,** Dekker, New York 1967, pp. 77 – 106.
136. E. M. Genies, C. Tsintavis, *J. Electroanal. Chem. Interfacial Electrochem.* **195** (1985) 109 – 128.
137. O. Glemser in R. E. Banks, D. W. A. Sharp, J. C. Tatlow (eds.): *Fluorine, the first Hundred Years,* Elsevier Sequoia, Lausanne 1986, pp. 45 – 69.
138. M. R. C. Gerstenberger, A. Haas, *Angew. Chem., Int. Ed. Engl.* **20** (1981) 647 – 667.
139. E. I. Du Pont, US 2 670 387, 1954 (H. B. Gottlieb, J. D. Park).
140. A. J. Rudge in A. T. Kuhn (ed.): *Industrial Electrochemical Processes,* Elsevier, Amsterdam 1971, pp. 1 – 69.
141. H. Imoto, K. Ueno, N. Watanabe, *Bull. Chem. Soc. Jpn.* **51** (1978) 2822 – 2825.
142. N. Watanabe, H. Touhara, T. Nakajima, N. Bartlett et al. in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 331 – 369.
143. Union Carbide, US 3 444 058, 1969 (G. W. Mellors, S. Senderoff).
144. K. Schulze, M. Krehl, *Nucl. Instrum. Methods Phys. Res., Sect. A,* **236** (1985) 609 – 616; *Chem. Abstr.* **103** (1985) 94588 t.
145. J. Portier in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 553 – 563.
146. D. T. Meshri in R. E. Banks, D. W. A. Sharp, J. C. Tatlow (eds.): *Fluorine, the First Hundred Years,* Elsevier Sequoia, Lausanne 1986, pp. 195 – 226.
147. The Brush Beryllium Co., US 2 381 291, 1945 (B. R. F. Kjellgren).
148. Eastman Kodak, US 3 294 878, 1966 (E. Carnall Jr., S. E. Hatch, L. S. Ladd, W. F. Parsons).
149. W. M. Phillips, J. W. Stearns, *Proc. Intersoc. Energy Convers. Engl. Conf. 20th,* 1985, vol. 2, p. 2384 – 2391; Chem. Abstr. 103 (1985) 180950 m.
150. Matsushita Electric Industrial Co., JP-Kokai 75 139 395, 1975 (M. Matsuura, N. Nishi, M. Matsuoka, T. Masuyama).

151. Tashkent State University, SU 466 202, 1975 (Kh. U. Usmanov, A. A. Yul'chibaev, T. S. Sirlibaev, A. Akramkhodzhaev).
152. R. D. Peacock in F. A. Cotton (ed.): *Progress in Inorganic Chemistry,* **vol. 2,** Interscience, New York 1960, pp. 193 – 249.
153. W. Klemm, DE 813 847, 1951.
154. L. B. Asprey, *J. Fluorine Chem.* **7** (1976) 359 – 361.
155. Rockwell International Corp., US 4 108 965, 1978 (K. O. Christe).
156. Showa Aluminium Industries K. K., JP-Kokai 60 46 867, 1983.
157. Atlantic Richfield Co., US 3 766 147, 1973 (L. G. Wolgemuth).
158. Daikin Kogyo Co., DE-OS 3 027 229, 1981 (Y. Ohsaka, T. Tohzuku).
159. C. L. Liotta, H. P. Harris, *J. Am. Chem. Soc.* **96** (1974) 2250 – 2252.
160. *Journal Officiel de la République Française,* Nov. 28, 1985.
161. S. L. Kravtsov, N. I. Soboleva, SU 706 325, 1979.
162. Fertilizers and Chemicals Ltd., GB 891 276, 1962 (G. Stern, S. Harel).
163. Sumitomo Electric Industries, JP-Kokai 75 113 449, 1975 (K. Motoyoshi, M. Kume, Y. Amano).
164. Degussa, CH 650 437, 1985 (W. Hasenpusch, K. F. Zimmermann, A. Heilmann).
165. Convertex, US 3 849 208, 1974 (M. N. Marosi).
166. A. B. Vicente, BR 85 01 205, 1985.
167. D. Standfuss, H. Stange, W. Oese, H. Kirk et al., DD 216 893, 1985.
168. H. C. Hodge, F. A. Smith in J. H. Simons (ed.): *Fluorine Chemistry,* vol. 4, Academic Press, New York 1965.
169. S. L. Kravtsov, N. I. Soboleva, SU 767 019, 1980.
170. Tennessee Corp., US 2 602 726, 1952 (E. A. Winter).
171. Hashimoto Chemical Industries Co., JP 71 03 253, 1971 (M. Mizuta, K. Kanai, T. Eguchi).
172. V. E. B. Chemiewerk Coswig, DD 220 587, 1985 (W. Koelling, W. Kerscher, U. Lamm, G. Marotz).
173. T. A. O'Donnel, *Rev. Pure Appl. Chem.* **20** (1970) 159 – 174.
174. G. A. Olah, J. G. Shih, G. K. S. Prakash in R. E. Banks, D. W. A. Sharp, J. C. Tatlow (eds.): *Fluorine, the First Hundred Years,* Elsevier Sequoia, Lausanne 1986, pp. 377 – 396.
175. A. M. Bond, I. Irvine, T. A. O'Donnel, *Inorg. Chem.* **14** (1975) 2408 – 2412.
176. Comurhex, FR 2 574 988, 1986 (M. Bargues, R. Romano).
177. I. B. M. Corp., GB 1 515 763, 1978 (J. J. Cuomo, J. M. Woodall, J. F. Ziegler).
178. E. L. Pace, J. S. Mosser, *J. Chem. Phys.* **39** (1963) no. 1, 134.
179. D. D. Wagman et al., *J. Phys. Chem. Ref. Data Suppl.* **11** (1982) no. 2, 111.
180. W. I. Patnode, J. Papish, *J. Phys. Chem.* **34** (1930) 1494.
181. C. L. Yaws, C. E. Gorin, I. L. Dickens, G. Hsu, R. Lutwack, *J. Chin. Inst. Chem. Eng.* **12** (1981) 33 –44.
182. G. L. Flemmert, DE 1 028 099, 1955.
183. N. Miller, *Nature (London)* **158** (1946) 950.
184. J. W. Hill, R. V. Lindsay, R. H. Wiley, US 2 395 826, 1946.
185. G. S. Niewland, *J. Chem. Soc.* **58** (1936) 897.
186. G. V. Medoks, E. M. Joshestvenskaya, *Zh. Obshch. Khim.* **26** (1956) 116.
187. R. E. Driscoll, G. L. Decuir, DE 2 132 427, 1972.
188. Central Glass Co., GB 2 079 262, 1982 (T. Otsuka, N. Kitsugi, T. Fukinaga).
189. Allied, US 4 470 959, 1984 (J. S. Talwar, J. Hendrickson).
190. R. Caillat, *Ann. Chim. (Paris)* **20** (1945) 367 – 420.
191. Ocrit Fabrik, NL 73 584, 1955; NL 77 764, 1956; NL 86 875, 1957.
192. G. L. Flement, *Proc. Fert. Soc.* **163** (1977) 1 – 23.
193. *Chem. Eng. News* 1984, Jan. 9, 5.
194. *Jpn. Chem. Week* **24** (1983) no. 1220, 1 – 3.
195. *Chem. Mark. Rep.* **225** (1984) no. 1, 3, 57.
196. *Registry of Toxic Effects of Chemical Substances,* NIOSH 3 (1981 – 82) 597.
197. R. Stolba, *J. Proc. Chim.* **90** (1924) 193.
198. C. A. Jacobson, *J. Phys. Chem.* **28** (1924) 506.
199. E. Baur, *Z. Phys. Chem.* **48** (1904) 483 – 503.
200. B. I. Schramban, L. D. Padlukhina, N. D. Afomna, A. P. Gorskaya, *J. Appl. Chem. USSR (Engl. Transl.)* **52** (1979) 2220 – 2222.
201. G. Gambaretto, A. E. Pedler, *Ann. Chim. (Rome)* **64** (1974) 711 – 715.
202. Dow Chemical, US 4 008 130, 1976 (J. F. M. Leathers, D. W. Calvin).
203. G. E. Mapstone, *Chem. Aust.* **47** (1980) no. 1, 30 – 31.
204. D. D. Wagman et al., *J. Phys. Chem. Ref. Data* **11** (1982) no. 2, 111.
205. R. G. Rees, L. J. Hudleston, *J. Chem. Soc.* 1936, 1334.
206. I. G. Ryss, N. P. Bakina, *C. R. Akad. URSS* **2** (1936) 21; *Chem. Abstr.* **30,** 7058/2 30 7058 [2].

207. V. N. Krylov, E. V. Komarov, M. F. Pusklenkov, *Radiokhimiya* **13** (1971) no. 3, 430 – 434.
208. Central Glass Co., JP-Kokai 8 090 413, 1980 (Y. Kobayashi, Y. Nakaso, T. Nakamura).
209. BASF, DE 2 904 547, 1979 (A. Brunold, P. Kroetzsch, L. Diehl).
210. Swift and Co., US 3 273 713, 1969 (W. R. Parish).
211. Whiting Corp., US 2 926 690, 1965.
212. J. L. Medbery, *Proc. Ann. Meet. Fert. Ind. Round Table* **28** (1978) 109 – 125.
213. Boliden AB, SE 357 349, 1973 (G. B. Lagerstrom).
214. L. A. Mitchell Ltd., DE 1 767 323, 1968.
215. S.E.C.E.M. and A.E.U., FR 1 474 097, 1967 (P. Mollard).
216. J. Y. Morlock, B. Nocher, P. Mollard, *2nd Int. Congress on Phosphorous Compounds Proceedings,* 21. – 25. April 1980, Boston (USA), edit Imphos, p. 79.
217. V. Funasaka, M. Kowano, *Kogyo Kagaku Zasshi* **53** (1950) 326 – 328.
218. S. M. Thomsen, *Anal. Chem.* **23** (1951) 973 – 975.
219. J. C. Warf, W. D. Cline, R. D. Tevebaugh, *Anal. Chem.* **26** (1954) 342.
220. A. Laszlo, *Aluminium (Budapest)* **2** (1950) 7 – 8.
221. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH Product Review (Chemical Economics Handbook),* SRI International, August 1982, no. 739 – 1002 J – K.
222. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH Product Review (Chemical Economics Handbook),* SRI International, August 1982, no. 739 – 1001 P – S.
223. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH Product Review (Chemical Economics Handbook),* SRI International, August 1982, no. 739 – 1002 P.
224. *Registry of Toxic Effects of Chemical Substances,* NIOSH 3 (1981 – 82) 595.
225. W. Linke: *Solubilities Inorganic and Metal Organic Compounds,* 2nd ed., Am. Chem. Soc., Washington, D.C., 1958.
226. J. G. Ryss, N. M. Slutskaya, *Zh. Fiz. Khim.* **14** (1940) 701 – 707.
227. H. Bardt, H. Bendheim, DE 610 596, 1935.
228. R. Caillat, *Ann. Chim.* **20** (1945) 367 – 420.
229. N.A.S.A., Contract no. NAS 7.100; N.A.S.A., Tech. Brief **7** (1984) no. 4, from JPL Invention Report NPO 15721/SC – 1265.
230. Central Glass K. K., BE 863 501, 1977 (Y. Shiraki, H. Haraoka, H. Arai).
231. Gofman'ms, SU 648 516, 1976 (M. S. Gofman, E. M. Morgunova, I. G. Blyakher).
232. Smith Douglas CY Inc., US 2 556 064, 1952 (P. Cadwell, D. C. Kieffer).
233. A. E. Zolotarev, SU 823 280, 1976.
234. Henkel, DE 2 904 876, 1979 (E. Monterogar, D. Jahr, A. Rosas).
235. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH Product Review (Chemical Economics Handbook),* S.R.I. International, August 1982, no. 739 1002 j.
236. Elec Household Mach, SU 855 074, 1978 (M. T. Zagural'skıi, M. I. Kovner, V. Y. Iordanov).
237. Nippon Paint KK, JP 77 005 340, 1977.
238. K. Abe, JP 51 145 121, 1975.
239. Perm Cell Paper Ind., SU 983 162 A, 1981.
240. Rhône Poulenc Ind., DE 2 704 649, 1977 (C. Fouassier, B. Latourette).
241. F. A. Simonelli, US 3 899 616, 1973.
242. N. M. Levenson: "Fluorine Minerals and Inorganic Fluorine Compounds," in *CEH Products Review* (Chemical Economics Handbook), S.R.I. International, August 1982, no. 739 1002 P.
243. *Registry of Toxic Effects of Chemical Substances,* NIOSH 3 (1981 – 82) 595.
244. G. H. Cady in H. J. Emeléus, A. G. Sharpe (eds.): *Advances Inorg. Chem. Radiochem.,* **vol. 2,** Academic Press, New York 1960, pp. 105 – 157.
245. "Schwefel Halogenide," *Gmelin,* **9,**Suppl. 2, 5 – 201.
246. M. Centnerszwer, C. Strenk, *Ber. Dtsch. Chem. Ges.* **56** (1923) 2249 – 2253.
247. R. L. Kuczkowski, *J. Am. Chem. Soc.* **86** (1964) 3617 – 3621.
248. F. Seel in H. J. Emeléus, A. G. Sharpe (eds.): *Advances Inorg. Chem. Radiochem.,* **vol. 16,** Academic Press, New York 1974, pp. 297 – 333.
249. "Schwefel – Thionyl Halogenide," *Gmelin,* **9,** Suppl. 1, 3 – 18.
250. *Gmelin,* **B 3,** 1729 – 1733.
251. Dow Chemical, US 2 875 127, 1959 (E. E. Kenaga).
252. B. Cochet-Muchy, J. Portier in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 565 – 605.
253. Asahi Glass, GB 2 081 694, 1982 (Y. Oda, S. Morikawa, M. Ikemura, T. Yarita et al.).

254. Du Pont, US 2 937 123, 1960 (E. L. Muetterties).
255. Minnesota Mining and Manufacturing Co., US 3 345 277, 1967 (P. E. Ashley, J. D. La Zerte).
256. W. C. Schumb, *Ind. Eng. Chem.* **39** (1947) 421 – 423.
257. Kali-Chemie, DE-OS 2 407 492, 1975 (J. Massonne, W. Becher).
258. Montedison S.p.A., DE-OS 2 816 693, 1978 (A. Di Gioacchino, G. Tommasi, M. De Manuele).
259. Produits Chimiques Ugine Kuhlmann, EP-A 87 338, 1983 (M. Jaccaud, A. J. F. Ducouret).
260. H. Moissan, P. Lebeau, *C.R. Hebd. Séances Acad. Sci.* **130** (1900) 865 – 871.
261. F. Lombardo, *J. Fluorine Chem.* **18** (1981) no. 1, 1 – 24.
262. Hitachi Ltd., US 4 330 384, 1982 (S. Okudaira, K. Suzuki, S. Nishimatsu, I. Kanomata).
263. Allis-Chalmers Manufacturing Co., US 3 740 256, 1973 (J. P. Manion, D. J. Davies).
264. R. S. Busk, R. B. Jackson, *Proc. Annu. World Conf. Magnesium 37th,* (1980) pp. 1 – 4; Chem. Abstr. 97 (1982) 148943 h.
265. IEC Recommendation 376, DIN IEC 376 and VDE 0373,Part 1/4.80.
266. IEC Recommendation 376, 1st ed., Sect. 3 or DIN IEC 376, Chapter 3, and VDE 0373, Part 1, Chapter 3.
267. S. M. Williamson in F. A. Cotton (ed.): *Prog. Inorg. Chem.,* **vol. 7,** Interscience, New York 1966, pp. 39 – 81.
268. F. Brown, P. L. Robinson, *J. Chem. Soc.* 1955, 3147 – 3151.
269. Air Products and Chemicals Inc., US 3 399 036, 1968 (S. Kleinberg, J. F. Tompkins Jr.).
270. Asahi Glass Co., JP-Kokai 81 88 808, 1981.
271. Hoechst, DE-OS 2 925 540, 1981 (R. Franz).
272. R. Appel, DE-OS 2 363 679, 1975.
273. Commissariat à l'Energie Atomique, FR 1 586 833, 1970 (J. Aubert, B. Cochet-Muchy, J. P. Cuer).
274. G. A. Boswell, Jr., W. C. Ripka, R. M. Scribner, C. W. Tullock in W. G. Dauben (ed.): *Organic Reactions,* **vol. 21,** J. Wiley & Sons, New York 1974, pp. 1 – 124.
275. Sulfur Tetrafluoride Technical, New Product Information Bulletin 2 b, Du Pont 1961.
276. C. B. Colburn in M. Stacey, J. C. Tatlow, A. G. Sharpe (eds.): *Advances in Fluorine Chem.,* **vol. 3,** Butterworths, London 1963, pp. 92 – 116.
277. Air Products and Chemicals Inc., US 4 108 966, 1968 (A. J. Woytek, J. T. Lileck).
278. Air Products and Chemicals Inc., JP 8 008 926, 1980 (A. J. Woytek, J. T. Lileck).
279. Air Products and Chemicals Inc., US 3 235 474, 1966 (J. F. Tompkins Jr., E. S. J. Wang).
280. S. Nagase in P. Tarrant (ed.): *Fluorine Chemistry Reviews,* **vol. 1,** Dekker, New York 1967, pp. 77 – 106.
281. Handotai Energy Kenkyusho K. K., JP-Kokai 8 249 234, 1982 (Yamazaki, Shunpei).
282. *Chem. Eng. News* **43** (1965) no. 6, 47.
283. Stanford Research Institute, US 3 708 570, 1973 (W. E. Tolberg, R. S. Stringham, R. T. Rewick).
284. R. Bougon, T. Bui Huy, J. Burgess, K. O. Christe et al., *J. Fluorine Chem.* **19** (1982) 263 – 277.
285. A. G. Streng, *Chem. Rev.* **63** (1963) 607 – 624.
286. H. R. Leech in: *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol. 2, Suppl. 1, Longmans, London 1956, pp. 186 – 196.
287. P. Lebeau, A. Damiens, *C.R. Hebd. Séances Acad. Sci.* **185** (1927) 652 – 654.
288. A. Engelbrecht, E. Nachbaur, *Monatsh. Chem.* **90** (1959) 367 – 370.
289. P. Lebeau, A. Damiens, *C.R. Hebd. Séances Acad. Sci.* **188** (1929) 1253 – 1255.
290. M. Lustig, J. M. Shreeve in J. C. Tatlow, R. D. Peacock, H. H. Hyman (eds.): *Advances in Fluorine Chem.,* **vol. 7,** Butterworths, London 1973, pp. 175 – 198.
291. R. Schmutzler in M. Stacey, J. C. Tatlow, A. G. Sharpe (eds.): *Advances in Fluorine Chem.,* **vol. 5,** Butterworths, London 1965, pp. 31 – 285.
292. L. Kolditz in H. J. Emeléus, A. G. Sharpe (eds.): *Advances Inorg. Chem. Radiochem.,* **vol. 7,** Academic Press, New York 1965, pp. 1 – 26.
293. A. B. Burg in J. H. Simons (ed.): *Fluorine Chemistry,* **vol. 1,** Academic Press, New York 1950, pp. 77 – 123.
294. J. R. Van Wazer: *Phosphorus and its Compounds,* **vol. 1,** Interscience, New York 1958, p. 220.
295. Standard Oil, US 2 819 328, 1958 (H. C. Brown, W. S. Higley).
296. W. Lange in J. H. Simons (ed.): *Fluorine Chemistry,* **vol. 1,** Academic Press, New York 1950, pp. 125 – 188.

297. L. Cot, *Inf. Chim.* **189** (1979) 157 – 164.
298. Dow Chemical, US 2 846 412, 1958 (C. B. Havens).
299. H. S. Booth, J. T. Pinkston Jr. in J. H. Simons (ed.): *Fluorine Chemistry,* **vol. 1,** Academic Press, New York 1950, pp. 189 – 200.
300. W. K. R. Musgrave in M. Stacey, J. C. Tatlow, A. G. Sharpe (eds.): *Advances in Fluorine Chem.,* **vol. 1,** Butterworths, London 1960, pp. 1 – 28.
301. A. J. Downs, C. J. Adams in J. C. Bailar, H. J. Emeléus, R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorg. Chem.,* **vol. 2,** Pergamon, Oxford 1973, pp. 1107 – 1594.
302. H. Meinert, *Z. Chem.* **7** (1967) 41 – 57.
303. Yu. A. Buslaev, V. F. Sukhoverkhov, N. M. Klimenko, *Koord. Khim.* **9** (1983) 1011 – 1031; *Chem. Abstr.* **99** (1983) 110889 y.
304. W. B. Deichmann, H. W. Gerarde: *Toxicology of Drugs and Chemicals,* Academic Press, New York 1969, p. 651.
305. *Gas Encyclopedia,* Elsevier/l'Air Liquide, Amsterdam 1976, p. 849.
306. Allied Corp., US 3 367 745, 1968 (H. G. Tepp).
307. A. Commeyras, *Ann. Chim. (France)* **9** (1984) 673 – 682.
308. R. N. Haszeldine in R. E. Banks, D. W. A. Sharp, J. C. Tatlow (eds.): *Fluorine, the First Hundred Years,* Elsevier Sequoia, Lausanne 1986, pp. 307 – 313.
309. Novosibirsk State University, SU 1 051 047, 1983 (Yu. I. Nikonorov).
310. H. Selig, A. Pron, M. A. Druy, A. G. MacDiarmid et al., *J. Chem. Soc. Chem. Commun.* 1981, 1288 – 1289.
311. J. Terrel, D. Pratt, L. C. Sams, L. Robinson, *Oil Gas J.* **79** (1981) 108 – 118.
312. W. G. Sweetman, DE 1 029 770, 1958.
313. M. Schmeisser, K. Brändle in H. J. Emeléus, A. G. Sharpe (eds.): *Advances Inorg. Chem. Radiochem.,* **vol. 5,** Academic Press, New York 1963, pp. 41 – 89.
314. K. O. Christe, C. J. Schack in H. J. Emeléus, A. G. Sharpe (eds.): *Advances Inorg. Chem. Radiochem.,* **vol. 18,** Academic Press, New York 1976, pp. 319 – 398.
315. N. Bartlett, F. O. Sladky in J. C. Bailar, H. J. Emeléus, R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorg. Chem.,* **vol. 1,** Pergamon Press, Oxford 1973, pp. 213 – 330.
316. F. Sladky in V. Gutmann (ed.): *Inorg. Chem.,* **vol. 3,** Butterworths, London 1972, pp. 1 – 52.
317. K. Seppelt, D. Lentz in S. J. Lippard (ed.): *Prog. Inorg. Chem.,* **vol. 29,** Interscience, New York 1982, pp. 167 – 202.
318. H. Selig, J. H. Holloway in F. L. Boschke (ed.): *Topics in Current Chem.,* **vol. 124,** Springer Verlag, Berlin 1984, pp. 33 – 90.
319. J. H. Holloway in R. E. Banks, D. W. A. Sharp, J. C. Tatlow (eds.): *Fluorine, the First Hundred Years,* Elsevier Sequoia, Lausanne 1986, pp. 149 – 158.
320. N. Bartlett, D. H. Lohmann, *Proc. Chem. Soc. London* 1962, 115 – 116.
321. N. Bartlett, *Proc. Chem. Soc. London* 1962, 218.
322. J. Grannec, L. Lozano in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 17 – 76.
323. Messer Griesheim, DE-OS 2 506 475, 1976 (C. Goekcek).
324. R. Filler, *Isr. J. Chem.* **17** (1978) 71 – 79; *Chem. Abstr.* **89** (1978) 89755 h.
325. D. W. Setser, H. C. Brashears, T. D. Dreiling, *J. Phys. Colloq. (Orsay, Fr.)* 1980, no. C 3, 195 –202.
326. P. Kamarchik Jr., J. L. Margrave, *Acc. Chem. Res.* **11** (1978) 296 – 300.
327. N. Watanabe, T. Nakajima, H. Touhara: *Studies in Inorg. Chem.,* vol. 8 (Graphite Fluorides), Elsevier, Amsterdam 1988.
328. N. Watanabe, H. Touhara, T. Nakajima, N. Bartlett et al. in P. Hagenmuller (ed.): *Inorganic Solid Fluorides – Chemistry and Physics,* Academic Press, Orlando 1985, pp. 331 – 369.
329. N. Watanabe, T. Nakajima in R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Chapter 9, Ellis Horwood, Chichester (England) 1982.
330. H. Gisser, M. Petronio, A. Shapiro, *Lubr. Eng.* **28** (1972) 161 – 164; *Chem. Abstr.* **77** (1982) 37289 s.
331. H. E. Sliney, *Tribol. Int.* **15** (1982) no. 5, 303 – 315; *Chem. Abstr.* **98** (1983) 74913 v.
332. Matsushita Electric Industrial Co., DE-OS 1 917 907, 1969 (N. Watanabe, M. Fukuda).
333. Matsushita Electric Industrial Co., DE-OS 1 919 394, 1969 (N. Watanabe, M. Fukuda).
334. T. Nakajima, M. Kawaguchi, N. Watanabe, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **36 B** (1981) 1419 – 1423.

335. Union Carbide, US 3 956 018, 1974 (A. Kozawa).
336. Central Glass Co., DE-OS 3 235 596, 1982 (N. Watanabe, T. Nakajima, M. Masayuki).
337. N. Watanabe, T. Nakajima, M. Kawaguchi, DE-OS 3 330 695, 1984.
338. Kirk-Othmer, Vol. 2, p. 265, 1992
339. Kirk-Othmer, Vol. 11, p. 433, 1992

Fluorine Compounds, Organic

Fluoropolymers, Organic is a separate keyword

GÜNTER SIEGEMUND, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chaps. 1–4, 10, 11, and 12)

WERNER SCHWERTFEGER, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chaps. 3 and 4)

ANDREW FEIRING, E. I. DuPont de Nemours & Co., Wilmington, Delaware, United States (Chaps. 5–7)

BRUCE SMART, E. I. DuPont de Nemours & Co., Wilmington, Delaware, United States (Chaps. 5–7)

FRED BEHR, Minnesota Mining and Manufacturing Company, St. Paul, Minnesota, United States (Chap. 8)

HERWARD VOGEL, Minnesota Mining and Manufacturing Company, St. Paul, Minnesota, United States (Chap. 9)

BLAINE MCKUSICK, E. I. DuPont de Nemours & Co., Wilmington, Delaware, United States (Chap. 13)

1. Introduction . . . 3
2. Production Processes . . . 4
2.1. Substitution of Hydrogen . . . 4
2.2. Halogen – Fluorine Exchange . . . 5
2.3. Synthesis from Fluorinated Synthons . . . 6
2.4. Addition of Hydrogen Fluoride to Unsaturated Bonds . . . 6
2.5. Miscellaneous Methods . . . 6
2.6. Purification and Analysis . . . 6
3. Fluorinated Alkanes . . . 7
3.1. Fluoroalkanes and Perfluoroalkanes . . . 7
3.2. Chlorofluoroalkanes . . . 12
3.3. Bromofluoroalkanes . . . 16
3.4. Iodofluoroalkanes . . . 17
4. Fluorinated Olefins . . . 18
4.1. Tetrafluoroethylene . . . 19
4.2. Hexafluoropropene . . . 20
4.3. 1,1-Difluoroethylene . . . 21
4.4. Monofluoroethylene, Monofluoroethylene . . . 22
4.5. 3,3,3-Trifluoropropene . . . 22
4.6. 3,3,3-Trifluoro-2-(trifluoromethyl)-prop-1-ene . . . 22
4.7. Chlorofluoroolefins . . . 23
5. Fluorinated Alcohols . . . 23
6. Fluorinated Ethers . . . 24
6.1. Perfluoroethers . . . 24
6.1.1. Low Molecular Mass Perfluoroethers 24
6.1.2. Perfluorinated Epoxides . . . 24
6.1.3. High Molecular Mass Perfluoroethers . . . 25
6.2. Perfluorovinyl Ethers . . . 25
6.3. Partially Fluorinated Ethers . . . 26
7. Fluorinated Ketones and Aldehydes . . . 27
7.1. Fluoro- and Chlorofluoroacetones 27
7.2. Perhaloacetaldehydes . . . 28
7.3. Fluorinated 1,3-Diketones . . . 29
8. Fluorinated Carboxylic Acids and Fluorinated Alkanesulfonic Acids 31
8.1. Fluorinated Carboxylic Acids . . . 31
8.1.1. FluorinatedAcetic Acids . . . 31
8.1.2. Long-Chain Perfluorocarboxylic Acids . . . 31
8.1.3. Fluorinated Dicarboxylic Acids . . . 32
8.1.4. Tetrafluoroethylene – Perfluorovinyl Ether Copolymers with Carboxylic Acid Groups . . . 33
8.2. Fluorinated Alkanesulfonic Acids 33
8.2.1. Perfluoroalkanesulfonic Acids . . . 33
8.2.2. Fluorinated Alkanedisulfonic Acids 34
8.2.3. Tetrafluoroethylene – Perfluorovinyl Ether Copolymers with Sulfonic Acid Groups . . . 34
9. Fluorinated Tertiary Amines . . . 35
10. Aromatic Compounds with Fluorinated Side-Chains . . . 35
10.1. Properties . . . 36
10.2. Production . . . 36
10.3. Uses . . . 37
11. Ring-Fluorinated Aromatic, Heterocyclic, and Polycyclic Compounds . . . 38
11.1. Mono- and Difluoroaromatic Compounds . . . 38
11.1.1. Properties . . . 39

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_349

11.1.2. Production 39
11.1.3. Uses 41
11.2. Highly Fluorinated Aromatic Compounds 42
11.3. Perhaloaromatic Compounds . . . 43
11.4. Fluorinated Heterocyclic and Polycyclic Compounds 44
11.4.1. Ring-Fluorinated Pyridines 44
11.4.2. Trifluoromethylpyridines 44
11.4.3. Fluoropyrimidines 44
11.4.4. Fluorotriazines 44
11.4.5. Polycyclic Fluoroaromatic Compounds 45
12. Economic Aspects 45
13. Toxicology and Occupational Health 45
13.1. Fluorinated Alkanes 46
13.2. Fluorinated Olefins 47
13.3. Fluorinated Alcohols 47
13.4. Fluorinated Ketones 47
13.5. Fluorinated Carboxylic Acids . . . 47
13.6. Other Classes 47
14. References 48

1. Introduction

Organic fluorine compounds are characterized by their carbon – fluorine bond. Fluorine can replace any hydrogen atom in linear or cyclic organic molecules because the difference between the van der Waals radii for hydrogen (0.12 nm) and fluorine (0.14 nm) is small compared to that of other elements (e.g., chlorine 0.18 nm). Thus, as in hydrocarbon chemistry, organic fluorine chemistry deals with a great variety of species. When all valences of a carbon chain are satisfied by fluorine, the zig-zag-shaped carbon skeleton is twisted out of its plane in the form of a helix. This situation allows the electronegative fluorine substituents to envelop the carbon skeleton completely and shield it from chemical (especially nucleophilic) attack. Several other properties of the carbon – fluorine bond contribute to the fact that highly fluorinated alkanes are the most stable organic compounds. These include low polarizability and high bond energies, which increase with increasing substitution by fluorine (bond energies: C – F bond in CH_3F, 448 kJ/mol; C – H bond in CH_4, 417 kJ/mol; C – Cl bond in CH_3Cl, 326 kJ/mol; and C – F bond in CF_4, 486 kJ/mol).

The cumulative negative inductive effect of the fluorine in perfluoroalkyl groups may reverse the polarity of adjacent single bonds (e.g., in the pair $H_3C \lhd I$ and $F_3C \rhd I$) or double bonds (e.g., $CH_3C^{\delta+}H{=}C^{\delta-}H_2$ and $CF_3{-}C^{\delta-}H{=}C^{\delta+}H_2$). Fluorine substitution changes the reactivity of olefins and carbonyl compounds. Polyfluorinated olefins possess an electron-deficient double bond, which reacts preferentially with nucleophiles. Carboxy groups are affected by the presence of an adjacent perfluoroalkyl radical. In carboxylic acids, the acidity is markedly increased. In other carbonyl compounds, the reactivity is increased without any fundamental change in the chemistry of the compound. Correspondingly, the basicity of amines is reduced by the introduction of fluorine. Fluorine attached to the ring of aromatic compounds acts mainly as a para-directing substituent, whereas perfluoroalkyl groups behave as meta-directing substituents.

Naturally, the influence of fluorine is greatest in highly fluorinated and perfluorinated compounds. The fact that these compounds have a high thermal stability and chemical resistance and are physiologically inert makes them suitable for many applications for which hydrocarbons are not. Properties that are exploited commercially include high thermal and chemical stability, low surface tension, and good dielectric properties, for example, in fluoropolymers, perfluorinated oils and inert fluids.

Individual fluorine atoms or perfluoroalkyl groups do not change the technical properties of a hydrocarbon fundamentally. However, this is not the case with physiological properties. A fluorine atom in a bioactive material may simulate a hydrogen atom, and although this does not prevent metabolic processes from occurring, the end products may be ineffective or toxic. Accordingly, such fluorine compounds are important in, for example, pesticides and pharmaceuticals.

A bibliography of the scientific literature of organofluorine chemistry was published in 1986 [16]; commercial applications of fluorine products are reviewed in [7], [17], and [18].

Table 1. Nomenclature of organic fluorine compounds

Formula	CAS registry no.	IUPAC designation	Perfluoro designation
CF_3CF_3	[*76-16-4*]	hexafluoroethane	perfluoroethane, F-ethane
$CF_3CF_2CF_2CHO$	[*375-02-0*]	Heptafluoro-*n*-butyraldehyde	Perfluoro-*n*-butyraldehyde, F-*n*-butyraldehyde
$CF_3(CF_2)_6COOH$	[*335-67-1*]	Pentadecafluoro-*n*-octanoic acid	Perfluoro-*n*-octanoic acid, F-*n*-octanoic acid
$CF_3(CF_2)_2CHF_2$	[*375-17-7*]	1,1,1,2,2,3,3,4,4-Nonafluoro-*n*-butane	1*H*-Perfluoro-*n*-butane, 1-hydryl-F-*n*-butane
$CF_3(CF_2)_4CH_2OH$	[*423-46-1*]	2,2,3,3,4,4,5,5,6,6,6-Undecafluoro-*n*-hexanol	1*H*,1*H*-Perfluoro-*n*-hexanol, 1,1-dihydroperfluoro-*n*-hexanol

Nomenclature. Any organic fluorine compound can be named according to the rules of the International Union for Pure and Applied Chemistry (IUPAC) [19]. However, for highly fluorinated molecules with several carbon atoms, this nomenclature can be confusing. Therefore, the term "perfluoro" may be used when all hydrogen atoms bonded to the carbon skeleton have been replaced by fluorine. The designation of hydrogen atoms belonging to functional groups (e.g., CHO or COOH), of the functional groups themselves, and of other substituents is not affected [19]. Examples are given in Table 1.

In the case of highly fluorinated compounds with few hydrogen atoms (1 – 4), the perfluoro compound can be taken as the parent compound. The hydrogen atoms are named according to their number and position; the letter H or the prefix hydryl (hydro) are used for hydrogen. The symbol F was approved by the American Chemical Society as abbreviation for perfluoro [20].

Historical Development. The pioneering work in organofluorine chemistry dates from 1835 to 1940 [21]. Controlled production of organic fluorine compounds was started in 1892 by exchanging halogen for fluorine in hydrocarbons, using antimony(III) fluoride. The industrial phase began in 1929 in the United States with the discovery of the nonflammable, nontoxic refrigerants CCl_3F and CCl_2F_2 [22]. In Germany, commercial production of aromatic fluorine compounds started in 1930.

The first fluoropolymer, polychlorotrifluoroethylene, was synthesized in 1934 in Germany, followed by the discovery of polytetrafluoroethylene in 1938 in the United States. During World War II, thermally and chemically stable working materials for the separation of uranium isotopes were investigated by the United States Manhattan Project [23]. After World War II, numerous novel applications were discovered. The development of new organic fluorine compounds with novel applications continues undiminished.

2. Production Processes

The four principal methods for the preparation of organic fluorine compounds are as follows [1], [2], [24], [25]:

1) substitution of hydrogen in hydrocarbons using fluorine, high-valency metal or nonmetal fluorides, or electrochemical fluorination
2) halogen – fluorine exchange with hydrogen fluoride, hydrogen fluoride-base complexes, or metal fluorides
3) synthesis of higher molecular mass fluorine compounds from reactive fluorinated synthons
4) addition of fluorine, hydrogen fluoride, or reactive nonmetal fluorides to unsaturated bonds

Only a few of the many possibilities in each group have been developed commercially, with varying degrees of success.

2.1. Substitution of Hydrogen

Fluorination with Elemental Fluorine [26], [27]. The action of elemental fluorine on organic compounds normally leads to violent, mainly explosive, reactions. The substrate fragments into units with a varying degree of fluorination because the heats of formation of the

C – F bond (ca. 460 kJ/mol) and the H – F bond (566 kJ/mol) are greater than the heat of formation of the C – C bond (ca. 348 kJ/mol).

Therefore, direct fluorinations must take place with strict control of the reaction and removal of the heat generated. This may be achieved by dilution of the fluorine with inert gases (e.g., N_2 or CO_2), dilution of the organic substrates with inert solvents [28], intensive mixing, and reduction of the temperature to as low as − 150 °C.

Direct fluorination can also be carried out in the gas phase in a tubular reactor packed with silver- or gold-plated copper turnings [29]. Specialized methods are based on LaMar fluorination [26], aerosol fluorination [30], porous-tube fluorination [31], and jet fluorination [32]; high product selectivities are achieved at a laboratory scale. Commercial operation remains to be developed.

Fluorination with Metal Fluorides [33]. Metal fluorides that can transfer fluorine to organic substrates by changing the oxidative state of the metal, such as cobalt(III) fluoride (CoF_3) and silver(II) fluoride (AgF_2), serve as fluorinating agents in an oxidizing fluorination. The spent metal fluoride is regenerated with elemental fluorine.

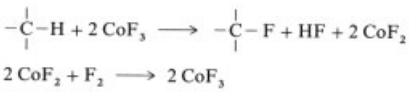

Fluorination and regeneration can be cyclical, permitting a commercial operation.

Electrochemical Fluorination. The *Simons process* [34], [35] is used commercially for the production of perfluorinated compounds. Solutions of organic compounds (mainly carboxylic acids, sulfonic acids, and tertiary amines) are electrolyzed in anhydrous hydrogen fluoride in a single cell without intermediate formation of free fluorine. Fluorination takes place at a nickel anode by a free-radical mechanism at current densities of 10 – 20 mA/cm^2 [36]. Selectivity decreases sharply as the number of carbons increases.

Volatile, hydrogen-containing compounds (hydrocarbons and chlorohydrocarbons) can be electrofluorinated on porous graphite anodes in KF-containing hydrogen fluoride in a process developed by Phillips Petroleum [37], [38], [39] and now referred to as CAVE (Carbon Anode Vapor Phase Electrochemical Fluorination), in operation at 3 M [40]. The organic compound is introduced into the cell through the anode. In the pores of the anode, i.e., at the phase boundary, partial or complete exchange of the hydrogen, but not the chlorine, takes place. To date this process has been used only on a small scale.

2.2. Halogen – Fluorine Exchange

Exchange of chlorine with hydrogen fluoride is used in many commercial processes both for the production of chlorofluoroalkanes and for the side-chain fluorination of aromatic and *N*-heterocyclic compounds [41]. The method involves exchange of chlorine for fluorine in polyhalogenated compounds using hydrogen fluoride [42]:

Whether the process takes place with or without a catalyst depends on the reactivity of the chlorine atoms to be exchanged. With compounds containing several chlorine atoms of differing reactivity, selective fluorination can be achieved by selecting suitable process conditions.

Fluorinations without a catalyst such as

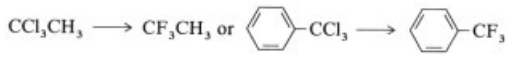

are carried out in liquid, anhydrous hydrogen fluoride at 100 – 150 °C in pressure vessels made of steel, alloy steels, or nickel. The process can be carried out by batch (e.g., autoclave) or continuous methods (autoclaves in series or a tubular reactor). In either case, the hydrogen chloride that is generated during fluorination is removed from the reactor to maintain the desired pressure.

Most liquid-phase fluorinations, e.g., of CCl_4, $CHCl_3$, CCl_3CCl_3, CCl_3CHCl_2, or CCl_3CH_2Cl, are carried out in the presence of a catalyst [43] to promote the exchange, which becomes increasingly difficult as fluorination progresses. The main catalysts used are antimony(III) and antimony(V) halides with low

volatility. Addition of chlorine oxidizes the antimony to the pentavalent state.

In addition to the liquid-phase processes, many commercially important gas-phase fluorinations employ hydrogen fluoride [43]. The components in the gas phase are passed through a tubular reactor containing the catalyst. The composition of the product can be controlled within wide limits by varying temperature, pressure, residence time, catalyst, and the proportions of the reactants. Various metal fluorides are suitable catalysts, e.g., aluminum fluoride [44] or basic chromium fluoride [45].

For further processing of the mixture produced by gas- or liquid-phase fluorination, the following criteria should be satisfied [46]:

1) The hydrogen chloride generated should be separated in a pure form to permit further use.
2) Unreacted hydrogen fluoride should be recovered.
3) Acid residues, water, and other impurities must be removed from the product.

Usually, the hydrogen chloride is separated from the crude fluorination mixture by fractional distillation. The bulk of the hydrogen fluoride may then be separated from the residue. Further treatment includes washing to remove traces of acid, drying, and fractional distillation.

Exchange of Chlorine with Nonoxidizing Metal Fluorides [41]. Alkali fluorides, especially potassium fluoride, are often used to exchange chlorine in carboxylic acid chlorides, sulfonic acid chlorides, α-chlorocarboxylic acid derivatives (esters, amides, and nitriles), aliphatic monochloro compounds, or activated aromatic chloro compounds (Halex process) [47].

The dry, finely powdered metal fluoride is employed in a solvent-free process at 400–600 °C, e.g., with polychlorinated aromatic compounds, or, in most other cases, in the presence of a solvent. For slow reactions, polar, aprotic solvents are used.

2.3. Synthesis from Fluorinated Synthons

The variety of organic fluorine compounds can be greatly increased by the use of easily accessible, low molecular mass fluoroalkanes and olefins to synthesize higher molecular mass products. Halofluoromethanes add to halogenated ethylenes to form halogenated fluoropropanes [48]. An industrially applied reaction is the addition of iodopentafluoroethane to tetrafluoroethylene yielding a homologous series of long-chain 1-iodoperfluoroalkanes (see Section 3.4). The pyrolysis of chlorodifluoromethane is the industrial source of tetrafluoroethylene, hexafluoropropene, and the corresponding oligomers and polymers (see Sections 4.1 and 4.2). These examples illustrate the importance of this synthetic method, especially for the production of organic fluorine compounds containing more than two carbon atoms where the above-mentioned fluorination methods fail to give high yields of the desired products.

2.4. Addition of Hydrogen Fluoride to Unsaturated Bonds

Addition of hydrogen fluoride to alkenes and alkynes takes place below 0 °C with formation of mono- or difluoroalkanes; ethylene and acetylene are exceptions [24]. Ethyl fluoride is produced from ethylene and hydrogen fluoride at 90 °C; catalytic processes have been developed for the addition of hydrogen fluoride to acetylene to produce vinyl fluoride or 1,1-difluoroethane. Unsymmetrical olefins obey Markovnikov's rule. Chloroolefins can undergo chlorine – fluorine exchange after addition of hydrogen fluoride. Addition of hydrogen fluoride to the electron-deficient double bond of perfluoroolefins can be performed using trialkylamine trishydrofluorides at moderate temperatures [49].

2.5. Miscellaneous Methods

Substitution of Amino Groups in Aromatic Compounds [50]. Introduction of one or two fluorine atoms into aromatic rings is carried out commercially by diazotization of aromatic amines in anhydrous hydrogen fluoride with solid sodium nitrite and decomposition of the dissolved diazonium salt (see Section 11.1.2).

Fluorination with Nonmetal Fluorides. Reactions of nonmetal fluorides with certain substrates are predominantly restricted to laboratory operations. Sulfur tetrafluoride and the following compounds can be used for the controlled introduction of fluorine into organic compounds: dialkylaminosulfur(IV) fluorides (R_2NSF_3) [51–53], fluoroalkylamines (e.g., 2-chloro-1,1,2-trifluoroethyldiethylamine or 1,1,2,3,3,3-hexafluoropropyldiethylamine), tetra-*n*-butylammonium fluoride, trialkylamine trishydrofluorides, nitrosyl fluoride, perchloryl fluoride, fluoroxyfluoroalkanes (e.g., CF_3OF) [54], xenon difluoride [55], or CH_3COOF [56]. They are of commercial value for the fluorination of complex organic compounds such as pharmaceuticals.

2.6. Purification and Analysis

Impurities are usually removed from organic fluorine compounds by fractional distillation, fractional crystallization, or chromatographic methods. This does not apply to fluoropolymers and high-boiling perfluorinated oils, which require special measures, i.e., the use of extremely pure starting materials.

Quantitative determination of fluorine is possible in most cases by combustion and subsequent analysis of the hydrogen fluoride generated. Wet chemical methods are used to determine fluoride ions [57].

Because of the high volatility of organic fluorine compounds, purity can be readily determined by gas chromatography. ^{19}F-Nuclear magnetic resonance spectroscopy is a valuable tool for determining the structure of organic fluorine compounds. Structure determinations, even of mixtures, are often easier using this method than with ^{1}H-NMR spectra due to the larger chemical shifts of ^{19}F-NMR spectra. The ^{19}F signals can be integrated and used for quantitative analysis [58].

3. Fluorinated Alkanes

The hydrogen atoms of alkanes may be partially or totally replaced by fluorine. Partially fluorinated alkanes are hydrofluorocarbons (HFCs); fully fluorinated alkanes (perfluoroalkanes) are perfluorocarbons (PFCs). In chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HCFCs), the alkane hydrogens are replaced by both chlorine and fluorine.

A special nomenclature [59] has been introduced to identify smaller chain length fluoroalkanes (up to four carbon atoms) used in refrigerants. It consists of a three-digit number combined with various letters. The first figure of the three-digit number indicates the number of carbon atoms minus one (for methane derivatives, the figure 0 is omitted); the second figure indicates the number of the hydrogen atoms plus one; and the third figure indicates the number of fluorine atoms. All other bonds are saturated with chlorine. The letter R before the code number is an abbreviation for refrigerant; the letter C indicates a cyclic compound. The complete number is called the refrigerant number. The American Society of Heating, Refrigerating, and Air Conditioning Engineers (Atlanta, Georgia) ASHRAE Standard 34 – 78 describes the method of coding. The abbreviation F is sometimes used and stands for fluorohydrocarbon.

This system does not allow isomerisms to be expressed for ethane derivatives; for such cases, a letter (a, b, . . . etc.) is added to isomers as their asymmetry increases, e.g.,

$CF_2Cl-CCl_2F$ = CFC 113 (R 113)
CF_3-CCl_3 = CFC 113 a (R 113 a)

The compound with the highest degree of symmetry is not given a letter.

A special isomer in the series of propane derivatives is designated by adding two letters to the numbers derived from the standard rules [60]. The first letter attached to the number refers to the central carbon atom, coding the total atomic mass of the two substituents attached (a = CCl_2, b = CFCl, c = CF_2, d = CHCl, e = CHF, f = CH_2). The second appended letter is derived from the symmetry rule applied to the two terminal carbon atoms combined to an imaginary ethane unit. This unit is then treated like an ethane derivative with the difference that the letter a is given to the most symmetric combination. Examples for the codes of the isomeric propane derivatives $C_3HCl_2F_5$ are:

$CF_3–CF_2–CHCl_2$	HCFC 225 ca
$CF_2Cl–CF_2–CHFCl$	HCFC 225 cb
$CF_3–CHCl–CF_2Cl$	HCFC 225 da
$CF_3–CHF–CFCl_2$	HCFC 225 eb.

Codes for the butane derivatives contain three letters appended to the numeral (e.g., $CF_3–CH_2–CF_2–CH_3$ is coded as HFC 365 mfc [61]).

3.1. Fluoroalkanes and Perfluoroalkanes

Properties. *Monofluoroalkanes* are attacked by bases and sometimes by heat; however, chemical resistance increases with increasing fluorine substitution, especially multiple substitution at the same carbon atom.

Perfluoroalkanes have distinct properties [62], [63]. Their physical properties differ from those of the corresponding hydrocarbons: densities and viscosities are higher, whereas surface tensions, refractive indices, and dielectric constants are lower. At room temperature perfluoroalkanes are attacked only by sodium in liquid ammonia. At 400 – 500 °C, they are degraded by alkali metals and silicon dioxide; the former produce a metal fluoride and carbon, while the latter produces silicon tetrafluoride and carbon dioxide. Thermal decomposition starts above 800 °C (compounds with tertiary carbon atoms above 600 °C) with the formation of saturated and unsaturated decomposition products and some carbon. In addition to their chemical and physical stability, perfluoroalkanes are characterized by nonflammability and physiological inertness.

Whereas partially fluorinated alkanes dissolve in many common solvents, perfluoroalkanes have low solubility, which decreases with their chain length. Only ethers, ketones, esters, chlorohydrocarbons, and chlorofluorocarbons have the power to dissolve perfluoroalkanes [62]. Boiling points, melting points, and densities of fluoroalkanes and perfluoroalkanes are listed in Table 2. For other physical properties of HFCs, see Table 3. Physical data for liquid perfluoralkanes are given in Tables 4 and 5.

Production. Mono- and difluoroalkanes can be produced by addition of hydrogen fluoride to olefins or alkynes (e.g., $CF_2H–CH_3$, HFC 152 a). Another synthetic pathway is the exchange of chlorine (or bromine) for fluorine using hydrogen fluoride or metal fluorides such as antimony fluoride. Monohydroperfluoroalkanes can be obtained by adding hydrogen fluoride to perfluoroalkenes (e.g., $CF_3–CHF–CF_3$, HFC 227 ea) or by decarboxylation of perfluorocarboxylates in the presence of proton donors. The novel commercially interesting hydrofluorocarbons are produced by processes, that have been developed for the production of chlorofluorocarbons and hydrofluorocarbons and optimized during the last decades (see Section 3.2).

Higher temperatures and higher hydrogen fluoride : substrate ratios are necessary to achieve complete replacement of all chlorine atoms in the starting chlorocompounds by fluorine. Both liquid-phase halogen exchange in the presence of catalysts such as antimony(V) or tin(IV) chlorofluorides and vapor phase reactions using solid-phase catalysts based on chromium are employed. Preferred starting materials are chloroform for HFC 23 [45], [65], dichloromethane for HFC 32 [66] and 1,1,1-trichloroethane for HFC 143 a [67]. The conversion of tetrachloroethylene to HFC 125 [68] and trichloroethylene to HFC 134 a [69] involves initial HF-addition across the double bond followed by a series of chlorine – fluorine exchange reactions. Vapor-phase hydrogenolysis of chlorofluorocarbons and hydrochlorofluorocarbons is also proposed for the production of hydrofluorocarbons [70], [71].

Perfluoroalkanes can be produced by a variety of routes. Indirect fluorination of hydrocarbons with cobalt(III) fluoride or silver(II) fluoride is carried out in a steel or nickel tube with stirring. The hydrocarbon vapors are passed at 150 – 450 °C over the fluorinating agent, which is regenerated in a fluorine stream [72], [73]. This process is suitable for the production of perfluoroalkanes containing up to 20 carbon atoms. It affords better process control and yields than the direct gas-phase fluorination using dilute fluorine and a metal catalyst, especially for longer-chain compounds [74].

Fluoroalkanes and perfluoroalkanes can also be produced electrochemically by the Phillips Petroleum process or the electrochemical fluorination of alcohols, amines, carboxylic acids, and nitriles by the Simons process (see Section 2.1).

Table 2. Boiling points, melting points, and densities of fluoroalkanes and perfluoroalkanes

Compound	CAS registry no.	Formula	M_r	Refrigerant no.	Code no.	*bp*, °C	*mp*, °C	d_4^θ, g/cm^3 (θ, °C)
Fluoromethane	[*593-53-3*]	CH_3F	34.03	R 41	HFC 41	− 78.5	− 141.8	0.8428 (− 60)
Difluoromethane	[*75-10-5*]	CH_2F_2	52.03	R 32	HFC 32	− 51.7	− 136	1.100 (20)
Trifluoromethane	[*75-46-7*]	CHF_3	70.02	R 23	HFC 23	− 82.1	− 160	1.246 (− 34)
Tetrafluoromethane	[*75-73-0*]	CF_4	88.01	R 14	PFC 14	− 128	− 183.6	1.33 (− 80)
Fluoroethane	[*353-36-6*]	CH_3CH_2F	48.06	R 161	HFC 161	− 37.1	− 143.2	0.8176 (− 37)
1,2-Difluoroethane	[*624-72-6*]	CH_2FCH_2F	66.05	R 152	HFC 152	30.7		0.913 (19)
1,1-Difluoroethane	[*75-37-6*]	CH_3CHF_2	66.05	R 152a	HFC 152a	− 24.7	− 117	0.966 (19)
1,1,2-Trifluoroethane	[*430-66-0*]	CHF_2CH_2F	84.04	R 143	HFC 143	5.0	− 84	
1,1,1-Trifluoroethane	[*420-46-2*]	CH_3CF_3	84.04	R 143a	HFC 143a	− 47.6	− 111	0.942 (30)
1,1,2,2-Tetrafluoroethane	[*359-35-3*]	CHF_2CHF_2	102.03	R 134	HFC 134	− 19.7	− 89	
1,1,1,2-Tetrafluoroethane	[*811-97-2*]	CF_3CH_2F	102.03	R 134a	HFC 134a	− 26.3	− 101	1.2078 (25)
Pentafluoroethane	[*354-33-6*]	CF_3CHF_2	120.03	R 125	HFC 125	− 48.5	− 103	1.250 (20)
Hexafluoroethane	[*76-16-4*]	CF_3CF_3	138.02	R 116	PFC 116	− 78.1	− 100.6	1.607 (− 78)
1,1,2,2,3-Pentafluoropropane	[*679-86-7*]	$CHF_2CF_2CH_2F$	134.05	R 245ca	HFC 245ca	26	− 82	
1,1,1,2,2-Pentafluoropropane	[*1814-88-6*]	$CF_3CF_2CH_3$	134.05	R 245cb	HFC 245cb	− 18		
1,1,1,3,3-Pentafluoropropane	[*460-73-1*]	$CF_3CH_2CHF_2$	134.05	R 245fa	HFC 245fa	15.3		1.320 (25)
1,1,1,2,3,3,3-Heptafluorpropane	[*431-89-0*]	CF_3CHFCF_3	170.03	R 227ea	HFC 227ea	− 16.5	− 131	1.394 (25)
Octafluoropropane	[*76-19-7*]	$CF_3CF_2CF_3$	188.03	R 218	PFC 218	− 36.7	− 183	1.350 (20)
Octafluorocyclobutane	[*115-25-3*]	F_2C-CF_2 / F_2C-CF_2	200.04	RC 318	PFCC 318	− 6.06	− 40.7	1.5241 (20)
1,1,1,3,3-Pentafluoro-*n*-butane	[*406-58-6*]	$CF_3CH_2CF_2CH_3$	148.07	R 365mfc	HFC 365mfc	40.2		1.264 (20)
Decafluoro-*n*-butane	[*355-25-9*]	$CF_3CF_2CF_2CF_3$	283.02	R 610		− 22	− 128.2	1.517 (20)

Tetrafluoromethane (carbon tetrafluoride, CF_4) can be produced by reaction of CCl_2F_2 or CCl_3F and hydrogen fluoride in the gas phase [75] or by direct fluorination of carbon [76]. Hexafluoroethane (PFC 116) is often obtained as a byproduct in the production of CFC 115. Octafluoropropane (PFC 218) can be produced by direct [77], electrochemical [78], or CoF_3-fluorination [79] of commercially available hexafluoropropene (see Section 4.2). Octafluorocyclobutane is obtained by dimerization of tetrafluoroethylene [80] or by passing 1,2-dichloro-1,1,2,2-tetrafluoroethane, $CClF_2CClF_2$, over a nickel catalyst at 590 °C [81].

Uses. Hydrofluorocarbons have been gaining increasing commercial interest as substitutes for chlorofluoro- and hydrochlorofluorocarbons since governmental regulations banned the worldwide production and consumption of CFCs by 1996 and introduced a specific timetable for the phase out of HCFCs [82]. In contrast \011to CFCs hydrofluorocarbons have no adverse effect on the ozone layer and only a low contribution to global warming (see below). The latter effect could be further minimized by avoiding leakages in refrigeration and air-conditioning equipment and by refrigerant recycling.

In all applications involving considerable emissions to the atmosphere, e.g., as propellants in aerosols (except medicinal aerosols), in open cell foams, or in extruded foams CFCs will be replaced by nonhalogenated compounds in the future.

Due to the presence of hydrogen in the molecule, the stability of HFCs is reduced. In the atmosphere they are degraded below the stratospheric level, leading to zero ozone depletion potentials (ODP = 0) and to reduced halogen global warming potentials (HGWPs). However, depending on the structure of the HFCs this re-

Table 3. Physical properties of hydrofluorocarbons

Property	CH_2F_2	CHF_3	CHF_2CH_3	CF_3CH_3	CF_3CH_2F	CF_3CHF_2	$CF_3CH_2CHF_2$	CF_3CHFCF_3	$CF_3CH_2CF_2CH_3$
Critical temperature, °C	78.2	26.3	113.3	73.6	101.1	66.3	154.1	101.8	187.7
Critical pressure, MPa	5.80	4.87	4.52	3.83	4.06	3.63		2.93	2.75
Critical density, g/cm^3		0.527	0.365	0.434	0.515		0.517	0.582	
Heat of evaporation at *bp*, kJ/kg		240.8	326.0	230.0	213.46		177.03	208.96	131.8
Specific heat, kJ/kg^{-1} K^{-1}						876			
Refractive index, n_D	1.200 (− 50)	1.442 (− 80)	1.023 (− 30)			1.53 (− 48.5)		1.407 (25)	
Surface tension, N/cm	6.93×10^{-5}		9.94×10^{-5}	4.60×10^{-5}	8.02×10^{-5}	3.70×10^{-5}			
Vapor pressure (kPa) at									
− 120 °C		5.9							
− 100 °C		31.5							
− 80 °C		113.9							
− 60 °C		314.0							
− 40 °C	177.3	712.0	141.7	52.0	148.4		32.4		
− 20 °C	405.8	1403.0	120.7	316.4	134.0	337.6	19.7	86.7	6.3
0 °C	813.4	2504.3	264.2	620.2	293.0	670.6	53.6	196.2	18.8
20 °C	1474.6	4184.3	513.4	1104.3	572.0	1204.6	123.8	390.2	46.7
40 °C	2477.4		909.7	1831.5	1017.0	2008.1	251.8	702.9	100.9
60 °C	3933.4		1501.0	2884.7	1681.0		463.5	1174.7	194.9
80 °C			2344.1		2631.0		788.8	1857.2	344.6
100 °C			3511.4		3970.0		1261.0	2824.6	567.3
120 °C							1920.0		880.2
140 °C									1300.2
Lower ignition limit (25 °C), vol %,	12.7	none	3.1	7.1	none	none	none	none	3.5
Atmospheric lifetime, a					15.6	40.5	7.4	40	10.8
HGWP ∗ (CFC 11 = 1)	0.15	8.4	0.03	1.0	0.29	0.67		0.69	

∗ HGWP = Halogen Global Warming Potential

Table 4. Boiling points, melting points, and densities of liquid perfluorocarbons

Compound (also mixtures)	CAS registry no.	Molecular formula	M_r	Commercial designation [a]	*bp*, °C	*pour point*, °C	d_4^{25}, g/cm^3
F-pentanes [b]	[*678-26-2*]	C_5F_{12}	288	PP50	29	− 120	1.604
F-methylcyclopentanes		C_6F_{12}	300	PP1C	48	− 70	1.707
F-hexanes	[*355-42-0*]	C_6F_{14}	338	PP1	57	− 90	1.682
F-methylcyclohexanes		C_7F_{14}	350	PP2	76	− 30	1.778
F-decalin (cis/trans)	[*306-94-5*]	$C_{10}F_{18}$	462	PP5/PP6	142	− 8	1.917
F-perhydrofluorene		$C_{13}F_{22}$	574	PP10	194	− 40	1.984
F-perhydrophenanthrene	[*306-91-2*]	$C_{14}F_{24}$	624	PP11	215	− 20	2.03
F-perhydrofluoranthene		$C_{16}F_{26}$	686	PP24	244	0	2.052
F-cyclohexylmethyldecalin		$C_{17}F_{30}$	774	PP25	260	− 10	2.049

[a] Flutec notation (BNFL Fluorochemicals); [b] F = perfluoro

Table 5. Further physical properties of liquid perfluoroalkanes [64]

Property	PP50	PP1C	PP1	PP2	PP5/6	PP10	PP11	PP24	PP25
Critical temperature, °C	148.7	180.8	177.9	212.8	292.0	357.2	377	388.7	400.4
Critical pressure, MPa	2.05	2.26	1.83	2.02	1.75	1.62	1.46	1.51	1.13
Critical volume, L/kg	1.626	1.567	1.582	1.522	1.521	1.59	1.58	1.606	1.574
Heat of evaporation at *bp*, kJ/kg	90.8	75.8	85.5	85.9	78.7	71	68	65.8	67.9
Specific heat, kJ/kg^{-1} K^{-1}	1.05	0.878	1.09	0.963	1.05	0.92	1.07	0.93	0.957
Refractive index, n_D^{20}	1.283	1.2650	1.2509	1.2781	1.3130	1.3289	1.3348	1.3462	1.3376
Surface tension, N/cm	9.4×10^{-5}	12.6×10^{-5}	11.1×10^{-5}	15.4×10^{-5}	17.6×10^{-5}	19.7×10^{-5}	19×10^{-5}	22.2×10^{-5}	
Viscosity (dynamic), mPa · s	0.465	1.049	0.656	1.561	5.10	9.58	28.4	31.5	114.5
Vapor pressure (kPa) at 25 °C	8.62	3.68	2.94	1.41	0.09	< 0.01	< 0.01	< 0.01	< 0.01

duced stability can entail an increased flammability and thus make safe handling more difficult. Therefore, alternatives to the CFCs need to retain the attractive properties of CFCs like low toxicity, nonflammability, good thermodynamic properties, and accessibility via economically and ecologically viable manufacturing processes, but avoid any adverse effect to the environment.

In refrigeration and air-conditioning systems CFC 12 is replaced by HFC 134 a, HCFC 22 by the azeotropic mixtures HFC 507 (HFC 125/HFC 143 a 1 : 1) or HFC 410 (HFC 32/HFC 125 1 : 1) and CFC 13 by HFC 23. HFC 134 a and HFC 227 ea can be used as propellants in medicinal aerosols instead of CFC 114. HFC 245 fa and HFC 365 mfc are proposed as blowing agents for foams in replacing CFC 11, CFC 113 and HCFC 141 b [83]. Until now no HFC candidate for replacement of CFC 11 and CFC 113 as solvents, degreasing agents, or cleaning agents for textiles or metal surfaces has been identified. In the period 1990–1995 159.5 x 10^3 t of HFC 134 a have been produced [84].

Gaseous perfluorocarbons (PFC 14, PFC 116, PFC 218) are used in plasma etching processes in the microelectronic industry [85] and as gaseous dielectrics. Liquid perfluorocarbons [64] serve as heat-transfer media in transformers and in capacitors, as lubricants and hydraulic fluids, or in vapor-phase soldering [86] and vapor-phase sterilization [64].

Perfluoroalkanes, e.g., perfluorodecalin [*306-94-5*], are used in the production of blood substitutes [87].

Trade Names. Hydrofluorocarbons substituting chlorofluorocarbons and hydrochlorofluorocarbons are marketed worldwide under protected trade names. The individual product is characterized by the HFC code number following the trade name. Some worldwide applied trade names are:

France	
Rhône-Poulenc	Isceon
Elf Atochem	Forane
Germany	
Hoechst	Reclin
Solvay	Solkane
United Kingdom	
ICI	Klea
Italy	
Montefluos	Algogrene
Japan	
Asahi Glass	Asahiflon
Daikin	Daiflon
United States	
Allied Signal	Genetron
DuPont	Suva
Great Lakes Chemical	FM
3 M	3M Brand

Perfluorocarbons are offered under trade names such as Freon C-51-12; Perfluorokerosene FCX-329, FCX-330;Perfluorolube oil FCX-512, FCX-412 (DuPont); Flutec PP-1, PP-2, PP-3, PP-9, etc. (BNFL Fluorochemicals), Multifluor Inert Fluids (Air Products and Chemicals).

3.2. Chlorofluoroalkanes

For more than 50 years chlorofluorocarbons and hydrochlorofluorocarbons have been the most important organic fluorine compounds commercially. The five products listed in Table 6 have been by far the most important of these with regard to the field of applications and to the amount produced [84].

Table 6. Economically most important organic fluorine compounds

Compound	Total production through 1995, 10^6 t	Maximum annual production (year), 10^3 t
Trichlorofluoromethane (R 11, CFC 11)	8.62	382 (1987)
Dichlorodifluoromethane (R 12, CFC 12)	11.28	425 (1987)
Chlorodifluoromethane (R 22, HCFC 22)	3.85	243 (1995)
1,2,2-Trichloro-1,1,2-trifluoroethane (R 113, CFC 113)	2.97	251 (1989)
1,2-Dichloro-1,1,2,2-tetrafluoroethane (R 114, CFC 114)	0.51	19 (1986)

The results of the investigation on postulated atmospheric changes caused by CFCs [88] with respect to the ozone balance and the greenhouse effect [89], [90] verified the adverse impact on the stratospheric ozone layer and the significant contributions to global warming due to the long atmospheric lifetimes of the CFCs. In 1987, a United Nations agreement, called the Montreal Protocol – revised in 1992 during the Copenhagen Intergovernmental Conference – set the deadline for the phase out of CFCs in developed countries. Since 1996 production and consumption of CFCs are prohibited, except as intermediates in the production of fluorine chemicals, especially fluoropolymers. For HCFCs with reduced atmospheric stabilities a timetable has been introduced for their phase out. HCFCs are allowed to be used as drop-in alternatives for CFCs until suitable HFC substitutes (see Section 3.1) will have been developed, but no longer than until 2015 to 2030 [82].

High molecular mass chlorofluorocarbons, a small but significant class of CFCs will not be affected by the ban.

Properties. Chlorofluoroalkanes are characterized by high chemical and thermal stabilities, which increase with their fluorine content. Low flammability (or nonflammability) and low toxicity are additional commercial advantages. Most of these compounds have a pleasant, weak odor; some are mild anesthetics [91].

Boiling points, freezing points, and densities of formerly commercially important chlorofluoroalkanes are shown in Table 7; other physical properties are listed in Table 8. Physical properties of some high molecular mass CFCs produced from chlorotrifluoroethylene are given in Table 9.

Production. Commercial production of chlorofluoroalkanes employs halogen exchange, with hydrogen fluoride in the liquid phase in the presence of a catalyst. The production scheme for dichlorodifluoromethane shown in Figure 1 is typical [93]:

$$CCl_4 + 2\,HF \longrightarrow CCl_2F_2 + 2\,HCl$$

A steam-heated steel autoclave (a) lined with stainless steel (V2A) serves as the reactor. The seals are made from aluminum or copper. The autoclave (capacity $2-5\,m^3$) is filled with 500 kg of hydrogen fluoride, 1540 kg of car-

Table 7. Boiling points, melting points, and densities of chlorofluoroalkanes

Compound	CAS registry no.	Formula	M_r	Refrigerant number	Code no.	*bp*, °C	*mp*, °C	d_4^θ, g/cm^3 (θ,°C)
Trichlorofluoromethane	[*75-69-4*]	CCl_3F	137.38	R 11	CFC 11	23.7	– 111	1.490 (20)
Dichlorodifluoromethane	[*75-71-8*]	CCl_2F_2	120.93	R 12	CFC 12	– 29.8	– 155	1.328 (20)
Chlorotrifluoromethane	[*75-72-9*]	$CClF_3$	104.47	R 13	CFC 13	– 81.1	– 181	0.924 (20)
Dichlorofluoromethane	[*75-43-4*]	$CHCl_2F$	102.93	R 21	HCFC 21	8.9	– 135	1.366 (20)
Chlorodifluoromethane	[*75-45-6*]	$CHClF_2$	86.48	R 22	HCFC 22	– 40.8	– 160	1.213 (20)
Tetrachloro-1,2-difluoroethane	[*76-12-0*]	CCl_2FCCl_2F	203.85	R 112	CFC 112	92	27.4	1.634 (30)
Tetrachloro-1,1-difluoroethane	[*76-11-9*]	$CClF_2CCl_3$	203.85	R 112 a	CFC 112 a	91.5	40.8	1.649 (20)
1,1,2-Trichlorotrifluoroethane	[*76-13-1*]	CCl_2FCClF_2	187.39	R 113	CFC 113	47.7	– 33	1.582 (20)
1,1,1-Trichlorotrifluoroethane	[*354-58-5*]	CF_3CCl_3	187.39	R 113 a	CFC 113 a	45.9	14	1.579 (20)
1,2-Dichlorotetrafluoroethane	[*76-14-2*]	$CClF_2CClF_2$	170.94	R 114	CFC 114	3.8	– 94	1.473 (20)
1,1-Dichlorotetrafluoroethane	[*374-07-2*]	CF_3CCl_2F	170.94	R 114 a	CFC 114 a	– 2	– 56.6	1.478 (21)
Chloropentafluoroethane	[*76-15-3*]	CF_3CClF_2	154.48	R 115	CFC 115	– 38	– 106	1.291 (25)
1,1,2-Trichloro-2,2-difluoroethane	[*354-21-2*]	$CClF_2CHCl_2$	169.39	R 122	HCFC 122	71.9	– 140	1.544 (25)
1,1-Dichlor-2,2,2-trifluoroethane	[*306-83-2*]	CF_3CHCl_2	152.94	R 123	HCFC 123	28.7	– 107	1.475 (15)
1-Chloro-1,2,2,2-tetrafluoroethane	[*2837-89-0*]	CF_3CHClF	136.48	R 124	HCFC 124	– 12	– 199	1.364(25)
1,2-Dichloro-1,1-difluoroethane	[*1649-08-7*]	$CClF_2CH_2Cl$	134.94	R 132 b	HCFC 132 b	46.8	– 101.2	1.4163 (20)
1-Chloro-2,2,2-trifluoroethane	[*75-88-7*]	CF_3CH_2Cl	118.49	R 133 a	HCFC 133 a	6.9	– 101	1.389 (0)
1,1-Dichloro-1-fluoroethane	[*1717-00-6*]	CCl_2FCH_3	116.95	R 141 b	HCFC 141 b	32	– 103.5	1.250 (10)
1-Chloro-1,1-difluoroethane	[*75-68-3*]	$CClF_2CH_3$	100.49	R 142 b	HCFC 142 b	– 9.2	– 130.8	1.120 (25)
1,1-Dichloro-2,2,3,3,3-pentafluoropropane	[*422-56-0*]	$CF_3CF_2CHCl_2$	202.9	R 22 5ca	HCFC 225 ca	51.1	– 94	1.550 (25)
1,3-Dichloro-1,2,2,3,3-pentafluoropropane	[*507-55-1*]	$CClF_2CF_2CHClF$	202.9		HCFC 225 cb	56.1	– 97	1.560 (25)

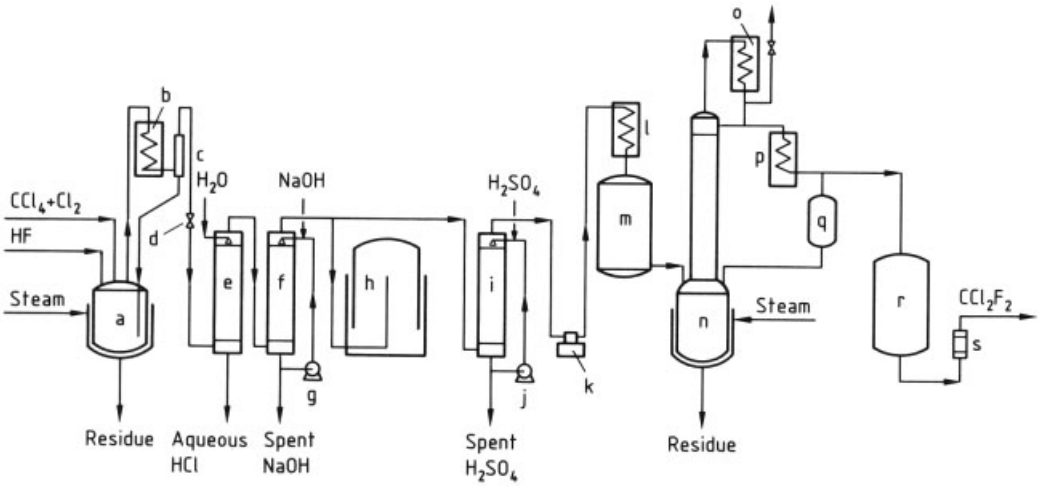

Figure 1. Manufacture of dichlorodifluoromethane (R 12) by fluorination of carbon tetrachloride in the liquid phase
a) Autoclave; b) Reflux condenser; c) Separator; d) Pressure valve; e) Wash column (water); f) Wash column (NaOH); g) Pump for NaOH; h) Gasometer; i) Wash column (H_2SO_4); j) Pump for H_2SO_4; k) Compressor; l) Condenser; m) Receiver for crude product; n) Distillation pot; o) Dephlegmator; p) Condenser; q) Forerun receiver; r) Tank for pure product; s) Caustic filter

Table 8. Physical properties of chlorofluoroalkanes [92]

Property	CCl_3F	CCl_2F_2	$CClF_3$	$CHClF_2$	CCl_2FCClF_2	$CClF_2CClF_2$	CF_3CHCl_2	CF_3CHClF	CCl_2FCH_3	$CClF_2CH_3$
Critical temperature, °C	198.0	112.0	28.8	96.0	214.1	145.7	185	122.2	210.3	137.1
Critical pressure, MPa	4.40	4.21	3.86	4.94	3.41	3.27	3.79	3.57	4.64	4.12
Critical density, g/cm^3	0.548	0.558	0.581	0.525	0.576	0.578				
Heat of evaporation at *bp*, kJ/kg	182.16	166.88	148.50	234.12	145.70	139.42	174.17	167.9	223.15	223.15
Specific heat at 101.3 kPa, J kg^{-1} K^{-1}	871	854	850	1088	946	971	1017.4	1130	1155.6	1297.9
Refractive index, $n_D^{26.5}$	1.384	1.285	1.263	1.252	1.355	1.290	1.3322 (15)		1.3600 (10)	
Surface tension, N/cm	19×10^{-5}	9×10^{-5}		9×10^{-5}	19×10^{-5}	13×10^{-5}			19×10^{-5}	
Solubility in water, g/100 g										
at 0 °C	0.0036	0.0025	0.0019	0.060	0.0036	0.0026				
at 30 °C	0.013	0.0125	0.0065	0.15	0.013	0.011	0.39 (25 °C)	1.71(25 °C)	0.021(25 °C)	0.14(25 °C)
Dielectric strength at 101.3 kPa, 23 °C, nitrogen = 1	3.1	2.4	1.4	1.3	2.6 (39.2 kPa)	2.8				
Dielectric constant										
liquid at 25 °C∗	2.5	2.1	2.3 (− 30 °C)	6.6	2.6	2.2				
vapor (*t*, °C)	1.0019 (26)	1.0016 (29)	1.0013 (29)	1.0035 (25.4)	1.0024 (27.5)	1.0021 (26.8)			7.9 (20 °C)	
Vapor pressure (kPa) at										
− 120 °C			6.96							
− 100 °C		1.18	33.14							
− 80 °C		6.12	109.8							
− 60 °C		22.7	282.5			3.6				
− 40 °C	5.1	64.3	607			13.0				
− 20 °C	15.7	151.0	1146	246	5.08	37.0				
0 °C	40.2	308.7	1969	500	14.79	87.9			28	
20 °C	89.0	566.9	3177	917	36.4	182			65	
40 °C	176	958.5		1549	78.3	340			184 (50 °C)	
60 °C	316	1518.7		2459	151.3	583				
80 °C	528	2284.7			268.0	935				
100 °C	830	3297.5			442.2	1423				
120 °C	1242					2083				
140 °C	1785					2964				
Atmospheric lifetime, a	55	116		15.8	110	220	1.71	7.0	10.8	22.4
ODP	1.00	1.00		0.055	1.07	0.80	0.02	0.022	0.11	0.065
HGWP (CFC11 = 1)	1.00	3.00		0.33	1.6	7.1	0.02	0.11	0.14	0.41

∗ Except where otherwise stated.

Table 9. Physical data of high molecular mass CFCs with the structure $Cl(CF_2CFCl)_nCl$

$n =$	2	3	4	5
M_r	304	420.5	537	653.5
bp, °C	136	205	255	300
Density, $d_4^{37.8}$	1.713	1.808	1.865	1.902
Viscosity (dynamic), mPa · s (37.8 °C)	1.35	3.4	10.8	48.9

bon tetrachloride, 220 kg of antimony(III) chloride, and 20 kg of chlorine, and the mixture is heated to 100 °C. After ca. 2 h and an increase in pressure to ca. 3 MPa, the fluorination products with lower boiling points are removed together with the hydrogen chloride that is generated and some hydrogen fluoride; higher-boiling products in the exit gases are condensed and recycled. The low-boiling fraction is first washed with water in a tower (e) lined with poly(vinyl chloride) and packed with graphite; it is then washed with caustic in a tower (f) filled with porcelain packing. After being washed to neutrality, the product is dried in a tower (i) containing concentrated sulfuric acid, compressed to a liquid, and fed into an intermediate storage tank (m). Each batch takes ca. 24 h to process. The antimony catalyst remains in the reactor and is regenerated before each subsequent batch by adding a small amount of chlorine to convert it to the catalytically active Sb(V) form.

The crude product is fractionally distilled under pressure (0.6 – 0.8 MPa). The lower-boiling fraction contains some chlorotrifluoromethane and most of the dichlorodifluoromethane (yield 90 % based on carbon tetrachloride, 80 % based on hydrogen fluoride). The higher-boiling fraction consists of trichlorofluoromethane (5 – 10 % based on carbon tetrachloride), which can be recycled. The distilled product is passed through a caustic filter (s). Steel bottles, pressure vessels, tank cars, and tank trucks are used for transport.

More recently developed exchange processes are carried out continuously in the gas phase at 100 – 400 °C, using catalysts based on chromium [45], aluminum [44], or iron [94]. Starting materials, which include carbon tetrachloride, chloroform, tetrachloroethylene, and trichloroethylene, are passed over the catalyst with excess hydrogen fluoride and, where necessary, chlorine. Further processing follows the same principles as in the liquid-phase process.

In the Montedison chlorofluorination process, reaction of C_1- and C_2-hydrocarbons with chlorine and hydrogen fluoride takes place in a single step in a fluidized-bed reactor. A suitable catalyst is a combination of aluminum chloride and other metals [95–97]:

$$CH_4 + 4\,Cl_2 + 2\,HF \xrightarrow{\text{Cat.}} CCl_2F_2 + 6\,HCl$$

Commercial production of chlorofluoroalkanes is also possible by the electrochemical fluorination process developed by Phillips Petroleum (see Section 2.1).

High molecular mass chlorofluoroalkanes are produced by fluorination with chlorine trifluoride [98], [99].

Other important processes for production of high molecular mass CFCs are based on the telomerization of chlorotrifluoroethylene (see Section 4.7) with carbon tetrachloride [100] or CFC 113 [101] as telogens. Stabilization and end-group fluorination are achieved using cobalt trifluoride as fluorinating agent [102], [103].

Specifications. Chlorofluoroalkanes (and also the alternative HCFCs and HFCs) produced on an industrial scale are subject to stringent standards. Impurities must not exceed the following limits (vol %):

acids	0
moisture	< 0.001
higher-boiling fractions	< 0.05
other gases	2

Uses. Up to the ban in the USA in 1978 chlorofluoroalkanes had been used mainly as aerosol propellants and as spraying and foam blowing agents (R 11, R 12, R 114). Further important applications up to 1996 had been in the area of refrigerants, where R 11, R 12, R 13, R 22, R 113, R 114, R 115, and the chlorine-free compounds R 23 and RC 318 were preferred. Of the drop-in

alternatives to CFCs HCFC 22 is the most important compound, used as refrigerant (annual production in 1995 243×10^3 t [84]), followed by HCFC 141 b (113×10^3 t in 1995 [84]) and HCFC 142 b (38×10^3 t in 1995 [84]) used as blowing agents for closed cell foams. For HFCs as alternatives see Section 3.1.

Chlorofluoroalkanes, especially R 11 and R 113, were also employed as solvents and degreasing and cleaning agents for textiles; HCFC 22, CFC 113, and HCFC 142 b will be important intermediates for the production of fluoroolefins also in the future.

Higher molecular mass perchlorofluoroalkanes are used as oils, greases, and waxes, as lubricants, hydraulic fluids, damping oils, heat-transfer media, impregnating agents, and plasticizers. Oligomers of chlorotrifluoroethylene have achieved special importance in this area [104].

Trade Names. Chlorofluoroalkanes were sold worldwide under protected trade names; the refrigerant numbers describing the chemical composition (see Table 7) are included to specify individual compounds. Some of the trade names that were applied worldwide are:

Australia	
Australian Fluorine Chemicals	Isceon
Czechoslovakia	
Slovek Pro Chemickov	Ledon
Federal Republic of Germany	
Hoechst	Frigen
Kali-Chemie	Kaltron
France	
Rhône-Poulenc	Flugene
Ugine Kuhlmann	Forane
German Democratic Republic	
Volkseigener Betrieb Chemiewerk Nunchritz	Fridohna
Volkseigener Betrieb Fluorwerke Dohna	Frigedohn
Italy	
Montedison	Algofrene
Japan	
Asahi Glass	Asahiflon
Daikin Kogyo	Daiflon
Mitsui Fluorochemicals	Flon
North America	
Allied Chemical	Genetron
DuPont	Freon
Kaiser Chemicals	Kaiser
Pennwalt	Isotron
Racon	Racon
Union Carbide	Ucon
Soviet Union	
	Khladon
	Eskimon
The Netherlands	
AKZO	FCC
United Kingdom	
ICI	Arcton
Imperial Smelting Corporation	Isceon

Trade names of higher molecular mass chlorofluoroalkanes include Florubes (ICI), Fluorolube oils, Fluorolube greases (Hooker Industrial Chemicals Division),Halocarbon oils, greases, and waxes (Halocarbon Products),Kel-F oils, greases, and waxes (3M), and Voltalef (Atochem).

3.3. Bromofluoroalkanes

Bromofluoro compounds of practical importance are found mainly in the methane and ethane series.

In the refrigerant (R) numbering system for bromofluoroalkanes, the corresponding chlorofluoroalkanes are taken as the basic structures (see also Section 3); the substitution of chlorine by bromine is expressed by the addition of B1, B2, etc. For example, bromotrifluoromethane is denoted as R 13B1 and 1,2-dibromotetrafluoroethane as R 114B2. In firefighting applications the Halon numbering system is used, specifying the number of carbon, fluorine, chlorine, and bromine atoms in the molecules when reading from left to right (e.g., Halon 1301 denotes CF_3Br; Halon 1211 is CF_2ClBr; and Halon 2402 denotes CF_2BrCF_2Br).

Bromofluorocarbons (BFCs) and hydrobromofluorocarbons (HBFCs) are involved in the depletion of stratospheric ozone and global warming like the CFCs and HCFCs; however the contributions of the bromine-containing compounds are distinctly higher. Therefore, the 1992 Copenhagen meeting agreed to phase out the production of BFCs and HBFCs by 1994, with the exception of Halon for some essential firefighting applications. HFC 125 and HCFC 123 (DuPont) or HFC 227 ea (Great Lakes) are announced as alternatives though they are significantly less efficient as fire extinguishing agents than the BFC Halons

Properties. Fully halogenated compounds with a high fluorine content have excellent thermal stability; they are nonflammable and some

(e.g., CF_3Br) are physiologically inert [92]. At high temperature, thermal cleavage of the C–Br bond into radicals occurs, which is responsible for the utility of some of these compounds in extinguishing fires (→ Fire Extinguishing Agents, Chap. 6.) [105]. Their chemical stability is slightly lower than that of the corresponding chlorofluoroalkanes. However, as with the chlorofluoroalkanes, stability increases with the fluorine : bromine ratio. Some compounds have a marked anesthetic effect [91]. Physical properties are listed in Table 10.

Production. Bromofluoromethanes are obtained by bromination of a stream of the appropriate fluoromethane [106] or chlorofluoromethane [107] at 300–600 °C. Ethane derivatives can also be obtained by thermal bromination [108] or by addition of bromine or hydrogen bromide [109] to fluoroolefins. In some cases hydrogen bromide can be used to exchange a chlorine atom in a chlorofluoroalkane for a bromine atom [110]. Iodine–bromine exchange in a fluoroiodoalkane can be effected with bromine [111].

Uses. The lower-boiling compounds $CBrF_3$ (R 13B1; Halon 1301) and $CBrClF_2$ (R 12B1; Halon 1211) had been used as fire extinguishing agents. Producers were Atochem, ICI, and Solvay in Europe, DuPont and Great Lakes in the USA and Asahi Glass, Daikin, and Nippon Halon in Japan. The total worldwide Halon production was estimated to be 25×10^3 t in 1986.

Perfluoro-1-bromo-*n*-octane [*423-55-2*] is physiologically inert and is useful as an X-ray contrast agent, especially for lung examinations [112]. With its low surface tension it penetrates small spaces and evenly wets healthy lung tissue.

2-Bromo-2-chloro-1,1,1-trifluoroethane [*151-67-7*], also known as halothane, has been used worldwide since 1956 as an effective, nonflammable inhalation anesthetic (→ Anesthetics, General). It is commonly produced by the ICI process:

$$CF_3CH_2Cl \xrightarrow[-HBr]{Br_2} CF_3CHBrCl$$

or the Hoechst process [43, pp. 208–210]:

$$CF_2{=}CClF \xrightarrow[h\nu]{+HBr} CBrF_2CHClF \xrightarrow{Cat.} CF_3CHBrCl$$

As alternatives for halothane a series of fluorinated ethers (containing in addition hydrogen and chlorine atoms or exclusively hydrogen atoms) have been developed, that retain or even surmount the desirable properties of halothane as inhalation anesthetic [113]. However, also these compounds have an adverse effect on the ozone layer.

Trade Names. Fluothane (ICI), Halothane "HOECHST" (Hoechst).

3.4. Iodofluoroalkanes

Iodofluoroalkanes have become important intermediates in the commercial production of compounds containing a perfluorinated moiety.

Properties. In contrast to chloro- and bromofluoroalkanes, iodofluoroalkanes readily undergo chemical reactions [24, Chap. 6], reacting preferentially by homolytic cleavage of the C–I bond.

The radical intermediates $C_nF_{2n+1}^{\bullet}$ and $I^{\bullet}$ can add to double bonds; thus, reaction with ethylene yields 1*H*,1*H*,2*H*,2*H*-1-iodoperfluoroalkanes, which are commercially important intermediates [114]:

$$C_nF_{2n+1}I + CH_2{=}CH_2 \longrightarrow C_nF_{2n+1}CH_2CH_2I$$

Control of the reaction between iodofluoroalkanes and fluoroolefins, especially tetrafluoroethylene, can result in oligomerization (telomerization) of the olefin [115]:

$$CF_3I + n\,CF_2{=}CF_2 \longrightarrow CF_3(CF_2CF_2)_nI$$

This reaction is employed commercially and is initiated by free radicals, UV irradiation, or heat [116].

Iodofluoroalkanes also form organometallic compounds, some of which are useful intermediates, e.g., for Grignard reactions [117].

Iodoperfluoroalkanes cannot be used as alkylating agents and their applications are therefore limited. However, derivatives of the FITS-type reagent are alkylating agents [118]:

Table 10. Boiling points, melting points, and densities of bromofluoroalkanes

Compound	CAS registry no.	Formula	M_r	Halon no.	*bp*, °C	*mp*, °C	d_4^{θ}, g/m^3 (θ,°C)	ODP R 11 = 1	Atmospheric lifetime,a
Tribromofluoromethane	[*353-54-8*]	CBr_3F	270.76	1103	106	− 74.5	2.7648 (20)		
Dibromodifluoromethane	[*75-61-6*]	CBr_2F_2	209.84	1202	24.5	− 110	2.3063 (15)		
Dibromochlorofluoromethane	[*353-55-9*]	CBr_2ClF	226.30	1112	80				
Bromotrifluoromethane	[*75-63-8*]	$CBrF_3$	148.93	1301	− 57.8	− 168	1.58 (21)	16	67
Bromochlorodifluoromethane	[*353-59-3*]	$CBrClF_2$	165.38	1211	− 4	− 160.5	1.850 (15)	4	19
Bromodifluoromethane	[*1511-62-2*]	$CHBrF_2$	130.92	1201	− 15.5	− 145	1.825 (20)	1.4	5.6
1,2-Dibromotetrafluoroethane	[*124-73-2*]	$CBrF_2CBrF_2$	259.85	2402	47.5	− 110.4	2.18 (20)	6	
1-Bromo-2-chloro-1,1,2-trifluoroethane	[*354-06-3*]	$CHClFCBrF_2$	197.40	2311 a	51.7		1.864 (20)		
2-Bromo-2-chloro-1,1,1-trifluoroethane	[*151-67-7*]	$CF_3CHBrCl$	197.40	2311	50.2		1.861 (25)		

Table 11. Boiling points, melting points, and densities of iodoperfluoroalkanes

Compound	CAS registry no.	Formula	M_r	*bp*, °C	*mp*, °C	d_4^{θ}, g/cm^3 (θ, °C)
Trifluoroiodomethane	[*2314-97-8*]	CF_3I	195.9	− 22.5		2.3608 (− 32)
Pentafluoroiodoethane	[*354-64-3*]	CF_3CF_2I	245.9	12.5	− 92	2.0850 (20)
Perfluoro-1-iodopropane	[*27636-85-7*]	$CF_3(CF_2)_2I$	295.9	41.2	− 95	2.0626 (20)
Perfluoro-2-iodopropane	[*677-69-0*]	CF_3CFICF_3	295.9	38		
Perfluoro-1-iodo-*n*-butane	[*423-39-2*]	$CF_3(CF_2)_3I$	345.9	67	− 88	2.07 (15)
Perfluoro-1-iodo-*n*-pentane	[*638-79-9*]	$CF_3(CF_2)_4I$	395.9	94.4	− 50	2.0349 (27.8)
Perfluoro-1-iodo-*n*-hexane	[*355-43-1*]	$CF_3(CF_2)_5I$	445.9	117	− 45	2.06 (20)
Perfluoro-1-iodo-*n*-heptane	[*335-58-0*]	$CF_3(CF_2)_6I$	495.9	137 – 138		
Perfluoro-1-iodo-*n*-octane	[*507-63-1*]	$CF_3(CF_2)_7I$	545.9	163	20.8	2.008 (25)
Perfluoro-1-iodo-*n*-decane	[*423-62-1*]	$CF_3(CF_2)_9I$	645.9	195 – 200	65.5	1.940 (70)
1,2-Diiodotetrafluoroethane	[*354-65-4*]	CF_2ICF_2I	353.8	112		2.629 (25)
1,4-Diiodooctafluoro-*n*-butane	[*375-50-8*]	$I(CF_2)_4I$	453.8	150	− 9.0	2.4739 (27)

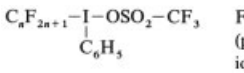

FITS reagent (perfluoroalkyl)phenyliodonium triflate

Physical constants of some iodofluoroalkanes are shown in Table 11.

Production. Iodofluoroalkanes can be produced by heating the silver salts of the perfluorocarboxylic acids with iodine [119] or the corresponding sodium salts with iodine in dimethylformamide [120]. Of commercial importance is the production of pentafluoroiodoethane (CF_3CF_2I) by reaction of tetrafluoroethylene with a mixture of iodine pentafluoride and iodine [121]:

$$5\,CF_2{=}CF_2 + IF_5 + 2\,I_2 \longrightarrow 5\,CF_3CF_2I$$

Heptafluoro-2-iodopropane (CF_3CFICF_3) is obtained similarly from hexafluoropropene.

The higher homologues $C_nF_{2n+1}I$ ($n = 4$ – 12) are produced commercially by reaction of the lower members with tetrafluoroethylene (telomerization).

α,ω-Diiodoperfluoroalkanes are obtained from tetrafluoroethylene and iodine [122]:

$$I_2 + CF_2{=}CF_2 \longrightarrow ICF_2CF_2I \xrightarrow{CF_2=CF_2} I(CF_2CF_2)_nI$$

Uses. 1-Iodoperfluoroalkanes and 1*H*,1*H*,-2*H*,2*H*-1-iodoperfluoroalkanes are intermediates in the production of surfactants and textile finishes [123]. Perfluorocarboxylic acids, especially perfluorooctanoic acid, are obtained from perfluoroiodoalkanes [124], and perfluori-

nated dicarboxylic acids are obtained from α,ω-diiodoperfluoroalkanes [125].

4. Fluorinated Olefins

The commercial importance of fluoro- and chlorofluoroolefins lies in the production of fluorinated plastics and inert fluids.

Properties. The chemical behavior of fluoroolefins [126] is governed by the number of vinylic fluorine atoms. In contrast to their hydrocarbon analogues, fluoroolefins are attacked by electrophiles only with difficulty [127], which increases with the degree of fluorination. However, fluoroolefins react readily with nucleophiles [128], [129], because as the number of vinylic fluorine atoms increases, the π-electron system of the double bond is destabilized. Thermodynamic calculations have shown that the strength of the C – C π-bond in tetrafluoroethylene

$$CF_2{=}CF_2 \rightleftharpoons {}^{\bullet}CF_2{-}CF_2^{\bullet}$$

is only ca. 160 kJ/mol as opposed to 241 kJ/mol in ethylene [130]. In unsymmetrically substituted fluoroolefins, the nucleophile attacks the carbon atom that is made strongly positive by the neighboring fluorine atoms and is shielded only weakly (sp^2 hybridization). The reactivity of fluoroolefins toward nucleophiles increases as follows:

$$CF_2{=}CF_2 < CF_2{=}CF{-}CF_3 < CF_2{=}C(CF_3)_2$$

Fluoroolefins are slightly to highly toxic and must be handled with care. The toxicity of fluorinated olefins is apparently proportional to their reactivity toward nucleophiles [131]. Perfluoroisobutene, $CF_2{=}C(CF_3)_2$ [*382-21-8*], for example, is far more toxic than its lower homologues.

Physical properties of commercial fluoro- and chlorofluoroolefins are given in Table 12.

Fluoroolefins also differ from hydrogen-containing olefins in their marked tendency to undergo cycloaddition [132].

Production. Fluoroolefins are produced by dehalogenation of chlorofluoro-, bromofluoro-, or iodofluoroalkanes with zinc and alcohol, by dehydrohalogenation of hydrogen-containing haloalkanes with alcoholic alkali, or by heating. Other common methods include addition of hydrogen halides to alkynes, decarboxylation of fluorocarboxylic acid salts, and pyrolysis of fluorohydrocarbons [133–135].

4.1. Tetrafluoroethylene

Properties. Tetrafluoroethylene (TFE), perfluoroethylene, $CF_2{=}CF_2$, a colorless, odorless gas, is flammable in oxygen, producing tetrafluoromethane and carbon dioxide. For physical properties, see Table 12. At low temperature in the presence of oxygen, explosive peroxides are formed [136], [137]. Tetrafluoroethylene must be handled with great care since, even in the absence of oxygen, it can decompose explosively into carbon and tetrafluoromethane under pressure above $-20\,°C$ ($\Delta H = -276$ kJ/mol at 298 K). If the polymerization to polytetrafluoroethylene (PTFE) ($\Delta H = -172$ kJ/mol at 298 K) is uncontrolled, a more strongly exothermic decomposition reaction can occur. Polymerization inhibitors include dipentene [*138-86-3*], α-pinene [*80-56-8*], which are added to liquid tetrafluoroethylene during purification and storage (at $-30\,°C$) [138]. In the United States, the transportation of liquid TFE containing stabilizers is permitted.

In the gas phase at ca. 300 – 500 °C, tetrafluoroethylene dimerizes to perfluorocyclobutane [139]. Above 600 °C, hexafluoropropylene (see Section 4.2) and the highly toxic perfluoroisobutene are formed [139].

Production. Many commercial processes for the production of tetrafluoroethylene are known, e.g., reaction of tetrafluoromethane in an electric arc [140], dechlorination of $CF_2Cl - CF_2Cl$ with a metal [141], and thermal decomposition of trifluoroacetic acid [142]:

$$2\,CF_3COOH \longrightarrow CF_2{=}CF_2 + 2\,HF + 2\,CO_2$$

The two principal commercial methods are pyrolysis of trifluoromethane [143]:

Table 12. Physical properties of fluoroolefins and chlorofluoroolefins

Property	Tetrafluoro-ethylene [*116-14-3*]	Hexafluoro-propene [*116-15-4*]	1,1-Difluoro-ethylene [*75-38-7*]	Fluoro-ethylene [*75-02-5*]	Chlorotri-fluoroethylene [*359-29-5*]	3,3,3-Tri-fluoro-prop-1-ene [*677-21-4*]
M_r	100.02	150.02	64.03	46.04	116.47	96.05
bp, °C	− 75.6	− 29.4	− 82	− 72.2	− 28.36	− 27
mp, °C	− 142.5	− 156.2	− 144	− 160.5	− 158.2	
d_4^{θ}, g/cm^3 (θ, °C)	1.519 (− 76.3)	1.292 (− 29.4)	0.617 (23.6)	0.775 (− 30)	1.51 (− 40)	
Critical temperature, °C	33.3	86.2	30.1	54.7	105.8	107
Critical pressure, MPa	3.82	2.75	4.29	5.43	3.93	4.14
Critical density, g/cm^3	0.58		0.417	0.320	0.55	
Heat of evaporation, J/mol (t, °C)	16 821 (− 75.6)	20 100 (− 29)	13 189 (− 40)	13 494 (− 20)	20 893 (− 28.4)	

$$2\,CHF_3 \xrightarrow{700-800\,^\circ C} CF_2{=}CF_2 + 2\,HF$$

and pyrolysis of chlorodifluoromethane [144]:

$$2\,CHClF_2 \xrightarrow{600-800\,^\circ C} CF_2{=}CF_2 + 2\,HCl$$

In the second method (Fig. 2), the chlorodifluoromethane gas is passed at atmospheric or reduced pressure through a heated platinum, silver, or carbon tubular reactor (a). The 28 % conversion obtained under these conditions is low (yield, 83 %), but can be increased to ca. 65 % with the same yield by adding steam [145]. In modification of this method, chlorodifluoromethane is treated with superheated steam at ca. 700 °C, which results in a conversion rate of 60 – 80 % and a selectivity of 84 – 93 % [146]. The pyrolysis gas is washed with water (b) to cool it and to remove HCl. After being washed with caustic soda (c) and dried with concentrated sulfuric acid (d), the crude product can be stored in liquid or gaseous form (e). It is a complex mixture from which tetrafluoroethylene is separated by distillation in the presence of dipentene or a similar stabilizer to prevent polymerization. An overhead fraction containing inert gases and trifluoromethane is obtained in a low-boiling fractionation column (f) before isolating pure TFE (column g).

The higher-boiling fractions are processed (i) to recover unreacted $CHCl_2F$ and to isolate hexafluoropropene, a byproduct. Extractants such as methanol are added because of the formation of azeotropes during distillation [147]. Methanol also reacts with the toxic perfluoroisobutene, $(CF_3)_2C{=}CF_2$, to give an addition product.

Uses. Currently, tetrafluoroethylene is the most important fluoroolefin; it is used mainly for the production of fluoropolymers (→ Fluoropolymers, Organic). It reacts with perfluoronitrosoalkanes to produce so-called nitroso rubbers [148]. Tetrafluoroethylene is also used in the production of low molecular mass compounds and intermediates, e.g., for the manufacture of iodoperfluoroalkanes (see Section 3.4).

Polytetrafluoroethylene [*9002-84-0*], (PTFE) is a homopolymer sold under the trade names Algoflon (Montefluos), Fluon (ICI), Halon (Allied Chemical), Hostaflon TFE (Hoechst), Teflon PTFE (DuPont), Fluon (Asahi Glass), and Polyflon TFE (Daikin). Copolymers with fluorinated or fluorine-free olefins or vinyl ethers are also commercially available.

4.2. Hexafluoropropene

Properties. Hexafluoropropene (HFP), $CF_3CF{=}CF_2$, a colorless, odorless gas, is nonflammable in air at room temperature. For physical properties, see Table 12. It exhibits no special tendency toward radical homopolymerization [149], but like other fluoroolefins, it reacts readily with nucleophiles [128], [129].

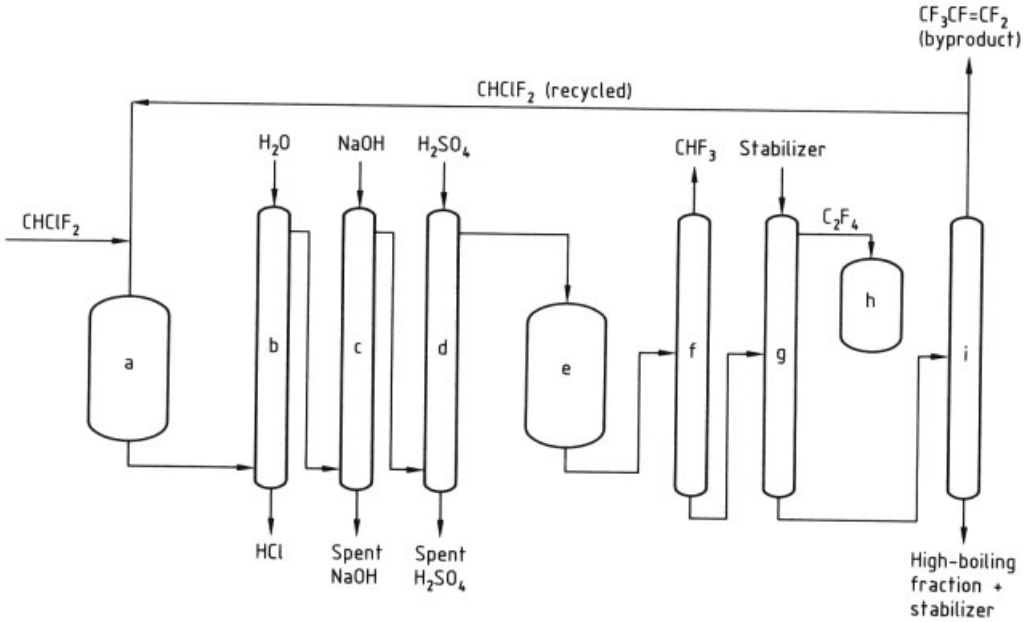

Figure 2. Flow sheet for tetrafluoroethylene production
a) Pyrolysis reactor; b) Quench column (water); c) Wash column (NaOH); d) Drier (conc. H_2SO_4); e) Intermediate storage tank for crude tetrafluoroethylene; f) Fractionation column for low-boiling constituents; g) Product distillation column; h) Tetrafluoroethylene storage tank; i) Fractionation column

Liquid hexafluoropropene has unlimited storage life under pressure at room temperature in steel containers, even without stabilizers. In many countries the liquid can be transported in steel cylinders or tank cars. Hexafluoropropene is toxic (LC_{50} 3000 ppm) and can decompose thermally to form highly toxic perfluoroisobutene.

Production. Hexafluoropropene is produced commercially by temperature-controlled pyrolysis of chlorodifluoromethane (cf. production of tetrafluoroethylene, see Section 4.1) [150]. Hexafluoropropene can also be obtained from tetrafluoroethylene by heating at normal or reduced pressure, preferably in the presence of an inert gas (e.g., CO_2) or water vapor [151]:

$$3\,CF_2{=}CF_2 \xrightarrow{700-900\,^\circ C} 2\,CF_3CF{=}CF_2$$

Uses. An important application of hexafluoropropene is the production of copolymers, e.g., with tetrafluoroethylene or 1,1-difluoroethylene (→ Fluoropolymers, Organic, Chap. 2.3., → Fluoropolymers, Organic, Chap. 3.2.). The versatile epoxide, hexafluoropropylene oxide [*428-59-1*] (see Section 6.1.2), can be obtained from hexafluoropropene by oxidation.

4.3. 1,1-Difluoroethylene

vinylidene fluoride (VDF) $CH_2{=}CF_2$, is a commercially important, partially fluorinated olefin. It is a colorless, flammable gas that undergoes homopolymerization and copolymerization. For physical properties, see Table 12.

Production. Currently, three basic methods are used for the commercial production of vinylidene fluoride:

1) Dechlorination of 1,2-dichloro-1,1-difluoroethane [*1649-08-7*], R 132 b in the gas phase and on a metal catalyst [142], [152]:

$$CH_2ClCClF_2 \xrightarrow[Ni]{500\,^\circ C} CH_2{=}CF_2 + Cl_2$$

2) Dehydrochlorination of 1-chloro-1,1-difluoroethane [*75-68-3*], R 142 b [144], [153]:

$$CH_3CClF_2 \xrightarrow{700-900\,^\circ C} CH_2{=}CF_2 + HCl$$

In the presence of steam the temperature can be reduced to 500 – 650 °C [154].

3) Dehydrofluorination of 1,1,1-trifluoroethane [*420-46-2*], R 143 a [155]:

$$CH_3CF_3 \xrightarrow{1100-1300\,^\circ C} CH_2{=}CF_2 + HF$$

Vinylidene fluoride is transported as a liquid in steel cylinders without stabilizers.

Uses. 1,1-Difluoroethylene is the starting material for the commercially important homopolymer poly(vinylidene fluoride) [*24937-79-9*], (PVDF) (→ Fluoropolymers, Organic, Chap. 2.8.).

Trade Names (PVDF). Kynar (Atochem USA), Florafon (Atochem), Solef (Solvay), Neoflon (Daikin), KF (Kureha Chem.), Vidar (Süddeutsche Kalkstickstoffwerke). Worldwide annual production (1988): 7.4×10^3 t.

The copolymer with hexafluoropropene [*116-15-4*] is marketed as Viton (DuPont) and Fluorel (3M).

4.4. Monofluoroethylene, Monofluoroethylene

vinyl fluoride (VF), $CH_2{=}CHF$, is a colorless, highly flammable gas; up to 0.2 % of polymerization inhibitors are added for stabilization during transport and storage. For physical properties, see Table 12.

Production. In the past, vinyl fluoride was produced by dehydrofluorination of 1,1-difluoroethane [*75-37-6*], obtained in two steps by addition of hydrogen fluoride to acetylene [24, p. 59 – 66], [156]:

$$CH{\equiv}CH + 2\,HF \longrightarrow CH_3CHF_2 \xrightarrow{Cat.} CH_2{=}CHF + HF$$

However, with mercury catalysts, vinyl fluoride can be produced directly from acetylene and hydrogen fluoride [24, p. 59 – 66], [157]:

$$CH{\equiv}CH + HF \xrightarrow{Hg\ Cat.} CH_2{=}CHF$$

The dehydrochlorination of 1-chloro-1-fluoroethane [*1615-75-4*], $CHClFCH_3$, and 1-chloro-2-fluoroethane [*762-50-5*], CH_2FCH_2Cl, are also utilized commercially [158].

Uses. The main use of monofluoroethylene is in the production of poly(vinyl fluoride) [*24981-14-4*] (PVF) (→ Fluoropolymers, Organic, Chap. 2.8., → Fluoropolymers, Organic, Chap. 2.9.).

Trade Names (PVF). Tedlar (DuPont), Dalbon (Diamond Shamrock, USA). Worldwide annual production (1988): 1.6×10^3 t

4.5. 3,3,3-Trifluoropropene

3,3,3-Trifluoropropene, $CF_3CH{=}CH_2$ (TFP), is produced almost exclusively by fluorination and dehalogenation of 1,1,1,3-tetrachloropropane [*1070-78-6*] (TCP), $CCl_3CH_2CH_2Cl$. With sodium fluoride at 400 – 475 °C, trifluoropropene is obtained in a single step, involving chlorine – fluorine exchange and dehydrochlorination [159]. Using hydrogen fluoride and oxygen, the reaction can be carried out at 300 °C over a chromium fluoride catalyst [160]. Liquid-phase fluorination of 1,1,1,3-tetrachloropropane with hydrogen fluoride in the presence of an antimony catalyst yields 1,1,1-trifluoro-3-chloropropane, $CF_3CH_2CH_2Cl$, which gives 3,3,3-trifluoropropene when treated with a base [161]. The conversion is effected in a single operation with a mixture of hydrogen fluoride and a tertiary amine [162].

$$\begin{tikzcd} CCl_3CH_2CH_2Cl \arrow[r] \arrow[rr, bend right] & CF_3CH_2CH_2Cl \arrow[r] & CF_3CH{=}CH_2 \end{tikzcd}$$

A multistep synthesis starts from vinylidene fluoride ($CF_2{=}CH_2$) [163].

Uses. The main use of trifluoropropene is in the production of fluorine-containing silicones used in hydraulic fluids [159], [164].

4.6. 3,3,3-Trifluoro-2-(trifluoromethyl)-prop-1-ene

3,3,3-Trifluoro-1-(trifluoromethyl)prop-1-ene [*382-10-5*], hexafluoroisobutene, (HFIB), $(CF_3)_2C{=}CH_2$, is a colorless, toxic gas (4-h LC_{50} 1700 ppm) with a *bp* of 14.1 °C at 101.3 kPa. A number of processes for its production have been described; the most important use hexafluoroacetone [165] or perfluoroisobutene [166] as the starting material.

$$(CF_3)_2{=}O + CH_2{=}C{=}O \longrightarrow (CF_3)_2C{-}O \atop H_2C{-}C{=}O$$

$$\longrightarrow (CF_3)_2C{=}CH_2 + CO_2$$

$$(CF_3)_2C{=}CF_2 + CH_3OH \longrightarrow (CF_3)_2CHCF_2OCH_3$$

$$\xrightarrow[\text{steps}]{\text{Several}} (CF_3)_2C{=}CH_2$$

Acetic anhydride can be used with hexafluoroacetone instead of ketene. The multistep reaction takes place in a single operation in a copper reactor above 300 °C [167]. Presumably, processes based on the highly toxic perfluoroisobutene are designed to remove it as a harmful byproduct of tetrafluoroethylene or hexafluoropropene production.

Hexafluoroisobutene is used for the production of fluoropolymers. The trade name of the copolymer [*34149-71-8*] with vinylidene fluoride is CM-X (Ausimont) [168].

4.7. Chlorofluoroolefins

Among the numerous known chlorofluoroolefins [133], [134], 1,1-dichloro-2,2-difluoroethylene [*79-35-6*] has some importance as a starting material in the production of methoxyflurane [*76-38-0*], an inhalation anesthetic. However, chlorotrifluoroethylene is the most important member of this class.

Chlorotrifluoroethylene [*359-29-5*] (CTFE), $CF_2{=}CFCl$, a colorless, flammable gas, is less reactive than tetrafluoroethylene. For physical properties, see Table 12. Although chlorotrifluoroethylene is more stable than tetrafluoroethylene, stabilizers such as tributylamine are used during transportation and storage in steel cylinders [169]. Chlorotrifluoroethylene is toxic.

Production. Chlorotrifluoroethylene is produced commercially by dechlorination of 1,1,2-trichloro-1,2,2-trifluoroethane [*76-13-1*], (R 113) with zinc in methanol [170]:

$$CClF_2CCl_2F + Zn \xrightarrow{CH_3OH} CF_2{=}CClF + ZnCl_2$$

An alternative route is dechlorination in the gas phase, e.g., on an aluminum fluoride – nickel phosphate catalyst; this catalyst is highly stable [171].

Uses. Chlorotrifluoroethylene is a starting material for homopolymers and copolymers (PCTFE) (→ Fluoropolymers, Organic, Chap. 2.6., → Fluoropolymers, Organic, Chap. 2.7.). In addition, chlorotrifluoroethylene is an intermediate in the production of the inhalation anesthetic halothane. (2-Chloro-1,1,2-trifluoroethyl)-diethylamine [*357-83-5*], $ClFHCCF_2N(C_2H_5)_2$ (an addition product of chlorotrifluoroethylene with diethylamine), is used as a fluorinating agent to replace hydroxyl groups in steroids and carbohydrates with fluorine [54], [172]. Another use of chlorotrifluoroethylene is in telomerization with carbon tetrachloride or chloroform. The products are stabilized with fluorine or CoF_3 and are used as inert fluids, hydraulic fluids, or lubricants [104], [173].

Trade Names (PCTFE). Aclon (Allied-Signal), Daiflon (Daikin), Kel-F (3M), Voltalef (Atochem). Worldwide annual production (1988): 10^3 t.

5. Fluorinated Alcohols [174]

Primary and secondary alcohols that have fluorine and hydroxyl groups on the same carbon atom are unstable, and readily lose hydrogen fluoride to form carbonyl compounds. Primary and secondary perfluoroalkoxides, however, can be prepared in polar solvents under aprotic conditions from carbonyl compounds and a source of ionic fluoride [175]; typical counterions are alkali-metal, tetraalkylammonium, or tris(dialkylamino)sulfonium cations. The perfluoroalkoxides are moderately nucleophilic and are used in situ to prepare compounds containing perfluoroalkoxy groups (see Section 6.3).

Tris(dialkylamino)sulfonium perfluoroalkoxides are unusually stable and, in some cases, have been isolated as crystalline solids [176]. Tertiary perfluoro alcohols, e.g., perfluoro-*tert*-butanol (1,1,1,3,3,3-hexafluoro-2-trifluoromethyl-2-propanol) [*2378-02-1*] [177] and alcohols containing CH_2 groups between the fluorinated segment and the OH group are also stable. Fluorinated alcohols are more acidic than their nonfluorinated analogues because fluorine is highly electronegative (see Table 13) [178].

Table 13. pK_a values of alcohols [178]

Alcohol	R = H	R = F
CR_3CH_2OH	15.9	12.8
$(CR_3)_2CHOH$	17.1	9.3
$(CR_3)_3COH$	19.2	5.4

Short-Chain Fluoroalcohols. Of the short-chain fluorinated alcohols (C_2–C_4), only 2,2,2-trifluoroethanol [*75-89-8*], *bp* 73.6 °C, and 1,1,1,3,3,3-hexafluoro-2-propanol [*920-66-1*], *bp* 58.2 °C, have achieved commercial importance. The former is prepared by catalytic hydrogenation of trifluoroacetamide [179], trifluoroacetyl chloride [180], or trifluoroacetic acid esters [181], or by reduction of trifluoroacetic acid with metal hydrides [182]. Similar processes are used to prepare the latter from hexafluoroacetone [183]. Because of their strong tendency to form hydrogen bonds, the fluorinated alcohols form stable complexes with hydrogen bond acceptors and are excellent solvents for polar polymers [184]. Ethers derived from these alcohols are used as inhalation anesthetics (see Section 6.3). 2-Aryl-1,1,1,3,3,3-hexafluoro-2-propanols are prepared by alkylation of aromatic compounds with hexafluoroacetone [185]. Tertiary diols, e.g., 1,3-bis(2-hydroxyhexafluoro-2-propyl) benzene [*802-93-7*] are used to prepare fluoroepoxy resins [186].

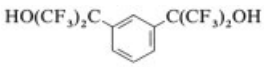

Long-Chain Fluoroalcohols. The unique hydro- and oleophobic properties of long perfluoroalkyl chains lead to many applications of long-chain fluoroalcohols and their derivatives as surfactants, antisoilants, and surface-treatment agents [187]. Structures can be represented by,

$$X-(CF_2)_n(CH_2)_mOH$$

where X is typically F, H, or CH_3, n is 6–12, and m is 1 or 2. The most important compounds are the 1*H*,1*H*-perfluoroalcohols (X = F, $m = 1$), prepared by reduction of the corresponding perfluorinated carboxylic acids [188], and the 1*H*,1*H*,2*H*,2*H*-perfluoroalcohols (X = F, $m = 2$); the latter are prepared by treatment of the corresponding 1-iodo-1*H*,1*H*,2*H*,2*H*-perfluoroalkanes with oleum [189]. Alcohols of the formula $H(CF_2)_nCH_2OH$, where n is an even number, are prepared by telomerization of tetrafluoroethylene with methanol [190].

For use as soil-repellent finishes, the long-chain fluoroalcohols are converted into derivatives to adjust the hydrophobicity, solubility in aqueous and organic solutions, surface retention, and stability for each specific application. These derivatives include esters, phosphates, carboxylates, and polyoxyethylenes. Polymeric alcohols are prepared by esterification with (meth)acryloyl chloride, followed by polymerization.

Trade Names. Teflon, Zepel, and Zonyl (DuPont), Scotchgard (3M), and Oleophobal (Chemische Fabrik Pfersee, Ciba-Geigy).

6. Fluorinated Ethers [191]

6.1. Perfluoroethers

Perfluorinated ethers are less basic than their hydrogen-containing analogues [192]. The saturated aliphatic and cycloaliphatic perfluoroethers are noncombustible, and, with the exception of the perfluorinated epoxides, display high chemical and thermal stability. Other properties of the stable ethers, such as a large difference between the melting and boiling points at ambient pressure, and a low pour point, surface tension, and dielectric constant, are the basis of applications such as dielectric and heat-exchange fluids in, for example, high power transformers. Higher molecular mass perfluoroethers are used as lubricants and hydraulic flu-

ids in extreme service conditions, i.e., at high temperature and/or in a corrosive environment.

6.1.1. Low Molecular Mass Perfluoroethers

The low molecular mass perfluoroethers are usually prepared by electrochemical fluorination of aliphatic ethers, alcohols, or carboxylic acids [193]. Typical empirical formulas are $C_6F_{12}O$, $C_7F_{14}O$, and $C_8F_{16}O$; these ethers often consist of isomer mixtures. Compound FC-75 [*11072-16-5*] (3M), mostly perfluorobutyltetrahydrofurans, is a useful solvent that has been explored as an oxygen-transport agent in artificial blood [194]. Perfluorinated ethers can also be prepared from partially fluorinated or nonfluorinated ethers by fluorination with cobalt fluoride or elemental fluorine under carefully controlled conditions [195].

Trade Names. Fluorinert liquids (3M) and Galden fluorinated fluids (Montedison).

6.1.2. Perfluorinated Epoxides

In contrast to other perfluorinated cyclic ethers, ring strain in perfluorinated epoxides results in high reactivity, making them versatile precursors to other fluorinated compounds [196]. The most important is hexafluoropropylene oxide [*428-59-1*] (HFPO), trifluoro(trifluoromethyl)oxirane [197]. It is prepared from hexafluoropropene by reaction with elemental oxygen [198], by electrochemical oxidation [199], or by reaction with hypochlorites [200] or hydrogen peroxide [201] in alkaline media.

Hexafluoropropylene oxide (*bp* $-27.4\,^{\circ}C$) is stable at room temperature, but decomposes above 150 °C to form trifluoroacetyl fluoride and difluorocarbene [202]. In the presence of strong Brønsted or Lewis acids, such as alumina or aluminum chloride, HFPO undergoes catalytic rearrangement to hexafluoroacetone, constituting a convenient synthesis of this compound [203]. Most significantly, HFPO reacts readily with nucleophiles. Attack usually occurs at the central carbon atom [204], resulting in formation of an acid fluoride by loss of a fluoride ion from the intermediate perfluoroalkoxide, which can react further with HFPO to form higher oligomers. Acid fluorides are precursors to the commercially important perfluorovinyl ethers and higher molecular mass perfluoroethers.

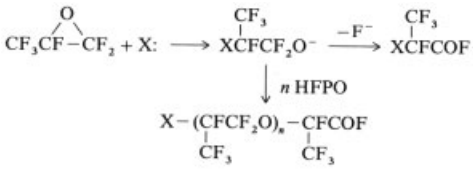

where X = nucleophile, e.g., fluoride ion.

Both tetrafluoroethylene oxide [*694-17-7*] (TFEO), also called tetrafluorooxirane [205], and epoxides of longer chain-length perfluoroolefins [206] are known; however, the former is unstable at room temperature and rearranges to form trifluoroacetyl fluoride, whereas the latter are prepared from inaccessible perfluoroolefins. Only HFPO has achieved commercial significance because it is used for the synthesis of hexafluoroacetone (see page 27), high molecular mass perfluoroethers, Section 6.1.3, and fluorinated vinyl ethers (Section 6.2).

6.1.3. High Molecular Mass Perfluoroethers

High molecular mass perfluorinated ethers are prepared by fluoride-catalyzed oligomerization of HFPO [207]. The resulting terminal acid fluoride group is removed by hydrolysis and decarboxylative fluorination with elemental fluorine. Chemically inert ethers are produced which have the formula,

These ethers are obtained in various molecular mass and viscosity ranges by controlling oligomerization conditions or by partial distillation of the oligomeric mixture. In an alternative method [208], perfluorinated olefins (e.g., tetrafluoroethylene or hexafluoropropene) react photochemically with oxygen to form oligomeric perfluoroethers with terminal acid fluoride groups and peroxide bonds [209]. These end groups and the unstable peroxide linkages are eliminated by fluorination. Perfluoroethers with molecular masses of ca. 500 – 6000 are used as inert fluids, lubricants, and hydraulic fluids in

applications that require resistance to high temperature or strongly corrosive environments.

Trade Names. Krytox (DuPont), Aflunox (SCM Specialty Chemicals), and Fomblin (Montedison).

6.2. Perfluorovinyl Ethers

Perfluorovinyl ethers are comonomers used in the preparation of melt-processable perfluoroplastics, fluorinated elastomers, and perfluoropolymers containing functional groups [210] (→ Fluoropolymers, Organic, Chap. 3.3.). The ethers are synthesized by reaction of a fluorinated alkoxide generated in situ, or of other nucleophiles with HFPO. The resulting acid fluorides are converted to acid salts, which lose carbon dioxide and metal fluoride when heated in an aprotic environment. For example, HFPO is treated with cesium trifluoromethoxide (from cesium fluoride and carbonyl fluoride). Hydrolysis and decarboxylation of the resulting acid fluoride gives perfluoro(methyl vinyl ether) [*1187-93-5*].

$$CF_2O + CsF \longrightarrow CF_3OCs \xrightarrow{\text{HFPO}}$$

$$CF_3OC(CF_3)COF \xrightarrow[\text{2) heat}]{\text{1) } H_2O} CF_3OCF{=}CF_2$$

This compound is a comonomer with tetrafluoroethylene in a perfluorinated elastomer [211].

Perfluoro(propyl vinyl ether) [*1623-05-8*], $CF_3CF_2CF_2OCF{=}CF_2$, prepared from the dimer of HFPO [212], is copolymerized with tetrafluoroethylene to a melt-processable perfluoroplastic.

Perfluorovinyl ethers containing specific functional groups and usually two or more moles of hexafluoropropylene oxide are prepared in a similar fashion. These compounds have the structure $RO(C_3F_6O)CF{=}CF_2$; they are used as functional groups in perfluoroelastomers which can undergo a crosslinking reaction, i.e., cure-site monomers ($R = CF_2CF_2CN$ or pentafluorophenyl) [213] and as monomers that provide the ionic groups in perfluorinated ion-exchange resins ($R = CF_2CF_2SO_2F$, $CF_2CF_2CF_2CO_2CH_3$, or $CF_2CF_2CO_2CH_3$) [214]. Synthesis of the sulfonyl fluoride-substituted vinyl ether is illustrative. Reaction of tetrafluoroethylene with sulfur trioxide gives a sultone, which rearranges to fluorosulfonyldifluoroacetyl fluoride. The anion formed after addition of fluoride ion to this acid fluoride gives a 2 : 1 adduct with HFPO [215]. Pyrolysis of the sodium salt over sodium carbonate gives the functional monomer [216].

$$CF_2{=}CF_2 + SO_3 \longrightarrow \underset{O}{CF_2}-\underset{SO_2}{CF_2} \longrightarrow FO_2SCF_2COF$$

$$FO_2SCF_2CFO + 2\,CH_3C(O)F-CF_2 \xrightarrow{F^-}$$

$$FO_2SCF_2CF_2OC(CF_3)CF_2OC(CF_3)FO \xrightarrow[\text{heat}]{Na_2CO_3}$$

$$FO_2SCF_2CF_2OC(CF_3)CF_2OCF{=}CF_2$$

Perfluorovinyl ethers can also be prepared by deiodofluorination of iodine-containing ethers, $XR_fCF_2OCF_2CF_2I$, where X is hydrogen, halogen, CO_2R, $CONR_2$, SO_2F, or $PO(OR)_2$ (R = alkyl) and R_f is a perfluoroalkyl group. Initially, an organometallic derivative of the iodide is formed by reaction with metals, such as Mg, Cu, Zn, Sn, or Sb. Heating in the absence of a proton source affords the vinyl ethers [217].

6.3. Partially Fluorinated Ethers

Partially fluorinated ethers are synthesized by several methods. Fluoroalkyl alkyl or fluoroalkyl aryl ethers are prepared from alkoxides or phenoxides and fluoroolefins [218]; for example, 1,1,2,2-tetrafluoro-1-methoxyethane [*425-88-7*] is prepared from tetrafluoroethylene and sodium methoxide – methanol. Higher perfluoroolefins can also be employed [219]. The methyl ethers are valuable intermediates, which can be converted to acid fluorides with sulfur trioxide [220] or other Lewis acids [221]. Primary and secondary perfluorinated alkoxides (see Chap. 5) are not sufficiently nucleophilic to react with fluorinated or nonfluorinated olefins. However, they react with a variety of olefins in the presence of halogen to give ethers [222], e.g.,

$$(CF_3)_2CFO^- + CH_2{=}CF_2 + ICl \longrightarrow (CF_3)_2CFOCF_2CH_2I + Cl^-$$

Perfluoroalkoxides react with benzyl halides to give perfluoroalkyl benzyl ethers [223].

In contrast to the physiological inertness of perfluoroethers, some partially fluorinated ethers act as inhalation anesthetics. Many cyclic and acyclic ethers containing varying amounts of fluorine, hydrogen, and sometimes other halogens have been evaluated as anesthetics in search of the ideal combination of potency, volatility, lack of short- and long-term toxicity, and nonflammability [224]. Some ethers have achieved clinical significance (→ Anesthetics, General, Chap. 2.5.).

The first fluorinated ether anesthetic to be introduced clinically was2,2,2-trifluoro-1-vinyloxyethane [*406-90-6*] [trade name: Fluromar (Anaquest)] [225]. It is prepared by addition of 2 mol of trifluoroethanol to acetylene, followed by thermolysis:

$$2\,CF_3CH_2OH + HC{\equiv}CH \longrightarrow CH_3CH(OCH_2CF_3)_2 \longrightarrow CH_2{=}CHOCH_2CF_3$$

The flammability behavior of the compound is not significantly better than that of nonfluorinated anesthetics, and it behaves as a mutagen in the Ames test [226]; therefore, the compound is not currently produced commercially.

2-Chloro-1-(difluoromethoxy)-1,1,2-trifluoroethane [*13838-16-9*] [trade name: Ethrane (Anaquest)], which is nonflammable, is prepared by successive chlorination and fluorination of the hydrocarbon ether [227]. 1-Chloro-1-(difluoromethoxy)-2,2,2-trifluoroethane [*26675-46-7*] [trade name: Forane (Anaquest)] is prepared in a similar fashion [228]. Other possible anesthetics include 2,2-dichloro-1,1-difluoro-1-methoxyethane [*76-38-0*] [229] and 1,1,1,3,3,3-hexafluoro-2-(fluoromethoxy)propane [*28523-86-6*] [230].

7. Fluorinated Ketones and Aldehydes

The carbonyl groups in partially or perfluorinated ketones and aldehydes are electron deficient owing to the inductive effect of the highly electronegative fluorine atom. Compared with their hydrocarbon counterparts, fluorinated ketones and aldehydes are, consequently, much more reactive toward nucleophilic reagents. Stable addition compounds with water, alcohols, and amines are commonly formed. By contrast, fluorinated ketones and aldehydes are relatively unreactive toward electrophilic reagents. An extreme case is hexafluoroacetone, which is not protonated by the FSO_3H-SbF_5 superacid [231].

A wide variety of fluoroalkyl ketones and aldehydes has been synthesized, often by special methods that are unique to polyfluorinated compounds [232]. Of the many known examples, a few fluorinated acetones, aldehydes, and 1,3-diketones are of practical importance.

7.1. Fluoro- and Chlorofluoroacetones

Several fluoro- and chlorofluoroacetones (propanones) are listed in Table 14. Hexafluoroacetone and, to a lesser extent, chloropentafluoroacetone and *sym*-dichlorotetrafluoroacetone are commercially important. Hexafluoroacetone is manufactured by DuPont in the United States and by Hoechst in the Federal Republic of Germany. The two chlorofluoroacetones were made in pilot-plant quantities by Allied in the 1960s [233], but their production has been discontinued.

Hexafluoroacetone, 1,1,1,3,3,3-hexafluoro-2-propanone, is made industrially by the vapor-phase reaction of hexachloroacetone with hydrogen fluoride in the presence of a chromium catalyst [234], [235]. The rearrangement of hexafluoropropylene oxide induced by Lewis acids [203] is an attractive new route that avoids the highly toxic *sym*-dichlorotetrafluoro- and chloropentafluoroacetones. A convenient laboratory synthesis uses hexafluoropropylene as a starting material [236]:

$$2\,CF_3CF{=}CF_2 + \tfrac{1}{4}S_8 \xrightarrow[\text{dimethylformamide}]{KF}$$

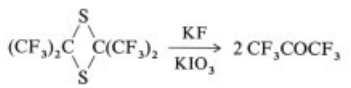

Table 14. Molecular masses and boiling points of fluoro- and chlorofluoropropanones

Name	CAS registry no.	Formula	M_r	*bp*, °C (101.3 kPa)
1,1,1-Trifluoro-2-propanone	[*421-50-1*]	CF_3COCH_3	112.05	21.5 – 22.5
1,1,3,3-Tetrafluoro-2-propanone	[*360-52-1*]	CHF_2COCHF_2	130.05	58
1,1,1,3,3,3-Hexafluoro-2-propanone	[*684-16-2*]	CF_3COCF_3	166.03	− 27.4
1-Chloro-1,1,3,3,3-pentafluoro-2-propanone	[*79-53-8*]	$CClF_2COCF_3$	182.48	7.8
1,3-Dichloro-1,1,3,3-tetrafluoro-2-propanone	[*127-21-9*]	$CClF_2COCClF_2$	198.93	45.2
1,1,3-Trichloro-1,3,3-trifluoro-2-propanone	[*79-52-7*]	$CCl_2FCOCClF_2$	215.38	84.5
1,1,3,3-Tetrachloro-1,3-difluoro-2-propanone	[*79-51-6*]	$CCl_2FCOCCl_2F$	231.83	123.9
1,1,1,3,3-Pentachloro-3-fluoro-2-propanone	[*2378-08-7*]	CCl_3COCCl_2F	248.28	163.7
1,1,1,3,3,3-Hexachloro-2-propanone	[*116-16-5*]	CCl_3COCCl_3	264.73	203.6

Hexafluoroacetone is used mainly for the manufacture of the solvent hexafluoro-2-propanol and high-performance fluoropolymers. For the properties, chemistry, and uses of hexafluoroacetone, see Acetone, Chap. 11.3..

Chloropentafluoroacetone, 1-chloro-1,1,3,3,3-pentafluoro-2-propanone [trade name: 5FK (Allied)] and *sym*-dichlorotetrafluoroacetone, 1,3-dichloro-1,1,3,3-tetrafluoro-2-propanone (4FK, Allied) can be made by the incomplete exchange of chlorine in hexachloroacetone, using hydrogen fluoride and a Cr^{3+} or Cr^{5+} catalyst [234], [235], [237]. Their properties and chemical reactivities are similar to those of hexafluoroacetone [238]. Like hexafluoroacetone, chloropentafluoroacetone and *sym*-dichlorotetrafluoroacetone form stable, acidic hydrates and hemiacetals, e.g., $CClF_2COCF_3 \cdot 3\,H_2O$ [*34202-28-3*], *bp* 105 °C, and $CClF_2COCClF_2 \cdot 2\,H_2O$ [*34202-29-4*], *bp* 106 °C. The hydrates are powerful solvents for acetal resins and various polar polymers [239], [240]. These chlorofluoroacetones and some of their derivatives possess herbicidal or fungicidal activity [241–243] and are useful intermediates for synthesizing repellants for textile fibers [244], [245], specialty polycarbonates [246], [247], and inhalation anesthetics [248].

1,1,1-Trifluoro-2-propanone can be prepared in quantitative yield by the acid hydrolysis of ethyl trifluoroacetoacetate, the ethyl ester of 4,4,4-trifluoro-3-oxobutanoic acid [*372-31-6*], which is made by the alkali-promoted condensation of ethyl trifluoroacetate with ethyl acetate [249]. 1,1,1-Trifluoro-2-propanone is easily made in the laboratory by the reaction of trifluoroacetic acid with methylmagnesium iodide [250]. Its aryl hydrazone derivatives show nematocidal and acaricidal activity [251].

1,1,3,3-Tetrafluoro-2-propanone is made by the acid hydrolysis of the ethyl ester of 2,2,4,4-tetrafluoro-3-oxobutanoic acid, $CHF_2COCF_2\ CO_2C_2H_5$ [249]. The ketone is used as an intermediate in the synthesis of inhalation anesthetics [252], [253].

7.2. Perhaloacetaldehydes

Perchlorfluoroacetaldehydes. Some physical properties of the *perchlorofluoroacetaldehydes* are shown in Table 15. Their chemical reactivities are similar to those of the perchlorofluoroacetones.

Chlorodifluoroacetaldehyde and dichlorofluoroacetaldehyde can be prepared by the lithium aluminum hydride reduction of the corresponding methyl chlorofluoroacetates [254], [255].

Trifluoroacetaldehyde (fluoral) and some of its derivatives have found practical importance as monomers and intermediates for biologically active compounds.

Properties. Trifluoroacetaldehyde is a colorless gas at ambient temperature and pressure. Like hexafluoroacetone, it reacts

Table 15. Molecular mass and boiling points of perchlorofluoroacetaldehydes

Name	Formula	M_r	CAS registry no.	*bp*, °C (101.3 kPa)
Trichloroacetaldehyde	CCl_3CHO	147.38	[*75-87-6*]	97.8
Dichlorofluoroacetaldehyde	CCl_2FCHO	130.93	[*63034-44-6*]	56
Chlorodifluoroacetaldehyde	$CClF_2CHO$	114.48	[*811-96-1*]	17.8
Trifluoroacetaldehyde	CF_3CHO	98.03	[*75-90-1*]	− 18 to − 19∗

∗ At 99.7 kPa

with water to form a stable, solid hydrate – 1,1-dihydroxy-2,2,2-trifluoroethane [*421-53-4*], $CF_3CH(OH)_2$, *mp* 69 – 70 °C. It reacts in a manner similar to hexafluoroacetone with alcohols to give stable hemiacetals: 2,2,2-trifluoro-1-methoxyethanol [*431-46-9*], $CF_3CH(OH)OCH_3$, *bp* 96 – 96.5 °C; and 2,2,2-trifluoro-1-ethoxyethanol [*433-27-2*], $CF_3CH(OH)OC_2H_5$, *bp* 104 – 105 °C (99.3 kPa). Unlike hexafluoroacetone, it readily homopolymerizes upon cationic, anionic, or free-radical initiation [254], [255].

Production. Several laboratory syntheses of trifluoroacetaldehyde have been developed, including reduction of trifluoroacetic acid and its alkyl esters, trifluoroacetic anhydride, and trifluoroacetyl chloride [256]. Trifluoroacetaldehyde can be manufactured on a large scale by the reaction of trichloroacetaldehyde (chloral) with hydrogen fluoride in the presence of chromium catalysts [235]. The product is a $CF_3CHO \cdot HF$ complex, *bp* 38 °C (133.3 kPa), which requires treatment with a hydrogen fluoride acceptor (e.g., sodium fluoride) to give free trifluoroacetaldehyde. Hydrolysis of the inhalation anesthetic Halothane, $CF_3CHBrCl$, by a mixture of 65 % oleum, mercuric oxide, and silver oxide also produces trifluoroacetaldehyde in high yield [257].

Trifluoroacetaldehyde is commercially available as its hydrate or methyl and ethyl hemiacetals, which liberate the pure aldehyde in polyphosphoric acid at 150 – 180 °C. A Hoechst process for the manufacture of its hemiacetals involves treating the product from the gas-phase fluorination of chloral with tetraalkoxyl silanes, or with alcohols and silicon tetrachloride. This process avoids the need to isolate or handle free trifluoroacetaldehyde [258], [259].

$$CCl_3CHO \xrightarrow[CrO_2F_2]{HF} CF_3CHO \cdot HF \xrightarrow{(RO)_4Si} CF_3CH(OH)OR$$

R = alkyl

Uses of Perhaloacetaldehydes. Certain oximes [260] and hydrazones [251] of trifluoroacetaldehyde have insecticidal or acaricidal activity. Its hemiacetal, 2,2,2-trifluoro-1-methoxyethanol, has been used as a starting material for the preparation of fluoroether inhalation anesthetics [261] (see Section 6.3), including isoflurane, $CF_3CHClOCHF_2$ [262], and its isomer $CF_3CHFOCHClF$ [263], [264].

The stereoregularity of perchlorofluoroacetaldehyde polymerizations has been an area of active research, although no commercial uses of the polymers have yet appeared.

Trifluoroacetaldehyde can be homopolymerized to give insoluble crystalline, or soluble, amorphous, polyoxymethylene polymers depending upon the polymerization conditions [254], [255]. This is in contrast with trichloroacetaldehyde which can only be polymerized to a crystalline, apparently isotactic polymer. Copolymers of perhaloacetaldehydes have been prepared [254], [255].

7.3. Fluorinated 1,3-Diketones

Fluorinated 1,3-diketones in which the two carbonyl groups are separated by a methylene or methine group form complexes with a wide variety of metal ions. This property is the basis of their utility in chromatographic analysis of metals, laser technology, NMR spectroscopy, and hydrometallurgical separations. The properties, preparation, and uses of fluorinated 1,3-diketones and theirmetal complexes have been extensively reviewed [265–267].

Properties and Production. Some physical properties of the industrially important fluorinated 1,3-diketones are shown in Table 16. These compounds are considerably more acidic than their nonfluorinated analogues, i.e., 1,3-pentanedione [*123-54-6*], $CH_3COCH_2COCH_3$ ($pK_a = 8.9$), $CF_3COCH_2COCH_3$ ($pK_a = 6.7$), and $CF_3COCH_2COCF_3$ ($pK_a = 4.6$) [267]. Fluorinated 1,3-diketones have a high enolic content, typically 92 – 100 %, in comparison with ca. 80 % for 1,3-pentanedione. Their enolic protons can be readily replaced by metals or metal salts to form 1,3-diketonates of the type

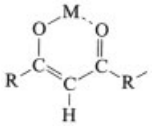

Fluorinated 1,3-diketones form hydrates with water and hemiketals with shorter alcohols. They are usually obtained by a Claisen condensation of a fluorinated carboxylic acid ester with a ketone, using a strong base such as a sodium alkoxide or sodium hydride.

Uses. The 1,3-diketones $RCOCH_2COR'$ ($R = CF_3$, n-C_3F_7 and $R' = t$-C_4H_9, CF_3) that form volatile, thermally and hydrolytically stable complexes are useful for liquid or gas chromatographic analysis of metals. These diketones, especially 1,1,1,2,2,3,3-heptafluoro-7,7-dimethyl-4,6-octanedione are widely used for the extraction of metal ions from aqueous solutions at variable pH [268].

4,4,4-Trifluoromethyl- 1-(2-thienyl)-1,3-butanedione is especially suited for the analysis of uranium and other radioactive elements. The *tetrakis*-chelates of 4,4,4-trifluoro-1-phenyl-1,3-butanedione with rare-earth elements are potential laser materials.

The paramagnetic lanthanide complexes of heptafluoro-7,7-dimethyl-4,6-octanedione have gained considerable importance as NMR shift reagents for simplifying the interpretation of complex NMR spectra [269]; the europium complex is the most widely used.

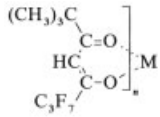

M = Ag; $n = 1$
M = Dy, Eu, Gd, Ho, Pr, Yb; $n = 3$

The chiral-shift reagents derived from 3-heptafluorobutyryl-(+)- or -(−)-camphor and 3-trifluoroacetyl-(+)- or -(−)-camphor are very useful for the NMR assay of enantiomeric purity in solution [269].

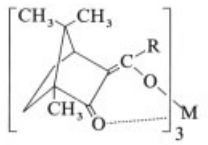

R = CF_3, n-C_3F_7
M = Eu, Pr, Yb

8. Fluorinated Carboxylic Acids and Fluorinated Alkanesulfonic Acids

8.1. Fluorinated Carboxylic Acids

8.1.1. FluorinatedAcetic Acids

Fluorination increases the strength of acetic acid as seen in the pK_a values of monofluoroacetic acid (2.66), difluoroacetic acid (1.24), and trifluoroacetic acid (0.23) compared to the pK_a of 4.74 for acetic acid [270].

Production. Trifluoroacetic acid has been prepared by the electrochemical fluorination of acetyl chloride or acetic anhydride in anhydrous hydrogen fluoride using the Simons process (Section 2.1) followed by hydrolysis of the resulting trifluoroacetyl fluoride. The yield is excellent (> 90 %) [271], [272]. The Phillips Petroleum electrochemical process (Section 2.1) employs acetyl fluoride as feed to produce trifluoroacetyl fluoride along with mono and difluoroacetyl fluoride [273]. Sulfur trioxide treatment of CF_3CCl_3, obtained by isomerization of Freon 113, $CF_2ClCFCl_2$, yields CF_3COCl [274]. The acid is purified by hydrolysis of the acyl halide with alkali, followed by acidification and distillation.

Uses. Most uses of fluorinated acetic acids are confined to trifluoroacetic acid, its anhydride, and its derivative, 1,1,1-trifluoroethanol.

Table 16. Molecular masses and boiling points of fluorinated 1,3-diketones

Name	CAS registry no.	Formula	M_r	*bp*, °C (kPa)
1,1,1-Trifluoro-2,4-pentanedione	[*367-57-7*]	$CF_3COCH_2COCH_3$	154.09	105 – 107 (101.3)
1,1,1,5,5,5-Hexafluoro-2,4-pentanedione	[*1522-22-1*]	$CF_3COCH_2COCF_3$	208.06	70 – 71 (101.3)
1,1,1-Trifluoro-5,5-dimethyl-2,4-hexanedione	[*22767-90-4*]	$CF_3COCH_2COC(CH_3)_3$	196.17	138 – 141 (101.3)
1,1,1,2,2,3,3-Heptafluoro-7,7-dimethyl-4,6-octanedione	[*17587-22-3*]	$CF_3CF_2CF_2COCH_2COC(CH_3)_3$	296.19	46 – 47 (0.67)
4,4,4-Trifluoro-1-phenyl-1,3-butanedione	[*326-06-7*]	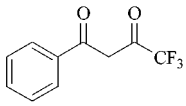	216.16	224 (101.3) [a]
4,4,4-Trifluoro-1-(2-thienyl)-1,3-butanedione	[*326-91-0*]	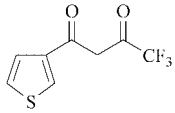	222.18	96 – 98 (1.07) [b]
3-(Trifluoroacetyl)camphor, 1,7,7-trimethyl-3-(trifluoroacetyl)-bicyclo[2.2.1]heptan-2-one	[*51800-98-7*]	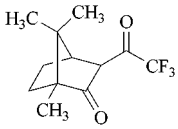	248.25	100 – 101 (2.13)
3-(Heptafluorobutyryl)camphor, 3-(2,2,3,3,4,4,4-heptafluoro-1-oxo-butyl)-bicyclo[2.2.1]heptan-2-one	[*51800-99-8*]	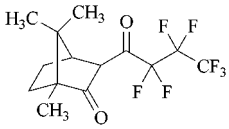	348.26	60 – 70 (0.03)

[a] *mp* 38 – 40 °C. [b] *mp* 42 – 43 °C.

Monofluoroacetic acid derivatives (salts, esters, amides, and alcohols) are toxic because they are metabolized to fluorocitric acid which inhibits respiration (see Section 13.5) [275]. ω-Fluoro acids of the formula $F(CH_2)_nCOOH$, where n is an odd number, are extremely toxic, because of degradation in vivo to monofluoroacetic acid and finally to fluorocitric acid [275]. Sodium monofluoroacetate has been used as a rodenticide, but is now banned.

8.1.2. Long-Chain Perfluorocarboxylic Acids

Properties. The boiling points and densities of straight-chain perfluorocarboxylic acids are shown in Table 17.

Production. Long-chain perfluorocarboxylic acids are prepared by the Simons electrochemical fluorination (see Section 2.1.) of the corresponding acyl halide:

Table 17. Boiling points and densities of perfluorocarboxylic acids

Acid	CAS registry no.	Formula	*bp*, °C (kPa)	d_4^{20}, g/cm^3
Perfluoroacetic	[*76-05-1*]	CF_3CO_2H	72.4	1.489
Perfluoropropionic	[*422-64-0*]	$C_2F_5CO_2H$	96	1.561
Perfluorobutyric	[*375-22-4*]	$C_3F_7CO_2H$	120	1.651
Perfluorovaleric	[*2706-90-3*]	$C_4F_9CO_2H$	130	
Perfluorocaproic	[*307-24-4*]	$C_5F_{11}CO_2H$	157	1.762
Perfluoroheptanoic	[*375-85-9*]	$C_6F_{13}CO_2H$	175	1.792
Perfluorocaprylic	[*335-67-1*]	$C_7F_{15}CO_2H$	189	
Perfluorononanoic	[*375-95-1*]	$C_8F_{17}CO_2H$	110 (2.1)	
Perfluorocapric	[*335-76-2*]	$C_9F_{19}CO_2H$	121 (2.0	
Perfluoroundecanoic	[*2058-94-8*]	$C_{10}F_{21}CO_2H$	245	
Perfluorododecanoic	[*307-55-1*]	$C_{11}F_{23}CO_2H$	270	

$$C_nH_{2n+1}COF + (2n+1)\,HF \longrightarrow C_nF_{2n+1}COF + (2n+1)H_2$$

The acids are obtained by hydrolysis of the perfluoroacyl fluoride, followed by distillation. Some carbon–carbon bond scission occurs to form lower homologous acids along with inert fluorocarbons and cyclic ethers. The acid yield decreases with increasing chain length; for example, perfluorobutyric acid yields are ca. 36 % compared to ca. 20 % for the industrially important perfluorooctanoic acid [271], [272]. Trifluoroacetic acid is also prepared by the Phillips electrochemical method, employing a $KF \cdot HF$ molten salt electrolyte [273]. Acids with higher boiling points are not easily or efficiently prepared by the Phillips process.

Perfluorocarboxylic acids are also prepared by nonelectrochemical methods. Treatment of perfluoroalkyl iodides (R_fI) with sulfur trioxide [276] or chlorosulfonic acid [277] gives the carboxylic acid in good yield. Another method employs fuming sulfuric acid (oleum) [278]. Another process involves the preparation of perfluorotetrahydroalkyl iodides, $R_fCH_2CH_2I$, obtained by the free-radical addition of ethylene to perfluoroalkyl iodides, followed by dehydroiodination and oxidation by dichromate [279] or ozonolysis [280]:

$$C_nF_{2n+1}I + CH_2{=}CH_2 \longrightarrow C_nF_{2n+1}CH_2CH_2I \xrightarrow{-HI}$$
$$C_nF_{2n+1}CH{=}CH_2 \xrightarrow[\text{or } O_3]{Cr_2O_7^{2-}} C_nF_{2n+1}COOH$$

This process produces carboxylic acids having one more carbon atom than the starting perfluoroalkyl iodide, in contrast to the electrochemical and oleum routes, which produce acids containing the same number of carbon atoms as the original acyl halide or telomeric iodide. In all processes utilizing fluorinated iodine-containing telomers, it is important to find uses for the telomers and to recover the expensive iodine.

Polyfluoroalkoxyacyl fluorides of the type

$$R_fCF_2O(\overset{\overset{\displaystyle CF_3}{|}}{C}FCF_2O)_nCF\overset{\overset{\displaystyle CF_3}{|}}{C}FO$$

are prepared by the addition of perfluoroacyl fluorides and hexafluoropropylene oxide catalyzed by alkali-metal fluorides [281]. Acids, salts, or esters are obtained by hydrolysis, neutralization, or esterification, respectively.

Addition reactions of perfluorodiacyl fluorides with hexafluoropropylene oxide [282] give ether-containing diacyl fluorides, such as

$$FCO\overset{\overset{\displaystyle CF_3}{|}}{C}FO(CF_2)_nO(\overset{\overset{\displaystyle CF_3}{|}}{C}FCF_2O)_m\overset{\overset{\displaystyle CF_3}{|}}{C}FCOF$$

where n is usually 2 – 4, and at least two units are derived from hexafluoropropylene oxide (i.e., $m \geq 2$) [282]. The addition reaction may occur at one or both acyl groups of the starting diacyl fluoride. Selectivity can be maintained by esterifying one of the acyl fluoride groups. Thus, $CH_3OCO(CF_2)_2COF$, prepared by addition of methanol to perfluoro-γ-butyrolactone or perfluorosuccinyl fluoride, reacts with hexafluoropropylene oxide to give ester acyl fluorides of the formula

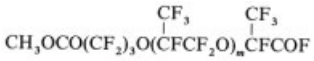

[283], [284]. The acyl halides are converted to acids or salts by hydrolysis or neutralization, respectively.

Uses. Long-chain perfluoroalkanecarboxylic acids and their salts are surface-active chemicals (surfactants), which greatly reduce the surface tension (surface energy) of water, aqueous solutions, and organic liquids even at low concentrations. These acids (C_6–C_{12}) and derivatives are used as wetting, dispersing, emulsifying, and foaming agents.

Ammonium perfluorooctanoate (FC-143, 3M) is used as an emulsifier in the polymerization of fluorinated monomers, especially tetrafluoroethylene. It has exceptional chemical stability and lowers the surface tension of water to ca. 3 μN · m at 0.5 wt %. In contrast with hydrocarbon emulsifiers, ammonium perfluorooctanoate does not interfere with the emulsion polymerization of tetrafluoroethylene.

Trade Names. Fluorad FC-26, 126, 143 (3M), Fluorowet CP (Hoechst), RM 350, 370 (Rimar), and Surflon S-111P (Asahi Glass).

8.1.3. Fluorinated Dicarboxylic Acids

Properties. The boiling points and densities of perfluorodicarboxylic acids are shown in Table 18.

Production. α,ω-Perfluoroalkanedicarboxylic acids are prepared by the electrochemical method, followed by hydrolysis, acidification, and extraction. Other methods include the oxidation of the appropriate chlorofluoroolefin or perfluoroolefin. Perfluoroglutaric acid is prepared from hexachlorocyclopentadiene by halogen exchange, followed by oxidation and acidification [285]. Perfluoroadipic acid is prepared by an analogous method from hexachlorobenzene [286]. The cyclic anhydrides from perfluorosuccinic and perfluoroglutaric acid are prepared by dehydration of the acids with phosphorus pentoxide. They are used to prepare the corresponding alcohols by reduction.

8.1.4. Tetrafluoroethylene – Perfluorovinyl Ether Copolymers with Carboxylic Acid Groups

Ion-exchange membranes, used in fuel cells and chloralkali production, are copolymers of tetrafluoroethylene and perfluorovinyl ethers that contain esters or other acid precursor groups [287]. These membranes have excellent thermal and chemical resistance to hot concentrated alkali (up to 40 %).

Production. The vinyl ethers are prepared by the reaction of hexafluoropropylene oxide and methyl-3-fluorocarbonyl perfluoropropionate [285], followed by pyrolysis to give:,

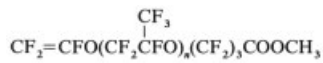

Copolymerization with tetrafluoroethylene followed by saponification produces a polymer with terminal carboxylic acid groups.

Trade Names. Flemion (Asahi Glass), Nafion (DuPont), Neosepta (Tokuyama Soda), and Aciplex (Asahi Chemical).

8.2. Fluorinated Alkanesulfonic Acids

8.2.1. Perfluoroalkanesulfonic Acids

Properties. The first member of the series, trifluoromethanesulfonic acid, was reported in 1954 [288]. Perfluoroalkanesulfonic acids are among the strongest acids known. Conductivity measurements in acetic acid show that the acid strength of trifluoromethanesulfonic acid is comparable to that of fluorosulfonic and perchloric acids [289]. Boiling points are listed in Table 19. Because of their ability to lower surface energy, the longer-chain perfluoroalkanesulfonic acids and sulfonyl fluoride derivatives have found utility as surface-active agents and water repellents and for antisoiling treatment of textiles and fabrics.

Table 18. Physical properties of perfluorodicarboxylic acids

Acid	CAS registry no.	Formula	*mp*, °C	*bp*, °C (kPa)
Perfluoromalonic	[*1514-85-8*]	$CF_2(CO_2H)_2$	117 – 118	
Perfluorosuccinic	[*377-38-8*]	$(CF_2)_2(CO_2H)_2$	86 – 87	150 (2.0)
Perfluoroglutaric	[*376-73-8*]	$(CF_2)_3(CO_2H)_2$	88	134 – 138 (0.4)
Perfluoroadipic	[*336-08-3*]	$(CF_2)_4(CO_2H)_2$	132 – 134	138 (0.5)
Perfluorosuberic	[*678-45-5*]	$(CF_2)_6(CO_2H)_2$	154 – 158	

Table 19. Boiling points of perfluoroalkanesulfonic acids

Acid Formula	CAS registry no.	*bp*, °C (kPa)
CF_3SO_3H	[*1493-13-6*]	60 (0.4)
$C_2F_5SO_3H$	[*354-88-1*]	81 (0.29)
n-$C_4F_9SO_3H$	[*59933-66-3*]	76 – 84 (0.13)
n-$C_5F_{11}SO_3H$	[*3872-25-1*]	110 (0.67)*
n-$C_6F_{13}SO_3H$	[*355-46-4*]	95 (0.46)
n-$C_8F_{17}SO_3H$	[*1763-23-1*]	133 (0.8)

* n-$C_5F_{11}SO_3H \cdot H_2O$

Production. Perfluoroalkanesulfonyl fluorides are usually made by the Simons electrochemical fluorination process (see Section 2.1), in which a hydrocarbon sulfonyl fluoride is electrolyzed in anhydrous hydrogen fluoride at nickel electrodes:

$$C_nH_{2n+1}SO_2F + (2n+1)HF \longrightarrow C_nF_{2n+1}SO_2F + (2n+1)H_2$$

The electrochemical yield is excellent for the first member of the series and decreases progressively with the increasing length of the carbon chain; the yield for octanesulfonyl fluoride is ca. 40 % [271], [290]. Alkaline hydrolysis of perfluoroalkanesulfonyl fluorides gives the corresponding salts, which when acidified and distilled from concentrated sulfuric acid yield the anhydrous sulfonic acids [288].

A nonelectrochemical method for the preparation of trifluoromethanesulfonic acid derivatives is shown below:

$$CF_3SSCF_3 \xrightarrow{Cl_2} 2\,CF_3SCl \xrightarrow[H_2O]{Cl_2} 2\,CF_3SO_2Cl$$

Alkaline hydrolysis, followed by acidification of the sulfonate salt, gives the acid [291].

Uses. Short-chain perfluoroalkanesulfonyl fluorides are used to prepare sulfonamides, which are employed as plant growth regulators and herbicides. Trifluoromethanesulfonic acid is used as an esterification catalyst (FC-24, 3M), and the lithium salt has been investigated as a fuel cell and battery electrolyte (FC-124, 3M). Perfluorobutanesulfonate salts are used as antistatic agents [292], [293]. The higher homologues exhibit good surfactant properties; the derivatives of perfluorohexanesulfonyl fluoride are employed in fire extinguishing formulations. Useful derivatives containing the sulfonamido group can be prepared by reaction of the sulfonyl fluoride with a diamine, followed by quaternization with alkylating agents such as methyl iodide to give a cationic surfactant. As an example, reaction of perfluorohexanesulfonyl fluoride with 3-dimethylaminopropylamine gives

$$C_6F_{13}SO_2NH(CH_2)_3N(CH_3)_3{}^+I^-.$$

The properties related to the low surface energy of the perfluorooctanesulfonyl fluoride, $C_8F_{17}SO_2F$, are utilized in many derivatives. These derivatives include alcohols and their acrylate and methacrylate esters; they are used as comonomers in polymers that impart oil-, water-, and soil-repellent properties to porous substrates such as paper and textiles [294], [295]. A typical reaction sequence for the synthesis of a perfluoroalkanesulfonamide acrylate monomer is shown below:

$$C_8F_{17}SO_2F + RCH_2NH_2 \longrightarrow C_8F_{17}SO_2NHCH_2R$$

$$\xrightarrow{ClCH_2CH_2OH,\ OH^-} C_8F_{17}SO_2N(CH_2R)CH_2CH_2OH$$

$$\xrightarrow{CH_2{=}CHCO_2H} C_8F_{17}SO_2N(CH_2R)CH_2CH_2OCOCH{=}CH_2$$

Perfluoroalkanesulfonamido alcohols are also used as mold-release agents. Esterification with phosphoric acid gives phosphate ester salts of the type

$$[R_fSO_2N(C_nH_{2n+1})CH_2CH_2O]_mPO(ONH_4)_l$$

R_f = perfluoroalkyl; $l = 1, 2$; $m = 1, 2$; and $n = 1, 2$

which are useful oil repellents for paper products [296].

Trade Names. Paper Treatment FC-807, 808 (3M), Textile Treatment Dic-Guard (Dainippon Ink), and Scotchgard (3M).

8.2.2. Fluorinated Alkanedisulfonic Acids

α,ω-Perfluoroalkanedisulfonyl fluorides are prepared by the electrochemical fluorination of hydrocarbon disulfonyl fluorides. The corresponding acids are prepared by alkaline hydrolysis of the fluoride and acidification, or by aqueous permanganate oxidation of the disulfone, $CH_3SO_2(CF_2CF_2)_nSO_2CH_3$ [297]. Distillation of the acids from phosphorus pentoxide yields the cyclic anhydrides [298].

8.2.3. Tetrafluoroethylene – Perfluorovinyl Ether Copolymers with Sulfonic Acid Groups

Ion-exchange membranes are prepared by the copolymerization of tetrafluoroethylene with perfluorovinyl ethers containing sulfonyl halide groups, followed by hydrolysis to yield sulfonic acids. These membranes have excellent chemical and thermal properties similar to those of ion-exchange membranes with terminal carboxylic acid groups (see Section 8.1.4) [299–301].

Production. These ethers are prepared by the condensation of hexafluoropropylene oxide with fluorosulfonyldifluoroacetyl fluoride, FSO_2CF_2COF, which is prepared from C_2F_4 and SO_3, followed by isomerization [302], [303]. Condensation of hexafluoropropylene oxide with γ-fluorosulfonylperfluoroalkyl carbonyl fluorides, $FSO_2(CF_2)_nCOF$ (prepared by electrochemical fluorination of the respective aliphatic sulfone [304], [305]) gives perfluoroether fluorosulfonyl acyl fluorides, e.g., [306]

$$FCOCF(CF_3)O(CF_2CF(CF_3)O)_m(CF_2)_{n+1}SO_2F$$

Subsequent conversion to the vinyl ether

$$CF_2{=}CFO(CF_2CF(CF_3)O)_m(CF_2)_{n+1}SO_2F$$

followed by copolymerization with tetrafluoroethylene gives polymers with fluorosulfonyl side chains, which are hydrolyzed to sulfonate group side-chains [307].

9. Fluorinated Tertiary Amines

Physical Properties. Relative to their molecular mass, perfluoroalkyl-*tert*-amines, like the perfluoroethers, have low boiling points and low freezing or pour points, as shown in Table 20. Low polarity and weak intermolecular forces are responsible for other unusually low values for properties such as viscosity, solubility, heat of vaporization, refractive index, dielectric constant, and surface tension [308]. Some perfluorobis(dialkylaminoalkyl) ethers have even lower pour points and liquid ranges as broad as 250 °C. These ethers exhibit increased internal flexibility through the combined effect of the nitrogen and oxygen atoms [311].

Chemical Properties. Perfluorinated *tert*-amines are chemically inert and thermally stable [308], [312]. The electron-withdrawing nature of the perfluoroalkyl substituents deprives the nitrogen atom of its basic character and reactivity. Fluorinated *tert*-amines do not form salts or complexes with strong acids and are not attacked by most oxidizing or reducing agents. With aluminum chloride they form chlorinated imines [313]. Because of their nonpolar nature, fluorinated *tert*-amines are poor solvents and are immiscible with water and alcohols [308]. Gases such as oxygen, nitrogen, and carbon dioxide have unusually high solubility in perfluorinated *tert*-amines. For example, perfluorotributylamine dissolves 40 vol % of oxygen at ambient conditions and has been used in artificial blood as an effective oxygen-transport medium [313].

Production. Electrochemical fluorination via the Simons process (see Section 2.1) is the preferred route to fluorinated tertiary alkylamines. The hydrogen atoms are completely

Table 20. Physical properties of fluorinated *tert*-amines [308–310]

Compound	CAS registry no.	Formula	M_r	*bp*, °C	Pour point, °C	d_4^{25}
Perfluorotrimethylamine	[*432-03-1*]	$N(CF_3)_3$	221	−11		
Perfluorotriethylamine	[*359-70-6*]	$N(C_2F_5)_3$	371	69		1.74
Perfluorotripropylamine	[*338-83-0*]	$N(C_3F_7)_3$	521	130	−52	1.82
Perfluorotributylamine	[*311-89-7*]	$N(C_4F_9)_3$	671	178	−50	1.88
Perfluorotriamylamine	[*338-84-1*]	$N(C_5F_{11})_3$	821	215	−25	1.93
Perfluorotrihexylamine	[*432-08-6*]	$N(C_6F_{13})_3$	971	256	33∗	1.90∗∗

∗ Freezing point. ∗∗ d_4^{35}

replaced by fluorine atoms. Perfluorotributylamine, for example, is synthesized as follows:,

$$(C_4H_9)_3N + 27\,HF \xrightarrow{e^-} (C_4F_9)_3N + 27\,H_2$$

The crude product contains a significant amount of perfluorinated isomers and cleavage products because of molecular rearrangement during electrolysis; it is purified by fractional distillation and treatment with base.

Uses. The combination of unusual physical and chemical properties, excellent dielectric properties, nonflammability, and lack of toxicity make the perfluorinated *tert*-amines useful for many fluid applications that involve direct contact with sensitive materials [308], [313]. The electronic industry relies heavily on these fluids in reliability testing of electronic components, as direct-contact coolants for integrated circuits, and as heating media in vapor-phase reflow soldering. Perfluorinated tributylamine, triamylamine, and trihexylamine are the main constituents of Fluorinert electronic liquids FC-43, FC-70, and FC-71 (3M).

Information on the preparation and utility of other nitrogen-containing fluoroaliphatic compounds can be found in the general references [1–15].

10. Aromatic Compounds with Fluorinated Side-Chains

The first aromatic compound with a fluorinated side-chain, benzotrifluoride [*98-08-8*], trifluoromethylbenzene, was synthesized in 1898 [314]. The perfluoroalkyl substituents of aromatic compounds are *meta*-directing; this influence can, of cause, be overcome by a stronger *ortho*- or *para*-directing substituent. These compounds are usually prepared from iodoperfluoroalkanes and halogenated aromatic compounds in the presence of a copper catalyst [315–317]. Except for the benzotrifluorides, aromatic compounds with fluorinated side-chains have only scientific interest [318]. Araliphatic compounds with one or two trifluoromethyl substituents on the benzene ring gained commercial importance in the early 1930s for two reasons: 1) the recognition of the advantageous properties of aromatic dyes with CF_3 substituents, and 2) the development of an economical production process [4]. Since then, their importance in the production of dyes, pharmaceuticals, and pesticides has increased.

10.1. Properties

The physical properties of benzotrifluorides are shown in Table 21.

If no other substituents are present in the benzene ring, the trifluoromethyl group is thermally stable up to 350 °C and resistant to bases up to 130 °C. It is inert toward reducing agents [319], [320] and inhibits oxidation of the benzene ring. However, in the presence of aluminum chloride the trifluoromethyl group undergoes chlorolysis to produce a trichloromethyl group [321], and acid hydrolysis forms a carboxy group [322]. Substituents such as amino or hydroxyl groups destabilize the trifluoromethyl group [323].

The characteristic reaction of benzotrifluorides is electrophilic substitution of the benzene ring; chlorination and nitration are commercially important. *Chlorination* of benzotrifluoride at 65 °C with $FeCl_3$ as a catalyst yields 83 % 3-chlorobenzotrifluoride [*98-15-7*] [324]; chlorination of 4-chlorobenzotrifluoride gives 3,4-dichlorobenzotrifluoride [*328-84-7*].

Table 21. Physical properties of benzotrifluorides

Compound	CAS registry number	Empirical formula	M_r	*bp*, °C	d_4^{θ} (θ, °C)	n_D^{θ} (θ, °C)
Benzotrifluoride	[*98-08-8*]	$C_7H_5F_3$	146.11	102.03	1.188 (20)	1.4114 (25)
2-Chlorobenzotrifluoride	[*88-16-4*]	$C_7H_4ClF_3$	180.56	152.5	1.367 (20)	1.4550 (20)
4-Chlorobenzotrifluoride	[*98-56-6*]	$C_7H_4ClF_3$	180.56	140	1.35 (15)	1.4444 (25)
2,4-Dichlorobenzotrifluoride	[*320-60-5*]	$C_7H_3Cl_2F_3$	215.0	117 – 118	1.377 (20)	1.4802 (20)
1,3-Bis(trifluoromethyl)benzene	[*402-31-3*]	$C_8H_4F_6$	214.11	116.1	1.379 (25)	1.3791 (20)
1,4-Bis(trifluoromethyl)benzene	[*433-19-2*]	$C_8H_4F_6$	214.11	117.1	1.381 (25)	1.3792 (20)
3-Trifluoromethylbenzoyl fluoride	[*328-99-4*]	$C_8H_4F_4O$	192.11	159 – 163		1.4350 (20)

Nitration of benzotrifluoride results in a 6 : 3 : 91 mixture of 2-nitro- [*384-22-5*], 4-nitro- [*402-54-0*], and 3-nitrobenzotrifluoride [*98-46-4*] [325]. Other commercially important nitrations are the conversion of 2-chlorobenzotrifluoride [*88-16-4*] to 2-chloro-5-nitrobenzotrifluoride [*777-37-7*] and the conversion of 4-chlorobenzotrifluoride to yield 99 % 4-chloro-3-nitrobenzotrifluoride [*121-17-5*] or 4-chloro-3,5-dinitrobenzotrifluoride [*393-75-9*] [326]. These derivatives are reduced to amines or are further processed.

10.2. Production

Benzotrifluorides are prepared on a laboratory scale by the reaction of aromatic compounds with iodotrifluoromethane [315]; by the reaction of aromatic carboxylic acids and their derivatives with sulfur tetrafluoride [51–53]; or by chlorine – fluorine exchange in trichloromethyl aromatic compounds with metal fluorides [327].

In the commercial process, known since 1931, chlorine is exchanged for fluorine with use of anhydrous hydrogen fluoride in the presence or absence of catalyst [328]:

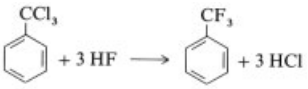

This reaction can be carried out as a batch process in autoclaves or continuously in a series of autoclaves [329] (Fig. 3) or tubular reactors [330]. Typical conditions for the production of benzotrifluoride are a temperature of 80 – 110 °C, pressure of 1.2 – 1.4 MPa, and a molar ratio of HF : benzotrichloride of 4 : 1. A yield of 70 % is obtained within 3 – 4 h [331]. In continuous processing, a yield of over 90 % is obtained in a nickel flow tube with a residence time of 1 h at 90 – 130 °C and at 3 – 5 MPa [332]. Yields are increased by using additives such as hexamethylenetetraamine [333] or by employing chlorine – fluorine exchange in the gas phase in the presence of a transition metal – aluminum oxide catalyst [334]. Corrosion is reduced by lowering the reaction temperature (< 60 °C) and adding iron or iron compounds [335], [336].

This method is used to produce benzotrifluoride, 2-chlorobenzotrifluoride, 4-chlorobenzotrifluoride, 2,4-dichlorobenzotrifluoride, 1,3-bis(trifluoromethyl)benzene, 1,4-bis(trifluoromethyl)benzene, 3-(trifluoromethyl)benzoyl fluoride, and 4-(trifluoromethyl)benzoyl fluoride [*368-94-5*] from the corresponding trichloromethyl compounds.

The trifluoromethyl group can also be introduced into aromatic compounds by the Friedel – Crafts reaction with carbon tetrachloride in the presence of hydrogen fluoride [337].

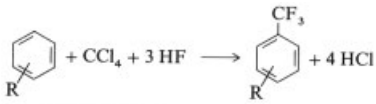

R = alkyl, aryl, halogen

10.3. Uses

Benzotrifluoride and 4-chlorobenzotrifluoride are key intermediates for the synthesis of dyes, pharmaceuticals, and pesticides.

Dyes. Trifluoromethyl aromatic compounds were first used in dyes [4] and are still important intermediates for azo, anthraquinone, and triphenylmethane dyes. The strongly electronegative

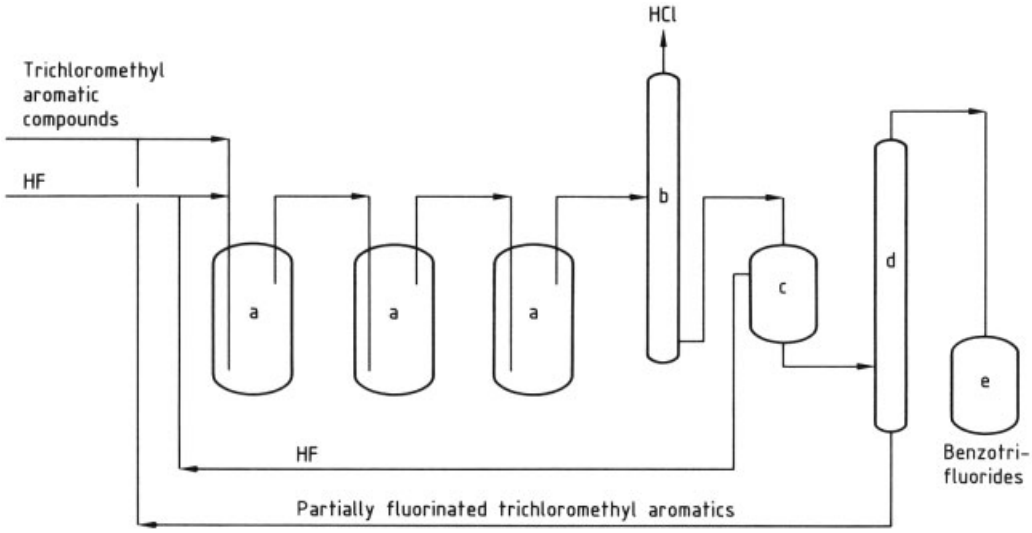

Figure 3. Production of benzotrifluorides in a series of autoclaves [329]
a) Reactor; b) Pressure distillation column; c) Separator; d) Product distillation column; e) Product storage tank

trifluoromethyl group improves color clarity and fastness to light and washing; it also shifts light absorption to the visible and ultraviolet ranges.

Some of these dye intermediates [338] for the production of anthraquinone and azo dyes (Naphtol AS, Hoechst AG) and azo pigments are still used, especially for polyesters and polyamides. These compounds include 2-aminobenzotrifluoride [*88-17-5*] for C.I. Pigment Yellow 154 [*63661-02-9*]; 3-amino-4-chlorobenzotrifluoride [*121-50-6*] for C.I. Pigment Orange 60 [*68399-99-5*]; 3-(trifluoromethyl)benzoyl fluoride for Indanthren Blue CLB [*6492-78-0*] (BASF) as well as 2-amino-5-chlorobenzotrifluoride [*445-03-4*], 3,5-bis(trifluoromethyl)aniline [*328-74-5*], and 3-amino-4-ethylsulfonylbenzotrifluoride [*382-85-4*].

Pharmaceuticals. The trifluoromethyl substituent is highly lipophilic; it increases the lipid solubility of pharmaceuticals and thus accelerates their absorption and transport in a living organism [339], [340]. In some cases, introduction of the CF_3 group also increases drug effectiveness and reduces undesirable side effects; therefore, benzotrifluorides are used in the synthesis of many pharmaceuticals [341]. These include the analgesics flufenamic acid [*530-78-9*] and niflumic acid [*4394-00-7*], the antidepressant fluoxetine [*54910-89-3*], the muscle relaxant flumetramide [*7125-73-7*], the appetite depressants fenfluramine [*458-24-2*] and fludorex [*15221-81-5*], and the tranquilizers trifluopromazine [*146-54-3*] and fluphenazine [*69-23-8*]. Bendroflumethiazide [*73-48-3*] is an effective diuretic and antihypertensive agent.

Pesticides. Benzotrifluorides are also important in the production of pesticides [342]. The key intermediate, 4-chloro-3,5-dinitrobenzotrifluoride [*393-75-9*], is obtained by dinitration of 4-chlorobenzotrifluoride. Reaction with secondary amines gives trifluralin [*1582-09-8*], profluralin [*26399-36-0*], and benfluralin [*1861-40-1*]. 4-Chlorobenzotrifluoride and 3,4-dichlo-robenzotrifluoride are intermediates for herbicides with a diphenyl ether structure – fluorodifen [*15457-05-3*] and acifluorfen [*50594-66-6*] and the insecticide fluvalinate [*69409-94-5*]. 3-Aminobenzotrifluoride [*98-16-8*] is obtained from benzotrifluoride by nitration and hydrogenation. It is used to make the selective herbicide fluometuron [*2164-17-2*] or is used as a component of the herbicide norflurazon [*27314-13-2*].

11. Ring-Fluorinated Aromatic, Heterocyclic, and Polycyclic Compounds

The compounds discussed in this chapter contain one or more fluorine atoms that are directly attached to aromatic, heterocyclic, or polycyclic rings. Unlike chlorination and bromination, fluorination with elemental fluorine is rarely employed for their production because its violence produces many side reactions (ring-opening, coupling, polymerization, and charring). Therefore, indirect methods are used, distributing the reaction enthalpy over several controllable steps [318].

Diazonium salts are suitable for introducing up to three fluorine atoms per ring. Chlorine – fluorine exchange (halogen exchange, Halex process) with alkali – metal fluorides is suitable in cases where activating substituents are present.

Ring-fluorinated aromatic, heterocyclic, and polycyclic compounds are predominantly used as intermediates for pharmaceuticals, pesticides, dyes and other products. The exceptional properties of fluorinated cyclic compounds (bioactivity spectrum, effectiveness, and solublity) justify the higher production costs compared to those of fluorine-free compounds.

11.1. Mono- and Difluoroaromatic Compounds

Research on ring-substituted aromatic fluorine compounds began in 1870. The fluorine atom attached to the benzene ring is strongly electronegative; it preferentially directs new substituents into the para position and rarely, if ever, into the ortho position [343]. Important aspects, especially for biological applications, are (1) the influence of the strongly electronegative fluorine substituent on adjacent groups, (2) its participation in the resonance system of the aromatic compound by returning electron density, and (3) the simulation of hydrogen or hydroxy substituents that is due to the similarity in space requirements and has led to the synthesis of a series of monofluoro aromatic enzyme inhibitors. The varying lability of the carbon – fluorine bond in mono- and difluoro aromatic compounds has been exploited to prepare indicators and analytical reagents, especially for biogenic and metabolic studies.

11.1.1. Properties

Replacement of aromatic hydrogen by fluorine has only a minor effect on boiling points. Density increases, whereas refractive index and surface tension decrease. Physical constants are shown in Table 22.

The reactivity of fluorine in benzene derivatives depends on the nature of the other ring substituents. Nucleophilic substitution occurs only where activating groups (e.g. nitro) are present in the ortho or para position. In other cases, e.g., with 1-bromo-4-fluorobenzene [*460-00-4*], the carbon – fluorine bond is stronger; hydrolysis gives 4-fluorophenol [*371-41-5*] [344], [345].

11.1.2. Production

Among the published processes for the production of mono- or difluoroaromatic compounds, ring fluorination with dilute fluorine or fluorinating agents such as xenon difluoride [346–349] does not follow a clear course (i.e., without side reactions) and has no commercial value. Fluorobenzene itself can be prepared by pyrolysis of chlorodifluoromethane or chlorotrifluoroethylene in the presence of cyclodipentadiene at 300 – 800 °C [350] and by anodic fluorination of benzene with tetraethylammonium fluoride in acetonitrile [351]. Promising methods are the decarbonylation of benzoyl fluorides in the presence of tris(triphenylphosphine)rhodium(I) chloride in boiling xylene [352] and the thermal decarboxylation of aryl fluoroformates [353].

Diazotization. Aromatic compounds containing one or two fluorine substituents are produced commercially by diazotization of aromatic amines and decomposition of the resulting diazonium fluorides [318],[354]. In one process, aromatic amines are diazotized with dry sodium nitrite in anhydrous hydrogen fluoride at 0 – 20 °C [355] (see Fig. 4).

Table 22. Physical properties of ring-fluorinated aromatic compounds

Compound	CAS registry number	Empirical formula	M_r	*bp*, °C (101.3 kPa)	d_4^{θ} (θ, °C)	n_D^{θ} (θ, °C)
Fluorobenzene	[*462-06-6*]	C_6H_5F	96.1	84.7	1.083 (25)	1.4629 (25)
2-Fluorotoluene	[*95-52-3*]	C_7H_7F	110.13	113 – 114	1.003 (21)	1.4727 (20)
3-Fluorotoluene	[*352-70-5*]	C_7H_7F	110.13	115	0.991 (25)	1.4691 (20)
4-Fluorotoluene	[*352-32-9*]	C_7H_7F	110.13	116	0.991 (25)	1.4688 (20)
4,4′-Difluorodiphenylmethane	[*457-68-1*]	$C_{13}H_{10}F_2$	204.22	263.5∗	1.145 (20)	1.5362 (20)
1,3-Difluorobenzene	[*372-18-9*]	$C_6H_4F_2$	114.09	82 – 83	1.1572 (20)	1.4410 (20)
1,4-Difluorobenzene	[*540-36-3*]	$C_6H_4F_2$	114.09	88 – 89	1.176 (20)	1.4421 (20)

∗ At 100.5 kPa

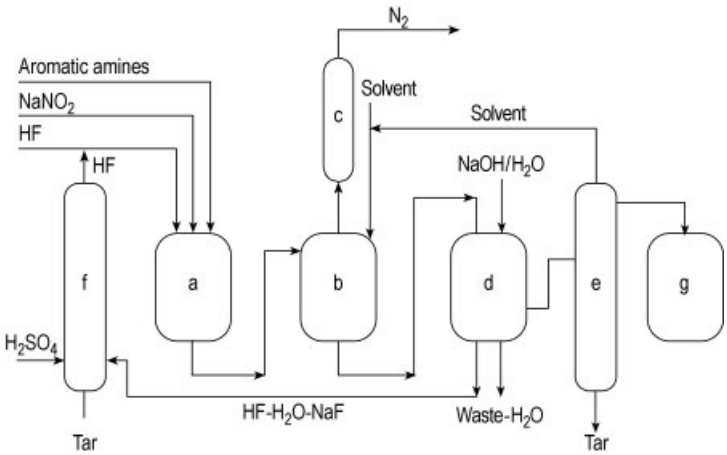

Figure 4. Production of fluoroaromatics by the HF diazotization – dediazoniation method [354]
a) Diazotation reactor; b) Dediazoniation vessel; c) Condenser; d) Separator; e) Distillation column; f) HF purification column; g) Product storage tank

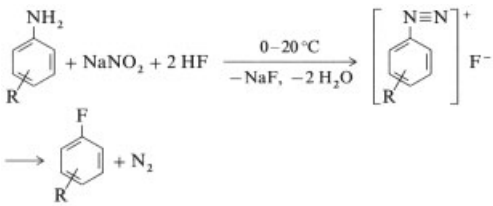

The temperature is increased to 30 – 120 °C, and the formed diazonium fluorides decompose to form fluoroaromatic compounds. A yield of over 90 % of mono- and difluorinated compounds is obtainable. Yields of ca. 81 % are reported for batch operations on a 1-t scale.

In a continuous process the three exothermic steps are controlled by separation; i.e., hydrofluorination of the aromatic amine, diazotization, and thermal decomposition [356]. This permits safe operation on a large scale. Diazotization with nitrosyl chloride [357] and nitrosyl fluoride – HF complexes [358] is also possible. Problems associated with these processes are hydrogen fluoride recovery and waste gas treatment.

In another method (Balz – Schiemann reaction), water-insoluble diazonium fluoroborates are prepared by diazotization of aromatic amines with sodium nitrite in the presence of 40 % fluoroboric acid or sodium or ammonium tetrafluoroborate in HCl [327], [359]. After filtration the diazonium salts are dried and thermolyzed [360] (Fig. 5):

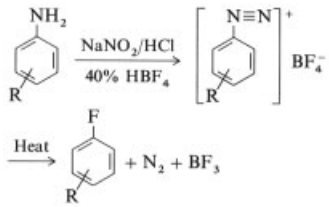

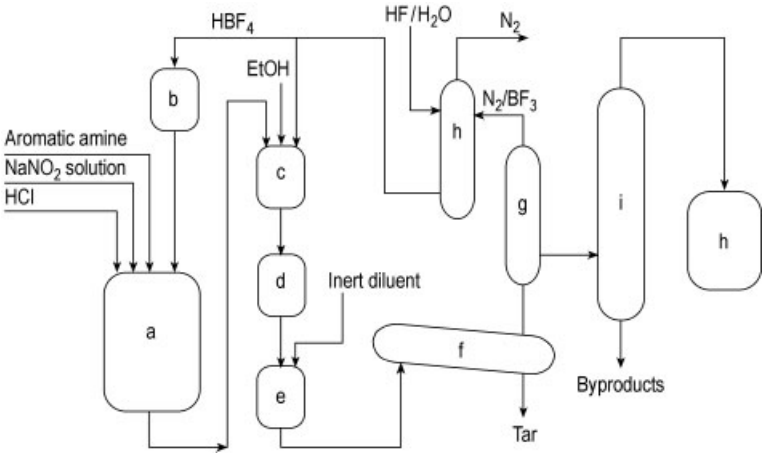

Figure 5. Production of fluoroarmatics by the Balz – Schiemann method
a) Stirred diazotization reactor; b) Storage for 50 % HBF_4; c) Filter; d) Drier; e) Mixer; f) Dediazoniation reactor; g) Condenser; h) Separation column; i) Product distillation column; k) Product storage tank

This thermal decomposition must be strictly controlled, especially when nitro substituents are present, to avoid an explosion. Diazonium tetrafluoroborates usually decompose at a higher temperature than the corresponding diazonium fluorides. The Balz – Schiemann process has rarely been used on a large scale because of the difficulties in handling diazonium tetrafluoroborates. However, it is convenient as a laboratory process because it does not require specialized apparatus. A plant of several 100 t/a has been reported [361].

A large number of fluoroaromatic compounds have been produced by using the two diazotization routes described [327], [362]. Those of commercial interest include fluorobenzene from aniline; 2-, 3-, and 4-fluorotoluene from the appropriate toluidines; 4,4′-difluorodiphenylmethane from 4,4′-diaminodiphenylmethane; 1,3-difluorobenzene from *m*-phenylenediamine; and 1,4-difluorobenzene from *p*-phenylenediamine.

The reaction sequence of nitration, reduction, diazotization, and thermal decomposition can be repeated to introduce up to four fluorine atoms into the benzene ring.

Chlorine – Fluorine Exchange. Of similar commercial importance to diazotization is the replacement of activated chlorine atoms with the aid of alkali-metal fluorides [318], [354]. The usual activating groups are *ortho* and *para* nitro, cyano, and trifluoromethyl groups [363]. Aprotic – polar solvents are preferred, such as dimethylformamide, dimethylacetamide, dimethyl sulfoxide, *N*-methyl-2-pyrrolidone, and tetrahydrothiophene-1,1-dioxide (sulfolane).

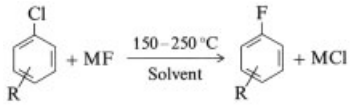

R = NO_2, CN, CF_3; M = K, Cs

Phase-transfer catalysts are sometimes used [364]. The effectiveness of the fluoride source decreases in the order CsF > KF > NaF > LiF.

In commercial batch or semicontinuous operations, the *Halex process* (Fig. 6), potassium fluoride and the activated chloroaromatic compound are thoroughly mixed (1 : 1) with a large volume of an aprotic solvent (dimethyl sulfoxide or sulfolane) and are heated to 150 – 250 °C [363], [365]. The reaction is 90 % complete within 48 h. Removal of KCl and distillation of the product present no difficulty, but efficient solvent recovery is important to reduce costs.

Cost is also reduced by regenerating the potassium fluoride by treating the KCl with HF [366]. The Halex process has the clear advantage of readily accessible raw materials, simplicity of a single-step procedure, and structural specificity. Certain products are accessible in no other way.

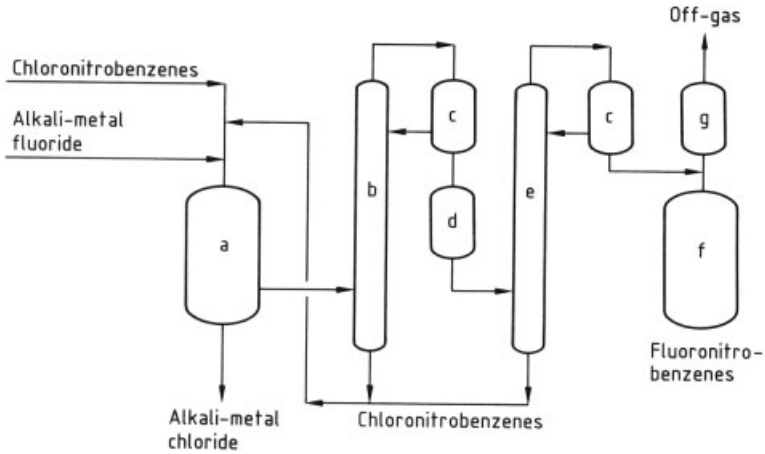

Figure 6. Flow sheet of the Halex process for the manufacture of fluoronitrobenzene [363]
a) Reactor; b) Distillation column; c) Condenser; d) Drier; e) Product distillation column; f) Product storage tank; g) Condenser for off-gas

The following commercially important intermediates are obtained by the Halex process: 1-fluoro-2-nitrobenzene [*1493-27-2*] and 1-fluoro-4-nitrobenzene [*350-46-9*] from the corresponding chloronitrobenzenes;1-chloro-2-fluoro-5-nitrobenzene [*350-30-1*] from 1,2-dichloro-4-nitrobenzene;1-chloro-4-fluoro-3-nitrobenzene [*345-18-6*] from 1,4-dichloro-2-nitrobenzene;1-fluoro-2,4-dinitrobenzene [*70-34-8*] from 1-chloro-2,4-dinitrobenzene;5-fluoro-2-nitrobenzotrifluoride [*393-09-9*] from 5-chloro-2-nitrobenzotrifluoride [*118-83-2*]; 1,3-difluoro-4-nitrobenzene [*446-35-5*] from 1,3-dichloro-4-nitrobenzene; and 2,6-difluorobenzonitrile [*1897-52-5*] from 2,6-dichlorobenzonitrile. These intermediates are mostly used for substituted anilines and compounds with carbonyl and carboxy functions.

11.1.3. Uses

Fluorobenzene, difluorobenzenes, and their derivatives are used widely in the synthesis of pharmaceuticals and pesticides, and as fine chemicals. Fluoroaromatics have played a special role in the development of drugs that act on the central nervous system [367]. The increase in lipid solubility due to fluorine atoms facilitates the absorption and transport of drugs through the blood – brain barrier into the central nervous system.

Fluorobenzene derivatives are used in neuroleptics such as haloperidol [*52-86-8*], tranquilizers such as fluspirilene [*1841-19-6*], sedatives such as flurazepam [*17617-23-1*], and antidepressants such as fluroxamin [*54739-18-3*].

Pesticidal [342] and fungicidal fluorobenzene derivatives [368] include the insecticide diflubenzuron [*35367-38-5*] (made from 2,6-difluorobenzonitrile), the herbicides flamprop [*58677-63-3*] and fluoronitrofen [*13738-63-1*], and the fungicides nuarimol [*63284-71-9*] and flurimid [*41205-21-4*].

4,4′-Difluorobenzophenone [*345-92-6*], a starting material for aromatic polycondensates, is produced by oxidation of 4,4′-difluorodiphenylmethane, which is obtained by diazotization of the corresponding diamine. 4,4′-Difluorobenzophenone undergoes polycondensation with hydroquinone, yielding a polyetherether ketone (PEEK) resin, a thermoplastic. 4-Fluoroaniline [*371-40-4*], 4-fluorobenzaldehyde [*459-57-4*], and 4-fluorobenzoic acid [*456-22-4*] are intermediates for liquid crystal polymers [369].

Aromatic compounds with reactive fluorine substituents are used for the characterization of amino acids (Sanger's reagent, 1-fluoro-2,4-dinitrobenzene [*70-34-8*]), the immobilization of enzymes (4-fluoro-3-nitrophenylazide [*28166-06-5*]), and peptide cross-linkage (1,5-difluoro-2,4-dinitrobenzene [*327-92-4*]).

Aromatic fluorine compounds have been developed for medical applications, such as ^{19}F-magnetic resonance imaging (MRI) and in ^{18}F-positron emission tomography (PET) [370].

11.2. Highly Fluorinated Aromatic Compounds

Compounds in which an aromatic ring is substituted by three to five fluorine atoms have little commercial importance. 1,2,4-Trifluorobenzene [*367-23-7*], *bp* 88 °C, and 1,3,5-trifluorobenzene [*372-38-3*], *bp* 75.5 °C, are produced by the Balz–Schiemann reaction (see Section 11.1.2) from 2,4-difluoroaniline [*367-25-9*] and 3,5-difluoroaniline [*372-39-4*], respectively. 1,3,5-Tricyano-2,4,6-trifluorobenzene [*3638-97-9*], *mp* 148–150 °C, is used as an intermediate for pesticides; it is obtained by chlorine–fluorine exchange from 1,3,5-tricyano-2,4,6-trichlorobenzene [371].

1,2,3,5-Tetrafluorobenzene [*2367-82-0*], *bp* 83 °C, is obtained by the Balz–Schiemann reaction from 2,3,5-trifluoroaniline [*363-80-4*], and 1,2,4,5-tetrafluorobenzene [*327-54-8*], *bp* 90 °C, from 2,4,5-trifluoroaniline [*57491-45-9*]. 1,2,3,4-Tetrafluorobenzene [*551-62-2*], *bp* 95 °C, pentafluorobenzene [*363-72-4*], *bp* 85 °C, and hexafluorobenzene are produced by the CoF_3 method (see Section 11.3).

The Halex process (Section 11.1.2) is used commercially to produce 1,4-dicyano-2,3,5,6-tetrafluorobenzene [*1835-49-0*], *mp* 197–199 °C, from the corresponding tetrachloro derivative [372]. This compound is used as a monomer for thermostable polymers.

11.3. Perhaloaromatic Compounds

Hexafluorobenzene [*392-56-3*], *bp* 80.3 °C, has been thoroughly investigated [373]. Nucleophilic substitution produces pentafluoroaromatic compounds such as

bromopentafluorobenzene [*344-04-7*], *bp* 135.6 °C
pentafluorophenol [*771-56-2*], *bp* 117–118 °C
pentafluoroaniline [*771-60-8*], *bp* 153 °C
pentafluorobenzene [*771-61-9*], *bp* 143 °C
pentafluorotoluene [*771-56-2*], *bp* 117 °C
pentafluorobenzaldehyde [*653-37-2*], *bp* 164–166 °C
pentafluorobenzoic acid [*602-94-8*], *bp* 220 °C

Production. Hexafluorobenzene was first produced in 1955 by pyrolysis of tribromofluoromethane [*353-54-8*] in a platinum tube at 640 °C under atmospheric pressure [374]:,

$$6\,CBr_3F \longrightarrow C_6F_6 + 9\,Br_2$$

Of greater commercial interest is the pyrolysis of an equimolar mixture of dichlorofluoromethane [*75-43-40*] and chlorofluoromethane [*593-70-4*] at 600–800 °C [375]:

$$3\,CHCl_2F + 3\,CH_2ClF \longrightarrow C_6F_6 + 9\,HCl$$

For many years a commercial multistep process employed CoF_3 as the fluorinating agent [376]. For example, at 150 °C benzene gives a mixture of cyclohexanes containing 8–11 fluorine atoms.

$$C_6H_6 \xrightarrow[CoF_2]{CoF_3} C_6H_{12-n}F_n \xrightarrow{MOH} C_6H_{11-n}F_{n-1} + C_6H_{10-n}F_{n-2} \xrightarrow{Fe\ or\ Fe_3O_4} C_6F_6,\ C_6HF_5,\ C_6H_2F_4$$

$n = 8-11$,
M = alkali metal

The CoF_2 formed in this reaction is fluorinated to CoF_3 with elemental fluorine, and reused. The organic products are heated with an alkali-metal hydroxide to form a mixture of polyfluorocyclohexenes and polyfluorocyclohexadienes. These products are aromatized by passage over iron or iron compounds at 400–600 °C to give hexafluorobenzene, pentafluorobenzene, and tetrafluorobenzenes. Octafluorotoluene [*434-64-0*] and the three perfluoroxylenes are also obtained by this method.

The disadvantages of this process are the technological difficulties and the low utilization of fluorine, of which a large part is converted to hydrogen fluoride and alkali-metal fluorides. As a consequence, aromatic fluorine compounds containing other halogens are currently produced by the Halex method (Section 11.1.2). Reaction of hexachlorobenzene [365] with potassium fluoride at 450 °C and 1.03 MPa gives a yield of 21 % hexafluorobenzene together

with chloropentafluorobenzene [*344-07-0*] (20 %), 1,3-dichloro-2,4,5,6-tetrafluorobenzene [*1198-61-4*] (14 %), and 1,3,5-trichloro-2,4,6-trifluorobenzene [*319-88-0*] (12 %).

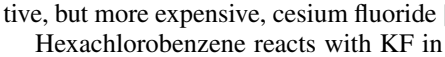

$$C_6Cl_6 + KF \longrightarrow C_6F_6,\ C_6ClF_5,\ C_6Cl_2F_4,\ C_6Cl_3F_3$$

The yield of hexafluorobenzene can be increased by recycling the other products for further reaction with potassium fluoride. A yield of 42 % hexafluorobenzene is obtained from chloropentafluorobenzene with the more reactive, but more expensive, cesium fluoride [377].

Hexachlorobenzene reacts with KF in aprotic solvents such as dimethylformamide, dimethyl sulfoxide, *N*-methyl-2-pyrrolidone, and sulfolane to give not hexafluorobenzene, but the above-mentioned mixed products. The reaction temperature is 150 – 250 °C and the residence time 5 – 36 h. Detailed discussions of the process are given in [363], [365].

Uses. Hexafluorobenzene has been investigated as an inhalation anesthetic in veterinary medicine [378] and as a working fluid in Rankine-cycle engines for temperatures above 350 °C [379]. Derivatives such as pentafluorobenzaldehyde or pentafluorophenyl dimethylsilyl ether are used in the chromatographic analysis of steroids [380] and catecholamines [381], and as intermediates for the production of liquid crystal polymers [382]. Pentafluorophenoxy perfluorovinyl ethers are cross-linking comonomers for perfluorinated elastomers [383].

11.4. Fluorinated Heterocyclic and Polycyclic Compounds

Fluorinated pyridines, pyrimidines, and triazines are heterocycles with commercial importance. Only some representative polycyclic fluorine compounds are discussed as examples.

11.4.1. Ring-Fluorinated Pyridines

Introduction of fluorine into the pyridine ring reduces the basicity of the latter [384]. 2-Fluoropyridine [*372-48-5*], *bp* 126 °C (100.4 kPa), has a labile fluorine substituent. It is produced in 74 % yield by reaction of 2-chloropyridine with potassium bifluoride at 315 °C in 4 h [385]. 3-Fluoropyridine [*372-47-4*], *bp* 105 – 107 °C (100.3 kPa), is produced in 50 % yield from 3-aminopyridine using the Balz – Schiemann process (see page 39). 4-Fluoropyridine [*694-52-0*], *bp* 108 °C (100 kPa) is produced in 54 % yield by diazotization of 4-aminopyridine in anhydrous hydrogen fluoride [386].

2,4-Difluoropyridine [*34941-90-7*], *bp* 104 – 105 °C, is obtained by the Halex process (see page 41) from 2,4-dichloropyridine and potassium fluoride in sulfolane [387]. 2,6-Difluoropyridine [*1513-65-1*], *bp* 124.5 °C (99.1 kPa), is obtained by the reaction of 2,6-dichloropyridine with potassium fluoride in the absence of solvent. The yield is 80 % after 18 h at 400 °C [388].

For pentafluoropyridine [*700-16-3*], *bp* 83 °C, the preferred method is the reaction of pentachloropyridine with KF at 480 – 500 °C; the yield is 83 % [389]. Nucleophilic substitution reactions have been thoroughly investigated [390].

Uses. Fluoropyridines are intermediates for pesticides. For example, 2-fluoro-4-hydroxypyridine [*22282-69-5*] is a precursor of 2-fluoro-3,5-dihalo-4-hydroxypyridine herbicides [391]; 2-fluoro-6-hydroxypyridine [*55758-32-2*] is used for insecticides and nematocides [392] and various derivatives of fluorinated pentahalogenpyridine herbicides [393]. Pentafluoropyridine in mixtures with hexafluorobenzene is used as the working fluid in Rankine-cycle engines up to 382 °C [379].

11.4.2. Trifluoromethylpyridines

2-Trifluoromethyl- [*368-48-9*], 3-trifluoromethyl- [*3796-23-4*], and 4-trifluoromethylpyridine [*3796-24-5*] can be obtained by the reaction of picolinic, nicotinic, or isonicotinic acid, respectively, with sulfur tetrafluoride [394]. Commercial processes use side-chain chlorination of the methylpyridines followed by chlorine – fluorine exchange with antimony chlorofluorides or anhydrous hydrogen fluoride [395]. 2-Chloro-5-trifluoromethylpyridine,

bp 190 °C, is produced commercially from 2-chloro-5-trichloromethylpyridine and hydrogen fluoride at 180 – 200 °C and at a pressure of 3 – 4.5 MPa. It is an intermediate for the synthesis of the selective herbicide fluazifop [*69335-91-7*], [396].

11.4.3. Fluoropyrimidines

Certain fluoropyrimidines have gained commercial importance [397]. 5-Fluoropyrimidines are employed in cancer chemotherapy; their biochemistry and pharmacology have been intensively studied [398]. 5-Fluorouracil [*51-21-8*] is obtained in a 80 – 92 % yield by the fluorination of 2,4-dihydroxypyrimidine with fluorine or trifluoromethyl hypofluorite [399], [400].

5-Chloro-2,4,6-trifluoropyrimidine [*697-83-6*], *bp* 114.5 °C (100 kPa), is produced commercially from 2,4,5,6-tetrachloropyrimidine by chlorine – fluorine exchange, using sodium fluoride at 300 °C [401] or anhydrous hydrogen fluoride in the liquid phase [402] or gas phase [403].

The 5-chloro-2,4-difluoropyrimidinyl radical acts as the reactive group in reactive dyes [397] for cellulose and cotton fibers such as Levafix EA (Bayer) and Drimarene K (Sandoz) and for wool, e.g., Verofix (Bayer) and Drimalene (Sandoz).

11.4.4. Fluorotriazines

2,4,6-Trifluoro-1,3,5-triazine [*675-14-9*], *bp* 72.4 °C (101.7 kPa), is produced commercially from 2,4,6-trichloro-1,3,5-triazine (cyanuric chloride) with either anhydrous hydrogen fluoride [404] or sodium fluoride in sulfolane [405]. It is used to manufacture reactive dyes by reaction between one or two fluorine substituents with amino groups of chromophores; the remaining fluorine binds to the fiber [406]. The reactive intermediates can be synthesized e.g., from the corresponding 6-substituted 2,4-dichlorotriazines and an alkali-metal fluoride or by reaction of cyanuric fluoride with anilines or phenols in organic solvents [406], [407]. Like the pyrimidine reactive dyes, the fluorotriazines are used on cellulose, polyesters, polyamides, and wool [397]. See also → Reactive Dyes.

11.4.5. Polycyclic Fluoroaromatic Compounds

Polycyclic fluoroaromatic compounds are used as intermediates for the production of pharmaceuticals.

4-Fluoro- [*324-74-3*] and 4,4′-difluorobiphenyl [*398-23-2*] are obtained by diazotization from the corresponding amines in yields of up to 80 % [408]. Other biphenyl derivatives are prepared from fluorobenzenes. The analgesic diflunisal [*22494-42-4*] is produced from 2,4-difluoroaniline [*367-25-9*] by diazotization and coupling with salicylic acid. The anti-inflammatory drugsflurbiprofen [*5104-49-4*] and fluprofen [*17692-38-5*] are prepared from 2-fluoroaniline [*348-54-9*].

The anti-inflammatory drug sulindac [*38194-50-2*] is a monofluorinated indole-3-acetic acid.

12. Economic Aspects

The compounds discussed in Chapters 10 and 11 are not produced in large quantities and often command high prices. 4-Chlorobenzotrifluoride, 3,4-dichlorobenzotrifluoride, 3-trifluoromethylphenyl isocyanate, and benzotrifluoride are produced in the United States by Occidental Chemical, in Western Europe by Hoechst, Rhône-Poulenc, MitEni and Dow-Elanco, and in Japan by Daikin. The worldwide production capacity is ca. 35 000 – 40 000 t/a.

Heterocyclic fluorine compounds have economic importance in the production of reactive dyes. The main producers in Western Europe are Bayer AG, Ciba-Geigy, ICI, and Sandoz.

Fluorobenzene and its derivatives are produced in Western Europe by AlliedSignal – Riedel-de-Haën (production capacity 1600 t/a), Zeneca (2000 – 2500 t/a), Rhône-Poulenc (1000 t/a), MitEni (1000 t/a) in Europe, DuPont (1400 t/a), Mallinckrodt (1200 t/a) in the USA and Asahi Glass (1000 t/a) in Japan. Fluoronitrobenzenes and fluoroanilines are produced by the Halex process by Hoechst, MitEni, Rhône-Poulenc, Asahi Glass, and others. The total capacity for fluoroaromatic intermediates is estimated to be several thousand tons per year.

13. Toxicology and Occupational Health

With few exceptions, organic fluorine compounds are physiologically inert and display insignificant toxicity. This is a consequence of the chemical stability of the carbon – fluorine bond and the increased stability of hydrogen and halogen bonds attached to a fluorinated carbon atom. Low toxicity is an important factor in many applications of these compounds.

The difference in toxicity between a chloro or bromo compound and the corresponding fluoro compound is often striking. Thus, carbon tetrachloride [*56-23-5*] is a powerful liver and kidney toxin and a weak carcinogen, as reflected in its TLV of 5 ppm. However, the product obtained by replacing one chlorine atom with fluorine, trichlorofluoromethane [*75-69-4*], has no adverse effects on the liver, kidney, or other organs, and no carcinogenic effects on animals exposed to high concentrations for a lifetime; the TLV is 1000 ppm. Another example is the chemical warfare agent mustard gas, $S(CH_2CH_2Cl)_2$ [*505-60-2*], which is a strong alkylating agent and notorious vesicant. The fluoro analog, bis(2-fluoroethyl) sulfide [*373-25-1*], is chemically and physiologically inert, with no vesicant properties [409].

The few highly toxic organofluorine compounds usually have easily replaceable fluorine atoms. Examples are diisopropyl fluorophosphate [*55-91-4*], a potent cholinesterase inhibitor, and perfluoroisobutylene [*382-21-8*], which causes pulmonary edema at low concentrations. Sodium fluoroacetate [*62-74-8*], a potent rodenticide, is an exception; it does not liberate fluorine, but interferes with metabolism by mimicking acetic acid.

13.1. Fluorinated Alkanes

Fluoroalkanes [410], [411]. Perfluoroalkanes have very low toxicity. Thus, rats exposed to an 80 : 20 mixture of perfluorocyclobutane [*115-25-3*] and oxygen for 4 h survived with no ill effects. Likewise, no ill effects were seen in four species of animals exposed to a 10 % concentration in air 6 h/d for 90 d. Partially fluorinated alkanes have similarly low toxicity, as shown by similar experiments with difluoromethane [*75-10-5*] and1,1-difluoroethane [*75-37-6*].

Chlorofluoroalkanes are more toxic than the corresponding fluoroalkanes; nevertheless, most chlorofluoroalkanes have low toxicity [410–412]. High concentrations (10 – 50 % in air) of many of them, like lower concentrations of many hydrocarbon and chlorohydrocarbon solvents, can cause cardiac sensitization, i.e., sensitization of the heart to the body's adrenalin. This can lead to cardiac arrhythmia (heartbeat irregularity) and sometimes cardiac arrest. Deaths have been caused by "aerosol sniffing".

The toxicity of dichlorodifluoromethane [*75-71-8*] has been thoroughly investigated. Rats survived a 6-h exposure to an 80 % mixture with oxygen. Five species of animals continuously exposed to 810 ppm for 90 d showed no effects except for slight liver damage in guinea pigs. Rats and dogs showed no significant health effects when fed a diet containing 0.3 % for 2 years. Teratogenic and reproductive tests in rats were also negative. Repeated exposure caused little or no irritation to rat skin or the rabbit eye. In a screening test, dogs injected with adrenalin showed cardiac sensitization on exposure to 50 000 ppm in air, but not to 25 000. On the basis of animal data and human experience, a TLV of 1000 ppm has been selected to provide an ample margin of safety against cardiac sensitization and other injury [413]. The MAK is also 1000 ppm [414].

Compared to dichlorodifluoromethane, trichlorofluoromethane [*75-69-4*] and 1,1,2-trichloro-1,2,2-trifluoroethane [*76-13-1*] are slightly more toxic, whereas 1,2-dichloro-1,1,2,2-tetrafluoroethane [*76-14-2*] is slightly less toxic. However, for all three a TLV of 1000 ppm is judged to provide an adequate margin of safety. On the other hand, 1,1,1,2-tetrachloro-2,2-difluoroethane [*76-11-9*] and 1,1,2,2-tetrachloro-1,2-difluoroethane [*76-12-0*] cause liver and lung damage to rats subjected to repeated exposure at 1000 ppm; therefore, a TLV of 500 ppm is recommended for these compounds. These data indicate that the toxicity of chlorofluorocarbons tends to increase with the chlorine : fluorine ratio and the number of carbon atoms.

Although the toxicity of most chlorofluoroalkanes bearing hydrogen (chlorofluorohy-

drocarbons) is also low, it tends to be higher than that of the closely related chlorofluorocarbons. The difference is usually slight, as in the case of 2-chloro-1,1,1,2-tetrafluoroethane [*2837-89-0*], which is a slightly stronger cardiac sensitizer than 1,2-dichloro-1,1,2,2-tetrafluoroethane, but is otherwise similar to it in toxic properties. Chlorodifluoromethane [*75-45-6*] is similar in toxicity to dichlorodifluoromethane in most respects, and has the same TLV of 1000 ppm. At 50 000 ppm it has a weak carcinogenic effect in male rats, but not at lower concentrations or in mice or female rats; therefore, this is considered of no practical significance [415].

The toxicity of dichlorofluoromethane [*75-43-4*] is more like that of chloroform than of dichlorodifluoromethane or chlorotrifluoromethane [*75-72-9*], especially with respect to injury on repeated exposure; its TLV, 10 ppm, is low for a chlorofluoroalkane.

Bromofluoroalkanes, some of which are fire extinguishing agents and anesthetics, are more toxic than the corresponding chlorofluoroalkanes, but generally are low in toxicity compared to other fire extinguishing agents and anesthetics [409], [413]. Trifluorobromomethane [*75-63-8*] produced no adverse effects on dogs and rats exposed to 23 000 ppm 6 h/d, 5 d/week, for 18 weeks; its TLV is 1000 ppm [413].

13.2. Fluorinated Olefins

Most fluorinated olefins have halogen atoms of low reactivity and a correspondingly low-to-moderate toxicity [409], [412]. The toxicities of the five most common members of the class, typical in these respects, are shown in Table 23.

Table 23. Toxicities of fluorinated olefins∗

Compound	CAS registry no.	Lethal conc., ppm
Vinyl fluoride	[*75-02-5*]	> 800 000 (ALC)∗∗
Vinylidene fluoride	[*75-38-7*]	> 800 000 (ALC)∗∗
Tetrafluoroethylene	[*116-14-3*]	40 000 (LC_{50})
Hexafluoropropene	[*116-15-4*]	3 000 (LC_{50})
Chlorotrifluoroethylene	[*79-38-9*]	1 000 (LC_{50})

∗ Inhalation by rats, 4-h exposure.
∗∗ Approximate lethal concentration.

In perfluoroisobutylene and 2,3-dichloro-1,1,1,4,4,4-hexafluoro-2-butene [*303-04-8*], halogens are readily displaced by nucleophilic reactants; thus, these two compounds exhibit high acute toxicity. Perfluoroisobutene, with a LC_{50} of 0.5 ppm, acts much like phosgene in causing death by pulmonary edema. However, perfluoroisobutene is ca. 10 times as toxic as phosgene, so exposure to it must be carefully avoided. Pyrolysis of tetrafluoroethylene or its polymers above 400 °C is one of its sources.

13.3. Fluorinated Alcohols

2-Fluoroethanol [*371-62-0*] has a high acute toxicity (LD_{50} 10 mg/kg), a consequence of its ready biological oxidation to fluoroacetic acid (see Section 13.6). Thus, contact with the skin and inhalation of vapors must be avoided. The di- and trifluoroethanols ([*359-13-7*] and [*75-89-8*], respectively) have relatively low acute toxicity, similar to the corresponding acetic acids [412]. The acute toxicity of 1,1,1,3,3,3-hexafluoro-2-propanol [*920-66-1*] is also low, but the substance is a strong skin and eye irritant.

13.4. Fluorinated Ketones

In 90-d inhalation studies in animals, hexafluoroacetone [*684-16-2*] caused severe damage to kidneys and other organs at 12 ppm, moderate damage at 1 ppm, and no damage at 0.1 ppm. Repeated skin exposure led to testicular damage in rats. These data and plant experience led to selection of a TLV of 0.1 ppm, with a warning against skin exposure [413]. Studies of hexafluoroacetone and three fully halogenated chlorofluoroketones indicated moderate acute toxicity [416].

13.5. Fluorinated Carboxylic Acids

Fluoroacetic acid [*144-49-0*] is highly toxic to mammals; its sodium salt is an effective, but indiscriminate, rodenticide; in rats the LD_{50} of the salt is only 1.7 mg/kg [413]. In contrast, difluoroacetic acid [*381-73-7*] and perfluoroalkanoic acids have low acute toxicity.

The unusually high toxicity of fluoroacetic acid compared to the more highly fluorinated acids is due to its unique ability to interfere with the citric acid cycle, the oxidation pathway used for energy production from amino acids, fatty acids, and carbohydrates. Fluoroacetic acid enters the cycle at the same site as acetic acid and is converted to fluorocitric acid [*387-89-1*] analogously to the conversion of acetic acid to citric acid. The fluorocitric acid inhibits aconitase, a key enzyme for the breakdown of citric acid, with the result that the citric acid concentration soon rises to lethal levels [417].

Substances that yield fluoroacetic acid on biochemical oxidation, such as straight-chain, even-numbered, ω-fluoro alcohols or alkanoic acids, are also very toxic.

13.6. Other Classes

Simple perfluoroethers, and the oily oligomers of hexafluoropropylene oxide with modified end groups, have the low toxicity expected from their chemical inertness. Hexafluoropropylene oxide itself, a reactive substance, is moderately toxic to rats (4-h LC_{50}, 3700 ppm) [418]. Several partially fluorinated ethers are used as anesthetics, e.g., Enflurane [*13838-16-9*], F_2CHOCF_2CHFCl, and Methoxyflurane [*76-38-0*], $CH_3OCF_2CHCl_2$. Their toxicity is low compared to that of most anesthetics (→ Anesthetics, General, Chap. 1.).

Perfluorinated tertiary aliphatic amines are inert both chemically and biologically. This is illustrated by perfluorotripropylamine [*338-83-0*]; in emulsion with perfluorodecalin [*306-94-5*], it has shown promise as a blood substitute in clinical trials [419].

Fluorine substituents usually have little effect on the toxicity of aromatic compounds, whether monocyclic, polycyclic or heterocyclic [416].

14. References

General References

1. *Houben-Weyl,* **5/3,** pp. 1 – 502.
2. *Kirk-Othmer,* **10,** 829 – 962.
3. R. C. Downing, F. W. Mader, T. W. Tomkowik: "Fluorocarbons" in J. J. McKetta, *Encyclopedia of Chemical Processing and Design,* vol. 23, Marcel Dekker, New York 1985, pp. 216 – 269.
4. O. Scherer: "Technische Organische Fluorverbindungen" in *Topics in Current Chem.,* vol. 14, Springer Verlag, Berlin – Heidelberg – New York 1970, pp. 127 – 234.
5. R. E. Banks, D. W. A. Sharp, J. C. Tatlow: *Fluorine, The First Hundred Years (1886 – 1986),* Elsevier Sequoia, Lausanne – New York 1986; *J. Fluorine Chem.* **33** (1986) 1 – 399.
6. R. E. Banks (ed.): *Organofluorine Chemicals and their Industrial Applications,* Ellis Horwood Ltd., Chichester 1979.
7. R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Ellis Horwood Ltd. & Wiley Halsted Press, New York – Brisbane – Chichester – Toronto 1982.
8. M. Sittig: "Fluorinated Hydrocarbons and Polymers," *Chemical Process Monograph* No. 22, Noyes Development Corp., Park Ridge N.J., 1966.
9. Process Economics Program (PEP) Report No. 166: *Fluorinated Polymers,* SRI International, Menlo Park, Ca. 1983.
10. J. H. Simons (ed.): *Fluorine Chemistry,* **vols. 1 – 5,** Academic Press, New York 1950 – 1964.
11. M. Stacey, J. C. Tatlow, A. G. Sharpe (eds.): *Advances in Fluorine Chemistry,* **vols. 1 – 7,** Butterworths, London 1960 – 1973.
12. P. Tarrant (ed.): *Fluorine Chemistry Reviews,* **vols. 1 – 8,** Dekker, New York 1967 – 1977.
13. R. E. Banks, M. G. Barlow (eds.): *Fluorocarbons and Related Chemistry,* **vols. 1 – 3,** The Chemical Society, London 1971 – 1976.
14. H. J. Emeleus, J. C. Tatlow (eds.): *J. Fluorine Chem.* Elsevier Sequoia S.A., Lausanne.
15. Chemical Abstracts Selects (CAS®): *Organofluorine Chemistry,* American Chemical Society.

Specific References

16. M. Hudlický, *J. Fluorine Chem.* **33** (1986) 302.
17. R. E. Banks (ed.): *Organofluorine Chemicals and their Industrial Applications,* Ellis Horwood Ltd., Chichester 1979.
18. R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994.

19. R. E. Banks, J. C. Tatlow, *J. Fluorine Chem.* **33** (1986) 285 – 286.
20. R. E. Banks, J. C. Tatlow, in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 1.
21. R. E. Banks, J. C. Tatlow, *J. Fluorine Chem.* **33** (1986) 71 – 108.
22. T. Midgley, A. L. Henne, *Ind. Eng. Chem.* **22** (1930) 542; T. Midgley, *Ind. Eng. Chem.* **29** (1937) 239.
23. H. Goldwhite, *J. Fluorine Chem.* **33** (1986) 109.
24. W. A. Sheppard, C. M. Sharts: *Organic Fluorine Chemistry,* W. A. Benjamin Inc., New York 1969.
25. A. Haas, M. R. C. Gerstenberger, *Angew. Chem.* **93** (1981) 659; *Angew. Chem. Int. Ed. Engl.* **20** (1981) 647.
26. R. J. Lagow, J. L. Margrave, *Prog. Inorg. Chem.* **26** (1979) 161.
27. R. J. Lagow, *J. Fluorine Chem.* **33** (1986) 321.
28. W. Bockemüller, *Justus Liebigs Ann. Chem.* **506** (1933) 20; W. T. Miller, *J. Am. Chem. Soc.* **62** (1940) 341.
29. G. H. Cady, et al., *Ind. Eng. Chem.* **39** (1947), 290. R. N. Haszeldine, F. Smith, *J. Chem. Soc.* 1950, 2689 – 94, 2787 – 89.
30. J. L. Adcock, K. Horita, E. B. Renk, *J. Am. Chem. Soc.* **103** (1981) 371. J. L. Adcock, *J. Fluorine Chem.* **33** (1986) 327.
31. Allied, EP 0 032 210, 1984 (H. R. Nykka, J. B. Hino, R. E. Eibeck, J. I. Braumen); EP-A 0 031 519, 1980 (H. R. Nykko, J. B. Hino, R. E. Eibeck, M. A. Robinson).
32. E. A. Tyczkowski, L. A. Bigelow, *J. Am. Chem. Soc.* **77** (1955) 3007. A. F. Maxwell, F. E. Detoro, L. A. Bigelow, *J. Am. Chem. Soc.* **82** (1960) 5287.
33. M. Stacey, J. C. Tatlow in [11], vol. 1, pp. 166 – 188.
34. T. Abe, S. Nagase in [7], pp. 19 – 43.
35. Y. W. Alsmeyer, W. V. Childs, R. M. Flynn, J. C. Schmeltzer in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 121.
36. G. P. Gambaretto, et al., *J. Fluorine Chem.* **27** (1985) 149.
37. H. M. Fox, F. N. Ruehlen, W. V. Childs, *J. Electrochem. Soc.* **118** (1971) 1246.
38. Phillips Petroleum Co., US 3 511 760, 1967 (H. M. Fox, F. N. Ruehlen); US 3 686 082, 1970 (F. N. Ruehlen); US 3 709 800, 1971 (H. M. Fox); US 3 882 001, 1973 (K. L. Mills).
39. W. V. Childs in "Technique of Electroorganic Synthesis," in A. Weissberger (ed.): *Techniques of Chemistry,* **vol. 5,** J. Wiley & Sons, New York 1982, Chapter 7.
40. W. V. Childs, L. Christensen, F. W. Klink, C. F. Kolpin in H. Lund, M. M. Baizer (eds.): *Organic Electrochemistry,* 3rd ed., Marcel Dekker, New York 1981, p. 1103.
41. A. K. Barbour, L. J. Belf, M. W. Buxton in [11], vol. 3, pp. 181 – 270.
42. Kinetic Chemicals Inc., DE 552 919, 1930.
43. O. Scherer in [4], vol. 14, p. 136.
44. Hoechst, DE 1 000 798, 1952 (O. Scherer, H. Kühn, E. Forche).
45. Hoechst, GB 1 025 759, 1963 (O. Scherer J. Korinth, P. Frisch).
46. J. M. Hamilton, Jr. in [11], vol. 3, pp. 117 – 180.
47. L. Dolby-Glover, *Chem. Ind. (London)* 1986, 518 – 523.
48. O. Paleta in [11], vol. 8, pp. 39 – 71.
49. Hoechst, EU 634 383, 1994 (R. Franz, G. Siegemund).
50. H. Suschitzky in [11], vol. 4, pp. 1 – 30.
51. G. A. Boswell, Jr., W. C. Ripka, R. M. Scribner, C. W. Tullock in Wm. G. Dauben et al. (eds.): *Organic Reactions,* **vol. 21,** J. Wiley & Sons, New York 1974, pp. 1 – 124.
52. Ch.-L. J. Wang in A. S. Kende et al. (eds.): *Organic Reactions,* **vol. 34,** J. Wiley & Sons, New York 1985, pp. 319 – 400.
53. Y. A. Fialkov, L. M. Yagupolski in I. L. Knunyants, G. G. Yakobson (eds.): *Syntheses of Fluoroorganic Compounds,* Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1985, pp. 233 – 289.
54. C. M. Sharts, W. A. Sheppard in [51], pp. 125 – 406.
55. R. Filler, *Jsr. J. Chem.* **17** (1978) 71.
56. D. Hebel, O. Lerman, S. Rozen, *J. Fluorine Chem.* **30** (1985) 141.
57. R. Perry, *J. Fluorine Chem.* **33** (1986) 293.
58. R. Fields, *J. Fluorine Chem.* **33** (1986) 287.
59. J. M. Hamilton in [11], p. 125, vol. 3.
60. *PCR Research Chemicals Catalogue (1990/91),* PCR, Inc., Gainesville, FL 1990, p. 176.
61. P. P. Barthélemy, A. Leroy in: *Proceedings of the UTECH 96 Conference*, Den Haag 1996, p. 404.
62. T. J. Brice in [10], vol. 1, p. 423.
63. T. M. Reed III in [10], vol. 5, p. 133.
64. S. W. Green et al. in [18], p. 89.

65. Halocarbon, DE 2 644 594, 1975 (L. L. Ferstandig).
66. Allied, US 3 644 545, 1986; ICI, US 2 452 975, 1948.
67. Dow, GB 805 503, 1958 (R. P. Ruh, R. A. Davis).
68. ICI, US 4 129 603, 1979 (S. L. Bell).
69. DuPont, US 3 258 500, 1966.
70. C. Gervasutti, L. Marangoni, W. Marra, *J. Fluorine Chem.* **19** (1981/82) 1.
71. Asahi Glass, JP Appl. 1-319 443, 1989 (Y. Furutaka, H. Aoyama, Y. Homoto).
72. R. D. Fowler et al., *Ind. Eng. Chem.* **39** (1947) 292, 319.
73. A. K. Barbour et al., *J. Appl. Chem.* **2** (1952) 127.
74. J. M. Tedder in [11], vol. 2, p. 104.
75. DuPont, US 2 458 551, 1947 (A. F. Benning, J. D. Park, S. E. Krahler).
76. A. J. Woytek in [5], p. 331.
77. Asahi Glass, JP 5 841 829, 1984 (S. Fukui, H. Yoneda).
78. Daikin Kogyo, JP 6 077 983, 1985.
79. Showa Denka, JP 6 081 134, 1985.
80. DuPont, US 2 404 374, 1943 (J. Harmon).
81. The National Smelting Comp., DE-OS 1 205 533, 1962 (J. C. Tatlow, P. L. Coe).
82. R. L. Powell, J. H. Steven in [18], p. 617.
83. L. Zipfel, W. Krücke, K. Börner, P. Barthélemy, P. Dournel in: *Polyurethanes World Congress 1997*, p. 176.
84. AFEAS Administrative Organization, Science & Policy Services, Inc., West Tower - Suite 400, 1333 H Street NW, Washington 1997.
85. A. J. Woytek, *J. Fluorine Chem.* **33** (1986) 331.
86. D. S. L. Slinn, S. W. Green in [7], pp. 45 – 82.
87. M. LeBlanc, J. G. Riess in [7], pp. 83 – 138.
88. M. J. Molina, F. S. Rowland, *Nature* **249** (1974) 810.
89. R. T. Watson, M. A. Geller, R. S. Stolarski, R. F. Hampson: *Processes That Control Ozone and Other Climatically Important Trace Gases,* NASA Reference Publication 1162, NASA, Washington 1986.
90. J. K. Hammitt, et al.: *Product Uses and Market Trends for Potential Ozone Depleting Substances 1985 – 2000, (R-3386-EPA),* Rand Corp., Santa Monica 1986.
91. E. R. Larsen in [12], vol. 3, p. 1.
92. Hoechst AG, *Frigen, Handbuch für Kälte- und Klimatechnik,* 1969.
93. BIOS Final Report Nr. 112, I.G. Werke Hoechst (1945).
94. VEB-Werk Nunchritz, DD 215014, 1983 (I. Lehms, R. Kaden, D. Mross, D. Hass).
95. M. Vecchio, *Hydrocarbon Process.* **49** (1970) 116.
96. M. Vecchio, *Hydrocarbon Process.* **52** (1973) 97.
97. Montecatini Edison S.p.A., US 3 442 962, 1966 (M. Vecchio, L. Lodi).
98. A. J. Rudge, *Chem. Ind. (London)* 1955, 452 – 461.
99. ICI, GB 633 678, 1947 (H. R. Leech, R. L. Burnett).
100. Occidental Chemical, EP 140 385, 1985 (B. F. Daniels).
101. Occidental Chemical, US 4 849 556, 1990 (B. F. Daniels, D. J. Olsen).
102. W. T. Miller, Jr., et al., *Ind. Eng. Chem.* **39** (1947) 333 – 337.
103. E. A. Belmore et al., *Ind. Eng. Chem.* **39** (1947) 338 – 342.
104. Occidental Chem. Corp., EP-A 93579, 93580, 1982 (M. S. Saran).
105. L. Scheichl, *VFDB* Z. 3 (1954) 37.
106. Hoechst, DE-OS 1 155 104, 1960 (O. Scherer, H. Kühn, G. Hörlein, K. H. Grafen); DE-OS 3 046 330, 1980 (H. Bock, J. Mintzer, J. Wittmann, J. Russow).
107. Hoechst, DE-OS 1 168 404, 1962 (O. Scherer, H. Kühn, G. Hörlein).
108. ICI, GB 767 779, 1954 (C. W. Suckling, J. Raventos).
109. *Houben-Weyl,* **5/4,** 51 – 55, 116 – 117.
110. DuPont, US 2 729 687, 1951 (J. D. Sterling).
111. Daikin Kogyo KK, JP-Kokai 85 184 033, 1984 (Y. Furutaka, T. Nakada).
112. *Chem. Eng. News* **53** (1975) no. 36, 18.
113. D. F. Halpern in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 543.
114. DuPont, US 3 145 222, 1961 (N. O. Brace).
115. R. N. Haszeldine, *J. Chem. Soc.* 1949, 2856; *J. Chem. Soc.* 1953, 3761.
116. DuPont, US 3 234 294, 1962 (R. E. Parsons); DE-OS 1 229 506, 1963 (R. E. Parsons).
117. R. E. Banks: *Fluorocarbons and their Derivatives,* McDonald Technical & Scientific, London 1970, pp. 102 – 202.
118. T. Umemoto et al., *J. Fluorine Chem.* **31** (1986) 37.
119. R. N. Haszeldine, *J. Chem. Soc.* 1951, 584.
120. D. Paskovich, P. Gaspar, G. S. Hammond, *J. Org. Chem.* **32** (1967) 833.
121. DuPont, US 3 132 185, 1962 (R. E. Parsons).
122. C. D. Bedford, K. Baum, *J. Org. Chem.* **45** (1980) 347.

123. N. S. Rao, B. E. Baker in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 321.
124. Hoechst, EP-A 52285, 1980 (K. v. Werner, A. Gisser).
125. Daikin Kogyo KK, JP-Kokai 84 048436, 1982 (S. Misaki, T. Kamifukikoshi, M. Suefuji).
126. *Beilstein* **I (4),** 694 ff.
127. B. L. Dyatkin, E. P. Mochalina, I. L. Knunyants in [12] vol. 3, p. 45.
128. R. D. Chambers, R. H. Mobbs in [11] vol. 4, p. 50.
129. A. Young in P. Tarrant (ed.): *Fluorine Chemistry Reviews,* **vol. 1** Dekker, New York 1967 p. 359. J. D. Park, R. J. McMurtry, J. H. Adams in P. Tarrant (ed.): *Fluorine Chemistry Reviews,* **vol. 2** Dekker, New York p. 55.
130. H. E. O'Neal, S. W. Benson, *J. Phys. Chem.* **72** (1968) 1866.
131. E. W. Cook, J. S. Pierce, *Nature* **242** (1973) 337.
132. W. H. Sharkey in *Fluorine Chemistry Reviews*, vol. 2, pp. 1 – 53. DuPont, US 3 316 312, 1959 (D. I. McCane, I. M. Robinson).
133. A. M. Lovelace, D. A. Rausch, W. Postelnek in: *Aliphatic Fluorine Compounds, A.C.S. Monograph* **138,** Reinhold Publ. Corp., New York 1958, Chap. III.
134. D. Osteroth: *Chemie und Technologie aliphatischer fluororganischer Verbindungen,* Chap. IB, Ferdinand Enke Verlag, Stuttgart 1964.
135. R. D. Chambers: *Fluorine in Organic Chemistry,* J. Wiley & Sons, New York – London – Sidney – Toronto 1973, pp. 144 – 148.
136. A. Pajaczkowski, J. W. Spoors, *Chem. Ind. (London)* 1964, no. 16, 659.
137. R. Müller, H. Fischer, *Z. Chem.* **7** (1967) 314.
138. DuPont, US 2 407 396, 1943 (M. M. Brubaker); US 2 407 405, 1943 (M. A. Deitrich, R. M. Joyce, Jr.).
139. B. Atkinson, V. A. Atkinson, *J. Chem. Soc.* 1957, 2086.
140. DuPont, US 3 081 245, 1960 (M. W. Farlow); Dow Chemical, US 3 133 871, 1963 (W. R. von Trees).
141. Kinetic Chemicals, US 2 401 897, 1946 (A. F. Benning, F. B. Downing, R. J. Plunkett).
142. DuPont, US 2 894 996, 1955 (M. W. Farlow, L. Muetterties).
143. Pennsalt Chem. Corp., US 3 009 966, 1960 (M. Hauptschein, A. H. Fainberg).
144. DuPont, US 2 551 573, 1945 (F. B. Downing, A. F. Benning, R. C. McHarness).
145. Hoechst, DE-OS 1 073 475, 1958 (O. Scherer, A. Steinmetz, H. Kühn, W. Wetzel, K. Grafen).
146. ICI, GB 960 309, 1962 (J. W. Edwards, S. Sherratt, P. A. Small).
147. Daikin Kogyo KK, GB 1 027 435, 1963.
148. M. C. Henry, C. B. Griffis, E. C. Stump in [12] vol. 1, pp. 1 – 75.
149. DuPont, US 2 958 685, 1957 (H. S. Eleuterio).
150. DuPont, DE-OS 1 170 935, 1961 (R. H. Halliwell); Asahi Glass Co., US 3 459 818, 1966 (H. Ukihashi, M. Hisasue).
151. DuPont, US 2 758 138, 1954 (D. A. Nelson); US 2 970 176, 1957 (E. H. Ten Eyck, G. P. Larson). NL 7 308 149, 1972 (N. E. West); Daikin, Kogyo KK, GB 1 016 016, 1963 (H. Niimiya).
152. Allied, US 2 734 090, 1950 (J. D. Calfee, C. B. Miller).
153. Kellogg Co., US 2 774 799, 1954 (M. M. Russel, W. S. Bernhart). Allied, US 2 628 989, 1951 (C. B. Miller). Hoechst, US 3 183 277, 1961 (O. Scherer, A. Steinmetz, H. Kühn, W. Wetzel, K. Grafen).
154. Bayer, DE-OS 2044370, 1970 (J. N. Meussdoerfer, H. Niederprüm).
155. Pennsalt Chem. Corp., US 3 188 356, 1961 (M. Hauptschein, A. H. Fainberg).
156. DuPont, US 2 599 631, 1945 (J. Harmon).
157. DuPont, US 2 419 010, 1943 (D. D. Coffman, T. A. Ford); Dynamit Nobel AG, DE-OS 1 259 329, 1965 (A. Bluemcke, P. Fischer, W. Krings), DE-OS 1 768 543, 1968 (A. Bluemcke, P. Fischer, M. Augenstein, H. J. Vahlensieck).
158. Montecatini, US 3 200 160, 1962 (D. Sianesi, G. Nelli). Atlantic Richfield Co., US 3 621 067, 1969 (J. W. Hamersma), US 3 642 917, 1969 (J. W. Hamersma).
159. Dow Corning Corp., US 3 739 036, 1971 (J. A. Valicenti, R. L. Halm, F. O. Stark).
160. Central Glass Co., JP-Kokai 84 108726, 1982 (T. Nakasora, H. Nakahara).
161. Halocarbon Products Corp., DE-OS 2 644 595, 1975 (L. L. Ferstanding).
162. DuPont, US 4 220 608, 1979 (A. E. Feiring).
163. A. E. Feiring, *J. Fluorine Chem.* **12** (1978) 471.
164. Dow Corning Corp., US 3 660 453, 1966 (E. D. Groenhof, H. M. Schiefer).
165. I. L. Knunyants, Y. A. Cheburkov, *Otdel. Khim. Nauk* 1960, 678.

166. Daikin Kogyo KK, JP-Kokai 81 090026, 1979 (S. Misaki, S. Takamatsu); JP-Kokai 81 138127, 1980 (S. Takamatsu, S. Misaki).
167. Allied, US 3 894 097, 1974 (N. Vanderkooi, H. J. Huthwaite).
168. A. B. Robertson, S. Chandrasekaran, C. S. Chen, *Int. SAMPE Tech. Conf. 1986,* 18 (Mater. Space-Gathering Momentum), 490 – 9.
169. *Matheson Gas Data Book,* 4th, 5th ed., Supplements, The Matheson Company, East Rutherford, N.J. 1966 – 1971.
170. Kellogg Co., US 2 590 433, 1949 (O. A. Blum); 3M, US 3 014 015, 1953 (J. W. Jewell).
171. Phillips Petroleum Co., US 3 789 016, 1966 (L. E. Gardner).
172. F. Liska, *Chem. Listy* **66** (1972) 189. E. D. Bergmann, A. M. Cohen, *Isr. J. Chem.* **8** (1971) 925.
173. A. J. Elliot in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 145.
174. S. K. De, S. R. Palit in J. C. Tatlow, R. D. Peacock, H. H. Hyman (eds.): *Advances in Fluorine Chemistry,* **vol. 6,** Butterworths, London 1970, pp. 69 – 81.
175. J. A. Young, *Fluorine Chem. Rev.* **1** (1967) 359. M. E. Redwood, C. J. Willis, *Can. J. Chem.* **45** (1967) 389.
176. W. B. Farnham et al., *J. Amer. Chem. Soc.* **107** (1985) 4565.
177. DuPont, US 3 317 616 (V. Weinmayr). R. Filler, R. M. Shure, *J. Org. Chem.* **32** (1967) 1217. B. L. Dyatkin, E. P. Mochalina, I. L. Knunyants, *Tetrahedron* **21** (1965) 2991.
178. R. D. Chambers: *Fluorine in Organic Chemistry,* Wiley-Interscience, New York 1973, pp. 65 – 66.
179. H. Gillman, R. G. Jones, *J. Amer. Chem. Soc.* **70** (1948) 1281.
180. Abbott Laboratories, US 3 970 710, 1976 (S. Wolownik).
181. Allied, US 4 072 726, 1978 (H. R. Nychka, R. E. Eibeck, M. A. Robinson, E. S. Jones).
182. D. R. Husted, A. H. Albrecht, *J. Amer. Chem. Soc.* **74** (1952) 5422.
183. I. L. Knunyants, SU 138 604, 1961. Allied, BE 634 368, 1963 (J. Hollander, C. Woolf).
184. Hoechst, DE 1 017 782, 1957 (P. Schlack). M. Goodman, I. G. Rosen, M. Safdy, *Biopolymers* **2** (1964) 503, 519, 537. DuPont, US 3 418 337, 1968 (W. J. Middleton).
185. DuPont, US 3 148 220, 1964 (D. C. England). B. S. Farah, E. E. Gilbert, J. Sibilia, *J. Org. Chem.* **30** (1965) 998.
186. J. R. Griffith, *Chemtech.* **12** (1982) 290.
187. H. C. Fielding in [6], Chap. 11.
188. Allied, US 4 273 947, 1981 (M. Novotny). Phillips Petroleum Co., US 4 396 784, 1983 (M. M. Johnson, G. P. Nowack, P. S. Hudson, B. H. Ashe, Jr.). Hoechst, DE-OS 1 944 381, 1971 (H. Fischer).
189. DuPont, US 3 283 012, 1966 (R. Day). F. Mares, B. C. Oxenrider, *J. Fluorine Chem.* **8** (1976) 373. Hoechst, DE-OS 2 318 677, 1974 (H. Millauer). N. O. Brace, *J. Fluorine Chem.* **31** (1986) 151.
190. T. Nguyen, M. Rubinstein, C. Wakselman, *Synth. Commun.* **13** (1983) 81. Daikin Kogyo Co., US 4 346 250, A 1982 (T. Satokawa, T. Fujii, A. Ohmori, Y. Fujita).
191. A. M. Lovelace, D. A. Rausch, W. Postelnek: *Aliphatic Fluorine Compounds,* Reinhold Publ. Co., New York 1958, Chap. 5.
192. A. L. Henne, S. B. Richter, *J. Amer. Chem. Soc.* **74** (1952) 5420.
193. T. Abe, S. Nagase in [7], Chap. 1.
194. L. C. Clark, F. Becattini, S. Kaplan, *Ala. J. Med. Sci.* **9** (1972) 16. E. P. Wesseler, R. Iltis, L. C. Clark, *J. Fluorine Chem.* **9** (1977) 137.
195. G. G. Gerhardt, R. J. Lagow, *J. Chem. Soc., Chem. Commun.* 1976, 259. R. D. Chambers, B. Grievson, F. G. Drakesmith, R. L. Powell, *J. Fluorine Chem.* **29** (1985) 323.
196. P. Tarrant, C. G. Allison, K. P. Barthold, E. C. Stump, Jr., *Fluorine Chem. Rev.* **5** (1971) 77.
197. H. Millauer, W. Schwertfeger, G. Siegemund, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 161.
198. DuPont, FR 1 322 597, 1963 (H. H. Biggs, J. L. Warnell). DuPont, US 3 536 733, 1970 (D. P. Carlson). Nippon Mektron, JP 57 134 473, 1982.
199. Hoechst, DE-OS 2 658 328, 1978 (M. Millauer, W. Lindner). M. Millauer, *Chem. Ing. Tech.* **52** (1980) 53.
200. Asahi Chemical, EP-OS 64 293, 1982 (M. Ikeda, M. Miura, A. Aoshima). Daikin Kogyo, JP 58 39 472, 1983; JP 58 49 372, 1983. Asahi Chemical, JP 59 78 176, 1984.
201. DuPont, US 3 358 003, 1967 (H. S. Eleuterio, R. W. Meschke). Hoechst, DE-OS 2 557 655, 1977 (R. Sulzbach, F. Heller).
202. R. C. Kennedy, J. B. Levy, *J. Fluorine Chem.* **7** (1974) 101. P. B. Sargeant, *J. Amer. Chem. Soc.* **91** (1969) 3061.
203. Asahi Glass, JP 58 62 131 A2, 1983; JP 58 62 130 A2, 1983. Hoechst, EP-A 54 227, 1982 (P. P. Rammelt, G. Siegemund). DuPont, US 4 302 608, 1981 (E. N. Squire). DuPont,

GB 1 019 788, 1966 (E. P. Moore, A. S. Milian). 3M, US 3 213 134, 1965 (D. E. Morin).
204. R. Watts, P. Tarrant, *J. Fluorine Chem.* **6** (1975) 481.
205. DuPont, US 3 125 599, 1964 (J. L. Warnell).
206. P. L. Coe, A. Sellers, J. C. Tatlow, *J. Fluorine Chem.* **23** (1983) 102. DuPont, GB 904 877, 1962.3M, US 3 213 134, 1965 (D. E. Morin). E. M. Rokhlin et al., *Dokl. Akad. Nauk SSSR* **161** (1965) 1356.
207. H. S. Eleuterio, *J. Macromol. Sci. Chem.* **A-6** (1972) 1027. J. T. Hill, *J. Macromol. Sci.* [A] Chem. 8 (1974) 499.
208. D. Sianesi, A. Pasetti, R. Fontanelli, G. C. Bernardi, G. Caporiccio, *Chim. Ind. (Milan)* **55** (1973) 221. D. Sianesi, A. Pasetti, C. Corti, *Macromol. Chem.* **86** (1965) 308. D. Sianesi, R. Tontanelli, *Macromol. Chem.* **102** (1967) 115.
209. A. Faucitano, A. Buttafava, G. Caporiccio, C. T. Viola, *J. Amer. Chem. Soc.* **106** (1984) 4172.
210. B. C. Anderson, L. R. Barton, J. W. Collette, *Science* **208** (1980) 807.
211. R. E. Uschold, *Polym. J.* **17** (1985) 253. A. L. Barney, W. J. Keller, N. M. von Gulick, *J. Polym. Sci. Part A* **1** (1970) 1091.
212. DuPont, US 3 250 808, 1966 (E. P. Moore, A. S. Milian, Jr., H. S. Eleuterio). DuPont, US 3 291 843, 1966 (C. G. Fritz, S. Selman). Hoechst, US 4 118 421, 1978 (T. Martini).
213. DuPont, US 4 281 092, 1981 (A. F. Breazeale). DuPont, US 3 467 638, 1969 (D. B. Pattison). G. H. Kalb, A. A. Khan, R. W. Quarles, A. L. Barney: *Advances in Chemistry Series, 129,* Amer. Chem. Soc., Washington, D.C., 1973.
214. A. Eisenberg, H. Yeager (eds.): "Perfluorinated Ionomer Membranes," *ACS Symposium Series, 180,* Amer. Chem. Soc., Washington, D.C., 1982.
215. DuPont, US 3 301 893, 1967 (R. E. Putnam, W. D. Nicoll).
216. DuPont US 3 282 875, 1966 (D. J. Connally, W. F. Gresham). W. Grot, *Chem. Ing. Tech.* **47** (1975) 617.
217. Asahi Glass, GB 2 029 827 A, 1979 (M. Yamabe, S. Kumai, S. Munekata).
218. D. C. England, L. R. Melby, M. A. Dietrich, R. V. Lindsey, Jr., *J. Amer. Chem. Soc.* **82** (1960) 5116.
219. I. L. Knunyants, A. I. Shchekotiklin, A. V. Fokin, *Izv. Akad. Nauk SSSR, Otd. Khim. Nauk* 1953, 282.
220. D. C. England, L. Solomon, C. G. Krespan, *J. Fluorine Chem.* **3** (1973/1974) 4.
221. D. C. England, *J. Org. Chem.* **49** (1984) 4007.
222. F. W. Evans, M. H. Litt, A.-M. Weidler-Kubanek, F. P. Avonda, *J. Org. Chem.* **33** (1968) 1839.
223. Dow Chemical, US 3 549 711, 1970 (C. I. Merrill, N. L. Madison).
224. W. G. M. Jones in [7], Chap. 5. R. Filler in R. E. Banks (ed.): *Organofluorine Chemicals and their Industrial Applications,* Ellis Horwood Ltd., Chichester 1979, Chap. 6; W. G. M. Jones in [6], Chap. 7. E. R. Larsen, *Fluorine Chem. Rev.* **3** (1969) 1.
225. Air Reduction Co., US 2 870 218, 1959 (P. W. Townsend).
226. J. M. Baden, M. Kelley, R. S. Wharton, B. A. Hitt, V. F. Simon, R. I. Mazze, *Anesthesiology* **46** (1977) 346.
227. Air Reduction Co., GB 1 138 406, 1969 (R. C. Terrell).
228. Air Reduction Co., US 3 535 425, 1970 (R. C. Terrell). A. P. Adams, *Br. J. Anaesth.* **53** (1981) Suppl. 1, 27.
229. Dow Chemical, US 3 264 356, 1966 (E. R. Larsen).
230. M. J. Halsey, *Br. J. Anaesth.* **53** (1981) Suppl. 1, 43.
231. G. A. Olah, C. V. Pittman, *J. Am. Chem. Soc.* **88** (1966) 3310.
232. *Houben-Weyl* **7 (2C),** 2147.
233. *Chem. Eng.* **75** (1968) 22.
234. Allied, US 3 164 637, 1965 (H. R. Nychka, C. Woolf).
235. DuPont, FR 1 372 549, 1964 (F. W. Swamer). *Chem. Abstr.* **62** (1965) 6397 a.
236. M. Van Der Puy, L. G. Anello, *Org. Synth.* **63** (1985) 154.
237. Allied, FR 1 369 784, 1964 (L. G. Anello, H. R. Nychka, C. Woolf). *Chem. Abstr.* **62** (1965) 1570 f.
238. I. Mellan: *Ketones,* Chemical Publ. Co., New York 1968, pp. 64 – 68.
239. W. J. Middleton, R. V. Lindsey, Jr., *J. Am. Chem. Soc.* **86** (1964) 4948.
240. Allied, US 3 620 666, 1971 (R. W. Lenz, L. Barish, V. L. Lyons).
241. USDA, US 3 998 620, 1976 (A. G. Pittman, W. L. Wasley).
242. Stauffer Chemical Co., US 3 594 362, 1971 (K. Szabo).
243. Stauffer Chemical Co., NE-A 6 410 536, 1965 (K. Szabo, G. P. Willsy, Jr.). *Chem. Abstr.* **63** (1965) 11577 g.
244. Allied, US 3 177 187, 1965 (J. Hollander, C. Woolf).

245. Allied, US 3 177 185, 1965 (J. Hollander, C. Woolf).
246. Allied, NL-A 6 407 623, 1965 (E. E. Gilbert, J. A. Otto).
247. Allied, NL-A 6 407 548, 1965 (E. E. Gilbert, B. C. Oxenrider, G. J. Schmitt).
248. L. Speers, A. J. Szur, R. C. Terrell, *J. Med. Chem.* **15** (1972) 606.
249. M. Hudlicky: *Chemistry of Organic Fluorine Compounds,* 2nd ed., Ellis Horwood, Chichester 1976, pp. 353, 712.
250. [249], p. 714.
251. Commonwealth Scientific and Industrial Research Organization, JP 54 64 728, 1979 (G. Holan). *Chem. Abstr.* **91** (1979) 56635 g.
252. Baxter, US 3 659 023, 1972 (B. M. Regan).
253. Allied, US 3 450 773, 1969 (R. E. A. Dear, E. E. Gilbert).
254. D. W. Lipp, O. Vogl in T. Saegusa, E. Goethals (eds.): *ACS Symposium Series 59,* Amer. Chem. Soc., Washington, D.C., 1977, Chapter 8, pp. 111 – 128.
255. B. Yamada, R. W. Campbell, O. Vogl, *Polym. J.* **9** (1977) 23.
256. *Beilstein* **1 (4),** 3133.
257. CS 183 467, 1980 (A. Posta et al.). *Chem. Abstr.* **94** (1981) 156316 j.
258. Hoechst, DE-OS 2 139 211, 1973 (G. Siegemund).
259. G. Siegemund, *Chem. Ber.* **106** (1973) 2960.
260. Sumitomo, JP-Kokai 60 64 949, 1985.
261. Hoechst, DE-OS 2 656 548, 1978 (G. Siegemund).
262. Hoechst, DE-OS 2 344 442, 1975 (H. Kuehn, F. Kluge, G. Siegemund).
263. Hoechst, DE-OS 2 361 058, 1975 (G. Siegemund, R. Muschaweck).
264. Hoechst, DE-OS 2 340 560, 1975 (G. Siegemund, R. Muschaweck).
265. K. C. Joshi, *J. Indian Chem. Soc.* **59** (1982) 1375.
266. K. C. Joshi, V. N. Pathak, *Coord. Chem. Rev.* **22** (1977) 37.
267. P. Mushak, M. T. Glenn, J. Savory, *Fluorine Chem. Rev.* **6** (1973) 43.
268. K. G. Seeley, W. J. McDowell, L. K. Felker, *J. Chem. Eng. Data* **31** (1986) 136.
269. K. A. Kline, R. E. Sievers, *Aldrichimica Acta* **10** (1977) 54.
270. A. L. Henne, C. J. Fox, *J. Am. Chem. Soc.* **73** (1955) 2323.
271. S. Nagase in [12], vol. 1, p. 77.
272. J. Burdon, J. C. Tatlow in [11], vol. 1, p. 129.
273. V. Childs in N. Weinberg, B. Tilak (eds.): *Technique in Electroorganic Synthesis,* **vol. 5,** Wiley-Interscience, New York 1982, Part 3, p. 341.
274. Kali-Chem, US 3 725 475, 1973 (H. Paucksch, J. Massonne, H. Bohm, G. Fernschild).
275. F. L. M. Pattison: *Toxic Aliphatic Fluorine Compounds,* Elsevier, Amsterdam 1959, p. 387.
276. M. Hauptschein, *J. Am. Chem. Soc.* **83** (1961) 2500.
277. Pennwalt, US 3 328 240, 1966 (M. Hauptschein).
278. Hoechst, US 4 400 325, 1983 (K. von Werner, A. Gisser).
279. Daikin Kogyo, US 3 525 758, 1970 (A. Katsushima).
280. Asahi Glass, US 4 138 417, 1979 (H. Ukihashi, T. Hayashi, Y. Takashi et al.).
281. DuPont, US 3 250 808, 1966 (E. Moore, A. Milian, E. Eleuterio); US 3 322 826, 1967 (E. Moore).
282. DuPont, GB 1 108 128, 1968.
283. S. Kumai, S. Samejima, M. Yamabe, *Rep. Res. Lab. Asahi Glass Co. Ltd.* **29** (1979) 105.
284. Asahi Glass, US 4 151 200, 1979 (M. Yamabe, S. Kumai).
285. A. L. Henne, W. J. Zimmersheid, *J. Am. Chem. Soc.* **69** (1947) 281.
286. E. T. McBee, P. A. Wiseman, G. B. Bachman, *Ind. Eng. Chem.* **39** (1947) 415.
287. S. Stinson, *Chem. Eng. News* 1982 (March 15) 22.
288. R. N. Hazeldine, J. M. Kidd, *J. Chem. Soc.* 1954, 4228.
289. T. Gramstad, *Tidsskr. Kjemi Bergves. Metall.* **19** (1959) 62.
290. Bayer, DE-OS 2 201 649, 1973 (P. Voss, H. Niederpruem, M. Wechsberg).
291. R. N. Hazeldine, J. M. Kidd, *J. Chem. Soc.* 1955, 2901.
292. Hercules, US 4 505 990, 1985 (S. Dasgupta et al.).
293. Eastman Kodak, US 4 582 781, 1986 (J. Chen, J. Kelly, J. Plakunov et al.).
294. 3M, US 2 803 656, 1957 (A. Ahlbrecht, H. Brown et al.).
295. 3M, US 2 803 615, 1957 (A. Ahlbrecht, H. Brown, S. Smith et al.).
296. 3M, US 3 094 547, 1963 (R. Heine et al.).
297. R. B. Ward, *J. Org. Chem.* **30** (1965) 3009.
298. 3M, US 4 329 478, 1982 (F. Behr).
299. DuPont, US 3 301 893, 1967 (R. Putnam, W. Nicoll et al.).
300. S. Torrey (ed.): *Membrane and Ultrafiltration Technology,* Noyes Data Corp., Park Ridge, N.J., 1980, p. 132.

301. A. Eisenberg, H. Yeager (eds.): *Perfluorinated Ionene Membranes,* Amer. Chem. Soc., Washington, D.C., 1982.
302. DuPont, US 3 282 875, 1964 (D. Connolly, W. Gresham).
303. DuPont, FR 1 406 778, 1963 (R. Putnam, W. Nicoll).
304. F. E. Behr, R. J. Koshar: *Abstracts Fifth Winter Fluorine Conference,* Daytona Beach, Fla., 1981, p. 5.
305. 3M, EP 0 058 466, 1982 (F. Behr, R. Koshar et al.).
306. Asahi Chemical Industry, GB 2 051 831, 1980 (K. Kimoto, H. Miyauchi, J. Ohmura, M. Ebisawa et al.).
307. C. Jackson (ed.): "Modern Chloro-alkali Technology," *3rd Int. Chlorine Symp. 1982,* Horwood, United Kingdom 1983, p. 79.
308. "Fluorinert Electronic Liquids," Product Manual, 3M Co., St. Paul, Minn., 1985.
309. R. N. Haszeldine, *Research (London)* **3** (1950) 430. R. N. Hazeldine, *Chem. Soc.* 1951, 102.
310. 3M Co., US 2 616 927, 1952 (E. A. Kauck, J. H. Simons).
311. G. G. I. Moore et al., *J. Fluorine Chem.* **32** (1986) no. 1, 41.
312. H. G. Bryce: "Industrial and Utilitarian Aspects of Fluorine Chemistry," in [10], vol. 5, pp. 302 – 346.
313. T. Abe, S. Nagase, R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Chap. 1, Ellis Horwood Ltd. & Wiley Halsted Press, New York – Brisbane – Chichester – Toronto 1982, p. 35; D. S. L. Slinn, S. W. Green, R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Chap. 2, Ellis Horwood Ltd. & Wiley Halsted Press, New York – Brisbane – Chichester – Toronto 1982, pp. 45 – 80; M. Le Blanc, J. E. Riess, R. E. Banks (ed.): *Preparation, Properties and Industrial Applications of Organofluorine Compounds,* Chap. 3, Ellis Horwood Ltd. & Wiley Halsted Press, New York – Brisbane – Chichester – Toronto 1982, pp. 83 – 116.
314. F. Swarts, *Bull. Cl. Sci. Acad. R. Belg.* **35** (1898) 375.
315. Y. Kobayashi, I. Kumadaki, *Tetrahedron Lett.* 1969, 4095.
316. Y. Kobayashi, I. Kumadaki, *Chem. Pharm. Bull.* **18** (1970) 2334.
317. V. C. R. McLaughlin, J. Thrower, *Tetrahedron* **25** (1969) 5921.
318. J. H. Clark, D. Wails, T. W. Bastock: *Aromatic Fluorination,* CRC Press Inc., Boca Raton, FL 1996.
319. F. Swarts, *Bull. Cl. Sci. Acad. R. Belg.* 1920, 399.
320. Y. Kobayashi, I. Kumadaki, *Acc. Chem. Res.* **11** (1978) 197.
321. A. L. Henne, M. S. Newman, *J. Am. Chem. Soc.* **60** (1938) 1697.
322. R. Filler, H. Novar, *Chem. Ind. (London)* 1960, 1273.
323. R. Grinter et al., *Tetrahedron Lett.* 1968, 3845.
324. Hooker Chemical, US 3 234 292, 1966 (St. Robota, E. A. Belmore).
325. R. J. Albers, E. C. Kooyman, *Rec. Trav. Chim. Pays Bas* **83** (1964) 930.
326. Montedison, DE-OS 2 746 787, 1976 (M. Bornengo).
327. G. Schiemann, B. Cornils: *Chemie und Technologie cyclischer Fluorverbindungen,* Ferdinand Enke Verlag, Stuttgart 1969, pp. 188 – 193.
328. Hoechst (IG Farbenindustrie), DE 575 593, 1931 (P. Osswald, F. Müller, F. Steinhäuser).
329. Hoechst, DE 2 356 257, 1973 (R. Lademann, H. Lindner, Th. Martini, P. Rammelt).
330. A. K. Barbour et al. in [11], vol. 3, p. 181.
331. Kinetic Chem. Inc., US 1 967 244, 1932 (L. C. Holt, E. Lorenzo). J. H. Brown et al., *J. Chem. Soc. Suppl. Issue I,* 1949, 95.
332. Hoechst, DE 1 618 390, 1967 (O. Scherer, G. Schneider, R. Hähnle).
333. Hoechst, DE 1 965 782, 1969 (H. Hermann, H. Lindner).
334. Hooker Chem. Corp., DE-AS 1 543 015, 1965 (St. Robota).
335. Kali-Chemie, DE-OS 2 161 995, 1971 (H. Böhm, J. Massonne).
336. Olin Mathieson Chem. Corp., US 3 136 822, 1961 (L. J. Frainier).
337. Bayer, DE-OS 2 837 499, 1978 (A. Marhold, E. Klauke).
338. *Ullmann,* 3rd ed., **7** 636 – 639. G. Wolfrum in [6], pp. 208 – 213.
339. T. Fujita, J. Iwasa, C. Hansch, *J. Am. Chem. Soc.* **86** (1964) 5175.
340. C. Hansch, R. M. Muir, T. Fujita et al., *J. Am. Chem. Soc.* **85** (1963) 2817. C. Hansch, J. Fukunaga, *CHEMTECH* **7** (1977) 120.
341. R. Filler in [7], pp. 123 – 153.
342. G. T. Newbold in [6], pp. 169 – 187.
343. W. A. Sheppard, C. M. Sharts: *Organic Fluorine Chemistry,* W. A. Benjamin, New York 1969, pp. 18 – 49.

344. G. G. Yakobson, T. D. Rubina, N. N. Vorozhtsov, Jr., *Dokl. Akad. Nauk SSSR* **141** (1961) 1395.
345. M. M. Boudakian et al., *J. Org. Chem.* **26** (1961) 4641.
346. US Atomic Energy Com., US 3 833 581, 1971 (D. R. McKenzie, J. Fajer, R. Snol).
347. V. Grakauskas, *J. Org. Chem.* **35** (1970) 723; **34** (1969) 2835.
348. M. J. Shaw et al., *Chem. Eng. News* **47** (1969) no. 40, 45; *J. Am. Chem. Soc.* **92** (1970) 6498.
349. K. Bottenberg, *Chem. Ztg.* **96** (1972) 84.
350. Institut Organicheskoi Khimii, Imeni Zelinskogo, DE-OS 1 618 564, 1967 (O. M. Nefedov, A. A. Ivashenko).
351. I. N. Rozhkov, A. V. Bukhtiarov, I. L. Knunyants, *Iszv. Akad. Nauk SSSR Ser. Khim.* 1972, 1130.
352. G. A. Olah, P. Kreienbühl, *J. Org. Chem.* **32** (1967) 1614.
353. K. O. Christe, A. E. Pavlath, *J. Org. Chem.* **31** (1966) 559.
354. J. S. Moilliet in R. E. Banks, B. E. Smart, J. C. Tatlow (eds.): *Organofluorine Chemistry, Principles and Commercial Applications,* Plenum Press, New York 1994, p. 195.
355. Hoechst, DE 600 706, 1933 (P. Osswald, O. Scherer). FIAT Final Rep. Nr. 998, I.G. Werke Hoechst (1947). R. L. Ferm, C. A. v. d. Werf, *J. Am. Chem. Soc.* **72** (1950) 4809.
356. Riedel-de-Haen, DE-OS 3 520 316, 1985 (E. Begemann, H. Schmand).
357. Harshaw Chemical Company, US 2 563 796, 1951 (W. J. Shenk, G. R. Pellon).
358. Allied, FR 611 545, 1960 (L. G. Anello, C. Woolf).
359. H. Suschitzky in [11], vol. 4, p. 1.
360. D. T. Flood in A. H. Blatt (ed.): *Organic Syntheses,* Collective Volume II, J. Wiley & Sons, New York 1943, p. 295.
361. J. P. Regan: *Meeting on Fluoro-organic Compounds in Industry,* Society of Chemical Industry, London 23rd January 1986.
362. A. E. Pavlath, A. J. Leffler: "Aromatic Fluorine Compounds," *ACS Monogr.* **155** (1962).
363. W. Prescott, *Chem. & Ind. (London)* 1986, 56.
364. Boots Co., DE-OS 2 938 939, 1978 (R. A. North).
365. L. Dolby-Glover, *Chem. & Ind. (London)* 1986, 518.
366. ISC Chemicals Ltd., GB 1 143 088, 1966 (D. W. Cottrell).
367. A. J. Elliott in R. Filler, Y. Kobayashi (eds.): *Biomedicinal Aspects of Fluorine Chemistry,* Kodansha Ltd., Tokyo & Elsevier Biomedical, Amsterdam – New York – Oxford 1982, pp. 55 – 74.
368. G. C. Finger, F. H. Reed, L. R. Tehon, *Illinois State Geol. Survey,* Circular No. 199, Urbana (1955). R. S. Shiley et al., *J. Fluorine Chem.* **5** (1975) 371.
369. Thomson-CSF, US 3 979 321, 1976 (A. Couttet, J. C. Dubois, A. Zaum). American Mikro-System, US 4 018 507, 1977 (N. V. V. Raaghavan). J. P. Van Meter, *Eastman Org. Chem. Bull.* **45** (1973) 1.
370. R. Filler, *J. Fluorine Chem.* **33** (1986) 361.
371. K. Wallenfels, F. Witzler, K. Friedrich, *Tetrahedron* **23** (1967) 1845.
372. Keishin Matsumoto, JP-Kokai 51 6940, 1974 (S. Fujii, K. Inoukai).
373. *Kirk-Othmer,* **10,** 915. L. S. Kobrina in [12], vol. 7, pp. 1 – 114.
374. Y. Desirant, *Bull Cl. Sci. Acad. R. Belg.* **41** (1955) 759.
375. Sperry Rand. Corp., US 3 158 657, 1962 (F. R. Callihan, Ch. L. Quatela).
376. M. Stacey, J. C. Tatlow in [11], vol. 1, p. 166.
377. G. Fuller, *J. Chem. Soc.* 1965, 6264.
378. L. W. Hall, S. R. K. Jackson, G. M. Massey, *Int. Congr. Ser. Exerpta Med.* 1974, 347.
379. D. R. Miller et al., *Intersoc. Energy Convers. Engl. Conf. Conf. Proc. 7th,* (1972) 315.
380. E. D. Morgan, C. F. Poole, *J. Chromatogr.* **104** (1975) 351.
381. J. C. Lhuguenot, B. F. Maume, *J. Chromatogr. Sci.* **12** (1974) 411.
382. M. M. Murza et al., *Zh. Org. Khim.* **13** (1977) 1046.
383. DuPont, FR 1 555 360, 1969 (D. B. Pattison).
384. H. C. Brown, D. McDaniel, *J. Am. Chem. Soc.* **77** (1955) 3752.
385. Olin, US 3 296 269, 1967 (M. M. Boudakian).
386. Olin, US 3 703 521, 1972 (M. M. Boudakian).
387. ICI, DE-OS 2 128 540, 1971 (A. Nicolson).
388. M. M. Boudakian, *J. Heterocycl. Chem.* **5** (1968) 683.
389. J. A. Burdon et al., *Nature (London)* **186** (1960) 231.
390. G. G. Yakobson, T. D. Petrova, L. S. Kobrina in [7] vol. 7, pp. 115 – 223.
391. ICI, GB 1 479 721, 1973 (D. W. R. Headford, J. W. Slater, R. L. Sunley, R. D. Bowden, M. B. Green).
392. Dow, DE 2 263 429, 1975 (R. H. Rigterink); US 3 810 902, 1974 (R. H. Rigterink).
393. ICI, US 4 063 926, 1977 (Sp. Tomlin, J. W. Slater, D. Hartley); US 3 850 943, 1974 (R. D. Bowdon, J. W. Slater, B. G. White).

394. M. S. Raasch, *J. Org. Chem.* **27** (1962) 1406.
395. E. T. McBee, H. B. Hass, E. M. Hodnett, *Ind. Eng. Chem.* 2nd ed. 39 (1947) 389.
396. Ishihara, DE-OS 2 812 571, 1979 (R. Nishiyama, T. Haga, N. Sakashita).
397. W. Harms in [6], pp. 188 – 207.
398. Acad. Sci. USSR, DE 2 014 216, 1971 (J. L. Knuny-ants, L. S. German, N. B. Kazmina).
399. PCR, US 3 954 758, 1976 (P. D. Schuman, P. Tarrant, D. A. Warner, G. Westmoreland); US 4 113 949, 1978 (P. D. Schumann, P. Tarrant, D. A. Warner, G. Westmoreland).
400. D. V. Santi, A. L. Pogolotti, Jr., E. M. Newman, Y. Wataya in [367], pp. 123 – 142.
401. Bayer, FR 1 546 305, 1968 (E. Klauke, H. S. Bien); LU 58 622, 1969 (E. Klauke, H. S. Bien).
402. Bayer, GB 1 158 300, 1966 (E. Klauke, H. S. Bien).
403. Bayer, GB 1 273 914, 1969 (H. U. Alles, E. Klauke, H. S. Bien).
404. Bayer, DE 1 044 091, 1957 (A. Dorlars).
405. C. W. Tullock, A. D. Coffman, *J. Org. Chem.* **25** (1960) 2016.
406. Bayer, BE 713 937, 1967 (H. S. Bien, E. Klauke, K. Wunderlich); BE 716 013, 1967 (H. S. Bien, E. Klauke, K. Wunderlich); BE 716 014, 1967 (H. S. Bien, E. Klauke, K. Wunderlich).
407. Bayer, GB 872 313, 1957 (A. Dorlars).
408. G. Schiemann, E. Bolsted, *Chem. Ber.* **61** (1928) 1403.
409. J. W. Clayton: "Fluorocarbon Toxicity and Biological Action," in P. Tarrant (ed.): *Fluorine Chemistry Reviews,* vol. 1, Dekker, New York 1967, pp. 197 – 252.
410. R. R. Montgomery, C. F. Reinhardt: "Toxicity of Fluorocarbon Propellants," in P. A. Sanders (ed.): *Handbook of Aerosol Technology,* 2nd ed., Van Nostrand Reinhold, New York 1979, pp. 466 – 511.
411. D. M. Aviado: "Fluorine-Containing Organic Compounds," in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., vol. 2B, Wiley-Interscience, New York 1981, pp. 3071 – 3115.
412. J. W. Clayton: "The Mammalian Toxicology of Organic Compounds Containing Fluorine," in F. A. Smith (ed.): *Handbook of Experimental Pharmacology,* vol. 20/1, Springer Verlag, New York 1966, pp. 459 – 500.
413. ACGIH (ed.): *Documentation of the Threshold Limit Values (TLV),* ACGIH, Cincinnati, Ohio 1986.
414. DFG (ed.): *Maximum Concentrations at the Workplace and Biological Tolerance Values for Working Materials,* Verlag Chemie, Weinheim 1984, Report 20.
415. M. H. Litchfield, E. Longstaff, *Food Chem. Toxicol.* **22** (1984) 465.
416. J. F. Borzelleca, D. Lister, *Toxicol. Appl. Pharmacol.* **7** (1965) 592.
417. L. Stryer: *Biochemistry,* Freeman, San Francisco 1981, p. 298.
418. H. Millauer, W. Schwertfeger, G. Siegemund, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 161.
419. R. Frey, H. Beisbarth, K. Stosseck: *5th International Symposium on Perfluorochemical Blood Substitutes,* Mainz, Mar. 1981, Zuckschwerdt Verlag, Munich 1981, pp. 86 – 90.

Fluorine-Containing Polymers → Fluoropolymers, Organic
Fluoroboric Acid → Boron Compounds

Fluoropolymers, Organic

Fluorine Compounds, Organic is a separate keyword

D. Peter Carlson, E.I. Du Pont de Nemours & Co., Polymer Products Dept., Experimental Station Laboratory, Wilmington, Delaware, United States

Walter Schmiegel, E.I. Du Pont de Nemours & Co., Polymer Products Dept., Experimental Station Laboratory, Wilmington, Delaware, United States

1. Introduction . . . 3
2. Fluoroplastics . . . 4
2.1. Introduction . . . 4
2.2. Polytetrafluoroethylene . . . 5
2.2.1. Production . . . 5
2.2.2. Properties . . . 6
2.2.3. Processing . . . 10
2.2.4. Uses . . . 11
2.3. Tetrafluoroethylene–Hexafluoropropylene Copolymers . . . 11
2.3.1. Production . . . 12
2.3.2. Properties . . . 12
2.3.3. Processing . . . 14
2.3.4. Uses . . . 14
2.4. Tetrafluoroethylene–Perfluoro-(Propyl Vinyl Ether) Copolymers . . . 14
2.4.1. Production . . . 15
2.4.2. Properties . . . 15
2.4.3. Processing . . . 16
2.4.4. Uses . . . 17
2.5. Tetrafluoroethylene–Ethylene Copolymers . . . 17
2.5.1. Production . . . 18
2.5.2. Properties . . . 19
2.5.3. Processing . . . 20
2.5.4. Uses . . . 21
2.6. Polychlorotrifluoroethylene . . . 21
2.6.1. Production . . . 21
2.6.2. Properties . . . 21
2.6.3. Processing . . . 22
2.6.4. Uses . . . 22
2.7. Chlorotrifluoroethylene–Ethylene Copolymers . . . 22
2.7.1. Production . . . 22
2.7.2. Properties . . . 23
2.7.3. Processing . . . 23
2.7.4. Uses . . . 24
2.8. Poly(Vinylidene Fluoride) . . . 24
2.8.1. Production . . . 24
2.8.2. Properties . . . 24
2.8.3. Processing . . . 25
2.8.4. Uses . . . 25
2.9. Poly(Vinyl Fluoride) . . . 26
2.9.1. Production . . . 26
2.9.2. Properties . . . 26
2.9.3. Processing . . . 27
2.9.4. Uses . . . 27
3. Fluoroelastomers . . . 27
3.1. Introduction . . . 27
3.2. Elastomers Based on Vinylidene Fluoride . . . 28
3.2.1. Production . . . 29
3.2.2. Properties . . . 30
3.2.3. Processing . . . 31
3.2.4. Uses . . . 34
3.3. Elastomers Based on Tetrafluoroethylene–Perfluoro(Methyl Vinyl Ether) Copolymers . . . 35
3.3.1. Production . . . 35
3.3.2. Properties . . . 35
3.3.3. Processing and Uses . . . 36
3.4. Elastomers Based on Tetrafluoroethylene–Propylene Copolymers . . . 36
3.4.1. Production . . . 37
3.4.2. Properties . . . 37
3.4.3. Processing . . . 37
3.4.4. Uses . . . 38
4. Toxicology and Occupational Health 38
4.1. Fluoroplastics . . . 38
4.2. Fluoroelastomers . . . 38
5. Economic Aspects . . . 39
6. References . . . 39

Abbreviations used in this article:

CTFE chlorotrifluoroethylene copolymer
ECTFE ethylene–chlorotrifluoroethylene copolymer
ETFE ethylene–tetrafluoroethylene copolymer
FEP fluorinated ethylene–propylene

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_393

	copolymer
HFP	hexafluoropropene
HFPI	hexafluoropropene index
LOI	limiting oxygen index
MI	melt-flow index
P	propene
PCTFE	polychlorotrifluoroethylene
PFA	perfluoroalkoxy copolymer
PMVE	perfluoro(methyl vinyl ether)
PPVE	perfluoro(propyl vinyl ether)
PTFE	polytetrafluoroethylene
PVC	poly(vinyl chloride)
PVDF	poly(vinylidene fluoride)
PVF	poly(vinyl fluoride)
SSG	standard specific gravity
TFE	tetrafluoroethylene
VDF	vinylidene fluoride
VF	vinyl fluoride

1. Introduction

History. The fluoropolymer industry owes its start to a fortunate accident in Du Pont's Jackson Laboratory in 1938. Roy Plunkett was attempting to prepare new refrigerant gases from tetrafluoroethylene (TFE). However, one morning, when the valve to the almost full TFE cylinder was opened, no gas came out. After being satisfied that the valve had not malfunctioned, he had the cylinder cut open. What he found was the first sample of polytetrafluoroethylene (PTFE), a remarkable polymer, which has become an essential material in the world today. Initially, no applications were known for this new material. However, the Manhattan Project of World War II required a material sufficiently resistant to the highly corrosive substance UF_6, an intermediate in the production of ^{235}U-enriched uranium. Polytetrafluoroethylene appeared to fill this need, spurring the development of processing and production methods for the new fluoroplastic. In 1946, PTFE was introduced commercially under the tradename Teflon. This new plastic, with its outstanding chemical, electrical, and surface properties, gained immediate acceptance in the marketplace. Since then, other fluoropolymers have been discovered and commercialized. These products include fluoroplastics and fluoroelastomers and serve an ever-expanding market, providing materials with the unique combination of properties which fluoropolymers alone can provide.

Monomers. Only four fluoroolefins have been polymerized to high molecular mass homopolymers of commercial importance. These monomers are tetrafluoroethylene (TFE), vinyl fluoride (VF), vinylidene fluoride (VDF or VF_2), and chlorotrifluoroethylene (CTFE). However, other monomers have been copolymerized with these fluoroolefins. Such copolymers have created a diversity of commercially available fluoropolymers, including both plastics and elastomers. The principal monomers from which most commercial fluoropolymers are made are listed in Table 1. The fluoromonomers are described in detail elsewhere in this volume (→ Fluorine Compounds, Organic). Several other monomers are used in small amounts as modifiers and to provide cure sites.

Polymers. Among the fluoroplastics (semicrystalline polymers with $mp > 150\,°C$), several forms of PTFE and a variety of TFE copolymers are commercially available. Others are based on VF, VDF, and CTFE. In the case of fluoroelastomers (amorphous polymers with $T_g < 0\,°C$), VDF–HFP (hexafluoropropylene) copolymers and VDF–HFP–TFE terpolymers constitute the bulk of today's commercial products. Other important fluoroelastomers include TFE copolymers with propene (TFE–P), and perfluoro(methyl vinyl ether), (TFE–PMVE), fluorosilicones, and fluorophosphazenes. A special group of TFE copolymers that contain sulfonic acid and carboxylic acid functions are commercially available in the form of ion-exchange membranes (→ Fluorine Compounds, Organic, Chap. 8.2.).

Production. The principal method for synthesizing organic fluoropolymers is free-radical polymerization. Because of the electrophilic nature of fluoroolefins, cationic polymerization catalysts are not effective. Fluoroolefins and epoxides can be polymerized by anionic catalysts, but termination by fluoride ion elimination prevents high molecular masses from being attained. Coordination catalysts do not effect polymerization of fluoroolefins.

The polymerization and processing techniques described in this article are described in

Table 1. Monomers used in commercial fluoropolymers

Compound	CAS registry number	Formula	Abbreviation
Ethylene	[*74-85-1*]	$CH_2{=}CH_2$	E
Tetrafluoroethylene	[*116-14-3*]	$CF_2{=}CF_2$	TFE
Chlorotrifluoroethylene	[*79-38-9*]	$CF_2{=}CClF$	CTFE
Vinylidene fluoride	[*79-38-7*]	$CF_2{=}CH_2$	VDF (VF_2)
Vinyl fluoride	[*75-02-5*]	$CFH{=}CH_2$	VF
Propene	[*115-07-1*]	$CH_3CH{=}CH_2$	P
Hexafluoropropene	[*116-15-4*]	$CF_3CF{=}CF_2$	HFP
Perfluoro(methyl vinyl ether)	[*1187-93-5*]	$CF_3OCF{=}CF_2$	PMVE
Perfluoro(propyl vinyl ether)	[*1623-05-8*]	$CF_3CF_2CF_2OCF{=}CF_2$	PPVE

detail elsewhere (→Polymerization Processes, → Plastics, Processing).

Properties. The properties of fluoropolymers vary with the fluorine content and with the distribution of fluorine atoms in the molecule. The carbon–fluorine bond is one of the strongest organic bonds. Substitution of fluorine for hydrogen in a hydrocarbon polymer generally increases chemical and oxidative stability, weatherability, and melting point. The presence of fluorine reduces flammability, solubility in hydrocarbons, adhesiveness, radiation resistance, and the coefficient of friction. Each fluoropolymer has its own unique combination of properties. Cost generally increases with fluorine content.

Uses. Numerous uses have been found for fluoropolymers, many of which cannot be satisfied by any other material.

Fluoroplastics find their greatest utility in electrical applications such as insulation, flexible printed circuits, and piezoelectric devices. Other applications include chemically resistant coatings, bearings, nonstick surfaces, architectural fabric (dome construction), laboratory ware, heat exchangers, heart valves and other prosthetic devices, and water-repellent fabric.

Fluoroelastomers have found their greatest use in sealing applications, such as O-rings, gaskets, valve stem seals, and shaft seals. They are also used in flue duct expansion joints, as linings for fuel hoses, binders for military flares, and processing aids for polyethylene film extrusion.

2. Fluoroplastics

2.1. Introduction

The family of thermoplastic fluoropolymers is headed by polytetrafluoroethylene (PTFE). Other members include TFE−HFP copolymers (FEP), TFE−PPVE copolymers (PFA), and TFE−ethylene copolymers (ETFE), as well as polychlorotrifluoroethylene (PCTFE) and CTFE−ethylene copolymers (ECTFE). Poly(vinyl fluoride) (PVF), poly(vinylidene fluoride) (PVDF), and copolymers of VDF are also important.

Since the high melt viscosity of PTFE resins prevents processing by conventional extrusion and molding techniques, they are processed by sintering techniques similar to those used in powder metallurgy. Although PTFE can be converted into practically any form (rods, billets, films, fibers, tubing, or coatings) by these techniques, the lack of melt processibility is frequently a handicap. Research on melt-processible forms of PTFE began almost immediately after its discovery [5]. The first to be developed was a TFE−HFP copolymer called FEP (fluorinated ethylene–propene) copolymer [6]. The FEP resins, commercialized in 1960, share nearly all the properties of PTFE, but are also processible by extrusion and injection molding. However, the melting point of FEP is ca. 60 – 70 °C lower than that of PTFE; its maximum use temperature is also lower. Further research on TFE copolymers led to the development of perfluoroalkoxy (PFA) resins, which are TFE−PPVE copolymers [7] that were introduced commercially in 1972. These have the same high-temperature properties as PTFE resins and an even lower melt viscosity than

FEP resins; they can be processed by extrusion and injection molding. In 1970 ETFE resins (TFE–ethylene copolymers) were commercialized [8]. They exhibit improved mechanical properties and far better processibility than other TFE copolymers. In addition, they can be cross-linked by high-energy radiation. The maximum use temperature, however, is lower than that of PFA or FEP.

Among the fluoroplastics that contain monomers other than TFE, PCTFE was the first to be commercialized. Extensive development work was carried out on this polymer during World War II in conjunction with the Manhattan Project [9]. Compared to TFE resins, PCTFE is harder, more resistant to creep, and less permeable. It has the lowest permeability to moisture of any plastic. Chlorotrifluoroethylene copolymerized with VDF provides an improved resin for film manufacture. However, ECTFE is the most important CTFE copolymer [10]. These resins are similar in properties and uses to ETFE copolymers.

Poly(vinyl fluoride) (PVF) contains the smallest amount of fluorine (41.3 %) of any commercial fluoropolymer, but possesses many of the properties of more highly fluorinated polymers [11]. It was commercialized in 1961 by Du Pont under the trade name Tedlar. It is used mainly in coatings for metal, plastic, paper, and similar substrates to provide resistance to weather, chemicals, staining, and abrasion.

Poly(vinylidene fluoride) (PVDF) shares many of the characteristics of other fluoropolymers such as thermal and oxidative stability, as well as outstanding weatherability [12]. However, the arrangement of alternate fluorine and hydrogen atoms leads to unusual polarity within the polymer chains, with a dramatic effect on dielectric properties and solubility. An unusual product made from PVDF is a film with piezoelectric properties. Copolymerization of VDF with a small amount (< 15 %) of HFP reduces stiffness and improves processibility for certain wire coating applications.

2.2. Polytetrafluoroethylene

Polytetrafluoroethylene [*9002-84-0*] (PTFE) is a straight-chain polymer of tetrafluoroethylene (TFE) of the general formula

$$-(CF_2CF_2)_n-$$

It has a high crystalline melting point (327 °C), very high melt viscosity (10–100 GPa · s at 380 °C), and a high maximum use temperature (> 260 °C). In addition, it exhibits unusual toughness down to very low temperatures (< −200 °C); its molecular mass is extremely high ($10^6 - 10^7$). It is insoluble in all known solvents and resists attack by most chemicals. Dielectric loss is low, whereas dielectric strength is high; antistick and antifriction properties are most unusual. Although these properties give PTFE great commercial value, they also rule out processing by conventional thermoplastic techniques.

Polytetrafluoroethylene is produced in three forms: *Granular PTFE* is prepared by suspension polymerization. The coarse product is cut to uniform particle size and processed by powder metallurgy techniques. *Fine powder resins* are prepared by dispersion polymerization and coagulation by high-speed agitation. The fine powder resins consisting of uniform particles of 0.2 µm diameter are mixed with a lubricant such as kerosene and processed by a cold extrusion method called paste extrusion. The third form of PTFE is an *aqueous dispersion.* The dispersion products are processed by latex processing techniques (dip coating or spray coating) followed by baking at high temperature which causes coalescence into a continuous film.

Manufacturers of PTFE resins include Ausimont (Algoflon and Halon), Daikin Kogyo (Polyflon), Du Pont (Teflon), Hoechst (Hostaflon), ICI (Fluon), and the former Soviet Union (Ftoroplast). The People's Republic of China also manufactures PTFE products.

2.2.1. Production

Polymerization. Tetrafluoroethylene must be polymerized with great care to avoid thermal decomposition of the monomer, which can disproportionate explosively to carbon and carbon tetrafluoride:

$$CF_2{=}CF_2 \longrightarrow C + CF_4 \quad \Delta H = -316.4\,\text{kJ/mol}$$

This reaction, which generates the same explosive force as gun powder, can be initiated by hot spots formed during polymerization. Because of the high heat of polymerization of TFE

and the excellent insulating qualities of the polymer, hot spots can develop if mixing is not adequate or polymerization is not controlled. Therefore, equipment and polymerization vessels must be adequately barricaded to prevent injury or death in the event of monomer decomposition. Polymerization in storage vessels can be prevented by the addition of inhibitors such as terpenes, mercaptans, or amines. However, these must be completely removed from the TFE before use.

Tetrafluoroethylene is polymerized in aqueous medium by two different processes. In the first, little or no dispersing agent is employed and vigorous agitation produces a coagulated granular resin. In the second procedure, a fluorinated surfactant is employed and agitation is very mild, producing small spherical particles dispersed in the aqueous medium. In the latter process, coagulation of the dispersion is avoided until after polymerization is completed. Subsequent precipitation of the particles produces fine powder resins which are converted to finished products by paste extrusion. The dispersion can also be concentrated and stabilized by addition of a nonionic surfactant. These stabilized aqueous dispersions can be applied by spray or dip coating.

Granular Resins. Granular resin was the first commercial PTFE product. The first granular resins were made from pure TFE with no other ingredients except initiator. Currently, however, some granular resins contain traces of certain comonomers such as PPVE [13], [14]. Sometimes granular polymerization is carried out in the presence of a small amount of a dispersing agent. Alkaline buffering agents are used occasionally. In the early stages of granular polymerization, a dispersion is formed, but the polymer soon coagulates. Subsequent polymerization occurs on the solid particles in a gas–solid reaction in which the water acts primarily as a heat-transfer medium. The polymerized product consists of stringy, irregularly shaped particles, which can exceed a millimeter in size. However, the specific surface area of these particles is over $2\,m^2/g$, which is ca. 1700 times the observed outer surface area. Therefore, the particles must be very porous and spongelike, as is confirmed in photomicrographs.

The "as-polymerized" products cannot be used as such, but must be cut to uniform particle size to provide sinterability. Coarse cut products have good powder flow, but leave a high void content after sintering. Finely cut resins give void-free moldings, but have reduced powder flow. A good compromise between ease of handling and ease of molding is achieved by reducing the particle size to 400 – 800 μm [15]. For some applications (ram extrusion of tubes and rods), the cut resins are partially presintered for better handling.

Fine Powder Resins. Fine powder resins are produced by polymerizing TFE in the presence of a dispersing agent such as ammonium perfluorooctanoate [16]. The process partially resembles emulsion polymerization. The polymer is highly crystalline and has little ability to solubilize the monomer. In addition, the polymer is insoluble in the monomer, as is typical of an emulsion polymerization. Many of the principles of emulsion polymerization apply to this process. Initially, only pure TFE polymers were prepared in this way, but products containing traces of comonomers such as HFP, CTFE, and PPVE are also made [17], [18]. The properties of these modified fine powders vary with the type and manner of addition of the comonomer.

The dispersion product must not coagulate during polymerization but must be sufficiently unstable to allow coagulation at the end of the polymerization cycle. Very gentle stirring prevents premature coagulation during polymerization. The average particle size is ca. 0.2 μm. The specific surface area is ca. $12\,m^2/g$, which agrees well with calculated values. Unlike the granular products, the dispersion particles display little, if any, porous structure. During the coagulation phase, the nonwettable polymer floats to the surface, where it is drawn off and dried.

Aqueous Dispersions. Aqueous dispersions are made by the same polymerization process used to prepare fine powder resins. The crude dispersions can be polymerized to give different average particle sizes. The most commonly used dispersion has an average particle size of ca. 0.2 μm. The "as-polymerized" dispersion must be stabilized and concentrated before use. A nonionic surfactant such as Triton X-100 is normally employed for stabilization. The dispersion

is concentrated thermally to 60 – 65 % solids. Other additives such as thiotropic agents and fillers may be used to modify the dispersion further.

Filled Resins. A variety of fillers are utilized by PTFE manufacturers and processors; they include glass fibers, graphite, asbestos, molybdenum sulfide, and bronze. Fillers can be blended dry with the PTFE powder before molding or can be coagulated with the dispersion to produce a filled fine powder. The filler serves to improve wear resistance, increase hardness, or reduce creep.

2.2.2. Properties

Physical Properties. The structure of PTFE chains is unusual in that they are completely linear. The branching that occurs during the free-radical polymerization of polyethylene cannot take place during the polymerization of TFE. Furthermore, PTFE chains are stiffer than polyethylene chains because their fluorine atoms are larger than hydrogen atoms. Steric hindrance prevents a PTFE chain from assuming a planar zigzag structure; instead, it is forced to adopt a zigzag structure with a helical twist along the chain axis. Crystals of PTFE are made up of chains with a 180° twist every 13 to 15 carbon atoms, depending on the temperature. The high melting point of PTFE (327 °C) is due to the small entropy change produced during melting, which in turn results from the stiffness of the chains [19].

The chemical bonding forces within the chains, between the chains, and the polymer surface are unusual:

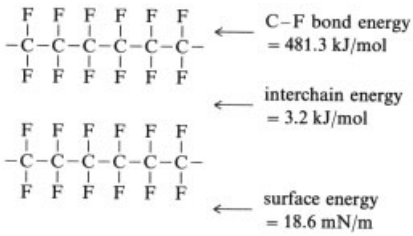

The C–F bond energy is among the highest known. However, the interchain bonding forces are very weak, and the surface energy is very low. This combination of unusually strong and weak forces results in many unique properties.

The melting point of once-melted PTFE is 327 °C, but high molecular mass, virgin unmelted PTFE melts at a higher temperature (342 °C). Differential scanning calorimetry (DSC) indicates that the 342 °C melting point is irreversible and the subsequent melting point is 327 °C. Electron-microscopic studies of dispersion particles indicate that the rodlike crystals in virgin PTFE are fully extended with few defects [20]. Virgin PTFE particles are 92 – 98 % crystalline. The crystallinity of once-melted PTFE varies with molecular mass and the rate of cooling from the melt. However, commercial grades are usually 45 – 75 % crystalline.

Themolecular mass of PTFE cannot be determined by the usual methods that require dissolution of the polymer because there is no known solvent for PTFE below 300 °C. However, the number-average molecular mass has been estimated from the determination measurements of end groups. Because the molecular mass of PTFE is very high, the concentration of end groups is very low. A quantitative method must be sensitive to less than 1 ppm of end groups. An elegant method tags the end groups with radioactive sulfur in the initiator [21]; the sulfur from the initiator must be retained in the polymer with no loss by hydrolysis. When persulfate is used as the initiator, the resulting sulfate end groups are hydrolyzed to carboxylic acid end groups and no radioactive sulfur is detected in the polymer. However, when an iron–bisulfite initiator is used, the sulfur is retained in the polymer because the initiating species (a bisulfite radical, $HOSO_2\cdot$) gives rise to hydrolytically stable sulfonic acid end groups.

A series of polymers was prepared, using the radioactive bisulfite–iron initiator, with molecular masses in the range of industrial interest. The number-average molecular masses were calculated from the radioactive sulfur content, assuming two sulfonic acid groups per polymer chain. The calculated molecular mass ranged from $3 \cdot 89 \times 10^5$ to $8 \cdot 9 \times 10^6$. These samples were used to calibrate other methods for measuring the molecular mass of PTFE resins, such as the standard specific gravity (SSG) method [22]. Since the rate of crystallization varies inversely with molecular mass, samples of high molecular mass have a lower SSG than those

of low molecular mass; typical values for polymers of commercial interest range from 2.14 to 2.20. The SSG method is widely used with PTFE resins. The correlation between number-average molecular mass $\bar{M}_n$ and the SSG is given by the following relationship:

$$\mathrm{SSG} = 2.612 - 0.058 \log_{10} \bar{M}_n$$

Studies on the melting behavior of virgin PTFE by DSC [23] has led to another method for measuring molecular mass. The following relationship was found between number-average molecular mass and heat of crystallization (ΔH_c, J/g):

$$\bar{M}_n = 2.1 \times 10^{10} \cdot \Delta H_c{}^{-5.16}$$

In addition to its crystalline melting point at 327 °C, PTFE has several transition temperatures [19]. Two crystalline transitions, at 19 and 30 °C, have a significant effect on product behavior. Below 19 °C, the polymer crystallites exist in an almost perfect three-dimensional array. Above 19 °C, the triclinic unit cell changes to a hexagonal unit cell. Slight untwisting of the chains is observed in the cell from 13 CF_2 groups to 15 CF_2 groups per 180° twist. The transition at 30 °C is only ca. one-tenth as large as that at 19 °C. The hexagonal unit cell disappears above 30 °C, but the rodlike hexagonal packing of the chains in the lateral direction is retained. As measured by torsion pendulum, two important amorphous transitions occur at ca. −97 °C and +127 °C, these transition temperatures are dependent on frequency. The lower transition ranges from −110 to −73 °C when measured at frequencies from 0.033 to 90 Hz. The higher temperature transition ranges from 120 to 140 °C over the same frequency range.

Mechanical Properties. The mechanical properties of PTFE are affected by processing variables such as preform pressure, sintering temperature, and cooling rate, as well as by polymer variables such as molecular mass, particle size, and particle-size distribution. Properties affected are the flex life, tensile strength, elongation, permeability, stiffness, resilience, and impact strength. Properties such as the coefficient of friction, low-temperature flexibility, and thermal stability are relatively unaffected. Typical values for the mechanical properties of granular and fine powder PTFE resins are listed in Table 2. Further data are given in [24–26].

The crystalline transition at 19 °C causes a volume change of 1.0 – 1.8 % and must be taken into account when precision parts made from PTFE are designed.

The coefficient of friction of PTFE is unusually low. Static friction decreases with increasing load and is less than the dynamic friction which frees PTFE from stick – slip problems. Because of the low surface energy, only liquids with extremely low surface tension (e.g., perfluorocarbon acids) can wet a PTFE surface. Surfaces can be made wettable by treatment with alkali-metal compounds (e.g., sodium naphthenate) which, however, also increases the coefficient of friction.

The effect of fillers on the mechanical properties of PTFE is shown in Table 3; tensile strength, elongation, and the coefficient of friction are not significantly affected. In contrast, wear resistance is dramatically improved: the wear factor of the bronze-filled material is only 1/250th that of unfilled PTFE.

Electrical Properties. The very low dielectric constant of PTFE (2.1) does not change over a wide range of temperature (−40 to +250 °C) and frequency (5 Hz to 10 GHz). Such electrical behavior is attributed to PTFE's highly symmetrical structure composed solely of carbon–fluorine bonds, in which all electrical dipoles are exactly balanced. Similarly, the dissipation factor is extremely low and independent of temperature or frequency. Resistivity and dielectric strength, on the other hand, are very high, as is arc resistance; PTFE leaves no carbonized path. The electrical properties of PTFE are given in Table 4.

Electrical properties can be adversely affected by voids, which may reduce the dielectric strength and corona resistance of fabricated parts. Voids must be minimized to achieve optimum electrical insulation.

Chemical Resistance. Polytetrafluoroethylene is resistant to attack by most chemicals, including aqua regia, hot fuming nitric acid, hot caustic, gaseous chlorine, chlorosulfonic acid, organic esters, ketones, and alcohols. The only materials known to attack PTFE are molten alkali metals, chlorine trifluoride, and gaseous

Table 2. Mechanical properties of PTFE resins

Property	Granular	Fine Powder	ASTM
Tensile strength, MPa	7 – 28	18 – 24	D 638
Elongation, %	100 – 200	300 – 600	D 638
Flexural modulus, MPa	350 – 630	280 – 630	D 747
Flexural strength, MPa	no break		D 790
Impact strength, J/m	160		D 256
Compressive stress at 1 % deformation, MPa	4.2		D 695
Hardness, Shore D	50 – 65	50 – 65	D 1706
Coefficient of thermal expansion, K^{-1}	12×10^{-5}		D 696

Table 3. Properties of unfilled and filled PTFE resins

Property	Unfilled resin	Filled resin	
		25 % glass	60 % bronze
Relative density	2.18	2.24	3.74
Tensile strength, MPa	28	18	14
Elongation, %	350	250	150
Impact strength, J/m	152	119	
Hardness, Shore D	51	57	70
Wear factor, mPa^{-1}	5013	26.2	12
Coefficient of friction			
Static, 3.4 MPa load	0.08	0.13	0.10
Dynamic, u=900 m/s	0.01	−0.24	−0.22
Thermal conductivity, mW m^{-1} K^{-1}	0.24	0.45	0.46

Table 4. Electrical properties of PTFE resins

Property	Granular	Fine powder	ASTM
Dielectric constant, 60 Hz to 2 GHz	2.1	2.1	D 150
Dissipation factor, 60 Hz to 2 GHz	0.0003		D 150
Volume resistivity, $\Omega \cdot$ cm	$>10^{18}$	$>10^{18}$	D 257
Surface resistivity, Ω/m^2	$>10^{16}$		D 257
Surface arc resistance, s	>300	>300	D 495
Dielectric strength, 2-mm thickness, kV/mm	23.60	23.60	D 149

fluorine at elevated temperature and pressure. Some highly fluorinated oils can swell or dissolve PTFE near its melting point. A list of chemicals that are compatible with PTFE is available [27].

Flame Resistance. Polytetrafluoroethylene is one of the most flame-resistant polymers known and does not support combustion in air. It has an extremely high limiting oxygen index (LOI) of 96 %, i.e., it burns only in almost pure (96 vol%) oxygen.

Weatherability. The weatherability of PTFE is remarkable; it is completely unaffected by all types of weather from dry desert heat to humid jungle conditions. Test films showed no change after 20 – 30 years of continuous exposure in Florida.

Radiation Resistance. On exposure to a high-energy electron beam or gamma radiation, PTFE undergoes degradation rather than cross-linkage. In the absence of oxygen, stable radicals are produced which slow down the rate of degradation. However, in the presence of air,

the radicals react with oxygen which accelerates scission and degradation. On exposure to 10^4 Gy (1 Mrad) of radiation in air, PTFE loses 50 % of its original tensile strength. Irradiation also affects its electrical properties: resistivity decreases, whereas the dielectric constant and dissipation factor increase.

Thermal Stability. In both air and nitrogen, PTFE has an extremely high thermal stability. Rates of decomposition are not measurable below ca. 440 °C, decomposition rates are high at 540 °C. In vacuum pyrolysis of thin films, first-order thermal decomposition occurs, the TFE monomer being the main product [28]; the activation energy is 347.4 kJ/mol. The melt viscosity decreases during pyrolysis which probably involves random chain cleavage followed by depolymerization (short kinetic chain length) and termination by disproportionation.

2.2.3. Processing

Granular Resins. Moldings are produced from granular resins in three steps: First, the dry powder is placed in a mold and compressed at moderate temperature to produce a preform that is strong enough for handling without breaking. The preform is than placed in an oven at 380 °C to allow the particles to coalesce. This operation is called sintering. The final step is controlled cooling to produce the desired crystallinity. Granular resins are available in a number of forms that are optimized for different types of molding. High-flow resins are used in small and automatic moldings. Finely divided resins are preferred for large billet moldings to which they impart superior properties. Presintered resins are easier to handle and are preferred in ram extrusion applications. Most PTFE manufacturers give detailed descriptions of molding equipment and procedures [29].

Automatically molded articles require no further finishing after molding. However, large billets and rods are subsequently skived or machined. Film and sheet are produced by skiving a billet on a lathe. Precision parts can be machined from ram-extruded rods. Another method, called coining, is used for articles that are too complicated to make by machining. In this operation, a sintered molding is heated to the melting point, quickly pressed into a mold cavity, and held under pressure until it resolidifies. Coined moldings are limited in their upper service temperature. After long periods at high temperature, they return to their original precoining shape.

Fine Powder Resins. The paste-extrusion process for fine powder resins represented an important advance in PTFE processing. It permits the manufacture of continuous PTFE tubing and continuous PTFE coatings on wires.

In paste extrusion, 15 – 25 wt % of lubricant (a petroleum fraction, usually kerosene) is mixed with the fine powder resin. The wetted powder is then gently shaped into a preform at low pressure. (Fine powder resins are extremely sensitive to shear; great care must be taken to avoid shearing the powder before extrusion, otherwise processibility may be lost.) The preform is forced through a die mounted in the extruder. The high shear exerted at the die fibrillates the powder and confers "green strength" to the extrudate. The lubricant is then evaporated and the extrudate is sintered at 380 °C to coalesce the elongated particles. Drying and sintering are performed consecutively by passing the extrudate through a multistage oven located directly after the extruder. Residence time in the oven can vary from a few seconds for thin-walled wire insulation to a few minutes for large-diameter tubing. Extrusion pressure depends on the reduction ratio, the extrusion rate, the lubricant content, and the extruder characteristics.

Paste extrusion is also used to produce unsintered tape. Lubricant mixing and extrusion are the same as above. The product is extruded into rods, which are calendered on hot rolls to the desired dimensions.

Different fine powder resins have been developed for different applications. Powders that are suitable for high reduction-ratio applications, such as wire coatings, are not usually suitable for medium reduction-ratio applications such as tubing. (The reduction ratio is the ratio of the die diameter to the product diameter.) Thread sealant tapes are generally produced with other special grades. New types of fine powder resins are constantly being developed, leading to both improved and new applications and processing techniques [30–32].

Dispersion Resins. In general, PTFE dispersions are concentrated to 50–60 % solids and stabilized by a nonionic surfactant. The dispersions are applied to various substrates by spraying, flow-coating, dipping, coagulating, or electrodepositing. Films can be made by casting the aqueous dispersion on a supporting surface, drying, baking, and cooling. If the film is too thick, cracks develop; the maximum film thickness per application is ca. 40 µm. Thicker coatings are made by casting and baking a series of layers.

An important application of PTFE dispersions is cloth impregnation. The woven cloth (e.g., fiber glass) is usually dipped into the dispersion and dried and sintered as in film formation. In some cases, the fiber is coated with the PTFE dispersion before weaving. Dispersions are also used in the production of PTFE fibers [33]. The dispersion is mixed with a matrix-forming medium such as viscose and spun into fibers through a spinneret and coagulating bath. The fibers are heated to remove the matrix polymer, sintered, and drawn for spinning.

2.2.4. Uses

About half of the PTFE resin that is produced is used for electrical purposes [34]. A major application is PTFE-insulated hookup wire used in military and aerospace electronic equipment. Polytetrafluoroethylene is also used in coaxial cables wrapped with PTFE tape, as insulation for airframe and computer wires, as well as in electrical components and "spaghetti" tubing.

Other important uses of PTFE resin are found in fluid-conveying systems, where it is used for gaskets, molded packings and seals, piston rings, bellows, overbraided hose, and lined pipe. Laboratory apparatus is also made of PTFE.

In static and dynamic load supports, large quantities of PTFE are used for bearings, ball- and roller-bearing components, and sliding bearing pads. The most commonly use known is as a release coating; PTFE-coated cookware utilizes a large amount of resin each year. Sheet and pressure-sensitive tape also serve for release applications.

Architectural fabrics are made by coating heavy-duty glass fabric with PTFE dispersions. These materials have outstanding weatherability and are resistant to chemicals, microorganisms, and fire. They have been used to enclose football stadiums and shopping centers.

Fibers and filaments are used in the production of PTFE filter cloth. Highly porous fabrics (e.g., Gore-Tex) are prepared by a process based on the fibrillation of high molecular mass PTFE. These fabrics have a high permeability for water vapor, but none for liquid water. They are widely used in outdoor wear and camping accessories.

A small but important use for PTFE is in medicine [35]; due to its inert and antiadhesive character, it is a highly suitable material for implants in the human body.

The main uses of the principal commercial PTFE resins are given in Table 5.

Table 5. Uses of PTFE resins

Type of resin	Uses
Granular	
Agglomerates	Gaskets, packing, seals, bearings, sheet, rod, heavy wall tubing, and tape
Coarse cut	Tape, molded shapes for chemical, mechanical, electrical, and nonstick applications
Fine cut	Molded sheets, tape, wire, wrapping, tubing, and gasketing
Presintered	Rods and tubes
Fine powder	
High reduction ratio *	Wire coating and thin-walled tubing
Medium reduction ratio *	Tubing, pipe, overbraided hose, and spaghetti tubing
Low reduction ratio *	Thread-sealant tape, pipe liners, tubing, and porous structures
Dispersion	
General purpose	Impregnation, coating, and packing
Coating	Film and coating

* Ratio of extrusion-die diameter to product diameter.

2.3. Tetrafluoroethylene–Hexafluoropropylene Copolymers

Copolymers of tetrafluoroethylene (TFE) and hexafluoropropene (HFP) are commonly referred to as fluorinated ethylene–propylene (FEP) resins [*25067–11-2*]. They are linear, perfluorinated copolymers:

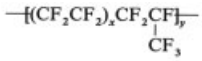

The presence of HFP in the polymer chain reduces the tendency to crystallize. Thus, a copoly-

mer with a much lower molecular mass than the corresponding homopolymer TFE retains toughness adequate for most applications. The molecular mass of commercial FEP resins is only ca. 1/100th that of commercial PTFE resins. This results in a ca. 10^6-fold reduction in melt viscosity. Resins made of FEP have melt viscosities in the range of 1 – 10 kPa · s at 380 °C, which enables standard plastics processing techniques, such as extrusion and injection molding, to be used. In spite of their much lower molecular mass, the properties of FEP resins are largely comparable to those of PTFE resins. The FEP resins have outstanding chemical resistance, nonflammability, weatherability, and electrical, antiadhesive, and low-friction properties. The maximum use temperature of 200 °C, however, is 60 °C below that of PTFE resins.

Several grades of FEP molding powder have been developed for different requirements. The lowest viscosity is shown by FEP 110, which is designed for injection molding. The melt viscosity of FEP 100 is 2 – 3 times higher than that of the 110 grade. It is intended for general purposes, but is used mainly for extrusion. Another extrusion-grade resin, FEP 140, with a melt viscosity about twice that of FEP 100, provides significantly higher stress–crack resistance. The highest viscosity grade, FEP 160, is used in making linings for chemical process equipment and other applications requiring maximum stress–crack resistance. An aqueous dispersion of FEP, containing 55 wt% solids, is stabilized with a nonionic surfactant.

Du Pont produces FEP resins under the Teflon trade name. A similar product is manufactured by Daikin Kogyo of Japan and sold under the Neoflon trade name. The FEP resins are available in the form of extruded pellets (molding powders) and as an aqueous dispersion.

2.3.1. Production

Copolymerization can take place in aqueous or nonaqueous media. Because it is much less reactive than TFE, HFP is present in larger proportions in the mixture. In aqueous media, a dispersing agent such as ammonium perfluorooctanoate is normally employed. Persulfate is usually preferred as a free-radical initiator. Dispersion polymerization conditions are very similar to those employed for PTFE fine powder resins (see also page 6) [6], [36]. Agitation must be gentle to avoid premature coagulation. The control of FEP polymerization is difficult because HFP content and melt viscosity must be maintained within narrow limits to achieve the proper balance between polymer properties and processibility.

In aqueous polymerization, unstable acid end groups are formed, which must be removed during finishing. Otherwise, they decompose during fabrication, causing bubble formation and discoloration [37]. Polymers synthesized in a nonaqueous medium, with fluorinated peroxides or azo compounds as initiators, do not contain unstable end groups [38].

2.3.2. Properties

Physical Properties. The FEP resins are random copolymers of TFE and HFP. The HFP content is usually expressed as an HFP index (HFPI), which is the ratio of IR absorbances at ca. 980 and 2365 cm^{-1}. Commercial FEP resins have HFPIs between ca. 3 and 5.5 [6]. The actual HFP content of FEP resins can be determined by a high-temperature ^{19}F-NMR technique [39]; a typical FEP resin contains 8.5 mol% of HFP.

The crystallinity of "as-polymerized" FEP is lower than that of "as-polymerized" PTFE (70 % vs. 98 %). After processing, crystallinity depends on the molecular mass and the rate of cooling from the melt. The crystallinity of FEP moldings is usually ca. 50 %.

Like PTFE resins, FEP resins are insoluble in all known solvents. This precludes the determination of molecular mass below 300 °C by the usual methods. Number-average molecular masses have been estimated by measuring the concentration of acid end groups before finishing. (Polymerization in aqueous media, gives 0.05 to 0.20 mol% acid end groups.) The number-average molecular masses of commercial grades are between ca. 125 000 and 300 000 [40]. However, molecular mass is normally specified in terms of the melt-flow index, measured at 380 °C in a melt rheometer of the type used for polyethylene equipped with a corrosion-resistant barrel, piston, and die (ASTM D-2116). Melt-flow index values range from ca. 1.5 for the

highest viscosity grade (FEP 160) to ca. 20 for the lowest viscosity grade (FEP 110).

Typical FEP resins have a crystalline melting point from ca. 260 to 280 °C, depending on their HFP content, and do not exhibit any other crystalline transitions (the 19 °C transition in PTFE is absent in FEP). An amorphous transition is observed at 90 – 150 °C, depending on frequency of measurement. A second amorphous transition occurs below −150 °C [41].

The thermal stability of FEP resins is lower than that of PTFE resins, but is adequate for processing up to at least 380 °C. Thermal decomposition occurs in two distinct stages. In the first, HFP is preferentially split out of the backbone at a rate that is about 4 times faster than that found in normal PTFE decomposition. The second stage involves the decomposition of the remaining PTFE backbone, at the same rate as that of PTFE decomposition.

Mechanical Properties. The mechanical properties of FEP are very similar to those of PTFE resins. The absence of the 19 °C transition in FEP resins eliminates the volume change observed in PTFE resins at this temperature. The coefficient of friction is slightly higher in FEP resins than in PTFE. However, the static coefficient is lower than the dynamic coefficient, which avoids stick–slip. Typical values of mechanical properties are given in Table 6.

Table 6. Mechanical properties of FEP resins ∗

Property	Value	ASTM
Relative density	2.14 – 2.17	D 792
Tensile yield, MPa	12	D 638
Tensile strength, MPa	21	D 1708
Elongation, %	300	D 1708
Flexural modulus, MPa	660	D 790
Compressive strength, MPa	15	D 695
Hardness, Shore D	59	D 2240
Izod impact strength, J/m	no break	D 256
Coefficient of thermal expansion, K^{-1}	9.3×10^{-5}	D 696

∗ FEP 100

The temperature dependence of the tensile properties of FEP and PTFE resins is similar from ca. −250 to +200 °C. However, above 200 °C, the FEP properties deteriorate rapidly, whereas the PTFE properties remain satisfactory up to at least 260 °C. Thus, the maximum use temperature of FEP is 60 °C below that of PTFE resins.

As with PTFE, fillers can be added to FEP resins to increase wear resistance and stiffness and to lower creep. The maximum loading of fillers is controlled by processing considerations. Available compositions include 30 % bronze filled, 20 % graphite filled, and 10 % fiber glass filled. These all have high *PV* (pressure–velocity) values desirable for use as bearings.

The surfaces of FEP can be made wettable by corona discharge or by treatment with sodium naphthenate or with amines in a hot oxidizing atmosphere.

Electrical Properties. The electrical properties of FEP resins (Table 7) and PTFE resins (Table 4) are virtually identical. However, the effect of voids on the electrical properties that is observed in PTFE resins is absent in the void-free FEP resins. Because of the stability of FEP in dielectric measurements, the National Bureau of Standards has selected Teflon FEP resins as reference standards. The electrical properties of FEP are unaffected by prolonged exposure to water or a temperature of 200 °C. Only corona ignition affects the dielectric life of FEP. If the voltage stress is too low to cause corona discharge, dielectric life is infinitely long at any frequency.

Table 7. Electrical properties of FEP resins

Property	Value	ASTM
Dielectric constant		
100 kHz	2.03	D 1531
1 MHz	2.02	
Dissipation factor		
100 kHz	0.0003	D 1531
1 MHz	0.0008	
Surface resistivity, Ω/m^2	$> 10^{15}$	D 257
Volume resistivity, $\Omega \cdot cm$	$> 10^{17}$	D 257
Dielectric strength, 3.2-mm sheet, kV/mm	20.0	D 149
Arc resistance, s	no carbon track	D 495

Chemical Resistance. Like PTFE resins, FEP resins are inert to practically all chemicals and solvents, with the exception of fluorine at elevated temperature and pressure and molten alkali metals or sodium hydroxide. Adsorption of chemicals is usually less than 1 % even after long

exposure at an elevated temperature. Adsorption is completely reversible and has no effect on the polymer. Permeation of gas and vapor through FEP is slower than through most other plastics.

Radiation Resistance. The effect of gamma-ray or electron-beam irradiation on FEP and PTFE is largely degradative. In the absence of air, the radiation resistance of FEP is 10 times greater than that of PTFE. In the presence of air, however, both degrade at about the same rate. In the absence of air, irradiation at low dose rates at a temperature above the alpha transition (>150 °C) results in cross-linking [42]. However, this treatment is not of practical importance.

Weatherability. Both FEP and PTFE resins exhibit the same outstanding weatherability. After 15 years' exposure in Florida, the tensile properties and light transmission of a crystal clear FEP film remained unchanged.

Flame Resistance. The flame resistance of both FEP and PTFE is outstanding; FEP does not support combustion in air and has a limiting oxygen index (LOI) of 96 %.

2.3.3. Processing

Extrusion and Injection Molding. Conventional methods such as extrusion and injection molding are employed for processing FEP. However, FEP has a higher melt viscosity and lower critical shear rate than many other thermoplastic materials, which makes processing more difficult. Molten FEP can corrode the metals used in extruders and injection molding machines. Equipment parts that come into contact with molten FEP resin should be constructed from more corrosion-resistant materials (e.g., Hastelloy C or Inconel) [43].

Critical shear rates, above which melt fracture occurs, are low and range from $20\,s^{-1}$ for FEP 100 to $2\,s^{-1}$ for FEP 160. The resulting low production rates preclude the use of pressure-extrusion techniques. Fortunately, FEP resins have a high melt strength and can tolerate high draw-down ratios without fracture. Melt-draw techniques make high production rates possible. In addition, FEP can be extruded into wire coating, tubing, rods, and film of various sizes. It can be extruded over such materials as silicone rubber, poly(vinyl chloride), glass braid, metal-shielded cables, twisted conductors, and parallel multiconductor cables. Concentrates of any desired color can be blended with standard resins.

Injection molding is best carried out with the lowest melt viscosity grade (FEP 110 resin). Recommendations for molding equipment, designs, and operating conditions are available from the manufacturers. Bottles, flasks, beakers, and other laboratory ware are typical FEP products that can be made by injection molding.

Dispersion Coating. Aqueous FEP dispersions are processed in essentially the same way as PTFE dispersions. They are usually coated on the substrate (e.g., glass cloth) and the water is evaporated. The coating is then fused by heating above the melting point. Because of the lower melt viscosity of FEP, fusion is quicker than with PTFE resins. In addition, continuous films, free of pin holes, are more readily achieved. Sometimes FEP dispersions are used to provide a final coating on a PTFE coating.

2.3.4. Uses

Electrical applications comprise most of the uses of FEP resins, including hookup wire, interconnecting wire, coaxial cable, computer wire, thermocouple wire, and molded parts. Other applications include lined pipes and fittings, lined tanks, overbraided hose, heat exchangers, and laboratory ware. Mechanical uses include antistick applications such as conveyor belts and roll covers. Film is used in solar-collector windows because of its light weight, excellent weatherability, and high solar transmission.

2.4. Tetrafluoroethylene–Perfluoro-(Propyl Vinyl Ether) Copolymers

Copolymers of TFE with perfluoro(propyl vinyl ether) (PPVE) are called perfluoroalkoxy (PFA) resins [*26655-00-5*]. They have the following general formula:

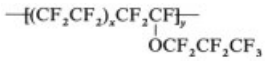

These resins were developed after FEP resins and were designed to overcome the disadvantages of FEP resins in comparison with PTFE resins, i.e., lower thermal stability and poorer mechanical properties at high temperature. The ether group in the side chain of PFA resins eliminates steric strain at the branching point, which is responsible for the reduced thermal stability of FEP resins [44]. In addition, the large perfluoropropoxy side group in PFA reduces the crystallinity much more than the trifluoromethyl side group in FEP; it is ca. 6 – 8 times more effective on a molar basis. Thus, the comonomer content of PFA resins is substantially less than that of FEP resins even though the former are tougher. The melting point of PFA (305 – 310 °C) is only ca. 20 °C below that of PTFE. The thermal stability and maximum use temperature of PFA and PTFE resins are the same. The melt-flow rates and critical shear rates of PFA resins are higher than those of their FEP counterparts, which makes PFA resins easier to process. In addition, PFA resins have all the other advantages of PTFE resins: antistick performance, resistance to virtually all chemicals, low coefficient of friction, nonflammability, and excellent electrical properties.

Du Pont manufactures PFA resins under the Teflon PFA tradename. Similar resins are produced by Hoechst (Hostaflon TFA) and Daikin (Neoflon AP).

2.4.1. Production

Copolymerization. The copolymerization of TFE and PPVE can be carried out in aqueous [45] or nonaqueous media [46]. The aqueous process resembles the dispersion polymerization used for PTFE fine powders and FEP resins (see page 6). However, in contrast to the FEP process, in which HFP is the principal reaction ingredient, only a small amount of PPVE is present in the PFA reaction mixture. This is because (1) PPVE is ca. 10 times more reactive than HFP and (2) a much smaller amount of comonomer is incorporated into PFA compared to FEP. This can lead to inhomogeneity in the copolymers produced in the aqueous process. However, the addition of a small amount of fluorocarbon diluent to the aqueous polymerization mixture gives a uniform distribution of PPVE in the product. Another complication of this copolymerization process is the tendency of the growing PPVE radical to terminate by a rearrangement chain transfer:

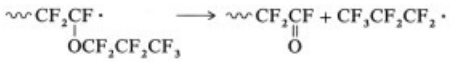

This reaction competes with the usual chain termination by coupling, and can result in a polymer with a very broad molecular mass distribution. The use of chain-transfer agents such as methanol can keep the molecular mass and molecular mass distribution within desired limits [47].

Nonaqueous polymerization uses fluorinated peroxide initiators to produce stable end groups. Again, chain-transfer agents are required to keep the molecular mass and distribution within the desired limits. However, because of the PPVE radical rearrangement, some acid fluoride end groups are always present in the nonaqueous product. These groups can be stabilized by reaction with methanol or ammonia [48] or preferably with elemental fluorine [49].

Polymer finishing is simpler for PFA than for FEP resins. The aqueous dispersion product is coagulated, washed, dried, and heated to remove volatile components and unstable end groups. The resulting powder can be used directly for coating and the like or extruded into molding powders for standard extrusion and molding applications. The dispersion is also concentrated and stabilized for spray or dip-coating applications. Nonaqueous products are dried to remove solvent and extruded directly into molding powders. Several grades of PFA resins are available for extrusion and molding applications.

2.4.2. Properties

Physical Properties. Perfluoroalkoxy (PFA) resins undergo a single first-order transition at their melting point (305 – 310 °C). They do not display the other first-order transitions present in PTFE resins. Torsion-pendulum measurements revealed three second-order transi-

tions at −100, −30, and 90 °C. The crystallinity of "as-polymerized" dispersion polymer was 65 – 75 %. Samples slowly cooled from the melt had 58 % crystallinity.

The thermal stability of PFA resins is comparable to that of PTFE resins. Although they can be processed at temperatures up to 425 °C [50], degradation occurs on prolonged heating at or above 425 °C. Below its melting point, PFA does not decompose. Samples of PFA aged in an air oven at 285 °C for 2 years showed no deterioration; in fact, tensile strength, ultimate elongation, and flex life increased.

Mechanical Properties. The mechanical properties of PFA resins resemble those of PTFE and FEP resins at room temperature, but above 200 °C, PFA resins are superior to FEP resins. The high-temperature mechanical properties of PFA resins are equal or superior to those of PTFE resins. Both PFA and PTFE resins have a maximum use temperature of 260 °C. The low-temperature mechanical properties of the two resins are equivalent. The mechanical properties of PFA are shown in Table 8.

The wear and friction properties of PFA and FEP resins are similar. The static coefficient of friction is lower than the dynamic coefficient; consequently the polymer is free of stick–slip behavior. Fillers such as fiber glass or bronze powder improve the wear resistance of PFA, FEP, and PTFE resins.

Electrical Properties. The electrical properties of PFA, PTFE, and FEP resins are essentially the same [51]. The dielectric constant of PFA (2.06) is slightly higher than that of FEP (2.03).

Flame Resistance. Perfluoroalkoxy resins have the same nonflammability characteristics as PTFE resins. The limiting oxygen index (LOI) is over 95 %. According to Underwriters Laboratories (UL 94), PFA resins are self-extinguishing and classified 94 VE–O.

Radiation Resistance. High-energy irradiation affects PFA, FEP, and PTFE resins similarly. The PFA resins have about the same rate of degradation as PTFE resins in air, but resist radiation better under vacuum.

Chemical Resistance. Perfluoroalkoxy resins have outstanding resistance to chemicals and solvents, and are unaffected by strong oxidizing acids, bases, halogens, organic acids, aromatic and aliphatic hydrocarbons, alcohols, aldehydes, ketones, ethers, amines, esters, chlorinated compounds, and polymer solvents. Perfluoroalkoxy resins can be attacked by alkali metals and elemental fluorine at elevated temperature and pressure.

Optical Properties. Perfluoroalkoxy films have good optical properties; they are colorless and transparent in thin sections, and are translucent in thicker sections. Energy transmission is best in the IR region (87 – 93 %) followed by the visible (80 – 87 %) and UV regions (55 – 80 %).

2.4.3. Processing

Perfluoroalkoxy resins may be processed by conventional techniques, such as extrusion [52], injection molding [53], transfer molding, and rotocasting. Dispersions are available for spray and dip coating of fabrics and other substrates. High-temperature PFA melts can be corrosive to conventional stainless steels; more corrosion-resistant metals such as Hastelloy-C or Xaloy 306 are recommended for equipment parts that come into contact with the melts. Processing temperatures of ca. 330 – 425 °C are recommended. Residence times at the upper temperature limit should be kept short (< 10 min) to avoid polymer degradation.

Perfluoroalkoxy resins have higher melt-flow indexes than their FEP counterparts and consequently have higher critical shear rates before the onset of melt fracture (Table 9). Because of their higher inherent thermal stability, PFA resins can be processed at a higher temperature than their FEP counterparts. This results in a further increase in melt-flow index and critical shear rate for PFA resins. Thus, PFA resins can also be processed at significantly higher rates than their FEP counterparts.

Intricate articles that are difficult, if not impossible, to mold from FEP 110 resin, can be readily injection-molded from PFA 340 resin. The superior molding behavior of PFA is due to its favorable high-temperature mechanical properties, allowing the use of higher mold temper-

Table 8. Mechanical properties of PFA resins

Property	PFA 340	PFA 350	ASTM
Tensile yield, MPa			
23 °C	14	15	D 1708
250 °C	3.5	4.1	
Tensile strength, MPa			
23 °C	28	31	D 1708
250 °C	12	14	
Elongation, %			
23 °C	300	300	D 1708
250 °C	480	500	
Flexural modulus, MPa			
23 °C	655	690	D 790
250 °C	55	69	
Hardness, Shore D	60	60	D 2240
MIT∗ flex life, cycles	50 000	500 000	D 2176
Coefficient of thermal expansion, K^{-1}	6.7×10^{-5}	6.7×10^{-5}	D 696

∗ Massachusetts Institute of Technology instrument.

Table 9. Melt-flow index and shear rate of FEP and PFA resins

Grade		Melt-flow index,	Critical shear rate,
FEP	PFA	g/10 min ∗	s^{-1}
100		6	20
	340	14	50
140		3	8
	310	6	16
160		1	2
	350	2	6

∗ The melt-flow index is the amount of resin extruded at 372 °C in 10 min (ASTM D2116).

atures (not far below the melting point of the resin). With FEP resins, mold temperatures must be below 200 °C because poor mechanical properties above 200 °C make ejection of the article from the mold without damage or distortion impossible.

Other molding procedures often used with PFA resins are transfer molding and rotocasting. Transfer molding is used for valve liners and similar products. Rotocasting is used for large structures such as drums and tanks. A special PFA powder, developed for rotocasting (TE-9749), gives a uniform, bubble-free casting. In this procedure, the heating cycle may range from 90 to 180 min at 355 – 365 °C. The thermal stability of PFA is more than adequate for this procedure.

Powders can be used to produce thin layers of PFA on various substrates. The surfaces to be coated are heated above the melting point of the PFA and the powder is sprayed on. The powder melts on the surface, which is then cooled.

A dispersion form (PFA 335) is available for coating metals, glass fabrics, and similar substrates. The dispersion can be processed by the techniques used with PTFE dispersions (see page 10).

2.4.4. Uses

Perfluoroalkoxy resins are used for producing high-quality electrical insulation and molded components. Other important uses are for bellows and expansion joints and liners for valves, pipes, pumps, and fittings. Tubing and film are also produced. Film can be oriented to give products with specialized properties. Equipment made from PFA is used extensively in the semiconductor industry for demanding chemical handling applications.

2.5. Tetrafluoroethylene–Ethylene Copolymers

Tetrafluoroethylene and ethylene have a strong tendency to alternate during polymerization. Thus, TFE–ethylene copolymers [*25038-71-5*] (ETFE resins) are composed mainly of alternating sequences of the two monomers and have the following structure:

$$-(CF_2CF_2CH_2CH_2\text{-})_n-$$

The alternating comonomer units are responsible for the unique properties of ETFE resins; they have a high tensile strength, moderate stiffness, high flex life, outstanding impact strength, good abrasion resistance, and very high "cut-through" resistance. Their electrical properties are similar to those of other TFE polymers. The ETFE resins have low dielectric constant, high dielectric strength, excellent resistivity, and a low dissipation factor; they also display good thermal stability and excellent chemical resistance. Unlike other TFE copolymers, ETFE resins have a high resistance to high-energy radiation. They are cross-linked by low doses of electron beam or gamma ray radiation. The cross-linked products have improved high-temperature properties, including resistance to cut-through by a hot soldering iron. The processibility of ETFE resins is superior to that of other TFE copolymers (the critical shear rate of ETFE resin is ca. 250 times higher than that of FEP resin). They are particularly well suited to extrusion and injection molding.

Unmodified TFE–ethylene copolymers have poor thermal stress–crack resistance, which severely limits their utility. This problem has been largely overcome by incorporating 0.1–10 mol% of termonomers, such as perfluoro(alkyl vinyl ethers) and perfluoroalkylethylenes, in the polymer backbone [54–56]. Commercial ETFE resins are all modified in this way. The thermal stress–crack resistance of modified ETFE resins is a function of the amount of termonomer and the melt viscosity of the resin. Processibility (extrusion rate), on the other hand, is mainly a function of the melt viscosity. Lowering the melt viscosity improves processibility, but reduces stress–crack resistance. Raising the melt viscosity improves stress–crack resistance, but reduces processibility. Several grades of commercially available modified ETFE resins provide different compromises between processibility and thermal stress–crack resistance to meet a variety of use and processing requirements.

Du Pont produces a series of modified ETFE resins under the Tefzel trade name. The 210 grade has the lowest viscosity and is designed for fast extrusion coating of small diameter wires that do not require maximum thermal stress–crack resistance. The 200 grade is a general-purpose resin designed for both extrusion and molding applications. The 280 grade has the highest viscosity and the best stress–crack resistance. It is designed for applications requiring maximum thermal stress–crack resistance. Other modified ETFE resins are produced by Asahi Glass (Aflon COP), Ausimont (Halon ET), Daikin (Neoflon EP), and Hoechst (Hostaflon ET).

2.5.1. Production

Tetrafluoroethylene and ethylene monomer mixtures can undergo thermal decomposition reactions that are even more energetic than those of pure TFE. Therefore, great care must be exercised during the copolymerization of these monomers. All equipment containing the monomers in a pure state under pressure must be barricaded to prevent injury or death in the event of violent decomposition.

Copolymers of TFE and ethylene can be prepared in aqueous, nonaqueous, or mixed systems. In the last case [55], the comonomers and initiator are dissolved in the nonaqueous phase. The water acts as a heat-transfer medium and dispersant for the viscous nonaqueous phase. Aqueous polymerizations have been carried out in the presence of fluorinated surfactants such as ammonium perfluorooctanoate with manganic acids as initiators [57]. These mixtures usually also contain a chain-transfer agent and stabilizer such as ammonium oxalate. Nonaqueous polymerizations employ a fluorinated solvent (e.g., 1,1,2-trichloro-1,2,2-trifluoroethane) with a fluorinated acyl peroxide as initiator. Chain-transfer agents are added to control molecular mass. The polymerization temperature is usually between 30 and 85 °C, with pressures below 1.7 MPa.

The following four propagation reactions take place in TFE–ethylene copolymerizations:

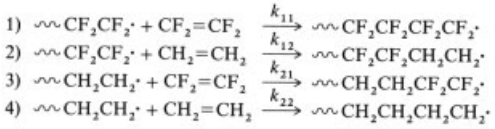

k = rate constant; $r_1 = k_{11}/k_{12}$; $r_2 = k_{22}/k_{21}$

Reactions 2 and 3 give alternating units, while reactions 1 and 4 give nonalternating units. The ratio of the rate constants of reaction 1 to reaction 2 is designated r_1. The corresponding ratio of reaction 4 to reaction 3 is designated r_2. Low values for r_1 and r_2 signify a highly alternating copolymerization. Values for r_1 and r_2, measured at 60 °C, are 0.033 and 0.258, respectively. By using these values, it can be calculated that a 1 : 1 copolymer of TFE and ethylene, prepared at 60 °C, contains ca. 92 % alternating units. This calculated value agrees well with experimentally determined values (e.g., by high-temperature NMR studies). More highly alternating copolymers are obtained at lower temperatures. A 1 : 1 copolymer prepared at 0 °C is ca. 96 % alternating. Theory predicts that an essentially 100 % alternating copolymer can be prepared at ca. −80 °C.

The high-temperature stress–crack resistance of ETFE resins can be improved by incorporating suitable termonomers such as vinyl monomers without telogenic activity that provide at least a two-carbon branch [54–56].

The copolymers are isolated from the aqueous polymerization medium by coagulating, filtering, washing, and drying. The product is recovered from nonaqueous polymerization by evaporating the fluorinated solvent. Finishing usually includes a melt-compaction step in which the dried powders are converted into extruded pellets (molding powders).

2.5.2. Properties

Physical Properties. Alternating copolymers of TFE and ethylene have a unique crystalline lattice [58]. The unit cell, which is orthorhombic or monoclinic, has the following parameters: $a = 0.96$ nm, $b = 0.925$ nm, and $c = 0.5$ nm ($\gamma = 96°$). The calculated crystalline density of the unit cell is 1.9 g/cm^3. The molecular conformation is an extended zigzag, with the molecular packing in the orthorhombic configuration. Each molecule has four nearest neighbors with the CH_2 groups of one chain adjacent to the CF_2 groups of another. Determination of percent crystallinity of ETFE resins by X-ray techniques is difficult because of numerous imperfections in the lattice. Typical copolymers have a degree of alternation of ca. 0.88, which means a departure from alternation once every 16 carbon atoms. Crystallinity probably ranges from 40 to 60 %, depending on quench rate and conditions after molding.

The melting points of ETFE resins can vary from 200 to 300 °C [8], depending on the TFE : ethylene ratio, degree of alternation, and termonomer content. A typical commercial ETFE resin, Tefzel 200, melts at ca. 270 °C. The effect of degree of alternation on the melting point of a 1 : 1 copolymer is as follows:

Degree of alternation	*mp*, °C
0.92	280
0.96	296
(1.0)	(318)

As indicated in Section 2.5.1, the degree of alternation varies with the polymerization temperature.

In addition to the crystalline melting point, ETFE resins exhibit three second-order transitions [59]. At 1 Hz, the α-, β-, and γ-relaxations were observed at 110, −25, and −120 °C, respectively, in a quenched sample. The α- and γ-relaxations are attributed to motions of long and short segments, respectively, in the amorphous regions. The β-transition appears to occur in the crystalline regions. These transition temperatures depend on frequency. The temperature of the α-transition is 145 °C when measured at 100 Hz, and 170 °C when measured at 10^5 Hz. The γ-transition occurs at −35 °C when measured at 10^5 Hz.

Under ambient conditions, ETFE resins are insoluble in every known solvent. This precludes the measurement of molecular mass by the usual methods. However, diisobutyl adipate dissolves ETFE resins above 230 °C. A high-temperature laser light-scattering apparatus can be used to measure the weight-average molecular mass $\bar{M}_w$ of ETFE resin in diisobutyl adipate solution at 240 °C [60]. Molecular masses, $\bar{M}_w$, are obtained ranging from 5.62×10^5 to 1.13×10^6, with a polydispersity index ($\bar{M}_w/\bar{M}_n$) of ca. 1.3. The correlation between melt viscosity (in Pa · s) and $\bar{M}_w$ was shown to be

$$\eta = 1.88 \times 10^{-20} \cdot \bar{M}_{w\ 3.42}$$

The molecular mass of ETFE resins is normally specified in terms of melt-flow index (MI), measured at 300 °C in a melt rheometer; values are between 3 and 30 g/10 min.

Mechanical Properties. Modified ETFE resins are tougher and stiffer and have higher tensile strength than other TFE polymers such as PTFE, FEP, or PFA. Typical properties are given in Table 10 [61]. As wire insulation, ETFE resins exhibit excellent cut-through and abrasion resistance.

Addition of fillers enhances creep and wear resistance. Unlike other TFE polymers, ETFE can be reinforced by glass fibers. The properties of a 25 % glass-reinforced Tefzel composition, HT-2004, are shown in Table 10; it has about twice the tensile strength and heat distortion temperature of the unreinforced resin, and its flexural modulus is almost 5 times higher. Its ultimate elongation is low, which is characteristic of reinforced compositions. However, its high toughness is indicated by the high Izod impact values at room temperature, as well as at −54 °C.

The operating temperature range of ETFE resins is broad, from ca. −100 °C to at least +150 °C. This upper limit is a conservative estimate based on extensive high-temperature testing. It is the temperature at which at least half of the original properties are retained after 20 000 h exposure. The compositions with higher stress–crack resistance such as Tefzel 280 tolerate temperatures up to ca. 240 °C without cracking. Cross-linking by high-energy radiation further enhances its high-temperature mechanical properties [62]. Some radiation-cross-linked ETFE wire constructions have been rated for continuous service at 200 °C.

Electrical Properties. The ETFE resins serve as excellent insulating materials. Their dielectric constant is low (2.6) and is essentially independent of the frequency. The dielectric strength ranges from 16 – 20 kV/mm, when measured on 3-mm thick specimens, to 160 – 200 kV/mm on films 25 – 75 µm thick.

Radiation Resistance. Resistance to gamma and electron beam radiation is excellent. Properties are still useful after exposure to more than 10^6 Gy (100 Mrad). Radiation is more harmful in air than in nitrogen.

Chemical Resistance. Resistance to chemicals and solvents is also excellent. Strong acids and bases, aromatic and aliphatic hydrocarbons, alcohols, ketones, and esters have no effect on ETFE resins. Strong oxidizing acids, organic bases, and sulfonic acids attack ETFE resins to varying degrees at elevated temperatures. Certain high molecular mass esters such as diisobutyl adipate are solvents for ETFE resins above 200 °C.

Weatherability. Like other TFE polymers, ETFE resins have outstanding weatherability. Extended exposure in Florida and Arizona had no effect.

Flame Resistance. The ETFE resins are nonflammable in air. Wires insulated with ETFE resins are rated 94V–O by the Underwriters Laboratories, with a LOI of 31 %.

2.5.3. Processing

Common techniques such as injection molding, compression molding, blow molding, transfer molding, rotational molding, extrusion, and coating techniques are used for processing ETFE resins. Processing temperatures are usually 300 – 340 °C. Molten ETFE can corrode some steels, and corrosion-resistant steel such as Hastelloy C is recommended for equipment parts that come in prolonged contact with molten ETFE resins.

High extrusion rates without melt fracture can be achieved [63]. The critical shear rates for ETFE resins are at least 250 times higher than those of other TFE resins. Typical values are as follows:

Product grade	Melt-flow index, g/10 min	Critical shear rate, s^{-1}
Tefzel 210	45	12 000
Tefzel 200	8	4 800
Tefzel 280	3	4 500

Table 10. Mechanical properties of ETFE resins

Property	Tefzel 200	Tefzel HT-2004 *	ASTM
Relative density	1.70	1.86	D 792
Tensile strength, MPa	44.8	82.7	D 638
Elongation, %	200	8	D 887
Flexural modulus, MPa	1380	6550	D 790
Izod impact strength, J/m	no break	9	D 621
Low temp. embrittlement, °C	< -100		D 746
Hardness, Shore D	75		
Coefficient of thermal expansion, K^{-1}	9×10^{-5}	3×10^{-5}	D 696
Coefficient of friction			
Static, 3.4 MPa load		0.3	
Dynamic, $u = 900$ m/s	0.4	0.3	

* 25 % Glass fiber reinforced composition.

2.5.4. Uses

The unique properties of ETFE resins make them useful in many applications. The general-purpose 200 grade is used for jacketing signal, control, communications, and power wiring for mass transport systems. It is also used for control and instrumentation wire for utilities, for critical wiring in chemical plants, and for injection-molded components such as sockets and connectors. Other products include seal glands, pipe plugs, fasteners, pump vanes, pump impellers, laboratory ware, and chemical packings. The high melt-flow 210 grade is used for extrusion coating of fine wires and injection molding of intricate shapes. The high stress–crack resistance 280 grade is used for insulating heater cables and automotive wiring and for oil-well logging cables. This resin is also used for transfer molding and injection molding of articles containing metal inserts or thick sections, and for stock shapes such as tubes and rods.

2.6. Polychlorotrifluoroethylene

Polychlorotrifluoroethylene [*9002-83-9*] (PCTFE) is a linear, high molecular mass polymer with the general formula:

It is semicrystalline (*mp* ca. 210 °C) and can be processed by conventional methods. It has good mechanical properties, chemical inertness, radiation resistance, and low vapor permeability. The quick-quenched form is optically clear; the slow-cooled form is translucent. It is supplied by Daikin Kogyo (Daiflon), 3M (Kel-F), Allied Signal (Aclar), and Ausimont (Aclon).

2.6.1. Production

Polychlorotrifluoroethylene can be polymerized to high molecular mass in bulk [64], solution [65], suspension [66] , or emulsion systems [67] using free-radical initiators. The emulsion process is carried out near room temperature over a persulfate–bisulfite redox catalyst. The polymer emulsion is coagulated by freezing or by the addition of salts. The isolated polymer is washed to remove initiator residues and dried.

The dried powders can be stabilized by the addition of phenols or salts of organic acids to help prevent discoloration and bubbling during processing. The tendency for PCTFE to become brittle during use can be reduced by incorporating a small amount (< 5 %) of vinylidene fluoride during polymerization.

2.6.2. Properties

The crystallinity of PCTFE ranges from 45 % (quick-quenched) to 65 % (slow-cooled). The molecular mass of PCTFE resin ranges from 1×10^5 to 5×10^5. The useful temperature range is between -200 and $+180$ °C. Other properties are as follows [68]:

mp	211 – 216 °C
Second-order transition (amorphous)	71 – 99 °C

Coefficient of thermal expansion	$(4-7) \times 10^{-5}\ K^{-1}$
Specific heat	$0.92\ kJ\ kg^{-1}\ K^{-1}$
Relative density	2.10 – 2.18
calcd. amorphous	2.075
crystalline	2.185
Tensile strength	
at 23 °C	30 – 40 MPa
at 100 °C	16 – 20 MPa
Elongation	
at 23 °C	30 – 80 %
at 100 °C	600 %
Flexural modulus	1000 – 2000 MPa
Hardness, Shore D	90
Dielectric constant	
at 100 Hz to 1 MHz	2.8 – 2.5
Dissipation factor	
at 1 kHz	0.023
Dielectric strength, 0.1 mm film	120 – 150 kV/mm

Resistance to attack by most chemicals is good. However, chlorinated compounds, ketones, and aromatic compounds such as benzene, toluene, and xylene swell PCTFE resins. They are impermeable to water vapor as well as other gases. They are nonflammable (LOI= 95 %), and exhibit outstanding stability to UV and gamma radiation. The coefficient of friction (like that of FEP resin) is low; antistick properties (like those of polyethylene) are excellent [69].

2.6.3. Processing

Polychlorotrifluoroethylene resins can be processed by the usual methods for thermoplastic materials such as extrusion, injection molding, and transfer molding. As with other fluoropolymers, corrosion-resistant materials should be used for equipment parts in contact with the melt. Processing temperatures are in the range of 280 – 350 °C. However, degradation begins at ca. 280 °C and so always accompanies the processing of PCTFE resins. Partial degradation does not seem to affect properties greatly, but severely degraded PCTFE becomes highly crystalline and its mechanical properties suffer. In injection molding, the mold should be heated to 70 – 130 °C. Corrosion-resistant coatings of PCTFE resins can be applied from aqueous dispersions (30 – 60 % solids) by dipping or spraying followed by drying and sintering. Coating thickness is in the range of 0.1 – 0.2 µm. Finely divided PCTFE powders (particle size < 150 µm) are used for powder coating.

2.6.4. Uses

Because of its high compressive strength, PCTFE is well suited to the construction of chemical apparatus. It is used in laboratory ware, pump parts, transparent sight glasses, flow meters, tubes, and linings. Films, wire and cable insulation, and components made from PCTFE are used in the electrical industry. Thin films are used as waterproof packaging for pharmaceuticals and expensive electronic parts.

2.7. Chlorotrifluoroethylene–Ethylene Copolymers

Copolymers of CTFE and ethylene (ECTFE resins) [*25101-45-5*] are linear, semicrystalline polymers composed of alternating monomer units:

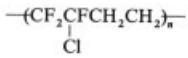

These resins have high tensile strength, moderate stiffness, and good flex life and impact strength. Electrical properties, chemical resistance, weatherability, radiation resistance, and nonflammability are all good. The usual processing methods, such as extrusion, injection molding, and transfer molding, are employed. The resins are readily converted into fibers, filaments, films, sheets, and wire and cable insulation. Powder forms are available for the preparation of void-free, low-permeability coatings on metals by electrostatic and fluidized-bed processes. The resins are manufactured by Ausimont (Halar).

2.7.1. Production

Ethylene and CTFE can be copolymerized by the techniques used for other fluoroolefins. However, to prepare commercially useful products, the copolymerization must be carried out at low temperature (preferably 10 °C or below). A typical polymerization employs a mixture of CTFE and ethylene in water at 10 °C, with trichloroacetyl peroxide as catalyst and chloroform as chain-transfer agent [70]. Under these

conditions, the copolymerization takes place in the condensed CTFE phase. Water acts as a dispersant and heat-transfer medium. Copolymerization has been carried out at even lower temperatures (between −40 and −80 °C), using oxygen-activated trialkylborons as initiators [71]. Terpolymers with improved thermal stress–crack resistance are prepared similarly. Termonomers for thermal stress–crack-resistant compositions include perfluoro(alkyl vinyl ethers) [54], perfluoroalkylethylenes [72], and 2-trifluoromethyl-3,3,3-trifluoroprop-1-ene [56]. The same techniques are used for isolation from the polymerization media as with other fluoropolymers.

2.7.2. Properties

Physical Properties. The crystallinity of commercial ECTFE resins is in the 50 – 60 % range. X-ray diffraction studies indicate that the unit cell of the alternating copolymer is hexagonal, with a chain repeat distance of $c = 0.50$ nm [70]. The molecular conformation is an extended zigzag, with an ethylene unit on one chain adjacent to a CTFE unit on another chain.

The melting point of an ECTFE copolymer is a function of the molar ratio and the degree of alternation of the two monomers. Commercial ECTFE resins are 1 : 1 copolymers of CTFE and ethylene, 80 – 90 % alternating. They have melting points in the range of 235 – 245 °C. A 1 : 1 copolymer prepared at −80 °C, essentially 100 % alternating, melts at 264 °C [71]; second-order transitions are observed at 140, 90, and −65 °C. The α- and γ-transitions are associated with the motion of chain segments in the crystalline regions, whereas the β-transition seems to be associated with the amorphous region.

The number-average molecular mass ($\bar{M}_n$) of ECTFE resins, determined by osmometry at 150 °C, ranges from 1×10^5 to 5×10^5; other properties are as follows:

mp	236 – 246 °C
Coefficent of thermal expansion	5×10^{-5} K^{-1}
Relative density	1.7
Tensile strength	50 – 65 MPa
Elongation	150 – 250 %
Flexural modulus	1800 – 2500 MPa
Hardness, Shore D	78
Izod impact strength	
at 23 °C	no break
at −40 °C	2 – 3 J/m
Dielectric constant (ASTM D 150, 60 Hz to 1 kHz)	2.55 – 2.60
Dissipation factor (ASTM D 150)	
at 60 Hz	< 0.005
at 1 MHz	0.0024
Dielectric strength 0.2 mm film	80 – 90 kV/mm
Volume resistivity	$> 10^{15}$ Ω·cm

Mechanical Properties. The ECTFE resins are tough, moderately stiff, and creep resistant. They are usable over a broad temperature range (−100 to +150 °C). Abrasion resistance and the heat distortion temperature are good and can be enhanced with fillers.

Chemical Resistance. Both ECTFE and PCTFE resins are similar in chemical resistance. They exhibit good stability toward acids, bases, and halogens. There are no known solvents below 120 °C, but amines, esters, and ketones can attack ECTFE resins, especially when warm. These resins have excellent barrier properties to water vapor and other gases. In this respect they are superior to the perfluorinated plastics.

Weatherability and Radiation Resistance. Like most fluoropolymers, ECTFE resins exhibit outstanding weathering resistance. They also resist high-energy gamma and beta emission. The polymer retains useful properties after absorbing more than 10^6 Gy (100 Mrads) of radiation. Cross-linking by exposure to low doses 50 – 150 kGy (5 – 15 Mrads) enhances the high-temperature mechanical properties [62].

Flame Resistance. The ECTFE resins are nonflammable; they burn when exposed to a flame, but are self-extinguishing when the flame is removed. They are nondripping and form a char; the limiting oxygen index (LOI) is 64 %.

2.7.3. Processing

The standard processing methods for thermoplastics, such as extrusion, injection molding, and transfer molding, are employed to process ECTFE. Extrusion is used to manufacture tubing, hose, and wire and cable insulation. Temperatures from 260 to 300 °C can be used, but

should be kept as low as possible to prevent degradation. Extrusion of ECTFE resins is possible at high rates without degradation. Like other fluoropolymers, molten ECTFE resins can be corrosive. A corrosion-resistant steel such as Hastelloy-C is recommended for metal surfaces in contact with the molten polymer. In injection molding of ECTFE resins, the melt temperature should be from 260 to 300 °C; for best results, the mold should be heated to 60 – 100 °C. Metals are coated by powders of average particle size below 150 µm using an electrostatic or fluidized-bed process. Coatings of 1 mm or thicker can be obtained free of voids and cracks.

2.7.4. Uses

The principal use of ECTFE resin is in flame-resistant wire and cable insulation [73]. Monofilament made from ECTFE resin is used in the manufacture of chemical-resistant filters and screens. Other applications include coatings and linings for chemical process equipment. Roto-molded tanks and containers made from ECTFE resins are used to store corrosive chemicals, e.g., nitric and hydrochloric acids. Extruded sheet is used for thermoforming various products, such as battery cases for heart pacemakers. Tubing made from ECTFE resin is used in laboratory and some automotive applications.

2.8. Poly(Vinylidene Fluoride)

Poly(vinylidene fluoride) [*9002-85-1*] (PVDF) is a linear, high molecular mass, semicrystalline polymer that can exist in three distinct crystalline forms and has the following structure:

$$-(CH_2CF_2\text{-})_n-$$

The alternating CH_2 and CF_2 groups create a dipole that confers solubility in highly polar solvents such as dimethylformamide and dimethylacetamide. In addition, PVDF has a high dielectric constant and exhibits piezoelectric properties under certain conditions. It has excellent resistance to abrasion, stress fatigue, and cold flow. Like PTFE and TFE copolymers, it has outstanding weatherability. It also exhibits excellent resistance to gamma and beta radiation. The usable temperature range is from ca. −40 to +150 °C. The resins are processible by extrusion, injection molding, and other molding techniques. Copolymerization with HFP (< 15 %) improves processibility, flexibility, and impact strength.

Uses for PVDF resins include durable finishes for exterior metal siding, electrical insulation, and chemical process equipment. The manufacturers are Pennwalt Corp. (Kynar), Solvay & Cie (Solef), Dynamit Nobel (Dyflor), Kureha Chemical Co. (Kureha KF), and Daikin Kogyo (Neoflon VDF).

2.8.1. Production

Vinylidene fluoride (VDF) can be polymerized by a variety of free-radical initiators, but the preparation of a commercially useful product requires the correct choice of initiator, dispersing agent, and reaction medium. In the first commercially successful process, developed by Pennwalt in the mid-1960s, VDF is polymerized in water at 120 – 130 °C and ca. 6.2 MPa with di-*t*-butyl peroxide as initiator [74]. Other successful processes use peroxy dicarbonates and peroxy esters as initiators [75]. Vinylidene fluoride can also be polymerized in aqueous emulsion at 80 °C and 5.2 MPa, using disuccinic acid peroxide as initiator and a perfluorocarbon acid as emulsifier. Polymerization in nonaqueous media using fluorocarbon acyl peroxides as initiators has also been reported [76].

Vinylidene fluoride–hexafluoropropene copolymers are prepared by an emulsion process, using a fluoroalkyl surfactant and CCl_3F as chain-transfer agent [77].

The products are isolated from the polymerization mixtures by the procedures described previously for other fluoropolymers.

Barium and strontium salts have been added to PVDF to improve its thermal stability.

2.8.2. Properties

Physical Properties. Poly(vinylidene fluoride) has a highly regular structure, with most VDF units joined head-to-tail and with ca. 5 % of the monomer units joined head-to-head. Crystallinity is between 40 and 60 %; three different crystalline forms exist.

In phase I (β-form), the polymer crystals are orthorhombic and the molecules adopt a planar zigzag conformation. In phase II (α-form) the crystals are monoclinic and the molecules adopt a gauche-trans-gauche-trans conformation. In phase III (γ-form), the crystals are also believed to be monoclinic, but the molecules adopt a planar zigzag conformation. Slow or rapid cooling of polymer from the melt yields only phase II. Conditions for converting one form to another have been established [78].

Because of its polymorphism, the melting point of PVDF is not precisely defined (154 – 184 °C). The melting point of the more common phase II form is ca. 170 °C; a glass transition temperature is observed at −40 °C. Properties of PVDF resin are shown in Table 11.

Mechanical Properties. Yield strength of PVDF is higher and elongation is lower than those of ETFE and ECTFE at room temperature. Modulus is similar but impact strength is lower. Modification with HFP (Solef 11010 or Kynar Flex 2800) lowers the modulus, but increases elongation and impact strength.

Electrical Properties. A high dielectric constant and dissipation factor impair the utility of PVDF as a primary wire and cable insulation. However, these "poor" electrical properties allow the production of PVDF films with their piezoelectric and pyroelectric properties [79]. These films are prepared from extruded films in the phase I (β-form) conformation, obtained by stretching the film as it is extruded; ultrapure polymer is required. Both surfaces of the film are metallized, and it is then subjected to a high voltage, which leaves it permanently polarized. Such films generate a voltage when stretched or compressed (piezoelectricity) or heated (pyroelectricity).

Chemical Resistance. Resistance of PVDF to halogens, most inorganic acids and bases, alcohols, halogenated hydrocarbons, and aromatics is good. It is, however, soluble in polar solvents such as acetone and ethyl, butyl, and amyl acetate.

Weather and Radiation Resistance. Resistance to UV light is excellent (no effect after 5000 h exposure in Weather-O-Meter). Tolerance to ionizing radiation is very high. Low radiation doses of 20 – 150 kGy (2 – 15 Mrads) cause cross-linking, which enhances the high-temperature modulus of PVDF.

Flame Resistance. When exposed to flame, PVDF is nonflammable and nondripping; it is self-extinguishing after the flame is removed. The limiting oxygen index (LOI) is 44 %.

2.8.3. Processing

Poly(vinylidene fluoride) can be processed by the usual techniques for thermoplastics such as extrusion and transfer, injection and compression molding. Metal coatings are applied from organisols (mixtures of resin and solvent) in a process called coil coating. Typically, the organisol is applied to coil stock of steel or aluminum. The film is formed on the metal surface as the solvent evaporates. In extrusion, attention should be paid to the elimination of dead spots where degradation of the polymer can start. The temperature should be between 215 and 270 °C. Similarly, in injection molding, the melt temperature should be between 200 and 270 °C. Runnerless molds are recommended.

2.8.4. Uses

An important use for PVDF resin is as a coating for metallic roofing, window frames, panel siding, and other construction units. It is also used as wire insulation in the aircraft and electronics industries. Heat-shrinkable tubing made from PVDF resin forms "solder sleeves" that are used in the electronic, aerospace, and aircraft industries. It is used on a large scale for solid PVDF pipe and fittings and as a liner for pipes and valves. Piezoelectric PVDF films are very sensitive transducers with various applications [80].

2.9. Poly(Vinyl Fluoride)

Poly(vinyl fluoride) [*24981-14-4*] (PVF) is a high molecular mass, semicrystalline polymer with a melting point of 185 – 210 °C and the following structure:

Table 11. Properties of PVDF resins

Property	Kynar PVDF	Kynar Flex 2800	Solef 11010
Relative density	1.76	1.78	1.77
Tensile yield, MPa	35.9	23.4	30.8
Tensile strength, MPa	35.9	41.4	24.8
Elongation, %	37	525	430
Tensile modulus, MPa	1503	414	1000
Flexural modulus, MPa	1393	552	1000
Dielectric constant			
at 1 kHz	9		
at 1 MHz	8		
Dissipation factor			
at 1 kHz	0.014		
at 1 MHz	0.18		
Volume resistivity, $\Omega \cdot$ cm	2×10^{14}	8×10^{13}	6×10^{14}

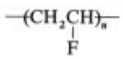

It is considered a thermoplastic but cannot be processed by conventional thermoplastic techniques because it is unstable above its melting point. It can be fabricated into films by mixing with "latent" solvents and extruding at a temperature below the melting point. (Dissolution in "latent" solvents takes place at an elevated temperature below the melting point of the PVF.) Films made from PVF have exceptional resistance to sunlight, chemical attack, and water adsorption. They are also extremely tough and have good abrasion resistance. These unusual properties make PVF highly suitable for outdoor coatings or laminates. Films of various tensile modifications and thicknesses are available with adhereable and nonadhereable surfaces. These films are manufactured by Du Pont under the Tedlar trade name.

2.9.1. Production

Vinyl fluoride is polymerized by free-radical processes like those used for other fluoropolymers. It is, however, more difficult to polymerize than TFE or VDF and requires higher pressures [81]. Polymerization is carried out in aqueous media between 50 and 150 °C at a pressure of 3.4 – 34.4 MPa. The usual catalysts are peroxides and azo compounds. A continuous process employs water, VF, and catalyst at 100 °C and 27.5 MPa [82]. Polymers have also been prepared at lower temperatures (−30 to +55 °C) using oxygen-activated trialkylboron catalysts [83]. The polymerization temperature affects the crystallinity and the melting point of the product; higher temperatures also increase branching.

2.9.2. Properties

Physical and Mechanical Properties. The monomer units in PVF are predominantly head-to-tail, with 12 – 18 % head-to-head [84]. The proportion of head-to-head units decreases with decreasing polymerization temperature. Conversely, randomness increases with increasing polymerization temperature.

The number-average molecular mass of PVF resin ranges from 7.6×10^4 to 2.34×10^5, the weight-average molecular mass from 1.43×10^5 to 6.54×10^5. The ratio of weight-average to number-average molecular mass is 2.5 : 5.6 [85].

The melting point of PVF varies with polymerization temperature, but is in the range of 185 – 210 °C. Torsion-pendulum studies indicate that PVF has second-order transitions at +41 °C and −20 °C. The temperature range for continuous use is −70 to +110 °C. The mechanical (ASTM D 882 and 1004) and physical properties of PVF film follow [86], [87]:

mp	185 – 190 °C
Relative density	1.38 – 1.57
Tensile strength	48 – 120 MPa
Elongation	115 – 250 %
Tensile modulus	1700 – 2600 MPa
Impact strength	10 – 22 kJ/m
Tear strength	174 – 239 kJ/m
Coefficient of thermal expansion	5×10^{-5} K^{-1}
Dielectric constant, 1 MHz	4.8
Dielectric strength, 0.1 mm film	120 – 140 kV/mm

Chemical Resistance. Poly(vinyl fluoride) is insoluble in ordinary solvents below 100 °C. At higher temperature, it is soluble in polar solvents such as amides, ketones, tetramethylene sulfone, and tetramethylurea.

Films show good resistance to acids and bases at room temperature. The resistance to aliphatic (petroleum), aromatic, and alcoholic liquids is also very good. Acetone attacks PVF, especially when warm.

Weather and Radiation Resistance. Films have excellent light and weather resistance, showing no change after 10 years outdoor exposure. They are transparent to both UV and visible light. A 25-μm thick film has over 90 % transmission between 400 and 1400 nm and over 80 % between 230 and 400 nm.

In addition, PVF films show good resistance to high-energy radiation. Usable properties are retained up to 1 MJ/kg exposure to electron beam radiation.

2.9.3. Processing

The polymer is dissolved in a "latent" solvent. Pigments, stabilizers, and plasticizers can be added to the PVF organosol before extrusion. The organosol is extruded into film, and the solvent is evaporated and recovered. The film can be oriented in varying degrees to achieve different levels of tensile and tear strength. Exposure to such treatments as corona discharge or alkali metal in liquid ammonia solution imparts adhesiveness.

2.9.4. Uses

Poly(vinyl fluoride) films are laminated to sheets of metal with special adhesives. The laminates are used for exterior siding on industrial and residential buildings. Pigmented films can be made in any color. Films can also be laminated to hardboard, paper, vinyl, polystyrene, polyurethane, and other substrates. These laminates are used for wall coverings, aircraft cabin interiors, pipe covering, duct liners, as well as other applications. A weather-resistant PVF barrier is also resistant to abrasion, chemicals, solvents, and staining. Another application for PVF film is as a release sheeting in the molding of phenolic, epoxide, and polyester plastics, and as a release film for bag molding. It does not become brittle and strips easily from the mold. Toughness is maintained at elevated temperatures at all degrees of humidity and in contact with common solvents. The film has been used for solar stills because of its high transparency and resistance to solar radiation.

3. Fluoroelastomers

3.1. Introduction

Fluorine-containing elastomers are designed to maintain rubber-like elasticity in extremely severe environments at high temperature and in contact with various chemicals. Their primary use is in seals, where they function to prevent leakage after prolonged compression. Fluoroelastomers are rarely subjected to tension, and then only at modest elongations. Technologically important fluoroelastomers can be divided into two groups: (1) the fluorocarbon elastomers discussed here and the fluorosilicones developed in the 1950's, and (2) the fluoroalkoxyphosphazines introduced in 1975, which represent only a very small share of the fluoroelastomer market.

At service temperatures above 0 °C, noncross-linked elastomers are highly viscous, incompressible liquids. To prevent permanent flow deformation under an imposed force and to develop reversible elastic properties, these flexible macromolecules must be cross-linked. The restoring force that acts on a uniaxially or biaxially deformed elastomeric network is based on the lower entropy of the extended, and hence, oriented, chains rather than on an increase in enthalpy. In general, cross-linking is performed chemically by vulcanization to produce thermoset elastomer networks. However, physical cross-linking by crystallization of hard segments in segmented or grafted copolymers can also provide a network structure in so-called thermoplastic elastomers. The few known thermoplastic fluoroelastomers have a much lower high-temperature service ceiling than the thermoset elastomers, nevertheless they possess certain advantages in processibility because they can be reversibly cross-linked. Furthermore,

they generally yield less extractable material because they contain no curatives.

To achieve high solvent resistance and thermal stability in fluorocarbon polymers, low-temperature flexibility and a high fluorine content are desirable; these requirements conflict to some extent. Low-temperature flexibility is governed by the glass transition temperature (T_g), which depends mainly on the conformational flexibility of individual chain elements. Therefore, if the various conformational states can be interconverted more freely, the T_g is lower and the low-temperature performance is better.

If linear perfluorocarbon macromolecules are cross-linked, they do not exhibit elastomeric properties at ambient or lower temperatures because of the severely restricted rotation about the carbon–carbon bonds and the consequent lack of alternative conformations. Thus, a tetrafluoroethylene (TFE) homopolymer cannot be an elastomer even if it has no crystallinity in the temperature range mentioned. Alternating difluoromethylene and methylene groups are the principal structural feature of poly(vinylidene fluoride). This polymer has a sufficiently low T_g (−40 °C) for elastomeric properties, but very strong intermolecular forces cause crystallization (latent heat of fusion ΔH_f, 6 kJ/mol; *mp,* 165 °C) [88] and, thus, restrict conformational freedom. Because of these properties of the homopolymers, most fluorocarbon elastomers are based on structures that interrupt this crystallinity and yet benefit from the low T_g of short vinylidene fluoride (VDF) sequences. For example, copolymerization of hexofluoropropene (HFP) and VDF in the molar ratio of 3.5 : 1.0 produces a fluoroelastomer [*9011-17-0*] with no crystallinity and a T_g of ca. −20 °C. This composition (60 wt% VDF) is the basis of the fluorocarbon elastomer industry:

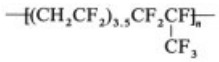

Copolymerization of a monomer that produces a crystalline but low-T_g polymer with a monomer that has a relatively bulky substituent produces an elastomer by interrupting the homopolymer crystallinity. Thus, copolymerization of VDF with HFP, HFP–TFE, chlorotrifluoroethylene (CTFE), perfluoro(methyl vinyl ether) (PMVE), or PMVE–TFE produces elastomers, copolymerization of TFE with PMVE or propylene also yields elastomers.

Fluoroelastomers with a higher fluorine content and superior low-temperature properties can be prepared by copolymerization with PMVE or similar perfluorinated vinyl ethers, although at greatly increased cost. With the exception of the perfluoropolymer poly(TFE–PMVE), fluorocarbon elastomers generally contain hydrogen and are also known as hydrofluoroelastomers.

The low-temperature properties of fluorosilicones (→ Silicones) and fluoroalkoxyphosphazenes (→ Inorganic Polymers) are superior to those of elastomers with backbones that consist only of carbon atoms substituted by hydrogen, fluorine, chlorine, fluoroalkyl groups, or simple fluoroalkoxy groups. Their outstanding chain flexibility is a result of the alternating oxygen or nitrogen backbone atoms. The low degree of substitution results in low activation energies for rotation about these atoms and adjacent chain atoms and, therefore, in low glass transition temperatures. However, fluorosilicones and fluoroalkoxyphosphazenes contain less fluorine than fluorocarbon elastomers; therefore, their solvent resistance and high-temperature stability are inferior.

Commercial fluorosilicone elastomers are polydimethylsiloxane homopolymers in which one methyl substituent of each silicon atom is replaced by a 3,3,3-trifluoropropyl group:

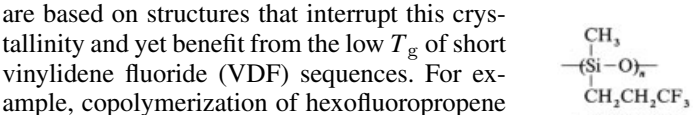

[*25791-89-3*]

Commercial fluoroalkoxyphosphazene elastomers are produced by replacing the chlorine in dichlorophosphazene polymers by fluorocarbon alkoxides [89]:

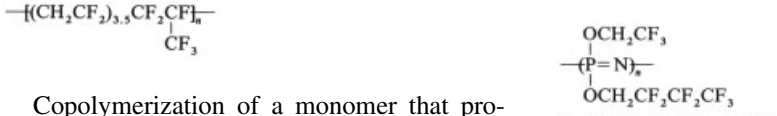

[*37002-15-6*] (T_g = −77 °C)

The service temperature range and fluorine and hydrogen content of fluoroelastomers are given in Table 12.

Table 12. Service temperature, fluorine content, and hydrogen content at fluoroelastomers

Type	Service temperature, °C ∗	Fluorine, %	Hydrogen, %
Hydrofluorocarbon			
VDF−HFP	−18 to 210	66	1.9
VDF−HFP−TFE	−12 to 230	66 – 69.5	1.1 – 1.9
VDF−PMVE−TFE	−27 to 230	64 – 66.5	1.1 – 1.7
TFE−P	5 to 200	54	4.3
Perfluorocarbon	− 0 to 260	73	0
Fluorosilicone	−65 to 175	37	4.5
Fluoroalkoxyphosphazene	−65 to 175	55	1.4

∗ Temperature of elastomeric side of glass transition to maximum temperature for continuous service.

3.2. Elastomers Based on Vinylidene Fluoride

Dipolymers of vinylidene fluoride (VDF) and chlorotrifluoroethylene (CTFE) [*9010-75-7*] prepared by M. W. Kellogg under contract with the United States Army Quartermaster Corps were first described in 1955 [90–92]. After vulcanization, this fluoroelastomer (Kel-F) exhibited previously unattainable resistance to heat and solvents. However, it was soon surpassed by an even more stable fluoroelastomer, a dipolymer of VDF and hexafluoropropene (HFP) [*9011-17-0*] , that was described in 1956 and commercialized by Du Pont in 1958 under the Viton A trade name [93–95]. In 1960, Du Pont introduced a TFE-containing terpolymer [*25190-89-0*] (Viton B) that was even more thermally stable and solvent resistant than Viton A [96], [97]. The Minnesota Mining and Manufacturing Co. (3M), who had acquired the fluorocarbon polymer assets of M. W. Kellogg, introduced a VDF−HFP copolymer under the Fluorel trade name in 1958 and continued its manufacture after 1962 under license from Du Pont.

Elastomeric copolymers of VDF and 1-hydropentafluoropropene [*32552-63-9*] and a TFE-containing terpolymer [*29830-35-1*] were introduced by Montecatini–Edison during the 1960s [98–100]. However, these polymers, sold under the Tecnoflon name, lacked the stability of similar HFP compositions, and were replaced by HFP analogs as early patents for poly(VDF–HFP) and poly(VDF–HFP–TFE) expired. Daikin Kogyo began production of VDF-based fluoroelastomers (Dai-el) in 1970 under license from Du Pont. A dipolymer, known as SKF-32 and produced in the Soviet Union, is not traded on the world market.

Trademarks and manufacturers of commercial VDF-based fluoroelastomers are shown in Table 13.

Table 13. Commercial VDF-based elastomers

Copolymer	Trade name	Manufacturer
VDF−HFP	Viton	Du Pont
	Fluorel ∗	3M
	Tecnoflon	Montefluos
	Dai-el ∗	Daikin
VDF−HFP−TFE	Viton ∗	Du Pont
	Fluorel ∗	3M
	Tecnoflon	Montefluos
	Dai-el ∗	3M
VDF−PMVE−TFE	Viton ∗∗	Du Pont

∗ Also available with proprietary cure site for peroxide curing.
∗∗ Contains perfluoro(methyl vinyl ether) and proprietary cure site; can be cured only with peroxide.

3.2.1. Production

Vinylidene fluoride based elastomers are usually prepared by continuous or batch polymerization in aqueous emulsion with free-radical initiators. Although free-radical solution polymerization is also possible at practical polymerization temperatures, chain transfer keeps molecular mass low in the solvents that must be used. The monomers are mainly derived from fluorocarbon refrigerants and their intermediates, and are relatively inexpensive because they are manufactured on a large scale.

Typical aqueous emulsion polymerizations are conducted at 5 – 7 MPa and 100 – 120 °C [101]. Commonly used initiators are inorganic peroxides such as ammonium persulfate; however, organic peroxides and redox systems are also used, especially at low temperature where the thermal decomposition of persulfate is slow. Emulsifying agents, when required, must be

compatible with the highly reactive fluorocarbon radicals of growing polymer chains; salts of fluorocarbon acids are used to prevent chain transfer. Molecular mass is regulated by controlling the rate of initiation and polymerization; chain-transfer agents may also be useful. Such agents generally have reactive carbon–hydrogen or carbon–halogen bonds; diethyl malonate and carbon tetrachloride are typical. Buffers control the pH of the polymerization and stabilize the emulsification system. Molecular mass ($\bar{M}_w$) ranges from 1.5×10^5 to 10^6.

In batch polymerization, the ingredients are heated in the reactor and the reaction is allowed to proceed until the pressure drop corresponds to the desired extent of reaction. In semibatch polymerizations, monomers can be added continually under pressure at rates that correspond to their consumption, providing better control of copolymer composition.

In a continuous process, water, monomers, initiator, and other components are fed to the reactor while polymer latex is removed at a corresponding rate. Composition and molecular mass can be precisely controlled. Monomers and other reactants of widely ranging reactivities can be used without undue variation in polymer structure. Unreacted monomers are evaporated from the latex and recycled.

In either process, the latex from the reactor is coagulated by the addition of salts or acids. The polymer crumb is filtered or centrifuged and washed to remove coagulant and water-soluble residues. After being dried, it may be mixed with curing (cross-linking) agents. The finished product is packaged as extruded pellets and lumps or in milled sheets.

3.2.2. Properties

Heat Resistance. Vinylidene fluoride based polymers are the largest group of fluoroelastomers in terms of production volume and variety. Their properties (Table 14) serve as a standard for comparing other fluoroelastomers that may be superior in certain respects such as lowtemperature flexibility (fluorosilicones and fluoroalkoxyphosphazenes), but are less satisfactory at high temperatures and in destructive fluids. An exception is the poly(TFE–PMVE–X) perfluoroelastomer (where X = cure-site monomer), which is virtually chemically inert and extremely stable to heat and oxidation. Vinylidene fluoride based elastomers can be used continuously at temperatures as high as 230 °C, depending on polymer type and formulation. Dipolymers retain over 50 % of their tensile strength for more than 1 year at 200 °C or 2 months at 260 °C. Functionalized hydrocarbon elastomers that are considered to be resistant to oil, such as nitrile or acrylate rubber, lose 50 % of their elongation aging at 150 °C for 10 and 25 days, respectively; in comparison VDF-based fluoroelastomers retain over 95 % of their elongation after 1 year at 200 °C [102].

Table 14. Physical properties of VDF-based fluoroelastomers

Property	Range	ASTM
Relative density of raw polymer	1.77 – 1.91	
Tensile properties		
100 % modulus, MPa	2 – 15	
Tensile strength, MPa	5 – 20	
Elongation at break, %	50 – 500	
Hardness, durometer A	60 – 95	
Compression set ∗, %		D-395
70 h at 25 °C	10 – 15	
70 h at 200 °C	10 – 40	
1000 h at 200 °C	50 – 70	
Low-temperature properties		
Clash–Berg torsional-stiffness temperature (69 MPa), °C	−31 to −5	
Retraction temperature (TR-10), °C	−31 to −6	D-1329
Brittle point, °C	−51 to −34	D-746

∗ Method B, 24.5 × 3.5 mm O-rings.

In the retraction-temperature test (Table 14) (ASTM D-1329), an elastic article is stretched at room temperature, cooled to a specified lower temperature, and allowed to retract. In general, it does not reach its original length but remains larger. The percent increase in length is denoted as TR, e.g., TR–10 for a 10 % increase.

Fluid and Chemical Resistance. The resistance of VDF–HFP and VDF–HFP–TFE elastomers to petroleum-based fuels and oils is excellent. Certain lubricants, however, contain additives that disperse sludge and inhibit metal corrosion, but are highly basic or decompose to basic products that attack the elastomer. Therefore, the performance of the fluoroelastomer must be evaluated in contact with the particular fluid containing its proprietary additives. Formulators of oil additives should examine not only lubricant stabilization and metal protection [103–

105], but also the behavior of seals. Fluoroelastomers to be used with hydrocarbon–alcohol fuels (gasohol) must be chosen with care [106]. A VDF–TFE polymer of 69.5 % fluorine content is used in methanol blends to prevent excessive swelling. This polymer is also recommended for certain phosphate ester based hydraulic fluids.

In contact with aromatic solvents, 68 or 69.5 % fluorine polymers give the best performance. Alcohols other than methanol do not greatly swell VDF-based elastomers despite their polar nature, but blends of alcohols and other fluids may act differently. Polar solvents that are not hydrogen bond donors, but only acceptors, tend to have a high affinity for these elastomers and may actually dissolve raw polymer or destructively swell vulcanizates. Examples of such solvents are low-boiling ketones and esters (acetone, methyl ethyl ketone, ethyl acetate), amides (*N,N*-dimethylformamide, *N,N*-dimethylacetamide), and ethers (tetrahydrofuran, 2-methoxyethyl ether). Mixtures that contain substantial amounts of these solvents cannot be used with this group of fluoroelastomers.

In addition to the destructive action of highly swelling solvents, the vulcanizate may be attacked chemically. Dehydrofluorination by primary and secondary aliphatic amines occurs readily and is the basis of amine cross-linking. The resulting unsaturated bonds are susceptible to nucleophilic attack, which can lead to cross-linking and embrittlement. With amine-based curing systems, which are seldom used, dehydrofluorination is a controlled process; however. uncontrolled exposure to soluble aliphatic amines and other soluble strong organic bases degrades these vulcanizates and must be avoided. Resistance to aqueous alkali is far greater, but will eventually cause surface cracks and failure. Phase-transfer catalysts such as quaternary ammonium ions accelerate degradation by aqueous alkali.

Resistance to aqueous acids is generally good, especially if acid acceptors such as calcium hydroxide, magnesium oxide, or calcium oxide are not present in large amounts. Lead-based acid acceptors such as sublimed litharge and dibasic lead phosphite are used for optimal acid resistance; these must be handled by safe procedures. The use of high-concentration dispersions of litharge in fluoroelastomers prevents litharge dust problems during compounding. Such dispersions must not be heated above 200 °C to avoid exothermic decomposition to metallic lead, which occurs with an 80–20 litharge–fluoroelastomer mixture [107].

Other Types of Resistance. Ozone resistance and weatherability are excellent [108] and permeability to gases is low. Vulcanizates can be formulated to be self-extinguishing or nonflammable. Abrasion resistance varies. Resistance to β-radiation is moderate and independent of VDF–copolymer composition. Radiation doses are cumulative and a total exposure to $10^4 - 10^5$ J/kg produces moderate to severe effects on physical properties. Radiation failure occurs by embrittlement rather than chain cleavage [109].

3.2.3. Processing

Formulation. Elastomer processing generally includes formulation. Optimum conditions must also be selected to provide the properties desired after vulcanization (cross-linking). The composition (in parts per hundred of rubber, phr) of a typical fluorocarbon elastomer follows:

Fluorocarbon elastomer	100
Acid acceptor (inorganic base)	3 – 15
Filler	0 – 50
Curing agents (sometimes included before formulation)	0 – 5
Processing aids	0 – 3

In O-ring formulations (Table 15), compression-set resistance is of critical importance, because sealing can be maintained only if the O-ring can retain some tendency to regain its uncompressed dimensions. Higher molecular mass and thus higher viscosity usually display better compression-set resistance. The choice of curing systems is also important because the stability of the cross-links limits the stability of the elastomer network. The bisphenol curing system offers greater compression-set resistance than the peroxide curing system and is generally preferred for O-ring and similiar sealing applications. However, peroxide-cured stocks usually have better resistance to steam and acids and are often specified [101].

Table 15. Fluoroelastomer molding and O-ring compounds

	General-purpose molding	O-ring material
Formulation		
Poly(VDF–HFP) [a]	100	
Poly(VDF–HFP) [b]		100
Benzyltriphenylphosphonium chloride	0.25	0.60
Hexafluoroisopropylidenediphenol	1.0	2.0
Calcium hydroxide	6.0	6.0
Magnesium oxide [c]	3.0	3.0
MT carbon black (N-908)	30	30
Compound viscosity [d]	27	35
Vulcanizate properties [e]		
Tensile, 25 °C		
100 % Modulus, MPa	2.2	5.3
Tensile strength, MPa	11.2	12.4
Elongation at break, %	360	190
Compression set, % [f]	47	16
Hardness, durometer A	66	72

[a] Viton E-45 (Du Pont), 100 °C Mooney viscosity, ML-10 = 45.
[b] Viton E-60 (Du Pont), 100 °C Mooney viscosity, ML-10 = 100.
[c] Maglite D (Merck & Co).
[d] Mooney scorch minimum viscosity, MS at 121 °C for 30 min.
[e] Press-cured 10 min at 177 °C, postcured 24 h at 232 °C.
[f] ASTM D-395, Method B (O-rings), 70 h at 200 °C.

In formulations for molded articles (Table 15), compression-set resistance is usually sacrificed in favor of elongation. A lower cure state and greater elongation are achieved by using less cross-linking agent or a lower molecular mass polymer; this has the important effect of reducing viscosity and increasing mold flow. High elongation is often required, because many mold designs require substantial elongation of the hot vulcanizate to facilitate removal from the mold. The tear strength at the demolding temperature, which increases with increasing elongation, is also an important processing requirement, especially if adhesion to the mold is high. If these processing requirements for elongation have been satisfied, molded articles are not usually used in tension, but rather under slight compression, although some elongation may occur during installation.

Vinylidene fluoride based elastomers are available in a range of molecular masses that correspond to Mooney viscosity values of 0–160 (ML-10, 121 °C). (The Mooney viscosity is a measure of polymer plasticity.) The choice of fillers and acid acceptor in fluoroelastomers has a major effect on viscosity and, therefore, influences processibility and properties after vulcanization [110], [111].

Specialized polymers can be compounded for exceptional resistance to fluids and low-temperature flexibility (Table 16). The PMVE polymer exhibits the best low-temperature properties of any commercial fluorocarbon elastomer [112]. Excellent resistance to methanol and methanol–fuel blends are characteristic of compounds based on the polymer containing 69.5 % fluorine [113].

Mixing. Standard rubber compounding equipment and techniques are used to mix VDF-based fluoroelastomer stocks. Rubber mills are suitable for most viscosity grades, although some reduction in molecular mass may occur when high-viscosity polymers are cold milled and the clearance between rolls is small. Banbury and other internal mixers are also suited for preparing precompounds that also contain fillers, inorganic acid acceptors, and processing aids. Producers commonly supply precompounds that are then fully compounded by a custom-mixing firm or by the molder.

Molding. Compression molding is usually the most economical and flexible form of vulcanization. When extruded or cut preforms are used, the amount of mold flashing is kept to a minimum and higher viscosity polymers can be used because mold-flow requirements are low.

Transfer molding requires lower viscosities than compression molding because the compound must flow more easily. Bisphenol-curable compounds are suited to transfer molding because of their processing safety (premature vulcanization during processing is low). Only small articles are normally produced by this technique.

Injection molding requires a higher capital investment than compression or transfer molding and is usually less labor intensive. Injection molding is used increasingly and is justified in large-scale production where labor savings offset equipment costs and increased waste from mold filling channels. Newer mold designs reduce this waste. Compound formulation is critical for injection molding because high flow rates and compound safety are required. Compounds must be characterized rheologically before processing conditions can be established. Injection molding of thermoset elastomers is much more demanding than injection molding of thermoplastics or thermoplastic elastomers because the

Table 16. Specialty compounds for increased fluid resistance and low-temperature flexibility

	Terpolymers			Standard
	methanol resistant	standard	low-temp. resistant	dipolymer
Formulation				
Poly(VDF−HFP−TFE) [a], 69.5 % F	100			
Poly(VDF−HFP−TFE) [b], 68.5 % F		100		
Poly(VDF−PMVE−TFE) [c]			100	
Poly(VDF−HFP) [d], 65 % F				100
Triallylisocyanurate	3		4	
Organic peroxide [e]	3		4	
Sublimed litharge	3			
MT black (N-908)	30	30	30	30
Calcium hydroxide		3	4	6
Magnesium oxide [f]		3		3
Benzyltriphenylphosphonium chloride		1.0		0.6
Hexafluoroisopropylidenediphenol		2.0		2.0
Compound viscosity [g]	40	72	32	35
Vulcanizate properties [h]				
100 % modulus at 25 °C, MPa	5.4	7.0	5.3	5.3
Tensile strength, MPa				
at 25 °C	17.6	15.0	13.8	12.4
at −30 °C			37.4	45.5
at −40 °C			39.9	28.4
Elongation at break, %				
at 25 °C	210	210	200	190
at −30 °C			160	60
at −40 °C			60	20
Brittle point (1.91-mm thick), °C [i]	−45	−42	−51	−34
Clash–Berg torsional-stiffness temperature (69 MPa), °C	−5	−12	−31	−17
Compression set, % [j]	38	30	30	16
Hardness, durometer A	76	80	67	72
Volume swell, 24 °C for 7 d, %				
Toluene	4	13		
Reference fuel C [k]	3	6		
Methanol	3	22		
Reference fuel C/methanol (85/15)	9	20		
Phosphate ester hydraulic fluid, 121 °C [l]	45	127		
Reference fuel C/ethanol (85/15), 100 °C	18	24		
Permeability to automotive fuel, g [m]				
Reference fuel C	1	3		6
Reference fuel C/ethanol (85/15)	11	19		28
Reference fuel C/methanol (85/15)	17	57		125

[a] Viton GF, contains peroxide-sensitive cure site (Du Pont).
[b] Viton B (Du Pont).
[c] Viton GLT, contains peroxide-sensitive cure site (Du Pont).
[d] Viton E-60 (Du Pont).
[e] Luperco 101-XL (Pennwalt Corp.).
[f] Maglite D (Merck & Co.).
[g] Mooney scorch minimum viscosity, MS at 121 °C for 30 min.
[h] Press-cured 10 min at 177 °C, postcured 24 h at 232 °C.
[i] ASTM D-2137.
[j] ASTM D-395, Method B (O-rings), 70 h at 200 °C.
[k] 50 % toluene, 50 % isooctane.
[l] Skydrol 500-B (Monsanto Co.).
[m] ASTM E-96, 0.45 mm membrane, 24 °C.

former case entails irreversible chemical reactions that increase viscosity and create cross-links during hot flow; in the latter case, the polymer flows easily in the melt and is simply cooled or crystallized in the mold.

Molding cycles are usually 1 – 5 min. at 160 – 200 °C and depend on the size of the article because of heat transfer limitations. In injection molding, however, frictional or shear heating, produced by high flow rates through small mold orifices, raises the temperature in the mold significantly. In large moldings produced by compression or injection molding, the heat of reaction of the vulcanization must also be considered.

Extrusion. Fluorocarbon elastomers can be shaped and cured by extrusion. The fully compounded stock is continuously forced through a shaping die at the end of an extruder below the curing temperature. The continuous extrudate, which may be hose or tubing, rod stock for spliced O-rings, or profile stock for cut gaskets, is rolled up or otherwise placed in a steam autoclave, where curing takes place at ca. 160 °C. Extruded stock is also used to cut preforms for subsequent curing by compression or transfer molding.

Postcuring. Fluorocarbon elastomers are usually postcured to develop optimum physical properties, such as compression-set resistance and tensile strength [114]. The postcure cycle subjects a press-cured article to circulating air at atmospheric pressure at 200 – 260 °C for 12 – 24 h. Small amounts of volatile substances are removed, and slight shrinkage occurs. Typical bisphenol–curable stocks undergo a 0.3 % mass loss on press curing and a further 1.2 % loss on postcuring [115], [116]. Postcuring is believed to complete the vulcanization and deactivate remaining intermediates. In this view, improved compression-set resistance on postcuring is seen as reduction of the formation of a second network under compression, with more complete recovery to the uncompressed dimensions. The conditions for a postcure cycle are established rather arbitrarily, because each elastomer-curing system has a point at which improvement of properties on heating a vulcanizate in air gives optimum properties, beyond which further heating or heating at a higher temperature affects physical properties undesirably with the onset of heat aging.

Vulcanization Chemistry. The chemical reactions that result in formation of networks in VDF–HFP-based fluoroelastomers compounded with bisphenols or other polyhydroxy aromatics, quaternary phosphonium or ammonium salts, and inorganic bases are believed to involve dehydrofluorination of HFP–VDF–HFP or HFP–VDF–TFE sites. The resulting double bonds are susceptible to nucleophilic substitution because of their electron-deficient character. Thus, a quaternary phosphonium or ammonium phenolate displaces a vinylic or allylic fluorine and forms the first attachment of an aromatic bisether cross-link [101], [117], [118].

Peroxide-initiated cross-linking of bromine-containing fluoroelastomers in the presence of radical traps such as triallyl isocyanurate is believed to proceed by radical-initiated polymerization of the radical trap to form a network that retains a substantial fraction of the allyl functions. When further polymerization of these functions becomes slow, the propagating radicals abstract bromine atoms from the fluoroelastomer; the resulting fluorocarbon radicals attack intact allylic double bonds. The driving force for this reaction is the addition of an electron-deficient free radical to an electron-rich double bond. Propagation reactions may involve redistribution of the free-radical intermediates from fluoroelastomer radical addition to a double bond of the partially polymerized radical trap [101], [119].

3.2.4. Uses

Vinylidene fluoride based elastomers are used widely in vehicle construction because of their excellent resistance to heat and oil. Important fluoroelastomer parts are engine oil seals of both the rotating and reciprocating types, drive-train seals, and fuel-handling components such as gaskets, O-rings, valve seals, and hoses. Other applications are hydraulic and pneumatic seals, hose and seals for chemical industry, flue-duct expansion joints for power plants and pollution control equipment, and fuser rolls for copying

machines. The aircraft industry uses more fluorosilicone than fluorocarbon elastomers because of the extreme low-temperature requirements of high-altitude flight [120].

3.3. Elastomers Based on Tetrafluoroethylene–Perfluoro(Methyl Vinyl Ether) Copolymers

Despite the vast improvement that VDF-based elastomers offered over conventional hydrocarbon elastomers in thermooxidative stability and oil resistance, even more stable elastomers were required for critical industrial and military applications. In response to this need, perfluorinated materials were developed at Du Pont in the 1960s. The goal was an elastomer with the inertness and solubility characteristics of PTFE or poly(TFE−HFP).

To create such a stable elastomer, a comonomer was required that would interrupt TFE crystallinity and impart good low-temperature properties. At this time, the epoxidation of HFP was developed and the resulting hexafluoropropylene oxide could be isomerized to perfluoro(methyl vinyl ether) (PMVE) (→ Fluorine Compounds, Organic, Chap. 6.2.).

Copolymerization of PMVE with TFE gave an elastomer [*26425-79-6*] whose thermal, oxidative, and chemical stability before cross-linking approached that of perfluorocarbon plastics [121]. However, the inertness of this base elastomer and the absence of cure sites prevented cross-linking, and most of the effort to develop a practical elastomer went into cure site monomers [122–127]. Several types of curing systems were developed, and articles made from them were commercialized by Du Pont in the mid-1970s under the Kalrez Perfluorocarbon Elastomer Part trade name.

3.3.1. Production

Copolymerization of TFE–PMVE and cure-site monomers is conducted in an aqueous emulsion like that used with VDF–HFP-based systems (see 3.2.1). Polymerization rates in continuous reactors are about one-tenth those of VDF–HFP-based systems at temperatures that give useful molecular mass. Initiation is effected by thermal dissociation of persulfate or by redox couples such as persulfate–sulfite. An important difference between the polymerization of VDF–HFP and TFE–PMVE is that the latter require a surfactant. Hydrocarbon-based surfactants cannot be used because of chain-transfer reactions with the highly reactive perfluorocarbon chain radicals; salts of C_8-C_{12} perfluorocarbon acids are satisfactory.

The polymer latex is coagulated by addition of polyvalent metal salt solutions. The polymer crumb is filtered or centrifuged and washed to remove coagulant, initiator, and surfactant and then dried [125].

3.3.2. Properties

Heat Resistance. The mass loss of a TFE−PMVE copolymer in air over 7 d at 316 °C is 6.5 %; an uncured terpolymer that contains the proprietary perfluorophenoxy ether-based cure-site monomer loses 12 % under the same conditions [125]. This typical behavior is to be expected in a perfluoroelastomer because the cure site has a more complex structure than the backbone and is more prone to degradation under these extreme conditions. Degradation of the cure site would also be expected to lead to some local degradation of the main chain.

Cured perfluoroelastomer specimens are also exceptionally stable and are differentiated by the stability of their cross-links. A perfluoroelastomer with a perfluorocarbon nitrile group as cure site retains physical properties best under extreme heat conditions. The triazine cross-links that are believed to form by trimerization of the nitrile functions are sufficiently stable to prevent any change from the original postcured properties with respect to 100 % modulus, tensile strength, and elongation at break over 18 d at 288 °C in air. Even at 316 °C, useful properties are maintained: 100 % modulus is reduced by less than 20 % and tensile strength by less than 40 %; elongation at break is unchanged.

Although perfluoroelastomers cross-linked through the perfluorophenoxy group are less stable to heat, they are still far more stable than hydrofluoroelastomers: after 20 d at 288 °C in air, 100 % modulus is reduced by less than 25 % and tensile strength by less than 15 %; elongation at break is increased 120 %.

Fluid and Chemical Resistance. Uncured TFE−PMVE-based perfluoroelastomers are insoluble in all solvents except certain perfluorocarbons and chlorofluorocarbons. Dilute solutions can be prepared in Freon 113 or 2,2,3-trichloroheptafluorobutane containing 2 – 3 % of a highly polar solvent such as methanol or tetrahydrofuran.

In 10 d at 25 °C, mass increases 61 % in diethylamine, 90 % in fuming nitric acid, and 41 % in Skydrol 500-B, a phosphoric ester based hydraulic fluid; Freon 113 swells the polymer, with a 40 % mass gain [125], [126]. Permeability values for a 0.127 mm calendered film at 24 °C, in g mm^{-1} h^{-1} cm^{-2} are as follows:

Benzene	1.5
Carbon tetrachloride	1.3
Ethyl acetate	1.7
Methanol	0.3
Diethylamine	0.5
Isooctane–toluene 1 : 1 (Reference fuel C)	1.1
Water	< 1.3

The swelling resistance of cured carbon black-containing TFE−PMVE−X perfluoroelastomers in various liquids at elevated temperature is shown in Table 17. Large differences exist between the triazine, K_2−AF, and peroxide curing systems. The triazine cure refers to trimerization of perfluorocarbon nitrile groups. The K_2−AF cure refers to the dipotassium salt of bisphenol AF that is used to cross-link polymers with perfluorophenoxy groups. The peroxide cure indicates free-radical cross-linkage of a bromine-containing perfluoroelastomer.

Typical physical properties of a K_2−AF-cured vulcanizate loaded with carbon black follow [126]:

100 % modulus	9 MPa
Tensile strength	16 MPa
Elongation at break	160 %
Hardness, Shore A	75
Compression set B (O-rings)	
70 h at 121 °C	23 %
70 h at 288 °C	45 %
Clash–Berg torsional-stiffness temperature (66.9 MPa)	−2 °C
Brittle point	−38 °C
Retraction temperature	
TR-10	−1 °C
TR-50	+8 °C
Relative density	2.05

3.3.3. Processing and Uses

Perfluoroelastomers based on TFE−PMVE are difficult to process because of the stiffness of the fully fluorinated chains. Since the polymer is also difficult to cross-link and expensive to produce, Du Pont commercialized these materials in the form of finished articles under the trade name Kalrez Perfluorocarbon Elastomer Parts. A substantial market was found for O-rings, flat and lathe-cut gaskets, rotating and reciprocating shaft seals, and diaphragms. Other products include coated fabrics, hose linings, tubing, industrial gloves, and extruded articles such as autoclave and oven seals. The high cost of these products is justified in applications where no other material is equally resistant to oxidative, thermal, fluid, and chemical exposure. They are used in the manufacturing, aerospace, and automobile industries.

3.4. Elastomers Based on Tetrafluoroethylene–Propylene Copolymers

Du Pont [131] and Asahi Glass [132–134] established independently in the 1960s that TFE and propene (P) undergo a nearly alternating polymerization. The resultant TFE−P fluoroelastomer [*27029-05-6*] exhibits very good thermal and chemical resistance, but has a high glass transition temperature (−2 °C) despite its low fluorine content (54 %). Although resistance to swelling by aromatic hydrocarbons is inadequate, these elastomers offer excellent resistance to certain polar solvents. An unusual property of TFE−P fluoroelastomers is their exceptional resistance to dehydrofluorination by organic bases. As a result, they are cured by peroxide-radical trap systems [135], unless a specific cure site is present [136], [137].

Asahi Glass introduced TFE−P polymers in the mid-1970s under the Aflas trade name. In 1987, TFE−P-based fluoroelastomers represented only a small fraction of the total fluorocarbon elastomer market.

Table 17. Chemical and fluid resistance of perfluoroelastomers [128–130]

Liquid	Temperature, °C	Time, h	Swell of cured elastomer, %		
			Triazine	K_2-AF	Peroxide
Nitric acid	85	70	3	8	9
Acetic acid	100	70	3	32	16
Aniline	100	166	5	2	6
Butyraldehyde	70	70	13	10	14
Toluene	100	70	7	7	7
Skydrol 500-B	125	70	9	5	6
Methyl ethyl ketone	70	70	6	4	4
Water	225	70	27	60	7

3.4.1. Production

Tetrafluoroethylene–propylene elastomers are prepared in aqueous emulsion with a perfluorocarbon surfactant. Batch copolymerization at 70 – 80 °C, using a conventional ammonium persulfate–sodium sulfite redox catalyst, produces elastomers of relatively low molecular mass ($\bar{M}_w$=70 000) because of chain transfer. Polymerization below 60 °C to increase molecular mass requires a change in initiator because the persulfate–sulfite couple is too unreactive at such temperatures. To maintain a reasonable polymerization rate, more active redox initiators were developed [134]. With these catalysts, combinations of ammonium persulfate, ferrous sulfate, ethylenediaminetetraacetic acid, and sodium hydroxymethanesulfinate, efficient batch polymerization to a $\bar{M}_w$ of 15×10^4 can be carried out in the 25 – 35 °C range. The emulsion is buffered with disodium hydrogen phosphate and sodium hydroxide to maintain the pH between 5.5 and 10.

The polymer latex is coagulated by addition of solutions of polyvalent salts, and the coagulated crumb is washed and dried.

3.4.2. Properties

Typical properties of a TFE−P elastomer vulcanized by a peroxide–radical trap curing system (press-cured at 160 °C for 30 min, postcured at 200 °C for 2 h) are as follows [134]:

Raw polymer	
Relative density	1.55
Mooney viscosity (ML-10, 100 °C)	85
Appearance	dark brown
Formulation	
Poly(TFE−P)	100 phr
α,α'-Bis(*t*-butylperoxy)diisopropylbenzene	2 phr
Triallylisocyanurate	3 phr
MT carbon black (N-908)	30 phr
Vulcanizate Properties	
Relative density	1.60
Tensile properties at 25 °C	
100 % modulus	3 MPa
Tensile strength	20 MPa
Elongation at break	300 %
Compression set at 200 °C	
1 d	40 %
30 d	65 %
Hardness, Durometer A	72
Brittle point	−40 °C
Retraction temperature (TR-10)	3 °C
Volume increase after immersion	
95 % H_2SO_4, 100 °C, 3 d	4.4 %
Water, 150 °C, 3 d	8.7 %
Steam, 160 °C, 3 d	4.6 %
Fuel Oil B, 25 °C, 7 d	59 %
Benzene, 25 °C, 7 d	40 %
Methanol, 25 °C, 7 d	0.2 %
Ethyl acetate, 25 °C, 7 d	88 %

3.4.3. Processing

Tetrafluoroethylene–propylene fluoroelastomers are formulated as follows: Dialkyl peroxides are used with radical traps such as polyallyl or polyvinyl compounds or *N,N′-m*-phenylenedimaleimide. An inorganic acid acceptor such as calcium hydroxide or magnesium oxide is also present. Fillers such as silicas can be used, but carbon blacks are usually chosen for reinforcement [134].

Tetrafluoroethylene–propylene elastomers process less smoothly on conventional rubber processing equipment than VDF–HFP-based fluoroelastomers. Processing aids such as carnauba wax or sodium stearate are widely employed. Similarly, longer press-cure cycles are used to develop optimum properties: the press-cured vulcanizates are usually postcured at 200 °C for 16 h [138], [139].

3.4.4. Uses

The TFE–P fluoroelastomers are used where high temperatures and destructive media are encountered. Applications are similar to those described for VDF–HFP-based polymers (Section 3.2.4), but the choice may depend on specific requirements. Thus TFE–P is recommended for exposure to strong organic bases and certain highly polar solvents, but not for maximum oil resistance and low-temperature properties.

4. Toxicology and Occupational Health

All commerical fluoropolymers today are low in toxicity and safe under normal conditions. They are physiologically inert and some have been approved by the FDA for use in contact with food or for human implants. Incineration of fluoropolymers should be avoided because the combustion products include hydrogen fluoride, a toxic and corrosive gas. Certain processing hazards are discussed separately for fluoroplastics and fluoroelastomers.

4.1. Fluoroplastics

Polytetrafluoroethylene resins are subjected to prolonged heating at 380 °C during sintering. Since a small quantity of toxic fumes can be emitted during this process, adequate ventilation should be provided to prevent inhalation of the fumes, which can cause a condition known as polymer fume fever. This temporary flulike disorder occurs several hours after exposure and lasts for 36 – 48 h. The effect is not cumulative. It is often caused by smoking tobacco contaminated with PTFE powder. Therefore, smoking should be forbidden in areas where PTFE is processed. Safe handling practice must be observed during processing [34]. Enormous quantities of PTFE resins have been manufactured, processed, and used without causing any known cases of serious illness or death.

The same precautions should be followed with all TFE-containing polymers. In the extrusion or molding of TFE copolymers, the resin is heated above 300 °C (to 425 °C in the case of PFA resins). Exhaust hoods should be used over the die and feed hopper to remove any toxic fumes that may be evolved.

The main hazard associated with the processing of PVDF, ECTFE, and PVF resins is the evolution of hydrogen fluoride; this is prevented by avoiding overheating. Exhaust hoods should be used over the feed and die sections of the extruder or molding machine to remove hydrogen fluoride.

4.2. Fluoroelastomers

Fluorocarbon elastomers are nontoxic and nonirritating when handled in accordance with manufacturers' recommendations. Detailed literature on safety is available from the manufacturers; the following general principles should be observed: (1) adequate ventilation of compounding, extrusion, curing, and postcuring areas should be provided; (2) dispersion of finely divided active metals or high concentrations of active dehydrofluorinating agents (e.g., aliphatic amines and combinations of phase-transfer catalysts with metal oxides or hydroxides) should be avoided; (3) fluoroelastomers should not be heated to or above their decomposition temperature; and (4) incineration should not be used as a waste disposal method because of the attendant evolution of hydrogen fluoride.

Certain compounding ingredients may pose specific hazards, and manufacturers' or suppliers' recommendations should be followed.

Certain fluoroelastomers have been approved for contact with foods under the Code of Federal Regulations, 21, Food and Drugs, Part 177.2600.

5. Economic Aspects

Fluoroplastics. Worldwide production capacity in 1987 for all fluoroplastics was ca. 45 000 t. Selling prices ranged from ca. $9 to 44/kg. Du Pont is the world's leading producer, with plants in Europe, Japan, and the United States. Other major producers include Asahi Glass (Japan), Ausimont (Italy), Daikin Kogyo (Japan), Dynamit Nobel (FRG), Hoechst (FRG), ICI (UK), Kureha Chemical Co. (Japan), 3M (USA), Pennwalt Corp. (USA), and Solvay & Cie. (Belgium).

Fluoroelastomers. Worldwide annual consumption of fluorocarbon elastomers was estimated to be 5000 t in 1987. Prices for non-PMVE fluorocarbon elastomers ranged from ca. $33 to 45/kg. Fluoroelastomers are produced by Du Pont (Deepwater, USA) and Dordrecht (Netherlands); by 3M (Decatur, USA) and Zwindrecht (Belgium); by Montefluos (Linate, Italy); by Daikin Kogyo (Osaka, Japan); and by Asahi Glass (Chiba, Japan.)

6. References

General References

1. R. G. Arnold, A. L. Barney, D. C. Thompson, *Rubber Chem. Technol.,* **46** (1973) 619.
2. M. W. Ranney: *Fluorocarbon Resins,* Noyes Data Corporation, Park Ridge, NJ, 1971.
3. C. A. Sperati, H. W. Starkweather, Jr., *Fortschr. Hochpolym. Forsch.* **2,** (1961) 465 – 495.
4. L. A. Wall: *Fluoropolymers,* vol. XXXV, Wiley-Interscience, New York 1972.

Specific References

5. Du Pont, US 2 230 654, 1941 (R. J. Plunkett).
6. Du Pont, US 2 946 763, 1960 (M. I. Bro, B. W. Sandt).
7. Du Pont, US 3 180 995, 1965 (J. F. Harris, D. I. McCane).
8. Du Pont, US 2 468 664, 1949 (W. E. Hanford, J. R. Rowland).
9. W. T. Miller in C. Slessar, S. R. Schramm (eds.): *Preparation, Properties and Technology of Fluorine and Organic Fluorine Compounds,* McGraw-Hill, New York 1951, pp. 567 – 713.
10. Du Pont, US 2 392 378, 1946 (W. E. Hanford).
11. H. W. Starkweather, *J. Am. Chem. Soc.* **56,** (1934) 1870.
12. Du Pont, US 2 435 537, 1948 (T. A. Ford, W. E. Hanford).
13. Du Pont, US 3 855 191, 1974 (T. R. Doughty, C. A. Sperati).
14. Allied, US 3 655 611, 1972 (M. B. Mueller et al.).
15. Du Pont, US 3 766 133, 1973 (R. Roberts, R. F. Anderson).
16. Du Pont, US 2 612 484, 1956 (S. G. Bankoff).
17. Du Pont, US 3 142 665, 1964 (A. J. Cardinal et al.).
18. ICI, US 4 038 231, 1977 (J. M. Douner et al.).
19. C. A. Sperati, H. W. Starkweather, Jr., *Fortschr. Hochpolym. Forsch.* **2** (1961) 465 – 495.
20. H. D. Chanzy, P. Smith, *J. Polym. Sci. Polym. Lett. Ed.* **24** (1986) 557.
21. R. C. Doban et al., Paper presented at 130th meeting of the ACS at Atlantic City, Sept. 1956.
22. ASTM D 1457-83,
23. T. Suwa, M. Takehisa, S. Machi, *J. App. Poly. Sci.* **17** (1973) 3253.
24. C. A. Sperati in J. Brandrup, E. H. Immergut (eds): *Polymer Handbook,* 2nd ed., J. Wiley & Sons, New York 1975, pp. V-29 – 36.
25. J. T. Milek, *A Survey Materials Report on PTFE Plastics,* AD 607 798, U.S. Dept. of Commerce, Washington, D.C., Sept. 1964.
26. Mechanical Design Data, *Teflon Fluorocarbon Resins,* Bulletin, Du Pont, Wilmington, Del.
27. J. Teflon (Du Pont), 11, Jan – Feb. (1970).
28. J. C. Siegle et al., *J. Poly. Sci. Part A* **2** (1964) 391.
29. *Teflon TFE-Fluorocarbon Resins Molding Techniques,* Bulletin, 2nd ed., Du Pont, Wilmington, Del., May 1966.
30. W. L. Gore, US 3 962 153, 1976 (R. W. Gore).
31. Kewanee Oil Co., US 3 993 584, 1976 (J. E. Owen, J. W. Vogt).
32. American Cyanamid, US 3 704 171, 1972 (H. P. Landi).
33. Du Pont, US 3 051 545, 1962 (W. Steuber).
34. D. I. McCane in N. M. Bikales (ed.): *Encyclopedia of Polymer Science and Technology,* Tetrafluoroethylene Polymers, vol. 13, J. Wiley & Sons, New York 1970, pp. 623 – 690.
35. V. A. Lis, I. F. Kashtanova, *Kauch. Rezina* 1985, no. 7, 20.
36. Du Pont, US 3 132 124, 1964 (M. J. Couture et al.).
37. Du Pont, US 3 085 083, 1963 (R. C. Schreyer).
38. Daikin Kogyo, DE-A 2 104 077, 1971 (Y. T. Kometani et al.).

39. A. D. English, O. T. Garza, *Macromolecules* **12** (1979) 352.
40. H. W. Starkweather, Jr., private communication.
41. R. K. Eby, F. C. Wilson, *J. Appl. Phys.* **33** (1962) 2951.
42. R. Y. M. Huang, P. J. F. Kanitz, *Polym. Prepr.* **10 (2),** Sept. 2, 1969.
43. "Teflon" FEP-Fluorocarbon Resin-Techniques for Processing by Melt Extrusion, 2nd ed., Information Bulletin, X-82, Du Pont, Wilmington, Del., 1960.
44. Du Pont, *Du Pont Innovation* **5** (1973), No. 1, p. 16.
45. Du Pont, US 3 635 926, 1972 (W. F. Gresham, A. F. Vogelpohl).
46. Du Pont, US 3 536 733, 1970 (D. P. Carlson).
47. Du Pont, US 3 674 758, 1972 (D. P. Carlson).
48. Du Pont, US 3 642 742, 1972 (D. P. Carlson); Du Pont, US 4 599 386, 1986 (D. P. Carlson et al.).
49. Du Pont, US 4 743 658, 1988 (J. F. Imbalzano, D. L. Kerbow).
50. R. L. Johnson: "Tetrafluoroethylene Copolymers with Perfluoroalkoxy Pendant Group", *Encyclopedia of Polymer Science and Technology,* Suppl. 1, J. Wiley and Sons, New York 1976.
51. E. W. Fasig et al., paper presented at the 22nd International Wire and Cable symposium, Atlantic City, N.J., December 1973.
52. "Teflon" PFA Fluorocarbon Resins: Melt Processing of "Teflon" PFA TE-9705, PIB #1, bulletin, Du Pont, Wilmington, Del., 1973.
53. "Teflon" PFA Fluorocarbon Resin: Injection Molding of "Teflon" PFA TE-9704, PIB #4, bulletin, Du Pont, Wilmington, Del., 1973.
54. Du Pont, US 3 624 250, 1971 (D. P. Carlson).
55. Asahi Glass Co., US 4 123 602, 1978 (H. Ukihashi, M. Yamabe).
56. Allied, US 3 893 978, 1975 (S. Chandrasekaran).
57. Hoechst, US 4 338 287, 1982 (H. Sulzbach et al.).
58. F. C. Wilson, H. W. Starkweather, Jr., *J, Polym. Sci. Polym. Phys. Ed.* **11** (1973) 919.
59. H. W. Starkweather, Jr., *J. Polym. Sci. Polym. Phys. Ed.* **11** (1973) 587.
60. B. Chu, C. Wu., *Macromolecules,* **20** (1987) 93.
61. Tefzel Fluoropolymer, *Design Handbook,* Du Pont, Wilmington, Del., 1973.
62. Du Pont, US 3 738 923, 1973 (D. P. Carlson, N. E. West).
63. "Technical Information" from the Fluoropolymers Division Technical Service Laboratory, No. 10, Jacketing Rate Calculation, bulletin, Du Pont, Wilmington, Del., August 1982.
64. Du Pont, US 2 626 254, 1954 (W. T. Miller, J. T. Maynard).
65. Union Carbide, US 2 700 662, 1955 (D. M. Young, B. Thompson).
66. 3M, US 2 842 528, 1958 (R. L. Herbst, B. F. Landrum).
67. 3M, GB 840 735, 1960.
68. H. Matsuo, *J. Polym. Sci.* **25** (1957) 234.
69. E. G. Safrin, W. A. Zisman, *Naval Res. Lab. Rept.* (1963) 5985.
70. J. P. Sibilia, L. Roldin, S. Chandrasekaren, *J. Polym. Sci., Part B.* **10** (1972) 549.
71. Montecatini, US 3 501 446, 1970 (M. Ragazzini et al.).
72. H. K. Reimschuessel, F. J. Rahl, *J. Polym. Sci., Pt. A., Poly. Chem.* **25** (1987) 1871.
73. A. B. Robertson, *Appl. Polym. Symp.* **21** (1973) 89 – 100.
74. Pennwalt, US 3 193 539, 1965 (M. Hauptschein).
75. Kureha Chem. Co., GB 1 094 558, 1967.
76. Du Pont, NL 68-15 435, 1969 (D. P. Carlson).
77. Pennwalt, EP 169 328, 1986 (L. A. Barber).
78. K. R. Hasegawa et al., *J. Polym. Sci. Polym. Phys. Ed.* **8** (1970) 1073.
79. *Machine Design,* April 16, 1987, p. 133.
80. M. G. Broadhorst, *J. Appl. Phys.* **49** (1978) 4992.
81. R. N. Haszeldine et al., *Polymer* **14** (1973) 221.
82. Du Pont, US 3 265 678, 1966 (J. L. Hecht).
83. Montecatini, GB 1 029 635, 1966.
84. C. W. Wilson, E. R. Santee, *J. Poly. Sci. Part C.* **1** (1965) 97.
85. F. S. Cohen, N. Kraft, *Encyclopedia of Polymer Science and Technology,* Vol. 14, J. Wiley & Sons, New York 1971, 522 – 540.
86. Technical Information Bulletin TD-31, Du Pont, Wilmington, Del., 1979.
87. Technical Information Bulletin TD-1A, Du Pont, Wilmington, Del., 1974.
88. G. J. Welch, R. L. Miller, *J. Polym. Sci. Polym. Phys. Ed.* **14** (1974) 1683.
89. H. R. Allcock et al., *Inorg. Chem.* **5** (1966) 1709.
90. M. E. Conroy et al., *Rubber Age N.Y.* **76** (1955) 543.
91. C. B. Griffis, J. C. Montermoso, *Rubber Age N.Y.* **77** (1955) 559.
92. W. W. Jackson, D. Hale, *Rubber Age N.Y.* **77** (1955) 865.

93. S. Dixon et al., *Ind. Eng. Chem.,* **49** (1957) 1687.
94. J. S. Rugg, A. C. Stevenson, *Rubber Age N.Y.* **82** (1957) 102.
95. Du Pont, US 3 051 677, 1962 (D. R. Rexford).
96. Du Pont, Report No. 59-4, 1959 (A. L. Moran).
97. Du Pont, US 2 968 649, 1961 (J. R. Pailthorp, H. E. Schroeder).
98. Montecatini-Edison S.p.A., US 3 331 823, 1967 (D. Sianesi et al.).
99. Montecatini-Edison S.p.A., US 3 335 106, 1967 (D. Sianesi et al.).
100. A. Miglierina, G. Ceccato, *Fourth Int. Syn. Rubber Symp.,* No. 2, (1969) 65.
101. W. W. Schmiegel, A. L. Logothetis in J. C. Arthur (ed.): "Polymers for Fibers and Elastomers", *ACS Symp. Ser.,* **260** (1984).
102. J. R. Dunn, H. A. Pfisterer, *Rubber Chem. Technol.* **48** (1976) 356.
103. A. Nersasian, *ASLE Trans.* **23** (1980) 343.
104. B. Frapin, *Rev. Gen. Caoutch. Plast.* **672** (1987) 125.
105. I. A. Abu-Isa, H. E. Trexler, *Rubber Chem. Technol.* **58** (1985) 326.
106. A. Nersasian, *Elastomerics* **112** (1980), no. 10, 26.
107. Bulletin VT-100.1, "Handling Precautions for "Viton" and Related Chemicals", Du Pont, Wilmington, Del.
108. E. W. Bergstrom, *Elastomerics* **109** (1977) no. 2, 21.
109. Bulletin VT-515.1, Radiation Resistance of Viton, Du Pont, Wilmington, Del.
110. Bulletin VT-000.1 (R2), *A Capsule View of the A, B, and E Types of Viton,* Du Pont, Wilmington, Del.
111. Bulletin VT-210.1, *Selection Guide for Viton Fluoroelastomers,* Du Pont, Wilmington, Del.
112. Bulletin VT-250.GLT, *Viton GLT,* Du Pont, Wilmington, Del.
113. Bulletin VT-250.GF, *Viton GF,* Du Pont, Wilmington, Del.
114. Bulletin VT-440.1, *Effect of Oven Post-Cure Cycles on Vulcanizate Properties,* Du Pont, Wilmington, Del.
115. A. W. Fogiel et al., *Rubber Chem. Technol.* **49** (1976) 34.
116. J. D. MacLachlan, A. W. Fogiel, *Rubber Chem. Technol.* **49** (1976) 43.
117. W. W. Schmiegel, *Kautsch. Gummi Kunstst.* **31** (1978) 137.
118. W. W. Schmiegel, *Angew. Makromol. Chem.* **76/77** (1979) 39.
119. D. Apotheker et al., *Rubber Chem. Technol.* **55** (1982) 1004.
120. R. H. Wehrenberg II, *Mater. Eng.,* **94** (1981) no. 7, 44.
121. Du Pont, US 3 069 401, 1962 (G. A. Gallagher).
122. Du Pont, US 3 467 638, 1969; US 3 876 654, 1975 (D. B. Pattison).
123. Du Pont, US 3 546 186, 1971 (E. K. Gladding, R. Sullivan).
124. Du Pont, US 3 114 778, 1963 (A. L. Barney, W. Honsberg).
125. A. L. Barney et al., *J. Polym. Sci., Polym. Chem. Ed.* **8** (1970) 1091.
126. G. H. Kalb et al., *"Adv. Chem. Ser.",* **129** (1973) 13.
127. A. L. Barney et al., *Rubber Chem. Technol.* **44** (1971) 660.
128. A. L. Logothetis, International Rubber Conf., Kyoto, Japan (1985) 73.
129. A. L. Logothetis, Centenary of the Discovery of Fluorine (Symposium), Paris, France, (1986) 226.
130. A. L. Logothetis, Japan – U.S. Polymer Symposium, Kyoto, Japan, (1985) 121.
131. Du Pont, US 3 467 635, 1969 (W. R. Brasen, C. S. Cleaver).
132. Y. Tabata et al., *J. Polym. Sci. Part A* **2** (1964) 2235.
133. G. Kojima, Y. Tabata, *J. Macromol. Sci., Chem.* **A 5** (1971) 1087.
134. G. Kojima et al., *Rubber Chem. Technol.* **50** (1977) 403.
135. G. Kojima, H. Wachi, *Rubber Chem. Technol.* **51** (1978) 940.
136. G. Kojima et al., *Rubber Chem. Technol.* **54** (1981) 779.
137. Du Pont, US 4 035 565, 1977; US 4 214 060, 1980 (D. Apotheker, P. J. Krusic).
138. D. E. Hull, *Elastomerics* **115** (1983) no. 10, 40.
139. D. E. Hull, *Elastomerics* **114** (1982) no. 7, 27.

Fluorosilicic Acid → Silicon Compounds, Inorganic

Fluorosulfuric Acid

Erhardt Tabel, Bayer AG, Leverkusen, Federal Republic of Germany

Eberhard Zirngiebl, Bayer AG, Leverkusen, Federal Republic of Germany

Joachim Maas, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Preparation 1
3. Analysis 3
4. Safety Measures and Regulations. 3
5. Transport, Storage, and Filling. 4
6. References 4

1. Introduction

Fluorosulfuric acid [*7789-21-1*], HSO_3F, M_r 100.07, is a colorless to pale yellow liquid of stinging odor, which fumes readily in moist air. The halogen is more firmly bonded than that of chlorosulfuric acid so that fluorosulfuric acid is hydrolyzed more slowly than chlorosulfuric acid. Under favorable conditions, it is even possible to prepare salts in aqueous solution [3].

Fluorosulfuric acid was first mentioned in 1892 [4], [5]. The method of preparation, which is the one still used, from hydrogen fluoride and sulfur trioxide, was also described. Fluorinations using HSO_3F and catalysis of alkylations and cyclization by HSO_3F have been known for a long time. Industrial processes include the production of boron trifluoride [6] and the polymerization of tetrahydrofuran [7]. The chemical polishing of lead crystal glass with the acid was described in the 1960s [8].

The composition of commercial fluorosulfuric acid (wt %) is $\geq 99.0\,\%$ HSO_3F, $\leq 1.0\,\%$ H_2SO_4, $\leq 0.6\,\%$ free SO_3, $\leq 0.1\,\%$ SO_2, and $\leq 0.003\,\%$ Fe. The physical properties are listed in Table 1.

Table 1. Physical properties of fluorosulfuric acid

Melting point *mp*, °C	− 88.98
Boiling point *bp*, °C	
at 101.3 kPa	162.7
16.0 kPa	110
2.5 kPa	77
Density ϱ, g/cm^3	
at 18 °C	1.740
at 25 °C	1.725
Vapor pressure p at 25 °C, Pa	330
Viscosity η at 25 °C, mPa s	1.56
Heat of formation ΔH_f (liquid), kJ/mol	792.45

The salts and derivatives of fluorosulfuric acid have had no industrial importance.

2. Preparation

Hydrogen fluoride and sulfur trioxide are introduced, with stirring, into fluorosulfuric acid, which serves as a solvent for the reaction. The fluorosulfuric acid is pumped through a cooler to keep the temperature in the reactor constant at $< 100\,°C$ (Fig. 1).

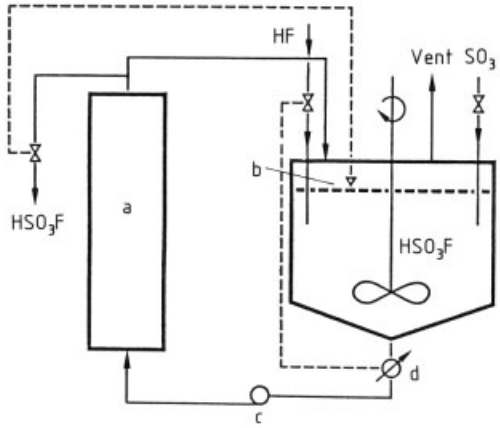

Figure 1. Preparation of HSO_3F
a) Cooler; b) Level control; c) Pump; d) Conductometer

Uses. In *inorganic chemistry,* fluorosulfuric acid is suitable for fluorination reactions, which proceed more gently than fluorination with hydrofluoric acid. Examples:

$SiO_2 \longrightarrow SiF_4$ [9]
$H_3BO_3 \longrightarrow BF_3$ [6]

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_431

$NO_x \longrightarrow FSO_3NO_x$
$As_2O_3 \longrightarrow AsF_3$ [10]
$KClO_4 \longrightarrow ClO_3F$ [11]

The last product, ClO_3F, is a reagent for the fluorination of organic compounds.

Fluorosulfuric acid is used to restore baths for chemically polishing lead glass to their initial strength. The polishing baths, as a rule, consist of 60 – 70 wt % H_2SO_4, 2 – 6 wt % HF, with the remainder water. During the polishing process, the glass surface is partially dissolved, which can be represented by the following equations:

$$PbO + H_2SO_4 \longrightarrow PbSO_4 + H_2O$$
$$SiO_2 + 6\,HF \longrightarrow H_2SiF_6 + 2\,H_2O$$

The acids consumed are replaced by fluorosulfuric acid, which is rapidly hydrolyzed in the strong acid, consuming some of the water formed during glass polishing:

$$HSO_3F + H_2O \longrightarrow H_2SO_4 + HF$$

In this way, the fluorosulfuric acid covers ca. 28 % of the total hydrofluoric acid requirement. When fluorosulfuric acid is added, 1.5 – 2 kg of waste acid is produced per kilogram of glass removed, but if a mixture of 96 % sulfuric acid and hydrofluoric acid are added to regenerate the polishing acid, 2.5 – 4.5 kg of waste acid is produced per kilogram of glass removed [12].

In organic chemistry, fluorosulfuric acid can be used as a catalyst for the alkylation of olefins, polymerizations, and isomerizations. The reactive alkyl ester of fluorosulfuric acid is evidently responsible for the catalytic activity [13]. Examples of alkylations:

Alkylation of olefins with ca. 13 % fluorosulfuric acid as catalyst [14]

Alkylation of isoparaffins with an olefin [15]

Formation of high-octane fuels from low-octane hydrocarbons by reaction with isobutane in the presence of fluorosulfuric acid and SbF_5 (magic acid) [16]

Formation of alkylated aromatics from aromatics (e.g., benzene) and ethylene with BF_3 and fluorosulfuric acid at ca. 10 kPa and room temperature [17]

Alkylation of phenol to *o*- and *p-tert*-butylphenol [18]

Examples of polymerization and isomerization reactions:

Polymerization of tetrahydrofuran [7]
Polymerization of ethylene with fluorosulfuric acid and BF_3 as catalyst to give long-chain unsaturated hydrocarbons [19]
Polymerization of trioxane to polyoxymethylene with fluorosulfuric acid and SbF_5 (magic acid) [20]
Isomerization of C_7 hydrocarbons to branched products [21]
Isomerization of pentene to isopentene [22]
Isomerization of methylpentene with sulfuric acid and fluorosulfuric acid to give isomethylpentene [23]

In the preparation of naphthol from naphthalene by catalytic hydroxylation with hydrogen peroxide, α-naphthol is formed selectively in classical acid systems, such as hydrogen fluoride, whereas the use of fluorosulfuric acid and SbF_5 as the catalyst produces β-naphthol with an isomeric purity between 92 and 98 % [24].

Fluorosulfuric acid has been proposed to remove hydrogen fluoride from elemental fluorine [25]. The customary purification with NaF, which forms slightly soluble $NaHF_2$, has the disadvantage that $NaHF_2$ is difficult to decompose. In contrast, hydrogen fluoride absorbed in fluorosulfuric acid can simply be distilled off.

3. Analysis

The principal impurities in fluorosulfuric acid can be determined by the following methods [2]:

1)	HF	A known excess of SO_3 is added. The free SO_3 is then back-titrated with water, with conductometric determination of the end point.
2)	SO_3	Direct titration with water, as in Method 1. Free SO_3 can be detected by the blue color upon addition of sulfur.
3)	H_2SO_4	The specific electrical conductivity at the end point of the titration in Method 1 or 2 is a function of the resulting sulfuric acid content of the fluorosulfuric acid. This allows determination of the sulfuric acid content if the quantity of sulfuric acid formed during titration is subtracted from the total.

Fluorosulfuric acid can be identified by infrared spectrometry [2], [26].

4. Safety Measures and Regulations.

Fluorosulfuric acid and its vapors are corrosive to eyes, mucus membranes, and skin. Reaction with water (humid air, perspiration, etc.) produces sulfuric and hydrofluoric acids by hydrolysis. As a result, contact with fluorosulfuric acid, especially if the eyes are involved, can result in injury from hydrofluoric acid, with all the accompanying complications: hydrofluoric acid also penetrates the skin, destroys the underlying tissues, and causes acute disruption of metabolism by binding Ca^{2+} and Mg^{2+} and inhibiting essential enzymes (Toxicology in → Fluorine Compounds, Inorganic).

Therefore, great care should always be taken in the handling of fluorosulfuric acid, and the safety regulations should always be followed exactly. In the Federal Republic of Germany, for example, the general rules of the Berufsgenossenschaft der chemischen Industrie, an association like the U.S. OSHA, to avoid accidents (Data Sheet M005) [27] should be observed.

In case of exposure to fluorosulfuric acid, all parts of the body that were exposed should be washed thoroughly with water and treated with a calcium gluconate solution (> 1 wt %), and the patient brought as quickly as possibly to a physician. (The physician should be informed and provided with Data Sheet M005.)

If fluorosulfuric acid is swallowed, milk of magnesia (20 wt % MgO in water) or calcium gluconate solution should be administered immediately. In no case should vomiting be induced.

Plant employees should wear appropriate protective clothing when working with fluorosulfuric acid, e.g.,

Breathing apparatus	air filter (type ABEK) if the MAK is exceeded
Gloves	fabrication out of Viton
Eye protection	protective glasses, face shield, possibly a protective helmet of Viton

If these devices are not adequate, then a complete protective Viton suit, providing protection for the head, face, and neck and including jacket and pants or overalls, and PVC boots must be worn.

Upon contact with organic material, in extreme cases, fluorosulfuric acid can induce ignition. Any spilled fluorosulfuric acid should be diluted with a large amount of water (the reaction with water is violent, and fogs can form). If spilled fluorosulfuric acid is foggy, surround the acid puddles with barriers of dry sand and cover with Diakon LG 156 Clear 001 (ICI) or Sedipur NF 102 (BASF) to avoid fogs. Small amounts of fluorosulfuric acid are best destroyed by mixing with tenfold the amount of cold, concentrated (96 %, better 98 %) sulfuric acid. This mixture is then treated slowly with ice until it is sufficiently dilute, and is then neutralized with lime.

In the Federal Republic of Germany, with respect to identification and handling, fluorosulfuric acid is subject to the *Decree on Hazardous Substances* of October 26, 1993.

The procedures necessary for the safe handling of fluorosulfuric acid at the Bayer works at Leverkusen, Federal Republic of Germany, have led the German chemical industry to expand the identification prescribed in the *Decree on Hazardous Substances* of October 26, 1993 by the symbols in Table 2.

Table 2. R and S phrases for fluorosulfuric acid

Symbol	Meaning
R 20	injurious to health when inhaled
S 26	upon contact with the eyes, flush thoroughly with water and consult a physician
S 45	if feeling unwell, obtain medical advice (if possible, showing container label)

The following safety regulations apply in the USA: Data Sheets SD20 and SD25 [28].

5. Transport, Storage, and Filling.

Samples of fluorosulfuric acid are supplied in Teflon or aluminum bottles. When opening the bottles, the following safety precautions should be taken: Viton gloves, safety glasses, face shield, and if necessary, appropriate clothing.

The following type of drum is permitted: air-tight closed steel or aluminum rolling hoop drums up to 220 L in capacity. Self-supporting plastic drums are not permitted. All drums must be protected against the sun and, above all, against heat. They should be opened once a week to dissipate any pressure that may have developed.

Fluorosulfuric acid may be shipped in road tankers, tank wagons, or tank containers. The tanks must be designed for a pressure of at least 1 kPa (= 10 bar) and tested at a pressure of at least 0.4 kPa (= 4 bar). For tanks in which fluorosulfuric acid is stored or transported a bottom outlet is not permitted. During unloading, in the Federal Republic of Germany, for example, Data Sheet T015 [29] must be complied with. In the transportation and loading data sheets fluorosulfuric acid is often called fluorosulfonic acid.

The following regulations apply for transport (as of 1995):

IMDG (international sea transport)	Class 8 UN No. 1777, Packing Group I
ADNR (European waterways)	Class 8 item 8 A
IATA-DGR (international air transport)	Class 8 UN No. 1777, Packing group I
RID/ADR (European railway and roads)	Class 8 item 8 A

6. References

General References

1. *Houben-Weyl.*
2. J. H. Karchmer (ed.): *The Analytical Chemistry of Sulfur and Its Compounds. Chemical Analysis,* vol. 29, part 1, Wiley-Interscience, New York – London – Sydney – Toronto 1970, pp. 264 – 273.

Specific References

3. W. Traube, J. Hoerenz, F. Wunderlich, *Ber. Dtsch. Chem. Ges.* **52** (1919) 1272 – 1284.
4. T. E. Thorpe, W. Kirman, *J. Am. Chem. Soc.* **61** (1892) 921 – 924.
5. A. A. Woolf, *J. Inorg. Nucl. Chem.* **14** (1960) 21 – 25.
6. General Chemical Co., US 2 416 133, 1944 (De Walt Secrist Young, J. H. Pearson).
7. Du Pont, US 4 115 408, 1977 (M. C. Baker).
8. E. Sälzle, DE 1 185 780, 1962.
9. National Smelting Co., CA 448 662, 1948. Standard Oil, GB 755 692, 1954 (A. J. Edwards).
10. A. Engelbrecht, A. Aignesberger, E. Hayek, *Monatsh. Chem.* **86** (1955) 470 – 473.
11. A. A. Woolf, *J. Inorg. Nucl. Chem.* **3** (1956) 250.
12. K. Döring, *Sprechsaal* **116** (1983) no. 5, 366.
13. P. J. Stang, M. Hanack, L. R. Subramanian, *Synthesis* 1982, no. 2, 85 – 126.
14. Standard Oil, GB 537 589, 1940.
15. Texaco, US 4 008 178, 1975 (J. W. Brockington).
16. Standard Oil, US 4 144 282, 1975 (D. A. McCaulay).
17. Universal Oil Products, US 2 428 279, 1944 (V. N. Ipatieff, C. B. Linn).
18. Beck, Koller & Corp., GB 640 485, 1946.
19. Universal Oil Products, US 2 421 946, 1945 (V. N. Ipatieff, C. B. Linn).
20. Du Pont, US 2 795 571, 1953 (A. K. Schneider).
21. Esso, US 2 956 095, 1958 (D. Schwartz).
22. Standard Oil, GB 551 961, 1941.
23. Standard Oil, US 2 564 080, 1949 (A. K. Roebuck, B. L. Evering).
24. PCUK-Produits Chimiques Ugine Kuhlmann, DE-OS 3 102 305, 1981.
25. ICI, GB 824 427, 1957 (H. R. Leech, W. H. Wilson).
26. R. Savoic, P. A. Guigere, *Can. J. Chem.* **42** (1964) 277 – 287.
27. *Fluorwasserstoff (Flußsäure)* und anorganische Fluoride [Hydrogen Fluoride (Hydrofluoric Acid) and Inorganic Fluorides], Merkblatt der Berufsgenossenschaft der chemischen Industrie, no. M005, 8/94.
28. *Chemical Safety Data Sheets SD20 (1973) and SD25 (1970),* Manufacturing Chemist's Association, Washington, D.C. DOT Hazardous Materials Table 49CFR 172 101 and RCRA Status Hazardous Waste, Corrosive (D002) – 40CFR, part 261.33.1980.
29. *Befüllen und Entleeren von Eisenbahnkesselwagen* [*Loading and Unloading of Railroad Tank Cars*], Merkblatt der Berufsgenossenschaft der chemischen Industrie, no. T015, 7/90. M004 Irritating Materials, Corrosive Materials. M053 General Measures for the Protection of Workers Handling Dangerous Materials.

Foamed Plastics

Heinz Weber, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 and 3; Sections 4.1 and 4.4 – 4.15; Chaps. 6 and 7)

Isidoor De Grave, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 4.2 and Chap. 5)

Eckhard Röhrl, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 4.3)

1. Introduction 2
2. Production 3
2.1. Principles 3
2.2. Blowing Agents 3
2.3. Additives 4
3. Properties 5
4. Specific Foamed Plastics 6
4.1. Polyolefin (PO) Foams 6
4.1.1. Low-Density Polyolefin Foams ($25 – 250\,kg/m^3$) 7
4.1.2. High-Density Polyolefin Foams ($250 – 700\,kg/m^3$) 11
4.2. Polystyrene (PS) Foams 12
4.2.1. Expandable Polystyrene (EPS) Foams 12
4.2.2. Extruded Polystyrene (XPS) Foams . 14
4.2.3. Structural Polystyrene Foams 15
4.3. Poly(Vinyl Chloride) (PVC) Foams 16
4.3.1. Rigid Poly(Vinyl Chloride) Foams . . 16
4.3.2. Flexible Poly(Vinyl Chloride) Foams 17
4.4. Foams from Phenol–Formaldehyde (PF) Resins 18
4.5. Foams from Urea–Formaldehyde (UF) Resins 20
4.6. Foams from Melamine–Formaldehyde (MF) Resins 20
4.7. Silicone (SI) Foams 21
4.8. Epoxy (EP) Foams 22
4.9. Polyimide (PI) Foams 23
4.10. Polymethacrylimide (PMI) Foams . 24
4.11. Poly(Methyl Methacrylate) (PMMA) Foams 24
4.12. Polyamide (PA) Foams 25
4.13. Poly(Phenylene Oxide) (PPE) Foams 25
4.14. Polysulfone (PSU) Foams 25
4.15. Polycarbonate (PC) Foams 26
5. Testing 26
5.1. Mechanical Properties 26
5.1.1. Rigid Foams 26
5.1.2. Flexible Foams 27
5.2. Thermal Properties 28
5.3. Miscellaneous Properties 29
6. Economic Aspects 29
7. Toxicology and Environmental Aspects 29
8. References 30

Abbreviations used in this article:

ABS	acrylonitrile–butadiene–styrene
BRR	ball rebound resilience
EP	epoxy
EPS	expandable polystyrene
EVM	ethylene–vinyl acetate copolymer
HDPE	high-density polyethylene
ILD	indentation load deflection
LDPE	low-density polyethylene
LLDPE	linear low-density polyethylene
MF	melamine–formaldehyde
NF	Norme Française
PA	polyamide
PC	polycarbonate
PE	polyethylene
PF	phenol–formaldehyde
PI	polyimide
PMMA	poly(methyl methacrylate)
PMI	polymethacrylimide
PO	polyolefin
PP	polypropylene
PPE	poly(phenylene oxide)
PS	polystyrene
PSU	polysulfone
PUR	polyurethane
PVC	poly(vinyl chloride)
RIM	reaction injection molding
RTV	room-temperature vulcanization
SAN	styrene–acrylonitrile
SI	silicone
TFE	thermoplastic foam extrusion
TFI	thermoplastic foam injection molding
UF	urea–formaldehyde
UL	Underwriter's Laboratories
XPS	extruded polystyrene

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_435

1. Introduction

Foamed plastics are materials that have cells distributed throughout their entire mass. A cell is a small void formed during the manufacture of the foam; it is enclosed partly or completely by *cell walls* or *struts*. In contrast to an open cell, a closed cell is totally surrounded by its walls and, therefore, is not interconnected with other cells via the gas phase.

The apparent density of a foamed plastic depends on the fraction of the total volume occupied by the polymer phase that forms the skeleton; it is always lower than the density of the skeletal substance [12].

Structural foams consist of a light, closed-cell core that is sandwiched between solid integral skins. *Syntactic foamed plastics* are gas-filled polymers that are not, however, foamed. They contain completely closed cells that are formed by hollow spheres distributed in the polymer matrix (see page 22).

The constituents of the polymer skeleton can be produced by polymerization, polyaddition, or polycondensation.

History. The first foamed polymers were produced in the 1920s from rubber (Dunlop). These were followed in the 1930s by foamed plastics from urea resin (BASF), flexible poly(vinyl chloride) (I.G. Farben), and polyurethane (Bayer). Polyethylene foams (Du Pont), extruded polystyrene foams (Dow), rigid poly(vinyl chloride) foams (Lonza), and the first phenolic resin foam (Dynamit Nobel) were developed during World War II. Foams from polystyrene beads (BASF), epoxy resins (Shell), and silicones (Dow Corning) were announced in the late 1940s and early 1950s. In the 1960s, polyamide (BASF), polycarbonate (General Electric), polysulfone (Pechiney-Saint-Gobain, ICI), polyimide (Du Pont), and polymethacrylimide (Röhm) foams were described. Poly(phenylene oxide) (General Electric), and melamine resin (BASF) foams have been known since the 1970s.

2. Production (→ Plastics, Processing)

Production processes for individual foams are described in Chapter 4. Only a brief overview is given here.

2.1. Principles

The three key starting materials for a foam are polymer, blowing agent, and additives.

In the manufacture of *thermoplastic foams,* the starting material is generally an existing polymer. Foams are produced in both continuous and batchwise processes which frequently involve an extrusion step. Thus, boards or films as well as profiled and shaped parts with various apparent densities can be produced (e.g., Section 4.2.2).

Foamed thermoplastic beads are fused with steam to form a variety of shaped parts, including blocks (Sections 4.1.1 and 4.2.1).

In the manufacture of *thermoset foams,* the macromolecules are cross-linked during the foaming process. The foams are produced in various ways. The pourable charge is distributed on a conveyor belt where the mixture foams and hardens continuously (Sections 4.4 and 4.6). The resulting paste can also be poured into block molds (Section 4.4) or other cavities where the reaction is completed (Sections 4.7 and 4.8). Whipping (Section 4.5) and spraying (Sections 4.4 and 4.8) processes may also be used.

2.2. Blowing Agents

In almost all foamed plastics, the volume fraction of the gas phase (a function of density) is far greater than that of the skeletal substance. Thus, for example, at an apparent density of 10 kg/m^3, the gas phase accounts for ca. 99 % of the volume. The gas phase is produced by using a blowing agent.

Physical blowing agents may be compressed gases (air, nitrogen) or low-boiling hydrocarbon or chlorofluorocarbon (CFC) liquids, e.g., butane, pentane, trichlorofluoromethane (CFC 11). They do not leave undesirable decomposition products.

In the oldest method, originally developed for rubber, air is whipped mechanically into the latex [13]. The resulting froth is fixed by cross-linking. This process is commonly used for urea–formaldehyde resin foams (Section 4.5) and flexible poly(vinyl chloride) (PVC) foams (Section 4.3.2).

In an alternative method, compressed nitrogen is dissolved in a polymer melt. The cellular structure is then obtained by reducing the pressure. Examples are the batchwise production of cross-linked polyethylene foam (Section 4.1.1) and injection molding.

Incorporation of the blowing agent pentane [*109-66-0*] into polystyrene beads (Section 4.2.1) renders them heat-expandable.

Low-boiling chloro- and chlorofluorocarbons are incorporated into polyethylene (Section 4.1.1) or polystyrene melts (Section 4.2.2) in an extruder. Trichlorofluoromethane [*75-69-4*] (CFC 11) and 1,2-dichloro-11,22-tetrafluoroethane [*76-14-2*] (CFC 114) are used as blowing agents for polyethylene. Methyl chloride [*74-87-3*], ethyl chloride [*75-00-3*], and trichlorofluoromethane [*75-69-4*] (CFC 11) are used for expandable polystyrene. The blowing agents evaporate and the polymer expands. Cooling increases the viscosity and sets the skeleton of the foam.

Thermoset foams (Sections 4.4 and 4.6) are also produced with low-boiling liquid blowing agents which are activated by the heat of the cross-linking reaction or by externally supplied energy.

The choice of blowing agent is based on a knowledge of physical and chemical properties [14].

Chemical blowing agents are solids which decompose at elevated temperature with the formation of gas. For optimum action, they must be distributed very uniformly in the substrate. Their decomposition temperature must be suitable for the processing temperature of the polymer. Numerous organic, nitrogen-containing substances with decomposition temperatures of 100 – 250 °C [15], [16] or higher [17] are known (e.g., azo compounds, nitroso compounds, and hydrazides). The most frequently used compound is azodicarbonamide [*123-77-3*] which splits off nitrogen and carbon monoxide at about 200 °C. The decomposition temperature can be lowered by various additives such as zinc oxide [18].

The compounds, which contain polymers and blowing agents, are then subjected to extrusion processing.

If chemical blowing agents are used, thermoplastic foam injection molding (TFI) and thermoplastic foam extrusion (TFE) are usually employed to yield heavy, foamed plastics known as *structural* (*integral*) *foams* (Section 4.2.3). The injection mold (TFI) or the die (TFE) is cooled so that the foam collapses at the wall of the mold or die and forms a compact surface layer. The compacted surface has nearly the same density as the polymer matrix, whereas the internal layers are foamed.

Inorganic blowing agents, such as carbonates, are of minor importance. A mixture of sodium bicarbonate [*144-55-8*] and citric acid [*77-22-9*] serves as an additional, nucleating blowing agent in combination with pentane (Section 4.2.2). Sodium borohydride [*16940-66-2*], which splits off hydrogen in the presence of water and is adsorbed on fumed silica, has been reported to be an effective blowing agent [19].

Blowing agents can also be formed as reaction side products. For example, in certain polyurethane foams and in thermoset-modified PVC foams, carbon dioxide is formed from isocyanate and water (Section 4.3.1). Polyimide foam is blown by the alcohol (usually ethanol) formed in the imidization reaction (Section 4.9).

2.3. Additives

Nucleating agents are important additives in foam manufacture, especially in physically blown, thermoplastic foams. They are finely dispersed, often silicate-like solids but they may also be waxlike. They act as boiling chips in the polymer melt which is supersaturated with dissolved blowing gas.

Minute, nucleating gas bubbles, called hot spots, are also formed during the initial decomposition of chemical blowing agents [20]. Gaseous side products of the reaction act in the same manner.

Surfactants are required in thermosetting plastics to emulsify the liquid blowing agent. They reduce interfacial tension and thus facilitate nucleation. The surfactant also stabilizes the foam during expansion until the foam is set by the simultaneous hardening process.

Other additives include dyes, inert or reinforcing fillers and fireproofing agents. Antioxidants and UV stabilizers are sometimes needed.

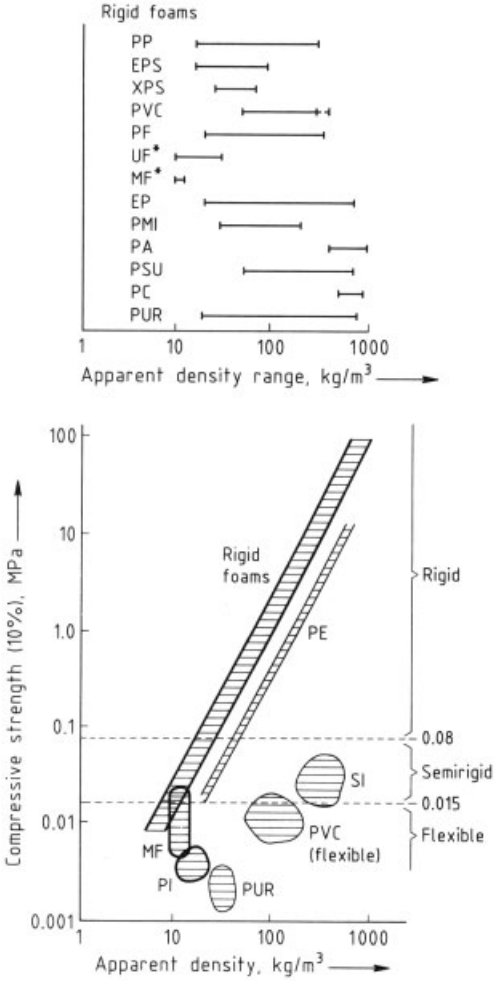

Figure 1. Compressive strength and apparent density of foamed plastics
∗ Also include flexible foams.
Abbreviations are defined in the abbreviation list at the beginning of this article.

3. Properties

The properties of foamed plastics are determined by their polymer substrate, apparent density, and morphology.

All three parameters, but especially apparent density, affect mechanical behavior. Resistance to elevated temperature and to chemical agents, as well as fire behavior, are primarily determined by the polymer. Morphology is responsible for the acoustical and thermal insulating properties. Both morphology and polymer substrate determine the water balance (water absorption, diffusion of water vapor).

Compressive Strength. One the basis of their compressive strengths determined in a compression test at 10 % deformation (DIN 53 421, ISO 844), foamed plastics are classified as *rigid* (> 0.08 MPa), *semirigid* (0.015 – 0.08 MPa) or *flexible* (< 0.015 MPa) [12]. The measured compressive strength depends primarily on the apparent density of the foamed plastic but also on the modulus of elasticity of the polymer [21].

Figure 1 provides an overview of the ranges of apparent density and compressive strength for individual foams. A linear relationship exists between the logarithms of the compressive strength and of the apparent density for all rigid plastic foams [15, p. 411], [22]. Polyethylene foams lie parallel to the scatter band; flexible plastic foams and some semirigid are outside it.

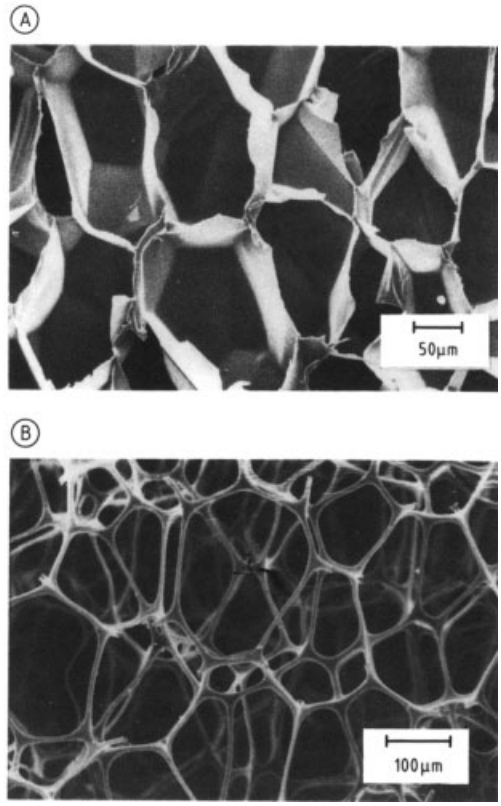

Figure 2. Scanning electron micrographs of (A) a closed-cell, extruded polystyrene foam and (B) an open-cell melamine–formaldehyde foam

Cell Structure. Rigid and semirigid plastic foams generally have *closed cells* which do not interconnect via the gas phase. They can, therefore, also contain gases other than air, e.g., fluorocarbons. Flexible foams, on the other hand, are usually *open-celled* and filled with air. Exceptions to this rule are, however, found. Figure 2 shows scanning electron micrographs of a closed-cell (A) and an open-cell (B) plastic foam.

The *cell size* of most plastic foams is 0.05 – 1 mm. Thermoset-modified PVC foams, and noncross-linked, extruded polyethylene and poly(methyl methacrylate) foams have cell sizes > 1 mm. Foams with smaller cells usually exhibit better mechanical behavior.

If the cells are oriented, e.g., stretched in the rising direction of the foam, the mechanical properties are no longer isotropic.

The thickness of the cell wall of closed-cell foams, the strut thickness of open-cell foams, and the cell size affect the radiant fraction of thermal conductivity. Each material has an optimum value [23], [24].

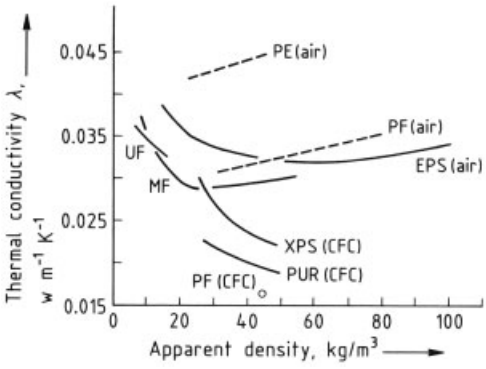

Figure 3. Thermal conductivity of foamed plastics as a function of apparent density
Values were measured at 10 °C according to DIN 52 612, ISO-DIS 8302, ASTM C 177. The gases used to fill the foams, air or a chlorofluorocarbon (CFC), are indicated in parentheses. Dashed lines indicate that values vary widely. EPS = expandable polystyrene; MF = melamine–formaldehyde resin; PE = polyethylene; PF = phenol–formaldehyde resin; PUR = polyurethane; UF = urea–formaldehyde resin; XPS = extruded polystyrene

Thermal Insulation. The thermal insulation of a foamed plastic depends primarily on the thermal conductivity of the filling gas, followed by the apparent density and then morphology. Foams filled with heavy chlorofluorocarbon gases, therefore, have better thermal insulation characteristics than air-filled foams. Figure 3 shows the density dependence of the thermal conductivities of several foams measured at 10 °C [25], [26].

Depending on prevailing conditions, open-cell products can absorb water and thus increase their thermal conductivity [27]. Since they are also open to diffusion, they must be provided with a barrier layer in some cases.

Sound Insulation. Open-cell foams are well suited for airborne sound insulation. Closed-cell products are useful for impact sound insulation, provided they have sufficient resilience [28]. Rigid foams are less suitable for acoustic applications.

Fire Behavior. Polymer-dependent properties (i.e., maximum long-term service temperature and oxygen index) are shown in Figure 4 for the individual foams discussed in Chapter 4. Thermoplastics differ from thermosets: when exposed to a flame, thermoplastic foams sinter above their softening range and melt, whereas thermosets form a carbon skeleton.

The oxygen index provides some indication of the flammability of a polymer. It measures the minimum oxygen concentration necessary to support combustion of the material [29]. Thus, higher values are better. The oxygen index of a foamed plastic may often be only slightly higher than that of the compact polymer, exceptions are urea–formaldehyde resins and poly(phenylene oxide).

Other important parameters for judging fire behavior are the smoke density and the toxicity of the combustion gases.

In compliance with the flammability requirements specified in guidelines for use in construction or transportation, some foams contain flame retardants. These are often halogen compounds, examples include chlorinated paraffins and hexabromocyclodecane for polyethylene and polystyrene, as well as boron trioxide for phenol—formaldehyde resins (cf. Chap. 5).

4. Specific Foamed Plastics

Polyurethane (PUR) foams are treated elsewhere (→ Polyurethanes). For a description of foamed

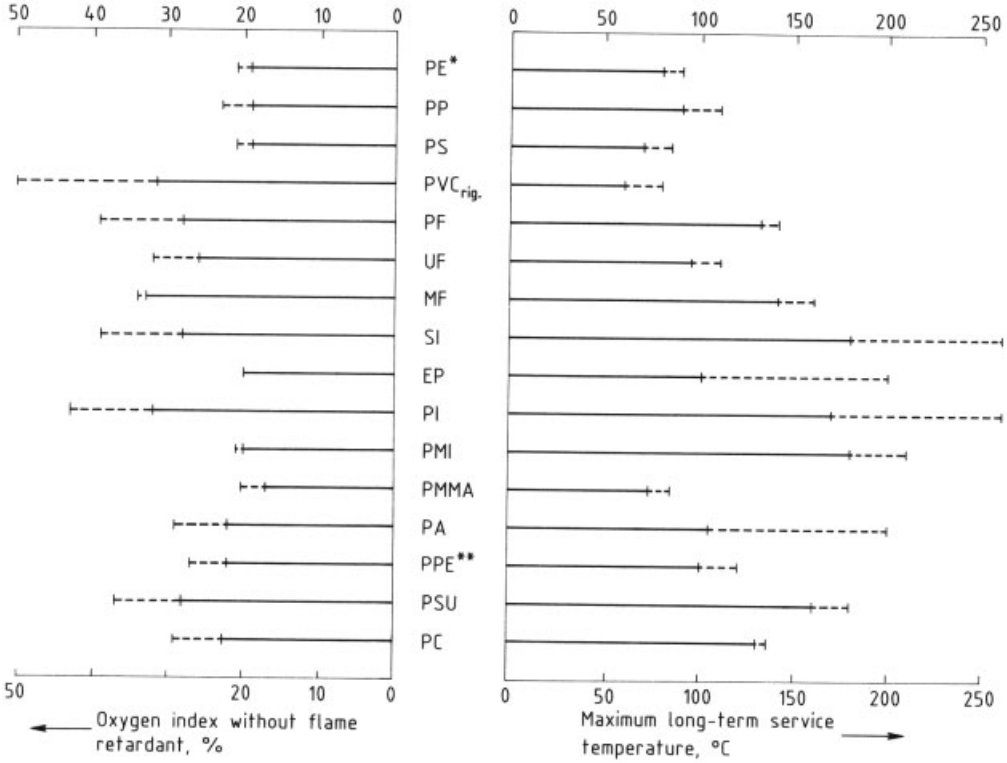

Figure 4. Oxygen index and maximum long-term service temperature of foamed plastics
The dashed lines indicate the range of variation (i.e., minimum and maximum values).
∗ Abbreviations are defined at the beginning of this article.
∗∗ Ratio of PPE : PS = 1 : 1.

PVC coatings, see → Leather Imitates. Trade names of individual foams are listed in Table 1.

4.1. Polyolefin (PO) Foams

Polyolefin (PO) foams are manufactured by many producers in a variety of processes and are classified as low density (25 – 250 kg/m^3) or high density (250 – 700 kg/m^3). The market has expanded relative to other foamed plastics, especially in Japan.

Most polyolefin foams are made from low-density polyethylene [*9002-88-43*] (LDPE). Copolymers such as the more flexible ethylene–vinyl acetate [*24937-78-8*] (EVM) and, more rarely, high-density polyethylene [*9002-88-4*] (HDPE) are also found. Finally, foams may be produced from linear low-density polyethylene (LLDPE), a copolymer of ethylene and other olefins (1-butene to 1-octene) which is synthesized with a Ziegler catalyst. Polypropylene [*9003-07-0*] (PP), which is more resistant to heat deformation, stiffer, and tougher than polyethylene, was first processed into foamed sheets in the 1960s. Today, a polypropylene copolymer with a few percent ethylene is also utilized. For a detailed description of these polyolefins, see → Polyolefins. Trade names of polyolefin foams are listed in Table 1.

Fire Behavior. Like their starting materials, polyolefin foams are flammable. The smoke density is low. The combustion gases contain only carbon dioxide and, in case of oxygen deficiency, carbon monoxide.

4.1.1. Low-Density Polyolefin Foams (25 – 250 kg/m^3)

Boards, sheets, or profiled shapes are manufactured by continuous processes; boards are also

produced by batch processes. In addition, beads can be made continuously or batchwise.

Table 1. Trade names of foamed and foamable plastics

Trade name and raw material	Producer
Polyolefin (PO)*	
Alveolit (PE)	Alveo (Switzerland)
Apro (PP)	Arco Chemical (USA)
Arpak (PE)	Arco Chemical (USA)
Eperan (PE)	Kanegafuchi Chemical Industry (Japan)
Ethafoam (PE)	Dow Chemical (USA)
Evazote (EVM)	Bakelite Xylonite Ltd. (UK)
Foam ace (PE)	Furukawa (Japan)
Frelen (PE)	Freudenberg (FRG)
Neopolen (PE)	BASF (FRG)
Opcel (PE)	Sentinel (France)
Plastazote (PE)	Bakelite Xylonite Ltd. (UK)
Softlon (PE)	Sekisui (Japan)
To-re-pepu (PE)	Toyo Rayon (Japan)
Trocellen (PE)	Dynamit Nobel (FRG)
Polystyrene (PS)	
Afcolene	Société Chimique des Charbonnages (France)
Arcel	Arco Chemical (USA)
Dylark	Arco Chemical (USA)
Dytherm	Arco Chemical (USA)
Eslen	Sekisui (Japan)
Extir	Montedison (Italy)
Kane Pearl	Kanegafuchi Chemical Industry (Japan)
Rigipore	British Petroleum (UK)
Roofmate	Dow Chemical (USA)
Styro-lite	3M (USA)
Styrocell	Shell Chemical (UK)
Styrodur	BASF (FRG)
Styrofoam	Dow Chemical (USA)
Styropor	BASF (FRG)
Vestypor	Hüls (FRG)
Poly(vinyl chloride) (PVC)	
Airex	Airex (Switzerland)
Conticell	Continental-Gummi (FRG)
Divinycell	Diab-Barracuda (Sweden)
Ensolite	Uniroyal (UK)
Foamcell	Pelf (Italy)
Klegecell	Kleber-Colombes (France)
Kömacell	Kömmerling (FRG)
Pliovic	Goodyear (USA)
PVC Foam	Shin-Etsu (Japan)
Rigicell	Goodrich (USA)
Solvic	Solvay (Belgium)
Vestolit	Hüls (FRG)
Vinoflex	BASF (FRG)
Phenol-formaldehyde (PF) resins	
Cellobond	British Petroleum (UK)
Dynapor	Dynamit Nobel (FRG)
Eurothane xtra	Recticel (Belgium)
Exeltherm xtra	Koppers (USA)

Continuous Process for Non-Cross-Linked Foam. Non-cross-linked LDPE foam is produced by applying the technique used for polystyrene (Styrofoam, see Section 4.2.2) which involves the gassing of a polymer melt with low-boiling halogenated hydrocarbons in an extruder. The blowing agent evaporates as it emerges from the nozzle; the melt then cools and its viscosity increases [30]. A strand of closed-cell foam (cell size 0.5 – 1.5 mm) is formed, with a thickness of 80 mm and an apparent density of 30 kg/m^3. Thin sheets of foam up to 5 mm thick are produced in the same manner or via the film blowing process (→ Films,). An example of such a film is Ethafoam produced by Dow Chemical.

Cross-Linked Foams. As shown in Figure 5, LDPE melts in a very narrow range [31]. Cross-linked LDPE exhibits a less precipitous drop in viscosity with increasing temperature. Thus, a broader temperature range for processing is obtained; in addition, the foamed plastic has smaller cells.

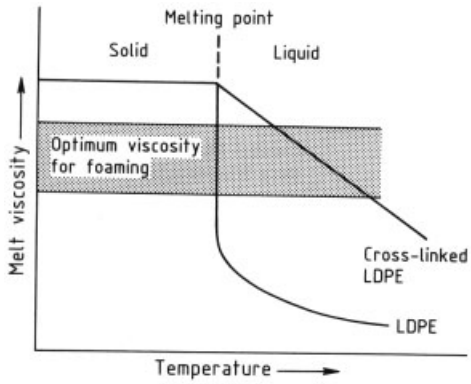

Figure 5. Melt viscosity of low-density polyethylene (LDPE) as a function of temperature

Cross-linking is accomplished either chemically (e.g., with dicumyl peroxide [*80-43-3*]) or with beams of electrons at doses up to 20 Mrad.

The gel content (degree of cross-linking) should be ca. 30 – 60 % [32]. Occasionally, divinylbenzene [*1321-74-0*] [33] or oximes [34] are added to the peroxide. Polyolefins grafted with silanes can be cross-linked catalytically in contact with water [35], [36]. For polypropylene, cross-linking agents with higher decomposition temperatures are used, e.g., polysulfone azides [37].

Batch Process for Molded, Cross-Linked Foam. In the in-mold, molded foam (batch) method, preformed polyethylene boards containing peroxide cross-linking agents are placed in a mold and exposed to nitrogen at ca. 38 MPa

and 180 °C. After cooling and careful pressure reduction, the nitrogen remains dissolved in the melt. When heated to 100 – 130 °C, the partially cross-linked board foams to yield a foamed plastic. Its cells are very fine (0.1 mm), and its density is 24 – 70 kg/m^3 [38], [39]. This process is used by Bakelite Xylonite Ltd. to manufacture Plastazote.

Table 1. (Continued)

Trade name and raw material	Producer
Fenomo	Isomo (Belgium)
Mosy	Compo (FRG)
Oasis	Smithers (USA)
Phenexpan	Société Chimique des Charbonnages (France)
Phenobil	Roland-Werke (FRG)
Phenocomp	Sekisui (Japan)
Phenolite	Bridgestone (Japan)
Pyrofoam	Union Carbide (USA)
Resiphen	Texaco (USA)
Urea–formaldehyde (UF) resins	
Aerolite	Ciba-Geigy (Switzerland)
Beetle UF Foam	British Industrial Plastics (UK)
Cascofoam	Borden Inc. (UK)
Hygromull	BASF (FRG)
Isoschaum	Schaumchemie (FRG)
Kobosan	Vorwerk (FRG)
Melamine–formaldehyde (MF) resins	
Basotect	BASF (FRG)
Silicones (SI)	
Cohrlastic	Connecticut Hard Rubber (USA)
Foamega	Bisco Prod. (USA)
Polyvoltac	Polyvoltac (UK)
Rau-Sik	Rehau (FRG)
Epoxy resins (EP)	
Araldit	Ciba-Geigy (Switzerland)
Scotchplay	3M (USA)
Polyimide (PI)	
Chem-Lon	Chem-Tronics (USA)
LMB 1907	Ciba-Geigy (Switzerland)
Polyimid 2080	Upjohn (USA)
Skybond	Monsanto (USA)
Solimide	International Harvester (USA)
Polymethacrylimide (PMI)	
Rohacell	Röhm (FRG)
Poly(methyl methacrylate) (PMMA)	
Isocryl	IMC Acrylguß (FRG)
Polyamide (PA)	
Maranyl	ICI (UK)
Rilsan	Société Nationale des Pétroles d'Aquitaine (France)
Poly(phenylene oxide) (PPE)	
Noryl	General Electric (USA)
Prevex	Borg-Warner (USA)
Polycarbonate (PC)	
Lexan	General Electric (USA)
Makrolon	Bayer (FRG)
Merlon	Mobay Chemical (USA)

∗ PE = polyethylene; PP = polypropylene; EVM = ethylene–vinyl acetate copolymer.

Instead of nitrogen, chemical blowing agents such as azodicarbonamide can also be used. The gel is removed from the mold rapidly while it is still hot and somewhat larger cells result.

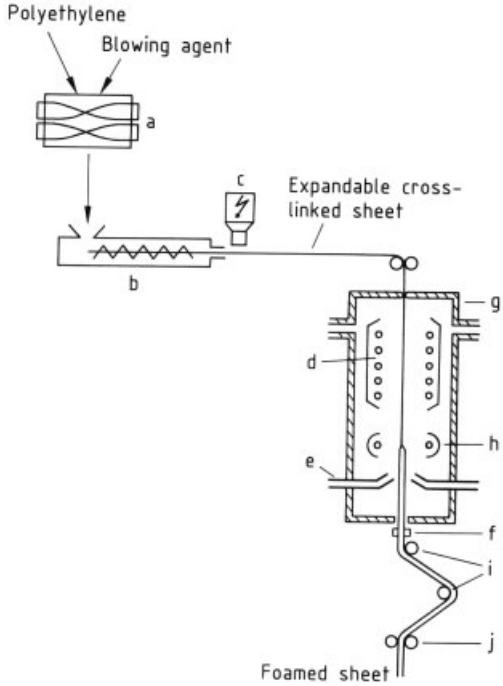

Figure 6. Production of radiation-cross-linked polyethylene by using a chemical blowing agent
a) Mixer; b) Extruder; c) Ionization irradiation unit; d) Preheater; e) Hot-air blower; f) Wrinkle-removing unit; g) Oven; h) Hot-air heater for foaming; i) Chill rolls; j) Take-up roll

Continuous Process for Cross-Linked Sheets. For the continuous manufacture of foam sheeting cross-linked with peroxides, the polyolefin (e.g., LDPE) is premixed carefully with up to 10 parts by weight of blowing agent (azodicarbonamide) and < 1 part by weight of cross-linking agent (dicumyl peroxide). The mixture is extruded to form a thin matrix which is then cross-linked by passing it through a heating tunnel at 180 – 230 °C. The sheet is finally foamed to a thickness of 4 – 20 mm and an apparent density of 25 – 250 kg/m^3. This process was developed by Furukawa and licensed to other companies (Bayer, Dynamit Nobel, Freudenberg) [32], [40], [41].

In a similar process, the matrix containing the blowing agent is cross-linked by irradiation with electrons. Subsequently, it is foamed in a

shaft by exposure to IR heaters and hot air at 200 – 250 °C (Fig. 6). Free expansion can also be carried out in a hot fluid bath. The depth of penetration of the electron beam limits the maximum thickness of the foamed sheet to 5 mm. Apparent densities are 30 – 200 kg/m^3. This process was developed by Sekisui and Alveo, it is used to produce Alveolit [32], [42], [43].

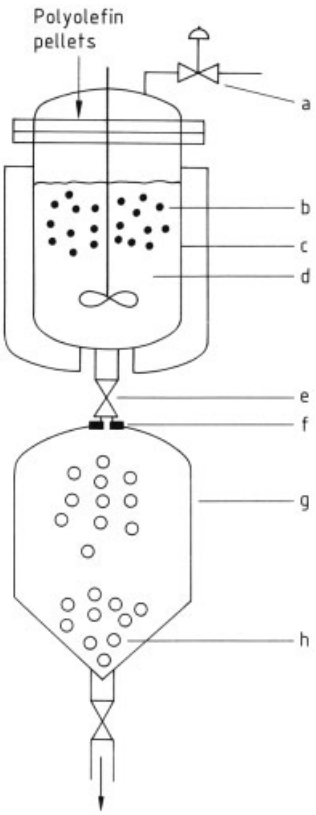

Figure 7. Production of expanded foam beads from polyolefin pellets
a) Pressure control valve; b) Polyolefin pellets; c) Pressure-resistant container; d) Water; e) Discharge valve; f) Orifice plate; g) Discharge tank; h) Expanded foam beads

Continuous Process Using Cross-Linked Beads. Strands of foam, produced by extruding LDPE with added butane, are comminuted to spherical beads. After the polyethylene has been cross-linked by an electron beam, the beads are fused continuously with hot air to form foam boards with thicknesses up to 120 mm and apparent densities in the range 28 – 45 kg/m^3. Cell size is about 0.5 mm. This process is used to manufacture Neopolen (BASF) [44], [45].

Batch Processes Using Beads. Pellets of weakly cross-linked LDPE, propylene–ethylene copolymers, or LLDPE in an aqueous dispersion are impregnated with low-boiling (halogenated) hydrocarbons as blowing agents. Impregnation is performed at a temperature near the softening range in the presence of a suspension stabilizer (e.g., tricalcium phosphate, aluminum oxide).

When the pressure is reduced, the particles expand into foam beads with bulk densities of 20 – 70 kg/m^3, depending on the amount of blowing agent and the temperature (Fig. 7) [46], [47]. Uniform cell structure and low bulk densities can be obtained, for example, by posttreatment with nitrogen under pressure [48], [49]. To ensure that the final molded parts have a satisfactory surface, the original pellets should be small (mass, 1 mg) and the beads should have a diameter of 3 – 6 mm. The beads can be processed into molded parts without further cross-linking. This process is used by Kanegafuchi Chemical Industries, Japan Styrene Paper, and Yuka Badische [44], [47].

Processing. Polyolefin foam boards can be cut or split readily with band saws or slicers. They can also be milled, cut thermally, or punched.

By heating and subsequent pressing (hot-melt lamination), cross-linked sheets can be fused to each other or to film, fabric, or sheet metal. They can also be cemented together with rubber-based contact adhesives or elastic melt adhesives. Thermoforming can be carried out effectively on commercial equipment [50–52].

Foam beads can be fused into parts of any desired shape, even boards, without waste by using steam or hot air; the process is similar to that used for expandable polystyrene beads (EPS) (see Section 4.2.1). Additional steps are required to completely fill the interstices between the spheres with foam:

1) In the Neopolen process for molded parts (Fig. 8), beads are placed in the mold cavity at elevated pressure and allowed to expand into the interstices by reducing the pressure. Only then are they exposed to steam, cooled, and removed from the mold [44], [53], [54].
2) The interstitial volume can also be reduced by a movable mold face which compresses the contents of the mold [55].

3) The internal pressure, and thus the expandability of the beads, can be increased immediately prior to mold foaming by using a compressed gas, e.g., nitrogen [56].

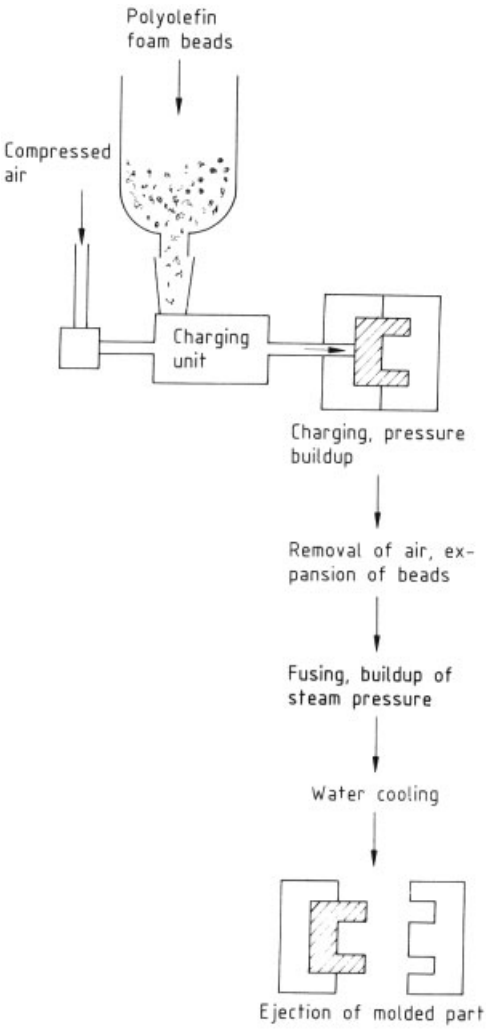

Figure 8. Manufacture of molded parts from polyolefin beads

Properties and Uses. Polyolefin foams are predominantly of the closed-cell type. They become semirigid as the crystallinity of the substrate and the apparent density decrease. Because of the air trapped in the cells, PO foams display excellent shock absorption properties which barely deteriorate under repeated impact. During prolonged loading, some of the air escapes from the cells. However, when the load is removed, air diffuses back into the cells and a slow recovery from deformation occurs, which depends on the apparent density and the load. Consequently, polyethylene foams are especially suitable for packaging impact-sensitive goods, e.g., electronic products. In packaging heavy goods, polypropylene foams can support higher loads and thus permit smaller support surfaces than the softer polyethylene foams. Polypropylene is particularly useful when greater dimensional stability under heat is required, e.g., in the automotive sector for items such as bumper cores [57].

The low moisture permeability of polyolefin foams combined with their low water absorption and relatively good weather resistance make them suitable for marine applications (e.g., ship fenders, life buoys, and life jackets) [54]. Protection against UV radiation is recommended for outdoor uses.

Because of their low thermal conductivity, which is only slightly greater than that of other foamed plastics, polyolefin foams are used as flexible pipe covers and insulating bags. More flexible types are used for impact soundproofing and resilient flooring in sports halls [50], [51], [58], [59].

4.1.2. High-Density Polyolefin Foams (250 – 700 kg/m^3)

Low-density polyethylene is foamed with chemical blowing agents (e.g., azodicarbonamide) to give apparent densities up to 450 kg/m^3. This material is used to insulate wires in telephone cables and coaxial cables; its dielectric constant is 3×10^{-4} [60].

Larger, shaped parts with apparent densities of 500 – 700 kg/m^3 can be produced from HDPE and polypropylene by the thermoplastic foam injection molding process. The integral skins of these structural foams gives a higher stiffness than is expected from their mean density [10, pp. 1 – 16].

Polypropylene is also foamed in extruders with chemical blowing agents to produce films and tapes in the density range of 400 – 700 kg/m^3. These products have high surface luster and are used in packaging [10, p. 15], [61].

4.2. Polystyrene (PS) Foams

(→ Polystyrene and Styrene Copolymers)

Polystyrene [*9003-53-6*] (PS) foamed plastics are important because of their physical properties, their broad spectrum of applications, and their favorable price–performance relationship. They can be produced by various processes and with different apparent densities; accordingly, they differ in properties and in possible uses. Trade names of foamed and expandable polystyrene are listed in Table 1.

4.2.1. Expandable Polystyrene (EPS) Foams

Production. The techniques used for the production of expandable polystyrene (EPS) beads and their processing to expanded polystyrene foams were developed at the end of the 1940s by BASF [62] who marketed the new raw material under the trade name Styropor. Due to licensing and the expiry of patents, other raw material manufacturers and trade names have appeared (see Table 1).

Expandable polystyrene is produced by suspension polymerization of styrene with addition of blowing agents; the resulting polymer beads are then sieved into fractions with various ranges of bead size. Depending on end use different coatings may be applied (→ Polymerization Processes).

In addition to pure polystyrene, foams are also made from polystyrene copolymers (e.g., with acrylonitrile or maleic anhydride). These foams possess special properties such as increased solvent or heat resistance. Incorporation of surface coating with organobromine compounds (e.g., hexabromocyclododecane [*25637-99-4*], pentabromophenyl allyl ether [*3555-11-1*]) act as flame retardants and result in flameproofed foams, e.g., Class B 1 building materials according to DIN 4102, M 1 according to NF P 92-505 (Norme Française), or MF-1 according to UL-94 (Underwriter's Laboratories). Blowing agents include low-boiling hydrocarbons, particularly pentane, but also butane, propane, and their mixtures; EPS beads contain 4 – 7 wt % of these substances.

Two other production methods for EPS consist of impregnating standard polystyrene with blowing agents in pressurized vessels and extrusion of EPS or standard polystyrene with added blowing agents. Cylindrical pellets are formed instead of beads.

Processing into Foam. Foam is obtained from EPS in three steps–prefoaming, intermediate storage, and final foaming (or molding)–as illustrated in Figure 9.

During *prefoaming,* the particles are heated and the polystyrene is thereby softened, the blowing agent in the polystyrene evaporates to form rapidly growing bubbles until either the heat supply is shut off or the expandability exhausted [63]. The bead diameter increases about threefold and the bead volume (bulk volume) about thirtyfold. The apparent density of the foamed product is determined by the bulk density obtained during prefoaming.

Intermediate storage is required prior to final foaming into blocks, boards, or molded parts to allow air to diffuse into the individual cells. Air is needed in subsequent molding: it acts as a supplementary blowing agent and also enables the soft cellular structure to withstand the external atmospheric pressure when the finished part is removed from the mold.

Final foaming is usually fully automated [64], [65]: perforated molds are completely filled with prefoamed beads and exposed to steam. The beads expand to fill the residual voids and are fused together. In modern installations, a vacuum is also applied. Some of the water condensed in the foam then evaporates and internal cooling reduces the pressure of the foamed plastic more rapidly, so that the parts can be quickly removed from the mold (short cycle times).

Boards are manufactured batchwise with block molds or with board molding machines; they may also be produced continuously on special moving-belt machines.

The glass transition temperature of polystyrene is reduced if pentane, which acts as a plasticizer, is used as blowing agent. The temperature range used during prefoaming and final foaming is then 85 – 120 °C. The most widely used heat-transfer medium is steam, which has a large heat capacity; it diffuses rapidly through the cell walls, and thus acts as a supplementary blowing agent. In some special cases, hot air is used instead of steam.

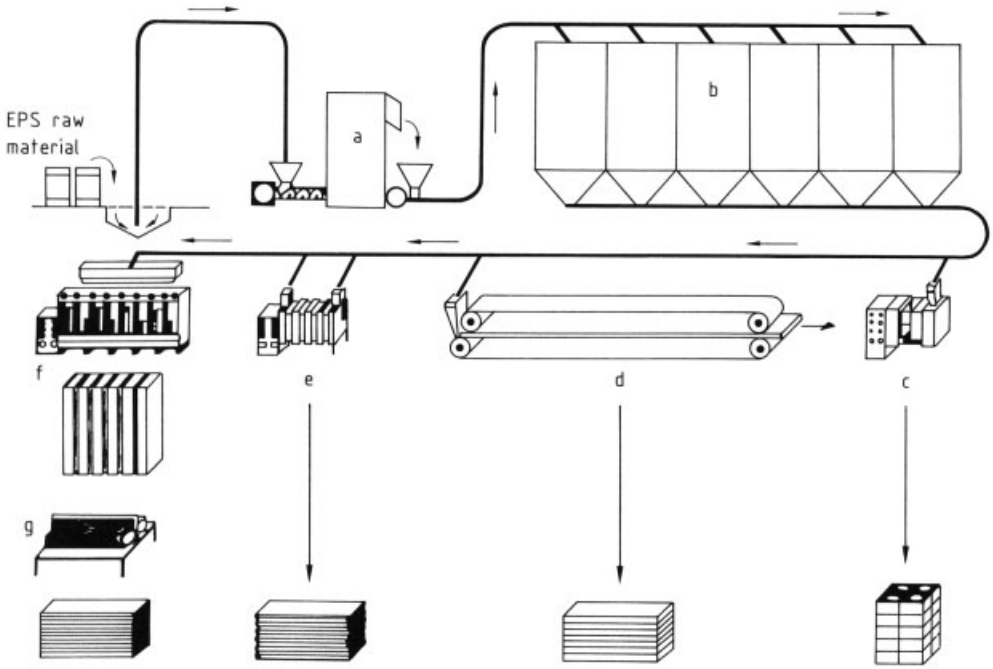

Figure 9. Processing of expandable polystyrene (EPS) foam into molded parts and insulation board
a) Prefoamer; b) Intermediate storage; c) Automatic molding machine; d) Continuous foaming unit; e) Automatic board molding machine; f) Block mold; g) Cutter

Foam boards for heat insulation are made primarily from foamed blocks 1 – 6 m^3 in volume. After stabilization for several hours to complete cooling and air uptake, the blocks are trimmed and cut into boards on cutters using electrically heated or oscillating wires. Posttreatment such as grooving or slotting is carried out on milling machines. Sheets of any desired length and 1 – 6 mm thick can be produced from special cylindrical blocks by using peeling machines with rotary blades. These peeled sheets are used under wallpaper and in packaging.

The resilience of boards for impact soundproofing is increased by briefly compressing the blocks prior to cutting or by rolling (or squeezing) the individual boards.

In a number of applications, the foamed plastic must be stored prior to use. This helps to remove internal moisture (steam condensate and residual cooling water) and to reduce postshrinkage of the foam.

Properties. Most of the physical properties of EPS foams depend strongly on the apparent density, the most important are summarized in Table 2. Boards for thermal insulation generally have densities between 15 and 30 kg/m^3, whereas the density of boards used for impact soundproofing is ca. 12 kg/m^3. The density of shape-molded parts is generally 20 – 30 kg/m^3, for special applications it may be much higher.

The chemical resistance of the foams corresponds largely to that of polystyrene.

Uses. Expandable polystyrene foams have many applications including thermal insulation, impact soundproofing [66], and formwork elements for concrete in the building industry; insulation of cold-storage depots and cooling cells, storage and transport tanks, pipes, etc.; and molded parts for packaging [67], promotion articles, toys, and sporting goods. It is also used for producing molded articles by lost-foam casting, a process that is similar to wax molding. A molded EPS form is embedded in sand and molten metal is poured into this composite. The foam shape evaporates and is replaced by metal.

Considerable amounts of nonfused, loose beads are also used as additives in bricks; (during firing they decompose to leave uniformly

Table 2. Physical properties of expandable polystyrene (EPS) foams as a function of apparent density

Property	Testing method	Apparent density, kg/m^3 *				
		15	20	30	60	100
Compressive stress at 10 % deformation, N/mm^2	DIN 53 421	0.06 – 0.10	0.10 – 0.14	0.18 – 0.25	0.40 – 0.65	0.8 – 1.3
Dimensional stability/ compressive creep, %	DIN 18 164	< 5	< 5	< 5		
Thermal conductivity at 10 °C, $W\,m^{-1}\,K^{-1}$	DIN 52 612	0.036 – 0.039	0.033 – 0.036	0.031 – 0.035	0.031 – 0.035	0.036 – 0.039
Specific heat, $kJ\,kg^{-1}\,K^{-1}$		1.21	1.21	1.21	1.21	1.21
Coefficient of linear thermal expansion, K^{-1}		7×10^{-5}	7×10^{-5}	7×10^{-5}	7×10^{-5}	7×10^{-5}
Water vapor transmission rate, $g\,m^{-2}\,h^{-1}$	DIN 53 429	1.7	1.6	0.8	0.3	0.15
Water absorption after 7 days, wt %	DIN 53 434	0.4 – 2	0.4 – 1.5	0.3 – 1.0	0.3 – 0.7	0.3 – 0.7

* Density determined according to DIN 53 420.

distributed pores). Further uses are as additives in light-weight concrete [68] and as renderings for thermal insulation or as drainage material.

Foamed EPS waste that accumulates during processing can be recycled; it is ground, mixed with preformed beads, and reused during molding. Ground EPS foam has also proved useful for loosening soil, for drainage, and as a filling material. These foams are inert and do not harm ground water, flora, or fauna.

4.2.2. Extruded Polystyrene (XPS) Foams

Foamed polystyrene boards and films are produced in large amounts by extrusion [69].

Production. In the original DOW process [70], polystyrene and a blowing agent were fed into a pressurized autoclave and melted. The molding batch was then cooled to about 100 °C; upon emerging into the atmosphere or into a mold chamber, it foamed and formed a uniform cell structure.

The melting of the polystyrene pellets and the gassing, homogenizing, and cooling of the melt were then performed in extruders [71]. The continuously emerging strand foams, is formed into a flat sheet, solidified in a cooling zone, and cut into boards.

Blowing agents are mainly halogenated hydrocarbons which increase the foam's thermal resistance (decrease thermal conductivity) because of their high molecular mass.

In addition to Styrofoam (Dow Chemical), similar XPS boards with apparent densities of ca. 25 – 60 kg/m^3 are also available, e.g. Styrodur (BASF).

Foam boards are generally dyed during extrusion and provided with flame retardants (e.g., hexabromocyclododecane).

Depending on the end use, the boards are produced with different apparent densities, thicknesses, surface textures (foam skin, milled), and edges (flat, step groove, or groove and tongue).

Properties. The most important physical properties of extruded polystyrene (XPS) foams at different densities are summarized in Table 3. The requirements for use as thermal insulation material in building applications are specified in ISO 4898. The chemical resistance corresponds to that of polystyrene.

Uses. Extruded polystyrene foam boards are expecially suitable for thermal insulation in special constructions where high compressive strength, good appearance in exposed locations, and very low water adsorption are required.

Extruded polystyrene films are produced exclusively by the film blowing process (→ Films). A tubular film is extruded from an annular spinneret. It is then expanded three- to sixfold by using a high internal gas pressure and is thus stretched transverse to the direction of drawoff. The film is then calibrated via a cooling mandrel, cut into one or more strips, flattened,

Table 3. Physical properties of extruded polystyrene (XPS) foams as a function of apparent density

Property	Testing method	Apparent density, kg/m^3 *			
		25	30	40	50
Compressive stress at 10 % deformation, N/mm^2	DIN 53 421	0.1 – 0.2	0.2 – 0.4	0.4 – 0.6	0.5 – 0.7
Compressive creep, %	DIN 18 164	< 5	< 5	< 5	< 5
Thermal conductivity at 10 °C, $W\,m^{-1}\,K^{-1}$	DIN 52 612	0.029 – 0.033	0.025 – 0.030	0.025 – 0.030	0.025 – 0.030
Specific heat, $kJ\,kg^{-1}\,K^{-1}$		1.45	1.45	1.45	1.45
Coefficient of linear thermal expansion, K^{-1}		7×10^{-5}	7×10^{-5}	7×10^{-5}	7×10^{-5}
Water absorption after 28 days, wt %	DIN 53 434	< 0.1	< 0.1	< 0.1	< 0.1

* Density determined according to DIN 53 420.

and either wound or processed immediately by thermoforming or lamination.

In the past films were produced in smaller, simple extruders from expandable polystyrene containing a blowing agent. However, the growing demand for such films required the use of large, direct-gassing installations that use standard polystyrene and an added blowing agent as in the extrusion of boards [72]. Low-boiling, preferably nonflammable, hydrocarbons are usually used as blowing agents. Chemical agents (e.g., citric acid plus sodium bicarbonate) and nucleating agents are also added to achieve a uniform cell structure.

Depending on the end use, film thicknesses range from several millimeters to a fraction of a millimeter and apparent densities from 100 kg/m^3 for thick films to 200 kg/m^3 for thin ones.

During thermoforming or lamination with cardboard and other layers, the films are often postfoamed; their apparent density is thus reduced and their thickness is increased. In this way, items such as egg cartons and packaging trays for fruit and meat are obtained with apparent densities of 60 kg/m^3 or less.

Polystyrene film laminated with cardboard is being tested as a replacement for corrugated cardboard where outstanding wet strength is required. Thin films are used for decorative purposes and as nonskid interlayers.

4.2.3. Structural Polystyrene Foams

Special starting materials are available for structural foams derived from polystyrene, its copolymers, and polymer blends. They already contain chemical blowing agents and are specially formulated for thermoplastic foam injection molding (TFI) and thermoplastic foam extrusion (TFE).

Products from butadiene-modified, impact-resistant standard polystyrene and from acrylonitrile–butadiene–styrene (ABS) polymers have the highest usage, followed by structural foams produced from standard polystyrene and styrene–acrylonitrile (SAN) polymers. For a detailed description of these polymers, see → Polystyrene and Styrene Copolymers.

Production. Physical blowing agents in the form of halogenated hydrocarbons are in principle suitable for structural polystyrene foams. However, they have not been widely used because more sophisticated processing technology is required than for chemical agents and control of foam structure is more difficult.

The following chemical blowing agents are commonly used: azo compounds (especially azodicarbonamide, sometimes modified), hydrazines (e.g., benzenesulfohydrazide), and sodium bicarbonate with citric acid.

Cell structure is affected not only by the blowing agents and the processing conditions, but also by the amount and type of nucleating additives such as talc or pigments.

Thermoplastic Foam Injection (TFI) Molding. Molded TFI parts can be produced on conventional injection molding machines with shut-off nozzles. However, low-pressure TFI machines are more economical. The mold is filled in two steps: (1) injection of the melt containing

the blowing agent and (2) postexpansion to fill the mold cavity. Thus, the clamping unit experiences only the expansion pressure which is about one-tenth of the injection pressure required for compact polystyrene. In this two-step process, removal of heat from the melt starts with evaporation of the blowing agent and surface streaks are created on the molded parts. Streaking can be prevented by various techniques, some of which are patented [8].

Thermoplastic Foam Extrusion (*TFE*). The formulations of raw materials used in TFE are similar to those used in TFI. However, since the temperature of the extrusion mass is generally lower, blowing agents with a lower decomposition temperature are used or catalysts are added to promote decomposition of the blowing agent. Although a number of patented extrusion processes exist (e.g., the Celuka process for very smooth surfaces, the Woodlite process for wood-like grain), conventional devices, profiled nozzles, and subsequent calibration are commonly employed [73]. Prior to use, the parts or profiles are frequently subjected to surface treatment such as spraying, printing, or lacquering.

Properties. The physical and chemical properties of structural polystyrene foams depend highly on the raw materials used. With their generally very high apparent densities and virtually compacted surface layers, the mechanical properties of structural foams are only slightly inferior to those of solid parts.

Uses [8]. The major end user of structural polystyrene foam is the furniture industry. The product is not used simply to imitate wood, but to construct structural parts for furniture of modern, functional design that exploits the properties of the raw material. The hi-fi and television industries are also important customers for housings and other TFI components. Significant quantities are also used in the building and transport industries, in the manufacture of sports equipment, and in the automotive sector.

4.3. Poly(Vinyl Chloride) (PVC) Foams

[→ Poly(Vinyl Chloride)]

Depending on the content of plasticizers, polymer processing aids, and thermosetting binders, poly(vinyl chloride) [*9002-86-2*] foams can be flexible, semirigid, or even rigid. Because of their range of properties, PVC foams have numerous uses. Frequently, they form part of structurally complex, multicomponent articles such as floor coverings or window frames.

The composition of PVC foams varies widely. Some consist mainly of PVC and contain only minor amounts of additives necessary for smooth processing and trouble-free application (stabilizers, lubricants, fillers, pigments, or blowing agents and their decomposition products). Others contain larger amounts of modifiers (plasticizers, cross-linking agents, processing aids, or impact modifiers.

Dry blends, granules and fluid or paste mixtures containing blowing agents are used to manufacture PVC foam. The most important processing technologies for dry blends and granules are molding, extrusion, calendering, and injection molding. Fluid or paste mixtures are usually processed by spread coating.

The PVC molding masses are foamed with both chemical and physical blowing agents (see Section 2.2) [74]. The foaming rate can be influenced by kickers or inhibitors. Poly(vinyl chloride) foams cannot be consistently classified because of the different manufacturing processes involved and their broad spectrum of properties and uses. However, division into rigid and flexible PVC foams has proved useful. Semirigid PVC foams are classified as rigid PVC foams although they contain plasticizers.

Trade names of PVC foams are listed in Table 1.

4.3.1. Rigid Poly(Vinyl Chloride) Foams

High mechanical strength, along with resistance to moisture, rot, and chemicals, is common to all rigid PVC foams. Thermal conductivity is $0.036-0.040\,\mathrm{W\,m^{-1}\,K^{-1}}$. These foams have a high oxygen index and their combustion gases contain carbon dioxide, carbon monoxide, and hydrogen chloride.

Most rigid PVC structural foams are produced by extrusion of PVC compound (dry blend or granules containing blowing agents) in the form of pipes, shaped parts, or boards [75–85]. Metal pipes or shapes can also be coated

with foamed PVC by extrusion [86], [87]. Furthermore, rigid PVC foam shapes can be reinforced with glass fibers. Very few rigid PVC foam components are produced by injection molding [88].

In the extrusion of molded rigid PVC parts, processing conditions are adjusted so that the blowing agent foams the molding mass only after it leaves the nozzle. Foamed parts have an outer skin whose thickness and density vary depending on how the melt foams and how rapidly the boundary phase is cooled. Complex shapes and boards of PVC foam can be produced by the so-called Celuka process [89] in which appropriate machine design allows foaming to proceed "inwards".

In the coextrusion process [90], [91], a compact PVC layer (which may be stabilized and pigmented as desired) is melted onto the foam during extrusion as an outer skin. Such profiles are preferred for outdoor use, e.g., as panels.

Rigid PVC foams range from predominantly closed-cell to open-cell, depending on the type and amount of blowing agent and the process used. These foams are used in the building and furniture industries and in the do-it-yourself sector for thermal insulation and as wood substitutes.

Thermoset-modified rigid PVC foams are generally manufactured by a two-step process [92–95]. In the first, so-called high-pressure stage, a paste containing PVC, plasticizer, isocyanate cross-linking agent, stabilizers, and a chemical blowing agent is heated in closed molds to about 180 °C under high pressure. The PVC and the plasticizer then form a gel, and the blowing agent decomposes to form tiny closed cells. After about 1 h, the molds are cooled while the pressure is maintained. In the second, so-called expansion stage, the molded articles containing the blowing agent are exposed to water vapor for several hours at 90 – 100 °C and reduced pressure. The material is still thermoplastic and thus softens; the expansion pressure of the occluded blowing agent increases the volume of the molded article. Some of the isocyanate additive reacts with water to form carbon dioxide which acts as an additional blowing agent; the remainder, together with a special auxiliary system, cross-links the PVC and significantly consolidates the foam structure.

The closed-cell foams produced are used primarily as core materials of high mechanical quality for composites in automotive, ship, and airplane construction. Important manufacturers are Kleber Colombes (France), Airex (Switzerland), and Conti-Gummiwerke (Federal Republic of Germany).

Rigid thermoplastic PVC foams can be produced by the same mold foaming process used for thermoset-modified foams. They are manufactured from polymer granules without using plasticizers, isocyanate, or cross-linking agents. These foams are used as core materials for composites and, because of their closed outer skin, are also used in marine buoyancy or flotation items.

4.3.2. Flexible Poly(Vinyl Chloride) Foams

Flexible PVC foams can be manufactured by various processes, either with or without a carrier.

Flexible PVC foams without support can be manufactured in the same way as rigid PVC foams, by a two-step process using dry blends, granules, or pastes containing blowing agents. The PVC foams obtained are predominantly closed cell; their flexibility and elasticity depend on the cell structure and the plasticizer content. Flexible PVC foams can also be produced continuously but with limited cross section by paste spreading or extrusion.

Flexible foams are used for sealing and insulation in the building trade, for heat insulation and buoyancy in clothing, and for impact protection in sports equipment.

Flexible PVC foams with support are produced by spreading pastes containing blowing agents on various supporting materials. Synthetic PVC leather, for example, consists of a textile carrier, a base layer of flexible PVC, and a layer of flexible PVC foam with a thin, compact surface. The flexible foam layer imparts a soft feel to synthetic PVC leather (→ Leather Imitates).

Modern, chemically embossed, color-printed floor and wall coverings have a similar structure. Inhibitors in the printing ink retard or prevent decomposition of the chemical blowing agent and thus create an embossed pattern that matches the printed pattern [96]. These materials are especially important as floor coverings (Congoleum process) because they are pleasant to walk on (→ Floor Coverings, Chap. 2.3.2.).

A spreadable whipped foam is obtained by intensively whipping gases (nitrogen, carbon dioxide, air) into PVC pastes. This foam is used mainly to back textile or rubber floor covering with an elastic layer that contains largely open pores after gelling [97–99].

Miscellaneous Flexible PVC Foams. Other flexible PVC foam items are produced from pastes containing blowing agents. They are used predominantly as sealing materials. Thus, for example, seals in crowned cork bottle caps used for beverages are produced by gelling and foaming a small amount of PVC paste applied directly to the cap. Similarly, flexible PVC foam seals for industrial components (e.g., air filters for automobiles) can be foamed in situ.

4.4. Foams from Phenol–Formaldehyde (PF) Resins (→ Phenolic Resins)

Production. Phenolic resin foams are produced by the acid-catalyzed, exothermic hardening of a phenol–formaldehyde resin [*9003-35-4*] condensed under alkaline conditions [100], [101] (→ Phenolic Resins). Evaporation of an added (halogenated) hydrocarbon blowing agent by the heat of the reaction produces foaming. Surfactants emulsify the blowing agent and stabilize the expanding foam. The foam is hardened with catalysts such as sulfuric acid, phosphoric acid, or arylsulfonic acids. Hydrochloric acid is no longer used [102]. For complete and thorough hardening, heat treatment may follow [103].

Hydrogen produced by reaction of powdered aluminum with an acid catalyst can also act as a blowing agent. This method is employed in the former Soviet Union where PF foams account for about one quarter of foamed plastic production [104].

Production is predominantly a batch process in which the foaming mixture is poured into block or other cavity molds. The use of spray guns has also been reported [105].

The well-known continuous production of laminated sheets [106] is used by Koppers [107], [108] and British Petroleum Chemicals [109], [110]. Closed-cell foams are presumably created. The starting materials are special resols with a high resin solids content. The blowing agent is a mixture of trichloromonofluoromethane [*75-69-4*] (CFC 11) and 1,1,2-trichloro-1,2,2-trifluoroethane [*76-13-1*] (CFC 113). Alkoxylated silicones [108] or glycerol ricinoleates [110] are used as surfactants.

For hardening, strong organic acids such as an anhydrous mixture of tolyl- and xylylsulfonic acids [108] or 50 % sulfuric acid [110] are used. Special mixing heads have been developed for intensive mixing of components [111].

The mixture is sandwiched between two surface layers of, for example, aluminum-coated cardboard; the foaming mass is shaped and hardened, while still plastic, in a heated, double-belt press (Fig. 10). The pressure exerted in the press is reported to have an important effect on foam quality. It reduces the open-cell content [107] and void formation underneath the upper layer [109].

The resulting laminates are between 30 and 55 mm thick and have apparent densities of 35 – 60 kg/m^3.

Properties. The hard, brittle, thermoset nature of cured PF resin is shared by the foam. It is, however, less pronounced in foams with smaller cells (cell size $< 80\,\mu$m) or with highly perforated cell walls [112].

All PF foams in existence at the end of the 1970s were open-cell types. Here, the cells are interconnected by holes in their walls which are several micrometers wide. Because the holes also allow communication with the surroundings, the cells are air-filled [113–115]. Like polyurethane and polystyrene foams, closed-cell PF foams can now be produced, whose cell walls are intact and free of pores [116], [117]. These cells can thus retain fluorohydrocarbon blowing agents. Because the thermal conductivity of these heavy gases is lower than that of air, the resulting foams are better insulators than air-filled foams (thermal conductivity $< 0.02\,\mathrm{W\,m^{-1}\,K^{-1}}$ and $0.033\,\mathrm{W\,m^{-1}\,K^{-1}}$, respectively).

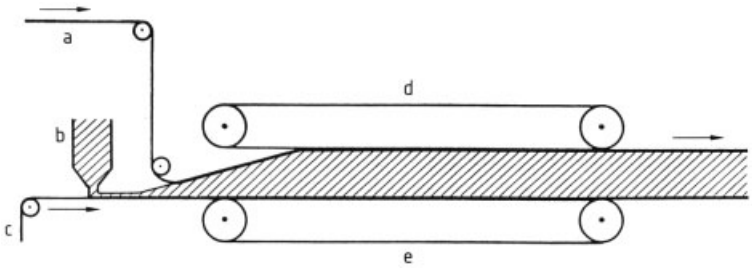

Figure 10. Production of laminated phenolic foam board
a) Upper surface layer; b) Dispenser for foamable phenolic resin; c) Lower surface layer; d) Upper conveyor belt; e) Lower conveyor belt

Water absorption naturally increases with the number of open cells. Light, largely open-cell PF foams (apparent density ca. 20 kg/m^3) may even be used as floral foam in flower arrangements [118].

If the foam is only partially closed-cell, expensive water-vapor barriers may be needed as, for example, in underground long-distance heating lines insulated with PF foam having an apparent density of 100 kg/m^3 [119]. Infiltration of water would not only impair heat insulation but also initiate corrosion in combination with the acidic hardening agents. Despite the generally high acid content (up to 20 wt %), corrosion problems are seldom encountered, especially with closed-cell products. Corrosion inhibitors such as calcium carbonate can be added.

Over short periods PF foams can resist temperatures up to 180 °C and, over longer periods, up to 130 °C. They are resistant to many solvents and attacked only by concentrated acids or bases.

Compared to other organic insulators, PF foam has a favorable fire behavior primarily because of the low smoke density of its combustion gases and its low flammability. Like other thermosets, the material does not melt. When ignited, a carbon skeleton is formed. With respect to the evolution of toxic gases (carbon monoxide), however, PF foams do not have any advantage over other foamed plastics [100], [101].

The afterglow of open-cell foams (punking) can be suppressed by the addition of flame retardants such as boron trioxide [*1303-86-2*] or aluminum hydroxide [*21645-51-2*]. Closed-cell PF foams hardened with sulfuric acid do not exhibit punking; when burnt, they "pop"; due to the noisy bursting of small particles of material [120].

Uses. In the building industry, for example, PF foam boards covered with glass-fiber webs are used as thermal insulation for flat roofs. They can be processed with hot bitumen because of their high heat resistance. When reinforced with an impregnated paper honeycomb, PF foam serves as the core of structural sandwich elements for walls and roofing, which are capable of bearing static loads. Sheet metal, polyester–glass fiber, plywood, or plaster may be used as surface layers [120–122].

Hemispheres cut from blocks are used for pipe insulation. A pour-in-place process with special molds has also been described [119].

In the former Soviet Union, PF foam is preferred for insulation in transport (motor vehicles, railroad cars, ships) as well as in aeronautic and space applications [104].

Special PF Foams. Inorganic additives (fillers) are added to PF foams to improve their mechanical strength and fire behavior. Depending on the type and amount of filler, *filled PF foams* with apparent densities of 100 – 300 kg/m^3 are obtained. Fillers include expanded silicates such as perlite [*7344-61-8*] [123]. Products that contain phenolic and furan resins are highly filled with materials such as aluminum hydroxide and are used as noncombustible building materials [124].

Phenolic foam with an apparent density of 20 – 30 kg/m^3, which is filled with polystyrene foam beads, can be obtained by dielectric heating [125].

Hybrid foams with good mechanical properties and low combustibility can be produced from PF resins and polyisocyanates [126].

4.5. Foams from Urea–Formaldehyde (UF) Resins (→ Amino Resins, Chap. 7.1.)

Production. A surfactant solution containing a hardener (e.g., phosphoric acid) is foamed with compressed air and blended with an aqueous 30 – 40 % solution of urea–formaldehyde resin [*9011-05-6*]. The resin is distributed throughout the surfactant foam, and cross-linking is initiated by the acidic hardener. A light, whipped-cream-like foam is formed which can be shaped easily and used to fill any cavity. The foam hardens rapidly to a water-insoluble, rot-resistant, foamed plastic without creating expansion pressure [127], [128]. A hardening time of 15 – 40 s is generally selected, which can be controlled within certain limits by changing external variables (temperature, hardener concentration).

Newly formed foam contains ca. 70 – 75 wt % of water. During drying, cracks are formed in the lamellae and the foam becomes open-celled [129]. This is accompanied by a ca. 8 % decrease in volume which can be reduced to 3 % by additives (e.g., polyols) [128], [130], [131]. Any free formaldehyde remaining is lost along with the water. Problems caused by formaldehyde release can be avoided largely by using more urea or improved resin formulations [132–134].

Stationary and portable equipment is available for foaming. Insulating layers can thus be introduced in cavities that are not easily accessible (e.g., in double walls).

Properties. Apparent densities of UF foam range from 10 to 30 kg/m^3 depending on the application. It can support little or no mechanical load and is, therefore, not used as semifinished product.

Urea–formaldehyde foam can withstand continuous exposure to temperatures up to 100 °C; at higher temperature, gradual shrinkage occurs. This material decomposes above 200 °C [128].

The thermal conductivity of UF foam (ca. 0.035 W m^{-1} K^{-1}) is in the same range as that of other air-filled insulating foams (cf. Fig. 3).

The material has a relatively favorable oxygen index (cf. Fig. 4). Although UF foam does not meet the flammability requirements of DIN 4102, it performs better in other flammability tests [ASTM D 1692-59 T, BS 2782 (1985) 508 D] [128].

Uses. The apparent density of in situ UF foam used in buildings must be at least 11 kg/m^3 (DIN 18 159). The foam serves mainly as thermal insulation in wall cavities [129], [135] and in plumbing ducts [136].

In mining, cavities (e.g., gallery roofs) are filled with UF foam to prevent accumulation of dangerous firedamp [137], [138]. The foam is also employed in the construction of airtight dams to seal off sections of mines (e.g., abandoned shafts) [139].

In agriculture and horticulture, UF foams with apparent densities > 20 kg/m^3 and preferably in the form of flocks (e.g., Hygromull, BASF) are mixed into the soil to improve heat balance. Because of their open-cell structure, the foams also store water and dissolved nutrients.

Kindling is obtained by emulsifying liquid petroleum products with UF foam during manufacture [140].

Foamed plastic with an apparent density of ca. 30 kg/m^3, ground to a particle size of 0.1 – 6 mm and containing a surfactant and an antistatic agent, is used in carpet cleaning [141].

Urea–formaldehyde foam is suitable for cushioning impacts only if no more than one violent impact is anticipated (parachute deployment) [142]. For trade names of UF foams, see Table 1.

4.6. Foams from Melamine–Formaldehyde (MF) Resins (→ Amino Resins, Chap. 7.2.)

The rigid, brittle melamine–formaldehyde resin [*9003-08-1*] foams known until the late 1970s had no practical significance [143]. However, a flexible, elastic foam can now be produced from the same resin. This foam (Basotect, BASF) has a completely different morphology [144], [145]; it is open-celled and has slender, flexible struts (cf. Fig. 2B). The length : diameter ratio of the struts is crucial to its elastic behavior. For strut

diameters from 7 to 15 µm, the ratio must exceed 10.

Production. A surfactant is used to emulsify pentane with a melamine–formaldehyde precondensate (molar ratio 1 : 3). Hardening occurs upon addition of formic acid. Controlled expansion in a continuous foaming tunnel is initiated by heating [146], [147].

Pentane and water from the reaction mixture act as blowing agents [148]. After removal of water and pentane, the foam is perfused continuously with hot air at 220 – 250 °C, and any cleavable formaldehyde is completely eliminated [149]. Flexible foams result only if the solids content exceeds 70 % [146]. Brittle products are obtained if the solids content is lower.

Properties. Density ranges from ca. 10 to 12 kg/m^3. Morphology and, hence, mechanical properties can be modified by pressing, which slightly bends the originally straight struts. The material becomes more flexible, with a negligible increase in density. For example, the compressive stress at 40 % deformation drops from 1.1 to 0.8 MPa [150]. Extreme mechanical–thermal treatment produces a highly consolidated, weblike material [151].

Because of the large proportion of IR radiation scattered by the thin struts, the thermal conductivity of the foam is in the lower range of air-filled insulators (Fig. 3).

Melamine resin foam can withstand short exposure up to 250 °C and long exposure up to 150 °C. No solvents are known for the foam; it is attacked by concentrated acids and bases.

Water absorption and diffusion properties are similar to those described in Chapter 3 for open-cell foams. The open, extremely fine structure and the flexibility of the foam result in especially favorable acoustical properties; despite the low apparent density, these properties are similar to those of open-cell polyurethane foams and mineral fibers.

The material meets the requirements of flammability class B 1 building materials, specified in DIN 4102 ("schwer entflammbar," i.e., of low flammability), without an added flame retardant. According to the tunnel test of ASTM E 84-82 A, it is a class A material as defined in the Standard Building Codes. The smoke density of the combustion gases is low but does not reach the minimum values of PF foam. The material carbonizes during burning, without melting, and decomposes above 350 °C [145].

Uses. Thermal insulation, heat resistance, and ease of processing have opened up many applications for MF foams in industrial heat insulation (e.g., pipe insulation).

The material is used for sound insulation in the form of decorative ceiling and wall paneling, as well as for casings in automotive engines. Because of its low dynamic stiffness, MF foam is also employed for impact sound insulation. Laminates with plasterboard provide good soundproofing in addition to the desired thermal insulation.

4.7. Silicone (SI) Foams

Production. Silicone foams are produced from hot- or cold-vulcanized silicone rubber (→ Rubber, 5.Technology, Chap. 2.5.2.2., → Silicones).

Heat-activated vulcanization systems contain mostly high molecular mass polysiloxanes incorporating methyl as well as vinyl groups. Additional components are chemical, heat-activated blowing agents, peroxide cross-linking agents, and silicate fillers [152]. These "single"-component blends are stable at room temperature and begin to foam when vulcanized by the action of heat.

Generally, preforms of the mixture or previously formulated dry blends are foamed in compression molds or extruders at 150 – 200 °C. The particles of silicone foam are then annealed at ca. 200 °C to remove any decomposition products derived from the cross-linking agent and to complete residual cross-linking.

Room-Temperature Vulcanization (*RTV*) *Systems.* Pourable mixtures that can be vulcanized and foamed at room temperature are required for producing molded articles and for filling cavities or coating parts with silicone foam.

Two-component systems are used in which short-chain polysiloxanes containing Si – H groups are allowed to react with low-viscosity polysiloxanes containing Si – OH groups. The reaction is performed at room temperature in the presence of a condensation catalyst:

$$\equiv Si-H + \equiv Si-OH \xrightarrow[\text{Room temp.}]{\text{Catalyst}} \equiv Si-O-Si\equiv + H_2$$

The hydrogen evolved acts as a blowing agent, but additional blowing agents may be necessary. Organotin compounds are used as catalysts [153], [154]; platinum catalysts yield products with improved heat and fire resistance [155–157].

The degree of cross-linking and the morphology of the foam can be controlled by the ratio of silanol to silane groups, especially by their functionality. Foams that range from rigid to flexible and from largely open-cell to closed-cell can be produced this way [158].

The above reaction takes only a few minutes but can be retarded by using acetylene alcohols or rhodium compounds as catalyst. The system is then stable for several hours and can be reactivated with heat [159].

Foam systems that can be cross-linked with atmospheric moisture have also been described [160].

A froth produced from a silicone emulsion can be converted to a light, elastomeric silicone foam in a microwave oven; no additional blowing agent is required [161].

Properties. Silicone foams are produced with apparent densities of 60 – 300 kg/m^3, but highly filled products are heavier. The heat distortion temperature is at least 200 °C; the material generally decomposes above 350 °C although higher limits have been recorded [162]. Thermal conductivity is reported to be 0.035 – 0.07 W m^{-1} K^{-1}.

Silicone foams display largely constant cushioning properties over a wide temperature range (− 80 to + 200 °C), and compression set is low. These materials exhibit good dielectric properties; are resistant to hydrolysis, radiation, aging, and weathering; and are physiologically inert.

The materials have a high flame resistance, with a favorable oxygen index of about 30 and a low smoke density. Combustion gases contain only carbon monoxide and carbon dioxide.

Uses [157]. Examples of the uses of silicone foams are

- sealing (shaped profiles, strings, rings) including heavy-duty applications in nuclear power plants;
- filling cracks or cavities, and sealing cable inlets and outlets in construction work;
- thermal and electrical insulation of fragile components in aerospace applications, with simultaneous vibration damping;
- production of carbonizing heat shields used in space vehicles (filled foams) [163];
- cushioning in airplane seats [162]; and
- fitting artificial limbs and producing inert implants for medical use [164].

4.8. Epoxy (EP) Foams (→ Epoxy Resins)

Production. The starting components for epoxide foams are epoxide oligomers, cross-linking agents, blowing agents, surfactants, and possibly also fillers. Silicone–glycol copolymers are described as surfactants [165]. The preferred epoxide oligomers are reaction products of 2,2-bis(4-hydroxyphenyl)propane [*80-05-7*] (bisphenol A) and epichlorohydrin [*106-89-8*] (→ Epoxy Resins, Chap. 2.1.). Cross-linking agents include fast-acting aliphatic amines (e.g., diethylenetriamine [*111-40-0*]) or slower acting aromatic amines (e.g., 1,3-phenylenediamine [*108-45-2*]) that are effective only at elevated temperature. The reaction is highly exothermic and thus activates the blowing agent which is often a fluorinated hydrocarbon or a compound that splits off nitrogen (e.g., sulfonyl hydrazides). Precise temperature control during foaming is important. Too high a temperature causes cracking and deformation of the foam.

Foams with apparent densities of 20 – 40 kg/m^3 are produced from fast-acting systems by a two-component spray process. However, mold-pouring or injection processes are more common and can also be used with slow-acting systems that require external heat. The resulting foams have apparent densities of 60 – 300 kg/m^3 [166]. Trade names are listed in Table 1.

Syntactic foams (see also 1) [167] often consist of epoxy resins and, therefore, are discussed in this section. Their cells are formed by hollow spheres (microballoons) that are distributed in the polymer matrix as a filler. The hollow spheres are composed of glass or phenolic resin and have a diameter of ca. 100 µm and a wall thickness of ca. 2 µm. Larger "macrospheres" composed of epoxy resin and with a diameter of over 1 mm have also been described [168].

The hollow spheres are mixed with a low-viscosity epoxy resin in the same way as fillers. When the fraction of spheres is 40 vol % or less, mold pouring is used for further processing; extrusion or compression molding can be employed for fractions up to 67 vol %. A vacuum is applied for degassing. The products have apparent densities of 200 – 800 kg/m^3.

Phenolic resins, polyimides [169], silicones [170], and polyether sulfones are also used as matrix resins.

Properties and Uses. Epoxy foams are usually rigid and primarily closed-cell materials. They have favorable mechanical properties, absorb very little water, and are not hydrolyzable. They are resistant to many chemicals but not to ketones.

The good adhesion of EP foams to many materials makes them suitable as laminate cores. They also serve as edge seals for honeycombs made from materials such as the aramid Nomex that are used in airplane construction.

Because of their low dielectric constant (< 2), EP foams are used in protective housings for radar equipment (radomes). Materials with a higher dielectric constant are obtained by addition of metal flocks (e.g., aluminum) and are used as radar lenses.

Sensitive electronic components are protected by encapsulation in EP foam [165].

The heat distortion temperature is reported to be 100 – 200 °C. The thermal conductivity of EP foams containing heavy gases is ca. 0.02 W m^{-1} K^{-1}, whereas that of foams manufactured with other blowing agents exceeds 0.035 W m^{-1} K^{-1}, depending on the density. Therefore, these foams are also used as insulation for hot-water pipes.

Epoxy foams that do not contain fire retardants are flammable. The oxygen index is 20.

Syntactic foams are very rigid and able to withstand very high hydrostatic pressures. They are employed in laminates for underwater vehicles, including a shuttle for deep-sea mining to a depth of 6000 m. Other uses include flotation devices with a density of 0.5 g/cm^3 for steel pipes in offshore oil exploration [168], and ablation materials for space flight.

4.9. Polyimide (PI) Foams (→ Polyimides)

The increasing importance of polyimides as temperature-resistant materials has led to the development of PI foams. Trade names are listed in Table 1.

Production. In a typical polyimide synthesis, the hemiester (**2**) is prepared from 3,3′,4,4′-benzophenonetetracarboxylic acid dianhydride [*2421-28-5*] (**1**) and ethanol. The polyimide (**4**) is obtained by reacting **2** with a diamine, in this case 4,4′-diaminodiphenylmethane [*101-77-9*] (**3**) (and if necessary with 2,6-diaminopyridine [*141-86-6*]), in the presence of suitable surfactants.

The powder obtained after spray-drying the precursor solution is foamed into large blocks in a microwave oven. The water and alcohol liberated upon imidization act as blowing agents.

Flexible, open-cell foams are formed [171]. Morphology and especially thermal properties can be modified by using other diamines containing, for example, hexafluoropropyl groups [172], [173].

Rigid panels can be produced by impregnation of fibrous webs or honeycombs with precursor solution and subsequent foaming. Syntactic PI foams (cf. page 22) and rigid PI foams [174] are also known.

Properties. The range of apparent densities of flexible PI foams is 6 – 18 kg/m^3; the thermal conductivity is ca. 0.035 – 0.040 W m^{-1} K^{-1} at 10 °C.

The materials are resistant to solvents. They do not undergo hydrolysis and can be used between − 195 and + 260 °C. Their fire behavior is particularly favorable: they have a very low smoke density and a high oxygen index [175–177].

Uses. Lower density PI foam is used for acoustic and thermal insulation. Its low mass and fire behavior make it highly suitable for aerospace applications [178] where material cost is not a primary concern.

Examples of uses are as a heat and vibration protection for sensitive military and aerospace electronics equipment, a lightweight fire barrier

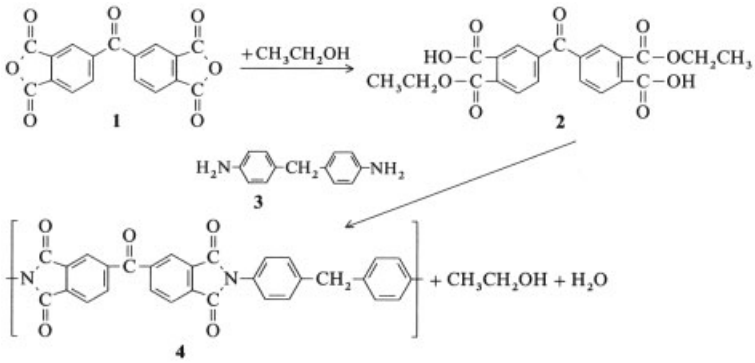

for airplane cabins, and cryogenic insulation for fuel tanks. The higher density foams can replace polyurethane foams in airplane seats because of their cushioning properties [179]. Uses in shipbuilding have also been reported.

4.10. Polymethacrylimide (PMI) Foams

(→ Polyimides)

Methacrylic acid [*79-41-4*] and methacrylonitrile [*126-98-7*] are subjected to radical copolymerization after the addition of blowing agents (e.g., formamide, or 1-propanol) that will come into play later. A transparent sheet is formed which is heated in a second step to 175 – 220 °C.

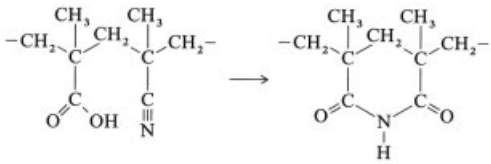

Copolymer of methacrylic acid and methacrylonitrile — Polymethacrylimide [*53112-45-7*]

Heating results in imide formation and the production of a rigid, closed-cell foam (Rohacell, produced by Röhm, Federal Republic of Germany) [180–182].

Properties. The relatively rigid material has apparent density of 30 – 200 kg/m^3. The heat distortion temperature (180 °C) can be increased to ca. 205 °C by annealing, which increases the degree of imidization. While warm, the foamed plastic can be reshaped and compacted. The thermal conductivity is 0.030 – 0.034 W m^{-1} K^{-1}.

Polymethacrylimide foam is resistant to organic solvents but attacked by alkaline media. The material burns, giving off little smoke; its oxygen index is 20.

Uses. Polymethacrylimide foam serves as a core material for lightweight laminates with high flexural rigidity which can be loaded dynamically. Because the foam is thermoplastic and can be compacted, complex forms can be produced with precision from a preform in a press. Therefore, the product frequently simplifies existing fabrication techniques. It is used in skis, tennis racquets, helicopter rotor blades, and load-bearing parts in airplanes.

Components with a large surface area (e.g., for railroad car construction) can be produced from sheet-molding compound sandwiches in short cycles at a molding pressure of ca. 1 MPa [183–185].

4.11. Poly(Methyl Methacrylate) (PMMA) Foams (→ Polymethacrylates)

Production. Monomeric methyl methacrylate [*80-62-6*] is subjected to radical polymerization in the form of a thin layer between glass plates. It contains trichlorofluoromethane or pentane as a blowing agent and butyl methacrylate as a plasticizer. A board is formed (typical thickness 4 mm) which is subsequently foamed at 110 °C [186].

Properties. The foamed sheets (typical thickness 16 mm) have very large closed cells (diameter 2 – 10 mm) and thus absorb very lit-

tle visible light. Material with an apparent density of 30 kg/m^3 has a thermal conductivity of 0.035 – 0.044 W m^{-1} K^{-1}. It is highly weather resistant.

The maximum long-term service temperature of the PMMA foam Isocryl (IMC Acrylguß) is 80 °C.

Uses. The transparent insulating material is used for double-glazing improvements to increase the thermal insulation of windows in large buildings and greenhouses. It is also used in ceiling construction [187].

4.12. Polyamide (PA) Foams

(→ Polyamides)

Production. A structural foam with only slightly lowered apparent core density (1.1 → 1.0 kg/m^3) is produced by the reaction injection molding (RIM) process. Caprolactam [*105-60-2*], for example, is injected into a mold with sodium lactamate [*2123-24-3*] as catalyst and an acyl lactam as activator [188–190]. Nitrogen serves as blowing agent [191]. Reaction injection molding permits good filling of the mold with low postshrinkage of the molded part. No sink marks are formed.

Foams with apparent densities down to ca. 400 kg/m^3 are made by conventional thermoplastic injection molding from polyamide (nylon 6) and a second batch containing blowing agents (e.g., azocarbonimide) [192], [193].

Polyamide foams with apparent densities below 200 kg/m^3 can also be obtained by alkaline lactam polymerization with low-boiling hydrocarbons as blowing agents [192–195]. However, they have no commercial significance.

Properties and Uses. Polyamide foams with higher apparent densities are used more widely than those with low densities, and their properties resemble more closely those of solid polyamides. Their toughness, rigidity, and abrasion resistance are worthy of mention. The high heat resistance, solvent resistance, and relatively favorable fire behavior are additional advantages. The inherent water absorption of polyamides can alter the mechanical and thermal behavior of the foam.

Some applications of structural PA foams are as housings for electrical apparatus, lamp sockets, furniture parts, pallets, containers, and bowling pins. Trade names are listed in Table 1.

4.13. Poly(Phenylene Oxide) (PPE) Foams [→ Poly(Phenylene Oxides)]

Blends of poly(phenylene oxide) [*25667-40-7*] with high-impact polystyrene are fabricated into structural foams by thermoplastic foam injection molding. Trade names are listed in Table 1.

The material is distinguished by good mechanical properties and a high heat resistance. The polystyrene content has an adverse effect on the favorable fire behavior of PPE (oxygen index = 30).

Applications in the automobile sector (crash pads) and in computer manufacture have been described [10, p. 16].

In the patent literature, the manufacture of extruded, lighter weight foams from PPE – PS blends, by using chlorinated hydrocarbons as blowing agents, is described [196].

4.14. Polysulfone (PSU) Foams

(→ Specialty Plastics)

Production. In the *dough-foam process,* particles of polysulfone [*25135-51-7*] are allowed to swell and form a gel in a solvent such as methylene chloride or a ketone. A cold-extruded dough sheet is comminuted to dough pellets, which can then be processed into foam-molded parts by using EPS technology (cf. Section 4.2.1). The gel can also be pressed directly to give a preform which foams when heated. Heating causes the solvent or blowing agent to escape and the low glass transition temperature of the gel increases to the high polysulfone value of ca. 200 °C [197–199].

Properties and Uses. The apparent densities of PSU foams are between 40 and 500 kg/m^3, depending on the process used. Unfilled products made by the dough-foam process, for example, have apparent densities of 390 kg/m^3; fiber-reinforced foams may have apparent densities as high as 700 kg/m^3.

Polysulfone foams display high heat resistance and a favorable fire behavior.

Possible applications include self-lubricating bearings, as well as special laboratory and medical products.

Highly filled syntactic polyether sulfone foams have been proposed for radomes [200].

4.15. Polycarbonate (PC) Foams

(→ Polycarbonates)

Production. Structural polycarbonate [*24936-68-3*] foams with apparent densities of 560 – 900 kg/m^3 are produced by low-pressure thermoplastic foam injection molding. Special chemical blowing agents [201] are admixed prior to processing which occurs at about 300 °C [202].

The manufacture of lighter weight foams by extrusion with chlorinated hydrocarbon blowing agents has also been described [203].

Properties and Uses. Polycarbonate foams, like the solid material, exhibit high impact resistance, rigidity, and heat resistance. They are flame-resistant and have a high electrical resistance. They are resistant to acid, but are attacked by bases, and by aromatic or halogenated hydrocarbons.

Examples of applications are housings for electrical apparatus, sports equipment (water skis), and furniture parts. Trade names are listed in Table 1.

5. Testing (→ Plastics, Properties and Testing)

Almost all major industrialized countries have adopted their own standards for testing foamed plastics (e.g., ASTM, United States; BS, Great Britain; DIN, Federal Republic of Germany; Association Française de Normalisation AFNOR, France). International standards (ISO) also exist. In some countries, test methods are specified in end-use-related standards. Thus, for example, foams for the building industry and their test methods are specified in ASTM C 578 "Standardized Specification for Preformed, Cellular Polystyrene Thermal Insulation."

In this overview, only the principles of the most important, generally valid test methods are described. Detailed information necessary for actual testing must be obtained by referring to the standards listed.

5.1. Mechanical Properties

Density (ASTM D 1622, ASTM C 303, ISO 845, DIN 53 420). The density is determined by using cubical test specimens or slabs; it is measured in kilograms per cubic meter.

Open- and Closed-Cell Content (ASTM D 2856). The content of open and closed cells is determined by measuring the apparent and true volumes with a special apparatus. The fraction of open or closed cells is reported in volume percent.

5.1.1. Rigid Foams

Compressive Properties (ASTM D 1621, ISO 844, DIN 53 421). The *compressive stress* is the compressive load per unit area of minimum original cross section carried by the test specimen at any given moment (Fig. 11). It is expressed in force per unit area, e.g., newtons per square meter or pascals. The *compressive strength* is the stress at the yield point (the first point on the stress–strain curve at which an increase in strain occurs without an increase in stress). In the absence of such a yield point, the compressive strength is the stress at 10 % deformation.

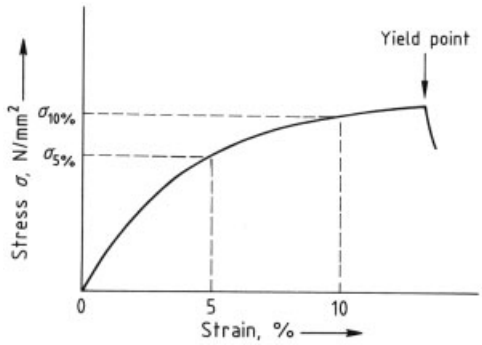

Figure 11. Typical stress–strain curve for a foamed plastic

Tensile Properties (ASTM D 1623, ISO 1962, DIN 53 430). The *tensile strength* σ_B is obtained by dividing the breaking load in newtons by the original minimum cross-sectional area of the specimen in square centimeters. The *elongation* is the extension of the material at the moment of rupture expressed as a percentage of its original length measured between gauge marks. The shape of the test specimen used depends on the amount of sample available, as well as the stiffness of the material. It may be round or square and have tapered ends or rounded shoulders; length varies from 10 cm to 1 m. A typical shoulder shape is shown in Figure 12. Speed of testing strongly influences the results; a standard speed is chosen such that rupture occurs in 1 – 6 min.

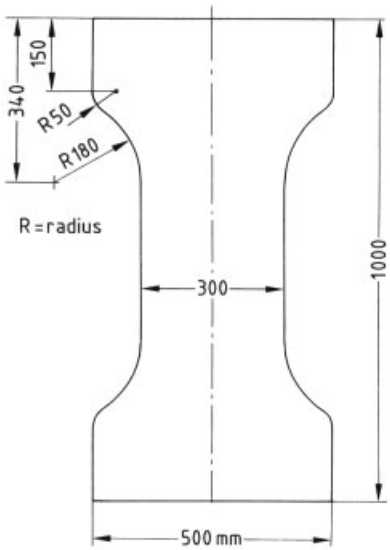

Figure 12. Specimen for tensile testing
All lengths are given in millimeters.

Breaking Load and Flexural Properties (ASTM C 203, ISO 1209, DIN 53 423). The breaking load and flexural properties are tested by different procedures on a bar of rectangular cross section which can be loaded in two ways:

1) The bar rests on two supports and is loaded midway between the supports.
2) The bar rests on two supports and is loaded at the two quarter points.

Normally, values are measured at the maximum load or breaking point. The *flexural strength* is equal to the maximum stress in the outer layers at break; it is calculated from the load applied at the break point by means of equations that depend on the testing procedure used. Flexural properties may vary with specimen span-to-thickness ratio, temperature, difference in rate of straining, and testing procedure followed. In comparing results, these parameters must be consistent.

Dynamic Stiffness (DIN 52 214). The dynamic stiffness of elastified rigid foams is easily measured and is used to estimate their potential as impact soundproofing material. A metal plate that applies a 2 kPa (kN/m^2) load uniformly to the test specimen is affixed to it with plaster. An alternating force is applied perpendicular to the metal plate. The resonance frequency f_r of the system is measured. The dynamic stiffness s' is proportional to f_r^2.

5.1.2. Flexible Foams

Several procedures are used to test the compressive behavior of flexible foams.

Compression Load Deflection Test (ASTM D 1564, ASTM D 2406, DIN 53 577). This test involves measuring the load necessary to produce a defined compression (20, 40, or 60 %) over the entire surface area of the foam specimen. The compressive stress at the specified compression is given in pascals or newtons per square meter. The test is also used to measure the relaxation of flexible foams after removal of the compressive load.

Indentation Load Deflection (ILD) (ASTM D 1564, ASTM D 2406, DIN 53 576). This test measures the load necessary to produce a defined indentation (e.g., 25, 65, or 70 %) in the foam product. A flat, circular indentor foot of specified diameter is connected to a load-measuring device by means of a swivel joint and is pressed into the specimen. The force in newtons required for 25 and 65 % indentation is known as the 25 and 65 % ILD value, respectively.

The ASTM tests also describe a procedure for measuring indentation on loading. These values are known as indentation residual gage

load (IRGL) values. The load deflection is determined by measuring the thickness of the foam under fixed loads on a circular indentor foot.

Compression Set Test (ASTM D 1564, ASTM D 2406, DIN 53 572). This test involves compressing the foam specimen under specified conditions of time and temperature, and noting the effect on the thickness of the specimen. The compression device consists of two or more flat plates arranged so that the plates are held parallel to each other by bolts or clamps; the space between the plates can be adjusted to the required thickness by means of spacers.

Standard testing conditions are a compression of 50 or 75 %, a temperature of 70 °C for 22 h, and a relaxation time of 30 min. The constant-deflection compression set is calculated and expressed either as a percentage of the original thickness or as a percentage of the original deflection.

Tensile Properties (ASTM D 1564, ASTM D 2406, DIN 53 571). Tensile stress, tensile strength, and elongation at break are measured when a tensile load is applied to the foam. The test specimen is stamped out with a die of a specified shape (shoulder shape, see Fig. 12). The *tensile stress,* in pascals or newtons per square millimeter, at a given elongation is the load applied at this elongation divided by the original cross-sectional area of the specimen. The *tensile strength* is the tensile stress at the breaking point.

Resilience (Ball Rebound) Test (ASTM D 1564). The ball rebound test is used to evaluate the resilience of foams; a steel ball is dropped on a foam specimen, and the height of rebound is noted. Values are given as a percentage of the dropping height and are known as ball rebound resilience (BRR) values.

5.2. Thermal Properties

Dimensional Stability (ISO 2796, DIN 53 431) consists of measuring the dimensions of a test specimen at a standard test temperature (e.g., − 25, 0, or +70 °C). The change in linear dimensions is expressed as a percentage of the values measured at room temperature.

Compressive Creep (ISO/TR 2799, ISO 4898, DIN 18 164) consists of measuring the change in thickness of a test specimen when it is subjected to a load at elevated temperature. The change is given as a percentage of the original thickness. Several testing procedures are used; common testing conditions follow.

t, °C	Load, kPa	Time, h
70	0	48
80	20	48
70	40	168

Deflection Temperature (ASTM D 648, DIN 53 424) consists of determining the temperature at which an arbitrary deformation occurs when a specimen is subjected to an arbitrary set of testing conditions. Either a compressive or a flexural load may be used.

Flexural Load. A bar of rectangular cross section is tested as a simple beam; the load is applied at its center, or the bar is held at one end while the load is applied at the other. The loaded specimen is immersed in a heat-transfer medium whose temperature can be increased at a constant rate (e.g., 2 °C/min or 50 °C/h). The temperature of the medium is measured when a certain deflection (e.g., 2 mm) has been reached and is recorded as the deflection temperature under flexural load.

Compressive Load. A test specimen of 50×50×20 mm is loaded with 25 kPa (0.025 N/mm^2). Then, as with flexural load, the temperature of the test specimen is raised at a constant rate (e.g., 50 °C/h). The temperature at which a certain deflection is reached is recorded as the deflection temperature under compressive load.

Thermal Conductivity (DIN 52 612, ASTM C 177, ASTM C 518, ASTM C 236). The thermal conductivity, expressed in watts per meter per kelvin, is the rate of heat flow through a unit thickness of an infinite slab of a homogeneous material in a direction perpendicular to the surface, induced by a unit temperature difference.

Thermal conductivity is one of the most important properties of foamed plastics and is generally low due to their cellular structure.

Because of the technical importance of this property and the complicated nature of the

measurements, different test methods exist. In principle, a steady-state heat flux is measured through flat slabs of the specimen using a guarded hot-plate apparatus. Single-sided and double-sided modes of measurement are used. An idealized double-sided system is shown in Figure 13. The specimen is sandwiched between two isothermal cold plates and a guarded hot plate. The guard portion of the isothermal hot plate provides the power necessary to create the proper thermal conditions within the test volume.

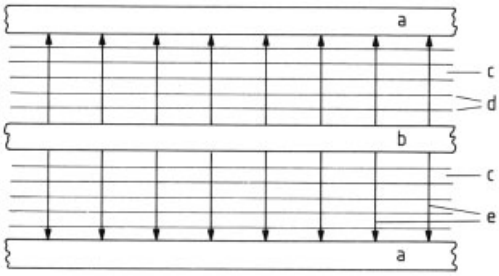

Figure 13. Schematic of the guarded hot-plate apparatus for determining thermal conductivity
a) Cold isothermal surface; b) Hot isothermal surface; c) Specimen; d) Isothermal planes; e) Heat flux lines

A good review of terms, definitions, symbols, and units pertaining to thermal insulating materials is given in ASTM C 168.

Fire Behavior. The fire behavior of foamed plastics depends on many parameters. Therefore, many different tests exist, each of which emphasizes different aspects of burning, i.e. ease of ignition, flame propagation, heat release, combustion products (smoke density and toxicity). The most important tests are DIN 4102, DIN 53 438, ASTM E 84, NF P92 501, NF P92 504, NF P92 505, NF P92 506, UL 94, ASTM D 2863-77.

5.3. Miscellaneous Properties

Water absorption (ASTM D 2842, ISO 2896, DIN 53 433) of cellular plastics is determined by measuring the change in buoyancy resulting from immersion under a head of water for a certain time (2 h, 4 d, 7 d, 28 d, or longer). Water absorbed by the specimen lowers buoyancy by reducing the volume of water displaced. Water absorption may be expressed in volume percent or weight percent.

Thermal and Humid Aging (ASTM D 2126). Because of the wide variety of potential uses of cellular plastics, artificial aging tests must be based on the intended application. Normally, responses to temperature and humidity are selected. Temperature ranges from − 70 to + 200 °C and humidity ranges from ambient to 97 % are used. The specimen is examined visually; changes in density and dimensions are also measured.

Electrical properties of foamed plastics are seldom measured. Testing methods are identical with those used for standard plastics (see → Plastics, Properties and Testing).

Table 4. Consumption of foamed plastics in 1986

Foamed plastics *	Consumption, 10^3 t		
	Western Europe	United States	Japan
Expanded polystyrene (EPS)	485	283	190
Extruded polystyrene (XPS)	135	293	300
Poly(vinyl chloride) (PVC)**	25	150	
Polyolefin (PO)	20	25	37
Urea–formaldehyde (UF) resin	12		
Phenol–formaldehyde (PF) resin	10	8	3

* Foamed plastics with an apparent density of $< 200\,kg/m^3$.
** Including foams with apparent density $> 200\,kg/m^3$.

6. Economic Aspects

Polyurethanes are the most important foams as far as consumption is concerned and are described elsewhere (→ Polyurethanes). The consumption of other major foamed plastics in Western Europe, the United States, and Japan is summarized in Table 4. The most important of these are foams derived from polystyrene beads and extruded polystyrene. Significant amounts of foam from polyolefins, poly(vinyl chloride), phenol–formaldehyde resins, and urea–formaldehyde resins are also used.

7. Toxicology and Environmental Aspects

The polymeric materials used for foamed plastic production generally do not present a toxicological hazard because they are not absorbed by

the human body. Of greater concern are residual monomers and additives, i.e., materials of lower molecular mass. All legal requirements pertaining to such constituents must be observed. For the benefit of processors and end users, the manufacturers of synthetic foams provide pamphlets, safety bulletins, and technical information on the processing and proper handling of synthetic foams.

Foamed plastic waste is bulky and is therefore often compacted before disposal. Since toxic substances are not released on storage, contamination of ground water does not occur. After being dumped the foamed plastics are slowly degraded, the rate depending on the nature of the plastic [204]. During the controlled aerobic degradation of noncomminuted thermoplasts and thermosets (composting), soil-inhabiting Actinomycetes were shown to attack aminoplasts most effectively [205].

Although foamed plastics have a high calorific value, they do not play an important role in refuse incineration plants because of their low mass. However, thermoplastic foams must be mixed with sufficient quantities of nonmelting materials with a low calorific value.

The combustion products of most foamed plastics are carbon dioxide and steam which do not pose any environmental problems. However, chlorine-containing products release hydrogen chloride which must be removed by washing the flue gas [206].

Chlorofluorocarbons (CFCs) are used in the production of many foamed plastics. They remain enclosed within the cells of the plastic and increase the thermal insulation of air-filled foams.

The ozone depletion hypothesis [207] suggests that CFCs damage the stratospheric ozone layer that prevents shortwave UV radiation from reaching the earth. Models of stratosphere chemistry, propose as many as 200 possible reactions that involve CFCs, carbon dioxide, nitrogen oxides, and methane. Chlorine atoms seem to play an important role in initiating ozone-consuming reactions [208–210].

"Soft" CFCs (e.g., chlorodifluoromethane [*75-46-6*]) are being introduced as alternative blowing agents. Since they are less stable than the customery CFCs, they are reported to decompose in the troposhere and thus have a much lower ozone depletion potential [211], [212].

8. References

General References

1. C. J. Benning: *Plastic Foams,* **vols. 1, 2,** Wiley-Interscience, New York 1969.
2. A. A. Berlin, F. A. Shutov, A. K. Zhitinkina: *Foam based on Reactive Oligomers,* Technomic Publ. Co, Westport 1982.
3. K. C. Frisch, J. H. Saunders: *Plastic Foams,* **vols. 1, 2,** Marcel Dekker, New York 1972/73.
4. N. C. Hilyard: *Mechanics of Cellular Plastics,* Applied Science Publ., London 1982.
5. D. Homann: *Kunststoff-Schaumstoffe,* Hanser Verlag, München 1966.
6. T. H. Ferrigno, *Rigid Plastic Foams,* Reinhold Publ. Co., New York 1967.
7. W. Knappe in R. Vieweg, D. Braun (ed.): *Kunststoff-Handbuch,* vol. 1, Chap. 5.1.4.6, Hanser Verlag, München 1975.
8. H. Piechota, H. Röhr, *Integralschaumstoffe,* Hanser Verlag, München – Wien 1975.
9. F. A. Shutov: *Syntactic Polymer Foams,* Springer Verlag, Berlin – Heidelberg 1986.
10. G. Trausch in VDI-Ges. Kunststofftechnik (ed.): *Schäume aus der thermoplastischen Schmelze,* VDI-Verlag, Düsseldorf 1981.
11. B. R. Wendle: *Structural Foam,* Marcel Dekker, New York 1985.

Specific References

12. DIN 7726,(1982): Cellular Materials: Definition of Terms, Classification.
13. F. Lober, *Angew. Chemie* **64** (1952) 65 – 92.
14. R. A. Gorski, R. B. Ramsey, K. T. , Dishart, *J. Cell. Plast.* **22** (1986) 21 – 52.
15. J. H. Saunders, R. H. Hansen in K. C. Frisch, J. H. Saunders (ed.): *Plastic foams,* Part I, Chap. 2, Marcel Dekker, New York 1972.
16. W. S. Fong, *Foamed Plastics, Suppl. A,* Process Economics Program Report 97A, SRI Int., Menlo Park, California, Jan. 1984, pp. 29 – 36.
17. *Mod. Plast. Int.,* **6** (1976) no. 9, 56 – 57. Ciba Geigy, EP 7 499, 1978.
18. R. J. Bathgate, K. T. Collington, *39th Annu. Tech. Conf.-Soc. Plast. Eng.* **27** (1981) 856 – 862.
19. J. A. Gribens, R. C. Wade, *39th Annu. Tech. Conf.-Soc. Plast. Eng.* **27** (1981) 852 – 855.
20. R. H. Hansen, W. M. Martin, *Ind. Eng. Chem.* **3** (1964) no. 2, 137 – 141.
21. G. Menges, F. Knipschild in N. C. Hilyard (ed.): *Mechanics of Cellular Plastics,* Applied Science Publ., London 1982, p. 27 – 72.

22. D. J. Doherty, R. Hurd, G. R. Lester, *Chem. Ind. London* **30** (1962) 1340 – 1356.
23. B. Koglin, *Kältetech. Klima.* **21** (1969) 122 – 125.
24. N. E. Hager, R. C. Steere, *J. Appl. Phys.* **38** (1967) 4663 – 4668.
25. W. Küster, *Kunststoffe* **60** (1970) 249 – 255.
26. H. Zehendner, 12. Int. Fachtagung für Schaumkunststoffe, Düsseldorf 1983, pp. 153 – 158.
27. H. Mittasch, *Plaste und Kautsch.* **16** (1969) 268 – 271.
28. DIN 18 164/2-1979: Insulating Materials for Impact Sound Insulation.
29. ASTM D 2863-77: Oxygen Index.
30. Dow Chemical, US 3 067 147, 1957.
31. A. Osakada, *Jpn. Chem. Q.* **5** (1969) 55.
32. N. Shiina, M. Tsuchiya, H. Nakae, *Jpn. Plast. Age* **10** (1972) no. 12, 37 – 48.
33. Toray Ind., US 3 542 702, 1967.
34. Sekisui, JP-Kokai 61-143 450, 1986.
35. Mitsubishi Petr., JP-Kokai 58-134 131, 1983.
36. Y. D. Lee, L. F. Wang, *J. Appl. Polym. Sci.* **32** (1986) 4639 – 4647.
37. Hercules, US 3 341 481, 1964.
38. M. B. Johnson, *Polym. Age* **6** (1975) 88 – 89.
39. Dynamit Nobel, DE 3 430 108, 1984.
40. Furukawa, DE-AS 1 694 130, 1967.
41. F. W. Alfter, *Plastverarbeiter* **29** (1978) 129 – 132.
42. Sekisui, DE 1 947 589, 1969.
43. *GAK Gummi, Asbest + Kunstst.* **36** (1983) 356 – 357.
44. BASF, DE-OS 2 542 452, 1975.
45. F. Stastny, R. Gäth, U. Haardt, *Kunststoffe* **61** (1971) 745 – 749.
46. Kanegafuchi, US 4 448 901, 1983.
47. Japan Styrene Paper, US 4 399 087, 1981.
48. Yuka Badische, DE-OS 3 431 245, 1983.
49. Japan Styrene Paper, US 4 440 703, 1982.
50. N. Shiina, M. Tsuchiya, H. Nakae, *Jpn. Plast. Age* **11** (1973) no. 3, 47 – 53.
51. F. Stastny, R. Gäth, U. Haardt, *GAK Gummi, Asbest + Kunstst.* **25** (1972) 398 – 402.
52. F. Kleiner, *Kautsch. + Gummi, Kunstst.* **28** (1975) 149 – 153.
53. BASF, DE 1 629 316, 1966.
54. U. G. Haardt, *Kunststoffe* **68** (1978) 468 – 469.
55. BASF, DE-AS 1 629 281, 1966.
56. Kanegafuchi, US 3 953 558, 1973.
57. Japan Styrene Paper, EP 097 504, 1983.
58. N. Shiina, M. Tsuchiya, H. Nakae, *Jpn. Plast. Age* **11** (1973) no. 1, 49 – 53, **11** (1973) no. 2, 45 – 50.
59. I. Tamai, K. Yamaguchi, *Jpn. Plast. Age* **10** (1972) no. 12, 22 – 27.
60. K. Leilich, K. E. Büscher, in R. Vieweg (ed.): *Kunststoff-Handbuch,* **vol. IV,** Hanser Verlag, München 1969, p. 709.
61. J. Helberg, W. Löw, T. Riedel, *Kunststoffe* **75** (1985) 342 – 345.
62. BASF, DE 845 264, 1951; DE 941 389, 1951.
63. K. Hinselmann, D. Stockburger, *Kunststoffe* **61** (1971) 152 – 157.
64. I. De Grave, *Kunststoffe* **70** (1980) no. 10, 625 – 629.
65. K. Stange, *Kunststoffe* **74** (1984) no. 10, 556 – 559.
66. E. Neufert: *Styropor-Handbuch,* 2nd ed., Bauverlag GmbH, Wiesbaden – Berlin 1971.
67. W. Prankel, *Tech. Rundsch.* **39** (1979) 9 – 15; **42** (1979) 5 – 7.
68. G. Baum, *Beton Information* **5** (1974) 58 – 67.
69. W. Albert, H. Gümpel, H. Röhr, *Kunststoffe* **61** (1971) 386 – 396.
70. Dow Chemical, US 2 450 436, 1947.
71. Dow Chemical, US 2 669 751, 1950.
72. A. Krämer, *Plastverarbeiter* **25** (1974) 73 – 78.
73. A. Krämer, *Kunststoffe* **64** (1974) 350 – 353.
74. R. Städter, *Kunststoff J.* **8** (1974) no. 7/8, 22 – 28.
75. R. P. Sacha, 29th Annu. Techn. Conf. – Soc. Plast. Eng., May 10 – 13 1971, pp. 326 – 328.
76. Gebr. Anger, DE 1 963 811-Q, 1971.
77. Dynamit Nobel, DE 2 158 673-Q, 1973.
78. *Kunstst. Berat.* **18** (1973) no. 11, 812.
79. *Kunstst. Berat.* **19** (1974) no. 4, 180 – 182.
80. K. Wetzel, *Plaste Kautsch.* **24** (1977) no. 3, 197 – 199.
81. H. Barth, *Kunststoffe* **67** (1977) no. 3, 130 – 135.
82. Dow Chemical, US 4 146 563, 1979.
83. *Mod. Plast. Int.* **12** (1982) no. 4, 18.
84. H. W. Matill, A. Schaefer, *Kunstst. Bau* **20** (1985) no. 2, 81 – 83.
85. P. Zingsheim in W. Becker, D. Braun (eds.): *Kunststoff Handbuch,* **vol. 2/2,** Carl Hanser Verlag,p. 977 –1000.
86. Vallourec, FR 2033 555-Q, 1970.
87. E. Morianz, *"Schäume aus der thermoplastischen Schmelze,"* in Reference 10 pp. 85 – 93.
88. *Plast. Ind. News* **24** (1978) no. 2, 26.
89. Ugine Kuhlmann, DE-OS 1 919 921, 1969.
90. L. F. Fow, *Soc. Plast. Eng. (Tech. Pap.)* **24** (1978) 29 – 33.
91. *Mod. Plast. Int.* **13** (1983) no. 6, 34 – 35.
92. Kleber Colombes, DE-OS 1 569 238, 1963.
93. *Elastomerics* **114** (1982) no. 4, 18.

94. M. Bouzim, *Rev. Gen. Caoutch. Plast.* **49** (1972) no. 6, 541.
95. Y. J. Landler, *J. Cellular Plast.* **3** (1967) no. 9, 400 – 404.
96. H. Boelle, *GAK Gummi, Asbest + Kunstst.* **34** (1981) no. 9, 546 – 552.
97. UCC, DE-AS 10 10 271, 1955.
98. Israel Dennis, US 2 763 475, 1954.
99. Vanderbilt, DE-AS 1 719 311, 1967.
100. K. Hillier, *Plast. Rubber Process Appl.* **1** (1981) 39 – 42.
101. C. J. Hilado, P. A. Huttlinger, *J. Cell. Plast.* **16** (1980) 285 – 292.
102. A. J. Lowe, A. Barnatt, *Chimica Ind. (Milan)* **61** (1979) 656 – 660.
103. M. Cherubim in *Schaumkunststoffe, Berichte von den FSK-Tagungen 1971 – 1975,* Hanser Verlag, München – Wien 1976, pp. 402 – 406.
104. F. A. Shutov, *Cell. Polym.* **3** (1984) 95 – 104.
105. V. D. Valgin, *Europlast. Mon.* **46** (1973) no. 7, 57 – 60.
106. K. Brüning in *Schaumkunststoffe,* cf. Reference 103, pp. 407 – 410.
107. Koppers, DE-OS 3 139 890, 1981.
108. Koppers, DE-OS 3 324 431, 432, 433, 1983.
109. BP Chemicals, EP-A 154 452, 1985.
110. BP Chemicals, EP-A 170 357, 1985.
111. Isomo, EP-A 108 167, 1982.
112. BP Chemicals, EP-A 068 673, 1982.
113. A. J. Lowe, A. Barnatt, E. F. Chandley, R. Dyke, K. Hillier, *Eur. J. Cell. Plast.* **1** (1978) no. 1, 42 – 49.
114. A. Vazquez, R. J. Williams, *Cell. Polym.* **5** (1986) 123 – 140.
115. F. A. Shutov, *Cell. Polym.* **1** (1982) 73 – 75.
116. G. C. Marks, 12. Int. Schaumkunststofftagung, Düsseldorf 1983, pp. 59 – 67.
117. A. S. Wood, *Mod. Plast. Int.* **11** (1981) no. 10, 48 – 50.
118. Smithers, US 2 753 277, 1953.
119. F. A. Shutov, V. V. Ivanov, *Cell. Polym.* **3** (1984) 369 – 382.
120. E. G. Norman: "Plastics and Rubber," *Processing* **4** (1979) no. 3, 41 – 45.
121. H. D. Schmidt, *Kunstst. Bau* **10** (1975) no. 2, 8 – 10.
122. *Mod. Plast. Int.* **6** (1976) no. 6, 8 – 10.
123. Sekisui Plast., JP-Koaki 59-155 016, 1983.
124. E. Rühl, EP-A 154 874, 1985.
125. B. Mosier, US 4 596 682, 1984.
126. Thermocell Dev., US 4 390 641, 1982.
127. L. Unterstenhöfer, *Kunststoffe* **57** (1967) 850 – 855.
128. H. Baumann, *Plastverarbeiter* **27** (1976) 235 – 243.
129. E. K. H. Wulkan, 9. Int. Fachtagung für Schaumkunststoffe, Düsseldorf 1979, pp. 59 – 82.
130. Schaumchemie, DE 1 054 232, 1956.
131. BASF, DE-OS 2 064 102, 1970.
132. H. Baumann, *Cell. Polym.* **3** (1984) 249 – 261.
133. Schaumchemie, DE 3 216 897, 1982.
134. Schaumchemie, DE-OS 3 108 279, 1981.
135. E. K. H. Wulkan, *Kunstst. Bau* **14** (1979) no. 4, 167 – 171.
136. L. Gerhardy, H. Lottner, *Baugewerbe* (1980) no. 3, 30 – 34.
137. R. Grossmann, *Glückauf* **112** (1976) no. 14, 803 – 807.
138. Bauer, DE-AS 1 014 508, 1953.
139. Schaumchemie, DE 1 298 488, 1967.
140. BASF, DE-AS 1 794 131, 1968.
141. Vorwerk, DE 2 732 011, 1977.
142. J. Miltz, G. Gruenbaum, *J. Cell. Plast.* **17** (1981) no. 7/8, 213 – 219.
143. CIBA, DE 870 027, 1948.
144. H. Weber, F. Krückau, *Kunststoffe* **75** (1985) 843 – 848.
145. H. Weber, 14. Schaumkunststofftagung, Düsseldorf 1986, pp. 61 – 77.
146. BASF, EP 074 593, 1979.
147. BASF, EP 037 470, 1981.
148. BASF, DE-OS 3 109 929, 1981.
149. BASF, DE-OS 3 138 862, 1981.
150. DIN 53 577 (1976); ISO 3386/1: Determination of Compression Stress/strain Characteristics and Compression Stress Value.
151. BASF, EP-A 111 860, 1983.
152. General Electric, US 2 875 163, 1956.
153. Dow Corning, US 3 070 555, 1959.
154. Wacker-Chemie, DE 1 224 040, 1963.
155. General Electric, US 3 425 967, 1965.
156. Dow Corning, US 3 923 705, 1974.
157. C. L. Lee, R. G. Niemi, K. M. Kelly, *J. Cell. Plast.* **13** (1977) no. 1/2, 62 – 66.
158. C. L. Lee, G. M. Ronk, *J. Cell. Plast.* **18** (1982) no. 5/6, 178 – 182.
159. C. L. Lee, G. M. Ronk, S. Spells, *J. Cell. Plast.* **19** (1983) no. 1/2, 29 – 33.
160. Dow Corning, US 4 567 212, 1985.
161. Dow Corning, EP-A 099 002, 1983.
162. Polyvoltac Ltd. (UK), Information sheet, Southall, England, May 1986.
163. US Air Force, US 3 429 838, 1966.
164. H. L. Vincent in K. C. Frisch, J. H. Saunders (ed.): *Plastic Foams,* Part I, "Silicone Foams," Marcel Dekker, New York 1972, pp. 385 – 396.
165. M. Hajimichel, A. Lewis, D. Scholey, C. Simmonds, *Br. Polym. J.* **18** (1986) no. 5, 307 – 311.

166. F. A. Shutov, A. A. Berlin: *Epoxy Oligomer Based Foams in Foam based on Reactive Polymers,* Technomic Publishing Co., Westport 1982, pp. 215 – 261.
167. F. A. Shutov: "Syntactic Polymer Foams," in *Advances in Polymer Science,* vol. 73/74, Springer Verlag, Berlin – Heidelberg 1986, pp. 63 – 123.
168. *Shell Polymers* **7** (1983) no. 3, 80 – 81.
169. M. Narkis, S. Kenig, M. Putermann, *Polym. Compos.* **5** (1984) 159 – 164.
170. S. Kenig, I. Raiter, M. Narkis, *J. Cell. Plast.* **20** (1984) 423 – 429.
171. Int. Harvester, US 4 296 208, 1980.
172. IMI-Tech Corp., US 4 535 101, 1984.
173. P. B. Rand, *J. Cell. Plast.* **9** (1973) 130 – 133.
174. *Mod. Plast. Int.* **10** (1980) no. 3, 21.
175. *Plast. Technol.* **28** no. 3, March 1982, pp. 15 – 17.
176. W. Frank, 14. Schaumkunststofftagung, Düsseldorf 1986, pp. 9 – 22.
177. J. Gagliani, *NASA Tech Briefs,* Fall 1980, pp. 384 – 387.
178. *Chem. Eng. News* **59** (1981) no. 1, 21.
179. *Kunststoffe* **77** (1987) 57.
180. Röhm, DE 1 817 156, 1968.
181. Röhm, DE 2 726 259, 1977.
182. Röhm, DE 2 726 260, 1977.
183. W. Pip, *Kunststoffe* **64** (1974) 23 – 27.
184. W. Pip, *Plastverarbeiter* **34** (1983) 629 – 632.
185. Röhm: *Rohacell-Handbuch,* Darmstadt 1986.
186. Imchemie, EP 068 439, 1982 (A. Hohnholz).
187. *Kunstst. Bau* **20** (1985) 75.
188. J. D. Gabbert, R. M. Hedrick, *Kunststoffe* **75** (1985) 416 – 420.
189. P. Wagner, *Kunststoffe* **73** (1983) 588 – 590.
190. J. Härting, *Kunstst. Berat.* **20** (1984) no. 11/12, 30 – 32.
191. Rogers Corp., US 4 464 491, 1983.
192. M. Kuga, K. Matsukura, *Jpn. Plast. Age* **10** (1972) 32 – 39.
193. UBE Industries, JP-Kokai 57 000 144, 1982.
194. BASF, DE-AS 1 159 643, 1961.
195. BASF, DE-AS 1 177 340, 1962.
196. Mobil Oil, US 4 598 101, 1985.
197. W. H. Smarook, *Soc. Plast. Ind., 32nd Annu. Tech. Conf.,* 1977 Section 11-G, 1 – 8.
198. D. A. Blackadder, A. Gupta, W. D. James, *Chem. Eng. Sci.* **38** (1983) no. 1, 181 – 182.
199. H. R. Smith, WO 82/00649 (1981).
200. ICI, EP 16 544, 1980.
201. Bayer, DE 1 252 891, 1961.
202. W. Cohnen, W. Meyer, *Kunstst. Rundsch.* **20** (1973) 425 – 434.
203. Mobil Oil, US 4 579 874, 1985.
204. E. Küster, *J. Appl. Polym. Sci. Appl. Polym. Symp.* **35** (1979) 395 – 404.
205. P. Chrometzka, F. X. Kneer, *Forum Umwelt Hyg.* **26** (1975) no. 8, 225 – 230.
206. H. Reimer, *Müll Abfall* **4** (1972) no. 1, 19 – 21.
207. F. S. Rowland, M. J. Molina, *Rev. Geophys. Space Phys.* **1** (1975) 1 – 35.
208. J. F. Mills, *Polyurethanes World Congress 1987,* Aachen, pp. 54 – 58.
209. ICI Brochure: *Chlorofluorocarbons and the Ozone Layer,* ICI, Runcorn, Ch, England, 1986.
210. L. R. Ember, P. L. Layman, W. Lepkowski, P. S. Zurer, *Chem. Eng. News* **64** (1986) no. 47, 14 – 64.
211. K. T. Dishart, J. A. Creazzo, M. R. Ascough in Reference 208, pp. 59 – 66.
212. *Mod. Plastics Int.* **17** (1987) no. 12, 57 – 59.

Foams and Foam Control

Rainer Höfer, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chaps. 1, 3, 4, 5 [in part], 6, and 8)

Franz Jost, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 2)

Milan J. Schwuger, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 2)

Rolf Scharf, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 5.2.1)

Jürgen Geke, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 5.3)

Josef Kresse, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 5.3)

Herbert Lingmann, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 5.3)

Rudolf Veitenhansl, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 5.8)

Werner Erwied, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Chap. 7)

1. Introduction 2
2. Theory of Foam and Foam Control 3
2.1. Characteristic Properties of Foam 3
2.2. Foam Formation 3
2.2.1. Surface Activity of Adsorbed Surfactants 3
2.2.2. Methods of Foam Generation 4
2.3. Foam Stability 5
2.3.1. Gibbs Film Elasticity and Marangoni Effect 5
2.3.2. Rheological Aspects of Foam Stability 5
2.3.3. Electrostatic Foam Stabilization . . 6
2.3.4. Mechanism of Lamellar Rupture . . 8
2.4. Foam Inhibition and Destruction 8
2.4.1. Chemical Foam Inhibition and Destruction 9
2.4.2. Mechanical Defoaming 10
3. Antifoaming Agent Composition . 10
3.1. Carrier Oils 10
3.2. Silicone Oils and Silicone Foam Inhibitors 10
3.3. Hydrophobic Silica 11
3.4. Hydrophobic Fat Derivatives and Waxes 11
3.5. Water-Insoluble Polymers 11
3.6. Amphiphilic Components 12
3.7. Emulsifiers 12
3.8. Coupling Agents 12
4. Mechanical Means of Combating Foam 12
5. Foam Problems in Specific Applications 13
5.1. Detergents 13
5.2. Food and Beverage Industries . . 13
5.2.1. Cleansing and Disinfecting 13
5.2.1.1. Tank Cleansing Based on the Cleaning-in-Place (CIP) Method . . 13
5.2.1.2. Cleansing of Returnable Bottles . . 13
5.2.2. Food Processing 14
5.2.2.1. Sugar 14
5.2.2.2. Yeast 15
5.2.2.3. Potatoes 15
5.3. Metal Treatment 15
5.3.1. Foam Problems Associated with Cooling Lubricants 15
5.3.2. Foam in Alkaline Cleansing Solutions 16
5.3.3. Foam Control in Neutral Cleansers 17
5.4. Polymer Industry 17
5.5. Paint and Coating Industry 18
5.6. Construction Industry 19
5.7. Adhesives Industry 19
5.8. Textile Industry 20
5.8.1. Pretreatment 21
5.8.2. Dyeing 21
5.8.3. Foam Application of Textile Auxiliaries 22
5.9. Leather Industry 22
5.10. Pulp and Paper Industry 22
5.10.1. Pulping 22
5.10.2. Pulping Liquor Disposal 23
5.10.3. Air Content of Pulp Suspensions . . 23
5.10.4. Conversion to Paper 23
5.10.5. Paper Coating 23
5.10.6. Waste Paper Deinking 23
5.11. Phosphoric Acid Manufacture . . 23
5.12. Wastewater Treatment 24
6. Testing Methods 24
7. Legal Aspects 25
8. Economic Aspects 25
9. References 26

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_465

1. Introduction

In a number of contexts, foaming is considered a desirable phenomenon. The food and beverage industries are well aware of the wholesome connotations associated with the foamy head on a glass of beer or such foams as whipped cream and meringue. Generations of housewives have assumed that foam is evidence of a detergent's cleaning power. *Meerschaum* provides another interesting example. It is a form of sepiolite [*61180-58-3*] prized as a decorative basis for pipes, and its common name means "sea foam" in German. Other solid foams are important for their use as insulating or lightweight construction materials (→ Foamed Plastics).

Nevertheless, foam is often very undesirable: it can severely hinder the filling of a container, reduce the efficiency of many high-speed operations, and cause surface defects in coating operations. Various terms such as *foam inhibitor, antifoaming agent, foam controller,* and *defoamer* are applied to substances designed to prevent or control foaming, but the distinctions are somewhat vague and a given material often serves both preventive and control functions.

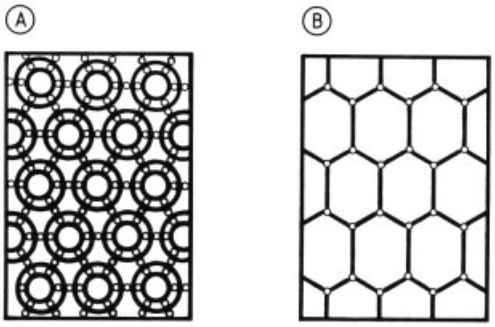

Figure 1. Schematic representations of the major types of foam [9]
A) Spherical-bubble foam; B) Polyhedral foam

2. Theory of Foam and Foam Control

2.1. Characteristic Properties of Foam

A foam consists of gas dispersed within a liquid. Pure liquids are incapable of foaming; soluble surfactants or impurities [8] are required to stabilize a foam by retarding the otherwise extremely rapid coalescence of gas bubbles.

Foams are thermodynamically unstable. A freshly prepared "wet" foam dissipates rapidly. Such a foam is of the spherical-bubble type [9] (Fig. 1 A), with individual bubbles having little influence upon one another. As a foam is dehydrated, the distance between bubbles progressively diminishes, leading ultimately to a "dry" polyhedral agglomerate of bubbles (Fig. 1 B). A metastable state can be established if equilibrium is reached among the various forces contributing to foam stabilization and destabilization [10].

Exceptionally regular polyhedral foam structures with a lifetime of several hours can be prepared if an appropriate system is thermostated and carefully protected from external influences [9], [11]. The physicochemical aspects of foam stabilization can be examined by preparing well-defined, isolated lamellae [10]. Nevertheless, a sound theoretical understanding of the various stabilization mechanisms has not yet been achieved.

2.2. Foam Formation

2.2.1. Surface Activity of Adsorbed Surfactants

The introduction and dispersion of a gaseous material under the surface of a foaming liquid (e.g., when air is brought into an aqueous surfactant solution) leads to the formation of rising bubbles. In the process, dissolved surfactant molecules diffuse out of the bulk phase toward the gas – liquid interfaces, to form reversibly adsorbed monolayers. Polar groups in the surfactant molecules remain directed toward the water, while hydrophobic chains extend into the air [12].

The spreading pressure π of an adsorbed surfactant increases to a limiting value as the surface concentration increases; this reduces the opposing surface tension of the initially pure liquid (γ_0) to a new value γ.

$$\pi = \gamma_0 - \gamma \tag{1}$$

The spreading pressure can be determined tensiometrically by the Wilhelmy method [13]. Because $d\pi = -d\gamma$, the spreading pressure characterizes the surface activity of a given surfactant or surfactant mixture at a particular concentration. At constant temperature, a dynamic

equilibrium is established between the surfactant dissolved in the bulk phase and the surfactant adsorbed at the boundary surface. Surface-active agents are repelled by the bulk phase because the van der Waals forces between water and the hydrophobic chains of the surfactant molecules are weaker than those between water molecules. At low concentrations, the amount of adsorbed surfactant Γ (surface concentration in moles per unit area) increases with increasing concentration c [14]. Simultaneously, the increasing spreading pressure π reduces the surface tension γ, as shown by Equation (1). For a nondissociating amphiphilic substance ($m = 1$), the following equation applies [15]:

$$G = -\frac{1}{mRT} \cdot \frac{\mathrm{d}\gamma}{\mathrm{dln}c} \tag{2}$$

Modified forms of this equation have been suggested [16] for ionic substances in which $1 < m \leq 2$. Equation (2) is known as the Gibbs adsorption equation; it describes an equilibrium situation and facilitates estimation of the surface area requirement per adsorbed molecule. Only at very low surface concentrations do hydrophobic groups lie flat on the surface, where they are free to rotate about their hydrophilic partners. A relatively low surfactant concentration often suffices to induce the formation of densely packed aggregates in which the hydrophobic chains are held erect.

Once the maximum spreading pressure has been reached, the adsorption capacity of the surface is fully utilized, so that the number of adsorbed molecules remains constant despite further increase in overall surfactant concentration. At this point, the two-dimensional structure is fully ordered. The critical micelle concentration (c_{M}), the concentration at which micelles begin to form, may be viewed as a measure of surface activity; above this value, the spreading pressure of a given pure surfactant is constant. Table 1 contains a selection of typical c_{M} values; these vary significantly as a function of the molecular structure of the surfactant. The effect of changes in the size of the hydrophobic group is illustrated in Figure 2. Nonionic ethoxylates with limited water solubility reach their maximum spreading pressure at extremely low volume concentrations. Concentration-dependent surface tension minima are commonly observed with commercial surfactants, but such minima appear to be caused by impurities. The time required to establish diffusion-controlled equilibrium increases markedly as the concentration differs from c_{M} [18] (Fig. 3).

Table 1. Critical micelle concentration c_{M} of surfactants at 25 °C

Surfactant	c_{M}, mol/L
$C_9H_{19}OSO_3Na$	59×10^{-3}
$C_{12}H_{25}OSO_3Na$	8.0×10^{-3}
$C_{16}H_{33}OSO_3Na$	0.38×10^{-3} (35 °C)
$C_{18}H_{37}OSO_3Na$	0.17×10^{-3} (50 °C)
$C_{12}H_{25}(OCH_2CH_2)_4OH$	4.3×10^{-5}
$C_{12}H_{25}(OCH_2CH_2)_8OH$	1.0×10^{-4}
$C_{12}H_{25}(OCH_2CH_2)_{12}OH$	1.4×10^{-4}
$C_{12}H_{25}N(CH_3)_3Cl$	1.8×10^{-2}
$C_{14}H_{29}N(CH_3)_3Cl$	5.0×10^{-3}
$C_{18}H_{37}N(CH_3)_3Cl$	3.5×10^{-4}

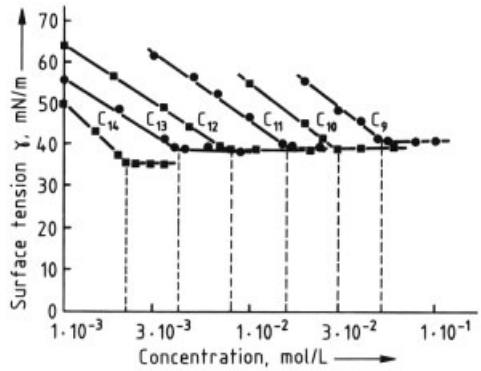

Figure 2. Influence of surfactant structure on surface activity [17]
Surface tension γ of aqueous sodium *n*-alkyl sulfate solutions as a function of concentration c at 25 °C.

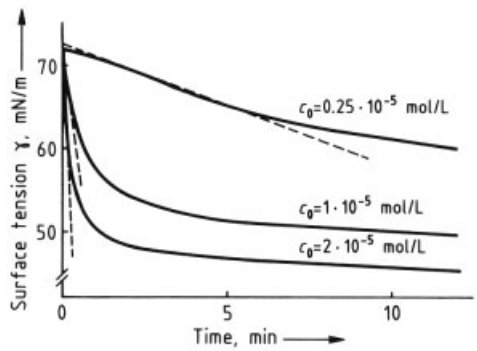

Figure 3. Influence of surfactant concentration on adsorption rate [18]
Surface tension vs. time curves for various concentrations of octylphenol dodecaglycol ether at 25 °C.

2.2.2. Methods of Foam Generation

Many techniques have been described for producing a foam [9], [19], but they fall into two main mechanistic categories.

Dispersion Methods. Mechanical action greatly increases the surface area that separates a liquid from a gas. The continuous phase may undergo extreme fragmentation, thereby establishing a foam with maximum homogeneity. Good results are often simply achieved in a long tube by establishing flow velocities in the turbulent range. In a commonly used procedure, the two phases are forced through a column tightly packed with glass beads or other appropriate material. Thin films of liquid form in the resulting narrow gaps, and these films are blown into bubbles by the passing gases. Foam generators and dispersion turbines operate in a similar fashion.

Condensation Methods. Foam generation methods based on condensation rely on spontaneous degassing of supersaturated, continuous liquid phases. The degassing may result from either pressure reduction or temperature increase. The most familiar example is the carbon dioxide foam that develops at the surface of carbonated soft drinks, beer, and sparkling wine. Although this method entails a minimum expenditure of effort, the associated costs often greatly exceed those of mechanical (dispersion) methods.

2.3. Foam Stability

The economic impact of practical foam problems is evident from the number of experimental and theoretical studies devoted to this subject [8], [20]. Unfortunately, most of the usual physicochemical measurement techniques are unsuitable for the direct study of such complicated agglomerates of draining gas bubbles in a moving, foaming liquid. More limited theories attempt to correlate foam stability with changes in a single dominant parameter. Several fundamental issues related to foam stability have been clarified by studies conducted on isolated foam lamellae. This is a legitimate approach, because foam degradation is a mechanical process consisting of the rupture of large numbers of individual lamellae [10].

2.3.1. Gibbs Film Elasticity and Marangoni Effect

Except in the case of pure liquids, foam bubbles and lamellae are stabilized relative to small changes in bubble size and lamellar thickness by two closely related mechanisms, the Gibbs film elasticity and the Marangoni effect.

Gibbs film elasticity E applies to extended lamellae with thicknesses several orders of magnitude smaller than their surface dimensions [21]. Film elasticity is defined as the quotient obtained by dividing an increment in the surface tension ($2\,\mathrm{d}\gamma$) by the corresponding increase in relative surface area of the film ($\mathrm{d}A/A$):

$$E = 2A \cdot \frac{\mathrm{d}\gamma}{\mathrm{d}A} = -2A \cdot \frac{\mathrm{d}\pi}{\mathrm{d}A} \tag{3}$$

Extension of a lamella results in a corresponding reduction in the equilibrium surface concentration of adsorbed surfactant molecules and thus in an increase in the surface tension γ, which acts as a restoring force (Fig. 4).

An additional stabilizing restorative force that diminishes with time also exists, this is known as the *Marangoni mechanism.* The time required for establishing the Gibbs adsorption equilibrium is determined by the diffusion-controlled transport of surfactant molecules out of the bulk phase.

The rapidly expanding, but still relatively thick lamellae characteristic of a low-viscosity, aqueous, spherical-bubble foam often exhibit surface depressions. These are caused by localized expansion and corresponding reductions in the spreading pressure. Adsorbed monolayers in undisturbed regions spread spontaneously into such depressions and compensate for the low surface concentration [22], [23]. Furthermore, solvated hydrophilic groups of spreading surface-active agents aid in the transport of interlamellar liquid from nearby layers, which helps compensate for the destabilizing effect of local dilution [24]. Both mechanisms are depicted schematically in Figure 4.

2.3.2. Rheological Aspects of Foam Stability

Foams tend to lose their interlamellar liquid under the influence of gravitational forces. As the

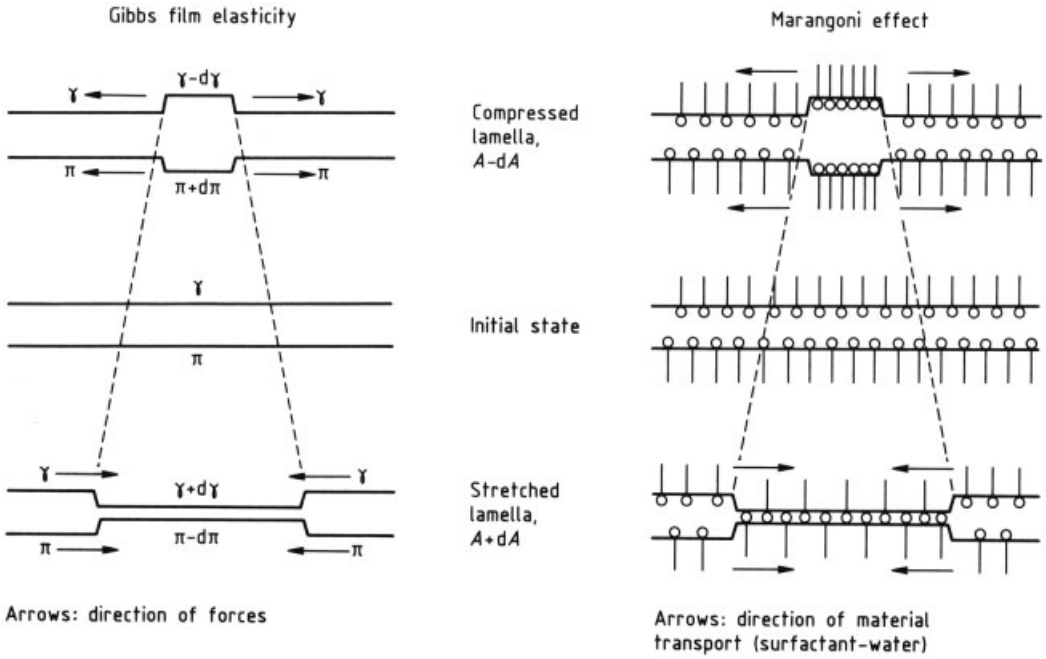

Figure 4. Schematic representation of the Gibbs and Marangoni foam-stabilization mechanisms

foam lamellae become thinner, their fragility increases. Thus, any factor causing resistance to flow should contribute to stabilization. Additionally, low-viscosity foams often exhibit yield points with important consequences from the standpoint of applications, as in the case of shaving cream [11].

Viscosity of the Interlamellar Liquid. An increase in viscosity diminishes the rate of escape of interlamellar liquid, thereby increasing foam stability and height [9], [18]. At the same time, however, high viscosity can also inhibit the process of foam formation.

Surface Shear Viscosity. Whereas three-dimensional rheology is the factor defining the dynamic viscosity of a bulk phase, the two-dimensional surface viscosity γ_s is responsible for the velocity gradient between parallel linear elements at a sheared surface [25]. Experimental determination of this parameter presents certain difficulties.

A modified couette system [26] is normally employed for measurements on non-Newtonian systems. The sample is placed in a rotating cup. An inverted bowl-shaped sensor is allowed to make concentric contact with the rotating liquid surface, thereby permitting accurate determination of the corresponding torque.

Absolute values can be determined even at elevated temperature, but their significance remains uncertain. Relative values are of more interest, particularly for surfactant mixtures, as are the significant increases in surface viscosity commonly observed with higher alcohols and other water-insoluble substances [27]. The expected proportionality between surface shear viscosity and foam stability has been confirmed to only a limited extent. Solutions of surface-active substances with a high surface shear viscosity do indeed exhibit significant foam stability, and foams in low-viscosity media subside rather rapidly; however, solutions with intermediate surface shear viscosities fail to yield satisfactory correlations.

Surface Dilatation Viscosity. Dilatation [28] and compression are modes of deformation that accompany pure shearing in certain types of flow. The surface dilatation viscosity K is determined with the aid of a modified film balance [29] by using the relationship

$$\Delta\gamma = K\frac{1}{A}\cdot\frac{\mathrm{d}A}{\mathrm{d}t} \tag{4}$$

The total surface of the surfactant solution under investigation is confined between two movable barriers whose position is varied such that the relative rate of expansion $d \ln A/dt$ remains constant. The resulting increase in surface tension $\Delta\gamma$ in the stationary zone between the barriers is measured tensiometrically by the Wilhelmy method [13].

Studies of the mechanism of lamellar rupturing suggest parallels between the stability of thin films and the corresponding surface dilatation viscosity [30].

2.3.3. Electrostatic Foam Stabilization

A fresh foam is a thermodynamically unstable system. However, thin lamellae can be transformed into a metastable state by electrostatic repulsion.

True Foam Lamellae. A solution that is supersaturated with a gas generally produces a spherical-bubble foam (see Fig. 1) [9], [11] containing considerable amounts of liquid. Progressive draining leads to rapid transformation of the spherical foam into a symmetrical polyhedral foam.

The lamellae within a polyhedral foam are separated from one another by surface regions exhibiting concave deformation as well as increased strength. These "plateau regions" [9], [10] have a destabilizing suction effect on the interlamellar liquid. An ideal unit cell of a uniform polyhedral foam resembles a regular pentagonal dodecahedron. However, true foams are ill-suited to the study of the opposing forces involved in progressive draining because of their instability.

Single Lamellae. Single lamellae can be withdrawn easily from a liquid with the aid of a glass frame [9], [10]. Unlike foams, they can be subjected to intensive investigation as long as they are protected from the effects of dust, evaporation, temperature change, and vibration. Lines of contact with the frame are characterized by plateau regions analogous to those encountered between the lamellae of a true foam.

Freshly drawn, vertically oriented lamellae are generally several micrometers thick. Liquid flows out of the center of such a lamella under the combined influence of gravity and the suction effect of plateau regions. A lamella stretched to a thickness between roughly 1000 and 100 nm displays a characteristic pattern of interference colors [9]. As soon as the lamellar thickness falls significantly below the wavelength of visible light, a "black film" results; which is so designated because of complete interference between incident and reflected light. Depending on a variety of factors, including the composition of the solution, lamellar shrinkage ceases at a thickness between 60 and 10 nm. The existence of such metastable black films implies that mechanical equilibrium is achieved between attractive and repulsive forces.

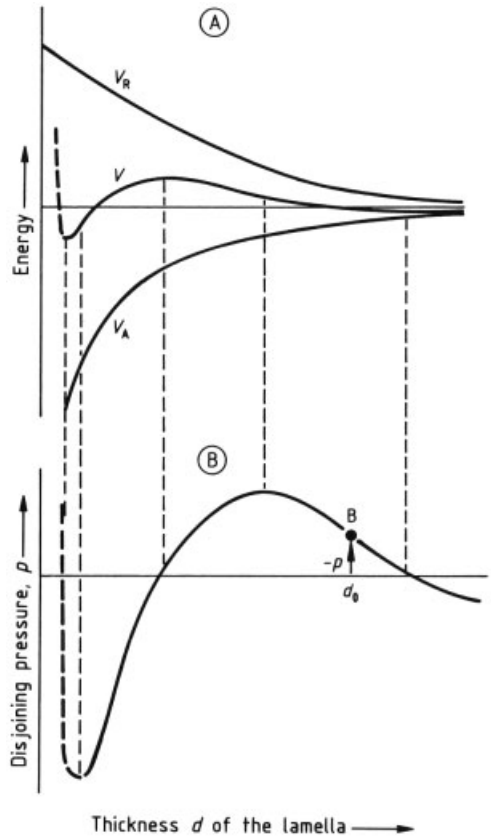

Figure 5. Force equilibria resulting in metastable lamellae [10]
A) Potential energy curves showing van der Waals attraction V_A, charge repulsion V_R, and the resultant energy V as a function of thickness d of the lamella
B) Disjoining pressure $p = -\delta v/\delta d$

A potential energy curve [10] can be used to illustrate the combined effects of repulsion caused by identically charged diffuse double layers adsorbed on the two faces of a surfactant film and the concurrent attraction due to van der Waals forces. Such a curve generally displays two minima (Fig. 5). Differentiation of the curve with respect to distance yields the *disjoining pressure* per unit area, a force which compensates in the metastable state (at equilibrium thickness d_0) for the destructive effect of suction.

The lamellar thickness corresponding to a black film represents the shallow secondary minimum on the potential energy curve. Conclusions derived from theoretical considerations can be largely confirmed at low salt concentrations [31], [32]. Thus, for a given disjoining pressure, the thickness of a metastable film decreases with increasing concentration of electrolytes (Fig. 6).

At higher electrolyte concentration, film thicknesses approach a limiting value of ca. 12 nm, a result which is essentially independent of the suction effect and may be due to additional separating forces of steric origin.

Metastable lamellae often transform themselves into "type II black films" with a thickness of only ca. 5 nm [10], [33]. These extremely thin "sandwich structures" probably consist of two surfactant films separated by a thin layer of water of hydration. They are thought to occupy the primary minimum in the potential energy diagram.

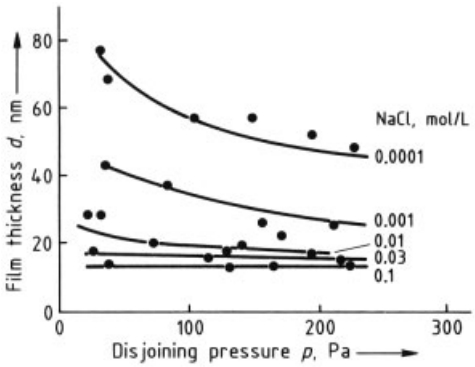

Figure 6. Disjoining pressure in a solution of sodium oleate (10^{-3} mol/L) at various sodium chloride concentrations

2.3.4. Mechanism of Lamellar Rupture

The surfaces of foam lamellae are under constant stress due to surface tension, a spontaneous force that tends to minimize surface extension. However, no direct relationship has ever been demonstrated between surface tension and foam stability. Rupture of the lamellae cannot, therefore, be regarded as a simple matter of overcoming some critical tear resistance; rather, it appears to be due to statistical fluctuations in lamellar thickness [34]. Reducing the thickness below a certain critical value ensures that the energy criteria for destabilization are met. Countless local instabilities can be expected to arise under practical working conditions as a result of vibration, temperature gradients, evaporation, and dust. Small particles of hydrophobic solids and liquids are particularly disruptive.

Lamellae containing relatively weak surface-active agents (alcohols or short-chain fatty acids) tend to decompose at an early stage of solvent exclusion and, therefore, never profit from the stabilization associated with thin black films. Instead, the concentration-dependent maxima observed for low-stability foams are attributed to the stabilizing effect of the Gibbs film elasticity (see Section 2.3.1) [35]. Lamellae associated with powerful surfactants rarely show such concentration-dependent maxima. Although the Gibbs stabilizing mechanism is inoperative, progressive thinning of the lamellae often terminates in a metastable black film rather than film rupture.

Many water-soluble polymers (e.g., gelatin and cellulose ethers) show very limited surface activity but nevertheless possess significant foam – stabilization power. The reasons for this are of rheological and especially steric origin. The most familiar example is the stabilization of beer foam by adsorbed protein – polysaccharide complexes.

2.4. Foam Inhibition and Destruction

A variety of chemical, biological, and technical processes are severely hindered by foam formation. Foams are controlled by chemical (Section 2.4.1), thermal, or mechanical (Section 2.4.2) means. In practical applications, foam inhibition, which aims to inhibit foam formation

in liquids, is distinguished from destruction of foams that have already formed. Thermal foam destruction primarily promotes bursting of enlarged gas bubbles and accelerates the destabilizing drainage of interlamellar liquids. However, this approach is often rather ineffective, and undesirable deterioration of the product is usually unavoidable.

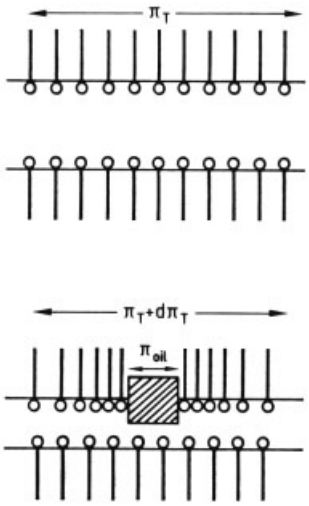

Figure 7. Effect of a spontaneously spreading oil on lamellar thickness (Marangoni mechanism)

2.4.1. Chemical Foam Inhibition and Destruction

Effective foam inhibitors and defoaming agents often spread spontaneously over liquid surfaces. Earlier studies [36] assumed the existence of a causal relationship between this characteristic and foam destruction, but more recent findings [37] demonstrate frequent deviations from any such correlation. A foam inhibitor may be introduced in either dispersed (e.g., oils or pigments) or molecular form. The principal mechanism of action of oils appears to involve local displacement of the adsorption film containing the foaming agent. Thus, spontaneous spreading is not a precondition for effectiveness. What matters is that the surface-active component in the foaming agent should not be adsorptively incorporated into the oil – water boundary surface. The stabilizing film is, therefore, perforated; the lamellae become seriously deformed; and the interfacial surface energy of the moistened portion of the oil droplet approaches a maximum. Intensive investigation of pure sodium dodecyl sulfate and a range of purified liquid foam inhibitors provides support for this mechanism [37].

Another approach attempts to explain the mechanisms and energetics of foam destruction as consequences of two consecutive spreading phenomena. Thus, if drifting oil particles within the bulk phase are to be effective on the surface of a foaming liquid, two fundamental energy criteria must be fulfilled [8]:

1) Hydrophobic materials, rising because of their buoyancy, must spontaneously penetrate the liquid surface and form droplets. In the process, an aqueous film is replaced by an oil – gas surface. This step is characterized by a negative spreading coefficient S_1, which implies a positive "entering" coefficient.
2) Floating oil droplets must then distribute themselves spontaneously over the lamellae that are to be destroyed ($S_2 > 0$). This second condition is best fulfilled by foam inhibitors whose surface and interfacial tensions are both low.

However, hydrophobic agents characterized by these two criteria (i.e., $S_1 < 0$ and $S_2 > 0$) often prove to be rather poor foam inhibitors or combatants. In particular, silicone oils are ineffective unless they have been "activated" with highly dispersed particles of hydrophobic solids [38].

Particles of hydrophobic silica alone have a destructive effect on polyhedral foams, although they show no tendency to inhibit the formation of wet spherical-bubble foams [38]. On the basis of numerous investigations [39], this activity is now assumed to be a consequence of the free hydrophobic surfaces becoming hydrophilic due to adsorbed surfactants. Silicone oils spread in most media and protect the surface of hydrophobic silica particles, thus they transport active silica to the breaking foam bubbles.

The converse of the Marangoni effect also has a destabilizing effect [10], [24] because it leads to the removal of interlamellar liquid (Fig. 7).

The foam-destroying properties of dispersed, hydrophobic solid particles (e.g., Teflon, silanized silica, solidified paraffins, or sulfur) can also be explained in another way [40]: as the particles penetrate an expanding liquid lamella, they contribute to its rupture by

markedly increasing the interfacial tension. This phenomenon can be viewed as a displacement of the aqueous phase resulting from a dewetting tension (cf. the condition $S_1 < 0$ in [8]).

Soap mixtures with a broad range of chain lengths (C_{12-22}) are often used in detergents based on alkylbenzenesulfonates and fatty alcohol polyglycol ethers [41], [42]. The foam-inhibiting action is due to calcium salts formed in situ (see Section 3.6). Therefore, builders must not completely remove the calcium ions.

2.4.2. Mechanical Defoaming

Mechanical defoaming is used when chemical foam inhibitors or defoamers might impair process or product quality [43]. Foam lamellae can be subjected to high shear forces with the aid of centrifuges, injectors and ejectors, and especially, rotating disks. The dominant foam-breaking mechanism here involves the rupture of individual lamellae, which becomes possible above a specific minimum shear velocity [44]. Optimization of this approach is complex [43–45].

3. Antifoaming Agent Composition

Traditionally, antifoaming agents were often single-component liquid systems or homogeneous solutions derived from vegetable oils or their constituent glyceride oils and fatty acids. Other useful materials included mineral oils and their derivatives [46], [47]. However, more complex formulations were also known, including aqueous sols derived from plant phosphatides [48], as well as other water-soluble or water-emulsifiable preparations [49].

Modern antifoaming agents (also called defoamers) are formulated to meet more diverse demands, including optimal effectiveness, easy application in an industrial context, convenient dosage, low volatility, and safety from both toxicological and ecological standpoints.

Most defoamers are composites rather than simple substances. Several groups can be distinguished:

1) Liquid single-component systems or homogeneous solutions
2) Dispersions of hydrophobic solids in a carrier oil
3) Aqueous or water-containing emulsions or suspension
4) Solid defoamer formulations

The functional components of a foam inhibitor can be divided into four categories:

1) Carrier oils
2) Active ingredients
3) Amphiphilic substances
4) Coupling and stabilizing agents

Foam-inhibiting emulsions contain water, which serves as the continuous phase [50]. Solid foam inhibitors are free-flowing powders in which the foam-inhibiting ingredients, as well as any emulsifying or dispersing agents, are bound adsorptively to an inorganic sorbent such as calcium silicate [51] or sodium triphosphate [52], or to a polymeric organic matrix such as methyl cellulose [53]. Particulate solid antifoaming ingredients suitable for incorporation into detergent powder compositions comprise a core that consists of an oily, solid, or gel-like carrier material. The core supports the antifoaming substances and is surrounded by a coating which is disrupted on contact with water to release the antifoaming substances [54].

3.1. Carrier Oils

The most important carrier oils are water-insoluble paraffinic and naphthenic mineral oils, together with vegetable oils such as tall oil [*8000-26-4*], castor oil [*8001-79-4*], soybean oil [*8001-22-7*], or peanut oil [*8002-03-7*]. These oils themselves have foam-control capability. Useful agents for prolonging antifoam effectiveness, which to some extent also show synergistic activity, include residues from the oxo alcohol synthesis, alkylbenzenes, and crude oils from the low-temperature carbonization of brown coal or other bituminous materials.

3.2. Silicone Oils and Silicone Foam Inhibitors

Silicone oils are particularly effective antifoaming agents because of their low surface tension, thermal stability, chemical inertness, and

total water insolubility [55], [56]. One of the largest applications for silicone oils is their use as antifoaming agents in the petroleum industry, particularly in gas – oil separation [57]. By far the most important silicone oils are the dimethylpolysiloxanes [*63148-62-9*], in which chain ends are saturated with trimethylsilyl groups. The number of siloxane units present ranges typically from 2 to 2000.

Dimethylpolysiloxanes by themselves show little foam-inhibiting activity with respect to aqueous surfactant solutions. Most silicone defoamers are in fact not pure silicone oils, but rather silicone formulations supplemented with finely divided hydrophobic solid particles. These particles presumably confer the actual foam-inhibiting properties, while the silicone oil acts as a hydrophobic carrier oil that offers unusually good spreading properties [58].

Silicone foam inhibitors are supplied in the form of anhydrous dispersions of pyrogenic or hydrophobized silica [*7631-86-9*] in silicone oil. Such mixtures are oily and cloudy. In some cases they have the consistency of a paste, but the most commonly used foam inhibitors for aqueous systems are 5 – 50 % emulsions. Silicone antifoaming agent are directed toward suppressing undesirable side effects silicone antifoaming agents such as the formation of "fish eyes" in surface coatings, investigators have studied the effect of introducing organic residues into the polymeric siloxane matrix [59–61]. Other goals have included the development of foam inhibitors containing only traces of silicone oil, as well as silicone antifoams based on hydrocarbon oils [62]. Certain polyether modifications of poly(methylsiloxanes) have achieved importance as foam stabilizers and cell regulators in polyether urethane foams (→ Polyurethanes).

3.3. Hydrophobic Silica

The development of hydrophobic silica [63] can be regarded as one of the most important innovations in the field of foam-control agents. Several methods have been perfected for conferring hydrophobic properties on this normally hydrophilic material [64], the most important of which are

1) spraying the silica with silicone oil and tempering at 250 – 350 °C,
2) treatment with organosilicon halide vapors in an autoclave, and
3) dispersing the silica in a silicone oil at elevated temperature and recovery of the solid by centrifugation.

Hydrophobic silica can also be produced by reacting silicia hydrogel with alcohol [65], [66], fatty amines, or wax. Amorphous, precipitated silica (particle size 1 – 2 μm) is the most suitable starting material for foam-inhibiting agents from the standpoint of both cost and performance.

3.4. Hydrophobic Fat Derivatives and Waxes

Hydrophobic fats and waxes include the following materials:

1) fatty acid esters of monofunctional and polyfunctional alcohols;
2) fatty acid amides and sulfonamides;
3) paraffinic hydrocarbon waxes [*8002-74-2*], ozokerite [*8001-75-0*], and montan wax [*8002-53-7*];
4) phosphoric acid mono-, di-, and triesters of short- and long-chain fatty alcohols (for the triesters, even tributyl phosphate [*126-73-8*] and tributoxyethyl phosphate [*78-51-3*] have foam-inhibiting properties [67]);
5) short- and long-chain natural or synthetic fatty alcohols [68], [69];
6) water-insoluble soaps of long-chain fatty acids, including aluminum stearate [*637-12-7*], calcium stearate [*1592-23-0*], and calcium behenate [*3578-72-1*] [70];
7) perfluorinated fatty alcohols [71].

3.5. Water-Insoluble Polymers

A number of polymeric substances have been reported to have antifoaming properties, including low molecular mass, fatty acid modified alkyd resins; low molecular mass novolaks; copolymers of vinyl acetate and long-chain maleic and fumaric acid diesters; and methyl methacrylate – vinylpyrrolidone copolymers. Other relevant polymeric materials include poly(propylene glycols) and high molecular mass propylene oxide adducts to glycerol

[*56-81-5*], trimethylolpropane (1,1,1-tris(hydroxymethyl)propane) [*77-99-6*], pentaerythritol [*115-77-5*], triethanolamine [*102-71-6*], dipentaerythritol [*126-58-9*], or polyglycerol [*25618-55-7*] [72–74].

Addition products of butylene oxide or long-chain α-epoxides with polyvalent alcohols show the same effects as propylene oxide polymers [75], [76].

3.6. Amphiphilic Components

The category of amphiphilic compounds includes antifoam components with varying water solubility whose foam-inhibiting effects are due to a variety of mechanisms. Examples include sodium oleate [*143-19-1*] and the hardened fish fatty acid soaps used as foam regulators in detergents [41], [42]. The activity of the latter depends on in situ formation of calcium salts, which restricts their application to water that has not been demineralized.

Some amphiphilic substances are water soluble at low temperature but flocculate upon warming beyond a particular point called the cloud point. Above the cloud point, these surface-active materials act as foam inhibitors. Their foam-breaking effectiveness reaches an optimum at the cloud point itself. Examples of this type are nonionic surfactants such as modestly ethoxylated fatty alcohols, fatty acids, rosin acids, fatty amines, and alkylphenol derivatives with HLB (hydrophilic – lipophilic balance) values < 10. (For further explanation of HLB values, see → Emulsions, Chap. 9.2.) However, the effectiveness of this group of amphiphilic substances can be increased markedly if the free, primary hydroxyl groups of the polyglycol ether portions are blocked through esterification or ether formation with a hydrophobic residue.

Silicone surfactants can also be classified as amphiphilic substances. Silicone surfactants are silicone oils to which polyether groups have been chemically bonded. Introduction of polyether residues increases the hydrophilic character of a siloxane; as a result, it is no longer totally water insoluble and, instead, possesses a cloud point [77].

3.7. Emulsifiers

Emulsifiers assure rapid distribution of active ingredients throughout an aqueous phase (→ Emulsions). Those surfactants already enumerated in the amphiphilic category fulfill this role to some extent, although more strongly hydrophilic, nonionic surfactants with an HLB value > 10 are also employed (e.g., ethoxylated sorbitan esters).

3.8. Coupling Agents

Coupling agents increase the mutual tolerance of the various components of a foam inhibitor and prevent gel formation or sedimentation during long-term storage at fluctuating temperatures. Most of these materials are glycols, low molecular mass alcohols, or other ingredients known as solubilizers such as naphthalenesulfonate [*1321-69-3*] or *p*-toluenesulfonate [*657-84-1*].

4. Mechanical Means of Combating Foam

Foam problems can often be prevented or at least minimized by proper design of production facilities and careful adjustment of operating conditions. Examples include reduction of stirring rates, avoidance of vigorous boiling or sudden evacuation, and elimination of rough surfaces. Foam formation can also be reduced by careful control of venting operations, and by diminishing surface areas and the corresponding potential for uptake of air (e.g., by selecting tall vessels of limited diameter). The proper arrangement of heating devices can also have a beneficial effect, as can the choice of reactor construction materials. Thus, the presence of highly conductive materials in the region above the fill level of a reactor vessel can result in more rapid dissipation of foams with low stability.

Foam can sometimes be countered by spraying with finely divided water [78], introducing air (perhaps supplemented by carbon dioxide), use of a spark discharge [79] or balancing electrical charge [80]. Other mechanical approaches to dealing with foam include the removal of gases or gas-laden liquids by means of

a porous filter. This method is particularly applicable to fermentation processes and, under appropriate conditions, to the monomer-stripping operation of vinyl chloride used in the manufacture of poly(vinyl chloride) [81]. Modern computer technology has created new means of optimizing the rate at which the defoamer is supplied to the processing vessel [82].

5. Foam Problems in Specific Applications

5.1. Detergents

See →Laundry Detergents, Chap. 3.4.3. and [83].

5.2. Food and Beverage Industries

5.2.1. Cleansing and Disinfecting

Any surface that comes in contact with food or beverages during their manufacture, processing, or transport must be cleaned and disinfected at regular intervals. Because the affected industries (e.g., breweries, soft drink manufacturers, dairies, meat packers, the sugar and preserves industries) are confronted with very different problems of residue removal, a wide range of processes, cleansing agents, and disinfectants is required.

The ever-increasing role of mechanization and automation in industry has focused increased attention on potential foam problems. In particular, inconvenient foam formation must be anticipated in all of the following activities:

1) CIP ("cleaning in place") cleansing and disinfecting of tanks, if the active solutions are to be recovered and stored for reuse;
2) cleansing and disinfecting drums, containers, milk collectors, and other vessels if active solution is sprayed over the contaminated surfaces;
3) cleansing of returnable bottles, regardless of whether the automatic bottle-cleaning assembly utilizes immersion or spray technology; and
4) lubrication of bottle transport lines.

5.2.1.1. Tank Cleansing Based on the Cleaning-in-Place (CIP) Method

The CIP technique involves application of a recirculating cleansing or disinfecting solution onto the walls of a soiled tank by means of spray heads or targeted nozzles. Generally, the process is conducted at low pressure. Recovered cleansing solution is led into storage tanks to permit its reuse (Fig. 8).

Foam may cause air to be drawn into the circulating pumps, leave residues that are difficult to remove, or result in solution losses due to reservoir overflow. Residues that can produce foaming in the tanks include

1) proteins and fats (dairy industry);
2) proteins, polysaccharides, tannin or polyphenols, hop resins, and bittering agents (breweries); and
3) pectin, cell components, ethereal oils, and polysaccharides (vegetable and fruit juice processors).

Components of the cleansing or disinfecting agent itself may also present problems. These include nonionic, anionic, and cationic surfactants (e.g., quaternary ammonium compounds).

5.2.1.2. Cleansing of Returnable Bottles

A distinction is generally made between two fundamental types of bottle washers: those that utilize separate soakers (soaking baths) and those that combine soaking with spraying (Fig. 9).

Machines of the latter type are subject to foam problems both when the bottles are initially flooded with cleansing solution and later during spray cleaning. Foam arises as a result of the following contaminants.

1) soaps and surfactants originating in lubricants applied to the conveyor belt;
2) cellulose, sizing, surfactants, dyes, and pigments derived from bottle labels;
3) dextrin, casein, and vegetable gum, as well as synthetic glue, again from bottle labels; and

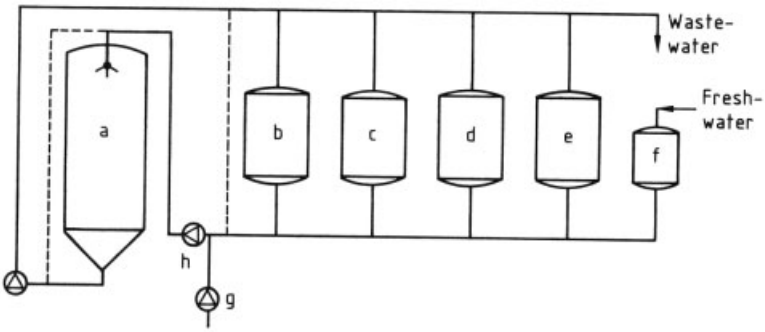

Figure 8. Cleaning-in-place (CIP) system for cleansing and disinfecting tanks
a) Soiled tank; b) Recycle water; c) Alkaline wash solution; d) Acid; e) Disinfectant; f) Fresh water; g) Alkaline solution pump (precleaning); h) Pressure pump

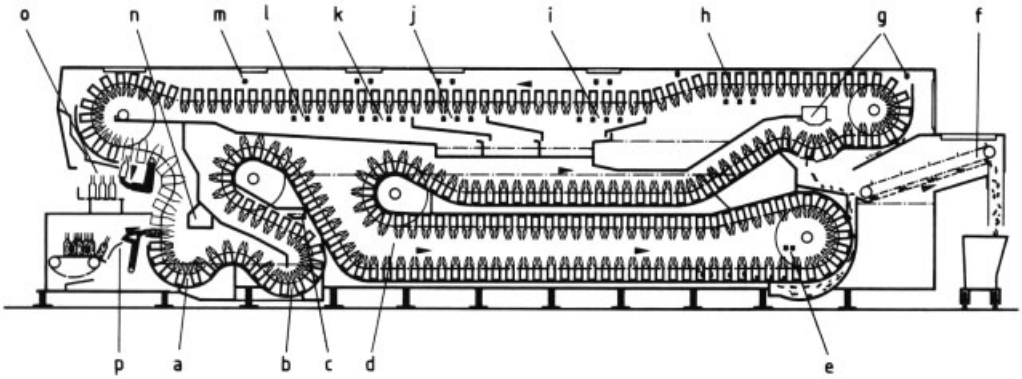

Figure 9. Automatic bottle-washing unit employing a combination soaking – spray system (Riwella-RK, Seitz Enzinger Noll Maschinenbau, FRG)
a) Preheating stage 1; b) Preheating stage 2; c) Preheating stage 3; d) Caustic 1 submersion soaking bath; e) Immersion flooding; f) Label removal assembly; g) Above-surface flushing; h) Wash II–interior (moving) and exterior spray; i) Warm-water intermediate spray, internal and external (internal synchronized); j) Warm-water spray, internal and external (internal synchronized); h) Cold-water spray, internal and external (internal synchronized); l) Fresh-water spray, internal synchronized; m) Fresh-water spray curtain; n) Guided vapor removal system; o) Bottle feed; p) Bottle discharge

4) sugar, protein, and fat residues from the original beverages.

Intensive foaming in a bottle-cleaning line can result in overflow of the cleansing baths and can hinder the discharge of bottle labels. The labels tend to float in or on the foam. They gradually disintegrate and may eventually clog the spray nozzles. Foam also causes the recirculating pumps to draw a mixture of detergent solution and air, which in turn reduces the effectiveness of the bottle-flooding operation. Label fragments carried into the rinse baths can cause further clogging, and bottles may exit the machine contaminated with cleansing solution.

The most important characteristic of a foam inhibitor intended for use in a bottle-washing system is that it be stable to alkali. In some countries, including the Federal Republic of Germany, such agents must also be biodegradable (to the extent of 80 %, according to the German surfactant law). The latter requirement has been addressed in 1987 by the introduction of nonionic surfactants bearing specially designed protective end groups [84].

5.2.2. Food Processing

5.2.2.1. Sugar

Modern methods for processing sugar beets are fully mechanized and automated (→ Sugar).

The raw beets are first washed and chopped. The sugar is then extracted with hot water in extracting towers. The crude extract, which also contains nonsugar components of the beets, is treated first with limewater and then with carbon dioxide. The "lime sludge" includes most of the nonsugar components and is removed by filtration. The filtrate is subjected to multistage evaporation–ultimately under vacuum–leaving a supersaturated crystalline broth, the massecuite. Sugar is finally separated by centrifugation from the residual syrup, known as molasses.

More or less intensive foam formation is observed in each of the above steps [125]. The foam is attributed to the numerous nonsugar materials present, such as cellulose, lignin, protein, vegetable bases (betaine and choline), and especially, saponin.

Foam controllers employed in the beet-washing process are likely to appear in wastewater; therefore, biodegradability is an important prerequisite. Because sugar is intended for human consumption and trace amounts of antifoaming agents may be present in the finished product, various legal and health issues are involved (see Chap. 7) [126]. Furthermore, steam-volatile components must be avoided during the evaporation and boiling steps, as must any material that might leave deposits which would diminish the effectiveness of the heat exchangers. Foam-control agents must also not inhibit the ion-exchange resins commonly employed to remove residual hardness from the dilute syrup. Finally, antifoaming agents must not affect subsequent use of the molasses in yeast production (see below).

5.2.2.2. Yeast

Baker's yeast is prepared commercially in massive (up to 200 000-L) fermentation tanks. The most commonly employed nutrient is the molasses obtained as a byproduct of sugar refining.

A large amount of air is required to ensure rapid multiplication of yeast cells; it is generally introduced through a series of tiny nozzles. After 12 – 14 h, the yeast is ready for harvesting. It is removed from the exhausted growth medium, reduced to the proper moisture content by means of a rotary filter, and pressed into tablets for final distribution.

The initial stage of yeast culture is an anaerobic process that is rarely subject to disruptive foam formation; by contrast, fermentation is aerobic, and is generally accompanied by extensive foaming. Foam suppressants are necessary to prevent overflow from the fermentor [127]. This heavy foaming is due mainly to residual nonsugar beet constituents that accompany the raw molasses (see Section 5.2.2.1).

In modern yeast production facilities, foam suppressants are introduced by means of automatic electrode-activated devices. The chemical defoaming preparation chosen must not interfere with the transport of oxygen because yeast growth would then suffer [128, 129]. In addition, the absence of detrimental effects on the final yeast cubes, such as premature softening or liquefaction must be ensured [130].

5.2.2.3. Potatoes

Whether the product is potato chips, French fries, mashed potatoes, or potato starch, potato processing is always beset by foam formation. A wash bath is always employed, and foaming can lead to serious processing difficulties. Starch manufacture, in particular, is plagued by intensive foaming in the potato juice, the starch slurry, and the processing water, caused by proteins and other nitrogenous compounds found in potatoes. Starch foam is very stable and difficult to counter. Mechanical methods of foam control have such limited effectiveness that the addition of foam-control agents may be regarded as the most practical and universally applicable solution [131].

5.3. Metal Treatment

5.3.1. Foam Problems Associated with Cooling Lubricants

Cooling lubricants mixed with water are used frequently in the metal-working industry in conjunction with operations that involve the removal of shavings or turnings (e.g., drilling, turning, milling, or grinding) as well as in such processes as wire drawing, rolling, and deep drawing.

Cooling lubricants are concentrates based on mineral oils and emulsifiers. The user dilutes

these with water to the desired concentration. Certain applications require the use of oil-free, transparent solutions. Both emulsifiers and corrosion inhibitors can induce foam formation, which in turn diminishes the filterability and cooling capacity of the lubricant and may lead to overflow. Foam behavior is, therefore, an important consideration in the selection of emulsifiers and other components for a cooling lubricant concentrate. However, considerations related to emulsion stability often so limit the choice that the only way of avoiding foam problems is to incorporate additional foam inhibiting agents.

A widely accepted approach utilizes calcium soaps as foam inhibitors. These are formed when the soaps present in the cooling lubricant concentrate react with hardening agents in the water used to dilute the concentrate. The calcium soaps are produced as a finely divided suspension of insoluble particles. The amount and type of soap must be adjusted to ensure an appropriate amount of residue and the correct particle size and to avoid disruptive buildup of sediment.

Cooling lubricants normally have a working lifetime of weeks or even months. As a result, external influences can alter their foam behavior and may require the addition of other foam inhibitors based on wax dispersions or silicone oil. However, these materials cling to metal surfaces, which can lead to irregularities in subsequent surface treatment. Their application must therefore be kept to a minimum and should preferably be continuous rather than on an as-needed basis.

5.3.2. Foam in Alkaline Cleansing Solutions

Alkaline cleansing agents are used primarily for difficult industrial cleaning tasks such as the removal of large amounts of oil, fat, or pigments. They are also employed in the final preparation of metallic surfaces prior to finishing. The combined action of the builders, the surfactants, and various additives removes pigments, metal shavings, metallic soaps, graphite, fats, and oils and helps in dispersing and emulsifying soil particles in the wash liquor. Such agents are widely employed in repair shops, enameling, galvanizing, electroplating, and sheet steel manufacture.

The major constituents of alkaline cleansing agents used in the metals industry are builders and surfactants. Builders are generally hydroxides or inorganic salts, such as silicates, phosphates, borates, or carbonates of sodium or potassium. The surfactants include anionic materials such as alkylbenzenesulfonates, and/or nonionic substances. Such products are often supplemented with other substances (e.g., chelating agents, corrosion inhibitors) in order to meet the needs of a particular application.

In many cases, foaming can be controlled by *cloud point adjustment.* The cloud point is defined as that temperature at which a second phase, enriched in wetting agent, forms; it is recognized by the onset of cloudiness. The foaming tendency decreases rapidly above the cloud point, but cleaning capacity also diminishes. For this reason, the cloud point should be set slightly below the expected application temperature. However, the fact that the cloud point decreases with increasing salt concentration and thus with increasing total concentration of the cleansing agent must also be taken into consideration. Careful choice of nonionic surfactants with various HLB values, perhaps in combination with appropriate concentrations of one or more anionic surfactants, ensures that a specific cloud point will be obtained at the dilution employed. If properly employed, cleansing solutions whose cloud point has been correctly set, are free of foaming problems even when large amounts of air are introduced, as in spray applications.

If the surfactant concentration is not constant, e.g., due to contamination by extraneous surface-active substances, foam control by cloud point adjustment is no longer possible.

Extraneous surface-active substances may be introduced into a cleansing bath directly in the form of soil on the objects being washed or indirectly as a result of chemical reactions with the alkaline solution. Relevant soils include nonionic emulsifiers as well as fatty acids and other long-chain carboxylic acids present either as anionic emulsifiers or because of their effective lubricating properties. Carboxylic acid esters, which are often included in lubricating agents as high pressure (HP) additives (→ Lubricants and Related Products), are saponified in the alkaline cleansing solution to alcohols and high-foaming soaps. The rate of hydrolysis increases with increasing alkalinity, temperature, and concentration of cleansing agent as well as with rising emulsification potential. An example of the

problem is the cleansing with orthosilicate solution of cold-rolled black plate contaminated with palm oil.

In this case, introduction of specific foam-control agents on an as-needed basis may be advantageous. Foams due to soap can be controlled by soluble calcium salts, perhaps in the form of hard water added to make up for evaporative losses. Substances belonging to the category of ethylene oxide – propylene oxide block polymers [*9003-11-6*] are also commonly employed. Silicone oils are inappropriate for applications involving the cleaning of metals because they are likely to interfere with subsequent surface treatments.

5.3.3. Foam Control in Neutral Cleansers

Neutral cleansers are liquid industrial cleansers that provide temporary corrosion protection. Their basic chemical constituents are nonionic and cationic surfactants, as well as organic corrosion inhibitors. Neutral cleansers are used preferably in automatic spray washers at temperatures between 5 and 90 °C to clean iron or steel, aluminum, nonferrous metal, and plastic parts after cutting or noncutting working [85]. The aqueous solutions have a pH of ca. 9. Foam produced by vigorous mechanical action during spraying can be controlled by cloud point adjustment (see Section 5.3.2) of the nonionic surfactants; foam-controlling nonionics based on mixed ethers are now preferred for both ecological and toxicological reasons [84]. The cationic surfactants present in neutral cleansers also inhibit the formation of foams caused by extraneous anionic surfactants during spraying and thus permit recycling of the cleansing solution [86], [87].

5.4. Polymer Industry

Foam problems are common in the production of plastics, synthetic rubber, and polymer dispersions. Foaming may accompany not only the actual polymerization step but also the recovery of unreacted monomer and the packaging, transfer, and transport of aqueous dispersions or latices.

Foam problems during polymerization are particularly common when heat generated by the reaction is removed by refluxing low-boiling monomers, e.g., in the preparation of poly(vinyl acetate) or poly(vinyl chloride).

Emulsion (E) polymerizations have a greater tendency to foam than suspension or dispersion polymerizations. Weakly foaming anionic emulsifiers [88], [89], oligomeric carboxylates [90], and low foaming nonionic coemulsifiers [91] have all been proposed as agents for the prevention of foaming during emulsion polymerization. The correct choice of the primary emulsifier for the polymerization is also important because substances in this category vary widely in their foaming tendencies (Fig. 10).

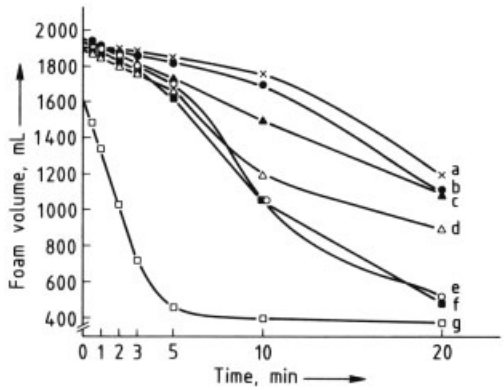

Figure 10. Foaming behavior of anionic primary emulsifiers at 25 °C
a) Nonylphenol + 4 EO sulfate; b) Lauryl alcohol + 4 EO sulfate; c) Lauryl alcohol + 50 EO sulfate; d) Lauryl alcohol + 30 EO sulfate; e) Nonylphenol + 10 EO sulfate; f) Lauryl alcohol + 10 EO sulfate; g) Nonylphenol + 25 EO sulfate
EO = ethylene oxide.

Differences are also observed, in various suspension (S) polymerizations, with regard to foaming. For example, S-PVC charges containing poly(vinyl alcohol) as a protective colloid foam more readily than those in which cellulose ethers are used to stabilize the suspension. Foam is commonly combated at this stage by spraying the stirring vortex with antifoaming agents, but mechanical means are also employed, including spraying with water [92] or bursting the bubbles with a propeller [93].

In the synthesis of both poly(vinyl chloride) and synthetic rubber, the polymerization reaction is usually terminated before complete conversion which necessitates the recovery of un-

reacted monomer. In the case of synthetic rubber, the latex is freed from monomeric butadiene in large decompression vessels. Less volatile monomers such as styrene are removed by passing the latex through a column against a flow of rising steam.

A significant improvement was introduced in 1965 by BASF, which resulted in simpler, quicker, more efficient monomer recovery. In the BASF process, steam is passed at high velocity through the boiling polymer dispersion to produce a steam – dispersion – foam mixture that carries residual monomer with it. This mixture is then subjected to rapid decrease in pressure, during which the foam is destroyed. Steam is separated from the broken dispersion foam, and the latter is returned to the boiling polymer dispersion [94].

In the case of S-PVC, demonomerization [95] involves either the use of a multicompartment vessel equipped for countercurrent vapor removal [96] or countercurrent vapor removal through a continuously operating column [97]. For E-PVC, the method of choice is spraying in conjunction with a pressure gradient [98].

Foaming during monomer removal [99] is caused by surface-active protective colloids or emulsifiers. Foam control by purely mechanical means may suffice, but in most cases antifoaming agents are employed. Among the criteria used in choosing them are

1) effectiveness with respect to the nonionic protective colloids in the case of suspension polymerization;
2) effectiveness with respect to the anionic emulsifiers in the case of emulsion polymerization;
3) stability to acidic or alkaline conditions, depending on the method of PVC manufacture employed; and
4) long-term stability and activity consistent with multistep, continuous stripping operations.

Silicone antifoaming agents are generally avoided, particularly for poly(vinyl chloride) destined to be made into film or where foam might be expected to play a role in further processing.

The choice of possible antifoaming agents for use in the manufacture of poly(vinyl chloride) or for the preparation of polymer dispersions is often severely restricted by the fact that proposed ingredients may have to be approved by a national health agency such as the FDA (United States) or the Bundesgesundheitsamt (Federal Republic of Germany).

5.5. Paint and Coating Industry

In the paint industry, foam formation can interfere with various processing steps. In addition, inadequate foam control (or use of the wrong foam-controlling agent) can negatively influence the decorative value or surface-protective function of the finished product. Foam problems may be encountered during pigment dispersion, introduction and mixing of the pigment dispersion and binder, transfer of the finished paint into cans, or final application–whether by brush, sprayer, or dipping [100].

Paint formulations are prepared in some types of stirring apparatus by blending a mixture of pigments and fillers with water to form a paste. The water is supplemented with dispersing agents [101], wetting agents, or a suitable combination of the two.

Dispersion is most commonly effected today with the aid of a dissolver. The shear forces required, coupled with the effects of dispersing and wetting agents, result in formation of a prodigious amount of foam, which is further stabilized by the presence of thickening agents such as poly(vinyl alcohol) or hydroxyethyl cellulose. The nature of the pigments themselves is also a factor. Finely divided pigment particles (e.g., carbon black or phthalocyanine blue) tend to promote foam formation. Foam control is complicated further by the presence of iron oxide pigments, which have a tendency to adsorb and thus neutralize the effect of antifoaming agents.

In the paint industry, the antifoaming agent should be tailored to the specific application at hand [102], [103]. Screening tests should be performed to verify that a particular antifoaming agent is effective and application tests should be conducted in its presence [104].

A pigment dispersion containing the defoamer is applied with a sponge roller to a piece of molded fiberboard, whose smooth side has been sealed with the appropriate polymer dispersion. After the paint has dried, the appear-

ance of the coating is compared with samples comprising a graduated test scale (Fig. 11). A given antifoaming agent must be evaluated further for microfoaming, a particularly frequent problem with modern aqueous coatings. This is done by examining paint applied to a glass plate and exposed to incident light from two neon tubes [103].

Air introduced into a nonaqueous paint produces a spherical rather than a polyhedral foam (see Chap. 2). If insufficient time is available for air to escape during the drying process, cavities or pinholes may form during baking: e.g., rapid evaporation of solvent near the surface may result in the surface becoming closed, and bubbles entrapped in the lower layers may burst at a later stage in baking when the coating has become so viscous it cannot flow into developing vacancies. The most effective remedy is incorporation of high-boiling solvents or solvent mixtures.

The decisive factor, with respect to foam behavior and foam control for water-borne coatings (e.g., acid-functional acrylic polymers cross-linked with melamine) is the bulk viscosity of the mixture, which is often dictated by the proposed application conditions [105]. The increasing use of such coatings in the automotive industry has resulted in concern about foaming during the removal of paint residues. These residues (oversprays) are withdrawn from the spray booth through a circulating water system. Special coagulating agents are added to cause pigment agglomeration, after which the pigment is precipitated with the aid of a flocculant (Fig. 12). Surface-active substances tend to cause foam formation in the constantly recirculated water; this, in turn, may clog the circulating system and reduce the effectiveness of the precipitating steps. In severe cases, the coating process itself can be affected [106]. Special silicone-free antifoaming agents have been developed exclusively for this application.

5.6. Construction Industry

Since 1945, polymer dispersions have played a rapidly increasing role in the construction industry. The plasticizing influence of such dispersions permits a reduction in the amount of water required to make cement and similar materials workable. More important, however, the addition of polymer dispersions to mortar leads to greater tensile strength in bending, and plaster containing both polymers and mineral binders has better adhesion and less tendency to crack.

Increased amounts of emulsifiers and dispersing agents are required to stabilize polymer dispersions. These increase the risk that foam produced during mixing of the latex and the cement or mortar will result in undesirable entrapment of air bubbles, thereby reducing the ultimate stability of the solidified mass. Antifoaming agents utilized in this context must be stable to alkali and must not have an adverse effect on key properties relevant to the construction trade (e.g., nonflammability requirements) [107].

In some cases, powder defoamers are added directly to the cement or mortar mix [108], [109]; silicone antifoaming agents are also used [110]. Appropriate low-foaming dispersions utilizing antifoaming agents with good long-term stability could be offered by latex manufacturers themselves, thereby obviating the need for subsequent introduction of foam-control substances [111].

5.7. Adhesives Industry

Most industrial adhesives are applied mechanically. The entrapment of tiny air bubbles during application may change the material flow characteristics and result in an uneven or damaged adhesive layer along with a corresponding reduction in the tensile strength of the adhesive bond.

The foaming tendency and the extent of foam formation depend upon the nature of the binding agents and other components added to the formulation. Natural binders such as starch, soy protein, cellulose ether, and natural rubber either are inherently surface-active materials or contain impurities that are surface active, such as saponin (in starch) or lipoproteins (in natural rubber latex). Synthetic binders contain surface-active stabilizers. Soaps based on fatty acids or rosin acids are present in synthetic rubber latex; poly(vinyl acetate) dispersions contain hydroxyethyl cellulose or poly(vinyl alcohol); synthetic anionic or nonionic emulsifiers are found in acrylate and methacrylate polymers. Foam stability is enhanced by thickeners such as methyl

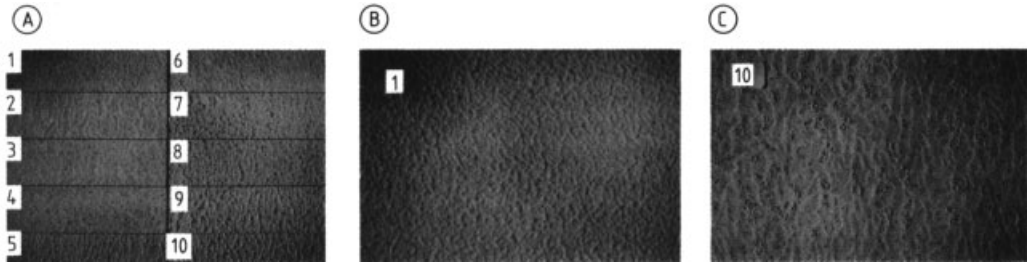

Figure 11. Comparison scale for evaluating the effects of foaming during application of an emulsion paint
A) Comparison scale; B) Grade 1; C) Grade 10

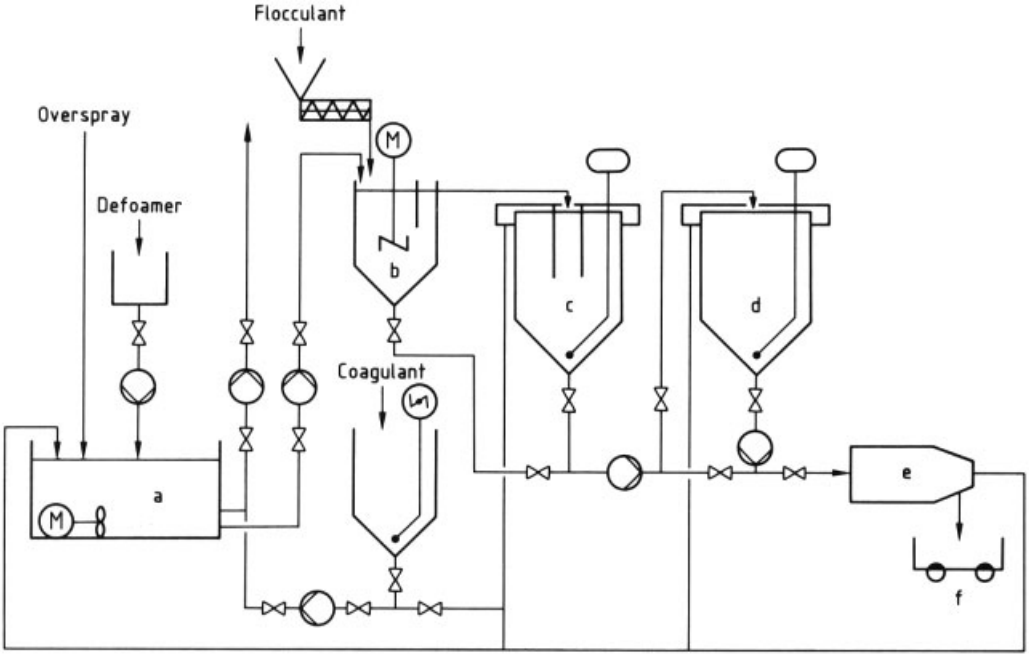

Figure 12. The ESKA (Entsorgung von Spritzkabinen Anlagen) system for removal of residual paint in the automotive industry (Dürr Anlagenbau, FRG)
a) Coagulation tank (pit); b) Flocculation tank; c) Settling tank; d) Sludge buffer; e) Decanter; f) Sludge cart

cellulose or polyacrylate, which increase the viscosity of the mixture and thus inhibit the degradation of foam once it has been produced.

Antifoaming agents for adhesive applications must be chosen for their suitability with respect to the particular surface-active agents in question. Adhesives that may come in contact with food or the human body must meet relevant legal requirements (Chap. 7).

Under certain conditions, foam can be turned to technical and economic advantage. For example, controlled foam formation in the Nordson Foam Melt process [112] results in hot melts with prolonged open time as well as improved flow and filling characteristics and more rapid setting. The amount of adhesive required is also reduced.

5.8. Textile Industry

Most chemical processes associated with the textile industry occur at the fiber surface and in aqueous medium. Surfactants improve the wettability of fiber surfaces, help displace adsorbed air, aid in the emulsification and dispersion of textile auxiliaries, and facilitate removal of soil from fabric surfaces (→ Textile Auxiliaries).

The increased rate of machine operation and material transport in modern textile mills implies an enormous expenditure of mechanical energy. Foam formation must be anticipated in any operation in which air is permitted to enter the treatment baths. The detrimental effects of foam may be qualitative, as in the case of "spoiled" goods (fabrics showing irregularities in surface treatment, undesirable depositions, spotting), or quantitative in the sense of requiring more complicated machinery or reduction in output.

Three general approaches have been taken to solving the textile industry's foam problems:

1) Mechanical – technological solutions
2) Introduction of foam inhibitors
3) Use of low-foaming surfactants in the formulation of textile auxiliaries

In most cases, a combination of measures is required.

Pretreatment and dyeing are the most important and critical textile processing steps with respect to foam control, but other stages in the process (e.g., fiber manufacture or fabric sizing, imprinting, and finishing) also present foam problems.

5.8.1. Pretreatment

The processing of cotton and spun rayon involves alkali treatment and mercerization; fabric is subjected to various concentrations of sodium hydroxide in the presence of wetting agents that are usually anionic in nature. Especially with the "dry-in-wet" process, but also with "wet-in-wet" technology, some introduction of air is inevitable. The preferred method of foam control relies on phosphate esters of alcohols containing four to eight carbon atoms (e.g., tributyl phosphate). These esters are stable in solutions containing high concentrations of alkali and electrolytes.

Other pretreatment steps conducted in a strongly alkaline medium rich in powerful anionic and nonionic surfactants include boiling off (alkali extraction of cotton) and bleaching, particularly when hydrogen peroxide is used as the bleach. Both processes are amenable to either continuous or batchwise implementation.

Batch treatment is increasingly carried out in jet or overflow systems. If the processing equipment is not completely filled with liquor, a more or less intensive–but, in any event, disruptive–foaming can be anticipated as a result of the high flow rates and rapid material transport. Foam at this stage interferes with homogeneous treatment of the fabric, which has a tendency to float. In addition, mixtures of surfactants and antifoaming agents tend to separate as a result of the high shear forces present in the circulating pumps. Foam residues (including silicone defoamers) may also remain on the fabric and cause spotting. Therefore, low-foaming wetting agents should be selected from the outset, and silicone-containing foam control agents must be avoided. The usual alternatives are formulations based on tributyl phosphate.

Continuous processing in an open-width unit also presents severe foam problems, because surfactant may be present in concentrations up to ten times that required for a batch process. Rotor jet agitation of the treatment bath and spray jet equipment aggravate the situation.

Mixed ether formulations [84] represent one modern solution to such foam problems, especially in bleach baths. Their foam-inhibiting power is particularly noteworthy at bleach-bath temperatures above the cloud point of the additives.

The use of low-foaming wetting agents can largely obviate the need for supplementary foam control agents in a peroxide bleaching bath. Moreover, these surfactants are biodegradable. Hydroxyalkane ether sulfonates prepared by SO_3 sulfonation of mixed ethers of unsaturated fatty alcohols are especially suitable [113].

5.8.2. Dyeing

Dyeworks are the largest consumers of foam-control agents in the textile industry. Because water is the only feasible solvent for dyeing and all water-insoluble dyes must be introduced as dispersions, dyeing involves many surface-active auxiliaries such as wetting agents, dye dispersants, and leveling agents.

The foam problems associated with dyeing are comparable to those encountered during pretreatment (Section 5.8.1), in part because much of the processing machinery (both batch and continuous) is similar. The same applies to dye posttreatments, which include steps designed to improve fastness, as well as various finishing operations using fabric softeners (usually cationic)

or other processing aids (e.g., finishes for improving sewability).

Dyers often utilize special silicone antifoaming agents, but preparations devoid of silicones are also used.

5.8.3. Foam Application of Textile Auxiliaries

Even though foam is generally regarded as a disruptive influence in textile manufacture, it can also serve a useful function, as in the case of foam application technology [114], [115], particularly because this opens the way to energy-efficient drying. Foams also play a structural role in certain finished products such as carpet backing, and foam is useful in situations calling for application of a substance to only one side of a fabric.

Foam technology has become an established part of such operations as rotary screen printing, fabric coating, bonding of nonwoven fabrics, backing (upholstery fabrics), and certain types of dyeing.

5.9. Leather Industry

Almost every stage of leather processing–from rawhide to consumer product–involves contact or treatment of the leather with surface-active compounds. During processing of the hides in the beam house, and subsequent tanning, dyeing, and fat-liquoring, inconvenient foam is countered largely with the aid of silicone foam-control agents. In many cases, the necessary treatment aids contain appropriate foam-regulating additives.

The problems faced during leather finishing are considerably more complex. Many finishing agents contain surfactants or are applied in the form of emulsions (e.g., casein solutions, silicone emulsion for use as a feel agent, wax emulsions, nitrocellulose emulsions). Pigmented dispersions contain wetting or dispersing agents, and aqueous acrylic resins are polymer emulsions that incorporate anionic or cationic emulsifiers. Foam formation during finishing and coating can lead to serious process disruption. The choice of a suitable antifoaming agent is often difficult because it must be both highly effective and, at the same time, compatible with other aspects of the process; it must in no way lower the quality of the final finish [116].

5.10. Pulp and Paper Industry

The paper industry is one of the major consumers of foam-control agents. Foam is a problem at every stage from pulp and paper fabrication through coating and printing, as well as in the reprocessing of wastepaper and the treatment of wastewater.

5.10.1. Pulping

The alkaline conditions required by the sulfate pulping process result not only in the solubilization of lignin but also in the production of soap. The soap originates from resin acids found in the resin ducts of coniferous trees and from fats and resins located in the parenchymal cells of deciduous trees. The soap is responsible for the generation of considerable amounts of foam. Depending on their degree of dispersion, colloidal and finely dispersed cellulose particles and fibrils can stabilize this foam. The resulting fiber suspension is washed in multistage brown-stock washers at a temperature of more than 82 °C and a pH of ca. 12, to remove most of the residual cooking liquor. The antifoaming agents must therefore be effective at high temperature and alkali-stable [117].

In addition to the sulfate process, sulfite pulping is still of some importance, particularly in Europe. In this process, lignin is solubilized and extracted from the wood in the form of ligninsulfonate. Sulfonation is performed using a hot aqueous solution of sulfites or hydrosulfites, sometimes in the presence of excess sulfur dioxide. Sulfite pulping liquor is acidic, and foam generation is due to the presence of ligninsulfonates and other surface-active materials containing sulfonate groups. The resulting demands on any potential antifoaming agent are, therefore, different from those in sulfate pulping [118].

Foam controllers for pulping must be effective against large-scale foam production. In addition, they must be capable of dissipating gas entrapped in small bubbles stabilized by minute

cellulose fibers. This gas consists of air as well as carbon dioxide arising from fermentation of raw materials or chemical reactions associated with hard water. Insufficient attention to this problem results in a pulping mixture that is difficult to filter, and the washed pulp is likely to contain undesirable tall oil residues.

5.10.2. Pulping Liquor Disposal

Both economic and environmental considerations require that spent liquor from the pulping operation be subjected to further, extensive treatment. Evaporative measures are complicated by foam due to the presence of dissolved soaps derived from fatty acids and abietates. Such foams can be controlled with the aid of mixtures of organic solvents and nonionic surfactant [119]. Suitable wastewater defoamers can also be obtained by reacting fatty acid esters, fatty alcohols, or paraffinic hydrocarbons with liquid or gaseous sulfur dioxide [120].

5.10.3. Air Content of Pulp Suspensions

A common problem in paper manufacture is entrapment of air as a result of the turbulence associated with rapidly moving machinery. Dispersed air can lead to tiny bubbles in the paper or to foam on the surface, thus reducing the rate of subsequent moisture removal and lowering product quality. The seriousness of the problem can be reduced by incorporating degassing devices but high-speed processes require supplementary chemical measures to further suppress foaming and minimize air content.

Most of the standard wetting agents are unsuitable for degassing purposes. The most effective air release agents are slightly water-soluble, nonionic surfactants such as water-in-oil emulsifiers. Ethoxylation products based on fatty alcohols, fatty acids, or ricinoleic acid play a particularly important role [121].

5.10.4. Conversion to Paper

Foam formation in the paper-making machine results from mechanically introduced air that interacts with surface-active compounds present in added sizing agents. Foam of this type interferes with the formation of a closed fiber structure. It also inhibits drying and causes irregularities in the paper surface. Choice of an appropriate foam-control agent is difficult because of adverse potential effects on the sizing process, contamination of the rollers, or further disruption of the integrity of the paper surface [122].

5.10.5. Paper Coating

The binders utilized in paper coating are water soluble starch, cellulose ethers, and casein or aqueous polymer dispersions based on copolymers of styrene – butadiene and acrylic esters. The synthetic binders are particularly likely to cause foam problems and surface disturbances. Air cannot be excluded to a sufficient degree from the coating materials themselves; hence, the addition of chemical foam-control agents is common practice [123]. These differ from the antifoam preparations employed in paper making and must conform to the synthetic emulsions used in coating materials.

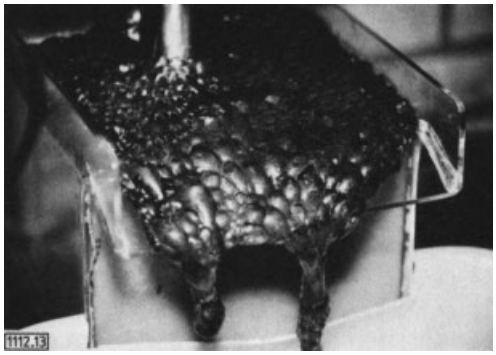

Figure 13. Deinking process
Flotation causes dislodged ink pigments to be collected by air bubbles and removed along with the foam.

5.10.6. Waste Paper Deinking

Air is deliberately introduced during the flotation deinking of paper waste. This induces the formation of air bubbles at the surface of the printing ink pigments and reduces the effective density of the pigment particles. The particles rise to the surface thus separating the ink from the fibers (Fig. 13). The solid particles in the foam are separated from the aqueous phase by centrifugation or passage through a belt filter press [124].

5.11. Phosphoric Acid Manufacture

The phosphoric acid used in fertilizers is produced primarily by the wet process in which phosphate-containing minerals are reacted with aqueous sulfuric acid. The accompanying release of carbon dioxide and silicon tetrafluoride leads to vigorous foaming, especially at the onset of the reaction. Foam may also be encountered during subsequent vacuum evaporation of the crude phosphoric acid. These foams are stabilized by insufficiently wetted, finely divided crystallites of the water-insoluble mineral gypsum. Powerful wetting agents provide a reliable solution to the problem. Anionic surfactants such as oleic acid, dioctyl sulfosuccinate, sulfonated tall oil, or mineral oil are especially useful [132]; they are often introduced in combination with fatty alcohols and nonionic surfactants (e.g., fatty alcohol and fatty acid ethoxylates). A mixture of a high molecular mass, water-insoluble, polyacrylamide flocculating agent and sulfonated tall oil fatty acids is particularly effective [133].

5.12. Wastewater Treatment

Wastewater entering a treatment plant is first freed of voluminous material by filtration through a coarse sieve. It is then passed through a series of channels designed to reduce its velocity so that sand and gravel settle and are left behind. The water then enters a large sedimentation container, from which the sludge is transferred into ponds or digesters.

The water emerging from a sedimentation unit still contains dissolved solids and colloidal particles derived from organic matter, and it may be subjected to further treatment by activated sludge. This requires that the liquid be brought into intimate contact with air, either by mechanical action or by using finely divided airstreams. Aeration is a rather violent process and may lead to serious foam problems.

Although foam is particularly undesirable at the stage of activated sludge treatment, it can also be encountered elsewhere in the process. In addition, foam buildup may occur in large bodies of water or streams. Such foam not only interferes with the control, observation, and management of water, but also reduces the effectiveness of water purification measures.

As the surface activity of the impurities present increases, so do the tendency toward foam formation and the stability of the resulting foam. Materials of low surface activity lead to formation of unstable foams, which dissipate without the need for special intervention.

Analysis of the water emerging from the treatment plant does not permit accurate determination of the cause of a particular foam episode. For this reason, foam-control agents are usually selected on the basis of laboratory experiments and field trials. Nonetheless, certain criteria can be applied in their selection. One of the criteria is the pH of the wastewater: an antifoaming agent that is effective only in alkaline medium is clearly of little interest if neutral water is to be treated. Second, an antifoaming agent that is to be introduced at the biological stage of wastewater treatment must not have a detrimental effect on the microorganisms responsible for purification. In addition, some antifoaming agents may have a positive or a negative effect on the flocculating process. Any foam control measures taken must be compatible with the particular flocculation procedure employed.

6. Testing Methods

Because of the variety of industrial processes in which foam formation occurs and the various effects of defoamers, three fundamental types of foam-related tests are used:

1) Tests to assess the foam-producing potential of a particular substance
2) Tests of the effectiveness of a given antifoaming agent
3) Tests to establish the effect of a particular antifoaming agent on the final product

Numerous standard test procedures have been developed. Usually, the foams are generated by one of the following techniques:

1) The test solution is confined within a cylindrical container and agitated with a perforated plate (DIN 53 902, p. 1).
2) A defined volume of the test solution is poured from a specific height onto the surface of the same liquid ("free-fall circulating

method" or Ross – Miles test; DIN 53 902, p. 2).

3) Air is introduced directly into a test liquid, through a capillary (DIN 51 381), a standardized paper disk (DIN 515 666), or an appropriate sintered-glass device (ASTM D 892-72).

In these procedures, either the time required to obtain a particular quantity of foam or the volume of foam produced under test conditions is measured. Additional information can be derived from measurement of the time required to dissipate the foam.

Additional methods for establishing the foaming capacity of surfactants, particularly those relevant to the detergent industry, are given in [12].

As many of the factors influencing foam formation as possible must be considered. Test conditions should be analogous to those expected in the field, especially with respect to temperature and the means of air introduction. Standard test conditions often have to be modified to ensure their relevance to a specific situation.

7. Legal Aspects

Numerous restrictions apply to the use of antifoaming agents in the manufacture of foods and other commercial products. Some take the form of prohibitive laws and ordinances, while others involve lists of permitted agents (positive lists). A multitude of regulations exist worldwide; the discussion that follows is confined largely to the Federal Republic of Germany and the United States.

Paragraph 11 of the German Food and Consumer Products Law of 15 August 1974 explicitly prohibits the use of additives in food products [134]. However, certain exceptions are permitted if the additives are removed completely prior to distribution or if they are removed as completely as possible and, at the same time, confer no undesirable taste or odor and pose no health risk. Examples of the unavoidable introduction of antifoaming agents are provided by the yeast, sugar, and potato chip industries.

The use of plastics in the context of food products is covered by recommendations of the Plastics Commission (Kunststoff-Kommission) of the Federal Health Office (Bundesgesundheitsamt, BGA), which must in turn comply with the Food and Consumer Products Law [134]. For example, recommendation XIV of the Bundesgesundheitsamt (Synthetic Polymer Dispersions) contains under the heading "Processing Aids" a section entitled "Antifoaming Agents."

The U.S. Code of Federal Regulations gives a list of permitted substances, along with their chemical constitutions and allowable levels in food products (Title 21, "Food and Drugs," § 173.340, Defoaming agents). Other relevant material includes defoaming agents used in coatings (§ 176.200) and defoaming agents used in the manufacture of paper and paperboard (§ 176.210).

Defoamers are also mentioned in § 240.1051 (Materials authorized for the treatment of wine). Those defoamers listed in § 176.200 are implicitly allowed under § 177.1200 (Cellophane) and § 177.1400 (Water-insoluble hydroxyethyl cellulose film), and the list in § 176.210 also applies to § 175.105 (Adhesives).

Under the auspices of the Joint Expert Committee on Food Additives of the FAO – WHO, efforts are being made to establish international guidelines. The use of foam-control additives in food products is thus subject to constant critical review with respect to both the nature and the amount of substances used.

8. Economic Aspects

Foam problems arising in industrial processes are a matter of immediate concern, because they can rapidly lead to serious economic consequences. For example, an automatic papermaking machine or bottle-washing unit may have to be shut down, and inadequate degreasing during metal treatment or surface irregularities resulting from foam or from use of the wrong antifoaming agent might lead to defective products. On the other hand, a certain amount of foaming may be permitted in a production line operating at less than full capacity to avoid the extra cost of foam-control measures.

These factors introduce a large measure of uncertainty into predictions of the market for antifoaming agents. The demand for foam-control agents in the United States has been estimated

at 136 000 t/a, and world demand is thought to be almost twice this value.

Table 2. Summary of commercial antifoam products

Trade name	Firm	Application
Afranil	BASF, FRG	paper industry
Agitan	Münzing Chemie GmbH, FRG	wide assortment for various industrial applications
Alkaterge	Angus Chemical Co., USA	fermentation
Antimussol	Sandoz, Switzerland	textile and paper industries, cutting oils
Antispumin	Chemische Fabrik Stockhausen, FRG	paper, textile, and sugar industries
Bayer-Entschäumer	Bayer, FRG	silicone defoamers
Baysilone-Entschäumer	Bayer, FRG	silicone defoamers
Bevaloid defoamers	Bevaloid Ltd., UK	wide assortment for various industrial applications
Byk	Byk – Mallinckrodt, FRG	paint and coating industry
Contraspum	Zschimmer & Schwarz, FRG	wide assortment for various industrial applications
Degressal	BASF, FRG	detergent and cleansing industry
Dehydran	Henkel KGaA, FRG	wide assortment for various industrial applications
Dow – Corning Antifoam	Dow – Corning, USA	silicone defoamers
Etingal	BASF, FRG	paper and textile industries
Foamkill	Lehmann & Voss, FRG	wide assortment for various industrial applications
Foammaster	Henkel KGaA, FRG	wide assortment for various industrial applications
Fumexol	Ciba – Geigy, Switzerland	textile industry
General Electric Silicone AF	General Electric, USA	silicone defoamers
Mazu	Mazer Chemicals, USA	wide assortment for various industrial applications
Nopco	Henkel KGaA, FRG	wide assortment for various industrial applications
Plurafac	BASF, FRG	low-foaming nonionic surfactants
Polymekon	T. Goldschmidt, FRG	paint and pigment industries, silicone defoamers
Ralulac	Raschig GmbH, FRG	emulsion paints, leather and textile industries
Repriman	Henkel KGaA, FRG	paper industry
Respumin	Bayer, FRG	silicone defoamers

Because of the multitude of industries serviced, the marketing of antifoaming agents necessitates a high level of specialization. Moreover, provision of an adequate level of customer service is also necessary. Suppliers of antifoaming agents differ markedly in their orientation and in their range of product lines. Thus, some suppliers specialize in specific types of material (e.g., silicone antifoaming agents), while others address the concerns of particular industries (e.g., suppliers to the textile industry). Table 2 lists commercial antifoaming products.

Table 2. (Continued)

Sentry Simethicone	Union Carbide, USA	pharmaceutical, cosmetic, and food processing industries
SAG-Silicone-Antifoams	Union Carbide, USA	silicone defoamers
Surfinol	Air Products, USA	low-foaming wetting agents
Tego-Entschäumer	T. Goldschmidt, FRG	paint and pigment industry, silicone defoamers
Tegosipon	T. Goldschmidt, FRG	paint and pigment industry, silicone defoamers
Texadril	Henkel KGaA, FRG	low-foaming wetting agents
Troykyd	Troy Chemical Corp., USA	wide assortment for various industrial applications
Viscasil	General Electric, USA	silicone defoamers
Wacker-Silikon-Antischaummittel	Wacker, FRG	silicone defoamers

9. References

General References

1. S. Berkman, G. Egloff: "The physical chemistry of foams," *Chem. Rev.* **15** (1934) 377 – 424.
2. S. Ross: *The Inhibition of Foaming Rensselaer Polytechn. Inst. Bull.,* Eng. Sci. Ser. No. 63, Troy, New York (1950).
3. T. Sasaki, "Foam phenomena," *Kogyo Kagaku Iasshi* **58** (1955) 809 – 814.
4. D. G. Dervichian: "Les Mousses: Formation, Stabilité, Destruction", Mémoires (Soc. Chimique de France, Conference, 25 March 1955), *Soc. Chim.* **5** (1956) 2 – 22.
5. *Houben-Weyl,* **I/2,** 167 – 185.
6. L. Louden: *Foam and foam Control,* The Institute of Paper Chemistry, Bibliographic Series No. 277, Appleton 1977.
7. H. T. Kerner: *Foam Control Agents,* Noyes Data Corp., Park Ridge, N.J. 1976.

Specific References

8. S. Ross, *Chem. Eng. Prog.* **63** (1967) no. 9, 41 – 47.

9. E. Manegold: *Schaum,* Straßenbau, Chemie und Technik Verlagsgesellschaft mbH, Heidelberg 1953.
10. H. Lange, *VDI-Berichte* **182** (1972) 71 – 77.
11. J. H. Aubert, A. M. Kraynik, P. B. Rand: *Spektrum der Wissenschaft,* **7** (1986) 126 – 138. Scientific American **254** (1986) no. 5, 58 – 66.
12. H. E. Tschakert, *Tenside* **3** (1966) 317 – 322, 359 – 365, 388 – 394.
13. J. G. L. Gaines: *Insoluble Monolayers at Liquid-Gas Interfaces,* Intersci. Publish, 1966.
14. J. W. Gibbs, *Trans. Connecticut Acad.* **3** (1876) 375.
15. K. Shinoda, K. Kinoshita, *J. Phys. Chem.* **63** (1959) 648 – 650.
16. M. J. Schwuger, *Kolloid Z. Z. Polym.* **233** (1969) 979 – 985.
17. H. Lange, M. J. Schwuger, *Kolloid Z. Z. Polym.* **223** (1968) 145 – 149.
18. M. J. Schwuger, W. D. Hell, *Prog. Colloid Polym. Sci.* **65** (1978) 6 – 76.
19. J. J. Bikerman: *Foams,* Springer Verlag, Berlin–Heidelberg–New York 1973.
20. A. Sheludko: *Colloidal Chemistry,* Elsevier Publishing Company, Amsterdam–London–New York 1966.
21. J. W. Gibbs: *Collected Works,* **vol. I,** Longmans, Green and Co., New York 1928, p. 300.
22. C. Marangoni, *Nuovo Cimento* **3** (1878) no. 3, 97, 193.
23. J. W. S. Rayleigh, *Proc. R. Soc. London* **48** (1890) 363.
24. W. E. Ewers, K. L. Sutherland, *Aust. J. Sci. Res. Ser. A* 5 (1952) 697 – 710.
25. A. W. Adamson: *Physical Chemistry of Surfaces,* Wiley-Interscience, New York 1976.
26. A. G. Brown, W. C. Thuman, J. W. McBain, *J. Colloid Sci.* **8** (1953) 491 – 507.
27. N. F. Djabbarah, D. T. Wasan, *Ind. Eng. Chem. Fundam.* **21** (1982) 27 – 31.
28. B. Stuke, *Chem. Ing. Tech.* **33** (1961) 173 – 181.
29. F. van Voorst Vader, T. F. Erkens, M. van den Tempel, *Trans. Faraday Soc.* **60** (1964) 1170 – 1177.
30. M. van den Tempel, Surf. Chem. Proc. Scand. Symp. **2nd 1964,** 306 – 315.
31. B. V. Derjaguin, J. A. Titijevska, I. I. Abricossova, A. D. Malkina, *Discuss. Faraday Soc.* **18** (1954) 24 – 41.
32. D. Exerowa, T. Kolarov, K. Khristov, *Colloids and Surf.* **22** (1987) 171 – 185.
33. J. S. Clunie, J. M. Corkill, J. F. Goodman, *Discuss. Faraday Soc.* **42** (1966) 34 – 41.
34. A. Vrij, *Discuss. Faraday Soc.* **42** (1966) 23 – 33.
35. O. Bartsch, *Kolloidchem. Beih.* **20** (1925) 1 – 2.
36. R. E. Pattle, *J. Soc. Chem. Ind. London* **69** (1950) 363 – 371.
37. S. Okazaki, T. Sasaki, *Tenside* **3** (1966) 115 – 118.
38. R. D. Kulkarni, E. D. Goddard, B. Kanner, *Ind. Eng. Chem. Fundam.* **16** (1977) 472 – 474.
39. R. E. Patterson, *Text. Chem. Color.* **17** (1985) no. 9, 181 – 184.
40. P. R. Garrett, *J. Colloid Interface Sci.* **69** (1979) 107 – 121.
41. Procter & Gamble, DE-AS 1056316, 1956 (W. L. St. John, W. L. Griebstein).
42. Procter & Gamble, DE-AS 1 080 250, 1957 (E. A. Schwoeppe).
43. M. Zlokarnik, *Chem. Ing. Tech.* **56** (1984) 839 – 844.
44. B. Furchner, A. Mersmann, *Chem. Ing. Tech.* **58** (1986) 332 – 333.
45. F. Müller, *Chem. Rundsch.* **28** (1975) no. 18, 3 – 7.
46. C. Philipp, *Allg. Oel Fett Ztg.* **39** (1942) no. 5, 167 – 170;no. 6, 203 – 207; no. 7, 235 – 239.
47. F. Ohl, *Seifen öle Fette Wachse* (1953) no. 5, 114 – 115; (1953) no. 6, 141 – 143.
48. Lehmann & Voß, DE 635881, 1935.
49. Chem. Fabrik Stockhausen, DE 694242, 1937 (A. Buch, K. Brunner, H. Lalleike).
50. R. C. Montani, F. J. Boylan, *Ann. Meet.–Techn. Assoc. Pulp Pap. Ind.* (1980) 293 – 300; Chem. Abstr. 92 (1981) 165525 n.
51. Diamond Shamrock, DE-OS 2 048 847, 1970 (M. D. Kellert, I. A. Lichtman).
52. Rhône-Poulenc Chimie de Base, EP 210 114, 1986 (D. Joubert).
53. Dow Corning, GB 892 787, 1962.
54. Unilever, EP 206 522, 1985 (R. Foret, H. T. Tai).
55. W. Noll, *Chemie und Technologie der Silikone,* Verlag Chemie, Weinheim (1960).
56. L. A. Cuenca, *Técnica Industrial* **5/6** (1976) 197 – 199. W. Kaiser, R. Riedle, *Winnacker-Küchler* **6/II** (1982) pp. 816 – 852.
57. P. G. Pape, *J. Pet. Technol.* **35** (1983) no. 7, 1197 – 1204.
58. G. Rossmy, *Fette-Seifen-Anstrichmittel* **71** (1969) 56 – 60.
59. G. Koerner, *Goldschmidt Informiert* **56** (1982) no. 1, 2 – 7.

60. Th. Goldschmidt AG, DE-AS 2 443 853, 1974 (H. F. Fink, H. Fritsch, G. Koerner, G. Rossmy, G. Schmidt).
61. Henkel KGaA, DE-OS 3 011 304, 1980 (M. Grünert, H. U. Hempel, H. Tesmann).
62. Union Carbide Corp., EP 121 210, 1984 (R. D. Kulkarni, G. E. Desmond, M. P. Aronson).
63. Hercules Powder Comp., US 3 076 768, 1963 (F. J. Boylan); US 3 408 306, 1968 (F. J. Boylan).
64. H. Ferch, A. Kreher, *Winnacker-Küchler,* **3/II** (1983) pp. 75 – 90.
65. DuPont de Nemours, US 2 801 185, 1952 (R. K. Iler).
66. W. R. Grace & Co., US 3 720 532, 1971 (E. A. Simpson, C. F. Doyle).
67. P. E. Levesque, N. S. Mount, *Plastverarbeiter* **18** (1967) 137 – 138.
68. Deutsche Hydrierwerke AG, GB 429 423, 1932.
69. Chem. Fabrik Stockhausen, DE-OS 3 001 387, 1980 (R. Peppmöller).
70. ICI, US 1 947 725, 1932 (A. Macarthur, A. Stewart).
71. Hoechst, DE 2 742 853, 1977 (F. J. Gohlke, W. Klinger).
72. Henkel KGaA, DE-AS 1 280 455, 1964 (H. Kasperl, G. Tischbirek, K. H. Worms).
73. Henkel KGaA, DE-AS 1 621 593, 1970 (R. Schoenemann, H. J. Schlüßler).
74. Henkel KGaA, DE-OS 3 531 212, 1985 (F. Lange, R. Höfer, K. H. Schmid, A. Asbeck, K. Herrmann).
75. Th. Goldschmidt AG, DE 3 201 478, 1982 (J. Fock).
76. Th. Goldschmidt AG, DE 3 201 479, 1982 (J. Fock, H. F. Fink).
77. G. Koerner, *Seifen öle Fette Wachse* **98** (1972) 335 – 338.
78. F. Ohl, *Chemische Rundschau* **4** (1951) 271.
79. R. Schnurmann, *Ind. Eng. Chem. Anal. Ed.* **11** (1939) 287.
80. Electro-Chemical Engineering Corp., US 1 984 210, 1930 (L. O. Gunderson). British Thomson-Houston Corp., GB 440 770, 1935.
81. Wacker-Chemie GmbH, GB 2 013 520, 1978 (O. Zajicek, J. Kirschner, H. Klapp, R. Rompeltien).
82. Int. ITT Corp., US 4 624 745, 1984 (W. E. Sande, B. D. Stroud, D. P. Dundie).
83. G. Jacobi, A. Löhr: *Detergents and Textile Washing,* VCH Verlagsgesellschaft, Weinheim (1987).
84. R. Piorr, R. Höfer, H. J. Schlüssler, K. H. Schmid, *Fette Wiss. Techn.–Fat Sci. Techn.* **89** (1987) 106 – 111.
85. J. Geke, *Metalloberfläche* **41** (1987) no. 5, 227 – 232.
86. Henkel KGaA, DE-OS 3 501 775, 1986 (J. Geke, H. Rutzen).
87. Henkel KGaA, DE-OS 3 247 431, 1982 (J. Geke, B. Stedry).
88. Soc. Belge de l'Azote et des Produits Chimiques de Marly, FR 977 296, 1948.
89. Chemische Werke Hüls, DE-AS 2 015 660, 1970 (U. Türk).
90. Uniroyal, DE-OS 1 745507, 1967 (L. E. W. Dannals, C. R. N. Peaker).
91. Henkel KGaA, DE-OS 3 530 301, 1985 (R. Höfer, A. Meffert, R. Piorr, B. Wegemund, U. Held).
92. Hoechst, DE 2 458 024, 1974 (H. Sonnenburg, B. Krämer).
93. Shin-Etsu Chemical Co., DE-OS 3 028 753, 1980 (S. Koyanagi et al.).
94. BASF, DE-AS 1 248 943, 1965 (K. Herrle, H. Fischer, A. Müller).
95. K. Flatau, K. Heidel in H. Felger (ed.): *Becker-Braun Kunststoff-Handbuch 2/1,* "Polyvinylchlorid", Carl-Hanser-Verlag, München, Wien 1985.
96. Hoechst, DE-OS 2 520 591, 1975 (C. Heinze, F. Bötsch, H. Wolff).
97. Wacker Chemie GmbH, DE-OS 2 757 168, 1977 (G. Nettesheim, F. Engel).
98. Chemische Werke Hüls, DE-OS 2 526 728, 1975 (J. Kalka, B. Hentschel, J. Ziebarth). Firestone Tire & Rubber Co., DE-OS 2 714 025, 1977 (D. M. Connor, J. M. Ogorzalek).
99. S. M. Eglund, *Chem. Eng. Prog.,* August 1981, 55 – 59.
100. H. J. Freitag, *Plaste und Kautsch.* **33** (1986) 72 – 76.
101. R. D. Athey, Jr., *J. Water Borne Coats,* August 1982, 7 – 24.
102. G. Newsome, *Paint & Resin* **52** (1982) no. 3/4, 33, 34, 38.
103. W. Gress, *Polym. Paint Colour J.* **174** (1984) 452 – 466.
104. J. Gärtner, *Defazet Aktuell* **26** (1972) 359 – 365.
105. P. Kuschnir, R. E. Eley, F. C. Floyd, *J. Coat. Technol.* **59** (1987) 75 – 87.
106. J. Geke, *Oberfläche JOT* **11** (1986) 43 – 46.
107. D. Zacharowski, *Polym. Paint Colour J.* **176** (1986) 850 – 858.
108. BASF, US 3 834 919, 1974 (E. T. Parker).
109. Diamond Shamrock, DE-AS 2 048 847, 1970 (I. A. Lichtman, M. D. Kellert).

110. T. Kawano, *Polym. Concr., Int. Congr. Proc., 3rd, 1981* **1** (1982) 147 – 162. *Chem. Abstr.* **98** (1983) 166012 d.
111. Hoechst, DE-AS 1 495 748, 1963 (G. Lorentz, G. Schultz, F. Girg).
112. "Das Foam-Melt-Verfahren," *Adhäsion* (1982) no. 6/7, 36 – 39.
113. R. Piorr, A. Meffert, *Fette Seifen Anstrichmittel* **87** (1985) no. 10, 391 – 394.
114. Denkendorfer Technologie-Berichte Nr. 3, *"Schaumauftrag,"* Konradin-Verlag, 1981.
115. G. M. Bryant, H. T. Walter, "Finishing with Foam," *Handbook of Fibre Science and Technology,* **vol. II, Part A,** Marcel Dekker, New York, Basel 1983.
116. J. Poré, *Rev. Tech. Ind. Cuir* **59** (1967) 328 – 338.
117. Associated Chemists Inc., US 4 032 473, 1975 (R. K. Berg, D. S. Smalley). R. T. Maher, "Water based brown stock defoamers," *Pulping Conf. Proc.* 1981, 461 – 470.
118. Nalco Chemical Comp., US 3 893 941, 1973 (R. J. Michalski, C. C. Cochrane).
119. K. Roberts, C. Axberg, R. Österlund: "Emulsion foam killers containing fatty and rosin acids" in *Foams,* An International Conference, Brunel University, Uxbridge 1975.
120. PWA Papierwerke Waldhof-Aschaffenburg, DE 2 509 626, 1975 (A. Höllerich).
121. R. Heyden, *ARP Allg. Pap. Rundsch.* **29** (1970) 1046 – 1048.
122. M. Humbert Rius, G. Campos Andreu, Prod. Quim. Aux Ind. Papelera (Curso), Barcelona (1982). *Chem. Abstr.* **101** (1984) 153641 p.
123. P. S. Hudson, *Pap. Technol.* **9** (1968) 185 – 186. H. F. Voss, *Wochenbl. Papierfabr.* **2** (1977) 51 – 56.
124. H. Ortner, R. F. Wood, H. Gartemann, *Wochenbl. Papierfabr.* **16** (1982) 597 – 601. H. Ortner, P. Schweiss, *Wochenbl. Papierfabr.* **10** (1982) 343 – 347.
125. Nalco, DE-OS 2 654 739, 1976 (R. Wachala, R. Svetic).
126. M. Potokar, H. Tesmann, A. Asbeck, *Zuckerindustrie (Berlin)* **107** (1982) 744 – 750.
127. Monsanto, US 2 666 017, 1950 (H. V. Moss, R. E. Morse).
128. J. I. Evans, M. J. Hall, *Process Biochem.* April 1971, 23 – 26.
129. B. Plichon, A. Decq, J. B. Guillaume, *Ann. Microbiol. (Paris)* **127 A** (1976)521 – 528.
130. Henkel KGaA, DE-OS 3 434 984, 1984 (A. Asbeck, K. H. Schmid, J. Schindler).
131. J. Malcher, K. Medal, *Stärke* **25** (1973) 346 – 349.
132. Cities Service Co., US 4 065 402, 1976, US 4 065 404, 1976, DE-OS 2 907 341, 1979 (W. A. Satterwhite, R. M. Leach, H. A. Stuhler).
133. United States Steel Corp., US 4 581 156, 1984 (D. M. Chatham).
134. Lebensmittel- und Bedarfsgegenständegesetz, *BG Bl I,* (1974) p. 1945.

Folic Acid → Vitamins
Food Colors → Foods, 3.Food Additives

Foods, 1. Survey

Amino Acids; Antioxidants; Beer; Benzoic Acid; Beverages, Nonalcoholic; Bread and Other Baked Products; Carbohydrates; Cellulose; Cellulose Esters; Cellulose Ethers; Cereals and Cereal Products; Cheese, Processed Cheese and Whey; Chocolate;Choline; Citric Acid; Coffee; Confectionery; Dairy Products, Imitation ; Dextran; Disinfectants; Enzymes; Fats and Fatty Oils; Flavors and Fragrances; Foods, 2. Food Technology; Foods, 3. Food Additives; Foods, 4. Food Packaging;Fructose; Gelatin; Gluconic Acid; Glucose and Glucose-Containing Syrups; Ice Cream and Frozen Desserts; Lactose and Derivatives; Lecithin; Margarine and Shortenings; Biocides; Milk and Dairy Products ; Monosodium Glutamate; Propionic Acid and Derivatives; Proteins; Seasonings; Single Cell Proteins; Sorbic Acid; Spirits; Starch and Other Polysaccharides; Sugar; Sugar Alcohols; Sweeteners; Tea; Tobacco; Vinegar; Vitamins; Wine; and Yeasts are separate keywords.

W. Frank Shipe, Cornell University, Ithaca, New York 14853, United States (Chaps. 2 – 5)

Charles G. Olentine, Watt Publishing, Mt. Morris, Illinois 61054, United States (Chap. 6)

1. Introduction 1
2. History of Food Production and Preservation 2
3. Components and Their Reactions . 4
3.1. Water 5
3.2. Proteins 6
3.3. Lipids 9
3.4. Carbohydrates 11
3.5. Mineral Components 14
3.6. Vitamins 14
3.7. Enzymes 17
3.8. Miscellaneous Components 17
3.8.1. Flavor Components 17
3.8.2. Natural Pigments 18
3.8.3. Undesirable or Potentially Undesirable Constituents 19
4. Food Law 19
5. Documentation 21
6. Feeds 21
6.1. Feed Definitions 22
6.2. Feed Utilization by the Animal . . . 23
6.3. Feed Components 25
6.4. Processing 25
6.5. Chemical Analysis 26
6.6. Quality Control 27
6.7. Survey of Common Feeds 28
6.8. Storage and Transportation 29
6.9. Economic Aspects 30
7. References 31

1. Introduction

Foods are mixtures of chemicals that are consumed by humans to satisfy their appetites for nourishment and pleasure. The nature and reactivity of the chemical constituents determine the properties of foods. The sensory properties, i.e., appearance, flavor, and texture, determine the acceptance of food and the pleasure derived from consuming it. Water-soluble components (salts, sugar, acids, and bitter substances) determine the taste, and volatile compounds determine the aroma. The texture is due to insoluble complexes or compounds such as proteins and polysaccharides; in some foods, lipids contribute to texture. Pigments are critical contributors to appearance. Most of the nutrients are provided by proteins, vitamins, and minerals, whereas carbohydrates and fats provide the energy. The ideal food should be both delicious and nutritious. Unfortunately, many nutritious foods are considered unappetizing by many people.

The following procedures or processes have been used to produce appealing, nutritious foods (for details, see → Foods, 2. Food Technology):

1) *mixing products,* e.g., fruit salads, vegetable salads, and fruit yogurts
2) *adding seasonings or other materials,* e.g., adding soy protein to sausage or smoking meats
3) *fractionating* to produce new products, e.g., separating milk to yield cream and skim milk

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_491

4) *homogenizing* to produce a more uniform product, e.g., peanut butter or milk
5) *fermenting* and *pickling* to produce new and more stable products, e.g., cheese manufacture
6) *enzymatic treatment,* e.g., enzymatic modification of starch to produce corn syrups of varying sweetness
7) *thermal processing* including cooking and baking, to destroy undesirable microbes, enzymes, and antinutrients such as trypsin inhibitors and enhance sensory appeal.

2. History of Food Production and Preservation

In the beginning, foods were selected from available natural products [1–8]. This led to the development of regional eating habits. The origin of modern economic plants was possibly as follows [1]:

Central Asia	apple, barley, broad bean, carrot, celery, cherry, cucumber, date, eggplant, lentil, lettuce, melon, mulberry, mustard, olive, onion, pea, pear, plum, pomegranate, quince, radish, rye, spinach, turnip, wheat
Mediterranean	artichoke, asparagus, cabbage, cauliflower, fig, horseradish, parsley, parsnip
Southeast Asia	banana, breadfruit, orange, peach, persimmon, rice, soybean, sugar cane, yam
Central or South America	avocado, cassava, corn, cranberry, kidney and lima bean, pineapple, potato, pumpkin, squash, sweet potato, tomato

At first, selection of food from the available supply was based primarily on sensory properties. By trial and error, people learned that some products satisfied their appetites while others adversely affected their health. The people who made the right choices lived to pass on the information to the next generation.

In the days when humans were merely hunters or gatherers of foods, the adequacy of diet depended on the types and quantity of available plants and animals. The need to develop methods for preserving food for use throughout the year became obvious. Concern over the quantity and quality of food increased as population increased. These food problems stimulated development of methods for producing and preserving foods (→ Foods, 2. Food Technology). The progression of civilization from the food gathering stage to the early food production era, i.e., from the Old Stone Age to 400 AD, is shown in Table 1.

In Germany, substantial increases in agricultural production occurred after the introduction of the iron plow in the 6th century, the horseshoe in the 9th century, and the horsecollar soon after. The threefold rotation system (one-third of the land remained fallow, summer crops were raised on another third, and winter crops on the other third) also increased production.

During the *Middle Ages* the variety of available foods increased as a result of increased traveling and trading. The discovery of America had a pronounced effect on the variety of food with the introduction of potatoes and corn, as well as other native American products, such as tomatoes, peanuts, lima beans, and turkeys.

The 16th century marked the beginning of the so-called *Golden Age of Science* [1], [6–10]. During this time, scholars investigated the nature and composition of foods. Some of the components discovered are listed in Table 2. These discoveries gave credibility to the scientific approach to biological problems and stimulated further research and technical developments (Table 3).

In the *last two centuries* considerable progress has been made in increasing the quantity and improving the quality of food [11], [12].

Food Quantity. Quantitative increases in the food supply have resulted from the following:

1) The use of better breeding practices and careful genetic selection, which led to the development of high yielding varieties of wheat and rice.
2) Replacing draft animals with mechanical power increased the acreage that could be farmed and released millions of hectares of land that had been devoted to raising feed for horses. However, mechanization of agriculture increased fossil fuel consumption.
3) Irrigation has converted vast areas from desert or semidesert to profitable food producing areas. Unfortunately, continuous irrigation has caused water logging and increased soil salinity.
4) The use of chemical fertilizers has created remarkable increases in production (Fertilizers). The gains have been partially offset by

Table 1. History of growing, processing, and preserving food [1]

Date (approximate) Period	Food	Farming–trading	Preserving–processing	Scientific–technological advances
Prior to 15 000 BC Paleolithic (second period of Stone Age)	eggs, fish, fruits, honey, insects, nuts, seeds, roots, small animals		drying, pounding, roasting	bags, baskets, clothes, language, "made" fire, painting, sculpture, stone and bone implements
15 000 BC Mesolithic (transitional period of Stone Age)	bigger selection of food, storing berries and wild fruits		boiling, drying fish, smoking, steaming, storing food	bow and arrow; dog, goat, reindeer, and sheep (domesticated); claycovered baskets
9000 BC Neolithic (last period of Stone Age) *villages*	domesticated animals, milk, butter, cheese, gruel, dates, olives, grapes, beer, vinegar, wine	cultivating cereals (seasonal) in permanent fields using hoes and hand plows, pruning	alcoholic fermenta-tion, adding acetic acid, salting, baking, making bread, sieving, pressing primitively, seasoning	pottery wheel; spinning; weaving; wood, flint, and bone sickles; saddle quern; mortar; fishing with hooks and nets
3500 BC Bronze *cities*	soybeans, figs, rice, olive oil, vegetables, lentils, cabbage, cucumbers, onions	irrigating, plowing with horses and oxen, much trading locally and externally	filtering, lactic acid fermentation, more types of flavoring, flotating, leavening bread, making sausage, frying, pressing (sophisticated and complex), clarifying	architecture, smelting, wheeled carts, ships, writing, bronze tools, mathematics, rotary millstones, bronze weapons, astronomy, shadoofs, medicine, chemistry
1500 BC Iron	artichokes, beans, fruits, lettuce, sauces, spices	trading by land and sea, using heavier plows	refinement of flavoring and cooking	pulleys, glass, improved and cheaper tools and weapons, currency
600 BC – 400 AD Roman	sugarcane, apples, asparagus, beets, oranges	using reaping machines, rotating legumes, using plows with wheels, trading extensively	food adulteration common	water mills, donkey mills, wooden cooperage

Table 2. Some components isolated from foods prior to 1800

Component	CAS registry number	Food	Investigator
Lactose	[*63-42-3*]	milk	F. Bartoletti (1586 – 1630)
Fructose	[*57-48-7*]	honey, raisins	J. R. Glauber (1604 – 1668)
Gluten	[*8002-80-0*]	flour	F. M. Grimaldi (1618 – 1663)
Tartar, ethanol	[*868-14-4*], [*64-17-5*]	wine	J. D. Portius (1636 – 1703)
Sucrose	[*57-50-1*]	beets	A. S. Marggraf (1709 – 1792)
Citric acid	[*77-92-9*]	lemon juice, gooseberries	C. W. Scheele (1742 – 1786)
Malic acid	[*6915-15-7*]	apples	
Glycerol	[*56-81-5*]	olive oil	

Table 3. Chronology of some developments in food technology

Date	Inventor	Process or product	Purpose
1679	Papin	digester	cooking in iron pot with clamped on lid
1809	N. Appert	appertization	cooking in sealed glass jars
1810	N. Appert	autoclave	cooking canned food under pressure
1830	A. Coffey	Coffey still	continuous distillation
1830	N. Rellieux	triple effect evaporator	energy efficient evaporation
1835	Sulzberger	roller milling	roller process for flour milling
1855	J. A. Just	roller dryer	energy-efficient drying
1856	G. Borden	evaporated milk	milk evaporated under vacuum and packed in hermetically sealed cans
1860	F. Carre	mechanical refrigerator	water–ammonia system for providing refrigeration
1865	L. Pasteur	heat treatment	initially to preserve wine, then used to destroy pathogens in grape juice and milk
1874	A. K. Shriver	retort	closed kettle using super-heated or live steam for canning
1877	G. de Laval	cream separator	rapid centrifugal separation of cream
1911	A. J. A. Ottesen	quick freezing process	minimizing adverse effects on food during freezing
1920	Kidd & West	controlled atmospheric storage	minimizing quality changes during storage of fruits
1931	C. Birdseye	multiple freezer	energy-efficient freezing
1945	E. W. Flosdorf	freeze-dryer	drying of frozen food under vacuum

the cost of the fertilizers in terms of dollars and expenditure of fossil fuel. In some cases the runoff of chemical fertilizers has polluted drinking water and increased the growth of algae in the streams, which has killed fish.

5) The use of pesticides has also contributed to the world food supply by eliminating pests that reduce yields or pests that consume harvested foods (Crop Protection). Because pesticides contribute to pollution and are hazardous to some species of animal life, alternative pest control methods have been and are being sought. Some success has been achieved in breeding disease-resistant varieties of plants and in developing biological methods, e.g., microbial parasites, that destroy the pests (Biological Control).
6) Reduction in postharvest losses has been achieved by using storage facilities and packages that protect the food from biological, chemical, and physical damage (Foods, 4. Packaging). In developing countries, postharvest losses can be as high as 50 % for some commodities [13].
7) Recovery of foods and ingredients from underutilized products. For example, recovery of proteins from whey and blood [14–17], underutilized meats from chicken [18], [19] and fish [20], [21].

Food Quality. Much of the progress in solving food quality problems can be attributed to basic scientific studies that have identified the causes for changes in quality, e.g., LOUIS PASTEUR's studies, which led to pasteurization (→ Foods, 2. Food Technology, Chap. 4.1.). In the process of studying the mechanism of thermal destruction of microbes, researchers noted that in some cases thermal processing also affected flavor, color, and texture [6], [7], [22–24]. Heating fluid dairy products produced cooked (custardlike) flavor, which was found to be due to the release of sulfhydryl groups from proteins [25]. More drastic heating of milk or other foods containing carbonyl and amino groups caused both flavor and color changes. In 1908, A. R. LING described this reaction, and in 1912 L. C. MAILLARD investigated the mechanism [26]. Direct heating of products with high sugar content produced both caramel color and flavor. Heating affected the texture of meat and eggs by altering the proteins. It was found that the activity of many endogenous food enzymes could be controlled by thermal processing [27], [28]. The importance and nature of enzymatic changes are discussed in Section 3.7 (→ Enzymes).

Scientific studies have shown that many changes in food quality are due to either oxidative or hydrolytic reactions, and therefore, controlling them is important [6], [7], [29–31]. *Oxidative changes* have been minimized by (1) storing food in a controlled atmosphere, (2) using packages that are impervious to oxygen, (3) inactivating oxidative enzymes, (4) avoiding contamination with catalytic metals, or (5) addition of antioxidants (→ Antioxidants; → Foods, 3. Food Additives, Chap. 3.2.). *Hydrolytic changes* have been inhibited by (1) inactivating hydrolytic enzymes, (2) maintaining the integrity of cell walls or membranes, which protect the substrate, or (3) adding inhibitors [32].

3. Components and Their Reactions

Foods range in complexity from a single compound (e.g., sugar) to multicomponent mixtures (e.g., meat). The properties of foods reflect the properties of the individual components and their interactions. *Food interactions* include all types of noncovalent (van der Waals, hydrogen, ionic, and hydrophobic) and covalent bonds. These interactions contribute to a variety of structural features such as cell walls, micelles, fat globules, fat globule membranes, three-dimensional gel structures, muscle fibers, and curds.

The *reactions of components* in foods are affected by pH, temperature, concentration of reactants, available water, catalysts, activators, and inhibitors. They may be catalyzed by light, enzymes, and nonenzymatic materials such as metals. Inhibitors can be substances that chelate metal catalysts or that interact with one of the reactants. Chelating agents can react with calcium or iron and thus reduce their nutritional availability. The nutritional quality of foods can also be reduced by oxidation of essential fatty acids or other nutrients, or by the interaction of proteins with tannins. The optimal sensory and nutritional properties can be obtained by controlling the reactions of the components.

3.1. Water

Occurrence. Water [*7732-18-5*] is one of the basic and most ubiquitous constituents of living organisms, [6], [7], [33–36]. The water content of raw fruits, vegetables, and meats exceeds 50 wt %, but the concentration in cereals, nuts and many processed foods is lower (→ Foods, 2. Food Technology, Chap. 2.).

Water exists naturally as an intracellular or extracellular component in plants and animal products. It is used as a solvent for sugars, salts, and acids in some foods, as a dispersing medium for hydrophilic macromolecular carbohydrates or proteins, and as a dispersed phase in emulsified products such as butter and margarine.

Properties. Compared to other substances with similar molecular masses, water is unique (→ Water). Its ability to engage in three-dimensional hydrogen bonding enables it to bond with organic molecules that contain nitrogen, oxygen, fluorine, or chlorine. The dissolution of sugar is because of bonding between water and the polar groups of the sugar. Dissolution of salts involves electrostatic forces between water and the positive ions which are greater than the attraction between ions.

Effect on Food. Water acts as a dispersing medium for amphipathic molecules, such as polar lipids, which have both hydrophilic and hydrophobic groups. The water associates with the hydrophilic moieties, such as phosphate groups, to "solubilize" the molecules. The amphipathic molecules form macromolecular aggregates, which are called micelles. Introducing hydrophobic substances, such as apolar groups of fatty acids, amino acids, and proteins, into water leads to an association of the apolar groups, i.e., hydrophobic interactions.

Water may be present in food as "free" water or "bound" water. *Free water* behaves like pure water, whereas the behavior of bound water is limited. *Bound water* has been defined as water that is not available as a solvent and does not freeze at − 40 °C, although some scientists have suggested other definitions. The term "water binding" has been defined as the tendency of water to associate, with various degrees of tenacity, to hydrophilic substances. "Water-holding capacity" has been used to describe the ability of a matrix of molecules to entrap a large amount of water in such a manner that exudation is prevented [7]. Although the entrapped water does not flow freely from the product, it behaves very much like pure water during processing. The water-holding capacity of meats and gels have a profound effect on rheological properties. The water-holding capacity of meats is affected by pH and salts, especially phosphates. A classification of water–protein thermodynamic associations is as follows [34]:

Structural water is water that is hydrogen-bonded to specific groups; it participates in stabilization of structure and is unavailable for chemical reaction.

Hydrophobic hydration water is structured cagelike water surrounding apolar residues; like structural water, it is very much involved in stabilizing protein structure.

Monolayer water is the first adsorbed water monolayer; it is hydrogen-bonded and is unavailable as solvent, but may be available for chemical reactions; ranges from 4 to 9 g/100 g protein.

Unfreezable water includes roughly all water (structural monolayer, and perhaps some adsorbed multilayer water) that does not freeze at normal temperature; amounts to 0.3 – 0.5 g/g of protein and corresponds to water up to $a_w = 0.9$; amount varies with polar amino acid content and includes some water available for chemical reactions.

Capillary water is water that is held physically in clefts or by surface forces in the protein molecule (e.g., water entrapped in gels or cheese curd); its physical properties are similar to those of bulk water.

Hydrodynamic hydration water "loosely" surrounds the protein and is transported with the protein during diffusion (centrifugation); it has properties typical of normal water.

Water Activity [6]. A complete discussion of water activity can be found under → Foods, 2. Food Technology, Chap. 1.4.

Perishability. A general correlation exists between the perishability of food and water content, but a better correlation exists between perishability and water activity (→ Foods, 2. Food Technology, Chap. 1.). Both enzymatic and

nonenzymatic reactions tend to increase as water activity increases. However, enzymatic reactions are less sensitive to water activity changes. Nonenzymatic browning reactions reach a maximum rate at $a_w \approx 0.80$. Lipid oxidation is inhibited by increasing water content. Under some conditions water may form a protective film around the fat. The destruction of vitamin C and chlorophyll increases rapidly as the water activity increases. The rate of vitamin B_1 destruction reaches a maximum at $a_w \approx 0.5$ [7].

Food Quality. A complete discussion of the effect of freezing the water in food has on food quality can be found under → Foods, 2. Food Technology, Chap. 3.2.

3.2. Proteins (→ Amino Acids; → Proteins);[6], [7], [29], [37–44]

Structure. Proteins are polymers composed of amino acids linked together by peptide bonds. Proteins constitute ca. 50 % of the dry matter of living cells and perform a vital role in the function and structure of the cells. Proteins containing only amino acids are called homoproteins, whereas those containing additional groups (i.e., prosthetic groups) are called heteroproteins. The heteroproteins include lipoproteins, glycoproteins, phosphoproteins, hemoproteins, metalloproteins, and nucleoproteins.

The complexity of proteins depends on the number and types of amino acids and other groups. The primary, secondary, and tertiary structures of proteins are discussed under → Proteins. The structural properties of some major food proteins are shown in Table 4.

Denaturation. The secondary and tertiary structures of proteins are fragile and, therefore, can be changed by a variety of treatments that do not cleave covalent bonds (with the exception of disulfide bonds). The term denaturation is applied to such processes. *Heating* is the most commonly used physical method for denaturing proteins. Most proteins are denatured or coagulated between 55 and 75 °C. The susceptibility of proteins to denaturation depends on the nature of the protein and environmental factors such as pH, water activity, ionic strength, and the types of ions present. The proteins of egg white are readily denatured by heat, whereas casein and gelatin are relatively resistant. This resistance is attributed to the presence of high amounts of proline and hydroxyproline and low levels of sulfur-containing amino acids. Thermal denaturation of proteins can result in the following:

1) inactivation of enzymes
2) destruction of microbes
3) detoxification of toxic proteins
4) increased protein digestibility
5) improved texture of protein-rich foods
6) increased exposure of hydrophilic groups

The exposure of hydrophobic groups can cause changes in water and flavor-binding capacity and in solubility.

Physical treatments other than heating can also cause protein denaturation or other protein alterations. *Freezing and low-temperature storage* can cause aggregation and precipitation of some proteins (e.g., gliadin, 11 S soy proteins, and some egg and milk proteins). Aggregation refers to clustering of protein molecules or particles. In food systems, aggregation usually involves hydrophobic forces, but other forces may be involved in some cases. Precipitation involves separation of the protein from the rest of the solution or suspension. In some cases aggregation leads to precipitation. After freezing, fish may become tough and rubbery and lose moisture as a result of protein destabilization. Some *mechanical treatments,* such as the rolling and kneading of bread, may denature proteins. *High hydrostatic pressure* may also have a denaturing effect. Ovalbumin and trypsin have been reported to be denatured at pressures of 50 and 60 kPa, respectively [7]. *Electromagnetic irradiation* can cause changes in the conformation of proteins. The extent of such changes depends on the wavelength and energy involved.

Proteins can also be denatured by *chemical reagents.* The structural stability of proteins is influenced by pH, with each protein having a characteristic stable range. Consequently, shifting the pH outside of this range with either acid or base denatures the protein. In some cases, native structure can be restored by readjusting the pH to within the stable range.

Alkaline-earth elements, such as calcium and magnesium, may react with proteins. Ions of copper, iron, mercury, and silver react even more

Table 4. Structural properties of some major food proteins [7]

Protein	CAS registry number	M_r	Type[a]	Secondary structure, %		Number of					Sequence[b]	Prosthetic group, wt %	Average hydro-phobi-city [x], kJ/mol
				α-Helix	β-Sheets	residues	S–S	SH	subunits	pI			
Myosin[c]	[*9066-74-4*]	475 000	F	high		4500	0	40	6	4 – 5	K	phosphorus	4.25[g]
Actin[c]	[*9064-37-3*]	42 000	G → F				0	5 – 6	1 – 300	4 – 5			4.4[g]
Collagen (tropocollagen)[c]	[*9064-68-0*]	300 000	F	collagen helix					3	~ 9	K		4.5[h]
αS_1-Casein B[c]	[*9000-71-9*]	23 500	R C			199	0	0	1	5.1	K	phosphorus (1.1)	5.0
β-Casein A[c]	[*9064-59-9*]	24 000	R C			209	0	0	1	5.3	K	phosphorus (0.56)	
K-Casein B[c]	[*9049-48-3*]	19 000	R C			169	0	2	1	4.1 – 4.5	K	carbohydrate (5) phosphorus (0.22)	
β-Lactoglobulin A[c]	[*9045-23-2*]	18 400	G	10	30	162	2	1		5.2	K		5.15
α-Lactalbumin B[c]	[*12585-12-5*]	14 200	G	26	14	123	4	0		5.1	K		4.8
Ovalbumin[d]	[*9006-59-1*]	45 000	G				1 or 2	4		4.6		carbohydrate (3.3) phosphorus	4.65
Serum albumin[c]	[*9048-46-8*]	69 000	G				17	1	1	4.8	K		4.7
Gliadins[e] (α,β,γ)	[*9007-90-3*]	30 000 – 45 000	G → F	30			2 – 4		1		K, P		4.5
Glutenins[e]	[*9061-41-0*]	1 000 000	F	15			50		15				
Glycinin[f]	[*9007-93-6*]	350 000	G	5	35		23	2	12	4.6	U		
Conglycinin[f]	[*104042-07-1*]	200 000	G	5	35		2		9	4.6	U	carbohydrate (4)	

[a] G = globular, F = fibrous, R C = random coil.
[b] K = known, U = unknown, P = partially known [44].
[c] Bovine.
[d] Hen's egg.
[e] Wheat.
[f] Soybean.
[g] Rabbit.
[h] Chicken.

readily, often forming stable complexes with thiol groups. Some of these ions become an integral part of protein molecules, micelles, or other molecular associations. Removal of ions by dialysis or chelating agents can reduce protein stability.

Organic solvents may denature proteins by altering the dielectric constant of the medium and, therefore, the electrostatic forces that contribute to protein stability. Urea and quinidine salts promote denaturation by disrupting hydrogen bonds. Nonpolar solvents can penetrate into hydrophobic regions, disrupting hydrophobic interactions and enhancing protein denaturation. Surface-active agents, such as sodium dodecylsulfate, also disrupt hydrophobic regions and promote the unfolding of native proteins. Unfolding can also be enhanced by anionic detergents. The conformation of proteins can be modified by reducing disulfide cross-links with reducing agents such as cysteine and ascorbic acid.

Maillard Reaction. The nonenzymic browning or Maillard reaction is one of the most important reactions in food. It involves reaction between the amino groups of proteins, peptides, or amino acids with reducing sugars and can produce either desirable or undesirable effects. The amino groups react with a glycosidic hydroxyl group to form an *N*-substituted glycosylamine. The glycosylamine undergoes an Amadori rearrangement to form an aldoseamine or ketoseamine. The aldoseamine could react with a second amino group to yield a diamino sugar, or the ketoseamine could react with a second aldose to form a diketoseamine. Both of these paths produce melanoidin pigments, which have pyrazine and imidazole rings in addition to hydroxymethyl furfural. The types of end products depend on the nature and quantity of the sugar and the molecules containing the amino groups. The browning reaction is affected by temperature, pH, moisture level, and the presence of oxygen, metals, phosphates, and inhibitors such as sulfur dioxide.

The impact of the browning reaction on flavor, color, and nutritive value depends on the type of reactants and end products. Desirable colors and flavors are produced during the broiling of meat or fish, baking of bread, or roasting of coffee. Frying potato chips containing high glucose levels can produce brown spots and some undesirable flavors. Undesirable flavors and colors develop during storage of dairy products that have been subjected to high heat, e.g., evaporated milk and ultrahigh-heat-treated milk. The color and flavor changes are usually accompanied by losses in nutritive values. Sugars destroy the nutritive value of lysine because they react readily with its ε-amino group. This loss can be significant because lysine is the limiting amino acid in many proteins (cf. essential amino acids). Other amino acids, especially the other basic amino acids, L-arginine and L-histidine, are also susceptible to degradation.

Degradation of amino acids via the Maillard reaction may occur before any visual browning. Furthermore, amino acid losses may result from the *Strecker degradation,* which does not produce browning. The Strecker degradation, which involves interaction of dicarbonyl compounds and amino acids, produces a variety of flavorful aldehydes and nitrogen-containing compounds such as pyrazines. These compounds contribute to the flavor of chocolate, honey, maple syrup, and bread [7].

Strong alkali treatment of proteins can induce changes that may result in reduction of their nutritive value [49]. This reduction may be due to the loss of limiting amino acids. The availability of some amino acids may be reduced by alkali-induced amino acid isomerization. Cystine can be converted to aminoacrylic acid (dehydroalanine), which can react with the amino group of lysine to form lysinoalanine. Under stronger alkaline conditions, arginine may be converted to ornithine and urea. Ornithine can then combine with dehydroalanine to yield ornithinoalanine. It has been postulated that lysinoalanine and ornithinoalanine may contribute to inter- or intramolecular cross-links, which reduce the accessibility of proteins to proteolytic enzymes. Although mild alkali treatments of proteins probably do not have a significant nutritional effect, processors should carefully control these treatments.

Lysinoalanine can produce toxic effects in rats, but these effects have not been observed in other experimental animals [7], [49].

Oxidation contributes to the loss of some amino acids. Methionine can react with lipid peroxides to yield methionine sulfoxide. Cysteine can be oxidized to cystine and then to alanine by

lipid free radicals. Light-induced oxidation of methionine produces methional and a mixture of sulfides, which imparts the so-called light-induced flavor to milk.

The physicochemical properties of proteins that affect their behavior in food systems during preparation, processing, storage, and handling are called *functional properties.* These properties and their role in food systems are listed in Table 5. Research on functional properties has been greatly stimulated by the use of proteins from plant sources such as soybeans. Modification of proteins by enzymatic, chemical [43], or texturization procedures [37], [42] are being used to improve protein functionality.

Table 5. Typical functional properties performed by functional proteins in food systems [40]

Functional property	Mode of action	Food system
Solubility	protein solvation	beverages
Water absorption and binding	hydrogen bonding of water; entrapment of water (no drip)	meat, sausages, breads, cakes
Viscosity	thickening; water binding	soups, gravies
Gelation	protein matrix formation and setting	meats, curds, cheese
Cohesion–adhesion	protein acts as adhesive material	meat, sausages, baked goods, pasta products
Elasticity	hydrophobic bonding in gluten; disulfide links in gels	meats, bakery
Emulsification	formation and stabilization of fat emulsions	sausages, bologna, soup, cakes
Fat absorption	binding of free fat	meats, sausages, doughnuts
Flavor binding	adsorption, entrapment, release	simulated meats, bakery, etc.
Foaming	form stable films to entrap gas	whipped toppings, chiffon desserts, angel cakes

Nutritive Value. The nutritive value of a protein depends on its amino acid content and its digestibility. The *essential amino acids* (i.e., those that must be supplied by the diet) are isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine for adults plus histidine for infants. A perfect protein must contain the proper concentrations of the essential amino acids, and they must be available (i.e., digestible). In general, animal proteins more nearly meet these criteria than plant proteins. Cereal proteins are low in lysine and some also lack tryptophan. Proteins from legumes often lack methionine, and oil seeds and nuts frequently lack both methionine and lysine. The amino acid deficiencies of individual proteins can be wholly or partially overcome by inclusion of another protein that has a complementary amino acid pattern. Cereal and legume proteins complement each other. Diets in which 30 – 40 % of the proteins come from animal sources usually contain a proper balance of amino acids.

The *availability of amino acids* depends on the ability of the digestive tract's proteases to digest the proteins. Poor digestibility may be due to several factors: (1) protein conformation; e.g., soluble globulins are more digestible than insoluble fibrous proteins; (2) presence of enzyme inhibitors, such as the trypsin inhibitors, in soybeans; (3) presence of substances that form indigestible complexes with proteins such as metals, nucleic acids, and tannins.

3.3. Lipids (→ Fats and Fatty Oils ; → Fatty Acids; → Fatty Alcohols), [6], [7], [29–31], [35]

Composition. Lipids include a variety of compounds that have been classified as follows [29, p. 140]:

1) *Simple lipids* (neutral lipids): esters of fatty acids with alcohols, e.g., fats and waxes
2) *Compound lipids:* compounds containing fatty acid esters plus other groups, e.g., phospholipids (phosphatides), cerebrosides (glycolipids), sphingolipids, and sulfolipids
3) *Derived lipids and lipid-soluble substances,* e.g., fatty acids, long-chain alcohols, sterols, hydrocarbons, carotenoids, and fat-soluble vitamins

Commercial fats and oils are mixtures of the three classes of lipids, with simple lipids (principally triglycerides) making up the major portion. Mixtures that appear solid are often called fats, whereas liquid mixtures are called oils. In this article, the term fat will be used to include both types.

Properties of the natural fats depend mainly on the fatty acid composition and the distribution of fatty acids in the glycerides. The fatty acid composition of a variety of fats and oils can

be found under → Fats and Fatty Oils. The differences in fatty acid composition partially explains the differences in melting points and melting point ranges. Butterfat, which has a large variety of fatty acids including several fatty acids with branched chains and odd-numbered carbon atoms, has a wide melting point range. Cocoa butter has a limited number of fatty acids and a narrow melting point range. Tallow with a relatively high concentration of long-chained saturated fatty acids has a high melting point, whereas the highly unsaturated safflower oil has a low melting point.

The distribution of fatty acids in the triglycerides of natural fats is shown in Table 6. Cocoa butter has very little trisaturated (GS_3) or triunsaturated (GU_3) types of triglycerides and is hard and brittle below the melting point, which makes it useful for chocolate and confectionery manufacture. The effects of the types of fatty acids and triglycerides on melting points are shown in Table 7.

Table 6. Triglyceride types and isomeric forms of natural fats [29]

	Types ∗, wt %				Isomers∗, wt %			
	GS_3	GS_2U	GSU_2	GU_3	SUS	SSU	USU	UUS
Pig fat	2.5	22.4	55.7	19.4	1.0	21.4	46.9	8.8
Peanut oil	0.1	9.9	42.5	47.5	9.3	0.6	0.7	41.8
Beef fat	12.6	43.7	35.3	8.4	30.6	13.1	3.4	31.9
Cocoa butter	7.1	67.5	23.3	2.1	65.0	2.5	0.2	23.1
Soybean oil	0	3.7	31.0	65.3	3.7	0	0	31.0

∗ G = Glyceride, S = saturated fatty acid, U = unsaturated fatty acid.

Table 7. Melting points of some triglycerides [6]

Triglyceride saturation	Fatty acids∗	*mp*, °C
Trisaturated	S–S–S	72
	P–P–P	65
	S–P–P	62
Disaturated	P–O–P	37
	O–P–P	34
	S–O–P	35
	S–P–O	39
	P–S–O	36
Diunsaturated	O–O–P	19
	O–O–S	23
Triunsaturated	O–O–O	5
	E–E–E	42
	L–L–L	− 10

∗ S = stearic; P = palmitic; O = oleic; E = elaidic; L = linoleic.

Solidified fats exhibit polymorphism; i.e., they can exist in different crystalline forms, depending on the modes of molecular packing in the crystal. Triglyceride may exist in three forms: α, β', and β, in order of increasing stability. The crystal form has a marked effect on the melting point. For example, the melting point of α, β', and β forms of tristearin are 54.7, 64.0, and 73.3 °C, respectively [6].

Chemical reactions of fats can have a significant impact on food quality. Hydrolysis of triglycerides, which releases glycerol and free fatty acids, can alter both flavor and texture. *Hydrolysis* of fats, such as coconut, palm kernel, and butterfat, which contain volatile, fatty acids (i.e., C_4–C_{10}), develops hydrolytic rancidity or lipolyzed flavor. This flavor is objectionable in fluid milk, but a low concentration is desirable in some cheeses. Extensive hydrolysis of any fat alters the texture because the melting point is changed. Hydrolysis rarely occurs in the absence of lipolytic enzymes. The rate of lipolysis depends on lipase activity and the availability of the substrate. In raw fluid dairy products, lipolysis can be a serious problem because these foods usually contain active lipase, and the substrate is available. Treatments that disrupt the fat globules increase substrate availability. Many psychotropic bacteria produce lipase, which may cause trouble during storage of foods.

Oxidation of fats contributes to development of undesirable flavor in foods. Fats with a high concentration of polyunsaturated fatty acids (Table 7) are very susceptible to oxidation, which leads to the production of a mixture of flavorful carbonyl compounds, mainly six- to ten-carbon aldehydes and ketones. Lipid oxidation is catalyzed by metals (e.g., copper and iron), ultraviolet light, and lipoxygenases, which occur in some plant sources. Metal-catalyzed oxidation can be inhibited by metal chelating agents such as citric acid and ethylenediaminetetraacetic acid (EDTA). The most commonly occurring natural antioxidants are the tocopherols, which inhibit the oxidative free radical chain mechanism (→ Vitamins). Carotenoids and a number of natural and synthetic phenolic compounds, e.g., butylated hydroxyanisole (BHA) and butylated hydroxytoluene (BHT), inhibit oxidation (→ Antioxidants).

Hydrogenation of the unsaturated fatty acid components of fats can be used (1) to convert liq-

uid oils to margarine or other semisolid products and (2) to decrease the susceptibility to oxidation. The extent and selectivity of hydrogenation can be controlled by the choice of reaction conditions (e.g., temperature, pressure and catalyst). A simplistic illustration of the possible reactions during hydrogenation of linolenate is as follows:

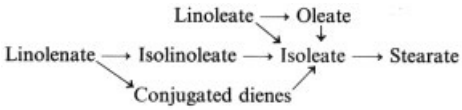

Normally fats are only partially hydrogenated because complete saturation would yield products that are too hard and brittle.

Polymerization of unsaturated fatty components may occur during commercial frying operations. Appreciable polymerization causes an increase in the viscosity and foaming of frying fats. Research with experimental animals indicates that polymers are poorly absorbed from the intestinal tract and some have been reported to be toxic [49].

Phospholipids or phosphatides are present in all fats and fat-containing foods; however, these compounds are removed during the refining of fats. Phospholipids consist of polyhydric alcohols (usually glycerol) combined with fatty acids, phosphoric acid, and a nitrogen-containing group. Phospholipids are surface active because they contain a lipophilic and hydrophilic portion. Consequently, they are emulsifying agents and are the natural emulsifiers in fluid dairy products. The fatty acid groups are usually more unsaturated than those found in triglycerides. The phospholipids in milk are particularly rich in polyunsaturated fatty acids and are on the surface of the fat globules. Consequently, they are more susceptible to oxidation than the unsaturated triglycerides.

Dietary Aspects. Fats play a significant dietary role by providing calories and essential fatty acids, by serving as vitamin carriers, and by increasing the palatability of food. However, excessive consumption can adversely affect health. High intake of cholesterol and saturated fatty acids may increase the risk of cardiovascular diseases in some susceptible individuals; in many others, no relationship has been found between diet and blood cholesterol. By contrast, increasing the relative amount of polyunsaturated fatty acids can be desirable. The changes can be achieved by eating less animal, coconut, or palm fats and more polyunsaturated vegetable or marine oils. However, some of this latter group, such as older varieties of rapeseed and herring oils may contain a high concentration of erucic acids [6]. Feeding rats diets with high erucic acid content caused an infiltration of fat into heart tissues. The oil from seeds of some rape varieties are low in erucic acid and processing treatments are being developed to remove it from marine oils.

3.4. Carbohydrates (→ Carbohydrates), [6], [23], [24], [35]

Composition. Carbohydrates are compounds containing carbon, hydrogen, and oxygen and may be classified as follows:

1) Monosaccharides: pentoses and hexoses
2) Oligosaccharides: di-, tri-, and tetrasaccharides
3) Polysaccharides (glycans) (→ Starch and Other Polysaccharides):
   Homoglycans: pentosans, hexosans (cellulose, dextrin, glycogen, and starch)
   Heteroglycans: gums, pectins, hemicellulose

The main carbohydrates in animal organisms are glucose and glycogen and in milk primarily the disaccharide lactose. Plants contain a wide range of all three classes of saccharides, with cellulose being the main structural carbohydrate and starch the main storage carbohydrate.

Properties. Monosaccharides are polyhydroxy compounds with either an aldehyde or ketone group and are the building blocks for oligo- and polysaccharides. The mono- and disaccharides (i.e., the sugars) are primarily responsible for the sweetness of foods. Sugars differ considerably in sweetness, with fructose and sucrose being the sweetest mono- and disaccharides, respectively [46]. The saccharides also vary with respect to solubility. Lactose (i.e., milk sugar) is much less soluble than sucrose and crystallizes during storage of products such as sweetened condensed milk or ice cream. However, the rate of crystallization and size of crystals can

be controlled by proper manufacturing and handling practices. During storage of products such as peas and sweet corn, the sugar is converted enzymatically to starch, which decreases sweetness and increases firmness. Storing potatoes at low temperature favors conversion of starch to sugar. Potatoes that are fried after low-temperature storage develop more brown color, resulting from the Maillard browning reaction between the sugars and the potato proteins (see Section 3.2).

Reactions. When sugars are heated in the absence of water or in concentrated solution, a complex series of reactions occurs that leads to *caramel formation.* Caramel pigment formation can be considered a nonenzymatic browning process in the absence of nitrogen. This process involves anomeric shift, ring size alteration, cleavage of some glycosidic linkages and creation of others, condensation, isomerization, dehydration, fragmentation, and browning [23], [46]. During the course of these reactions, a variety of compounds are produced, including 1,2-anhydro-α-D-glucose; 1,6-anhydro-β-D-glucose; 1,3′; 2,2′-dianhydro-β-D-glucopyranosyl-β-D-fructofuranose; caramelan ($C_{24}H_{36}O_{18}$), a bitter tasting pigment; caramelen ($C_{35}H_{50}O_{25}$); and volatile compounds such as diacetyl (2,3-butanedione), acetic acid, and formic acid. A typical caramel flavor has been attributed to 4-hydroxy-2,3,5-hexanetrione and 4-hydroxy-2,5-dimethyl-3(2*H*)-furanone [47].

The trisaccharide raffinose and the tetrasaccharide stachyose are found in legumes. These oligosaccharides are not readily absorbed, but they undergo *fermentation* in the large intestines [46]. This leads to gas production and flatulence, which tends to limit the consumption of legumes.

Starch. Starch, a polymer of D-glucose, is a major source of energy in the human diet (→ Starch and Other Polysaccharides). Starch can be classified into the following three types:

1) amylose: essentially a linear polymer of α-(1 → 4)-linked pyranose rings
2) amylopectin: mostly α-(1 → 4) linkages and 4 – 5 % α-(1 → 6) linkages to yield a branched structure
3) modified starch: a processed form of corn, wheat, or tapioca starch

Amylose represents ca. 17 – 30 % of the total starch in the commonly occurring natural starches. However, in some varieties of pea and corn starch, the amylose content may be as high as 75 %. By contrast, waxy corn starch contains only amylopectin. Some properties of various corn starches are shown in Table 8.

Starch occurs in plants in the form of discrete granules that have species-specific shapes and sizes. For example, potato starch granules are egg-shaped and large (15 – 100 µm), whereas rice starch granules are round and small (3 – 9 µm). Granules are insoluble in water and, when heated, they start to swell at the so-called *gelatinization temperature,* i.e., the temperature at which some of the intermolecular hydrogen bonds are broken. Generally starches with large granules have lower gelatinization temperatures than those with small granules. This temperature is also affected by pH, heating rate, and the presence of salts and sugars. Swelling is inhibited by C-16 or C-18 fatty acids or by monoglycerides that form inclusion complexes with helical amylose. Heating aqueous starch suspensions above the gelatinization temperature causes them to become viscous and translucent. When a cooked starch is cooled gradually, the linear portion may form a precipitate of spherical crystals. This phenomenon is called retrogradation and depends on the size, shape, and concentration of the starch molecules and on the temperature. Freezing greatly accelerates retrogradation [6, p. 157].

Starch can be modified by *chemical treatments* to enhance specific properties (some examples are shown in Table 8). Starch can be hydrolyzed to give a variety of products, depending on the composition of the raw starch and the method of hydrolysis, i.e., acid, enzymatic, or combined acid–enzymatic treatment. The extent of hydrolysis is often expressed in terms of its dextrose equivalent (DE) (→ Enzymes, Chap. 4.2.5.; → Glucose and Glucose-Containing Syrups). Low DE hydrolyzates are used to increase viscosity and to serve as flavor carriers; high DE hydrolyzates impart sweetness. The sweetness can be increased by treatment with D-glucose isomerase, which converts D-glucose to D-fructose.

Table 8. Properties of various corn starches [24]

Type	Amylose: amylopectin	Gelatinization temperature, °C	Distinguishing properties
Normal	1 : 3	62 – 72	poor freeze– thaw stability
Waxy	0 : 1	63 – 72	minimal retrogradation
High amylose	3 : 2 – 4 : 1	66 – 92	granules less birefringent than normal starches
Acid-modified	variable	69 – 79	decreased hot-paste viscosity compared with unmodified
Hydroxyethylated	variable	58 – 68 (DS 0.04)	increased paste clarity, reduced retrogradation
Phosphate, monoesters	variable	56 – 66	reduced gelatinization temperature, reduced retrogradation
Cross-linked	variable	higher than unmodified depends on degree of cross-linking	reduced peak viscosity, increased paste stability
Acetylated	variable	55 – 65	good paste clarity and stability

Table 9. Selected properties of some polysaccharide gums [24]

Name	Principal monosaccharide constituents	Source	Distinguishing properties
Guar	D-mannose, D-galactose	*Cyamopsis tetragonolobus*	high solution viscosity at low concentration
Locust bean gum	D-mannose, D-galactose	*Ceratonia siliqua*	synergistic with carrageenan
Gum arabic	D-galactose, D-glucuronic acid	*Acacia*	high water solubility
Gum tragacanth	D-galacturonic acid, and D-galactose, L-fucose, D-xylose, L-arabinose	*Astragalus*	stable over wide pH range
Agar	D-galactose, 3,6-anhydro-L-galactose	Red seaweed	forms extremely strong gels
Carrageenan	sulfated D-galactose, sulfated 3,6-anhydro-D-galactose	Irish moss (*Chondrus crispus*)	forms chemically set gels with K^+
Alginate	D-mannuronic acid, L-guluronic acid	Brown seaweed (*Macrocystis pyrifera*)	forms chemically set gels with Ca^{2+}
Dextran	D-glucose	*Leuconostoc mesenteroides*	crystallization inhibitor in candies and frozen desserts
Xanthan	D-glucose, D-mannose, D-glucuronic acid	*Xanthomonas campestris*	dispersions are highly pseudoplastic

Polysaccharides. Pectic substances and gums are two groups of plant polysaccharides that can be used to modify food texture (→ Polysaccharides). Pectic substances are polymers of $1 \rightarrow 4$-linked galacturonic acids, which are found in the middle lamella of plant cells. Part of the carboxyl groups of polygalacturonic acids are esterified, and part neutralized by one or more bases. Pectins with at least 50 % of the carboxyl groups methylated form gels in the presence of sugars. Pectins with a low methoxyl content can form gels, in the absence of sugar, because of cross-linking between carboxyl groups and divalent cations, especially calcium. The composition and properties of a number of gums are shown in Table 9.

Cellulose. Cellulose, the most abundant natural carbohydrate, is a high molecular mass polymer of anhydro-β-D-glucopyranose units linked by $(1 \rightarrow 4)$ glycosidic bonds (→ Cellulose). Because the human digestive tract lacks enzymes that can hydrolyze cellulose, it is indigestible, although it does provide fiber. Cellulose reacts with ethers to form products called methylcellulose, hydroxypropylcellulose (HPC), carboxymethyl cellulose (CMC), and microcrystalline cellulose (CM). These modified celluloses are used as stabilizers in a variety of products (→ Cellulose Ethers).

Hemicelluloses and *pentosans* are noncellulosic, nonstarchy complex polysaccharides that occur in plant tissues along with cellulose. Hemicelluloses are insoluble in water, whereas pentosans are soluble in water. These polysaccharides usually contain two to four different sugars. They frequently exist in single chains of β-$(1 \rightarrow 4)$-linked D-xylopyranose units and may have L-arabinofuranose units attached at C-3 or possibly at C-2. Because these compounds are not digestible, they contribute to the dietary fiber content of food. The hemicelluloses may improve the water-binding capacity of flour, the mixing quality of bread dough, and the loaf volume of bread.

Dietary Aspects. Carbohydrates provide an efficient and inexpensive energy supply. Many United States nutritionists recommend that United States consumers increase their carbohydrate consumption so that it supplies at least 58 % of the total energy intake (carbohydrates already supply more than this amount of energy in the diets of consumers in countries with

lower meat consumption). Furthermore, it has been recommended that the dietary intake of refined and processed sugar be reduced so that the energy intake from this source does not exceed 10 %. Although this recommendation has not received unanimous support, it is based on evidence that simple sugars may contribute to dental caries and that the more complex carbohydrates have more potential advantages. Consumers are also recommended to include enough fiber in their diet to minimize constipation and diverticular diseases. In addition, fiber is believed to (1) reduce the risk of cardiovascular disease by lowering serum cholesterol and (2) reduce cancer of the colon and rectum by binding carcinogens and speeding their excretion.

3.5. Mineral Components

Occurrence. Mineral or inorganic components occur in foods as salts or in combination with organic compounds [6], [48], [49]. Such components are commonly divided into two groups: the major minerals and trace elements. The *major minerals,* which are essential nutrients, include potassium, sodium, calcium, magnesium chloride, sulfate, and phosphate. The *trace elements* have been divided into three groups:

1) essential nutritive minerals, which include fluorine, silicon, vanadium, chromium, manganese, iron, cobalt, nickel, copper, zinc, selenium, molybdenum, tin, and iodine
2) nonnutritive, nontoxic elements, including aluminum and boron
3) nonnutritive, toxic elements, including mercury, lead, arsenic, cadmium, and antimony

Some of the minerals exist in food as soluble salts, especially the monovalent species. These minerals contribute to the taste and the osmotic pressure of the aqueous phase. Some of the polyvalent ions are present as an equilibrium mixture of ionic, soluble nonionic, and colloidal species. Such equilibria exist in milk and meat. Many minerals are present as chelates, including the iron-containing hemoglobin and myoglobin, the magnesium-containing chlorophyll, the cobalt-containing vitamin B_{12}, and the copper-containing polyphenoloxidase.

Calcium, an important nutrient, forms complexes with a variety of compounds, including oxalates, phosphates, citrates, alginates, pectates, and caseinates. Oxalates bind the calcium so that it is unavailable as a nutrient. Calcium citrates and phosphates affect the colloidal stability of proteins in products such as evaporated milk. Calcium alginates and pectates are involved in producing stable viscous products. USDA reports indicate that many American diets contain less than the recommended amount of calcium. Calcium affects bone structure and therefore calcium deficiency could contribute to osteoporosis, a disorder in which bones become less dense and more fragile. This disorder is particularly prevalent in older women (postmenopause). Although several factors affect osteoporosis, many scientists believe that extra calcium may help prevent osteoporosis. This information has led to the fortification of a variety of foods with calcium. Some experts question the wisdom of such a massive fortification effort.

Reactions. Canned foods may react with tin, iron, or lead in the can and produce color changes resulting from reactions between the metals and polyphenols in the food. Calcium and magnesium may be absorbed by peas if they are cooked in hard water. When dry beans are soaked in water, some minerals may leach into the water.

3.6. Vitamins

Vitamins are organic compounds that are required in small amounts in the diet for growth and maintenance of life (→ Vitamins) [7], [35], [48–50]. Originally vitamins were all believed to be amines, and hence the name "vitamines" was chosen. When some of them were found not to be amines, the "e" was dropped from the word.

Vitamins have been divided into two groups, namely, fat-soluble vitamins (A, D, E, and K) and water-soluble vitamins (C and the B complex). *Fat-soluble vitamins* are stored in the liver and can be retained for long periods of time. This can lead to vitamin toxicity if the accumulated level gets too high. *Water-soluble vitamins* are only retained for short periods of time, and thus, the supply can be depleted rapidly.

Scientists worldwide have studied vitamin requirements and published dietary guidelines. Ta-

Table 10. Recommended daily dietary allowances [a] [51]

	Age	Weight		Height		Protein	Fat-soluble vitamins			Water-soluble vitamins							Minerals					
							Vit. A	Vit. D	Vit. E	Vit. C	Thiamin	Ribo-flavin	Niacin	Vit. B_6	Folacin [f]	Vit. B_{12}	Ca	P	Mg	Fe	Zn	I
	years	kg	lbs	cm	in	g	µg R.E. [b]	µg [c]	mg α T.E.	mg	mg	mg	mg N.E. [e]	mg	µg	µg	mg	mg	mg	mg	mg	µg
Infants	0.0–0.5	6	13	60	24	kg×2.2	420	10	3	35	0.3	0.4	6	0.3	30	0.5 [g]	360	240	50	10	3	40
	0.5–1.0	9	20	71	28	kg×2.0	400	10	4	35	0.5	0.6	8	0.6	45	1.5	540	360	70	15	5	50
Children	1–3	13	29	90	35	23	400	10	5	45	0.7	0.8	9	0.9	100	2.0	800	800	150	15	10	70
	4–6	20	44	112	44	30	500	10	6	45	0.9	1.0	11	1.3	200	2.5	800	800	200	10	10	90
	7–10	28	62	132	52	34	700	10	7	45	1.2	1.4	16	1.6	300	3.0	800	800	250	10	10	120
Males	11–14	45	99	157	62	45	1000	10	8	50	1.4	1.6	18	1.8	400	3.0	1200	1200	350	18	15	150
	15–18	66	145	176	69	56	1000	10	10	60	1.4	1.7	18	2.0	400	3.0	1200	1200	400	18	15	150
	19–22	70	154	177	70	56	1000	7.5	10	60	1.5	1.7	19	2.2	400	3.0	800	800	350	10	15	150
	23–50	70	154	178	70	56	1000	5	10	60	1.4	1.6	18	2.2	400	3.0	800	800	350	10	15	150
	51+	70	154	178	70	56	1000	5	10	60	1.2	1.4	16	2.2	400	3.0	800	800	350	10	15	150
Females	11–14	46	101	157	62	46	800	10	8	50	1.1	1.3	15	1.8	400	3.0	1200	1200	300	18	15	150
	15–18	55	120	163	64	46	800	10	8	60	1.1	1.3	14	2.0	400	3.0	1200	1200	300	18	15	150
	19–22	55	120	163	64	44	800	7.5	8	60	1.1	1.3	14	2.0	400	3.0	800	800	300	18	15	150
	23–50	55	120	163	64	44	800	5	8	60	1.0	1.2	13	2.0	400	3.0	800	800	300	18	15	150
	51+	55	120	163	64	44	800	5	8	60	1.0	1.2	13	2.0	400	3.0	800	800	300	10	15	150
Pregnant						+30	+200	+5	+2	+20	+0.4	+0.3	+2	+0.6	+400	+1.0	+400	+400	+150	[h]	+5	+25
Lactating						+20	+400	+5	+3	+40	+0.5	+0.5	+5	+0.5	+100	+1.0	+400	+400	+150	[h]	+10	+50

[a] Revised 1980. The allowances are designed for the maintenance of good nutrition of practically all healthy people in the United States. They are intended to provide for individual variations among most normal persons under usual environmental stresses. Diet should be based on a variety of common foods to provide other nutrients for which human requirements have been less well defined.

[b] Retinol equivalents. 1 retinol equivalent = 1 µg of retinol or 6 µg of β-carotene.

[c] As cholecalciferol. 10 µg of cholecalciferol = 400 I.U. vitamin D.

[d] Tocopherol equivalents. 1 mg of *d*-α-tocopherol = 1 α T.E.

[e] 1 N.E. (niacin equivalent) is equal to 1 mg of niacin or 60 mg of dietary tryptophan.

[f] The folacin allowances refer to dietary sources as determined by *Lactobacillus casei* assay after treatment with enzymes (conjugases) to make polyglutamyl forms of the vitamin available to the test organism.

[g] The RDA for vitamin B_{12} in infants is based on average concentration of the vitamin in human milk. The allowances after weaning are based on energy intake (as recommended by the American Academy of Pediatrics) and consideration of other factors such as intestinal absorption.

[h] The increased requirement during pregnancy cannot be met by the iron content of habitual American diets nor by the existing iron stores of many women; therefore, the use of 30–60 mg of suplemental iron is recommended. Iron needs during lactation are not substantially different from those of nonpregnant women, but continued supplementation of the mother for 2–3 months after parturition is advisable to replenish stores depleted by pregnancy.

Table 11. Approximate percentage contribution of fruit–vegetable food group to total nutrient supplies [52]

Energy and nutrients	Fruit and vegetable contribution, %					
	Citrus fruits	Other fruits	Potatoes and sweet potatoes	Dark-green and deep-yellow vegetables	Other vegetables including tomatoes	Total, %
Energy	0.9	2.2	2.7	0.3	2.5	9
Calcium	0.9	1.2	0.9	1.6	4.9	10
Phosphorus	0.7	1.1	3.9	0.7	5.0	11
Iron	0.8	3.3	4.4	1.6	9.0	19
Magnesium	2.2	3.9	7.1	2.1	10.4	26
Vitamin A	1.5	5.5	5.3	21.2	15.5	49
Thiamin	2.8	1.8	6.2	0.9	6.9	18
Riboflavin	0.5	1.5	1.7	1.1	4.5	9
Niacin	0.9	1.7	7.1	0.7	6.1	16
Vitamin B_6	1.2	5.5	11.2	1.7	9.2	28
Ascorbic acid	26.3	11.4	18.0	8.3	27.6	92

Table 12. Stability of nutrients∗ [50]

	pH			Air or oxygen	Light	Heat	Max. cooking loss, %
	pH < 7 (acid)	pH 7 (neutral)	pH > 7 (alkaline)				
Vitamins							
Vitamin A	U	S	S	U	U	U	40
4-Aminobenzoic acid	S	S	S	U	S	S	50
Ascorbic Acid (C)	S	U	U	U	U	U	100
Biotin	S	S	S	S	S	U	60
Carotene (pro-A)	U	S	S	U	U	U	30
Choline	S	S	S	U	S	S	5
Cobalamin (B_{12})	S	S	S	U	U	S	10
Vitamin D	U	S	U	U	U	U	40
Folic acid	U	U	S	U	U	U	100
Inositol	S	S	S	S	S	U	95
Vitamin K	U	S	U	S	U	S	5
Niacin (PP)	S	S	S	S	S	S	75
Pantothenic acid	U	S	U	S	S	U	50
Pyridoxine (B_6)	S	S	S	S	U	U	40
Riboflavin (B_2)	S	S	U	S	U	U	75
Thiamin (B_1)	S	U	U	U	S	U	80
Tocopherol (E)	S	S	S	U	U	U	55
Essential amino acids							
Isoleucine	S	S	S	S	S	S	10
Leucine	S	S	S	S	S	S	10
Lysine	S	S	S	S	S	U	40
Methionine	S	S	S	S	S	S	10
Phenylalanine	S	S	S	S	S	S	5
Threonine	U	S	U	S	S	U	20
Tryptophan	U	S	S	S	U	S	15
Valine	S	S	S	S	S	S	10
Essential fatty acids	S	S	U	U	U	S	10
Mineral salts	S	S	S	S	S	S	3

∗ S = stable (no significant destruction); U = unstable (significant destruction)

ble 10 shows the Recommended Daily Dietary Allowances, which was published by the United States National Academy of Science [51].

Most natural foods contain several vitamins, but the number and concentration vary, as illustrated in Table 11. Just like other constituents, the concentration of vitamins in a particular food depends on factors such as genetics, climate, weather, soil fertility for plants or feed for animals, maturity at the time of harvest or slaughter, postharvest handling and storage, and processing or cooking procedures. The effects of pH, oxygen, light, and heat on vitamin stability are shown in Table 12.

Table 13. Types of enzyme-catalyzed food modifications∗

Type of modification	Enzyme	Consequence
Flavor	ribonucleases	producing nucleotide flavor enhancers
	lipase	producing strong flavored, volatile free fatty acids
	protease	producing bitter flavor in cheese
	sulfhydryl oxidase	reducing cooked flavor in milk
	naringinase	debitterizing grapefruit juice
	thioglycosidase	regenerating flavors in dehydrated cabbage
Color	polyphenoloxidase	enzymatic browning of fruits, tea, and coffee
	lipoxygenase	oxidizing carotenoid pigments (e.g., bleaching of flour)
	glucose oxidase	removing sugar from eggs and reducing their browning
	pectinases	clarifying fruit juices
Texture	bromelain, papain	tenderizing meats
	cellulases	tenderizing vegetables
	amylases	reducing viscosity of starch gels
	rennet	coagulating milk
	polygalacturonase	softening of fruits
Nutritive value	cellulases	removing inedibles
	tannase	removing inhibition of protein digestion by tannins
	lactase	compensating for lactase deficiencies of some people
	thiaminase	destroying thiamin

∗ Taken from information in [7], [27], [28].

3.7. Enzymes

Enzymes are proteins with highly specific catalytic properties (→ Enzymes), [6], [7], [27], [28], [43]. Enzymes are found in all living organisms and catalyze many of the reactions that occur in natural products. Foods may contain endogenous enzymes, enzymes from microbial sources, as well as enzymes that are added for specific purposes. Enzymes may affect flavor, color, texture, and nutritive value of foods (see Table 13).

The velocity of enzyme-catalyzed reactions depends on temperature, pH, ionic strength, and the concentrations of enzyme, substrate, activators, and inhibitors. A plot of velocity versus temperature, pH, or ionic strength usually produces a bell-shaped curve. The peak of the curve represents the optimal conditions. The optimal conditions for some enzymes are restricted to a narrow range, whereas others exhibit a broad range. Because enzymes are proteins, they can be altered by denaturating agents or treatments. Thermal processing and freezing can markedly reduce enzymatic activity.

Enzymes vary with respect to their heat sensitivity, but they each have a specific inactivation temperature. Therefore, processing temperatures that inactivate undesirable enzymes can be chosen. Vegetables may be scalded or blanched to inactivate enzymes prior to freezing. The temperature needed to inactivate enzymes in foods may be higher than that required in simple model systems. For example, the fat and other solids in milk protect the lipase so that the heat treatment required to inactivate lipase increases as the solids content increases [53]. Some enzymes may become reactivated during storage (e.g., phosphatase in cream and catalase in spinach).

The rate of an enzymatic reaction can be influenced by factors affecting the contact between enzyme and substrate. Rupture of cell walls or other protective membranes by mechanical means or by freezing can increase the reaction rate. Homogenization of fat-containing fluid dairy products increases lipolysis by disrupting the fat globule membrane and increasing the surface area of the substrate.

3.8. Miscellaneous Components

3.8.1. Flavor Components

The flavor components of foods consist of substances that elicit taste and odor responses (→ Flavors and Fragrances; → Foods, 3. Food Additives, Chap. 4.7.;→ Foods, 3. Food Additives, Chap. 4.8.;→ Foods, 3. Food Additives, Chap. 4.10.;→ Foods, 3. Food Additives, Chap. 4.11.) [6], [7], [25], [53–58]. Salt, acid, sweet, and bitter are the main taste substances. There are many more odor components, including those derived from the degradation of the major constituents (i.e., lipids, carbohydrates, and proteins) plus a variety of other volatile aliphatic, aromatic, and heterocyclic compounds. Thousands of odor or aroma compounds have been identified, but attempts to classify them have not been completely successful.

Lipids degrade to give potent flavor compounds, including volatile fatty acids, lactones, methyl ketones, and a variety of carbonyl compounds arising from oxidation. Protein hydrolysis yields bitter-tasting amino acids and peptides. Microbial degradation of proteins pro-

duces amines and other offensive-smelling compounds. Hydrolysis of polysaccharides can yield sweet-tasting monosaccharides and polysaccharides. Carbohydrate fermentation yields a variety of alcohols and esters. Heat-induced interactions between carbohydrates and proteins yield numerous flavorful Maillard browning reaction products, including nitrogen heterocyclics (pyrroles, pyrazoles, pyridines, pyrazines, and oxazolines) and oxygen heterocyclics (furans, pyrones, furfuryl alcohols, furanones, and maltols).

Sulfur-containing compounds, such as disulfides, trisulfides, isothiocyanates, and thiazoles, have potent odors and are found in vegetables such as cabbages, cauliflower, onions, and garlic. Sulfur compounds may also originate from light- or heat-induced degradation of sulfur-containing amino acids [53], [55].

Esters contribute to the flavor of many fruits, but each fruit has its unique blend of esters and other components. In some cases, specific components that contribute to the unique flavor are called character impact compounds. Isoamyl acetate, ethyl 2-methylbutyrate, methyl anthranilate, and γ-decalactone have been identified as impact compounds in bananas, apples, grapes, and peaches, respectively [55].

Astringency, the puckery sensation that is perceived in the mouth, results from stimulation of the trigeminal nerve system rather than the chemoreceptor system that responds to taste. Tannins or other polyphenols are responsible for the astringency of products like tea, wine, and cranberry juice. *Pungency,* the stinging sensation that is perceived in the nose, also involves trigeminal nerve stimulation. Many spices contain pungent components including capsaicin from chili peppers and gingerol from ginger.

3.8.2. Natural Pigments [6], [7], [35]

The appearance and color of food are important attributes because the consumer's first judgement of its quality is based on its appeal to the eyes. The color indicates the maturity of fruits and vegetables, the freshness of meats and salads, and the extent of nonenzymatic browning in the case of heated products.

The chlorophylls, hemoglobins, and myoglobins are *tetrapyrrole pigments,* which have a basis skeleton of four pyrrole rings connected by methine bridges with a metal in the center of the complexes. The chlorophylls are magnesium chelates, and the other two pigments are iron chelates. The chlorophylls are esterified with phytol (a C_{20} alcohol), which accounts for their solubility in organic solvents and insolubility in water. Myoglobin is an iron-containing porphyrin linked to a 153-residue polypeptide. Hemoglobin contains four polypeptide chains and four porphyrin or heme groups. The presence of iron in these pigments accounts for their sensitivity to oxygen.

Isoprene groups provide the basic structure for the *carotenoid pigments.* These pigments include the carotenes, which are hydrocarbons, and the xanthophylls, which are oxygenated derivatives of the carotenes. They are soluble in fat and are responsible for many of the yellow and red colors of plant and animal products and the red color of shells, exoskeletons, and skins of crustaceans, mollusks, and fishes. The carotenoids are unsaturated and, therefore, are susceptible to oxidation when exposed to oxygen and light. They become oxidized quite rapidly in the presence of unsaturated lipids and lipoxygenase.

A number of water-soluble *phenolic constituents* of plants contribute to the color of their leaves, flowers, and fruits. The red anthocyanins and yellow flavonoids are the principal phenolic chromophores. Both of these types are glycosidic derivatives of phenylbenzopyrylium (flavylium) salts, but they differ with respect to the substitution on the flavylium nucleus. Tannins, which range in appearance from colorless to yellow or brown, represent another group of phenolic compounds. The two types of tannins are condensed tannins, which are polymers of catechin or related compounds, and hydrolyzable tannins, which are polymers of gallic acid esters.

Betalains are a group of *nitrogen-containing compounds* with colors resembling those of the anthocyanins and flavonoid pigments. Betanin is a red betalain that is found in beets and has potential as a replacement for synthetic red colors.

A summary of some characteristics of pigments is listed in Table 14.

Table 14. Characteristics of natural pigments

Pigments	Color in		Sensitivity to [a]	
	Acid	Base	Oxygen	Heat
Anthocyanins [b, d]	red	blue	+	++
Flavonoids [b]	white	yellow	+	+
Carotenoids [c]	orange	orange	++	Ins.
Chlorophylls [c]	olive green	bright green	Ins.	+
Myoglobins [b]	red	darkens	+	+
Betalains [b] (from beets)	red	red [e]	+	+

+ = moderately sensitive; ++ = highly sensitive.
[a] Ins. = Insensitive;
[b] Water soluble.
[c] Water insoluble.
[d] React with Fe, Al, or Sn to give blue or green colors.
[e] Betanin, the major pigment of beets, is red over the pH range of 3.5 – 7.0; above pH 7 the color shifts toward blue, which becomes obvious above pH 9.

3.8.3. Undesirable or Potentially Undesirable Constituents

Food may contain constituents that are toxic or that reduce the availability of nutrients or impart unacceptable sensory characteristics [5–7], [49], [59]. Some of these undesirable constituents may be naturally present in food, may be produced by microbes, or may be unintentionally introduced or produced during processing, handling, or storage of food. Even some intentional additives, such as saccharin, may be considered undesirable by some individuals.

Some naturally occurring undesirable components are listed in Table 15. Fungi can produce potent toxins including ergot, which has been found on wheat, and aflatoxin on peanuts. Botulinum toxin is the most potent bacterial toxin, but the enterotoxins from *Staphylococcus aureus* are more common. Unintentional additives that are potentially dangerous include fumigants, such as ethylene oxide and methyl bromide; solvents, such as trichloroethylene, which reacts with cysteine to form toxic components; polycyclic aromatic hydrocarbons from wood smoke; and numerous pesticides, herbicides, and antibiotics that are used in the production of foods.

Table 15. Poisons and antinutritional properties of foods [5]

Poison and antinutritional factor	Disease, ailment, or disease potential	Food
Allyl isothiocyanate	skin irritant	horseradish, mustard
Amines, biological	can elevate blood pressure with certain tranquilizers; can cause death	natural cheese, Chianti (wine), yeast extracts, bananas, pineapples, tomatoes, pumpkins, squash, cucumbers
Carcinogens		
Safrole	found to cause cancer in mice	cinnamon, nutmeg, sassafras
Tannic acid	known to cause liver tumors in rats	tea, nuts, fruits
Benzopyrene	known carcinogen	cabbage, charcoal-grilled meat, spinach, lettuce, tea
Lemon and sesame oils	cocarcinogenic properties	lemon and sesame oils
Enzyme inhibitors of		
Trypsin	growth retardation in animals	soybeans, kidney beans, lima beans
Chymotrypsin	growth retardation	Irish potatoes, egg whites
Cholinesterase	block transmission of nerve impulses	radishes, carrots, celery, potatoes, broccoli, eggplant, sugar beet, asparagus, Stayman apple, Valencia orange
Solanine	cholinesterase inhibition, teratogen, abortion, and fetal resorption agent	potatoes (blighted or stressed)
Factor produced by Faba beans	hemolytic anemia in persons predisposed to toxic factor	Faba bean (*Vicia faba*)
Glycosides		
Cyanogens	cyanide poisoning	Lima beans, cassava, almonds, black-eyed peas, wild black cherries, kidney beans, peach pits, apricot pits, choke cherry pits

4. Food Law

Food regulations and laws emerged as soon as consumers began to depend on others for some of their food [1], [2], [49], [60–70].

Table 15. (Continued))

Goitrogens	goiter	soybeans, rapeseed, white turnips, cauliflower, peaches, pears, strawberries, Brussels sprouts, spinach, carrots, rutabaga, cabbage
Glycosides of some honey	potent heart stimulant	honey from bees obtaining sugars from mountain laurel, rhododendron, azalea, oleander
Gossypol (or sterculic acid)	poisoning of farm animals, reproduction problems in rats and chickens, cocarcinogen in trout	pigment gland of cottonseed
Hemagglutinins	growth inhibition	soybeans and other legumes
Metal binding factors		
Phytates	interfere with calcium absorption	cereals
Oxalates	interfere with calcium absorption	spinach, cashews, almonds, cocoa, rhubarb, Swiss chard, beet tops, parsley, celery, currants, prunes, beans, soybeans
Mushroom poisons	food poisoning	inedible mushroom varieties
Myristicin	hallucinogen	nutmeg, carrots, celery
Nitrates (and nitrites)	cocarcinogen methemoglobinemia	spinach, smoked foods, green leafy vegetables
Shellfish poisoning	lethal toxin of shellfish consuming certain algae	shellfish
Stimulants (caffeine, theobromine, theophylline)	hypertension	coffee, tea, cocoa
Vitamin A	found to cause disturbances in rat offspring	cheese, butter, yellow vegetables, fruits
Vitamin D	vascular lesions, mental retardation; vitamin D and nicotine-arterial thrombosis in susceptible rabbits	butter, egg yolks, fish oil, liver, salmon, tuna

History. In 1700 BC the Babylonians created laws prohibiting adulteration of wines. Mosaic and Egyptian laws outlined procedures for selecting and slaughtering cattle and the subsequent handling of meat. The Egyptians also described the proper procedure for making bread. Several centuries before Christ, India had regulations prohibiting adulteration of grain and edible fats and oils. Roman laws required products to be unspoiled. They also created an inspection service to oversee commercial meat and bread operations. The Romans in Constantinople established stringent measures to prevent adulteration of spices. In 1120 the death penalty was threatened in Germany for those who adulterated wine. In 1319 a butcher was put in pillory for selling putrid beef. HENRY VIII forbade the use of hops in beer. In Germany in 1532 under CHARLES V, penalties were established for adulterating spices and other merchandise.

Public opinion about food quality was aroused in England in 1820 by FREDERICK ACCUM's treatise on food adulteration [71]. According to him, it was not uncommon to find sulfuric acid in vinegar; gravel, leaves, and twigs in black pepper; roasted grains, carrots, or scorched beans and peas in coffee; water in milk; alum, burned sugar, lime, and lead salts in wine; and sand, dust, lime, and pulp in sugar [72]. Similar adulterations also occurred in other countries, which stimulated the proliferation of regulations.

The rate and sequence of development of food laws varied from country to country. However, in all countries the main objective was to ensure a pure, safe food supply. Most of the early laws were designed to prevent adulteration of foods and misrepresentation of quality. This led to the development of standard weights and measures, as well as qualitative methods to detect impurities. At first, purity was determined by visual inspection, sieving, or filtering to separate foreign objects and by tasting to detect flavor abnormalities [72]. The demand for more regulations grew as consumers became more dependent on foods from distant places and on processed or formulated foods. Laws that defined product contents were developed. At first these definitions were expressed in terms of ingredients used or methods of preparation. Regulatory definitions became more specific and detailed as science progressed. Standards of identity were established for foods, including cereal and dairy products, salad dressings, vegetables oils, margarines, butter, and canned fruits and vegetables. Many of these standards stipulated the composition of the products so that failure to meet these specifications constituted a violation of the law. These standards put pressure on the food producers and processors to maintain quality.

During this century there has been a tremendous increase in concern about the wholesomeness and safety of food and a concomitant in-

crease in rules and regulations [70]. Regulations cover every aspect of production, processing, and handling of foods. At the production level, regulations govern the use of pesticides, herbicides, antibiotics, and in some cases, hormones. With perishable products like milk, regulations govern the type of milking facilities and equipment, as well as storage conditions. At the processing level, regulations are even stricter and include rules pertaining to type and design of processing equipment, processing and sanitizing procedures, packages, packaging procedures, and storage conditions.

Research showing the relationship between nutrient intake and health increased public concern about the nutritional quality of foods. These concerns led to the development of dietary guidelines in the United States and to rules governing the use of nutritional claims on food labels and in food advertisements. Regulations establishing maximum fat contents for products such as hamburgers arose from concern over excessive fat consumption. On the other hand, laws were introduced requiring fortification of low-fat milk with vitamin A because of concern over the impact that lowering fat intake might have on vitamin A intake. This simply illustrates the risks and benefits associated with dietary modification. Many regulatory issues address similar but more complex problems and are the subject of numerous publications [49], [59], [69], [70].

The "Delaney Clause" is one of the most controversial pieces of U.S. food regulations. It is part of the 1958 food additive amendment, which forbids the addition to food of any material shown to produce cancer in animals in any dose. This restriction has helped to prevent the approval of potentially harmful additives, but in some cases it may have prevented the use of beneficial additives. As analytical methods have become more sensitive, previously undetected substances are now being detected. In some cases the analytical threshold may be lower than the toxicity threshold.

The agencies responsible for making and enforcing laws vary from country to country. In the United States, some foods are also regulated by state, county, and municipal governments. Even within the same government, regulatory functions are divided. For example, in the United States the Food and Drug Administration regulates most foods and beverages, but the U.S. Department of Agriculture (USDA) regulates meat and poultry products, and the Bureau of Alcohol, Tobacco, and Firearms (BATF) regulates alcoholic beverages [62]. The Federal Trade Commission (FTC) regulates advertising. Each country has its own regulatory organization so that a food company wishing to export its product must determine local regulations. Many trade organizations and governments try to standardize regulations to facilitate trading, although occasionally they deliberately create barriers.

The International Dairy Federation (IDF) has contributed much to the standardization of dairy products and regulations. Regional alliances, such as the European Economic Community [68], and worldwide organizations, such as the International Standard Organization (ISO) [67], Food and Agriculture Organization (FAO), and World Health Organization (WHO) are working on standardization of all food products and laws [61]. The FAO/WHO Codex Alimentarius Commission has probably contributed the most toward standardization.

5. Documentation

Besides numerous scientific and trade journals, a variety of other sources contain information pertaining to foods. Textbooks [6], [7], [29], [35], [49], and encyclopedias [73–75] have extensive bibliographies. However, the most complete listing of food organizations and resources, including computer data banks, are contained in reference guides [76–79].

6. Feeds

History. The scientific feeding of livestock and poultry started as a means of disposing of refuse and byproducts of grains and other human foods. For over 200 years, water-powered grist mills were used to grind and process wheat and other grains destined for human consumption [80]. As society became more industrialized, oilseed processing, milk processing, and meat rendering yielded undesirable byproducts, the disposal of which was routinely accomplished through dumping in local lakes and rivers.

As livestock and poultry producers gradually learned about the physiological functions of

their animals, they became interested in using the inexpensive byproducts. As these products began to gain economic value to the producers, the commercial feed industries arose. To evaluate the *feeding value* of feedstuffs, two basic scientific approaches needed to be developed:

1) A definition of the nutritional requirements for various production methods needed to be established.
2) Reliable, predictable chemical analyses of feedstuffs needed to be correlated to the nutritional requirements of the animals.

Most of the early feeding studies were conducted in Germany [81]. The first set of scientific *feeding standards* is generally attributed to the German scientist Thaer, who in 1810 used meadow hay as a standard to which extractable nutrients from other feeds were compared and assigned "hay values." In 1827, Prout established protein, fat, and carbohydrates as essential organic nutrients. An extension of these principles can be found in Grouven's work (1859), which used this system of protein, fat, and carbohydrates to develop a sophisticated feeding standard system for farm animals [82]. Another German scientist, Wolff, in 1864, developed a feeding standard system based on actual farm trials in which digestibility of nutrients was measured. In 1897, Lehmann published the *Mentzel Lengerke Calendar,* which adapted Wolff's system so that standards were stated on a dry-matter basis, giving rise to what was to become known as the Wolff–Lehmann feeding standards [81].

Throughout the latter part of the 19th century and the first 50 years of the 20th century, other feeding standard systems arose throughout the world (available fuel values, Atwater, 1890; starch values, Kellner, 1907; Scandinavian Feed Unit, Woll, 1912; dairy standards by Haecker, 1914; digestible true protein and net energy, Armsby, 1915). In the United States, Henry first published a landmark book entitled *Feeds and Feeding* in 1898 [83]. The standards developed by Henry, as well as those carried on and refined by his later collaborator Morrison, were the most widely used in North America for the first half of the 20th century [81].

Currently, a wide range of feeding standards can be found throughout the world, largely reflecting the feeding systems of the country or region. In the United States and Latin America, the recommendations of the National Research Council (NRC) of the National Academy of Sciences provide the basis for nutritional research and livestock feeding. In the British Commonwealth countries, recommendations of the Technical Committee of the Agricultural Research Council (ARC) are used. Because of rapidly expanding information on livestock and poultry nutrition, nutritional standards are under continual review and modification, the results of which are then provided in revisions of the NRC and ARC standards.

6.1. Feed Definitions

In general, *feed* can be defined as edible material(s) that are consumed by animals and contribute energy or nutrients to the animal's diet (usually refers to animals rather than humans) [84]. *Feed ingredients*, on the other hand, are defined as component parts or constituents of any combination or mixture making up a commercial feed. In many countries, the terms *feedstuffs* or *feeding stuffs* can be used interchangeably with the term *feed ingredient.*

When ingredients are mixed in proper proportions as designed by scientific formulation, a *commercial feed* or compound feed is produced. Such feed may or may not be the sole item fed to livestock and poultry because various forms of commercial feeds may be designed to be mixed with locally provided ingredients to make a *complete feed,* i.e., one that is nutritionally adequate for maintaining life or promoting production.

Terminology describing various types of commercially available feeds is imprecise and varies not only from region to region, but also from company to company. However, Figure 1 provides a generalized concept of how feeds are categorized [85]. As is discussed in Section 6.2, various forms of livestock have anatomical and physiological differences that permit or prevent them from using different types of feedstuffs. For example, ruminants have the ability to digest and effectively use fibrous materials such as hays and forages, whereas nonruminants are restricted in their ability to use such feeds. A *complete feed–ruminant* refers to a commercially prepared product that provides all nutrients required to support the form of production for which it was designed, including roughage.

Complete feed–nonruminant refers to a commercially prepared feed that provides all nutrients required to support the form of production for which its was designed; this may or may not include roughage.

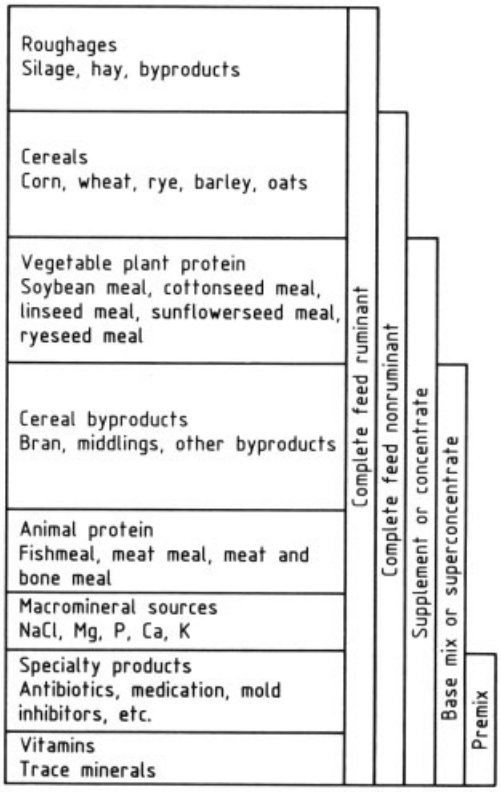

Figure 1. Feed classification

A *supplement* or *concentrate* is a commercially prepared feed that is formulated to be mixed with a primary grain or roughage to provide all the nutrients required to support the form of production for which it was designed. A *base mix* or *superconcentrate* is a low inclusion commercially prepared feed designed in such a way as to require the addition of a primary source of energy (grains), a primary source of protein (such as soybean meal), or roughage. *Premixes* are commercially prepared feeds designed to provide microingredients or specialty products to a mixture of feed ingredients to make a nutritionally complete feed.

6.2. Feed Utilization by the Animal

Feeds are formulated to address specific physiological requirements of the animal for which they are intended. Each species of livestock and poultry has its own unique metabolic requirements; this fact makes a feed designed for the laying hen different from that of the growing turkey, which is different from that of the milking cow. These basic differences can be understood by comparing the digestive tracts of various commercially raised livestock and poultry species.

From a simplistic view, mammalian livestock can be divided into two general categories: ruminants and nonruminants. Additionally, the digestive system of fowl is dramatically different from that of either the ruminant or the nonruminant.

Ruminants (cattle, sheep, and goats) have the unique ability to digest and use fibrous roughages and forages such as silage, grasses, and hays because of their compartmentalized stomach. Composed of four compartments, i.e., the rumen, reticulum, omasum, and abomasum, this stomach provides a complex digestive system that makes the ruminant an effective user of byproduct types of feeds. In effect, the rumen (the largest compartment of the system) is a huge fermentation vat harboring a wide variety of microorganisms (bacteria, yeasts, protozoa, etc.). It can effectively digest cellulose and complex carbohydrates to give simple volatile fatty acids essential for energy requirements. Additionally, a complex symbiotic relationship of the microflora allows the ruminant to form amino acids and, hence, proteins from nitrogenous compounds such as urea (Fig. 2) [86]. As is the case in most fermentation systems, microorganisms also produce a wide range of B-complex vitamins, therein reducing (and in most cases, eliminating) the need for B-vitamin supplements in the feed. Once the digested materials pass through the rumen complex, digestive processes in the small and large intestines are essentially analogous to those of nonruminants.

For *nonruminants* with simple gastrointestinal tracts, such as the pig and the dog, fiber digestion and microbial synthesis of B-vitamins is restricted. These types of animals are typically

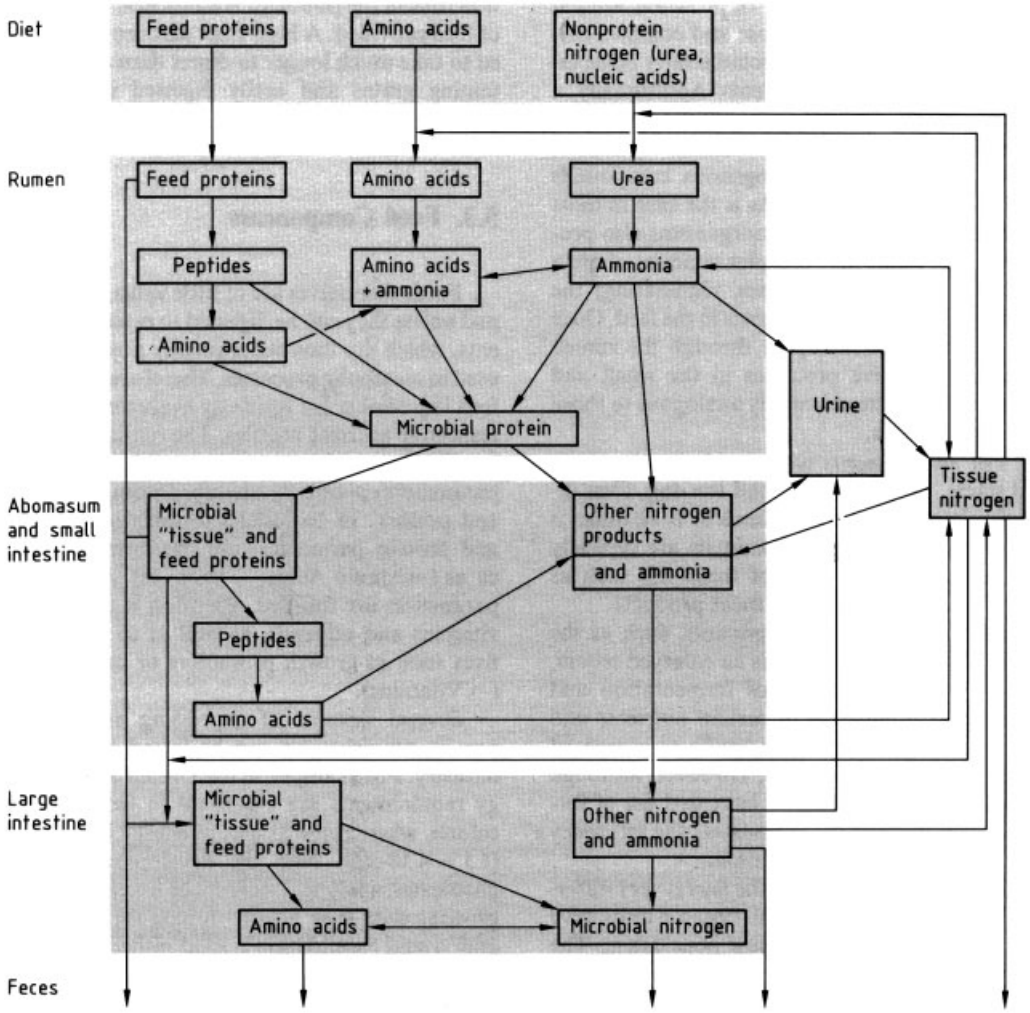

Figure 2. Protein and ammonia metabolism in the ruminant

fed easily digested types of feedstuffs such as grains, milk products, and meat products.

Several types of nonruminants, such as the horse and the rabbit, possess an enlarged cecum, which provides a means of fermentation and subsequent production of certain nutrients and increased use of fibrous feedstuffs as compared to the simple nonruminant. However, although the large cecum allows this increased use of forages, such animals do not possess the efficiency of digestion seen in the ruminant.

The digestive system of the *fowl* is very different from that of the mammal. Because birds have no teeth, mastication is almost nonexistent. The esophagus provides a passageway for feed to the crop, where the material is soaked with digestive enzymes and stored temporarily. The feed then passes to the proventriculus, which could be compared to the simple stomach of the mammal, where more digestive enzymes and acids are secreted to enhance chemical digestion. The feed is passed to the gizzard, a tough, muscular organ where the feed is mechanically crushed and ground, providing more surface area for digestion. The feed then passes to the small intestine, to the ceca, and then to the large intestine.

Feed retention is largely a function of the complexity of the digestive system. Because the digestive tract of the fowl is relatively small in

capacity and simple in structure, feed passage is rapid, as fast as 2.5 h in the laying hen. Conversely, in the ruminant where fermentation in the rumen is critical to the digestion process, feed retention time is long (up to several days). Retention time in the ruminant is a function of the type of feed provided. A high-fiber diet can be expected to take much longer to digest than a diet containing grains and easily digested ingredients [87].

6.3. Feed Components

Feeds themselves are of little value to the animal unless they can be digested to essential nutrients, which are then subsequently absorbed and used in metabolic processes. Therefore, feeds and feed ingredients are routinely evaluated for their respective nutrient profiles. The outline in Table 16 illustrates the analytical and nutritional parameters commonly used for feeding livestock and poultry. In basic feed formulation, energy and protein parameters are the primary references (→ Amino Acids; → Proteins). After these parameters are fulfilled, attention is directed to vitamins and minerals, as well as to feed additives such as growth promoters or medications (→ Vitamins).

Several systems of expressing energy and protein requirements can be found in the feed industry. For example, in the United States energy requirements are expressed in terms of the calorie, whereas the SI unit of energy is the joule (1 J = 4.18 cal). With the development of sophisticated analytical techniques for feed and the growing data base for fine-tuning formulations, amino acid formulation is used rather than total protein. In addition, many formulators use digestible, metabolizable or net energy, or protein values rather than gross energy or total protein.

6.4. Processing

Because feed is comprised of a number of raw commodities and industrial processing byproducts, the following methods of processing are used:

Particle size reduction (→ Size Reduction): grinding and rolling (crimping, cracking, and crushing)

Heating: pelletizing, extruding, flaking, roasting, cooking, and popping

Compaction: pelletizing, blocking, cubing, and crumbling

Chemical processing: tempering, fermentation

The advantages of processing are variable because of the type of animal being fed, the source of raw materials for the feed, and the wide variety of bulk densities and handling characteristics of feed ingredients. Some of the reasons for processing feed are as follows [88]:

1) Increase feed efficiency and production efficiency, thereby improving profit. However, processing is generally energy intensive, and the economic benefits must outweigh the costs of processing.
2) Alter particle size. Some ingredients need to be reduced in size for effective use by the animal and also to produce a feed that is homogeneous in particle size to prevent individual ingredients from segregating.
3) Change moisture content. Feeds can be too wet or too dry. A feed with too much moisture can physically restrict intake because of the limitation of gut "fill". Processing allows for removal or addition of water to enhance use.
4) Change the density of the feed. Bulky feeds may provide a physical "filling" of the animal, not allowing the animal to consume enough feed to meet its needs. Blocking or pelletizing produces a compact feed form, promoting increased consumption.
5) Change palatability. Often feeds containing unpalatable ingredients can have those tastes masked or blocked through processing.
6) Increase nutrient availability and eliminate detrimental effects of feed ingredients. Often raw commodities contain antinutritional factors (such as trypsin inhibitors in soybeans and in other pulses) and toxins (aflatoxins) or have limited nutritional values resulting from the physical structure of the ingredient particle (e.g., grains). Processing (heat treatment) can assist in minimizing the presence of antinutritional factors, and physical alteration of particles can increase surface area and improve digestion and absorption.

Table 16. Nutrient partitioning of feed

Energy	Gross	digestible	metabolizable	net
Protein∗	Nonprotein nitrogen	Protein		
		Nonessential amino acids		Essential amino acids
	Purines/pyrimidines Urea Ammonia	alanine asparagine aspartic acid cysteine cystine glutamic acid glutamine glycine proline serine tyrosine		arginine histidine isoleucine leucine lysine methionine phenylalanine threonine tryptophan valine
Vitamins∗∗	Fat soluble		Water soluble	
	A D E K		B-12 biotin choline folic acid niacin pantothenic acid pyridoxine riboflavin thiamine vitamin C	
Minerals	Macro		Micro	
	Calcium Phosphorus Potassium Magnesium Sulfur		cobalt copper iron iodine manganese molybdenum selenium zinc	

∗ Ruminants can effectively synthesize protein from nonprotein nitrogen, whereas poultry and nonruminants cannot; nonruminant ration formulation uses a combination of total protein and specific amino acid inclusion (methionine and lysine are the most commonly incorporated commercial amino acids).
∗∗ Ruminants can synthesize adequate amounts of water-soluble vitamins to fulfill nutrient requirements in most cases; vitamin C is generally not added to commercial feed.

6.5. Chemical Analysis

Although a feed may be complete chemically, if its constituent nutrients cannot be digested and metabolized by the animal, that feed is of little value. For example, a nonruminant has a very limited ability to convert roughages to energy because of the nature of the digestive system, while a ruminant is routinely fed fibrous materials.

In this age of computerization and sophisticated analytical technology, the primary means of chemical analysis of feed is still *proximate analysis,* developed over 100 years ago at the Weende Experiment Station in Germany. This method analyzes moisture, crude protein, ether extract, ash, crude fiber, and nitrogen-free extract.

Moisture. The water content of feed can be measured by oven drying at 135 °C until a constant mass is reached. In cases where feeds have a high content of volatile compounds, freeze drying, vacuum drying, and distillation with toluene are used.

Crude Protein. Because protein contains an average of 16 % nitrogen, crude protein values

are derived by determining the amount of nitrogen in feed by the Kjeldahl method of nitrogen analysis and multiplying this figure by 6.25. A "crude" approximation is obtained because not all proteins contain 16 % nitrogen and because not all the nitrogen in feed is associated with protein. However, this determination has been effectively used despite its inherent imprecision.

Ether Extract. By extraction with ether, fat content of feed can be approximated. The procedure also extracts other compounds present in the feed, but the magnitude of these components is generally negligible.

Ash. To determine the mineral content of feed, samples are placed in a muffle furnace and exposed to a temperature of ca. 600 °C. The material left after ignition is called the ash content. This fraction is often further analyzed for specific mineral content through such procedures as atomic absorption spectroscopy.

Crude Fiber. Because many animals have restricted capabilities for digesting fibrous materials, feeds are routinely analyzed for crude fiber. The extraction procedure produces a fraction that contains cellulose, hemicellulose, and lignin, all components not readily digested in nonruminant feeds. The Van Soest analysis has gradually replaced the crude fiber method of fiber determination. In the Van Soest analysis, feed samples are fractionated into neutral, detergent-soluble and -insoluble fractions. The soluble fraction is readily digested by most animals. The insoluble fraction is then further fractionated with an acid detergent procedure. Hemicellulose compounds are found in the soluble fraction, whereas lignocellulose (that of restricted availability) can be found in the insoluble fraction.

Nitrogen-Free Extract. This component of feed is considered to provide a crude representation of the soluble carbohydrates and nutrients not extracted by the previously mentioned procedures. It is derived mathematically by subtracting the percent moisture, ash, ether extract, crude protein, and crude fiber from 100.

In addition to the proximate analysis methods, a wide range of chromatographic, biological, and colorimetric procedures are used to identify and quantify individual nutrients in feed. *Gross energy values* of feed are routinely determined by bomb calorimetric procedures.

In vivo and in situ procedures using fistulation and cannulation surgical techniques are commonly used in experimental work to determine *digestibility* of feed components. Samples of feed are given to the animal in these evaluations, and material recovered from the stomach or intestinal regions is analyzed to see how the feed has been digested and absorbed.

6.6. Quality Control

Because feed ingredients are derived from plant sources or processing byproducts, they are inherently variable in nutrient content. Genetics plays an important role in the nutrient content of grains, as does the rate of fertilization, the climate, and the environment. Grain obtained from the same plot usually varies in nutrient composition from year to year. Additionally, mold and fungus growth is commonly found to arise from field and storage conditions. Toxins from these molds adversely affect the nutritional value of the feed.

Byproducts from industrial processing of oilseeds and rendered meat products are often highly variable, depending on the conditions and the processes used to make the primary products. Although the byproducts themselves have economic value to the manufacturer, they are not the primary products and, hence, are subject to a wide range of managerial influences.

Because of the high variability in the nutrient content of feed ingredients, feed manufacturers have developed considerable safeguards for their quality control programs. They must adhere to label guarantees of nutrient content and must demonstrate that their products are free of hazardous contaminants, such as pesticide residues or chemical contaminants.

For any quality control program to be successful, a detailed and accurate sampling program must be established. Probes are commonly used to inspect and sample dry materials, whether they are ingredients or finished feed. The size of sample, the number of samples, and the location of sampling are three primary considerations for such testing.

Devices which provide automated sampling include the bomb sampler for liquid bulk materials, diverter samplers for gravity spouts, auger samplers, and pneumatic probes [89].

With regard to actual feed or ingredient analysis, a combination of microscopic and visual inspections of products, as well as chemical assays, is routinely run for quality assurance. Microscopic analysis allows determination of the specific components of the feed source and also permits visual inspection both for mold or fungal contamination and for evidence of adulteration. Once the microscopic inspection is passed, routine chemical analyses, such as proximate analysis, are performed to verify that the product meets nutrient specifications.

6.7. Survey of Common Feeds

The profile of feed varies considerably from country to country and even from region to region within a country. Much of this variation can be attributed to the availability of locally grown or processed ingredients. In addition, governmental policies largely dictate priorities for feedstuffs. For example, corn is the most prominently used feed ingredient for animals in the United States. However, in Mexico, corn is considered to be a primary source of energy for the human population and is discouraged, therefore, from being used in animal feeds. Furthermore, governmental subsidies and tariffs play an important role in dictating availability of certain ingredients to be incorporated in feed.

All *feed ingredients* contribute a range of nutrients to the complete feed. However, ingredients are generally classified as to their respective primary contribution of nutrients, i.e., feed ingredients having high energy values are classified as energy feeds; those high in protein are classified as protein feeds. The following is a list of commonly used feed ingredients throughout the world:

High Energy Feeds
Grains: corn, barley, oats, rye, sorghum, rice, millet, and wheat (→ Cereals and Cereal Products)
Byproducts: manioc (cassava), molasses, potatoes, sugar beets, and bakery byproducts
Fats: acidulated soapstocks, tallow, and grease (→ Fats and Fatty Oils)

Protein Feeds (→ Proteins)
Oilseed meals: soybean, rapeseed, linseed, coconut, peanut, cottonseed, safflower, sesame, and sunflower (→ Fats and Fatty Oils)
Pulses: beans, chick-pea, cowpeas, peanuts, and soybeans
Meat byproducts
Poultry byproducts
Dairy processing byproducts
Marine byproducts
Nonprotein nitrogen: urea and biuret (→ Urea)
Distillers byproducts
Single-cell protein

Roughages
Hay
Silage
Pasture
Plant processing byproducts: pulp, skins, peels, and shells

Minerals

Vitamins: commercially prepared (→ Vitamins)

Feed Additives
Medications
Anthelmintics (→ Anthelmintics)
Growth promoting antibiotics (→ Antibiotics)
Hormones (→ Hormones)
Antioxidants and preservatives (→ Antioxidants)
Mold inhibitors
Flavors (→ Flavors and Fragrances)
Pellet binders
Buffers
Microbial cultures

In addition to the byproducts of food processing and the raw commodity type ingredients (e.g., grains and forages), *feed additives* are commonly used to enhance production, prevent disease, improve the commercial appeal of the product, stabilize the shelf life of the feed and improve acceptability to the animal. The list of feed additives is extremely long and varies from country to country because many countries place restrictions as to what can be added to feed, especially in the area of medications.

Table 17. Compound feed production for both developed and developing countries between 1974 and 1981 [90]

	Average production, 10^3 t/a						Growth rates
	1974 – 1976	1977	1978	1979	1980	1981	1974 – 1976 to 1981, %/a
World	289 855	316 871	342 033	360 495	371 442	376 913	4.5
Developing countries	20 385	27 772	32 867	37 011	40 683	43 236	13.4
Africa	595	744	947	1 165	1 510	1 834	20.8
Morocco	80	84	137	165	210	274	27.5
Nigeria	135	210	260	300	400	460	22.5
Tunisia	130	200	250	350	450	600	29.0
Latin America	13 260	16 270	18 833	20 370	21 567	22 600	9.3
Argentina	1 100	1 300	1 500	1 800	1 850	1 900	9.5
Brazil	6 900	8 600	10 300	11 100	11 400	11 700	9.2
Columbia	290	450	600	700	1 100	1 200	26.2
Mexico	3 167	3 780	3 930	4 400	4 600	5 000	7.9
Peru	320	340	403	320	467	500	7.7
Venezuela	1 233	1 500	1 700	1 700	1 700	1 800	6.5
Near East	784	1 532	2 287	3 031	3 861	4 212	32.5
Egypt	150	490	800	1 100	1 640	1 700	49.5
Iran	100	160	190	200	200	220	14.1
Libya	2	4	6	10	12	22	49.5
Saudia Arabia	15	90	180	250	300	400	73.0
Turkey	484	712	937	1 271	1 449	1 500	20.5
Far East	5 676	9 026	10 500	12 045	12 725	13 140	15.0
India	700	900	1 000	1 100	1 150	1 150	8.6
Korea (Republic)	1 070	1 890	2 700	3 300	3 500	3 600	22.5
Malaysia	319	386	444	465	475	490	7.4
Philippines	747	830	961	1 100	1 150	1 200	8.2
Sri Lanka		50	90	100	100	100	
Thailand	700	900	1 005	1 150	1 100	1 100	7.8
Asia	70	200	300	400	1 020	1 450	66.0
China	70	200	300	400	980	1 400	66.0
Developed countries	269 470	289 099	309 166	323 484	330 759	333 677	3.6
North America	114 667	115 500	121 100	123 800	125 000	125 400	1.5
Canada	9 567	10 000	10 300	11 000	11 200	11 500	3.1
United States	105 100	105 500	110 800	112 800	113 800	113 900	1.3
Western Europe	77 222	86 141	92 639	100 232	104 190	104 877	5.2
EEC	60 433	67 400	71 720	77 170	79 100	79 890	4.7
Belgium-Luxembourg	4 933	5 000	5 000	5 000	4 900	4 778	− 0.5
Denmark	3 000	3 700	4 200	4 750	4 842	4 753	8.0
Federal Republic of Germany	11 767	14 000	14 900	16 438	16 796	17 199	6.5
France	11 500	12 500	13 300	14 000	14 700	15 150	4.7
Ireland	1 100	1 400	1 600	2 000	1 800	1 860	9.1
Italy	6 600	8 100	8 800	10 000	10 400	10 457	8.0
Netherlands	10 867	12 300	12 700	14 000	14 500	14 570	5.0
United Kingdom	10 633	10 800	11 000	11 600	11 100	11 007	0.6
Austria	833	900	1 000	1 000	1 200	1 067	4.2
Norway	1 267	1 300	1 300	1 500	1 500	1 400	1.7
Portugal	1 933	2 800	2 600	2 800	3 500	3 637	11.1
Spain	8 133	7 400	8 900	10 500	11 200	11 000	5.2
Switzerland	706	740	750	750	780	793	2.0
Yugoslavia	1 467	2 801	3 269	3 412	3 410	3 400	15.0
Eastern Europe and USSR	55 467	62 400	68 300	72 300	74 700	76 300	5.5
Czechoslovakia	2 450	2 100	2 300	2 100	2 100	2 100	− 2.5

6.8. Storage and Transportation

The economics of feed production are such that transportation and storage are critical to making a profit. Because margins are relatively small on most feeds, transporting ingredients long distances is economically infeasible, especially when local products can be effectively substituted. Inventories are designed for rapid product turnover to prevent spoilage and to minimize the time raw materials and finished products are stored.

Transportation of raw materials and finished products is a function of the local/regional transportation facilities. In the Netherlands, almost all

Table 17. (Continued)

	Average production, 10^3 t/a						Growth rates
	1974 – 1976	1977	1978	1979	1980	1981	1974 – 1976 to 1981, %/a
Poland	4 767	4 200	4 700	4 200	4 200	3 800	– 3.8
USSR	44 050	51 300	56 000	60 000	62 000	64 000	6.4
Oceania	1 150	1 680	2 210	2 350	2 390	2 420	13.2
Australia	1 000	1 500	2 000	2 100	2 100	2 100	13.2
New Zealand	150	180	210	250	290	320	13.5
Other developed	20 964	23 378	24 917	24 802	24 479	24 680	2.7
Israel	1 570	1 600	1 650	1 702	1 579	1 580	0.1
Japan	17 594	19 878	21 067	20 800	20 500	20 500	2.6
South Africa	1 800	1 900	2 200	2 300	2 400	2 600	6.3

Table 18. Consumption of feed by livestock group

Feed	Consumption, %					
	Cattle [a] and buffalos	Sheep and goats	Pigs	Poultry [b]	Other [c]	All livestock
Total feed	56	12	10	7	15	100
Cereals	35	2	32	27	4	100
Oilmeals	21	3	28	45	3	100
Other concentrates	37	7	39	13	4	100
Total roughage	63	15	2	1	19	100

[a] For meat and milk production only.
[b] For egg and meat production.
[c] Mainly draught animals.

Table 19. Pattern of feed consumption in terms of metabolizable energy

Animal	Consumption, %				
	Cereals	Oilmeals	Other concentrates	Roughages∗	Total
Cattle and buffalos	10.4	0.9	4.2	84.5	100.0
Sheep and goats	2.3	0.5	3.4	93.8	100.0
Pigs	52.0	6.3	24.0	17.7	100.0
Poultry	64.2	15.1	12.1	8.6	100.0
Other∗∗	4.4	0.5	1.8	93.3	100.0
Total	16.5	2.3	6.3	74.9	100.0

∗ Including pastures, harvested roughages, crop residues, and household wastes.
∗∗ Mainly draught animals.

raw materials are transported by barge, and the finished products are shipped locally by truck. In the United States, raw materials are generally shipped by rail or truck, but where applicable, barge freight is used.

Another factor affecting storage and transportation is centered around whether the area uses complete feeds or uses low inclusion feeds, such as base mixes or concentrates, in conjunction with materials produced on the site of feeding (such as grain). If a complete feed is being shipped, the pricing structure is such that shipping distances must be shorter than if a low inclusion feed with higher margins is being shipped.

Because raw materials and finished products are perishable, storage is a major consideration, especially where high heat and humidity are prevalent. Environmental factors can render feed unusable in a short period when mold growth or oxidation can take place.

6.9. Economic Aspects

Growth in the worldwide feed industry is highly variable depending on respective regions. Feed tonnage in developed countries has remained relatively static resulting from slow growth in human population and leveling incomes. In rapidly developing economies, a combination of

increasing per capita income and growing human population has created new demands for livestock and poultry products.

Tables 17, 18, 19 show FAO/UN estimates of worldwide feed production and consumption for the periods from 1974 to 1981 [90]. Approximately 1.6×10^9 livestock units (1 LU = 500 kg livestock) represent the world's animal population, 80 % of which are ruminants.

7. References

1. G. F. Stewart, M. A. Amerine: *Introduction to Food Science and Technology,* Academic Press, New York – London 1973.
2. M. S. Peterson, D. K. Tressler: *Food Technology the World Over,* AVI Publ., Westport, Conn., 1963.
3. M. E. Lowenberg, E. N. Todhunter, E. D. Wilson, M. C. Feeney, J. R. Savage: *Food and Man,* J. Wiley & Sons, New York – London – Sydney 1968.
4. M. Pyke: *Food and Society,* Murray, London 1968.
5. V. S. Packard, Jr.: *Processed Foods and the Consumer,* Univ. of Minn. Press, Minneapolis 1975.
6. J. M. DeMan: *Principles of Food Chemistry,* AVI Publ., Westport, Conn., 1980.
7. O. R. Fennema (ed.): *Food Chemistry,* 2nd ed., M. Dekker, New York – Basel 1985, 991 ff.
8. N. N. Potter: *Food Science,* 4th ed., AVI Publ., Westport, Conn., 1986.
9. C. A. Browne: *A Source Book of Agricultural Chemistry,* Chromica Botanica Co., Waltham, Mass., 1944.
10. A. J. Ihde: *The Development of Modern Chemistry,* Harper & Row, N.Y. 1964.
11. P. H. Abelson (ed.): *Food: Politics, Economics, Nutrition and Research,* Am. Assn. for Advancement of Science, Washington, D.C., 1975.
12. L. R. Brown: *Human Food Production as a Process in the Biosphere, Readings from Scientific American Food,* W. H. Freeman & Co., San Francisco 1973.
13. National Research Council: *Post Harvest Food Losses in Developing Countries,* National Academy of Sci., Washington, D.C., 1978.
14. F. A. Glover in R. K. Robinson (ed.): *Modern Dairy Technology I Advances in Milk Processing,* Elsevier Applied Sci. Publ., London – New York 1986, pp. 235 – 271.
15. J. G. Zadow in, [14] pp. 273 – 316.
16. P. T. Tybor, C. W. Dill, W. A. Landmann, *J. Food Sci.* **40** (1975) 155 – 159.
17. J. Wismer-Pedersen, *Food Technol. (Chicago)* **33** (1979) no. 8, 76 – 80.
18. R. C. Baker, D. C. Goodrich: *Poult. Sci.* **60** (1981) 332 – 335.
19. R. C. Baker, D. S. Kline, *Poult. Sci.* **63** (1984) 274 – 278.
20. R. E. Martin, *Food Technol. (Chicago)* **30** (1976) no. 9, 64.
21. J. M. Regenstein, *Food Technol. (Chicago)* **40** (1986) no. 3, 101 – 106.
22. R. J. Priestly (ed.): *Effects of Heating on Foodstuffs,* Applied Publ., London 1979.
23. J. E. Hodge, E. Osman in O. Fennema (ed.): *Principle of Food Science, Part I, Food Chemistry,* M. Dekker, New York – Basel 1976, pp. 41 – 138.
24. R. L. Whistler, J. R. Daniel, [7] pp. 69 – 137.
25. R. Bassette, D. Y. C. Fung, V. R. Mantha, *CRC Crit. Rev. Food Sci. Nutr.* **24** (1985) 1 – 52.
26. C. Eriksson (ed.): *Maillard Reactions in Food: Chemical Physiological and Technological Aspects,* Pergamon Press, Oxford 1981.
27. R. L. Ory, A. J. St. Angelo: *Enzymes in Food and Beverage Processing,* ACS Symposium Series 47, Washington, D.C., 1977.
28. G. Reed: *Enzymes in Food Processing,* 2nd ed., Academic Press, New York 1975.
29. O. R. Fennema (ed.): *Principles of Food Science, Part I, Food Chemistry,* M. Dekker, New York – Basel 1976, 792 ff.
30. H. W. S. Chan (ed.): *Autoxidation of Unsaturated Lipids,* Academic Press, San Diego, Calif., 1987.
31. W. O. Lundberg (ed.): *Autoxidation and Antioxidants,* Interscience, New York 1961.
32. W. F. Shipe, G. F. Senyk, K. J. Boor, *J. Dairy Sci.* **65** (1982) 24 – 27.
33. R. B. Duckworth (ed.): *Water Relations of Foods,* Academic Press, London – New York – San Francisco 1975.
34. D. H. Chou, C. V. Morr, *J. Am. Oil Chem. Soc.* **56** (1979) 53 A.
35. C. Zapsalis, R. Anderele Beck: *Food Chemistry and Nutritional Biochemistry,* J. Wiley & Sons, New York 1985.
36. R. Davies, G. G. Birch, K. J. Parker (eds.): *Intermediate Moisture Foods,* Applied Science Publ., London 1976.
37. P. F. Fox, J. J. Condon: *Food Proteins,* Applied Science Publ., London – New York 1982.

38. B. J. F. Hudson (ed.): *Developments in Food Proteins-5,* Elsevier Appl. Sci., London – New York 1987.
39. J. E. Kinsella, *CRC Crit. Rev. Food Sci. Nutr.* **21** (1984) 197 – 262.
40. J. E. Kinsella in P. F. Fox, J. J. Condon (eds.): *Food Proteins Applied Science Publ.,* London – New York 1982, pp. 51 – 103.
41. J. M. Regenstein, C. E. Regenstein: *Food Protein Chemistry,* Academic Press, New York 1984.
42. K. C. Rhee, C. K. Kuo, E. W. Lusas in J. P. Cherry (ed.): *Protein Functionality in Foods,* Am. Chem. Soc., Washington, D.C., 1981, pp. 51 – 88.
43. R. F. Feeney, J. R. Whitaker (eds.): *Food Proteins Improvement through Chemical and Enzymatic Modification,* Am. Chem. Soc., Washington, D.C., 1977.
44. T. R. Richardson, J. W. Finley (eds.): *Chemical Changes in Foods During Processing,* AVI Publ., Westport, Conn., 1986.
45. C. C. Bigelow, *J. Theoret. Biol.* **16** (1967) 187.
46. R. S. Shallenberger, G. G. Birch: *Sugar Chemistry,* AVI Publ., Westport, Conn., 1975.
47. G. R. Jurch, Jr., J. H. Tatum, *Carbohydr. Res.* **15** (1970) 233 – 239.
48. E. M. A. Hamilton, E. N. Whitney: *Nutrition-Concepts and Controversies,* 2nd ed., West Publ., St. Paul – New York – Los Angeles – San Francisco 1982.
49. S. R. Tannenbaum (ed.): *Nutritional and Safety Aspects of Food Processing,* M. Dekker, New York – Basel 1979.
50. R. S. Harris, E. Karmas: *Nutritional Evaluation of Food Processing,* AVI Publ., Westport, Conn., 1975.
51. Committee on Dietary Allowances of the Food & Nutrition Board (ed.): *Recommended Dietary Allowances,* 9th ed., National Academy of Sciences, Washington, D.C., 1980.
52. *National Food Situation,* Economic Research Service, U.S. Dept. of Agric. (USDA), Washington, D.C., 1974, p. 28.
53. W. F. Shipe in G. Charalambous (ed.): *Analysis and Control of Less Desirable Flavors in Foods and Beverages,* Academic Press, Orlando, Fla., 1980, pp. 201 – 239.
54. M. A. Amerine, R. M. Pangborn, E. B. Roessler: *Principles of Sensory Evaluation of Foods,* Academic Press, New York 1965.
55. R. Teranishi, R. A. Flath, H. Sugisawa (eds.): *Flavor Research Recent Advances,* M. Dekker, New York – Basel 1981.
56. D. G. Land, H. E. Nursten (eds.): *Progress in Flavour Research,* Applied Sci Publ., London 1978.
57. C. G. Birch, J. C. Breennan, K. J. Parker (eds.): *Sensory Properties of Foods,* Applied Science Publ., London 1978.
58. J. C. Boudreau (ed.): *Food Taste Chemistry, ACS Symposium Series 115,* Washington, D.C., 1979.
59. H. D. Graham (ed.): *Safety of Foods,* 2nd ed. AVI Publ., Westport, Conn., 1980.
60. F. L. Gunderson, H. W. Gunderson, E. R. Ferguson, Jr.: *Food Standards and Definitions in the United States,* Academic Press, New York – London 1963.
61. A. Gerard: *An Outline of Food Law: Structure, Principles, Main Provisions,* FAO, Rome 1975.
62. Y. H. Hui: *United States Food Laws, Regulations and Standards,* J. Wiley & Sons, New York – Chichester – Brisbane – Toronto 1979.
63. H. W. Schultz: *Food Law Handbook,* AVI Publ., Westport, Conn., 1981.
64. J. B. Dubeck, J. S. Elched: *Practical Food Law: Course Manual,* Federal Publications, Washington, D.C., 1983.
65. Ministry of Agriculture, Fisheries and Food (MAFF): *Selected Bibliography on Food Hygiene and Food Law,* MAFF, London 1983.
66. British Food Manufacturing Industries Res. Assn.: *Overseas Food Legislation,* 2nd ed., British Food Manufacturing Industries Research Assn., Leatherhead, England, 1984.
67. International Organization for Standardization: ISO Catalogue, Geneva, Switzerland, issued annually.
68. European Economic Community, *Food Legislation Manual,* Leatherhead Food Research Assn., Leatherhead, England, updated biannually.
69. T. P. Labuza, J. W. Erdman, Jr.: *Food Science and Nutritional Health,* West Publ., St. Paul, Minn., 1984.
70. I. C. Munro, R. Zeckhauser, B. N. LaDu, Jr., S. A. Miller et al., *Food Technol.* **33** (1979) no. 11, 42 – 67.
71. F. C. Accum: *A Treatise on Adulteration of Food and Culinary Poisons,* J. Mallett, London 1820.
72. F. A. Filby: *A History of Food Adulteration and Analysis,* George Allen & Unwin, London 1934.

73. D. N. Lapedes (ed.): *Encyclopedia of Food, Agriculture and Nutrition,* McGraw Hill, New York 1977.
74. M. S. Peterson, A. H. Johnson: *Encyclopedia of Food Science,* AVI Publ., Westport, Conn., 1978.
75. A. E. Ensminger, M. E. Ensminger, J. E. Konlande, J. R. K. Robson: *Foods and Nutrition Encyclopedia,* vol. 1 and 2, Pegus Press, Clovis, Calif., 1983.
76. I. S. Scarpa, H. C. Kiefer, R. Tatum (eds.): *Sourcebook on Food & Nutrition,* 3rd ed., Marquis Who's Who, Chicago, Ill., 1982.
77. R. C. Frank (ed.): *Directory of Food and Nutrition Information Services and Resources,* Onyx Press, Phoenix, Ariz., 1984.
78. S. Green: *Food Science and Technology,* Mansell Publ., London – New York 1985.
79. P. Szilard: *Food and Nutrition Information Guide,* Libraries Unlimited, Littleton, Colo., 1987.
80. R. Schoeff in R. R. McEllhiney (ed.): *Feed Manufacturing Technology,* American Feed Industry Ass., Arlington, Va., 1985, pp. 2 –12.
81. M. E. Ensminger, C. G. Olentine: *Feeds and Nutrition,* Ensminger Publ. Co., Clovis, Calif., 1978, pp. 557 –561.
82. L. A. Maynard, J. K. Loosli, H. Hintz, R. G. Warner: *Animal Nutrition,* McGraw Hill, New York, 1979, pp. 371 – 383.
83. W. Henry: *Feeds and Feeding,* Univ. of Wisconsin, Madison, Wisc., 1898.
84. Ass. of American Feed Control Officials, Official Publication, 1987.
85. C. G. Olentine, *Feed Manage.* **35** (1984, Jan.), no. 1, p. 6.
86. C. G. Olentine, *Feed Manage.* **37** (1986, Mar.) pp. 26 – 32.
87. D. C. Church: *Digestive Physiology and Nutrition of Ruminants,* D. C. Church, Publ., Oregon State Univ., Corvallis, Oreg., 1969, pp. 85 – 89.
88. M. E. Ensminger in, [80]pp. 529 – 533.
89. R. Schultz in, [80]pp. 335 –346.
90. *Feed Int.* **7** (1986, Jan.) no. 1.

Foods, 2. Food Technology

PAVEL JELEN, University of Alberta , Edmonton, Canada T6G 2P5

1. Introduction 2
1.1. Food Manufacture and Preservation 2
1.2. Food Spoilage by Microorganisms . 2
1.3. Chemical Deterioration of Food . . 3
1.4. Water Content and Water Activity 3
2. Food Preservation by Dehydration . 4
2.1. Concentration and Drying 4
2.2. Food Concentration Technology . . 5
2.3. General Aspects of the Food Drying Processes 6
2.4. Food Drying Technology 8
2.4.1. Convection Drying 8
2.4.2. Contact, Vacuum, and Freeze Drying 10
2.4.3. Radiant-Heat Drying 11
2.5. Instantization 11
2.6. Moisture Sorption Isotherms of Foods 11
2.7. Intermediate-Moisture Foods 12
3. Food Preservation at Low Temperature 12
3.1. Refrigeration and Freezing 12
3.2. General Aspects of the Food Freezing Process 13
3.3. Food-Freezing Technology 14
3.3.1. Convective Processes 14
3.3.2. Cryogenic Freezing 15
3.3.3. Contact Processes 16
3.3.4. Freezing System Capacities 16
3.4. Effects of Freezing on Foods 19
3.5. Frozen Storage and Thawing 20
3.6. Freeze Concentration and Freeze Texturization 20
4. Food Preservation by Heating 21
4.1. Pasteurization and Sterilization . . 21
4.2. Theory of Microbial Deactivation by Heating 22
4.2.1. Heat Resistance of Microorganisms . 22
4.2.2. Process Calculations for Canned Foods 24
4.3. Food Canning Technology 25
4.3.1. Sanitary Can Architecture 26
4.3.2. Conventional Retorting Equipment . . 27
4.3.3. Hydrostatic Sterilizers 27
4.3.4. Flame Sterilization 28
4.3.5. Retortable Pouches 28
4.3.6. Aseptic Canning 29
4.4. Ultra-High-Temperature Aseptic Processing 29
4.4.1. Theoretical Basis of UHT Processing 29
4.4.2. Ultra-High-Temperature Sterilization 30
4.4.3. Aseptic Packaging 30
4.4.4. Quality of UHT-Processed Foods . . . 31
4.5. Spoilage in Heat-Sterilized Foods . 32
5. Food Preservation by Ionizing Radiation 33
5.1. Radurization, Radappertization, and Radicidation 33
5.2. Theoretical Aspects of Food Irradiation 34
5.3. Food Irradiation Technology 35
6. Other Food Preservation Methods . 36
6.1. Chemical Preservatives 36
6.2. Modified Atmosphere Packaging . . 36
7. Nutrient Retention in Processed Foods 37
8. References 38

1. Introduction

1.1. Food Manufacture and Preservation

In the broadest sense, the field of food technology encompasses aspects of manufacturing foods from raw agricultural materials as well as the preservation of both agricultural and industrial food products. Many *primary* and *secondary food manufacturing processes* include some elements of preservation, as in milling of grains for flour, crystallization of sugar from sugar cane or sugar beet juices, smoking of sausages, baking of bread, fermentation of grape juice for wine, and many other examples. On the contrary, the *tertiary food processing operations* are concerned mainly with preservation of both raw agricultural materials and the foods manufactured from them. The various aspects of the main food manufacturing technologies are discussed more appropriately under the various food commodity headings elsewhere in

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_523

this encyclopedia; more complete descriptions are given in textbooks [1–3] and other sources. This article discusses the basic principles and some industrial aspects of the main food preservation technologies applicable to many agricultural commodities and food products.

Table 1. Growth requirements for microorganisms in foods

Environmental factor	Microorganism type	General growth requirements
Temperature	psychrotrophic	will grow at refrigerated temperatures
	mesophilic	20 – 36 °C
	thermophilic	40 – 55 °C
Oxygen	aerobic	O_2 required
	anaerobic	will not grow with O_2 present
Water *	bateria	$a_w \geq 0.9$
	yeasts	$a_w \geq 0.88$ (some will grow at $a_w = 0.7$)
	molds	$a_w \geq 0.8$ (some will grow at $a_w = 0.6$)
Acidity	all	approximate range of tolerance pH 4 – 8; some yeasts and molds very tolerant to acid
Nutrients	all	wide range of specific requirements for different organisms

* Expressed in terms of water activity (a_w).

1.2. Food Spoilage by Microorganisms

Raw agricultural materials and the foods manufactured from them are composed of organic matter that is susceptible to biological transformations by various forms of microbial life. The main food components (carbohydrates, proteins, lipids, minerals, and water) are nutrients that are required not only by humans and higher animals, but by bacteria, yeasts, and molds as well. Growth of these microorganisms in the various agricultural and food products is the primary cause of food spoilage. The mere presence of microbial life in the various foods is not deleterious; however, when these mostly unicellular organisms multiply in the optimal conditions needed for their growth, food spoilage occurs. Most food manufacturing and preservation techniques (→ Foods, 3. Food Additives) are designed to produce conditions that are unfavorable for microbial growth. The main growth requirements of the various microorganisms that are important to food preservation are summarized in Table 1. More detailed information can be found in the many textbooks of food microbiology [4–7].

The common principle of food preservation by dehydration, freezing, and sterilization by heat or by irradiation, is the complete elimination of microbial growth. This does not necessarily mean elimination of the microorganisms themselves; indeed, none of these three food preservation techniques removes the microorganisms from the foods. Rather, all three preservation approaches create such conditions in the foods themselves or the environment in which the food is held as to completely arrest any possible multiplication of the microbial cells that may be present. In the case of dehydration, this is achieved by elimination of one of the main nutrients required by all microorganisms, the water. Likewise, in the freezing preservation of food, water is made unavailable to the microbial life by crystallization as ice; in addition, the temperature used for storage of frozen foods is much too low to support growth of common forms of microbial life. Only in the food sterilization processes (canning, ultra-high-temperature processing, and irradiation) are the spoilage microorganisms killed; however, even in these processes the microbial cells are not removed from the foods, and in many instances, additional precautions are necessary, such as a storage environment that is not conducive to growth of any surviving forms of microbial life (the notion of commercial sterility; see Section 4.1).

The processing conditions needed to eliminate microbial spoilage in dried, frozen, or sterilized foods are often the cause of other types of food damage and deterioration, mainly because of physical effects and deleterious chemical reactions. Thus, the final choice of a proper food preservation technology is a compromise that gives adequate protection against microbial spoilage while minimizing the physical damage and the chances for undesirable chemical reactions.

1.3. Chemical Deterioration of Food

The main components of agricultural materials and foods made from them are carbohydrates, proteins, lipids, minerals, and water. Within or outside these broad chemical groups, there are many other chemical components including vi-

tamins, enzymes, acids, dissolved gases, and other compounds, many of which can be very reactive in the conditions used for food processing and storage. The chemical changes that occur in food during processing or storage include hydrolytic and oxidative reactions of fats, enzymatic breakdown of proteins to bitter peptides, and the various forms of food browning caused either by enzymatic reactions in the presence of oxygen as in fresh fruits or vegetables or by reactions of carbohydrates with protein or other carbohydrate molecules on high heating (the Maillard reaction). Some of the most important chemical reactions deleterious to food quality are enhanced by preservation processes such as removal of moisture, freezing, or concentration of the reactive components. More detailed information can be found in various food chemistry textbooks [8–10].

For the short-term preservation of fresh agricultural produce or manufactured foods, these reactions are much less important than microbial spoilage. However, in storage of food processed by the long-term preservation technologies, these chemical reactions often become the main factor determining the specific processing and storage conditions. In particular, fat oxidation, nonenzymatic (Maillard) browning, or enzyme-catalyzed flavor changes involving various food components are typical undesirable chemical reactions associated with dried, frozen, or sterilized foods. The rate of these and other deleterious reactions often depends on the presence and the state of water in the system.

1.4. Water Content and Water Activity

The most abundant component of raw agricultural products and many manufactured foods is water. (→Foods, 1. Survey, Chap. 3.1.). Although the total content of water in a given food may be useful as an indicator of nutritive value and a determinant of the necessary processing operations, it is the state in which this water is present that affects the shelf life and the susceptibility to various microbial and chemical spoilage reactions. The most common approach used to describe the state of water in a complex food material is through the water activity of the system.

The water activity (a_w) of a given food is defined as

$$a_w = \frac{p_f}{p_w}$$

where p_f and p_w are the pressures of water vapor in equilibrium with the given food system and with pure water, respectively, both at the same temperature. Because p_f can vary from 0 (for a completely dry system) to p_w (for an infinitely diluted aqueous solution), the range of a_w values varies from 0 to 1. Examples of typical a_w values for various foods are shown in Table 2; also listed are the requirements of the main food spoilage microorganisms for a_w conducive to their growth, while the general effect of a_w on the various food spoilage reactions is shown in Figure 1.

The main factor influencing p_f is the type and the amount of substances dissolved in the aqueous phase of a food, as approximated by Raoult's law:

$$p_f = p_w \cdot \left(\frac{n_w}{n_s + n_w}\right) = p_w \cdot X_w$$

where n_w and n_s are the number of moles of water and solute in the food, and X_w is the mole fraction of water. Although Raoult's law is exact only for ideal solutions, it gives a sufficiently close approximation of water activity of many common food systems. Considering the typical mass percentages found in foods, small molecules (some salts and acids) have a much greater effect at lowering a_w than do larger solutes, such as sugars, as shown by Raoult's law. Because of their very large molecular masses, food proteins have a negligible effect on a_w even in foods with high protein content. Although the water binding property of food proteins may be important for structural aspects of food systems, the water activity of typical protein-based foods, such as meat products or most cheeses, is not very different from that of pure water. Because fat is insoluble in water, it does not participate in the a_w food interactions. More specific information on a_w in foods is available in various books [11–13] and review articles [14, 15].

Individual food components may exert different effects on a_w before, during, and after a particular food preservation process; this, in turn, determines processing conditions (or even the process suitability) for each individual food. Some a_w relationships (Fig. 1) are particularly important in the case of food drying, where the

Table 2. Approximate values of moisture content and water activity (a_w) for selected foods and growth requirements of selected microorganisms (orientation data)

Food product	Average moisture content, wt %	Water activity (a_w)	Microbial activity
Water	100	1.00	
Fresh fruits, meats, vegetables	95 – 60	0.99	many pathogenic bacteria
Bread, aged cheese, salami	50 – 40	0.95	many spoilage bacteria
Ham, processed cheese, baked foods	40 – 20	0.90	*Lactobacillus, Bacillus*, and *Staphyloccus* species
Jams, jellies, parmesan cheese	30 – 20	0.80	sugar-tolerant yeast and bacteria
Salted meat, fish, condensed milk	30 – 20	0.70	osmophilic yeast and molds
Dried fruits, candy, honey, pasta products	20 – 10	0.60	a few osmophilic yeasts and molds
Dried mushrooms, dried milk, other dried foods	8 – 2	(0.3 – 0.2)	no growth

food product typically undergoes changes spanning the complete range of a_w, often in conditions of high heat input. Careful control of a_w and its manipulation is used in other traditional food manufacturing and preservation technologies, including preparation of fruit preserves, salting of fish and red meats, or in the new technology of intermediate-moisture foods (see Section 2.6).

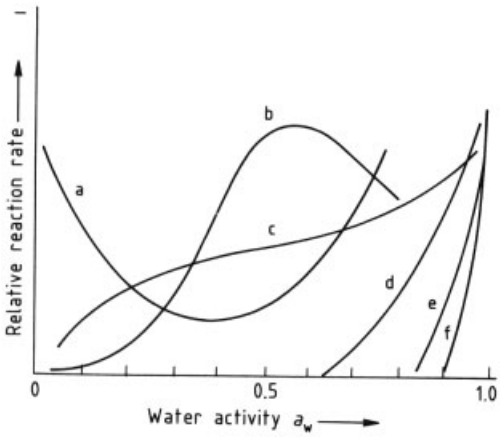

Figure 1. Generalized effects of water activity on relative reaction rates for various types of food spoilage
a) Lipid oxidation; b) Nonenzymatic browning; c) Enzymatic reactions; d) Mold growth; e) Growth of bacteria; f) Bacterial toxin production

2. Food Preservation by Dehydration

2.1. Concentration and Drying

Because water is one of the main nutrients required by all microorganisms for their growth, water removal (dehydration) offers theoretically one of the simplest opportunities for long-term food preservation. In industrial practice, the long shelf life of dry foods is important primarily for grains and some of their endproducts (flour, starch, pasta, and breakfast cereals), other agricultural commodities with a generally low moisture content (spices, herbs, dry pulses, and other seeds), and processed foods that can be dehydrated with little deleterious effect (e.g., skim milk or coffee extract). Many agricultural commodities and food products are available in the dehydrated form (dry fruits, vegetables, fish, eggs, cheese, completely prepared meals, baby formulas, and many specialty products), although in many instances the quality difference compared to fresh foods may be substantial.

Conceptually, food dehydration describes any process resulting in removal of water from food. In industrial practice, "food dehydration" is often used to describe "food drying," i.e., almost complete removal of water from either a liquid or a solid food to obtain a product with a very low a_w. In many industrial food drying processes (especially in those used for drying liquids such as milk, coffee, or fruit juices), the drying itself is preceded by a concentration process, whereby a major portion of the water is removed from the food solution more economically and

without changing the physical nature of the material, thus maintaining its fluidity. Concentration processes are special cases of food dehydration. They are important in industrial practice, but by themselves, with a few exceptions, such as sweetened condensed milk, do not result in products with a long shelf life.

2.2. Food Concentration Technology

The traditional processes for concentration of liquid foods are based on the principle of "boiling off" some of the constituent water under reduced pressure (the term evaporation is often synonymous to concentration in these cases). The evaporation is carried out under partial vacuum to minimize heat damage to the many heat-sensitive food components and to reduce the energy needs of the process (→Evaporation). The simplest production units, the *single-effect evaporators*, are relatively inexpensive to install, but the operating costs are high and productivity is low. The latent heat requirements for the evaporation are usually supplied by condensing steam.

For higher production economy, *multiple-effect evaporators* (Fig. 2) use the water vapors produced in one evaporation stage as the energy source in the next effect after appropriate elevation to a higher energetic level by "thermocompression" (enrichment with fresh steam) or by "mechanical recompression" with a steam compressor [16]. The installation costs of multiple-effect evaporators increase with the number of stages, but the effectiveness and operation economy improve substantially. Often the vapors generated in the food concentration process carry various volatile compounds characteristic of the given food flavor; in some cases, the flavor volatiles are stripped from the vapors and returned to the finished product.

Newer food concentration processes include freeze concentration [17] (see also Section 3.6) and reverse osmosis (RO). In *freeze concentration*, the liquid product is concentrated by freezing out some of the constituent water as ice and separating the ice from the concentrate. Technical advances have resulted in significant improvements, especially in the concentrate separation from the ice crystals, a problem that has hindered a more widespread use of this quality-efficient concentration technology [17], [18]. Because vacuum is not used to remove any volatile constituents, flavor retention in the freeze-concentrated foods is high. Industrial uses of freeze-concentration technology include plants for processing fruit juices, coffee, vinegar, and liquid beer concentrate for reconstitution "on the tap."

Reverse osmosis (RO) is an example of a membrane separation process based on selective molecular transport through a filtration membrane (→Membranes and Membrane Separations) [19]. The specially constructed membranes for RO are permeable to water, but essentially impermeable to any solutes that may be present; when pressure that exceeds the osmotic pressure of the liquid is applied, the solvent water permeates the membrane, thus leaving a more concentrated solution. The uses of RO in food concentration processes include specialized application for sensitive food materials (e.g., concentration of enzyme solutions) and as a preconcentration step to increase capacity of existing evaporators.

The related *fractional concentration* or "fractionation" process of ultrafiltration (UF) is based on the same principle, but uses a much "looser" membrane, permeable to other small molecules in addition to water. Typical applications of UF are in the recovery and concentration of protein fractions from protein-containing liquids such as milk, whey, blood serum, or various plant protein extracts. The industrial uses for UF have been increasing rapidly, principally in fluid milk processing, in cheese-making [20], for recovery of various valuable food proteins and in the production of fruit juices from whole pureed fruits [21].

In many cases where long shelf life or savings in transportation costs are important, the concentrates produced by the membrane processes must be dehydrated further in a suitable concentration or drying process, as the concentration efficiencies of RO and UF are limited to ca. 20 – 30 % total solids.

2.3. General Aspects of the Food Drying Processes

The preservation principle used in the drying of foods is the creation of a very low a_w system, which will not support microbial growth. This

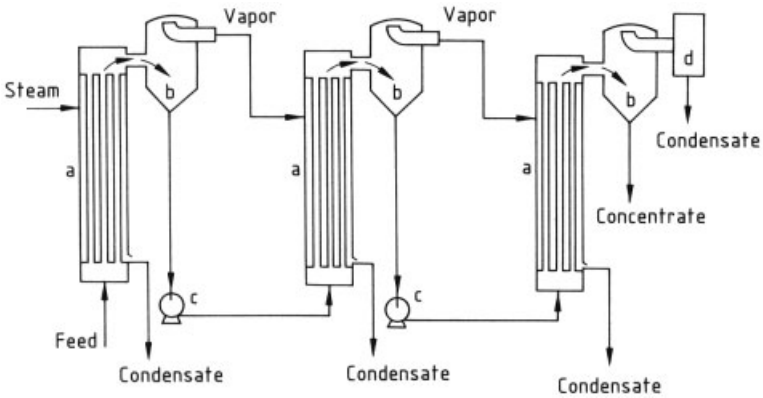

Figure 2. A principle of a multi-effect evaporator for concentrating liquid foods
a) Steam calandria; b) Vapor – liquid separator; c) Pump; d) Condenser

requires substantial removal of the product constituent water, often to a final moisture content $< 3-5\,\%$.

Food drying is a combined heat and mass transfer process (→Drying of Solid Materials). Heat energy must be supplied to the food to change the liquid water contained in the food to water vapor, which is subsequently removed. In most industrial food drying operations, heated air is used for both the heat and mass transfer functions; notable exceptions include roller (drum) drying and freeze drying, popular for several specialized food drying operations, as well as the seldom used vacuum drying.

Theoretically, most of the heat and mass transfer in food drying where hot air is used occurs on the surface of the food. As long as the moisture mass transfer rate from the food surface to the air is in equilibrium with the diffusion rate of moisture from the interior of the food to the surface, the drying process is said to be in the *constant rate period.* In this stage, the rate of moisture removal does not change with time and is governed by the external drying conditions (mainly the air temperature, humidity, and velocity). As a result, the temperature of the food surface is much lower than the temperature of the drying air because of the evaporative cooling effect of moisture removal. In the second drying stage, the *falling rate period*, the migration of the water from the interior to the surface is slower than the mass transfer of the water vapor to the air, and the temperature of the food rises. Table 3 lists the characteristic aspects of the constant and falling rate periods important to food drying operations. A more detailed description of food drying theory is covered in specialized texts [22], [23].

Table 3. Characteristics of the constant and falling rate periods in drying

Constant rate period	Falling rate period
Surface completely wetted	surface partly dry
No effect of material on drying rate, no effect of water diffusion through food	drying rate controlled by diffusion of water in material
Rate of diffusion is greater than rate of evaporation	rate of diffusion is lower than rate of evaporation; water may evaporate within the food
Heat transfer only between heat source and food surface	heat transfer through food may be important
Temperature of surface is approximately wet bulb	temperature of surface approaches temperature of drying air

Proper control of the air drying process, particularly in the falling rate period, is critical to the quality of the dried food product. The following quality defects [22] are common to many dried foods; in most cases, these need to be minimized by properly conducted drying processes.

Tissue Shrinkage. As water is removed in the course of drying, many foods shrink because of the elastic nature of the tissue, the surface tension exerted by the water, and the pressure difference between the outside and the voids left by the removed water. Shrinkage can be particularly severe in the case of slow drying.

Browning and Heat Damage. Many foods contain both proteins and sugars, the main reactants in the Maillard browning reaction. When heat transfer is not controlled properly, much heat damage can occur particularly in the falling rate period when a_w of the drying food may be in the intermediate zone most conducive to rapid browning (Fig. 1). Vitamin destruction, protein denaturation, and other typical heat-dependent reactions may also occur in improperly controlled drying processes.

Case Hardening. This term denotes a defect caused by a hard crust formed on the food surface, which interferes with further dehydration and which can be severely damaged by heat. As a result, the food appears to be dry while still containing substantial amounts of moisture. The defect, which may be characteristic for drying of certain foods, is often aggravated by migration of soluble constituents to the food surface together with the solvent water.

Loss of Solubility or Rehydratability. As a result of these effects (i.e., shrinkage, heat damage, and case hardening) the dry food may be difficult to reconstitute to its original state. In the case of liquid foods (milk and fruit juices), the drying effects appear as loss of solubility and formation of a sediment on reconstitution; for solid foods, reabsorption of water can be exceedingly difficult and may require up to several hours. Many air-dried foods (mushrooms, fruits, etc.) never rehydrate to their original moisture content.

Loss of Flavor Volatiles. Because drying is based on volatilization of water, other components may also be removed in the drying process. Many food flavor components are volatile, and loss of flavor is indeed one of the most common defects of all dried foods. Although many alternative designs of food drying processes have been developed to minimize other quality defects, loss of flavor is a difficult problem, often hindering wider acceptance of many industrially dried foods.

Many alternative engineering designs of basic air drying technology have been developed for foods to minimize some of these defects. In addition, food drying processes based on heat or mass transfer mechanisms other than the use of hot air have been successful in several specialized situations.

2.4. Food Drying Technology

Some of the important dry food products worldwide are milk, coffee, some vegetables (e.g. mushrooms, potatoes, and onions), pasta products, eggs, fish, meat, herbs, and spices, as well many dry food preparations for direct use by consumers (e.g., soups, gravies, condiments, complete meals for mountaineers, army rations, and baby foods) or for industrial and institutional uses (dry ingredients for the meat industry, dry cheese, and nutritional formulas for hospitals).

Different drying processes are available for drying solid or liquid foods (→ Drying of Solid Materials). The choice of a particular drying system often depends on the characteristics of the material being dried, its heat sensitivity, flavor retention properties, physical structure, designated end use, and, often most importantly, the cost of the process. Some techniques, such as freeze drying, are suitable for almost any product, but its high expense limits more widespread use; in other cases, custom-built equipment may be required for specialized applications.

2.4.1. Convection Drying

The majority of food drying technologies are based on the use of hot air for the convective heat transfer as well as for removal of the evaporated water. A great variety of batch or continuous air driers exists; most of the air drier designs vary mainly in the configuration of the drying chamber, i.e., the position or relative movement of the food and the air stream. Basic drier designs for solid food items include cabinet, kiln, tunnel, belt, tray, and other types of driers. According to the air movement, the design may by cocurrent, countercurrent, or crosscurrent. Figure 3 shows schematically several basic common air drier designs. More detailed descriptions can be found in specialized texts [22], [23].

Fluidized-Bed Drying. This is a special case of hot air drying suitable for drying of particulate foods (e.g., peas, corn, carrot cubes, and potato granules). The relatively high velocity of the cross-flow stream of air through the particulate bed provides controlled "air lift" to the

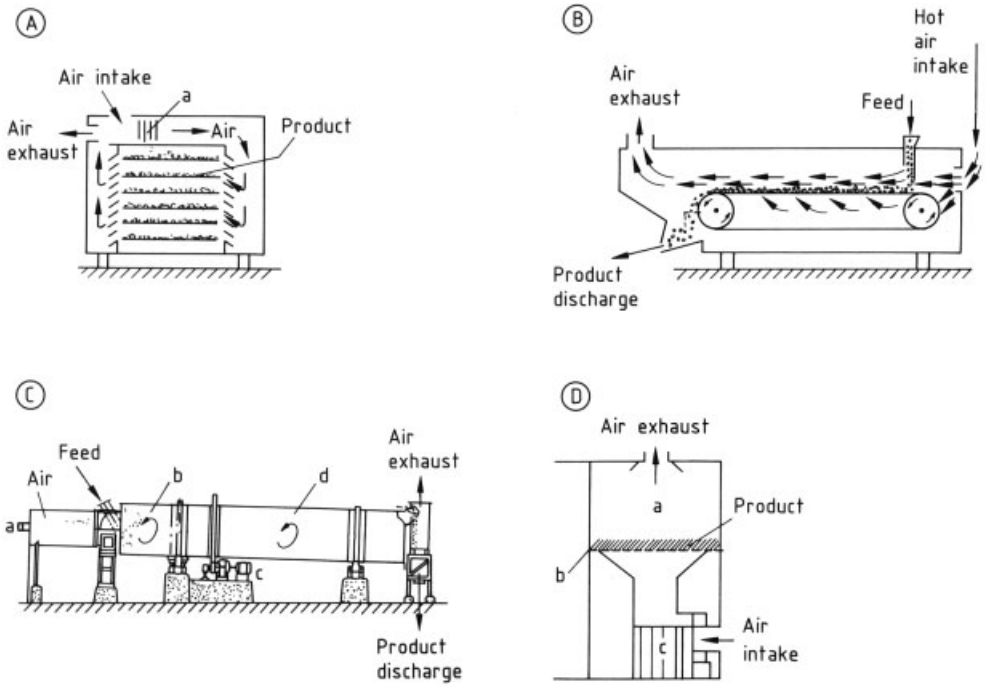

Figure 3. Principles of various air-drying processes
A) Tray: a) Air heaterB) BeltC) Tunnel: a) Burner; b) Higher end; c) Drive; d) Lower endD) Kiln: a) Drying chamber; b) Perforated bottom; c) Air heater

individual food particles, resulting in "fluidization" of the bed. Thus, more efficient drying is achieved because of increased total product surface area available to air contact. Both batch and continuous fluidized-bed driers are used (Fig. 4).

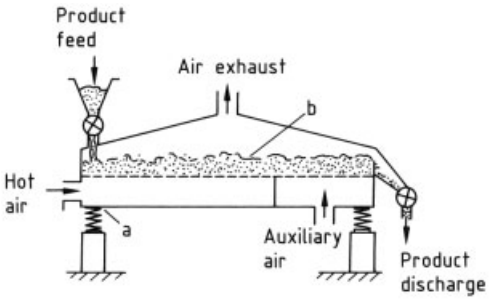

Figure 4. Principle of a fluidized-bed drier
a) Vibrating mechanism; b) Fluidized product

Spray-Drying. Although limited mainly to liquid foods, spray-drying is probably the most widely used food-drying technique. The principle of the continuous spray drying process is the dispersion (atomization) of the liquid being dried into a stream of hot air entering the drying chamber. The diameters of the individual droplets in the liquid spray are in the micrometer range and the inlet air temperature may be in excess of 200 °C; this results in almost instantaneous evaporation of the constituent water from the food particles into the air. The stream of the drying air exiting from the drying chamber carries the water vapor, absorbed from the food, as well as the dry food particles themselves. The dry food powder is subsequently removed from the air stream in a suitable powder-separation step.

Any food material that can be converted to a liquid or a slurry can be spray-dried; the resulting product is primarily a powder. The main difference between spray-drying, which is a special case of hot air drying, and other food drying processes is the relatively short drying time (several seconds) that does not allow for regulation of the drying process by time control. Because of the very rapid drying of the small droplets dispersed in the drying atmosphere of hot air, some of the

deleterious changes common to many other drying processes can be minimized.

The *basic spray-drier design* (Fig. 5) consists of three major parts: the drying chamber (a), the fluid dispersal (atomization) mechanism (b), and the dry powder collection assembly (c). The choice of atomization method influences the basic construction of the drying chamber (drying tower in the case of vertical driers). The most widely used atomization systems are the pressure nozzle and the centrifugal atomizing wheel. Whereas centrifugal atomization is much more complex and more expensive than nozzle atomization, the atomizing wheel provides greater versatility as it can handle viscous and particulate liquid systems.

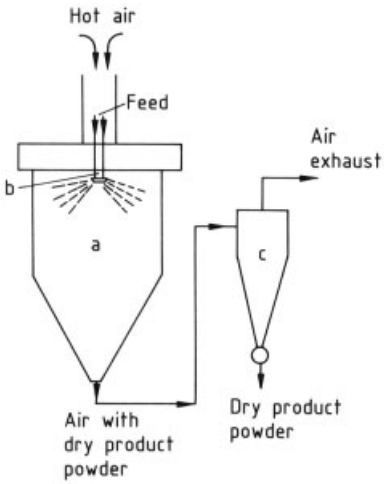

Figure 5. Principle of a spray-drier
a) Drying tower; b) Spray (atomization) mechanism; c) Dry powder collection (air cyclone)

Because of the fine nature of the dry particles, the powder collection assembly must be designed effectively to minimize entrainment of the dry material in the escaping drying air. The powder collectors used most often include the dry cyclone, the "baghouse," and the agglomerating trough (see Section 2.5). The "liquid collector" assembly, sometimes installed to recover the finest particles from the exhaust air, can be a source of microbial contamination and is thus unsuitable for modern food-drying operations.

Developments in spray-drier design include multiphase drying, where the primary spray-drying phase is followed by secondary belt drying of the collected powder; incorporation of heat exchangers for heat recuperation from the exit air; and a more compact design combining the drying and collection chambers. Additional information is available in sources [24], [25].

2.4.2. Contact, Vacuum, and Freeze Drying

In several specialized food-drying processes, the necessary heat transfer to the product being dried is accomplished by means other than hot air. For atmospheric contact drying (represented by the drum-drying process popular especially in the dairy industry) the heat is supplied through a hot surface in contact with the food product, while the evaporated water is dissipated to the surrounding atmosphere. To decrease the temperature for the phase change of liquid water to vapor, drying of heat sensitive foods may sometimes be carried out under vacuum. In the absence of air as both the heat transfer and mass transfer medium, heating is accomplished by contact with heated shelves or by radiation, whereas the water vapor removal requires vacuum pumps, refrigerated condensers, or other complicated designs.

Drum-Drying. For drying liquid or slurried products where heat damage is not detrimental, drum (roller) drying offers an economical alternative to spray-drying (for an example of a drum drier, see → Drying of Solid Materials). Water removal occurs typically at atmospheric pressure (vacuum drum driers have not been successful) by spreading the liquid being dried onto the surface of slowly rotating stainless steel rolls filled with saturated steam. After a full revolution of the drum, the sheet of dry material is scraped off the surface just before it reaches the area where the liquid is applied. Drum drying can be used in some specialized applications, such as for milk drying for confectionery use or for lower value inexpensive products (feed mixes, waste products from potato processing, etc.)

Freeze-Drying. The principle of freeze-drying (lyophilization) is the removal of water from frozen food by sublimation. Because the food is held in a solid state during the whole drying process, shrinkage is avoided; thus, the dry products are easily rehydratable. However, a similar fast rehydration effect has been achieved

with mushrooms using simple high vacuum drying [26], thus indicating that the need for prior freezing followed by sublimation drying may not be necessary in all cases, especially for delicate products where freezing results in substantial textural damage (see Section 3.4).

The need for controlled heat application in the freeze-drying process (to avoid melting the ice during drying) results in little heat damage, but requires long drying times under high vacuum, the main cost factor in this very expensive process.

Whereas retention of flavor in freeze-dried foods has been cited as a significant advantage over other food dehydration technologies, variable flavor losses may still occur, primarily because of the high vacuum needed for the sublimation. Because of the absence of air, heat transfer must be accomplished either by conduction from the heated shelves through the poorly conducting dry layers of the food to the frozen core, or by radiation, which may cause heat damage to the dry outer layers (Fig. 6). As a result, processing costs are high, and the freeze-drying process is used only in specialized applications involving foods of high value (e.g., coffee or functional food protein preparations) or for unique requirements such as lightweight rations for mountaineers, backpackers, or explorers.

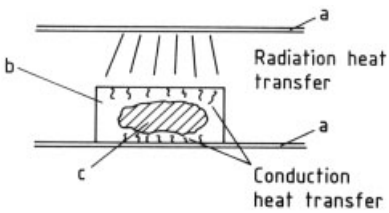

Figure 6. Illustration of the heat transfer processes in freeze-drying
a) Heated shelves; b) Dry food layer; c) Frozen food core

2.4.3. Radiant-Heat Drying

Use of radiant-heat transfer for the drying of foods is not common. The only two examples of practical significance are the radiant-heat supply in some cases of freeze-drying and the traditional drying of food materials by exposing them to the effects of the hot sun.

Sun Drying. Leaving food in the sun to dry was the original food preservation technique used by ancient Egyptians, as well as by North American Indians. Despite significant technological developments, sun drying of food is still an important technique used in many countries and in specialized situations (production of raisins, dried fish, dates, household drying of wild mushrooms, and other items). A new meaning of the term "sun drying" may be emerging with the development of driers for delicate products requiring low heat input (e.g., hops and mushrooms); in these driers, solar heating is used to heat the drying air, which is then conveyed to a modern drying chamber. The main problems of traditional, open air sun drying are the fluctuating supply of the heat, which is dependent on weather conditions and the time of the year, as well as proper hygienic control of the drying process.

2.5. Instantization

Spray-dried powders produced from liquid foods (e.g., milk, coffee, and chocolate) often show poor water-reconstitution properties, resulting in the formation of clumps. This defect is usually related to the very small particle size of the powders produced by the spray nozzle or atomizing wheel mechanisms. The poor ability of otherwise soluble powders to be reconstituted can be improved by using an instantization process, whereby the surface of the individual small particles is partially wetted in an atmosphere of moist steam in agglomerating chambers or "troughs." The particles are allowed to form loose aggregates, which are subsequently redried, thus creating large porous agglomerates that can be readily stirred into water during reconstitution.

2.6. Moisture Sorption Isotherms of Foods

The main requirement for successful food preservation by dehydration is sufficient lowering of a_w to prevent microbial growth. However, the final moisture content to which a food is dried is an important quality determinant of the dried product. Drying to a minimum moisture content

may not be possible and is not always desirable. A low moisture content in some foods may lead to oxidation of fat, whereas in other foods the same moisture content gives the longest shelf life. The optimum moisture content of dried food can be determined from experimentally established relationships between the moisture content and the corresponding water activity, known as moisture sorption isotherms (Fig. 7). Each food, depending on its particular composition, shows a different sorption isotherm [27]. Comparison of sorption isotherms for the same fresh foods and foods previously dried (desorption vs. adsorption) often shows a phenomenon of sorption hysteresis (area a, Fig. 9). The difference in the amount of moisture giving the same a_w for fresh versus previously dried foods is related to the food structure and its damage in the drying process, the affinity of water for the various food components, and other effects. For the longest shelf life, different foods require drying to a different optimum final moisture content, which can be determined by the classic Brunauer, Emmett, and Teller (BET) technique [28] or by similar more accurate modern procedures [29], [30]. In general, the optimum final moisture content of dried foods is in the vicinity of a theoretical value corresponding to a monomolecular layer of water molecules bound to all reactive sites of the food. Typical moisture contents and monomolecular layer values for several dried foods are as follows (in wt %):

Potato flakes	6.51
Freeze-dried salmon	6.84
Onion powder	3.58
Spray-dried cheese	2.22
Freeze-dried whey protein	8.03
Dried beans	4.50

Although the shelf life of dried foods is generally measured in months or years, the main quality determinants of properly dried and packaged foods are the deteriorative chemical reactions caused by removing water from the system.

2.7. Intermediate-Moisture Foods

With a few exceptions, dried foods are unsuitable for direct consumption without reconstitution with water. The low moisture content of dried foods makes them hard, brittle, and difficult to chew. A new technology of intermediate-moisture foods (IMF) was developed during the manned moon-exploration activities (1960 – 1970) to overcome the need for water to reconstitute dried foods while maintaining the sufficiently low a_w needed for preservation without refrigeration. Because microbial growth generally does not occur below $a_w = 0.6 - 0.55$, foods formulated to this a_w range may be stored without refrigeration or other special preservation techniques. The moisture content of these foods is usually high enough to ensure acceptable sensory properties, including chewability, not encountered in most dry foods.

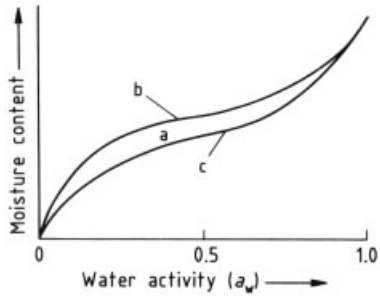

Figure 7. Moisture sorption isotherms for foods
a) Sorption hysteresis loop; b) Desorption (fresh food); c) Adsorption (dry food)

The concept of IMF is not new; jams, jellies, fruit cakes, some dry sausages, or even pemmican, the traditional food of American Indians, are examples of ready-to-consume foods that need not be refrigerated, yet do not spoil because of their low a_w. Modern IMF technology is based on the use of various food-grade, water-soluble compounds that lower a_w when added to a food. Some a_w-lowering compounds, generally known as *humectants*, have a long history of traditional use for food preservation (especially salts and sugars), whereas others (e.g., glycerol and sorbitol) are common food ingredients now being used for purposes other than a_w control. Alternative technological processes used include partial drying, infusion of dried foods with a liquid mixture of humectants, or careful formulation of ingredients according to the Ross equation [31]:

$$a_w\,(\mathrm{f}) = a_w\,(1) \cdot a_w\,(2) \cdot a_w\,(3) \ldots a_w\,(n)$$

where a_w(f) is the final a_w desired and $a_w(i)$ are water activities of binary solutions of the

same concentrations. Although minor discrepancies often exist between experimental values and those predicted by the Ross equation, for many routine industrial applications the use of this empirical equation is satisfactory.

Uses of IMF technology are limited to a few special applications including pet foods and granola-type breakfast bars. The main limitation is the need for relatively large amounts of humectants to achieve the required a_w lowering, which results in undesirable sensory impact.

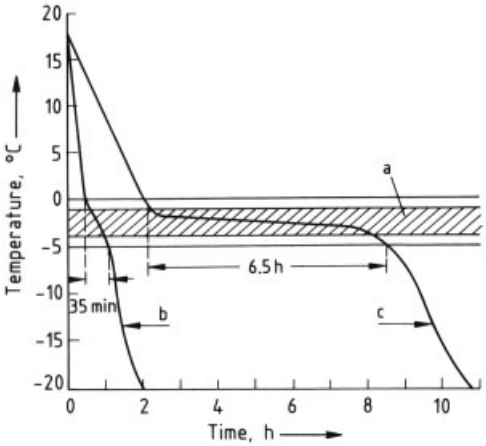

Figure 8. Time – temperature profile of a freezing process
a) Zone of maximal ice crystal growth; b) Fast freezing; c) Slow freezing

3. Food Preservation at Low Temperature

3.1. Refrigeration and Freezing

Food preservation by cooling is the most common technique used for food storage in all parts of the world. In industrialized countries, widespread use of mechanical refrigeration gave cold temperature preservation its prominence as the principal short-term food preservation technique for fresh foods in households as well as in industrial processing and retailing.

As long as foods are held at storage temperatures above their freezing point, no phase change in the constituent water occurs, and some of the cold-tolerant microorganisms (psychrotrophs) can grow. Thus, cold storage without freezing retards but does not stop the food spoilage process. For a long-term effect, food preservation by low temperature must include freezing and frozen storage. In frozen food, most of the water is present as solid ice and this makes it unavailable to support microbial growth. With the very low temperature of frozen storage, which further inhibits any biological activities, food freezing offers a much longer shelf life than can be achieved with cold storage above the freezing point. However, freezing does not destroy microbial life; after thawing, frozen foods are just as perishable as fresh foods. Some psychrotrophic bacteria can grow at temperatures as low as −5 to −10 °C; thus, since not all water is frozen in foods stored close to their freezing point, slow microbial growth can occur in improperly stored frozen foods.

3.2. General Aspects of the Food Freezing Process

Whereas pure water freezes at 0 °C, the freezing points of foods depend on the amounts of solutes dissolved in their constituent water (see Raoult's law, Section 1.4). Typical *apparent freezing points* of various foods are as follows:

Vegetables (e.g., peas, cauliflower, onions, carrots, and tomatoes)	−0.9 to −1.4 °C
Fruits (e.g., pears, plums, and apricots)	−1.8 to −2.5 °C
Berry fruits (e.g., strawberries and raspberries)	−0.8 to −1.2 °C
Milk	−0.5 °C
Beer	−2.2 °C
Meat, poultry, and fish	−0.6 to −2.0 °C
Ice cream	−4.0 to −6.0 °C

An approximate freezing point (t, in °C) of a solution of given molality M can be calculated as

$$t = -1.86 \cdot M \cdot n$$

where n is the number of ions or particles formed by dissociation of the solute molecules.

As a result of the freezing process, pure *ice crystals* are formed in food placed in a cold environment. The size, amount, and location of the ice crystals depends on the rate of the heat transfer from the food to the environment; the rate, in turn, is determined by the temperature difference between the apparent freezing point of the food and the environment temperature and

also the heat transfer coefficient characterizing the freezing process used. Faster freezing rates result in larger amounts of smaller crystals and a more uniform distribution of them. In biological materials where preservation of viability is important (frozen spermatozoa or embryos, red blood cells, organs for transplant operations, or microbial cultures for dairy and other food industries), the position of the formed ice crystals is of great significance. Appearance of ice crystals within the individual cells, resulting from very fast freezing rates, usually causes loss of viability; however, for optimum frozen food quality, the fastest possible freezing rate is normally preferred.

The *time – temperature profile* of a food frozen at two different rates is shown in Figure 8. At a slow freezing rate, the food temperature remains constant near its freezing point as all the heat removed is the latent heat of solidification. Ice crystal formation is slow, and a smaller number of larger crystals are formed, in contrast to a fast freezing rate, which favors crystal formation rather than crystal growth. In the freezing of fruits, vegetables, and other foods with recognizable cellular structure, slow freezing favors extra-cellular crystal formation. Migration of water to the growing crystals from the interior of the cells produces highly concentrated unfrozen solutions of salts, organic acids, and other component solutes within the cell. This, together with the volume increase after freezing (caused by the 9 % volume difference between liquid water and ice), may lead to severe textural damage in those products where the cell walls and membranes provide the structural network. On the contrary, rapid freezing may result in intracellular freezing, which affects the viability of the biological cell but does not necessarily affect individual cell integrity. However, volume changes may be important in products with a high moisture content because they produce cracks throughout the tissue. A generalized view of the two opposing effects on final viability of a biological cell is shown in Figure 9. Although this theoretical representation has been confirmed experimentally for the viability of various tissues [32], its importance for frozen food quality remains to be established. In general, the study of the food freezing process and of its effects on food quality is difficult because effects of freezing, frozen storage, and thawing cannot be easily separated. However, much theoretical and practical information is available on the various technological aspects of the frozen food industry [33], [34].

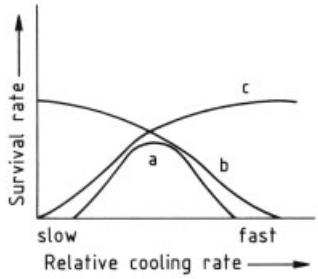

Figure 9. Survival of cellular biological systems as affected by different freezing rates
a) Net survival rate; b) Intracellular freezing; c) Solute effects

3.3. Food-Freezing Technology

Although many different engineering process designs have been developed for freezing food, they are all based on the same principle of placing the food in contact with a cold environment. The cold is generated either by a mechanical refrigeration system or by the use of liquefied gases with low boiling points (usually below −100 °C) (→ Cryogenic Technology, → Refrigeration Technology). Mechanical refrigeration systems are most often based on repeated use of ammonia either with a compressor or in an absorption system. In both cases the systems are operated in a closed cycle as ammonia is toxic, and thus an indirect means of heat transfer from the food to the ammonia must be accomplished: either through cold air, a cold liquid such as salt brine or alcohol, or a cold solid surface. On the contrary, cryogenic freezing systems using N_2 or CO_2 are operated in an open cycle, and the heat transfer from the food to the liquefied cryogenic gas is direct.

Table 4. Classification of freezing processes

Process	Frozen front movement, cm/h	Temperature change within food, measured in
Slow	0.1 – 0.2	°C/h
Intermediate	0.2 – 0.5	
Fast	0.5 – 3	°C/min
Very fast	3 – 10	°C/s

One of the principal differences distinguishing the various food freezing systems is the freezing rate achievable. On the recommendation of various bodies, such as the International Institute of Refrigeration (IIR), the freezing processes can be classified as slow, intermediate, fast, and very fast according to the velocity of movement of the frozen front from the surface to the thermal center of a product (Table 4). Alternatively, freezing velocity can also be defined as the change of temperature in the food per unit time.

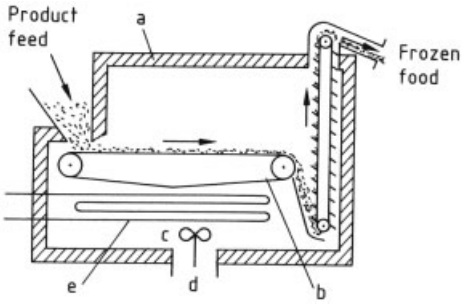

Figure 10. Principle of air-blast freezing tunnel
a) Insulation; b) Conveyor; c) Air fan; d) Air intake; e) Air cooling

3.3.1. Convective Processes

Still Air and Blast Freezing. Cold air at −10 to −40 °C is the most commonly used medium for freezing foods. The simplest freezing method (still air, also sharp freezing) consists of placing the food material in an environment of stagnant cold air, as practiced in typical household freezers. This is the least efficient freezing method, resulting in long freezing times, very slow freezing rates, and the most damaging effects on food structure due to formation and growth of large ice crystals. Still air technology is used industrially for storage of frozen foods; however, freezing is usually accomplished by a different means.

Movement of the cold air around the foods being frozen increases freezing rate. Air blast tunnels (Fig. 10) are commonly used for freezing small size food products, such as ice cream, fish, french fries, fruits, seafood, etc; typically, the product is carried through the tunnel on a conveyor (b) in a countercurrent or crosscurrent arrangement relative to the high velocity (3 – 10 m/s) stream of cold (−10 to −30 °C) air. Freezing in air blast tunnels is efficient and economical; however, for improperly packaged foods blast freezing may cause excessive dehydration, resulting in fresh weight losses and need for frequent defrosting of the freezing equipment.

Fluidized-Bed Freezing. In a manner similar to fluidized-bed drying (see Section 2.4.1), small particulate foods can be frozen by a stream of cold air passing through a stationary or moving bed of the material being frozen. The air-lift effect of the cold air results in improved contact of the air with the food surface, thus increasing heat transfer and correspondingly the freezing rate. Because of the rapid freezing rates, moisture losses can be lower than in comparable blast freezing operations. Fluidized-bed freezing is particularly suitable for products of uniform size and high commercial value (e.g., shrimps and blueberries).

Immersion Freezing. For limited industrial applications, nontoxic liquids with a low freezing point are being used for freezing packaged or unpackaged foods by direct immersion. This method is sometimes used for freezing fowl, tightly packaged in thin plastic film; the rapid surface freezing of the submerged poultry results in the attractive, uniform white appearance of the finished package. The heat-removal bath may be a cold (−10 to −20 °C) aqueous medium containing a water-miscible material (ethanol, propylene glycol, or glycerol), resulting in substantial freezing point lowering, or the bath may be an aqueous salt solution (sodium or calcium chloride), a nontoxic freezing agent (freon), or a liquefied gas such as nitrogen or carbon dioxide. However, use of a liquefied gas for immersion freezing is not practiced industrially; rather, these gases are used as sprays in cryogenic freezing tunnels.

3.3.2. Cryogenic Freezing

See also → Cryogenic Technology. The use of liquid nitrogen or liquid carbon dioxide is advantageous for attaining very fast freezing rates, for infrequently used small installations, or for mass production of specialized products in dedicated

freezing lines. The principal difference distinguishing cryogenic freezing from other food freezing methods is the direct contact of the food with the primary freeze-producing agent (N_2 or CO_2), which undergoes a change of state. Because of the low boiling points (−196 °C for N_2, −78.5 °C for CO_2 at 100 kPa), both the latent heat of vaporization and the sensible heat removed by the cold gas contribute significantly to the fast heat transfer from the food being frozen. Typically the basic design of a cryogenic freezer installation consists of a conveyor for the food to be carried under the cryogenic liquid spray in a simple insulated tunnel (Fig. 11); the cold gas escaping through the tunnel both concurrently and countercurrently with the conveyor movement provides either the precooling of the fresh material or the postfreezing heat removal from the frozen product.

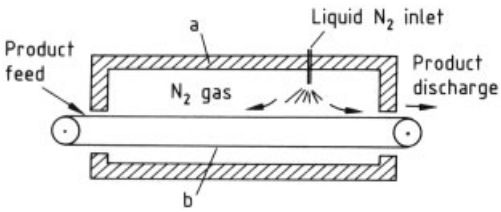

Figure 11. Principle of a cryogenic freezing tunnel
a) Insulated tunnel; b) Conveyor belt

The main *advantages* of cryogenic freezing systems are the very fast freezing rates, resulting in superior quality of most frozen food products, the very low weight losses because of negligible evaporation of water, and the low cost of the freezing equipment installation. The main *disadvantage* is the high overall processing cost, resulting from the high cost of the nonrecoverable freezing agent; in addition, the special safety precautions needed for installation and handling of hazardous compressed gas materials and equipment on otherwise nonhazardous industrial premises may discourage many potential applications.

3.3.3. Contact Processes

Plate and Scraped Surface Freezers. The principle of these processes is heat transfer from the food through a cold stainless steel surface. Rectangular packages (e.g., ice cream cartons, boxed vegetable products, and blocks of fish) can be frozen by contact freezing between two plates in plate freezers (Fig. 12). Either cold salt brine or other liquid freeze-producing agents (ammonia or freons) may be used for cooling the plates of the freezer. These often fully automated multiplate machines are popular because they require little floor space, have high loading capacity, and are effective for freezing prepackaged goods. However, the typical paper packaging used for the rectangular containers being frozen constitutes a hindrance to the required heat transfer so that the achieved freezing rate is not particularly fast, especially for foods containing large voids.

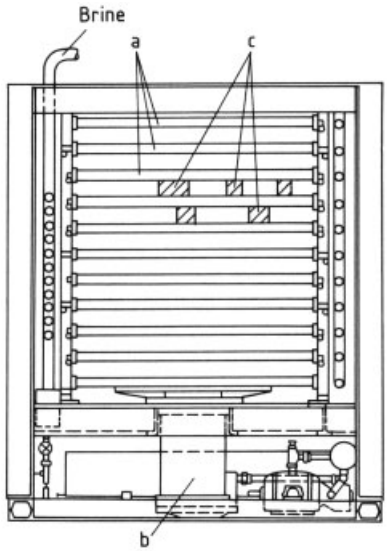

Figure 12. Principle of a contact plate freezer
a) Freezing plates; b) Plate compression mechanism; c) Food product

Much more effective freezing can be accomplished in contact freezing of unpackaged goods, such as fruit juices, freezer jams, or ice cream in scraped-surface continuous freezers (Fig. 13), where freezing occurs on the inside surface of a barrel cooled by a proper liquid coolant (ammonia, freon, or brine). The continuous scraping of the surface by rotating blades usually results in the incorporation of some air; this is often desirable as the semifrozen paste must possess enough flowability for effective product movement through the machine.

In the manufacture of ice cream products, the scraped-surface ice cream freezer is used for pro-

duction of the semifrozen (soft serve) ice cream mass. This may be further solidified by filling into stainles steel forms submerged in cold liquid (on-stick novelties) or into other suitable containers (boxes, cones, biscuit sandwiches, and cups), which are then "hardened" in blast or other type freezers.

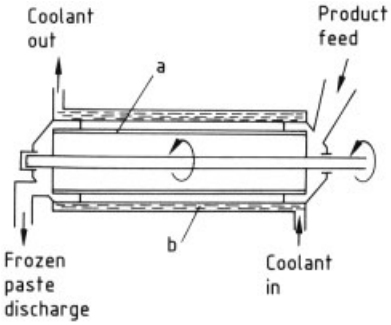

Figure 13. Principle of a scraped-surface food freezer
a) Surface scraping blades; b) Outer jacket filled with coolant

3.3.4. Freezing System Capacities

The theoretical freezing capacity of the various technological equipment designs is related to the primary capacity of the cold generating unit used for removal of the latent heat of fusion from the food. The capacity of the mechanical compressors used in most freezing systems is usually expressed in practical units referred to as a "ton of refrigeration"; this is defined as the rate of heat transfer provided by melting 1 short ton (2000 lb) of ice in 24 h; thus, in SI units, a ton of refrigeration represents a rate of heat removal equivalent to 210 kJ/min. The actual rate of a food freezing process is dependent on the total heat-removal requirement (determined by the mass and the composition of the food), and on the heat-transfer properties in the system, affecting the achievable freezing time.

Heat-Removal Requirements. The total amount of heat to be removed by the freezing equipment consists of three separate segments: (1) the sensible heat change needed to cool the food to its freezing point, (2) the latent heat of fusion evolved in solidifying the liquid water in the food to ice, and (3) further sensible heat change to cool the frozen piece to its final exit temperature. Estimates of the heat removal requirements Q (in kJ) can be obtained by using the basic calorimetric energy balance approach and by assuming that all water in the food freezes at a single freezing point:

$$Q = mc_p \Delta t + mL + mc'_p \Delta t'$$

where m = mass of food, kg; c_p and c'_p = specific heats of the unfrozen and frozen material, kJ kg^{-1} °C^{-1}; Δt and $\Delta t'$ = the temperature difference for the unfrozen and frozen material (from the initial temperature to the apparent freezing point and from the freezing point to the final exit temperature, respectively), °C; and L = latent heat of fusion of the food (kJ/kg). Whereas the proper values of c_p, c'_p and L for some foods can be found in the literature [35], [36], they are determined to a large degree by the water content of the product being frozen. Satisfactory estimates of the c_p and c'_p values can be made if the weight fraction of water in the product (w) is known, using the following relationships [36]:

$$c_p = 4.2w + 0.84\,(1-w)$$
$$c'_p = 2.1w + 0.84\,(1-w)$$

(alternatively for a product with a high moisture content, $c'_p = 1/2\, c_p$ may be used, as for water) and

$$L = 335\text{W}$$

Because in reality the constituent water in a food freezes over a range of temperatures, a more accurate approach to determining heat-removal requirements is based on *total enthalpy changes.* In this concept, −40 °C is used as a reference temperature, and the enthalpy (total energy content) of all foods at −40 °C is arbitrarily taken as 0 kJ/kg; the total amount of heat to be removed Q (in kJ) is then simply obtained as

$$Q = \Delta H = H_1 - H_2$$

The appropriate enthalpy values H_1 and H_2 (corresponding to the initial and final temperatures of the food at the beginning and the end of the freezing process) need to be known for each specific food. Some numerical data are available in the literature [35]; enthalpy – H_2O concentration diagrams for a few specific food materials can also be found [35].

An *example* of using such a diagram is shown for lean beef in Figure 14. The enthalpy difference ΔH can simply be read on the y-axis for a given moisture content (74 wt % is shown in this example) between any two temperatures of interest (e.g., for the freezing process beginning at +5 °C and ending at −25 °C, the total heat to be removed is ca. 275 kJ per kilogram of meat). In addition, information about the proportion of the total water existing as ice at each particular temperature can be read off the diagram as the intersect of the vertical composition line with the α-lines giving the wt % of water frozen. In this example, the beef with 74 wt % moisture starts to freeze at ca. −1 °C; at −2 °C, ca. 50 % of the constituent water (i.e., $0.5 \times 0.74 = 0.37$ kg of water per kilogram of meat) is in the form of ice, whereas at ca. −25 °C practically all the freezable water of the meat is solidified. The remaining 12 wt % water can be considered as unfreezable water, because it is tightly bound to the protein of the meat.

Information about the percent of frozen water contained in these diagrams can be useful also for estimating temperature fluctuation effects in frozen storage. Some foods can suffer severe quality changes due to recrystallization of the water melted by a slight temperature increase followed by a decrease in improperly controlled storage equipment.

Determination of the total heat-removal requirement is important for selecting appropriate freezing equipment. However, the food freezing capacity of a system also depends on the type of food, its composition and physical properties, the shape in which it is packaged, and the properties of the packaging material, and the heat transfer properties of the system used, resulting in substantial potential variations of freezing capacities of the same system for different foods. The main determinant of the freezing capacity is the time needed to accomplish the required freezing process, which must be found experimentally or calculated by using various approximations.

Prediction of Freezing Times for Foods. Because of the nonsteady state of the freezing process and the compositional changes occurring during freezing, which result in constantly changing theoretical freezing points, an exact mathematical model for calculation of freezing time is not available. Various predictive approaches for determining the industrially important freezing time data have been developed. The most commonly used formula is *Planck's equation* for prediction of freezing time (t_f) for rectangular and other simple geometrical objects:

$$t_f = \frac{\varrho \cdot \lambda}{\theta_f - \theta_e} \left(\frac{P \cdot a}{h} + \frac{R \cdot a^2}{k} \right)$$

where θ_f = apparent freezing point of the food, °C; θ_e = temperature of the freezing medium, °C; ϱ = product density, kg/m^3); λ = latent heat of fusion, J/kg; a = characteristic dimension (thickness of slab, diameter of sphere or cylinder), m; k = thermal conductivity of frozen material, W m^{-1} °C^{-1}; h = surface heat transfer coefficient, W m^{-2} °C^{-1}, determined by the freezing process used (Table 5); and P and R = constants for material geometry, determined from Table 6 or Figure 15.

Table 5. Approximate values of surface heat transfer coefficients h for various food freezing processes

Freezing process	h, W m^{-2} °C^{-1}
Still air	6 – 9
Air blast	18 – 30
Contact plate	120
Liquid immersion	600

Table 6. Constants P and R for Planck's equation for prediction of freezing times of basic geometrical forms

Constants	Geometrical form		
	Semi-infinite slab	Cylinder	Sphere or cube
P	1/2	1/4	1/6
R	1/8	1/16	1/24

Modifications of Planck's equation have been reviewed [37]. Because the assumptions used in Planck's equation result in underprediction of freezing times by 10 – 40 % [37], many of the alternative approaches aim to improve prediction accuracy. However, most of the formulas include additional terms and constants, resulting in rather complex computational requirements, especially for other than simple geometrical shapes. In practice, the freezing time required for food is usually determined empirically, using existing equipment; however, for design purposes, freezing time calculations must often be made.

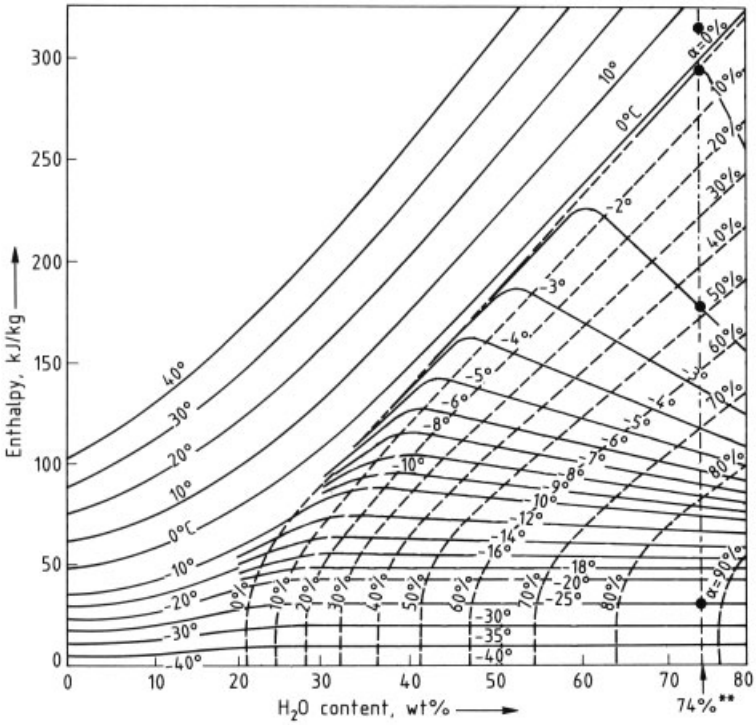

Figure 14. Enthalpy – H_2O concentration diagram for lean beef ∗
∗ = % Total water frozen as ice.
∗∗ Encircled points are used as examples in the text for beef at 74 wt % H_2O.

3.4. Effects of Freezing on Foods

Besides the main effect of freezing on food texture (see Section 3.2), freezing can result in many other important changes in foods and agricultural materials. Generally, foods can be classified in four broad groups according to their response to the freezing process:

1) Foods that can be frozen without major difficulties, e.g., bread and other bakery products, apple sauce, fruit juices, as well as the various frozen desserts based on ice cream, which is a food specifically formulated to withstand freezing.
2) Foods that need to be carefully processed to avoid problems in freezing, e.g., soups, gravies, other food emulsions, ready-to-serve frozen entrees, and certain fresh meats; without certain precautions, these foods can undergo deleterious changes resulting from the freezing process itself.
3) Foods that can be frozen without major difficulties but which will deteriorate in frozen storage; e.g., frozen fish, poultry, or certain formulated meat products may deteriorate as a result of fat oxidation or chemical changes in the protein structure.
4) Foods that cannot be processed by freezing without permanent quality changes, e.g., most fresh fruits, particularly those with a delicate texture such as strawberries or raspberries, lettuce, some cheeses, and beer.

In freezing fresh meats, care must be taken to avoid freezing during rigor mortis because, after thawing, the meat shows considerable and permanent toughening caused by the phenomenon of thaw rigor [38]. Meat frozen before rigor mortis sometimes does not experience this major defect; however, the safest practice is to freeze postrigor, well-aged muscle.

Cryoprotectants. A concept of cryoprotective additives or cryoprotectants has been introduced in cold temperature biology and medicine to minimize undesirable effects of freezing [32],

[34]. Dimethyl sulfoxide (DMSO) is an example of a cryoprotectant being used by cryobiologists because of its significant effect on freezing point lowering, thus effectively permitting cold temperature storage of biological materials without the damaging effects of ice crystal formation. Whereas DMSO is not an approved food material, the use of sugar for freezing strawberries, raspberries, and other delicate fruits is based on the same general concept. Discovery of other food-approved cryoprotectants could significantly increase the applicability of freezing to other fresh foods. An ideal food-grade cryoprotectant should be nontoxic, fully miscible with water, and easily penetratable through the cell walls and membranes, because its primary functions are to lower the freezing point, to dilute the "corrosive" concentrated solution inside the cells, as well as to minimize cell shrinkage by filling the voids left by the water deposited on the ice crystals once the freezing process begins.

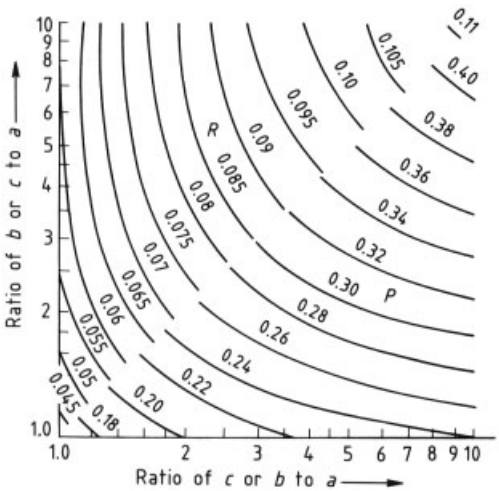

Figure 15. Material geometry constants for a brick-shaped object ∗ for the Planck's freezing times equation
∗With dimensions $a \cdot b \cdot c$, where a is the thickness (shortest side).

3.5. Frozen Storage and Thawing

In addition to the various effects of the freezing process, the quality of frozen foods can be altered by the frozen storage or by the thawing regime used. Major changes can occur on *extended storage of frozen foods* including loss of protein solubility, oxidative reactions of fat, or various enzymatic reactions if the enzymes were not properly deactivated before freezing. Because some of the constituent water remains unfrozen at conditions common to frozen food storage, some chemical and enzymatic reactions occurring in the unfrozen solution can be accelerated as a result of the concentration effect [34], particularly if the temperature of the frozen storage increases suddenly. Fluctuations in storage temperature, occurring commonly in industrial or household conditions, can have a profound effect on changes in crystal size distribution throughout the food. Increased temperature causes a certain amount of melting, especially of the smallest crystals. When followed by a temperature drop, repeated freezing subsequently favors growth of larger crystals.

Frozen storage of improperly packaged food can result in drying of the food surface, creating the industrially important *freezer burn.* This is caused by sublimation of the frozen water from the surface layers, resulting from the temperature difference between the food and the coldest spot in the frozen storage, i.e., the coils of the freezing agent evaporator of the mechanical refrigeration system. As a result, the coils become covered with an increasingly thicker layer of ice, which decreases the efficiency of the cold storage, thus necessitating periodic defrosting.

Further damage to frozen food quality can be caused by an improperly conducted *thawing process.* Because of the four-times lower thermal conductivity of liquid water as opposed to ice, thawing by contact with the outside environment is slower than freezing under otherwise equal driving force conditions. Acceleration of thawing by warm air (as in defrosting at room temperature overnight) may cause rapid growth of microorganisms on the surface layers of the thawed product, while hindering reabsorption of thawed water, thus creating the unsightly and often nutritionally wasteful drip loss. The use of microwave energy, which is not transferred by conduction through the thawed food layers, is the fastest and probably least damaging thawing process.

3.6. Freeze Concentration and Freeze Texturization

Freeze Concentration. Besides the primary use as a long-term preservation technology, the

freezing process can be used in other areas of food processing. Freeze-concentration technology was mentioned briefly in Section 2.2 as one alternative for liquid-food concentration. The freezing out of water and subsequent removal of the ice crystals from heat-sensitive solutions have been accomplished in various industrial or pilot-plant designs (Fig. 16); the practical use of the concept has been limited in the past by high loss of concentrated solution in the ice crystal removal step. For efficient concentrate removal in the wash columns or other devices used, large enough ice crystals must be developed to permit effective separation.

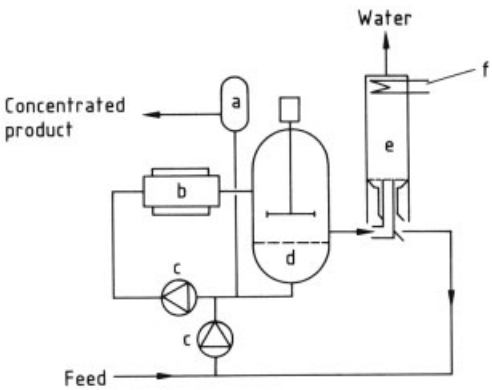

Figure 16. Principle of a freeze-concentration system
a) Expansion vessel; b) Scraped-surface cooler; c) Pumps; d) Ice crystallizer; e) Wash column; f) Ice melting

Freeze Texturization. Freezing of proteinaceous slurries and gels may result in formation of a fibrous structure as a result of controlled ice crystal development. The concentration of the unfrozen protein slurry by ice crystals unidirectionally grown in the proteinaceous matrix may cause formation of a rigid protein network resembling meat muscle or other fiberlike structure; when freezing is multidirectional, a spongelike structure may result as in the traditional Japanese process of tofu freezing to produce a modified "kori-tofu" product. Although the scientific explanation of the freeze texturization process is still incomplete, research has resulted in several patents and much new information [39]. The freezing process itself is not sufficient for long-term preservation of the protein texture, which must often be permanently stabilized by dehydration with alcohol or by heating the thawed protein matrix.

4. Food Preservation by Heating

4.1. Pasteurization and Sterilization

Microorganisms in their various biological forms are sensitive to elevated temperature. Growth of vegetative microorganisms in foods ceases at temperatures >45 °C, as evidenced by the common practice of "hot counters" used by fast-food service establishments; prolonged exposure to temperatures above 62 – 65 °C is lethal to most viable microorganisms. Although heating of many foods is practiced for reasons other than to retard their spoilage, all processes using heat, such as boiling, frying, stewing, smoking, or baking, result in foods with extended shelf lives. The vegetative cells of the main pathogenic organisms that may be associated with foods (several species from the genera *Salmonella, Staphylococcus, Clostridium, Lysteria*) are not very resistant to heat and are killed by short heating at temperatures above 72 – 75 °C. This principle (first discovered in 1864 by the famous French microbiologist Louis Pasteur, 1822 – 1895) is the basis of *pasteurization*, a controlled heating process used to eliminate any dangerous pathogens that may be present in milk, fruit-based beverages, some meat products, and other foods which are commonly subjected to this treatment. A similar controlled heating process, referred to as *blanching*, is used in the processing of fruits and vegetables; its main purpose is to deactivate the many enzymes present in the plant materials belonging to this food category. Both pasteurization and blanching are based on the use of the minimum heat requirement needed to deactivate specific microorganisms or enzymes, thus minimizing any quality changes in the foods themselves.

The effectiveness of controlled heating processes can be increased by prolonged treatment or by increasing the treatment temperature; however, increasing the severity of the treatment may have a negative effect on some food quality parameters such as flavor, texture, or content of heat-labile vitamins. Typical time – temperature combinations used by the food processing industry are shown in Table 7. In general,

pasteurization and similar processes prolong the product shelf life especially in refrigerated storage, but do not result in shelf-stable products because many heat-resistant spoilage microorganisms can survive the heating conditions employed.

Table 7. Examples of heat treatment combinations used in food processing

Process	Heating temp. holding time	Use
Bulk liquid pasteurization	63 °C/30 min	Vat pasteurization of milk
High-temperature – short-time pasteurization (HTST)	72 °C/15 s	Continuous pasteurization of milk for safety
Ultrapasteurization	138 °C/2 s	Pasteurization of milk for extended shelf life
Cooking of meat products	to 65 – 75 °C internal temperature	Ready-to-eat products (ham, meat loafs, frankfurters, etc.)
Blanching of vegetables	variable (e.g., 75 °C/5 min)	Enzyme deactivation, tissue softening
In-bottle pasteurization	60 °C/10 min	Shelf-life extension of beer

For long-term preservation of foods by heating, the treatment used must be severe enough to destroy all forms of microbial life that could be activated in the processed foods during prolonged storage. Whereas heat-treated shelf-stable foods are often referred to as sterile, a more appropriate term used by the industry is *commercial sterility*, denoting the absence of microbial growth in the processed foods in commercially achievable handling and storage conditions, rather than total absence of any evidence of microbial life in laboratory tests.

The main difference between pasteurization and sterilization treatments is the need to deactivate the bacterial spores. These sometimes very heat-resistant dormant forms of certain types of bacteria can germinate in improperly sterilized foods, causing rapid outgrowth of the viable cells thus formed.

4.2. Theory of Microbial Deactivation by Heating

4.2.1. Heat Resistance of Microorganisms

The basis for design and conduct of the various heat preservation processes are the experimentally determined and theoretically generalized thermal destruction relationships for the various microorganisms of interest. The data obtained from thermal destruction experiments indicate that heat deactivation of microorganisms generally follows *first-order reaction kinetics*, the rate of thermal deactivation being proportional to the concentration of the microorganisms present:

$$\frac{dc}{dt} = -kc$$

where c is the concentration (usually expressed as the number of organisms per gram or milliliter of food) and t the heating time at a given temperature. Thus, the time of a heating process required to deactivate a certain proportion of the microbial population present initially (c_0) is obtained as

$$t = D \log \frac{c_0}{c}$$

where D is a death rate constant referred to as the decimal reduction time (DRT), i.e., the time needed to accomplish a tenfold (1 log cycle or 90 %) reduction of the microorganisms present (Fig. 17 A).

For a given microorganism, different *reaction temperatures* result in different D values; as shown experimentally, the effect of temperature on the D value can also be generalized by a first-order line, as shown in Fig. 17 B. This plot, referred to as thermal death time (TDT) curve, is of fundamental importance for all calculations required for thermal deactivation of microorganisms in foods. The slope of the TDT curve indicates the sensitivity of the given microorganism to temperature changes; the steeper the TDT curve, the more sensitive the organism is. The negative reciprocal of the slope, referred to as the z value, gives the temperature interval for a tenfold change in the D value. All points along the TDT line for a given microorganism represent time – temperature combinations of equal lethal effect.

The heat resistance of microorganisms (as well as of enzymes, vitamins, and other heat-sensitive food chemical components) is adequately characterized by the *D and z constants*. The typical D value given in literature is for processing at the reference temperature of 121.1°C or 250°F, denoted as D_0. The D_0 and z values are different for the many microorganisms of

importance in canning, as shown in Table 8. For heat processing of shelf-stable foods, the thermophilic spore-forming spoilage organisms are of primary importance because these often show the highest heat resistance as indicated by D_0 values of 4 – 5 min, with a few exceptional cases of D_0 >50 min recorded in the literature [40]. Both D and z values are affected by environmental conditions such as pH, food composition, or the presence of compounds with a "shielding" effect, e.g., fat. From the standpoint of food safety, the heat destruction characteristics of spores of the toxin-producing anaerobic bacteria *Clostridium botulinum* must be included in the process calculations, although their relatively low heat resistance (D_0 ca. 6 – 12 s) usually provides for a large margin of safety as compared to the heat-resistant thermophiles.

Table 8. Heat resistance parameters for microbial spores in food sterilization ∗

Microbial systems	D_0, min	z, °C
Thermophilic spore formers:		
Bacillus stearothermophilus	4 – 5	8 – 13
Desulfotomaculum nigrificans ∗∗	2 – 3	9 – 12
Clostridium thermosaccharolyticum	3 – 4	9 – 13
Bacillus coagulans	0.02 – 0.07	7 – 10
Bacillus subtilis	0.5 – 0.8	7 – 13
Mesophilic spore formers:		
Clostridium botulinum	0.15 – 0.3	8 – 12
Clostridium sporogenes	0.1 – 1.5	8 – 13
P.A. 3679	0.7 – 2.5	9 – 12

∗ Orientation data, generalized from various sources.
∗∗ Heat-resistant strains may show D_0 > 50 min in some types of foods.

The effectiveness of a process aimed at microbial destruction by heat is determined by the *total thermal death time value*, denoted as F_0 and calculated as a multiple of D_0 values for the most detrimental organisms in each specific case:

$$F_0 = n \cdot D_0$$

The constant n, sometimes referred to as the sterilizing value, is defined as the logarithmic span of the survivor curve for the initial and final numbers of the microorganisms c_0 and c, respectively, thus

$$n = \log\left(\frac{c_0}{c}\right)$$

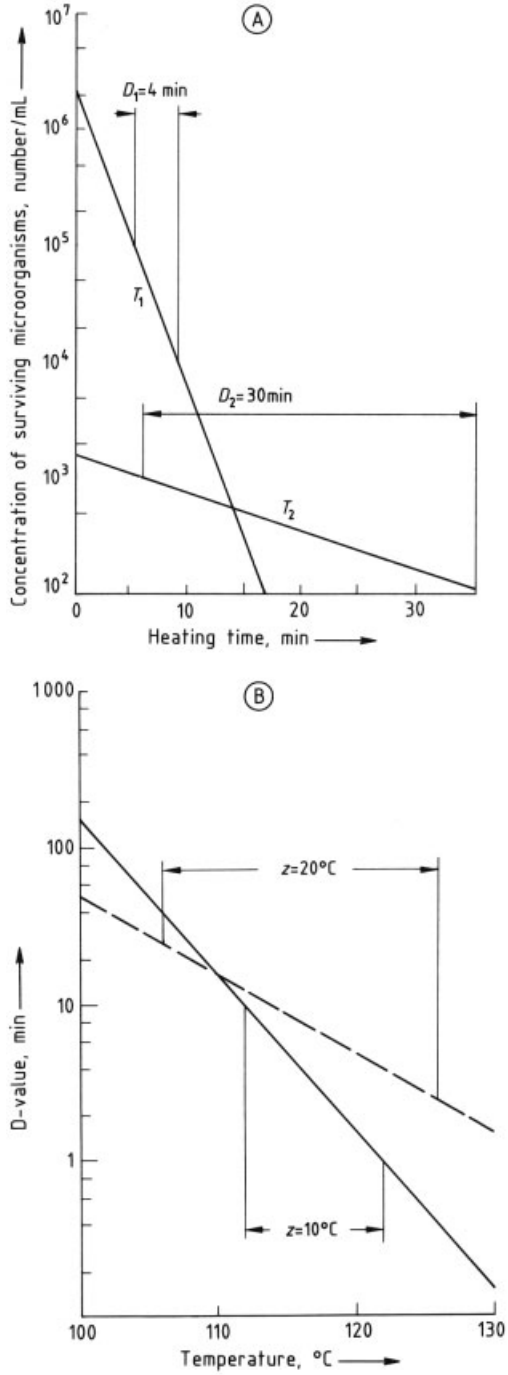

Figure 17. Thermal destruction relationships in heat processing
A) Microbial survivor curves and corresponding D values at two different heating temperatures T_1 and T_2.
B) Thermal death time curves and corresponding z values for microorganisms and enzymes. ⟶Microorganisms; – – – Enzymes

Whereas the initial number of the microorganisms of concern (c_0) is usually estimated for each specific food, the real meaning of the final number c is the desired probability of survival of the organism. For *Clostridium botulinum*, the *12 D concept* is used in heat processing of all shelf-stable foods with pH $\geq$ 4.5; this means that if the initial population of *C. botulinum* spores is 1 spore per gram of food (not an unreasonable assumption based on experimental observations), a process satisfying the $F_0 = 12\,D_0$ requirement should result in $<10^{-10}$ probability of spore survival in a 100-g container, or, alternatively, in one surviving spore in 10^{10} of these containers. Because D_0= ca. 0.25 min for the most heat-resistant strain of this dangerous mesophilic pathogen, the minimum thermal death time required of all processes is $F_0 = 12 \cdot 0.25 = 3$ min. In actual industrial practice, the margin of safety for *C. botulinum* is usually much greater than 12 D because the processes are designed for destruction of much more heat-resistant thermophilic spore forming organisms with D_0 values of 4 – 5 min.

Because *C. botulinum* growth with toxin production does not occur at pH $\leq$4.5, *food acidity* is an important determinant of the sterilizing process conditions used. Low-acid products with pH >4.5 must be processed by the "pressure" cook method at a temperature >100 °C and must satisfy the 12 D concept for *C. botulinum*. High-acid foods (pH <4.5) need not be processed at elevated pressure; heating at atmospheric conditions, i.e., at 100 °C (the atmospheric cook) is usually sufficient to destroy spores of acid-tolerant spoilage organisms.

4.2.2. Process Calculations for Canned Foods

The typical F_0 values used by many processors correspond to a ca. five-to-sixfold reduction in D_0 for the most heat-resistant organisms. This minimizes any substantial quality changes in excessively heated canned foods while still achieving an acceptable rate of spoilage from nonsterile cans. For cans with a typical microbial load of ca. 1000 spores per can, the resulting spoilage rate is <1 per 1000 containers.

While the basic heat process design calculations are usually carried out for the theoretical F_0 values related to the reference temperature of 121 °C (250°F), the actual heating of the can content shows a time-dependent *heat penetration profile* (Fig. 18). The type of food product being processed, its thermal properties, and the can size and shape determine the heat penetration characteristics of the process. Because of the conduction heating mode, solid and semisolid foods (e.g., luncheon meats, gels, pastes, and purees) require a much longer heating time than liquid foods and liquids containing particulates (peas, corn, beans, etc. canned in liquid brines), which heat by convection. The experimentally determined heat penetration curve, with the TDT constants (D_0 and z) of the most deleterious organism of concern for the given food and the desired probability of spoilage are used to determine the required total lethal effect and, thus, the necessary time of heating.

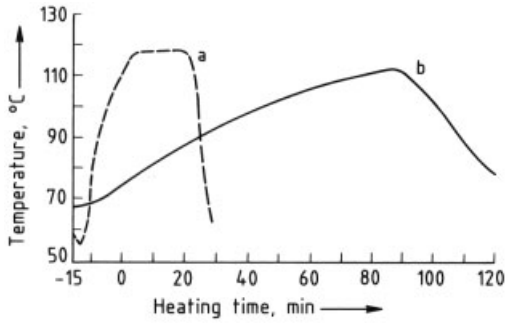

Figure 18. Heat penetration curves
a) Convection heating (peas in brine); b) Conduction heating (luncheon meat)

The principle of the thermal process calculations can be illustrated by the straight-line formula describing the TDT curve (Fig. 17) for the heating times t_0 and t_e at the reference temperature θ_0 and any other experimental temperature θ_e, respectively:

$$\log \frac{t_e}{t_0} = -1/z\,(\theta_e - \theta_0) = \frac{\theta_0 - \theta_e}{z}$$

Using this formula for t_0= 1 min and θ_0= 121 °C, the lethal effect of any heating time t_e at temperature θ_e can be expressed in equivalent lethal time E_t defined as

$$\log E_t = \frac{\theta_0 - \theta_e}{z} \quad \text{or} \quad E_t = 10 \exp \frac{121 - \theta_e}{z}$$

The reciprocal values of the equivalent lethal times are referred to as lethality units L (thus,

$L = \frac{1}{E_t}$), indicating the lethality of heating for 1 min at a given temperature in terms of equivalent time required at 121 °C. For the customarily used $z = 10$ °C and common process temperatures θ (in °C), calculated values oflethality units L (in min equivalents at 121 °C) are as follows:

$\theta = 111$	$L = 0.100$	$\theta = 121$	$L = 1.000$
112	0.126	122	1.259
113	0.159	123	1.585
114	0.199	124	1.995
115	0.251	125	2.512
116	0.316	126	3.162
117	0.398	127	3.981
118	0.501	128	5.012
119	0.631	129	6.310
120	0.794	130	7.943

For more specific lethal tables including values other than $z = 10$ °C, see [41]. The total lethality of a given process can then be determined by integration of the lethality effects of the given heat penetration curve (shown in Fig. 18) for the total process time; thus,

$$F_0 = \int_{t_0=0}^{t} L\,dt$$

This is the basis for the *general method* developed by BIGELOW in 1920.

Determination of total process time to achieve a desired lethality (or the lethality evaluation of a given process time for each specific product or can size) involves considerable computational complexities, and several mathematical approaches, as described in specialized texts, have been developed to simplify the task [41], [42]. These include the improved general method, the formula method, the nomogram method, and various computer-based algorithms. In all cases, the data required for the process evaluation calculations include information about the type of microbial populations of the food being processed, the D_0 and z values of the most detrimental organism (i.e., the most heat-resistant or most populous contaminant or pathogen), the desired rate of survival/spoilage, the processing equipment temperature, and the heat penetration for the specific food and container being used. On the basis of the relative population numbers of the various microorganisms present and their D_0 and z values, the most detrimental organism against which the process is designed must be identified, and the final process lethality must be verified experimentally. Often the verification experiments include processing of packs inoculated with spores of an appropriate indicator organism such as the putrefactive anaerobe *Clostridium sporogenes*, which is similar to *C. botulinum* in terms of its thermal death time parameters. (One of its more heat-resistant strains used for the inoculation pack studies is often referred to as P.A. 3679.) Some orientation F_0 values of commercially processed canned foods are as follows:

Vegetables in brine	3 – 4
Baked beans	4 – 6
Fish products	5 – 8
Custards and puddings	4 – 10
Meat stews	12 – 15

The experimental determination of heat penetration data is carried out by inserting specially designed thermocouples in the coldest spot of an experimental can. The location of the cold spot depends on the heat transfer characteristics of the food being processed. From this standpoint, foods can be divided into three categories:

1) conduction heating foods, including mainly solid and pasty products, for which the coldest spot is usually in the geometrical center of the can
2) convection heating foods, typically liquids and liquids with small solid particles, for which the coldest spot is usually near the can bottom because of the convection currents generated in packs filled with these foods
3) intermediate or mixed heating mode systems, including liquids that may gelatinize on heating (starch-based foods, heat-sensitive protein foods such as egg white) or solids and highly viscous pastes that may liquefy at higher temperatures (high fat products, products stabilized with hydrocolloids), for which the position of the cold spot varies and changes as heating progresses.

Whereas the basic heating patterns are determined by the physical properties of the foods being heated, the heating rates are influenced by the heating medium used (steam, superheated water, or direct flame), and the particular process equipment used.

4.3. Food Canning Technology

Processing of foods in hermetically sealed metal cans is the most often used preservation tech-

nology for food sterilization by heating. In the broader sense, the term canning may be used to denote the general in-container sterilization principle for various other containers (glass jars, bottles, and plastic pouches) as well (→ Foods, 4. Food Packaging).

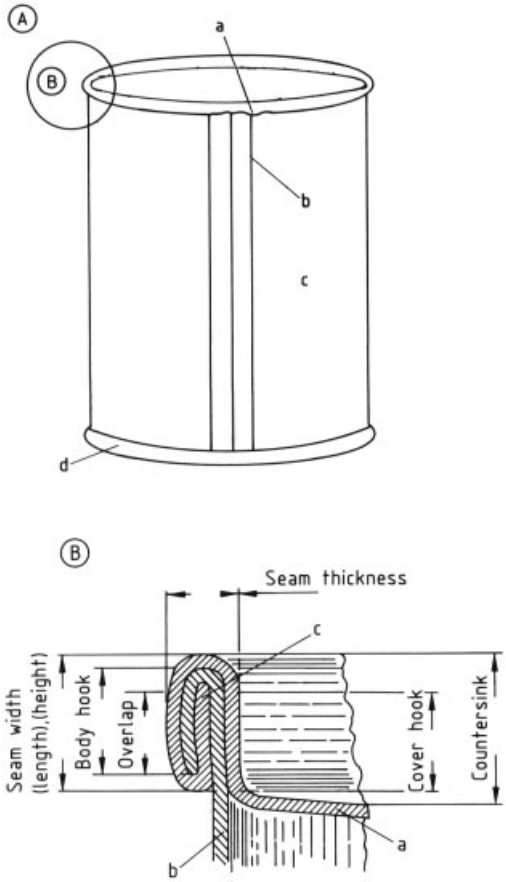

Figure 19. Diagrammatic representation of sanitary can components
A) Sanitary can architecture: a) Top seam (top lid); b) Side seam; c) Can body; d) Bottom seam (bottom lid) B) Detail of the double seam construction with the main measurement specifications: a) Cover lid; b) Can body; c) Sealing compound

The main steps involved in conventional canning operations are (1) the preparatory treatments (cutting, cleaning, pasteurization, blanching, ingredient mixing, etc.), (2) filling and sealing the containers, and (3) heating the hermetically sealed containers in conditions that prevent any contact of the food with the outside environment. The container seals must be strong enough to withstand the heating – cooling treatment with no mechanical damage that might allow for post-processing contamination. The typical cylindrical metal can used in conventional thermoprocessing of foods is specifically constructed for this purpose.

4.3.1. Sanitary Can Architecture

The metal containers used for traditional canned foods are referred to as "sanitary cans" (tins); this term denotes a three-piece container made from steel. Sometimes the steel sheet is plated with a thin layer of tin to inhibit the interaction of the food with the steel can material. Protective coatings commonly called "enamels" are also used for the same purpose. The basic aspects of sanitary can architecture are illustrated in Figure 19. Other types of cans may be used by the food industry for foods not subjected to the heat sterilization process (e.g., beer and pop cans, containers for pasteurized hams, or sweetened condensed milk cans); although these cans also fully satisfy the requirements of food hygiene and sanitation, they are not termed sanitary cans from the standpoint of construction.

The most important aspect of sanitary can architecture is the *double seam* (Fig. 19 B), which allows for attachment and hermetic sealing of the bottom lid (at the can producing factory) and the top lid (attached in the can sealing operation after filling with food). Routine periodic measurements of the double seams produced by the can filling – closing machinery must be carried out as one of the most important quality control procedures determining the safety of the finished product. Improperly formed seams can lead to mechanical defects created during heat sterilization processing and ultimately to nonsterility of the finished product.

Chances for oxidation reactions in the sterilized food in storage are minimized by the vacuum in the can created during filling. Several alternative evacuation procedures used by the industry include mechanical vacuum pumps and the "hot fill" and "steam exhaust" methods (in the last two methods the vacuum is created after cooling the container when the can content volume decreases or the steam in the head space of the can condenses). The creation of the vacuum eliminates reactive oxygen from the can. In addition, it also helps minimize potential seam

damage due to pressure buildup inside cans processed under higher than atmospheric pressure.

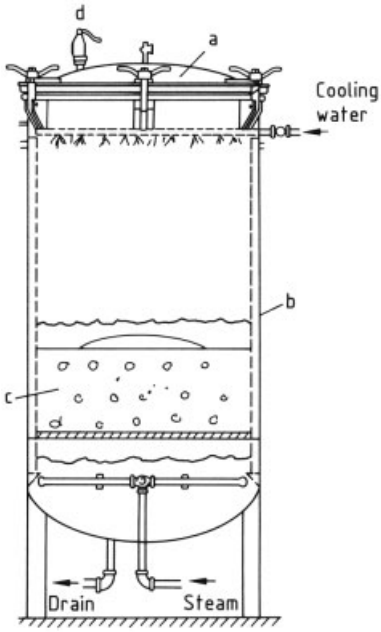

Figure 20. Principle of a still vertical retort
a) Retort lid; b) Retort body; c) Basket with cans; d) Safety vent

4.3.2. Conventional Retorting Equipment

The basic concept of the can sterilizing equipment is a pressurized vessel (retort) in which the cans are heated by saturated steam, superheated water under overpressure, or by steam – air mixtures [43]. After completion of the required heating process, the heating medium is replaced by cold water. The initial heating and cooling rates must be carefully controlled to avoid damage to the containers due to the pressure differences inside and outside the can.

The actual engineering design of the retorts varies widely in basic configuration; the vertical "still" retort (Fig. 20) is the simplest type of equipment used. Horizontal retorts that agitate the cans have been developed to increase the efficiency of heat transfer to the cans in the retort, whereas for increased capacity large "tunnel-like" horizontal designs can be used for whole pallet or crate loads of cans (Fig. 21). In general, conventional retorts are operated in a batch mode with the total required heating time determined according to the procedure outlined, and the heating – cooling cycle is carried out in the closed vessel in a predetermined processing sequence.

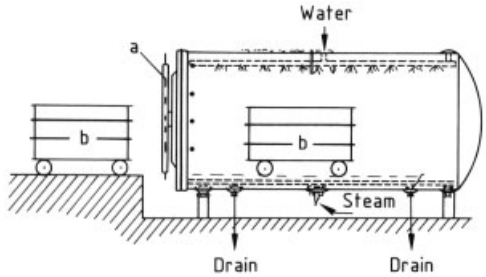

Figure 21. Horizontal retort with a crate carriage
a) Retort door; b) Crate carriage

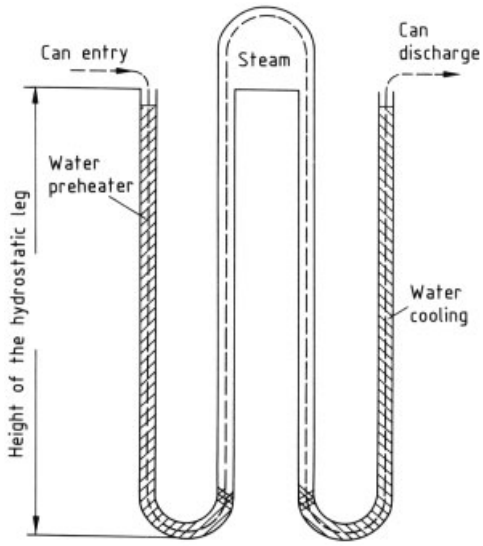

Figure 22. The hydrostatic sterilizer principle

4.3.3. Hydrostatic Sterilizers

Continuous retorts have been developed and are becoming popular because they increase process efficiency, reduce labor costs, and improve the heat transfer rate by can agitation. In general, the preheating, heating, and cooling operations are carried out simultaneously in various parts of the equipment, with the cans travelling through the retort. In particular, continuous "hydrostatic cookers" (Fig. 22) have been successful industrially. The concept of the hydrostatic sterilizing tower is based on the use of steam as in a classical retort combined with the control of the proper

steam pressure needed for the processing temperature by preheating and cooling hydrostatic legs, placed before and after the steam sterilizing area. The cans are transported through the various sections of the tower by a conveyor belt, thus allowing for some agitation of the contents. Because of the height of the hydrostatic legs required, these sizeable installations are suitable for large throughputs.

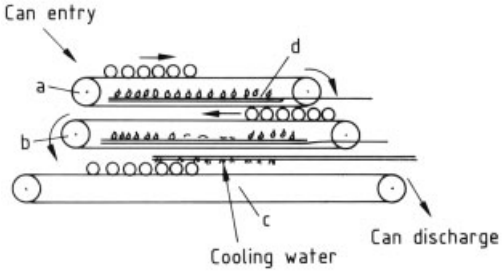

Figure 23. Principle of the flame sterilization equipment
a) Heating conveyor; b) Holding conveyor; c) Cooling conveyor; d) Gas flame burners

4.3.4. Flame Sterilization

Direct heating of sealed cans over an open flame provides an effective thermoprocessing alternative, particularly for products packaged under high vacuum [44]. However, despite the potential advantages of the flame sterilization process, industrial use is not widespread and is presently limited mainly to canned mushrooms and some other acidified fruit and vegetable products. The *Steriflamme machinery design* (Fig. 23) consists of a set of conveyors carrying the cans with forward – backward rotational movement over rows of "heating" and "holding" natural gas burners followed by water spray cooling. An important aspect of flame sterilizers is the requirement for properly designed process control because the heating medium temperature may fluctuate, unlike in steam-based retorts. Various approaches to monitoring the adequacy of the heating process have been developed; one possibility is automatic surface-temperature measurement of each can by contact thermocouples at the beginning and the end of the temperature holding section. With a computerized data collection system, the F_0 of each can can be assessed against a predetermined necessary minimum F_0 value. The present state of this new thermoprocessing technology [44] appears to make it suitable mainly for specialized applications in small-output situations.

4.3.5. Retortable Pouches

As an alternative to the expensive and bulky steel can, thin-profile flexible containers made of plastic film (Foods, 4. Food Packaging) are rapidly gaining popularity in some countries. The principle of the retort pouch technology is the same as that of traditional canning: foods hermetically sealed in a plastic bag are subjected to a heating – cooling process adequate to achieve commercial sterility. However, because the pouch is constructed as a thin package, the heat transfer to the cold spot is much faster with less overheating of the outside food layers, in particular for convection heating products. The plastic films that are used must be sturdy enough to withstand the abuses of the handling and retorting treatments. For added mechanical protection and to minimize the oxygen and water vapor transmission through the plastic film in storage, the laminated packaging materials often include aluminum foil, and the finished pouch is marketed in a secondary paper box. The heat-sealability of plastic films is an advantage for easy seam formation. The main disadvantage of retort pouch technology appears to be the low capacity of the filling – sealing machinery due to the lightness of the package, resulting in higher labor requirements. The basic principle of pouch-pack technology was modified further for marketing a line of heat-sterilized complete entrees with an 18-month shelf life [45]. The dishes are vacuum packaged and sterilized in a patented, complex container consisting of a plastic tray sealed with a semirigid lid. The thickness of the complete Menu-Box container is ca. 40 mm, thus requiring a less damaging heat process similar to the standard pouch-pack.

Retortable pouches are usually heated by immersion in water with overpressure or in steam – air retorts. Both of these systems allow for proper maintenance and control of the external pressure at a level higher than that in the pouch. This is needed to protect the package from bursting because of internal vapor pressure buildup and to counteract expansion of the gases contained in the food. In the early stages of

pouch-pack technology development, the main concern was potential underprocessing of certain packages within the retort load due to a lack of homogeneity of heating environment and production of cold spots. However, the development of special retort racks for the pouches and the use of circulating and positive flow retorts have largely overcome these problems, as evidenced by the widespread use of retort pouch technology in Japan and other industrialized countries [43].

4.3.6. Aseptic Canning

Although built for the use of standard metal cans or composite cans made of different materials, the *James Dole aseptic canning system* differs from conventional canning. Rather than being heated inside sealed cans, foods are sterilized outside any package and the sterile food is then filled into presterilized cans. The filling and sealing of the cans is carried out under aseptic conditions provided usually by a steam-lock atmosphere or by dry, hot air. Thus, this system uses the same approach as Ultra-high-temperature processes (see Section 4.4.) for milk and other liquids and is, in fact, a special case of food sterilization technology known as aseptic processing.

4.4. Ultra-High-Temperature Aseptic Processing

Ultra-high-temperature (UHT) technology is widely used for processing milk and related dairy products, fruit juices, various beverages, and other mainly liquid foods. Although the term "UHT processing" in its strict meaning refers to the outside-the-package sterilization step only, it is commonly used to denote the technological sequence of two distinctly different basic unit operations: food sterilization and aseptic packaging. In addition, several auxiliary installations are often required in these continuous processing lines for specialized applications.

4.4.1. Theoretical Basis of UHT Processing

Typically, the *process temperatures* used in UHT sterilizing equipment are substantially higher than those for conventional retorting. In the dairy industry, temperatures of 135 – 150 °C are suitable for most of the products manufactured. At these very high temperatures, the desired thermal death time values (F_t) for the typical spore-forming spoilage microorganisms are on the order of several seconds, whereas the corresponding F_0 values would be on the order of several minutes. This is so because the microbial z values are generally small (Table 8), indicating a high sensitivity of microorganisms to temperature change. In contrast, the z values characterizing many undesirable chemical reactions caused by heat (vitamin destruction, nonenzymatic browning, and flavor changes) are larger (Table 9), resulting in a much flatter TDT curve. As illustrated in Figure 24, the use of ultra-high processing temperatures requires only a very short processing time, but still gives adequate lethality for microbial destruction while minimizing most of the deleterious chemical changes. Thus, the sensory qualities of UHT-processed products are usually higher than those of comparable in-container sterilized items. For this reason, the International Dairy Federation has defined UHT milk as a product distinctly different from both pasteurized and traditional "in-bottle" sterilized milk [40].

Table 9. Heat resistance parameters for chemical reactions in food sterilization ∗

Reaction	D_0, min	z, °C
Vitamin destruction	120 – 1200	25 – 45
Nonenzymatic browning (milk)	1 – 2	25 – 30
Color compound changes (chlorophyll and anthocyanins)	10 – 120	20 – 60
Enzyme deactivation	1 – 15	6 – 60
Denaturation of whey proteins (milk)		10 – 35

∗ Orientation data, generalized from [40] and other sources.

The *shelf life* of UHT-processed dairy products and other foods is not indefinite, as can be the case with properly processed canned foods. The main cause of spoilage is usually incomplete deactivation of some of the indigenous enzymes present. As seen in Table 9 and Figure 24, because of the higher z values for enzymes, the very short processing times do not achieve the same level of enzyme destruction as the much longer equivalent processing times required for sterilization at lower temperature. The slow but

noticeable residual enzymatic activity often limits the shelf life of UHT milk to ca. 6 months or less, depending on the type of heating equipment used.

The two main categories of UHT sterilizers are the direct or the indirect systems. The common features of all UHT heating processes are the very rapid heat-up and cool-down rates achievable with the liquids being processed. As a result, the total amount of heat imparted to the food can be controlled with great accuracy.

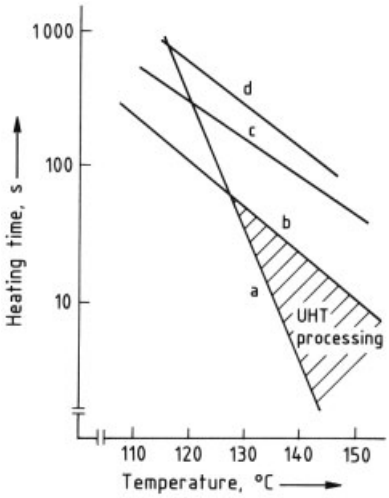

Figure 24. Optimum process conditions for UHT sterilization
a) Microbial spores destruction, 5 *D*; b) Vitamin destruction, 0.1 *D*; c) Enzyme deactivation, 1 *D*; d) Limit for browning

4.4.2. Ultra-High-Temperature Sterilization

The equipment used for continuous sterilization is designed to achieve the fastest possible heat transfer in the product. This is accomplished either by direct mixing of the heating medium, i.e., the culinary steam, with the food, or by heating a thin film of the food indirectly across a heat-transfer wall. The idealized time – temperature profiles of the two systems are illustrated in Figure 25.

The direct systems are capable of practically instantaneous heating because the condensing steam is mixed with the product providing a large amount of heat as the latent heat of condensation. The product being heated can be distributed as a thin free-falling film in a steam-filled chamber, as in the DASI system (Fig. 26), or the steam can be injected into the product in a special mixing head. In both cases the condensed steam must be subsequently evaporated to avoid product adulteration. Because heating an aqueous system to temperatures above 100 °C must be carried out under elevated pressure, the subsequent pressure drop caused by bringing the heated product to a vacuum chamber results in flash evaporation of some of the constituent water. Proper vacuum control insures that the amount of water evaporated equals that condensed in the product from the steam. As a result of evaporative cooling, the heat recovery into the incoming product is relatively poor in direct systems.

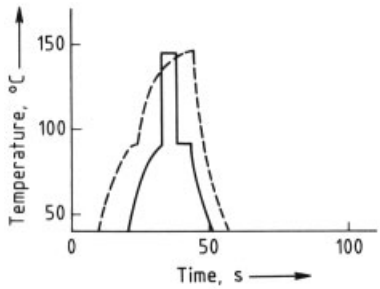

Figure 25. Idealized heating curves for direct and indirect UHT sterilizers [40]
– – – – indirect

The heating medium used in indirect UHT sterilizers is either superheated (pressurized) water or steam. The heat transfer into the product occurs across a heat-transfer wall of a pressurized plate heat exchanger or by a stainless steel pipe running through a steam chamber (Fig. 27). Although the heating rate is usually somewhat slower than in direct systems (Fig. 25), the possibility for heat regeneration from the heated product (Fig. 27) makes the indirect system more economical. However, product quality comparison for some dairy industry applications indicated less heat-damage-related changes in products obtained from direct systems [46]

4.4.3. Aseptic Packaging

The predominant aseptic packaging system used for the UHT-processed products is that developed by the Tetra-pak company *(Sweden)*.

The brick-shaped packagings are made of paper-based laminated material, which includes several layers of plastic film and aluminum foil (→ Foods, 4. Food Packaging). The packagings

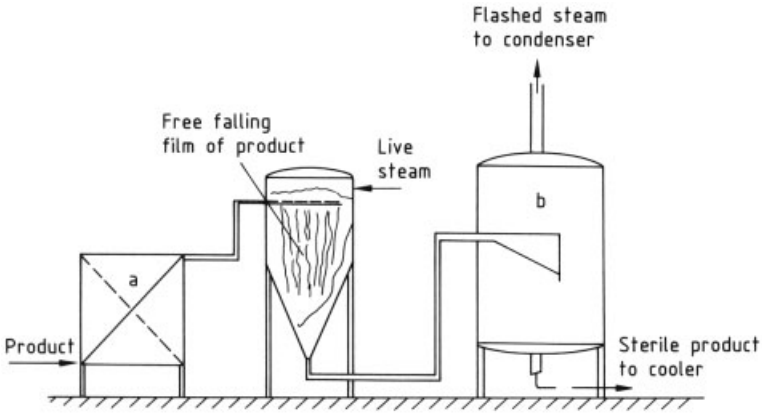

Figure 26. Principle of the milk-into-steam direct heating system DASI
a) Preheater; b) Flash evaporation vacuum chamber

are formed from a continuous strip of the material, which enters the filling machinery through a hydrogen peroxide bath for package sterilization. Subsequently, the strip is formed into a tube around the sterilized product feed line, and appropriate longitudinal and cross seams are made by heat-sealing the plastic inner surfaces as the product is being filled. This continuous aseptic packaging system is unique in that it does not require an aseptic chamber, as in the case of some competing systems like the preformed carton-based Combibloc, the various form-fill-seal or preformed cup systems [47], or the James Dole aseptic canning (see Section 4.3.6). Because the filling machinery may experience more frequent interruptions as compared to UHT sterilizers, aseptic processing systems customarily include an aseptic storage tank for the sterilized product before the filler.

Although UHT aseptic processing is predominantly used with liquid foods, systems allowing for handling liquids with particulates are also available. Some of these seldom used or experimental systems have been reviewed [44]. The technological approaches include separate heating of the liquids and the particles in a regular UHT machine and in a batch reactor, respectively, followed by aseptic mixing and packaging, or a scraped surface heating of the liquid – solid mixture with positive displacement feed pumps controlling the required pressures. Heating of particulates in a liquid food material that forms a part of the final product (e.g., meat in a sauce or vegetables in a soup) may result in substantial overheating of the liquid phase, contributing to possible quality defects [40].

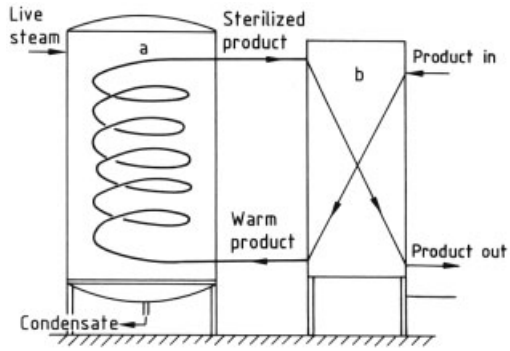

Figure 27. Principle of an indirect tubular UHT sterilizer
a) Sterilizing chamber; b) Regenerative heater–cooler

4.4.4. Quality of UHT-Processed Foods

Although UHT-sterilized products are shelf-stable and need not be refrigerated, their shelf life may be limited due to the *slow deteriorative chemical reactions.* In the case of milk and other dairy products, residual proteolytic activity may cause product gellation or sweet curdling [40], whereas incompletely inactivated lipases (especially the more heat resistant ones of microbial origin) cause hydrolytic fat breakdown, resulting in rancidity after prolonged storage. Because whey protein denaturation, which is the chemical reaction responsible for cooked flavor development in dairy products, has a relatively steep TDT curve (Table 9) with z values similar

to those for microorganisms, the cooked flavor development in UHT-sterilized milk cannot be fully eliminated by any time – temperature combinations. Direct heating systems appear to produce less cooked flavor because of better control of the heating process as well as subsequent vacuum treatment of the heatedmilk. The strong cooked flavor in freshly processed UHT milk diminishes after ca. 10 – 12 d storage.

The *nutritional quality* of UHT-processed foods, such as milk, is comparable to that of a fresh rather than a canned product [48]. The destruction of heat-labile vitamins by the UHT sterilization process is minimal, particularly in direct systems. However, some loss may occur during subsequent prolonged storage, especially because the plastic-to-plastic seams of the paper-based containers do not give absolute protection against oxygen penetration.

Nonsterility of UHT products may result from insufficient heat treatment, often related to presence of unexpected heat-resistant organisms, or because of postprocessing contamination particularly in the aseptic storage or the aseptic filling operation. The vacuum treatment of the sterilized product in direct systems increases the potential for postprocessing contamination in these machines compared to indirect processes. Any mechanical fault in the vacuum chamber, which allows entry of outside air, can cause major problems as in the case of imperfect seams in cans. The typical rate of spoilage in successful industrial UHT systems ($<0.1\,\%$) is comparable to the spoilage rate of conventionally processed canned foods.

4.5. Spoilage in Heat-Sterilized Foods

The shelf life of canned or UHT-processed foods depends on deactivation of all microorganisms capable of growing in the foods rather than on altering the environmental conditions necessary for microbial growth. Thus, any presence of viable microorganisms in a processed product stored without refrigeration may result in product spoilage, or even, in the case of surviving pathogens, in an outbreak of food-borne illness.

Two main reasons for microbial growth in sterile products are insufficient sterilization (underprocessing) or postprocessing contamination. *Underprocessing* may be caused by poor time – temperature control of an otherwise properly designed process, by the unexpected presence of highly heat-resistant strains [40], or by unusually high loads of microbial contaminants. Underprocessing often results in initial survival of small numbers of a specific organism giving rise to single-culture growth. This can be particularly dangerous in the case of toxin-producing pathogens growing without the competition of other spoilage organisms. Some of the typical spoilage bacteria found in underprocessed cans are homofermentative thermophilic spore formers, producing only acids as a byproduct of their growth; the resulting type of spoilage is termed "flat-sour," as no visible bulging of the can ends occurs. When the surviving organisms are heterofermentative, producing CO_2 gas in addition to the acid, the spoilage is easily recognized by the swollen package appearance. Organisms producing H_2S during their growth give rise to the "sulfur stinker" type of spoilage. Examples and characteristics of various spoilage types are as follows:

Flat-sour:	acid spoilage, no gas
Flipper, springer hard swell:	can swollen due to CO_2 gas produced
Sulfur stinker:	H_2S gas produced, darkening of can content
Postprocess contamination:	mixed microflora
Hydrogen swell:	chemical reaction of food with can material, H_2 gas only, no microbial spoilage

In contrast to underprocessing, *postprocessing contamination* typically produces a mixed population growth pattern. In the case of canned products, postprocessing contamination occurs because of imperfections in can seams, which allow entry of bacteria either during the cooling cycle in the retort resulting in creation of a vacuum in the can, or during storage. In aseptically processed foods, postprocessing contamination may be caused by failure of the various aseptic systems, needed for the food handling downstream from the sterilizer, as well as by faulty package seams.

Occasionally, *container swelling* may be experienced with heat-processed sterile foods without any signs of microbial growth. In canned products, this may be caused by the chemical reaction of an acid food with the metal of the container where the protective enamel or tin layers may have microscopic imperfections. If the resulting gas is exclusively hydrogen, the defect is

termed "hydrogen swell." Sterile foods in paper-based containers or in plastic retort pouches may show weak signs of swelling caused by gases liberated by reactions of incompletely deactivated enzymes. Incomplete degassing of products during filling before pouch-pack processing can also be a cause of package swelling. Despite the harmless nature of these "chemical swells," the potential health hazards involved in ingesting any sterilized food showing signs of spoilage should be considered when swollen shelf-stable food containers are handled.

To minimize spoilage problems at the retail level, canned foods as well as UHT-processed products are usually held in processors' warehouses for several weeks. During this time, randomly selected containers are subjected to quality control tests involving incubation at various temperatures and microbiological analyses. If a potential spoilage or a food-borne illness problem is discovered after the release of the product batch to commercial networks, elaborate and costly recalls may be necessary. However, despite a worldwide annual production of shelf-stable food containers in large amounts, the safety record of this important segment of the food industry is outstanding.

5. Food Preservation by Ionizing Radiation

5.1. Radurization, Radappertization, and Radicidation

The use of radiation energy for food preservation has been the subject of ongoing research since 1945. Much of the basic scientific information has been amassed at several centers around the world, including the U.S. Army Research Center at Natick, Mass.; Michigan State University; the Bundesforschungsanstalt für Ernährung (Federal Research Center for Nutrition), Karlsruhe, FRG; and similar centers in Holland, the former states of USSR, Israel, and other countries. Because irradiation can destroy microorganisms without any damaging effects on the main food nutrients or food quality attributes of many foods, early research has been focused on development of irradiation technology to replace heat sterilization. Other promising applications of irradiation included shelf-life extension of vegetables due to sprouting inhibition, or increased food safety by elimination of potential pathogens such as *Salmonella*. However, despite the main advantage of the process in terms of final product quality, the experimentally confirmed safety of irradiated food and changes in the regulatory position of many countries, which now consider irradiation a process (rather than a chemical additive, as in the past), food irradiation is a controversial issue and is not likely to gain widespread public acceptance for some time.

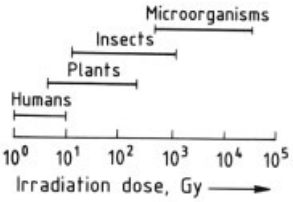

Figure 28. Relative lethality scale for ionizing radiations

The principle of food preservation by irradiation is the same as in heating: however, radiation energy (rather than heat energy) is used to deactivate microbial contaminants, resulting in elimination of microbial growth. When a low amount of radiation energy is used, elimination of only selected, radiation-sensitive organisms results in prolonged shelf life, but the product is not shelf stable; because this is the same as in pasteurization, the term *radurization* has been proposed for this use. For complete food sterilization by irradiation, the proposed term is *radappertization* (in honor of the inventor of the canning process, Frenchman Nicolas Appert, 1752 – 1841); however, the radiation energy requirement for this treatment exceeds permissible doses. Radiation energy can also be used for deinfestation of foods subject to insect contamination (tropical fruits, herbs, spices, grains, etc); the term *radicidation* is sometimes used to describe this treatment. Other food-related uses of irradiation have been proposed or approved in many countries; these include the inhibition of sprouting in stored vegetables (potatoes, onions, and garlic); the treatment of food processing wastes and other wastewaters for microbiologically safe disposal; low-dose irradiation for delayed ripening of fruits; sterilization of complete meals for hospital patients experiencing adverse

immunoreactions or requiring an aseptic environment; and sterilization of food packaging materials. A similar use for sterilization of laboratory and medical disposables is currently the largest industrial application of the bactericidal effects of irradiation.

5.2. Theoretical Aspects of Food Irradiation

When foods are irradiated to prolong their shelf life, the process is aimed primarily at *destruction of microbial viability*, similar to heat processing. The lethal effects of ionizing radiations are well-known since the times of W. K. Röntgen (1845 – 1923), Henri Becquerel (1851 – 1908), and the nuclear attack on Hiroshima (Japan). In general, the more complex the biological system, the lower energy dose is needed for a lethal effect. Deactivation of microorganisms requires much higher energy doses than insect deinfestation or sprouting inhibition (Fig. 28). Simpler chemical systems, such as single protein or enzyme molecules, are not greatly affected by the radiation energy.

The lethal effect of irradiation can be expressed in the same way as the effects of heat; whereas in heat-induced deactivation reactions, both time and temperature were important (see Section 4.2.1) as determinants of total energy dose, the irradiation dose is simply defined as the amount of radiation energy absorbed by the mass. The SI-unit of radiation dose is 1 gray (Gy), defined as the radiation energy of 1 J absorbed by 1 kg of mass; thus, 1 Gy = 1 J/kg. An older unit still widely in existence in the literature is 1 Rad (also rad or rd) and its multiples kRad (= 10^3 Rad) or MRad (10^6 Rad); the relationship between Gy and Rad is 100 Rad = 1 Gy. Although the radiation dose effect may be influenced by the rate of exposure or "dose rate" (the amount of energy absorbed per unit time), food-related irradiation effects can be accounted for by the total dose only.

The recommended *maximum dose* for food irradiation approved by the Joint Expert Committee on Food Irradiation of the United Nations (with internationally recognized authorities representing the FAO, the WHO, and the International Atomic Energy Commission) is 10 kGy. Practical doses considered adequate for specialized food application are ca. 0.02 – 0.15 kGy for sprouting inhibition, 0.15 – 0.3 kGy for insect deinfestation of fruits or for parasite control in meats (such as trichinosis in pork), and 1.0 – 2.0 kGy for radurization of meat, poultry, and seafood [49].

In a manner analogous to heat deactivation of microorganisms, a *decimal reduction dose* (DRD) has been determined for various types of organisms as the amount of radiation energy needed to deactivate 90 % (1 log cycle) of the populations present. Some orientation DRD values (in kGy) for several types of microorganisms are as follows:

Spores	
Clostridium botulinum	3.3 – 3.7
P.A. 3679	2.2
Vegetative cells	
Salmonella typhimurium	0.2 – 0.8
Staphylococcus aureus	0.1
Yeasts and molds	0.4 – 0.5

As can be seen, *Clostridium botulinum* is one of the most radiation-resistant organisms; this alone is a serious obstacle to development of shelf-stable irradiated foods for general use in the retail market.

The exact *mode of action* of the irradiation energy with respect to its lethal effects is not well understood. Possible alternatives include "direct hit" of the biological cell nucleus by the high-energy particles, or indirect chemical effects of the radiation energy on proteins, carbohydrates, or lipids, resulting in formation of highly reactive free radicals. These molecules, formed, e.g., by splitting a hydrogen atom from a hydrocarbon chain, can cause chemical changes within the living organisms. Similar chemical reactions also occur in the food components themselves, giving rise to various "radiolytic off-flavors" noticeable especially in foods irradiated at higher doses.

The *industrially applicable types of ionizing radiations* are β-rays or γ-rays. However, for practical food applications, β-rays (a stream of electrons generated by an electron accelerator) are less suitable than γ-rays (electromagnetic radiation of very short wavelength) because of the high cost of the equipment and the low penetrability of the β-rays energy. The γ-ray sources used commercially are predominantly the radioactive elements ^{60}Co or ^{127}Ce. The penetra-

tion depth of γ-rays (30 – 50 cm as compared to ca. 5 cm for β-rays) is a major advantage for irradiation processing of foods, as conceptual equipment designs can be based on processing of whole crate or pallet loads of products in commercial packages.

5.3. Food Irradiation Technology

The potential for food irradiation as a food preservation technology is hampered not only by unfavorable public image (common to any process based on radioactivity), but also by the high installation cost of the γ-irradiation equipment. Because of the need for highly sophisticated safeguards in handling the radioactive elements, the high throughputs achievable with even low emitting sources are of little consequence in reducing cost.

Typical plant installation proposals illustrate the basic common principle of a conveyorized product passage through a heavily shielded irradiation room [1], [50]. The bank of radiating rods is held submerged in a pool of heavy water, and a combination of safety switches and other precautionary measures prevent the constantly emitting source from being engaged in the working position before the food product is introduced. Various single- or multiple-pass designs can be considered, depending on the size of the packages to be irradiated, the limiting exposure dose to prevent any undesirable radiolytic product changes, and other technical factors [50].

Specialized equipment installations for individual food processors appear uneconomical at the present time. Where a whole segment of a given industry can be served in one location, cost-sharing may prove feasible. For example, deinfestation irradiation of papaya fruit shipped from Hawaii to the continental United States has been considered as a viable industrial option to prevent epidemic outbreaks of adult fruit flies [50]; in Canada, irradiation of fish for prolonged shelf life, allowing wide market distribution without freezing, has been studied in a pilot-plant project. In Japan, an irradiator for treatment of potatoes to prevent sprouting has been installed with a capacity of 15 000 t/month. In some cases, existing irradiator installations serving other than food industries could be used in co-processing agreements for selected food applications. Examples of industrial uses of food irradiation technology are listed in Table 10.

Table 10. Examples of industrial uses of food irradiation

Product	Reasons for irradiation	Dose, kGy
Potatoes, onions, and garlic	sprouting inhibition	0.02 – 0.15
Bananas, avocado, mango, papaya, and other tropical fruits	delayed ripening	0.2 – 0.4
Papaya, mango, peaches, other fresh or dried fruits, grain, and flour	insect deinfestation	0.15 – 0.75
Mushrooms, strawberries, asparagus, and endive	prolonged storage, delay of spoilage	1.0 – 2.5
Poultry, fish, seafood, pork, and beef	*Salmonella* control, prolonged shelf life	0.5 – 6.0
Pork	control of trichinosis	0.15 – 0.30
Herbs and spices	reduction of microbial contamination	1.0 – 2.0
Special hospital meals	sterilization for patients with specific needs	$\geq$25

Potential advantages of radiation-processed foods are prolonged shelf life without any deleterious sensory effects, greater retention of heat-sensitive nutrients (since the processing is carried out at low temperature), and increased safety of foods potentially associated with pathogens (such as poultry products, mussels, or other seafoods). Various foods have different sensitivities to radiolytic quality changes; some foods (e.g., herbs) can be irradiated at relatively high dose without any noticeable flavor defects, whereas dairy foods appear quite sensitive even to low-dose irradiation [51]. In general, the irradiation-related flavor and texture changes developed in foods are not completely understood, although animal as well as volunteer human feeding tests with even high-dose irradiated products did not indicate any toxic or other deleterious health effects. For certain experimental applications (e.g., radappertization of meat products) irradiation in the frozen state eliminates any undesirable flavor development and still achieves the desired lethal effect on the microorganisms. However, because irradiation is not effective for enzyme destruction, a mild heating step is often required as a radappertization pretreatment, not unlike blanching of vegetables before sterilization by canning.

Although the food irradiation process has been subjected to research and testing scrutiny unprecedented in the history of food technology, prospects for its widespread use are not bright,

especially with the current post-Chernobyl apprehensions. However, specialized applications will undoubtedly increase, including even the possibility for occasional undeclared low-dose treatments of international raw material shipments to decrease microbial contamination. Currently, no effective chemical tests are available to analyze foods for possible low-dose irradiation treatments as no detectable chemical changes occur.

6. Other Food Preservation Methods

Shelf-life extension of processed foods may be accomplished by several other means in addition to the principal technologies discussed in Chapters 2, 3, 4, 5. Refrigerated storage of fresh or processed food items is perhaps the most universal method used to retard food spoilage. Any process involving heat (pasteurization, blanching, cooking, baking, or frying) produces foods that can be kept for an extended time at proper storage conditions (temperature and packaging). Processes such as fermentation, pickling, smoking, and salting or sugaring of fish, sausage, dairy products, vegetables, fruits, etc. are used for preservation, as well as for diversification of the market selection. All these processes offer limited extension of shelf life by altering the environmental conditions conducive to microbial growth, as summarized in Table 11. Often several aspects of these techniques are combined in the processing – storage conditions applied (e.g., refrigerated storage of yogurt, highly acidic environment for packaging fermented fish, etc,). In many processed foods today, chemical preservatives are used as one of the complementary methods contributing to extension of shelf life.

6.1. Chemical Preservatives

Many naturally occurring components of food have antimicrobial properties. For example, cranberries contain ca. 5 % benzoic acid, which is commonly added as a preservative to jams and similar products. Propionic acid and its salts (calcium or sodium propionates) are effective antimold compounds added to bread, but produced naturally in Swiss cheese. Some lactic acid bacteria used in fermented dairy or meat products appear to produce metabolites with weak antibiotic properties. In many countries, selected antibiotics are allowed as food preservatives, particularly when applied to the surface of packaged foods such as cheese and fish.

The effectiveness of chemical preservatives is usually limited, especially in the low concentrations that are approved for various applications. However, their use enhances the effect of other food preservation methods such as refrigeration, packaging, or modified atmosphere storage. As one class of the large group of food additives, chemical preservatives are subject to stringent regulations (→ Foods, 3. Additives, Chap. 3.1.).

Table 11. Food preservation principles of various food processing operations

Process	Product example	Preservation principle
Fermentation	dairy products, some sausages, herring, and other fish	lowering of pH below tolerance limits of many microorganisms
Pickling	vegetables, fish, and mushrooms	pH control or increased salt content inhibits microbial growth
Smoking	meat products, cheese, and fish	some bacteriostatic effect of smoke; drying of the surface layers; sometimes heat effects
Salting	fish, mushrooms, and meat	lowering of a_w below microbial tolerance limits
Sugaring	candied fruits, jams, and jellies	lowering of a_w as in salting

6.2. Modified Atmosphere Packaging

The use of a modified gaseous environment to retard spoilage of fresh foods has emerged in the 1930 s mainly for transoceanic shipments of fresh meat to the United Kingdom. Later, development of controlled atmosphere storage systems to eliminate spoilage due to respiration of freshly harvested fruits extended the availability of certain seasonal produce (apples and pears) to many months. Inert gas (principally N_2) has been used for a long time in packaging of dried foods, such as whole milk or instant coffee, to prevent oxidative reactions during prolonged storage. The principles of these modified atmosphere storage techniques are well understood [52–54]. However, extension of the shelf life of perishable foods by modifying the atmosphere inside the retail containers (with or without concomitant use of refrigeration) is a relatively new concept that is still in the development stage [55–58].

Carbon Dioxide. The principal aspect of modified atmosphere packaging is the use of CO_2 as the primary component of the gaseous package atmosphere. The CO_2 gas acts as an inhibitor of microbial growth for many spoilage organisms, especially some of the psychrotrophs capable of growth at refrigerated temperatures, as well as many molds. Other bacteria, such as the lactic acid producing organisms, are not greatly inhibited by CO_2. Most of the food-related pathogens, including *Salmonella* species, *Staphylococcus aureus*, and *Yersinia enterocolitica*, do not multiply in a high CO_2 atmosphere; however some anaerobic pathogens, including *Clostridia*, are capable of growth with the CO_2 present. So far, modified atmosphere packaging appears suitable for shelf life extension of refrigerated products including meat, fish, and poultry, but it cannot be used as a substitute for refrigeration in foods where growth of the toxin-producing *Clostridium botulinum* can occur. For certain foods, especially bakery products, modified atmosphere packaging without refrigeration has been successful in industrial applications for shelf life extension to several weeks [59].

The exact mode of action of CO_2 in retarding microbial growth is not well understood. However, packaging atmospheres with up to 60 – 75 % CO_2 (with the balance consisting of O_2 and N_2 in varying proportions) have been used successfully to retard spoilage of many fruits, vegetables, meats, and other products. Because high CO_2 content in fresh meat packages can result in discoloration of the red meat pigment, incorporation of CO has been proposed (but not approved) as an experimental solution.

As a predecessor to controlled atmosphere packaging, the *vacuum packaging* of fresh or processed foods in gas-impermeable plastic containers has been used since the late 1950 s. This practice, in fact, establishes an atmosphere with up to 30 % CO_2 within the package because of metabolic and respiratory processes [56]. Fresh produce, such as cut-up green salads, packaged for institutional use in gas-impermeable plastic bags, creates its own modified atmosphere conditions by using the O_2 from the air in the package and producing CO_2 as a result of respiration [58].

7. Nutrient Retention in Processed Foods

The main objective of long-term food preservation is to extend the availability of seasonal food materials and thus, to maximize the use of their component nutrients. Apart from water, which is the most abundant component of most fresh food materials, the principal nutrients found in all foods are carbohydrates, proteins, and lipids. With the water, these macronutrients often account for more than 95 – 99 % of the food content. The remainder of the food components are the micronutrients, including minerals, vitamins, indigestible fiber ("bulk"), and other minor components such as organic acids, dissolved gases, etc.

The processing and storage techniques used to extend food shelf life affect the individual nutrients to a varying degree. Nutritional properties of the macronutrients, including the food energy values of the proteins and carbohydrates (16.5 kJ/g) and lipids (37 kJ/g) are not greatly affected by any of the long-term preservation techniques. The chemical reactions imparted by exposure to heat in drying or canning may cause denaturation or alter functionality of certain proteins (solubility, coagulation, and emulsifying properties), but this alone does not affect their nutritive value. However, exposure of systems containing both proteins and reducing sugars to severe heating may induce the nonenzymatic (Maillard) browning reaction, during which some of the essential amino acids from the proteins (principally lysine) become less available as nutrients for human body metabolism [60]. Similarly, oxidative reactions in lipids may result from processing (e.g., by irradiation) or prolonged storage without exclusion of oxygen (especially dried or frozen foods). These changes may lead to sensory defects in the products (e.g. rancidity or bitterness) but not necessarily to losses of nutritive value. Physical losses of these macronutrients may sometimes occur as an indirect result of the long-term preservation processes (e.g., loss of the water-soluble sarcoplasmic meat protein in the drip loss after thawing of frozen meats) or because of the various preparatory treatments needed for optimum quality retention (e.g., loss of water-soluble carbohydrates in water blanch-

ing of fruits and vegetables prior to drying, freezing, or canning). The same physical losses may also occur with water-soluble micronutrients, especially the minerals and vitamins, whereas fat-soluble vitamins are removed in preprocessing operations like skimming of milk before drying.

The main concern over nutrient losses in long-term food preservation usually concerns the *vitamins*, many of which are sensitive to heat, light, oxygen, and other factors encountered in the processing and storage conditions that are required. Continually expanding knowledge of the specific properties of the individual nutrients allows process technology improvements targeted specifically for maximization of nutrient retention. Technological advances (such as the UHT processing or food irradiation) also contribute to industrially realistic process modifications leading to less nutrient destruction. However, vitamin losses occur in prolonged food storage even under the best achievable conditions. In general, processing and storage effects on micronutrients in dried, frozen, or canned food must be expected as the trade-off for availability of macronutrients well beyond the growing season or the limited shelf life of fresh food materials.

8. References

General References

1. P. Jelen: *Introduction to Food Processing,* Reston – Prentice-Hall, Englewood Cliffs, N. J., 1985.
2. N. N. Potter: *Food Science,* 4th ed., AVI Publ., Westport, Conn., 1986.
3. N. W. Desrosier: *The Technology of Food Preservation,* AVI Publ., Westport, Conn., 1970.
4. T. A. Roberts, F. A. Skinner: *Food Microbiology – Advances and Prospects,* Academic Press, London 1983.
5. W. C. Frazier, D. C. Westhoff: *Food Microbiology,* McGraw-Hill, New York 1978.
6. G. J. Banwart: *Basic Food Microbiology,* AVI Publ., Westport, Conn., 1979.
7. T. D. Brock, L. M. Brock, D. W. Ward: *Basic Microbiology with Applications,* Prentice-Hall, Englewood Cliffs, N. J., 1986.
8. C. Zapsalis, R. A. Beck: *Food Chemistry and Nutritional Biochemistry,* J. Wiley & Sons, New York 1985.
9. O. R. Fennema: *Food Chemistry,* 2nd ed., Marcel Dekker, New York 1985.
10. H. D. Belitz, W. Grosch: *Food Chemistry (English translation by D. Hadziyev),* Springer Verlag, Berlin 1987.
11. J. A. Troller, J. H. Christian: *Water Activity and Food,* Academic Press, New York 1978.
12. L. B. Rockland, L. R. Beuchat: *Water Activity: Theory and Applications to Food,* Marcel Dekker, New York 1987.
13. D. Simatos, J. L. Multon: *Properties of Water in Foods,* Martinus Nijhoff Publ., Dodrecht, The Netherlands, 1985.

Specific References

14. T. P. Labuza, *Food Technol.* **34** (1980) no. 4, 36 – 41.
15. L. B. Rockland, S. K. Nishi, *Food Technol.* **34** (1980) no. 4, 42 – 51.
16. G. Hahn: "Evaporator Design," in D. McCarthy (ed.): *Concentration and Drying of Foods* , Elsevier Appl. Sci. Publ., London 1986, pp. 113 – 132.
17. W. H. J. M. Van Pelt, and W. J. Swinkels: "Recent Developments in Freeze Concentration," in M. Le Maguer, P. Jelen (eds.): *Food Engineering and Process Applications*, vol. 2, Elsevier Appl. Sci. Publ., London 1986, pp. 261 – 273.
18. R. K. Merle, R. F. Kozlik, S. F. Sapakie: "A New Freeze-Concentration Device," [17] pp. 275 – 280.
19. S. Sourirajan: "Reverse Osmosis: A New Field of Applied Chemistry and Chemical Engineering," in A. F. Turbak (ed.): *Synthetic Membranes*, vol. 1, ACS Symposium series no. 153, Amer. Chem. Soc., Washington, D.C., 1981, pp. 11 – 62.
20. R. C. Lawrence: *The Use of Ultrafiltration Technology in Cheesemaking,* Int. Dairy Fed., Brussels, Belgium, 1987, B-Doc 136.
21. R. C. Thomas, P. H. Westfall, Z. A. Louvieri, N. D. Ellis, *J. Food Sci.* **51** (1986) 559 – 563.
22. W. B. Van Arsdel, M. J. Copley, A. I. Morgan: *Food Dehydration,* **vol. 1,** AVI Publ. Co., Westport, Conn., 1973, pp. 82 – 200.
23. A. S. Mujumdar: *Handbook of Industrial Drying,* Marcel Dekker, New York 1987.
24. K. Masters: *Spray Drying Handbook,* J. Wiley & Sons, New York 1979.
25. I. Filkova, A. Mujumdar: "Industrial Spray Drying Systems", in [23] pp. 243 – 294.
26. B. Li-Shing Tat, P. Jelen, *Food Microstructure* **6** (1987) no. 2, 135 – 142.

27. H. A. Iglesias, J. Chirife: *Handbook of Food Isotherms,* Academic Press, New Jork 1982.
28. H. Salwin, *Food Technol.* **13** (1959) 594 – 595.
29. M. Caurie, *J. Food Sci.* **47** (1981) 332 – 333.
30. H. Weisser: "Influence of temperature on sorption isotherms," in [17] pp. 189 – 200.
31. D. P. Bone, E. L. Shannon, K. D. Ross: "The Lowering of Water Activity by Order of Mixing in Concentrated Solutions," in R. B. Duckworth (ed.): *Water Relations of Foods*, Acad. Press, New York 1975, pp. 613 – 626.
32. S. P. Leibo, P. Mazur, S. C. Jackowski, *Exp. Cell Res.* **89** (1974) 79 – 88.
33. R. L. Earle: "Freezing of Foods – An Overview," in [17] pp. 3 – 20.
34. O. R. Fennema, W. D. Powrie, E. H. Marth: *Low Temperature Preservation of Foods and Living Matter,* Marcel Dekker, New York 1973, pp. 207 – 222, 476 - 503.
35. C. K. Rha: "Thermal Properties of Food Materials," in C. K. Rha (ed.): *Theory, Determination and Control of Physical Properties of Food Materials* , D. Reidel Publ. Co., Dodrecht, The Netherlands, 1975, pp. 311 – 355.
36. R. L. Earle: *Unit Operations in Food Processing,* Pergamon Press, Oxford, U.K., 1983, pp. 187 – 188.
37. A. C. Cleland: "A Review on Methods for Predicting the Duration of the Freezing Process," in, [17] pp. 41 – 53.
38. R. A. Lawrie: "The Interaction of Freezing and Post-mortem Changes in Muscle," in J. Hawthorn, E. J. Rolfe (eds.): *Low Temperature Biology of Foodstuffs* , Pergamon Press, London 1968, pp. 73, 105, 359 – 380.
39. R. Lawrence, F. Consolacion, P. Jelen, *Food Technol.* **40** (1986) no. 3, 77 – 82.
40. P. Jelen, *Can. Inst. Food. Sci. Technol. J.* **16** (1983) 159 – 166.
41. C. R. Stumbo, K. S. Purohit, T. V. Ramokrishnan, D. A. Evans et al.: *Handbook of Lethality Guides for Low Acid Canned Foods*, vol. 1 and 2, CRC Press, Boca Raton, Fla., 1983.
42. National Canners' Assn., *Laboratory Manual for Food Canners and Processors,* **vol. 1,** AVI Publ., Westport, Conn., 1968, pp. 220 – 251.
43. M. A. Tung, H. S. Ramaswamy: "Steam/Air Media for Retort Pouch Processing," in M. Le Maguer, P. Jelen (eds.): *Food Engineering and Process Applications*, vol. 1, Elsevier Appl. Sci. Publ., London 1986, pp. 521 – 546.
44. R. L. Merson, T. K. Wolcott: "Recent Developments in Thermal Process Design," in [43] pp. 501 – 520.
45. "New Product Offers 'Freezer-freedom'," *Food Technol.* **41** (1987) no. 4, 102.
46. J. Mottar, M. Naudts, *Le Lait* **59** (1979) 476 (in French).
47. V. R. Carlson, *Food Technol.* **38** (1984) no. 12, 47 – 50.
48. New Monograph on UHT Milk, Int. Dairy Fed., Brussels, Belgium, 1981, Bulletin 133.
49. J. H. Moy, "Low Dose Irradiation of Foods – Prospects and Problems," in [43] pp 623 – 634.
50. E. T. O'Sullivan, A. L. Gunby: "Some Design Trade-Offs in a First Generation Food Irradiator," in [43] pp. 635 – 642.
51. D. A. E. Ehlermann: "Future Prospects for Radiation Processing of Food, in P. S. Elias, A. J. Cohen (eds.): *Recent Advances in Food Irradiation*, Elsevier Biomedical, Amsterdam 1983, pp. 331 – 354.
52. P. E. Brecht, *Food Technol.* **34** (1980) no. 3, 45 – 50.
53. A. A. Kader, *Food Technol.* **34** (1980) no. 3, 51 – 54.
54. S. T. Coulter. R. Jenness: "Dry Milk Products," in W. B. Van Arsdel, M. J. Copley, A. F. Morgan (eds.): *Food Dehydration*, vol. 2, AVI Publ. Co., Westport, Conn., 1973, pp. 337 – 338.
55. S. K. Wolfe, *Food Technol.* **34** (1980) no. 3, 55 – 58.
56. J. H. Silliker, S. K. Wolfe, *Food Technol.* **34** (1980) no. 3, 59 – 63.
57. W. Jamieson, *Food Technol.* **34** (1980) no. 3, 64 – 71.
58. C. B. Hintlian, J. H. Hotchkiss, *Food Technol.* **40** (1980) no. 12, 70 – 76.
59. J. P. Smith, B. Ooraikul, W. J. Koersen: "Novel Approach to Modified Atmosphere Packaging of Bakery Products," in I. D. Morton (ed.): *Cereals in a European Context*, VCH Publishers, Weinheim pp. 332 – 343.
60. A. Renz-Schauen: "Nutritional Value of Protein in Lactose-Hydrolyzed UHT Milk," in W. Kaufman (ed.): *Role of Milk Proteins in Human Nutrition*, Verlag Th. Mann KG, Gelsenkirchen 1983, pp. 325 – 327.

Foods, 3. Food Additives

Erich Lück, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany

Gert-Wolfhard von Rymon Lipinski, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany

1. Introduction . . . 2
1.1. History . . . 2
1.2. Safety Aspects . . . 3
2. Substances with Nutritive and Other Dietary Effects . . . 5
2.1. Vitamins and Provitamins . . . 5
2.2. Amino Acids . . . 5
2.3. Minerals and Trace Elements . . . 6
2.4. Bulking Agents . . . 6
3. Substances with Stabilizing Effects 6
3.1. Preservatives . . . 6
3.1.1. Mode of Action . . . 6
3.1.2. Methods of Evaluation . . . 7
3.1.3. Uses and Individual Substances . . . 7
3.2. Antioxidants . . . 8
3.2.1. Mode of Action . . . 8
3.2.2. Methods of Evaluation . . . 8
3.2.3. Uses and Individual Substances . . . 8
3.3. Synergists and Sequestrants . . . 9
3.4. Packaging Gases . . . 9
3.5. Stabilizers . . . 10
3.6. Emulsifiers . . . 10
3.6.1. Mode of Action . . . 10
3.6.2. Uses and Individual Substances . . . 10
3.7. Thickeners . . . 11
3.7.1. Mode of Action . . . 11
3.7.2. Methods of Evaluation . . . 11
3.7.3. Uses and Individual Substances . . . 11
3.8. Gelling Agents . . . 12
3.9. Foam Stabilizers . . . 12
3.10. Clouding Agents . . . 12
3.11. Humectants . . . 12
3.12. Anticaking Agents . . . 12
3.13. Coating Agents . . . 13
4. Substances with Sensory Effects (Organoleptic Substances) . . . 13
4.1. Coloring Agents . . . 13
4.1.1. Specifications . . . 13
4.1.2. Uses and Individual Substances . . . 13
4.2. Color Stabilizers . . . 14
4.3. Bleaching Agents . . . 14
4.4. Intense Sweeteners . . . 14
4.5. Nutritive Sweeteners . . . 15
4.6. Acidulants . . . 15
4.7. Substances with a Salty Taste . . . 15
4.8. Substances with a Bitter Taste . . . 15
4.9. Substances with an Alkaline Taste . 15
4.10. Flavor Enhancers . . . 15
4.11. Spices and Flavorings . . . 16
4.12. Chewing Gum Bases . . . 16
5. Processing Aids . . . 16
5.1. Extractants . . . 17
5.2. Clarifying Agents . . . 17
5.3. Filter Aids . . . 17
5.4. Propellants . . . 18
5.5. Cooling Agents and Cryogens . . . 18
5.6. Mold-Release Agents . . . 18
5.7. Antifoaming Agents . . . 18
5.8. Acidity Regulators . . . 18
5.9. Emulsifying Salts . . . 18
5.10. Dough Conditioners and Flour Improvers . . . 18
5.11. Leavening Agents . . . 19
5.12. Enzymes . . . 19
5.13. Microbial Cultures . . . 20
6. Legal Aspects . . . 20
6.1. Food Law Authorizations . . . 20
6.2. Definitions . . . 20
6.2.1. Codex Alimentarius . . . 20
6.2.2. United States of America . . . 20
6.2.3. Japan . . . 21
6.2.4. Federal Republic of Germany . . . 21
6.2.5. Regulations and Specification . . . 21
7. References . . . 22

1. Introduction

Food additives are substances that are intentionally added to food, without themselves being considered food in the ordinary sense of the term. The term "food additive" is defined differently by the food laws of individual countries. As examples, the definitions used in the Codex Alimentarius, in the United States, in Japan, and in the Federal Republic of Germany are given in Chapter 6. Food additives are indispensable for the production and processing of many foods. Some are essential for the economic production and distribution of foods. Together with contin-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_561

ued improvements in food technology, additives ensure the general availability of high-quality food with a satisfactory shelf life. The four basic reasons for using additives are

1) to influence the nutritive value of food,
2) to improve the stability of food,
3) to affect the sensory properties of food, and
4) to make certain technological processes possible.

Food additives are sometimes used to produce food more cheaply. However, they may not be used to deceive the consumer, for example, about the character or freshness of food. Furthermore, additives must not adversely affect food quality.

The properties and production methods of some additives are described in other articles; the reader is advised to consult the cumulative index for specific products. The following discussion is, therefore, confined to more general and universal aspects.

1.1. History

Two preservatives, smoke and salt, are probably the oldest food additives. They were also used to give food a good taste. Like spices, they have been used since prehistoric times to improve the taste of food from plant or animal sources. These foods were either eaten raw or prepared over an open fire. Until the 18th century, the use of additives remained low. However, with the beginning of industrialization in the 19th century, an increasing number of people moved to towns. This meant that they had less time available to provide and process their food. More and more food was produced either by specific craftsmen or industrially. Additives became increasingly necessary; especially those that protected food against spoilage. Some food additives introduced in the 19th century are still being used today, e.g., baking powder, benzoic acid, and saccharin. However, others in use at that time do not comply with current toxicological requirements and have thus disappeared from food processing.

In the 20th century, the industrial production of food developed further and the importance of additives increased. Thus, the invention of processed cheese (→ Cheese, Processed Cheese, and Whey, Chap. 6.) in 1913 would have been impossible without the development of emulsifying salts, such as citrates, phosphates, or similar substances. During this time emulsifiers were developed which improved and simplified the production of margarine. Additives used in certain food for physiological reasons have also become important, e.g., trace elements, vitamins, and essential amino acids. More recently, the importance of aroma and flavor enhancers, as well as flavoring compounds, has increased substantially.

Hygienic and legal controls have also been tightened considerably in this century. One example of this is in coloring agents. Many substances used in the first half of the 20th century have not survived modern toxicological requirements.

Despite all the advances, unresolved problems still remain in the area of food additives, e.g., finding substitutes for two specific types of preservative, nitrite and sulfite, which are considered undesirable for various reasons. The search also continues for substances or processes that can be used to reduce the undesirably high levels of sugar and common salt in some foods, without impairing their quality. The financial risks involved in developing new food additives are high because development of a new food additive is now almost as expensive as developing a new drug. Even when additives have been evaluated successfully and necessary authorization under food law has been obtained, the cost of development can be recovered only if the additive can be used widely.

1.2. Safety Aspects

Some people still harbor an aversion to food additives. This can be explained by earlier, uncontrolled use of chemicals in food processing. Today, food additives are authorized only if no harmful effects of any kind can be shown after extensive toxicological testing. Many additives have undergone more rigorous testing than some foods or food components.

To date, no internationally binding standards exist for the toxicological testing of additives. The most widely accepted guidelines are those laid down by expert committees of the WHO

[1]. The U.S. regulations [2] are also observed widely.

Additives are tested for their acute, subchronic, and chronic toxicity, as well as for carcinogenicity, mutagenicity, teratogenicity, and biochemical behavior. Before testing begins, the substance must be chemically and physically homogeneous and pure. As a matter of principle, a food additive should not be pharmacologically active at the concentration at which it is used. For obvious reasons, the testing of additives is carried out largely on animals. The tendency away from animal testing is increasing, but at present, this is only possible in a few areas.

Acute toxicity, expressed as LD_{50}, and subchronic toxicity, established by the so-called 90-day test, are only crude measures of the toxicological properties of a substance. The acute toxicity of food additives should always be as low as possible to prevent poisoning even in the case of accidental overdosing or misuse.

Chronic toxicity is more important. The specifications for food additives are far more strict in this respect than those for drugs, because additives may be consumed continuously for a very long period of time. Also, in contrast to drugs, side effects are not considered an acceptable risk with food additives.

Food additives must not be mutagenic or teratogenic. Because they are absorbed exclusively via the gastrointestinal tract, carcinogenicity is tested mainly by feeding.

Biochemical tests are performed to determine the extent to which a food additive is absorbed by the organism, the factors affecting its absorption, its distribution within the body, and the manner in which it is excreted. Four basic possibilities exist:

1) A substance is excreted quickly and unchanged. Examples of this are the sweeteners saccharin [*128-44-9*] and acesulfame-K [*55589-62-3*].
2) A substance is quickly metabolized within the body and excreted in this metabolized form. An example is the preservative benzoic acid [*65-85-0*], which is converted in the body to hippuric acid [*495-69-2*], by reaction with glycine [*56-40-6*], and excreted in this form.
3) A substance is utilized quickly as a nutrient. Some examples are the preservative sorbic acid [*110-44-1*] and the sweetener aspartame [*22839-47-0*].
4) A substance is excreted more slowly than it is absorbed and thus accumulates in the body. An example of this is boric acid, used as a preservative in the past. Such substances are considered undesirable and are no longer used as food additives.

Table 1. Acceptable daily intake (ADI) values of some food additives

Additive	CAS registry no.	ADI, mg/kg per day
Acesulfame-K	[*55589-62-3*]	0 – 9
Alginates		0 – 50
Amaranth	[*915-67-3*]	0 – 0.75 (temporary)
Ascorbic acid	[*50-81-7*]	0 – 15
Aspartame	[*22839-47-0*]	0 – 40
Azo Rubine	[*3567-69-9*]	0 – 4
Benzoic acid	[*65-85-0*]	0 – 5
Butylated hydroxyanisole	[*25013-16-5*]	0 – 0.5
Butylated hydroxytoluene	[*128-37-0*]	0 – 0.5
Caramel (ammonia process)		0 – 100 (temporary)
Cellulose ethers		0 – 25
Cyclamate	[*100-88-9*]	0 – 11
Erythrosine	[*16423-68-0*]	0 – 2.5
Ethyl maltol	[*4940-11-8*]	0 – 2
Fumaric acid	[*110-17-8*]	0 – 6
Gallates		0 – 0.2
Nitrates		0 – 5
Nitrites		0 – 0.2 (temporary)
Phosphates		0 – 70
Quinoline Yellow	[*8004-92-0*]	0 – 0.5 (temporary)
Saccharin	[*128-44-9*]	0 – 2.5 (temporary)
Sorbic acid	[*110-44-1*]	0 – 25

Animal trials are conducted to determine the concentration to which an additive remains nontoxic. Expert WHO committees use this value to define the *acceptable daily intake* (ADI) by reducing it by a factor of 100 (except in special cases). This takes into consideration the following facts: the food intake of some animals, relative to their body weight, is greater than that of humans; that some people (e.g., children, the sick, and the elderly) show special metabolism; and the ADI applies over the entire life-span.

The ADI is related to kilograms of body weight and is the amount of additive in milligrams that can be consumed safely per day (this

includes the reduction by a factor of 100). The ADI values of some food additives are given in Table 1. The actual consumption of food additives is generally far lower than the ADI. For preservatives in the Federal Republic of Germany, consumption is about 1 – 10 % of the ADI. Only sulfur dioxide is consumed in larger amounts [3].

2. Substances with Nutritive and Other Dietary Effects

The human organism requires regular supplies of about 40 – 50 nutrients if all bodily functions are to proceed smoothly. If even one of them is missing or present in insufficient amounts, deficiency symptoms result. A particularly well-known example of this is scurvy, i.e., vitamin C deficiency.

All nutrients and nutritive supplements needed by the body are normally present in sufficient amounts in food. However, in exceptional cases or with unbalanced diets, the amount of these substances may be insufficient. Some can be destroyed or lost during processing or storage. In such cases, these nutrients can be added in a controlled manner.

Substances with nutritive and dietetic functions are added to food to increase its nutritive quality or, in special cases, to reduce it in a controlled way. Some of these substances may also ameliorate metabolic disorders with a dietary or physiological cause and correct various deficiencies.

Incorporating these substances homogeneously is sometimes difficult because they must be added at very low concentrations to relatively large quantities of food. In such cases, special premixes are used to allow easier incorporation into the appropriate food. Some substances with nutritive and dietary effects are particularly susceptible to degradation, especially by oxidation. Therefore, they must be stabilized. In practice, this is achieved by the synthesis of especially stable molecules which will remain physiologically and dietetically effective, by the use of stabilizing additives, or by encapsulation with suitable carrier materials (e.g., microencapsulation).

2.1. Vitamins and Provitamins

(→ Vitamins)

Vitamins are used as food additives for dietary reasons, when sufficient vitamins are not supplied by the food itself or when losses have occurred during processing. Natural variation in the vitamin content of food can also be adjusted in this way. Vitamins may also be added when the body is unable to absorb sufficient amounts of the vitamins contained in food [4–6]. The controlled addition of vitamins to food is called *vitaminization,* and the addition to compensate for processing losses is *revitaminization. Enrichment* or *fortification* means the addition of vitamins beyond the natural content.

Standardization means adjustment of natural variations. Oversupply is harmful only with the fat-soluble vitamins A and D, because it can lead to so-called hypervitaminoses.

Examples of *revitaminization* are the addition of vitamins A and D to skimmed milk and the addition of B vitamins to cereal. Examples of *standardization* or *enrichment* are the addition of vitamin C to beverages or fruit products and the addition of vitamin D to milk. Vitamins are added to some foods that naturally contain few or no vitamins because they constitute a particularly suitable carrier for a particular vitamin. An example of this is the addition of fat-soluble vitamins A and D to margarine (→ Margarine and Shortenings).

Vitamins and provitamins can also be added to foods for technological reasons. For example, ascorbic acid [*50-81-7*] and tocopherols have an antioxidative effect (see Section 3.2), and some carotenoids with vitamin functions are important food colorants (see Section 4.1).

2.2. Amino Acids (→ Amino Acids)

Some amino acids are added to foods for technological reasons, e.g., flavoring or flavor enhancers (Chap. 4) or baking aids (Section 5.11). Their use as food additives for *dietetic reasons* is restricted to the essential amino acids, i.e., those which the body cannot synthesize. These are added to food for the same reason as vitamins. During food processing, lysine [*56-87-1*], in particular, can react with carbohydrates by the Maillard reaction to form compounds from which it

is no longer physiologically available. Deficiencies in low-protein foods which also have a low lysine ratio (e.g., some cereals) can be adjusted by enrichment with lysine. Mixtures of amino acids are sometimes administered with food in special diets, especially low-fiber diets. The importance of amino acids as human food additives is far less than in animal feed.

2.3. Minerals and Trace Elements

Sufficient amounts of minerals and trace elements are generally present in the average diet. In addition, they are not damaged by industrial processes so they are added to foods only in special cases. Iron salts are of some importance because iron is sometimes not absorbed in sufficient quantity from food. The same is true of calcium, magnesium, copper, and zinc salts [7]. The importance of magnesium for the cardiovascular system encouraged the addition of magnesium to food. In areas where the iodine content of drinking water is low, a sufficient iodine intake can be achieved by addition of iodides and iodates to common salt. A difference of opinion exists about the need for fluorine in human nutrition. In principle, fluorine can be added to common salt or to drinking water in the form of fluorides.

2.4. Bulking Agents

Bulking agents are inert substances with little or no nutritive value. They are used in reduced-calorie foods to generate a feeling of satiety.

The cheapest bulking agents are water and air. If these are to be incorporated into foods, the addition of emulsifiers (Section 3.6) or thickening agents (Section 3.7) is usually necessary. Other bulking agents include crystalline cellulose, polydextrose (a polycondensate of sucrose with sorbitol) and the disaccharide alcohol isomalt [*64519-82-0*], which have low calorie values.

3. Substances with Stabilizing Effects

Products with increased shelf life are in demand for many reasons: the increasing extent to which food is produced industrially, higher food-quality specifications, contemporary shopping and consumption patterns, as well as certain health reasons. Therefore, prevention of food spoilage–whether due to microbiological, chemical, biochemical, or physical factors–is a necessity. Improved shelf life can often be achieved by physical methods such as heating. However, some of these methods change the structure or taste of food, which is the reason they are sometimes supplemented or reinforced by the use of additives.

3.1. Preservatives

Preservatives are compounds that delay, or ideally prevent, microbiological spoilage of food [8], [9]. Not only do these act against visible spoilage of food by yeast, molds, and bacteria, they also prevent the formation of toxins, especially those produced by bacteria and molds. Today, this effect is more important than the protection of food against economic losses. For example, the antimicrobial agent nitrite destroys *Clostridia* which excrete botulinum toxin, and sorbic acid [*110-44-1*] acts against molds that can form carcinogenic aflatoxins.

The use of preservatives, also referred to as chemical preservation of food, does not compete with physical methods of preservation (→ Foods, 2. Technology), e.g., sterilization and chilling; rather, it complements them. Sometimes the combined use of preservatives and physical measures makes possible a reduction of either the intensity of physical measures or the concentration of preservatives. A typical example is the production of intermediate-moisture, partially dried, plums protected by sorbic acid from mold development. Their quality is superior to that of fully dried products.

The combination of physical and chemical preservation processes of any kind is referred to as the "hurdle concept" [10]. The addition of preservatives also protects food against subsequent spoilage once packages are opened or if they are not hermetically sealed. This is independent of, and in addition to, any physical preservation measures that may be used.

3.1.1. Mode of Action

In practice, a distinction is often drawn between bacteriostatic and bactericidal (antibacterial effect) or between fungistatic and fungicidal (antiyeast and antimold effect), i.e., between germ-inhibiting and germ-killing effects. Strictly speaking, this distinction is not justified [11]. Preservatives present in low-germ foods at sufficiently high dosages kill the microorganisms present more or less rapidly. If no reinfection occurs, this final state is achieved in a period of several days to several weeks. This is the crucial difference between preservatives and disinfectants (→ Disinfectants). The latter must destroy microorganisms in a very short period of time. The course per unit time of microbial death due to preservatives corresponds to a first-order reaction [12]:

$$k=\frac{1}{t}\cdot\ln\frac{Z_0}{Z_t} \text{ or } Z_t=Z_0\cdot e^{-\mathrm{kt}}$$

where k is the death constant, t is time, Z_0 is the number of viable cells at the start of the preservative action, and Z_t is the number of viable cells after time t.

The *antimicrobial effect* of preservatives can essentially be explained by the following phenomena:

1) interference with the genetic microstructure of the protoplast,
2) inhibition of protein synthesis,
3) inhibition of enzyme activity,
4) damage to the cell membrane, and
5) damage to the cell wall.

Substances that lower the *water activity* (→ Foods, 2. Technology, Chap. 1.4.) of a substrate and thereby impede microbial growth work differently. Packages, coatings, oil, and some packaging gases block access of oxygen to foods and thus inhibit the growth of aerobic microorganisms. As a rule, none of these substances kill microorganisms but, at best, damage them and stop them from multiplying. When the influence of these inhibitors is removed, the contaminating germs start multiplying again, and the product eventually spoils. Finally, preservatives such as acetic and lactic acid lower the pH of food and thereby inhibit the growth of some microorganisms, especially bacteria.

No preservative is equally effective against all microorganisms. Most substances used today are more effective against yeasts and molds than against bacteria. Unlike antibiotics, preservatives do not lead to the development of organisms resistant to them [13], [14].

Occasionally, two or more preservatives are used in tandem. In the past, such combinations were widely used commercially. However, their effectiveness has been overrated. Synergism of sufficient practical importance is unknown for preservatives. However, a combination of preservatives can sometimes broaden the spectrum of activity, e.g., the combined use of substances that are individually effective against molds or bacteria.

3.1.2. Methods of Evaluation

The results of tests based on nutrient media, commonly used in microbiology, are difficult to transfer to a food. This is because substrate characteristics influence the effect of preservatives in a variety of ways. For example, the pH and the redox potential of a food, the presence or absence of fats, and its water activity are of great practical importance for the effectiveness of a preservative. Some preservatives (e.g., nitrite) react with food components.

In practice, the preservative is added in different concentrations to batches of the food. These are then packaged commercially, and their shelf life is monitored under realistic storage conditions. Here, the microbiological status of the food is observed as are any changes in its sensory qualities. A common problem is inoculation with appropriate microorganisms. Not only is this a question of inoculating with suitable types of microorganisms, but the number of organisms must also represent the actual conditions found in practice. The role of preservatives is not to combat very high numbers of germs but rather to attack the relatively small population that can be expected to be present in food produced under hygienic conditions.

3.1.3. Uses and Individual Substances

Preservatives are used in cheese, meat products, fruit-based products, beverages, and baked goods. Depending on the type of food and the spoilage expected, different preservatives are used. For example, meat products are preserved mainly with common salt and nitrite; fruit-based products, with sulfur dioxide; beverages, with sorbic acid; and baked goods, with sorbic acid or propionates.

The most important preservatives commonly used in foods are summarized in Table 2. In addition to these, the following have some importance as preservatives although they are used partially for other reasons, such as flavoring: sodium chloride, nitrogen, silver, ozone, hydrogen peroxide, chlorine, ethanol, sucrose [*57-50-1*], and smoke.

3.2. Antioxidants (→ Antioxidants)

Antioxidants delay oxidative spoilage of food. They interfere with the early stages of oxidative and autoxidative processes to prevent formation of unwanted reaction products. Broadly speaking, antioxidants are substances that react quickly with oxygen and thereby quench it.

During food processing and storage, autoxidation is particularly troublesome with fats that contain unsaturated fatty acids, which can form highly odoriferous, unsaturated aldehydes. These cause the fats to smell and taste unpleasant; i.e., they become rancid. The formation of polymeric reaction products is also undesirable.

3.2.1. Mode of Action

The autoxidation of fats proceeds via a radical mechanism. Hydroperoxides are formed as intermediates, which then decompose to form radicals and can thereby initiate further reactions. The chain reactions can be terminated by scavengers that react with radicals to form inert compounds [15–17]. Radical scavengers constitute, therefore, the most important group of antioxidants.

If a mixture of two or more antioxidants is used, a stronger effect can be achieved than with a single antioxidant. To optimize this synergism, the necessary ratio of the two antioxidants must be established experimentally. In addition, some antioxidants inhibit lipoxygenase and, by this side effect, can prevent the enzymatic oxidation of fats.

Antioxidants have the desired effect only within a certain concentration range. At excessive concentration, they may even act as prooxidants. If the aim is to protect fats in the finished food product after it has been heated, antioxidants must be used which maintain their effectiveness after such processing. For example, certain phenolic antioxidants show this carry through effect [18].

Antioxidants, in the widest sense, scavenge oxygen and thereby prevent the occurrence of enzymatic reactions that require oxygen. Sometimes they also inhibit the enzymes involved such as polyphenol oxidases.

3.2.2. Methods of Evaluation

The most important means of evaluating antioxidants used for fats are the Swift test, the Barcroft–Warburg test, and the thiobarbituric acid test. In the *Swift test,* samples to which antioxidants have been added are aerated at 97 °C under defined conditions. The peroxide content is measured and compared to that of a sample without antioxidants. In the *Barcroft–Warburg test,* a sample is held under oxygen in a closed system and the oxygen uptake is monitored manometrically. In the *thiobarbituric acid test,* the presence of aldehydes is measured. Under suitable test conditions, this test is said to correlate highly with sensory properties of the fat [17].

3.2.3. Uses and Individual Substances

Antioxidants are added mainly to fats or fat emulsions and to foods containing fats in a more or less fine distribution. Antioxidants can also prevent unwanted oxidative changes in essential oils, aroma compounds, and chewing gum. In practice, antioxidants are combined with synergists to increase their effectiveness. Fruit- and potato-based products, in particular, can be protected against discoloration through addition of ascorbic acid [*50-81-7*], sulfur dioxide, and sulfites. Sulfur dioxide or sulfites alone, and in combination with ascorbic acid, considerably delay

Table 2. Concentrations of preservatives found in foods

	Preservative	Concentration, ppm
E∗ 200 – E 203	sorbic acid [*110-44-1*] and sorbates	500 – 2000
E 201 – E 213	benzoic acid [*65-85-0*] and benzoates	500 – 1000
E 214 – E 219	4-hydroxybenzoic acid esters (parabens) and their sodium derivatives	500 – 1000
E 220 – E 227	sulfur dioxide and sulfites	200 – 2000
E 230	biphenyl (diphenyl) [*92-52-4*]	50 – 70
E 231 – E 232	-phenylphenol [*90-43-7*]	10 – 12
E 233	thiabendazole [*148-79-8*]	3 – 6
E 236 – E 238	formic acid [*64-18-6*] and formates	3000 – 4000
E 239	hexamethylenetetramine [*100-97-0*]	20 – 200
E 249 – E 250	nitrites	50 – 100
E 251 – E 252	nitrates	200 – 600
E 260 – E 263	acetic acid [*64-19-7*] and acetates	5000 – 30 000
E 270	lactic acid [*50-21-5*]	5000 – 10 000
E 280 – E 283	propionic acid [*79-09-4*] and propionates	2000 – 3000
E 290	carbon dioxide	

∗ Serial number for use on labels

undesirable oxidation processes in beverages, e.g., the madeirization of wine [19]. However, here sulfur dioxide and sulfites are used because of their preservative effect and for other reasons as well, e.g., to scavenge fermentation byproducts.

The most important antioxidants commonly used in foods are listed below. In addition, the following substances are important antioxidants in some countries: 3,3′-thiodipropionic acid [*111-17-1*], dilauryl thiodipropionate [*123-28-4*], trihydroxybutyrophenone [*52262-23-4*] (THBP), ethoxyquin [*91-53-2*], and *tert*-butylhydroquinone [*1948-33-0*] (TBHQ).

E 220 – E 227	sulfur dioxide and sulfites
E 300 – E 304	ascorbic acid [*50-81-7*] and ascorbates isoascorbic acid [*89-65-6*] (erythorbic acid)
E 306 – E 309	tocopherols
E 310 – E 312	gallates
E 320	butylated hydroxyanisole [*25013-16-5*] (BHA)
E 321	butylated hydroxytoluene [*128-37-0*] (BHT)

3.3. Synergists and Sequestrants

Sequestrants are compounds that form complexes with metal ions and thereby convert them into an inactive form. Of special importance is the inactivation of traces of heavy metals, such as copper and iron, which accelerate oxidative changes catalytically.

Sequestrants also inactivate certain enzymes by binding the metal ions needed to activate them and thereby support the microbiological stabilization of foods. Sequestrants can sometimes suppress discoloration caused by interaction between metals and amines or mercapto compounds in delicatessen products and fish or shellfish preserves. Undesirable cloudiness in beverages, such as that caused in wine by iron and copper, is avoided by using sequestrants to convert the appropriate metal ions into ineffective complexes.

Sequestrants used in combination with antioxidants are called *synergists* because their reaction with metallic ions indirectly supports the antioxidative effect. Compounds that regenerate spent antioxidants are also called synergists.

The most important synergists and sequestrants commonly used in foods are summarized below. In addition, the following substances are of some importance as synergists and sequestrants: phytic acid [*83-86-3*] and its salts, calcium gluconate [*299-28-5*], calcium polyphosphate, and potassium ferrocyanide [*13943-58-3*].

E 270	lactic acid [*50-21-5*]
E 322	lecithin [*8002-43-5*]
E 324 – E 327	lactates
E 330 – E 333	citric acid [*77-92-9*] and citrates
E 334 – E 337	tartaric acid [*147-71-7*] and tartrates
E 338 – E 341	phosphoric acid [*7664-38-2*] and phosphates
E 342	calcium and sodium salts of ethylenediaminetetraacetic acid [*60-00-4*] (EDTA)

3.4. Packaging Gases

Packaging gases are used essentially to exclude oxygen from stored food [20] and thus primarily protect against oxidative changes. Furthermore,

they indirectly act as preservatives by displacing the oxygen which is vital for strictly aerobic microorganisms.

The main packaging gases used in commercial food processing are nitrogen, carbon dioxide, and mixtures of these. Besides displacing oxygen, carbon dioxide at higher concentration specifically inhibits some microorganisms, especially molds.

3.5. Stabilizers

Stabilizers are substances that protect from, or counteract, changes in the structure of food which, in a wider sense, can be considered spoilage. Stabilizers include emulsifiers, thickeners and gelling agents, foam stabilizers, humectants, anticaking agents, and coating agents.

3.6. Emulsifiers (→ Emulsions)

Emulsifiers are substances that facilitate, or may be essential to, the production of emulsions. Furthermore, they improve the structure of foods that naturally contain small amounts of fat, e.g., dough for baked goods. Emulsifiers improve and stabilize the consistency of foods and sometimes their viscosity, texture, and mouthfeel as well. They improve the shelf life of some food such as baked goods.

Thickeners (see Section 3.7) often stabilize emulsions by greatly increasing the viscosity of the aqueous phase. The formation and stabilization of emulsions in foods that contain protein and fat can also be achieved by addition of phosphates. The properties and effects of phosphates depend on the degree of condensation. Phosphates cause hydration and swelling of proteins, thereby increasing the viscosity of the system.

3.6.1. Mode of Action

A stable emulsion can be achieved only with emulsifiers or emulsifying systems that are specifically tailored for the components involved. Not every emulsifier is suitable for every food. The HLB system can be used to help select a suitable non-ionic emulsifier (→ Emulsions, Chap. 9.2.) [21–23]. Fat-soluble, hydrophobic emulsifiers have HLB values between 0 and 9; water-soluble compounds have HLB values between 11 and 20. Ionic emulsifiers do not fit into this system. For water-in-oil emulsions (W/O emulsions), emulsifiers or emulsifier mixtures with HLB values of 3 – 8 are suitable; for oil-in-water emulsions (O/W emulsions), HLB values between 8 and 12 are good. Compounds with HLB values of 12 to 18 can be used as solution promoters (solubilizers).

Emulsifier mixtures are preferred when the object is to achieve a particularly good emulsification. The HLB values of such mixtures are obtained by addition from the proportions of individual emulsifiers contained in the mixture. A suitable emulsifier is not the only requirement for stable food emulsions; the emulsification conditions and the components of the phases to be emulsified also strongly influence the quality of an emulsion.

3.6.2. Uses and Individual Substances

Fat spreads with high water content (e.g., margarine, a water-in-oil emulsion), cannot be produced unless a suitable emulsifier is used. In fat emulsions for frying, emulsifiers prevent spattering, which is why they are referred to as antispattering agents. For mayonnaise and salad dressings (oil-in-water emulsions), the emulsifying effect of the lecithin contained in egg yolk is often exploited. When these products are made with a reduced egg content or without any egg yolk, other emulsifiers must be added.

Particularly effective emulsifiers are needed for ice cream manufacture. Ice cream is the foam of an oil-in-water emulsion, in which the aqueous phase is a colloidal solution of milk proteins and, at the same time, a true solution of various sugars and salts as well as a suspension of ice crystals. Here, emulsifiers have a major effect on stability and melting behavior.

In yeast-leavened baked goods, certain emulsifiers such as monoglycerides delay staling through clathrate formation with starch. This inhibits the retrogradation of the starch and delays softening of the crust. Interaction between emulsifiers and proteins in the dough increases the elasticity of the dough and thereby improves its gas-retaining capacity. The finished baked goods

have a more uniform pore size and distribution and a larger volume [24], [25].

Shortenings are special baking fats with added emulsifiers. In pastry, emulsifiers cause the formation of foamy oil-in-water emulsions and thereby affect its pore structure.

In chocolate (→ Chocolate) manufacture, emulsifiers lower the viscosity of the chocolate mass and delay bloom development in the finished chocolate. Instant foods containing fats require the addition of emulsifiers. This leads to a particularly fine distribution of the fat and good flowability of the finished product.

In meat products, especially sausages, emulsifiers improve the distribution of meat, fat, and water (→ Foods, 2. Technology). In this context, the stability of emulsions on heating is also important. The sausage binders used to make sausage-type products are emulsifiers in a wider sense; they are used in the manufacture of frankfurter type sausages. They influence the swelling properties of the protein and thus improve the water- and juice-retention capacity of the meat so that it can be processed like freshly slaughtered meat. Under their influence, the denatured muscle protein coagulates uniformly during heat treatment, so that the final product can be sliced. Sausage binders are mixtures of common salt with salts of acetic, lactic, tartaric, or citric acid, and especially, with certain phosphates [26], [27].

The most important emulsifiers commonly used in foods are summarized below:

E 322	lecithins [*8002-43-5*]
E 339 – E 341	phosphates
E 450 a – c	diphosphates, triphosphates, and polyphosphates
E 470	sodium, potassium, and calcium salts of fatty acids
E 471	mono- and diglycerides of fatty acids
E 472 a – f	esters of mono- and diglycerides of fatty acids
E 473	sucrose esters of fatty acids
E 474	sucroglycerides
E 475	polyglycerol esters of fatty acids
E 477	propane-1,2-diol esters of fatty acids
E 481 – E 482	sodium and calcium stearoyl-2-lactylate
E 483	stearyl tartrate

3.7. Thickeners

Thickeners are hydrocolloids that are soluble in water or can be readily hydrated or dispersed. They form viscous solutions. The thickeners commonly used in food are polysaccharides (→ Polysaccharides) of plant, microbial, or semisynthetic origin, e.g., milled seeds, extracts of sea algae, and plant extracts or exudates. Thickeners are used to maintain or improve the viscosity and flow characteristics of food [28–32].

3.7.1. Mode of Action

In contrast to emulsifiers, thickeners are not fat-soluble. They interact with the water in food through their fibrous or cross-linked structure and their many polar groups, especially hydroxyl groups. Water molecules, which are polar, orient themselves around the polar groups of the thickeners. Formation of a hydration layer of this kind, often accompanied by unfolding of the molecules, limits the mobility of the water so that the viscosity of the system increases.

Thickeners compete for the available water with other food components that can also be readily hydrated. They therefore influence the swelling of other food components, such as proteins. Different thickeners can associate among themselves or with other high molecular mass food components to give a particularly strong increase in viscosity.

3.7.2. Methods of Evaluation

Viscosity measurements are used to characterize and evaluate swollen or dissolved thickeners. At commonly used concentrations, solutions and suspensions of hydrocolloids exhibit non-Newtonian, specifically pseudoplastic, flow behavior. However, the measured viscosities are material constants only for liquids with Newtonian flow behavior. As a practical rule, relative measurements within a particular class of substances are sufficient. For this reason, measuring apparatus and experimental conditions must be standardized.

3.7.3. Uses and Individual Substances

Thickeners are generally added under defined conditions to water or to a food component with a high water content. For successful application, the method by which they are incorporated, the

water temperature, and the stirrer speed are very important.

In fermented milk products such as yogurt, the addition of thickeners leads to a very finely flocculated precipitation of the casein and thus a particularly uniform structure. In the manufacture of ice cream, thickeners prevent the formation of large ice and sugar crystals and thereby lead to uniform melting behavior. In addition, the ice cream mixture can be whipped more readily into a foam because of stabilization of the finely dispersed air. In fat emulsions, thickeners can exercise an indirect emulsifying effect or improve the effect of emulsifiers, which leads to improved temperature stability. Thickeners also serve as clouding (Section 3.10) and bulking agents for reduced-calorie food (see Section 2.4).

The most important thickeners commonly used in foods are listed below:

E 400 – E 405	alginic acid [*9005-32-7*], alginates, propane-1,2-diol alginate [*12698-40-7*]
E 406	agar [*9002-18-0*]
E 407	carrageenan (Irish moss extract) [*9000-07-1*]
E 410	carob gum (locust bean gum) [*9000-40-2*]
E 412	guar gum [*9000-30-0*]
E 413	tragacanth gum [*9000-65-1*]
E 414	acacia gum (gum arabic) [*9000-01-5*]
E 415	xanthan gum [*11138-66-2*]
E 440 a – b	pectin [*9000-69-5*] and amidated pectin
E 461 – E 466	cellulose ethers, starch, starch esters, starch ethers, modified starches

3.8. Gelling Agents

Gelling agents combine with water to form pseudogels or gels. They maintain or improve the structure, consistency, or elasticity of a food. Gelling agents facilitate development of the gel structure desired in some food. They are used in confectionery, desserts, jams, preserves, and coating for fish and meat products. They also increase the freeze – thaw stability of frozen food. The action of some gelling agents depends on the calcium ion content and pH of the food. The most important gelling agents used in food are gelatin [*9000-70-8*], alginates, agar [*9002-18-0*], carrageenan [*9000-07-1*], and pectin [*9000-69-5*].

3.9. Foam Stabilizers

Foam stabilizers are used particularly in the baking and confectionery industries. They impart greater stability to foamy preparations. In sugar-containing preparations, hydrocolloids like methyl cellulose [*9004-67-5*] and tragacanth [*9000-65-1*] can be used; in foamy water-in-fat emulsions, emulsifiers may be used instead. Phosphates stabilize whipped protein foam. The foam of soft drinks and beer can also be stabilized by certain hydrocolloids.

3.10. Clouding Agents

Clouding agents are thickeners that increase the viscosity of a beverage and, thereby, retard or prevent the deposition of finely suspended particles. Certain emulsifiers are also used to prevent the separation of essential oils in beverages.

3.11. Humectants

Humectants are added to food to maintain a predetermined moisture level. They thereby prevent excessive drying out, or any changes in texture associated with this, and hardening. In intermediate-moisture food intended to have a long shelf life without expensive packaging, the role of humectants is to standardize water activity to 0.85 or less. This prevents the multiplication of bacteria which, as a rule, require water activity of 0.90 or more for growth. Growth of molds and yeasts is still possible below this limit, but it can be suppressed easily by addition of sorbic acid [8].

In confectionery, humectants can prevent undesirable crystallization of the sugar. They thereby contribute to the softness or chewiness of confectionery.

The primary humectants are sorbitol [*50-70-4*] and glycerol [*56-81-5*]. Common salt and sugar are added to intermediate-moisture foods to decrease water activity.

3.12. Anticaking Agents

Anticaking agents maintain the flowability of formed, powdered, or fine-grained products by preventing caking and adhesion. Mold-release agents (see Section 5.6), which keep food from sticking to molds or other containers, are different from anticaking agents.

Anticaking agents are inert organic or inorganic compounds which are generally insoluble in water. They coat the particles to be protected with a thin layer, increase the distance between them, and sometimes reduce their cohesion. In addition, they may prevent electrostatic attraction between particles with opposite charges.

The primary uses of anticaking agents are in common salt; spices; vegetable, beverage, and fruit powders; powdered soups; powdered sauces; leavening agents; and confectionery such as hard candy. The most important anticaking agents used in food are calcium stearate [*1592-23-0*] and magnesium stearate [*557-04-0*], silicon dioxide, silicates, talc [*14807-96-6*], flour, starch, and for common salt, alkali-metal ferrocyanides.

3.13. Coating Agents

Coating agents are compounds that are used to cover or coat foods, or that are added to food surfaces for their protection. They are distinct from packaging materials, which are not food additives. Coating agents are used to protect foods or their surfaces against undesirable changes, such as drying out and loss of aroma. In some food, coating agents may prevent oxidation or the growth of undesirable microorganisms on the surface. Other coating agents give an attractive look to the surface of a food that would not otherwise look very appealing.

Coatings are used mainly for citrus fruit, confectionery, meat products, and cheese. The most important coating agents used in food are waxes, resins, oils, cellulose esters or acetic acid esters of the monoglycerides of edible fatty acids, and talc.

4. Substances with Sensory Effects (Organoleptic Substances)

Not only should food have a nutritive value and be stable, it should stimulate the appetite by its appearance, taste, and aroma. Some substances have properties that influence positively the senses of smell, taste, and vision before, during, and after the consumption of food.

Foods naturally contain sensorially active compounds. However, many of these are volatile or unstable. They may, therefore, be lost during food processing and storage. The main purpose of food additives that produce sensory effects is to compensate for such losses as authentically as possible. In addition, some substances stimulate the secretion of saliva and digestive fluids in the gastrointestinal tract. Others may impart sweet, salty, or sour taste nuances or may tone down or mask undesirable odor and taste. Under no circumstances should such food additives be used to mask unsatisfactory quality or spoilage. Excessive flavoring and unnaturally high coloring are objectionable and are not in the interest of the consumer.

4.1. Coloring Agents

An attractively colored food stimulates the appetite more than a discolored one. Apparently, there is a relationship between the eye and the gustatory nerves. Another purpose of food coloring is to provide a more variable range of products, which is especially important in the confectionery industry. Despite some controversy in this area, the coloring of food is unavoidable [33–35].

Coloring fresh food is not permitted as a matter of principle. As a rule, coloring is used only for processed food with no color of its own or in which only residual amounts of color remain. Food must not be colored to simulate a higher level of nutritionally important components or, worse, to mask poor quality or spoilage.

4.1.1. Specifications

Coloring is generally added to food at a very early stage in processing. Coloring agents must therefore be stable to heating, cooling, acid, or oxygen. In particular, they must remain stable during the storage of the food when they are often exposed to light. Many natural colorants lack this stability so that, despite other advantages, their use is limited. Sulfur dioxide, which is used in a number of foods, can destroy many colors.

4.1.2. Uses and Individual Substances

Coloring agents are distinguished on the basis of their solubility, i.e., insoluble coloring, pigments, and water-soluble or fat-soluble coloring.

Pigments are used mainly for surface coloration of confectionery; *fat-soluble colorants,* for fatty foods such as margarine and cheese, and *water-soluble colorants,* for food with a high water content such as fruit products and beverages.

Depending on their origin, coloring agents may be classified as self-coloring food, natural coloring, or synthetic coloring. Examples of *natural coloring* are extracts of beetroot and turmeric, as well as carotene and other carotenoids, although as a rule the latter are produced synthetically [36]. Among purely *synthetic coloring* agents, triphenylmethane and azo dyes are used most often because they are the most stable.

The main uses for food coloring are in confectionery, cakes, fruit-based products, beverages, margarine, cheese, and certain fish products.

The most important colorants commonly used in foods are summarized below. In contrast to other food additives, major differences exist in the food-coloring laws of the EEC countries, Japan, and the United States (see Chap. 6).

E 100	curcumin [*458-37-7*]
E 101	riboflavin (lactoflavin) [*83-88-5*]
E 102	tartrazine [*1934-21-0*]
E 104	Quinoline Yellow [*8004-92-0*]
E 110	Sunset Yellow FCF [*2783-94-0*]
E 120	cochineal (carminic acid) [*1260-17-9*]
E 122	carmoisin (Azo Rubine) [*3567-69-9*]
E 123	amaranth [*915-67-3*]
E 124	Ponceau 4R (Cochineal Red A) [*2611-82-7*]
E 127	Erythrosin [*16423-68-0*]
E 131	Patent Blue V [*129-17-9*]
E 132	Indigo Carmine (indigotine) [*860-22-0*]
E 140 – E 141	chlorophyll [*1406-65-1*] and its copper complexes
E 142	Green S (Acid Brilliant Green BS or Lissamine Green) [*3087-16-9*]
E 150	caramel
E 151	Black PN (Brilliant Black BN) [*2519-30-4*]
E 153	carbon black (vegetable carbon)
E 160	carotenoids (carotene, annatto [*1393-63-1*], bixin [*6983-79-5*], norbixin [*542-40-5*], capsanthin [*465-42-9*], capsorubin [*470-38-2*], lycopene [*502-65-8*], β-apo-8′-carotenal and its ethyl ester)
E 161 a – g	flavoxanthin [*512-29-8*], lutein [*127-40-2*], cryptoxanthin [*472-70-8*], rubixanthin [*3763-55-1*], violaxanthin, [*126-29-4*], rhodoxanthin [*116-30-3*], canthaxanthin [*514-78-3*]
E 162	Beetroot Red (betanin) [*7659-95-2*]
E 163	anthocyanins
E 171	titanium dioxide [*13463-67-7*]
E 172	iron oxides and hydroxides
E 173 – E 175	metals (aluminum, silver, gold)
E 180	Pigment Rubine (Lithol Rubine BK) [*5281-04-9*]

4.2. Color Stabilizers

Color stabilizers help foods retain their natural color during processing and storage and prevent undesirable discoloration. Color stabilizers are especially important with meat products. The red pigment myoglobin can be oxidized by oxygen in the air to form metmyoglobin which is brown to grayish brown. Nitrates or nitrites stabilize the desired red meat color through formation of nitrosomyoglobin. This complex has good storage, cooking, and baking stability [37].

4.3. Bleaching Agents

Discoloration in light-colored food can be removed in some cases with bleaching agents. Sulfur dioxide and sulfites are effective color stabilizers for enzymatic oxidation processes because they are enzyme inhibitors.

Bleaching agents not only remove undesirable discoloration but also destroy vitamins. Because of this, they are added only to products in which this side effect is unimportant.

4.4. Intense Sweeteners (→ Sweeteners)

Intense sweeteners are compounds with a far more intensely sweet taste than sugar [38–40]. They can be synthetic or derived from plants. Because of their intense sweetness, they need be added only in small amounts and do not contribute (or do not contribute significantly) to the caloric value of food. The sweeteners can be tolerated by diabetics.

Intense sweeteners are judged by two criteria, their sweetening power and the quality of their sweetness (→ Sweeteners). The sweetening power is the factor by which the substance tastes sweeter than sucrose; it is concentration dependent. Data quoted in the literature usually refer to a comparison with 0.1 mol/L sucrose solution.

Intense sweeteners are used in foods for three main reasons: because of their lack of calories, because they are not harmful to diabetics, and because they are cheaper than sugar.

In the early 1970s, the two traditional sweeteners, saccharin and cyclamate, were suspected of being carcinogenic. Because of this, their use as food additives was prohibited in some countries. In other countries their use was restricted or made difficult because labels warning of their potential danger were required. This led to the

development of new sweeteners; among these, aspartame, other peptides, and acesulfame-K have achieved some importance. Possible sweeteners currently being discussed include chlorinated sugars.

4.5. Nutritive Sweeteners

Nutritive sweeteners are noncarbohydrate materials with a degree of sweetness similar to that of sugar; they can be used as substitutes for sucrose, glucose, and other sugars [38–40]. Like sugars they are metabolized. However, in contrast to sucrose and glucose, they can be metabolized independently of insulin and are, therefore, suitable for diabetics. Nutritive sweeteners increase osmotic pressure, sometimes to an even greater degree than sugars. When consumed in large amounts, they can therefore act as laxatives.

Nutritive sweeteners are used largely in place of sucrose or glucose in foods designed for diabetics. Because unlike sugars, they are not readily fermented, oral bacteria are unable to convert them into acids. Consequently, the risk of caries is lower with confectionery made from nutritive sweeteners than with that made from sugars.

The most frequently used nutritive sweeteners are the polyols (→ Sugar Alcohols) sorbitol [*50-70-4*], xylitol [*87-99-0*], and mannitol [*69-65-8*]. In some countries, fructose is included among the nutritive sweeteners because it is tolerated by diabetics.

4.6. Acidulants

Acidulants are used to impart a sour taste to foods in a controlled way. Besides being sour, most of them have a characteristic taste of their own. This can be desirable, for example, in the case of citric acid. Acids that taste only sour or nearly so are malic acid [*6915-15-7*] and orthophosphoric acid. Acidulants can highlight or suppress other aroma components of a food. The subjective perception of a sour taste is strongly influenced by salts with a buffering effect, by sweet compounds, and by other flavorings. In special cases, an acidulant may be chosen according to the speed with which the sour taste is registered or with which it fades.

Acidulants are added to beverages, especially nonalcoholic soft drinks, various fruit products, desserts, confectionery, pickles, mayonnaise-type products, salads, and fish products [41].

Vinegar and acetic acid are the most commonly used acidulants for pickles and mayonnaise-type or fish products. Orthophosphoric acid is used specifically to acidify cola beverages. Citric acid [*77-92-9*] is very widely used. Besides these, the most important acidulants are tartaric acid [*147-71-7*], lactic acid [*50-21-5*], malic acid [*6915-15-7*], succinic acid [*110-15-6*], and fumaric acid [*110-17-8*].

4.7. Substances with a Salty Taste

Common salt is the most important substance with a salty taste used in food. Because it has been used for thousands of years, in many cases it is no longer considered a food additive in legal terms. Besides its sensory function, salt is very important as a preservative (see Section 3.1).

Salt is an essential taste component of many foods such as meat products, certain fish products, vegetables, ready-to-eat meals, and most of all, bread. Without salt, many foods would taste bland or, in the view of some, be unpalatable.

If the intake of sodium must be restricted because of illness, substitutes for table salt, such as salts of potassium, calcium, magnesium, or choline with organic or inorganic acids, can be used.

4.8. Substances with a Bitter Taste

A bitter or slightly bitter taste in foods is desirable only in certain cases, generally in beverages. Quinine is used as an additive for tonic; hops resins, and extracts from hops and hops flowers are used as additives in beer. Caffeine, which is added to some cola beverages, also has a slightly bitter taste.

4.9. Substances with an Alkaline Taste

Among flavorings with an alkaline taste, the only one that has even a limited practical role is sodium hydroxide solution. It is used in some countries in the manufacture of pretzels.

4.10. Flavor Enhancers

Flavor enhancers are compounds that particularly enhance certain tastes or reduce undesirable flavors without having an especially strong taste of their own. They harmonize taste components and make food preparations more palatable. The Japanese expression *Umami* substances has come into general use [42].

Important Umami substances include glutamic acid [*56-86-0*] and glutamates, as well as purine-5′-ribonucleotides, especially inosine [*4691-65-0*], guanosine [*5550-12-9*], and adenosine [*61-19-8*] 5′-monophosphates. The latter are effective at far lower concentrations than glutamates. Lately, thaumatin [*53850-34-3*] has been of interest; this substance was originally developed as a sweetener. Common salt is also a flavor enhancer in the wider sense.

The main uses for flavor enhancers are in meat, soups, and vegetable products; gravies; ready-to-eat meals; seasoning; confectionery; beverages; and fruit products. Maltol [*118-71-8*] and ethyl maltol [*4940-11-8*] are flavor enhancers in confectionery, soft drinks, and fruit products.

4.11. Spices and Flavorings

In addition to the substances with a certain taste already described, certain materials, mostly of plant origin, have been used from ancient times to improve the flavor of food. These include spices, herbs, herb extracts, and essences. Aroma compounds are substances with a more or less pronounced odor that impart a specific aroma to foods [43–45]. Chemically pure compounds, as well as plant extracts and microbial metabolites or extracts, are used. Essential oils and preparations made from them are especially important (→ Flavors and Fragrances). Raw materials of animal origin play only a subordinate role in flavoring food.

Aroma compounds are perceived mainly through the nose; compounds with a taste and substances that affect food consistency are perceived more via the tongue. However, these groups of substances cannot always be considered separately. Aromagrams are now being discussed as a systematic method for overall evaluation of odor and taste. An aromagram is a form that can be used to sketch the aroma of any food and that allows the interaction between taste and odor to be illustrated, while the effect of consistency is also taken into account. The areas representing compounds that have overall sensory impact are shaded. In addition, the key substance (i.e., the primary aroma component) is given. Figure 1 shows the aromagram for Emmentaler cheese [43].

Flavoring is classified as natural, nature-identical, and artificial. *Natural flavorings* are obtained through purely physical processes, usually from plant material, or through microbiological processes. In many countries, synthetically produced substances which are chemically identical to the natural ones are called *nature-identical. Artificial flavorings* are substances synthetically produced that have not been found in any natural products suitable for human consumption. In some countries, only artificial flavorings are considered additives within the meaning of the food law.

Some aroma compounds are formed by the controlled use of technological methods. These may be strengthened or simulated by additives. An example involves flavor compounds produced by the Maillard reaction, which develop during the roasting of meat or baking of bread.

Both naturally occurring flavors and flavors produced during roasting and baking, as a rule, consist of numerous individual components. So far it has not been possible to synthetically reproduce natural flavors in their entirety in such a way that they are fully identical with the natural product. However, flavor impressions of a very high quality can be generated with synthetically produced flavoring.

4.12. Chewing Gum Bases

Chewing gum bases, also called masticatory substances, are the underlying components of chewing gum. At body temperature, they must be plastic, but they must also offer a certain bite resistance. Important chewing gum bases include gums, natural and synthetic resins, synthetic polymers, paraffins, and waxes.

5. Processing Aids

The Codex Alimentarius defines *processing aids* as “any substance or material, not including ap-

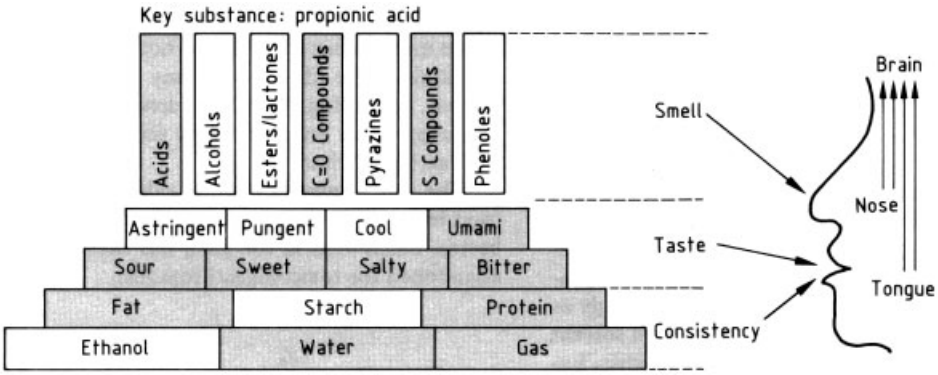

Figure 1. Aromagram of Emmentaler cheese

paratus or utensils, and not consumed as a food ingredient by itself, intentionally used in the processing of raw materials, foods or its ingredients, to fulfil a certain technological purpose during treatment or processing and which may result in the non-intentional but unavoidable presence of residues or derivatives in the final product" [46]. Processing aids have a special status among additives in the food laws of practically every country. Some countries do not consider them additives at all.

5.1. Extractants

Extractants are solvents used to extract certain components of food. This is done so that these components may be used separately or so that the food can be freed of certain unwanted components. As a rule, these solvents are removed from the food when extraction is complete, unless the solvent is necessary as a diluent for the extract.

The most commonly used solvents are water, ethanol, and other lower aliphatic alcohols, as well as hexane, other hydrocarbons, and in special cases, dichloromethane. Supercritical carbon dioxide has been used increasingly as an extraction solvent because of its good solvent properties and because it leaves no residue. Extractants are used in fat extraction from fatty raw materials, in the extraction of aroma compounds from plant materials, and in the extraction of caffeine from unroasted coffee beans.

Carrier solvents are distinct from extractants and are used to work small amounts of other additives into a food, e.g., to dissolve vitamins, antioxidants, or aroma compounds. Unlike extractants, carrier solvents are food additives in the true sense.

5.2. Clarifying Agents

Clarifying agents speed up the separation of suspended matter from those beverages in which clarity is desired, such as wine and beer. They also attract colloids and precipitate them. In addition to adsorbents, colloids of opposite charge can be effective against other colloids. If added in the right amounts, these precipitate together with the substances to be eliminated. Clarifying agents must be removed as far as possible with the suspended particles, mostly by sedimentation and ultrafiltration. The most important clarifying agents for food are bentonite [*1302-78-9*], activated charcoal, gelatin [*9000-70-8*], tannin, polyvinylpyrrolidone [*9003-39-8*], casein [*9000-71-9*], and caseinates.

Other clarifying agents include substances that form compounds of poor solubility with traces of soluble heavy metals. These can then be filtered readily. Thus, potassium ferrocyanide [*13943-58-3*] forms insoluble precipitates with traces of iron and copper. This process is especially useful in removing traces of iron from wine and is called blue or Moeslinger fining.

5.3. Filter Aids

To facilitate filtration, on suitable supports filter aids form a filter layer of appropriate pore size and maximum filtering capacity. These materials must be inert and should not contaminate the liquid to be filtered with any soluble compounds. In settling filters, mixtures of filter aids

and adsorbents are sometimes used, especially with activated charcoal.

The most important filter aids used in the food industry are kieselguhr and cellulose. Asbestos, much used in the past, has been largely abandoned for toxicological reasons.

5.4. Propellants

Propellants are used in the production of aerosols. Whipping cream, frying and baking and mold release oils, as well as spice extracts are used in aerosol form. Carbon dioxide, mixtures of propane and butane, dinitrogen monoxide, and certain fluorohydrocarbons and mixtures of these are used as propellants.

5.5. Cooling Agents and Cryogens

Cooling agents and cryogens are substances with a low boiling point which, when in direct contact with foods, extract heat from them and thereby cool them. The oldest *cooling agent* is crushed ice, and it continues to be important in chilling fish and other marine products at sea. The use of cryogens is more recent. The most important of these is liquid nitrogen which is sprayed onto food in special freezing tunnels. The advantage of liquid nitrogen is that the freezing process is particularly fast when oxygen is absent.

5.6. Mold-Release Agents

Release agents facilitate the removal of food such as confectionery or baked goods from molds. The most common mold-release agents are lecithins, stearic acid [*57-11-4*], calcium and other stearates, and waxes.

5.7. Antifoaming Agents

Antifoaming agents prevent unwanted foam development that may occur during food processing. Dimethylpolysiloxane, fatty acid esters of sorbitan and ethoxylated sorbitan, and various emulsifiers are suitable for this purpose.

5.8. Acidity Regulators

Acidity regulators generate a pH that is favorable or even essential for certain food processes. They are used, for example, to optimize the activity of enzymes or preservatives. The action of some gelling agents also depends on a particular range of pH. The most important regulators are the acids and acid salts described in Section 3.1. In addition, glucono-δ-lactone [*90-80-2*] may be used; it is gradually converted to gluconic acid, something that is occasionally desirable.

5.9. Emulsifying Salts (→ Cheese, Processed Cheese, and Whey, Chap. 6.)

Emulsifying salts are necessary for the production of processed cheese. They inactivate calcium which is important for the stability of the cheese gel. The action of emulsifying salts causes the insoluble paracasein gel to be converted to homogeneously flowing paracasein sol. After the melt has cooled down, the sol reconverts into a gel which, however, remains homogeneous and is stable. Some emulsifying salts are antimicrobial; thus, processed cheese is less susceptible to mold attack than natural cheese [47].

The most important emulsifying salts used in food are citrates and ortho-, di-, or polyphosphates. They are used mainly as mixtures that are tailored to the type and maturity of the original cheese to be processed.

5.10. Dough Conditioners and Flour Improvers

Dough conditioners and flour improvers enhance the processing and baking properties of flour and dough. Basically, they affect the gluten in flour and balance variations in the flour, especially in enzyme activity. Both oxidizing and reducing compounds can be used as dough conditioners. The effectiveness of *reducing* dough conditioners is attributed to their splitting of hydrogen or disulfide bridges. This breakdown improves the rheological properties of doughs so that less energy is used in working them.

Important dough conditioners include cysteine hydrochloride [*52-89-1*] and ascorbic acid [*50-81-7*]. Some countries, also allow azodicarbonamide [*123-77-3*] and bromates to be used.

Certain emulsifiers and amylolytic enzymes can also improve the properties of dough.

5.11. Leavening Agents (→ Bread and Other Baked Products, Chap. 10.3.)

Leavening agents lighten baked products, mostly by liberating carbon dioxide. Carbon dioxide can be generated biologically (e.g., by yeast or sourdough) or chemically (by baking powder). The latter is a mixture of sodium hydrogen carbonate and an acid, e.g., a solid organic acid such as tartaric or citric acid, or sour salts such as certain phosphates.

5.12. Enzymes (→ Enzymes)

Enzymes are widely distributed throughout food of plant or animal origin. They are involved in food spoilage and, therefore, must often be inactivated during food processing, usually by heat. On the other hand, some enzymes are deliberately added to foods to convert or remove certain components in a controlled and careful way, or to generate particularly desirable end products [48–51]. Often, enzymes make possible processes that either could not be carried out at all without them or would lead to profound and usually detrimental changes in food. The action of an enzyme is greatly influenced by its concentration, its activity, the length of time of application, pH optimum, the possible presence of enzyme inhibitors, and the process temperature. An advantage of using enzymes is that simple heating to 50 – 70 °C inactivates the enzyme and stops the reaction instantaneously. Enzymes generally have marked specificity and, when optical isomers are involved, usually react with only one enantiomer.

Uses and Individual Substances. Hydrolases, lyases, and isomerases are the most important classes of enzymes used as food additives. Enzymes are added to food either as standardized powders or in solution.

The addition of *pectinases* (→ Enzymes, Chap. 4.1.4.) to fruit increases juice yields by destroying the cellular structure of the fruit. Pectin-containing suspended matter, which may impede the filtration of fruit juice, can be flocculated by use of suitable pectinases and then removed. For this reason, such enzymes are called carifying enzymes.

Amylases (→ Enzymes, Chap. 4.1.1.) hydrolyze starch both during dough production and in the manufacture of glucose syrup and glucose. They are used in alcohol and beer manufacture to break down starch so as to obtain readily fermentable oligosaccharides. Of the enzymes that hydrolyze disaccharides, lactase and invertase have practical importance. *Lactase* (E.C. 3.2.1.108) [*9031-11-2*] cleaves lactose which has poor solubility and low sweetness, and cannot be tolerated by many people. It thereby improves milk and dairy products. *Invertase* (E.C. 3.2.1.26) [*9001-57-4*] cleaves sucrose into glucose and fructose (inversion). The addition of invertase makes possible the production of sweets with a solid coating and a soft center. Under the action of invertase, an originally solid sucrose-containing mass with a suitable coating takes on a soft, creamy, or liquid consistency. *Glucose isomerase* (E.C. 5.3.1.18) [*9055-00-9*] isomerizes glucose to fructose (→ Fructose). Because of the greater sweetness of fructose compared to sucrose and glucose, its use leads to sweeter products. *Naringinase* [*9068-31-9*] makes orange and grapefruit juices more palatable by cleaving their bitter component, naringin, into substances that do not taste bitter.

Among the *proteolytic enzymes,* exogenous proteases are used to tenderize meat. Like the proteases that occur naturally in meat, they split proteins into shorter chain proteins and peptides. Papain (E.C. 3.4.22.2) [*9001-73-4*], bromelain (E.C. 3.4.22.4) [*37189-34-7*], and ficin (E.C. 3.4.22.3) [*9001-33-6*] are proteases of plant origin that also break down proteins of connective tissues. In beverages, proteases can prevent flocculation of proteins and clouding due to protein – tannin complex formation.

The milk-clotting enzyme rennin (E.C. 3.4.23.4) [*9001-98-3*], obtained from the mucous membrane of the stomach of young calves has great commercial importance. It is indispensable for the manufacture of many cheeses. Microbial proteases are also used in cheese manufacture.

5.13. Microbial Cultures

Cultures of microorganisms resemble enzymes in their application and mode of action [51]. Frequently, their activity is due to the enzymes released by these cultures into food. Cultures of microorganisms may be used if they do not produce toxins or other undesirable substances. Commercial preparations are tested routinely for this.

From the earliest times, yeast has been very important as a dough-leavening agent, and pure cultures of certain yeasts are used in wine making. Certain mold cultures are used to make blue cheese, and pure cultures of certain molds are employed in the manufacture of salami. So-called starter cultures are used in certain areas of dairy and meat technology.

6. Legal Aspects

6.1. Food Law Authorizations

In many countries, into the 20th century food law was limited to prohibiting the use of additives that were damaging to health. This was achieved partly by listing certain substances on so-called negative lists. In the long run, this procedure did not protect consumers and now has generally been replaced by so-called positive lists. According to these, all substances are prohibited unless specifically permitted by decree. Generally, the law also states which substances may be used for which foods. Sometimes, the maximum amount permitted in a food and other conditions are also fixed. Both authorization and maximum amount are based on technological necessities, with calculation of a sensible safety margin. An obvious prerequisite is that the substance be innocuous.

6.2. Definitions

In the following material, the term "food additive" is defined according to the Codex Alimentarius and the food legislation of the United States, Japan, and the Federal Republic of Germany.

6.2.1. Codex Alimentarius

In the Codex Alimentarius, *food additive* means any substance not normally consumed as a food by itself, and not normally used as a typical ingredient of the food, whether or not it has nutritive value, the intentional addition of which to food for a technological (including organoleptic) purpose in the manufacture, processing, preparation, treatment, packing, packaging, transport or holding of such food results, or may be reasonably expected to result (directly or indirectly) in it or its by-products becoming a component of or otherwise affecting the characteristics of such foods. The term does not include "contaminants" or substances added to food for maintaining or improving nutritional qualities [46].

6.2.2. United States of America

In the United States, a food additive is any substance the intended use of which results or may reasonably be expected to result, directly or indirectly, in its becoming a component or otherwise affecting the characteristics of any food (including any substance intended for use in producing, manufacturing, packing, processing, preparing, treating, packaging, transporting, or holding food; including any source of radiation intended for any such use), if such substance is not generally recognized, among experts qualified by scientific training and experience to evaluate its safety, as having been adequately shown through scientific procedures (or, in the case of any substance used in food prior to January 1, 1958, through either scientific procedures or experience based on common use in food) to be safe under the conditions of its intended use; except that such term does not include a pesticide chemical in or on a raw agricultural commodity, or a pesticide chemical to the extent that it is intended for use or is used in the production, storage, or transportation of any raw agricultural commodity, or a color additive, or any substances used in accordance with a sanction or approval granted prior to the enactment of this paragraph pursuant to this Act, the Poultry Products Inspection Act or the Meat Inspection Act of March 4, 1907, as amended and extended [52].

Thus in the United States, the term additive is more widely defined than anywhere else in the world, yet food coloring and GRAS (generally recognized as safe) substances are not classified as food additives. These GRAS substances include salt, sugar, vinegar, and baking powder, as well as many substances classified as food additives in other countries, which are considered to be especially safe, such as citric and sorbic acids. However, regulations covering the use of coloring and GRAS substances in the United States are very similar to those for food additives. In the United States, additives include practically all materials that come in contact with food in any way.

Distinction is made between three types of food additive. (1) Food additives permitted for direct addition to food for human consumption are those substances which, in other countries, are considered food additives as such. (2) Secondary direct food additives permitted in food for human consumption include polymers and polymer adjuvants for food treatment, enzyme preparations and microorganisms, solvents, lubricants, release agents and related substances, and other products that come in contact with food only temporarily. (3) Indirect food additives are adhesives, components of coatings, food-packaging materials, components of paper and paperboard, polymers, processing aids, and sanitizers.

6.2.3. Japan

In Japan, the term "additive" means anything added to, mixed with, permeating, etc., food in the process of manufacturing, processing, or preserving it [53].

In Japanese food law, synthetic and naturally occurring additives are treated differently. The latter, in particular naturally occurring flavors and vitamins, do not require any special permission for use. This explains, for example, why sweeteners isolated from plants must be specifically allowed as additives everywhere else in the world, while they can be used freely in Japan.

6.2.4. Federal Republic of Germany

For the purpose of German food legislation, additives are substances destined to be added to foods to influence their character or to achieve particular properties or effects. This definition does not include substances of natural origin or those which are chemically identical to natural ones and are, according to general commercial practice, used predominantly because of their nutritive value, aroma, or flavor characteristics, or which are used as stimulants. Drinking or table water is also included in this category. The following are equivalent to additives:

1) minerals and trace elements and their compounds except common salt,
2) amino acids and their derivatives,
3) vitamins A and D and their derivatives,
4) nutritive sweeteners except fructose,
5) intense sweeteners,
6) adipic acid,
7) nicotinic acid and nicotinamide,
8) nitrite salt for curing,
9) substances for use in the manufacture of edible coatings,
10) substances to be added to nonedible food surfaces,
11) substances to be used in the treatment of food in such a way as to find their way into or onto the food, and
12) propellants or similar substances to be applied by pressure to food whereby they come in contact with it [54].

A peculiarity of the food law in the Federal Republic of Germany is that substances of natural origin or their chemically identical synthetic analogues are not considered additives if, by common commercial practice, they are used predominantly because of their nutritive, aroma, or taste values. This means, for example, that in the Federal Republic of Germany, flavorings, including nature-identical ones as well as citric and acetic acids, do not constitute additives within the framework of food law.

6.2.5. Regulations and Specification

In most countries, additives must be listed with the ingredients.

In many countries, all additives permitted are summarized in a single list. Examples are relevant regulations in the Scandinavian countries,

the Federal Republic of Germany, and Switzerland. In countries such as the United Kingdom, authorization of food additives is controlled by individual regulations, with reference to specified groups of additives. Some countries have authorizations within product regulations for certain foods. In addition, mixed systems exist, e.g., in the United States.

The international consensus is largely that, in principle, authorization for food additives should be given only if justified technologically. This means that without such additives, the required effect could only be achieved either uneconomically or not at all. In every country, food additives are subject to purity specifications which are fixed by law.

7. References

1. World Health Organization Geneva: "Principles for the Safety Assessment of Food Additives and Contaminants in Food" *Environ. Health Criter.* **70** (1987).
2. US Food and Drug Administration Bureau of Foods: *Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives Used in Food,* Washington 1982.
3. E. Lück, K.-H. Remmert: *Z. gesamtes Lebensmittelrecht* **3** (1976) 115 – 143.
4. Hoffmann-La Roche: *Vitamin-Compendium,* 2nd ed., Grenzach–Wyhlen 1980.
5. R. A. Morton: *Fat-soluble Vitamins,* Pergamon Press, Oxford 1970.
6. J. N. Counsell, D. H. Hornig: *Vitamin C (Ascorbic Acid),* Applied Science Publ., London 1981.
7. H. Zumkley: *Spurenelemente–Grundlagen–ätiologieDiagnose–Therapie,* Thieme Verlag, Stuttgart–New York 1983.
8. E. Lück: *Antimicrobial Food Additives,* Springer Verlag, Berlin–Heidelberg–New York–Tokyo 1980; *Chemische Lebensmittelkonservierung,* 2nd ed., Springer Verlag, Berlin–Heidelberg–New York–Tokyo 1986.
9. A. L. Branen, P. M. Davidson: *Antimicrobials in Foods,* Marcel Dekker, New York–Basel 1983.
10. L. Leistner, W. Röder, K. Krispien in L. B. Rockland, G. F. Stewart (eds.): *Water Activity Influences on Food Quality,* Academic Press, New York–London–Toronto–Sidney–San Francisco 1981, p. 855.
11. M. von Schelhorn, *Arch. Mikrobiol.* **19** (1953) 30 – 44.
12. M. von Schelhorn: *Fette Seifen Anstrichm.* **56** (1954) 221 – 224.
13. D. A. A. Mossel: *Proc. 21st Symp. Soc. Gen. Microbiol.,* 1971, pp. 177 – 195.
14. E. M. Lukas: *Zentralbl. Bakteriol. Parasitenkd. Infektionskr. Hyg., Abt. 2 Naturwiss. Allg. Landwirtsch. Tech. Mikrobiol.* **117** (1964) 485 – 509.
15. J. C. Allan, R. J. Hamilton: *Rancidity in Foods,* AVI, Westport 1977.
16. R. R. Hiatt: *CRC Crit. Rev. Food Sci. Nutr.* **7** (1975) 1 – 12.
17. WHO: *A Review of the Technicological Efficacy of Some Antioxidants and Synergists.* WHO: Food Additives Series No. 3, Geneva 1972.
18. E. R. Sherwin: *J. Am. Oil Chem. Soc.* **49** (1972) 468 – 472.
19. L. C. Schroeter: *Sulfur Dioxide,* Pergamon Press, Oxford 1966.
20. G. Dinglinger: *Z. Lebensmitteltechnol. Verfahrenstechn.* **35** (1984) 49, 50, 52, 54, 55.
21. G. Schuster: *Emulgatoren für Lebensmittel,* Springer Verlag, Berlin–Heidelberg–New York–Tokyo 1984.
22. *Emulgatoren–Ihre Wirkung in Lebensmitteln,* B. Behr's, Hamburg 1983.
23. N. Krog: *J. Am. Oil Chem. Soc.* **54** (1977) 124 – 131.
24. V. Bade: *Getreide Mehl Brot* **28** (1974) 296 – 299.
25. F. C. Greene: *Baker's Dig.* **49** (1975) no. 3, 16 – 18, 20 – 24, 26.
26. R. H. Ellinger: *Phosphates as Food Ingredients,* CRC Press, Cleveland 1972.
27. *Phosphate–Anwendungen und Wirkung in Lebensmitteln,* B. Behr's, Hamburg 1983.
28. H. D. Graham: *Food Colloids,* AVI, Westport 1977.
29. W. Burchard: *Polysaccharide,* Springer Verlag, Berlin–Heidelberg–New York–Tokyo 1985.
30. M. Glicksman (ed.): *Food Hydrocolloids,* **vol. I,** vol. II, CRC Press, Boca Raton 1982, 1983.
31. H. Neukom, W. Pilnik (eds.): *Gelier- und Verdikkungsmittel in Lebensmitteln,* Forster, Zürich 1980.
32. *Hydrokolloide–Stabilisatoren, Dickungs- und Geliermittel in Lebensmitteln,* B. Behr's, Hamburg 1984.

33. J. N. Counsell (ed.): *Natural Colours for Food and Other Uses,* Applied Science, London 1981.
34. *Food Technol. (Chicago)* **34** (1980) 77 – 84.
35. *Farbstoffe für Lebensmittel,* Boldt, Boppard 1978.
36. J. C. Bauernfeind: *Carotenoids as Colorants and Vitamin A Precursors,* Academic Press, New York 1981.
37. G. G. Giddings: *CRC Crit. Rev. Food Sci. Nutr.* **9** (1977) 81 – 114.
38. C. A. M. Hough, K. J. Parker, A. J. Vlitos (eds.): *Developments in Sweeteners–1,* Applied Science, London 1979.
39. T. H. Grenby, K. J. Parker, M. G. Lindley: *Developments in Sweeteners–2,* Applied Science, London 1983.
40. L. O'Brien Nabors, R. C. Gelardi (eds.): *Alternative Sweeteners,* Marcel Dekker, New York–Basel 1986.
41. M. H. M. Arnold: *Acidulants for Foods and Beverages,* Food Trade Press, London 1975.
42. J. A. Maga: *CRC Crit. Rev. Food Sci. Nutr.* **18** (1983) 231 – 312.
43. K. H. Ney: *Lebensmittelaromen,* B. Behr's, Hamburg 1987.
44. W. Auerswald, B. M. Brandstetter: *Natürliche, natur-identische und synthetische Geschmacksstoffe und deren lebensmittelrechtliche Aspekte,* Maudrich, Wien 1973.
45. R. C. Lindsay: *Food Technol. (Chicago)* **38** (1984) 76 – 81.
46. Joint FAO/WHO Food Standards Programme. Codex Alimentarius Commission: Codex Alimentarius, vol. 14 *Food Additives,* Food and Agriculture Organization of the United Nations, Rome 1983.
47. A. Meyer: *Joha-Schmelzkäsebuch,* Benckiser-Knapsack, Ludwigshafen 1970.
48. S. Schwimmer: *Source Book of Food Enzymology,* AVI, Westport 1981.
49. G. G. Birch, N. Blakebrough, K. J. Parker (eds.): *Enzymes and Food Processing,* Applied Science, London 1981.
50. *Enzympräparate–Standards für die Anwendung in Lebensmitteln,* B. Behr's, Hamburg 1983.
51. *Starterkulturen und Enzyme für die Lebensmitteltechnik,* VCH Verlagsgesellschaft, Weinheim 1987.
52. Y. H. Hui: *United States Food Laws, Regulations and Standards,* J. Wiley & Sons, New York–Chichester–Brisbane–Toronto 1979.
53. The Federation of Food Additives Associations: *Food Sanitation Law,* "Food Additives in Japan," Japan 1981.
54. Gesetz über den Verkehr mit Lebensmitteln, Tabakerzeugnissen, kosmetischen Mitteln und sonstigen Bedarfsgegenständen (Lebensmittel- und Bedarfsgegenständegesetz) vom 15. 8. 1974.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Foods, 4. Food Packaging

Standard Article

Norbert Buchner[1], Horst Weisser[2], Heinrich Vogelpohl[3], Albert L. Baner[4], Rainer Brandsch[5], Otto Piringer[6]

[1]Robert Bosch GmbH, Waiblingen, Federal Republic of Germany
[2]Technische Universität München, Lehrstuhl für Brauereianlagen und Lebensmittel-Verpackungstechnik, Freising – Weihenstephan, Federal Republic of Germany
[3]Technische Universität München, Lehrstuhl für Brauereianlagen und Lebensmittel-Verpackungstechnik, Freising – Weihenstephan, Federal Republic of Germany
[4]Friskies R & D Center (Nestlé), Missouri, USA
[5]FABES Forschungs GmbH, München, Federal Republic of Germany
[6]FABES Forschungs GmbH, München, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a11_583.pub2
Article Online Posting Date: January 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Historical Aspects
2. Definitions
3. Economic Significance
3.1. Functions and Effects
3.2. Characteristic Data
4. Objectives
4.1. Miscellaneous Objectives
4.2. Protective Function of Packaging
4.2.1. Protecting Contents
4.2.2. Protecting Humans and the Environment
5. Properties of Packaging Materials
5.1. General Mechanical Data
5.2. Barrier Properties
5.3. Thermal Properties
5.4. Biological Properties
5.5. Electrical Properties
5.6. Chemical Properties
5.7. Processing Properties
6. Types of Packaging Materials
6.1. Paper
6.2. Cardboard
6.3. Corrugated Cardboard
6.4. Coated Paper and Cardboard
6.5. Plastics for Packaging
6.5.1. Cellophane (Regenerated Cellulose)
6.5.2. Polyethylene
6.5.3. Ethylene Ionomers
6.5.4. Polypropylene
6.5.5. Poly(Vinyl Chloride)
6.5.6. Polystyrene and Styrene Copolymers

6.5.7. Thermoplastic Polyesters
6.5.8. Polyamides
6.5.9. Polycarbonates
6.5.10. Barrier Polymers
6.5.11. Foamed Plastics
6.5.12. Polymer Composites
6.5.13. Aluminum-Polymer Laminates
6.5.14. Metallized Polymers
6.5.15. Polyester Films with Vacuum-Deposited Silicon Dioxide
6.6. Metal Cans
6.7. Glass Bottles and Wide-Neck Containers
7. Packaging Machines
7.1. Unit Operations in Packaging Machines
7.1.1. Overwrapping Articles
7.1.2. Welding and Heat Sealing of Packaging Materials
7.1.3. Cold-Sealing Packaging Materials
7.1.4. Cutting Packaging Materials
7.1.5. Web Tension in Packaging Materials
7.1.6. Thermoforming of Polymers
7.1.7. Processing Folded Boxes
7.1.8. Closing and Sealing Boxes
7.2. Product Dosage in Packaging Machines
7.2.1. Gravimetric Dosage
7.2.2. Volumetric Dosage
7.3. Robots in Packaging
7.4. Aseptic Packaging
8. Interaction between Packaging and Food
8.1. Physical Interactions
8.1.1. Permeation
8.1.2. Migration
8.1.3. Prediction of Diffusion Coefficients
8.1.4. Multilayer Structures
8.1.5. Non-Fickian Processes
8.2. Chemical Interactions
8.2.1. Production of Off-Odors
8.2.2. Other Reactions
8.3. Measurement Methods and Migration Limits
8.3.1. Measurement Methods
8.3.2. Migration Limits
8.4. Legal Aspects
8.4.1. General Requirements
8.4.2. European Community
8.4.3. United States
8.4.4. Japan
9. Packaging and the Environment

[Top of Page]

1. Historical Aspects

Norbert Buchner, Horst Weisser, Heinrich Vogelpohl

The packaging requirements of our ancestors were met by using materials that were either provided by nature or could easily be adapted for packaging purposes: leaves, straw, leather, and clay. Glass, which was known to the ancient Egyptians, did not become an important packaging substrate in the Western Hemisphere until the 15th century, in the form of blown-glass articles [1]. In 1810, the discovery by NICOLAS APPERT of the heat sterilization of foods contained in glass articles, provided a considerable boost to the concept of packaging. In the same year, two patents were granted in England which described the utilization of tin-coated iron containers for the purpose of packaging foodstuffs [2]. In 1895, M. J. OWENS discovered a process for mass-producing glass containers. Similarly, in the 19th century, paper and cardboard were used increasingly in packaging applications. At the turn of the century, the first brand-name products featuring their own unique packaging were developed [3]. In 1938 – 1939, HUBERT ECKLÖH founded the first European self-service market in Osnabrück (Federal Republic of Germany); the self-service supermarket, along with the development of thermoplastics before and after World

War II, provided the impetus for the evolution of modern packaging concepts.

The first attempts at mechanization of packaging appeared around the turn of the century, with the development of bottling machines for beverages. However, these were not fully automatic systems, and many of the operations (e.g., the feeding of empty bottles or the removal of filled ones) had to be carried out manually. Only the valves were operated and controlled by the machine. Today, bottling and canning systems are manufactured that can automatically fill and seal up to 1500 containers per minute. Capacity increases are not expected in the near future.

The development of specialty machines was necessary in many cases in order to incorporate specific product properties or special packaging requirements. This lead to essentially task-specific machines and, as a result, low production in many cases. Hence, even today the design of packaging machines is a specialty within the field of mechanical engineering. Empirical considerations are predominant; basic research exists in only a few small areas.

[Top of Page]

2. Definitions

Norbert Buchner

Over 800 terms that are commonly encountered in the packaging industry are included in the standards DIN 55 405 and DIN 8740. Accordingly, packaging is a general term encompassing the whole range of packaging substrates and accessories. It is identified further by a preceding word describing the ware to be packed (e.g., fruit packaging) or the type of material employed (e.g., sheet-metal packaging). The *packaging material* is the substrate from which the packaging is manufactured. A *packaging* is defined as a finished article made from packaging material, which is ready to receive a commodity. The commodity is the article that is to be packed or that has already been packed. The *package* is thus a unit capable of being stored or transported, which includes both the packaging and the contents thereof.

[Top of Page]

3. Economic Significance

Norbert Buchner

3.1. Functions and Effects

Apart from its obvious functions (see Section Objectives), packaging has a variety of important effects and represents an indispensable component of the economic infrastructure. Packing contributes to a reduction in manufacturing costs because it enables large central production plants to be organized which operate at high economic efficiencies.

Packaging reduces waste. Scrap from large production sites can frequently be used as a raw material for byproducts (pectin from fruit purees, gelatin from bones). In the household, however, it represents waste. A comparison between Mexico City and three major cities in United States has shown that the waste per capita is reduced where highly developed packaging is available [4].

Packaging prevents spoilage. In Third World countries, between 20 and 50 % of the harvest is lost because of deficiencies in storage, transportation, and packaging facilities. Before World War II, losses of foodstuffs in Germany amounted to 13.4 % (in the 1980s only 1 %) [5].

Packaging reduces storage costs. Processed and packaged foods can often be stored efficiently without the need for special environmental conditions.

Packaging reduces transportation and marketing costs. Processed food is frequently lighter and more compact than fresh products.

Packaging makes new, practical, and time-saving foodstuffs possible with new forms of nutrition. Estimates are that, as a direct consequence of packaging, one hour of housework is saved per day in every modern household. Articles can be adapted to suit their designated application (e.g., sports or leisure).

Packaging makes a contribution to the improvement of nutrition and health. Valuable nutrients are available in the same high quality throughout the year. Foods can also be transported, without considerable loss, from areas of excess to famine-stricken regions.

3.2. Characteristic Data

The world market for packaging substrates and packaging accessories is approximately 3×10^{11} \$, which includes ca. 6×10^{10} \$ in Europe (EEC and EFTA) and 1.7×10^{10} \$ in the Federal Republic of Germany in 1987 [6]. Independent of the development stage, the capacity of the packaging sector in a political economy lies between 1.5 and 2 % of the gross national product. Economically, the packaging industry is very important, occupying twelfth position both in the Federal Republic of Germany and in France.

Packaging costs for individual items vary a great deal. For foods, the packaging costs in many countries amount to ca. 8 % of the retail price.

[Top of Page]

4. Objectives

Norbert Buchner

4.1. Miscellaneous Objectives

The objectives of packaging are listed in Table 1.

Table 1. Objectives of packaging

Stages	Protection				Information	Miscellaneous uses
	Contents		Consumer	Environment		
	Loss	Damage				
Retail unit	leakage	mechanical	fraud	odor	advertising	service
Consumer unit	theft	aroma damage	chemical threat	contamination	statutory labeling	secondary uses
Transportation unit		water vapor	injury	chemical threat	additional information	
Storage unit		oxygen				
Shelf unit		light				
		microbiological				
		pests				
		“breathing”				

Organizing and Itemizing. Packaging can organize, and hence reduce the cost of, various stages in production, wholesale, and retail.

The *retail unit* (single packs, multipacks, combination packs) simplifies selling the product. The *consumer unit* (portion pack, bulk pack for canteens) assists meal preparation and provides problem-free snacks between meals. The *transportation and storage units*, frequently loaded pallets, facilitate storage and transport. The *shelf unit* simplifies filling shelves and displaying prices. Many recommendations have been made regarding shelf units [7].

Advertising, Labeling, Information. In self-service stores, packaging constitutes a silent salesman. By using the advertising function of packaging, turnovers can be achieved which reduce the manufacturing costs of the article in question. Additionally, however, packaging must provide information pertaining to contents, price, use, manufacturer, amount, ingredients, and storage details.

Additional Functions. Packaging is increasingly assuming a service role (apportioning; foaming, baking, boiling, or heating by microwave, all in the packaging). Alternatively, it can be employed for secondary uses (drinking vessels or storage containers).

4.2. Protective Function of Packaging

Protection of the contents, the consumer, and the surroundings is the most important function of packaging.

4.2.1. *Protecting Contents*

Preventing Loss of Contents. Packaging must be designed so that under typical end-use conditions, leakage will not occur. Typically, packagings for small articles are oversized to help reduce shoplifting losses in self-service stores (between 0.4 and 1 %).

Preventing Damage to an Article. Packaging must protect the contents from many forms of damage:

Hygienic Protection. Packaging represents a safeguard against unhygienic conditions generated by people or the environment.

Mechanical Protection. Packaging prevents mechanical damage (breakage, crushing, abrasion) caused by pressure, impact, or vibrations.

Aroma Damage. Packaging protects the contents from damage with regard to aroma. This can arise from loss of the intended aroma, uptake of foreign odors, or formation of aroma-intensive materials.

Damage Caused by Diffusion of Water Vapor [8]. Commodities contained in a packaging material can often be affected adversely by the uptake or loss of unacceptable amounts of water vapor. Damage caused by the *uptake of water vapor* can be physical (loss of crispness, agglomeration, or progressive stickiness; physicochemical (crystal growth); chemical (Maillard reactions and oxidation [9]); and enzymatic or microbiological in nature. Following the *loss of water vapor*, undesirable

physical changes (staleness) or color and oxidation damage (freezer burn) can result.

Oxygen Damage. Damage may be caused by an excess or a deficiency of oxygen. Many ingredients in foodstuffs are oxidized on contact with oxygen: e.g., fats, some amino acids (arginine, histidine, lysine, methionine), a range of aromatic additives, several vitamins (A, C, E), and colorants. The sensitivity toward oxygen increases as the quality of the food improves, Hence, the demand for vacuum and protective gas packing increases. Damage due to *oxygen deficiency* can occur with red-colored fresh meat. Under conditions of oxygen depletion, and depending on the partial pressure of oxygen, the desirable pink-colored oxymyoglobin can turn to either purple-colored myoglobin (reversibly) or gray metmyoglobin (irreversibly). Packagings containing fresh meat, which are designed for a storage time of 5 ± 1 d, therefore contain approximately 30 % oxygen to preserve the oxymyoglobin [10].

Damage by Light. Light can cause damage to food by promoting chemical processes resulting from (1) the absorption of energy, most typically shortwave radiation, or (2) the warming effects of long-wave radiation. In so-called photooxidation processes (where light absorption and extent of oxidation are related), light provides the necessary energy. Autoxidation (a chain reaction with a propagation phase) is frequently initiated by the effects of light.

Hazards Caused by Microorganisms. Sufficiently moist foodstuffs are endangered by microbes. The threat depends upon the water activity a_w (⟶ Foods, 2. Food Technology), pH, temperature, and atmosphere. The lower limits of a_w are 0.7 – 0.72 for molds, 0.88 for yeasts, and 0.91 – 0.95 (absolute threshold 0.86) for bacteria. All microorganisms can grow above pH = 4.5, including the thermally resistant spore-forming and pathogenic species. Below this limit, growth is restricted to molds, yeasts, and a few non-spore-forming, acid-producing bacteria, as well as some pathogenic microorganisms. When the temperature decreases, the number of microorganisms that are able to reproduce, along with the rapidity at which they reproduce, decreases. Below –12 °C, no further growth of microorganisms can be observed. In deep-freeze packages, the number of microorganisms is only slightly reduced. By deoxygenation only the growth of aerobic microorganisms can be hindered. However, the addition of sufficient carbon dioxide impedes the growth of all microorganisms to various degrees. The addition of carbon dioxide is limited because of the danger of an undesirable increase in the acidity of the contents, as well as the risk of vacuum formation due to the high solubility of carbon dioxide. Packaging that includes a modified atmosphere has been developed for fish, meat, oven-ready meals, fruit, and vegetables.

Threat of Insects. Insects, in their various stages of development (eggs, larvae, etc.), and their excrement can endanger stored food. In temperate climate they mostly represent a hygiene problem. Insects can enter packages in various ways: they can be introduced by the packaging substrate; they may be carried by the food itself; or they can overcome the packaging by invasion through defects or by actually penetrating the material. The introduction of insects via the packaging substrate can be prevented by using film from a roll, whereas invasion can be counteracted by aroma-tight substrates with a hard surface [11].

Damage Caused by Ripening. Fresh foods, such as fruit and vegetables, can suffer damage due to loss of water vapor and “breathing” (exhaustion of oxygen and development of carbon dioxide). This can be reduced by evacuation of oxygen and addition of carbon dioxide to such an extent that the oxygen concentration during storage does not fall below that at which breathing turns to fermentation. Apples, for example, are ideally stored with 3 % O_2, 5 % CO_2, and 92 % N_2 [12]. Changes in this controlled atmosphere can be compensated by suitable storerooms. Choosing packaging materials with appropriate gas permeability from available materials represents only a compromise.

4.2.2. *Protecting Humans and the Environment*

Packaging also fulfills a protective function for humans and the environment. After several crimes involving the poisoning of medicine and food, so-called tamperproof (or at least tamper-indicating) packaging has become increasingly important. An intact seal or cap for a packaging precludes either partial removal or total exchange of the contents.

[Top of Page]

5. Properties of Packaging Materials

Norbert Buchner

The following sections represent a review of the most important properties of packaging materials. The standards given in parentheses provide information about test methods. Packaging substrates which are sensitive to climatic changes are tested for normal applications under standardized test conditions, i.e., 23 °C, 50 % R. H. (standard climates—see DIN 50 014; ISO 554).

5.1. General Mechanical Data

The *thickness* of a packaging material is characterized by using the actual thickness (plastic film, aluminum foil, sheet metal—DIN 53 370, 53 105; ISO 4593) or the unit weight (paper, cardboard, cellophane—DIN 53 352, 53 104; ISO 4591). *Breaking strength* and *strain at break* (DIN 53 455, 53 112; ISO/DIS 1184) are quoted as the dry strength for normal applications or the wet strength for applications involving the packaging of fat – water emulsions (butter, margarine).

The *bursting strength* (DIN 53 141, 53 113) characterizes the static area loadability of a stack of wrappings. *Impact strength* and *perforation resistance* (DIN 5337, 53 142) provide a measure of the dynamic loadability of a packaging material.

The *bending stiffness* of a packaging substrate in the elastic region (DIN 53 121, 53 123) is an important characteristic which indicates the static loadability of packaging as well as its processing properties on a packaging machine.

The static or dynamic *coefficient of friction* (DIN 53 375), which is generally velocity dependent, exerts an influence on the static loadability (the displacement of stacks), on dynamic processes, and on the machine processibility of packaging substrates.

5.2. Barrier Properties

The *watertightness* of a material is important for packaging liquids and fat – water emulsions. In addition, for fat-containing substances such as coffee, the *fat* and *oil barrier properties* (DIN 53 116) are significant.

Gas permeability (DIN 53 380, 53 536; ISO 2556) *water vapor transmission* (DIN 53 122; ISO R 1195) and *aroma barrier properties* of packing materials are particularly important for packaging sensitive commodities. *Air barrier properties* (DIN 53 120) play a role in the processing and stacking procedures. *Transparency* (DIN 10 050 Bl. 9) is important for reasons associated with presentation of the goods, in combination with light-induced damage to the products. The *barrier toward microbes* (DIN 6730) can generally be guaranteed by using plastic films because these are nonporous. Special paper products with very low pore sizes also generate an effective microbe barrier.

The *insect resistance* of packaging is promoted by the smoothness and hardness of the surface, as well as the thickness, strength, and aroma barrier properties of the substrate. Repellants, insofar as allowed, that are included in or on the packaging material, further increase insect resistance.

5.3. Thermal Properties

The range of application temperatures is limited at the lower end of the scale by material embrittlement and at the upper end by softening of the complete film or of the coating. The latter can lead to unacceptable reductions in stiffness and strength or to the onset of undesirable sticking or shrinkage.

The thermoforming (orientation blow molding) (see Section Welding and Heat Sealing of Packaging Materials) range and the region of pronounced shrinkage for films lie between the softening temperature and the melting temperature. During the manufacture of shrinkable films, prestressing (uniaxial or biaxial) is carried out in the thermoelastic region above the crystalline melting point. These stresses are frozen into the system by quenching and can subsequently be relaxed by raising the temperature. The shrinking forces and the actual degree of shrinkage (DIN 53 369) are determined by the polymer type, thickness, amount of prestressing, and shrinking temperature.

In the temperature range between melting and decomposition, processes leading to melt formation can be carried out, e.g., welding, heat sealing, extrusion, injection molding, and blowing of hollow articles (see Section Thermoforming of Polymers).

5.4. Biological Properties

Biological properties include the total content of bacteria and the degree of surface colonization with bacteria (DIN 54 378, 54 379), along with resistance to insects, rodents, and microbes.

5.5. Electrical Properties

Electrical charging of packaging substrates and dried commodities can lead to processing disruption (see Section Overwrapping Articles). The propensity to electrostatic charging (DIN 53 486) is governed by the type of material involved, the surface and breakdown resistance, and with hydrophilic substrates, the humidity conditions. For the welding of polymers such as poly(vinyl chloride) by using a high-frequency field, the dielectric loss angle is decisive.

5.6. Chemical Properties

Important chemical properties include chemical stability (DIN 53 476; ISO R 175/462), resistance of metals to corrosion or of polymers to stress corrosion (DIN 53 449), content of heavy metals or other toxic materials, stability to light (DIN 53 388, 53 389; ISO 877/R879) and radiation (DIN 53 750), and delamination resistance of laminates.

5.7. Processing Properties

Processing properties include heat-sealing, welding, and adhesion characteristics (see Section Welding and Heat Sealing of Packaging Materials). For the adhesion of packaging materials, aqueous adhesives that solidify on drying and, increasingly, hotmelt adhesives that vitrify on cooling are employed. A parallel requirement for satisfactory bonds is adequate wetting of the packaging material in the liquid state, together with strong adhesion and cohesion of the adhesive itself after hardening. For cardboard, the creasing and cracking characteristics (reduction in bending stiffness in the area of a fold) play an important role. Many of the properties outlined in Sections General Mechanical Data, Barrier Properties, Thermal Properties, Biological Properties, Electrical Properties, Chemical Properties are also significant with regard to processing.

[Top of Page]

6. Types of Packaging Materials

Norbert Buchner

The relative proportion of various types of packaging material has remained essentially unchanged over a number of years in the Federal Republic of Germany. A significant growth has been recorded only for plastics. In a total manufacturing value of packaging materials amounting to 17×10^9 \$ (1986), paper, cardboard, and corrugated cardboard together account for ca. 40 %, polymers ca. 28 %, metals ca. 22 %, and glass almost 8 % [13]. In the United States, the proportion of paper, cardboard, and metal is higher, and of plastics lower, than in Germany. Accordingly, a considerable increase in the share of plastics is expected [14].

6.1. Paper

Papers are differentiated by:

1. weight per unit area (generally between 20 and 150 g/m^2);
2. type of raw material (sulfate or sulfite cellulose, wood-free or wood-containing, primary fiber or recycled paper);

3. surface (machine-coated, one-side calendered, i.e., postsmoothed, calendered, coated); and
4. special type (simple packing paper, chipboard, natronkraft paper, kraft liner paper, tissue paper, crepe paper, microcrepe paper, wet tensile paper, greaseproof paper).

The proportion of secondary fiber materials (recycled paper) in the Federal Republic of Germany amounts to almost 50 % of all packaging paper.

Paper is suitable for packing foods that do not require much protection, such as sugar and salt. Similarly, paper is used for dried cereal, potatoes and rice products, where water-permeable packaging is necessary so that in the event of a reduction in temperature, and resulting increase in relative humidity, the threshold marking the onset of mold growth is not exceeded (often incorrectly referred to as “breathing packaging”).

The term *special paper* includes silk papers (20 g/m^2) and the relatively greaseproof and aroma-impermeable variations such as artificial parchment (consisting of a cellulose structure that has been compressed by grinding rollers), glassine (in which the surface is subsequently smoothed by calendering), and vegetable parchment (crude paper that has been broken down in sulfuric acid, neutralized, and cleaned). This last type is used frequently for relatively odor- and greaseproof inner pouches or wrappings. In several applications, however, it has been replaced by coated paper.

6.2. Cardboard

Cardboard is a fibrous material which has normally been couched together from numerous layers while still wet. An exception to this is cellulose cardboard, which is manufactured as a monolayer from cellulose fibers. Multilayer cardboard is also available, such as chromo and imitation chromo cardboard (made from primary materials with a wood-free top layer) as well as duplex or triplex cardboard (incorporating a reverse side or an intermediate layer of gray secondary materials). All these types of cardboard are also available with a lacquered surface [15]. The proportion of secondary fibers present in cardboard in the Federal Republic of Germany is approximately 85 %.

Cardboard is used for food-packaging applications similar to those described for paper. In addition, cardboard packets can be subjected to mechanical loads, stacked, or designed to meet advertising requirements as outer casings for pouches, tubes, boxes, etc.

6.3. Corrugated Cardboard

Solid cardboard materials have been replaced in a variety of applications by material-saving corrugated cardboard, which has been developed over the past hundred years [16]. Single-side corrugated cardboard (one layer of corrugated paper with a top layer, usually bleached on one surface) is often employed as a padding or cushioning material in packaging fragile articles, such as pastry and cookies. Singly corrugated cardboard (a single layer of corrugated paper incorporating a top layer on both surfaces) is the typical material used in food packaging to transport packets, trays, and fruit boxes, e.g., where the packets are formed by subjecting the articles to wraparound processes. When heavy articles are involved (e.g., for “bag in box” applications with liquids), multiple corrugated cardboard is employed. The proportion of corrugated cardboard manufactured from secondary fibers is as high as 95 % in the Federal Republic of Germany.

6.4. Coated Paper and Cardboard

Paper and cardboard are coated with waxes to impart water vapor impermeability and render them waterproof. Additionally, heat-sealing properties can be gained by coating with hotmelts (mixtures of waxes and polymers); with polyethylene (see Section Polyethylene) which is watertight in addition to that; and ethylene ionomers (see Section Ethylene Ionomers) which are especially well heat sealable and generate a barrier to fats and aromas. Similarly, coatings based on poly(vinylidene chloride) copolymers (extremely strong barrier properties with respect to odors, fats, and water vapor, along with heat-sealing characteristics), rubber latex (cold-sealing properties), and fluorinated hydrocarbons (grease and fat resistance), are relatively common. Cellulose cardboard coated on both surfaces with polyethylene is employed in large quantities for containers used in packaging milk and other dairy products, as well as products suitable for freezing. The inner polyethylene layer provides the necessary liquid barrier to contain the product, whereas the outer coating ensures the hydrophobic qualities necessary to prevent any problems with condensed water. By additionally incorporating aluminum foil, the aseptic packaging of products such as milk, fruit juice, still beverages, and more recently, tomato concentrate, apple sauce, and soup, has been made possible. Special paper that includes a coating of partially crystalline thermoplastic polyester can be employed for packaging and baking trays suitable for use at oven temperatures of 220 °C. Duplex papers with an intermediate layer of wax, hotmelts, or polyethylene have excellent adhesion properties.

6.5. Plastics for Packaging

The average costs (1988) of different types of plastics used in food packaging are given in the following material.

Plastic	Costs, $/kg
ABS	1.81
Acrylonitrile (barrier plastic with high acrylonitrile content)	1.76
EVOH	5.30
HDPE	0.77
LDPE	0.75
LLDPE	0.77
PA 6	3.53

PA 66	4.19
Amorphous PA	3.75
Oriented PET	1.32
PP	1.04
PVC	0.73
PVDC	2.54
Polycarbonate	4.08
Polystyrene	1.15

6.5.1. *Cellophane (Regenerated Cellulose) (⟶)*

Cellophane [9005-81-6] was originally used for packaging applications in 1924 and coated with a heat-sealing layer for the first time in 1929 (Du Pont). Since about 1970, cellophane has been increasingly replaced by alternative products, particularly heat-sealable, oriented polypropylene film. This development is quite advanced in both the United State and Japan [17], [18]. Because of its adaptability, however, cellophane will remain irreplaceable in certain sectors in the future.

The advantages of cellophane lie in its transparency and gloss, as well as favorable barrier properties toward many aromas and fats. Its inherent brittleness is overcome by hygroscopic, moisture-retaining agents (polyhydric alcohols); the properties, however, remain strongly dependent upon environmental conditions; especially the humidity of the environment. This can be reduced by coating with various, normally heat-sealable, materials [nitrocellulose, vinyl copolymers, poly(vinylidene chloride) copolymers], which improve the barrier properties of cellophane to odors, water vapor, and gases and enable films to be heat-sealed.

This gives one special advantage to cellophane: the base film is not thermoplastic, and the coatings allow heat sealing to be carried out over a wide temperature range. Another advantage is that optimal choice of coating allows barrier properties to be tailored to suit specific requirements. The greatest barrier properties are achieved with poly(vinylidene chloride) (PVDC) lacquers or coatings. Cellophane is also employed as a base film in laminates, where again it is facing strong competition from oriented films based on polyester (O-PET), polyamide (OPA) and polypropylene (OPP).

Applications of coated cellophane products include the packaging of baked goods, snacks, cheese, coffee, candy, and spices. Cellophane is also used as a wrapping to improve appearance.

6.5.2. *Polyethylene (⟶ Polyolefins)*

Based on figures for the 1980s, polyethylene (PE) represents approximately two-thirds of the total usage of plastics in the packaging industry. The various products available are normally broken down according to their density (⟶ , ⟶ ; ⟶ Polyolefins). Polymer properties can be varied by mixing the components.

Alternatively, ethylene can be copolymerized with polar comonomers. The resulting polymers include EVM [24937-78-8] (ethylene – vinyl acetate copolymer), EAA [9010-77-9] (ethylene – acrylic acid copolymer), and EMA [25103-74-6] (ethylene – methacrylic acid copolymer). Apolar comonomers include butene, hexene, octene, and methylpentene which are present in low concentrations in LLDPE (linear low-density polyethylene). By mixing, particularly the homopolymers, a multitude of products is available.

Polyethylene is very inexpensive (see Section Plastics for Packaging), chemically stable, relatively easy to weld, tough with a large strain potential, and insensitive to reduced temperatures. It can be employed over a wide temperature range. Additionally, it is an excellent barrier to water vapor, but only average with regard to the permeability of gases, aromas, and fats. Barrier properties increase with increasing density, as do the degree of crystallinity, strength, stiffness, hardness, and melting temperature. Contrary to this, the toughness, impact strength, transparency, sensitivity to reduced temperature, and resistance to stress cracking decrease with increasing density.

Low-density polyethylene [9002-88-4] (LDPE) is the least expensive product and is used most frequently for films, coatings, and laminating purposes. Linear low-density polyethylene is increasingly employed as a replacement for LDPE because of its increased strength and toughness (which allows a reduction in thickness), the increased temperature range for welding, and the better hot-tack properties. Typical applications include sealing layers in laminates, freezer bags, shoppery bags, sacks for heavy goods, and stretch films.

High-density polyethylene [9002-88-4] (HDPE) is employed for applications where increased impermeability to fats (meat packaging etc.) and aromas, increased stiffness (injection-molded articles such as crates), or increased temperature stability (milk bottles that can withstand sterilization in an autoclave) is important. The special high molecular mass polyethylene (HMPE) is used mainly for small, stiff carrier bags.

Ethylene – vinyl acetate copolymers with vinyl acetate contents below 10 % are also used in packaging applications. Compared to polyethylene homopolymer products, these copolymers show increased strain properties, transparency, stability to reduced temperature, and welding properties (lower melting point). Typical applications include sealing layers in laminates, freezer bags, elastomeric stretch films, and sacks for heavy goods.

Ethylene – acrylic acid copolymers with acrylic acid contents up to 10 % are similarly employed in food-packaging applications. These copolymers can be welded over a wider temperature range than the homopolymer and also show increased hot tack properties (see Section Welding and Heat Sealing of Packaging Materials), strength of welded seams, adhesion to other materials, and resistance to stress cracking.

Ethylene – methacrylic acid copolymer (EMA) has a lower melting temperature as well as improved welding properties and adhesion in laminates. The elongation at break is increased, albeit at reduced strength, and the sensitivity to low temperatures is extended to –76 °C.

Applications. The main applications for homo- and copolymers of ethylene, with a share amounting to ca. 70 %, are associated with films. Extrusion coating and laminating amount to 7 %. Films are used as wrapping, as materials to produce bags and sacks, as shrink and stretch films and as heat-sealing layers in laminates. Many substrates are extrusion coated with PE, e.g., cardboard (see Section Corrugated Cardboard), paper, cellophane, and oriented films based on polyester, polyamide, and polypropylene. During this process, barrier layers can also be included to permit the manufacture of packaging that is impermeable to water vapor, gases or odors, that can be sterilized by using heat, and that can be hot- or aseptically filled (see Section Barrier Polymers).

Coextrusion with similar or dissimilar polymers is continually gaining in importance. With dissimilar polymers, the objective is usually to include reinforcing or barrier materials, whereas with similar polymers, an improvement in processing properties is generally sought. Typical examples include composites of stiff, highly heat-resistant HDPE with the excellent welding LDPE (or, possibly, LLDPE or ethylene copolymers). Similarly, composites with an LLDPE core and surface layers of EVM are common as stretch films with autoadhesion ("cling") properties [19-21].

6.5.3. *Ethylene Ionomers*

Ethylene ionomers are copolymers made from ethylene and acrylic acid, in which the latter is partially present as either the sodium or the zinc salt. The polymer thus contains both covalent and ionic bonds. Ionomers have a major disadvantage compared to polyethylene because they are expensive and can be used only within a narrow temperature range. In contrast, they have the advantage, depending upon variant, of introducing further improvements, such as excellent adhesion to other substrates (useful laminating agents with good delamination resistance), increased hot tack with simultaneous low melting temperatures, (see Section Welding and Heat Sealing of Packaging Materials) (this leads to increated welding speed and safety), improved toughness and perforation strength, as well as favorable fat impermeability.

6.5.4. *Polypropylene (⟶ ; ⟶ Polyolefins)*

Polypropylene [9003-07-0] (PP) films have been available since the beginning of the 1960 s. Like polyethylene, polypropylene is one of the least expensive mass-produced polymers.

As with PE, the water vapor barrier properties of PP are good (lying between those of LDPE and HDPE). However, it has only average impermeability to gases and aromas. Polypropylene acts as a better barrier to fats than LDPE. It also has a much higher melting temperature (165 °C) than PE and can be employed up to ca. 140 °C. When crystallization is hindered by quenching, its transparency is also considerably greater than that of polyethylene.

Standard PP film (homopolymer) is rarely used for packaging purposes because it is only stable down to 0 °C. A typical application is for packaging bread. Copolymers with ethylene are employed to improve the stability to reduced temperature, welding properties (by reducing the melting point), thermal stability, and toughness, particularly of welding seams. Typical applications include injection-molded containers for frozen foods such as ice cream, heat-sealing layers for sterilizable packaging substrates, and steam-sterilizable cups and menu trays that can be sterilized in an autoclave and warmed in a microwave oven. More recent developments include multilayer bottles and cans incorporating a barrier layer that can be hot-filled or sterilized in an autoclave.

The *orientation* of polypropylene, mostly biaxial, (⟶), improves its mechanical strength, impermeability, transparency, and stability to reduced temperature (to –50 °C); it also permits shrinkage. *Monoaxial orientation* is employed particularly for tapes (straps, tapes for woven sacks). Uncoated *biaxially oriented* PP film (OPP) is an inexpensive base film for laminates.

Coextruded OPP films have taken a major step in replacing two-sided heat-sealable, coated OPP film (e.g., with PVDC, to increase barrier properties to gases and aromas) and cellophane. The coextruded fims, with surface layers between 2 and 3 µm thick, consist of copolymers with ethylene or terpolymers. The outstanding properties include high transparency, stiffness, strength, and—given the limited temperature loadability—good processing characteristics [22], [23].

6.5.5. *Poly(Vinyl Chloride) (⟶ ; ⟶ Poly(Vinyl Chloride))*

Poly(Vinyl chloride) [9002-86-2] (PVC) is a material with very well-balanced properties; its costs are in the medium range; it offers a relatively good barrier to gases, aromas, and fats; and frequently, it alone is sufficiently impermeable to water vapor. Additionally, PVC is highly transparent and glossy, has adequate strength and stiffness, and has excellent thermoforming properties with a wide thermoelastic range (80 –150 °C). Whereas several PVC products soften as low as 60 °C or 70 – 75 °C, modifications also exist that can be hot-filled at 80 – 85 °C (e.g., blends with chlorinated PE).

Films of PVC are manufactured mainly in a calender process, but they can also be produced by extrusion. The advantage of PVC films lies in their high elastic strain and high oxygen permeability, which retains the desired red coloration of oxymyoglobin in packaged meat .

Other major applications are the thermoforming of containers and trays for cooking oil, baked goods, and snack articles; sorting trays for chocolates; and the manufacture of corresponding heat-sealed coated lids. In several cases, replacement by PP has already begun. Biaxial orientation, improves the impermeability and strength of PVC. Such films can be employed as decorative shrink-wrapped films [24], [25].

Bottles made of PVC—also biaxially oriented—are used particularly for mineral water and cooking oils. Here, replacement by polyester is underway.

The use of poly(vinyl chloride) has been criticized on several occasions because it is manufactured from carcinogenic

monomers and because of its combustion products, which include chlorine gas and, if water is present, hydrochloric acid. Additionally, PVC is suspected of contributing to the formation of dioxins. The use of stretch films containing plasticizers, to package fresh meat and fruit in trays, is disputed as well. In this case, to minimize the transfer of materials present in the film, approved film thickness and applications have been regulated. Direct contact with foods of high fat content is to be avoided [26].

To pose no health hazard, the content of vinyl chloride monomer should be limited to less than 10 ppb in foods and 1 ppm in PVC. In incinerators, ca. 50 % of the chlorine content is originated from PVC; the exhaust gas washing stages (which are employed during the incineration of all materials) capture the hydrochloric acid. Combustion trials involving excess amounts of PVC have proved that dioxin is not formed during correct incineration. More recent investigations carried out in the Kernforschungszentrum in Karlsruhe (Federal Republic of Germany) have suggested that dioxin formation can be attributed to inorganic chlorine compounds.

6.5.6. *Polystyrene and Styrene Copolymers (⟶ Polystyrene and Styrene Copolymers)*

The following styrene polymers and copolymers are used for packaging purposes:

1. standard polystyrene (PS),
2. styrene – butadiene copolymer (SB, impact-modified polystyrene),
3. styrene – acrylonitrile copolymer (SAN), and
4. acrylonitrile – butadiene – styrene copolymer (ABS elastomer, modified SAN).

Additional variations and blends are available.

Standard Polystyrene. The advantage of PS [9003-53-6] lies in its great stiffness combined with average cost, along with favorable processing properties during injection molding and thermoforming (110 – 160 °C). A disadvantage is its high permeability to gases and vapor. Hence, polystyrene is typically employed as a packaging substrate for the short-term storage of relatively insensitive products, often under refrigerated conditions, e.g., dairy products that do not have a high fat content (yogurt, cottage cheese, ice cream, cream). Polystyrene is also used in sorting trays for fruit, eggs, baked goods and candy. Standard polystyrene is highly transparent and stiff, although it is also brittle and prone to stress cracking. Biaxially oriented products are employed within limits. Copolymers can be hot-filled up to approximately 80 °C.

Styrene – Butadiene Copolymers. Depending on the butadiene content, the stiffness, and transparency of SB copolymers [9003-55-8] decrease while impact toughness increases. A block copolymer with a fine distribution of the polystyrene or polybutadiene phase is crystal clear. The complex manufacturing process required leads to high costs. Styrene – butadiene copolymers are used as materials for cups and containers for the packaging of dairy products.

Styrene – Acrylonitrile Copolymers. Relative to standard PS, several properties of SAN copolymers [9003-54-7] are improved, including stability to oils and aroma substances, the upper limit of application temperature range, and gas barrier properties. An increased acrylonitrile content makes the polymer a suitable barrier toward gases.

Acrylonitrile – Butadiene – Styrene Copolymer. The ABS copolymer [9003-56-9] has high impact strength but is not transparent. Thermoforming properties, stability towards fats, impact toughness, and stability to reduced temperature (as low as –40 °C) are all favorable. Because of its high cost, the polymer has limited use as a packaging substrate, e.g., as a component in coextruded multilayer films [27-30].

Polystyrene and styrene copolymers with PE as an inner barrier for steam, along with additional barrier and tie layers, are coextruded to a thermoformable composite film possessing excellent impermeability to water vapor, gases, and odors. Such composites are increasingly employed for the manufacture of single-portion containers under aseptic conditions (e.g., for desserts) and permit a storage time up to one year at room temperature.

6.5.7. *Thermoplastic Polyesters (⟶ ; ⟶ Polyesters)*

The term thermoplastic polyesters refers to a group of high-quality polymers of which poly(ethylene terephthalate) [25038-59-9] (PET or PETP) is the most significant. Additionally, poly(butylene terephthalate) [24968-12-5] (PBT) and glycol-modified PET (PETG) are also important in packaging.

Properties. At present poly(ethylene terephthalate) occupies the mid- to high-price range, although the long-term trend is toward lower prices due to increased application. Films made of PET represent an average to good barrier to gases, aromas, and fats, and possess adequate to good impermeability to water vapor. Poly(ethylene terephthalate) is a high-strength material that can be employed over a wide range of temperatures (–60 to 200 °C in short-term loading, e.g., during heat sealing). Poly(ethylene terephthalate) is partially crystalline; its strength and temperature stability increase with increasing crystallinity. A disadvantage of polyesters is the relatively low glass transition temperature (PET: 73 – 79 °C; PBT: 36 – 49 °C), which means that in a highly amorphous condition, the materials are unsuitable for applications involving hot filling. This property can be obtained however by crystallization at elevated temperature in a forming mold (PBT crystallizes more rapidly), as well as by blending or coextruding with a material of higher thermal stability.

Areas of Application. Thermoformable, unoriented films are available, including a heat-seal layer when required, but these materials do not play a significant role in packaging. Tubular polyester films are used as baking films.

Biaxially oriented PET films (⟶ Films – Polyester) account for a very important range of products. Orientation improves transparency, strength, and impermeability. High short-term thermal loadability has resulted in O-PET film (most frequently, 12 µm thick) becoming one of the most important base films for the manufacture of impermeable, flexible composites, for use over a wide temperature range. Because long-term thermal loadability is guaranteed to >150 °C, these can also be employed as base films in the production of composites capable of being sterilized in an autoclave.

Polyesters find further application as baking trays for hot-air ovens at temperatures ranging from 200 to 220 °C. Such articles are manufactured from crystallizable PET or from special cardboard that has been coated with PBT.

The utilization of biaxially oriented bottles, wide-neck containers, and cans made from polyester is increasing. Bottles for sparkling beverages were developed at the beginning of the 1970 s, particularly for 1.5- and 2-L sizes, because of the weight problem and danger associated with breakage of equivalent glass containers. Meanwhile, PET is used for bottles of other sizes to contain commodities such as water, cooking oil, and alcoholic beverages. For packaging beer and wine, improvement of gas barrier properties is necessary. This can be attained by a PVDC coating on the outer surface or by coextruding a barrier layer of amorphous polyamide [31-34].

6.5.8. *Polyamides (⟶ ; ⟶ Polyamides)*

From the approximately 20 different polyamide (PA) products presently available, the most important in terms of packaging are PA 6 [25038-54-4], PA 12 [24937-16-4], PA 6/66 [32131-17-2], PA 6/12, and PA 11 [25035-04-5] (numbers correspond to the number of carbon atoms between NH groups). While PA is partially crystalline, PA 6/66 and PA 6/12 have the lowest degree of crystallinity. An increase in the degree of crystallinity increases stiffness and impermeability but reduces transparency.

Properties. Polyamides are high-quality polymers in the mid- to high-price range. Depending on the degree of crystallinity, they vary from highly to relatively transparent; they have high strength and toughness, average hardness and stiffness, and are relatively impermeable to gases, aromas, and fats. Because of their hydrophilic character, polyamides constitute only an average barrier to water vapor. Melting point varies, according to product, from 175 to 255 °C; stability to low temperature varies from –50 to –70 °C.

Applications. Polyamides, as monofilms, are of little importance because the rapid transition to a low-viscosity melt during welding hinders processing. Tubes made from PA 12, or coextruded from PA 6 and PA 12, are used for frankfurter-type sausages. The primary use of polyamides is in the manufacture of laminates, particularly in combination with polyolefins (PE, ethylene copolymers, PP) which complement the properties of PA with favorable water barrier and welding characteristics. Combinations with unoriented PA, including additional barrier layers, are often employed as thermoformable laminates in tray-type packaging (under vacuum or incorporating a protective gas) for meat products, fish, and cheese.

Biaxial orientation leads to improvements in impermeability, strength, stiffness, and transparency. Oriented polyamide films are important base materials for laminates incorporating a heat-seal layer and, possibly, further barrier layers. Such laminates are used in vacuum and protective gas packaging for coffee, powdered milk, meat products, etc. Another application is as the inner bag in “bag in box” system for liquids [33], [35].

6.5.9. *Polycarbonates (⟶ ; ⟶ Polycarbonates)*

Polycarbonates (PC) [25037-45-0] are relatively expensive and, because they are amorphous, crystal clear. They are stable as low as –90 °C and as high as 135 °C (for sterilizing); however, they are permeable to gas and water vapor. Investigations are underway to determine the possible use of PC for hollow packaging, in combination either with barrier layers or with thermally sensitive structures to increase the heat stability in applications involving pasteurization or heat sterilization.

6.5.10. *Barrier Polymers*

Barrier polymers are highly impermeable to gases and usually also to flavors [36-40]. This field is developing rapidly because in some cases, barrier polymers can partially replace glass and sheet metal. Examples are poly(vinylidene chloride) copolymer, poly(vinyl alcohol) (PVAL) ethylene – vinyl alcohol copolymer, high-barrier polyamide, and polyacrylonitrile (PAN) blends with high acrylonitrile content. Products from poly(amide – imide) are also being developed. Figure 1 shows oxygen permeability ranges for some of these plastics.

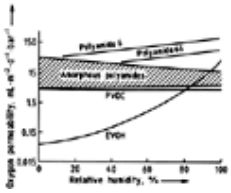

Figure 1. Oxygen permeability range of barrier polymer membranes at 30 °C (membrane thickness = 25 µm)

Poly(Vinylidene Chloride) Copolymers (⟶ ; ⟶). Copolymers of PVDC [9002-85-1] were the first barrier polymers. They are partially crystalline thermoplastics with a composition of at least 50 wt % vinylidene chloride (VDC), copolymerized with other comonomers such as vinyl chloride (VC), acrylic acid methyl ester (methyl acrylate), methacrylic acid methyl ester (methyl methacrylate), acrylonitrile (AN), and unsaturated carboxylic acids. They are employed as aqueous dispersions, lacquers, or extrudable granulates. The melting point of pure PVDC lies above its decomposition temperature; therefore, comonomers must be copolymerized into extrudable systems to bring about a reduction in melting temperature. In these cases, the content of VDC is ca. 75 – 90 %. In dispersions and lacquers, it can reach 94 %. Barrier properties increase with increasing VDC content.

The advantage of PVDC is the extremely low permeability of gases, water vapor, aromas, and fats under dry and moist conditions. Particularly high barrier properties are possible with more recent products that include AN as a comonomer. Disadvantages include the relatively difficult processing characteristics, the chlorine content, and associated problems in recycling production scrap with mixed polymers. Poly(vinylidene chloride) copolymer dispersions are used to coat paper, cardboard, cellophane, OPP, PET, and PA films, and to improve the barrier properties of PET bottles. Lacquers can be similarly employed as heat-sealable barrier coatings on such films. The plastic is extruded to form monofilms and barrier layers in flexible films, as well as thick thermoformable composites.

Poly(Vinyl Alcohol) and Ethylene– Vinyl Alcohol Copolymers (⟶ , ⟶ ; ⟶ Polyvinyl Compounds, Others). In ethylene – vinylalcohol (EVOH) [25067-34-9] copolymers, the moisture sensitivity of poly(vinyl alcohol) (PVAL) [9002-89-5] is reduced by copolymerization with ethylene. The vinyl alcohol content ranges from 66 to 82 %. A variety of specialty products are available which can be processed be extrusion techniques, e.g., with increased barrier properties, improved thermoformability, and greater resistance to fracture on bending.

Depending upon the ethylene content, which generally lowers barrier properties, EVOH copolymers have far superior gas impermeability than PVDC determined under dry conditions. A disadvantage for use in humid conditions or in packaging aqueous liquids is that EVOH is strongly hygroscopic, and water uptake causes a drastic increase in permeability. The extrudability and recyclability of mixed production scraps are an advantage.

Ethylene – vinyl alcohol copolymer is present in coextruded flexible, high-barrier composite films (e.g., PA – EVOH – PE) and also in thick thermoformable structures for containers and trays, which can be heat-sterilized and are suitable for use in microwave ovens. It is also employed as a barrier layer in multilayer bottles. For packaging moist articles, a water vapor barrier layer (thick polyolefin) is selected for the inner surface, along with an outer surface of a less impermeable material (a layer permeable to water vapor or a thin PP layer) to keep the EVOH as dry as possible.

Amorphous Amorphous Polyamides (⟶ Polyamides). A new group of materials still in the development stage and not yet approved for applications involving foodstuffs, the new amorphous polyamides have all the advantages of PA 6 but also are highly transparent and have favorable stiffness properties at high humidity levels. They can be processed by injection molding, extrusion, blow molding, and thermoforming techniques because a well-defined crystalline melting point is no longer present. Relative to partially crystalline PA materials, the gas, water vapor, and aroma barrier properties are increased, while gas permeability under humid conditions decreases slightly. Because of their high dimensional stability at elevated temperature, amorphous polyamides are particularly suitable for applications involving hot filling or heat sterilization. Current development is directed toward applications involving films and uni- or multilayer containers. In the last case, amorphous PA would govern the barrier properties and also improve the high-temperature dimensional stability. In Japan, wine bottles have been manufactured which consist of a PET – amorphous PA – PET structure.

Poly(Acrylonitrile) Multicomponent Blends. At the beginning of the 1970 s, a range of blends with high AN content (e.g., 70 – 95 %) were developed with properties including gas impermeability, transparency, high strength and stiffness, favorable processing characteristics (injection molding, extrusion, blow molding, thermoforming), and thermal stability up to 90 °C. Areas of applications involved the packaging of fruit juices, alcohol, beer, cooking fat, and sauces. These materials are accepted by the European authorities. In 1977, the U.S. Food and Drug Administration (FDA) banned the material because of possible health hazard due to the migration of AN monomer. Since 1984, however, PAN has been accepted in the United States, albeit with defined limits relative to the monomer components (1 ppm).

A recent blend of PET and EVOH from Japan can also be included in a list of barrier materials. This may mark the onset of a new developmental route: generating barriers not by clearly defined individual layers but by integration into a single layer. Processing will undoubtedly be simplified by this trend.

6.5.11. *Foamed Plastics (* ⟶ Foamed Plastics*)*

The foaming of polystyrene (expanded polystyrene, EPS) can be attributed to developments at BASF in 1951. Expanded polystyrene films are used in thermally insulating drinking cups, trays for fresh meat, and packaging for dairy products. They are also employed as padding materials in the form of boxes, matrices, inserts, and inert filler chips [41-43]. Like EPS, foamed PE films are also important for padding.

6.5.12. *Polymer Composites*

In many applications, a monofilm cannot meet the demands made of packaging substrates. Hence, to reach a desired property profile, several materials are combined. This can be achieved by:

1. introducing additional layers onto base film by coating with lacquers, melts (polymers, hot-melts), or dispersions;
2. laminating of webs which have been manufactured previously (pastes, lacquers, reactive adhesives, polymer melts, hotmelts, waxes); or
3. coextruding different polymers. The last leads to increased process economy (utilization of expensive components as thinner layers, incorporation of reclaimed material) and savings in subsequent processing (see Sections Polyethylene), as well as improved thermal behavior (see Section Thermoplastic Polyesters) and barrier properties (see Section Barrier Polymers) [44]. At present, thick thermoformable composites are manufactured via coextrusion with up to six separate layers and a maximum width of 1.8 m.

6.5.13. *Aluminum-Polymer Laminates*

Today, aluminum foil for packaging applications can be rolled to thicknesses as low as 7 µm. Besides the discontinuous rolling of blocks, a continuous process involving casting and rolling is used. Typical foil thicknesses lie between 7 and 15 µm. Depending upon the alloy, freedom from pinholes is possible at thicknesses greater than 15 – 18 µm. After being rolled, most foils are annealed at ca. 400 °C. The extension at break, which is only 2 – 3 % with thinner foils, increases with increasing thickness to a value of 25 – 30 %.

Aluminum foil that is merely coated with a protective lacquer is rarely used for other than decorative applications (wrapping for chocolate, candy, cheese). After being coated with sealing layers, foils can also be employed as film lids for cups. In the absence of pinholes, aluminum foils are impermeable to gas, vapor, aromas, liquid, fat, visible light, and UV or IR radiation. Coatings and polymers cover any pinholes so that with suitable aluminum – polymer composites, extraordinarily high barrier properties can be obtained. If adhesive layers are ignored, high-barrier laminates often have a three-layer structure. The outer surface consists of the mechanically strong base film made of O-PET, and OPA OPP, or cellulose. The aluminum foil is in the middle, and the inner surface represents the heat-sealing layer (usually PE, ethylene copolymers, or PP). With correct

property matching between the aluminum and polymer films, as well as high-quality laminating, the danger of failure at the grain boundaries of the aluminum crystallites caused by mechanical loading is reduced. In fact, the crystallites may be elongated. Shallow containers, which can be sterilized at 121 °C, are manufactured by drawing such laminates (see Section Thermoforming of Polymers). These are used along with semirigid trays made of aluminum from 70 – 150 µm thick (with a heat-sealing layer), which are manufactured similarly by using thermoforming techniques [45], [46].

Aluminum tubes are manufactured by compression. The inner surface is lacquered with a material that prevents contact with the food; the outer surface is printed and lacquered. Their advantage over polymer tubes results from the fact that aluminum tubes are gastight and free of any recoil effects; this means that after the product is removed from the tube, air is not drawn into the free volume. Such tubes are used to pack mustard, mayonnaise, and tomato concentrate. An interesting variation is in tubes made of a high-quality printed PE – Al – PE laminate; this is wound in a parallel manner and sealed on the longitudinal overlap seam, so that the aluminum cut face on the inner surface is protected by PE that is squeezed out of the seal. The body of the tube is normally equipped with a PE shoulder piece and sealed with a screw cap.

6.5.14. *Metallized Polymers*

Metal evaporation onto paper and PET films in a high vacuum was originally employed with wound capacitors. This discontinuous process has been transferred to packaging materials such as paper and films (O-PET, OPA, OPP, PC, PVC, and cellophane). The layer thickness is generally less than 0.1 µm. The attainable barrier properties (e.g., toward gases) depend not only on the thickness of the layer but also on the morphology and, consequently, on deposition conditions.

Packaging substrates with vacuum-deposited aluminum layers are used for decorative purposes (packaging for candy, bottle labels, sorting trays for chocolates), and also as the barrier layer in laminate structures having relatively high gas impermeability (e.g., for roasted coffee and snack such as nuts, potato chips, or crackers). The barrier properties of laminates incorporating aluminum foil are not achieved; However, the films are often less prone to damage caused by flex cracking.

6.5.15. *Polyester Films with Vacuum-Deposited Silicon Dioxide*

Vacuum deposition of silicon dioxide onto PET films, which was developed in Japan, provides high gas barrier properties equivalent to those of EVOH measured under dry conditions. In contrast to EVOH, however, the permeability does not increase with increasing humidity.

6.6. Metal Cans

Initially, cans were made in three parts, from tin-plated steel; the longitudinal seam was folded and soldered with tin solder (containing lead), which was normally coated with an epoxy or an acrylic resin. Today, this sector has become considerably more diversified: sheet steel is increasingly being manufactured by casting into continuous bars and rolling out to a minimum thickness of 0.14 mm. Electrolytic tin plating has displaced annealing methods, which allows savings of tin amounting to approximately 80 %. The thickness of tin, which often varies according to whether the inner or outer surface is being prepared, is between 2.8 and 11.2 g/m^2. Apart from tin-plated sheet steel, tin-free steel (TFS), meaning sheet steel with a surface of chromium – chromium oxide, is being used increasingly for food packaging. Cans with welded longitudinal seams, as well as two-part cans (either singly or doubly drawn, or, alternatively, drawn and ironed to a minimal thickness of 0.09 mm) have been introduced.

Apart from the precoating of sheet-metal webs, the coating of finished cans—either completely or only in the seam area—is widespread. For reasons of environmental conservation, lacquers containing organic solvents are supplemented by coating compositions in which the solvent has been either partially or fully replaced by water. Additionally, reactive systems have been developed, which are cured by using UV or electron-beam techniques, along with powder coatings that are applied electrostatically and subsequently fused.

Cans made from aluminum sheet play a major role in the manufacture of shallow containers, for example, for packing fish and sensitive meat products. Beverage cans are manufactured from aluminum as well as iron-based metals. A number of products exist which represent alternatives to tin cans. For packaging dried products, combination cans with cardboard as a structural element in the casing and barrier layers made from aluminum foil or plastic can be used. The casing is wound in either a parallel or a spiral configuration, and metal lids are normally attached by rolling. Cans made from plastics (e.g., PP) are also available.

For packaging fruit juices, special composite cans exist in which the casing and the lid are heat-sealed; the cans are filled either while still hot or under aseptic conditions (see Section Aseptic Packaging). The walls of these composite cans are made of PE – cardboard – PE – Al – PE (starting from the outside layer, the lids consist of PVC – Al – PE, or cardboard – PE – Al – PE). For sparkling drinks, two- and three-part cans have been developed with a body or casing made from PET with rolled-on metal lids.

For sterilized products, a two-part can has been developed in the United States, with a body made of PP – EVOH – PP composite and a rolled-on metal lid. Apart from this, a newer can manufactured in Sweden is available in which the body and both end pieces are made from PP – Al – PP. Further competitors include sterilizable trays made from PP that incorporates intermediate polymeric barrier layers, as well as aluminum foil coated with PP [47-50].

6.7. Glass Bottles and Wide-Neck Containers

Blown-glass bottles for beverages, salad oil, sauces, and vinegar have been manufactured since about the middle of the 19th century. In the past two decades, careful analysis of this process, optimization of conditions, and willingness to accept certain compromises in the shape of the bottle (partly to promote standardization and interchangeability) have made possible a weight reduction of up to 50 %.

The press-blow process permits the manufacture of lighter narrow-neck bottles with relatively constant wall thickness and in

the production of wide-neck containers, weight reductions greater than 50 % are possible.

The surface of the bottles is modified to avoid their premature fracture due to notch effects caused by the presence of microcracks in the surface. During the hot end-treatment, the outer surface is coated with tin chloride or titanium chloride which reacts with glass and hardens between 500 and 600 °C. In the subsequent, final, cold end-treatment, which is essentially intended to increase frictional properties, aqueous solvents or dispersions of organic materials (e.g., waxes and plastics) are sprayed onto the bottles at temperatures between 80 and 150 °C. Because material transport to the inner surface cannot be excluded, these materials must conform to Recommendation XLVIII of the Health Administration of the Federal Republic of Germany (BGA). Bottles with shrinkable sleeves made from polystyrene foam or shrinkable labels made from PVC film, along with other polymer coatings applied by immersion techniques, are available to meet the following requirements: labeling, decoration, noise insulation, protection against shattering, or prevention of surface damage.

Because glass bottles are completely airtight, gas permeation is possible only through the cap or, when metal caps are involved, through the seal that is pressed against the mouth of the bottle. This seal is usually made from plasticized PVC or PE. For crown corks, screw caps, and vacuum lids, minimum oxygen transmission rates of 0.004 cm^3/d (23 °C) have been reported [51]. Due to the danger of tampering, caps incorporating safety devices, such as tear-off or break-off rings, are being employed increasingly or even stipulated in some countries for certain applications.

Wide-neck glass containers for preserved foods, marinades, salads, and instant powders are manufactured according to the optimized press-blow process. Screw cap, twist-off, and press-on lids made from metal, with polymeric seals, are employed as lids. For packaging dry products, that are sensitive to water vapor, plastic screws and spring caps, including sealing membranes which are glued or heat-sealed to the lip of the glass, are also used.

Great effort has been made to seal wide-neck containers by using plastic-coated aluminum membranes. Pretreatment of the container surface in the vicinity of the mouth is a prerequisite if the seal on the hydrophilic glass surface is to be stable to liquids. Sealing equipment must be able to take into account any irregularity of the glass surface [52].

Bottles and wide-neck containers made from glass are being replaced by plastics. Bottles made from PVC for salad oil, vinegar, and mineral water, as well as sterilizable bottles made from HDPE for milk, have been used since the beginning of the 1960s. Since 1975, carbonated beverages are also available in PET bottles, either with or without an outer PVDC coating. Newer developments include gastight bottles with barrier layers made from EVOH or amorphous PA within a structure of sterilizable PP or PET (see Section Barrier Polymers) [53], [54].

[Top of Page]

7. Packaging Machines

Horst Weisser and Heinrich Vogelpohl

According to DIN 8741 Part 1, *packaging machines* are those machines which carry out either completely or partially the processes involved in packaging (e.g., forming, erecting, filling, sealing, and wrapping), as well as those processes prior or subsequent to the actual packaging steps which nevertheless contribute in some way to the manufacture of packages ready for distribution, storage, and retail [55]. Machines which are involved only in the manufacture of packaging materials or packaging accessories, and do not form an integral part of a packing line (e.g., machines to manufacture bottles, cans, tubes, folded boxes, etc.), do not constitute packaging machines.

A breakdown of packaging machines according to the technology involved is shown in Figure 2.

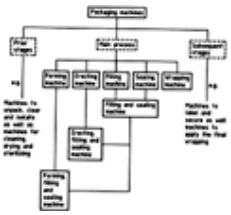

Figure 2. Breakdown of packaging machines

Normally, a packaging machine completes many sequential functions. A typical multifunctional packaging machine can, for example, form a tubular packaging substrate from a film material wound on a roll, fill this with the product, and then generate a filled tube by sealing in the transverse direction before cutting to produce individual bags or pouches. The individual tubular pouches can subsequently be packed in folded boxes which, in turn, are collected in containers for transportation. The term *packaging line* is applied to such a system [55]. A packaging line can also include machines that are not themselves packaging machines, e.g., closing and sealing units or weight control aggregates.

Whereas general packaging technology represents an extremely comprehensive area ranging from the packaging of tablets to packing bulk containers, food-packaging technology is limited to so-called retail packaging. According to DIN 55 405, Part 5, this is described as packaging in which the design is governed by the requirements imposed during the distribution of the goods and the selling to the final consumer [56]. As a rule, food packaging deals with complete packaging, meaning packaging that has been stipulated by law, involving certain commodities and certain capacities, which is closed and sealed in the absence of the final consumer. The contents of the commodity incorporated in the complete packaging cannot be altered without opening or noticeably tampering with the packaging.

7.1. Unit Operations in Packaging Machines

A special characteristic of many fully automatic packaging machines is the discontinuous mechanism. Within one machine

cycle, the packaging substrate must be formed, filled, and closed. Theoretical studies have been published regarding the movement in drives during a discontinuous operation in machines typically employed in the packaging industry [57]. Complex mechanical processes often take place in packaging machines; on the other hand, the machines are designed in many cases for processing highly specialized packages or products. Therefore, the mechanical functions should be classified logically in terms of unit operations. If a vertical tubular bag machine is used as an example, these unit operations can be described as follows (Figure 3):

1. Unwinding the packaging substrate from a roll.
2. Generating a constant web tension by means of a dancing roller system.
3. Forming the flat packaging substrate into a tube by using a forming shoulder.
4. Sealing the packaging substrate in the longitudinal and transverse directions.
5. Removing the tube with the aid of grips or conveyor belts, depending on the dimensions of the bag.
6. Dispensing the products to be filled into the tube of packaging material either gravimetrically or volumetrically.
7. Sealing the filled tube.
8. Separating the packing into individual bags by means of a guillotine.

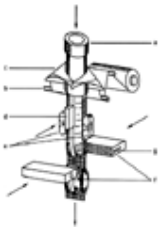

Figure 3. Vertical tubular-bag machine

a) Charging pipe; b) Deflector roller; c) Forming shoulder; d) Longitudinal sealing jaw; e) Conveyor belt; f) Transversal sealing jaw; g) Cross-cutter

The routines associated with all packaging machines can be broken down similarly into individual steps. During various stages, the machines come in contact with the packaging material as well as the contents. In most cases, these interactions are of a mechanical nature. Partially the operations are carried out pneumatically or hydraulically.

During these unit operations, many disturbances can occur. For example, with some packaging materials [e.g., cellophane (see Section Cellophane (Regenerated Cellulose)) or cardboard (see Section Cardboard)], mechanical properties vary with the environmental humidity. In these cases, the processing stages of the machine must be adapted to meet processing conditions. During the processing of plastic films, electrostatic effects often appear as unpredictable disturbances. These effects hinder not only the packaging process but, in many cases, filling the packaging as well. Particularly difficult products include ground coffee and polished rice.

7.1.1. *Overwrapping Articles*

The term *overwrapping* refers to a packaging process in which articles of a given shape are wrapped by using a thin planar packaging material and the shape of the article determines the shape of the package itself. In contrast to other forms of packaging overwrapping utilizes almost without exception two-sided, heat-sealable films, preferably films of coated cellophane, polyethylene (see Section Polyethylene), biaxially oriented polypropylene (see Section Polypropylene), poly(vinyl chloride), and cellophane laminates (Section Aluminum-Polymer Laminates). The packaging materials employed must possess a high degree of stiffness, must have a low tendency to revert to the original shape, and must be sealable in short sealing times. A special case of overwrapping is twist overwrapping which is employed mainly by the confectionery industry for packing candy. The packaging materials used need not be sealable; however, they must have good twisting properties and low propensity to untwist.

The most important unit operations and the possibilities of varying them to obtain a smoothly running machine are described in [58]. Here, unit operations including unwinding the packaging material, adjustment of web tension, packaging material takeoff, cutting, forming, and heat sealing are described. Because the properties of the packaging material are essentially fixed and can be influenced to only a small extent by the processing stages in the machine, mechanical processing techniques must by optimally adapted to the packaging material being used.

7.1.2. *Welding and Heat Sealing of Packaging Materials*

The term *welding* is used to describe the bonding of thermoplastic packaging materials by means of heat, either with or without pressure. In contrast, *heat sealing* is used when only the thermoplastic surface layers of the base materials are bonded under the influence of temperature and pressure. The base materials themselves are not plasticized.

A knowledge of the sealing properties of packaging materials is of major importance in the packaging industry. The welded seams should be airtight and strong, and ideally should form within a fraction of a second. The necessary melting energy is provided either by heating or by using ultrasonic techniques. With heated contact processes, the energy is transferred from heated sealing jaws (see Fig. 4) via the external layers and the bulk of the packaging material to the actual sealing layers on the inner surfaces of the substrate. With ultrasonic techniques, the melting energy is introduced by using a high-frequency horn; the necessary heat is generated by the intrinsic friction.

Figure 4. Cross section of a permanently heated sealing jaw

a) Jaw; b) Cutter; c) Cartridge heaters; d) Thermoelement; e) Channels for cooling air; f) Teflon coating; g) Top ejector

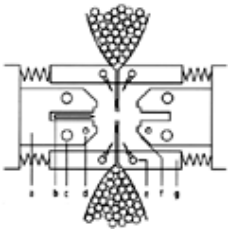

The actual sealing time, sealing pressure, and sealing temperature or, in the case of ultrasonic sealing, the intensity of the high-frequency signal are determined by the construction and adjustment of the machine. In addition, the material and the surface finish of the jaws are important. The quality of the seal is judged by the tensile strength, the hot tack (i.e., the tensile strength measured immediately after the sealing process), the degree of air tightness of the seal, and the optical impressions. The hot-tack values are greater with ultrasonic sealing than with heated contact techniques [58]; as far as the tensile strength of the cooled seal is concerned, the sequence is reversed. Operation at optimum sealing pressure is considerably more important with ultrasonic than with heated contact techniques.

To obtain high production rates, the sealing times must be reduced. With heated contact processes, this is achieved by increasing the sealing temperature; however, this is possible only within the limits imposed by the thermal stability of the packaging material. According to theoretical and experimental analysis, damage to seals is induced at excessive sealing temperatures by the generation of bubbles, discoloration, and creases. The packaging substrate may also adhere to the surface of the welding jaws [59]: Simultaneous increase in sealing temperature and decrease in sealing time lead to a reduction in the seal strength of most packaging materials. If extremely low sealing times (0.05 s) are combined with increased temperature, the seal strength must be expected to be dependent on sealing pressure.

Hot-Tack Properties. The term *hot tack* refers to the strength, and hence the loadability, of the joint while it is still warm, immediately after the sealing jaws are released. All heat-sealing and bonding processes are based on the same principle: the heat-seal layers (or, similarly, hotmelt adhesive) are compressed in a melted condition, causing the substrate to fuse and form a single homogeneous layer. The initial adhesion in the melt viscous state is very low; hence, bond damage is likely to be induced during sealing if the seal is loaded immediately after release of the heat-sealing jaws. Such loading can be caused by adhesion of the packaging material to the sealing jaws, by the weight of the product during the filling of tubular pouches, or even by the mechanical resistance of the packaging material itself (e.g., in wrapping machines). The damage results in either a partial opening of the bond or even an immediate failure of the seal. Latent seal damage can be recognized in the cooled seal by the reduction in bond strength or by leakage at the seams. Latent damage leads to subsequent separation of the seal and, hence, failure of the packaging.

The hot-tack behavior of a seal is influenced not only by the properties of the packaging material but also by the actual sealing conditions. Accordingly, hot tack can also be defined as the cohesive and adhesive strength of a thermoplastic material in the temperature region of the heat-sealing process. The various methods described to measure hot tack can be divided into those that monitor the force required to cause separation and those that measure the separation itself [60]. Data from different tests cannot be compared directly, and no standardized test procedure has yet been described.

7.1.3. *Cold-Sealing Packaging Materials*

According to DIN 55 405, Part 6, cold sealing or pressure-sensitive adhesion describes the bonding of substrates, typically packaging materials, whose surfaces are either completely or partially coated with an adhesive that causes bonding only after the two layers have been compressed [61]. Cold-sealing techniques were applied to food packaging in the early 1970 s, after the development of adhesives that were both suitable for machine processing and approved for use with foodstuffs. In terms of processing, the technology that had been employed in heat-sealing applications was utilized; the temperature of the sealing jaws was simply reduced, or the heat was switched off. The relationships between the effects of processing parameters (sealing pressure, temperature, time and jaw design) and the influence of adhesive composition, base material, and storage conditions on the quality of the seal are only partially known to date.

The starting material for all cold-seal adhesives is natural rubber latex. To coat packaging materials intended for use with foods, a modified latex together with other polymer dispersions is employed. Furthermore, protective colloids which increase the stability to heat, oxygen, light, and mechanical stress are added. The mechanism of adhesion of a cold-seal material can be explained by the freedom of movement of the linear molecular chains in the surface of the cold-seal layer: under the influence of jaw pressure, these become entangled with chains in the opposing surface and thus form a bond.

Investigation into the relationship between seal strength and mechanical or processing conditions has shown that seal strength increases with increasing sealing temperature and pressure [62]. Similarly, seal strength is a function of the thickness of the applied adhesive. The surface profile of the sealing jaws is particularly important; smooth surfaces are undesirable. An improvement in seal strength can be obtained by increasing the sealing time.

7.1.4. *Cutting Packaging Materials*

After the incoming web of packaging materials is unwound from a roll, the cross-cutter stations (see Fig. 5) of the packaging machine cut it into the individual sections required to manufacture the final packaging. The wear mechanisms that occur in the packaging machine during cutting must be known if slitting processes are to be reliably controlled. Shear cutting of packaging films by using straight blades (“open-drawing – intersecting cutting”), according to DIN 8588, has proved to be particularly important in packaging machines [63]. In investigating blade wear in such cross-cutters, the packaging material was found not to lead to blade wear but, in fact, to exert a wear-inhibiting effect [64]. Optimal blade service life can be achieved only by correct adjustment of machine parameters, angle of mesh, dihedral angle, and cutting pressure. An

oscillating relative movement between web and blade during the intervals separating the individual cuts is especially important for long blade service life.

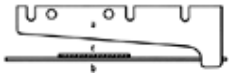

Figure 5. The two knives of a cross-cutter station

a) Revolving knife; b) Bed knife; c) Packaging material

7.1.5. *Web Tension in Packaging Materials*

The machinability of flexible packaging materials is strongly influenced by the web tension on the line. This is particularly true during the forming, filling, and sealing stages of processing machines for tubular bags. In such machines, the packaging material experiences high mechanical stress as it is drawn over the forming shoulder. The takeoff forces that are generated can be considered a measure of the processing quality. Material properties including *bending stiffness, dynamic coefficient of friction,* and *tensile strength* exert the strongest influence on the machinability of a film on a tubular-bag machine [65]. Bending stiffness plays the most decisive role of all.

Thin packaging materials with a *bending stiffness* up to 10 mN m^2 m^{-1} pose few problems in takeoff. Problems arise with materials of bending stiffness greater than 80 mN m^2 m^{-1}. Measures to reduce the takeoff forces include applying a powder to the packaging material or coating the machine components with Teflon. With machines that operate in cycles, webs of materials with large bending stiffness can easily break during processing, because the takeoff forces do not exist continuously but arise suddenly and generate high peak values of tensile stress. As a rule, the increased tensile strength associated with high bending stiffness cannot compensate for the greater peak loading to which the material is exposed during takeoff. To avoid any problems, the tensile strength of the packaging material should be approximately five times the expected takeoff forces.

The running properties of the packaging material, however, also depend on the adjustment and construction characteristics of the machine. The web tension generated is influenced by the line speed, the length of the bag, the geometry and material characteristics of the forming shoulder, and the design of the dancing roller system.

7.1.6. *Thermoforming of Polymers*

Thermoforming of packaging materials is used especially in the milk-processing industry. As shown in Figure 6, the packaging machine is fed with a roll of substrate that is drawn through a heating zone (contact or radiation heating) and passed to the forming station in which bowls or cups are molded via mechanical – pneumatic drawing. The material, which is firmly clamped in the forming station, is expanded and the wall thickness decreases in proportion to the relative increase in surface area. The characteristic dimension for the forming process is the so-called deep-draw ratio $H : D$, defined as the relationship between the height of the article H and the diameter D. The attainable deep-draw ratio is material dependent and also limited by the desired minimum strength of the package. Consistent wall thickness is difficult to obtain, which is of utmost importance because the thinnest area in the packaging has the greatest effect upon its gas barrier properties and mechanical loadability during stacking. Because the container must be removed from the form after thermoforming, the walls must be conical, with a taper between 1 and 2°. With typical packaging substrates, thermoforming temperatures lie between 110 and 170 °C. Immediately after thermoforming, the containers are filled and sealed by a film lid.

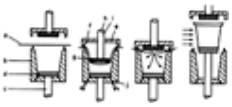

Figure 6. Principles of thermoforming

a) Film; b) Deep drawing die; c) Piston; d) Ejector; e) Cutting plunger; f) Cutting edge; g) Stretching plunger; h) Air channel; i) Plunger; j) Air vent

Comprehensive empirical investigations of thermoforming are discussed in [66]. Various quality criteria are important: these include expressions of the processing versatility, the deep-draw ratio relative to the drawing temperature and process, the attainable sharpness of the edges as representative of flow characteristics, and the wall thickness distribution. During thermoplastic deformation, an isotropic biaxial stretch is a prerequisite if the container is to demonstrate the necessary impact strength. The relationship between deep-draw ratio and draw temperature is characteristic for each type of polymer; it also depends upon the film thickness and the drawing process. The optimum drawing temperature for PVC is 110 °C and for PS, 140 °C. Deep-drawing properties of films, determined under laboratory conditions, should always be confirmed by machine trial carried out under production conditions; as a rule, larger temperature deviations can be expected because of the increased film widths and multicavity forms involved in scale up.

Deep Drawing of Aluminum Laminates. Film laminates that include aluminum foil are frequently employed in food packaging because the aluminum layer imparts a barrier to gases, flavors, and light (see Section Aluminum-Polymer Laminates). On thermoforming such laminates however, problems arise due to the extremely low extension at break of the aluminum, which can thus be torn easily. Although an aluminum rod can have an extension at break of ca. 30 %, this is reduced to ca. 3 % for a 12-µm foil because of the grain structure of the aluminum. The strain properties of an aluminum – polymer laminate are governed by the thickness of the aluminum, the quality of lamination, and the extension at break of the polymer substrate. The lower the extension at break of the polymer, the higher is the potential elongation of the aluminum foil within the laminate. Depending upon the film substrate, strains between 5 and 25 % can be achieved with laminates to a 15-µm aluminum foil, without the formation of visually detectable pinholes [67]. Clearly, however, in comparison to polymeric films, only low deep-draw ratios are possible with aluminum laminates. For this reason, such packaging materials are used preferably to manufacture shallow dishes and bowl-shaped containers. The greatest drawing depth is obtained by using a rounded form, although rectangular containers are more economical because of the lower amounts of scrap generated during stamping. Depending on the construction of the laminate, the packaging container may be filled with hot material or, after filling, may be sterilized in an autoclave.

7.1.7. *Processing Folded Boxes*

During the processing of cardboard panels or collapsed folded boxes on binding machines, the erectile behavior of the cardboard is the decisive characteristic for determining machinability. The quality of the creases is primarily responsible for this behavior. During creasing, the areas corresponding to the edges of the finished box are exposed to shear forces that weaken the bonds between individual layers of fibers and cause ply separation on bending. Hence, when the box is being set up, a hinge effect is generated along the crease. Tests designed to judge the quality of creases have shown that the bending-moment ratio (i.e., the relationship between the bending moments of a creased and a noncreased cardboard sample) is a reliable estimate of machinability [68], [69].

7.1.8. *Closing and Sealing Boxes*

Boxes made of corrugated cardboard are often sealed automatically with gummed or adhesive tape. Gummed tapes are tapes made of kraft paper with a coating of water-activated gum. Today, the adhesive normally contains chemically modified vegetable glues. Adhesive tapes based on polymer films coated with an autoadhesive, generally a synthetic material, can also be used. The strength of the adhesive tape, regardless of type, is decisive in determining the mechanical load that a carton can withstand. The strength of a sealed box is determined by means of falling and stacking tests [70].

7.2. Product Dosage in Packaging Machines

The centerpieces of all filling plants are the dosage units or filling machines. According to DIN 8740, Part 2, filling machines are packaging machines that introduce the product—in precisely determined amounts—into the packages [71]. Solid and liquid products can be metered quantitatively according to weight or volume. An exception with solid materials is the dosage based on number of items. Accordingly, filling machines are divided into gravimetric filling machines, volumetric filling machines, and counting filling machines.

7.2.1. *Gravimetric Dosage*

Gravimetric filling machines are subdivided into *net weight filling machines*, which weigh out a predetermined amount before introducing the contents into the packaging substrate, and *gross weight filling machines*, which weigh out a predetermined amount together with the packing substrate during filling – with or without tare adjustment. Figure 7 shows a gravimetric filling machine. According to the type of monitoring characteristics gravimetric filling machines may be mechanical or electrical. In most cases, the products is fed to the weighing mechanism via a two-stream principle. Most of the contents come from a coarse stream, and the target weight is made up subsequently from a fine stream (see Fig. 7). The dosage precision increases with decreasing mass flow of the fine stream.

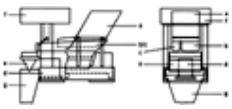

Figure 7. Gravimetric filling machine

a) Feed funnel; b) Coarse filling; c) Fine filling; d) Electronic dosage; e) Weighing tank; f) Controller; g) Discharge pipe

With discontinuous dosage, inaccuracies in the amount of material fed can occur due to inconsistencies in the transport mechanism in the weighing chamber [72]. Apart from this, dosage errors caused by the product sticking to the weighing cells are also possible.

7.2.2. *Volumetric Dosage*

Filling machines that operate on volumetric principles are divided into *dosage filling machines* and *level filling machines.* Dosage filling machines include measuring cup, displacement, piston, and rotary filling machines. Level filling machines, in which a certain dosage is determined by the packaging substrate, include vacuum, gravitational, and pressure filling machines.

The dosage capacity of a packaging machine is described primarily by the dosage rate and precision. Investigations of the bulk material behavior associated with measuring cup dosage incorporated into vertical tubular-bag machines are described in [73]. Filling time calibration curves, which illustrate the relationship between dosage size and required filling time, are helpful in determining product-specific properties when attempting to optimize the line speed of a packaging machine.

Auger filling is another means of volumetric filling. Here, the desired dosage is determined by the number of revolutions of a horizontally or vertically positioned metering screw. Auger filling is employed preferentially with materials that lead to high levels of dust formation but, at the same time, are not easily broken. Dosage inaccuracies can be caused by uncontrolled metering of the product by the screw or by the product sticking to the screw. In comprehensive studies of the technology and accuracy of auger dosage aggregates, both short-term precision and long-term errors were considered in judging various systems [74].

7.3. Robots in Packaging

According to VDI guideline 2860, industrial robots are described as universally applicable pieces of apparatus capable of movement in many axes, where the movement as defined by the order of movements or the distance and angle of movement is freely programmable (i.e., can be altered without mechanical modifications), and which in certain circumstances are controlled by sensors [75]. Robots can be equipped with grips, tools, or other equipment necessary for manufacturing purposes and can carry out handling or manufacturing routines. In the packaging sector, robots can be employed to fill packaging substrates with solid articles, to stack packed items, and to commission transport units. Product handling is carried out either mechanically with the aid of grip mechanisms or pneumatically with suction caps. Working cycles can be classified into those that are controlled by programming and those in which optical or acoustic sensors are involved. In the food-packaging sector, robots equipped with optical detector systems are used which can recognize chocolates being transported on a conveyor belt and individually place them in a tray. Packaging robots are particularly common where individual packages (e.g., plastic bottles) must be placed in boxes for transporting. In general, however, robot technology is rarely

employed in the packaging industry. The reason for this lies in the relatively low working speeds compared to conventional packaging aggregates, coupled with the high investment costs for machines and programs.

7.4. Aseptic Packaging

The principal aim in food packaging is to maintain product quality. To avoid thermal damage to foods during sterilization of the package, techniques have been developed over the past three decades which enable packages (originally of liquids) to be filled under aseptic conditions. This has been extended more recently to include partially solid foodstuffs. During aseptic packaging, the contents and package are sterilized in separate operations. Subsequent dosage and sealing processes are then carried out under aseptic conditions. The aseptic packaging of foods is currently one of the fast-developing areas of packaging technology [76]. Sterilization of the packaging material is a central step in aseptic packaging. It is normally carried out in hydrogen peroxide solution at increased temperature. Sterilizing processes involving ozone, sulfur dioxide, peracetic acid, heated air, and steam are also available which can be used for certain products as well as packaging substrates. The sterilizing effect of concentrated hydrogen peroxide solution was investigated as a function of concentration and temperature by using *Bacillus subtilis* spores [77]. Optimal sterilization was observed at hydrogen peroxide concentrations between 45 and 50 %.

Because in aseptic plants that function on a continuous basis, sterilization must be ensured within several seconds, the hydrogen peroxide solution has to be heated to 80 – 90 °C. An employer utilizing such aseptic packaging equipment is required to take the necessary measures to guarantee the health and safety of his employees. This involves determining the hydrogen peroxide emission from the system [78].

[Top of Page]

8. Interaction between Packaging and Food

Albert L. Baner, Rainer Brandsch, Otto Piringer

Today, food and packaging constitute an indivisible unit. Proper food packaging requires that the properties of the components — the food and the packaging — are combined in an optimal way. The interaction between packaging and food consists of the transfer of substances from the environment through the packaging and from the packaging into the food and vice versa (see Fig. 8). Chemical reactions inside the packaging or food are also possible. Not only can the quality of the packed food change as a result of such interactions, but the properties of the packaging may change as well, and this can lead to further loss of its protective function. Hence, the main criterion for an optimal packaging system is, in most cases, minimal interaction. General references are given in: [123-131].

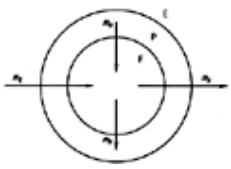

Figure 8. Mass transfer in packaged food

m_E= mass transfer from the environment (E); m_P= mass transfer from the packaging (P); m_F= mass transfer from the food (F)

8.1. Physical Interactions

Kinetic (diffusion) and thermodynamic (partitioning) processes are the most important physical interactions. In a diffusion process, the mass transport occurs in the direction of a concentration gradient as a result of random molecular motion. One can write the general equation for the mass transport by diffusion in the direction of the *x*-axis, according to Fick's 2nd law [132]:

$$\frac{\delta c}{\delta t} = D\frac{\delta^2 c}{\delta x^2} \quad (1)$$

where *c* is concentration, *t* time, *x* distance, and *D* the diffusion coefficient (constant).

Solving the diffusion equation gives the concentration contribution at each point in time and at each location on the *x*-axis. Simplifying approximations make analytical solutions possible.

8.1.1. *Permeation*

Under steady state conditions, $\delta c/\delta t = 0$ and, as in the case of permeation, the equation of diffusion reduces to Fick's 1st law [132], which expresses the flux *J* as the amount of matter passing through a unit area of a film in unit time:

$$J = -D \cdot \frac{dc}{dx} \quad (2)$$

A typical example for a packaged food is the permeation of gases into or out of the package. From Fick's 1st law one can obtain:

$$J = P \cdot \frac{\Delta p}{d} \quad (3)$$

where Δp is the partial pressure difference of the permeating gas between the exterior and interior of the packaging, *d* is the thickness of the packaging, and *P* is the permeability coefficient. If the solubility coefficient *S* in the packaging is defined as $S = c/p$ (the ratio of the concentration *c* of the gas in the packaging to the partial pressure *p* of the gas in equilibrium with the packaging), then $P = D \cdot S$, where *D* is the diffusion coefficient.

Many different units can be used to express *P*. Values can vary over several orders of magnitude, depending on the properties of the material. The values are tabulated in polymer handbooks and manufacturers literature [133]. For a given shelf life of a particular foodstuff, Eq. (3) can be used to calculate the thickness of packaging material and the barrier properties required for an allowable flux of gas into or out of the package (see Section Barrier Polymers). A frequently used example is calculation of the maximum water-vapor flux for a packaging containing a water-sensitive product.

8.1.2. *Migration*

Migration from Packaging into Food. Most interactions involving mass transfer between the packaging and food take place under non-steady-state conditions and are referred to as migration. Migration from the packaging (e.g., polymer film) into the food (e.g., a liquid) can be described by using Fick's 2nd law. As mentioned above, analytical solutions for the diffusion equation can be obtained under simplifying assumptions. In the simplest case it is assumed that the thickness of the packaging is infinite, and the solubility of the migrant in the well-mixed liquid is high. Assuming that the migration process is far away from equilibrium, one can obtain the simple expression:

$$\frac{m_{\mathrm{F,t}}}{A} = \frac{2}{\sqrt{\pi}} c_{\mathrm{P},0} \rho_{\mathrm{P}} \sqrt{D_{\mathrm{P}} t} \tag{4}$$

where $m_{F,t}$ is the amount of the migrated compound, *A* the contact area, $c_{P,0}$ the initial concentration of the migrant in the polymer, ρ_P the density of the polymer, D_P the diffusion coefficient of the migrant in the polymer, and *t* the contact time.

Equation (4) shows that the material transport from the packaging material into a well-mixed liquid is proportional to the initial concentration of the migrant in the polymer, the square root of time, and the square root of the diffusion coefficient. Most stringent is the timeframe (small *t* values) or, in other words, the limit of applicability beyond a mass transfer of more than ca. 60 %. If this conversion is exceeded, which often happens, and the system runs close to the equilibrium state, high overestimations or even aberrations of the calculated mass transfer into the liquid can occur.

To overcome this limitations one must use the more general solution of the diffusion equation [134]. It is assumed that at the beginning of mass transfer the migrant is homogeneously distributed in the polymer, there is no boundary resistance for transfer between polymer P and liquid food F, and the total amount of the migrant in P and F remains constant during the migration process. The diffusion of the migrant in the polymer is assumed to be several orders of magnitude slower than diffusion in the liquid phase (e.g., mixed during transport), and the diffusion coefficient of the migrant in the polymer is constant.

$$\frac{m_{\mathrm{F,t}}}{A} = c_{\mathrm{P},0}\, \rho_{\mathrm{P}} d_{\mathrm{P}} \left(\frac{\alpha}{1+\alpha}\right) \left[1 - \sum_{n=1}^{\infty} \frac{2\alpha(1+\alpha)}{1+\alpha+\alpha^2 q_n^2} \exp\left(-D t \frac{q_n^2}{d_{\mathrm{P}}^2}\right)\right] \tag{5}$$

$$\alpha = \frac{V_{\mathrm{F}} / V_{\mathrm{P}}}{K_{\mathrm{P,F}}}$$

where $m_{F,t}$ is the amount of migrated compound, *t* the contact time, *A* the contact area, $c_{P,0}$ the initial concentration of the migrant in the polymer, ρ_P, ρ_F the density of the polymer and the liquid food, d_P the thickness of the polymer, D_P the diffusion coefficient of the migrant in the polymer, V_P and V_F the volumes of the polymer and the liquid food, $K_{P,F} = c_{P,\infty}\rho_P/c_{F,\infty}\rho_F$ the partition coefficient [the ratio of the migrant concentration (w/v) in the polymer and the liquid food at equilibrium], and $\tan q_n = -\alpha q_n$ (q_n are the positive roots of the trigonometric identity).

Due to good mixing in the initial stage of migration, the total amount of the material is transported away from the contact layer into the volume of the liquid, and the concentration in the contact layer drops to zero. The rate of diffusion of the migrant out of the package is the rate-determining step. With longer migration times, the partition coefficient begins to play a deciding role through the α value. Especially for $\alpha \ll 1$ ($K_{P,F} \gg 1$), only a very small fraction of $c_{P,0}$ migrates into the liquid.

As an example, estimations were made for a conventional EC packaging (cube of 1 kg with a surface area of 6 dm^2), made from a high-density polyethylene (HDPE) film 50 µm thick containing 1000 ppm of the antioxidant Irganox 1076 (M_r = 531 dalton). The assumed test conditions are 10 d at 40 °C and the diffusion coefficient of the antioxidant in the HDPE matrix is $D_P = 2.9 \times 10^{-10}$ cm^2/s. The resulting migrant concentration in the foodstuff as a function of time is shown in Figure 9. It is evident that the estimation made with Equation (4) gives fair values for conversions below 60% of the maximum amount of 3 mg/kg. As the system approaches the equilibrium state an increasing overestimation is made. The estimation made with the more general Equation (5) gives precise results under the assumption that $K_{P,F} = 1$. The main reason for this discrepancy is the assumption of an infinite polymer volume made for Equation (4). Only by taking into account the finite volume of the polymer, as in Equation (5), can the correct migrant concentration in the foodstuff in the vicinity of the equilibrium state can be estimated.

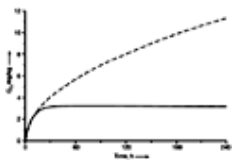

Figure 9. Migration estimation for the antioxidant Irganox 1076 (1000 ppm) into foodstuff from a conventional EEC packaging made from HDPE (50 µm). Dashed line: estimation made with Equation (4); solid line: estimation made with Equation (5).

For viscous or solid food the simplifying assumption of a well-mixed liquid must be abandoned, and the system, as well as

the possibility of finding an analytical solution for it, becomes more complicated. One must take into account the diffusion coefficient of the migrant in the solid foodstuff, which is often in the same range as in the packaging material. This has been done in a simplified way in the following expression derived from Equation (4):

$$\frac{m_{F,t}}{A} = \frac{2}{\sqrt{\pi}} c_{P,0} \rho_P \frac{\beta}{1+\beta} \sqrt{D_P t} \tag{6}$$

$$\beta = \frac{1}{K_{P,F}} \sqrt{\frac{D_F}{D_P}}$$

where D_F is the diffusion coefficient of the migrant in the food.

For $D_F \gg D_P$ and $K_{P,F} \leq 1$ we obtain Equation (4). If the diffusion coefficients in the packaging and in the food are approximately equal, the partition coefficient, $K_{P,F}$ determines the transport through the system. The packaging determines the rate of the whole process if the migrant dissolves much better in the food than in the packaging ($K_{P,F} \leq 1$) and the food determines the rate of the whole process, if the migrant dissolves much better in the packaging than in the food ($K_{P,F} \gg 1$). If $D_F < D_P$ the mass transport is determined by the diffusion coefficient in the food D_F and the partition coefficient $K_{P,F}$. This leads to the buildup of a concentration profile in the foodstuff.

An exact analytical solution of the differential equation that takes into consideration the diffusion in food and finite values for V_P and V_F is not available, and as a consequence the application of numerical methods is necessary.

Migration from Food into Packaging. In principle there is no difference in the mathematical treatment of the mass transfer from food into packaging compared to the reverse phenomenon described above. While in the initial phase the mass transfer of the migrant from the package into the liquid is independent of the partition coefficient, the migration from the liquid into the package is proportional to K. The partition coefficient plays a decisive role in the dissolution (sorption) of a migrant in the packaging layer in contact with the liquid. For high K values the total amount of sorbed material is concentrated in a thin layer of packaging material in contact with the liquid, and the transport process in the initial stage is independent of the material thickness.

Food packaging interactions include the uptake of major food components, such as fats and oils or water, by the packaging material. If high enough, such absorption can cause swelling and changes in the properties of the polymer. This swelling can lead to increased migration of additives from the polymer because their diffusion coefficient increases by several orders of magnitude [135]. Other consequences of swelling are structural changes that occur in the polymer, which may lower the mechanical properties of the packaging. The possibility of swelling is an important consideration in food packaging compatibility.

8.1.3. *Prediction of Diffusion Coefficients*

To use Equation (5) in practical cases data for two fundamental constants are needed: (1) the partition coefficient $K_{P,F}$ of the migrating compound between the polymer P and the foodstuff or simulating liquid F and (2) the diffusion coefficient D_P of the migrant in P. So far upper limits for migration amounts are needed from the regulatory standpoint; predictions of “worst-case” scenarios can start with the assumption of good solubility of the migrant in F, and consequently $K_{P,F} = 1$ can be used. Much more complicated is the prediction of an adequate value for D_P.

Much work has been done in correlating polymer structure with migrant structure to predict migrant diffusion in the polymer [136]. From a practical viewpoint, most of this work has limited value in predicting the diffusion behavior of substances in food packaging systems because of their complex nature.

In the food packaging sector polyolefins, especially polyethylene (PE) and polypropylene (PP), are the most widely used plastic materials. Because PE and PP have a rather low glass transition temperature T_g, it is legitimate to assume that under normal conditions of use the migration of substances in and from packaging made from these materials obeys the laws of diffusion. It has been shown that the migration in some non-polyolefin polymers such as polystyrene (PS), poly(ethylene terephtalate) (PET) and polyamide (PA) also follows the laws of diffusion.

To estimate upper limits of migration values from Equation (5), a simple semi-empirical formula for the estimation of D_P was developed recently [129], [137], [138]:

$$D_P = D_0 \exp\left(A_P - 0.1351\, M_r^{2/3} + 0.003\, M_r - \frac{10454}{T}\right) \tag{7}$$

where $D_0 = 1\ m^2/s = 10^4\ cm^2/s$, M_r is the relative molecular mass of the migrant, and T the temperature in kelvin.

The parameter A_P has the role of a “conductance” of the polymer matrix towards the diffusion of the migrant. Higher values of A_P in such polymers as PE lead to increased D_P values, while in stiff-chain polymers, such as polyesters and polystyrene, lower A_P values take into account the smaller diffusion coefficients for the same migrant. The equation can be used for $M_r \leq 4000$ dalton.

To calculate migration rates with a sufficient safety margin from the regulation standpoint (see Section Legal Aspects) it is

possible to match the parameter A_P in Equation (7) to yield a "worst-case" value A_P^*, which leads to a corresponding D_P^* instead of the real D_P. Due to the scatter of experimental values, the use of $D_P^* \geq D_P$ can avoid underestimations with a certain statistical degree of assurance (e.g., 95 %). The greater the scatter, the higher the upper limit for D_P^* must be set. This statistical assurance is a matter of convention and can be established from a regulatory point of view.

A consequence of the above consideration is a periodical change in A_P^* values depending on available data collections. A_P^* is a function of temperature: $A_P^* = A_P' - \tau / T$, where the parameter τ, with the dimension K, accounts for a specific contribution of the polymer matrix to the diffusion activation energy. Table 2 lists values of A_P' and τ for different polymers.

Table 2. Values of A_P' and τ different polymers

Polymer*	A_P'	τ
LDPE	11.5	0
HDPE	14.5	1577
PP	13.1	1577
PS	0.0	0
HIPS	1.0	0
PET	6.0	1577
PA	2.0	0

* LDPE = low-density polyethylene, HDPE = high-density polyethylene, PP = polypropylene, PS = polystyrene, HIPS = high-impact polystyrene, PET = poly(ethylene terephthalate), PA = polyamide.

For the polyolefins another semi-empirical diffusion model has been developed [139], from which it is it is possible to arrive at the following relationship:

$$D_P = a \cdot \exp\left(bM^{1/2} - k \cdot \frac{M^{1/3}}{T}\right) \; [\mathrm{cm^2/s}] \tag{8}$$

where *a*, *b*, and *k* are empirical constants determined from experimental diffusion data. The values for polyolefins are given in Table 3.

Table 3. Values for *a*, *b*, and *k* for Equation (8)

Polymer*	ln *a*	*b*	*k*
LDPE	– 4.16	0.555	1140.5
HDPE	0.90	0.819	1760.7
PP	– 2.10	0.597	1335.7

* For explanation of abbreviations, see Table 2.

The main limitation in estimating migrations for legislative purpose with the diffusion model of **Limm** and **Hollifield** is underestimation in some cases. Various programs are available for estimating the specific migration of packaging compounds [140-142].

8.1.4. *Multilayer Structures*

In Section Prediction of Diffusion Coefficients a practical situation is reduced to an idealized system in which the plastic material is a single-layer structure with a finite thickness and a homogeneously distributed migrant. Hence, the initial and boundary conditions for the analytical solution of the diffusion equation (5) are fulfilled. In the general case of structures with more than one layer or variable diffusion coefficients, only numerical mathematics lead to the desired result. Basically, this means restricting the solution of the diffusion problem to a set of grid points, conveniently distributed within the integration domain, and approximating the involved derivatives by discrete schemes [12].

The diffusion problem in a multilayer polymer system can be solved numerically, each layer $k = 1, 2, \ldots n$ having diffusion coefficient D_k, density ρ_k, thickness d_k, the laminate being in contact with a foodstuff medium of mass m_F and density ρ_F. The migration of a compound is governed by partial differential equations of diffusion for each layer *k*.

Under steady-state conditions the permeation through the multilayer structure is dominated by the laminate layer with the lowest permeability:

$$\frac{d}{P} = \frac{d_1}{P_1} + \frac{d_2}{P_2} + \cdots + \frac{d_n}{P_n} \tag{9}$$

where d_k is the thickness of layer *k* and P_k the permeability of layer *k*.

Therefore, the presence of a thin barrier layer with low *P* can determine the permeation characteristics of a multilayer structure. Figure 10, shows schematically the development of the migrant concentration with time in the foodstuff. Due to the barrier layer, the system needs a period of time, the time lag θ, until the steady state regime is reached and a linear dependence of the concentration with time is observed. The slope of the linear part of the graph is an indication for the efficiency of the barrier layer.

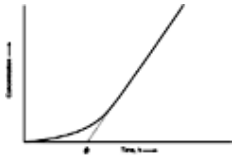

Figure 10. Schematic representation of the permeation through an multilayer structure with a barrier layer (θ = time lag)

8.1.5. *Non-Fickian Processes*

The preceding discussion holds true only if the diffusion and partition coefficients are constants. However, in some cases D_P and *K* can be functions of a concentration gradient in the polymer and can vary with time. If the migrant causes extensive swelling, diffusion can no longer be described by the concentration dependence of Fick's 2nd law [134], [135].

At very low migrant concentrations in the polymer, the migration rate may no longer be proportional to its concentration in the polymer. For example, active sites within the polymer matrix may tend to bind a migrant and thus decrease the possibility of its transfer from the polymer to the food-contacting phase.

In many cases, the observed transfer of additives is due not to migration but to removal of material from the polymer surface in contact with the food. During manufacturing of the packaging material, some of the additives are forced to the surface of the polymer. Subsequent contact with food, often in the initial period of contact (e.g., pasteurization), removes this surface layer of additives. After removal of this layer, a much slower diffusion-controlled migration occurs.

In the case of paper and board, physical interactions and the transfer of migrants to foods are controlled by sorption processes. Migration is governed by sorption isotherms, and the prediction of migration from paper materials is difficult because of their very different structures, compositions, and manufacturing processes.

8.2. Chemical Interactions

Much is known about reactions involving gases and water vapor with food (⟶ Foods, 2. Food Technology) [143]. From the packaging point of view, only the maximum allowable transmission of gases during the shelf life needs to be known to specify the package. With this knowledge, the maximum allowable package permeation rate (see Section Permeation) can be estimated.

8.2.1. *Production of Off-Odors*

Because of the complex structure and chemical composition of package systems, a variety of chemical reactions can occur during package manufacture and use. These reactions occur between some packaging components, whereby other components act as catalysts, to produce substances with low odor thresholds. In the manufacture of packaging materials, various processes can produce off-odors, for example, the surface treatment of polyolefins by corona discharge. These substances then migrate through the material during storage and lower the quality of the packed product. Such odor-producing substances need not to be formed during manufacture but may develop over the shelf life of the product (e.g., a continuous reaction taking place in the presence of a metal). These substances diffuse slowly into the product and their presence is not detected initially. However, defects and possible complaints arise after the product is on the market. Identification and analysis of these products is very difficult; many isolation and concentration steps are required because the odor thresholds are too low to be analytically detectable. Furthermore, a combination of instruments (e.g., GC-MS) coupled with sensory evaluation is necessary to identify and quantify these substances [129].

Unsaturated carbonyl compounds, particularly $C_6 – C_{10}$, are an important class of odor-producing substances because they occur in nature, have low thresholds, and are relatively stable due to their conjugated double bonds. Other oxygen- and sulfur-containing compounds and some hydrocarbons are important off-odor compounds.

8.2.2. *Other Reactions*

A list of important classes of reactions that can occur between food and packaging materials follows. This field of research is relatively new, and for this reason, in combination with the complexity of analysis and detection, very little information and few data are available on actual reaction products.

1. Hydrolysis or decomposition of additives migrating into food can occur. Little is known about these reactions because they are very complex and occur only at trace levels [144].
2. Reactions occurring at high temperature are becoming more important as packages are used in high-temperature applications in microwave and conventional ovens.
3. Chemical sterilization and irradiation of package materials can generate decomposition products [145].
4. Corrosion of metals and their reaction products are important factors in metal containers [146].

Despite the many possibilities for food packaging interactions mentioned above, in practice complaints and problems arising from these interactions occur relatively seldom.

8.3. Measurement Methods and Migration Limits

8.3.1. *Measurement Methods*

The purpose of migration testing is to ensure food safety and protect the consumer by controlling the extent of mass transfer of undesirable constituents from the packaging to the foodstuff. A distinction is made between migration testing, in which mass transfer is subject to the laws of diffusion and partitioning, realized under conditions in which the packaging material withstands food contact without altering its mechanical and diffusion properties, and extraction, in which the solvent penetrates the packaging material and changes its diffusive and mechanical properties.

The control of material transfer from plastic packaging materials into food is manly based on the measurement of substances in the food or a food simulant after certain specified and in most cases standardized contact conditions. Adequate migration cells made of glass or stainless steel are used. The packaging material is completely immersed in the liquid or contacts the simulant from one side. Conventional migration measurements keep a sample in contact with the food or simulant in a manner that represents the actual conditions of use. In alternative migration tests, the sample is kept in contact with an appropriate simulant in such a manner that a strong interaction between packaging and simulant occurs. By using shorter contact times and/or higher temperatures (more severe test conditions) at least equal or even exaggerated extents of migration are obtained.

Beside the geometric dimensions of a given food packaging system which influence migration, mass transfer from a plastic into food and vice versa is governed by kinetics (diffusion in the polymer and foodstuff) and thermodynamics (equilibrium partitioning between packaging and food; see Section Physical Interactions). It is common practice in migration evaluations to start as a worst-case assumption with a total mass transfer scenario in which the initial concentration of a migrant in the packaging is known. However, if this relatively simple calculation provides a scenario which exceeds a migration limit, then it is advisable and necessary to achieve further refinements to the total mass-transfer approach.

When selecting a method of analysis, some pre-considerations are obligatory. The volatility of the analyte must be taken into account, as well as the nature of the matrix to be analyzed and, in particular, the expected concentration range of the substance being measured. Generally, these methods should allow one to quantify the monomer or plastic additive at the required restriction limit in all relevant food simulants and/or in the polymer. It is generally recognized and accepted that analytical methods must be suitable for the intended use. The analytical procedures for compliance testing with food laws are to be carried out on the basis of validated methods as specified in ISO 5725 based on round-robin tests, with well-defined repeatability *r* and reproducibility *R*.

Validation and standardization of analytical methods is a recognized basic task of CEN. The Technical Sub-Committee CEN TC194/SC1 “General Chemical Methods of Testing for Materials Intended to Come into Contact with Food” is active in developing standardized analytical methods and specific migration test methods. There have been many efforts to standardize migration measurements as laid down in the EN/ENV Standards. The EN/ENV 1186 series is concerned with the testing required by the European Commission Directive relating to plastic materials and articles intended to come into contact with food (90/128/EEC), and is intended to give advice on the selection of the most appropriate test conditions and test methods for a given application of a plastics article [147]. Fully validated methods have been developed for 15 plastics monomers and published as European pre-norms in the ENV 13130 series [148].

The sample for testing should be taken from the final article in its ready-for-use state. Different cell types for one-side testing are described in Part 1 of EN/ENV 1186. The most commonly used test conditions are 10 d at 40 °C, which simulates conditions of actual use in which the contact time is longer then 24 h at 5 – 40 °C. More detailed conditions for overall and specific migration tests, are given in Parts 2 – 15 of EN/ENV 1186 and the EEC directive 82/711/EEC and its amendments (see Section Legal Aspects).

For specific migration testing there is a clear need for pragmatic and cost-effective solutions. To reduce the amount of analytical work without compromising the requirements of consumer protection the concept of functional validation and precision is introduced. This means that the number of validation steps to be applied and the amount of precise data needed is a function of the analytical demands and the relevance of the migrant concerned.

State of the art analytical methods for qualitative and quantitative determination of migrants use gas chromatography (GC), high-performance liquid chromatography (HPLC), mass spectrometry (MS), and combinations thereof.

8.3.2. *Migration Limits*

Any migration control procedure must be linked either directly or indirectly to the measurement or evaluation of the concentration of all undesired substances. This concentration represents the toxicologically relevant target parameter for any migration-control methodology. The toxicologically relevant molecular mass range for migrants is considered to end at a molecular mass of 1000 g/mol. Larger molecules are considered from a biophysical point-of-view not to be physiologically absorbable in the gastrointestinal tract. The toxicologically relevant target parameter can be measured, controlled, or assessed in different ways and must then be compared to relevant migration limits set by regulations such as the European Commission Directive 90/128/EEC and its amendments, the United States Code of Federal Regulations (CFR), title 21, parts 170 – 199, the Food Sanitation Law in Japan, or by nonmandatory but practically binding recommendations (see Section Interaction between Packaging and Food).

In the EC migration limits are set in terms of “global migration limits” (GML = 60 mg/kg food) and “specific migration limits” (SML, in µg/kg or mg/kg food) from a package into a foodstuff or food simulant under given contact conditions. SMLs are directly calculated from the toxicological characteristics of the specific substances, expressed as acceptable daily intake (ADI) or tolerable daily intake (TDI) in µg/kg body weight, by multiplying them by the factor 60 (the average weight of a European is assumed to be 60 kg). Because it is further assumed that one person consumes 1 kg of foodstuff per day, the SML value can be compared directly with the migration data if expressed in milligrams or micrograms per kilogram of food simulant or real foodstuff. The correlation between area- and mass-related migration data can be made by the accepted assumption that 1 kg food is contacted with 6 dm^2 of packaging material (conventional EC packaging). If the SML is set as “not detectable” in the foodstuff, the SML corresponds to the “limit of detection” (LOD) of the method indicated for the determination of the migrant, which is usually in the range of 10 µg/kg food simulant or foodstuff. (For further details see

Section European Community).

On the other hand, food packaging legislation or regulation has also set up restrictions on the packaging material side. Such compositional limits in plastics are given as “maximum quantity limits” (QML).

In the United States a somewhat different approach is used. Instead of calculating from the SMLs toxicological data and comparing them with migration data, the FDA uses migration data to calculate the “estimated daily intake” (EDI), which is then compared directly with the corresponding ADI or TDI values. The EDI is calculated by multiplying the concentration in the daily diet, based on “consumption factors” (CF) and “food type distribution factors” (f_T), by a factor of three (it is assumed that one individual consumes 3 kg liquid and solid food per day). Finally the FDA established a concentration in the daily diet of 0.5 μg/kg (ppb) as the threshold of regulation for substances used in food contact articles. (For more details see Section United States).

Migration of packaging components must be also controlled for sensory reasons. Residual solvents and other components from printing, laminating, and lacquering operations are examples of nontoxic but sensorially active substances. These substances can produce off-odors, and their limits must be established in accordance with their odor thresholds. Migration of these nontoxic substances is acceptable as long as their level remains under the sensory threshold. No official limits exist because they are too difficult to establish owing to the wide range of thresholds, which also depend on the type of food product. For unsaturated carbonyl compounds the odor threshold is 10^{-4} mg/m^3 and even lower for other substances [149]. Many factors influence the odor threshold in packaged foods, such as the partition coefficient, the ratio of the package area to the food mass, the temperature, and the composition and intensity of the aroma, to name a few [129]. In some countries, standards for testing the odor intensities of packaging materials have been developed [150].

8.4. Legal Aspects

8.4.1. *General Requirements*

The role of packaging in food legislation has been realized since the early 1900s. However, the importance of mass transfer of substances from the packaging to the food has been considered only since the 1950s, when the variety and complex chemical composition of packaging materials were recognized.

Realistic evaluation of the health risk associated with packaging is necessary. With the present state of the art, the amounts of substances transferred from the packaging to the food are so low that health hazards can be considered under the aspect of the chronic toxicity of the migrants. The risk-to-benefit ratio for pre-packed food is negligible compared to that for unpacked food. Even though this is unarguable, care must be taken to ensure consumer safety. Thus every regulation should consider the following points: 1) the transfer of substances from packaging into food is, in principle, undesirable and 2) unavoidable mass transfer in state-of-the-art packaging should not be prohibited if the substances do not constitute a health hazard and are technically unavoidable in an optimized package system. The principal optimization criteria for packaging are the protection of foods, consumer safety, environmental compatibility, and low cost. Meeting the above two points benefits the consumer and encourages technological progress.

To minimize mass transfer from packaging to food, the mass transfer must be quantified. Great uncertainty exists about setting limits on the allowable quantity of migration into foods, and the following points must be considered. Besides the acute toxicological properties of substance in packaged food, the chronic mutagenic and carcinogenic activity of transferred substances is most important. The application to humans of the results of no toxic effect level (NTEL) experiments on animals and the validity of subsequently derived ADI values are very uncertain. EDI levels have additional uncertainties arising from regional and temporal considerations. The kinds and amounts of food consumed vary with season, geography, and socioeconomic or ethnic background.

The amount of substances transferred from the packaging to the foodstuff can vary over a large range. The theory of diffusion can be used to estimate mathematically the quantity of material transferred (see Section Migration). However, many other complex factors affecting migration cannot be anticipated (see Section Non-Fickian Processes).

The permissible level of components from packaging in food are a function of the above considerations. As a consequence of these uncertainties, all substances can be divided into categories according to their health risk, which than can be used as a basis for regulation [128]. One end of the spectrum consists of substances that poses no risk (substances which are inert or naturally occurring and are toxicologically inactive), and at the other end are substances that must be excluded from packaging because they are recognized as toxicologically dangerous. Between this two extremes are the remaining substances, which require more or less detailed toxicological or migration studies from which useful limits can be derived. When the need to limit the use of a substance has been recognized, statistically reliable analytical methods are required which can be used by industry and control laboratories around the world. Once the limit for a substance is set, the means for maintaining this limit in practice must be determined.

8.4.2. *European Community*

The European Commission initially drew up a first framework directive 76/893/EEC, which was amended by a new framework directive 89/109/EEC and establishes two general principles:

1. The principle of the “inertness” of the material and the “purity” of the foodstuffs, whereby the materials and articles must not transfer any of their constituents to foodstuffs in quantities which could “endanger human health and bring about an unacceptable change of the composition of the foodstuff or deterioration in the organoleptic characteristics thereof”.
2. The principle of “positive labeling”, whereby materials and articles intended to come into contact with foodstuff must be accompanied by the words “for food” (or an appropriate symbol 80/590/EEC) [151].

The new framework directive laid down, in confirmation of the above-mentioned principles, the sectors in which the

Commission is asked to establish Community rules and the criteria and procedures to be followed in the drafting in specific directives. The Commission must, as far as possible, satisfy rigorous health criteria and thus consult the Scientific Committee for Food (SCF, set up in 1974 by a Commission Decision, is a independent advisory body consisting of leading figures from the fields of toxicology, nutrition, medicine, etc.). Specific directives and amendments to existing directives will be adopted by the Regulatory Committee procedure, in this instance the Standing Committee on Foodstuffs (CPDA, Council Decision 69/414/EEC).

First Framework Directive	76/893/EEC
New Framework Directive	89/109/EEC
Symbol Directive	80/590/EEC
Plastics: Directive on Vinyl Chloride Monomer (VCM)	78/142/EEC
Plastics: Directive on the Methods for Determining VCM in PVC	80/766/EEC
Plastics: Directive on the Methods for Determining VCM in Food	82/432/EEC
Plastics: Directive on the Basic Rules for Migration Testing	82/711/EEC
1st Amendment	93/8/EEC
2nd Amendment	97/48/EEC
Plastics: Directive on the List of Simulants	85/572/EEC
Plastics: Base Directive: Monomers	90/128/EEC
1st Amendment	92/39/EEC
2nd Amendment	93/9/EEC
3rd Amendment	95/3/EEC
4th Amendment	96/11/EEC
5th Amendment	99/91/EC
6th Amendment	2001/62/EC
Regenerated Cellulose	83/229/EEC
1st Amendment	86/388/EEC
2nd Amendment	92/15/EEC
Regenerated Cellulose Codification	93/10/EEC
Ceramics: Base Directive	84/500/EEC
Elastomers: Nitrosamines in Teats and Soothers	93/11/EEC

The list of authorized substances in Directive 90/128/EEC and following amendments concerns monomers and most types of additives, including substances which constitute the means of polymerization. Colors and catalysts are excluded from the Community lists. All of the more important information's on the evaluation of substances can be found in the document entitled “Synoptic” which is to be distributed on the Internet under http://cpf.jrc.it/webpack/.

The legislation in force provides for two types of restrictions. The first, applying to all substances, provides that they may not be released alone or together with others in quantities greater than the overall migration limit, which is designed on the one hand to impart a certain inertness to material intended to come into contact with food so as to guarantee its purity and on the other hand to avoid setting a special migration limit for each substance.

The second type of restriction, which applies to isolated substances, provides that they must not migrate in quantities higher than a certain value fixed in accordance with the toxicological characteristics of each term. The SML value is generally calculated on the basis of the ADI or TDI laid down by the SCF for these substances. SMLs established in this way contain inbuilt safety margins in the range of 100 – 1000 relative to the NTEL and range typically from the low ppb level (µg/kg) for specific substances of toxicological concern such as acrylamide (10 ppb), butadiene (20 ppb), acrylonitrile (20 ppb), and others up to the lower ppm level (mg/kg) for less toxic principals such as bisphenol A (3 ppm), ethylenediamine (12 ppm) or 1-octene (15 ppm).

The European Commission Directive lists QM limits in cases where the migrant is not chemical stable enough in food simulants or may be lost during migration testing due to its volatility. A typical example is 2,2-bis(4-hydroxyphenyl)propane bis(2,3-epoxypropyl) ether (BADGE) which is listed in 90/128/EEC with both a SML value of 20 ppb and a QM value of 1 mg/kg polymer.

For new substances to be included in the community lists the industry has to submit a special request accompanied by a technical file containing a set of data which will enable the SCF to evaluate the risk associated with the use of that substance. These data were laid down by the SCF in the “Guidelines of the Scientific Committee for Food”.

Migration Measurements. Because of the variety and complexity of chemical components in polymers, the directive 82/711/EEC was adopted to establish the general rules for migration testing. As a consequence of the analytical difficulty of measuring migration into food, standardized measurement procedures have been developed which employ food-simulating liquids with properties very close to those of foods, as well as contact time and temperatures. The following simulants can be used:

Simulant A: distilled water
Simulant B: 3 % (w/v) acetic acid in water
Simulant C: 10 vol % ethanol in water
Simulant D: distilled olive oil

The first amendment to Council Directive 93/8/EEC made the standard conditions for migration testing more flexible by allowing a greater number of possible combinations of times and temperatures and the use of other simulants for the “fat test”. A second amendment, Directive 97/48/EEC, laid down the conditions of use of volatile solvents, e.g., isooctane and ethanol, as testing liquids in the “fat test”. It provides that these solvents may replace olive oil or other fat simulants (HB 307, sunflower oil, etc.) if the fat test is not applicable in the basic version for technical reasons or in routine checking. The migration of substances under these simulation conditions should not exceed the limits given in the positive list.

Directive 85/572/EEC includes a table of correlations between food groups and their food-simulating liquids. For some food groups the migration into Simulant D is too high and must be divided by the reduction factor specified in the table.

Directive 85/374/EEC should be mentioned because it affixes the responsibility of interaction problems to all who are involved in manufacture of packaged food (i.e., packaging and food producers or fillers). The producers of packaged food are responsible for problems (e.g., interactions) arising from the use of their product if they cannot demonstrate that the state-of-the-art in science and technology was considered at the time of production.

8.4.3. *United States*

The laws governing food packaging in the United States is the Federal Food, Drug, and Cosmetic act of 1958, as amended, and the Fair Packaging and Labeling Act. These regulations are specified in the Code of Federal Regulations (CFR). The CFR is organized under 50 titles. Of practical importance for packaging is Title 21, which contains general regulations for enforcement of the Federal Food, Drug, and Cosmetic Act and the Fair Packaging and Labeling Act (CFR 21, Part 1). Regulations for current manufacturing, processing or holding of human food (CFR 21, Part 110) and for additives in packaging materials (indirect food additives) can be found in the CFR 21, Parts 170 – 199 [152], [153].

Regulations for food additives are established by petitioning the FDA for permission to manufacture food-packaging material. The petition generally contains the following information:

1. Identity of the additive/polymer.
2. Use.
3. Intended technical effect.
4. Migration and analytical methodology.

The petition should provide sufficient information to estimate the daily intake of the additive, i.e., consumer exposure. From analyzed or estimated levels of an additive in food or food-simulating solvent, the concentration of the additive expected in the daily diet is calculated. An estimation of the exposure to an additive in the diet is determined by combining migration data from food simulants or actual foods with information on the use of packaging that may contain this additive.

The consumption factor (CF) is used to describe the fraction of the daily diet expected to contact specific packaging materials. The CF represents the ratio of the weight of all food contacting a specific packaging material to the weight of all food packed. These values were derived from information on the types of food consumed, the types of food contacting each packaging surface, the number of food-packaging units in each food-packaging category, the distribution of container sizes, and the ratio of the weight of food packaged to the weight of the package. The minimum CF used for exposure estimates involving low-consumption applications is 0.05. This value may, however, be modified as new packaging information is received.

Before migration levels can be combined with CF values to derive estimates of probable consumption, the nature of the food that is likely to contact the packaging material must be known. Migration into a fatty food simulant, for example, will be of little use in estimating probable exposure if the packaging material is used exclusively to package aqueous food. To account for the variable nature of the food contacting each packaging material, food-type distribution factors f_T have been calculated for each packaging material to reflect the fraction of all food contacting each material that is aqueous, acid, alcoholic, and fatty. The total dietary concentration is calculated as follows:

$$\text{Dietary Concentration} = \text{CF} \cdot \sum_{i=1}^{4} (M_i f_T) \qquad (10)$$

were M_i is the concentration of the migrant in the *i*th food-simulating solvent (aqueous, acidic, alcoholic, fatty) expressed in mg/kg (ppm). Simulant concentrations are equivalent to food concentrations if the simulant volume to exposed surface area ratio is 10 g/in^2 (under the assumption that 10 g of food is in contact with 1 in^2 of surface area).

The threshold value of regulation is 2000 times lower than the dietary concentration at which the vast majority of studied compounds are likely to cause noncarcinogenic toxic effects and 200 times lower than the chronic exposure level at which potent pesticides induce toxic effects. FDA believes that these safety margins, which are larger than the 100-fold safety factor that is typically used in applying animal experimentation data to human (21 CFR 170.22), support a conclusion that substances consumed in a dietary concentration at or below 0.5 μg/kg are not of concern.

Migration Measurements. As a consequence of the analytical difficulty of measuring migration into food, standardized measurement procedures have been developed which employ food-simulating liquids whose properties are very close to those of foods, as well as contacting time and temperatures. The following simulants can be used:

10 % ethanol (aqueous and acid foods),
10–50 % ethanol (low- and high-alcohol foods),
Edible oil (e.g., corn oil), HB307, or Migliol 812 (fatty foods).

8.4.4. *Japan*

Food Chemistry and Administration is executed under the Food Sanitation Law [154]. This law applies not only to food and additives, but to food apparatus and containers/packages such as kitchen utensils, tableware, bottles, cans, and wrapping paper; even toys and detergents. Standards and specifications for apparatus and containers/packages were established under “Specifications and Standards for Foods, Additives, etc., Announcement No. 370” including additional establishments and partial amendments. All apparatus and containers/packages which come into direct contact with foods are governed by regulation of the Food Sanitation Law. Briefly “apparatus” means articles used for manufacturing, processing, and otherwise handling foods, and “containers/packages” means articles used for wrapping and packaging foods for delivery. The sale of toxic or injurious apparatus or containers/packages is prohibited by law (Article 9).

The specifications for apparatus and containers/packages are mainly divided into three types:

1. Specifications applying generally to apparatus and containers/packages and their raw materials. These specifications cover those containing metals and colors. Those concerning metals are targeted at tin for plating, solder, and products including copper, lead, and antimony. The specifications concerning colors applies to all products. Any color used for apparatus and containers/packages must be one designated by the Minister of Health and Welfare under Article 6 of the law. However, this does not apply where no color dissolves or migrates into foods.
2. Specifications for apparatus and containers/packages applying to individual materials. The law specifies six materials which become targets of specifications for apparatus and containers/packages. They are glass, ceramics, enamel, synthetic resins, rubber, and metal. Paper and wood are not specified as target materials of specifications. However, products made of these two materials are subject to Article 9 of the law (see above).
3. Specifications for apparatus and containers/packages according to use. This uses include containers and packages which are pasteurized under pressure, or for nonalcoholic beverages; apparatus used for manufacturing ice candies and other similar products; food vending machines and containers for selling foods from vending machines; or transporting apparatus or container/packages of stock solutions kept in cup-filling type vending machines or in automatic preparing machines.

Migration Measurements. The specification includes material tests and elution tests. Material tests determine ingredients derived from the raw material used. Several types of materials are determined, such as metals (e.g., cadmium and lead) and volatile substances. Elution tests determine the amounts of substances that dissolve in solutions prepared under specific conditions. These tests determine various items such as residues on evaporation, heavy metals, and consumption of potassium permanganate.

The specifications for synthetic resins are divided into specifications that apply generally to all resins and specifications that apply specifically to individual resins. In both cases, material tests and elution tests are specified. The elution tests are performed with *n*-heptane, 20 % ethanol, water, and 4 % acetic acid and are evaluated by the residue on evaporation, expressed in ppm. The conditions for extraction are 60 min at 25 °C for *n*-heptane, and 30 min at 60 °C (90 °C for high-temperature use) for the other solvents.

[Top of Page]

9. Packaging and the Environment

Norbert Buchner

Packaging causes waste, which represents a major problem particularly in *urban* regions that lack disposal areas. The total amount of refuse from an average household in an economic system with high packaging rates has proved to be lower than from the equivalent household in a less developed system (see Section Functions and Effects.). However, the composition of the waste differs, and the former are more dependent on public refuse systems. In the Federal Republic of Germany, for instances, the amount of solid household waste resulting from packaging (approximately 8×10^6 t/a) represents about 30 % of the total household refuse (26×10^6 t/a). However, with respect to the total solid waste of 5×10^8 t/a, this amounts to only ca. 1.6 %. Disposable packages for beverages, whose refuse of 4×10^5 t/a represents ca. 1.5 % of total household waste, are particular objects of attention. Many of the conventional measures designed to reduce packaging costs also lead to *refuse reduction*. In the Federal Republic of Germany the fraction of packaging material in the gross national product dropped from ca. 2 % to 1.5 % within three decades.

Within the field of *waste recycling*, a parallel exists between material and energy recycling. Very high, but still increasing, rates of recycling occur with glass (the Netherlands leads with almost 50 %, compared to the Federal Republic of Germany which recycles around 40 %) [120]. When compared with the expenditure for new glass, this can lead in an ideal case to an energy saving of almost 25 % in manufacturing. In the Federal Republic of Germany, ca. 40 % of tin cans are recycled, amounting to a theoretical energy saving of 60 % in comparison with new cans. In the United States, where aluminum recycling is particularly attractive due to the widespread use of aluminum cans, a return system has been built up that allows a possible energy saving of over 90 %. A high percentage of packaging papers, cardboard, and corrugated cardboard is manufactured from secondary fibers (see Sections Cardboard, Corrugated Cardboard, Coated Paper and Cardboard); however, these originate to a low degree in the packaging sector.

Methods of *refuse disposal* include dumps, composting, incineration, and pyrolysis. With correct disposal or composting of packaging materials, danger to the environment (soil, water, air) can be ruled out. Thermal methods of disposal (refuse power stations, pyrolysis) are increasingly regarded as, in effect, energy-recycling systems, which can also reduce the amount of deposited waste by 90 – 95 %. When paper and cardboard are incinerated in refuse power stations, up to 65 % of the expenditure of energy can be recycled, with plastics more than 50 % of the primary energy (fuel value and production energy) is regained [122]. Refuse power stations are today able to meet more stringent exhaust gas regulations than conventional heating power stations [121].

[Top of Page]

References

1. K. G. Bergner, F. Kiermeier: *Handbuch der Lebensmittelchemie*, vol. **9**, Springer Verlag, Berlin, Heidelberg, New York 1970, p. 128.
2. R. Fürst: “Das Verpackungswesen—Ein Spiegel der Zeit-Geschichte,” *Neue Verpack.* **34** (1981) no. 8, 1152 – 1158.
3. Festschrift der Fa. Wolff, Walsrode (D) 1965. *Vom Ballenbinder zur Selbstbedienung*, Walsrode (D) 1965.
4. P. Pink, *Verpackgs. Rundsch.* **37** (1986) no. 11, 69 –76.
5. G. Grundke, *Lebensm. Technol.* **38** (1987) no. 4, 265 – 276.
6. *Verpackungsmarkt Europa, Packg. u. Transp.* **20** (1987) 18.
7. Corporate Identitiy, *Neue Verpack.* **39** (1986) 86 – 87.
8. R. Heiss, *Verpackung feuchtigkeitsempfindlicher Güter*, Springer Verlag, Berlin, Göttingen, Heidelberg 1956.
9. R. Heiss *Lebensm. Wiss. u. Technol.* **10** (1977) 123 – 130.
10. J. S. Segura, *Packaging (Boston)* **58** (1987) no. 2, 74 – 75.
11. H. A. Highland, *Package Dev. & Sys.* (1977), 36 – 38.
12. W. Guise, *Packaging (Boston)* **58** (1987) no. 4, 14 – 18.
13. F. J. Möller, G. Schricker, *Lebensm. Techn.* **10** (1987) 561 – 566.
14. G. F. C. Barrett *Packaging* (*Boston*) **58** (1987) 21 – 25.
15. D. Berndt: “Arbeitsmappe für den Verpackungspraktiker,” *Neue Verpack.* **35** (1982) no. 3, 331 – 336.
16. D. Berndt, *Neue Verpack.* **33** (1980) no. 10, 1275 – 1278.
17. *Neue Verpack.* **36** (1983) no. 9, 1063.
18. Th. Schotte: “Zellglas in der Lebensmittel-Verpackung,” *Gordian* **67** (1968) no. 5, 6.
19. M. Herner, H. J. Wolter, *Kunststoffe* **77** (1987) no. 10, 959 – 962.
20. C. Gondro, *Kunststoffe* **77** (1987) no. 10, 962 – 966.
21. D. Berndt: “Kunststoffe für Packstoffe, Packmittel und Packhilfs-Mittel, Part 1: Polyethylen und Polypropylen,” *Neue Verpack.* **40** (1987) no. 12, 32 – 41.
22. W. Strobel, *Kunststoffe* **77** (1987) no. 10, 967 – 971.
23. D. Berndt, *Neue Verpack.* **32** (1979) no. 2, 169 – 174.
24. D. Berndt, *Neue Verpack.* **31** (1978) no. 6, 957 – 960; no. 8, 1207 – 1210; no. 9, 1361 – 1362; no. 10, 1541 – 1544; **33** (1980), no. 7, 859 – 860.
25. PVC und Umwelt, *Neue Verpack.*, **40** (1987), no. 12, 52 – 59.
26. R. Frank: *Kunststoffe im Lebensmittelverkehr, Loseblatt-Sammlung*, C. Heymans Verlag, Köln-Berlin-Bonn-München.
27. W. A. Neitzert, *Neue Verpack.* **37** (1984) no. 1, B 65 – B 70.
28. G. Lindenschmitt, R. Theysohn, *Kunststoffe* **77** (1987) no. 10, 982 – 988.
29. H. Jenne, *Kunststoffe* **77** (1987) no. 10, 972 – 975.
30. H. Jenne, *Verpack. Rundsch.* **39** (1986) no. 10, 1163 – 1172.
31. H. M. Caesar, *Kunststoffe* **77** (1987) no. 10, 1004 – 1009.
32. D. Berndt, *Neue Verpack.* **35** (1982) no. 11, 1435 – 1437.
33. M. Bakker (ed.): *The Wiley Encyclopedia of Packaging Technology*, J. Wiley & Sons, New York 1986.
34. M. Larson, *Packaging (Boston)*, **57** (1987), no. 6, 58 – 64.
35. D. Berndt, *Neue Verpack.* **37** (1986) no. 8, B 89 – B 94.
36. D. Flieger, *Verpackgs. Rundsch.* **39** (1986) no. 10, 1182 – 1184.
37. M. Boysen, *Kunststoffe* **77** (1987) no. 5, 522 – 525.
38. U. A. Neitzert: *Neue Verpack.* **40** (1987) no. 2, 50 – 63.
39. H. O. Roppel, *Kunststoffe* **77** (1987) no. 12, 1259 – 1263.
40. *Neue Verpack.* **40** (1987) no. 12, 44 — 47.
41. R. Vieweg, G. Daumiller: *Kunststoff-Handbuch*, vol. 5, Hanser Verlag, München 1969.
42. D. Berndt, *Neue Verpack.* **34** (1981) no. 2, 159 – 163; no. 3, 309 – 312.
43. I. De Grave, *Kunststoffe* **70** (1980) no. 10, 625 – 629.
44. W. Guise, *Packaging (Boston)* **58** (1987) no. 12, 4 – 9.
45. N. Buchner, D. Liede, *Verpack. Rundsch.* **28** (1977) no. 4, 494 –498; no. 5, 604 – 612.
46. D. Berndt, *Neue Verpack.* **36** (1983) no. 8, B 35 – B 44; no. 10, B 45 – B 52; no. 11, B 53 – B 56.
47. E. Klenke, *Neue Verpack.* **38** (1985) no. 6, 34 – 37.
48. *Verpackgs. Rundsch.* **34** (1983) no. 9, 1043 – 1047.

49. D. Berndt, *Neue Verpack.* **40** (1987) no. 2, B 203 – B 206; no. 3, B 207 – 210; no. 4, B 211 – B 214; no. 5, B 215 – B 218; no. 6, B 219 – B 222; no. 9, B 231 – 234; no. 10, B 235 – B 238; no. 11, B 239 – B 242.
50. P. M. Erlandson, *Package Dev*. & Sys. (1977 no. 314, 31 – 38).
51. N. Buchner, *Neue Verpack.* **38** (1985) no. 4, 26 – 36.
52. D. Schreiber, *Neue Verpack.* **40** (1987) no. 7, 22 – 30,
53. D. Berndt, *Neue Verpack.* **32** (1979) no. 8, 935 – 938; **34** (1981) no. 5, 881 – 884; **37** (1984) no. 5, B 81 –B 84; no. 6, B 85 – B 88.
54. *Packung und Transport* (1985) no. 4, 21 – 22.
55. DIN 8740, *Teil 1 05.76. Verpackungsmaschinen,Begriffe und Einteilung.*
56. DIN 55 405, *Teil 5 04.77. Begriffe für das Verpackungswesen.*
57. B. Heise, *Verpack. Rundsch.* **35** (1984) no. 6, 35 – 42.
58. G. Schönbuch, *Verpack. Rundsch.* **22** (1971) no. 6, 49 – 54.
59. U. Ernst, *Verpack. Rundsch.* **36** (1985) no. 1, 1 – 7.
60. H. J. Hohmann, *Neue Verpack.* **40** (1987) no. 8, 36 – 45.
61. G. Schönbach, *Verpack. Rundsch.* **30** (1979) no. 1, 1 – 7.
62. L. Placzek, *Coating* (1984), Nr. 8, 198 – 204.
63. DIN-Entwurf 8588, *12.83. Fertigungsverfahren Zerteilen – Einordnung, Unterteilen, Begriffe.*
64. Th. Hülsmann: *Messerverschleiß in Querschneideeinrichtungen von Verpackungsmaschinen*, Universität Stuttgart, Diss., 1988.
65. H. J. Hohmann, Ch. Woyke, *Verpack. Rundschau.* **31** (1980) no. 10, 67 – 73.
66. W. A. Neitzert in E. Schmidt (ed.): RGV-Handbuch Verpackung, "*Die Thermoformung von Kunststoffverpackungen*" **vol. 2**, Berlin 1986.
67. F. Zeppelzauer, *Verpack. Rundsch.* **13** (1962), no. 7, 53 – 58.
68. D. J. Hine, R. Ward, *Packaging Technology* **17** (1971) no. 116, 16 – 20, 27.
69. H. J. Hohmann, *Verpack. Rundsch.* **28** (1977) no. 12, 95 – 104.
70. Y. Dagel, U. Sandberg, *Verpack. Rundsch.* **20** (1969) 101 – 104.
71. DIN 8740, *Teil 2 04.81. Begriffe für Verpackungsmaschinen, Füllmaschinen.*
72. H. Gericke, A. Wagner, *Chem. Ing. Tech.* **57** (1985) no. 4, 274 – 280.
73. H. Vogelpohl, H. J. Hohmann, *Verpack. Rundsch.* **38** (1987) no. 5, 35 – 40.
74. G. Vetter, H. Gericke, D. Fritsch, *Aufbereit. Technik* **25** (1984) no. 12, 705 – 717.
75. VDI-Richtlinie 2860, Entwurf 10.82. Montage und Handhabungstechnik.
76. N. Buchner, ZFL *Intern. Z. Lebensm. Technol. Verfahrenstech.* **38** (1987) no. 5, 371 – 381.
77. G. Cerny, *Verpack. Rundsch.* **27** (1976) no. 4, 27 – 32.
78. A. Schutz, *Neue Verpack.* **40** (1987) no. 10, 52 – 57.

General References

79. *Food Cosmetics and Drug Packaging*, Elsevier International Bulletins, P.O. Box 211, 1000 AE Amsterdam, The Netherlands.

Specific References

80. *Ullmann*, 4th ed., **1**, 140.
81. *Ullmann*, 4th ed., **15**, 249.
82. *Ullmann*, 4th ed., **11**, 680.
83. R. Heiss: *Verpackung von Lebensmitteln*, Springer-Verlag, Heidelberg 1980, p. 159.
84. K. Eichner in M. Mathlouthi (ed.): *Food Packaging and Preservation* "Theory and Practice," Elsevier Applied Science Publishers, London 1986, p. 67.
85. R. C. Reid, K. R. Sidman, A. D. Schwope, D. E. Till, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 580 – 587.
86. C. G. vom Bruck, W.-D. Bieber, K. Figge in M. Mathlouthi (ed.): *Food Packaging and Preservation*, "Theory and Practice", Elsevier Applied Science Publishers, London 1986, p. 39.
87. N. T. Crosby: *Food Packaging Material* "Aspects of Analysis and Migration of Contaminantes," Applied Sci. Pub., London 1981, p. 112.
88. J. Crank: *The Mathematics of Diffusion*, Clarendon Press, Oxford 1975, p. 58.
89. J. Koszinowski, *J. Appl. Polym. Sci.* **31** (1986) 1805 — 1826. Links
90. J. Crank, in [88], p. 58.
91. F. B. Rudolph, *J. Polym. Sci. Polym. Phys. Ed.* **19** (1980) 2323 – 2336.
92. W.-D. Bieber, K. Figge, J. Koch, *Food Addit. Contam.* **2** (1985) no. 2, 113 – 124. Links
93. C. E. Rogers in J . Comyn (ed.): *Polymer Permeability*, Elsevier Applied Science Pub. LTD., Essex, England 1986, p. 11.
94. A. Keinhorst, H. Niebergall, *Dtsch. Lebensm. Rundsch.* **82** (1986) no 10, 325 – 333.
95. M. Salame, in Proceedings from 1986 Polymer, Laminations and Coating Conference, Book 2, TAPPI Press, Atlanta, Georgia, p. 363.
96. J. Crank, in [88], p. 267.
97. J. Crank, in [88], p. 254.

98. P. G. Demertzis, M. G. Kontominas, S. G. Gilbert, *J. Food Sci.* **52** (1987) no. 3, 747 – 750. Links
99. W. Karcher, G. Haesen, B. Le Goff, K. Figge, *Dtsch. Lebensm. Rundsch.* **80** (1984) no. 5, 141 – 143.
100. G. Schricker in M. Mathlouthi (ed.): *Food Packaging and Preservation*, "Theory and Practice," Elsevier Applied Science Publishers, London 1986, p. 341.
101. O. Till, A. D. Schwope, D. J. Ehnholt, K. R. Sidman et al., CRC Crit. Rev. Toxicol. **18** (1987) no. 3, 161 – 188.
102. W.-D. Bieber, W. Freytag, K. Figge, C. G. vom Bruck, Chem. Toxicol. **22** (1984) no. 9, 737 – 742.
103. J. Crank, G. S. Park in J. Crank and G. S. Park (eds.): *Diffusion in Polymers*, Academic Press Inc., London. 1968, p. 1.
104. R. Hernandez, J. Giacin, A. Baner, *Packaging Technology (Hillsdale N.J.)* **16** (1986) no. 4, 12 – 15.
105. ISO 5725, 3534,
106. L. J. van Gemert, A. H. Nettenbreijer: *Compilation of odour threshold values in air and water*. National Institute for Water Supply, Central Institute for Nutrition and Food Research TNO, Netherlands 1977.
107. R. Granzer, J. Koszinowski, L. Robinson-Mand, O. Piringer, *Verpack. Rundsch.* **37** (1896) no. 8, 53 – 58.
108. DIN 10 955, *Sensorische Prüfungen; Prüfungen von Packmitteln für Lebensmittel, Beuth Verlag GmbH, Berlin* 1983.
109. J. Koszinowski, O. Piringer, *J. Plast. Film Sheeting* **2** (1986) 40 – 50.
110. J. D. McGuineess, *Food Addit. Contam.* **3** (1986) no. 2, 95 – 102. Links
111. C. Mannheim, N. Passy, *CRC Crit. Rev. Food Sci. Nutr.* **17** (1982) no. 4, 371.
112. L. L. Katan, A New Approach To Migration Legislation, presented at the 5th IAPRI Conference, Bristol, England, October 7 – 9, 1986, Proceedings vol. 3.
113. R. Franck, H. Wieczorek: *Kunststoffe im Lebensmittelverkehr*, Carl Heymanns Verlag, Köln, 37th update, July 1987.
114. L. Rossi in L. L. Katan (ed.): *Food Packaging Legislation*, U.S./E.E.C. Perspectives, Elsevier International Bulletins, Oxford, UK. 1982, p. 118.
115. Code of Federal Regulations: *Food and Drugs*, Title 21 parts 170 – 199, U.S. Government Printing Office, Washington, D.C. 1982.
116. P. Schwartz in L. L. Katan (ed.): *Food Packaging Legislation*, U.S./E.E.C. Perspectives, Elsevier International Bulletins, Oxford, UK 1982, p. 102.
117. L. L. Katan (ed.): *Food Packaging Legislation*, U.S./E.E.C. Perspectives, Elsevier International Bulletins, Oxford, UK 1982.
118. J. Gilbert, J. R. Startin, J. D. McGuinness, *Food Addit. Contam.* **3** (1986) no. 2, 133 – 144. Links
119. A. D. Little, Inc.: Final summary report: Stud. of indirect food additive migration, FDA contract 223–77–2360, Cambridge, Mass., July 1983. Chapter 9.
120. *Verpack. Rundsch.* **38** (1987) no. 12, 1486 – 1487.
121. *Verpack. Rundsch.* **38** (1987) no. 12, 1502.
122. *Verpack. Rundsch.* **38** (1987) no. 12, 1482 – 1485.
123. R. Asby, I. Cooper, S. Harvey, P. Tice: *Food Packaging Migration and Legislation*, Pira International, Leatherhead 1997.
124. R. J. Hernandez, in "Food Packaging Materials, Barrier Properties, and Selection", E. Rostein, R. P. Singh, K. J. Valentas (eds.): *Handbook of Food Engineering Practice*, Chapter 9, CRC Press, Boca Raton, FL 1997.
125. J. H. Hotchkiss (ed.): Food and Packaging Interactions. ACS Symposium Series 365, American Chemical Society, Washington, DC 1988.
126. J. H. Hotchkiss, S. J. Risch (eds.): Food and Packaging Interactions II. ACS Symposium Series 473, American Chemical Society, Washington, DC 1991.
127. S. J. Risch (ed.): Food Packaging—Testing Methods and Applications. ACS Symposium Series 753, American Chemical Society, Washington, DC 2000.
128. L. L. Katan: *Migration from Food Contact Materials*, Aspen Publishers, Inc., Frederick, Maryland 1996.
129. O.-G. Piringer, A. L. Baner: *Plastic Packaging Materials for Food*, Wiley-VCH, Weinheim, New York 2000.
130. G. L. Robertson: *Food Packaging: Principles and Practice*, Marcel Dekker, N.Y. 1993.
131. J. M. Vergnaud: *Liquid Transport Processes in Polymeric Materials. Modeling and Industrial Applications*, Prentice Hall, Englewood Cliffs, New Jersey 1991.
132. P. W. Atkins: *Physical Chemistry*, 6th ed., Oxford University Press, Oxford 1998.
133. J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook*, 3rd ed., John Wiley & Sons, 1989.
134. J. Crank: *The Mathematics of Diffusion*, 2nd ed., Clarendon Press, Oxford 1975.
135. A. Reynier, P. Dole, A. Feigenbaum, *Food Addit. Contam.* **16** (1999) no. 4, 137 – 152. Links
136. P. Mercea, in "Models of diffusion in polymers", O.-G. Piringer, A. L. Baner (eds.): *Plastic Packaging Materials for Food*, Chapter 5, Wiley-VCH, Weinheim, N.Y. 2000.
137. O. Piringer, "Evaluation of Plastics for Food Packaging", *Food Addit. Contam.* **11** (1994) 221 – 230. Links
138. J. Brandsch, P. Mercea, O.-G. Piringer: "Modeling of Additive Diffusion Coefficients in Polyolefins", S. J. Risch (ed.): Food Packaging—Testing Methods and Applications. ACS Symposium Series 753, American Chemical Society, Washington, DC 2000.
139. W. Limm, H. C. Hollifield, *Food Addit. Contam.* **13** (1996) no. 8, 949 – 967. Links
140. MIGRATEST Lite 2000, A software to estimate the migration from polymeric packaging into foodstuffs and food simultants, FABES Forschungs-GmbH, Schragenhofstr. 35, D-80992 München, Germany.
141. SMEWISE, Freeware downloadable on http://www.inra.fr/Internet/Produits/securite-emballage. INRA-CPCB, Moulinde de la Housse, B.P. 1039, F-51687 Reims Cedex2, France.

142. Prediction of Specific Migration from Packaging Material into Food. Swiss Federal Office of Public Health, Division of Food Science, P.O. Box, CH-3003 Bern, Switzerland.
143. P. Ackermann, M. Jägerstadt, T. Ohlsson (eds.): *Foods and Packaging Materials—Chemical Interactions*, The Royal Society of Chemistry, Cambridge 1995.
144. H. Zweifel (ed.): *Plastics Additives Handbook*, 5th ed., Carl Hanser Verlag, Munich 2001.
145. P. G. Demertzis, R. Franz, F. Welle, *Packag. Technol. and Sci.* **12** (1999) 119 – 130.
146. C. Mannheim, N. Passy, *CRC Crit. Rev. Food Sci. Nutr.* **17** (1982) no. 4, 371.
147. CEN, European Standards EN/ENV 1186, Materials and articles in contact with foodstuff – Plastics, Part 15. European Committee for Standardization, Brussels 1998.
148. CEN, European Standards ENV 13130, Materials and articles in contact with foodstuff – Plastics, Part 1–8. European Committee for Standardization, Brussels 1999.
149. L. J. van Gemert, A. H. Nettenbreijer, “Compilation of odor threshold values in air and water”, National Institute for Water Supply, Central Institute for Nutrition and Food Research TNO, Netherlands, 1977.
150. DIN 10955, Sensorische Prüfung; Prüfung von Packmitteln für Lebensmittel, Beuth Verlag GmbH, Berlin 1983.
151. L. Rossi, in “European Community Legislation on Materials and Articles Intended to come into contact with Food”, O.-G. Piringer, A. L. Baner (eds.): *Plastic Packaging Materials for Food*, Chapter 12, Wiley-VCH, Weinheim, New York 2000.
152. Code of Federal Regulations: Food and Drugs Title 21, parts 170–199, US Government Printing Office, Washington DC. 2001.
153. T. H. Begley, in “Migration from food packaging: Regulatory Considerations for Estimating Exposure”, O.-G. Piringer, A. L. Baner (eds.): *Plastic Packaging Materials for Food*, Chapter 11, Wiley-VCH, Weinheim, New York 2000.
154. “Administration of Food Chemistry”, 12th Food Sanitation Administration Experts Study Program, 2000, prepared by Food Chemistry Division, Environmental Health Bureau, Minister of Health and Welfare, Japan.

[Top of Page]

Abbreviations used in this article:

ABS: acrylonitrile – butadiene – styrene copolymer

AN: acrylonitrile

EAA: ethylene – acrylic acid copolymer

EMA: ethylene – methacrylic acid copolymer

EPS: expanded polystyrene

EVM: ethylene – vinyl acetate copolymer

EVOH: ethylene – vinyl alcohol copolymer

HDPE: high-density polyethylene

HIPS: high impact polystyrene

HMPE: high molecular mass polyethylene

LDPE: low-density polyethylene

LLDPE: linear low-density polyethylene

PA: polyamide

PAN: poly(acrylonitrile)

PBT: poly(butylene terephthalate)

PE: polyethylene

PET: poly(ethylene terephthalate)

PETG: glycol-modified poly(ethylene terephthalate)

PP: polypropylene

PS: polystyrene

PVAL: poly(vinyl alcohol)

PVC: poly(vinyl chloride)

PVDC: poly(vinyliden chloride)

SAN: styrene – acrylonitrile copolymer

SB: styrene – butadiene copolymer

VC: vinyl chloride

VDC: vinylidene chloride

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Formaldehyde

GÜNTHER REUSS, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1, 2, 4, and 5)

WALTER DISTELDORF, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 3, 6 – 9)

ARMIN OTTO GAMER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 10)

ALBRECHT HILT, Ultraform GmbH, Ludwigshafen, Federal Republic of Germany (Chaps. 11, 12)

1. Introduction 2
2. Physical Properties 2
2.1. Monomeric Formaldehyde 2
2.2. Aqueous Solutions 3
3. Chemical Properties 5
4. Production 6
4.1. Silver Catalyst Processes 7
4.1.1. Complete Conversion of Methanol (BASF Process) 8
4.1.2. Incomplete Conversion and Distillative Recovery of Methanol 9
4.2. Formox Process 11
4.3. Comparison of Process Economics 12
4.4. Distillation of Aqueous Formaldehyde Solutions 13
4.5. Preparation of Liquid Monomeric Formaldehyde 14
5. Environmental Protection 14
6. Quality Specifications and Analysis 17
6.1. Quality Specifications 17
6.2. Analysis 17
7. Storage and Transportation 18
8. Uses 19
9. Economic Aspects 20
10. Toxicology and Occupational Health 21
11. Low Molecular Mass Polymers . . 22
11.1. Linear Polyoxymethylenes 22
11.2. Cyclic Polyoxymethylenes 25
11.2.1. Trioxane 25
11.2.2. Tetraoxane 28
11.2.3. Higher Cyclic Polyoxymethylenes . 28
12. Formaldehyde Cyanohydrin 28
13. References 29

1. Introduction

Formaldehyde occurs in nature and is formed from organic material by photochemical processes in the atmosphere as long as life continues on earth. Formaldehyde is an important metabolic product in plants and animals (including humans), where it occurs in low but measurable concentrations. It has a pungent odor and is an irritant to the eye, nose, and throat even at a low concentration; the threshold concentration for odor detection is 0.05 – 1 ppm. However, formaldehyde does not cause any chronic damage to human health. Formaldehyde is also formed when organic material is incompletely combusted; therefore, formaldehyde is found in combustion gases from, for example, automotive vehicles, heating plants, gas-fired boilers, and even in cigarette smoke. Formaldehyde is an important industrial chemical and is employed in the manufacture of many industrial products and consumer articles. More than 50 branches of industry now use formaldehyde, mainly in the form of aqueous solutions and formaldehyde-containing resins. In 1995, the demand for formaldehyde in the three major markets – Northern America, Western Europe, Japan – was 4.1×10^6 t/a [1].

History. Formaldehyde was first synthesized in 1859, when BUTLEROV hydrolyzed methylene acetate and noted the characteristic odor of the resulting solution. In 1867, HOFMANN conclusively identified formaldehyde, which he prepared by passing methanol vapor and air over a heated platinum spiral. This method, but

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a11_619

with other catalysts, still constitutes the principal method of manufacture. The preparation of pure formaldehyde was described later by KEKULÉ in 1882.

Industrial production of formaldehyde became possible in 1882, when TOLLENS discovered a method of regulating the methanol vapor : air ratio and affecting the yield of the reaction. In 1886 LOEW replaced the platinum spiral catalyst by a more efficient copper gauze. The German firm, Mercklin und Lösekann, started to manufacture and market formaldehyde on a commercial scale in 1889. Another German firm, Hugo Blank, patented the first use of a silver catalyst in 1910.

Industrial development continued from 1900 to 1905, when plant sizes, flow rates, yields, and efficiency were increased. In 1905, Badische Anilin & Soda-Fabrik started to manufacture formaldehyde by a continuous process employing a crystalline silver catalyst. Formaldehyde output was 30 kg/d in the form of an aqueous 30 wt % solution.

The methanol required for the production of formaldehyde was initially obtained from the timber industry by carbonizing wood. The development of the high-pressure synthesis of methanol by Badische Anilin & Soda-Fabrik in 1925 allowed the production of formaldehyde on a true industrial scale.

2. Physical Properties

2.1. Monomeric Formaldehyde

Formaldehyde [*50-00-0*], CH_2O, M_r 30.03, is a colorless gas at ambient temperature that has a pungent, suffocating odor and an irritant action on the eyes and skin.

Formaldehyde liquefies at −19.2 °C, the density of the liquid being 0.8153 g/cm^3 at −20 °C and 0.9172 g/cm^3 at −80 °C. It solidifies at −118 °C to give a white paste. The liquid and gas polymerize readily at low and ordinary temperatures up to 80 °C. Pure formaldehyde gas does not polymerize between 80 and 100 °C and behaves as an ideal gas. For the UV absorption spectra of formaldehyde, see [2]. Structural information about the formaldehyde molecule is provided by its fluorescence [3], IR [4], RAMAN [5], and microwave spectra [6]. Following are some of the thermodynamic properties of gaseous formaldehyde:

Heat of formation at 25 °C	−115.9 ± 6.3 kJ/mol
Gibbs energy at 25 °C	−109.9 kJ/mol
Entropy at 25 °C	218.8 ± 0.4 kJmol^{-1} K^{-1}
Heat of combustion at 25 °C	−561.5 kJ/mol
Heat of vaporization at −19.2 °C	23.32 kJ/mol
Specific heat capacity at 25 °C, c_p	35.425 Jmol^{-1} K^{-1}
Heat of solution at 23 °C	
in water	−62 kJ/mol
in methanol	−62.8 kJ/mol
in 1-propanol	−59.5 kJ/mol
in 1-butanol	−62.4 kJ/mol
Cubic expansion coefficient	2.83×10^{-3} K^{-1}
Specific magnetic susceptibility	-0.62×10^6
Vapor density relative to air	1.04

The vapor pressure p of liquid formaldehyde has been measured from −109.4 to −22.3 °C [7] and can be calculated for a given temperature T (K) from the following equation:

$$p\,(\mathrm{kPa}) = 10^{[5.0233-(1429/T)+1.75\log T-0.0063T]}$$

Polymerization in either the gaseous or the liquid state is influenced by wall effects, pressure, traces of humidity, and small quantities of formic acid. Formaldehyde gas obtained by vaporization of paraformaldehyde or more highly polymerized α-polyoxymethylenes, which is ca. 90 – 100 % pure, must be stored at 100 – 150 °C to prevent polymerization. Chemical decomposition is insignificant below 400 °C.

Formaldehyde gas is flammable, its ignition temperature is 430 °C [8]; mixtures with air are explosive. At ca. 20 °C the lower and upper explosive limits of formaldehyde are ca. 7 and 72 vol % (87 and 910 g/m^3), respectively [9]. Flammability is particularly high at a formaldehyde concentration of 65 – 70 vol %.

At a low temperature, liquid formaldehyde is miscible in all proportions with nonpolar solvents such as toluene, ether, chloroform, or ethyl acetate. However, solubility decreases with increasing temperature and at room temperature polymerization and volatilization occur, leaving only a small amount of dissolved gas. Solutions of liquid formaldehyde in acetaldehyde behave as ideal solutions [10]. Liquid formaldehyde is slightly miscible with petroleum ether and *p*-cymene [11].

Polar solvents, such as alcohols, amines or acids, either catalyze the polymerization of formaldehyde or react with it to form methylol compounds or methylene derivatives.

2.2. Aqueous Solutions

At room temperature, pure aqueous solutions contain formaldehyde in the form of methylene glycol $HOCH_2OH$ [*463-57-0*] and its oligomers, namely the low molecular mass poly(oxymethylene) glycols with the following structure

$$HO(CH_2O)_nH \ (n = 1-8)$$

Monomeric, physically dissolved formaldehyde is only present in low concentrations of up to 0.1 wt %. The polymerization equilibrium

$$HOCH_2OH + n\,CH_2O \rightleftharpoons HO(CH_2O)_{n+1}-H$$

is catalyzed by acids and is shifted toward the right at lower temperature and/or higher formaldehyde concentrations, and toward the left if the system is heated and/or diluted [12], [13] (see also Section 11.1).

Dissolution of formaldehyde in water is exothermic, the heat of solution (− 62 kJ/mol) being virtually independent of the solution concentration [14]. Clear, colorless solutions of formaldehyde in water can exist at a formaldehyde concentration of up to 95 wt %, but the temperature must be raised to 120 °C to obtain the highest concentrations. Concentrated aqueous solutions containing more than 30 wt % formaldehyde become cloudy on storage at room temperature, because larger poly(oxymethylene) glycols ($n \geq 8$) are formed which then precipitate out (the higher the molecular mass of the polymers, the lower is their solubility).

Table 1. Calculated distribution of oligomers of methylene glycol, $HO(CH_2O)_nH$, in an aqueous 40 wt % formaldehyde solution at 35 °C [12]

n	Proportion, %	n	Proportion, %
1	26.80	7	3.89
2	19.36	8	2.50
3	16.38	9	1.59
4	12.33	10	0.99
5	8.70	> 10	1.58
6	5.89		

Equilibrium constants have been determined for the physical dissolution of formaldehyde in water and for the reaction of formaldehyde to give methylene glycol and its oligomers [12]. These parameters can be combined with other data to calculate the approximate equilibria at any temperature from 0 to 150 °C and at a formaldehyde concentration of up to 60 wt % [13]. Table 1 gives the calculated oligomer distribution in an aqueous 40 wt % solution of formaldehyde.

A kinetic study of the formation of methylene glycol from dissolved formaldehyde and water shows that the reverse reaction is 5×10^3 to 6×10^3 times slower than the forward reaction [15], and that it increases greatly with the acidity of the solution. This means that the distribution of the higher mass oligomers ($n > 3$) does not change rapidly when the temperature is increased or the solution is diluted; the methylene glycol content then rises at the expense of the smaller oligomers ($n = 2$ or 3). In aqueous solutions containing ≤ 2 wt % formaldehyde, formaldehyde is entirely monomeric.

Methylene glycol can be determined by the bisulfite method [16] or by measuring the partial pressure of formaldehyde [17]. Molecular masses and monomer contents can be determined by NMR spectroscopy [13], [18].

The approximate amount of monomeric formaldehyde present as formaldehyde hemiformal and methylene glycol in aqueous solutions containing formaldehyde and methanol, can be calculated from data at 25 – 80 °C [19] by using the following equation:

$$\text{Monomer (mol\%)} = 100 - 12.3\sqrt{F} + (1.44 - 0.0164F)\,M$$

where F is the formaldehyde concentration (7 – 55 wt %) and M is the methanol concentration (0 – 14 wt %).

The partial pressure p_F of formaldehyde above aqueous solutions has been measured by Ledbury and Blair and computed by Walker and Lacy [20]. The parameter p_F for solutions in which F is in the range 0 – 40 wt % can be calculated with a relative error of 5 – 10 % in the temperature range $T = 273 - 353$ K by using the following equation :

$$p_F\,(\text{kPa}) = 0.1333F e^{-F^{\alpha}(a_0 + a_1/T + a_2/T^2)}$$

$$\alpha = 0.08760 \pm 0.00950$$

$$a_0 = -12.0127 \pm 0.0550$$

$$a_1 = 3451.72 \pm 17.14$$

$$a_2 = 248257.3 \pm 5296.8$$

Results of such calculations are given in Table 2 and agree well with the measured values.

Table 3 gives the partial pressures and concentrations of formaldehyde in the liquid and

Table 2. Partial pressure p_F of formaldehyde (kPa) above aqueous formaldehyde solutions

t, °C	Formaldehyde concentration, wt %								
	1	5	10	15	20	25	30	35	40
5	0.003	0.011	0.016	0.021	0.025	0.028	0.031	0.034	0.037
10	0.005	0.015	0.024	0.031	0.038	0.043	0.049	0.053	0.056
15	0.007	0.022	0.036	0.047	0.057	0.066	0.075	0.083	0.090
20	0.009	0.031	0.052	0.069	0.085	0.099	0.113	0.125	0.137
25	0.013	0.044	0.075	0.101	0.125	0.146	0.167	0.187	0.206
30	0.017	0.061	0.105	0.144	0.180	0.213	0.245	0.275	0.304
35	0.022	0.084	0.147	0.203	0.256	0.305	0.353	0.398	0.442
40	0.028	0.113	0.202	0.284	0.360	0.432	0.502	0.569	0.634
45	0.037	0.151	0.275	0.390	0.499	0.604	0.705	0.803	0.899
50	0.047	0.200	0.371	0.531	0.685	0.833	0.978	1.119	1.258
55	0.059	0.262	0.494	0.715	0.929	1.137	1.341	1.541	1.740
60	0.074	0.340	0.652	0.953	1.247	1.536	1.820	2.101	2.378
65	0.093	0.437	0.852	1.258	1.657	2.053	2.443	2.831	3.218
70	0.114	0.558	1.104	1.645	2.182	2.717	3.250	3.780	4.310

gaseous phases of aqueous formaldehyde solutions. The partial pressures and concentrations were measured at the boiling points of the solutions at a pressure of 101.3 kPa [21].

Table 3. Concentration and partial pressure of formaldehyde measured at the boiling points (101.3 kPa) of aqueous formaldehyde solutions [21]

Formaldehyde concentration, wt %		Partial pressure
Liquid phase (F_l)	Gaseous phase (F_g)	(p_F), kPa
3.95	3.68	2.35
8.0	7.3	4.75
12.1	10.6	7.0
15.3	13.2	8.65
20.1	16.95	11.2
25.85	21.45	14.45
30.75	24.9	16.8
35.65	27.4	18.8
42.0	30.5	21.4
47.5	33.1	23.4
49.8	34.0	24.1

Aqueous Formaldehyde – Methanol Solutions. Technical-grade formaldehyde solutions contain a small amount of methanol as a result of the incomplete methanol conversion during formaldehyde production. The amount of methanol present depends on the production process employed. The presence of methanol is often desirable in aqueous solutions containing more than 30 wt % formaldehyde because it inhibits the formation of insoluble, higher mass polymers. Methanol concentrations of up to 16 wt % stabilize the formaldehyde.

The approximate *density* ϱ (in grams per cubic centimeter) of aqueous formaldehyde solutions containing up to 13 wt % methanol at a temperature of 10 – 70 °C can be calculated by using the following equation [22]:

$$\varrho = a + 0.0030\,(F - b) - 0.0025\,(M - c) - 10^4\,[0.055\,(F - 30) + 5.4]\,(t - 20)$$

where

F = formaldehyde concentration in wt %
M = methanol concentration in wt %
t = temperature in °C
a, b, and c = constants

The following values can be assumed when F is in the range 0 – 48: $a = 1.092$, $b = 30$, and $c = 0$. The corresponding values in the range $F = 48 - 55$ are $a = 1.151$, $b = 50.15$, and $c = 1.61$.

The *boiling points* of pure aqueous solutions containing up to 55 wt % formaldehyde are between 99 and 100 °C at atmospheric pressure [23]. In dilute aqueous solutions, formaldehyde lowers the freezing point of water. If solutions containing more than 25 wt % formaldehyde are cooled, polymer precipitates out before the freezing point is reached. According to Natta [22], the approximate *refractive index* n_D^{18} of aqueous 30 – 50 wt % formaldehyde solutions containing up to 15 wt % methanol can be calculated from the following equation:

$$n_D^{18} = 1.3295 + 0.00125F + 0.000113M$$

where F and M are wt % concentrations of formaldehyde and methanol, respectively.

In close agreement with measurements of commercial solutions, the *dynamic viscosity* η

of aqueous formaldehyde – methanol solutions may be expressed by the following equation [24]:

$$\eta\,(\mathrm{mPa{\cdot}s}) = 1.28 + 0.039F + 0.05M - 0.024t$$

This equation applies to solutions containing 30 – 50 wt % formaldehyde and 0 – 12 wt % methanol at a temperature t of 25 – 40 °C.

Detailed studies on chemical reactions, vapor–liquid equilibria and caloric properties of systems containing formaldehyde, water, and methanol are available [216–226].

3. Chemical Properties

Formaldehyde is one of the most reactive organic compounds known and, thus, differs greatly from its higher homologues and aliphatic ketones [25], [26]. Only the most important of its wide variety of chemical reactions are treated in this article; others are described in [27]. For a general discussion of the chemical properties of saturated aldehydes, see → Aldehydes, Aliphatic and Araliphatic, Chap. 2.

Decomposition. At 150 °C, formaldehyde undergoes heterogeneous decomposition to form mainly methanol and CO_2 [28]. Above 350 °C, however, it tends to decompose into CO and H_2 [29]. Metals such as platinum [30], copper [31], chromium, and aluminum [32] catalyze the formation of methanol, methyl formate, formic acid, CO_2, and methane.

Polymerization. Anhydrous monomeric formaldehyde cannot be handled commercially. Gaseous formaldehyde polymerizes slowly at temperatures below 100 °C, polymerization being accelerated by traces of polar impurities such as acids, alkalis, or water (see paraformaldehyde, Section 11.1). Thus, in the presence of steam and traces of other polar compounds, the gas is stable at ca. 20 °C only at a pressure of 0.25 – 0.4 kPa, or at a concentration of up to ca. 0.4 vol % at ca. 20 °C and atmospheric pressure.

Monomeric formaldehyde forms a hydrate with water; this hydrate reacts with further formaldehyde to form polyoxymethylenes (see Section 2.2). Methanol or other stabilizers, such as guanamines [33] or alkylenebis(melamines) [34], are generally added to commercial aqueous formaldehyde solutions (37 – 55 wt %) to inhibit polymerization.

Reduction and Oxidation. Formaldehyde is readily reduced to methanol with hydrogen over a nickel catalyst [27], [35]. For example, formaldehyde is oxidized by nitric acid, potassium permanganate, potassium dichromate, or oxygen to give formic acid or CO_2 and water [27], [36].

In the presence of strong alkalis [37] or when heated in the presence of acids [38], formaldehyde undergoes a Cannizzaro reaction with formation of methanol and formic acid [39]. In the presence of aluminum or magnesium methylate, paraformaldehyde reacts to form methyl formate (Tishchenko reaction) [27].

Addition Reactions. The formation of sparingly water-soluble sodium formaldehyde bisulfite is an important addition reaction of formaldehyde [40]. Hydrocyanic acid reacts with formaldehyde to give glycolonitrile [*107-16-4*] [27]. Formaldehyde undergoes an acid-catalyzed Prins reaction in which it forms α-hydroxymethylated adducts with olefins [24]. Acetylene undergoes a Reppe addition reaction with formaldehyde [41] to form 2-butyne-1,4-diol [*110-65-6*]. Strong alkalis or calcium hydroxide convert formaldehyde to a mixture of sugars, in particular hexoses, by a multiple aldol condensation which probably involves a glycolaldehyde intermediate [42], [43]. Mixed aldols are formed with other aldehydes; the product depends on the reaction conditions. Acetaldehyde, for example, reacts with formaldehyde to give pentaerythritol, $C(CH_2OH)_4$ [*115-77-5*] (→ Alcohols, Polyhydric, Chap. 4.).

Condensation Reactions. Important condensation reactions are the reaction of formaldehyde with amino groups to give Schiff's bases, as well as the Mannich reaction [27]. Amines react with formaldehyde and hydrogen to give methylamines. Formaldehyde reacts with ammonia to give hexamethylenetetramine, and with ammonium chloride to give monomethylamine, dimethylamine, or trimethylamine and formic acid, depending on the reaction conditions [44]. Reaction of formaldehyde with diketones and ammonia yields imidazoles [45].

Formaldehyde reacts with many compounds to produce methylol ($-CH_2OH$) derivatives. It reacts with phenol to give methylolphenol, with urea to give mono-, di-, and trimethylolurea, with melamine to give methylolmelamines, and with organometallic compounds to give metal-substituted methylol compounds [27].

Aromatic compounds such as benzene, aniline, and toluidine combine with formaldehyde to produce the corresponding diphenylmethanes. In the presence of hydrochloric acid and formaldehyde, benzene is chloromethylated to form benzyl chloride [*100-44-7*] [46]. The possible formation of bis(chloromethyl)ether [*542-88-1*] from formaldehyde and hydrochloric acid and the toxicity of this compound are reported elsewhere (→ Ethers, Aliphatic).

Formaldehyde reacts with hydroxylamine, hydrazines, or semicarbazide to produce formaldehyde oxime (which is spontaneously converted to triformoxime), the corresponding hydrazones, and semicarbazone, respectively. Double bonds are also produced when formaldehyde is reacted with malonates or with primary aldehydes or ketones possessing a CH_2 group adjacent to the carbonyl group.

Resin Formation. Formaldehyde condenses with urea, melamine, urethanes, cyanamide, aromatic sulfonamides and amines, and phenols to give a wide range of resins (→ Amino Resins; → Phenolic Resins; → Resins, Synthetic).

4. Production

Formaldehyde is produced industrially from methanol [*67-56-1*] by the following three processes:

1) Partial oxidation and dehydrogenation with air in the presence of silver crystals, steam, and excess methanol at 680 – 720 °C (BASF process, methanol conversion = 97 – 98 %).
2) Partial oxidation and dehydrogenation with air in the presence of crystalline silver or silver gauze, steam, and excess methanol at 600 – 650 °C [47] (primary conversion of methanol = 77 – 87 %). The conversion is completed by distilling the product and recycling the unreacted methanol.
3) Oxidation only with excess air in the presence of a modified iron – molybdenum – vanadium oxide catalyst at 250 – 400 °C (methanol conversion = 98 – 99 %).

Processes for converting propane, butane [48], ethylene, propylene, butylene [49], or ethers (e.g., dimethyl ether) [50] into formaldehyde are not of major industrial significance for economic reasons. Processes that employ partial hydrogenation of CO [51] or oxidation of methane [52] do not compete with methanol conversion processes because of the lower yields of the former processes.

The specifications of the methanol, used for formaldehyde production according to processes 1 – 3 are listed in Table 4. However, crude aqueous methanol obtained by high- [54], medium-, or low-pressure [55] synthesis can also be used for process 1. This methanol contains low concentrations of inorganic impurities and limited amounts of other organic compounds. The methanol must be first subjected to purification processes and preliminary distillation to remove low-boiling components.

Table 4. Specifications of commercial methanol (grade AA) used for the production of formaldehyde [53]

Parameter	Specification
Methanol content	> 99.85 wt %
Relative density, d_4^{20}	0.7928 g/cm^3
Maximum boiling point range	1 °C
Acetone and acetaldehyde content	< 0.003 wt %
Ethanol content	< 0.001 wt %
Volatile iron content	< 2 µg/L
Sulfur content	< 0.0001 wt %
Chlorine content	< 0.0001 wt %
Water content	< 0.15 wt %
pH	7.0
$KMnO_4$ test, minimum decolorization time	30 min

4.1. Silver Catalyst Processes

The silver catalyst processes for converting methanol to formaldehyde are generally carried out at atmospheric pressure and at 600 – 720 °C. The reaction temperature depends on the excess of methanol in the methanol – air mixture. The composition of the mixture must lie outside the explosive limits. The amount of air that is used is also determined by the catalytic quality of the silver surface. The following main reactions occur during the conversion of methanol to formaldehyde:

$$CH_3OH \rightleftharpoons CH_2O + H_2 \quad \Delta H = +84\ \text{kJ/mol} \quad (1)$$

$$H_2 + 1/2\ O_2 \rightarrow H_2O \quad \Delta H = -243\ \text{kJ/mol} \quad (2)$$

$$CH_3OH + 1/2\ O_2 \rightarrow CH_2O + H_2O \quad \Delta H = -159\ \text{kJ/mol} \quad (3)$$

The extent to which each of these three reactions occurs, depends on the process data.

Byproducts are also formed in the following secondary reactions:

$$CH_2O \rightarrow CO + H_2 \quad \Delta H = +12.5\ \text{kJ/mol} \quad (4)$$

$$CH_3OH + 3/2\ O_2 \rightarrow CO_2 + 2\ H_2O \quad \Delta H = -674\ \text{kJ/mol} \quad (5)$$

$$CH_2O + O_2 \rightarrow CO_2 + H_2O \quad \Delta H = -519\ \text{kJ/mol} \quad (6)$$

Other important byproducts are methyl formate, methane, and formic acid.

The endothermic dehydrogenation reaction (1) is highly temperature-dependent, conversion increasing from 50 % at 400 °C to 90 % at 500 °C and to 99 % at 700 °C. The temperature dependence of the equilibrium constant for this reaction K_p is given by

$$\log K_p = (4600/T) - 6.470$$

For detailed thermodynamic data of reactions (1) – (6) see [56]. Kinetic studies with silver on a carrier show that reaction (1) is a first-order reaction [57]. Therefore, the rate of formaldehyde formation is a function of the available oxygen concentration and the oxygen residence time on the catalyst surface:

$$\frac{dc_F}{dt} = kc_o$$

where

c_F = formaldehyde concentration
c_O = oxygen concentration
k = rate constant
t = time

A complete reaction mechanism for the conversion of methanol to formaldehyde over a silver catalyst has not yet been proposed. However, some authors postulate that a change in mechanism occurs at ca. 650 °C [58]. New insight into the reaction mechanism is available from spectroscopic investigations [227–229], which demonstrate the influence of different atomic oxygen species on reaction pathway and selectivity. The synthesis of formaldehyde over a silver catalyst is carried out under strictly adiabatic conditions. Temperature measurements both above and in the silver layer show that sites still containing methanol are separated from sites already containing predominantly formaldehyde by only a few millimeters.

The oxygen in the process air is shared between the exothermic reactions, primarily reaction (2) and, to a lesser extent depending on the process used, the secondary reactions (5) and (6). Thus, the amount of process air controls the desired reaction temperature and the extent to which the endothermic reactions (1) and (4) occur.

Another important factor affecting the yield of formaldehyde and the conversion of methanol, besides the catalyst temperature, is the addition of inert materials to the reactants. Water is added to spent methanol – water-evaporated feed mixtures, and nitrogen is added to air and air – off-gas mixtures, which are recycled to dilute the methanol – oxygen reaction mixture. The throughput per unit of catalyst area provides another way of improving the yield and affecting side reactions. These two methods of process control are discussed in [59].

The theoretical yield of formaldehyde obtained from Reactions (1) – (6) can be calculated from actual composition of the plant off-gas by using the following equation:

$$\text{Yield (mol\%)} = 100\left[+r + \frac{(\%CO_2)+(\%CO)}{0.528(\%N_2)+(\%H_2)-3(\%CO_2)-2(\%CO)}\right]^{-1}$$

Percentages signify concentrations in vol % and r is the ratio of moles of unreacted methanol to moles of formaldehyde produced [60]. The equation takes into account the hydrogen and oxygen balance and the formation of byproducts.

4.1.1. Complete Conversion of Methanol (BASF Process)

The BASF process for the complete conversion of methanol to formaldehyde is shown schematically in Figure 1 [61]. A mixture of methanol

and water is fed into the evaporating column. Fresh process air and, if necessary, recycled off-gas from the last stage of the absorption column enter the column separately [60]. A gaseous mixture of methanol in air is thus formed in which the inert gas content (nitrogen, water, and CO_2) exceeds the upper explosive limit. A ratio of 60 parts of methanol to 40 parts of water with or without inert gases is desired. The packed evaporator constitutes part of the stripping cycle. The heat required to evaporate the methanol and water is provided by a heat exchanger, which is linked to the first absorption stage of the absorption column [62]. After passing through a demister, the gaseous mixture is superheated with steam and fed to the reactor, where it flows through a 25 – 30 mm thick bed of silver crystals. The crystals have a defined range of particle sizes [63] and rest on a perforated tray, which is covered with a fine corrugated gauze, thus permitting optimum reaction at the surface. The bed is positioned immediately above a water boiler (cooler), which produces superheated steam and simultaneously cools the hot reaction gases to a temperature of 150 °C corresponding to that of the pressurized steam (0.5 MPa). The almost dry gas from the gas cooler passes to the first stage of a four-stage packed absorption column, where the gas is cooled and condensed. Formaldehyde is eluted countercurrent to water or to the circulating formaldehyde solutions whose concentrations increase from stage to stage.

The product circulating in the first stage may contain 50 wt % formaldehyde if the temperature of the gas leaving this stage is kept at ca. 75 °C; this temperature provides sufficient evaporation energy for the feed stream in the heat exchanger. The final product contains 40 – 55 wt % formaldehyde, as desired, with an average of 1.3 wt % methanol and 0.01 wt % formic acid. The yield of the formaldehyde process is 89.5 – 90.5 mol %. Some of the off-gas is removed at the end of the fourth stage of the column [60] and is recycled due to its extremely low formaldehyde content (Fig. 1, route indicated by dashed-dotted lines). The residual off-gas is fed to a steam generator, where it is combusted [64] (net calorific value = 1970 kJ/m^3). Prior to combustion the gas contains ca. 4.8 vol % CO_2, 0.3 vol % CO, and 18.0 vol % H_2 as well as nitrogen, water, methanol, and formaldehyde. The combusted off-gas contains no environmentally harmful substances. The total steam equivalent of the process is 3 t per ton of 100 wt % formaldehyde.

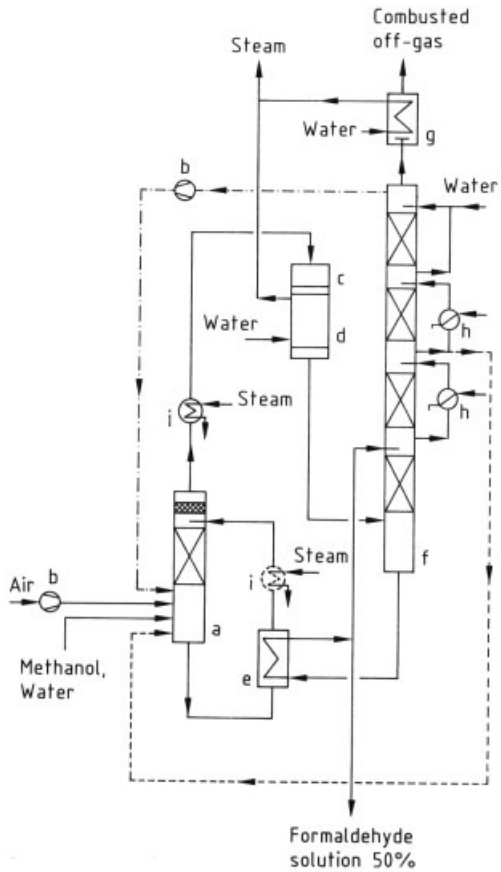

Figure 1. Flowchart of formaldehyde production by the BASF process
a) Evaporator; b) Blower; c) Reactor; d) Boiler; e) Heat exchanger; f) Absorption column; g) Steam generator; h) Cooler; i) Superheater
Recycling schemes : –·–·– off-gas, – – – – formaldehyde solution.

In an alternative procedure to the off-gas recycling process (Fig. 1, dashed lines) the formaldehyde solution from the third or fourth stage of the absorption tower is recycled to the evaporator; a certain amount of steam is used in the evaporation cycle. The resulting vapor is combined with the feed stream to the reactor to obtain an optimal methanol : water ratio [65]. In this case, the temperature of the second stage of the absorption column is ca. 65 °C.

The yields of the two processes are similar and depend on the formaldehyde content of the recycled streams.

The average life time of a catalyst bed depends on impurities such as inorganic materials in the air and methanol feed; poisoning effects caused by some impurities are partially reversible within a few days. The life time of the catalyst is also adversely affected by long exposure to excessively high reaction temperatures and high throughput rates because the silver crystals then become matted and cause an increase in pressure across the catalyst bed. This effect is irreversible and the catalyst bed must be changed after three to four months. The catalyst is regenerated electrolytically.

Since formaldehyde solutions corrode carbon steel, all parts of the manufacturing equipment that are exposed to formaldehyde solutions must be made of a corrosion-resistant alloy, e.g., certain types of stainless steel. Furthermore, tubes that convey water or gases must be made of alloys to protect the silver catalyst against metal poisoning.

If the throughput and reaction temperature have been optimized, the capacity of a formaldehyde plant increases in proportion to the diameter of the reactor. The largest known reactor appears to be that of BASF in the Federal Republic of Germany; it has an overall diameter of 3.2 m and a production capacity of 72 000 t/a (calculated as 100 wt % formaldehyde).

4.1.2. Incomplete Conversion and Distillative Recovery of Methanol

Formaldehyde can be produced by partial oxidation and distillative recovery of methanol. This process is used in numerous companies (e.g., ICI, Borden, and Degussa) [66]. As shown in Figure 2, a feed mixture of pure methanol vapor and freshly blown-in air is generated in an evaporator. The resulting vapor is combined with steam, subjected to indirect superheating, and then fed into the reactor. The reaction mixture contains excess methanol and steam and is very similar to that used in the BASF process (cf. Section 4.1.1). The vapor passes through a shallow catalyst bed of silver crystals or through layers of silver gauze. Conversion is incomplete and the reaction takes place at 590 – 650 °C, undesirable secondary reactions being suppressed by this comparatively low temperature. Immediately after leaving the catalyst bed, the reaction gases are cooled indirectly with water, thereby generating steam. The remaining heat of reaction is then removed from the gas in a cooler and is fed to the bottom of a formaldehyde absorption column. In the water-cooled section of the column, the bulk of the methanol, water, and formaldehyde separate out. At the top of the column, all the condensable portions of the remaining formaldehyde and methanol are washed out of the tail gas by countercurrent contact with process water. A 42 wt % formaldehyde solution from the bottom of the absorption column is fed to a distillation column equipped with a steam-based heat exchanger and a reflux condenser. Methanol is recovered at the top of the column and is recycled to the bottom of the evaporator. A product containing up to 55 wt % formaldehyde and less than 1 wt % methanol is taken from the bottom of the distillation column and cooled. The formaldehyde solution is then usually fed into an anion-exchange unit to reduce its formic acid content to the specified level of less than 50 mg/kg.

If 50 – 55 wt % formaldehyde and no more than 1.5 wt % methanol are required in the product, steam addition is restricted and the process employs a larger excess of methanol. The ratio of distilled recycled methanol to fresh methanol then lies in the range 0.25 – 0.5. If a dilute product containing 40 – 44 wt % formaldehyde is desired, the energy-intensive distillation of methanol can be reduced, leading to savings in steam and power as well as reductions in capital cost. The off-gas from the absorption column has a similar composition to that described for the BASF process (in Section 4.1.1). The off-gas is either released into the atmosphere or is combusted to generate steam, thus avoiding environmental problems caused by residual formaldehyde. Alternatively, the tail gas from the top of the absorber can be recycled to the reactor. This inert gas, with additional steam, can reduce the excess methanol needed in the reactor feed, consequently providing a more concentrated product with less expenditure on distillation. The yield of the process is 91 – 92 mol %.

Process variations to increase the incomplete conversion of methanol employ two-stage oxidation systems [67]. The methanol is first partly

converted to formaldehyde, using a silver catalyst at a comparatively low temperature (e.g., 600 °C). The reaction gases are subsequently cooled and excess air is added to convert the remaining methanol in a second stage employing either a metal oxide (cf. Section 4.2) or a further silver bed as a catalyst.

Formaldehyde solutions in methanol with a relatively low water content can be produced directly by methanol oxidation and absorption in methanol [68]. Anhydrous alcoholic formaldehyde solutions or alcoholic formaldehyde solutions with a low water content can be obtained by mixing a highly concentrated formaldehyde solution with the alcohol (ROH) and distilling off an alcohol – water mixture with a low formaldehyde content. The formaldehyde occurs in the desired solutions in the form of the hemiacetals $RO(CH_2O)_nH$.

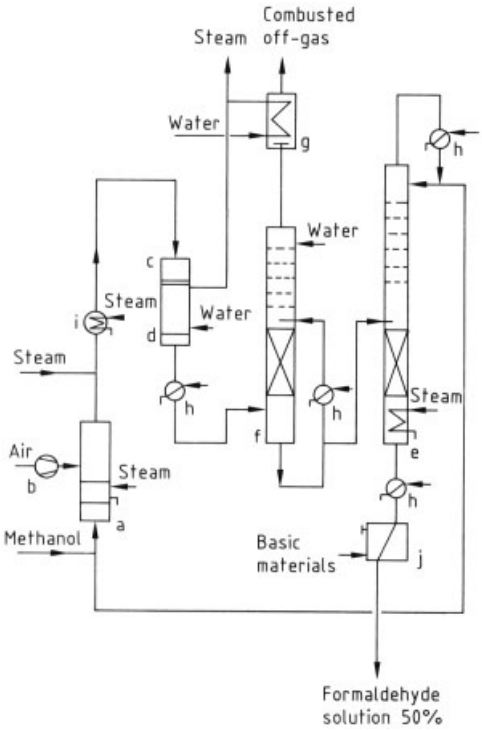

Figure 2. Flowchart of formaldehyde production with recovery of methanol by distillation
a) Evaporator; b) Blower; c) Reactor; d) Boiler; e) Distillation column; f) Absorption column; g) Steam generator; h) Cooler; i) Superheater; j) Anion-exchange unit

4.2. Formox Process

In the Formox process, a metal oxide (e.g., iron, molybdenum, or vanadium oxide) is used as a catalyst for the conversion of methanol to formaldehyde. Many such processes have been patented since 1921 [69]. Usually, the oxide mixture has an Mo : Fe atomic ratio of 1.5 – 2.0, small amounts of V_2O_5, CuO, Cr_2O_3, CoO, and P_2O_5 are also present [70]. Special conditions are prescribed for both the process and the activation of the catalyst [71]. The Formox process has been described as a two-step oxidation reaction in the gaseous state (g) which involves an oxidized (K_{OX}) and a reduced (K_{red}) catalyst [72]:

$$CH_3OH_{(g)} + K_{OX} \rightarrow CH_2O_{(g)} + H_2O_{(g)} + K_{red}$$

$$K_{red} + 1/2\ O_{2(g)} \rightarrow K_{OX}\ \Delta H = -159\ \text{kJ/mol}$$

$$CH_2O + 1/2\ O_2 \rightleftharpoons CO + H_2O\ \Delta H = -215\ \text{kJ/mol}$$

In the temperature range 270 – 400 °C, conversion at atmospheric pressure is virtually complete. However, conversion is temperature-dependent because at >470 °C the following side reaction increases considerably:

$$CH_2O + 1/2\ O_2 \rightleftharpoons CO + H_2O\ \Delta H = -215\ \text{kJ/mol}$$

The methanol oxidation is inhibited by water vapor. A kinetic study describes the rate of reaction to formaldehyde by a power law kinetic rate expression of the form [230]

$$r = kP^x_{CH_3OH} P^y_{O_2} P^z_{H_2O}$$

where $x = 0.94 \pm 0.06$; $y = 0.10 \pm 0{,}05$ and $z = -0.45 \pm 0.07$. The rate is independent of the formaldehyde partial pressure. The measured activation energy is 98 ± 6 kJ/mol.

As shown in Figure 3, the methanol feed is passed to a steam-heated evaporator. Freshly blown-in air and recycled off-gas from the absorption tower are mixed and, if necessary, preheated by means of the product stream in a heat exchanger before being fed into the evaporator. The gaseous feed passes through catalyst-filled tubes in a heat-exchanging reactor. A typical reactor for this process has a shell with a diameter of ca. 2.5 m that contains tubes only 1.0 – 1.5 m

in length. A high-boiling heat-transfer oil circulates outside the tubes and removes the heat of reaction from the catalyst in the tubes. The process employs excess air and the temperature is controlled isothermally to a value of ca. 340 °C; steam is simultaneously generated in a boiler. The air – methanol feed must be a flammable mixture, but if the oxygen content is reduced to ca. 10 mol % by partially replacing air with tail gas from the absorption tower, the methanol content in the feed can be increased without forming an explosive mixture [73]. After leaving the reactor, the gases are cooled to 110 °C in a heat-exchange unit and are passed to the bottom of an absorber column. The formaldehyde concentration is regulated by controlling the amount of process water added at the top of the column. The product is removed from the water-cooled circulation system at the bottom of the absorption column and is fed through an anion-exchange unit to reduce the formic acid content. The final product contains up to 55 wt % formaldehyde and 0.5 – 1.5 wt % methanol. The resultant methanol conversion ranges from 95 – 99 mol % and depends on the selectivity, activity, and spot temperature of the catalyst, the latter being influenced by the heat transfer rate and the throughput rate. The overall plant yield is 88 – 91 mol %.

Well-known processes using the Formox method have been developed by Perstorp/Reichhold (Sweden, United States, Great Britain) [74], [75], Lummus (United States) [76], Montecatini (Italy) [77], and Hiag/Lurgi (Austria) [78].

The tail gas does not burn by itself because it consists essentially of N_2, O_2, and CO_2 with a few percent of combustible components such as dimethyl ether, carbon monoxide, methanol, and formaldehyde. Combustion of Formox tail gas for the purpose of generating steam is not economically justifiable [79]. Two alternative methods of reducing atmospheric emission have been developed. The off-gas can be burned either with additional fuel at a temperature of 700 – 900 °C or in a catalytic incinerator at 450 – 550 °C. However, the latter system employs a heat exchanger and is only thermally self-sufficient if supplementary fuel for start-up is provided and if an abnormal ratio of oxygen : combustible components is used [80].

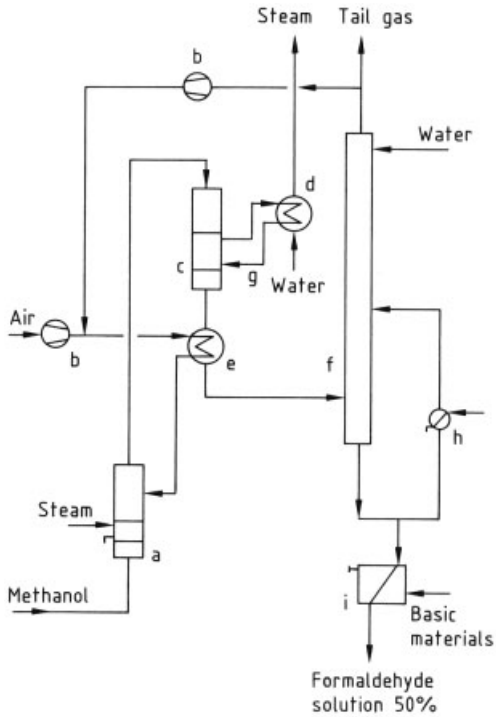

Figure 3. Flowchart of formaldehyde production by the Formox process
a) Evaporator; b) Blower; c) Reactor; d) Boiler; e) Heat exchanger; f) Formaldehyde absorption column; g) Circulation system for heat-transfer oil; h) Cooler; i) Anion-exchange unit

4.3. Comparison of Process Economics

Considering the economic aspects of the three formaldehyde processes in practice, it becomes obvious that the size of the plant and the cost of methanol are of great importance. Generally, the Formox process proves to be advantageous regarding the attainable formaldehyde yield. However, in comparison with the silver process this process demands a larger plant and higher investment costs. For the purpose of a cost comparison, a study was undertaken based on the cost of methanol being $ 200 /t and a plant production capacity of 20,000 t/a of 37 % formaldehyde (calculated 100 %) [1]. Table 5 summarizes the economic data. According to these data the silver process, without the recovery of methanol (cost of formaldehyde $ 378/t), offers the most favorable production costs, followed by the Formox process ($ 387/t) and the silver process with

Table 5. Comparison of economic factors in formaldehyde production processes [1]

		Complete methanol conversion (BASF process)	Incomplete conversion and methanol recovery	Formox process
Total capital investment, $ 10^6		6.6	8.6	9.6
Methanol consumption, t/t		1.24	1.22	1.15
Raw materials, $/t		255	252	227
	Methanol	250	247	232
	Catalyst and chemicals	5	5	7
	Byproduct credit (steam)	not mentioned	not mentioned	12
Utilities, $/t		12	20	13
	LP Steam	3.4	9.5	
	Power purchased	3.4	4.3	8.0
	Cooling water	2.9	2.8	4.0
	Process water	2.4	3.3	1.0
Variable costs, $/t		267	272	240
Direct fixed costs, $/t		27	29	30
Total allocated fixed costs, $/t		18	20	21
Total cash cost, $/t		312	321	291
Depreciation, $/t		33	43	48
Cost of production, $/t		345	364	339
Return of total capital investment (ROI), $/t		33	43	48
Cost of production and ROI, $/t		378	407	387

recovery ($ 407/t). The two latter processes produce a product with $< 1\,\%$ methanol whereas the methanol content in the silver process without recovery lies between 1 – 5 %.

The study takes into consideration the benefit of the production of steam only in the case of the Formox process. If the production of steam is included in the silver process (3 t per tonne CH_2O without and 1.5 t per tonne CH_2O with methanol recovery) better results than demonstrated in Table 5 can be obtained (costs per tonne $ 24 and $ 12 lower, respectively). The proven capacity limits of a plant with only one reactor are about 20 000 t/a (calculated 100 %) with the metal oxide process and about 72 000 t/a with the silver process.

The key feature of the BASF process for the production of 50 wt % formaldehyde is a liquid circulation system in which heat from the absorption unit of the plant is transferred to a stripper column to vaporize the methanol – water feed. Therefore, the process produces excess steam, with simultaneous savings in cooling water.

Plant operation and start-up are simple; the plant can be restarted after a shutdown or after a short breakdown, as long as the temperatures in the stripping cycle remain high. The BASF process has several other advantages. Formaldehyde is obtained from a single pass of the methanol through the catalyst. If a lower formaldehyde concentration is needed (e.g., 40 wt %) the yield can be increased by employing a feedstock of suitably pretreated crude aqueous methanol instead of pure methanol (cf. Section 4.1.1). Deacidification by means of ion exchangers is not necessary. The off-gas does not present any problems because it is burned as a fuel gas in power stations to generate steam or steam and power. The catalyst can be exchanged within 8 – 12 h of plant shutdown to restart and can be regenerated completely with little loss. The plant is compact due to the small volume of gas that is used and the low space requirements; both factors result in low capital investment costs.

Formaldehyde production processes based on incomplete methanol conversion employ a final distillation column to recover the methanol and concentrate the formaldehyde. As shown in Table 5, this means that more steam and cooling water is consumed than in the BASF process. The BASF process has a somewhat lower yield but all other aspects are roughly comparable. Other distinctive features of the incomplete conversion of methanol are the relatively large amount of direct steam introduced into the

feedstock and the lower reaction temperature, which give a somewhat larger amount of hydrogen in the off-gas with a net calorific value of 2140 kJ/m^3. The additional ion-exchange unit also increases production costs.

The Formox process uses excess air in the methanol feed mixture and requires at least 13 mol of air per mole of methanol. A flammable mixture is used for the catalytic conversion. Even with gas recycling, the process must handle a substantial volume of gas, which is 3 – 3.5 times the gas flow in a silver-catalyzed process. Thus, the equipment must have a large capacity to accommodate the higher gas flow. The main disadvantage of the Formox process is that the off-gas is noncombustible, causing substantial costs in controlling environmental pollution. To reduce air pollution to the levels obtained in the silver-catalyzed processes, a Formox plant must burn the tail gas with sulfur-free fuel, with or without partial regeneration of energy by means of steam production. Advantages of the process are its very low reaction temperature, which permits high catalyst selectivity, and the very simple method of steam generation. All these aspects mean in easily controlled process. Plants based on this technology can be very small with annual capacities of a few thousand tons. As a result, plants employing Formox methanol oxidation are most commonly encountered throughout the world. However, if higher capacities are required and a small number of reactors must be arranged in parallel, the economic data favor the processes employing a silver catalyst.

Although approximately 70 % of existing plants use the silver process, in the 1990s new plant contracts have been dominated by the metal oxide technology [1].

4.4. Distillation of Aqueous Formaldehyde Solutions

Since formaldehyde polymerizes in aqueous solutions, the monomer content and thus the vapor pressure of formaldehyde during distillation are determined by the kinetics of the associated reactions.

Vacuum distillation produces a more concentrated bottom product and can be carried out at a low temperature, an extremely low vapor pressure, and an acid pH value of 3 – 3.4 [81]. However, the distillation rates are low, making this procedure uneconomical.

High-pressure distillation at 0.4 – 0.5 MPa and above 130 °C with long columns produces a relatively concentrated overhead product. Efficiency is high, but yields are limited due to the formation of methanol and formic acid via the Cannizzaro reaction [82].

If formaldehyde solutions are subjected to slow *distillation at atmospheric pressure* without refluxing, the distillate has a lower formaldehyde content than the bottom product [21]. If the condensate is refluxed, the ratio of condensate (reflux) to distillate determines the formaldehyde content of the distillate removed [81].

In the case of aqueous formaldehyde solutions that contain methanol, a virtually methanol-free product can be obtained by using distillation columns with a large number of plates and a relatively high reflux ratio. The product is taken from the bottom of the column [83].

4.5. Preparation of Liquid Monomeric Formaldehyde

Two methods have been described for the preparation of liquid monomeric formaldehyde from paraformaldehyde, the first was developed by F. Walker [11] and the second by R. Spence [84]. In Walker's method, liquid formaldehyde is prepared by vaporizing alkali-precipitated α-polyoxymethylene. The resultant vapor is then condensed and the crude liquid condensate is redistilled. The process is performed in an apparatus made of Pyrex glass. A vaporizing tube is charged to about one-half its height with the polymer. The thoroughly dried system is then flushed with dry nitrogen. The vaporizing tube is heated to 150 °C in an oil bath and the condensing tube is chilled in a bath of solid carbon dioxide and methanol. The polymer is vaporized in a slow stream of nitrogen by gradually raising the temperature. Formation of polymer on the tube walls is minimized by winding wire round the tubes and heating with electricity. The crude liquid product, which is opalescent due to precipitated polymer, is then distilled in a slow current of nitrogen.

According to the method of Spence, paraformaldehyde is dried over sulfuric acid in

a vacuum desiccator and introduced into a distillation flask. This flask is connected to a glass condenser via glass tubes with relatively long hairpin turns designed to separate traces of water (Fig. 4). The system is first evacuated by means of a mercury diffusion pump, and the distillation flask is then heated to 110 °C in an oil bath to remove traces of oxygen. The distillate is heated electrically to 120 °C when it flows through the upper parts of the hairpin turns; in the lower parts of the loops, it is cooled to −78 °C by means of a cooling bath. After the valve to the pump is shut and the condenser flask is cooled in liquid air, a colorless solid product condenses. The inlet and outlet tubes of the condenser flask are then sealed with a flame. The contents of the condensing flask liquefy when carefully warmed. The procedure can be repeated to obtain an even purer substance. The liquid formaldehyde that is prepared does not polymerize readily and, when vaporized, leaves only very small traces of polymeric product.

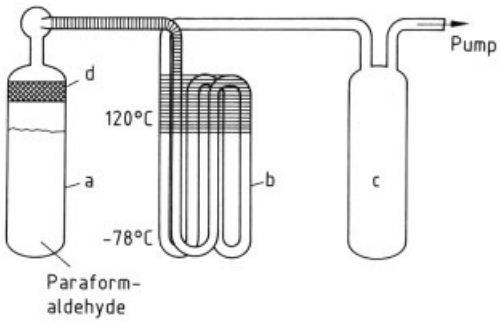

Figure 4. Apparatus for the preparation of liquid monomeric formaldehyde
a) Distillation flask; b) Glass tube with hairpin turns; c) Condenser; d) Glass wool

5. Environmental Protection

As already stated, formaldehyde is ubiquitously present in the atmosphere [85]. It is released into the atmosphere as a result of the combustion, degradation, and photochemical decomposition of organic materials. Formaldehyde is also continuously degraded to carbon dioxide in processes that are influenced by sunlight and by nitrogen oxides. Formaldehyde washed out of the air by rain is degraded by bacteria (e.g., *Escherichia coli*, *Pseudomonas fluorescens*) to form carbon dioxide and water [86].

The major source of atmospheric formaldehyde is the photochemical oxidation and incomplete combustion of hydrocarbons (i.e., methane or other gases, wood, coal, oil, tobacco, and gasoline). Accordingly, formaldehyde is a component of car and aircraft exhaust fumes and is present in considerable amounts in off-gases from heating plants and incinerators. The main emission sources of formaldehyde are summarized in Table 6.

Table 6. Sources emitting formaldehyde into the atmosphere [87]

Emission source	Formaldehyde level
Natural gas combustion	
Home appliances and industrial equipment	2400 – 58 800 μg/m^3
Power plants	15 000 μg/m^3
Industrial plants	30 000 μg/m^3
Fuel-oil combustion	0.0 – 1.2 kg/barrel oil
Coal combustion	
Bituminous	< 0.005 – 1.0 g/kg coal
Anthracite	0.5 g/kg coal
Power plant, industrial, and commercial combustion	2.5 mg/kg coal
Refuse incinerators	
Municipal	0.3 – 0.4 g/kg refuse
Small domestic	0.03 – 6.4 g/kg refuse
Backyard (garden refuse)	up to 11.6 g/kg refuse
Oil refineries	
Catalytic cracking units	4.27 kg/barrel oil
Thermofor units	2.7 kg/barrel oil
Automotive sources	
Automobiles	0.2 – 1.6 g/L fuel
Diesel engines	0.6 – 1.3 g/L fuel
Aircraft	0.3 – 0.5 g/L fuel

The formaldehyde in the exhaust gases of motor vehicles is produced due to incomplete combustion of motor fuel. Formaldehyde may be produced directly or indirectly. In the indirect route, the unconverted hydrocarbons undergo subsequent photochemical decomposition in the atmosphere to produce formaldehyde as an intermediate [88]. The concentration of formaldehyde is higher above densely populated regions than above the oceans as shown in Table 7 [89]. According to a 1976 report of the EPA [89], the proportions of formaldehyde in ambient air are derived from the main emission sources as follows:

Exhaust gases from motor vehicles and airplanes (direct production)	53 – 63 %
Photochemical reactions (derived mainly from hydrocarbons in exhaust gases)	19 – 32 %
Heating plants, incinerators, etc.	13 – 15 %
Petroleum refineries	1 – 2 %
Formaldehyde production plants	1 %

Formaldehyde in confined areas comes from the following sources:

1) Smoking of cigarettes and tobacco products [88], [90], [91]
2) Urea--, melamine--, and phenol–formaldehyde resins in particle board and plywood furniture
3) Urea – formaldehyde foam insulation
4) Open fireplaces, especially gas fires and stoves
5) Disinfectants and sterilization of large surfaces (e.g., hospital floors)

Table 7. Geographical distribution of formaldehyde in ambient air

Location	Formaldehyde concentration (max.), ppm *
Air above the oceans	0.005
Air above land	0.012
Air in German cities	
normal circumstances	0.016
high traffic density	0.056
Air in Los Angeles (before the law on catalytic combustion of exhaust gases came into effect)	0.165

* 1 ppm = 1.2 mg/m^3

Sources generating formaldehyde must be differentiated into those which release formaldehyde for a defined period, cases (1), (4), and (5) and those which release formaldehyde gas continuously, i.e., decomposition of resins as in cases (2) and (3). Many regulations have been issued to limit pollution of the atmosphere with formaldehyde in both general and special applications [92]. Protection against pollution of the environment with formaldehyde must be enforced with due attention to its sources.

The most effective limitation of atmospheric pollution with formaldehyde is the strict observation of the maximum allowable concentration indoors and outdoors. A maximum workplace concentration of 0.5 ppm (0.6 mg/m^3) has, for example, been established in the Federal Republic of Germany [93]. Other limit values and guide values have been specified for formaldehyde levels in outdoor and indoor air. Emission limits for stationary installations have also been established and regulations for specific products have been formulated. Table 8 gives a survey of regulations valid in some countries of the Western world in 1987.

In the Federal Republic of Germany formaldehyde levels and emissions are subjected to stringent regulations. Plants operating with formaldehyde must conform to the plant emission regulations introduced in 1974 which limit formaldehyde in off-gases to a maximum of 20 mg/m^3 for mass flow rates of 0.1 kg/h or more [94]. This presupposes a closed handling procedure. For example, industrial filling and transfer of formaldehyde solutions is carried out by using pressure compensation between communicating vessels. Discharge of formaldehyde into wastewater in Germany is regulated by law since it endangers water and is toxic to small animals [95]. Formaldehyde is, however, readily degraded by bacteria in nonsterile, natural water [96].

A maximum limit of 0.1 ppm formaldehyde in indoor living and recreation areas has been recommended by the BGA (German Federal Health Office) [97]. To avoid unacceptable formaldehyde concentrations in room air, the German Institute for Structural Engineering has issued guidelines for classifying particle board into emission categories E1, E2, and E3, class E3 having the highest emission [98]. The lowest class (E1) is allowed a maximum formaldehyde emission of 0.1 ppm and a maximum formaldehyde content of 10 mg per 100 g of absolutely dry board (as measured by the DIN EN-120 perforator method) [99]. Furthermore, the uses and applications of urea – formaldehyde foams, which are used to some extent for the heat insulation of cavity walls, have been controlled by DIN 18 159 [99] since 1978. No formaldehyde emission is permitted after the construction has dried.

Cigarette smoke contains 57 – 115 ppm of formaldehyde and up to 1.7 mg of formaldehyde can be generated while one cigarette is smoked. If five cigarettes are smoked in a 30 m^3 room, with a low air-change rate of 0.1 (i.e., 10 %) per hour, the formaldehyde concentration reaches 0.23 ppm [88], [91].

The best protection against accumulation of formaldehyde in confined spaces is, however, proper ventilation. The strong smell of formaldehyde is perceptible at low concentration and thus provides adequate warning of its presence. If all manufacturing and application regulations are strictly observed, possible emission of formaldehyde from consumer products

Table 8. International regulations restricting formaldehyde levels

Country	Emission limit		Product-specific regulations
	Outdoor air, ppm	Indoor air, ppm	
Canada		0.1 (1982)	Urea – formaldehyde (UF) foam insulation prohibited. Voluntary program of particle board manufacturers to reduce emission, no upper limit. Registration of infection control agents
Denmark		0.12 (1982)	Guidelines for particle board: max. 10 mg/100 g of absolutely dry board (perforator value). Guidelines for furniture and in situ UF foam. Cosmetic regulations. Prohibited for disinfecting bricks, wood, and textiles if there is contact with food
Federal Republic of Germany	0.02 (MIK_D, 1966) [a] 0.06 (MIK_K, 1966) [b]	0.1 (1977)	Particle board classification. Guidelines (GefStoffV, Gefahrstoffverordnung) for wood and furniture: upper emission limit 0.1 ppm, corresponding to 10 mg/100 mg of absolutely dry board (perforator value); detergents, cleaning agents, and conditioners: upper limit 0.2 %; textiles: compulsory labeling if formaldehyde content >0.15 %. Guidelines for in situ UF foam: upper limit 0.1 ppm. Cosmetic regulations
Finland		0.12 0.24 for pre 1983 buildings (1983)	Upper limit for particle board: 50 mg/100 g absolutely dry board (perforator value). Prohibited as an additive in hairsprays and antiperspirants. Guidelines for cosmetics, but as yet (1987) no EEC directives
Great Britain			Upper limit for particle board : 70 mg/100 g of absolutely dry board (perforator value)
Italy		0.1 (1983)	Cosmetic regulations (July 1985)
Japan			Prohibited as an additive in foods, food packaging, and paints. Guidelines for particle board, textiles, wall coverings, and adhesives
The Netherlands		0.1 obligatory for schools and rented accommodation (1978)	Particle board quality standard on a voluntary basis: upper limit 10 mg/100 g of absolutely dry boad (perforator value). Particle board regulations in preparation
Sweden		0.4 – 0.7 (1977)	Particle board and plywood quality standards: upper limit 40 mg/100 g of absolutely dry board (perforator value)
Switzerland		0.2 (introduced 1984, came into force 1986)	Particle board quality standard on a voluntary basis: upper limit 10 mg/100 g of absolutely dry board (perforator value, Oct. 1985); quality symbol "Lignum CH 10"
Spain			Regulations for in situ UF foam (1984): upper limit 1000 µg/m^3 = 0.8 ppm, 7 days after installation; 500 µg/m^3 = 0.4 ppm, 30 days after installation
United States		0.4 (Minnesota, 1984) [c] 0.4 (Wisconsin, 1982) [c]	UF foam insulation prohibited in Massachusetts, Connecticut, and New Hampshire; upper limit for existing UF-insulated houses in Massachusetts 0.1 ppm (1986). FDA limit for nailhardening preparations: 5 %. Department of housing and urban development (HUD) guidelines for emission from particleboard and plywood for the construction of mobile houses: upper limit 0.3 ppm.

[a] MIK_D = Maximum allowable concentration for constant immission (mean annual value).
[b] MIK_K = Maximum allowable concentration for short-term immission (30 min or 24 h).
[c] Replaced by HUD product standards, 1985.

is very low and will not therefore constitute a human health hazard.

Formaldehyde concentrations in cosmetic products have been limited since 1977, they must be appropriately labeled if they contain > 0.05 wt % formaldehyde [100]. Below this level, formaldehyde does not cause allergic reactions even in sensitive subjects.

6. Quality Specifications and Analysis

6.1. Quality Specifications

Formaldehyde is commercially available primarily in the form of an aqueous (generally 30 – 55 wt %) solution, and in solid form as paraformaldehyde or trioxane (cf. Chap. 11). Formaldehyde solutions contain 0.5 – 12 wt % methanol or other added stabilizers (see Chap. 7). They have a pH of 2.5 – 3.5, the acid reaction being due to the presence of formic acid, formed from formaldehyde by the Cannizzaro reaction. The solutions can be temporarily neutralized with ion exchangers. Typical product specifications for formulations on the European market are listed in Table 9. Other manufacturers' specifications are described in [102–108].

6.2. Analysis

The chemical reactivity of formaldehyde provides a wide range of potential methods for its

Table 9. Typical specifications of commercial formaldehyde solutions [101]

Formaldehyde content, wt %	Methanol content (max), wt %	Formic acid content (max), mg/kg	Iron content (max), mg/kg	Density t, °C	Density g/mL	Added Stabilizer
30	1.5	150	0.8	20	1.086 – 1.090	
37	1.8	200	1	20	1.107 – 1.112	
37	8 – 12	200	1	20	1.082 – 1.093	Methanol
37	1.8	200	1	20	1.108 – 1.112	Isophthalobisguanamine, 100 mg/kg
50	2.0	200	1	55	1.126 – 1.129	
50	2.0	200	1	40	1.135 – 1.138	Isophthalobisguanamine, 200 mg/kg

qualitative and quantitative determination in solutions and in the air.

Qualitative Methods. Qualitative detection of formaldehyde is primarily by colorimetric methods, e.g., [109], [110]. Schiff's fuchsin – bisulfite reagent is a general reagent used for detecting aldehydes. In the presence of strong acids, it reacts with formaldehyde to form a specific bluish violet dye. The detection limit is ca. 1 mL/m^3. Further qualitative detection methods are described in [111].

Quantitative Methods. Formaldehyde can be quantitatively determined by either physical or chemical methods.

Physical Methods. Quantitative determination of pure aqueous solutions of formaldehyde can be carried out rapidly by measuring their specific gravity [27]. Gas chromatography [112], [113] and high-pressure liquid chromatography (HPLC) [114–116] can also be used for direct determination.

Chemical Methods. The most important chemical methods for determining formaldehyde are summarized in [111]. The sodium sulfite method is most commonly used. This method was developed by Lemmé [117] and was subsequently improved by Seyewetz and Gibello [118], Stadtler [119], and others. It is based on the quantitative liberation of sodium hydroxide produced when formaldehyde reacts with excess sodium sulfite:

$$CH_2O + Na_2SO_3 + H_2O \rightarrow HOCH_2SO_3Na + NaOH$$

The stoichiometrically formed sodium hydroxide is determined by titration with an acid [27].

Formaldehyde in air can be determined down to concentrations in the µL/m^3 range with the aid of gas sampling apparatus [120], [121]. In this procedure, formaldehyde is absorbed from a defined volume of air by a wash liquid and is determined quantitatively by a suitable method. The quantitative determination of formaldehyde in air by the sulfite/pararosaniline method is described in [122].

A suitable way of checking the workplace concentration of formaldehyde is to take a relevant sample to determine the exposure of a particular person and to use this in combination with the pararosaniline method. The liquid test solution is transported in a leakproof wash bottle [111]. A commercial sampling tube [123], [124] can also be used, in which the formaldehyde is converted to 3-benzyloxazolidine during sampling. Evaluation is carried out by gas chromatography.

Continuous measurements are necessary to determine peak exposures, e.g., by the pararosaniline method as described in [125].

7. Storage and Transportation

With a decrease in temperature and/or an increase in concentration, aqueous formaldehyde solutions tend to precipitate paraformaldehyde. On the other hand, as the temperature increases, so does the tendency to form formic acid. Therefore, an appropriate storage temperature must be maintained (Table 10). The addition of stabilizers is also advisable (e.g., methanol, ethanol, propanol, or butanol). However, these alcohols can be used only if they do not interfere with further processing, or if they can be separated off; otherwise, effluent problems may be encountered.

The many compounds used for stabilizing formaldehyde solutions include urea [126], melamine [127], hydrazine hydrate [128], methylcellulose [129], guanamines [130], and bismelamines [33]. For example, by adding as little as 100 mg of isophthalobisguanamine [*5118-80-9*] per kilogram of solution, a 40-wt % formaldehyde solution can be stored for at least 100 d at 17 °C without precipitation of paraformaldehyde, and a 50-wt % formaldehyde solution can be stored for at least 100 d at 40 °C [32].

Table 10. Storage temperatures for commercial formaldehyde solutions

Formaldehyde content, wt %	Methanol content, wt %	Storage temperature, °C
30	≤ 1	7 – 10
37	< 1	35
37	7	21
37	10 – 12	6 – 7
50	1 – 2	45 *
50	1 – 2	60 – 65

* Stabilized with 200 mg/kg of isophthalobisguanamine

Formaldehyde can be stored and transported in containers made of stainless steel, aluminum, enamel, or polyester resin. Iron containers lined with epoxide resin or plastic may also be used, although stainless steel containers are preferred, particularly for higher formaldehyde concentrations. Unprotected vessels of iron, copper, nickel, and zinc alloys must not be used.

The flash point of formaldehyde solutions is in the range 55 – 85 °C, depending on their concentration and methanol content. According to German regulations for hazardous substances (Gefahrstoffverordnung, Appendix 6) and Appendix 1 of the EEC guidelines for hazardous substances, aqueous formaldehyde solutions used as working materials that contain ≥ 1 wt % of formaldehyde must be appropriately labeled. The hazard classifications for the transport of aqueous formaldehyde solutions with a flash point between 21 and 55 °C containing > 5 wt % formaldehyde and < 35 wt % methanol are as follows [131]:

GGVS/GGVE, ADR/RID	Class 8, number 63 c
CFR	49: 172.01 flammable liquid
IMDG Code (GGVSee)	Class 3.3
UN No.	1198

Formaldehyde solutions with a flash point > 61 °C and aqueous formaldehyde solutions with a flash point > 55 °C that contain > 5 wt % formaldehyde and < 35 wt % methanol are classified as follows:

GGVS/GGVE, ADR/RID	Class 8, number 63 c
CFR	49: 172.01 combustible liquid
IMDG Code (GGVSee)	Class 9
UN No.	2209

8. Uses

Formaldehyde is one of the most versatile chemicals and is employed by the chemical and other industries to produce a virtually unlimited number of indispensable products used in daily life [132].

Resins. The largest amounts of formaldehyde are used for producing condensates (i.e., resins) with urea, melamine, and phenol and, to a small extent, with their derivatives (see also → Amino Resins; → Phenolic Resins; → Resins, Synthetic). The main part of these resins is used for the production of adhesives and impregnating resins, which are employed for manufacturing particle boards, plywood, and furniture. These condensates are also employed for the production of curable molding materials; as raw materials for surface coating and as controlled-release nitrogen fertilizers. They are used as auxiliaries in the textile, leather, rubber, and cement industries. Further uses include binders for foundry sand, rockwool and glasswool mats in insulating materials, abrasive paper, and brake linings. A very small amount of urea – formaldehyde condensates are used in the manufacture of foamed resins (→ Foamed Plastics, Chap. 4.4., → Foamed Plastics, Chap. 4.5., → Foamed Plastics, Chap. 4.6.) that have applications in the mining sector and in the insulating of buildings.

Use as an Intermediate. About 40 % of the total formaldehyde production is used as an intermediate for synthesizing other chemical compounds, many of which are discussed under separate keywords. In this respect, formaldehyde is irreplaceable as a C_1 building block. It is, for example, used to synthesize 1,4-butanediol [*110-63-4*], trimethylolpropane [*77-99-6*], and neopentyl glycol [*126-30-7*], which are employed in the manufacture of polyurethane

and polyester plastics, synthetic resin coatings, synthetic lubricating oils, and plasticizers. Other compounds produced from formaldehyde include pentaerythritol [*115-77-5*] (employed chiefly in raw materials for surface coatings and in permissible explosives) and hexamethylenetetramine [*100-97-0*] used as a cross-linking agent for phenol – formaldehyde condensates and permissible explosives).

The complexing agents nitrilotriacetic acid [*139-13-9*] (NTA) and ethylenediaminetetraacetic acid [*60-00-4*] (EDTA) are derived from formaldehyde and are components of modern detergents. The demand for formaldehyde for the production of 4,4′-diphenylmethane diisocyanate [*101-68-8*] (MDI) is steadily increasing. This compound is a constituent of polyurethanes used in the production of soft and rigid foams and, more recently, as an adhesive and for bonding particle boards.

The so-called polyacetal plastics (→ Polyoxymethylenes) produced by polymerization of formaldehyde are increasingly being incorporated into automobiles to reduce their weight and, hence, fuel consumption. They are also used for manufacturing important functional components of audio and video electronics equipment [232].

Formaldehyde is also a building block for products used to manufacture dyes, tanning agents, dispersion and plastics precursors, extraction agents, crop protection agents, animal feeds, perfumes, vitamins, flavorings, and drugs.

Direct Use. Only a very small amount of formaldehyde is used directly without further processing. In the Federal Republic of Germany, ca. 8000 t/a are used in this way, which corresponds to ca. 1.5 % of total production. It is used directly as a corrosion inhibitor, in the metal industry as an aid in mirror finishing and electroplating, in the electrodeposition of printed circuits, and in the photographic industry for film development. However, formaldehyde as such is used mainly for preservation and disinfection, for example, in medicine for disinfecting hospital wards, preserving specimens, and as a disinfectant against athlete's foot (→ Disinfectants).

Modern hygiene requires preservatives and disinfectants to prevent the growth of microorgansims which can produce substances that may be extremely harmful to man. As an antimicrobial agent, formaldehyde displays very few side effects, but has a broad spectrum of action. All alternative agents have unpleasant or even dangerous side effects. Moreover, their toxicity has not been as thoroughly investigated as that of formaldehyde, and their spectrum of action is limited (i.e., they do not provide comprehensive disinfectant protection). Another advantage of formaldehyde is that it does not accumulate in the environment since it is completely oxidized to carbon dioxide within a relatively short time. In the cosmetics industry, formaldehyde is employed as a preservative in hundreds of products, for example, soaps, deodorants, shampoos, and nail – hardening preparations; in some of these items, upper limits have been set for the formaldehyde concentration due to its sensitizing effect (cf. Table 8). Formaldehyde solutions are also used as a preservative for tanning liquors, dispersions, crop protection agents, and wood preservatives. Furthermore, formaldehyde is required in the sugar industry to prevent bacterial growth during syrup recovery.

9. Economic Aspects

Formaldehyde is one of the most important basic chemicals and is required for the manufacture of thousands of industrial and consumer products. It is the most important industrially produced aldehyde.

Formaldehyde can seldom, if ever, be replaced by other products. Substitutes are generally more expensive; moreover, their toxicities have been less thoroughly investigated than that of formaldehyde.

Worldwide capacity [1], [231] is approximately 8.7×10^6 t/a in 1996 (see Table 11; the values are based on 100 % formaldehyde); the five largest manufactures account for ca. 25 % of this capacity:

Borden	0.66×10^6 t / a
BASF	0.444×10^6 t / a
Hoechst Celanese	0.38×10^6 t / a
Georgia Pacific	0.38×10^6 t / a
Neste Resins	0.37×10^6 t / a

The three leading countries with a capacity share of about 45 % are:

United States	1.77×10^6 t/a
Germany	1.46×10^6 t/a
Japan	0.65×10^6 t/a

Table 11. Worldwide formaldehyde production capacities in 1996 [1], [231]

Country	Total capacity, 10^3 t/a
Western Europe	3119
Germany	1464
Italy	389
Spain	265
United Kingdom	197
France	126
Sweden	124
Netherlands	115
Others	439
Eastern Europe	1850
North America	2008
United States	1772
Canada	236
South America	253
Mexico	65
Chile	63
Brazil	48
Argentina	44
Others	33
Japan	651

Formaldehyde consumption is ca. 6×10^6 t/a, although present data about capacity use in Eastern Europe are not available. The demand and the estimated average annual growth rate in the Western hemisphere is summarized in Table 12.

Table 12. Consumption of formaldehyde in the United States, Western Europe, and Japan in 1995 [1], [231]

	United States	Western Europe	Japan
Consumption, 10^6 t/a	1.37	2.22	0.52
Use, %			
Urea – formaldehyde resin	32	50	27
Melamine – formaldehyde resin	4	6	
Phenol – formaldehyde resin	24	10	6
Polyoxymethylenes	10	10	24
1,4-Butanediol	11	7	
MDI	5	6	4
Others	14	11	39
Average annual growth rate, %	2.5	1.5	2.0

Formaldehyde and its associated products are used in ca. 50 different branches of industry, as described in Chapter 8.

In the mid 1980s the sales of industrial products derived from formaldehyde was more than DM 300×10^9 in the Federal Republic of Germany [132]. At least 3×10^6 people in the Federal Republic of Germany work in factories that use products manufactured from formaldehyde.

10. Toxicology and Occupational Health

Formaldehyde toxicity was investigated extensively during the last decades and comprehensive reviews are available [233–235]. Formaldehyde is an essential intermediate in cellular metabolism in mammals and humans, serving as a precursor for the biosynthesis of amino acids, purines and thymine. Exogenously administered formaldehyde is readily metabolized by oxidation to formic acid or reacts with biomolecules at the sites of first contact. Inhalation exposure of rats, monkeys and humans to irritant concentrations did not increase blood formaldehyde levels, which were found to be around 80 mM (= 2.4 ppm) in these species.

Formaldehyde gas is toxic via inhalation and causes irritation of the eyes and the mucous membranes of the respiratory tract. Concentration – response relationship following human exposure is given in Table 13.

Table 13. Dose – response relationship following human exposure to gaseous formaldehyde [133], [134]

Effect	Exposure level, ppm
Odor threshold	0.05 – 1.0
Irritation threshold in eyes, nose, or throat	0.2 – 1.6
Stronger irritation of upper respiratory tract, coughing, lacrimation, extreme discomfort	3 – 6
Immediate dyspnea, burning in nose and throat, heavy coughing and lacrimation	10 – 20
Necrosis of mucous membranes, laryngospasm, pulmonary edema	> 50

Aqueous formaldehyde solutions cause concentration dependent corrosion or irritation and skin sensitization. There is no evidence for Formaldehyde to cause respiratory allergy [236].

In *chronic inhalation* studies with rats, mice, hamsters, and monkeys no systemic toxicity occurred in irritant concentrations. Upper respiratory tract irritation ceased at concentrations < ca. 1 ppm. At concentrations above 1 – 2 ppm changes in the nasal mucosa (respiratory epithelium) occur. At high concentrations (15 – 20 ppm) olfactory epithelium, laryngeal mucosa, and proximal parts of the tracheal epithelium might also be affected. The lesions are characterized by epithelial hyperplasia and metaplasia. Studies using other routes of administration also failed to show systemic toxicity or reproductive effects.

Formaldehyde was *genotoxic* in several in vitro test systems. In animals, there are some indications of in vivo genotoxicity in tissues of initial contact (portal of entry) but not in remote organs or tissues. In workers exposed to formaldehyde no systemic genotoxicity and no convincing evidence of local genotoxicity was found.

No evidence of systemic *carcinogenicity* was found after oral dermal and inhalative administration of formaldehyde. Several chronic inhalation studies in rats showed development of nasal tumors starting at concentrations at or above 6 ppm, causing in addition severe chronic epithelial damage in the nasal epithelium [237]. The nonlinear concentration response curve shows a disproportionately high increase in tumor incidence at concentrations of 10 and 15 ppm. The same nonlinear concentration response was observed for DNA protein cross-link (DPX) formation in nasal mucosa, which is a surrogate of formaldehyde tissue "dose", and for increase in cell proliferation in nasal epithelium. This leads to the suggestion that increased cell proliferation is a prerequisite for tumor development [237]. Chronic inhalation studies in mice failed to show statistically significant increases in tumor incidences at similar concentrations while in hamsters no nasal tumors were found. This may be attributed to differences in local formaldehyde tissue dose or lower susceptibility of the species for nasal tumor formation.

In humans numerous *epidemiological studies* failed to give convincing evidence of carcinogenicity [235]. IARC [234] concluded that the epidemiologic data available represent "limited evidence of carcinogenicity" and classified formaldehyde as "probably carcinogenic to humans (Group 2A)". The European Union categorizes the compound as "possibly carcinogenic to humans (Class 3)".

Current occupational exposure limits in different countries vary between 0.3 and 2 ppm [238]. Proposed limit values for indoor air are in the range of 0.1 ppm [239].

11. Low Molecular Mass Polymers

The ability of formaldehyde to react with itself to form polymers depends directly on the reactivity of formaldehyde as a whole. Two different types of formaldehyde polymers are possible and are based on the following structural elements:

1) $-CH_2-O-$
2) $-CH(OH)-$

The polyhydroxyaldehydes consist of the structural unit (2). The highest molecular mass representatives of this group are the sugars. Although these substances can be made by aldol condensation of formaldehyde, they do not revert to formaldehyde on cleavage, and are not discussed in this article.

The representatives of group (1), the real formaldehyde polymers (polyoxymethylenes), revert to formaldehyde on cleavage and, therefore, can be considered as a solid, moisture-free form of formaldehyde. If these linear or cyclic compounds contain no more than eight formaldehyde units, they are defined as low molecular mass polymers. The high molecular mass substances are the real polymers (paraformaldehyde, acetal plastics, see also → Polyoxymethylenes). Chemical and physical analyses of these low and high molecular mass compounds as well as investigation of their chemical reactions led to the elucidation of the molecular structure of polymers in general [135].

11.1. Linear Polyoxymethylenes

Apart from the poly(oxymethylene) glycols, also called poly(oxymethylene) dihydrates or simply polyoxymethylenes, of the formula $HO-(CH_2O)_n-H$, derivatives such as poly(oxymethylene) diacetates $CH_3COO(CH_2O)_nCOCH_3$ and

poly(oxymethylene) dimethyl ethers $CH_3O(CH_2O)_nCH_3$ should be mentioned. Some of their physical properties are given in Table 14. The *n* values of the real low molecular mass polyoxymethylenes are 2 – 8; the *n* values of paraformaldehyde are 8 – 100. However, high polymers with a degree of polymerization $n \geq 3000$ are also obtained. The polyoxymethylenes are also classified as α, β, γ, δ, and ε polymers which are of historical importance. They differ in their degrees of polymerization and in their chemical structures (Table 15). Their toxicology is the same as that of formaldehyde (see Chap. 10).

The lower poly(oxymethylene) glycols are colorless solids with melting points between 80 and 120 °C (Table 14). In contrast to the high molecular mass materials, they dissolve in acetone and diethyl ether without or with only slight decomposition; they are insoluble in petroleum ether. When dissolved in warm water, they undergo hydrolysis to give a formaldehyde solution. The low molecular mass polymers constitute a homologous series, whose properties change continuously with the degree of polymerization.

A freshly prepared, aqueous formaldehyde solution polymerizes to the lower polymers when allowed to stand (see also Section 2.2). Indeed, formaldehyde exists in dilute solution as dihydroxymethylene (formaldehyde hydrate), which in turn undergoes polycondensation to yield low molecular mass poly(oxymethylene) glycols:

$$CH_2O + H_2O \Leftrightarrow HO{-}CH_2{-}OH + n\,HOCH_2OH \Leftrightarrow HO{-}CH_2O{-}(CH_2O)_n{-}H + n\,H_2O$$

Equilibrium is attained under the influence of a hydrogen ion catalyst. At a low temperature and a high concentration, equilibrium favors formation of high molecular mass polymers. However, the major product is of lower molecular mass when the system is heated. The polymers partially separate out, crystallize, and slowly undergo further condensation polymerization [140]. The low molecular mass substances can be further precipitated and isolated by concentrating the solution at low temperature under vacuum conditions; the polymers can be further precipitated by evaporation [141]. The resulting mixture can be separated into the individual substances by exploiting their varying solubilities in different solvents [135].

The transformation of poly(oxymethylene) dihydrates to diacetates and, above all, to diethers produces a remarkable increase in thermal and chemical stability. This is because the unstable hemiacetals at the ends of the chains are eliminated through saturation of the hydroxyl groups. The diethers are stable up to 270 °C in the absence of oxygen and up to 160 °C in the presence of oxygen. These diethers and diacetates are resistant to hydrolysis under neutral conditions, the diethers are also stable in the presence of alkali. Similar to the dihydrates, the properties of the diacetates and diethers also change continuously as the degree of polymerization increases (see Table 14).

Poly(oxymethylene) diacetates are produced by the reaction of paraformaldehyde with acetic anhydride [135]. Pure products are isolated by vacuum distillation, solvent extraction, and crystallization.

The formation of poly(oxymethylene) dimethyl ethers involves the reaction of poly(oxymethylene) glycols or paraformaldehyde with methanol at 150 – 180 °C in the presence of traces of sulfuric or hydrochloric acid in a closed vessel [135]. Alternatively, they can be synthesized by the reaction of formaldehyde dimethyl acetal with either paraformaldehyde or a concentrated formaldehyde solution in the presence of sulfuric acid. This synthesis can be varied by substituting other formaldehyde dialkyl acetals for the dimethyl compound [142].

Paraformaldehyde [*30525-89-4*] was first produced in 1859. This polymer, at first mistakenly called dioxymethylene and trioxymethylene, consists of a mixture of poly(oxymethylene) glycols $HO{-}(CH_2O)_n{-}H$ with $n = 8 - 100$. The formaldehyde content varies between 90 and 99 % depending on the degree of polymerization *n* (the remainder is bound or free water). It is an industrially important linear polyoxymethylene.

Properties. Paraformaldehyde is a colorless, crystalline solid with the odor of monomeric formaldehyde. It has the following physical properties: *fp* 120 – 170 °C, depending on the degree of polymerization; heat of combustion 16.750 kJ/kg (product containing 98 wt % formaldehyde); energy of formation

Table 14. Physical properties of low molecular mass poly(oxymethylene) glycols HO−$(CH_2O)_n$−H and their derivatives

n	Poly(oxymethylene) glycols		Poly(oxymethylene) diacetates			Poly(oxymethylene) dimethyl ethers		
	fp, °C	Solubility in acetone	*fp*, °C	*bp*, °C	ϱ, g/m^3 (13 Pa, 24 °C)	*fp*, °C	*bp*, °C (101.3 kPa)	ϱ, g/cm^3 (25 °C)
2			−23	39 – 40	1.128			
3	82 – 85	Very soluble in the cold	−13	60 – 62	1.158	−69.7	105.0	0.9597
4	82 – 85	Very soluble in the cold	− 3	84	1.179	−42.5	155.9	1.0242
5	95 – 105 (decomp.)	Very soluble in the cold	7	102 – 104	1.195	− 9.8	201.8	1.0671
6			17	124 – 126	1.204	18.3	242.3	1.1003
7		Soluble in the cold	ca. 15	180 – 190				
8		Soluble in the cold	ca. 15	180 – 190				
9	115 – 120 (decomp.)	Soluble when heated	32 – 34		1.216 ∗			

∗ Value at 13 Pa and 36 °C.

Table 15. Structure and synthesis of $\alpha - \varepsilon$ polyoxymethylenes [136]

Polymer	Formula	Synthesis
Paraformaldehyde	HO $(CH_2O)_n$H $n = 8 - 100$	from aqueous formaldehyde solution [137]
α-Polyoxymethylene	HO $(CH_2O)_n$H $n > 100$	from aqueous formaldehyde solution [137]
β-Polyoxymethylene	HO $(CH_2O)_n$H $n > 200$	by heating paraformaldehyde [138]
γ-Polyoxymethylene	$H_3CO\,(CH_2O)_nCH_3$ $n = 300 - 500$	from a methanolic paraformaldehyde solution in the presence of sulfuric acid [139]
δ-Polyoxymethylene	$H_3CO\,[CH_2OC(OH)HO]_nCH_3$ $n = 150 - 170$	by prolonged boiling of γ-polyoxymethylene with water [138]
ε-Polyoxymethylene	HO $(CH_2O)_n$H $n > 300$	by sublimation of 1,3,5-trioxane [138]

177 kJ/mol formaldehyde (product containing 93 wt % formaldehyde); flash point 71 °C; ignition temperature of dust 370 – 410 °C; lower explosion limit of dust 40 g/m^3 (the last three values strongly depend on the particle size); minimum ignition energy 0.02 J.

Table 16. Vapor pressure of formaldehyde (p) released from paraformaldehyde

t, °C	10	21	25	33	37
p, kPa	0.112	0.165	0.193	0.408	0.667
t, °C	43	47	51	58	65
p, kPa	0.943	1.096	1.376	1.808	20.8
t, °C	80	90	100	110	120
p, kPa	32.8	44.1	49.6	53.5	78.3

∗ Values up to 58 °C according to [143] and from 65 – 120 °C according to [144].

Even at ambient temperature, paraformaldehyde slowly decomposes to gaseous formaldehyde (Table 16), a process which is greatly accelerated by heating. Depolymerization is based on a chain "unzipping" reaction which starts at the hemiacetal end groups of the individual molecules. The rate of decomposition therefore depends on the number of end groups, i.e., on the degree of polymerization.

Paraformaldehyde contains only a few acetone-soluble components (lower diglycols). It dissolves slowly in cold water, but readily in warm water where it undergoes hydrolysis and depolymerization to give a formaldehyde solution. Although this solution is generally cloudy as a result of impurities, these can be removed by filtration. Indeed, the resulting solution is identical with the solution obtained on dissolving gaseous formaldehyde in water. Not only heat, but also dilute alkali or acid increase the solubility of paraformaldehyde. Alkali catalyzes formaldehyde cleavage at the chain ends, acid causes additional splitting at the oxygen bridges. The rate constant of dissolution passes through a minimum between pH 2 and 5, it increases rapidly above and below this pH range [145]. The situation becomes more complex at higher concentrations of formaldehyde in the solution

due to the incipient back reaction. Solubility in nonaqueous solvents is also improved in the presence of acidic or alkaline agents as a result of the onset of depolymerization. Paraformaldehyde also dissolves in alcohols, phenols, and other polar solvents, in which depolymerization and solvation can occur.

Production. Paraformaldehyde is prepared industrially in continuously operated plants by concentrating aqueous formaldehyde solutions under vacuum conditions. At first, colloidal, waxy gels are obtained, which later become brittle. The use of a fractionating column through which gases were passed dates from about 1925 [146], [147].

Paraformaldehyde is currently produced in several steps which are carried out at low pressure and various temperatures [148], [149]. Highly reactive formaldehyde is produced under vacuum conditions starting with solutions that contain 50 – 100 ppm of formic acid and also 1 – 15 ppm of metal formates where the metals have an atomic number of 23 – 30 (e.g., Mn, Co, and Cu) [150]. The solutions are processed in thin-layer evaporators [151] and spray dryers [152].

Other techniques such as fractional condensation of the reaction gases in combination with the formaldehyde synthesis process [153] and very rapid cooling of the gases [154] are also applied. Alternatively, formaldehyde-containing gas is brought into contact with paraformaldehyde at a temperature that is above the dew point of the gas and below the decomposition temperature of paraformaldehyde [155]. The product is obtained in the form of flakes when a highly concentrated formaldehyde solution is poured onto a heated metal surface. The hardened product is subsequently scraped off and thoroughly dried [156]. Paraformaldehyde beads are produced by introducing a highly concentrated melt into a cooling liquid (e.g., benzene, toluene, cyclohexane [157]. Acids [158] and alkalis [148], [159] are also added; they apparently accelerate polymerization and lead to the formation of higher molecular mass but less reactive paraformaldehyde.

Highly soluble, highly reactive paraformaldehyde with a low degree of polymerization is very much in demand. It is produced from concentrated, aqueous – alcoholic formaldehyde solutions [160], [161].

Stabilizing agents include hydantoins [162], nicotinamide, and succinamide [163].

Producers. The main producers of paraformaldehyde are Degussa (Federal Republic of Germany), Ticona (owned by Hoechst, United States), Mitsui Toatsu, and Mitsubishi Gas (Japan). Smaller producers include Derivados Forestales (Spain), Synthite (United Kingdom), Alder (Italy), and Lee Chang Yung (Taiwan).

Quality Specifications. An important factor regarding the quality of paraformaldehyde is its tendency to "age", especially under the influence of heat, which results in decreased reactivity. The paraformaldehyde loses its residual water and becomes brittle. This may be prevented by storage below 10 °C; in a moist atmosphere regeneration is possible. After production, the residual moisture in the product should therefore be retained.

Reactivity is determined with the aid of the resorcinol test in which the rate at which paraformaldehyde condenses with resorcinol in an alkaline medium is measured [162]. The formaldehyde content is measured by using the sodium sulfite method (cf. Section 6.2) and the water content by the Karl Fischer method.

Transport and Storage. Paraformaldehyde is stored and transported as granules or a free-flowing coarse powder in bags, containers, and in silos or silo wagons. Cool, dry conditions must be maintained.

Uses. Paraformaldehyde is used in place of aqueous formaldehyde solutions, especially in applications where the presence of water interferes, e.g., in the plastics industry for the preparation of phenol, urea, and melamine resins, varnish resins, thermosets, and foundry resins. Other uses include the synthesis of organic products in the chemical and pharmaceutical industries (e.g., Prins reaction, chloromethylation, Mannich reaction), the production of textile auxiliaries (e.g., for crease-resistant finishes), and the preparation of disinfectants and deodorants.

11.2. Cyclic Polyoxymethylenes

11.2.1. Trioxane

1,3,5-Trioxane [*110-88-3*], trioxymethylene, or 1,3,5-trioxycyclohexane, $C_3H_6O_3$, M_r 90.1, is

the cyclic trimer of formaldehyde and was first produced and identified in 1885. A laboratory oddity for many years, this substance is very important today, especially as an intermediate in the production of acetal resins (polyoxymethylenes).

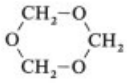

Physical Properties. Trioxane is a white crystalline solid with a characteristic odor resembling that of chloroform. Other physical properties are as follows (see also [164]):

fp	62–63 °C
bp	115 °C
Density of crystals	1.39 kg/m^3
Density of liquid at 65 °C	1.17 kg/m^3
Refractive index n_D^{20} at 65 °C	1.3891
Specific heat	1.23 kJ kg^{-1}K^{-1}
Heat of fusion	222 kJ/kg
Heat of evaporation	452 kJ/kg
Trouton constant	25
Heat of formation	180 kJ/mol formaldehyde
Heat of combustion	16.850 kJ/kg

Vapor pressure

t, °C	25	37.5	86	87	90	114.5	129
p, kPa	1.69	4.16	37.7	39.5	44.0	101.2	161.8

Flash point	45 °C
Ignition temperature	410 °C
Lower explosion limit (38 °C)	3.57 vol %
Upper explosion limit (75 °C)	28.7 vol %
Dynamic viscosity η	
at 65 °C	2.05 mPa s
at 85 °C	0.91 mPa s
Dielectric constant	
at 20 °C	3.2–3.4
at 70 °C	8
Conductance at 78 °C	<1 µS

For the spectroscopic analytical identification of trioxane, see [164]. The compound forms an azeotropic mixture with 30 wt % water, *bp* 91.3 °C.

Trioxane is soluble in water (0.172 g/cm^3 at 18 °C, 0.211 g/cm^3 at 25 °C and completely soluble at 100 °C), alcohol, ketones, organic acids, ethers, esters, phenols, aromatic hydrocarbons, and in chlorinated hydrocarbons; however, it is only sparingly soluble in aliphatic hydrocarbons.

Like dioxane and other cyclic ethers, melted trioxane, (mixed with water if necessary) is an excellent solvent for organic substances. It forms addition compounds with phenol and 1,3,5-trinitrobenzene.

Studies of vapor-liquid equilibria of systems containing formaldehyde, water, methanol, and trioxane are available [220], [224], [226].

Chemical Properties. Trioxane is stable up to 224 °C. Like acetals, it is hydrolyzed by aqueous solutions of strong acids. However, trioxane is inert under neutral and alkaline conditions. For this reason, the usual detection methods for formaldehyde, such as the sodium sulfite method, cannot be applied directly. Formaldehyde in acetals, and thus trioxane, can be determined by heating in water in the presence of a strong acid (e.g., sulfuric acid) to convert the product to monomeric formaldehyde. The formaldehyde can then be determined by the sulfite method (see Section 6.2).

The acid-catalyzed hydrolysis of trioxane is accelerated by using stronger acids, higher acid concentrations, and nonaqueous solvents. When trioxane is heated in nonaqueous systems in the presence of either strong acids (sulfuric acid, hydrochloric acid, zinc chloride, iron chloride etc.) or comonomers, it is converted to high molecular mass oxymethylene homopolymers or oxymethylene copolymers, respectively. This method is applied today in the large-scale production of acetal resins. If additional substances which react with formaldehyde (e.g., phenols, melamine) are also present during these hydrolytic or moisture-free reactions, then the nascent formaldehyde reacts very vigorously with them.

Production. Trioxane is prepared by the trimerization of formaldehyde. At first, the starting materials used were paraformaldehyde or polyoxymethylenes which were heated with acid, usually sulfuric acid [165]. As the commercial importance of trioxane grew, aqueous formaldehyde solutions and high formaldehyde concentrations [166] were introduced. World production in 1987 was 250 000 t, in 1997 about 380 000 t.

The currently preferred process for trioxane production is illustrated in Figure 5 (see also [167]). Commercial aqueous formaldehyde is

concentrated to 60 – 65 wt % in a column (a) under vacuum conditions. The 60 – 65 % formaldehyde solution is fed into the reactor (b). A dilute (6 – 10 %) formaldehyde solution is removed and, if necessary, can be reconcentrated by pressure distillation and returned to the process. In the reactor (b), formaldehyde is trimerized in the presence of a catalyst (e.g., sulfuric acid, phosphoric acid, ion exchanger [168], or heteropoly acid [169]) which is present at a concentration of up to 25 wt % [170]. Technical and kinetic experiments designed to achieve optimal reactor conditions are described in [171]. A mixture of trioxane, formaldehyde, and water is distilled off and, if necessary, concentrated in a distillation column [172]. In the extraction column (c) [173], the trioxane is extracted from the aqueous mixture by using methylene dichloride [*75-09-2*] [166] or another inert, water-immiscible solvent (benzene [174], 1,2-dichloroethane [175], or nitrobenzene [176]). The formaldehyde – water mixture is returned to the concentrating column (a). The trioxane – methylene dichloride mixture is separated in the distillation column (d). The crude trioxane is obtained at the bottom of the column and the methylene dichloride as overhead product. The trioxane is subsequently purified by distillation in column (e). The solvent methylene dichloride is freed of impurities in the solvent purification step (f) and recycled to the extraction column (c).

Impurities and byproducts (formic acid, methyl formate, dimethylformal, and formaldehyde) can be eliminated by washing with sodium hydroxide solution after the extraction step [177]. Emulsification of the reaction mixture in an inert, nonvolatile liquid, e.g., paraffin oil, may reduce the formation of paraformaldehyde in the reactor [178]. Addition of inert organic substances with low dielectric constants to the trimerization mixture increases conversion [179]. Special processes in a synthesis reactor have also been described [180]. Crystallization of liquid trioxane and contact with an inert gas stream can be used for further purification, especially from higher homologues of bis (methoxymethyl) ether [181].

In the last years a new method has been proposed to remove trioxane from the mixture of trioxane, formaldehyde, and water. After trimerization formaldehyde is separated by pervaporation, e.g., with a polyether block amide membrane [182], [183].

Since water is added with the aqueous formaldehyde solution, the energy requirement for trioxane production is very high (14 – 17 t of steam per ton of trioxane). This may be reduced by using a process based on the trimerization of gaseous formaldehyde with Lewis acids [184] and catalyzed by vanado- and/or molybdo and/or wolframo-phosphoric acids [185–187].

Trioxane can be produced by recycling of poyoxymethylenes (POM) [188], [189].

Producers. Ticona (owned by Hoechst, Federal Republic of Germany and United States), Ultraform (owned by BASF and Degussa, Federal Republic of Germany and United States), Polyplastics (owned by Hoechst and Daicel, Japan), Mitsubishi Gas Chemical (Japan), Asahi (Japan), Korean Engineering Plastics (owned by Tong Yang Nylon and Mitsubishi Gas Chemical, Korea), Taiwan Engineering Plastics, Tepco (owned by Polyplastics and Chang Chun Group, Taiwan), Formosa Plastics (Taiwan), Tarnoform (Poland). The world capacity is about 400 000 t.

Uses. The production of plastics (polyoxymethylene plastics, POM) is the most important commercial application of trioxane and requires extremely pure material. Since trioxane depolymerizes to produce formaldehyde, it can be used in almost all formaldehyde reactions, especially when anhydrous formaldehyde is desired. Trioxane is used, e.g., as a textile auxiliary; in various cross-linking agents, for instance for pitch fibers in manufacturing of carbon fibers; for stabilization and reinforcement of wood to produce musical instruments; and in distillable binders in ceramic compositions for shaped articles, e.g., in an injection moulding process. It is also employed in the stabilization and deodorization of waxes, as an anticorrosive additive for metals (especially when corrosion is caused by chlorinated hydrocarbons) [190], as an additive in photographic developers, as a solid fuel, in coating removers and, because of its hydrolysis to formaldehyde in presence of acid, in disinfectants.

Storage and Transport. Liquid trioxane is stored and transported in heated tanks. Solid tri-

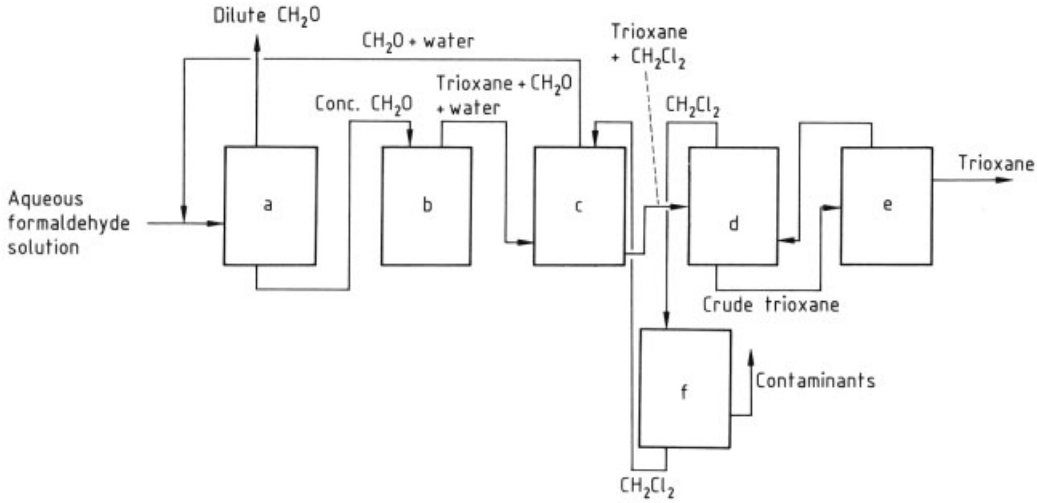

Figure 5. Trioxane production
a) Concentration column; b) Reactor; c) Extraction column; d, e) Distillation columns; f) Solvent purification

oxane is transported in drums, trioxane flakes are transported in drums and bags.

Toxicology. Trioxane is not known to display any special toxic reactions; its oral LD_{50} value in rats is > 3200 mg/kg. Hydrolysis in the presence of acid produces formaldehyde (see Chap. 10).

11.2.2. Tetraoxane

1,3,5,7-Tetraoxane [*293-30-1*], tetraoxymethylene, or 1,3,5,7-tetroxocane, $C_4H_8O_4$, M_r 120, is the cyclic tetramer of formaldehyde and is a crystalline solid (*fp* 112 °C). Tetraoxane vapors are stable even at 200 °C (*bp* at 101.3 kPa, 163 – 165 °C).

Tetraoxane is produced by heating a water-insoluble, high molecular mass poly(oxymethylene) diacetate [165]. In the tetramerization of formaldehyde, metal sulfates may be used as catalyst [191].

Tetraoxane accumulates as a high-boiling substance in the distillation residue obtained during the production of trioxane. It is also obtained in the polymerization of trioxane and can be isolated by extraction with, for example, ethylbenzene [192]. The formation of tetraxane by irradiation of carbon monoxide and hydrogen has also been described [193].

Tetraoxane is a more favorable textile auxiliary for achieving crease resistance, e.g., for cellulose fibers, than trioxane because of its higher boiling point. Like trioxane, tetraoxane can also be employed in the production of acetal plastics.

11.2.3. Higher Cyclic Polyoxymethylenes

Higher cyclic polyoxymethylenes can also undergo polymerization. The tendency to polymerize in the melt decreases with increasing ring size.

1,3,5,7,9-Pentoxecane [*16528-92-0*], pentaoxane, pentaoxacyclodecane, or pentaoxymethylene, $C_5H_{10}O_5$, the cyclic pentamer of formaldehyde, is a crystalline product (*fp* 55 – 56 °C). The polymerization of trioxane with the aid of a tin tetrachloride catalyst produces several cyclic homologues including pentoxecane [194].

The fractional distillation and subsequent preparative gas chromatography of residues from the commercial trioxane polymerization affords 1,3,5,7,9-pentoxecane [195].

Although *hexoxecane* (six formaldehyde units) has not clearly been identified in the residue, *Pentadecoxane* (15 formaldehyde units) has been unambiguously detected (*fp* 68 – 70 °C). Pentoxecane is used in addition

to tetraoxane in patent formulations for the shrinkproofing of cotton [196].

12. Formaldehyde Cyanohydrin

Formaldehyde cyanohydrin [*107-16-4*], $CH_2(OH)CN$, M_r 57.05, is also known as hydroxyacetonitrile, glycolonitrile, or glycolic nitrile. It is less commercially important than acetone cyanohydrin (→ Acetone, Chap. 11.1.) but of about the same importance as the cyanohydrin derivatives of acetaldehyde, benzaldehyde, and ethylene.

Physical Properties. Formaldehyde cyanohydrin is a colorless liquid with an odor similar to that of hydrogen cyanide. It is soluble in water, ethanol, and diethyl ether but insoluble in chloroform and benzene. For IR spectrum see [197]. Other physical properties follow:

mp	−72 °C
bp at	
101.3 kPa	183 °C (slight decomp.)
3.2 kPa	119 °C
0.4 kPa	76.5 – 78 °C
Relative density d_4^{19}	1.1039
n_D^{20}	1.4112
Electrolytic dissociation constant k at 25 °C	0.843×10^{-5}
Dielectric constant	178

Chemical Properties. Like all cyanohydrins, reactions can occur at both the nitrile and the hydroxyl groups of formaldehyde cyanohydrin. Heating at 100 °C with water yields hydroxyacetamide, acid-catalyzed hydrolysis yields hydroxyacetic acid. Reaction with absolute ethanol in the presence of hydrochloric acid produces ethyl hydroxyacetate [198]. *N*-Substituted amides can be synthesized by heating formaldehyde cyanohydrin with amines in water [199]. Catalytic hydration (Ni – Al catalysts) of the nitrile group gives amines. The resulting mixtures can be separated into primary, secondary, and tertiary β-hydroxyethyleneamines [200].

The hydroxyl group of formaldehyde cyanohydrin can be replaced by other electronegative groups. For example, reaction with ammonia gives amino nitriles including nitrilotriacetonitrile $N(CH_2CN)_3$ which is an intermediate in the synthesis of nitrilotriacetic acid (NTA) [201]. Ethylenediaminetetraacetonitrile is similarly synthesized from formaldehyde cyanohydrin and ethylene diamine in the presence of sulfuric acid as a catalyst. Subsequent hydrolysis yields ethylenediaminetetraacetic acid (EDTA) [201]. The reaction of cyanohydrins with ammonia is a step in the Strecker synthesis of amino acids. Hydration of formaldehyde cyanohydrin and subsequent treatment with an aqueous solution of sodium cyanide and ammonia (Bucherer reaction, a variant of the Strecker synthesis) leads to *DL*-serine [202].

Formaldehyde cyanohydrin reacts with hydrogen halides or phosphorus pentachloride to form α-halonitriles. In aqueous solutions at pH 8 and a temperature of 10 °C or less, it trimerizes to form 4-amino-2,5-bis(hydroxymethyl)-pyrimidin-5-ol [203]. Hydantoin can be synthesized by reacting formaldehyde cyanohydrin with carbon dioxide, ammonia, or ammonium carbonate at high pressure [204], [205].

Production. Formaldehyde cyanohydrin is formed in 79.5 % yield when 37 % formaldehyde is mixed with a stoichiometric amount of aqueous hydrogen cyanide for 1 h at 2 °C in the presence of sodium hydroxide as a catalyst [206]. Patents also describe synthesis by (1) heating formaldehyde with hydrogen cyanide at 250 °C [207] or in the presence of sulfur dioxide at pH 1.5 – 2.0 in an aqueous medium [208] and (2) by reacting acetonitrile with oxygen and with or without water at high pressure in the presence of various vanadium oxides [209–211].

Formaldehyde cyanohydrin is generally handled as an aqueous solution but it can also be isolated in the anhydrous state by ether extraction, drying, and vacuum distillation [212]. Although the extremely pure product is reported to be very stable [213], addition of a small amount of stabilizer (e.g., monochloroacetic, mercaptoacetic, lactic, sulfonilic, and phosphoric acids) is advisable [214]. Monochloroacetic acid is especially suitable because it codistils with the product. Cyanohydrin with a lower degree of purity is not very stable, especially at high pH values, this should be borne in mind when handling anhydrous formaldehyde cyanohydrin (see also [203]).

Transport and Storage. Formaldehyde cyanohydrin is used almost exclusively as an

intermediate and is normally produced at the site where it is to be processed. However, it is also supplied as a 70 % aqueous solution and is transported in steel containers and tank cars. Attention should be paid to stabilization during transport and storage. Impure formaldehyde cyanohydrin can decompose to give formaldehyde and hydrogen cyanide, attention should therefore be paid to explosion and fire hazards.

Toxicology. Formaldehyde cyanohydrin, like all cyanohydrins, is extremely toxic when inhaled or ingested and moderately toxic when absorbed via the skin [215]. All forms of contact should be avoided because of the possible formation of hydrogen cyanide or cyanides.

13. References

1. Chem. Systems Inc.: "Formaldehyde" (April 1996).
2. J. C. D. Brand, *J. Chem. Phys.* **19** (1951) 377; *J. Chem. Soc.* 1956, 858 – 872; *J. Chem. & Ind.* 1955, no. 7, 167. G. H. Dieke, G. B. Kistiakowsky, *Proc. Natl. Acad. Sci., U.S.A.* **18** (1932) 367 – 372; *Phys. Rev.* **45** (1934) 4 – 28. P. J. Dyne, *J. Chem. Phys.* **20** (1952) 811 – 818. W. C. Price, *Phys. Rev.* **46** (1934) 529; *J. Chem. Phys.* **3** (1935) 256 – 259. R. I. Reed, *Trans. Faraday Soc.* **52** (1956) 1194. G. W. Robinson, V. E. Digiorgio, *Can. J. Chem.* **36** (1958) 31 – 38. S. A. Schou, V. Henri, *C. R. Hebd. Seances Acad. Sci.* **182** (1926) 1612 – 1614; **186** (1928) 690 – 692, 1050 – 1052. *Nature London* **118** (1926) 225; *Z. Phys.* **49** (1928) 774 – 826. V. Henri, *J. Chim. Phys. Phys. Chim. Biol.* **25** (1928) 665 – 721; **26** (1929) 1 – 43. P. Torkington, *Nature London* **163** (1949) 446. R. S. Mulliken, *J. Chem. Phys.* **3** (1935) 514.
3. Th. Förster, *Fluoreszenz organ. Verbindungen,* Vandenhoeck u. Rupprecht, Göttingen 1951, pp. 49, 97, 125, 143.
4. H. H. Blau, H. H. Nielsen, *J. Mol. Spectrosc.* **1** (1957) 124 – 132. D. W. Davidson, B. P. Stoicheff, H. J. Bernstein, *J. Chem. Phys.* **22** (1954) 289 – 294. K. B. Harvey, J. F. Ogilvie, *Can. J. Chem.* **40** (1962) 85 – 91. I. C. Hisatsune, D. F. Jr. Eggers, *J. Chem. Phys.* **23** (1955) 487 – 492. R. H. Pierson, A. N. Fletcher, E. St. Clair Gantz, *Anal. Chem.* **28** (1956) 1218 – 1239. H. M. Randall, R. G. Fowler, N. Fuson, J. R. Dangl: *Infrared Determination of Organic Structures,* D. Van Nostrand Company, New York 1949, pp. 49, 51, 53, 55, 56. W. G. Schneider, H. J. Bernstein, *Trans. Faraday Soc.* **52** (1956) 13 – 18. W. E. Singer, *Phys. Rev.* **71** (1947) 531 – 533.
5. J. H. Hibben: *The Raman Effect and its Chemical Applications,* Reinhold Publ. Co., 1939, p. 187.
6. J. K. Bragg, A. H. Sharbaugh, *Phys. Rev.* **75** (1949) 1774 – 1775. R. B. Lawrance, M. W. P. Strandberg, *Phys. Rev.* **83** (1951) 363. T. Oka, H. Hirakawa, K. Shimoda, *J. Phys. Soc. Japan* **15** (1960) 2265 – 2279.
7. R. Spence, W. Wild, *J. Chem. Soc.* 1935 506 – 509.
8. J. Legrand, R. Delbourgo, P. Lafitte, *C. R. Hebd. Seances Acad. Sci.* **249** (1959) 1515 – 1516. M. Vanpée, *Bull. Soc. Chim. Belg.* **64** (1955) 235 – 263.
9. National Fire Protection Association: *Fire Protection Guide on Hazardous Materials,* 5th ed., 49 Hazardous Chem. Data, Boston 1973, p. 150.
10. S. Sapgir, *Bull. Soc. Chim. Belg.* **38** (1929) 392 – 408.
11. J. F. Walker, *J. Am. Chem. Soc.* **55** (1933) 2821 – 2825.
12. S. Bezzi, A. Iliceto, *Chim. Ind. (Milan)* **33** (1951) 212 – 217. A. Iliceto, *Gazz. Chim. Ital.* **81** (1951) 915 – 932, **84** (1954) 536 – 552. F. J. Walker, *Formaldehyde,* 3rd ed., Reinhold Publ. Co., New York 1964, p. 62. E. Koberstein, K. P. Müller, G. Nonnenmacher, *Ber. Bunsenges. Phys. Chem.* **75** (1971) 549 – 553.
13. J. F. Walker, *J. Phys. Chem.* **35** (1931) 1104 – 1113.
14. A. Iliceto, *Gazz. Chim. Ital.* **81** (1951) 786 – 794.
15. H. Schecker, G. Schulz, *Z. Phys. Chem. (NF)* **65** (1969) 221 – 224.
16. A. Iliceto, *Ann. Chim. (Rome)* **39** (1949) 703 – 716.
17. E. W. Blair, W. Ledbury, *J. Chem. Soc.,* 1925, 26 – 40. W. Ledbury, E. W. Blair, *J. Chem. Soc.* 1925, 2832 – 2839.
18. P. Skell, H. Suhr, *Chem. Ber.* **94** (1961) 3317 – 3327.
19. W. Dankelman, J. M. H. Daemen, *Anal. Chem.* **48** (1976) 401. Z. Fiala, M. Navratil, *Collect. Czech. Chem. Comm.* **39** (1974) 2200 – 2205. D. A. Young, *unpublished results,* Celanese Research Company, Summit, N. J. 1978.

20. W. Ledbury, E. W. Blair, *J. Chem. Soc.* 1925, 33 – 37, 127, 2834 – 2835. J. F. Walker, T. J. Mooney, *unpublished data,* Du Pont. B. S. Lacy, unpublished data, Du Pont.
21. E. L. Piret, M. W. Hall, *Ind. Eng. Chem.* **40** (1948) 661 – 672.
22. G. Maue, *Pharm. Ztg.* **63** (1918) 197. H. Gradenwitz, *Chem. Ztg.* **42** (1918) 221; *Pharm. Ztg.* **63** (1918) 241. G. Natta, M. Baccaredda, *G. Chim. Ind. Appl.* **15** (1933) 273 – 281.
23. S. J. Green, R. E. Vener, *Ind. Eng. Chem.* **47** (1955) 103 – 109.
24. *Kirk-Othmer,* **11,** 231 – 250.
25. *Organikum,* VEB Deutscher Verlag der Wissenschaft, Berlin 1986.
26. H. Gault, *Chim. Ind. (Paris)* **67** (1952) 41 – 64.
27. J. F. Walker: *Formaldehyde, ACS Monographic Series,* 3rd ed., Reinhold Publ. Co., New York, Amsterdam, London 1967.
28. J. G. Calvert, E. W. R. Steacie, *J. Chem. Phys.* **19** (1951) 176 – 182.
29. W. A. Bone, H. L. Smith, *J. Chem. Soc.,* 1905, 910 – 916.
30. M. J. Marshall, D. F. Stedman, *Trans. R. Soc. Can.* Sect. 3, 17 (1923) 53.
31. Y. Miyazaki, J. Yasumori, *Bull. Chem. Soc. Japan,* **40** (1967) 2012.
32. H. Tropsch, O. Roehlen, *Abh. Kenntnis Kohle* **7** (1925) 15.
33. BASF, DE 2 365 180, 1973 (H. Diem et al.); DE 2 358 856, 1973 (H. Diem et al.)
34. Degussa, DE-OS 3 143 920, 1981 (P. Werle et al.).
35. P. Sabatier, J. B. Senderens, *C. R. Hebd. Seances Acad. Sci.* **137,** (1903) 301 – 303.
36. *Beilstein,* **E III, 1,** 2549 ff.
37. O. Loew, *Ber. Dtsch. Chem. Ges.* **22** (1889) 470 – 478.
38. H. Staudinger et al., *Justus Liebigs, Ann. Chem.* **474** (1929) 254 – 255. H. J. Prins, *Rec. Trav. Chim. Pays Bas* **71** (1952) 1131 – 1136.
39. H. S. Fry, *Rec. Trav. Chim. Pays Bas* **50** (1931) 1060 – 1065.
40. W. M. Lauer, L. M. Langkammerer, *J. Am. Chem. Soc.* **57** (1935) 2360 – 2362.
41. GAF, US 2 232 867, 1941 (W. Reppe, E. Keyssner).
42. L. Orthner, E. Gerisch, *Biochem. Z.* **259** (1933) 30 – 52.
43. E. Katzschmann, *Ber. Dtsch. Chem. Ges.* **77** (1944) 579 – 585.
44. H. Gilman: *Org. Synthesis Coll.,* **vol. I,** J. Wiley & Sons Inc., New York 1932, p. 514.
45. R. Behrend, J. Schmitz, *Justus Liebigs Ann. Chem.* **277** (1893) 338. A. Windus, *Chem. Ber.* **42** (1909) 760.
46. G. Blanc, *Bull. Soc. Chim. Belg.* **33** (1923) 313 – 319.
47. A. R. Chauvel et al., *Hydrocarbon Process.* **52** (1973) 179.
48. J. V. Hightower, *Chem. Eng. (N.Y.)* **55** (1948) 136 – 139;Celanese Co., US 2 570 216, 1949 (H. K. Dice, R. L. Mitchell) US 2 570 217, 1949 (H. K. Dice, R. L. Mitchell).
49. Ruhrchemie, US 3 255 238, 1962 (O. Roelen, W. Rottig).
50. H. Tadenumar et al., *Hydrocarbon Process* . 45 (1966), 195 – 196.
51. V. N. Ipatieff, G. S. Monroe, *J. Am. Chem. Soc.* **67** (1945) 2168 – 2171.
52. F. Fischer, *Oel Kohle* **39** (1943) 521 – 522. Chemische Werke Hüls, DE-OS 2 201 429, 1972 (R. Bröckhaus).G. E. Haddeland, *Formaldehyde,* Stanford Research Institute, Menlo Park, California 1967, p. 56, 107 – 135. F. J. Walker, *Formaldehyde,* 3rd ed., Reinhold Publ. Co., New York 1964, pp. 25 – 26.
53. D. D. Mehta, W. W. Pan, *Hydrocarbon Process.* **45** (1971), 115 – 120.
54. BASF, DE 1 277 834, 1966 (V. Gerloff et al.).
55. BASF, DE 2 034 532, 1970 (H. Hohenschutz et al.).
56. E. Jones, G. G. Fowlie, *J. Appl. Chem.* **3** (1953) 206 – 213.
57. V. I. Atroshchenko, I. P. Kushnarenko, *Int. Chem. Eng.* **4** (1964) 581 – 585.
58. V. N. Gavrilin, B. I. Popov, *Kinet. Catal. (Engl. Transl.)* **6** (1965) 799 – 803.
59. H. Sperber, *Chem. Ing. Tech.* **41** (1969) 962 – 966. Heyden Chem. Corp., US 2 465 498, 1945 (H. B. Uhl, I. H. Cooper).
60. BASF, DE 2 442 231, 1974 (G. Halbritter et al.).
61. H. Diem, *Chem. Eng. N.Y.* **85** (1978) 83.
62. BASF, DE 0 150 436, 1984 (A. Aicher et al.).
63. BASF, DE 2 322 757, 1973 (A. Aicher et al.).
64. BASF, DE 2 655 321, 1976 (A. Aicher et al.).
65. BASF, DE E 2 444 586, 1974 (A. Aicher et al.).
66. *Chem. Week* **105** (1969) 79. D. G. Sleemann, *Chem. Eng. N.Y.* **75** (1968) no. 1, 42 – 44. M. Weimann, *Chem. Eng. N.Y.* **77** (1970) no. 3, 102 – 104. *Hydrocarbon Process.* **52** (1973) 135, 179. J. H. Marten, M. T. Butler, *Oil Gas J.* **72** (1974) no. 10, 71 – 72. Du Pont, FR 1 487 093, 1967.

67. Du Pont, US 2 519 788, 1947, (W. A. Payne); 3 959 383, 1974 (E. S. Northeimer); 4 076 754, 1978 (G. L. Kiser, B. G. Hendricks). Allied Chemical, US 2 462 413, 1943 (W. B. Meath).
68. Borden, US 3 629 997, 1970 (C. W. DeMuth).
69. Barrett Comp., US 1 383 059, 1921 (G. C. Bailey, A. E. Craver). Bakelite Corp., US 1 913 405, 1933 (V. E. Meharg, H. Adkins). Montecatini, US 3 198 753, 1965 (F. Traina). Perstorp AB, GB 1 080 508, 1967 (S. A. Bergstrand).
70. G. D. Kolovertnov, G. K. Boreskov et al., *Kinet. Catal./Engl. (Tansl.)* **6** (1965) 950 – 954; **7** (1966) 125 – 130. Reichhold Chemicals, US 2 973 326, 1954 (T. S. Hodgins, F. J. Shelton). Montecatini, AT 218 513, 1959 (G. Natta et al.).
71. Lummus Comp., US 3 408 309, 1964 (A. W. Gessner). Reichhold Chem. Inc., US 2 812 309, 1954 (C. L. Allyn et al.); 3 855 153, 1973 (G. M. Chang).
72. P. Jiru et al., *Proc. 3rd Intern. Congr. Catalysis* Amsterdam, 1 (1965) 199 – 213. M. Dente et al., *Chim. Ind. (Milan)* **46** (1964) 1326 – 1336; **47** (1965) 359 – 367; **47** (1965) 821 – 829.
73. Du Pont, US 2 436 287, 1948 (W. F. Brondyke, J. A. Monier Jr.); GB 589 292, 1947 (W. F. Brondyke, J. A. Monier Jr.).
74. *Chem. Eng. N.Y.* **61** (1954), 109 – 110.
75. *Chem. Process Eng. (London)* **51** (1970) 100 – 111.
76. Lummus Comp., US 3 277 179, 162 (M. C. Sze).
77. Montecatini, G. Greco, U. Soldano, *Chem. Ing. Tech.* **31** (1959) 761 – 765.
78. W. Exner et al., *Chem. Anlagen + Verfahren,* 1973, 87 – 92.
79. H. R. Gerberich et al., Celanese Chemical Comp., Inc., *Kirk-Othmer,* **11,** 240.
80. C. W. Horner, *Chem. Eng. N.Y.* **84** (1977) no. 14, 108 – 110.
81. A. Iliceto, *Chim. Ind. (Milan)* **36** (1954) 523 – 528. B. Olsson, S. G. Svennson, *Trans. Inst. Chem. Eng.* **53** (1975) 97 – 105. Cities Service Oil Co., US 2 665 241, 1948 (L. G. Willke et al.).
82. M. I. Faberov, V. A. Speranskaya, *Zh. Prikl. Khim. (Leningrad)* **28** (1955) 205 – 208. J. Meissner, FR 1 255 022, 1960.
83. S. J. Green, R. E. Vener, *Ind. Eng. Chem.* **47** (1955) 103 – 109. Du Pont, GB 931 688, 1961 (C. H. Manwiller, J. B. Thompson). Sumitomo, GB 869 764, 1959. M. W. Hall, E. L. Piret, *Ind. Eng. Chem.* **41** (1949) 1277 – 1286.
84. R. Spence, W. Wild, *J. Chem. Soc.* 1935, 338 – 340.
85. *Formaldehyd,* Gemeinsamer Bericht des Bundesgesundheitsamtes, der Bundesanstalt für Arbeitsschutz und des Umweltbundesamtes, Familie und Gesundheit, vol. 148, Verlag Kohlhammer, Stuttgart 1984.
86. G. Bringmann, R. Kühn, *Gesund. Ing.* **11** (1960) 337 – 339.
87. IARC (Int. Agency f. Res. on Cancer): *Formaldehyde,* **vol. 29,** Lyon 1982, pp. 346 – 389.
88. US National Academy of Sciences, *Formaldehyde – An Assessment of its Health Effects,* Washington, D.C. 1980.
89. J. F. Kitchens et al., *Investigation of Selected Potential Environmental Contaminants, Formaldehyde,* EPA 560/2-76-002, US Environmental Protection Agency, Office of Toxic Substances, Washington, D.C. 1976.
90. A. Weber-Tschopp et al., *Int. Arch. Occup. Environ. Health* **39** (1977) 207 – 218.
91. WKI-Report No. 13/82, Fraunhofer Institut für Holzforschung, Wilhelm-Klauditz-Institut, Braunschweig (1982).
92. W. Lohrer, N. J. Nantke, R. Schaaf, *Staub-Reinhalt, Luft* **45** (1985) no. 5, 239 – 247.
93. Technische Regeln für Gefahrstoffe TRGS 900; Oktober 1996; BArb. Bl. Nr. 10/1996, p. 88.
94. Technische Anleitung zur Reinhaltung der Luft, TA-Luft, Feb. 27, 1986, GMBl. 1986, p. 95.
95. Verwaltungsvorschrift wassergefährdender Stoffe, April 18, 1996, GMBl. 1996, p. 327.
96. U. Pagga, *Vom Wasser* **55** (1980) 313 – 326 Bundesministerium für Forschung und Technologie, Forschungsbericht (02-WA 822), Wassertechnologie, Dec. 1980.
97. H. Petri, H. L. Thron, J. Wegner, *Grenzwert-Bestimmung für Formaldehyde in der Innenraumluft,* BGA Jahresbericht 1977; Bundesgesundhbl. 28 (1985) no. 6.
98. Ausschuß für Einheitliche Technische Baubestimmungen: *Richtlinien über die Verwendung und Klassifizierung von Spanplatten bezüglich der Formaldehydabgabe,* Beuth Verlag, Berlin, Köln 1980.
99. ‚Beuth Vertrieb , DIN EN 120,Perforator-Methode.
100. *Kosmetikverordnung,* Dec. 16 1977, BGBl. I, p. 2589.

101. *Formaldehyde Product Bulletin,* Badische Anilin- & Soda-Fabrik AG, Ludwigshafen, Sept. 1975.
102. *Formaldehyde Product Bulletin,* Celanese Chemical Co., Dallas.
103. *Formaldehyde Product Bulletin,* Du Pont, Wilmington, Del.
104. *Formaldehyde Product Bulletin,* Georgia-Pacific Corp., Portland, Oregon.
105. *Formaldehyde Product Bulletin,* Deutsche Gold- und Silber-Scheideanstalt AG, Frankfurt, Germany.
106. *Formaldehyde Product Bulletin,* Borden Chemical Co., Columbus, Ohio, Feb. 1978.
107. *Formox Process for Producing Formaldehyde, Product Bulletin,* Reichhold Chemicals, White Plains, N.Y.
108. *Formaldehyde Product Bulletin,* ICI, Birmingham, UK, Feb. 1977.
109. G. Denigès, *J. Pharm. Chim.* **6** (1896) 193.
110. G. Denigès, *C. R. Hebd. Seances Acad. Sci.* **150** (1910) 529 – 531.
111. H. Petersen, N. Petri, *Melliand Textilber.* **66** (1985) 217 – 222, 285 – 295, 363 – 369.
112. H. L. Gruber, H. Plainor, *Chromatographia* **3** (1970) 490.
113. L. Gollob, J. D. Wellons: "Analytical Methods for Formaldehyde", *For. Prod. J.* **30** (1980) 27 – 35.
114. K. Mopper, W. L. Stahovec, *J. Chromatogr.* **256** (1983) 243 – 252.
115. K. Kuwata, H. Uebori, H. Yamasaki, Y. Kuge, *Anal. Chem.* **55** (1983) 2013 – 2016.
116. K. Fung, D. Grosgean, *Qual. Chem.* **53** (1981) 168 – 171.
117. G. Lemmé, *Chem. Ztg.* **27** (1903) 896; *Chem. Zentralbl.* **II** (1903) 911.
118. A. Seyewetz, Gibello, *Bull. Soc. Chim. Fr.* **31** (1904) 691 – 694.
119. S. S. Stadtler, *Am. J. Pharm.* **76** (1907) 84 – 87; *Chem. Zentralbl.* **I** (1904) 1176.
120. W. Leithe: *Die Analyse der Luft und ihre Verunreinigungen,* Wissenschaftl. Verlags GmbH, Stuttgart 1974.
121. D. Henschler, *Luftanalyse, Analytische Methoden zur Prüfung gesundheitsschädlicher Arbeitsstoffe,* vol. 1, Verlag Chemie, Weinheim 1976.
122. Verein Deutscher Ingenieure: 1, *Messen gasförmiger Immissionen, Bestimmen der Formaldehydkonzentration nach dem Sulfit-Pararosanilin-Verfahren,* Richtlinie VDI 3484, Blatt 1, Düsseldorf 1979.
123. Supelco, Bulletin 794, Supelchem GmbH, 8399 Friesbach, Federal Republic of Germany.
124. NIOSH Manual of Analytical Methods, vol. 7, Analytical Method P & CAM 354 E, Cincinatti, Ohio, 1981.
125. Antechnika, Karlsruhe, Formaldehyd-Monitor TGM 555.
126. Du Pont, US 2 000 152, 1932 (J. F. Walker).
127. American Cyanamid, US 2 237 092, 1939 (R. C. Swain, P. Adams).
128. G. Altieri, IT 509 976, 1954.
129. Celanese, US 3 137 736, 1959 (R. H. Prinz, B. C. Kerr); US 3 532 756, 1968 (R. H. Prinz, B. C. Kerr).
130. Süddeutsche Kalkstickstoff-Werke, DE 1 205 072, 1962 (J. Seeholzer); DE 1 205 073, 1964 (P. Bornmann, H. Michaud). Koei Chemical Co., CA 809 691, 1963 (S. Matsuura, Y. Hikata).
131. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin, Heidelberg 1970, 1974, 1980, 1987, Merkblatt 95, 95 a.
132. Formaldehyd, Porträt einer Chemikalie, BASF, Ludwigshafen 1984.
133. WHO: *Environmental Health Criteria (EHC) for Formaldehyde. Interantional Program on Chemical Safety (IPS),* Hannover 1987.
134. D. Henschler: *Formaldehyd, MAK-Nachtrag 1987,* VCH Verlagsgesellschaft, Weinheim1987.
135. H. Staudinger: *Die hochmolekularen organischen Verbindungen,* Springer Verlag, Berlin 1932. W. Kern et al.: *Angewandte Chemie* **73** (1961) 177.
136. *Houben-Weyl,* **E 20,** 1390.
137. *Houben-Weyl,* **XIV/1 a,** 408 – 410, 413 – 421.
138. J. F. Walker, *Formaldehyde,* 3rd ed., Reinhold Publishing Corp., New York 1964, pp. 52 ff., 158 ff.
139. Du Pont, US 2 512 950, 1950 (T. E. Londergan).
140. O. Vogl, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 12** (1975) 109.
141. *Houben-Weyl,* **7/1,** 429.
142. Du Pont, US 2 449 469, 1948 (W. F. Gresham, R. E. Brooks).
143. G. Nordgren, *Acta Pathol. Microbiol. Scand.* **40** (1939) 21.
144. H. H. Nielsen, E. S. Ebers, *J. Chem. Phys.* **5** (1937) 824.
145. J. Löbering, *Ber. Dtsch. Chem. Ges. (B. Abhandlungen)* **69** (1936) 1846.
146. Consortium f. Elektrochem. Ind., DE 489 644, 1925 (M. Mugdan, J. Wimmer).
147. Du Pont, US 2 527 654, 1947 (C. Pyle, J. A. Lane); US 2 527 655, 1948 (C. Pyle, J. A. Lane), US 2 581 881, 1948 (C. Pyle, J. A. Lane).

148. Celanese, US 2 568 016 – 18, 1949 (A. F. McLean, W. E. Heinz).
149. Cities Service Oil Co., US 2 498 206, 1948 (B. W. Greenwald, R. K. Cohen).
150. Degussa, DE 1 260 143, 1968 (H.-J. Mann, H. Murek).
151. Société Chimique des Charbonages, FR 2 067 169, 1969 (Y. Moreaux).
152. Jos. Meissner KG, DE 2 037 759, 1972 (F. M. Deisenroth).
153. Celanese, GB 682 737, 1950.
154. Petric V. N. et al., SU 675 051, 1979 (V. N. Petric, N. V. Kudrina, A. E. Obraztsov, G. G. Vidorenkov).
155. Du Pont, US 2 529 269, 1948 (J. F. Walker); US 2 992 277,1959 (H. F. Porter).
156. Sumitomo, US 3 001 235, 1959 (D. Komiyama, T. Takaki, T. Ando, T. Nii).
157. Degussa, DE 1 795 551, 1963 (H. Junkermann, F. Löffler).
158. Heyden Newport Chem. Corp., US 2 915 560, 1957 (D. Steinhardt, D. X. Klein, R. H. Barth).
159. Societa Italiana Resine, US 3 772 392, 1972 (T. Paleologo, J. Ackermann).
160. Degussa, DE 884 947, 1951 (H. Leyerzapf).
161. Pan American Petroleum Corp., US 2 823 237, 1954 (J. F. McCants).
162. Du Pont, US 2 481 981, 1948 (R. L. Craven); US 2 519 550, 1948 (R. L. Craven).
163. Mitsubishi Gas Chemical, JP 73 21 082, 1973 (M. Katayama, S. Yahara, T. Endo).
164. *Beilstein,* **19 (3, 4)** 4710.
165. H. Staudinger, H. Luthy, *Helv. Chim. Acta* **8** (1925) 65.
166. Du Pont, US 2 304 080, 1940 (C. E. Frank).
167. J. Mahieux, *Hydrocarbon Process.* **48** (1969) no. 5, 163.
168. G. Rotta, DE 2 225 267, 1972.
169. Asahi Kasei Kogyo, DE 3 106 476, 1981 (K. Yoshida, T. Iwaisako, J. Masamoto, K. F. Hamanaka et al.).
170. BASF, DE 1 543 340, 1966 (H. Buchert, H. Sperber).
171. E. Bartholomé, W. Köhler, H. G. Schecker, G. Schulz, *Chem. Ing. Tech.* **43** (1971) 597.
172. ICI, GB 1 012 372, 1963 (W. R. Bamford).
173. BASF, DE 1 668 867, 1968 (H. Sperber, H. Fuchs).
174. Houillères du Bassin du Nord, FR 1 459 000, 1965 (E. Comber, H. Montanbric).
175. BASF, DE 1 668 867, 1968 (H. Sperber, H. Fuchs, H. Libowitzky).
176. Asahi, JP 95 215 961, 1995 (J. Masamoto, H. Morishita).
177. Péchiney-Saint-Gobain, FR 1 449 675, 1965 (M. P. Raoul).
178. Degussa, FR 1 377 169, 1963.
179. Polyplastics, US 3 732 252, 1971 (H. Komazawa, O. Matsumo).
180. Hoechst, DE 2 853 091, 1978 (K. F. Mück, G. Sextro, K. H. Burg); DE 2 912 767, 1979 (H. Bär, K. H. Burg, H. Mader, K. F. Mück, Cr. Sextro); DE 2 943 984, 1979 (H. Bär, H. Mader, K. F. Mück, P. Zorner).
181. Mitsubishi Gas Chem., DE 2 855 710, 1978 (A. O. S. Sugio et al.).
182. Asahi, JP 95 33 762, 1995 (J. Masamoto).
183. Hoechst, EP 596 381, 1994 (D. Arnold).
184. Perstorp, AT 252 913, 1964 (P. G. M. Flodin, P. Komfeldt, J. I. Gardshol).
185. G. Emig, F. Kern, St. Ruf, H.-J. Warnecke, *App. Catal. A.* **118** (1994) L17 – L20.
186. Hoechst, EP 604 884, 1994 (G. Emig et al.).
187. Hoechst, EP 691 338, 1996 (M. Hoffmockel, G. Sextro, G. Emig, F. Kern).
188. D. Fleischer, K. F. Mueck, G. Reuschel, *Kunststoffe* **82** (1992) no. 9, 763 – 766.
189. Hoechst, EP 484 786, 1992 (K.-F. Mueck, G. Reusch, D. Fleischer).
190. Solvay, DE 2 142 920, 1971 (A. Ryckaert). Péchiney-Saint-Gobain, DE 1 793 235, 1968 (Y. Correia).
191. Toyo Koatsu Industries, US 3 426 041, 1965 (Y. Miyake, S. Adachi, N. Yamanchi, T. Hayashi et al.).
192. Houillères du Bassin du Nord, FR 1 548 554, 1967 (E. Gombar, J. Mahieux).
193. S. Sugimoto et al., *Int. J. Appl, Radiat. Isot.* **34** (1983) no. 11, 559.
194. Miki Tetsuro et al., *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) no. 11, 3031.
195. K. H. Burg et al., *Makromol. Chem.* **111** (1968) 181.
196. Mitsui Toatsu Chem., JP 74 24 836, 1974 (K. Yamamoto, M, Naito, M. Hata).
197. *American Petroleum Institute Res. Proj 44,* no. 1264 (1951).
198. V. P. Belikov et al., *Izv. Akad. Nauk SSSR Ser. Khim* **8** (1967) 1862.
199. Coastal Interchemical Co, US 3 190 916, 1965 (N. B. Rainer).
200. Du Pont, GB 598984, 1948.
201. Hampshire Chemical, US 2 855 428 1958 (J. J. Sin-ger, M. Weisberg).
202. Degussa DE 3 242 748, 1984 (A. Kleemann, B. Lehman, K. Deller).
203. D. B. Lake, T. E. Londergan, *J. Org. Chem.* **19** (1954) 2004.
204. Mitsui Toatsu Chem., JP 61 72 761, 1986 (H. Inagaki et al.).

205. Mitsui Toatsu Chem., JP 61 83 164, 1986 (K. Takeuchi et al.).
206. Dow Chemical, US 2 890 238, 1959 (A. R. Sexton).
207. Monsanto, US 2 752 383, 1953 (S. F. Belt).
208. Mitsubishi Chemical, JP-Kokai 76 100 027, 1976 (Y. Ono).
209. Standard Oil Co., US 4 634 789, 1987 (R. G. Teller, J. F. Brazdil, L. C. Glaeser).
210. J. F. Brazdil et al., *J. Catal.* **100** (1986) no. 2, 516.
211. Standard Oil Co., US 4 515 732, 1985 (J. F. Brazdil, W. A. Marrit, M. D. Ward).
212. R. Gaudry; *Organic Synthesis, Coll.* vol. 3, J. Wiley & Sons, New York 1955, 436.
213. Rohm and Haas Co., US 3 057 903, 1962 (J. W. Nemec, C. H. McKeever).
214. Röhm und Haas, DE 811 952, 1949 (H. Beier); US 2 623 896, 1952 (H. Beier).
215. H. E. Christensen (ed.): *Registry of Toxic Effects of Chemical Substances,* U.S. Dept. of Health, Education and Welfare, Rockville 1976.
216. M. Albert, I. Hahnenstein, H. Hasse, G. Maurer: "Vapor-Liquid Equilibrium of Formaldehyde Mixtures: New Data and Model Revision," *AIChE J.* **42** (1996) no. 6, 1741 – 1752.
217. I. Hahnenstein, H. Hasse, Y.-Q. Liu, G. Maurer: "Thermodynamic Properties of Formaldehyde Containing Mixtures for Separation Process Design," *AIChE Symp. Ser.* **90** (1994b) no. 298, 141 – 157.
218. I. Hahnenstein, H. Hasse, C. G. Kreiter, G. Maurer: "^{1}H- and ^{13}C-NMR Spectroscopic Study of Chemical Equilibria in Solutions of Formaldehyde in Water, Deuterium Oxide, and Methanol," *Ind. Eng. Chem. Res.* **33** (1994b)no. 4, 1022 – 1029.
219. I. Hahnenstein, M. Albert, H. Hasse, C. G. Kreiter, G. Maurer: "NMR-Spectroscopic and Densimetric Study of Reaction Kinetics of Formaldehyde Polymer Formation in Water, Deuterium Oxide, and Methanol," *Ind. Eng. Chem. Res.* **34** (1995)no. 2, 440 – 450.
220. H. Hasse: "Dampf-Flüssigkeits-Gleichgewichte, Enthalpien und Reaktionskinetik in formaldehydhaltigen Mischungen," PhD Thesis, Universität Kaiserslautern, 1990.
221. H. Hasse, G. Maurer: "Kinetics of the Poly(oxymethylene) Glycol Formation in Aqueous Formaldehyde Solutions," *Ind. Eng. Chem. Res.* **30** (1991a)no. 9, 2195 – 2200.
222. H. Hasse, G. Maurer: "Vapor-Liquid Equilibrium of Formaldehyde-Containing Mixtures at Temperatures below 320 K," *Fluid Phase Equilib* **64** (1991b) 185 – 199.
223. H. Hasse, G. Maurer: "Heat of Dilution in Aqueous and Methanolic Formaldehyde Solutions," *Ber. Bunsenges. Phys. Chem.* **96** (1992)no. 1, 83 – 96.
224. H. Hasse, I. Hahnenstein, G. Maurer: "Revised Vapor – Liquid Equilibrium Model for Multicomponent Formaldehyde Mixtures," *AIChE J.* **36** (1990)no. 12, 1807 – 1814.
225. Y.-Q. Liu, I. Hahnenstein, G. Maurer: "Enthalpy Change on Vaporization of Aqueous and Methanolic Formaldehyde Solutions," *AIChE J.* **38** (1992)no. 11, 1693 – 1702.
226. G. Maurer: "Vapor-Liquid Equilibrium of Formaldehyde- and Water-Containing Multicomponent Mixtures," *AIChE J.* **32** (1986)no. 6, 932 – 948.
227. H. Schubert, U. Tegtmeyer, R. Schlögl, *Catalysis Letters* **28** (1994) 383 – 395.
228. H. Schubert et al., *Catalysis Letters* **33** (1995) 305 – 319.
229. G. J. Millar, J. B. Metson, G. A. Bowmaker, R. P. Cooney: *J. Chem. Soc. Faraday Trans.* **91** (1995) no. 22, 4149 – 4159.
230. W. L. Holstein, C. J. Machiels, *J.Catal.* **162** (1996) 118.
231. *Chemical Week* (1996) Feb. 28.
232. J. Eckenberger, *Kunststoffe* **86**(1996)no. 10, 1514.
233. WHO International Program on Chemical Safety (IPCS), Environmental Health Criteria Document 89, Formaldehyde, 1989.
234. WHO international Agency for Research on Cancer (IARC), IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 62, Wood Dust and Formaldehyde, 1995.
235. European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), Technical Report no.65, Formaldehyde and Human Cancer Risk, 1995.
236. J. Hilton et al.: "Experimental assessment of the sensitizing properties of formaldehyde," *Food Chem. Toxicol.* **34** (1996) 571 – 578.
237. T. M. Monticello et al: "Correlation of regional and nonlinear formaldehyde induced nasal cancer with proliferating population of cells," *Cancer Res.* **56** (1996) 1012 – 1022.
238. NIOSH Registry of Toxic Effects of Chemicals, July 1996.
239. Bundesgesundheitsblatt 9, 1992, p. 482 – 483; WHO Air Quality Guidelines (in press).

Formamides

HANSJÖRG BIPP, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 4)

HEINZ KIECZKA, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 5)

1. Introduction 1
2. Formamide 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.3. Production 3
2.4. Environmental Protection 4
2.5. Quality Specifications and Analysis . 4
2.6. Storage and Transportation 4
2.7. Uses 5
2.8. Economic Aspects 5
3. *N,N*-Dimethylformamide 5
3.1. Physical Properties 5
3.2. Chemical Properties 6
3.3. Production 6
3.4. Environmental Protection 7
3.5. Quality Specifications and Analysis . 8
3.6. Storage and Transportation 8
3.7. Uses 9
3.8. Economic Aspects 9
4. Other Formamides 9
5. Toxicology and Occupational Health . 10
6. References 12

1. Introduction

Formamides are derivatives of the smallest of the aliphatic carboxylic acids, formic acid, and are thus structurally and chemically rather straightforward substances. Two of the simplest members of the family–formamide itself [*75-12-7*], $HCONH_2$, and *N,N*-dimethylformamide [*68-12-2*], $HCON(CH_3)_2$–have considerable industrial importance because of their chemical bifunctionality and high polarity. Formamide is especially valued for its chemical reactivity; *N,N*-dimethylformamide is used widely for its solvent properties.

Formamide (methanamide) was first prepared by A. W. HOFFMANN in 1863 by reaction of ethyl formate with ammonia. The compound became commercially important only after the development of high-pressure syntheses based on carbon monoxide and ammonia, to yield formamide directly or via the intermediate methyl formate. *N,N*-Dimethylformamide (formyldimethylamine) was reported in 1893 by VERLEY, who obtained it upon distillation of a mixture of calcium formate and dimethylamine hydrochloride. Referred to commonly as "DMF," this compound is a key solvent in numerous operations, especially the spinning of synthetic fibers (a development of the 1950s).

No other member of the formamide family has achieved the industrial importance of these two substances.

2. Formamide

2.1. Physical Properties

Formamide [*75-12-7*], CH_3NO, M_r 45.04, is a hygroscopic, clear, colorless, odorless, and pH-neutral liquid with a consistency like that of glycerol. It melts at 2.55 °C and boils (decomp.) at 210 °C (101.3 kPa). Below 80 °C, the substance exhibits a greater degree of association than water, although this association diminishes rapidly with increasing temperature. The most important physical properties of formamide may be summarized as follows:

mp	2.55 °C
bp (at 101.3 kPa)	210 °C (decomp.)
Density ϱ (at 20 °C)	1.1334 g/cm^3
(at 50 °C)	1.1078 g/cm^3
Refractive index n_D^{15}	1.4491
n_D^{20}	1.4468
Dynamic viscosity η (at 15 °C)	4.320 mPas
(at 30 °C)	2.926 mPas
Surface tension σ (at 20 °C)	0.0582 N/m
Coefficient of expansion α	0.000742 cm^3/K
Specific heat $c^{19}{}_p$	2300 J kg^{-1}K^{-1}
Thermal conductivity λ (at 40 °C)	0.352 W m^{-1}K^{-1}
Heat of vaporization	1673 kJ/kg
Heat of combustion	12 530 kJ/kg
Heat of formation (at 18 °C)	−3529 kJ/kg
Electrical conductivity (at 25 °C)	$2\times10^{-7}\,\Omega^{-1}\,cm^{-1}$
Dielectric constant (at 25 °C) (wavelength 1 – 5 m)	109 ± 1.5
Dipole moment (at 30 °C)	3.37 D
Flash point (Abel Martens)	175 °C
Ignition temperature (DIN 51 794)	>500 °C
Explosive limits in air	2.7 – 19 vol %

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_001

Vapor pressure of formamide at various temperatures

t, °C	70.5	122.5	157.5	193.5	210
p, kPa	0.13	2.67	13.33	53.32	101.3 (decomp.)

Because of its high polarity and dielectric constant, formamide is an effective nonaqueous solvent for many inorganic salts, including chlorides, iodides, nitrates, phosphates, and carbonates. Formamide itself is very soluble in polar solvents such as water, lower aliphatic alcohols, carboxylic acids, esters, glycols, acetone, phenol, chloroform, and ether. Many high molecular mass polymers, natural products, and dyes also exhibit solubility or swelling in formamide. Examples include peptones and proteins (e.g., keratin), saccharides (e.g., cellulose and starch), and poly(vinyl alcohol). By contrast, hydrocarbons, fats, and oils have little solubility in formamide.

2.2. Chemical Properties

Pure formamide is a slightly hygroscopic liquid that is stable to the effects of light and atmospheric oxygen below ca. 100 °C. Above this limit, and especially above 160 °C, significant decomposition occurs. The decomposition rate reaches 0.5 % per minute at the atmospheric pressure boiling point. Principal decomposition products are carbon monoxide and ammonia, along with smaller amounts of hydrogen cyanide and water. Further transformation products are also detectable, including polymeric hydrogen cyanide derivatives. Above 500 °C, and with the aid of catalysts such as aluminum oxides or aluminum silicates, decomposition can be so accelerated that hydrogen cyanide yields may exceed 90 %. Formamide is stabilized by the presence of traces (e.g., 0.001 wt %) of purine bases [7].

Formamide is quite stable to water at room temperature. However, at elevated temperatures it undergoes hydrolysis to formic acid and ammonia, a combination that appears as ammonium formate. Both acid and alkali accelerate this hydrolysis, and the reaction of sulfuric acid with formamide at high temperature produces a nearly quantitative yield of formic acid. Until recently, this process served BASF as the basis for industrial-scale preparation of formic acid; today, however, direct synthesis of formic acid by high-pressure hydrolysis of the formamide precursor methyl formate is more economical, because the formation of ammonium sulfate as a side product is avoided (→ Formic Acid, Chap. 4.). Alkaline hydrolysis of formamide leads to alkali formates and ammonia, a reaction that can be used for quantitative analysis of formamide.

Formamide has attracted commercial interest not only because of its favorable solvent characteristics but also because of its bifunctional nature, which makes it a valuable component in many synthetic sequences. It is particularly useful as a formylating reagent, e.g., in the N-formylation of primary and secondary amines. Formamide can also serve as the starting point for substituted formamides (e.g., formanilide [*103-70-8*]), which are obtained by transamidation. It is used in the preparation of nitrogen-containing heterocycles such as theophylline [*58-55-9*], caffeine [*58-08-2*], pyrimidine [*289-95-2*], and imidazole [*288-32-4*]. Formate esters are produced by the reaction of formamide with the corresponding alcohols in the presence of an acid catalyst. Formamide reacts with acyl chlorides to give triacylamines [8]. Reaction with sodium methoxide produces sodium diformylamide [*18197-26-7*] [9].

2.3. Production

The production of formamide when carbon monoxide and ammonia are heated under pressure and subjected either to dark electrical discharge or to the catalytic effects of ceramic fragments has been known since the beginning of this century. However, the work of R. Wietzel in 1924 led to the first economical, continuous process for the manufacture of formamide. His results showed that alkyl formates, especially methyl formate, could be produced in satisfactory yield at high pressure (25 MPa) and low temperature (30 – 140 °C) with the aid of alkali alkoxide catalysts, thus establishing what is now referred to as the two-step formamide synthesis [10]. Wietzel also introduced a direct synthesis of formamide from carbon monoxide and ammonia in the presence of alcohols as a solvent, again promoted by alkali alkoxides [11].

Modern facilities for the preparation of formamide utilize either this *direct synthesis*,

$$CO + NH_3 \xrightarrow{NaOCH_3} HCONH_2$$

or the indirect *two-step process*:

$$CO + CH_3OH \xrightarrow{NaOCH_3} HCOOCH_3$$

$$HCOOCH_3 + NH_3 \longrightarrow HCONH_2 + CH_3OH$$

Direct Synthesis. In the direct synthesis of formamide, carbon monoxide or a gas mixture containing carbon monoxide is mixed with ammonia (either in stoichiometric amounts or in excess) and subjected to the catalytic action of alkoxides dissolved in alcohol (usually methanol). Other potential catalysts have been examined, including organometallic compounds from groups 4 – 10 of the periodic table (i.e., titanium to nickel), but these have not proved satisfactory so far [12]. The amount of catalyst required depends on the extent to which catalyst poisons are present, i.e., substances such as water that lead to undesirable side reactions. Poisons of this type render a portion of the catalyst inactive, causing it to precipitate from the reaction mixture in the form of sodium formate. Requisite pressures are 0.8 – 1.7 MPa at 75 – 80 °C [13]. Ammonia may be introduced in either liquid or gaseous form. Pure carbon monoxide is suitable for the process, but so is a gas mixture containing carbon monoxide, such as the coke-oven gas or blast-furnace gas obtained from a steel mill [14]. Water, hydrogen sulfide, and carbon monoxide are potential catalyst poisons and must be removed from the starting material, but hydrogen and inert gases pose no problem.

The direct synthesis of formamide actually consists of a series of steps, including synthesis, methanol recovery, catalyst recovery, salt removal, and final product purification. A separate purification step (flushing with water or separation of formamide by distillation) is necessary because impurities collect during salt removal [15], [16]. Because starting materials and product are both corrosive and stringent purity standards exist with respect to iron contamination, exposure to iron must be avoided. According to that, stainless steel is used as material of construction.

Two-Step Process. The two-step approach to formamide synthesis (as practiced, for example, by BASF) begins with ordinary commercial (96 %) methyl formate and ammonia. No other starting materials are involved, so this process avoids the problem posed in direct synthesis by residual salts and other impurities. Thus, no overhead distillation of formamide is required. Apart from excess starting materials, the only byproducts are traces of entrained water and inert gas, both of which can be removed easily by distillation to yield formamide of >99.5 % purity. The process divides itself naturally into three phases: the reaction itself, the recovery of methanol and other low-boiling materials, and vacuum evaporation of trace residues of volatile substances [17].

The BASF Process. The BASF process utilizes a continuous reactor equipped with a circulating pump and an external heat exchanger. Equimolar amounts of liquid methyl formate and gaseous ammonia [18] are introduced into the reactor [19], and the excess heat that results is removed by circulating the mixture over the external heat exchanger. Temperature is maintained in the range of 40 – 100 °C at 0.1 – 0.3 MPa. The gaseous phase is washed to remove excess ammonia and methyl formate. The inert gases methane, nitrogen, and a small amount of carbon monoxide (the latter arising as a decomposition product) are subsequently combusted together with off-gases from the evaporation step that also contain carbon monoxide. The reaction mixture, which consists of formamide, methanol, methyl formate, and ammonia then passes through a methanol recovery unit. This is a distillation system in which volatiles are isolated and separated into two fractions: methanol, and the low-boiling substances ammonia and methyl formate. Methanol accumulates at the bottom of the separator in sufficient purity to permit its use as a starting material in the synthesis of methyl formate. The low-boiling fraction contains virtually all the excess ammonia and methyl formate, together with a small amount of methanol, and this is recycled to the reactor. The residual formamide contains trace amounts of methanol which are removed under vacuum, leaving a product of sufficient purity to meet all standards applicable to both end-product and intermediate uses.

2.4. Environmental Protection

Flammability poses few problems, since formamide has very high ignition and flash points (>500 and 175 °C, respectively), as well as a low vapor pressure (see data in Section 2.1). Thus, the risk of combustion is quite low. If the substance should happen to ignite, the resulting fire may be quenched with water, extinguishing foam, carbon dioxide, or dry chemical extinguishers. Formamide should not be allowed to enter surface or groundwater sources but it can be readily biodegraded in appropriate wastewater treatment systems.

Serious effort should be made to prevent uncontrolled release of formamide from production and storage areas. The preferred method for transferring the liquid within an enclosed area involves the use of sealed pumps, although these should not be the stuffing-box type. Ideally, production should be carried out in the open, with exhaust gases washed or passed through a combustion unit prior to their release into the environment. If production or processing must be conducted within an enclosed facility, exceptionally good ventilation is essential because inhalation of formamide may be hazardous.

2.5. Quality Specifications and Analysis

Market demands require that formamide be available in a purity exceeding 99 %. The maximum permissible color number is 10 APHA, which implies a very low content of polymeric hydrocyanic acids and iron compounds. The material supplied by BASF is certified as meeting the following specifications.

Formamide content	min. 99.5 %	(by difference)
Methanol	max. 0.1 %	(photometric)
Ammonium formate	max. 0.1 %	(titrimetric)
Water content	max. 0.05 %	(DIN 51 777)
Iron	max. 1 ppm	(photometric)
Hazen color number	max. 10	(DIN 53 409)
pH	6.5	
Freezing point	≥2.1 °C	(DIN)

Methanol is determined photometrically after oxidation to formaldehyde, which in turn reacts with chromotropic acid to a purple dye. Ammonium formate content is measured by releasing the ammonia, trapping it with formaldehyde, and titrating the residual formic acid. The Karl Fischer method is employed for measuring water content. Iron is determined photometrically as the bipyridyl complex. The most accurate estimate of total formamide content results from subtracting the combined impurities from 100 %. Estimation by chromatography is less precise because of the compound's thermal instability, a consideration which also limits the usefulness of Kjeldahl determinations. Purity determination can also be based on melting point, which is related directly to the content of methanol and water.

2.6. Storage and Transportation

Unalloyed steel, copper, brass, and lead are all subject to corrosion by formamide, especially in the presence of water. Moreover, most protective paints and plastic coatings are attacked by the substance. In applications where iron contamination is undesirable aluminum and stainless steel containers are used instead of iron tanks. Formamide is shipped in containers made of stainless steel, in polyethylene drums, or in iron drums coated with polyethylene. Typical containers are designed to hold either 60 or 220 kg.

The physical and chemical properties of formamide are such that it is exempt from classification as a hazard with respect to highway, rail, inland waterway, or ocean transport.

When kept in closed containers, formamide is stable almost indefinitely. It is neither explosive nor subject to spontaneous combustion, although it is flammable. Those involved in the storage or transport of formamide should be familiar with the safety considerations discussed in Section 2.4, and Chapter 5.

2.7. Uses

Until relatively recently, the major application of formamide involved its reaction with sulfuric acid to produce formic acid and ammonium sulfate. However, alternative syntheses of formic acid based on methyl formate and water have now largely displaced the formamide method (→ Formic Acid, Chap. 4.). Currently, the principal application relates to the so-called formamide vacuum synthesis of hydrogen cyanide

[*74-90-8*] [20], in which formamide is pyrolyzed to hydrogen cyanide and water.

Formamide is also used as a synthetic intermediate in the preparation of various pharmaceuticals and fungicides. Examples include caffeine [*58-08-2*], theophylline [*58-55-9*], and theobromine [*83-67-0*] [21–24]. It serves as a catalyst in certain synthetic reactions, including carbonylation [25].

Formamide is employed widely as a *solvent*. It is used to dissolve lacquers [26] and plays a role in the spinning of acrylonitrile mixed polymerizates [27]. Together with nitromethane [*75-52-5*], it is used to dissolve polyacrylonitrile [*25014-41-9*] [28], and it is also employed in the polymerization of unsaturated amines to produce ion-exchange resins [29]. Formamide is a good solvent for wood stains, as well as for the ink used in felt- and fiber-tip pens. It was formerly employed in conjunction with silicates as a floor sealant, and also in de-icing airport runways, but these uses have been abandoned because of its teratogenic properties.

2.8. Economic Aspects

Only four major producers of formamide can be found worldwide. The leading manufacturer is BASF, with an annual capacity of 100 000 t in 1998. Most of the industrial output is utilized directly by the manufacturers. Despite the potential diversity of applications for formamide, market demand is relatively low, which explains why the number of production sites is so limited.

3. *N,N*-Dimethylformamide

3.1. Physical Properties

N,N-Dimethylformamide [*68-12-2*] (DMF), C_3H_7NO, M_r 73.09, is a colorless, high-boiling, nonviscous, highly polar, distillable, hygroscopic liquid with a weak, characteristically amine-like odor. Its most important physical characteristics can be summarized as follows:

mp	−61 °C
bp (at 101.3 kPa)	152.5 – 153.5 °C
Density ϱ (at 20 °C)	0.9500 g/cm^3
(at 50 °C)	0.9215 g/cm^3
Refractive index n_D^{20}	1.4310
n_D^{25}	1.4282
Dynamic viscosity η (at 15 °C)	0.87 mPa · s
(at 50 °C)	0.63 mPa · s
Surface tension σ (at 20 °C)	0.0355 N/m
Specific heat $c^{20}{}_p$	2062 J kg^{-1}K^{-1}
Thermal conductivity λ (at 40 °C)	0.180 Wm^{-1}K^{-1}
Heat of vaporization	573.5 kJ/kg
Heat of combustion	26 197 kJ/kg
Heat of formation (at 18 °C)	−3342 kJ/kg
Electrical conductivity (at 25 °C)	6×10^{-8} Ω^{-1} cm^{-1}
Dielectric constant (at 25 °C) (wavelength 1 – 5 m)	36
Dipole moment (at 20 °C)	3.82 D
Flash point (DIN 51 755)	58 °C
Ignition temperature (DIN 51 794)	410 °C
Explosive limits in air	2.2 – 16 vol %

Vapor pressure of DMF

t, °C	−50	−10	0	10	40	80	120	153
p, kPa	0.0003	0.033	0.081	0.182	1.34	9.67	42.0	101.3

Dimethylformamide is miscible in all proportions with water, alcohols, ethers, ketones, esters, and carbon disulfide, as well as with chlorinated and aromatic hydrocarbons. By contrast, it has only limited solubility or complete insolubilty in aliphatic hydrocarbons. Many high molecular mass natural products and synthetics are quite soluble in DMF. Examples include poly(vinyl alcohol) [*9002-89-5*], poly(vinyl chloride) [*9002-86-2*], poly(vinylidene chloride) [*9002-85-1*], poly(vinyl acetate) [*9003-20-7*] poly(vinyl ether), polyacrylate, polyacrylonitrile [*25014-41-9*], polystyrene [*9003-53-6*], polyester resins, alkyd resins, urea–formaldehyde resins, melamine – formaldehyde resins, phenol – formaldehyde resins, epoxy resins, and natural resins. It is also a good solvent for a number of drying oils, plasticizers, and cellulose derivatives, as well as for both natural and chlorinated rubber. The solubility of polyurethanes in DMF is a function of the conditions under which they are prepared. Most halogenated hydrocarbons are also soluble in DMF.

Various inorganic salts are highly soluble in DMF [30–38]. For example, the following solubilities (in 100 g DMF) are reported at 25 °C: 11.43 g of $FeCl_3$ [*7705-08-0*], 13.71 g of

Table 1. Solubility of organic gases in dimethylformamide (partial pressure 101.3 kPa)

Gas	CAS registry no.	Solubility (in 100 g of DMF), g 20 °C	100 °C
Vinylacetylene	[*689-97-4*]	112	18
Acetylene	[*74-86-2*]	5.1	0.9
1,3-Butadiene	[*106-99-0*]	11.2	2.5
Isobutene	[*115-11-7*]	7.3	1.5
Butane	[*106-97-8*]	4.7	1.3
Propene	[*115-07-1*]	0.78	0.46
Propane	[*74-98-6*]	0.85	0.33
Ethylene	[*74-85-1*]	0.29	0.14

KI [*7681-11-0*], 15.97 g of KSCN [*333-20-0*], 27.53 g of LiCl [*7447-41-8*], and 39.6 g of $NaClO_4$ [*7601-89-0*]. Several organic and inorganic gases are soluble in DMF, as indicated in Tables 1 and 2. Other gases such as ammonia, carbon monoxide, carbon dioxide, oxygen, and hydrogen dissolve in DMF to only a limited extent at atmospheric pressure.

Table 2. Solubility of inorganic gases in dimethylformamide (partial pressure 101.3 kPa)

Gas	Solubility (in 100 g of DMF), g at 25 °C
Hydrogen bromide	47.55
Hydrogen chloride	39.37
Hydrogen sulfide	1.32
Sulfur dioxide	179.00

3.2. Chemical Properties

N,*N*-Dimethylformamide is stable to air and light and, unlike its parent compound formamide, can be distilled without decomposition at atmospheric pressure. Nevertheless, decomposition is observed at temperatures above the boiling point, the products being carbon monoxide and *N,N*-dimethylamine. The compound also shows less tendency than formamide to hydrolyze. It is hygroscopic, however, and water absorbed from the air will produce traces of hydrolysis products; these produce the faint amine odor that frequently accompanies commercial DMF. The resistance of DMF to hydrolysis is demonstrated by the fact that a refluxing 5 % aqueous solution was found to have decomposed only 0.17 % after 120 h.

Pure DMF can be handled quite safely. Nevertheless, a few reports exist of reactions in which DMF displays a reactivity bordering on the explosive. Thus, with certain halogenated hydrocarbons, e.g., carbon tetrachloride or hexachlorobenzene, at elevated temperature and in the presence of ionic iron, explosive decomposition has occurred. Explosion during the reaction of DMF with thionyl chloride has also been described [39]. Furthermore elemental bromine and triethylaluminum [*97-93-8*] react violently with DMF. Vigorous oxidation reactions have been described in DMF systems containing chromic acid, as well as when nitrates are dissolved in DMF. Finally, DMF vapors form a flammable mixture with air.

3.3. Production

The first experiments directed toward the production of *N*-alkylformamides were conducted in the 19th century. Thus, A. Behal reported in 1899 that *N*-alkylformamides could be prepared by reacting mixed anhydrides of formic and other aliphatic acids with *N*- alkylamines. Since then, *N,N*-dimethylformamide has been synthesized by the reaction of hydrocyanic acid with methanol [40] or methylamines in the presence of water [41]. Other approaches include the reaction of ammonia or formamide with hydrogen and carbon dioxide in the presence of catalysts [42], as well as the reaction of *N,N*-dimethylamine with hydrogen and carbon dioxide [43]. Processes have also been described starting with carbon monoxide and *N,N*-dimethylamine, either alone [44] or together with trimethylamine [*75-50-3*] [45].

As with formamide, two methods dominate present-day commercial preparation of DMF: *direct synthesis* and a *two-step process.*

Direct Synthesis. The direct or one-step synthesis [46] of DMF begins with either pure carbon monoxide or a gas stream containing carbon monoxide. This is reacted in a continuous process with *N,N*-dimethylamine, by using a solution of alkali alkoxide (usually sodium methoxide) in methanol as catalyst. Methyl formate is presumably formed as an intermediate. The reaction mixture passes over an external heat exchanger to remove the excess heat generated and to ensure thorough mixing of the components [47].

The reaction is conducted between 0.5 and 11 MPa at 50 – 200 °C. The reaction mixture exits the reactor through a decompression chamber. In addition to *N,N*-dimethylformamide, the crude product contains methanol, a certain amount of unreacted *N,N*-dimethylamine, dissolved carbon monoxide, and residual catalyst. The addition of acid or water deactivates any catalyst present, resulting in the formation of sodium formate [*141-53-7*]. Dissolved carbon monoxide, together with inert gases, escapes from the mixture during decompression, and the off-gases are removed by combustion. Preliminary distillation is followed by a second distillation in a separate column; here, dimethylformamide is separated from methanol which contains traces of *N,N*-dimethylamine. Further distillation results in a product of 99.9 % purity.

Two-Step Process. The two-step process [48] for the synthesis of *N,N*-dimethylformamide differs from direct synthesis because methyl formate is prepared separately and introduced in the form of ca. 96 % pure (commercial-grade) material. Equimolar amounts of methyl formate and *N,N*-dimethylamine are subjected to a continuous reaction at 60 – 100 °C and 0.1 – 0.3 MPa. The resulting product is a mixture of *N,N*-dimethylformamide and methanol. The purification process involves distillation and is analogous to that described for direct synthesis. However, no separation of salts is required because no catalysts are involved in the process. According to the corrosive properties of both starting materials and products, stainless steel has to be used as material of construction for production facilities.

3.4. Environmental Protection

The physical properties of *N,N*-dimethylformamide (i.e., its autoignition temperature and flash point) are such that the compound is exempt from classification under the ordinance dealing with combustible liquids in the Federal Republic of Germany (VbF, Verordnung für brennbare Flüssigkeiten). If fire involving DMF does occur, it can be extinguished with water, dry chemical agents, or carbon dioxide.

N,N-Dimethylformamide is regarded as a hazardous liquid with respect to water pollution, and direct discharge into streams and lakes must be avoided. Nevertheless, it is easily biodegradable in appropriate wastewater treatment plants.

Regulations governing the use and disposal of *N,N*-dimethylformamide are essentially the same as those for formamide. The advisory covering atmospheric emissions in the Federal Republic of Germany (TA Luft, Verwaltungsvorschrift zum Bundesimmissionsschutzgesetz, 28th August 1974) places DMF in Category II. This means that DMF emissions exceeding 3 kg/h must be diluted to such an extent that the maximum concentration in air does not exceed 150 mg/m^3. If this limit cannot be met, the corresponding off-gases must be treated with water or combusted. The content of DMF in air can be measured either by gas chromatography or with an infrared detector.

3.5. Quality Specifications and Analysis

Most applications of DMF require that it be available in high purity. Table 3 contains the specifications applicable to DMF marketed by BASF.

The acid and base content of a DMF sample is established easily by potentiometric titration. The method of choice for determining DMF in an aqueous solution depends upon the anticipated concentration: those above 10 % may be quantified by gas chromatography, while lower concentrations are best measured with a double-beam infrared photometer. If only traces of DMF are expected, as in wastewater, a combination of chemical and physical methods of analysis

should be employed. Thus, any DMF present can be hydrolyzed with acid, after which the solution is made alkaline and the resultant *N,N*-dimethylamine is distilled. The latter is then converted to the corresponding nitroso compound, which can be determined polarographically. In the absence of impurities, DMF content can be calculated from simple physical measurements such as refractive index or density.

Table 3. Sales specifications for *N,N*-dimethylformamide from BASF

Sales specifications	Limit	Test method
DMF content	min. 99.9 %	gas chromatography
Methanol content	max. 100 mg/kg	gas chromatography
Water content	max. 0.03 wt %	DIN 51 777
Boiling range	152.5 – 153.5 °C	DIN 53 171
Acid content calcd. as HCOOH	max. 20 mg/kg	titrimetry
Base content calcd. as $NH(CH_3)_2$	max. 20 mg/kg	titrimetry
Hazen color no.	max. 10	DIN 53 409

3.6. Storage and Transportation

If it is pure and anhydrous, at ordinary temperatures, *N,N*-dimethylformamide does not tend to corrode metals during storage and transportation. Even with standard steel (ST 37), the linear corrosion rate for DMF is less than 0.01 mm per year. Exceptions to this generalization include copper, tin and their alloys.

Because DMF is hygroscopic, it readily absorbs moisture from the air, which renders it more corrosive. To prevent iron contamination, DMF should be packaged under an atmosphere of dry nitrogen if iron containers are used. To avoid all risk of contamination, it should be stored in stainless steel containers, although aluminum containers are recommended for DMF used in synthetic fiber production. Suitable gasket materials for pipelines, industrial fittings, and other apparatus exposed to DMF include poly(tetrafluoroethylene) [*9002-84-0*], high molecular weight polyethylene [*9002-88-4*], polypropylene [*9003-07-0*], and high molecular weight poly(vinyl alcohol) [*9002-89-5*]. *N,N*-Dimethylformamide may be shipped in tank cars, containers, and tank trucks fabricated from stainless steel. Iron drums with a capacity of 200 kg are also utilized. However, DMF should not be allowed to come in contact with chlorinated hydrocarbons because of the potential for explosive reaction, especially with such compounds as carbon tetrachloride [*56-23-5*] or hexachlorobenzene [*118-74-1*] in the presence of dissolved iron.

N,N-Dimethylformamide is subject to the following international transport designations (UN No. 2265):

Rail transport	GGVE/RID	3,32 C
Highway transport	GGVS/ADR	3,32 C
Water transport	GGVSee/IMDG-Gde	3,3
Air carriers	IATA-DGR/ICAO-TI	

Designations applicable to its transport and handling in the Federal Republic of Germany are derived from the ordinance "Verordnung über gefährliche Stoffe (Gefahrstoffverordnung GefStoffV)" of 26th August 1986, in which DMF is among the substances classified as having minimal toxicity. The corresponding designations include

Hazard symbol	Xn
R	20/21-36
S	26-28-36
EEC Guideline	616-001-00X

3.7. Uses

The principal applications of *N,N*-dimethylformamide are as a solvent and as an extractant, particularly for salts and compounds with high molecular mass. This role is consistent with its interesting combination of physical and chemical properties: low molecular mass, high dielectric constant, electron-donor characteristics, and ability to form complexes. The use of DMF as a component in synthesis is of relatively minor significance, at least commercially.

The most important use of DMF as a solvent for polymers is in the preparation of polyacrylonitrile solutions for the manufacture of fibrous polyacrylonitrile [*25014-41-9*], (→ Fibers, 4. Synthetic Organic, Chap. 5.). Residual DMF can be recovered from the spun fibers with relatively little expenditure of energy. In addition DMF is used as a solvent in the preparation of high-quality surfaces based on polyurethanes, which are otherwise difficult to solubilize, and it is employed as a solvent for coating surfaces with polyamides.

The high boiling point and high solvating power of DMF lead to its further use as a solvent for wire enamel formulated with polyamides, polyesterimides, or polyurethanes. Conversely, its solubilizing ability also makes it an effective component of paint removers, and it is used as a selective solvent in a wide range of manufacturing processes. For example, DMF is useful in separating acetylene from ethylene; butadiene from C_4 cuts; and hydrogen chloride, hydrogen sulfide, or sulfur dioxide from CO_2-containing gases. *N,N*-Dimethylformamide or mixtures containing it are used by oil refiners to extract aromatic compounds from hydrocarbon mixtures.

Table 4. Manufacturers of *N,N*-dimethylformamide

Manufacturer	Country	Annual capacity, t
Air Products	USA	7 000
BASF AG	Fed. Rep. of Germany	60 000
BASF	Brazil	6 000
Celanese	Mexico	6 000
Chinook	Canada	10 000
Du Pont	USA	40 000
Ertisa	Spain	5 000
ICI	UK	15 000
Korea Fertilizer Co.	Republic of Korea	8 000
LCY	Taiwan	10 000
Mitsubishi Gas Chemicals	Japan	20 000
Nitto Chemicals	Japan	25 000
UCB	Belgium	16 000
VEB Leuna	German Democratic Republic	19 000

Solutions of salts in specially purified DMF serve as fillers for electrolytic capacitors, a role made possible by the high dielectric constant of DMF. This substance is also used either alone or with other solvents in the recrystallization of aromatics, heterocyclics, dyes, fluorescent whitening agents, and pigments. The pharmaceutical industry also utilizes its broad solvent and extractant properties

N,N-Dimethylformamide could potentially be employed in the synthesis of aldehydes, acetals, amides, amidines, esters, and heterocycles, but thus far its use has been limited [49–53].

3.8. Economic Aspects

N,N-Dimethylformamide is produced by a number of manufacturers worldwide; total capacity is estimated at 250 000 t per year. Much of the production is utilized directly by the manufacturers, the most important of whom are listed in Table 4.

The principal market for DMF is as a solvent for polyacrylonitrile, followed by its use in processing polyurethanes. The remaining applications are remarkably diverse but minor relative to the principal uses.

4. Other Formamides

While many other *N*-substituted formamides are known, only three have achieved a measure of commercial significance; *N*-methylformamide, *N*-formylmorpholine, and formanilide. These are prepared by a two-step process: either the corresponding nitrogen compound is reacted with methyl formate, or a transamidation is performed, typically involving formamide.

***N*-Methylformamide.** *N*-Methylformamide [*123-39-7*], $HCONHCH_3$, M_r 59.1, is a colorless, nearly odorless liquid that is infinitely soluble in water and relatively stable to hydrolysis at low temperature. Commercial material is available with a purity of at least 99.7 %. Its most important physical properties include the following: *mp* −3.2 °C; boiling range 199.3 – 200.5 °C; ϱ (at 20 °C) 1.003 g/cm^3; n_D^{20} 1.432 – 1.433; flash point 111 °C; ignition temperature 425 °C; explosive limits in air 3.8 – 17.8 vol %; vapor pressure (at 20 °C) 0.014 kPa (0.14 bar). The substance is used as an intermediate in the synthesis of the insecticide formothion. It also has limited use in oil refineries as a selective solvent for aromatic hydrocarbons.

***N*-Formylmorpholine.** *N*-Formylmorpholine [*4394-85-8*], $HCON(CH_2)_4O$, M_r 115.13, is a colorless liquid with a faint amine-like odor. It is very resistant to hydrolysis at low temperature and has sufficient thermal stability to be distilled at atmospheric pressure. Commercial material has a minimum purity of 99.5 %. The following are important physical properties of this substance: *mp* 20 – 21 °C; *bp* 244 °C; ϱ (at 20 °C) 1.152 g/cm^3; n_D^{20} 1.4870; flash point 125 °C; ignition temperature 370 °C; explosive

limits in air 1.2 – 8.5 vol %; vapor pressure (at 80 °C) 0.17 kPa.

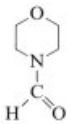

N-Formylmorpholine is used by petroleum refiners as a solvent for the extraction of aromatics and also as an anticorrosive agent in fuel oil.

Formanilide. Formanilide [*103-70-8*]; $HCONHC_6H_5$, M_r 121.14, is used as an intermediate in the pharmaceutical industry and as an antioxidant (anti-ager) in the rubber industry.

Formamides containing various aliphatic substituents on nitrogen have acquired limited significance as solvents and extractants. Thus, *N,N*-di-*n*-butylformamide [*761-65-9*] or *N*-methyl-*N*-2-heptylformamide [*68018-42-8*] can be used to extract organic acids from aqueous media [53].

5. Toxicology and Occupational Health

Formamide.

Acute Toxicity. Acute toxicity studies demonstrate that single doses of formamide, administered by a variety of routes, have low toxicity. In rats the acute *oral* LD_{50} is >5000 mg per kilogram body weight [55], [56] and the acute dermal LD_{50} is 13.5 g per kilogram body weight [56]. The acute inhalation hazard is low: no mortality was observed in rats exposed for 8 h at 20 °C to an atmosphere enriched in or saturated with formamide [55]. In the rabbit skin and eye the substance is not irritating to slightly irritating [55], [56]. Skin sensitization was not observed in guinea pigs [57].

Subacute, Subchronic Toxicity. Twenty oral doses of 300 µL formamide per kilogram body weight, given to rats via stomach tube, over four weeks were lethal in 50 % of the animals. The affected organs were the stomach, spleen thymus, adrenals, and testis. Treatment with 100 µL per kilogram body weight resulted in reduction of body weight gain and a reduced food intake whereas 30 µL per kilogram body weight did not produce any observable adverse effects [58]. Rats were subjected to doses of 30 – 3000 mg per kilogram body weight, applied dermally for three months (6 h per day, 5 d per week), semi-occlusive on the intact skin. Mortality occurred after treatment with 3000 mg per kilogram body weight. The affected organs were the liver, kidney, adrenals, spleen and testis. Hematological investigation revealed polycytemia, a change in the red blood picture, as the only effect observed. This effect also occurred at doses of 1000 and 300 mg/kg but not at 100 and 30 mg/kg body weight [59].

Mutagenicity. In several mutagenicity tests on bacterial systems (Ames test [60–62]), on mammals (dominant lethal test, mouse [64]), or in an in vitro short term test for carcinogenic activity (cell transformation test in rat embryo cultures [64], [65]) formamide did not demonstrate a mutagenic potential.

Teratogenicity. Formamide has been tested in a variety of studies for its teratogenic effects. In rats and mice the compound produced adverse effects on the offspring after oral and dermal administratation. The degree of fetal damage depends on the stage of pregnancy at which the substance was given. Doses that had adverse effects on the dams also had distinct effects on the offspring; at lower doses the effects were less pronounced [56], [66].

TLV Value. The TLV value (1988 – 1989) was established to be 10 ppm (15 mg/m^3).

Dimethylformamide.

Acute Toxicity. Dimethylformamide which has been tested in different animal species using various administration methods is slightly toxic. The affected organs, especially after repeated administration are the liver and kidney.

The acute *oral* LD_{50} in rats and mice is 2200 – 7550 mg per kilogram body weight [67–69]. The acute *dermal* LD_{50} in rats is between 800 and 17 000 mg per kilogram body weight and in rabbits between 1000 and 5000 mg per kilogram body weight. The acute inhalative LC_{50} in rats is given as >2500 ppm for a four or six hour exposure and 2475 ppm for a one-hour exposure [69]. In mice LC_{50}-values are 9.4 mg/L (3133 ppm, two-hour exposure) or 686 ppm (two-hour exposure) [56]. Rats survived a four-hour inhalation of an atmosphere

enriched or saturated with DMF at room temperature, but deaths occurred after 6 h [67]. In other studies cats, rabbits, guinea pigs, and rats survived an eight-hour exposure to a saturated atmosphere without any adverse effects. Rabbits, cats, and dogs survived a 6 h-inhalation of 1000 ppm [67].

Irritation. DMF slightly irritates the rabbit skin. The liquid is absorbed by the skin and can facilitate the absorption of other substances [56], [67]. The compound has a pronounced irritating effect on the eye of rabbits, the effects being reversible within 14 d [56], [67], [69]. Washing the eye with water immediately after application increased the irritating effect and produced opacity of the cornea [69]. Although DMF has not been shown to cause sensitization [67], a case of allergic contact dermatitis has been reported [70].

Subacute and Subchronic Toxicity. In subacute and subchronic animal studies (rats, mice, rabbits, cats, and dogs) in which repeated doses of DMF were given orally, dermally, or by inhalation, nephrotoxicity and hepatotoxicity were the prominent effects. The doses and concentrations used varied widely depending on the method and duration of administration. Changes in the red blood picture were observed in a few cases (for further details, see [56], [67]).

Mutagenicity. DMF has been investigated in several mutagenicity tests covering all different end points of mutagenic activity and has in all these tests been found to be non-mutagenic [56].

Carcinogenicity. Animal studies published so far have not provided any evidence of carcinogenicity [56]. Rats receiving 150 mg DMF per kilogram body weight in their drinking water for 500 d did not develop tumors. A weekly subacute treatment with doses of 200 or 400 mg per kilogram body weight (total dose 8 – 20 g per kilogram body weight) did not increase the incidence of tumor rate in the rat. A single intraperitoneal injection of 0.5 g per kilogram body weight to rats with or without simultaneous partial hepatectomy did not cause formation of tumors or "nodules" as early stages. In an insufficiently designed, inadequately described study, rats developed multiple tumors, liver necrosis, and ulceration of the intestinal mucosa after ten weekly intraperitoneal injections of 1 mL per rat. In comparison to all other mentioned studies the latter result seems to be inadequate.

Teratogenicity. DMF has been tested in different animal species for its adverse effects on the offspring using various administration methods. Fetotoxic and embryotoxic effects (fetal death and/or reduced fetal weight) were observed, especially at doses that produced toxic effects in the dams. In some studies DMF increased the incidence of malformations.

Oral doses that are toxic to the dams increased the incidence of malformations in rats and rabbits, whereas lower doses which were not maternally toxic are still embryotoxic [71], [72]. In mice oral administration of DMF leads to embryotoxicity and to an increase in the incidence of spontaneous malformations without detectable maternal toxicity [72]. Dermal application of DMF (200 mg kg^{-1} d^{-1}) in the rabbit caused fetal death; maternal toxicity was not discussed [73]. Other investigators found that dermal administration of a 200 mg kg^{-1} d^{-1} dose of DMF, applied semi-occlusively, did not have any effects in rats but a dose of 400 mg kg^{-1} d^{-1} increased the incidence of skeletal anomalies and showed slight maternal toxicity [72]. In other studies in rats, dermal application of DMF at doses which were slightly toxic to the dams had adverse effects on the offspring [72], [73]. In inhalation studies, embryotoxic effects were observed in rats at doses that produced maternal toxicity; in a few cases these effects already occurred in the absence of maternal toxicity [56].

Dermal administration of 400 mg kg^{-1} d^{-1} caused a higher incidence of anomalies in the offspring of rabbits, parallelled by reduced body weight in the dams. The next lower dose of 200 mg kg^{-1} d^{-1} did not cause any effects [74].

Inhalative exposure of rabbits (6 h/d, 7 – 19 day of pregnancy) to a DMF concentration of 150 and 450 ppm in the respiratory air produced dose-dependent maternal toxic, embryotoxic, and teratogenic effects. A concentration of 50 ppm did not cause any effects [75]. According to these investigations DMF seems to be teratogenic after inhalation only in higher concentrations which normally also lead to observable maternal toxicity. The threshold concentration for embryotoxic and teratogenic effects after inhalative exposure is between 50 – 150 ppm.

MAK Value. The MAK value is established to be 20 ppm (60 mg/m^3) refering to the danger of skin resorption.

6. References

General References

1. The Sigma Aldrich Library of Chemical Safety Data, Edition I, 1985.
2. Robert Kühn, Karl Birett: *Merkblätter gefährliche Arbeitsstoffe,* Verlag Moderne Industrie, München.
3. *Beilstein* **2,** 26 – 27; **2 (1),** 20 – 21; **2 (2),** 36 – 37, **2 (3),** 55 – 59; **2 (4),** 45 – 50; system no. 156.
4. BASF, *Technisches Merkblatt über Formamid,* 1983. (Technical Leaflet Formamide, 1983.) for *N,N-Dimethylformamide*
5. *Beilstein* **4,** 58; **4 (1),** 171 – 174; **4 (3),** 122; system no. 335.
6. BASF, *Technisches Merkblatt über N,N-Dimethylformamid,* 1984. (Technical Leaflet Dimethylformamide, 1986.)

Specific References

7. Du Pont, US 2 346 425, 1944 (E. C. Kirkpatrik)
8. Q. E. Thompson, *J. Am. Chem. Soc.* **73** (1951) 5841.
9. I. C. Gramain, R. Rémuson, *Synthesis* 1982, 264.
10. I. G. Farben, DE 495 935, 1930 (R. Wietzel).
11. I. G. Farben, DE 550 749, 1932 (R. Wietzel).
12. Ethyl Corporation, US 3 099 689, 1961 (H. J. Cragg).
13. Du Pont, DE 624 508, 1936.
14. O. Bankowski, M. Köthnig, K. Klemt, DD 12 113, 1953.
15. VEB Leuna-Werke, DD 13 619, 1954 (K. Smeykal, M. Köthnig).
16. VEB Leuna-Werke, DD 61 536, 1967 (G. Peinze).
17. BASF, DE 1 215 130, 1966 (E. Germann, H. Friz)
18. BASF, DE 2 623 173, 1980 (H. Hohenschutz et al.)
19. Buss AG, CH 370 057, 1963 (W. Steiner)
20. I. G. Farben, FR 906 114, 1944.
21. H. Bredereck, *Angew. Chem.* **56** (1943) 328 – 329.
22. H. Bredereck et al., *Chem. Ber.* **83** (1950) 201 – 211.
23. H. Bredereck, H. G. v. Schuh, DE 864 870, 1953.
24. H. Bredereck, et al., *Chem. Ber.* **86** (1953) 333 – 351.
25. Esso Research and Engineering Co., US 2 834 812, 1958 (L. Hughes, C. Township, J. Kirshenbaum).
26. Radio Corp. of America, US 2 737 465, 1956 (L. Pessel).
27. Allied Chemical and Dye Corp., US 2 776 945, 1957 (F. J. Rahl, H. H. Weinstock).
28. P. Halbig, CH 295 072, 1953.
29. State Inst. of the State Florida, US 2 687 382, 1954 (G. B. Butler, R. L. Bunch).
30. R. Ch. Paul et al., *Indian J. Chem.* **4** (1966) 382.
31. R. Ch. Paul et al., *Indian J. Chem.* **2** (1964) 97.
32. W. Gerrard et al., *J. Chem. Soc.* 2 (1960) 2144.
33. T. S. Piper et al., *J. Am. Chem. Soc.* **76** (1954) 4318.
34. E. L. Muetterties et al., *J. Am. Chem. Soc.* **75** (1953) 490.
35. J. Archambault et al., *Can. J. Chem.* **36** (1958) 1461.
36. R. Ch. Paul et al., *Indian J. Chem.* **3** (1965) 300.
37. A. J. Jungbauer, *Nature (London)* **202** (1964) 290.
38. M. Glavas et al., *J. Inorg. Nucl. Chem.* **31** (1969) 291.
39. S. J. Mukund, *Chem. Eng. News* **64** (1986) no. 14, 2.
40. Y. Fukuoka, N. Kominami, *Chem. Technol.* 1972, 670 – 674.
41. Asahi Chem. Ind. Co., JP 7 208 051, 1968.
42. Air Products and Chemicals Inc., US 4 558 157, 1985 (J. A. Marsella, G. P. Pez).
43. Shell Oil Co., US 3 530 182, 1970 (P. Haynes, J. F. Kohnle, L. H. Slaugh).
44. The Leonard Process Co., *Hydrocarbon Process.* **32** (1983) no. 11, 90.
45. Nitto Chemical Industry Co., JP 160 075, 1978.
46. Du Pont, US 2 204 371, 1937 (D. J. Loder).
47. UCB S. A., DE 2 710 725, 1987 (W. Couteau, J. B. Ramioulle).
48. I. G. Farben, DE 714 311, 1937 (R. Seidler, H. Klein).
49. R. S. Kittila: *Dimethylformamide Chemicals, Uses,* Du Pont de Nemours, Wilmington, Del., 1967.
50. H. Bredereck et al., *Angew. Chem.* **74** (1962) 354.
51. H. Bredereck et al., *Chem. Ber.* **96** (1963) 3265.
52. H. Bredereck et al., *Angew. Chem.* **75** (1963) 825.
53. H. Bredereck et al., *Chem. Ber.* **97** (1964) 61.
54. BASF, DE 2 545 658, 1985 (H. Hohenschutz, J. E. Schmidt, H. Kiefer).
55. BASF Aktiengesellschaft, unpublished results 1963.
56. G. L. Kennedy, Jr., CRC Crit. Rev. Toxicol. **17** (1986) no. 2, 129 – 182.

57. P. H. Chanh et al., *Therapie* **26** 409 – 424 (1971).
58. BASF Aktiengesellschaft, unpublished results 1975.
59. BASF Aktiengesellschaft, unpublished results 1982 –1984.
60. National Toxicology Program, Fiscal Year 1984,Annual Plan.
61. National Toxicology Program, *Technical Bulletin* no. 9, April1983.
62. K. Mortelmans et al., *Environ. Mutagen.* Suppl. 7, 8 (1986) 1 – 119.
63. BASF Aktiengesellschaft, unpublished results 1970.
64. A. E. Freeman et al., *J. Nat. Cancer Inst.* **51** (1973) no. 3, 799 – 808.
65. Ishidate, Kada: *Environmental Mutagens Data Book,* 1980.
66. BASF Aktiengesellschaft, unpublished results 1963 – 1969.
67. Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten der Kommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe der Deutschen Forschungsgemeinschaft, Stand 16.12.1974, Nachtrag 20.12.1977, Verlag Chemie, Weinheim.
68. J. Lundberg, et al., *Environ. Res.* **40** (1986) 411 – 420.
69. G. L. Kennedy, Jr., H. Sherman, *Drug Chem. Toxicol.* **9** (1986) 2, 147 – 170.
70. J. G. Camarasa, *Contact Dermatitis* **16** (1987) no. 4, 234.
71. J. Merkle et al., *Arzneim. Forsch.* **30** (1980) 1557 –1562.
72. BASF Aktiengesellschaft, unpublished results 1970 – 1976.
73. E. F. Stula et al., *Toxicol. Appl. Pharmacol.* **41** (1977) 35 – 55.
74. BASF Aktiengesellschaft, unpublished results 1984.
75. BASF Aktiengellschaft, unpublished results 1988.

Formazan Dyes → Metal-Complex Dyes

Formic Acid

Werner Reutemann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 2)

Heinz Kieczka, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 13)

1. Introduction 2
2. Physical Properties 2
3. Chemical Properties 4
4. Production 5
4.1. Methyl Formate Hydrolysis 6
4.1.1. Kemira – Leonard Process 8
4.1.2. BASF Process 9
4.1.3. USSR Process 11
4.1.4. Scientific Design – Bethlehem Steel Process 11
4.2. Other Processes 11
4.2.1. Oxidation of Hydrocarbons 11
4.2.2. Hydrolysis of Formamide 11
4.2.3. Production of Formic Acid from Formates 13
4.2.4. Direct Synthesis from Carbon Monoxide and Water 13
4.2.5. Use of Carbon Dioxide 14
4.3. Concentration of Formic Acid . . . 14
4.4. Construction Materials 15
5. Environmental Protection 15
6. Quality Specifications 15
7. Chemical Analysis 16
8. Storage and Transportation 16
9. Legal Aspects 16
10. Uses 17
11. Economic Aspects 17
12. Derivatives 18
12.1. Esters 18
12.1.1. Methyl Formate 18
12.1.2. Ethyl Formate 19
12.1.3. Isobutyl Formate 19
12.1.4. Miscellaneous Esters 19
12.2. Salts 19
12.2.1. Alkali-Metal and Alkaline-Earth Metal Formates 19
12.2.2. Ammonium Formate 19
12.2.3. Aluminum Formate 19
12.2.4. Nickel Formate 19
12.2.5. Copper Formate 20
13. Toxicology and Occupational Health 20
14. References 21

1. Introduction

Formic acid [*64-18-6*] HCOOH, M_r 46.03, is a colorless liquid with a pungent odor, which is completely miscible with water and many polar solvents but only partially miscible with hydrocarbons. Formic acid derived its name from the red ant, *Formica rufa*, in which it was discovered around 1670. Formic acid has been detected in the poison or defense systems of ants, bees, and other insects and also of cnidarians.

Formic acid is used primarily in dyeing, in the textile and leather industries; in rubber production; and as an intermediate in the chemical and pharmaceutical industries. The use of formic acid as an aid in the ensilage of green forage has increased sharply.

The worldwide production of formic acid is about 260 000 t/a in 1987 and 390 000 t/a in 1995 [1]. Formic acid is produced by hydrolysis of methyl formate or formamide or from its salts. In addition, formic acid is a byproduct of acetic acid production by liquid-phase oxidation of hydrocarbons.

2. Physical Properties

Formic acid, *mp* 8.3 °C [2], *bp* 100.8 °C (at 101.3 kPa) [3], is a colorless, clear, very corrosive liquid with a pungent odor. With a pK_a of 3.739, formic acid is the strongest unsubstituted fatty acid, about ten times stronger than acetic acid [4]. Pure formic acid is hygroscopic.

The temperature dependence of the *density* of formic acid is given in Table 1. The density of formic acid – water binary mixtures, as a function of formic acid concentration, is shown in Table 2.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_013

Table 1. Density of pure formic acid as a function of temperature [5]

t, °C	ϱ, g/cm^3	t, °C	ϱ, g/cm^3
0	1.244 *	40	1.195
10	1.232	50	1.182
15	1.226	60	1.169
20	1.220	70	1.156
25	1.214	80	1.143
30	1.207	90	1.130
		100	1.117

* Supercooled liquid.

Table 2. Density of aqueous formic acid solutions at 20 °C [6]

Formic acid content, wt %	ϱ, g/cm^3	Formic acid content, wt %	ϱ, g/cm^3
2	1.003	50	1.118
5	1.011	60	1.141
10	1.024	70	1.163
20	1.048	80	1.185
30	1.072	90	1.204
40	1.095	100	1.221

The *freezing point* diagram for the formic acid – water binary mixture exhibits a eutectic point (Table 3) at −48.5 °C and 41.0 mol % of formic acid [2]. Formic acid does not increase in volume when it solidifies and has a tendency to undergo supercooling.

Table 3. Freezing points of various formic acid – water mixtures [2]

Formic acid molar fraction, %	fp, °C	Formic acid molar fraction, %	fp, °C
100.0	8.3	41.0	−48.5
80.0	−5.6	39.8	−46.8
74.8	−8.2	37.0	−43.0
69.5	−13.0	31.7	−33.5
64.5	−17.3	28.0	−29.3
59.0	−23.2	23.1	−22.8
54.2	−28.3	20.0	−19.7
50.0	−35.3	12.0	−10.3
44.6	−42.0		
42.0	−46.7		

Table 4 shows the *vapor pressure* curve of pure formic acid. The vapor of formic acid deviates considerably in behavior from an ideal gas because the molecules dimerize partially in the vapor phase. At room temperature and normal pressure, 95 % of the formic acid vapor consists of dimerized formic acid [8]:

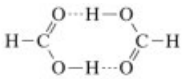

The enthalpy of the gas-phase dimerization is −63.8 kJ/mol [8]. The thermodynamic properties of monomeric and dimeric formic acid have been investigated by Waring [9].

Table 4. Vapor pressure of pure formic acid [7]

Liquid		Solid	
t, °C	p, kPa	t, °C	p, kPa
−5.23	1.083	−5.07	0.664
0.00	1.488	0.00	1.096
8.25	2.392	8.25	2.392
12.57	3.029		
20.00	4.473		
29.96	7.248		
39.89	11.357		
49.93	17.347		
59.98	25.693		
70.04	37.413		
79.93	52.747		
100.68	101.667		
110.62	135.680		

The ring-type dimeric structure exists both in the vapor phase and in solution. Liquid formic acid consists of long chains of molecules linked to each other by hydrogen bonds. Solid formic acid can also be isolated in two polymorphic forms (α and β) [10]:

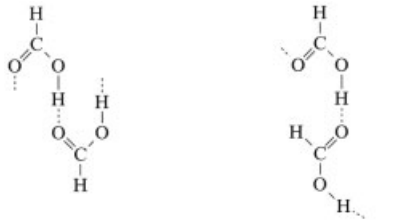

Portion of the chain in an α-type formic acid crystal

Portion of the chain in a β-type formic acid crystal

The thermodynamic properties of monomeric and dimeric formic acid have been investigated [10].

Vapor – liquid equilibrium data for formic acid – water mixtures and for mixtures of formic acid with organic compounds have been collected in [11]. Formic acid and water form a maximum-boiling azeotropic mixture whose boiling point is 107.6 °C at 101.3 kPa; it consists of 77.6 wt % formic acid and 22.4 wt % water. The composition and boiling point of formic acid – water azeotropic mixtures are shown as a function of pressure in Table 5. Formic acid can form azeotropic mixtures with many other substances [14].

Table 5. Azeotropic data for the formic acid – water system

p, bar	bp of azeotropic mixture, °C	Formic acid content, wt %	Reference
0.093	48.6	66.2	[12]
0.267	72.3	70.5	[12]
1.013	107.6	77.6	[12]
2.026	128.7	84	[13]
3.140	144	85	[13]

The variation of dynamic *viscosity* with temperature is shown in Table 6. The dynamic viscosity of formic acid – water mixtures decreases approximately linearly as the water content of formic acid increases [15]. The *thermal conductivity* of formic acid is markedly higher than that of comparable liquids, because of its pronounced polarity.

Table 6. Dynamic viscosity of pure formic acid as a function of temperature [16]

t, °C	η, mPa · s
10	2.262
20	1.804
30	1.465
40	1.224
50	1.025

The variation of specific *heat capacity* with temperature is shown in Table 7. The specific heat capacity of formic acid – water mixtures decreases approximately linearly as the concentration of formic acid increases [6].

Table 7. Specific heat capacity of formic acid as a function of temperature at constant pressure [17]

t, °C	c_p, J g^{-1}K^{-1}	t, °C	c_p, J g^{-1} K^{-1}
Solid		*Vapor*	
−150	0.921	25	1.058
−100	1.114	100	1.192
−75	1.193	200	1.348
−50	1.285	300	1.480
−25	1.411	400	1.589
0	1.800	600	1.757
		800	1.870
		1000	1.953
		1200	2.037
Liquid			
20	2.169		
50	2.202		
100	2.282		

The physical properties of formic acid are listed below:

Heat of fusion		276 J/g [7]
Heat of vaporization (at bp)		483 J/g [7]
Dielectric constant		
Liquid (at 20 °C)		57.9 [21]
Solid (at −10.1 °C)		11.7
Refractive index n_D^{20}		1.37140 [19]
Surface tension σ		
(at 20 °C)		37.67×10^{-3} N/m [20]
(at 40 °C)		35.48×10^{-3} N/m
(at 60 °C)		33.28×10^{-3} N/m
(at 80 °C)		31.09×10^{-3} N/m
Heat of formation ΔH_f^0		
Liquid	(at 25 °C)	−425.0 kJ/mol [22]
Vapor	(at 25 °C, monomer)	−378.57 kJ/mol [8]
	(at 25 °C, dimer)	−820.94 kJ/mol
Heat of combustion ΔH_c^0		
Liquid	(at 25 °C)	−254.8 kJ/mol [22]
Entropy S^0		
Liquid	(at 25 °C)	129.0 J K^{-1} mol^{-1} [10]
Vapor	(at 25 °C, monomer)	248.88 J K^{-1} mol^{-1} [8]
	(at 25 °C, dimer)	332.67 J K^{-1} mol^{-1} [8]
Heat of neutralization		56.9 kJ/mol [23]
Critical data p_{crit}		7.279 MPa [18]
	T_{crit}	581 K [18]
	ϱ_{crit}	0.392 g/cm^3 [18]
Thermal conductivity λ		
Liquid	(at 20 °C)	0.226 W m^{-1} K^{-1} [18], [24]
	(at 60 °C)	0.205 W m^{-1} K^{-1}
	(at 100 °C)	0.185 W m^{-1} K^{-1}
Vapor	(at 50 °C)	0.0136 W m^{-1} K^{-1}
	(at 100 °C)	0.0176 W m^{-1} K^{-1}
	(at 200 °C)	0.0267 W m^{-1} K^{-1}
Electrical conductivity (at 25 °C)		6.08×10^{-5} Ω^{-1} cm^{-1} [25]
Coefficient of cubic expansion α (at 30 °C)		0.001 [3]

In 1968, GALLANT reported the important physical properties of formic acid and other low molecular weight fatty acids as a function of temperature [18].

3. Chemical Properties

To some extent, formic acid reacts differently from its higher homologues because it is both a carboxylic acid and an aldehyde.

Aldehyde Reactions. In some reactions, formic acid behaves like an aldehyde. Thus if an aqueous solution of formic acid is heated with a silver nitrate solution, metallic silver is deposited. Mercury, gold or platinum ions are reduced in solution, as are organic compounds. For

example, formic acid reduces triphenylmethanol to triphenylmethane. The reduction of imines (Schiff bases) by formic acid has been known for a long time. Primary amines can be prepared from ketones, ammonia, and formic acid (Leuckart reaction). Like other aldehydes, formic acid has bactericidal properties.

Carboxylic Acid Reactions. Magnesium, zinc, and iron dissolve in formic acid with the evolution of hydrogen. Because of its strongly acidic nature, many alcohols can be esterified by formic acid, without the addition of mineral acids. Primary and secondary alcohols are esterified in pure formic acid 15 000 – 20 000 times more rapidly than in pure acetic acid [26]. The rate of esterification of primary, secondary, and tertiary alcohols in formic acid has been determined [27], [28]. Because of its extremely reactive carboxyl group, amines can be formylated by formic acid, with high yields, for example, the yield of *N*-methylformanilide [*93-61-8*] is 93 – 97 %. Formic acid adds to the double bonds of olefins to form esters. Acetylenes react with formic acid in the vapor phase to yield vinyl formates [29].

Formic acid may also be used as a source of carbon monoxide. Thus, carboxylic acids can be prepared from olefins and formic acid in the presence of sulfuric or hydrofluoric acid (Koch carboxylic acid synthesis). Formic acid and tertiary organic bases form addition compounds (3 : 1 and 2 : 1 ratio of formic acid : base). If, for instance, 1 mole of gaseous triethylamine is passed in 3 moles of anhydrous formic acid, the addition product is obtained in quantitative yield. In these compounds formic acid is present in an extremely active form. Addition compounds of formic acid and trimethylamine or triethylamine can be used as liquid reducing agents in many selective reductions. The reduction of sulfur dioxide to sulfur is approximately quantitative [30].

Formic acid reacts with hydrogen peroxide in the presence of an acidic catalyst to form unstable performic acid [*107-32-4*] (HCOOOH); performic acid (→ Peroxy Compounds, Organic) may decompose explosively when heated to 80 – 85 °C. Performic acid exhibits typical peroxide properties [31]. The reactions of formic acid were summarized by Gibson in 1969 [32].

Although formic acid is not very stable, it can be distilled at atmospheric pressure in the absence of catalysts. Formic acid may decompose into either carbon monoxide and water (dehydration) or carbon dioxide and hydrogen (dehydrogenation); the nature of the decomposition and the reaction rate depend strongly on the presence of catalysts, the temperature, and the concentration of formic acid. *Dehydrogenation* is catalyzed preferentially by metals (platinum sponge, copper, nickel, silver). *Dehydration*, on the other hand, is catalyzed by aluminum oxide, silicon dioxide, and charcoal. Dehydration is promoted by mineral acids but inhibited by water; it also occurs on activated metal surfaces which must be considered during the preparation and processing of formic acid, as well as during transportation and storage of the highly concentrated acid. Tables 8 and 9 show the formation of carbon monoxide as a function of formic acid concentration and temperature [33]. A summary of the catalytic decomposition of formic acid has been published by Mars [34], along with an extensive list of references. The temperature dependence of formic acid decomposition at low temperature (between 40 – 100 °C) has been investigated [35].

Table 8. Volume of carbon monoxide produced per liter of formic acid (99 wt %) as a function of storage time and temperature [33]

Time, *d*	*V*, L		
	at 20 °C	at 30 °C	at 40 °C
15	0.012	0.064	0.303
30	0.024	0.128	0.597
60	0.049	0.255	1.161
90	0.073	0.381	1.700

Table 9. Volume of carbon monoxide produced per liter of aqueous formic acid as a function of storage time and concentration [33]

Time, *d*	*V*, L ($t = 40$ °C)		
	90 wt %	98 wt %	99 wt %
15	0.010	0.110	0.303
30	0.019	0.220	0.597
60	0.039	0.437	1.161
90	0.058	0.653	1.700

4. Production

The worldwide installed capacity for producing formic acid was about 330 000 t/a in 1988. The installed formic acid processes can be classified in four groups:

1) methyl formate hydrolysis,
2) oxidation of hydrocarbons,
3) hydrolysis of formamide, and
4) preparation of free formic acid from formates.

The significance of individual processes has shifted considerably in the last decade because the economic data for production and raw products have changed and new processes have been developed. In the past, a large amount of the formic acid utilized was a byproduct in the manufacture of acetic acid. However, the proportion of formic acid produced by dedicated processes has increased.

Formic acid is a byproduct in the production of acetic acid [*64-19-7*] by liquid-phase oxidation of butane or naphtha (→ Acetic Acid, Chap. 4.2.). For many years, oxidation of hydrocarbons was the most important method of producing acetic acid. Over the past decade, however, the preferred process in the establishment of new acetic acid plants has been the carbonylation of methanol [36], which is expected to remain the preferred method in the future. As a result, the proportion of formic acid produced worldwide, which results as a byproduct in the manufacture of acetic acid, will continue to decrease.

The production of formic acid by hydrolysis of formamide [*75-12-7*] played an important role in Europe; in 1972, about one-third of the world production was obtained by this process [3]. However, the consumption of ammonia and sulfuric acid, along with the unavoidable production of ammonium sulfate, have in many cases made this process economically inferior. As a result, the direct hydrolysis of methyl formate is currently preferred.

Another industrial production method involves formation of the free acid from its salts. Sodium formate [*141-53-7*] and calcium formate [*544-17-2*] are used for this purpose.

The economic disadvantages of the methods discussed led to the development of a process specifically dedicated to the production of formic acid, with no undesirable byproducts. In the 1970s, the hydrolysis of methyl formate [*107-31-3*] to methanol and formic acid was developed commercially by various firms into an economically feasible method. This process involves carbonylation of methanol and subsequent hydrolysis of the methyl formate produced. The methanol resulting from this process is returned to the first stage. Formic acid plants based on this process were started up at BASF (Federal Republic of Germany) in 1981 and Kemira (Finland) in 1982 [37]. Most of the formic acid plants begun in recent years employ this principle of methyl formate hydrolysis.

Table 10 shows the world capacity for various formic acid production processes in 1986. Although other methods for producing formic acid have been patented, they do not appear to have been implemented industrially. Direct synthesis of formic acid from carbon monoxide and water cannot be carried out economically because of the unfavorable position of the equilibrium.

Table 10. Formic acid production capacity in 1986 (without standby plants; BASF assessment)

Process	Capacity, t/a	Percentage of total capacity, %
Hydrolysis of methyl formate	160 000	49
Byproduct in hydrocarbon oxidation	74 000	22
Hydrolysis of formamide	10 000	3
Acidolysis of alkali formates	86 000	26
Total	330 000	

In recent years, various ways of producing formic acid from carbon dioxide and hydrogen have been investigated extensively, but this method has not yet been applied industrially. The oxidation of methyl *tert*-butyl ether [*1634-04-4*] with oxygen to form *tert*-butyl formate [*762-75-4*] and formic acid has been described [38]. Various methods have been developed for the electrochemical production of formic acid, but they have not been used commercially [3], [39]. The combined production of formic acid and acetic acid from methyl formate, carbon monoxide, and water with methyl iodide [*74-88-4*] as promoter and anhydrous rhodium chloride [*13569-65-8*] as catalyst, has been described by BP [40].

4.1. Methyl Formate Hydrolysis

Synthesis of formic acid by hydrolysis of methyl formate is based on a two-stage process: in the

first stage, methanol is carbonylated with carbon monoxide; in the second stage, methyl formate is hydrolyzed to formic acid and methanol. The methanol is returned to the first stage:

$$CH_3OH + CO \longrightarrow HCOOCH_3$$

$$CH_3OOCH + H_2O \longrightarrow CH_3OH + HCOOH$$

$$CO + H_2O \longrightarrow HCOOH$$

Although the carbonylation of methanol is relatively problem-free and has been carried out industrially for a long time [3], [41], only recently has the hydrolysis of methyl formate been developed into an economically feasible process. The main problems are associated with work-up of the hydrolysis mixture. Because of the unfavorable position of the equilibrium, reesterification of methanol and formic acid to methyl formate occurs rapidly during the separation of unreacted methyl formate. Problems also arise in the selection of sufficiently corrosion-resistant materials.

Older publications suggest the reaction of methyl formate and a dicarboxylic acid, with subsequent distillation of formic acid from the higher-boiling esters [3]. Degussa suggested hydrolyzing methyl formate at 200 °C above atmospheric pressure and then reducing the pressure suddenly to less than 0.1 MPa. Separation of the methyl formate–methanol mixture flash-evaporated from the aqueous formic acid can then take place in a liquid separator [5].

Industrial methods involving carbonylation of methanol and hydrolysis of methyl formate, followed by isolation of the formic acid, were developed by (1) the Leonard Process Company (United States), (2) BASF (Federal Republic of Germany), (3) Halcon – Scientific Design – Bethlehem Steel (United States), and (4) the Scientific Research Institute for the Chlorine Industry of the Ministry of the Chemical Industry (former Soviet Union).

Key patents relating to the carbonylation of methanol and the hydrolysis of methyl formate are listed in [43].

Basic Principles. *Carbonylation of Methanol*. In the four processes mentioned, the first stage involves carbonylation of methanol in the liquid phase with carbon monoxide, in the presence of a basic catalyst:

$$CH_3OH + CO \longrightarrow HCOOCH_3 \quad \Delta H_R^0 = -29\ \text{kJ/mol}$$

This reaction was first described by BASF in 1925 [41]. As a rule, the catalyst is sodium methoxide [*124-41-4*]. Potassium methoxide [*865-33-8*] has also been proposed as a catalyst; it is more soluble in methyl formate and gives a higher reaction rate [44]. Although fairly high pressures were initially preferred, carbonylation is carried out in new plants at lower pressure. Under these conditions, reaction temperature and catalyst concentration must be increased to achieve acceptable conversion. According to published data, ca. 4.5 MPa, 80 °C, and 2.5 wt % sodium methoxide are employed. About 95 % carbon monoxide, but only about 30 % methanol, is converted under these circumstances. Nearly quantitative conversion of methanol to methyl formate can, nevertheless, be achieved by recycling the unreacted methanol. The carbonylation of methanol is an equilibrium reaction, and Table 11 specifies the equilibrium concentrations of methyl formate.

Table 11. Maximum methyl formate concentration (in weight percent) achievable by carbonylation of methanol (unpublished BASF data)

t, °C	p, MPa			
	1	5	9	17
50	59	95	98	99
60	43	91	96	99
70	31	85	94	98
80	21	78	90	96
90	15	69	85	94
100	10	59	79	91
110	7	49	71	87
120	5	39	62	82

The reaction rate can be raised by increasing the temperature, the carbon monoxide partial pressure, the catalyst concentration, and the interface between gas and liquid:

$$r = C \cdot a \cdot e^{-E/RT} \cdot c \cdot p_{CO}$$

where

r = reaction rate,
C = proportionality coefficient,
a = interface between gas and liquid,
E = activation energy,
R = molar gas constant,
c = catalyst concentration in solution, and
p_{CO} = CO partial pressure in the gas phase [45].

This equation reveals the dependence of reaction rate on the partial pressure of carbon monoxide. Therefore, to synthesize methyl formate, gas mixtures with a low proportion of carbon monoxide must first be concentrated; this can be achieved by low temperature distillation, pressure-swing adsorption, prism separators (Monsanto), or the Cosorb process (Tenneco).

In a side reaction, sodium methoxide reacts with methyl formate to form sodium formate and dimethyl ether, and becomes inactivated:

$$NaOCH_3 + HCOOCH_3 \longrightarrow HCOONa + CH_3OCH_3$$

The substances used must be anhydrous; otherwise, sodium formate is precipitated to an increasing extent:

$$NaOCH_3 + H_2O + HCOOCH_3 \longrightarrow 2\,CH_3OH + NaOOCH$$

Sodium formate is considerably less soluble in methyl formate than in methanol. The risk of encrustation and blockage due to precipitation of sodium formate can be reduced by adding poly(ethylene glycol) [46]. The carbon monoxide used must contain only a small amount of carbon dioxide; otherwise, the catalytically inactive carbonate is precipitated.

Basic catalysts may reverse the reaction, and methyl formate decomposes into methanol and carbon monoxide. Therefore, undecomposed sodium methoxide in the methyl formate must be neutralized [47].

Hydrolysis of Methyl Formate. In the second stage, the methyl formate obtained is hydrolyzed:

$$HCOOCH_3 + H_2O \rightleftarrows CH_3OH + HCOOH \qquad \Delta H_R^0 = 16.3\ \text{kJ/mol}$$

The equilibrium constant for methyl formate hydrolysis depends on the water: ester ratio [48]. With a molar ratio of 1, the constant is 0.14, but with a water: methyl formate molar ratio of 15, it is 0.24. Because of the unfavorable position of this equilibrium, a large excess of either water or methyl formate must be used to obtain an economically worthwhile methyl formate conversion. If methyl formate and water are used in a molar ratio of 1 : 1, the conversion is only 30 %, but if the molar ratio of water to methyl formate is increased to 5 – 6, the conversion of methyl formate rises to 60 %. However, a dilute aqueous solution of formic acid is obtained this way, and excess water must be removed from the formic acid with the expenditure of as little energy as possible.

Another way to overcome the unfavorable position of the equilibrium is to hydrolyze methyl formate in the presence of a tertiary amine, e.g., 1-(*n*-pentyl)imidazole [*19768-54-8*], [49]. The base forms a salt-like compound with formic acid; therefore, the concentration of free formic acid decreases and the hydrolysis equilibrium is shifted in the direction of products. In a subsequent step formic acid can be distilled from the base without decomposition.

A two-stage hydrolysis has been suggested, in which a water-soluble formamide is used in the second stage [50]; this forms a salt-like compound with formic acid. It also shifts the equilibrium in the direction of formic acid.

To keep undesirable reesterification as low as possible, the time of direct contact between methanol and formic acid must be as short as possible, and separation must be carried out at the lowest possible temperature. Introduction of methyl formate into the lower part of the column in which lower boiling methyl formate and methanol are separated from water and formic acid, has also been suggested. This largely prevents reesterification because of the excess methyl formate present in the critical region of the column [51].

The hydrolysis of methyl formate is catalyzed by strong acids, but the efficiency of strong mineral acids [52] is restricted because they also promote the decomposition of formic acid. In the processes described by BASF, Leonhard, and Halcon – Scientific Design – Bethlehem Steel, autocatalysis by formic acid is therefore used. In the Soviet Union, a process has been described in which hydrolysis is carried out in a vertical column reactor divided into two zones [53]; the upper zone is filled with a strongly acidic cation exchanger which partially hydrolyze methyl formate. In the lower part of the reactor, hydrolysis is carried out with autocatalysis by formic acid produced in the upper section.

Dehydration of the Hydrolysis Mixture. Formic acid is marketed in concentrations exceeding 85 wt %; therefore, dehydration of the hydrolysis mixture is an important step in the production of formic acid from methyl formate. For dehydration, the azeotropic point must be

overcome. The concentration of formic acid in the azeotropic mixture increases if distillation is carried out under pressure (Table 5), but the higher boiling point at high pressure also increases the decomposition rate of formic acid. At the same time, the selection of sufficiently corrosion-resistant materials presents considerable problems. A number of entrainers have been proposed for azeotropic distillation [43].

BASF describes various energy-saving processes for the dehydration. One process involves extractive distillation with *N*-formylmorpholine [*4394-85-8*] [54]. Another method of increasing formic acid concentration in dilute solutions without a considerable input of energy is to extract formic acid from water by liquid – liquid extraction. Secondary amides have been proposed [55] as extractive agents. Formic acid binds weakly to the amide; therefore, water can be distilled to overcome the azeotropic point. In a subsequent column, formic acid can then be distilled from the extractant in vacuo. The energy consumed can be reduced further by optimizing the exchange of heat [56].

4.1.1. Kemira – Leonard Process

A 20 000-t/a formic acid plant based on a method developed by the Leonard Process Co. [57] was built at Kemira in Finland and put into operation in 1982. The process has been developed further by Kemira, and licences for it have been issued in Korea, India and Indonesia.

In the Kemira – Leonard process, Methyl formate and carbonylation is carried out at about 4 MPa and a temperature of approximately 80 °C, with additive-containing alkoxides used as catalyst [58]. Hydrolysis is carried out in two reactors with different operating conditions. Methyl formate and water react in the preliminary reactor in approximately equimolar proportions. The formic acid produced catalyzes the hydrolysis in the main reactor. An excess of methyl formate is employed in the main reactor, and hydrolysis is carried out at ca. 120 °C and 0.9 MPa. The reactor discharge is brought to atmospheric pressure in a flash tank; reesterification is largely prevented by the cooling that occurs during flash-evaporation. methanol are separated under vacuum. Although the reverse reaction is attenuated further as a result, condensation of the low-boiling methyl formate in vacuo presents a problem. The formic acid is dehydrated by distillation. If 85 wt % formic acid is desired, dehydration must be carried out at ca. 0.3 MPa. Even higher concentrations can be achieved by connecting an additional dehydration column downstream under atmospheric pressure, and formic acid concentrations up to ca. 98 wt % can then be drawn off as the distillate.

Process Description [57–60] (Fig. 1). Compressed carbon monoxide and methanol are converted into methyl formate in reactor (a). Catalyst is fed into the reactor in a methanol solution. The amount of methanol introduced in this way makes up for methanol losses in the process. The discharge from reactor (a) is flashed and fed into the methyl formate column (b) from which methyl formate is drawn off as the distillate. Methanol and the dissolved catalyst are returned to the reactor; inactivated catalyst (primarily sodium formate) is crystallized and discharged.

Off-gas from column (b) and waste gas from reactor (a) are burned. Methyl formate reacts partially with water in the preliminary reactor (c), and discharge from the preliminary reactor is fed into the main reactor (d) along with recycled methyl formate, methanol, and water. The contact time in this reactor is largely sufficient for equilibrium to be established. Reactor discharge is flashed to approximately atmospheric pressure in the flash tank (e); methyl formate, methanol, and small quantities of formic acid evaporated in this process are recycled to the main reactor (d). Methyl formate and methanol are distilled in vacuo in the acid separation column (g). Contact time is minimized by internals with a small liquid holdup. The distillate is separated into methyl formate and methanol in the recycle column (f). Formic acid can be concentrated under pressure in a column; if the pressure is ca. 0.3 MPa, 85 % formic acid is drawn off as the bottom product. Distillation in two columns is more industrially controllable. Water is distilled overhead in the first product column (h). The bottom product is concentrated further in the second product column (i), and formic acid with a maximum concentration of ca. 98 wt % is drawn off as distillate. The bottom product is recycled to the first product column.

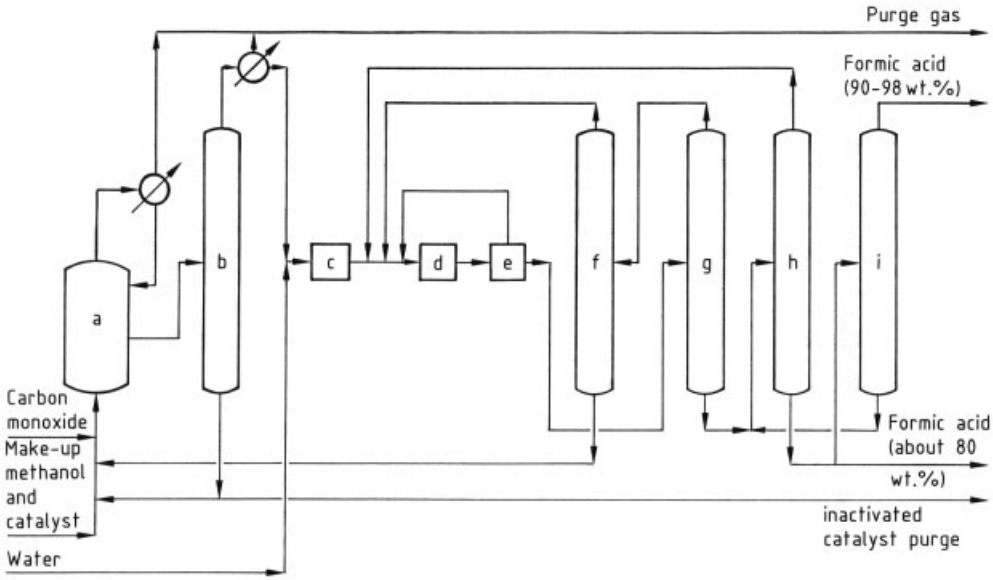

Figure 1. Production of formic acid (Kemira – Leonard process)
a) Methyl formate reactor; b) Methyl formate column; c) Preliminary reactor; d) Main reactor; e) Flash tank; f) Recycle column; g) Acid separation column; h) First product column; i) Second product column

4.1.2. BASF Process

A 100 000-t/a formic acid plant began operating in Ludwigshafen (Federal Republic of Germany) in 1981. In this plant, a technology for the hydrolysis and dehydration was used for the first time.

The production of methyl formate by carbonylation of methanol has been carried out on a large scale for many years at BASF [41], [61]. The carbonylation stage is largely identical to that of the Kemira – Leonard process, but the hydrolysis stage and the dehydration of formic acid are noticeably different.

In the BASF process, hydrolysis is carried out with a large excess of water (about 5 mol of water per mole of methyl formate) to shift the equilibrium in the direction of formic acid. Much of the water is separated by liquid – liquid extraction with a secondary amide.

Process Description (Fig. 2) Carbon monoxide and methanol react in the methyl formate reactor (a) in the presence of sodium methoxide. Methyl formate is fed, as a distillate, from the methyl formate column (b) into the formic acid reactor (c) together with recycled methyl formate. Methanol and dissolved catalyst are drawn off from the bottom of column (b) and returned to reactor (a); catalyst decomposition products are discharged by crystallization. In reactor (c), methyl formate is hydrolyzed with excess water at elevated temperature and increased pressure. The reaction product is flashed into the low-boiler column (d). Methyl formate is removed as the distillate, with methanol as a side stream, and dilute aqueous formic acid is drawn off from the bottom into the extraction unit (e). Here, the formic acid and some of the water are extracted by the secondary amide. Most of the water (largely free of formic acid) is recycled to reactor (c). The extract–a mixture of extractant, formic acid, and some water–is distilled in the dehydration column (f). Enough water is distilled via the head for the required formic acid concentration to be obtained in the pure acid column (g). This column is operated in vacuo. The extraction agent is recycled from the bottom of column (g) to the extraction unit.

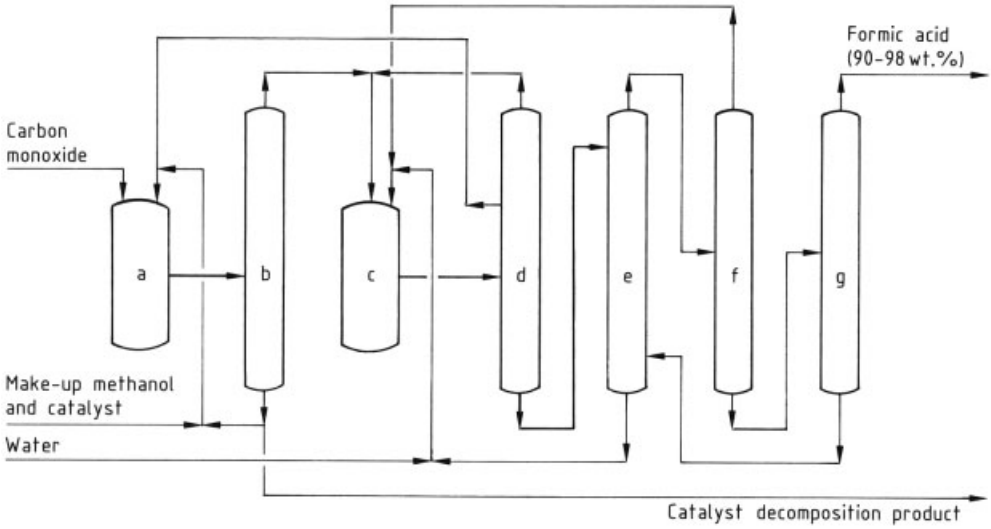

Figure 2. Production of formic acid (BASF process)
a) Methyl formate reactor; b) Methyl formate column; c) Hydrolysis reactor; d) Low-boiler column, e) Extraction unit; f) Dehydration column; g) Pure acid column

4.1.3. USSR Process

A 40 000-t/a formic acid plant based on a process [53] developed in the former Soviet Union is being built in Saratov (Ukraine) [62] and is expected to be operational in 1989.

According to published material, this formic acid process differs from the above-mentioned processes [63], [64] in the hydrolysis stage. Hydrolysis is carried out at 80 °C and 0.6 MPa in a fixed-bed reactor equipped with bubble-cap trays in the upper section. Reesterification is prevented by decompressing and cooling to about 45 °C. Formic acid is produced in a pressure column as an azeotropic mixture with water which contains 85 wt % formic acid.

Process Description [63], [64] (Fig. 3). Carbon monoxide reacts with methanol in column reactor (a), in the presence of the catalyst and a stabilizer, at about 3 MPa, to yield methyl formate. Separation into methyl formate (distillate) and methanol plus catalyst takes place in the methyl formate column (b) at 0.2 MPa. The exhaust gas produced is partially recycled. In reactor (c), methyl formate is hydrolyzed with water in two stages; in the first stage, packed acidic cation exchanger is used as catalyst (f), which partially hydrolyzes the methyl formate. In the second stage, hydrolysis is carried out with autocatalysis by the formic acid produced in stage one. Methanol and methyl formate are distilled from the hydrolysis product in the low-boiler column (d); these compounds are separated, along with the carbonylation reaction mixture, in the methyl formate column (b). Aqueous formic acid is drawn off the bottom of column (d) and further dehydrated in the formic acid column (e) under pressure. Formic acid with a content of 85 wt % is drawn off in the side stream, and the concentration can be increased to about 98 wt % in the column situated downstream.

4.1.4. Scientific Design – Bethlehem Steel Process

The process developed jointly by Scientific Design and Bethlehem Steel [65] has not yet been implemented industrially on a large scale.

The carbonylation stage is essentially similar to the processes described previously. If the concentration of carbon monoxide is higher than 90 mol %, a single back-mixed reactor is provided, but if the carbon monoxide concentration is between 50 and 90 mol %, a two-stage, countercurrent, back-mixed reactor system is recommended. If the carbon monoxide concentration

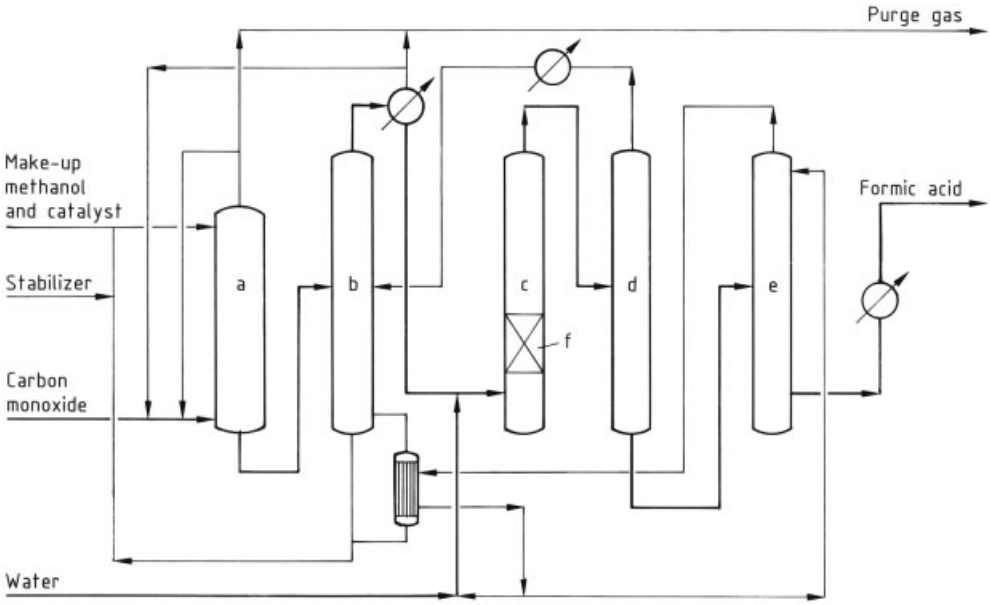

Figure 3. Production of formic acid (USSR process)
a) Methyl formate reactor; b) Methyl formate column; c) Hydrolysis reactor; d) Low-boiler column; e) Formic acid column; f) Fixed-bed catalyst

is low, however, the total pressure must be increased considerably [66]. The hydrolysis is catalyzed by formic acid, and the dehydration is carried out by distillation alone.

4.2. Other Processes

4.2.1. Oxidation of Hydrocarbons

Formic acid is produced as a byproduct in the liquid-phase oxidation of hydrocarbons to acetic acid (→ Acetic Acid, Chap. 4.2.). In the United States, butane [*106-97-8*] is used as the hydrocarbon, and ca. 50 kg of formic acid is produced per ton of acetic acid. In Europe, the oxidation of naphtha is preferred, and up to 250 kg of formic acid is produced per ton of acetic acid in this process. For the mechanism of formic acid production, see [59].

Unreacted hydrocarbons, volatile neutral constituents, and water are separated first from the oxidation product. Formic acid is separated in the next column; azeotropic distillation is generally used for this purpose. The entrainers preferred in this process are benzene or chlorinated hydrocarbons. The formic acid contains about 2 wt % acetic acid, 5 wt % water, and 3 wt % benzene [59]. Formic acid with a content of about 98 wt % can be produced by further distillation.

4.2.2. Hydrolysis of Formamide

In 1972, 35 % of the formic acid produced worldwide was still made by the formamide process developed by BASF. However, because of the direct hydrolysis of methyl formate, this process has lost much of its significance (see Table 10).

Formic acid is produced this way in a three-stage process. In the first stage, methanol is carbonylated to yield methyl formate:

$$CO + CH_3OH \rightarrow HCOOCH_3 \qquad \Delta H_R^0 = -29\ \text{kJ/mol}$$

In the second stage, formamide is produced by ammonolysis of ethyl formate:

$$HCOOCH_3 + NH_3 \rightarrow HCONH_2 + CH_3OH \qquad \Delta H_R^0 = -72\ \text{kJ/mol}$$

In the third stage, sulfuric acid is used to hydrolyze formamide to formic acid and ammonium sulfate:

$$HCONH_2 + 1/2\,H_2SO_4 + H_2O \longrightarrow HCOOH + 1/2\,(NH_4)_2SO_4 \qquad \Delta H_R^0 = -48\ \text{kJ/mol}$$

Methanol is carbonylated as in the first stage of the processes already described (Section 4.1). Conversion to formamide is usually carried out at 0.4 – 0.6 MPa and 80 – 100 °C; methanol is distilled and recycled to the methyl formate stage (see → Formamides). Formamide is hydrolyzed continuously in the third stage by using 68 – 74 % sulfuric acid at temperatures between 85 °C and the boiling point of formic acid [3]. This reaction is carried out preferably in stirred containers and provides the heat for the distillation of formic acid. A hot slurry consisting essentially of ammonium sulfate and formic acid flows out of the stirred container into a rotary kiln where the residual formic acid is distilled so that dry, pure ammonium sulfate is produced at the kiln end. The yield of formic acid exceeds 90 %. The economic efficiency of this process is determined by the commercial value of ammonium sulfate.

Hydrochloric acid, nitric acid, or phosphoric acid can also be used to hydrolyze formamide [3].

4.2.3. Production of Formic Acid from Formates

The reaction of sodium formate [*141-53-7*] or calcium formate [*544-17-2*] with strong mineral acids, such as sulfuric and nitric acids, is the oldest known process for producing formic acid commercially. If formates or sodium hydroxide are available cheaply or occur as byproducts in other processes, formic acid can still be produced economically in this manner.

Formates as Byproducts in the Production of Polyhydric Alcohols. Sodium formate and calcium formate occur as byproducts in the production of pentaerythritol [*115-77-5*], trimethylolpropane [*77-99-6*], and 2,2-dimethyl-1,3-propanediol [*126-30-7*] (neopentyl glycol) (→ Alcohols, Polyhydric). The formates unavoidably produced in this manner are used in several plants as feed for manufacturing formic acid.

Formates from Sodium Hydroxide and Carbon Monoxide. The production of sodium formate from sodium hydroxide and carbon monoxide once had considerable significance, but today this process is carried out simply to utilize excess sodium hydroxide solution in low-capacity plants. The economic efficiency of producing formic acid from its salts suffers because one equivalent of a low-value inorganic salt (e.g., sodium sulfate), is produced per mole of formic acid.

Carbon monoxide reacts with alkaline compounds, even in aqueous solution, to yield the corresponding formates:

$$CO + NaOH \xrightarrow{H_2O} HCOONa$$

Carbon monoxide is mixed countercurrently with aqueous sodium hydroxide, for example, in a tower reactor at 1.5 – 1.8 MPa and 180 °C [59]. Sodium formate crystallizes and reacts with strong mineral acids, (e.g., concentrated sulfuric acid) at normal pressure in a cooled, stirred reactor at 35 °C:

$$2\,HCOONa + H_2SO_4 \rightarrow 2\,HCOOH + Na_2SO_4$$

The reaction mixture is separated, for example, in a thin-film evaporator at normal pressure and 100 – 120 °C, to produce formic acid and dry sodium sulfate [67]. The suggestion has been made that reaction with sulfuric acid be carried out in two stages. In this case, the first stage takes place in vacuo in a horizontal tubular reactor with rotating internals, and some of the formic acid is evaporated at the same time. In the second stage, the remaining formic acid is evaporated, also in vacuo, at 30 – 60 °C in a smaller tubular reactor [68].

Norsk Hydro describes the production of formic acid from calcium formate by reaction with nitric acid. The calcium nitrate produced may be used, for example, as a fertilizer.

4.2.4. Direct Synthesis from Carbon Monoxide and Water

The simplest theoretically possible method of producing formic acid is to react carbon monoxide with water:

$$H_2O + CO \rightleftarrows HCOOH$$

As pressure increases and temperature decreases, the equilibrium of this exothermic reaction shifts in favor of formic acid. Below 150 °C, the reaction rate is very slow, and although equilibrium is reached rapidly at higher temperature, the pressure must be increased drastically to obtain acceptable formic acid concentrations. Table 12 lists the calculated equilibrium concentrations at various temperatures and pressures.

Inorganic acids and salts have been described as catalysts in the literature. Unless the reaction rate can be increased markedly by the development of new catalysts, direct synthesis cannot be used for the economical production of formic acid.

Table 12. Calculated equilibrium concentrations of formic acid at different temperatures and pressures [69]

Carbon monoxide pressure, MPa			Formic acid concentration, wt %
at 25 °C	at 100 °C	at 217.9 °C	
0.027	0.63	6.95	1
0.30	7.15	79.1	10
0.72	16.95	187	20
3.7	85.8	950	50
108	2540	28170	90

4.2.5. Use of Carbon Dioxide

Interesting developments in carbon dioxide chemistry have resulted in formic acid syntheses, but so far these have not been used industrially [70].

Zinc selenide [*1315-09-9*] and zinc telluride [*1315-11-3*] have been proposed as catalysts for the hydrogenation of carbon dioxide [71]. If homogeneous transition-metal catalysts are employed, carbon dioxide, hydrogen, and water can be converted to formic acid. If alcohol is used instead of water, the corresponding ester is produced. Ruthenium and palladium complexes have proved particularly active; either inorganic bases [72] or organic bases [73] (e.g., aliphatic tertiary amines) can be used as the alkaline material.

BP Chemicals has developed a process in which formic acid is produced from carbon dioxide and hydrogen via several process stages [74–76]. In the first stage, a nitrogen base such as triethylamine [*121-44-8*] reacts with carbon dioxide and hydrogen in the presence of a noble-metal catalyst (e.g., a ruthenium complex) to yield the formate of the nitrogen base. In the second and third stages, the formate is separated from the catalyst and the low-boiling constituents. In the fourth stage, the formate reacts with a high boiling base, for example, 1-(*n*-butyl)imidazole [*4316-42-1*], to yield a formate that can be thermally decomposed. Simultaneously, the low-boiling base is liberated and distilled. In the fifth stage, the formate is thermally decomposed, formic acid distilled, and the high-boiling base is returned to the fourth stage.

4.3. Concentration of Formic Acid

Formic acid – water mixtures cannot be concentrated to more than the azeotropic composition by simple distillation (→ Distillation and Rectification). Further dehydration is carried out by ternary azeotropic distillation, extractive distillation, or extraction. Additional concentration can then be achieved by simple distillation. Through suitable choice of distillation conditions, formic acid with a content exceeding 99 wt % can be drawn off as distillate.

Azeotropic Distillation. Propyl and butyl formates have been proposed as entrainers for azeotropic distillation [3]. A particularly economical separation of water from the water – formic acid – acetic acid mixture obtained by oxidizing butane is achieved by azeotropic distillation with ethyl *n*-butyl ether [*628-81-9*] [77].

Extractive Distillation. In extractive distillation, formic acid is extracted in a distillation column by means of a basic extracting agent introduced countercurrently and fed into the bottom of the column. If this mixture is heated in a column downstream, the formic acid is liberated from the salt-like compound and distilled. *N*-Formylmorpholine [*4394-85-8*] has been suggested as extractant [54], and various sulfones have been described as auxiliary liquids for extractive distillation [78].

Extraction. A number of extractants have been proposed for extracting acetic and formic acids [79]. According to BASF, secondary amides are efficient extracting agents [55]; Hüls recommends compounds in the series tri-*n*-

octylamine [*1116-76-3*] to tri-*n*-dodecylamine [*102-87-4*] [80].

4.4. Construction Materials

The use of sufficiently corrosion-resistant materials in the individual process stages is part of the expertise of formic acid production. Tests on materials under laboratory conditions do not always guarantee that corrosion resistance can be achieved in practice. A low impurity content or a change in concentration or temperature may produce a decisive change in the corrosion resistance of a material toward formic acid.

If austenitic CrNi steels are used, those that contain molybdenum are preferred to molybdenum-free types [81]. Above 20 °C, stainless steel AISI 304 (Cr 18 wt %, Ni 8 wt %) has only adequate resistance, if the formic acid concentration is below 10 % or above 90 %. Stainless steel AISI 316 (Cr 18 wt %, Ni 10 wt %, Mo 2 wt %) is resistant up to ca. 50 °C to formic acid below 40 wt % or over 80 wt % [82]. For higher temperatures and formic acid concentrations between 40 and 80 wt %, steels with higher Cr, Ni, and Mo contents (for example, AISI 317 or 317 L) or special nickel-rich alloys must be used. Aluminum ist not sufficiently resistant to formic acid, except at concentrations below 30 wt % or above 80 wt % and at temperatures below 20 °C [82]. Titanium is resistant under oxidizing conditions but is attacked very strongly under reducing conditions [81]. The corrosion behavior of titanium is considerably improved by addition of platinum or palladium. The more exotic metals tantalum, niobium, and zirconium are also corrosion resistant [81].

Of the nonmetallic materials, graphite and glass are sufficiently corrosion-resistant, but their use is limited because of their poor mechanical properties. Among the common plastics, polyethylene is resistant up to about 50 °C. Polytetrafluorethylene (PTFE) and poly(vinylidene fluoride) (PVDF) are resistant to formic acid. Hard poly(vinyl chloride) (PVC) is resistant only at room temperature, and soft PVC should be used only for formic acid concentration below 10 wt %.

Liquid-crystal polymers (LCP) are recommended as packing materials for columns because their resistance to formic acid is reportedly much better than that of ceramics. A survey of construction materials used in formic acid production is given in [83].

5. Environmental Protection

Like most other simple organic acids, formic acid can be degraded rapidly and completely by biological methods. For this reason, formic acid is totally mineralized in a short time in purification plants and streams. The ecological equilibrium of the streams can, however, be disturbed by a change in pH. For formic acid, the theoretical biological oxygen demand (BOD) is 350 mg/g. The chemical oxygen demand has been determined to be 86 mg/g after 5 d (COD_5) and 250 mg/g after 20 d (COD_{20}). In the modified OECD screening test (OECD method 30 IE), formic acid is degraded to >90 % [84]. If an additional carbon source is available, degradation proceeds more quickly.

Toxicity to fish was tested on the blue gill sunfish (*Lepomis macrochirus*). The lethal dose (TLm 24) was found to be 175 ppm [85]. Toxicity in invertebrates was determined on *Daphnia magna*. The TLm 48 value was 120 ppm [86]. In the Federal Republic of Germany, formic acid is classified among the substances that slightly endanger water (Class 1 of water endangerment, WGK 1) [87].

Formic acid vapor can be removed from exhaust streams by washing with water. Fog results when formic acid vapor is emitted during periods of high humidity. In the Federal Republic of Germany emission is controlled by the technical guidelines for air (TA Luft) [88]. Formic acid belongs to Class 1, which means that for emission rates of 0.1 kg/h or higher, the formic acid concentration must be <20 mg/m^3 in the exhaust gas. Upon filling of formic acid, the displaced air is recycled via a gas-transfer line or it is emitted after purification (e.g., by water washing or passage through an incinerator).

6. Quality Specifications

Formic acid is marketed in concentrations of 85, 90, 95, 98, and >99 wt %. The impurity content depends on the production process. Formic

acid produced by oxidation of hydrocarbons still contains a small proportion of acetic acid. A typical specification for formic acid produced by hydrolysis of methyl formate is given in Table 13; this acid meets the specifications required in the Food Chemicals Codex (FCC) [90].

Table 13. Sales specification for formic acid [89]

Specification	Limits	Analytical method
Content	min. 85 wt % min. 90 wt % 99 wt %	titrimetric, ISO 731/II, 5.2
Color number	max. 10 APHA	ASTM D 1209-79
Acetic acid	max. 150 mg/kg	ISO 731/VII
Sulfate	max. 1 mg/kg	ISO 731/V
Chloride	max. 1 mg/kg	ISO 731/IV
Iron	max. 1 mg/kg	ISO 731/VI
Other heavy metals	max. 1 mg/kg	FCC III
Evaporation residue	max. 20 mg/kg	DIN 53 172
Quality	the formic acid complies with DAB 7 and FCC III	

7. Chemical Analysis

The concentration of formic acid is normally determined by titration with sodium hydroxide solution. Volatile impurities can be determined by gas chromatography on packed columns. Polyesters or Carbowax 20 M, both treated with phosphoric acid and adsorbed on chromosorb as a carrier, have proved successful as the stationary phase. Formic acid content is measured with a thermal-conductivity cell because the substance cannot be detected with flame ionization detectors (FID).

Test methods for formic acid have been standardized by the International Organization for Standardization (ISO) [91]. Other standard analytical methods for formic acid are described in [92].

8. Storage and Transportation

Formic acid is stored in tanks made of austenitic CrNi steels (18 % Cr, 10 % Ni, 2 % Mo), for example, AISI 316 L. The use of other materials, such as AISI 304 L, polyethylene-lined carbon steel, glass-lined carbon steel, and polypropylene, is restricted to specific concentrations and low temperatures. Tank cars and tank trucks made of AISI 316 L are used for transportation. Small canisters may be made of polyethylene; glass bottles are used for laboratory quantities. When aqueous formic acid mixtures are stored or transported the melting points at various concentrations (Table 3) should be considered. Tanks are best heated with hot water or subsidiary electrical heating systems. To prevent corrosion, even of AISI 316, wall temperatures above 50 °C must be avoided. Formic acid decomposes to highly toxic carbon monoxide and water. For this reason, formic acid with a concentration of 99 wt % or higher is transported in canisters without gastight seals [3]. Containers of highly concentrated formic acid should be protected against exposure to heat and stored in a well-ventilated place. Before entering ship tanks or storage tanks that have contained highly concentrated formic acid, personnel should monitor the carbon monoxide content of the air. The storage method should ensure that contact with strong alkali solutions (evolution of heat) and strong acids or oxidizing agents (decomposition) does not occur. Table 14 contains important safety data.

9. Legal Aspects

Formic acid is marked according to the regulations of the European Economic Community (council directive 76/907/EEC including the latest amendment) and to the regulations of the Federal Republic of Germany (Gef Stoff V) [94] as: Corrosive, R34 (formic acid content 25 – 90 wt %), R35 (formic acid content > 90 wt %), and S 2-23-26. The EEC numbers of formic acid are 607-001-00-0 (formic acid content >90 wt %) and 607-001-01-8 (formic acid content 25 – 90 wt %). Hazard classifications for the transport of formic acid are listed in the following:

RID/ADR	class 8, item 32 b
Fed. Rep. of Germany: GGVE/GGVS	class 8, item 32 b
ADNR	class 8; item 21 b
IMDG Code (Fed. Rep. of Germany: GGVSee)	8
IATA-DGR	1779
UN Number	1779
DOT Hazard Classification (USA)	corrosive material
HazChem Code (UK)	2 R

Table 14. Safety data [93] for formic acid

	Formic acid content, wt %				
	85	90	94	99	Method
Flashpoint, °C	59	58	55	48	DIN 51 755
Autoignition point, °C	500	480	480	480	DIN 51 794
Flammable limits in air, vol %	15 – 47	14 – 44	14 – 38	12 – 38	

10. Uses

Because of its acidity, its aldehyde nature, and its reducing properties, formic acid is used in a variety of fields (Table 15).

Table 15. Applications of formic acid (1988, BASF assessment)

Textiles, leather	25 %
Pharmaceuticals, crop-protection agents	10 %
Latex, rubber auxiliaries	10 %
Silage	35 %
Miscellaneous	20 %

In contrast to mineral acids, formic acid evaporates without leaving any residue. The leather and textile industries, in particular, make use of its strongly acidic nature. Animal skins and hides are acidified with a mixture of formic acid, sulfuric acid, and sodium chloride before chrome tanning (→ Leather).

Formic acid is used for adjusting the pH of dye baths in dyeing of natural and synthetic fibers. Use is made of its highly reactive nature in the manufacture of pharmaceuticals and crop-protection agents [32], [43]. Formic acid plays an important role in the coagulation of rubber latex.

In Europe, most of the formic acid is used as a silage aid. If freshly mown, damp grass is sprayed with formic acid to establish a pH of ca. 4, such fodder can be used directly for silage. Formic acid promotes the fermentation of lactic acid and suppresses the harmful formation of butyric acid. If formic acid is used, fermentation processes occur more rapidly and at lower temperature so that fewer nutrients are lost [95], [96]. Mixtures of formic acid with formaldehyde or with ammonium formate are also used as silage aids. Silage that utilizes formic acid is restricted largely to Europe, but it should be of interest in all regions where climate makes *green forage* difficult to dry.

Other applications of formic acid are as an additive for cleaning agents (replacement of mineral acid for environmental reasons), in the synthesis of the sweetener aspartame [*22839-47-0*] [97] (→ Amino Acids, Chap. 5.1.2.), and in the desulfurization of flue gas by the Saarberg – Hölter process (pH stabilization of the wash liquid) [98]. A large potential market for formic acid is anticipated by the Leonard Process Co. in steel pickling and wood pulping [58]. In both fields, however, formic acid has not yet replaced the standard techniques (steel pickling with hydrochloric acid or wood pulping by the Kraft process).

Table 16. Worldwide formic acid production capacity (BASF assessment)

Company	Basis of process	Capacity, t/a
BASF (Fed. Rep. of Germany)	methyl formate	100 000
BP (UK)	naphtha	50 000
Kemira Oy (Finland)	methyl formate	40 000
Norsk Hydro (Norway)	calcium formate	13 000
Chem. Werke Hüls (Fed. Rep. of Germany)	sodium formate	10 000
Perstorp (Sweden)	sodium formate	10 000
Far (Italy)	sodium formate	10 000
CSSR	formamide	10 000
Celanese (United States)	butane	13 000
PT Pupuk Kujang (Indonesia)	methyl formate	10 000
GNFC (India)	methyl formate	5 000
Korea Fertilizer (Republic of Korea)	methyl formate	5 000
Daicel (Japan)	naphtha	5 000
Others	sodium formate calcium formate formamide	50 000
Under construction or planned (start-up):		
USSR	methyl formate	40 000 (1989)
USSR	methyl formate	40 000 (1990)

11. Economic Aspects

A breakdown of production capacity for formic acid is given in Table 16.

12. Derivatives

Among the derivatives of formic acid, amides (→Formamides), various esters, and salts are the most important.

12.1. Esters

Esters of formic acid have the typical physical and chemical properties of esters; they are readily volatile and flammable, and have a sweet, often fruity smell. They are only partially soluble in water. A summary of their chemical properties has been published by GIBSON [32].

12.1.1. Methyl Formate

Methyl formate [*107-31-3*], $C_2H_4O_2$, M_r 60.05, is an intermediate in the production of formic acid (Sections 4.1, 4.2) and formamide (→ Formamides, Chap. 2.3.) in large manufacturing plants. With increasing interest in the production of synthesis gas by coal gasification, methyl formate will become important as a versatile key intermediate. It can also be used as a starting material in the production of high-purity carbon monoxide [99]. Physical properties, safety data, and marking regulations for methyl formate are summarized below.

Boiling point (101.3 kPa)	31.7 °C [17]
Melting point	−99 °C [17]
Specific gravity d_4^{20}	0.975 [17]
Refractive index n_D^{20}	1.3433 [99]
Flash point (open cup)	−20 °C [99]
Ignition temperature	450 °C [99]
Flammable limits in air	5.0 – 23 vol % [99]
MAK value	100 mL/m^3; 250 mg/m^3 (category I)
TLV–TWA	100 mL/m^3; 250 mg/m^3
TLV–STEL	150 mL/m^3 ; 375 mg/m^3
UN number	1243
EEC number	607-014-00-1
Marking according to Gef Stoff V (FRG)	symbol : F R12, S 9-16-33
Hazchem Code (UK)	2 SE
RID/ADR	class 3, item I a
ADNR	class 3, item I a
IMDG Code (F.R. of Germany GGVSee)	3.1
IATA-DGR	1243
DOT Hazard Classification (USA)	flammable liquid

Production. Methyl formate is produced by base-catalyzed carbonylation of methanol (Section 4.1). Processes have been developed for dehydrogenating methanol, but these have not yet been implemented commercially [100]:

$$2\,CH_3OH \longrightarrow HCOOCH_3 + 2\,H_2 \quad \Delta H_R^0 = 98.9\,\text{kJ/mol}$$

Catalysts containing copper are used.

Mitsubishi Gas Chemical (MGC) has developed highly active catalysts which contain Cu – Zr – Zn or Cu – Zr – Zn – Al. The conversion of methanol is approximately 50 %, with a methyl formate selectivity of ca. 90 % and a space – time yield of 3000 g L^{-1} h^{-1} [101]. Air Products uses a copper – chromite catalyst for the dehydrogenation of methanol [102]. Methyl formate can also be produced by oxy-dehydrogenation of methanol:

$$2\,CH_3OH + O_2 \longrightarrow HCOOCH_3 + 2\,H_2O \quad \Delta H_R^0 = -472.8\,\text{kJ/mol}$$

Soluble chromium compounds are used as catalyst [103].

The direct synthesis of methyl formate from synthesis gas has been investigated:

$$2\,CO + 2\,H_2 \longrightarrow HCOOCH_3 \quad \Delta H_R^0 = -157.2\,\text{kJ/mol}$$

The high pressure of 100 – 200 MPa (1000 – 2000 bar) required makes economical implementation impossible [100]. The dimerization of formaldehyde has likewise not been exploited industrially [100]:

$$2\,CH_2O \longrightarrow HCOOCH_3 \quad \Delta H_R^0 = -146.4\,\text{kJ/mol}$$

Uses. Most of the methyl formate produced is used as an intermediate for the production of formic acid and formamide. Dimethylformamide [*68-12-2*] is produced by reacting methyl formate with dimethylamine [*124-40-3*] (→ Formamides, Chap. 3.3.). A new use has been found for methyl formate in the production of foundry molds. The mold is formed from granular refractory material and a binder consisting of phenol – formaldehyde resin which is cured by exposure to methyl formate vapor. This process makes possible the production of rapidly cured foundry molds at low temperature [104].

Small quantities of methyl formate are used as solvents and insect control agents. With the anticipated extension of C_1 chemistry, methyl formate may be used as an intermediate for a number of products [99], [100].

12.1.2. Ethyl Formate

Ethyl formate [*109-94-4*], $C_3H_6O_2$, M_r 74.08, *bp* 54.1 °C, *mp* −80 °C, $d_{4^0}^2$ 0.923, flash point −20 °C, ignition point 440 °C, UN number 1190, has a sharp rumlike smell and taste.

Ethyl formate is produced by carbonylation of ethanol with carbon monoxide or by esterification of ethanol with formic acid.

Ethyl formate is a solvent for acetyl cellulose and nitrocellulose. Ethyl formate is used in alcohol-free drinks, ice cream, chewing gum, and other confectionery as a component of apple-, pineapple-, banana-, and peach-type flavors.

12.1.3. Isobutyl Formate

Isobutyl formate [*542-55-2*], M_r 102.13, *bp* 98 °C, *mp* −96 °C, $d_{4^0}^2$ 0.88, flash point 5 °C, ignition point 320 °C, UN number 2393, is produced by carbonylation of isobutyl alcohol [*78-83-1*] with carbon monoxide or esterification of isobutyl alcohol with formic acid. Isobutyl formate is used as a solvent for colors, lacquers, adhesives, and cleansing agents.

12.1.4. Miscellaneous Esters

A number of formates produced by esterification of the corresponding alcohols with formic acid are used as perfumes and flavorings (→ Flavors and Fragrances). Some representative products are citronellyl formate [*105-85-1*], geranyl formate [*61759-63-5*], phenylethyl formate [*104-62-1*], benzyl formate [*104-57-4*], and isopentyl formate [*110-45-2*].

Orthoformates are used in the preparation of acetals and ketals. Chloroformic esters are used as intermediates in organic syntheses.

12.2. Salts

12.2.1. Alkali-Metal and Alkaline-Earth Metal Formates

Sodium formate [*141-53-7*] and calcium formate [*544-17-2*] are byproducts in the synthesis of polyols such as pentaerythritol. However, they are also produced directly from the corresponding hydroxides and carbon monoxide (Section 4.2.3) [3].

Sodium or calcium formate is used to produce formic acid (Section 4.2.3). An important process for manufacturing sodium dithionite [*7775-14-6*] starts with sodium formate. Oxalic acid [*144-62-7*] production employs sodium formate as an intermediate. Sodium formate is used in chrome tanning and as a mordant in the dyeing and printing of fabrics by the textile industry. The reducing power of sodium formate is utilized in electroplating baths and photographic fixing baths. Calcium formate is also used in tanning. Adding calcium formate to concrete reduces the setting time considerably.

12.2.2. Ammonium Formate

Ammonium formate solutions are used as a low-corrosion silage aid; the complex salts ammonium diformate [*64165-14-6*] and ammonium tetraformate [*70179-79-2*] are stable in aqueous solution [105].

12.2.3. Aluminum Formate

Stepwise replacement of the hydroxyl groups in aluminum hydroxide [*21645-51-2*] yields dibasic aluminum formate [*18748-09-9*], $(HO)_2Al(OOCH)$; Almonobasic aluminum formate [*51575-25-8*], $(HO)Al(OOCH)_2$; and aluminum formate [*22918-74-7*], $Al(OOCH)_3$. Aluminum hydroxy acetate formate [*34202-30-7*] $(HO)Al(OOCH)(OOCCH_3)$, is also manufactured commercially. The various formates are produced either directly from the metal or by reaction with an aluminum compound, for example, aluminum hydroxide [106].

uminum formates are used to impregnate textiles and to make paper water-resistant. Formates are also employed mordants in the textile industry. Because of the low toxicity and antiseptic, astringent, and basic properties of the aluminum salts, aluminum formates are used in pharmaceuticals.

12.2.4. Nickel Formate

Nickel(II) formate [*15694-70-9*] usually exists in the form of a green crystalline dihydrate

$Ni(HCOO)_2 \cdot 2\,H_2O$; it is prepared by dissolving nickel(II) hydroxide [*12054-48-7*] or nickel(II) carbonate [*16337-84-1*] in formic acid or by reaction of nickel(II) sulfate with sodium formate. Nickel(II) formate is used for the production of nickel hydrogenation catalysts.

12.2.5. Copper Formate (→ Copper Compounds , Chap. 5.1.)

Copper(II) formate [*544-19-4*] is produced by reacting copper(II) hydroxide [*20427-59-2*], copper(II) oxide [*1317-38-0*], or copper(II) carbonate [*12069-69-1*] with formic acid. It is used as an antibacterial agent for treating cellulose.

13. Toxicology and Occupational Health

Formic acid occurs naturally in small amounts in plants, animal poisons, and higher organisms. In humans, it is formed in toxic amounts during methanol poisoning. After intake of large amounts of methanol, the metabolic capacity to detoxify it to carbon dioxide is insufficient, and the high amount of formic acid that results leads to metabolic acidosis. Under physiological conditions, formic acid which is observed under these conditions in low concentration is a component of human blood and tissues, and plays an important role in the transfer of C_1 compounds during intermediate metabolism. Most of the formic acid taken up by the body is metabolized; a small portion is excreted unchanged in the urine. After oral intake, the biological half-life of formic acid is about 45 minutes in human blood plasma [107].

The most prominent toxic property of this compound is its corrosive effect upon skin and mucous membrane. When applied to the clipped skin of rabbits, formic acid causes necrosis, which heals slowly [108]. Following oral intake of aqueous dilutions of formic acid, corrosion of the oral cavity and esophagus occurs even at concentrations as low as 6 % [108–110]. The LD_{50} after oral administration is reported to be 1830 mg/kg in rats [111] and 1076 mg/kg in mice [112]. Formic acid, whose solubility in water is good, can be absorbed easily by the skin [107], [108] and causes serious local injuries upon inhalation. In rats, increasing concentrations of aqueous acid solutions, inhaled as saturated vapors (20 °C), lead to death during decreasing exposure times: a 10 % aqueous solution of formic acid is tolerated for 7 h. Eyelid closure and watery nasal discharge as a result of pronounced irritation is no longer observed after termination of the exposure period. A 25 % aqueous solution is lethal after an exposure of more than 3 h; a 50 % aqueous solution, after 30 min; and undiluted (98 %) formic acid, after only 3 min inhalation. In rats, the LC_{50} is 7.4 mg/L after inhalation of the vapor [113]. The predominant symptoms of inhalative intake are irritation of the eyes, irritation and corrosion of the nasal mucous membranes, and corneal opacity.

According to earlier investigations [108], formic acid affects the central nervous system (CNS). In rabbits, intravenous administration of buffered formic acid in the range of 0.46 – 1.25 g per kilogram of body weight leads to CNS depression; higher dosages cause convulsions. Subcutaneously administered doses of 0.8 g/kg in dogs and over 0.4 g/kg in cats lead to staggering, while a dosage of 0.4 g/kg leads to sleepiness in cats. In dogs, lethal doses administered intravenously result in tonic and clonic convulsions [108]. The mechanism of CNS effects was investigated by exposing rats to formic acid vapors (20 ppm) for 6 h a day, five days per week, for two or three weeks. This treatment caused changes in glial cell metabolism, which could be the reason for neurological effects [114].

Rats administered formic acid in concentrations of 0.5 and 1.0 % in feed and drinking water, respectively, showed a slight increase in body weight gain. In addition the weights of liver, kidneys, adrenals, and (for the lower dose group) spleen were lower than those of control animals [111]. In rats, the administration of dosages of 8.2, 10, 25, 90, 160, and 360 $mg\,kg^{-1}\,d^{-1}$ in drinking water for 2 to 27 weeks led to a lower intake of food and delay in body weight gain only in the highest dose group [115].

The influence of formic acid on biochemical parameters was investigated in rats by exposing them for 6 h a day to vapors of 20 ppm for three or eight days [116]. This treatment led to a decrease in the concentration of glutathione in the liver and kidneys.

Few experimental results are available regarding the mutagenic effects of formic acid. The substance was weakly mutagenic when tested on *Escherichia coli* [117]. A mutagenic effect was also observed on the germ cells of drosophila [118]. However, no primary genotoxic mechanism of action appears to be involved; rather, the mutagenicity is probably associated with the acidity of the compound, which leads to radical formation after inhibition of the enzyme catalase.

The TLV-TWA exposure limit (ACGIH) for formic acid is 5 ml/m^3 (9 mg/m^3); its MAK value is also 5 ml/m^3 (9 mg/m^3) (short term exposure maximum: category I). Formic acid has a PDK (former Soviet Union) of 0.5 ml/m^3 (1 mg/m^3).

14. References

1. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998.
2. M. N. Kuznetsova, A. G. Bergmann, *J. Gen. Chem. USSR (Engl. Transl.)* **26** (1956) 1497 – 1504; *Chem. Abstr.* **50** (1956) 14337.
3. *Ullmann,* 4th ed. **7** 362 – 373.
4. J. E. Prue, A. J. Read, *Trans. Faraday Soc.* **62**(1966) 1271.
5. The Texas A & M University System, Table 23-2-1-(1.1210)-d, p 1.
6. Landolt-Börnstein, *New Series,* **vol. I**Group IV , Part b, Springer Verlag, Heidelberg 1977, pp. 114, 301.
7. A. S. Coolidge, *J. Am. Chem. Soc.* **52** (1930) 1874 – 1887.
8. J. Chao, B. J. Zwolinski, *J. Phys. Chem. Ref. Data* **7** (1978) no. 1, 363 – 377.
9. W. Waring, *Chem. Rev.* **51** (1952) 171 – 183.
10. R. J. Jakobsen, Y. Mikawa, J. W. Brasch, *Spectrochim. Acta Part A:* **23 A** (1967) 2199 – 2209.
11. J. Gmehling, U. Onken, et al.: *Vapor-Liquid Equilibrium Data Collection,* **Dechema Chemistry Data Series, vol. I,** part 1 (1977), 1 a (1981), 2 d (1982), 3 + 4 (1979), 5 (1982).
12. T. Ito, F. Yoshida, *J. Chem. Eng. Data* **8** (1963) 315 – 320.
13. M. M. Gil'burd, F. B. Moin et al., *Zh. Prikl. Khim. (Leningrad)* **57** (1984) 915 – 917; *Chem. Abstr.* **100** (1964) 216, 549.
14. L. H. Horsley: "Azeotropic Data III," Adv. Chem. Ser. **n116** (1973) 66 – 69.
15. P. B. Davis, H. C. Jones, *J. Am. Chem. Soc.* **37** (1915) 1194 – 1198.
16. *Landolt-Börnstein,* 6th ed., **II/5 a,** 209.
17. VDI-Wärmeatlas, 4th ed., VDI-Verlag, Düsseldorf 1984.
18. R. W. Gallant, *Hydrocarbon Process.* **47** (1968) no. 6, 139 – 148.
19. R. R. Dreisbach, R. A. Martin, *Ind. Eng. Chem.* **41** (1949) no. 12, 2875 – 2878.
20. J. J. Jasper, *J. Phys. Chem. Ref. Data* **1** (1972) no. 4, 851.
21. J. F. Johnson, R. H. Cole, *J. Am. Chem. Soc.* **73** (1951) 4536 – 4540.
22. G. C. Sinke, *J. Phys. Chem.* **63** (1959) 2063.
23. W. J. Canady, H. M. Papée, K. J. Laidler, *Trans. Faraday Soc.* **54** (1958) 502 – 506.
24. W. Jobst, *Int. J. Heat Mass Transfer* **7** (1964) 725 – 732.
25. Th. C. Wehman, A. I. Popov, *J. Phys. Chem.* **72** (1968) 4031.
26. A. Kailan, N. H. Friedmann, *Monatsh. Chem.* **62**(1933) 284 – 316.
27. A. Kailan, F. Adler, *Monatsh. Chem.* **63** (1933) 155 – 185.
28. A. Kailan, G. Brunner, *Monatsh. Chem.* **51** (1929) 334 – 368.
29. E. N. Rostovskii, J. A. Arbuzova, *Met. Khim. Prom. Kazakhstana Nauchn. Tekhn. Sb.* **71** (1961) *Chem. Abstr.* **58** (1963) 5489.
30. K. Wagner, *Angew. Chem.* **82** (1970) no. 2, 73 – 77.
31. D. Swern, *Chem. Rev.* **45** (1949) 3 – 5.
32. H. W. Gibson, *Chem. Rev.* **69** (1969) 673 – 692.
33. W. Harder (BASF), private communication, Sept. 1984.
34. P. Mars, J. J. F. Scholten, P. Zwieterling, *Adv. Catal.* **14** (1963) 35 – 113.
35. H. N. Barham, L. W. Clark, *J. Am. Chem. Soc.* **73** (1951) 4638 – 4640.
36. S. F. Dickson, J. Bakker, A. Kitai: "Acetic Acid", in *Chemical Economics Handbook*, SRI International, Menlo Park, Calif. March 1982, revised July 1982, pp. 602.5020, 602.5021.
37. *Eur. Chem.* 1982 no. 5, 64.
38. Mitsubishi Gas Chem. KK, JP-Kokai 58 167 529, 1982.
39. *Chem. Eng.,* (1979, July 30), 17.
40. BP Chemicals Ltd. EP 60 695, 1981 (R. D. J. Mathieson).
41. IG Farben, GB 252 848, 1925 (R. Witzel).
42. Deutsche Gold- und Silber-Schneideanstalt, DE 1 035 637, 1956 (L. Hüther, M. Petzold).

43. R. Hoch, H. W. Scheeline, SRI International, Report No. 156, "Formic Acid," private report by the Process Economics Program, Aug. 1983.
44. Shell International Petroleum Maatschappij, N. V. GB 1 047 408, 1966.
45. Union Chimique Belge, US 4 216 339, 1980 (W. Couteau, J. Ramioulle).
46. BASF, EP 48 891, 1980 (F. J. Müller, W. Steiner, K. H. Ross, O. Kratzer).
47. D. M. Pond, *Chem. Eng. News* **60** (1982) no. 37, 43.
48. R. F. Schultz, *J. Am. Chem. Soc.* **61** (1939), 1443 – 1447.
49. BASF, EP 1432, 1981 (H. Hohenschutz, H. Kiefer, J. E. Schmidt).
50. BASF, EP-A 161 544, 1985 (H. Kiefer, H. Hohenschutz, J. E. Schmidt).
51. Hüls AG, DE-OS 3 428 321, 1984 (M. Zölffel).
52. Carbide & Carbon Chemicals, US 2 160 064, 1936 (J. F. Eversole).
53. I. I. Moiseev, O. A. Tagaev, N. M. Zhavoronkov et al., DE 3 220 555, 1982.
54. BASF US 4 076 594, 1978 (H. Bülow, H. Hohenschutz, W. Sachsze, J. E. Schmidt).
55. BASF, DE 2 545 658, 1975 (H. Kiefer, H. Hohenschutz, J. E. Schmidt).
56. BASF, EP 17 866, 1980 (H. Schoenmakers, D. Wolf, K. Bott, et al.).
57. J. D. Leonard, EP 5 998, 1978.
58. L. J. Kaplan, *Chem. Eng. (N.Y.)* **89** (1982) no. 14, 71 – 73.
59. A. Aguilo, Th. Horlenko, *Hydrocarbon Process.* **59** (1980) no. 11, 120 – 130.
60. *Hydrocarbon Process.* **62** (1983) no. 11, 104.
61. BASF, DE 1 147 214, 1961 (E. Germann).
62. *Eur. Chem. News* **42** (1984) May 7, 22.
63. Salzgitter Industriebau GmbH: *Technical Information Formic Acid,* 1982.
64. Salzgitter Industriebau GmbH: *Technical Information Formic Acid,* 1985.
65. Bethlehem Steel Corp., US 3 907 884, 1973 (J. B. Lynn, O. A. Homberg, A. H. Singleton).
66. A. Peltzman, *Oil Gas J.* **79** (1981) no. 46, 103 – 109.
67. Skanska Attikfabriken AB (Perstorp), GB 1 049 013, 1964 (E. Lindkvist).
68. Davy Mc Kee, DE 3 137 356, 1981 (W. Dürkoop, E.-W. Wisskirchen).
69. Commercial Solvents, US 1 606 394, 1926 (W. C. Arsem).
70. A. Behr, *Chem. Ing. Tech.* **57** (1985), no. 11, 893 –903.
71. Omsk Poly., SU 1 097 366, 1983.
72. Adogar Kogyo KK, JP-Kokai 53 46 820, 1978,
73. Mitsubishi Gas Chem. Ind., JP-Kokai 51 138 614, 1976.
74. BP Chemicals Ltd., EP-A 0095 321, 1982 (D. J. Drury, J. E. Hamlin).
75. BP Chemicals Ltd., EP-A 0126 524, 1983 (J. J. Anderson, J. E. Hamlin).
76. BP Chemicals Ltd., EP-A 0181 078, 1984 (J. J. Anderson, D. J. Drury, J. E. Hamlin, A. G. Kent).
77. W. Hunsmann, K. H. Simmrock, *Chem. Ing. Tech.* **38** (1966) no. 10, 1053 – 1059.
78. L. Berg, An-I Yeh, US 4 642 166, 1986.
79. P. Eaglesfield, B. K. Kelly, J. F. Short, *Ind. Chem.* **29** (1953) April, 147 – 151; **29** (1953) June, 243 – 250.
80. Hüls AG, DE-OS 3 428 319, 1984 (F. v. Praun, H.-U. Hög, G. Bub, M. Zölffel).
81. National Association of Corrosion Engineers Publication 5A 174, July 1974, pp. 13 – 18.
82. F. F. Berg: *Korrosionsschaubilder, Corrosion Diagrams,* VDI-Verlag, Düsseldorf 1965.
83. D. Behrens (ed.): *DECHEMA-Werkstoff-Tabelle,* Frankfurt, 1988.
84. OECD-Guidelines for Testing of Chemicals (1981), Paris.
85. B. F. Dowden, H. J. Bennett, *J. Water Pollut. Control Fed.* **37** (1965) 1308 – 1316.
86. G. Bringmann, R. Kühn, *Gesund. Ing.* **80** (1959) 115 – 120.
87. *Katalog wassergefährdender Stoffe,* Gemeinsames Ministerialblatt Ausgabe A, 36. Jahrgang, Nr. 11, Bonn April1985, pp. 173 – 186.
88. T. A. Luft: *Gemeinsames Ministerialblatt* Ausgabe A, 37. Jahrgang, Nr. 7, Bonn Febr.1986, pp. 93 – 143.
89. BASF Data sheet D 027 d, e (Sept. 1986).
90. *Food Chemicals Codex,* 3rd. ed., National Academy Press, Washington D.C., 1981.
91. International Standard ISO 731/I-VII, 1977.
92. R. M. Speights, A. J. Barnard, Jr. in F. D. Snell, L. S. Ettre (eds.): Formic Acid, *Encyclopedia of Industrial Chemical Analysis,* **vol. 13,** Interscience Publishers, New York 1971, pp. 117 – 138.
93. BASF DIN-Sicherheitsdatenblatt für Ameisensäure 85 %, 90 %, 95 %, 99 % (1986).
94. Verordnung über gefährliche Stoffe (Gefahrstoffverordnung-Gef Stoff V), 26. 08. 86 (BGBl. I S. 1470), Carl Heymans Verlag KG, Köln 1986.
95. P. Mc Donald: *The Biochemistry of Silage,* J. Wiley & Sons, New York 1981.

96. M. P. Czaikowski, A. R. Bayne, *Hydrocarbon Process.* **59** (1980) no. 11, 103 – 106.
97. C. A. M. Hough, K. J. Parker, A. J. Vlitos (eds.): *Developments in Sweeteners-1,* Applied Science Publ., London 1979.
98. W. Kaminsky, *Chem. Ing. Tech.* **55** (1983) no. 9, 667 – 683.
99. T. Ikarashi, *CEER Chem. Econ. Eng. Rev.* **12** (1980) no. 8, 31 – 34.
100. M. Röper, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) no. 11, 506 – 511.
101. Mitsubishi Gas Chemical Ind., DE 2 753 634, 1978 (M. Yoneoka, M. Osugi).
102. Air Products and Chemicals, EP 26 415, 1979 (J. V. Martinez de Pinillos, G. B. De La Mater, H. Ladenheim).
103. BP Chemicals Ltd., EP 60 718, 1981 (D. J. Drury, J. Pennington).
104. Borden, EP 86 615, 1982 (P. H. Lemon, J. D. Rallton, P. R. Ludlam, T. J. Reynolds).
105. BP Chemicals Ltd., DE-OS 2 653 448, 1975 (J. J. Huitson).
106. *Kirk-Othmer,* **2,** 202 – 204
107. G. Malorny, *Z. Ernährungswiss.* **9** (1969) no. 4, 340 – 348.
108. W. F. von Oettingen., *Arch. Ind. Health* **20** (1959) 517 – 531.
109. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* Thieme-Verlag, Stuttgart – New York 1980, p. 273.
110. K. E. von Mühlendahl et al., *Arch. Toxicol.* **39**(1978) 299 – 314.
111. A. Sporn et al., Igenia **11** (1962) 507 – 515, cited in: "Evaluation of the Health Aspects of Formic Acid, Sodium Formate and Ethyl Formate as Food Ingredients," Life Sci. Res. Office, Fed. Am. Soc. Exp. Biol., Bethesda, MD., 1976, NTIS No. 03 266 282.
112. C. Malorny, *Z. Ernährungswiss.* **9** (1969) no. 4, 332 – 339.
113. BASF Aktiengesellschaft, Ludwigshafen, unpublished results, 1979 – 1981.
114. H. Savolainen et al., *Acta Pharmacol. Toxicol.* **47** (1980) 239 – 240.
115. T. Sollmann, *J. Pharmacol. Exp. Ther.* **16** (1921), 463 – 474.
116. A. Zitting et al., *Res. Commun. Chem. Pathol. Pharmacol.* **27** (1980) no. 1, 157 – 162.
117. D. M. Demerec et al., *Am. Nat.* **85** (1951) no. 821, 119 – 136.
118. B. F. A. Stumm-Tegethoff, *Theor. Appl. Genet.* **39**(1969) 330 – 334.

Foundry Technology

Heiner Träger, Chemetall GmbH, Frankfurt, Federal Republic of Germany

Andreas Bührig-Polaczek, Institut für Giessereikunde, Montanuniversität Leoben, Leoben, Austria

1. Introduction 1
2. The Casting Process 2
2.1. Fundamentals 2
2.2. Design of Castings 4
2.3. Casting in Nonpermanent Molds . . 5
2.3.1. Molding Materials, Binders, Additives, and Coatings 5
2.3.2. Molding Sand Preparation 6
2.3.3. Clay Bonded Sand Molds 6
2.3.4. Chemically Bonded Sand Molds . . . 6
2.4. Casting in Permanent Molds 10
3. Melting, Pouring, and Solidification of Metals 11
3.1. Melting, Holding, and Pouring . . . 11
3.2. Solidification Morphology 11
4. Aftertreatment of Castings 12
4.1. Cleaning of Castings (Fettling) . . . 12
4.2. Heat Treatment 12
5. Control of the Casting Process . . . 13
5.1. Quality Assurance 13
5.2. Protection of the Workplace and the Environment 13
6. References 13

1. Introduction

Casting is the formation of a solid body of predetermined size and shape (a casting) from a material without shape (a melt) by a process of cohesion (solidification in a mold). Foundry technology is the art and science of the processes and materials used in manufacturing castings from metals and alloys. It deals with the manufacture of molds and patterns, melting practice, metal treatment, and pouring methods, together with the processes for finishing castings (fettling) and the equipment used.

Most metallic materials may be formed by the casting process. To secure the properties required of the castings it is of crucial importance that the design is suitable for the casting process and that the metal is correctly chosen. The process is suitable for mass production of standard parts, as well as for small series and individual components.

History. The history of metal working and the extraction of metals from their ores goes back many thousands of years, although the earliest beginnings cannot be dated precisely. The oldest known castings date from about 3000 B.C. (copper alloys, lead, and tin bronzes). Around 1500 B.C. the technique of reducing iron ores was understood, and the lumps of iron so obtained were forged into artifacts. The casting of iron began in the late Middle Ages, steel casting became an industrial process in the 19th century, and now, in the 20th century, light metal casting is becoming of increasing importance [10–13].

There is continuous development of new alloys and casting techniques on the one hand and new mold-making methods and materials on the other. There are ever more stringent requirements, especially concerning mechanical properties of the castings, their surface quality, torsion resistance, and machinability, as well as lightweight construction, abrasion resistance, and suitability for series production. Tolerances become ever narrower (mass and physical properties), and quality control becomes increasingly elaborate and must include statistical procedures. Modern computer technology is now used in the entire foundry process, in design and planning, and in process and quality control. The impetus towards the development of new technology now comes not only from the need to compete with other materials and processes; occupational health and environmental aspects must also be considered.

2. The Casting Process

2.1. Fundamentals

Basically, foundry processes can be divided into two categories: casting of semi-finished products and casting of components (Fig. 1). Casting

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_035

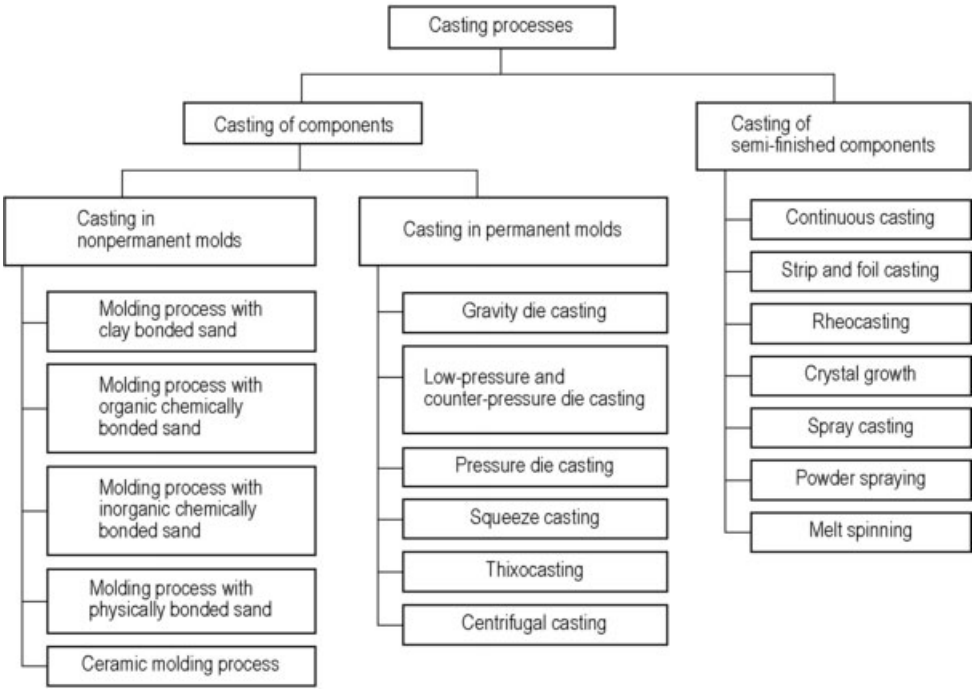

Figure 1. Overview of casting processes

of semi-finished products can be further subdivided into continuous casting, ingot casting, strip and foil casting, melt spinning, rheocasting, crystal growth, spray casting, and powder spraying.

The most important process in industry is continuous casting. Continuous casting is the production of symmetrical solid or hollow profiles (billets, rods, tubes) by pouring the liquid metal into a water-cooled mold which determines its shape and from which it is extracted in a continuous or semi-continuous process. The profiles of continuously cast parts are usually given their final shape in subsequent forming processes. However, it is also possible to produce profiles directly. Strip, wire, and foil castings are also normally produced by continuous casting.

Component casting is used for the production of metal products with various specific contours and wall thicknesses. They are cast directly, as a liquid melt, into molds consisting of one or more parts. Cast components largely or completely meet the requirements for their final shape. The mold is a hollow form, the interior contours of which determine the shape of the casting, i.e., it is a negative of the desired casting. The processes for the casting of components can be further subdivided into nonpermanent- and permanent-mold methods.

Molding deals with the production of molds and cores; casting is concerned with the methods of filling the molds with molten metal. Molds and cores may be produced by machine or hand (individual parts).

Nonpermanent molds are generally made of sand or ceramics. After casting and cooling, the mold must be destroyed so that the casting may be removed. To produce the mold, a pattern (positive model of the casting) is needed. The molding material is applied to the pattern and set by compression or by hardening. The hollow space for the future casting is produced by pulling or lifting the pattern out of the mold. Hollow spaces within the casting itself are created by means of cores, produced in the same way as sand molds (Fig. 2).

The cores are placed onto previously prepared surfaces in the sand molds. (Different procedures used in investment casting and lost-foam

casting are described under these headings in Section 2.3.4) Patterns or coreboxes are made of wood, plastic, metal, or inorganic materials such as plaster. They can, if necessary be coated with abrasion-resistant coatings. The patterns may also be treated with release agents to facilitate their removal from the molding material.

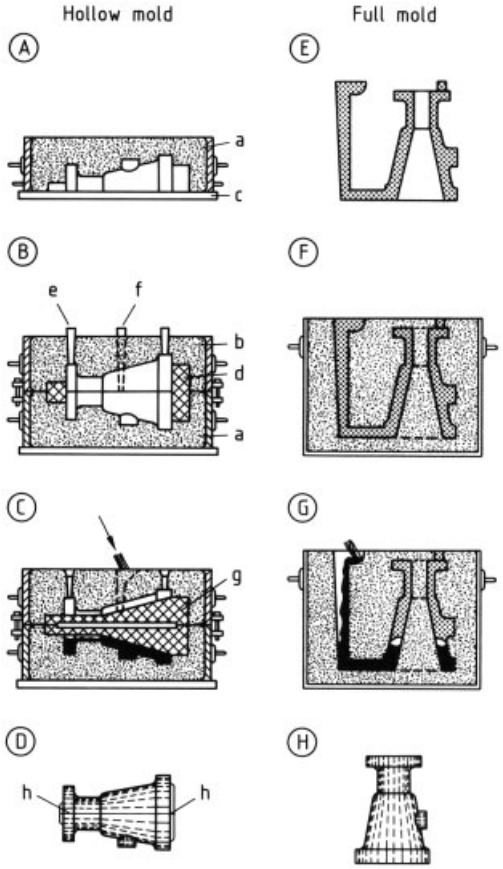

Figure 2. Molding and casting with the hollow mold process and full mold processes [14]
A) Lower half of mold; B) Upper half of mold added; C) Pouring; D) Casting with flash; E) Foamed plastic pattern; F) Pattern in a one-piece box; G) Pouring, no cores; H) Casting without flash
a) Lower box (drag); b) Upper box (cope); c) Pattern plate; d) Core print; e) Riser; f) Ingate; g) Core with core irons; h) Flash

Permanent molds (dies) consist mainly of durable, heat-resistant metallic materials, usually special steel and cast-iron alloys that have undergone specific heat treatment. In permanent molds the hollow space for the future casting is created by machining, electrical-discharge machining, or by casting and subsequent machining. Permanent molds have the advantage that they can be used repeatedly from thousands to hundreds of thousands of times or more, depending on the cast alloy. Permanent molds often consist of several segments and have core pullers to allow even complex castings to be removed.

The casting processes, the cast alloys, and the weight limits which generally apply today are illustrated in Table 1. Castings range from only a few grams to several hundred tonnes in weight. Typical fields of application for castings are: the automobile industry; train, airplane, and ship building; the chemical industry; industrial furnaces; iron and steel works; foundries; energy production; the electrotechnical industry; precision engineering; transmission systems; construction and housing; mechanical engineering; and medicine [1], [8], [11], [15], [18], [34], [59–65], [105–111], [121–127].

2.2. Design of Castings

The design of a casting must be in accord with accepted standards and material information sheets. Choice of material, shape, mass, and other specifications, such as later surface finishing, are given in the technical drawing. This information is needed for designing the pattern, which must take into account the degree of shrinkage (see Section 3.2), degrees of taper, and machining allowances, as well as the position and type of cores and the arrangement of loose pieces (of pattern) necessary to form undercut parts of the mold. The pattern drawing also shows its position in the mold and the way in which it is usually divided into two parts, one for the upper molding box and one for the lower. At the same time the gating and, if required, feeding systems (see Section 3.2) are determined (Fig. 3).

In addition to the choice of suitable material and molding process, the design must be suitable for the actual process of casting so that both the manufacture and the desired physical and mechanical properties of the cast product are achieved. For this reason, the founder should participate in the design of the casting. It must not only be designed correctly as regards construction material and service requirements, but also be dimensionally accurate, clean, and able to pass the required tests.

Table 1. Range of application of the most common nonpermanent- and permanent-mold casting processes

Casting process	Casting alloys	Weight of casting	Series size
Hand molding	all cast alloys (mainly gray iron, ductile iron, steel, bronze)	up to over 100 t (limited only by melting and transport capacities)	individual parts, small series
Machine molding	all cast alloys (mainly gray iron, ductile iron, Al alloys)	1 kg to several tonnes	small and large series
Shell molding process	all cast alloys (mainly gray iron, ductile iron)	1 – 150 kg	medium and large series
Lost foam process	gray iron, ductile iron, steel, Al alloys	< 1 kg to several tonnes	individual parts, small and large series
Investment and precision casting	all cast alloys	1 g to100 kg	individual parts, small and large series
Gravity die casting	Al, Mg, Cu, Zn alloys; gray iron, ductile iron	< 1 kg to100 kg	medium and large series
Low- and counterpressure die casting	Al, Mg, Cu alloys	< 1 kg to 70 kg	medium and large series
Pressure die casting	Al-, Mg-, Cu-, Zn-,Sn-, and Pb-based pressure die casting alloys	several grams to 60 kg, depending on the alloy	medium and large series

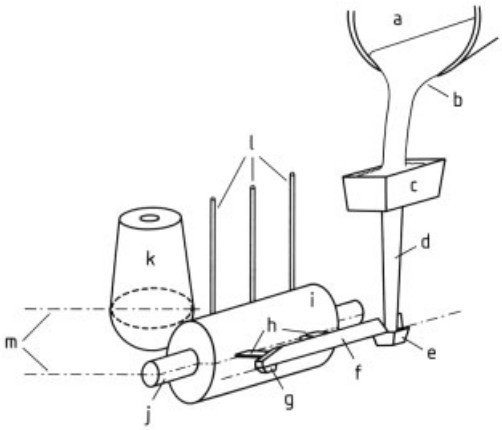

Figure 3. Gating and feeding system for a hollow cylinder [2] (1981)
a) Ladle; b) Pouring lip; c) Pouring basin; d) Downsprue; e) Sprue base; f) Runner; g) Skim bob; h) Gate; i) Casting; j) Core; k) Feeder; l) Air vents; m) Parting line: upper box (cope), lower box (drag)

Modern near-net-shape production of components requires close cooperation between designer and foundryman to reduce costs, aftertreatment, and weight. In addition, all the possibilities offered by the casting process must be taken advantage of.

Ever-increasing demands make the use of modern simulation methods an indispensable tool for mastering the complexity required. The prediction of casting processes by numerical simulation has become indispensible. Computer simulations of the complex casting process often allow invisible problem areas to be located both qualitatively and quantitatively. Variations of parameters and the resulting effects can be investigated by computer, thus allowing purposeful optimization (Fig. 4). Software is available for linking various program packages (including CAD systems) and enabling direct communication between the foundryman and the designer [1], [2] (1981), [9], [12], [16], [59], [100], [101], [103], [104], [110], [111], [117], [133].

2.3. Casting in Nonpermanent Molds

2.3.1. Molding Materials, Binders, Additives, and Coatings

Nonpermanent molds and cores consist of the basic particulate material, a binder system, and additives. Additionally, mold coatings may be necessary.

The mold and the molding material have a decisive influence on the quality of the casting, for example, its dimensional accuracy, its surface quality, and its microstructure. The main demands made on the molding material are:

- Good workability
- Sufficient strength after shaping or during pouring and solidification
- Good disintegration characteristics after casting
- Highest possible accuracy of detail
- High refractoriness
- Sufficient gas permeability

– Negligeable undesirable interaction with the melt
– Uncomplicated reusability

The basic materials are granular substances of adequate refractoriness (washed, classified, and dried silica sand of the highest purity, zircon sand, chromite, olivine, corundum, magnesite, sillimanite, mullite, or chamotte). Binder systems may be inorganic or organic, natural or synthetic (Table 2). Additives to the molding materials improve their properties: flowability, compactability, bond improvement, prevention of sand expansion defects, breakdown properties. Common additives are sawdust, iron oxide, sugars, and glycol, among many others.

Some molding sand additives produce lustrous carbon, e.g., coal dust and hydrocarbons. These are vaporized at high temperatures and are simultaneously decomposed (cracked). The molding sand particles become coated with lustrous carbon and then are not easily wet by molten metal.

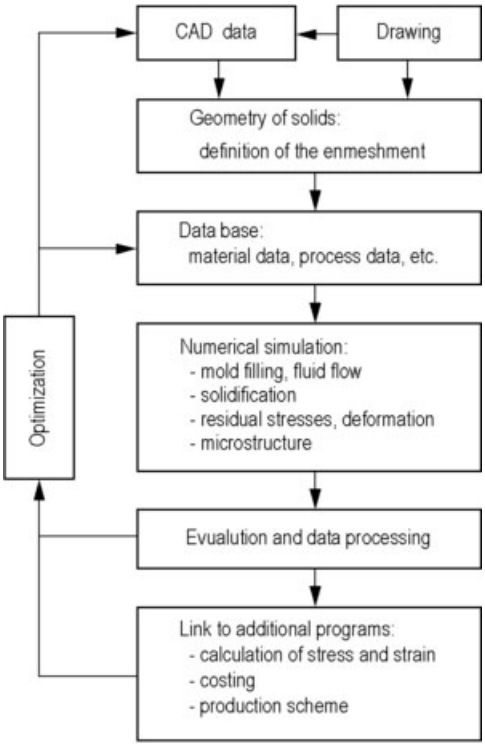

Figure 4. Design optimization by numerical simulation of casting processes

Table 2. Sand Binders

	Natural	Synthetic
Inorganic (in combination with water)	clays bentonites (mont-morillonite)	sodium silicate cement sulfite lye phosphate binders plaster
Organic (in combination with hardeners, catalysts and/or heat)	oil binders/ core oils sugar	condensation resins: aminoplasts (UF resins) phenoplasts (PF resins) furan resins reaction resins epoxy resins polyester resins polyurethane resins

Ready-to-use mold coatings are dispersions of finely ground minerals or refractory materials, and/or powdered coke in water or organic solvents, with a binder system and other additives to improve their application properties. They are applied to the surfaces of molds or cores, or if necessary to the surface of polystyrene patterns, the aims being to reduce metal penetration into the sand, to improve the casting surface finish, to facilitate separation of the casting from the molding sand, and to reduce fettling costs [2] (1981), [12], [15], [17] , [19–24], [25], [31], [32], [101], [128], [130], [131].

2.3.2. Molding Sand Preparation

Molding sand in a condition ready for use is produced by intensive mixing of the individual components or the use of pre-mixes with additives.

Usually, sand molds and, if possible, cores are made from used, reconstituted sand (old sand). The new sand supplied to the system should be used only for core making and for critical parts of the mold (facing sand). The amount of new sand should not be appreciably greater than the quantity lost in sand preparation. A mechanized foundry requires 5 – 12 t molding sand for every tonne of cast iron poured. The new sand also serves to maintain the sieve grading of the system sand.

Routine analytical control of the molding sand, including the testing of specimen cores, is at least as important as the quality control of the molten metal. Scrap rate due to mold defects can considerably exceed that due to metallur-

gical deficiencies; pinholes are often due to a combination of the two [20], [24], [26–29].

2.3.3. Clay Bonded Sand Molds

Natural sand (i.e., a naturally occurring sand – clay mixture) is now seldom used in foundries as a molding material. The more common synthetic sands are produced by mixing together the pure individual ingredients. The clay component is usually a bentonite (montmorillonite); natural sodium bentonites are preferred, or bentonites activated by ion exchange with sodium carbonate. Synthetic clay-bonded molding sand consists mainly of sand, bentonite, and a lustrous-carbon former (and, in the form of reclaimed sand, is usually already coated with lustrous carbon). It contains the amount of water necessary to give a bond (green sand). The water content is not decreased by any subsequent treatment of the mold. The quality of the molding sand is very dependent on the amount of fines present (arising from sand abrasion, active and dead-burnt bentonite, and carbonized lustrous-carbon former), and lastly on temperature, because the water content must be maintained at a low but precisely controlled level.

Green sand molds are produced by compaction of the prepared sand by, e.g., ramming, slinging, jolting, or squeezing. Compaction can also be achieved by gas pressure (explosion and air-impulse processes) or by an air stream.

Green sand molds that have been oven-baked or at least air-dried on the surface have much greater strength than undried molds and may be used for production of heavy castings, especially of steel, provided in this case no lustrous carbon is present.

Molds made using chamotte are used for heavy, thick-walled steel castings. Loam molds are now seldom used (template molding) [20], [30], [33], [37–43].

2.3.4. Chemically Bonded Sand Molds

Chemical binders can give well bonded molds with relative ease. The development of these products has made it possible to produce hard molds, often using simple methods. Good casting yields are obtained even without feeders (spheroidal graphite cast iron) and dimensional accuracy is also good. However, special care must be taken to protect the environment when organic binders are used [44].

Inorganic Binder Systems. Sodium silicate solution in the presence of an ester acts as a binder in a two-stage reaction (hydrolysis of the ester followed by gelling of the silicate at the lower pH). The bond strength and reaction velocity are dependent on the $SiO_2 : Na_2O$ ratio of the sodium silicate and the type and amount of ester used. The disadvantages are poor breakdown properties and difficult reuse of old sand. Better results are obtained with aluminum phosphate binders, which can be hardened with warm air or by adding strongly basic oxides, such as magnesium oxide powder. A typical application is mold manufacture with olivine sand, which is a self-hardener.

Other, less widely used sodium silicate systems use hardeners such as silicon–ferrosilicon powders (Nishiyama process, now obsolete because of the danger of hydrogen evolution), dicalcium silicate, or organic esters.

The importance of sodium silicate as a binder was first established by the hardening process involving gassing of the sand mixture with carbon dioxide. This process became of particular importance for core making on account of its rapid turn-around time (10 – 25 s), and it is environmentally extremely beneficial. Additives are used which reduce moisture pick-up on storage of the molds or cores, and also improve breakdown properties.

Cement is used as a sand binder in combination with hardening accelerators (e.g., molasses). This process is normally restricted to large, simply shaped molds. Dust problems render the process unsuitable for core making. In the fluidized sand process, the molding sand is prepared in a pumpable condition by the addition of surfactants and the incorporation of air bubbles [20], [45–48].

Heat Hardening Organic Binder Systems. In these systems molds and cores are produced by heat hardening of sands bonded with synthetic resins. They include the thermo-shock, hot box, warm box, and shell (Croning) processes, but not those processes that seek merely to improve hardening or to produce other effects. The

equipment used (pattern plates and core boxes) must be made of metal. The use of these processes is widespread; they give great dimensional accuracy and good breakdown properties.

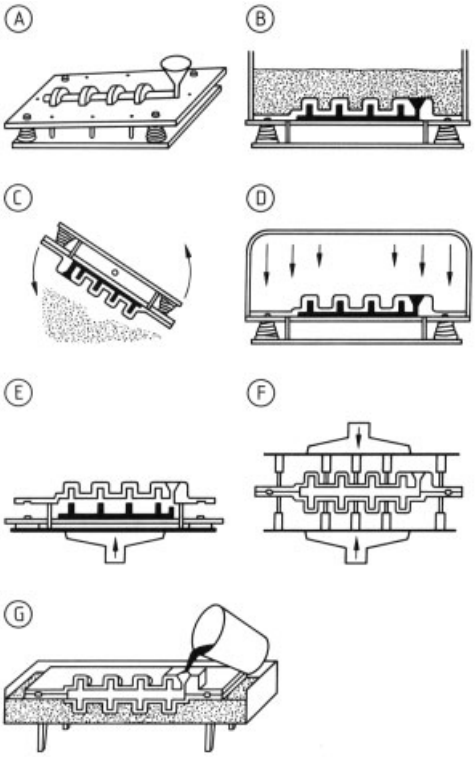

Figure 5. Casting production by the shell process (Croning process) [11]
A) Metal pattern equipment heated to 250 °C; B) Dump box filled with coated sand; C) Pattern inverted to remove excess sand; D) Hardening of the shell mold; E) Stripping of the shell mold; F) Joining the two halves with adhesive; G) Pouring the mold on a bed of sand (or freestanding or in backing sand with mold vertical)

The *thermo-shock process* makes use of urea – furan – phenol resins with aqueous solutions of acid salts as catalysts. The sand mixture, which can include kaolin to improve the green sand strength, is hardened by heating in a mold at ca. 400 °C for a few minutes.

The *hot box process* uses similar liquid phenolic or furan resins, but with a latent acid catalyst, such as a solution of an ammonium salt. The hardening process, an exothermic reaction, takes place in 20 – 120 s in a core box preheated to 180 – 340 °C. The mixed sand has a limited bench life. There is no need for further hardening by oven baking.

The *warm box process* is an extension of the hot box process to lower working temperatures.

The *shell process* (Fig. 5) makes use of sand precoated with synthetic resin. This sand is dry, free flowing, and has a practically unlimited storage life. The resin is a condensation polymer not unlike that used in the hot box and warm box processes. The sand is coated with novolac resin using either the hot melt or cold solvent process. The pattern is preheated to ca. 250 °C and covered with the precoated sand that becomes hard where it is in contact with the hot surface, the depth of hardening (4 – 10 mm) being determined by the dwell time. The shell mold that is produced from the pattern plate in this way is bonded by an adhesive to a mold, usually another shell mold. Casting then takes place, either with the completed mold freestanding or embedded to withstand the metal pressure. When cores are made by the shell process, the free-flowing sand inside the cores may be emptied out (hollow cores) and re-used, or baked through (full cores) in an oven [12], [16], [34–36], [42], [45], [48].

Cold setting organic binders are systems with liquid hardeners and catalysts (no bake process), which are particularly suitable for making molds and larger cores, and systems with rapid-acting gas hardening, which are usually used for core making (cold box process).

No Bake Process. The mixed sands have a limited life, and distinction should be made between the time available for working with the sand (bench life) and the optimum time that should elapse before separating the hardened sand from the pattern or core box (strip time).

The cold setting oil process makes use of unsaturated vegetable oils (drying oils) in conjunction with organometallic accelerators (metallic dryers such as cobalt or lead naphthenate) and oxidizing agents such as potassium permanganate or sodium perborate. Hardening takes place at room temperature over a long period. Oven drying is used to complete the process. Alkyd resins may be used to supplement or replace the drying oils.

Furan resins are condensates between furfuryl alcohol and formaldehyde (urea modified). The hardening catalysts are acids such as phosphoric and sulfuric acids, or, particularly for re-

claimed sand systems, sulfonic acids. The addition of sulfuric acid increases setting speed.

Phenolic resins are hardened by sulfonic acids. These have only a limited storage life but are usually nitrogen-free and thermally more stable when hardened through. Other types of phenolic resins are formed from highly alkaline solutions of resoles by addition of an ester, which hydrolyzes in the alkaline medium and forms the acid necessary for gelling to take place. A combination of furan and phenolic resins is also possible.

The first urethane binder systems consisted of isocyanate and an alkyd resin, which are hardened by atmospheric oxygen in the presence of metal-ion catalysts. The more modern practice is to use a phenolic-based polyol, which reacts with a polymeric isocyanate in the presence of a basic catalyst [45], [49–51].

Gas Hardening Processes. In the cold box process, a mixed sand, usually including a multicomponent binder, is charged into a core box and then swept with a gaseous reagent or catalyst (with or without a carrier gas), thereby causing hardening of the sand, or speeding up a hardening process. The equipment must be sealed as the carrier gas (air) contains added or generated harmful gases or vapors. Heat may also accelerate the hardening process and can be provided either externally or by an exothermic reaction. Details of the common processes are given in Table 3 [20], [42], [45], [47], [49], [50], [52–54].

Physically Bonded Molding Materials. In the frozen mold process, moist sand without a binder is hardened by cooling, e.g., with liquid nitrogen.

In the magnetic mold process, a pattern made of foamed plastic (e.g., polystyrene) is surrounded by a ferromagnetic molding material (e.g., iron powder). During pouring and solidification of the metal, the molding material is held firmly in place by a strong applied magnetic field; the pattern is vaporized by the molten metal.

Another version of the full mold process with foamed plastic patterns uses dry sand. This is held in place merely by the gas pressure produced when the molten metal causes the pattern material to vaporize. The process may be improved by applying reduced pressure to the mold and by covering exposed sand surfaces as far as possible, e.g., with a plastic film.

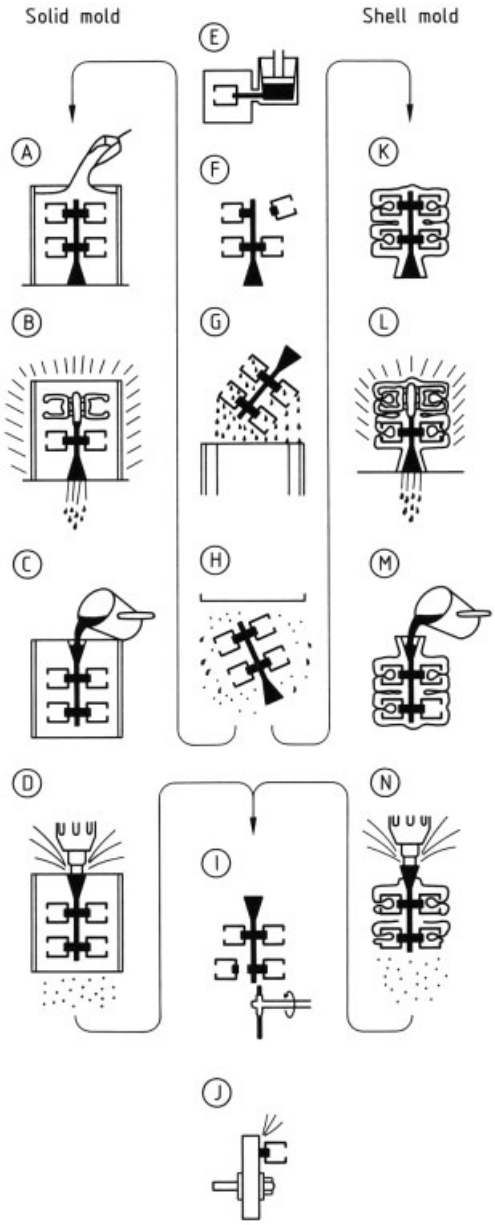

Figure 6. Schematic diagram of the investment process [11]
A) Back filling; B) Melting out; C) Casting; D) Knockout ; E) Pattern manufacture; F) Assembly; G) Dipping; H) Enrobing; I) Separation; J) Grinding; K) Shell formation by multiple dipping and enrobing; L) Melting out; M) Casting; N) Knockout

Table 3. Common gas hardening (cold box) processes

	Gassing medium					
	Tertiary amine ∗		SO_2 (+N_2/CO_2)		Methyl	Acetal ∗∗
Resin component	phenolic resin polyisocyanate	furan resin	epoxy resin	acrylic resin	resole (highly alkaline)	polyphenol
Other component		peroxide	peroxide	peroxide (acid)	(ester)	acid
Polymer	polyurethane	furan resin	epoxy resin	acrylic resin	phenolic resin	phenolic resin

∗ Dimethylethylamine, dimethylisopropylamine, or triethylamine.
∗∗ For example, dimethoxymethane

In the vacuum process a film of thermoplastic material is placed over a special pattern containing many holes through which a vacuum is applied. The film is thereby stretched to fit the contours of the pattern exactly. A coating is applied, and, after inverting the pattern and placing it in a molding box, the molding box is filled with dry sand, covered with a second plastic film and a vacuum applied to the sand, enabling the pattern to be removed. A second molding box with its sand mold is then joined to the first in the usual way. Cores may be included. During the entire process up to the solidification of the metal, the vacuum must be maintained [20], [34], [45], [55, pp. 441 – 457], [56].

Precision Casting Processes. Molding shells are made by pouring a high-quality molding mixture onto pattern plates. In the Shaw process, a slurry of powdered refractories and ethyl silicate is used. This solidifies and is flamed off; the shells are then cemented together and heated to 800 – 1000 °C. In the Unicast process, the shells are dipped into a hardening bath instead of flaming off. Other variations are the Schott and Ceramcast processes [57].

Calcined gypsum is a suitable medium for both light and heavy metal casting. Combined water may be driven off by baking or calcining, giving modifications of the process such as the Capaco, Antioch, or foamed plaster processes [58].

Investment Casting. In this process, previously prepared one-piece wax patterns are coated or enrobed with the molding material, either one at a time or in clusters (Fig. 6). The patterns may be made of urea or of vaporizable foamed plastic. The running system (channels for the molten metal) is always part of the pattern. The process of covering the pattern is carried out by repeated application of coatings by dipping into a slurry or slurries consisting of refractory – binder systems, with a drying stage after each dip. The pattern material is then removed by melting, dissolving, or vaporizing, as appropriate; the mold is strongly heated and casting is carried out while the mold is still hot. Thin-walled molds require back-filling; thick shells do not [16], [19], [20], [42].

2.4. Casting in Permanent Molds

Permanent-mold casting is mainly used for low-melting nonferrous alloys (aluminum, zinc, magnesium, and copper alloys). In Germany, more than half of all nonferrous alloy castings are produced solely by pressure die casting. The main permanent-mold casting methods are gravity die casting, low- and counter-pressure die casting, and pressure die casting. The centrifugal casting process is mainly used for the production of spun iron pipes.

In *gravity die casting*, the melt enters the metallic permanent mold through the ingate simply by force of gravity. In *low-pressure die casting*, the melt is transferred from the furnace into the permanent mold above it through a pipe electromagnetically or by means of a slight overpressure of about 0.2 – 0.8 bar. In *counterpressure die casting*, not only the furnace but also the permanent mold are subjected to a pressure of up to 10 bar. Here too, mold-filling takes place due to a slightly higher pressure build-up in the furnace than in the mold. These two pressure-assisted methods allow very smooth, homogeneous filling of the mold. The permanent mold is usually treated with a ceramic coating to protect it from the melt and to regulate heat loss. This layer must be replaced at regular intervals. Maintaining the heat balance usually also requires the thermal energy brought into the die by the melt to be removed by cooling. Proper die design and well-planned location of the cooling channels allow

for controlled solidification in the direction of the gatings or the filling pipe and thus ensure a sound casting. In order reproduce undercuts and as-cast holes, the dies can be divided into several sections and be furnished with core pullers. It is also possible to place sand cores in the mold in order to form hollow spaces within the casting.

In *pressure die casting*, the molten metal is filled into a casting chamber and then pressed into the metallic permanent mold at high pressure. Because the filling time is very brief (20 – 500 ms, depending on wall thickness and size of the casting) and due to the pressure which the molten metal is subjected to even in thin cross sections, it is possible to attain a wall thickness of under 1mm and highly exact reproduction of the desired contours. Today, except in rare cases, hot and cold chamber die-casting machines are used. Hot-chamber machines are well suited to low-melting metals. Heavy metals such as zinc, lead, and tin alloys, as well as magnesium alloys, are cast by this method. The casting chamber sits in a bath of molten metal and takes on its temperature (this explains the term ″hot chamber″). Cold-chamber machines are suitable for higher-melting metals such as aluminum, magnesium, and copper alloys. In the cold-chamber machine, the metal is kept hot in a separate furnace next to the machine and then filled into the pressure chamber by means of a ladle or a dosing system. (The term cold chamber refers to the fact that it is ″cold″ compared to the melt in the furnace). The two halves of the permanent mold have water or oil channels to attain the optimal working temperature by regulating the heat balance in the mold. Properly dimensioned and located channels can influence localized heat distribution, a factor of great importance, especially for complex castings.

To achieve higher quality, mainly for aluminum components, the thixocasting and the squeeze casting processes have been developed. In thixocasting semi-solid slugs with a melt content of about 50 % are processed by the pressure die casting method. The behavior of the semi-solid material during mold filling differs from that of a liquid melt. This results in a better quality casting and makes it possible to use high-strength alloys which normally are not suitable for the pressure die casting method. In squeeze casting, the melt solidifies under a very high pressure of 1000 bar and more. Slow filling of the mold and high melt quality are important. Castings made by this method are of very high quality [4], [34], [59–61], [63–65], [111], [116], [118], [121–123].

3. Melting, Pouring, and Solidification of Metals

3.1. Melting, Holding, and Pouring

Melting shops have various methods of producing and casting molten metal. On average, their costs amount to about one-third of total foundry costs. Therefore, the melting shop must be optimally coordinated with the rest of the plant. Flexibility in the production of the melt and in carrying out necessary metallurgical procedures, as well as the ability to provide molten metal over longer periods of time, require well-adapted production units (furnaces specifically adapted to the alloys, the lot sizes, the cycle time, etc.). Another essential aspect is that the quality of the cast part depends decisively on the melting technique. The most widely used industrial furnaces are currently the cupola furnaces, used by iron foundries to produce molten iron, and induction and resistance furnaces, which are used especially for nonferrous metals but also for cast-iron materials.

The cupola furnace is based on the principle of the vertical shaft furnace. The charge, which is loaded in layers, undergoes continuous melting in the cupola furnace. The charge mixture consists mainly of the fuel (usually coke, but oil or natural gas firing is also possible), the metal charge (pig iron, scrap and recycled iron) and (slag-forming) additives. The coke provides the energy, that is, the energy required for the melting process is produced within the cupola furnace. As a result of the continuous melting process, the metal charge gradually slides into the furnace shaft where it is heated by the rising gases. In the adjacent melting zone, the iron liquefies and, after superheating, collects on the floor of the furnace before leaving the furnace through the tapping hole. The intense contact between the molten iron and the coke causes carbon pick-up in the liquid iron.

In induction melting, heat is produced directly within the metal charge by the magnetic field of a primary coil. Industry uses

mainly medium-frequency and, occasionally, low-frequency crucible furnaces. Both types of furnace are very economical compared to other melting units. The metal charge is placed directly in the crucible and it is possible to begin the melting process even while the metal is still cold. All common cast iron materials can be melted in the induction furnace. The intensive movement caused in the molten material by the eddy current leads to a high melt-down rate for the metal charge and the alloying metals, and produces high regularity and homogeneity in melt composition.

In the resistance crucible furnace, heat is produced in electric heating elements and transferred to the metal by radiation, convection, or conduction. The advantages of the resistance furnace are that it can be used universally and requires minimal maintenance. Due to its low efficiency, it is now widely being replaced by small medium-frequency induction furnaces. It is, however, still in use as a holding furnace for lightweight metals in pressure casting processes [6], [7], [14], [66–69], [101], [102], [119–121].

3.2. Solidification Morphology

The solidification process of the metallic materials plays an important role both in determining the casting procedure and in the properties of the casting. During solidification, metals form crystals. The actual casting does not consist of a single ideal crystal but of many crystalline grains. The crystals form as the melt solidifies, developing in random fashion in various areas of the melt and increasing in size until they meet and limit each other's growth. The size, growth, and position of the crystals which make up the casting material are referred to as the microstructure of the casting. In some alloys, such as cast iron, additional microstructural transformations occur on cooling to room temperature.

In practice, inoculations and grain-refinement treatments are carried out to influence microstructure formation. These change the nucleation conditions in the melt or the growth conditions of individual crystals. Regulating heat transfer from the solidifying metal to the mold material can also modify the microstructure of the casting.

Another important property of metals that has a major influence on the casting process is the volume of contraction during cooling of the melt, solidification, and final cooling of the solidified casting (Fig. 7). The feeding, which is an essential aspect of the casting process, is conceived to balance out the differences in volume which arise during cooling and solidification of the melt. The differences in volume depend on the alloys involved. Gray cast iron and nodular cast iron increase in volume during solidification. Whereas only liquid and solidification shrinkage are of importance for casting and feeding technology, pattern and mold construction are decisively influenced by volume shrinkage in the solid state. The extent of this shrinkage also depends on the alloy. The dimensions of pattern and mold must be increased such that after shrinkage, the required size is attained at room temperature [26], [113–115].

4. Aftertreatment of Castings

4.1. Cleaning of Castings (Fettling)

In nonpermanent-mold casting, separation of the casting from the mold is carried out by breaking down the bond of the molding material with vibrating grates, tumbling mills, heavy hammers, or water jets. The surfaces are then cleaned and descaled by shot or sand blastin, or tumbling, and ground smooth. In permanent-mold casting, the surface is normally not treated.

The gating and running systems are removed by knocking off, abrasive cutting, or flame cutting. The removal of surface unevenness and defects, especially flash and fins, is carried out by grinding with abrasive wheels; this is often expensive, as these operations even today are very labor intensive [55, pp. 458 – 461], [70–89], [129].

4.2. Heat Treatment

Heat treatment is carried out, if required, to modify the metallic structure of the whole casting (e.g., annealing, tempering, hardening, hardening plus tempering, age hardening) or the surface (e.g., surface hardening or nitriding by gas or salt bath treatment). Stress-relief annealing removes shrinkage stresses (cooling stresses) without altering the structure of the metal [90–92], [112].

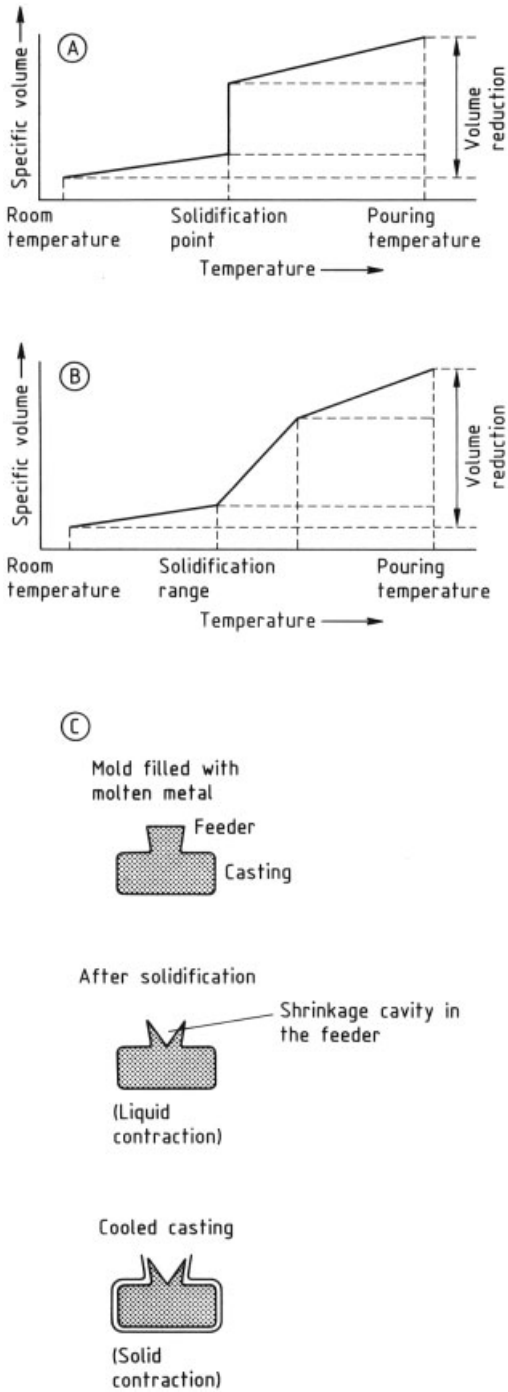

Figure 7. Schematic representation of the volume contraction of metallic materials on cooling from the molten state A) Pure metals and eutectic alloys; B) Noneutectic alloys; C) Schematic representation of the solidification of a casting

5. Control of the Casting Process

5.1. Quality Assurance

The manufacture of a casting is complete only when it meets the set requirements. Quality assurance is vital at all stages, from raw materials to casting design. Quality assurance prevents casting defects, or, failing this, it finds and rejects defective castings. Defective castings may sometimes be salvaged, e.g., by welding or heat treatment. The statistical analysis of production data and test results forms the foundation of every type of production control procedure. The test methods themselves and the associated costs are determined by the particular quality requirements (Table 4). Corrosion protection of the castings is another aspect of quality assurance [16], [93].

Table 4. Quality control of castings

Properties	Tests
Test castings	wedge test wall thickness test fracture surface, microstructure, hardness
Composition	chemical analysis
Surface	visual inspection (roughness, defects) crack tests – penetrant materials – magnetic powder
Measurements	dimensions, weights, distortion
Density,	weight
inclusions	tests using sound waves radiographic methods
Hardness	file test (machineability) point load tests (Rockwell, Brinell, etc.)
Structure	micrography physical methods
Mechanical properties	test pieces taken from the casting or
(static and dynamic)	specially cast nondestructive tests

5.2. Protection of the Workplace and the Environment

While the cast product, like any other metal product, has always been completely recyclable as scrap, the production process does leave

residues that are currently not reusable. In terms of environmental protection, foundries do need to pay specific attention particularly to dust, noise, sludge, exhaust, effluent and solid wastes produced especially by casting and melting, by the mold (treatment, regeneration, preparation, release agents, and coatings) and in the cleaning department. Used sand, which makes up about 75 % of the waste produced by foundries, has become an important cost factor, primarily due to rising disposal-site costs. Modern developments in the foundry business include waste reduction by optimization of the production process, recycling and reprocessing of materials, and, as a last resort, treatment of the waste materials before disposal to render them innoccuous [94–96], [97–99], [131], [132], [134–140].

6. References

General References

1. G. Spur, Th. Stöferle (eds.): *Handbuch der Fertigungstechnik,* vol. 1, Hanser Verlag, München-Wien 1981.
2. S. I. Karsay: *Gusseisen mit Kugelgraphit* QIT–Fer et Titane Inc., 1st ed., 1976, 2nd ed., 1981, HMS Druckhaus GmbH, Dreieich.
3. D. Altenpohl: *Aluminium von innen,* Aluminium-Verlag, Düsseldorf 1994.
4. E. Brunhuber: *Praxis der Druckgußfertigung,* 4th ed., Schiele & Schön, Berlin 1991.
5. *Jahresübersichten des Gießereiwesens*, vol. 32, Verein Deutscher Gießereifachleute, Düsseldorf 1996 (published once a year; preprints of particular articles in *Giesserei*).
6. J. H. Brunklans, F. J. Stepanek: *Industrieöfen, Bau und Betrieb,* Vulkan-Verlag, Essen 1986.
7. E. Flemming, W. Tilch: *Formstoffe und Formverfahren,* Deutscher Verlag für Grundstoffindustrie, Leipzig – Stuttgart 1993.
8. *Aluminum Casting Technology,* American Foundrymen's Society Inc., Des Plaines, Illinois, 1986.
9. P. R. Sahm, P. N. Hansen: *Numerical Simulation and Modelling of Casting and Solidification Processes for Foundry and Cast-House,* CIATF, Zürich 1984.

Specific References

10. W. W. Krysko, J. Krysko, *Giesserei* **70** (1983) no. 26, 692 – 698.
11. *Ullmann,* 4th ed., **12,** 271 – 287.
12. K. E. Höner in [1] pp. 9 – 58.
13. H. Wübbenhorst: *5000 Jahre Giessen von Metallen,* Giesserei-Verlag, Düsseldorf 1984.
14. K. Röhrig et al., *Legiertes Gusseisen,* **vol. 1,** 2. Giesserei-Verlag, Düsseldorf, 1970, 1974.
15. J. Flimm: *Spanlose Formgebung,* 4th ed., Hanser Verlag, München 1984.
16. Verein Deutscher Giessereifachleute: *Konstruieren mit Gusswerkstoffen,* Giesserei-Verlag, Düsseldorf 1966.
17. W. Wiebelhaus, in [1] pp. 240 – 245.
18. H. Wübbenhorst, in [1] pp. 228 – 238.
19. G. Matthée: *Lexikon der Fertigungstechnik und Arbeitsmaschinen,* 4th ed., **vol. 8,** vol. 9, DVA, Stuttgart 1967, 1968
20. I. Bindernagel: "Formstoffe and Formverfahren in der Giessereitechnik," *VDG-Taschenbuch 12,* Giesserei-Verlag, Düsseldorf 1983.
21. C. Ivanov: *Fonderie-Fondeur d' Aujourd'hui* **33,** (1978) no. 375, Feb., 55 – 58.
22. "Zusammensetzung und Auftragungsmöglichkeiten der Form- und Kernüberzugsstoffe," *Br. Foundryman* **76** (1983) no. 3, XII – XIV.
23. G. Vingas, *Mod. Cast.* **72** (1982) no. 3, 43 – 46.
24. G. Trinkl, in [1] pp. 490 – 499.
25. F. Hoffmann, *Giesserei* **72** (1985) no. 4, Feb. 18, 92 – 96.
26. S. Engler in [1] pp. 59 – 63.
27. C. C. Sharp, *Foundry Trade J.* **152** (1982) no. 3229, 18 – 20.
28. H. Taubmann, in [1] pp. 500 – 541.
29. G. Steinbauer, *AFS Int. Cast Met. J.* **1** (1976) no. 1, March, 8 – 10.
30. H. Sindermann, N. Damm, *Giesserei* **70** (1983) no. 22, Oct. 31, 589 – 593.
31. R. Prumbaum et al., *Giesserei* **70** (1983) no. 7, 203 – 209.
32. M. Gawlikowska, B. Ryglidzi, *Giessereiforschung* **34** (1982) no. 4, 147 – 152.
33. D. Boenisch, *Giesserei* **69** (1982) no. 21, 593 – 598;no. 23, 664 – 668.
34. A. Kopecky, R. Schamschula: *Mechanische Technologien,* 4th ed., Springer-Verlag, Wien 1977.
35. J. G. Kowalew, *Giessereitechnik* **20** (1974) no. 3, March, 88 – 91.
36. W. Tilch, *Giessereitechnik* **32** (1986) no. 4, 123 – 129.
37. R. G. Joseph, *Mod. Cast* **70** (1980) no. 1, Jan., 64 – 67.
38. G. A. Schofield, *Br. Foundryman* **72** (1979) no. 9, Sept., 209 – 220.

39. U. Kleinheyer, *Giesserei* **65** (1978) no. 20 Sept., 561 – 567.
40. J. Jablonski, L. Lewandowski, *Giessereitechnik* **17** (1971) no. 4, April, 117 – 119.
41. T. Bewick, *Foundry Trade J.* **160,** (1986) no. 3323, March 27, 235 – 241.
42. K. Roesch: *Stahlguss* 2nd ed., Stahleisen mbH, Düsseldorf 1982.
43. D. Boenisch, K. Daume, *Giesserei* **71** (1984) no. 10, 381 – 393.
44. P. F. Ambridge, P. D. E. Biggins, *Foundry Trade J.* **160** (1986) no. 3325, April 24, 332 – 344.
45. R. L. Naro, A. Dorfmueller jr.: "The Evolution of chemically-bonded sands: past, present and future," *49th International Foundry Congress,* Chicago, April1982.
46. R. Mollenkott, W. Hanke, in [1] pp. 292 – 301; 372 – 380.
47. U. Kleinheyer, in [1] pp. 302 – 308.
48. R. Döpp, R. Deike, G. Gettwert, *Giesserei* **72** (1985), no. 22, 626 – 635.
49. J. M. Fuqua, R. L. Smith: *Practical Applications of the Free Radical Curing Cold Box Binder System,* Ashland Chemical Company, Foundry Products Division, Dublin, Ohio.
50. C. Johnson: "The next step in shell and urethane technology," *AFS Mold/Core Binder, Technology Conference,* Oakland, Calif., 1985.
51. B. Witowska, *Polimery (Warsaw)* **18** (1973) no. 6, June, 317 – 318.
52. D. Boenisch: "Das Coldbox-plus-Verfahren–höhere Kernqualität bei geringerem Binderbedarf," *Giesserei* **72** (1985) no. 15, Jul. 22, 435 – 444, no. 18, Sept. 2, 516.
53. H. Kögler, in [1] pp. 346 – 358.
54. W. Ellinghaus, M. Torbus: "Ein neues Gashärteverfahren zur Kernherstellung," *Giesserei* **74** (1987) no. 10, May 11, 310 – 312.
55. A. Wittmoser, in [1], pp. 441 – 457 and pp. 458 – 461.
56. G. Engels, in [1], pp. 461 – 476.
57. A. Engels, in [1], pp. 381 – 394.
58. W. Ruegg, in [1], pp. 395 – 409.
59. W. Büchen, in [1], pp. 649 – 665.
60. W. Tappen, L. Porep, W. Reinemann, in [1], pp. 608 – 648.
61. W. Venus, in [1], pp. 665 – 690.
62. F. Klein, *Giesserei* **70** (1983) no. 23, Nov. 14, 615 – 619.
63. H. Berndt, in [1], pp. 690 – 709.
64. H. A. Krall, in [1], pp. 715 – 728.
65. C. J. Evans: *Foundry Trade J.* **158** (1985) no. 3301, March 14, 165 – 172.
66. F. Neumann, in [1], pp. 83 – 227.
67. K. James, *Foundry Trade J.* **155** (1983) no. 3269, Aug., 143 – 161.
68. J. M. Greenhill: "Metallurgical Aspects of Melting and Holding in Electrical Furnaces," *Foundry Trade J. Int.* **5** (1982) no. 14, June, 34 – 46.
69. "Developments in electric-melting research," *Foundry Trade J.* **142** (1977) no. 3112, May, 1361 – 1369.
70. W. Riege, in [1], pp. 735 – 751.
71. G. Karabasch, *Giesserei* **71** (1984) no. 4, 161 – 165.
72. G. Asser, *Giesserei* **71** (1984) no. 4, 165 – 169.
73. H. Ritter, J. Bognar, *Giesserei* **71** (1984) no. 24, 923 – 927.
74. J. C. Miske, *Foundry Manage. Technol.* **113** no. 11, Nov., 38 –46.
75. D. L. Maln, *Mod. Cast* **74** (1984) no. 10, Oct., 34 – 36.
76. B. Beckett, *Foundry Trade J.* **155** (1983) no. 3270, Sept. 15, 266 – 268.
77. G. Bewick: "Fettling and Cleaning–A Review of Some Recent Developments," *Foundry Trade J.* **149** (1980) no. 3196, Aug. 28, 307 – 327.
78. W. Gesell, *Giesserei* **64** (1977) no. 22, Oct. 27, 584 – 592.
79. W. Gesell, *Giesserei* **63** (1976) no. 22, Oct. 28, 633 – 640.
80. W. Gesell, *Giesserei* **62** (1975) no. 22, Oct. 30, 579 – 586.
81. H. Jansen, *Foundry Trade J.* **153** (1982) no. 3241, 11 – 23.
82. A. Kolorz, I. Bindernagel, *Giesserei* **67** (1980) no. 7, 157 – 178.
83. E. L. Kotzin, R. G. Godding, R. N. Hutson: "Cleaning room technology–an update for the 80's," *Proc. AFS/CMI Conf.,* Rosenmont, Ill., 1981.
84. H.-J. Müller, D. Weber: "Einzelmaßnahmen zur Verbesserung der Putzarbeiten," *Giesserei* **71** (1984) no. 11, 429 – 440.
85. N. Göbl, L. Monath: "Elektrohydraulisches Putzen," *Giessereitechnik* **29** (1983) no. 7, 205 – 210;no. 8, 239 – 244.
86. A. Vörös: *Gussputzen,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1982.
87. I. Horowitz: *Oberflächenbehandlung mittels Strahlmitteln,* 2nd ed., Vulkanverlag, Essen 1982.
88. R. Bailey, *Mod. Cast.* **72** (1982) no. 1, 46 – 49.

89. W. E. Dougherty, "Modernisierung der Putzerei," (Conference on increasing productivity and profitability) *Proc. AFS/CMI Conf.,* Des Plaines, Ill., 1980.
90. D. Pal, "Heat treatment of iron castings," *FWP J.* **25** (1985) no. 25, April, 43 – 52.
91. S. N. Monteiro et al., *Metalurgia (Sao Paulo)* no. 306, May, (1983) 255 – 258.
92. J. Mazumder, "Laser Heat Treatment: The State of the Art," *J. Met.* **35** (1983) no. 5, May, 18 – 26.
93. H. G. Trapp, in [1], pp. 752 – 766.
94. National Society for clean air: *50th annual conference,* Brighton, England, 1983.
95. R. T. Sohr, *Trans. Am. Foundrymen's Soc.* **83** (1975) 229 – 232.
96. J. C. Hungate: "Cupola Emmission Controls & Problems,"*Proc. Ont. Ind. Waste. Conf.* Toronto, Ottawa 1975, pp. 121 – 145.
97. "Noise Reduction in the Casting Shop–Some practical examples," *Foundry Trade J.* **159** (1985) no. 3314, Oct. 10, 292 – 294.
98. M. H. Quinn: Troubleshooting Cleaning Problems *Die Cast. Eng.* **29** (1985) no. 3, 32 – 38.
99. Deutsche Forschungsgemeinschaft: *Maximale Arbeitsplatzkonzentrationen (MAK) und Biologische Arbeitsstofftoleranzwerte (BAT) 1985,* VCH-Verlagsgesellschaft, Weinheim 1985.
100. E. Ambos, R. Hartmann, H. Lichtenberg: *Fertigungsgerechtes Gestalten von Gußstücken,* Hoppenstedt Technik Tabellen Verlag, Darmstadt 1992.
101. J. Guan, P. R. Sahm: "Berechnung der Restspannungen abgekühlter Gußstücke mit Hilfe der Finiten-Elemente-Methode," *Gießereiforschung* **43** (1991) no. 1, 10 – 17.
102. F. Neumann: *Gußeisen: Schmelztechnik, Metallurgie, Schmelzbehandlung,* Expert-Verlag, Renningen – Malmsheim 1994.
103. R. Richter: *Form- und gießgerechtes konstruieren,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1976.
104. P. Schneider: *Kokillen für Leichtmetallguß,* Gießerei-Verlag, Düsseldorf 1986.
105. H. Schrewe: *Stranggießen von Stahl, Einführung und Grundlagen,* Verlag Stahleisen, Düsseldorf 1987.
106. E. Lossak (Ed.): *Stranggießen,* Deutsche Gesellschaft für Metallkunde e.V., Frankfurt 1991.
107. E. Lossak (Ed.): *Stranggießen: Schmelzen – Gießen – Überwachen,* Deutsche Gesellschaft für Metallkunde e.V., Frankfurt 1987.
108. *Continuous Casting,* **vols. 1 – 3**, Iron and Steel Society of AIME, Warrendale PA, 1983, 1984.
109. E. Hermann: *Handbook of Continuous Casting,* Aluminium Verlag, Düsseldorf 1980.
110. Entwicklungsaufgaben optimal gelöst mit Gußkonstruktionen, VDI-Berichte 1173, VDI Verlag, Düsseldorf 1995.
111. Fortschritt mit Gusskonstruktionen 97', VDI-Berichte 1324, VDI Verlag, Düsseldorf 1997.
112. J. Feikus, W. Schneider, *Giesserei* **83** (1996) no. 1, 20 – 23; no. 9, 17 – 22; no. 15, 17 – 22; no. 19, 23 – 28.
113. A. Oho: *Solidification,* Springer Verlag, Berlin-Heidelberg 1987.
114. P. R. Sahm (Ed.): *Erstarrung metallischer Schmelzen,* Deutsche Gesellschaft für Metallkunde e.V., Frankfurt 1988.
115. R. W. Cahn, P. Haasen, E. J. Kramer (Eds.): *Processing of Metals and Alloys, Materials Science and Technology,* **vol.15,** VCH, Weinheim 1991.
116. *4th International Conference on Semi-Solid Processing of Alloys and Composites*, Conference Proceedings, 19 – 21 June 1996, University of Sheffield, England 1996.
117. *Modelling of Casting, Welding and Advanced Solidification Processes VII*, Conference Proceedings, London 10 – 15 Sept. 1995, TMS Warrendale PA, USA 1995.
118. 18th International NADCA Die Casting Congress 1995, Conference Proceedings, Indianapolis 2-5 Oct. 1995, North American Die Casting Association 1995.
119. T. Enzenbach in [5], pp. 103 – 114.
120. E. Dötsch in [5], pp. 115 – 121.
121. L. Heusler, F. Feikus in [5], pp. 55 – 74.
122. W. Leis, P. Koch in [5], pp. 83 – 102.
123. D. Wolters in [5], pp. 4 – 27.
124. R. Deike in [5], pp. 28 – 34.
125. K. Röhrig in [5], pp. 35 – 39.
126. D. Wittekopf in [5], pp. 40 – 49.
127. A. Engels in [5], pp. 50 – 54.
128. K. H. Schütt in [5], pp. 122 – 137.
129. W. Riege in [5], pp. 138 – 143.
130. E. Brümmer, E. Strümps in [5], pp. 144 – 165.
131. R. Prumbaum in [5], pp. 175 – 182.
132. C. Wilherlm in [5], pp. 191 – 197.
133. A. Bührig-Polaczek, A. Ludwig in [5], pp. 183 – 190.
134. Unfallverhütungsvorschriften "Gießereien" (VBG 230) 1993.
135. Unfallverhütungsvorschriften "Druckgießmaschinen", (VBG7n8) 1993.

136. Bundesimmissionsschutzgesetz, Fassung 1990.
137. Kreislaufwirtschafts- und Abfallgesetz 1994.
138. Anforderung an die stoffliche Verwertung von mineralischen Reststoffen/Abfällen- technische Regeln- Länderarbeitsgemeinschaft Abfall 1996.
139. Rahmen Abwasserverwaltungsvorschrift, Appendix 24: Eisen-, Stahl-, und Tempergießereien 1996; Appendix 29: Nichteisenmetallgießereien 1990.
140. European Waste Catalog (EWC); in 75/442/EWG modified by 91/156/EWG, Brussels 1993.

Fractionation → Distillation and Rectification
Fragrances → Flavors and Fragrances
Francium → Radionuclides
Friedel–Crafts Reaction → Acylation and Alkylation

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Fructose

Standard Article

Wolfgang Wach[1]
[1]Südzucker AG Mannheim/Ochsenfurt, Offstein, Germany

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a12_047.pub2
Article Online Posting Date: January 15, 2004

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Structure, Physical and Chemical Properties
2.1. Structure
2.2. Physical Properties
2.3. Chemical Properties
3. Raw Materials
4. Production
4.1. General Aspects
4.2. Production from Starch
4.3. Production from Sucrose
4.4. Production from Inulin
4.5. Crystalline Fructose
4.6. Further Developments
5. Uses
6. Quality Specifications and Legal Aspects
7. Analysis
8. Storage and Handling
9. Trade Names
10. Economic Aspects
11. Physiological Properties

[Top of Page]

1. Introduction

Fructose, also known as fruit sugar, is a ketohexose ($C_6H_{12}O_6$) monosaccharide reducing sugar found widely in the plant kingdom. Fructose is the sweetest naturally occurring sugar. In its free monosaccharide form fructose is present in notable quantities only in honey and a few fruits (dates, figs, apples, grapes, and most berries). As a constituent of the disaccharide sucrose, which consists of D-fructose and D-glucose residues in glycosidic union, fructose is an important part of everyday sugar usage [1]. Oligo- and polysaccharides of D-fructose (fructans) are present in many plants.

The common name “levulose” was adopted for D-fructose in 1890 after the observation that D-fructose rotates polarized light to the left in aqueous solution [2].

Production of fructose by alkaline isomerization of glucose was reported in 1895 by Lobry de Bruyn and van Ekenstein, but various attempts at commercializing such a process never materialized. Today fructose is produced on a large scale from the raw materials starch, sucrose, and inulin. The industrial-scale production of fructose was made possible by significant advances in enzymatic processes and the development of chromatographic separation technologies in the 1960s and 1970s. Fructose contributes many useful physical and functional attributes to food and beverage applications, and thus fructose can be used as a sweetener in any food.

[Top of Page]

2. Structure, Physical and Chemical Properties

2.1. Structure

D-Fructose [57-48-7] is a monosaccharide with the chemical composition $C_6H_{12}O_6$, M_r 180.159 g/mol. According to IUPAC nomenclature fructose is (–)-D-arabino-hex-2-ulopyranose [3]. Fructose is a 2-ketohexose, that is, a six-carbon pentahydroxyketone with a keto group at C-2. Due to the keto functionality fructose acts as a reducing sugar.

In the solid state and in solution fructose exists in a cyclic hemiacetal form, the preferred five-membered ring known as furanose. Cyclization results in the generation of a new asymmetric center, the anomeric carbon C-2, thereby giving rise to diastereoisomeric hemiacetals which are called anomers and labeled α and β.

The structure of fructose contains three asymmetric centers, thus eight stereoisomers exist. They can be grouped into four pairs of enatiomers, and classified as D- and L-sugars. The isomers of fructose are the ketohexoses sorbose [3615-56-3] (D), tagatose [87-81-0] (D), and psicose [551-68-8] (D).

Anomeric configurations (α and β forms) and different ring sizes (pyranose and furanose) interconvert in solution through a ring-opening reaction called mutarotation. Fructose dissolved in water consists of a mixture of five tautomers (Fig. 1).

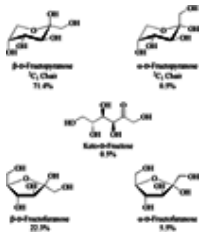

Figure 1. Distribution of frutose tautomers in water at room temperature.

At room temperature, the composition of a 10 wt % solution is 0.5 % α-fructopyranose, 71.4 % β-fructopyranose, 5.5 % α-fructofuranose, and 22.3 % β-fructofuranose, the acyclic keto form being on the order of 0.5 %. The β-pyranose: β-furanose ratio increases with increasing temperature (0 to 60 °C) [4]. As an explanation for the large fraction of β-fructopyranose proportion in D-fructose, the presence of an intramolecular hydrogen bond between the primary 1-OH and the 3-OH groups is invoked [5].

β-D-Fructopyranose is the normal crystalline form of fructose and it is the only tautomer which has been isolated. Other tautomeric forms are isolated as derivatives. β-Fructofuranose is most prevalent for fructose bound in naturally occurring compounds like sucrose [57-50-1], raffinose [512-69-6], and inulin [9005-80-5].

Different conformations for pyranoid sugars are described as chair (C), boat (B), half chair (H), and skew (S). The chair form of fructose follows a similar pattern as that of glucose with a few exceptions. Since fructose has a keto functional group, ring closure occurs at C-2. In the case of fructose a five-membered ring is formed. The OH group on C-5 is converted to an ether linkage to close the ring with C-2. The preferred conformation for β-D-fructopyranose is the $^2C_5$ chair conformation. Envelope (E) and twist (T) arrangements are the principal conformation of the furanose ring. The exact conformation of an individual furanose is not known.

The rare ketohexose L-fructose [7776-48-9], which does not exist in nature, can be synthesized by chemical procedures or by microbiological methods such as enzymatic preparation from L-mannose or bacterial oxidation of L-mannitol to L-fructose. All methods are very elaborate and not suitable for mass production of this sugar [6].

2.2. Physical Properties

The most comprehensive tabulation of physical properties of D-fructose can be found in *Sugar Technologists Manual* [7].

Fructose crystallizes as colorless crystals with a white appearance. The crystals are anhydrous. The density of the solid crystalline phase is 1.60 g/cm^3. The specific enthalpy of combustion is 15 600 kJ/kg. The melting temperature of D-fructose is 378.15 K, and the molar enthalpy of melting 33.0 kJ/mol [8]. Due to its relatively low enthalpy of solution of –37.69 kJ/kg (–6.79 kJ/mol) no cooling effect in the mouth is caused by fructose.

The fundamental knowledge of the macroscopic and molecular behavior of concentrated aqueous carbohydrate solutions is also the basis for the understanding of the process of crystallization [9]. The nucleation, growth kinetics, and the extent of oversaturation are influenced by the liquid film surrounding the crystals [10]. Beside the well-known and preferred anhydrous D-fructose other crystalline phases such as D-fructose hemihydrate and D-fructose dihydrate were determined [11].

Solvents for fructose are water, pyridine, quinoline and a series of polar aprotic solvents. The solubility of fructose in alcohols is limited [31]. Fructose has a solubility of 4 g per gram H_2O at 25 °C, the highest solubility of most, if not all, sugars and sugar alcohols. The dry substance content w_{DS} of a saturated fructose solution is 78.9 % at 20 °C and 88.1 % at 55 °C [7]. It is for this reason that fructose is so difficult to crystallize from aqueous solution [12].

Because of the high solubility of fructose in water industrial crystallization from aqueous solution has to cope with high viscosities. In comparison with a dynamic viscosity η = 96.5 mPa·s of a saturated sucrose solution at 50 °C a saturated fructose solution at this temperature shows a viscosity of η = 1630 mPa·s. Successful crystallization can only be achieved by

defined process control. Such a control may be improved by modeling based not only on viscosity data but also on diffusion coefficients of water and the different tautomers of this sugar.

The water activity and solubility of fructose, other sugars, and polyols are listed in Table 1. Other properties are summarized in Table 2. [7], [13]. Table 3 lists the dynamic viscosity and density of aqueous fructose solutions[12].

Table 1. Comparison of water activity (a_W– value) of saturated solution and solubility at 25 °C for selected crystalline sugars and polyols[12]

Compound	Water activity at 25 °C	Solubility at 25 °C, g/g water
Mannitol	0.977	0.22
Maltose	0.952	0.85
Lactose (monohydrate)	0.931	0.23
Dextrose (monohydrate)	0.891	1.04
Sucrose	0.844	2.07
Sorbitol (gamma)	0.725	2.7
Fructose	0.634	4

Table 2. Physical properties of D-fructose

Property	Unit	Value	Ref. [7], page
Molar mass (anhydride)	g/mol	180.159	117
Density (β-anhydride)	g/cm^3	1.60 (20 °C)	117
Melting point (β-anhydride)	°C	103–105	117
Enthalpy of solution	kJ/mol (kJ/kg)	–6.79 (–37.69)	117
Specific rotation $[\alpha]_D^{20}$	°	–133 to –92	117
Dissociation coefficient pK_1		11.90 (20 °C)	120
Osmotic pressure Π (100 g/L H_2O)	kPa	1275	172
Density of pure fructose solution (w_{DS} = 70 %)	kg/m^3	1344.6	171, 173, 187
Refractive index n_D^{20} (w_{DS} = 70 %)		1.461654	166, 177, 188
Dynamic viscosity η (w_{DS} = 70 %)	mPa·s	178.00	175/176
Physiological calorific value	kJ/kg	15 600	117
Freezing point depression (t_f (w_{DS} = 20 %)	K	2.64	183
a_W– value (saturated solution)		0.63	183

Table 3. Dynamic viscosity and density of pure aqueous fructose solutions in kg/m3 at 20 °C [7]

w_{DS}	20 %	30 %	40 %	50 %	60 %	70 %
η, mPa·s	1.82	2.84	5.01	10.70	32.3	178.00
ϱ, kg/m^3	1081.6	1127.7	1177.1	1229.7	1285.6	1344.6

For the dynamic viscosity η, density ϱ, and refractive index n_D, algorithms for a broad range of concentrations and temperatures are given in [7], [14], [15]. The dependence of the viscosity on concentration and temperature has been investigated for a better understanding of crystallization process [9].

Water vapor sorption is the most relevant property concerning the storage characteristics. The predominant interaction of crystalline sugars with water vapor is by surface adsorption. At critical relative humidity (RH_0) they may deliquesce. For fructose RH_0 is 64 % at 25 °C [16].

Fructose is the sweetest natural sugar. The sweetness of crystalline fructose is 1.8 times that of sucrose, whereas the sweetness of fructose solutions is not as high. In this respect, fructose would be the ideal nutritive sweetener because it would give the same texture and sweetness to food as sucrose but with a lower energy content. However, when fructose is in solution the relative sweetness varies with temperature, pH, and concentration [17]. The relative sweetness of fructose

increases with decreasing temperature and vice versa. This phenomenon could be explained by greater formation of furanose forms at higher temperatures.The sweetness of fructose also increases with increasing acidity [18].

2.3. Chemical Properties

Fructose exhibits the same reactions as other carbohydrates to form esters, ethers, and acetals [19], [20]. Typical carbonyl reactions with nucleophiles such as amino acids, peptides, or other amino-group donating compounds could be realized. The mild acid-catalyzed alkylation of fructose with alcohol results in fructose glycosides [21].

Under alkaline conditions fructose undergoes *isomerization* via the 1,2-enediol to the aldose glucose, which can epimerize to mannose. The total rearrangement is known as the Lobry de Bruyn – van Ekenstein reaction [22]. Fructose is converted to the stable aldohexoses glucose [50-99-7] and mannose [345-28-4] and a small amount of psicose [551-68-8]. Isomerization under alkaline conditions and elevated temperatures is followed by fragmentation, and the resulting products are glyceraldehydes, lactic acid, methyl glyoxal, etc.

Reduction of D-fructose by the enzyme ketose reductase can lead to either D-sorbitol [50-70-4] or D-mannitol [69-65-8] [19]. Commercial mannitol production is based on catalytic hydrogenation of the keto group of fructose with the help of heterogeneous metal catalysts such as Ni or Cu. Depending on the selectivity of the metal catalyst 40–70 % of mannitol is obtained from a 100 % fructose solution. Mannitol can be purified by crystallization or chromatographic separation, since sorbitol is formed as well. Mannitol is used in chewing gums, for example, and in pharmaceutical preparations [25].

Heterogeneous catalytic *oxidation* of fructose is performed with air or molecular oxygen in the presence of supported noble metal catalysts [25]. Well-known catalytic systems are Pt/C and Pd/C. Apparently, in ketoses like fructose the primary hydroxyl group on C-1 is activated by the adjacent carbonyl group, and this OH group is attacked preferentially to the other primary hydroxyl group at C-6. Catalytic oxidation of D-fructose leads to two major products: 2-keto-D-gluconic acid [328-50-7] (60 %) and D-*threo*-hexo-2,5-diulose (“5-ketofructose”) [26].

Heating fructose above the boiling point of concentrated solution or the sintering range of the crystalline substance leads to *dehydration* with anhydride formation or caramelization [27]. The formation of condensation products such as difructose dianhydride by elimination of water is dependent on the concentration of fructose, pH, and temperature.

Thermal decomposition of fructose in the presence of amino acids leads to Maillard products, color- and flavor-giving components. This reaction is responsible for the nonenzymatic browning of foodstuffs.

The acid-catalyzed conversion of concentrated aqueous fructose to 5-hydroxymethylfurfural (HMF) [67-47-0] has been known for a long time [28]. The ketose fructose is much more suitable to opening of the ring followed by rearrangement to form the acyclic 1,2-enediol intermediate than glucose. Subsequent dehydration of the enediol leads to 5-hydroxymethylfurfural with removal of three molecules of water. Thus fructose is a good starting material for production of HMF. The HMF can be rehydrated in the presence of two equivalents of water to form levulinic and formic acids [29].

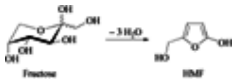

Figure 2. Formation of HMF.

Due to its wide range of diverse industrial applications, many suggestions have been made to utilize this dehydration reaction in a commercially practical manner for the manufacture of HMF. Based on renewable resources, HMF is suggested to be a very interesting and versatile intermediate product [30]. However, despite the fact that the technology of HMF production has been elaborated on a pilot-plant scale, it is not produced on industrial scale.

Monosaccharides like fructose are starting materials in the *enantiospecific synthesis* of natural products. The advantage of using monosaccharides like fructose as a chiral pool are commercial availability, low cost, valuable carbon source, enantiopure materials, and several defined chiral centers.

Rare sugars such as *D-psicose* could be made from D-fructose by using D-tagatose 3-epimerase from *Pseudomonas cichorii* ST-24. After epimerization the substrate D-fructose was removed by treatment with baker’s yeast [32].

[Top of Page]

3. Raw Materials

Fructose occurs widely in the plant kingdom in its free form or in free sucrose, which consists of D-fructose and D-glucose residues in glycosidic union. Polysaccharides of D-fructose (fructans) are common in many plants. The largest amount of fructose is found in honey (40 g/100 g), apples (6–8 g/100 g), pears (5–9 g/100g), prunes (15 g/100g) and their juices [P 993].

Under the aspect of economy the primary raw materials used worldwide for the production of fructose are starch, sucrose, and inulin.

Starch is a storage carbohydrate in many plants such as corn, barley, wheat, rice, potato, tapioca, and other starch-bearing grains and roots. Starch is a polysaccharide consisting only of glucose units. A precondition for making fructose from sources which consist only of glucose is easy conversion to fructose. The most important plant for fructose production from starch is corn.

Sucrose, a disaccharide consisting of fructose and glucose, is produced from sugar beet or sugar cane on a very large scale (140×10^6 t/a) worldwide. Sucrose is almost an ideal raw material because of its extremely good availability, high purity, and the high content of fructose (equimolar mixture with glucose) after hydrolysis.

Materials that are naturally rich in fructose include the storage carbohydrates fructans such as inulin or levan. Fructans are made up of D-fructose residues linked to a terminal sucrose residue. Planting, harvesting, and scale of production have to be economically efficient [33], [34]. Materials such as roots of chicory, tubers of Jerusalem artichoke, tubers of dahlia, and hearts of globe artichoke are rich in fructan [24], [35].

Levan consists mainly of β(2,6) linked fructosyl-fructose bonds and is today irrelevant as raw material for fructose production, whereas fructose production based on the raw material inulin is carried out on a large scale. The primary sources for industrial-scale inulin and inulin/fructose production are artichokes and chicory. The chicory root, for example, has an inulin content of 15–20 %. Inulin is a polydisperse mixture of molecules with the basic chemical structure GF_n where G is the glucosyl moiety, F is the fructose moiety and *n* is the number of fructose moieties linked by β(2,1) linkages. The degree of polymerisation (DP) of inulin typically ranges from 3 to 60.

Virtually, all of the inulin produced in the EU is derived from the chicory root. The European inulin industry is relatively new, having only been established in 1984 [36]. In addition to the industrial application of inulin, it is used as a source of fructose and oligofructose. Oligofructose, which is partially hydrolyzed inulin, consists of between 2 and 20 fructose units and provides nutritional and functional advantages.

[Top of Page]

4. Production

4.1. General Aspects

Production of fructose by alkaline isomerization of glucose was reported in 1895 by Lobry de Bruyn and van Ekenstein, but various attempts at commercializing such a process never materialized due to low fructose yield and formation of color, off-flavor, and other byproducts [20].

The industrial-scale production of fructose syrup was made possible by significant advances in industrial isomerization and the development of chromatographic separation technologies in the 1960s and 1970s. The most sophisticated technological step in the production of fructose from the raw material sucrose or starch is chromatographic separation of the fructose/glucose syrup. Separation is achieved on a strongly acidic cation-exchange resin in the calcium form. In the past batch systems were used, but more recntly the simulated moving bed (SMB) process became more and more mature, and today it is the state of the art. All newly built plants will use SMB technology for the separation of fructose and glucose. The main advantages are the lower water consumption, lower steam consumption, and high recovery rate. Suppliers of classical SMB systems use resin beds with different depths and different types of resin compartments. One supplier has installed the compartments on a carousel in order to replace the on/off valves with a rotary valve The carousel rotating under the distribution ports moves up to 20 large columns [37]. In a sequential simulated moving bed system (SSMB), the streams do not flow continuously. Each step is divided into two periods: one with production flows and one with just circulating stream flowing in a loop. The division of each step allows much better defined fraction cutpoints. Depending on the capacity in the sugar industry for sugar separation, a bed size of several 100 m^3 is required [38], [39].

For reliable modeling and design of a large-scale SMB operation it is crucial to understand the properties of the adsorbent, adsorption equilibria (partition coefficients), and mass transfer rates. The resins have a volume-weighted median pore size of 1 nm, roughly equal to the diameter of the hydrated sugar molecules [40].

Another technological improvement that made the production of fructose syrups more effective was the use of monodisperse ion-exchange resins for the chromatographic separation. The use of these resins leads to higher recovery, lower dilution, and improved overall operating efficiency.

4.2. Production from Starch

To produce fructose syrup from the glucose polymer starch, enzymatic processes (see also ⟶ Enzymes – Immunoassays) had to be first developed for liquefying starch, converting liquefied starch to glucose, isomerizing glucose to fructose, and separating fructose from the solution. Fructose, however, is not a starch hydrolysis product per se [20]. The starch is converted to glucose by using enzymatic processes with α-amylase (E.C. 3.2.1.1) [9000-90-2] and glucoamylase (EC 3.2.1.3) [9032-08-0] and the glucose syrup is epimerized to a mixture of fructose and glucose by using glucose isomerase.

Glucose isomerase (D-xylose isomerase, E.C. 5.3.1.5) [9055-00-9] catalyzes the isomerization of D-glucose to D-fructose. The enzyme is actually a xylose isomerase, but also isomerizes D-xylose to D-xylulose, as well as D-ribose to D-ribulose. The isomerase is specific for the α-anomeric form of the substrate. The kinetics of the glucose isomerase are dependent on various metal ions and pH values [41]. Glucose isomerase is activated by magnesium, and calcium acts as enzyme inhibitor. Isomerization of glucose to fructose is a thermodynamically controlled reaction. The equilibrium constant of this reaction is approximately unity at 60 °C.

In the 1950s, several enzymes were discovered that had the ability to convert three-, four-, and five-carbon aldoses to the corresponding ketoses. An industrial process for isomerization of glucose to fructose was developed and commercialized in 1966. Since 1968, a 42 % fructose syrup, high-fructose corn syrup (HFCS) has been manufactured by the enzymatic route. The soft drink industry began to substitute high-fructose syrups for sucrose from the late 1970s onwards. By chromatographic separation the fructose purity is increased to at least 90 %, followed by crystallization.

4.3. Production from Sucrose

The first commercial process for the production of crystalline fructose was based on sucrose as raw material. (Fig. 3). Sucrose is hydrolyzed and subsequently separated into glucose and fructose by chromatography.

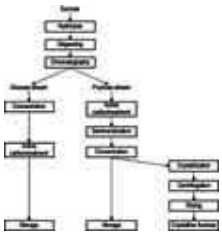

Figure 3. Production of fructose.

The hydrolysis of the disaccharide sucrose to its free component sugars fructose and glucose can be affected by different kind of technologies like enzymatic reaction or homogeneous or heterogeneous acid-catalyzed reaction.

Hydrolysis can be performed by a highly specific enzymatic reaction using an invertase from *Saccharomyces cerevisea*, namely, β-D-fructofuranosidase (E.C. 3.2.1.26) [9001-57-4]. Drawbacks of this technology are the costs of enzyme and the batch process. Until now immobilized invertase has not been used in an industrial production process.

Homogeneously catalyzed hydrolysis by acids is accomplished by free mineral acids permitted by food regulations, such as hydrochloric acid or sulfuric acid. The degree of inversion is precisely adjustable by means of a neutralization step at the end of the reaction with NaOH or KOH. High ash content is caused by this method. The reaction conditions with high temperatures and low pH are quite harsh and lead to undesirable byproducts.

State of the art for sucrose hydrolysis is the use of immobilized acids such as strongly acidic cation exchangers in protonated (H^+) form in a fixed-bed column. The resins have legal approval for food production. In a continuous bulk production sucrose solutions with 50-60 wt % can be completely hydrolyzed at an inversion temperature of 30-45 °C, and a practically ash-free product with avoidance of undesirable materials is obtained.

The resulting hydrolyzate is the feed for the chromatographic separation. The batch process is being replaced by semicontinuous or continuous SMB technology. The two fractions of this separation have a minimum purity of 95 %. The following crystallisation step is described in Section Crystalline Fructose.

4.4. Production from Inulin

Chicory processing has evolved from laboratory to full industrial scale since the mid-1990s. In the EU this route has become very important for the production of fructose but also for inulin and oligofructose [24].

For the production of fructose from inulin a hydrolysis step is needed. This can be done by acid catalysts or with the help of enzymes. Acids such as sulfuric, hydrochloric, phosphoric, citric, oxalic, and tartaric acids perform the hydrolysis reaction effectively but they suffer from the drawbacks of promoting the decomposition of fructose and of contaminating the product. To overcome these drawbacks, academic projects were performed with acidic zeolites as catalysts for hydrolysis of inulin to fructose. This reaction is relatively fast and does not decompose fructose. Presently this route of production is only of academic interest [33].

Given the drawbacks of acid hydrolysis the use of enzymes is the preferred strategy. A mixture of the enzymes exo-inulinase (E.C. 3.2.1.80) and endo-inulinase (E.C. 3.2.1.7) from fungi (e.g., *Aspergillus niger*, Fructozyme from Novozymes) hydrolyzes inulin to fructose in a single-step reaction [35]. In contrast, conventional fructose production from starch needs at least three enzymatic steps and yields only 45 % of fructose [42]. Partial hydrolysis yields oligofructose, which consists of between 2 and 20 fructose units.

The hydrolysis of inulin can be carried out batchwise in large tanks with holding times of between 12 and 24 h. A solution of 10–20 wt % dry substance is preferred. The fructose content of the syrup produced varies with the source of inulin and the harvesting time of the plant, but is usually between 85 and 95 %, the rest being 3–13 % glucose and 1–2 % oligosaccharides. A temperature of 60 °C is recommended. Activity is optimal in the interval between 57 and 63 °C. Above 63 °C the stability of the enzyme is markedly reduced. At lower temperatures the rate of hydrolysis is reduced and the risk of microbial infection is increased. The enzymatic hydrolysis is carried out in neutral or slightly acidic medium. After hydrolysis the syrup is demineralized, decolorized, and further concentrated. The resulting fructose syrup has a high fructose content of 85 % or more. The advantage of the enzymatic route is the avoidance of aftertaste and degradation products.

If higher purities are required a chromatographic separation process is possible. A process without additional fractionation or refining giving purities of at least 93.5 wt % is described. The fructose content is strongly dependent on the degree of polymerization of the raw material [35].

4.5. Crystalline Fructose

The crystallization of fructose is the most difficult part of the whole process due to the physicochemical properties and a number of unfavorable reactions that can take place. The crystallization process itself can be batch or continuous. Independent of the raw material a fructose syrup with at least 90 % fructose in the total solids content is needed as starting point for successful crystallization of pure fructose. A number of different aqueous and solvent processes have been proposed for the crystallization of fructose. Due to the very costly removal and recovery of the solvent from the product the most successful crystallization processes in use today are based on the crystallization of fructose from aqueous solution [20].

Various combinations of concentration under atmospheric or reduced-pressure conditions and seeding to initiate crystallization followed by a cooling step to allow for crystal growth are utilized. One process involves cooling an aqueous solution with a solids content of 85–95 % from 65 to 25 °C at a rate of < 1 °C/h. Batch and continuous crystallization processes have been devised that incorporate preprogrammed concentrating, cooling, warming, and recooling cycles. To produce a high purity product that will remain free-flowing and granular during storage, the crystals must be thoroughly freed of mother liquor. This is done in centrifuges, generally of the basket screen type, which include a washing zone or cycle to rinse the crystals free of any remaining mother liquor. The crystals are then exposed to a further drying step to remove surface moisture to prevent lumping and give a good storage quality [12].

Impurities such as glucose and difructose dianhydrides have a negative influence on the crystallization of anhydrous fructose from aqueous solution. Glucose is the major impurity likely to be present in fructose syrup from starch or sucrose, while several difructose dianhydrides are formed in situ under crystallization conditions. Both sets of impurities were found to inhibit crystal growth, but the mechanisms responsible in each case are different [43]. The presence of glucose increases the solubility of fructose in water and thus lowers the supersaturation of the solution. This is probably the main effect responsible for the decrease of crystal growth. The decrease of crystal growth by difructose dianhydrides is probably caused by the incorporation of these impurities into or adsorption on the crystal surface, which would accept fructose molecules in the orientation that existed in the difructose dianhydride [43].

4.6. Further Developments

Research and development activities for improvement in fructose production are focused on raw materials, enzymatic conversion, and chromatographic separation.

The combination of reaction catalyzed by enzymes and a chromatographic separation in a single unit operation is proposed for the simultaneous inversion of sucrose [44] or the enzymatic isomerization of glucose to fructose and the separation of glucose and fructose. In the case of the isomerization the idea is to shift the reversible reaction towards the favorable product fructose by combining an immobilized enzyme reaction and adsorption. The bioreactor was applied to production of higher fructose syrup. This technology is not mature and work has been done on lab scale or modeling working groups [23]. One of the aims of the last few decades has been to directly obtain fructose purity of 90 % by genetically modified enzymes in the isomerization step. Another aim is to crystallize and separate fructose and glucose directly from lower purity high-fructose syrups [8].

The production of enriched fructose syrups and ethanol from a medium with high sucrose concentration was studied in a batch process using a mutant of *Saccharomyces cerevisiae* [45].

The production of inulin with a high degree of polymerization from sucrose by microorganisms could be an alternative to native inulin [46].

[Top of Page]

5. Uses

Fructose was positioned early as a nutritionally advantageous sweetener because of the special metabolic properties. Therefore, many of the products first sold that contained fructose were dietetic products such as meal replacers and diabetic foods. These included various powdered diet beverages, nutritional candy bars and other specialty food items [12]. Fructose contributes many useful physical and functional attributes to foods and beverages, and thus fructose can be used as a sweetener in any food.

As a result of the exceptional sweetening power fructose is used primarily for added sweetness in low-calorie foods. Its tendency to confer “body” on a product is a major advantage over intense sweeteners. The synergetic effects with other sweeteners such sucrose allow a reduction in amount of sweetener and calories without a reduction in perceived sweetness. Its ability to enhance flavors is remarkable, especially for fruit flavors. Consequently, reduced levels of sweeteners and flavors could lead to saving of costs. High solubility at low temperature and large freezing-point depression of fructose are particularly interesting for ice cream and frozen-dessert manufacturers, since these properties influence product taste and texture. Fructose is superior to all nutritive sweeteners and humectants in controlling the water in frozen systems.

For baked goods the benefits of fructose can be seen in its influence on moisture management which leads to extended shelf life, color development due to intense browning, and flavor development.

Fructose is used extensively in carbonated beverages/soft drinks, frozen desserts, baked goods, canned fruits, confectionery (chocolate, jellies, and gums), and diary products. Typical formulations include sports drinks and reduced-calorie beverages. Fructose has the ability to mask bitter and metallic tastes and aftertastes, especially with intense sweeteners [47]. Due to its extremely good solubility fructose is suitable for use in dry-mix beverages and desserts. Fructose is also more soluble in alcohol than sucrose, enabling the production of sweetened liqueurs with higher contents of dry substances [47]. Synergy with starches or other gel-forming products leads to increased gel strength and improved texture. Pharmaceutical uses include tablets or solutions for injections and infusions.

[Top of Page]

6. Quality Specifications and Legal Aspects

Fructose has been considered, like sucrose, to be de facto generally recognized as safe (GRAS). FAO-WHO Codex Alimentarius Commission [48], Food Chemicals Codex (FCC) [49], Council Directive 2001/111/EEC [50], European Pharmacopeia (E.P.4.05) [3], United States Pharmacopeia (U.S.P. XXVI) [51] define purity standards for crystalline fructose. The FAO-WHO Codex describes fructose as “purified and crystallized D-fructose with a fructose content of not less than

98 % m/m, and a glucose content of not more than 0.5 % m/m".

Only crystalline fructose is designated by the term "fructose". Dissolved in water it results in "fructose solution". A product of lesser purity, containing about 95 % fructose on total solids is referred to as "fructose syrup".

Fructose is regarded as a food ingredient, not an additive. In several countries it has the further status of sugar substitute and is permitted for sweetening foods for diabetics.

The composition and quality requirements for crystalline fructose in the Food Chemicals Codex are listed in Table 4 [49]:

Table 4. Fructose composition and quality requirements according to FCC.

Description*	
Assay**	
Arsenic (as As)	≤ 1 mg/kg
Chloride	≤ 0.018 %
Glucose	≤ 0.5 %
Heavy metals (as Pb)	≤ 5 mg/kg
Hydroxymethylfurfural	≤ 0.1 %, calculated on dried basis.
Lead	≤ 0.1 mg/kg
Loss on drying	≤ 0.5 %
Residue on ignition	≤ 0.5 %
Sulfate	≤ 0.025 %

* "Fructose occurs as white, hygroscopic, odourless, purified crystals or as a purified crystalline powder having a sweet taste. It is a natural constituent of fruit (hence, the name "fruit sugar") and is obtained from glucose in corn syrup by the use of glucose isomerase. Its density is about 1.6. It is soluble in methanol and in ethanol, freely soluble in water and insoluble in ether."

** "Not less than 98.0 % and not more than 102.0 % of fructose ($C_6H_{12}O_6$), after drying."

Typical composition of commercial grades of crystalline fructose is generally specified by the manufacturer (Table 5):

Table 5. Typical specification of crystalline fructose by manufacturer.

	Crystalline fructose
Solids, %	99.5
Moisture, %	≤ 0.5
Carbohydrate (dry solids basis), %	
Fructose	≥ 99.5
Dextrose	≤ 0.5
Oligosaccharides	trace to none
Ash, %	≤ 0.01
Heavy metals, mg/kg	<1 mg/kg

[Top of Page]

7. Analysis

Chromatographic methods are the most powerful analytical techniques for the analysis of fructose samples (raw material, food, production control, etc.) [52-55]. They allow the simultaneous analysis of the type and concentration of different kinds of saccharides in the sample. Thin layer chromatography (TLC) [56], high-performance liquid chromatography (HPLC), and high-performance anion exchange chromatography (HPAEC) with pulsed amperometric detection (PAD) [57], [58] are commonly used to separate and identify carbohydrates. Well-established for fructose determination is a HPLC system with strongly acidic cation exchanger in the calcium form as separation phase and water as eluent, or with amino-bonded phases with acetonitrile/water as eluent, using a refractive index (RI) detector. Gas chromatography (GC) is strongly recommended if impurities or minor components are of interest. GC can be applied after conversion of carbohydrates into volatile derivatives (e.g., oximation and subsequent trimethylsilylation) [59].

The determination of individual sugar contents in fresh fruits and vegetables and their juices is an important chemical analysis carried out to evaluate quality and to detect adulteration or contamination [60]. Up to now, analytical techniques such as liquid chromatography or enzymatic methods have become established [61], [62].

Chemical methods are based on the fact that fructose reacts as reducing agent. The principal methods are based upon the property of reducing alkaline solutions of certain metallic salts, especially alkaline copper solutions, and on the property of forming colored products with aromatic compounds in the presence of acid. The disadvantage of these reactions is the nonspecificity for fructose. Various carbohydrates can be detected and determined by these methods. In the reduction methods the reduced or unreduced copper can be determined by titration [63-65] (Lane-Eynon or the Luff-Schoorl procedure) or the reduced copper is determined gravimetrically [66]. The amount of reducing sugar, the fructose, is calculated from the required titer or the amount of reduced copper. An example for detection and determination of fructose by a colorimetric method is the reaction with anthron.

Fructose can be specifically determined by an enzymatic method, i.e., the hexokinase method [67], [68]. D-Fructose is phosporylated by hexokinase and adenosine-5′-triphosphate to D-fructose-6-phosphate, which is converted by phosphoglucose isomerase to D-glucose-6-phosphate. The glucose-6-phosphate is oxidized in the presence of the enzyme D-glucose-6-phosphate dehydrogenase by NADP to D-gluconate-6-phosphate. Reduced NADP (NADPH) is formed by the oxidation reaction and is measured photometrically by its characteristic absorbance at 340 nm. The amount of NADPH is equivalent to the amount of fructose. When the sample contains glucose, first the glucose must be determined by the hexokinase method. The enzymatic method is rapid, specific, and sensitive to low concentrations and is therefore ideal for determination of carbohydrates. Liquid foods can be tested directly, whereas solid foods have to be dissolved in water first. Enzyme assay kits for fructose-glucose-sucrose mixtures can be purchased commercially.

Physical methods like polarimetry, refractive index, IR, and density are commonly used. Polarimetry is a well established and quick method for purity testing. In the U.S.P. XXVI the polarimeter test is prescribed as assay for fructose determination. The concentration of carbohydrate in an unknown sample is determined by measuring its angle of rotation and comparing it with the calibration curve.

[Top of Page]

8. Storage and Handling

Crystalline fructose is a high-purity, free-flowing solid, which can be mixed with all types of dry ingredients. Storing and handling of crystalline fructose is determined by its chemical and physical characteristics. For handling and storage of crystalline fructose a maximum temperature of 25 °C and a relative humidity of ≤ 60 % [69] is recommended to prevent subsequent caking and flow problems. When stored and handled under the recommended conditions, crystalline fructose is shelf-stable and retains its free-flowing characteristics. Processing conditions should be controlled to avoid hardening and lumping problems. Most bagged fructose is packed in foil-lined paper bags to ensure minimal transfer of moisture to the product. Big bags are filled with up to 1000 kg of crystalline fructose.

Commercial fructose solutions with a dry matter content of 70 % are more resistant to microbial spoilage than equivalent concentrations of sucrose, owing to the greater osmotic pressure and the lower water activity [47]. It can be stored without any fear of crystallisation.

[Top of Page]

9. Trade Names

All commercial fructose products are characterized by the fructose content and residual other carbohydrates, which depend on the raw material: glucose from sucrose, fructooligosaccharides or glucose from chicory, and glucose, maltose, or oligosaccharides from starch.

Fructose as crystalline sweetener is often simply called “crystalline fructose” (e.g., Archer Daniels Midland, Südzucker). Liquid products with high purities (≥ 95 %) are also commercially successful.

Trade names include Krystar Liquid or Cristallin (Staley in the USA and Amylum in Europe, as part of the Tate & Lyle Group), Fructofin (Danisco), Fruisana (retail; Danisco), Fructose N (Amylum), Fruitose (Galam), fructose syrup CornSweet 95 (Archer Daniels Midland) and Fruchtzucker kristallin and Fruchtzuckersirup F 70/95 (Südzucker).

Fructose products derived from the hydrolysis of chicory inulin are produced and distributed by the Orafti Group. Their fructose syrup is named Raftisweet. Orafti is one of the four activity groups of Raffinerie Tirlemontoise which is part of the Südzucker group.

Products based on inulin from the producer Cosucra are named Fructoline L 85 (liquid), Fructoline L 95 (liquid) and Fructoline CR (crystallized). Frutasun, another high-fructose syrup based on chicory, is produced by Sensus.

[Top of Page]

10. Economic Aspects

Fructose is marketed as a healthy fruit sugar for added sweetness in low-calorie foods. Depending on the market functionality and/or nutritional aspects of fructose as a natural sweetener can be emphasised [70].

During the 1960s industrial quantities of crystalline fructose were produced in Europe with sucrose as raw material. Efforts in reducing production time and costs, especially in starch-based fructose production led to initial growth in the late 1980s. It was thought that crystalline fructose had the potential to make a breakthrough (as a sugar substitute) in granular uses similar to that which high-fructose corn syrup had achieved in liquid applications, notably the soft-drinks market. Until now, this has not proven to be the case.

Due to the complexity of the described production processes and the cost of the raw material this product is relatively expensive and is often viewed as a costly specialty caloric sweetener. However, over the last decade technological advances in corn-based sweetener processing techniques and chromatographic separation systems have lowered production costs.

The total sales of crystalline fructose are about 240 000 t worldwide. The largest producer of fructose has a stake of about half of the world production. The tonnage for total sales represents a little over 0.1 % of the global caloric sweetener market. In the years between 1985 and 1992, there was a strong growth of the crystalline fructose market. After several years crystalline fructose sales reached a plateau. The USA is the biggest producer and exporter of crystalline fructose. The largest amounts are sold to Canada, Japan, and the EU, the second largest market for crystalline fructose. Israel is permitted to export crystalline fructose to the EU under preferential trade agreement duty free. In the EU fructose syrups are very successful.

[Top of Page]

11. Physiological Properties

According to the structural classification of carbohydrates as suggested by the Food and Agriculture Organization and the World Health Organization (FAO/WHO) in 1998 fructose belongs to the group of sugars. Its energy value is specified by the EEC as 17 kJ/g (4 kcal/g) [71].

Ingestion of fructose is followed by its absorption in the small intestine as a consequence of facilitated diffusion mediated by Glut5, a member of the glucose transporter family. This absorption proceeds more slowly than the active absorption of glucose. Absorbed fructose enters the bloodstream via the portal vein, which transports it to the liver, where fructose is preferentially taken up. At high intakes, part of it (30–40 %) is rapidly metabolized in the intestinal mucosa to lactate, which passes to the liver and is transformed into glucose [72].

Fructose is metabolized by the human body fairly readily. Fructose metabolism (Fig. 4) involves the phosphorylation of fructose to fructose-1-phosphate by fructokinase (E.C. 2.7.1.3), cleavage of the resulting fructose 1-phosphate into dihydroxyacetone phosphate and D-glyceraldehyde by the liver isoenzyme of aldolase (E.C. 4.1.2.13) and phosphorylation of D-glyceraldehyde to its 3-phosphate by triose kinase (E.C. 2.7.1.28) [73].

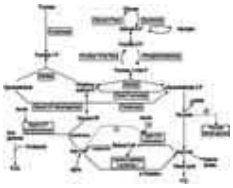

Figure 4. Fructose metabolism.

The pathways of glucose and fructose metabolism in the liver converge at the triose phosphate stage as the glyceraldehyde-3-phosphate enters glycolysis or gluconeogenesis.

Thus, major products of fructose metabolism in the liver are glucose, glycogen, and lactate. Smaller quantities are oxidized to carbon dioxide, ketone bodies, or converted to triacylglycerol. Because the majority of the absorbed fructose is taken up by the liver and requires conversion to intermediates before entering the blood and undergoing metabolism, the effect of fructose on blood glucose level is very small [74].

Blood glucose levels following an oral fructose load are lower than after a comparable oral glucose load, i.e., fructose is a weakly glycemic carbohydrate. Thus, the glycemic index (GI) of fructose (using white bread as a standard) is only 32, which is explained by the partial conversion of fructose into glucose in the liver.

In addition, fructose has only a modest affect on the stimulation of insulin secretion, and uptake of fructose into the cells does not require the presence of insulin. Together with its weakly glycemic properties, this is why fructose is recommended as a dietary sweetener for diabetics, especially for people with non-insulin-dependent (type II) diabetes mellitus (NIDDM).

One of the most controversial areas in the dietary fructose literature is the effect of fructose on serum lipids and in particular on triglyceride levels. This issue has arisen because of alterations in plasma lipid concentrations following long-term excessive consumption of large amounts of carbohydrates, particularly fructose. The significance in humans is not clear and is still a matter of debate.

In hereditary fructose intolerance, a rare genetic defect, the absence of liver fructose 1-phosphate aldolase isoenzyme leads to an accumulation of fructose 1-phosphate in the liver [75]. Blockage of glycolysis and gluconeogenesis occurs and leads to hypoglycemia and eventually to hypoglycemic shock. Patients suffering from this disorder must avoid fructose in their diet.

[Top of Page]

References

1. J. E. Riby, T. Fujisawa, N. Kretchmer, *Am. J. Clin. Nutr.* **58** (1993) suppl., 748 – 753.
2. R. Bernetti, *Cereal Foods World* **35** (1990) 390 – 393.
3. European Pharmacopeia 4.05, Europäisches Arzneibuch, 4. Ausg., Deutscher Apotheker Verlag Stuttgart, Govi-Verlag-Pharmazeutischer Verlag, Eschborn, 2002, 1927 – 1928.

4. M. Cockman, D. G. Kubler, A. S. Oswald, L. Wilson, *J. Carbohydr. Chem.* **6** (1987) 181 – 201.
5. F. W. Lichtenthaler, S. Rönninger, *J. Chem. Soc. Perkin Trans.* **2** (1990) 1489 – 1497.
6. H. Itoh, K. Izumori, *J. Ferment. Bioeng.* **81** (1996) 351 – 353.
7. Z. Bubnik, P. Kadlec, D. Urban, M. Bruhns: *Sugar Technologists Manual*, 8th ed., Verlag Dr. Albert Bartens, Berlin, 1995.
8. A. Flood, *Proc. Regional Symp. Chem. Eng. (RSCE)*, Singapore, 2000.
9. M. Rampp, C. Buttersack, H.-D. Lüdemann, *Carbohydr. Res.* **328** (2000) 561 – 572. Links
10. L. D. Shiau, K. A. Berglund, *AIChE J.* **33** (1987) 1028 – 1032. Links
11. F. E. Young, F. T. Jones, H. J. Lewis, *J. Phys. Chem.* **56** (1952) 1093 – 1096.
12. L. M. Hanover, J. S. White, *Am. J. Clin. Nutr.* **58** (1993) 724S – 732S. Links
13. A. Rapaille, *Fruit Process.* **6** (2000) 206 – 212.
14. A. Emmerich, L. Emmerich, *Zuckerindustrie* **111** (1986) 441 – 448. Links
15. A. Emmerich, B. Prowe, K. J. Rosenbruch, *Zuckerindustrie* **109** (1984) 525 – 537. Links
16. B. C. Hancock, S. L. Shamblin, *PSST* **1** (1998) 345 – 351.
17. M. C. Zamora, F. M. Buratti, M. E. Otero-Losada, *J. Sensory Studies* **13** (1998) 213 – 228.
18. A. M. Fontvieille, A. Faurion, I. Helal, S. W. Rizkalla, S. Falgon, M. Letanoux, G. Tchobroutsky, G. Slama, *Diabetes Care* **12** (1989) 461 – 486.
19. B. Fraser-Reid, K. Tatsuta, J. Thiem (eds.): *Glycoscience: Chemistry and Chemical Biology I-III*, Springer-Verlag, Berlin, Heidelberg, New York, 2001.
20. R. E. Hebeda, *Biotechnol. Enzym. Food* **9** (1995) 737 – 765.
21. A. T. J. W. de Goede, F. van Rantwijk, H. Van Bekkum, *Starch/Stärke* **47** (1995) 233 – 237.
22. P. M. Collins, R. J. Ferrier: *Monosaccharides*, John Wiley & Sons, Chichester, 1995.
23. K. Hashimoto, S. Adachi, Y. Shirai in G. Ganetsos, P. E. Barker (eds.): *Preparative and Production Scale Chromatography*, Marcel Dekker., New York, 1993, pp. 395 – 419.
24. L. De Leenheer: “Production and Use of Inulin: Industrial Reality with a Promising Future” in H. Van Bekkum, H. Röper, F. Voragen (eds.): *Carbohydrates as Raw Materials III*, CRF, The Hague, 1996.
25. A. Abbadi, H. van Bekkum: “Metal-Catalyst Oxidation and Reduction of Carbohydrates” in H. Van Bekkum, H. Röper, F. Voragen (eds.): *Carbohydrates as Raw Materials III*, CRF, The Hague, 1996.
26. A. W. Heinen, J. A. Peters, H. van Bekkum, *Carbohydr. Res.* **304** (1997) 155 – 164. Links
27. M. J. Antal, Jr., W. S. L. Mok, G. N. Richards, *Carbohydr. Res.* **199** (1990) 91 – 109. Links
28. B. F. M. Kuster, *Starch/Stärke* **42** (1990) 314 – 321.
29. P. M. Mulvihill: “Crystalline and Liquid Dextrose Products” in F. W. Schenk, R. E. Hebeda, (eds.): *Starch Hydrolysis Products*, VCH, New York, Weinheim, Cambridge, 1992.
30. H. Schiweck, M. Munir, K. M. Rapp, B. Schneider, M. Vogel: “New Developments in the Use of Sucrose as an Industrial Bulk Chemical” in F. W. Lichtenthaler, (ed.): *Carbohydrates as Organic Raw Materials*, VCH, Weinheim, Germany, 1991.
31. J. Chen, B. F. M. Kuster, K. van der Wiele, *Biomass Bioenergy* **1** (1991) 217 – 223.
32. K. Takeshita, A. Suga, G. Takada, K. Izumori, *J. Biosci. Bioeng.* **90** (2000) 453 – 455. Links
33. A. E. Abasaeed, Y. Y. Lee, *Chem. Eng. Technol.* **18** (1995) 440 – 444. Links
34. J. J. Rumessen, *Scand. J. Gastroenterol.* **27** (1992) 819 – 828. Links
35. Raffinerie Tirlemontoise, EP 0 822 262, 1997 (L. De Leenheer, K. Booten).
36. *Sweetener Analysis*, Landell Mills Commodities Studies, Oxford, England, July 1993.
37. B. Ahlgren, *Int. Sugar J.* **99** (1997) 31 – 34. Links
38. D. Paillat, M. Cotillon, Starch Convention, Detmold, Germany, 1999.
39. H. Paananen, F. Rousset, *Zuckerindustrie* **126** (2001) 601 – 604.
40. M. Saska, S. J. Clarke, M. D. Wu, K. Iqbal, *Int. Sugar J.* **93** (1991) 223 – 228. Links
41. M. Rangarajan, B. S. Hartley, *Biochem. J.* **283** (1992) 223 – 233. Links
42. D. Vranesic, Z. Kurtanjek, A. M. P. Santos, F. Maugeri, *Food Technol. Biotechnol.* **40** (2002) 67 – 73
43. Y. D. Chu, L. D. Shiau, K. A. Berglund, *J. Cryst. Growth* **97** (1989) 689 – 696. Links
44. M. Meurer, U. Altenhöner, J. Strube, A. Untiedt, H. Schmidt-Traub, *Starch/Stärke* **48** (1996) 452 – 457.
45. H. Atiyeh, Z. Duvnjak, *J. Chem. Technol. Biotechnol.* **76** (2001) 1017 – 1022. Links
46. Südzucker, DE 1 010 6163, 2001 (D. Engels, A. Haji Begli, M. Kunz, R. Mattes, M. Munir, M. Vogel).
47. M. J. Tatlock: “Crystalline Fructose” in J. M. Dalzell (ed.): *Ingredients Handbook Sweetener*, Leatherhead Food RA, 1996.
48. Codex Standard for Sugars, Codex Stan 212 – 1999 (Amd.1-2001).
49. Food Chemicals Codex, 4th ed. Nat. Acad. Press, Washington DC (1996) 162 – 163.
50. Council Directive 2001/111/EEC of 20 December 2001 Relating to Certain Sugars for Human Consumption, Official Journal L 10/54, 12/01/2002.
51. The United States Pharmacopeia XXVI, 26th revision, 2002.
52. Y. C. Lee, *J. Chromatogr. A* **720** (1996) 137 – 149. Links
53. K. B. Hicks, *Adv. Carbohydr. Chem. Biochem.* **46** (1988) 17 – 72. Links
54. “HPLC Determination of Carbohydrates in Foods” In: L. M. L. Nolled (ed.): *Food Analysis by HPLC*, Marcel Dekker.,

New York, 1992, "Chap. 8".

55. K. Robards, M. Whitelaw, *J. Chromatogr.* **373** (1986) 81 – 110. Links
56. H. Jork, W. Funk, W. Fischer, H. Wimmer: *Dünnschichtchromatographie Reagenzien und Nachweismethoden*, 1. Aufl., VCH, Weinheim, Germany, 1989.
57. K. W. Swallow, N. H. Low, D. R. Petrus, *J. Assoc. Offic. Anal. Chem.* **74** (1991) 341 – 345.
58. M. Jochum, R. Bakry, J. Wartusch, C. W. Huck, H. Engelhardt, G. K. Bonn, *Chromatographia* **56** (2002) 263 – 267. Links
59. N. H. Low, W. South, *J. AOAC Int.* **78** (1995) 1210 – 1218. Links
60. N. H. Low, *J. AOAC Int.* **84** (2001) 486 – 492. Links
61. L. E. Rodriguez-Saona, F. S. Fry, M. A. McLaughlin, E. M. Calvey, *Carbohydr. Res.* **336** (2001) 63 – 74. Links
62. K. Zürcher, H. Hadorn, *Dtsch. Lebensm.-Rundsch.* **72** (1976) 197.
63. Verein der Zuckerindustrie (Hrsg.): *Analytische Betriebskontrolle der Zuckerindustrie*, Verlag Dr. Bartens, Berlin 1978, A 44 ff, B1.3.1, C2.R.
64. Richtlinie der EWG Kommission zur Festlegung gemeinschaftlicher Analysenmethoden für die Kontrolle von zur menschlichen Ernährung bestimmter Zuckerarten (79/796/EWG), Anlage II, Methode 6, Methode 7.
65. *Handbuch der Lebensmittelchemie*, Springer Verlag, Berlin 1967, Bd. II/2, S. 355.
66. JECFA 5 Rev. 1, Rome 1983; guide to specifications, p. 87.
67. Official Methods of Analysis of the Association of Official Analytical Chemists (1997), 16th ed., 3rd rev., vol. 2, chap. 28.1.17, p. 5 (AOAC Official Method 985.09).
68. Europäische Norm/European Standard EN 1140 (Dec. 1994) Fruit and Vegetable Juices; Enzymatic Determination of D-Glucose and D-Fructose Content.
69. T. F. Osberger: “Crystalline Fructose” in L. O'Brien Nabors, R. C. Gelardi (eds.): *Alternative Sweeteners*, 2nd ed., Marcel Dekker, New York, 1996.
70. M. E. Daly, C. Vale, M. Walker, K. G. M. M. Alberti, J. C. Mathers, *Am. J. Clin. Nutr.* **66** (1997) 1072 – 1085. Links
71. Council Directive 90/496/EEC of 24 September 1990 on nutrition labelling for foodstuffs, Official Journal L 276, 06/10/1990, 0040 – 0044.
72. J. Gray, *Carbohydrates: Nutritional and Health Aspects*, ILSI Europe Concise Monograph Series, ILSI Press, Brussels, 2003.
73. P. A. Mayes, *Am. J. Clin. Nutr.* **58** (1993) suppl. 754 – 765.
74. R. R. Henry, P. A. Crapo, *Annu. Rev. Nutr.* **11** (1991) 21 – 39. Links
75. J. E. Schmitz, A. Grünert, F. W. Ahnefeld, *Aktuel. Ernährnungsmed*. **14** (1989) 253 – 262.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Fuel Cells

Batteries and Electrochemistry are separate keywords

Marcella Cappadonia, Technische Universität München, Physics Department, Garching, Germany

Ulrich Stimming, Technische Universität München, Physics Department, Garching, Germany

Karl V. Kordesch, Technische Universität Graz, Graz, Austria (Section 2.4)

Julio Cesar Tambasco de Oliveira, Technische Universität Graz, Graz, Austria (Section 2.4)

1. Introduction 1
2. Principles of Fuel Cells 2
2.1. Functions 2
2.2. Set-Up 4
2.3. Efficiency 6
2.4. History and Development 7
3. Classification of Fuel Cell Systems . 8
3.1. Low Temperature Fuel Cells 9
3.1.1. Alkaline Fuel Cells (AFC) 9
3.1.2. Proton Exchange Membrane Fuel Cells (PEMFC) 11
3.1.3. Phosphoric Acid Fuel Cells (PAFC) . 14
3.2. Intermediate Temperature Fuel Cells 15
3.3. High Temperature Fuel Cells 15
3.3.1. Molten Carbonate Fuel Cells (MCFC) 15
3.3.2. Solid Oxide Fuel Cells (SOFC) 17
4. Examples of Application of Fuel Cells 21
4.1. Zero Emission Vehicle 21
4.2. PAFC Power Plants 21
4.3. Regenerative Systems 22
5. References 22

1. Introduction

Energy conversion systems are the principal source of pollution on our planet. In the industrialized countries, the electric power plants which convert chemical energy in electrical energy are the main producers of CO_2 and SO_2 emissions, private cars, on the other hand, are the principal producers of CO and NO_x emissions. Figure 1 illustrates this situation for an industrialized country such as Germany (given for Western Germany in 1990) [1]. The concentration of CO_2 in the earth's atmosphere increased considerably since the last century as a consequence of industrialization [2] (see Fig. 2); it is considered the main reason for the greenhouse effect leading to an increase of the earth's surface temperature. Power plants and cars are important consumers of mineral oils and natural hydrocarbons, which have a higher potential as a raw material for the chemical industry than as fuel for, e.g., the production of electricity. Nowadays the air pollution and the diminution of the earth's resources of hydrocarbons are two main topics in politics and economics and regenerative energy sources are sought after. An alternative energy conversion technology such as the fuel cell technology offers many attractive possibilities for reducing both the air pollution and the indiscriminate destruction of precious chemicals. Fuel cells convert the chemical energy of a fuel such as methane directly into electrical energy with a better yield and much lower emission levels than any other energy conversion system [3], [4]. Although the principal operation of a fuel cell is known since the last century, the necessary technology for producing fuel cell systems of high efficiency, acceptable costs and at an industrial level has only been developed in the last few decades.

2. Principles of Fuel Cells

2.1. Functions

The principle operation of a fuel cell is comparable to that of a battery. In contrast to batteries – where the chemical energy is stored in substances inside the battery – fuel cells are just converting systems, the reagents have to be supplied continuously to the fuel cell in order to obtain electricity. Thus, fuel cells are systems which convert chemical energy directly into electricity in an invariant electrochemical set-up.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_055

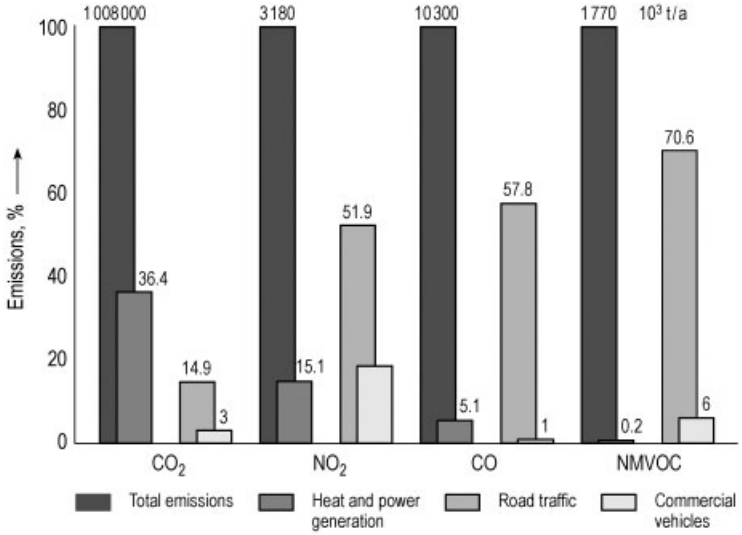

Figure 1. Emissions generated in Western Germany in 1990 with emphasis on power plants and road traffic in relation to overall emissions NMVOC = Non-methane volatile organic compounds

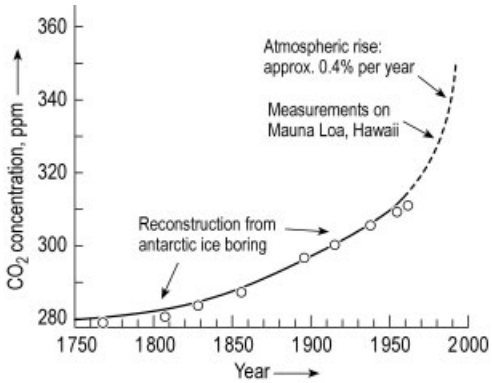

Figure 2. CO_2 concentration in the earth's atmosphere as a function of time [2]

Figure 3 schematically shows the principle of a fuel cell with a proton conducting electrolyte, hydrogen as fuel and oxygen as oxidant. Hydrogen and oxygen are supplied to porous electrodes, the gases reach the electrode – electrolyte interface and react such that water and an electrical current are produced.

A comparison between a fuel cell and a battery is shown in Figure 4. Batteries, both primary cells and rechargeable batteries, are best suited for smaller power requirements while fuel cells are more suited for larger power needs. In the 0.1 – 10 kW range an application of both systems may be possible depending on the specific conditions.

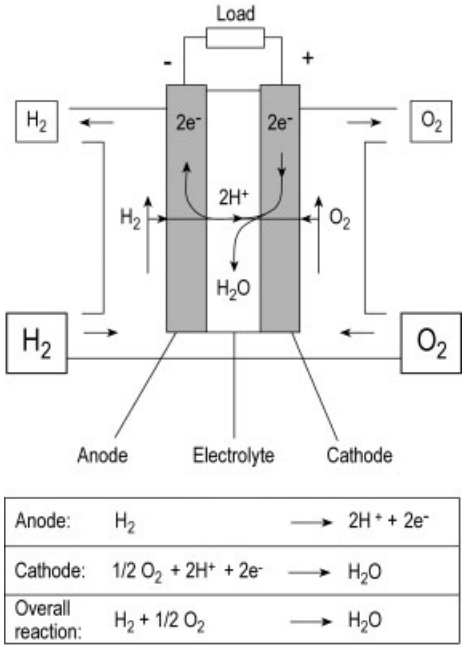

Figure 3. Sketch of the principal operation of a fuel cell: Two electrodes are separated by an gastight electrolyte. The fuel (H_2) is supplied at the anode, where it is oxidized releasing electrons. The oxidant (O_2) is supplied at the cathode, where it is reduced consuming electrons.

Because of the direct energy conversion fuel cells work at a higher efficiency than energy conversion processes which involve a conventional heating cycle. The latter are, therefore, limited

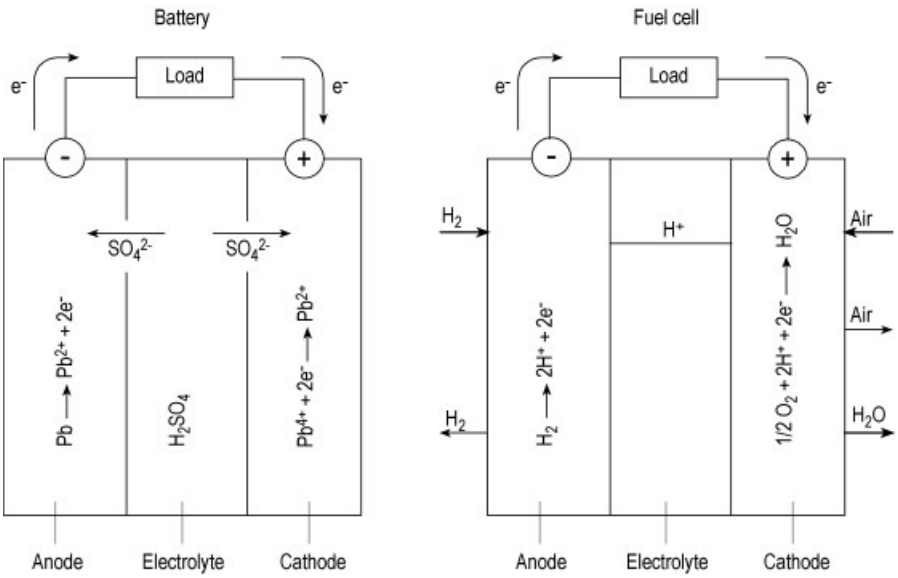

Figure 4. Comparison between two electrochemical devices – battery and fuel cell – for electricity production. A battery already contains the reagents, which have to be supplied to a fuel cell from an external reservoir.

by the Carnot's cycle. Figure 5 shows a comparison of the conversion routes of fossil fuels to electricity in a fuel cell and in a conventional system such as a heat engine. One obvious aspect is that fewer steps are involved in the fuel cell process than in the conventional process. In addition, the conventional process exhibits a lower conversion efficiency due to thermodynamic limitations. In a conventional process the chemical energy of fossil fuels is converted in electrical energy after several steps comprising a chemical reactor, mechanical engines and electrical generators, each step causes energy losses; a fuel cell can directly convert chemical energy in electrical energy, minimizing the losses.

This is illustrated in Figure 6 which compares the theoretical efficiencies as a function of temperature of a fuel cell supplied with pure hydrogen and pure oxygen and a Carnot process as the most efficient heat cycle operating between a heat reservoir at T and one at 300 K. Only at temperatures higher than 1000 °C is the efficiency of a Carnot process higher than that of a fuel cell. However, practical efficiencies of fuel cells remain higher even if the thermal cycles operate at much higher temperatures.

The efficiencies of the different energy conversion systems are compared in Figure 7 as a function of size of power plants. Figure 7 shows that the efficiencies of two types of fuel cell systems (phosphoric acid fuel cell, PAFC, see Section 3.1.3 and solid oxide fuel cell, SOFC, see Section 3.3.2) are higher than those of engines and conventional power plants of comparable size.

In addition to the high efficiency, fuel cells also exhibit an emission level, which is notably lower than present legal requirement, in Germany for instance, stipulated by "TA Luft". Fuel cells can convert up to 90 % of the energy of the fuel into electricity and heat, the electricity conversion efficiency is 45 % and higher, and the ratio between the electricity and the heat produced is about 1. The heat generated by a fuel cell can be used for producing hot water or steam. The highest temperature reached in a fuel cell is about 1000 °C which is the operating temperature of a solid oxide fuel cell. Even at this temperature, highly polluting nitrogen oxides are not produced, CO_2 is produced in reduced quantities, ashes and solid combustion products are not formed. Finally fuel cells themselves do not have moving mechanical parts, therefore they do not make noise.

2.2. Set-Up

The electrodes of a fuel cell exhibit a porous structure; through the pores fuel and oxidant

reach the so-called three-phase zones, where the reagents get in contact with both the catalyst deposited on the electrodes and the electrolyte. The reactions occur in these regions. Referring to Figure 3, the fuel (H_2) is oxidized at the anode, to form protons which move through the gas-impermeable electrolyte. At the cathode oxygen is reduced and water is produced, while electrons flow in an external circuit. Water is the only byproduct of the energy conversion in a hydrogen – oxygen fuel cell. This water is produced at the cathode in a fuel cell with proton-conducting electrolytes such as a proton exchange membrane or phosphoric acid, and at the anode in fuel cells with an alkaline electrolyte or an oxygen-ion conducting electrolyte such as yttrium stabilized zirconia.

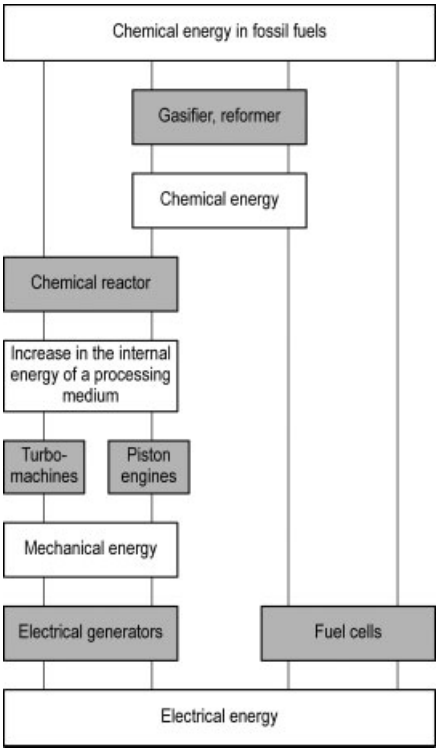

Figure 5. Comparison between the two conversion routes of fossil fuel into electric energy. The chemical energy of the fuel is directly converted into electrical energy by fuel cells, in conventional systems the chemical energy of the fuels is converted first into thermal energy, than in mechanical energy and finally into electrical energy.
In the reformer hydrocarbons or methanol are converted to CO and H_2, according to $C_nH_m + n\,H_2O \longrightarrow n\,CO + (m+n/2)\,H_2$, or $CH_3OH + H_2O \longrightarrow CO_2 + 3\,H_2$

Figure 3 shows only the electrochemical core of a fuel cell. However, a single electrode – electrolyte assembly is not sufficient for electricity generation, since the typical voltage of a single cell is below 1 V. In order to obtain suitable voltages a number of cells are serially connected. Thus, in addition to the electrochemical components a fuel cell needs interconnector plates, which connect electrically two adjacent electrochemical units and usually also act as the reactant distributor. Several fuel cells connected by interconnectors form a stack. Such a stack is shown in Figure 8 for a fuel cell with polymer electrolyte.

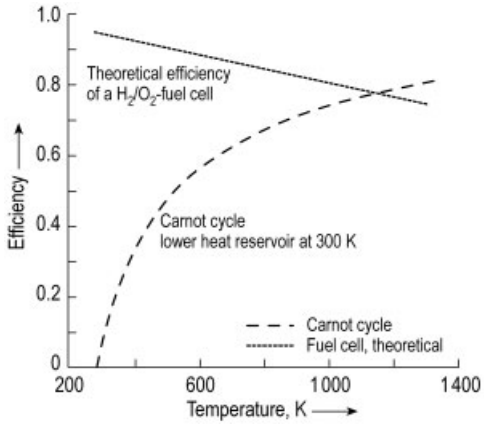

Figure 6. Efficiency as a function of temperature of the energy conversion of a fuel cell and a conversion process limited by the Carnot's factor.

The supply of pure hydrogen and pure oxygen is not easy to realize, since no distribution system exists for these pure substances. For some niche applications, such as storage of regenerative energy, reactants in the form of pure hydrogen and oxygen may be available. In usual applications, a fuel cell should work with air as oxidant and with hydrogen produced in or near the system by reforming of other fuels. Typical reactions are reforming of hydrocarbons (e.g., methane) or alcohols (e.g., methanol) (see Section 3.1.2). Thus, a reformer is needed in the periphery of a fuel cell system. Operation with pure chemicals involves the advantages of working at low temperature and almost problem-free operation of the catalytic materials of the electrodes, whose performance depends strongly on the presence of impurities in the reagents. In recent years many efforts have been put into the

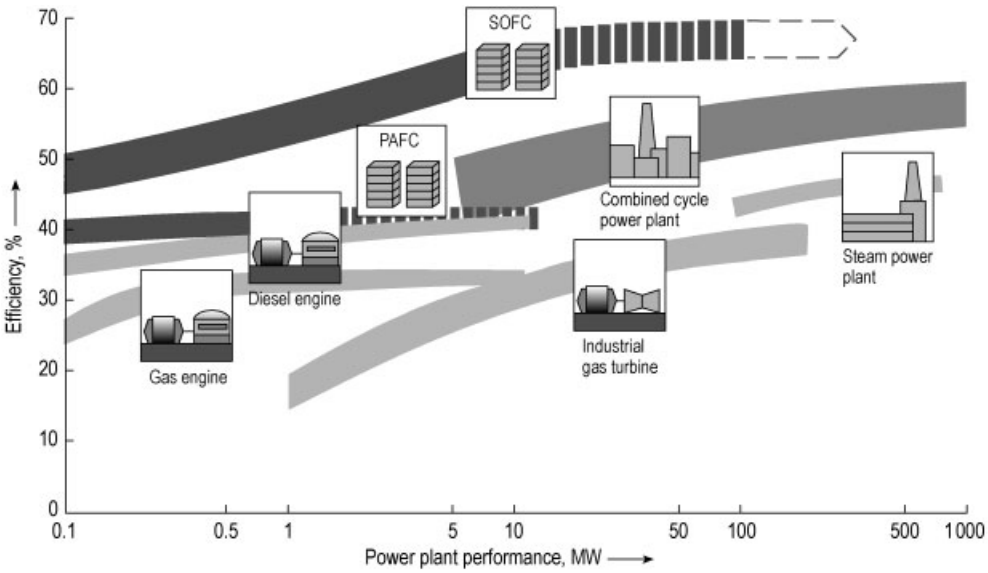

Figure 7. Efficiency of energy conversion as a function of the size of the power plant; conventional systems are compared with projected fuel cell systems such as PAFC and SOFC

improvement of the electrocatalysts and into the gas separation process after reforming in order to use cheaper and more common fuels and air in the fuel cell. A catalytic burner for the combustion of unreacted gases and a d.c./a.c. convertor also belong to the periphery of a fuel cell system, a more complete picture of a system for stationary electricity and heat generation is given in Figure 9.

Figure 9 shows a complete fuel cell system for stationary production of electricity and heat. Before being supplied to the fuel cell the fuel is compressed, heated, chemically processed in presence of water vapor (for example by reforming), the air is compressed and preheated. The residual gases after exploitation in the fuel cell stack for energy production are burned in a catalytic burner, which supplies the necessary heat for preheating and reforming. The residual heat is eventually removed from the system.

The materials used in a fuel cell must exhibit properties which can be generally summarized as:

The *electrolyte* has to have a high conductivity at operating temperature, i.e. it should cause a low ohmic drop ($< 0.15\ \Omega/\text{cm}^2$). It should have negligible electronic conductivity, be gas impermeable, and chemically stable in an oxidative and reductive environment (1.2 V and 0.0 V, respectively).

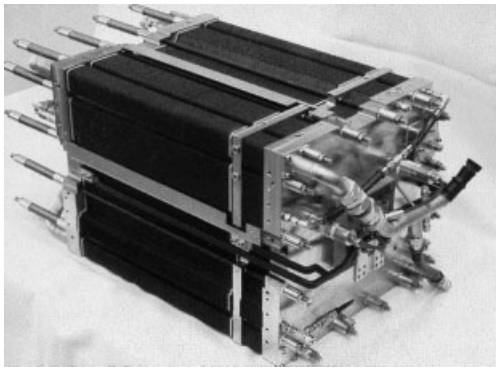

Figure 8. A Siemens PEMFC stack

Rated power	40 kW
Voltage at rated power	109 V
Open circuit voltage	160 V
Rated current	350 A
Size (without tierals)	$41 \times 41 \times 56\,\text{cm}^3$
Volume	94 L
Weight	280 kg
Number of cells	160
Integrated H_2 humidifier	

The *electrodes* must have a high electrocatalytic activity for fuel oxidation or oxidant reduction, a high electrical conductivity (elec-

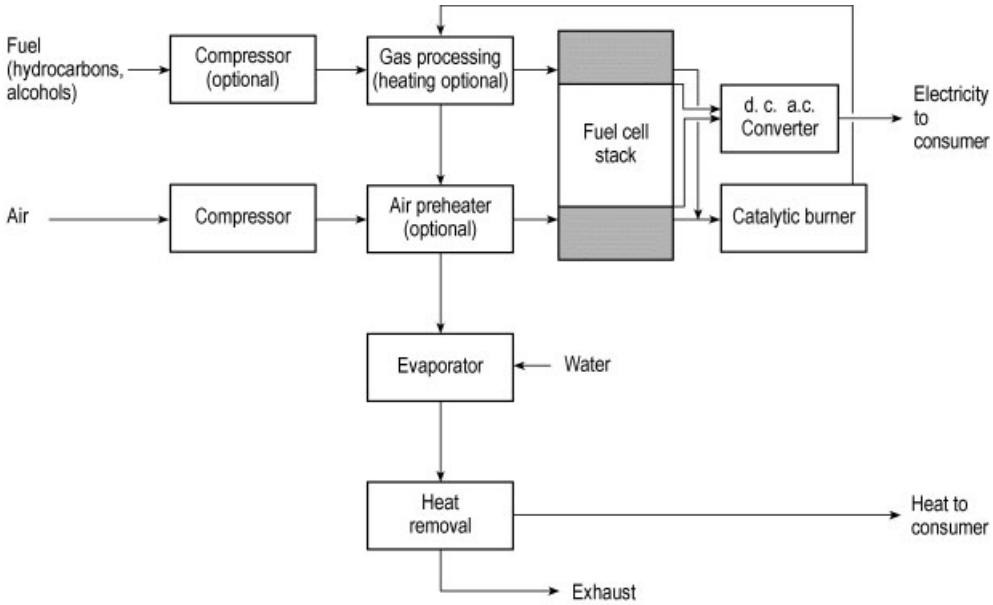

Figure 9. Fuel cell system for stationary production of electricity and heat
A complete fuel cell system comprises besides the electrochemical device other components such as gas compressors, reformers, catalytic burner, and d.c./a.c. converter.

tronic or mixed ionic and electronic), and must be chemically stable to reduction or oxidation and gas permeable.

The *interconnector* must have a high electronic conductivity and must be stable in oxidative environment on the cathode side and reductive environment on the anode side.

The *electrochemical system* interconnector – anode – electrolyte – cathode must exhibit high thermal stability to sufficiently withstand frequent thermal cycles, have a long operation life (5000 – 150 000 h, depending on application), be light, compact, not expensive, and easy to produce. In order to reach these targets, the design of the cell is of crucial importance, in fact, it can strongly affect the production costs and the efficiency of a fuel cell system. An important factor for the design is the dimension of the three-phase zones, where the electrochemical reactions take place. The three-phase zone has a thickness of just a few micrometers, therefore cell designs have to consider the design of the components (electrodes, electrolytes, and interconnectors) in such a way that the specific power density is increased by minimizing the volume of the stack for the same volume of the reaction zone.

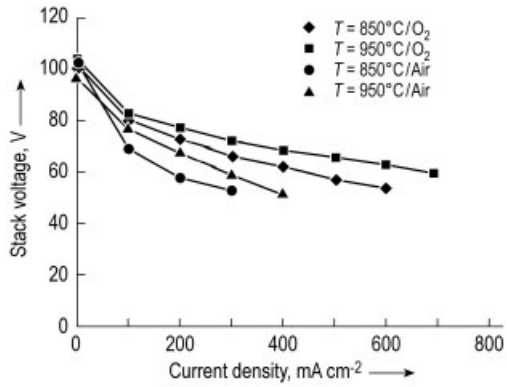

Figure 10. Examples of current density – voltage curves for a solid oxide (see Section 3.3.2) fuel cell stack as a function of temperatures. These curves are measured directly. Parameter list: Interconnector $CrFe_5Y$; Anode Ni cermet ($Ni + ZrO_2$); Cathode lanthan strontium manganite (LSM); Electrolyte YSZ ($ZrO_2 - Y_2O_3$); Fuel H_2 (1 bar); Oxidant O_2/air (1 bar); stacks of 80 cells in serial connection

The usual way to characterize a fuel cell is to record the current density – voltage curve of single cells and stacks under different conditions of temperature, catalyst loading, reagent composition, concentration and pressure and to study how the curves vary with time in order to determine the possible operation time of cells

and stacks. From the current density – voltage curves the current density – power curve can be obtained which usually serves to compare different systems.

Examples of current density – voltage curves for a solid oxide fuel cell are given in Figure 10.

2.3. Efficiency

Another important parameter of a fuel cell is the efficiency of the conversion of the chemicals to electricity. This *electrical efficiency* is given by the product

$$\eta_{\mathrm{el}} = \eta_{\mathrm{id}} \ \eta_{\mathrm{U}} \ \eta_{\mathrm{F}} \ \mathrm{U}_{\mathrm{F}} \tag{1}$$

where the ideal efficiency $\eta_{\mathrm{id}} = \Delta G/\Delta H$ is the thermodynamic efficiency and corresponds to the total conversion of the free energy change, ΔG, to electrical energy in reference to the heating value of the fuel, ΔH.

The *voltage efficiency* $\eta_{\mathrm{U}} = U_{\mathrm{cell}}/U_{\mathrm{theoretic}}$ reflects voltage losses due to polarization phenomena at the electrodes and ohmic potential drops at the interfaces, in the electrolyte, and in the interconnectors (the theoretic cell voltage for water production is 1.23 V, voltages between 0.6 and 0.9 V are usually observed in real systems depending on current densities).

The *faradaic efficiency* $\eta_{\mathrm{F}} = I/I_{\mathrm{m}}$ is defined as ratio between the observed current and the theoretically expected current calculated on the basis of complete reaction. The faradaic efficiency is lower than 1 if a parallel reaction or chemical short circuit occur in the cell.

U_{F} is the utilization of the fuel in the cell or stack.

The electrical efficiency of the complete fuel cell system comprises also the efficiency of the periphery and is given by

$$\eta_{\mathrm{el}}^{\mathrm{system}} = \eta_{\mathrm{el}}^{\mathrm{stack}} \ \eta_{\mathrm{P}} \tag{2}$$

2.4. History and Development

The fuel cell history dates back to the 19th century. It began in 1839 with the experiments of the British scientist Grove, which were at that time just a laboratory curiosity [5]. Grove electrolyzed sulfuric acid using two platinum electrodes and noticed that the gases evolved (hydrogen and oxygen) were electrochemically active and provided a potential difference of ca. 1 V between the electrodes. Later Ostwald (1894) realized, on the basis of thermodynamic considerations, that an electrochemical cell can provide a higher energy conversion efficiency than heat engines, which are limited by Carnot's law [6].

At the turn of the century, eminent physicochemists like Nernst and Haber dedicated a lot of effort to the study of direct carbon-oxidizing fuel cells. The difficulty in obtaining appropriate materials at that time limited the success of their investigations. Baur proposed (for the first time in 1933) an electrochemical system working at room temperature, with alkaline electrolyte, using hydrogen as fuel [7]. Baur's proposal caught the attention of Bacon, a British engineer [8]. He successfully developed porous nickel electrodes in the 1950s and built the first high pressure, medium temperature alkaline fuel cell system in the kilowatt-range. Bacon's work was the basis of the later developed fuel cell for the NASA Apollo program.

Highly active porous metal electrodes with a dual porous structure like the Bacon electrodes were produced in the mid-1960s by Justi, Bishoff, Spenger, and Winsel in Germany, at the Technical University of Braunschweig [9]. These electrodes operated already at room temperature and at atmospheric pressure, and were based on pressed Raney nickel and Raney silver powder. The hydrogen and oxygen electrodes later developed at Varta and Siemens are based on this work.

In 1958, Broers [10] and Ketelaar, in Holland began to improve the high temperature molten carbonate fuel cell and achieved a longer life expectancy. Their research was based on the far earlier studies of Baur, Preiss (1938) and Davtyan (1946). Broers' work led to the present-day power plant systems which can oxidize hydrocarbon fuels directly, without a reforming step.

Aiming at low temperature air operation with low-cost carbon materials, other scientists in the 1950s continued the studies of Schmid, about gas diffusion electrodes (1923), of Berl (1943), about the peroxide mechanism of carbon-oxygen-electrodes and the work

on carbon – air electrodes by HEISE and SCHUMACHER (1932 – 1947).

KORDESCH and MARKO (University of Vienna) prepared special metal oxide catalysts (cobalt – aluminum spinels) for carbon electrodes (1951) [11]. The carbon tubes or plates were steam-activated to enlarge the active surface. The carbon – hydrogen electrodes contained small amounts of a platinum catalyst and could oxidize fuels such as alcohols and aldehydes directly at low temperature. These new carbon electrodes provided high current densities in alkaline electrolytes, at low temperatures, confirming BAUR's predictions of 1933.

Fuel cells experienced worldwide attention in the 1950s and the 1960s because of the "space race". Hydrogen – oxygen fuel cells were found to be the best choice for manned space flights not only because of their high energy density, but also because the reaction product is pure water that can be used by the crew for drinking and for cooling purposes.

After a short period of decline in the number of space projects, a new wave of fuel cell research and development (R & D) programs started in 1973 because of the oil crisis. Since then, investments in fuel cell projects have increased steadily, with more efficient energy conversion schemes and strong environmental control being demanded.

During the 1980s fuel cell R & D programs have spread out in many countries around the world. Large efforts are dedicated to the development of the technology of the five technically most promising systems: the alkaline, phosphoric acid, molten carbonate, solid oxide and the proton exchange membrane system (names referring to electrolytes).

3. Classification of Fuel Cell Systems

Fuel cell systems can be divided in two groups depending on the operation temperature. Low temperature fuel cells operate at a temperature lower than 250 °C, while high temperature fuel cells operate at a temperature above 600 °C. The fuel cell systems are usually classified and named after the electrolyte present in the cell, with one exception: the direct methanol fuel cell (DMFC) which is identified by the fuel used. Except for some alkaline fuel cells, a common characteristic of fuel cells is the use of an immobilized electrolyte, which means that the electrolyte is solid or consists of a liquid retained on a solid matrix. Immobilized electrolytes offer many practical advantages such as fewer corrosion problems, no leakage of the liquid, easier design, planning and construction. An overview of the material used for the components of the different fuel cell system is given in Table 1.

The other parameters such as reagents, temperature, pressure and application of the fuel cell system depend on the electrolyte in the cell. Table 2 lists these parameters for different types of fuel cells.

Several fuels can be used; depending on the type of systems they can be directly supplied to the anode or first undergo a reforming reaction. The thermodynamic efficiency of the combustion of different fuels are given in Figure 11.

The temperature dependence of the theoretical efficiencies $\Delta G^{o}/\Delta H^{o}$ of various reactions are plotted on the right scale. The corresponding cell voltages calculated from the Nernst equation are shown as well (left scale).

3.1. Low Temperature Fuel Cells

3.1.1. Alkaline Fuel Cells (AFC)

Alkaline fuel cells were the first type of fuel cells produced as a commercial product. The first alkaline fuel cell was build by BACON in the mid 1950s, it was a 5 kW system with 30 % aqueous KOH as electrolyte operating at 200 °C and 5 MPa, reactants were pure gases [8]. NASA used slightly modified Bacon cells for the Apollo program [12], the technical features of the Apollo fuel cells were: 85 % KOH as electrolyte, 230 °C operating temperature, 0.33 MPa pressure, pure gases as reagents, stack power 1.5 kW, and up to 11 000 h operation at the mission end. NASA improved the AFCs for the space shuttle program. The stack power was increased to 12 kW, the operating temperature lowered to 90 °C. They operated with pure gases and a specific power of 275 W/kg. Nickel plates were used as interconnectors, Pt – Pd mixtures were employed as anodic, Au – Pt mixtures as cathodic catalysts. Lifetimes up to 15 000 h were observed. In the 1970s many research projects on AFCs involved companies such as Union

Table 1. Materials used for the components of different fuel cell systems

Fuel cell	Electrolyte	Electrodes (anode/cathode)	Interconnector	Construction materials
Alkaline fuel cell (AFC)	conc. KOH in porous matrix	Ni/Ag (Pt/Pt) on porous Ni	Nickel	graphite, Inconel
Proton exchange membrane fuel cell (PEMFC)	proton exchange membrane	Pt on graphite	graphite sheet metal (stainless steel)	steel
Phosphoric acid fuel cell (PAFC)	conc. phosphoric acid in porous matrix	Pt/Pt	graphite	Inconel
Molten carbonate fuel cell (MCFC)	molten Li and K carbonate in porous matrix	Ni/NiO	Inconel	Inconel, Ni, ceramic
Solid Oxide fuel cell (SOFC)	yttrium stabilized zirconia ZrO_2 (Y_2O_3)	Ni cermet, $SrLaMnO_3$, other perovskites	chromium based alloys or chromium based ceramics	ceramic, Ni and Fe alloys

Table 2. Operating parameters of different types of fuel cells

Characteristic	AFC	PEMFC	PAFC	MCFC	SOFC
Temperature	60 – 90 °C	50 – 90 °C	160 – 220 °C	620 – 660 °C	800 – 1000 °C
Fuel	pure hydrogen	pure hydrogen, reformate ∗	pure hydrogen, reformate	natural gas, reformed or directly fed, biogas, coal gas	natural gas, reformed or directly fed, biogas, coal gas
Oxidant	pure oxygen	pure oxygen or air	air	air	air
Application	space and military	space, military, automotive, and stationary	cogeneration power plant	cogeneration or combined cycle power plants, depending on size	
System power at present		50 kW – 250 kW	11 MW	2 MW demonstration plant	100 kW demonstration plant
Electrical efficiency					
Stack	69 – 70 %	50 – 68 %	55 %	65 %	60 – 65 %
System	62 %	43 – 58 %	40 %	54 % (cogeneration) 60 – 65 % (combined cycle)	> 50 % (cogeneration) 65 – 70 % (combined cycle)

∗ Produced by reforming of hydrocarbons or methanol.

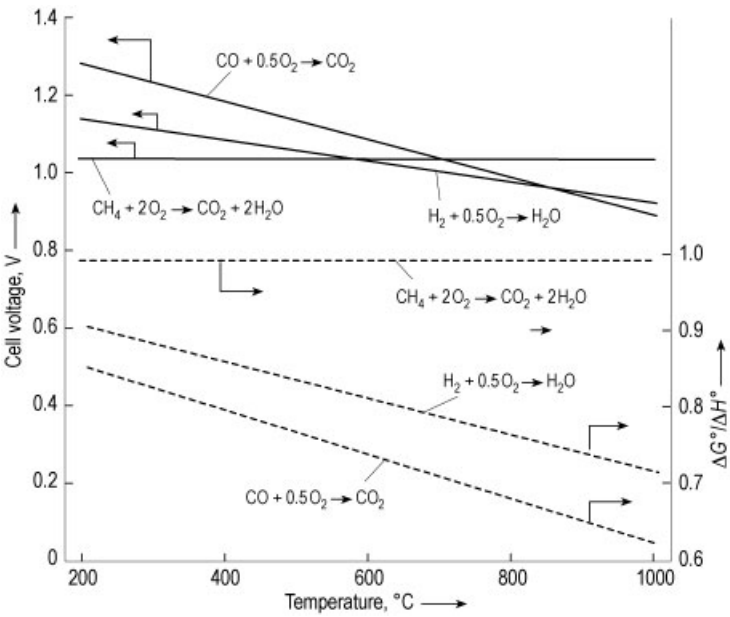

Figure 11. Thermodynamic equilibrium voltages and theoretical efficiencies for the combustion of different fuels in a solid oxide fuel cell

Carbide [13], Varta [14], Siemens [15], Elenco [16], and universities such as Graz [17], [18]. Siemens built an AFC with a power density of 12 kg/kW and current density of 0.4 A/cm^2 at 0.78 V [15]. The newest ACFs use immobilized electrolytes consisting of a porous matrix impregnated with 7 mol/L KOH; they work at temperatures $< 150\,°C$ [19].

The operating principles of an AFC comprise the following reactions:

$$H_2 + 2\,OH^- \rightarrow 2\,H_2O + 2\,e^-$$

(oxidation of the fuel at the anode)

$$1/2\,O_2 + 2\,e^- + H_2O \rightarrow 2\,OH^-$$

(reduction of the oxidant at the cathode)

The mobile ions are the OH^- ions of the alkaline solution which move from the cathode to the anode. The water is produced at the anode, it can be removed together with unreacted hydrogen or with the circulating electrolyte depending on stack design.

Reagents. The AFCs work only with pure reagents, the presence of impurities and nitrogen reduces the electrochemical efficiency of the cell. The presence of CO_2 which reacts easily with the alkaline electrolyte, to form carbonates may be detrimental for the cell, resulting in occlusion of the pores of the gas diffusion electrodes.

Electrodes. The AFCs use porous carbon electrodes, platinum is a typical catalyst for both the anodic and cathodic reaction, but also Ni and Ag are used for the anodic and the cathodic reactions, respectively [20]. The kinetics of both reactions, but especially of oxygen reduction, is favorable in the basic environment, the charge transfer overpotentials for the pure gases are less than 0.1 V for both electrode reactions.

Nickel is used as material for the *interconnectors*.

The operation temperature of an AFC depends on the electrolyte concentration but it is usually below 100 °C. The possibility to work at low temperatures and the use of pure gases allows the use of cheaper materials for electrodes and periphery. The pure gases allow for a dead-end configuration, where no gas outlets at the electrodes are necessary.

Although AFCs exhibit the highest electrical efficiency (60 %) of all fuel cell systems, their applications are limited to space and military programs, where high costs are permissible and pure hydrogen and pure oxygen can be supplied. The small regular amounts of CO_2 contained in air are sufficient to cause problems in the operation of a AFC, thus, air must be avoided. In spite of several attempts, operation with air was not successful, since even traces of CO_2 cause carbonate formation in the pores.

3.1.2. Proton Exchange Membrane Fuel Cells (PEMFC)

The PEMFCs are also known as *polymer electrolyte fuel cells* (PEFC), *ion exchange membrane fuel cells* (IEMFC) and *solid polymer electrolyte fuel cells* (SPEFC). The first prototype of PEMFC was produced by General Electric (GE) in 1963 [21]. A 1 kW system was realized for the Gemini mission of NASA with a sulfonated polystyrene membrane as electrolyte, the cell was operated with pure gases. In 1969 Nafion (Du Pont) was used as electrolyte in a GE fuel cell for the first time [22]. In the 1980s Ballard Power System began with a project on PEMFC and also used a proton exchange membrane produced by DOW Chemical [23]. A 5 cell stack using air as oxidant was presented in 1988; similar systems of varying power were used by Ballard [24] and by Daimler-Benz [25] in demonstration projects for the traction of buses and passenger cars since 1993.

The operating principles of a PEMFC comprise the following reactions:

$$H_2 \rightarrow 2\,H^+ + 2\,e^-$$ (oxidation of the fuel at the anode)

$$1/2\,O_2 + 2\,e^- + 2\,H^+ \rightarrow H_2O$$

(reduction of the oxidant at the cathode)

The mobile ions are the H^+ which move from the anode to the cathode, the water is produced at the cathode. Water is removed together with nitrogen and unreacted oxygen, if air is used as oxidant (see Fig. 12). The PEM single cell is often called *membrane – electrode assembly* (MEA).

Reagents. Air is generally used as oxidant, pure oxygen can also be used, with the possibility to have a dead-end configuration. The PEMFCs have been proposed for using hydrogen obtained from reforming of organic fuels

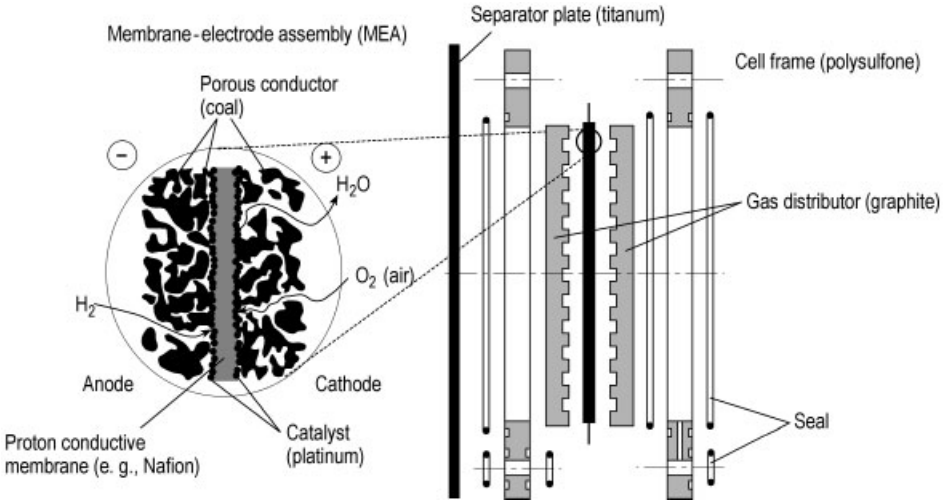

Figure 12. Proton exchange membrane fuel cell: the membrane – electrode assembly (MEA) consists of the Nafion membrane with the electrocatalyst on the surface in contact with porous carbon electrodes. Each MEA is mounted between two gas manifolds in an electrical insulating MEA frame and it is separated by an interconnector plate from the adjacent MEAs.

since the membrane is insensitive to CO_2. [The main problem in using hydrogen from a reformer is the presence of CO, which blocks the active sites of the Pt catalysts even at low concentrations (> 5 ppm).]

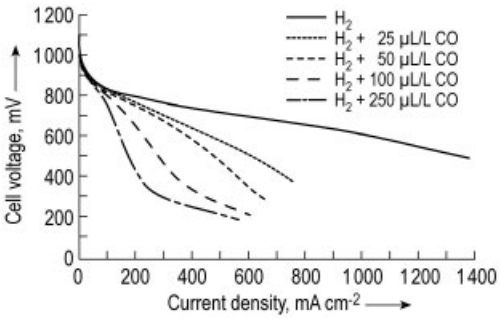

Figure 13. Current density – voltage curves of a PEMFC with Pt as electrocatalyst as a function of CO content of the fuel; the poisoning of the electrocatalyst is evident in the reduction of the voltage of the cell at constant current density. Anode: 1 mg/cm^2 Pt; Cathode: 1 mg/cm^2 Pt; Electrolyte: Nafion 117; $p(H_2)$: 2.2 bar (abs.); $p(O_2$:2.4 bar (abs.)

Electrodes. The porous electrodes consist of carbon cloths loaded with a mixture of platinum and Nafion. In order to achieve a good contact between the electrode, the catalyst and the electrolyte, the electrodes are pressed on the electrolyte membrane which acts as the supporting component. Other catalysts such as binary Pt – Ru mixtures and ternary systems such as Pt – Ru – Sn have been studied in view of a better CO tolerance [26]. Typical catalyst loadings are now below 0.5 mg/cm^2, but even lower loadings have been proposed [27]. The current density voltage curves of Figure 13 show the effect of CO on the performance of a cell with platinum as catalyst, Figure 14 shows the higher catalytic activity of Pt – Ru mixtures in presence of CO as compared with pure Pt. The positive effect of Ru is due to its ability to provide surface oxygen at low potentials; Figure 15 sketches the possible mechanism for the oxidation of CO on Ru which reduces the CO blocking of Pt and enables better hydrogen oxidation.

Graphite and stainless steel are used as materials for the *interconnectors*.

PEMFCs are characterized by the use of proton exchange membranes as electrolyte which are good electronic insulators and very good proton conductors in the presence of water.

Nafion (DuPont) has a fluorocarbon polymer backbone with fixed sulfonic groups (→ Fluorine Compounds, Organic, Chap. 8.1.4.). Nafion was developed in the 1960s by Grot as a separator membrane in chlor – alkali cells [22], it is until now the most frequently used membrane for this application, although it is expensive (and environmentally questionable because it is a fluorocarbon). Nafion membranes are produced as films with

thicknesses between 0.05 and 0.18 mm and different degree of sulfonation. They exhibit high mechanical and chemical stability and can be used as support for the porous electrodes, which can be applied by hot isostatic pressing onto the membrane [28]. A disadvantage of Nafion membranes is that their conductivity depends strongly on the water content of the membrane which is given by the ratio of number of water molecules per available number of sulfonic acid groups (see Fig. 16) [29]. If the membrane dries out, the ohmic drop becomes large and power and efficiency of the fuel cell decrease. The water content of proton exchange membranes depends on temperature, concentration of fixed charges, and the electro-osmostic drag, which causes the transport of water molecules with the proton migrating from the anode to the cathode during the operation of the fuel cell. Thus, the membrane can dry out on the anode side causing an increase in resistance of the electrical contact electrode – electrolyte because of the loss of ionic conductivity in the dry membrane. Water flooding of the pores can occur at the cathode. Both processes have to be avoided by a proper water management in the fuel cell, which comprises the use of humidified fuel gas.

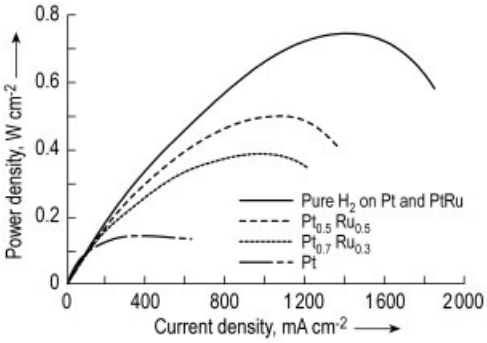

Figure 14. Performance plots of PEM single cells with Pt, $Pt_{0.7}Ru_{0.3}$, and $Pt_{0.5}Ru_{0.5}$ anodes. The performance of different electrocatalysts in presence of CO in the fuel are shown as current density – power curves. Anode: 1 mg/cm^2 Pt or PtRu; Cathode: 1 mg/cm^2 Pt; Electrolyte: Nafion 117; Fuel gas: H_2 + 100 ppmv CO; Temperature: 80 °C; $p(H_2)$: 2.2 bar (abs.); $p(O_2)$: 2.4 bar (abs.)

Current densities of more than 1 A/cm^2 at 0.5 V have been observed at PEMFC with hydrogen as fuel and air as oxidant, even higher current densities can be obtained with pure oxygen as oxidant [30].

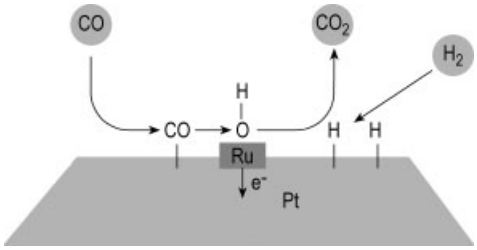

Figure 15. CO oxidation on a Pt – Ru catalyst. The presence of Ru in the catalyst composition causes a more facile oxidation of CO to CO_2 with removal of the adsorbed species from the catalytic sites.

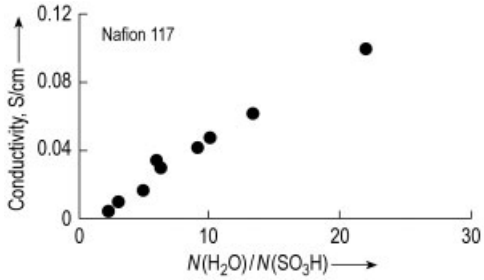

Figure 16. Conductivity of Nafion 117 as a function of the water content of the membrane given as number of water molecules per sulfonic group. The conductivity of Nafion increases linearly with the water content.

In order to provide hydrogen *the reforming of a carbon containing fuel* may be necessary. The reactions of the reforming of hydrocarbons for the production of hydrogen are:

$$C_nH_m + nH_2O \rightarrow nCO + (m + n/2)H_2$$

$$CO + H_2O \rightarrow CO_2 + H_2$$

$$CO + 3\,H_2 \rightarrow CH_4 + H_2O$$

The steam reforming of methanol proceeds according to:

$$2\,CH_3OH + H_2O \rightleftharpoons 5\,H_2 + CO$$

$$CO + H_2O \rightleftharpoons H_2 + CO_2$$

The PEMFCs can also be used for *direct methanol conversion* (DMFC), in this case the following main reactions occur in the fuel cell:

Anode reaction $CH_3OH + H_2O \rightarrow CO_2 + 6\,H^+ + 6\,e^-$

Cathode reaction $3/2\,O_2 + 6\,H^+ + 6\,e^- \rightarrow 3\,H_2O$

The oxidation occurs in several steps, with the formation of adsorbed intermediate, such as –CHO or –CO on the catalyst. These adsorbates

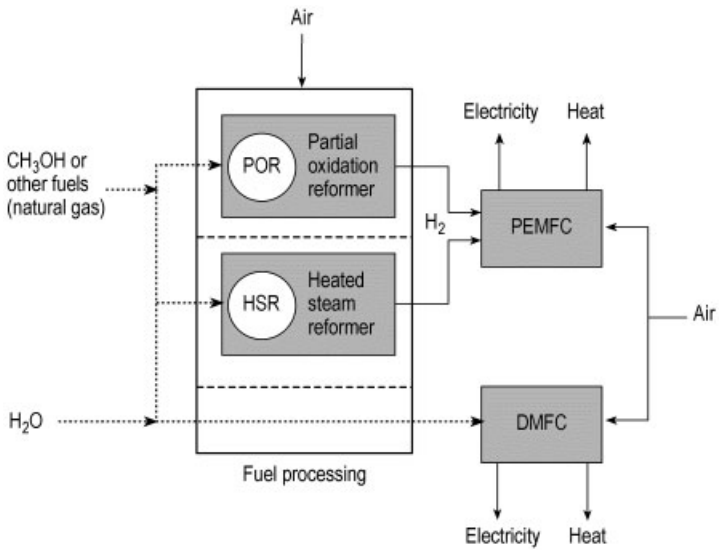

Figure 17. Two different routes for using methanol as fuel for a fuel cell.

cause a deactivation of the catalyst with a consequent loss of cell efficiency. Depending on the electrolyte, a DMFC can work with a liquid methanol – water mixture at ca. 80 °C or with a gaseous mixture at higher temperatures.

A DMFC offers several advantages such as the absence of the reformer in the periphery of the system which reduces costs and increases efficiency. A comparison between a PEMCF and a DMFC is shown in Figure 17, which shows the two different way to supply methanol or other fuels such as natural gas to a low temperature fuel cell with proton exchange membrane as electrolyte.

The development of DMFC dates back to the 1920s when Müller studied the possibility of electrochemical oxidation of methanol [31], his results were the basis of the work of Kordesch and Makro, who tried to build a DMFC in 1951 [11]. In the 1960s, industrial research resulted in the construction of a 100 W – methanol oxygen fuel cell by Bosch in 1963 [32], with a faradaic efficiency of 50 – 60 %, and current densities of 50 – 100 mA/cm^2 at 0.4 V.

At the beginning of the 1980s, the interest for DMFC rose again and methanol was used as fuel for high temperature fuel cells [33]. Nowadays, PEM fuel cells with Nafion as electrolyte appear to be well suited for the direct oxidation of methanol. There are, however, some major problems in adapting a PEMFC to operate with methanol. The catalytic material of the anode has to be improved in order to avoid the loss of activity because of the formation of byproducts that block the active sites of the catalyst. As in the case of a PEMFC using hydrogen from a reformer as fuel, a method to improve the anodic material is the use of Pt – Ru or Pt – Ru – Sn mixtures [26]. A particular problem of the DMFC is cross-diffusion of methanol through the electrolyte. The methanol is then oxidized at the cathode which causes a reduction of the cell voltage, because of the mixed potential at the cathode due to methanol oxidation and oxygen reduction. In addition, the efficiency of the electrochemical system is reduced, because part of the fuel is oxidized without the production of electrons in the external circuit.

Another possibility of avoiding the poisoning of the catalysts is to operate at higher temperature as long as the electrolyte is stable and the cross over of methanol is tolerable. The interaction between Nafion and methanol is complex, methanol causes an increase of the ohmic drop of the membrane and a simultaneous modification of the pore structure of Nafion. Although the PEMFC is not yet optimized for direct methanol oxidation and the high efficiency obtained in a PEM with pure hydrogen as fuel is not yet reached (see Fig. 18), current densities of

400 mA/cm^2 at 0.5 V have been observed in a PEMFC at 130 °C for direct methanol oxidation [34].

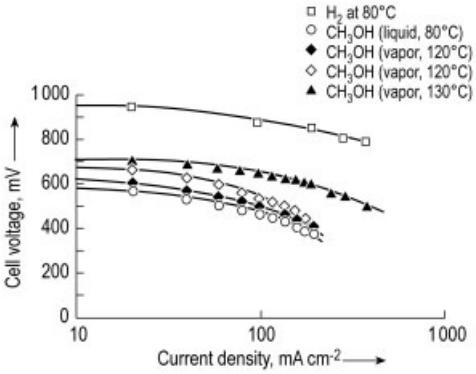

Figure 18. Current density – voltage curves for PEMFC operating with hydrogen and methanol. The efficiency of a PEMFC operating as direct methanol fuel cell is about half of that of a PEMFC operating with pure hydrogen.

3.1.3. Phosphoric Acid Fuel Cells (PAFC)

In contrast to alkaline electrolytes, acid electrolytes tolerate CO_2, therefore, in the 1950s, fuel cells containing acid electrolytes have been proposed for the oxidation of hydrocarbons. After first attempts with sulfuric acid, the interest of researchers focused on phosphoric acid as electrolyte, and at the end of the 1960s the first PAFC was realized with the acid immobilized in a matrix. In the 1970s, several companies in USA and Europe tried to improve the PAFC technology, but only with the Japanese Moonlight program PAFCs became a commercial system. PAFC systems with a total capacity of 20 000 kW are currently operating in Japan. A 11 MW PAFC power plant has been built as pilot plant by Tokyo Electric Power Corporation, in cooperation with Toshiba and International Fuel Cells (IFC). Other important Japanese companies such as Fuji electric, Toyota and Mitsubishi are involved in fuel cell projects as manufacturers [4].

The operating principles of a PAFC comprise the following reactions:

$$H_2 \rightarrow 2\,H^+ + 2\,e^- \text{ (oxidation of the fuel at the anode)}$$

$$1/2\,O_2 + 2\,e^- + 2\,H^+ \rightarrow H_2O$$

(reduction of the oxidant at the cathode, air is the most common oxidant)

The mobile ions are the H^+ which move from the anode to the cathode through 100 % phosphoric acid immobilized in a SiC – poly(tetrafluoroethylene) (PTFE) matrix, the water is produced at the cathode, it is removed with nitrogen and unreacted oxygen.

Reagents. Air is used as oxidant and hydrogen obtained by reforming of hydrocarbons such as natural gas, as fuel.

Electrodes. Graphite paper is used for the porous carbon electrodes and the catalyst is a mixture of platinum and PTFE based resin, for the use of $CO + H_2$ mixtures (reformate) Pt – Ru is used as catalyst for the anodic reaction.

Graphite is the material of the *interconnectors*.

The phosphoric acid fuel cell works at a temperature between 160 and 220 °C, the CO contamination of the catalysts is rather low at these temperatures; up to 1 % CO are tolerated in the feed gas. The conversion efficiency of a complete PAFC system is ca. 40 %, the emissions of a PAFC systems are much lower than the emissions caused by a gas engine and far below the emission limits stipulated by law in Germany (TA Luft), (see Fig. 19).

3.2. Intermediate Temperature Fuel Cells

No fuel cell system has been developed up to now working in a temperature range between 250 °C and 600 °C, although this range is favorable for some applications. Waste heat is obtained at a temperature level where it can be used for methanol reforming. In other applications high grade heat may be available without having the material problems of high temperature fuel cells. The main problem seems to be the development of a suitable electrolyte.

3.3. High Temperature Fuel Cells

3.3.1. Molten Carbonate Fuel Cells (MCFC)

The development of molten carbonate fuel cells started in 1951, when Boers and Ketlaar built a fuel cell with a mixture of sodium carbonate and tungsten oxide as electrolyte; later they changed it to a mixture of lithium, sodium, and

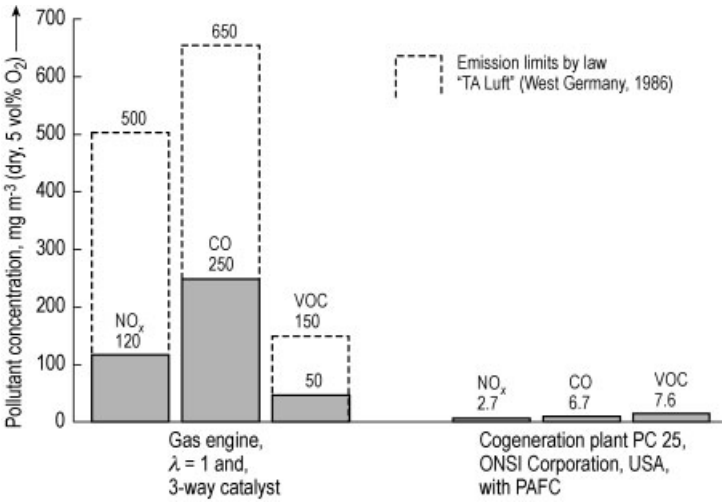

Figure 19. Comparison of emissions of a cogeneration plant with a gas engine and with a phosphoric acid fuel cell
VOC = volatile organic compounds

potassium carbonate in a sintered MgO matrix [10]. Such mixtures exhibit high ionic conductivities at about 700 °C. Fuel cells with circulating molten carbonate were built in the 1960s, when also the partial oxidation concept (POC) for the reformer was introduced with about 70 % efficiency of the energy conversion. In the field of energy conversion, partial oxidation means understoichiometric thermal or catalytic combustion of hydrocarbons or alcohols with oxygen from air in a reformer.

Theoretical (stoichiometric) conversion for the exothermic hydrogen production is as follows:

$$C_mH_n + m\,O_2 \longrightarrow m\,CO_2 + n/2\,H_2$$

$$C_mH_{2m+1}OH + (m-0.5)\,O_2 \longrightarrow m\,CO_2 + (m+1)\,H_2$$

The produced gas contains 20 – 40 vol % H_2, with large amounts of nitrogen (up to 50 vol %), water and carbon oxides.

In 2000 several projects on MCFC are being developed in USA, Japan, and Europe. Energy Research Corp. (ERC) and M-C Power Corporation are working on a MCFC plant with a power output of 2 MW/a; which are supposed to commercialize 500 kW to 3 MW systems. Mitsubishi, Fuji and Hitachi in Japan are involved in MCFC projects, cells of an area of about 1 m^2 are planned. Italian Ansaldo, Dutch ECN, and German MTU formed a development consortium for MCFC systems.

The operating principles of a MCFC comprise the following reactions:

$$H_2 + CO_3^{2-} \rightarrow H_2O + CO_2 + 2\,e^-$$

(oxidation of the fuel at the anode)

$$1/2\,O_2 + CO_2 + 2\,e^- \rightarrow CO_3^{2-}$$

(reduction of the oxidant at the cathode, air is the most common oxidant)

The mobile ions are CO_3^{2-} ions which move from the cathode to the anode through the molten Li, K, and Na carbonates immobilized in a MgO matrix, the water is produced at the anode. Water is removed together with CO_2, possible traces of CO react at the anode with H_2O, forming CO_2 and H_2. Therefore, it is very convenient to use MCFC with fuels such as natural gas or other hydrocarbons.

CO_2 needs to be recycled in order to keep the electrolyte composition constant, CO_2 is formed at the anode and has to be supplied at the cathode with the oxygen.

Reagents. Hydrogen, reformed alcohols, or hydrocarbons are used as fuel, oxygen or air as oxidant.

Electrodes. The electrocatalytic material of a MCFC is nickel. The cathode becomes oxidized and lithiated during the first hours of operation. Nickel oxide is soluble in molten carbonates, thus in the course of operation two undesirable effects may occur: (1) metallic nickel particles

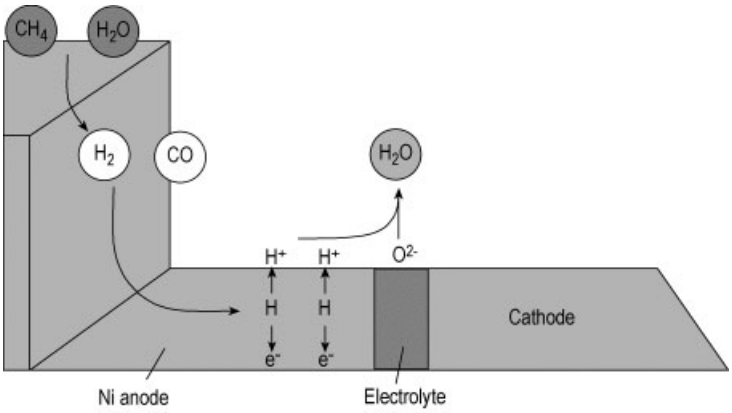

Figure 20. Methane reforming reaction at the anode of a molten carbonate fuel cell. The reaction between fuel and water takes place in the outer part of the anode, the produced hydrogen reacts at the interface anode – electrolyte.

are formed in the electrolyte which can lead to an electronic short circuit of the electrodes, (2) the size of the cathode diminishes. Two approaches have been proposed for solving these problems: the use of less corrosive molten carbonate mixtures and more stable cathodes containing iron and cobalt. The nickel anodes usually contain chromium which promotes the sintering process.

Inconel is used for the *interconnectors*.

The molten carbonate fuel cells work at ca. 650 °C, this temperature requires special materials stable at such high temperatures and towards corrosion, which is severe at this temperature. The electrocatalysts are mixed with ceramic materials and the cathode acts as the supporting element in MCFC. The cathode has the structure of sintered porous tiles, the electrolyte is produced by a tape cast technique, the anodic material is like a paste and is deposited on the electrolyte.

Molten carbonate fuel cells do not need external reforming of methane, reforming takes place at the anode of the fuel cell. The reactions occurring at the anode are as follows (see Fig. 20):

$$CH_4 + H_2O \rightarrow 3\,H_2 + CO$$

$$3\,H_2 + 2\,CO_3^{2-} \rightarrow 3\,H_2O + 3\,CO_2 + 6\,e^-$$

$$CO + CO_3^{2-} \rightarrow 2\,CO_2 + 2\,e^- \text{ (minor side reaction)}$$

The internal reforming takes place at temperatures of about 650 °C, which is lower than the usual temperature of a reformer (about 800 °C). The conversion, however, is fast enough, moreover it is not necessary to supply heat for the external reformer, since the heat for the internal reforming is provided by the heat losses of the reactions in the fuel cell. In this way the internal reforming works as a heat sink which helps to reduce the size of the heat exchanger in the power plant. Advantages and problems of internal reforming referring to higher temperature fuel cells in general are summarized below:

Advantages. The efficiency of the power plant increases with increasing degree of internal methane reforming from 2 at 50 % internal reforming to 8 at 99 % internal reforming.
Disadvantages. The H_2 producing endothermic reforming reaction is much faster than the H_2 consuming electrochemical oxidation; temperature gradients cause material problems; a sharp cooling effect occurs at the cell inlet.

3.3.2. Solid Oxide Fuel Cells (SOFC)

Solid oxide fuel cells have experienced a recent renaissance. The most common solid oxide, which is used as electrolyte in SOFC, is a mixture of zirconia and yttria, which is very similar to the Nernst mixture, an oxide studied by Nernst at the beginning of the century because of its ionic conductivity [35]. Yttrium stabilized zirconia (YSZ) contains between 8 and 10 mol % yttrium and has a stable cubic structure; it is an oxide ion conductor with a sufficiently high conductivity at temperatures around 1000 °C. This is an advantage for the direct oxidation of fuels

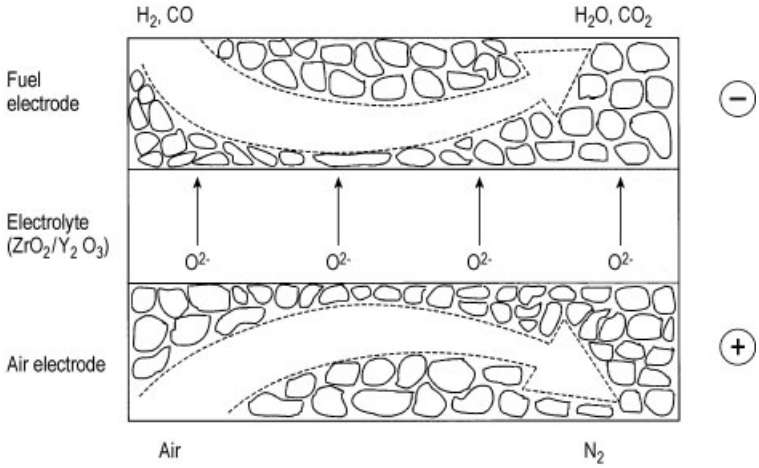

Figure 21. Structure of the three-phase zones at anode and cathode of a solid oxide fuel cell

in the fuel cell, but is a very challenging temperature for any material. Thermodynamics predict an efficiency of 80 %. A further advantage of the high operation temperature is the possibility to efficiently use the waste heat in a combined cycle power plant.

SOFC developing programs are carried out by Siemens – Westinghouse [36], Allied – Signal, Mitsubishi, Fuji Electric, and Sulzer.

The operating principles of a SOFC comprise the following reactions:

$$H_2 + O^{2-} \rightarrow H_2O + 2\,e^-$$

(oxidation of the fuel at the anode)

$$1/2\,O_2 + 2\,e^- \rightarrow O^{2-}$$

(reduction of the oxidant at the cathode)

The mobile O^{2-} ions in YSZ move from the cathode to the anode, the water is produced at the anode, it can be removed together with unreacted hydrogen and CO_2 (in the case of hydrocarbon combustion). For internal reforming the anode reaction also provides the necessary water for steam reforming. The operation principles of a SOFC with YSZ as electrolyte is shown in Figure 21.

Gd-doped CeO_2 has been proposed by several researchers as electrolyte with high conductivity at lower temperatures; gallates have also investigated. The Arrhenius plots of the ionic conductivity of some mixed oxides are shown in Figure 22 [37].

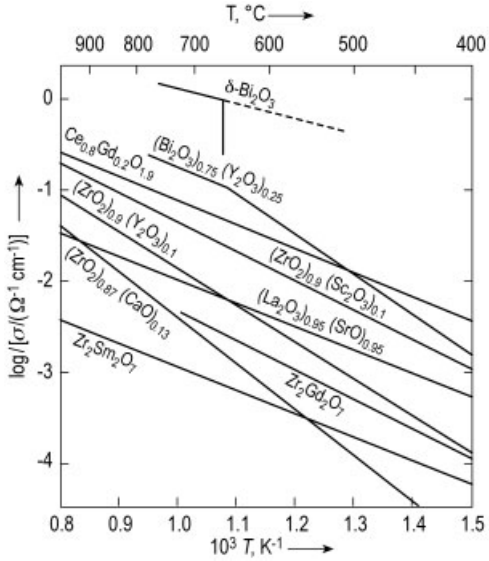

Figure 22. Arrhenius plots of the conductivity of several oxides, which exhibit high ionic conductivity and could be suitable electrolytes for SOFC

Reagents. Any kinds of fuels and oxidants are suitable, no reformer is needed for methane, the reagents can be directly supplied to the cell.

Electrodes. The anodes of SOFC consist of Ni cermet, a composite of metallic Ni and YSZ, Ni provides the high electrical conductivity and catalytic activity, zirconia provides the mechanical, thermal, and chemical stability. In addition, it confers to the anode the same expansion coefficient of the electrolyte and renders compatible

anode and electrolyte. The electrical conductivity of such anodes which is predominantly electronic, is shown in Figure 23 as a function of the nickel content, while Figure 24 shows the three-phase boundary at the interface porous anode YSZ and the reactions which take place. The cathode of the SOFC consists of mixed conductive oxides with perovskite crystalline structure. Sr doped lanthanum manganite is mostly used, it is a good *p*-type conductor and can contain noble metals.

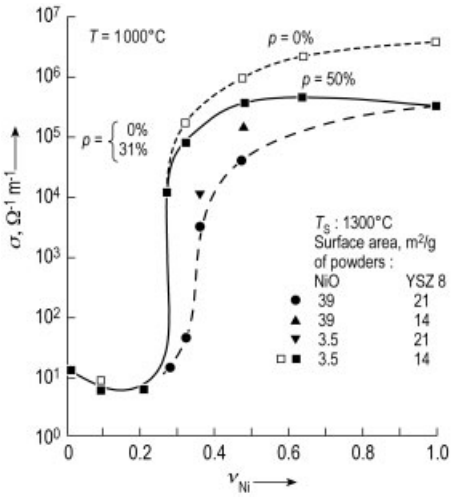

Figure 23. Electrical conductivity of Ni cermet, the anodic material of SOFC
$T = 1000\,°C$; $T_s = 1300\,°C$

Interconnectors. A ceramic consisting of lanthanum chromite or a chromium based alloy are normally used; austenitic steel is also suitable at $T < 850\,°C$.

The high operating temperature of SOFC renders crucial the choice of the materials, which have to possess not only the abovementioned typical features but also a very similar expansion coefficient in order to allow for thermal cycling. Thus, only a few materials can sustain the working conditions of SOFC without significant changes after long term operation. Most of the materials used in SOFC are relatively expensive. The expected high efficiency of a SOFC is quite attractive and many cell designs have been proposed. Many attempts have also been made to lower the operating temperature in order to reduce the demands on the materials.

Three designs for SOFC stacks are described below:

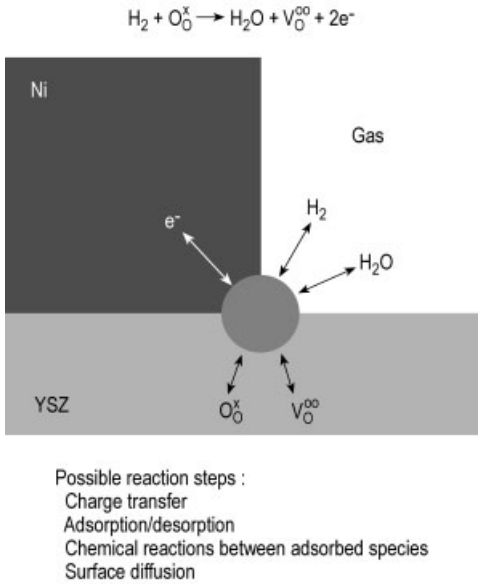

Figure 24. Anodic reactions at the three-phase boundary of a SOFC
$V_O^{\circ\circ}$ = oxygen vacancy in solid conducting oxide

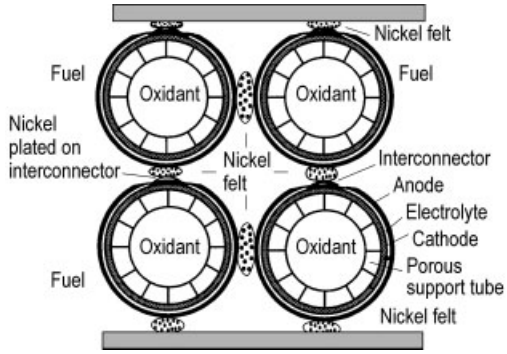

Figure 25. Tubular SOFC stack concept of Westinghouse
Advantages: No sealing; Disadvantages: Ohmic loses, low power density

Tubular Design of Westinghouse. Figure 25 shows the principles of the tubular design developed by Westinghouse. A single cell has the form of a long tube. The tube consists of a porous ceramic substrate, usually on zirconia basis. The components of the electrochemical cell are deposited on the tubular support, starting with the cathode. The electrolyte is deposited on the cathode, but the cathode is not completely covered by the electrolyte material, a strip of cathodic material is covered by the interconnector material for the electrical connection of the cathode. The anode is deposited on the electrolyte.

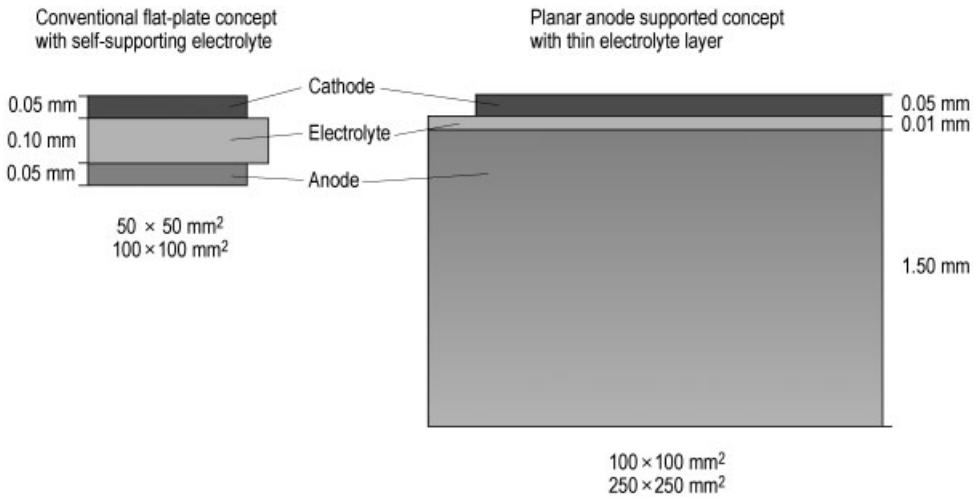

Figure 26. Conventional flat-plate (Siemens, left) and planar (Research Center Jülich, Germany, right) SOFC stack concepts

Several tubes are arranged in a case separated by nickel felts, air is supplied inside the tubes, the fuel is supplied outside the tubes.

The manufacturing of the tubular cells is relatively easy, the subsequent depositions of the electrode materials occur by simple coating and masking techniques, the gastight materials of electrolyte and of interconnector are more difficult to obtain and more sophisticated techniques such as plasma spray or electrochemical vapor deposition are necessary, which are rather expensive manufacturing techniques.

The main advantage of the tubular design is the absence of sealing in the stack. Westinghouse built a 25 kW stack and tested single cells up to 40 000 h. Currently a 100 kW demonstration power plant is being operated in the Netherlands.

Planar Designs.

Siemens Concept (Fig. 26, left). The planar concept of SOFC developed by Siemens is similar to the sandwich structure of the PEM single cell (see Figs. 8 and 12), a foil of YSZ with thickness of 150 – 200 µm acts as the supporting component on which the electrode materials are deposited, the single cells are then stacked with interconnectors with the gas channels and sealing elements in between. The gases can be supplied in co-, cross- or counterflow and can be let to and out the stack in internal or external manifolding. The differences in the gas management cause different temperature pattern in the single cells and in the stack. The thickness of the electrolyte, which is defined by the necessary mechanical stability, causes ohmic losses and renders it difficult to reduce the operating temperature considerably below 900 °C. Another disadvantage of this type of planar cell is the relative small size of the single cell (commonly $5 \times 5\,cm^2$, $10 \times 10\,cm^2$ possible), which is mainly due to the difficulty to cast large uniform ceramic foils and the necessity to increase the thickness of the electrolyte in the case of larger cells which results in higher ohmic losses.

Other developers also use the flat cell concept which was recently discontinued by Siemens because of the merge with Westinghouse.

IEV Concept (Fig. 26, right). In the planar SOFC developed by IEV-Forschungszentrum Jülich the anode is chosen as the supporting element [38]. In this cell, the anode has a thickness of approximately 1 mm, electrolyte and cathode are deposited in 2 – 10 µm and 40 – 50 µm thick layers, respectively, on the anode. The anode exhibits a varying composition. Typical characteristics of the anode are (Fig. 27): the reforming reactions occur in the outer layer of the anode, the central part of the anode has a high electronic conductivity and high porosity, the electrochemical reactions take places as usual at the interface with the electrolyte. This planar concept overcomes the disadvantages of the first planar concept (Siemens), it is, in fact, possible to work at lower temperature (< 700 °C) and to produce

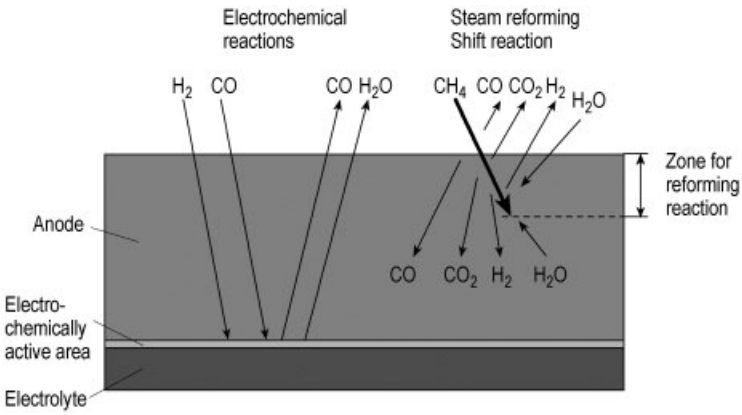

Figure 27. Model of the multi-zones of a substrate anode.

large single cells ($25 \times 25\,cm^2$). In addition, the reaction zones for reforming methane and electrochemically oxidizing the resulting hydrogen are spatially separated and can be optimized separately.

Figure 28. Sketch of the electric car Necar III developed by Daimler-Benz

4. Examples of Application of Fuel Cells

4.1. Zero Emission Vehicle

Demonstration cars equipped with fuel cells were produced already in the 1960s in USA by Union Carbide and Allis Chalmers, and in Europe by Siemens and VARTA, the fuel cells used in these vehicles were alkaline fuel cells. Since the beginning of the 1990s fuel cell cars have become again very attractive because they offer the possibility to reduce the engine emissions. Two of the most relevant projects on zero emission cars are carried out in Canada by Ballard and in Europe by Daimler-Benz. Both projects use PEMFCs. Beside other advantages PEMFCs exhibit a more advantageous weight : power ratio than AFCs of about 2 kg/kW vs. 15 kg/kW, the former is similar to combustion engine values. Ballard and Daimler tend to use methanol as fuel and air as oxidant. In the first stage, methanol will be reformed to hydrogen. However, the possibility of using a DMFC is not excluded, since it would have several advantages such as reducing the size and the weight of the system because of the elimination of the reformer and of the shift reactor necessary for the CO removal.

In 1993 the costs of a drive train for an electric car furnished with fuel cells are about 10 times higher than those for a conventional one (150 – 260 €/kW) [39]. However, the prices of fuel cells are supposed to decrease rapidly in the next years. The first Daimler-Benz car was a van named NECAR (new electric car), which was presented to the public in 1994. The van was mostly occupied by the fuel cell and by the hydrogen tank. In 1996 NECAR II was presented, NECAR II is also a van which seats six to eight passengers, has a compact fuel cell and the hydrogen tank on the roof. NECAR II can reach 110 km/h and can travel up to 250 km without refueling.

Necar III is a small car (Mercedes A-Klasse) with a PEMFC for indirect methanol combus-

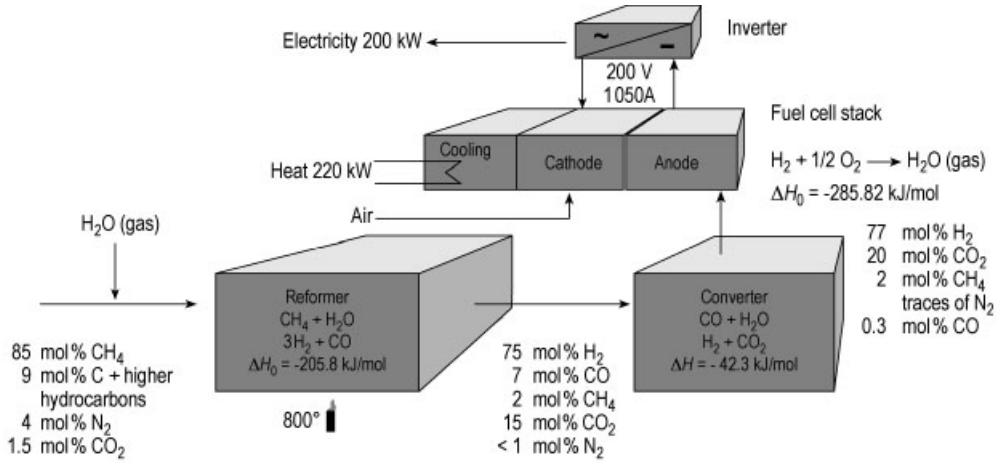

Figure 29. Phosphoric acid fuel cell cogeneration system PC 25 developed by ONSI and tested by Thyssengas

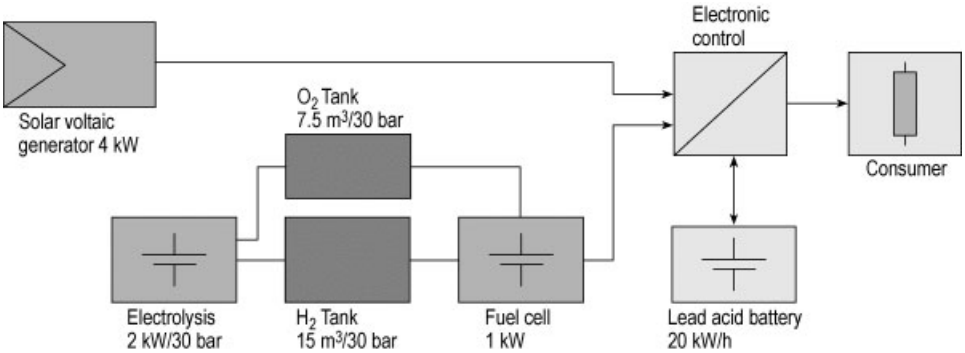

Figure 30. Block diagram of the energy generation and storage system for the energy autonomous house developed by FhG-ISE, Germany

tion with a methanol reformer and a gas clean-up system. Figure 28 shows the main features of the vehicle [40].

4.2. PAFC Power Plants

Cogeneration plants simultaneously produce electricity and heat in the form of steam or hot water. Fuel cells are suitable for being used in such systems, especially with natural gas as fuel. The advantage of natural gas is the low emission and the high energy content. An example of a cogeneration plant with PAFC is shown in Figure 29 [41]. This system comprises a reformer for the production of hydrogen from natural gas and a shift reactor for the removal of carbon monoxide down to ca. 1 % from the gas mixture to be supplied to the fuel cell. The fuel cell of this system is a phosphoric acid fuel cell stack PC 25 A, produced by ONSI Inc.; it has a maximum electric power of 200 kW and a maximum thermal power of 220 kW. The electric efficiency of the stack is about 40 %. PC 25 A are applied in several countries for on-site energy production and are the first fuel cell systems commercially available.

4.3. Regenerative Systems

Fuel cells can also be used in regenerative systems as a component of a long term energy storage system. An example of this application is the energy autonomous house in Freiburg, Germany, a project of the Fraunhofer Society. The house is a solar, independent, emission-free electricity source, which can work continuously throughout the year because of an electrochemical storage system. It comprises a photovoltaic solar

system for the conversion of light into electricity, the produced electricity which is not immediately used, can be stored in batteries for day-to-day fluctuation of the power requirements. For seasonal storage (store in the summer use in the winter) an electrolyzer is employed for the production of hydrogen and oxygen. The highly pure gases can be stored and later supplied to a fuel cell for the production of electricity (see Fig. 30) [42]. The quality of the gases allows the application of any low temperature fuel cell such as AFC or PEMFC.

5. References

1. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (ed.): "Umweltpolitik, Klimaschutz in Deutschland", Bonn 1993.
2. C. D. Schönwiese, B. Diekmann: *"Der Treibhauseffekt – Der Mensch ändert das Klima",* Rowohlt, Reinbek 1991.
3. K. Kordesch, G. Simader: *"Fuel Cells",* VCH, Verlagsgesellschaft, Weinheim 1996.
4. K. Ledjeff (ed.): *"Brennstoffzellen",* C. F. Müller, Heidelberg 1995.
5. W. R. Grove, *Philos. Mag.* **14** (1839) no. 3, 127.
6. W. Ostwald, *Z. Elektrochemie* **1** (1894) 122.
7. E. Baur, J. Tobler, *Z. Elektrochem.* **36** (1933) 169.
8. F. T. Bacon, *BEAMA J.* **6** (1954) 61.
9. K. Bishoff, E. Justi, *Jahrb. Akad. Wiss. Lit. (Mainz)* **250** (1955) 30.
10. G. H. J. Broers, *"High Temperature Galvanic Fuel Cells,"* Thesis, University of Amsterdam 1958.
11. K. Kordesch, A. Marko, *Österr. Chem. Ztg.* **52** (1951) 125.
12. Pratt & Whitney Aircraft, Design and Development of H_2/O_2-Fuel cell power plants for Apollo Command Module, NASA/Houston Final report 6/16/1962. R. B. Ferguson: "Apollo Fuel Cell System", Proceedings of 23rd Annual Power Sources Conference, 1969, pp. 11 – 13.
13. M. B. Clark et al., *Electrochem. Technol.* **3** (1965) 166.
14. VARTA, US 3 597 275, 1971.
15. Siemens AG: "20 kW Fuel cell systems of compact design," Report BMFT-FB-T 80-055 to 80-059, 7 parts, October 1980, 940pp.
16. H. van den Broeck et al.: "Status of Elenco's Alkaline Fuel Cells Development", *Int. J. Hydrogen Energy* **7** (1986) no. 21, 471 – 474.
17. K. V. Kordesch: "1951 – 1975, 25 Years of Fuel Cell Development," *J. Electrochem Soc.* **125** (1978) 77C – 91C.
18. K. Kordesch et al.: "Fuel cell research and development projects in Austria", Proceedings of the 7th world Hydrogen Energy conference Moscow 1988 T. N. Veziglu (ed.) *Hydrogen Energy Progress VII* Pergamon Press, Oxford, 1988.
19. M. Schautz, G. Dudley: "Testing of Space fuel cells at ESTEC" reprinted article from ESA Bulletin no. 60, 1992.
20. O. Führer et al., *Int. J. Hydrogen Energy* **19** (1994) no. 3, 343 – 348.
21. R. Cohen: "Gemini Fuel Cell System", *Proc. 20th Power Sources Conference* 24 – 26 May 1996, p. 21 – 24.
22. US 3 784 399, 1974 (W. Grot).
23. K. Prater, *J. Power Sources* **29** (1990) 410.
24. Ballard Power Systems Inc. Information Summary 1993.
25. *Automotive Engineering* Dec. 1997, p. 112.
26. A. Aramata, M. Masuda, *J. Electrochem. Soc.* **138** (1991) 1949.
27. A. C. Ferriera, S. Srinivasan, *ECS, Spring Meetings, San Francisco, CA,* May 22 – 27, 1994, p. 969.
28. Brochure DuPont Nafion perfluorinated membranes, DuPont, Faytetteville 1994.
29. T. E. Springer, T. A. Zawodzinski, S. Gottesfeld, *J. Electrochem. Soc.* **138** (1991) 2334.
30. T. E. Springer, M. S. Wilson, S. Gottesfeld, *J. Electrochem. Soc.* **140** (1993) 3513.
31. E. Müller, *Z. Electrochem.* **28** (1992) 101.
32. W. Haecker et al., *Metalloberfläche* **24** (1979) 185.
33. D. J. Wheeler et al., Proc. Workshop on Direct Methanol – Air Fuel Cells, The Electrochemical Soc., Pennington NJ, vol 92 – 14, 1992, p. 193.
34. M. Waidhas et al.: "Research and development of low temperature fuel cells at Siemens" 1994-Fuel cell seminar San Diego, CA, Nov/Dec 1994, p. 477.
35. W. Nernst, *Z. Elektrochem.* **5** (1899) 41.
36. SOFC, Production information from Westinghouse Electric Corporation December 1992.
37. B. C. H. Steele: "Ceramics for SOFC and Electrochemical Reactor Technology" in R. Waser et al. (eds.): Proc. 4th Intern. Conf. on Electronic Ceramics and Applications, Electroceramics VI, Aachen 1994, pp. 663 – 676.

38. EP-A1 0 829 103, 1998 (L. G. J. Haart, J. Divisek, E. Riensche, U. Stimming).
39. K. Letjeff in K. Letjeff (ed.): *Energie für Elektroautos,* Verlag C. F. Müller, Karlsruhe 1993, p. 227.
40. http://www.daimler-benz.com/research/specials/zukunft/necar_g.htm
41. H. Knappstein, *Euroheat and power* **25** (1996) 58.
42. Brochure "Das energieautarke Solarhaus", Fraunhofer-Institut für Solare Energiesysteme, FhG-ISE, Freiburg 1992.

Fuel Oil → Automotive Fuels
Fuel Oil → Heating Oil
Fullerenes → Carbon
Fumaric Acid → Maleic and Fumaric Acids
Fumigants → Insect Control
Fumigants → Nematocides

Fungicides, Agricultural

Antimycotics, Dermatotherapeutic Agents, Disinfectants are separate keywords.

PETER ACKERMANN, Novartis Crop Protection AG, Basel, Switzerland (Chaps. 1 – 8)

PAUL MARGOT, Novartis Crop Protection AG, Basel, Switzerland (Chaps. 1 – 9)

FRANZ MÜLLER (formerly Novartis Crop Protection AG), Allschwil, Switzerland (Chaps. 6 – 10)

1. Introduction . . . 3
2. Phytopathogenic Fungi . . . 3
3. Types of Fungicide and Mechanism of Action . . . 4
4. Method of Fungicide Application 5
4.1. Seed Treatment . . . 6
4.2. Foliage and Fruit Sprays . . . 6
4.3. Soil Treatment . . . 7
4.4. Trunk Injections . . . 7
4.5. Postharvest Treatment . . . 7
5. Fungicide Resistance . . . 7
6. Nonsystemic (Contact) Fungicides . . . 8
6.1. Inorganic Compounds . . . 8
6.2. Organometallic Compounds . . 9
6.3. Dithiocarbamates . . . 9
6.4. *N*-Trihalomethylthio Derivatives (Phthalimides) . . . 12
6.5. Dicarboximides . . . 13
6.6. Miscellaneous Organic Compounds . . . 14
7. Fungicidal Antibiotics . . . 21
8. Systemic Fungicides . . . 22
8.1. Organophosphates . . . 22
8.2. Benzimidazoles . . . 23
8.3. Carboxanilides . . . 24
8.4. Phenylamides . . . 25
8.5. Phosphites . . . 27
8.6. Inhibitors of Sterol Biosynthesis 28
8.6.1. Triazoles . . . 28
8.6.2. Imidazoles . . . 34
8.6.3. Pyridines . . . 35
8.6.4. Pyrimidines . . . 35
8.6.5. Piperazines . . . 36
8.6.6. Morpholines . . . 37
8.6.7. Strobulins . . . 38
8.7. Miscellaneous . . . 39
9. Economic Aspects . . . 42
10. Toxicology and Occupational Health . . . 42
10.1. Nonsystemic (Contact) Fungicides . . . 42
10.1.1. Inorganic Compounds . . . 42
10.1.2. Organometallic Compounds . . . 43
10.1.3. Dithiocarbamates . . . 43
10.1.4. *N*-Trihalomethylthio Derivatives (Phthalimides) . . . 45
10.1.5. Dicarboximides . . . 46
10.1.6. Miscellaneous Organic Compounds . . . 47
10.2. Fungicidal Antibiotics . . . 52
10.3. Systemic Fungicides . . . 52
10.3.1. Organophosphates . . . 52
10.3.2. Benzimidazoles . . . 53
10.3.3. Carboxanilides . . . 54
10.3.4. Phenylamides . . . 55
10.3.5. Phosphites . . . 56
10.3.6. Inhibitors of Sterol Biosynthesis . 56
10.3.6.1. Triazoles . . . 56
10.3.6.2. Imidazoles . . . 60
10.3.6.3. Pyridines . . . 61
10.3.6.4. Pyrimidines . . . 61
10.3.6.5. Piperazines . . . 62
10.3.6.6. Morpholines . . . 62
10.3.6.7. Strobilurins . . . 63
10.3.7. Miscellaneous . . . 63
11. References . . . 65

1. Introduction

This article is concerned only with fungicides used in agriculture. Fungicides for human and veterinary therapy are described in → Antimycotics, → Dermatotherapeutic Agents. For the use of fungicides in the protection of wood and other materials, see → Biocides; → Wood, Surface Treatment and → Wood, Preservation.

Agricultural fungicides are chemicals that are used to prevent or minimize crop losses caused by phytopathogenic fungi. Successful compounds must be sufficiently fungitoxic at doses that are well tolerated by the crop. De-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_085

pending upon the concentration of the fungicide, its activity may be *fungistatic,* i.e., fungal growth is only temporarily inhibited in its presence, or it may be *fungicidal,* i.e., the fungus is killed on contact.

Fungal plant diseases frequently develop during periods of high humidity accompanied by favorable temperatures; thus the period of most intensive fungicide use generally coincides with the onset of the rainy summer season. This is particularly true for fungicides that are sprayed to protect foliage and fruit of the growing crop. The use of fungicides in seed and postharvest treatment is a relatively low-cost method of preventing crop loss, even in situations where the risk is considered to be low.

2. Phytopathogenic Fungi

Plant diseases may be induced by noninfectious or by infectious agents. Noninfectious diseases can arise as a result of unfavorable environmental factors (e.g., temperature extremes, poor water relations, a deficiency or excess of minerals). Infectious diseases are caused by living organisms including fungi, bacteria, viruses and nematodes that can be transmitted to healthy plants.

Biology of Fungi. (See also → Biotechnology, Chap. 1.3.2.). Fungi are considered to be primitive (eukaryotic) microorganisms that lack chlorophyll. The body (thallus) of the fungus usually consists of branched threads (hyphae) which form a network (mycelium). The hyphae may be divided into cells by cross-walls (septae). Reproduction and dispersal take place by means of asexually or sexually produced spores that are borne directly on the hyphae or are produced in fruiting bodies. Asexually produced spores (in conidia, sporangia) are usually formed in large numbers and generally cannot survive under unfavorable environmental conditions (e.g., temperature extremes or low humidity). Their function is to ensure rapid dissemination of the fungus and colonization of new hosts. Sexually produced spores (e.g., oospores, ascospores, basidiospores), on the other hand, usually ensure the survival of the fungus under adverse weather conditions in the absence of a suitable host. The asexual part of the life cycle is referred to as the imperfect stage whilst the sexual phase is referred to as the perfect stage.

Classification. Because the spectrum of activity of fungicides can often be linked to the taxonomic grouping of the fungi, a basic understanding of the classification system is necessary when discussing activity. Fungi are classified on the basis of the type of spore or fruiting body they produce and are divided into four major classes: Oomycetes, Ascomycetes, Basidiomycetes, and Deuteromycetes. (In Deuteromycetes the perfect stage is not known.) Not only are the spore forms and life cycles of the fungi in these classes very different, the composition of their cell walls and their metabolic pathways also vary considerably. Examples of some important plant pathogens in the four major classes are given below.

Oomycetes include the downy mildews, late blights, white rusts, and root and collar rots caused by two soil-borne genera *Pythium* and *Phytophthora.* Important examples are

Phytophthora infestans, late blight of potato and tomato
Plasmopara viticola, downy mildew of grape vine
Phytophthora nicotianae var. *nicotianae*, tobacco black shank
Peronosclerospora sorghi, maize and sorghum downy mildew
Pythium ultimum, seedling damping-off of various crops

Ascomycetes include the powdery mildews and anthracnoses, as well as various leaf and fruit spot diseases. Important examples are

Erysiphe graminis, powdery mildew of cereals
Podosphaera leuchotricha, powdery mildew of apples
Venturia inaequalis, apple scab
Sclerotinia fructicola, brown rot of stone fruit
Gaeumannomyces graminis, take-all of cereals

Basidiomycetes include the true rusts, smuts, and bunts. Important examples are

Puccinia triticina, brown rust of wheat

Puccinia striiformis, yellow smut of cereals
Tilletia caries, wheat bunt
Pellicularia oryzae, rice sheath blight

Deuteromycetes is a large group of plant pathogens responsible for leaf and fruit spots, among others. Important examples are

Pyricularia oryzae, rice blast
Botrytis cinerea, grey mold of grapes
Alternaria solani, early blight of tomatoes
Cercospora arachidicola, leaf spot of groundnuts
Fusarium oxysporum, Panama disease of bananas

Mode of Action. Phytopathogenic fungi attack higher plants to obtain nutrients for their own growth and reproduction, i.e., they are parasitic. *Facultative parasitic fungi* (e.g., *Pythium ultimum*) are able to grow on both living plant material and nonliving substrates. They invade and kill host tissue by producing enzymes or toxins. *Obligate parasites,* such as rusts, are only able to obtain nutrients from living hosts; they are adapted to carry out this function, without killing the plant outright.

All plant parts are liable to fungal attack, and the symptoms can be expressed in many forms such as wilt, rot, blight, or mildew. Because these terms describe only the symptom and not the causal agent of the disease, the pathogen must often be isolated from the diseased material and identified before a diagnosis can be made. Once the pathogen has been identified, a knowledge of its life cycle and epidemiology provides an appropriate basis for selecting

1) the most appropriate time to implement chemical control measures
2) the method of application, and
3) the most suitable fungicide.

Crop Losses. (See also → Crop Protection.) The damage that fungi cause to agricultural crops may be of a quantitative or a qualitative nature. Diseases caused by Oomycetes tend to be "explosive" and, under appropriate conditions, can result in total crop loss within a matter of weeks (e.g., potato late blight). In regions where such diseases pose a threat, protective fungicide sprays must be applied routinely before disease symptoms appear.

Other fungi may not cause total crop loss but may produce leaf lesions or necrosis which reduces the photosynthetic leaf area. This can lead to premature senescence of the foliage and decreased yield (e.g., *Cercospora* leaf spot of groundnuts). The timing of fungicide application is not as critical in this case and the decision might be taken to delay spraying until the disease is first reported in a particular region.

The economic effects of a disease such as powdery mildew of apples may not be obvious at first, particularly under conditions of moderate disease pressure. In the long term, however, failure to implement suitable control measures leads to a decline in the vigor of affected plants, which results ultimately in death.

Qualitative damage is caused by diseases that affect the appearance of the marketed crop (e.g., apple scab). Lesions may be superficial and not affect deeper tissues, but blemished fruit is sharply downgraded and the grower then receives a much lower price. A strict, preventative fungicide spray regime should, therefore, be followed to prevent economic losses.

3. Types of Fungicide and Mechanism of Action

Fungicides can be divided conveniently into nonsystemic and systemic types, although some overlap occurs with compounds that have a local penetrating effect.

Nonsystemic fungicides (also called contact or residual protective fungicides) do not penetrate the plant tissues. They form a protective barrier on the plant surface and prevent the fungus from entering the host by inhibiting spore germination or mycelial growth. They only serve a protective function. Most nonsystemic fungicides react with thiol groups that are present in the enzymes of the fungus and so inhibit many metabolic processes. They are therefore considered as multisite inhibitors.

Systemic fungicides enter plant tissues and are transported in the plant sap. They may be transported upward (i.e., acropetal translocation) or downward, from leaves to roots (i.e., basipetal translocation). These compounds usually inhibit fungal growth after the pathogen has penetrated the host tissues; thus, they have curative properties. The spectrum of activity may be

narrow and confined to specific types of fungi, as in the case of ethirimol and powdery mildew, but it may also be relatively broad as in the case of the triazoles. Systemic fungicides generally interfere with biosynthetic processes (e.g., respiration, protein synthesis, membrane functions) and are specific-site inhibitors. They inhibit fungal growth and spore production more effectively than spore germination.

Locally systemic fungicides penetrate into the plant tissue but are not transported in the plant sap.

The earliest nonsystemic fungicides were inorganic and based on either sulfur or copper. Later, organic mercury compounds were introduced for seed treatment. Fungicides based on these three elements are still important, although mercury based compounds are being replaced gradually by safer, more environmentally acceptable chemicals.

From the 1930s to the late 1960s, many new organic compounds with fungicidal activity were discovered. Of great importance was the discovery of dithiocarbamates in 1934, which finally led to the patenting of mancozeb in 1964. Other major fungicide types discovered during this period include chlorothalonil and the phthalimides, which are still important today.

Modern systemic fungicides were discovered in the mid- to late 1960s when activity against rice diseases was detected in organophosphorus compounds. In 1967, a new group of broad-spectrum fungicides, the benzimidazoles, was discovered. This was followed quickly by the identification of several major new groups including carboxanilides, pyrimidines, morpholines, and antifungal antibiotics. In the 1970s, further breakthroughs occurred with the identification of phenylamides, ethyl phosphite triazoles, and strobilurins.

The action of fungicides may be described as

1) *curative:* the fungicide is able to stop pathogen development during the incubation period (the time from when infection occurs to the appearance of first symptoms)
2) *eradicative:* the fungicide is able to stop the progress of the disease after the appearance of symptoms
3) *protective:* the fungicide prevents infection (penetration into the plant) from taking place. It should be present on the plant surface before spore germination occurs.

Systemic fungicides have resulted in considerable progress in the standards of disease control in crops. Compared to nonsystemic fungicides, significantly smaller amounts are needed per hectare at less frequent intervals. Because systemic fungicides inhibit fungal growth inside plant tissues, they can be used curatively or even eradicatively. However, the spectrum of diseases controlled by certain systemic fungicides is narrow, and some fungi are prone to development of resistance (cf. Chap. 5).

Quantitative analysis of most agricultural fungicides relies on HPLC and GC. A description of these methods is, however, beyond the scope of this article; for details, see [9].

Some older fungicides are no longer in use and have therefore been omitted from this update, e.g., Chloroneb, Anizaline, Dichlone, Phosdiphen, SSF-109, ICIA 0001, Furmecyclox, Pyracarbolid, Ethoxyquin. Other older fungicides which are no longer under patent protection are still listed in this edition because they are marketed as generics by different companies or used in combination products. Only a selection of trade names is given for each fungicide.

4. Method of Fungicide Application

Because of the close association between pathogenic fungi and their hosts, fungal diseases are often more difficult to control than insect pests, which are more obvious and accessible targets. To be effective, fungicides must come into contact with the pathogen and, especially in the case of nonsystemic compounds, must be distributed uniformly on the surface of all susceptible plant parts so as to form a protective barrier.

The efficiency of the fungicide is influenced by the type of equipment that is chosen for its application. Likewise, the formulation can influence retention and persistence of the chemical on the target surface.

Plants are treated with fungicides primarily in five ways, which are described in Sections 4.1, 4.2, 4.3, 4.4, 4.5.

Legal Aspects. Before any fungicide can be used in agriculture, it must be registered with the

responsible authorities in the country concerned to ensure that it complies with the necessary health, environmental, and safety requirements (→ Crop Protection, Chap. 6., → Crop Protection, Chap. 7.)

4.1. Seed Treatment

Fungal spores adhering superficially to the surface of seeds (e.g., *Tilletia caries* on wheat) can be controlled easily by nonsystemic fungicides. Others such as *Helminthosporium* spp., which colonize the seed coat in the form of a mycelium, can also be controlled by nonsystemic fungicides, but better control may be obtained with systemic products. Some pathogens establish themselves deeply within the embryonic tissue of the seed (e.g., *Ustilago* spp. on wheat and barley) and can only be controlled effectively with systemic fungicides that can penetrate the seed. Seed lots may be contaminated by dormant structures of fungi (e.g., sclerotia or oospores). Damage to seedlings caused by these fungi, as well as by pathogens living in the soil, can be minimized by appropriate seed treatment.

Systemic fungicides (e.g., metalaxyl used for the control of downy mildew of peas) protect the foliage for several weeks after treatment of the seed.

Fungicides for seed treatment are formulated as dry dusts, wettable powders, or ready-to-use liquids. Dusts are applied directly to the seed but the amount that can be retained on the seed surface may be limited. This may be overcome by mixing the fungicide with the seed during sowing – the so-called planter-box treatment. Wettable powders and liquids are mixed with water and applied as slurries. Liquid formulations are also sprayed onto the seeds with special equipment. Formulations usually contain "stickers" to aid retention on the seed, as well as dyes to indicate that the seed has been treated.

Modern technology employs inert carriers to coat the seed with fungicide; this permits larger amounts of fungicide to be carried on the seed.

4.2. Foliage and Fruit Sprays

The most common method of applying fungicides to foliage and fruit is by a water-based spray solution in which fungicide particles are finely dispersed. This allows uniform coverage of the leaves or fruit. Dust formulations are preferred for some compounds (e.g., sulfur) in certain countries, but they are becoming less popular.

Because nonsystemic fungicides are purely protective, and act by inhibiting spore germination or penetration of the fungus into plant tissue, they must be applied before the arrival of the pathogen. Application must be repeated at short, often weekly, intervals to protect new growth and to replace losses resulting from weathering or wash-off.

Since systemic fungicides enter the plant and are redistributed by translocation, the need for an efficient spray deposit on all susceptible plant surfaces is less critical than for nonsystemic compounds. However, in most cases systemic fungicides are only translocated upward (i.e., acropetally) by means of transpiration; therefore, it is important to ensure good penetration into the lower foliage. Systemic fungicides do not have to be applied as frequently as nonsystemic fungicides because they persist longer within tissues, are less liable to be washed off, and also often protect the new growth. Intervals of 14 – 21 days are often possible, compared with the 7 – 10 days required for nonsystemic compounds. Furthermore, the timing of spraying is not as critical because systemic fungicides can block pathogen development after infection has occurred.

Formulations for use as foliar and fruit sprays include wettable powders, emulsifiable concentrates, suspension concentrates (flowables), and more recently, water-dispersible granules. They all contain agents that ensure uniform dispersion of the active ingredient in the spray mixture and on the plant surface.

4.3. Soil Treatment

The soil-inhabiting fungi responsible for diseases such as seedling damping-off, vascular wilt, and root and basal stem rot are particularly difficult to control. Until recently, all-purpose soil fumigants (e.g., methyl bromide [*74-83-9*]) were the only effective means of chemical control. Fumigants act as general soil sterilants and

kill fungi, insects, and weed seeds. They are phytotoxic and can be used only for a defined period several weeks prior to planting.

Since the mid-1970s, a number of soil-applied fungicides have become available (e.g., metalaxyl, etridiazole, and propamocarb) which permit highly effective control of soil-borne *Phytophthora* and *Pythium* spp.

Soil fungicides are formulated as granules for ease of application, especially in perennial crops, but they are also available in formulations that can be sprayed onto the soil surface. Soil fungicides should be relatively mobile so that they move down to the root zone, but not so mobile that they are leached out of the upper soil layers. Compounds that are applied directly in the furrow during planting to control seedling diseases are placed accurately where needed.

4.4. Trunk Injections

A recent, novel development is the control of *Phytophthora* root rot disease of avocado and citrus trees by means of fungicides injected into the trunk [10]. At present, this technique can be applied only to phosphite fungicides such as fosetyl Al and its degradation product phosphorous acid, which are translocated downward (i.e. basipetally) in appreciable quantities (cf. Chap. 3). Trunk injection is laborious because several holes must be bored in the trunk of each tree. The fungicide solution is then injected by syringe into the trunk, under pressure, over a period of a few hours. If other compounds with similar properties but a different spectrum of activity are discovered, use of this elegant method could be extended to control root diseases of other tree crops.

4.5. Postharvest Treatment

Fungal pathogens are responsible for the majority of postharvest losses in perishable produce such as fruit, vegetables, and starchy root crops with high moisture content. Infection may occur prior to, or during, harvest as well as during storage. Although fungicides that are applied in the field can prevent losses which develop during storage, the term *postharvest fungicides* refers only to compounds that are applied to the harvested crop.

One or two fungi may be initially responsible for disease. However, they are often followed by a number of nonspecific, secondary pathogenic fungi and bacteria which cause extensive tissue breakdown and rotting of the produce. If mechanical injury occurs during harvesting, these secondary microorganisms can establish themselves without the help of primary invaders. Postharvest treatment should not only prevent infection and subsequent spread of disease to healthy produce, but also stop the development of disease after infection has taken place. The chemicals used should, therefore, have penetrating properties (e.g., fumigants or systemic fungicides).

Most postharvest fungicides are used in the fruit export industry to treat crops such as citrus fruit, bananas, and tropical fruit. Compounds such as benzimidazole derivatives, dicarboximides, and triazoles are commonly used for this purpose. The fungicides are generally applied in aqueous solution by dipping, spraying, or brushing, but they are sometimes incorporated into the standard waxes used to treat fruit.

5. Fungicide Resistance

Although fungicides have been used on a large scale for over a century, fungicide resistance has only become an important problem since the early 1970s. The fungicides used until 1970 were protective, nonsystemic substances with a multisite or nonspecific mode of action. The onset of resistance problems coincided with the introduction of systemic compounds having a specific or single-site mode of action. Resistance has, however, not been limited entirely to the systemics; it has also developed in some contact fungicides with site-specific activity, such as the organotin compounds used for the control of *Cercospora beticola* on sugar beet.

Systemic fungicides are especially prone to the development of resistance because of their site-specific mode of action. Phytopathogenic fungi have been able to overcome this specificity because a slight change at the target site results in decreased sensitivity to the fungicide. Such genetic variants may arise spontaneously but occur more frequently in fungi which pro-

duce large numbers of spores and undergo many reproductive cycles during the cropping season.

Resistance to nonsystemic fungicides is rare because they are mostly multisite inhibitors; changes would be required at many sites for resistance to occur.

The mechanisms by which resistance to fungicides can arise are outlined in [8], [11].

The majority of cases in which there has been a reduction in the level of disease control under practical conditions (i.e., field resistance) have occurred when specific systemic fungicides have been used exclusively and often intensively over an extended period of time. Experience has shown that certain types of systemic fungicide are especially prone to sudden, large shifts in sensitivity of part of their target population, which results in complete loss of disease control. This type of resistance development has occurred, for example, in *Venturia inaequalis* (apple scab) treated with benzimidazoles and in *Phytophthora infestans* (potato late blight) treated with phenylamides.

The triazole fungicides belong to a large group of compounds whose mode of action is to inhibit the C-14 demethylation step in ergosterol biosynthesis (for further details, see Section 8.6.1). Their increasing use has been associated with progressive shifts in fungal sensitivity over a period of several years. Field performance was not affected initially, but the levels of control of some diseases [e.g., *Erysiphe graminis* (powdery mildew) of barley] are now lower than when the products were first introduced. The morpholines, which also inhibit ergosterol biosynthesis, have not yet been associated with a resistance problem; they inhibit $\Delta^8 \rightarrow \Delta^7$-isomerization and/or Δ^{14}-reduction.

Measures taken by fungicide manufacturers to manage the resistance problem include (1) supplying "at-risk" compounds solely as prepacked mixtures with nonsystemic fungicides, (2) restricting use to a limited number of sprays or to a well-defined period of the growing season, and (3) discouraging curative use and soil application for the control of air-borne pathogens. Alternate use of fungicides with different modes of action also hinders the development of resistance, and such usage has been actively promoted in many countries both by academic and government departments. These combined measures reduce the selection pressure that is placed on the fungus population.

The activities of the Groupement International des Associations Nationales de Fabricants de Produits Agrochimiques (GIFAP) and the Fungicide Resistance Action Committee (FRAC) [12] have been important in persuading the manufacturers of fungicides that have a common mode of action to agree on common antiresistance strategies. These strategies have, however, tended to erode the unique benefits of systemic fungicides, such as long intervals between spraying, curative activity, low dose rates, and ease of application. Nevertheless, even with the compromises that must be made to manage resistance, the standards of disease control that can be obtained by following these strategies are much higher than those of the 1970s.

6. Nonsystemic (Contact) Fungicides

6.1. Inorganic Compounds

Sulfur (→ Sulfur) was one of the first fungicides used in agriculture and is still widely employed today; it accounts for one-third of inorganic fungicide sales. It is formulated for use as a dust, a colloidal suspension, or a wettable powder. Sulfur's spectrum of activity is narrow and is confined to the powdery mildews, but it also has a side effect on mites. It is used mainly on grapevines but also on orchards, sugar beet, and vegetables. At high temperature, sulfur can be phytotoxic.

Sulfur [*7704-34-9*], S, A_r 32, *mp* 114.5 °C, *bp* 445 °C, *d* 2.06, vapor pressure 527 µPa at 30.4 °C, is a yellow solid which is practically insoluble in water, however, crystalline forms are soluble in carbon disulfide [9, p. 499], [13, pp. 17, 51], [14, p. 123].

Trade Names. Cosan (AgrEvo); Kumulus DF (BASF); Thiovit (Novartis); Microthiol Special (Elf Atochem).

Copper was originally discovered and used in the form of Bordeaux mixture; more than 30 types of copper fungicides are now available (→ Copper Compounds). In terms of market share, copper oxychloride dominates because of its efficacy, coupled with its good crop tolerance; other important types include copper hydroxide and copper oxide. The spectrum of activity is broad, and copper compounds are used on a wide

range of crops including grapevines, coffee, vegetables, fruit, and cocoa. Copper fungicides are relatively persistent on foliage or fruit and are readily redistributed by water, which gives an improved surface cover. Fungitoxicity increases, as does phytotoxicity, with smaller particle size. Although not as active as more modern, organic fungicides, copper compounds are attractive because of their low price and broad spectrum of activity. They are relatively safe and have few official restrictions on their use. Some concern exists, however, regarding the accumulation of copper in soil, particularly on perennial crops.

Copper(II) oxychloride [*1332-40-7*], dicopper chloride trihydroxide, $Cu_2Cl(OH)_3$, M_r 213.6, is a green to blue-green powder which is insoluble in water or organic solvents but soluble in dilute acids and ammonium hydroxide [9, p. 134], [14, p. 121], [15, p. 371].

Copper(II) oxychloride is produced by air oxidation of cuprous salts and scrap copper in sodium chloride solution. Properties and production are described in detail elsewhere (→ Copper Compounds, Chap. 4.3.).

Trade Names. Coprantal; Recop (Novartis); Cupravit (Bayer); Cobax (BASF); Biltox (Rallis); Ossiram (Sipcam); Pasta Caffero (Caffero).

Cuprous oxide [*1317-39-1*], copper(I) oxide, Cu_2O, M_r 143.1, is a yellow to red powder which is insoluble in water and organic solvents but soluble in dilute mineral acid and aqueous solutions of ammonia [9, p. 139], [13, p. 48], [14, p. 121].

Cuprous oxide is produced by reduction of copper(II) salts under alkaline conditions or by electrochemical oxidation of metallic copper.

On exposure to moist air, copper(I) oxide oxidizes to copper(II) compounds. Properties and production are described in detail elsewhere (→ Copper Compounds, Chap. 3.1.).

Trade Names. Copper-Sandoz (Novartis); Copper Nordox (Nordox).

6.2. Organometallic Compounds

Organotin compounds such as triphenyltin hydroxide and triphenyltin acetate act as fungicides. They are used mostly for the control of *Cercospora beticola* on sugar beet and late blight on potatoes. They are phytotoxic if used at high doses and are, therefore, often not used throughout the growing season or, if so, are mixed with dithiocarbamates. Restrictions on their use are enforced in some countries for toxicological or environmental reasons. Organotin compounds exert their fungitoxicity as specific inhibitors of respiration [16, p. 149].

Fentin hydroxide [*76-87-9*], triphenyltin hydroxide, $C_{18}H_{16}OSn$, M_r 367, *mp* 116 – 120 °C, vapor pressure 47 µPa, consists of colorless crystals which are practically insoluble in water but moderately soluble in acetone, dichloromethane, 1,2-dichloroethane, diethyl ether, ethanol, and benzene [9, p. 271], [13, p. 149], [14, p. 130].

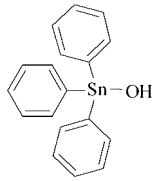

Fentin hydroxide is produced by treatment of triphenyltin chloride with potassium hydroxide [17].

Trade Names. Brestan Flow (AgrEvo); Suzu-H (Nihon Nohyaka).

6.3. Dithiocarbamates

Dithiocarbamates are the most widely used group of fungicides, even though mancozeb, the most important compound, was discovered in the 1950. Mancozeb and maneb together account for the bulk of consumption but other products such as zineb, propineb, polyram, and ziram are also commonly used (see also → Dithiocarbamic Acid and Derivatives, Chap. 5.). They have a very broad spectrum of activity and are generally well tolerated by many crops. They are cheap and effective, and constitute the major component of many routine, protective, spray programs. They are being used increasingly as mixtures with systemic fungicides to reduce the risk of resistance and to broaden the spectrum of activity of the systemic compounds.

In the plant, dithiocarbamates are probably converted to their isothiocyanate radicals, which then combine non-specifically with cell constituents such as sulfhydryl-containing enzymes [10, p. 148].

Mancozeb [*8018-01-7*], manganese ethylenebis(dithiocarbamate) (polymeric) complex with zinc salt; the complex contains about 20 % manganese and 2.55 % zinc, flash point 137.8 °C; it is a greyish yellow powder which is insoluble in water and most organic solvents [9, p. 339], [13, p. 66], [14, p. 140].

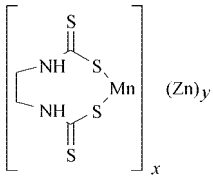

Mancozeb is produced by reaction of a concentrated solution of zinc sulfate with a suspension of maneb (see below) in water [18].
Trade Names. Dithane M 45 (Rohm & Haas); Manzate (Du Pont); Micane (Sipcam); Vondozeb (Elf Atochem).

Maneb [*12427-38-2*], manganese ethylenebis (dithiocarbamate) (polymeric), $C_4H_6MnN_2S_4$, M_r 265.3, *mp* decomposes on heating without melting, d_4^{25} 1.92, is a yellow crystalline solid which is almost insoluble in water and most organic solvents [9, p. 340], [13, p. 66], [14, p. 139].

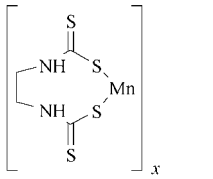

Maneb is produced by reaction of disodium ethylenebis(dithiocarbamate) and manganese sulfate [19].
Trade Names. Manzate (Du Pont); Dithane M 22 (Rohm & Haas); Bresten (AgrEvo).

Thiram [*137-26-8*], tetramethylthiuram disulfide, $C_6H_{12}N_2S_4$, M_r 240.4, *mp* 155 – 156 °C, consists of colorless crystals which are practically insoluble in water but soluble in chloroform and acetone [9, p. 534], [13, p. 62], [14, p. 137], [20].

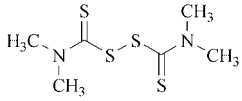

Thiram is produced by careful oxidation of a *N,N*-dimethyl dithiocarbamate salt with hydrogen peroxide, chlorine, or air [21].

Thiram was discovered in 1931 but is still popular today, especially as a broad-spectrum seed treatment for many crops. It is also used as a protective foliar fungicide on vegetables, turf, apples, and bananas, as well as a preplant dip treatment for ornamental bulbs. Crops on which it is used as a soil fungicide include vegetables and cotton.
Trade Names. Pomarsol (Bayer); Aetiram (AgrEvo); Rhodiason (Rhône-Poulenc); Thiram (Uniroyal).

Ferbam [*14486-64-1*], iron tris(dimethyldithiocarbamate), $C_9H_{18}N_3S_6Fe$, M_r 416.5, *mp* > 180 °C (decomp.), is a black powder which is slightly soluble in water and soluble in organic solvents with a high dielectric constant [22, p. 541].

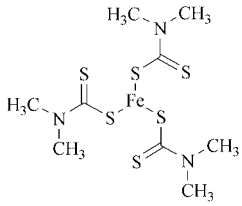

Ferbam is produced by precipitation from an aqueous solution of sodium dimethyldithiocarbamate in the presence of a water-soluble iron salt [23].

Ferbam (announced in 1965) is a foliar fungicide with protective action and secondary action against chlorosis. It is used to control scab on pome fruits, peach leaf curl (*Taphrina deformans*), blue mold on tobacco, and other diseases on other crops.
Trade Name. Ferbam Granuflo (UCB).

Metiram [*9006-42-2*], zinc ammoniate ethylenebis(dithiocarbamate) poly(ethyl-

enethiuram disulfide), $(C_{16}H_{33}N_{11}S_{16}Zn_3)_x$, M_r (1088.7)$_x$, *mp* 156 °C (decomp.), is a yellowish powder which is practically insoluble in water, and most organic solvents but soluble in pyridine (with decomposition) [22, p. 828].

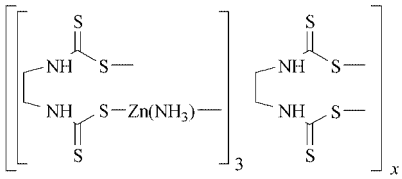

Metiram is produced by addition of a mixture of concentrated sulfuric acid, hydrogen peroxide, and zinc chloride with vigorous stirring to an aqueous solution of *N,N*-tetramethylenedithiocarbamate and sodium ethylenebis(dithiocarbamate) [24].

Metiram (announced in 1958) inactivates fungal proteins by binding to essential thiol groups. It is used to control a wide range of diseases of many crops: scab on pome fruits; shot-hole on stone fruit; rust on currants and plums; downy mildew, red fire disease, and black rot on vines; *Phytophthora* late blight on potatoes and tomatoes; *Septoria* leaf spot on celery and celeriac; and downy mildews on hops and tobacco.
Trade Name. Polyram (BASF).

Propineb [*9016-72-2*], polymeric zinc propylenebis(dithiocarbamate), $(C_5H_8N_2S_4Zn)_x$, M_r (289.8)$_x$, *mp* > 150 °C (decomp.), is a white to yellowish powder which is practically insoluble in all common organic solvents but fairly soluble in a mixture of DMF and dimethyl sulfoxide [22, p. 1032].

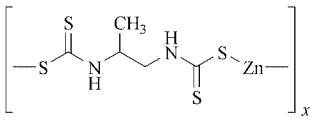

Propineb is produced by reaction of an aqueous solution of 1,2-propylenediamine with carbon disulfide in NaOH. The zinc salt is obtained by precipitation with zinc sulfate [25].

Propineb (announced in 1963) is a foliar fungicide used to control downy mildew, black rots, red fire disease, and grey mold on vines; leaf spot diseases on stone fruits; and *Alternaria* and *Phytophthora* blights, downy mildew, *Septoria* leaf spot, and leaf mold in tomatoes.
Trade Name. Antracol (Bayer).

Zineb [*12122-67-7*] polymeric zinc ethylenebis(dithiocarbamate), $(C_4H_6N_2S_4Zn)_x$, M_r (275.8)$_x$, *mp* 157 °C (decomp.), is a light colored powder which is practically insoluble in water and common organic solvents [22, p. 1276.]

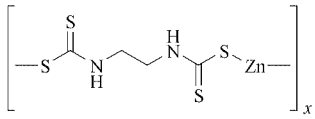

Zineb is produced by mixing ethylenebis(dithiocarbamate) with an aqueous zinc sulfate solution [26].

Zineb (announced in 1943) is used to control downy mildews in vines, hops, lettuce, onions, spinach, brassicas, oilseed rape, tobacco and ornamentals; rusts in currants, berry fruits, plums, vegetables, and ornamentals; red fire disease on ornamentals; potato and tomato blights; *Cercospora* on bananas; *Botrytis*; and needle cast in forestry.
Trade Names. Dithane Z-78 (Rohm & Haas); Aphytora (AgrEvo); Amitan (Sipcam).

Ziram [*137-30-4*], zinc bis(dimethyldithiocarbamate), $C_6H_{12}N_2S_4Zn$, M_r 305.8, *mp* 246 °C, is a colorless powder which is practically insoluble in water, slightly soluble in ether and ethanol; and soluble in carbon disulfide and chloroform [22, p. 1277.]

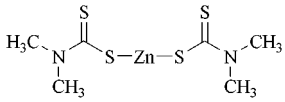

Ziram is produced by precipitation from an aqueous solution of sodium diemthyldithiocarbamate in the presence of a water-soluble zinc salt [27].

Ziram (announced in 1962) is an inhibitor of enzymes containing copper ions or sulfhydryl groups. It is used to control fungicidal diseases in pome fruits, stone fruits, nuts, vines, vegetables, and ornamentals (especially scab in apples and pears), *Monilinia*, *Alternaria*, *Septoria*, peach leaf curl, shot-hole, rusts, black rot, and anthracnose. It is also used as a wildlife repellent.
Trade Names. Cuman (Novartis); Fuclasin (AgrEvo); Pomarsol Z (Bayer).

6.4. *N*-Trihalomethylthio Derivatives (Phthalimides)

Four main phthalimide fungicides were introduced between 1952 and 1961. Captan and folpet are the most important, followed by captafol and ditalimfos; they are used widely as protectant fungicides. Concern has arisen about their safety, especially that of captafol. Restrictions on use have been imposed in some countries, while total bans have been imposed in others.

The fungitoxicity of these compounds is nonspecific and is probably due to enzyme inactivation caused by blocking of sulfhydryl groups.

Captan [*133-06-2*], *N*-(trichloromethylthio)cyclohex-4-ene-1,2-dicarboximide, $C_9H_8Cl_3NO_2S$, M_r 300.6, *mp* 178 °C, vapor pressure < 1.3 mPa at 25 °C, consists of colorless crystals which are practically insoluble in water but partially soluble in acetone, chloroform, cyclohexanone, and xylene [9, p. 87], [13, p. 108], [14, p. 190].

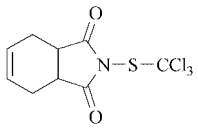

Captan is produced by reaction of 1,2,3,6-tetrahydrophthalimide with trichloromethanesulfenyl chloride (perchloromethylmercaptan) in the presence of sodium hydroxide [28].

Captan is used as a foliar fungicide for the control of scab (*Venturia inaequalis*) of apples and pears, as well for other diseases of fruit, vegetable, and ornamental crops. It is also used as a seed or planter-box treatment, as a preplant soil treatment, or as a postharvest fungicide.

Trade Names. Phytocape (Bayer); Captof (Rallis).

Folpet [*133-07-3*], *N*-(trichloromethylthio)-phthalimide, $C_9H_4Cl_3NO_2S$, M_r 296.6, *mp* 177 – 180 °C (decomp.), consists of white crystals which are practically insoluble in water and slightly soluble in organic solvents [13, p. 109], [14, p. 191], [29, p. 518].

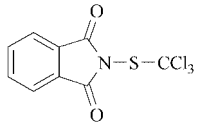

Folpet is produced by reaction of the potassium salt of phthalimide with trichloromethanesulfenyl chloride in organic solvents [28].

Folpet is used as a foliar fungicide on grapevines, fruit trees, and vegetables, and is active against a wide range of diseases including downy mildew, leaf spot, powdery mildew, and rust as well as black rot and dead arm of grapes.

Trade Name. Folpan (Makhteshim-Agan, Philagro).

Dichlofluanid [*1085-98-9*], *N*-dichlorofluoromethylthio-*N′*,*N′*-dimethyl-*N*-phenylsulfamide, $C_9H_{11}Cl_2FN_2O_2S_2$, M_r 333.2, *mp* 105 – 106 °C, vapor pressure 133 µPa at 20 °C, is a colorless powder which is practically insoluble in water, slightly soluble in methanol and xylene, and soluble in dichloromethane [9, p. 175], [13, p. 95], [14, p. 193].

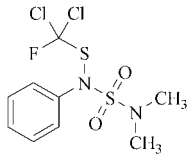

Dichlofluanid is produced by reaction of *N′*,*N′*-dimethyl-*N*-phenylsulfamide with dichlorofluoromethanesulfenyl chloride in the presence of a base [30].

Dichlofluanid (announced in 1965) is used as a preventative foliar fungicide against *Botrytis* and downy mildew on grapevines. It has largely been replaced by more effective compounds (e.g., dicarboximides), but interest has been renewed recently because of resistance to dicarboximides.

Trade Names. Euparen, Elvaron (Bayer).

Captafol [*2425-06-1*], *N*-(1,1,2,2-tetrachloroethylthio)cyclohex-4-ene-1,2-dicarboximide, $C_{10}H_9Cl_4NO_2S$, M_r 349.1, *mp* 160 – 161 °C, is a light tan powder which is practically insoluble in water and slightly soluble in most organic solvents [22, p. 175].

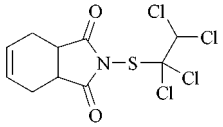

Captafol is produced by reaction of the sodium salt of tetrahydrophthalimide in benzene with tetrachloroethylsulfenyl chloride [31].

Captafol (announced in 1962) acts by inhibiting germination of spores. It is used to control scab on pome fruits, shot-hole in some stone fruit; peach leaf curl, downy mildew and black rot of vines; *Alternaria* and mildew of carrots; *Septoria* of wheat; *Rhynchosporium* of barley; and various diseases of vegetables and fruits.
Trade Names. Difoltan, Foltaf (Rallis).

Tolylfluanid [*731-27-1*], *N*-dichlorofluoromethylthio-*N′*,*N′*-dimethyl-*N*-*p*-tolylsulfamide, $C_{10}H_{13}Cl_2FN_2O_2S_2$, M_r 347.2, *mp* 93 °C, is a colorless to pale yellow powder which is practically insoluble in water, and soluble in organic solvents [22, p. 1208.]

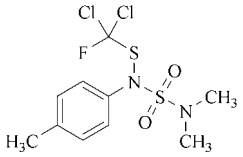

Tolylfluanid is produced by reaction of the potassium salt of *N′*,*N′*-dimethyl-*N*-*p*-tolylsulfamide with dichlorofluoromethylthiochloride in xylene with heating [32].

Tolylfluanid (announced in 1967) is used in deciduous fruit crops to control *Venturia* spp. and *Podosphaera leucotricha*. It is also used against storage diseases; in grapes against *Uncinula necator*, *Plasmopara viticola*, and *Botrytis cinerea*; and in vegetables against *B. cinerea*, *Peronospora* spp., and powdery mildew.
Trade Names. Elvaron M, Euparen M (Bayer).

6.5. Dicarboximides

The dicarboximide fungicides include iprodione, procymidone, vinclozolin, and metomeclan, which are all *N*-3,5-dichlorophenyl derivatives of cyclic imides. They are used to control diseases caused by the taxonomically related genera *Sclerotinia, Monilinia*, and *Botrytis*. They are generally considered nonsystemic, protectant fungicides, but evidence exists for systemic transport in some plants. They share a common mode of action with a probable primary effect on nuclei, but they may also affect the cytoplasmic membrane and/or the cell wall [16, pp. 207 – 237]. Resistance problems have occurred in several field crops, notably in *Botrytis* spp. [33, p. 101].

Iprodione [*36734-19-7*], 3-(3,5-dichlorophenyl)-*N*-isopropyl-2,4-dioxoimidazolidine-1-carboxamide, $C_{13}H_{13}Cl_2N_3O_3$, M_r 330.2, *mp* 136 °C, vapor pressure 133 µPa at 20 °C, consists of colorless crystals which are practically insoluble in water, slightly soluble in ethanol and methanol, and soluble in many organic solvents such as acetone or dichloromethane [9, p. 322], [14, p. 209].

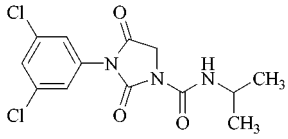

Iprodione is produced by reaction of 3,5-dichlorophenyl isocyanate with glycine, followed by ring closure to give the hydantoin, then treatment with isopropyl isocyanate [34].

Iprodione (announced in 1970) has both preventative and curative properties. It is used against turf grass diseases, *Alternaria* diseases of apples, pears, and oilseed rape, as well as *Botrytis, Monilinia,* and *Sclerotinia* diseases of grapevines, fruit, and vegetables.
Trade Names. Rovral, Kidan, Verisan (Rhône-Poulenc).

Vinclozolin [*50471-44-8*], (*RS*)-3-(3,5-dichlorophenyl)-5-methyl-5-vinyl-1,3-oxazolidine-2,4-dione, $C_{12}H_9Cl_2NO_3$, M_r 286.1, *mp* 108 °C, vapor pressure < 10 mPa at 20 °C, consists of colorless crystals which are slightly soluble in water but more soluble in acetone, benzene, chloroform, and ethyl acetate [9, p. 558], [14, p. 217].

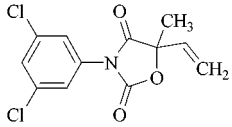

Vinclozolin is produced by reaction of 3,5-dichlorophenyl isocyanate with an alkyl ester of

2-hydroxy-2-vinylpropionic acid; ring closure is achieved at elevated temperature [35].

Vinclozolin (announced in 1973) has a similar spectrum of activity to iprodione and is used to control specific diseases of rape, grapevines, apples, and pears. Although it is a nonsystemic fungicide, it also has a local systemic effect and displays curative properties up to 36 h after infection.

Trade Names. Ronilan (BASF); Flotilla (Barclay).

Procymidone [*32809-16-8*], *N*-(3,5-dichlorophenyl)-1,2-dimethylcyclopropane-1,2-dicarboximide, $C_{13}H_{11}Cl_2NO_2$, M_r 284.1, *mp* 164–166 °C, vapor pressure 10.5 mPa at 20 °C, is a light-brown solid which is practically insoluble in water but slightly soluble in methanol or xylene [9, p. 453].

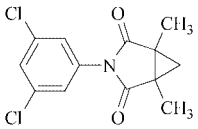

Procymidone is produced by reaction of 1,2-dimethylcyclopropanedicarboxylic anhydride and 3,5-dichloroaniline; the resulting monoamide is then cyclized with acetic anhydride and a catalytic amount of anhydrous sodium acetate at 100 °C [36].

Procymidone (announced in 1976) has a similar spectrum of activity to vinclozolin and iprodione. It has systemic properties and is used as a foliar fungicide with protective and eradicative properties.

Trade Names. Sumisclex, Sumilex (Sumitomo).

Chlozolinate [*72391-46-9*], ethyl(±)-3-(3,5-dichlorophenyl)-5-methyl-2,4-dioxo-1,3-oxazolidine-5-carboxylate, $C_{13}H_{11}Cl_2NO_5$, M_r 332.1, *mp* 112.6 °C, is an almost colorless powder which is practically insoluble in water, and soluble in most organic solvents [22, p. 244].

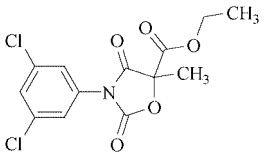

Chlozolinate is produced by reaction of 3,5-dichlorophenylisocyanate with hydroxymethyl diethylmalonate in benzene followed by the addition of triethylamine [37].

Chlozolinate (announced in 1980) is used to control *Botrytis* spp., *Sclerotina* spp., and *Monilinia* spp. in ornamentals, vegetables, vines, pome fruits, stone fruits, sunflowers, and strawberries.

Trade Name. Serinal (Isagro).

6.6. Miscellaneous Organic Compounds

Quintozene [*82-68-8*], pentachloronitrobenzene, $C_6Cl_5NO_2$, M_r 295.3, *mp* 146 °C, d_4^{25} 1.718, vapor pressure 1.8 Pa at 25 °C, consists of colorless crystals which are practically insoluble in water but soluble in benzene and chloroform [9, p. 484], [13, p. 73], [14, p. 154].

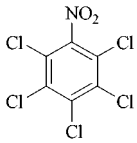

Quintozene is prepared by chlorination of nitrobenzene at 60–70 °C in chlorosulfuric acid, with iodine as a catalyst [38].

Quintozene was discovered in the 1930s but is still used frequently as a soil fungicide or seed treatment. It is particularly effective for the control of soil-borne *Rhizoctonia* and *Sclerotinia* spp. Very high rates are required when it is applied to the soil during or prior to planting.

Trade Names. Brassicol (AgrEvo); Terrachlor, Folosan (Uniroyal); Kobutol (Hokko).

Dodine [*2439-10-3*], 1-dodecylguanidinium acetate, $C_{15}H_{33}N_3O_2$, M_r 287.4, *mp* 136 °C, consists of colorless crystals which are practically insoluble in water and most organic solvents but soluble in hot alcohol and hot water [9, p. 229], [13, p. 70], [14, p. 145], [15, p. 374].

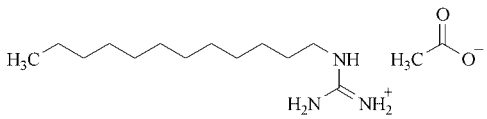

Dodine is produced by reaction of 1-dodecylamine with cyanamide and acetic acid [39].

Dodine (announced in 1959) has preventative as well as eradicative properties because of its

local systemic activity. It is used mainly for control of apple scab but also for fruit rot and leaf spot diseases of deciduous fruit and nuts.
Trade Names. Cyprex, Venturol, Melprex (Cyanamid, Rhône-Poulenc); Dodene (Sipcam).

Dithianon [*3347-22-6*], 5,10-dihydro-5,10-dioxonaphtho-[2,3-*b*]-1,4-dithiin-2,3-dicarbonitrile, $C_{14}H_4N_2O_2S_2$, M_r 296.3, *mp* 225 °C, vapor pressure 66 µPa at 25 °C, consists of brown crystals which are practically insoluble in water and only slightly soluble in organic solvents [9, p. 225], [13, p. 105].

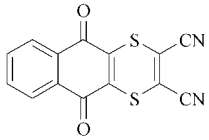

Dithianon is produced by reaction of hydrogen cyanide with carbon disulfide in dimethylformamide to yield dimercaptomaleic acid dinitrile; this intermediate is then treated with 2,3-dichloronaphthoquinone [40].

Dithianon (announced in 1963) is a protective contact fungicide with a broad spectrum of activity against, for example, leaf and fruit spot diseases, downy mildew, late blight, anthracnose, and rust. It is relatively expensive and is used on crops (e.g., deciduous fruit trees and coffee) only when it produces significant benefits that are not obtained with cheaper contact fungicides such as dithiocarbamates.
Trade Names. Actuan, Delen (Cyanamid); Cluster (Barclay).

Chlorothalonil [*1897-45-6*], tetrachloroisophthalonitrile, $C_8Cl_4N_2$, M_r 265.9, *mp* 250–251 °C, *bp* 350 °C, vapor pressure 1.3 Pa at 40 °C, consists of colorless crystals which are practically insoluble in water but slightly soluble in organic solvents such as acetone, cyclohexanone, and xylene [9, p. 120], [13, p. 91], [14, p. 163].

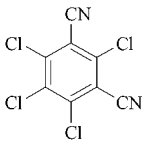

Chlorothalonil is produced by reaction of tetrachloroisophthalyl chloride with ammonia to give the diamide which, on treatment with phosphorus oxychloride, yields the dinitrile [41].

Chlorothalonil (announced in 1964) has a broad spectrum of activity and is used as a preventative foliar fungicide on a wide range of crops including vegetables, turf, top fruit, coffee, groundnuts, and cereals. It is relatively expensive and is usually used on more valuable cultures.
Trade Names. Bravo (ISK Biosciences); Daconil (SDS Biotech, Zeneca, ISK Biosciences).

Etridiazole [*2593-15-9*], 5-ethoxy-3-trichloromethyl-1,2,4-thiadiazole, $C_5H_5Cl_3N_2OS$, M_r 247.5, *mp* 20 °C, *bp* 95 °C at 0.13 kPa, vapor pressure 13 mPa at room temperature, is a yellow liquid with a persistent odor; it is practically insoluble in water but soluble in many organic solvents [9, p. 252], [14, p. 219], [15, p. 397].

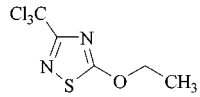

Etridiazole is produced by reaction of trichloroacetamidine hydrochloride with trichloromethanesulfenyl chloride and then with sodium hydroxide in ethanol [42].

Etridiazole (announced in 1969) is a soil fungicide with preventative and curative properties. It is effective against *Phytophthora* and *Pythium* spp. and is used on turf, ornamentals, vegetables, avocados, and other crops. It can be applied to the soil surface or mixed with the soil prior to planting.
Trade Name. Terrazole (Uniroyal).

Phthalide [*27355-22-2*], 4,5,6,7-tetrachlorophthalide, $C_8H_2Cl_4O_2$, M_r 271.9, *mp* 209–210 °C, consists of colorless crystals which are practically insoluble in water and only slightly soluble in organic solvents [9, p. 517], [14, p. 197].

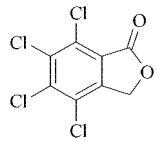

Phthalide is produced by treatment of 1,2-bis(dichloromethyl)-3,4,5,6-tetrachlorobenzene with 75 % sulfuric acid [43].

Phthalide (announced in 1971) is a foliar fungicide with preventative properties for the control of blast in rice.
Trade Names. Rabcide (Kureha); Kasurabcide (Hokko).

Toclofos-methyl [*57018-04-9*], *O*-2,6-dichloro-*p*-tolyl-*O,O*-dimethyl phosphorothioate (or thiophosphate), $C_9H_{11}Cl_2O_3PS$, M_r 301.1, *mp* 78 – 80 °C, vapor pressure 57 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in acetone, cyclohexanone, and xylene [4, p. 536].

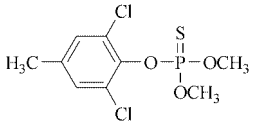

Toclofos-methyl is produced by reaction of 2,6-dichloro-*p*-cresol with (*O,O*-dimethyl) chlorothiophosphate in the presence of a base [44].

Toclofos-methyl (announced in 1973) is a contact fungicide for the control of soil-borne diseases caused by *Rhizoctonia solani* and by *Sclerotium* and *Typhula* spp. on potatoes, sugar beet, cotton, groundnuts, vegetables, cereals, and turf. It is used as a seed treatment and a soil fungicide.
Trade Name. Rizolex (Sumitomo, Du Pont).

Cymoxanil [*57966-95-7*], 1-(2-cyano-2-methoxyiminoacetyl)-3-ethylurea, $C_7H_{10}N_4O_3$, M_r 198.2, *mp* 160 – 161 °C, d^{25} 1.31, vapor pressure 80 µPa at 25 °C, is a colorless crystalline solid which is slightly soluble in water and benzene but soluble in acetone, chloroform, and dimethylformamide [9, p. 148], [14, p. 146], [15, p. 405].

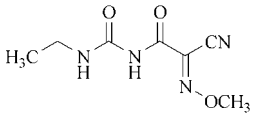

Cymoxanil is produced by condensation of cyanoacetic acid with ethylurea in the presence of acetic anhydride; the resulting 2-cyano-*N*-ethylcarbamoyl acetamide is converted to the hydroximino derivative which is methylated with dimethyl sulfate [45].

Cymoxanil (announced in 1977) has local penetrating properties and a curative mode of action. Persistence in the plant is relatively short (3 – 5 days), especially at higher temperatures. Therefore, it is used only in mixtures with contact fungicides to which it gives a "kick-back" (curative) action. It is used against downy mildew of grapes and late blight of potatoes.
Trade Names. Curzate (Du Pont); Pulsan (Novartis).

Pencycuron [*66063-05-6*], 1-(4-chlorobenzyl)-1-cyclopentyl-3-phenylurea, $C_{19}H_{21}ClN_2O$, M_r 328.8, vapor pressure < 1 mPa at 20 °, consists of colorless crystals which are practically insoluble in water but soluble in dichloromethane [9, p. 422].

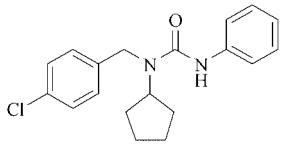

Pencycuron is produced by reaction of *N*-(4-chlorobenzyl)cyclopentylamine with phenyl isocyanate [46].

Pencycuron (announced in 1981) is used for the control of diseases caused by *Rhizoctonia solani,* such as rice sheath blight, black scurf of potatoes, and damping-off of ornamentals.
Trade Name. Monceren (Bayer).

Fenpiclonil [*74738-17-3*], 4-(2,3-dichlorophenyl)pyrrole-3-carbonitrile, $C_{11}H_6Cl_2N_2$, M_r 237.1, *mp* 153 °C, vapor pressure 42 µPa at 20 °C, consists of colorless crystals which are practically insoluble in water, slightly soluble in methanol and dichloromethane, and soluble in acetone and cyclohexanone [47, p. 65].

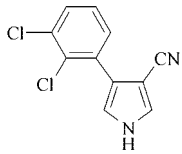

Fenpiclonil can be produced by reaction of either 2-(2,3-dichlorophenyl)acrylonitrile or 1-cyano-2-(2,3-dichlorophenyl)ethyl cinnamate with *p*-toluenesulfonylmethyl isocyanate in dimethoxyethane in the presence of a base such as sodium hydride or sodium methoxide [48].

Fenpiclonil was announced in 1988 as a seed treatment fungicide. On cereals it has been introduced for the control of *Gerlachia nivalis (Fusarium nivale)* and *Tilletia caries* as well as seed-borne *Septoria nodorum.* In combination with imazalil it allows control of *Pyrenophora graminea* and *Pyrenophora teres* in barley. It is also being developed for use on noncereal crops where it shows activity against a broad spectrum of Ascomycetes, Deuteromycetes and Basidiomycetes pathogens.
Trade Names. Beret, Gambit (Novartis).

Biphenyl [*92-52-4*], $C_{12}H_{10}$, M_r154.2, *mp* 69 – 71 °C, consists of colorless crystals which are practically insoluble in water but readily soluble in most organic solvents [22, p. 128.]

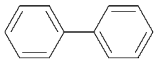

Biphenyl is produced by reaction of phenylmagnesium bromide with 1,4-dimethyl-2,3-dibromobutane in an inert solvent [49].

Biphenyl (announced in 1944) is a fungistat that inhibits sporulation. It is used to impregnate citrus fruit wraps to prevent fungal attack.

2-Phenylphenol [*90-43-7*], biphenyl-2-ol, $C_{12}H_{10}O$, M_r 170.2, *mp* 57 °C, is a colorless crystalline solid which is slightly soluble in water and soluble in most organic solvents [22, p. 955.]

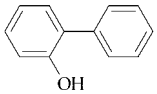

2-Phenylphenol is produced by heating 6-phenyl-2,5-hexadienoic acid in a mixture of acetic hydride, acetic acid, and sodium acetate [50].

2-Phenylphenol (announced in 1936) is used as a desinfectant fungicide in post-harvest control of storage dieseases of apples, citrus fruits, tomatoes, cucumbers, and capsicums.

Diphenylamine [*122-39-4*], $C_{12}H_{11}N$, M_r 169.2, *mp* 53 – 54 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in organic solvents [22, p. 435].

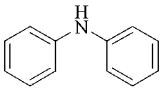

Diphenylamine is produced by heating a mixture of aniline hydrochloride and aniline, and extraction of the product with hydrchloric acid, followed by distillation of the extract [51].

Diphenylamine (announced in 1940) is used as a post-harvest fungicide protectant and scald inhibitor for pome fruit.
Trade Names. Coraza (Productos OSA); No Scald (Elf Atochem).

***sec*-Butylamine** [*13952-84-6*], (*R,S*)-*sec*-butylamine, $C_4H_{11}N$, M_r 73.1, *bp* 63 °C, is a colorless liquid which is miscible with water and with most organic solvents [22, p. 169].

sec-Butylamine is produced by is produced by heating a mixture of *sec*-butylbromide, potassium isocyanate, and ethanol in DMF, followed by hydrolysis of the resulting urethane [52].

sec-Butylamine (announced in 1962) is used in fungicidal fumigation of seed and ware potatoes for control of skin spot and gangrene; and against citrus blue, green molds and stem rot of oranges.
Trade Name. Deccotane (Elf Atochem).

Chinomethionat [*2439-01-2*], *S,S*-(6-methylquinoxaline-2,3-diyl) dithiocarbamate, $C_{10}H_6N_2OS_2$, M_r 234.3, *mp* 170 °C, consists of yellow crystals which are sparingly soluble in water but soluble in cyclohexanone and DMF [22, p. 198].

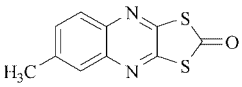

Chinomethionat is produced by reaction of 7-methyl-2,3-quinoxaline thiol with carbonyl chloride [53].

Chinomethionat (announced in 1960) is used to control powdery mildews and spider mites on fruit, ornamentals, cucurbits, cotton, coffee, tea, tobacco, walnuts, vegetables, and glasshouse crops.
Trade Name. Morestan (Bayer).

Dichlorophen [*97-23-4*], 4,4′-dichloro-2,2′-methylenediphenol, $C_{13}H_{10}Cl_2O_2$, M_r 269.1,

mp 177 – 178 °C, consists of colorless crystals which are sparingly soluble in water but soluble in acetone and ethanol [22, p. 364].

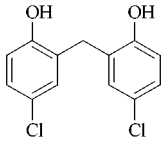

Dichlorophen is produced by gradually adding an aqueous solution of formaldehyde to a mixture of *p*-chlorophenol and sulfuric acid in methanol [54].

Dichlorophen (announced in 1946) is used to control moss, red thread, *Fusarium* patch, and dollar spot in turf; and against moss on paths, walls, roofs, and noncrop land.
Trade Name. Super Mosstox (Rhône-Poulenc).

Diclomezine [*62865-36-5*], 6-(3,5-dichloro-4-methylphenyl)pyridazin-3(2*H*)-one, $C_{11}H_8Cl_2N_2O$, M_r 255.1, *mp* 250.5 – 253.5 °C, consists of colorless crystals which are sparingly soluble in water and fairly soluble in methanol and acetone [22, p. 377].

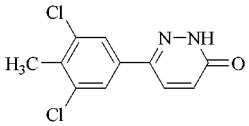

Diclomezine is produced by the Friedel – Crafts reaction of toluene with succinic anhydride, followed by chlorination to give 4-methyl-3,5-dichlorobenzoylpropionic acid, which is further chlorinated to yield 4-methyl-3,5-dichlorobenzoyl-2-chloropropionic acid. The latter is then treated with hydrazine [55].

Dichlomezine (announced in 1988) inhibits septum formation and mycelial growth. It is used to control *Rhizoctonia* and *Sclerotinium* spp. in rice and white mold and twig rot on peanuts.
Trade Name. Monguard (Sankyo).

Dicloran [*99-30-9*], 2,6-dichloro-4-nitroaniline, $C_6H_4Cl_2N_2O_2$, M_r 207.0, *mp* 195 °C, is a yellow crystalline solid which is slightly soluble in water and soluble in acetone, benzene, chloroform, cyclohexane, and dioxan [22, p. 378].

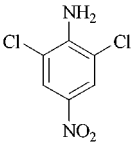

Dicloran is produced by chlorination of 4-nitro-*o*-anilinesulfonic acid to form the monochloro product, which is further chlorinated to yield dichloran [56].

Dicloran (announced in 1960) interferes with the transfer of ribonucleic acid complex to the ribosome and inhibits protein synthesis without affecting respiration. It is used to control *Botrytis*, *Monilinia*, *Rhizopus*, *Sclerotina*, and *Sclerotium* spp. on stone fruits, berry fruits, vines, rhubarb, potatoes, ornamentals, and lettuce.
Trade Names. Allisan, Botran (Gowen).

Fenfuram [*24691-80-3*], 2-methyl-3-furanilide, $C_{12}H_{11}NO_2$, M_r 201.2, *mp* 109 – 110 °C, is a cream-colored solid which is slightly soluble in water and soluble in acetone, cyclohexane, and methanol [22, p. 512].

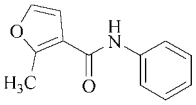

Fenfuram is produced by heating 2-methyl-4,5-dihydrofuran-3-carboxylic acid with sulfonyl chloride, followed by reaction with aniline [57].

Fenfuram (announced in 1970) is used as a seed treatment to control bunts and smuts (*Tilletia* and *Ustilago* spp.) in cereals.
Trade Name. Pano-ram (Rhône-Poulenc).

Ferimzone [*89269-64-7*], (*Z*)-2′-methylacetophenone-4,6-dimethylpyrimidin-2-yl hydrazone, $C_{15}H_{18}N_4$, M_r 254.3, *mp* 175 – 176 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in ethanol, ethyl acetate, chloroform, xylene, and acetonitrile [22, p. 543.]

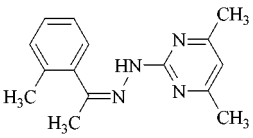

Ferimzone is produced by condensation of 2-methylacetophenone with 4,6-dimethyl-2-hydrazinopyrimidine [58].

Ferimzone (announced in 1984) is used to control *Helminthosporium oryzae*, *Cercospora oryzae*, and *Pyricularia oryzae* in rice.
Trade Name. Blasin (Takeda).

Fludioxonil [*131341-86-1*], 4-(2,2-difluoro-1,3-benzodioxol-4-yl)pyrrole-3-carbonitrile, $C_{12}H_6F_2N_2O_2$, M_r 248.2, *mp* 199.8 °C, consists of colorless crystals which are sparingly soluble in water, and hexane, moderately soluble in toluene, *n*-octanol and ethanol, and readily soluble in acetone [22, p. 566.]

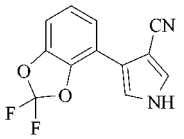

Fludioxonil is produced by reactionof 4-amino-2,2-difluorobenzodioxide in hydrochloric acid-containing acetic acid with sodium nitrite at 0 °C, followed by addition of ethylenenitrile, methyl ethyl ketone, and a solution of copper(I) chloride in hydrochloric acid. The product is extracted with dichloromethane and the thus-obtained solution refluxed in the presence of triethylamine to give 2,3-(difluoromethylenedioxy)cinnamic acid nitrile, which is treated with *p*-toluenesulfonylmethylisocyanide and potassium *tert*-butoxide in THF [59].

Fludioxonil (announced in 1991) is used as seed treatment for control of *Gibberella* in rice and to control *Fusarium*, *Rhizoctonia*, *Tilletia*, *Helminthosporium*, and *Septoria*, in cereal and noncereal crops.
Trade Names. Celest, Maxim (Novartis).

Fluoroimide [*41205-21-4*], 2,3-dichloro-*N*-4-fluorophenylmaleimide, $C_{10}H_4Cl_2FNO_2$, M_r 260.1, *mp* 140 – 214 °C, consists of yellow crystals which are sparingly soluble in water and soluble in methanol and acetone [22, p. 582.]

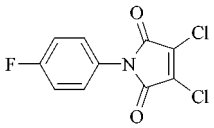

Fluoroimide is produced by reaction of 2,3-dichloromaleic acid in aqueous solution with *p*-fluoroaniline [60].

Fluoroimide (announced in 1978) inhibits sulfhydryl-containing enzymes in the pathogen. It controls *Monilinia mali* and *Mycosphaerella pomi* in apples and *Doaporthe citri* and *Elsinoe fawcetti* in citrus fruits.
Trade Name. Sparticide (Mitsubishi Chemical).

Flusulfamide [*106917-52-6*], 2',4-dichloro-α,α ,α-trifluoro-4′-nitro-*m*-toluenesulfonamide, $C_{13}H_7Cl_2F_3N_2O_4S$, M_r 415.2, *mp* 169-171 °C, consists of light yellow crystals which are sparingly soluble in water, fairly soluble in methanol, acetone, and THF, and readily soluble in xylene [22, p. 605].

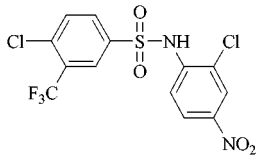

Flusulfamide is produced by reaction of 2-chloro-4-nitroaniline with 3-chloro-4-trifluoromethylphenylsulfenyl chloride in 1,3-dimethyl-2-imidazolidone in the presence of sodium hydride [61].

Flusulfamide (announced in 1987) is used as a soil treatment for control of *Plasmodiophora brassica* in brassica and *Polymyxa betae* in suger beet.
Trade Name. Nebijn (Mitsui Toatsu).

Mepanipyrim [*110235-47-7*], *N*-(4-methyl-6-prop-1-ynylpyrimidin-2-yl)aniline, $C_{14}H_{13}N_3$, M_r 223.3, *mp* 132.8 °C, is a colorless crystalline solid which is slightly soluble in water and fairly soluble in most organic solvents [22, p. 784].

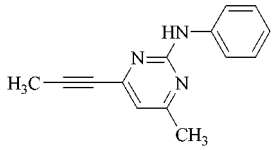

Mepanipyrim is produced by reaction of 4-methyl-6-prop-1-ynylpyrimidine with a mixture of NaOH and formanilide in THF at room tem-

perature, followed by the addition of 4N HCl and refluxing the mixture [62].

Mepanipyrim (announced in 1990) inhibits secretion of pathogen proteins such as the cell-wall degrading enzyme pectinase. It is used to control gray mold on vines, strawberries, tomatoes and cucumbers; scab on apples and pears; and brown rot on peaches.
Trade Name. Frupica (Kumiai).

Nitrothal-isopropyl [*10552-74-6*], diisopropyl 5-nitroisophthalate, $C_{14}H_{17}NO_6$, M_r 295.3, *mp* 65 °C, consists of yellow crystals which are sparingly soluble in water and soluble in acetone, benzene, chloroform, and ethyl acetate [22, p. 884].

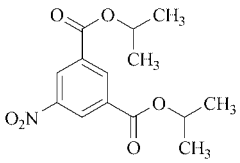

Nitrothal-isopropyl is produced by reaction of 5-nitroisophthaloyl chloride with isopropanol in an inert solvent [63].

Nitrothal-isopropyl (announced in 1973) is used in combination with other fungicides for control of powdery mildews on apples, vines, hops, vegetables, and ornamentals and scab on apples.
Trade Names. Kumulan, Pallined (BASF).

Oxycarboxin [*5259-88-1*], 5,6-dihydro-2-methyl-1,4-oxathi-ine-3-carboxanilide 4,4-dioxide, $C_{12}H_{13}NO_4S$, M_r 267.3, *mp* 119.5 – 122.5 °C, consists of colorless crystals which are moderately soluble in water, and fairly soluble in acetone, DMF, ethanol, and methanol [22, p. 916].

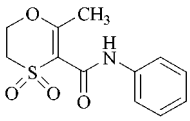

Oxycarboxin is produced by oxidation of oxathiin, which is prepared by reaction of acetoacetanilide with sulfonyl chloride, followed by treatment with a potassium hydroxide solution and then with 2-mercaptoethanol [64].

Oxycarboxin (announced in 1966) is used to control rust diseases on ornamentals, cereals, and nursery trees; and fairy rings on turf.
Trade Names. Plantvax (Uniroyal); Carbojet (Maughet).

Triazoxide [*72459-58-6*], 7-chloro-3-imidazol-1-yl-1,2,4-benzotriazine 1-oxide, $C_{10}H_6ClN_5O$, M_r 247.7, *mp* 182 °C, consists of light yellow crystals which are sparingly soluble in water and moderately soluble in hexane, isopropanol, toluene, and dichloromethane [22, p. 1228].

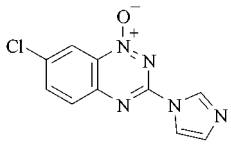

Triazoxide is produced by reaction of 3,7-dichloro-1,2,4-benzotriazine 1-oxide with imidazole in pyridine [65].

Triazoxide (announced in 1980) is used as a seed dressing for control of seed-borne *Pyrenophora graminea* and *Pyrenophora teres* in barley. Only used in combination with other fungicides.
Trade Names. Brio, Gaucho Orge, Raxil S (Bayer).

7. Fungicidal Antibiotics

The antibiotics dealt with in this chapter are described in more detail elsewhere (→ Antibiotics).

The earliest antibiotic used in agriculture, griseofulvin, was developed in England in 1939 by the British Ministry of Agriculture. It is a systemic fungicide that acts against *Botrytis* spp. and powdery mildew diseases but has only been used to a very limited extent on vegetable crops. The next important discovery was that of cycloheximide in 1946 by The Upjohn Company. Although its spectrum of activity was broader than that of griseofulvin, practical use was limited by its phytotoxicity. The success of antibiotics in human medicine stimulated a more intensive search for agricultural fungicides in the 1950s, especially for compounds that could be used to control rice diseases. This led in 1959 to the discovery of blasticidin S, a compound with

nonsystemic, preventative properties which was used for the control of rice blast. This compound was superseded by kasugamycin (announced in 1965). Like blasticidin, kasugamycin acts by inhibiting protein synthesis and is used to control rice blast. It, in turn, has largely been replaced by more modern compounds but remains important in some countries because of its low price. Resistance to kasugamycin has been found under field conditions, but resistant strains are somewhat less biologically fit than sensitive ones [33, p. 207].

The limited number of antibiotics developed to control plant pathogens is due partially to their high cost and to the incompatibility of using them in both medicine and agriculture. In some countries, however, streptomycin is used for both purposes. The most successful antibiotics for agricultural use are the polyoxins and validamycin.

Polyoxin B [*19396-06-6*], $C_{17}H_{25}N_5O_{13}$, M_r 507.4, *mp* 160°C, is an amorphous powder which is soluble in water but insoluble in most organic solvents [9, p. 450], [14, p. 230], [15, p. 226], [66]. See also → Antibiotics.

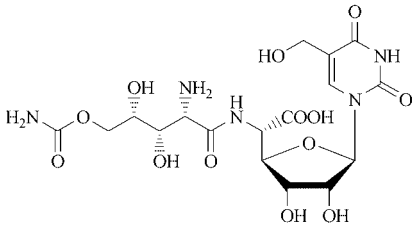

Polyoxin B is produced by fermentation of *Streptomyces cacaoi* var. *asoensis* [67]. About 15 structurally related polyoxins are known, with polyoxins B and D having the greatest commercial importance.

Polyoxin B, a streptomycin antibiotic, was introduced in 1965 as a foliar fungicide and is somewhat systemic in activity. It is used to control black spot of Japanese pear caused by *Alternaria kikuchiana* and also to control rice blast. Four years after it was first used on pears, poor field performance occurred as a result of resistance developed. However, the proportion of fungicide-resistant strains dropped when use was discontinued [16, p. 105]. Polyoxins act by inhibiting chitin synthesis.

Trade Name. Polyoxin AL (Kaken, Kumiai, Nihon, Nohyaku, Hokko).

Validamycin [*37248-47-8*], $C_{20}H_{35}NO_{13}$, M_r 497.5, decomposes over a wide temperature range around 130 °C, $[\alpha]_D^{24}$ +110° (water); it is a colorless powder which is soluble in water, dimethylformamide, dimethyl sulfoxide, and methanol [9, p. 554], [14, p. 228], [15, p. 231]. See also → Antibiotics.

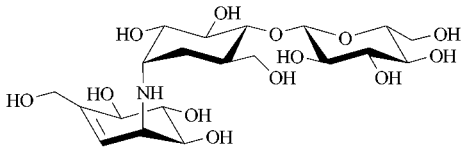

Validamycin is produced by fermentation of *Streptomyces hygroscopicus* var. *limoneus* [68]; for structural elucidation see [69].

Validamycin (announced in 1970) is a nonsystemic antibiotic with fungicidal properties against Basidiomycetes. It is used widely against *Rhizoctonia solani* of the anastamosis group I, which causes sheath blight of rice. In addition, it is used in some countries to control black scurf of potatoes, caused by *R. solani* anastamosis group IV. Validamycin has a short half-life in the soil (< 4 h) and is, therefore, environmentally favorable. It is sold in dust and liquid formulations and is more effective as a curative than as a protective treatment. The mode of action may be direct by preventing the pathogen from penetrating the host, or indirect by slowing pathogen growth so that host defense mechanisms can be mobilized [16, p. 118]. Resistance to validamycin has not appeared in practice.

Trade Names. Validacin (Takeda); Solacol (AgrEvo); Mycin (Sanonda).

8. Systemic Fungicides

8.1. Organophosphates

Iprobenfos [*26087-47-8*], *S*-benzyl-*O,O*-diisopropyl phosphorothioate, (thiophosphate), $C_{13}H_{21}O_3PS$, M_r 288.3, *bp* 126 °C at 5.3 Pa, is a yellow oil which is slightly soluble in water but soluble in most organic solvents [9, p. 44], [14, p. 232].

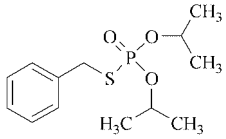

Iprobenfos is produced by reaction of the ammonium or alkali salt of diisopropyl thiophosphate and benzyl chloride [70].

Iprobenfos (announced in 1964) is a systemic fungicide with protective and eradicative properties. It is used against rice diseases and is applied to the foliage or into paddy water. Resistance in *Pyricularia oryzae* has been reported from experimental fields after intensive use over a number of years [33, p. 133].

Iprobenfos inhibits the biosynthesis of phosphatidylcholine, an important constituent of the cell membrane [33, p. 35].

Trade Names. Kitazin P, Kitazin (Kumiai).

Edifenphos [*17109-49-8*], *O*-ethyl-*S,S*-diphenyl phosphorodithioate, (dithiophosphate), $C_{14}H_{15}O_2PS_2$, M_r 310.4, *bp* 154 °C at 1.33 Pa, is a yellow to light-brown liquid which is practically insoluble in water but readily soluble in acetone, xylene, and methanol [9, p. 231], [13, p. 132], [14, p. 234].

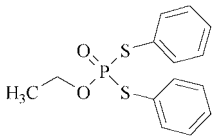

Edifenphos is produced by reaction of *O*-ethylphosphoryl dichloride with thiophenol in the presence of a base [71].

Edifenphos (announced in 1967) has preventative and eradicative properties. It is used as a foliar fungicide on rice for the control of blast, sheath blight, and stem rot. Its mode of action is to inhibit lipid biosynthesis [33, p. 35].

Trade Names. Hinosan, Hinorebcide (Bayer).

Pyrazophos [*13457-18-6*], ethyl 2-diethoxythiophosphoryloxy-5-methylpyrazolo[1,5-α]-pyrimidine-6-carboxylate, $C_{14}H_{20}N_3O_5PS$, M_r 373.4, *mp* 50–51 °C, vapor pressure 220 µPa at 50 °C, consists of colorless crystals which are practically insoluble in water but soluble in acetone, ethyl acetate, and toluene [9, p. 477], [14, p. 233].

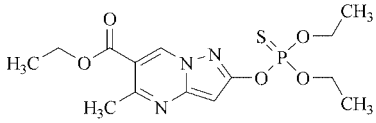

Pyrazophos is produced by condensation of cyanoacetic acid hydrazide with 2-formylacetoacetic acid to form ethyl 2-hydroxy-5-methylpyrazolo[1,5-α]pyrimidine-6-carboxytate followed by reaction with (*O,O*-diethyl) chlorothiophosphate [72].

Pyrazophos (announced in 1970) is a systemic fungicide with preventative and curative properties. Its spectrum of activity is confined to the powdery mildews, and it is used on cucurbits, apples, and other crops. Pyrazophos has side effects against some insect pests. Resistance in cucumber powdery mildew was detected in greenhouses after several years of use, but resistant isolates were less pathogenic [33, p. 133].

Trade Name. Curamil (AgrEvo).

8.2. Benzimidazoles

The systemic benzimidazole fungicides were originally developed as anthelmintics and include the derivatives benomyl, thiophanate methyl, thiabendazole, carbendazim, and fuberidazole. Benomyl is the most important, followed by thiophanate methyl; together they account for 75 % of all sales of benzimidazole fungicides. They have a broad spectrum of activity against plant-pathogenic fungi in the classes Ascomycetes, Deuteromycetes, and Basidiomycetes. However, they are particularly prone to development of resistance, which has led to a decline in their use for many crops. Nevertheless, they continue to be the most important group of systemic fungicides, in terms of world sales. Benzimidazoles act by binding to tubulin, a protein found in the microtubules [10, p. 299].

Benomyl [*17804-35-2*], methyl [1-(butylcarbamoyl)-1*H*-benzimidazol-2-yl]carbamate, $C_{14}H_{18}N_4O_3$, M_r 290.3, decomposes on heating without melting, is a colorless crystalline solid which is practically insoluble in water but soluble in heptane, chloroform and dimethylformamide [9, p. 37], [13, p. 117], [14, p. 175].

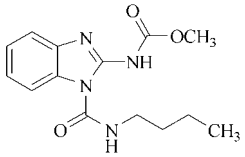

Benomyl is produced by reaction of carbendazim with butyl isocyanate [73].

Benomyl, introduced commercially in 1968, is used as a foliar fungicide with preventative and curative properties against a large range of diseases including *Botrytis* of grapes, soft fruit, and vegetables; scab and powdery mildew of deciduous fruit; and leaf spot diseases of field crops. It is also used as a post-harvest fungicide, a trunk injection for the control of Dutch elm disease, and for dipping planting material in pineapples and sugar cane. The activity in vivo can be ascribed to its hydrolysis product carbendazim described below.

Trade Name. Benlate (Du Pont).

Thiophanate methyl [*23564-05-8*], dimethyl 4,4′-(*o*-phenylene)bis(3-thioallophanate), $C_{12}H_{14}N_4O_4S_2$, M_r 342.4, *mp* 172 °C (decomp.), is a colorless crystalline solid which is sparingly soluble in water and only slightly soluble in acetone, chloroform, and methanol [9, p. 533], [14, p. 173].

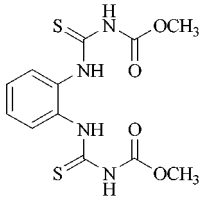

Thiophanate methyl is produced by reaction of *o*-phenylenediamine and methoxycarbonyl isothiocyanate [74].

Thiophanate methyl, introduced in 1969 by Nippon Soda, is used as a preventative and curative fungicide. It is applied as a foliar fungicide, seed treatment, postharvest treatment, and soil drench on a wide range of crops including deciduous fruit, stone fruit, vegetables, grapes, sugar beet, rice, and turf. The thiophanate molecule is degraded in the plant to carbendazim, the active moiety.

Trade Names. Topsin M, (Nippon Soda); Mildothane (Rhône-Poulenc); Cycosin (Cyanamid).

Carbendazim [*10605-21-7*], methyl (1*H*-benzimidazol-2-yl)carbamate, $C_9H_9N_3O_2$, M_r 191.2, *mp* 310 °C (decomp.), vapor pressure < 100 nPa at 20 °C, is a colorless powder which is sparingly soluble in all solvents [9, p. 89], [14, p. 175].

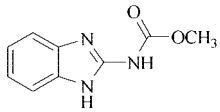

Carbendazim is produced by the reaction of *o*-phenylenediamine with methoxycarbonylcyanamide or with methoxycarbonyl isocyanide dichloride (formed from cyanogen chloride, phosgene, and methanol) in the presence of a base [75], [76].

Carbendazim was introduced in 1973 and has a spectrum of activity and uses similar to benomyl and thiophanate methyl.

Trade Names. Bavistan (BASF); Derosal, Troika (AgrEvo); Delsene (Du Pont).

Fuberidazole [*3878-19-1*], 2-(2-furyl)benzimidazole, $C_{11}H_8N_2O$, M_r 184.2, *mp* 292 °C (decomp.), is a fine crystalline powder which is slightly soluble in water, soluble in toluene, and fairly soluble in dichloromethane, light petroleum, 2-propanol, 0.1 M hydrochloric acid, and 0.1 M sodium hydroxide [22, p. 632].

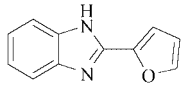

Fuberidazole is produced by reaction of *o*-phenylenediamine in copper acetate-containing methanol with furfural to give the copper salt of the desired compound, from which the free compound is liberated by treatment with hydrogen sulfide [77].

Fuberidazole (announced in1968) is used in seed treatment for control of *Fusarium* spp. in cereals, only in combination with other fungicides.

Trade Names. Baytan, Sibutol (Bayer).

Thiabendazole [*148-79-8*], 2-(1,3-thiazol-4-yl)benzimidazole, $C_{10}H_7N_3S$, M_r 201.2, *mp* 304 – 305 °C, is a colorless powder which is moderately soluble in water (pH dependent) and soluble in most organic solvents [22, p. 1183].

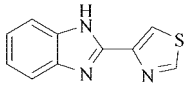

Thiabendazole is produced by heating thiazolyl-2-formamide with *o*-phenylenediamine in the presence of polyphosphoric acid [78].

Thiabendazole (announced in 1968) is used to control *Aspergillus*, *Botrytis*, *Cenospora*, *Cellotrichum*, *Corticum*, *Diaposthe*, *Dipodia*, and *Fusarium* in asparagus, avocados, bananas, barley, and beans. It is also used to control storage diseases of fruit and vegetables.

Trade Names. Mertect, Tecto, Storite (Merck, DowElanco); Decco (Elf Atochem); LSP (Gustafson).

8.3. Carboxanilides

The systemic antifungal properties of carboxin and oxycarboxin were discovered in the mid-1960s. These compounds and the related substance pyracarbolid act only against plant pathogens of the class Basidiomycetes. They interfere with respiration by inhibiting succinate oxidation [16, p. 155].

Carboxin [*5234-68-4*], 5,6-dihydro-2-methyl-1,4-oxathiine-3-carboxanilide, $C_{12}H_{13}NO_2S$, M_r 235.3, *mp* 91.5 – 92.5 °C, is a colorless solid which is practically insoluble in water but soluble in acetone, dimethyl sulfoxide, ethanol, and methanol [9, p. 95], [13, p. 104], [14, p. 199].

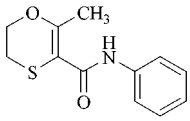

Carboxin is produced by reaction of 2-chloro-3-oxobutyricanilide with 2-mercaptoethanol, followed by acid-catalyzed ring closure with elimination of water [79].

Carboxin (announced in 1966) is used as a seed treatment for the control of smuts and bunts on cereals and for the control of *Rhizoctonia* spp. on cotton, vegetables, and groundnuts. Resistance in *Ustilago nuda* has been reported in samples isolated from barley varieties in France [80].

Trade Names. Vitavax (Uniroyal); Oxatin (Dichem); Kemikay (Kemira FC).

Flutolanil [*66332-96-5*], 3′-isopropoxy-α,α,α-trifluoro-*o*-toluanilide, $C_{17}H_{16}F_3NO_2$, M_r 323.3, *mp* 102 – 103 °C, vapor pressure 1.77 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in chloroform and methanol [9, p. 287].

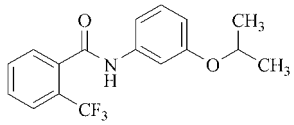

Flutolanil is produced by reaction of 2-(trifluoromethyl)benzoyl chloride with 3-isopropoxyaniline [81].

Flutolanil (announced in 1981) acts against Basidiomycetes. It is used as a foliar fungicide for the control of *Typhula* on barley and sheath blight on rice, as well as a seed dressing against *Rhizoctonia solani* on potatoes and cucurbits.

Trade Names. Moncut (Nihon Nohyaku); Prostar (AgrEvo, Nihon Nohyaku).

Mepronil [*55814-41-0*], 3′-isopropoxy-*o*-toluanilide, $C_{17}H_{19}NO_2$, M_r 269.3, *mp* 92 – 93 °C, consists of colorless crystals which are slightly soluble in water and soluble in most organic solvents [22, p. 787].

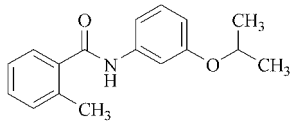

Mepronil is produced by reaction of 3-isopropoxyaniline with 2-methylbenzoyl chloride or with 2-methylbenzoic acid in the presence of phosphorous trichloride and triethylamine [82].

Mepronil (announced in 1979) inhibits succinic acid oxidation during metabolic respiration. It is used to control diseases caused by *Basidiomycetes* in rice, cereals, potatoes, and vegetables.

Trade Name. Basitac (Kumiai).

Thifluzamide [*130000-40-7*], 2′,6′-dibromo-2-methyl-4′-trifluoromethoxy-4-trifluoromethyl-1,3-thiazole-5-carboxanilide,

$C_{13}H_6Br_2F_6N_2O_2S$, M_r 528.1, *mp* 177.9 – 178.6 °C, is a colorless to light brown powder which is sparingly soluble in water [22, p. 1191].

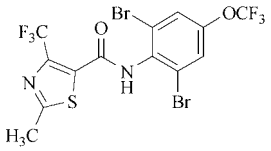

Thifluzamide is produced by reaction of 2-chloro-3-carbonyl-4-trifluoroethyl butylate with thioacetamide in DMF; the resulting thiazolecarboxylate is converted in two steps to thifluzamide [83].

Thifluzamide (announced in 1992) inhibits succinate dehydrogenase in the tricarboxylic acid cycle of fungi. It is used to control a wide range of *Basidiomycetes* fungi on rice, cereals, field crops, and turf by foliar application and as a seed treatment.

*Trade Name.*Not yet announced (Rohm & Haas).

8.4. Phenylamides

Discovered in the mid-1970s and previously known as acylalanines, the phenylamides are highly active systemic fungicides which act specifically against Peronosporales (Oomycetes). They include the compounds furalaxyl, metalaxyl, ofurace, oxadixyl, benalaxyl, and cyprofuram, which all have preventative, curative, and eradicative properties. They are prone to development of resistance, which has been observed in a number of target pathogens including those affecting cucumbers, grapevines, and potatoes.

Metalaxyl [*57837-19-1*], methyl ester of *N*-(2-methoxyacetyl)-*N*-(2,6-xylyl)-DL-alanine, $C_{15}H_{21}NO_4$, M_r 279.3, *mp* 71.8 – 72.3 °C, vapor pressure 293 µPa at 20 °C, consists of colorless crystals which are slightly soluble in water but readily soluble in benzene, dichloromethane, and methanol [9, p. 352], [15, p. 410].

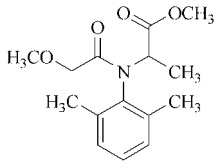

Metalaxyl is produced by reaction of 2,6-dimethylaniline with an excess of methyl 2-bromopropionate in the presence of sodium bicarbonate, followed by acetylation with methoxyacetyl chloride [84].

Metalaxyl (announced in 1977) is used as a seed treatment, foliar spray, soil application, or postharvest fungicide. It is used as a seed dressing to control *Pythium* damping-off and systemic downy mildews which occur on brassicas, sunflowers, peas, and tropical *Gramineae.* As a foliar fungicide, it is used to control downy mildew and late blight on a wide range of crops including vegetables, grapes, and potatoes. Applied to the soil, it is used to control rot of the root and stem base of many crops, caused by *Pythium* and *Phytophthora* spp.

For foliar use, metalaxyl is available only in prepacked mixtures with contact fungicides, to reduce the risk of resistance and broaden the spectrum of activity.

Trade Names. Ridomil, (Novartis); Milor (Rotam).

Oxadixyl [*77732-09-3*], 2-methoxy-*N*-(2-oxo-1,3-oxazolidin-3-yl)-*N*-(2,6-xylyl)-acetamide, $C_{14}H_{18}N_2O_4$, M_r 278.3, *mp* 104 – 105 °C, vapor pressure 3.3 mPa at 20 °C, consists of colorless crystals which are slightly soluble in water but readily soluble in acetone, dimethyl sulfoxide, and ethanol [29, p. 755].

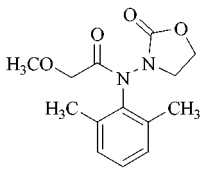

Oxadixyl is produced by intramolecular condensation of 1-chloroethyl-2-(methoxyacetyl)-2-(2,6-xylyl)-hydrazine carboxylate with sodium hydride [85].

Oxadixyl (announced in 1983) has a similar spectrum of activity to metalaxyl. It is used mostly as a foliar fungicide in mixtures with

contact fungicides and cymoxanil. It is also employed as a seed and soil treatment.
Trade Names. Sandofan, (Novartis); Anchor (Gustafson).

Ofurace [*58810-48-3*], α-[2-chloro-*N*-(2,6-xylyl)acetamido]-γ-butyrolactone, $C_{14}H_{16}ClNO_3$, M_r 281.74, *mp* 145 – 146 °C, vapor pressure 0.13 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in chloroform, acetone, ethyl acetate, and cyclohexanone [15, p. 413].

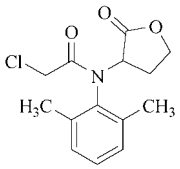

Ofurace is produced by refluxing 3-*N*-(2,6-dimethylphenyl)amino-γ-butyrolactone with chloroacetyl chloride in the presence of pyridine [86].

Ofurace shows a similar spectrum of activity to metalaxyl and is used as a foliar fungicide in mixtures with contact fungicides on crops such as potatoes, grapes, and vegetables.
Trade Names. Vamin, Patafol (AgrEvo).

Benalaxyl [*71626-11-4*], methyl *N*-phenylacetyl-*N*-(2,6-xylyl)-DL-alaninate, $C_{20}H_{23}NO_3$, M_r 325.4, *mp* 78 – 80 °C, is a colorless solid which is slightly soluble in water, sparingly soluble in hexane and saturated hydrocarbons, and soluble in most other organic solvents [22, p. 87].

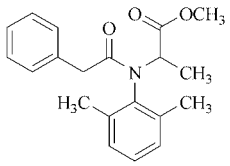

Benalaxyl is produced by reaction of the methyl ester of *N*-(2,6-xylyl)-DL-alanine with benzoyl chloride in the presence of a base [87].

Benalaxyl (announced in 1981) inhibits nucleic acid RNA polymerase. It is used to control *Oomycetes*, particularly fungi of the family *Peronosporaceae*, *Phytophthora*, *Plasmopara*, *Pseudoperonospora*, *Sclerospora*, *Bremia*, and *Phythium* spp. Some uses include control of late blights of potatoes and tomatoes and downy mildew of hops, vines, lettuce, onions, and other crops.
Trade Name. Galben (Isagro).

Furalaxyl [*57646-30-7*], methyl *N*-(2-furoyl)-*N*-(2,6-xylyl)-DL-alaninate, $C_{17}H_{19}NO_4$, M_r 301.3, *mp* 70 °C, consists of colorless crystals which are sparingly soluble in water, and soluble in many organic solvents [22, p. 634].

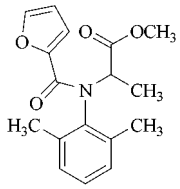

Furalaxyl is produced by reaction of *N*-methylpropionate-2,6-dimethylaniline in toluene and DMF with 2-furancarbonyl chloride [88].

Furalaxyl (announced in 1977) is used to control soil-borne diseases caused by *Phytophthora* and *Pythium* spp., and other *Oomycetes* on ornamentals.
Trade Name. Fongarid (Novartis).

Metalaxyl-M [*70630-17-0*], methyl *N*-(methoxyacetyl)-*N*-(2,6-xylyl)-D-alaninate, $C_{15}H_{21}NO_4$, M_r 279.3, *mp* -38.7 °C, is a pale yellow to light brown viscous liquid which is moderately soluble in water and miscible with many organic compounds [22, p. 794].

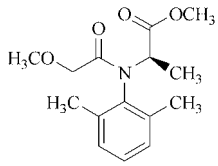

Matalaxyl-M is produced by reaction of 2,6-dimethylaniline with methyl 2-bromopropionate in the presence of sodium carbonate, followed by acetylation with methyloxyacetyl chloride and separation of the isomers [89].

Matalaxyl-M (announced in 1996) inhibits protein synthesis in fungi by interference with the synthesis of ribosomal RNA. It is used to

control diseases caused by air- and soil-borne *Peronosporales* on a wide range of temporate, subtropical, and tropical crops.
Trade Names. Apron XL, Ridomil Gold (Novartis).

8.5. Phosphites

The phosphites are a group of fungicides with unique systemic properties because they are translocated both acropetally and basipetally in the plant; important examples are fosetyl Al and its degradation product, phosphorous acid.

Fosetyl Al [*39148-24-8*], aluminum tris(ethyl phosphite), $C_6H_{18}AlO_9P_3$, M_r 354.1, decomposes above 200 °C, consists of colorless crystals which are soluble in water but practically insoluble in most organic solvents [9, p. 294], [15, p. 407].

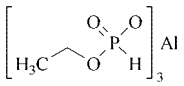

Fosetyl Al is produced by reaction of aluminum nitrate with an aqueous solution of sodium ethyl phosphite [90] or by reaction of ethyl phosphite and hydrargillite [$Al(OH)_3$] at 60 °C [91].

Fosetyl Al (announced in 1977) is used as a seed treatment, foliar fungicide, trunk injection, and soil treatment. Its spectrum of activity is largely confined to the Oomycetes, but it is not sufficiently active against all the important pathogens of this class, e.g., *Phytophthora infestans* of potato. As a seed treatment, it is used against downy mildew of peas. It is applied as a foliar spray to control airborne downy mildew of crops such as grapes and vegetables and to control root rot disease of pineapples, strawberries, and subtropical crops. It is also effective as a tree trunk injection in the control of soil-borne *Phytophthora* spp. of tree crops such as avocado because of its basipetal properties.
Trade Name. Aliette (Rhône-Poulenc).

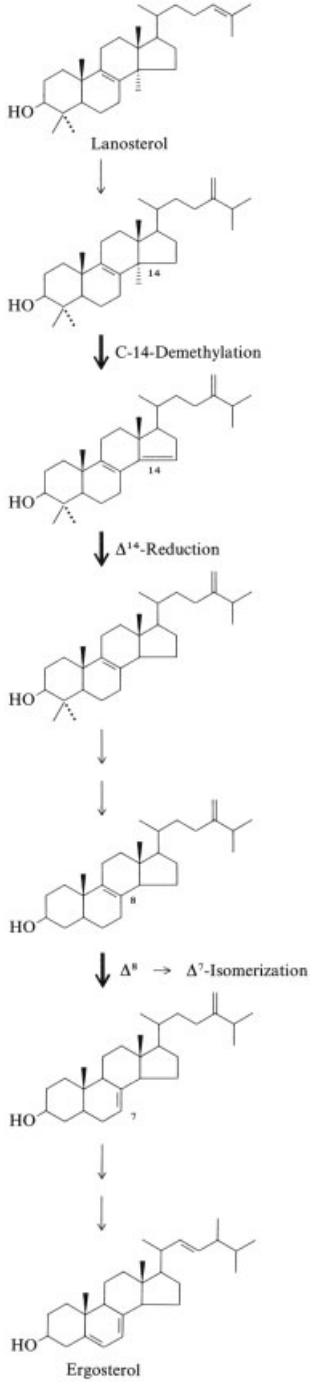

Figure 1. Biosynthesis of ergosterol from lanosterol showing the three reactions inhibited by sterol biosynthesis inhibitor (SBI) fungicides

8.6. Inhibitors of Sterol Biosynthesis

Ergosterol is a principal sterol of most fungi and is indispensable for membrane stabilization. Most species of Ascomycetes, Basidiomycetes, and Deuteromycetes synthesize ergosterol from acetic acid. Sterol biosynthesis inhibitors (SBI) are thus active against plant pathogens of these classes, but not of Oomycetes which depend on an exogenous supply of suitable sterols. They block the multistep conversion of lanosterol to ergosterol mainly at the three positions shown in Figure 1. Many compounds such as substituted triazoles, imidazoles, pyridines, pyrimidines, and piperazines inhibit the cytochrome P450 assisted C-14-demethylation step. Morpholine and piperidine derivatives, however, inhibit the $\Delta^8 \rightarrow \Delta^7$-isomerization and can also block the Δ^{14}-reduction [92].

8.6.1. Triazoles

Triazole fungicides are used for foliar and seed treatment; they have protective and curative properties. They can be used alone or in mixtures with other fungicides.

Triadimefon [*43121-43-3*], 1-(4-chlorophenoxy)-3,3-dimethyl-1-(1*H*-1,2,4-triazol-1-yl)-2-butanone, $C_{14}H_{16}ClN_3O_2$, M_r 293.8, *mp* 82.3 °C, vapor pressure < 100 µPa at 20 °C, is a colorless solid which is practically insoluble in water but soluble in cyclohexanone, dichloromethane, 2-propanol, and toluene [9, p. 539], [14, p. 208], [92, p. 41].

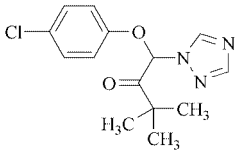

Triadimefon can be produced by dichlorination of pinacolone followed by substitution with sodium *p*-chlorophenoxide and triazole in a one-pot reaction [93].

Triadimefon (announced in 1973) is used as a foliar spray for the control of powdery mildew and rust diseases of cereals, coffee, grapes, stone fruit, vegetables, and ornamentals. In the plant, triadimefon is converted into triadimenol.
Trade Name. Bayleton (Bayer).

Triadimenol [*55219-65-3*], 1-(4-chlorophenoxy)-3,3-dimethyl-1-(1*H*-1,2,4-triazol-1-yl)butan-2-ol, $C_{14}H_{18}ClN_3O_2$, M_r 295.8, *mp* 121 – 127 °C, vapor pressure < 1 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in dichloromethane, cyclohexane, and 2-propanol [9, p. 540], [92, p. 47].

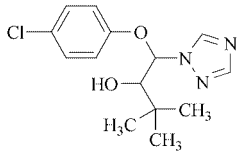

Triadimenol is produced by reduction of triadimefon with aluminum isopropoxide [94].

Triadimenol (announced in 1977) is used as a seed treatment and foliar fungicide. As a seed dressing, it is used on cereals primarily for the control of powdery mildew and rust diseases but also for *Rhynchosporium secalis* on barley. It is used as a foliar fungicide to control powdery mildew in grapes, vegetables, hops, sugar beet, and tobacco, as well as banana sigatoka and coffee rust.
Trade Names. Baytan, Bayfidan, Silvacur (Bayer).

Bitertanol [*55179-31-2*], *rac*-1-(biphenyl-4-yloxy)-3,3-dimethyl-1-(1*H*-1,2,4-triazol-1-yl)-butan-2-ol, $C_{20}H_{23}N_3O_2$, M_r 337.4, *mp* 123 – 129 °C, vapor pressure 1 µPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in dichloromethane and partially soluble in toluene and 2-propanol [9, p. 54], [92, p. 47].

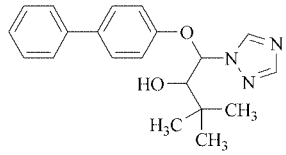

The production of bitertanol is similar to that of triadimenol [94].

Bitertanol (announced in 1977) is used as a foliar fungicide for the control of leaf spot diseases of groundnuts as well as apple scab and *Sclerotinia* spp. on stone fruit. As a seed dressing on wheat, it is used to control dwarf

bunt (*Tilletia controversa*), *Fusarium nivale,* and *Septoria nodorum*.
Trade Names. Baycor, Sibutol (Bayer).

Propiconazole [*60207-90-1*], (±)-1-[2-(2,4-dichlorophenyl)-4-propyl-1,3-dioxolan-2-ylmethyl]-1*H*-1,2,4-triazole, $C_{15}H_{17}Cl_2N_3O_2$, M_r 342.2, n_D^{20} 1.5468, *bp* 180 °C at 13.3 Pa, vapor pressure 133 µPa at 20 °C, is a yellowish viscous liquid which is practically insoluble in water but readily soluble in acetone, methanol, and 2-propanol [9, p. 468], [92, p. 44].

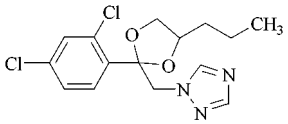

Propiconazole can be produced by ketalization of 2-bromo-2′,4′-dichloroacetophenone with 1,2-pentanediol, followed by nucleophilic substitution with sodium 1,2,4-triazolide in dimethylformamide at 160 °C [95].

Propiconazole (announced in 1979) is a broad-spectrum foliar fungicide used to control a range of cereal diseases, including powdery mildew, rust, *Septoria* spp., *Rhynchosporium,* and net blotch. It is also used against banana sigatoka, coffee rust, and various diseases of turf, rice, peanuts, pecans, maize, stone fruit, and other crops.
Trade Names. Tilt (Novartis); Radar (Zeneca).

Penconazole [*66246-88-6*], 2-(2,4-dichlorophenyl)-1-(1*H*-1,2,4-triazol-1-yl)pentane, $C_{13}H_{15}Cl_2N_3$, M_r 284.2, *mp* 60 °C, vapor pressure 0.21 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in most organic solvents [92, p. 40].

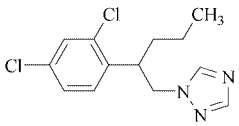

Penconazole can be produced by alkylation of 2,4-dichlorophenylacetonitrile with bromopropane, followed by conversion of the nitrile to the acetate. Reduction with lithium aluminum hydride gives the corresponding alcohol, which is first reacted with methanesulfonyl chloride and then with sodium triazolide [96]. An alternative synthesis is the reaction of 2′,4′-dichlorobutyrophenone with triazolylmethyl phosphonate, followed by hydrogenation [97].

Penconazole (announced in 1983) is used as a foliar fungicide to control powdery mildew of pome fruits, grapes, stone fruit, tobacco, vegetables, and ornamentals; scab of apples and pears; and black rot of grapes.
Trade Name. Topas (Novartis).

Flutriafol [*76674-21-0*], (*RS*)-2,4′-difluoro-α-(1*H*-1,2,4-triazol-1-ylmethyl)benzhydryl alcohol, $C_{16}H_{13}F_2N_3O$, M_r 301.3, *mp* 130 °C, vapor pressure 400 nPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in acetone, dichloromethane, methanol and xylene [92, p. 50].

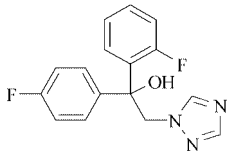

Flutriafol is produced by reaction of 2,4′-difluorobenzophenone and trimethyloxosulfonium iodide; the resulting epoxide is then opened by triazole in the presence of a base [98].

Flutriafol (announced in 1983) is used as a foliar fungicide to control diseases such as powdery mildew, rust, *Rhynchosporium secalis, Helminthosporium,* and *Septoria* spp. It is also employed as a seed treatment on cereals.
Trade Names. Impact, Vincit (Zeneca).

Flusilazole [*85509-19-9*], bis(4-fluorophenyl)(methyl)(1*H*-1,2,4-triazol-1-ylmethyl) silane, $C_{16}H_{15}F_2N_3Si$, M_r 315.38, *mp* 55 °C, is a white crystalline solid which is practically insoluble in water but soluble in many organic solvents [22, p. 603].

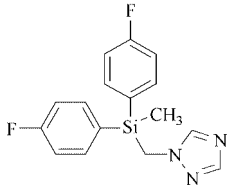

Flusilazole can be produced by addition of butyllithium to a solution of chloromethyldi-

chlorosilane and 1-bromo-4-fluorobenzene and subsequent reaction with triazole [99].

Flusilazole (announced in 1984) is a broad-spectrum foliar fungicide which is used for the control of powdery mildew, rust, and leaf spot diseases of cereals; black rot and powdery mildew of grapes; and powdery mildew and scab of apples.

Trade Names. Punch, Nustar, Olymp, Sanction, Captan (Du Pont).

Hexaconazole [*79983-71-4*], (*RS*)-2-(2,4-dichlorophenyl)-1-(1*H*-1,2,4-triazol-1-yl)hexan-2-ol, $C_{14}H_{17}Cl_2N_3O$, M_r 314.2, *mp* 111 °C, vapor pressure 0.02 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in many organic solvents [29, p. 562].

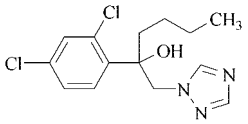

Hexaconazole is produced by reaction of 1-(2,4-dichlorophenyl)pentan-2-one and trimethyloxosulfonium iodide, followed by opening of the resulting epoxide with triazole in the presence of a base [100].

Hexaconazole (announced in 1986) is being developed as a broad-spectrum foliar fungicide. It is used for the control of powdery mildew and black rot of grapes, powdery mildew and scab of apples, coffee rust, and leaf spot diseases of groundnuts [100, p. 19]

*Trade Names.*Contaf (Rollis); Anvil, Planeta (Zeneca).

Diniconazole [*76714-88-0*], (*E*)-1-(2,4-dichlorophenyl)-4,4-dimethyl-2-(1*H*-1,2,4-triazol-1-yl)pent-1-en-3-ol, $C_{15}H_{17}Cl_2N_3O$, M_r 326.2, *mp* 134 – 156 °C, is a white crystalline solid which is slightly soluble in water and partially soluble in acetone or methanol [92, p. 48].

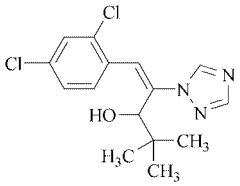

Diniconazole is produced by condensation of 2,4-dichlorobenzaldehyde and triazolylpinacolone, followed by photochemical isomerization of the *E* – *Z* mixture to the *E* isomer and reduction with sodium borohydride [102].

Diniconazole (announced in 1986) is being developed as a foliar spray and a seed treatment. As a foliar spray, it controls leaf spot, rust, and white mold of groundnuts; powdery mildew of apples, grapes, and small grain cereals; and rust diseases of apples and cereals. As a seed treatment, it controls bunts and smuts of cereals.

Trade Names. Spotless, Sumi-8 (Sumitomo).

Myclobutanil [*88671-89-0*], 2-(4-chlorophenyl)-2-(1*H*-1,2,4-triazol-1-ylmethyl)hexanenitrile, $C_{15}H_{17}ClN_4$, M_r 288.8, *mp* 63 – 68 °C, *bp* 202 – 208 °C at 133 Pa, vapor pressure 213 mPa, is a pale-yellow solid which is practically insoluble in water but soluble in ethanol, benzene, ethyl acetate, and acetone [22, p. 854].

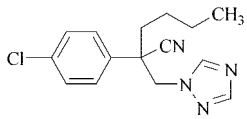

Myclobutanil is produced by alkylation of 4-chlorophenylacetonitrile with butyl chloride, followed by treatment with dibromomethane and sodium hydroxide in dimethyl sulfoxide. The resulting intermediate is then reacted with the potassium salt of 1,2,4-triazole in dimethylformamide [103].

Myclobutanil (announced in 1986) acts against powdery mildew and scab of apples, as well as black rot and powdery mildew of grapevines. It is also being tested as a cereal seed dressing [101, p. 55].

Trade Name. Systhane (Rohm & Haas).

Tebuconazole [*80443-41-0*], (*RS*)-1-(4-chlorophenyl)-4,4-dimethyl-3-(1*H*-1,2,4-triazol-1-ylmethyl)pentan-3-ol, $C_{16}H_{22}ClN_3O$, M_r 307.81, *mp* 104.7 °C, consists of colorless crystals which are practically insoluble in water but soluble in many organic solvents [22, p. 1144].

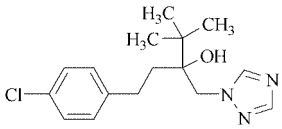

Tebuconazole is produced by opening the 2-*tert*-butyl-2-(4-chlorophenethyl) oxirane with 1,2,4-triazole in the presence of a base [104].

Tebuconazole (announced in 1986) is being developed as a foliar spray and seed treatment against a range of fungal pathogens. It is an effective seed treatment against bunts, smuts, and *Septoria* spp. on cereals. As a foliar spray, it acts against rust, powdery mildew, and leaf and head diseases and is also being tested on a range of field crops [101, p. 41].

Trade Names. Folicur, Raxil, Elite, Horizon, Lynx, Matador (Bayer).

Cyproconazole [*94361-06-5*], (2*RS*,3*RS*)-2-(4-chlorophenyl)-3-cyclopropyl-1-(1*H*-1,2,4-triazol-1-yl)butan-2-ol, $C_{15}H_{18}ClN_3O$, M_r 291.8, *mp* 103 – 105 °C, vapor pressure 34.7 µPa at 20 °C, consists of colorless crystals which are practically insoluble in water but readily soluble in acetone, ethanol, xylene, and dimethyl sulfoxide [101, p. 33].

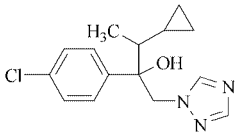

Cyproconazole is produced by reaction of 1-(4-chlorophenyl)-2-cyclopropylethanone with methyl iodide in the presence of sodium hydride followed by treatment with dodecyldimethylsulfonium methyl sulfate to yield the epoxide intermediate which is then opened by triazole in the presence of a base [105].

Cyproconazole (announced in 1986) is being developed as a foliar spray against rust diseases of cereals and coffee; powdery mildews of cereals, fruit trees and grapes; leaf spot diseases of peanuts and sugar beet; apple scab; and white mold.

Cyproconazole is a broad spectrum foliar fungicide with activity against diseases caused by powdery mildews, rusts, *Monilinia, Cercospora*, *Venturia, Rhizoctonia* and *Sclerotium*. It shows long-lasting protective as well as curative and eradicative activity. It penetrates into the plant very rapidly and is translocated acropetally. Crops on which it is used or under development include cereals, fruit trees, grapevines, peanuts, sugar beet, bananas, and coffee.

Trade Names. Alto (Novartis); Shandon (Barclay).

Difenconazole [*119446-68-3*], 3-chloro-4-[4-methyl-2-(1*H*-1,2,4-triazol-1-ylmethyl)-1,3-dioxolan-2-yl]phenyl-4-chlorophenyl ether, CGA-169374, $C_{19}H_{17}Cl_2N_3O_3$, M_r 406.3, *mp* 76 °C, vapor pressure 0.12 µPa at 20 °C, is a white crystalline solid that is practically insoluble in water but very soluble in most organic solvents [29, p. 328].

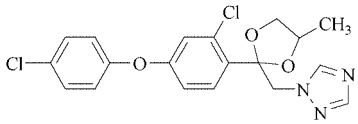

Difenconazole is produced by ketalization of 2-bromo-2′-chloro-4′-(4-chlorophenyl)acetophenone with 1,2-propanediol, followed by nucleophilic substitution with sodium 1,2,4-triazolide in dimethylformamide under reflux [106].

Difenconazole (announced in 1988) is a broad-spectrum foliar and seed treatment fungicide. It shows translaminar systemicity and provides long-lasting protective and curative activity against Ascomycetes, Deuteromycetes, and Basidiomycetes.

Difenconazole is particularly effective against pathogens such as *Alternaria,* and pathogen complexes against which other SBIs are relatively weak. The spectrum includes species of *Septoria, Cercospora, Cercosporidium, Ascochyta, Alternaria, Ramularia, Venturia, Guignardia, Phoma, Collectotrichum,* rust fungi, and powdery mildews).

The crops on which it is being developed include wheat, sugar beet, peanuts, potatoes, pome fruit, grapes, and vegetables.

Trade Names. Score, Dividend (Novartis).

Azaconazole [*60207-31-0*], 1-[{2-(2,4-dichlorophenyl)-1,3-dioxolan-2-yl}methyl]-*H*-1,2,4-triazole, $C_{12}H_{11}Cl_2N_3O_2$, M_r 300.1, *mp* 112.6 °C, is a beige to brown colored powder

which is moderately soluble in water and hexane and readily soluble in acetone, methanol, and toluene [22, p. 58].

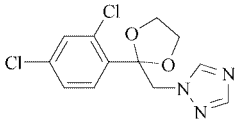

Azoconazole is produced by reaction of 1*H*-1,2,4-triazole with bromomethyl-2,4-dichlorophenyl ethylene ketal [107].

Azoconazole (announced in 1983) is used against wood-destroying and sapstain fungi. It is also used as a disinfectant in mushroom cultivation and on storage boxes for fruit and vegetables.

Trade Names. Rodewod, Safetray SL (Jansson).

Bromuconazole [*116255-48-2*], 1-[(2*RS*, 4*RS*; 2*RS*, 4*SR*)-4-bromo-2-(2,4-dichlorophenyl)tetrahydrofurfursaveyl]-1*H*-1,2,4-triazole, $C_{13}H_{12}BrCl_2N_3O$, M_r 377.1, *mp* 84 °C, is a colorless powder which is sparingly soluble in water and moderately soluble in organic solvents [22, p. 152].

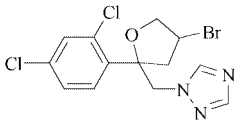

Bromuconazole is produced by reaction of 1-bromomethoxyacetylene in toluene with ethylmagnesium bromide in THF, followed by the addition of 2,4-dichlorobenzoylmethyl chloride to give 2,4-dichlorophenyl-2′-chloromethyl-2′-hydroxy-2′-bromomethoxyacetylene. The product is then hydrogenated and treated with 4-methylphenylsulfonic acid in toluene to give 2-chloromethyl-2-(2,4-dichlorophenyl)-2,5-dihydrofuran, which is treated with 1,2,4-triazole in *N*-methylpyrrolidone in the presence of potassium carbonate to yield the desired product [108].

Bromuconazole (announced in 1990) is used against *Ascomycetes*, *Basidiomycetes* and *Fungi imperfecti* including *Alternaria*, *Fusarium*, black and yellow sigatoke, and *Pseudocercosporalia* spp. in cereals, fruits, vines, vegetables, tropical crops, turf, and ornamentals.

Trade Names. Granit, Vectra (Rhône-Poulenc).

Epoxiconazole [*106325-08-0*], (2*RS*, 3 *SR*)-1- [3-(2-chlorophenyl)-2, 3-epoxy-2- (4- fluorophenyl)propyl]-*H*-1,2,4-triazole, $C_{12}H_{17}ClFN_3O$, M_r 329.8, *mp*136.2 °C, consists of colorless crystals which are sparingly soluble in water, moderately soluble in *n*-heptane, and soluble in acetone, and dichloromethane [22, p. 79].

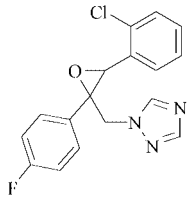

Epoxiconazole is produced by reaction of 2,4-dichlorobenzylphosphonium triphenylchloride with 4-chlorophenyl acetate and potassium *tert*-butoxide in refluxing methanol, and bromination of the resulting product, followed by epoxidation with 3-chloroperbenzoic acid [109].

Epoxiconazole (announced in 1990) is used as a broad-spectrum fungicide for control of diseases caused by *Ascomycetes, Basidiomycetes*, and *Deuteromycetes* in cereals, sugarbeet, peanuts, oilseed rape, and ornamentals.

Trade Name. Opus (BASF).

Fenbuconazole [*114369-43-6*], 4-(4-chlorophenyl)-2-phenyl-2-(1*H*-1,2,4-triazol-1-ylmethyl)butyronitrile, $C_{19}H_{17}ClN_4$, M_r 336.8, *mp* 124-126 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in common organic solvents [22, p. 508].

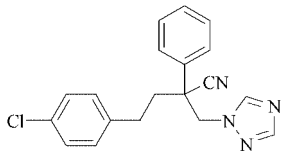

Fenbuconazole is produced by reaction of benzylnitrile with potassium hydride in toluene, followed by the addition of *p*-chlorophenethyl chloride in diemthylformamide to give α-[2-(*p*-chlorophenyl)ethyl]phenylacetonitrile. This is treated with bromomethane in dimethylsulfoxide in the presence of sodium hydrox-

ide to yield 1-bromo-2-cyano-2-phenyl-4-(*p*-chlorophenyl)butane, which upon heating with triazole, in KOH-containing dimethyl sulfoxide, yields fenbuconazole [110].

Fenbuconazole (announced in 1988) is used to control *Septoria* and *Puccinia* rusts, bunt, and smut and *Rhynchosporium secalis* on cereals; and powdery mildew and scab on many crops.
Trade Names. Inder (Rohm & Haas); Troika (AgrEvo).

Fluquinconazole [*136426-54-5*], 3-(2,4-dichlorophenyl)-6-fluoro-2-(1*H*-1,2,4-triazol-1-yl)quinazolin-4(3*H*)-one, $C_{16}H_8Cl_2FN_5O$, M_r 376.3, *mp* 184 – 192 °C, is a pale crystalline solid which is practically insoluble in water, moderately soluble in ethanol, and soluble in acetone, xylene, and dimethyl sulfoxide [22, p. 588].

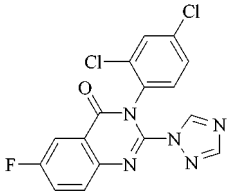

Fluquinconazole is produced by reaction of 2-chloro-6-fluoro-*N*-3-*p*-chlorophenylquinazoline-4(3*H*)-one with 1,2,4-triazole in an inert solvent in the presence of potassium carbonate [111].

Fluquinconazole (announced in 1992) is used to control a wide range of *Ascomycetes*, *Deuteromycetee*, and *Basidiomycetes* by foliar application to broad-leaved and cereal crops.
Trade Names. Castellan, Vista (AgrEvo).

Metconazole [*125116-23-6*], (1*RS*,5*RS*; 1*RS*,5*SR*)-5-(4-chlorobenzyl)-2,2-dimethyl-1-(1*H*-1,2,4-trizol-1-ylmethyl)cyclopentanol, $C_{17}H_{22}ClN_3O$, M_r 319.8, *mp* 110-113 °C, is a colorless crystalline solid which is sparingly soluble in water but readily soluble in acetone, and methanol [22, p. 803].

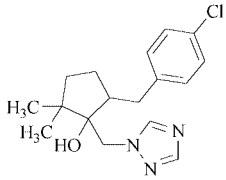

Metconazole is produced by reaction of 3-dimethyl-2-oxocyclopentanecarboxylate in benzene with sodium hydride, followed by addition of 4-chlorobenzoyl chloride, and the thus obtained product decarboxylated by heating with hydrogen bromide [112].

Metconazole (announced in 1992) is applied post-emergence to control a wide range of foliar diseases on cereals and other crops.
Trade Name. Carambo (Cyanamid).

8.6.2. Imidazoles

Imidazole fungicides with protective and curative properties are used for foliar, postharvest, and seed treatment. They include the compounds prochloraz, imazalil, and triflumizole.

Prochloraz [*67747-09-5*], *N*-propyl-*N*-[-2-(2,4,6-trichlorophenoxy)ethyl]imidazole-1-carboxamide, $C_{15}H_{16}Cl_3N_3O_2$, M_r 376.7, *mp* 38.5 – 41 °C, vapor pressure 80 nPa at 20 °C, is a colorless crystalline solid which is practically insoluble in water but readily soluble in acetone, chloroform, diethyl ether, and toluene [9, p. 452], [92, p. 53].

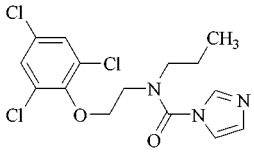

Prochloraz is produced by reaction of 2-(2,4,6-trichlorophenoxy)ethyl bromide with propylamine, followed by phosgenation and formation of the substituted urea with imidazole [113].

Prochloraz (announced in 1977) is used as a foliar fungicide against a range of diseases occurring on cereals, field crops, fruit, vegetables, and turf. On cereals it is active against eye spot, net blotch, powdery mildew, and *Rhynchospo-*

rium and *Septoria* spp. Prochloraz is also employed as a postharvest treatment against storage diseases of tropical and subtropical fruit.
Trade Names. Sportak (AgrEvo); Eyetak (Barclay).

Imazalil [*35554-44-0*], allyl 1-(2,4-dichlorophenyl)-2-(imidazol-1-yl)ethyl ether, $C_{14}H_{14}Cl_2N_2O$, M_r 297.2, d^{23} 1.2429; n_D^{20} 1.5643, vapor pressure 9.3 µPa at 20 °C, is a brownish oil which is slightly soluble in water and freely soluble in organic solvents [9, p. 315], [14, p. 210], [92, p. 50].

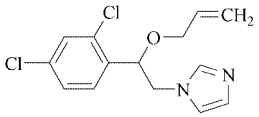

Imazalil is produced by bromination of 2′,4′-dichloroacetophenone, followed by reduction with sodium borohydride to form an epoxide intermediate. The epoxide is opened with imidazole, and the resulting alcohol is then etherified [114].

Imazalil (announced in 1973) is active as a seed treatment against *Fusarium, Helminthosporium,* and *Septoria* spp. on cereals and as a postharvest fungicide against storage rot of fruit.
Trade Names. Fungaflor, Fungazil (Jansson); Deccozil (Elf Atochem); Nuzone (Wiebur-Ellis); Freshguard (FMC).

Pefurazoate [*101903-30-4*], pent-4-enyl *N*-furfuryl-*N*-imidazol-1-ylcarbonyl-DL-homoalaninate, $C_{18}H_{23}N_3O_4$, M_r 345.4, *bp* 235 °C (decomp.), is a pale brownish liquid which is slightly soluble in water but readily soluble in most organic solvents [22, p. 932].

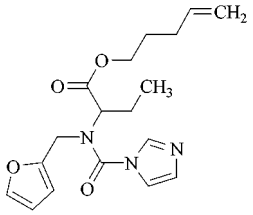

Pefurazoate is produced by reaction of [*N*-(pent-4-enyloxy)carbonyl] butylfurfurylcarbamoyl chloride with imidazole in dimethyl sulfoxide containing triethylamine [115].

Pefurazote (announced in 1990) is used to control seed-borne fungal diseases, particularly *Fusarium moniliforme*, *Pyricutaria oryzae*, and *Cochliobulus miyaheanna*. It is also used for control of rice seedling blight caused by soil-borne pathogens.
Trade Name. Healthied (Hokko, Ube).

Triflumizole [*68694-11-1*], (*E*)-4-chloro-α,α,α-trifluoro-*N*-(1-imidazol-1-yl-2-propoxyethylidene)-*o*-toluidine, $C_{15}H_{15}ClF_3N_3O$, M_r 345.7, *mp* 63 °C, consists of colorless crystals which are sparingly soluble in water and hexane but highly soluble in chloroform, acetone, and acetonitrile [22, p. 1245].

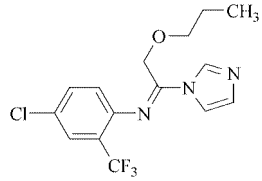

Triflumizole is produced by reaction of 2-trifluoromethyl-4-chloropropoxyacetanilide with imidazole in the presence of phosphorous pentachloride [116].

Triflumizole (announced in 1982) is used to control *Gymnosporangium* and *Venturia* spp. in pome fruits, against powdery *Erysiphaceae* in fruit and vegetables, and against *Fusarium*, *Fulvia* and *Monilinia* spp. It is also used as a seed treatment against *Helminthosporium*, *Tilletia* and *Ustilago* spp. in cereals.
Trade Name. Trifmine (Nippon Soda).

8.6.3. Pyridines

Pyrifenox [*88283-41-4*], 2′,4′-dichloro-2-(3-pyridyl)acetophenone *O*-methyloxime, $C_{14}H_{12}Cl_2N_2O$, M_r 295.16, *bp* > 150 °C at 13.3 Pa, vapor pressure 1.9 mPa, is a tan liquid which is practically insoluble in water but soluble in organic solvents [29, p. 884].

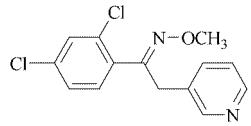

Pyrifenox is produced by condensation of an alkyl ester of 2,4-dichlorobenzoic acid with ethyl 3-pyridylacetate, followed by reaction with *O*-methylhydroxylamine [118].

Pyrifenox (announced in 1986) has protective and curative properties. It is active as a foliar fungicide against powdery mildew of apples and grapes, apple scab, and groundnut leaf spot diseases.

Trade Names. Rondo, Furado (Novartis); Padigrol (Novartis, Nippon Kagoku).

8.6.4. Pyrimidines

Pyrimidine fungicides include the compounds fenarimol and nuarimol; they all show protective and curative properties, and are active primarily against powdery mildew fungi.

Fenarimol [*60168-88-9*], (±)-2,4′-dichloro-α-(5-pyrimidinyl)benzhydryl alcohol, $C_{17}H_{12}Cl_2N_2O$, M_r 331.2, *mp* 117 – 119 °C, vapor pressure 13 µPa at 25 °C, consists of colorless crystals which are practically insoluble in water but soluble in acetone and methanol [9, p. 257], [14, p. 214], [92, p. 29].

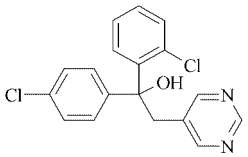

Fenarimol is produced by reaction of 2,4′-dichlorobenzophenone with 5-lithiopyrimidine at – 125 °C [117].

Fenarimol (announced in 1975) is used as a foliar spray for the control of powdery mildew of apples, cucurbits, grapes, and ornamentals, as well as apple scab.

Trade Name. Rubigan (Dow Elanco).

Cyprodinil [*121552-61-2*], 4-cyclopropyl-6-methyl-*N*-phenylpyrimidin-2-amine, $C_{14}H_{15}N_3$, M_r 225.3, *mp* 75.9 °C, is a beige powder which is fairly soluble in water and soluble in ethanol, acetone, toluene, *n*-hexane, and *n*-octanol [22, p. 319].

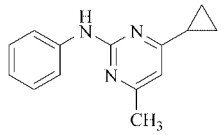

Cyprodinil is produced by cyclocondensation of phenylguanidinium carbonate with 1-cyclopropyl-1,3-butenedione [119].

Cyprodinil (announced in 1994) inhibits methionine biosynthesis. It is used as a foliar fungicide in cereals, grapes, pome fruits, stone fruits, strawberries, and vegetables and as a seed dressing on barley; it controls a wide range of pathogens, including *Pseudocercosporella herpotricloides*, *Erysiphe* spp., *Pyrenophora teres, Rhynchosporium* secalis*, Septoria nodorum.*

Trade Names. Unix, Chorus, Stereo, Switch (Novartis).

Dimethirimol [*5221-53-4*], 5-*n*-butyl-2-dimethylamino-4-hydroxy-6-methylpyrimidine, $C_{11}H_{19}N_3O$, M_r 109.3, *mp* 102 °C, consists of colorless needles which are moderately soluble in water and soluble in acetone, chloroform, ethanol, and xylene [22, p. 412].

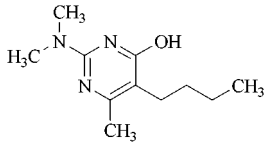

Dimethirimol is produced by reaction of 5-*n*-butyl-4-hydroxy-6-methyl-2-methylthiopyrimidine with dimethylamine acetate [120].

Dimethirimol (announced in 1968) is used in soil application for control of powdery mildews in cucurbits, tobacco, capsicums, tomatoes, and some ornamentals.

Trade Name. Milcurb (Zeneca).

Nuarimol [*63284-71-9*], (±)-2-chloro-4′-fluoro-α-(pyrimidin-5-yl)benzhydryl alcohol, $C_{17}H_{12}ClFN_2O$, M_r 314.7, *mp* 126 – 127 °C, is a colorless crystalline solid which is sparingly soluble in water and readily soluble in acetone, methanol, and xylene [22, p. 889].

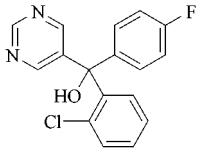

Nuarimol is produced by reaction of 5-bromopyrimidine with 4-fluorobenzoyl-2-chlorobenzene in the presence of butyllithium [121].

Nuaromil (announced in 1980) is used to control a wide range of pathogenic fungi, e.g., *Cerosporella* spp., *Septoria* spp., *Ustilago* spp., powdery mildew, leaf spot in cereals, both as foliar spray and as a seed treatment; powdery mildews on pome fruits, stone fruits, vines, and hops.
Trade Name. Trimidal (DowElanco).

Pyrimethanil [*53112-28-0*], *N*-(4,6-dimethylpyrimidin-2-yl)aniline, $C_{12}H_{13}N_3$, M_r 199.3, *mp* 96.3 °C, is a colorless crystalline solid which is sparingly soluble in water but soluble in most organic solvents [22, p. 1068].

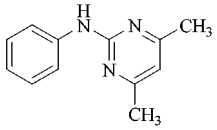

Pyrimethanil is produced by reaction of 2-bromo-4,6-dimethylpyrimidine with aniline [122].

Pyrimethanil (announced in 1992) is used to control gray mold (*Botrytis cinerea*) on vines, fruit, vegetables, and ornamentals.
Trade Names. Scala, Mythos (AgrEvo).

8.6.5. Piperazines

Triforine [*26644-46-2*], 1,4-bis(2,2,2-trichloro-1-formamidoethyl)piperazine, $C_{10}H_{14}Cl_6N_4O_2$, M_r 435.0, *mp* 155 °C, vapor pressure 27 µPa at 25 °C, consists of colorless crystals which are practically insoluble in water but soluble in dimethylformamide or 1-methyl-pyrrolid-2-one [9, p. 553], [14, p. 213], [92, p. 26].

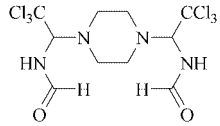

Triforine is produced by reaction of piperazine with (1,2,2,2-tetrachloroethyl)formamide in the presence of triethylamine [123].

Triforine (announced in 1969) is used as a foliar and postharvest fungicide. It is active against powdery mildew, rust, and leaf spot diseases of fruit, ornamentals, vegetables, and cereals.
Trade Name. Saprol (Cyanamid).

8.6.6. Morpholines

The morpholine fungicides include tridemorph, fenpropimorph, and fenpropidin. They are used as foliar fungicides with protective and curative properties for the control of powdery mildew and rust on cereals. Although the first representatives of this group were launched in the late 1960s, interest in them has been renewed because of the development of resistance to the C-14-demethylation inhibitors in powdery mildews. Morpholine fungicides are used alone and in mixtures with contact and systemic fungicides.

Tridemorph [*24602-86-6*], 2,6-dimethyl-4-tridecylmorpholine, $C_{19}H_{39}NO$, M_r 297.5, *bp* 134 °C at 66 Pa, vapor pressure 40 mPa at 20 °C, is a colorless oil which is practically insoluble in water but miscible with acetone, benzene, chloroform, and ethanol [9, p. 550], [13, p. 123], [14, p. 219], [92, p. 55].

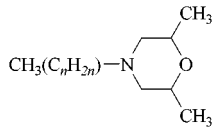

Tridemorph is produced as the main product (60 – 70 %) in the reaction of tridecylamine and propylene oxide, followed by dehydration and cyclization in the presence of concentrated sul-

furic acid [124]. Other homologues ($n = 10$, 11, or 13) are also formed.

Tridemorph (announced in 1969) is used as a foliar fungicide for the control of powdery mildew of cereals.
Trade Name. Calixin (BASF).

Fenpropimorph [*67564-91-4*], (±)-*cis*-4-[3-(4-*tert*-butylphenyl)-2-methylpropyl]-2,6-dimethylmorpholine, $C_{20}H_{33}NO$, M_r 303.5, *bp* 120 °C at 6.6 Pa, n_D^{20} 1.4940, is a colorless oil which is practically insoluble in water but miscible with acetone, chloroform, and ethanol [9, p. 265], [92, p. 56].

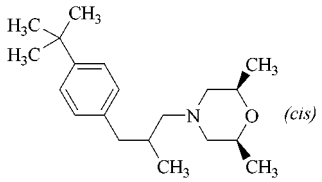

Fenpropimorph is produced by reductive amination of 3-(4-*tert*-butylphenyl)-2-methylpropanal and *cis*-2,6-dimethylmorpholine [125].

Fenpropimorph (announced in 1979) is used as a foliar fungicide in cereals for the control of powdery mildew, rust, and *Rhynchosporium secalis*.
Trade Name. Corbel (BASF, Novartis).

Dimethomorph [*110488-70-5*], (*E,Z*)-4-[3-(4-chlorophenyl)-3-(3,4-dimethoxyphenyl)acryloyl]morpholine, $C_{21}H_{22}ClNO_4$, M_r 387.9, *mp* 127 – 148 °C, is a colorless crystalline solid that is practically insoluble in water, slightly soluble in acetone, toluene and methanol, and soluble in dichloromethane [47, p. 17].

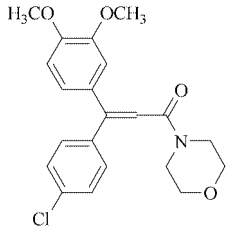

Dimethomorph is produced by Friedel – Crafts acylation of vertrole with 4-chlorobenzoyl chloride, followed by condensation with triethyl phosphonoacetate in the presence of a base such as sodium hydride in dimethoxyethane. The resulting ethyl cinnamate is then converted to the corresponding morpholide [126].

Announced in 1988, dimethomorph is active against downy mildew fungi (Peronosporaceae) and the genus *Phytophthora*. It is being developed in mixtures with protectant fungicides for the control of downy mildew of grapevine and late blight of potato and tomato. It shows translaminar movement following foliar application as well as protectant, curative, and antisporulant activity.
Trade Names. Acrobat, Forum (Cyanamid).

Dodemorph acetate [*31717-87-0*], *N*-cyclododecyl-2,6-dimethylmorpholinium acetate, $C_{20}H_{39}NO_3$, M_r 341.5, *mp* 63-64 °C, is a colorless solid which is sparingly soluble in water and soluble in chloroform, ethanol, acetone, ethyl acetate, and cyclohexane [22, p. 449].

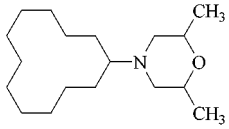

Dodemorph acetate is produced by reaction of 2,6-dimethyltetrahydro-1,4-oxazin with dodecyl ketone to give the corresponding enamine, followed by hydrogenation [127].

Dodemorph (announced in 1968) is used to control powdery mildews on roses and other ornamentals.
Trade Name. Meltatox (BASF).

8.6.7. Strobulins

The strobulins are a new group of broad-spectrum, systemic, preventive and eradicant fungicides active against a wide range of *Ascomycete* and *Deuteromycete* fungi, but especially *Botrytis* spp. and *Venturia* spp.

Azoxystrobin [*131860-33-8*], (*E*)-2-{2-[6-(2-cyanophenoxy)pyrimidin-4-yloxy]phenyl}-3-methoxyacrylate, $C_{22}H_{17}N_3O_5$, M_r 403.4, *mp* 116 °C, is a colorless solid which is sparingly soluble in water, hexane, and *n*-octanol, moderately soluble in methanol, toluene, and

acetone, and readily soluble in ethyl acetate, acetonitrile, and dichloromethane [22, p. 70].

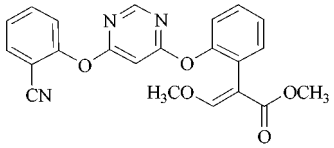

Azoxystrobin is produced by reaction of 4,6-dichloropyrimidine with (*E*)-2-hydroxymethoxyacrylate in DMF containing potassium carbonate, followed by reaction with 2-cyanophenol in DMF containing potassium carbonate [128].

Azoxystrobin (announced in 1992) inhibits mitochondrial respiration by blocking electron transfer between cytochrome b and cytochtome c_1. It is used to control a wide range of pathogens.

Trade Names. Amistar, Heritage, Quadris (Zeneca).

Kresoxim-methyl [*143390-89-0*], methyl (*E*)-2-methoxyimino[α-(*o*-tolyloxy)-*o*-tolyl]acetate, $C_{18}H_{19}NO_4$, M_r 313.4, *mp* 97.2 – 101.7 °C, consists of colorless crystals which are sparingly soluble in water [22, p. 743].

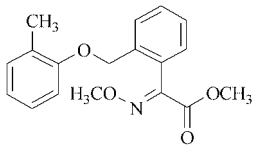

Kresoxim-methyl is produced by by the Grignard reaction of 1-(2-methylphenoxy)-2-bromobenzene in THF with methyloxalylimidazole to give 2-(2-methyloxalylimidazole)phenylglyoxylic acid methyl ester, which upon reaction with methylhydroxylamine hydrochloride in methanol containing sodium carbonate yields the desired product [129].

Kresoxim-methyl (announced in 1992) is used to control scab in apples and pears and powdery mildew on apples, vines, cucurbits, and sugar beets.

Trade Names. Allegro, Stroby, Mentor (BASF).

Metominostrobin [*133408-50-1*], (*E*)-2-methoxyimino-*N*-methyl-2-(2-phenoxyphenyl)-acetamide, $C_{16}H_{16}N_2O_3$, M_r 284.3, *mp* 87 – 89 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in dichloromethane and chloroform [22, p. 1114].

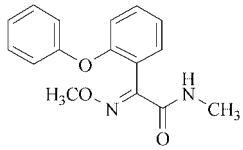

Metominostrobin is produced by reaction of 1-bromo-2-phenoxybenzene with magnesium in THF in the presence of dibromethane, followed by chlorination with oxalyl chloride in THF, and amidation with methylamine in methanol [130].

Metominostrobin (announced in 1994) is used to control *Pyricularia orzyae* on rice.

Trade Name. Oribright (Shionogi).

8.7. Miscellaneous

Propamocarb [*24579-73-5*], propyl 3-(dimethylamino)propylcarbamate, $C_9H_{21}ClN_2O_2$, M_r 224.7, *mp* 45 – 55 °C (hydrochloride), vapor pressure 800 mPa at 25 °C; the hydrochloride is a colorless crystalline solid, which is hygroscopic and soluble in water, methanol, and dichloromethane [9, p. 461], [15, p. 404].

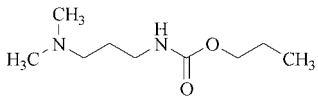

Propamocarb is produced by reaction of *N,N*-dimethylpropane-1,3-diamine with propyl chloroformate in the presence of a base, followed by treatment with hydrogen chloride [131].

Propamocarb (announced in 1978) is used as a soil and seed treatment, a dip treatment for bulbs and tubers, and a foliar fungicide. It should be used preventatively; its spectrum of activity is confined to Oomycetes.

Trade Names. Previcur, Banol (AgrEvo).

Isoprothiolane [*50512-35-1*], diisopropyl (1,3-dithiolan-2-ylidene)malonate, $C_{12}H_{18}O_4S_2$, M_r 290.4, *mp* 50 – 54.5 °C, *bp* 167 – 169 °C at 70 Pa, consists of colorless crystals which are slightly soluble in water [9, p. 328], [14, p. 198].

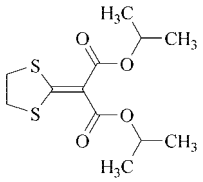

Isoprothiolane is produced by condensation of diisopropyl malonate with carbon disulfide in the presence of sodium hydroxide, followed by alkylation with 1,2-dichloroethane [81].

Isoprothiolane (announced in 1973) is used as a foliar spray and an "into-water treatment" for the control of rice blast. It is translocated to the leaves and heads after uptake by the roots. Resistant strains of *Pyricularia oryzae* have been detected in field populations in Japan.

Trade Name. Fuji-one (Nihon Nohyaku).

Pyroquilon [*57369-32-1*], 1,2,5,6-tetrahydro-4*H*-pyrrolo[3,2,1-*ij*]-quinolin-4-one, $C_{11}H_{11}NO$, M_r 173.2, *mp* 112 °C, vapor pressure 0.16 mPa at 20 °C, consists of colorless crystals which are practically insoluble in water but soluble in methanol, dichloromethane, benzene, and acetone [22, p. 1075].

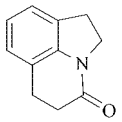

Pyroquilon is produced by reaction of indoline and chloroacetyl chloride. The resulting intermediate is then cyclocondensed with dimethylformamide in the presence of phosphorus oxychloride and hydrogenated [132].

Pyroquilon (announced in 1979) is used in the preventive control of rice blast as a seed and seedling box treatment, and as an "into-water treatment" in the field. The mode of action is by inhibition of melanin synthesis; the lack of melanin prevents the fungus from penetrating the leaf.

Trade Names. Coratop, Fongorene (Novartis).

Hymexazol [*10004-44-1*], 5-methylisoxazol-3-ol, $C_4H_5NO_2$, M_r 99.15, *mp* 86 °C, vapor pressure < 133 mPa at 25 °C, consists of colorless crystals which are soluble in water and most organic solvents [9, p. 314], [14, p. 215], [15, p. 399].

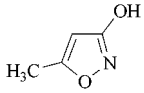

Hymexazol is produced by reaction of ethyl 2-butynoate with hydroxylamine [133].

Hymexazol (announced in 1973) is used as a seed and soil treatment. It is active against seedling diseases caused by *Aphanomyces, Pythium* and *Fusarium* spp. It is used as a soil fungicide on ornamentals, rice, and sugar beet and as a seed dressing on sugar beet. It stimulates root growth in rice seedlings.

Trade Name. Tachigaren (Sankyo).

Tricyclazole [*41814-78-2*], 5-methyl-1,2,4-triazolo[3,4-*b*]-benzothiazole, $C_9H_7N_3S$, M_r 189.2, *mp* 187 – 188 °C, vapor pressure 27 µPa at 25 °C, is a crystalline solid which is slightly soluble in water, acetone, methanol, and xylene [9, p. 549], [14, p. 220].

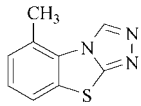

Tricyclazole is produced by reaction of 2-hydrazino-4-methylbenzothiazole with triethyl orthoformate at 100 – 200 °C [134].

Tricyclazole (announced in 1976) is used as a seedling box treatment and as a foliar spray for the control of rice blast. The mode of action is considered to be by the inhibition of melanin synthesis. The lack of melanin prevents the fungus from penetrating the leaf.

Trade Names. Beam (Dow Elanco); Sezola (Sananda).

Ethirimol [*23947-60-6*], 5-butyl-2-ethylamino-6-methylpyrimidin-4-ol, $C_{11}H_{19}N_3O$, M_r 209.3, *mp* 159 – 160 °C, vapor pressure 267 µPa at 25 °C, consists of colorless crystals which are practically insoluble in water, slightly soluble in ethanol, and soluble in chloroform or trichloroethylene [9, p. 245], [14, p. 211].

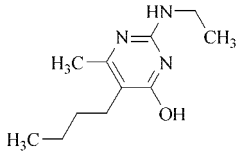

Ethirimol is produced by reaction of ethyl 2-butylacetoacetate and *N*-ethylguanidine [135].

Ethirimol (announced in 1969) is used as a foliar spray and seed treatment. It has protective and curative activity but only against powdery mildew. It is used on cucurbits and cereals. Resistant strains have been detected in field populations of powdery mildew affecting both these crops. The mode of action is by interference with purine metabolism [16, pp. 185 – 205].
Trade Name. Milgo (Zeneca).

Bupirimate [*41483-43-6*], 5-butyl-2-ethylamino-6-methylpyrimidin-4-yl dimethylsulfamate, $C_{13}H_{24}N_4O_3S$, M_r 316.4, *mp* 50 – 51 °C, vapor pressure 67 µPa at 20 °C, is a pale-tan, waxy solid which is slightly soluble in water but soluble in most organic solvents [9, p. 70].

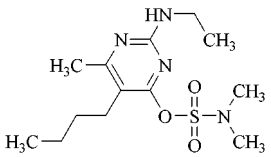

Bupirimate is produced by refluxing the sodium salt of 2-ethylamino-4-methyl-5-butyl-6-hydroxypyrimidine with dimethylsulfamoyl chloride in toluene [136].

Bupirimate (announced in 1975) is used as a foliar fungicide for the control of powdery mildew of apple and ornamentals. It has both protective and curative properties; its mode of action is the same as that of ethirimol.
Trade Name. Nimrod (Zeneca).

Probenazole [*27605-76-1*], 3-(2-propenyloxy)-1,2-benzisothiazole 1,1-dioxide, $C_{10}H_9NO_3S$, M_r 223.24, *mp* 128 – 129 °C, consists of colorless crystals which are slightly soluble in water, readily soluble in most common organic solvents [29, p. 821].

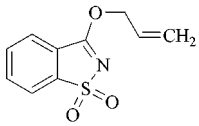

Probenazole is produced by reaction of 3-chloro-1,2-benzoisothiazole 1,1-dioxide with allyl alcohol [137].

Probenazole (announced in 1977) is used in granular form as an "into-water treatment" to control blast and bacterial leaf blight of rice. The compound is effective only when administered via the root system. It has little activity against the target pathogens in vitro and confers protection by the induction of host resistance mechanisms [138].
Trade Name. Oryzemate (Meiji Seika, Hokko).

Carpropamid [*104030-54-8*], 2,2-dichloro-*N*-[1-(4-chlorophenyl)ethyl]-1-ethyl-3-methylcyclopropanecarboxamide, $C_{15}H_{18}Cl_3NO$, M_r 334.7, *mp* 145-149 °C, is a colorless powder which is sparingly soluble in water [22, p. 744].

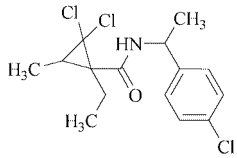

Carpropamid is produced by reaction of 1-ethyl-2,2-dichloro-3-methylcyclopropanecarboxylic acid chloride with 4-chloro-α-methylbenzylamine in toluene in the presence of triethylamine [139].

Carpropamid (announced in 1994) acts specifically against *Pyricularia oryzate*; it does not provide curative properties, so protective treatment is essential. It is also used for seed treatment.
Trade Names. Win, Win Admire, Arcado (Bayer); Seed One, Zabara (Hannona).

Spiroxamine [*118134-30-8*], 8-*tert*-butyl-1,4-dioxaspiro[4,5]decan-2-ylmethyl(ethyl)(propyl)amine, $C_{18}H_{35}NO_2$, M_r 297.5, is a yellowish liquid which is fairly soluble in water (pH 3) [22, p. 746].

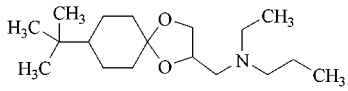

Spiroxamine is produced by ketalization of 4-*tert*-butylcyclohexanone with 1-chloro-2,3-dihydroxypropane, followed by treatment with ethylpropylamine [140].

Spiroxamine (announced in1996) is a new sterol biosynthesis inhibitor acting mainly by inhibition of Δ^{14} reductase. Intended for use as

a sytemic foliar fungicide in wheat and barley, with rapid initial effect and prolonged activity.
Trade Names. To be announced (Bayer).

Fenpropidin [*67306-00-7*], (*RS*)-1-[3-(4-*tert*-butylphenyl)-2-methylpropyl]piperidine, $C_{19}H_{31}N$, M_r 273.5, *bp* >250 °C, is a pale yellow, slightly viscous liquid which is slightly soluble in water but completely miscible with acetone, ethanol, toluene, *n*-octanol, and *n*-heptane [22, p. 526].

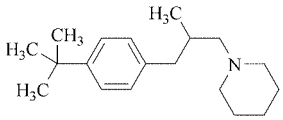

Fenpropidin is produced by condensation of 4-*tert*-butylbenzaldehyde with propanal to give 4-*tert*-butylphenyl-2-methylpropylinealdehyde, the hydrogenation product of which is condensed with piperidine, followed by hydrogenation [141].

Fenpropidin (announced in 1986) is used to control powdery mildew, rusts, and *Rhynchosporium secalis* in barley and wheat.
Trade Names. Term (Novartis); Columbia (Zeneca); Sponsor (AgrEvo).

Famoxadone [*131807-57-3*], 3-anilino-5-methyl-5-(4-phenoxyphenyl)-1,3-oxazolidine-2,4-dione, $C_{22}H_{18}N_2O_4$, M_r 374.4, *mp* 140.3 – 141.8 °C, is practically insoluble in water [22, p. 500].

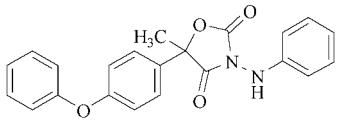

Famoxadone is produced by reaction of 2-(4-phenoxyphenyl)lactate with a mixture of phenoxyphenyl and 1,1′-carbonyldiimidazole in dichloromethane, followed by treatment with phenylhydrazine [142].

Famoxadone (announced in 1996) inhibits mitochondrial electron transport by blocking ubiquinone:cytochrome c oxidoreductase at complex III. It is used to control a broad spectrum of fungi, being particularly effective against grape downy mildew, potato and tomato late and early blights, wheat leaf and glume blotch, and barley net blotch.
Trade Name. To be announced (Du Pont).

Quinoxyfen [*124495-18-7*], 5,7-dichloro-4-quinolyl-4-fluorophenyl ether, $C_{15}H_8Cl_2FNO$, M_r 308.1, [22, p. 1083].

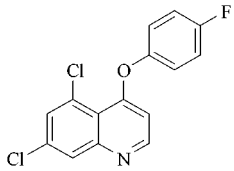

Quinoxyfen is produced by reaction of 4,5,7-trichloroquinoline with *p*-fluorophenol in the presence of 4-dimethylaminopyridine as catalyst in refluxing xylene [143].

Quinoxyfen (announced in 1996) is used to control cereal powdery mildew and offers long-term protection.
Trade Name. To be announced (Dow Elanco).

Table 1. World fungicide market in 1987 [144]

Type of fungicide	Share of world sales∗, %
Nonsystemic fungicides	
Dithiocarbamates	11.2
Inorganics	7.6
Others	12.2
Total nonsystemics	31.0
Systemic fungicides	
Benzimidazoles	6.7
Organophosphates	3.3
Triazoles	23.9
Substituted anilides	11.0
Morpholines	4.8
Strobilurins	3.7
Others	15.6
Total systemics	69.0

∗ Total world sales = $ 5515 × 10^6

9. Economic Aspects

In 1997, fungicides accounted for 18 % of the world market for agrochemicals with an end user value of $ 5515 × 10^6. Fungicide sales were highest in value in Western Europe (43 %), followed by Japan (19 %), North America (14 %), and Latin America (12 %).

Fruit and vegetables accounted for the largest share (47 %) in terms of value of the fungicide market in 1997. This was followed by cereals (25 %) and rice (13.6 %) (→ Crop Protection,

Chap. 3.). Distribution of market shares between the different types of fungicides is shown in Table 1.

10. Toxicology and Occupational Health

A summary of the toxicology of nonsystemic fungicides, fungicidal antibiotics, and systemic fungicides currently in use is given below. Because most of these compounds are proprietary products, many of the toxicological data are not published. Information published in monographs by WHO/FAO was given preference over unreferenced data in handbooks, because the original data were available to the experts of the WHO/FAO panels and the monographs are fully referenced. Furthermore the data of the corresponding studies were conducted according to internationally accepted test protocols and good laboratory practice.

10.1. Nonsystemic (Contact) Fungicides

10.1.1. Inorganic Compounds

Sulfur [*7704-34-9*] has a low toxicity; colloidal sulfur is more toxic than crystalline or powdered sulfur. Irritation of the skin and mucous membranes occurs in persons sensitive to it. In guinea pigs, enzymes were affected only after oral administration of very high dosages for more than 60 d [145], [146]. Sulfur is partly converted to hydrogen sulfide, some of which enters the general metabolic pool. Acute inhibition of cellular respiration by inactivation of enzymes containing heavy metals was observed only at very high doses; this effect is therefore of no significance when sulfur is used as a fungicide.

LD_{50} (mouse, oral)	209 – 245	mg/kg	[22, p. 1193]
LC_{50} (rat, 4-h inhalation)	0.06	mg/L air	[22, p. 1193]
Chronic (rat, 2-year feed)	0.01	$mg\,kg^{-1}\,d^{-1}$	[22, p. 1193]

Copper [*7440-50-8*]. Copper compounds were first employed as a fungicide in the form of bordeaux mixture in the late 1800s. The acute toxicity of the mixture is low. Generally the copper compounds used as fungicides owe their mammalian toxicity to a high overdose of copper ions, in particular copper(II) ion. In high doses, it leads to gastrointestinal disorders [147]. Poisoning due to long-term exposure to copper fungicides, on the average for 18 years, has been described. A specific lung condition produced after long periods of exposure (3 – 15 years) to copper sulfate is known as vineyard sprayer lung. Such an effect has not been described for copper oxychloride [148], [149].

Copper(II) Oxychloride [*1332-40-7*]

LD_{50} (rat, oral)	700 – 800	mg/kg	[22, p. 269]
LD_{50} (rat, dermal)	>2000	mg/kg	[22, p. 269]
LC_{50} (rat, 4-h inhalation)	>30	mg/L air	[22, p. 269]

Copper(I) Oxide [*1317-39-1*]

LD_{50} (rat, oral)	470	mg/kg	[22, p. 277]
LD_{50} (rat, dermal)	>2000	mg/kg	[22, p. 277]

10.1.2. Organometallic Compounds

Organomercury fungicides are no longer used in agriculture because of their toxicity; only organotin compounds are worth mentioning.

Fentin Hydroxide [*76-87-9*]

LD_{50} (rat, oral)	110 – 171	mg/kg	[150]
LD_{50} (dog, oral)	100	mg/kg	[150]
LD_{50} (rat, dermal)	1600	mg/kg	[150]
LD_{50} (rabbit, dermal)	127	mg/kg	[150]
LC_{50} (rat, 4-h inhalation)	0.06	mg/L air	[150]

Rabbit: slightly irritating to skin [150]
Rabbit: severely irritating to eyes [150]
Guinea pig: not a skin sensitizer [150]
Nonmutagenic [150]
NOEL/NOAEL:

Reproduction.			
Rat (2-generation feed):	0.4	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[150]
Teratogenicity.			
Rat (gavage)	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[150]
	2.8	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[150]
Rabbit (gavage):	0.1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[150]
	0.3	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[150]

Chronic.

Rat (2-year feed):	< 0.3	$mg\,kg^{-1}\,d^{-1}$	[150]
Mouse (18-month feed):	1	$mg\,kg^{-1}\,d^{-1}$	[150]
Dog (one-year feed):	0.2	$mg\,kg^{-1}\,d^{-1}$	[150]

ADI: $0.0005\,mg\,kg^{-1}\,d^{-1}$ [150]

Ingestion of Fentin hydroxide by humans causes gastrointestinal disorders. After several hours or even days, headache, giddiness, vomiting, and abdominal pain occur. In serious cases, paralysis, cerebral edema, and ocular changes result [150].

10.1.3. Dithiocarbamates

In general, the acute and dermal toxicity of dithiocarbamate fungicides is low. The chronic toxicity of dithiocarbamates is related to a common metabolite of ethylenebisthiocarbamates, namely, ethylene thiourea (ETU) [*96-45-7*]. It is formed in plants and animals, and in food during processing and cooking. It is carcinogenic in animals, inducing thyroid tumors (follicular adenomas and carcinomas) and liver tumors. Thyroid neoplasia appears to be related to a nongenotoxic, epigenetic mechanism involving inhibition of thyroid peroxidase, a common effect of dithiocarbamates in rats [151].

ETU [*96-45-7*]

LD_{50} (rat, oral)	940 – 2400	mg/kg	[152]
LD_{50} (mouse, oral)	4000	mg/kg	[152]
LD_{50} (hamster)	> 3000	mg/kg	[152]

Nonmutagenic [152]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	0.16 – 0.38	$mg\,kg^{-1}\,d^{-1}$	[152]

Teratogenicity.

Rat (gavage)	40	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152]
	5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152]
Rabbit (gavage):	80	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152]
	40	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152]

Chronic.

Rat (2-year feed):	0.37	$mg\,kg^{-1}\,d^{-1}$	[152]
Mouse (3-month feed):	1.7	$mg\,kg^{-1}\,d^{-1}$	[152]
Dog (one-year feed):	0.18	$mg\,kg^{-1}\,d^{-1}$	[152]

ADI: $0.004\,mg\,kg^{-1}\,d^{-1}$ [152]

Ferbam [*14484-64-1*]

LD_{50} (rat, oral)	> 4000	mg/kg	[153]
LD_{50} (mouse, oral)	1000	mg/kg	[153]
LD_{50} (guinea pig, oral)	450 – 2000	mg/kg	[153]
LD_{50} (rabbit, oral)	2000 – 3000	mg/kg	[154, p. 460]

Rabbit: irritating to skin and eyes [153]
Nonmutagenic [153]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[155]
Dog (2-year feed):	5	$mg\,kg^{-1}\,d^{-1}$	[155]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [154, p. 460]

Mancozeb [*8018-01-7*]

LD_{50} (rat, oral)	> 5000	mg/kg	[152, p. 167]
LD_{50} (mouse, oral)	> 5000	mg/kg	[152, p. 167]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[152, p. 167]
LC_{50} (rat, 4-h inhalation)	5.14	mg/L air	[152, p. 167]

Rabbit: slight irritant to skin [152, p. 167]
Rabbit: irritating to eyes [152, p. 167]
Guinea pig: skin sensitizer [152, p. 167]
Mutagenicity, equivocal [152, p. 167]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	7.0	$mg\,kg^{-1}\,d^{-1}$ (parental)	[152, p. 167]
	70	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 167]

Teratogenicity.

Rat (gavage):	60	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[152, p. 167]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152, p. 167]
	80	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 167]

Chronic.

Rat (2-year feed):	4.8	$mg\,kg^{-1}\,d^{-1}$	[152, p. 167]
Mouse (78-week feed):	17	$mg\,kg^{-1}\,d^{-1}$	[152, p. 167]
Dog (one-year feed):	7.0	$mg\,kg^{-1}\,d^{-1}$	[152, p. 167]

Neuropathology.

Rat (90-day feed):	8.2	$mg\,kg^{-1}\,d^{-1}$	[152, p. 167]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [152, p. 167]

Maneb [*12427-38-2*]

LD_{50} (rat, oral)	> 5000	mg/kg	[152, p. 257]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[152, p. 257]
LC_{50} (rat, 4-h inhalation)	7.38	mg/L air	[152, p. 257]

Rabbit: slightly irritating to skin [152, p. 257]
Rabbit: irritating to eyes [152, p. 257]
Guinea pig: not a skin sensitizer [152, p. 257]
Nonmutagenic [152, p. 257]

NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5.6	$mg\,kg^{-1}\,d^{-1}$ (parental)	[152, p. 257]
	ca. 22	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 257]

Teratogenicity.

Rat (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[152, p. 257]

Chronic.

Rat (31-month feed):	20	$mg\,kg^{-1}\,d^{-1}$	[152, p. 257]
Mouse (79-week feed):	11	$mg\,kg^{-1}\,d^{-1}$	[152, p. 257]
Dog (52-week feed):	6.4	$mg\,kg^{-1}\,d^{-1}$	[152, p. 257]
Monkey (6-month feed):	7.3	$mg\,kg^{-1}\,d^{-1}$	[152, p. 257]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [152, p. 257]

Metiram [*9006-42-2*]

LD_{50} (rat, oral)	> 6810	mg/kg	[152, p. 311]
LD_{50} (rat, dermal)	> 2000	mg/kg	[152, p. 311]
LC_{50} (rat, 4-h inhalation)	5.7	mg/L air	[152, p. 311]

Rabbit: nonirritating to skin and eyes [152, p. 311]
Guinea pig: skin sensitizer [152, p. 311]
Nonmutagenic [152, p. 311]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	1.8	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[152, p. 311]

Teratogenicity.

Rat (gavage):	80	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[152, p. 311]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152, p. 311]
	40	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 311]

Chronic.

Rat (111-119-week feed):	3.1	$mg\,kg^{-1}\,d^{-1}$	[152, p. 311]
Mouse (88-week feed):	24	$mg\,kg^{-1}\,d^{-1}$	[152, p. 311]
Dog (one-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[152, p. 311]
Monkey (26-week gavage):	5	$mg\,kg^{-1}\,d^{-1}$	[152, p. 311]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [152, p. 311]

Propineb [*9016-72-2*]

LD_{50} (rat, oral)	> 5000	mg/kg	[152, p. 369]
LD_{50} (mouse, oral)	> 5000	mg/kg	[152, p. 369]
LD_{50} (cat, oral)	> 500	mg/kg	[152, p. 369]
LD_{50} (sheep, oral)	2500	mg/kg	[152, p. 369]
LD_{50} (rat, dermal)	> 5000	mg/kg	[152, p. 369]
LC_{50} (rat, 4-h inhalation)	> 0.69	mg/L air	[152, p. 369]
LC_{50} (hamster, 4-h inhalation)	> 0.39	mg/L air	[152, p. 369]
LC_{50} (mouse, 4-h inhalation)	> 0.39	mg/L air	[152, p. 369]

Rabbit: nonirritating to skin and eyes [152, p. 369]
Guinea pig: a skin sensitizer [152, p. 386]
Nonmutagenic [152, p. 369]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[152, p. 369]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152, p. 369]
	30	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 369]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[152, p. 369]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[152, p. 369]

Chronic.

Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[152, p. 369]
Mouse (2-year feed):	26	$mg\,kg^{-1}\,d^{-1}$	[152, p. 369]
Dog (2-year feed):	75	$mg\,kg^{-1}\,d^{-1}$	[152, p. 369]
Thyroid function (62-day gavage):	0.74	$mg\,kg^{-1}\,d^{-1}$	[152, p. 369]

ADI: $0.007\,mg\,kg^{-1}\,d^{-1}$ [152, p. 369]

Thiram [*137-26-8*]

LD_{50} (rat, oral)	1800 – 4000	mg/kg	[156]
LD_{50} (mouse, oral)	4000	mg/kg	[156]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[156]
LC_{50} (rat, 4-h inhalation)	> 0.1	mg/L air	[156]

Rabbit: irritating to skin and eyes [154, p. 989]
Nonmutagenic [156]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1.5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[156]
	8.9	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156]

Teratogenicity.

Rat (gavage):	2.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[156]
	7.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156]
Rabbit (gavage):	2.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[156]
	5.0	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156]

Chronic.

Rat (2-year feed):	1.2	$mg\,kg^{-1}\,d^{-1}$	[156]
Mouse (97-week feed):	3	$mg\,kg^{-1}\,d^{-1}$	[156]

Dog (one-year feed):	0.84	$mg\,kg^{-1}\,d^{-1}$	[156]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [156]

Zineb [*12122-67-7*]

LD_{50} (rat, oral)	> 5200	mg/kg	[152, p. 395]
LD_{50} (guinea pig, oral)	> 4800	mg/kg	[152, p. 395]
LD_{50} (rat, dermal)	> 5000	mg/kg	[152, p. 395]

Rabbit: slightly irritating to skin [152, p. 395]
Nonmutagenic [152, p. 395]
NOEL/NOAEL:
Teratogenicity.

Rat (gavage):	630	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[152, p. 396]
Mouse (gavage):	2000	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[152, p. 395]

ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [152, p. 395]

Ziram [*137-30-4*]

LD_{50} (rat, oral)	1400	mg/kg	[153]
LD_{50} (rabbit, oral)	100 – 1020	mg/kg	[153]
LD_{50} (guinea pig, oral)	100 – 150	mg/kg	[153]
LD_{50} (rat, dermal)	> 6000	mg/kg	[154, p. 1050]

Rabbit: irritating to skin and eyes [154, p. 1050]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1050]
Dog (one-year capsule):	5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1050]

ADI: 0.02 $mg\,kg^{-1}\,d^{-1}$ [154, p. 1050]

10.1.4. *N*-Trihalomethylthio Derivatives (Phthalimides)

The acute oral and dermal toxicity of *N*-trihalomethylthio fungicides is low. Controversy continues regarding their mutagenic and teratogenic effects. Sperm mobility was inhibited in rats and mice after intake for longer periods. Particular attention has been paid to the teratogenic effects of *N*-trihalomethylthio derivatives because of their close structural relationship to thalidomide. Captan, folpet, and their primary metabolites did not produce teratogenic changes in rabbits. In chronic feeding experiments with captan in mice, tumor formation was reported in the small intestine. This could be because the compound is probably metabolized in the gastrointestinal tract. High dosages inhibit mitochondrial respiration in isolated liver cells; lower doses cause oxidative uncoupling.

Captafol [*2425-06-1*]

LD_{50} (rat, oral)	2500 – 6200	mg/kg	[157]
LD_{50} (rabbit, dermal)	> 15 000	mg/kg	[157]

Nonmutagenic [158]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[158]
Dog (one-year capsule):	10	$mg\,kg^{-1}\,d^{-1}$	[158]

Captan [*133-06-2*]

LD_{50} (rat, oral)	2650	mg/kg	[159]
LD_{50} (mouse, oral)	130	mg/kg	[159]
LD_{50} (rabbit, oral)	740	mg/kg	[159]

Nonmutagenic (in-vivo) [160]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[160]

Teratogenicity.

Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160]
	40	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160]

Chronic.

Mouse (20-month feed):	60	$mg\,kg^{-1}\,d^{-1}$	[160]
Dog (66-week capsule):	100	$mg\,kg^{-1}\,d^{-1}$	[160]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [160]

Dichlofluanid [*1085-98-9*]

LD_{50} (rat, oral)	500 – 1000	mg/kg	[161]
LD_{50} (chicken, oral)	> 1000	mg/kg	[161]
LD_{50} (guines pig, oral)	250	mg/kg	[161]

Rabbit: slightly irritating to skin [162]
Rabbit: moderately irritating to eyes [162]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	75	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[161]

Teratogenicity.

Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[162]

Chronic.

Rat (2-year feed):	30	$mg\,kg^{-1}\,d^{-1}$	[162]
Mouse (2-year feed):	252	$mg\,kg^{-1}\,d^{-1}$	[162]
Dog (2-year feed):	25	$mg\,kg^{-1}\,d^{-1}$	[162]

ADI: 0.3 $mg\,kg^{-1}\,d^{-1}$ [162]

Folpet [*133-07-3*]

LD_{50} (rat, oral)	> 5000	mg/kg	[160, p. 127]
LD_{50} (rabbit, dermal)	> 23 000	mg/kg	[160, p. 127]
LC_{50} (rat, 4-h inhalation)	0.39	mg/L air	[160, p. 127]

Guinea pig: irritating to skin [160, p. 127]
Guinea pig: a skin sensitizer [160, p. 127]
Nonmutagenic (in vivo) [160, p. 127]

NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	13.7	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[163]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[160, p. 127]

Chronic.

Rat (2-year feed):	9.5	$mg\,kg^{-1}\,d^{-1}$	[163]
Mouse (2-year feed):	16	$mg\,kg^{-1}\,d^{-1}$	[160, p. 127]
Dog (one-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[160, p. 127]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [160, p. 127]

Tolylfluanid [*731-27-1*]

LD_{50} (rat, oral)	> 5000	mg/kg	[164]
LD_{50} (mouse, oral)	> 1000	mg/kg	[164]
LD_{50} (rabbit, oral)	> 1000	mg/kg	[164]
LD_{50} (rat, dermal)	> 5000	mg/kg	[164]

Rabbit: nonirritating to skin [164]
Rabbit: irritating to eyes [164]
Guinea pig: a skin sensitizer [164]
Nonmutagenic (in vivo) [164]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	15	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[164]

Teratogenicity.

Rat (gavage):	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[164]

Chronic.

Rat (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[164]
Mouse (2-year feed):	714	$mg\,kg^{-1}\,d^{-1}$	[164]
Dog (one-year capsule):	12.5	$mg\,kg^{-1}\,d^{-1}$	[164]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [164]

10.1.5. Dicarboximides

Dicarboximides do not have a very high acute oral or dermal toxicity in rats, rabbits, and some species of wildlife. In animal experiments, apathy and respiratory disturbances predominate; dogs undergo a temporary loss of appetite.

Chlozolinate [*84332-86-5*]

LD_{50} (rat, oral)	> 4500	mg/kg	[154, p. 208]
LD_{50} (mouse, oral)	> 10 000	mg/kg	[154, p. 208]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 208]
LC_{50} (rat, 4-h inhalation)	> 10	mg/L air	[154, p. 208]

Rabbit: nonirritating to skin and eyes [154, p. 208]
Guinea pig: a skin sensitizer [154, p. 208]

Iprodione [*36734-19-7*]

LD_{50} (rat, oral)	> 2000	mg/kg	[156, p. 211]
LD_{50} (mouse, oral)	1870 – 2670	mg/kg	[156, p. 211]
LD_{50} (dog, oral)	> 2000	mg/kg	[156, p. 211]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[156, p. 211]
LC_{50} (rat, 4-h inhalation)	> 3.3	mg/L air	[156, p. 211]

Rabbit: mildly irritating to eyes [156, p. 211]
Rabbit: nonirritating to skin [156, p. 211]
Guinea pig: not a skin sensitizer [156, p. 211]
Nonmutagenic [156, p. 211]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	21	$mg\,kg^{-1}\,d^{-1}$ (parental)	[156, p. 211]
	71	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156, p. 211]

Teratogenicity.

Rat (gavage):	200	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[156, p. 211]
	90	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156, p. 211]
Rabbit (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[156, p. 211]
	60	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156, p. 211]

Chronic.

Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[156, p. 211]
Dog (one-year feed):	18	$mg\,kg^{-1}\,d^{-1}$	[156, p. 211]

ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$ [156, p. 211]

Procymidone [*32809-16-8*]

LD_{50} (rat, oral)	> 5000	mg/kg	[165]
LD_{50} (mouse, oral)	> 5000	mg/kg	[165]
LD_{50} (rat, dermal)	> 5000	mg/kg	[165]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[165]
LC_{50} (rat, 4-h inhalation)	> 1.5	mg/L air	[165]

Rabbit: nonirritating to skin and eyes [165]
Guinea pig: not a skin sensitizer [165]
Nonmutagenic [165]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[165]

Teratogenicity.

Rat (gavage):	300	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[165]
Rabbit (gavage):	750	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[165]

Chronic.

Rat (2-year feed):	4.6	$mg\,kg^{-1}\,d^{-1}$	[165]
Mouse (2-year feed):	14.3	$mg\,kg^{-1}\,d^{-1}$	[165]
Dog (6-month capsule):	100	$mg\,kg^{-1}\,d^{-1}$	[165]

ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$ [165]

Vinclozolin [*50471-44-8*]

LD_{50} (rat, oral)	> 15 000	mg/kg	[160, p. 181]
LD_{50} (mouse, oral)	> 15 000	mg/kg	[160, p. 181]
LD_{50} (guinea pig, oral)	8000	mg/kg	[160, p. 181]
LD_{50} (rat, dermal)	> 5000	mg/kg	[160, p. 181]
LC_{50} (rat, 4-h inhalation)	> 29.1	mg/L air	[160, p. 181]

Rabbit: nonirritating to skin and eyes [160, p. 181]
Guinea pig: a skin sensitizer [160, p. 181]
Nonmutagenic [160, p. 181]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	4	$mg\,kg^{-1}\,d^{-1}$ (parental)	[160, p. 181]
	15	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 181]
Teratogenicity.			
Rat (gavage):	400	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 181]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 181]
Rabbit (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 181]
	200	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 181]
Chronic.			
Rat (2-year feed):	1.4	$mg\,kg^{-1}\,d^{-1}$	[160, p. 181]
Mouse (18-month feed):	24	$mg\,kg^{-1}\,d^{-1}$	[160, p. 181]
Dog (one-year feed):	2.4	$mg\,kg^{-1}\,d^{-1}$	[160, p. 181]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [160, p. 181]

10.1.6. Miscellaneous Organic Compounds

Biphenyl [*92-52-4*]

LD_{50} (rat, oral)	2410	mg/kg	[157, p. 1433]
LD_{50} (rabbit, oral)	3280	mg/kg	[157, p. 1433]

Rabbit: nonirritating to skin [157, p. 1433]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	50	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[157, p. 1433]
Chronic.			
Rat (2-year feed):	25	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]
Dog (one-year feed):	20	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]
Monkey (one-year feed):	20	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]

ADI: 0.125 $mg\,kg^{-1}\,d^{-1}$ [154, p. 105]

Chlorothalonil [*1897-45-6*]

LD_{50} (rat, oral)	10 000	mg/kg	[163, p. 101]
LD_{50} (dog, oral)	> 5000	mg/kg	[163, p. 101]
LD_{50} (rabbit, dermal)	10 000	mg/kg	[163, p. 101]
LC_{50} (rat, 4-h inhalation)	0.6	mg/L air	[154, p. 193]

Rabbit: mildly irritating to skin [154, p. 193]
Rabbit: severely irritating to eyes [154, p. 193]
Nonmutagenic [156, p. 103]
Nonteratogenic [156, p. 103]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	75	$mg\,kg^{-1}\,d^{-1}$ (parental)	[156, p. 103]
Chronic.			
Rat (2-year feed):	1.5	$mg\,kg^{-1}\,d^{-1}$	[156, p. 103]
Mouse (2-year feed):	1.6	$mg\,kg^{-1}\,d^{-1}$	[156, p. 103]
Dog (2-year feed):	3.0	$mg\,kg^{-1}\,d^{-1}$	[156, p. 103]

ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [156, p. 103]

Diphenylamine [*122-39-6*]

LD_{50} (rat, oral)	3280	mg/kg	[157, p.1433]
LD_{50} (rabbit, oral)	2410	mg/kg	[157, p.1433]

NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	50	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]
Chronic.			
Rat (2-year feed):	25	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]
Mouse (80-week feed):	15	$mg\,kg^{-1}\,d^{-1}$	[166]
Dog (one-year feed):	> 20	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]
Monkey (one-year feed):	20	$mg\,kg^{-1}\,d^{-1}$	[157, p. 1433]

ADI: 0.025 $mg\,kg^{-1}\,d^{-1}$ [166]

2-Phenylphenol [*90-43-7*]

LD_{50} (rat, oral)	2700	mg/kg	[154, p. 794]
LD_{50} (mouse, oral)	2000	mg/kg	[154, p. 794]

Rabbit: mildly irritating to skin [154, p. 794]
Nonmutagenic [167]
NOEL/NOAEL:
Reproduction.

Rat (2-generation gavage):	40	$mg\,kg^{-1}\,d^{-1}$ (parental)	[167]
	49	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[167]
Teratogenicity.			
Rat (gavage):	150	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[167]
	300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[167]

Chronic.

Rat (2-year feed):	2	$mg\,kg^{-1}\,d^{-1}$	[167]
Dog (one-year gavage):	300	$mg\,kg^{-1}\,d^{-1}$	[167]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [167]

Quintozene [*82-68-8*]

LD_{50} (rat, oral)	> 1500	mg/kg	[160, p. 173]
LD_{50} (dog, oral)	> 1500	mg/kg	[160, p. 173]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[160, p. 173]
LC_{50} (rat, 4-h inhalation)	> 1.7	mg/L air	[160, p. 173]

Rabbit: nonirritating to skin [160, p. 173]
Rabbit: mildly irritating to eyes [160, p. 173]
Guinea pig: a mild skin sensitizer [160, p. 173]
Human, a mild skin sensitizer [160, p. 173]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[160, p. 173]

Teratogenicity.

Rat (gavage):	> 1200	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[160, p. 173]
Rabbit (gavage):	12	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 173]
	120	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 173]

Chronic.

Rat (2-year feed):	1	$mg\,kg^{-1}\,d^{-1}$	[160, p. 173]
Mouse (103-week feed):	390	$mg\,kg^{-1}\,d^{-1}$	[160, p. 173]
Dog (one-year feed):	4.2	$mg\,kg^{-1}\,d^{-1}$	[160, p. 173]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [160, p. 173]

sec-Butylamine [*13952-84-6*]

LD_{50} (rat, oral)	350	mg/kg	[168]
LD_{50} (mouse, oral)	660	mg/kg	[168]
LD_{50} (dog, oral)	250	mg/kg	[168]
LD_{50} (rabbit, dermal)	2500	mg/kg	[168]
LC_{50} (rat, 4-h inhalation)	3.5	mg/L air	[168]

Rabbit: irritating to skin [168]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	35	$mg\,kg^{-1}\,d^{-1}$	[168]
Dog (2-year capsule):	69	$mg\,kg^{-1}\,d^{-1}$	[168]

ADI: $0.2\,mg\,kg^{-1}\,d^{-1}$ [168]

Chinomethionat [*2439-01-2*]

LD_{50} (rat, oral)	> 1000	mg/kg	[158, p. 63]
LD_{50} (rat, dermal)	> 2000	mg/kg	[158, p. 63]
LC_{50} (rat, 4-h inhalation)	3	mg/L air	[154, p. 163]

Rabbit: slightly irritating to skin and eyes [154, p. 163]
Nonmutagenic [169]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	0.75	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[169]

Teratogenicity.

Rat (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[169]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[169]

Chronic.

Rat (2-year feed):	0.6	$mg\,kg^{-1}\,d^{-1}$	[169]
Mouse (21-month feed):	16	$mg\,kg^{-1}\,d^{-1}$	[169]
Dog (one-year feed):	0.6	$mg\,kg^{-1}\,d^{-1}$	[169]

ADI: $0.006\,mg\,kg^{-1}\,d^{-1}$ [169]

Cymoxanil [*57966-95-7*]

LD_{50} (rat, oral)	1196 – 1390	mg/kg	[154, p. 257]
LD_{50} (guinea pig, oral)	1096	mg/kg	[154, p. 257]
LD_{50} (rabbit, dermal)	> 3000	mg/kg	[154, p. 257]
LD_{50} (dog, dermal)	> 3000	mg/kg	[154, p. 257]
LC_{50} (rat, 4-h inhalation)	> 5.06	mg/L air	[154, p. 257]

Rabbit: mildly irritating to eyes [154, p. 257]
Rabbit: nonirritating to skin [154, p. 257]
Guinea pig: not a skin sensitizer [154, p. 257]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 257]

Dichlorophen [*97-23-4*]

LD_{50} (rat, oral)	2690	mg/kg	[154, p. 306]
LD_{50} (mouse, oral)	1000	mg/kg	[154, p. 306]
LD_{50} (guinea pig, oral)	1250	mg/kg	[154, p. 306]
LD_{50} (dog, oral)	2000	mg/kg	[154, p. 306]

NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	133	$mg\,kg^{-1}\,d^{-1}$	[154, p.]306

Diclomezine [*62865-36-5*]

LD_{50} (rat, oral)	> 12 000	mg/kg	[154, p. 317]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 317]
LC_{50} (rat, 4-h inhalation)	0.82	mg/L air	[154, p. 317]

Rabbit: nonirritating to skin [154, p. 317]
Nonmutagenic [154, p. 317]
Nonteratogenic [154, p. 317]

Dicloran [*99-30-9*]

LD_{50} (rat, oral)	4000 – 10 000	mg/kg	[170]
LD_{50} (mouse, oral)	1500 – 2500	mg/kg	[170]
LD_{50} (cat, oral)	> 500	mg/kg	[170]
LD_{50} (guinea pig, oral)	1450	mg/kg	[170]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[170]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 318]
LC_{50} (rat, one-h inhalation)	> 21.6	mg/L air	[154, p. 318]

Rabbit: nonirritating to skin and eyes [154, p. 318]
Guinea pig: not a skin sensitizer [154, p. 318]
Nonmutagenic [158, p. 39]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	25	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[158, p. 39]
Chronic.			
Rat (2-year feed):	50	$mg\,kg^{-1}\,d^{-1}$	[158, p. 39]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 39]

ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [158, p. 39]

Dithianon [*3347-22-6*]

LD_{50} (rat, oral)	678 – 720	mg/kg	[156, p. 165]
LD_{50} (mouse, oral)	492 – 528	mg/kg	[156, p. 165]
LD_{50} (rat, dermal)	> 2000	mg/kg	[156, p. 165]
LD_{50} (mouse, dermal)	> 3200	mg/kg	[156, p. 165]
LC_{50} (rat, 4-h inhalation)	2.1	mg/L air	[156, p. 165]

Rabbit: nonirritating to skin [156, p. 165]
Rabbit: severely irritating to eyes [156, p. 165]
Guinea pig: not a skin sensitizer [156, p. 165]
Nonmutagenic [156, p. 165]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	10	$mg\,kg^{-1}\,d^{-1}$ (parental)	[156, p. 165]
Teratogenicity.			
Rat (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 165]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[156, p. 165]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[157, p. 165]
Chronic.			
Rat (2-year feed):	1	$mg\,kg^{-1}\,d^{-1}$	[156, p. 165]
Mouse (80-week feed):	2.9	$mg\,kg^{-1}\,d^{-1}$	[156, p. 165]
Dog (2-year feed):	1	$mg\,kg^{-1}\,d^{-1}$	[156, p. 165]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [156, p. 165]

Dodine [*2439-10-3*]

LD_{50} (rat, oral)	750 – 1540	mg/kg	[161, p. 265]
LD_{50} (rabbit, oral)	535	mg/kg	[161, p. 265]
LD_{50} (guinea pig, oral)	176	mg/kg	[161, p. 265]
LD_{50} (rabbit, dermal)	2100 – 10 000	mg/kg	[161, p. 265]

Rabbit: severely irritating to skin and eyes [161, p. 265]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	< 40	$mg\,kg^{-1}\,d^{-1}$ (parental)	[161, p. 265]
	> 40	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[161, p. 265]
Chronic.			
Rat (2-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[161, p. 265]
Dog (one-year feed):	1.25	$mg\,kg^{-1}\,d^{-1}$	[161, p. 265]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [161, p. 265]

Etridiazole [*2593-15-9*]

LD_{50} (rat, oral)	1028	mg/kg	[154, p. 419]
LD_{50} (mouse, oral)	2000	mg/kg	[154, p. 419]
LD_{50} (rabbit, oral)	779	mg/kg	[154, p. 419]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[154, p. 419]
LC_{50} (rat, 4-h inhalation)	> 5.7	mg/L air	[154, p. 419]

Rabbit: nonirritating to skin [154, p. 419]
Rabbit: slightly irritating to eyes [154, p. 419]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	4	$mg\,kg^{-1}\,d^{-1}$	[154, p. 419]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 419]

ADI: 0.025 $mg\,kg^{-1}\,d^{-1}$ [154, p. 419]

Fenfuram [*24691-80-3*]

LD_{50} (rat, oral)	12 900	mg/kg	[154, p. 473]
LD_{50} (cat, oral)	2450	mg/kg	[154, p. 473]
LC_{50} (rat, 4-h inhalation)	> 10.3	mg/L air	[154, p. 473]

NOEL/NOAEL:
Chronic.

Rat (2-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[154, p. 473]
Subchronic.			
Dog (3-month feed):	7.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 473]

Fenpiclonil [*74738-17-3*]

LD_{50} (rat, oral)	> 5000	mg/kg	[154, p. 444]
LD_{50} (mouse, oral)	> 5000	mg/kg	[154, p. 444]
LD_{50} (rabbit, oral)	> 5000	mg/kg	[154, p. 444]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 444]
LC_{50} (rat, 4-h inhalation)	> 1.5	mg/L air	[154, p. 444]

Rabbit: nonirritating to skin and eyes [154, p. 444]
Guinea pig: not a skin sensitizer [154, p. 444]
ADI: 0.0125 mg kg^{-1} d^{-1} [154, p. 444]

Ferimzone [*89269-64-7*]

LD_{50} (rat, oral)	642 – 721	mg/kg	[154, p. 461]
LD_{50} (mouse, oral)	542 – 590	mg/kg	[154, p. 461]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 461]
LC_{50} (rat, 4-h inhalation)	> 3.8	mg/L air	[154, p. 461]

Fludioxonil [*131341-86-1*]

LD_{50} (rat, oral)	> 5000	mg/kg	[171]
LD_{50} (mouse, oral)	> 5000	mg/kg	[154, p. 482]
LD_{50} (rat, dermal)	> 2000	mg/kg	[171]
LC_{50} (rat, 4-h inhalation)	> 2.6	mg/L air	[171]

Rabbit: nonirritating to skin and eyes [171]
Guinea pig: not a skin sensitizer [171]
Nonmutagenic [171]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	15	mg kg^{-1} d^{-1} (parental)	[171]
	150	mg kg^{-1} d^{-1} (developmental)	[171]

Teratogenicity.

Rat (gavage):	100	mg kg^{-1} d^{-1} (maternal)	[171]
	1000	mg kg^{-1} d^{-1} (developmental)	[171]
Rabbit (gavage):	10	mg kg^{-1} d^{-1} (maternal)	[171]
	300	mg kg^{-1} d^{-1} (developmental)	[171]

Chronic.

Rat (2-year feed):	50	mg kg^{-1} d^{-1}	[171]
Mouse (18-month feed):	143	mg kg^{-1} d^{-1}	[171]
Dog (one-year feed):	2.5	mg kg^{-1} d^{-1}	[171]

ADI: 0.025 mg kg^{-1} d^{-1} [171]

Fluoroimide [*41205-21-4*]

LD_{50} (rat, oral)	> 15 000	mg/kg	[154, p. 494]
LD_{50} (mouse, oral)	> 15 000	mg/kg	[154, p. 494]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[154, p. 494]
LC_{50} (rat, 4-h inhalation)	0.52 – 0.72	mg/L air	[154, p. 494]

NOEL/NOAEL:
Chronic.

Rat (2-year feed):	30	mg kg^{-1} d^{-1}	[154, p. 494]

Flusulfamide [*106917-52-6*]

LD_{50} (rat, oral)	132 – 180	mg/kg	[154, p. 511]
LD_{50} (mouse, oral)	245 – 254	mg/kg	[154, p. 511]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 511]
LC_{50} (rat, 4-h inhalation)	0.47	mg/L air	[154, p. 511]

NOEL/NOAEL:
Chronic.

Rat (2-year feed):	0.1	mg kg^{-1} d^{-1}	[154, p. 511]

Mepanipyrim [*110235-47-7*]

LD_{50} (mouse, oral)	> 5000	mg/kg	[154, p. 652]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 652]
LD_{50} (rat, oral)	> 2000	mg/kg	[154, p. 652]

Rabbit: nonirritating to skin and eyes [154, p. 652]
Guinea pig: not a skin sensitizer [154, p. 652]

Nitrothal-isopropyl [*10552-74-6*]

LD_{50} (rat, oral)	> 6400	mg/kg	[154, p. 739]
LD_{50} (rat, dermal)	> 2500	mg/kg	[154, p. 739]
LD_{50} (rabbit, dermal)	> 4000	mg/kg	[154, p. 739]
LC_{50} (rat, 8-h inhalation)	> 2.8	mg/L air	[154, p. 739]

Rabbit: slightly irritating to eyes [154, p. 739]
NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	25	mg kg^{-1} d^{-1}	[154, p. 739]
Dog (90-day feed):	500	mg kg^{-1} d^{-1}	[154, p. 739]

Oxycarboxin [*5259-88-1*]

LD_{50} (rat, oral)	2000	mg/kg	[154, p. 761]
LD_{50} (rabbit, dermal)	> 16 000	mg/kg	[154, p. 761]
LC_{50} (rat, one-h inhalation)	> 20	mg/L air	[154, p. 761]

Rabbit: irritating to eyes [154, p. 762]
Rabbit: slightly irritating to skin [154, p. 761]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[154, p. 761]
Dog (2-year feed):	75	$mg\,kg^{-1}\,d^{-1}$	[154, p. 761]

ADI: $0.15\,mg\,kg^{-1}\,d^{-1}$ [154, p. 761]

Pencycuron [*66063-05-6*]

LD_{50} (rat, oral)	> 5000	mg/kg	[154, p. 777]
LD_{50} (mouse, oral)	> 5000	mg/kg	[154, p. 777]
LD_{50} (dog, oral)	> 5000	mg/kg	[154, p. 777]
LD_{50} (cat, oral)	> 1000	mg/kg	[154, p. 777]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 777]
LD_{50} (mouse, dermal)	> 2000	mg/kg	[154, p. 777]
LC_{50} (rat, 4-h inhalation)	> 0.27	mg/L air	[154, p. 777]

Rabbit: nonirritating to skin and eyes [154, p. 777]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	1.8	$mg\,kg^{-1}\,d^{-1}$	[154, p. 777]
Mouse (2-year feed):	71	$mg\,kg^{-1}\,d^{-1}$	[154, p. 777]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 777]

ADI: $0.018\,mg\,kg^{-1}\,d^{-1}$ [154, p. 777]

Phthalide [*27355-22-2*]

LD_{50} (rat, oral)	> 10 000	mg/kg	[154, p. 810]
LD_{50} (mouse, oral)	> 10 000	mg/kg	[154, p. 810]
LD_{50} (rat, dermal)	> 10 000	mg/kg	[154, p. 810]
LD_{50} (mouse, dermal)	> 10 000	mg/kg	[154, p. 810]
LC_{50} (rat, 4-h inhalation)	> 0.004	mg/L air	[154, p. 810]

Rabbit: nonirritating to skin and eyes [154, p. 810]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	100	$mg\,kg^{-1}\,d^{-1}$	[154, p. 810]
Mouse (2-year feed):	14	$mg\,kg^{-1}\,d^{-1}$	[154, p. 810]

Tolclofos methyl [*57018-04-9*]

LD_{50} (rat, oral)	~ 5000	mg/kg	[172]
LD_{50} (mouse, oral)	3500	mg/kg	[172]
LD_{50} (dog, oral)	> 1000	mg/kg	[172]
LD_{50} (rat, dermal)	> 5000	mg/kg	[172]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[172]
LC_{50} (rat, 4-h inhalation)	> 3.32	mg/L air	[172]

Rabbit: nonirritating to skin and eyes [172]
Guinea pig: not a skin sensitizer [172]
Nonmutagenic [172]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	> 100	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[172]

Teratogenicity.

Rat (gavage):	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[172]
	1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[172]
Rabbit (gavage):	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[172]
	3000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[172]

Chronic.

Rat (122-129-week feed):	41	$mg\,kg^{-1}\,d^{-1}$	[172]
Mouse (104-week feed):	6.5	$mg\,kg^{-1}\,d^{-1}$	[172]
Dog (one-year feed):	1.1	$mg\,kg^{-1}\,d^{-1}$	[172]

ADI: $0.07\,mg\,kg^{-1}\,d^{-1}$ [172]

Triazoxide [*72459-58-6*]

LD_{50} (rat, oral)	~ 150	mg/kg	[154, p. 1009]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 1009]
LC_{50} (rat, 4-h inhalation)	0.75	mg/L air	[154, p. 1009]

Rabbit: nonirritating to skin and eyes [154, p. 1009]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	0.05	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1009]
Mouse (2-year feed):	0.7	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1009]
Dog (one-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1009]

ADI: $0.0005\,mg\,kg^{-1}\,d^{-1}$ [154, p. 1009]

10.2. Fungicidal Antibiotics

Fungicidal antibiotics have a low toxicity. The acute oral LD_{50} of polyoxin B [*19396-06-6*] and validamycin [*37248-47-8*] is reported to be more than 20 000 mg/kg for rats and mice. The acute dermal LD_{50} is also higher than 20 000 mg/kg. In a chronic feeding experiment, the NOEL for polyoxin B was more than 2400 mg/kg/d for rats and > 6800 mg/kg/d for mice [173], [174]. To prevent the development of microbial resistance, antibiotics that are used in human and veterinary medicine should not be employed as agricultural fungicides.

10.3. Systemic Fungicides

10.3.1. Organophosphates

Organophosphates form a complex with cholinesterase that is analogous to the complex formed between the enzyme and its substrate acetylcholine. Enzymatic cleavage of acetylcholine requires only microseconds, and the enzyme is available again almost immediately. However, organic phosphates bind to cholinesterase more strongly than acetylcholine. Depending on the active compound, cleavage may require several days to several weeks.

The inhibitor-enzyme complex can be reactivated by nucleophilic substances within a few hours of its formation; pralidoxime iodide [*94-63-3*] (2-PAM) and obidoxime chloride [*114-90-9*] are the best known antidotes [175], [176]. Atropine, which must be applied in advance, does not affect this blockage but acts on the excess acetylcholine available after degradation of the latter has been prevented [177], [178].

Edifenphos [*17109-49-8*]

LD_{50} (rat, oral)	100 – 340	mg/kg	[179]
LD_{50} (mouse oral)	214	mg/kg	[179]
LD_{50} (guinea pig, oral)	350	mg/kg	[179]
LD_{50} (rabbit, oral)	350	mg/kg	[179]
Nonmutagenic [179]			
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[179]
Mouse (2-year feed):	1.53	$mg\,kg^{-1}\,d^{-1}$	[179]
Dog (2-year feed):	0.58	$mg\,kg^{-1}\,d^{-1}$	[179]
ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [179]			

Iprobenfos [*26087-47-8*]

LD_{50} (rat, oral)	490	mg/kg	[154, p. 602]
LD_{50} (mouse oral)	1760 – 1830	mg/kg	[154, p. 602]
LD_{50} (mouse, dermal)	4000	mg/kg	[154, p. 602]
LC_{50} (rat, 4-h inhalation)	0.34 – 1.12	mg/L air	[154, p. 602]
NOEL/NOAEL:			
Chronic.			
Rat (2-year feed):	0.036 – 0.45	$mg\,kg^{-1}\,d^{-1}$	[154, p. 602]

Pyrazophos [*13457-18-6*]

LD_{50} (rat, oral)	151 – 778	mg/kg	[156, p. 363]
LD_{50} (mouse, oral)	205 – 413	mg/kg	[156, p. 363]
LD_{50} (dog, oral)	> 100	mg/kg	[156, p. 363]
LD_{50} (rat, dermal)	> 2000	mg/kg	[156, p. 363]
LD_{50} (mouse, dermal)	4000	mg/kg	[156, p. 363]
LC_{50} (rat, 4-h inhalation)	> 1.22	mg/L air	[156, p. 363]
Rabbit: nonirritating to skin [156, p. 363]			
Rabit, slightly irritating to eyes [156, p. 363]			
Guinea pig: not a skin sensitizer [156, p. 363]			
Nonmutagenic [156, p. 363]			
NOEL/NOAEL:			
Reproduction.			
Rat (2-generation feed):	1	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[156, p. 363]
Teratogenicity.			
Rat (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 363]
Rabbit (gavage):	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 363]
Chronic.			
Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[156, p. 363]
Dog (2-year feed):	0.4	$mg\,kg^{-1}\,d^{-1}$	[156, p. 363]
ADI: $0.004\,mg\,kg^{-1}\,d^{-1}$ [156, p. 363]			

10.3.2. Benzimidazoles

Benzimidazoles have moderate to low toxicity. They range from relatively simple compounds to very large molecules with one or more rings containing a nitrogen or a sulfur atom.

In experiments on rats, more than 90 % of a single dose of benomyl or thiophanate is excreted within 72 h via the feces and urine. Methyl-2-benzimidazole carbamate is the primary metabolite of benzimidazoles. They are metabolized by the microsomal hepatic enzymes, as is reflected in liver findings in chronic experiments. No accumulation occurs.

The active fungicide principle is carbendazim.

Benomyl [*17804-35-2*]

LD_{50} (rat, oral)	> 10 000	mg/kg	[162, p. 7]
LD_{50} (rabbit, oral)	> 10 000	mg/kg	[162, p. 7]
LD_{50} (rabbit, dermal)	> 10 000	mg/kg	[162, p. 7]
LC_{50} (rat, 4-h inhalation)	> 4.01	mg/L air	[162, p. 7]

Rabbit: irritating to skin and eyes [162, p. 7]
Guinea pig: a skin sensitizer [162, p. 7]
Nonmutagenic [162, p. 7]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	37	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[160, p. 34]

Teratogenicity.

Rat (gavage):	80	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[160, p. 34]
Rabbit (gavage):	90	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[160, p. 34]

Chronic.

Rat (2-year feed):	109	$mg\,kg^{-1}\,d^{-1}$	[160, p. 34]
Mouse (2-year feed):	< 64	$mg\,kg^{-1}\,d^{-1}$	[160, p. 34]
Dog (2-year feed):	13	$mg\,kg^{-1}\,d^{-1}$	[160, p. 34]

ADI: $0.1\,mg\,kg^{-1}\,d^{-1}$ [160, p. 34]

Carbendazim [*10605-21-7*]

LD_{50} (rat, oral)	$> 10\,000$	mg/kg	[160, p. 54]
LD_{50} (rat, dermal)	> 2000	mg/kg	[160, p. 54]
LC_{50} (rat, 4-h inhalation)	> 5.9	mg/L air	[160, p. 54]

Rabbit: irritating to skin and eyes [160, p. 54]
Guinea pig: not a skin sensitizer [160, p. 54]
Nonmutagenic [160, p. 54]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	50	$mg\,kg^{-1}\,d^{-1}$ (parental)	[160, p. 54]
	120	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 54]

Teratogenicity.

Rat (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 54]
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 54]
Rabbit (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 54]
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 54]

Chronic.

Rat (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[160, p. 54]
Mouse (96-week feed):	34	$mg\,kg^{-1}\,d^{-1}$	[160, p. 54]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[160, p. 54]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [160, p. 54]

Fuberidazole [*3878-19-1*]

LD_{50} (rat, oral)	~ 500	mg/kg	[154, p. 533]
LD_{50} (mouse, oral)	825	mg/kg	[154, p. 533]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 533]
LC_{50} (rat, 4-h inhalation)	> 0.3	mg/L air	[154, p. 533]

Rabbit: nonirritating to skin and eyes [154, p. 533]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	4	$mg\,kg^{-1}\,d^{-1}$	[154, p. 533]
Dog (2-year feed):	0.25	$mg\,kg^{-1}\,d^{-1}$	[154, p. 533]

ADI: $0.0025\,mg\,kg^{-1}\,d^{-1}$ [154, p. 533]

Thiabendazole [*148-79-8*]

LD_{50} (mouse, oral)	3600	mg/kg	[180]
LD_{50} (rabbit, oral)	2400	mg/kg	[180]
LD_{50} (rat, oral)	3850	mg/kg	[180]

Rabbit: nonirritating to skin and eyes [180]
NOEL/NOAEL:
Reproduction.

Rat (3-generation gavage):	40	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[180]
Mouse (5-generation feed):	30	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[180]

Chronic.

Rat (2-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[180]
Dog (2-year gavage):	20	$mg\,kg^{-1}\,d^{-1}$	[180]

ADI: $0.05\,mg\,kg^{-1}\,d^{-1}$ [180]

Thiophanate methyl [*23564-05-8*]

LD_{50} (rat, oral)	7000	mg/kg	[160, p. 185]
LD_{50} (mouse, oral)	3510	mg/kg	[154, p. 987]
LD_{50} (rabbit, oral)	2270	mg/kg	[154, p. 987]
LD_{50} (oral, dermal)	$> 10\,000$	mg/kg	[160, p. 185]
LC_{50} (rat, 4-h inhalation)	1.8	mg/L air	[160, p. 185]

Rabbit: nonirritating to skin [160, p. 185]
Rabbit: mildly irritating to eyes [160, p. 185]
Guinea pig: a skin sensitizer [160, p. 185]
Nonmutagenic [160, p. 185]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	8	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[160, p. 185]

Teratogenicity.

Rat (gavage):	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[160, p. 185]
	> 1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[160, p. 185]
Rabbit (gavage):	2	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmetal)	[160, p. 185]

Chronic.

Rat (2-year feed):	9	$mg\,kg^{-1}\,d^{-1}$	[160, p. 185]
Mouse (2-year feed):	29	$mg\,kg^{-1}\,d^{-1}$	[160, p. 185]

Dog (2-year gavage):	10	$mg\,kg^{-1}\,d^{-1}$	[160, p. 185]

ADI: 0.02 $mg\,kg^{-1}\,d^{-1}$ [160, p. 185]

10.3.3. Carboxanilides

Carboxin [*5234-68-4*]

LD_{50} (rat, oral)	3820	mg/kg	[154, p. 155]
LD_{50} (rabbit, dermal)	> 4000	mg/kg	[154, p. 155]
LC_{50} (rat, one-h inhalation)	> 20	mg/L air	[154, p. 155]

Rabbit: nonirritating to eyes [154, p. 155]
Guinea pig: not a skin sensitizer [154]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	30	$mg\,kg^{-1}\,d^{-1}$	[154, p. 155]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [154, p. 155]

Flutolanil [*66332-96-5*]

LD_{50} (rat, oral)	> 10 000	mg/kg	[154, p. 512]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 512]
LC_{50} (rat, 4-h inhalation)	> 5.98	mg/L air	[154, p. 512]

Rabbit: nonirritating to skin and eyes [154, p. 512]

Mepronil [*55814-41-0*]

LD_{50} (rat, oral)	> 10 000	mg/kg	[154, p. 656]
LD_{50} (mouse, oral)	> 10 000	mg/kg	[154, p. 656]
LD_{50} (rat, dermal)	> 10 000	mg/kg	[154, p. 656]
LD_{50} (rabbit, dermal)	> 10 000	mg/kg	[154, p. 656]
LC_{50} (rat, 6-h inhalation)	> 1.32	mg/L air	[154, p. 656]

Rabbit: nonirritating to skin and eyes [154, p. 656]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	0.3	$mg\,kg^{-1}\,d^{-1}$	[154, p. 656]

Thifluzamide [*130000-40-7*]

LD_{50} (rat, oral)	> 5000	mg/kg	[154, p. 978]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[154, p. 978]

Rabbit: moderately irritating to eyes [154, p. 978]
Rabbit: slightly irritating to skin [154, p. 978]

10.3.4. Phenylamides

Ofurace [*58810-48-3*]

LD_{50} (rat, oral)	2600 – 3500	mg/kg	[154, p. 745]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[154, p. 745]

Rabbit: severely irritating to eyes [154, p. 745]
Rabbit: mildly irritating to skin [154, p. 745]
Guinea pig: not a skin sensitizer [154, p. 745]
Nonmutagenic (Ames test) [154, p. 745]

Oxadixyl [*77732-09-3*]

LD_{50} (rat, oral)	1860 – 3480	mg/kg	[154, p. 755]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 755]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 755]
LC_{50} (rat, 6-h inhalation)	> 5.6	mg/L air	[154, p. 755]

Rabbit: nonirritating to skin and eyes [154, p. 755]
Guinea pig: not a skin sensitizer [154, p. 755]
Nonmutagenic [181]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	22	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[181]

Teratogenicity.

Rat (gavage):	250	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[181]
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[181]
Rabbit (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[181]
	> 200	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[181]

Chronic.

Rat (2-year feed):	10.9	$mg\,kg^{-1}\,d^{-1}$	[181]
Mouse (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[181]
Dog (2-year feed):	> 12.5	$mg\,kg^{-1}\,d^{-1}$	[181]

ADI: 0.11 $mg\,kg^{-1}\,d^{-1}$ [181]

Benalaxyl [*71626-11-4*]

LD_{50} (rat, oral)	4200	mg/kg	[169, p. 9]
LD_{50} (mouse, oral)	680	mg/kg	[169, p. 9]
LD_{50} (rabbit, dermal)	2000	mg/kg	[169, p. 9]
LC_{50} (rat, 4-h inhalation)	> 10	mg/L air	[169, p. 9]

Rabbit: nonirritating to skin and eyes [169, p. 9]
Guinea pig: not a skin sensitizer [169, p. 9]
Nonmutagenic [169, p. 9]

NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[169, p. 9]

Teratogenicity.

Rat (gavage):	200	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[169, p. 9]
Rabbit (gavage):	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[169, p. 9]
	5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[169, p. 9]

Chronic.

Rat (2-year feed):	50	$mg\,kg^{-1}\,d^{-1}$	[169, p. 9]
Mouse (78-week feed):	40	$mg\,kg^{-1}\,d^{-1}$	[169, p. 9]

ADI: $0.05\,mg\,kg^{-1}\,d^{-1}$ [169, p. 9]

Furalaxyl [*57646-30-7*]

LD_{50} (rat, oral)	940	mg/kg	[154, p. 534]
LD_{50} (mouse, oral)	603	mg/kg	[154, p. 534]
LD_{50} (rat, dermal)	>3100	mg/kg	[154, p. 534]
LD_{50} (rabbit, dermal)	5508	mg/kg	[154, p. 534]

Rabbit: mildly irritating to skin and eyes [154, p. 534]
Guinea pig: not a skin sensitizer [154, p. 534]
NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	85	$mg\,kg^{-1}\,d^{-1}$	[154, p. 534]
Dog (90-day feed):	1.8	$mg\,kg^{-1}\,d^{-1}$	[154, p. 534]

ADI: $0.0009\,mg\,kg^{-1}\,d^{-1}$ [154, p. 534]

Metalaxyl [*57837-19-1*]

LD_{50} (rat, oral)	669	mg/kg	[182]
LD_{50} (mouse, oral)	788	mg/kg	[182]
LD_{50} (rabbit, oral)	697	mg/kg	[182]
LD_{50} (hamster, oral)	7170	mg/kg	[182]
LD_{50} (rat, dermal)	>3100	mg/kg	[182]
LD_{50} (rabbit, dermal)	>6000	mg/kg	[182]

Rabbit: mildly irritating to skin and eyes [182]
Guinea pig: not a skin sensitizer [182]
Nonmutagenic [183]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	62.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[182]

Teratogenicity.

Rat (gavage):	>120	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[182]
Rabbit (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[182]
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[182]

Chronic.

Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[182]
Mouse (2-year feed):	187.5	$mg\,kg^{-1}\,d^{-1}$	[182]
Dog (6-month feed):	7.4	$mg\,kg^{-1}\,d^{-1}$	[182]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [182]

Metalaxyl-M [*70630-17-0*]

LD_{50} (rat, oral)	667	mg/kg	[22, p. 274]
LD_{50} (rat, dermal)	>2000	mg/kg	[22, p. 274]
LC_{50} (rat, 4-h inhalation)	2.29	mg/L air	[22, p. 274]

Rabbit: nonirritating to skin [22, p. 274]
Rabbit: strong irritant to eyes [22, p. 274]
Guinea pig: not a skin sensitizer [22, p. 274]
Not mutagenic [22, p. 274]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[22, p. 274]
Mouse (18-month feed):	35.7	$mg\,kg^{-1}\,d^{-1}$	[22, p. 274]
Dog (one-year feed):	8.0	$mg\,kg^{-1}\,d^{-1}$	[22, p. 274]

ADI: $0.025\,mg\,kg^{-1}\,d^{-1}$ [22, p. 274]

10.3.5. Phosphites

Fosetyl [*39148-24-8*]

LD_{50} (rat, oral)	5800	mg/kg	[154, p. 530]
LD_{50} (mouse, oral)	3700	mg/kg	[154, p. 530]
LD_{50} (rat, dermal)	>3200	mg/kg	[154, p. 530]
LD_{50} (rabbit, dermal)	>2000	mg/kg	[154, p. 530]
LC_{50} (rat, 4-h inhalation)	>0.0017	mg/L air	[154, p. 530]

Rabbit: nonirritating to skin [154, p. 530]
NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	250	$mg\,kg^{-1}\,d^{-1}$	[154, p. 530]
Dog (90-day feed):	1250	$mg\,kg^{-1}\,d^{-1}$	[154, p. 530]

ADI: $3.0\,mg\,kg^{-1}\,d^{-1}$ [154, p. 530]

10.3.6. Inhibitors of Sterol Biosynthesis

The azole nitrogen of sterol biosynthesis inhibitors reacts with heme iron. These inhibitors have a high affinity for cytochrome P450. The synthesis of ergosterol is often inhibited at the C-14 demethylation step [183], which results in the accumulation of metabolite precursors. Some of these compounds have been reported to produce irritation of the gastrointestinal tract [184–193].

10.3.6.1. Triazoles

Azaconazole [*60207-31-0*]

LD_{50} (rat, oral)	308	mg/kg	[194]
LD_{50} (rabbit, oral)	324	mg/kg	[194]
LD_{50} (mouse, oral)	1123	mg/kg	[194]
LD_{50} (rat, dermal)	> 2560	mg/kg	[194]
LC_{50} (rat, 4-h inhalation)	1.22 – 1.84	mg/L air	[195]

Rabbit: nonirritating to eyes [194]
Rabbit: slightly irritating to skin [194]
Guinea pig: not a skin sensitizer [194]
Nonmutagenic [194]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	80	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[195]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[195]
	60	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[195]
Rabbit (gavage):	80	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[195]
	160	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[195]

Chronic.

Rat (2-year feed):	12.5	$mg\,kg^{-1}\,d^{-1}$	[195]
Dog (6-month feed):	10	$mg\,kg^{-1}\,d^{-1}$	[195]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [195]

Bitertanol [*55179-31-2*]

LD_{50} (rat, oral)	3300	mg/kg	[162, p. 47]
LD_{50} (mouse, oral)	4202	mg/kg	[162, p. 47]
LD_{50} (dog, oral)	> 5000	mg/kg	[162, p. 47]
LD_{50} (sheep, oral)	> 1000	mg/kg	[162, p. 46]
LD_{50} (rat, dermal)	> 5000	mg/kg	[162, p. 47]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[162, p. 47]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[162, p. 47]

Guinea pig: not a skin sensitizer [209]
Nonmutagenic [162, p. 47]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	∼ 12	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[162, p. 47]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[162, p. 47]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[162, p. 47]

Chronic.

Rat (2-year feed):	1	$mg\,kg^{-1}\,d^{-1}$	[169, p. 19]
Mouse (2-year feed):	3	$mg\,kg^{-1}\,d^{-1}$	[162, p. 47]
Dog (2-year capsule):	0.25	$mg\,kg^{-1}\,d^{-1}$	[169, p. 19]

ADI: 0.003 $mg\,kg^{-1}\,d^{-1}$ [169, p. 19]

Bromuconazole [*116255-48-2*]

LD_{50} (rat, oral)	365	mg/kg	[154, p. 124]
LD_{50} (mouse, oral)	151	mg/kg	[154, p. 124]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 124]
LC_{50} (rabbit, 4-h inhalation)	> 5	mg/L air	[154, p. 124]

Rabbit: nonirritating to skin and eyes [154, p. 124]
Guinea pig: not a skin sensitizer [154, p. 124]
Nonmutagenic [154, p. 124]

Cyproconazole [*94361-06-5*]

LD_{50} (rat, oral)	1020 – 1330	mg/kg	[196]
LD_{50} (mouse, oral)	200 – 218	mg/kg	[196]
LD_{50} (rabbit, oral)	460	mg/kg	[196]
LD_{50} (rat, dermal)	> 2000	mg/kg	[196]
LC_{50} (rat, 4-h inhalation)	> 5.65	mg/L air	[196]

Rabbit: nonirritating to skin and eyes [196]
Guinea pig: not a skin sensitizer [196]
Nonmutagenic [196]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1.65	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[196]

Teratogenicity.

Rat (gavage):	6	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[196]
	12	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[196]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[196]

Chronic.

Rat (118-121-week feed):	2.48	$mg\,kg^{-1}\,d^{-1}$	[196]
Mouse (81-88-week feed):	2.2	$mg\,kg^{-1}\,d^{-1}$	[196]
Dog (one-year feed):	1.0	$mg\,kg^{-1}\,d^{-1}$	[196]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [196]

Difenoconazole [*119446-68-3*]

LD_{50} (rat, oral)	1453	mg/kg	[197]
LD_{50} (mouse, oral)	> 2000	mg/kg	[197]
LD_{50} (rabbit, dermal)	> 2010	mg/kg	[197]
LC_{50} (rat, 4-h inhalation)	3.46	mg/L air	[197]

Rabbit: nonirritating to skin and eyes [197]
Guinea pig: not a skin sensitizer [197]
Nonmutagenic [197]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	12 – 28	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[197]

Teratogenicity.

Rat (gavage):	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[197]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[197]
Rabbit (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[197]

Chronic.

Rat (2-year feed):	1.0 – 1.3	$mg\,kg^{-1}\,d^{-1}$	[197]
Mouse (78-week feed):	4.7 – 5.6	$mg\,kg^{-1}\,d^{-1}$	[197]
Dog (one-year feed):	3.4 – 3.7	$mg\,kg^{-1}\,d^{-1}$	[197]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [197]

Diniconazole [*76714-88-0*]

LD_{50} (rat, oral)	639	mg/kg	[154, p. 356]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 356]
LC_{50} (rat, 4-h inhalation)	> 2.77	mg/L air	[154, p. 356]

Rabbit: nonirritating to skin [154, p. 356]
Rabbit: slightly irritating to eyes [154, p. 356]
Guinea pig: not a skin sensitizer [154, p. 356]

Epoxiconazole [*106325-08-0*]

LD_{50} (rat, oral)	> 5000	mg/kg	[198]
LD_{50} (rat, dermal)	> 2000	mg/kg	[198]
LC_{50} (rat, 4-h inhalation)	> 5.3	mg/L air	[198]

Rabbit: nonirritating to skin and eyes [198]
Guinea pig: not a skin sensitizer [198]
Nonmutagenic [198]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	2.3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[198]
Rat (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[198]
Rabbit (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[198]
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[198]

Chronic.

Rat (2-year feed):	1 – 2	$mg\,kg^{-1}\,d^{-1}$	[198]
Mouse (78-week feed):	0.74 – 1.01	$mg\,kg^{-1}\,d^{-1}$	[198]
Dog (one-year feed):	1.1	$mg\,kg^{-1}\,d^{-1}$	[198]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [198]

Fenbuconazole [*114369-43-6*]

LD_{50} (rat, oral)	> 2000	mg/kg	[154, p. 428]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 428]
LC_{50} (rat, 4-h inhalation)	2.1	mg/L air	[154, p. 428]

Rabbit: nonirritating to skin and eyes [154, p. 428]
Nonmutagenic [154, p. 428]
NOEL/NOAEL:
Subchronic.

Rat (3-month feed):	ca. 2	$mg\,kg^{-1}\,d^{-1}$	[155, p. 428]
Mouse (3-month feed):	ca. 8.6	$mg\,kg^{-1}\,d^{-1}$	[154, p. 428]
Dog (3-month feed):	ca. 5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 428]

Fluquinconazole [*136426-54-5*]

LD_{50} (rat, oral)	112	mg/kg	[154, p. 498]
LD_{50} (mouse, oral)	180 – 325	mg/kg	[154, p. 498]
LD_{50} (rat, dermal)	625	mg/kg	[154, p. 498]
LC_{50} (rat, 4-h inhalation)	0.75	mg/L air	[154, p. 498]

Rabbit: nonirritating to skin and eyes [154, p. 498]
Guinea pig: not a skin sensitizer [154, p. 498]
Nonmutagenic [154, p. 498]

Flusilazole [*85509-19-9*]

LD_{50} (rat, oral)	674	mg/kg	[199]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[199]
LC_{50} (rat, 4-h inhalation)	> 5.0	mg/L air	[199]

Rabbit: slightly irritating to skin and eyes [199]
Guinea pig: not a skin sensitizer [199]
Nonmutagenic [199]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	10	$mg\,kg^{-1}\,d^{-1}$ (parental)	[199]
	2	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[199]

Teratogenicity.

Rat (gavage):	2	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[199]
Rabbit (gavage):	12	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[199]

Chronic.

Rat (2-year feed):	0.46	$mg\,kg^{-1}\,d^{-1}$	[199]
Mouse (18-month feed):	3.4	$mg\,kg^{-1}\,d^{-1}$	[199]
Dog (one-year feed):	0.2	$mg\,kg^{-1}\,d^{-1}$	[199]

ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$ [199]

Flutriafol [*76674-21-0*]

LD_{50} (rat, oral)	1140	mg/kg	[154, p. 514]
LD_{50} (rat, dermal)	> 1000	mg/kg	[154, p. 514]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 514]
LC_{50} (rat, 4-h inhalation)	1.65	mg/L air	[154, p. 514]

Rabbit: nonirritating to skin [154, p. 514]
Rabbit: mildly irritating to eyes [154, p. 514]
Guinea pig: not a skin sensitizer [154, p. 514]
Nonmutagenic [154, p. 514]

NOEL/NOAEL:
Subchronic.

Rat (3-month feed):	ca. 2	$mg\,kg^{-1}\,d^{-1}$	[154, p. 514]
Dog (3-month feed):	ca. 0.3	$mg\,kg^{-1}\,d^{-1}$	[154, p. 514]

Hexaconazole [*79983-71-4*]

LD_{50} (rat, oral)	2189	mg/kg	[163, p. 81]
LD_{50} (mouse, oral)	>557	mg/kg	[163, p. 81]
LD_{50} (rat, dermal)	>2000	mg/kg	[163, p. 81]
LC_{50} (rat, 4-h inhalation)	>5.9	mg/L air	[163, p. 81]

Rabbit: nonirritating to skin [163, p. 81]
Rabbit: mildly irritating to eyes [163, p. 81]
Guinea pig: a skin sensitizer [163, p. 81]
Nonmutagenic [163, p. 81]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[163, p. 81]

Teratogenicity.

Rat (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[163, p. 81]
	2.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[163, p. 81]
Rabbit (gavage):	>50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[163, p. 81]
	12.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[163, p. 81]

Chronic.

Rat (105-week feed):	ca. 0.5	$mg\,kg^{-1}\,d^{-1}$	[163, p. 81]
Mouse (2-year feed):	4.7	$mg\,kg^{-1}\,d^{-1}$	[163, p. 81]
Dog (one-year feed):	2	$mg\,kg^{-1}\,d^{-1}$	[163, p. 81]

ADI: $0.005\,mg\,kg^{-1}\,d^{-1}$ [163, p. 81]

Metconazole [*125116-23-6*]

LD_{50} (rat, oral)	1459	mg/kg	[154, p. 699]
LD_{50} (rat, dermal)	>2000	mg/kg	[154, p. 699]
LC_{50} (rat, 4-h inhalation)	>5.6	mg/L air	[154, p. 699]

Rabbit: nonirritating to skin [154, p. 699]
Rabbit: slightly irritating to eyes [154, p. 699]
Guinea pig: not a skin sensitizer [154, p. 699]

Myclobutanil [*88671-89-0*]

LD_{50} (rat, oral)	1600	mg/kg	[156, p. 267]
LD_{50} (rat, dermal)	>5000	mg/kg	[156, p. 267]

Rabbit: nonirritating to skin [156, p. 267]
Rabbit: mildly irritating to eyes [156, p. 267]
Guinea pig: not a skin sensitizer [156, p. 267]
Nonmutagenic [156, p. 267]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[156, p. 267]
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[156, p. 267]

Teratogenicity.

Rat (gavage):	93.8	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 267]
Rabbit (gavage):	60	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 267]

Chronic.

Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[156, p. 267]
Mouse (2-year feed):	2.7	$mg\,kg^{-1}\,d^{-1}$	[156, p. 267]
Dog (one-year feed):	3.1	$mg\,kg^{-1}\,d^{-1}$	[156, p. 267]

ADI: $0.025\,mg\,kg^{-1}\,d^{-1}$ [156, p. 267]

Penconazole [*66246-88-6*]

LD_{50} (rat, oral)	2125	mg/kg	[156, p. 291]
LD_{50} (mouse, oral)	2444	mg/kg	[156, p. 291]
LD_{50} (hamster, oral)	>5000	mg/kg	[156, p. 291]
LD_{50} (rabbit, oral)	971	mg/kg	[156, p. 291]
LD_{50} (rat, dermal)	>3000	mg/kg	[156, p. 291]
LC_{50} (rat, 4-h inhalation)	>4.05	mg/L air	[156, p. 291]

Rabbit: slightly irritating to skin and eyes [156, p. 291]
Guinea pig: not a skin sensitizer [156, p. 291]
Nonmutagenic [156, p. 291]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[156, p. 291]

Teratogenicity.

Rat (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 291]
Rabbit (gavage):	75	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[156, p. 291]

Chronic.

Rat (2-year feed):	3.8	$mg\,kg^{-1}\,d^{-1}$	[156, p. 291]
Mouse (2-year feed):	17	$mg\,kg^{-1}\,d^{-1}$	[156, p. 291]
Dog (one-year feed):	3	$mg\,kg^{-1}\,d^{-1}$	[156, p. 291]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [156, p. 291]

Propiconazole [*60207-90-1*]

LD_{50} (rat, oral)	1517	mg/kg	[169, p. 111]
LD_{50} (mouse, oral)	1490	mg/kg	[169, p. 111]
LD_{50} (rabbit, oral)	1344	mg/kg	[169, p. 111]
LD_{50} (rat, dermal)	>4000	mg/kg	[169, p. 111]
LD_{50} (rabbit, dermal)	>6000	mg/kg	[169, p. 111]

Rabbit: slightly irritating to skin and eyes [169, p. 111]
Guinea pig: not a skin sensitizer [169, p. 111]
Nonmutagenic [200]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[201]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[201]

Teratogenicity.

Rat (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[200]
Rabbit (gavage):	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[200]
	>400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[200]

Chronic.

Rat (2-year feed):	5	$mg\,kg^{-1}\,d^{-1}$	[200]
Mouse (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[200]
Dog (one-year feed):	1.25	$mg\,kg^{-1}\,d^{-1}$	[200]

ADI: $0.013\,mg\,kg^{-1}\,d^{-1}$ [200]

Triadimefon [*43121-43-3*]

LD_{50} (rat, oral)	363	mg/kg	[204]
LD_{50} (hen, oral)	5000	mg/kg	[205]
LD_{50} (quail, oral)	1720	mg/kg	[205]
LD_{50} (mouse, oral)	1071	mg/kg	[204]
LD_{50} (dog, oral)	500	mg/kg	[205]
LD_{50} (rat, dermal)	>1000	mg/kg	[204]
LC_{50} (rat, 4-h inhalation)	0.29	mg/L air	[204]

Rabbit: mildly irritating to skin [204]
Rabbit: nonirritating to eyes [204]
Guinea pig: not a skin sensitizer [204]
Nonmutagenic [204]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	2.5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[206]

Teratogenicity.

Rat (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[204]
Rabbit (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[204]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[204]

Chronic.

Rat (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[162, p. 447]
Mouse (2-year feed):	40	$mg\,kg^{-1}\,d^{-1}$	[162, p. 447]
Dog (2-year feed):	8.3	$mg\,kg^{-1}\,d^{-1}$	[205]

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$ [206]

Tebuconazole [*80443-41-0*]

LD_{50} (rat, oral)	>5000	mg/kg	[202]
LD_{50} (mouse, oral)	1615	mg/kg	[202]
LD_{50} (rabbit, oral)	>1000	mg/kg	[202]
LD_{50} (sheep, oral)	625 – 1250	mg/kg	[202]
LD_{50} (dog, oral)	625 – 1250	mg/kg	[202]
LD_{50} (rat, dermal)	>5000	mg/kg	[202]
LC_{50} (rat, 4-h inhalation)	0.82	mg/L air	[202]

Rabbit: nonirritating to skin and eyes [202]
Guinea pig: not a skin sensitizer [203]
Nonmutagenic [202]
NOEL/NOAEL:
Reproduction.

Rat (2-generation):	15	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[203]

Teratogenicity.

Rat (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[203]
Rabbit (gavage):	30	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[203]

Chronic.

Rat (2-year feed):	7.4	$mg\,kg^{-1}\,d^{-1}$	[203]
Mouse (21-month feed):	5.9	$mg\,kg^{-1}\,d^{-1}$	[203]
Dog (one-year feed):	1	$mg\,kg^{-1}\,d^{-1}$	[203]

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [203]

Triadimenol [*55219-65-3*]

LD_{50} (rat, oral)	>700	mg/kg	[165, p. 215]
LD_{50} (rat, dermal)	>5000	mg/kg	[165, p. 215]
LC_{50} (rat, 4-h inhalation)	>1.56	mg/L air	[165, p. 215]

Rabbit: nonirritating to skin and eyes [165, p. 215]
Guinea pig: not a skin sensitizer [165, p. 215]
Nonmutagenic [165, p. 215]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[207]

Teratogenicity.

Rat (gavage):	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[207]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[207]
Rabbit (gavage):	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[207]
	125	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[207]

Chronic.

Rat (2-year feed):	6.25	$mg\,kg^{-1}\,d^{-1}$	[207]
Mouse (2-year feed):	18	$mg\,kg^{-1}\,d^{-1}$	[207]
Dog (2-year feed):	3.75	$mg\,kg^{-1}\,d^{-1}$	[207]

ADI: $0.038\,mg\,kg^{-1}\,d^{-1}$ [207]

10.3.6.2. Imidazoles

Imazalil [*35554-44-0*]

LD_{50} (rat, oral)	227 – 343	mg/kg	[154, p. 580]
LD_{50} (dog, oral)	> 640	mg/kg	[154, p. 580]
LD_{50} (rat, dermal)	4200 – 4880	mg/kg	[154, p. 580]
LC_{50} (rat, 4-h inhalation)	16	mg/L air	[158, p. 283]

Nonmutagenic [158, p. 283]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	40	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[155, p. 244]

Teratogenicity.

Mouse (gavage):	40	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[208]
Rabbit (gavage):	2.5	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[208]
Rat (30-month feed):	5	$mg\,kg^{-1}\,d^{-1}$	[198, p. 99]
Mouse (18-month drinking water):	40	$mg\,kg^{-1}\,d^{-1}$	[155, p. 244]
Dog (one-year capsule):	2.5	$mg\,kg^{-1}\,d^{-1}$	[150, p. 283]

ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [150, p. 283]

Pefurazoate [*101903-30-4*]

LD_{50} (rat, oral)	981 – 1051	mg/kg	[154, p. 774]
LD_{50} (mouse, oral)	946 – 1299	mg/kg	[154, p. 744]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 744]
LC_{50} (rat, 4-h inhalation)	3.45	mg/L air	[154, p. 744]

Rabbit: nonirritating to skin [154, p. 744]
Rabbit: slightly irritating to eyes [154, p. 744]
Guinea pig: not a skin sensitizer [154, p. 744]
Nonteratogenic in rat and rabbit [154, p. 744]
NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 774]

Prochloraz [*67747-09-5*]

LD_{50} (rat, oral)	1600 – 2400	mg/kg	[162, p. 365]
LD_{50} (mouse, oral)	2400	mg/kg	[162, p. 365]
LD_{50} (rat, dermal)	5000	mg/kg	[162, p. 365]
LD_{50} (rabbit, dermal)	3000	mg/kg	[162, p. 365]
LC_{50} (rat, 6-h inhalation)	> 0.42	mg/L air	[162, p. 365]

Rabbit: irritating to skin and eyes [162, p. 365]
Nonmutagenic [162, p. 365]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1.3	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[162, p. 365]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[162, p. 365]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[162, p. 365]
Rabbit (gavage):	12	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[162, p. 365]
	48	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[162, p. 365]

Chronic.

Rat (111-115-week feed):	1.3	$mg\,kg^{-1}\,d^{-1}$	[162, p. 365]
Mouse (106-week feed):	7.5	$mg\,kg^{-1}\,d^{-1}$	[162, p. 365]
Dog (2-year feed):	0.92	$mg\,kg^{-1}\,d^{-1}$	[162, p. 365]

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [162, p. 365]

Triflumizole [*68694-11-1*]

LD_{50} (rat, oral)	695 – 715	mg/kg	[155, p. 1023]
LD_{50} (rat, dermal)	> 5000	mg/kg	[155, p. 1023]
LC_{50} (rat, 4-h inhalation)	> 3.2	mg/L air	[155, p. 1023]

Nonmutagenic [155, p. 1023]

10.3.6.3. Pyridines

Pyrifenox [*88283-41-4*]

LD_{50} (rat, oral)	2912	mg/kg	[154, p. 884]
LD_{50} (mouse, oral)	> 2000	mg/kg	[154, p. 884]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 884]
LC_{50} (rat, 4-h inhalation)	> 2.05	mg/L air	[154, p. 884]

Rabbit: slightly irritating to skin [154, p. 884]
Rabbit: nonirritating to eyes [154, p. 884]
Guinea pig: not a skin sensitizer [154, p. 884]
Nonmutagenic [154, p. 884]

10.3.6.4. Pyrimidines

Cyprodinil [*121552-61-2*]

LD_{50} (rat, oral)	> 2796	mg/kg	[209]
LD_{50} (rat, dermal)	> 2000	mg/kg	[209]
LC_{50} (rat, 4-h inhalation)	> 1.2	mg/L air	[209]

Rabbit: minimally irritating to eyes [209]
Rabbit: slightly irritating to skin [209]
Guinea pig: a weak skin sensitizer [209]
Nonmutagenic [209]

NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	5	mg kg^{-1} d^{-1} (parental)	[209]
	50	mg kg^{-1} d^{-1} (developmental)	[209]

Teratogenicity.

Rat (gavage):	200	mg kg^{-1} d^{-1} (maternal and developmental)	[209]
Rabbit (gavage):	150	mg kg^{-1} d^{-1} (maternal)	[209]
	400	mg kg^{-1} d^{-1} (developmental)	[209]

Chronic.

Rat (2-year feed):	3.75	mg kg^{-1} d^{-1}	[209]
Mouse (18-month feed):	300	mg kg^{-1} d^{-1}	[209]
Dog (one-year feed):	62.5	mg kg^{-1} d^{-1}	[209]

ADI: 0.0375 mg kg^{-1} d^{-1} [209]

Dimethirimol [*5221-53-4*]

LD_{50} (rat, oral)	2350	mg/kg	[154, p. 347]
LD_{50} (mouse, oral)	800 – 1600	mg/kg	[154, p. 347]
LD_{50} (guinea pig, oral)	500	mg/kg	[154, p. 347]
LD_{50} (rat, dermal)	> 400	mg/kg	[154, p. 347]

Rabbit: nonirritating to skin and eyes [154, p. 347]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	15	mg kg^{-1} d^{-1}	[154, p. 347]
Dog (2-year feed):	0.63	mg kg^{-1} d^{-1}	[154, p. 347]

Fenarimol [*60168-88-9*]

LD_{50} (rat, oral)	2550	mg/kg	[160, p. 63]
LD_{50} (mouse, oral)	4500	mg/kg	[154, p. 425]
LD_{50} (dog, oral)	> 200	mg/kg	[154, p. 425]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[160, p. 63]
LC_{50} (rat, one-h inhalation)	> 2.04	mg/L air	[160, p. 63]

Rabbit: nonirritating to skin and eyes [160, p. 63]
Guinea pig: not a skin sensitizer [160, p. 63]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	1.3	mg kg^{-1} d^{-1} (parental and developmental)	[160, p. 63]

Teratogenicity.

Rat (gavage):	35	mg kg^{-1} d^{-1} (maternal)	[160, p. 63]
	13	mg kg^{-1} d^{-1} (developmental)	[160, p. 63]
Rabbit (gavage):	50	mg kg^{-1} d^{-1} (maternal and developmental)	[160, p. 63]

Chronic.

Rat (2-year feed):	1.3	mg kg^{-1} d^{-1}	[160, p. 63]
Mouse (2-year feed):	24	mg kg^{-1} d^{-1}	[160, p. 63]

ADI: 0.01 mg kg^{-1} d^{-1} [160, p. 63]

Nuarimol [*63284-71-9*]

LD_{50} (rat, oral)	1250 – 2500	mg/kg	[154, p. 742]
LD_{50} (mouse, oral)	2500 – 3000	mg/kg	[154, p. 742]
LD_{50} (dog, oral)	500	mg/kg	[154, p. 742]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 742]
LC_{50} (rat, one-h inhalation)	> 0.37	mg/L air	[154, p. 742]

Rabbit: nonirritating to skin [154, p. 742]
Rabbit: slightly irritating to eyes [154, p. 742]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	2.5	mg kg^{-1} d^{-1}	[154, p. 742]
Mouse (2-year feed):	7	mg kg^{-1} d^{-1}	[154, p. 742]

Pyrimethanil [*53112-28-0*]

LD_{50} (rat, oral)	4150 – 5971	mg/kg	[154, p. 885]
LD_{50} (mouse, oral)	4665 – 5359	mg/kg	[154, p. 885]
LD_{50} (rat, dermal)	> 5000	mg/kg	[154, p. 885]
LC_{50} (rat, 4-h inhalation)	>> 1.98	mg/L air	[154, p. 885]

Rabbit: nonirritating to skin and eyes [154, p. 885]
Guinea pig: not a skin sensitizer [154, p. 885]
Nonmutagenic [154, p. 885]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	20	mg kg^{-1} d^{-1}	[154, p. 885]

ADI: 0.2 mg kg^{-1} d^{-1} [154, p. 885]

10.3.6.5. Piperazines

Triforine [*26644-46-2*]

LD_{50} (rat, oral)	> 16 000	mg/kg	[210]
LD_{50} (mouse, oral)	> 6000	mg/kg	[210]
LD_{50} (dog, oral)	> 2000	mg/kg	[210]
LD_{50} (quail, oral)	> 6000	mg/kg	[210]
LD_{50} (rat, dermal)	> 10 000	mg/kg	[210]
LC_{50} (rat, one-h inhalation)	> 4.5	mg/L air	[210]

Rabbit: nonirritating to skin and eyes [210]
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed):	25	mg kg^{-1} d^{-1} (parental)	[210]
	125	mg kg^{-1} d^{-1} (developmental)	[210]

Teratogenicity.

Rat (gavage):	800	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[210]
	400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[210]

Chronic.

Rat (2-year feed):	30	$mg\,kg^{-1}\,d^{-1}$	[210]
Dog (2-year feed):	2.5	$mg\,kg^{-1}\,d^{-1}$	[210]

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$ [210]

10.3.6.6. Morpholines

Dimethomorph [*110488-70-5*]

LD_{50} (rat, oral)	3900	mg/kg	[154, p. 351]
LD_{50} (mouse, oral)	3700	mg/kg	[154, p. 351]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 351]
LC_{50} (rat, 4-h inhalation)	> 4.2	mg/L air	[154, p. 351]

Rabbit: nonirritating to skin and eyes [154, p. 351]

Dodemorph acetate [*31717-83-0*]

LD_{50} (rat, oral)	2465 – 3944	mg/kg	[154, p. 380]
LD_{50} (rat, dermal)	> 4000	mg/kg	[154, p. 380]
LC_{50} (rat, 4-h inhalation)	5	mg/L air	[154, p. 380]

Rabbit: severely irritating to skin and eyes [154, p. 380]

Fenpropimorph [*67564-91-4*]

LD_{50} (rat, oral)	3515	mg/kg	[172, p. 77]
LD_{50} (rat, dermal)	4291	mg/kg	[172, p. 77]
LC_{50} (rat, 4-h inhalation)	> 3.6	mg/L air	[172, p. 77]

Rabbit: severely irritating to skin [172, p. 77]
Rabbit: irritating to eyes [172, p. 77]
Guinea pig: not a skin sensitizer [172, p. 77]
Nonmutagenic [172, p. 77]
NOEL/NOAEL:
Reproduction.

Rat (2-generation feed):	0.62	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[172, p. 77]

Teratogenicity.

Rat (gavage):	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[172, p. 77]
	40	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[172, p. 77]
Rabbit (gavage):	1.5	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[172, p. 77]

Chronic.

Rat (2-year feed):	0.3	$mg\,kg^{-1}\,d^{-1}$	[172, p. 77]
Mouse (95-week feed):	3	$mg\,kg^{-1}\,d^{-1}$	[172, p. 77]
Dog (one-year feed):	3.2	$mg\,kg^{-1}\,d^{-1}$	[172, p. 77]

ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$ [172, p. 77]

Tridemorph [*24602-86-6*]

LD_{50} (rat, oral)	400	mg/kg	[154, p. 1019]
LD_{50} (rat, dermal)	> 4000	mg/kg	[154, p. 1019]
LC_{50} (rat, 4-h inhalation)	4.5	mg/L air	[154, p. 1019]

Rabbit: severely irritating to skin and eyes [154, p. 1019]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	1.8	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1019]
Mice (2-year feed):	> 13	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1019]

Fenpropidin [*67306-00-7*]

LD_{50} (rat, oral)	> 1400	mg/kg	[154, p. 447]
LD_{50} (rat, dermal)	> 1800	mg/kg	[154, p. 447]
LC_{50} (rat, 4-h inhalation)	> 1.22	mg/L air	[154, p. 447]

Rabbit: irritating to skin and eyes [154, p. 447]
Guinea pig: not a skin sensitizer [154, p. 447]
Nonmutagenic [154, p. 447]
Nonteratogenic [154, p. 447]
ADI: $0.005\,mg\,kg^{-1}\,d^{-1}$ [154, p. 447]

10.3.6.7. Strobilurins

Azoxystrobulin [*131860-33-8*]

LD_{50} (rat, oral)	> 5000	mg/kg	[22, p. 70]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 70]

Rabbit: slight skin and eye irritant [22, p. 70]
Guinea pig: not a skin sensitizer [22, p. 70]
Nonmutagenic [22, p. 70]
ADI: $0.2\,mg\,kg^{-1}\,d^{-1}$ [22, p. 70]

Kresoxim-methyl [*143390-89-0*]

LD_{50} (rat, oral)	> 5000	mg/kg	[22, p. 743]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 743]
LC_{50} (rat, 4-h inhalation)	> 5.6	mg/L air	[22, p. 743]

Nonmutagenic [22, p. 743]
ADI: $0.4\,mg\,kg^{-1}\,d^{-1}$ [22, p. 743]

Metominostrobin [*133408-50-1*]

LD_{50} (rat, oral)	708 – 776	mg/kg	[22, p. 1147]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 1147]
LC_{50} (rat, 4-h inhalation)	> 1.88	mg/L air	[22, p. 1147]

Rabbit: not a skin irritant [22, p. 1147]

10.3.7. Miscellaneous

Bupirimate [*41483-43-6*]

LD_{50} (rat, oral)	> 4000	mg/kg	[154, p. 127]
LD_{50} (mouse, oral)	> 4000	mg/kg	[154, p. 127]
LD_{50} (rabbit, oral)	> 4000	mg/kg	[154, p. 127]
LD_{50} (guinea pig, oral)	> 4000	mg/kg	[154, p. 127]
LD_{50} (rat, dermal)	4800	mg/kg	[154, p. 127]
LC_{50} (rat, 4-h inhalation)	>> 0.04	mg/L air	[154, p. 127]

Rabbit: mildly irritating to skin and eyes [154, p. 127]
Guinea pig: a moderate skin sensitizer [154, p. 127]
NOEL/NOAEL:
Subchronic.

Rat (90-day feed):	50	$mg\,kg^{-1}\,d^{-1}$	[154, p. 127]
Dog (90-day feed):	15	$mg\,kg^{-1}\,d^{-1}$	[154, p. 127]

Chronic.

Rat (2-year feed):	5	$mg\,kg^{-1}\,d^{-1}$	[154, p. 127]

Ethirimol [*23947-60-6*]

LD_{50} (rat, oral)	6340	mg/kg	[154, p. 409]
LD_{50} (mouse, oral)	4000	mg/kg	[154, p. 409]
LD_{50} (guinea pig, oral)	500 – 1000	mg/kg	[154, p. 409]
LD_{50} (rabbit, oral)	1000 – 2000	mg/kg	[154, p. 409]
LD_{50} (rat, dermal)	> 2000	mg/kg	[154, p. 409]
LC_{50} (rat, 4-h inhalation)	> 4.92	mg/L air	[154, p. 409]

Rabbit: nonirritating to skin [154, p. 409]
Rabbit: mildly irritating to eyes [154, p. 409]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[154, p. 409]
Dog (2-year feed):	30	$mg\,kg^{-1}\,d^{-1}$	[154, p. 409]

Hymexazol [*10004-44-1*]

LD_{50} (rat, oral)	3909 – 4678	mg/kg	[154, p. 576]
LD_{50} (mouse, oral)	1968 – 2148	mg/kg	[154, p. 576]
LD_{50} (rat, dermal)	> 10 000	mg/kg	[154, p. 576]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 576]

Rabbit: nonirritating to skin [154, p. 576]
Rabbit: irritating to eyes [154, p. 576]
Nonmutagenic [154, p. 576]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	19 – 20	$mg\,kg^{-1}\,d^{-1}$	[154, p. 576]
Dog (2-year feed):	15	$mg\,kg^{-1}\,d^{-1}$	[154, p. 576]

Isoprothiolane [*50512-35-1*]

LD_{50} (rat, oral)	1190	mg/kg	[154, p. 610]
LD_{50} (mouse, oral)	1340	mg/kg	[154, p. 610]
LD_{50} (rat, dermal)	> 10 250	mg/kg	[154, p. 610]

Rabbit: nonirritating to skin and eyes [154, p. 610]

Propamocarb [*24579-73-5*]

LD_{50} (rat, oral)	2000 – 8550	mg/kg	[154, p. 843]
LD_{50} (mouse, oral)	1960 – 2800	mg/kg	[154, p. 843]
LD_{50} (dog, oral)	1450	mg/kg	[154, p. 843]
LD_{50} (rabbit, dermal)	> 3920	mg/kg	[154, p. 843]
LC_{50} (rat, 4-h inhalation)	0.0057	mg/L air	[154, p. 843]

NOEL/NOAEL:
Chronic.

Rat (2-year feed):	10	$mg\,kg^{-1}\,d^{-1}$	[155, p. 179]
Dog (2-year feed):	25	$mg\,kg^{-1}\,d^{-1}$	[155, p. 179]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [155, p. 179]

Pyroquilon [*57369-32-1*]

LD_{50} (rat, oral)	321	mg/kg	[154, p. 889]
LD_{50} (mouse, oral)	581	mg/kg	[154, p. 889]
LD_{50} (rat, dermal)	> 3100	mg/kg	[154, p. 889]
LC_{50} (rat, 4-h inhalation)	5.1	mg/L air	[154, p. 889]

Rabbit: mildly irritating to skin [154, p. 889]
Rabbit: slightly irritating to eyes [154, p. 889]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	30	$mg\,kg^{-1}\,d^{-1}$	[154, p. 889]

ADI: 0.015 $mg\,kg^{-1}\,d^{-1}$ [155, p. 889]

Tricyclazole [*41814-78-2*]

LD_{50} (rat, oral)	314	mg/kg	[154, p. 1017]
LD_{50} (mouse, oral)	245	mg/kg	[154, p. 1017]
LD_{50} (dog, oral)	> 50	mg/kg	[154, p. 1017]

LD_{50} (rabbit, dermal)	> 2000	mg/kg	[154, p. 1017]
LC_{50} (rat, one-h inhalation)	1.46	mg/L air	[154, p. 1017]

Rabbit: nonirritating to skin [154, p. 1017]
Rabbit: slightly irritating to eyes [154, p. 1017]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	13.8	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1017]
Mouse (2-year feed):	57	$mg\,kg^{-1}\,d^{-1}$	[154, p. 1017]

Carpropamid [*104030-54-8*]

LD_{50} (rat, oral)	> 5000	mg/kg	[22, p. 744]
LD_{50} (mouse, oral)	> 5000	mg/kg	[22, p. 744]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 744]
LC_{50} (rat, 4-h inhalation)	> 5.0	mg/L air	[22, p. 744]

Rabbit: nonirritating to skin and eyes [22, p. 744]
Guinea pig: not a skin sensitizer [22, p. 744]
Nonmutagenic [22, p. 744]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	400	$mg\,kg^{-1}\,d^{-1}$	[22, p. 744]
Dog (one-year feed):	200	$mg\,kg^{-1}\,d^{-1}$	[22, p. 744]

ADI: 0.014 $mg\,kg^{-1}\,d^{-1}$ [22, p. 744]

Spiroxamine [*118134-30-8*]

LD_{50} (rat, oral)	500 – 600	mg/kg	[22, p. 746]
LD_{50} (rat, dermal)	> 1600	mg/kg	[22, p. 746]
LC_{50} (rat, 4-h inhalation)	1.98 – 2.78	mg/L air	[22, p. 746]

Rabbit: severe skin irritant [22, p. 746]
Rabbit: not an eye irritant [22, p. 746]
Nonmutagenic [22, p. 746]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	3.50	$mg\,kg^{-1}\,d^{-1}$	[22, p. 746]
Mice (18-month feed):	2.85	$mg\,kg^{-1}\,d^{-1}$	[22, p. 746]
Dog (one-year feed):	1.88	$mg\,kg^{-1}\,d^{-1}$	[22, p. 746]

ADI: 0.025 $mg\,kg^{-1}\,d^{-1}$ [22, p. 746]

Probenazole [*27605-76-1*]

LD_{50} (rat, oral)	2030	mg/kg	[22, p. 999]
LD_{50} (mouse, oral)	2750 – 3000	mg/kg	[22, p. 999]
LD_{50} (rat, dermal)	> 5000	mg/kg	[22, p. 999]

Nonmutagenic [22, p. 999]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	22.5	$mg\,kg^{-1}\,d^{-1}$	[22, p. 999]

Teratogenicity.

Rat (gavage):	150	$mg\,kg^{-1}\,d^{-1}$	[22, p. 999]

Fenpropidin [*67306-00-7*]

LD_{50} (rat, oral)	> 1447	mg/kg	[22, p. 526]
LD_{50} (rat, dermal)	> 4000	mg/kg	[22, p. 526]
LC_{50} (rat, 4-h inhalation)	> 1.22	mg/L air	[22, p. 526]

Rabbit: skin and eye irritant [22, p. 526]
Guinea pig: not a sensitizer [22, p. 526]
Nonmutagenic [22, p. 526]
NOEL/NOAEL:
Chronic.

Rat (2-year feed):	0.5	$mg\,kg^{-1}\,d^{-1}$	[22, p. 526]
Mouse (18-month feed):	4.5	$mg\,kg^{-1}\,d^{-1}$	[22, p. 526]
Dog (one-year feed):	2.0	$mg\,kg^{-1}\,d^{-1}$	[22, p. 526]

ADI: 0.005 $mg\,kg^{-1}\,d^{-1}$ [22, p. 526]

Famoxadone [*131807-57-3*]

LD_{50} (rat, oral)	> 5000	mg/kg	[22, p. 500]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 500]

Rabbit: nonirritating to skin and eyes [22, p. 500]
Nonmutagenic [22, p. 500]

Quinoxyfen [*124495-18-7*]

LD_{50} (rat, oral)	> 5000	mg/kg	[22, p. 1083]
LD_{50} (rat, dermal)	> 2000	mg/kg	[22, p. 1083]
LC_{50} (rat, 4-h inhalation)	> 3.38	mg/L air	[22, p. 1083]

Rabbit: mildly irritating to eyes [22, p. 1083]
Rabbit: not a skin irritant [22, p. 1083]
Guinea pig: not a sensitizer [22, p. 1083]
Nonmutagenic [22, p. 1083]
ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$ [22, p. 1083]

11. References

General References

1. D. C. Torgeson: *Fungicides, An Advanced Treatise,* 1 + 2, Academic Press, New York-London 1967/1969.
2. *Plant Pathologist's Pocketbook,* 2nd ed., Commonwealth Mycological Institute, England, 1983.
3. W. T. Thompson: *Agricultural Chemicals,* **rev., vol. IV, Fungicides** Thompson Publications, Calif. 1985.

4. *Farm Chemicals Handbook,* Meister Publishing Co., Willoughby, Ohio 1985.
5. P. H. B. Talbot: *Principles of Fungal Taxonomy,* Macmillan Press, London 1971.
6. K. A. Jeffs (ed.): *Seed Treatment,* The British Crop Protection Council, 1986.
7. G. N. Agrios: *Plant Pathology,* 2nd ed., Academic Press, New York-London 1978.
8. R. W. Marsh (ed.): *Systemic Fungicides,* 2nd ed., Longman, London 1977.

Specific References

9. C. R. Worthing: *The Pesticide Manual,* 7th ed., British Crop Protection Council 1983.
10. J. Darvas, J. Torien, L. Milne, "Control of Avocado Root Rot by Trunk Injection with Phosetyl Al.," *Plant Dis.* **68** (1984) 691 – 693.
11. S. G. Georgopoulus in M. R. Siegel, H. D. Sisler (eds.): *Antifungal Compounds,* **vol. 2,** Marcel Dekker, New York-Basel 1977, pp. 409 – 495.
12. M. Wade, "The aims and activities of Industry's Fungicide Resistance Action Committee," *Plant Dis.* **71** (1987) no. 7, 652 – 653.
13. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 2,** Springer Verlag, Berlin – Heidelberg – New York 1970.
14. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 4,** Springer Verlag, Berlin – Heidelberg – New York 1977.
15. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 6,** Springer Verlag, Berlin – Heidelberg – New York 1981.
16. A. Trinci, J. Ryley (eds.): *Mode of Action of Anti-Fungal Agents,* Cambridge Univ. Press. 1984.
17. E. Krause, R. Pohland, *Ber. Dtsch. Chem. Ges.* **57** (1924) 540. G. J. M. Van der Kerk, J. G. A. Luijten, *J. Appl. Chem.* **4** (1954) 314; **6** (1956) 56.
18. Rohm & Haas Comp., DE-AS 1 234 704, 1961 (C. B. Lyon, J. W. Nemec, V. H. Unger).
19. Du Pont, US 2 504 404, 1950 (A. L. Flenner) US 2 710 822, 1955 (D. R. V. Golding, B. L. Richards Jr.).
20. G. D. Thorn, R. A. Ludwig: *The Dithiocarbamates and Related Compounds,* Elsevier, Amsterdam-New York 1962.
21. Du Pont, US 1 972 961, 1931 (W. H. Tisdale, J. Williams). I. G. Farbenindustrie, DE 642 532, 1934 (H. Günzler, F. Hedemanns, E. Urbschat).
22. C. D. S. Tomlin: *The Pestizide Manual,* 11th ed., British Crop Protection Council 1997.
23. Du Pont, US 1 972 961, 1964.
24. BASF, BG 840 211, 1960 (O. Fleig, H. Windel).
25. Bayer, DE 1 157 027, 1964 (H. Lehmann, F. Grewe, W. Lautenschlager).
26. Rohm & Haas, US 2 457 674, 1949 (J. W. Hueberger).
27. R. Weigler, *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **4** (1977) 136.
28. Standard Oil Development Co., US 2 553 770, 1951 (A. R. Kittleson).
29. C. Tomlin: *Pesticide Manual,* 10th ed., British Crop Protection Council, 1994.
30. Bayer, DE-OS 1 193 498, 1960 (E. Klauke, E. Kühle, F. Greve, H. Kaspers).
31. W. D. Thomas, *Phytopathology* **52** (1962) 754.
32. Bayer, DE 1 193 498, 1963 (E. Klenke, E. Kuehle, F. Grewe, H. Kasper, R. Wegler).
33. J. Dekker, S. G. Georgopoulos (eds.): *Fungicide Resistance in Crop Protection,* Centre for Agric. Publishing, Wageningen 1982.
34. Rhône-Poulenc, DE-OS 2 149 923, 1970 (M. Sauli).
35. BASF, DE-OS 2 207 576, 1972 (D. Mangold, B. Zeeh, E. H. Pommer).
36. Sumitomo, US 3 903 090, 1969 (Akira Fujinami et al.).
37. Montedison, GE 2 906 574, 1980 (V. Di Toro, F. Gozzo, M. Cecere, S. Lorusco, C. Gavanaglia).
38. I. G. Farbenindustrie, DE 682 048, 1932 (A. Steindorff, K. Pfaff, M. Erlenbach, W. Staudermann).
39. Amer. Cyanamid, US 2 425 341, 1947 (J. H. Paden, A. F. MacLean).
40. E. Merck AG, DE-OS 1 158 056, 1963 (H. Hahn, G. Mohr, A. v. Schoor).
41. Diamond Alkali Co., US 3 290 353, 1963 (R. D. Battershell, H. Bluestone).
42. Olin Mathieson Chemical Corp., US 3 260 725, 3 260 588, 1966 (H. A. Schroeder).
43. Bayer, DE-OS 1 643 347, 1967 (E. Roos, K. Wagner).
44. Sumitomo, US 4 039 635, 1975 (T. Kato, M. Sasaki, T. Ooishi, K. Mukai).
45. Du Pont, US 3 957 847, 1976 (S. H. Davidson).
46. Nikon Tokushu Noyaku Seizo K.K., DE-OS 2 732 257, 1976 (Y. H. Yamada, J. M. Saito, T. T. Tamura, Y. H. Kurahashi).
47. *Proc. British Crop Protection Conf. Pests and Diseases,* Brighton 1988.
48. Nippon Soda, GB 2 024 824, 1979 (K. Ohkuma, H. Tagaki, A. Nakata, S. Kosaka). Nippon Soda, GB 2 185 013, 1986 (Y. Genda et al.).

49. Union Carbide, *Tetrahedron* **21** (1965) 1357 (L. Skatteboel).
50. G. Chiusoli, G. Agnas, C. A. Caselli, S. Merzoni, *Chim. Ind.* **46** (1964) 21.
51. L. F. Fieser, M. Fieser, *Lehrbuch der Organischen Chemie* (1957) 689.
52. N. Atsuo, K. Kisao, F. Kenichi, *Nippon Kayaku Zasshi* **87** (1966) 485.
53. Bayer, DE 1 100 372 (F. Grewe, K. Sasse, R. Wegler).
54. Sindar and BDH, US 2 334 408, 1944 (W. S. Gump, M. Luthy).
55. Sankyo, DE 2 810 267, 1978 (T. Jojima, H. Takeshiba).
56. Celaneste Corp., US 2 501 831, 1950 (G. W. Seymour, V. S. Salvin).
57. Shell, DE 1 914 954 (1970).
58. Takeda Chemicals, EP 19 450, 1981 (K. Konishi, K. Matsuura).
59. Novartis, EP 206 999, 1991 (R. Nyfeler, J. Ehrenfreund).
60. Mitsubishi Chemicals, DE 2 521 159, 1976 (K. Matsui, T. Shigematsu, T. Hibohara, M. Nakazawa).
61. Mitsui Toatsu Chemicals, EP 199 433, 1987 (T. Yasimoto, M. Umemoto, K. Igarashi, Y. Kubora, H. Yamazaki, Y. Enomoko, H. Yanagida).
62. Kumiai Chemical Industry, EP 224 339, 1987 (S. Ito, K. Masuda, S. Kusano, I. Nagata, Y. Kojima, N. Sawai).
63. Du Pont, US 3 541 213, 1967.
64. Uniroyal, Neth. 6 605 527, 1968.
65. Bayer, DE 2 802 488, 1980 (K. Sasse, W. Gauss, P. E. Frohberger, P. Kraus, V. Paul).
66. J. Berdy: *Handbook of Antibiotic Compounds,* **vol. 5,** CRC Press, 1981.
67. K. Isone et al., *J. Am. Chem. Soc.* **91** (1969) 7490.
68. T. Isawa et al., *J. Antibiot.* **23** (1970) 595; **24** (1971) 107, 119.
69. T. Suami et al., *J. Antibiot.* **33** (1980) 98.
70. Kumiai Chemical Industry, US 3 708 573, 1969 (E. Yoshinaga, Y. Takahashi, M. Kado).
71. Bayer, BE 686 048, 1967 (G. Schrader, K. Mannes, H. Scheinpflug).
72. Hoechst, DE-OS 1 545 790, 1965 (O. Scherer, H. Mildenberger).
73. Du Pont, US 3 738 995, 1968 (C. de Witt Adams, R. Schlatter).
74. Nippon Soda Co., DE-OS 1 930 540, 1969 (T. Noguchi et al.).
75. Du Pont, US 3 637 733, 1968 (de Witt Adams, R. Schlatter).
76. Bayer AG, DE-OS 1 922 297, 1969 (A. Widding, E. Kühle).
77. R. Weidenhagen, *Ber.* **69** (1936) 2263.
78. Merck, US 3 017 415, 1962 (L. H. Sarett, H. D. Brown).
79. U.S. Rubber Comp., US 3 249 499, 1965 (B. v. Schmeling, M. Kulka, D. S. Thiara, W. A. Harrison).
80. P. Le Roux: "Caractéristiques des souches *d'Ustilago nuda*, agent du charbon nu de l'orge, résistantes à la carboxine," *Agronomie* (*Paris*) **6** (1986) no. 2, 225 – 226.
81. Nihon Noyaku K.K., DE-OS 2 316 921, 1972 (K. Taninaka et al.).
82. Kumiai Chemical Industra, DE 2 434 430, 1975 (I. Chiyomura, S. Kawada, K. Takita).
83. Monsanto, EP 371 950, 1990 (G. H. Alt, J. K. Pratt, W. G. Phillips, G. H. Sranji).
84. Ciba-Geigy, DE-OS 2 515 091, 1975 (A. Hubele).
85. Sandoz, DE-OS 3 030 026, 1980 (J. Harr, R. Sandmeier, H. P. Schelling).
86. Chevron Research Co., US 3 933 860, 1976 (D. C. K. Chan).
87. Montadison, DE 2 903 612, 1980 (E. Bosone, G. Camaggi, L. De Vries, C. Garavoglia, F. Gozzo, J. C. Overeem).
88. Novartis, DE 2 513 788, 1976 (A. Hubele).
89. Novartis, CH 609 964, 1979 (A. Hubele).
90. Pepro, FR 2 254 276, 1973, FR 2 288 463, 1974 (J. Ducret, G. Lacroix, J. M. Gaullirad).
91. Philagro, S.A., FR 2 370 750, 1976, FR 2 402 663, 1977 (J. Abblard, M. Ghazarian, R. Viricel).
92. W. S. Bowers, W. Ebing, T. R. Fukuto, D. Martin et al.: *Chemistry of Plant Protection,* vol. 1, Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1986.
93. Bayer, DE-OS 2 406 665, 1974 (C. Stölzer, W. Krämer, K. H. Büchel, W. Meiser).
94. Bayer, DE-OS 2 324 010, 1973 (W. Krämer, K. H. Büchel, W. Meiser, H. Kaspers et al.).
95. Janssen AG, US 4 079 062, 1974 (G. von Reet, J. Heeres, L. Wals).
96. Janssen AG, DE-OS 2 735 872, 1976 (J. Heeres, L. J. Backx, J. A. Mostmans).
97. Ciba-Geigy, EP 63 099, 1981 (L. Maier, W. Kunz).
98. ICI, EP 15 756, 1979 (K. P. Parry, W. G. Rathmell, P. A. Worthington).
99. Du Pont, EP 68 813, 1982 (W. K. Moberg).
100. ICI, EP 47 594, 1980 (K. P. Parry, W. G. Rathmell, P. A. Worthington).
101. *Proc. British Crop Protection Conf., Pests and Diseases,* Brighton 1986.

102. Sumitomo, DE-OS 3 010 560, 1979 (Y. Funaka et al.).
103. Rohm & Haas Co., EP 145 294, 1983 (T. T. Fujimoto).
104. Bayer, DE-OS 3 018 865, 1980 (G. Holmwood, K. H. Büchel, M. Plempel, I. Haller).
105. Sandoz, DE-OS 3 406 993, 1983 (F. Schaub).
106. Ciba-Geigy, GB 2 098 607, 1982 (A. Hubele, P. Riebli).
107. Janssen, US 4 160 838, 1979 (G. Van Reet, J. Heeres, L. Wals).
108. Rhône-Poulenc, EP 258 161, 1988 (A. Greiner).
109. BASF, EP 94 564, 1984 (B. Janssen, N. Meyer, E. H. Pommer, E. Ammermann).
110. Rohm & Haas, DE 3 721 786, 1988 (S. H. Shaber, K. E. Flynn, B. Weinstein).
111. FBC, EP 183 458, 1986 (D. E. Grem, A. Parcival).
112. Kureha Chemical Industry, EP 267 778, 1990 (S. Kamazawa, S. Shimizu, H. Enari, A. Ito, S. Ikeda, N. Sato, T. Saishoji).
113. Boots Co., US 3 991 071, 1973 (R. F. Brooks et al.).
114. Janssen AG, BP 1 244 530, 1969 (E. F. Godefroi, J. Heeres).
115. Hoffmann-La Roche, EP 49 854, 1982 (F. Dorn).
116. Hokko Chemicals and Ube Industries, JP 90 260 572, 1986 (Y. Hiroja, N. Kuroda, T. Wada, K. Tsujimoto).
117. Eli Lilly and Co., FR 1 569 940, 1967 (J. D. Davenport, R. E. Hackler, H. M. Taylor).
118. Nippon Soda, DE 2 814 041, 1979 (K. Ikura, K. Kahsuura, M. Kataoka, A. Nakada, M. Mizuna).
119. Novartis, EP 310 550, 1989 (A. Hubele).
120. R. S. Elias, M. C. Shepard, B. K. Snell, J. Stubbs, *Nature* **219** (1968) 1160.
121. Eli Lilly, F 1 569 940, 1970 (J. J. Davenport, R. E. Hackler, M. Harald).
122. DE 151 404, 1982 (F. Franke, M. Klepel, G. Krause, H. Lehmann, B. Braemer).
123. Böhringer & Sohn, DE-OS 1 901 421, 1968 (N. Ost, U. Thomas, D. Jerchel, K. Appel).
124. BASF, BE 614 214, 1961 (J. Kradel, E. H. Pommer, H. Effland).
125. Hoffmann-La Roche, DE-OS 2 752 135, 1978 (A. Pfiffner, K. Bohnen). BASF, DE-OS 2 656 747, 1978 (W. Himmele, E. H. Pommer, N. Goetz).
126. Celamerck GmbH, EP 208 999, 1985 (J. Curtze et al.).
127. K. H. Küng, E. H. Pommer, W. Sanne, *Angew. Chem.* **77** (1965) 327.
128. Zeneca, EP 382 375, 1991 (J. M. Clough, C. R. A. Godfrey, I. T. Streeting, R. Cheetham).
129. BASF, DE 3 623 921, 1988 (B. Wenderoth, C. Rentzea, E. Ammermann, E. Pommer, W. Steglich, T. Anke).
130. Shionogi, EP 535 928, 1993 (A. Takase, H. Kai, K. Nishida, M. Masui).
131. Schering AG, DE-OS 1 567 169, 1966, DE-OS 1 643 040, 1967 (G. A. Hoyer, E. A. Pieroh).
132. Ciba-Geigy AG, EP 90 769, 1982 (P. Martin).
133. Sankyo Co., US 3 544 584, 1965 (I. Ivai, Y. Kishida, T. Hiraoka, N. Nakamura).
134. Eli Lilly and Co., DE-OS 2 250 077, 1971 (C. J. Paget Jr.).
135. ICI, GB 1 182 584, 1966 (B. K. Snell, R. S. Elias, P. F. H. Freeman).
136. ICI, GB 1 400 710, 1971 (A. M. Cole, J. A. W. Turner, B. K. Snell).
137. Meiji Confectionary Co., JP 7 014 301, 1967 (I. Seki); US 3 629 428, 1967 (S. Seki et al.).
138. T. Ohashi: *Oryzemate: Novel Systemic Antibacterial and Anti-Fungal Agrochemical,* Japan Pesticide Information 1980, pp. 37 – 39.
139. Nihon, EP 170 842, 1986 (Y. Kurahashi, K. Shiotawa, S. Kagabu, S. Sajiawa, K. Moriga).
140. Bayer, DE 3 735 555, 1989 (W. Kraemer, J. Weissmueller, D. Berg, S. Dutzmann).
141. Hoffmann-La Roche, CH 635 076, (A. Pfiffner, K. Bohnen).
142. Du Pont, WO 94 11 359, 1994 (J. A. Sternberg, K. M. Sun, M. Toji, V. G. Witterholt).
143. DowElanco, US 5 245 036, 1994 (R. L. Robery, C. A. Alt, C. V. DeAminis).
144. Wood Mackenzie and Co.: *Agrochemical Service,* Edinburgh, October 1988.
145. A. Pestano, A. Sols, *Biochem. Biophys. Res. Commun.* **39** (1970) 522.
146. M. Weselowski, L. Rozewicka, *Med. Press* **27** (1976) 103.
147. A. Semple, W. H. Parry, D. E. Phillips, *Lancet* **2** (1960) 700.
148. J. C. Pimentel, J. C. Menezes, *Am. Rev. Respir. Dis.* **111** (1975) 189.
149. J. C. Pimentel, J. C. Menezes, *Gastroenterology* **72** (1975) 275.
150. Pesticide Residues in Food – 1991. *WHO/PCS/* **92.52** (1992) 173.
151. A. W. Hayes (ed.): *Principles and Methods of Toxicology,* 3rd ed., Raven Press, New York, 1994, p. 403.
152. Pesticide Residues in Food – 1991. *WHO/PCS/* **94.4** (1994) 167.
153. *Environmental Health Criteria* **78,** WHO Geneva 1988.

154. C. Tomlin (ed.): *The Pesticide Manuel,* 10th ed., British Crop Protection Council 1994, p. 460.
155. Pesticide Residues in Food – 1980. *FAO Plant Prod Prot. Paper 26 Sup.* (1981) 180.
156. Pesticide Residues in Food – 1992. *WHO/PCS/* **93.34** (1993) 391.
157. W. J. Hayes, E. R. Laws (eds.): *Handbook of Pesticide Toxicology,* Academic Press, New York 1991, p. 1415.
158. Pesticide Residues in Food – 1977. *FAO Plant Prod Prot. Paper 10 Sup.* (1978) 33.
159. 1973 Evaluations of Some Pesticide Residues in Food. *WHO Pesticide Residues Series,* **No. 3,** (1974) 132.
160. Pesticide Residues in Food – 1995. *FAO Plant Prod Prot. Paper* **133** (1996) 52.
161. 1974 Evaluations of Some Pesticide Residues in Food. *WHO Pesticide Residues Series,* **No. 4,** (1975) 185.
162. Pesticide Residues in Food 1983. *FAO Plant Prod Prot. Paper* **61** (1985) 179.
163. Pesticide Residues in Food – 1990. *FAO Plant Prod Prot. Paper* **102** (1990) 37.
164. Pesticide Residues in Food – 1988. *FAO Plant Prod Prot. Paper* **93/2** (1989) 61.
165. Pesticide Residues in Food – 1989. *FAO Plant Prod Prot. Paper* **100/2** (1990) 161.
166. 1966 Evaluations of Some Pesticide Residues in Food. *FAO/WHO* (1970) 91.
167. Pesticide Residues in Food – 1990. *WHO/PCS/* **91.47** (1991) 119.
168. 1975 Evaluations of Some Pesticide Residues in Food. *WHO Pesticide Residues Series,* **No. 5,** (1976) 59.
169. Pesticide Residues in Food – 1987. *FAO Plant Prod Prot. Paper* **86/2** (1988) 27.
170. 1974 Evaluations of Some Pesticide Residues in Food. *WHO Pesticide Residues Series,* **No. 4,** (1975) 215.
171. *Federal Register* **62** (1997) 5403.
172. Pesticide Residues in Food – 1994. *WHO/PCS/* **95.2** (1995) 189.
173. M. Bakkeren, S. Jong, J. Krul, *Proc. Br. Insectic. Fungic* . Conf. 9 (1977) 541.
174. T. Misato, *Residue Rev.* **25** (1969) 93,
175. O. R. Klimmer: *Pflanzenschutz- und Schädlingsbekämpfungsmittel, Abriss einer Toxikologie und Therapie von Vergiftungen,* Hundt-Verlag, Hattingen 1971, pp. 39 – 46.
176. C. Klotzsche, *Int. Arch. Gewerbepathol. Gewerbehyg.* **21** (1964) 92.
177. A. F. Grapov, *Usp. Khim.* **42** (1973) 1681.
178. Y. Uesugi, *Shokubutsu Boeki* **26** (1972) 103.
179. Pesticide Residues in Food – 1981. *FAO Plant Prod Prot. Paper* **42** (1981) 195.
180. 1970 Evaluations of Some Pesticide Residues in Food. *FAO/WHO* **1971,** 479.
181. *Federal Register* **57** (1992) 8842.
182. Pesticide Residues in Food – 1982. *FAO Plant Prod Prot. Paper* **49** (1982) 259.
183. D. Berg, M. Plempl (eds.): *Sterol Biosynthesis Inhibitors: Pharmaceutical and Agrochemical Aspects,* Ellis Horwood, 1988.
184. P. Gadher, E. C. Mercer, B. C. Baldwin, T. E. Wiggins, *Pesticide Biochem. Physiol.* **19** (1983) 1.
185. S. S. Kukalenko, S. D. Volodkivich, *Kim. Sel'sk Khoz.* **17** (1979) 30.
186. D. Berg, W. Draber, H. von Hugo, W. Hummel, D. Mayer, *Z. Naturforsch.* **36 c** (1981) 798.
187. M. R. Gajendragat, S. J. Seshadri, *Indian J. Anim. Sci.* **53** (1983) 185.
188. R. A. Martin, L. v. Edgington, *Pestic. Biochem. Physiol.* **17** (1983) 1.
189. J. B. Knaak, K. Yee, C. R. Ackermann, G. Zweig, B. W. Wilson, *Toxicol. Appl. Pharmacol.* **72** (1984) 406.
190. J. L. Rivière, P. Leroux, J. Bach, M. Gredt, *Pestic. Sci.* **15** (1984) 317.
191. T. E. Wiggins, B. C. Baldwin, *Pestic. Sci.* **15** (1984) 206.
192. G. O. Rankin, D. J. Yang, K. Cressey-Veneziano, R. T. Wang, P. I. Brown, *Toxocology* **34** (1985) 1.
193. Y. Yoshida, Y. Aoyama, *Biochem. Pharmacol.* **36** (1987) 229.
194. UK MAFF, *Evaluation of Azaconazole,* London1988.
195. *Azaconazole Doc.* 11 Agricultural Canada, Pesticide Directive Ottawa1989.
196. UK MAFF, *Evaluation of Cyproconazole,* London 1991, Doc. 45.
197. UK MAFF, *Evaluation of Difenoconazole,* London 1994, Doc. 106.
198. UK MAFF, *Evaluation of Epoxyconazole,* London 1994, Doc. 108.
199. *Federal Register* **53** (1988) 34513.
200. *Federal Register* **59** (1994) 8135.
201. *Propiconazole,* Agricultural Canada, Pesticide Directive, Ottawa 1987, 87 – 05.
202. UK MAFF, *Evaluation of Tebuconazole,* London 1993, Doc. 1.
203. *Federal Register* **60** (1995) 40499.
204. *Triadimefon,* Agricultural Canada, Pesticide Directive, Ottawa 1987, 87 – 03.
205. US-EPA, *Tox One – Liners,* Bayleton, Washington 1985.
206. Pesticide Residues in Food – 1985. *FAO Plant Prod Prot. Paper* **72/2** (1985) 199.

207. *Federal Register* **58** (1993) 46084.
208. Pesticide Residues in Food – 1986. *FAO Plant Prod Prot. Paper* **78/2** (1996) 77.
209. *Federal Register* **62** (1997) 5406.
210. Pesticide Residues in Food – 1978. *FAO Plant Prod Prot. Paper* **15** (1978) 257.

Fungicides, Pharmaceutical → Antimycotics

Furfural and Derivatives

H. E. Hoydonckx, TransFurans Chemicals b.v.b.a, Belgium

W. M. Van Rhijn, TransFurans Chemicals b.v.b.a, Belgium

W. Van Rhijn, TransFurans Chemicals b.v.b.a, Belgium

D. E. De Vos, Centre for Surface Chemistry and Catalysis, KULeuven, Belgium

P. A. Jacobs, Centre for Surface Chemistry and Catalysis, KULeuven, Belgium

1. Introduction 1
2. Furfural 3
2.1. Properties 3
2.1.1. Physical Properties 3
2.1.2. Chemical Properties 4
2.2. Resources and Raw Materials 6
2.3. Production 7
2.3.1. Chemistry of Furfural Production . . . 7
2.3.2. Production of Furfural 8
2.4. Uses and Economic Aspects 10
2.5. Quality Specifications, Transportation, Toxicology 11
3. Furfuryl Alcohol 12
3.1. Properties 12
3.1.1. Physical Properties 12
3.1.2. Chemical Properties 12
3.2. Resources and Raw Materials 15
3.3. Production 15
3.4. Uses and Economic Aspects 17
3.5. Quality Specifications, Transportation, Toxicology 18
4. Furan 18
4.1. Properties 18
4.2. Production 20
4.3. Uses, Economic Aspects, Transportation, Toxicology 21
5. Tetrahydrofurfuryl Alcohol 22
5.1. Properties 22
5.2. Production 22
5.3. Uses, Economic Aspects, Transportation, Toxicology 22
6. Furfurylamine and Tetrahydrofurfurylamine 23
7. Methylfuran and Methyltetrahydrofuran 24
8. Furoic Acid 25
9. References 25

Detailed information is given on the production, properties, and economical aspects of furfural [*98-01-1*] and furfuryl alcohol [*98-00-0*], and the most important specialty chemicals derived from furfural and furfuryl alcohol are summarized: furan [*110-00-9*], tetrahydrofurfuryl alcohol [*97-99-4*], 3,4-dihydro-2*H*-pyran [*110-87-2*], furfurylamine [*617-89-0*], tetrahydrofurfurylamine [*4795-29-3*], 2-methylfuran [*534-22-5*], 2-methyltetrahydrofuran [*96-47-9*], and furoic acid [*88-14-2*].

1. Introduction

This article discusses the properties, production, and industrial uses of furfural and compounds derived from furfural on a commercial basis. The only commercial route to furfural is acid-catalyzed digestion of pentosan sugars in biomass. Nowadays, furfural is the sole precursor for all molecules containing a furyl, furfuryl, furoyl, or furfurylidene radical in the chemical industry.

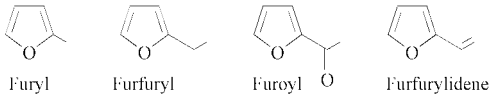

History. In 1831 J. Döbereiner isolated a yellow oily substance when he treated carbohydrates with H_2SO_4 and MnO_2. In 1845 the same substance with molecular formula $C_5H_4O_2$ was prepared by G. Fownes by acid digestion of bran and named furfural (Latin *furfur* = bran). Almost 100 years after its discovery, commercial production of furfural from oat hulls was started by the Quaker Oats Company in 1922 at their cereal mill in Cedar Rapids, Iowa producing several

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_119.pub2

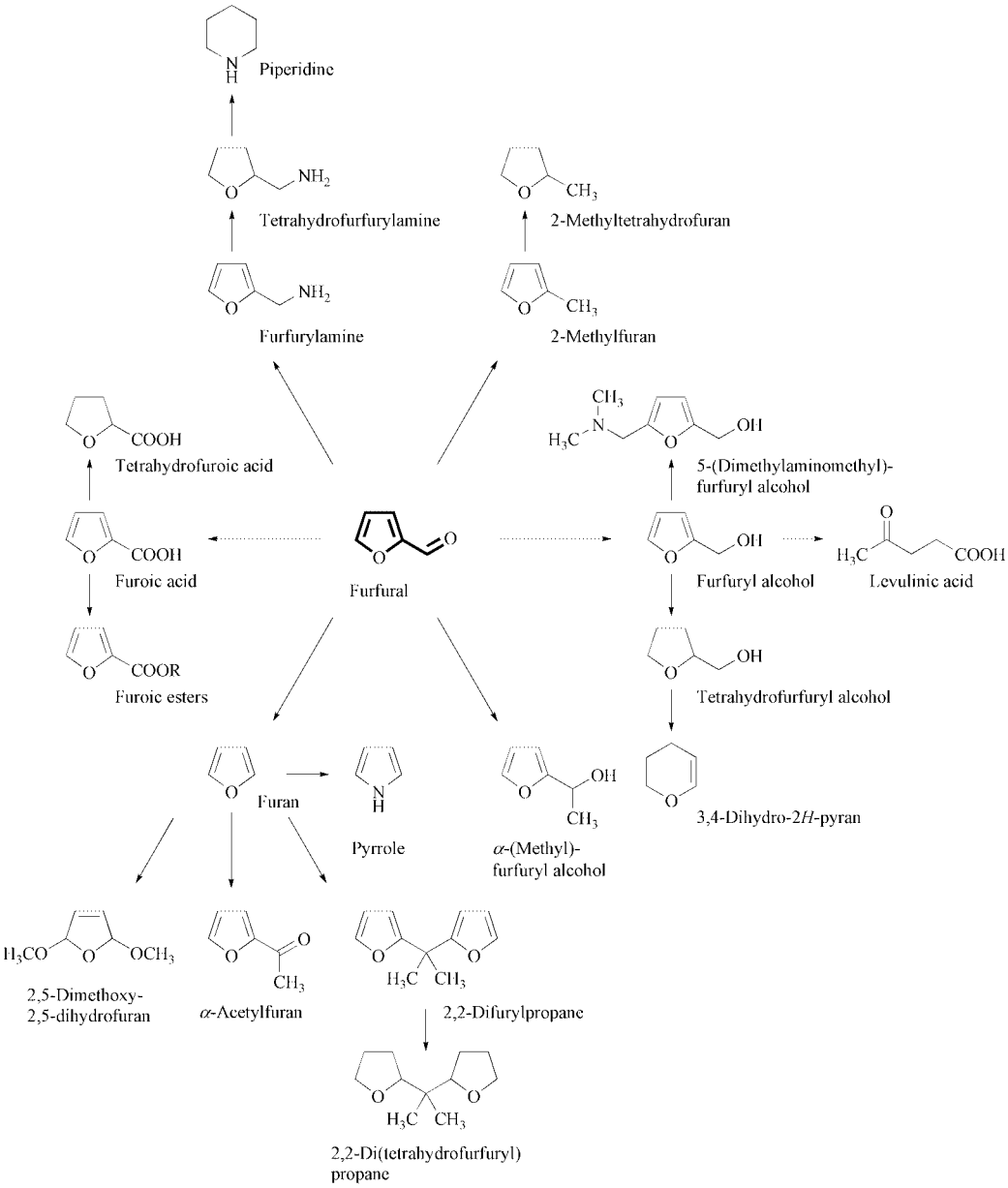

Scheme 1. Chemicals derived from furfural

tons per month. Discovery of the applicability of furfural in oil refining and phenol–furfural resins led to an important increase in furfural production capacities in the 1930s and 1940s. In 1934 furfuryl alcohol was made commercially available by the Quaker Oats Company by high-pressure reduction of furfural. Today this alcohol is still the largest volume derivative of furfural. In 1949 Du Pont disclosed a new route to adiponitrile for the manufacture of nylon 66 from furfural. This four-step process with furan, tetrahydrofuran, and dichlorobutane as intermediates led to an important additional increase in furfural production during the 1950s. In 1961 Du Pont abandoned this furfural-based production process because tetrahydrofuran, the key intermediate, could be made from petrochemical C_4 hydrocarbons or acetylene and formaldehyde

(Reppe process). In 1955 tetrahydrofurfuryl alcohol, a water-miscible solvent, was made commercially available from furfuryl alcohol. After the introduction of the furan no-bake process for the production of casting molds in 1958, the foundry industry became a major consumer of resins based on furfural and furfuryl alcohol. Improved versions of these resins became available in the furan hot-box (1962) and furan warm-box (1978) processes. In 1985 furfuryl alcohol became a key intermediate in the pharmaceutical industry for the production of ranitidine, an antiulcer drug. In today's chemical industry furfural is the precursor for (Scheme 1)

- Furfuryl alcohol and tetrahydrofurfuryl alcohol
- 3,4 Dihydro-2*H*-pyran
- 5-(Dimethylaminomethyl)furfuryl alcohol,
- α-Methylfurfuryl alcohol,
- Levulinic acid
- 2-Methylfuran and 2-methyltetrahydrofuran
- Furan
- 2,2-Difurylpropane and 2,2-bis(tetrahydrofurfuryl)propane
- 2,5-Dimethoxy-2,5-dihydrofuran
- α-Acetylfuran
- Pyrrole
- Furfurylamine and tetrahydrofurfurylamine
- Piperidine
- Furoic acid, furoic esters, and tetrahydrofuroic acid

2. Furfural

2.1. Properties

2.1.1. Physical Properties

Furfural [*98-01-1*], 2-furancarbonal, 2-furaldehyde, furfuraldehyde, $C_5H_4O_2$, M_r 96.08, has an aromatic odor reminiscent of almonds. Its exceptional physical properties make this heteroaromatic aldehyde a selective extractant for removing aromatics from lubricating oils and diesel fuels. Detailed information on the solvent applications of furfural can be found in [1]. Furfural forms an azeotrope with water at atmospheric pressure (65 wt % water, *bp* 97.85 °C). General physical properties of furfural are given in Table 1. Equations for calculation of some physical properties are presented in Table 2.

Table 1. Physical properties of furfural

M_r	96.08
bp at 101.3 kPa, °C	161.7
fp, °C	−36.5
n_D^{20}	1.5261
n_D^{25}	1.5235
ρ at 20°C, g/cm^3	1.1598
Vapor density (air=1)	3.3
Critical pressure P_c, MPa	5.502
Critical temperature T_c, °C	397
Solubility, wt% (25°C), in water	8.3
ethyl alcohol, diethyl ether	∞
Spectroscopic polarity (E_T^N)	0.426
Dielectric constant at 20°C	41.9
Heat of vaporization (liquid), kJ/mol	42.8
Heat capacity (liquid, 20–100 °C), J g^{-1} K^{-1}	1.74
Heat of combustion (liquid), kJ/mol	2344
ΔH_f (l), kJ/mol	−201.65
ΔH_f (g), kJ/mol	−151.05
Viscosity, mPa·s, 25 °C	1.49
Surface tension at 29.9°C, mN/m	40.7
Explosion limits (in air), vol %	2.1/19.3
Flash point, °C, tag closed cup	61.7
Autoignition temperature, °C	315

Table 2. Equations for physical properties of furfural

Physical property	Equation	Constants
Vapor pressure	$p_i = \exp(a + b/T + c \ln T + dT^2)$	$a = 78.653$
		$b = -8043$
		$c = -8.1424$
	(p_i in Pa and T in K)	$d = 4.509 \times 10^{-6}$
Thermal conductivity liquid	$\lambda = a + bT$	$a = 0.2295$
		$b = -1.936 \times 10^{-4}$
vapor	$\lambda = (aT^b)/(1 + c/T + d/T^2)$	$a = 8.3221 \times 10^{-3}$
	(λ in W m^{-1} K^{-1} and T in K)	$b = 0.4228$
		$c = 637.87$
		$d = 5.502 \times 10^5$
Surface tension	$\sigma = a(1–T)^b$	$a = 0.08429$
	(σ in N/m and T in K)	$b = 1.1124$
Heat capacity at constant pressure liquid	$C_p = a + bT$	$a = 73.9$
		$b = 0.2994$
vapor	$C_p = a + b[(c/T)/\sinh(c/T)]^2 + d[(e/T)/\cosh(e/T)]^2$	$a = 47.3$
		$b = 198.3$
		$c = 1040.6$
	(C_p in J/mol.K and T in K)	$d = 109$
		$e = 472.6$
Enthalpy of evaporation	$\Delta H_{vap} = a(1-T)^b$	$a = 59610$
	(ΔH_{vap} in J/mol and T in K)	$b = 0.3126$

2.1.2. Chemical Properties

Furfural undergoes reactions typical for aldehydes like acetalization, acylation, aldol and Knoevenagel condensations, reduction to alcohols, reductive amination to amines, decarbonylation, oxidation to carboxylic acids, and Grignard reactions. The furan ring can be subjected to alkylation, hydrogenation, oxidation, halogenation, and nitration reactions. Due to the electron-withdrawing effect of the carbonyl group, the furan ring of furfural is less susceptible to hydrolytic ring cleavage and Diels–Alder cycloaddition reactions.

Alkylation (Arylation). Furfural undergoes common Friedel–Crafts alkylations, for example, with chlorinated hydrocarbons [2–4]. The 5-position is most susceptible to alkylation reactions. Use of large amounts of acid catalyst and high reaction temperature should be avoided to prevent extensive resinification of furfural. Another method to alkylate furfural in its 5-position is oxidative coupling catalyzed by Pd acetate in the presence of an olefin as alkenylating agent. In the case of acrylic esters this reaction proceeds with moderate yields and Pd-based turnover numbers [5, 6]. Furfural and 5-bromofurfural can be coupled by photochemical aryl coupling to form 5,5′-diformyl-2,2′-difuran [7]. If this reaction is carried out in the presence of an acid scavenger the dialdehyde can be produced in high yields. Photochemical production of 5,5′-diformyl-2,2′-difuran has been patented [8].

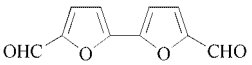

5,5-Diformyl-2,2'-difuran

Cannizzaro Reactions. The Cannizzaro reaction of furfural in alcoholic KOH solution yields equimolar amounts of potassium furoate and furfuryl alcohol [9]. Cannizzaro reactions of furfural can also be performed under solvent-free conditions [10] and under microwave irradiation [11]. The Cannizzaro reaction of furfural is currently used for the production of furoic acid (see Chapter 8).

Acetalization and Acylation. Like any other aldehyde furfural undergoes acetalization with various alcohols [12, 13].

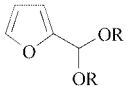

Furfuryl acetal

If the reaction water is removed, the reaction can be brought to completion with or without an acid catalyst [14–16]. More recently Ce^{3+}-exchanged montmorillonite clays and mesoporous aluminosilicates were found to be efficient and reusable inorganic heterogeneous catalysts for the production of acetals of furfural [17, 18].

Furfural can be acylated with acid anhydrides. In the presence of acetic anhydride and stannous chloride as catalyst, furfuryl diacetate is formed [19]. More recently other acid catalysts have been applied, such as H-Nafion [20], $FeCl_3/SiO_2$ under microwave irradiation [21], Fe^{3+}-exchanged montmorillonite [22], HY zeolite [23], and heteropolyacids such as $H_6P_2W_{18}O_{62}\cdot H_2O$ [24].

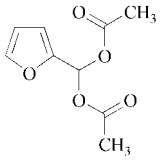

Furfuryl diacetate

Condensation Reactions. In the presence of alkaline catalysts, generally NaOH or KOH, furfural undergoes a crossed aldol reaction with aldehydes having α-hydrogen atoms [25–31]. A typical reaction is the condensation of furfural with α-phenylacetaldehyde to give 2-phenyl-3-(2-furyl)propenal.

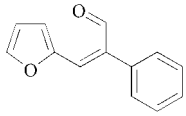

2-Phenyl-3-(2-furyl)-propenal

When the compound with an active α-hydrogen atom is an ester, acid, or a nitrile, this base-catalyzed reaction is called a Knoevenagel

condensation [32–34]. Microwave-assisted and solvent-free procedures in the presence of heterogeneous catalysts have received some attention [35, 36].

Reduction and Reductive Amination. Furfural reduction, the principal reaction in today's furfural industry, is used for the production of furfuryl alcohol (Section 3.2), tetrahydrofurfuryl alcohol (Section 5.1), furfurylamine, and tetrahydrofurfurylamine (Chap. 6).

Decarbonylation. Decarbonylation of furfural to furan and CO or CO_2 and H_2 is discussed in Section 4.2.

Oxidation. The aldehyde group of furfural easily undergoes catalytic oxidation to furoic acid in the presence of oxygen, air, or hydrogen peroxide. Details of these liquid-phase reactions are given in Chapter 8.

At high temperatures (320 °C) furfural can be oxidized with air to maleic anhydride with a maximum yield of 25 % over MoO_3–V_2O_5 catalysts [37, 38].

Photooxidation of furfural with singlet dioxygen yields 5-hydroxy-2(5*H*)-furanone (80 %), a highly versatile compound in organic synthesis and a key constituent of a number of biologically active compounds [39]. In this reaction, singlet oxygen is generated from O_2 with sensitizers like rose Bengal or methylene blue under sunlight irradiation. 5-Hydroxy-2(5*H*)-furanone can be used directly as fungicide or insecticide. This reaction was patented by BASF in 1971 [40].

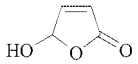

5-Hydroxy-2(5*H*)-furanone

Halogenation and Nitration. Furfural can be halogenated and nitrated at the 5-position to form 5-halo-2-furfuraldehyde and 5-nitro-2-furfuraldehyde [41]. Chlorination occurs in CS_2 by direct reaction of furfural with chlorine, whereas 5-bromo-2-furfuraldehyde can only be formed by the action of bromine on furfuryl diacetate. Nitration of furfural can be performed with fuming nitric acid in acetic anhydride at −10 °C. Hydrolysis of the nitrated furfural diacetate results in 5-nitro-2-furfuraldehyde.

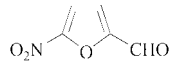

5-Nitro-2-furfuraldehyde

5-nitro-2-furfuraldehyde and 5-nitro-2-furfuraldehyde diacetate are the key components in the manufacture of various synthetic antimicrobial nitrofuran pharmaceuticals.

Grignard Reaction. The reaction between furfural and an organomagnesium halide (Grignard reagent) yields secondary furfuryl alcohols. This reaction is currently used in manufacturing α-methylfurfuryl alcohol, the precursor for maltol.

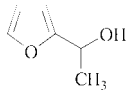

α-(Methyl)furfuryl alcohol

Alternatively, α-methylfurfuryl alcohol can be obtained by reduction of α-acetylfuran, but this route is not employed in today's furfural-based industry.

Resinification. Furfural is very sensitive to resinification [42]. On heating, furfural undergoes self-condensation reactions in both acid and alkaline media. In the initial stage of polymerization furfural undergoes two different reactions, namely, the formation of difuryl ketonic aldehyde (**10**) and trifurylic dialdehyde (**11**). The latter is responsible for further polymerization, which yields a black cross-linked polymer. Because of the high reactivity of the tertiary H atom in **2**, conjugated systems with strongly delocalized electron pairs are formed, which lead to the black color of the resin.

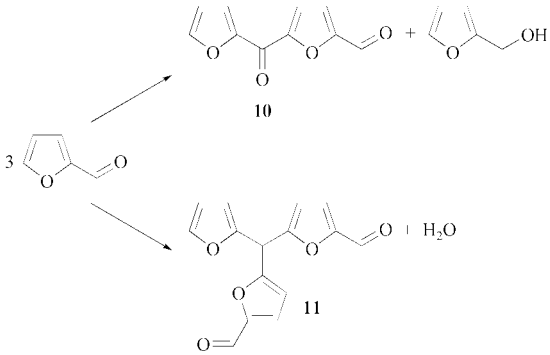

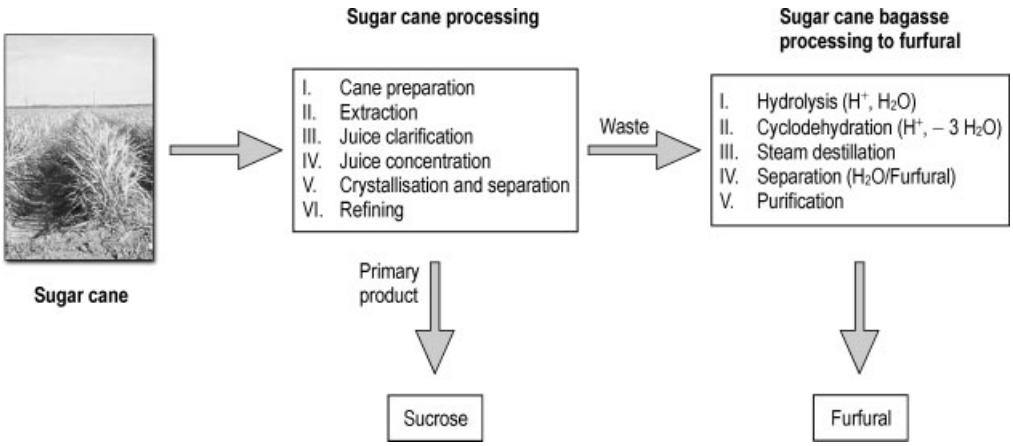

Figure 2. Integration of furfural production in the manufacture of sucrose from sugar cane

2.2. Resources and Raw Materials

The only route to furfural in today's chemical industry is the acid-catalyzed digestion of hemicellulose-rich agricultural wastes. Hemicelluloses are next to cellulose the most abundant biobased polymers in the biosphere and mainly consist of C_5 polysaccharides, or pentosans. Biomass with a substantial amount of pentosan sugar can be used as raw material for furfural production (Fig. 1). The C_5 polysaccharides are hydrolyzed by an acid catalyst to monosaccharides, which are dehydrated to furfural.

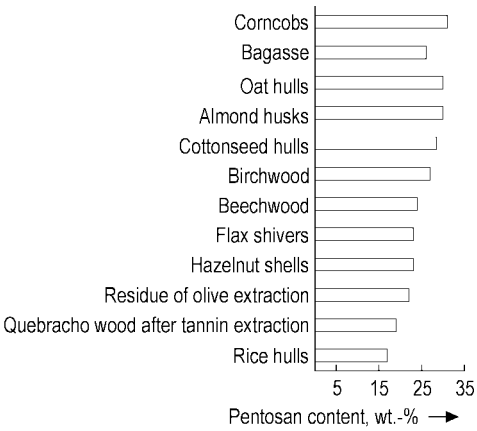

Figure 1. Pentosan content (dry substance) of raw materials for furfural production

On industrial scale the pentosan content, price, local availability, and costs for collection, transportation, and handling of the raw material greatly determine the profitability. Corncobs and bagasse from sugar cane are the major industrial feedstocks for furfural production. The high bulk density and pentosan content make furfural production from corncobs feasible. Nevertheless, high collection and transportation costs can be a drawback. Although bagasse is difficult to handle (low bulk density), the largest furfural production plants are located at sugar mills where sugar cane is being processed. In this process sucrose is extracted from the sugar cane. The resulting waste, bagasse, is digested to furfural (Fig. 2).

The common raw materials used for furfural production differ by region (Table 3). Next to the use of these renewable resources, furfural is produced as a byproduct from sulfite pulping of wood. During the pulping process a minor quantity of pentosan sugars is dehydrated to furfural, which can be recovered by distillation. The rest of the pentosan fraction can be extracted and purified to obtain xylose for xylitol production [43].

Table 3. Use of raw materials for furfural production by region (2004)

Raw material	Region
Bagasse	Dominican Republic, Thailand, India; South Africa
Corncobs	China
Almond shells, apricot and olive seeds	Spain
Quebracho, chestnut, birch and aspen wood	Argentina; Slovenia; Russia
Sulfite black liquor of wood pulping	Austria

2.3. Production

2.3.1. Chemistry of Furfural Production

Furfural is produced from the pentosan sugar fraction of biomass. This pentosan fraction consists of polysaccharides of C_5 sugars and is generally built up from a backbone of 1,4-linked β-D-xylopyranosyl residues or xylan. The xylan backbone contains α-L-arabinofuranose groups attached at various positions. Therefore, these structures are called arabinoxylans (**1**). In the production of furfural, arabinoxylans are first hydrolyzed by acids to the monosaccharides L-arabinofuranose (**2**) or L-arabinose (**3**) and D-xylopyranose (**4**) or D-xylose (**5**) (see below).

Some hydroxyl groups of arabinoxylan (**1**) are acylated with short fatty acids like acetic acid. These esters are also hydrolyzed to the corresponding carboxylic acids **6**.

In the second step monosaccharides are dehydrated to furfural. The dehydration pathway is presented for D-xylose; a similar pathway can be considered for conversion of arabinose to furfural. A generally accepted pathway for the dehydration occurs through the 1,2-enediol form (**7**) of xylose. A first water molecule is abstracted at the 3-position to produce 3-deoxy-D-xylosulose (**8**) in its enolic form. Further dehydration at the 4-position yields 3,4-dideoxy-D-xylo-3-pentenosulose (**9**), which is converted to the cyclic acetal and finally to furfural by loss of a third water molecule. The hydrolysis of the pentosan fraction to monomeric C_5 sugars is approximately 50 times faster than the dehydration to furfural. The kinetics of these reactions have been studied in detail [44–48].

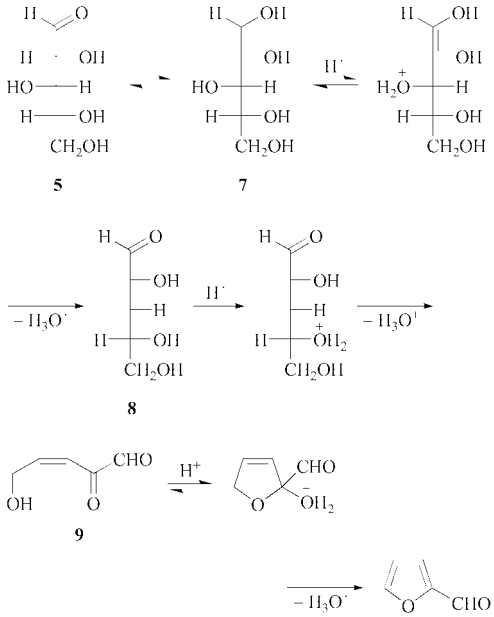

Some pentosan monosaccharides contain additional methyl groups at the 1- or 5-position. On dehydration these sugars yield respectively 2-furyl methyl ketone (2-acetylfuran) and 5-methylfurfural. These two products are common impurities (<1 wt %) in commercial furfural. Acetaldehyde, another byproduct, results from the acid digestion of coniferylaldehyde, which is present in some raw materials. Under industrial reaction conditions acetaldehyde can be converted to 2,3-butanedione and 2,3-

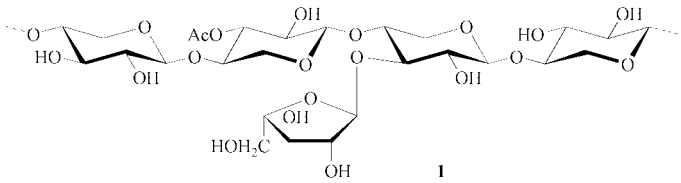

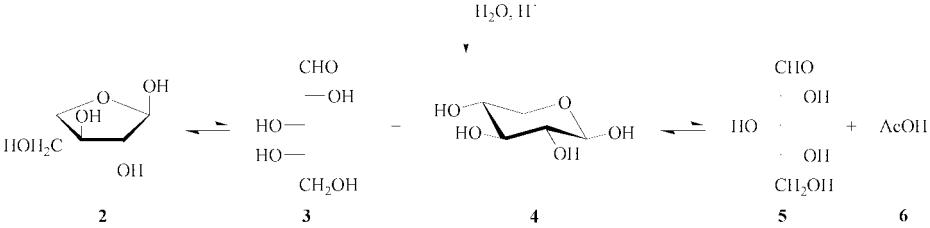

pentanedione by a radical reaction with oxygen. If they are formed in furfural production, these highly valuable diketones are generally recovered from crude furfural by distillation.

Under industrial conditions not all of the pentosan sugars are converted to furfural. As a substantial amount of product and substrate are lost in side reactions, yields never exceed 55 % of the theoretical maximum. Acetalization of pentose with furfural is the most important yield-lowering side reaction.

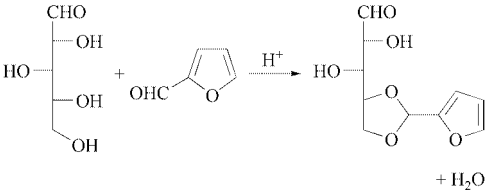

The choice of acid catalyst used under industrial conditions is restricted to the mineral acids H_2SO_4 and H_3PO_4. HNO_3 cannot be used due to corrosion problems and loss of product by formation of nitrated and oxidized products. Some processes are operated without addition of acid catalyst and rely on the acidity of the thermally produced carboxylic acids.

The analytical method for the determination of the total pentosan content of the raw materials can also rely on the acid dehydration of pentoses to furfural [49, 50]. In this system the formed furfural is instantaneously vaporized due to its insolubility in the boiling reaction medium. In common industrial reactors, the reaction medium cannot be brought to its boiling temperature because heating is always effected by the introduction of superheated steam, the condensation temperature of which is below the boiling point of the reaction medium.

2.3.2. Production of Furfural

Furfural production processes can be roughly divided into low- and high-temperature processes. Most commercial units are operated under low-temperature (< 190 °C) conditions. Reactors are always heated by introduction of superheated steam. The furfural yield of these processes, based on the pentosan content, is typically around 50 %.

Industrial Processes. The oldest furfural process was started in 1921 by the Quaker Oats Company [51]. In the *Quaker Oats batch process*, reactors are lined with carbon bricks sealed by acid-proof cement. Air-dried biomass, for example, dried oat hulls, is mixed with aqueous sulfuric acid at a H_2SO_4 to dry biomass weight ratio of 0.02. Steam injection is used to heat the rotary reactor to 153 °C and to remove the formed furfural. After 5 h the reaction is completed and the acid residue is separated into spent liquor and digested biomass. Quaker Oats issued some variants of this basic design regarding reactor setup [52–54], furfural removal [55], and material pretreatment [56]. Optimized versions of the Quaker Oats batch process are still in use in furfural manufacturing.

The simplest variant of the Quaker Oats batch process is the current *Chinese batch method*. In this process, the reactors have no lining; instead, mild-steel reactors with 50 mm wall thickness are employed. Additionally, fouling by furfural-derived polymers protects the inner wall of the reactor against corrosion.

The *Agrifurane process*, now abandoned, was operated with slurry mixtures rather than with loose material to permit the use of fines. Several batch reactors (161–177 °C) are operated in series in order to reduce steam consumption [57–59].

After using the batch process for almost 50 years the Quaker Oats Company developed the *Quaker Oats continuous process*, with a full-scale continuous plant at Belle Glade, Florida (1966–1997). The central part of this process was a brick-lined horizontal reactor which was fed with moisturized raw material. Admission of superheated steam and dilute sulfuric acid brought the contents to 184 °C. The Quaker Oats continuous process produced furfural with the same yield as the batch process, but without continuous steam stripping and with a fivefold decrease in residence time. This innovative plant was operative for more than 30 years before it was shut down in 1997 because the cost of raw materials, labor, and maintenance rendered furfural production in the USA unprofitable.

Other continuous processes are the Escher Wyss process and the Rosenlew process. In the now abandoned *Esher Wyss process* [60] furfural was produced in a fluidized bed of biomass in a vertical column reactor. With continuous charge

and discharge of raw material, this resulted in a wide distribution of residence times, as is also observed in common continuous stirred tank reactors. This caused a dramatic yield loss compared to simple batch reactions. *The Rosenlew process* [61] is comparable to the Escher Wyss process. The reactor is also a column-type reactor with charge at the top and discharge at the bottom and a countercurrent flow of steam, but the raw material is not mixed by fluidization. Instead, as in a moving-bed reactor, slow transport of material from top to bottom results in a narrow residence-time distribution and acceptable furfural yields. Both Esher Wyss and Rosenlew processes can be operated without the addition of extra acid; the reactions are catalyzed by autocatalytically formed carboxylic acids.

Novel and Future Processes. Challenges for future processes are to increase furfural yield and to optimize the energy efficiency of furfural plants.

Early work [47] showed that furfural yield from aqueous mixtures of xylose increases with increasing reaction temperature. However, continuous steam stripping of the reaction does not significantly improve yields. Implementation of this knowledge can be found already in the Quaker Oats continuous process, which operated at 184 °C without steam stripping of the formed furfural. More recently patented processes utilize even higher reaction temperatures, preferentially above 220 °C. At these high temperatures residence times are decreased to a few minutes; this allows the use of small reactors.

The *Suprayield process* [62–64] operates at high temperature and pressure. However, the design allows the reaction mixture to boil by gradual decompression of the reactor at high temperature. Furfural is insoluble in the boiling reaction mixture and is spontaneously transferred to the vapor phase. In this manner furfural is removed more effectively than by stripping steam. Since the formed furfural is quickly removed from the liquid phase, it cannot participate in yield-lowering side reactions. In comparison with traditional methods superior yields (70 %) were obtained in a pilot-scale setup.

Biorefinery Processes In all the processes mentioned above, refining of the biomass is restricted to furfural formation. After reaction, the digested raw material is burned to meet the energy demand of the furfural production plant or used as fertilizer after neutralization. Hence, only the C_5 carbohydrate fraction of the biomass is utilized. Next to the acid digestion of pentosans to furfural, the hexose polysaccharides (e.g., cellulose) can be used for the production of levulinic acid. The hexosan fraction needs a hydrolysis temperature of 200 °C in the presence of dilute aqueous H_2SO_4 in order to release hexose monosaccharides. Under these conditions hexoses are dehydrated to 5-(hydroxymethyl)furfural which is hydrolyzed to levulinic and formic acids [65–70]. A plant for the production of levulinic acid from cellulose-rich materials is currently (2006) being set up by Le Calorie s.p.a in Caserta, Italy. Levulinic acid is a potential precursor for 2-methyltetrahydrofuran (Chap. 7).

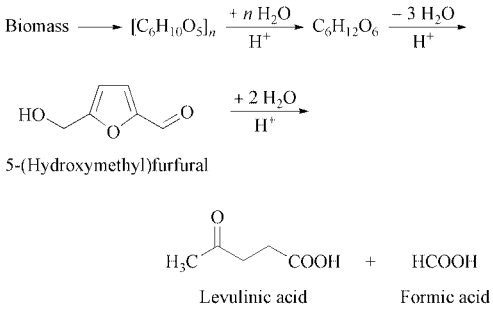

Another example of utilization of the C_6 carbohydrate fraction is the production of ethanol and furfural from deciduous wood and agricultural waste. In 1926 G. Scholler reported the percolation hydrolysis of coniferous wood to obtain glucose and subsequently ethanol by fermentation. In this process, dilute sulfuric acid was continuously filtered through a layer of wood chips at elevated pressure and temperature to depolymerize cellulose to glucose. Whereas coniferous wood has a total pentosan content of 5 wt %, deciduous wood like beech- and birchwood contains up to 25–28wt % pentosan sugars. Therefore, the production of ethanol and furfural from deciduous wood and agricultural waste is very interesting. Nevertheless, in traditional furfural production processes and the percolation-hydrolysis method, 40–50 % of the cellulose is destroyed during furfural formation when biomass with high pentosan content is

used. Hence, these processes cannot be adapted for the combined economic production of ethanol and furfural. According to the process developed by N. Vedernikov, furfural can be produced in 70 % yield from biomass with preservation of the cellulose part [71–73]. The new technology, which allows the fermentative production of ethanol from the lignocellulosic residue after the production of furfural by using the cooperative catalytic effect of concentrated sulfuric acid and thermally produced carboxylic acids, was implemented in 1997 in Russia. In the case of birch and aspen wood as raw materials, the plant capacity is 11 (10^6 L/a of ethanol and 4000 t/a of furfural.

Distillation of Water–Furfural Product Vapor. In all known furfural manufacturing processes, the output of the acid digestion reactor is furfural in water vapor containing various byproducts. The process diagram of a furfural distillation plant is presented in Figure 3. Product vapors are condensed before they are injected into the azeotropic distillation column (A) at atmospheric pressure. In this column the water–furfural azeotrope is purified from low boilers such as methanol and high boilers comprising carboxylic acids such as acetic acid. The purified water–furfural vapor is then condensed and fed to a decanter (C) where an aqueous and an organic phase are formed. The light aqueous phase, which contains a small amount of furfural, is recycled to the azeotrope distillation column. The heavy layer (94 % furfural) is neutralized with alkali (D) and undergoes an additional vacuum distillation (E) to yield pure anhydrous furfural. The head fraction of this vacuum column contains the water–furfural azeotrope, which is recycled to the static decanter. The bottom fraction consists of a small amount of furfural polymers.

2.4. Uses and Economic Aspects

Uses. Of the world production of furfural 60–70 % is converted to furfuryl alcohol. The remaining part (50 000 t in 2003) is used as:

- Extractant for aromatics from lubricating oils
- Purification solvent for C_4 and C_5 hydrocarbons
- Reactive solvent and wetting agent
- Chemical feedstock for other furan derivatives
- Nematode control agent

Furfural is used as selective solvent in the refining of lubricating oils. The first refining unit was built by a subsidiary of Texaco, the Indian Refining Company, in 1933. This technology led to production of high-quality motor oil with improved temperature–viscosity properties. Aromatics, polar components, mercaptans, and others are removed from petroleum by means of furfural extraction. Furfural can also be used as decolorizing agent to refine crude wood rosin. During World War II purification technology for butadiene was developed in the United States for the manufacture of synthetic rubber. By extractive distillation with furfural, butadiene and isoprene can be separated from saturated C_4 and C_5 hydrocarbons, respectively. Detailed information about solvent uses of furfural can be found in [1].

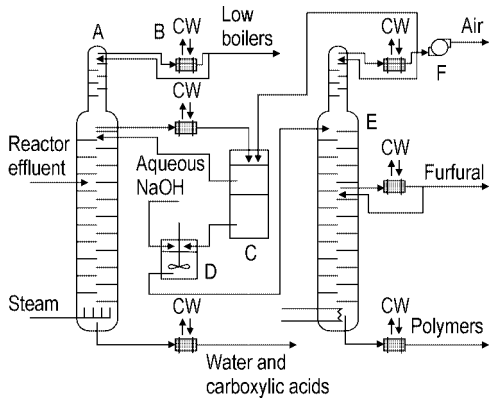

Figure 3. Furfural distillation plant
A) Azeotropic distillation column; B) Heat exchangers (CW = cooling water); C) Static decanter; D) Stirred neutralizer; E) Vacuum distillation column; F) Vacuum pump

Furfural is unusually effective as a solvent for phenolic resins. In the manufacture of abrasive wheels, brake linings, and refractory products for the steel industry furfural is known as a reactive solvent and excellent wetting agent.

Next to the production of furfuryl alcohol, furfural is a chemical feedstock for the production of other 5-membered oxygen-containing heterocycles such as furan, methylfuran, acetylfuran, furfurylamine, and furoic acid.

Table 4. Evolution of the furfural production in Asia, South and North America, and Europe excluding the former USSR

	Production, 10^3 t/a					
Country	**1987**	**1991**	**1994**	**1999**	**2001**	**2003**
China	21.7	23	45	105	100	120
Thailand		1.8	4	9	9	7
India	1.8	1.4	3			
Argentina	1.8	4.5	4	4	4	3
Brazil	3.2					
Dominican Republic	33.5	31	35	35	35	35
Mexico	2.7	1.8				
USA	49.4	49.0	37.8	26	8	6
Tunisia	1.4	0.5			0.5	
South Africa	13.6	13	12	18	18	20
Austria	1.8	2–3	2–3	4	5	4
France	8.2					
Italy	4.1	4.5				
Spain	2.3	2.7	2	2–3	2.7	3
Slovenia	0.5	1.4	1	2	2	1.5
Total	146	138	147	206	194	200

In South Africa furfural has recently been introduced by Illovo Sugar Ltd as an active ingredient in Crop Guard to control plant-parasitic nematodes in the soil. Existing nematicides are effective but generally more toxic (e.g., methyl bromide). Extensive research has shown that Crop Guard is selective in suppressing parasitic nematodes. Crop Guard reacts with the cuticle of the nematode. In this way nematodes are immobilized and destroyed. Crop Guard finds application for the cultivation of tomatoes, groundnuts, peppers, and maize [74].

Economic Aspects. In 2003, world production of furfural excluding the former USSR was 200 000 t. The last two decades have witnessed an enormous expansion of furfural producing facilities in the China. Today, small furfural producers in eight different provinces represent half of world capacity. In the USA, too-high manufacturing costs rendered furfural production economically impossible. The last furfural plant, at Cedar Rapids, was closed in 2004. Nowadays, furfural activities are totally concentrated in developing countries with cheap labor, favorable climate, and fertile land. The largest single-site producer is located in the Dominican Republic. The evolution of furfural production in Asia, South and North America, and Europe excluding the former USSR is presented in Table 4.

2.5. Quality Specifications, Transportation, Toxicology

Quality Specifications. On the international market furfural must meet the specifications listed in Table 5.

Table 5. Furfural specifications on the international market

Furfural (GC assay) wt %	99 (min.)
Moisture, wt %	0.2 (max.)
Acidity, meq/L	20 (max.)
Residue, wt%	0.5 (max.)

Transportation. Furfural is shipped in tankers, steel tank trucks, aluminum tank trucks, ISO containers, IBCs, and steel drums. Drum lots may be stored for months without appreciable change in physical properties. Bulk storage under nitrogen is always recommended to prevent darkening of furfural. Due to its relatively high boiling point (161.7°C), it is difficult to build up substantial concentrations of furfural in the atmosphere. If ignited, the fire can be controlled by the use of water, foam, carbon dioxide, or dry chemicals. Furfural should not be mixed with strong acids or alkalis nor stored near strong oxidizing chemicals

Toxicology. Furfural is virtually ubiquitous in nature; together with furfuryl alcohol, it occurs naturally in many fruits and in tea, coffee (55–255 mg/kg), cocoa and alcoholic beverages

(1–33 mg/kg), and whole grain bread (up to 26 mg/kg). Furfural is listed in the register of chemically defined flavoring substances (EC, 2002).

Table 6. Toxicological data for furfural

LD_{50} (oral), mg/kg	400–500 (mouse); 50–149 (rat)[a]; 650–950 (dog); 541 (guinea pigs)
LD_{50} (i.v.), mg/kg	152 (mouse); 250 (dog)
LD_{50} (s.c.), mg/kg	148 (rat); 119–223 (rabbit); 214–850 (dog)
LD_{50} (dermal), mg/kg	>310 (rabbit); < 10 000 (guinea pigs)
LD_{50} (i.p.), mg/kg	102 (mouse); 20–121 (rat)
LC_{50} (inhalation, 1-h exposure), mg/m^3	4075 (rat)
OEL (8 h, TWA), mg/m^3	8

[a] More recent rat studies indicate that the acute oral LD_{50} is to be found at the higher end of this range.

Table 7. Physical properties of furfuryl alcohol

M_r	98.10
bp at 101.3 kPa (1 atm), °C	170
fp, °C	
metastable crystalline form	−29
stable crystalline form	−14.6
n_D^{20}	1.4868
n_D^{25}	1.4843
ρ at 20°C, g/cm^3	1.1285
Vapor density (air=1)	3.4
Critical pressure P_c, MPa	5.35
Critical temperature T_c, °C	359
Solubility in, wt%	
water	∞
ethyl alcohol, diethyl ether	∞
Spectroscopic polarity (E_T^N)	0.605
Heat of vaporization (liquid), kJ/mol	50.1
Heat capacity (liquid, 25°C), J g^{-1} K^{-1}	2.10
Heat of combustion (liquid), kJ/mol	2548
ΔH_f (l), kJ/mol	−276.23
ΔH_f (g), kJ/mol	−211.80
Viscosity, mPa·s, 25°C	4.62
Surface tension at 25°C, mN/m	38.2
Explosion limits (in air), vol%	1.8/16.3
Flash point, °C, tag closed cup	65
Autoignition temperature, °C	391

Furfural is harmful in contact with skin, toxic by inhalation or swallowing, irritating to eyes and respiratory system, but not sensitizing to skin. There is only limited evidence of any carcinogenic effect. In an in vivo gene mutation study on transgenic Muta-mice, furfural was not found to be a direct-acting carcinogen. Orally applied furfural was unable to induce gene mutations in vivo [75]. Furfural is not toxic to reproduction.

In rats and mice furfural is rapidly metabolized; the major part is oxidized to furoic acid, which is excreted either free or conjugated with glycine (in humans). The smaller part condenses with acetic acid [76]. Detailed toxicological data are summarized in Table 6. Furfural is readily biodegradable in both aerobic and anaerobic degradation tests.

3. Furfuryl Alcohol

3.1. Properties

3.1.1. Physical Properties

Furfuryl alcohol [*98-00-0*], furanol, 2-furanmethanol, 2-(hydroxymethyl)furan, $C_5H_6O_2$, M_r 98.10, is a mobile colorless or pale yellow liquid with mild odor. Furfuryl alcohol forms an azeotrope with water at atmospheric pressure (80 wt % water, *bp* 98.5 °C). General physical properties of furfuryl alcohol are given in Table 7. Equations for calculation of some physicochemical properties are presented in Table 8.

3.1.2. Chemical Properties

Furfuryl alcohol can be derivatized by alkylation, cycloaddition, (trans)esterification, ether-

Table 8. Equations for calculation of physical properties of furfuryl alcohol.

Physical property	Equation	Constants
Vapor pressure	$p_i = \exp(a + b/T + c\ln(T) + dT^2)$ (p_i in Pa and T in K)	$a = 77.274$ $b = -8715.5$ $c = -7.9113$ $d = 4.7922 \times 10^{-3}$
Surface tension	$\sigma = a.(1 - T)^b$ (σ in N/m and T in K)	$a = 0.07217$ $b = 1.006$
Heat capacity at constant pressure		
liquid	$C_p = a + bT$	$a = 101.2$ $b = 0.337$
vapor	$C_p = a + b[(c/T)/\sinh(c/T)]^2 + d[(e/T)/\cosh(e/T)]$ (C_p in J/mol.K and T in K)	$a = 169.17$ $b = 117.3$ $c = 1229$ $d = -78470$ $e = 7.742$
Enthalpy of evaporation	$\Delta H_{vap} = a.(1\text{-}T)^b$ (ΔH_{vap} in J/mol and T in K)	$a = 68820$ $b = 0.2876$
Viscosity liquid	$\mu = \exp(a + bT + cT^2)$	$a = -37.008$ $b = 3555$ $c = 3.4664$
vapor	$\mu = (a \cdot T^b)/(1 + c/T)$ (μ in Pa·s and T in K)	$a = 2.9819 \times 10^{-7}$ $b = 0.6658$ $c = 233.43$

ification, hydrogenation and hydrogenolysis, oxidation, and halogenation reactions.

Alkylation. Furfuryl alcohol is generally alkylated at the 5-position when it undergoes electrophilic aromatic substitution. Under acid conditions, furfuryl alcohol is hydroxymethylated to 2,5-bis(hydroxymethyl)furan with formaldehyde.

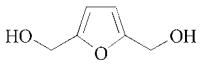

2,5-Bis(hydroxymethyl)furan

At low molar formaldehyde/furfuryl alcohol ratios (2–5) this reaction can be catalyzed by homogeneous carboxylic acids [77, 78] and by acidic ion-exchange resins [79], though the selectivity towards 2,5-bis(hydroxymethyl)furan is low. At high molar ratios (> 110) the reaction can be catalyzed with high yield using dealuminated H-mordenite zeolites [80, 81]. This zeolite-based process was patented in 1995 [82].

Furfuryl alcohol can also undergo Mannich-type reactions. In the presence of formaldehyde and dimethylamine, 5-(dimethylaminomethyl)furfuryl alcohol can be prepared with 70 % yield. 5-(Dimethylaminomethyl)furfuryl alcohol is the precursor for the production of ranitidine, an anti-ulcer drug. This reaction can be performed with or without acetic acid as catalyst at moderate temperature (70–100 °C) [83, 84].

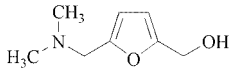

5-(Dimethylaminomethyl)furfuryl alcohol

Cycloadditions. Furfuryl alcohol readily undergoes Diels–Alder cycloaddition with dienophiles like maleic anhydride. In the case of maleic anhydride the reaction solvent determines whether the normal Diels–Alder (NDA) product or the derived lactone is formed [85–87].

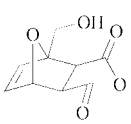

NDA

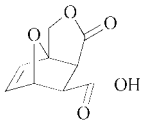

Lactone

Self-Resinification. The primary alcohol function of furfuryl alcohol can be protonated to form an electrophilic carbocation, which can attack another furfuryl alcohol molecule in an electrophilic aromatic substitution reaction to yield polyfurfuryl alcohol.

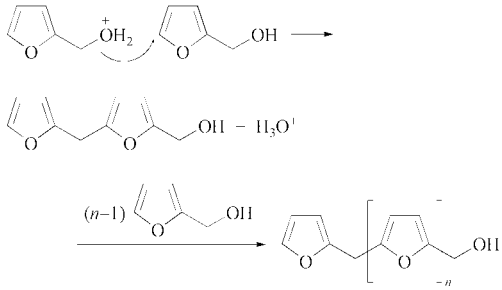

Polyfurfuryl alcohol is a black, highly cross-linked thermoset polymer which is of great interest in the foundry industry. To avoid a violent explosive reaction due to exothermicity, the polymerization should not be carried out with strong Brønsted or Lewis acids. Belgacem and Gandini elucidated the complicated structure of polyfurfuryl alcohol [88, 89]. Based on reactions with model compounds mimicking properties of monomer and oligomer fractions, they proposed the formation of highly conjugated structures by consecutive hydride and proton abstractions from the polyfurfuryl alcohol chain. These conjugated systems are responsible for chromophore formation and are believed to be the active species in cross-linking by Diels–Alder-type reactions (Fig. 4).

Esterification (Transesterification). In view of the acid sensitivity of furfuryl alcohol, base-catalyzed transesterification is preferred to prepare furfuryl esters. Nevertheless, recently developed Lewis and Brønsted acid catalysts have been found useful for the synthesis of furfuryl esters in esterification and transesterification reactions. β-Keto esters can be prepared by using montmorillonite K10 clay or yttria–zirconia as transesterification catalysts [90, 91]. Furfuryl alcohol can be acetylated with acetic anhydride in presence of $Bi(OTf)_3$ [92] and $LiClO_4$ [93], whereas acetic acid can be used as acyl donor in an esterification catalyzed by yttria–zirconia [94].

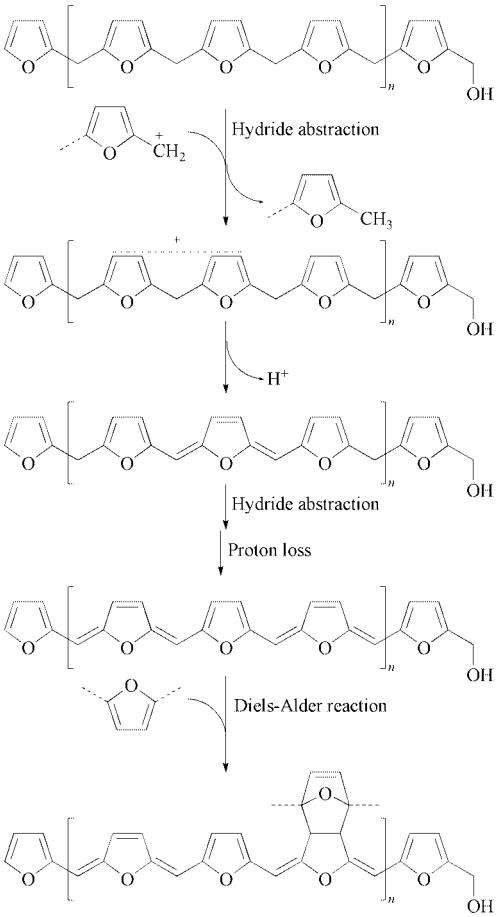

Figure 4. Mechanisms of chromophore formation and cross-linking during the self-resinification of furfuryl alcohol

Etherification. Furfuryl ethers can be formed by normal Williamson ether synthesis in strongly basic media. This method is frequently used in organic synthesis [95–99]. Tetrahydropyranyl ethers of furfuryl alcohol can be formed with 3,4-dihydro-2*H*-pyran in the presence of a catalyst such as HY zeolite [100] , $LiClO_4$ [101], or $LiBF_4$ [102]. For the preparation of 3-(2-furfuryloxy)propionitrile by addition of furfuryl alcohol to the unsaturated double bond of acrylonitrile, the use of a strong basic catalyst has been patented [103]. In the presence of ethylene oxide, furfuryl alcohol can be ethoxylated to yield high-boiling, water-soluble ethers.

Hydrolysis. Levulinic esters can be prepared in high yield from furfuryl alcohol in dilute alcoholic HCl mixtures [104].

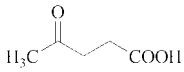

Levulinic acid

Pure levulinic acid can be obtained by subsequent hydrolysis of the esters. Methods for direct production of levulinic acid (4-oxopentanoic acid) from furfuryl alcohol have also been patented [105, 106]. Based on ^{13}C NMR spectroscopy a possible mechanism of levulinic acid formation from furfuryl alcohol has been proposed [70].

Hydrogenation and Hydrogenolysis. Metal-catalyzed hydrogenation of furfuryl alcohol can yield various products (Scheme 2). Of these, 2-methylfuran and tetrahydrofurfuryl alcohol are the most important compounds. In view of their economic importance, the production of tetrahydrofurfuryl alcohol and 2-methylfuran from furfuryl alcohol and other feedstocks is discussed in Chapters 5 and Chap. 8, respectively.

Open-chain compounds like 1,2- and 1,5-pentanediol are formed at elevated temperatures by hydrogenolysis of the furan ring with a Cu chromite catalyst [107, 108]. Low yields of 1,2,5-pentanetriol can be obtained in the presence of acid and a Raney Ni catalyst [109, 110].

Oxidation. Furfuryl alcohol can be oxidized to maleic anhydride and CO_2 in an O_2 flow over a V_2O_5 catalyst at 320 °C [111, 112]. Recent discoveries in oxidation chemistry have led to selective reactions for the oxidation of furfuryl alcohol to furfural [113, 114] and the versatile organic synthon 6-hydroxy-2*H*-pyran-3(6*H*)-one [115–118].

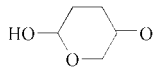

6-Hydroxy-2*H*-pyran-3(6*H*)-one

A titanium-substituted silicalite (TS-1) was reported for the heterogeneously catalyzed oxidation of furfuryl alcohol to 6-hydroxy-2*H*-pyran-3(6*H*)-one with H_2O_2 as oxidant [119].

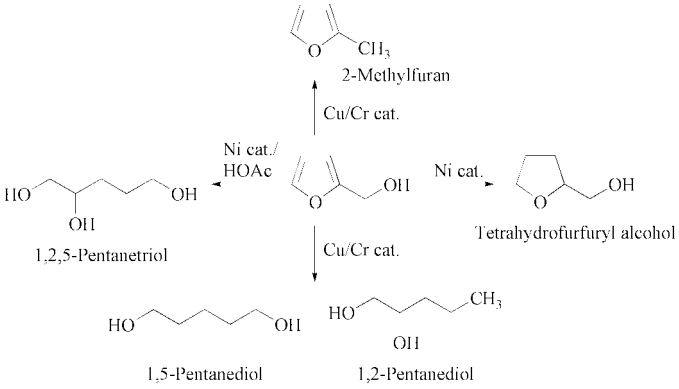

Scheme 2. Hydrogenation of furfuryl alcohol

Halogenation. Furfuryl halides can be synthesized at moderate temperatures (0–25 °C) by use of $SOCl_2$, PBr_3, and PI_3 to obtain respectively furfuryl choride [120, 121], furfuryl bromide, and furfuryl iodide [122].

3.2. Resources and Raw Materials

Furfuryl alcohol is produced by catalytic hydrogenation of furfural.

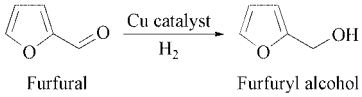

For successful furfuryl alcohol production, a minimal furfural purity of 99.5 % is required. The maximum water content allowed is 0.2 wt % and the acidity should be below 0.02 eq/L. On vacuum distillation of furfural the solid residue should be less than 0.3–0.5 wt %.

3.3. Production

Hydrogenation of furfural to furfuryl alcohol with hydrogen is catalyzed by a reduced Cu catalyst. The composition of this catalyst is of utmost importance for selective hydrogenation of furfural without formation of side products or coke and without rapid catalyst deactivation. Furfuryl alcohol can be prepared from furfural by vapor-phase reaction or by liquid-phase reaction. Both routes are economically viable, though the vapor-phase reaction is currently most widely employed in industry. Reactions in the vapor phase are usually carried out at atmospheric pressure; hence, reactor investment is quite low compared with the high-pressure equipment needed for liquid-phase hydrogenation. On the other hand, vapor-phase hydrogenation uses high molar H_2/furfural ratios; hence, large amounts of H_2 must be circulated.

Liquid-Phase Production of Furfuryl Alcohol. Only small-scale manufacturers produce furfuryl alcohol by liquid-phase furfural reduction. The reactions are carried out in stirred tank reactors rather than in tubular plug-flow reactors.

The hydrogenation of furfural in the liquid phase was first reported in the patent literature by the Quaker Oats Company in 1928 [123]; a precipitated Ni on MgO catalyst was used. Because of subsequent hydrogenation of the aromatic ring, delicate process control was needed to prevent formation of tetrahydrofurfuryl alcohol (THFA). THFA formation could be circumvented by use of a precipitated Cu chromite catalyst at temperatures between 100 and 160 °C and H_2 pressures of 10–20 MPa. Under these conditions furfuryl alcohol yields of 95–98% can be obtained in solvent-free batch furfural hydrogenation. Reaction temperatures higher than 175 °C should be avoided in order to prevent formation of substantial amounts of resinous products and methylfuran. This method was patented by Rohm & Haas [124] and by Quaker Oats [125]. The selectivity of Cu-chromite-catalyzed hydrogenation can be further increased by addition of CaO to the reaction mixture to prevent acid-catalyzed resinification and etherification. With CaO, the reaction temperature can be increased to 180 °C without loss of furfuryl alcohol yield. At this high temperature an H_2 pres-

sure of 3 MPa is sufficient for successful hydrogenation; a furfural conversion and furfuryl alcohol selectivity of respectively 99.9 and 99.6 % were reported by Ashland Oil [126, 127]. Due to environmental considerations and costs associated with chromium salts, chromium-free Cu hydrogenation catalysts have been developed. CuFe on Al_2O_3 and Cu on ZnO or MgO proved to be efficient catalysts [128, 129].

Besides supported Cu catalysts, Ni-based catalysts have been recently reevaluated for furfuryl alcohol production. With Raney Ni impregnated with Mo polyacids [130, 131] and with Ni–P–B [132] or Ni–Fe–B amorphous alloy catalysts [133], THFA formation at high furfural conversion is significantly reduced in comparison with unmodified Ni catalysts.

The reaction heat for furfural hydrogenation is 63 kJ/mol. Particularly in a continuous fixed-bed reactor, special care must be taken to dissipate the heat. This can be achieved by diluting the catalyst bed with unreactive ceramics [134] or by diluting the furfural feed with furfuryl alcohol [128]. The kinetics of Cu-chromite-catalyzed furfural reduction have been reported [135].

Vapor-Phase Production of Furfuryl Alcohol. In all large furfuryl alcohol production plants, furfural is currently hydrogenated in the vapor phase. Gas-phase reduction of furfural on a supported Cu catalyst at 140 °C was first reported in 1927 [136]. The use of Cu chromite catalysts in the vapor-phase reaction was patented in 1931 by Du Pont de Nemours [137]. To compensate for the loss of activity of the Cu catalyst after a certain time on stream, the temperature of the catalyst bed is gradually increased. At temperatures between 155 and 180 °C, significant amounts of 2-methylfuran are produced. This problem was circumvented by Quaker Oats by the use of a reduced Cu on $Na_2O \cdot x\ SiO_2$ catalyst [138]. The use of basic sodium silicate lowers the hydrogenolysis activity towards 2-methylfuran, and furfuryl alcohol yields of 99 % are obtained in the temperature range between 132 and 177 °C. The $Cu/Na_2O \cdot x\ SiO_2$ catalyst can also be deposited on high surface area carriers [139, 140]. Use of Ca silicate instead of Na silicate leads to formation of basic oxides at the catalyst surface [141]. In the open literature, Pt catalysts on various supports have been reported for the vapor-phase reduction of furfural [142, 143].

Figure 5 presents the process diagram of a furfuryl alcohol production unit. Furfural is vaporized in a stream of hydrogen in a tubular vaporizer at 120°C (A). After demisting (B), this stream is superheated before it enters the oil-heated multitubular reactor (D). The liquid furfural from the demister pad is recovered in tank C. After passage through the reactor furfural conversion exceeds 99 %. The crude furfuryl alcohol is condensed from the reactor effluent and collected in tank E. The hydrogen stream is passed through another demister (F) before it is recycled to the vaporizer (A). Some of the hydrogen is continuously vented as a bleed stream to prevent accumulation of impurities. The crude furfuryl alcohol from tank E is reheated and distilled in a vacuum column (G) to obtain a furfuryl alcohol purity of more than 99 %. The bottom fraction of the vacuum column consists of high-boiling oligomers, whereas water and methylfuran are withdrawn from the top of the column.

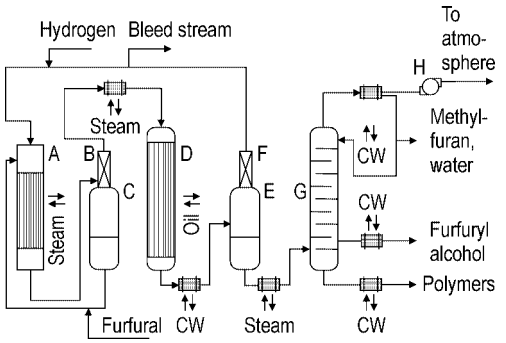

Figure 5. Furfuryl alcohol production plant
A) Tubular vaporizer; B) Demister; C) Furfural recovery tank; D) Tubular reactor; E) Crude furfuryl alcohol tank; F) Demister; G) Vacuum distillation column; H) Vacuum pump; CW = cooling water

Production of furfuryl alcohol starts at a reactor temperature of 120 °C, which is maintained until furfural conversion drops below 99 %. To maintain the rate, the catalyst bed temperature is then gradually increased to 160–170 °C. For catalyst regeneration, the temperature of the reactor is increased to 220 °C and coke deposits are removed by oxidation. After pre-reduction at 160°C, the reaction with the Cu catalyst is restarted at 120 °C.

Table 9. Furfuryl alcohol production from 1987 until 2003 in Asia, South and North America, and Europe excluding the former USSR

Country	Production, 10^3 t/a						Hydrogenation
	1987	1991	1994	1998	1999	2003	
China	1.8	16.7	18	52	50	70	liquid/gas
Korea (South)		2.7	5				
Thailand		2.3	5	15	15	13	gas
Japan	9.5–12.7	11.3	6.2				
Argentina					2	2	
Brazil	1.8	0.9	2				
USA	24.5	25	25	25	18	5	gas
South Africa	11.3	10.4	8.5	11	11	11	gas
Belgium	32	32	32	32	32	32	gas
France	7.3	4.1					
Hungary		0.5					
Turkey		0.5		2	2–3		liquid (slurry)
Total	88–92	107	103	138	132	133	

3.4. Uses and Economic Aspects

Uses. Furfuryl alcohol plays a vital role in the production of foundry sand binders (→ Tetrahydrofuran). For over thirty years it has been extensively used to produce high-quality cores and molds for metal casting. The major part of furfuryl alcohol is purchased by foundry binder suppliers.

Furfuryl alcohol is the major ingredient in furan foundry binders. Furan is a generic term for binders containing furfuryl alcohol and either urea or phenol–formaldehyde or mixtures of both. Today furfuryl alcohol is used in binders for furan hot-box, warm-box, and gas-hardened processes, and especially in the traditional Furan no-bake system.

The *furan no-bake process* was introduced in 1958. It is suitable for making all types of metal castings in all sizes, and is particularly used for the production of molds and larger cores. This acid-catalyzed cold-setting binder consists of hardening catalysts such as sulfuric acid, sulfonic acids, or phosphoric acid and a reactive furan-type resin which contains typically 75 wt % furfuryl alcohol. The furan no-bake process is known for its superior shake-out characteristics, and the sand can be reused by thermal and/or mechanical reclamation [144].

The *furan hot-box process* uses furan resins in combination with a latent acid catalyst. The *warm-box* process is operated at lower temperatures and was developed by the Quaker Oats Company for the rapid production of cores in existing hot-box equipment. The binder for this process contains more furfuryl alcohol than hot-box furan binders. A copper salt catalyst is used to cure the binder very rapidly on heating.

The *furan SO_2 process* is a gas-cured binder system for the rapid production of small molds and cores. Curing of the furanic resin occurs immediately when the sand mix is exposed to SO_2 gas at room temperature.

Apart from its use in foundry resins, furfuryl alcohol finds application in corrosion-resistant fiber-reinforced plastics, in corrosion-resistant cements and mortars, in wood protection [145, 146], as reactive solvent for phenolic resins in the refractory industry, in paint stripper and cleaning compound formulation, as chemical building block for drug synthesis (ranitidine), as viscosity reducer for epoxy resins, and as chemical feedstock for tetrahydrofurfuryl alcohol and 3,4-dihydro-2*H*-pyran.

Economic Aspects. Worldwide production of furfuryl alcohol increased significantly since the mid-1980s due to the great expansion of furfuryl alcohol production capacity in China. The largest furfuryl alcohol production plant is located in Belgium and produces 32×10^3 t/a. Smaller scale producers are located in Thailand and South Africa. In the USA production decreased fivefold since ca.1990. In 2003 wordwide production amounted to 133×10^3 t. Annual worldwide demand is estimated at (120–140) $\times 10^3$ t. Table 9 summarizes furfuryl alcohol production from 1987 to 2003 in Asia, South and North America, and Europe excluding the former USSR

3.5. Quality Specifications, Transportation, Toxicology

Quality Specifications. On the international market furfuryl alcohol and high-purity (HP) furfuryl alcohol must meet the specifications listed in Table 10.

Table 10. International market specifications for furfuryl alcohol and high-purity (HP) furfuryl alcohol

	Furfuryl alcohol	HP furfuryl alcohol
Furfuryl alcohol (GC assay), wt %	98.0 (min.)	99.0 (min.)
Furfural, wt %	0.7 (max.)	0.1 (max.)
Moisture, wt %	0.3 (max.)	–
Color, Gardner		1 (max.)
Cloud point, °C	10 (max.)	10 (max.)

Transportation. Furfuryl alcohol is shipped in steel or aluminum tanks or drums, ISO containers, and IBCs. All should be clean and must not contain any traces of acidic materials to prevent acid resinification of furfuryl alcohol, which can be violent. The use of hoses made of cross-linked polyethylene, fluorocarbons, ethylene–propylene rubber, or styrene–butadiene rubber is recommended for transferring furfuryl alcohol.

Containers for furfuryl alcohol storage need phenolic resin coating. Lacquer, vanish, and epoxy-resin coatings are not acceptable. Containers need to be tightly closed and stored away from open flames, acids, and strong oxidizers. The buildup of substantial concentrations of furfuryl alcohol in the atmosphere usually does not occur due to the relatively high boiling point (170 °C). Fires can be extinguished with water, foam, carbon dioxide, or dry chemical.

Toxicology. Furfuryl alcohol is harmful by inhalation, if swallowed, and in contact with skin. In case of acute overexposure furfuryl alcohol is irritating to eyes, skin, and/or respiratory tract. Furfuryl alcohol does not have genotoxic potential (in vivo); evidence of carcinogenic activity of furfuryl alcohol in rats is inconclusive.

LD_{50} values are 160 mg/kg (mouse, oral), 275 mg/kg (rat, oral), 400 mg/kg (rabbit, dermal) The LC_{50} has been determined as 592 ppm (rat, inhalation, 1 h). The occupational exposure limit is 10 ppm (8 h, TWA). The MAK value is 41 mg/m^3 (10 ppm). Furfuryl alcohol is readily biodegradable in aquatic environment. Detailed information on toxicity of furfuryl alcohol can be found in [147].

4. Furan

4.1. Properties

Physical Properties. Furan [*110-00-9*], furfuran, oxole, C_4H_4O, M_r 68.07, is a colorless

Table 11. Physical properties of specialty chemicals derived from furfural and furfuryl alcohol

	M_r	*mp*, °C	*bp*, °C	ρ at 20 °C, g/cm^3	n_D^{20}	Flash point[b], °C	Solubility in H_2O at 25°C, wt %	Vapor pressure, mm Hg	Ignition temperature, °C	LEL and UEL, vol %
Furan [*110-00-9*]	68.07	−85	31.36	0.9378	1.4214	−35.5	1	493 at 20°C	413	2.3/14.3
Tetrahydrofurfuryl alcohol [*97-99-4*]	102.13	<−80	178	1.0511	1.4520	83.9[c]	∞	0.22 at 20°C	282	1.5/9.7
3,4-Dihydro-2*H*-pyran [*110-87-2*]	84.1	−70	85	0.923	1.4380	−17.8	1.6	90 at 20°C	240	
Furfurylamine [*617-89-0*]	97.12	−75	145	1.057	1.4908	37	∞	10 at 49–50°C	390	2.0/15.0
Tetrahydrofurfuryl amine [*4795-29-3*]	101.15		153	0.98	1.4551	43	∞		270	1.6/14.1
2-Methylfuran [*534-22-5*]	82.1	−88.7	63.2–65.6	0.913	1.4320	−27	<0.3	122 at 20°C	450	1.9/14.0
2-Methyltetrahydrofuran [*96-47-9*]	86.13	−136	80.2	0.854	1.4025	−11	15.1	102 at 20°C	270	1.5/8.9
Furoic acid [*88-14-2*]	112.08	133	231 (158)[a]	0.55	/	137	5.8 at 30°C, 1.8 at 96°C 335 at 80°C			

[a] Decarboxylation temperature.
[b] Tag, closed cup.
[c] Tag, open cup.

liquid with strong ethereal odor. It is low boiling and highly flammable. Some physical properties of furan are listed in Table 11.

Chemical Properties. Furan or oxole is the parent compound of the five-membered heterocycles containing one oxygen atom. The furan ring is a cyclic, fixed cisoid dienic ether and is stabilized by resonance. The aromatic properties of the furan ring are evident from its orbital structure. In the planar pentagonal structure each carbon atom has a single electron in its p orbital while the oxygen atom has electron pairs in p and sp^2 orbitasl. The electrons in the p orbital system form a delocalized 6π-electron system which gives the furan ring its aromatic character. The furan ring is less aromatic than benzene, as expressed in the respective resonance energies of 96 and 155 kJ/mol. In comparison with benzene and other heterocycles the aromaticity decreases in the following order.

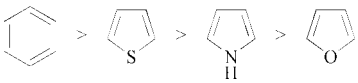

The furan heterocycle is extremely susceptible to electrophilic aromatic substitution. Substitutions take place at the α- rather than at the β-position, as is evident from the resonance structures of the transition states. Hence, bromination, nitration, and alkylation yield respectively 2-bromofuran, 2-nitrofuran, and 2-alkylfurans.

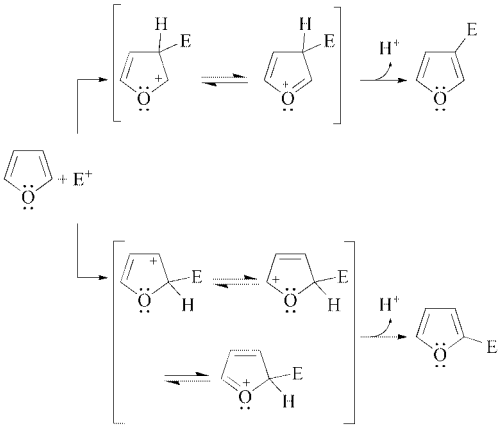

If the furan ring is substituted with electron-donating substituents, electrophilic aromatic substitution reactions are accelerated. In the furan series the following reactivity towards electrophilic aromatic substitution is observed:

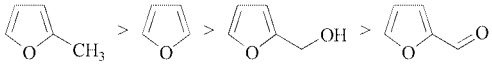

On heating in the presence of mineral acids furan and alkylfurans are hydrolytically opened to diketones. Furans with electron-withdrawing substituents are quite stable to hydrolytic cleavage.

At elevated temperature in the presence of steam and NH_3 or H_2S, furan and alkylfurans can undergo O/N and O/S exchange reactions leading to pyrrole and thiophene derivatives.

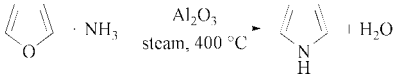

If the furan ring is not substituted with electron-withdrawing groups, it readily undergoes Diels–Alder reactions with activated dienophiles like maleic acid or anhydride.

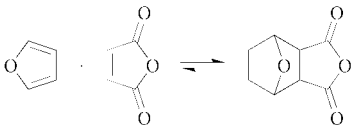

Hydrogenation of furan to tetrahydrofuran is preferably carried out with a supported Ni catalyst. Supported noble metals like Ru and Pd, and to lesser extent Pt, catalyze both ring hydrogenation and ring hydrogenolysis reactions. The ratio between hydrogenolysis and hydrogenation products increases with increasing temperature.

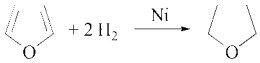

Under mild oxidative conditions, e.g. with aqueous *m*-chloroperbenzoic acid (mCBPA) at room temperature, furan is oxidized to malealdehyde (or fumaraldehyde). At high temperatures in the presence of V_2O_5 or MoO_3 catalysts, furan is oxidized to maleic anhydride. These high temperature oxidations can also be carried out

with furoic acid, furfural, and furfuryl alcohol to produce maleic anhydride and CO_2.

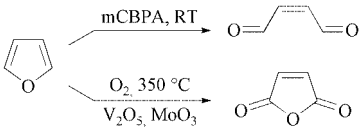

Though the production of tetrahydrofuran via catalytic furan hydrogenation seems logical, this route is considered as less economical in comparison to tetrahydrofuran manufacture via maleic anhydride hydrogenation or the Reppe process. Nevertheless, furan hydrogenation to tetrahydrofuran has recently been reintroduced in China.

α-Acetylfuran, the precursor for the antibiotic Cefuroxime, is a common impurity in commercial furfural, but the costs for its recovery from crude furfural are economically prohibitive. a-Acetylfuran can be prepared by Friedel–Crafts acylation of furan with acetic anhydride. This reaction can be carried out in the liquid phase without solvent at 20–60°C with homogeneous catalysts like $ZnCl_2$ [148], H_3PO_4 [149], HI [150], and BF_3 [151]. In 1986 BASF patented the vapor-phase acylation of furan with Ce- and Cs-exchanged borosilicate zeolites. At furan conversions of 10–20 % selectivity for α-acetylfuran is up to 99 % at reaction temperatures of 150–200 °C [152].

2,2-Difurylpropane is obtained by electrophilic aroamtic substitution of furan with acetone.

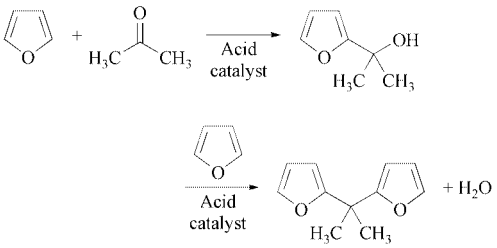

2,2-Difurylpropane and 2,2-bis(tetrahydrofurfuryl)propane are currently used as rubber modifiers.

2,5-Dimethoxy-2,5-dihydrofuran, a versatile organic synthon, is prepared by bromine or chlorine oxidation of furan in methanol. If Na_2CO_3 is added to neutralize the produced HCl, a yield of 50 % can be obtained [153]. In the presence of Br_2, reported yields are somewhat higher [154]. Bromine or chlorine can also be generated in situ by electrolysis of an appropriate halide electrolyte. This results in the following overall reaction equation:

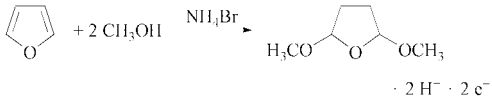

The yields obtained by electrolytic oxidation are generally in the range of 80–90 % [155–157]. Reduction leads to 2,5-dimethoxytetrahydrofuran, which can be used for the production of pyrrole derivatives [158].

Direct production of pyrrole from furan and NH_3 over a γ-Al_2O_3 catalyst in the presence of steam at 400 °C was published by Du Pont in 1952 [159]. Later patents by Japanese companies reported the use of alkali metal or alkaline earth metal exchanged amorphous silica alumina [160] or zeolites [161]. Exchanged zeolites are not only useful for pyrrole formation, but also for reaction of furan with H_2S to give thiophene [162–164].

4.2. Production

Furan is produced by Pd-catalyzed decarbonylation of furfural. It can also be manufactured by partial oxidation of 1,3-butadiene.

Decarbonylation of Furfural. Decarbonylation of furfural can be carried out via two catalytic routes. The oldest method is the decomposition of furfural into furan, H_2, and CO_2 in the presence of steam, catalyzed by Zn Fe or Zn Mn chromite [165, 166]. A Zn Fe chromite catalyst with a Zn chromite to Fe chromite ratio of 9:1 is reported to be the best performing material.

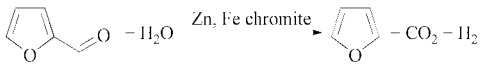

In a typical reaction, steam and furfural in a molar ratio of 1:5–1:10 are converted over the chromite catalyst at 400 °C with a GHSV of 500–1000 h^{-1}. Furfural to furan decarbonylation yields of more 90 % were claimed at total conversion of furfural. Addition of 1 wt % alkali

metal carbonate promoter to the catalyst matrix has a beneficial effect on the selectivity towards furan. Zn molybdite catalysts have also been reported to be active but furan yields are inferior in comparison with the chromite catalysts [167].

In 1959 Du Pont disclosed a Pd-catalyzed route to furan from furfural in the absence of steam. Furfural decomposes on the Pd surface to furan and CO.

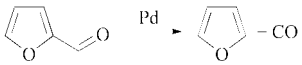

In a typical procedure furfural is refluxed under pressure (200 °C) in the presence of a supported Pd catalyst in the presence of alkali metal carbonate. Furan and CO are continuously removed from the reaction mixture by distillation. A typical productivity of 4–10 kg of furan per gram of Pd can be obtained [168]. When the alkali metal carbonate is replaced with alkali metal acetate, the base-catalyzed resinification of furfural can be minimized [169]. Irradiative pretreatment (230–270 nm) of the Pd catalyst is reported to have a beneficial effect on the catalyst lifetime [170, 171]. The Pd-catalyzed decarbonylation can also be performed in the vapor phase [172]. In a typical example furfural is fed to a 0.3 wt % Pd on Al_2O_3 catalyst with a furfural WHSV of 230 h^{-1} in the presence of a molar furfural/H_2 ratio close to one. At 300°C, a typical furan yield of 95 % can be obtained. Typical productivities of 10–15 kg of furan per gram of Pd can be reached before the catalyst needs to be regenerated. By using supported Pt and Rh catalysts these productivities can be increased [173].

Partial Oxidation of 1,3-Butadiene. Partial oxidation of 1,3-butadiene in the vapor phase was first reported by Du Pont in 1957 with MoO_3 and WO_3 catalysts at 500 °C [174]. With the use of V–P oxide catalysts, 1,3-butadiene can be converted to furan in a yield of 7 % in the presence of steam and oxygen (1,3-butadiene GHSV of 100 h^{-1}, 480–590°C) [175]. Addition of promoters like Co, Ni, and Mo [176] and supporting the V P oxides on high surface area carriers results in more active catalysts [177]. Other catalyst compositions comprise Bi–Mo oxide on SiO_2 [178], Sb oxide promoted with Sn, Cu, and W [179], Ag Mo oxide supported on Al_2O_3 [180, 181], and Mo Bi Pb oxide promoted with Cu, V, and Au [182]. Yields of furan generally do not exceed 10–15%.

In 1987 Du Pont issued a low-temperature oxidation process (90–120 °C) for the production of furan from 1,3-butadiene. In this process 1,3-butadiene vapor is oxidized by $CuCl_2$ to furan in an aqueous mixture of $CuCl_2$, CuCl, NaCl, KI, and HCl. On total reduction of $CuCl_2$ to CuCl the 1,3-butadiene flow is interrupted and replaced by an oxygen flow to reoxidize CuCl. Consecutive production and reoxidation cycles render this process catalytic in Cu. Furan can be produced with high selectivity (90 %) at maximum 1,3-butadiene conversion of 20 %. Higher conversions and hence longer residence time result in extensive resinification of the produced furan. This catalyst composition can be used dissolved in an aqueous mixture [183–185] or supported on a solid carrier [186, 187]. Due to the extreme corrosiveness of the reaction mixture, reactors need a glass, carbon, or ceramic lining.

4.3. Uses, Economic Aspects, Transportation, Toxicology

Uses. Furan is mainly used in the preparation of specialty chemicals such as α-acetylfuran, 2,5-dimethoxy-2,5-dihydrofuran, 2,2-difurylpropane, 2,2-di(tetrahydrofurfuryl)propane, and pyrrole derivatives.

Transportation. Due to its low boiling point furan is a potentially hazardous chemical, and to prevent buildup in work areas should be properly vented. Drums or tanks containing furan must be earthed, able to withstand furan vapor pressure at ambient temperature, and stored away from oxidizers, strong acids, and heat sources. In the case of fire or explosion dry chemical foam or carbon dioxide should be utilized as extinguisher medium.

Toxicology. Furan is a narcotic, toxic substance that can be adsorbed though the skin and is suspected to be carcinogenic. LD_{50} (i.p.) values are 7 mg/kg for mice and 5.2 mg/kg for rats. The LC_{50} is 3464 ppm (rat, inhalation, 1 h).

5. Tetrahydrofurfuryl Alcohol

5.1. Properties

Physical Properties. Tetrahydrofurfuryl alcohol [*97-99-4*], 2-(hydroxymethyl)tetrahydrofuran, 2-tetrahydrofuranmethanol, THFA, $C_5H_{10}O_2$, M_r 102.13 is a transparent, mobile, high-boiling liquid with mild odor and is completely miscible with water. Physical properties of THFA can be found in Table 11.

Chemical Properties. Tetrahydrofurfuryl alcohol readily undergoes reactions typical of primary alcohols. Its saturated nature renders it quite stable towards mineral acids. However the tetrahydrofuryl ring can be cleaved by acetolysis or hydrogenolysis reactions. In the presence of $ZnCl_2$ and acetic anhydride, THFA is opened to 1,2,5-triacetoxypentane. Hydrolysis of this triacetate yields 1,2,5-pentanetriol [188, 189].

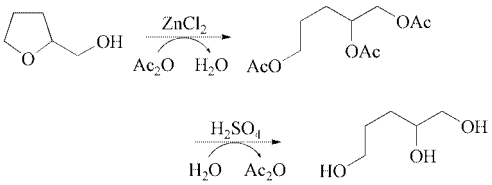

Hydrogenolysis of THFA can be effected by Cu chromite catalysts at high temperatures and pressures with a maximum 1,5-pentanediol yield of 35 %. This reaction is considered commercially uninteresting due to the extreme reaction conditions and only moderate diol yield [190, 191].

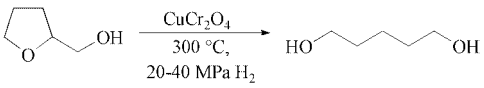

Dehydration of THFA to 3,4-dihydro-2*H*-pyran has been the subject of various patents in the last six decades. The reaction is always carried out in the vapor phase over an acid catalyst supported on alumina [192–198], or an alumina-doped acid oxide such as titania [193]. Reaction temperatures vary between 250 and 400°C with sometimes high yields of 3,4-dihydropyran; high temperatures tend to promote catalyst coking. The state of the art in THFA dehydration to 3,4-dihydro-2*H*-pyran can be found in Table 12.

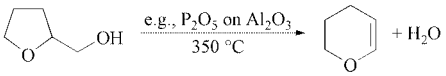

Generally, simple tubular reactors are used, although a fluidized-bed reactor has been applied in one case [197]. Dilution of THFA vapors with a N_2 carrier generally increases 3,4-dihydro-2*H*-pyran yield. In recent patents, optimal operation temperatures are lower than 300 °C [197, 198]. The optimal catalyst composition in the latter patents is not mentioned, but it is likely that the γ-alumina support is loaded with acid metal oxides like MoO_3, TiO_2, or V_2O_5/P_2O_5.

3,4-Dihydro-2*H*-pyran is a key intermediate in the synthesis of 1,5-difunctional open-chain aliphatic compounds. With alcohols, tetrahydropyranyl ethers are formed under mild acidic conditions. These ethers are useful protecting groups for alcohols.

5.2. Production

Tetrahydrofurfuryl alcohol (THFA) can be produced directly from furfural or by the intermediate production of furfuryl alcohol. Supported Ni catalysts are preferred for both furfural and furfuryl alcohol hydrogenation. In early patents, Ni chromite was used for the production of THFA from furfural in the liquid phase, but yields were only moderate (ca. 65 %) [199–201]. With Raney Ni, THFA is formed quite selectively from furfural, but excessive operating pressures and temperatures are needed (25 MPa H_2, 150 °C) [202]. Production of THFA from furfuryl alcohol was patented by Quaker Oats in 1954 [203]. This patent described a vapor-phase process at 80 °C, in which furfuryl alcohol was completely and selectively hydrogenated to THFA by a Ni on kieselguhr catalyst. In today's industrial practice THFA is produced from furfuryl alcohol with various supported Ni catalysts. At moderate temperatures of 50–100°C both vapor- and liquid-phase processes are economically feasible.

5.3. Uses, Economic Aspects, Transportation, Toxicology

Uses. THFA is a water-miscible biodegradable compound and can be used as high-boiling solvent for dyes, printing inks, pesticides and herbicides, as paint-stripping agent, industrial cleaner, and reactive diluent for epoxy resins.

THFA is also used in the preparation of pharmaceuticals. Next to Nafronyl (a vasodilator)

Table 12. State of the art in THFA dehydration to 3,4-dihydro-2*H*-pyran

Patent	Catalyst	*T*	THFA feed	DHP yield
ICI, 1941 [192]	P_2O_5 (30 wt %) on alumina	350 °C	0.5 mL mL^{-1} h^{-1}	64 % at 98 % THFA conversion
Quaker Oats, 1959 [193][a]	TiO_2 (85–95 wt %) and Al_2O_3 (15–5 wt %)	320 °C	0.1–0.07 mL mL^{-1} h^{-1}, N_2 carrier	95 % at 97 % THFA conversion
Olin Mathieson, 1963 [194][b]	MoO_3 (10 wt %) on Alumina	367 °C	1.1 mL mL^{-1} h^{-1}, N_2 carrier	74 %
Olin Mathieson, 1963 [195]	V_2O_5 (10–20 wt %) on alumina	400 °C	1 mL mL^{-1} h^{-1}	56 %
Emery Ind., 1968 [196]	ε-alumina doped with oxides of Mo,V, and Ti	315 °C	6 g $g^{-1}h^{-1}$	90 %
BASF, 1995 [197][c]	γ-alumina doped with non-specified metal oxides	250 °C	0 g $g^{-1}h^{-1}$, N_2 carrier	77 %
BASF, 2001 [198]	γ-alumina doped with unspecified metal oxides	250 °C	0 34 g $g^{-1}h^{-1}$, N_2 carrier	92 %

[a] DHP yield was determined after distillation; molar ratio N_2/THFA=4.2.
[b] Molar ratio N_2/THFA=1.7.
[c] Fluidized-bed reactor, molar ratio N_2/THFA=9.8.

THFA is an intermediate for the preparation of disinfectants and antimetabolite drugs.

3,4-Dihydro-2*H*-pyran and to a lesser extent tetrahydrofurfurylamine (see Chap. 6) are the main applications of THFA in the production of specialty chemicals.

Transportation. THFA is shipped and stored in steel or aluminum containers. Due to its high solvent power, piping is preferably made of steel. If ignited, fire can be extinguished by alcohol foam, CO_2, and dry chemical diluted with water spay. THFA must be stored away from open flames, strong acids, and oxidizers.

Toxicology. THFA is water-soluble, biodegradable compound with an atmospheric half-life of 13 h. Oral LD_{50} in rat, mouse, and guinea pig are respectively 1600, 2300, and 800 mg/kg. The LD_{50} (i.p.) for rats and guinea pigs is 400 mg/kg.

6. Furfurylamine and Tetrahydrofurfurylamine

Furfurylamines are produced by reductive amination of furfural. Supported Ni catalysts [204–206], Raney Ni catalysts [205, 207, 208], and supported Co catalysts [206] have been found to be effective for this reaction. When furfural is contacted with NH_3 the imine is formed. Furfurylimine (**12**), which is unstable, readily undergoes a condensation reaction to give the trimeric hydrofuramide (**13**). Hydrogenation of hydrofuramide yields difurfurylamine (**14**) and furfurylamine (**15**). Undesired formation of difurfurylamine can be avoided by using an excess of NH_3, by using primary amines as cosolvent, or by gradual admission of furfural during reaction. On heating, hydrofuramide is converted to furfurine (**16**). Furfurine is a stable unwanted byproduct which cannot be converted to amines by hydrogenation. If the reductive amination is carried out at temperatures higher than 110 °C, furfurylamines are converted to tetrahydrofurfurylamine (**17**) [209].

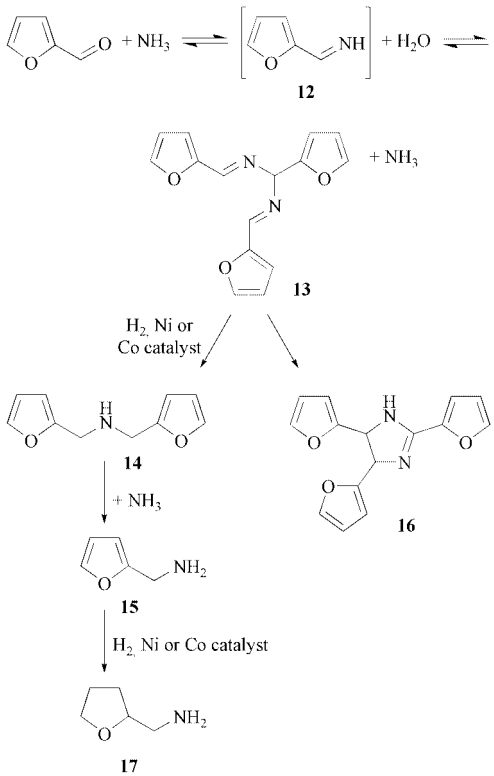

Direct production of tetrahydrofurfurylamine from tetrahydrofurfuryl alcohol with a Raney Ni catalyst has also been reported [210].

At 210 °C and 6 MPa H_2 pressure, piperidine can be formed from furfurylamine or tetrahydrofurfurylamine by using a Co–Re–Mo catalyst [211, 212]. In the case of tetrahydrofurfurylamine conversion and piperidine selectivity exceed 90 %.

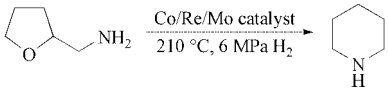

Furfurylamine and tetrahydrofurfurylamine are high-boiling colorless to pale yellow liquids with typical amine odor. Physical properties of these compounds are listed in Table 11.

Furfurylamine and tetrahydrofurfurylamine find application in the synthesis of herbicides, pesticides, pharmaceuticals, and piperidine derivatives. Furfurylamine is a key intermediate for the production of the diuretic furosemide. An LD_{50} of 200 mg/kg (i.p., mouse) is reported for both furfuryl and tetrahydrofurfurylamine.

7. Methylfuran and Methyltetrahydrofuran

2-Methylfuran or sylvan is a common byproduct in the vapor-phase production of furfuryl alcohol from furfural. In this process the Cu or Cu chromite catalysts are slowly deactivated by coking during furfural production at low temperature (130 °C). To maintain the same production rate the reactor temperature is gradually increased to 150 or 160 °C, at which 2-methylfuran is produced by hydrogenolysis. This is the only source of 2-methylfuran in the chemical industry.

Though 2-methylfuran is currently made as a byproduct of furfuryl alcohol manufacture, several attempts have been made towards selective 2-methylfuran production from furfural [213–215] and furfuryl alcohol [216, 217]. Most attempts have focused on the use of Cu chromite catalysts at elevated temperatures (175–200 °C). At these temperatures the catalyst is rapidly deactivated, and this prohibits the commercial implementation of these processes.

2-Methyltetrahydrofuran is produced by Ni-catalyzed hydrogenation of 2-methylfuran. This reaction can be performed in the liquid phase under H_2 pressures of 2 MPa or in the vapor phase at atmospheric pressure with common supported Ni catalysts.

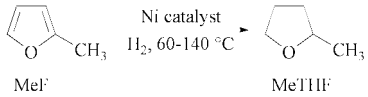

In the future 2-methyltetrahydrofuran could also be produced by cyclization and hydrogenation of levulinic acid.

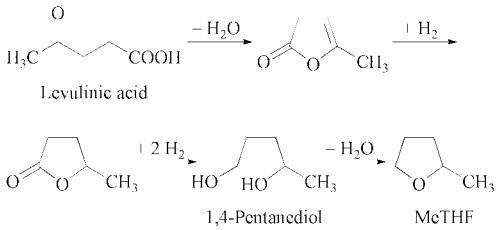

A method was patented for the single-stage production of 2-methyltetrahydrofuran from levulinic acid with a Re/Pd on carbon catalyst [218]. Though these processes have not been brought to industrial practice they hold the potential for the use of hydrolyzed C_6 carbohydrate biomass (i.e., a potential source of levulinic acid) for the production of base bulk chemicals ([218] and references therein).

2-Methylfuran and 2-methyltetrahydrofuran are both colorless mobile liquids with high solvent power. Some of their physical properties are listed in Table 11. 2-Methyltetrahydrofuran is inversely soluble in water, i.e., the solubility decreases with increasing temperature. The chemical properties of 2-methylfuran and 2-methyltetrahydrofuran are comparable to those of furan and tetrahydrofuran, respectively. 2-Methylfuran is used as solvent and as feedstock for the production of antimalarial drugs (chloroquine), methylfurfural, nitrogen and sulfur heterocycles, and functionally substituted aliphatic compounds. 2-Methyltetrahydrofuran is mainly used as a higher boiling substitute for tetrahydrofuran as specialty solvent. 2-Methyltetrahydrofuran can also be used in the electrolyte formulation for secondary lithium electrodes [219] and as a component of alternative fuels [220].

2-Methylfuran and 2-methyltetrahydrofuran are both hazardous highly flammable liquids and should be handled with the same precautions as furan and tetrahydrofuran, respectively. LC_{50} values are 500 ppm (rat, inhalation, 4 h) for 2-methylfuran and 6000 ppm (rat, inhalation, 4 h) for 2-methyltetrahydrofuran. The LD_{50} of 2-methyltetrahydrofuran is 4500 mg/kg (rabbit, dermal).

8. Furoic Acid

Furoic acid can be produced by the Cannizzaro reaction or oxidation of furfural. In the Cannizzaro reaction of furfural, furoic acid salts and furfuryl alcohol are formed in equimolar amounts. When sodium hydroxide is used as base, sodium furoate crystals partially precipitate during the reaction. After filtration and acidification of these salts, crude furoic acid is purified by sublimation. Though the maximum theoretical furoic acid yield is 50 % and product purification is difficult due to the presence of furfuryl alcohol and resinous material, furoic acid is still produced by this route today.

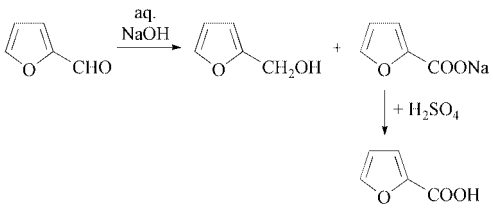

Furoic acid
2-Furancarboxylic acid

In the past, selective oxidation of furfural with molecular oxygen or hydrogen peroxide has been investigated. High yields of furoic acid (salts) can be obtained with these oxidation processes, but they have not been implemented for large-scale furoic acid production.

In the presence of catalytic amounts of silver oxide and oxygen high yields (95 %) of furoic acid salts were reported for the oxidation of furfural in strongly alkaline media. Furoic acid is then obtained by acidification of the salts with H_2SO_4 or HCl. The difficult removal and reactivation of the silver catalyst are the major drawbacks of this method [221, 222]. The Quaker Oats Chemical Co. expanded the use of Ag catalysts to Au, Pt, and Pd and supported these noble metals on metal oxides like Fe_2O_3 [213, 223]. A bimetallic Pt/Pb on carbon catalyst has also been reported for the efficient selective oxidation of furfural to furoic acid with molecular oxygen in aqueous alkaline media [224]. The process issued by Interox Chemicals features biphasic oxidation of furfural with a concentrated aqueous solution of hydrogen peroxide in the presence of a secondary or tertiary amine [225].

Furoic acid can be converted to furoyl chloride, which is used in drug and insecticide manufacturing. Its esters can be produced by common esterification reactions. Tetrahydrofuroic acid is produced by hydrogenation of furoic acid or by hydrogenation and hydrolysis of furoic esters. Methyl furoate and methyl tetrahydrofuroate are used as intermediates for cardiovascular and cancer drugs. Furoic acid is a precursor for synthesis of antibiotic, alpha blockers, and corticosteroid intermediates.

Physical properties of furoic acid are listed in Table 11. Furoic acid has a LD_{50} (i.p., mouse) of 100 mg/kg.

9. References

1. A. P. Dunlop, F. N. Peters, *ACS Monograph Series: The Furans*, Reinhold Publishing Corp., New York 1953, pp. 733–773.
2. A. Padwa et al., *J. Org. Chem.* **66** (2001) 3119–3128.
3. H. Gilman, N. O. Calloway, *J. Am. Chem. Soc.* **55** (1933) 4197–4200.
4. H. Gilman, R. R. Burtner, *J. Am. Chem. Soc.* **57** (1935) 909–911.
5. T. Itahara, F. Ouseto, *Synthesis* **6** (1984) 488–489.
6. J. Tsuji, H. Nagashima, *Tetrahedron* **40** (1984) 2699–2702.
7. B. Taljaard and G. J. Burger, *Adv. Synth. Catal.* **344** (2002) 1111–1114.
8. Illovo Sugar Ltd., US 6506285, 1999 (G. J. Burger, A. D. Wills).
9. V. N. Ulrich, *Chem. News* **3** (1861) 116.
10. K. Yoshizawa, S. Toyota, F. Toda, *Tetrahedron Lett.* **42** (2001) 7983–7985.
11. A. Sharifi, M. M. Mojtahedi, M. R. Saidi, *Tetrahedron Lett.* **40** (1999) 1179–1180.
12. A. P. Dunlop, F. N. Peters: *ACS Monograph Series: The Furans*, Reinhold Publishing Corp., New York 1953, pp. 341–348.

13. H. Adkins, J. Semb, L. M. Bolander, *J. Am. Chem. Soc.* **47** (1931) 1853–1858.
14. M. J. Astle, J. A. Zaslowski, P. G. Lafyatis, *Ind. Eng. Chem.* **46** (1954) 787–788.
15. American Soap and Glycerine Producers, US 1934309, 1933 (K. H. Hoover).
16. Norton Company, US 2526601, 1950 (L. Coes).
17. J. Tateiwa, H. Horiuchi, S. Uemura, *J. Org. Chem.* **60** (1995) 4039–4043.
18. Y. Tanaka, N. Sawamura, M. Iwamoto, *Tetrahedron Lett.* **39** (1998) 9457–9460.
19. H. E. Burdick, H. Adkins, *J. Am. Chem. Soc.* **56** (1934) 438–442.
20. G. A. Olah, A. K. Mehrotra, *Synthesis* **11** (1982) 962–963.
21. C. Wang, M. Li, *Synth. Commun.* **32** (2002) 3469–3473.
22. T. S. Li, Z. H. Zhang, Y. J. Gao **28** (1998) 4665–4671.
23. C. Pereira et al., *Synthesis* (1995) 1077–1078.
24. G. P. Romanelli et al., *Tetrahedron Lett.* **44** (2003) 1301–1303.
25. M. L. Kantom et al., *Chem. Commun.* **1998**, 1033–1034.
26. Quaker Oats Company, US 2527714, 1950 (A. P. Dunlop, E. L. Washburn).
27. Shenley Industries Inc., US 2770619, 1956 (E. Schraufstatter, M. Bock).
28. Eastman Kodak Company, US 2768175, 1956 (D. S. Young, G. P. Rogers).
29. Azoplate Company, US 2768077, 1956 (W. Neugebauern, M. Tomanek, T. Scherer).
30. Les Usines de Melle S.A., US 2848498 1958, (M. Mention).
31. Du Pont de Nemours, US 2448520, 1948 (M. E. Cupery).
32. E. Knoevenagel, DE 164296, 1904.
33. Schering Corp. US 2594355, 1952 (E. Schwenk, D. Papa).
34. Schering Corp. US 2606922, 1952 (D. Papa, E. Klingsberg).
35. A. Loupy et al., *J. Chem. Soc. Perkin Trans.* **1** (2001) 1220–1222.
36. P. K. Kwon et al., *Synth. Commun.* **27** (1997) 4091–4100.
37. N. A. Milas, W. L. Walsh, *J. Am. Chem. Soc.* **57** (1935) 1389–1390.
38. Research Co., US 2118567, 1935 (N. A. Milas, W. L. Walsh).
39. P. Esser, B. Pohlmann, H. D. Scharf, *Angew. Chem.* **106** (1994) 2093–2108.
40. BASF, DE 2111119, 1971, G. Bolz, W. W. Wiersdorff.
41. A. P. Dunlop, F. N. Peters: *ACS Monograph Series: The Furans*, Reinhold Publishing Corp., New York 1953, pp. 102–105, 152–156.
42. A. Gandini, M. N. Belgacem, *Prog. Polym. Sci.* **22** (1997) 1203–1379.
43. H. Harms: European Conference "Crops for a Green Industry", Gmunden, Austria October 6–8, 1998.
44. D. R. Arnold, J. L. Buzzard: Proceedings of the South African Chemical Engineering Congress, Sun City, South Africa, September 3–5, 2003.
45. C. D. Hurd, L.L. Isenhour, *J. Am. Chem. Soc.* **54** (1932) 317–330.
46. G. Barale et al., *Inform. Chim.* **228/229** (1982) 151–159.
47. E. F. Root et al., *For. Prod. J.* **9** (1959) 158–165.
48. K. J. Zeitsch: "Sugar Series", in *The Chemistry and Technology of Furfural and its Many By-products*, vol. 13, 1st ed, Elsevier Science, Amsterdam, 2000, pp. 14–22.
49. E. E. Hughes, S. F. Acree, *J. Res. Nat. Bureau Standards* **21** (1938) 327–336
50. K. J. Zeitsch: "Sugar Series", in *The Chemistry and Technology of Furfural and its Many By-products*, vol. 13, 1st ed, Elsevier Science, Amsterdam, 2000, pp. 262–263.
51. Quaker Oats Co, US 1735084, 1922 (C. S. Miner, H. J. Brownlee).
52. Quaker Oats Co, US 1919878, 1928 (H. J. Brownlee).
53. Quaker Oats Co, US 1919877, 1928 (H. J. Brownlee).
54. Quaker Oats Co, US 2140572, 1936 (H. J. Brownlee).
55. Quaker Oats Co, US 2536732, 1947 (A. P. Dunlop).
56. Quaker Oats Co, US 2884428, 1956 (H. C. Wamsley, P. A. Wells).
57. K. J. Zeitsch: "Sugar Series", in *The Chemistry and Technology of Furfural and its Many By-products*, vol. 13, 1st ed, Elsevier Science, Amsterdam 2000, pp. 41–43.
58. Agrifurane S.A., GB 842689, 1958.
59. Agrifurane S.A., FR 1161087, 1956 (J. Guyonnaud).
60. Escher Wyss G.m.b.H., US 3035962, 1962 (E. Schinn).
61. Oy. W. Rosenlew Ab., US 4029515, 1975 (K. Kiminki, R. Kulmala, S. Sipila).
62. WO00/47569, 2000 (P. D. Steiner, K. J. Zeitsch).

63. K. J. Zeitsch: "Sugar Series", in *The Chemistry and Technology of Furfural and its Many By-products*, vol. 13, 1st ed., Elsevier Science, Amsterdam 2000.
64. P. D. Steiner, K. J. Zeitsch, WO 00/63488, 2000.
65. E. Ramos, L. A. Carlo, R. Vazquez-Romero, *Proc. Int. Soc. Sugar Cane Technol* **13** (1969) 1900–1911.
66. Commw of Puerto Rico, US 3701789, 1970 (Ramos, Rodrigez).
67. Biofine Inc., US 4897497, 1988 (S.W. Fitzpatrick).
68. Biofine Inc., US 5608105, 1995 (S.W. Fitzpatrick).
69. Arkenol Inc., US5892107, 1996 (W. A. Farone, J. E. Cuzens).
70. J. Horvat, B. Klaic, B. Metelko et al., *Tetrahedron Lett.* **26** (1985) 2111–2114.
71. N. Vedernikov: Proc. 10th Int. Symp. Wood and Pulping Chemistry, Yokohama, Japan, June 7–10, 1999, 3, 468.
72. J. Gravitis et al.: "Furfural and Levoglucosan Production from Deciduous Wood and Agricultural Wastes", in *Chemicals and Materials from Renewable Resources*, American Chemical Society, Washington, DC, 2001, ACS Symp. Ser. N 784, chap. 9, pp. 110–122
73. SU1365674, 1994 (N. A. Vedernikov et al.).
74. www.cropguard.co.za.
75. Steenwinkel 1993, TNO report V3934.
76. NIH publication No. 90-2837, 1990.
77. Quaker Oats Co., US 4426460 (1980, W. J. Pentz).
78. QO Chemicals Inc., WO 9418187, 1993 (G. Everett, G. Maclennan, M. Chen).
79. I. Iovel, Y. Goldberg, M. Shymanska, *J. Mol. Catal.* **57** (1989) 91–103.
80. J. Lecomte, A. Finiels, P. Geneste, C. Moreau, *Appl. Catal. A* **168** (1998) 235–241.
81. A. Finiels, P. Geneste, J. Lecomte, J. Marchinez, C. Moreau, *J. Mol. Catal.* **148** (1999) 165–172.
82. Agrichimie S.A., FR 2741343, 1995 (O. Boulet, R. Emo, P. Faugeras et al.).
83. S. Manfredi et al., *Chem. Heterocycl. Compd.* (1999) 457–495.
84. E. Gill, H. Ing, *J. Am. Chem. Soc.* (1958) 4728–4731.
85. T. Imagawa, T. Nakagawa, K. Matsuura, et al., *Chem. Lett.* (1981) 903–904.
86. A. Pelter, B. Singaram, *Tetrahedron Lett.* **23** (2) (1982) 245–248.
87. A. Pelter, B. Singaram, *J. Chem. Soc. Perkin Trans 1* **1983** 1383–1386.
88. A. Gandini, M. N. Belgacem, *Prog. Polym. Sci.* **22** (1997) 1203–1379.
89. M. Choura, N. M. Belgacem, A. Gandini, *Macromolecules* **29** (1996) 3839–3850.
90. D. E. Ponde, V. H. Deshpande, V. J. Bulbule, et al., *J. Org. Chem.* **63** (1988) 1058.
91. P. Kumar, R. K. Pandey, *Synlett.* **2** (2000) 251.
92. A. Orita, C. Tanahashi, A. Kakuda, J. Otera, *Angew. Chem. Int. Ed.* **39** (2000) 2877.
93. Y. Nakae, I. Kusaki. T. Sato, *Synlett.* **10** (2001) 1584.
94. P. Kumar, R. K. Pandey, M. S. Bodas, S. D. Dagade, et al., *J. Mol. Catal. A* **181** (2002) 207.
95. M. Lautens, E. Fillion, *J. Org. Chem.* **62** (1997) 4418–4427.
96. H. Garcia, A. Gilbert, O. Griffiths, *J. Chem. Soc. Perkin Trans* **2** (1994) 247–252.
97. O. Arjona, F. Iradier, R. M. Manas, *Tetrahedron* **54** (1998) 9095–9110.
98. O. Achmatowicz, M. H. Burzynska, *Tetrahedron* **38** (1982) 3507–3513.
99. J. Moursounidis, D. Wege, *Tetrahedron Lett.* **27** (1986) 3045–3048.
100. P. Kumar, C. U. Dinesh, R. S. Reddy, et al., *Synthesis* **11** (1993) 1069–1070.
101. B. S. Babu, K. K. Balasubramanian, *Tetrahedron Lett.* **39** (1998) 9287–9288.
102. B. S. Babu, K. K. Balasubramanian, *Syn. Lett.* **8** (1999) 1261–1262.
103. Resinous Prod. & Chem. Co., US 2280790, 1940 (H. A. Bruson).
104. US 2763665, 1952 (R. H. Lock, K. Reynolds).
105. US 3752849, 1971 (M. Otsuka, T. Kinoshita, T. Masawa).
106. US 5175358, 1989 (B. Capai, G. Latigau).
107. H. Adkins, R. Conner, *J. Am. Chem. Soc.* **53** (1931) 1091–1097.
108. H. Adkins, R. Conner, *J. Am. Chem. Soc.* **54** (1932) 4678–4689.
109. Quaker Oats Co., US 2097493, 1937 (G. J. Leuck, J. Pokorny, F. N. Peters).
110. R. M. Lukes, L. S. Nelson, *J. Org. Chem.* **21** (1956) 1096–1098.
111. N. A. Milas, W. L. Walsh, *J. Am. Chem. Soc.* **57** (1935) 1389–1393.
112. US 2118567, 1938 (N. A. Milas, W. L. Walsh).
113. H. Friedrich, F. Khan, N. Singh et al., *Synlett.* **6** (2001) 869–871.
114. H. Friedrich, B. Holger, N. Singh, *Tetrahedron Lett.* **41** (2000) 3971–3974.
115. H. M. R. Hoffman et al., *Tetrahedron* **49** (1993) 8999–9018.
116. B. M. Adger et al., *J. Chem. Soc. Chem. Commun.* **1991**, 1553–1554.

117. A. De Mico, R. Margarita, G. Piancatelli, *Tetrahedron Lett.* **36** No. 20 (1995) 3553–3556.
118. J. Finlay, M. A. McKervey, H. Q. N. Gunaratne, *Tetrahedron Lett.* **39** (1998) 5651–5654.
119. J. Wahlen et al., *Adv. Synth. Catal.* **346** (2004) 333–338.
120. W. R. Kirner, *J. Am. Chem. Soc.* **50** (1928) 1955.
121. L. P. Kyrides, M. F. Zienty, *J. Am. Chem. Soc.* **71** (1949) 1122.
122. J. Zanetti, *J. Am. Chem. Soc.* **49** (1927) 1065.
123. Quaker Oats Co., US 1906873, 1928 (F. N. Peters).
124. Röhm, Haas Co., GB 410148, 1932.
125. Quaker Oats Co., US 2094975, 1932 (H. Adkins, R. Conner).
126. Ashland Oil Inc., US 4251396, 1979 (L. J. Franier, H. H. Fineberg).
127. Ashland Oil Inc., US 4302397, 1980 (L. J. Franier, H. H. Fineberg).
128. DE2526198, 1975 (S. Ciborowski et al.).
129. Kao Soap Co., Ltd., US 4252689, 1979 (M. Bunji).
130. Gaf Corp., US 4153578, 1978 (W. De Thomas, E. Hort).
131. L. Baijun et al., *Appl. Catal. A* **171** (1998) 117–122.
132. S. P. Lee, Y. W. Chen, *Ind. Eng. Chem. Res.* **38** (1999) 2548–2556.
133. H. Li et al., *J. Mol. Catal.* **203** (2003) 276–275.
134. US 4261905, 1979 (E. Preobrashenskaya et al.).
135. R. Rao, A. Dandekar, R.T.K. Baker et al., *J. Catal.* **171** (1997) 406–419.
136. S.A. des Dist. des Deux Serves, US 1739919, 1927 (E. Ricard, H. M. Guinot).
137. Du Pont de Nemours, US 2077422, 1931 (W. Lazier).
138. Quaker Oats Co., US 2754304, 1952 (S. Swadesh).
139. Quaker Oats Co., US 4185022, 1977 (A. A. Kozinski).
140. Degussa, US 5591873, 1995 (M. Bankmann, J. Ohmer, T. Tacke).
141. Engelhard Co., US 5977010, 1995 (B. D. Roberts, W. J. Carrick, D. S. Thakur).
142. J. Kijenski, P. Winiarek, *Appl. Catal. A* **193** (2000) L1–L4.
143. J. Kijenski et al., *Appl. Catal. A* **233** (2002) 171–182.
144. P.R. Carey, M. Lott, *Foundry Management & Technology*, July 1995, 26–30.
145. Wood Polymer Technologies AS, WO 02060660, 2002 (M. L. Schneider).
146. Wood Polymer Technologies AS, WO 0230638, 2001 (M. L. Schneider).
147. NIH publication No. 99-3972, 1999.
148. H. D. Hartough, A. I. Kosak, *J. Am. Chem. Soc.* **69** (1947) 1012).
149. Socony-Vacuum Oil Co., US 2460823, 1946 (H. D. Hartough, A. I. Kosak).
150. Socony-Vacuum Oil Co., US 2460822, 1946 (H. D. Hartough, A. I. Kosak).
151. Socony-Vacuum Oil Co., US 2515123, 1947 (H. D. Hartough, A. I. Kosak).
152. BASF AG, DE 3618964, 1986 (H. Lermer, W. Hoelderigh, M. Schwartzmann).
153. ICI Ltd., US 2475097, 1945 (D. G. Jones).
154. D. M. Burness, A. E. Black, H. E. Baumgarten, *Org. Synth.* **CV5**, 403.
155. Kemisk Vaerk Koge, GB 682736, 1949.
156. Sadolin Holmblad, US 2714576, 1953 (N. K. F. W. Clauson-Kaas, F. Limborg).
157. BASF AG, DE 2710420, 1977 (D. Degner, H. Nohe).
158. BASF AG, US 4906758, 1990 (W. Hoelderich, M. Hesse, H. Siegel).
159. Du Pont de Nemours & Co., US 2600289, 1952 (C. A. Bordner).
160. Daicel KK, JP 58090548, 1983 (K. Masayoshi).
161. Asahi Chem. Ind. Co. Ltd., JP 1301658, 1989 (Y. Tadatsugu).
162. P. B Venuto, *Microporous Mater.* **2** (1994) 297.
163. Y. Ono, *Heterocycles* **16** (1988) 1755.
164. T. I. Gubina, V. G. Karchenko, *Chem. Heterocycl. Compd.* **31** (1995) 900
165. Du Pont de Nemours & Co., US 2374149, 1943 (G.M. Whitman).
166. Du Pont de Nemours & Co., US 2776981, 1954 (L.W. Tyran).
167. Du Pont de Nemours & Co., US 2634276, 1951 (J. E. Carnahan).
168. Du Pont de Nemours & Co., US 3007941, 1959 (H. B. Copelin, D. I. Garnett).
169. Quaker Oats Co., US 3257417, 1963 (A. P. Dunlop, G. W. Huffman).
170. P. Lejemble, Y. Maire, A. Gaset et al., *Chem. Lett.* (1983) 1403–1406
171. Agrifurane S.A., EP 0096913, 1983 (P. Lejembe, A. Gaset, P. Kalck et al.).
172. Quaker Oats Co., US 3223714, 1963 (D. G. Manly, J. P. O'Halloran).
173. BASF Ag., US 4780552, 1988 (M. Fischer, M. Irgang, L. Wambach).
174. Du Pont de Nemours & Co., US 2900396, 1957 (J. R. Harisson).

175. Phillips Petroleum Co., US 3906009, 1973 (F. E. Fahra, Jr.).
176. Phillips Petroleum Co., US 3912763, 1973 (F. E. Fahra, Jr., M. M. Johnson, D. C. Tabler).
177. Du Pont de Nemours & Co., WO 0012208, 1998 (K. Kourtakis).
178. Du Pont de Nemours & Co., US 4322358, 1979 (T. Bither, W. R. McClellan).
179. The Standard Oil Co., US 4342698, 1979 (J. P. Bartek, R. K. Grasselli, R. L. Cepulis).
180. GAF Corp., US 4293444, 1980 (R. Parthasarathy, E. V. Hort).
181. GAF Corp., US 4309356, 1981 (R. Parthasarathy, E. V. Hort).
182. Du Pont de Nemours & Co., WO 03/062217 A1, 2003 (K. Kourtakis, P. Mills).
183. Du Pont de Nemours & Co., US 4172838, 1978 (D. I. Garnett, M. L. Peterson).
184. Du Pont de Nemours & Co., US 4268421, 1980 (D. I. Garnett, M. L. Peterson).
185. Du Pont de Nemours & Co., US 4387236, 1982 (R. B. Pinkerton).
186. Du Pont de Nemours & Co., US 4278563, 1980 (J. M. Fremont, D. I. Garnett).
187. Du Pont de Nemours & Co., US 4243593, 1980 (J. M. Fremont, D. I. Garnett).
188. O. Grummitt, J. A. Stearns et al., *Org. Synth.* **CV3**, 833.
189. C. Wilson, *J. Chem. Soc.* **1945**, 48.
190. D. Kaufman, W. Reeve, *Org. Synth.* **CV3**, 693.
191. Du Pont de Nemours & Co., US 2768978, 1952 (J. A. Robertson).
192. ICI ltd., US 2365623, 1941 (J. G. Mackay Bremner, D. Mc-Neil).
193. Quaker Oats Co., GB 858626, 1959.
194. Olin Mathieson Chemical Co., US 3240790, 1963 (W. I. Denton).
195. Olin Mathieson Chemical Co., US 3240791, 1963 (W. I. Denton).
196. Emery Industries Inc., US 3518281, 1968 (D. C. Holtman).
197. BASF, EP0691337, 1995 (J. Schröder, K. Ebel).
198. BASF, US 2003/0162980 (H. Weigl, K. Ebel, J. Simon).
199. Du Pont de Nemours & Co., US 1794453, 1928 (G. D. Graves).
200. Du Pont de Nemours & Co., US 2077422, 1937 (W. A. Lazier).
201. ICI Ltd., GB 337296, 1930.
202. ICI Ltd., GB 608540, 1946 (F. Starkey, J. G. Mackay Bremner).
203. Quaker Oats Co., US 2838523, 1954 (A. P. Dunlop, H. Schegulla).
204. Goodyear Tire and Rubber Co, GB 412914, 1932.
205. Du Pont de Nemours & Co., US 2112715, 1935 (C. Sly).
206. Melle-Bezons, US 3812161, 1969 (A. Bouniot).
207. Wingfoot Co, US 2109159, 1935 (C. F. Winans).
208. US 2175585, 1936 (H. Adkins, C. F. Winans).
209. Mitsubishi Petrochemical Co, US 4598159, 1984 (T. Ayusawa, S. Mori, T. Aoki et al.).
210. ICI ltd., US 2636902, 1949 (A. W. C. Taylor, P. W. Reynolds).
211. Mitsubishi Petrochemical Co, US 4544749, 1984 (T. Ayusawa, T. Aoki, R. Hamana et al.).
212. Mitsubishi Petrochemical Co, US 4605742, 1985 (S. Moiro, T. Aoki, R. Hamana et al.).
213. L. W. Burnett, *IE& C* **40** (1948) no. 3, 502–505
214. US 2445714, 1945 (Holdren R.).
215. Pure Energy Co., US 6479677, 2000 (I. Ahmed).
216. D. G. Manly, A. P. Dunlop, *J. Org. Chem.* **23** (1958) 1093–1095.
217. R. Lukes, C. Wilson, *J. Am. Chem. Soc.* **73** (1951) 4790–4794.
218. Batelle Memorial Institute, US 5883266, 1998 (D. C. Elliott, J. G. Frye).
219. V. R. Koch, J. H. Young, *Science* **204** (1979) 499–501.
220. US Department of Energy, 6450-01-P, May 17, 1999.
221. Quaker Oats Company, US 2041184, 1932 (L. L. Isenhour).
222. Union Carbide Co., GB 782430, 1954 (A. E. Montagna).
223. Quaker Oats Company, US 2407066, 1942 (A. P. Dunlop).
224. P. Verdeguer, N. Merat, A. Gaset, *Appl. Catal. A* **112** (1994) 1–11.
225. Interox Chemicals Limited, GB 1288927, 1986 (K. T. Rowbottom, D. A. Cummerson).

Furs

THILO NUNGESSER, (BASF Aktiengesellschaft), Ludwigshafen, Federal Republic of Germany

1. Introduction . . . 1
1.1. Definitions . . . 1
1.2. Fur-Bearing Animals . . . 1
1.3. Technological Data . . . 2
2. Nature and Production of Fur Pelts . 2
2.1. Morphology . . . 2
2.2. Flaying . . . 2
2.3. Preservation . . . 2
3. Fur Dressing . . . 3
3.1. Soaking . . . 3
3.2. Scouring . . . 3
3.3. Mechanical Treatment . . . 3
3.4. Pickling . . . 3
3.5. Tanning . . . 3
3.6. Oiling and Fatliquoring . . . 4
3.7. Degreasing in Organic Solvents . . . 5
4. Fur Finishing . . . 5
4.1. Bleaching . . . 5
4.2. Dyeing . . . 5
4.3. Fur Dyes . . . 6
4.4. Special Dyeing Processes . . . 7
5. Mechanical Operations . . . 8
6. Economic Aspects . . . 8
7. References . . . 9

1. Introduction

1.1. Definitions

This article surveys the chemical and physical treatment of fur skins from the raw material to the dressed and finished individual furs. The term *fur dressing* (Chap. 3) describes operations (e.g., soaking, washing, tanning, and fatliquoring) that do not significantly change the hair of the fur skins. The term *fur finishing* (Chap. 4) encompasses processes (e.g., bleaching, dyeing, and mechanical operations) that modify the appearance of the hair.

1.2. Fur-Bearing Animals

The raw materials used for the manufacture of leather are mainly large skins and hides of domestic mammals (e.g., cow, calf, sheep, pig, and goat). For the production of furs, however, the skins of more than one hundred different species are used. These animals are almost exclusively mammals and include cloven-hoofed animals, predators, and rodents. The economic significance of the various fur skins varies widely (Chap. 6).

In order to meet steadily growing demand, valuable fur skins, e.g., sable, chinchilla, mink, fox, and karakul (Persian lamb), are obtained mainly from animals bred in captivity. Fur skins obtained from animal farms are usually more valuable than the corresponding skins obtained from the wild. Animal farms now account for 90 % of fur skin production.

1.3. Technological Data

The countries in which fur skins are dressed and finished on an industrial scale are mainly those with a large production or consumption of such skins. Fur skins are often dressed and finished on a commission basis. From prehistoric times, dressing and finishing have been carried out in aqueous solutions of certain chemicals. The vessels most frequently used for wet treatment are *paddles,* large trough-like containers with a paddle in which the fur skins float in long dilute liquors; the chemicals are only incompletely taken up. For reasons of economy (to save time, and to reduce consumption of energy, water, and chemicals) and to reduce environmental pollution, equipment used in leather production has been introduced, e.g., tanning drums (often specially equipped) and other machines based on the same principles as drum washing machines (but with Y-shaped partitions) or cement mixers (→ Leather).

The term *liquor ratio* denotes the ratio of the volume of liquor to the weight of fur skin. In processing raw, dried pelts in paddles, the liquor ratio may vary between 12 : 1 and 30 : 1;

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_135

for moisture-containing, centrifuged fur skins, it may vary from 5 : 1 to 10 : 1; and in drums from 3 : 1 to 6 : 1. The amount of chemicals used is usually expressed in g/L, mL/L, or sometimes as a fraction (wt %) of the weight of the fur skins.

2. Nature and Production of Fur Pelts

2.1. Morphology

Morphologically, there is essentially no difference between the skins and hair of animals that are processed for leather and those that are used for fur skins (→ Leather). The corium or true skin of some fur-bearing animals, e.g., chinchilla, is thin, whereas in other species, such as the seal, the corium is very thick.

The *hair* of many different types of fur skins varies widely in form and structure, and is species-specific. Even the hair of an individual skin does not have a uniform structure. There are three different kinds of hair:

1) Guard-hair, which is long, smooth, straight, and relatively sparse
2) Top-hair, which is also straight, somewhat shorter than the guard-hair, and thicker in the center
3) Underhair, underwool, fur-hair, or underfur, which is short, silky, usually crimped, and far more numerous than the other types of hair; lamb- and sheepskins possess only this type of hair.

Hair is a product of the thin outer layer of the skin, the epidermis. In the production of leather, the hair and the outer layer are removed by liming. (For details of the fine structure of hair, see → Hair Preparations, → Wool.)

2.2. Flaying

The pelt should be stripped from the animal carcass as soon as possible, but not before the body has cooled and bleeding stopped. The skins and hides used for the production of leather, and also the pelts of larger species, e.g., sheep, beaver, seal, and the big cats, are almost always flayed in the flat form. In contrast, the smaller types of pelt are cased: the skins are cut on the under side of the tail and along the hind legs across the body, and the skin is then removed by pulling it over the head of the body like a glove. The latter method is easier to carry out and is also beneficial for the fur.

2.3. Preservation

Freshly flayed fur pelts are seldom processed immediately. In most cases they are preserved; the water content of the skin is reduced in order to prevent putrefaction by microorganisms.

The customary method of preservation is to dry the skin in air at a not too high temperature (ca. 25 °C, with some air convection) and in a lightly stretched state. The hair itself is resistant to microorganisms, but if the skin begins to putrefy owing to inadequate or careless preservation bacterial destruction of the hair roots causes the hair to become loose, so that the complete fur skin becomes worthless.

The method of preserving the skins by salting, or by salting followed by drying, is used only for the skins of large fur-bearing animals in the flat form. Common salt is used to remove moisture from the skin, inhibiting putrefaction. Storage of the fur skins in cold rooms or cold storehouses hinders development of microorganisms and also protects the skins from attack by pests such as grease beetles and moths.

3. Fur Dressing

3.1. Soaking

The object of soaking is to moisten the preserved skins thoroughly, returning them to the condition in which they were immediately after flaying. The soak also removes dirt, blood, and preserving salt as well as water- and salt-soluble proteins. The pelts are soaked in water at 15 – 25 °C in vats, paddles, or drums, depending on the sensitivity of the raw material. Salted skins are generally more readily wetted than dried skins. Small amounts of bactericides are usually added to the soaking baths, and in wetting dried skins, common salt, and surfactants are added to accelerate the soaking process.

3.2. Scouring

Dirty and very greasy skins, particularly those of lamb and sheep, are given one or two washes at 25 – 35 °C in paddles or drum washing machines. The detergents commonly used for scouring fur skins are mixtures based on alkyl ether sulfates and alkyl sulfates with nonionic alkyl(phenyl)polyglycol ethers. Fat-dissolving organic solvents and soda are often added to the wash baths. The amount of detergent used is 1 – 3 g/L, and the maximum washing time is 2 h. Scouring is followed by rinsing, draining, and centrifugation or partial removal of water by passing the fur skin between rollers (*samming*).

3.3. Mechanical Treatment

After scouring, the rinsed skins are fleshed, i.e., freed from subcutaneous tissue and residues of flesh and fat. This process is carried out either manually on the fur dresser's bench or mechanically on fleshing machines. The smaller fur skins are shaved with a circular knife to level the thickness of the leather and to reduce the weight.

3.4. Pickling

Before the actual dressing process (tanning), the fur skins are pickled. The pickling serves several purposes: to loosen and open up the collagen of the skin by the hydrolytic effect of an acidic salt solution, to remove soluble protein, and to prepare the leather part for the subsequent dressing with mineral tannins.

The most common pickling acid was sulfuric acid, but as it is liable to damage the leather, it has now been replaced largely by organic acids, e.g., formic [*64-18-6*], glycolic [*79-14-1*], lactic [*598-82-3*], or sulfophthalic acid [*26657-75-0*], and also certain short-chain dicarboxylic acids. The concentration of the acid or the acid mixture in the pickle is 3 – 10 g/L. In order to prevent any acid-induced swelling of the collagen, and according to the type of pickling acid used, 40 – 80 g/L of common salt is added to the pickling liquor before immersing the fur skins. The temperature of the pickling bath is 20 – 30 °C, and the pickling time is usually 12 – 48 h. The fur skins are normally pickled in paddles or drums.

3.5. Tanning

Tanning converts the skin or hide substance of the pelt into leather. Tanning causes the peptide chains of the collagen to become differentiated or isolated from each other, cross-linked, and thereby strengthened; tanning markedly raises the shrinkage temperature. (For the theory of tanning, see → Leather.)

Fur skins must not be made heavy and leathery by the tannage; almost no vegetable tannins and few synthetic tanning agents are used.

Leipzig dressing is the oldest method of dressing and was once common; it consists merely of a common salt and sulfuric acid pickle, followed by oiling. Leipzig dressing cannot claim to be a tannage in the real sense of the word. Although it produces a lightweight and elastic leather that can be readily processed, the water resistance and the storage stability are quite unsatisfactory.

Dressing with Aluminum Salts. One of the oldest methods of dressing fur skins, still often used, is treatment with ammonium [*7784-26-1*] or potassium aluminum sulfate [*7784-24-9*], and aluminum sulfate [*7784-31-8*]. Solutions of these salts (applied in amounts of 20 – 50 g/L at ca. 30 °C) produce white and fairly elastic leathers, but the water resistance is unsatisfactory. During tanning, free acid is formed by hydrolysis; it is therefore necessary to add common salt to the tanning float to prevent acid-induced swelling of the collagen. At the end of the tanning process, the liquor is adjusted to a pH of 3.8 – 4 (preferably with sodium acetate). Chlorides and nitrates are superior in tanning behavior to the much more acidic sulfate salts of aluminum. The former are less readily washed out and they have greater tanning activity. The chlorides are often stabilized with masking agents and placed on the market as prepared tanning salts.

Trade Names: Lutan (BASF), Pellutax (Henkel), Novaltan (Zschimmer & Schwarz).

In order to improve the water resistance of the leather, accelerate leathering, and raise the shrinkage temperature, the use of aluminum salts in dressing is often combined with formaldehyde [*50-00-0*] or chromium salts.

Dressing with Aldehydes. Formaldehyde alone is rarely used for tanning because it produces fur skins that are very flat, that have poor stretch, and that are liable to lose tensile strength during storage. Formaldehyde is often used, however, in combination with other tanning and fatliquoring agents.

Dressing with Chromium Salts. In contrast to dressing with aluminum salts, tanning with chromium(III) salts is irreversible. Tanning or retanning with chromium salts produces a leather that has good resistance to water and heat, but somewhat inferior elastic properties. This method of tanning is indispensable whenever the fur skins are to be dyed with synthetic dyes (cf. Chapter 4.3). Chromium salts cannot be used, however, for fur skins that are to be dyed in dark shades with oxidation dyes or that are to undergo an oxidative bleach, because this would result in damage to or complete destruction of the leather.

The products used for chrome tanning of fur skins are the same as those recommended for the production of leather.

Trade Names: Chromosal (BAYER), Chromitan (BASF) Salcromo (STOPPANI), Chromodul (BRITISH CHROME).

For details of the theory of chrome tanning, basicity, masking, etc., see → Leather.

Chrome dressing is carried out with 5 – 15 g/L of chrome tannin (with a chromium oxide content of 25 – 30 %) together with common salt at 30 – 35 °C. The exact amount of chrome tannin used depends on the specific fur skin, the treatment that it has already received, and the shrinkage temperature desired. The tanning process, which takes 12 – 24 h, commences with low-basic chromium salt solutions and the bath is subsequently basified moderately by slow addition of solutions of weak alkali. After the dressing, the fur skins are drained and then centrifuged or partially dried.

3.6. Oiling and Fatliquoring

The purpose of oiling or fatliquoring is to render the leather of the fur skin soft and supple. During this process, the skin fibers are enveloped with fatty material to prevent them from sticking together during drying. Certain types of fur skins are treated for this purpose with water-insoluble fats in hammer or faller (crank) stocks. Fatliquoring is carried out with oiling agents emulsified in water. The starting materials for these products are in most cases liquid derivatives of animal and vegetable, nondrying oils, and also synthetic products (higher molecular mass chlorinated hydrocarbons). The natural oils are rendered emulsifiable in water by being partially sulfated or sulfonated, the synthetic substances by partial chlorosulfonation followed by saponification. In many cases medium viscosity mineral oils are also added to the fatliquors.

Fur skins can be oiled or fatliquored at all stages of processing and by various techniques. The oldest method is *brush oiling;* the oiling agent is applied with a brush to the leather side either manually or by machine. A growing trend is the application of electrolyte-stable fatliquors (in amounts up to 15 g/L) either in the tanning bath or immediately after tanning. If the fur skins are not to undergo any further wet treatment, fatliquoring is followed by drying at ca. 35 °C, depending on the relative humidity and the amount of air movement.

3.7. Degreasing in Organic Solvents

Degreasing removes fatty matter and solvent-soluble substances from the hair and the leather, thus reducing the weight of the fur skins. Degreasing also improves the dyeing properties, e.g., for suede. The customary method is to treat dry fur skins, dyed or undyed, dressed with mineral tannins, in degreasing machines for a few minutes at 20 – 30 °C with solvents (chlorinated hydrocarbons or fluorohydrocarbons). After degreasing, and in some cases refatliquoring, the fur skins are centrifuged in the machine and dried. Complete degreasing is never intended; on the contrary, a certain amount of fat should always remain in the leather. In order to ensure this, fatliquoring during the dressing process is carried out with a large proportion of nonextractable substances.

4. Fur Finishing

4.1. Bleaching

Reductive Bleach. Naturally white, but slightly yellowed fur skins of sheep and lamb, and certain types of kid and rabbit skins for which a pure white shade of hair is highly desirable but impossible to achieve simply by washing, are usually given a reductive bleach. The skins are treated during or after the wash (or for chrome-tanned skins also in the neutralizing bath) with sulfites, bisulfites, or most commonly with dithionite. Bleaching is carried out with 3 – 10 g/L of these products for up to 24 h at 25 – 35 °C. In order to achieve an exceptionally good effect, the reductive bleach may be preceded by an oxidative bleach with hydrogen peroxide (→ Bleaching, Chap. 2.2.4.).

Oxidative Bleach. A real bleach in the sense of an intensive destruction of the natural pigments in the hair can only be achieved by means of an oxidative bleach with peroxides, e.g., hydrogen peroxide, or persulfates, catalyzed by iron(II) salts; this treatment destroys the pigments in the hair. This method, or one of its numerous variations, is used to bleach the hair of naturally dark skins, e.g., black karakul or muskrat, in order to dye them subsequently in fashionable shades. For the catalytic bleach, the fur skins must be in faultless condition and must not have undergone any chrome tannage.

The catalytic bleach is a very difficult process to control and must be monitored carefully; the bleaching conditions (type and amount of chemicals, time, temperature, and pH) must be controlled precisely. A reductive afterbleach is often given in order to remove the iron salts. Following the bleach, the fur skins are retanned and in most cases fatliquored or oiled.

4.2. Dyeing

General. Although many types of fur skins, particularly the more valuable furs, are further processed without dyeing, the proportion of skins that is dyed has increased markedly in recent years (e.g., sheepskin articles for decorations, automobile seat covers, and clothing).

There are many variations in the process of fur dyeing, and the method used depends on the kind of fur. Certain types of fur skins are dyed only lightly in the hair and the leather by means of a bath treatment (*blending:* adjustment of natural shades; *reinforcing:* to intensify natural colors by treatment with metal salt solutions; or *blueing:* treatment of naturally yellowish fur skins with small amounts of blue dye) or only dyed partially in the hair by a brush or spray process (*tipping:* application of dye to the tips of guard hairs; *striping:* application of a dye solution to the central, or dorsal, region of certain fur skins in order to imitate a natural darkness in that region). The dull shade of naturally black karakul is intensified by bath dyeing, which simultaneously promotes gloss. The cheaper furs, such as lamb, sheep, kid, and rabbit, can be made to look like valuable furs by suitable dyeing and mechanical treatment. These cheaper fur skins are also printed to produce imitations of more valuable fur skins.

The presence of different types of hair, and variations in the chemical composition and morphology of the hair and skin, exerts a marked influence on the different behavior of the two proteins keratin and collagen in the dyeing process.

The keratin of the hair contains the same amino acids as the collagen of the skin, with the exception of L-hydroxyproline [*51-35-4*]. In contrast to collagen, keratin contains cysteine [*52-90-4*], which cross-links the polypeptide chains of the keratin and imparts stability. In addition, the proportion of acidic amino acids in keratin is far larger than in collagen. Consequently, the isoelectric points of the two polypeptides, an important factor for dyeing, lie at different pH; for untreated collagen, it is at ca. pH 7, whereas for keratin it is at ca. pH 5.5. Apart from this, the isoelectric point of collagen is shifted by tannage, whereas that of keratin remains practically unchanged (except when the skins are treated with aldehydes).

In order to accommodate these phenomena, certain rules have been established for dyeing fur skins. A dyeing temperature higher than 60 – 70 °C cannot be used because of the limited hydrothermal resistance of the leather. As both the hair and the leather are affected by alkali, dyeing cannot be carried out at high pH. Dyed skins must have a level surface and satisfactory

fastness to light, rubbing, and sublimation; in many cases dyed sheepskins are also required to have good fastness to ironing (see Chap. 5).

Pretreatment.

Killing. The purpose of killing is to clean the hair and open it up to take up the dye by spreading out the scales. The chemicals used for killing include ammonia, soda, or (less often) caustic soda solution together with wetting agents or detergents. The intensity of killing depends on the type of fur skin and the color effect desired. The brush process or the dip process may be used for this purpose. Dip killing takes 1 – 3 h at 20 – 35 °C.

Mordanting. Before dyeing with oxidation dyes the fur skins are treated with metal salt solutions, usually potassium dichromate or iron(II)sulfate, or (less often) copper(II)sulfate, or mixtures of these products. These mordants convert the dyes into color lakes and thus improve their fastness and the depth of the dyed shade. The mordants act on the killed and thoroughly rinsed fur skins in weakly acid solution; they must penetrate into the hair to ensure that the lacquer is adsorbed in the interior of the hair. Each mordant imparts a characteristic shade to the color lake. Mordanting is usually carried out in baths at 25 – 30 °C with 0.5 – 3 g/L of potassium dichromate or 3 – 6 g/L of iron(II)sulfate, and the process usually takes several hours. Adjustment of the pH, an important factor, is effected with formic, acetic, or tartaric acid or with potassium monotartrate (in amounts of 0.5 – 2 g/L). In most cases the mordanted fur skins are merely centrifuged or sammed (partially dried by rollers).

4.3. Fur Dyes

Vegetable Dyes. The oldest method of dyeing fur skins was to treat them with extracts of logwood, red wood, fustic, or sumach leaves. This method is seldom used today; it is restricted almost exclusively to the black dyeing of karakul with logwood and iron or copper salts. This dyeing process is complicated and usually has to be repeated several times; between the treatments, the fur skins are hung up in humid rooms at ca. 30 °C to ensure satisfactory oxidation of the dye.

Oxidation Dyes (→ Azine Dyes, Chap. 2.2.6.). These colorants were placed on the market towards the end of the last century and are still commonly used. They are not strictly dyes, but chemicals from which a pigment is formed in the hair by oxidation and conversion with metal salts. Examples of oxidation dyes include 1,4-phenylenediamine [*106-50-3*], pyrocatechol [*120-80-9*], resorcinol, aminophenols, and analogous naphthalene derivatives.

Trade Names: BASF Ursol (BASF), Nako (HOECHST).

Oxidation dyes produce dull colors from yellow to orange, brown, blue-grey and grey, to black with good levelness, and in full shades with good lightfastness, on all types of fur skins; the fastness to ironing is only limited.

The dye bath is prepared in a paddle at 30 – 35 °C, and the mordanted fur skins are then added. The dye (up to 8 g/L) is added with approximately the same amount of 30 % H_2O_2, usually in several portions. The actual pigment dye is developed in the course of several hours.

Fur skins that have been dyed with oxidation dyes must be washed and rinsed thoroughly to remove all loose and incompletely oxidized dye. Since the various bath treatments and rinses remove much of the tanning and fatliquoring agents from the fur skins, they are usually tanned and oiled or fatliquored again after dyeing. In working with oxidation dyes, certain precautions have to be taken because some people are allergic to specific products.

Synthetic Dyes. The large number of synthetic dyes, together with the use of new technology for dyeing the hair of fur skins, has now made it possible to meet demands for furs dyed in fashionable shades. Synthetic dyes are only taken up by the material at elevated temperature: for this reason, the fur skins must be chrome tanned or chrome retanned, so that the shrinkage temperature is at least 20 °C higher than the dyeing temperature of 60 – 65 °C. With these dyes, it is possible to dye the whole range of shades relatively easily by the bath process and in a far shorter time (1 – 2 h) than that necessary for dyeing with oxidation dyes. No mordant is required and killing is far less intensive; it is actually a hair cleaning rather than a killing process.

Disperse Dyes. Disperse dyes are azo and anthraquinone dyes, e.g., Cellitone (BASF), which, with the aid of dispersing agents, yield colloidal solutions in water (→ Disperse Dyes, Chap. 1.). They can be combined with one another, and they are taken up by the material at a temperature of less than 60 °C. Disperse dyes produce level dyeings in pale to medium shades with a medium level of lightfastness, but their fastness to ironing is unsatisfactory. These colorants are very suitable for use as brightening dyes in combination with oxidation dyes and as blending dyes for valuable fur skins.

1 : 2 Metal-complex dyes are also used in fur dyeing; for more details, see → Metal-Complex Dyes. When applied to lamb- and sheepskins at a minimum of 60 °C, these dyes produce shades with highly satisfactory fastness to light and ironing. They can be combined with one another (in amounts up to 2 g/L). With the aid of carriers (certain ethoxylation products and emulsions of chlorobenzenes or esters of phosphoric acid) they can also be used for dyeing dark shades.

Trade Names: Eukesolar (BASF), Irgalan (Ciba-Geigy), Dermalicht (Sandoz).

Anionic Dyes. From the comprehensive range of anionic dyes used in the textile and leather industries, only a few simple nitro, monoazo, and anthraquinone dyes have established themselves for dyeing the hair of fur skins (→ Azo Dyes, Chap. 3.). They are mainly low molecular mass dyes, which stain the leather only lightly or can be stripped from the leather by reduction, and are readily miscible with each other.

Anionic dyes often have unsatisfactory lightfastness when applied in pale shades. They are used on a fairly large scale as mixtures on the principle of trichromatic dyeing for dark shades.

Trade Names: Lurazol (fur) (BASF), Sellacid (Ciba-Geigy), Dermapelz (Sandoz).

Additions, Drying. The dyed fur skins are washed and rinsed, and if necessary tanned and fatliquored again.

As all treatments carried out in baths are liable to damage the hair of fur skins, it is customary to add fiber protective agents based on protein degradation products: these shorten the treatment by virtue of their surface action and function as protective colloids. For protection against the attack of moths, the fur skins may be treated in the bath with preparations such as Eulan (Bayer).

All wet treatments are usually followed by mechanical removal of water from the skins and drying at moderate temperature.

4.4. Special Dyeing Processes

Suede. Tone-in-tone dyeings (with the wool and the leather dyed in the same shade) can be produced by first dyeing the wool with oxidation dyes and then the leather with special leather dyes (→ Leather). When the hair is first dyed at elevated temperature with synthetic dyes, the leather is usually dyed in a fresh bath after the fur skins have been neutralized and rinsed.

For suede with white wool the intermediately dried, buffed skins are treated with a wetting agent and then dyed at room temperature with selected leather dyes that do not stain the wool at such a temperature.

Printing of Fur Skins. Imitations of valuable furs, such as leopard, ocelot, or tiger, can be produced on fur skins that have been suitably treated, in most cases lamb, sheep, calf, or horse, by the printing methods used in the textile industry. The printing inks most commonly used are solutions of oxidation dyes that contain a thickening agent and hydrogen peroxide.

Resist Process. Two-color effects can be produced on lamb- or sheepskins by brushing the hair with solutions of certain plastics at the parts that are to be resisted. The resist paste is allowed to dry, the fur skins are dyed in a hot bath, and the resist paste is then removed by rinsing with cold water.

5. Mechanical Operations

In addition to the chemical processes discussed above, certain mechanical operations also play a critical role in the dressing and dyeing of fur skins.

Drumming and Cageing. In this important operation, the dressed or dyed fur skins are tumbled in batches in large drums for several hours with moist, later with dry, sawdust from

deciduous trees (beech, poplar, or birch). This operation removes excess fat, salts, and other impurities, and improves the suppleness of the leather and the gloss of the hair. In addition to water, organic solvents (tetrachloromethane or tetrachlorethylene) and surfactants may be added to the sawdust. After this treatment, the adhered dust and shavings are shaken out of the fur skins using a rotating drum similar to the drum used for cleaning, but which consists of a wide-meshed wire cage.

Stretching and Staking (→ Leather). These operations render the leather softer and more elastic.

Plucking or Dehairing. Certain kinds of fur skins, e.g., nutria, beaver, and rabbit, have top-hair which often has to be removed, either by hand with special instruments or on plucking machines.

Shearing. Special effects can be obtained on some types of fur skins, e.g., sheep, lamb, rabbit, and nutria, by trimming the hair on shearing machines.

Ironing. The wool of sheep and other skins with dull and crimped hair can be stretched and made more lustrous by ironing on machines provided with heated, rapidly rotating cylinders. The hair is brushed with water, with or without the addition of a wetting agent, and the hair side is then ironed at ca. 200 °C, and the process repeated several times. In order to fix the stretching effect on the hair and to make the effect resistant to moisture, the ironing solutions applied later contain increasing amounts of formaldehyde together with water-soluble alcohols and acid. This process can be used, for instance, to produce imitations of beaver on lamb skins.

Beating and Combing. The hair of the dried skins is finally beaten and combed, either by hand or mechanically, in order to eliminate felting.

Sorting. The final sorting of the fur skins according to the kind of fur, color, general appearance, fall and fineness of the hair, appearance of the curls, and the size is an operation that requires considerable experience.

6. Economic Aspects

The main suppliers of valuable fur skins are the former Soviet Union, Scandinavia, and North America. Annual production in 1983/84 is given in Table 1.

Table 1. World production of fur skins in 1983/1984

Fur	Country	Approximate number of skins, $\times 10^6$
Mink	Soviet Union	9
	Scandinavia	13
	United States	3
	World total	*28*
Nutria (coypu)	Argentina	3
	North America	1.25
	Poland	1
	World total	*6*
Fox	Scandinavia	3.4
	World total	*4.1*
Karakul (Persian lamb)	Afghanistan	2
	South Africa	1
	Soviet Union	7
	World total	*11*

7. References

General References

1. W. Pense: "Rauchwaren" in W. Grassmann (ed.): *Handbuch der Gerbereichemie und Lederfabrikation,* vol. 3, 2nd Part, Springer, Vienna 1955, pp. 517 – 638.
2. F. Stather: *Gerbereichemie und Gerbereitechnologie,* 4th ed., Akademie-Verlag, Berlin 1967, pp. 805 – 885.
3. *Rauchwarenherstellung und Pelzkonfektion,* VEB Fachbuchverlag, Leipzig 1979.
4. J. Kroll, C. Franke: *Jury Fränkel's Rauchwaren-Handbuch,* 5th ed., Rifra-Verlag, Murrhardt 1976.

Gallbladder and Liver Therapy

Géza Csomós, Hamburg, Federal Republic of Germany (Chaps. 1 and 3)

Ulrich Leuschner, Johann Wolfgang Goethe Universität, Frankfurt, Federal Republic of Germany (Chap. 2)

1. Choleretics 2
1.1. Natural Products 2
1.2. Synthetic Products 3
1.2.1. Benzyl Alcohol Derivatives 3
1.2.2. Cinnamic Acid Derivatives 3
1.2.3. Miscellaneous Synthetic Choleretics . 4
2. Dissolution of Gallstones 6
2.1. Gallstone Pathophysiology 6
2.2. Oral Dissolution of Gallbladder Stones with Bile Acids 6
2.3. Direct Percutaneous – Transhepatic Lysis (PTL) of Gallbladder Stones . 8
2.4. Oral Litholysis after Extracorporeal Shock-Wave Lithotripsy (ESWL) . 8
2.5. Direct Dissolution of Bile Duct Stones 9
2.6. Critical Assessment of Chemical Stone Dissolution 9
3. Liver Therapeutics 10
3.1. Introduction 10
3.2. Therapeutic Agents 11
3.2.1. Inhibitors of Viral Nucleic Acid Synthesis 11
3.2.2. Bile Salts used for Primary Biliary Liver Disease 12
3.2.3. Natural Substances 12
3.2.4. Miscellaneous Substances 14
3.2.5. Antifibrinogenic Drugs 17
4. References 18

Abbreviations:

ALT	alanine aminotransferase
ara-A	9-β-D-arabinofuranosyladenine
ara-AMP	9-β-D-arabinofuranosyladenine 5-monophosphate
AST	aspartate aminotransferate
BSP	bromsulphalein
cheno	chenodeoxycholic acid
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
EPL	essential phospholipids
ESWL	extracorporeal shock-wave lithotripsy
GLDH	glutamate dehydrogenase
GOT (AST)	glutamic – oxalacetic transaminase
GPT (ALT)	glutamic – pyruvic transaminase
γ-GT	γ-glutamyltransferase
HBV	hepatitis B virus
HMG-CoA	β-hydroxy-β-methylglutaryl-coenzyme A
MAST	microprocessor-assisted solvent transfer system
MTBE	methyl *tert*-butyl ether
PTC	percutaneous – transhepatic cholangiography
PTL	percutaneous – transhepatic lysis
RNA	ribonucleic acid
urso	ursodeoxycholic acid

This article is divided into three chapters that deal with substances that increase bile flow (choleretics, Chap. 1), the dissolution of gallstones in the gallbladder and bile duct (Chap. 2), and drugs used for liver therapy (Chap. 3).

1. Choleretics

Although the gallbladder is not one of the most vital organs of the human body, it causes many subjective complaints and objective changes, so that it continues to be a subject of great interest in both research and medicine. Numerous substances have been tested in animals with a view to stimulating bile flow or inhibiting lithogenesis (i.e., stone formation). Stimulation of bile flow has been attributed to several substances [1].

A *cholagogue* is a substance that increases bile flow. Cholagogues may be either *choleretics,* which stimulate secretion of bile by the liver, or *cholekinetics,* which lead to emptying of the gallbladder [2].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_143

The major constituents of bile are the surface-active bile acid salts; other components include cholesterol, phospholipids, bilirubin, mucus, and inorganic salts. Bile is produced in liver cells and secreted into the bile ducts. It then passes through a shunt into the gallbladder, where it is concentrated to about one-tenth of its original volume. After ingestion of food, the gallbladder contracts and forces the concentrated bile into the duodenum via the common bile duct [3].

On the average, about 500 – 600 mL of bile is produced daily. Secretion of water and electrolytes into the bile duct system is stimulated by

1) secretin,
2) the autonomous nervous system,
3) changes in pressure in the bile duct, and
4) changes in the flow of blood through the liver.

The most important function of bile is the digestion of fats: it emulsifies ingested fats and activates pancreatic lipase.

Cholagogic drugs have not become established in English-speaking and Scandinavian countries, where they are viewed rather skeptically by the medical field. However, in other European countries where medicine is more inclined to speculation, these substances still play a surprisingly important role. Many of them have been tested only in animals, and their effects were found to be species dependent. However, results obtained in animal experiments cannot be extrapolated to humans, and investigations on humans have generally used unsatisfactory methods (e.g., duodenal probes). When reliable methods were employed to test choleretics in humans, only the bile acid preparations were found to be effective [4]. The choleresis observed on administration of bile acids involves an increased secretion of water and electrolytes (i.e., hydrocholeresis) and can also be observed during spontaneous fluctuations of bile secretion or after intake of food. True choleresis, i.e., increased secretion of bile-specific substances, cannot be induced by drugs. In addition to these drawbacks, unlimited therapeutic use of choleretics cannot be recommended for any single disease of the digestive tract [5]. Substances native to the body (bile preparations), plant drugs, and synthetic compounds are prescribed as therapeutic choleretics. However, they must be viewed in light of the above statements.

1.1. Natural Products

Bile Preparations. Substances native to the body are supplied as choleretics in the form of bile preparations and their derivatives. They stimulate the secretion of bile in the liver. The active agents in these preparations are sodium glycocholate [*475-31-0*] and sodium taurocholate [*81-24-3*]. (Glycocholic and taurocholic acids are formed by conjugation of the bile acid cholic acid with glycine or taurine, respectively.) Further choleretics are the naturally occurring bile acids (e.g., cholic and deoxycholic acids) and dehydrocholic acid, which is obtained by oxidizing cholic acid .

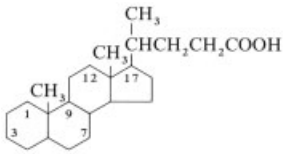

Cholanic acid [*25312-65-6*]

Deoxycholic acid [*83-44-3*]: 3,12-dihydrocholanic acid
Cholic acid [*81-25-4*]: 3,7,12-trihydroxycholanic acid
Dehydrocholic acid [*81-23-2*]: 3,7,12-trioxocholanic acid

Trade Names. Decholin (Casella-Riedel, Frankfurt), Agicholit (Brunnengräber, Lübeck), Chol-Arbuz (Schwag, München), Gallo Sanol (Sanol, Monheim).

Plant Drugs. Drugs obtained from plants constitute the largest number of cholagogues and are used in tea mixtures or other preparations. In addition, special products also contain bile acids, spasmolytic agents, and substances that stimulate digestion.

The most frequently used medicinal herbs are *Berberis vulgaris* (barberry), *Chelidonium majus* (celandine), different types of turmeric, *Cynara scolymus* (artichoke), *Peumus boldus* (a South American plant), *Mentha piperita* (peppermint), *Raphanus sativus* (black radish), and

Taraxacum officinale (dandelion). Very few of the drugs extracted from these plants have been investigated chemically or pharmacologically. The active ingredients of turmeric and artichoke are described in Sections 1.2.1 and 1.2.2.

Trade Names. Cholagogum capsules (Nattermann, Köln), Cholagutt-A (Albert-Roussel, Wiesbaden), Choleodoron (Weleda, Schwäbisch Gmünd), Galleb (Hover, Neuss), Heparaxal (Roland, Essen), Bilicura (Müller, Göppingen), Hepata (Madaus, Köln).

1.2. Synthetic Products

1.2.1. Benzyl Alcohol Derivatives

The choleretic ingredient of the plant drug turmeric (*Curcuma*) is *p*-tolylmethylcarbinol. Synthetic choleretics based on this structure have been developed by lengthening the aliphatic carbon chain or modifying the substituents of the aromatic ring. Reaction of the carbinol with a dicarboxylic acid gives a monoesterified acid whose sodium and diethanolamine salts are water-soluble [6]. Among the large number of compounds synthesized, the following are of practical importance.

1-Phenylpropanol [*93-54-9*], $C_9H_{12}O$, M_r136.19, *bp* (101.3 kPa) 219 °C.

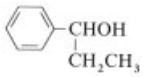

1-Phenylpropanol is obtained (1) by reaction of ethylsodium with benzaldehyde in toluene and decomposition of the reaction product with water or (2) from benzaldehyde and ethylmagnesium halogenides [5]. Its toxicity is low; the oral LD_{50} in rats is 1500 mg/kg [1], [2], [6].
Trade Name. Felicur (Asche, Hamburg).

1-Phenylpentanol [*583-03-9*], fenipentol, C_{11} $H_{16}O$, M_r164.25, *bp* (1.62 kPa) 123 – 124 °C, is obtained by condensing styrene oxide with ethyl acetoacetate to form α-aceto-γ-phenyl- γ -butyrolactone. Decarboxylation and Wolff –Kishner reduction of this compound give 1-phenylpentanol [7] (see next page, on top).

The oral LD_{50} in mice is 3.10 g/kg [3], [8], [9].
Trade Name. Cholipin (Thomae, Biberach).

1.2.2. Cinnamic Acid Derivatives

Cynarin [*30964-13-7*], 1,4-bis(3,4-dihydroxycinnamoyloxy)-3,5-dihydroxycyclohexanecarboxylic acid, $C_{25}H_{24}O_{12}$, M_r516.44, *mp* 225 – 227 °C.

For structural elucidation and synthesis, see [10]. Cynarin is a cinnamic acid derivative found in artichokes. It possesses choleretic properties, lowers the cholesterol level in blood, and is a mild diuretic [11].

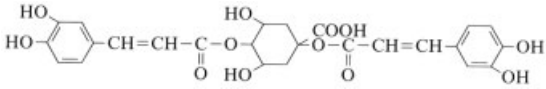

Toxicity is low, although use of high dosages for long periods can cause a pathological decrease in liver and kidney function [12].
Trade Name. Cynarix (Sagitta, Feldkirchen).

Dimecrotic acid [*7706-67-4*], 2,4-dimethoxy-β-methylcinnamic acid.

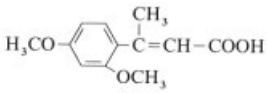

Dimecrotic acid is a choleretic agent with spasmolytic properties. It is obtained by reaction of resorcinol with acetoacetate to give 4-methyl-7-hydroxycoumarin. Treatment with dimethyl sulfate and sodium carbonate gives the final product [13].
Trade Name. Hepadial (Lab. Biocodex, Montrouge Cedex).

1.2.3. Miscellaneous Synthetic Choleretics

Cyclobutyrol [*512-16-3*], α-ethyl-1-hydroxycyclohexaneacetic acid, $C_{10}H_{18}O_3$, M_r186.24, *mp* 81 – 82 °C; sodium salt, *mp* 299 – 300 °C.

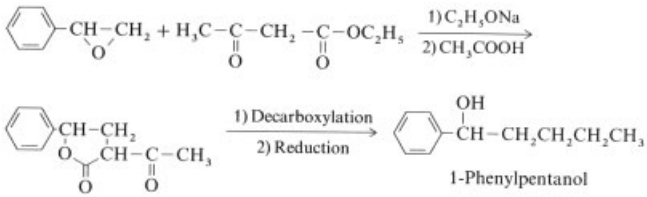

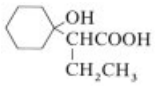

Cyclobutyrol is a choleretic without cholekinetic action [14]. The ethyl ester is obtained via the Reformatsky reaction by heating cyclohexanone with ethyl α-bromobutyrate in anhydrous benzene in the presence of zinc oxide. The ester is saponified to give the product [15].
Trade Name. Hebucol (Logeais, Issy-les-Moulineaux).

β-(1-Methoxy-4-naphthoyl)propionic acid [*3562-99-0*], $C_{15}H_{14}O_4$, M_r258.26, *mp* 172 – 173 °C.

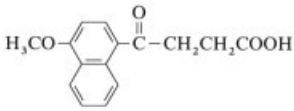

This substance proved to be the most effective compound in tests carried out on ketocarboxylic acids derived from naphthalene. It is obtained by reaction of 1-methoxynaphthalene with succinic anhydride in the presence of Friedel – Crafts catalysts [16].
*Trade Names.*Magnesium salts: Icteryl, Hepalande (Delalande, Courbevoie).

***N*-(Hydroxymethyl)nicotinamide** [*3569-99-1*], $C_7H_8N_2O_2$, M_r152.15, *mp* 141 – 142 °C.

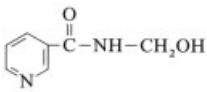

N-(Hydroxymethyl)nicotinamide is obtained by reaction of nicotinamide with formaldehyde in the presence of potassium carbonate [17]. It is a chemotherapeutic agent used against various cocci, *Escherichia coli, Proteus vulgaris,* and other pathogenic agents involved in inflammation of the biliary tract and the gallbladder. The nicotinamide moiety is responsible for its choleretic and spasmolytic action [18].
Trade Name. Bilamid (Cilag-Chemie, Alsbach-Bergstraße).

Anethole trithione [*532-11-6*], 5-(4-methoxyphenyl)-3*H*-1,2-dithiole-3-thione, $C_{10}H_8OS_3$, M_r240.36, *mp* 111 °C, is obtained as follows [19]:

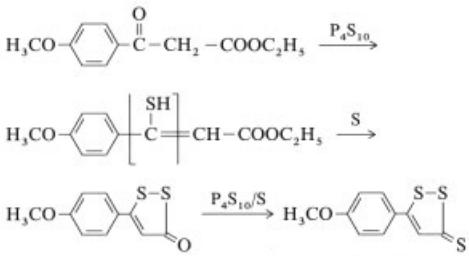

Anethole trithione has choleretic and mild cholekinetic properties. It also increases the formation and excretion rates of urea and thus has a diuretic effect [20].
Trade Name. Mucinol (Plantorgan, Bad Zwischenahn).

Azintamide [*1830-32-6*], 2-[(6-chloro-3-pyridazinyl)thio]- *N,N*-diethylacetamide, $C_{10}H_{14}ClN_3OS$, M_r259.77, *mp* 97 – 98 °C, can be obtained by treatment of 3,6-dichloropyridazine with hydrogen sulfide and subsequent reaction of the 6-mercapto compound with *N,N*-diethylchloroacetamide:

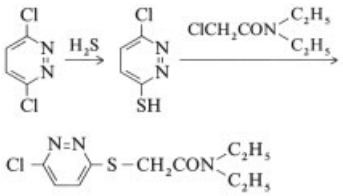

Azintamide can also be obtained by reacting the acetyl chloride derivative of the mercapto compound with diethylamine [21]:

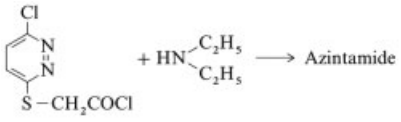

Azintamide is highly effective and is quickly absorbed from the intestine. It increases bile flow and has neither an excitatory nor an inhibitory effect on the central nervous system [22].
Trade Name. Orgallin (Hormon Chemie, München).

Hymecromone [*90-33-5*], 7-hydroxy-4-methylcoumarin, $C_{10}H_8O_3$, M_r176.16, *mp* 194 –195 °C.

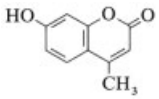

Hymecromone is obtained by reaction of resorcinol with acetoacetate [23]. It is a spasmolytic agent that selectively acts on the bile ducts, particularly on Oddi's sphincter. Sphincter diastole is lengthened, and bile flow into the intestine increases. Hymecromone also possesses choleretic properties [24].
Trade Name. Cantabilin (Medica, Lipha, Lyon).

Moquizone [*19395-58-5*], 2,3-dihydro-1-(4-morpholinylacetyl)-3-phenyl-4(1 *H*)-quinazolinone, $C_{20}H_{21}N_3O_3$, M_r351.41, *mp* 135 – 137 °C; hydrochloride, *mp* 210 – 214 °C.

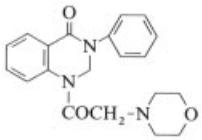

Moquizone is synthesized by reacting 2,3-dihydro-4(1 *H*)-quinazolinone with chloroacetyl chloride and subsequently with morpholine [25].

Moquizone exhibits choleretic effects in animal experiments [26].
Trade Name. Peristil (Recordati, Milano).

2. Dissolution of Gallstones

Chemical gallstone dissolution (litholysis) can be achieved by four methods:

1) oral therapy with the bile acids chenodeoxycholic or ursodeoxycholic acid,
2) direct percutaneous – transhepatic lysis with methyl *tert*-butyl ether,
3) oral litholysis after extracorporeal shockwave lithotripsy,
4) direct dissolution of bile duct stones via a nasobiliary tube or T-tube.

2.1. Gallstone Pathophysiology

Classification of Gallstones. Gallstones consist of a mixture of cholesterol, pigments, calcium bilirubinate, inorganic calcium salts, electrolytes, heavy metals, fatty acids, mucins and mucoproteins. The predominant component in the stone provides its name. The most common gallstone found in patients in industrialized countries is the *cholesterol stone; pigment stones* occur less frequently. Cholesterol stones contain at least 75 % cholesterol, whereas pigment stones contain no more than 25 % cholesterol [27]. An interesting pigment stone is the black pigment stone; little is known about its pathogenesis and composition. Concretions consisting predominatly of fatty acids are extremely rare. The composition, incidence, and physiological sites of occurrence of gallstones are summarized in Table 1.

Table 1. Classification, incidence, and localization of human gallstones

Type of stone	Composition	Prevalence	Localization
Cholesterol stone	$\geqq$70 % cholesterol, pigments, calcium salts, mucins, proteins	70 – 90 % of all gallbladder stones [28]	gallbladder; in 15 % of patients, gallbladder plus common bile duct
Composition stone (mixed stone)	like cholesterol stones, but with significantly lower cholesterol content (30 – 70 %)	20 %; higher in older patients	like cholesterol stones
Calcium bilirubinate stone (brown pigment stone)	$<$ 30 % cholesterol, high calcium bilirubinate content	60 % of all common bile duct stones after cholecystectomy	common bile duct; rare in the gallbladder
Black pigment stone	black pigment, mucins, proteins	5 – 30 % of all gallbladder stones [28]	gallbladder; rare in the common bile duct

Stone Formation. Formation of cholesterol stones is induced by a hepatic metabolic disorder. In the liver of gallstone carriers, the activity of β-hydroxy-β-methylglutaryl coenzyme A (HMG-CoA) reductase – the key enzyme of cholesterol synthesis – is increased. In addition, 7-α-hydroxylase, which catalyzes one of the steps in the conversion of cholesterol to bile acids, is diminished [29]. This results in high secretion of cholesterol and low secretion of bile acids into bile. Bile acids are essential for the micellar dissolution of water-insoluble cholesterol. Thus, when micellar capacity is exhausted by increased cholesterol output, cholesterol precipitates and forms cristals. A glycoprotein, probably secreted by the gallbladder mucosa, serves as a nucleation factor in initial stone formation [30]. Mucin production by the gallbladder wall and also gallbladder function play an important role in stone formation. Furthermore, stone formation depends on the sex of the patient and is affected by diet, body weight, race, and diseases of the liver and the intestinal tract [30–33].

Little is known about the formation of black pigment stones. Their incidence is higher in urban populations and in patients who consume a protein-rich diet. Hemolytic diseases associated with a surplus of unconjugated bilirubin are relevant. Brown calcium bilirubinate stones in the billiary tree form after increased bacterial deconjugation of bilirubinate – diglucuronide in bile duct infections. Therefore, these stones are found mostly after cholecystectomy or as a relapsing (recurrent) stone after endoscopic papillotomy. They often occur in rural populations with a low-protein diet.

2.2. Oral Dissolution of Gallbladder Stones with Bile Acids

Oral administration of the bile acids *chenodeoxycholic acid* [*474-25-9*] (cheno) or *ursodeoxycholic acid* [*128-13-2*] (urso) is indicated only in patients with cholesterol gallbladder stones. Chenodeoxycholic acid is given in a dosage of 15 mg per kilogram body weight daily. Dissolution rate is dose dependant and is augmented by a low-cholesterol diet [34]. Urso preparations are given in a daily dosage of 10 mg per kilogram body weight daily. Higher doses are of no avail [35].

Mode of Action of Bile Acids. Chenodeoxycholic acid acts by inhibiting HMG-CoA reductase and the 7-α-hydroxylase in the liver (cf. Section 2.1) [29]; this diminishes the secretion of cholesterol and bile acid into bile. The administered chenodeoxycholic acid is also secreted into bile, where it solubilizes the cholesterol incorporated in the stones by means of micelle formation.

Ursodeoxycholic acid has little influence, if any, on HMG-CoA reductase and the micellar dissolution of stone cholesterol. It acts by forming cholesterol liquid crystals

(phospholipid – bile salt – cholesterol vesicles) and inhibiting cholesterol reabsorption in the intestine [36], [37].

Cheno- and ursodeoxycholic acid act differently; both mechanisms are used for chemical stone dissolution. Dosage reduction of the individual bile acid is possible in combination therapy [38] and diminishes their side effects, i.e., diarrhea and elevation of transaminases with chenodeoxycholic acid, or stone calcification with ursodeoxycholic acid.

Patient Selection. Therapeutic success depends largely on patient selection. Criteria for successful treatment are

1) radiolucent cholesterol stones in the gallbladder,
2) gallbladder less than 50 % filled,
3) stone diameter $\leq$1.5 cm,
4) functioning gallbladder after stimulation,
5) patent (i.e., unobstructed) cystic duct,
6) patent common bile duct, and
7) patient compliance.

Contraindications for oral stone dissolution are as follows:

1) pigment stones and calcified stones,
2) stone diameter >1.5 cm,
3) stones in the cystic duct or common bile duct,
4) acute or chronic cholecystitis,
5) suspected bile duct carcinoma,
6) marked gallbladder septa,
7) acute or chronic liver disease,
8) chronic intestinal inflammation,
9) acute gastric or duodenal ulcer, and
10) frequent severe colics or constant pain.

Ultrasonic imaging reveals whether stones are present. A subsequent X-ray shows ring-shaped or totally calcified stones. Oral or intravenous cholecystography allows differentiation between pigment and cholesterol stones and provides information regarding stone number, age, and central spotlike calcifications.

According to [39], only 24 % of gallstone patients are suitable for conservative medical treatment; a second study reports a figure of 59 % [40]. This discrepancy is a result of the composition of the patient cohort: in the first study, only patients with advanced gallstone disease were examined; in the second study, early forms of stone disease were investigated.

Clinical Results. Complete stone dissolution is achieved with monotherapy (cheno: 15 mg kg^{-1} d^{-1}; urso: 10 mg kg^{-1} d^{-1}) in ca. 60 % of all patients treated (Table 2) [41–43]. Combination therapy (cheno + urso, half dosage each) is reported to induce dissolution more rapidly than monotherapy, according to a study in 120 patients [38]. The therapeutic result depends mostly on stone size and treatment time. Small concretions of about 0.5 cm can be dissolved in 80 – 90 % of patients, but stones with a diameter of 2 cm rarely dissolve [41], [42], [44]. The mean treatment time for stones up to 1.5-cm diameter is 19 months [45].

About 40 % of successfully treated patients experience stone recurrence [46]; these stones are usually cholesterol stones that can be dissolved within a few months. Recently it has been shown that the recurrent stone should be treated immediately after its detection and not just after colics or abdominal pain has developed. Using this treatment scheme 80 – 90 % of the patients remain stone free after 5 years.

2.3. Direct Percutaneous – Transhepatic Lysis (PTL) of Gallbladder Stones

Cholesterol gallbladder stones were first dissolved with percutaneous – transhepatic lysis (PTL) in 1985 [45]. Dissolution was performed with 2 – 10 mL of methyl *tert*-butyl ether [*1634-04-4*] (MTBE) which was instilled directly into the gallbladder and reaspirated immediately. Treatment requires 1 – 2 d and produces a few mild side effects (e.g., nausea, vomiting, and mild upper abdominal pain) [47]. Leakage of bile into the peritoneal cavity after withdrawal of the perfusion catheter presents a certain risk.

The therapeutic success of PTL is greater than 90 %, at a cost of ca. 200 DM ($ 100) per patient [48]. Further therapeutic results are summarized in Table 3. MTBE-therapy is contraindicated by the following:

1) calcified stones and pigment stones,
2) obstruction of the cystic duct,
3) acute or chronic cholecystitis,
4) small gallbladder bed area, and
5) gallbladder anomalies.

Table 2. Gallstone dissolution in patients treated with chenodeoxycholic acid (cheno), ursodeoxycholic acid (urso), or cheno vs. urso.

Bile acid ∗	Patients		Dosage	Stone		Patients with
	No.	Body weight, kg	mg kg^{-1} d^{-1}	No. (mean)	Mean size, cm^2	complete stone dissolution, %
Cheno	28	65.4	15.0	8	4.9	60.7
Urso	135	63.8	11.3	12	4.0	62.2
Cheno vs.	20	60.9	15.8 cheno	8.5	1.1	61.0
urso	28	64.5	11.5 urso	17	0.3	85.0

∗ Patients were treated with Chenofalk and Ursofalk (Falk Pharma, Freiburg, Federal Republic of Germany).

2.4. Oral Litholysis after Extracorporeal Shock-Wave Lithotripsy (ESWL)

Extracorporeal shock-wave lithotripsy (ESWL) is based on the concentration of an energy wave produced by spark discharge to a radiologically or sonographically localized gallbladder stone [49]. However, because the resulting stone fragments remain in the gallbladder, ESWL is possible only in combination with oral litholysis; therefore, patients must be selected carefully (cf. Section 2.2). After approximately six months, the combination method ESWL – litholysis completely removes solitary stones with a diameter of 2 cm in about 80 % of patients. Patients with three 1-cm stones become stone free after 18 months. Disadvantages of the treatment are its very high cost (the gallstone lythotriptor costs ca. 3.2×10^6 DM $\approx$ \$ 2×10^6), and only three concretions can be fragmented in any one patient. Furthermore, ESWL is unsuitable for calcified cholesterol and pigment stones.

Table 3. Results of stone dissolution with MTBE – data from 21 European centres

Number of patients scheduled for therapy	803
Placement of catheter successfully	761 (94.8 %)
Mean stone number and range	7.0 (1 - 68)
Mean stone size (cm) and range	1.6 (0.3-4.5)
Patients with solitary stones	197/761 (25.9 %)
Patients with multiple stones	564/761 (74.1 %)
Stone dissolution (> 90 % of stone burden)	724/803 (95.1 %)
	724/724 (90.2 %)
Residual sludge in the gallbladder after catheter removal	315/724 (43.5 %)
Bile leak after catheter removal	39/761 (5.1 %)
Patients with minor complications	129/803 (16.1 %)
Method related lethality	0/803 (0 %)
30-day lethality	3/803 (0.4 %)

2.5. Direct Dissolution of Bile Duct Stones

Bile duct stones are removed endoscopically in over 90 % of patients [50], [51]. Chemical dissolution is seldom indicated [52]. For cholesterol stones, 2 – 5 mL of glyceryl monooctanoate is infused via a T-tube or a nasobiliary tube into the bile ducts. Complete stone dissolution has been achieved in 50 – 60 % of 600 patients treated. Side effects are epigastric pain and diarrhea. Average treatment time is three weeks. Methyl *tert*-butyl ether is not suitable (or only suitable in the hands of an expert) for dissolution because it flows into the intestine where it is reabsorbed and has an anesthetizing effect (see also Table 3).

Calcium bilirubinate concrements account for up to 60 % of all bile duct stones found after cholecystectomy and can be dissolved by perfusion with an alkaline, buffered ethylenediaminetetraacetate solution (10 – 15 mL/h) [53]. In vitro, this solution dissolves almost all calcium bilirubinate stones and even black pigment stones within 30 h [54]. The aqueous solution is adjusted to pH 9.4 with sodium hydroxide and contains the following:

disodium ethylenediaminetetraacetate	1.0 %
cholic acid	0.5 %
ursodeoxycholic acid	0.5 %
carnosine	0.5 %
arginine	1.9 %

The ethylenediaminetetraacetate solution acts by chelating calcium from calcium bilirubinate to produce sodium bilirubinate which is water soluble [55]. In combination with PTL – MTBE therapy (described in Section 2.3), it can also be used for the direct treatment of pigment stones in the gallbladder.

2.6. Critical Assessment of Chemical Stone Dissolution

Any new method for treatment of gallstones must be compared with surgery (cholecystectomy) which has been practiced since 1882. Cholecystectomy has an age-dependent lethality of 0.1 – 4 % (mean 0.4 %) and is a well-established, low-risk method. However, despite modern presurgery diagnostics, 25 – 40 % of patients who have undergone surgery suffer from the postcholecystectomy syndrome (i.e., from the same complaints as before surgery) [56], [57]; in ca. 10 – 15 % of all patients, recurrent stones develop in the bile ducts [50].

The advantages of oral stone dissolution with cheno- or ursodeoxycholic acid are the easy administration of the preparation and the absence of fatalities. A disadvantage is the high rate of stone recurrence (ca. 40 %). The cost is identical to that of surgical therapy.

A final assessment of PTL – MTBE treatment is still premature. Extracorporeal shock-wave lithotripsy with subsequent oral litholysis may be useful in the treatment of stones with diameters >1.5 cm.

A logical therapeutic concept for treating gallstones is as follows:

1) Silent gallbladder stones should not be treated but should be monitored.
2) Cholesterol stones in a functioning gallbladder, which have a diameter of 1.5 – 2.0 cm and produce mild symptoms, are suitable for oral litholysis.
3) Cholesterol stones with a diameter of 1.5 – 2.0 cm in a functioning gallbladder are suitable for the combination for ESWL therapy and oral dissolution with bile salt.
4) Cholesterol stones in a functioning gallbladder, which measure up to 3 – 4 cm or fill the gallbladder to more than 50 % and produce marked symptoms, are suitable for PTL – MTBE therapy.
5) Surgery is indicated in patients with frequent colics and all complications (e.g., cholecystitis, shrunken gallbladder, empyema, obstructed cystic duct, calcified stones, and pigment stones).
6) Since laparoscopic cholecystectomy has been introduced in 1985, oral litholysis, ESWL therapy and contact litholysis with MTBE are less important in the treatment of gallstone disease. The advantages of laparoscopic operations are: a lower lethality rate, a shorter hospital stay, rapid reconvalescence. The disadvantages are: a higher morbidity rate and higher costs.

3. Liver Therapeutics

3.1. Introduction

Drug therapy for liver disease is unsatisfactory from many points of view. Because the liver occupies a key position in metabolism, it has an enormous amount of reserve energy that is used to only a small extent under normal circumstances. It receives a more than adequate supply of the building blocks required for its varied, rapid metabolic reactions via the portal vein and the hepatic artery. Like all other parenchymatous organs, the liver requires stimulation of its functions to maintain its normal structure [58].

Partial, temporary, or permanent collapse of liver function caused by serious injury to the organ cannot be overcome by drug administration. Similarly, irreversible structural changes cannot be reversed by pharmacotherapy [59].

Prophylactic therapy can be used to prevent liver damage. Additional, medicinal protection is possible to only a limited extent and is normally unnecessary because of the liver's enormous reserve energy supply and regenerative capacity. This might appear improbable in view of the range of preparations available on the European market for the "protection" of the liver. However, critical analysis of these preparations shows that they are:

1) experimental products used without any scientific basis,
2) substances uncritically recommended for use in humans on the basis of animal experiments, or
3) products chosen on the basis of speculation and inadequate knowledge of intermediary metabolism.

Liver disease is caused by many different pathogenic agents; consequently, no single drug can be effective against all types of disease. The task of the clinician is to find out which

therapeutic measures can be employed to restore liver functions that have failed partially or totally. Potentially successful therapeutic substances should possess the following pharmacological properties:

1) Stimulation of the lowered cell metabolism
2) Stimulation of the reticuloendothelial system, particularly in the hepatic region, thereby causing an increase in the phagocytic capacity of the reticuloendothelial Kupffer cells
3) Maintenance or recovery of the intra- and extracellular structure of the membrane and its functions
4) A positive regenerative effect
5) An anti-inflammatory effect
6) Antitoxic properties
7) Good absorption from the digestive tract
8) Preferential excretion in the enterohepatic circulation
9) Lack of properties that are toxic or impose a stress on liver cells
10) Enhancement of detoxification by effective entrapment of radicals
11) Absence of side effects

The vitamins, amino acid cocktails, and other preparations available for protection of the liver should be considered obsolete. Only a few substances are important for the treatment of liver disease; these include the corticosteroids, immunosuppressive agents, and silymarin [60], [61].

Almost all the therapeutic measures discussed below had their origins in the United States, where they have long since been forgotten. However, they seem to have an undeserved ability to survive in continental Europe [59], [62], [63].

The classical experiments of Opie and Alford proved that animals fed a carbohydraterich diet could withstand chloroform-induced narcosis better than those receiving a low-carbohydrate diet. Liver glycogen has since been considered to be one of the factors responsible for the protection of liver cells. Subsequent animal experiments demonstrated that chloroform sensitivity was associated with protein, rather than sugar, metabolism; thus proteins became of interest. In rats, deficiency or lack of the lipid-mobilizing (lipotropic) substances choline and methionine caused fatty liver and led to cirrhosis, whereas a diet lacking cystine (or cysteine), methionine, and vitamin E resulted in liver necrosis [60], [64], [65]. These animal experiments, however, were carried out under extreme conditions that are unlikely to occur in humans. Choline, cysteine, and methionine were used either singly or in combination in most of the so-called preparations for "protection" of the liver. As expected, they never produced convincing results [66–68].

Different vitamins, particularly those belonging to the B group, form the "hard core" of liver protection therapy. Of these, vitamin B_{12} is the most important. Although the ineffectiveness of vitamin B_{12} therapy in acute hepatitis, chronic hepatitis, and liver cirrhosis has been proved biochemically and clinically [60], it is still remarkably popular. The effectiveness of various vitamin combination drugs has not yet been proved by controlled chemical trials.

3.2. Therapeutic Agents

Liver diseases are heterogeneous and differ in their etiology. This fact must be taken into consideration in their treatment. After careful diagnosis, a suitable substance must be selected to achieve a therapeutic effect.

3.2.1. Inhibitors of Viral Nucleic Acid Synthesis

Acyclovir [*59277-89-3*], 9-(2-hydroxyethoxymethyl)guanine, is a nucleoside with very low toxicity and strong virostatic properties (→ Chemotherapeutics).

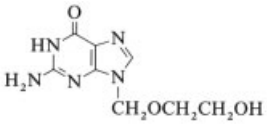

In herpes virus infections, acyclovir must first be activated to acyclovir monophosphate by the viral thymidine kinase and then converted to acyclovir triphospate by cellular phosphotransferases. Acyclovir triphosphate inhibits viral DNA polymerases by competitively displacing deoxyguanosine. Although the hepatitis B virus (HBV) does not possess any endogenous thymidine kinase, a transient decrease in HBV-specific

DNA polymerase has been described in some cases [69]. Intravenous administration was accompanied by some serious side effects, while very low concentrations in blood plasma could be detected after oral administration of the drug. Proacyclovir (6-desoxyacyclovir) BW515 is absorbed very well from the intestine. Results indicate that acyclovir and proacyclovir could become valuable adjuvant drugs in a virostatic combination therapy for chronic hepatitis B [70–73] .
*Trade Name.*Zovirax (Wellcome, Burgwedel)

Intercalary Substances. Some substances that can inhibit DNA polymerase in vitro (quinacrine, chloroquine, chlorpromazine, and primaquine) have also been tested in clinical studies on patients with chronic hepatitis B. Therapeutic efficiency has not yet been proved [72], [74], [75].

Interferon. Interferon comprises a set of biological proteins with antiviral and immunomodulatory activity (→ Interferons). Interferons can alter the course of viral infection by inhibiting intracellular virus replication and by modifying the immune responses to viral antigens in body fluids and on cellular surfaces [76–78]. The α- and β-interferons are induced by viral replicative events in all types of cells and are thus probably important during the early phases of an infection [58]. γ-Interferon is produced as a result of the formation of antigen-sensitized T cells and probably exerts its effect during later stages or during chronic infection [77], [79], [80].

Phosphonoformic Acid [*4428-95-9*].
Sodium phosphonoformate [*63585-09-1*] is a pyrophosphate analogue that selectively inhibits viral nucleic acid synthesis (→ Chemotherapeutics). The formation of triphosphate nucleotides by polymerases, which is essential for the synthesis of RNA and DNA, is prevented by this substance [78]. The HBV DNA polymerase is also inhibited by phosphonoformic acid in vitro and in experiments on chimpanzees. However, the substance does not appear to be suitable in the treatment of humans, because the phosphate moiety may accumulate in the bones and delayed toxic effects cannot be ruled out. Moreover, it must be administered intravenously at relatively short intervals, which makes long-term treatment difficult [72], [74], [77].

Ribavirin [*36791-04-5*], 1-β-D-ribofuranosyl-1*H*-1,2,4-triazole-3-carboxamide, is a nucleoside analogue that acts against a broad spectrum of RNA and DNA viruses in vitro by inhibiting the biosynthesis of guanosine monophosphate. It accumulates in the liver but does not appear to affect the reproduction of HBV. Therapeutic tests show that the substance does not favorably influence the course of chronic hepatitis B [80], [81].

Vidarabine [*24356-66-9*], 9-β- D-arabinofuranosyl-adenine (ara-A) and its monophosphate (ara-AMP) [*29984-33-6*] are purine nucleosides with strong virostatic properties (→ Chemotherapeutics). They selectively inhibit the activity of viral DNA polymerase and thereby prevent synthesis of the complementary minus-strand DNA from the viral pregenome [82], [83].

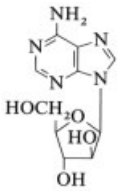

Vidarabine

The decrease in DNA polymerase activity specific to hepatitis B virus is faster and greater after treatment with ara-A or ara-AMP than with interferon. However, when treatment is discontinued, DNA polymerase activity usually reaches the initial values observed before treatment. Continued inhibition of HBV replication after discontinuation of therapy occurs about as often as after treatment with interferon [84]. In comparison with the monophosphate ester ara-AMP, ara-A has the disadvantage of requiring large infusion volumes in clinical therapy. Generally, therefore, the more water-soluble derivative ara-AMP is used. The monophosphate ara-AMP can also be administered intramuscularly and is suitable for treatment of outpatients. However, daily dosages of 5 – 10 mg per kilogram of body weight for a period of 28 days are only

slightly effective. Longer treatment periods or higher doses lead to serious, extremely painful and long-lasting polyneuropathies that disappear very slowly after discontinuation of treatment. This neuromuscular pain syndrome begins mainly in the lower extremities in the form of a "burning feet syndrome." Other side effects include tiredness, listlessness, loss of appetite (occasionally with mild nausea), and slight depletion of bone marrow which is reversible after discontinuation of the drug [73], [82].
Trade Name. Vidarabin (Dr. Thilo, Sauerlach)

3.2.2. Bile Salts used for Primary Biliary Liver Disease

Ursodeoxycholic acid is a naturally occurring bile salt in humans. It is rapidly absorbed in the small intestine and excreted with the bile. Ursodeoxycholic acid is the drug of choise for the treatment of primary biliary cirrhosis and primary sclerosis cholanyitis. The daily dose amounts to 10 – 15 mg/kg. It has no side effects even after a treatment period of up to 20 years.

In primary biliary cirrhosis, it improves the laboratory data significantly by 50 – 80 % and the symptoms in 30 – 40 % of the patients. Ursodeoxycholic acid prevents or delays the development of oesophageal varices and prolongs life expectancy. Better results are obtained when ursodeoxycholic acid is combined with the glucocorticoid budesonide.

3.2.3. Natural Substances

Cyanidanol [*154-23-4*], Catergen, (+)-catechin, (+) 3-cyanidanol, (2*R*,3*S*)-3,3′,4′,5,7-pentahydroxyflavan, was isolated in 1830 from a series of plants of African, American and Indian origin (*Uncaria gambir*). Unequivocal elucidation of its stereochemical structure was carried out in the late 1950s [85].

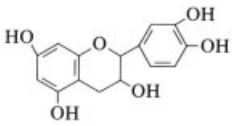

Cyanidanol is absorbed rapidly from the gastrointestinal tract. Its concentration in blood reaches a maximum after 1 – 2 h. Excretion takes place mainly as the glucuronide and sulfate conjugates in the urine [86–88].

Unchanged cyanidanol is present in only small amounts in the liver because almost all of it is metabolized rapidly during its first passage through the organ. Eleven metabolites have been isolated [89–92].

Investigations with radioactively labeled cyanidanol have shown that about 27 % of the administered dosage is excreted in the urine after 3 h and between 70 and 88 % after 27 h. This catechol lacks liver specificity because of its rapid urinary elimination. Furthermore, the substance is inactivated where it is supposed to act, i.e., in the liver. Because of its serious, sometimes fatal side effects, this compound was withdrawn from the market in 1987 [92–94].

Silymarin is obtained from the milk thistle – *Silybum marianum* (L.) Gaertn. or *Carduus marianus* (L.) – a medicinal plant known since ancient times. For structural elucidation and total synthesis of the constituents of silymarin, see [94], [95]. Silymarin is a general term used to describe three substances derived from the milk thistle fruit: silybin [*22888-70-6*], silydianin [*29782-68-1*], and silychristin [*33889-69-9*]. For structural elucidation of these compounds, see [96], [97].

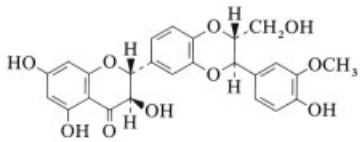

Silybin

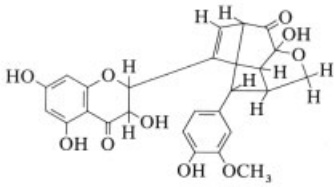

Silydianin

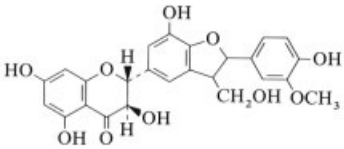

Silychristin

The effects of silymarin and its main constituent, silybin, were investigated on different models of liver damage in acute, subchronic, and chronic experiments on animals. Treatment was either prophylactic or curative [98–100].

Silymarin is active against a large number of hepatotoxic substances. It exerts a positive influence on several basic cellular functions, namely, membrane integrity, of ribosomal RNA and thus of proteins, and the regenerative capacity of the liver [101], [102]. Most of these functions are influenced only in the damaged liver and not in a healthy liver [103–105].

Preparations like silymarin have a protective and stabilizing effect on the liver cell membrane due to their radical scavenger function and the resultant preventation of lipid peroxidation. Furthermore, silymarin has an immunomodulatory effect (normalization of the reduced lymphoblastic transformation activity of cirrhotic patients). Both the immunomodulatory and antioxidant activities of silymarin (Legalon) play an important role in the mode of action of this hepatoprotective preparation [64], [65], [101], [106–108].

Pharmacokinetics. About 90 % of silymarin is excreted in the bile; renal elimination accounts for ca. 10 %. The compound is conjugated in the form of glucuronides, sulfates, and sulfaglucuronides in the liver. These conjugates are then cleaved by the intestinal flora, and the active ingredient is reabsorbed. About 40 % of the absorbed substance enters this cycle, i.e., the enterohepatic circulation. Silymarin does not induce enzyme synthesis; interaction with other drugs has not been observed [64], [65], [109], [110].

Use in Alcohol-Induced Liver Disease. The therapeutic action of silymarin in diseases of the liver caused by alcohol has been well established [111–113].

The efficacy of Legalon was investigated in a double-blind study of alcohol-induced liver damage. The serum (GPT) glutamic – pyruvic transaminase and glutamic – oxalacetic transaminase (GOT) values, bromsulphalein (BSP) retention, and histological changes normalized more rapidly and more frequently in patients receiving the drug than in the control group [114].

The efficacy of silymarin was shown in patients with alcohol-induced liver disease in a randomized double-blind study. The serological parameters bilirubin, GOT, GPT, glutamate dehydrogenase (GLDH), γ-glutamyltransferase (γ-GT) and alkaline phosphatase were chosen for monitoring. Other serological and sonographic findings were used as general clinical controls. The results, particularly changes in γ-GT, GPT, GOT showed that Legalon produced significant improvement [113], [115].

Use in Occupational Liver Disease. Silymarin treatment of patients suffering from occupational liver disease led to considerable improvement of parameters that showed elevated pathological values: alkaline phosphatase, bilirubin, GOT, GPT, γ-GT, and cholinesterase [116–118].

The action of Legalon on liver function after cholecystectomy was investigated in a double-blind study. The values for GPT and γ-GT showed a statistically significant difference in favor of Legalon [61], [63].

The above results confirm other investigations on the postoperative behavior of cholinesterase and various hepatic enzymes in the presence of Legalon [112], [119].

Silymarin produced an antihepatotoxic effect, as well as a decrease in lipoprotein and triglyceride concentrations, in patients with liver damage resulting from industrial toxicity caused by silicon dioxide, lead, trinitrotoluene, and trichloroethylene. In workers from the chemical industry who were exposed continuously to xylene and/or toluene, Legalon had a signifcant protective effect on the liver [59], [60], [64].

Use in Liver Cirrhosis. In a randomized double-blind study, 170 patients with cirrhosis of the liver were treated with silymarin: 87 patients received 140 mg Legalon three times daily, 83 patients received a placebo. The average observation time was 41 months for the placebo group and 50 months for the group receiving Legalon. The survival rate was 76 ± 6 % (standard error) in patients treated with silymarin and 53 ± 8 % in the placebo group. Analysis of the subgroups indicated that treatment was effective in alcoholic cirrhosis as well as in cases classified as a cholangic biliary cirrhosis (an ailment affecting children up to 12 years of age). These results show that mortality in cirrhosis patients is lowered by treatment with silymarin [120].

In a multicenter study of *Amanita phalloides* intoxication in the Federal Republic of Germany,

Switzerland, Austria, and France, a survival rate of 91.6 % was observed in 220 patients treated with silybinin. Silybinin infusions in combination with conventional treatment represent an effective therapeutic supplement in patients with *Amanita phalloides* intoxication [106], [121–123].

Trade Names. Legalon 70, Legalon 140, Legalon Sil Ampullen, Legalon Liguidum (Dr. Madaus, Köln).

3.2.4. Miscellaneous Substances

"Essential" Phospholipids. Lecithin has been used in therapy as a hepatotropic substance ever since the liver was shown to incorporate choline into lecithin. "Essential" phospholipids (EPL) are derived from a special fraction of lecithin and isolated by chromatographic methods from a crude mixture of phospholipids obtained from soybean. They consist of naturally occurring phosphatidylcholine with an especially high content of polyunsaturated, essential fatty acids, particularly linoleic acid [124–128].

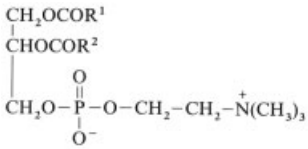

More than 50 % of EPL consists of phosphatidylcholine in which the substituents R^1 and R^2 are derived from linoleic acid:

$$CH_3(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$$

The essential phospholipids differ from the phospholipids present in the body in that R^1 is derived from a polyunsaturated fatty acid (mainly linoleic acid) and not a saturated one [129], [130].

Production. Commercially available crude phosphatides of plant origin are dissolved in ethyl acetate, a lower dichlorohydrocarbon, or a mixture of the two; up to 6 vol % of alcohol may also be present. Stirring with a quantity of aluminum oxide corresponding to at least five times the crude mass of phosphatides and subsequent elution of the adsorbent with alcohol give a product that contains >90 % phosphatidylcholine. Pure phosphatidylcholine can be obtained by using propylene dichloride for desorption of the aluminum oxide, instead of alcohol. After evaporation of the solvent in an inert atmosphere under reduced pressure, phosphatidylcholine is obtained in yields of up to 88 % as an almost colorless, pasty mass.

Mode of Action. Phospholipids are constituents of biological membranes and, therefore, play an important role in membrane-dependent metabolic processes. On the one hand, the properties of therapeutically employed phospholipids containing polyunsaturated fatty acids should correspond to those native to the body. However, the cis double bonds in the polyunsaturated fatty acids prevent parallel alignment of the hydrocarbon chains of the membrane phospholipids. This should lead to a loosening of the tightly packed, micellar phospholipids and thus to a higher transmembrane exchange rate [129–131].

Therapeutic effects. Lipemia caused by ethanol is antagonized by EPL in acute as well as chronic animal experiments. Serum alkaline phosphatase and plasma protein fractions of the plasma tend to return to normal values. Cirrhosis caused by chronic administration of small doses of carbon tetrachloride is largely prevented by the simultaneous administration of EPL. However, these are less effective in treating poisoning due to galactosamine, carbon disulfide, and allyl isothiocyanate [132], [133].

In rats, orally administered EPL were absorbed completely within 24 h and reached the liver almost exclusively via the lymphatic system [134].

The liver absorbs between 10 and 15 % of the administered EPL. Excretion occurs slowly via urine and bile.

Acute viral hepatitis is considered to be amenable to EPL therapy. Administration of EPL shortens the duration of hepatitis A and B. Chronic hepatitis, chronic-aggressive hepatitis, and liver cirrhosis show clear signs of improvement following treatment with EPL [135–137].

Based on the clinical picture and laboratory results obtained after EPL therapy, this substance appears to increase the biochemical efficiency of residual parenchyma tissues. Iatrogenic injuries to the liver, caused by treatment of tuberculosis with isoniazid and rifampicin, can also be alleviated. Treatment with EPL significantly reduces fatty degeneration of liver cells

and alleviates cell membrane injury occurring during the protein-deficiency disease kwashiorkor [138], [139].

The effectiveness of EPL has been proved pharmacologically, and they are also well tolerated clinically. Thus, EPL may be regarded as a liver curative that intervenes in the form of a "substitution therapy" in hepatocyte anabolism and liver disease in which an activation of anabolic processes appears desirable [140–142].
Trade Name. Essentiale (Nattermann, Köln).

Orotic acid [*65-86-1*], 1,3,5-trihydro-2,6-dioxopyrimidine-4-carboxylic acid, uracil-6-carboxylic acid, whey acid, $C_5H_4N_2O_4$, M_r 156.10, *mp* 345 – 346 °C [143–145].

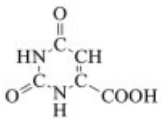

Production. The oldest known method for synthesizing orotic acid is the condensation of urea with diethyl oxalacetate and subsequent alkaline hydrolysis of the reaction product. Free orotic acid is obtained from the alkali salt by treatment with mineral acid [124–126].

A newer synthesis starts by condensing chloroacetoacetyl chlorine, obtained from diketene and chlorine, with urea to give 6-chloromethyluracil. Oxidation of the latter with hydrogen peroxide in an alkaline medium yields orotic acid [134], [146].

In another method, hydantoin or hydantonitrile is treated with glycolic acid in an alkaline medium to give carboxymethylene hydantoin, which rearranges to form orotic acid [145], [147].

Orotic acid is also produced in fermentation by using microorganisms such as *Corynebacterium, Brevibacterium, Arthrobacter,* and *Candida* [148], [149].

Mode of Action. In the liver, orotic acid is the precursor of the pyridine nucleotides of nucleic acid (→ Nucleic Acids). As a parenchymatous, glandular organ, the liver is rich in cell nuclei, where the nucleic acids are localized. Cow's milk also contains a relatively high concentration of orotic acid (6 – 10 mg/L).

Orotic acid is synthesized in the liver from aspartic acid and carbamyl phosphate in two steps. It is the precursor of the simplest pyrimidine nucleotide, uridine monophosphate, and of deoxythymidine monophosphate. The corresponding triphosphates are building blocks for ribonucleic and deoxyribonucleic acids (RNA and DNA, respectively; → Nucleic Acids) [127–131].

Uridine triphosphate is a coenzyme for the transfer of uronic acids. Uridine diphosphate glucuronic acid, and thus orotic acid, play a vital role as glucuronate donor in detoxification and elimination of substances such as steroids, drugs, and bilirubin. Although orotic acid is an essential compound for many microorganisms, it does not have any vitamin-like character in mammals [129], [130], [132], [133].
Trade Name. Hepatofalk (Dr. Falk GmbH, Freiburg).

Thioctic acid [*62-46-4*], 5-(1,2-dithiolan-3-yl)valeric acid, α-lipoic acid, $C_8H_{14}O_2S_2$, M_r 208.31, *mp* 47.5 – 48 °C, forms yellowish flakes and is sparingly soluble in water, but soluble in methanol, ethanol, diethyl ether, and chlorofom.

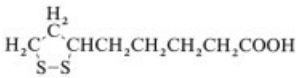

Thioctic acid is widespread in the plant and animal kingdom. It can be prepared by reacting 6,8-dichlorooctanoic acid with sodium disulfide (Na_2S_2) (for this and other methods of synthesis, see [150–152]).

Production. Industrial synthesis of thioctic acid is performed in three stages. Monomethyl or monoethyl adipate is converted into the corresponding acid chloride by reaction with thionyl chloride. Treatment with ethylene in the presence of anhydrous aluminum chloride yields 8-chloro-6-ketooctanoate. In the second stage, this ester is reduced to 8-chloro-6-hydroxyoctanoate by using thionyl chloride in pyridine. In the third stage, the chlorine atoms are replaced by sulfur atoms by treatment with sodium disulfide. Hydrolysis of the resulting dithionooctyl ester formed with alcoholic potassium hydroxide gives thioctic acid [153], [154].

Mode of Action and Pharmacology. The concentration of thioctic acid is highest in the liver. Its main biochemical role is as a coen-

zyme in the oxidative decarboxylation of α-keto acids. It is a constituent of acid reductase transacetylase, liponamide oxidoreductase, and the α-ketoglutarate dehydrogenase complex. Detoxifying properties can be expected from the five-membered disulfide ring, particularly that function as enzyme inhibitors [152], [155–157].

Thioctic acid is a growth factor for only certain protozoa and bacteria. However, in vertebrates it is formed by biosynthesis and thus a deficiency does not occur in humans. Nevertheless, it has been recommended in the treatment of hepatic coma, chronic hepatitis, and cirrhosis. Expectedly, no reproducible clinical success has been achieved. The antitoxic properties of thioctic acid have been utilized with partial success, for example, in treatment of poisoning caused by the deadly amanita (death cap) mushroom [158].

In humans, thioctic acid is excreted mainly in the urine within 24 h of a single oral or parenteral dose. In rats, the mucous membrane of the isolated duodenum rapidly absorbs thioctic acid within 45 min [159–162].

Trade Name. Thioctacid (Chemiewerk Homburg, Frankfurt)

3.2.5. Antifibrinogenic Drugs

Antifibrinogenic drugs inhibit the formation of fibrin.

Colchicine [*64-86-8*], isolated from the meadow saffron (*Colchicum autumnale* ; for structure and production, see → Alkaloids, Chap. 9.23.), was the first natural substance recognized for its antimitotic effect. This water-soluble alkaloid is quickly absorbed from the intestine and binds strongly to plasma proteins and cell proteins ; it is deacetylated in the liver. Colchicine accumulates in the body due to its slow breakdown, entrance into the enterohepatic circulation, and high protein binding. Colchicine inhibits microtubular protein assembly in fibroblasts, and thereby collagen formation. Toxic dosage produces acute renal damage and ascending paralysis, as well as chronic agranulocytosis, hair loss, and myopathy [163–165].

Over 4 years, 23 patients with liver cirrhosis received colchicine (1 mg/day, 5 days/week) and 20 a placebo [166]. Mortality decreased from 40 % with placebo to 17 % in the treatment group, but this decrease was not statistically significant. Ascites, encephalopathy, and splenomegaly improved significantly in the group receiving colchicine. Although the results of this study are encouraging, confirmation in a larger population sample is needed [166–171].

d-Penicillamine [*52-67-5*], $C_5H_{11}NO_2S$, M_r149.21, *mp* 198.5 °C, is a dimethyl derivative of the amino acid cysteine, which contains a sulfhydryl group (for production, see → Amino Acids).

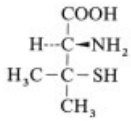

The biochemical and pharmacological effects of D-Penicillamine can be explained by three important reactions:

1) Chelation of heavy metals
2) Exchange reaction with disulfides
3) Reaction with aldehydes

Treatment of chronic – aggressive hepatitis with D-Penicillamine is based on the following mechanism of action [172], [173]:

D-Penicillamine reacts with the aldehyde groups of the collagen precursor tropocollagen to form thiazolidines. This prevents cross-linking of tropocollagen and thus the synthesis of insoluble collagen. Inhibition of collagen synthesis has been shown in animal experiments where D-penicillamine disturbs wound healing, decreases the tensile strength of the skin, and increases dilatory capacity of the lungs due to lowered elastic resistance [173], [174].

The importance of these effects is evident in diseases in which the major pathological symptom is an excessive proliferation of mesenchyma. The inhibitory influence of D-penicillamine on the inflammatory mesenchymal reaction (i.e., on cell proliferation) can be seen clearly in liver punctates [175–177].

D-Penicillamine does not dissolve or "soften" cirrhotic cicatrizations (scars). However if treatment is started at an early stage, the histologically active process may be retarded to allow healing of the residual fibrosis. This possibility assumes particular significance when other

methods of treatment are unsuccessful [178–180].
Trade Name. Metalcaptase (Knoll, Ludwigshafen; Heyl, Berlin)

Insulin and Glucagon. In animal experiments, the presence of hepatotrophic factors in portal blood has repeatedly been documented [58]. Insulin and glucagon were shown to be particularly potent stimuli of hepatocellular regeneration [181]. Administration of these hormones ameliorated fulminant murine hepatitis [182]. Stimulation of regeneration by infusion of these hormones may be a useful approach for treatment of severe alcoholic hepatitis in humans.

A randomized, double-blind controlled trial of insulin and glucagon infusion has been conducted on patients with alcoholic hepatitis [183]. Twenty five patients received 24 units of insulin and 2.4 mg glucagon over 12 h daily for 3 weeks; another 25 patients received dextrose solution as placebo. Mortality in the treatment and placebo group was not significantly different. However, in a subgroup of 34 patients with prolonged prothrombin times, fewer deaths were observed in the treatment group. In addition, total serum bilirubin and prothrombin time improved more rapidly in the insulin- and glucagon-infused group. Complications included hypoglycemia, superficial thrombophlebitis at the site of infusion, and generalized macular skin eruptions. This type of therapy, although still experimental, seems to be a promising approach for the seriously ill patients, in whom the rate of hepatocellular regeneration might be critical for survival [184].

4. References

1. W. Schmidt, B. Huber, *Med. Klin. (München)* **52** (1957) 2257.
2. C. Concilio, G. Murari, A. Petri, A. Tezza, *Arzneim. Forsch.* **12** (1962) 498.
3. F. W. Koss, R. Beisenherz, U. Engelhorn, U. Chuchra, *Arzneim. Forsch.* **12** (1962) 1026.
4. E. Kohlstaedt, GB 513 818, 1939.
5. G. Zambon, DE 1 136 703, 1962.
6. H. Brügel, O. Docker, D. P. Mertz, *Dtsch. Med. Wochenschr.* **81** (1956) 1679.
7. R. M. Adams, C. A. Vander-Werf, *J. Am. Chem. Soc.* **72** (1950) 4368.
8. R. Engelhorn, *Arzneim. Forsch.* **10** (1960) 255.
9. H. Scheffler, R. Engelhorn, GB 915 815, 1963.
10. L. Panizzi, M. L. Scarpati, *Gazz. Chim. Ital.* **95** (1965) 71; *Chem. Abstr.* **63** (1965) 5555 h.
11. L. Panizzi, M. L. Scarpati, *Nature (London)* **174** (1954) 1062.
12. H. Horn, *Muench. Med. Wochenschr.* **98** (1956) 160.
13. S. A. Uncler, DE 1 915 023, 1968.
14. J. Maillard, R. Morin, M. Benard, *Sem. Ther.* **33** (1957) 836.
15. G. A. R. Kon, J. Narayanan, *J. Chem. Soc.* 1927, 1536.
16. R. R. Burtner, J. M. Brown, *J. Am. Chem. Soc.* **73** (1951) 897; *J. Am. Soc.* **75** (1953) 2334.
17. B. Chechelska, T. Urbanski, *Rocz. Chem.* **27** (1953) 396; *Chem. Abstr.* **49** (1955) 1022 f.
18. H. Hoerster, *Dtsch. Med. Wochenschr.* **76** (1951) 277.
19. U. Schmidt, A. Lüttringhaus, H. Trefzger, *Liebigs Ann. Chem.* **631** (1960) 129.
20. B. N. Halpern, O. Gaudin, *R. Seances Soc. Biol.* **142** (1948) 779.
21. Österr. Stickstoffwerke, BE 624 848, 1963; AT 227 710, 1963.
22. H. Stormann, *Arzneim. Forsch.* **14** (1964) 266.
23. L. L. Woods, J. Sapp, *J. Org. Chem.* **27** (1962) 3703.
24. P. Defrance, *Ars Med.* **30** (1975) 455.
25. G. Bonola et al., *J. Med. Chem.* **11** (1968) 1136.
26. J. Setnikar, M. Conti, R. Guerci, *Arzneim. Forsch.* **22** (1972) 1994.
27. R. D. Soloway, B. W. Trotman, J. D. Ostrow, *Gastroenterology* **72** (1977) 167.
28. B. W. Trotman, R. D. Soloway, *Am. J. Dig. Dis.* **20** (1975) 735.
29. M. J. Coyne, G. G. Bonorris, L. I. Goldstein et al., *J. Lab. Clin. Med.* **87** (1976) 281.
30. M. J. Burnstein, R. G. Ilson, C. N. Petrunka, *Gastroenterology* **85** (1983) 801.
31. L. Den Besten, W. E. Connor, S. Bell, *Surgery* **73** (1973) 266.
32. J. E. Dotty et al., *Gastroenterology* **85** (1983) 168.
33. E. Sampliner et al., *N. Engl. J. Med.* **283** (1970) 1358.
34. R. M. Kupfer, D. P. Maudgal, T. C. Northfield, *Dig. Dis. Sci.* **27** (1982) 1025.
35. S. Erlinger et al., *Hepatology Baltimore* **4** (1984) 308.
36. Y.-H. Park, H. Igimi, M. C. Carey, *Gastroenterology* **87** (1984) 150.

37. O. Leiss et al., *Gastroenterology* **87** (1984) 144.
38. M. Podda, "Die Kombinationsbehandlung von Gallenblasensteinen mit Ursodesoxycholsäure und Chenodesoxycholsäure," in: *Neue wissenschaftliche Erkenntnisse der Gallensteinauflösung* , Falk Foundation e.v., Freiburg 1987, p. 21.
39. C. Wolpers, *Leber Magen Darm* **6** (1976) 43.
40. M. Leuschner et al., *Hepato Gastroenterology* **31** (1984) 140.
41. R. H. Dowling et al., in G. Paumgartner, A. Stiehl, W. Gerok (eds.): *Bile Acids and Cholesterol in Health and Disease,* "Efficacy of Bile Acid Treatment in Dissolving Gallstones: Collateral Benefits of CDCA and UDCA Therapy; Gallstone Recurrence and Postdissolution Management," MTP Press, Lancaster, England 1983, p. 345.
42. L. Barbara et al., "A Comparative Study of Chenodeoxycholic Acid and Ursodeoxycholic Acid as Cholesterol Gallstone Dissolving Agents," in: G. Paumgartner, A. Stiehl, W. Gerok (eds.): *Bile Acids and Cholesterol in Health and Disease* , MTP Press, Lancaster 1983, p. 367.
43. U. Leuschner et al., in L. Okolicsanyi, G. Czomos (eds.): *Assessment and Management of Hepato Biliary Diseases,* "Medical treatment of Cholesterol Gallstones: Present and Future," Springer Verlag, Berlin – Heidelberg – New York 1987, p. 107.
44. T. Northfield: *Patientenauswahl für die medikamentöse Gallensteinauflösung mit Gallensäuren,* Falk Foundation e.V., Freiburg 1987, p. 5.
45. M. J. Allen et al., *N. Engl. J. Med* **312** (1985) 217.
46. D. C. Ruppin, R. H. Dowling, *Lancet* 1982, vol. I, 181.
47. U. Leuschner et al., *Am. J. Gastroenterol.* **83** (1988) 169.
48. U. Leuschner et al., *Dig. Dis. Sci.* (1998) in press.
49. T. Sauerbruch et al., *N. Engl. J. Med.* **314** (1986) 385.
50. E. Seifert, K. Gail, J. Weismüller, *Dtsch. Med. Wochenschr.* **107** (1982) 610.
51. J. F. Riemann et al., *Endoscopy* **15** (1983) 165.
52. U. Leuschner, *Clin. Gastroenterol.* **15** (1986) 333.
53. U. Leuschner et al., *Digestion* **34** (1986) 36.
54. U. Leuschner, H. Baumgärtel, U. Wosiewitz, *Gastroenterology* **88** (1985) 1674.
55. U. Leuschner, H. Baumgärtel, in K. Okuda, F. Nakayama, J. Wong (eds.): *Intrahepatic calculi,* "Chemical Dissolution of Common Bile Duct Stones," Alan R. Liss, New York 1984, p. 193.
56. P. Stefanini et al., *Surg. Gynecol. Obstet.* **139** (1974) 734.
57. K. W. Warren, E. G. C. Tan, in L. Schiff (ed.): *Diseases of the liver,* "Diseases of the Gallbladder," 4th ed., Lippincott, Philadelphia, Toronto 1975, p. 1278.
58. T. E. Starz, J. Terblanche, *Prog. Liver Dis.* **6** (1979) 135 – 152.
59. G. Csomos, *Medica* **12** (1984) 134 – 142.
60. G. Csomos, *Orvosképés* **62** (1987) 158.
61. G. Csomos, H. Thaler: *Clinical Hepatology,* Springer Verlag, Berlin – Heidelberg – New York 1983.
62. V. Fintelmann: *Intuitive Medizin,* Hippokrates-Verlag, Stuttgart 1986.
63. R. Kirsch, J. Kruskal, G. Csomos, J. Terblanche: *Liver Update* 2/1985 Bailliere Tindall, 1985.
64. J. Feher, G. Csomos, A. Vereckei, *Free Radical Reactions in Medicine,* Springer Verlag, Berlin – Heidelberg – New York 1987.
65. L. Okolicsanyi, G. Csomos, G. Crepaldi: *Assessment and Management of Hepatobiliary Disease,* Springer Verlag, Heidelberg 1987.
66. I. Arias, H. Popper, D. Schachter, D. Schafritz: *The Liver,* Raven Press, New York, 1982.
67. H. Thaler: *Leberkrankheiten,* Springer-Verlag, Berlin – Heidelberg – New York 1987.
68. W. Gerok: *Hepatologie,* Urban & Schwarzenberg, Stuttgart 1987.
69. G. J. M. Alexander et al., *J. Med. Virol.* **21** (1987) 81 – 87.
70. J. Dienstag, K. J. Isselbacher, *Arch. Intern. Med.* **141** (1981) 1419 – 1423.
71. J. H. Hoofnagel: *Therapy of Chronic Hepatitis B,* Academic Press, London – New York 1985.
72. H. Popper, F. Schaffner, *Progress in Liver Diseases,* Grune & Stratton Press, New York 1985.
73. C. J. Smith et al., *Am. J. Med.* **73** (1982) 267 – 270.
74. S. Sherlock: *Diseases of the Liver and Biliary System,* Blackwell-Scientific Publ., London 1985.
75. R. Wright, G. H. Millvard-Sadler, H. G. M. N. Alberti, S. Karran, (eds.): *Liver and Biliary Disease,* Saunders, Philadelphia 1985.
76. T. Hüttenroth, H. Poralla, K. H. Meyer zum Büschenfelde, *Internist* **26** (1985) 607 – 613.
77. S. W. Schalm, H. R. Buuren, R. A. Heytink, R. A. de Mann, *Lancet* 1985, vol. II, 358 – 360.

78. R. D. Solway, W. H. J. Summerskill, A. H. Baggenstoss, *Gastroenterology* **63** (1972) 820 – 833.
79. M. Rizetto, G. Verme, *Internist* **26** (1985) 628 – 633.
80. S. Jain, H. C. Thomas, J. S. Oxford, S. Sherlock, *J. Antimicrob. Chemother.* **4** (1978) 367 – 373.
81. R. Müller, *Internist* **25** (1984) 728 – 734.
82. C. Trepo et al., *Gastroenterology Clin. Biol.* **5** (1981) 93 A.
83. C. J. Smith, T. C. Merigan, *Prog. Liver Dis.* **7** (1982) 481 – 494.
84. G. Strohmeyer, H.-J. Lübke, *Dtsch. Med. Wochenschr.* **110** (1985) 104 – 105.
85. A. Gajdos, M. Gajdos-Török, R. Horn, *Biochem. Pharmacol.* **21** (1972) 595.
86. T. Omura, R. Sato, *J. Biol. Chem.* **239** (1964) 2370.
87. J. Sipos et al., *Int. J. Tissue Reakt.* **2** (1980) 167 – 171.
88. G. Csomos, G. Pagliaro, *Hepatology (Baltimore)* **6** (1972) 22 (41615).
89. M. Jounes, C. P. Sixgers, *Arch. Toxicol.* **45**(1980) 61.
90. A. L. Blum et al. *Lancet* 1977, vol. II, 1153 – 1155.
91. F. Di Nola, *Lancet* 1980, vol. II, 1379 – 1380.
92. D. Eichelberg, W. Schmutzler, *Immun. Infekt* **11** (1983) 109 – 122.
93. K. A. Neftel, P. Diem, H. Gerber, A. L. De Weck, P. Stucki, *Schweiz. Med. Wochenschr.* **110** (1980) 380 – 383.
94. W. H. Down, L. F. Chasseaud, R. K. Grundy, *Arzneim. Forsch.* **24** (1974) no. 12, 1986 – 1988.
95. H. Wagner, L. Hörhammer, R. Münster, *Arzneim. Forsch.* **18** (1968) no. 6, 688 – 696.
96. G. Csomos, *Hepatology* **6** (1979) 20 (41614).
97. G. Vogel et al., *Arzneim. Forsch. (Drug-Res.)* **25** (1975) 82, 179.
98. H. W. Leber, S. Knauff, *Arzneim. Forsch.* **26** (1976) no. 8, 1603 – 1605.
99. L. Okolicsanyi et al., in F. de Ritis, G. Csomos, R. Braatz (eds.): *Der toxisch-metabolische Leberschaden,* Hans. Verl.-Kontor, Lübeck 1981, p. 83.
100. J. D. Berenger, D. Caarasco, *Muench. Med. Wochenschr.* **119** (1977) 240 (span. ed).
101. J. Sonnenbichler, A. Pohl, *Hoppe-Seyler's Z. Physiol. Chem.* **361** (1980) 1757 – 1761.
102. J. Sonnenbichler et al., *Hepatology (Baltimore)* **6** (1980) 362.
103. G. Martines, G. C. Mingrino, G. Cagnetta, V. Copponi, *Clin. Ther. (Rome)* **92** (1980) 33 – 58.
104. H. M. Rauen, H. Schriewer, *Arzneim. Forsch.* **24** (1971) 1201.
105. J. Sonnenbichler, I. Zetl, *Hoppe-Seyler's Z. Physiol. Chem.* **365** (1984) 555 – 566.
106. G. Csomos, K. Hruby, H. Thaler, *Quarterly Bulletin,* **vol. 2 (3),** 1985.
107. J. Sonnenbichler, J. Mattersberger, G. Hanser, *Hoppe- Seyler's Z. Physiol. Chem.* **361** (1980) 1751 – 1756.
108. H. Schriewer, R. Badde, G. Roth, H. M. Rauen, *Arzneim. Forsch. (Drug Res.)* **23** (1973) no. 1 a, 157 – 158.
109. B. Tuchweber, I. M. Yousef, *Hepatology (Baltimore)* **6** (1980) 360.
110. R. Seeger, *Arzneim. Forsch.* **28** (1978) no. 6, 1013.
111. V. Schreiber, F. X. Wohlzogen, *Int. J. Clin. Pharmacol. Biopharm.* **16** (1978) 533 – 535.
112. V. Fintelmann, *Med. Klin. (Munich)* **68** (1973) 809 – 815.
113. V. Fintelmann, A. Albert, *Therapiewoche* **30** (1980) 5589 – 5591.
114. H. A. Salmi, S. Sarna, *Scand. J. Gastroenterol.* **17** (1982) 517.
115. E. Kiesewetter, I. Leodolter, H. Thaler, *Leber Magen Darm* **7** (1977) 318 – 323.
116. B. Magliulo, B. Gagliardi, G. P. Fiori, *Med. Klin. (Munich)* **73** (1978) 1060 – 1065.
117. S. Cavalieri, *Gazz. Med. Ital.* **133** (1974) 628 – 653.
118. P. Saba et al., *IRCS Med. Sci. Libr. Compend.* **8** (1980) 46.
119. J. Lahtinen, H. Hendolin, T. Tuppurainen, *XI. International Congress for Gastroenterology AECGE Hamburg 1980,* Georg Thieme Verlag, Hansisches Verlagskontor Lübeck 1981 pp. 79 – 82.
120. P. Ferenci et al. in L. Okolicsanyi, G. Csomos (eds.): *Assessment and Management of Hepato Biliary Diseases,* Springer Verlag, Berlin – Heidelberg – New York 1987.
121. H. Thaler, *Wien. Klin. Wochenschr.* **95** (1983) 224.
122. K. Hruby, M. Fuhrmann, G. Csomos, H. Thaler, *Wien. Klin. Wochenschr.* **95** (1983) 225 – 231.
123. K. Hruby et al. (eds.): *Diagnose, Verlaufskontrolle und Therapie schwer exogener Vergiftungen* Schattauer-Verlag, Stuttgart – New York 1984 p. 267.
124. E. E. Belenky, J. A. Runichin, T. A. Tunizkaja, *Farmakol. Toksikol (Moscow)* **34** (1971) 472.
125. G. F. Combs, G. H. Arscott, H. L. Jones, *Poult. Sci.* **33** (1954) 71.

126. K. Davis, *Int. J. Vitam. Nutr. Res.* **41** (1971) 360.
127. E. Hoffmann: *Dynamische Biochemie* Part II, Springer Verlag, Berlin 1966.
128. E. Jassmann, L. Lücke, I. Hauer, *Pharmazie* **34** (1979) 305.
129. E. Morgenstern, *Zbl. Pharmaz. Pharmakotherap. Laboratoriumsdiagnostik* **122** (1983) 188.
130. E. Nemesanszky, I. Szellenyi, *Arzneim. Forsch.* **21** (1971) 685.
131. F. A. Kheinonen, C. Makeera, *Kardiologiya* **10** (1970) 31.
132. H. Matthies, *III. Neurobiol. Sympos.,* Magdeburg, May 5 – 7, 1971.
133. F. Z. Meerson, E. J. Chazov, K. N. Bykowskaja, *Experientia* **23** (1967) 546.
134. T. Ott, *Pharmazie* **34** (1979) 306.
135. D. Müting, *Ther. Ggw.* **111** (1972) no. 11, 1660.
136. B. Stoll, U. Scholz, W. Haase, *Med. Welt* **27** (1976) 707 – 712.
137. H. Wallnöfer, M. Hanusch, *Wien. Med. Wochenschr.* **117** (1967) no. 17/18, 488 – 492.
138. F. Knüchel, *Med. Welt.* **30** (1979) 411 - 416.
139. H. D. Kuntz et al., *Med. Welt* **29** (1978) 452.
140. R. Airo et al., *Med. Heute (Osaka)* **10** (1961) no. 1, 16 – 23.
141. R. Arendt et al., *Zschr. Inn. Med* 25. Jg., H. 13 (1970), 587 – 589.
142. E. Kautzsch, *Med. Klin. (Munich)* **60** (1965) no. 35, 1401 – 1406.
143. N. Reuter, *Arzneim. Forsch.* **23** (1973) 1062.
144. S. Schulzeck, S. Schmidt, W. Pohle, N. Popov, *Z. Med. Labortech.* **12** (1971) 193.
145. K. H. Simon: *Magnesium Physiologie, Pharmakologie,* Klinik Stuttgart 1967.
146. P. Nuhn: *Chemie der Naturstoffe,* Springer Verlag, Berlin 1981.
147. I. Szelenyi, J. Pucsok, E. Nemesanszky, J. Sos, *Arzneim. Forsch.* **21,** (1971) 777.
148. E. Szirmai, J. Klosa, *Arzneim. Forsch.* **27,** (1977) 2414.
149. E. Wenzel, *Wien. Med. Wochenschr.* **121** (1971) 226.
150. P. Boni et al., *Clin. Ther. (Rome)* **9** (1955) 129.
151. H. Griesebach, *Angew. Chem.* **68** (1956) no. 17/18, 554.
152. G. Eisfeld et al., *ärztl. Forsch.* **12** (1958) no. 10, 490.
153. P. Prager et al., *Acta Physiol. Hung.* **64** (1984) no. 3 – 4, 489.
154. H. Taegtmeyer, J. M. Passmore, *Lancet* **19** (1985) 139.
155. H. Petzold, *Acta Biol. Med. Ger.* **3** (1959) 1.
156. H. Thölen et al., *Helv. Med. Acta* **6** (1966) 492.
157. L. J. Reed, R. M. Oliver, *Brookhaven Symp. Biol.* **21** (1968) no. 2, 397.
158. C. E. Becker et al., *West. J. Med.* **125** (1976) no. 2, 1831.
159. A. W. Marshall et al., *Gut* **23** (1982) 1088.
160. R. Plotzker et al., *Am. J. Med. Sci.* **283** (1982) no. 2, 79.
161. H. Thaler, *Leber Magen Darm* **12** (1982) no. 4, 171.
162. W. Schimmelpfennig et al., *Dtsch. Gesundheitswes.* **38** (1983) no. 18, 690.
163. E. Barrett-Connor, K.-T. Khaw, S. S. C. Yen, *N. Engl. J. Med.* **315** (1986) 1519 – 1524.
164. M. Rojkind, M. Uribe, D. Kershenobich, *Lancet 1973,* **vol. I,** 38 – 39.
165. E. W. Taylor, *J. Cell. Biol.* **25** (1965) 145 – 160.
166. D. Kershenobich, F. Vargas, G. Garcia-Tsao, R. P. Tamayo et al., *N. Engl. J. Med.* **318** (1988) 1709 – 1713.
167. P. Gines et al., *Hepatology (Baltimore)* **7** (1987) 122 – 128.
168. M. M. Kaplan et al., *N. Engl. J. Med.* **315** (1986) 1448 – 1454.
169. T. W. Warnes et al., *J. Hepatol.* **5** (1987) 1 – 7.
170. R. M. Naidus, R. Rodvien, C. H. Mielke, Jr. *Arch. Intern. Med.* **137** (1977) 394 – 396.
171. Y. H. Caplan, K. G. Orloff, B. C. Thompson, *J. Anal. Toxicol.* **4** (1980) 153 – 155.
172. M. Alexander, *Ges. Inn. Med.* **77** (1971) 1303 – 1306.
173. H. Wesch, H. Przuutek, D. Feist, *Dtsch. Med. Wochenschr.* **105** (1980) 483 – 488.
174. S. Sherlock, *Am. J. Med.* **65** (1978) 217 – 219.
175. W. Endres, *Klin. Wochenschr.* **59** (1981) 535 – 537.
176. W. Siegmuth, in K. Chlud (ed.): *D-Penicillamin 1970 – 1980, Konsequenzen für die Praxis* Aesopus Verlag, Zug 1981.
177. M. Franke, *Therapiewoche* **31** (1981) 2131 – 2132.
178. O. Epstein et al., *Lancet* 1981, vol. I, 1275 – 1277.
179. D. Mathoff, E. Alpert, R. Resnick, M. Kaplan, *N. Engl. J. Med* **306** (1982) 318 – 326.
180. M. Bassendine, A. Maklon, R. Mulcaly, O. James, *Gut* **23** (1982) A 909.
181. N. Bucher, M. Swaffield, *Proc. Nat. Acad. Sci. USA* **72** (1975) 1157 – 1160.
182. M. Farivar, J. Wands, K. Isselbacher, N. Bucker, *N. Engl. J. Med.* **295** (1976) 1517 – 1519.
183. A. Baker et al., *Gastroenterology* **80** (1981) 1410 – 1414.
184. J. Feher et al., *Hepatology (Baltimore)* **5** (1987) 224.

Gallium and Gallium Compounds

Jörg Friedrich Greber, INGAL International Gallium GmbH, Schwandorf, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Occurrence and Raw Materials** 2
5. **Production** 2
6. **Storage and Transportation** 5
7. **Uses** 5
8. **Economic Aspects** 5
9. **Toxicology and Industrial Medicine** . . 6
10. **References** 6

1. Introduction

Gallium [*7440-55-3*] is an element of group 13. Its existence was predicted by Mendeleev, and it was discoveredby Lecoq de Boisbaudran in 1875. The metal was mainly of academic interest until 1970, when it was found that compounds of gallium with group 15 elements have semiconducting properties (→ Semiconductors), and so began the extensive industrial use of the element. Gallium compounds are of particular value in optoelectronics, especially for light-emitting diodes.

2. Physical Properties

Gallium is a silvery-white metal, *mp* 29.78 °C, *bp* 2403 °C. The naturally occurring element contains two stable isotopes: ^{69}Ga (60.4 %) and ^{71}Ga (39.6 %). The density of the liquid is 6.095 g/cm^3 at 29.8 °C, and the density of the solid is 5.904 g/cm^3 at 24.6 °C. There is a volume expansion of 3.2 % on solidification.

Gallium forms orthorhombic crystals with the lattice constants $a = 0.4523$ nm, $b = 0.45198$ nm, and $c = 0.76602$ nm, and the physical properties are consequently strongly anisotropic. The thermal expansion coefficients are 1.65×10^{-5} K^{-1} along the a axis, 1.13×10^{-5} K^{-1} along b, and 3.1×10^{-5} K^{-1} along c. The electrical properties are equally anisotropic, but it is not possible to quote reliable figures owing to the great influence of purity and temperature on the electrical conductivity.

3. Chemical Properties [1–4]

Metallic gallium dissolves only slowly in dilute mineral acids, but rapidly in aqua regia and concentrated aqueous sodium hydroxide. It also dissolves slowly in solutions of hydrogen halides in ether.

In its compounds, the valency of gallium is usually 3. The monovalent state is unstable, although the monovalent gallium compounds Ga_2O and GaCl can be isolated. An oxidation state of +1 has not yet been detected in aqueous solution, but some reactions of gallium indicate that it exists.

The oxygen compounds of gallium resemble those of aluminum in that there are high and low temperature forms of Ga_2O_3 and two hydroxides, $Ga(OH)_3$ and $GaO \cdot OH$. Gallium halides have covalent character and therefore have good solubility in many nonpolar solvents in which they exist in dimeric form.

With elements of group 15, gallium forms binary compounds, of which GaP [*12063-98-8*] and GaAs [*1303-00-7*] areof great industrial importance. They may be prepared by direct combinationof the elements at high temperature.

In aqueous solution, gallium forms octahedrally coordinated aquo ions $[Ga(H_2O)_6]^{3+}$.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_163

There are many salts of oxoacids, such as the sulfate, nitrate, and perchlorate, and also hydrated halides.

In aqueous solution, the hexaaquo complex has an acid reaction owing to hydrolysis. Gallium salts of weak acids cannot exist in the presence of water for the same reason. Of the numerous anionic and cationic complexes, the most industrially important are the halide complexes, the $[GaX_4]^-$ ions, whose oxonium salts have good solubility in several organic solvents. This property is utilized in the extraction and purification of the metal. Other gallium complexes of industrial importance are the octahedrally coordinated chelates with β-diketones and 8-hydroxyquinolines, these also being soluble in organic solvents. Alcoholates and organic derivatives may be prepared similarly to the analogous aluminum compounds.

Gallium forms a series of alloys with other metals, sometimes even at low temperatures. Some of these, containing up to 3 % gallium, are useful in dentistry.

4. Occurrence and Raw Materials

With its natural abundance of 16 ppm, gallium is one of the rarer elements in the earth's crust. Gallium-containing minerals, the best known of which is a variety of germanite, do not have economic significance. Of the known locations, only the Apex mine in Utah has been commercially mined.

The greatest part of the world production of gallium is as a byproduct of the production of aluminum oxide. The gallium concentration in bauxite ranges between 0.003 and 0.008 %, depending on the location. The high contents occur in tropical bauxites such as Surinam bauxite, which contains 0.008 % Ga, the highest average content known. The world supply of recoverable gallium in bauxite is estimated at 1.6×10^6 t based on bauxite reserves and their gallium contents [5].

Because of the low concentration of gallium in bauxite, it is not economic to mine bauxite to extract gallium. The amount of gallium available from bauxite is then determined by the amount of aluminum oxide that is produced. For the year 1986, complete utilization of the gallium from aluminum oxide production would have corresponded to a theoretical production of ca. 2000 t of gallium. The gallium contained in zinc ore (sphalerite) has lost economic importance due to the changeover to hydrometallurgical methods of zinc extraction. The reserves of gallium in zinc ores worldwide has been estimated to be 6500 t [5].

The largest reserves of gallium are contained in phosphate ores and in coal of various kinds. When phosphorus is produced electrothermally from phosphate, the gallium is concentrated in the flue dust and may be recovered from it. Likewise, gallium may be obtained from the fly ash of coal, another material in which gallium is concentrated. Neither source is exploited at present because the gallium concentration is only 0.01 – 0.1 wt %. It is not economic to treat these materials merely to extract the gallium. The total gallium reserves in phosphates and coals have been estimated at several million tonnes. They thus considerably exceed those in bauxite [5].

5. Production

The most important process for gallium production is the extraction of the metal from the circulating liquors in the Bayer process for aluminum oxide manufacture (→ Aluminum Oxide, Chap. 3.). These contain 70 – 150 mg/L of gallium, depending on the bauxite and on the concentration of the caustic liquor. The gallium content of the liquor is also determined by the dissolution process used. Bauxite which is high in boehmite ($AlO \cdot OH$) requires high temperatures for the dissolution process, leading to more complete release of the gallium and higher concentrations in the circulating liquors.

The gallium extraction plant is usually located within or very near to the aluminum oxide works. Part of the liquor stream is diverted for gallium extraction and then recycled. In some instances the liquor undergoes a preliminary purification process to remove certain heavy metals.

Three types of process are used to extract the gallium from the circulating liquor: (1) fractional precipitation, (2) electrolytic processes, and (3) extraction with chelating agents.

Fractional Precipitation. The basis of this process is that when the gallium-containing sodium aluminate solution is treated with carbon dioxide the first product precipitated is pure

aluminum hydroxide. This increases the gallium content of the liquor. Further treatment then causes the gallium to separate along with more aluminum hydroxide. After this, fractional precipitation gives an aluminum hydroxide richer in gallium, and this is followed by one or more further fractional precipitations. The gallium-enriched aluminum hydroxide is dissolved in aqueous sodium hydroxide to give a solution of sodium aluminate and sodium gallate, from which gallium is obtained by electrolysis.

The precipitation process is used only to a small extent because of the expense of removing the byproduct sodium carbonate unless thermal decomposition is used in the aluminum oxide works. Also, on account of the many precipitation stages involved, the process is labor intensive and can therefore be used only when energy and labor costs are low. The People's Republic of China produces gallium at a rate of 5 t/a by this process.

Electrolytic Processes. The electrolytic process depends on the fact that mercury forms an amalgam with gallium from which the gallium may be extracted with a caustic soda solution. The gallium amalgam can be produced either directly (Alusuisse process) or indirectly (VAW/INGAL process).

In the Alusuisse process, the gallium is deposited directly onto the mercury cathode, using the appropriate conditions of temperature, agitation, and current density, and an amalgam is formed. In the VAW process, sodium amalgam is used to reduce the gallium present in the Bayer liquor. This gives an amalgam from which the gallium is removed by treatment with caustic alkali, and the mercury is then used again to produce sodium amalgam, i.e., the mercury circulates in a closed system.

Reduction processes have in recent years become less attractive because they require special conditions to be commercially viable. In particular, the aluminate solutions must contain over 150 g/L sodium and be very low in vanadium, tungsten, molybdenum, and iron so that the electrolytic current efficiency reaches economically acceptable values. However, such highly concentrated liquors are not produced in modern aluminum oxide plants, and therefore the electrolytic process can no longer be used in new installations. The environmental problems associated with mercury further discourage the use of electrolytic processes. These processes are, however, in operation in Hungary, Czechoslovakia, the German Democratic Republic, and the Federal Republic of Germany. About 15 – 20 t/a of gallium is produced.

Extraction with Chelating Agents. The discovery was made in 1974 that hydroxyquinolines could be used to extract gallium from sodium aluminate liquors (the Rhône-Poulenc process, Fig. 1) [6], these substances having earlier been developed for the hydrometallurgical extraction of copper. The aluminate liquor is brought into contact with a solution of Kelex 100 [*29171-27-5*] in kerosene.Gallium is thereby extracted from the liquor along with aluminum and sodium.Most of the elements extracted with the gallium may be stripped from the extraction medium by treatment with dilute acid. The gallium is then extracted with concentrated acid; either hydrochloric or sulfuric acid may be used. Gallium concentrations of 0.1 – 10 g/L may be reached in the acid extract, depending on the method used. The gallium to aluminum concentration ratio in the extract can reach values as high as 1, which corresponds to enrichment factors of up to 1000 in the primary extraction stage.

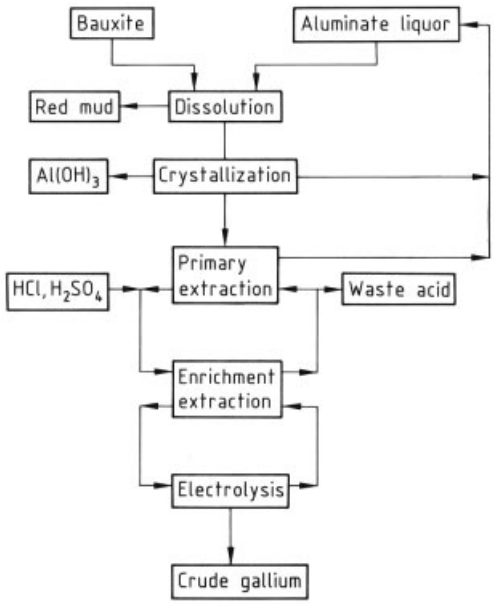

Figure 1. Flow diagram showing gallium extraction by the Rhône-Poulenc and Sumitomo processes

Further concentration of the gallium is achieved by anion- or cation-exchange treatment

of the acid extract from the first stage. The metal can be isolated from this concentrate by direct electrolysis.

Although the process solutions are recirculated to a large extent, considerable quantities of waste acid are generated. This waste acid is contaminated with other extracted metals and residues from the extraction media. An important economic factor in the extraction process is therefore the possibility to dispose of these solutions cheaply or to make some use of them.

In the Sumitomo process (Fig. 1) the first stage is extraction with the ion-exchange resin Duolite CS-346 (Diamond Shamrock) [7]. This resin contains amidoxime groups, which form chelate bonds to gallium. Unlike Kelex 100, this resin has the drawback that any vanadium extracted from the aluminate liquors blocks the ion exchange sites. Vanadium must therefore be removed from the circulating liquor of the aluminum oxide plant before gallium extraction. The rest of the Sumitomo process corresponds to the Rhône-Poulenc process.

Gallium is produced by extraction processes in France and Japan.

Purification. The gallium from extraction processes has a purity of 99 – 99.9 wt %. For most applications, however, much higher purities are necessary, reaching 99.999999 % (8N). Various procedures are used to remove the impurities. Volatile metals, such as mercury and zinc, are distilled off under vacuum. Further purification is effected by washing with aqueous acids and alkalies. The purest gallium is obtained by fractional crystallization, zone melting, or single crystal growth. Other methods of purification include the extraction of gallium chloride from acid solution and fractional distillation of liquid gallium compounds.

Environmental Protection. No harmful effects of gallium or its compounds have been reported. There are therefore no regulations concerning gallium concentrations in exhaust gases from gallium manufacturing plants. For the materials used in gallium manufacture, such as mercury or 8-hydroxyquinoline, the relevant legal requirements must be observed.

Quality. The commercially available materials include a crude grade and various pure grades of gallium. The crude metal contains 99 – 99.9 % gallium. The material for semiconductor manufacture is used in purities from 99.9999 % (6N) to 99.999999 % (8N).

Analysis. Gallium is determined in raw materials and partially enriched intermediate products either by complexometric titration (high concentrations) or atomic absorption (low concentrations).

For the analysis of the purest grades of gallium, it is necessary to have methods of determining impurities at parts per million to parts per billion levels. The most important technique is spark source or glow discharge mass spectrometry, which enables heavy metals to be determined with sufficient accuracy. Lighter elements can be determined by activation analysis or by special mass spectrometric techniques.

A general assessment of gallium purity may be made from its residual resistance ratio. This is determined from the electrical conductivity of a gallium single crystal at room temperature and at liquid helium temperature. This ratio is very sensitive to the presence of impurities but is nonspecific. Typical values for high purity gallium are better than 55 000.

6. Storage and Transportation

The low melting point of gallium and the high purity requirements mean that care must be taken to prevent gallium from melting or being contaminated by the packing materials during storage and transportation. The metal is usually sold in lumps of 10 g – 2 kg, individually packed and kept cold in plastic containers.

For short distances, gallium may be carried by truck or rail, provided that the temperature is not allowed to exceed 20 °C. For carriage over long distances, the method chosen is usually air transport, in which uninterrupted cooling is to be maintained. Regulations governing packaging for air transport have been laid down by IATA owing to the ability of gallium to form alloys with aluminum at normal temperatures, these alloys having very low mechanical strength. Gallium must be contained inside seven layers of packing material and in warm weather the container must be cooled. In this way, the danger of an accident in which the structural metal of

Table 1. Gallium production in kilograms of gallium for 1978, 1984, and 1988.

Region and country	Year		
	1978	1984	1988
Europe			
Czechoslovakia		2 000	2 000
France	2 000	10 000	8 000
Federal Republic of Germany	4 000	8 000	8 000
Hungary	1 000	3 000	2 000
Others	1 000	1 200	2 000
Asia			
Japan	200	13 000	12 000
China	600	5 000	5 000
United States	8 000	3 000	
Others	1 000	1 000	1 000
Total	*17 800*	*46 200*	*40 000*

the aircraft is attacked by liquid gallium can be reduced. In addition, there is a limit imposed on the amount of gallium which may be carried in an aircraft.

7. Uses

The main use of gallium is as a raw material in the manufacture of semiconductors which are formed by the combination of gallium with elements of group 15 (especially P and As). They have certain advantages over other semiconductor materials: (1) faster operation with lower power consumption, (2) better resistance to radiation and, most importantly, (3) they may be used to convert electrical into optical signals.

The largest proportion of the gallium produced worldwide goes to the manufacture of the arsenide or phosphide for the production of lightemitting diodes (LEDs, → Display Technology; → Electroluminescent Materials and Devices). The manufacture of integrated circuits using gallium arsenide is still at the development stage. A small proportion of the gallium produced is converted into the oxide for manufacture of gallium – gadolinium garnets (GGG) [8]. These are used to produce magnetic bubble memories.

8. Economic Aspects

In 1986, 50 t gallium was used worldwide (Fig. 2), mainly in the manufacture of lightemitting diodes and other optoelectronic components.

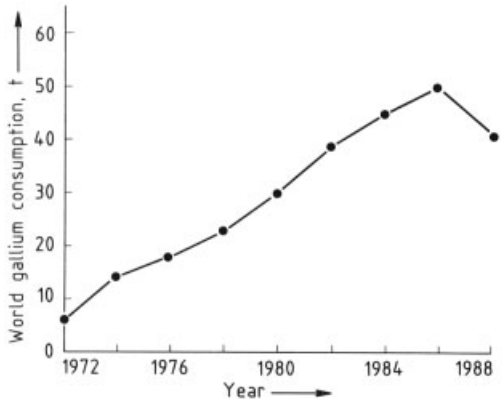

Figure 2. World gallium consumption

About 65 % of the market is in Japan, 20 % in the United States, and 15 % in Europe. The price has fallen continuously since 1970 and in 1986 reached $ 450 or 77 000 yen per kg for 6N gallium. The total value of the gallium used worldwide in 1986 was therefore $ 22×10^6 (3.85×10^9 yen). Gallium is extracted in Western Europe, Japan, and several countries of the Eastern Bloc. Production levels in the Soviet Union, the German Democratic Republic, Hungary, and Czechoslovakia are not known exactly as only a small part of the gallium produced there is exported.

The following companies produce gallium: INGAL, (Federal Republic of Germany), Rhône-Poulenc (France), Sumitomo (Japan), and Dowa Mining (Japan). Musto Mining (USA), Eagle Picher (USA), Alcan (Canada), and Cominco (Canada) have ceased production.

In addition to these producers of primary gallium, a number of companies recover pure gallium from gallium scrap.

World gallium production increased from 17 800 kg in 1978 to 46 200 kg in 1984, but decreased to 40 000 kg in 1988 (Table 1).

9. Toxicology and Industrial Medicine

Gallium is only slightly toxic. No industrial injury caused by the metal or its compounds during gallium production and processing has been reported. An investigation in 1984 and 1985 by the Chemical Trade Association did not reveal any evidence of damage to health caused by gallium.

10. References

1. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 4th ed., Wiley, New York 1980, pp. 326 – 351.
2. *Gmelin,* **8,** System no. 36.
3. *Gmelin,* Organogallium Compounds, Part 1 (1987).
4. I. A. Sheka, I. S. Clans, T. T. Mityureva: *The Chemistry of Gallium,* Elsevier, Amsterdam 1966.
5. F. E. Katrak, J. C. Agarwal, *J. Met.* **33** (1981) no. 9, 33.
6. Rhône-Poulenc, FR 2307 047, 1974 (J. Helgorsky, A. Leveque).
7. Sumitomo Chem. Comp., EP 0076 404, 1983.
8. Vereinigte Aluminiumwerke AG, DE 25 17 292, 1975, (H. Pfundt, M. Fuchs, P. Voiß).

Gas Production

Heinz Hiller, Lurgi, Frankfurt am Main, Germany (Chap. 1)

Rainer Reimert, (formerly Lurgi) Engler-Bunte-Institut der Universität Karlsruhe, Karlsruhe, Germany (Chap. 1, 4 and 7.3)

Friedemann Marschner, Lurgi, Frankfurt am Main, Germany (Chap. 2)

Hans-Joachim Renner, Lurgi, Frankfurt am Main, Germany (Chap. 2)

Walter Boll, Lurgi, Frankfurt am Main, Germany (Chap. 2 and 5.3)

Emil Supp, Lurgi, Frankfurt am Main, Germany (Chap. 3, 5.1 and 5.2)

Miron Brejc, Lurgi, Frankfurt am Main, Germany (Chap. 3)

Waldemar Liebner, Lurgi, Frankfurt am Main, Germany (Chap. 3)

Georg Schaub, (formerly Lurgi) Engler-Bunte-Institut der Universität Karlsruhe, Karlsruhe, Germany (Chap. 4)

Gerhard Hochgesand, Lurgi, Frankfurt am Main, Germany (Chap. 5 and 6.3)

Christopher Higman, Lurgi, Frankfurt am Main, Germany (Chap. 5.1, 5.2 and 7.2)

Peter Kalteier, Lurgi, Frankfurt am Main, Germany (Chap. 5.1 and 5.2)

Wolf-Dieter Müller, Lurgi, Frankfurt am Main, Germany (Chap. 5.3)

Manfred Kriebel, Lurgi, Frankfurt am Main, Germany (Chap. 5.4)

Holger Schlichting, Lurgi, Frankfurt am Main, Germany (Chap. 5.4)

Heiner Tanz, (formerly Lurgi, Frankfurt), IVTP, Dreieich, Germany (Chap. 5.5, 5.6, 5.7 and 5.8)

Hans-Martin Stönner, Lurgi, Frankfurt am Main, Germany (Chap. 1.5, 6.1 and 6.2)

Helmut Klein, Lurgi, Frankfurt am Main, Germany (Chap. 6.2)

Wolfgang Hilsebein, Lurgi, Frankfurt am Main, Germany (Chap. 7.1)

Veronika Gronemann, Lurgi, Frankfurt am Main, Germany (Chap. 7.1)

Uwe Zwiefelhofer, Lurgi, Frankfurt am Main, Germany (Chap. 7.2)

Johannes Albrecht, Lurgi, Frankfurt am Main, Germany (Chap. 7.3)

Christopher J. Cowper, British Gas, London, United Kingdom (Chap. 8)

Hans Erhard Driesen, Engler-Bunte-Institut der Universität Karlsruhe, Karlsruhe, Germany (Chap. 8)

Related Articles → Coal Liquefaction and → Coal Pyrolysis are separate keywords.

1. Introduction 3
1.1. Types of Gases; General Overview of Production Methods and Characteristics 3
1.1.1. Water Gas and Producer Gas 3
1.1.2. Synthesis Gas and Reduction Gas . 4
1.1.3. Town Gas and Medium-Btu Gas . . 7
1.1.4. Biogas and Landfill Gas 9
1.1.5. Rich Gas and Substitute Natural Gas (SNG) 9
1.2. Raw Materials for Gasification . . 9
1.3. Physicochemical Basis for Gas Production 12
1.4. Characteristics of the Basic Processes 16
1.5. Product Gas Treatment 18
1.5.1. Purification Processes 18
1.5.2. Conditioning 19
1.5.3. Byproducts 20
2. Steam Reforming of Natural Gas and Other Hydrocarbons 21
2.1. Feedstocks 21
2.2. Natural Gas and Other Gaseous Hydrocarbons 22
2.2.1. Principles 24

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_169.pub2

2.2.2. Catalysts, Catalyst Poisons, Desulfurization . . . 26
2.2.3. Tubular Reformers . . . 27
2.2.4. Production of Fuel Gas and Synthesis Gas . . . 29
2.2.5. Special Reforming Processes . . . 30
2.3. Tubular Steam Reforming of Liquid Hydrocarbons . . . 31
2.3.1. Commercial Processes . . . 32
2.3.2. Fuel Gas and Synthesis Gas from Liquid Hydrocarbons . . . 32
2.3.3. Special Processes . . . 32
2.4. Prereforming . . . 33
2.4.1. Principles . . . 33
2.4.2. Catalysts . . . 34
2.4.3. Prereforming of Natural Gas . . . 34
2.4.4. Prereforming of Naphtha; Rich Gas Process . . . 35
2.5. Autothermal Catalytic Reforming 37
3. Noncatalytic Partial Oxidation and Special Gasification Processes for Higher-Boiling Hydrocarbons 40
3.1. Raw Materials . . . 40
3.2. Partial Oxidation of Hydrocarbons . . . 42
3.2.1. Principle . . . 42
3.2.2. Types of Processes . . . 43
3.2.3. Influencing Raw Gas Composition . 47
3.2.4. Submerged Flame Process . . . 51
3.3. Hydrogenating Gasification . . . 51
4. Gas Production from Coal, Wood, and Other Solid Feedstocks . . . 52
4.1. Fundamentals . . . 53
4.1.1. Thermodynamics of Chemical Reactions . . . 53
4.1.2. Kinetics . . . 54
4.2. Classification and General Characteristics of Gasification Processes . . . 56
4.2.1. Criteria for Classification . . . 56
4.2.2. Criteria for Process Assessment . . 57
4.2.3. Mathematical Modeling of Gasification Reactors . . . 57
4.3. Characterization of Solid Feedstocks for Gasification . . . 59
4.4. Moving- or Fixed-Bed Processes . 62
4.5. Fluidized-Bed Processes . . . 66
4.6. Entrained-Flow Processes . . . 70
4.7. Molten-Bath Processes . . . 72
4.8. Underground Coal Gasification . 74
4.9. Environmental Aspects of Gasification . . . 75
5. Gas Treating . . . 76
5.1. Carbon Monoxide Shift Conversion . . . 77
5.1.1. Fundamentals . . . 77
5.1.2. Catalysts . . . 78
5.1.3. Clean Gas Shift Conversion . . . 79
5.1.4. Raw Gas Conversion . . . 81
5.1.5. General Comments on Reactor Arrangements . . . 82
5.1.6. Noncatalytic Quench Conversion . 82
5.2. Carbonyl Sulfide Conversion . . . 82
5.3. Methanation and Methane Synthesis . . . 83
5.3.1. Definitions and Applications . . . 83
5.3.2. Principles of Methanation . . . 84
5.3.3. Methanation as a Step in Hydrogen Purification . . . 85
5.3.4. Methanation of Rich Gas . . . 86
5.3.5. Methane Synthesis from Gases with High Carbon Monoxide Content . . 86
5.4. Absorption Processes . . . 90
5.4.1. General . . . 90
5.4.2. Processes for Carbon Dioxide and Sulfur Compound Absorption . . . 93
5.4.2.1. Physical Absorption Processes . . . 94
5.4.2.2. Chemical Absorption Processes . . 101
5.4.2.3. Physical – Chemical Absorption Processes . . . 107
5.4.2.4. Comparison of Various Absorption Processes for Hydrogen Production and IGCC . . . 108
5.4.3. Liquid-Phase Oxidation Processes . 112
5.4.4. Removal of Gas Impurities of Low Concentration . . . 116
5.5. Adsorption Processes . . . 118
5.5.1. Fundamentals . . . 118
5.5.2. The “Classic” Method . . . 120
5.5.3. Pressure-Swing Adsorption . . . 121
5.5.4. Adsorption on Activated Carbon . . 122
5.6. Cryogenic Processes . . . 122
5.6.1. Partial Condensation . . . 123
5.6.2. Liquid Methane Wash Process . . . 124
5.6.3. Liquid Nitrogen Wash Process . . . 125
5.7. Gas Separation by Membranes . 126
5.8. Addition of Inerts or Other Substances . . . 127
6. Handling of Byproducts . . . 128
6.1. Aqueous Condensates . . . 128
6.1.1. Mechanical Treatment . . . 129
6.1.2. Extraction and Adsorption of Organic Substances . . . 131
6.1.3. Removal and Recovery of Ammonia and Sulfur . . . 132
6.1.4. Biological and Final Treatment . . . 134
6.1.5. Removal of Heavy Metals . . . 135

6.1.6. Example of an Industrial Application 135
6.2. Hydrocarbon Condensates 136
6.2.1. Distillation 137
6.2.2. Hydrorefining 137
6.3. Gaseous Byproducts 139
7. Typical Examples of Complex Gas Production Plants 141
7.1. Methanol Production from Natural Gas 141
7.1.1. Methanol Production Based on Catalytic Autothermal Reforming . 142
7.1.2. Comparison of Conventional Steam Reforming and Combined Reforming Processes for Methanol Production 143
7.2. Hydrogen Production Based on Heavy Residues 143
7.3. Combined Cycle Power System Based on Coal 147
7.3.1. Introduction 147
7.3.2. Fundamentals 148
7.3.3. Installations and Design Studies . . 150
8. Analysis and Quality Control . . . 152
8.1. Quality Specifications 152
8.1.1. Combustion Characteristics 153
8.1.2. Minor Constituents 153
8.2. Test Methods 153
8.2.1. Determination of Combustion Characteristics 153
8.2.1.1. Calorimetry 153
8.2.1.2. Density 154
8.2.1.3. Wobbe Index 154
8.2.2. Analytical Methods 155
8.2.2.1. General Methods for Determination of Several Components 155
8.2.2.2. Specific Methods for Determination of Individual Components 155
8.2.2.3. Methods for Determination of Minor Components 156
8.2.2.4. Determination of Trace Constituents 158
9. Acknowledgement 158
10. References 158

1. Introduction

1.1. Types of Gases; General Overview of Production Methods and Characteristics

Gases for industrial or heating uses can be classified in various ways. The most appropriate classification appears by calorific (heating) value ranges, in kJ/m^3 (STP):

Water and producer gas	4600 – 12 500
Synthesis and reduction gas	ca. 12 500
Town gas and medium-Btu gas	16 700 – 20 000
Biogas and landfill gas	18 000 – 29 000
High-Btu gas and substitute natural gas (SNG)	25 000 – 37 000

A distinction has deliberately been made between *water gas* and *synthesis gas*, although water gas was formerly used in large quantities for synthesis as well. However, the so-called water gases are based on very specific raw materials and belong to a different generation of production methods compared to synthesis gases produced in modern plants. *Reduction gas* refers to gases used in iron production.

Synthesis gas is the basis for most of the hydrogen (→ Hydrogen) produced, and it is also used as a fuel gas, e.g., in gas turbines (see Section 7.3).

Biogas and *landfill gas* (→ Methane) are produced during biological degradation of organic matter. They may be collected and purified either to be used further (renewable energy) or to prevent air pollution.

1.1.1. Water Gas and Producer Gas

The terms water gas and producer gas were first used in solid fuel gasification technology. Both types of gas were very important from the 1920s to the 1940s. They were the basis for producing synthesis gas, town gas, and fuel gas. The technology is of only minor importance today because the raw material basis has shifted largely from solid fuels to liquid and gaseous fuels, and high-performance processes have been developed for coal-based synthesis gas, town gas, and grid gas.

Production of Water Gas. Water gas is usually produced in an alternating operation, i.e., by alternately hot blowing a fuel layer with air and gasifying it with steam. Among its various modifications, the production of so-called *blue water gas* from coke was the most common method, yielding gas with high hydrogen and

carbon monoxide contents which was suitable for use as synthesis gas. By addition of higher hydrocarbons, the calorific value of water gas could be increased (*carbureting*) so it could be used as town gas.

Although most water gas plants were based on coke, plants also operated with anthracite and noncaking highly volatile types of coal (coal water gas or double gas). Coal with a high volatile matter content gives gases with higher calorific values because gasification gas is supplemented by a considerable amount of devolatilization gas.

In the water gas process with its cyclic mode of operation, coke has two functions: (1) it reacts with blowing air during warm-up and with steam during gasification, and (2) serves as a heat regenerator.

Because of the expensive raw material (coke), complicated batch operation, and relatively low output per gasifier (6000 – 9000 m^3 (STP)/h), this method of producing synthesis and town gases has been replaced by more up-to-date processes. The gas production of synthesis gas plants based on either natural gas (in tubular reformers; see Chap. 2) or heavy oil (by partial oxidation; see Chap. 3), and of modern coal gasification plants (see Chap. 4) is ten times higher than that of water gas plants.

Production of Producer Gas. Producer gas is obtained by supplying a gasification agent consisting of air and steam continuously, rather than cyclically, to a gas producer of the same general design as a water gas producer. This gas has a lower calorific value than water gas because it contains atmospheric nitrogen. Producer gas was obtained from a variety of rotary grate gas producers from the 1920s to the 1940s. Because of its high nitrogen content, which cannot be eliminated at economically justifiable cost, this gas is not suitable for use as synthesis gas; it was employed almost exclusively as fuel gas and as an additive to town gases in municipal works.

In addition to coke, noncaking coals were also used as feedstocks, and a few plants were also built for the gasification of caking bituminous coals under atmospheric conditions and under pressure (Lurgi pressure gasification).

Nature of Water Gas and Producer Gas. The following is a typical analysis of a water gas, in percent by volume:

Carbon dioxide	5	Hydrogen	50
Carbon monoxide	40	Nitrogen	5

A typical analysis of a producer gas, in percent by volume, is

Carbon dioxide	5.5	Hydrogen	10.5
Carbon monoxide	29.0	Nitrogen	55.0

The higher heating values (HHV) for water gas based on coke and for producer gas are 10 500 – 12 500 kJ/m^3 (STP) and 4600 – 6700 kJ/m^3 (STP), respectively.

The nitrogen content of water gas, amounting to ca. 5 vol %, does not originate from coke or coal for the most part, but is due to the fact that air remains in the coke pores on transition from the blowing to the gasification period. This nitrogen content can be lowered by discarding a small fraction of the gas produced at the beginning of the gasification period.

1.1.2. Synthesis Gas and Reduction Gas

The gas production method used depends on the raw materials. The raw materials are mainly natural gas, naphtha, heavy vacuum residue, and coal (see Chaps. 2, 3, 4). In addition, unconventional fuels such as solid or liquid waste materials or biomass gain interest. However, in the recent decades most of the world production of synthesis gas was based on natural gas and sulfur-rich heavy vacuum residues. Because of more stringent environmental protection requirements, these residues may no longer be burned in boiler plants in many countries, so their market value is often low. Countries that do not have these two raw materials, particularly countries in which no natural gas is available, often use naphtha. Depending on certain boundary conditions such as the relationship between oil and coal prices, as well as local factors and strategic considerations, plants based on coal will continue to be built.

Production of Synthesis Gas. Because synthesis gas is processed under elevated pressure, a trend toward operating gas production plants at increasingly high pressure could be observed. Gas production at higher pressure reduces the

cost of compressing synthesis gas to the required synthesis pressure.

Natural Gas as Raw Material (Chap. 2). Synthesis gases are produced from natural gas mainly in tubular reformers. If the atural gas contains virtually no nitrogen, a temperature of 750 – 800 °C is sufficient for gas suitable in *ammonia synthesis.* Residual unreacted methane is reformed in a secondary reformer (catalytic partial oxidation with air) with simultaneous adjustment of its nitrogen content.

For normal *methanol synthesis*, reforming is performed in *one step* in a tubular reactor at 850 – 900 °C in order to leave as little methane as possible in the synthesis gas.

For large methanol synthesis plants, Lurgi has introduced a *two-step combination* that gives better results. In the primary tubular reformer, lower temperature (ca. 800 °C) but higher pressure (2.5 – 4.0 MPa instead of 1.5 – 2.5 MPa) are applied. The secondary reformer operates with a fixed catalyst bed as in the case of ammonia synthesis gas, but pure oxygen is used for partial oxidation.

More recently, Lurgi developed another two-step gas production scheme. It is based on catalytic autothermal reforming with an adiabatic prereformer and has economical advantages for very large methanol plants (Section 7.1).

The various tubular reformer designs differ primarily in the method of firing the catalyst-containing tubes (from the side or top, terraced, or from the bottom). Subsequent developments of combined reforming resulted in the integration of the primary and the secondary reforming step in one apparatus using the thermal energy of the effluent of the autothermal process for heating the steam reformer part (M.W. Kellogg, Uhde). Exxon tested the simultaneous reforming and partial oxidation of natural gas in a reactor with fluidized catalyst. Apart from tubular reformer technology with or without a secondary reformer, partial oxidation (in one step) is still used sporadically for reforming natural gas. This can be performed either noncatalytically (Shell, Texaco) or catalytically (Lurgi). The latter is used especially when low ratios of hydrogen to carbon monoxide are required.

Naphtha as Raw Material (Chap. 2). The purity of naphtha allows the use of catalysts. Therefore, the production of synthesis and reduction gases from naphtha is carried out almost exclusively catalytically and at elevated pressure.

Externally heated tubular reformers containing special catalysts are used to produce *hydrogen-rich gases* (e.g., ammonia synthesis). Two processes stand out from the others: they differ in the types of tubular reformers recommended and, primarily, in the type of nickel catalyst used. Topsøe uses a nickel catalyst without alkali; ICI employs one with a definite alkali content.

In future naphtha might replace methanol as a raw material for hydrogen produced on board of vehicles driven by *fuel cells*. Due to the much smaller size of such production units as compared with plants in the petrochemical industry and due to restrictions by volume and by utilities (e.g., steam and water), the production scheme and above all the reactors and heat exchangers for naphtha reforming in vehicles look rather different.

Apart from single-stage tubular reforming, two-stage processes have also proved successful for the production of *gases rich in carbon monoxide*, such as those required for methanol and oxo synthesis. An example of a two-stage process is the Recatro process, developed by BASF – Lurgi. Here, naphtha is converted with relatively little steam in an initial low-temperature adiabatic stage over a special nickel catalyst to produce a rich gas containing ca. 60 – 65 vol % methane plus hydrogen and carbon dioxide. This gas is supplied together with the remaining steam to a conventional tubular reformer to be reformed to carbon monoxide-rich synthesis gas in the presence of a conventional catalyst. The two-stage operation makes carbon dioxide recycling unnecessary in many cases or reduces it considerably.

A few plants using partial oxidation with (Lurgi) or without (Shell, Texaco) catalysts have been built.

Heavy Fuel or Residual Oil as Raw Material (Chap. 2). For the production of synthesis and reduction gases from heavy and residual oils, two processes are accepted worldwide: the Shell and the Texaco process, both operating under elevated pressure and noncatalytic. The Texaco process was originally designed for partial oxidation of natural gas; the Shell process, developed later, was tailored from the beginning to gasification of heavy oil. The differences bet-

ween the two processes are virtually negligible and involve only details of equipment design.

Coal as Raw Material (Chap. 4). Processes employing coal that are used in practice operate on a noncatalytic basis. Various processes operating under *elevated pressure* are available: Lurgi, BGL (British Gas/Lurgi), HTW (High Temperature Winkler), U-Gas, Texaco, Shell and Prenflo. The Koppers – Totzek process and the original Winkler process operate at *atmospheric pressure.* Coal gasification plants for the production of synthesis gas built since the 1950s are based essentially on the Koppers or Lurgi processes.

In view of the distribution of fossil fuels in the world — almost ten times as much coal exists as petroleum and natural gas, and it is distributed relatively uniformly — and the sharp increase in oil prices in the 1970s, efforts have been made in many countries to develop new coal gasification processes tailored to the requirements of particular countries and their coal deposits.

In most processes, the heat required for gasification is provided by partial oxidation with pure oxygen. These processes are based essentially on the following three process arrangements:

1) moving bed,
2) one or more fluidized beds, and
3) entrained-flow systems in which coal dust and the gasification agent are conveyed cocurrently.

Combinations of these three process arrangements are being investigated, as are processes applying high-temperature melts.

The overall arrangement of various coal gasification plants, including downstream units, is determined essentially by the composition of the raw gases. The gases produced at increased pressure by the Lurgi and by the BGL gasification processes contain methane, which can be converted to carbon monoxide and hydrogen for downstream ammonia or methanol synthesis, or can be removed before entering the high-pressure compressor. The Koppers – Totzek, the Texaco, as well as the other entrained-flow processes yield a gas without methane, but with a higher specific oxygen consumption. As in all high-temperature gasification processes, gas from the latter two processes has a high carbon monoxide content, requiring considerable conversion measures for certain syntheses. The gasification processes applied to coal are also suitable for petrol coke if they can cope with its low reactivity and they are also applied when biomass or waste materials are used as fuels.

Production of Reduction Gas. In principle, reduction gas production is very similar to that of synthesis gas. However, iron ore is usually reduced under atmospheric conditions, so reduction gas can be produced at atmospheric or slightly increased pressure. Reduction gases used for the production of sponge iron by various direct reduction processes (→ Iron, Chap. 2.6) must meet certain requirements. Generally, the proportion of oxidizing components such as steam and carbon dioxide in the gas mixture produced should be as low as possible, always $< 10\,\%$. This goal can be achieved by cooling the gases produced from coal, oil, naphtha, or natural gas and removing carbon dioxide. However, because reduction gases in the shaft furnaces of steel works are needed at high temperature, more cost-effective solutions are being sought and used. In the *Midrex process* natural gas is converted at ca. 1000 °C in the tubular reformer, without the addition of steam, thereby producing only hydrogen and carbon monoxide; the conversion is such that the reformed gases can be used directly in the shaft furnace.

As the recovery of byproducts from coke oven raw gas (benzene, phenol) loses economical attractiveness, reduction gas is considered to be produced from coke oven raw gas by catalytic partial oxidation.

Composition and Aftertreatment. Synthesis gas requirements depend on the intended application.

Ammonia synthesis units now generally operate between 13.0 and 22.0 MPa and require gas with a hydrogen: nitrogen ratio of 3: 1. Impurities are limited in modern processes to < 1 ppm for sulfur and 2 – 10 ppm for oxygen-containing components such as oxygen itself, carbon dioxide, and water. With a view to the economy of synthesis (i.e., the purge gas quantities to be branched off), the methane content of synthesis gas must not exceed 0.5 – 0.8 vol %.

This means that the reformer and gasification gases produced from various raw materials must be conditioned appropriately. If the methane content is originally considerably high it can

be lowered by secondary reforming with oxygen or air. Because all reformed gases contain carbon monoxide, they have to be converted, irrespective of the raw materials from which they are produced; this is done by catalytically reacting carbon monoxide with steam to form hydrogen and carbon dioxide (shift conversion). Depending on its residual content after shift conversion, either the carbon monoxide concentration must be lowered (to the ppm range) in a methanation stage or it must be removed at low temperatures by a liquid nitrogen wash. Hydrogen sulfide, organic sulfur compounds, and carbon dioxide have to be removed by appropriate washing procedures.

Similar conditioning and purification processes have to be applied for the production of hydrogen for *fuel cells* (→ Fuel Cells).

Methanol is generated from carbon monoxide as well as from carbon dioxide:

$$2\,H_2 + CO \rightleftharpoons CH_3OH$$

$$3\,H_2 + CO_2 \rightleftharpoons CH_3OH + H_2O$$

The preferred reaction path goes via carbon monoxide, some carbon dioxide is required for kinetic reasons. The ratio $(c_{H_2} - c_{CO_2})/(c_{CO} + c_{CO_2})$ should generally be between 2.0 and 2.2. The theoretical value derived from adding both reaction equations after pondering each of them arbitrarily is 2.0. A higher carbon monoxide content is required for fuel methanol synthesis (Lurgi Octamix and others), in which a mixture of methanol and higher alcohols is produced as an additive to motor fuel. For high-pressure methanol synthesis (35.0 MPa), which is still used in a few cases, the presence of methane adds an economic burden (as in ammonia synthesis). It is less important in low-pressure (5.0 – 10.0 MPa) methanol synthesis (ICI, Lurgi) because compression costs for this gas component are practically negligible because of the lower compression ratio required. The content of sulfur compounds in the gas used for low-pressure methanol synthesis, which is based on a special copper-containing catalyst, must be extremely low. The amount of hydrogen sulfide and organic sulfur compounds in the synthesis gas produced from raw materials with relatively high sulfur content can be lowered to ca. 0.1 ppm by special processes (Rectisol for coal and oil gasification gases; hydrogenation of the organic sulfur compounds in the presence of Comox and Nimox catalysts, and subsequent adsorption prior to the reforming of naphtha and natural gas in the tubular reformer).

Purities applicable in methanol synthesis are also required for *oxo synthesis*. However, because the latter often requires a hydrogen: carbon monoxide ratio of ca. 1: 1, the carbon: hydrogen ratio of the feedstock may require that carbon dioxide be recycled or fed to the gasification unit.

For iron ore *reduction gas*, the sulfur content must be reduced to a few parts per million. Table 1 shows typical analysis data for raw gases from various starting materials.

Table 1. Typical raw gas analyses

Feedstocks and processes	Content, vol %						
	H_2	CO	CH_4	C_{2+}	N_2+ Ar	CO_2	H_2S + COS
Coal							
Lurgi pressure gasification	43	12	11.5	1	0.3	32	0.2
BGL	28	57	7.1	0.4	4.2	3	0.3
Koppers – Totzek	30	55	0.1		1.4	13.2	0.3
Texaco	37	41.3	0.1		0.6	20.7	0.3
Oil gasification							
Shell – Texaco	46	47	0.5		0.6	5.5	0.4
Naphtha							
Tubular reformer	67	19	3			11	
Recatro (BASF – Lurgi)	66	22	4			8	
Natural gas							
Tubular reformer	73	16	4			7	

1.1.3. Town Gas and Medium-Btu Gas

Because natural gas (→ Natural Gas) has become the preferred fuel gas for households and industry, the production of town and medium-Btu gas from coal and liquid fuels is of minor importance today. High-temperature coking — formerly general practice in town gas works — has disappeared almost completely (→ Coal Pyrolysis). The following is a brief outline of gasification processes for the production of town gas.

Basis of Production. When charcoal was replaced in the last century by coke from bituminous coal to smelt iron in blast furnaces, large quantities of degasification gas were produced by the carbonization of caking coals. Initially, these gases were largely flared; later on, they were distributed through piping systems supplying gas particularly to the municipalities in the

neighborhood of steelworks. The gas grids were large enough to provide sufficient gas from the coke ovens even at times of peak demand; on Sundays and holidays, the gas was flared in the coking plants.

With the increasing demand for grid gas, utilization of the gas system improved, so that towns had to build their own gas production facilities to meet peak requirements. These facilities were based mainly on water gas, which in some cases was carbureted.

Both solid and liquid or gaseous fuels were used as raw materials for town gas production. Liquid hydrocarbons such as fuel oil and, later on, naphtha were preferred, but coal was also used.

Plants using fuel oil are based mainly on a cyclic process arrangement and became very successful since the 1950s. Some of them use straight thermal processes without catalysts, but most employ a catalyst bed. During a blowing phase, a bed of catalytic or noncatalytic material is heated by the combustion of fuel oil with air. After switching to the gasification period, fuel oil is passed over the hot bed along with steam. The ONIA GEGI and SEGAS processes and the Koppers Cowper process have become widely known [1].

The disadvantages of these cyclic processes are their relatively low output and difficulties caused by excessive wear of the switchover mechanism. In addition, as environmental laws tightened toward the end of the 1950s and beginning of the 1960s, plant operators turned to processes using lighter hydrocarbons and continuous operation. The development of special catalysts (ICI, Topsøe) permitted gas production from readily available liquid feedstocks (such as lighter hydrocarbons and naphtha having a boiling end point up to 200 °C) to be carried out in externally heated tubular reformers. Apart from its higher efficiency, this mode of operation had an additional advantage because the gases produced were available at increased pressure, thus reducing the expenditure for their distribution in pressure gas systems. The methane content of the gases could be increased by keeping reaction temperatures in the tubular reformer low, so that after a cold carbureting stage with propane – butane, standard town gas could be produced.

Cold carbureting was often not applicable because either propane – butane was not available at a reasonable price, or certain standards with respect to dew-point were not met, therefore, modified processes aiming at increased production of methane from naphtha (BASF/Lurgi, British Gas, Japan Gasoline Corp.) were developed. Because methane content increases with decreasing temperature according to thermodynamic equilibrium, appropriate catalysts had to be developed. In these processes, a so-called *rich gas* having a methane content of ca. 65 vol % is produced from naphtha and steam by a slightly exothermic reaction in a shaft reactor at 400 – 490 °C. Because the calorific value and specific gravity of this gas are too high, compared with pertinent standards for town gas, part of it must be reformed in a downstream tubular reformer using commercial catalysts and then remixed with the residual methane-rich gas. Another method of obtaining a leaner gas is to add nitrogen or air. Many European cities built plants of this type with town gas capacities up to 1×10^6 m^3 (STP)/d.

Plants that converted propane – butane to town gas via externally heated tubular reformers were built particularly for small capacities. Part of the propane – butane was added to increase the calorific value of the reformed gas after carbon monoxide shift conversion. A few plants produced town gas by reforming natural gas in externally heated tubular reformers for transitional periods. This was done mainly when the consumers' existing piping systems and burners were to be used, even if a loss in efficiency had to be accepted in converting natural gas to town gas.

Oil gasification by partial oxidation (Shell, Texaco, Lurgi; see Section 3.2.2) has not become important in town gas production because of the high investment cost of such plants. The gases produced, consisting essentially of hydrogen and carbon monoxide must be carbureted to comply with standards.

Along with the development of gas production processes, downstream gas conditioning and gas cleaning units were adapted as well. Carbon monoxide shift conversion plants often reduced the carbon monoxide content to < 3 vol %, sometimes even to < 1 vol %, to eliminate the danger of poisoning.

The *composition* of town and grid gas is given in Tables 42 and 43 (Chap. 8). Group A in these tables corresponds to town gas; Group B, to grid

gas; and the first gas family, to town and grid gases combined.

1.1.4. Biogas and Landfill Gas (→ Methane)

Biogas is the gaseous reaction product from anaerobic digestion of wastewater having a high BOD value or of sludge from wastewater treatment plants. It is used solely as fuel for combustion engines to meet the power requirements of the treatment plant.

Landfill gas is produced by the degradation of organic matter in the waste disposed of in landfills under anaerobic conditions. Because access to oxygen is not strictly controllable, aerobic degradation may occur simultaneously, producing water and carbon dioxide. Landfill gas is recovered via pipelines buried in the landfill. For existing landfills, wells are drilled and lined with tubes to allow the gases to be extracted. Landfill gas is used mostly as fuel gas for cogeneration and in combustion engines after suitable purification.

Composition. Biogas is composed of 60 – 70 vol % methane and 30 – 40 vol % carbon dioxide, depending on the feed and digester temperature (mesophilic – thermophilic).

Landfill gas is composed of 50 – 55 vol % methane, 41 – 45 vol % carbon dioxide, and 1 – 4 vol % nitrogen.

Depending on the nature of landfill waste, the gas may also contain considerable amounts of halocarbons.

1.1.5. Rich Gas and Substitute Natural Gas (SNG)

Basis of Production. As a result of the doubts concerning gas supplies and their availability to the industrial nations in the 1970s efforts were made to produce synthetic natural gas from the plentiful coal supplies. These efforts were feasible where natural gas accounted for a high percentage of primary energy and cheap domestic coal was available. Thus, a plant was built in North Dakota (United States) which has gone into operation in 1984, producing ca. 3.9 $\times 10^6$ m^3 (STP)/d of SNG from lignite. In this plant, lignite is gasified in Lurgi pressure gasifiers, and the methane content required is adjusted by reacting carbon monoxide and hydrogen from the raw gas in a methanation stage downstream of the gas cleaning unit. In addition, the Lurgi gasification process has been developed further to produce SNG more economically by increasing gasification pressure to 10.0 MPa (Ruhr 100, see Section 4.4). Here, more than 50 vol % of methane recoverable from coal is produced in the gasification stage proper.

Heavy vacuum residues can also be used as feedstock for SNG production. However, as in coal-based plants, capital investment is relatively high. The discovery of huge new natural gas resources in some areas of the world during the past few years has made the production of SNG less attractive.

Composition. *Rich gases* have a higher calorific value than grid gas, with gross calorific values from about 25 000 kJ/m^3 (STP) upward. A rich gas produced from naphtha (British Gas or BASF – Lurgi process) has the following typical analysis, in percent by volume:

Carbon dioxide	21.6	Hydrogen	12.5
Carbon monoxide	0.3	Methane	65.6

Substitute natural gas should differ as little as possible from natural gas. It must be conditioned so that it can be substituted for natural gas. This depends on the following parameters: higher heating value, density, and Wobbe index (see Section 8.2.1.3).

In addition the specified content of impurities, particularly sulfur, must not be exceeded. Typical SNG produced in a coal gasification plant in North Dakota has the following analysis, in percent by volume (unless otherwise stipulated):

Methane	96.3	Nitrogen + argon	1.1
Carbon monoxide	0.02	Carbon dioxide	1.2
Hydrogen	1.38	Hydrogen sulfide	4 ppm

1.2. Raw Materials for Gasification

Gaseous Fuels. Natural gas as raw material is of paramount importance for gasification or reforming (the term reforming is used more

commonly for gases) in externally heated tubular reformers to yield synthesis gas for various applications. Plants also reform natural gas by partial oxidation. The composition of natural gases is not important for the technology to be applied, as long as their inerts content is negligible. However, if the nitrogen content rises above 5 – 10 vol %, this may have consequences for the design of ammonia synthesis gas plants operated on the tubular reformer principle.

In certain new energy scenarios, natural gas still plays an important role. Thus, the energy required for natural gas reforming may be supplied — at least partially — by solar collectors or high-temperature nuclear reactors. Liquid motor fuels can be synthesized from synthesis gas. In this way, thermal energy, which is difficult to store and utilize, is converted to energy carriers better adapted to the requirements.

Coke oven gas (→ Coal Pyrolysis) is of minor importance for producing synthesis gas because special expenditures are necessary to remove various impurities from the coke oven gas, and compression costs are high. In older plants, coke oven gas was separated into its individual components by fractionated condensation and distillation at low temperature rather than being reformed (the content of methane and higher hydrocarbons is low anyway); these components were then used for synthesis gas production.

Liquid Fuels. In most cases, various fractions of crude oil are used for gasification. Only in very rare cases is unrefined crude oil used directly for gasification because neither its economic nor its technical considerations are optimum.

Liquefied petroleum gas (LPG, a propane – butane mixture; → Liquefied Petroleum Gas) is used for the production of town gas, particularly in smaller capacity plants, and for peak shaving. It is virtually sulfur-free and can be processed without significant problems. *Light naphtha* is utilized for larger town gas plants and for the production of synthesis gas. Generally, fractions boiling at 40 – 110 °C are used, but full-range *straight-run naphtha* with a boiling end point of 150 – 200 °C may also be employed. Depending on its origin, this naphtha has a sulfur content between 30 and 300 ppm. Naphtha must be completely desulfurized because almost all conventional processes for the production of town or synthesis gas include a reaction stage involving catalysts that are very sensitive to sulfur poisoning.

In some cases, LPG and naphtha have also been converted to town and synthesis gas by partial oxidation (thermally or catalytically) with oxygen or oxygen-enriched air. For catalytic partial oxidation, desulfurization to 20 – 50 ppm is sufficient.

Synthesis gas is produced from *high-boiling residues* by partial oxidation with oxygen. Originally, the high-boiling fraction (initial boiling point above 350 °C) from atmospheric distillation units of refineries (→ Oil Refining) — the heavy fuel oil — was used for this purpose. To date even higher viscosity fractions from vacuum distillation and conversion plants are being used increasingly. Because of the more stringent requirements imposed by environmental legislation, these heavy oils can no longer be burned in boiler houses and are, therefore, cheap raw materials for gasification, which also eliminates various pollutants in the residue such as sulfur and heavy metals. To be used in gasification plants, the heavy oil must be suitable for pumping and spraying. Therefore, the oils are preheated to attain the required viscosity.

The ash content of the heavy oils and the composition of the ash affect the operating temperature of gasification plants and the service life of the refractory lining. Vanadium and nickel in ash are particularly harmful.

Apart from petroleum fractions, other liquid fuels may be used for gasification, although they are not very important. These include:

1) tars from bituminous coal and lignite for gasification by partial oxidation;
2) fractions from shale oil (→ Oil Shale) production, particularly heavy fractions, for gasification by partial oxidation;
3) methanol for producing peak shaving gas in catalytic town gas plants;
4) hydrocarbon condensates from natural gas production for gasification by partial oxidation; and
5) waste from chemical plants (e.g., isobutyraldehyde).

Purity Requirements for Liquid and Gaseous Fuels. Raw materials for gasification plants using catalysts must generally be very

pure. Therefore, in most cases, such plants are equipped with special upstream purification units. A certain tolerance regarding the content of impurities can be accepted in cyclic processes that regenerate catalysts in the respective cycles. However, continuous processes important to modern technology require exceptionally pure fuels, especially with respect to sulfur compounds. Catalysts used for these processes thus permit the use only of gaseous fuels (natural gas) and of liquid hydrocarbons up to a boiling end point of ca. 200 °C.

Generally, the lower the operating temperature of the catalytic process, the lower is the residual sulfur content that can be tolerated in the raw materials. How adverse an effect sulfur compounds in the gas have on the active sites of catalysts can be derived from equilibrium conditions governing the reaction of the metallic phase of the catalyst with the sulfur compounds. Most gasification catalysts contain nickel as an active substance. At constant partial pressure of sulfur compounds in the gas, nickel sulfide formation increases with decreasing temperature; a sulfided catalyst no longer provides the desired results regarding reaction direction (danger of carbon formation), speed, and selectivity.

Therefore, light hydrocarbons are desulfurized to < 0.1 ppm of sulfur upstream of plants operated at low catalyst temperature (400 – 700 °C). The sulfur content may often be higher by one or two orders of magnitude in catalytic processes operating above 850 °C.

Solid Fuels. Various solid fuels are not all equally suitable for gasification. In addition, different gasification processes have different fuel requirements.

Because endothermic gasification reactions should be as complete as possible, the suitability of solid fuels for gasification improves as their reactivity increases. This reactivity can be correlated with the oxygen content, which decreases with increasing carbonization (i.e., with increasing age of the fuel).

The classification and designation of various types of coal are different in individual industrial countries (→ Coal, Chap. 4). As a rule, however, they depend on volatile matter, which almost always means classification by age of the coal.

Additionally, the following properties of fuels are essential for assessing their suitability for gasification:

Ash content	Coal grain size
Water content	Melting behavior of the ash
Sulfur content	Behavior of the coals on heating
Caking properties	

Because coal gasification adds considerably to the costs of upgraded energy (i.e., the gas), coal used for this purpose should be as cheap as possible. Wherever possible, only coal that is not suitable for other applications and, therefore, has a low market price should be used. Coal produced by open-pit mining is of special interest for gasification.

Almost all solid fuel such as wood, peat, lignite, young and old bituminous coal, and coke, including petrol coke, has been used for gasification. In the future, more use should be made of lignites and younger noncaking or caking bituminous coal, as well as coal with a high ash content (20 – 50 wt %) or sulfur-rich coal which is rarely suitable for other applications.

With regard to various types of petrol coke (delayed, fluid, Lurgi – Ruhr gas (LR) coke, → Oil Shale,), their high sulfur and heavy-metal content (→ Petroleum Coke) renders combustion increasingly difficult in view of pollution control measures. Petrol coke is, therefore, turning into a new potential raw material for gasification.

Gasification of wood has been revived because of increasing interest in renewable raw materials. Emphasis is mainly on the production of clean fuel gas (low sulfur content of wood), but synthesis gas production is also possible with modern fluidized-bed processes.

Suitability of Fuels for Gasification. Considerable differences exist in the complexity of gasification technology applied, depending on whether solid, liquid, or gaseous fuels are used. A general rule is that capital cost and consumption figures are higher for solid than for liquid fuels, and for liquid higher than for gaseous fuels. Therefore, in the production of particular gases, mere comparison of specific energy prices of various fuels is insufficient; individual costs for gasification must be considered as well (see Table 2).

Table 2. Comparison of capital cost and energy requirements for an ammonia plant and a methanol plant with a capacity of 1000 t/d based on gaseous, liquid, and solid fuel

	Natural gas	Heavy oil	Coal
Ammonia plant			
Capital cost, %	100	170	225
Energy requirements, %	100	115	135
Methanol plant			
Capital cost, %	90	150	200
Energy requirements, %	95	105	125

To make end products competitive, the specific energy price of coal must be lower than that of heavy oil, and the latter must be lower than that of natural gas. In principle, this also applies to other solid and liquid fuels with regard to the complexity of the conversion plant based on these particular feedstocks.

Even if fuels of the same physical state are compared, great differences exist in the properties of individual substances, which also affect their suitability for gasification.

1.3. Physicochemical Basis for Gas Production

Gasification can be described by overall chemical reactions, from which conclusions may be drawn regarding the conversion of materials and heat of reaction.

Gasification of solid and liquid fuels involves heterogeneous reactions at the phase boundary of particles or droplets and homogeneous reactions in the gas phase. The latter include reactions of components that devolatilize at operating temperature.

The following principles form the basis for calculating gasification processes:

1) *Material Balance.* In establishing a material balance, the total number of atoms of the individual elements in the reactants is equalized to that in the products, irrespective of the distribution of the elements between the individual components (principle of conservation of mass).
2) *Energy Balance.* Energy balance is based on the principle of conservation of energy. Consequently, energy supplied to the system (enthalpy of reactants and heat supplied) is equal to energy removed from it (enthalpy of products and heat removed). In determining enthalpy, formation and phase change energies must be considered.
3) *Chemical Equilibrium.* The equilibrium composition of the gasification product is determined by the temperature-dependent equilibrium constants of reactions going on in the system and by the operating pressure.
4) *Chemical Kinetics.* Because equilibrium is not attained when residence time is too short (for example, because the temperature is too low), the composition of the gasification product in this case is determined by the kinetics of the reactions. In addition, transport phenomena, such as boundary layer or pore diffusion, may have some influence (see page 15.

Practical calculation leads to nonlinear equation systems in which the number of unknown quantities is determined essentially by the number of components in the raw gas. The solution may be found iteratively or — if appropriate routines are used — simultaneously. Calculations are particularly complicated when chemical kinetics must be considered.

Reaction Enthalpy of Gasification. Gasification reactions can be described by many different equations. To calculate the reaction enthalpy of gasification, however, a limitation to certain characteristic reaction groups is sufficient. The reaction enthalpy of all other reactions can be derived by combining these basic equations, and individual reactions have no influence on the total heat balance of a sequence of reactions if the same feedstocks are turned into the same products (Hess's law).

The reaction groups characteristic of gasification can be subdivided as follows (ΔH refers to 0 °C and 0.1 MPa as reference state; ΔH is negative for exothermic reactions and positive for endothermic reactions):

1) *Reactions with molecular oxygen* (*combustion*)

$$\mathrm{C}+1/2\,\mathrm{O_2} \rightleftharpoons \mathrm{CO}\ \Delta H = -110.62\,\mathrm{kJ/mol} \quad (1)$$

$$\mathrm{CO}+1/2\,\mathrm{O_2} \rightleftharpoons \mathrm{CO_2}\ \Delta H = -283.15\,\mathrm{kJ/mol} \quad (2)$$

$$\mathrm{H_2}+1/2\,\mathrm{O_2} \rightleftharpoons \mathrm{H_2O}\ \Delta H = -242.00\,\mathrm{kJ/mol} \quad (3)$$

$$\mathrm{C}_n\mathrm{H}_m+(n+m/4)\ \mathrm{O_2} \rightleftharpoons n\,\mathrm{CO_2}+m/2\,\mathrm{H_2O} \quad (4)$$

For example,

$$CH_4 + 2\,O_2 \rightleftharpoons CO_2 + 2\,H_2O \quad \Delta H = -802.86\,\mathrm{kJ/mol} \tag{4a}$$

$$C_2H_6 + 7/2\,O_2 \rightleftharpoons 2\,CO_2 + 3\,H_2O \quad \Delta H = -1428.8\,\mathrm{kJ/mol} \tag{4b}$$

$$C_3H_8 + 5\,O_2 \rightleftharpoons 3\,CO_2 + 4\,H_2O \quad \Delta H = -2045.4\,\mathrm{kJ/mol} \tag{4c}$$

The reaction enthalpy of combustion of other gaseous or liquid fuels is described in the literature (e.g., [2 – 4]).
For *partial combustion* of gaseous or liquid fuels to carbon monoxide and steam,

$$C_nH_m + (n/2 + m/4)\,O_2 \rightleftharpoons n\,CO + m/2\,H_2O \tag{5}$$

the reaction enthalpy can be calculated from the enthalpies of Equations (4) and (2) as follows:

$$\Delta H\,(\mathrm{Eq.\,5}) = \Delta H\,(\mathrm{Eq.\,4}) - n\,\Delta H \tag{Eq. 2}$$

2) *Reactions with steam*

$$C + H_2O \rightleftharpoons CO + H_2 \quad \Delta H = +131.38\,\mathrm{kJ/mol} \tag{6}$$

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad \Delta H = -41.16\,\mathrm{kJ/mol} \tag{7}$$

(homogeneous water gas reaction, shift conversion)

$$C_nH_m + n\,H_2O \rightleftharpoons n\,CO + (m/2 + n)\,H_2 \tag{8}$$

For example,

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2 \quad \Delta H = +206.28\,\mathrm{kJ/mol} \tag{8a}$$

$$C_2H_6 + 2\,H_2O \rightleftharpoons 2\,CO + 5\,H_2 \quad \Delta H = +347.50\,\mathrm{kJ/mol} \tag{8b}$$

$$C_3H_8 + 3\,H_2O \rightleftharpoons 3\,CO + 7\,H_2 \quad \Delta H = +498.06\,\mathrm{kJ/mol} \tag{8c}$$

The enthalpy of the reaction of fuels with steam to produce carbon dioxide and hydrogen

$$C_nH_m + 2\,n\,H_2O \rightleftharpoons n\,CO_2 + (m/2 + 2\,n)\,H_2 \tag{9}$$

can be determined by

$$\Delta H\,(\mathrm{Eq.\,9}) = \Delta H\,(\mathrm{Eq.\,8}) + n\,\Delta H \tag{Eq. 7}$$

because Equation (9) may be considered a combination of Equations (8) and (7).

3) *Reactions with carbon dioxide*

$$C + CO_2 \rightleftharpoons 2\,CO \quad \Delta H = +172.54\,\mathrm{kJ/mol} \tag{10}$$

(Boudouard reaction)

$$C_nH_m + n\,CO_2 \rightleftharpoons 2\,n\,CO + m/2\,H_2 \tag{11}$$

For example,

$$CH_4 + CO_2 \rightleftharpoons 2\,CO + 2\,H_2 \quad \Delta H = +247.44\,\mathrm{kJ/mol} \tag{11a}$$

$$C_2H_6 + 2\,CO_2 \rightleftharpoons 4\,CO + 3\,H_2 \quad \Delta H = +429.82\,\mathrm{kJ/mol} \tag{11b}$$

$$C_3H_8 + 3\,CO_2 \rightleftharpoons 6\,CO + 4\,H_2 \quad \Delta H = +621.53\,\mathrm{kJ/mol} \tag{11c}$$

Generally, the reaction enthalpies for hydrocarbon gasification with carbon dioxide (Eq. 11) are obtained from the enthalpies of Equations (8) and (7) as follows:

$$\Delta H\,(\mathrm{Eq.\,11}) = \Delta H\,(\mathrm{Eq.\,8}) - n\,\Delta H \tag{Eq. 7}$$

4) *Hydrocarbon decomposition reactions* (*soot formation*)
Hydrocarbon decomposition is described by

$$C_nH_m \rightleftharpoons n\,C + m/2\,H_2 \tag{12}$$

For example,

$$CH_4 \rightleftharpoons C + 2\,H_2 \quad \Delta H = +74.91\,\mathrm{kJ/mol} \tag{12a}$$

$$C_2H_6 \rightleftharpoons 2\,C + 3\,H_2 \quad \Delta H = +84.74\,\mathrm{kJ/mol} \tag{12b}$$

$$C_3H_8 \rightleftharpoons 3\,C + 4\,H_2 \quad \Delta H = +103.92\,\mathrm{kJ/mol} \tag{12c}$$

may be considered the reverse of hydrocarbon formation from its elements and is essential for the gasification of hydrocarbons.
The reaction enthalpy can be calculated from the combustion enthalpies of the reactants:

$$\Delta H\,(\text{Eq. 12}) = \Delta H\,(\text{Eq. 4}) - n\,\Delta H\,(\text{Eq. 1}) - n\Delta H\,(\text{Eq. 2}) - m/2\,\Delta H\,(\text{Eq. 3})$$

In calculating the main gas components, side reactions of sulfur compounds (H_2S, COS, etc.) and nitrogen compounds (NH_3, HCN) are generally neglected because they have no great effect on the numerical distribution of the main gas components. If detailed information on the distribution of secondary components is required, further calculations can be made and the results used for the main gas components.

Chemical Equilibrium. The gasification reactions represented by Equations (1) – (12) are never complete in either of the two possible directions but tend to reach an equilibrium expressed by the equilibrium constant K_p. For Equation (8 a), for example, this constant is given by

$$K_p = \frac{p_{\mathrm{CO}} \cdot p_{\mathrm{H_2}}^3}{p_{\mathrm{CH_4}} \cdot p_{\mathrm{H_2O}}} = \frac{x_{\mathrm{CO}} \cdot x_{\mathrm{H_2}}^3}{x_{\mathrm{CH_4}} \cdot x_{\mathrm{H_2O}}} P^2 = f\,(T)$$

where x_i represents the mole fraction and p_i the partial pressure of each of the four components, P the total pressure, and T the temperature.

For Equation (7), the carbon monoxide shift conversion (which in nearly any case has to be considered) the equilibrium constant is given by

$$K_p = \frac{p_{\mathrm{CO_2}} \cdot p_{\mathrm{H_2}}}{p_{\mathrm{CO}} \cdot p_{\mathrm{H_2O}}} = \frac{x_{\mathrm{CO_2}} \cdot x_{\mathrm{H_2}}}{x_{\mathrm{CO}} \cdot x_{\mathrm{H_2O}}} = f\,(T)$$

In comparison to the steam reforming reaction (8 a) the equilibrium composition of the shift conversion is not influenced by the total pressure, because simply spoken, the number of molecules on the right and on the left side of Equation (7) is identical. The temperature dependence of the equilibrium constant (log K_p vs. $1/T$) is shown in Figure 1.

According to the laws of thermodynamics, the equilibrium constant K_p of a reaction can be calculated by means of the Gibbs free energy of the reaction or the free formation energies of the reactants.

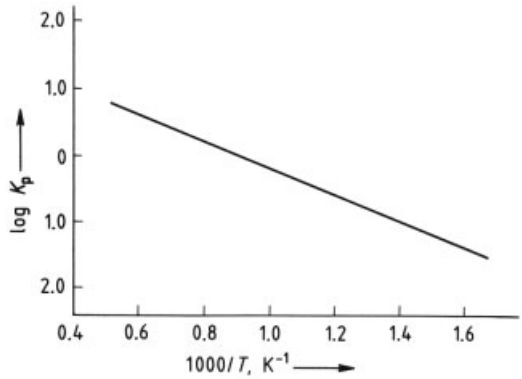

Figure 1. Water gas equilibrium, $K_p = f\,(1000/T)$

Strictly speaking, the equilibrium constant K_p in the above form applies only to ideal gases. At increased pressure (total of ca. 1.0 MPa or more) and near the boiling points of the liquids, the appropriate partial fugacities should be used instead of partial pressures. These fugacities allow for the real behavior of individual gas components. However, in practice, partial pressures can be used for the calculation with sufficient accuracy in most cases.

In heterogeneous reactions, partial pressures of the solids are not taken into account because they are regarded as a function of temperature alone, which is included in the equilibrium constant K_p. However, the free formation energy of the solids must be considered in a thermodynamic calculation of the equilibrium constants.

The three reactions described by Equations (7), (8 a) and (10) are almost always sufficient to calculate the theoretical product gas composition of various gasification processes. Table 3 shows the equilibrium constants of these three reactions at different temperatures. In addition equilibrium constants for heterogeneous methane decomposition (Eq. 12 a) are given as an example of hydrocarbon decomposition. This equilibrium may also be derived from Equations (7), (8 a), and (10). Additionally equilibrium constants for COS hydrogenation and hydrolysis are included (see Section 5.2).

$$\mathrm{COS + H_2 \rightleftharpoons H_2S + CO}\qquad \mathrm{COS + H_2O \rightleftharpoons H_2S + CO_2}$$

The temperature dependence of the equilibrium constants can be calculated with an equation of the following type

Table 3. Temperature-dependent equilibrium constants of various reactions

Equation	(7)	(8 a)	(10)	(12 a)	COS hydrogenation	COS hydrolysis
Unit for K_p		bar^2	bar	bar		
Temperature, °C						
100					12.4	4.39×10^4
200					19.1	4.26×10^3
300	38.5	6.13×10^{-8}	3.60×10^{-7}	4.36×10^{-3}	24.4	940
400	11.5	5.54×10^{-5}	8.55×10^{-5}	5.56×10^{-2}	28.0	323
500	4.82	9.16×10^{-3}	4.99×10^{-3}	3.76×10^{-1}	30.3	146
600	2.51	4.88×10^{-1}	1.16×10^{-1}	1.66	31.5	79.1
700	1.52	11.8	1.42	5.41	32.1	48.7
800	1.02	159	11.0	14.1	32.2	32.9
900	7.41×10^{-1}	1.38×10^3	59.8	31.2	32.0	23.7
1000	5.71×10^{-1}	8.63×10^3	249	60.6		
1100	4.60×10^{-1}	4.12×10^4	845			
1200	3.84×10^{-1}	1.59×10^5	2.43×10^3			
1300	3.30×10^{-1}	5.16×10^5	6.08×10^3			
1400	2.91×10^{-1}	1.48×10^6	1.36×10^4			
1500	2.61×10^{-1}	3.69×10^6	2.79×10^4			

$$\log K_p = A+B/T+C\log T+D\cdot T+E\cdot T^2$$

where T is the absolute temperature in Kelvin. For the various equilibria used in this article the values of the constants A – E are given in Table 4. (The data of Tables 3 and 4 were taken from the ASPEN program [5] which uses in this case the correlations cited in [6]. Minor differences with respect to other literature data may occur.)

Table 4a. Values of coefficients for calculating the equilibrium constants of various reactions as a function of absolute temperature

Equation	(7)	(8a)
A	−8.000980	−18.98177
B	2456.620	−9520.569
C	1.984098	10.91012
D	-0.3329441×10^{-3}	-0.319556×10^{-2}
E	0.563315×10^{-7}	0.3631894×10^{-6}

Table 4b. Values of coefficients for calculating the equilibrium constants of various reactions as a function of absolute temperature

Equation	(10)	(12a)
A	2.220419	−21.61666
B	−8709.775	−2875.50
C	2.495100	9.505546
D	-0.6399432×10^{-3}	-0.3694027×10^{-2}
E	0.5280003×10^{-7}	0.5233708×10^{-6}

Table 4c. Values of coefficients for calculating the equilibrium constants of various reactions as a function of absolute temperature

Equation	COS hydrogenation	COS hydrolysis
A	5.956145	4.485945
B	−566.2154	1568.825
C	−1.309130	−1.681947
D	0.5766548×10^{-4}	0.7452208×10^{-3}
E	-0.1222679×10^{-7}	-0.1160356×10^{-6}

A general assessment of tendencies in gasification reactions is facilitated by the Le Châtelier – Braun principle that a system in equilibrium tries to evade a change forced upon it. For example, in Equation (8 a), an increase in total pressure results in increased methane formation because the number of gaseous molecules on the right side of Equation (8 a) is higher than that on the left. A pressure increase in reactions proceeding without a change in the number of molecules, such as the homogeneous water gas reaction (Eq. 7) has almost no effect.

Reaction Sequence for Various Gasification Processes. No statements can be made on the basis of thermodynamics regarding the chronological sequence of gasification reactions. Calculations must include reaction kinetics. Frequently, however, the required data are not available. Approximation methods must then be used to calculate product gas composition. A common method involves use of K_p values which, according to experience, reflect the actual final state but pertain to a temperature other than that prevailing at the end of the reactor.

Some processes employ catalysts to accelerate the reaction and enhance reaction kinetics in order to approach equilibrium. Selective catalysts promote desirable reactions at the expense of undesirable ones (e.g., carbon formation according to Boudouard). Generally, reaction rate increases as temperature increases. Gasification,

therefore, approaches equilibrium more easily at higher than at lower temperature.

Tubular Reactor Reforming. When reforming hydrocarbons, especially methane, in externally heated tubular reactors, conversion of the hydrocarbons with steam to yield carbon oxides and hydrogen (according to Eqs. 8 and 9) must be considered basic reactions. The composition of the product gas is determined by the simultaneous equilibrium between Equations (7) and (8 a).

Because soot formation should be suppressed, the Boudouard reaction (Eq. 10) and hydrocarbon decomposition (Eq. 12 and, especially, 12 a) are essential for an assessment of the reaction system.

Rich-Gas Process. When naphtha is reformed catalytically with steam in a shaft reactor to yield gases with a high methane content, the endothermic gasification reactions (Eqs. 8 and 9) must be considered as starting reactions, whose products are then transformed in an exothermic methanation reaction described by Equation (8). The final condition of the reforming process is determined by Equations (7) and (8 a).

Oil Gasification. In oil gasification with oxygen and steam, partial combustion of hydrocarbons (Eq. 4), as well as steam and carbon dioxide gasification of carbon (Eqs. 8 and 11), must be considered to proceed more or less simultaneously. Product gas composition is determined by the final state of the simultaneous equilibrium among the reactions described in Equations (7), (8 a), and (10).

Coal Gasification. The first step in a coal gasification reaction (e.g., with oxygen and steam in countercurrent mode of operation) is exothermic combustion of carbon to carbon dioxide, according to Equations (1) and (2). The carbon dioxide formed then reacts with hot carbon to yield carbon monoxide in an endothermic reaction (Eq. 6). This in turn reacts with steam to produce either hydrogen and carbon dioxide (according to Eq. 7) or methane and carbon dioxide (by the reverse of Eq. 11 a). The latter occurs particularly at increased gasification pressure. Carbon monoxide and hydrogen are also produced by direct endothermic carbon gasification with steam (Eq. 6). In entrained-flow gasification processes, the reaction sequence described is mostly overlapping. The mode of reaction essentially determines the temperature profile in a gas producer.

The composition of gasification gas is determined by a more or less accurate adjustment of the simultaneous equilibrium among the shift conversion reaction (Eq. 7), the methane reforming reaction (Eq. 8 a), and the Boudouard reaction (Eq. 10).

1.4. Characteristics of the Basic Processes

In most cases, the aim is to produce gases with high hydrogen and carbon monoxide content. This applies to all synthesis gases and also to town gas. Depending on the particular requirements, the carbon monoxide content may be reduced with steam to yield hydrogen and carbon dioxide. To produce such gases from various solid, liquid, and gaseous fuels, heat must be supplied to support the endothermic reactions. Thus, different processes can be classified according to the type of heat supply.

On the other hand, processes for producing gases with high methane content, such as rich gas and SNG, in principle require no external heat supply because methane formation is exothermic. Therefore, the process characteristics of this group are different.

Trends in the past decades have shifted from processes operated under atmospheric conditions to those operated at elevated pressure. This has resulted in economic advantages because the subsequent chemical syntheses [ammonia, methanol, oxo (→ Oxo Synthesis) and Fischer –Tropsch synthesis (→ Coal Liquefaction)] operate at more or less elevated pressure, which removes the cost of compressing the high-volume synthesis gas. Moreover, plants operating under pressure can be designed for higher capacity due to their smaller gas volume, which again means savings in capital costs.

An essential criterion is whether catalysts are used. With catalysts, a better approximation to reaction equilibria is attained and undesired byproducts are mostly avoided. However, catalysts can thus far be used successfully only for gaseous feeds and for distillate liquid feeds.

Cyclic Processes. Cyclic processes use heat carriers consisting of refractory brick lining (in

regenerators, air heaters, etc.) or beds of ceramic material, which may be appropriately impregnated to impart catalytic properties to them. Regenerative heating is provided by hot flue gases from the combustion of liquid or gaseous fuels. During gasification, the heat carrier transfers its heat to the raw material to be converted. The process is suitable for liquid and gaseous feedstocks. Any carbon black deposited during the reforming period can be burned off after the system is switched back to heating. The process is not suitable for operation at elevated pressure.

Processes using circulating heat carriers are similar to cyclic gasification except for being continuous processes circulating the heat carrier in a closed loop. This carrier is heated by hot flue gases and then fed to a reaction section in which its heat is used for gasification. In addition to sensible heat, some heat carriers used in the past also produced heat through reaction with gas components (producing $CaCO_3$ from burned lime and CO_2). Any carbon formed in the process can be burned off in the heating zone. The process is generally suitable for gaseous, liquid, and solid fuels. The circulating heat carrier may consist of granular catalytic or noncatalytic solids, or of liquids (such as slag baths) and circulating gas. Although system design does not altogether exclude elevated operating pressure, the necessary design and operating conditions would be difficult to meet.

Autothermal Processes. Autothermal processes produce the heat required for gasification through partial combustion of the fuel to be converted in the reactor. Oxygen or air — if nitrogen in the air is not a ballast in the product gas — is added for this purpose. The process may operate with or without catalysts. If catalysts are used, the temperature can be kept lower because a closer approach to equilibrium is ensured. However, the catalytic process cannot be used to gasify heavy and residual oils because most of these oils contain ash particles that contaminate the catalyst. Also, because of the ash content, catalysts for coal gasification are not applied, so far. Either the catalyst would have to be very cheap so that it could be disposed of along with the ash, or it would have to be recovered by a very complex procedure.

Processes applying baths of molten slag, iron, or salts as heat carriers and reaction promoters have been used and are still being developed. A characteristic of autothermal processes is that the oxygen added to generate heat is chemically bound in the product gas, so the ratio of hydrogen to carbon monoxide in the product is lower than in other processes. This type of process is applied to solid, liquid, and gaseous fuels.

Processes Involving Externally Heated Walls. Processes using externally heated walls include tubular reforming, which is used widely today. To keep the tube stresses resulting from elevated pressure and temperature to a minimum, all tubular reformers employ a catalyst filling in the tubes, which allows lower reaction temperature.

Modern materials permit tubes to be designed for wall temperatures of ca. 900 °C (in special cases even 1000 °C at atmospheric pressure) and pressures up to 4.0 MPa. All gaseous hydrocarbons (with low unsaturated hydrocarbon content) and distillate liquid fuels are suitable feedstocks. When liquid hydrocarbons are used, demands on the catalysts are extremely high. Virtually ash-free fuels are used almost exclusively to heat the tubes because ash constituents (salts, metals) would damage tube walls.

Reactors with externally heated walls are also used in gas production from solid feedstocks. This process technology offers the opportunity to employ nonfossil energy (heat from nuclear reactors, solar energy) for gasification.

Rich-Gas Processes. Because no external heat supply is required for rich-gas processes, the equipment and process loops can be chosen freely. In most systems, a catalyst is applied because methane formation is favored by lower temperature. Apart from the unfavorable equilibrium position, the risk of carbon deposits resulting from the Boudouard reaction is higher at increased temperature, and other byproducts are formed as well. All the processes operate at elevated pressure because this promotes methane formation.

Because catalysts are highly sensitive with respect to attrition, reactors are normally of the fixed- rather than the fluidized- or entrained-bed type. Raw materials must be virtually free of all

substances that may poison catalysts, particularly sulfur compounds.

1.5. Product Gas Treatment

Each gasification plant requires gas treatment facilities to purify product gases and also, in many cases, to condition them. Whereas the gas purification system eliminates the components that would affect downstream processing or utilization of these gases, the purpose of a conditioning system is to adjust gas components to the appropriate ratio. The ratio of hydrogen to carbon monoxide, especially, must be matched to the requirements of downstream synthesis or reduction units and other gas consumers.

The task of gas purification and conditioning is most complex downstream of plants gasifying solid fuels. Capital investment in these purification units can be much higher than that required for the gasification plant itself. Treatment of gases produced through gasification of high molecular mass liquid hydrocarbons (heavy oils, vacuum residues) is also rather expensive. Generally, the smaller the molecules from which a gas is produced, the lower is the capital investment required in treatment facilities. The costs of gas purification are also strongly dependent on the amount and type of impurities in the feedstock, particularly its content of sulfur and trace elements.

Whereas most catalytic gasification processes have some purification facilities (removing sulfur, lead, chlorine) upstream of the gasification stage, the impurities in noncatalytic processes should be removed only from the product gas.

Impurities in the Gasification Gas. Depending on the type and composition of feedstock and the type of gasification process used, a gas purification plant must handle the following impurities:

- Dust (coal gasification)
- Carbonization products, tar, gas liquor (coal gasification)
- Carbon (Heavy-oil gasification)
- Oxygen (coal gasification, oil gasification)
- Cyanic compounds (coal gasification, oil gasification)
- Nitrogen oxides (coal gasification, oil gasification)
- Carbon dioxide (all gasification processes)
- Hydrogen sulfide, carbonyl sulfide, organic sulfur
- compounds (all noncatalytic gasification processes)
- Ammonia (all gasification processes)

Other gas components, even if they occur only in trace amounts, require specific measures to prevent long-range damage or malfunction of the plant. This is particularly true for the gasification of coal and heavy oils. Such components may include volatile metal compounds and high molecular mass organic compounds.

1.5.1. Purification Processes

Because of the large number of impurities, along with their widely varying content and combinations, a variety of gas purification processes are available to meet the requirements of downstream processing (synthesis, in particular) and of industrial and municipal consumers (town gas). *Ammonia synthesis*, for instance, tolerates only ca. 5 ppm (mL/m^3) of oxygen compounds (e.g., O_2, CO, CO_2, and H_2O), and the sulfur content must be < 1 ppm (mL/m^3).

Methanol synthesis, on the other hand, requires a sulfur content of < 0.1 ppm (mL/ m^3) and the stoichiometric ratio $(c_{H_2} - c_{CO_2})/(c_{CO} + c_{CO_2})$ must be adjusted to ca. 2 (see Section 7.1.2). To achieve this ratio, the gas must contain a certain amount of carbon monoxide — a factor that influences the selection of gasification and purification processes and also raises the question of whether an additional process stage is required to convert carbon monoxide into hydrogen and carbon dioxide. Carbon dioxide need not be removed completely so the general plant design can be simplified considerably.

Reduction gases for direct reduction of iron ore may contain sulfur on the order of 7 ppm (mL/m^3), but the content of water and carbon dioxide should not exceed 5 vol %.

If several gas purification plants are combined for optimum effect, perhaps including gas conditioning facilities as well, possible interaction should be checked carefully. Interactions could occur because of solvents contained in the

gas, depending on their vapor pressure, different operating behavior, etc. If fuel gas is produced for a combined power cycle (see Section 7.3), the gas purification plant must meet the load change requirements of such plants.

Gas purification processes may be classified as follows:

Adsorption. Adsorption describes the retention of impurities by solids, either through physical deposition or through chemical reaction (chemisorption) (→ Adsorption , see also Section 5.5). Whenever solids have to be reused (e.g., for economic reasons), batchwise operation is necessary because the laden solids must be regenerated.

Sulfur compounds are eliminated by using beds of iron hydroxide and zinc oxide (especially for desulfurizing gaseous and liquid feedstocks for catalytic gasification). Activated carbon is favored in special cases for cleaning gases containing high molecular mass organic sulfur compounds.

Molecular sieves (→ Adsorption; → Zeolites) have been developed for selective removal of particular gases. They can be used not only for carbon dioxide, sulfur compounds, and steam, but also for diatomic gases.

Gas purification systems based on the adsorption (and chemisorption) principle are preferred when small quantities of impurities must be eliminated or small gas throughputs must be purified.

Absorption. A general distinction is made between physical and chemical absorption processes. Physical processes (→ Absorption, Chap. 1.4 , see also Section 5.4) are particularly suitable for high pressure, high impurity content, and large gas throughput. The processes available differ in solvents used, mode of plant operation, operating temperature, method of energy recovery, and potential for recovering several components more or less separately. The selection of appropriate solvents is determined to a large extent by their specific solubility coefficients for the desired gases and by their selectivity.

Chemical absorption systems (→ Absorption, Chap. 1.5) can be used at elevated or atmospheric pressure. They are particularly suitable whenever the absolute quantity of impurities to be removed is small. Conventional processes differ mainly in the degree of purity that can be achieved and the amount of energy (normally steam) required for regeneration. Numerous processes also use some combination of physical and chemical absorption principles.

Special Wash Processes. Special applications include the removal of solids entrained with gases in the form of dust or carbon black, dust-laden tar, and high molecular mass or unsaturated hydrocarbons that tend to form gums.

1.5.2. Conditioning

For nearly all applications, gasification gases not only must be purified but also conditioned to obtain the optimum proportions of desirable gas components (Sections 5.1, 5.2, 5.3).

Carbon Monoxide Shift Conversion. Prominent among gas conditioning processes is catalytic carbon monoxide shift conversion, which is based on water gas equilibrium (Eq. 7). By this process, the hydrogen content of a gas can be increased as desired at the expense of its carbon monoxide content. This is important not only for ammonia synthesis but also for town and grid gas production. The process permits the gas to be adjusted to the standard density and Wobbe index, and also enables nontoxic, low-carbon monoxide town gas to be produced, in conformity with current regulations.

Carbon monoxide shift conversion requires the addition of steam, which can be done in one of several ways. Steam at the appropriate pressure may be added directly to the gas to be converted. If steam is not available in sufficient quantities and under economically justifiable conditions, a so-called saturator cooler loop is often used. With this arrangement, gas entering the conversion section is humidified by circulating hot water, and unreacted steam in the gas leaving this stage is recondensed and recycled to the saturator. A distinction is also made between high- and low-temperature conversion, the latter being provided downstream of the former to reduce the residual carbon monoxide content of the gas to a minimum (e.g., for ammonia synthesis). Shift conversion catalysts for sulfur-free gases are based on iron (high temperature

shift) or copper (low temperature shift). They are supplied as oxides and reduced during operation. For gases containing sulfur compounds the shift catalysts are based on cobalt and molybdenum. They are active only as sulfides.

Methanation. Another gas conditioning method is methanation, the conversion of carbon monoxide, carbon dioxide, and hydrogen to methane according to Equations (8 a) and (9). A methanation stage is included in many ammonia synthesis plants to ensure a high degree of purity of the synthesis gas. Similarly, a methanation step is indispensable if a high-methane gas for SNG production is to be obtained.

Other Processes. The following processes may also be regarded as gas conditioning.

If the calorific value of town gas, grid gas, or SNG is insufficient, the gasification gas may be carbureted by adding C_3 and C_4 hydrocarbons or, less frequently, small quantities of high molecular mass hydrocarbons. On the other hand, too high a calorific value can be reduced by adding nitrogen, air, or oxygen.

For certain syntheses, gases must be dried. Drying is also necessary in pipelines, to protect the piping from corrosion where a risk of water condensation in the presence of carbon dioxide exists. In addition, the formation of gas hydrates is avoided.

Unlike town gas produced in the past from coke oven batteries, the town gas obtained today through gasification no longer contains any chemical compounds that produce a manifest odor. Therefore, such gases are now odorized artificially so that leaks occurring in the piping or at the consumer's facilities can be detected easily.

To recover particular components from a gas stream (e.g., H_2 from purge gas), pressure-swing adsorption (PSA) (→ Adsorption) or membrane processes (→ Membranes and Membrane Separation Processes) can be used. The PSA principle is based on the fact that the adsorption capacity of molecular sieves (→ Zeolites) differs for different gases. Generally, heavier or polar gas components are adsorbed preferentially. When a gas mixture is passed over a molecular sieve bed, the component least amenable to adsorption (e.g., H_2) remains in the gas whereas other components are adsorbed.

Gas separation by membranes is based on the differences in permeation coefficients of different gas components. Permeation denotes the sequence of absorption (solution) of the component in the membrane material, diffusion through the membrane, and desorption. The permeation rate of a species increases with its solubility and its diffusion coefficient in the membrane material. Membranes may consist of ceramic or metallic materials, organic liquids, or polymers.

Both processes, PSA and membrane separation, can be used to advantage for gases at elevated pressure.

To remove traces of higher hydrocarbons (e.g., compressor lubricants) from gases, beds of activated carbon are used. For recovery of low-boiling compounds, as well as purification from traces of higher boiling compounds, cryogenic processes (→ Cryogenic Technology) can be applied. These employ primarily liquid methane or nitrogen to absorb higher boiling gases such as carbon monoxide.

1.5.3. Byproducts

Some gasification processes yield byproducts along with the desired gas (see also Chap. 6). A comparison of different gasification processes should, therefore, always consider such byproducts and assess their importance. Environmental protection, too, plays an important role because it imposes stringent conditions on pollutant discharge.

Many gasification plant feedstocks contain sulfur, and sulfur often constitutes the main byproduct. Whether this is best recovered in the form of elemental sulfur or sulfuric acid is worth considering.

Other coal gasification processes have coke as a byproduct.

Gasification involving coal carbonization processes can produce a large number of byproducts such as tar, oil, gasoline, phenols, cresols, ammonia, and organic acids. Certain liquid byproducts may even contain solids whose elimination would result in additional costs. Because of this, processes have been developed to recycle certain liquid fractions to the gasification plant if their market value is low.

Heavy-oil gasification plants yield carbon as a byproduct. Attempts to use it for special industrial applications have been successful, but demand for the specific qualities of carbon is limited. Therefore, processes are being developed to recycle most of this carbon to the gasification stage. These and processes where carbon is burnt in conventional power stations at little extra cost are of major interest.

Gas liquor is a product of almost all gasification processes. Depending on its composition and on plant requirements, this gas liquor can either be treated for use as boiler feed water or cooling water, or be discharged to receiving water via chemical or biological effluent treatment plants.

2. Steam Reforming of Natural Gas and Other Hydrocarbons

In the steam reforming process hydrocarbons are catalytically converted by reaction with steam into hydrogen and carbon oxides. This is the most common method for producing hydrogen or hydrogen: carbon oxide mixtures in the manufacture of important basic chemicals (e.g., ammonia and methanol), oil refining, and in many other industrial applications (e.g., iron ore reduction, hydrogenation of fats, production of oxo alcohols).

Nowadyas natural gas predominates by far as a feedstock over other hydrocarbons (e.g., naphtha, LPG, refinery gases) or coke oven gas which are used only under special circumstances.

The catalytic steam reforming process in tubular furnaces was invented in 1926 – 1928 by BASF [7]. This process was applied in the United States for the first time in the early 1930s at two commercial plants: (1) to produce hydrogen from natural gas for hydrogenation purposes and (2) to synthesize ammonia. The process was conducted under low pressure (0.4 – 1 MPa) and at temperatures close to 800 °C until the early 1950s; pressures up to 4 MPa and temperatures up to 950 °C are used today.

A special type of steam reforming which has been in use even earlier than tubular steam reforming is *autothermal reforming*, also called *catalytic partial oxidation*. This process differs from catalytic steam reforming in that the required reaction heat is not supplied from outside (furnace) but by internal partial combustion of the feedstock with oxygen or air, admixed to the process feed. As autothermal reforming is carried out in refractory-lined vessels (cold shell), higher pressures can be applied than in tubular steam reforming. Outlet temperatures up to 1000 °C are usual. Due to the additional expenditure for oxygen supply and safe process control this technology was of minor general importance in the past. Of common use, however, is the process modification with air as oxidant in ammonia synthesis. More recently, the oxygen based gasification has been introduced into methanol synthesis. "Combined reforming", a combination of tubular and autothermal reforming, is recognized as the most efficient syngas technology for large scale methanol plants at the moment.

A comprehensive overview of recent developments in steam reforming technology is given in [8].

Whereas in former years steam reformers have been built nearly exclusively at locations where there was requirement for hydrogen or synthesis gas and products derived therefrom, a big change has occurred since the 1970s, in that particularly world-scale production plants for ammonia and methanol have been built at locations, where natural gas is available in abundance (e.g., Trinidad, Borneo, Cape Horn).

Total hydrogen production by catalytic reforming currently (ca. 2000) amounts to over 400×10^9 m^3/a worldwide. Its uses can be broken down approximately as follows: 60 % for ammonia production, 10 % for methanol production, 25 % for petrochemical processes (hydrotreating, hydrodesulfurization, hydrocracking), and 5 % for other industrial purposes. The cost of commercial production is about $ 0.10 per cubic meter of hydrogen. The catalytic steam reforming process is especially low in emissions. Only the off-gas from the steam reformer contains, apart from carbon dioxide, a certain amount of NO_x and, depending on the feedstock, sulfur dioxide.

2.1. Feedstocks

Natural Gas. Natural gases (→ Natural Gas) are used as gasification feedstocks, either directly or after appropriate pretreatment. Table 5

gives examples of pretreated natural gases and other hydrocarbon feedstocks. High-boiling hydrocarbons are largely eliminated before transportation to prevent condensation in the distribution system.

Particular attention must be paid to sulfur compounds in natural gas. A gas suitable for catalytic reforming has to be virtually free of these compounds as sulfur is a strong poison to the employed catalysts. High percentages of nitrogen are undesirable as well, except in ammonia synthesis gas. For hydrogen – carbon monoxide gases, nitrogen must be removed in a complex operation. A certain percentage of carbon dioxide may be tolerated and even be beneficial in methanol production.

The condensation fractions recovered from casinghead gases (see Table 5) extend into the naphtha range and contain higher hydrocarbons that may create problems in catalytic reforming. These fractions are handled similar to naphtha.

Reevaporated liquefied natural gas (LNG) is free of all accompanying components such as carbon dioxide, steam, sulfur compounds, and C_{5+} hydrocarbons, which must be removed for process reasons prior to liquefaction. It does not contain any gas with a lower boiling point than methane, such as nitrogen or noble gases (helium). However, when obtained from the public grid, LNG may contain sulfur compounds (used as odorants; e.g., tetrahydrothiophene) and nitrogen (from calorific value – density conditioning).

Other Gases (see Table 6). Gases of different qualities are derived from refineries, coking plants, the chemical industry, steelworks, etc. They are suitable not only as fuel gases but, increasingly, as gasification feedstocks or as raw materials for the chemical industry. As a typical example, hydrogen used in a refinery can be produced by the conversion of refinery gases or butane.

purge gases) from ammonia and methanol synthesis or olefin production are also used for gas production. Unsaturated hydrocarbons in the residual gases are usually saturated by hydrogenation before they are catalytically gasified with steam. The top gas from some processes for direct reduction of iron ore is upgraded by adding fresh natural gas and used for gas production. Coke oven gas, which was formerly used mainly as town gas, is utilized today for ammonia production or the reduction of iron ore.

Naphtha. At the beginning of the 1960s, light naphtha was available in large quantities for the production of town gas, ammonia, and methanol. Since the 1973 oil crisis, it has been used almost exclusively as a raw material for the production of chemicals (especially ethylene). Exceptions are town gas and synthesis gas production in isolated areas where no natural gas is available.

A distinction must be made between straight-run virgin naphtha as obtained directly from crude oil distillation and naphtha cuts from the production facilities of a refinery, for example, from a reforming unit (high-aromatic naphtha).

Naphthas differ according to geographical origin of the crude. The main criteria for assessing the suitability of a naphtha as a raw material for gas production are its boiling range (low boilers can be converted more easily); carbon: hydrogen ratio; aromatic, olefin, and naphthene contents; sulfur content; and content of metals or other substances (e.g., lead, chlorine, which will poison the catalyst). Table 7 shows the principal characteristics of various naphthas.

Generally, low-boiling naphthas rich in alkanes are preferred for catalytic gasification because they are easier to convert and the catalyst has a longer service life. In individual cases, naphtha with a boiling end point up to 220 °C, mixtures of naphtha and kerosene, or kerosene itself may be considered as raw materials.

Ammonia and Methanol. Ammonia and methanol are occasionally used as feedstocks for hydrogen production (in small units and for peak shaving gas).

2.2. Natural Gas and Other Gaseous Hydrocarbons

In the 1920s BASF developed a catalytic process for endothermic conversion of gaseous hydrocarbons with steam in externally heated tubes, using nickel as a catalyst. Exchanges of experience with Standard Oil (New Jersey) and ICI in the 1930s led — particularly in the United States and England — to the first commercial application of this process in the production of hydro-

Table 5. Properties and composition of (treated) natural gases used as feedstocks for catalytic gas production

	Composition, vol%							
	Ekofisk	Groningen	Saudi Arabia Al Jubail	Indonesia[a]	Soviet Union[a]	United States[a]	Casinghead gas Fed. Rep. of Germany[a]	LNG (typical)
He		0.05				0.07		
N_2	0.52	14.00	0.6	0.34	1.80	2.1	2.7	
O_2				0.21				
CO_2	1.61	1.00	1.0	8.35	1.30	0.2	5.1	
CO								
H_2								
CH_4	86.08	81.60	71.1	83.19	95.74	85.9	63.2	89.7
C_2H_6	8.02	2.78	25.6	3.36	1.0	5.8	8.3	7.7
C_3H_8	2.66	0.37	1.5	2.65	0.07	3.3	9.6	2.6
i-C_4H_{10}	0.29	0.06	0.2	0.79	0.02	1.6		
n-C_4H_{10}	0.57	0.07		0.64				
i-C_5H_{12}	0.10	0.07		0.18	0.07	1.0	11.1	
n-C_5H_{12}	0.09			0.29				
C_6	0.05							
C_{7+}	0.01							
γ_0[b]	0.84	0.83	0.92	0.92	0.75	0.87		0.80
HHV[b]	44 420	35 170	48 150	40 985	39 090	45 430		43 800
Total S[c]		150		10	80			
H_2S[c]		5			20			
Mercaptans[c]					20			
Other[c]					40			

[a] Typical natural gas composition from these countries. [b] Standard density γ_0, kg/m^3; higher heating value (HHV), kJ/m^3. [c] Figures from supply contracts (actual figures are sometimes considerably smaller), mg/m^3.

Table 6. Residual gases (Properties and composition of various residual gases and of a butane fraction

	Composition, vol%							
	Coke oven grid gas	Residual catalytic cracker gas	Residual hydro-cracker gas	Residual hydro-treater gas	Residual hydro-treater gas	Residual gas from ethylene plant[a]	Butane fraction	Mixed refinery gas
N_2	8.2							2.9
O_2	0.5							
CO_2	2.0				0.05			
CO	5.5				1.5	0.4		1.5
H_2	56.3	49.3	53.96	43.19	77.2	3.8		25.7
CH_4	25.0	26.8	36.93	27.94	15.2	94.5		37.4
C_2H_4	2.5	3.4			5.7	1.3		7.9
C_2H_6		12.7	7.16	2.46		0.1		19.6
C_3H_6								1.6
C_3H_8		6.0	1.22	2.46			1.3	1.3
C_4H_8							i, traces[b]	1.7
C_4H_{10}		1.7	0.47	10.45			i = 30.4[b]	
C_5H_{10}							n = 58.2[b]	0.4
C_5H_{12}		0.06	0.13	13.15			0.1	
C_{6+}		0.05	0.13	0.35				
γ_0[c]	0.46		0.45			0.70		0.82
HHV[c]	19 580					40 360		42 580
Total S[d]		65				5	3	36
H_2S[d]								18

[a] Hydrogen removed. [b] i = iso; n = normal. [c] Standard density γ_0, kg/m^3; higher heating value (HHV), kJ/m^3. [d] Figures from supply contracts, mg/m^3.

Table 7. Properties and composition of various naphtha feedstocks for gasification

	Light naphtha				Boiling range of full-range naphtha				Heavy naphtha	Treated naphtha [a]		Condensate from casing head gas
										Hydrotreated	Raffinate	
Density (15 °C), kg/L	0.66	0.68	0.68	0.72	[b]	0.70 – 0.73	0.71 – 0.73	0.71	0.76	0.73		0.74
Initial boiling point, °C	31	41	39	35 – 46	60	39 – 45	43 – 65	44	95	45	99	31
Intermediate boiling point, °C [c]	[b]	70	69	96 – 105	[b]	98	127	91	134	110		138
Final boiling point, °C	89	108 – 125	123	150 – 165	66	180 – 185	165 – 187	200	190	197	206	344
Molar mass				98 – 102								
C/H mass ratio	5.17	5.24	5.35	5.70 – 5.77	[b]	5.4	5.8	5.54		6.0	6.0	5.61
Alkanes, vol %	[b]	ca. 75.5	[b]	[b]	56.3	95 [d]	[b]	81.3	48	72	[b]	71
Olefins, vol %	0	0	0	0.10 – 0.2	0.1	0.3 – 0.6	1	1.5			[b]	[b]
Naphthenes, vol %	[b]	ca. 22	20	[b]	37.4		[b]	7.8	41	8	40	[b]
Aromatics, vol %	1.5	2.5	3.1	3.7 – 6.2	6.3	3.2 – 4.8	11 – 15	9.4	11	20	11	[b]
Chlorides, mg/kg	1.2	3.0 – 4.0	1.2	1 – 3	4	16 – 30	[b]	0.1		[b]	[b]	[b]
Lead, mg/kg	0.02	0.02	0.9	0.01	0.1	0	[b]	< 0.001		[b]	[b]	[b]
Sulfur, mg/kg	3.0	15 – 20	14	11 – 54	120	8 – 10	560 – 635	100		9	[b]	68
Higher heating value, kJ/kg	48 235	47 960	47 855	47 270 – 47 400	[b]	47 835 – 48 275	46 685 – 46 895	[b]		46 520	[b]	[b]
Lower heating value, kJ/kg	[b]	44 465	[b]	44 090 – 44 170	[b]	44 630	[b]	[b]		43 335	[b]	[b]

[a] From reformate.
[b] Not measured.
[c] Boiling point after 50 % reduction in volume.
[d] Paraffins and naphthenes.

gen for hydrogenation and ammonia synthesis. These plants operated at atmospheric pressure. New reformer tube materials, especially the development of centrifugally cast tubes, permitted higher pressure in the gas production section, which in turn improved plant efficiency considerably.

2.2.1. Principles

Catalytic conversion of hydrocarbons with steam to produce gases with high hydrogen and carbon monoxide content, according to the reactions (8) and (8 a) — and always accompanied by the shift conversion (Eq. 7) (see Section 1.3) — is known as steam reforming. This conversion is highly endothermic. It occurs in a number of parallel, externally heated tubes (reformer tubes) filled with a nickel catalyst.

After being preheated in heat exchangers or fired heaters to ca. 500 °C, the steam – hydrocarbon mixture is heated further over nickel catalyst in the tubes to the desired end temperature of normally 800 – 900 °C. Higher hydrocarbons (Eq. 8) are completely converted to carbon monoxide and hydrogen, methane only partially — corresponding to the reaction conditions. The composition of the reformed gas at the reformer outlet reflects the equilibria of Equations (8 a) and (7). For specified conditions (pressure, reformed gas temperature, steam – hydrocarbon ratio), the composition of the reformed gas can, therefore, be calculated from component balances and equilibrium conversion rates. The steam: hydrocarbon ratio is an essential parameter in the steam reforming reaction and must not be lowered arbitrarily, because low ratios lead to carbon deposits on or in the catalyst. Figure 2 shows the influence of pressure and temperature on the composition of the reformed gas for a defined steam: methane ratio in a methane reforming unit.

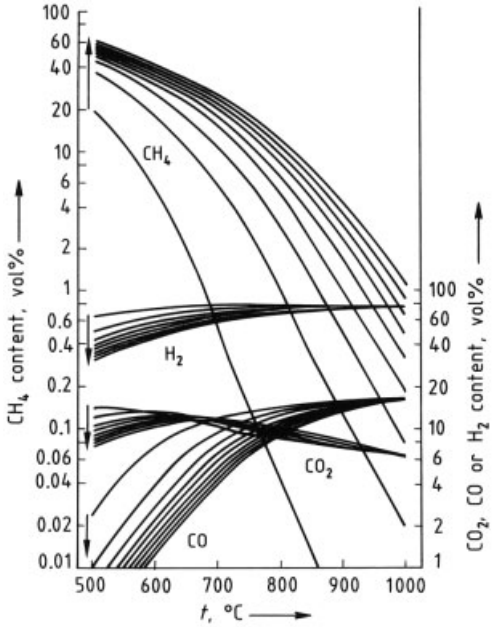

Figure 2. Equilibrium composition for methane reforming at 500 – 1000 °C and pressures (in the direction of the arrows) of 0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4 MPa. Molar steam: methane ratio 3.333 corresponding to 5 kg H_2O per kilogram of carbon

Kinetics. The kinetics of catalytic steam reforming of methane are summarized in [9, 10] and [11]. Many authors observe that the shift conversion reaction described by Equation (7) proceeds very fast so that the gas composition may be assumed to be in equilibrium at every point in the reformer tube — a proposition that is not true for methane conversion which approaches to equilibrium only towards the outlet of the tubes. Sulfur in the gas has no decisive influence on the shift conversion, but can decrease the reaction rate of the hydrocarbon conversion dramatically, especially at the inlet side of the tube.

Methane reforming can be described as a first-order reaction, irrespective of operating pressure. At low temperature, the molecular diffusion rate is much higher than the reaction velocity so that, theoretically, the catalyst activity can be fully utilized. If the temperature is high, the conversion rate is determined by pore diffusion, and catalyst efficiency is reduced. Hence, large-surface catalysts offer an advantage with regard to catalytic activity. For further details of reaction mechanism and kinetics, see [12] and [13].

Carbon Deposition. Carbon deposition is a frequently experienced phenomenon with steam reforming. Though generally various reactions may cause solid carbon formation from carbon compounds containing gaseous phases [e.g., Boudouard reaction ($2\,CO \rightleftharpoons C + CO_2$), hydrocarbon decomposition ($C_nH_m \rightleftharpoons n\,C + m/2\,H_2$), the heterogeneous water gas reaction ($CO + H_2 \rightleftharpoons C + H_2O$)], hydrocarbon decomposition is practically the exclusive cause for carbon formation in the course of reactions in the steam reforming process. Therefore carbon deposition is observed mainly in the first half of the reformer tubes where hydrocarbon concentrations are still high and the approach to equilibrium low.

Investigations have shown [14] that the thermodynamic limits of carbon formation (graphite) do not apply below 700 °C because there the relevant reactions are dominated by kinetics. So the carbon-free area may increase, depending on catalyst properties. Figures for a uranium – nickel – α-alumina catalyst are given in [15] and for a number of other catalysts in [13].

Figures 3 and 4 show experimental "equilibrium constants" for the Boudouard reaction and methane decomposition measured by various authors, as well as calculated figures for thermodynamic equilibrium according to [16]. Figure 4 also depicts the "working line" for methane steam reforming according to [17], by applying two catalysts with different activities. Points on the curves indicate the location of the measuring point as a proportion of tube length (viewed from the gas inlet).

In steam reforming of natural gas, the temperature at the catalyst inlet is such that carbon might theoretically be deposited. This does not occur, however, below ca. 650 °C, but only at higher temperature when an inactive catalyst fails to decompose sufficient methane. Generally, the formation of carbon deposits is not inevitable even at operating temperatures where it might be expected.

Figures 3 and 4 also illustrate the more favorable initial situation and possible further reforming of a naphtha-based rich gas with added carbon dioxide (to obtain high-CO gases) if catalyst activity remains unimpaired (line (c) in Fig.

3 or line (a) in Fig. 4) or if no further methane conversion occurs (line (b) in Figs. 3 and 4).

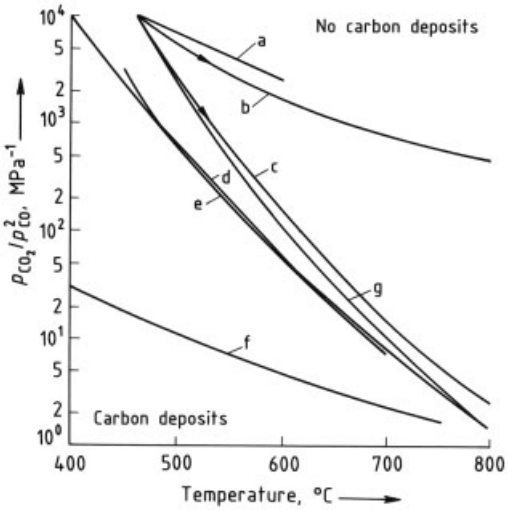

Figure 3. Equilibrium constant of the Boudouard reaction ($2\,CO \rightleftharpoons C + CO_2$)
a) Condition of rich gas from naphtha at inlet to tubular reformer; b) Equilibrium reached for reaction (7) only; c) Equilibria reached for reactions (7) and (8 a); d) According to [14]; e) According to [13]; f) For nickel – uranium catalyst according to [15]; g) For sulfided catalyst according to [13]

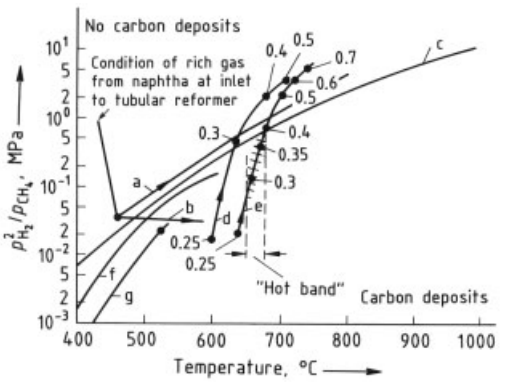

Figure 4. Methane decomposition ($CH_4 \rightarrow C + 2\,H_2$)
a) Equilibria reached for reactions (7) and (8 a); b) Equilibrium reached for reaction (7) only; c) Equilibrium constant for carbon in graphite form according to [16]; d) and e) Working lines for methane reforming in a reactor tube; d) High-activity catalyst; e) Low-activity catalyst with red hot tube section from ca. 650 °C to equilibrium; figures at the data points are proportions of tube length as viewed from inlet; f) Equilibrium constant determined by experiment according to [13]; g) Connecting line between equilibrium constants determined by experiment at 410 and 525 °C according to [14]

2.2.2. Catalysts, Catalyst Poisons, Desulfurization

Steam reforming is generally carried out in the presence of nickel catalysts. These catalysts are usually in the form of thick-walled Raschig rings, with 16-mm diameter and height and a 6 – 8-mm hole in the middle. However, especially in modern steam reformers with high heat flux, catalysts with high geometric surface (HGS-type) are preferred. As already mentioned in Section 2.2.1, the access to the inner surface of catalysts continuously decreases with increasing temperature due to diffusion limitation. This is compensated by introducing catalysts with a higher surface to volume ratio. Typical shapes of this type of catalysts are spoked wheels, gear wheels, or rings with several holes. These catalysts are additionally advantageous with regard to their low pressure drop.

A general distinction is made between impregnated and precipitated catalysts. The latter are the more traditional types which are produced by coprecipitation of all constituents. *Precipitated catalysts* have a Ni-surface which is by one order of magnitude higher than that of impregnated catalysts and for this reason have an excellent activity already at moderate temperatures. But they are more prone to sintering at higher temperatures and do not have the mechanical strength of supported catalysts.

Impregnated catalysts are manufactured by separately impregnating a support with nickel as the active constituent. The main advantage of this type of catalyst is that the support can be calcined at high temperatures before impregnation, resulting in predetermined excellent mechanical properties which are important especially at high temperature application. As the conditions of steam reforming have generally become more severe, impregnated catalysts predominate by far at present. Frequently used support materials are alumina (the α-modification is preferred today), magnesium oxide, or spinel type substances (e.g., magnesium aluminum oxide). In some cases, catalysts are stabilized by addition of a hydraulic binder (calcium aluminum oxide). Crushing strength is decreased by the hydration of magnesium oxide at low temperature (< 500 °C) and by changes from the γ to

the α-modifications of alumina above 500 °C. A silicon dioxide content of < 0.3 wt % in the catalyst is essential, because silicon dioxide becomes volatile in the presence of steam and may foul the downstream heat exchangers through which the reformed gas flows.

Nickel is at first present in the form of nickel oxide; after the system has been heated, it is reduced in the reformer tube by hydrogen formed when natural gas is added to steam. Some catalysts require a special method of reduction for which a definite steam: hydrogen ratio must be maintained (see Fig. 5), according to [12] and [13].

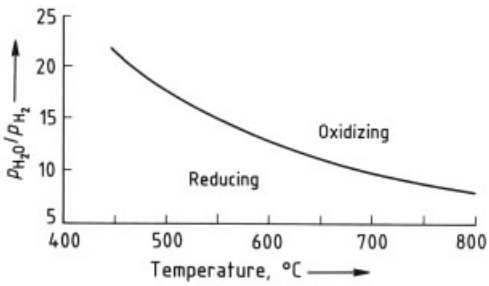

Figure 5. Reduction conditions for nickel oxide

Catalyst Poisons and Desulfurization. Catalytic activity is affected seriously even by very low concentrations of catalyst poisons in gases to be reformed. Such catalyst poisons are sulfur, arsenic, copper, vanadium, lead, and chlorine or halogens in general. Sulfur, found in practically all hydrocarbon feedstocks, in particular lowers catalyst activity [18].

Earlier desulfurization systems used impregnated activated carbon as adsorbent and operated at ambient temperature. However, because their efficiency differs according to the particular sulfur compounds in the gas and because the requirement with regard to the residual sulfur content has increased, zinc oxide desulfurization systems operating at 350 – 400 °C are generally preferred today. Iron-based adsorbents are used in only a few cases where special conditions (i.e., low operating temperature, possibility of regeneration, sulfur composition) prevail.

Zinc oxide reactors are very reliable in absorbing hydrogen sulfide and, with limitations, sulfur compounds such as carbonyl sulfide and mercaptans. To achieve optimum desulfurization conditions, certain space velocities and effective flow rates must be ensured. Sulfur pickup ranges from 15 to > 30 %, that is almost complete conversion of zinc oxide to zinc sulfide is possible.

Organic sulfur compounds such as mercaptans and thiophenes require hydrogenation over cobalt – molybdenum or nickel – molybdenum catalysts. These catalysts are often arranged in a separate vessel. Hydrogen or hydrogen containing gas are added to the process feedstock at temperatures of ca. 350 – 380 °C. The organic sulfur compounds are converted to hydrogen sulfide and the corresponding saturated hydrocarbons. Hydrogen sulfide produced in the hydrogenation stage is then absorbed by zinc oxide. The hydrogenation reactor can be used simultaneously to hydrogenate unsaturated hydrocarbons in the raw gas. This reaction is strongly exothermic. Because the temperature range for the hydrogenation stage is limited to 250 – 400 °C, unsaturated hydrocarbon content is limited as well. Ammonia, carbon monoxide, or carbon dioxide impurities in hydrogen affect desulfurization and can lead to undesirable side reactions such as methanation [12]. An economic solution particularly for natural gases with low sulfur content is the combination of hydrogenation and zinc oxide absorption. This means that both catalysts can be combined in one bed in the same reactor. Residual sulfur content is normally < 0.2 mg/m^3 [19].

2.2.3. Tubular Reformers

Tubular reformers are mostly top- and wall-fired box-type units, apart from older types or special designs (Fig. 6). Whereas in *wall-fired reformers*, rows of reformer tubes are heated mainly by the radiant side wall, *top-* or *bottom-fired reformers* provide most of the heat through radiation of the burner flame and the hot flue gases. Top-fired reformers may, therefore, have several parallel rows of tubes, whereas wall-fired reformers can have only one row. Wall-fired reformers with a large number of tubes consist of several units adjacent to one another; waste heat from the flue gases is recovered in a common system. Modern reformers in large plants that produce ammonia and methanol synthesis gas from natural gas contain hundreds of tubes arranged in rows.

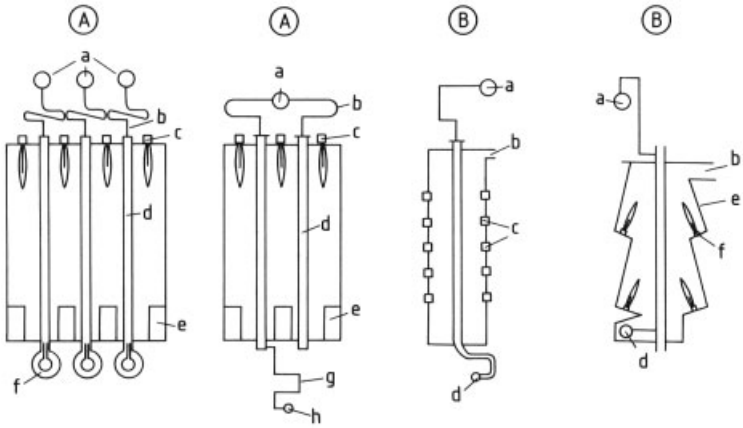

Figure 6. Reformer types
A) Top-fired reformer; a) Inlet header(s); b) Pigtail; c) Burner; d) Catalyst-filled reformer tube; e) Flue-gas collector; f) Refractory lined outlet manifold; g) Outlet pigtail; h) Outlet manifold
B) Wall-fired reformers; a) Inlet header; b) Flue-gas duct; c) Wall burners; d) Outlet manifold; e) Terraced wall; f) Burner

Firebox. The reformer box consists of several layers of refractory insulation and insulating mats with an exterior metal skin. The hot face refractory must be resistant to temperatures > 1200 °C; the temperature is reduced to ca. 75 °C on the outside of the metal skin.

Burner geometry, flame length and diameter, tube-to-tube and row-to-row spacing, fired tube length, and distance from the flame to the reformer wall determine the homogeneity of heat transfer to the tubes. Modern computers are capable of linking up this reformer geometry with heat-transfer processes inside the reformer tubes (radial heat conductivity) to derive complex equation systems from which heat transfer at all points of the firebox can be calculated (see Fig. 7). These equations yield local temperature profiles for flame – flue gas (g – f), reformer wall (d), tube wall (b), (c), reformed gas (a) or catalyst, and local heat flux (e). These figures form the design basis for the reformer. If the tube wall temperature and local heat flux are known, and heating and cooling rates are estimated, the wall thickness of the reformer tubes can be calculated.

One important factor is spacing between the reformer tubes, referred to as tube pitch. In calculating tube pitch, irregularities in heat flux around the tube must be considered [20].

Burners and Firing Systems. Top-fired reformers use turbulent open-jet burners with forced air supply. Appropriate selection of parameters such as jet diameter, fuel and air supply rates, and burner geometry produces a relatively short, stable flame ca. 2 – 2.5 m in length; the reformer tubes, therefore, receive their heat to a considerable degree from flue gas radiation. Accurate calculation of heat transfer requires a knowledge of flue gas backflow and flame emission behavior.

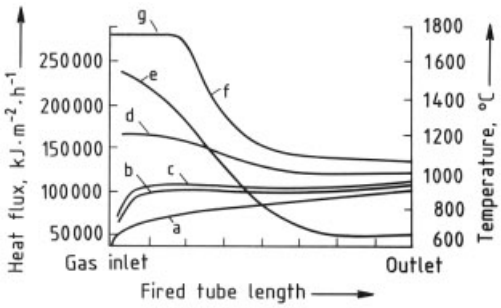

Figure 7. Approximate temperature profiles across a reformer
a) Reformed gas (average over tube cross section); b) Tube wall inside; c) Tube wall outside; d) Heater wall; e) Local heat flux; f) Flue gas; g) Flame

Flue gases are discharged via ducts inside the firebox and cooled to ca. 200 °C in heat exchangers, which are used to preheat the gaseous feedstock, generate steam, etc.

Wall-fired reformers employ other burner designs or flame shapes, such as resist burner cups

into which the fuel is burned or flames burning parallel to the side wall.

Reformer Tubes. Normally, the tubes have inside diameters of 75 – 125 mm, wall thicknesses of 10 – 20 mm, and fired lengths of 9 – 15 m, depending on reformer type (top- or wall-fired). The fixed tube support is usually located outside the firebox at the top or bottom; when heated, the tube expands longitudinally from this support. Counterweights or spring hangers often relieve the tubes of ca. 50 % of their dead weight.

Tube wall thickness is calculated on the basis of internal pressure and of 100 000-h creep-to-rupture strength; see [20]. More sophisticated calculations allowing for mechanical and thermal stresses, and in some cases even establishing a link between them, are being used increasingly [21, 22].

Table 8 summarizes tube materials currently preferred, along with their rupture strengths. Tubes are welded to the required overall length from centrifugally cast tube sections > 3 m in length. The calculated "sound wall thickness" must be increased by an allowance for fabrication tolerances.

The pressure drop inside a reformer tube depends on the gas flow rate and geometry of the catalyst; it is normally between 0.15 and 0.5 MPa.

Pressurized Convection Reformers. A major factor determining the size of reformer tubes is the pressure difference between the reformed gas and the nearly pressureless flue gas. Obviously, the wall thickness of the reformer tubes can be reduced if the pressure in the firebox is increased. This concept is now mature enough for application on an industrial scale [23]. Reformer tubes with a much larger diameter than usual can be employed; the catalyst is arranged in an annular space, and reformed gas is recycled internally by straight or helically coiled tubes (to increase internal heat exchange). Such reformers can be designed for hydrogen output up to 5000 m^3 (STP)/h; beyond this, the wall thickness of the pressure vessel makes the concept economically unattractive.

Higher output can be achieved by connecting several units in parallel. The pressurized combustion reformer requires little space and can, therefore, be prefabricated in modules. However, combustion air and off-gas from the PSA unit required for the firing system must be compressed to high pressure, whereas the flue gas pressure energy can be recovered only partly in expanders.

A promising route for the future might be to increase process pressures to a level at which compression of synthesis gas is no longer required [24].

Inlet Header and Outlet Manifold Systems. The preheated hydrocarbon: steam mixture (400 – 530 °C) is fed to the reformer tube rows via inlet headers with branches to each individual reformer tube. These inlet pigtails are designed to take up the expansion of the inlet header and of the tubes (if the fixed support is underneath the reformer), as well as their own expansion. Header and pigtails are made of carbon steel (or alloyed carbon steel). Flexible metal hoses are used instead of inlet pigtails only in special cases.

Hot reformed gases are discharged from below the catalyst supporting grid in the tubes to outlet manifolds. The outlet manifolds are made of centrifugally cast (see Table 8) or drawn material (Incoloy 800 H). Brick-lined manifolds are used as well; in this case, reformer tubes are mounted directly on the manifolds or connected by small-diameter straight tubes, or "flexitubes" [25]. Another design has manifolds installed in the bottom of the firebox; product gas is extracted through the reformer top by means of one or more risers.

On leaving the outlet manifold, the hot reformed gases are cooled in heat exchangers to generate steam or to preheat boiler feedwater, raw gas, etc. They are then fed to carbon monoxide shift conversion units or carbon dioxide scrubbers for further conditioning.

2.2.4. Production of Fuel Gas and Synthesis Gas

Operation of the reformer under defined reaction conditions, in conjunction with additional process steps, produces the desired gas composition for fuel gas or synthesis gas applications. The variety of possibilities is indicated in Table 9, which lists typical reaction parameters for tubular reforming.

Table 8. Creep-to-rupture strength (average for minimum values) σ_{Rmin} 100 000 for centrifugal castings, N/mm^2

Material designation (DIN 17 006)	Trade names	Temperature, °C						
		800	850	900	950	1000	1050	1100
G-X 40 CrNiSi 2520	HK 40	25.0	18.0	12.0	8.0	5.0	3.0	
G-X 35 CrNiSiNb 2424	IN 519	35.0	25.0	17.0	11.0	6.6		
G-X 50 CrNiSiNb 3030	Märker RE30N	40.0	30.0	21.0	13.5	8.2	4.7	2.4
G-X 45 NiCrNb 3525	Pyrotherm G 35/25 NbTZ, KHR 35C	38.0	29.0	21.0	13.5	5.0	4.7	2.3
G-X 40 NiCrNbTi 3425	Manaurite XM	47.0	35.7	25.7	17.5	11.2	6.8	3.9
G-NiCr 48 W	NA 22 H, Paralloy H48T	34.0	25.0	17.5	12.0	7.8	4.8	2.8
G-NiCr 50 Nb *	IN 657	28.5	20.5	13.0	7.5	3.8		
G-X 10 NiCrNb 3220 **	Pyrotherm G20 32Nb	27.0	18.0	11.0	6.4			

* This material (which has good resistance to oil ash corrosion) is normally used as the hot face layer of centrifugally cast compound tubes; the cold face is made of high-strength material.
** Material for hot but unheated outlet manifolds.

Table 9. Production of various types of gas by tubular reforming of gaseous hydrocarbons

Product gas for producing	Typical process stages *	Typical H_2O/C ratio, mol/mol	Typical pressure at reformer tube outlet, MPa	Typical reformed gas temperature, (primary reformer), °C
CO	desulfurization, tubular reforming with recycled CO_2, CO_2 scrubbing, CO separation (e.g., low-temperature separation, CoSorb, PSA, or membranes)	2 – 3	1 – 3	850 – 900
H_2	desulfurization, tubular reforming, HT and LT conversion, CO_2 scrubbing, methanation (standard arrangement)	4 – 5	1.5 – 3	800 – 900
	desulfurization, tubular reforming, HT conversion, PSA, or membranes	3	2 – 2.5	800 – 900
Oxosynthesis gas	desulfurization, tubular reforming, CO_2 scrubbing, partial separation of hydrogen by PSA, addition of imported CO_2 upstream of reforming if necessary	2.5	1 – 2	850 – 950
Methanol synthesis gas	desulfurization, tubular reforming, or additional secondary reformer	2.5 – 2.8	2 – 3	850 – 900 700 – 800
NH_3 synthesis gas	desulfurization, tubular reforming (primary reformer), secondary reforming, HT and LT conversion, CO_2 scrubbing, methanation	3.5	3 – 4	780 – 830
Town gas	desulfurization, tubular reforming, HT conversion, CO_2 scrubbing	3.0	1 – 2.5	650 – 750
Reduction gas	desulfurization, tubular reforming, or top gas recycling upstream of tubular reforming	1.25 – 1.5 0.65	0.2 – 0.3 0.15 – 0.2	850 – 1000 900 – 1000

* PSA = pressure swing adsorption; HT = high temperature; LT = low temperature.

2.2.5. Special Reforming Processes

Reforming with Additional Carbon Dioxide. Reforming with additional carbon dioxide is used to produce gases with an increased carbon monoxide content. By recycling carbon dioxide from the reformer product gas, partially or totally, the hydrogen: carbon monoxide ratio can be adjusted in a range defined by the composition of the natural gas. By total carbon dioxide recycle the complete carbon content of the process feed is converted to carbon monoxide. In this case the shift conversion reaction (Eq. 7) is eliminated and the hydrocarbon conversion follows exclusively Equation (8). Product gases with particularly high content of carbon monoxide are obtained by adding carbon dioxide from an external source in addition to the carbon dioxide recycle to the reformer feed. This combined steam – carbon dioxide reforming entails a greater risk of carbon deposits, which must be countered by selecting appropriate reaction parameters and catalysts (see Section 2.3.3).

Whereas reduction gases having high hydrogen and carbon monoxide content are normally produced by methane reforming with only a slight amount of steam being added (< 1.5 mol of steam per mole of methane), the Midrex process in particular (direct reduction of iron oxide, → Iron, Chap. 2.6.1) relies heavily on combined steam: carbon dixoide reforming.

Midrex Reforming. In the Midrex process, iron ore reduction and the production of reduction gas are integrated. Natural gas is mixed with off-gas from iron reduction (cleaned and humid-

ified gas from the direct reduction furnace). The latter contains unreacted hydrogen and carbon monoxide in addition to steam and carbon dioxide produced by oxidation of carbon monoxide in the furnace. Typical for the process is a steam: carbon ratio (regarding the molar carbon content of only methane and higher hydrocarbons) of 0.65 and a carbon dioxide: carbon molar ratio (see above) of ca. 0.8. Pressure at the reformer outlet is ca. 0.2 MPa. Special catalysts (Midrex and others) ensure that gases can be reformed without carbon formation.

To this end, either catalysts of different activities must be arranged successively in the reformer tubes, or the activity of the catalyst must be adjusted by means of a controlled sulfur content so that it will be lower at the tube inlet than at the outlet. Reformed gas is conditioned preferably to a hydrogen: carbon monoxide molar ratio of ca. 1.6. Processes for the production of direct reduction gases are described in [26].

Steam Reforming with Addition of Air. In the past, town gas was produced with addition of process air to adjust gas density; examples are the Koppers natural gas reforming process and the Dr. Otto butane reforming process. However, with the conversion of fuel gas systems to natural gas, this technology has become less important. It is still used to make synthesis gas for ammonia production from low-boiling hydrocarbons up to the naphtha boiling range in small compact plants containing a combination of primary and secondary reformers (see below).

Lately, natural gas reforming with addition of air has been used temporarily to adjust gas density during the changeover period from town gas to natural gas until the conversion of gas appliances to natural gas is completed.

The standard technology for producing ammonia synthesis gas from low-boiling hydrocarbons is the two-stage reforming process. The first stage is catalytic steam reforming of hydrocarbons in a tubular reformer (*primary reformer*). Product gas from the primary reformer, which has a temperature of ca. 800 °C, is then reformed with air in an autothermal *secondary reformer* at an outlet temperature of ca. 1000 °C (Section 2.5).

Methane Reforming with Nuclear or Solar Energy. In high-temperature nuclear reactors, recycled helium is heated from 350 to 950 °C. For catalytic reforming, hot helium is circulated in indirectly heated heat exchangers in countercurrent to methane and steam flowing through the reformer tube, releasing its sensible heat and being cooled from 950 to 600 °C. The preferred reformer tube design has an inner helical tube through which reformed gas is discharged to heat the catalyst-filled tube; see [27, 28]. Such reactors, termed EVA reactors, have been tested in pilot plants since 1971 and are considered suitable for commercial reformed gas production. Energy is supplied to the heating helium in the core of a high-temperature nuclear reactor. In *CRS* (i.e., Central Receiver System) *solar plants*, this occurs in the receiver.

Reformed gas can be used to produce basic chemicals (H_2, NH_3, CH_3OH) or, in conjunction with methanation, as a heat-transfer system (ADAM – EVA Systems, see Section 5.3.4). No known commercial applications have yet been found.

2.3. Tubular Steam Reforming of Liquid Hydrocarbons

In the mid-1950s, ICI initiated widespread use of naphtha as feedstock for the steam reforming process. This was enabled by the development of special nickel catalysts in conjunction with a very efficient naphtha desulfurization. Initially, the ICI process was frequently used for town gas production. Due to the abundance of natural gas, naphtha reforming is meanwhile of minor importance. Only at some locations with no access to natural gas there is still some hydrogen and syngas production based on naphtha.

Principles. Direct reforming of naphtha is not generally different from natural gas reforming. The conversion follows Equation (8) in the same way as already described for the higher hydrocarbons contained in natural gas (see Section 2.2.1). Like natural gas reforming naphtha reforming is carried out in externally heated tubes over a nickel catalyst and yields, via a complex reaction mechanism, a gas mixture consisting of hydrogen, carbon monoxide, carbon dioxide, methane, and steam. The gas composition reflects the equilibria of Equations (8 a) and (7) at

outlet conditions. If less active or partially deactivated catalysts are used, the equilibrium concentration of methane (as per Eq. 8 a) is not attained, and the methane content of the reformed gas is higher than at equilibrium. Due to the higher carbon content in the process feed naphtha reforming results in a product gas with increased carbon monoxide (and dioxide) content compared to natural gas feed.

2.3.1. Commercial Processes

A mixture of desulfurized and evaporated naphtha with superheated steam (molar steam: carbon ratio 3: 1 or higher) is supplied to reformer tubes at 450 – 550 °C.

Catalyst. Due to the high carbon: hydrogen ratio in the feedstock naphtha direct reforming is characterized by an increased tendency of carbon formation. This is mainly caused by naphtha cracking and dehydrogenation reactions, especially on acidic centers of catalysts based on alumina which is the most frequently used support material. The breakthrough of naphtha reforming has been enabled by the development of alkalized catalysts. The acidic centers of the support are neutralized by adding potassium to suppress naphtha cracking. Moreover, the steam conversion of carbon once formed by naphtha cracking is strongly enhanced by the presence of potassium. However, alkali also delays complete reforming to carbon monoxide and hydrogen. Therefore, the portion of the reformer tube that is filled with potassium-laden catalyst is restricted to the minimum required for reliable naphtha conversion; the rest of the tube is filled with more active alkali-free catalyst.

Today, a controlled continual release of potassium is achieved by incorporating a special potassium compound ($KAlSiO_4$) in the catalyst, thereby reducing the potassium vapor pressure [12]. Alternatively, alkali-free catalysts mostly based on magnesium oxide are used.

Desulfurization. Naphtha generally contains organic sulfur compounds, and the sulfur content may be as high as several hundred parts per million. It must, therefore, undergo hydrogenating desulfurization with hydrogen at 350 – 400 °C. Cobalt – molybdenum catalysts or, if gases with a certain carbon oxide content are intended for hydrogenation, nickel – molybdenum catalysts are used. The sulfur converted to hydrogen sulfide is then adsorbed on zinc oxide, which has been used commercially as a hot desulfurization agent since the 1940s.

2.3.2. Fuel Gas and Synthesis Gas from Liquid Hydrocarbons

The description given in Section 2.2.4 and Table 9 applies analogously to the production of gas from naphtha. These plants are distinguished from those based on natural gas by the following features:

1) More complex desulfurization systems;
2) Use of a special catalyst in the tubular reformer and, in conjunction with this, a special start-up system;
3) Fewer reformer tubes per quantity of hydrogen and carbon monoxide produced at equal heat loads per unit area (natural gas requires considerably more reforming energy than hydrocarbons in the naphtha boiling range); and
4) Larger carbon dioxide washing systems (naphtha has a higher carbon content than methane).

Details of modern integrated *ammonia plants*, particularly the steam system, are described elsewhere (→ Ammonia) [9].

In *town gas production* by the ICI process, the reformed gas is produced with a heating value below that of town gas. The calorific value of the reformed gas must, therefore, be increased

1) By adding low-boiling hydrocarbons (e.g., LPG) and
2) By producing gases with a higher calorific value in special process stages for addition to the reformed gas. This has been achieved by means of a downstream gas recycle hydrogenator (GRH) or a parallel rich-gas stage based on naphtha (see Section 2.4), or by reacting reformed gas from the tubular reformer with additional naphtha and steam in a secondary reformer [29].

2.3.3. Special Processes

Steamless Carbon Dioxide Reforming. Feedstocks with high carbon: hydrogen ratios

(such as naphtha) and low steam: carbon ratio favor the production of synthesis gas with high carbon monoxide and low hydrogen content (as required, for instance, for the oxo synthesis). In practical applications, however, this system soon reaches the point at which carbon formation occurs, particularly at high pressure. This difficulty has until now been overcome by reforming the gases under established process conditions (i.e., CO_2 recycling with less steam than in regular steam reforming), washing out carbon dioxide, separating hydrogen and carbon monoxide in a cryogenic or PSA unit, and finally remixing them in the desired ratio. In this concept undesirable hydrogen surplus can often be used only to fire the reactor.

The hydrogen production can be minimized by converting hydrocarbons only with carbon dioxide according to the reactions

$$CH_4 + CO_2 \rightleftharpoons 2\,CO + 2\,H_2$$

or as example for higher hydrocarbons

$$C_3H_8 + 3\,CO_2 \rightarrow 6\,CO + 4\,H_2$$

This leads to hydrogen: carbon monoxide ratios of 1: 1 or even less.

Steamless carbon dioxide reforming requires still more sophisticated conditions and catalysts than naphtha direct reforming [30]. It has been realized in the Calcor C process that employs special catalysts and ambient or only slightly elevated pressure. This system recovers the necessary carbon dioxide from the reformed gas in a monoethanolamine (MEA) or diethanolamine (DEA) wash unit, recycles it to the reforming stage, and adds imported carbon dioxide as required [31].

Downstream of the reformer and carbon dioxide wash unit, raw gas from a C_3– C_4 feedstock, for example, will contain hydrogen and carbon monoxide at a ratio of ca. 0.65, with a methane content below 1 vol %. For higher carbon monoxide concentrations, ranging to > 99.5 vol %, carbon monoxide and hydrogen are separated in a PSA unit or cold box. Plants designed for 160 and 250 m^3 (STP) of carbon monoxide per hour have been constructed to date.

Higher Feedstocks. Catalytic reforming of hydrocarbons with a higher boiling range has often been investigated. In most cases, only short-term tests were made, which did not reach the stage of commercial application. The use of very high-boiling hydrocarbons is described in [32].

In other tests, naphtha, crude oil, and heavy residues have reportedly been converted to reformed gases with a hydrogen content of ca. 60 – 65 % without desulfurization or removal of heavy metals [33]. The alternative of noncatalytic conversion of high boiling hydrocarbons or even residues and the high availability of natural gas have slowed down the interest in these feedstocks at present.

2.4. Prereforming

Prereforming is a term which has been introduced into common nomenclature of gasification processes in the 1980s. In the most used sense it means the application of an adiabatic low temperature steam reforming step upstream of a conventional steam reformer using natural gas as feedstock. Regarding effect and conditions this process step corresponds to the steam conversion of naphtha in the so-called rich gas process which has been applied for nearly 40 years. Even the same or very similar catalysts are used [34].

Prereforming can be defined as a steam reforming process at limited temperatures ($<$ 700 °C) resulting in an intermediate product gas, the main compound of which is methane besides steam. The process is in thermodynamic equilibrium with respect to the methane (Eq. 8 a) and shift conversion reaction (Eq. 7). The product does not contain higher hydrocarbons. Normally this intermediate gas is further processed in a steam reformer, but in special cases it can also be used for other applications, particularly for the production of fuel gas (e.g., town gas, SNG).

2.4.1. Principles

Prereforming follows the same sequence of reactions as tubular steam reforming characterized by Equations (8), (8 a), and (7). But in contrast to the latter, the equilibrium is established at far lower temperatures. The main feature of prereforming is the irreversible, complete conversion of the higher hydrocarbons fed to the process. Due to the comparably low temperature the main compound of the product gas is — besides unconverted steam — methane. For this reason the

steam conversion of naphtha is called "rich gas process", which reflects the fact that the resulting product gas is rich in methane. The remaining compounds are hydrogen, carbon dioxide, traces of carbon monoxide, and the inerts contained in the feed.

Depending on the feedstock prereforming can be endo- or exothermic. Generally the steam conversion of hydrocarbons to carbon oxides and hydrogen is endothermic. But as prereforming is carried out at only moderate temperatures, the carbon oxides are further converted to methane, a reaction that is highly exothermic. So, if naphtha is used as process feed, prereforming is exothermic whereas in the case of natural gas, the main compound of which is already methane, the overall reaction is endothermic.

2.4.2. Catalysts

As prereforming is a steam reforming process at low temperatures, special catalysts are required to achieve sufficiently high reaction rates. Common properties of these catalysts are a high nickel content (30 – 70 wt %) and a high specific surface (e.g., 200 m^2/g); the support material may be made, for instance, of alumina with minor additions of potassium or magnesium silicate. The geometric form is also adjusted to the special requirements among which the conversion of higher hydrocarbons is the most important. To restrict diffusion ways small pellets (e.g., 5 × 5 mm tablets) are generally used.

Because of the lower temperature, prereforming catalysts are still more sensitive to poisoning than conventional steam reforming catalysts. This is especially true for sulfur poisoning, which is at least partly reversible at the higher temperature level of tubular steam reforming, but irreversible at prereforming conditions. For this reason special care has to be taken in the desulfurization step, the residual sulfur content of which should not exceed 0.2 mg per kilogram of gas. Naphtha is desulfurized by using the refinery-proven technology of hydrogenating desulfurization (see Section 2.2.2).

2.4.3. Prereforming of Natural Gas

Of the different variations of this process, prereforming of natural gas is most frequently used at present. Driving force for the application of this technology is the general constraint for increased economy.

The prereformer is installed upstream of the tubular steam reformer to ease its operation. As all higher hydrocarbons contained in the natural gas are converted, the risk of carbon formation, which is the most critical point in steam reforming, is considerably lowered. This allows the decrease of the steam: carbon ratio and an increase in the heat load on the reformer tubes, which means less energy consumption and smaller equipment. In addition hydrogen is produced by partial conversion of natural gas and furthermore residual traces of catalyst poisons are ab- or adsorbed in the prereformer. This gives, effectively, a catalyst of optimum activity at the inlet of the tubular reformer, where high activity is required most.

Prereforming is generally carried out in adiabatic reactors with a typical inlet temperature close to 500 °C. Due to endothermic hydrocarbon conversion the outlet temperature is lower by 25 to 40 °C, depending of the content of higher hydrocarbons in natural gas. Figure 8 shows a characteristic temperature profile of an adiabatic prereformer.

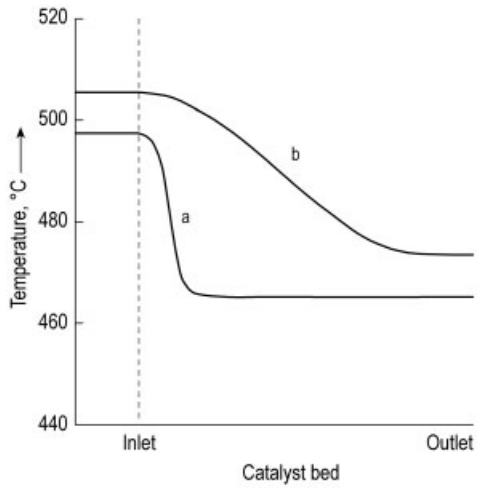

Figure 8. Characteristic temperature profile in a natural gas prereformer
a) Fresh catalyst; b) Spent catalyst

An option that is frequently made use of in prereforming is the reheating of prereformed gas

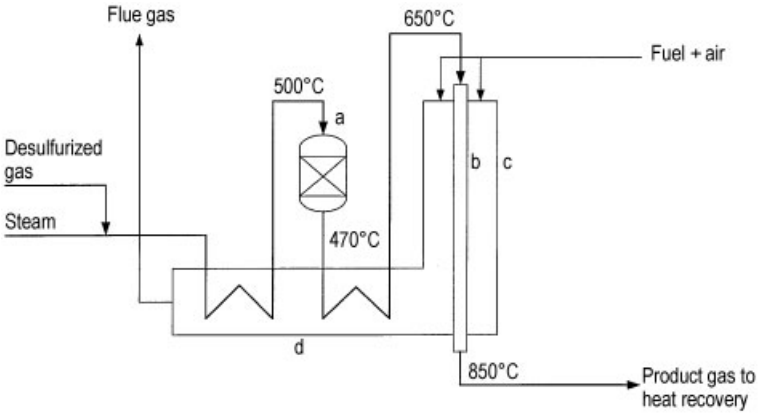

Figure 9. Tubular steam reforming with prereformer and reheating of prereformed gas
a) Prereformer; b) Reformer tube; c) Furnace; d) heat recovery

before feeding it to the steam reformer. As steam reforming technology generally has a surplus of thermal energy which can only be utilized for steam generation, this is an additional possibility to improve the overall energy efficiency of a gas production plant. In conventional steam reforming the reintegration of available process heat either from flue or from product gas is restricted by the risk of cracking, that may occur when heating natural gas – steam mixtures to temperatures above ca. 550 °C. This risk depends essentially on the content of higher hydrocarbons. Due to the absence of hydrocarbons except methane and the increased hydrogen content, prereformed gas can be reheated to 650 °C without any problem (Fig. 9).

An alternative to adiabatic operation is carrying out prereforming in heat exchanger type reactors by supplying heat from an outside source. This source can be the product gas of the downstream main reformer (steam reformer or secondary reformer) or the flue gas of the firebox of the steam reformer. In this case the outlet temperature of the prereformer can be up to 700 °C. By this arrangement even more available sensible heat can be reintegrated in the process than by reheating.

2.4.4. Prereforming of Naphtha; Rich Gas Process

The rich gas process is the traditional implementation of prereforming. In addition to prereforming in its actual sense this process has been used particularly for the production of fuel gases during the decades of its application.

Development of the rich-gas (high-methane gas) process started in England in the mid-1950s [35]. Parallel developments took place in Germany and Japan from about 1960. Today, the commercially available processes are those of British Gas, BASF – Lurgi, and Japan Gasoline.

Initially, the process was used mainly for the production of town gas. Soon its advantages were also applied to the production of high-carbon monoxide synthesis gases. In the early 1970s, the rich-gas process was used in the United States to produce substitute (or synthetic) natural gas (SNG).

The rich gas process is carried out commercially at 350 – 550 °C and 1 – 5 MPa. The outlet temperature of a rich-gas reactor is ca. 30 – 100 °C higher than the inlet temperature of the hydrocarbon/steam mixture. Like with prereforming of natural gas the composition of the rich gas can be calculated from the equilibrium constants of the reversible reactions (Eqs. 8 a, 7) according to [36] (see Fig. 10).

The sequence of reactions occurring typically in an adiabatic shaft reactor can be observed by following the temperature profile. At the inlet of the catalyst bed the temperature profile is at first determined by endothermic cracking and gasification reactions, which cause the temperature to drop by ca. 20 °C. Further in the course of the reaction, exothermic hydrogenation reactions prevail, leading to a temperature increase of ca. 60 °C above inlet temperature (Fig. 11).

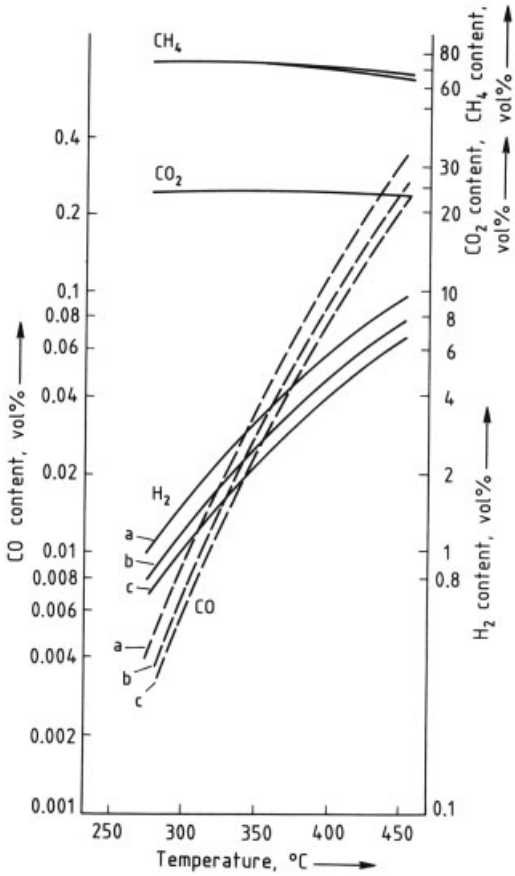

Figure 10. Equilibrium composition of rich gas C: H = 6; 2.0 kg H_2O per kilogram naphtha
a) 2.5 MPa; b) 4.0 MPa; c) 5.5 MPa

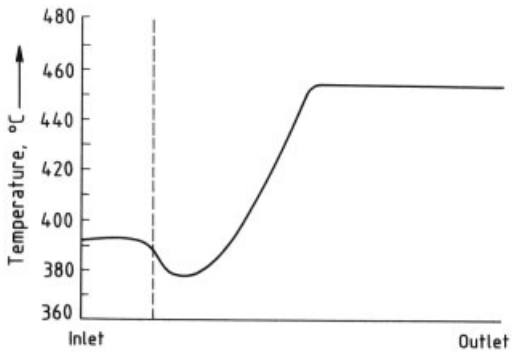

Figure 11. Characteristic temperature profile in a rich gas reactor

During operation, the temperature profile moves through the catalyst bed to the gas outlet. When the upper temperature point (i.e., the end point of the temperature increase) approaches the end of the catalyst bed, naphtha breaks through. Depending on the downstream reaction stage (e.g., tubular reformer or methanation), naphtha breakthrough can be tolerated only up to a certain concentration.

Migration of the temperature profile is caused by exhaustion or blocking of the active nickel centers because of recrystallization, polymer deposits, or catalyst poisoning (sulfur, chlorine, arsenic, etc.). If caused by polymer deposits, this profile migration can be reversed to some extent by regenerating with steam or hydrogen.

Three processes have predominated commercially in the past:

BASF – Lurgi Process. This process is used commercially for the production of town gas, high-carbon monoxide synthesis gas, carbon monoxide, hydrogen, and SNG. The first commercial plants have been in operation since 1965 [37].

This process uses a relatively stable rich-gas catalyst which is fully active as low as 400 °C. Another characteristic of this process is the preferred use of indirectly heated naphtha evaporators.

CRG Process. The catalytic rich-gas (CRG) process is based on developmental work by F. J. Dent (Midland Research Station of the Gas Council). The process was used mainly for producing town gas and SNG over a nickel catalyst on an alumina carrier material. The temperature profile moves almost unchanged through the entire CRG catalyst bed. The first commercial plants have been in operation since 1965 [36].

MRG Process. A special feature of the methane rich-gas (MRG) process is the use of a selective hydrogenation catalyst in the hydrogenating naphtha desulfurization stage, which is capable of operating with hydrogenation gas having a low hydrogen content [38].

Substitute Natural Gas (SNG). The first commercial plants for production of SNG were built in the United States. They use a rich-gas

reactor, generally followed by two methanation stages, carbon dioxide washing, and gas drying. Steam: naphtha ratios of 2.0 kg/kg or slightly higher are applied. An improvement was the installation of a second rich gas stage after the first one. To the second rich gas stage only naphtha is added (in about the same proportion as to the first rich gas stage) or naphtha and only a small amount of steam. Thus, the overall steam: naphtha ratio decreases to ca. 1.1 – 1.4 kg/kg. The temperature range of the second stage varies according to the process. The usual inlet temperature is 350 – 380 °C; the outlet temperature is about 100 °C higher. The two-stage mode lowers steam consumption (i.e., increases efficiency) and the amount of steam to be condensed downstream. It improves the rich gas composition by decreasing the hydrogen content after the second rich gas stage.

Usually, rich gas from the rich-gas stage(s) is cooled to ca. 300 °C and methanated in a shaft reactor under adiabatic reaction conditions by using nickel catalysts. Because undecomposed residual steam in the rich gas is present during this reaction, the process stage is termed *wet methanation.* After cooling and steam condensation, the methane content can be increased further in a second methanation stage (*dry methanation*). The dry methanation stage can be arranged upstream or downstream of the carbon dioxide scrubber (see Section 5.3.3).

The possibility of multistage rich-gas production and multistage methanation offers a great variety of combinations and process alternatives for SNG production. Figure 12 shows the overall diagram of an SNG plant with combined two-stage rich-gas production and single-stage methanation.

Town Gas. The rich-gas process was originally used for town gas production. The standard production concept of town gas from naphtha, with total conversion of naphtha in a rich-gas stage followed by partial reforming of the rich gas in an externally heated tubular reformer to yield a hydrogen-rich gas (see Fig. 13), was developed quite early. Hydrogen-rich reformed gas from the tubular reformer is treated further in the usual manner (CO shift conversion, CO_2 scrubber, etc.).

The ICI method of injecting evaporated methanol into town gas upstream of the high-temperature conversion stage has been applied in some town gas plants as an elegant way of increasing capacity to cover peak requirements in winter. This has enabled the capacity of existing plants to be increased by 20 – 30 %.

Methanol can be converted to town gas in a rich-gas stage. However, in view of the comparatively high price of methanol, this is economically justifiable only for covering peak requirements in winter. Although capital investment costs for the plant are low, the process has found only limited application.

2.5. Autothermal Catalytic Reforming

Catalytic reforming with air or oxygen has been in use longer than either tubular reforming or partial oxidation. Representatives of this first generation technology are the ONIA-GEGI and the SEGAS processes, both used for the production of fuel (town) gas. Feedstock in these cases was gas oil or even crude oil, resulting in a high deactivation rate of the catalysts in use by coking. Correspondingly these processes were operated discontinuously with periodic regeneration. Plant designs today, however, differ substantially from this type of plant, as target product (syngas) as well as feedstock (natural gas) have changed completely and continuous operation is a "must".

Characteristic for autothermal reforming is the high flexibility of the process. This generally enables a wide range of reaction conditions with regard to feedstock, steam: carbon ratio, temperature and pressure. Typical operating conditions are outlet temperatures close to 1000 °C and pressures between 3 – 4 MPa. Suitable feedstocks are natural gas, LPG, and naphtha up to a boiling end point of ca. 200 °C. Furthermore autothermal reformers are often fed with gases, which have already passed a steam reformer ("primary reformer"). In this case, the autothermal reformer is called "secondary reformer". Today, the combination of tubular steam reforming and autothermal reforming is the standard technology for the production of ammonia synthesis gas. Autothermal reforming is more flexible than tubular reforming because the higher allowable operating temperature can compensate for any increase in methane slip which higher pressure would otherwise cause. Furthermore, the higher

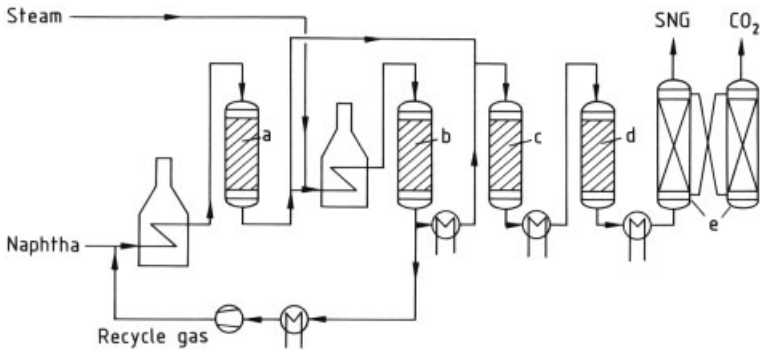

Figure 12. SNG plant
a) Desulfurization; b) Rich gas reactor; c) Hydrogenating gasification stage; d) Methanation; e) CO_2 wash

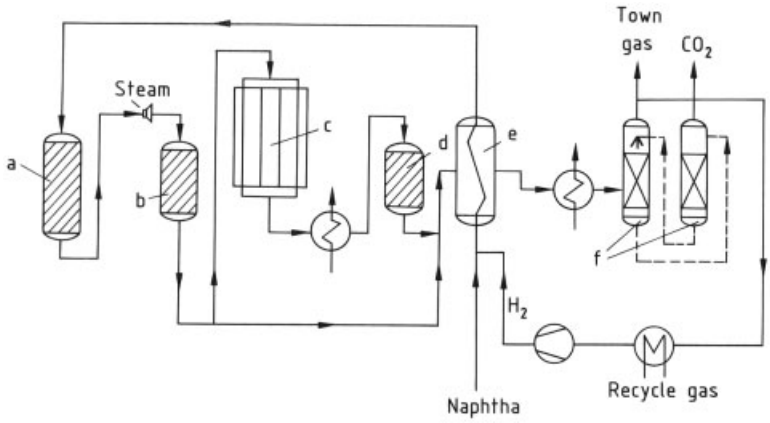

Figure 13. Diagram of a town gas plant operation on the rich gas process based on naphtha
a) Desulfurization; b) Rich gas reactor; c) Tubular reformer; d) HT conversion; e) Naphtha evaporator; f) CO_2 wash

operating temperatures can also compensate for a wider equilibrium approach (difference between actual and equilibrium temperature at the reactor outlet) which may be caused by mild sulfur poisoning of the catalyst. The maximum operating temperature is not limited by the tube material, as in tubular reformers, but only by the stability of the catalysts and the refractory lining of the reactor.

Oxidizing agents may be oxygen or air. The selection depends on whether the presence of nitrogen in the reformed gas is desirable, e.g., for the production of ammonia. By proper partition of primary reforming and secondary reforming with air, a final syngas with stoichiometric hydrogen: nitrogen ratio is obtained. The application as secondary reforming in ammonia synthesis is by far the most frequent commercial application of autothermal reforming.

There are two versions of the process differing in reactor design and arrangement of the catalyst bed. In one type of reformer the reactants (feedstock, steam, and oxygen or air) are discharged directly from a mixer to the catalyst in a fixed bed reactor. In the second type the mixer acts as a burner discharging into an empty, usually conical space above the catalyst bed, which only fills the lower portion of the reactor. The latter modification has generally proven as the more versatile and is used predominantly.

Typical applications of the two versions of the process are compared in Table 10. A methane tail gas was selected as feedstock for the all-catalytic version, whereas reformed gas from a tubular reformer for the production of ammonia synthesis gas (typical use of a secondary reformer) was employed for the version with precombustion.

Table 10. Data for two reformed gases obtained by autothermal reforming

	Methane tail gas reforming with oxygen addition	Reforming with air addition and precombustion (secondary reformer)
Gas for reforming		
Composition, vol %		
CO_2		7.11
CO	3.90	4.25
H_2	1.80	34.32
CH_4	80.00	3.90
C_2H_4	0.25	
C_2H_6	0.05	
N_2	1.70	
Ar	12.30	
H_2O		50.42
Flow rate, m^3/h	55 000	160 000
Pressure, MPa	3.0	3.04
Temperature, °C	500	780
Oxygen or air		
Flow rate, m^3/h	27 725	30 400
Temperature, °C	240	550
Steam		
Flow rate, kg/h	113 200	
Temperature, °C	485	
Reformed gas (dry)		
Composition, vol %		
CO_2	13.23	10.59
CO	16.99	11.67
H_2	63.23	55.52
CH_4	1.08	0.19
N_2	0.65	21.77
Ar	4.82	0.26
Flow rate, m^3/h	148 400	109 000
Reformed gas (moist)		
Flow rate, m^3/h	251 450	196 000
Pressure, MPa	2.65	2.97
Temperature, °C	925	965

The catalysts used consist of thick-walled Raschig rings 16 mm in diameter and height, having a hole of 6 – 8 mm in the middle (standard size). The nickel content is about 3 – 12 wt %. Aluminum oxide (preferably the α-modification) is used as carrier material, and the content of silicon dioxide should be < 0.2 % because of its volatility with steam.

Figure 14 shows the design of a modern commercial autothermal reformer. The refractory lining is almost always multi-layered; high-strength bricks or other lining materials (castables) are used on the hot face, whereas materials of lower bulk weight and high insulating capacity are used on the cold face (inner wall of pressure vessel). Special attention must be paid to ensuring a close fit between lining and pressure vessel (vapor stops), because the considerable pressure drop in the catalyst bed involves the danger of bypass streams between lining and pressure vessel, which can lead to overheating (see Fig. 14). In one frequently used modification the reformer is provided with a water jacket that generally cools at atmospheric pressure but occasionally is slightly pressurized.

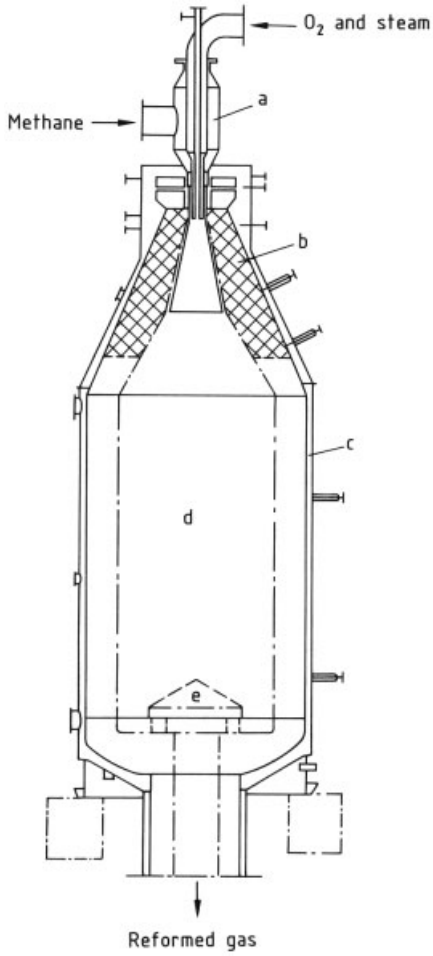

Figure 14. Autothermal reformer (secondary reactor)
a) Burner; b) Refractory lining; c) Water jacket; d) Catalyst; e) Grid

The catalyst grid is usually made of heat-resistant refractory material. Catalyst can be removed by suction from the top.

The burner (mixer) is the most critical part of the autothermal reformer. This is true for both types of reactor design. In most cases, the burner is designed as a single-tube mixer. Various types of burner nozzles are available, all intended to

produce a homogeneous mixture of the oxidant (oxygen or air) and the gasification agent before they enter the catalyst bed. The burner internals are made of highly heat-resistant steel or cast steel; some burner types are cooled.

To achieve sufficient mixing of the different inlet gas streams is a special problem in the *flameless version* of the process, because a certain free space below the mixer is required. As linear gas velocities have to be high in this area to avoid ignition, the mechanical stress on the catalyst is high, leading to abrasion and disintegration. This effect causes an increase in the empty space below the mixer that has to be compensated for by periodical refilling of catalyst during operation which is difficult to perform. The advantage of this operation mode is that the maximum temperature in the reactor does not essentially exceed the adiabatic end temperature (reactor outlet temperature).

In the *open flame version* of autothermal reforming generally a good mixture of the feed gases is achieved. The problem to be solved in this case is to keep the flame, which has temperatures even in excess of 3000 °C, away from the reactor wall and the catalyst surface. An enormous progress has been made in the last years by applying model calculations (Computational Fluid Dynamics) for burner and flame [39]. Therefore, this modification of autothermal reforming has become the clearly superior technology.

Besides ammonia synthesis — where autothermal reforming (secondary reforming) is standard technology — and a specific application in the gasification of Fischer – Tropsch tail gas, autothermal reforming is also applied in some large-scale methanol plants. The syngas production in these cases follows the mode of "combined reforming", a combination of tubular steam reforming and autothermal reforming comparable to ammonia syngas production. Combined reforming enables the production of a syngas with an optimum ratio of hydrogen to carbon oxides and minimizes the feed gas consumption in this way. However, pure oxygen is required as oxidant. This requires additional investment cost, if the oxygen has to be produced on site. Therefore combined reforming is economical particularly for large-size plants with a capacity above 1000 t/d (see Section 7.1).

As already stated, a main advantage of autothermal reforming is the "cold" shell of the reactor, which allows the application of high reaction temperatures and pressures. Whereas a temperature close to 1000 °C is already a certain limit in view of the heat resistance of the used catalysts, the potential of applying higher pressures is not yet finally utilized. Pressures up to 8 MPa, usual in noncatalytic partial oxidation, seem to be possible also for the catalytic version and will make it increasingly attractive in the near future. The design of a plant with one pressure level for syngas production (only autothermal reforming) and methanol synthesis has been patented.

3. Noncatalytic Partial Oxidation and Special Gasification Processes for Higher-Boiling Hydrocarbons

3.1. Raw Materials

Partial oxidation is the most commonly used process for gasification of heavy oil although virtually all hydrocarbon mixtures regardless of origin, are suitable feedstocks.

Crude Oil, Crude Oil Distillates and Residual Oils. Crude oil (→ Oil Refining) is a mixture of many different types of hydrocarbon. These can be classified in one of three ways: as alkanes, (and isoalkanes), naphthenes, and aromatics; by the number of carbon atoms in the molecule; or by their boiling point, which is approximately proportional to this number. However, crude oils also contain varying quantities of other substances, which also play a role in the usability of the oil and, especially, of the residue. The most important impurities are sulfur compounds; others are oxygen- and nitrogen-bearing components, as well as minerals (sand and dissolved salts).

In addition, some impurities not naturally present in the crude stem from its transportation; the main example is water, usually seawater with a high salt content. Residual oils may also contain impurities that arise from wear of catalyst, refractory material, and equipment in conversion units.

As a general rule, low-boiling components are distilled from the crude oil and marketed as

light naphtha, heavy naphtha, kerosene, and gas oil. However, some crudes contain only small amounts of light ends; the extremes are plastic asphalts (→ Asphalt and Bitumen). These resources, from which marketable products can be obtained only by applying expensive refining techniques, are suitable as feedstocks for partial oxidation with practically no pretreatment, except removal of water, salts, or solids.

Crude oil distillates in the boiling range of ca. 35 – 360 °C are valuable raw materials for the chemical industry and are also processed into motor fuels or used as light fuel oil. Like crude oil, these distillates are used for gas production only in special cases, for example, if a surplus of various distillates cannot be balanced out in the refineries.

The gas recycle hydrogenator (see page 51), in which only hydrocarbons with a final boiling point $\leq$ 325 °C can be gasified, is an example of the way in which process considerations can restrict feedstock suitability.

Residual oils are the preferred raw material for partial oxidation. First, they are generally cheaper than distillates; second, the percentage of residues that cannot be used for combustion purposes is increasing because of pollution control. For example, in some places, power stations are only permitted to burn heavy oils with a sulfur content $<$ 0.8 %.

Refinery technology, too, contributes greatly to the increasing quantity of high-boiling, high-sulfur residues. To recover an ever-increasing proportion of light and middle distillates from crude oil, processes have been developed that also enable recovery of naphtha and gas oil from distillation residues. Among these processes are liquid-phase cracking (→ Oil Refining, Chap. 3.2), catalytic reforming (→ Oil Refining, Chap. 3.4), solvent deasphalting (→ Oil Refining, Chap. 3.6), and hydrocracking (→ Oil Refining, Chap. 3.5). In some cases, the residue from these processes may even be solid coke; it contains practically all the impurities in crude oil, apart from those sulfur compounds that have already been driven off during distillation and salt from seawater, which is removed beforehand in the desalter.

On assessing the suitability of residual oils for gasification, the quantity and type of ash, feedstock viscosity, and above all, the ratio of alkalis to vanadium which is important for the lifetime of the refractory lining must be considered. Properties such as boiling range, carbon – hydrogen ratio, and sulfur content do not influence suitability but are important for assessing design parameters. Some typical residual oils and other feedstocks currently used for partial oxidation are listed in Table 11.

Other Feedstocks. In addition to crude oil and its manifold products, a number of other hydrocarbon mixtures, as well as carbon- and hydrogen-rich compounds, are used for gasification of heavy oils. Chief among these are certain liquid byproducts from coal gasification and upgrading.

In gasification of bituminous coal by the Lurgi pressure gasification process (see Section 4.4) aqueous condensates as well as medium oils and coal tar are obtained as byproducts. Medium oils have a variety of uses, but the generally fluid tar can only be recycled to the gasification reactor or used as a feedstock for heavy-oil gasification. Precleaning, at least by settling, is necessary because of the generally high ash content of the tar, which may be as much as 15 %.

Lignite tar oil is a byproduct of the carbonization of lignite or lignite briquettes. Like coal tar, it may contain a larger or smaller quantity of dust, depending on the nature of the raw material and the characteristics of the carbonization process. Carbonization of lignite is uneconomic today. However, certain plants are still operated; in some cases, the char and lighter condensation products are sold and the lignite tar oil is used as a feedstock for gas production by partial oxidation. In gasifying coal and lignite tars, attention must be paid to the content of low-melting ash, which may require slagging operation.

In a few cases, undesirable or unsalable *byproducts of synthesis processes* are used for partial oxidation, preferably when the gas produced by oil gasification is suited to such synthesis processes. An example of this is oxo synthesis, in which oxo alcohols and various byproducts are produced from propylene and carbon monoxide: hydrogen mixtures. One of these byproducts is isobutyraldehyde, which is formed in considerable quantities. It has already been used for partial oxidation with great success.

The heavy products of *tar sand extraction* (→ Tar Sands) and *oil shale retorting* (→ Oil Shale)

Table 11. Specific data for typical feedstocks (basis: Lurgi oil gasification, gasification pressure: 6 MPa; no soot recycle)

Feedstock properties	Feedstock type				
	Natural gas	Vacuum residue	Propane asphalt	Petroleum coke (35 % water)	Orimulsion [a] (29 % water)
Specific gravity (15 °C)		1.05	1.14	1.03	1.002
Kinematic viscosity, m Pa · s		180	620	150	300
at °C		200	200	80	20
C/H ratio, kg/kg	3.01	8.72	9.02	26.65	8.88
S content, wt %		4.1	6.1	3.4	4.00
V content, mg/kg		210	280	1000	300
Ni content, mg/kg		70	120	220	50
Na content, mg/kg		30	40	20 – 120	80
Ash content, wt %		0.07	0.12	0.05	0.03
Higher heating value HHV (dry), MJ/kg	53.32	39.82	39.65	42.63	42.84
Feed flow rates, kg per 1000 m^3 (STP) [b] of CO + H_2					
Hydrocarbon	286	360	372	426	368
Oxygen [c]	338	378	387	416	403
Steam (380 °C)		126	173	20	16
Gasification mode	boiler	boiler	boiler	quench	quench
Product gas composition, mol % (dry)					
H_2	62.0	45.4	44.0	32.9	43.9
CO	33.5	49.1	48.5	58.0	48.7
CO_2	3.1	3.9	4.5	5.9	4.8
CH_4	0.8	0.3	0.2	0.2	0.2
N_2 + Ar	0.6	0.2	1.3	1.4	1.4
H_2S		1.0	1.4	0.9	0.9
COS		0.1	0.1	0.1	0.1
Product steam [d] (saturated)					
Gross, ex boiler, kg	755	929	990	755	793
Pressure, MPa	9	13	13	1.1	1.1

[a] Aqueous emulsion of a highly viscous crude from the Orinoco basin in Venezuela.
[b] 1 m^3 (STP) = 1 m^3 at 0 °C and 0.101325 MPa.
[c] Expressed as 100 % oxygen; actual O_2-purity: 99.5 mol %.
[d] Total steam from heat recovery boilers (basis: boiler feed water at 105 °C); medium pressure boilers in quench mode are optional, product gas may be fed directly to CO shift unit.

may also be suitable as feedstocks. Their use, however, lies in the future.

Natural gas is still used as a feedstock for the production of synthesis gases with a higher ratio of carbon monoxide to hydrogen than is achievable by steam reforming. For economical reasons, partial oxidation of natural gas is used in small plants and in regions where natural gas is cheap. Lately it is considered as a means of converting "*stranded gas*" into marketable liquid products like methanol or synthetic fuel.

3.2. Partial Oxidation of Hydrocarbons

3.2.1. Principle

Partial oxidation is, in principle, the reaction of hydrocarbons with an amount of oxygen insufficient for complete combustion at temperatures between 1600 and 1350 °C and pressures up to 15 MPa. It operates as a continuous process. The basic reactions are as follows:

$$C_nH_m + n/2\,O_2 \rightleftharpoons n\,CO + m/2\,H_2 \quad (13)$$

$$C_nH_m + n\,H_2O \rightleftharpoons n\,CO + (m/2+n)\,H_2 \quad (14)$$

$$C_nH_m + n\,O_2 \rightleftharpoons n\,CO_2 + m/2\,H_2 \quad (15)$$

The minimum amount of oxygen required for complete conversion of the hydrocarbons is indicated by Equation (13); 0.5 kmol of oxygen is required for every kilomole of carbon. If the reaction proceeds according to this equation, when high-boiling distillates or residual oils (i.e., feedstocks with a large number of carbon atoms per molecule) are being gasified sufficient heat is generally produced to warm the reactants to about 1500 °C. Carbon monoxide and hydrogen are the main products until the hydrocarbons have been completely converted; only then can carbon dioxide and water be formed from

surplus oxygen. Opinions differ on the reaction sequence; some test results indicate that carbon dioxide and water are primary reaction products [40].

To prevent excessive temperature increase, steam is usually added, which reacts endothermically with the hydrocarbons according to Equation (14). This leads to the formation of more hydrogen than would be expected from conversion according to Equation (13).

The main elements of the hydrocarbon mixtures — carbon, hydrogen, oxygen, and sulfur — are converted independent of the nature of the feedstock, to the following thermodynamically stable compounds: carbon monoxide, carbon dioxide, hydrogen, water, methane, hydrogen sulfide, and carbonyl sulfide. The proportions of various components in the gas mixture are determined by a range of equilibria, which include the shift conversion:

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{7}$$

the methane equilibrium:

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2 \tag{8a}$$

the hydrogen sulfide – carbonyl sulfide equilibrium:

$$H_2S + CO_2 \rightleftharpoons H_2O + COS \tag{16}$$

and

$$CO + 1/2\,O_2 \rightleftharpoons CO_2 \tag{2}$$

$$CH_4 + CO_2 \rightleftharpoons 2\,CO + 2\,H_2 \tag{11a}$$

These equilibria are largely established in the gasification reactor between 1500 and 1350 °C. They remain practically unchanged during very rapid cooling by quenching with water or by indirect heat transfer to water boiling at a comparatively low temperature. Below 900 °C, even an approximation to equilibrium could be achieved only by accepting very long residence times or by using catalysts. The latter cannot yet be applied commercially because of soot formation. The principal advantages of partial oxidation are (1) the possibility of converting hydrocarbons that cannot be vaporized and (2) the complete destruction of all carbon – carbon bonds; the only hydrocarbon remaining in the raw gas is methane.

Methane formation is about ten times as high as indicated by Equation (8 a); this is probably due to the slowness of the reaction according to Equation (11 a). As expected, increasing pressure promotes methane formation; at 1400 °C and 3.0 MPa, the methane content of the product gas is ca. 0.3 vol %, whereas it is 0.55 vol % at 6.0 MPa. Methane content is also influenced by changes in reaction temperature and addition of larger or smaller amounts of steam.

Under conditions prevailing in the reaction zone, no free carbon should be present, according to either the Boudouard equilibrium

$$C + CO_2 \rightleftharpoons 2\,CO \tag{10}$$

or the reaction

$$C + H_2O \rightleftharpoons H_2 + CO \tag{6}$$

Soot can be present in the raw gas only at below 1200 °C and around 5.0 MPa. Whereas the amount of soot produced during partial oxidation of methane is practically zero, in heavy-oil gasification about 0.5 – 2.5 mass % of feedstock is contained in the raw gas as free carbon. Some authors assume that this soot formation is due solely to thermal cracking of a proportion of the hydrocarbons; others attribute it partially to the Boudouard limit being reached during cooling. Ash particles in the heavy oils also seem to act as nuclei of condensation or catalysts for soot formation [41, 42].

In heavy-oil gasification, soot formation can be nearly eliminated by increasing the reaction temperature, at the cost of a considerable increase in oxygen consumption and a corresponding increase in carbon dioxide content at the expense of carbon monoxide.

Sulfur components of the gas mixture are largely hydrogenated according to Equation (16). Under normal reaction conditions, ca. 95 % of the sulfur is converted to hydrogen sulfide and the remaining 5 % to carbonyl sulfide. No sulfur di- or trioxide is detectable in the raw gas from partial oxidation.

3.2.2. Types of Processes

Three processes are commercially established for the production of gas mixtures (primary components: H_2 and CO) from gaseous or liquid hydrocarbons by partial oxidation: the Texaco, the

Shell, and the Lurgi process. The main steps of all three processes are now practically identical.

The original differences between the processes were that Texaco used natural gas as the sole feedstock and changed to liquid hydrocarbons only after the process of steam reforming of natural gas became available, whereas Shell used heavy oils from the beginning. Lurgi's process in contrast was developed for lignite tar oils and slurries and covers now the broadest range of feedstocks.

What still distinguishes the processes today are the methods of heat recovery and soot management. While all three use boiler configurations for high-grade heat recovery, only Texaco and Lurgi have available quench configurations for heavy-duty services. On the other hand, all three can offer some kind of soot recirculation system while only Lurgi and Shell have built more economic once-through systems, which are also environmentally favorable.

Over 320 plants operating on these three processes have been built to date.

Texaco Process. Originally, the Texaco process did not have a waste-heat recovery boiler (Fig. 15). Feedstocks are preheated as in the other two processes. Fired heaters are often preferred to steam-heated ones for heavy oil as well, probably by analogy with natural gas cracking practice. Besides partial oxidation of natural gas and residue oil, this process has been developed further for partial oxidation of coal slurries (for details, see Chap. 4).

Process steam is used to atomize the heavy oil in the reactor, so the oil must be pressurized to only a few bars above reactor pressure.

Oxygen is supplied to the reactor through a central tube; the oil – steam mixture is fed coaxially through an annular gap. The gas mixture leaving the reaction zone is cooled to 250 – 300 °C by quenching with hot water in a quenching section of the reactor. Most of the soot and ash is removed from the gas at the same time, whereas the water vapor content increases considerably due to evaporation of part of the quench water. The water vapor content of the gas is sufficient for conversion of the carbon monoxide to a residual content of ca. 1.8 – 2.5 vol % in a downstream high-temperature conversion unit.

After leaving the quenching tank, the gas is scrubbed again in a special venturi tube and a soot scrubber (a packed column), which remove the soot to a residual content of 1 – 2 mg/m^3. If the gas is to be fed to a high-temperature conversion unit, the soot is scrubbed out with hot water.

Shell Process. The basic flow sheet of this process resembles the one presented in Figure 16. The processes differ in the burner technology, the soot removal and the soot/ash management. The Shell process was right from the beginning equipped with a waste-heat boiler and is still available in this configuration only.

By means of reciprocating or centrifugal pumps the heavy oil feedstock (as single stream) is fed through a preheater to the reactor. The so-called co-annular burner is of the steam-atomizing type: oil, steam – oxygen mixture and steam are routed through concentric, alternating channels (slits) whose number determines the capacity. Oil viscosities at the burner of up to 300 mm^2/s are manageable. The burner is cooled by a pressurized circulating cooling water system.

The refractory-lined reactor and the waste-heat recovery boiler are similar to those of the other processes. Soot is removed from the raw gas in two stages: most of it is washed out with water in a quench tube [instead of the venturi scrubber d) of Figure 15] and collected in a soot separator. In the scrubber, a packed column, the remaining soot is removed to a residual content of < 1 mg/m^3. Preferably soot-free water from the soot recovery unit is supplied to the top of the scrubber, and soot-laden water from the scrubber outlet is used for the quench tube. By cooling the water fed to the scrubber, the gas is cooled to slightly above ambient temperature. The soot slurry is treated in a once-through soot/ash recovery and management system as described below.

Lurgi Process. This process was originally developed for the gasification of tars, oils, and slurries formed as byproducts of fixed bed gasification. This lends to an inherent robustness (as in slagging service) and flexibility (digesting unmixable feeds) which the refinery-derived processes are lacking. The Lurgi process is available in both, quench and boiler configurations.

Figure 16 shows the boiler configuration of the Lurgi process with metals ash recovery system. The feedstock is fed through preheaters

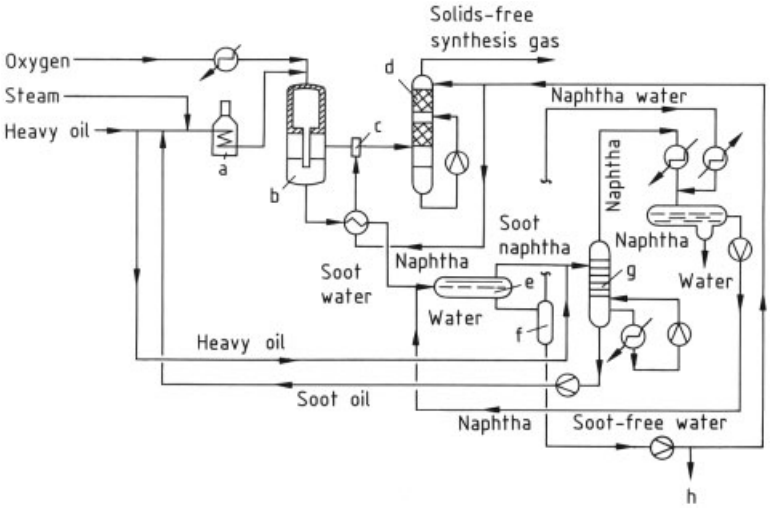

Figure 15. Texaco gasification
a) Preheater; b) Reactor; c) Venturi scrubbing section; d) Soot scrubber; e) Decanter; f) Flash distillation; g) Naphtha column; h) Water clarification

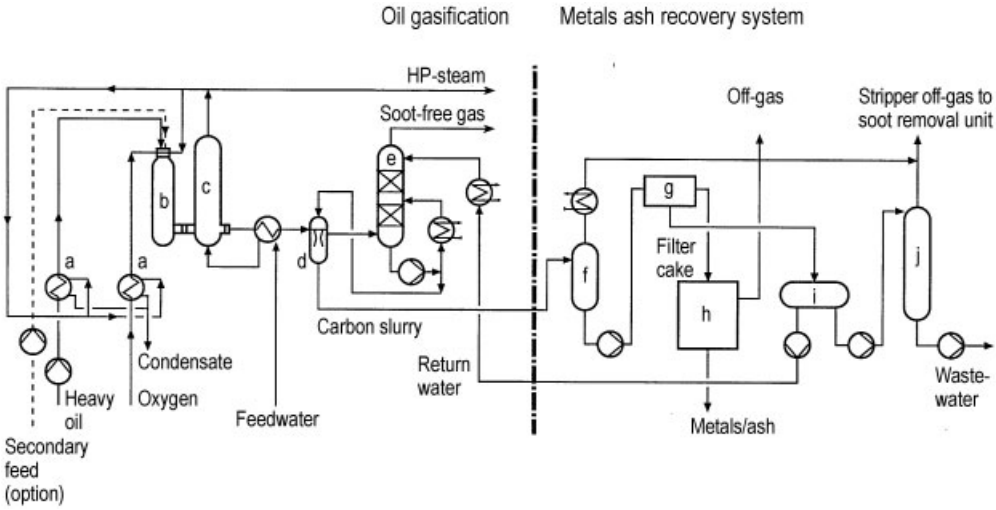

Figure 16. Lurgi oil gasification (boiler configuration) with metals ash recovery system (MARS)
a) Preheater; b) Reactor; c) Waste heat boiler; d) Venturi scrubber; e) Fine wash scrubber; f) Slurry tank; g) Filter; h) Multiple hearth furnace; i) Return water vessel; j) Wastewater stripper

(a) to the reactor (b) by centrifugal pumps (or reciprocating pumps for smaller capacities). A pressure of only about 0.5 MPa above reactor pressure is needed for operation of the steam-atomizing burner. Preheating of feedstock and oxygen is mainly done for optimization of the process heat balance, i.e., to save on the costs of heavy oil and oxygen otherwise consumed for internal heating. The lowering of feedstock viscosities is an additional — but welcome — effect. The burner itself works well even with viscosities up to 500 mm^2/s, meaning that every pumpable feedstock can be gasified. This top mounted multi-gun burner has the capability of handling different, unmixable fluids, e.g., emulsions and slurries in addition to heavy oils. The broken feed line represents such an optional secondary feed.

The oxidant is preheated as well and mixed with steam as moderator prior to being fed to the

burner. The burner and reactor are tuned by fluid-dynamic simulation to achieve mixing of this mixture with the feedstock within the smallest possible volume. Thus, the reactor space is optimally used for gasification to completion. The burner is water-cooled using a dedicated circulating cooling water system at a pressure above that prevailing in the reactor. The entire reactor is refractory-lined. In the nonslagging service associated with boiler operation a refractory material with high alumina content has achieved the longest lifetimes.

The raw gas leaves the reactor at about 1350 °C and passes directly into the heat recovery boiler (c) where 10 – 14 MPa steam is formed while cooling the gas to a few degrees above saturated steam temperature. This boiler is specifically designed for the high gas inlet temperatures and the particulates-charged gas at high velocities. A small part of the steam thus generated is used for feedstock and oxidant preheating, the bulk is superheated for use in steam turbine drives or in a combined cycle power plant. The rest of the heat is recovered in a boiler feed water economizer. Commonly, most of the heat content of the gas below the dew point is used to heat up make-up water or turbine condensate in a water preheater after the venturi scrubber.

The raw gas from the reactor contains a certain amount of free carbon (soot). The plant is normally designed for a soot content in the gas equivalent to 0.5 wt % of hydrocarbon reactor feedstock. The soot in the gas is removed in a two-stage water wash, which consists of a venturi scrubber (d) and the fine wash scrubber (e). The gas leaves the scrubber at 40 °C at a residual soot content < 1 mg/m^3. The soot slurry, which has an ash-soot content of 1 wt % carbon, is routed to the metals ash recovery system (MARS, see below).

The Lurgi oil gasification is also available in the *quench configuration* where the hot reaction gas is brought down rapidly to saturation temperature by direct water quench. This quench can be selected for slagging service as in the original application (coal – oil slurries, wastes) or for the less severe high-ash, high-salt residue service. In slagging service the reactor is lined with a special high-chromium refractory which has proven its resistance in many years of operation.

Like in the Shell process, the gas mixture produced contains approximately equal proportions of carbon monoxide and hydrogen, ca. 5 % carbon dioxide, and small amounts of methane, nitrogen, argon, and hydrogen sulfide. Although not equally well suited to all syntheses (meaning those that require CO and H_2, as well as those that require one of the two, either alone, or either mixed with a third gas such as N_2), its composition is nevertheless neutral for many uses. Methanol synthesis gas or hydrogen can be produced by converting part or all of the carbon monoxide. If the necessary amount of nitrogen is added, ammonia synthesis gas can be produced. If the hydrogen – carbon monoxide ratio remains unchanged, the gas is suitable, after purification, for oxo synthesis and for production of maximum quantities of carbon monoxide by low-temperature separation or absorption.

Soot/Ash Recovery and Management. In the Texaco process, soot is extracted from the soot water with naphtha in a closed system and recovered for recirculation to the reactor. Shell used an open system based on pelletizing in its original process but was subsequently obliged to change to a closed system also based on extraction with naphtha.

Due to difficulties with increased viscosities of very heavy feedstocks, which makes carbon recycle impossible, and due to poorer economics of the recycle processes, once-through systems were developed by Lurgi and Shell. These systems basically incinerate the soot after filtration and recover the heavy metals as oxides.

Texaco Method (Fig. 15). Soot water is mixed with naphtha, the soot and some of the ash passing into the naphtha phase. After the two phases are separated, practically soot-free water from the decanter is depressurized in a flash tank, to drive off dissolved naphtha. Naphtha is recovered by condensation and recirculated to the process; the same is true for naphtha-free water.

The naphtha – soot mixture from the decanter is mixed with heavy oil, and naphtha is distilled off overhead in a distillation column. Oil collects at the bottom of the distillation column together with some ash and a residual amount of naphtha. The oil is mixed with more heavy oil and fed to the reactor via the oil heater.

Lurgi Metals Ash Recovery System (MARS, Fig. 16). The problem of uneconomic carbon recycle was solved by Lurgi through application of its vanadium roasting technology. A proprietary multiple hearth furnace is used for controlled combustion of the soot under conditions where the vanadium oxides neither melt nor corrode.

The soot slurry from the gasification (in both, quench and boiler configurations) is flashed to atmospheric pressure in a slurry tank (f). The slurry is filtered (g) leaving a filter cake with about 80 wt % residual moisture and a clear water filtrate, which is partly recycled as quench or scrubbing water. The filter cake is subjected to a controlled incineration process in a multiple hearth furnace (h). The product contains typically about 75 wt % vanadium pentoxide (plus the nickel and iron from the crude), which can be sold to metal reclaimers. The process is autothermal, the heat of combustion being sufficient to evaporate the moisture content of the filter cake. Excess water is stripped of dissolved gases in a wastewater stripper (j) and pumped to final treatment.

Shell Method. Shell uses a similar system of filtration and incineration. The main distinctions to the Lurgi process as described above are in the selection of key equipment like furnace and filters.

Selection Criteria for Quench Configuration Versus Heat Recovery Configuration. In addition to process-specific marketing strategies there are technical and economic criteria for the selection between quench and boiler configurations (see Table 12). These must be studied carefully for every application of the Texaco and Lurgi processes that alone offer alternatives.

Generally speaking, the heat recovery configuration can be selected when the feedstock contains low amounts of precipitation-prone contaminants, which could block the boiler tubes. Most refinery heavy residues fall in this category. In this case it is possible to run the gasification at highest efficiency by generating considerable amounts of high-pressure steam, which also yields highest efficiency in a possible downstream integrated gasification combined cycle (IGCC) power plant. The additional investment in the boiler is recovered quickly.

Table 12. Comparison of quench and heat recovery configuration

	Quench configuration	Heat recovery configuration
Feedstocks		
Gas, residues, wastes, slurries (sludges, coal, coke), extreme ash and/or salt contents	highest flexibility	limited by possible salt precipitation
Product range:		
$H_2 + CO$, H_2, CO	fastest (cheapest) route to H_2 and NH_3	highest flexibility
Energy utilization:		
Steam production	MP * steam available trade off efficiency vs. cost	HP ** steam, heat recovery at highest temperature, highest efficiency for IGCC possible
Investment cost:	lowest cost for gasification unit	HP ** waste-heat boiler (high efficiency) at extra cost

* MP = medium pressure.
** HP = high pressure.

On the other hand, should the feedstock contamination be above the waste heat boiler's limits, the quench configuration has to be chosen. It also may be selected when such contaminated feeds are expected in later operation of the refinery or, when the customer wants to keep open all options for feedstocks like coke/coal slurries and wastes. The quench configuration thus offers highest (potential) flexibility at lowest investment cost traded off against a loss in energy efficiency.

The product side also influences the selection, e.g., for hydrogen or ammonia as product it is more economical to route the quenched, i.e., hot saturated gas directly into a raw gas shift conversion unit (see Section 5.1) instead of cooling it down and desulfurizing it before the shift. This is because the water quench raises all the steam necessary for the CO shift reaction directly within the reactor effluent. If a syngas rich in CO is required, the heat recovery configuration is the appropriate choice.

3.2.3. Influencing Raw Gas Composition

Different types of gas are produced not so much by changing operating conditions (e.g., pressure and temperature) as by changing the proportions of reactants, adding carbon dioxide, and employing further process steps immediately downstream of actual gasification. Commercially produced gases can be classified in

three large groups: synthesis, fuel, and reduction gases.

An overview of gas yields and compositions in relation to the feedstock and gasifying media, together with other important data, is presented in Table 13.

Synthesis Gas. Synthesis gas must nearly always be low in undesirable gas components such as methane, nitrogen, and argon, and high in either carbon monoxide or hydrogen, or both. Medium to high pressure is desirable as well. Whereas the nitrogen and argon contents of unpurified synthesis gas depend almost exclusively on the quality of air separation (nitrogen components of heavy oils generally cause a rise of less than 0.1 % in the nitrogen content of the gas), the methane content is largely a function of gasification pressure (see Section 3.2.1).

The proportions of gasification steam, oxygen, and heavy oil are the same for almost all synthesis gases, regardless of whether hydrogen- or carbon monoxide-rich gases or gases with a defined ratio of hydrogen to carbon monoxide are produced. For the last category, which includes primarily synthesis gases for the production of oxo alcohols, carbon dioxide must be recycled to the reaction zone to increase the carbon monoxide content of the synthesis gas and obtain the correct ratio of hydrogen to carbon monoxide (which can range from 1: 1 to 1: 1.5 for certain oxo processes).

To manufacture synthesis gases in which no (or only partial) conversion of carbon monoxide to hydrogen and carbon dioxide with steam occurs, use of the waste-heat boiler to cool the hot reformed gas is preferred to quenching with water or steam. Examples are gases for carbon monoxide production, some oxo processes, and methanol synthesis.

A special case is *partial conversion* of carbon monoxide immediately following the gasification reaction. This can be brought about by quenching with defined quantities of water or steam, either in the bottom section of the gasification reactor or in a quench tube located between the reactor and the waste-heat boiler.

Here, best results are achieved by using steam and cooling the gas to ca. 1000 °C. At lower temperature, the water gas equilibrium, (Eq. 7) cannot be reached without a catalyst. A very flexible and economic alternative is the use of a direct quench reactor in parallel to a reactor/waste-heat boiler train.

If, on the other hand, production of hydrogen-rich synthesis gas is desired, *direct quenching* of the hot synthesis gas, followed by catalyst-based shift conversion is the preferred route. An example is the production of ammonia from heavy oil in countries with cheap coal reserves and little oil of their own, (e.g., India), which are able to supply coal-based energy at low cost (whether directly from steam or indirectly via electric power).

Oxygen with a purity of $> 95\,\%$ is used for gasification. In fact, the oxygen content is usually $> 99\,\%$, because the cost of achieving the higher purity is always less than the extra cost of gas purification and synthesis (e.g., by additional consumption of synthesis gas) to compensate for hydrogen or carbon monoxide losses from increased discharge of purge gas.

The possibility of manufacturing ammonia synthesis gas from heavy oil by gasification with enriched air (i.e., with a $N_2 - O_2$ mixture containing $> 21\,\%\ O_2$) has also been proposed but not exploited commercially. The higher costs of purifying and conditioning the gas that result from the increased volume of synthesis gas because of its nitrogen content outweigh the savings in gasifying medium. (Enriched air can be produced from air by mixing with pure oxygen or by adsorptive depletion of nitrogen.)

Fuel Gas. In contrast to synthesis gas production, a low content of inerts is not particularly important for the manufacture of fuel gas. Only the required heating value determines the gasifying medium, pure oxygen or air. Combustion of larger volumes of gas instead of gas with an increased heating value is often justifiable economically. However, not only the heating value but also the density and thus the Wobbe index (see Section 8.2.1.3) of the fuel gas are almost always stipulated for reasons of interchangeability with fuel gases from different sources. When fuel gases containing carbon monoxide are used in the private sector, it is mandatory to minimize their toxicity by limiting the carbon monoxide content, generally to between 1 and 3 vol %.

Hot Carburetion. Fuel gas specifications for the private sector require heating values ranging from 17 500 (town gas) to 34 000 kJ/m^3

Table 13. Typical results of partial oxidation (basis: Shell gasification)

Type of gas:	Untreated raw gas		Synthesis gas, (H_2-rich)	Fuel gas, carbureted	Fuel gas, low-calorific	Reduction gas
Feedstock:	Bunker-C fuel (without soot recirculation)	Vacuum residue (with soot recirculation)	Bunker-C fuel (with soot recirculation)			
Operating pressure, MPa	6	6	6	6	1.5	35
Preheating temperature, °C						
Heavy oil – natural gas	245	280	245	245	245	245
O_2– air	245	260	245	245	400	245
Steam	310	360	310	310	310	310
CO_2						245
Feedwater	160	105	160	160	160	160
Feed quantities						
Feedstock, pure kg	100	100	100	100	100	100
O_2 (98 – 99.5 %), m^3	77.6	70.7	74.1	73.7		81.2
Air, m^3					450	
Steam, kg	45	50	45	45		30
Quenching steam – water, kg			100 *	60 **		
CO_2, m^3						50
Light naphtha, kg				65		
Product gas volume, m^3	303	290	327	354	602	305
Gas composition, vol %						
CO_2	4.32	4.63	13.34	13.32	2.36	7.68
CO	46.55	48.92	33.29	31.02	22.34	52.29
H_2	47.15	44.94	51.49	26.68	16.71	38.03
CH_4	0.60	0.3	0.60	27.79	0.03	0.60
N_2+ Ar	0.56	0.2	0.54	0.49	57.70	0.60
H_2S + COS	0.81	1.01	0.73	0.70	0.16	0.80
Lower heating value (LHV), kJ/m^3	11 365		10 105	16 897	4670	11 077
Steam output, kg	275	258	301	252	580	320
Steam pressure, MPa	10.5	8	10.5	10.5	10.5	10.5
Steam exported, kg	230	213	118	207	540	290

* Steam.
** Liquid.

(natural gas). However, gasification of heavy oil produces heating values of only 4000 – 11 000 kJ/m^3, depending on whether air or oxygen is used as the gasifying medium. Thus, the heating value must be improved. This can be achieved — after appropriate gas conditioning — either by methanation without addition of other fuels or by hot carburetion with addition of liquid hydrocarbons having a low final boiling point. Methanation is necessary if heating values > 20 000 kJ/m^3 are required. For the production of town gas with heating values between 16 500 and 19 000 kJ/m^3, however, hot carburetion is more beneficial from the heat economy aspect, because it operates with practically no addition or subtraction of heat. This is in contrast to the strongly exothermic methanation process, in which about 10 500 kJ (i.e., ca. one-fifth the heating value of the feedstock) must be dissipated per cubic meter of methane formed.

Due to the widespread availability of natural gas — in some places as LNG — carburetion has lost most of its significance as a commercial process.

As can be calculated with the aid of Table 13, the thermal efficiency (i.e., ratio of the heating value of the product gas to the heating value of the fuel) is 81 % for normal gasification and increases to 89 % with hot carburetion, at the expense of reduced generation of steam from waste heat.

Fuel Gas for IGCC Power Plants. A special area of fuel gas production is for gas turbines in power plants. Due to increasingly stringent air pollution control regulations, operating coal- or oil-fired steam boilers without flue gas desulfurization systems is impossible. Because cleaning large gas volumes at slightly above atmospheric pressure is very expensive, attempts have been made to find a more economic solution by

applying fuel gasification with subsequent gas cleaning under pressure. As shown in Figure 17, partial oxidation of heavy oil can be combined beneficially with a gas turbine-driven generator, not only from the point of view of gas cleaning but also with regard to generation of high-pressure steam. For the arrangement illustrated, an efficiency well above 46 % is achieved, 2 – 4 % better than for a conventional steam boiler power plant [43].

IGCC Power Plants (Fig. 17). A heavy refinery residue is gasified with oxygen from an air separation unit (i) in a Lurgi oil gasification plant (a) at 6.5 MPa, producing a lean gas with a heating value of 11 000 kJ/m^3. Oxygen and residue are preheated to 280 °C. Cooling in the waste-heat recovery boiler produces high-pressure steam at 13.0 MPa. After the soot is removed as described above and the gas is cooled, carbonyl and hydrogen sulfides and all other trace contaminants are removed to 20 ppm in one step in a Lurgi Rectisol unit (b) (see Section 5.4.2.1). The desulfurized gas passes through a gas expander (f) generating electric power and refrigeration to cool the methanol within Rectisol. This *dual use expansion* adds significantly to the overall efficiency of power generation.

The low-calorific gas is burned in a gas turbine (d). For NO_x-control the gas is further diluted by saturation with water and — optionally — with nitrogen from the air separation unit. Waste heat from the gas turbine off-gas and the saturated steam from the gasification are used in the waste-heat steam generator (h) to produce additional steam and to superheat steam. The energy of the superheated steam is used in a condensing steam turbine (g) that drives a generator (e). For maximum efficiency the steam system will be designed for three levels with one or two reheat stages.

The gas turbine is also coupled to an air compressor in which combustion air for the gas turbine is compressed. A part of this compressed hot air can be routed through a cooler – saturator system producing steam for fuel gas saturation before entering a medium pressure air separation saving the separate air compressor. The nitrogen from the air separation in turn can be recycled back to the gas turbine for NO_x-control and to maintain the mass flow of the turbine. Air and nitrogen integration as described yields higher efficiency than a nonintegrated system with independent air separation unit (ASU) compressor, but involves increased complexity of the plant. Therefore different designs with varying degrees of integration exist to accommodate for all requirements. Many versions of this IGCC process exist, some of which export steam or synthesis gas together with electrical power. Coproduction of this type results in better overall economics for a plant.

A good example for this concept is a reference plant for the Lurgi oil gasification (SVZ, Schwarze Pumpe, Germany) where the oil gasifier together with coal-waste gasifiers produces a synthesis gas which is fed to a combined cycle and a methanol plant [44]. Another example is a large refinery in Pernis (The Netherlands) where a Shell gasification feeds two thirds of its capacity to a hydrogen production plant and one third to an IGCC power plant [45].

Reducing Gas. Gas produced from heavy oil is also used in ore reduction. The carbon dioxide and water vapor content or reducing gases must be as low as possible. Water vapor can be removed easily from the gas by condensation, which presents no problem if the gas has been freed of soot, sulfur, and carbon dioxide. However, for reduction of less sensitive ores, uncooled or only partly cooled, soot-laden gas manufactured from low-sulfur oils is used whenever possible. Thus, water vapor content may be an important element of the gas composition. The quality of reducing gases is measured by the degree of oxidation η, which is defined as follows:

$$\eta_0 = \frac{(c_{H_2O}+c_{CO_2})\cdot 100}{c_{H_2O}+c_{CO_2}+c_{H_2}+c_{CO}}\ (\%)$$

Degrees of oxidation $\leq 5\,\%$ are the goal (see Table 13).

Reducing gases, which are always produced directly alongside the ore reduction furnace, differ from synthesis and fuel gases in pressure and temperature. Synthesis and fuel gases must be maintained at high pressure and below ambient temperature: for synthesis gases, with regard to gas purification and compression; for fuel gases, with regard to gas purification, distribution, and storage. For admitting reducing gas to the shaft furnace, a slight overpressure is sufficient; on the other hand, a temperature up to 1000 °C is desirable. The pressure range for manufacture

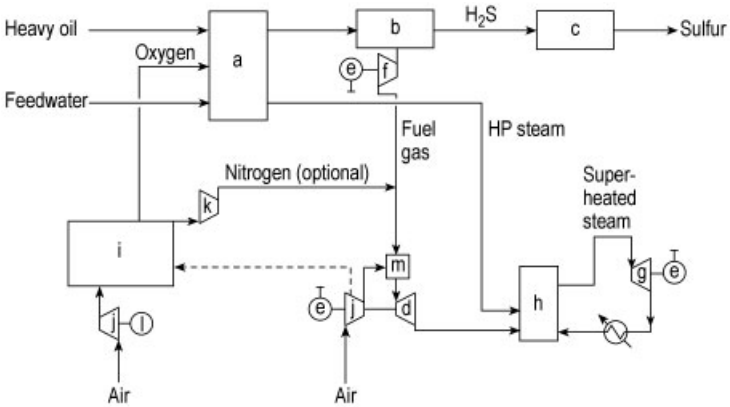

Figure 17. Lurgi oil gasification for combined cycle power plants
a) Oil gasification; b) Gas purification; c) Sulfur recovery; d) Gas turbine; e) Generator; f) Gas expander; g) Steam turbine; h) Heat recovery steam generator; i) Air separation; j) Air compressor (compressed air for air separation may be extracted from air compressor for gas turbine); k) Nitrogen compressor (optional); l) Motor; m) Combustion chamber

of reducing gas is currently 0.3 – 0.5 MPa, but efforts are in progress to achieve ore reduction at ca. 2 MPa.

3.2.4. Submerged Flame Process

The BASF submerged flame process is also a partial oxidation process; it was developed to produce acetylene (→ Acetylene, Chap. 4.2.3) and ethylene from natural gas and crude oil. Its distinctive feature is maintenance of a flame beneath the surface of the oil sump, to ensure high temperature and short residence time. Hot gases produced by partial combustion and cracking, which have a temperature of ca. 1500 °C, are quenched almost instantly to 250 °C in the hot oil sump. Thus, the thermodynamic equilibrium for the high temperature is "frozen". As in other partial oxidation processes, soot is formed, which collects in the oil sump and is recirculated to the burner. After condensable cracking products are removed, the cracked gas, which still contains a small amount of soot, is scrubbed with condensate. The condensate is then recirculated to the submerged burner.

The process operates between 0.2 and 1.0 MPa and yields a cracked gas with the following approximate composition, in percent by volume, which can be processed into synthesis gas:

Carbon dioxide	7 – 9	Ethylene	6 – 7
Carbon monoxide	40 – 45	Propylene	1 – 2
Hydrogen	23 – 30	Higher olefins and acetylenes	0 – 5
Methane	4 – 5		

From the residual gas, 3.5 t of ammonia, for instance, can be produced per tonne of acetylene.

3.3. Hydrogenating Gasification

Hydrogenating gasification, like partial oxidation, is a continuous process.

Principles. In contrast to partial oxidation, hydrogenating gasification is not based on combustion of part of the feedstock; instead, a hydrogenation reaction which proceeds exothermically overall takes place with addition of hydrogen. Liquid hydrocarbons are cracked primarily to form methane and ethane. This process, therefore, is beneficial for the production of *fuel gases* with a high heating value: on the other hand, direct manufacture of carbon monoxide- and hydrogen-rich synthesis gas is not possible [46].

Gas Recycle Hydrogenator (GRH) Process. The British Gas Corp. began to develop the GRH process in the early 1960s [47]. This enables the conversion of liquid hydrocarbons with a final boiling point up to ca. 325 °C to a gas mixture consisting mainly of hydrogen, methane, and ethane under controlled temperature and pressure in a shaft reactor, termed the

hydrogenator. Hydrogenating gasification produces no soot. The reaction takes place at 1.8 – 7.0 MPa and 350 – 500 °C at the inlet or ca. 750 °C at the outlet, depending on the boiling range of the feedstock. The latter must contain a certain amount of sulfur (about 10 ppm), which acts as a poison on reactor components made of nickel alloy, to prevent them from acting as catalysts that might lead to soot formation from the hydrocarbons. After desulfurization, aromatics removal and methanation, an SNG may be produced with a higher heating value (HHV) of 39 100 kJ/m^3, and a pressure of 6.0 MPa. The thermal efficiency of the process is 85 %.

Fluidized-Bed Hydrogenation (FBH) Process. The FBH process, also developed by the British Gas Corp., enables gasification of crude and extremely heavy oils. Feed oil and hydrogen are superheated and fed into the FBH reactor, in which hydrogenation of the hydrocarbons takes place in a fluidized bed formed from coke between 700 and 820 °C. The most important process variables are temperature, residence time, and oil: hydrogen ratio. Part of the hydrogen is supplied to the reactor through a grate from below to fluidize the coke.

Most of the hydrogen is mixed with the oil and sprayed from below into a central tube. By this means, rapid circulation is achieved in the reactor. The oil evaporates at the outlet of the spray nozzle. Any oil that does not evaporate adheres immediately to the flowing coke particles and is entrained with them. If the raw material contains residue-forming components, coke particles are also formed. Therefore, a small amount of coke must continually be withdrawn from the reactor.

Gas leaving the reactor is quenched with light aromatics, and the condensate is removed from the rich gas. Most of the sulfur in the oil is converted to hydrogen sulfide. A gas produced from Algerian crude with the addition of ca. 6.5 kg of hydrogen per kilogram of oil has a HHV of 39 400 kJ/m^3 at a pressure of 5.0 MPa. For SNG production, gas from the FBH plant must undergo a second hydrogenation step to reduce its hydrogen content [48]. More details on both processes, GRH and FBH, may be found in [49].

4. Gas Production from Coal, Wood, and Other Solid Feedstocks

This chapter deals with the production of gases from solid carbonaceous materials such as *coal*, *peat*, *wood*, *solid wastes* and solid residues from petroleum upgrading (*delayed coke*, etc.). The gasification process occurs via pyrolysis reactions (thermal decomposition) and subsequent heterogeneous (gas – solid) reactions of the solid residue from pyrolysis with reactive gases (O_2, steam, CO_2, H_2).

Besides carbon, solid feedstocks contain other elements such as hydrogen, oxygen, nitrogen, sulfur, heavy metals, and trace elements in varying amounts. The feedstocks are converted to a gaseous product suitable for use either as a source of energy or as a raw material for the synthesis of chemicals and liquid or gaseous fuels. The gaseous product can be handled with maximum convenience and purified easily from compounds that would cause pollution problems. Thus, gasification is an upgrading process that greatly extends the uses of solid fuels.

During their long history, the science and technology of solid feedstock gasification have advanced in various stages induced by external factors. At first directed toward the production of fuel or town gas for lighting and heating, the interest in gasification shifted in the first half of this century to synthesis gas for the production of chemicals and liquid motor fuels.

After a period of increasingly intense competition from oil and natural gas, gas production from solid feedstocks has become less important since the 1940s. In the late 1960s and early 1970s, some industrialized countries began to expect natural gas shortages which, together with oil price increases, changed the attitude toward utilization of solid fuels, particularly coal available in many parts of the world. More and more, the finite nature of fossil fuel reserves, the need for conservation, and good resource management became accepted and the significance of the world's coal reserves was recognized. Interest in *coal* utilization increased, primarily in the fields of gasification and liquefaction (→ Coal Liquefaction), and large sums of money became available from governments for the research, development, and demonstration of new processes and the improvement of existing ones. More recently, since large resources of natural

gas have become available, the interest in coal gasification has decreased considerably.

Petroleum-derived residues have gained some attention as feedstocks for gasification, with the increasing yield of residues in petroleum refining because of decreasing oil quality and the need for their economic and environmentally acceptable utilization, for example, for H_2 production. Use of *wood* as a feedstock has played a continuous role in small-scale gasification units (e.g., for the production of gaseous fuels) and more recently gained interest in the context of biomass energy applications.

Advances in gasification have been made possible by progress in other fields. Fundamental research on the mechanism and kinetics of chemical reactions, mass and heat transfer, fluid dynamics, and mathematical modeling has contributed to a better understanding of the phenomena occurring in gasification processes long after their successful operation started.

This chapter introduces fundamental aspects and process technologies of gas production from solid feedstocks. Given the variety of new developments in the past decades, only a selection can be presented here to help identify and discuss basic principles, state of the art, and future trends. For more details, the reader is referred to review articles on gasification fundamentals [50 – 52], gasification processes [53 – 59], or both [60, 61], and to numerous publications on this subject.

4.1. Fundamentals

Gasification of solid feedstocks involves the complex interaction of chemical reactions and physical processes such as heat and mass transfer.

Figure 18 shows a scheme of the main chemical reactions that occur when carbonaceous materials are exposed to reactive gases at high temperature. The yield and composition of the product are determined by the combined effects of feedstock properties and reaction conditions on the thermodynamics and kinetics of individual reactions. For example, the yields of liquid byproducts (e.g., tar, phenols) depend on the reaction conditions that prevail during pyrolysis and secondary reactions of the volatile products (e.g., temperature, gas-phase residence time).

4.1.1. Thermodynamics of Chemical Reactions

Some reactions of the scheme in Figure 18 are endothermic, in particular the gasification reactions of char with steam or carbon dioxide, which are significant in most gasification processes. (For heat effects of individual reactions, see Chap. 1)

The resulting heat requirement is met either by reactions with oxygen (in autothermic processes) or by an external heat source (allothermic process). Minimization of overall heat requirements has been one incentive among others for improving process efficiencies and economics.

For example, the idea of applying reaction conditions that favor the overall reaction

$$2\,C + 2\,H_2O \rightarrow CH_4 + CO_2$$

which is nearly thermoneutral, has affected gasification process development efforts during the times when production of CH_4 from coal appeared attractive.

In some of the chemical reactions of Figure 18, conversion is limited by equilibrium constraints under typical gasification conditions (temperature, pressure). Most significant in this respect are the reactions of char with steam, carbon dioxide, and hydrogen. Equilibrium constants for these and other reactions, as well as the effect of temperature, are presented in Chapter 1.

Thermodynamic analysis may aid in examining the limits for conversion and product gas composition as affected by such process variables as temperature, pressure, and stoichiometry. For example, the effect of temperature and pressure on equilibrium gas composition for individual char – gas reactions with steam, carbon dioxide and hydrogen is shown in Figure 19. The curves reflect two general equilibrium principles: (1) higher temperature favors products of endothermic reactions and (2) higher pressure shifts the equilibrium in the direction of lower amounts of gaseous components. Some past incentives of process development were based on thermodynamic analyses of gasification. For example, to produce methane-rich gases (substitute natural gas, SNG), high pressure and low temperature are advantageous under equilibrium

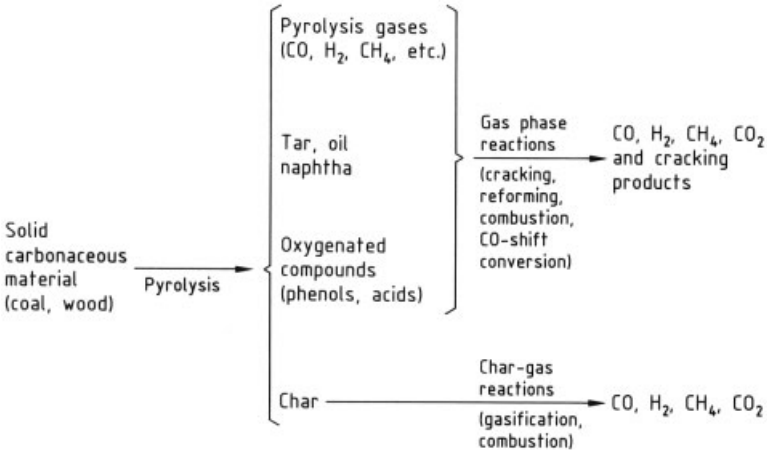

Figure 18. Main chemical reactions during gasification of solid carbonaceous materials

aspects. As a consequence, research and development efforts were directed toward applying higher pressure and enhancing feedstock reactivity (e.g., via catalysis), thereby lowering gasification temperature.

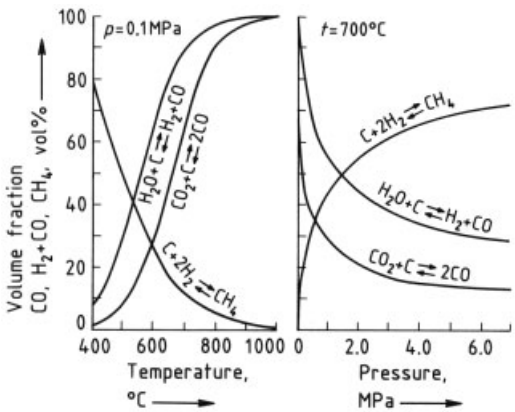

Figure 19. Effect of temperature and pressure on equilibrium gas composition in the presence of solid carbon

However, only the limits of conversion and product gas composition can be identified by thermodynamic analysis. In real gasification systems, equilibrium is generally not reached because of kinetic constraints.

4.1.2. Kinetics

Pyrolysis Reactions. Among the chemical reactions that occur during gasification, pyrolysis is generally fast. This is illustrated in Figure 20, which presents first-order rate constants of some reactions shown in Figure 18. Underlying values for activation energies and preexponential factors are taken from various literature sources (Table 14). Formation of tar, methane, and hydrogen as volatile products from Montana lignite is presented as part of pyrolysis.

The rates of primary pyrolysis product formation can be described by

$$\frac{\mathrm{d}V_i}{\mathrm{d}\,t} = k_{0i}\exp\left(-E_{\mathrm{A}i}/RT\right)\left(V_i^* - V_i\right) \tag{14}$$

where V_i is the total amount of product evolved from reaction *i* until time *t*, and V_i * is the value of V_i at $t \to \infty$.

Kinetic parameters for pyrolysis product formation have been determined for various kinds of coal and experimental conditions. Extensive compilations of pyrolysis data and kinetic information drawn from them can be found in review papers [51, 62].

Char – Gas Reactions. Of the char – gas reactions, combustion is generally fastest, and gasification is generally faster with steam than with carbon dioxide and hydrogen.

Table 14. Examples of first-order rate constants for reactions occurring during gasification (for nomenclature, see text)

		k_0, s^{-1}	E_A, J/mol	V *, wt %, (maf) [b]	Reference
Pyrolysis product formation	tar	2.0×10^{17}	315 281	3.5 (+ 2.9)	[63] (lignite)
	CH_4	4.7×10^{14}	290 578	1.1 (+ 0.4)	[63] (lignite)
	H_2	1.6×10^{18}	371 806	0.6	[63] (lignite)
Char – gas reactions	$C-O_2$	1.8×10^{6} [a]	113 050		[64] (bituminous coal)
	$C-H_2O$	4.1×10^{2} [a]	146 545		[65] (subbituminous coal), [66]
	$C-CO_2$	2.5×10^{2} [a]	146 545		[65] (subbituminous coal), [66]
	$C-H_2$	8.4×10^{-4} [a]	67 201		[65] (subbituminous coal), [66]
Tar cracking	coal tar	1.2×10^{9}	164 968		[67] (bituminous coal)
		8.0×10^{12}	272 992		[67] (bituminous coal)

[a] For $(p_i - p_i^{eq}) = 1.013\times10^5$ Pa (see Eq. 15).
[b] maf = moisture and ash free.

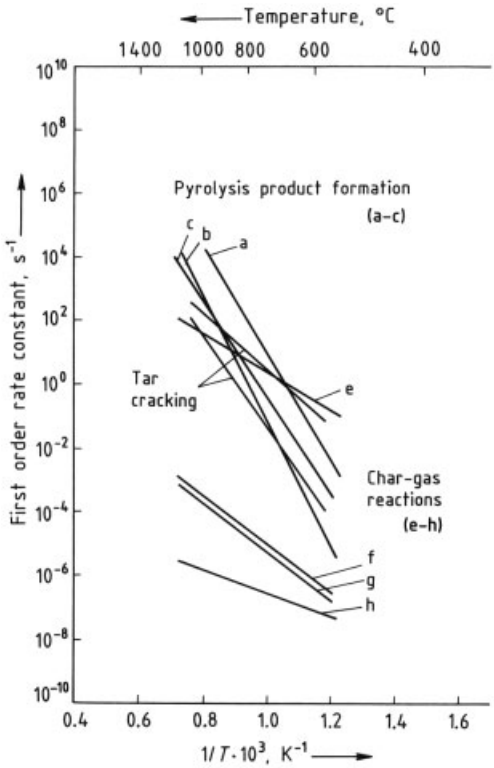

Figure 20. Examples of first order rate constants for reactions during gasification (pyrolysis product formation: Montana lignite [63], tar cracking: bituminous coal [67], char – gas reactions: bituminous and subbituminous coal [66]
Pyrolysis product formation: a) Tar; b) H_2; c) CH_4;Char – gas reaction: e) $C-O_2$; f) $C-H_2O$; g) $C-CO_2$; h) $C-H_2$

Although detailed kinetics include noninteger reaction orders and inhibiting effects of reaction products, simple first-order rate laws can be used for estimating and modeling purposes:

$$\frac{dm_c}{dt} = -k_0 \exp\left(-E_A/RT\right) m_c \left(p_i - p_i^{eq}\right) \quad (15)$$

where p_i is the partial pressure and p_i^{eq} the equilibrium partial pressure of reactant i, and m_c is the mass of solid carbon in the char. Values of k_0 and E_A for char – gas reactions vary with the kind of solid feedstock, according to its reactivity. Values of gasification rate constants for chars derived from various feedstocks vary with up to eight orders of magnitude. High transient reactivities can be observed under rapid heating conditions, (i.e., when pyrolysis and gasification of the resulting char occur simultaneously). Kinetic parameter values for char – gas reactions in Table 14 are taken from a modeling study of coal gasification in moving-bed reactors.

A characteristic of gas – solid reactions is the existence of a regime of temperature and pressure in which mass transfer affects overall kinetics. Extra- and intraparticle mass-transfer resistance can be incorporated in Equation (15) as follows:

$$\frac{dm_c}{dt} = -k_{eff} m_c \left(p_i - p_i^{eq}\right) \quad (16)$$

$$k_{eff} = \frac{1}{\frac{1}{k_g} + \frac{1}{\eta k_r}} \quad (17)$$

where k_g is an extraparticle (film) mass-transfer coefficient; k_r, a first-order chemical reaction rate constant; and η, an effectiveness factor that includes resistance to pore diffusion. The theoretical background can be found in reaction engineering textbooks (e.g., [68]; → Fluidized-Bed Reactors).

The temperature dependence of Equation (17) in its general form is shown in Figure 21. Accordingly, the chemical reaction is rate-controlling at lower temperature because it has the highest activation energy and extraparticle diffusion is limiting at high temperature.

In between ranks the intraparticle mass-transfer regime. Because these regimes may all play a significant role in technical gasification their limits must be defined by means of kinetic reactor analysis. The effects of feedstock reactivity, particle size, gas flow rate, and total pressure are discussed in [52].

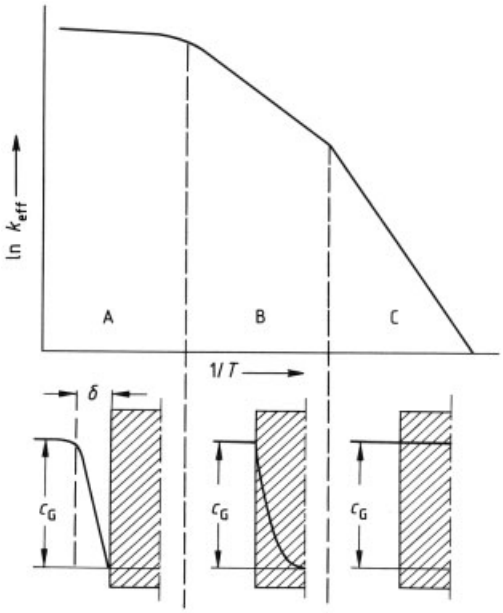

Figure 21. Temperature dependence of effective rate constant for gas – solid reactions with porous particles, schematic representation
A) Film-diffusion controlling (c_G gas concentration, δ film thickness)
B) Pore diffusion controlling
C) Chemical reaction rate controlling

Gas-Phase Reactions of Volatiles. Gas-phase reactions of volatile pyrolysis products include a variety of individual reactions. Among them are cracking, reforming, combustion of tarry species and hydrocarbons, dealkylation, dehydroxylation, and shift conversion of carbon monoxide.

The rates of these reactions depend, in a complex way, on the reactivity of the reacting species and on reaction conditions (partial pressures, temperature, catalytic effects).

Again, first-order laws are sufficient for estimating and modeling purposes:

$$\frac{\mathrm{d}\,c_i}{\mathrm{d}\,t} = -k_0 \exp\left(-E_A/RT\right) c_i \tag{18}$$

where c_i is the molar (or mass) concentration of species i in the gas phase.

Values of kinetic parameters for tar cracking given in Table 14 were determined with freshly formed tar from bituminous coal [67].

Equations (14) – (16) and (18) can be used for describing source or sink terms in the balance equations in reactor models (see Section 4.2.3).

4.2. Classification and General Characteristics of Gasification Processes

4.2.1. Criteria for Classification

Processes for gasification of solid carbonaceous materials may be classified in the following ways (the abbreviations given are used in subsequent tables that present information on individual gasification processes):

1) Method of contacting solid and gaseous reactants (type of gasification reactor)
   - moving fixed bed
   - fluidized bed
   - entrained flow
   - molten bath
2) Kind of solid feedstock
   - coal
   - wood – biomass
   - solid waste materials
   - petroleum-derived residues
3) Particle size of solid feedstock
   - fine
   - medium
   - coarse
4) Gasifying medium
   - air (air)
   - oxygen (ox)
   - steam (st)
   - steam – air (st/a)
   - steam – oxygen (st/ox)
   - hydrogen (hydr)
5) Method of supplying heat required by gasification reactions
   - internal heating
     - autothermic (I/A)
     - cyclic (I/C)
     - heat-carrying fluids or solids (I/HC)
   - external heating (E)

heat transferred through walls of reaction vessel or heat exchanger (allothermic)

6) Operating pressure
   atmospheric
   elevated
7) Condition of gasification residue to be removed
   dry ash in nonslagging operation
   slag in slagging operation
8) Application of product gas
   fuel gas (medium or low Btu) (FG)
   fuel gas for combined cycle power plant (FG/ IGCC)
   synthesis gas (SG)
   substitute natural gas (SNG)
   source of hydrogen (H_2)
   gas for direct reduction of iron (RG)
9) Developmental status
   commercially available process
   process in developmental stage
      bench scale
      pilot-plant scale
      demonstration scale

Obviously, these criteria are not independent, and others could be used instead. An example is a characteristic gasification temperature, which determines product gas yield and composition, utility consumption, and the approach for predicting product gas composition and process performance (see Section 4.2.3).

For the purpose of this chapter, processes are presented according to the method of bringing the reactants into contact. This appears to be the most appropriate way of identifying process principles and characteristics, and is the most widely applied and accepted practice. Figure 22 shows examples of reactor types that represent different methods of contacting reactants.

4.2.2. Criteria for Process Assessment

Besides economics and availability, efficiency and other process performance criteria characterize individual gasification processes and aid in their comparison and assessment. Some commonly used criteria for practical purposes are defined in the following material.

Efficiencies

$$\text{Cold gas efficiency} = \frac{\text{higher heating value product gas}}{\text{higher heating value solid feedstock}}$$

$$\text{Gasification efficiency} = \frac{\text{higher heating value} \left\{\begin{matrix}\text{product} + \\ \text{byproducts}\end{matrix}\right\} (+ \text{ recoverable product heat content})}{\text{higher heating value solid feedstock}}$$

(Lower heating values are often used instead of higher heating values.)

$$\text{Carbonconversion} = 1 - \frac{\text{carboningasificationresidue}}{\text{carboninsolidfeedstock}}$$

Specific Gasification Rates

- Gas production rate per unit volume of gasifier
- Gas production rate per unit cross section of gasifier

Specific Production and Consumption Figures

- Gas (or byproduct) production per unit mass of solid feedstock
- Utility consumption per unit mass of solid feedstock
- Utility consumption per unit volume of product gas

4.2.3. Mathematical Modeling of Gasification Reactors

Mathematical models have gained importance as practical tools for designing gasification reactors and improving their operation and control. The general aim is to predict and interpret gasifier performance features such as

Product gas yield and composition
Utility consumption figures
Efficiency
Response time to input changes

as affected by

1) Solid feedstock characteristics (e.g., proximate and ultimate analysis, reactivity), and
2) Design and operational parameters (e.g., gasifier size, pressure, composition of gasifying medium).

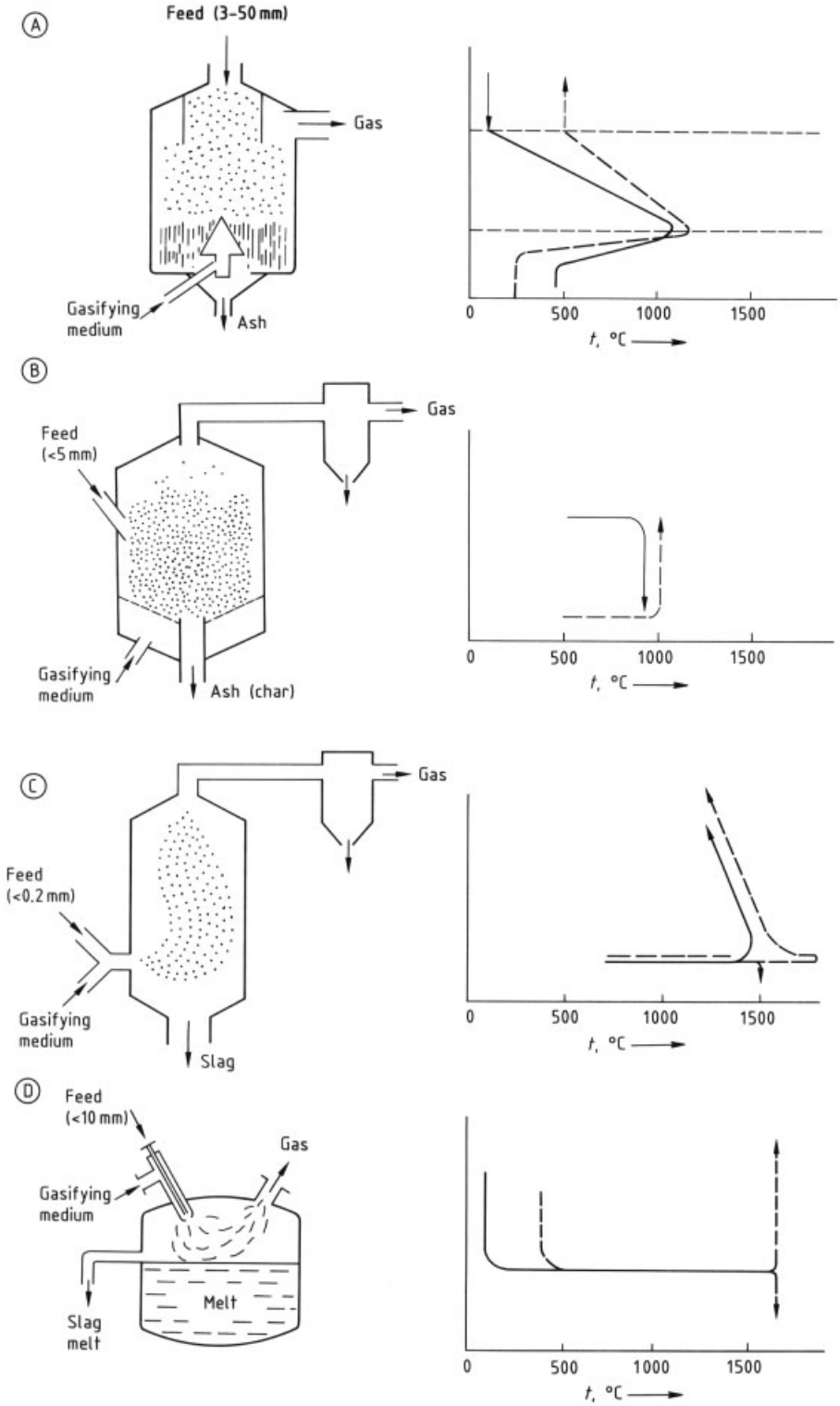

Figure 22. Types of gasification reactor representing different methods of contacting reactants and operating temperatures A) Moving bed (nonslagging, countercurrent); B) Fluidized bed; C) Entrained flow; D) Molten bath

The degree of detail and the complexity of mathematical models depend on the particular type of gasifier and the task to be addressed. Progress in computing and in efficient algorithms, as well as the interest in coal gasification, led to numerous modeling activities and higher sophistication of models during the 1970s and later.

For gasification reactors, the following types of models are most widely applied:

1) Blackbox models (empirical correlations),
2) Equilibrium models, and
3) Kinetic models.

Blackbox Models (Empirical Correlations). Experimental data form the basis for correlation of commercial-size gasifier performance with feedstock characteristics and operational parameters. Such correlations are of great practical importance for all types of gasification processes. They usually require little fundamental knowledge and mathematical effort. However, because they are not based on fundamental principles, their usefulness for inter- or extrapolation is limited.

Equilibrium Models. Equilibrium models are valuable for identifying limits of conversion and product yields and for predicting reactor outlet conditions. This description is approximate because chemical equilibrium is usually reached only after long reaction times. Reaction rates become slow when equilibrium is approached, leading to long residence times or large reactor volumes to attain equilibrium conversion. How well an equilibrium approach can describe reactor outlet conditions in a real system depends on the rates of chemical reactions, mass transfer, and convective flow of reactants.

Equilibrium models are well established for high-temperature gasification systems (entrained-flow and molten-bath reactors), where chemical reactions are sufficiently rapid to approach thermodynamic equilibrium (e.g., [69]).

In gasifiers where product gas composition results from reactions occurring at lower temperature (e.g., fluidized-bed or moving-bed reactors), a defined deviation from chemical equilibrium can be incorporated by means of corrective factors for equilibrium constants or temperature differences between actual and equilibrium values [70, 71]. However, these restricted equilibrium parameters generally cannot be correlated with feedstock properties and operating variables on a fundamental basis. Therefore, capabilities for extrapolation are limited.

Kinetic Models. Kinetic models offer the possibility of describing local conditions in a chemical reactor, thus representing the process that eventually leads to reactor outlet conditions. Mathematical representation consists of differential equations based on balance equations for individual compounds and energy, including rate expressions for chemical reactions (e.g., Eqs. 14 – 18) and transfer of mass and heat. Because kinetic models provide the most detailed description of gasification reactors, they require a large number of parameters (activation energies, Arrhenius factors, mass- and heat-transfer parameters, etc.). Some parameters depend on feedstock characteristics and can be determined experimentally only with difficulty. Typical results of kinetic models are temperature and component concentration profiles in axial and radial directions. Examples of kinetic models can be found for moving-bed [66, 72], fluidized-bed [73, 74], and entrained-flow gasifiers [75, 76].

4.3. Characterization of Solid Feedstocks for Gasification

An overview of the molar ratios of carbon: hydrogen: oxygen for solid feedstocks is presented in Figure 23. Table 15 lists typical analytical data and Table 16 gives an overview of characteristic properties of different solid biomass and waste materials. Some of the most important properties that characterize solid carbonaceous materials are discussed briefly in the following. The significance of these properties depends on the type of gasification process.

Moisture Content. Solid carbonaceous materials have moisture contents ranging from $<$ 5 % (petroleum-derived residues and anthracitic coals) up to about 50 % (lignitic coals, peat, and wood). *Moving-bed gasifiers* can accommodate the highest moisture content (ca. 35 %) because the feedstock is dried in the gasifier by

Table 15. Data from typical solid feedstocks for gasification

	Delayed coke	Anthracitic coal (Ruhr)	Bituminous coal (Pittsburgh)	Lignitic coal (North Dakota)	Wood (redwood)	Municipal solid waste
Analysis as-received, wt %						
C	82.6	88.6	75.4	48.7	35.8	47.6
H	3.1	3.2	2.2	3.0	4.1	6.0
O	0.2	3.3	8.1	14.6	29.7	32.9
N	1.6	0.1	1.2	0.6	0.1	1.2
S	3.4	0.9	2.2	0.2	0.1	0.3
Ash	0.4	2.9	6.5	6.6	0.2	12
Moisture	8.6	1.0	1.4	26.1	30.0	20
V [a]	1000	30 [b]	30 [b]	30 [b]	15	n.a. [c]
Ni [a]	220	20 [b]	20 [b]	20 [b]	15	n.a. [c]
Molar ratio						
H/C	0.46	0.43	0.83	0.74	1.37	1.51
O/C	0.002	0.03	0.08	0.22	0.62	0.58
Volatile matter, wt %	10.2	5.6	38.0	27.8	58.1	59.5
Fixed carbon, wt %	80.8	90.5	54.1	39.5	11.7	8.5
Reactivity [e] relative to bituminous coal	0.29	0.48	1.00	5.90	3.93	n.a. [c]
Ash softening temperature [e], °C	1155	1240	1260	1240	1150	n.a. [c]
Grindability, HGI [f]	50 – 100	50	60	60		
General characteristics	wide distribution of particle size, varies with cutting procedure	particle-size distribution depends on mining – cutting and preparation/beneficiation				

[a] Parts per million (by weight).
[b] Typical value for coal.
[c] n.a. = not available.
[d] Gasification rate in CO_2 (standard conditions).
[e] Oxidizing conditions.
[f] HGI = hard growth index.

Table 16. Characteristic properties of nonconventional solid feedstocks for gasification

	Reactivity	Moisture content	Inorganics content	Inorganics composition	Inorganics melting temp.	Fixed carbon content	Volatiles content	Critical compounds
Biomass								
Wood	high	high	low	Fe, Al, Si, Ca	low	low	high	
Waste wood	high	low	varying	varying	low	low	high	insecticides pesticides
Energy crops	high	high	moderate	Si	low	low	high	
Agriculture waste	high	low	low to high	Si, Na, K	low	low	high	
Sewage sludge	high	high	high	Si, Al, Fe, Ca, P	normal	low	high	Hg, Cd, Pb, Cu, Cl, (N)
Waste materials								
Municipal waste	high	varying	medium to high	Si, Ca, Al, Fe	normal	low	high	Cl, F, Zn, Pb, Cu, Cr, S, Hg
Hazardous waste	varying	varying	varying	Si, Fe	varying	low	varying	Zn, Cu, Ba, Sn
Used tires	low	low						Zn
Petroleum residues								
Coke	low	low	moderate		low	high	low	Ni, V
Hydrogenation residue *	low	low	medium		medium	high	low	Ni, V

* Example: Veba Combi Cracking (using lignite coke as additive).

exiting product gas. *Entrained-flow* and *molten-bath gasifiers* generally require that the moisture content be adjusted by predrying to maintain free-flowing behavior during crushing, conveying, and charging to the gasifier. *Fluidized-bed processes* can cope with higher moisture content which, however, in any case leads to lower process efficiencies.

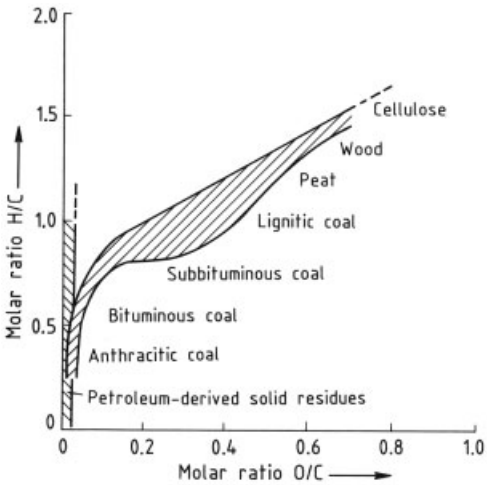

Figure 23. Molar ratios C–H–O of feedstocks for gasification

Ash Content. Ash is present in petroleum-derived residues and wood in very low amounts ($< 1\,\%$) and in coals from 1 to $> 50\,\%$. For gasification, ash should be at a minimum because of the following reasons. High ash content is detrimental to carbon conversion. It furthermore must be heated to the gasifier outlet temperature, which leads to heat losses. In addition, provision must be made for withdrawing ash from the system, which increases in complexity and cost with increasing operating pressure. An exception is the *dry ash moving-bed gasifier* which can cope with high ash content (up to ca. 50 %) and utilizes sensible heat in the exiting ash to preheat the gasifying medium.

Ash fusion behavior is significant because in some gasification processes the ash must remain solid for removal (moving-bed nonslagging and fluidized-bed processes), whereas in others it is removed as a liquid slag (moving-bed slagging, entrained-flow, and molten-bath processes). Fusion temperature, therefore, must be thoroughly taken into account in the design and operation of gasifiers. Ash composition must be considered when selecting materials of construction, particularly for slagging gasifiers. Ash can vary in character from acidic to basic and can contain elements such as iron, which may separate or require special handling. Also, in slagging operation addition of reagents such as fluxes may be desirable to improve flow characteristics of the slag or to adjust its chemical behavior.

Ash can be beneficial when particular constituents like potassium enhance the kinetics of gasification, tar cracking, or other reactions.

Volatile Matter. Volatile matter refers to gas, tar, oil, and phenolic compounds formed by pyrolysis reactions of the feedstock. The yield of these volatile products determines the remaining amount of char that must be gasified in (relatively slower) gas – solid reactions. Pyrolysis product distribution depends strongly on feedstock characteristics and, to some extent, on reaction conditions during pyrolysis (temperature, heating rate, pressure, ambient gas composition, particle size [51]).

Fixed Carbon. The residual char remaining after pyrolysis contains the fixed carbon that must be gasified in gas – solid reactions with steam, hydrogen, carbon dioxide, air, or oxygen. Because these gas – solid reactions are endothermic for most gasification processes, the specific heat requirement per unit mass of feedstock (which in autothermic processes means specific oxygen or air consumption) depends on the amount of fixed carbon in the feedstock. Of the different feedstocks presented in Table 15, wood and municipal solid waste have the lowest value of fixed carbon, whereas delayed coke, petroleum-derived residue from a high-severity coking process, and anthracitic coal have the highest values.

Caking Behavior. Caking and swelling occur with certain coals, a phenomenon related to pyrolysis reactions. Certain bituminous coals with a high content of volatile material become fluid, and the particles may adhere together and swell. Other coals with similar volatile matter content may retain the original shape, with individual pieces caking together weakly. Still other solid carbonaceous materials (e.g., lig-

nitic coals) can disintegrate during pyrolysis into small pieces. Accordingly, feedstocks are referred to as strongly or weakly caking or noncaking. Fluidized-bed processes tend to be sensitive to strongly caking coals, whereas entrained-flow, molten-bath and moving-bed processes (with sophisticated mechanical stirrer systems) can generally accept feedstocks with caking properties.

Reactivity. Solid carbonaceous materials and their residual char from pyrolysis exhibit major variations in their reactivity to steam, hydrogen, carbon dioxide, and oxygen. This can be attributed basically to three factors [77]:

1) Concentration of active sites (i.e., carbon atoms located at the edges of crystallites or building blocks);
2) Diffusional limitations on how rapidly reactive gas molecules can reach active sites; and
3) Presence of catalytic inorganic impurities.

In technical gasification processes, reactivity affects the temperature level at which the gasifier must be operated to ensure acceptable carbon conversion. As a consequence, specific heat requirements (i.e., oxygen or air consumption in autothermic processes) are related to feedstock reactivity, to an extent dependent on the particular type of process. For example, low reactivity in combination with low ash fusion temperature may be prohibitive for dry ash removal gasification processes (e.g., fluidized-bed processes). Highly reactive feedstocks are particularly suited for methane production because low operating temperatures favor the formation of methane according to chemical equilibrium.

Of the feedstocks in Tables 15 and 16, wood, biomass waste, and lignitic coal are the most reactive, whereas petroleum-derived residues and anthracitic coal have the lowest reactivity.

Particle-Size Distribution. The particle-size distribution of a given carbonaceous material depends on its origin and characteristics: (1) system of mining (for coals), (2) type of coking process (for petroleum-derived residues), (3) kind of cutting (for wood), (4) kind of pretreatment (for waste materials), (5) mechanical strength, and (6) method of preparation and transport. For example, strip-mined bituminous coal may contain up to 30 % fines (< 3 mm) whereas heavily mechanized deep-mined coal may have 40 % or more, and strip-mined lignite can reach 60 % [55].

In *moving-bed gasifiers*, relatively coarse feedstock particles (3 – 50 mm) are preferred. The acceptance of fines is limited. Developments have moved, however, toward acceptance of a greater proportion of fines, either as agglomerates or by direct injection of fines into the gasifier. *Fluidized-bed, entrained-flow*, and *molten-bath processes* generally require pretreatment steps such as drying, crushing, and grinding to achieve an optimum feedstock size distribution. Coal mining techniques that result in high fines production offer an advantage in this case, because they help reduce costs of the coal preparation plant necessary for fine coal processes. Where a grinding step is needed, feedstock grindability is an important property that affects the kind and cost of equipment.

4.4. Moving- or Fixed-Bed Processes

In moving-bed gasifiers (also called fixed-bed gasifiers), the gasifying medium passes through a bed of granular or lump fuel. Figure 22 A shows a *nonslagging, countercurrent version*, which exhibits excellent thermal efficiencies because outgoing ash heats the incoming gases and outgoing product gas heats the incoming solid feedstock. The long residence times of solid particles moving through the bed (typically 1 – 2 h), together with the temperature profile of the countercurrent system, allow high carbon conversion efficiencies.

Tables 17 and 18 contain information on selected moving-bed gasification processes and typical performance data with coal and petroleum-derived residues as feedstocks. Dry ash and slagging pressure gasifiers are shown as examples in Figures 24 and 25.

In countercurrent moving-bed gasifiers, the solid fuel descends through four not distinctly separate zones with varying temperatures (Fig. 22 A) and gas compositions, where the chemical reactions outlined in Figure 18 can occur.

Drying and Pyrolysis Zone. Raw feedstock comes in contact with hot product gases, and moisture is driven off. Subsequently, pyrolysis of the carbonaceous material occurs to yield gaseous products and char.

Table 17. Selection of moving-bed gasification processes [a]

Process name	British Gas – Lurgi (BGL) slagging gasifier	Lurgi pressure gasification	Wellman	Wellman Galusha	Imbert	Thermoselect
Licensor or contributor [b]	British Gas Corp., Lurgi	Lurgi	Wellman Thermal Systems	Dravo	Imbert Energietechnik	Thermoselect S.A.
Feedstock	coal, petroleum coke, waste RDF [c]	coal, waste RDF [c]	coal	coal	wood	municipal solid waste
Gasifying medium	st/ox	st/ox, st/air	st/air	st/air, st/ox	air (downdraft)	ox
Method of heat supply	I/A	I/A	I/A	I/A	I/A	I/A
Operating pressure, MPa	elevated (2.0 – 3.0)	elevated (2.0 – 10.0)	atmospheric	atmospheric	atmospheric	atmospheric
Condition of residue	slag	dry ash	dry ash	dry ash	dry ash	slag
Application of product gas	FG, FG/IGCC, SG	FG, FG/IGCC, SG	FG	FG	FG	FG
Developmental status	commercial	commercial	commercial	commer-cial	commercial	commercial
Number of commercial-size gas generators	2	> 165	> 40	> 150	very high	3
Reference	[55, 78 – 80]	[55, 83]	[84]	[78]	[54, 87]	[86]

[a] For abbreviations, see Section 4.2.1.
[b] in some cases historical.
[c] RDF = Refuse-derived fuel.

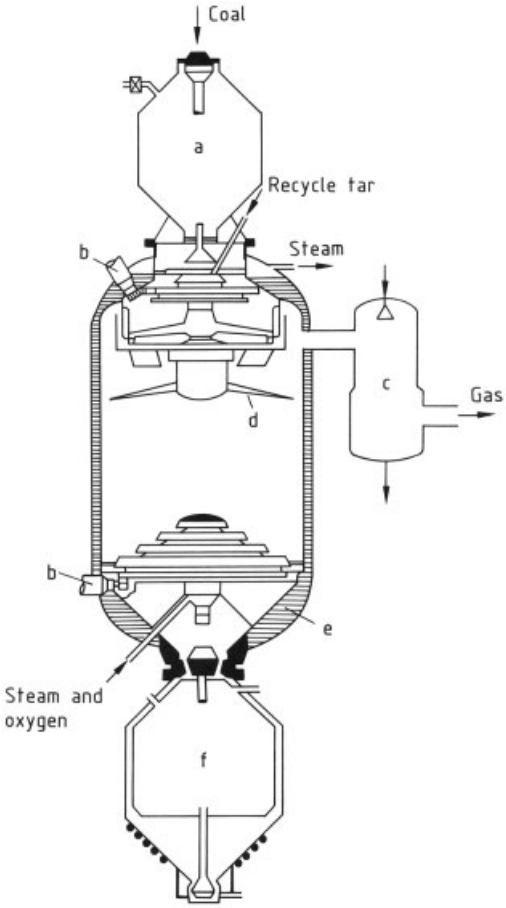

Figure 24. Schematic of Lurgi dry ash pressure gasifier
a) Coal lock; b) Drive; c) Scrubbing cooler; d) Stirrer; e) Water jacket; f) Ash lock

Gasification Zone. Char from the pyrolysis zone comes in contact with hot combustion products and steam from the zone directly below. Char – gas reactions occur mainly with steam, carbon dioxide, and to a lesser extent, hydrogen, so that the overall reaction is endothermic.

Combustion Zone. The combustion zone supplies heat for the gasification zone directly above. The key reaction is that of carbon (in the remaining char from gasification) with oxygen to produce heat and carbon oxides. Here, the temperature rises to a maximum, which in nonslagging operation must be kept below the fusion point of the ash by supplying steam in excess of that required to gasify carbon. In slagging operation, the steam supplied together with oxygen only slightly exceeds that required to gasify carbon. This results in temperatures of 1500 – 2000 °C, at which the ash melts and drains off as a liquid slag.

Ash Zone. In dry ash removal gasifiers, the ash bed, located at the bottom, heats the incoming gasifying medium by direct heat exchange and acts as a distributor.

Lurgi Pressure Gasification Process. A widely used version of the moving-bed gasifier is the Lurgi dry ash pressure gasifier (Fig. 24), which has been used commercially since the 1930s. The main gasifier shell is surrounded by a water jacket in which process steam is pro-

Table 18. Typical performance data of moving-bed gasification processes, feedstocks: coal and petroleum-derived residues [a]

Process name	British Gas – Lurgi		Lurgi				Wellman
Feedstock	bituminous coal Pittsburgh	delayed coke	bituminous coal South Africa	lignite North Dakota	bituminous coal Germany	bituminous coal Germany	bituminous coal
Size of gasifier	demonstration	demonstration	commercial	commercial	commercial	pilot	commercial
Volatile matter, wt %	34.1	8.9	20.6	26.9	31.7	23.8	32.0
Fixed carbon, wt %	50.2	85.7	50.9	30.6	52.6	38.6	49.0
Moisture, wt %	5.0	3.6	10.0	36.3	3.0	5.0	7.0
Ash, wt %	10.7	1.8	18.5	6.3	12.7	32.6	12.0
C, wt % (daf [b])	83.6	92.6	80.6	73.9	82.3	83.6	75.1
H, wt % (daf)	5.7	3.8	4.0	5.2	5.1	6.0	5.9
O, wt % (daf)	6.8	0.2	13.0	18.9	9.4	6.8	16.5
N, wt % (daf)	1.6	1.4	2.1	1.0	1.3	1.1	1.3
S, wt % (daf)	2.0	2.0	0.2	1.0	1.8	2.3	1.1
Molar ratio							
H/C	0.818	0.492	0.596	0.844	0.744	0.861	0.943
O/C	0.061	0.002	0.121	0.192	0.086	0.061	0.165
Fischer assay tar, wt % (daf) [b]	17.8	1.2	5.5	8.9	14.8		
Feed throughput, t/h (daf) [b]	10.5	8	23.8	25.1	13.0	6.0	2.5
Reactor diameter, m	1.8	1.8	3.8	3.8	3.4	1.5	3.0
Operating pressure, MPa	2.5	2.5	2.8	3.1	2.0	9.0	atmospheric
Gasifying medium	st/ox	st/ox	st/ox	st/ox	st/air	st/ox	st/air
Ratio of steam to O_2	1.0	1.1	5.8	6.9		7.0	
Ratio of steam to air					0.4		0.2
Crude gas composition, vol %							
CO_2	3.0	1.0	27.8	32.0	10.0	30.2	4.0
CO	56.4	60.0	23.0	15.8	21.5	17.5	29.0
H_2	28.0	30.0	28.6	39.8	20.0	35.0	16.0
CH_4	7.1	3.0	9.1	11.8	3.5	16.5	2.0
C_nH_m	0.4	0.2	0.4	0.8	0.8	0.8	
N_2	4.2	5.4	0.4		44.2		49.0
Crude gas yield, m^3 (STP)/t (daf [b])	2190	2640	2381	1980	3800	2075	4300
Tar – oil yield, kg/t (daf [b])			42.0	63.7		33	
Oxygen consumption, m^3/m^3 (STP)	0.20	0.17	0.14	0.13		0.18	
Air consumption, m^3/m^3 (STP)					0.56		
Cold gas efficiency, [c] %	85.1	88.1	80.9	86.8	76.3	81.9	ca. 80
Reference	[88]	[89]		[90], [91]	[92]	[93]	[84]

[a] For abbreviations, see Section 4.2.1.
[b] daf = dry and ash free.
[c] For definition, see Section 4.2.2.

duced. A coal lock hopper is mounted on top of the gasifier to feed the coal (typically 3 – 50-mm particles), and a motor-driven distributor is used to spread incoming coal evenly over the coal bed. A mechanical stirrer is included in some designs to allow the gasifier to handle caking coals. A motor-driven grate at the bottom of the gasifier is used to withdraw the ash formed, which drops

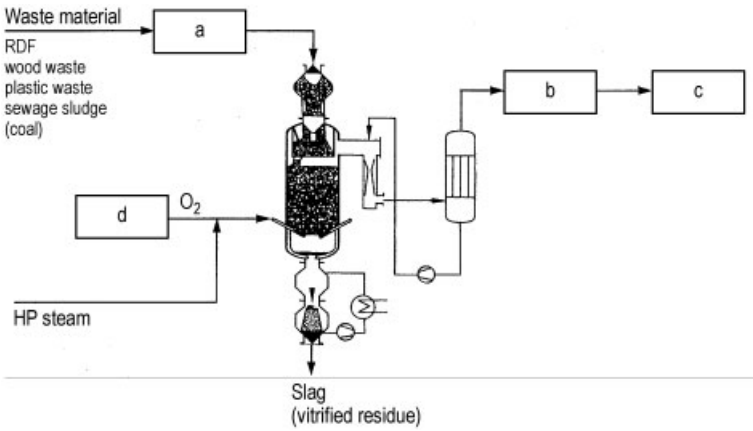

Figure 25. Schematic of British Gas – Lurgi slagging gasifier (syngas production from solid waste materials, based on [80])
a) Feed preparation (metal removal, drying, pelletizing); b) Gas purification and conditioning; c) Clean gas utilization (methanol synthesis, combined cycle power generation); d) Air separation

into the ash lock hopper. Steam and oxygen (or air) are introduced at the bottom of the gasifier and distributed through the coal bed via the rotating grate. The grate supports the coal bed and is continuously rotated to ensure constant, even withdrawal of ash. Raw product gas leaves from the top of the gasifier (typically at 400 – 600 °C) and flows through a scrubber – cooler where it is washed by circulating gas liquor. Further cooling, cleaning, and conditioning of the gas must be done according to the desired application.

In the *British Gas – Lurgi Slagging Gasifier*, a slagging version of the Lurgi gasifier, the ash melts as a eutectic with potentially added flux to form slag. The high temperature in the combustion zone is a result of low steam oxygen ratios in the gasifying medium. Molten slag is tapped discontinuously through a taphole into a quench vessel, where the slag granulates as it drips into the quench water. One application of this process is the production of synthesis gas from various waste materials, as shown in Figure 25.

Compared to the dry ash process, the slagging version has higher capacities per unit volume (see Fig. 32), has lower hydrogen carbon monoxide ratios, and produces less aqueous condensate, tar, phenols, and hydrocarbons. These can be recycled and destroyed by gasification.

For *gasification of wood*, most moving-bed processes apply the *cocurrent* (or *downdraft*) system in which product gas flows through a hot zone before leaving the gasifier (Fig. 26). The advantage of this system is the low tar content of the product gas, because most of the condensable pyrolysis products are cracked while passing through the combustion zone. A higher tar content in the product gas does not cause a problem when a gas turbine is situated close to the gasifier since tars are kept in the vapor phase. An example of a wood gasification process is given in Table 17. Moving-bed gasifiers for wood gasification are suited to small-scale application (generally less than 200 kg of wood per hour). Performance data for a moving-bed downdraft wood gasifier are given in Table 21.

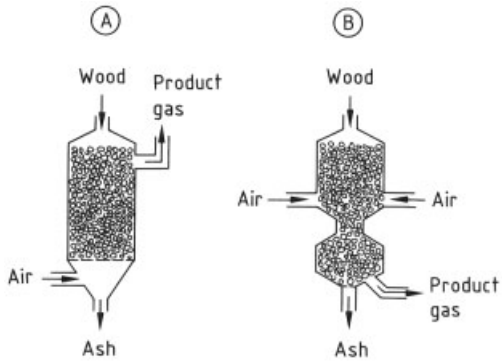

Figure 26. Types of moving bed gasifier for wood gasification
A) Countercurrent updraft; B) Cocurrent downdraft

For gasification of municipal waste together with other waste materials, practical experience is available with pressurized moving-bed gasifiers and with the new Thermoselect process (see

Table 21). In the latter case, raw municipal waste is pyrolyzed before being gasified under slagging conditions. In the case of moving bed gasification, pretreatment of the waste feedstock is required (extrusion/pelletizing, see Figure 25). Slagging operation ensures vitrification of the residue containing the inorganic material in the feedstock.

Some *advantages of moving-bed gasifiers* are: (1) the technology is mature with many commercial designs available, including pressure operation; (2) the large fuel inventory provides safety, reliability, and stability; (3) the dry ash gasifier can cope with high ash content feedstocks; (4) the gasifier may be banked for long periods; (5) the carbon conversion efficiency is high; (6) the thermal efficiency is high because of countercurrent flow and thus the specific oxygen consumption is low (see Fig. 33); and (7) the capability for turndown is high.

Some *limitations* of moving-bed gasifiers include (1) internal moving parts with some mechanical complexity are employed; (2) gasifier capacity is limited by gas flow rates (see Fig. 32); (3) feedstock fines must be handled separately (e.g., via agglomeration); (4) the product gas contains tar, oil, phenols, and ammonia, which require separation and cleanup systems; and (5) excess steam for temperature control leads to thermal losses and requires special condensate treatment.

Besides the processes mentioned in Tables 17 and 18, other moving-bed processes exist, some of which are of historical interest only: BCURA, Chevron, Demag, Foster Wheeler, Gegas, Gas Integrale, Grand Forks, Kellogg, KGN [81], Kilngas [82], Leuna, Morgas, Otto, Pintsch, Riley – Morgan, Wilputte, Woodall Duckham for coal gasification [55, 58, 78], and various processes for wood gasification [85, 94].

4.5. Fluidized-Bed Processes

Fluidized beds for gasification of solids are characterized by linear velocities of gasifying medium sufficient to lift the solid particles. This requires smaller particle sizes than moving-bed processes, typically in the 0.1 – 5-mm range (Fig. 22 B). As a result of fluidization, the solid particles move about randomly so that mixing of solids within the bed approaches that of a perfectly mixed reactor. Therefore, the solids composition and the temperature of the bed are nearly uniform, bed temperature being chosen according to feedstock reactivity. The maximum bed temperature is limited by the ash softening and sticking behavior. The temperature of reactor exit gases is about the same as that of the bed; a heat recovery device is required for energy economy.

Tables 19, 20 and 21 contain information on selected fluidized-bed gasification processes and typical performance data with various feedstocks. The number of processes in these tables reflects the fact that fluidized beds have been favored in many recent process development projects, mainly because they combine some advantages of moving-bed (e.g., large fuel inventory) and entrained-flow (e.g., control of hydrocarbon formation) gasifiers. In addition to the processes listed, others combine fluidized beds with various contacting methods, such as dilute phase entrainment or moving bed. An example is the *Hygas gasification process*. In Tables 19 and 20 the DMT process is listed as an example for an allothermic process, with heat from an external source (e.g., from a nuclear reactor or from hot flue gas) introduced via heat exchange bundles in the fluidized bed. These data are included although presently there is little commercial interest in the process.

Fluidized beds with gas velocities far above the minimum fluidizing velocity (3 – 10 m/s), called *circulating* or *fast fluidized beds* [95], are also used for gasification. They operate at lower bed densities and can thus cope with caking coals without particle agglomeration or sticking.

Winkler Gasification Process. The Winkler gasifier, the first commercially applied fluidized-bed process, has been in service since the 1920s for producing fuel and synthesis gases from coal (Fig. 27). The coal feed inlet is located in the lower part of the gasifier. Variable-speed screws feed crushed coal (< 9 mm) into the gasifier. Steam and oxygen (or air) are fed through nozzles located at several levels in the fluidized bed which occupies only the lower third of the gasifier; the remainder serves as a disengaging zone. Secondary steam and oxygen injection occur above the bed level to gasify unconverted carbon leaving the bed, tars, and hydrocarbons. Heavier particles drop through the fluidized bed and pass into the ash discharge unit at the gasifier bottom,

Table 19. Selection of fluidized-bed gasification processes; feedstocks: coal and petroleum-derived residues, wood, and biomass waste *

Process name	DMT (catalytic)	Foster Wheeler (Ahlstrom)	DMT	High temperature Winkler	Lurgi CFB	U-gas	Winkler
Licensor or contributor **	DMT		DMT	Rheinbraun	Lurgi	IGT	Davy McKee
Feedstock	coal	wood chips	agriculture forestry waste	lignite, brown coal	coal, wood, waste	coal	coal
Gasifying medium	st	air	st	st/air, st/ox	st/air, st/ox	st/air, st/ox	st/air, st/ox
Method of heat supply	E	I/A	E	I/A	I/A	I/A	I/A
Operating pressure, MPa	4.0	2.0	0.3 – 2.0	1.0 – 2.5	< 0.2	< 3.5	< 0.4
Condition of residue	dry ash	dry ash	dry ash	dry ash	dry ash	dry ash (agglomerated)	dry ash
Application of product gas	SG	FG (GT)	FG (GT, GM, FC)	FG, FG/IGCC, SG	FG, FG/IGCC, SG	FG, FG/IGCC	FG,SG,H_2
Developmental status	development	commercial	development	commercial	commercial	commercial	commercial
Number of commercial-size gas generators		≥ 1		2	3	≥ 8	ca. 70
References	[96, 97]	[98]	[99]	[78, 102, 108]	[89, 103, 104, 109]	[55, 78, 85, 105]	[55, 78, 110]

* For abbreviations, see Section 4.2.1.
** In some cases historical.

whereas lighter particles are carried up through the bed with product gas. Coal residence time is 20 – 30 min, based on coal feed rate. Carbon conversion is limited by feedstock reactivity and continuous loss of carbon with the ash.

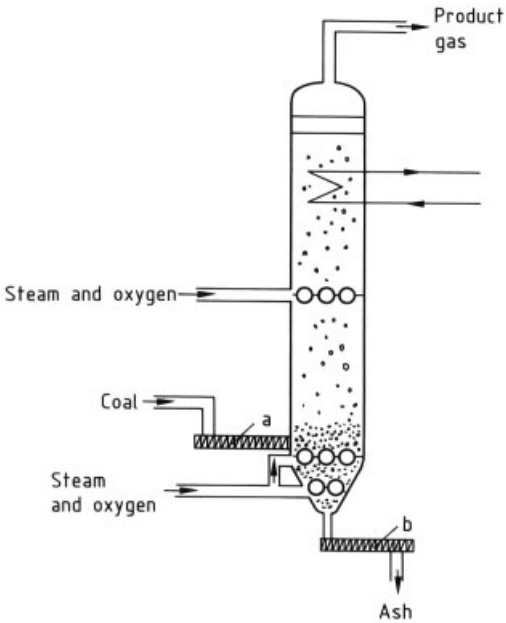

Figure 27. Schematic of Winkler gasifier
a) Coal screw; B) Ash screw

A modification of the atmospheric Winkler gasification process has been developed (high-temperature Winkler process, HTW): It operates at elevated pressure (up to 1.0 MPa) and slightly increased temperature, i.e., conditions that considerably improve process efficiency and gas quality with respect to the requirements of a synthesis gas.

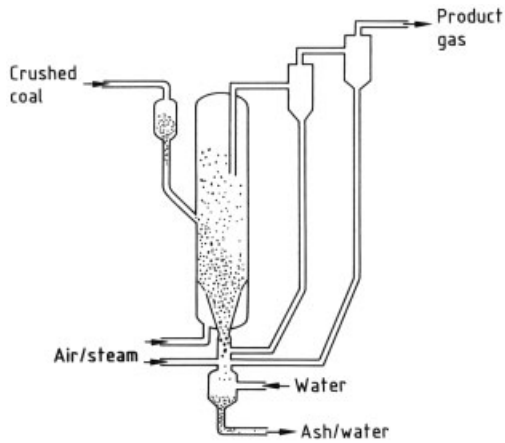

Figure 28. Schematic of U-Gas gasifier

U-Gas Gasification Process. A pressurized version of a fluidized-bed gasifier with controlled ash agglomeration is shown in Figure 28. The U-gas gasifier is a vertical cylinder with two cyclones for returning elutriated fines to the bed. A sloping grid at the bottom serves as the air and steam distributor and the agglomerated ash outlet. Noncaking coals can be fed directly to the gasifier from the crusher (particle size < 6 mm). Caking coals may require pretreatment by con-

Table 20. Typical performance data of fluidized-bed gasification processes; feedstocks: coal and peat [a]

Process name	DMT	High-temperature Winkler		U gas	Winkler	
Feedstock	bituminous coal	brown coal	peat	bituminous coal	lignite	lignite
Size of gasifier	pilot	commercial	commercial	pilot, commercial	commercial	commercial
Volatile matter, wt %	12.2	44 – 46	57 – 61			
Fixed carbon, wt %	68.2	37 – 40	20 – 24			
Moisture, wt %	1.2	12 (10 – 25)	15 (12 – 18)		8.0	8.0
Ash, wt %	18.4	4 (2 – 18)	4 (3.5 – 4.5)			
C, wt %, (daf [e])	86.9	69	60		71.1	71.2
H, wt %	4.0	5.0	5.0		5.5	4.8
O, wt %	4.9	25	31		18.9	20.2
N, wt %	2.0	0.9	2.8		0.9	2.2
S, wt %	2.2	0.1	0.2		3.8	1.6
Molar ratio						
H/C	0.559	0.870	1.200		0.920	0.818
O/C	0.042	0.272	0.387		0.199	0.213
Feed throughput, t/h (daf [e])		22	21	1 – 5	29	41
Reactor diameter, m		2.71	2.4		6	6
Operating pressure, MPa	4.0	1.0	1.0	1.6 – 2.4	0.105	0.105
Gasification temperature, °C	830	950	950	ca. 1000	815 – 1200	815 – 1200
Gasifying medium	st	st/ox	st/ox [c]	st/air	st/air	st/ox
Ratio of steam to O_2						1.43
Ratio of steam to air				ca. 0.3	0.06	
Crude gas composition, vol %						
CO_2	22.3	22.8	21.3	9.9	7.7	19.5
CO	13.7	35.7	31	19.6	22.5	36.0
H_2	56.2	35.9	29.7	17.5	12.6	40.0
CH_4	7.8	4.9	5.6	3.4	0.7	2.5
C_nH_m		0.1	0.1			
N_2		0.6	12.3	48.9	55.7	1.7
Crude gas yield, m^3 (STP)/t (daf [e])		2000	2100	3596	3058	1760
Tar – oil yield, kg/t (daf [e])						
Oxygen consumption, m^3/m^3 (STP)		0.19	0.19			0.14 – 0.24
Air consumption, m^3/m^3 (STP)					0.43 – 0.76	
Cold gas efficiency, [d] %		80	73	71 – 79	61.9	74.4
Carbon conversion, [d] %	95	94	92	96 – 98	83.0	81.0
Reference	[96, 97]	[102]	[102]	[105]	[55, 56]	

[a] For abbreviation, see Section 4.2.1.
[b] Predried.
[c] Gasification with oxygen-enriched air.
[d] For definition, see Section 4.2.2.
[e] daf = dry and ash free.

tact with air in a second fluidized bed operating at gasifier pressure and 370 – 430 °C. Here, an oxidized outer layer forms on the coal particles and prevents agglomeration and possible blockage in the gasifier.

Coal is gasified with a mixture of air (or oxygen) and steam in a single-stage fluidized bed. Residence time of the particles is ca. 45 – 60 min. The fluidizing velocity is about 1 m/s. The key to gasifier operation is the agglomeration and separation of low-carbon ash from the bed. This is accomplished by keeping the temperature near the gasifying medium inlet close to the softening point of the ash particles. Thus, ash particles preferentially stick together, and the agglomerates grow until they are heavy enough to move downward and fall out of the fluidized bed.

The fluidized-bed system is particularly attractive for gasification of wood and other biomass because it allows greater feedstock flexibility (e.g., with respect to particle size) and easier scaleup than moving-bed gasifiers. There-

Table 21. Typical performance data of wood, biomass, and waste gasification processes [a]

Process name	Imbert	Lurgi/SVZ	Foster Wheeler Ahlstrom	Lurgi CFB			Thermoselect
System	moving bed downdraft	grate gasifier	fluidized bed	fluidized bed			moving bed
Feedstock	wood	RDF/various waste/coal mix	wood chips	wood bark	wood waste	lignite combustion residue	municipal solid waste [b]
Size of gasifier	commercial	commercial	commercial		commercial		commercial
Volatile matter, wt %				(70)			
Fixed carbon, wt %				(15)			
Moisture, wt %			10 – 20	8 – 10	19	10	20 – 30
Ash, wt %			2 – 5		2	78	20 – 30
Feed throughput, t/h (daf [e])		ca. 10		3.5 – 4.5	mix (70 – 100 MWth)	mix	1.5 – 5
Reactor diameter, m		3.6					
Operating pressure, MPa	atmospheric	2.4	1.8	atmospheric	atmospheric		atmospheric
Operating temperature, °C	[f]	[f]	950 – 1000	ca. 800	900 – 950		600 – 1200
Gasifying medium	air	st/ox	air	air	air		ox
Ratio of steam to O_2			–	–	–		0.0
Ratio of steam to air	0.0		0.0	0.0	0.0		–
Crude gas composition, vol %							
CO_2	10.0	31 – 35	12	15			22 – 27
CO	20.0	15 – 22	16	15.4			35 – 39
H_2	19.0	29 – 35	13	14.8			32 – 36
CH_4	1.5	9 – 13	3	4.2			< 0.1
C_nH_m				2.0			< 0.1
N_2	50.0		44	39.6			3 – 4
Crude gas yield, m^3 (STP)/kg (daf [e])				2000 – 3000			1588 – 1926
Oxygen consumption, m^3/m^3 (STP)				–			0.45 – 0.52 [c]
Air consumption, m^3/m^3 (STP)				(0.48) [g]			–
Cold gas efficiency, [d] %					> 70		
Carbon conversion, [d] %				93	> 96		
Reference	[115]	[80]	[98]	[116]	[103]		[86]

[a] For abbreviations, see Section 4.2.1.
[b] Without pretreatment.
[c] With natural gas as additional fuel (ca. 135 m^3 (STP)/t daf).
[d] For definition, see Section 4.2.2.
[e] daf = dry and ash free.
[f] Temperature distribution.
[g] Estimated value.

fore, several fluidized-bed concepts have been proposed and developed; a selection is presented in Table 21. As an example, Figure 29 shows a flow diagram for fuel gas production from biomass and waste feedstocks in a circulating fluidized-bed reactor integrated in a cement production plant. The feed material is fed to the lower part of the gasifier via a screw feeder. Air as gasifying agent is injected at the bottom. Because of excellent heat-transfer conditions, the temperature is uniform throughout the reactor and is high enough to convert the tarry and oxygenated pyrolysis products. Typical plant capacities for fluidized-bed wood and waste gasifiers are in the range of several tons per hour. The product gas is particularly suitable as feedstock for boilers, lime kilns, gas engines, and gas turbines.

Advantages of fluidized-bed gasifiers are: (1) commercial designs are available; (2) the technology is less complex than that of moving beds and does not involve moving parts; (3) the large fuel inventory provides safety, reliability, and stability; (4) a variety of particle sizes can be handled; (5) the amount of tar and phenols is low; (6) gas composition is steady due to uniform conditions in the bed; and (7) moderate gasification temperatures can be used.

Limitations of fluidized bed-gasifiers include the following: (1) capacity flexibility is limited by entrainment at high gas velocities; (2) the inventory of solid carbon is lowered

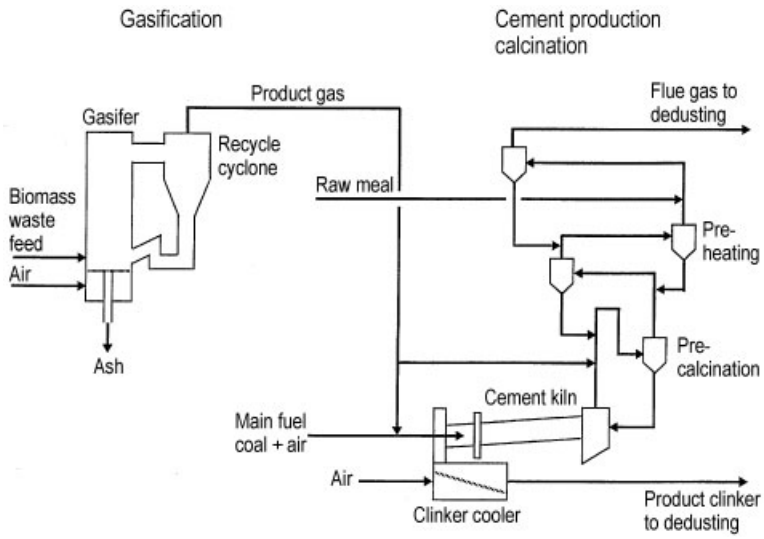

Figure 29. Flow diagram of a circulating fluid bed process for the production of fuel gas from biomass and waste feedstocks, integrated in a cement production plant, based on [103].

by the high content of inert material in the bed; (3) caking coal may require pretreatment; (4) conflicting temperature requirements exist for low-reactivity feedstocks with low-softening point ash; (5) inevitable loss of carbon in ash occurs due to uniform solids composition of the bed; (6) the turndown range is limited by the gas velocity required to maintain fluidization.

Besides the processes mentioned in Tables 19 20 21, other fluidized-bed processes are BASF – Flesch – Demag, Battelle – Union Carbide, Hydrogasification, *BCR*, CGT, CO_2-*Acceptor*, Electrofluidic process, Exxon catalytic process, Hydrane, Hydrogasification (HKV), Hygas, KRW, Synthane, Tosco, University of Wyoming, West Virginia University, Winkler – Flesch for coal or petroleum coke [55, 58, 117, 119], IGT's RENUGAS and various other processes for wood and waste gasification [54, 94, 118].

4.6. Entrained-Flow Processes

In entrained-flow gasifiers, solid particles are carried or entrained by the reacting gases. Thus, solids and gases move in the same direction with approximately the same velocity. To achieve this, the particles must be smaller than in other systems (typically 70 % $<$ 0.1 mm). This, together with high temperature (see Fig. 22 C), allows gasification rates high enough to ensure acceptable carbon conversion during the short solids residence time in the gasifier (a few seconds).

The reactants (e.g., coal, oxygen, and steam) are typically introduced into the gasifier at high velocity through one or more burners or nozzles. The burners can be oriented in many different ways, for instance, tangentially, radially opposed, and axially. Gasifier performance is affected to a large degree by the flow characteristics and mixing efficiency of the burners. The solid feedstock is either injected in the form of a water slurry or as dry particles.

Flame temperatures at the burner outlets can be as high as 2000 °C. This results in extremely rapid conversion of feedstock particles and destruction of virtually all higher hydrocarbons. Outside the immediate flame region, heat losses, further mixing with steam, and endothermic reactions combine to lower the gas-phase temperature. Because of the high reaction temperature, oxygen consumption is usually greater than for other gasification systems (see Fig. 33). A significant portion of the feedstock ash melts because of the high reaction temperature and is removed from the reaction zone as liquid slag.

Tables 22 and 23 contain information on selected entrained-flow gasification processes and typical performance data with coal or petroleum-derived residues as feedstock. Figures 30 and 31

show, as examples, a Koppers – Totzek gasifier and a scheme of the Texaco coal gasification process with gas cooler mode.

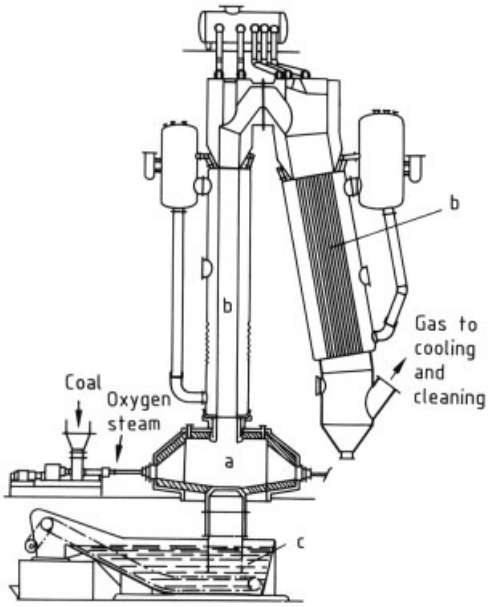

Figure 30. Schematic of Koppers – Totzek gasifier
a) Gasifier; b) Waste heat boiler; c) Slag quench tank

Koppers – Totzek Gasification Process. The two-headed Koppers – Totzek gasifier (Fig. 30) is horizontal and ellipsoidal in shape, with two heads mounted on the ends. A waste-heat boiler is mounted on top to recover heat from the hot effluent gases. The usual waste-heat boiler system, which produces high-pressure steam as a byproduct of the process, is a radiant surface boiler followed by a fire tube boiler. The gasifier shell is protected from the high gasification temperatures by a castable refractory lining. Coal is dried to 2 – 8 % moisture content (depending on coal rank) and simultaneously pulverized to 70 – 90 % below 80 µm. A variable-speed screw feeder feeds the pulverized coal into a mixing nozzle, where a premixed stream of oxygen and steam is introduced. Steam and oxygen entrain the pulverized coal, and the coal – steam – oxygen mixture is injected into the gasifier burners at injection speeds above the speed of flame propagation to prevent flashback. Two adjacent burners on each head improve turbulence behavior and provide a safety feature to ensure continuous ignition if one burner is temporarily blocked. Coal particles are gasified in approximately 1 s or less, producing a temperature in the flame zone of ca. 1900 °C. Heat losses and endothermic reactions cool the gas mixture to ca. 1300 °C. Ash in the coal is liquefied because the gasifier temperature is maintained above the ash fusion temperature. Most of the ash flows down the gasifier walls as molten slag and drains into a slag quench tank. Part of the ash leaves the gasifier as fine slag particles entrained in the exit gas. Pressurized modifications of the Koppers – Totzek process are the Prenflo and the Shell gasification processes (the latter jointly developed by Shell and Koppers), both using dry solid particles as feed.

TTexaco Gasification Process. The exaco coal gasification process (based on experience with oil gasification; see Section 3.2.2) uses a pressurized version of an entrained-flow reactor and a slurry system (Fig. 31). The gasifier is a vertical, cylindrical vessel with a refractory-lined upper portion where solid feedstock reacts with steam and oxygen. Hot product gas is cooled upon leaving the gasifier, either by direct quenching with water, resulting in internal steam production, or by indirect cooling in gas coolers with external high-pressure steam generation. The mode of operation is determined by the method of downstream processing and by the end use of the product gas. The quench mode is generally selected for the production of hydrogen or ammonia; carbon monoxide in the raw product gas reacts with steam in a carbon monoxide shift step to produce additional hydrogen. The gas cooler mode of operation is selected for applications that do not require carbon monoxide conversion (e.g., syngas for oxo synthesis or clean fuel for power generation). The solid gasifier feed is processed through two stages of dry grinding (the second under inert atmosphere) and pneumatically conveyed to slurry mix tanks where it is mixed with water to form a pumpable slurry (slurry solids loading 50 – 70 %). The slurry is fed continuously to the top of the gasifier by a positive displacement pump.

Advantages of Entrained-Flow Gasifiers. Among the advantages of entrained-flow gasification processes are: (1) commercial designs are available; (2) the gasifier has no moving parts and a simpler geometry than a fluidized bed;

Table 22. Selection of entrained-flow gasification processes *

Process name	Koppers – Totzek	Prenflo	Shell	Texaco	DesTec (DW)
Licensor or contributor	Krupp – Koppers Uhde	Krupp – Koppers Uhde	Shell	Texaco	Dynegy
Feedstock	coal petroleum coke (dry)	coal petroleum coke (dry)	coal petroleum coke (dry)	coal petroleum coke (slurry)	coal petroleum coke, (slurry)
Gasifying medium	ox (st/ox)	ox (st/ox)	st/ox	st/ox	ox
Method of heat supply	I/A	I/A	I/A	I/A	I/A
Operating pressure, MPa	atmospheric	3.0	2.0 – 4.0	< 8.5	< 1.5
Condition of residue	slag	slag	slag	slag	slag
Application of product gas	FG,SG,H_2	FG,SG,H_2,FG/IGCC	FG,SG,RG,FG/IGCC,H_2	FG,SG,H_2,FG/IGCC	FG,SG,H_2, FG/IGCC
Developmental status	commercial	development commercial	development commercial	commercial	commercial
Number of commercial-size gas generators	ca. 80	1	≥ 1	≥ 5	≥ 2
Reference	[55, 78, 120]	[78, 120, 124]	[78, 121, 122, 125]	[55, 78, 126 – 129]	[123]

* For abbreviations, see Section 4.2.1.

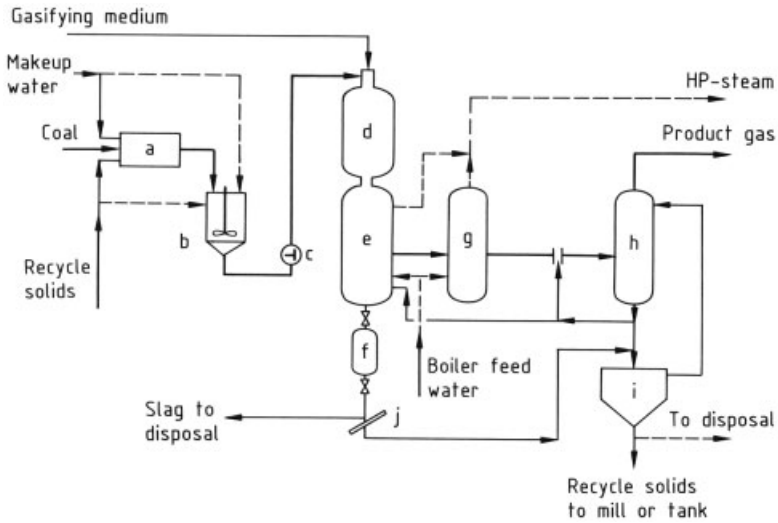

Figure 31. Schematic of Texaco coal gasification process with gas-cooler mode
a) Mill; b) Slurry tank; c) Slurry pump; d) Gasifier; e) Radiation cooler; f) Lock hopper; g) Convection cooler; h) Gas scrubber; i) Clarifier; j) Slag screen

(3) the gasifier has the highest capacity per unit volume (see Fig. 32); (4) any type of coal may be used without pretreatment; (5) no fines are rejected; (6) the product gas is free of tar and phenols (see Fig. 33); and (6) the slagged ash produced is inert and has a low carbon content.

Limitations of entrained-flow gasifiers include: (1) combustor nozzles and heat recovery in the presence of molten slag are critical design areas; (2) a small solids inventory requires advanced control techniques to ensure safe, reliable operation; (3) pulverizing and drying of surface moisture are required; and (4) the high gasification temperature causes thermal losses, necessitates higher quality construction materials, and leads to increased oxygen consumption (see Fig. 33).

Besides the processes mentioned in Tables 22 and 23, other entrained-flow processes are Babcock Wilcox, BiGas, Combustion Engineering, Foster Wheeler, GSP, and ORC for coal or petroleum coke gasification [55, 131].

4.7. Molten-Bath Processes

In molten-bath gasification processes, solid feedstock and gasifying agents are injected into

Table 23. Typical performance data of entrained-flow gasification processes; feedstocks: coal and petroleum-derived residues [a]

Process name	Koppers – Totzek	Prenflo		Shell		Texaco	
Feedstock	coal	mixture of coal/ petroleum coke	coal	bituminous coal (Illinois no. 6)	Texas lignite	bituminous coal (Illinois no. 6)	delayed coke
Size of gasifier	commercial	commercial	pilot	pilot [g]	pilot [g]	pilot [h]	pilot
Volatile matter, wt %		18.5	33	34	31	33.5	
Fixed carbon, wt %		56	52	41	27		
Moisture, wt %	1.0 [b]	9.4	6.0	16 [c]	33 [c]	13.0	
Ash, wt %	20.2	24.5	8.8	9	9	11.1	0.5
C, wt % (daf [f])	79.7	84.7	81.5	77.0	72.0	78.1	90.7
H_2, wt %	4.6	4.0	5.4	6.0	6.0	5.1	3.9
O_2, wt %	13.5	4.7	10.1	11.0	19.0	11.6	0.9
N_2, wt %	1.8	1.9	1.9	1.4	1.2	1.3	2.6
S, wt %	0.5	4.7	1.1	4.3	1.0	3.8	1.9
Molar ratio							
H/C	0.688			0.935	1.000	0.784	0.516
O/C	0.127			0.107	0.198	0.111	0.007
Feed throughput, t/h (daf [f])		71	1.7			5.7	
Operating pressure, MPa [i]		2.5	2.4		2.1 – 2.8		
Gasification temperature, °C	1870	> 2000	> 2000			> 1250	
Gasifying medium	st/ox	st/ox	st/ox	st/ox	ox	ox	ox
Ratio of steam to O_2	0.196	0.14	0.04	0.080	0.0	1.0 [d]	0.75 [d]
Crude gas composition (dry), vol %							
CO_2	6.9	3.8	1.9	2.0	4.6	20.7	13.3
CO	64.3	61.9	58.4	65.1	61.8	41.1	53.4
H_2	27.1	22.5	26.2	30.5	32.4	36.6	31.8
CH_4	0.1	0.1	0.1			0.1	0.02
C_nH_m							
N_2	0.9	11.7	13.4	0.9	0.9	0.4	0.95
Crude gas yield, m^3 (STP)/t (daf [f])		2578	2230	2109	2029	2192	2667
Tar – oil yield, kg/t (daf [f])							
Oxygen consumption, m^3/m^3 (STP)		0.30	0.30	0.297	0.294	0.321	0.304
Cold gas efficiency, [e] %	69.3	78.4	81.8	81	81	70.3	
Carbon conversion, [e] %	90.0	99.3	99.3		99	94	
Reference	[55]	[124, 130]	[130]	[122]	[122]	[128]	[126]

[a] For abbreviations, see Section 4.2.1.
[b] Predried.
[c] Before drying.
[d] H_2O as water in slurry.
[e] For definition, see Section 4.2.2.
[f] daf = dry and ash free.
[g] Commercial-scale data in [125].
[h] Commercial-scale data in [129].
[i] Exact pressure not always available, see also Table 22.

a hot liquid bath of molten slag, metal, or salt (Fig. 22 D). The high temperature normally required to keep the bath molten promotes high reaction rates, which are often enhanced by catalytic properties of the melt. The material employed for the melt largely determines the characteristics of each particular process. Slag acts mainly as a heat-transfer medium but can sometimes have a catalytic effect. Iron is used as a bath material because of its strong affinity for sulfur, its high solubility of carbon in molten iron, and because any iron oxide present in the melt will donate oxygen. Molten salts, which are generally less corrosive and have lower melting points than molten metals, may strongly catalyze the basic char – steam reaction, permitting prac-

tically complete gasification of some feedstocks at relatively low temperature.

As an example, the Molten Iron Puregas (MIP) Gasification Process uses an iron inventory of 20 t for a coal throughput of 10 t/h. The reactor is designed for gasification pressure up to 0.3 MPa and is brick lined. It contains liquid iron into which the feed coal is injected via an inert carrier gas, as well as facilities for feeding oxygen and other media that may be required such as lime for adjusting the basicity of the slag and steam for cooling the iron. Slag floating on the surface of the iron bath is discharged continuously and subjected to secondary treatment. Product gas leaves the reactor at ca. 1400 – 1500 °C and passes through a waste-heat recovery system followed by a dust removal system. Pilot-plant design enables optimum location of feeding points, either at the top (top blowing) or at the bottom (bottom blowing) of the reactor, to be investigated. This process draws on experience available from steelmaking technology.

Advantages of molten-bath gasifiers Among the advantages of molten-bath gasifiers are (1) the large fuel or heat inventory provides safety, reliability, and stability; (2) the high-temperature bath ensures safe ignition; (3) any grade of coal in a wide size range can be utilized (i.e., crushed run-of-mine coal); (4) no pretreatment or pulverization is required for gasifier operation; and (5) product gas contains no sulfur (retained in molten bath) and no tar or phenols at high temperature.

Limitations of molten-bath gasifiers include the following: (1) cleanup of molten media (if required) is complicated; (2) capacity flexibility is limited by melt entrainment in the product gas at high flow rates; (3) ash is removed as liquid slag, leading to loss of sensible heat; (4) the catalyst (if required) must be separated from ash for recycle, and make-up catalyst must be added; (5) the melt may be highly corrosive to refractory at the gasifier temperature; and (6) bottom blowing gasifiers require sophisticated cooling systems for the tuyeres.

Process variations include besides the Molten Iron Puregas Process (MIP) other processes for coal or petroleum gasification: Rummel [61], Rockgas Molten Salt [132, 133], Saarberg/Otto [78], Atgas, Kellogg, Klöckner, Sun [55, 134].

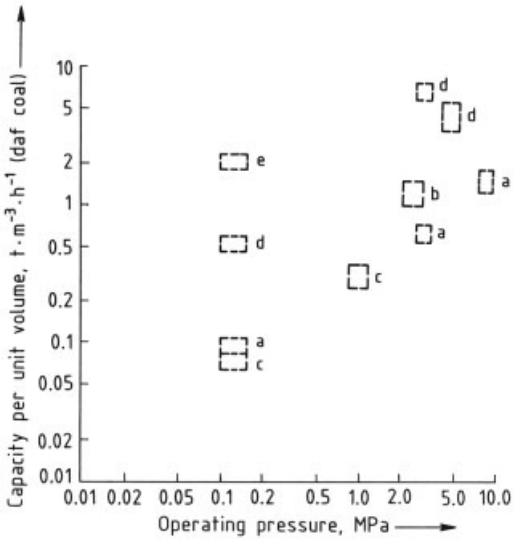

Figure 32. Effect of pressure and type of gasification reactor on capacity per unit volume, based on published data, feedstock: various coals
a) Moving bed dry ash; b) Moving bed slagging; c) Fluidized bed; d) Entrained flow; e) Molten bath iron

4.8. Underground Coal Gasification

[135 – 138]

The purpose of coal gasification in situ is to recover the energy or chemical values of the coal without mining. Gaseous products may be utilized for electric power generation, manufacture of organic chemicals, and synthesis of liquid or gaseous fuels. Advantages of in situ coal gasification are the elimination of underground mining operations and, possibly, the utilization of coal in beds that cannot be mined profitably. Included among past attempts are projects in Europe, the United States, and the Soviet Union. Although details vary from project to project, mixtures of oxygen and steam (or air and steam) must always be admitted into a hot zone in the coal seam for gasification to occur, and raw product gas must be withdrawn from the coal seam for surface processing to make the gas environmentally and technically acceptable. The main difficulties in underground coal gasification are creating and maintaining adequate permeability, maintaining and controlling the reaction zone, minimizing the detrimental effects of fissuring and channeling, and maintaining steady product

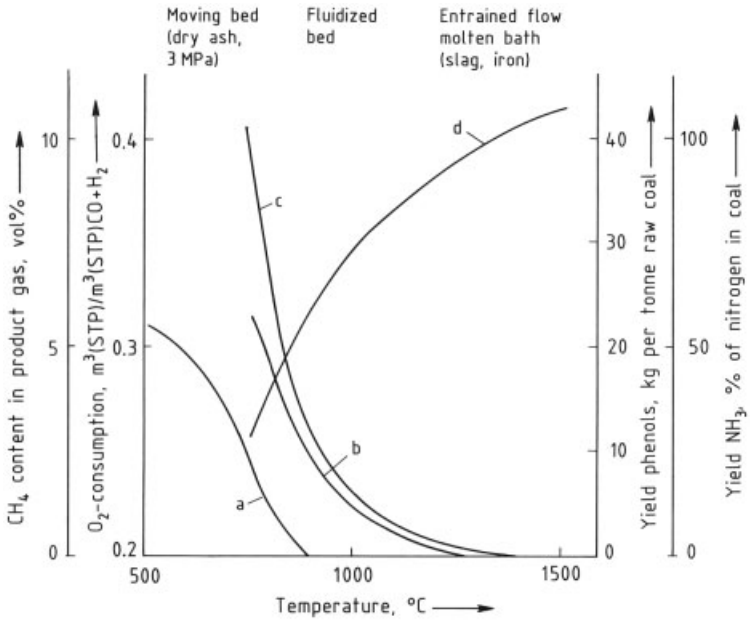

Figure 33. Characteristic temperatures of different gasification systems and correlation with selected performance data, based on published data
a) Phenols; b) NH_3; c) CH_4; d) O_2-consumption

gas composition and yield. Despite past projects, underground gasification of coal is not used widely at present and is not expected to play a significant role in the near future.

4.9. Environmental Aspects of Gasification

Gasification offers a practical means of utilizing coal, petroleum-derived residues, wood, or solid waste material while at the same time meeting stringent environmental control requirements. Although the type and severity of environmental control problem vary with gasification process, routine commercial operations have demonstrated successful cleanup of the most difficult streams.

Air Emissions. Among the sulfur compounds, reduced species (H_2S, COS, CS_2) predominate in gasifier raw gases and can be removed by a variety of processes (Chap. 5). Sulfur compound concentrations are higher in gasifier gases than in combustion gases, which facilitates their removal. For example, synthesis gas streams can be cleaned sufficiently for use in processes that require almost total removal of sulfur compounds because of the use of sulfur-sensitive catalysts. In addition, studies indicate that virtually complete control of sulfur in coal gasification increases product cost only moderately [53].

Particulates are routinely removed from synthesis gas produced by coal gasification to be fed to catalyst beds for chemical synthesis. In this case, particulates are removed to a far greater extent for catalyst protection than would be necessary to comply with environmental requirements. Of greater concern are emissions from feed lock hoppers or feed handling and storage systems; these, however, can be controlled if the equipment is designed properly.

Although nitrogen oxides are not formed to any appreciable extent in the reducing atmosphere of gasification, use of the product gas as fuel for combustion could pose an NO_x emission problem due to nitrogen compounds in the gas. However, well-cleaned product gas from gasification should yield an improvement over, for example, the direct firing of coal.

Water Effluents. Gasification processes in which raw gas leaves the gasifier at relatively

low temperature (moving bed) result in greater requirements for wastewater treating because of the tar and phenols produced, yet such wastewater can be purified adequately by conventional technology. For example, water effluent streams from Lurgi gasifiers have been well controlled on a large commercial scale at the Sasol II complex (→ Coal Liquefaction) in South Africa. Some gasification plants are designed for zero discharge in arid areas where discharges might seriously affect the quality of the receiving water.

Solid Wastes. All coal-consuming processes require the disposal of ash or slag resulting from the mineral content of coal. Some concern has been expressed that solid waste from gasification might be declared hazardous because of its leaching characteristics. However, leaching of gasification waste generally results in lower concentration of metals than leaching of power plant ash, which has not been classified as hazardous.

5. Gas Treating

Gas treating described in this chapter is often preceded by a separation stage for condensates obtained from cooling stages immediately following gas production (see Chap. 6). An overall gas treating sequence is described in Section 7.2.

Objectives. The method of gas treating is determined essentially by the following conditions:

1) origin (i.e., state and composition of the raw gas) and
2) utilization (i.e., specifications for the treated gas).

As described in previous chapters, the composition of raw gases differs considerably depending on the mode of production; this applies also to raw gas pressure and temperature (see Table 24).

Requirements for the treated gas are manifold. They are more or less stringent depending on whether the gas is to be used as a feedstock for synthesis or as fuel gas. Generally, no special requirements exist for fuel gas composition as long as the limits for calorific value and density are met (see Chap. 8). Purity is specified with a view to avoiding corrosion and deposits in the piping system or toxicity.

When syntheses follow downstream of gasification, the stoichiometry of the reaction requires a very specific composition, and the catalysts, which are generally sensitive, set very exacting standards for purity of the reactants. In addition, economic aspects must be considered concerning the inerts contained in the gas, which unnecessarily reduce the partial pressure of reactants and tend to concentrate in recycle processes so that a considerable amount of purge gas must be branched off.

Stoichiometric Conditions. Most syntheses of basic chemicals in very large quantities start from hydrogen and/or carbon monoxide, in conjunction with further reactants if necessary. The ratio of hydrogen to carbon monoxide offers a simple characterization (see Table 25).

Purity. Sulfur compounds, which are severe respiratory and catalyst poisons and may cause dangerous corrosion, are the subject of almost all purity specifications. The content of hydrogen sulfide (measured as sulfur) is generally limited to 2 mg/m^3 in long-distance pipeline systems. For syntheses, particularly those using sensitive copper catalysts, the permissible concentration is much lower: e.g., for the low-pressure methanol process, 0.1 ppm of total sulfur, measured as hydrogen sulfide.

In some syntheses, oxygen compounds are also undesirable: thus a content of

$$O_2+CO_2+2\,CO<20\ \text{ppm}$$

is usually required for ammonia synthesis (→ Ammonia). For reasons of corrosion, the oxygen content in gas grids is limited to 0.5 – 3 %. Nitrogen oxides are also undesirable because they may form unwanted or dangerous compounds with unsaturated hydrocarbons. This applies not only to syntheses but also to gas piping systems and cryogenic gas treatment processes (→ Cryogenic Technology; → Natural Gas). In view of the high process pressures, the nitrogen content of hydrogen used for hydrocracking is often limited to avoid formation of ammonia.

A number of other impurities, such as hydrogen cyanide, organic acids (e.g., formic acid),

Table 24. Typical raw gas data

Raw material:	Coal		Vacuum residue		Natural gas	
Example:	1	2	3	4	5	6
Production process:	Koppers – Totzek	Lurgi Pressure Gasification	Shell Gasification Process	Texaco Gasification Process	Tubular reformer	
	ex gasifier		downstream of scrubber	downstream of CO shift	upstream of CO shift	downstream of CO shift
Pressure, MPa	0.105	2.9	5.4	8.5	2.2	2.1
Temperature, °C	45	45	35	35	350	230
Water dew point, °C	45	45	35	35	170	162
Analysis, vol %						
H_2	28.7	43.2	45.9	63.0	76.1	78.3
CO	57.0	11.6	48.6	1.8	10.6	0.4
CH_4	0.1	10.7	0.5	0.3	3.3	3.0
C_{2+}		1.0				
N_2+ Ar	1.4	0.3	0.2	0.1		
CO_2	12.6	32.9	4.0	34.2	10.0	18.3
H_2S	0.2	0.3	0.8	0.6		
COS, ppm	300	150	500	10		
Other components	O_2, SO_2, HCN, NH_3	O_2, C_mH_n, HCN, NH_3	O_2, HCN, NH_3	NH_3		

Table 25. Typical hydrogen – carbon monoxide stoichiometric numbers for a few basic chemicals

Product	Acetic acid [a]	Oxo alcohols [b]	Direct reduction [c]	Fischer – Tropsch [d]	Methanol [e]	Ammonia
C_{H_2}/C_{CO}	0 – 0.02	1.0 – 1.2	1.3 – 1.5	1.6 – 2.3	2.0 – 2.3	$>10^5$
Additional reactants	methanol	propylene	iron oxides	CO_2	CO_2	N_2(=1/3 H_2)

[a] Monsanto process (→ Acetic Acid).
[b] BASF and Ruhrchemie process; Shell process: 1.6 – 1.8.
[c] Midrex process.
[d] Lower range for fixed-bed synthesis (ARGE); upper range for Synthol (SASOL).
[e] The reaction is reflected more accurately by $\left(C_{H_2}-C_{CO_2}\right) / \left(C_{CO}+C_{CO_2}\right)$ (→ Methanol).

or unsaturated and higher boiling hydrocarbons, are normally not included in purity specifications for synthesis processes because they are already undesirable in upstream process stages of synthesis gas production.

The total content of inerts (N_2, Ar, CH_4), which in most cases are not dangerous to synthesis but undesirable for economic reasons, must often be $< 2\,\%$.

To meet all these requirements, several process stages are necessary for gas purification or separation. They include catalytic processes (CO shift conversion, COS conversion, methanation; see Sections 5.1, 5.2, 5.3), as well as absorption, adsorption, and other fractionation processes (Sections 5.4, 5.5, 5.6, 5.7, 5.8).

5.1. Carbon Monoxide Shift Conversion

Processes for the manufacture of synthesis gas from hydrocarbon feedstocks produce mixtures of hydrogen, carbon dioxide, methane, and varying amounts of carbon monoxide. For the majority of industrial processes, this carbon monoxide content is higher than that required for synthesis. Exceptions include, for instance, the synthesis of oxo alcohols and of acetic acid from methanol and pure carbon monoxide.

The following are typical examples of commercial processes in which carbon monoxide in raw synthesis gas is partly or largely converted:

1) Adjustment to the desired hydrogen: carbon monoxide ratio in carbon monoxide-rich gases, which are formed, for instance, by partial oxidation (e.g., methanol synthesis);
2) Conversion of toxic carbon monoxide to carbon dioxide (e.g., town gas); and
3) Production of hydrogen from steam used for the carbon monoxide shift reaction (e.g., hydrogen production, ammonia synthesis).

5.1.1. Fundamentals

The water gas shift equilibrium among carbon monoxide, carbon dioxide, hydrogen, and steam

depends on temperature but is almost independent of pressure in the industrial range (elevated pressures up to 7.0 MPa).

Carbon monoxide shift conversion is an exothermic reaction:

$$CO + H_2O \rightleftharpoons H_2 + CO_2 \quad \Delta H_R^0 = -41.2\,\mathrm{kJ/mol}$$

The equilibrium constant is described by

$$K_p = \frac{p_{H_2} \cdot p_{CO_2}}{p_{CO} \cdot p_{H_2O}}$$

A number of models for the temperature dependency of the equilibrium constant and the heat of reaction appear in the literature. The temperature dependence of the equilibrium constant is shown in Chapter 1 (Fig. 1 and Table 3). Pressure influences activity and selectivity of the CO shift catalysts.

The various formulas, which are generally used for calculation of $K_p(T)$ and $\Delta H_R(T)$, give only slight differences for the temperature range of commercial processes. For the low-temperature range around 250 °C, the heat of reaction of shift conversion ΔH_R is ca. − 39.6 kJ/mol, it is about − 37.3 kJ/mol for the high-temperature range around 500 °C [139].

Above 950 – 1000 °C, the water gas equilibrium is established rapidly enough without a catalyst. In reforming processes using nickel-based catalysts, equilibrium is still attained at much lower temperatures. However, because nickel-based catalysts promote methanation of carbon oxides, they cannot be used for process steps in which only carbon monoxide shift conversion is desired.

The temperature ranges used in commercial carbon monoxide shift conversions are:

a) High-temperature shift (HTS): ca. 300 °C (start of run, SOR, with new catalyst) to 510 °C
b) Low-temperature shift (LTS): ca. 180 °C (SOR) to 270 °C
c) Raw gas shift (RGS): ca. 200 °C (SOR) to 500 °C.

Some catalyst manufacturers offer a so-called medium-temperature shift catalyst, which is normally a further stabilized LTS catalyst to operate at temperatures up to about 320 °C. Maximum temperature for *high-temperature shift* (HTS) is determined by the temperature resistance of the catalyst (sintering); minimum temperature, by the decrease in activity of this catalyst (large catalyst volumes are required). In *low-temperature shift* (LTS), the upper temperature limit is set by accelerated recrystallization of the copper catalyst (deactivation) above 270 °C; the lower limit by both the decline in catalyst activity and the need to keep the gas inlet temperature enough above the dew point of the gas – steam mixture to prevent condensation of water and subsequent damage to the catalyst. *Raw gas shift* (RGS) has a wide temperature range of application and the above explanations of temperature limits for HTS and LTS are similarly applicable for RGS.

In practice, at start of run equilibrium conversion can be reached whereas at end of run conditions typically an approach to equilibrium of about 40 °C is realized.

5.1.2. Catalysts

For carbon monoxide shift conversion in so-called clean gases (i.e., gases containing small quantities of impurities such as sulfur or condensable hydrocarbons) in the high-temperature range, iron – chromium oxide-based catalysts with trace additions of compounds containing sodium and potassium are used. At present HTS catalysts are often copper promoted. In the low-temperature range (LTS) copper – zinc – aluminum oxide-based catalysts with traces of potassium, etc. are commercially available [140, 141]. Some of these added elements act as promoters (enhancing activity, selectivity and providing self-guarding, i.e., chemically binding trace elements in the feed gas that would lead to catalyst poisoning) and some as stabilizers (increasing mechanical strength and active surface compositions at elevated temperatures and pressures) [142].

When new, catalysts often contain small amounts of sulfur. At present, the catalysts are sold mainly in tablet form and are stronger than the broken catalysts (iron oxide) formerly in use. In addition, their smooth surface and regular shape result in a smaller pressure drop through the catalyst bed.

High-temperature shift catalysts are supplied in the oxidized condition (i.e., iron is present as Fe_2O_3 and chromium as Cr_2O_3). Generally, the reduction necessary to activate the catalyst is not carried out before it is put into

the reactor, because the reduction process decreases the mechanical strength by up to 30 %. Therefore, the catalyst is reduced in situ with carbon monoxide- and hydrogen-laden gases, usually the clean gas itself. To control the reduction rate, the reducing gas is diluted with superheated steam or nitrogen (or both) to a mixture containing below 5 vol % ($CO + H_2$). During reduction the temperature profile of the catalyst bed has to be monitored carefully.

Iron oxide catalysts are resistant to appreciable amounts of impurities; even sulfur levels in the range of 10 to > 100 ppm (mL/m^3) can be tolerated at the cost of activity decrease. Rapid temperature and pressure changes caused, for example, by water coming into contact with the hotter catalyst (water evaporates in the pores) must be avoided. These maloperations lead to disintegration of the structure and, generally, to a rapid increase in pressure drop because of the accumulation of dust and pellet parts. All heating and cooling steps as well as pressure changes should be carried out gradually.

Low-Temperature Shift Catalysts. Stringent gas purity requirements must be met when using low-temperature shift catalysts because they are very susceptible to poisoning by sulfur or halogen compounds. Therefore, the gas must be cleaned to sulfur and chlorine contents of < 0.1 ppm (dry) upstream of LTS conversion. If this purity is not already met by the feed (e.g., from steam reforming of natural gas) or achieved by washing the raw gas (partial oxidation followed by Rectisol wash), a "guard bed" (usually a vessel filled with zinc oxide or LTS guard catalyst) must be installed upstream of the low-temperature conversion unit.

Raw Gas Shift Catalysts. Carbon monoxide shift conversion of gases containing appreciable amounts of sulfur or heavy hydrocarbons such as tar is known as raw gas shift conversion. For such gases catalysts consisting mainly of cobalt and molybdenum are used instead of the iron – chromium type. The activity of cobalt – molybdenum-based catalysts increases significantly between 4.0 and 8.0 MPa, whereas with HTS iron oxide catalysts, increasing the operating pressure above 4.0 MPa has practically no effect on the volume of catalyst required. RGS catalysts attain their full activity only in the sulfided condition. Therefore, they must be presulfided before installation or sulfided during start-up. A minimum content of sulfur compounds in the dry raw gas of ca. 100 to 1500 ppm (for gases from partial oxidation; the value depends on the steam: sulfur ratio) is required in normal operation to maintain catalyst activity. The minimum sulfur content in the feed gas also depends on other operating conditions (e.g., temperature). The sulfur content has no upper limit. The RGS offers the advantage of allowing sulfur and carbon dioxide removal to be accomplished in one step downstream the RGS unit [140].

The volume of catalyst required depends on the partial pressures of reacting components, the desired carbon monoxide content of the converted gas, and the reaction temperature. Normally, operating parameters are chosen to achieve space velocities of 1000 – 3000 m^3 (STP) of gas per cubic meter of catalyst per hour in the HTS range and 2000 – 5000 m^3 (STP) per cubic meter of catalyst per hour in the RGS and LTS range. If necessary, the process is divided into several steps with intermediate cooling or water quenching (see Section 5.1.5).

5.1.3. Clean Gas Shift Conversion

If the carbon monoxide-containing gas is substantially free of sulfur and heavy hydrocarbons, clean gas shift conversion is applied. In case the gas had to be desulfurized in an upstream purification stage, a cooler – saturator system is favorable. For naturally clean gases, (e.g., steam reformer outlet gas) only a simple HTS reactor is necessary.

Desulfurized feed gases derived from upstream partial oxidation and a physical or chemical washing unit usually enter the CO shift unit at a temperature between 20 and 120 °C. Thus, even if they are saturated with water vapor, the addition of a considerable quantity of high-pressure steam is required to effect satisfactory carbon monoxide shift conversion. For example, to lower the carbon monoxide content of a gas available at ca. 5.4 MPa and ambient temperature from 45 vol % (dry) to a residual ca. 3 vol % (dry) in two stages at a final reaction temperature of 380 °C, the addition of 1.18 kg of external HP-steam per cubic meter (STP) of gas is necessary. The cooler – saturator system, which generates steam in the saturator by using the reaction heat of the carbon monoxide shift con-

version, enables the quantity of external process steam to be lowered to 0.33 kg/m^3 (STP) of gas.

Cooler – Saturator System. Shift feed gas enters the saturator (a) at ambient temperature and rises through the tower packing while being sprayed with hot water (see Fig. 34). The gas becomes saturated with steam at ca. 225 °C. Heating of the gas – steam mixture is continued, more steam (0.33 kg/m^3 (STP) dry gas) is added, and the resulting mixture is fed to the first reaction zone (c). The temperature of the mixture rises due to the heat of reaction to about 500 °C. The first converter outlet CO concentration is about 6 to 8 vol % (dry). The hot gas is fed first to a gas – gas heat exchanger (b) to preheat the incoming mixture and then into a gas – water heat exchanger (e) to heat circulating water; by this means, it is cooled to ca. 350 – 360 °C. At this temperature, it enters the second reaction zone (d), where carbon monoxide is converted further to a residual content of ca. 3 vol % (dry), the gas being simultaneously heated to ca. 380 °C.

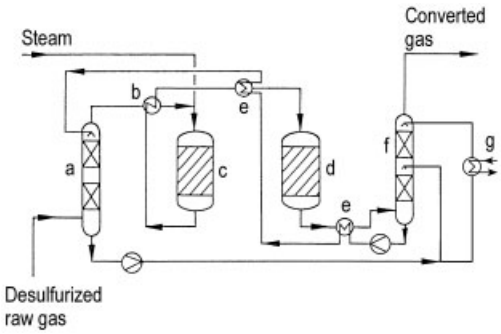

Figure 34. Flowsheet of a two-stage CO conversion plant with cooler – saturator system
a) Saturator; b) Gas – gas heat exchanger; c) Reactor 1; d) Reactor 2; e) Gas – water heat exchanger; f) Cooler; g) Waste Heat Recovery

Converted gas leaving the second reaction zone is cooled in a second gas – water heat exchanger (e). Then it enters the cooler (f) at ca. 230 °C. Here it is irrigated with colder water while it flows upward through a packed bed. By this means, most of the steam condenses, and water leaving the cooler is heated approximately to the dew point of the converted gas. The water leaves the cooler at ca. 210 °C. The cooler is often divided into two sections: the bottom section is fed directly with most of the water leaving the saturator; the top section is fed with a side stream that has been cooled to ca. 50 °C. The heat thus removed is used e.g., for boiler feedwater preheating, demineralized water preheating, etc. (g). By using such an arrangement, the gas can be cooled to about 60 °C.

Water from the cooler is heated to ca. 250 °C in the two gas – water heat exchangers (e) before entering the saturator (a), which it leaves at ca. 160 °C. Alternatively to the shown arrangement, in larger plants, the first gas – gas heat exchanger (b) can be installed inside reactor 1, to avoid the hot transfer piping.

Countercurrent Cooling Reactor. In smaller plants, a countercurrent cooling reactor (Fig. 35) is often used instead of two reaction steps in series. Whereas the reaction in a simple fixed-bed reactor proceeds adiabatically, which necessitates the two-stage mode of operation, a much more favorable temperature profile is achieved in the countercurrent cooling reactor.

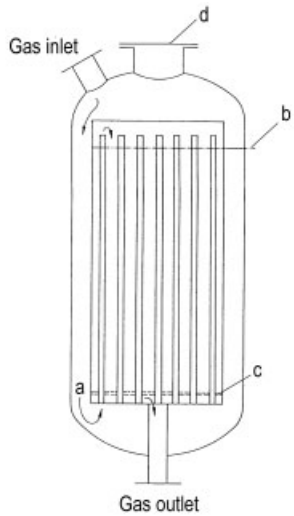

Figure 35. Countercurrent cooling reactor
a) Gas entry into the cooling tubes; b) Surface of catalyst bed; c) Catalyst support grid; d) Manhole

The gas – steam mixture, which is not preheated to catalyst inlet temperature, enters the annular space between reactor shell and catalyst vessel, flows downward through this space, and enters the tubes of the catalyst vessel. While flowing through these tubes which are surrounded with catalyst, the mixture is heated

to ca. 380 °C. It then enters the catalyst bed from above. Due to the interaction between the temperature rise caused by the heat of reaction and the cooling effect of the gas mixture flowing through the cooling tubes, the temperature curve illustrated in Figure 36 develops. This type of reactor combines a high reaction rate at high temperature in the top section with a favorable equilibrium in the bottom section.

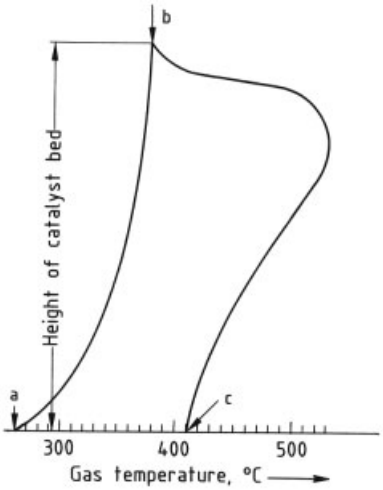

Figure 36. Temperature profile in countercurrent cooling reactor
a) Gas entry into cooling tubes; b) Gas entry into catalyst; c) Gas outlet (denotations a to c correspond with Fig. 35)

Carbon Monoxide Shift Conversion of Steam Reformer Outlet Gas. Shift conversion can be carried out more easily on product gases of a steam reformer for the following reasons: (1) the content of carbon monoxide is far lower (ca. 12 – 15 vol % in dry gas) compared to syngas produced by partial oxidation or coal gasification; that means generation of less reaction heat in CO conversion; (2) the required steam is already present in the steam reformer outlet gas, i.e., the gas has only to be cooled in the downstream waste-heat boiler to the inlet temperature of the shift converter (ca. 300 – 350 °C); (3) the gas is essentially free of catalyst poisons, as the reformer feed has to be purified before, because of the sensitivity of the reformer catalyst.

An example is the production of synthesis gas for ammonia manufacture from natural gas. After reforming in the primary and secondary reformer, and initial cooling in a waste-heat boiler, the reformed gas is, e.g., available at 2.9 MPa and 350 °C, and a steam content of 0.64 kg/m^3 (STP). It contains 10.5 vol % (dry) carbon dioxide, 11.7 vol % (dry) carbon monoxide, and 55.5 vol % (dry) hydrogen, the remainder being nitrogen and methane. After passing through the HTS conversion step, a carbon monoxide content of 2.2 vol % (dry) is obtained at 405 °C. The gas, which still contains 0.52 kg of steam per cubic meter (STP), is cooled to 220 °C and then fed to the LTS step. There the carbon monoxide content is decreased to ca. 0.3 vol % (dry), while the temperature increases to 232 °C.

Shift conversion of these gases is carried out either in one step in a HTS converter or in two steps, with intermediate cooling, in HTS and LTS converters. More recently quasi isothermal LTS reactors have been installed, reaching low CO contents at their outlet in a single step [143]. As a byproduct a significant amount of methanol is normally produced [141, 144].

5.1.4. Raw Gas Conversion

Raw gas conversion (sometimes also termed "sulfided CO shift") denotes the conversion of gases containing significant quantities of sulfur, soot, or condensable hydrocarbons. There are two principal industrial applications: (1) shift conversion of gases from heavy-oil gasification that have not been cooled and desulfurized but have only undergone hot quenching to add the required steam and remove soot; and (2) shift conversion of the carbon monoxide in gases from coal pressure gasification that contain not only sulfur but also saturated and unsaturated hydrocarbons, including tars. Cobalt – molybdenum based catalysts have proved particularly effective for converting both types of gas.

The carbon monoxide content (ca. 45 vol %, dry) of a gas from partial oxidation of heavy oil enters the carbon monoxide shift reactor at about 250 °C with a steam: dry gas volume ratio of ca. 2.4. The carbon monoxide content can be reduced to ca. 1.6 vol % (dry) over a cobalt – molybdenum based catalyst. Residual carbon monoxide contents of 0.5 – 1.0 vol % (dry) are attainable in a two-stage system even at lower steam: dry gas ratios. The sensible heat of the product gas is used to generate steam and to

preheat feedwater. Impurities in the gases for raw gas conversion, especially formic acid necessitate the use of mostly austenitic materials for streams with temperatures below the water dew point. Carrying out shift conversion on raw gases produced by coal pressure gasification is a good way of "detoxifying" town gas. The amount of steam present in the gas at 3.0 MPa and 180 °C, ca. 0.6 kg/m^3 (STP), is sufficient to reduce the carbon monoxide content from, e.g., 15.5 to ca. 5.0 vol % (dry). The cobalt – molybdenum based catalyst has a very useful side effect: unsaturated hydrocarbons in the gas are hydrogenated almost completely. Parallel to the CO shift conversion reaction COS is converted to H_2S; i.e., COS contents in the shifted gas reach low values close to equilibrium at catalyst outlet conditions.

5.1.5. General Comments on Reactor Arrangements

Apart from one-reactor applications several two-reactor systems are of commercial interest: two HTS (or RGS) reactors in series or one HTS followed by one LTS reactor, in each case with intermediate cooling. In the latter case a "guard bed" filled with zinc oxide is often installed between the HTS and LTS reactors to ensure that gas entering the LTS step is largely sulfur and halogen-free. Alternatively, the "guard bed" is filled with a smaller amount of LTS guard catalyst. Some HTS processes include water quenching between two reactor beds to cool the gas on the one hand and to add steam as a reactant on the other hand. In these concepts special caution has to be taken to avoid wetting of the catalyst.

5.1.6. Noncatalytic Quench Conversion

A less frequently used method of carbon monoxide shift conversion on raw gases is noncatalytic quench conversion. Here, advantage is taken of the fact that the water gas equilibrium is attained fairly rapidly above ca. 950 °C, even without catalysts. If, for example, oxo synthesis gas with a hydrogen: carbon monoxide volume ratio of 1.25 is to be produced from raw gas obtained by partial oxidation of asphalt and containing 48.9 vol % (dry) carbon monoxide and 43.9 vol % (dry) hydrogen, then 0.4 kg of cold water per kilogram of asphalt is sprayed into the hot (ca. 1570 °C) gas. As a result of cooling caused by the heating and evaporation of water and the associated carbon monoxide shift conversion (heat of reaction), the water gas equilibrium is attained at 1250 °C. The carbon monoxide content is then 38.3 vol % (dry) and the hydrogen content 47.8 vol % (dry), with a H_2: CO ratio as desired.

5.2. Carbonyl Sulfide Conversion

In all gasification processes based on sulfur-containing feedstocks, carbonyl sulfide is produced as well as hydrogen sulfide. For example, by partial oxidation of vacuum residues with a sulfur content of 3 wt %, a raw gas containing 0.71 vol % (dry) of hydrogen sulfide and 0.03 vol % (dry) of carbonyl sulfide is formed. Many modern gas washing processes remove carbonyl sulfide together with hydrogen sulfide. A selective carbonyl sulfide conversion (i.e., without parallel carbon monoxide conversion in a HTS or RGS converter, see Section 5.1) is limited to applications where there is no need for deep desulfurization (e.g., fuel gas for gas turbines).

Conversion of organic sulfur compounds in feedstocks for tubular reforming by hydrogenation over cobalt – molybdenum and nickel – molybdenum catalysts is described in Section 2.2.2.

Hydrogenation of carbonyl sulfide over iron oxide and cobalt – molybdenum based (see Section 5.1) catalysts occurs simultaneously with carbon monoxide shift conversion, and equilibrium in the form

$$H_2 + COS \rightleftharpoons H_2S + CO \quad \Delta H_R^0 = +10.9\,\mathrm{kJ/mol}$$

being established. In the CO shift conversion range at high temperatures of 350 – 500 °C, the equilibrium and the kinetics of this reaction lead to high COS conversion rates. When the gas mentioned above undergoes carbon monoxide shift conversion, the concentration of carbonyl sulfide in the product gas typically reaches values below 10 ppm (dry). (Table 3 lists a number of values of the equilibrium constant for this reaction in relation to temperature.)

A selective transformation of carbonyl sulfide to hydrogen sulfide is required in gases in which carbon monoxide shift conversion is not

desired. Catalysts have been developed for the selective conversion of carbonyl sulfide to hydrogen sulfide which are inactive for the water gas shift reaction. These catalysts promote hydrolysis of carbonyl sulfide according to

$$COS + H_2O \rightleftharpoons H_2S + CO_2 \quad \Delta H_R^0 = -30.2\,\mathrm{kJ/mol}$$

This reaction is catalyzed by chromium – aluminum oxide-based promoted catalysts, which work in a sulfided state. The commercially available catalysts are mainly not subject to poisoning by heavy metals and arsenic; but halogen compounds will lead to reduction of activity, selectivity, and lifetime. These catalysts are used at 150 – 320 °C; satisfactory equilibrium is attained at space velocities between 500 and 2000 m^3 (STP) per cubic meter of catalyst per hour [145]. As at lower temperatures the equilibrium conditions favor COS hydrolysis, less steam in the feed gas is needed to reach desired hydrolysis rates. (Values for equilibrium constants as a function of temperature are shown in Table 3.)

A typical reduction of the carbonyl sulfide content in gases from partial oxidation [assumed gas compounds (dry): 45 vol % H_2, 45 vol % CO, 5 vol % CO_2, 0.71 vol % H_2S, 300 ppm (mL/m^3) COS] to about 10 ppm is achievable by COS hydrolysis at 170 °C. A steam content of 0.086 kg per cubic meter (STP) of gas in this example corresponds to a saturation temperature of 150 °C at 5.0 MPa. In parallel to COS hydrolysis the contents of CS_2 and HCN are also significantly reduced. For some applications a separate bed for hydrogenation/hydrolysis of these compounds has to be installed additionally. For gases with low total sulfur content there are zinc oxide-based "absorption catalysts" commercially available, which can hydrolyze COS in the presence of steam (low concentrations) before H_2S is absorbed in the same bed.

5.3. Methanation and Methane Synthesis

5.3.1. Definitions and Applications

Methanation is generally defined as the catalytic reaction of hydrogen with carbon monoxide and/or carbon dioxide, forming methane and water, according to the description given in Section 5.3.2. Two main applications of methanation exist, which require very different technical solutions:

1) Methanation for the elimination of carbon oxides in hydrogen production: upstream of this final purification step the concentrations of harmful carbon oxides are rather low and the exothermic methanation reaction therefore causes only a moderate temperature increase, i.e., a single methanation step is sufficient (see Section 5.3.3).
2) Methanation for the production of SNG (Substitute Natural Gas): in this application, also called methane synthesis, hydrogen shall be converted to methane as completely as possible. Due to the strongly exothermic reaction and the high concentrations of all reactants measures have to be taken to limit temperature rises. As also the thermodynamics, i.e., the relevant equilibrium, ask for relatively low temperatures to yield high methane and correspondingly low hydrogen contents in the product, usually multiple-stage reaction systems will be required (see Sections 5.3.4 and 5.3.5).

Application (1) must be regarded as a well established technique with little incentive to improvements. It is widely used in hydrogen production, although nowadays pressure-swing adsorption (PSA) systems for hydrogen purification dominate the market. Catalytic systems may be damaged by too high temperatures due to breakthrough of carbon oxides as a result of sudden malfunctions of the upstream purification systems.

Application (2), the production of SNG, has lost commercial interest at present due to the increased availability of natural gas. The price of SNG cannot compete with that of natural gas. The development of new techniques for this kind of plants has therefore been terminated.

SNG production plants have demonstrated the reliability of processes for conversion of naphtha to SNG (Baltimore, USA) or coal to SNG (North Dakota, USA). Special local conditions — such as isolated distribution systems or missing storage capabilities — may be the reason for temporary implementation of SNG plants (for instance in West-Berlin before the German reunification, where only naphtha could

be stored at the time). A methanation system at low temperatures (with ruthenium-based catalysts instead of normally applied nickel catalysts) may find application in space technology [146]. In this projected application carbon dioxide from the breathing air is methanated in combination with a water electrolysis, thereby regaining oxygen for the biological cycle.

5.3.2. Principles of Methanation

The principle of catalytic synthetic production of methane from carbon monoxide and hydrogen was discovered in 1902 by Sabatier and Senderens. It is described by

$$CO + 3\,H_2 \rightleftharpoons CH_4 + H_2O$$

Carbon dioxide can also be converted to methane according to

$$CO_2 + 4\,H_2 \rightleftharpoons CH_4 + 2\,H_2O$$

Both reactions are linked by shift conversion, which is always observed simultaneously whenever active catalysts are used:

$$CO + H_2O \rightleftharpoons CO_2 + H_2$$

Many observations indicate that the transformation of carbon dioxide to methane is initiated by a reverse shift conversion reaction with hydrogen to yield carbon monoxide and steam. The carbon monoxide formed then reacts to yield methane:

$$\begin{array}{l} CO_2 + H_2 \rightleftharpoons CO + H_2O \\ CO + 3H_2 \rightleftharpoons CH_4 + H_2O \\ \hline CO_2 + 4H_2 \rightleftharpoons CH_4 + 2H_2O \end{array}$$

Both reactions for the formation of methane are strongly exothermic, as shown by Table 26. This table also indicates that low temperature and high pressure are required or favorable for high methane yield.

Kinetics. The mechanism and kinetics of methanation have been investigated, but these studies do not yield uniform results. The rate of conversion of carbon monoxide is often described formally by

$$r_{CO} = k_1 \cdot p_{CO}^m \cdot p_{H_2}^n$$

with $m = 0.7 - 1.0$ and $n = 0.3 - 0.5$.

Table 26. Equilibrium constants and heats of reaction (according to [147] for

1. $CO + 3\,H_2 \rightleftharpoons CH_4 + H_2O$, $K_1 = p_{CH_4} \cdot p_{H_2O}/p_{CO} \cdot p_{H_2}^3$
2. $CO_2 + 4\,H_2 \rightleftharpoons CH_4 + 2\,H_2O$, $K_2 = p_{CH_4} \cdot p_{H_2O}^2/p_{CO_2} \cdot p_{H_2}^4$

Temperature, K	Reaction 1		Reaction 2	
	K_1, bar^{-2}	ΔH_R, kJ/mol	K_2, bar^{-2}	ΔH_R, kJ/mol
400	0.409×10^{16}	−210.828	0.276×10^{13}	−170.172
500	0.115×10^{11}	−214.722	0.867×10^{8}	−174.862
600	0.198×10^{8}	−217.984	0.730×10^{5}	−179.070
700	0.372×10^{4}	−220.667	0.413×10^{3}	−182.771
800	0.321×10^{2}	−222.807	0.794×10^{1}	−185.953
900	0.766×10^{0}	−224.465	0.347×10^{0}	−188.658
1000	0.377×10^{-1}	−225.692	0.274×10^{-1}	−190.894
1100	0.319×10^{-2}	−226.609	0.338×10^{-2}	−192.816

In the purification of synthesis gas with a large excess of hydrogen, the reaction rate law is simplified to

$$r_{CO} = k_2 \cdot p_{CO}^m$$

Summaries and details can be found in the literature; see [148 – 150].

Kinetics must also be considered when investigating the possibility of undesirable side reactions. Kinetic effects have to be used to avoid the formation of carbon, because methanation usually begins in the temperature region where suitable thermodynamic conditions prevail for the Boudouard reaction to occur:

$$2\,CO \rightleftharpoons C + CO_2$$

(see Section 1.3). If thermodynamic equilibrium calculations indicate a possible buildup of carbon in the outlet region of the reactor, this undesirable side reaction cannot be excluded and may call for changes in process conditions. At the outlet zone of methanation reactors, attention must also be paid to the methane decomposition reaction

$$CH_4 \rightleftharpoons C + 2\,H_2$$

(see also Section 1.3.). However, a small residual hydrogen content and moderate temperature normally prevent this reverse reaction.

Catalysts. For the removal of carbon oxides from synthesis gases, catalysts with usually < 15 wt % nickel are used predominantly. Activated alumina, magnesium oxide and mixtures thereof are widely used as carrier materials. For the synthesis of methane from gases rich in carbon monoxide, catalysts with a high nickel content, similar to those used in reforming naphtha to a methane-rich gas, are usually employed. Catalyst poisons, must be eliminated carefully to the same level as in naphtha reforming (see Section 2.4.4) or even lower, because deactivation with time depends primarily on such impurities in the synthesis gas. Catalysts of the naphtha reforming type usually reach their limit of operation at about 500 °C. High-nickel catalysts with special aluminum oxide carriers and stabilization by zirconium oxide, which are capable of operating at 300 – 700 °C, are also available. The catalysts are used preferably in the form of cylindrical tablets measuring 5 × 5 or 8 × 8 mm.

Catalysts based on ruthenium have been tried repeatedly for methanation but have not found the broad application of nickel-based catalysts. Raney nickel coated reaction tube systems are also of little interest, because replacement of the spent catalyst appears complicated.

Metal Carbonyl Formation. Precaution must be taken with nickel-containing catalysts to prevent formation of highly poisonous *nickel carbonyl* [$Ni(CO)_4$] (TLV 0.1 ppm (mL/m^3) or 0.7 mg/m^3; see [151]).

In particular, low temperatures must be avoided because they favor the formation of carbonyls, as shown by Table 27. Generation of carbonyls is favored by nickel in reduced form, high total pressure, or high partial pressure of carbon monoxide [152]. The reaction is also promoted by sulfur, selenium, and tellurium. In practical operation of methanation plants, temperatures ≤ 200 °C at the nickel catalyst must be avoided, and procedures for start-up and shutdown must be adjusted accordingly. Formation of *iron carbonyl* also has to be prevented because of its toxicity and general corrosion problems, and because iron carbonyl decomposes and forms iron layers on the active surface of the catalyst as the temperature increases. Thus, carbon monoxide-laden synthesis gases must be heated in stainless steel heat exchangers. Many years of plant operation have shown that with adequate precautions, the formation of carbonyls can be suppressed successfully.

Table 27. Calculated nickel carbonyl content at equilibrium for various carbon monoxide concentrations as a function of temperature for a total pressure of 1.48 MPa (according to [152]), ppm (mL/m^3)

Temperature,	CO content, mol%				
°C	0.2	0.5	1.0	2.0	3.2
66	0.3	12	190	3 000	20 000
93	0	0.2	3	49	320
121	0	0	0.1	1.6	11
149	0	0	0	0.1	0.5
177	0	0	0	0	0.03
204	0	0	0	0	0

5.3.3. Methanation as a Step in Hydrogen Purification

Carbon oxides are often poisonous to *catalytic hydrogenation* processes. Their simultaneous elimination is possible by means of a methanation process step, which has been used widely since the 1950s and is nowadays competing with pressure-swing adsorption processes. Methanation remains the normal purification method for *ammonia synthesis gas* production, especially by secondary air reforming. Here, the nitrogen content prevents the use of other methods.

As shown in Figure 37, the inlet gas, containing ca. 0.5 – 0.8 mol % of carbon monoxide and the same amount of carbon dioxide, is heated by methanated gas in a heat exchanger (b). The final inlet temperature to the adiabatic reactor — typically 300 – 330 °C — may be achieved by a fired trim heater (c). In the adiabatic shaft reactor (a) filled with nickel catalyst, carbon monoxide and carbon dioxide are converted with hydrogen to methane and water vapor. Because the temperature increase in the reactor is ca. 70 K for each percent of carbon monoxide and about 55 K for each percent of carbon dioxide, the resulting outlet temperature can reach ca. 400 °C and the temperature gradient is then large enough to shut off the trim heater in continuous operation. Instead of the fired trim heater, other arrangements of heat exchangers in combination with the reforming plant are possible, as well as electrically operated start heaters. Purified hydrogen is then cooled to ambient temperature in the heat exchanger (b) and a final cooler (d), which also condenses most of the vapor formed.

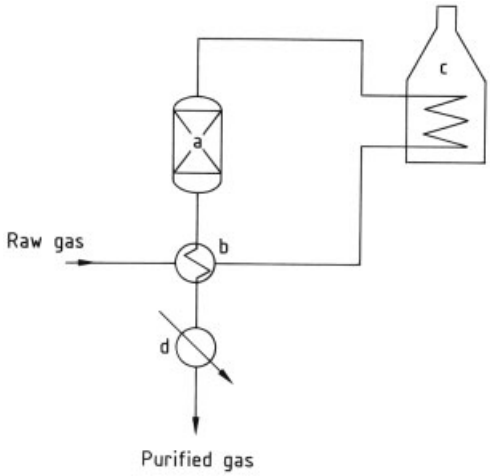

Figure 37. Methanation as a purification step for hydrogen or synthesis gas
a) Reactor; b) Heat exchanger; c) Fired trim heater; d) Cooler

In synthesis gas purification of this type, the total content of carbon oxides is limited to ca. 2 vol %; otherwise, the temperature limits of the catalyst would be exceeded and the final content of methane in the hydrogen would be too high. Because of the loss of hydrogen by methanation, carbon oxides contents no greater than 0.8 % total are aimed at. Additionally, the residual content of carbon oxides is increased by higher outlet temperature and thus will not remain below the accepted tolerance of ca. 10 ppm (mL/m^3) for the sum of carbon monoxide and carbon dioxide.

5.3.4. Methanation of Rich Gas

Although methanation as a method of purifying synthesis gas yields only small absolute quantities of methane, methanation of rich gas and methane synthesis from gases high in carbon monoxide (e.g., produced by coal gasification) intentionally generate methane as product. Production of SNG by gasification of coal is being considered as a long-term replacement for gaseous hydrocarbons of natural origin when the latter are no longer available in sufficient quantity. In contrast, production of methane from rich gas (see Section 2.4.4) is restricted to specific areas where naphtha is abundant but no natural gas is available.

When SNG is produced from *naphtha*, the methane content of the rich gas is typically ca. 70 vol % (see Fig. 10, Section 2.4.4), whereas hydrogen amounts to ca. 7 % and carbon dioxide 25 %. Less than 0.5 % carbon monoxide is present, and unreacted steam from naphtha gasification amounts to ca. 1.30 mol of H_2O per mole of dry gas. Such a gas mixture can be enriched further by admitting it at an inlet temperature around 300 °C (i.e., after cooling downstream of gasification) to an adiabatic shaft reactor containing a nickel catalyst identical or similar to that used in naphtha reforming. Because the portion of unreacted steam is quite high and further steam is formed during the reaction, the degree of methane formation is limited in this so-called *wet methanation step*.

After cooling of the gas and condensation of steam, the methane content can be increased further by another methanation step (*dry methanation*), which may be up- or downstream of the required carbon dioxide removal. The *upstream* arrangement yields better methane quality, but the danger of carbon formation is higher due to the thermodynamic possibility of a reverse shift reaction and consequent formation of carbon according to the Boudouard reaction. Installation of the dry methanation step *downstream* of the carbon dioxide scrubbing system allows milder reaction conditions, but the carbon dioxide and hydrogen content of the product is increased slightly because of the need to provide excess carbon dioxide for the reaction of hydrogen.

A typical *analysis* of the final SNG is: methane 98.5 vol %, carbon dioxide 0.5 vol %, hydrogen 1.0 vol % and carbon monoxide < 0.1 vol %.

Figure 12, Section 2.4.4, shows a one-stage methanation system for rich gas installed upstream of the CO_2 removal system as part of the total plant concept for the production of SNG from naphtha.

5.3.5. Methane Synthesis from Gases with High Carbon Monoxide Content

Synthesis gases high in carbon monoxide are generated by various coal and oil gasification processes. These processes naturally result in different qualities of synthesis gas according to the feedstock used, the upgrading processes for

synthesis gas preparation, and the gasification process itself. Whereas high-temperature gasification processes yield high concentrations of carbon monoxide and little methane, interest in SNG production is concentrated on gasification processes that yield high methane content in the raw gas. This is possible, for instance, with Lurgi pressure gasification of coal, especially when gasification pressures of 8.0 – 10.0 MPa are applied, which has been successfully tested (see Section 4.4). Methanation processes with little methane in the raw gas suffer principally from (1) high exothermic heat release during methanation, (2) the very large quantities of synthesis gas that must be handled (four to five volumes of dry synthesis gas yield one volume of methane), and (3) the high proportion of steam formed during methane synthesis, which limits the directly achievable SNG quality in wet methanation steps.

Synthesis gases for methane synthesis are often classified by their stoichiometric number (SN), i.e., the ratio of the reactable hydrogen volume to the pro rata volume of carbon oxides:

$$\mathrm{SN} = \frac{\eta \cdot v_{\mathrm{H_2}}}{3v_{\mathrm{CO}} + 4v_{\mathrm{CO_2}}}$$

where η stands for an equivalent hydrogen consumption required for the hydrogenation of unsaturated and higher hydrocarbons in producing methane.

Because SNG specifications demand a hydrogen content of <10 vol %, a general classification may be SN $\leq$ 1.05. Because of restrictions in carbon dioxide content as well, the synthesis gas must meet the condition $0.98 < \mathrm{SN} < 1.03$ for stoichiometric methane synthesis, which is suitable for direct production of SNG. Such stoichiometric methane synthesis allows synthesis gas to be prepared by shift conversion and carbon dioxide removal. After methane synthesis, only compression to pipeline pressure (if the gasification pressure is not high enough) and drying (e.g., by a glycol system) are necessary. The quality of SNG varies slightly because fluctuations of the stoichiometric number are inevitably caused by variables and control functions of the gasification and upgrading processes.

In stoichiometric methane synthesis, little risk exists of carbon formation according to the Boudouard reaction, even in high-temperature methanation up to 700 °C. These problems become more serious at lower SN numbers. Lower stoichiometric numbers always require carbon dioxide removal as the final step before drying. The allowable load of carbon monoxide for the stoichiometric number of the synthesis gas must be determined experimentally for each gasification process. Individual optimization is necessary to attain lowest product costs, because shift conversion may be expensive if the necessary vapor is not provided with the raw gas. Among a vast variety of process configurations developed in the 1970s, the following are reportedly available for commercial operation. However, only the process with recycle has demonstrated reliability under the harsh conditions of commercial production, due to the lack of interest in SNG.

Besides the stoichiometric number, the highly exothermic reaction generally creates a problem for the design of methane synthesis plants: either the temperature increase must be limited by recycling of reacted gas or steam dilution, or special techniques such as isothermal reactors or fluidized beds, each with indirect cooling by evaporating water, must be used.

Recycle Methane Synthesis. Figure 38 shows a schematic of a methane synthesis plant with *hot recycle*. The type shown is designed as a two-stage recycle system with a final countercurrent-cooled reactor. These plants were originally designed for a recycle quantity such that 500 °C is not exceeded at the outlet of the adiabatic reactors. Current systems, optimized with regard to low secondary energy requirements, are available for maximum temperatures up to 650 °C. Inlet temperatures are ca. 300 °C, which is also the operating range of the recycle compressor. An advantage of the hot recycle is that water vapor formed during the methanation reaction is not condensed and, therefore, the stoichiometric number is not changed for the stages processing feed gas. Recycle systems show a high flexibility for accepting synthesis gases of very different natures [153, 154]. The waste-heat recovery system is normally designed to generate saturated steam up to 10.0 MPa, but superheated steam may also be withdrawn. Replacing the recycle compressor by a gas ejector reportedly produces positive results [155].

The higher the temperature selected for the first stage of methane synthesis, the more stages

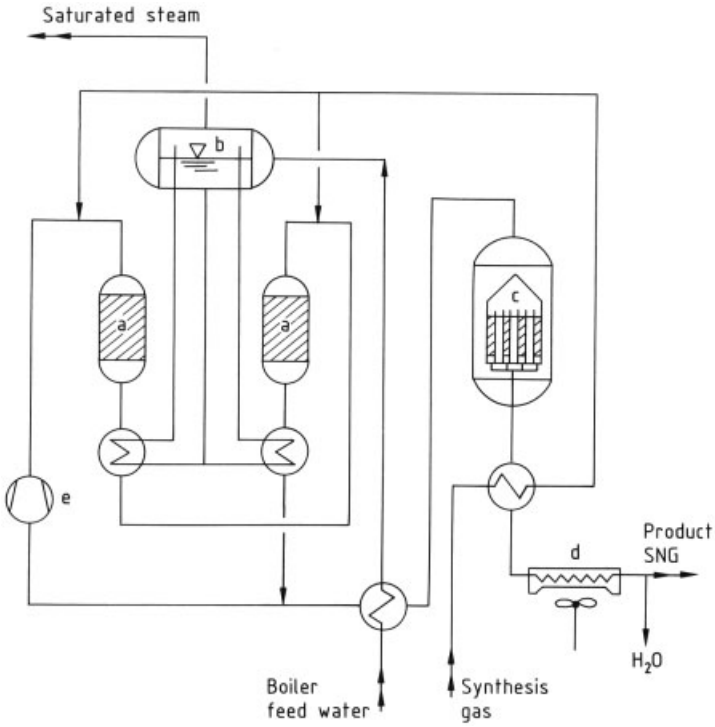

Figure 38. Recycle methane synthesis
a) Recycle reactors; b) Steam drum; c) Countercurrent cooling reactor; d) Air-cooled cooler; e) Hot recycle compressor

must be applied to reach specification-grade SNG, for thermodynamic reasons. Therefore, with high-temperature methanation in the recycle stage(s), an intermediate step without feed gas addition at ca. 300 – 460 °C is required upstream of a final dry or wet methanation stage. The countercurrent cooling reactor in Figure 38, has a similar design to that described in Section 5.1.3 (Fig. 35). It allows product quality to be reached in one final wet step; otherwise, more stages are required to achieve the same analysis.

Steam Quenching Methane Synthesis. For synthesis gases with low stoichiometric numbers (low H_2: CO ratio), steam must always be added to prevent carbon formation. For optimum results, the required quantity of steam is mixed with a proportion of the feed gas in such a way that the temperature increase in the first methanation stage is kept within allowable limits and, after heat removal, a further portion of the feed gas is admitted to the first intermediate product gas stream as quenching stream. The quenching procedure after cooling is repeated until all the feed gas is consumed. Thus, a multistage concept results, depending on the inlet gas composition, as shown in Figure 39 for the RM process [156]. With the addition of steam, the equilibrium composition of any intermediate product gas is not favorable enough to enable direct production of specification-grade SNG in a wet methanation step. The removal of carbon dioxide, therefore, is followed by a final, dry methanation stage.

Fluidized-Bed Methanation. Intensive research and demonstration have been reported for the methanation of feed gases of unfavorable composition, with little or no addition of steam, in a cooled fluidized-bed reactor in one step without an additional cleanup reaction (Comflux process) [157]. The dangers of carbon formation, catalyst deactivation by high carbon monoxide partial pressure, and excessive catalyst consumption by the fluidized-bed process are said to have been overcome, and operation up to 6.0 MPa has been demonstrated successfully in the pilot plant.

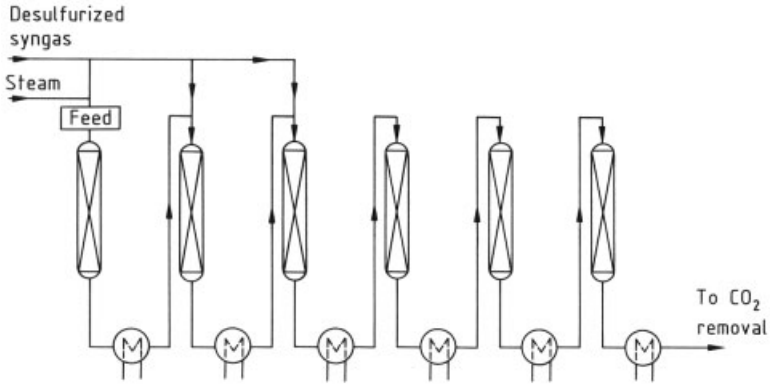

Figure 39. Flow principle of multi stage quenching system – RM process

Isothermal Operation. By increasing the allowable temperatures for methanation catalysts, methane synthesis can be performed by a once-through method in quasi-isothermal reactors cooled by evaporating water, generating saturated steam. Under favorable conditions, such systems produce specification-grade SNG in only one catalytic step (IRMA process), but superheated steam can be produced only by combining isothermal and adiabatic reactors [158]. Any traces of sulfur compounds must be eliminated carefully, because poisoning of the catalyst will displace the temperature profile in the reactor and the reaction will not go to completion as the catalyst lifetime expires. Tests with isothermal reactors have been performed for methane synthesis within the ADAM – EVA systems for energy transport [159].

The ADAM – EVA Concept (Long-Distance Energy Transport System). Generally, methane synthesis is used to generate methane, the heat of reaction being an inevitable byproduct. However, a proposal by the Jülich Nuclear Research Center (KFA-Jülich, Germany) and the Rheinische Braunkohlenwerke (Cologne, Germany) uses the energy release of the methanation reaction as the essential product. The energy demand for the steam reforming of methane is provided by a nuclear high-temperature reactor; the resulting synthesis gas is then retransformed into methane a few hundred kilometers away, thereby providing thermal energy up to 650 °C. The methane is routed back to the nuclear system by a second pipeline to restart the cycle; thus, no fossil energy is used up.

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2$$

Endothermic Exothermic

By the combination of power generation and energy distribution at a level of ca. 150 °C, an overall efficiency of about 60 % can be reached for such a chemical heat pipe system. A general flow diagram of this process is given in [160]. The use of high-temperature solar heating instead of nuclear power has been proposed for steam reforming of methane, however, solar ADAM – EVA systems may be restricted to a few favorable locations. Plans have been reported to provide energy that has been recovered and converted by high-temperature reforming in the Negev Desert to the industrial zones of Israel.

Studies have been made on the replacement of steam in the reforming step by carbon dioxide as oxidant [161]:

$$CH_4 + CO_2 \rightarrow 2\,CO + 2\,H_2$$

Although this reaction sequence significantly reduces the load in steam generation and can improve overall efficiency, transport of the thermodynamically necessary excess carbon dioxide in pipelines for both directions reduces this advantage. In addition, the danger of carbon formation in both reforming and methane synthesis at relatively high pressure must be examined carefully.

Other Synthesis Processes.

Methanation of Raw Gases. The dry product gas volume from methanation is only 20 – 35 % of the synthesis gas volume, depending on the gasification process and its yield of primary methane. Because nickel catalysts require complete removal of poisonous sulfur compounds,

relatively large amounts of gas must be handled in the upstream treating units. The Gas Research Institute in Chicago has proposed direct methanation of raw gases containing sulfur and carbon dioxide on special catalysts, to reduce gas volumes in acid gas removal. Savings in overall steam balance and in capital investment are claimed for such process routes. A final methanation step for cleanup in a sulfur-free atmosphere yields specification-grade SNG.

Liquid-Phase Methanation. A proposal by Chem Systems, Hackensack, N.J., is based on conventional gas preparation with methanation in the liquid phase [162]. Similar to processes employing fluidized-bed methanation or isothermal reactors, liquid-phase methanation appears to be able to handle high carbon monoxide concentrations and to yield high-quality methane in a single reactor. An inert liquid (aliphatic oil) containing dispersed catalyst is circulated and the feed gas is passed through this. The exothermic reaction heat is taken up by the liquid, mainly as sensible heat and partly by vaporization, resulting in excellent temperature control.

5.4. Absorption Processes

5.4.1. General

Principles. Processes that use a liquid solvent to remove one or more compounds from a gas stream are called absorption processes [140, 163 – 165] (→ Absorption).

The main purpose of absorption processes in gas production is the removal of acid gas components. They can also be used successfully to purify gases containing low concentrations of other components such as ammonia, higher hydrocarbons, water, cyanide, and organic sulfur compounds. Absorption processes are further applied to the withdrawal of gas constituents poisonous to catalysts or to the environment and also for the recovery of liquid gas or gasoline components from refinery gases.

In electric power generation from coal or liquid hydrocarbons by the combined power cycle process (see Section 7.3), absorption is used to selectively remove sulfur compounds and other impurities such as cyanide and mercury from the gas in order to meet gas turbine requirements and environmental regulations. So-called zero-emission IGCCs are discussed in view of CO2 emissions regulations and greenhouse gas abatement. Carbon of the feedstock is converted to carbon dioxide, which is removed in the absorption process and must be concentrated in a pure carbon dioxide stream for sequestration (see Fig. 59).

In the absorber, gas and liquid are brought into contact countercurrently. The solvent removes one or more components from the raw gas, more or less selectively. The absorber is usually equipped with trays or packing to provide a large gas – liquid contact area. The types most frequently used are valve and sieve trays. Bubble cap or tunnel trays are used only in special cases (a comprehensive treatment of trays and packings used in absorption is presented elsewhere; → Absorption).

The solvent laden with the absorbed components is withdrawn from the bottom of the absorber, routed to a regeneration system, where it is freed of absorbed gas components, and returned as lean solvent to the absorber. In most gas treating processes, absorption is reversible, and dissolved components are released chemically unchanged during regeneration. Liquid-phase oxidation processes for hydrogen sulfide removal differ in this respect (see Section 5.4.3).

The different absorption processes are characterized as *physical* or *chemical* depending on whether gas components are simply dissolved physically or are bound chemically to the solvent. Typical equilibrium lines for chemical and physical absorption are shown in Figure 40. Loading capacity c (kilomoles per ton or per cubic meter of solvent) is represented as a function of the partial pressure of the dissolved component. The loading in physical absorption, initially following Henry's law, is almost directly proportional to the partial pressure in the gas phase, which results in a nearly linear equilibrium line. The equilibrium line in chemical absorption is bowed sharply during saturation of the chemically active solvent component. After saturation of the chemical capacity of the solution, only weak physical absorption in the solvent's main component (usually water) is possible which results in a steep linear equilibrium curve.

At low partial pressure (e.g., p_2 in Fig. 40) the absorption capacity c_{ch} of the chemical solvent

is much higher than the absorption capacity c_{ph} of the physical solvent, whereas the physical solvent gives better results at high partial pressure, (e.g., p_1).

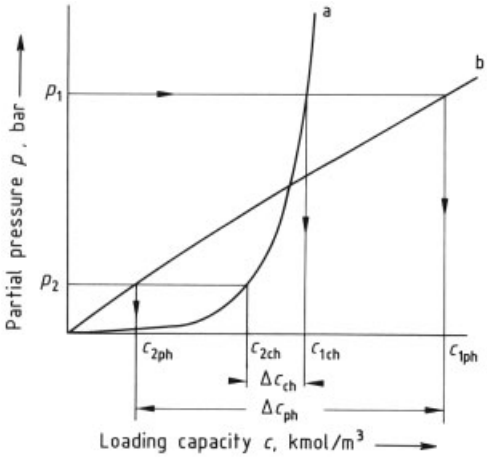

Figure 40. Gas – liquid equilibrium lines for (a) chemical and (b) physical absorption

The strong dependence of physical solubility on partial pressure can be utilized efficiently for solvent regeneration by means of simple flashing (i.e., pressure reduction). This improves the economy of operation considerably, particularly in case of bulk removal of impurities at high operating pressure.

If the dissolved components are chemically bound, less gas is released ($\Delta c_{\mathrm{ch}} < \Delta c_{\mathrm{ph}}$) when the solvent is depressurized. Thus, regeneration in chemical absorption processes is almost always accomplished by reboiling.

Fewer trays are generally required for chemical absorption than for physical absorption due to the acceleration of mass transfer by chemical reaction in the liquid phase and the low acid gas equilibrium pressure over the solvent at low loading.

Selection of Solvents. Selection of the appropriate solvent in a gas treating plant is most important because the solvent circulation rate determines equipment size and thus capital and operating costs. The following simplified rules are applicable, if linear dependence of the loading capacity on partial pressure across the entire concentration range in physical absorption and saturation of a chemical solvent above a certain partial pressure are assumed.

The solvent circulation rate required for *physical absorption* is largely independent of the concentration of the component to be removed but proportional to the total quantity G of crude gas and inversely proportional to the total pressure p and the absorption factor α_i. The minimum solvent circulation rate $L_{\min}$ required to remove component i completely can be estimated from the component balance around the absorber.

$$L_{\min} = \frac{G}{\alpha_i \cdot p}$$

The solvent circulation rate for *chemical absorption* is proportional to the quantity of the component to be removed and thus highly dependent on the concentration of this component in the gas, but practically independent of pressure. The maximum allowable concentration of the absorbed component in chemical solvents is often set by practical limits like increasing corrosion rate, increasing viscosity or tendency of foaming at higher loading.

In addition to these basic rules, the following points should be considered in selecting an absorption process.

Achievable Purity. High product gas purity can be obtained with many absorption processes, but only a few are suitable from an economic viewpoint. The minimum concentration of the component to be removed in the clean gas is mainly determined by the residual loading of that component in the lean solvent fed to the absorber top. Physical solvents can easily be regenerated to low residual loadings by pressure reduction and mild reboiling, whereas chemical solvents require high energy input for reboiling to achieve low loading.

Selectivity. The selectivity of a solvent is its tendency to absorb primarily one or a group of components from a gas mixture. Selectivity is often defined as the concentration ratio of component i to component j in the regenerator off-gas to the concentration ratio of component i to component j in the raw gas.

$$S = \frac{y_{\mathrm{i,off-gas}}}{y_{\mathrm{j,off-gas}}} \Big/ \frac{y_{\mathrm{i,raw-gas}}}{y_{\mathrm{j,raw-gas}}}$$

A high selectivity between the components to be removed and the product gas components is advantageous. High absorption rates of the

valuable product gas components reduce process economy and may affect treatment of the off-gas. If two or more components are to be removed from the raw gas and these components must be obtained separately, the solubility of the various components in the solvent should be different so that the selectivity is high.

Utilities. For chemical absorption, cheap low-pressure steam or waste heat with a sufficiently high temperature is generally required, whereas cheap electric power and cooling water at low temperature are an advantage for physical absorption.

Crude Gas Impurities. Impurities such as carbonyl sulfide, carbon disulfide, hydrogen cyanide, ammonia, mercaptans, thiophenes, phenols, heavy hydrocarbons, metal carbonyls, chlorides, and inorganic components such as mercury directly influence selection of both solvent and process [166]. Some trace compounds of the raw gas can react with the solvent or other absorbed compounds to form nonvolatile reaction products, which cannot be removed during regeneration. They accumulate in the solvent circuit and can lead to operational problems like plugging, foaming, corrosion and to solvent losses by degradation. Many commercial solvents are not capable of removing organic sulfur components and cyanide to the required low levels to meet emission regulations or downstream process requirements. Raw gases from coal gasification contain mercury which must be trapped. The alternative to the selection of a solvent which allows the removal of all these trace components in the absorption process step is the installation of a process chain consisting of COS hydrolysis and cyanide conversion reactors and adsorption guard beds.

Therefore trace components often determine the suitability of a solvent.

Other Solvent Properties. If the raw gas has to be dried or a low water dew point of the product gas has to be achieved, absorption processes using non-aqueous organic solvents are favorable. Aqueous chemical washes are preferable when upstream process stages produce water as a component of the treated gas mixture and this water is required in the product gas downstream. The laden and the lean solvent should be noncorrosive to avoid expensive materials of construction. High chemical and thermal stability reduce solvent losses due to degradation and avoid expensive solvent reclaiming. Toxicity affects solvent handling, storage, and safety equipment. The solvent should be easily available on the market.

The selection of an appropriate solvent can be supported by computer tools such as ABSORPERT (Ruhr-University Bochum [167]). This program contains heuristic rules, thermodynamic properties of a large number of solvents and a databank of industrial absorption processes. All the information is linked to provide an expert system for a systematic solvent selection and a basic process design based on the information of the feed stream and the desired separation.

Regeneration Methods. Absorption processes with solvent recovery consist, in principle, of absorption and regeneration (desorption) process steps. Several possible solvent regeneration methods are available; these are shown in Figure 41.

The simplest and cheapest method is flash regeneration (Fig. 41 A), in which the pressurized laden solvent, is depressurized in one or more stages. The residual content of dissolved components depends on the pressure in the last stage; therefore, this is often reduced to vacuum. Regeneration efficiency can be increased considerably by inert gas stripping (see Fig. 41 B). If the inert gas is completely free of the gas to be removed, the residual load of the solvent can be reduced as desired, provided the volume of stripping gas is large and the stripping rate is high enough. It is, however, unavoidable that the removed gas component is diluted by the inert gas. This may be desirable in some cases, whereas in others, further processing becomes economically prohibitive. Extremely high purity can be obtained without using inert gas if the impurities are removed in a hot regeneration stage (Fig. 41 C). This method makes use of the fact that, in general, the solubility of all gases decreases sharply with increasing temperature. The solvent is stripped by its own vapor, which is then condensed from the overhead gas and recycled as reflux. However, this method is rather costly in terms of capital investment (1) because it requires heat exchangers for the lean and laden solvent and a regeneration column with reboiler and condenser, and (2) because the utility demand for heating and cooling is rather high.

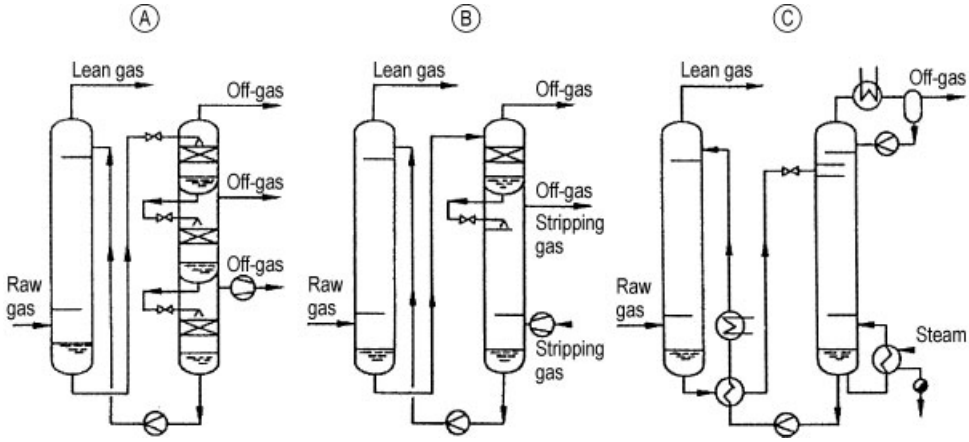

Figure 41. Absorption with different regeneration methods
A) Flashing; B) Stripping; C) Reboiling

5.4.2. Processes for Carbon Dioxide and Sulfur Compound Absorption [168 – 171]

Numerous chemical and physical absorption processes are available for removing carbon dioxide and various sulfur compounds from raw gases. Hydrogen sulfide is the most widespread sulfur component and, like carbon dioxide, is an acidic compound. Other sulfur impurities that may be present in a raw gas are carbonyl sulfide, carbon disulfide, mercaptans, and thiophenes. A summary of chemical and physical absorption processes is given in Table 28.

Physical absorption processes are used predominantly to remove carbon dioxide and hydrogen sulfide from gases with high carbon dioxide partial pressure, especially converted gases produced in partial oxidation plants based on coal or residual oil. Here the partial pressure of carbon dioxide is above the level appropriate for economic application of chemical absorption systems.

Chemical absorption process (normal amine based) are used widely, especially for crude gas containing carbon dioxide and hydrogen sulfide at low partial pressure. As shown above, various amines are employed. The selection of amine is influenced by specific process conditions such as the presence of further impurities in the raw gas, which may cause solvent degradation, corrosion and foaming. Low coabsorption of valuable gas components in the aqueous amine solution is an advantage of all amine processes.

Table 28. Absorption processes for acid gas removal

Trade name	Solvent	Licensor
Chemical absorption with aqueous amine solutions		
MEA	1 – 3 N monoethanolamine	free process
DEA	2 – 4 N diethanolamine	Elf Aquitaine and others
DGA/ Econamine	4 – 6 N diglycolamine (2-(2-aminoethoxy)ethanol)	Fluor Daniel, Huntsman
DIPA	2 N diisopropanolamine	Shell and others
MDEA	2 – 5 N methyldiethanolamine	free process
ADIP	DIPA or MDEA	Shell
Amine Guard/ Ucarsol	formulated MDEA	UOP/Union Carbide
aMDEA	MDEA + activator for enhanced CO_2 absorption	BASF
Gas Spec	formulated MDEA	DOW Chemical
Flexsorb	hindered amine	Exxon
Chemical absorption with alkaline solutions (hot potassium carbonate)		
Benfield	K_2CO_3 + activator (DEA)	UOP
Catacarb	K_2CO_3 + catalyst	Eickmeyer
Giammarco-Vetrocoke	K_2CO_3 + activator (arsenic trioxide)	Giammarco
Vacasulf	K_2CO_3 + NaOH	Krupp Uhde
Physical absorption		
Rectisol	methanol	Linde/Lurgi
Selexol	polyethylene glycol dimethyl ether (DMPEG)	UOP
Purisol	*N*-methyl-2-pyrrolidone (NMP)	Lurgi
Morphysorb	*N*-formylmorpholine (NFM)	Krupp Uhde
Physical – chemical absorption		
Sulfinol D/M	sulfolane (tetrahydrothiophene oxide) + DIPA or MDEA	Shell
Amisol	methanol + MEA, DEA or DETA	Lurgi

Sulfur components like hydrogen sulfide, carbonyl sulfide, and mercaptans have a higher solubility in physical solvents than carbon diox-

ide. This facilitates production of a sufficiently hydrogen sulfide rich gas for the sulfur recovery process. Compared to aqueous amine solutions, selectivity with regard to valuable components is lower in organically based physical solvents, which leads to higher losses of these components.

Amine solutions do not show a selectivity of hydrogen sulfide over carbon dioxide at equilibrium. Therefore, very high sulfur purity can be obtained only if carbon dioxide is removed simultaneously to a very low level. However, the reaction rate of certain amines with hydrogen sulfide is much higher than with carbon dioxide. Hydrogen sulfide can therefore be selectively removed by utilizing this difference in reaction kinetics. In the absorber a large gas liquid interfacial area has to be provided for a fast hydrogen sulfide mass transfer to achieve the required low sulfide concentration in the purified gas. Simultaneously a short contact time between gas and liquid is required so that only a small portion of the absorbed carbon dioxide can react with the amine [172].

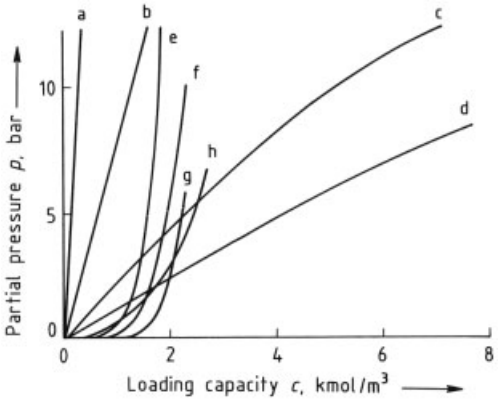

Figure 42. Equilibrium curves of CO_2 in various solvents a) H_2O (30 °C); b) *N*-methyl-2-pyrrollidone (40 °C); c) Methanol (− 15 °C); d) Methanol (− 30 °C); e) Hot potassium carbonate solution (110 °C); f) Sulfinol solution (50 °C); g) 2.5 M Diethanolamine solution (50 °C); h) 3 M Amisol DETA solution

In an attempt to combine the advantages of chemical and physical absorption, processes using chemical absorbents mixed with organic solvents have been developed. These processes, termed *physical – chemical absorption processes* or *mixed solvent processes*, are suitable both for absorbing carbon dioxide and hydrogen sulfide and for removing organic sulfur compounds, giving high end purities. A survey of the equilibrium behavior of carbon dioxide in common solvents is shown in Figure 42.

Initial selection of a gas purification process based on the partial pressure of the acid gas in the raw gas versus the product gas specification can be done with the aid of Figure 43.

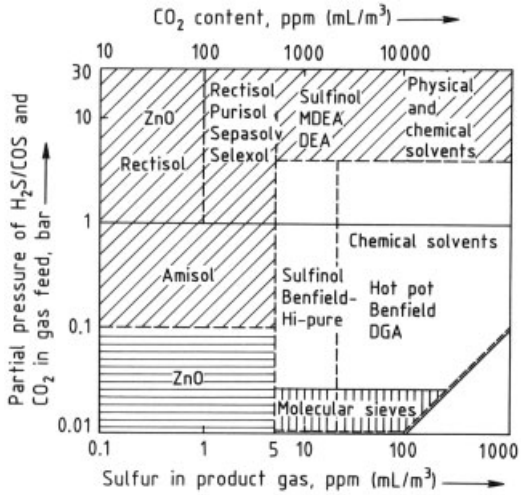

Figure 43. Selection of appropriate gas purification process for simultaneous H_2S/COS and CO_2 removal

If the acid gas partial pressure in the raw gas is very low, *dry-bed adsorption processes* (e.g., solid beds of zinc oxide, iron oxide, activated carbon, or molecular sieves) or nonregenerable scavenger solutions are used for gas cleaning.

5.4.2.1. Physical Absorption Processes

Water was an industrially important solvent for carbon dioxide absorption, particularly in ammonia synthesis, until organic solvents were introduced. Organic solvents have a much higher solvent capacity than water, which improves process economy.

Organic solvents are especially suitable for treating converted raw gases obtained from partial oxidation of oil or coal. The carbon dioxide partial pressure in these gases can be as high as 2.5 MPa. At this high acid gas partial pressure the solvent capacity, i.e. solvent loading,

of organic solvents exceeds that of chemical absorbents considerably.

Rectisol Process. The Rectisol process, invented by Lurgi around 1950 and developed further together with Linde, is particularly suitable for the specific requirements of high-pressure synthesis gas production [173].This process achieves syngas quality with very low total sulfur content in the ppb range and removal of all impurities and bulk CO2 in one process step. More than 60 Rectisol units were reported to be operating worldwide in 2005 purifying more than 210×10^6 m^3/d (STP) of synthesis gas, which is equivalent to more than 75 % of the gas produced from coal, oil, and waste gasification.

Originally designed and applied in conjunction with Lurgi pressure gasification of coal, the Rectisol process is now also used successfully in oil-based synthesis gas plants and other coal gasification processes, as well as in special cases of natural gas treatment.

The Rectisol process uses cold methanol as solvent. The solubility of acid gas in methanol increases strongly with decreasing temperature whereas the solubility of the valuable light components is almost temperature independent. Temperatures as low as − 75 °C are therefore applied in order to minimize the required solvent circulation rate and the loss of valuable gas components. The low operation temperature also reduces solvent losses into the product streams by lowering the vapor pressure of methanol. Good mass and heat transfer are provided by methanol due to its low viscosity. Operating the process at low temperature does not necessarily require large refrigeration units. The major part of the cooling is done by internal heat exchange and by utilizing the temperature decrease by gas desorption and expansion. In Figure 44, the temperature dependence of the absorption coefficient λ, which is defined by $c = \lambda p$, is shown for various gases at 1 bar. The absorption coefficient is not entirely pressure-independent but rises above the value measured at 1 bar as the partial pressure of the gas approaches its saturation vapor pressure. The nonideal behavior of the gas phase at higher total pressure also must be considered to model the gas liquid equilibria correctly (→ Absorption).

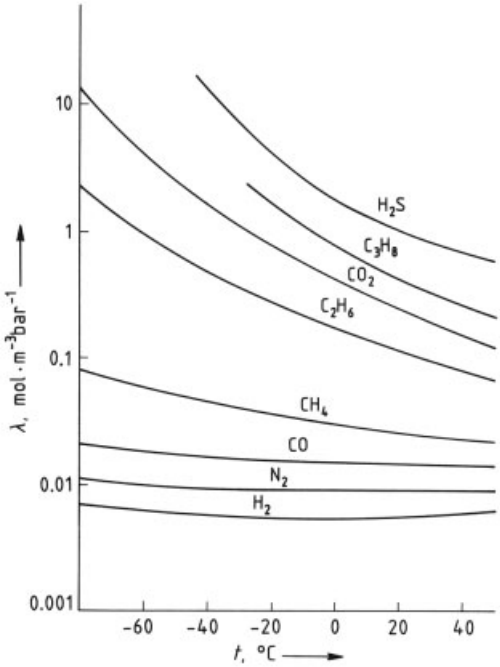

Figure 44. Absorption coefficient λ of various gases in methanol at low pressure (1 bar)

Standard Rectisol Process. In gas production, the Rectisol process is used in two basic configurations: the nonselective standard process and the selective process configuration. All acid gas constituents are absorbed simultaneously in the standard Rectisol process. The basic flow diagram of the standard process with two absorption sections is shown in Figure 45. Flash-regenerated methanol, fed to the lower part of the absorber, largely removes hydrogen sulfide and carbon dioxide at bottom temperatures between − 10 to − 30 °C. The remaining hydrogen sulfide and carbon dioxide are removed from the gas to the required purity by heat-regenerated cold methanol in the absorber top section.

Under these process conditions, solvent loading is very high at the *bottom of the absorber*, which results in a low solvent rate. The heat of absorption raises the temperature of the solvent; the absorber bottom temperature may be as much as 50 °C higher than the solvent inlet temperature. During flash regeneration of the solvent, this heat of absorption is given off. and the solvent temperature drops accordingly. The applied low absorption temperatures are generated

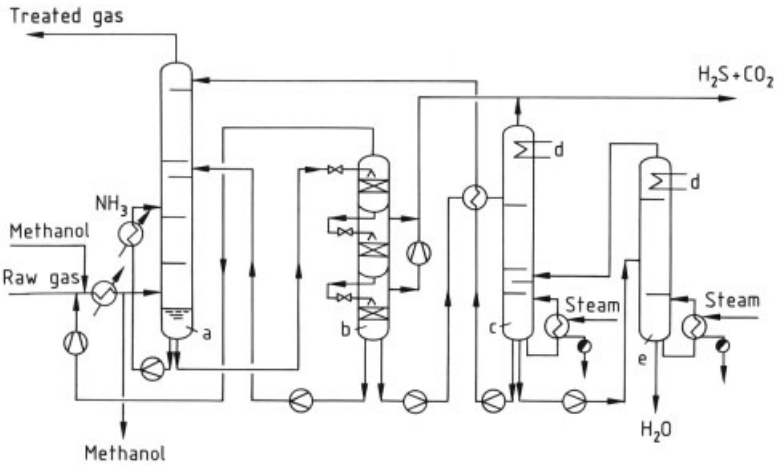

Figure 45. Standard Rectisol process for simultaneous removal of H_2S and CO_2
a) Absorber; b) Flash tower; c) Hot regeneration; d) Condenser; e) Methanol – water column

in this desorption step and not by the refrigeration unit. This phenomenon is often referred to as the "autorefrigeration effect". In the *fine wash section* of the absorber, only the residual acid gases are removed. Therefore, little heat of absorption is released in the fine wash section and the temperatures remain low. Moreover, there is a smaller quantity of gas to be treated in the fine wash section. Therefore, solvent demand in the fine wash section is considerably lower than that in the main section of the absorber. This is desirable in view of the high investment and utility costs for hot regeneration.

In Fischer – Tropsch synthesis, SNG, and town gas plants, carbon dioxide must only be removed to a final content of 1 – 2 %. This value can even be reached in the main wash part with flash-regenerated methanol. The fine wash part then has to be designed only for removal of hydrogen sulfide and carbonyl sulfide, which results in a lower solvent rate, normally one-tenth of the total amount of solvent withdrawn from the bottom, due to the higher solubility of these components.

Downstream of hot regeneration, a methanol – water distillation column is used to separate water entering with the raw gas. To prevent icing in the cooling section of the raw gas, a small amount of methanol is injected upstream of the heat exchanger. The mixture of methanol and water is withdrawn and fed directly to the distillation column. A refrigeration unit is used to cover cold losses in the plant due to surge energy, incomplete heat exchange, unrecovered heat of absorption, and insulation losses. An ammonia absorption refrigeration unit can be advantageously applied if low temperature (130 – 200 °C) waste heat is available.

The standard Rectisol process combines all absorbed gases in a single acid gas stream. The sulfur compounds are thus diluted with the entire absorbed volume of carbon dioxide. (Referred to the raw gas specification in Table 24, only 0.9 vol % of hydrogen sulfide [example 2] and only 1.8 vol %, of hydrogen sulfide [example 4] in the off-gas are typical of the standard process with complete removal of carbon dioxide.) The off-gas with this low sulfur content is not suitable for sulfur recovery in a Claus process. Other sulfur recovery processes must be applied such as liquid-phase oxidation (see 5.4.3) or adsorption. The standard Rectisol process is therefore advantageously applied for gases from low-sulfur feedstock.

Selective Rectisol Process. Initially the selective Rectisol process was introduced to produce an off-gas rich in hydrogen sulfide which could be processed in a Claus plant for sulfur recovery. Pure carbon dioxide is produced as a second product with a total sulfur content in the low ppm range. This stream can be vented to atmosphere or utilized, for example in urea plants, which are often built in conjunction with ammonia plants, or recovered in food grade (see Section 6.3),or can be used for enhanced oil recov-

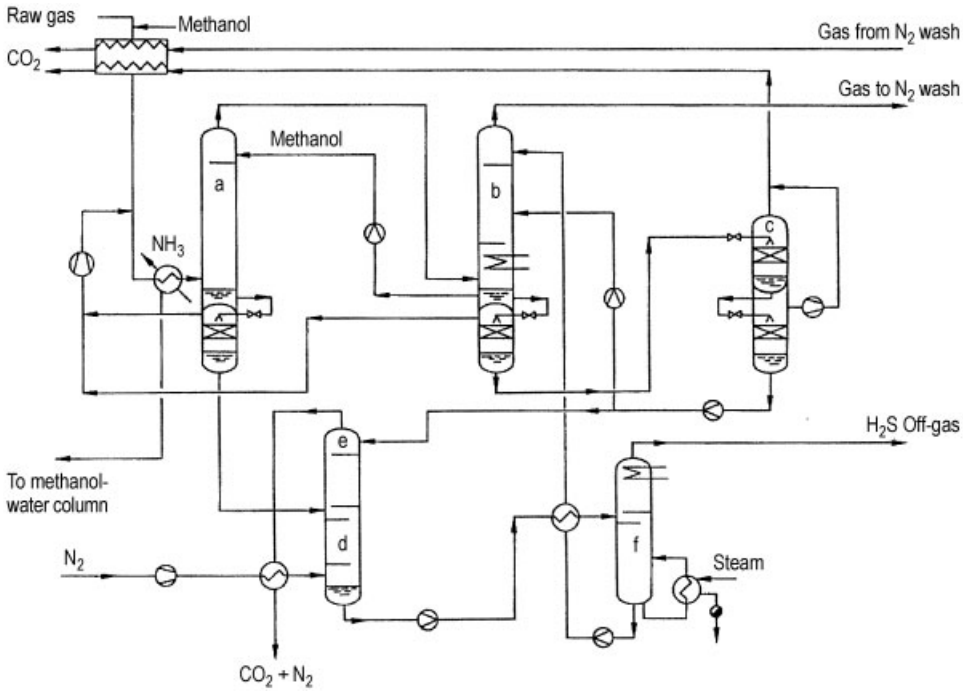

Figure 46. Selective Rectisol process for separate removal and recovery of H_2S and CO_2
a) H_2S absorption; b) CO_2 absorption; c) Flash regeneration; d) Stripper; e) Reabsorber; f) Hot regeneration

ery and sequestration. The solvent loops shown in Figure 45 are interchanged for that purpose (see Fig. 46). The raw gas is first desulfurized with a small amount of heat-regenerated methanol. The carbon dioxide coabsorption is low because of its low solubility and the presaturation of methanol in the carbon dioxide absorption column.

Most of the carbon dioxide and the remaining traces of sulfur compounds are removed in the downstream carbon dioxide scrubber in two stages, by using flash-regenerated methanol in the middle section and heat-regenerated methanol in the top section.

Depending on the requirements, carbon dioxide may be removed down to a few percent or even a few parts per million. Removal of both carbonyl and hydrogen sulfides as well as mercaptans to 0.1 ppm is ensured in all cases.

According to Figure 44, the selectivity between carbon dioxide and hydrogen sulfide for the raw gas indicated in column 4 of Table 24 is not sufficiently high to give an off-gas rich enough in hydrogen sulfide for the Claus process. Therefore, an additional concentration stage is required before the solvent passes to hot regeneration. In the bottom section of the reabsorber column (d in Fig. 46) carbon dioxide is stripped out of the laden solvent by simple pressure reduction or stripping with inert gas. In the top section (e) the small amount of hydrogen sulfide which is stripped together with the carbon dioxide in the bottom section is reabsorbed with methanol to yield a sulfur-free off-gas.

In the selective Rectisol process, much more solvent must be treated by hot regeneration than in the standard Rectisol process. Additional methanol is required in the reabsorber, and more methanol is necessary in the hydrogen sulfide absorber because the entire gas rate must be treated at higher temperature.

To avoid the extra load on the regeneration unit, the two Rectisol wash steps can be separated by removing the sulfur compounds upstream of carbon monoxide conversion, with carbon dioxide removal located downstream of the conversion unit. The solvent loops of both absorption steps are still combined. A classic application of this case is synthesis gas production by residual oil gasification.

As shown in Table 24, example 3, the ratio of hydrogen sulfide to carbon dioxide upstream of shift conversion is approximately ten times higher than downstream (example 4). Therefore, the natural selectivity of methanol is sufficient to produce a sulfur-rich off-gas suitable for sulfur recovery in a Claus process. In addition, the gas rate upstream of carbon monoxide conversion is only 70 % of the rate downstream; therefore, less solvent is required for desulfurization. This simplification of the gas treating unit, combined with the advantage of passing a sulfur-free gas to the carbon monoxide conversion unit and the possibility to apply low-temperature shift conversion for minimum residual carbon monoxide in the raw hydrogen, must be paid for by cooling the gas twice.

Because the selective Rectisol process is much more complex than the standard process, a decision must be made in each individual case as to which route gives the most economical solution in conjunction with the downstream plant and the sulfur recovery unit.

For Fischer – Tropsch synthesis and SNG plants based on coal with a low sulfur content (example 2 in Table 24; pure carbon dioxide is not required in either case), the standard Rectisol process has mostly been used, with the off-gas being treated in a special chemical system.

A two train Rectisol plant is shown in Figure 47. Part of the raw gas from a residual oil gasification is shifted and purified in the first train to produce pure hydrogen. The remaining raw gas is purified in the second train giving gas for oxo synthesis. The solvent of both trains is regenerated in a common hot regenerator column.

Applications [174, 175]. In view of (1) the high flexibility with regard to carbon dioxide, hydrogen sulfide, and carbonyl sulfide removal, as well as the removal of various trace compounds, (2) the high purity of the treated gas, and (3) the absence of corrosion at low temperature, the Rectisol process is used in the following cases:

Standard processes are used for raw and converted gas from coal gasification (in this case, combination with prewash systems for other impurities is also possible).

Selective processes with two *absorption steps immediately following* are used for raw and converted gas from coal or oil gasification plants.

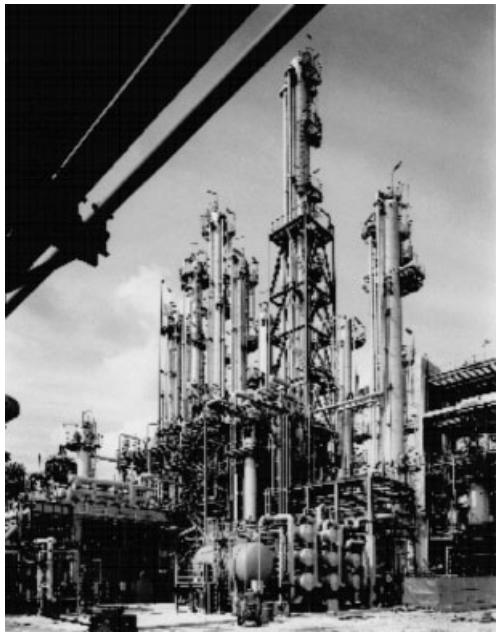

Figure 47. Two train selective Rectisol plant downstream of residual oil gasification producing 40 000 m^3/h (STP) gas for oxo synthesis from unshifted gas and 30 000 m^3/h (STP) hydrogen from shifted raw gas (Courtesy of Celanese and Lurgi AG)

The selective process is used *with two separate absorption steps* for raw and converted gas from oil and coal gasification by partial oxidation processes. Absorber – stripper systems (Fig. 41) are used to remove small quantities of carbon dioxide and hydrogen sulfide as integral parts of low-temperature units.

Plants with a one train throughput of more than 15.6×10^6 m^3 (STP)/d are already operating. The utility requirements for a Rectisol unit vary widely depending on process conditions. Typical consumption figures for selective removal of hydrogen sulfide and carbon dioxide from raw gas, yielded by partial oxidation relative to 100 000 m^3 (STP)/h (H_2 and CO) are

Power for drive shaft (without power recovery)	2000 – 2500 kW	
Steam (0.5 MPa)	5 – 7 t/h	
Waste heat above 110 °C	$33 - 42 \times 10^6$ kJ/h	using absorption
Cooling water (Δt = 10 °C)	1500 – 2000 m^3/h	refrigeration
Methanol	60 – 70 kg/h	

The application of the Rectisol Process for hydrogen production and for Integrated Gasification Combined Power Cycles (IGCC) [176]is described in 5.4.2.4(see Sections 7.2 and 7.3).

Processes with Other Physical Solvents. The Rectisol process needs a refrigeration unit to maintain the low absorption temperatures. Other processes employ organic solvents with high boiling points and low vapor pressures. These solvents provide a high absorption capacity for acid gases at ambient or slightly subambient temperature without appreciable vapor losses.

Although higher absorption temperature lowers carbon dioxide solubility, at a high partial pressure of carbon dioxide the loading capacity of organic solvents is still attractive and similar to the capacity of chemical solvents under comparable conditions.

Processes using high-boiling organic solvents are

1) Selexol process developed by Allied Chemical Corporation (now licensed by Norton and UOP) and the Genosorb process of Clariant and Krupp Uhde, all using a mixture of dimethyl ethers of polyethylene glycol (DMPEG);
2) Purisol process, developed by Lurgi, using *N*-methyl-2-pyrrolidone (NMP);
3) Fluor solvent process using propylene carbonate (PC);
4) Morphysorb process developed by Krupp Uhde using *N*-formylmorpholine (NFM);
5) Sepasolv process by BASF applying a mixture of oligoethylene glycol methyl isopropyl ethers (MPE); and
6) the Gaselane process using *N*-methylcaprolactam (NMC) developed by VEB Leuna [178]

The normal boiling point of these solvents is between 200 and 350 °C; their relative molecular mass, between 100 and 320. The solubility of carbon dioxide is similar for all solvents.

If selective removal of hydrogen sulfide from gas containing carbon dioxide is required, NMP is superior to Selexol, Sepasolv, and propylene carbonate. The solubility of carbonyl sulfide in all solvents is remarkably lower than the solubility of hydrogen sulfide.

This means that if these processes are used to purify gases containing both carbonyl and hydrogen sulfides to synthesis gas quality, a carbonyl sulfide hydrolysis step is generally required. With regard to removal of carbon dioxide from sulfur-free synthesis gas, the process flow sheet is basically the same for all processes, consisting of a simple absorber – stripper system. In addition to nitrogen or other inert gases, air is a suitable stripping medium.

Selexol Process [177]. This process was developed by Allied Chemical Corporation and is now licensed by Norton and UOP. The Selexol solvent is a trademark of the DOW Chemical Company and consists of a mixture of dimethyl ethers of polyethylenglycol (DMPEG). A similar mixture of DMPEG is used as solvent in the Genosorb process of Krupp Uhde and Clariant (solvent supplier). More than 50 Selexol plants had been constructed by 2005 purifying more than 110×10^6 m^3/d (STP) of natural gas and synthesis gas. Most commercial applications of the Selexol process are for treating natural and synthesis gas with high contents of hydrogen sulfide and carbon dioxide or high concentrations of carbon dioxide alone [175]. To purify gases containing carbon dioxide and hydrogen sulfide, the process may be used for either bulk removal of both components or selective removal of hydrogen sulfide, with the desired quantity of carbon dioxide being retained in the lean gas. Whereas the solubility of hydrogen sulfide is approximately nine times that of carbon dioxide, the solubility of carbonyl sulfide is only twice that of carbon dioxide.

This means that the difference in solubility between hydrogen sulfide and carbon dioxide is large enough to provide a good basis for selective absorption of the former. The solubility difference between carbonyl sulfide and carbon dioxide, on the other hand, is less favorable.

If carbonyl sulfide must be removed together with hydrogen sulfide, the solvent picks up more carbon dioxide and the feed to the Claus unit will be diluted with more carbon dioxide than for removal of hydrogen sulfide alone. Hydrolyzing carbonyl sulfide upstream of the absorption process can circumvent this problem. Nearly total conversion of carbonyl to hydrogen sulfide can be attained by vapor-phase catalytic hydrolysis, leaving only hydrogen sulfide and carbon dioxide as key components.

A certain amount of water in the Selexol solvent does not greatly decrease the solubility of

acid gases. Thus, in the case of hot regeneration by reboiling under normal pressure, the boiling point of the solvent can be decreased by addition of a suitable quantity of water.

A disadvantage of the Selexol solvent is that its lower boiling components are distilled off. Care must be taken that solvent properties do not change in the course of time.

The Selexol process is applied for following purposes:

1) Selective removal of hydrogen sulfide from natural gas and synthesis gas
2) Removal of high carbon dioxide content to a low residual level from natural gas and synthesis gas
3) Purification of landfill gas

Purisol Process. The Purisol process applies a mixture of NMP and water as solvent. It can be operated either at ambient temperature or — in conjunction with a refrigeration unit — at temperatures down to ca. − 15 °C. The Purisol process has been operated successfully for several decades in selective natural gas desulfurization and bulk CO_2 removal from converted synthesis gas. The process has the advantage of a higher hydrogen sulfide – carbon dioxide selectivity than Selexol, permitting the production of sulfur-rich Claus gas with about 50 % hydrogen sulfide, even if the ratio of carbon dioxide to hydrogen sulfide in the raw gas reaches extremely high values.

The solubility of carbonyl sulfide is only 20 % that of hydrogen sulfide, but the former is partly hydrolyzed to the latter by the NMP solvent resulting in a 60 – 80 % carbonyl sulfide removal. If the absorber is designed with carbonyl sulfide reaction zones, COS can be removed quantitatively. The absorption of valuable gas components is approximately 10 % lower compared to Selexol. Therefore a smaller amount of flash gases has to be recompressed.

For carbon dioxide removal, the NMP solvent can be regenerated with air, if nitrogen is not available. The carbon dioxide concentration in the purified gas can be reduced to 1000 ppm or less, with almost no steam demand for regeneration.

A typical example of a Purisol unit for removing carbon dioxide from converted gas is shown in Figure 48. Two identical units of this type operate in the same plant, treating synthesis gas for hydrogen production.

Pre- and afterwash steps are shown for both the absorber and the stripper in the flow sheet of Figure 48. Initially, raw and stripping gases are dried with a small amount of NMP to keep the water content of the solvent in the carbon dioxide absorption section low. A water wash on top of the columns reduces NMP losses. In general, NMP recovery with water is not necessary if the Purisol process is operated at subambient temperature. If a low water dew point of the product gas is required, NMP can be recovered by means of a glycol afterwash step.

The water – NMP mixture is separated in a third column, which simultaneously recovers NMP from the carbon dioxide off-gas at the flash stage. Carbon dioxide itself reduces the boiling point of the NMP – water mixture, which is a desirable effect.

Applications. The Purisol process is excellently suited to fours typical applications for high-pressure gases:

1) Selective removal of hydrogen sulfide
2) Removal of high carbon dioxide content to a low residual level
3) bulk removal of acidic compounds down to moderate product purity by using a simplified flash regeneration system, and
4) Removal of organic sulfur components (mercaptans)

Fluor Solvent Process. In the late 1950s, the Fluor solvent process using propylene carbonate was commercialized by Fluor and El Paso [180]. Removal of carbon dioxide from natural gas at mildly subambient temperature was the first application. The main criterion for selecting PC in the Fluor solvent process was its high carbon dioxide solubility, together with a relatively low solubility of methane. Propylene carbonate has limited water solubility, and slowly reacts irreversibly with water and carbon dioxide around 90 °C. Thus, it is unsuitable for controlling the water content by distillation under atmospheric pressure. Pre- or afterwash steps are, therefore, not applicable.

Morphysorb [181, 182]. Krupp Uhde applied their experience with *N*-formyl morpholine (NFM) as solvent to develop the Morphysorb process for selective sulfur compounds removal and bulk CO_2 absorption. Lower meth-

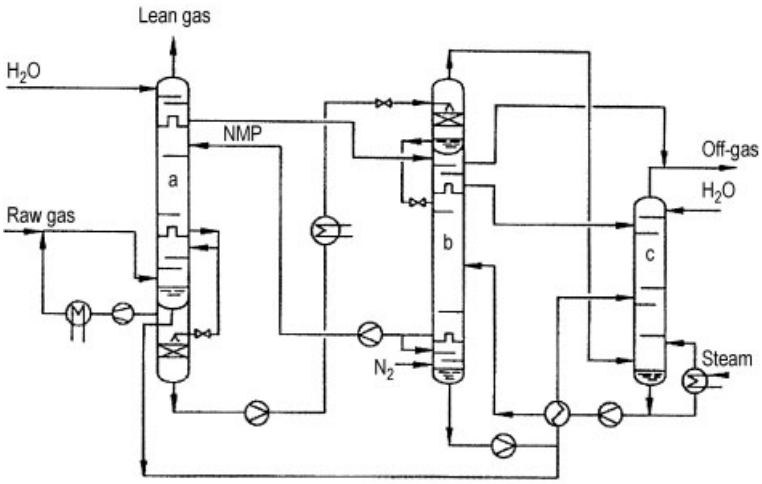

Figure 48. Purisol process for CO_2 removal from high pressure converted gas
a) Absorber; b) Stripper; c) NMP-water separator

ane solubility compared to other commercial solvents is claimed. NFM has a freezing point of 23 °C. To take advantage of the higher acid gas solubility at lower temperatures a mixture of NFM and *N*-acetyl-morpholine (NAM) is proposed.

5.4.2.2. Chemical Absorption Processes

Amine Processes. Aqueous amine solutions have been used on an industrial scale for the absorption of carbon dioxide and hydrogen sulfide for a long time. They are especially suited for gases with low organic sulfur content and less stringent purity requirements regarding hydrogen sulfide, as for example natural gas of pipeline quality or fuel gases. Synthesis gas qualities with total sulfur contents below 1 mL/m^3 require an additional sulfur adsorption guard bed downstream of the amine absorber.

Mono- and diethanolamine (MEA, DEA), as well as diglycolamine (DGA, 2-(2-aminoethoxy)ethanol), diisopropanolamine (DIPA), and methyldiethanolamine (MDEA), are well known in this field. Several so-called formulated amines are available on the market such as Ucarsol, Gas-Spec, and aMDEA (see Table 28). These solvents contain specific additives in order to allow higher amine concentrations, to enhance carbon dioxide absorption rate and to reduce corrosion potential and foaming tendency. The selectivity of the solvent towards hydrogen sulfide can be improved by sterically hindered amines such as Flexsorb SE.

Alkanolamines can be regarded as derivatives of ammonia in which one, two, or three hydrogen atoms have been replaced by a C_nH_{2n}-OH group. According to the number of organic groups attached to the nitrogen atom they are classified as primary, secondary, and tertiary alkanolamines. The primary alkanolamines applied for gas treating contain one ethanol group (MEA) or one glycolic group (DGA) The secondary alkanolamines employed contain two ethanol (DEA) or isopropyl alcohol groups (DIPA). MDEA as an *N*-alkylated alkanolamine has two ethanol groups and one methyl group attached to the nitrogen. The amino group provides the necessary alkalinity in the water solution to absorb the acid gases. The hydroxyl group in the alcohol moiety increases the water solubility and reduces the amine vapor pressure.

In the amine processes the acid gas components react reversibly with the alkaline solvent. The reactions of hydrogen sulfide and carbon dioxide with amines can be represented as follows:

Reaction of amine and water (protonation):

$$R_1R_2R_3N + H_2O \rightleftharpoons R_1R_2R_3NH^+ + OH^-$$

Sulfide formation:

$$H_2S + R_1R_2R_3N \rightleftharpoons R_1R_2R_3NH^+ + HS^-$$

Carbamate formation:

$$CO_2 + 2\,R_1R_2R_3N \rightleftharpoons R_1R_2NCOO^- + R_1R_2R_3NH^+$$

Bicarbonate formation:

$$CO_2 + OH^- \rightleftharpoons HCO_3^-$$

Hydrogen sulfide reacts instantaneously with all amines to form sulfide. Carbon dioxide can only react with primary and secondary amines to form carbamate. The carbamate reaction rate has a moderate velocity. The bicarbonate formation is slow. The different reaction rates of hydrogen sulfide and carbon dioxide with the amines are the basis for selective hydrogen sulfide absorption. MDEA has a higher selectivity compared to primary and secondary amines since no carbamate is formed with tertiary amines and carbon dioxide can only be absorbed via the slow bicarbonate reaction. In order to utilize this selectivity based on the reaction kinetics, the absorber must provide sufficient mass transfer area for complete hydrogen sulfide removal in combination with short residence time to suppress carbon dioxide absorption [183].

When the temperature of the solvent is increased and the pressure reduced, the chemical equilibrium of the above equations is shifted to the left side and the acid gases are released.

The flow diagrams for all amine process are basically the same (Fig. 49). Feed gas is brought in contact with the solvent in an absorber (a). Laden solvent flows through a heat exchanger (d) to the regeneration system (e), in which acid gases are desorbed from the solvent at reduced pressure with vapor generated by reboiling. Acid gases leave the regenerator overhead after the stripping steam has condensed (f). Regenerated solvent is returned to the absorber by way of the solvent heat exchanger (d) and solvent cooler (b). Amine losses with the lean gas have to be reduced by a water wash step. If the absorption operates at elevated pressure it is often economical to recover the absorbed valuable gases in an additional flash stage downstream of heat exchanger (d).

The amine concentration determines the capacity of the solvent and thus the solvent circulation rate. The maximum amine concentration of the solution is limited with regard to corrosivity of the rich solvent, viscosity, and other properties. Alkanolamines are susceptible to degradation by several trace components present in the raw gas and by carbon dioxide. Carbon dioxide catalyzes the degradation of MEA and DEA to form heat stable salts and inactive polymer reaction products, which increase solvent viscosity and foaming tendency. DIPA reacts with carbon dioxide to form oxazolidone. The capacity of the primary and secondary amines has to be maintained by continuous solvent reclaiming. MDEA as tertiary amine and hindered amines, which do not form carbamates, are not affected by carbon dioxide.

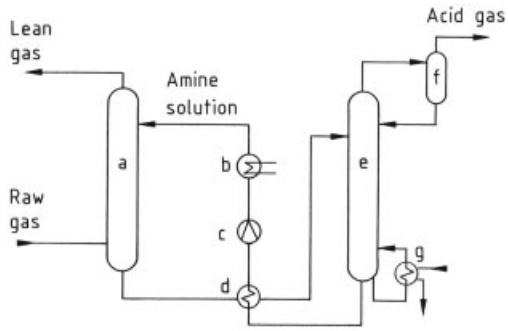

Figure 49. Basic flow diagram for amine processes
a) Absorber; b) Solvent cooler; c) Solvent pump; d) Heat exchanger; e) Regenerator; f) Condenser; g) Reboiler

Carboxylic acids (e.g., formic acid, acetic acid, oxalic acid) are frequently found in amine solvents. They are formed directly from cyanide and by degradation of the amine with oxygen. At high carbon monoxide partial pressure formic acid is directly produced in alkaline solutions. Carboxylic acids tie up amines as heat stable salts (HSS). Sulfur dioxide also reacts irreversibly with the amine to form HSS. Heat stable salts reduce solvent capacity and are a major source of corrosivity [184]. Common practice is the addition of caustic soda to liberate the amine and to restore solvent capacity. The effect of caustic addition on the corrosivity of the solvent is small because the acid anions remain in the solution [185]. The salts of the carboxylic acids have to be removed by ion exchange or distillation.

Primary Amines. Solutions of *MEA* were used almost exclusively for many years to remove hydrogen sulfide and carbon dioxide from natural gas and certain synthesis gases because of their high reactivity and the high purity obtainable.

The disadvantages of MEA are the formation of nonregenerable and corrosive degradation products with carbonyl sulfide and carbon disulfide. Because of the high vapor pressure of MEA a water wash step of the purified gas is normally required. The high heat of reaction causes high steam demand for hot regeneration. Thus MEA units are rapidly being replaced by other systems. However, the MEA solution is still preferred for gas streams containing small amounts of hydrogen sulfide and carbon dioxide with practically no carbonyl sulfide or carbon disulfide (e.g., the removal of carbon dioxide from reformed gas produced in a tubular reformer).

Newly developed corrosion inhibitors allow aqueous concentration of 30 wt % MEA, which results in a lower solvent circulation rate. An example is the Amine Guard process from Union Carbide. The Amine Guard process is suitable only for removal of carbon dioxide because sulfur components react with the inhibitors.

DGA is used in Fluor's Econamine process. The high DGA concentration (typical 40 – 60 wt %) reduces the required solvent circulation rate compared to MEA. Carbonyl sulfide is partially removed from the gas and the degradation products of DGA can be reclaimed at elevated temperature. No water wash step is required due to its lower vapor pressure [186].

Secondary amines DEA and DIPA are widely used to treat refinery gases and high pressure natural gases.

DEA is applied in the 20 – 25 wt % concentration range for refinery applications. DEA is superior to MEA for treating of refinery gases, which contain appreciable amounts of carbonyl sulfide and carbon disulfide, because DEA reacts with these components to a much lesser extent and the reaction products are less corrosive. The *SNEA-DEA process* for natural gases was developed by Societé National Elf Aquitaine (SNEA) [187]. A 25 – 30 wt % DEA solution is used resulting in higher solvent capacity. Acid gas loadings up to 0.7 – 1 mol/mol of DEA is possible with sufficiently high acid gas partial pressure. The heat of reaction of DEA and carbon dioxide is about 25 % lower than that of MEA and carbon dioxide. Therefore, DEA solutions require less steam for hot regeneration.

The *ADIP process* of Shell International Oil Products, licensed by JGC, Stork Engineers & Contractors, and Chiyoda [188], utilizes DIPA in concentrations up to 50 wt %. ADIP units are suitable for sweetening natural gas, refinery gas, and Claus tail gas. MDEA is also used in ADIP units if higher hydrogen sulfide selectivity is required. DIPA is claimed to be more stable against degrading and corrosive trace components like cyanide, oxygen, and sulfur dioxide than MDEA [212]. Therefore in cases where selectivity is not critical, DIPA is the preferred solvent for hydrogen sulfide removal. Raw gas purification and Claus tail-gas treatment can be integrated into one plant (Fig. 50). One single regeneration column (d) provides lean amine solvent for the fine wash section of the high pressure absorber (a) and the Claus tail-gas absorber (e). The tail-gas absorber operates at low pressure and hydrogen sulfide has to be removed selectively to avoid a large carbon dioxide recycle through the Claus plant. Therefore the rich solvent leaving the tail-gas absorber is not completely saturated and can absorb acid gas in the lower section of the high pressure absorber (a).

Hydrogen sulfide concentrations in the purified gas below 5 ppm (mL/m^3) are achievable with DIPA since it is easier to regenerate with less steam than MEA and DEA. DIPA has a lower tendency to form corrosive degradation products than DEA.

Tertiary Amines. MDEA is the most widely used tertiary amine. It is suitable to purify a large variety of gases. The carbon dioxide absorption rate is slower in tertiary amines because they do not form carbamates This makes MDEA particularly suited for selective hydrogen sulfide removal from gases with elevated carbon dioxide contents. Many formulated MDEA based amines are available on the market, which are claimed to have improved hydrogen sulfide selectivity (e.g., Gas Spec, Ucarsol) [189]. The low heat of reaction facilitates regeneration down to low acid gas contents in the lean amine required to achieve hydrogen sulfide concentrations below 5 mL/m^3 in the purified gas. MDEA is applied in the 30 – 50 wt % range. Acid gas loadings up to 0.8 mol/mol of MDEA is possible in carbon steel equipment. The high MDEA concentration in conjunction with the high attainable acid gas loading result in a high capacity of the solvent and a low solvent circulation rate. For this reason many amine units originally designed and operated with primary or secondary

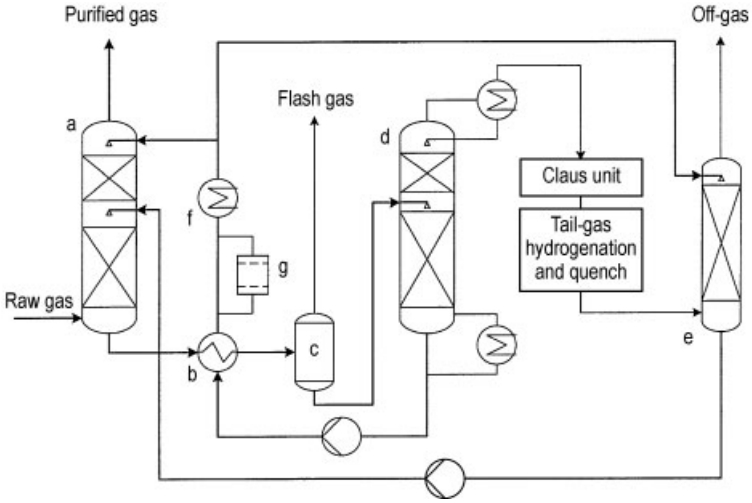

Figure 50. Integration of high pressure gas purification with Claus tail-gas treating
a) High pressure absorber; b) Solvent heat exchanger; c) Flash drum; d) Regenerator; e) Claus tail-gas absorber; f) Solvent cooler; g) Amine filter

amines were converted to use MDEA as solvent in order to increase plant capacity and reduce utility consumption [190]. As a guideline, the steam required for regeneration per volume solvent is approximately 80 kg/m^3 for bulk carbon dioxide removal and 120 kg/m^3 for selective hydrogen sulfide removal.

MDEA is widely used in refineries for hydrogen sulfide absorption from various medium- to high-pressure off-gas streams from hydrotreating units and the hydrocracker. MDEA is also suitable for selective absorption of hydrogen sulfide from low-pressure gases and is therefore applied in the Claus tail gas treating process.

The carbon dioxide absorption rate in MDEA solutions is enhanced by addition of small amounts of primary and secondary amines [191]. This effect is employed for example in the aMDEA process of BASF and the activated MDEA process of Elf Aquitaine. The carbamate formed with the activator component is hydrolyzed, giving bicarbonate and recovering the activator. The consecutive carbamate formation and hydrolysis reactions are much faster than direct bicarbonate formation with water.

Activator-carbamate reaction:

$$CO_2 + R_1R_2NH \rightleftharpoons R_1R_2NCOO^- + H^+$$

Carbamate hydrolysis:

$$R_1R_2NCOO^- + H_2O \rightleftharpoons R_1R_2NH + HCO_3^-$$

Part of the absorbed carbon dioxide can be liberated from aqueous MDEA solution by pressure reduction. Flash regeneration is sufficient if only bulk carbon dioxide removal is required. Lower carbon dioxide concentrations in the purified gas are achievable with an additional fine wash section, which has to be fed only with a small stream of steam regenerated solvent (Fig. 51). This makes the activated MDEA process attractive in terms of steam consumption. A typical aMDEA plant for the removal of carbon dioxide from steam reformer gas is shown in Figure 52.

The largest acid gas absorbers were built for a liquefied natural gas (LNG) terminal with a single train capacity of 1.6×10^6 m^3(STP)/h. Production of LNG requires bulk removal of all sulfur and carbon dioxide to avoid freezing in the liquefaction units. The aMDEA process of BASF is used to remove simultaneously hydrogen sulfide and carbon dioxide to the low ppm level. Additional molecular sieve adsorbers trap organic sulfur components such as mercaptans.

Degradation products of MDEA are secondary and primary amines, which also serve as activator for carbon dioxide absorption. Therefore in selective MDEA applications the carbon dioxide absorption is often higher than expected.

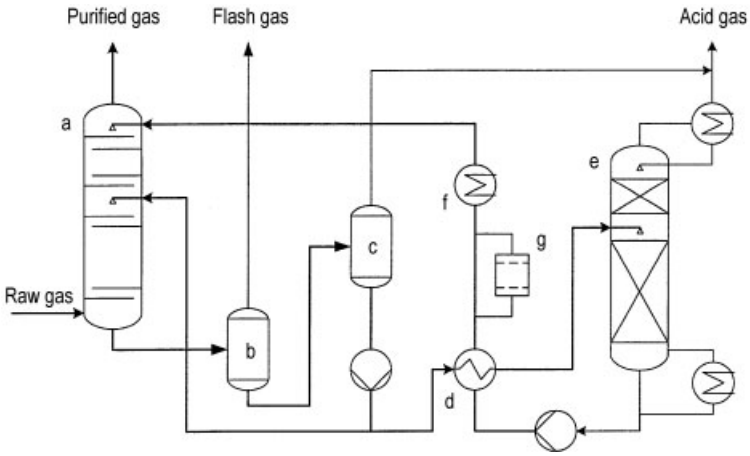

Figure 51. Activated MDEA process for carbon dioxide removal with semi lean amine regenerated by flash a) High pressure absorber; b) First flash drum; c) Second flash drum; d) Solvent heat exchanger; e) Regenerator; f) Solvent cooler; g) Amine filter

Figure 52. aMDEA plant removing carbon dioxide from 15 000 m^3(STP)/h raw gas of a steam reformer (Courtesy of BASF AG and Lurgi Oel Gas Chemie GmbH)

Sterically Hindered Amines. Solvents based on sterically hindered amines are licensed by Exxon Research and Engineering Co. under the trade name *Flexsorb* [192, 193]. These sterically hindered amines are secondary amine molecules bearing one bulky side group. This steric hindrance reduces the carbamate reaction rate by more than an order of magnitude compared to secondary amines. The basicity of this hindered secondary amine is higher than that of MDEA resulting in a higher hydrogen sulfide absorption capacity. Therefore, very high selectivity is attained in absorbers with short contact time [194].

Flexsorb SE solvent is a strongly hindered amine designed for high hydrogen sulfide selectivity [195]. It is therefore best suited for gases with high carbon dioxide: hydrogen sulfide ratio and in applications where low carbon dioxide absorption is required, such as Claus tail-gas treating. The benefits of Flexsorb have been demonstrated in various low pressure retrofit applications. Higher gas throughputs were achieved with lower solvent circulation rates compared to the original solvent. In addition the steam requirement for regeneration is reduced due to the lower solvent circulation rate.

With *Flexsorb SE Plus* hydrogen sulfide concentrations as low as 10 ppm (mL/m^3) at atmospheric absorber pressure are possible at the cost of 25 % higher circulation rates relative to Flexsorb SE. Acid buffer is added to the solution, which facilitates solvent regeneration.

Hot Potassium Carbonate Process. The potassium carbonate process has been successfully applied for carbon dioxide removal since 1960. The hot potassium carbonate process (HP) uses a potassium carbonate concentration in aqueous solution of 20 – 40 wt %. The reaction with carbon dioxide that occurs in the process also shows an equilibrium behavior that is favorable to absorption even at elevated temperature. As a result of the high absorber temperature (near the atmospheric boiling point of the solution), relatively low heat input is required for regeneration by reboiling.

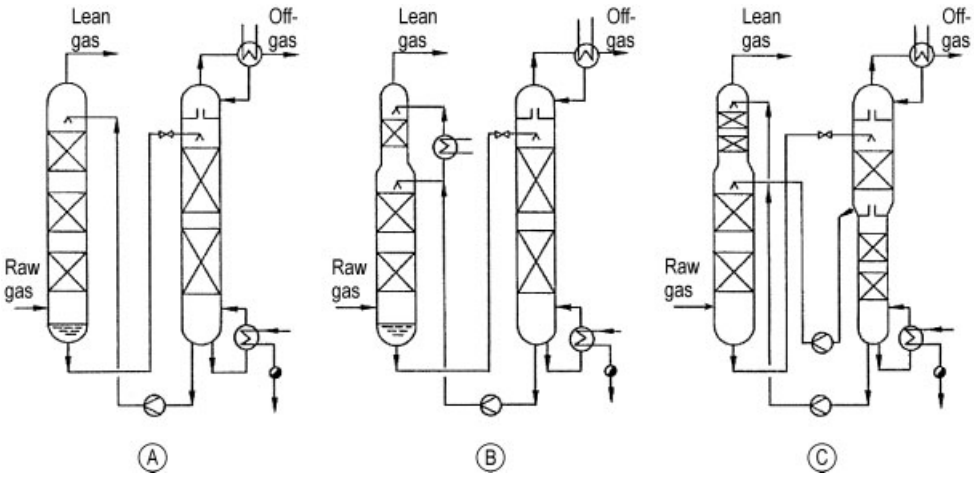

Figure 53. Typical flow diagrams of the HP process for CO_2 removal
A) Single stage; B) Single stage with splitflow; C) Two stage process

Investment costs are lower than for MEA and DEA because solvent heat exchangers are not required. Figure 53 shows three typical versions of the process. The single stage type (Fig. 53 A) can be varied by cooling part of the solvent to lower the vapor pressure of carbon dioxide so that a higher purity of the treated gas can be obtained (Fig. 53 B). A carbon dioxide content of less than 0.5 % in the treated gas can be obtained economically only by using a two-stage design (Fig. 53 C).

The main application of the HP wash is removal of carbon dioxide from converted reforming gas, because process conditions are almost ideal for this. The carbon dioxide partial pressure after conversion is normally in the range of 0.4 – 0.7 MPa and, therefore, in the optimum range for equilibrium behavior of the solution (see Fig. 42, line e). In the production of hydrogen and synthesis gas for ammonia, a methanation stage follows as a fine purification stage for converting carbon monoxide and carbon dioxide residues. This step is operated at high temperature and in the presence of water, as in carbon monoxide conversion. This is in keeping with the high absorption temperature, as well as the aqueous solvent basis of the HP process. In view of the 0.25 – 0.4 % carbon monoxide in the gas downstream of the shift conversion, the gas in the absorption stage does not have to be cleaned to a carbon dioxide content of less than 0.1 %. The converted gas contains enough steam to cover the heat requirements of the HP process.

Figure 54 shows a characteristic working diagram for a two-stage HP absorption process. In addition to the equilibrium curve, which is practically identical for absorption and regeneration, operating lines of the absorber and the regenerator, together with determination of the theoretical stages according to McCabe – Thiele, are shown. The slopes of the operating lines are determined by the split fraction of semi-lean and lean solvent. Optimum adjustment of the split fraction results in the minimum number of theoretical stages required.

The process is available in different versions. Modifications differ less in the basic principles of the process than in the additives used as corrosion inhibitors or activators to increase mass transfer.

Modifications. The *Benfield process* uses mainly solutions with 25 – 30 % potassium carbonate and vanadate as an additive. Several developments have led to its widespread commercial application, especially with regard to ammonia plants based on natural gas and naphtha [196].

The solution used in the *Catacarb process* contains proprietary additives. Amine promoters enhance mass transfer rates and inhibit corrosion in applications such as purification of ammonia synthesis gas, hydrogen production, and natural gas. If no carbon dioxide is present a solution of potassium borate is used for hydrogen sulfide removal instead of potassium carbonate.

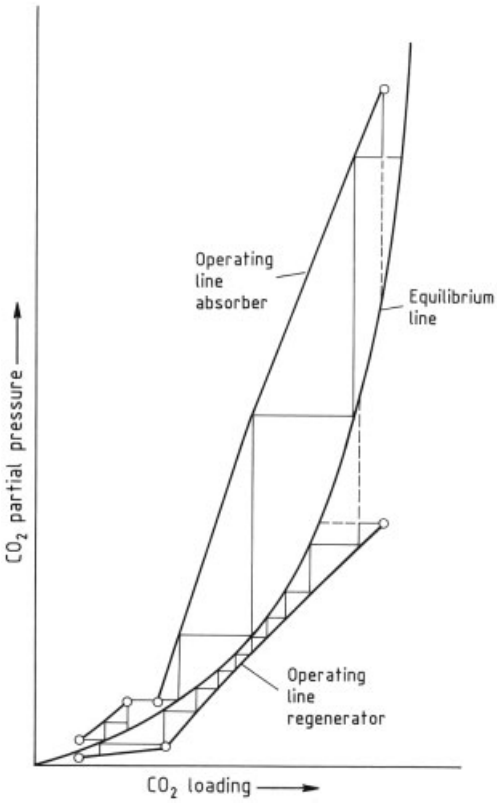

Figure 54. McCabe – Thiele diagram for a two-stage HP process

The *Lurgi HP process* uses borate as an additive, which accelerates absorption and is effective as a corrosion-inhibiting compound. In this process modification, both carbonyl and hydrogen sulfides can also be absorbed in the hot potassium carbonate solution. In the *high-purity version*, for example, hydrogen sulfide is removed to < 1 ppm. It can be absorbed selectively against carbon dioxide under certain conditions because of their different reaction velocities. The selectivity, together with other properties of the HP process (high absorption temperature, low investment and utility costs), is particularly advantageous for purifying coal gasification gas to fuel gas quality. Because coal gas usually contains additional impurities that cannot be removed in an HP wash, a precleaning step is required (see Section 5.4.4).

The regeneration operates under vacuum in the *Vacasulf process* of Krupp Uhde. The process is mainly applied for low pressure desulfurization of coke oven gas. Stringent lean gas qualities are achieved by a fine wash step with caustic soda.

Absorption in Ammonia Solution and Caustic Soda. Desulfurization with *ammonia solution* to purify gases containing ca. 500 mg of hydrogen sulfide per cubic meter was formerly very common for treating coke oven gas and is still used in some cases. However, the process has not become established for cases with more stringent purity requirements. On the other hand, it is an integral part of new developments for separating hydrogen sulfide – carbon dioxide and ammonia for sulfur recovery from sour gases containing ammonia, and for recovering ammonia from gas condensates.

Absorption of carbon dioxide and hydrogen sulfide in *caustic soda* produces bicarbonate and bisulfide the bonding of which is so strong that the vapor pressure of the sour components practically disappears. Therefore, gas purities as low as 0.1 mL/m^3 (CO_2 and H_2S) or less can be attained. However, regeneration, which is possible in principle but very complex, is dispensed with. Caustic soda is thus used only for fine purification.

5.4.2.3. Physical – Chemical Absorption Processes

Processes utilizing a solvent mixture of a physical organic component and a chemically acting component are termed physical – chemical absorption processes or mixed solvent processes. Mixed solvents provide a higher capacity than the pure physical solvent because of the additional chemical bonding of the acid gases with the chemical acting constituent. Fewer stages are normally required in the absorber due to the acceleration of mass transfer by the chemical reaction. The physically acting constituent of the mixed solvent provides absorption capacity for trace components such as carbonyl sulfide and organic sulfur components, which can not be removed with aqueous amines.

The best known processes are the Shell Sulfinol process and the Lurgi Amisol process.

Sulfinol Process. The Sulfinol solvent is a three-component solution containing the following constituents: an alkanolamine, usually

DIPA (diisopropanolamine), sulfolane (tetrahydrothiophene dioxide), and water. The ratio of the constituents varies according to the requirements of the application. The process, introduced in 1964, was used in its early years to remove hydrogen sulfide and carbon dioxide from gasification gases in steam reforming, as well as from natural gases [197]. More than 180 plants are reported to be in operation in 1998.

The process displays many properties and advantages of both chemical and physical processes. Acid gas loading is higher at elevated pressure and less steam for regeneration is required compared to purely chemical solvents. The physical character of the solvent facilitates to achieve hydrogen sulfide concentration as low as 1 ppm (mL/m^3) in the purified gas. Sulfinol is capable of removing other sulfur compounds including mercaptans, carbonyl sulfide, and carbon disulfide. The process flow sheet is very similar to chemical absorption processes (Fig. 49). Flashing of the rich solvent is often included as additional step to recover valuable gases and to increase the hydrogen sulfide concentration in the Claus gas.

The *Sulfinol-M process* uses MDEA instead of DIPA as chemical constituent in the solvent. This process can be designed for selective hydrogen sulfide absorption while a large percentage of the carbon dioxide from the raw gas slips to the treated gas.

Amisol Process. The Amisol solvent consists of an alkanolamine, normally MEA or DEA, in conjunction with methanol. This solvent mixture was successfully used in the first commercial Amisol plant in the early 1970s to remove carbon dioxide, hydrogen sulfide, and carbonyl sulfide from raw gas produced in a partial oxidation plant handling vacuum residues [198]. A clean gas with very low total sulfur content (hydrogen sulfide, carbonyl sulfide, mercaptans) of < 0.1 mL/m^3, as required for synthesis gas, can be obtained by the Amisol process. Good absorption of trace components such as hydrogen cyanide, and higher hydrocarbons is also achieved. Because of the low boiling point of methanol, absorption and regeneration temperatures are not far apart; therefore, no lean – rich solvent heat exchangers are normally required. Selective removal of both carbonyl and hydrogen sulfides versus carbon dioxide is possible by using *aliphatic alkylamines* such as diisopropylamine (DIPAM) and diethylamine (DETA) [199].

The Amisol process is suitable for treating various raw gases to absorb hydrogen sulfide, carbonyl sulfide, and carbon dioxide along with trace contaminants or to selectively remove sulfur compounds versus carbon dioxide. An example of a commercial Amisol plant using diethylamine involves purification of coal gas produced by an entrained-bed gasifier. A concept was developed for clean fuel power stations.

Flexsorb PS. The Flexsorb PS solvent is a mixture of a moderately hindered amine, a physical solvent, and water. This solvent is primarily suited for natural gas clean up. More than 95 % of carbonyl sulfide and mercaptans can be removed from the gas.

5.4.2.4. Comparison of Various Absorption Processes for Hydrogen Production and IGCC

Hydrogen Production Based on Residue Oil Gasification (see also 7.2)**.** Refineries require large amounts of hydrogen for hydrotreating and hydrocracking to meet stringent fuel specifications and to process heavier opportunity crude oils. The residues from visbreaker or hydrocracker units are an inexpensive feedstock for hydrogen production by gasification. The carbon monoxide from the raw gas is converted to hydrogen in a CO shift unit and the resulting raw hydrogen must be purified from acid gases and trace impurities. The selection of the appropriate gas purification technology depends on the required hydrogen purity, the value of the fuel gas and CO2 byproducts, and the plant emission regulations.

Acid gases must be removed from the raw hydrogen downstream of the sulfur-tolerant shift unit. This can be achieved with various processes.

In the first case (see Fig. 55A), the Rectisol process produces an H_2S-rich sour gas suitable for the Claus unit. In addition, CO_2 bulk removal is economical and a pure CO_2 stream is generated which can be used for sequestration or as industrial gas, or vented without further treatment. A pressure swing adsorption (PSA) unit is

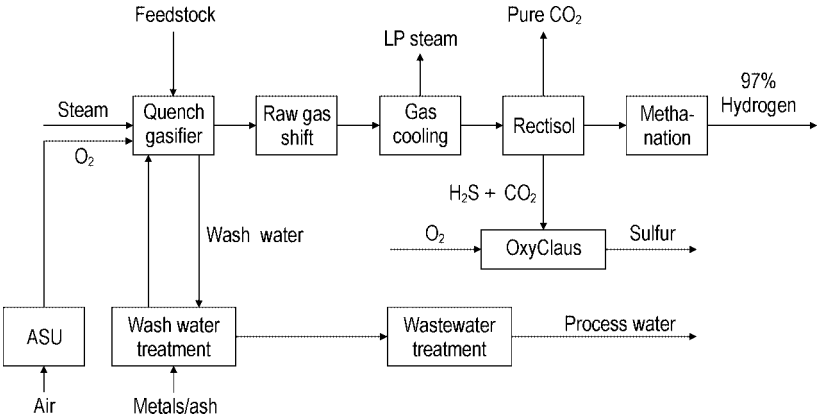

Figure 55. Gas purification options for hydrogen production from gasification of heavy residue oil
A) Rectisol-based acid gas removal and PSA for hydrogen fine purification; B) Selective acid gas removal and PSA for hydrogen purification; C) Acid gas fine removal by Rectisol and methanation for hydrogen purification

used to free the hydrogen from impurities such as methane, nitrogen, carbon monoxide, and argon, which are concentrated in the PSA off-gas. The PSA off-gas is a high-BTU, low-sulfur gas, since Rectisol removes besides H_2S also economically virtually all the CO_2 and other trace acid components.

In the second case (see Fig. 55B) a selective acid gas removal (AGR) process is applied such as Purisol or Selexol, which also produces an H_2S-rich sour gas suitable for the Claus unit. Typically, 20–30 % of the CO_2 content of the raw gas is removed in a selective gas purification process. This CO_2 is produced as a separate, impure CO_2 stream with a significant sulfur concentration. The sulfur impurity in the CO_2 stream is normally not a problem for sequestration purposes or enhanced oil recovery. Simple venting, however, is usually not permitted.

Since AGR is selective, a considerable amount of CO_2 passes to the PSA unit in this case and is separated there. The PSA unit also separates the remaining sulfur impurities. Due to the high CO_2 content the PSA off-gas has a low heating value. It also contains a significant amount of sulfur because both Selexol and Purisol leave traces of H_2S in the low ppm range in the purified gas and remove carbonyl sulfide (COS) only partly.

In the third case (see Fig. 55C), if hydrogen production is to be maximized but high purity is not required and additional fuel gas is of low value, the PSA unit can be replaced by a methanation reactor which converts CO and CO_2 to methane. Since all impurities end up in the product hydrogen, the gas purification unit must remove all sulfur species to the ppb level and remove bulk CO_2 to the low ppm range. Rectisol is the process of choice for this configuration. Hydrogen purity exceeding 97 % can be achieved, which is normally sufficient for hydrotreating and hydrocracking applications, and no fuel gas is produced. This process configuration gives the highest specific hydrogen production.

Gas Purification for Power Production by Combined Cycle IGCC. In an integrated gasification combined power cycle (IGCC), fuel gas produced by partial oxidation of coal or residual oil is used to power a gas turbine (see 7.3). A gasification process is advantageously applied which recovers the energy of the hot gas leaving the reactor in a waste-heat boiler to produce high-pressure steam. The heat recovered from gasification and the exhaust gas of the gas turbine are utilized for a steam cycle (see Fig. 56). If the IGCC is installed in a refinery, additional hydrogen can easily be produced. Part of the desulfurized gas downstream of the acid gas removal is passed to a CO shift unit and the hydrogen is purified in a PSA unit.

Due to mechanical restrictions gas turbines usually operate at pressures below 30 bar, whereas gasification and gas cleaning operate advantageously at pressures above 50 bar. Higher operating pressures have the benefit of smaller equipment size. Additionally the increased solubility of the sulfur compounds in the

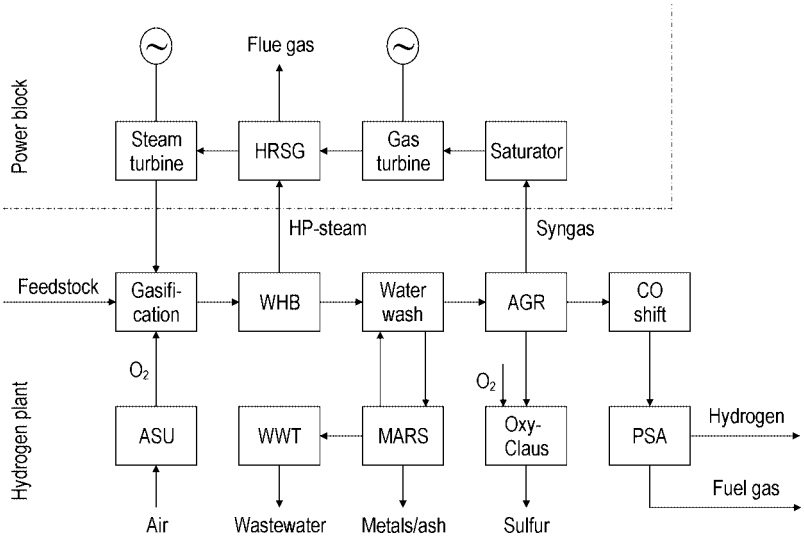

Figure 56. Cogeneration of power (IGCC) and hydrogen WHB-waste-heat boiler, AGR-acid gas removal, HRSG-heat recovery steam generation, PSA-pressure swing adsorption, MARS-metal/ash recovery system

absorbing solvent at higher pressure results in a lower solvent circulation rate that reduces utility demand for regeneration and solvent cooling.

The pressure difference between gasification and gas turbine can be utilized for additional power generation by expansion of the purified gas downstream of gas cleaning in a gas expander (see Fig57c). During expansion with power generation the gas temperature is reduced considerably. The expanded cold gas can be used for solvent cooling (d). The Rectisol process can be advantageously used for fuel gas production due to its ability to remove all sulfur components, including carbonyl sulfide. The required cooling to low temperatures can be done completely by the cold gas downstream of the expansion, if the pressure difference between gasification and gas turbine is high enough.

A comparison of a process with low-pressure gasification, gas purification, and gas turbine and a process with high-pressure gasification and gas cleaning, gas expansion, and low-pressure gas turbine is compiled in Table 29. The calculation is based on a residual oil gasification of 100 t/h and a net electrical power generation of 500 MW. Capital expenditures only for the gas cleaning unit are increased by 17 % by the gas expansion equipment, but for the complete IGCC plant the capital expenditures are increased only by 2 %. The utility demand is decreased by 13 %, and this leads to an overall increase of the electrical power generation efficiency of the IGCC from 45.4 to 46.2 %.

Table 29. Comparison of low pressure IGCC and high pressure IGCC with gas expansion

	High pressure gas cleaning with expansion	Low pressure gas cleaning
Gasifier pressure	58 bar	24 bar
Gas turbine pressure	22 bar	22 bar
Utility consumption		
Cooling water for gas cleaning	180 m^3/h	1230 m^3/h
Steam for regeneration	8.6 t/h	13 t/h
Electrical power for gas cleaning	− 0.4 MW	8.1 MW
Electrical power for air separation	32.6 MW	29.1 MW
Total electrical power	32.2 MW	37.2 MW
Capital expenditures (relative)		
Gas cleaning unit	117	100
Complete IGCC plant	102	100
Efficiency of electrical power generation		
Overall efficiency	46.2 %	45.4 %

An off-gas-free selective desulfurization for fuel gas production within an IGCC applying a selective absorption process such as Selexol and Purisol is shown in Figure 58. This process configuration operates successfully in an IGCC plant in Italy based on heavy residue gasification. Undesired trace components of the gasification such as cyanides or higher hydrocarbons are removed in a prewash stage (a). The rich solvent of the main wash is flashed in the reabsorp-

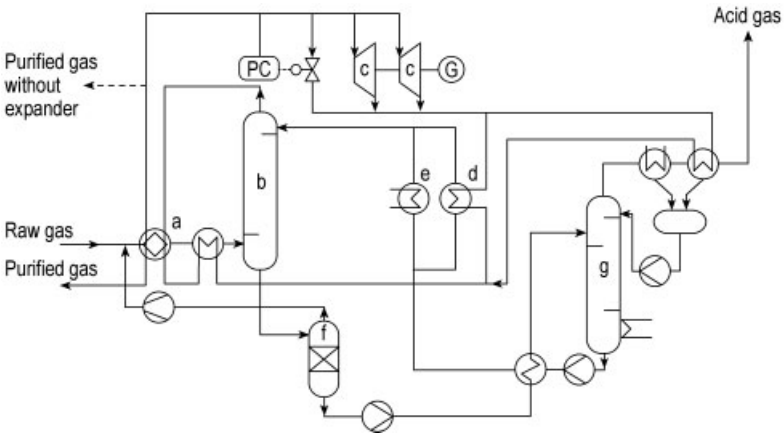

Figure 57. Rectisol for integrated gasification combined power cycle (IGCC) with gas expansion a) Crude gas cooling; b) H_2S absorber; c) Gas expansion turbines; d) Solvent cooling by expanded gas; e) Solvent cooling by refrigerant, f) Flash regeneration, g) Hot regeneration

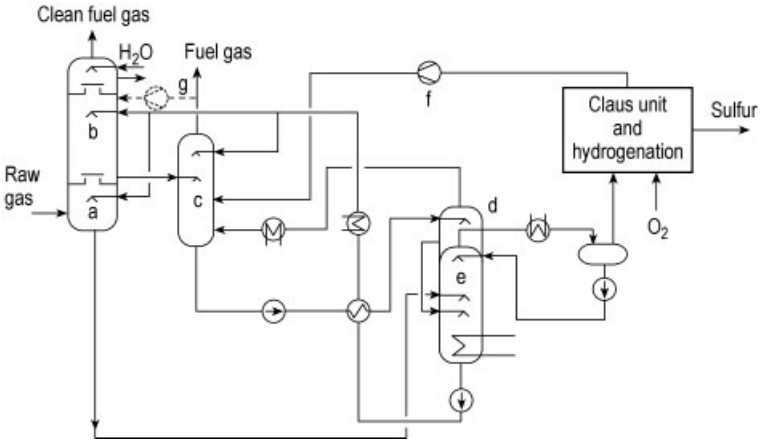

Figure 58. Off-gas free Purisol process for selective hydrogen sulfide removal in an integrated combined power cycle (IGCC) a) Prewash section, b) H_2S absorber, c) Reabsorber, d) Flash regeneration, e) Hot regeneration, f) Claus tail gas recycle blower, g) Fuel gas recycle compressor (optional)

tion column (c). Carbon dioxide and valuable components are recovered as fuel gas, whereas hydrogen sulfide is reabsorbed in the solvent in order to increase the hydrogen sulfide concentration in the sour gas to the Claus unit. The fuel gas leaving the reabsorption column overhead can either be recompressed and added to the high-pressure purified gas or can be used as low-pressure fuel gas for the steam boiler of the IGCC. The typical sulfur recovery rate of a two-stage Claus unit is 95 - 96 %. The off-gas of the Claus unit contains beside unconverted hydrogen sulfide considerable amounts of SO_2 and COS, which are converted to hydrogen sulfide in a catalytic hydrogenation step. The hydrogenated off-gas is compressed and fed to the reabsorption step of the acid gas removal unit.

So-called zero-emission IGCC (Fig. 59) is discussed in view of the Kyoto protocol for reducing carbon dioxide emissions. In addition to total sulfur removal, carbon from the feedstock is also captured and concentrated in a second pure carbon dioxide product stream. The fuel gas to the gas turbine is pure hydrogen. The gasification process converts the carbon of the feedstock mainly to carbon monoxide, which is further converted to carbon dioxide and hydrogen in a carbon monoxide conversion unit (→ 5.1). The carbon dioxide can easily be separated from the hydrogen in the absorption process and

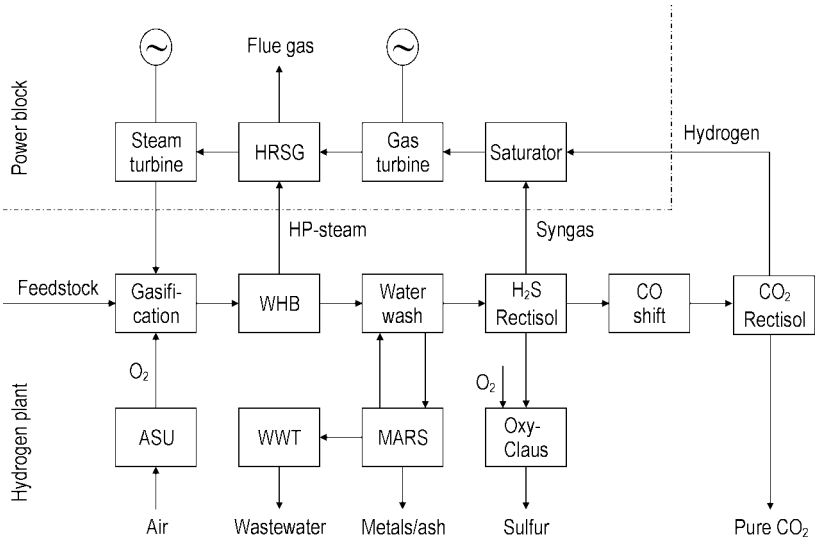

Figure 59. "Zero-emission" IGCC concept WHB-waste-heat boiler, AGR-acid gas removal, HRSG-heat recovery steam generation, PSA-pressure swing adsorption, MARS-metal/ash recovery system

a pure carbon dioxide stream is produced, which can be used for industrial purposes such as urea production, for enhanced oil recovery, or can be sequestered for carbon dioxide emission control. The economically achievable carbon recovery rate and the purity of the carbon dioxide product stream depend on the selected absorption process. The Rectisol process allows the production of ultralow-sulfur fuel gas (hydrogen) to the gas turbine and generation of very pure carbon dioxide. Selective-solvent processes such as Selexol and Purisol leave more sulfur in the fuel hydrogen. As expressed by the designation selective, this type of solvent requires greater effort for removal of bulk carbon dioxide. In addition, the carbon dioxide product stream contains larger amounts of hydrogen sulfide. Chemical absorption processes are not particularly suitable for removing these large amounts of carbon dioxide due to the large amount of steam required for solvent regeneration.

5.4.3. Liquid-Phase Oxidation Processes

In liquid phase oxidation processes hydrogen sulfide is absorbed from a gas stream in a liquid solution and directly oxidized to sulfur in the solution. The overall reaction can be expressed as follows:

$$H_2S + 1/2\,O_2 \rightarrow S + H_2O$$

The oxidizing agent in the solvent is reduced and solid sulfur is formed in the gas liquid contactor. In the subsequent regeneration step the solution is oxidized with air converting the oxidant to its oxidized state.

The main process steps are shown in Figure 60. The feed gas is contacted with the redox solution in a suitable gas – liquid contactor such as a countercurrent packed bed absorber, a spray tower, a cocurrent sparged tower, or a venturi scrubber. The sulfur formation takes place in the absorber and is completed in a downstream reaction vessel which may also serve as gas – liquid separator if the absorber works in the cocurrent mode. In the oxidizer the active component is converted from the reduced to the oxidized state and the sulfur particles are separated either as froth by flotation or as slurry by settling. The regenerated solution is recycled to the gas – liquid contactor. The sulfur froth or slurry is either separated in a filter unit or a sulfur melter.

Liquid-phase oxidation processes are especially suitable for selective removal of hydrogen sulfide versus carbon dioxide, because carbon dioxide does not react with the redox solution. Essentially complete hydrogen sulfide removal with concentrations below 1 ppm (mL/m^3) in the treated gas is attainable. Contrary to physical or chemical absorption processes in which the off-gas of the regenerator must be generally treated

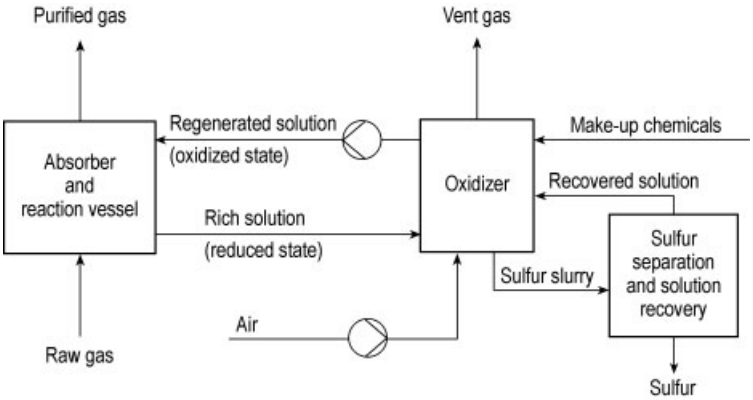

Figure 60. Schematic flow diagram of liquid phase oxidation processes

in an additional unit to produce sulfur or sulfuric acid, in liquid-phase oxidation processes the formation of elemental sulfur occurs within the purification unit itself.

The main applications of liquid-phase oxidation processes are

1) Desulfurization of coke oven gas, natural gas, and synthesis gas with low hydrogen sulfide concentration,
2) Treatment of gases that cannot be processed economically in a Claus plant (e.g., due to low hydrogen sulfide concentration, small total amount of sulfur),
3) Treatment of Claus tail gases,
4) Replacement of dry purification processes, and
5) Treatment of very low raw gas rates (e.g., landfill gas).

Economical operation is limited to sulfur production rates less than 20 t/d due to high utility costs. Continuous solvent make-up is required because the solvent is susceptible to chemical and thermal degradation and because of losses with the sulfur filter cake. Continuous solvent purging is required due to the accumulation of side products such as thiosulfate in the solvent circulation. Thiosulfate production can be kept below 1 % of the hydrogen sulfide converted, if operating conditions such as temperature and solution pH are carefully controlled. A buffering agent in the solution avoids large pH swings due to carbon dioxide absorption.

Carbonyl sulfide and organic sulfur components are only partly removed in liquid oxidation processes. Hydrogen cyanide and sulfur dioxide are almost quantitatively absorbed and react to thiocyanate and sulfate in the solution. These salts accumulate in the circulating liquor and increase the required purge stream and the amount of hazardous waste byproducts generated. Ammonia is also absorbed and causes pH changes. Ammonium polysulfide is formed with solid sulfur, increasing the salt concentration in the solution. Ammonia is stripped in the oxidizer and may cause odor problems in the vent stream.

Sulfur quality is often not crucial because of the small amount of sulfur produced. Residual solvent impurities have to be removed from the filter cake by washing with demineralized water. An additional sulfur washing and melting step is required to achieve sulfur quality similar to Claus plant sulfur.

A drawback of all liquid-phase oxidation processes is the relatively low oxidation capacity of the solution for hydrogen sulfide. This necessitates large solvent circulation rates. The heat of reaction generated by the oxidation of hydrogen sulfide must be dissipated at low temperature and cannot be used, as in the Claus process, to generate valuable steam. Only part of the heat of reaction can be used for the heat demand of the process.

The development of liquid phase oxidation processes started with iron oxide suspended in aqueous alkaline solutions (e.g., Ferrox, Gluud, and Manchester). The second generation of processes applied thioarsenate (e.g., Thylox, Giammarco–Vetrocoke [200]). They have lost market acceptance because of the toxicity of the solution. An aqueous mixture of vanadium pentoxide, naphthoquinone, and alkali carbon-

ate is the basis of the widely used *Stretford* process [201], the *Takahax* process [202] and the *Sulfolin* process. Presently, environmental concerns about the vanadium content in the purge stream have resulted in the selection of iron chelate-based processes such as *Lo-Cat* [203], *SulFerox* [204] and *Sulfint* [205] for most applications. Processes using sulfur dioxide are *Townsend* [206] and *IFP Clauspol* [207]. In the end they have not been commercially successful. A new class of processes apply microorganisms to convert hydrogen sulfide to sulfur (e.g., Shell-Thiopaq [208] and Bio-SR).

Vanadium-Based Solutions. The *Stretford* process, originally developed for the removal of hydrogen sulfide from coke oven gas, has proved to be equally suitable for desulfurization of other gas streams such as natural, synthesis, and tail gases.

The Stretford process uses an aqueous solution containing sodium carbonate, sodium bicarbonate, anthraquinonedisulfonic acid (ADA), and sodium metavanadate. The process can be considered to proceed in the following steps:

1) Absorption of hydrogen sulfide in the alkali solution,

$$H_2S + Na_2CO_3 \rightarrow NaHS + NaHCO_3$$

2) Oxidation of hydrogen sulfide to sulfur by sodium metavanadate,

$$4\,NaVO_3 + 2\,NaHS + H_2O \rightarrow Na_2V_4O_9 + 2\,S + 4\,NaOH$$

3) Reoxidation of the reduced vanadate by ADA

$$Na_2V_4O_9 + 2\,NaOH + H_2O + 2\,ADA \rightarrow 4\,NaVO_3 + 2\,ADA\,(reduced)$$

4) Oxidation of reduced ADA by air in the regenerator

$$2\,ADA\,(reduced) + O_2 \rightarrow 2\,ADA + 2\,H_2O$$

Figure 61 shows a simplified flow diagram of the Stretford process for selective desulfurization. The raw gas is brought in contact countercurrently with the solution in the absorber (a) and purified to a level of < 1 ppm of hydrogen sulfide. The solution flows from the absorber bottom to the reaction vessel (c) to complete the conversion of hydrogen sulfide to elemental sulfur; then the solution is regenerated in the oxidizer (d) by contact with air.

Sulfur is separated by flotation and removed at the top of the oxidizer (d) as froth containing 5 – 10 % solids. The sulfur froth is collected in a vessel and subsequently processed in a filter or centrifuge. The sulfur has to be washed with water to separate solution and sulfur and to produce relatively pure sulfur. Though installed in numerous commercial plants, the Stretford process suffers from the drawbacks of formation of byproducts and environmental concern about vanadium in the effluent streams.

In the solvent used in the *Sulfolin process* ADA is replaced by an organic nitrogen vanadium promoter [209]. Like the Sulfolin process the *Unisulf process* [210] claims a reduced formation of byproducts and lower corrosivity, resulting in lower chemical consumption.

Iron-based Solutions. Iron is kept in the aqueous solution by organic ligands or chelating agents (L). The basic chemistry in the iron chelate solution can be summarized as follows:

Absorption:

$$2\,Fe^{3+}L + H_2S \rightarrow 2\,Fe^{2+}L + S + 2\,H^+$$

Regeneration:

$$2\,Fe^{2+}L + 1/2\,O_2 + 2\,H^+ \rightarrow 2\,Fe^{3+}L + H_2O$$

In iron chelate processes sulfur is generally separated by settling. The slurry containing 10 – 15 wt % sulfur is further processed by filtration in order to recover the solvent. Stainless steel equipment is required due to the high solvent power of the solution for iron.

The *Lo-Cat* process is licensed by U.S. Filter. The flow sheet of the conventional Lo-Cat process is similar to the SulFerox process shown in Figure 63. Oxidation of the rich solvent and sulfur settling take place in the same vessel. The solution contains up to 0.3 wt % iron resulting in large solvent circulation rates with high power demand for solvent pumping especially in high pressure gas purification.

In the Lo-Cat *autocirculation design*, hydrogen sulfide absorption, the conversion to sulfur, and the oxidation of the redox solution take place in one vessel in separate compartments (Fig. 62). The autocirculation design is applicable for low pressure gases such as acid gases from amine units and Claus tail gases. Due to the compact design hydrogen sulfide is not totally

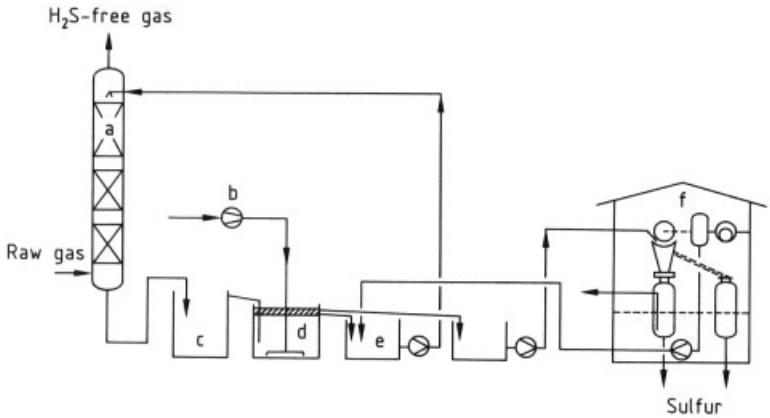

Figure 61. Schematic of a Stretford plant for selective H_2S removal
a) Absorber; b) Air compressor; c) Reaction vessel; d) Oxidizer; e) Solvent vessel; f) Sulfur filter

converted to sulfur before the solution is contacted with air. This causes an increase in the thiosulfate formation of up to 5 % of the converted hydrogen sulfide.

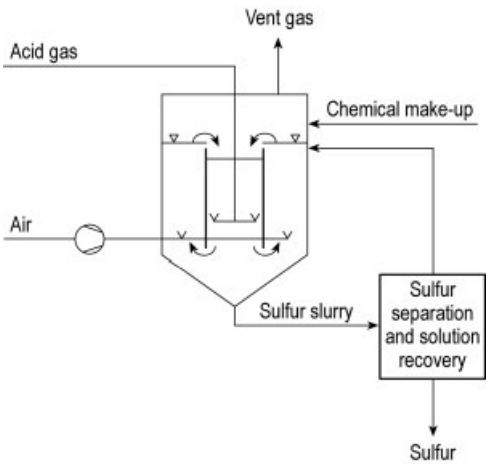

Figure 62. Lo-Cat autocirculation design

The Lo-Cat process operates with the stoichiometric ratio of iron(III): hydrogen sulfide of 2: 1. In the Lo-Cat II process the ratio of iron(III) to hydrogen sulfide is maintained at 1: 1. An improved solution chemistry enhances the reaction kinetics. The rich solvent from the absorber is mixed with oxidized solution upstream of the oxidizer. Residual HS^- in the rich solvent is thus totally converted to sulfur by iron(III) before the solution is contacted with oxygen resulting in thiosulfate formation below 1 %. Due to the substoichiometric operation less solvent has to be circulated to the high pressure absorber, reducing the power requirement for pumping. The Lo-Cat II process is also available in the autocirculation design.

The *SulFerox* process was developed by Shell and is licensed by DOW and Shell (Fig. 63). Special feature of this process is the high iron concentration of 2 to 4 wt %, reducing equipment sizes and pumping costs. The high iron concentration accelerates the reaction kinetics and thus the hydrogen sulfide absorption rate. One possible design for the absorption section is the patented cocurrent pipeline contactor with proprietary mixing elements which makes use of the fast reaction kinetics. Low pressure drop, nearly infinite hydrogen sulfide turndown capability, and low fouling tendencies are provided in a small diameter vessel. Due to the fast reaction kinetics all of the hydrogen sulfide absorbed is converted to sulfur and no HS^- ions are present in the rich solution when entering the oxidizer. Thiosulfate formation is claimed to be below 0.1 %.

The pH of the SulFerox solution is kept below 8. Carbon dioxide solubility in the SulFerox solution is therefore similar to that in water, resulting in a very high hydrogen sulfide selectivity. The SulFerox solution is also capable of removing about 70 % of mercaptans present in the raw gas.

A comparison between one of the amine processes plus Claus plant, autocirculation Lo-Cat, direct Lo-Cat, Lo-Cat II, and the SulFerox pro-

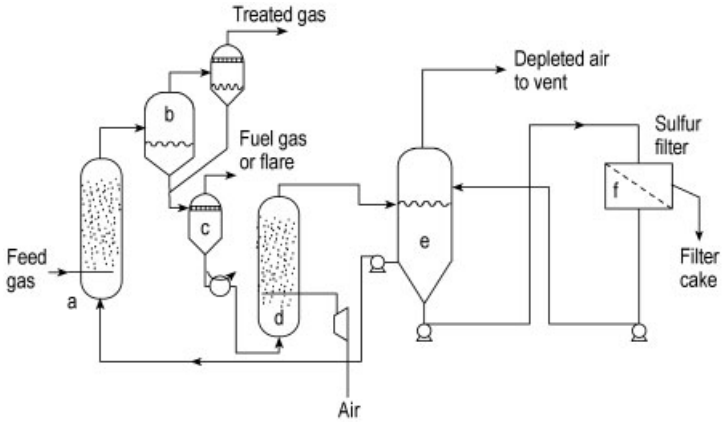

Figure 63. Schematic flow diagram of the SulFerox process. A) Absorber, b) Gas-liquid separator, c) Flash drum, d) oxidizer, e) Sulfur settling vessel, f) Sulfur filter (Courtesy of Shell International Oil Products B.V.)

cess in technical and economical terms was performed by the Gas Research Institute (GRI) [213]. Lo-Cat II and SulFerox perform very similar for the two high pressure gas cleaning cases studied. The two Lo-Cat modifications and an amine process plus Claus plant result in significantly higher sulfur production costs.

Biological Processes. Suitable bacteria convert hydrogen sulfide under certain conditions to sulfur. The processes consist of the same process steps as conventional liquid phase oxidation processes (Fig. 60). In the *Shell-Thiopaq process* hydrogen sulfide is absorbed into an iron-free caustic solution. The dissolved sulfide is subsequently oxidized to sulfur in the Thiopaq reactor by bacteria of the genus Thiobacillus. The *Bio-SR process* of NKK Corporation uses a nonchelated iron(III) sulfate solution. The bacteria act as catalyst in the oxidizer to enhance the kinetic of regeneration from iron(II) to iron(III) sulfate. Both processes have demonstrated to be capable for removing hydrogen sulfide from high pressure natural gases.

5.4.4. Removal of Gas Impurities of Low Concentration

Raw gases contain numerous gaseous impurities besides the acid gases. High boiling hydrocarbons have to be removed because of dew point specifications in the purified gas. Organic sulfur compounds are harmful to most catalysts. Unsaturated hydrocarbons tend to polymerize in the solvent circulation. Acids, hydrogen cyanide, and ammonia tend to form salts in the absorption solution. Gases from residual oil or coal gasification often contain volatile metal compounds such as mercury and metal carbonyls.

The absorption of impurities depends on the solvent. Higher hydrocarbons will only be partly absorbed in water-based amine solvents and cause foaming problems. Low-boiling impurities are trapped in organic solvents such as Selexol and Purisol and can not be removed by solvent stripping. Rectisol has the advantage that the impurities absorbed are purged together with the process water from the solvent loop. Although the concentration of impurities in the gas is often in the ppm range they may accumulate in the solvent circulation to levels that cause problems such as fouling, foaming, corrosion, and solvent degradation. Therefore they are often removed in a separate prewash process upstream of the acid gas removal.

Water Wash Process. Even a simple water prewash can remove acids, cyanide compounds, and ammonia very effectively from the gas, because their water solubility (e.g., expressed by the absorption coefficient) is very high (see Fig. 64).

However, high solubility also means difficult regeneration, especially because purity requirements of wastewater are stringent and fresh water is scarce. Therefore, the laden water must often undergo complex stripping with steam and impurities must then be incinerated.

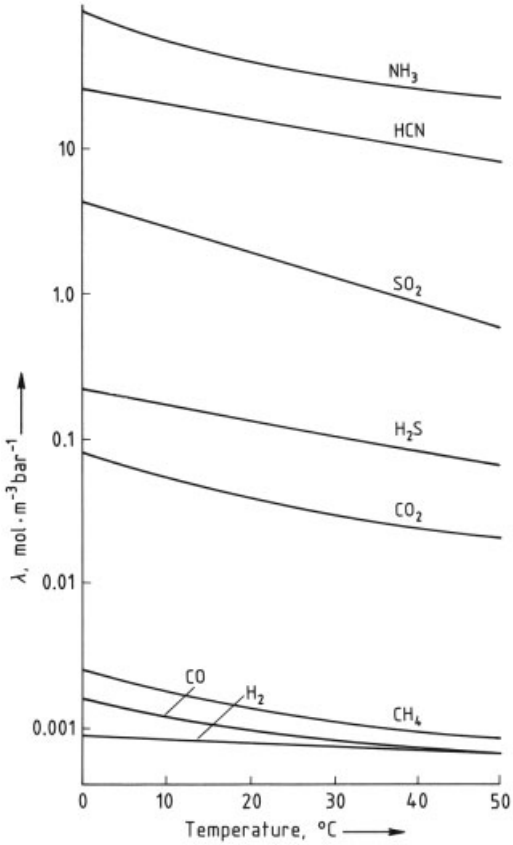

Figure 64. Solubility of different gases in water at low pressure 0.1 MPa

Oil Wash Process. Hydrocarbons, especially the normally undesirable unsaturated ones that tend to polymerize, as well as saturated high-boiling ones, are best removed by using an oil wash, which generally is designed with a hot regeneration stage (see Fig. 41 C). Occasionally, the vapor pressure of the washing agent in the absorber is too high so that a higher boiling oil fraction must be used. If the boiling temperature of this fraction is too high, it cannot be boiled indirectly in the regenerator with steam but only directly with live steam or a fired reboiler.

In some types of raw gas — especially that generated from coal — removal of the main gas impurities can require a whole series of prewash steps before highly concentrated gas components such as carbon dioxide are separated and chemically converted. Examples are prewashes with water for hydrogen cyanide and ammonia, followed by absorption of hydrogen sulfide in aqueous ammonia. In a final fine wash step benzene, naphthalene, gum formers, and other unsaturated hydrocarbons are absorbed in an oil wash.

Rectisol Prewash. The previously described prewash systems were particularly important for raw gas produced from coal pressure gasification, which contained almost all the impurities named, extending to gas naphtha. Therefore, attempts were made to find a simpler solution to this problem, which succeeded in conjunction with the development of the Rectisol process: all impurities are removed with only one solvent, methanol, which is also used downstream for removal of hydrogen sulfide and carbon dioxide.

Figure 65 shows a simplified scheme of such a prewash in which only a small quantity of cold methanol removes water, hydrogen cyanide, ammonia, gum formers, gas naphtha, and other impurities from the raw gas. Methanol is regenerated by extraction with water and two-stage distillation.

Condensation. Low boiling substances and water have to be removed from natural gases to meet pipeline dew point specification. This is often accomplished by condensation at low temperature. The *Ifpexol* process (Fig. 66) of Institut Français du Petrol (IFP) removes water and recovers liquid hydrocarbons from raw natural gas [211]. Part of the raw gas is brought into contact with a methanol rich stream in the methanol stripper (a). The methanol concentration in the water leaving the stripper as bottom product can be reduced to 50 mg/kg. The methanol vapor in the overhead raw gas prevents hydrate and ice formation in the gas chiller (b). The liquid condensate is separated in a settler vessel (d) into the liquid hydrocarbon fraction and the methanol – water fraction. The methanol – water fraction is recycled to the methanol stripper. The pretreated gas can be further purified by any suitable process.

Mercury Removal. Mercury emissions from coal-based power plants are of environmental concern. Conventional coal-fired power stations have to treat a huge volume of flue gas in an activated-carbon adsorber. Power production based on coal gasification (IGCC) allows easier removal of mercury. The Rectisol process

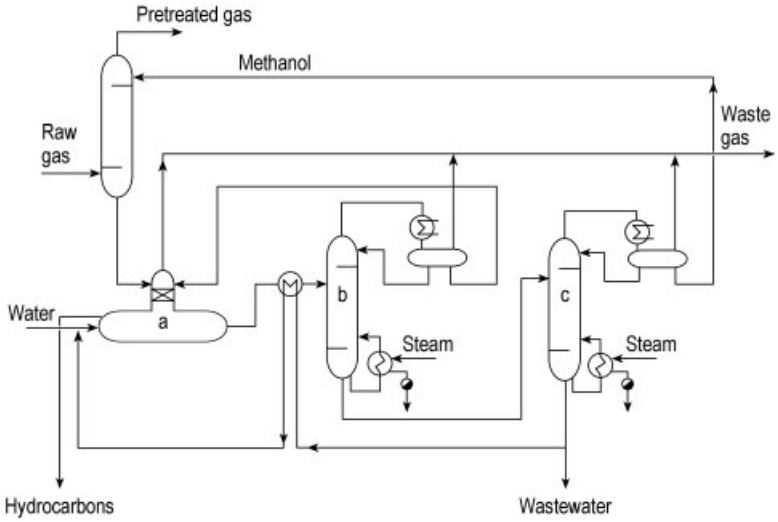

Figure 65. Rectisol prewash for raw gas from coal
a) Extraction system; b) Azeotropic distillation column; c) Methanol – water distillation column

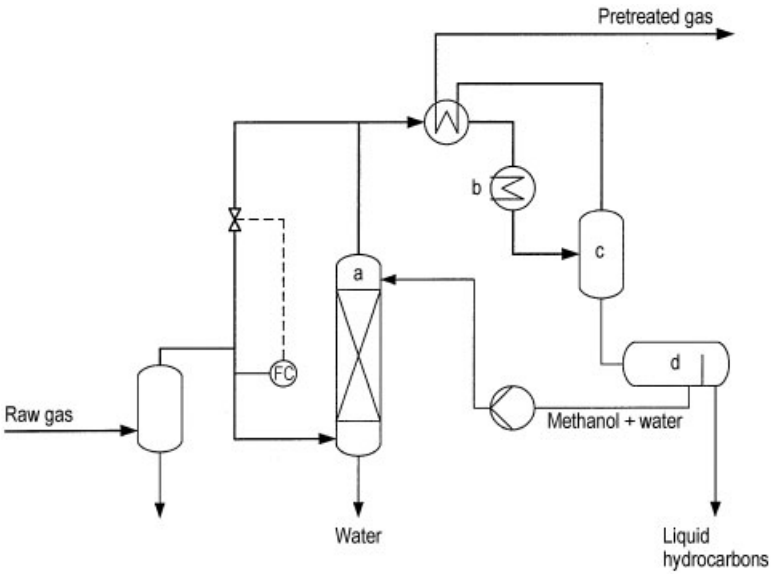

Figure 66. Gas pretreating by condensation with Ifpexol process. a) Methanol stripper, b) Gas chiller, c) Gas – liquid separator, d) Liquid – liquid separator

for acid gas removal also traps mercury with no additional adsorption step required. Other absorption processes require an additional activated carbon guard bed, which has to treat the smaller amount of fuel gas to the gas turbine.

5.5. Adsorption Processes

5.5.1. Fundamentals (→ Adsorption)

Three types of adsorbents are particularly suitable for the adsorption of gaseous components in fixed beds on an industrial scale:

1) activated aluminum oxide or silica gel,
2) molecular sieves, and
3) activated carbon.

Apart from process conditions such as pressure and temperature, the degree of adsorption on the above-mentioned solids is determined also by the structure of the adsorbent and the properties of gaseous components, such as volatility and polarity. Adsorption of substances that are characterized by high volatility and low polarity — typical examples are hydrogen and helium — is very poor or almost imperceptible, whereas water and carbon dioxide, for example, are well adsorbed because of their low volatility and high polarity. If process conditions are near the dew point of a gas component, that component is adsorbed preferentially.

Adsorption Mechanisms. The qualities of adsorbents are attributable in part to their characteristic material structures. Activated carbon (→ Carbon, Chap. 5), for instance, traps and retains gas molecules in minute pores measuring only a few molecular diameters.

In addition to this minipore effect, water molecules are adsorbed on silica gel or activated aluminum oxide in the form of water of crystallization, filling vacancies in the crystal lattice. In the case of molecular sieves (→ Zeolites), electrostatic forces come into play in addition to the minipores and surface-enlarging micropores, increasing their adsorption capacity, especially for polar and polarizable gases.

Adsorption Conditions. A look at adsorption processes in the molecular range suggests that these effects are improved by increasing process pressure and decreasing temperature.

For a particular gas and a selected adsorbent, adsorption capacity B is shown in Figure 67 as a function of pressure p at varying temperature T. This group of curves, which is characteristic of each adsorbate and adsorbent, cannot be calculated quantitatively but must be determined by empirical means.

Basic Modes of Operation in Adsorbent Beds. Figure 67 also shows the basic modes of operation possible for a fixed-bed adsorber. Three possibilities exist:

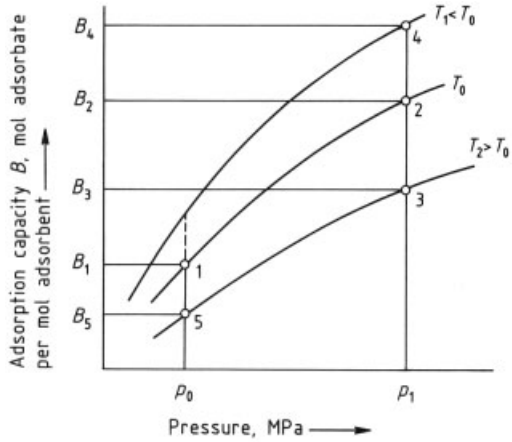

Figure 67. Adsorption capacity B of adsorbents as a function of pressure and temperature

First Basic Mode of Operation: Path 1 – 2. At T_0 and p_0, the adsorbent is loaded to the degree B_1. If the temperature T_0 is kept constant and the pressure is increased from p_0 to p_1, the adsorbent can collect the higher load B_2. The difference $B_2 - B_1$ is the adsorption capacity, which can be used for *pressure-swing adsorption* (PSA).

Second Basic Mode of Operation: Path 3 – 4. While the pressure is kept constant at p_1, the temperature is reduced from T_2 to T_{14}. This operation forms the basis of a *temperature-swing adsorption* (TSA) process.

Third Basic Mode of Operation: Path 4 – 5. This classic method of adsorption and regeneration uses the low temperature T_1 and high pressure p_1 for adsorption, and then regenerates the system by decreasing the pressure to p_0 and increasing the temperature to T_2 by heating. The adsorption capacity, described by the difference $B_4 - B_5$, is clearly the highest of the three cases.

Possible Applications. For gas conditioning, the adsorption of gases on fixed beds has three possible applications:

1) elimination of undesirable trace impurities from gas mixtures,
2) removal of contaminating gaseous components to yield a highly pure finished product, and
3) separation of substances.

Industrially useful applications can be derived only from an appropriate combination of gas components to be adsorbed and effective adsorbents. The following text describes a few typical processes used in practice.

5.5.2. The "Classic" Method

The classic method (third mode of application, above) may be illustrated by means of the removal of water and carbon dioxide traces.

If gas mixtures such as hydrogen – carbon monoxide, hydrogen – methane, nitrogen – methane, or air are separated into their individual components in cryogenic processes, traces of water and carbon dioxide will lead to icing and clogging of equipment. For this reason, such substances must be removed completely from the gas mixture at ambient temperature.

High carbon dioxide contents in the gas mixture are normally reduced by chemical or physical wash processes (see Section 5.4) to a level determined by the optimum arrangement of the wash process with a view to efficient subsequent adsorption.

The moisture content of the gas — as shown in the example in Figure 68 — can be significantly decreased by lowering the temperature to near the freezing point (ca. + 5 °C). The water and carbon dioxide traces remaining in the gas mixture can then be adsorbed efficiently and completely on a fixed bed.

Process Description. As shown in Figure 68, most carbon dioxide is removed from the gas mixture in a carbon dioxide scrubber, and the gas is cooled to ca. +5 °C. Condensed water is then removed in a downstream gas separator.

Loading Period. The gas mixture containing a residual quantity of 100 – 150 ppm of carbon dioxide and water is fed to adsorber (a), flowing through it from bottom to top. For water adsorption, adsorbers (a) and (b) may be filled with activated aluminum oxide, silica gel, or molecular sieves. The latter have been found most effective for adsorbing carbon dioxide. After carbon dioxide and water have been removed, the gas is fed to the cryogenic section.

Partial Regeneration by Flashing. At the same time, the pressurized adsorber (b) is isolated from the gas flow at working temperature and depressurized to regeneration pressure. The gas in the adsorber is lost to the process. Adsorbed water and carbon dioxide, which are liberated from the adsorbent as the system is depressurized, are removed from the adsorption bed by the regeneration gas flowing in the opposite direction to the gas flow during adsorption.

Fine Regeneration by Reheating. To complete regeneration, the adsorber bed is purged with hot regeneration gas containing neither water nor carbon dioxide. In the process, the adsorbent is heated to ca. 250 – 300 °C. It is then recooled with cold regeneration gas in the final stage. Regenerated adsorber (b) is pressurized again with clean process gas and can be used for another adsorption cycle.

Changeover. When adsorber (a) has been fully loaded, the gas flow is switched to adsorber (b) while adsorber (a) is regenerated.

Process Conditions. The gas composition and the length of the adsorption – regeneration cycles determine the size of the adsorption unit. The following conditions in particular must be met: (1) The total quantity of impurities in the gas mixture must not be too large so that the adsorbent volume — and, along with it, the size of the adsorbers — as well as the recycled regeneration gas rate are kept within reasonable limits. (2) The adsorbent quantity must be adequate for the adsorption stage to ensure that the heating and cooling time is sufficient for the required amount of regeneration gas during the simultaneous regeneration stage. With this arrangement, the adsorption and regeneration cycles normally require 4, 6, or 8 h.

Other Applications of the Classic Adsorber Loop. The system is also used to adsorb traces of water from gases before they are sent through pipeline systems. Moisture is an undesirable ballast, leading to corrosion along with carbon dioxide, hydrate formation with methane, and freezing when cooled. The maximum allowable moisture content, normally specified for water dew points from −5 to −30 °C, can be achieved for small or moderate gas flow rates at reasonable cost with adsorber units that use activated aluminum oxide, silica gel, or molecular sieves in long adsorption cycles, as well as the classic method of regeneration. Natural or town gas may be dried in this way before being fed to the gas grid. Similarly, adsorbers are

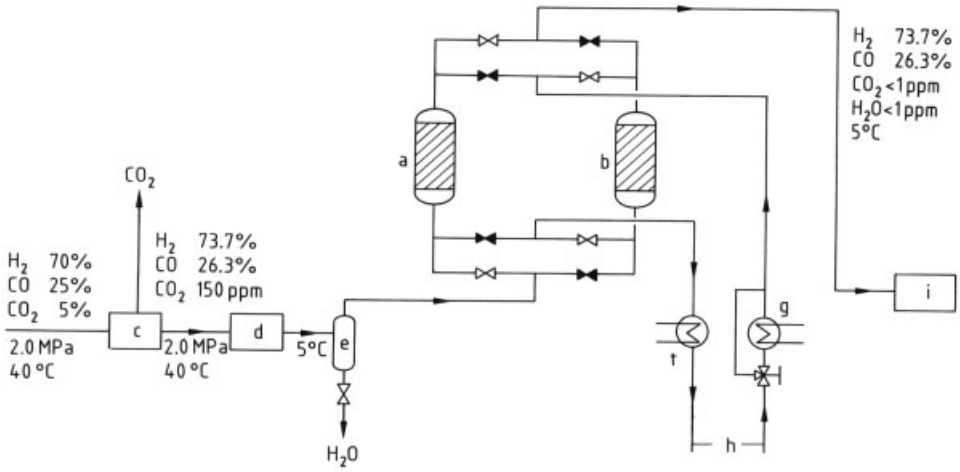

Figure 68. Reversible adsorber unit to eliminate residual carbon dioxide and water
a) Adsorber; b) Adsorber; c) CO_2 scrubber; d) Chiller; e) Gas separator; f) Cooler; g) Heater; h) Regenerator gas; i) Cryogenic process

used to adjust the residual moisture content of the control air, instrument air, and compressed air used in plants to a dew point normally of −25 to −30 °C.

5.5.3. Pressure-Swing Adsorption (→ Adsorption, Chap. 7.2)

The principle of the PSA process can be illustrated by the recovery of highly pure hydrogen. If a gas mixture containing hydrogen passes through a bed of molecular sieves at elevated pressure and ambient temperature, these sieves preferentially adsorb the high-density and polar gas components. Hydrogen, being the component least amenable to adsorption, remains alone in the gas phase, provided the adsorber bed is large enough for the gas rate flowing through it.

Process Description. At least two adsorbers are required to ensure continuous production of highly pure hydrogen. While one adsorber is in the adsorption stage, the other is regenerated by depressurization.

To purge the adsorbent bed, the flashed gases are discharged from the bed in the opposite direction to the flow during the adsorption stage. In this way, the pure gas released toward the end of the regeneration stage expels from the adsorber those components that are adsorbed first during the adsorption stage. The regenerated adsorbent bed is repressurized with pure product gas or with the gas mixture to prepare it for the next adsorption stage.

Process Conditions. Depressurizing and repressurizing adsorbers requires less time than heating and cooling the adsorbent bed during classic regeneration. Therefore, keeping the adsorber mass, and hence also the vessels, small by using short adsorption – regeneration cycles is cost-effective. Cycles between 5 and 10 min are efficient in practical applications. However, a drawback of the process is that one vessel volume of recoverable hydrogen is lost during each regeneration phase when the adsorber is flashed. This is why yields of highly pure hydrogen are generally poor relative to the recoverable hydrogen content of the gas mixture.

Special Process Loops. To improve the yield, the number of adsorbers may be increased from two to four, six, or ten.

In principle, these adsorbers or adsorber groups simultaneously operate in the following four modes (→ Adsorption):

1) adsorption in the first adsorber,
2) depressurizing the second adsorber,
3) purging the third adsorber at low pressure, and
4) pressurizing the fourth adsorber.

With this loop, losses of recoverable hydrogen are reduced by using the gas released when the second adsorber is depressurized to repressurize the fourth adsorber to medium pressure.

Product Purity. A four-adsorber loop can yield hydrogen of 99.999 % purity and produces a uniform product flow, achieving yields as high as 72 – 78 % of the hydrogen in the gas mixture. The types and quantities of contaminants in the gas mixture influence the size of the adsorber beds and, to a certain extent, the yields. However, they do not influence the function of the pressure-swing adsorption unit or the purity of the recovered hydrogen. Contaminants are discharged at ambient pressure together with the hydrogen used for purging. In the last ten years, this process has been used in several plants designed for an output up to 100 000 m^3 (STP)/h of highly pure hydrogen.

Multicomponent Separation [214]. Further developments of current adsorption processes aim, on the one hand, at improving the adsorbents and optimizing process loops and adsorption – regeneration cycles and, on the other, at utilizing various desorption effects depending on the respective components. In the past, adsorption processes were tailored to recover one product in a highly pure form (e.g., H_2 from gas mixtures, O_2 or N_2 from air). Modern developments take advantage of the different properties of individual components during desorption to increase the content of certain components in the outgoing flows. To this end, the origin of the desorption flash gas is considered, for example, raw gas from the inactive volume of the vessel and the intergranular volume, raw gas from the pore volume, and desorbed gas from the inner surfaces of the adsorbent at various desorption pressures (medium pressure, normal pressure, vacuum). The upgrading effect is reinforced by suitable process arrangements during desorption (e.g., cocurrent or countercurrent desorption), in combination with the passing of adsorbed gases to other adsorbent beds.

This process arrangement can be used successfully to separate gas mixtures consisting of hydrogen, carbon monoxide, and carbon dioxide into highly pure product hydrogen, a carbon monoxide fraction, and a residual gas fraction with an increased carbon dioxide content. Analogously, a mixture consisting of nitrogen, methane, and carbon dioxide can be separated into nitrogen waste gas, a high-methane fraction, and a carbon dioxide fraction. In both cases, the components participating in the process behave very differently because of their properties, not only during individual adsorption cycles but also during desorption cycles.

5.5.4. Adsorption on Activated Carbon

Activated carbon (→ Carbon, Chap. 5) shows a strong adsorptive effect because of its extensive pore system. Suitable preparation of carbonaceous materials can produce different pore structures, so that the distribution of pore radii may be matched to specific adsorption requirements. A variety of pore sizes down to micropores and minipores with diameters on the order of small gas molecules can be produced.

Applications. For gas purification, adsorption on activated carbon is used to protect downstream process stages or systems from traces of compressor lubrication oil or other low-volatility hydrocarbons, such as toluene and washing oil, which may have entered the gas flow through a special wash process. These adsorbers are characterized by high adsorptivity, especially for substances that are near their dew point or have been entrained with the gas flow in the form of minute condensate droplets. Activated carbon adsorbers can usually be regenerated by heating them and simultaneously purging the adsorbent bed with hot steam. Coking of the adsorbent, however, cannot be avoided entirely, so that the adsorption capacity of the activated carbon adsorber decreases with each regeneration cycle. Frequently, greater economy results from providing a single activated carbon bed of sufficient size, operating it to breakthrough, and refilling it with fresh activated carbon during a brief shutdown.

5.6. Cryogenic Processes (→ Cryogenic Technology)

Cryogenic processes are used to recover pure gas fractions from gas mixtures. Typical examples are *carbon monoxide* and *hydrogen*. The required purity of both products depends on their intended use. The accompanying substances and contaminants are more or less the same in all cases: hydrogen, carbon monoxide, carbon dioxide, residual methane, nitrogen, oxygen, argon,

steam, contaminants such as oils, saturated and unsaturated higher hydrocarbons, sulfur compounds, nitrogen oxides, and acetylene. Their concentrations vary according to the method used to produce the synthesis gas. Because the gas mixture is further processed, admixtures such as hydrogen or carbon monoxide, carbon dioxide or steam may lead to undesirable secondary reactions. Especially when unused gases are recycled to the process, nitrogen, argon and methane will lead to a build-up of inert components and thus impair the process conditions. Contaminants especially in the ppm range such as sulfur compounds, steam and acetylene, but also carbon monoxide in hydrogen, frequently poison the catalyst or produce secondary reactions at the catalyst.

In practice, a carbon monoxide product gas contaminated with not more than 1 – 3 % hydrogen, methane, or nitrogen meets the quality requirements for most downstream chemical processes. Hydrogen used for hydrogenation or ammonia synthesis must meet even more stringent purity requirements.

Sulfur compounds, acetylene, carbon dioxide, and water are removed from synthesis gas by washing, adsorption, and absorption. This reduces the problem to the recovery of pure hydrogen or carbon monoxide from a gas mixture essentially containing hydrogen, carbon monoxide, and methane. Cryogenic separation is a promising route, in addition to selective absorption and adsorption, for separating the components of this gas mixture.

Plant engineering has benefited from several cryogenic processes, and various process combinations have been used for the respective applications. Two principal process loops seem to be promising for *carbon monoxide* recovery from synthesis gas mixtures: (1) partial condensation and (2) the liquid methane wash process. A third process — the liquid nitrogen wash process — is used to recover *hydrogen* for ammonia synthesis.

5.6.1. Partial Condensation

The first stage of the partial condensation process for recovering hydrogen or carbon monoxide from synthesis gas mixtures is the liquefaction of higher boiling gas components such as carbon monoxide and methane. Depending on the pressure applied, liquefaction begins at ca. −100 to −180 °C, and ends when the sublimation temperatures of methane and carbon monoxide are reached.

The first components to liquefy in a hydrogen atmosphere are methane and carbon monoxide. Separation of the gaseous and liquid phases alone yields relatively pure product flows.

Because carbon monoxide recovery requires temperatures as low as −200 °C, for a cryogenic process to work properly all components that may freeze and clog the piping must be removed completely.

In partial condensation, carbon monoxide is condensed along "cold walls." This leaves a quantity of carbon monoxide, corresponding to the carbon monoxide vapor pressure, uncondensed. The "cold wall" is produced in a heat exchanger by flashing liquid carbon monoxide to the lowest possible pressure and evaporating it in countercurrent. This defines the coldest point of the partial condensation process: it is the temperature of liquid carbon monoxide evaporating at the lowest possible pressure [215, 216].

If the carbon monoxide condensate contains an unacceptably high percentage of methane, this must be removed by distillation. The carbon monoxide yield is 75 – 80 %. Losses are due to uncondensables in the hydrogen flow and to carbon monoxide removed along with methane. However, recovery of only part of the carbon monoxide may be desirable if hydrogen gas containing a certain proportion of carbon monoxide is required.

Process Description. Figure 69 shows a typical process flow diagram of a partial condensation unit. After all contaminants that might freeze and clog the piping have been removed from in a pretreatment stage, the gas mixture is cooled in heat exchangers (a) in countercurrent to the cold products coming from the plant. Part of the carbon monoxide and methane is condensed, and the gas and liquid are separated in a separator (b). Hydrogen gas, which still contains some carbon monoxide, is reheated. The liquid containing traces of dissolved hydrogen and methane is depressurized; dissolved hydrogen boils off and is removed in the downstream separator (c). The liquid, which no longer contains any hydrogen, is fed to a column (d) where the carbon monoxide – methane mixture is sep-

arated into pure carbon monoxide and a methane fraction containing a certain percentage of carbon monoxide. All cold products are evaporated at ambient pressure and heated in countercurrent to the hot incoming gas mixture. The carbon monoxide yield may be increased by modifying the process by additional process steps and recycling of carbon monoxide containing flash gases.

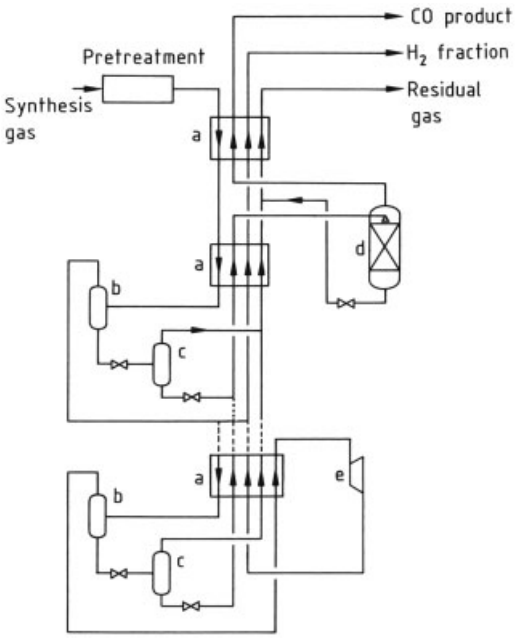

Figure 69. Cryogenic partial condensation process
a) Heat exchanger; b) Separator; c) Degasser; d) CO – CH_4 separator; e) Expander

In this process, hydrogen is obtained at its full pressure, which is not always required in a carbon monoxide recovery unit. The hydrogen produced can, therefore, also be used for flash refrigeration in an expander (e). This process element enables the coldest temperature to be lowered by ca. 20 to −200 °C, so that the amount of carbon monoxide remaining in the gaseous phase is greatly decreased and the carbon monoxide yield increased [215, 216]. Process results are given in Table 30.

Table 30. Typical analyses for the process flow diagram shown in Figure 69

	Feed gas	H_2 fraction	CO product	Residual gas
Without expander stage				
Analysis, mol%				
H_2	70	88	1.5	40
CO	27	12	98.5	50
CH_4	3			10
Volume parts	1000	650	50	300
With expander stage				
Analysis, mol%				
H_2	70	96.5	1.5	40
CO	27	3.5	98.5	50
CH_4	3			10
Volume parts	1000	600	100	300

5.6.2. Liquid Methane Wash Process

Liquid methane is used to wash a pressurized hydrogen flow slightly above the freezing point of methane (i.e., ca. −180 °C). During this process, carbon monoxide carried along with hydrogen is dissolved in methane; 100 % of carbon monoxide in the feed is recovered [217] so that hydrogen leaving the column overhead is free of carbon monoxide. For calculation methods applicable to isothermal wash processes, see [218].

Spent methane is regenerated by distillation, expelling carbon monoxide overhead. By using the heat pump principle and employing a flash refrigerator, carbon monoxide can be conveniently recycled and used to maintain the refrigeration balance.

Process Description (see Fig. 70). Pretreated synthesis gas is cooled in a heat exchanger (a) to the washing temperature of −180 °C. High-boiling components are thus condensed, and only the gaseous carbon monoxide in the hydrogen gas is washed out with methane in column (b). This washing methane, methane from the gas mixture, and carbon monoxide that has been condensed or washed out are fed to the carbon monoxide – methane separator (c). There, the carbon monoxide product is discharged overhead while the washing methane remains as the bottom product. A liquid methane pump (d) delivers the washing methane via a supercooler (e) to the top of the wash column (b), from which carbon monoxide-free hydrogen is discharged overhead. The product streams of carbon monoxide, hydrogen, and residual meth-

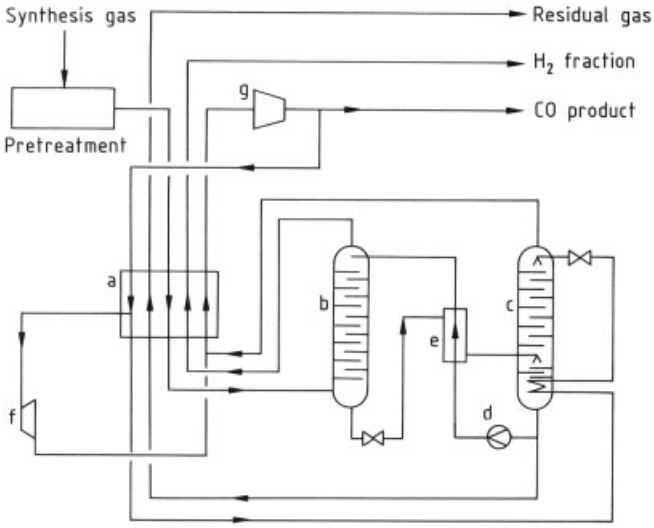

Figure 70. Liquid methane wash process
a) Heat exchanger; b) Liquid methane wash column; c) CO – CH_4 separator; d) Liquid methane pump; e) Liquid methane subcooler; f) Recycled CO expander; g) Recycled CO compressor

ane are heated in the exchanger (a), flowing in countercurrent to the feed gas, and recovered individually.

The carbon monoxide compressor (g) closes the refrigeration loop with the expander (f). The same compressor may also be used to adjust the carbon monoxide product to the required battery limit pressure. Process results are given in Table 31.

Table 31. Typical analyses for the process flow diagram shown in Figure 70

	Feed gas	H_2 fraction	CO product	Residual gas
Analysis, mol%				
H_2	70	98.5	1.7	
CO	27		98.5	
CH_4	2	1.5	0.1	100
Volume parts	1 000	705	275	20

5.6.3. Liquid Nitrogen Wash Process

A finishing treatment in a liquid nitrogen wash unit is the last process stage in the recovery of hydrogen for ammonia synthesis from hydrocarbons. In preceding stages, carbon monoxide is converted to carbon dioxide, producing additional hydrogen, and carbon dioxide is then eliminated completely from the synthesis gas. Hydrogen gas entering the liquid nitrogen wash unit contains the following undesirable substances and contaminants:

1) traces of unreacted oxygen from the gasification process,
2) residual methane from unreacted hydrocarbons,
3) some unconverted carbon monoxide, and
4) argon that entered the system together with the oxygen required for gasification

Whereas argon and methane in the synthesis gas are inerts and would unnecessarily burden the ammonia synthesis loop, the residual oxygen and carbon monoxide are poisonous to the ammonia synthesis catalyst. Except for an unavoidable residue of argon, the above-mentioned contaminants are all eliminated by means of a liquid nitrogen wash unit.

The refrigeration needed for this process is obtained by flashing the required amount of nitrogen for washing the synthesis gas into the purified hydrogen already in the cryogenic unit. By mixing the nitrogen with the hydrogen stream, the pressure of the cryogenic nitrogen is lowered from the full process pressure at which the pure nitrogen is cooled to the nitrogen partial pressure in the hydrogen stream, which more or less reflects the proportion of nitrogen (24 – 25 %) in the synthesis gas. Essentially, the positive excess enthalpy ΔH occurring during this process pro-

vides the refrigeration necessary for the entire process [219].

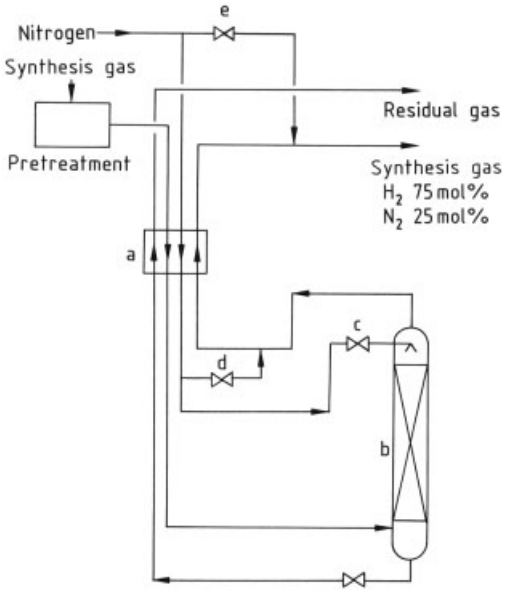

Figure 71. Liquid nitrogen wash unit
a) Heat exchanger; b) Wash column; c) Valve for wash nitrogen; d) Refrigeration valve; e) Trim valve

Process Description (see Fig. 71). Hydrogen gas containing contaminants such as oxygen, argon, carbon monoxide, and methane is cooled to ca. −180 °C in a heat exchanger (a). At a temperature around its dew point, the gas is fed to the bottom of the wash column (b). Liquid nitrogen trickling down countercurrently to the hydrogen gas dissolves the contaminants and is discharged with them from the bottom of the column. The overhead product is purified hydrogen with an nitrogen content that reflects the equilibrium stage of a hydrogen – nitrogen mixture at the of temperature and pressure prevailing at the top of the column. Nitrogen required in the wash unit is compressed to the process pressure, cooled in an exchanger (a), and fed to the wash column through a valve (c). Nitrogen needed for refrigeration is cooled in the same way, flashed through a valve (d) into the purified hydrogen, and evaporated in the exchanger (a) in countercurrent to the uncooled and untreated hydrogen and the nitrogen flow. The valve (e) is used to add nitrogen to establish the hydrogen – nitrogen ratio (3: 1) required for synthesis.

Wash nitrogen containing the dissolved contaminants is discharged from the column bottom, flashed to a lower pressure, evaporated, heated, and discharged as residual gas (for process results, see Table 32).

Table 32. Typical analyses for the process flow diagram shown in Figure 71

	H_2 gas	N_2 gas	Synthesis gas	Residual gas
H_2	94.11 mol %		75 mol %	8.31 mol %
N_2	0.05 mol %	100 mol %	25 mol %	26 mol %
O_2	ppm	ppm	2 ppm	ppm
Ar	0.60 mol %		50 ppm	7.14 mol %
CO	4.44 mol %		5 ppm	52.68 mol %
CH_4	0.50 mol %		10 ppm	5.87 mol %
Volume parts	1000	337	1253	84

5.7. Gas Separation by Membranes (→ Membranes and Membrane Separation Processes)

Fundamentals. Certain materials, when produced in the form of nonporous membranes of defined thickness, absorb gas components in their molecular structure at different rates and allow them to diffuse through the membrane with a certain pressure drop. The diffusion rate depends on (1) the type of material, (2) the size of the molecules, (3) the pressure differential across the membrane, and (4) the diffusion length (membrane thickness) and active area of the membrane. The following types of membrane can be distinguished:

Polymer membranes
Carbon membranes with micropores
Liquid membranes
Metallic membranes
Ceramic membranes

Whereas membranes based on ceramics, metals, and liquids have been known for a long time and have been developed to an industrial standard for some applications (e.g., palladium ceramics for the production of highly pure H_2), the development of polymer membranes for gas separation [220 – 225] has been accelerated since 1977 by the rapid development of reverse osmosis membranes for seawater desalination [226].

Several gas separation membrane systems, which compete in materials and design of the modules, can be found on the market:

1) *Prism separator* (Monsanto), an asymmetric porous membrane on a polysulfone carrier with a nonporous silicon-based coating, in the form of hollow-fiber bundles;
2) *Delsep* (Delta Engineering Corporation), an asymmetric porous membrane based on cellulose acetate with a nonporous cellulose acetate-based coating, in the form of a coil-type module with a flat membrane;
3) *Separex* (Air Products Company); and
4) *RO membranes* (DuPont), asymmetric porous membranes based on aromatic polyamide or cellulose acetate with a nonporous aramid-based coating, in the form of hollow-fiber bundles.

Applications. Hydrogen and helium have the best sorption and diffusion behavior of all gas components, and their separation efficiency is very good in comparison to nitrogen, methane, carbon dioxide, or carbon monoxide. Such gas separation units are, therefore, employed to recover hydrogen from reaction gas mixtures by using the same process principle as shown in Figure 72 in all cases.

Process Description. *Case 1: Gas Separation Membrane installed in a Slipstream.* Purge gas which must be extracted continuously from a synthesis gas loop (as installed for the ammonia production) to keep down the level of undesirable ballasts and nonreacting components essentially contains the same valuables as the loop (e.g., H_2, N_2, or CO). To improve overall efficiency of the synthesis loop, these valuables can be recovered by means of a gas separation membrane and recycled at low pressure to the suction side of the raw gas compressor. Those gas components that do not permeate the membrane (e.g., CH_4, Ar, CO, higher hydrocarbons) remain in the waste gas. Typical gas composition is shown in Table 33.

Table 33. Typical gas composition for purge gas treatment

	Raw gas	Permeate	Waste gas
H_2, vol %	74	96	57
CO, vol %	26	4	43
Flow rate, kmol/h	100	43.6	56.4
Yield		H_2: 56 %	CO: 93 %

Case 2: Gas Separation Membrane in the Main Stream. Hydrogen sulfide and carbon dioxide exhibit much better sorption and diffusion behavior than the product gases carbon monoxide and methane. This favorable separation efficiency can be used in a pretreatment stage to reduce the size of downstream conventional fine treatment (see Fig. 73). Most of the sour gases are removed by the gas separation membrane, although some valuable components are lost as well. Such losses of valuable components can be minimized by splitting the process into several stages and recycling the gases — a method which, of course, requires additional energy. Being preceded by a membrane, the absorption unit is turned into a fine treatment stage, whose dimensions and complexity can be reduced considerably.

5.8. Addition of Inerts or Other Substances

Normally, the quality of a gas is described not only by its composition but also by such characteristics as its calorific value and Wobbe index (see Chap. 8).

These characteristics are adjusted as necessary, either by adding inerts such as nitrogen or air to decrease the calorific value and increase the density of a gas, or by adding high-calorific substances such as liquefied gas, butane, or gasoline to increase the calorific value. Proper mixing is achieved by installing flow turbulators, providing spray injection of liquids into the gas stream, or adding densifiers.

If air is added, the oxygen level in the gas must be watched closely. Explosive air – gas mixtures, which may occur briefly in areas where air is injected into the gas, can be eliminated by flow turbulators.

If liquefied natural gas is used as feedstock, it can be conditioned to the required calorific value by adding liquefied nitrogen in the cryogenic liquid phase. Proper mixing is ensured as the gas evaporates at discharge pressure. This conditioning method saves energy because only a minimum of energy is required to pump the liquefied natural gas and liquefied nitrogen, rather than necessitating installation of compressor stations for natural gas, nitrogen, or the gas mixture.

Substances with a manifest odor—usually higher mercaptans—are added to normally odorless gases to warn of any leaks.

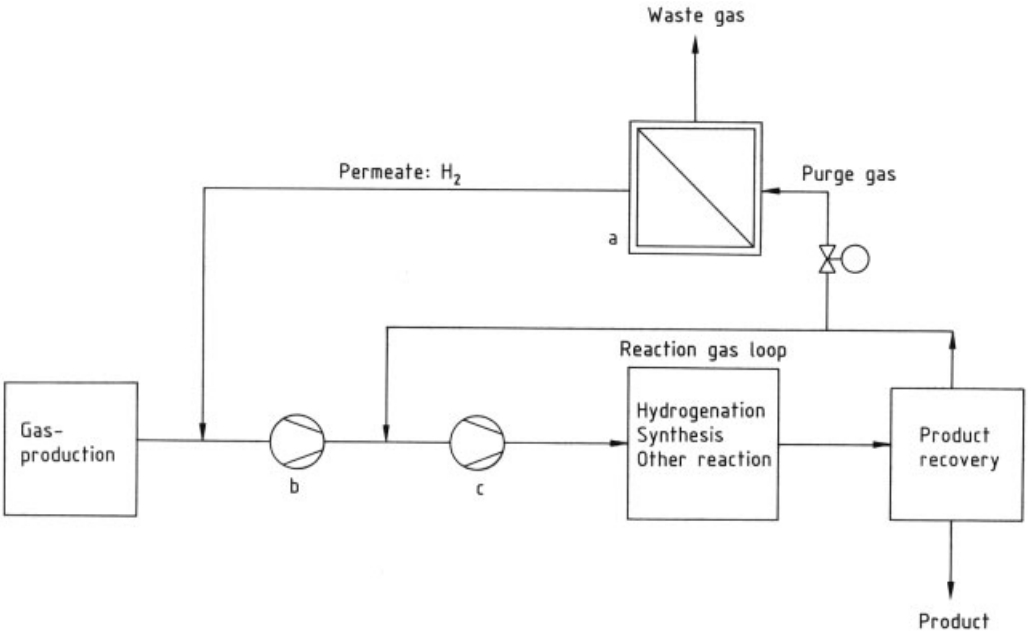

Figure 72. Principle of H_2 recovery from reaction gas mixtures by means of gas separation membranes
a) Gas separation membrane; b) Raw gas compressor; c) Recycle gas compressor

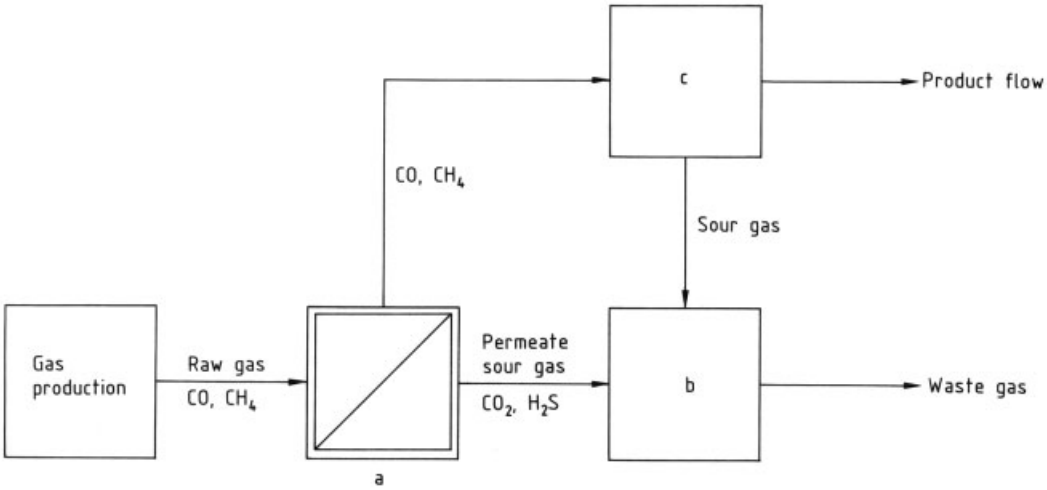

Figure 73. Pretreatment by means of gas separation membrane
a) Gas separation membrane (pretreated stage); b) Sour gas treatment; c) Absorption (fine treatment)

6. Handling of Byproducts

Cooling the crude product gas leads to the formation of aqueous and hydrocarbon condensates. Usually, water-insoluble hydrocarbons are removed from the aqueous phase and handled separately. Treatment of the condensates and the crude gas also results in gaseous byproducts.

6.1. Aqueous Condensates

Rates and Components. All production methods for gases containing carbon monoxide and hydrogen primarily generate a hot crude gas mixed with water vapor, regardless of feedstock. The water content of the crude gas originates

from the feedstock moisture (coal), unconverted gasification steam, or reaction water (oxidation of part of the feedstock hydrogen). Cooling the crude gas produces aqueous condensates laden to a greater or lesser extent with organic or inorganic substances, depending on the type of feedstock and the gasification method. Table 34 shows typical condensate rates and contents of dissolved or suspended components for some feedstocks and selected gasification methods. The numbers may change depending on gasification conditions. As a rule, condensate rate and impurity content increase with decreasing gasification temperature and increasing feedstock complexity [227 – 231]. In addition to the classical feedstocks nowadays a variety of waste materials are used. These have to be prepared, e.g., by pelletizing using suitable binders such as bituminous coal and/or by mixing with coal in certain ratios between 50 % and 85 %. The waste includes solid and liquid materials like plastics, sewage sludge, rubber, fluff, contaminated wood, residues of paint, household wastes, industrial waste, etc., as well as tar/oil from fixed bed gasifiers, used oils, solvents etc. [232]. This changes the composition of the condensates slightly. The modern practice is to select gasification types and conditions in such a way that the condensates obtained are as clean as possible and/or show low specific rates, in order to reduce the wastewater treatment cost (see Chap. 4).

Typical Specifications of Treated Wastewater. Condensates are recycled as far as possible within the gas production plant. Only excess condensate must be specially treated. Treatment methods depend on the final destination of condensates:

1) Use as boiler feedwater (liquid disposal),
2) Use as cooling water make-up (gaseous disposal),
3) Burning of contents and evaporation of water (gaseous disposal), or
4) Release to receiving water (liquid disposal).

The first two are possible only if the condensates contain easily removable components (e.g., using steam reforming condensate as boiler feedwater after degassing) or are cleaned by various treatment processes (e.g., using coal pressure gasification condensate as cooling water after tar – dust separation, filtration, extraction, stripping, adsorption, biological oxidation, etc.) [233, 234]. Burning is limited by the amount of water to be evaporated, the environmental impact of sulfur and nitrogen oxides, and economic aspects (proportion of recoverable heat). In practice, incineration of gas condensates has been considered only for small rates [235] or for the concentrated residues after treatment [236].

The last method is the most common: gas condensates are treated to meet wastewater specifications (local authority regulations) so that the effluent can be discharged into receiving water (river, canal, lake, sea).

No universal wastewater disposal standards exist. Specifications must be negotiated with the local water control authority, which usually considers the plant location, type of feedstock, conversion process, and type of receiving water [237, 238]. Typical wastewater specifications for authorized effluent disposal are listed in the following material [BOD_5 = biological oxygen demand (five days)], indicating the targets for various wastewater treatment processes described in the following sections.

Temperature	max. 30 °C
pH	6.5 – 8.5
Suspended solids	max. 30 mg/L
BOD_5	max. 30 mg/L
COD	max. 500 mg/L
Hydrogen sulfide	max. 0.1 mg/L
Ammonia (measured as nitrogen)	max. 200 mg/L
Free cyanide	max. 0.5 mg/L
CNS^-	max. 15 mg/L
Monohydric phenols	max. 0.1 mg/L
Total phenols	max. 10 mg/L
Substances extractable with petroleum ether	max. 20 mg/L

6.1.1. Mechanical Treatment

Gravitational Separation. Suspended solids such as soot, char, or ash and entrained tar or oil are removed mechanically from gas condensates. In coal pressure gasification plants, the first cleaning step is usually carried out in tar separators, often followed by additional oil separators. In oil gasification plants, soot is usually extracted with oil or naphtha and recycled to the process.

Table 34. Specific rates and composition of gas condensates

Process examples	Lurgi coal pressure gasifier	British Gas – Lurgi slagging gasifier	Texaco, Shell,Lurgi MPG	Steam reforming
Type	moving bed	moving bed	partial oxidation	catalytic
Feedstock	bituminous coal	bituminous coal	heavy vacuum residue	naphtha
Specific rate of feed, m^3/t	0.7 – 1.2	0.15 – 0.2	0.4	0.5 – 0.7
COD, mg/L	25 000		2400	
BOD_5, mg/L	8000		1000	
Composition, mg/L				
Suspended solids	100		120	
Suspended oil	1000		20	
Free ammonia	15 000	20 000	1500	
Fixed ammonia	350	3 000	150	
CO_2	30 000	40 000	2400	400
H_2S	700	600	1600	
HCN	50	100	300	
Monohydric phenols	4500	4000	<1	
Total phenols	6000	5000	<1	
Dissolved organic substances	500			
Halides (Cl^-, F^-)	250	4000		
Fatty acids	800	500	400	
Nickel	0		2 – 5	
Iron	4		6 – 15	
Vanadium	0		3 – 8	
pH	8 – 9	8 – 9	8 – 9	5 – 7

Primary tar separators may be simple tanks provided with a feed nozzle, an expansion gas exit to release dissolved gases, an oil skimmer, and drawoffs for tar, solids, and cleaned water. Oil has a lower density, and tar – dust a higher density, than water so they form top and bottom layers that are withdrawn while water is obtained in the intermediate layer. The settling area of tar and oil separators is usually designed to remove all particles having a settling velocity higher than 1 – 3 m/h from the aqueous phase. Remaining particles are removed in the following filtration step.

Part of the mixture of tar and dust (heavy tar) is recycled to the gasifiers by positive displacement pumps. Excess tar sometimes undergoes further dust removal and is sold as a product. After primary tar separation, the gas condensate still contains up to 100 mg/L of solids and 500 – 5000 mg/L of tar and oil [239].

The gas condensate is cleaned further in *oil separators* down to a content of oil and solids in the range of 20 – 50 mg/L. The efficiency of these separators can be enhanced by fitting a large number of parallel corrugated plates, which not only increases the settling area but also achieves laminar flow conditions, resulting in low Reynolds numbers. An example of this separator type is the Shell CPI [240].

Filtration. The following treatment steps (e.g., extraction, stripping, adsorption) require that tar and solids content be reduced to 10 mg/L or lower. This can be done in simple gravel filters operated at low specific loads of ca. 5 – 7 $m^3\,m^{-2}\,h^{-1}$. Modern dual- or multimedia filters can operate up to 20 $m^3\,m^{-2}\,h^{-1}$. In dual-media filters, the gas condensate flows first through an upper layer of lower density, coarse filter media, which allows finer solids to penetrate deeply into the filter bed, and then through a lower layer of finer filter media to ensure final cleaning.

Filters operate batchwise. Filter beds loaded with tar and solids are regenerated by backflushing with (hot) water. Usually, the bed is also loosened by blowing gas through it. After cleaning, filter bed particles are classified according to Stokes' law by means of an upward stream of water at a defined rate. The filters not only remove solids to < 10 mg/L but, at low temperatures and low velocities, remove tar and oil as well. These partly stick to the filter media and partly collect in the bottom of the filter, from which they can be removed prior to backflushing. The backflushing liquid (mud liquor) is normally recycled to the primary tar separator.

Phenol	45
Cresols	30
Xylenols	6
Polyhydric and other phenols	10
Pyridine bases	1
Neutral oil, pitch	5
Water	3

6.1.2. Extraction and Adsorption of Organic Substances

Filtered gas condensates containing high levels of dissolved organic substances cannot directly be treated biologically. First, the phenol content must be reduced to an acceptable level, usually < 700 mg/L. At low feed rate or concentration, this is done by dilution. At high concentration of steam volatile phenols, the condensate is dephenolized prior to stripping the dissolved gases. This is carried out either by liquid – liquid extraction using organic solvents — e.g., benzene, methyl isobutyl ketone (MIBK), butyl acetate, or diisopropyl ether (DIPE) — or by adsorption on activated carbon. Combinations of these processes are also possible [241 – 243]. Adsorption on activated carbon is mainly used for posttreatment after dephenolization and/or biological treatment of condensates in gas production plants [244 – 246].

Phenol Recovery by Solvent Extraction. Figure 74 shows a typical phenol extraction process used in gasification plants, as well as in coke oven and carbonization plants, in the phenol industry, and in coal hydrogenation plants. In the Phenosolvan process the filtered and cooled phenolic effluent is treated countercurrently with a suitable solvent such as DIPE in a multistage extractor. The extract is separated by fractional distillation into pure solvent (overhead product) and crude phenols (bottoms product). The solvent is recycled to the extractor. The raffinate still contains a small amount of solvent which is recovered by stripping with recycled gas. The solvent is then removed from the gas by absorption in cooled and recycled crude phenols, from which it is subsequently recovered in the stripping section of the fractionator. The Phenosolvan process can be operated with different feedstocks, different solvents or solvent mixtures (e.g., benzene and DIPE or DIPE plus MIBK) [247], and different solvent recovery systems (e.g., steam, gas, or ammonia stripping). All steam volatile phenols and neutral oils can be recovered almost completely, but only partial recovery of pyridines, fatty acids, and dihydric phenols is possible. The suitability of this process for hard-to-extract organics is limited only by economics, because residual phenols and other organics can be removed more cheaply by adsorption or biological treatment. The intermediate phenol content between extraction and final treatment is best determined by means of an economic optimization study, an example of which is shown in Fig-

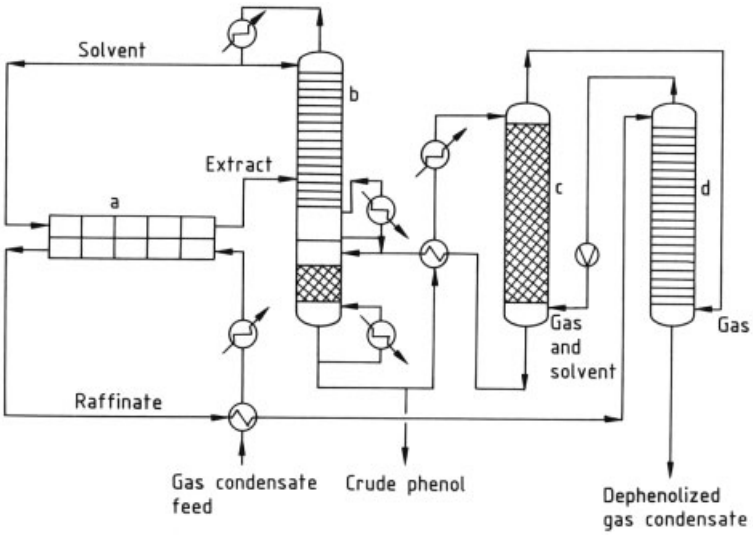

Figure 74. Lurgi Phenosolvan process for phenol recovery (with steam-saving solvent recovery by gas recycle)
a) Extractor (mixer – settler); b) Fractionator; c) Solvent absorber; d) Solvent stripper

ure 75. An advantage of solvent extraction is the recovery of crude phenol by fractionating the extract which contains valuable byproducts. Typical composition of a Phenosolvan extract from coal pressure gasification is given below, in percent by weight:

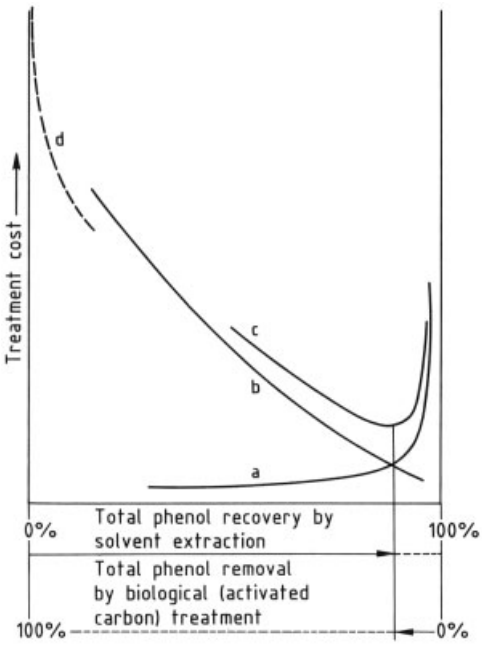

Figure 75. Economical phenol recovery point between phenol extraction and final treatment (typical, e.g., for coal gasification condensates containing ca. 4500 mg/L monohydric and ca. 1500 mg/L polyhydric phenols)
a) Extraction; b) Biological treatment; c) Total cost; d) Biological and activated carbon treatment

If large gas condensate flows must be dephenolized, extraction must be performed in mixer settlers, exhibiting high stage efficiencies. Lower flow rates can be treated in extraction columns by using many stages to compensate for low efficiency.

Recovery by High-Pressure Extraction. Organics can also be recovered by using supercritical solvents [248]. This method might be economical for small condensate rates with high feed concentrations of phenols and other valuable organics, but it is not yet being used in gas production plants [249]

6.1.3. Removal and Recovery of Ammonia and Sulfur

Before gas condensates rich in dissolved gases (NH_3, CO_2, HCN, H_2S) can be treated in biological oxidation or in heavy-metal precipitation plants, they must be stripped with gas or steam to decrease free ammonia to < 50 mg/L, hydrogen sulfide to < 1 mg/L or hydrogen cyanide to $<$ 10 mg/L, or to lower the pH (previously 8 – 9) to 6. This is done in the *total stripper* common to all removal or recovery processes. Its primary purpose is to produce stripped gas condensate as bottom product. The overheads contain all the free gases and water vapor. Processes differ in the method of treating this gas mixture.

Total Stripper. Depending on the major feed contaminant, this stripper has different names such as *ammonia still* and *sour water, hydrogen sulfide*, or *hydrogen cyanide stripper* [250]. It usually operates at normal atmospheric pressure, but the pressure may be adjusted depending on the destination of the top gases. Typical steam consumption is 140 – 250 kg of steam per cubic meter of gas condensate.

The stripper is usually equipped with a top condenser to reduce the water content of the gases from > 90 to $< 50\,\%$.

Incineration, Catalytic Oxidation and Sulfur Recovery. The mixture of ammonia, sour gases, and water can be flared off only if the environmental impact (e.g., of SO_2, NO_x) is acceptable. Otherwise, a special incineration plant with flue gas treatment is required. Ammonia can also be removed by catalytic oxidation [251].

If a Claus plant (→ Sulfur, Chap. 6) is available, the strip gas can be mixed with the feed for sulfur recovery and ammonia oxidation. If the ammonia flow is too high compared to the Claus sulfur stream, the strip gas must be divided into an ammonia-free sour gas and an ammonia stream low in hydrogen sulfide [252]. Both streams are fed separately to the Claus plant; first the ammonia stream is oxidized above stoichiometric conditions and then mixed with the sour gas stream at below stoichiometric conditions to produce sulfur. Any NO_x formed in the first step reacts with sulfur produced in the second step; thus, both sulfur dioxide and NO_x emissions are eliminated [253].

Recovery of Ammonia. If the ammonia quantity exceeds a certain level, incineration is unacceptable but recovery becomes economical compared to (catalytic) oxidation. Formerly, the strip gases were washed with sulfuric acid to obtain ammonium sulfate, and the remaining acid gases were fed to sulfur recovery plants. Because ammonium sulfate sales today are limited, the recovery of anhydrous or aqueous ammonia is preferred in most cases. Generally, two different methods are available for the difficult separation of basic and acid gases from aqueous solution:

1) Chemical methods using agents that selectively absorb ammonia under certain conditions, but release ammonia under regeneration conditions [254]; and
2) Physical methods (without agents), based on different vapor pressures of ammonia and acid gases at different concentrations in aqueous solution and at different temperatures or pressures [242, 255, 256][257].

The first method is represented by the Phosam-W process (U.S. Steel, USX Engineers & Consultants, Aristech Phosam-W), and the second by the CLL ammonia recovery process (Chemie Linz/Lurgi) or by the WWT process (Chevron); both types of processes are used in coal pressure gasification plants [258 – 261]. The largest plant worldwide for the recovery of clean liquid ammonia from coal pressure gasification wastewater operates in Secunda at Sasol II and III since 1980. Both CLL trains together have a total capacity of 1000 t NH_3 per day formed as a byproduct.

Phosam-W Process (see Fig. 76). The total stripper (a) overheads are treated in the ammonia absorber (b) with lean ammonium phosphate solution, which selectively absorbs ammonia under prescribed conditions. The wet overhead acid gases are dried (c) and sent to a sulfur recovery unit. The rich ammonium phosphate solution is regenerated in the Phosam stripper (d) at high temperature and recycled to the absorber. The stripper overheads (mainly ammonia and water) are condensed, treated with sodium hydroxide to bind coabsorbed acid gases, and fractionated to yield pure anhydrous ammonia as final product. Hot (475 K) fractionator bottom water and sodium hydroxide are flashed into the total stripper.

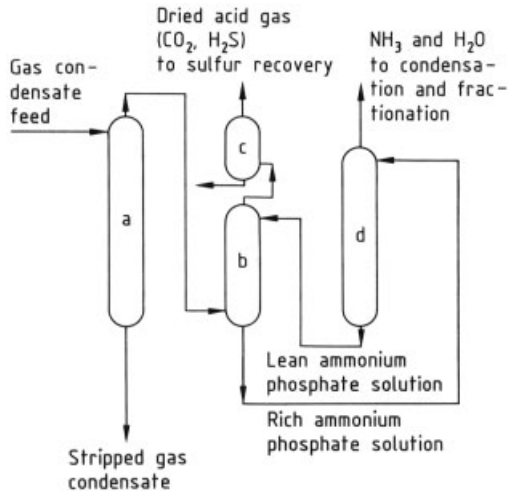

Figure 76. Phosam-W process for chemical separation of aqueous ammonia – acid gas mixture
a) Total stripper; b) Phosam ammonia absorber; c) Acid gas drier; d) Phosam ammonia stripper

Clean Liquid Ammonia Recovery Process (*CLL*; see Fig. 77). The total stripper (a) overheads are treated in the acid gas absorber (b) and ammonia stripper (c) column. Ammonia is cleaned by fractional condensation and fed to a liquefaction unit, to yield anhydrous ammonia as final product. Alternatively, ammonia can be absorbed in water or fed into the ammonia pipeline of the factory (Chemie Linz). The acid gas absorber bottoms are concentrated in the ammonia stripper (c) and fed to the deacidifier (d) where clean carbon dioxide and hydrogen sulfide are obtained as top product, suitable for use directly as sulfur recovery feed gas.

Deacidification is enhanced by dilution with water or elevation of temperature and pressure. Deacidified gas condensate is fed to the total stripper, closing the CLL loop.

The gas condensate is fed either to the deacidifier if it contains more acid gases and/or inert gases and/or traces of solvent from the dephenolization step or to the total stripper if it contains more ammonia. Diluted condensates are also fed to the total stripper.

Both processes described yield ammonia qualities comparable to synthetic ammonia (Table 35); hence, process selection depends on economic considerations.

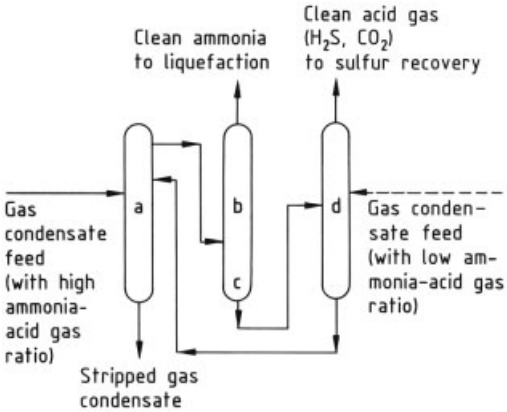

Figure 77. CLL Ammonia recovery process for physical separation of aqueous ammonia – acid gas mixtures
a) Total stripper; b) CLL ammonia stripper; c) CLL acid gas absorber; d) Deacidifier

Table 35. Typical composition of anhydrous ammonia

	Clean liquid ammonia recovered from gas condensates	Synthetic ammonia
Water (added), wt %	0.01 – 0.2	0.2
Ammonia, wt %	99.99 – 99.8	99.8
Oil, ppm	<2	<5
Organic substances, ppm	<8	not determined
H_2S, ppm	<1	not determined
CO_2, ppm	<50	not determined

Economic Considerations. *Chemical absorption – regeneration plants* usually have low capital costs but high operating costs because of the use of stronger agents and the need for high-pressure steam to regenerate the chemicals. Chemical absorption plants are less flexible with respect to overload, since the agents have a distinct limit of loading capacity. *Physical absorber – stripper plants* have higher capital costs but lower operating costs because of the use of cheaper steam. The higher the gas condensate flow rate or the ammonia feed concentration, the more economical are physical methods; chemical methods are more economical in the case of low flow rates or low ammonia feed concentrations.

6.1.4. Biological and Final Treatment

After extraction and stripping, gas condensates still contain organic and inorganic substances such as fatty acids, dihydric phenols, and anions (e.g., chlorides and fluorides). Usually, the COD and BOD of pretreated gas condensates are decreased biologically. Aerobic microorganisms break down and metabolize most of the organics to carbon dioxide and produce sludge.

Essentially, two methods exist for *aerobic treatment* of gas condensates. The *activated-sludge process* is most common because it is less sensitive to variations in feed BOD. The *trickle bed process*, in which a film of microorganisms oxidizes the organics, is seldom used for gas condensates [262 – 264]. Sometimes PAC (powdered activated carbon) is used, mainly in activated sludge processes [265 – 269].

If the biotreated gas condensate is to be (partly) reused (e.g., as boiler feedwater), final filtration for removal of suspended solids and, if necessary, ion exchange (→ Ion Exchangers), reverse osmosis, or multiple flash evaporation for removal of strong anions may be employed as well.

Activated Sludge Process. For aftertreatment of gas condensates, a single stage consisting of an activated sludge basin equipped with aerators, to provide the required oxygen and effect mixing of the basin contents, and a clarifier for separation and thickening of the activated sludge is usually adequate. Nutrients (phosphorus and nitrogen compounds) for the microorganisms are needed in the ratio of BOD: N: P = 100: 5: 1. Gas condensates normally contain the required nitrogen in the form of fixed ammonia; phosphorus must usually be added in the form of phosphoric acid.

Typical design figures for biological aftertreatment of gas condensates to achieve a residual BOD_5 of 20 ppm are listed in the following material (MLSS = mixed liquor – suspended solids, water-free; BOD_5 = biological oxygen demand (five days); DO = dissolved oxygen).

Activated sludge basin:

Specific volumetric load, kg COD m^{-3} d^{-1}	ca. 1.00
Specific sludge load, kg BOD_5 kg^{-1} MLSS d^{-1}	ca. 0.25
Specific sludge content, kg MLSS/m^3	ca. 4.00
Oxygen content, ppm DO	1 – 2
Oxygen demand, kg O_2/kg BOD_5	ca. 1.50
Specific oxygen transfer, kg O_2 kW^{-1} h^{-1}	ca. 2.50

Postclarification basin:	
Retention time, h	ca. 2.50
Rise rate (water), $m^3\ m^{-2}\ h^{-1}$	ca. 0.80
Specific area load, kg MLSS $m^{-2}\ h^{-1}$	4 – 5

Addition of PAC to the activated sludge basin improves removal efficiency in some cases, because of various favorable properties of PAC: (1) substances toxic or inhibitory to biodegradation are adsorbed; (2) settling, thickening, and dewatering characteristics of the sludge are improved; and (3) the sludge age can be increased, resulting in higher retention times of difficultly biodegradable organics. Typical PAC consumption is 0.1 – 0.3 kg per cubic meter of gas condensate.

6.1.5. Removal of Heavy Metals

Wastewater from different processes (refining, coking, metallurgy, fossil fuel conversion, power generation, and others) contains greater or lesser amounts of heavy metals. The type and content depend mainly on feedstock composition [270, 271] and the conditions under which products come in contact with water (process condensates, quench water, cleaning liquid, etc.). Some of the wastewater must be treated for removal or recovery of heavy metals. Typical methods are liquid – liquid ion extraction, fixed-bed ion exchange, and precipitation by adding suitable chemicals such as metal salts, lye, or acid [272 – 274].

In the majority of all cases, gas condensates contain only small quantities of heavy metals. However, a considerable nickel and vanadium content is found, for instance, in gas condensates from partial oxidation of heavy fuel or vacuum residues. These metals are environmentally harmful. The new Lurgi Multi Purpose Gasification (MPG) process provides a special treatment of the heavy metals (MARS, Metals Ash Recovery System, see Sections 3.2.2 and 7.2). This reduces the metals concentration in the effluents considerably.

Precipitation of Vanadium and Nickel. After removal of hydrogen cyanide, carbon dioxide, hydrogen sulfide, and ammonia, the stripped condensate typically has a pH of ca. 6. It is cooled to 40 °C and fed to the heavy-metal removal plant, which consists mainly of precipitation steps, sludge dewatering and filter cake storage.

Precipitation of vanadium and nickel can be carried out in *one step* at pH 9 if all metals are to be removed in one mixture [273]. This mixture is clarified in a Sedimat (an automatic clarifier) [275], equipped with a central flocculation chamber to which a coagulant (polyelectrolyte) is added. The flocculated sludge settles to the thickening section in the bottom of the Sedimat, from which it is discharged.

The metal-free effluent is filtered to remove suspended solids, neutralized with sulfuric acid, and discharged to receiving water.

Thickened sludge streams from the clarifiers are filtered in order to reduce the water content to 70 %.

6.1.6. Example of an Industrial Application

Gas condensate from a typical Lurgi moving-bed coal pressure gasification with the following approximate composition, in grams per liter:

Tar (suspended portion only)	11.0
Oil (suspended portion only)	3.5
Suspended solids	1.5
Steam volatile phenols	4.6
Dihydric phenols	0.9
Fatty acids as acetic acid	0.7
Soluble organic substances	0.4
Free ammonia	15.6
Fixed ammonia	0.7
Carbon dioxide	32.0
Hydrogen sulfide	0.4
Free cyanide	0.1
Water	balance

and a flow rate of 500 m^3/h is to be treated so that 20 % can be released to receiving water, 75 % used as cooling tower make-up water, and 5 % used as boiler feedwater. A typical process sequence is shown in Figure 78. Intermediate battery limit conditions should be selected in such a way that total treatment cost is minimized whereas final conditions are still satisfied [276]. This leads to the following typical design criteria, in milligrams per liter, for intermediate product streams (after treatment):

Tar separation	
Total suspended matter	1000
CPI oil separators	
Total suspended matter	350

Dual media filters (for 500 m^3/h)	
Suspended matter	10
Phenosolvan plant	
Steam volatile phenols	10
Dihydric phenols	280
Ammonia recovery	
Free ammonia	50
Hydrogen sulfide	1
Activated sludge treatment	
BOD_5 reduction	from 700 to 20
Final dual media filtration	
Activated sludge content	20
Activated carbon posttreatment (using GAC, Granular Activated Carbon; [277])	
(for 400 m^3/h) Biologically refractory substance	5
Ion exchanger (for 25 m^3/h)	
Strong ions	10

In addition, the aqueous condensate treatment units will produce the following valuable byproducts, in tons per hour:

Tar	5.4
Oil	1.7
Crude phenol	2.7
Sulfur (via Claus plant)	0.1
Clean liquid ammonia	7.9

6.2. Hydrocarbon Condensates

The yield and composition of hydrocarbon condensates depend very strongly on the type of coal and the gasification conditions. In a countercurrent gasification system (e.g., Lurgi pressure gasification), coal travels through a retorting zone where volatile components are vaporized and leave the gasifier along with the gas. When the product gas is cooled in subsequent heat exchangers and coolers, different condensate fractions are recovered. When the temperature drops below the dew point of water, water condenses along with the hydrocarbons and forms an aqueous phase in the final separators (see Section 6.1).

Two fractions are recovered from Lurgi pressure gasification: a *tar fraction* containing the high boilers and an *oil fraction* boiling in the medium range. A third *light naphtha fraction* is obtained when the gas stream is cooled further for low-temperature purification in a Rectisol plant.

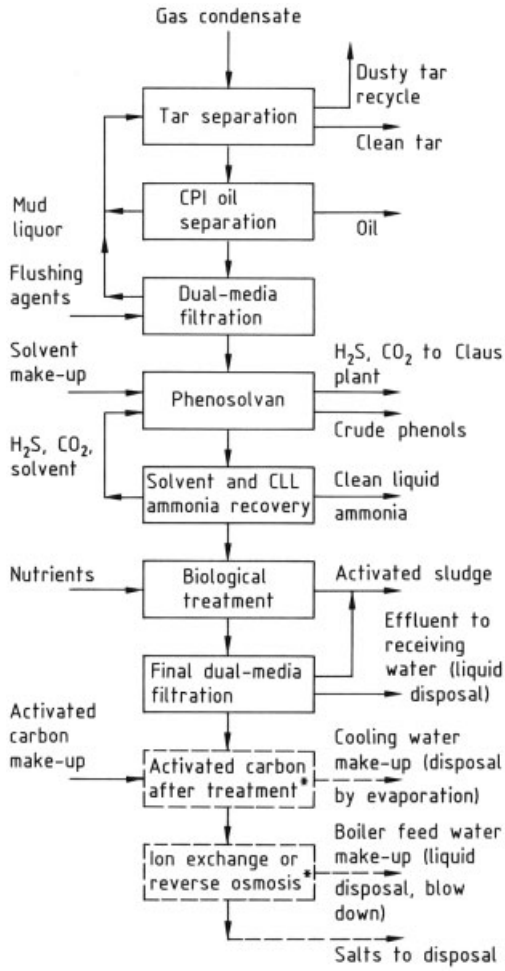

Figure 78. Treatment of aqueous condensates from moving bed coal gasifier (example given in the text)
* These processes are optional

Whereas the naphtha fraction can be treated directly in downstream units, the tar and oil fractions must be pretreated by conventional methods to remove suspended solid material (ash) and undissolved water. Typical properties and compositions of tar and oil are shown in Table 36.

To obtain products with suitable properties for use as fuel oil and gasoline, the combined pretreated medium- and high-boiling condensates are separated into their respective fractions by distillation [278]. Distillate boiling in the gasoline range is then refined along with the light naphtha fraction by hydrogenation to remove unsaturated compounds (which are potential gum formers) and to eliminate sulfur and ni-

trogen. Recovery of chemically pure compounds such as naphthalene or cresols (isomeric mixture) is usually uneconomic because of the small amounts present in the condensates [279].

Table 36. Typical composition of condensates from Lurgi pressure gasification

	Tar	Oil
Rate, wt % of total	60	40
Total water, wt %	10 – 15	15 – 20
Suspended solids, wt %	15 – 25	
Properties after removal of solids and undissolved water		
Specific gravity, d_{30}	1.05 – 1.1	0.94 – 0.96
Solidification temperature, °C	+ 15	< − 20
Flash point, °C	+ 60	+ 19
Water, wt %	4	2
Nitrogen, wt %	1	1
Sulfur, wt %	0.4	0.4
C_6– C_9 aromatics, wt %	6	16
Naphthalene, wt %	3	5
Methylnaphthalene, wt %	3	4
Phenolic compounds as cresol, wt %	21	19
Initial boiling point, °C	98	96
Portion distilled at 360 °C, wt %	60	75
Distillation residue, wt %	40	25
Properties of distillation residue		
Softening point, °C	67	63
Portion insoluble in toluene, wt %	10	7
Portion insoluble in quinoline, wt %	2	0.1
Coke (Conradson coking test), wt %	29	28

6.2.1. Distillation

In a first dewatering column, the water content is lowered by distillation. The dehydrated hydrocarbon condensate is then distilled in an atmospheric column. A typical distillation unit yields six fractions (see Table 37). Light and heavy naphtha are obtained as top and first side fractions; they are combined with the Rectisol naphtha and refined in a hydrogenation unit (see Table 38).

Two more side fractions, heavy and medium creosote, are recovered from the atmospheric column. The bottoms fraction is flashed into a drum under vacuum where it separates into vaporous residue oil and pitch.

Fractions boiling above 210 °C are used (e.g., as fuel or impregnation oil for wood). Pitch is suitable as a binder for briquetting of fine coal or may be fed to a partial oxidation plant to produce synthesis gas.

Table 37. Typical fractions from a coal tar distillation unit

	Approximate total, wt %	Approximate boiling range, °C
Light naphtha	4	≤170
Heavy naphtha	7	160 – 210
Medium creosote	25	190 – 290
Heavy creosote	21	260 – 400
Residual oil	13	max. 5 vol % ≤300
Pitch *	30	max. 15 vol % ≤350

* Softening point of pitch is ca. 70 °C; nonvolatile residue of pitch, max. 30 wt %.

Table 38. Typical feedstocks for hydrorefining (approximate figures)

	Heavy naphtha	Light naphtha from tar distillation	Light naphtha from Rectisol unit	Total blend
Rate, wt % of total	15	25	65	100
Specific gravity, d_{20}	0.93	0.84	0.83	0.85
Bromine consumption, g/100 mL	114	33	37	70
Total sulfur, wt %	0.35	0.49	0.42	0.4
Mercaptans, mg/kg	20	70	80	68
Phenolic oil, vol %	25			10
Alkaline nitrogen, mg/kg	10 100	7600	270	5000
Naphthalene, wt %	10			1.5
Initial boiling point, °C	130 – 140	90 – 95	50 – 55	55
Boiling end point at 95 wt %, °C	200 – 210	150 – 160	140 – 145	220
Chemical hydrogen consumption, m^3 (STP)/t	360	135	60	120

6.2.2. Hydrorefining

Hydrorefining is used (1) to saturate olefins and diolefins to their respective alkanes; (2) to convert sulfur in mercaptans and thiophenes to hydrogen sulfide; (3) to convert nitrogen in pyridines and amines to ammonia; (4) to convert organic chlorine to hydrogen chloride; and (5) to convert oxygen in phenols to water. The feedstock is a blend of the various naphtha streams recovered from condensates (see Table 38).

The *Benzoraffin process*, developed around 1960 jointly by BASF, Veba-Chemie, and Lurgi [280, 281], is commercially successful in refining the pretreated naphtha fractions derived from coal gasification. It is a fixed-bed, catalytic, gas-phase hydrogenation process using a cobalt – molybdenum catalyst operating at ca. 5 MPa and 350 – 420 °C (Fig. 79).

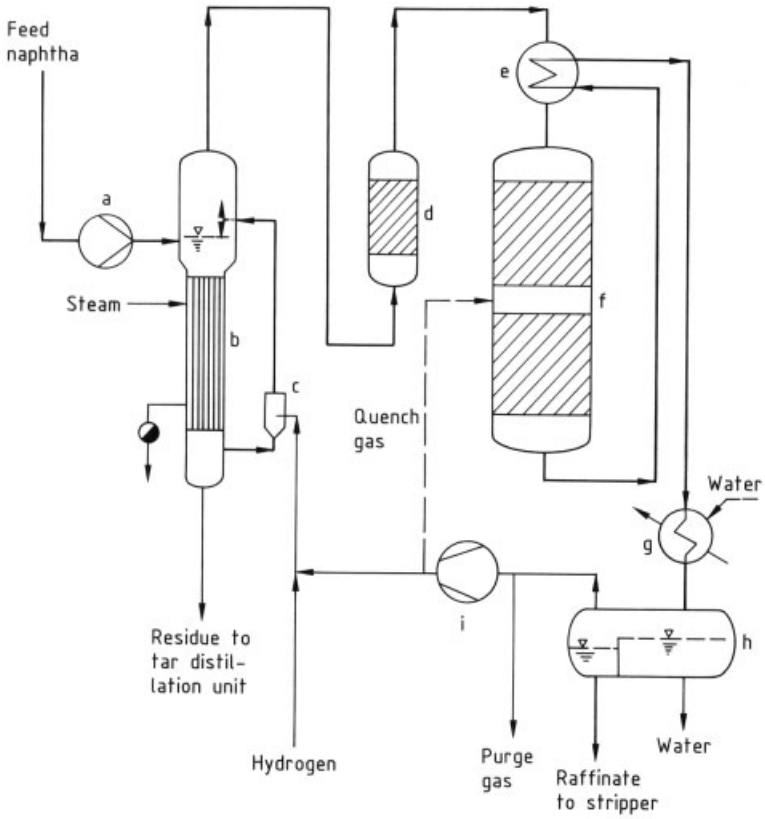

Figure 79. Benzoraffin process for hydrorefining of naphtha fractions
a) Feed pump; b) Vaporizer; c) Mixer; d) Prereactor; e) Heat exchanger; f) Reactor; g) Cooler; h) Separator; i) Recycle compressor

Modern alternatives to this process (e.g., by BASF, Lurgi-Südchemie, IFP et al.) — also used for the preparation of feedstocks for extractive distillation of toluene (→ Toluene, Chap. 3.1) — also work with two reactors, but the first reactor contains a palladium catalyst and operates in the liquid phase at low temperatures, preventing gum formation from diolefins or styrene.

The feed naphtha contains highly reactive chemical compounds with a strong tendency to form resinous deposits in heaters and vaporizers. A special vaporizer developed by Lurgi avoids fouling and frequent shutdowns for cleaning.

Process Description. Feed naphtha is vaporized under pressure and mixed with hydrogen-rich gas in a specially designed vaporizer (b) [281]. A small stream is withdrawn as residue from the bottom and returned to the tar distillation unit. The mixture of gas and naphtha vapor passes through a prereactor (d) filled with cobalt – molybdenum catalyst. Here, the most reactive unsaturated compounds are hydrogenated to prevent fouling in the heat exchanger (e). Reaction occurs at 200 – 250 °C. The outlet temperature is ca. 5 – 10 °C higher due to the exothermic reaction. After further heating in the heat exchanger (e) to about 350 °C, the prereactor effluent enters the main reactor (f). This reactor has two catalyst beds and a facility for injecting recycle gas to adjust the reaction temperature. The first bed contains cobalt – molybdenum catalyst pellets; the second bed is filled with a molybdenum catalyst. The exothermic reaction leads to an effluent temperature > 400 °C. The gas – vapor mixture is cooled to 30 – 40 °C in a heat exchanger (e) and cooler (g). Water should be injected into the cooler to prevent salt deposits, because hydrogen chloride and ammonia formed

during the reaction tend to solidify at lower temperature in the form of ammonium chloride.

The condensed raffinate is separated from the gas in a separator (h). After depressurization to ca. 1 MPa, the raffinate is stripped in a stripper column to remove hydrogen sulfide, ammonia, carbon dioxide, and water. If it still contains some phenolic compounds, a caustic wash and a subsequent water wash are recommended. A portion of the separated gas is withdrawn to remove sufficient light byproducts and inerts for the partial pressure of hydrogen to be adjusted properly. The main stream is compressed and recycled to the vaporizer after addition of fresh hydrogen.

A typical specification for a hydroraffinate from the naphtha blend in Table 38 is shown below.

Appearance	clear and free of suspended solids
Color, Saybolt	0
Water	max. 0.1 wt %
Total sulfur	max. 0.1 wt %
Existent gum	max. 4 mg per 100 mL
Aromatics	ca. 75 wt %

Because of its high aromatic content, this raffinate is a valuable blending component for gasoline. Alternatively, pure benzene, toluene, and other aromatics may be recovered from the raffinate by appropriate separation processes. Solvents with specified boiling ranges for various applications in the chemical industry can be obtained by fractionation.

6.3. Gaseous Byproducts

Sulfur Compounds (→ Hydrogen Sulfide, see → Sulfur). Hydrogen sulfide constitutes the main part of the gaseous sulfur components. It usually accounts for 90 – 95 % of the sum of all sulfur compounds. The remainder is primarily COS; CS_2, SO_2 and mercaptans are usually present in trace amounts only.

These sulfur components occur in various waste gases and are more or less strongly diluted with CO_2. Hence, processing depends totally on their concentration and usually aims at recovering elemental sulfur. In special cases, the direct production of sulfuric acid can also be of interest and even low concentrated H_2S gases can be used for this purpose.

There are various methods available for the recovery of elemental sulfur from low concentrated waste gases (e.g., 1 % H_2S, 98 % CO_2 and 1 % inert gases). These methods include concentration by selective absorption or by adsorption, and sometimes intermediate steps with (partial) combustion to SO_2 are also involved. Processes of this type have been occasionally employed.

From a theoretical standpoint, the most economical method should be direct oxidation, either in the gaseous or in the liquid phase as performed in an oxidation wash, e.g., the Lo-Cat process (see Section 5.4.3). However, it should be taken into consideration that organic sulfur compounds in the gas, primarily COS and CS_2, remain practically unchanged and this could represent an undue environmental burden. On the other hand, precautionary measures must be taken if the waste gas contains trace components, such as HCN, that could cause solvent degradation. Therefore, low investment cost are often counterbalanced by high operating cost for solvent reclaiming and/or make-up of chemicals. Today oxidation wash systems are mainly used for low-pressure off-gases with low sulfur contents, with limitations in capacity.

If hydrogen sulfide is present in concentrated form, i.e., in concentrations of more than 20 % (in some cases 10 % are sufficient) sulfur can be recovered more economically by using the Claus process. However, the Claus tail gas almost always requires an additional treating step; it usually still contains 3 to 6 % of the sulfur, mainly in the form of H_2S and SO_2. A modern solution is the integration of a three stage Doxosulfreen process (→ Sulfur, Chap. 7.2.1.1) shown schematically in Figure 80. Here, the first two catalytic reactors after the combustion chamber are continuously operated according to the classical Claus principle. After condensing the elemental sulfur formed in the Claus reactors, COS and CS_2 are converted by catalytic hydrolysis at elevated temperature. The following reactors — each containing a Sulfreen stage and a direct oxidation stage — are operated in cycles: in one, the Claus reaction is first, at a slight surplus of H_2S, continued below the sulfur dew point with adsorption of the newly formed sulfur, and then, at further reduced temperature, the residual H_2S is almost completely converted by catalytic oxidation with air. In the second reactor, operated

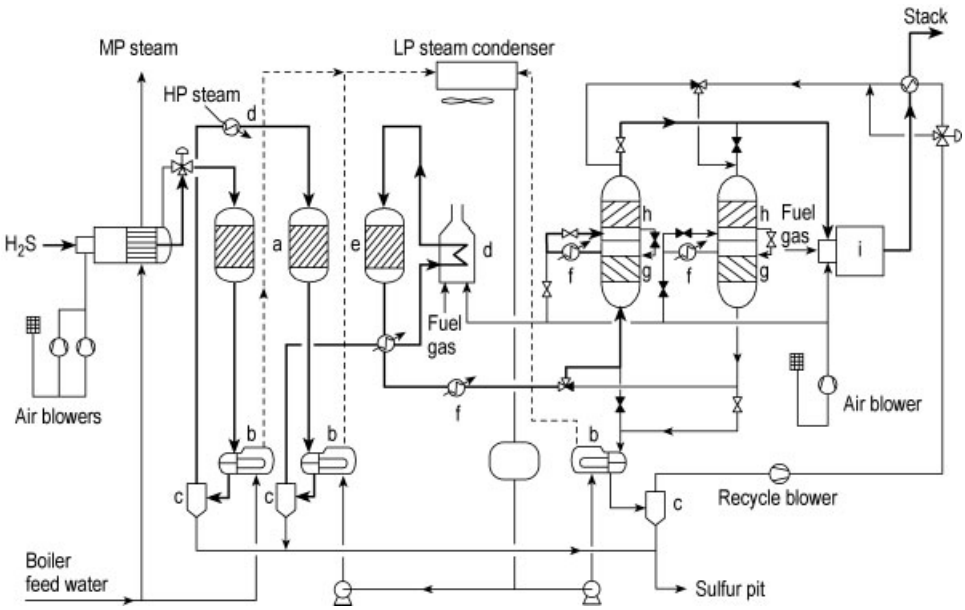

Figure 80. Two-stage Claus plant with internal hot gas bypass, integrated Doxosulfreen plant and thermal incineration
a) Claus reactors; b) Sulfur condensers; c) Separators; d) Heater; e) Hydrolysis reactor; f) Cooler; g) Sulfreen and h) Direct oxidation stages; i) Thermal incineration

in parallel, the sulfur is desorbed by acid recycle gas and is subsequently, as in the first two steps, separated in liquid form in coolers, with the generation of LP steam [282].

Carbon Dioxide. Today, there are basically four different ways of releasing sulfur-free carbon dioxide:

- Release into the atmosphere (there might be restrictions in the future),
- Use for enhanced oil recovery (EOR) or
- As chemical raw material, and
- Recovery as pure CO_2 for the food and pharmaceutical industry.

Usually a large part of the CO_2 waste gas is released into the atmosphere for lack of other uses, although it should be taken into consideration that even sulfur-free CO_2, in larger concentrations, is dangerous (in Germany the MAK value is 5000 ppm). However, in the case of furnaces, a sufficiently high flue-gas temperature normally ensures adequate dilution in the atmosphere; special precautions are required for cold process waste gases and the high carbon dioxide density relative to air. If introduction into a flue gas is not possible (dew point!), cooling towers can be used provided that the danger of corrosion is taken into account, otherwise, only special stacks may be employed.

In connection with the general discussion about the various factors influencing the earth's climate and the necessity to decrease the CO_2 emission into the atmosphere, several solutions have been considered theoretically, e.g., the depositing of CO_2 in absorbed form in deep layers of the oceans. However, none of these suggestions have been realized so far on a technical scale.

One of the more modern applications of CO_2 is in enhanced oil recovery. The specification for the CO_2 to be injected in aging oil wells (to reduce the oil's viscosity) is usually less stringent: CO_2 concentration 98 %, C_1/C_{3+} hydrocarbons up to 0.3 %, total sulfur in the 1 % range. This purity may often be accomplished simply by dehydration of the raw CO_2 [283].

Various requirements have to be met for the use of carbon dioxide as a *chemical raw material*: in the production of carbon monoxide or synthesis gas from natural gas, which has the lowest C/H ratio of all raw materials, carbon dioxide is frequently removed from the process gas by absorption and recycled to the reformer. Care must be taken that the catalyst used is not

damaged by solvent vapors or other trace components. Waste gases from hot potash, amine, and physical washes operated with sulfur-free solvents generally cause no problems, as long as effective separators are employed which are usually followed by a re-compression step. This compression step also contributes to residual separation.

Carbon dioxide used in the *synthesis of urea* has to meet more stringent requirements. Indeed, a maximum concentration of inert gases, usually 1 – 2 % of hydrogen and carbon monoxide, should not be exceeded and sometimes even traces of solvents can cause problems. However, these small amounts can generally be separated by a simple adsorption step with an appropriate molecular sieve.

The highest demands are made on the purity of carbon dioxide used in the *pharmaceutical* or *food industry*, e.g., in the form of dry ice for cooling or packaging purposes or in compressed or liquid form for the production of beverages. Depending on the impurities to be separated (even a few ppm of odorous substances are often sufficient to cause an unpleasant smell), usually several different, usually adsorptive, purification steps are necessary. More modern installations include also distillation stages (to separate "sour CO_2" together with heavier hydrocarbons and light inerts) as well as oxidation stages to convert the reminder of undesirable hydrocarbons (Fig. 81) [283, 284].

7. Typical Examples of Complex Gas Production Plants

7.1. Methanol Production from Natural Gas

The production of methanol (→ Methanol) with natural gas (→ Natural Gas) as feedstock is described by the following process steps: (1) gas production, (2) methanol synthesis, and (3) distillation. Details of the three single process steps are described in other chapters; this chapter deals with specific features of the overall arrangement.

Basically two different gas production technologies, steam reforming and catalytic autothermal reforming, or a combination of both are used to produce methanol from natural gas, depending on plant size and availability of carbon dioxide from outside battery limits.

Synthesis gas produced by *steam reforming* is characterized by a relatively low pressure and a surplus of hydrogen. By addition of carbon dioxide the composition of the synthesis gas can be adjusted to be more favorable for methanol production.

Pure catalytic *autothermal reforming* results in a synthesis gas with a relatively high pressure but with a deficiency in hydrogen for methanol synthesis. A carbon dioxide removal unit is required or additional hydrogen has to be provided to adjust the appropriate gas composition. Such hydrogen may be obtained either from the methanol synthesis purge gas and from additionally produced reformed gas or from a conditioned methanol synthesis purge gas.

The combination of both technologies, the so-called *combined reforming process*, has the advantage of an optimum composition and a high pressure of the produced synthesis gas.

For the methanol synthesis itself a wide range of technologies are available, which differ mainly in the design of the methanol converter. In principle, three different types of converters are used in the industry; the quench-cooled, the gas-cooled, and the water-cooled reactor. For optimization of the reaction and for capacity increase also combinations of these technologies are applied.

In the distillation section the raw methanol produced in the synthesis section is purified to meet the required specification. Depending on plant capacity and optimization with regard to cost or energy savings, a two or three column concept can be used.

At locations where no carbon dioxide is available most of the methanol plants are based on the following gas production technologies, depending on their capacities: steam reforming for capacities up to 2000 t/d, combined reforming from 1800 to 2500 t/d [285, 286].

Erection of plants with a capacity of 5000 or even 10 000 t/d is also discussed nowadays, e.g., for minimizing flaring of associated gas in oil fields by producing methanol as a "liquid storage". For this size of plants pure catalytic autothermal reforming in combination with a high efficient conversion of synthesis gas to methanol is the most economical way. The process is

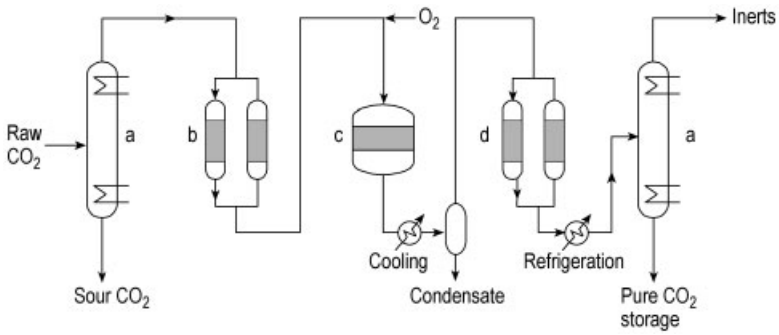

Figure 81. Purification of CO_2 for the production of beverages
a) Fractionation; b) Adsorptive sulfur guard; c) Catalytic oxidation; d) Adsorptive dryers

designed in such a way that conditioned purge gas with the required hydrogen content is obtained to adjust the reformed gas composition. This technology is illustrated in the simplified process flow diagram in Figures 82 and 83 and described in Section 7.1.1.

Steam reforming technology and combined reforming technology are discussed in Section 7.1.2.

7.1.1. Methanol Production Based on Catalytic Autothermal Reforming (Figs. 82 and 83)

Gas production, methanol synthesis, and subsequent distillation are the primary process steps. The relevant units are combined with an integrated steam system usually on several pressure levels. Steam is generated and conditioned in the synthesis, in the reformed gas and flue gas waste-heat system, and in the auxiliary boiler.

Natural gas is preheated and desulfurized. After desulfurization the gas is saturated with a mixture of preheated process water from the distillation section, process condensate, and make-up water in the saturator (d). The gas is further preheated and mixed with process steam as required for the prereforming process. In the prereformer (f) the gas is converted to hydrogen, carbon dioxide, and methane. Final preheating of the gas is achieved in the fired heater (b). In the autothermal reactor (g) the gas is reformed with steam and oxygen, the product gas contains hydrogen, carbon monoxide, carbon dioxide, and a very small amount of non-converted methane and inerts together with undecomposed steam. The reformed gas leaving the autothermal reactor possesses a considerable amount of heat, which is recovered in form of high pressure steam, preheating energy, and energy for providing heat for the reboilers in the distillation section.

The reformed gas is mixed with hydrogen from the pressure swing adsorption unit to adjust the synthesis gas composition. Synthesis gas is pressurized to 5 – 10 MPa by the synthesis gas compressor (o). Subsequently, this compressed gas is mixed with recycle gas from the synthesis loop. With the recycle gas compressor (t), the recycle: make-up gas ratio can be adjusted. The inlet gas to the reactor system consists mainly of hydrogen, carbon monoxide, and carbon dioxide as reactive components, along with methane and nitrogen as inerts. This gas mixture is converted to methanol according to the following exothermic reactions:

$$CO + 2\,H_2 \rightleftharpoons CH_3OH$$

$$CO_2 + 3\,H_2 \rightleftharpoons CH_3OH + H_2O$$

Details of the reaction mechanism are described elsewhere (→ Methanol). The synthesis reactor system (q) is the key section of the loop. Here superheating of the catalyst is avoided by generation of medium pressure steam and the reaction parameters are selected in such a way that properly conditioned purge gas is obtained.

In the pressure swing adsorption unit hydrogen is separated from the purge gas and routed to the reformed gas. The remaining gas is used as fuel in the fired heater.

The distillation unit is a three-column energy-saving system. In the light-ends column (u), low-boiling impurities are removed overhead.

Pure methanol distillation is divided into two sections. The first section (v) operates at elevated pressure, and ca. 50 % of the methanol is distilled overhead. In the second section (w), which operates at ambient pressure, the remaining methanol is separated from water and high-boiling impurities.

Whereas the first section is heated with low-pressure steam, heat for the second section is provided by condensing methanol vapor from the first section.

7.1.2. Comparison of Conventional Steam Reforming and Combined Reforming Processes for Methanol Production

The main reason for applying the combined reforming process in methanol production is adjustment of the stoichiometric number (SN) as defined in Table 39, closer to the theoretical value of 2.0. Whereas the practical SN is between 2.6 and 2.9 for conventional steam reforming, it is decreased to 2.03 in the combined reforming process.

Table 39. Conventional steam reforming versus combined process for methanol production

	Conventional steam reforming	Combined process
Process natural gas feed to tubular reformer, %	100	40 – 60
Process natural gas feed to autothermal reactor, %		60 – 40
Outlet pressure tubular reformer, MPa	1.9	3.7
Outlet pressure autothermal reactor, MPa		3.5
Outlet temperature tubular reformer, °C	875	780
Outlet temperature autothermal reactor, °C		950
Hydrocarbons converted in tubular reformer, %	83	22
Stoichiometric number, SN = $(C_{H_2} - C_{CO_2}) / (C_{CO} + C_{CO_2})$	2.6 – 2.9	2.03
Total natural gas lower heating value demand, GJ *	31.6	29.5
Thereof fuel consumption of reformer, GJ *	12.9	3.1
Discharge – suction pressure ratio (synthesis make-up gas compressor)	4.4	2.2
Instrument air **	separate unit	byproduct
Plant air **	separate unit	byproduct
Nitrogen **	separate unit	byproduct

* Per t of methanol.
** With oxygen generation in an air separation plant.

Besides reducing the SN, placement of an autothermal reactor downstream of the tubular reformer means that the tubular reformer can be drastically reduced in size for two reasons: first, it has to treat only about half of the natural gas feed while the other half is by-passed directly to the autothermal reactor and, second, it does not have to convert all of the methane but leaves an essential part for the secondary reforming step. So the fired steam reformer may be ca. 75 % smaller, can operate at lower temperature, and therefore, can operate at higher pressure than a steam reformer in the conventional steam reforming process.

Because of the lower SN, less surplus hydrogen must be compressed and introduced into the synthesis loop, which results in a significant reduction of equipment size and compression energy. Table 39 shows the main operating data of the combined reforming process in comparison with the conventional steam reforming process.

7.2. Hydrogen Production Based on Heavy Residues

Economic and environmental considerations have led petroleum refiners to intensify efforts to maximize production of high-quality products from crude oil. To meet this goal, a number of refinery processes have been developed to upgrade heavy residues to, for instance, naphtha and gas oil. Some of these conversion processes (e.g., hydrotreating or hydrocracking) require considerable quantities of hydrogen. Today most of the hydrogen is produced by steam reforming of natural gas (83 % of total production) or naphtha (14 %). Only 3 % is obtained by gasification of heavy residues [287]. However, though the capital investment for such a plant is considerably higher than for a unit based on steam reforming of natural or refinery gas, gasification of residues can still be advantageous, particularly where natural gas or other light hydrocarbons are not readily available or expensive.

In some refining schemes, particularly those for treating heavy crudes, the hydrogen balance of the refinery may require utilization of the residues for hydrogen production. In addition, partial oxidation of heavy residues can contribute to reducing the size of the fuel oil pool

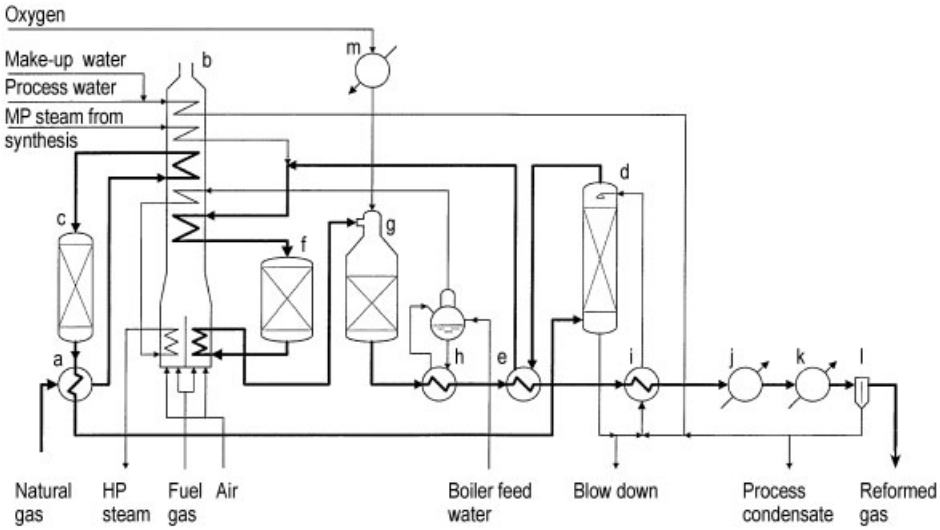

Figure 82. Methanol from natural gas, catalytic autothermal reforming route — production of synthesis gas
a) Natural gas interchanger; b) Fired heater; c) Desulfurizer; d) Saturator; e) Natural gas preheater; f) Prereformer; g) Autothermal reactor; h) Waste-heat boiler; i) Circulation water heater; j) Distillation reboilers; k) Final cooler; l) Final separator; m) Oxygen preheater

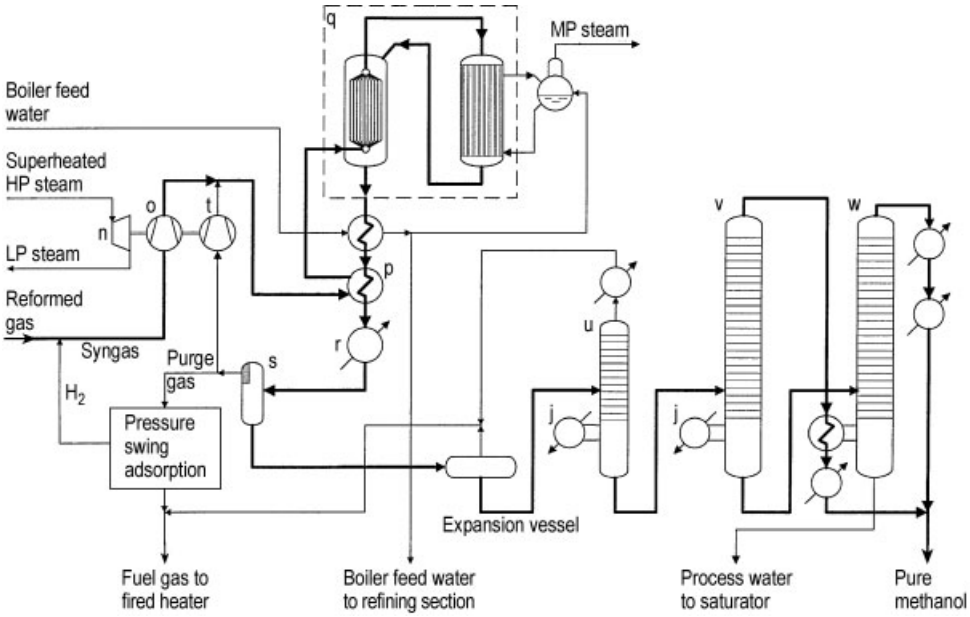

Figure 83. Methanol from natural gas, catalytic autothermal reforming route — methanol synthesis
n) Synthesis make-up gas compressor turbine; o) Synthesis make-up gas compressor; p) Heat exchanger; q) Synthesis reactor system; r) Cooler; s) Crude methanol separator; t) Recycle gas compressor; u) Light ends column; v) Pure methanol pressure column; w) Pure methanol atmospheric column

and increasing its quality by consuming some of the highest sulfur- and metal-containing components of the pool [288].

A typical hydrogen specification for a hydrocracker is

Hydrogen	$\geq$98 %
Carbon monoxide and carbon dioxide	$\leq$10 – 50 ppm
Oxygen	$\leq$100 ppm
Inerts (N_2, Ar, CH_4)	balance

The raw gas supplied by a partial oxidation unit fed with a heavy residue contains nearly 50 % carbon monoxide, as well as other components (CO_2, H_2S, COS) which must be converted or removed in a number of gas treatment stages to yield acceptable product quality (see Table 24, example 3). In practice, gas plants based on partial oxidation are often multiproduct plants, which in the refinery environment could imply cogeneration of clean fuel gas or, for example, methanol for methyl-*tert*-butyl ether (MTBE) (→ Methyl Tert-Butyl Ether) production.

Selection of gas treating processes must be made carefully, because ca. 20 – 30 % of the total investment for an overall hydrogen plant is concentrated in these units. Plant size and possible cogeneration of other products also influence the final choice of process.

For the partial oxidation unit, either of the three commercially proven processes (Shell, Texaco, Lurgi) is suitable.

Conventional Gasification-Based Hydrogen Production. The principal process steps for hydrogen production based on heavy residues as feedstock are as follows (see Figure 55):

- Gasification
- Acid gas removal (H_2S, CO_2), in one or two separate units
- CO shift conversion
- Final purification

The byproducts have to be handled within the refinery (steam, fuel gas) with the appropriate specifications or should be salable (sulfur, metal ash).

As outlined in Section 3.2.2, the gasification process itself can be applied in two different configurations followed by different downstream treatment modes of the raw synthesis gas:

1) Gasification in *boiler configuration* with downstream CO shift of desulfurized syngas.
2) Gasification in *quench configuration* with raw gas shift (see Section 5.1) prior to desulfurization (and CO_2 removal).

In the *boiler configuration*, the sensible heat of the hot raw gas leaving the gasification reactor is utilized to generate high-pressure steam in a waste-heat boiler before the gas enters the scrubber. After the subsequent desulfurization and a clean gas CO shift conversion, CO_2 must be removed from the syngas by a second washing step. This route has been chosen for the Shell Pernis plant [289]. The disadvantages of this concept are high capital costs due to the expensive gasification stage (waste-heat boiler), CO shift conversion (with cooler-saturator circuit) and the two-stage acid gas removal unit along with limited feedstock flexibility due to the boiler specification. An asset is a rather high thermal efficiency.

In the *quench configuration*, the raw gas leaving the gasifier is shock-cooled by direct injection and subsequent scrubbing with water. It can thus be routed directly to the sulfided CO shift conversion (raw gas shift). The steam content of the quenched gas is normally sufficient for the shift reaction, which leads to a quite simple layout of the shift conversion plant. Furthermore, trace components like COS and HCN are destroyed in the shift reactor, which facilitates the subsequent acid gas removal. This concept has been proposed by Texaco [290] and Lurgi [291, 292]. The advantage of this concept is lower capital cost for both the gasification unit and the shift. A disadvantage is the loss in thermal efficiency.

Improved Route for Hydrogen Production. In order to further increase the economics of a gasification-based hydrogen production unit compared to a further steam reforming step, the above described quench configuration route has been improved by Lurgi introducing the Multi Purpose Gasification together with an advanced ash handling system and an optimized CO shift conversion. The new concept consists of the following process units (Fig. 84):

- Air separation and oxygen compression,
- Multi Purpose Gasification (MPG) in quench configuration,
- Metals Ash Recovery System (MARS),
- Sulfided CO-Shift (single-stage raw gas shift),
- Sulfur removal (Rectisol process),
- Pressure swing adsorption (for remaining CO_2 and low concentrated impurities),

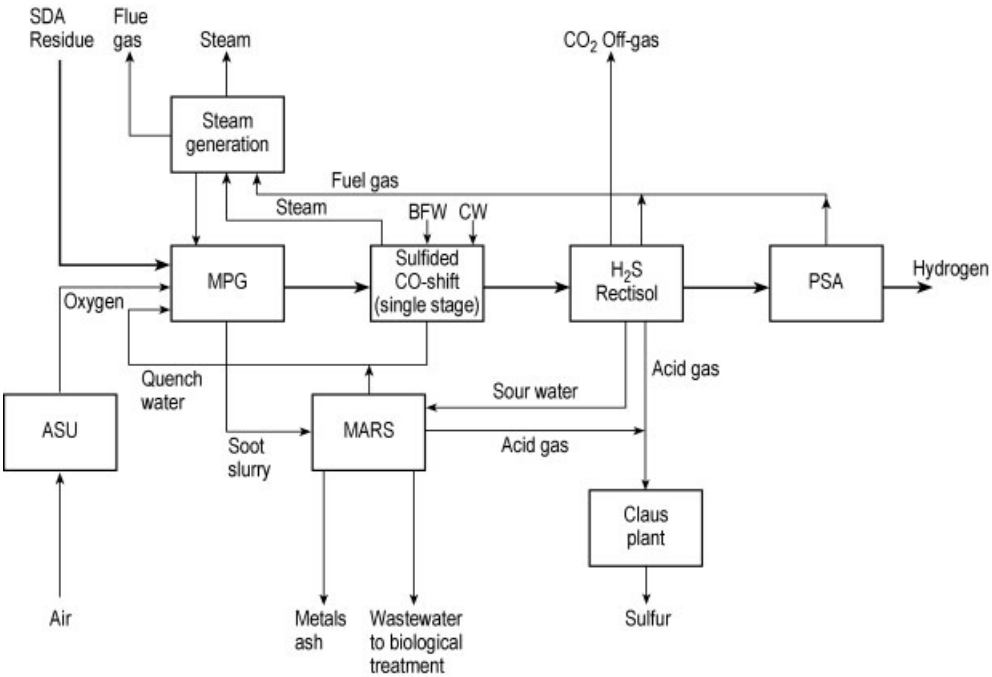

Figure 84. Block flow diagram of the MPG based hydrogen production
ASU = Air separation unit; MPG = Multi Purpose Gasification; MARS = Metals Ash Recovery System; PSA = Pressure swing adsorption; SDA = Solvent deasphalting; BFW = Boiler feed water; CW = Cooling water

– Sulfur recovery unit (Claus process).

Two main features make this concept most economic compared to conventional process routes:

1) The MPG process with its high feedstock flexibility allows the simultaneous processing of even the heaviest refinery residues and wastes, which can reduce the costs for feedstock supply (see below).
2) In the sulfided CO shift, the raw gas is converted in a single-stage reactor which results in an increased CO content in the converted gas. On the one hand, this leads to a lower H_2 yield and therefore requires more feedstock to be processed in gasification. The advantages, however, are:
   – A self-burning PSA off-gas which can directly be used as fuel gas in an auxiliary boiler or sent to the refinery's fuel gas system. There is no need for fuel import.
   – The capital costs for the overall plant are reduced because of the cheaper CO shift conversion and the cheaper CO_2 removal.

A simplified process flow diagram of the gasification and shift conversion section is shown in Figure 85. The raw gas is produced by partial oxidation of heavy residues using Lurgi's Multi Purpose Gasification Process in quench configuration at 6.0 MPa. The water-saturated raw gas from the quench vessel enters the scrubber at ca. 240 °C, where residual soot/ash traces are removed by injection of additional quench water. No further raw gas cooling occurs in the scrubber. The water leaving the quench section contains most of the soot and ash from the raw gas and is further treated in the Metals Ash Recovery System (see Section 3.2).

The raw gas from the MPG unit is saturated with water and routed to the CO shift conversion unit. After preheating of the raw gas, its CO content is reduced in a single-stage reactor to a residual 6 to 9 %, depending on the composition of the feedstock. No further steam injection into the shift feed is necessary to maintain the shift reaction. The sensible heat of the shifted gas is recovered by producing steam of different pressure levels and by preheating several water streams (quench water, demineralized water). The con-

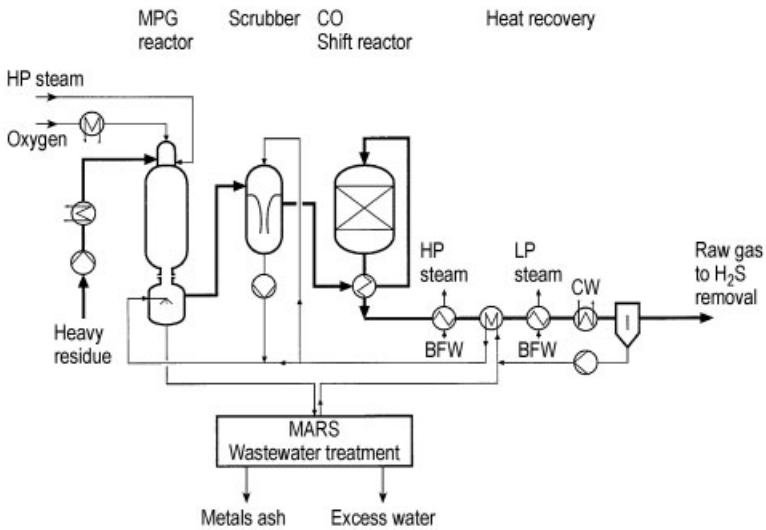

Figure 85. Simplified process flow diagram for the gasification (quench configuration) and the CO shift section in Figure 84

densate from the cooled shift product gas is sent to the MPG unit as quench and scrubbing water.

Both carbonyl and hydrogen sulfides are washed out of the gas to a residual 1 ppm in a *one—stage Rectisol unit*. The factors determining the selection of the Rectisol process instead of a chemical absorption system are

- Low steam consumption,
- High selectivity for H_2S versus CO_2 for the generation of a rich acid gas to facilitate the sulfur recovery.(see 5.4.2.4) for detailed discussion on acid-gas removal options).

The acid gas, which contains 25 – 60 % hydrogen sulfide, depending on the sulfur content of the feedstock, is normally discharged to a central Claus plant that also treats other sulfur containing gas streams originating in the refinery to produce elemental sulfur.

In the *final purification* of the raw hydrogen in the pressure-swing adsorption unit CO_2 and other impurities are reduced to 10 ppm of carbon dioxide, 10 ppm of carbon monoxide, and 0.1 % (maximum) of inerts, yielding hydrogen with a purity > 99.9 % (see Section 5.5) at a pressure between 4.7 and 4.8 MPa.

Plant Performance. For easier comparison of the improved concept with conventional process routes, three cases are considered for the economic evaluation:

Case 1: = MPG (quench) + single-stage raw gas shift + H_2S absorption + PSA

Case 2: = MPG (quench) + two-stage raw gas shift + acid gas absorption (H_2S, CO_2) + PSA (for remaining impurities only)

Case 3: = Texaco Gasification (TGP, quench) + two-stage raw gas shift + acid gas absorption (H_2S, CO_2) + PSA (based on [290], scaled to hydrogen production without power generation)

The characteristics of the respective plant performance are summarized in Table 40.

To compensate for the lower H_2 yield of the MPG/single-stage CO shift process variant, the plant capacity must be by 6 % higher which leads to higher variable operating costs. On the other hand the simple plant configuration of this route results in lower capital costs and fewer maintenance costs and therefore reduced fixed costs which makes this concept the most economical one.

7.3. Combined Cycle Power System Based on Coal

7.3.1. Introduction

Combined cycle power plants are an attractive option to raise the efficiency of electric power generation. This power plant concept combines a gas turbine and a conventional steam turbine

Table 40. Performance of improved-concept plants for hydrogen production from heavy residue

	Case 1 MPG, single-stage CO shift	Case 2 MPG, two-stage CO shift	Case 3 TGP [290]
Inputs	SDA pitch (S: 5 wt %, ash: 0.07 wt %)	SDA pitch (S: 5 wt %, ash: 0.07 wt %)	Asphalted vacuum residue
Feed oil, t/d	960	905	1050
Oxygen (99.5 %), t/d	925	875	1120
Process steam to gasifier, t/d	430	405	475
Intermediate products			
Fuel gas from PSA (LHV, kJ/m^3 (STP)	> 4500	ca. 8000	ca. 1000
Acid gas from Rectisol unit (vol % H_2S)	< 60	< 60	not available
Products			
Hydrogen, t/d	188.4	188.4	218
Byproducts, t/d			
Sulfur	47.8	45.1	58
Metals, ash	0.6 (76 wt % V_2O_5)	0.56 (76 wt % V_2O_5)	8.5 (filter cake, dry)
Steam export	1920	1810	2110 (calculated)

cycle. Based on available industrial gas turbines early plant concepts were developed and introduced in the early 1950s, these plants were fueled by conventional oil-derived gas or almost any kind of fuel gas. Although basic process concepts are well established, novel concept variations are still under development [293].

Due to its relatively low price as well as regional availability, coal is the major fuel for fossil-fueled power plants. To use combined cycle technology, coal must be transformed into a clean fuel applicable to gas turbine operation (the same applies to residual oil, petroleum coke, or wood). Impurities, which must be removed, are ash, alkali, halogens, and sulfur. Typical gas turbine specifications are given below:

Inlet temperature, °C (according to ISO)	1190
Na + K, mg/kg	0.08
V, mg/kg	0.1
Pb, mg/kg	0.25
C, mg/kg	2.5
Cl, mg/kg	2.45
Dust < 2 µm, mg/kg	4.4

The only way to meet these requirements is gasification because

1) Ash remains in the gasification residue,
2) Alkali is trapped within the ash or removed in one of the purification steps, and
3) Halogens and sulfur are removed by dry or wet absorption.

The combined effect of high efficiency and low pollution makes the combined cycle with upstream coal gasification an attractive approach for power generation from coal. In principle, two designs are possible: integrated and nonintegrated plants. In the latter, the gasification and gas purification process units are installed and operated separately from the power plant. This makes power plant operation easier, but an integrated plant (IGCC) allows for more sophisticated energy management and higher efficiency [294]. Hence, most engineering and economic studies and first-of-its-kind plants refer to this design principle. In this section only the principles of the integrated combined cycle based on the gasification of solid fuels are discussed [295, 296].

Environmental Aspects. The emissions of combined power cycles are inherently lower than that of coal combustion plants [297 – 299]. This is due mainly to the following characteristics:

1) Desulfurization of the fuel gas can be accomplished to any extent necessary at a relatively low additional cost. For applicable processes, see Section 5.4.
2) The byproduct of desulfurization is elemental sulfur, which can be sold or easily disposed of, in contrast to flue gas scrubber residues.
3) Low NO_x emissions are achievable due to the tightly controllable combustion conditions in a gas turbine.
4) Extensive particulate removal is essential for gas turbine operation, resulting in dust emissions far below existing requirements.

7.3.2. Fundamentals

Thermodynamics [300, 301]. Electric power is conventionally generated by combustion to produce high-temperature thermal

energy and converting that thermal energy into work to drive a generator. The working fluid can be the combustion gas itself or a special working medium (steam, air) to which heat is transferred from the combustion gas.

The upper limit to the proportion of thermal energy converted to work is set by Carnot's law

$$\eta \approx 1 - \frac{t_2}{t_1}$$

where t_1 and t_2 are the upper and the lower absolute temperatures of the working fluid and η is the conversion efficiency. For a *steam turbine cycle* construction materials usually limit the upper temperature to ca. 560 °C, the lower temperature depends on the available cooling water temperature. With sea water cooling lower temperatures of ca. 30 °C are achievable. For an open cycle *gas turbine* (stationary application), t_1 may be as high as 1050 °C, with t_2 being ca. 500 °C depending on the pressure ratio.

When gas and steam turbines are combined by using the thermal energy of the gas turbine exhaust gas to produce and superheat the steam from the steam cycle, t_1 may be as high as 1050 °C and t_2 as low as 30 °C. This, in simplified terms, is the basis for the efficiency increase when a *dual cycle* combining gas and steam turbines is used instead of a *single* working fluid cycle. Some losses occur along the route from coal to electrical energy. For simplification the overall efficiency of the process can be written as the product of individual efficiencies:

$$\eta = \eta_{CG} \cdot \eta_{GC} \cdot \eta_{CC}$$

where η_{CG} is the efficiency of coal gasification; η_{GC}, the efficiency of gas cleaning; and η_{CC}, the efficiency of the combined cycle.

Efficiency of Coal Gasification η_{CG}. Coal gasification processes suffer losses of useful energy due to the following mechanisms:

1) The heat of evaporation of gasification steam, coal moisture or slurry water cannot be recovered.
2) The bottom or fly ash inevitably contains unconverted carbon which is lost for the process. The same applies for the sensible heat of the ash.
3) Heat is lost to the environment, e.g., through the walls of apparatus and pipes.
4) The operation of drives, pumps and compressors consumes electric energy, which can be recovered only partly.
5) The gasification reactions require elevated temperatures. During cleaning of the hot raw gases, according to the requirements of the gas turbine operation and of the environment, the raw gas has to be cooled down by heat exchange which involves a certain degree of irreversibility.

Each of the mechanisms leads to a different amount of losses for the individual gasification processes. Reference figures for η_{GC} may be taken from Section 4.4, 4.5, 4.6, 4.7 (Tables 18, 20, 21, 23).

Efficiency of Gas Cleaning η_{GC}. The losses of useful energy during gas cleaning can mainly be attributed to

1) Nonreversible heat exchange during cooling of the raw gas and reheating of the clean gas,
2) The energy required for the regeneration of the solvent or adsorbent, which in most instances is lost,
3) The removal of sulfuric components and the co-removal of other combustible constituents of the raw gas which, therefore, cannot contribute to the combined gas/steam turbine operation,
4) The removal of inerts such as H_2O and CO_2 which reduces the mass flow through the gas-turbine and hence its energy output, and
5) The consumption of electric energy.

In most cases η_{GC} reaches 96 – 98 %.

Efficiency of Combined Cycle η_{CC}. For thermodynamic reasons the efficiency of the conversion of thermal energy into work and consequently electric power is limited as has been explained earlier. In addition to that, other losses of useful energy must be considered, e.g.:

1) The heat of evaporation of boiler feed water is only partly used for generation of power.
2) The nonideal operation of gas and steam turbines increases the entropy of the working fluids.
3) The flue gases leave the stack with a temperature above 100 °C, their thermal energy is not used for the process.
4) Heat losses through the insulation.

Based on *natural gas*, a combined cycle with a gas turbine inlet temperature of 1100 °C (according to ISO) may reach 48 % efficiency [299, 302].

At a *coal gasification* efficiency of ca. 90 % a combined cycle power generation unit based on coal gasification achieves an overall efficiency $\eta > 42\,\%$.

General Plant Arrangement. In principle, any high-pressure coal gasification process mentioned in Chapter 4 can be used as the front-end process. The selection of a gasification process depends on several boundary conditions, the most important of which are

1) Type of available coal (rank, particle size, ash, and moisture content);
2) Gasifier efficiency; and
3) Maturity of the process and its proven reliability.

Coal gasification can be integrated with combined cycle power generation in numerous ways. Modifications differ in the degree of carbon conversion (partial or total gasification), the gasification medium (air or oxygen), the measures taken to adjust the turbine inlet temperature (excess air or steam production), and the general arrangement (stand-alone plant or repowering). Advantages and limitations of the individual variants are described in [294, 297], and [303 – 307].

7.3.3. Installations and Design Studies

Installations. The first commercial-scale IGCC plants have been built at Lünen by Steag in the Federal Republic of Germany [308] and near Barstow, California (The Cool Water plant). The Lünen plant with a design output of 170 MW operated between 1972 and 1977; the Cool Water plant (100 MW) began operation in 1984. At present both plants are no longer in operation.

Lünen IGC Plant. The general arrangement of the Lünen plant is shown in Figure 86. Gasification takes place in 4 + 1 Lurgi moving bed gasifiers (see Section 4.4). High volatile subbituminous coal is gasified with steam and air countercurrently at a pressure of 2 MPa (c). The raw gas leaves the gasifiers with a temperature of 500 – 600 °C. It is cleaned in a scrubbing unit (d) where dusty and tarry compounds are removed from the gas. Simultaneously the gas is cooled to ca. 160 – 170 °C and saturated with water. The dusty tar removed from the gas is recycled in total to the gasifiers after separation from the wash water. The clean gas is passed through a first gas turbine (e) which drives the booster compressor (b) for the gasification air. Here the pressure of the clean gas is reduced from 2 to ca. 1 MPa. A bypass stream of 1 % of the clean gas is desulfurized with a hot potassium wash solution for pilot test purposes.

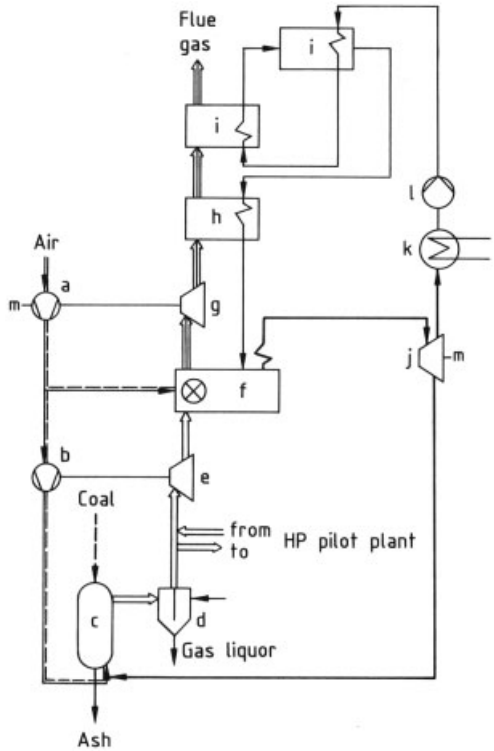

Figure 86. Simplified flow diagram of the IGCC Lünen plant
a) Air compressor; b) Gasification air compressor; c) Lurgi gasifiers (in all five); d) Wet gas cleaning; e) Expansion turbine; f) Supercharged boilers (in all two); g) Gas turbine; h), i) Boiler feed water preheaters; j) Steam turbine; k) Condensor; l) Boiler feed water pump; m) Generator

The clean gas is combusted in two supercharged boilers (f) at 1 MPa which is a particular feature of this plant. In these boilers 340 t/h steam are generated (13 MPa, 525 °C) powering a steam turbine (j) with an output of 96 MW.

The flue gases leaving the supercharged boilers drive a second gas turbine (g) with an output of 74 MW. The compressor for the total plant air is connected to the same shaft. In the gas turbine the flue gas enthalpy is reduced from 1.0 Pa, 810 °C down to 0.1 MPa and 400 °C. The residual thermal energy of the exhaust gases is used for boiler feed water heating (i, k), the stack inlet temperature is 170 °C [297, 309].

As a first-of-its-kind plant, the Lünen plant did not operate continuously; altogether it produced electrical energy for the grid for ca. 10 000 h. During this time, the data in Table 41 were obtained.

Table 41. Performance data, in percent, of the Lünen plant at 70 % partial load

Efficiency	Partial load			100 % load
	Design	Operation		Design
	Design coal	Coal 1 *	Coal 2 **	Design coal
η_{CG}	91.7	90.2	86.1	92.9
η_{CC}	35.8	35.2	35.0	39.4
η	32.8	31.7	30.5	36.7

* Coal 1 = design.
** Coal 2 = low grade (40 % < 5 mm, 28 % ash).

The plant generally operated at partial load and with low-grade coal. Table 41 also shows estimated full load design data which were not achieved because only 70 % of the necessary gasification air could be supplied due to limitations in air compression.

The second commercial-scale installation is the Cool Water plant [310, 311]. The plant is of commercial size, with a net output of ca. 100 MW.

In this plant, a Texaco entrained-flow gasifier (see Section 4.6) was used, which gasified a coal – water slurry with oxygen. According to prevailing emission standards, the gasification raw gas was purified by a physical absorption process (Selexol, see Section 5.4), which reduces the content of carbonyl and hydrogen sulfide in the purified gas to 50 ppm. Besides demonstrating the IGCC process principles the Cool Water plant was used to test construction materials and general control strategies.

Design Studies. The two plants have not been designed to exploit the full potential of IGCC. Their main purpose was the demonstration of the technical feasibility and operability of plants of this kind.

Hence the overall efficiency of these early IGCC plants is limited compared to modern power stations based on the Rankine process.

The successful introduction of natural gas fired IGCC plants stimulated numerous design studies based on gasification of solid fuels. Gasification processes which were considered are the Lurgi fixed-bed gasification [297], the British Gas/Lurgi slagging gasifier [304, 312, 313], entrained-flow gasification [304, 305], and fluidized-bed gasification [305, 314 – 316]. Depending on constraints like available fuel quality and sophisticated heat recovery systems efficiencies $\eta \gg 40$ % can be expected.

The most recent commercial scale IGCC power plants based on coal gasification have been commissioned at Buggenum (NL) (250 MW) by DEMKOLEC B.V. [317 – 319], at Puertollano (Spain) (320 MW) [320]; and Tampa (USA) [322]. The DEMKOLEC Buggenum plant applies entrained-flow gasification for turbine fuel production; standard fuel is conventional hard coal (ca. 2000 t/d). Commercial operation started in 1994. In a single train 250 MW of electric power are generated; based on the experience gained so far, capacities of 500 MW are regarded as feasible. The block diagram (Fig. 87) depicts the main building blocks of a typical IGCC layout following the DEMCOLEC design; Figure 88 presents an aerial view of the total plant . A major part of the setup is taken up by the gas cleaning and desulfurization systems. A conventional Claus – SCOT plant converts H_2S to elemental sulfur. The IGCC plant at Puertollano uses an entrained-flow gasifier to convert low grade subbituminous coal into fuel gas for a gas turbine [320, 321]. This project is jointly executed by several companies of different European countries and funded by the European commission. The Tampa IGCC plant was commissioned in 1996, using entrained-flow gasification; in addition to hard coal refinery residues can be used for gasification [322].

Besides these plants in operation, numerous studies investigated the suitability of IGCC plants for a wide range of feedstocks [314, 316, 323]. IGCC concepts have also been investigated with regard to the gasification of unconventional fuels [319, 323 – 325]. The use of biomass de-

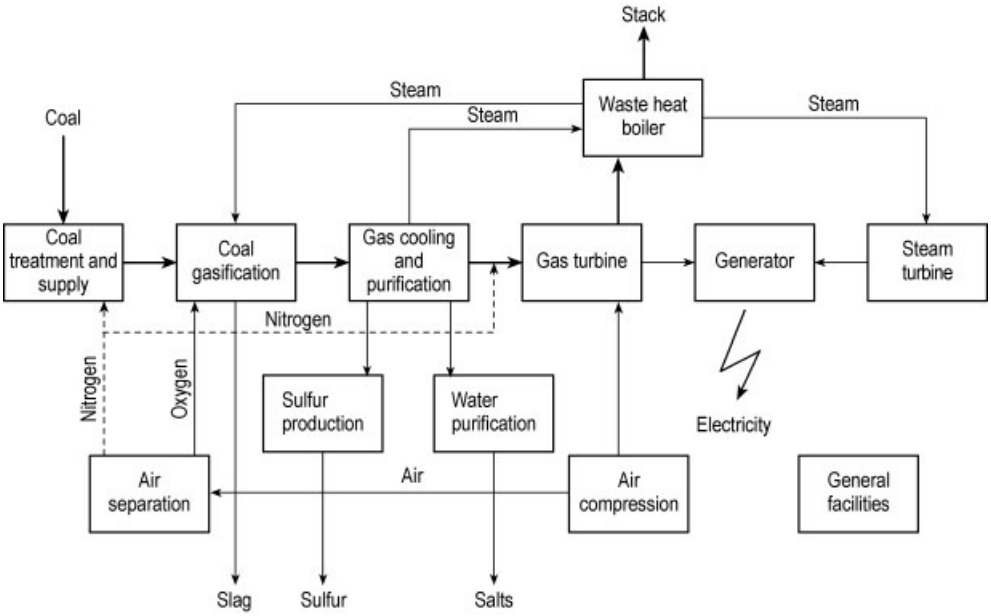

Figure 87. Principle layout of an IGCC plant

rived fuels opens the possibility for a reduction of CO_2 emissions whilst retaining high power generating efficiencies.

Figure 88. Aerial view of the DEMKOLEC plant

8. Analysis and Quality Control

Fuel gases [336] are sold principally as a source of energy, but in units described by volume. Consequently, considerable demand exists for specification of the thermal content per unit volume. The range of fuel gases that will burn satisfactorily on a particular appliance is limited and best described by the *Wobbe index* of the gas. The Wobbe index is a measure of the heat input to an appliance at constant pressure, and is expressed as the ratio of the calorific value to the square root of the relative density. This property should also be specified. *Safety requirements* relate to the flammable nature of the fuels, and engineering considerations relate to their distribution and storage.

Specifications for fuel gases cover various properties and may be devised or imposed by different authorities. Requirements for the specification of and means of ensuring *satisfactory odorization* are commonly a responsibility of government. Control of Wobbe index and of particular components that may have operational implications, such as water and condensable hydrocarbons, are usually the responsibility of the supply and distribution company. Different countries have different specifications [337], although common interests ensure a degree of consistency.

If a fuel gas is to be used as a *chemical feedstock*, quite different specifications will be used; total carbon content and the quantity and types of sulfur compounds present might be of much greater importance.

8.1. Quality Specifications

Natural gas supplies by far the greater part of the fuel gas market. The quality specifications

Table 42. Classification of fuel gases according to gas family (DVGW [a] Rule G 260) [340]

	Symbol	Unit	First Family (town gas)		Second family (natural gas)		Third family (LPG)
			Group A	Group B	Group L [b]	Group H [c]	
Wobbe index	W_s						
Total range		MJ/m^3 (STP)	23.0 – 28.1	28.1 – 33.5	37.8 – 46.8	46.1 – 56.5	requirements for
Nominal range		MJ/m^3 (STP)			44.6	54.0	propane and
Range of variation in location distribution	W_s	MJ/m^3 (STP)			+ 0.6 to − 1.4	+ 0.7 to − 1.4	propane – butane mixtures ($\geq$ 60 wt % C_4 hydrocarbons)
Gross calorific value	H_s						
Total range		MJ/m^3 (STP)	16.6 – 19.8	18.0 – 21.2			according to DIN
Nominal range		MJ/m^3 (STP)	17.6	19.8			51 622
Range of variation in location distribution	H_s	MJ/m^3 (STP)	± 0.3	± 0.3			
Relative density	d		0.40 – 0.60	0.32 – 0.55	0.55 – 0.704		
Appliance inlet pressure	p_e						
Total range		kPa	0.75 – 1.5	0.75 – 1.5	1.8 – 2.4		4.25 – 5.75
Nominal range		kPa	0.8	0.8	2.0	5.0	

[a] DVGW = Deutscher Verein des Gas- und Wasserfaches e.V.
[b] L = low.
[c] H = high.

of natural gas have been reviewed in both European and North American contexts [338, 339].

8.1.1. Combustion Characteristics

Fuel gases are categorized in one of three families: (1) manufactured gas (high hydrogen content), (2) natural gas and (3) liquefied petroleum gas [340]. Table 42 shows these families and the subdivisions for manufactured and natural gas, within which interchangeability [341, 342] may be assumed. Test gases for evaluation of appliances within these families are described [343].

Other combustion properties are *flame lift* [344 – 346], *incomplete combustion* (yellow tipping) and *flashback* [347].

8.1.2. Minor Constituents

Specifications from several (mainly European) countries have been compared [337]. Typical national requirements for maximum levels of impurities (Table 43) [340] and odorization [348] are described.

Table 43. Recommended limits of gas impurities according to DVGW * Rule G 260 [340]

	Gas family	
	First	Second
Benzene hydrocarbons, g/m^3 (STP)	10	
Naphthalene, mg/m^3 (STP)	50/p (p in bar)	
Hydrocarbon dew point, °C	soil temperature at	
Water dew point, °C	pipeline pressure	
Fog, dust, liquids	technically pure	
O_2 (dry distribution grid), vol %	3	3
O_2 (wet distribution grid), vol %	0.5	0.5
NO, cm^3/m^3 (STP)	0.2	
Total sulfur, mg/m^3 (STP)	200	120
Short-term value, mg/m^3 (STP)		150
Mercaptan sulfur, mg/m^3 (STP)		6
Short-term value, mg/m^3 (STP)		16
H_2S, mg/m^3 (STP)	2	5
Short-term value, mg/m^3 (STP)		10
NH_3, mg/m^3 (STP)	3	
CN, mg/m^3 (STP)	150	
CO, vol %	6	

* DVGW = Deutscher Verein des Gas- und Wasserfaches e.V.

8.2. Test Methods

8.2.1. Determination of Combustion Characteristics

8.2.1.1. Calorimetry

The calorific value (heating value) H of a fuel gas is defined as the amount of heat released by complete combustion in air, at constant pressure p_1, with the combustion products being returned to the same temperature t_1 as that of the reactants prior to combustion. Normally the *superior calorific value* H_S (synonymous *higher heating value*, HHV) is experimentally determined with all the water formed by combustion being condensed to the liquid state at t_1. The *inferior calorific value* H_I (synonymous *lower heating value*, LHV) is defined as the amount of heat released with all the water formed remaining in the vapor phase at t_1 and p_1.

The calorific value H can be measured or inferred in three ways: (1) direct combustion with measurement of heat released, (2) stoichiometric combustion with measurement of gas – air ratio, and (3) calculation from composition.

Direct Measurement. The calorific value can be measured directly in a *reference calorimeter*, in which the heat liberated from a discrete amount of gas is measured thermometrically [349]. More frequently, however, *continuous recording* calorimeters are used [350 – 354]. The principle of continuous recording calorimeters is combustion of a controlled stream of gas, followed by heat exchange of the products of combustion with another fluid, whose temperature rise is recorded.

Stoichiometric Combustion. Stoichiometric combustion relies upon the constancy of the ratio of calorific value to air – gas ratio for saturated hydrocarbons. For this reason, it can be used for natural gas, but not manufactured gas [355]. Stoichiometric combustion devices have been used as part of an energy flow meter [356], where the rate of supply of gas to the meter is kept exactly proportional to the flow in the main stream from which it is taken.

Calculation from Composition. The composition of fuel gases can be defined very well by gas chromatographic measurement. Knowledge of calorific values for the pure components and an appropriate calculation routine [357, 358] lead to the calorific values H_S and H_I of the mixture.

8.2.1.2. Density

Gas Density Balance. The gas density ϱ (mass per unit volume) or relative density d (ratio of the density of the fuel gas to the density of dry air) [359] is most commonly determined by a gas density balance [360, 361]. The method is based on the buoyancy of a nitrogen-filled sphere in the test gas. Instruments of this type also allow continuous monitoring of gas density.

Oscillating Device. Another method for determining density employs an oscillating device (e.g., a tuning fork). The frequency of oscillation is a function of the density of the gas surrounding it [362].

Both types of density recorders are checked and calibrated by standard gas mixtures of known (e.g., gravimetrically determined) density [363]. Density and relative density may also be calculated from gas analysis [357].

8.2.1.3. Wobbe Index

The Wobbe index W is defined as the ratio of calorific value H to the square root of relative density at the same specified metering reference conditions. As with the calorific value it must be distinguished between *superior Wobbe index* W_S and *inferior Wobbe index* W_I. Normally, the Wobbe index W is not measured directly, although it is one of the most important combustion characteristics for describing gas quality.

Calculation of Wobbe Index. The index can be calculated from the calorific value and the relative density more accurately than by direct determination. The Wobbe index may also be obtained by combining the signals of a calorimeter and a densimeter electronically, according to its definition equation.

Direct Measurement of Wobbe Index. Direct measurement of the Wobbe index with relatively simple and rapid instruments is recommended only for mixing gases of different quality to yield a mixture of constant Wobbe index. These *Wobbe recorders* are automatic calorimeters whose gas flow rate is adjusted so that it is proportional to $1/\sqrt{d}$.

8.2.2. Analytical Methods

The analytical methods listed in the following sections are (1) general methods (i.e., capable of measuring several components, Section 8.2.2.1); (2) specific methods for individual components (Section 8.2.2.2); or (3) methods suitable for minor components (Section 8.2.2.3).

Sampling is the first, and a most important, stage in gas analysis. Therefore, great care should be taken to obtain representative samples for determining gas properties and composition [364]. Continuous automatic sampling techniques have also been developed [365].

8.2.2.1. General Methods for Determination of Several Components

Orsat Method. The main constituents of manufactured gases (CO_2, O_2, CO, unsaturated hydrocarbons, H_2, saturated hydrocarbons and N_2) were determined volumetrically by the classic Orsat method involving successive absorption of the constituents in specific reagents and selective combustion of hydrogen and saturated hydrocarbons [366]. A simplified Orsat apparatus is used at present only for flue gas analysis or in combination with the chromatographic Janak method for carbon dioxide, oxygen, and unsaturated hydrocarbons.

Janak Method. The Janak procedure, which is suitable for manufactured gases as well as for natural gas components up to C_4H_{10}, is a simple chromatographic method using carbon dioxide as carrier gas [367]. The carrier gas emerging from the column passes to a burette filled with concentrated potassium hydroxide solution, in which carbon dioxide is absorbed and the remaining components can be measured consecutively by volume. The apparatus available commercially registers the volume ratio of the different components with an automatic burette.

Gas Chromatography (GC). Besides the Janak method, GC [368 – 370] with a thermal conductivity or flame ionization detector is most often recommended for fuel gas analysis. A large number of standardized methods exist for the analysis of manufactured gas [371] and natural gas [372 – 374] by using an *adsorption column* (e.g., molecular sieve) for the separation of hydrogen, helium, oxygen, argon, nitrogen, methane, and carbon monoxide, and a *partition column* (e.g., organic stationary phase on a solid support) for the separation of higher hydrocarbons. For the analysis of natural gas, determination of hydrocarbons up to pentane [372] or octane [373, 374] is normally sufficient. For specific needs, methods have been developed for analysis up to hexadecane [375].

For onstream applications, these analyses have been completely automated (process GC): this includes periodic calibration of the total system with a calibration gas mixture [376] and evaluation of the chromatogram by an electronic integrator. Gas analyses can also be performed by mass spectrometry [377]. The results of these analyses can be used to calculate the physical, thermodynamic, and combustion characteristics of the gas (H_S, H_I, W_S, W_I, ϱ, etc.) [357, 358].

8.2.2.2. Specific Methods for Determination of Individual Components

Hydrogen. In binary or quasibinary gas mixtures, the hydrogen content can be monitored easily by measuring *thermal conductivity*, because the conductivity of hydrogen is much higher than that of all other likely components.

Oxygen. Besides volumetric [366], chemical [378], and GC methods [373], oxygen (concentration >0.1 vol %) may be determined without significant interference from other components by using instruments based on the measurement of *magnetic susceptibility* [379].

Carbon Dioxide. The carbon dioxide concentration [380] of fuel and flue gases is normally recorded by nondispersive *infrared* instruments [381, 382]. The carbon dioxide absorption band is overlapped only slightly by the carbon

monoxide band, and for a wide range of concentration ratios this interference can be eliminated by using a filter cell containing carbon monoxide. Several instruments are available that use single- or double-beam techniques and different types of detectors. For discontinuous determination, aforementioned volumetric [366] or GC methods [371, 373] may be used.

Carbon Monoxide. *Infrared absorption* is also the recommended method for monitoring the carbon monoxide concentration of gases [383]. Interference from carbon dioxide must be filtered out in the manner described above. Gas chromatography [384] and detector tubes [385] may be used for consecutive measurements or for a single check. Carbon monoxide detector tubes are subject to interference from propane and higher hydrocarbons, which may be found in natural gas. This interference can be eliminated by first leading the gas through an adsorption tube filled with activated carbon.

8.2.2.3. Methods for Determination of Minor Components

Water Content. For discontinuous *gravimetric determination* of water, a drying tube containing a suitable desiccant (e.g., magnesium perchlorate) can be used [386]. Alcohols such as methanol or glycols are also absorbed. Water and methanol (concentration > 20 mg/m^3) can be determined by GC [387].

The *Karl Fischer method* [388], in which water reacts with iodine in an alcoholic solution of sulfur dioxide and pyridine, is suitable for the determination of low water content of the order of milligrams per cubic meter. Newer equipment in which iodine is generated coulometrically [389] is particularly suitable for this application.

A variety of *hygrometers* use different detectors [390, 391], such as *electrolytic cells* with phosphorus pentoxide as absorbent, *conductivity cells* with glycerol [392], a *capacitor* with aluminum oxide as the moisture-sensitive dielectric medium [393], *silicon chips*, and an *oscillating quartz crystal*. By absorption or adsorption of water from the surrounding gas, the electrolytic current, conductivity, dielectric properties, or frequency of oscillation is changed. Better instruments use a sophisticated purge and calibration technique to overcome contamination problems caused by hydrocarbon condensates and other impurities such as glycols or ethanolamines derived from gas treatment processes [394]. Some of these instruments measure water content in parts per million (mL/m^3 or mg/m^3); others indicate dew point temperature.

Although water content and dew point can be related theoretically, some uncertainty exists as to whether this can be applied with any degree of accuracy to uel gases. Measurement from any instrument is more safely related to the technique used for calibration.

Water Dew Point. Because of the risk of hydrate formation [395], gas temperature should not drop below the dew point. The water dew point can be determined directly when a constant stream of gas passes over a polished metal mirror that is cooled slowly [396]. The dew point corresponds to the temperature at which the first signs of condensation appear on the mirror. The measurement can be performed under pressures up to 10 MPa. Completely automated instruments employing photocells are also used [390]. If the gas contains methanol or glycols, the dew point corresponds to the simultaneous condensation of water and alcohol. If the methanol content is known, a correction can be applied to the observed dew point [396]. Condensation of hydrocarbons may also interfere, when the hydrocarbon dew point is higher than that of water. This interference can be reduced significantly by first passing the gas through high-boiling liquid paraffin.

Sulfur Compounds.

Total Sulfur. The *combustion method* is applicable to all fuel gases. A measured volume of gas is burned with excess air or oxygen in a specially designed apparatus (e.g., Wickbold [397], lamp [398], Lingener method [399]). Combustion products are absorbed in an oxidizing solution. The sulfuric acid formed is determined by *volumetric, gravimetric*, or *turbidimetric* methods [400]. Sulfur dioxide can also be titrated *coulometrically* [401].

In the *hydrogenation method*, sulfur compounds are catalytically hydrogenated to hydrogen sulfide at high temperature over platinum gauze in the presence of excess hydrogen and

water vapor. Subsequent determination of hydrogen sulfide by the *methylene blue procedure* is very sensitive [402]. The *lead acetate paper method* in combination with hydrogenation is less precise but can be automated for continuous photometric recording [403].

Simultaneous Determination of Hydrogen Sulfide, Carbonyl Sulfide, and Mercaptan Sulfur. A measured volume of gas is passed through two sets of bubblers in series. The first set contains concentrated potassium hydroxide solution in which hydrogen sulfide and mercaptans are absorbed; the second contains alcoholic monoethanolamine solution, which absorbs carbonyl sulfide. Potentiometric titration with silver nitrate measures each species [404].

Individual Sulfur Compounds. The preferred method for determining individual organic and inorganic sulfur compounds is GC. Although sulfur compounds can be detected by thermal conductivity or flame ionization detectors (the latter for organic sulfur compounds only), sulfur-specific detectors are preferred for the analysis of naturally occurring or added odoriferous sulfur compounds to avoid interference from other components. The *flame photometric detector* [368, 405] measures the light emitted at 396 nm when sulfur compounds are burned in a hydrogen-rich flame.

The *electrochemical detector* widely used for the control of gas odorization registers the potential change of an electrolytic cell when chromatographically separated sulfur compounds are oxidized by concentrated chromium trioxide solution [406]. Carbonyl sulfide, which is not indicated by this detector in its normal mode of operation, may also be detected when potassium hydroxide solution is used instead of chromium trioxide.

The *Hall detector* measures the change in conductivity caused by absorption of sulfur dioxide in an electrolyte after the effluent of the chromatographic column has been burned with excess oxygen by passing the gas stream through a narrow nickel tube at high temperature [407]. The change in conductivity is directly proportional to the sulfur content of the effluent. The same principle has also been used for monitoring the total sulfur content of fuel gases.

Hydrogen sulfide can be determined by the formation of brown or black stains of lead sulfide on lead acetate paper. This can be a simple semiquantitative test, a manual quantitative determination [408], or a completely automated control [403]. For *spot measurements*, detector tubes [385] may also be sufficient. Wet chemical determinations are based on iodometric titration [409] or the very sensitive Methylene Blue method [410]. The coulometric method uses the oxidation of hydrogen sulfide by electrochemically generated bromine. Interference from other oxidizable sulfur compounds can be eliminated by using a difference method with selective absorption solutions (e.g., $CdSO_4$ for H_2S and KOH for H_2S and mercaptans) [411]. Specific electrochemical cells are used for monitoring the hydrogen sulfide concentration in air or fuel gases.

Carbon disulfide is normally determined by GC [368], but colorimetric methods and detector tubes [385] may also be used.

Besides the aforementioned GC and potentiometric methods, carbonyl sulfide can be hydrolyzed in dilute potassium hydroxide solution. The resulting sulfide is determined colorimetrically by the Methylene Blue method.

Nitrogen Compounds.

Ammonia. Ammonia in manufactured gas is determined by absorption in standard sulfuric acid and titration of the excess acid [412]. *Low concentrations* are indicated by reaction with Nessler's reagent, which is also used for colorimetric determination. For *spot measurements*, gas detector tubes are available [385].

Hydrogen Cyanide. Hydrogen cyanide in coal gas with low hydrogen sulfide content can be determined after absorption in a nickel – sodium carbonate suspension [413] or in an ammoniacal solution of zinc sulfate by titration with silver nitrate. Gases with a higher content of hydrogen sulfide can be analyzed by the cyanogen bromide method [414].

Nitrogen Oxides. Nitrogen dioxide forms a red dye with Saltzman solution [415]. Nitric oxide present in coal gas, therefore, must first be oxidized by passing the gas through a sodium permanganate solution. Both oxides can be determined together, or nitric oxide alone, after selective absorption of nitrogen dioxide.

For the analysis of *flue gases*, the chemiluminescent reaction between nitric oxide and ozone is used for NO_x determination. If nitrogen dioxide is included, the gas must pass through a con-

verter where nitrogen dioxide thermally decomposes to nitric oxide and oxygen [416].

Condensable Hydrocarbons. Natural or manufactured gases may contain amounts of condensable hydrocarbons that can cause problems by forming gas condensates. In dry gas, the *hydrocarbon dew point* may be determined by the mirror method [396]. The potential hydrocarbon liquid content in natural gas can be evaluated either by weighing or volumetrically [417].

Gas benzene (i.e., BTX aromatics and higher boiling hydrocarbons) was formerly determined by adsorption on activated carbon and subsequent displacement by steam. Today, chromatographic methods are preferred.

Naphthalene and other condensed aromatic components can be determined chromatographically or in total by precipitation of the corresponding picrates [418].

8.2.2.4. Determination of Trace Constituents

For the determination of trace constituent (e.g., nickel and iron carbonyl, chlorine [419], and mercury [420]), several methods have been developed, using atomic absorption spectrometry (AAS) to determine heavy metals.

9. Acknowledgement

The entire topic was coordinated by Gerhard Hochgesand.

10. References

1. E. Michel, *GWF Gas Wasserfach* **95** (1954) 598 – 603.
2. H. Zeise: *Thermodynamik* vol. III/1, Leipzig 1954.
3. R. C. Weast: *Handbook of Chemistry and Physics*, CRC Press, Cleveland, Ohio.
4. Nat. Ber. Stand: *Selected Values of Chemical Thermodynamic Properties*, Washington 1952/53.
5. L. B. Evans et al., *Comput. Chem. Eng.* **3** (1979) 319 – 327.
6. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids*, McGraw-Hill, New York 1987.
7. BASF, DRP 296 866 1912 (Mittasch, Schneider).
8. *Nitrogen* **214** (1995) March/April, 38 – 56.
9. M. Appl, *Nitrogen* **100** (1976) March/April, 47 – 58.
   M. Appl, H. Gössling, *Chem. Ztg.* **96** (1972) no. 3, 135 – 153.
10. J. P. van Hook, *Catal. Rev. Sci. Eng.* **21** (1980) no. 1, 1 – 51.
    W. W. Akers, D. P. Camp, *Aiche J.* **1** (1955) no. 4, 471 – 475.
11. A. M. De Groote, G. F. Froment, *Rev. Chem. Eng.* **11** (1995) no. 2, 145 – 183.
12. ICI: *Catalyst Handbook*, Wolfe Scient. Books, London 1970.
13. J. R. Rostrup-Nielsen: *Steam Reforming Catalysts*, Teknisk Forlag, Kopenhagen 1975.
14. F. J. Dent, *Trans. Inst. Gas Eng.* **95** (1945 – 1946) 602 – 709.
15. T. Nicklin, R. J. Whittaker, *Inst. Gas Eng.* **8** (1968) no. 1, 15 – 39.
16. D. D. Wagmann et al., *J. Res. Nat. Bur. Stand. (U.S.)* **34** (1945), Feb. 143, 161.
17. G. W. Bridger, *Chem. Process Eng. (London)* **53** (1972) 38.
18. J. R. Rostrup-Nielsen, P. E. Højlund Nielsen: *Deactivation and Poisoning of Catalysts*, Marcel Decker, New York-Basel (1985) pp. 259 – 323.
19. J. R. Rostrup-Nielsen, J. Wrisberg: "Steam Reforming of Natural Gas at Critical Conditions," *Natural Gas Processing and Utilisation Conference*, Dublin 1976.
20. Supplement to API RP 530, Am. Petrol. Inst., Washington 1973.
21. H. F. Niessen, K. Kugeler, M. Kugler, *Rev. Int. 4. Temp. Réfract.* 1976 (13) 204 – 211.
22. F. Pons, J. Thuillier: *Life Conditions of Reformer Tubes*, Acieries du Manoir Pompey 1975.
23. *TOPSOE-topics* 1993 Feb.
24. R. G. Minet, O. Olesen: "Technical and Economic Advances in Steam Reforming of Hydrocarbons." *A. C. S. Symp. Ser.* **116** (1980) 160 – 174.
25. F. W. Hohmann, *Hydrocarbon Process.* **75** (1996) no. 3, 71 – 74.
26. R. L. Stephenson,
    R. M. Smailer (eds.): *Direct Reduced Iron*, copyr. by the Iron and Steel Society of AIME, Warendale 1980, pp. 64 – 95.

27. Programmstudie, *Auf dem Weg zu neuen Energiesystemen*, Bundesministerium für Forschung und Technologie, Bonn, Teil I bis VI 1975.
28. St. H. Fistedis, T. B. Belytschko, K. Kussmaul, R. T. Lahey jr. et al. (eds.), *Nuclear Engin. and Design, Special issue on High-Temperature Reactor and Nuclear Process Heat Application*, vol. 78, no. 2, North-Holland Physics Publishing, Amsterdam 1984.
29. L. Pott, L. Langfield, T. Eddlestone, *Inst. Gas Eng.* **9** (1969) no. 2, 73 – 82.
30. G. Kurz, S. Teuner, *Erdoel Kohle Erdgas Petrochem.* **43** (1990) no. 5, 171 – 172.
31. S. Teuner: "Make CO from CO_2," *Hydrocarbon Process.* **53** (1985) no. 5, 106 – 107.
32. D. J. Ward, *Hydrocarbon Process.* **53** (1974) no. 4, 85 – 88.
33. T. Tomita, M. Kitagawa: "A New Steam Reforming Process of Heavy Hydrocarbons," *Lecture at the Achema 1976*, Frankfurt, Federal Republic of Germany.
34. B. J. Cromarty, B. J. Crewdson, *Nitrogen* **191** (1991) May/June, 30 – 34.
35. R. G. Cockerham, G. Percival, *Trans. Inst. Gas Eng.* **107** (1957/58) 390 – 433.
36. H. S. Davies, K. J. Humphries, D. Hebden, D. A. Perey, *Inst. Gas Eng. J.* **7** (1967) no. 10, 707 – 736.
37. G. Baron, H. Hiller, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) no. 3, 196 – 200.
38. T. Ishiguro, *Hydrocarbon Process.* **47** (1968) no. 2, 87 – 91.
39. P. W. Farnell, *Fertiliser Industry* (1996) 53 – 58.
40. L. W. ter Haar, *Ind. Chem. Belge* **33** (1968) no. 7 – 8, 655 – 668.
41. S. Strelzoff, *Hydrocarbon Process.* **53** (1974, Dec.) 79 – 87.
42. H. Fricke, Chem. Ztg. **96** (1972) no. 3, 123 – 134.
43. D. Das, W. Liebner: "Optimizing/Costing Study for a 500 MW IGCC Power Plant based on the Lurgi MPG Process", mg engineering transactions, E 1704, 1998, June.
44. B. Buttker, H. Hirschfelder, G. Steiner: "Concept and Realization of the Schwarze Pumpe FRG *Waste to Energy and Chemicals Centre*", Paper presented at IChemE conference, Milan, Italy, 1997, Feb.
45. L. O. Koenders, S. A. Posthuma, P. L. Zuideveld: "The Shell Gasification Process for Conversion of Heavy Residues to Hydrogen and Power", Paper presented at EPRI/GTC conference, San Francisco, CA, 1996, Oct.
46. E. S. Pettyjohm, H. R. Linden, W. G. Blair, *Gas Age* **113** (1954) no. 10, 19 – 26, 68, 70.
47. R. W. Tittle, W. Hartley, SNG Symposium I, March 12 – 16, Chicago 1973.
48. J. F. McMahon, SNG Symposium I, March 12 – 16, 1973, Chicago.
49. *Ullmann*, 5th ed., **vol. A 12**, pp. 213 – 214.
50. J. L. Johnson: "Fundamentals of Coal Gasification" in M. A. Elliott (ed.): *Chemistry of Coal Utilization*, 2nd Suppl., New York 1981, pp. 1491 – 1598.
51. J. B. Howard: "Fundamentals of Coal Pyrolysis and Hydropyrolysis" in Elliott M. A. (ed.): *Chemistry of Coal Utilization*, 2nd Suppl. vol., New York 1981, pp. 665 – 784.
52. K. Hedden, *Ullmann* 3rd ed., **10**, 362 – 376.
53. D. R. Simbeck, R. L. Dickenson, E. D. Oliver, EPRI-Report AP-3109, Research Project 2207, 1983.
54. B. Miller, EPRI-Report AP-3101, Research Project 986–9, 1983.
55. D. Hebden, H. J. F. Stroud: "Coal Gasification Processes" in M. A. Elliott (ed.): *Chemistry of Coal Utilization*, 2nd Suppl., New York 1981, pp. 1599 – 1752.
56. O. Nowacki (ed.): *Coal Gasification Processes, Energy Technology Review No. 70*, Noyes Data Corp., Park Ridge 1981.
57. L. G. Massey: "Coal Gasification for High and Low Btu Fuels" in C. Y. Wen, E. S. Lee (eds.): *Coal Conversion Technology*, London 1979, pp. 313 – 427.
58. H. W. v. Gratkowski, *Ullmann* 3rd ed. **10**, 376 – 455.
59. Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
60. H. V. Bridgwater (ed.): *Thermochemical Processing of Biomass*, Butterworths London 1984.
61. C. G. van Fredersdorff, M. A. Elliott: "Coal Gasification" in H. H. Lowry (ed.): *Chemistry of Coal Utilization*, Suppl. vol., New York 1963 pp. 892 – 1022.
62. H. Jüntgen, K. H. van Heek, *Fuel Process. Technol.* **2** (1979) 261 – 293.
63. E. M. Suuberg, W. A. Peters, J. B. Howard, *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) 37 – 46.
64. G. D. Sergent, J. W. Smith, *Fuel* **52** (1973) 52 – 57.
65. M. A. Gibson, C. A. Euker Jr., Paper presented at AIChE Meeting, Los Angeles, CA Nov. 1975.

66. H. Yoon, J. Wei, M. M. Denn, *AIChE J.* **24** (1978) 885 – 903.
67. M. A. Serio, W. A. Peters, K. Sawada, J. B. Howard, *Prepr. Am. Chem. Soc. Div. Fuel Chem.* **29** (1984) no. 2, 65 – 76.
68. O. Levenspiel: *Chemical Reaction Engineering*, John Wiley and Sons, New York, London 1972.
69. K. K. Neumann, F. Keil, *Brennst. Wärme Kraft* **33** (1981) 22 – 25.
70. W. Gumz, *Vergasung fester Brennstoffe*, Springer Verlag, Berlin 1952.
71. D. E. Woodmansee, *Energy Commun.* **2** (1976) 13 – 44.
72. W. Mengis, S. Schulz, *Chem. Ing. Tech.* **56** (1984) 254 – 255.
73. H. S. Caram, N. R. Amundson, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) 80 – 102.
74. H. Kubiak, PhD Thesis, University of Essen 1981.
75. A. J. Lupa, H. C. Kliesch, EPRI Report AF-1179 vol. 1, Res. Proj. 1037–1, 1979.
76. R. Govind, J. Shah, *AIChE J.* **30** (1984) 79 – 92.
77. O. P. Mahajan, P. L. Walker: "Reactivity of Heat-Treated Coals" in Clarence Karr jr. (ed.): *Analytical Methods for Coal and Coal Products*, vol. 2 Academic Press, New York 1978, pp. 465 – 495.
78. "Gas Processing Handbook," *Hydrocarbon Process.* **58** (1979, April) 99 – 170; **61** (1982, April) 85 – 164; **63** (1984, April) 51 – 114.
79. D. Hebden, *7th Synthetic Pipeline Gas Symposium* Chicago, Ill. (Oct. 1975).
80. W. Seifert, B. Buttker, M. D. de Souza, H. Vierrath: "Clean Waste Recycle — the BGL Route at Schwarze Pumpe", in Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
81. J. P. Brown, P. Pietrogrande, J. Bauer, W. Kaimann, *Energy Prog.* **5** (1985) 9 – 14.
82. G. T. Peterson, *3rd EPRI Contractors Conf. on Coal Gasification*, July 1984, EPRI-AP-3586 CONF-831199, 4.1 – 4.8.
83. P. Rudolph, *4th Synthetic Pipeline Gas Symposium* Chicago, Ill. (1972).
84. J. M. Coffin, *Energy Prog* **4** (1984) 131 – 137.
85. F. C. Schora, J. G. Patel, *Petrole Informations* (1982) 33 – 38 (March).
86. G. Häßler: "Thermoselect — Der neue Weg, Restmüll umweltgerecht zu behandeln", *Conf. Proc. THERMOSELECT, Rastatt*, 1995, Jan. p. 71, 74.
87. K. Hedden, Th. Heike, B. R. Rao, *GWF Gas Wasserfach Gas Erdgas* **123** (1982) 482 – 495.
88. J. E. Scott, EPRI Report AP-1922, Research Project 1267–1, August 1981.
89. G. Schaub, R. Reimert, K. Janssen, P. R. Rodriguez, *3rd International Conference on Heavy Crude and Tar Sands*, Long Beach USA, July 1985.
90. M. J. Mujadin, P. K. Herbert, J. Loeffler, *Coal Gasification and Synthetic Fuels for Power Generation Conference*, San Francisco, USA (April 1984).
91. W. R. Miller, R. A. Lang, Report DOE/CH/10088–1874, vol. 1, (July 1985).
92. U. Dorstewitz, H. Wien, *Conference Coal Gasification in Energy Technology*, Dortmund, FR Germany, March 1979.
93. W. Schäfer, G. Heinrich, C. Lohmann, H. P. Peyrer, *Erdöl Kohle, Erdgas Petrochem.* **36** (1983) 557 – 562.
94. W. Bierter, Ch. Gaegauf: *Holzvergasung*, C. F. Müller Verlag, Karlsruhe, FR Germany 1982.
95. L. Reh, *Ger. Chem. Eng. (Engl. Transl.)* **1** (1978) 319 – 329.
96. K. H. van Heek, H. Jüntgen, W. Peters, *Atomenerg. Kerntech.* **40** (1982) 225 – 245.
97. R. Kirchhoff, K. H. van Heek, H. Jüntgen, W. Peters, *Nuc. Eng. Des.* **78** (1984) 233 – 239.
98. K. Stahl, M. Neergard: "Das Kombi-Kraftwerk mit integrierter Biomassen-Vergasung in Vörnamo/Schweden", *VGB Kraftwerkstechnik* **76** (1996) 327 – 330.
99. M. Yaqub Chughtai, H.-J. Mühlen: "Gas-, Strom- und Wärmeerzeugung aus Biomasse durch allotherme Vergasung", in Energetische und stoffliche Nutzung von Abfällen und nachwachsenden Rohstoffen, *DGMK Conf. Proc.*, 1998, April.
100. L. Schrader, H. Teggers, K. A. Theis, *Chem. Ing. Tech.* **52** (1980) 794 – 802.
101. R. Fladerer, L. Schrader, *Chem. Ing. Tech.* **54** (1982) 884 – 892.
102. Personal Communication from Rheinbraun Aktiengesellschaft, March 1999.
103. J. Albrecht, B. Gafron, P. Scur, R. Wirthwein: "Vergasung von Sekundärbrennstoffen in der Zirkulierenden Wirbelschicht zur Energetischen Nutzung für die Zementherstellung", in Energetische und stoffliche Nutzung von Abfällen und nachwachsenden Rohstoffen, *DGMK Conf. Proc.* 1998, April.
104. C. Greil, H. Hirschfelder: "Brenngas aus Biomasse — Innovative Anlagenkonzepte auf Basis der Zirkulierenden

Wirbelschichtvergasung", in Energetische und stoffliche Nutzung von Abfällen und nachwachsenden Rohstoffen, *DGMK Conf. Proc.* 1998, April.
105. B. G. Bryan: "Status of the Shanghai Coking and Chemical Company's U-Gas Coal Gasification Plant", *Ann. Int. Pittsburgh Coal Conf. 15th, Pittsburgh*, 1998, Sept.
106. C. A. Euker, R. D. Wesselhoft, *Energy Prog.* **1** (1981) 12 – 16.
107. R. A. Cardello, R. A. Reitz, F. C. R. M. Smits, *11th World Petroleum Congress*, London, August 1983.
108. K. A. Theis, E. Nitschke, *Hydrocarbon Process.* **61** (1982) 233 – 237. Uhde-publication CK J 5 19 2000 82 Rheinbraun-HTW-Verfahren.
109. G. Wargalla, G. Daradimos, BMFT Forschungsbericht T 85–080, August 1985.
110. W. Flesch, G. Velling, *Erdöl Kohle, Erdgas, Petrochem.* **15** (1962) 710 – 713.
111. J. M. Lemasle, in [60]
112. H. H. Lienhard, P. Mehrling, Fifth EPRI Contractors Conf. on Coal Gasification October 1985.
113. W. H. Blackadder, E. Rensfelt, in [60]
114. E. D. Oliver, EPRI Report AP-2567, Research Project 986–10, August 1982.
115. W. Zerbin in J. Gehrmann (ed.): *Thermochemische Gaserzeugung aus Biomasse – Vergasung und Entgasung*, Berichte der Kernforschungsanlage Jülich, Jül-Conf-46, 1981, pp. 102 – 121.
116. A. Schweizer, P. K. Herbert, J. C. Löffler: "Heizgas aus Baumrinde zum Betrieb eines Kalkdrehofens in einem Sulfat-Zellstoffwerk", *Zement-Kalk-Gips* **44** (1991) 217 – 220.
117. G. Pabst, *Bergbau-Rohstoffe-Energie* **23** (1984) 263 – 274.
118. Ahlström-Wärtsila Publication 5949/1983, "Diesel Power from Biomass."
119. Carbon Gas Technology Publication "CGT Kohlevergasung."
120. H. Schmidt-Traub, H. Ch. Pohl, *Chem. Ing. Tech.* **55** (1983) 850 – 855.
121. G. R. Mc Cullough, S. C. Roberts, M. J. van der Burgt, *Energy Prog.* **2** (1982) 69 – 72.
122. M. J. van der Burgt, J. E. Naber, *3rd BOC Priestley Conference*, London, September 1983.
123. P. Amick: "Altemate Fuel Testing at the Wabash River Coal Gasification Repowering Project", in Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
124. W. Schellberg: "First Results of the PRENFLO based IGCC Power Plant in Puertollano, Spain", in Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
125. J. N. Phillips, M. B. Kiszka, U. Mahagaokar, A. B. Krewinghaus: "Shell Coal Gasification Project Final Report on Eighteen Diverse Feeds", EPRI TR-100687, 1992.
126. W. G. Schlinger, J. H. Kolaian, M. E. Quintana, T. G. Dorawala, *Energy Prog.* **5** (1985) 234 – 238.
127. J. H. Kolaian, W. G. Schlinger, Energy Prog. **2** (1982) 228 – 233.
128. P. Ruprecht, R. Dürrfeld, EPRI Report AP-2607, Research Project 1799–1, Sept. 1982.
129. P. Curran: "Feedstock Versatility for Texaco Gasification", in Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
130. Personal Communication from Krupp-Uhde GmbH, April 1999.
131. DD 161 080, DD 240 571, June 1985. H. Brandt, M. Schingnitz, F. Berger, "The GSP Process-Developments in the Field of Pressure Gasification of Pulverized Brown Coal in the GDR," *16th World Gas Conference*, Munich 1985.
132. G. Henrich et al., *Chem. Tech. (Heidelberg)* **10** (1984) 45 – 53.
133. A. L. Kohl, J. A. Ashworth, *Hydrocarbon Process.* **62** (1983, July) 97 – 100.
134. W. Beckh, BMFT-Report FB-T-85–105, 1985.
135. T. F. Edgar, D. W. Gregg: "Underground Gasification of Coal" in M. A. Elliott (ed.): *Chemistry of Coal Utilization*, 2nd Suppl., New York 1981, pp. 1801 – 1843.
136. P. Ledent, T. K. Li, V. Chandelle, V. Fabry, M. Kurth, C. Sonntag, *Proc. Int. Gas Res. Conf., London* 1983, pp. 539 – 550.
137. W. B. Krantz, R. D. Gunn (eds.), *AIChE Symp. Ser.* **226** (1983)
138. J. L. Conchello, P. Fievez, A. Obis: "Underground Coal Gasification Under European Conditions", in Gasification — The Gateway to a Cleaner Future, *Chem. E. Conf. Proc., Dresden* 1998, Sept.
139. A. V. Russel, J. G. Slack: *Ammonia; Part 2*, Marcel Dekker Inc., New York 1974.
140. A. L. Kohl, R. Nielsen: *Gas Purification*, 5th ed., Gulf Pub. Co., Houston 1997.
141. J. Carstensen, *Hydrocarbon Process.* (1998) March, 100-C ff.

142. H. Roos et al., *Ammonia Plant and Related Facilities Safety*, AIChE, New York 1990.
143. H. Short, *Chem. Eng.* (1989) July, 41 ff.
144. J. Carstensen et al.: *Ammonia Plant and Related Facilities Safety*, AIChE, New York 1991.
145. A. L. Kohl, F. C. Riesenfeld: *Gas Purification*, 4th ed., Gulf Pub. Co., Houston 1985.
146. Y. Traa, J. Weitkamp, *Chem. Ing. Tech.* **70** (1998) 1428 – 1430.
147. D. D. Wagman, J. E. Kilpatrick, W. J. Taylor, K. S. Pitzer, F. D. Rossini, *J. Res. Nat. Bur. Stand (U.S.)* **34** (1945) 143 – 161. N.B.S. Research Paper R P 1634.
148. G. A. Mills, F. W. Steffgen, *Catal. Rev.* **8** (1973) 159 – 210.
149. A. L. Lee: *Methanation for Coal Gasification*, Inst. Gas Technol., Chicago.
150. P. Schoubye, *J. Catal.* **14** (1969) 238 – 246; **18**, (1970) 118 ff.
151. Bundesarbeitsblatt, Fachbeilage *Arbeitsschutz*, no. 10 (1976) pp. 362 – 371.
152. D. W. Allen, W. H. Yen, *Chem. Eng. Prog.* **69** (1973) no. 1, 75 – 79.
153. F. W. Möller, H. Roberts, B. Britz, *Hydrocarbon Process.* **53** (1974) no. 4, 69 – 74.
154. K.-H. Eisenlohr, F. W. Möller, M. Dry, *Adv. Chem. Ser.* **146** (1975) 113 – 122. *Pet. Times* **19** (1973, Okt.) 17.
155. J. Range, H. G. Harms, *Chem. Tech. (Heidelberg)* **11** (1982) 894 – 899.
156. G. A. White, T. R. Roszkowski, D. W. Standbridge, *Adv. Chem. Ser.* **146** (1975) 138 – 148.
157. Gregor Friedrichs, P. Proplesch, W. Lommerzheim, *Gas Wärme Int.* **31** (1982) 261 – 254.
158. Kernforschungsanlage Jülich GmbH, US 4 431 751, 1982 (B. Höhlein, M. Vorwerck, U. Boltendahl).
159. H. Harms, B. Höhlein, E. Jorn, A. Skov, *Oil Gas J.* **78** (1980, April 14) 120 – 135.
160. Bundesministerium für Forschung und Entwicklung: Studie, *Auf dem Wege zu neuen Energiesystemen*, vol. IV, Bonn 1975.
161. N. Sandler, *Chem. Eng (N.Y.)* **92** (1985) 16, 12 K.
162. D. B. Blum, M. B. Sherwin, M. E. Frank, *Adv. Chem. Ser.* **146** (1975) 149 – 159.
163. T. K. Sherwood, R. L. Pigford, Ch. R. Wilke: *Mass Transfer*, McGraw-Hill, New York 1975.
164. P. V. Danckwerts: *Gas-Liquid Reactions*, McGraw-Hill, New York 1970.
165. G. D. Astarita, D. W. Savage, A. Bisio: *Gas Treating with Chemical Solvents*, J. Wiley and Sons, New York 1983.
166. R. S. Smith: "Improve Economics of Acid Gas Treatment," *Oil Gas J.* **73** (1975, March 10) 78 – 82.
167. A. Kobus, H.-J. Roehm: "Ein heuristisch-numerischer Ansatz zum systematischen Entwurf und Design von Absorptionsverfahren," *Chem. Ing. Tech.* **71** (1999) no. 8, 849 – 852.
168. A. Kohl, R. Nielsen: *Gas Purification*, 5th ed., Gulf Publishing Company, Houston, Texas 1997.
169. D. K. Fleming, H. S. Primack: "Purification Processes for Coal Gasification," *Chem. Eng. Prog.* **3** (1977) 66.
170. St. A. Newman: *Acid and Sour Gas Treating Processes*, Gulf Publ Co., Houston 1985.
171. F. C. Vidaurri, L. C. Kahre: "Recover H_2S Selectively from Sour Gas Streams," *Hydrocarbon Process.* **56** (1977) no. 11, 333 – 337.
172. R. N. Tennyson, R. P. Schaaf: "Guidelines can help choose proper process for gas treating plants," *Oil Gas J.* **75**, (1977, Jan. 10) 78 – 86.
173. G. Ranke, *Linde Ber. Tech. Wiss.* **33** (1973) 6 – 12.
174. *Rectisol for Gas Treating*, Lurgi brochure, Frankfurt, Germany.
175. G. Hochgesand, *Ind. Eng. Chem.* **62** (1970) no. 7, 37 – 43.
176. U. Zwiefelhofer, H.-D. Holtmann, "Energierückgewinnung mit Entspannungsmaschinen", Paper presented at VDI Conference, Magdeburg 1996, Reprint in brochure of Lurgi Oel Gas Chemie GmbH.
177. R. W. Bucklin, R. L. Schendel: "Comparison of Fluor Solvent and Selexol Processes," *Energy Prog.* **4** (1984) no. 3, 137 – 142.
178. K. Wehner, B. Schlicht, P. Minok, W. Burk "Das Leuna-Geselan Verfahren, ein neuer technischer Gasentschwefelungsprozeß," *Chem. Tech.* (*Leipzig*) **34** (1982) no. 6, 291 – 294.
179. J. E. Johnson, "Selexol Solvent reduces lean, high CO_2 Natural Gas Treating Costs," *Energy Prog.* **4** (1984) no. 4, 241 – 248.
180. *Gas Conditioning Fact Book*, The Dow Chemical Company, Midland Michigan 1962.
181. M. Gross, J. Menzel: "Acid gas removal for upgrading natural and synthesis gas," *Hydrocarbon Eng.* (1997) no. 11, 72.
182. M. Gross, J. Menzel, O. Tondorf: "Acid Gas Removal from Natural and Synthesis Gas," *Hydrocarbon Eng.* (1998) no. 6, 30.

183. J. A. Lagas: "Acid Gas Removal with Selective Amine Process," *Sulphur* **159** (1982) 4, 35 – 39.
184. S. R. Hays, C. J. Schultz: "Corrosion and Fouling of refinery MDEA system", NACE Corrosion 92, paper no. 447.
185. S. M. Mecum, F. C. Veatch, A. L. Cummings: "Why caustic addition is bad for amine systems," *Hydrocarbon Process.* (1997) Oct., 115 – 118.
186. M. Huval, H. Van De Venne: "DGA Proves Out as a Low Pressure Gas Sweetener," *Oil Gas J.* (1981) Aug. 17, 91 – 99.
187. C. J. Wendt, L. W. Dailer, *Hydrocarbon Process.* **46** (1967) no. 10, 155 – 157.
188. E. C. Heyman, P. F. A. van Grinsen, "Shell Regenerable Acid Gas Removal Processes", Brochure of Shell Global Solutions.
189. C. Blanc, J. Elgue: "MDEA process selects H_2S," *Hydrocarbon Process.* **60** (1981, Aug.) 111 – 116.
190. M. L. Spears, K. M. Hagan, J. A. Bullin, C. J. Michalik: "Converting to DEA/MDEA Mix Ups Sweetening Capacity," *Oil Gas J.* (1996) Aug. 12, 63 – 67.
191. P. Bauvin: "The Elf activated MDEA process: New developments and industrial results", International Gas Research Conference, 1995, Reprint of Elf Aquitaine.
192. A. M. Goldstein, A. M. Edelmann, W. D. Beisner, P. A. Rusizka: "Hindered amines yield improved gas treating," *Oil Gas J.* **82** (1984) 70 – 76.
193. G. R. Chludzinski, S. Wiecert: "Commercial experience with Flexsorb SE absorbents", reprint of paper 58e, AIChE Meeting, 1986 April.
194. G. Sartori, D. W. Savage, *Ind. Engng. Chem. Fundam.* **22** (1983) 239 – 249.
195. "Selective H_2S removal with Flexsorb SE Solvents", brochure of Exxon Research and Engineering Company 1997.
196. H. E. Benson, J. H. Field, R. M. Jimeson, *Chem. Eng. Prog.* **50** (1954) 356.
197. C. L. Dunn, E. R. Freitas, E. S. Hill: "First Plant Data from Sulfinol Process," *Hydrocarbon Process.* **44** (1965) no. 4, 137 – 141.
198. K. Bratzler, A. Doerges, *Hydrocarbon Process.* **53** (1974) no. 4, 78 – 80.
199. M. Kriebel, "Improved Amisol Process for Gas Purification," *Energy Prog.* **4** (1984) no. 3, 143 – 146.
200. F. Giammarco, IT 537564, 1955; 560161, 1956; 56320, 1957.
201. B. M. Wilson, R. D. Newell, "H_2S-Removal by Stretford Process," *Chem. Eng. Prog.* **10** (1984, Oct) 40 – 47.
202. B. G. Goar, *Oil Gas J.* **69** (1971, July 19) 84 – 86.
203. L. C. Hardison, *Chem. Eng.* **21** (1985) 62 – 63.
204. "New equipment for the SulFerox process," *Sulphur* **232** (1994) May, 42 – 45.
205. H. Mackinger, F. Rosatti, G. Schmidt: "Sulfint Process," *Hydrocarbon Process.* **61** (1982, March) 169 – 172.
206. R. N. Maddox, M. D. Burns, *Oil Gas J.* (1984, March 13) 100 – 103.
207. Y. Barthal, A. Dechamps, C. Dezael, H. Gruier: "The IFP Clauspol 1500 process for sulfur recovery plant tail gas treatment," *Proceedings Sulphur '82* (1982) Vol. 1, 323 – 335.
208. A. Janssen, R. Sleyster, C. van der Kaa, A. Jochemsen, J. Bontsema, G. Lettinga: "Biological oxidation an a fed-batch reactor," *Biotechnol. Bioeng.* **47** (1995) 327 – 333.
209. M. P. Heisel, F. J. Marold: "New gas scrubber removes H_2S," *Hydrocarbon Process.* (1987) April, 35 – 37.
210. H. W. Gowdy, M. R. Anderson: "The Commercialization of the Unisulf Process," Paper presented at the 1986 Stretford User's Conference, Austin, Texas, 1986.
211. Institut Français du Petrol: "Ifpexol," *Hydrocarbon Process.* (1996) April, 124.
212. P. F. A. van Grinsen, G. J. v. Heeringen, E. C. Heyman, M. L. Dillon: "DIPA as the preferred solvent for amine treatment in refinery applications," Laurence Reid Gas Conditioning Conference 1999.
213. M. P. Quinlam, L. W. Echterhoff: "Technical and economic comparison of Lo-Cat II with other iron based liquid redox processes", Presented at 1992 GRI Liquid Redox Sulfur Recovery Conference, Austin Tex., GRI Contract No. 5088–221–1753.
214. E. Richter, K.-B. Harder, K. Knoblauch and H. Jüntgen: "Neue Entwicklungen zur Druckwechsel-Adsorption," *Chem. Ing. Tech.* **56** (1984) no. 9, 684 – 691.
215. M. Streich: "Die Konzentrierung von Wasserstoff durch Tieftemperaturverfahren" *Dechema Monographien* vol **58**, 195 – 204, Verlag Chemie Weinheim, 1968.
216. W. Förg: "Herstellung von Kohlenmonoxid und Wasserstoff hoher Reinheit als Zwischenprodukt für die Polymethanchemie auf Basis Erdgas," *GWF Gas Wasserfach Gas/Erdgas* **113** (1973), no. 11, 538.

217. H. Jockel, E. Supp: "Erdgas als Rohstoff für Me-thanol, Oxoalkohol und Kohlenmonoxid," *Gas Wärme Int.* **19** (1970) no. 6, 213.
218. K. Storck: "Neue Absorptionsverfahren für die Tieftemperaturtechnik," *Kältetechnik* **14** (1962) 274 – 276.
219. M. Streich, E. Kistenmacher, "Einfluß der Ungenauigkeit physikalischer Stoffwerte auf die Auslegung von Tieftemperaturanlagen," *Linde Ber. Tech. Wiss.* **44**, (1978).
220. A. M. Robinson: "Neue Möglichkeiten in der industriellen Gastrennung," *Chemie-Technik Sonderdruck* **10** (1981) 399 – 401.
221. D. J. Stookey, T. E. Graham, W. A. Pope: "Natural Gas Processing with Prism Separators," *AICE 1984 Winter National Meeting*, Atlanta.
222. C. S. Goddin: "Pick treatment for High CO2 Removal", *Hydrocarbon Process.* **62** (1983, May) 125 – 130.
223. D. L. MacLean, Y. C. Chae, E. C. Prince: "Energy Saving Modification in Ammonia Plants," *Curr. Lab. Pract.* **12** (1980) no. 3, 98 – 104.
224. A. E. Maciula: *Oil Ges J.* **78** (1980) 63.
225. H. Knieriem Jr., "Membrane Separation saves energy," *Hydrocarbon Process.* **60** (1980, July) 65 – 67.
226. L. E. Applegate: "Membrane separation processes," *Chem. Eng.* (N.Y.) 91 (1984, June 11) 66 – 88.
227. W. Herbert, *öl Kohle* **38** (1942) 525 – 531.
228. G. N. Richter, W. G. Schlinger: *Proc. Conf. Air Qual. Management Electr. Power Ind.*, 2nd ed., vol. **2**, 1980 pp. 1571 – 85; *Chem. Abstr.* **97** (1982) 97788 b.
229. J. F. Clausen, C. A. Zee, *DOE Symp. Ser.* **54** (1981) 558 – 575.
230. M. M. El-Halwagi, V. Manousiouthakis: "Automatic (analytic) synthesis of mass-exchange networks (MEN) single-component targets," *Chem. Eng. Sci.* **45** (1990) no. 9, 2813 – 2831.
231. F. Weyrich, B. Rumpf, G. Maurer: "Untersuchungen zum kalorischen Verhalten wässriger Lösungen, die Ammoniak, saure Gase und starke Elektrolyte enthalten," GVC Annual Meeting (Dresden, Germany 9/24–26/97) Paper no.169; *(Abstract) Chem. Ing. Tech.* **69** (1997) no. 9, 1325.
232. W. Seifert, B. Buttker, M. D. de Souza, H. Vierrath: "Clean Waste Recycle — the BGL Route at Schwarze Pumpe IChemE Dechema VDI VCA, Symposium," Gasification — The Gateway to a Cleaner Future, Dresden, Sept. 23 – 24, 1998.
233. M. S. Nie, B. Z. Wang, Y. G. Zhao: "Reuse of pretreated coal gasification waste — water as cooling tower makeup," ACS 206th National Meeting (Chicago 8/22 – 26/93) ACS Division of Fuel Chemistry Preprints **38** (1993) no. 3, 945 – 952.
234. B. H. Cooke, M. R. Taylor: "The environmental benefit of coal gasification using the BGL [(British Gas-Lurgi)] gasifier," British Gas Conference Newcastle upon Tyne 3/31–4/2/93); *Fuel* **72** (1993) no. 3, 305 – 314.
235. D. Klockner, *Tech. Mitt.* **77** (1984) no. 5, 238 – 243. Heurtey Efflutherm, BE 838 608, 1976 (R. Wang, D. Brun).
236. G. G. Mayer, C. D. Turner, T. A. Moe, C. A. Wentz: "Innovative Use of Cooling Tower Technology for Waste Reduction," Papers Presented at the Summer National Meeting of the American Institute of Chemical Engineers. Minneapolis, MN, USA, Aug. 16 – 19, 1987; Published by AIChE, New York, NY, USA, Paper 41f, 12p.
237. Environmental Protection Agency (Washington, DC, USA). Fed. Regist. 47, 18 Oct (1982) 201, 4634 – 57.
238. J. C. Lamb, D. B. Bull, *J. Water Pollut. Control Fed.* **57** (1985) no. 10, 930 – 998.
239. *Ullmann*, 5th ed., **A12**, 276, Fig. 65.
240. Bataafse Internat. Petrol Maatschappij: Deoiling Ind. Wastewater, Design and Operation of the "CPI". MF Report no. 68.700.
241. H. J. Wurm, *Gewässerschutz, Wasser, Abwasser* **60** (1982) 429 – 441.
242. F. Wöhler, *Spec. Publ. Brit. Carbonization Res. Assoc.* **24** (1979) 46 – 51; *Chem. Abstr.* **94: 108603 d** (1980).
243. D. Mackay, M. Medir: "*Industrial Effluent Treatment*" in T. C. Lo, M. H. I. Baird, C. Hanson (ed.): Handbook of Solvent Extractioin, Wiley-Interscience, New York 1983. pp. 619.
244. D. Rolke, *Tech. Mitt.* **77** (1984) no. 11, 511 – 513.
245. H. Juentgen, J. Klein, *Energy Sources* **2** (1976) no. 4, 311 – 30.
246. *Ullmann*, 5th ed., **A12**, 279, Fig. 67.
247. Metallgesellschaft Aktiengesellschaft, US 5 716 524, 1998 (H.-M. Stönner, M. Linicus, G. Ungar).
248. Critical Fluid Systems, Inc., an Arthur D. Little Co., Cambridge, Mass., *Chem. Eng.* (1985, June 24) 27.

249. S. Gupta, D. Ghonasgi, K. M. Dooley, F. C. Knopf: "Supercritical carbon dioxide extraction of a phenolic mixture from an aqueous waste stream," *J. Supercrit. Fluids* **4** (1991) no. 3, 181 – 185.
250. D. Glassmann, (U.S. Steel Corp.) *Proc. Ironmaking Conf.* **35** (1976) 121 – 33.
251. NKK Corporation, US 5 139 756, 1992 (T. Shgikada, M. Asanuma, Y. Tachibana).
252. Metallges., DE 22 29 213, 1972 (P. Wiesner, F. Wöhler, H. Stönner).
253. Metallges., DE 25 10 213, 1975 (H. Fischer).
254. R. D. Rice, J. V. Busa: "Recovering NH_3 by the Phosam-W Process," *Chem. Eng. Proc.* (1984, Oct) 61 – 63.
255. A. Schmidt, *Chem. Ing. Tech.* **42** (1972) 521 – 523.
Metallges., DE 25 27 985, 1975, US 4060591, 1975 (A. Garber, H. Stönner, P. Wiesner, A. Sinclair, A. Schmidt),
DE 35 20 934, 1985 (W. Hilsebein, G. Ungar, H. Stönner).
256. Metallgesellschaft Aktiengesellschaft, EP 0 654 443 B1, 1993/1999 (G. Ungar, H.-M. Stönner, J. Schlauer).
257. Metallgesellschaft AG, US 6 558 452, 1999, Priority: DE 198 48 499, 1998 (H.-M. Stönner, G. Ungar, M. Linicus)
258. R. D. Doctor, K. E. Wilzbach, J. B. Levenson: "Sulfur emissions reduction at the Great Plains(GP) Gasification Project (at Beulah, ND): A case for new technologies," 1988 AIChE Summer Natl. Meet. (Denver 8/21 – 24/88) Prepr. N.78f 22P.
259. A. H. Kuhn, G. G. Baker: "Pilot plant and commercial-size testing to resolve SO_2 emission problem at Great Plains," Conference: Opportunities in the synfuels industry: SynOps 88, Bismarck, ND, Aug. 29 – 31 1988, pp. 189 – 201.
260. C. L. Kimtantas: "Here are ways to increase sulfur processing capability," *Oil Gas J.* **87** (1989) no. 21, 58 – 64.
261. S. H. Chou, S. Y. Huang, S. C. Wu, T. K. Chuang, L. Y. Chen: "Treatment of sour water," *Shih Yu (Taipei)* **24** (1988) no. 4, 197 – 202.
262. W. W. Eckenfelder, J. Patoczka, A. T. Watkin, *Chem. Eng. (N.Y)* **92** (1985, Sept. 2) 60 – 74.
263. R. G. Luthy, D. J. Sekel, J. T. Tallon, *J. Environ. Eng. Div. (Am. Soc. Civ. Eng.)* **106** (1980) EE 3, 609 – 29.
264. G. Santer (Stahlwerke Röchling, Ger.), *Circ. Inf. Tech. Cent. Doc. Sider.* **33** (1976) no. 2, 273 – 283.
265. C. G. Grieves, M. K. Stenstrom, J. D. Walk, J. F. Grutsch, *Hydrocarbon Process.* **56** (1977) no. 10, Sect. 1 125 – 130.
266. R. Fisher, *Inst. Chem. Eng. Symp. Ser.* **77** (1983) 163 – 182.
267. T. L. Randall et al., *Environ. Progr.* **3** (1984) no. 4, 237 – 247.
268. P. M. Fedorak, E. Knettig, S. E. Hrudey, *Environ. Technol. Lett.* **6** (1985) no. 5, 181 – 188.
269. K. A. Khan, M. T. Suidan, W. H. Cross, *J. Water Pollut. Control Fed.* **53** (1981) no. 10, 1519 – 1532.
270. M. P. Czytko, K. Gaupp, R. Müller, *Hydrocarbon Process.* **62** (1983) no. 9, 115 – 119.
271. E. G. Graf, R. K. Thauer, *FEBS Lett.* **136** (1981) no. 1, 165 – 169.
272. J. Weitkamp, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 10, 360 – 466.
273. A. Bursik, S. Buseck, *VGB Kraftwerktechnik* **60** (1980) no. 10, 780 – 787.
274. Intern. Coal Ref., Allentown PA, USA, Report DOE/OR/-3054–69, *Energy Res. Abstr.* **9** (1984) no. 18, 35032;
Chem. Abstr. **102** (1985) 48 517 a.
275. Sedimat(R), Lurgi GmbH, T 132 Mai (1975).
276. H. Richter, *Haus Tech. Vortragsveröff.* **505** (1986) 24 – 30.
277. "Hydraffin Powdered and Granular Activated Carbon for Water Treatment," company brochure, Lurgi Aktivkohle GmbH, Frankfurt a. M., 1597e/12.94/lO.
278. H. G. Franck, G. Collin: *Steinkohlenteer*, Springer Verlag, Berlin-Heidelberg-New York 1968.
279. G. P. Blümer, G. Collin, *Erdöl Kohle Erdgas Petrochim.* **36** (1983) no. 1, 22 – 27.
280. H. Nonnenmacher, O. Reitz, P. Schmidt, *Erdöl Kohle* **8** (1955) no. 6, 407 – 411.
281. F. Sonntag, *Erdöl Kohle* **13** (1960) no. 10, 752 – 757.
282. S. Sevin, J. Nougayrede, W. Willing, G. Bandel, *Hydrocarbon Eng.* **3** (1998) no. 4, 54/56 and 80.
283. J. E. Nobles, J. W. Stancik, *Oil Gas J.* (1983) Dec. 26, 156 – 158.
284. Leuna-Werke, DE 43 05 386 A1, 1994 (P. Szargan et al.).
285. E. Supp, W. Hilsebein, *CEER Chem. Econ. Eng. Rev.* **17** (1985) no. 5, (no. 188) 18 – 21.
286. E. Supp, *Pres. at AIChE Spring Meeting*, Anaheim, Ca. May 20 – 23, 1984.

287. "Upgrading Heavy Crude Oils and Residues to Transportation Fuels: Technology, Economics and Outlook," multisponsored reprint Phase 6 SFA Pacific, Inc., Mountain View, CA, February 1999.
288. R. K. Olson, V. A. Gembicki: *Maximizing Distillate from Reduced Crude*, National Petroleum Refiner's Association, San Antonio, Texas 1984.
289. L. O. M. Koenders, S. A. Posthuma, P. L. Zuideveld: "The Shell Gasification Process for Conversion of Heavy Residues to Hydrogen and Power," Gasification Technol. Conf., San Francisco, Oct. 1996.
290. E. O. Gerstbrein, J. S. Falsetti: "Refinery Upgrading using Texaco Refining Technologies and the Texaco Gasification Process," 2nd Int. Conf. on the Refinery Industry, Vienna, Dec. 1994.
291. H. Fellner, W. Liebner: "Multi Purpose Gasification (MPG), Lurgi's Proven Technology for Refinery Bottoms Management," 4th Forum on Advances in the Refining Industry , Mexico City, Aug. 1998.
292. W. Liebner: "MPG – Lurgi/SVZ Multi Purpose Gasification, a New and Proven Technology," Gasification — A Gateway to a Cleaner Future, Dresden, Sept. 1998.
293. H. Edelmann, F. Stuhlmüller: "EFCC — Ein zukünftiges Konzept für Kohle-Kombikraftwerke?,," *VGB Kraftwerkstechnik* **77** (1997) no. 7, 537 – 543.
294. R. D. Lessard et al., *27th ASME Int. Gas Turbine Conf.*, London, UK, April 1982.
295. S. Kjaer, H. Koetzer, J. van Liere, I. Rasmussen: "Neue Konzepte für kohlebefeuerte Kraftwerke Vergleich von Wirkungsgrad, Wirtschaftlichkeit, Umwelt- und Betriebsaspekten," *VGB Kraftwerkstechnik* **74** (1994) no. 7, 561 – 568.
296. K. Weinzierl: "Kohlekraftwerke der Zukunft," *VGB Kraftwerkstechnik* **74** (1994) no. 2, 109 – 114.
297. U. Dorstewitz, H. Wien, *Conf. Coal Gasification in Energy Technology*, Dortmund, FR Germany, March 1979.
298. H. Maghon, B. Becker, *VGB Kraftwerkstechnik* **64** (1984) 999 – 1004.
299. B. Becker, A. Kreutzer, K. Goebel, *1983 Tokyo Int. Gas Turbine Congress*, Tokyo, Oct. 1983, pp. 791 – 798.
300. I. J. Hartill, J. A. Lacey, *The Inst. of Gas Engineers, 123rd Ann. General Meeting*, Torquay, UK, May 1986.
301. B. Robson, *Energy Res. Abstr.* **1** (1977) 157 – 177.
302. J. S. Joyce, W. E. Zoerner, *5th CEPSI Meeting*, Manila, Phillipines, Nov. 1984.
303. K. Weinzierl, *VGB Kraftwerkstechnik* **62** (1982) 365 – 379.
304. D. R. Simbeck, R. L. Dickenson, E. D. Oliver, EPRI-Rep. AP-3109, Research Project 2207, June 1983.
305. C. W. Schwartz, *Energy Progr.* **2** (1982) 207 – 212.
306. J. van Liere, M. van der Burgh: "Gasification — an alternative to natural gas," *Chem. Eng.* **603** (1995) 15 – 18.
307. R. P. Wieber, J. S. Halow: "Advanced IGCC Power Systems for the United States," *Energy Progress* **7** (1987) no. 2, 119 – 125.
308. H. Meyer-Kahrweg, *Energie* **29** (1977) 227 – 231.
309. R. Reimert, *CIMAC, 14th Int. Congress on Comb. Engines*, Helsinki, FI, 1981, pp. 19 – 27/ESC-DF.
310. N. A. Holt: *5th Int. Conf. and Exhibition on Coal Utilization and Trade*, **vol. 1**, London, UK, Dec. 1985, pp. 153 – 177.
311. D. R. Plumley: "Cool Water coal gasification — A progress report," *J. Eng. Gas Turbines Power* **107** (1985) 856 – 860.
312. J. Veen, D. Ceelen, G. Kreutzkamp, BMFT Forschungsbericht T-86–047, 1986.
313. H. Vierrath et al., *Int. Gas Turbine Conf. and Exhibition*, Düsseldorf, FR Germany, 1986, 86-GT-294.
314. J. Keller, W. Adlhoch, H. Vierrath: "High Temperature Winkler Coal Gasification Process, A proven Technology for Production of Syngas or Fuel Gas," Int. Gas Union 9th World Gas Conf., Milan 1994.
315. J. Keller, W. Adlhoch, H. Vierrath, R. Rizzi: "Low rank Fuel HTW Gasification as an Integral Part of Combined Cycle Power Stations," EERC 17th Biannual Low Rank Fuel Symposium, St. Louis, MO, 1993.
316. K. Shippers et al.: "Cobra will demonstrate High Temperature Winkler IGCC," *MPS Modern Power Systems* 1993 Feb., 41 – 50.
317. G. D. Zon, H. H. J. de Winter, W. Willebroer: "Kohlevergasung im Dienste der Stromerzeugung, die Demo- KV STEG," *VGB Kraftwerkstechnik* **74** (1994) no. 6, 436 – 441.
318. D. Borgmann, B. Schotter: "Kohlegas-Kraftwerk Buggenum," *VGB Kraftwerkstechnik* **74** (1994) no. 6, 532 – 536.
319. J. Th. G. M. Eurlings: "The future for the IGCC: scale, fuels and products," *Modern Power Systems* 1997, Nov., 39 – 44.

320. R. Cortes, W. Schellberg: "The Puertollano IGCC Plant, Commissioning status and preoperational records," EPRI Gasification Technology Conference, San Francisco 1998.
321. W. Schellberg, R. Wetzel: "Entwicklugstendenzen in Kombikraftwerken mit integrierter Flugstromvergasung, Wirkungsgrade, Umweltdaten, Kosten," *Brennstoff Wärme Kraft* **47** (1995) no. 9, 354 – 358.
322. J. Mc Daniel, Ch. Shelnut, T. Berry: "Tampa Electric Company Polk Power Station IGCC Project, Project status," EPRI Gasification Technology Conference, San Francisco 1998.
323. B. Gericke, J. C. Löffler, M. A. Percavec: "Biomassenverstromung durch Vergasung und integrierte Gasturbinenprozesse," *VGB Kraftwerkstechnik* **74** (1994) no. 7, 595 – 604.
324. J. Albrecht, C. Greil: "Fuel Gas Production from Biomass for Industrial Plants and Energy Generation," 10th European Conference Biomass for Energy and Industry, Würzburg 1998.
325. J. Albrecht: "Experiences from CFB Gasification with Biomass and Waste Fuel," Joint EU- VTT Seminar Power Production from Biomass, Espoo, Finland 1998.

General References

326. International Gas Union, Sub-Committee B: Methods of testing gases and raw materials, *14th World Gas Conference*, Toronto 1979.
327. British Gas Corporation: *Analytical Methods*, London Research Station, 1970 ff.
328. British Standard (BS) 3156 Part 1 – 5 and 10 – 12.
329. Gaz de France: *Essai de Chemie Gaziere*, Paris 1954, 1964.
330. DVGW-Regelwerk, Arbeitsblatt G 261: *Prüfung der Gasbeschaffenheit*, ZfGW-Verlag, Frankfurt 1985, under revision 1997.
331. G. J. van Rossum (ed.): *Gas quality; proceedings of the congress of gas quality – specification and measurement of physical and chemical properties of natural gas*, Elsevier, Amsterdam, Oxford, New York, Tokyo 1986.
332. H. Guerin: *Traite de manipulation et d'analyse des gaz*, 2nd ed., Masson, Paris 1981.
333. H. Seebaum, K. Wienke: *Analytische Methoden für die Untersuchung von Ferngas*, 4th ed., Vulkan-Verlag, Essen 1965.
334. J. Hengstenberg, B. Sturm, O. Winkler: *Messen, Steuern und Regeln in der chemischen Technik*, 3rd ed., vol. 2, Springer-Verlag, Berlin – Heidelberg – New York 1980.
335. P. Profos: *Handbuch der Industriellen Messtechnik*, 2nd ed., Vulkan-Verlag, Essen 1978.

Specific References

336. DIN 1340, ed. Dec. 1990.
337. International Gas Union: *Committee A*, Lausanne 1982, pp. 14 – 24.
338. K. Hedden, Th. Heike in [331], pp. 761 – 782.
339. G. F. Steinmetz in [331], pp. 11 – 25.
340. DVGW-Regelwerk: *Arbeitsblatt G 260, Gasbeschaffenheit*, draft, Wirtschafts- und Verlagsgesellschaft Gas und Wasser mbH, Bonn 1997;
DIN 51622, ed. Dec. 1985.
341. J. A. Rijnaarts in [331], pp. 27 – 35.
342. D. Vernet, Y. Deschamps, in [331], pp. 791 – 808.
343. DIN EN 26, ed. Jan. 1981.
DIN EN 437, ed. March 1993.
344. A. G. A. Testing Laboratories, Research Bulletin 36.
345. E. R. Weaver, *J. Res. Nat. Bur. Stand. U.S.* **46** (1951) no. 3, 213 – 245.
346. B. C. Dutton, E. Grymziesky, *J. Inst. of Energy* **56** (1983) no. 427, 107 – 108.
347. B. C. Dutton, Gas *Commun. Inst. Gas Eng.* **1246** (1984).
348. DVGW-Regelwerk, Arbeitsblatt G 280: *Gasodorierung*, Wirtschafts- und Verlagsgesellschaft Gas und Wasser mbH, Bonn 1990.
349. P. Cowan, A. E. Humphreys in [331], pp. 169 – 179.
350. ASTM D 1826–94.
351. G. J. van Rossum, G. J. Benes, *GWF Gas Wasserfach Gas/Erdgas* **122** (1981) 12 – 19.
352. L. J. Mondeil, F. M. Robert in [331], pp. 111 – 119.
353. D. Buwalda, M. W. C. de Jonge, J. Reints in [331], pp. 153 – 161.
354. A. E. Humphreys, *MESUCORA Conggress*, Session 11, Paris 1982, pp. 95 – 112.
355. H. F. A. Trimbos, G. J. van Rossum, W. R. Schierbeek in [331], pp. 139 – 151.
356. C. H. Griffis et al. in [331], pp. 121 – 127.
357. ISO 6976- 1995.
AFNOR X 20–522, ed. June 1982.
ASTM D 3588–91. BS 3156 Part 10.8;
DIN 51857, ed. March 1997.
358. G. J. van Rossum, H. Koning in [331], pp. 809 – 830.
359. DIN 1871, Draft April 1997.
360. ASTM D 1070–85.
361. Verein Deutscher Ingenieure: *VDI-Handbuch Reinhaltung der Luft*, 3490, Part 15, Draft Nov. 1985.

362. M. Jaescke, *GWF Gas Wasserfach Gas/Erdgas* **126** (1985) 93 – 102.
363. DIN 51870, ed. April 1981.
364. ISO 10715- 1997.
AFNOR NF X 20–251, ed. Dec. 1982.
ASTM D 5287–93.
DIN 51853, ed. Nov. 1995.
365. Th. F. Welker, G. J. De Nobel, D. Stirnberg, K. Homann in [331], pp. 415 – 442.
366. BS 1756 Part 2.
DIN 51872 Part 2, ed. July 1980, withdrawn 1997.
367. DIN 51872 Part 3, ed. Feb. 1978, withdrawn 1997.
368. C. J. Cowper, A. J. de Rose: *The analysis of gases by Chromatography*, Pergamon Press, Oxford 1983.
C. J. Cowper, P. A. Wallis, B. R. A. Wybrow in [331], pp. 481 – 493.
369. J. Tranchant (ed.): *Manuel pratique de chromatographie en phase gazeuse*, 3rd ed., Masson, Paris 1982.
370. DIN 51405, ed. May 1987.
371. ASTM D 1946–90.
372. ISO 6568- 1981.
AFNOR X 20–303, ed. Nov. 1974.
373. ISO 6974- 1984;
ISO/DIS 6974 Part 1–6, 1997.
AFNOR NF X 20–501, ed. July 1985.
ASTM D 1945–91.
DIN 51872 Part 4, ed. June 1990,
DIN 51872 Part 5, ed. Aug. 1996.
BS 3156 Part 11.1.2.
374. L. Huber, *Erdöl Kohle, Erdgas, Petrochem.* **37** (1984) 249 – 257.
R. Keuter, M. Struis, A. L. C. Smit, *Process Control Quality* **1** (1991) 127 – 136;
GWF Gas Wasserfach Gas/Erdgas **134** (1993) 21 – 27.
C. Busch, *GWF Gas Wasserfach Gas/Erdgas* **136** (1995) 23 – 30.
375. ISO 6975- 1997.
BS 3156 Part 11.1.2.
376. ISO 6142- 1981,
ISO/DIS 13275- 1995,
DIN-ISO 6142, ed. May 1985.
AFNOR NF X 20–202, ed. Dec. 1979.
R. Keuter, R. J. Beks, H. J. Bardewel, A. L. C. Smit, *GWF Gas Wasserfach Gas/Erdgas* **135** (1994) 164 – 170.
377. ASTM D 2650–93.
AFNOR X 20–302, ed. Oct 1974.
378. DIN 51856, ed. June 1979, withdrawn 1997.
379. AFNOR X 20–377, ed. Dec. 1980.
380. ISO/TR 6567/1- 1980.
AFNOR X 20–380, ed. Nov. 1979.
381. AFNOR X 20–301, ed. Jan. 1978. BS 4314 Part 1.
382. H. Karthaus, H. Engelhardt: *Grundlagen der physikalischen Gasanalyse*, Hartmann und Braun, Frankfurt 1971.
383. AFNOR X 20–361, ed. Sept. 1978. Verein Deutscher Ingenieure: VDI-Handbuch Reinhaltung der Luft 2459, Part 6, ed. Nov. 1980.
384. AFNOR X 20–363, ed. Sept. 1976.
385. K. Leichnitz: *Prüfröhrchen-Meßtechnik, (La technique de mesure par tubes reactifs)*, Ecomed-Verlag, Landsberg 1981.
Dräger-Röhrchen Handbuch 11th ed., Drägerwerk, Lübeck 1997.
Gas Processors Association 2377–77.
386. ISO 11541- 1997.
BS 1756 Part 4.
387. AFNOR NF X 20–530, ed. June 1984.
388. ASTM E 700–79.
BS 3156 Part 10.4; DIN 51777 Part 1, ed. March 1983, Part 2, ed. Sept. 1974.
389. ISO 10101, Part 1–3-1987; DIN 51869, ed. May 1986.
390. M. A. Berliner: *Feuchtemessung*, VEB Verlag Technik, Berlin-Ost 1980.
391. A. W. Jamieson, H. J. Sikkenga in [331], pp. 289 – 299.
392. J. Rötter in [331], pp. 311 – 325.
393. P. L. Cremonesi in [331], pp. 301 – 310.
394. M. Mathews, *Erdöl Kohle, Erdgas, Petrochem.* **37** (1984) 95 – 96.
C. B. Blakemore, J. C. Steichen, G. Dallas in [331], pp. 327 – 335.
395. DVGW-Regelwerk: Arbeitsblatt G 285, *Hydratinhibierung in Erdgasen mit Methanol*, ZfGW-Verlag, Frankfurt 1974.
396. ISO 6327- 1981.
AFNOR NF X 20–521, ed. Dec. 1981.
ASTM D 1142- 95.
BS 3156 Part 10.3. DIN 51871, ed. Dec. 1985.
397. ISO 4260- 1987.
DIN-EN 24260, ed. May 1994.
398. ASTM D 1072–90.
399. ISO 6326 Part 5–1989; DIN 51400 Part 4, ed. Oct. 1990.
400. DIN 51400 Part 1, ed. Dec. 1992,
DIN 51400 Part 2, ed. Feb. 1978.
401. ASTM D 3246–92.
DIN 51400 Part 7, ed. June 1992.
402. ASTM D 4468- 85.
403. L. Below, *GWF Gas Wasserfach Gas/Erdgas* **120** (1979) 59 – 65.

404. ISO 6326, Part 3- 1989.
DIN 51855 Part 6, ed. March 1989.
405. P. A. Gibbons, *Chromatographia* **19** (1984) 254 – 256.
ISO 6326 Part 4- 1994.
DIN 51855 Part 8, ed. 1997.
406. ISO 6326 Part 2- 1981.
AFNOR NF X 20–511, ed. Jan. 1982.
DIN 51855 Part 7, ed. Dec. 1986.
407. C. A. Risk, R. A. Hall, *J. Chromatogr. Sci.* **15** (1977) 156 – 159.
DIN 51855 Part 8, ed. 1997.
408. ASTM D 2420–91.
ASTM D 4084–94.
DIN 51855 Part 3, ed. June 1979, withdrawn 1991.
409. ASTM D 2385–81.
DIN 51855 Part 4, ed. June 1995.
I. P. 103/70.
410. ASTM D 2725–87.
DIN 51855 Part 4, ed. June 1995.
411. G. Herbst, A. Romanski, *GWF Gas Wasserfach Gas/Erdgas* **121** (1980) 105 – 108.
412. DIN 51854, ed. Sept. 1993.
413. DIN 51863 Part 1, ed. Sept. 1983.
414. DIN 51863 Part 2, ed. Sept. 1983.
415. DIN 51864, ed. Sept. 1986.
416. AFNOR NF X 43–018, ed. March 1983.
Verein Deutscher Ingenieure: *VDI-Handbuch Reinhaltung der Luft* 2456, Part 5 + 6., ed. May 1978.
417. ISO 6570 Part 1 – 3 1983/84.
DIN 51859, ed. June 1990.
BS 3156 Part 11.2.
418. DIN 51862, ed. Sept. 1982.
419. DIN 51408 Part 1, ed. June 1983.
420. W. Lommerzheim, *GWF Gas Wasserfach Gas/Erdgas* **117** (1976) 430 – 432.

Gelatin

Wilfried Babel, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Dieter Schulz, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Monika Giesen-Wiese, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Uwe Seybold, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Herbert Gareis, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Eberhard Dick, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Reinhard Schrieber, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Annelore Schott, Deutsche Gelatine-Fabriken Stoess AG, Eberbach, Federal Republic of Germany
Winfried Stein, Fritz Häcker GmbH & Co., Vaihingen/Enz, Federal Republic of Germany

1. Introduction . . . 1
2. Structure and Properties . . . 2
2.1. Structure . . . 2
2.2. Physical Properties . . . 4
2.3. Chemical Properties . . . 6
2.4. Microbiological Analysis . . . 6
2.5. Quality Specifications . . . 6
3. Raw Materials and Production . . . 7
3.1. Raw Materials . . . 7
3.2. Production . . . 7
4. Uses . . . 9
4.1. General Aspects . . . 9
4.2. Food Industry . . . 9
4.3. Pharmaceuticals . . . 11
4.4. Cosmetics . . . 12
4.5. Photography . . . 12
4.6. Technical Gelatin and Glue . . . 13
4.6.1. Matches . . . 13
4.6.2. Coated Abrasives . . . 13
4.6.3. Paper Sizing . . . 14
4.6.4. Adhesives . . . 14
4.6.5. Other Uses . . . 14
5. Economic Aspects . . . 14
6. References . . . 15

1. Introduction

Gelatin is a mixture of high molecular mass polypeptides produced from collagenous animal tissues such as hide splits, pigskin, and bones. Collagen is the most commonly occurring protein in the human and animal body, accounting for about 30 % of the total protein content. Since animal horns and hoofs are not composed of collagen they cannot serve as raw materials, contrary to popular belief. The name gelatin has been used since about 1700 and is derived from the Latin *gelatus* which means frozen [1]. There is evidence that gelatin has been used for at least 4000 years [2]. The great variety of applications of gelatin depend on its solubility in hot water, its availability in a wide range of qualities, and its ability to form thermally reversible gels, which is unique for a protein and is due to the fact that the primary structure of collagen still exists in collagen-derived polypeptides.

2. Structure and Properties

2.1. Structure

Collagen is not a uniform substance but an entire protein family. To date, some 19 polymorphs of collagen have been identified. All types of collagen have a triple-helix structure in which three protein chains are intertwined to form a rigid strand (Fig. 1). The length of the triple helix and the type and position of the nonhelical regions vary from one collagen type to an other.

The "classical" type I collagen, the most important type in the manufacture of gelatin, occurs

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_307

in skin and bone. Type II collagen occurs almost exclusively in cartilage tissue, and type III collagen occurs besides type I in skin. The concentration of type III collagen is strongly dependent on age: embryonic skin can contain up to 50 %, but older skin only 5 – 10 %. The other collagen types occur only in low concentrations and fulfil mostly organ-specific functions.

Collagen consists of three helical polypeptide chains wound around each other and connected by intermolecular cross-links. The triple helix is formed intercellularly from three individual chains (α-chains). Chain selection occurs via intra- and intermolecular disulfide bridges in the region of the N- and C-terminal peptides, additionally supported by glycopeptide structural elements. While collagen types II and III contains three identical chains (homotrimer $[\alpha_1(\mathrm{II})]_3$), type I collagen consists of two identical and one somewhat differing chain (heterotrimer $[\alpha_1(\mathrm{I})]_2\alpha_2(\mathrm{I})$). The triple helix molecule is 300 nm long and 1.4 nm in cross section. The amino acid sequences of the α_1 and α_2 chains are highly homologous (ca. 90 % correspondence).

Collagen contains large amounts of two unusual amino acids: 4-hydroxyproline and ε-hydroxylysine, which are the result of post-translational modification: special enzyme systems facilitate the intracellular hydroxylation of proline and lysine on the growing collagen chain. Other enzymes are responsible for the glycosylation of individual hydroxylysine residues. Type I collagen, for example, contains 0.5 wt % carbohydrate, and type III about 10 %. The carbohydrate portion comprises galactose and a 1,2-linked galactose – glucose disaccharide. The ether-like bond between collagen and the saccharide occurs between the C1 atom of galactose and the ε-hydroxyl group of hydroxylysine.

The enzyme-catalyzed modification steps (hydroxylation and glycolyzation) are restricted to growing individual chains; as soon as triple-helix formation is initiated, modification ceases. As stability of the helix is dependent on the hydroxyproline content, helix formation and proline hydroxylation are interdependent and self-regulating.

The helical part of all three collagen types contains 1014 amino acids per chain, and the N- and C-terminal nonhelical extension peptides (telopeptides) are between 9 and 26 amino acids long. The triple helix molecule has a molecular weight of about 290 000, including the short N- and C-terminal globular extension peptides.

Collagen consists of one-third glycine and 22 % of the imino compounds proline and hydroxyproline; the remaining 45 % comprises some 17 other amino acids (Table 1). The content of acidic and basic amino acids is relatively high. About one-third of the acidic amino acids glutamic acid and aspartic acid is present in the amidated form as glutamine and asparagine. Cysteine is completely absent, and of the sulfur-containing amino acids only methionine is present at low concentration (exception: type III which has two cysteine residues per 1000 amino acids).

The "classical collagen" types are characterized by repetitive tripeptide units – (glycine-X-Y-)$_n$ – where proline occurs in the X and Y positions and 4-hydroxyproline exclusively in the Y position. Several other amino acids also have preferred positions, presumably for steric reasons. For example, glutamic acid, phenylalanine, and leucine occur frequently in the X position, and arginine mostly in the Y position. Generally, collagen types from different species are to a high degree homologous (usually about 90 %); and, within the same species, type I collagen from skin, bones, and tendons are identical.

Triple-helix, water-soluble collagen (types I to III) is converted extracellularly by cross-linking and fibril formation into highly associated, water-insoluble collagen. These intra- and inter-molecular cross-links are responsible for a dramatic increase in stability of the collagen molecule relative to thermal, mechanical, and enzymatic influence. The type and extent of the covalent cross-links are age-dependent, an aspect that is of considerable importance for the gelatin industry. Process parameters and extraction conditions must hence be adapted to the age of the raw materials (e.g., young pig or calf skin / old cattle hide).

2.2. Physical Properties

Commercial gelatin is a vitreous solid with a faint color; it is almost tasteless and odorless and usually contains 9 – 13 % moisture. It gives typical protein reactions and is hydrolyzed by

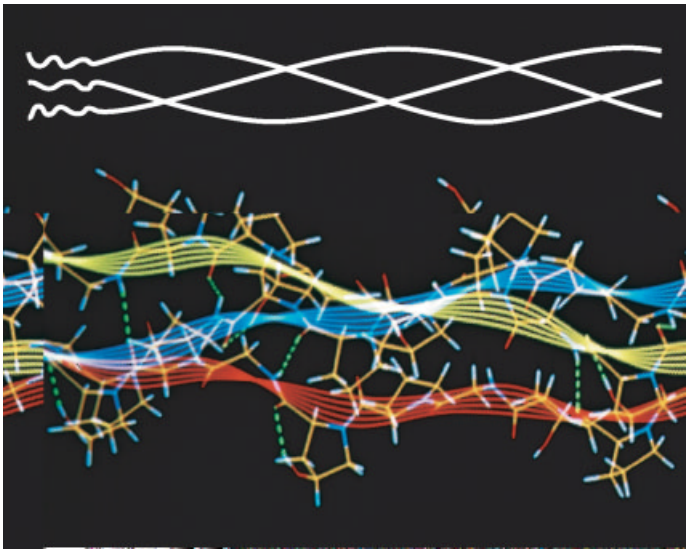

Figure 1. Schematic representation of the triple helix

Table 1. Comparison of the amino acid composition of collagen and gelatin: number of amino acids per 1000 (rounded off)

Amino acid	Type I collagen (bovine)	Type A gelatin ∗	Type B gelatin ∗∗
Alanine	114	112	117
Arginine	51	49	48
Aspargine	16	16	-
Aspartic acid	29	29	46
Glutamine	48	48	-
Glutamic acid	25	25	72
Glycine	332	330	335
Histidine	4	4	4
4-Hydroxyproline	104	91	93
ε-Hydroxylysine	5	6	4
Isoleucine	11	10	11
Leucine	24	24	24
Lysine	28	27	28
Methionine	6	4	4
Phenylalanine	13	14	14
Proline	115	132	124
Serine	35	35	33
Threonine	17	18	18
Tyrosine	4	3	1
Valine	22	26	22

∗ Type A gelatin: acid-pretreated pigskin gelatin. ∗∗ Type B gelatin: alkali-pretreated bone gelatin [1]

most proteolytic enzymes to give its peptide or amino acid components.

Depending on the degree of hydrolysis, two different types of gelatin are obtained:

1) Gelatin, gelling type [*9000-70-8*]: When the granules are immersed in cold water they are hydrated to discrete swollen particles. On heating, these swollen particles dissolve to form a solution.
2) Gelatin hydrolysate (also known as collagen hydrolysate), nongelling type [*68410-45-7*]: In this product range the molecular mass is reduced to such an extent ($<$ 15000 D) that no gelation is observed. This type, mostly produced as a fine powder is soluble in cold water. Gelatin and gelatin hydrolysate are in-

soluble in less polar organic solvents such as benzene, petroleum ether, ethanol, acetone and tetrachloromethane.

Gelatin and gelatin hydrolysate are insoluble in less polar organic solvents such as benzene, petroleum ether, ethanol, acetone and tetrachloromethane. Gelatin is amphoteric: in acidic solution gelatin is positively charged and migrates as a cation in an electric field; in alkaline solution it is negatively charged and migrates as an anion. The intermediate point, where the net charge is zero and no migration occurs, is known as the isoionic or isoelectric point. Type A gelatin, produced by acidic processing of collagenous raw materials, has a broad isoelectric region between pH 6.0 and 9.5. Type B gelatin, produced by the alkaline processing of collagenous raw materials, has an isoelectric point between pH 4.7 and 5.6. The difference in the isoelectric points of type A and type B gelatins is the result of partial desamidation of glutamine and asparagine to the corresponding glutamic acid and aspartic acid. Mixtures of type A and B, as well as gelatins produced by modifications of the above-mentioned processes, may exhibit isoelectric points outside these ranges. Gelatin in solutions containing no ions other than H^+ and OH^- is known as isoionic gelatin [3]. These solutions can be readily prepared by the use of ion-exchange resins.

One of the most important properties of gelatin (gelling type) is the formation of heat-reversible gels in water. Gelatin readily forms gels over a wide range of pH with a variety of solutes. The formation and stability of a gelatin gel is influenced by a number of factors. When an aqueous solution of gelatin with a concentration greater than ca. 1 % is cooled to about 35 – 40 °C it first increases in viscosity and on further cooling it forms a gel. The strength of the gel depends on concentration and the intrinsic strength of gelatin used, which is a function of structure and molecular mass. Gel formation is not fully understood, but is believed to result from hydrogen bonding; the gelatin molecules are arranged in micelles that form a semisolid gel and bind water. Quantitative measurement of this property is important, both for control and to determine the amount of gelatin required for a given purpose because gelatin is used in many products for its gel-forming properties. Modern theory proposes that the first step in gelation is the formation of locally ordered regions caused by the partial random return (renaturation) of gelatin to collagen-like helices (collagen fold). Next, a continuous fibrillar three-dimensional network of fringed micelles forms throughout the system, probably due to nonspecific bond formation between the more ordered segments of the chains. Hydrophobic, hydrogen, and electrostatic bonds may be involved in cross-linking. Since these bonds are disrupted on heating, the gel is thermoreversible. Formation of cross-links is the slowest part of the process, so that under ideal conditions the strength of the gel increases with time as more cross-links are formed. The total effect is a time-dependent increase in average molecular mass and in order. Renaturation involves association between components (peptide chains) that differ in degree of cross-linking, chain length, and chemical composition [4], [5].

Gelatin is stable for long periods of time when suitably stored in sealed containers, at normal ambient temperatures to prevent ingress or loss of moisture. When gelatin is heated above 45 °C in air at relatively high humidity it may lose its ability to swell and dissolve [6].

Gel Strength is determined with a Bloom gelometer [7], [8], [10–13]. A 6.67 % solution of the gelatin sample is prepared in a special wide-mouthed test bottle, which is then cooled to 10.0 ± 0.1 °C and kept for 17 ± 1 h for maturation at this temperature. The firmness of the resulting gel is then measured with a gelometer. This instrument impresses a standard plunger (12.7 mm diameter, plane surface, sharp edges) into the surface of the gel. The force required to depress the plunger 4 mm into the gel is the gel strength or Bloom value of the gelatin. Commercial gelatins vary from ca. 50 to 300 g Bloom.

Viscosity is usually determined by using a calibrated viscosity pipette (Bloom pipette), which measures the efflux time of 100 mL of a 6.67 % solution at 60 °C.

For gelatin hydrolysate, the viscosity of a 10 % or 20 % solution at 25 °C is determined, but other concentrations or temperatures are possible, according to the different uses. For measurement of higher viscosity values, rotation viscometers are also used.

Molecular mass fractions (≥ 200 000 D) have some impact on viscosity, medium molecular mass fractions (55 000 – 300 000 D) have an influence on gel strength. The viscosity of gelatin solutions increases with increasing gelatin concentration and with decreasing temperature; viscosity is at a minimum at the isoionic point. Gelatin solutions should not be exposed to temperatures above 60 °C for prolonged periods of time because the gel strength and viscosity decrease [14].

Color and clarity of gelatin and gelatin hydrolysate can be important for certain applications. These parameters can be measured with a spectrophotometer, but visual assessment against standard gelatins is also common practice. The color of gelatin depends on the raw material and the production process. Pigskin gelatins usually are less strongly colored than those made from bone or hide. Turbidity may be due to insoluble or foreign matter in the form of emulsions or dispersions that are stabilized by the protective-colloid action of gelatin, or can be caused by an isoelectric haze. This haze is at a maximum at the isoelectric point in ca. 2 % solution [15].

Gelatin typically contains ca. 9 – 13 % *moisture*, which can be determined by drying a sample at 105 °C for 17 h.

Technical gelatins are normally specified by their gelling power and viscosity. The test methods employed are generally similar to those used for edible or pharmaceutical gelatins, but different gelatin concentrations may be employed. For lower grades of technical gelatins (and animal glues) 12.5 % solutions (15 g in 105 mL water) are used.

Methods for sampling and testing of animal glues are described in an international standard [16]. In addition to gel strength and viscosity the measurement of moisture content, melting and setting point, foam characteristics, pH and fat content are also described.

In some countries the viscosity of technical gelatin is measured with a cup viscometer, normally the Engler viscometer [17], [18], [19], and using a 17.75 % solution at 40 °C. The viscosity (in degrees Engler, °E) is the flow time of the gelatin solution compared with that of an equal volume of water at 20 °C. Relationships between viscosity measured by capillary and by the Engler method are available [19], [20].

Gelatin also undergoes a phenomenon called coacervation when colloidal particles separate from the liquid phase. Coacervates are formed when two hydrophilic sols carrying opposite charges are mixed in suitable amounts [21]. If gelatin from an acid-treated precursor is mixed with gelatin from an alkali-treated precursor and the resulting gel is prepared at pH 5 – 7, different degrees of turbidity result. This effect depends on the ratio of the two gelatins in the mixture and the pH. This phenomenon arises from the formation of a complex coacervate between the gelatin micelles with opposite charge. Adjustment of the pH above or below the isoelectric range of both gelatins gives a clear gel; both gelatins are now either positively or negatively charged and hence are mutually compatible. A common application of complex coacervation is the use of gelatin and gum arabic to produce oil-containing microcapsules for carbonless paper manufacture [22], [23] (see → Microencapsulation, Chap. 4.1.).

By addition of plasticizers such as glycerin, sorbitol and lubricants like stearate it is possible to produce a blend with thermoplastic properties that can be used in extrusion and injection molding processes [24].

2.3. Chemical Properties

Gelatin is formed by the particle hydrolysis of collagen protein resulting in a mixture of protein fragments of varying molecular mass (gelatin 15 000 – 400 000 Dalton, gelatin hydrolysate < 15 000 D). The amino acid composition (see Table 1) corresponds to that of the collagen from which it is derived. Hydroxyproline and hydroxylysine are two unusual amino acids found in gelatin. The determination of the hydroxyproline content can be used for identification purposes [25].

The quantitative analysis of the elements gives about 50.5 % carbon, 25.2 % oxygen, 17.0 % nitrogen and 6.8 % hydrogen. The ash content of commercial gelatin varies with the origin of the raw material and the method of processing. Gelatins derived from pigskin contain small amounts of chlorides and sulfates resulting from acid treatment before extraction; gelatin from bone and hide contains calcium and sodium salts from the lime used in pretreatment.

Dry gelatin stored in airtight containers at room temperature remains unchanged for long periods. Degradation in solution may be caused by extremes of pH, temperature and by proteolytic enzymes such as papain or trypsin.

Gelatin can be chemically modified to change its properties. It undergoes typical protein reactions, including acylation and carbamylation [26]. Such products include succinylated, phthalated, and carbamylated gelatin. These products are used for special pharmaceutical and photographic applications. A new type is methacrylated gelatin, which is suitable for technical applications. Permanent cross-linking of gelatin can be achieved by reaction with aldehydes such as formaldehyde, glyoxal, or glutaraldehyde [26].

2.4. Microbiological Analysis

Microbiological testing is in most cases carried out according to methods described in the European Pharmacopeia (Ph. Eur.) and the United States Pharmacopeia (U.S.P.). Important test parameters are total aerobic count, including molds and yeasts, after an incubation period of up to 5 d at incubation temperatures of 30 – 35 °C and 20 – 25 °C and counts of colony-forming units on tryptic soy agar and Sabouraud agar. Tests for *Escherichia coli* and Salmonella species include at least one enrichment step and detection of typical growth on selective nutrition media.

Table 2. Gelatin specifications

Parameter	Ph. Eur. limits	U.S. P. limits
pH	3.8 – 7.6	
Total ash, %	2	2
Loss on drying, %	15	
Sulfur dioxide, mg/kg	200	40 ∗
Peroxides, mg/kg	100	
Arsenic, mg/kg	1	0.8
Heavy metals, mg/kg	50	50
Phenolic preservatives	0	
Total viable count CFU ∗∗	1000/g	1000/g
Escherichia coli	0 in 1 g	0 in 10 g
Salmonella	0 in 10 g	0 in 10 g

∗ Gelatin used for capsule manufacture can contain up to 1500 mg/kg. ∗∗ CFU = colony-forming units.

2.5. Quality Specifications

Gelatin, as a protein, is subject to contamination by microorganisms; good manufacturing practice must be followed to ensure a clean product. In addition to rendering gelatin unacceptable for human consumption, bacteria can degrade gelatin to a point at which it loses its gel-forming property. Conductivity and pH have a major impact on gelatin quality due to their importance regarding optimal growth conditions for bacteria [27].

Most gelatin manufacturing plants meet the strict requirements of ISO 9000 [28] and ensure consistent high quality by the implementation of HACCP (hazard analysis and critical control points) techniques [29]. The result is one of the purest proteins available to the food, pharmaceutical, and photographic industries. As a pure, natural protein, gelatin is a food, not an additive and, does not carry an E number.

The majority of commercial gelatins contain less than 3000 nonpathogenic bacteria per gram. For pharmaceutical-grade gelatin the pharmacopeia (Ph. Eur., U.S.P.) set a limit of 1000 bacteria per gram, and Salmonella species and *Escherichia coli* must be absent in 10 g and 1 g, respectively [9], [10]. Physical and chemical characteristics of gelatin are listed in Table 2, together with microbiological specifications for pharmaceutical-grade gelatin.

Up to now there are no official limits for microbiological, physical or chemical parameters for food-grade gelatin. Therefore the Gelatin Manufacturers of Europe (GME) cooperate with official authorities to set microbiological, physical, and chemical limits for food-grade gelatin. It is planned to release a monograph for food-grade gelatin which contains not only the quality limits but describes also the test methods to ensure compliance with those limits.

Since the raw materials are exclusively obtained from animals fit for human consumption, and the production process includes numerous purification steps like washing, filtration, ion exchange and a final UHT sterilization step, gelatin is a safe product with regard to bacteriological, viral, and TSE (transmissible spongiform encephalopathies, e.g., Scrapie and BSE) safety [30], [31] and therefore does not carry any risk for the health of consumers.

3. Raw Materials and Production

3.1. Raw Materials

Gelatin is commercially derived from animal collagen of skins and bones. The principle raw materials worldwide are cattle hides, bones and pigskins. In principle all raw materials are fit for human consumption. Sources from other mammalian species and fish sometimes are also used. Each raw material requires special treatment to remove noncollagenous extraneous substances such as fat and minerals. The resulting pure collagen is then hydrolyzed to gelatin, which is soluble in hot water.

Bones. The so-called green bones are supplied by slaughterhouses and meat packers. The fresh raw material is transported directly to the degreasing plants of the gelatin producer. After preselection the bones are crushed into small particles, degreased, dried and sorted according to size. The mineral components of bones are hydroxyapatite and calcium carbonate. To remove the inorganic components the bones are treated for about a week or more with cold diluted acid. The resulting collagen is called ossein, the starting material for subsequent gelatin extraction.

Cattle Hides. Skins from calves and beef are obtained from animals slaughtered for human use. The subcutaneous, fat-containing tissue is mechanically separated and the skins are split horizontally into two parts: the part previously covered with hair is tanned and subsequently used as a high-quality leather. The parts previously in contact with the flesh are the so called hide splits. These residual thinner sections remain untanned and are dried or preserved with salt or calcium hydroxide to become the raw material used in the production of gelatin.

Pigskin. Pigskin is the most important raw material for production of edible gelatin in Europe and United States. Pigskins are supplied by slaughterhouses and meat processing plants and are frequently cooled or even frozen to prevent deterioration.

3.2. Production

Each raw material requires a special pretreatment of the collagen containing material to render it soluble in hot water. During pretreatment procedures the collagen swells and softens, peptides and cross-links are hydrolyzed, and various substances, regarded as impurities, are extracted. Two main types of gelatin are distinguished: type A from acid pretreated raw material and type B from alkaline processing. The production process is shown in Figure 2.

Acid Process. The acid process is used for pigskin and sometimes for special types of ossein. Washed pigskins are treated with diluted mineral acids for approximately 24 h at low temperature. This pretreatment is sufficient to break acid-labile peptide bonds in pigskin collagen. After partial neutralization the gelatin is then extracted with hot water. Usually, the gelatin is extracted stepwise with successive increase of temperature and time. Gelatins from successive extractions show different physical and chemical properties. The first extract has the highest gel strength and molecular mass and a very low color. Later extracts, obtained at elevated temperatures, contain a higher amount of low molecular weight peptides and therefore have lower gel strength and are more intensive in color and clarity.

Alkali Process. For the production of type B gelatin, cattle hides and ossein are pretreated with alkali (lime or sodium hydroxide) at ambient temperature. The liming time varies from several weeks to several months. The treatment removes impurities and splits cross links and peptides. The liming time is a critical parameter for the further processing.

The dilute solutions obtained by hot-water extraction contain 3 – 10 % gelatin. The liquors are filtered and deionized to remove suspended matter and undesirable amounts of inorganic ions. The clarified, dilute gelatin solution is concentrated to 25 – 35 wt % by vacuum evaporation, filtered, and sterilized. The concentrated liquor is rapidly chilled to a gel, extruded as noodles, which are deposited on a stainless steel net for drying with hot, sterile air. The net passes

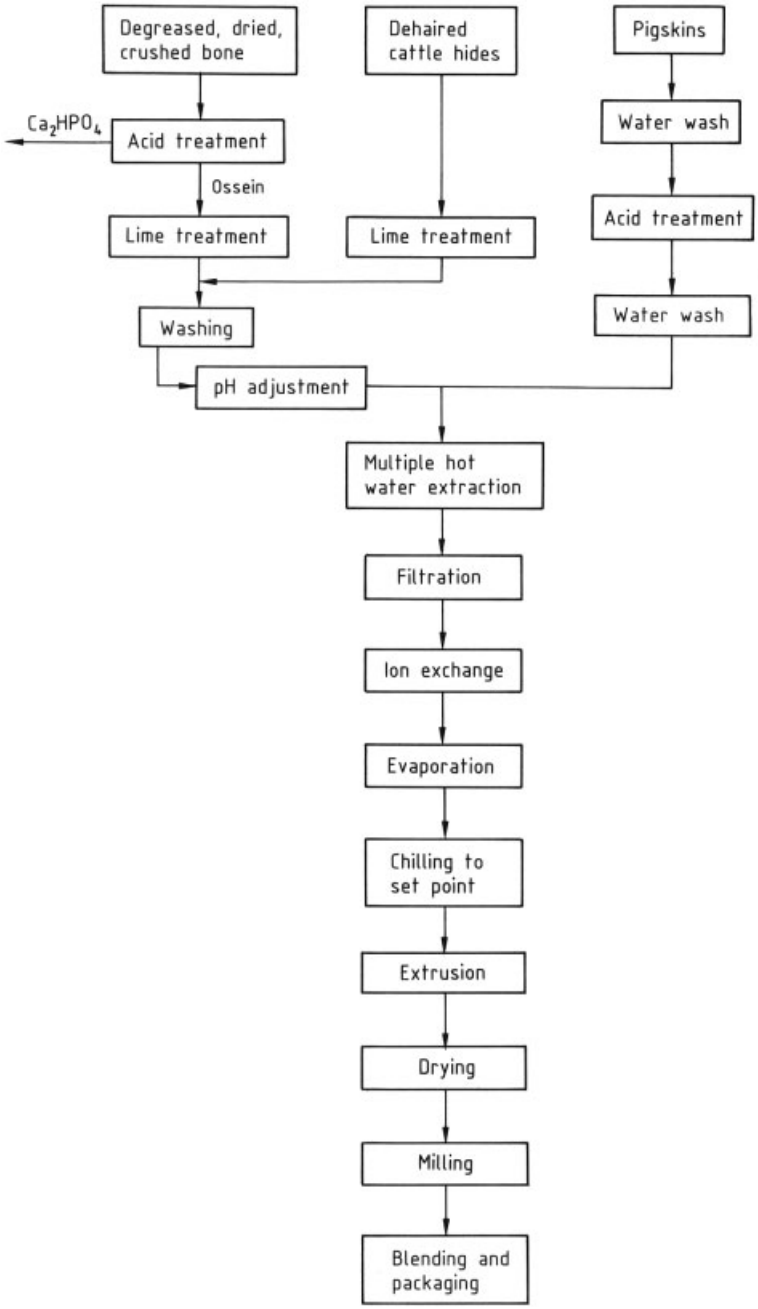

Figure 2. Gelatin production process

slowly through a drying chamber, which is divided into several zones with controlled temperature and humidity. Typical temperature range is from about 30 °C at the beginning up to about 70 °C in the final zone. The dried gelatin, having a moisture content of ca. 10 %, is broken and milled. The physical, chemical and microbiological properties of each extract are tested.

4. Uses

4.1. General Aspects

Although basically a food stuff, gelatin has long been valued not only for its nutritional contribution but also for its functional qualities. Therefore, the use of gelatin is not limited to the food industry. Very important applications are found in the pharmaceutical, cosmetic, photographic and technical area.

The use of gelatin in all these application fields depends largely on the structure function relationships that influence the functionality [33]. The main influencing molecular parameters are:

- Amino acid composition and sequence
- Secondary, tertiary, quaternary conformation
- Net charge and charge distribution
- Hydrophobicity/hydrophilicity ratio
- Size and shape
- Secondary reactions (intra- and intermolecular)
  - Hydrogen bonding
  - Ionic bonding
  - Van der Waal's bonding

The functional role of gelatin does not arise from a single physicochemical property, rather it is a manifestation of a complex interaction of multiple properties. In spite of this the amino acid composition and sequence is the main reason for several functional properties. Gelatin is composed of about 65 % of polar or ionized amino acids. This high proportion of these polar amino acids influences hydration properties such as solubility and water-binding capacity. The type of hydrophobic amino acid residues and the hydrophilicity ratio are associated with amphiphilic behavior, which is important for foaming and emulsifying properties.

Due to its molecular structure and its structure – function relationship gelatin, shows the following properties:

- Gel formation
- Water binding
- Thickening
- Emulsion formation and stabilization
- Foam formation and stabilization
- Film formation
- Adhesion / cohesion
- Protective colloid function

4.2. Food Industry

Edible gelatin and gelatin hydrolysate are widely used ingredients in the production of food. Different properties and different applications require the production of suitable gelatin types [32]. Because of the ease of measuring out the required quantity, *leaf gelatin* is sold worldwide for household use. For cream cakes and cream desserts, special instant gelatins are available for preparing foods without heating. In the food industry powdered gelatin and gelatin hydrolysate are used in the processing of confectionery, dairy products, desserts, meat products, beverages, and so on.

Confectionery. For the confectionery industry, gelatin and gelatin hydrolysate are used due to their different properties:

- In the production of gums and jellies, gelatin is responsible for gel formation, texture, and the bright shiny appearance.
- Aerated confectioneries such as marshmallows owe their light and soft character to the ability of gelatin to form and stabilize foams.
- In fruit chews, gelatin or gelatin hydrolysate is used to prevent products becoming sticky.
- The film-forming properties of gelatin are used in the production of dragees.
- In caramel and liquorice sweets, gelatin provides excellent texture and mouthfeel.
- In pastilles, an elastic structure and excellent melting properties are provided by gelatin.
- In lozenges and compressed products, the binding properties of gelatin guarantee retention of shape.

Dairy Products. Gelatin enhances the texture of dairy products in an optimal way. Many product properties can be controlled by using different quantities and types of gelatin:

- Gelatin acts as a protective colloid in yogurts, thus preventing syneresis. The consistency can be adjusted from creamy to almost solid.
- Soft cheese can be adjusted in consistency from creamy to sliceable.
- Cream and toppings can be stabilized to retain their shapes.
- Sour cream retains its spreading and good melting qualities.
- The ability of gelatin to bind water, form emulsions and provide stability is exploited in the production of low-fat dairy products.
- The melting behavior of ice cream is substantially enhanced by the improved emulsion and finer crystal structure made possible by the addition of gelatin. In this application, gelatin is frequently combined with other hydrocolloids.

Desserts. Crystal-clear, firm to soft gels are required for dessert and fruit jellies; these can be fulfilled by selecting the most suitable gelatin type. In the production of whipped mousse and cream desserts, excellent foam formation and stability of the gelatin are important selection criteria. Different types of instant gelatin with different gelling properties can be used in the production of ready-to-eat desserts, both in mass production and in the households.

Meat and Sausage Products. Gelatin is used as gelling agent in the production of crystal-clear aspic meat and sausage products. Depending on the gelatin type the texture of the gel can range from soft to sliceable. In addition gelatin guarantees the brilliance and color of these products. Nongelling gelatin hydrolysates are used in edible dips and in combination with gelatin as adhesives in meat and sausage coatings.

Additionally both gelatin and gelatin hydrolysate are used to optimize the following technological and sensory quality parameters:

- Reduction of jelly and fat residues in canned sausage
- Improvement of spreading, quality, and softness in emulsified sandwich spreads
- Whipping agent for low-calorie sandwich spreads
- Improved homogeneity of binding in cooked sausage
- Protein enrichment in cured meats
- Rapid reduction of a_w values (water activity, see → Foods, 2. Food Technology, Chap. 1.4.) and shorter maturation times in raw sausage preparations
- Stabilization of emulsions, dispersions and suspension

Beverages. Gelatin and gelatin hydrolysate are used for the clarification of beverages, especially wine, fruit juices and beer. Clarification removes, substances causing turbidity and/or bitterness and have a detrimental effect on optics and taste. Secondly, such pre-clarification can enhance the efficiency of subsequent centrifugation. If necessary, gelatin treatment can be enhanced by adding silicic acid solution or bentonite. This helps to give the wines and juices the desired degree of brilliance.

Bakery Products. In the bakery industry, powder gelatin, leaf gelatin and instant gelatin are primarily used for the binding or gelling of fillings, as well as for the stabilization of creams.

Other Food Applications. Other industrial uses of gelatin in the food area are:

- Binder and gelling agent in canned seafood and seafood aspic products.
- Protein enrichment of food such as in beverages and dietetic products.
- Film former in panned confectionery products such as sugared almonds and hard-panned chewing gum varieties.
- Stabilizer and film former in icings and glazings for bakery products.
- Thickener and emulsifier in soups, sauces, and gravies.
- Emulsifier in half-fat margarines and butter.

4.3. Pharmaceuticals

Gelatin is widely used in the pharmaceutical industry: for the manufacture of soft and hard capsules, as a bulking agent for the coating of water-insoluble vitamins, as a starting material for the

production of human plasma expanders, as a primary component for surgical sponges and as a binder in tablets. In addition to its use as an excepient (i.e., auxiliary agent in pharmaceuticals, see above) gelatin is recommended as a food supplement in the case of joint diseases.

The highest demand for pharmaceutical gelatin is for the production of soft and hard capsules. Soft capsules are made by feeding two plasticized gelatin ribbons between two revolving dies. Pharmaceuticals – either as a liquid or dispersed in an oil – are injected into the capsules as they are formed. Subsequently the capsules are dried in a two-step process.

Gelatin types used for the manufacture of soft capsules are 150 – 175 Bloom limed bone or limed hide, 190 – 210 Bloom acid pigskin or acid bone or a blend of 150 – 170 Bloom acid pigskin / limed bone and hide gelatin. Viscosities are in the range of 2.8 – 4.5 mPa · s. Acid-processed hide and bone gelatin types are preferred for water-miscible or hygroscopic contents. Due to their lower viscosity, less water can be used for the gel mass formulation. Glycerin or polyhydric alcohols like sorbitol are frequently used as plasticizers. In general, the gel mass is composed of 40 – 50 % gelatin, 20 – 30 % plasticizer and water.

Hard capsules, consisting of gelatin, dyes or pigments, and sometimes some processing aids, are produced as two separate parts, known as bodies and caps, by dipping stainless steel pins into a preheated 25 – 30 % gelatin solution. Dried and prelocked hard capsules are forwarded to pharmaceutical companies or contract laboratories to be filled mainly with powdered or granulated pharmaceuticals and sealed. Gelatin types used for the manufacture of hard capsules are 220 – 260 Bloom limed bone or blends of 220 – 260 limed bone / acid pigskin gelatin types. The use of gelatin to coat tablets is a growing market. In a process similar to the manufacture of hard capsules, the tablets are dipped into a concentrated gelatin solution and dried. In another process, the tablets are pressed into prefabricated hard capsules and sealed. Gelatin capsules and coatings protect pharmaceuticals from moisture, oxygen, and microbial deterioration.

Low-bloom gelatins (70 – 140 Bloom) are frequently used to coat vitamins A, D, and E for animal feed and human vitamin preparations. Emulsions of oil-soluble vitamins in gelatin solution containing carbohydrates are spray-dried to give free-flowing powders. Vitamins are thus available in a convenient and stable form that offers protection from light and oxidation.

Specially processed gelatin types have been successfully used as plasma extenders capable of maintaining osmotic pressure. The osmotic pressure of gelatin is strongly dependent on the average molecular weight. Thus the retention time of the plasma extender (generally: 24 – 48 h) is influenced by the size of the molecules. Commercial products are manufactured by partial thermal degradation of high-bloom limed bone gelatin, followed by chemical modification with succinic anhydride or cross-linking the resulting polypeptides with glutardialdehyde or isocyanates. Products containing ca. 3 % of degraded and modified gelatin are liquid at room temperature. Hemostatic sponges are manufactured by whipping a sterile gelatin solution into a foam, which is rendered insoluble by adding formaldehyde on drying. Gelatin sponges impregnated with thrombin can be left in place after surgery or extraction of teeth. Proteolytic enzymes completely digest the sponge.

Gelatin can act as a binder in the manufacture of tablets. Pharmaceuticals are dispersed in a gelatin solution and dried. The ground powder is subsequently compressed into tablets. Glycerin/gelatin masses are used for the production of zinc / gelatin dressings and suppositories.

4.4. Cosmetics

Gelatin and more frequently its hydrolysates find wide application in skin and hair-care products. Hydrolyzed collagen is nontoxic and produces no skin irritation, sensitization, or phototoxicity.

In skin-cleansing preparations, gelatin or gelatin hydrolysates reduce the irritation potential of anionic surfactants.

Due to their affinity to hair keratin, gelatin hydrolysates protect the structure of hair. Particularly in preparations for permanent waving, bleaching, or coloring damage of hair structure can be minimized.

4.5. Photography

Gelatin is an important raw material for all silver halide based photographic materials. For this application mostly alkaline-processed bone gelatin types are used. As a hydrocolloid and a polyelectrolyte, gelatin controls growth of silver halide crystals and prevents coagulation [34]. In addition, gelatin contains substances which influence the photographic properties of the light-sensitive material [35]. Gelatin also stabilizes hydrophobic additives in formulations (e.g., dye couplers) [36], [37]. The ability to undergo reversible sol – gel transition combined with the rheological properties of gelatin enable coating of several emulsion layers at the same time on a paper or film base [38], [39]. Gelatin can be cross-linked by a variety of chemical substances [40] to increase mechanical strength and stability of the coated layers to higher temperatures under acid or alkaline conditions without losing the ability to swell. Therefore chemicals can still diffuse through the different layers. This is essential for the development of the latent image and dye formation during processing of the exposed photographic material.

According to their application and their use in photographic materials, gelatin can be classified as follows:

- Emulsion gelatin
- Oil-in-water emulsion gelatin
- Support gelatin

Emulsion Gelatin. During the *precipitation* of the silver halide crystals, gelatin acts as a protective colloid. It stabilizes the silver halide crystals and influences their growth and their crystal habit. For this application highly purified lime-treated bone gelatin is used to prevent contamination of the silver halide by photographically active substances that may otherwise be present in gelatin. Specially modified gelatin may be used as well. By adjusting the pH, the modified gelatin forms a coacervate. This process is reversible and allows washing of the emulsion.

During *chemical ripening*, gelatin is used to control chemical reactions of added components with the silver halide surface in order to achieve desired photographic properties (e.g., sensitivity and contrast). Since gelatin itself could contain sensitizing or restraining substances, a consistent manufacturing process with a reliably high level of purification is required.

Oil in Water Emulsion Gelatin. Gelatin is used to aid formation of stable dispersions of organic substances (e.g., dyes and color couplers) as finely dispersed oil droplets. For this application the surface active properties of gelatin are important.

Support Gelatin. The physical properties of support gelatin are more important in these application areas than the chemical properties. In antihalation and antistatic coatings gelatin serves as a binding medium for light absorbing and electrically conducting materials. Other uses are as follows:

Before a photographic emulsion is applied to a film base, adhesion must be provided between the hydrophobic base material and the hydrophilic gelatin emulsion. For this purpose so-called *scrubbing gelatin*, mixed with organic solvents or acids, is applied to the substrate as an intermediate layer [41]. Gelatin used for this application must be free of any electrolytes that can flocculate with the organic solvents.

Backing gelatin is applied as an additional layer on the back side of the film base to prevent curling, which is caused by the different swelling and shrinkage behavior of the emulsion coatings and the film base material.

Top-coat gelatin is an abrasion-resistant layer which protects the emulsion layers against mechanical stress, rupture, and friction. For some applications, a glossy surface may also be desirable.

Baryta gelatin is used as a nearly colorless binder for a suspension of barium sulfate which is applied to photographic papers in order to provide a highly reflective white appearance.

4.6. Technical Gelatin and Glue

Technical gelatin is a convenient term for gelatin that is used for nonedible purposes. Chemically, like pure gelatin, it is a mixture of proteins obtained by partial hydrolysis of waste collagen from animal connective tissue, bone, and skin. Technical gelatins, however, usually have inferior physical properties (gelling strength and

viscosity of standard aqueous solutions, measured under standard conditions) and contain a higher proportion of nonprotein material than pure gelatin. The nonprotein component varies with the origin of the raw material and is usually residual mucopolysaccharides and soluble inorganic salts arising from the animal protein waste from which the gelatins are derived.

Technical gelatin has many applications based on its gelation and surface properties, for example: adhesives, remoistenable adhesive tapes, abrasives, matches, paper impregnation, plaster manufacture, production of stabilizers for emulsions, flocculation agents, and many more.

4.6.1. Matches

Only very clean glues are suitable for the production of matches, where they are applied as a binder for the matchhead composition. The inclusion of very fine air bubbles in the matchhead is very important to ensure the presence of additional oxygen and good ignition behavior. Therefore the glue must show well-defined foaming behavior as far as stability and structure of the foam is concerned. Special test methods and instruments have been developed to control these important properties. As fat and grease act as undesirable defoamers, they must be carefully eliminated during production. Furthermore, the glue should contain only very small amounts of salts. Especially chlorides lower the setting point of the glue. In countries with high humidity in the air this leads to softening of the matchhead, rendering ignition more difficult or even impossible. Cross-linking with other components of the matchhead may counteract this undesirable performance to a certain extent. Technical gelatin, due to its high gelation power, is also used as an adhesive to produce the slide cover of the matchbox, which is produced at speeds of several thousands of pieces per minute.

4.6.2. Coated Abrasives

In coated abrasives technical gelatin is used as a binder between the paper substrate and the abrasive component. Gelatin-impregnated papers and unglued papers are used as substrates. Usually two coats are applied by two-roller coating devices. As a base coat a concentrated warm gelatin solution is used to apply a thick adhesive film with good spreading behavior and good flat-lying properties. To achieve good spreading properties, gelatin with certain thixotropic properties is required. In some cases starch is added. After the first coat the abrasive material is added, and a high-voltage electric charge is used to align the particles to improve the abrasive performance.

In addition to gelatin fillers, the second coat may contain pigments and additives with a cross-linking effect to toughen the dried film and to reduce moisture sensitivity, thereby increasing the life cycle of the coated abrasive. The coated paper is usually dried in loops in an oven. Compared with abrasive materials coated with phenolic resins, gelatin provides better and more elastic binding to the substrate. A disadvantage is the lower water resistance. The main application for gelatin coated abrasives is therefore the traditional glass or sandpaper used for smoothing wood, leather, and paint surfaces.

4.6.3. Paper Sizing

Technical gelatin is utilized as a size coating in the production of high-quality all-wood, rag – wood, and rag-based papers:

- Currency and other security papers must possess high performance for multiple use, good moisture and abrasion resistance, and good adherence to printing inks.
- In aquarelle papers gelatin provides the desired reduction in surface fluff.
- Most hand-made papers contain gelatin as sizing agent.

4.6.4. Adhesives

Gelatin has many functions in the adhesive field. A rough surface is required to allow mechanical grab of the glue, and at least one substrate must be permeable to water vapor to accomplish proper bonding. Gelatin glues are generally applied as a solution at ca. 60 °C. They have a unique two-step setting behavior. If applied in

normal thickness to a substrate at room temperature, gelation takes place within parts of a second. A bond is formed which will allow handling of the joined materials almost immediately. The much higher final strength and the high temperature resistance (up to 190 °C) comes after complete evaporation of the free water.

Modified gelatin glues with special additives to develop the desired properties are used for:

- Bookcase making on fast-running machines, where stiff cover materials with high resilient strength must often be turned in.
- Book-spine reinforcement: Gelatin gives the book spine the necessary stability. It is also suitable for bonding the gauze or other material. The glue has to form an elastic film after drying and should have good adherence to previously applied synthetic dispersion glues.
- Lamination of paper to cardboard: glue based on gelatin exhibits good flat-lying properties and long open pot life, and only very little shrinkage of the paper is observed. It is used for the production of posters, jig saw puzzles and games.
- Box covering: cardboard boxes are covered with decorative cover materials. The gelatin-based glue must have good flat-lying properties and strong gelation power.
- Gummed tapes and papers: here often unmodified warm gelatin solutions are used. It is important to dry the solution with hot air to form a glasslike film which can be activated with cold water.
- Musical instruments and furniture: only for high-quality products gelatin based glues are still used today. In mass production, synthetic dispersion glues have replaced technical gelatin.

4.6.5. Other Uses

- Very fine ground gelatin is combined with gypsum to provide special rheological properties to surfacer compounds.
- Gelatin is employed for partly replacing casein in glues.
- Gelatin films exhibit antifogging behavior and are therefore used in the production of gas masks.
- As gelatin films are nonflammable, they are used for compostable grave lights.
- The resistance of gelatin to most organic solvents opens up applications in paintbrush production and the manufacture of paper gaskets.
- Gelatin is used in rubber mixtures to smooth the surface.

5. Economic Aspects

The gelatin market in total has shown only a moderate growth of about 2 to 3 % per annum over the last 25 years. Nevertheless, large differences could be seen from year to year with regard to different areas of application or geographical regions. Pharmaceutical use of gelatin in the late 1990s increased by more than 10 % per year in North America, South America, and the Far East, but there was practically no growth in Europe. The use of edible gelatin declined in the Far East and showed no growth in Western Europe.

Worldwide production of high-grade gelatin in 1997 was ca. 245 000 t and ca. 50 000 t of technical gelatin (glue). The distribution of the production is listed in Table 3.

Table 3. Gelatin production by region (in 10^3 t)

Territory	Gelatin	Technical gelatin (glue)
Western Europe	110	8
North America	42	
Asia/Australia/Africa	43	30
Latin America	42	9
East Europe	8	3

The raw material situation is summarized in Table 4.

Table 4. Raw materials for gelatin in production

Raw material	Gelatin	Technical gelatin (glue)
Pigskin	38 %	10 %
Bones	33 %	40 %
Hide splits	29 %	50 %

The worldwide market shares of the different applications were the following:

Edible gelatin	55 %
Pharmaceutical gelatin	17 %
Photographic gelatin	11 %
Technical gelatin (glue)	17 %

Regarding production capacity the leading manufacturers rank as follows:

High-grade gelatin

1) Deutsche Gelatine-Fabriken Stoess AG, Germany
2) SKW Biosystems, France
3) Leiner Davis Gelatin International, Australia
4) Tessenderlo B.V., Belgium

Technical gelatin (glue)

1) Fritz Häcker GmbH & Co., Germany
2) Trobas B.V., Netherlands
3) Ind. Cola e Gelatina Campo Belo Ltda., Brazil
4) Colas e Gelatinas Rebière Ltda., Brazil

6. References

1. P. I. Rose: "Gelatin", in *Encyclopedia of Polymer Science and Engineering,* **vol. 7,** Wiley, London 1987, pp. 488 – 513.
2. A. Lucas in H. Carter (ed.): *Tut-ench-Amun,* Brockhaus Verlag, Leipzig 1927.
3. A. Veis: *The Macromolecular Chemistry of Gelatin,* Academic Press, New York 1964, pp. 6 – 8.
4. G. Stainsby: "Gelatin Gels," in A. M. Pearson, T. R. Dutson, A. J. Bailey (eds.): *Advances in Meat Research, Collagen as Food,* vol. 4, Van Nostrand Reinhold Company, New York 1987.
5. L. Slade, H. Levine: "Polymer-Chemical Properties of Gelatin in Foods," in [4]
6. R. T. Jones in K. Ridgway (ed.): *Hard Capsules Development and Technology,* The Pharmaceutical Press, London 1987, pp. 41 – 42.
7. Gelatin Manufacturers Institute of America, Inc., *Standard Methods for the Sampling and Testing of Gelatins,* New York 1986.
8. S. Williams (ed.): *Official Methods of Analysis of the Association of Official Analytical Chemists,* 14th ed., 23, AOAC, Inc., Arlington, Virginia 1984, p. 429.
9. National Formulary XXIII, United States Pharmacopeial Convention, Inc., Rockville, Maryland 1997.
10. *European Pharmacopeia,* 3rd ed., European Council, Strasbourg1997.
11. *Supplementary Standardised Methods for the Testing of Edible Gelatine*, Gelatine Manufactures of Europe, Brussels 1998.
12. O. T. Bloom, US 1 540 979, 1925.
13. British Standard 757: "Methods for sampling and testing gelatine (physical and chemical methods)," 1975.
14. R. Schrieber, "Edible Gelatine: Types, Uses and Application in the Food Industry," *Gordian* **76**(1976) 356 – 364.
15. Gelatin Manufacturers Institute of America, Inc., *Gelatin,* New York 1986.
16. ISO 9665, Animal glues – Methods for sampling and Testing, 1993.
17. Bulgarian Standard BDS 543-72 "Bone glue", 1972; BDS 1560-63 "Technical gelatin", 1963.
18. USSR Standard GOST 2067-71 "Bone glue", 1971; GOST 3252-46 "Hide glue", 1946.
19. C. A. Finch, S. Ramachandran: *Matchmaking: Science, Technology, and Manufacture,* John Wiley & Sons, Chichester 1983.
20. E. Sauer: *Tierische Leime und Gelatine,* Springer Verlag, Heidelberg 1958.
21. J. E. Vandegaer (ed.): *Microencapsulation: Processes and Applications,* Plenum Press, New York 1974, pp. 21 – 27.
22. National Cash Register, US 2 800 457, 1957 (B. K. Green, L. Schleicher).
23. M. H. Gutcho: *Microcapsules and Microencapsulation,* Noyes Data Corporation, Park Ridge, New Jersey 1976, pp. 3 – 60.
24. Deutsche Gelatine-Fabriken Stoess AG, EP 0 354 345, 1993 (Deutsche Gelatine-Fabriken Stoess AG).
25. *Standardised Methods for the Testing of Edible Gelatine*, Gelatine Manufactures of Europe, Brussels 1998.
26. R. C. Clark, A. Courts: "The Chemical Reactivity of Gelatin," in A. G. Ward, A. Courts (eds.): The Science and Technology of Gelatin, Academic Press, London, pp. 209 – 241.
27. U. Seybold: "Mikrobiologie von Gelatine und Gelatineprodukten," in H. Weber (ed.): *Mikrobiologie der Lebensmittel: Fleisch und Fleischerzeugnisse,* Behr's Verlag, Hamburg 1997.
28. International Standard Organization ISO 9000.
29. Codex Alimentarius, Food and Agricultural Organization of the United Nations, Rome.
30. R. Schrieber, U. Seybold: "Gelatin production, the six steps to maximum safety," in F. Brown (ed.): *Transmissible Spongiform*

Encephalopathies – Impact on Animal and Human Health, **vol. 80,** Karger, Basel 1993.
31. U. Manzke, G. Schaf, R. Poethke, K. Felgenhauer, M. Muder, "On the Nervous Proteins from Materials Used for Gelatin Manufacture During Processing," *Pharm. Ind.* **58** (1996) 9, 837 – 841.
32. A. Schott, "Speisegelatine, Herstellung, Eigenschaften und Einsatzgebiete," *Food. Technologie Magazin* **7**(1996) 3, 14 – 20.
33. E. Dick, "Gelatine a look at functionality," *IFI* **21** (1998) 2, 42 – 46.
34. M. Szücs: "Die Rolle der Gelatine während der Fällung und des Wachstums von Silberhalogenidkristallen," in H. Amman-Brass, J. Pouradier, F. Moll (eds.): *Photographic Gelatin,* IAG, Fribourg 1996, p. 338.
35. T. Ohno et al., *J. Photogr. Sci.* **33** (1985) 207 – 211.
36. R. Schrieber, *Gordian* **73** (1973) 282 – 287.
37. http://www.dgfstoess.com
38. F. J. Moll: "Gelatin in the Photographic Industry," in H. Amman-Brass, J. Pouradier (eds.): *Photographic Gelatin,* IAG, Fribourg 1994, p. 181.
39. T. Kobayashi et al.: "On the Rates of Gelation and the Setting points of Gelatin Solutions," in M. Austin (ed.): *J. Photogr. Sci.* **40** (1992) p. 181.
40. T. Takahashi: "The Mechanism of the Crosslinking Reaction of Gelatin with Hardener," in H. Amman-Brass, J. Pouradier, F. Moll (eds.): *Photographic Gelatin,* IAG, Fribourg 1996, p. 593.
41. W. Krafft in K. Keller (ed.): *Science and Technology of Photography,* Verlag Chemie, Weinheim 1993, p. 66.

Gems → Crystal Growth
General Anesthetics → Anesthetics, General

Genetic Engineering

Nucleic Acids is a separate keyword

Tattanahalli Nagabhushan, Schering Corporation, Bloomfield, New Jersey 07003, United States

Satwant Narula, Schering Corporation, Bloomfield, New Jersey 07003, United States

1. Introduction 2
2. DNA Cloning Techniques 3
2.1. Cloning Genomic DNA Fragments in *Escherichia coli* 3
2.1.1. Plasmids as Cloning Vectors 4
2.1.2. Bacteriophage Lambda as a Cloning Vector 5
2.1.3. Cosmids as Cloning Vectors 8
2.1.4. Bacteriophage M13 as a Cloning Vector 8
2.2. Synthesis of cDNA 8
3. Methods for Screening and Identifying Recombinant Clones from a Library 9
3.1. Nucleic Acid Hybridization 9
3.2. Immunodetection Techniques . . . 10
3.3. Sequencing Techniques for DNA . 12
4. Expression of Recombinant Clones 12
4.1. Heterologous Gene Expression in Prokaryotes 13
4.1.1. *Escherichia coli* 13
4.1.2. *Bacillus subtilis* 15
4.2. Heterologous Gene Expression in Lower Eukaryotes (Fungi) 16
4.2.1. Yeast (*Saccharomyces cerevisiae*) . 16
4.2.2. Other Yeast Systems 18
4.2.3. Filamentous Fungi 18
4.3. Heterologous Gene Expression in Higher Eukaryotes 18
4.3.1. Mammalian Cells 18
4.3.1.1. Vectors 19
4.3.1.2. Promoters and Enhancers 20
4.3.1.3. Chromosomal Gene Amplification . 21
4.3.2. Insect Systems 21
4.3.3. Plant Systems 21
4.3.3.1. Plasmid Vectors 21
4.3.3.2. Plant Promoters 23
5. Applications of Recombinant DNA Technology 23
5.1. Therapeutic Agents 23
5.2. Vaccines 25
5.3. Genetically Altered Organisms . . 26
5.3.1. Transgenic Animals 26
5.3.2. Transgenic Plants 27
5.4. Gene Therapy 28
5.5. Diagnostics 28
6. Regulatory and Safety Aspects . . 29
7. Glossary of Special Terms 30
8. References 31

Abbreviations :

A glossary of special terms is given at the end of the article (Chap. 7).

A	adenosine
AIDS	acquired immune deficiency syndrome
ADH	alcohol dehydrogenase
ARS	autonomously replicating sequence
bp	base pair
BKV	bovine kidney virus
BPV	bovine papilloma virus
cDNA	copy DNA
C	cytidine
CEN	centromere DNA
CHO	Chinese hamster ovary
cos	cohesive
CSF	colony stimulating factor
dA	deoxyadenosine
dC	deoxycytidine
dG	deoxyguanidine
dT	deoxythymidine
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
EGF	epidermal growth factor
FGF	fibroblast growth factor
G	guanosine
GF	growth factor
IL	interleukin
kb	kilobase
lac	lactose
lpp	lipoprotein promoter

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_319

LTR	long terminal repeat
MLV	murine leukemia virus
mRNA	messenger RNA
MSV	murine sarcoma virus
ompA	outer membrane protein A
ORI	origin of replication
PDGF	platelet derived growth factor
PGK	phosphoglycerate kinase
rDNA	ribosomal DNA
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RNase	ribonuclease
SV	simian virus
T	thymidine
Tag	SV40 T antigen
Ti	tumor-inducing
TNF	tumor necrosis factor
t-PA	tissue plasminogen activator
tRNA	transfer RNA
trp	tryptophan
UPE	upstream regulatory element
vir	virulence

1. Introduction

Genetic engineering or recombinant deoxyribonucleic acid (DNA) technology has almost become a household word. The concept of genetic manipulation of living organisms is not new. Early horticulturists, animal breeders, and microbial geneticists practiced the science without today's detailed knowledge at a molecular level. The awesome precision of the new technology developed by modern molecular biologists has raised heated debates in recent years concerning its possible benefits and dangers.

Historical Aspects. Deoxyribonucleic acid, the genetic material that determines heredity, was first isolated in 1869, by the German scientist Miescher. When DNA was later discovered to consist of only four basic components, it was generally disregarded as a candidate for genetic material because it was simple. Seventy years after the first successful isolation of DNA, Avery conducted his historic DNA transformation experiments with *Pneumococcus,* which proved that DNA and not protein carried genetic information [9]. The discipline of modern molecular biology came into existence in 1953 when Watson and Crick published their brilliant elucidation of DNA structure based on X-ray crystallographic data [10]. The remaining pieces of the puzzle rapidly fell into place. Messenger ribonucleic acid (mRNA), transfer RNA (tRNA) and ribosomes were characterized and their functions elucidated. The mysteries of DNA replication and DNA synthesis were revealed. Major discoveries leading to the rapid development of recombinant DNA technology were the establishment of conditions and rules governing DNA hybridization [11], the existence and purification of DNA restriction enzymes [12], elucidation of the genetic code [13], and the discovery of DNA ligase [14], an important enzyme that enabled in vitro joining of DNA fragments. In 1972, DNA cloning methodology using the above techniques was put into effect by Boyer, Cohen, and Berg [15].

The powerful potential of this technology was immediately apparent to the medical and pharmaceutical community, although genuine, well-meaning concern existed over its misuse. As a result, leading scientists and representatives of government and industry in the United States met in 1975 at Asilomar to draw up a referendum governing the conduct of genetic engineering experiments in academia and industry. The Recombinant DNA Advisory Committee (RAC) was formed. This committee publishes rules and regulations for recombinant DNA experiments in the Federal Register of the United States annually.

In the United Kingdom, guidelines are formulated by the Genetic Manipulation Advisory Group (GMAG), but enforcement is under the jurisdiction of the Health and Safety Executive (HSE). In Switzerland the Commission of Experimental Genetics oversees recombinant DNA research, whereas in Japan this is the task of the Ministry of Education. In the Federal Republic of Germany, the Central Commission for Biological Safety promulgates guidelines for recombinant DNA research.

2. DNA Cloning Techniques

The first major step in cloning, (i.e., replicating) a gene is to be able to isolate and manipulate a copy of the gene. Several different routes may be

followed in cloning, depending on such factors as the information available about the base sequence of the gene, the amino acid sequence of the protein for which it codes, and the functional activity of the protein. Some standard cloning techniques are described in this chapter. They are based either on cloning DNA fragments in *Escherichia coli* or on synthesizing copy DNA (cDNA). The choice of cloning system is strictly dictated by the method of detection available.

The basic steps in gene cloning are outlined in Figure 1. The "foreign" DNA is cut with a specific enzyme known as a *restriction endonuclease,* in this case *Bam*HI. The resulting fragments are joined in vitro to a plasmid vector at a unique restriction endonuclease recognition site (*Bam*HI) by using DNA ligase. Generally, the recognition site lies within a drug-selection marker (e.g., for tetracycline resistance, Tet^r); this facilitates identification of the clone because insertion of DNA into this site inactivates drug resistance. The resulting recombinant plasmid is transformed (i.e., transferred) into bacteria (*Escherichia coli*), and the recombinants are screened by plating on selective drug media. In the case illustrated, the *E. coli* is first plated on medium containing ampicillin to select the ampicillin-resistant organisms. The tetracycline-sensitive, ampicillin-resistant recombinant clones are then detected by comparing replica plates, one containing ampicillin and the other tetracycline. Because the plasmid vector replicates at a high rate, large quantities of cloned DNA can be purified from a bacterial culture. Similar steps are involved in cloning a gene in yeast cells. For a comprehensive account of cloning vectors, see [16]; for genetic loci, see [17].

Most cloning techniques have been refined extensively and are available as convenient kits from several companies. For a description of the enzymes used in genetic engineering, see → Enzymes, Chap. 6.

2.1. Cloning Genomic DNA Fragments in *Escherichia coli*

One of the first requirements for cloning is the availability of a vector that can carry an extraneous fragment of DNA in an autonomous fashion. Therefore, a vector must be an independent replicon, i.e., a single, autonomously replicating genetic unit. It may be a plasmid, a bacterial virus (bacteriophage) such as lambda or M13, or a cosmid. Each of these vectors is easily distinguishable from the genomic DNA of *E. coli* and, hence, easy to purify and manipulate. They all have specific advantages, and choice of a system usually depends on the size of the DNA to be cloned. For short fragments, a few kilobases (kb) in length, a plasmid cloning system is generally adequate. Bacteriophages such as lambda or M13 are excellent for cloning 10 – 15-kb stretches of DNA. For longer DNA fragments of 30 – 40 kb, the cosmid approach is highly recommended.

2.1.1. Plasmids as Cloning Vectors

Plasmids are circular, double-stranded, autonomously replicating, extrachromosomal DNA elements common in several species of bacteria. For the purpose of cloning, a plasmid must have certain features, the most important being a *small size* and a *relaxed mode of replication,* which confer a higher copy number on the plasmid. In the relaxed mode, plasmid replication is independent of chromosomal replication. Consequently, the plasmid continues to replicate even during the stationary phase following chromosomal replication. In addition, one or more *selection markers* are useful for both selection of the transformants and maintenance of the plasmid. The selection markers are usually genes that code for drug resistance (e.g., tetracycline resistance, ampicillin resistance). Alternatively, they may be auxotrophic markers; in this case, the transformants are unable to synthesize certain substances (e.g., the amino acids leucine or tryptophan) and cannot grow on media that lack these compounds. The plasmid must carry unique *restriction enzyme sites,* some of which should be located within a selection marker.

An extensive list of cloning vectors is given in [18]. One of the most widely used plasmids is pBR322 [19] (Fig. 2) or its derivatives. Deoxyribonucleic acid from the desired source is cut at specific sites with an appropriate restriction enzyme. More than 100 restriction endonucleases are now available; they are described in detail in → Enzymes, Chap. 6.1. In nature, these enzymes prevent illegitimate recombination of

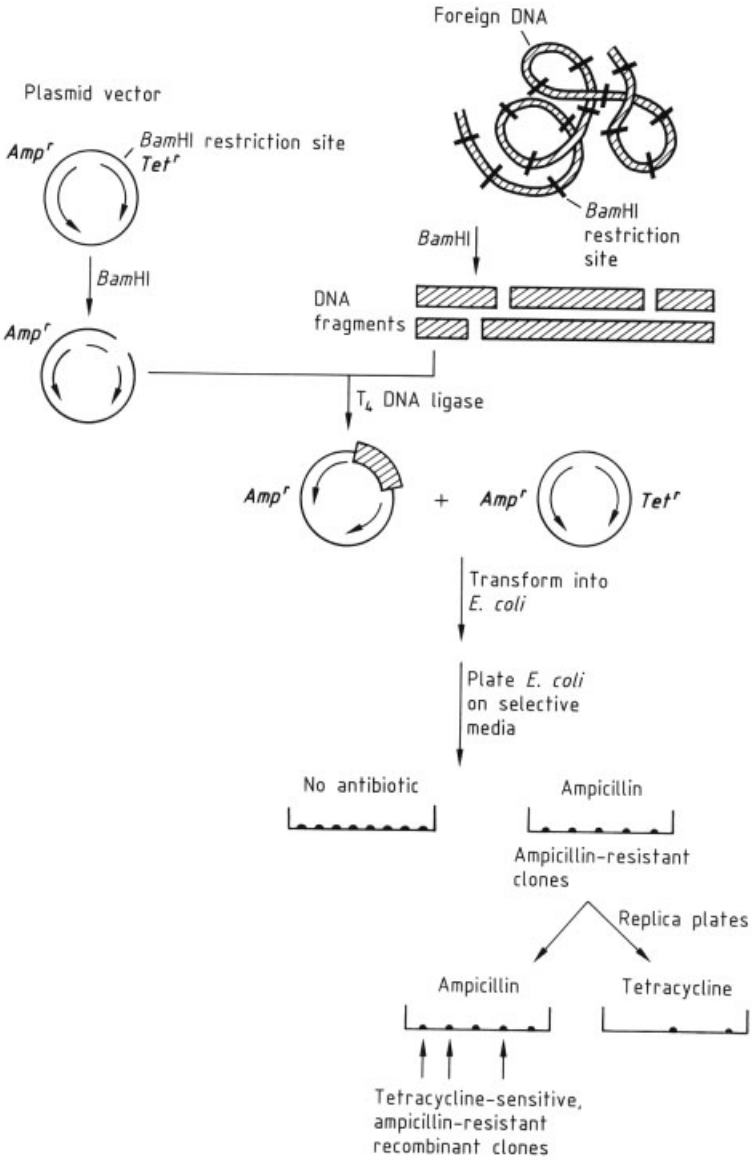

Figure 1. Cloning of foreign DNA in *Escherichia coli* by using a plasmid vector and the restriction endonuclease *Bam*HI
Amp^{r} = ampicillin resistance gene; *Tet*^{r} = tetracycline resistance gene

endogenous DNA with any extraneous foreign DNA by degrading it. They recognize very specific base sequences in DNA, usually four to six nucleotides long, and cut open the double DNA strand to produce two single-stranded overhangs, also known as sticky ends (Fig. 3). Therefore, restriction endonucleases provide not only a means of cutting the source DNA into specific fragments but also a means of recognizing and inserting DNA into the vector. Insertion into the vector can be performed within a selection marker to facilitate selection of recombinant clones by insertional inactivation of the marker, or it may be achieved by directional cloning. In directional cloning the source DNA fragment is cut with two enzymes (e.g., *Hin*dIII and *Bam*HI)

and inserted into a vector cut with *Hin*dIII and *Bam*HI, for both of which it carries unique sites. Because the *Hin*dIII and *Bam*HI recognition sequences are not compatible, the vector ends do not ligate. However, the DNA fragment that does have both cohesive (compatible) ends is inserted in a specific orientation within the plasmid. The circular ligated plasmid transforms *Escherichia coli* with a high frequency, and the recombinants can be picked by means of a selection marker.

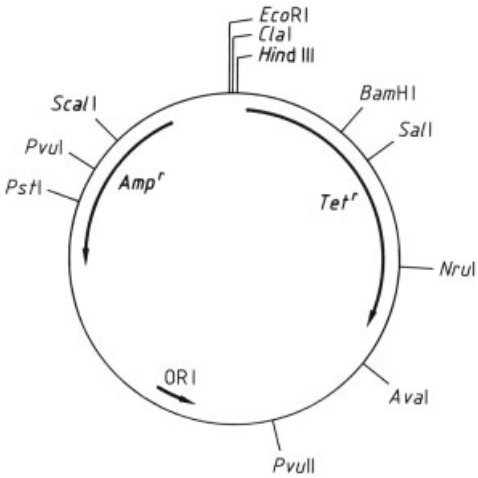

Figure 2. *Escherichia coli* plasmid pBR322 showing the positions of drug resistance genes and some of the unique restriction enzyme recognition sites
Amp^r = ampicillin resistance gene; *Tet*^{r} = tetracycline resistance gene; *ORI* = origin of replication

2.1.2. Bacteriophage Lambda as a Cloning Vector

A virus that infects a bacterial cell is known as a bacteriophage or, simply, phage. Bacteriophage lambda (λ) is a complex lysogenic phage of *Escherichia coli*. It is, however, one of the most well-characterized phages both genetically and physiologically [20], having a typical head-and-tail structure (Fig. 4). The λ genome is a double-stranded linear molecule about 47 kb in length. Each end of the DNA carries 5′ overhangs that are complementary to each other and contain 12 nucleotides. Upon injection into the *E. coli* cell, these cohesive ends, appropriately termed *cos L* and *cos R,* come together and circularize the genome. Replication of the DNA occurs via a bidirectional mode, and the head-to-tail, tandemly arranged DNA (concatenate DNA) is packaged into the head (Fig. 4). The DNA is cut at the *cos L* and *cos R* sites by a protein coded by λ gene *A*. Assembly of the tail with the head completes the formation of an infectious phage particle, which is released by lysis of the host cell. This is called the *lytic cycle.* Alternatively, the phage can also undergo a *lysogenic cycle* in which it is integrated into the host genome and exists as a stable prophage.

The lytic cycle has been adopted for cloning purposes by isolating appropriate mutants of the λ phage. The process has now been so simplified that the various components are supplied commercially as kits by several companies such as Invitrogen, Amersham, Pharmacia, Stratagene, and Clonetech.

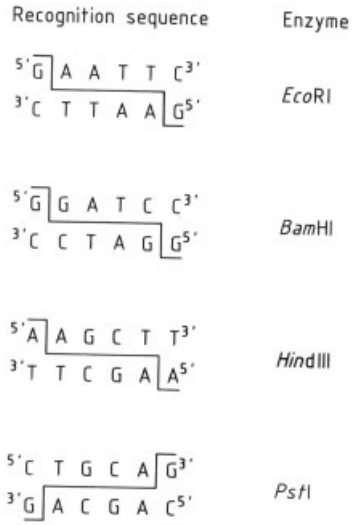

Figure 3. Recognition sequences of *Eco*RI, *Bam*HI, *Hin*dIII, and *Pst*I
The line drawn through the double-strand sequence indicates sites at which the enzyme cuts. Note that *Eco*RI, *Bam*HI, and *Hin*dIII yield 5′ overhangs, whereas *Pst*I yields a 3′ overhang.

For lytic infection, only about 60 % of the λ genome is essential. This includes the left arm, which carries the genes $A-I$ for the head and the tail proteins, and the right arm, which carries the λ rightward promoter (pR) and the *cos R* site. The middle portion is involved in replication and lysogeny, and is not crucial for the lytic cycle. This feature is important because it allows this portion to be replaced by a "foreign" DNA fragment.

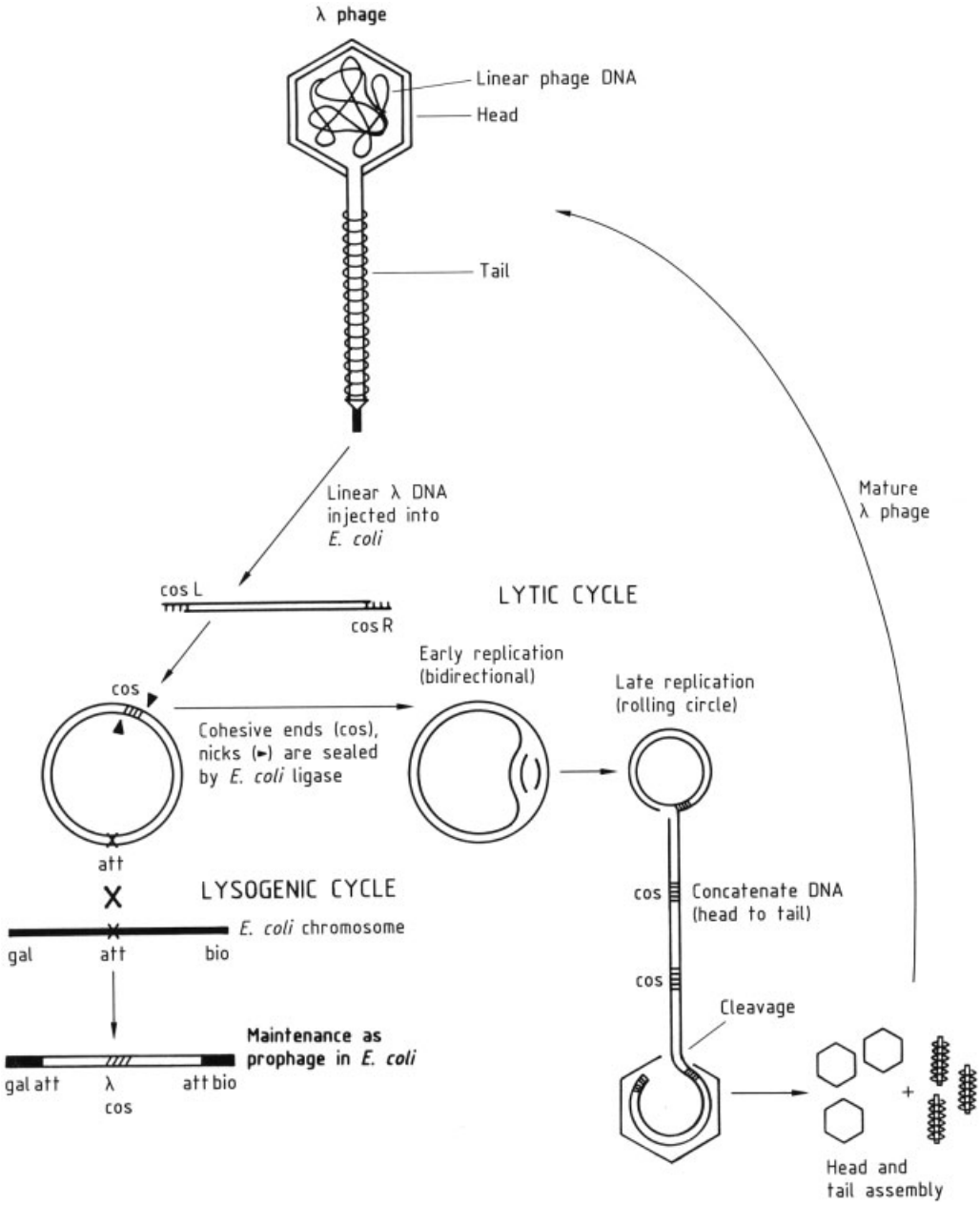

Figure 4. Lytic and lysogenic life cycles of the lambda (λ) phage

To permit introduction of "foreign" DNA, the wild-type λ phage has been modified so as to have a single target recognition site for a restriction enzyme (*insertional vector*) or a pair of such sites that allow part of the phage fragment to be replaced by the insert (*replacement vector*). Several variants of such vectors were constructed by BLATTNER [21], who appropriately named them *Charon phages* after the mythological greek ferryman.

Two of the most popular vectors of which several derivatives are now available are λgt10

and λgt11 [22]. The λgt10 vector contains a single *Eco*RI site in the *CI* gene [23], which codes for the repressor of the phage lytic cycle. Insertion at this site inactivates the *CI* function. As a result, recombinant vectors (CI^-) produce clear plaques, whereas nonrecombinant vectors (CI^+) slow down the growth of the host to produce turbid plaques when grown in *E. coli.* Recombinants derived from λgt10 cloning are usually screened by using nucleic acid hybridization (see Section 3.1). For λgt11 [24], the *Eco*RI site is located within the *lacZ* gene which codes for the enzyme β-galactosidase. Therefore, insertion into this site inactivates the *lacZ* gene. In this case, when grown in *E. coli* that cannot ferment lactose (lac^-), the recombinants can be distinguished on the basis of their ability to form colorless plaques on X-gal plates carrying the chromogenic substrate 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) [25]. A functional *lacZ* gene results in the formation of dark blue plaques due to the hydrolysis of X-gal. Recombinant phage particles obtained in this manner are generally representative of a single genetic source, such as the *E. coli* genome, human genome, or hamster genome, and are called *libraries.* In the case of λgt11 libraries, the *lacZ* protein is expressed. Insertion of a cDNA sequence (see Section 2.2) into the coding region of the *lacZ* gene results in a chimeric or fusion protein, whose proximal end represents part of the *lacZ* protein. Because the "foreign" DNA is then expressed as a protein, the λgt11 libraries can be screened for a particular protein by means of a specific antibody.

Almost every practicing scientist prefers to obtain his or her λDNA in a ready-to-use package. Most companies provide the λDNA as precut left and right arms, which obviates the need to perform this in the laboratory. Thus, genomic DNA fragments or cDNA fragments with appropriate ends can be ligated directly into the arms. The ligated DNA is then packaged into the phage particle. A minimum limit exists for the length of DNA that can be packaged into the λ head. The packaging efficiency decreases dramatically when the insert size is more than 105 % or less than 78 % of the phage genome size [26]. Therefore, if the two arms were ligated directly with each other, they would be packaged very inefficiently.

The success rate in cloning a gene is tied intimately to the final number of recombinant plaques screened. To obtain a complete library, at least 10^6 plaques are required, which is difficult to achieve by transfection of competent bacteria. However, use of an in vitro system for packaging phage DNA into empty phage heads and direct infection of the host cell yield nearly 10^6 recombinant plaques per microgram of ligated vector DNA. The in vitro packaging system consists of cell extracts of two kinds of heat-induced lysogenic strains. One of these is a *tail donor strain* and has a mutation in gene *D,* one of the genes required for head assembly. The other strain is a *head donor strain* and has a mutation in a gene required for tail assembly (gene *E*). Each of these is independently incapable of forming phage particles; in the mixed extract, however, genetic complementation occurs and exogenous DNA is packaged in the phage heads [18].

2.1.3. Cosmids as Cloning Vectors

Cosmids are a cross between the λ phage and a plasmid vector; they contain the *cos* site (hence the name cosmid) of a λ phage ligated into a plasmid vector [27]. Thus, "foreign" DNA can be inserted into a cosmid at a target site. Upon infection of *E. coli,* the cosmid circularizes but replication follows the autonomous plasmid mode and no infectious particles are produced because these functions are not encoded on the vector. The resulting DNA can be packaged into phage particles by using the in vitro packaging system described in Section 2.1.2. Recombinant selection may be based on a drug-selection marker encoded on the plasmid backbone.

Cosmids are extremely useful for cloning very large eukaryotic genomic fragments. Because the size of the parent vector is several times smaller than that of λgt10 or λgt11, for example, a very large foreign DNA fragment may be inserted and packaged with efficiency in the phage particles.

2.1.4. Bacteriophage M13 as a Cloning Vector

Phage M13 is a single-stranded coliphage that goes through a double-stranded replicative

phase in *Escherichia coli,* producing about 300 copies per cell. The phage does not lyse the cell but continuously extrudes phage particles containing single-stranded DNA into the medium. The replicating phage, however, does slow down the growth of the infected cells, producing turbid plaques.

MESSING has adapted this phage successfully for genetic engineering experiments such as DNA sequencing and site-specific mutagenesis [28]. Several variants have been constructed and are available commercially. Introduction of the *lacZ* gene and its promoter into a nonessential region of the phage is extremely useful. The coding sequence of *lacZ* is interrupted by a polylinker (a synthetic DNA fragment that contains sites for several different restriction enzymes). Any one of these sites may be used to clone either cDNA or a genomic fragment. The ligated replicative form is then transfected into a *lac*$^-$ *E. coli* host, and the recombinant colonies are picked as clear plaques on X-gal medium. The background colonies appear as blue plaques because the *lacZ* function is not interrupted. The M13 vectors are indispensable today for the Sanger method of DNA sequencing which utilizes single-stranded templates (see Section 3.3) and oligomer-directed site-specific mutagenesis.

2.2. Synthesis of cDNA

The synthesis of cDNA is used for cloning when the primary aim is to obtain the protein product of the gene.

Preparation of cDNA. The enzyme reverse transcriptase (E.C. 2.7.7.49) is a kind of DNA polymerase found only in retroviruses (see → Enzymes, Chap. 6.2.4.) [29]. It differs from other DNA polymerases by synthesizing DNA from an RNA, and not a DNA, template. Because most eukaryotic mRNA molecules contain a 3′-hydroxyl polyadenosine (polyA) tail, an oligodeoxythymidine (dT) primer is used to prime the reaction. The resultant structure is a DNA · RNA hybrid. The RNA strand can be destroyed with alkali, leaving a single-stranded cDNA copy of the original mRNA. The single-strand cDNA is apparently capable of self-priming due to formation of a transient hairpin loop at its 3′ end. A second polymerase reaction using either reverse transcriptase or DNA polymerase I completes the second strand synthesis. The hairpin and any other single-stranded regions are trimmed away by using the single-strand-specific nuclease S1, which results in a flush-ended duplex structure [30].

In a variation of this scheme, known as the Gübler procedure, RNaseH (endoribonuclease, E.C. 3.1.26.4) is used at the RNA · DNA hybrid stage (Fig. 5.) This nuclease creates nicks in a RNA · DNA molecule. These nicks are extended and filled in by using *Escherichia coli* DNA polymerase I, which has both a 3′ → 5′ exonuclease activity and a 5′ → 3′ polymerase activity. In the final step, T4 DNA polymerase is used to eliminate any persisting 3′ overhangs. This method gives a superior yield compared to the original method [31].

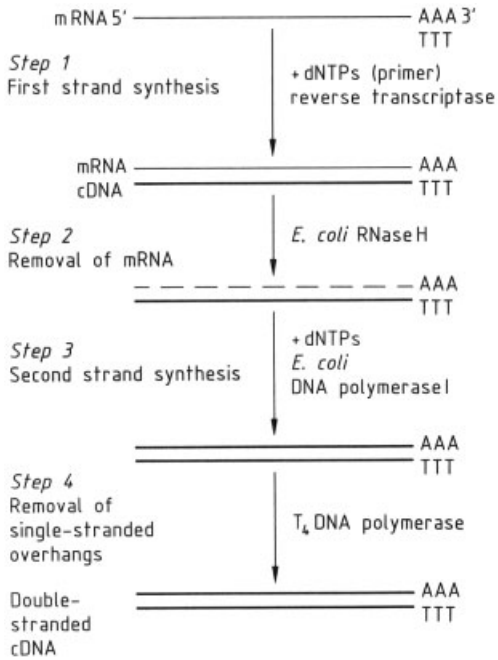

Figure 5. Scheme for cDNA synthesis according to the Gübler procedure
See text for explanation.

Insertion of cDNA into the Vector. Double-stranded cDNA may be inserted into the vector of choice. For insertion into *plasmids, cosmid, or M13 vectors,* synthetic oligomeric DNA fragments (linkers) that carry the recog-

nition site for a desired restriction enzyme are ligated to the ends of the cDNA. The resulting fragment is then cleaved and ligated into the vector cut with the specified restriction endonuclease.

For insertion into *plasmid vectors,* a homopolymer tailing reaction is used which employs the enzyme terminal deoxyribonucleotidyl transferase (E.C. 2.7.7.31) [32]. This enzyme catalyzes the addition of deoxynucleotides to the 3′ hydroxyl terminus of a DNA primer. In most cases, the plasmid pBR322 is cut at the *Pst*I site and tailed with oligodeoxyguanidine (dG), and the cDNA is tailed with oligodeoxycytidine (dC). When the cDNA is annealed to the linear plasmid backbone, the *Pst*I site is regenerated, providing a convenient means of excising the inserted DNA by using *Pst*I [33]. Alternatively, dA – dT tailing may be used. Because dA – dT polymers have a lower melting point than dC – dG polymers, the inserted DNA may be excised with S1 nuclease under partially denaturing conditions [34].

The *Okayama and Berg cloning system,* first described in 1982 [35], is becoming increasingly popular because it yields full-length or almost full-length cDNA clones with a very high efficiency. The main feature of this strategy is that the entire cDNA synthesis is carried out on the plasmid. The process is initiated by using a plasmid primer tailed with oligo(dT) to which the polyadenylated mRNA is annealed (Fig. 5). The cDNA is tailed with oligo(dC) and annealed to an oligo(dG)-tailed, plasmid-derived linker. All subsequent steps are essentially as described above. This procedure has also been adapted to a mammalian cloning vector pCD [36]. The mammalian system provides an added advantage in that the cloned cDNAs can be screened directly for expression, because the vector carries all the essential regulatory sequences such as promoters, splice signals, and polyA signals.

3. Methods for Screening and Identifying Recombinant Clones from a Library

3.1. Nucleic Acid Hybridization

Double-stranded DNA can be melted at high temperature (>90 °C) to yield two single strands, a process termed *denaturation.* Under suitable conditions these single strands reassociate to give the original double-stranded structure; this is termed *renaturation.* Renaturation can take place between any single-stranded nucleic acids as long as their nucleotide sequences are complementary, i.e., DNA – DNA, RNA – DNA, or RNA – RNA. This provides a sensitive method for detecting recombinant clones. Several quick and convenient hybridization methods have been devised for library screenings and detecting specific DNA or RNA sequences.

Plaque-Lift Hybridization. When screening phage libraries, the phage plaques obtained in agar plates can be lifted onto a nitrocellulose filter by simply placing the filter on the agar surface [37]. The filters are then treated to lyse the phage particles, and the released DNA is denatured and fixed. The filters can then be hybridized by using a radioactively labeled DNA or RNA probe under appropriate hybridization conditions. The positive plaques can be identified by autoradiography (Fig. 6). A plaque that gives a positive signal in the primary screen (Fig. 6 A) is isolated, and the putative recombinant phage derived from this plaque is used to infect a fresh *Escherichia coli* culture. The plaques are once again analyzed as described previously; in this second step, a severalfold increase in plaque purification occurs (Fig. 6 B). The process is repeated a third time with a single isolated plaque; in the third step, total purification is usually achieved (Fig. 6 C).

Colony Hybridization. In a process similar to plaque-lift hybridization, nitrocellulose replicas of crowded bacterial colonies may be made [38]. The filters are then treated with detergents such as sodium dodecyl sulfate to lyse the cells and release the cellular contents. The filters are baked to fix the nucleic acids, at which point they are ready for hybridization with the radioactively labeled probe. The positive colonies are evident on autoradiography and can be picked from the original master plate for further investigation.

Southern Blotting. Southern blots provide a means of localizing specific DNA fragments within a mixture [39]. The fragments are first

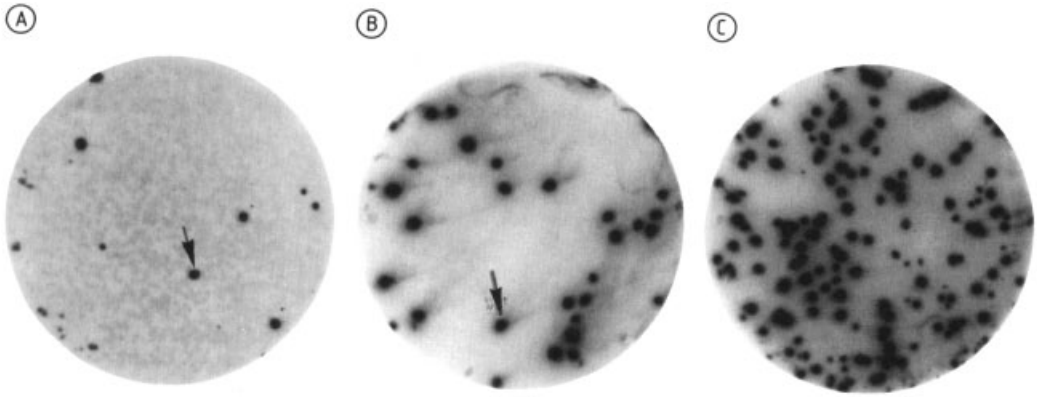

Figure 6. Three-step purification of a recombinant plaque by using plaque-lift hybridization and autoradiography for identification
Positive plaques are blackened (arrow) and their proportion increases in successive purification stages (A, B, and C). In the final step (C), all plaques give a positive signal, indicating a pure clonal preparation.

separated by gel electrophoresis. After denaturation, the DNA is transferred by passive diffusion to a nitrocellulose filter by placing the gel on the nitrocellulose sheet and then immobilized by baking the filter for 2 h at 65 – 80 °C in a vacuum oven. A radioactive DNA or RNA probe can be used to locate the band of interest. The technique is applied mainly to confirm the presence of a specific sequence in cloned DNA. It is also extremely useful for locating sequences in total eukaryotic DNA (Fig. 7).

Northern Blotting. Northern blots are usually applied after confirming the identity of a clone. The RNA, not DNA, isolated from the appropriate cell sources is electrophoresed and transferred to a nitrocellulose filter as described for Southern blotting [40]. Hybridization with the cloned DNA probe can yield useful information such as confirming the transcription of the cloned piece of DNA as well as the number of copies of RNA that are synthesized. This procedure may also be used to test various tissue sources for the presence of a particular RNA. By using appropriate probes, the start and stop sites of transcription and regions that are spliced out of RNA during RNA processing can also be identified [41].

Preparation of Radioactively Labeled Probes. Probes for nucleic acid hybridization are labeled primarily with radioactively tagged nucleotides. A double- or single-stranded DNA or RNA fragment may be labeled at the 5′ end by using a γ^{32}P-labeled ribonucleotide. Probes of very high sepecific activity can be prepared by using a technique called *nick translation* [42]. Here, nicks are produced in a double-stranded DNA fragment with DNase I and subsequently filled in with radioactively labeled nucleotide triphosphates by using DNA polymerase.

3.2. Immunodetection Techniques

Detection techniques based on immunochemical methods are extremely useful in screening expression libraries such as λgt11. The recombinant plaques are lifted onto nitrocellulose as described in Section 3.1. A specific antibody is used in the first step to bind to the recombinant protein. In the second step, this antibody reacts with a species-specific antibody that is either radioactively labeled or tagged with an indicator enzyme (e.g., alkaline phosphatase or horseradish peroxidase). In the former case, autoradiography is used to detect the positive clones; with the latter, the substrate gives a chromogenic response on reaction with the enzyme. A variation of this technique is popularly referred to as *Western blotting* [43]. Cellular extracts or mixtures of proteins are electrophoresed on a polyacrylamide gel and transferred by electroblotting techniques to a nitrocellulose sheet. Detection techniques are identical to those described for plaque-lift assays. The Western tech-

nique is extremely useful for identifying a specific protein in a mixture of proteins and also yields information on approximate protein size. It is used widely as an identification test in the course of expressing a protein.

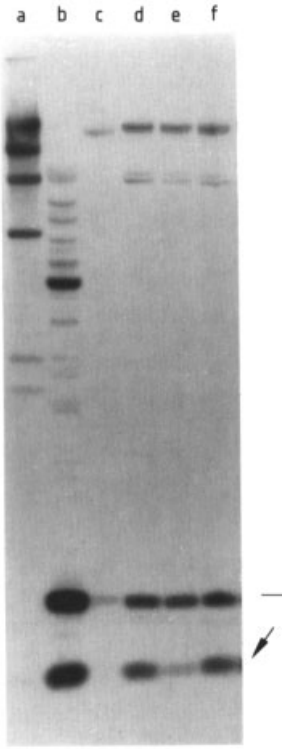

Figure 7. Southern blotting
The amplification of human interleukin-4 (IL-4) is followed on increasing concentrations of methotrexate (see Section 4.3.1.3) by using a "nick-translated" radioactively labeled probe (see below)
a) Lambda DNA *Hin*dIII marker; b) Mouse fibroblast cell line (L cells) transfected with Hu IL-4 cDNA plasmid; c) "Unamplified" Chinese hamster ovary (CHO) cell line transfected with Hu IL-4 cDNA plasmid; d) – f) Sister clones of CHO cell line transfected with Hu IL-4 cDNA plasmid and selected on 500 nmol/L methotrexate for DNA amplification (see Section 4.3.1.3.). Arrow points to the band corresponding to the 1L-4 coding sequence.

3.3. Sequencing Techniques for DNA

The development of rapid DNA sequencing techniques has facilitated characterization of a cloned sequence. The DNA sequence reveals important features such as the location of the restriction sites and open reading frames which may be used to deduce the amino acid sequence of the protein. Sequencing of DNA is carried out either by the method of Maxam and Gilbert [44] or by the dideoxy method described by Sanger [45]. Several automated DNA sequencers are now available. Most of these use Sanger's dideoxy method of sequencing. An innovative approach employs fluorescent dyes to label the nucleotides, thus allowing direct readout of the gels [46].

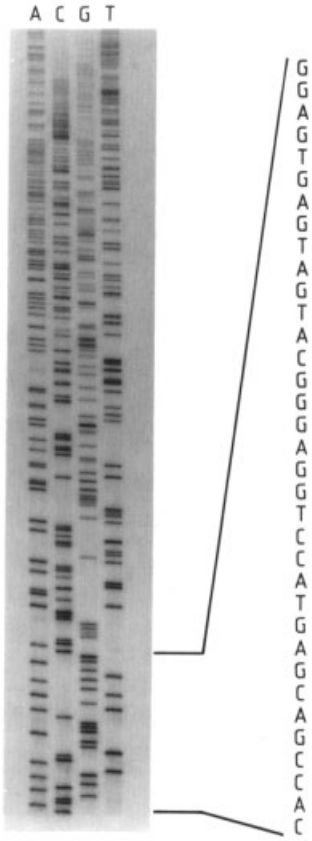

Figure 8. Autoradiogram of a DNA sequencing gel run according to Sanger's dideoxy method
The gel is "read" as described in the text (Section 3.3). Part of the base sequence is shown.

The *Maxam and Gilbert method* is based on the random cleavage of a DNA fragment by chemicals that cleave specifically at A, G, C, or T residues. Each of these reactions is carried out separately, and the four mixtures are electrophoresed on adjacent lanes of a polyacrylamide gel. The resulting fragments are separated strictly on the basis of their size. Because the starting DNA fragment is radioactively la-

beled at the 5′ end, the fragments on the gel can be detected by autoradiography. The sequence is read by starting with the lowest band and moving upward in a serial fashion, identifying bases by the lanes into which they fall.

The *Sanger method* uses a chain-terminating dideoxynucleotide. Here, the reaction depends on enzymatic synthesis of a complementary copy of a single-stranded DNA template (hence, the importance of cloning into single-stranded phage vectors such as M13; see Section 2.1.4). The substrate contains a mixture of the four nucleotide triphosphates, one of which is radioactively labeled. The reaction is performed on four mixtures, each of which contains one of the four specific dideoxy nucleotides at concentrations calculated to yield random terminations. Reaction is initiated by addition of the Klenow fragment of DNA polymerase (see → Enzymes, Chap. 6.2.2.). The reaction mixtures are electrophoresed in parallel lanes, and the gel is read as described for the Maxam and Gilbert method (Fig. 8).

4. Expression of Recombinant Clones

The production of proteins from recombinant DNA (i.e., *expression*) is an important aspect of genetic engineering. *Escherichia coli* has been extremely useful in the expression of foreign proteins that do not require post-translational modification for activity [47], [48]. However, in the case of complex proteins, (e.g., blood factor VIII), or proteins requiring post-translational modification (e.g., glycosylation or phosphorylation) for bioactivity, alternatives must be sought because *E. coli* is not capable of carrying out this task. Other major problems associated with *E. coli* are instability of the foreign mRNA or protein. Also, although expression levels as high as 40 % of the total cellular protein are achievable, the "foreign" protein often forms large, insoluble aggregates called inclusion bodies. This presents the problem of denaturing and refolding the foreign protein in vitro. Therefore, research on gene expression in lower eukaryotes (e.g., yeast) and in mammalian cells has gained prominence. The expression of foreign genes in a heterologous organism (i.e., from a different species) has drawn attention to the intermediate stages in expression of a protein, each of which may be rate limiting in the case of a foreign protein. Because the expression of high levels of a foreign protein is not essential to the functions of the host organism and may even be deleterious to its survival, a major area of research has been directed toward optimization of regulatory parameters for transcription, translation, and protein turnover.

Another aspect of heterologous gene expression concerns large-scale culture of the host organisms such as *E. coli, Bacillus subtilis,* fungi, and mammalian cell lines. Because the pharmaceutical industry already had extensive experience in large-scale fermentation of several antibiotic-producing bacteria, *E. coli* fermentations were standardized very quickly. Similarly, the brewery industry's experience with yeast was useful in dealing with recombinant yeasts. The greatest challenge was the scaleup of mammalian and plant cell lines. Many specialty houses mushroomed quickly to deal with this problem, which resulted in several ingenious technical innovations in this area. These include growth of cells on a synthetic microcarrier surface, gel entrapment of cells, and the growth of cells on solid matrices such as hollow fibers, collagen, or glass (→ Immobilized Biocatalysts) [49]. Most modern biotechnology companies can deal with large-scale growth of recombinant mammalian cell lines. For further details of biotechnological considerations, see → Biotechnology.

4.1. Heterologous Gene Expression in Prokaryotes

4.1.1. *Escherichia coli*

Escherichia coli is, by far, the most widely exploited prokaryote host for gene expression. The primary reason for this is that for decades *E. coli* has been the "workhorse" of microbial geneticists and, hence, its genetics and physiology are well-known. Early vector systems were based on the lambda bacteriophage and ColEI, an indigenous *E. coli* plasmid carrying the gene for the antibiotic colicin. Some of the earliest gene fragments to be cloned were well-studied *E. coli* promoters such as *lac* (lactose) [50], *trp* (tryptophan) [51], and λP_L [52].

Plasmid Vectors. Both inducible and constitutive plasmids with high copy numbers have been developed [53]. Constitutive high-copy plasmids arise from mutations in the control elements that regulate plasmid replication. Inducible high-copy mutants also have mutations in these elements but maintain a low copy number until they are induced. Inducible systems are generally preferred because the lower copy number of the uninduced plasmid allows more rapid cell growth. Both thermally and chemically inducible plasmid vectors have been used successfully in *Escherichia coli* [54]. On induction, these plasmids lose control over their replication and are termed runaway replication, high copy number mutants. The induced high copy number results in a dosage increase of the recombinant gene, generally with a concomitant increase in expression levels. However, the increase in expression level depends on the nature of the protein being expressed and may not be directly proportional to the copy number [55]. Most runaway-replication plasmids reach a copy number of 2000 per cell. Such intensive replicative activity tends to exhaust the cell's synthetic machinery and may even lead to plasmid instability. Also, in the case of inducible promoters such as *lac*UV5 and λP_L, it can very quickly dilute the repressor, making the system uncontrollable. Overexpression of certain genes such as tetracycline resistance can lead to plasmid instability. Plasmid stabilization technology has been developed and assures that plasmid loss is minimized to $10^{-4} - 10^{-6}$ cells per generation [56]. A plasmid stabilization system based on valine synthetase has also been developed [57]. Here, an *E. coli* host without valine synthetase activity is used, and its viability is assured by the presence of valine synthetase on the plasmid. This may be especially important when expression of a protein that is toxic to *E. coli* is desired because cells which have lost the plasmid possess a distinct advantage over plasmid-containing cells and could rapidly overgrow the culture.

Promoter Systems. A promoter is a conserved regulatory genetic sequence that is recognized by RNA polymerase for the initiation of mRNA transcription. The expression level of a heterologous protein must be directed by a host-specific promoter to achieve a high level of expression. One of the major efforts in genetic engineering has been to identify and construct strong promoters that have a high rate of transcriptional initiation.

A number of strong, inducible promoters have been developed based on the knowledge of endogenous *Escherichia coli* promoters such as *trp, lac, lpp* (lipoprotein promoter), and λP_L. In addition to natural promoters, several genetically engineered promoters have also been developed. An example is the *tac* promoter, which is a hybrid composed of the "−35 region" of the *trp* promoter and the "−10 region" of the *lac*UV5 promoter. It is stronger than either of the promoters from which it was constructed [58]. An extremely elegant, well-controlled expression system has also been developed in which the promoter and the gene to be expressed are maintained on separate plasmids [59]. Thermal induction of specific λ bacteriophage functions, that are maintained on the host chromosome results in specific recombination between the two plasmid species and expression of the recombinant gene. Bacteriophage T7 has also been exploited in several gene expression systems which rely on the fact that T7 RNA polymerase recognizes only T7 promoters and that *E. coli* RNA polymerase cannot recognize T7 promoters. Therefore, a gene cloned under the control of a T7 promoter is expressed only in the presence of T7 RNA polymerase. By placing the expression of T7 RNA polymerase under the control of an inducible *E. coli* promoter, the system can be turned on for heterologous gene expression. One such system relies on the presence of the T7 RNA polymerase gene which has been placed on the host cell chromosome [60].

Translational Efficiency. The efficiency with which a heterologous protein is translated from mRNA in *Escherichia coli* may be affected by parameters such as the ribosome binding site, the nucleotide sequences surrounding the ribosome binding site and the initiation codon [61], as well as codon usage within the gene [62]. In some cases, the secondary structure of the heterologous mRNA impedes efficient translation [63]. Addition of DNA sequences coding for several amino acids 3′ of the start codon increases the expression level of bovine growth hormone from less than 0.5 % to ca. 30 % of the total cell protein, with either a *trp* or an *lpp* promoter [64]. This effect is probably mediated

by an alteration in mRNA secondary structure that results in enhanced amenability to translation. Computer-assisted molecular modeling can be used to predict the secondary structure and introduce modifications that will yield better translation.

Secretory Systems. *Escherichia coli* secretory systems sequester the recombinant protein within the periplasmic space. The pINIII ompA system relies on the *lpp – lac* hybrid promoter along with the ompA signal peptide [65]. Another system is based on the *E. coli* alkaline phosphatase promoter and signal sequence [66]. The ompA signal peptide is derived from the gene encoding ompA (outer membrane protein A) in *E. coli.* The signal peptide is responsible for translocating the protein to the outer membrane. Both of these systems are inducible and have the following advantages:

1) Proteins are removed from the cytoplasm, thereby reducing toxic effects on the cell.
2) Exposure of the recombinant protein to cytoplasmic proteases is reduced, thus facilitating purification.
3) Proteins are obtained with their natural amino termini.

Escherichia coli does not usually secrete proteins into the medium. Attempts using hemolysin [67] and staphylococcal protein A amino terminal sequences [68], however, show promise and are being pursued in several laboratories.

Proteolytic Degradation and Fusion Proteins. One of the problems encountered during the expression of heterologous genes in *Escherichia coli* is the proteolysis of the foreign protein. With small proteins, this problem may be overcome by using a *lon*$^-$ or *htpR*$^-$ host background. The gene *lon* encodes the primary proteolytic activity, protease La, expressed during heat shock [69]; the *lon*$^-$ mutation inactivates this protease. The *htpR*$^-$ mutation results in inactivation of the repressor regulating the expression of the heat-shock operon. This operon codes for several proteases, including protease La. Because expression of a "foreign" protein can also trigger the heat-shock response, *lon* may play a major role in protein degradation. The *htpR* locus encodes a positive regulator of heat shock [70]. Therefore, an *htpR* mutation leads to nonexpression of heat-shock genes. The additive effect of *lon* and *htpR* suggests that other proteolytic enzymes may exist besides the protease La encoded by the *lon* gene.

The rate of proteolysis can also frequently be reduced if the recombinant protein is expressed as a *fusion protein.* This may be accomplished by using either an assembly of several copies of the gene of interest arranged one behind the other (*tandem copies*) [71] or an additional peptide sequence that can be cleaved after purification of the protein, e.g., collagen sequences [72]. Alkaline phosphatase fusion proteins have also been used to identify secretory mutants that show an increased level of expression of human growth hormone [73].

4.1.2. *Bacillus subtilis*

Bacillus subtilis, a gram-positive sporulating microorganism, has been explored extensively for the expression of foreign proteins. One of its most attractive features is the potential for secretion of foreign proteins into the culture medium. The surface of these organisms is enclosed by a simple, single cell-membrane system; they can thus secrete very high levels of some intrinsic enzymes. For example, up to 10 g/L of α-amylase (10 – 20 % of the total cellular protein) can be secreted by some industrial strains. Most of the proteins secreted into the growth medium by *B. subtilis* contain a *signal peptide* at the amino terminus. Translocation of the protein is guided by the signal peptide. The signal peptide is cleaved prior to release of the mature protein into the culture medium.

Another major advantage of *B. subtilis* is that it does not secrete any endotoxins. Furthermore, the fermentation industry has experience with this organism.

Plasmid Vectors. *Bacillus subtilis* itself has very few endogenous, naturally occurring plasmids; hence, most vectors have been derived from other gram-positive bacteria, mainly *Staphylococcus aureus.* These multicopy vectors replicate autonomously in *Bacillus* in a relatively stable manner. The most common vectors and plasmid progenitors available today are pUB110 [74], pC194 [75], and pE194 [76]; they

are all derived from *S. aureus.* The most important selection markers used are resistance genes to chloramphenicol, erythromycin, and kanamycin. Several bifunctional plasmids that are capable of replication in both *E. coli* and *B. subtilis* (e.g., pVH33 [77] and pBS42 [78]) are also available.

Promoter Systems. The α-amylase promoter–signal sequence is the most widely used regulatory sequence in *Bacillus subtilis.* Proteins such as staphylococcal protein A [79], human α-2-interferon [80], human γ-interferon, and *Escherichia coli* β-lactamase [81] have been expressed by using this system. Levels of secreted proteins range from 1 mg/L to 2 g/L. One of the major problems encountered, at least in the expression of human α-interferon, is sequestration of the precursor molecule as insoluble aggregates in the cytoplasm. Secretion of heterologous proteins hooked up with α-amylase signal is not always as efficient as that of the natural protein and is very dependent on the nature of the "foreign" protein. In addition to α-amylase, signal sequences from *B. amyloliquefaciens* alkaline protease and serine proteases [82] and *B. licheniformis* penicillinase [83] have also been used for the expression of heterologous proteins.

Extracellular Proteases. *Bacillus subtilis* secretes large quantities of extracellular proteases, especially during the onset of sporulation. This has been a major hurdle to expressing large quantities of heterologous proteins. To overcome this problem, attention has turned to obtaining host mutations that eliminate these proteases. Some mutants are now available, and their use has improved secretion levels of some heterologous proteins [82], [83].

4.2. Heterologous Gene Expression in Lower Eukaryotes (Fungi)

Fungi are a valuable source of commercially important natural products such as enzymes, vitamins, and amino acids. They are also widely used for fermentation processes. Both yeast and filamentous fungi (Fig. 9) have been exploited widely for these purposes. The vast experience of industry in the growth of these organisms has, in many cases, expedited industrial use of recombinant strains.

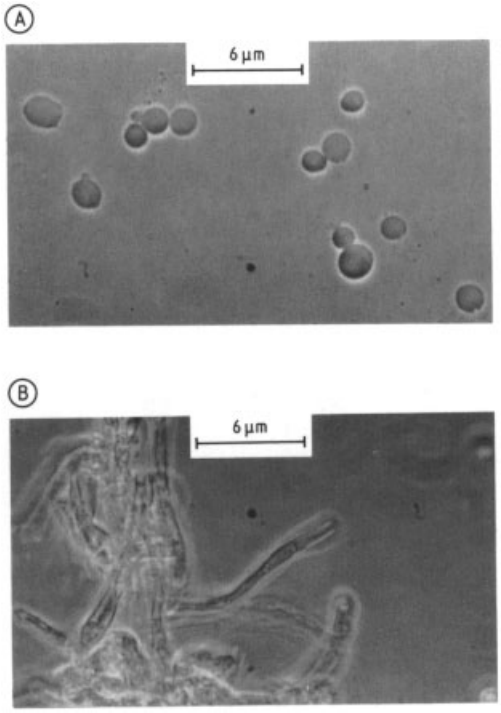

Figure 9. Photomicrographs of fungi used in genetic engineering
A) A commonly used laboratory strain of the unicellular yeast *Saccharomyces cerevisiae.* Notice the various stages of "budding" which indicate cell division; B) *Aspergillus nidulans,* a filamentous fungus currently being investigated as a host for production of heterologous proteins

4.2.1. Yeast (*Saccharomyces cerevisiae*)

Heterologous gene expression systems in yeast were pursued partly for the same reasons as the *Escherichia coli* systems. The genetics of the yeast *Saccharomyces cerevisiae,* chosen for this purpose, have been well studied. Yeast also offers the facility of a secreted product. The most appealing feature of the yeast system, however, is that it is eukaryotic; thus, yeast cells can splice mRNA precursors and carry out the posttranslational modifications (e.g., glycosylation, acetylation, proper folding) that are essential if a eukaryotic protein is to be fully functional.

Although a number of recombinant products have been expressed in yeast at high levels [84], expression levels are usually low com-

pared to those in *E. coli* or mammalian cells [85]. Another problem is the unfaithful cleavage of the heterologous product [86], which inevitably leads to severe problems in downstream protein purification. Although yeast was initially thought to be the ideal host, several promising features have fallen short of expectations; e.g., splicing functions are limited and do not follow the course generally observed in mammalian cells [87]. A major problem, especially in *Saccharomyces,* is hyperglycosylation of the heterologous product. Also, in certain cases, *Saccharomyces* has failed to glycosylate a given protein at its known glycosylation sites [86]. Yeast does not conform to mammalian systems in the manner in which it glycosylates a given protein [88]. Furthermore, the codon usage of yeast is very different from that of either *E. coli* or mammalian systems [89].

Plasmid Vectors. Two types of plasmid have been used widely for the expression of heterologous proteins in yeast: the 2-micron ($2\,\mu$) and the *ARS* plasmids.

2 μ Plasmid. The $2\,\mu$ plasmid is derived from an endogenous yeast plasmid that exists in the cell at a high copy number ranging from 50 to 100 [90]. By combining this plasmid with *Escherichia coli* plasmid pBR 322, several *shuttle vectors* have been constructed. These plasmids can replicate autonomously in both *E. coli* and yeast. They are extremely useful because they facilitate use of *E. coli* for all preliminary genetic manipulation, in addition to the ease of obtaining DNA from *E. coli.* The major disadvantage of using autonomously replicating plasmids is their inherent instability. This problem can be partly overcome by using 0^- strains of yeast which lack endogenous $2\,\mu$ plasmids. Such plasmids have been used successfully to produce large quantities of desired proteins.

ARS Plasmid. The *ARS* plasmids contain an *A*utonomously *R*eplicating *S*equence which was originally isolated from the yeast genome and confers autonomous replication function on plasmids in this organism [90]. Several *ARS* sequences have been isolated (e.g., *ARS1* [91], *ARS2* [92], and *ARS3* [93]). Sequence analysis indicates limited homology between the elements.

Like the $2\,\mu$ plasmids, the inherent problem with *ARS* plasmids is their relatively low stability, especially during mitotic growth [94]. This is generally attributed to competition for the initiation factors for DNA replication and their preferential segregation into the parent cell [95]. The *ARS* plasmids can be stabilized by introduction of centromeric DNA (*CEN*) [96] and become true minichromosomes. Mendelian segregation of the circular minichromosome then occurs during meiosis and confers plasmid stability in mitotic segregation. Several *CEN* elements have been identified and display significant homology [97]. The stability of *CEN* plasmids approaches, but does not equal, that of normal chromosomes.

Integrating Plasmids. Yeast can be transformed with linear plasmid DNA that is cleaved within a region showing homology with the DNA of the yeast genome. Plasmid integration is then targeted to the homologous genomic sequence [98]. Furthermore, multiple copies of the plasmid are found integrated into the yeast genome in a tandem array [99]. By using the *URA3* gene, this approach has been employed successfully to express high levels of heterologous proteins [100].

Linear Plasmids. Linear plasmids that are capable of autonomous replication in yeast are also called *artificial chromosomes.* They were initially constructed by flanking linear plasmids with terminal sequences of extrachromosomal ribosomal DNA (rDNA) from *Tetrahymena* [101]. Large linear plasmids that can exist in one or two copies per cell were found to be more stable than short linear plasmids or circular *ARS* and *CEN* plasmids. Their behavior during meiosis follows a strict Mendelian pattern of segregation.

Promoter Systems. A number of endogenous yeast promoters have been developed for the expression of heterologous proteins. The phosphoglycerate kinase (*PGK*) promoter and the 5′ coding region for PGK have been fused to the 5′ end of the human α-interferon gene. This inducible system has resulted in the expression of a PGK–α-interferon fusion protein as $1-2\,\%$ of the total cellular protein [84]. The *PGK* promoter has also been used to express human γ-interferon [86]. Another inducible promoter, derived from the acid phosphatase gene, has

been used in the expression of hepatitis B surface antigen [102]. The promoter derived from the yeast alcohol dehydrogenase gene (*ADH1*) has been used to express human α D-interferon [85]. Other useful promoters are the *GAL1* promoter, which is positively regulated by galactose [103], the extremely strong glyceraldehyde-3-phosphate dehydrogenase promoter [104], and the α-factor promoter [105].

Secretory Systems. A number of yeast secretory systems have been developed. One such system relies on the signal sequence derived from the yeast mating factor, mat-α, and has been used to express human epidermal growth factor [105]. Interestingly, this gene could not be expressed intracellularly. When fused to the α-factor signal sequence, the protein is secreted and correctly folded. The system, using a 2 μ-based plasmid, has been shown to process and secrete several proteins such as murine granulocyte – macrophage colony stimulating factor (GM-CSF) [106] and human interleukin-2 (IL-2) [107]. However, this system does not work with every protein. Smaller proteins such as growth factors seem to be more amenable to secretion than larger molecules. Other secretion signals such as *SUC2* and *Pho5* have also been employed to secrete heterologous proteins into the culture medium [88].

In most of the secretory systems, although a major proportion of the protein is directed into the secretory pathway, only a small fraction is secreted into the medium [85], [86], [88]. Obviously, rate-limiting steps exist. One way to overcome this problem is to isolate mutants with elevated levels of secretion–supersecreting mutants–which has been accomplished for chymotrypsin [88]. An encouraging aspect of the study is that these "supersecretory" mutants can be used with other expression – secretion systems, as well as for expressing heterologous proteins other than chymotrypsin [88].

4.2.2. Other Yeast Systems

The methanol utilization pathway of *Pichia pastoris* consists of genes that are responsive to methanol. The product of one of the genes, alcohol oxidase, makes up as much as 30 % of the cell's soluble protein. The regulatory sequence of this gene has been isolated and used for expression of heterologous proteins at high levels [108]. Expression can be controlled by switching carbon sources from glucose to methanol. Another appealing feature of *Pichia* is that, unlike *Saccharomyces,* it does not hyperglycosylate the proteins [108].

4.2.3. Filamentous Fungi

The direct appeal of filamentous fungi as hosts for heterologous gene expression derives from the fact that they secrete vast amounts of enzymes. Under optimal fermentation conditions, *Aspergillus niger* secretes up to 20 g/L of the enzyme glucoamylase from a single copy of the gene! By using the framework provided by yeast, the technology for gene expression in filamentous fungi is undergoing rapid development. Both *A. nidulans* (Fig. 9 B) and *Neurospora crassa* are eligible candidates because they are the most widely studied filamentous fungi.

Plasmid Systems. Attempts to isolate endogenous natural plasmids from *Aspergillus* and *Neurospora* have not been successful. The plasmid pDJB3 was the prototype for subsequent work [109]. It contains the ampicillin resistance gene and the origin of replication sequence from the *Escherichia coli* plasmid pBR325. Replication in filamentous fungi is made possible by the *ansl* sequence, which was isolated by following procedures employed for the *ARS* sequences of yeast. The *Pyr4* gene of *Neurospora* is used as a selection marker. However, several *Aspergillus* genes that match available mutants are now available and are being used in the development of vectors. The rate of transfer of DNA into the fungi (i.e., the transformation frequency) is high (5000 transformants/µg of DNA); most of the transformants are mitotically stable, indicating chromosomal integration.

Promoter Systems. A number of *Aspergillus* and *Neurospora* genes have been isolated, and their regulatory sequences are currently under investigation. The most sucessful application to date has been use of the glucoamylase promoter for the expression of bovine

chymosin in *A. nidulans* [110]. Glucoamylase promoter is induced by the presence of starch in the growth medium. The chymosin so produced was cleaved faithfully, with at least 90 % of the product being secreted efficiently. Although new, the filamentous fungi hold great promise as efficient secretors. However, a great deal of groundwork is still required, especially in the areas of vectorology and regulatory signals in order to exploit the full potential of this system.

4.3. Heterologous Gene Expression in Higher Eukaryotes

4.3.1. Mammalian Cells

Gene expression in mammalian cell cultures has an advantage over both the prokaryotic and the lower eukaryotic systems because it yields a product that most closely resembles the natural protein. Recombinant proteins expressed in mammalian cells are likely to be more properly glycosylated, assembled, and folded to yield functional proteins (post-translational modification). One of the major drawbacks of using mammalian cells as host organisms was considered to be the high cost involved in growing various cell lines. Extensive work has gone into developing cheaper culture media, and modern cell-culture technology is highly amenable to scaleup. For proteins such as erythropoietin, in which the activity is closely linked to glycosylation [111], expression in mammalian cells is the only available recourse. Also, *Escherichia coli* may not efficiently express extremely large molecules such as blood factor VIII or tissue plasminogen activator; therefore, these molecules are best expressed in mammalian cell lines. So far, almost every cell line used has shown faithful mRNA processing; i.e., the precursor mRNA is correctly spliced to yield the expected mRNA. When cDNAs with their signal sequences are cloned, faithful cleavage of the precursor protein and efficient secretion rates have been obtained.

The significance of glycosylation has still not been resolved. Glycosylation patterns appear to vary from one cell line to another. How critical is glycosylation of a protein to its functionality? What are the implications of altered glycosylation patterns? These questions have not yet been answered definitively. Many proteins that are glycosylated in their native state function equally well as unglycosylated forms expressed in *E. coli* [112], [113]. However, for some proteins such as erythropoietin, glycosylation is essential for bioactivity and alteration of the glycosylation pattern affects the range of bioactvity [114]. Furthermore, some cell lines cannot carry out all the different kinds of post-translational modifications. For example, several commonly used mammalian cell lines do not efficiently perform the γ-carboxylation step required by blood factor IX [115].

4.3.1.1. Vectors

Higher eukaryotes do not carry extrachromosomal (episomal) DNA molecules equivalent to bacterial plasmids. Most vectors are derived from eukaryotic DNA viruses or retroviruses.

Vectors Based on DNA Viruses. All eukaryotic vectors fall within two classes: integrating (intrachromosomal) or episomal (extrachromosomal).

Integrating Vectors. The most popular integrating vectors are derived from the *simian virus* SV40, a widely studied double-stranded DNA polyoma virus [116]. These shuttle vectors contain sequences from the bacterial plasmid pBR322 that allow replication and segregation in *Escherichia coli.* The SV40 sequences consist of a promoter unit, an intervening sequence, and a polyadenylation site. The cDNA to be expressed is cloned adjacent to the SV40 early promoter. A variation of this vector containing the gene for the SV40 T antigen (Tag) is also available [116]. The Tag is involved in SV40 DNA replication and promotes transient replication of vectors containing the SV40 origin. After transfection into COS-7 cells, the plasmid replicates rapidly. Due to the increased gene dosage, most heterologous proteins are expressed at easily detectable levels. This provides a quick method of screening recombinant clones within 72 h whereas 2 – 3 weeks are required with stable transformants that arise by random integration of the plasmid into the host genome. The COS-7 monkey cell line has the Tag integrated in its genome and is a popular cell line for transient transfections [117].

Episomal Vectors. The following are major advantages of an episomal vector:

1) Discrepancies or aberrations in expression of cloned genes resulting from random indiscriminate integration into the host genome can be avoided.
2) A standardized environment is available for comparison of related genes or their mutants.
3) A higher level of gene expression may be obtained due to the multicopy nature of the plasmid.

Of the episomal vectors, those based on the *bovine papilloma virus* (BPV) have been studied extensively. The BPV DNA is a circular duplex about 8 kb in length which is maintained episomally at about 100 copies per cell [118]. It can transform mouse, rat, and hamster cell lines and give rise to foci that can be visually scored [119]. A 69 % subgenomic fragment of BPV is sufficient to achieve transformation in the mouse fibroblast cell line C127 [120]. The plasmid is maintained in an episomal state.

Recombinant vectors have been constructed from this fragment and the plasmid pBR322 or pML-2, and are widely used for expression of heterologous genes [121]. If the vector containing the heterologous gene is maintained episomally, the copy number is generally low [121]. Although the BPV virus exists as an episome in its native state, the presence of a heterologous gene in the BPV vector may result in its integration into the host chromosome. The BPV vector has been used in conjunction with the metallothionein promoter to express biologically active dimeric human choriogonadotropin [122].

Bovine kidney virus (BKV) based vectors have also been employed to achieve high-level expression of heterologous genes in an episomal state [123].

The prospects for developing stable, nonintegrating plasmids may also receive an impetus from the discovery of indigenous episomal elements (*L factors*) in certain mammalian cell lines [124]. The L factors are circular duplex molecules that appear to have their own origins of replication. Whether they will maintain these characteristics after introduction of heterologous sequences remains to be seen.

Retroviral Vectors. A retrovirus has a RNA genome that acts as a template for DNA synthesis. The intermediate DNA is then used as a template for mRNA transcription. The demonstration that cloned retroviral DNA can produce infectious virus particles (virions) on transfection into cultured cells set the stage for their use as vectors [125]. Retroviruses have several advantages over DNA viruses. Integration of retroviral DNA into the host chromosome is obligatory for their reproduction and is carried out very efficiently. The long terminal repeat (LTRs) contain all the necessary information required in cis (i.e., on the same strand of the RNA genome) for transcription, replication, and integration [126]. The rest of the viral DNA can, therefore, be replaced by foreign DNA. By using a helper cell line to supply other viral functions in trans (i.e., from an independent genome), an infectious virion can be produced. In the case of *murine leukemia virus* (MLV) based vectors, two types of helper cell lines are available: those that produce a mixture of wild-type and recombinant virions, and Ψ^- mutants (the Ψ protein is required for the packaging of wild-type virion DNA) which produce only the recombinant virions. Retroviruses are capable of infecting various cell types; this fact has been exploited to introduce extraneous DNA sequences into bone marrow cells and embryonic cells. Retroviral vectors have been used to introduce cloned genes into cells both in culture and in vivo. Successful transfer of genes into normal bone marrow cells and their subsequent transplantation into recipient animals have been achieved [127]. Retroviruses are currently used to create transgenic animals by infecting preimplantation embryos [128]. With the recent understanding of tissue-specific expression of some mammalian promoters, the goal of tissue-specific expression (e.g., in milk and pancreas) has also been achieved [129], [130].

Most of the retroviral vectors are based on *Moloney murine sarcoma virus* (Mo-MSV) and *Moloney murine leukemia virus* (Mo-MLV). Many of the vectors such as MVGT [131] and M-MuLVneo [132] use the long terminal repeats for direct transcription of the cloned gene and the *gpt* and *neo* genes of *Escherichia coli* for selection. Others, such as MLVtK and MMuLV-SuTK-neo [133], carry additional promoters for expression of cloned genes. Presence of the SV40 origin of replication in the latter case allows amplification in monkey COS cells.

4.3.1.2. Promoters and Enhancers

A typical eukaryotic regulatory sequence contains a promoter and enhancer(s). The *promoter* guides the RNA polymerase to the correct initiation site for the eukaryotic messenger RNA. A typical eukaryotic promoter is about 100 base pairs (bp) long and consists of an AT-rich region designated the TATA box. The TATA box is generally preceded by one or more upstream regulatory elements (UPEs), which are about 12 bp in length and active in either direction. A well-known example of a UPE is the CCAAT box. The distance between the UPEs and the TATA box is critical for the rate of transcription.

An *enhancer* is a cis-acting element of DNA that stimulates transcription of adjacent genes from the promoter in either direction and over distances up to several kilobases. Viral enhancers were originally identified in SV40 and polyoma viruses and, subsequently, in a variety of others [134]. In addition, enhancer sequences have also been identified in normal cellular genes. Immunoglobulin enhancer sequences have been used to stimulate the transcription of nearby genes cloned on plasmids in lymphoid cells [135]. The ability of these sequences to stimulate transcription appears to be nonspecific because they increase the transcription of a number of different genes. The enhancer element has also been linked to several genes cloned in tandem, resulting in the high-level expression of all of them [136]. Both the enhancers and UPEs must apparently be recognized by one or more different soluble factors to elicit a full-fledged response. Viral enhancer sequences such as those of SV40 and adenovirus are active in most cells, and the level of their activity can be controlled by the availability of these factors. Some enhancers are inducible, for example, by heat shock [137]. The metallothionein enhancer is responsive to heavy metals [138], the β-interferon enhancer responds to viral induction [139], and elements in the long terminal repeat of Mo-MLV [140] and Mo-MSV [141] respond to steroids.

4.3.1.3. Chromosomal Gene Amplification

Plasmids containing sequences coding for dihydrofolate reductase (DHFR) are transfected and integrated into the chromosomes of *dhfr*$^-$ mutants of Chinese hamster ovary (CHO) cells [142]. After selection of *dhfr*$^+$ clones, these clones are placed under the selective pressure of the toxin methotrexate (a competitive inhibitor of DHFR), and the *dhfr* sequences undergo *amplification* (i.e., their copy number is increased several hundredfold). Almost always, the cDNA coding the recombinant protein of interest also undergoes amplification (see also Fig. 7). The resulting increase in gene dosage is accompanied by a simultaneous increase in expression level. This strategy has been employed successfully for high-level expression of recombinant proteins such as erythropoietin [143], blood factor IX [144], and human β-interferon [145].

4.3.2. Insect Systems

The baculovirus *Autographa californica* nuclear polyhedrosis virus (AcMNPCV) has been used to express heterologous genes in insect cells cultured from *Spodoptera frugiperda.* This virus is nonpathogenic to mammals and plants, and demonstrates a wide host range infecting at least 30 different Lepidoptera species. This system is helper-independent and uses the highly expressed promoter of the *A. californica* coat protein, polyhedrin, for expression of foreign genes. The level of expression obtained is consistently higher than that in *Escherichia coli,* yeast, or mammalian cells for a variety of genes [146]. Expression levels ranging from 1 to 500 mg/L have been observed. The system is still in a developmental phase, and several important issues remain unresolved. Polyhedrin expression varies depending on the cell lines infected and the composition of the media. For example, *S. frugiperda* cells accumulate high levels of polyhedrin, whereas the level produced by *Bombyx mori* cells is at the detection limit [146]. Expression of different genes in this system often varies widely. The effects of RNA structure, codon usage, and various control elements on gene expression are being investigated and may lead to improvements in the near future.

One of the major issues is the quality of the protein made by these systems. Insect cells appear to be extremely versatile in their ability to carry out post-translational processes such as secretion [147], localization in cellular organelles

[148] and at cell surfaces [149], proteolytic cleavage [147], phosphorylation [148], N- and O-linked glycosylation [147], [148], myristylation, and palmitylation [149]. Proteins such as human interleukin-2 [150], human β-interferon [147], and human α-interferon [151] have been successfully expressed in this system. Human β-interferon is glycosylated, the pre-β-interferon signal peptide is removed, and the mature protein is efficiently secreted. Human interleukin-2, however, is reported to be secreted in an unglycosylated form [152]. Glycosylation with carbohydrates such as sialic acid, galactose, and fucose may turn out to be inefficient in insect systems [146], especially in mosquito cell lines such as *Aedes aegypti* and *A. albopictus.* Efforts have also been directed at developing a modified baculovirus for direct introduction into live caterpillars. Expression levels are extremely high in this system, and it may prove to be a feasible economical alternative to cell culture [153].

4.3.3. Plant Systems

The availability of Ti (tumor-inducing) plasmid vectors opened up the possibility of studying genetic regulation at a molecular level in plants. Plant systems are adept at processing RNA, carrying out post-translational modification, and targeting proteins to their specific subcellular locations [154].

4.3.3.1. Plasmid Vectors

Development of efficient vectors for the introduction of DNA sequences into plants has been achieved primarily through the use of the Ti plasmid of the soil bacterium *Agrobacterium tumefaciens* [155]. This phytopathogen has evolved a sophisticated method of infecting its plant host cells in which a region of the Ti plasmid called T-DNA is integrated into the plant genome resulting in two specific physiological alterations: (1) diversion of carbon and nitrogen supplies for synthesis of derivatives of basic amino acids called opines that *A. tumefaciens* specifically catabolizes [156] and (2) proliferation of a tumor called crown gall [157]. The T region (T-DNA) of the plasmid contains the tumor-inducing genes [158], whereas the T-*vir* region is responsible for insertion of the plasmid DNA into the host genome [159]. Thus, any foreign DNA inserted in the T region is cotransferred to the plant cells.

Several approaches to introduce genes into plants have been pursued. Because *A. tumefaciens* cannot easily be transformed by naked DNA, alternative strategies had to be developed. Early methods involved insertion of DNA into a transposon such as Tn5 or Tn7 carried on a broad host-range plasmid in *Escherichia coli. Transposons* are mobile genetic elements that move from site to site in a bacterial, yeast, or mammalian genome. The selection of the introduced DNA segment was assured by a cumbersome, time-consuming "homogenotization" procedure [160]. Alternative vectors, such as the cointegrating and binary vectors, were developed to circumvent these problems.

An example of a cointegrating vector is the plasmid pMON200 which carries the pBR322 origin of replication and a selection marker, i.e., a Tn7-derived spectinomycin – streptomycin resistance gene [155]. A triparental mating procedure is used to transfer the new DNA into *A. tumefaciens.* Only the essential regions of T-DNA responsible for integration are retained; the T-DNA regions responsible for tumor formation are eliminated. *Agrobacterium tumefaciens* 3111-SE cells carrying the "disarmed" T-DNA plasmid pTi B6S3-SE are mixed with *E. coli* carrying modified pMN200 and *E. coli* carrying a helper plasmid pRK2013. The helper plasmid provides factors in *trans* that help to mobilize pMN200 into *A. tumefaciens.* The mixture is incubated overnight and then spread on a plate carrying a mixture of spectinomycin and streptomycin to select for the cointegrate plasmid that arises by homologous recombination with the Ti plasmid [161].

Cointegration into the Ti plasmid is an obligatory step for the replication of pMON200.

The binary vector system yields stable transformants [162], [163]. Both the T region and the *vir* region can be isolated from the Ti plasmid without loss of function. Therefore, the T-DNA can be manipulated in *E. coli* and then transformed into *A. tumefaciens* with a helper plasmid that carries the *vir* region. The presence of a complete T region is not essential for functioning of the binary vector system. In binary vectors such as pMON505 [164], the LIH region is re-

placed by the origin of replication derived from the broad host range plasmid pRK290. "Disarmed" T region DNA is equally efficient. Transfer into *Agrobacterium* cells is carried out either by transformation or by a triparental mating procedure using the mobilization helper plasmid as described above.

Gene transfer into plant cells can be carried out by infecting *A. tumefaciens* plant wounds with cultures of *A. tumefaciens* carrying recombinant Ti plasmids [165]. Direct DNA transfer into plant protoplasts can also be carried out by using transformation [166], electroporation [167], liposome fusion [168], microinjection [169], and leaf disk transfer [170] methods. The great advantage of gene transfer into plant protoplasts is that the transformed cells can be regenerated to give whole plants. This provides a powerful tool for carrying out genetic and molecular analysis of transgenic plants and their progeny.

The major advantage of the Ti system is its broad host range. It can induce tumors in most dicotyledonous species. In monocotyledonous plants, DNA transfer and integration take place, but no tumors are formed.

Naked DNA can be used to transform protoplasts derived from monocotyledons and dicotyledons. However, the DNA is scrambled, resulting in incomplete or rearranged copies [171]. Furthermore, plants cannot always be regenerated from protoplasts. Important issues still to be resolved include the effect of integration site on expression; the establishment of site-directed integration systems analogous to those available for yeast, and conversion of a Ti plasmid into an episomal vector by introduction of centromeres, plant autonomously replicating sequences, dominant selection markers, and regulatory promoter systems for expressing genes in specific cell or tissue types.

4.3.3.2. Plant Promoters

The most commonly used plant promoter is the nopaline synthase (*nos*) promoter [172]. Plant promoters possess the familiar TATA box found in most other eukaryotic promoters. Most plant genes placed under their normal promoters are fully active. Many of the promoters are specifically inducible; for example, a short, highly conserved, sequence of 33 base pairs appears to be responsible for the light inducibility of the gene encoding ribulose 1,5-bisphosphate carboxylase in the pea [173]. Similarly, signals that direct tissue-specific expression of the β-phaseolin gene of bean are localized to a 800-bp region at the 5′ end of the gene [174]. Many of these genetic regulatory pathways are being elucidated and offer the potential of using inducible systems to express cloned genes in a tissue-specific fashion.

5. Applications of Recombinant DNA Technology

5.1. Therapeutic Agents

The largest impact of the technological revolution produced by genetic engineering has been on the pharmaceutical industry. Almost every major pharmaceutical house was immediately attracted by the possibility of generating unlimited quantities of rare, valuable, therapeutically important proteins. This also led to a highly productive liaison between academia and industry. Distinguished scientists have joined forces with industrial firms to realize the commercialization of an academic research project in an industrial setting. However, while the replacement or improvement of traditional drug manufacture is attractive, it also requires standardization of cell growth and fermentation processes, as well as training of personnel to handle downstream processing and purification. Many of the companies concerned have invested substantial capital in this area. In addition to developing in-house capabilities, most of them also liaise with smaller biotechnology companies via contract research and product licenses. Although genetic engineering has simplified the manufacture of several established therapeutics such as insulin, growth hormone, serum albumin, and clotting factors, that were previously produced through purification from source tissues, it has also provided a host of new molecules that have substantial promise of being therapeutically important. In addition, development of safer vaccines derived from antigenic determinants such as surface proteins of pathogenic organisms has also become possible. The technology may also be

extended to enhance processes used for everyday pharmaceutical products such as vitamins, steroids, antibiotics, and other bioactive peptides.

Thus far (1988), four genetically engineered drugs have been approved for the United States market by the FDA, namely, human insulin, human growth hormone, human α-interferon, and tissue plasminogen activator.

Insulin derived from porcine or bovine sources was an already established product in the worldwide market for diabetes treatment. The cloned product was developed by Eli Lilly in collaboration with Genentech by using an *Escherichia coli* expression system [175]. The insulin produced by traditional methods is inexpensive and does not have immunogenic tolerance problems. However, submitting recombinant insulin for FDA approval provided industry with the experience needed in obtaining approval for other recombinant DNA products. For a detailed description of insulin, see → Peptide and Protein Hormones.

Interferon was discovered as an antiviral agent [176]. Leukocyte interferons constitute a multigene family, for detailed information see → Interferons [177]. The production of various interferons depends on the nature of infection; not all of them are induced at the same time. The first interferon to be cloned was human leukocyte interferon α-1 [178], followed almost immediately by human leukocyte interferon α-2. Interferon α-2 exhibited a more dramatic antiviral activity in human cells than interferon α-1 which was more potent in bovine cells [179]. Both Hoffman La Roche and Schering Plough proceeded with the expression of human interferon α-2 in *Escherichia coli.* In 1986, interferon was approved by the FDA for use against hairy cell leukemia in the United States and other countries such as the United Kingdom, the Federal Republic of Germany, France, Italy, and Canada. It is also approved for the treatment of Kaposi's sarcoma in the United States, Austria, Denmark, Switzerland, France, and the United Kingdom, as well as for multiple myeloma and condyloma acuminatum in Denmark, Portugal, Ireland, and Argentina.

The complexity of the interferon system was unknown at the time of its discovery. Therefore, cloning of the first interferon was hailed worldwide as a major breakthrough in medical science. Interferon was likened to penicillin as a broad-spectrum antiviral agent. Cloning of the first interferon led to the discovery that interferons are a finely tuned, complex, intricate network of molecules acting either in concert or independently to combat viral infections and modulate immune response. Besides hairy cell leukemia, human interferon α-2 may be therapeutically potent against renal cell carcinoma, warts, and multiple myelomas. Human infections caused by viruses, such as rhinovirus, herpes (eye and generalized infections), adenovirus (eye infections), papilloma virus, and hepatitis B virus, are also responsive to interferon treatment. Growing evidence indicates that this complex system may also protect the human body against certain bacterial and protozoan infections [180].

Human tissue plasminogen activator (t-PA) is a protein with a clear-cut function and is one of the most effective agents in dissolving blood clots. The protein is a multidomain serine protease that converts plasminogen to plasmin, which in turn lyses the fibrin network, thus helping to maintain hemostasis (→ Blood, Chap. 2.2.3.) [181]. It promises to be an effective therapeutic agent for thrombolysis in myocardial infarction, deep vein thrombosis, and similar indications. The cDNA for t-PA was cloned and expressed in CHO cells by Genentech using the DHFR amplification system (see Section 4.3.1.3.) [182]. Tissue plasminogen activator must be produced in mammalian cell lines because of its large size and extensive glycosylation. Clinical studies demonstrate decreased reocclusion rates and lower incidence of bleeding complications in patients treated with recombinant t-PA in comparison to other thrombolytic agents such as streptokinase [183]. Despite initial rejection, by late 1987 Genentech succeeded in gaining approval for t-PA in the United Kingdom, the United States, and Canada.

The systemic half-life of t-PA in humans is only 2–5 min [184]. The efficacious therapeutic dose in humans is, therefore, high and exceeds 100 mg [183]. This causes proteolysis of plasma proteins. A t-PA molecule with a longer half-life could overcome these problems and reduce the production cost per therapeutic dose. Several companies aimed to genetically engineering t-

PA to obtain a second-generation molecule with a longer half-life [185].

Human growth hormone was one of the earliest therapeutically important genes to be cloned [186]. Cloning allowed replacement of the tedious and cumbersome methodology involved in purifying this protein from pituitary gland. Genentech's application for the therapeutic use of human growth hormone was awarded by the FDA in 1986. The FDA also granted approval to Eli Lilly for a modified version of the hormone that lacked the N-terminal methionine residue. This modified hormone is claimed to reduce the antigenic problems posed by the Genentech product, which retains the N-terminal methionine. For a detailed description of growth hormone, see → Peptide and Protein Hormones.

Development of New Products. *Erythropoietin* is a hormone that stimulates precursor bone marrow cells to differentiate into red blood cells. Trials with recombinant erythropoietin indicate positive results for treatment of renal disease associated with anemia [187]. Furthermore, it may also be useful for treating sickle cell anemia, an inherited disorder of the red blood cell [188]. *Colony stimulating factors* are also generating strong interest in the treatment of several hematopoietic deficiencies [189]. Several of these have been cloned and are at various stages of development. Granulocyte colony stimulating factor (G-CSF) [190], granulocyte macrophage colony stimulating factor (GM-CSF) [191], macrophage colony stimulating factor (M-CSF or CSFI) [192], interleukin-3 (multi-CSF) [193], and interleukin-4 (B-cell stimulating factor 1) [194] are all being developed as products that could restore immunocompetence when used as an adjunct to chemotherapy. By the same criteria, these factors could prove to be extremely useful in treatment of systemic infections or burns and in bone marrow transplantation.

Interleukin-2 (IL-2) is also in advanced stages of testing, especially for cancer therapy in combination with lymphokine-activated killer cells or tumor-infiltrating lymphocytes. As with the interferons, the cloning of several of the cytokines has shed light on their role in the body's immune defense system. Interleukin-2, for example, appears to be the major inducer of γ-interferon. Interleukin-1, tumor necrosis factor, and γ-interferon are believed to play a role in several human diseases such as arthritis, diabetes, and atherosclerosis [195]. Therefore, interest exists in using the cloned products to set up screens for antagonist molecules.

Other products in the developmental stage are *growth factors* such as the platelet derived growth factor (PDGF), epidermal growth factor (EGF), and fibroblast growth factor (FGF) mainly in the areas of wound healing and angiogenesis. *Blood clotting factors,* such as tissue factor and factors VII, VIII, and IX, have been cloned and are in the developmental stage. Factor VIII is already being studied in clinical trials for inducing clotting in hemophiliacs [196]. *Hormones* such as thrombopoietin, follicle stimulating hormone [197], gonadotropin [198], and somatotropin [199] have been cloned and are currently being investigated (→ Peptide and Protein Hormones).

5.2. Vaccines

Control of deadly diseases such as smallpox, cholera, and polio can be attributed mainly to vaccines. For a detailed description of vaccines, see → Immunotherapy and Vaccines. Conventional methods of vaccine production have saved millions of lives. Despite the large market for vaccines, few companies have entered the field because they are intimidated by the immense legal costs. Genetic engineering, however, provides a safe alternative because it rules out vaccination by live or attenuated organisms and also offers the possibility of producing vaccines to control diseases that are beyond the scope of traditional methodology. Isolated gene products, especially surface antigens, can elicit an immune response from the body.

Several vaccines for use in humans are currently being investigated, some of the more important being those against hepatitis B, malaria, polio, herpes, gonorrhea, influenza, leprosy, and acquired immune deficiency syndrome (AIDS).

Hepatitis B Vaccine. The hepatitis B vaccine is the first genetically engineered vaccine to enter the U.S. market. The vaccine is produced by expressing the cloned hepatitis B virus surface antigen (HbsAg) in *Escherichia coli* and

yeast. The product physically and antigenically resembles the virus particles found in the serum of infected patients [200], [201]. Its efficacy as an immunization agent against hepatitis B infection was demonstrated in both chimpanzees [202] and humans [203].

Malaria Vaccine. The malaria vaccine is a prime example of what genetic engineering can accomplish where conventional approaches have failed. Because of the inherent antigenic diversity of the malaria parasite, the whole organism cannot be used as a source of vaccine. Cloning of some of the genes of this parasite has shown that some gene products may be invariant and, therefore, suitable candidates for a vaccine. An example of one such apparently invariant protein is the repeat sequence epitope from the circumsporozoite gene of *Plasmodium falciparum* [204]. A vaccine based on this sequence is currently undergoing clinical trials. If successful, it will replace or augment methods such as mosquito control and chemotherapy that are currently used to check the spread of this disease.

Acquired Immune Deficiency Syndrome. Acquired immune deficiency syndrome (AIDS) is caused by a retrovirus called the human immunodeficiency virus (HIV), which has been cloned and sequenced [205]. The virus is spread via body fluids, mostly by sexual intercourse or the use of contaminated syringes and needles among drug addicts. The alarming rate at which this fatal disease is spreading has caused great anxiety in the world health and medical community. AIDS has become the world's most deadly infectious disease, ranking fourth after ischemic heart disease, strokes and acute respiratory infections in terms of mortalities (2.28 million deaths worldwide in 1998).

A number of pharmaceutical and biotechnology firms are working closely with academic institutions to find methods to control this disease. A synthetic peptide fragment, encompassing the amino acid sequence from residues 735 to 752 of the viral envelope protein subunit gp41 was used to produce antibodies. The resulting antiserum reacted with the envelope subunit [206], indicating that viral envelope proteins are very immunogenic and may be used as a vaccine. A vaccine based on the recombinant version of the AIDS surface antigen protein gp160 produced in insect cells is already undergoing clinical trials [207]. The gp120 envelope protein of HIV-1 is also a prime target for use as a vaccine. Another approach is being followed by Smith Kline and French Laboratories and Genentech to develop an AIDS vaccine. Here, "soluble receptor" molecules are used that bind tightly to the virus, thus preventing its interaction with a lymphocyte or a brain cell [208].

Polio Vaccine. A novel approach has been used by Connaught Research Laboratories to develop a candidate polio vaccine [209]. An antigenic epitope for poliovirus type 3 was inserted into the 3′ end of the tobacco mosaic virus coat protein and expressed in *Escherichia coli.* Assembly of this protein was expected to produce a highly immunogenic, high molecular mass complex that carries several copies of the antigenic epitope. When injected into rats, this complex elicited a strong immune response that induced formation of neutralizing antibodies to poliovirus [209].

Other Vaccines. Production of vaccines via genetic engineering provides a method of obtaining proteins that undergo correct post-translational modification (e.g., glycosylation) by expressing them in mammalian cell lines. These modifications may be required for the protein to attain its native configuration. This is true in the case of the herpes simplex virus glycoprotein D, which when expressed in Chinese hamster ovary cells [210] can protect guinea pigs against genital viral infection.

Vaccinia virus vectors for polyvalent vaccines have also been developed [211]. Although this approach may allow one-shot immunization against several diseases, it is not without risk, because wide-scale use may elicit a severe response in patients with low immunocompetence. The safest vaccines are those that contain no DNA whatsoever, and the only way to achieve this is through synthetic and genetically engineered vaccines.

5.3. Genetically Altered Organisms

5.3.1. Transgenic Animals

Introduction of the human growth hormone gene under the control of metallothionein promoters

in mice was accomplished in 1983 [212]. This experiment caused a sensation because the mice inheriting the gene were much larger in size than control animals. Two methods are employed for introduction of foreign DNA into the fertilized mouse egg: (1) retroviral-mediated transfer and (2) direct microinjection into the male pronucleus of the fertilized egg. The DNA is integrated randomly and may form concatemers. Improved methods for targeting DNA to specific sites in the chromosome are still being perfected. Transgenic mice are widely used to study DNA sequences that mediate tissue- or cell-specific gene expression [137], [138]. In addition, they are employed as a routine tool to study oncogenes [213] and to set up animal models for genetically determined disease states [214].

The success achieved with transgenic mice promoted experiments with other animals such as fish [215], rabbits, sheep, and pigs [216], with a view to carrying out studies in improvement of livestock. One of the earliest experiments utilized human growth hormone in pigs. Despite high expression of the gene, no evidence of enhanced growth appeared [217]. Instead, the pigs developed arthritis-like symptoms. Growth hormone is normally produced in a highly regulated fashion by the pituitary gland. The development of pathological conditions was probably due to unregulated expression of the gene in all tissue types. When porcine growth hormone is expressed in a regulated fashion in transgenic pigs, their rate of growth is increased substantially. Furthermore, the meat is lean. Steps are being taken to commercialize breeding lines of such transgenic pigs [218].

The fertility of Booroola merino sheep increased severalfold upon introduction of the fertility (F) gene into transgenic sheep [218].

The latest use of transgenic animals has been in the production of heterologous proteins in milk. Using promoters of genes coding for milk-specific proteins (e.g., β-lactoglobulin and whey acid protein), researchers have expressed blood factor IX [219] and t-PA [129] in the milk of transgenic sheep and mice, respectively. Sheep β-lactoglobulin has also been expressed in transgenic mice, which produced up to 23 mg/mL of β-lactoglobulin in their milk [219].

Transgenic animals may soon become an attractive alternative for the production of genetically engineered, biologically active proteins. Several regulatory issues would have to be resolved, but expression levels and the quality of the product may excel those achieved via conventional methods.

5.3.2. Transgenic Plants

The concept of manipulating plant genetics has direct appeal for the agricultural industry. It provides a tool for introducing specific genes to protect plants against potentially lethal agents such as herbicides, viruses, and insects. Other advantages include the improvement of crop yield and quality.

The development of plasmid vectors carrying selectable markers for tranformation of plant cells brought about a revolution in plant breeding. By using the soil bacterium *Agrobacterium tumefaciens* with its associated Ti plasmid as a vector (cf. Section 4.3.3.1.), several transgenic strains of plants were generated. Kanamycin resistance carried on the *A. tumefaciens* plasmid is inherited stably in progeny of transgenic plants [154]. The host range of the *A. tumefaciens* is quite large among the dicotyledonous species ranging from petunias [220] to pines (e.g., the Douglas fir [221]), deciduous species (e.g., poplar [222]), and cotton [223]. Transfer of single gene traits has met with considerable success. Transgenic tobacco plants that carried a mutant *aro A* gene from *Salmonella typhimurium* conferred tolerance to the herbicide glyphosate [224]. The *aro A* locus codes for 5-enolpyruvyl-3-phosphokimate synthetase which is inhibited by the herbicide glyphosate. The mutant enzyme is resistant to the herbicide. By engineering a *Streptomyces* acetylating enzyme gene into tobacco, tomato, and potato plants, resistance to the herbicide Basta has been achieved [220]. Acetylation of the herbicide renders it ineffective. Resistance to tobacco mosaic virus was achieved in tomato and tobacco plants that were engineered to carry tobacco mosaic virus coat protein [225]. The presence of the coat protein provides cross-protection against superinfecting viruses because it encapsidates (packages) the viral RNA and renders it noninfectious. Another route to achieving viral resistance has been through transformation of tobacco plants with a satellite RNA species derived from a benign strain of cucumber mosaic virus. The pres-

ence of this RNA protects the plant from virulent infection [226]. Resistance against several insect pests has been achieved by creation of transgenic plants carrying a gene encoding cowpea trypsin inhibitor [227]. Proteinase inhibitors defend plants by inhibiting the proteolytic activity of the pathogens; the pathogens require this activity for the breakdown of plant protein and thus for their nutrition.

Direct transfer of naked DNA into protoplasts has also been employed in plant species that do not fall within the host range of *A. tumefaciens* [173]. This method is used particularly for monocotyledonous species, including cereals, which constitute the most important group of plants providing nutritional needs in humans. Protoplast fusion [228], liposome-mediated DNA transfer [229], and electroporation [230] have been used to transform these plant cells. Problems concerning the successful regeneration of plants from protoplasts have been partially solved, at least in the case of rice [231]. Attempts to repeat this success with other cereals such as maize, wheat, and barley are under way.

Although transfer of single gene traits has been accomplished successfully, many desirable traits such as disease resistance, stress tolerance, and photosynthetic efficiency probably depend on the activity of several genes. Intensive effort must, therefore, be mounted to characterize these genes. The availability of vectors, improved methods of gene transfer, and improvement in plant regeneration techniques have opened the door to what may be a revolution in plant breeding.

5.4. Gene Therapy

Gene therapy is considered by many to be the ultimate frontier still to be conquered by recombinant DNA technology. Despite intense research, progress has been slow. Gene therapy, as currently envisaged, differs from creation of transgenic animals in that it involves somatic gene manipulation as opposed to germ line manipulation. Work thus far has been restricted to laboratory animals. Approaches used for gene transfer include retroviral-mediated transfer [232], microinjection [233], electroporation [234], uptake of DNA precipitated by calcium phosphate [233], and homologous recombination [235]. The most popular of these methods is retroviral transfer. Concern exists over the use of retroviral vectors, and they must be shown not to pose a cancer hazard as a result of either recombination events or random insertion into (and activation of) oncogenes.

Hemopoietic stem cells have been the favorite candidates for gene transfer experiments. After retroviral transduction, the cells are transplanted into mice that have been lethally irradiated to eliminate existing stem cells. Expression of the introduced gene is checked four to six months later to allow for differentiation of stem cells [236]. Another method utilizes cultured murine fibroblasts; the cells are transfected with the desired DNA and implanted at several anatomical sites [237]. Expression of the gene of interest can potentially be regulated, depending on the site of implantation. This type of implantation has been termed "transkaryotic implantation."

Most of the early experiments used the retroviral LTRs for expression of the foreign DNA. The major problem has been low expression in the reconstituted cells. The LTR-based expression seems to be irreversibly blocked, especially in bone marrow cells. A breakthrough has been achieved [238]. A retroviral vector carrying the β-globin gene under the control of its natural promoter was transfected into hemopoietic stem cells. Measurable quantities of β-globin were found in all mice that had undergone bone marrow transplantation. These results illustrate the importance of the choice of promoters to achieve tissue-specific expression.

The aim of the above studies is to provide basic knowledge so that experimentation can be extended to humans. The most obvious candidates for gene therapy in humans are genetic hemoglobin disorders, the best known and most commonly inherited human diseases. Other candidates are diseases caused by inborn errors of metabolism such as the Lesch – Nyhan syndrome, which is due to a deficiency of hypoxanthine phosphoribosyltransferase, and severe combined immune deficiency syndrome, which is caused by defective production of adenosine deaminase.

Gene therapy is still viewed with skepticism and concern by the public. The FDA and the NIH Recombinant DNA Advisory Committee intend to scrutinize all gene therapy proposals closely

prior to approval [239]. However, the following criteria proposed by MULLIGAN may help researchers choose a disease suitable for gene therapy [240]:

1) The disease must be life-threatening and incurable.
2) The normal counterpart of the defective gene must be isolatable.
3) The organ or tissue responsible for disease-related pathophysiology must be identified.
4) Introduction of corrective sequences into a significant fraction of the affected system should be achievable.
5) The gene must be adequately expressed.

Although, in theory gene therapy should work, this goal has not been achieved in reality. The greatest detriments to progress have been inadequate expression levels, an inability to infect large numbers of cells, and a lack of methods for identifying the derived cell population in a mixed population of cells.

5.5. Diagnostics

One of the offshoots of streamlining DNA hybridization techniques is the establishment of diagnostics based on DNA probe technology. By using the exquisite specificity with which complementary sequences on single-stranded DNA or RNA find each other, cloned gene fragments are employed to detect infectious agents, hereditary markers, oncogenes, and antibiotic resistance genes. Currently, efforts are focused on replacing labor-intensive, time-consuming DNA hybridization methods with simpler and faster techniques. This is essential for successful commercialization of this technology, especially if it is to compete favorably with the immunodiagnostic market. Several of the currently employed tests use radioactively labeled probes. One of the first assays to be marketed was the *Salmonella detection* assay that utilizes a radioactive probe [241]. Simplification of DNA immobilization techniques and proprietary reactions that accelerate hybridization have helped to reduce the time required for such assays. The FDA approved the first isotopic probe in 1987 for the diagnosis of legionnaires' disease by use of patient specimens [241].

Because the use of radioactive isotopes is a cause for concern, especially in clinical laboratories that handle thousands of samples, many companies are trying to find nonisotopic tags. Suitable alternatives include fluorophore-based tests and enzymes linked to the probe either directly or via a biotin – avidin complex. The fluorophore-based tests are best suited to automation; inexpensive, easy-to-use fluorometers are being developed. Enzyme-based assays use dye conversion for detection, and these can be read on colorimeter or spectrophotometer. The low sensitivity of the nonisotopic methods presents a severe problem in that assay time is prolonged because the samples often require further culturing. Cetus has developed a selective DNA amplification technique called the polymerase chain reaction (PCR) that can amplify given DNA samples a million times [242]. This method relies on the hybridization of an oligomeric primer to a single-stranded DNA molecule and the ability of the DNA polymerase to complete the synthesis of the second strand. The double-stranded DNA is then denatured, and the whole process is repeated several times, resulting in a 200 000-fold amplification of the DNA sequence. This could overcome problems of low DNA content in samples and has been shown to be effective for detection of HIV-1 in DNA of peripheral blood mononuclear cells [243]. Several companies are involved in automating these assays.

Besides detecting infectious bacteria and viruses, DNA probe technology provides the means for diagnosing several inherited single gene disorders. Genetic screening by this method is based on analysis of restriction fragment length polymorphism (RFLP). Polymorphism in the DNA sequence is quite high (one in every 200 bp in the human genome). Because of the variance in DNA sequence, recognition sites for restriction endonucleases may also differ giving rise to altered patterns of restriction fragment lengths. Distinctive RFLP patterns can be linked to several genetic diseases such as thalassemia [244] or insulin [245] and growth hormone disorders [246]. The DNA from the patient's blood is digested with a specific restriction endonuclease, and the RFLP pattern is obtained by hybridization with a marker gene. Because deletions or mutations within the gene can produce an altered RFLP pattern, this type of

analysis can be used to predict the inheritance of a disease. A number of RFLP markers associated with specific diseases such as cystic fibrosis, sickle cell anemia, Duchenne-Type muscular dystrophy, thalassemia, hemophilia, retinoblastoma, and Huntington's chorea have been identified [247]. Some of these tests are available commercially, but still expensive (cost ranging from $ 800 to $ 1500) and time-consuming (8 – 10 d). Tests are also available for Philadelphia chromosome, chronic myelogenous leukemia, and T- or B-cell cancers.

6. Regulatory and Safety Aspects

Following the recommendations drawn up after the Asilomar meeting [248], the National Institutes of Health (NIH) set up the Recombinant DNA Advisory Committee (NIH RAC) in 1976. This committee proposed several levels of biological containment for experiments involving gene manipulation [249], with rigorous containment procedures. In addition, Institutional Biosafety Committees (IBCs) were set up by pharmaceutical companies and academic institutions to govern the conduct of everyday recombinant DNA research. Today, most countries involved in recombinant DNA research have implemented nationally monitored programs to control research in genetic engineering. Experience has shown that most genetic engineering experiments pose no greater threat than does conventional microbiological research. Consequently, containment levels have been relaxed in recent years. "High-risk" experiments such as those involving cloning of toxins and lethal viral genomes into harmless organisms are still strictly regulated. A surge of activity occurred in the study of transgenic animals. Some work in this area has also necessitated establishment of rigorous codes to prevent these animal populations from entering the common gene pool. For instance, the creation of transgenic mice carrying the entire genome of the AIDS virus has caused concern [250]. As a result, stringent measures have been advocated to prevent accidental release of these animals.

Another issue discussed frequently is the use of recombinant products as therapeutic agents. The main concern here is the authenticity of the product. Many factors such as codon usage, glycosylation, post-translational modification, protein folding, and methionyl derivatives (heterologous proteins expressed cytoplasmically in *Escherichia coli* that retain the N-terminal methionine) can all change the functionality of a protein. Some changes may be apparent immediately, but others may escape notice because of methodological limitations of the assay used.

An additional problem is the homogeneity of the preparation. Rigorous control of purification schemes used in the isolation of a recombinant product is required. The final product should be free of contaminants such as proteolytic fragments, host pyrogens, and host nucleic acid, which can elicit an immune respone in patients [251]. This may be deleterious to the efficacy of the drug or the general health of the patient. Therefore, studies must determine whether the resultant antibodies are specific, neutralizing, or associated with hypersensitivity. The use of cell lines and transgenic animals that synthesize the desired protein as production hosts is proposed as a method of overcoming these problems. Despite these problems, a number of recombinant products such as insulin, α-interferon, human growth hormone, and t-PA have made their way into the market and are being used successfully.

The application of recombinant DNA technology in agriculture has created microbes that protect plants from natural threats such as frost and insect pests, as well as transgenic plants engineered to withstand harmful herbicides and viruses. Controversy over the release of such organisms into the environment has reached dramatic proportions. Of primary concern are the mixing of gene pools in the environment and the inadvertent introduction of such "hardy" characteristics as herbicide resistance into "weedy" plants, with disastrous consequences. The NIH guidelines deal with the issue of field testing and deliberate release of such organisms with great caution on a case-by-case basis. The Environmental Protection Agency also provides recommendations in this case. Field tests for plants are closely monitored to ensure that they are grown in a controlled, isolated environment to prevent cross-pollination or their permanent establishment.

The benefits of engineered organisms for agriculture, health, and environmental purposes are obvious. The enormous promise of this technology is bound to be realized in coming years.

As in the past, all new issues will have to be addressed with prudence and wisdom so that rational guidelines for conducting such research can be drawn up and the fears of the general public can be allayed.

7. Glossary of Special Terms

amplification: generally refers to an increase in the copy number of a gene (this can sometimes be induced by selection pressure, as in DHFR-based amplification)

copy DNA: DNA strand synthesized on an mRNA template by the enzyme reverse transcriptase

cosmid: hybrid vector composed of part phage and part plasmid

enhancer: DNA sequence that can increase transcriptional initiation, sometimes even at a considerable distance from the promoter

episome: usually a double-stranded circular DNA molecule that can exist extrachromosomally within a cell

expression: generally refers to the biosynthesis of a protein directed by cellular components (mRNA, tRNA, ribosomes, etc.)

fusion protein: generally refers to a recombinant protein expressed as part of a protein that is abundantly expressed in the host organism (e.g., lac, Trp)

library: cDNA or genomic DNA, is a population of chimeric DNA molecules representing the genomic content or the mRNA content of a specific cell

ligation: the process that involves in vitro joining of double-stranded DNA molecules by the enzyme T4 DNA ligase

operon: cluster of genes that encode a specific biosynthetic pathway or function (e.g., lactose, heat-shock, or tryptophan operon)

phage: bacterial virus (e.g., λ, M13, T4, T7)

plasmid: double-stranded circular DNA molecule that can exist extrachromosomally and replicate autonomously

promoter: conserved sequence that is recognized by the host-specific RNA polymerase to initiate mRNA synthesis

replicon: an extrachromosomal DNA or RNA molecule that can replicate independently of the genome (automous replication)

repressor: protein molecule that can bind to a specific conserved sequence, the operator sequence, in certain promoter systems to turn off transcription

restriction endonuclease: enzyme that recognizes a specific sequence and catalyzes the formation of cuts in double-stranded DNA; used as a tool in genetic engineering

retrovirus: RNA virus that uses a DNA intermediate for replication

selection marker: gene such as drug resistance or an auxotrophic requirement used to select for a plasmid in a particular host background

shuttle vector: plasmid that has the requisite DNA sequences to permit its existence in more than one host such as bacteria and mammalian cell lines

transfection: introduction of naked DNA, generally of viral origin, into a cell

transformation: introduction of naked DNA, generally of bacterial origin, into a cell

translation: step in protein biosynthesis that involves the transfer of information from mRNA into the synthesis of a protein

transposon: genetic element capable of moving from site to site within a genome

8. References

General References

1. S. M. Kingsman, A. J. Kingsman: *Genetic Engineering: An Introduction to Gene Analysis and Exploitation in Eukaryotes,* Blackwell Scientific Publ. Ltd., Osney Mead, Oxford 1988.
2. R. A. Zilinkas, B. K. Zimmerman (eds.): *The Gene-Splicing Wars: Reflections on the Recombinant DNA Controversy,* MacMillan, Inc. for AAAS, Washington, DC 20005, 1988.
3. E. L. Winnacker (ed.): *From Genes to Clones, Introduction to Gene Technology,* VCH Publishers, Weinheim 1987.
4. W. Reznikoff, L. Gold (eds.): *Maximizing Gene Expression,* Butterworths, Boston 1986.
5. R. Rodriguez (ed.): *Vectors: A Survey of Molecular Cloning Vectors and Their Uses,* Butterworths, Boston 1988.

6. S. L. Berger, A. R. Kimmel (eds.): *Methods in Enzymology: Guide to Molecular Cloning Techniques,* vol. 152, Academic Press, New York 1987.
7. M. Inouye (ed.): *Experimental Manipulation of Gene Expression,* Academic Press, New York 1984.
8. C. T. Caskey, R. L. White (eds.): *Recombinant DNA Applications to Human Disease,* Banbury Report 14, Cold Spring Harbor Laboratory, New York 1983.

Specific References

9. O. T. Avery, C. M. Macleod, M. McCarty, *J. Expl. Med.* **79** (1944) 137.
10. J. D. Watson, F. H. C. Crick, *Nature (London)* **171** (1953) 736.
11. J. G. Wetmur, N. Davidson, *J. Mol. Biol.* **31** (1968) 349.
12. D. Nathans, H. O. Smith, *Ann. Rev. Biochem.* **44** (1975) 273.
13. F. H. C. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, *Nature (London)* **192** (1961) 1227.
14. P. E. Lobban, A. D. Kaiser, *J. Mol. Biol.* **78** (1973) 453.
15. S. N. Cohen, A. C. Y. Chang, H. W. Boyer, R. B. Helling, *Proc. Nat. Acad. Sci. USA* **70** (1973) 3240.
16. P. H. Pouwels, B. E. Enger-Volk, W. J. Brammar (eds.): *Cloning Vectors,* Elsevier Publications, New York 1985.
17. S. J. O'Brien (ed.): *Genetic Maps: A Compilation of Linkage and Restriction Maps of Genetically Studied Organisms,* Cold Spring Harbor Laboratory, New York 1984.
18. J. Weigle, *Proc. Nat. Acad. Sci. USA* **55** (1966) 1462.
19. F. Bolivar, R. L. Rodriguez, M. C. Betlach, H. W. Boyer, *Gene* **2** (1977) 75.
20. R. W. Hendricks, J. W. Roberts, F. W. Stahl, R. A. Weisberg: *Lambda II,* Cold Spring Harbor Laboratory, New York 1983, pp. 3 – 19.
21. B. G. Williams, F. R. Blattner, *J. Virol.* **29** (1979) 555.
22. T. V. Huyh, R. A. Young, R. W. Davis in D. Clover (ed.): *DNA Cloning Techniques: A Practical Approach,* IRL Press, Oxford 1985 p. 49.
23. N. E. Murray, *Mol. Gen. Genet.* **150** (1977) 53.
24. R. A. Young, R. W. Davis, *Proc. Nat. Acad. Sci. USA* **80** (1985) 1194.
25. J. H. Miller (ed.): *Experiments in Molecular Genetics,* Cold Spring Harbor Laboratory, New York, 1972, p. 47.
26. T. Maniatis, E. F. Fritsch, J. Sambrook: *Molecular Cloning. A Laboratory Manual,* Cold Spring Harbor Laboratory, New York, 1982, p. 17.
27. J. Collins, B. Hohn, *Proc. Nat. Acad. Sci. USA* **75** (1978) 4242.
28. J. Vieira, J. Messing, *Gene* **19** (1982) 259.
29. H. Temin, D. Baltimore, *Adv. Virus Res.* **17** (1972) 129.
30. A. Efstratiadis, F. C. Kafatos, A. M. Maxam, T. Maniatis, *Cell* **7** (1976) 279.
31. U. Gubler, B. J. Hoffman, *Gene* **25** (1983) 263.
32. F. J. Bollum, *The Enzymes,* **vol.** 10, Academic Press, New York 1974, p. 145.
33. L. Villa-Komaraft et al., *Proc. Nat. Acad. Sci. USA* **75** (1978) 3727.
34. P. H. Johnson, M. Laskowski, *J. Biol. Chem.* **245** (1970) 891.
35. H. Okayama, P. Berg, *Mol. Cell. Biol.* **2** (1982) 161.
36. H. Okayama, P. Berg, *Mol. Cell. Genet.* **3** (1983) 280.
37. W. D. Benton, R. W. Davis, *Science* **196** (1977) 180.
38. M. Grunstein, D. S. Hogness, *Proc. Nat. Acad. Sci. USA* **72** (1975) 3961.
39. E. M. Southern, *J. Mol. Biol.* **98** (1975) 503.
40. P. S. Thomas, *Proc. Nat. Acad. Sci. USA* **77** (1980) 5201.
41. J. F. Burke, *Gene* **30** (1984) 63.
42. P. W. J. Rigby, M. Dieckmann, C. Rhodes, P. Berg, *J. Mol. Biol.* **113** (1977) 237.
43. H. Towbin, T. Staehelin, J. Gordon, *Proc. Natl. Acad. Sci. USA* **76** (1979) 4350.
44. A. Maxam, W. Gilbert, *Proc. Nat. Acad. Sci. USA* **74** (1977) 560.
45. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **74** (1977) 5463.
46. J. M. Prober et al., *Science (Washington, D.C., 1883-)* **238** (1987) 336.
47. D. V. Goeddel et al., *Nature (London)* **287** (1980) 411.
48. D. V. Goeddel et al., *Nucleic Acids Res.* **8** (1980) 4057.
49. J. van Brant, *Bio/Technol.* **4** (1986) 701.
50. R. C. Dickson, J. Abelson, W. M. Barnes, W. S. Reznickoff, *Science* **182** (1975) 27.
51. G. N. Bennett et al., *Proc. Nat. Acad. Sci. USA* **73** (1976) 2351.
52. V. Pirotta, *Nature (London)* **254** (1975) 114.
53. A. R. Stuitje, G. E. Spelt, E. Veltkamp, H. J. J. Nijkamp, *Nature (London)* **290** (1981) 264.
54. B. E. Uhlin, K. Nordstrom, *Mol. Gen. Genet.* **165** (1978) 167.

55. N. Panayatotos, *Nucleic Acids Res.* **12** (1984) 2641.
56. K. Gerdes, P. B. Rasmussen, S. Molin, *Proc. Nat. Acad. Sci.* **83** (1986) 3116.
57. J. Nilsson, S. G. Skogman, Bio/Technol. **4** (1986) 901.
58. H. A. deBoer, L. J. Comstock, M. Vasser, *Proc. Nat. Acad. Sci. USA* **80** (1983) 21.
59. K. Backman, Bio/Technol. **2** (1984) 1045.
60. F. W. Studier, B. A. Moffat, *J. Mol. Biol.* **189** (1986) 113.
61. G. D. Stormo, T. D. Schneider, L. M. Gold, *Nucleic Acids Res.* **10** (1982) 2971.
62. H. A. deBoer, R. A. Kastelein in W. Reznikoff, L. Gold (eds.): *Maximizing Gene Expression,* Butterworths, Stoneham, MA 02180, United States 1986 p. 225.
63. N. Panayotatos, K. Truong, Nucleic Acids Res. **13** (1985) 2227.
64. B. E. Schoner, H. M. Hsiung, R. M. Belagaje, N. G. Mayne et al., *Proc. Nat. Acad. Sci. USA* **81** (1984) 5403.
65. J. Ghrayeb, H. Kimura, M. Takahara, H. Hsiung et al., *EMBO J.* **3** (1984) 2437.
66. T. Miyake, *J. Biochem.* **97** (1985) 1429.
67. K. Erb, M. Vogel, W. Wagner, W. Goebel, *Mol. Gen. Genet.* **208** (1987) 88.
68. L. Abrahmsen, T. Moks, B. Nilsson, M. Uhlen, *Gene* **14** (1986) 7487.
69. C. H. Chung, A. L. Goldberg, *Proc. Nat. Acad. Sci. USA* **78** (1981) 4931.
70. T. A. Baker, A. D. Grossman, G. A. Gross, *Proc. Nat. Acad. Sci. USA* **81** (1984) 6779.
71. S. H. Shen, *Proc. Nat. Acad. Sci. USA* **81** (1984) 4627.
72. J. Germino, D. Bastia, *Proc. Nat. Acad. Sci. USA* **81** (1984) 4692.
73. M. Matteucci, H. Lipesky, Bio/Technol. **4** (1986) 51.
74. T. J. Gryczan, S. Contente, D. Dubnau, *J. Bacteriol.* **134** (1978) 318.
75. T. J. Gryczan, D. Dubnau, *Proc. Nat. Acad. Sci. USA* **75** (1978) 1428.
76. J. Scheer-Abramowitz, T. J. Gryczan, D. Dubnau, *Plasmid* **6** (1981) 67.
77. S. B. Primrose, S. D. Ehrlich, *Plasmid* **6** (1981) 193.
78. L. Band, D. J. Henner, *DNA* **3** (1984) 17.
79. S. R. Fahnestock, K. E. Fisher, *J. Bacteriol.* **165** (1986) 764.
80. I. Palva et al., *Gene* **22** (1983) 229.
81. I. Palva et al., *Proc. Nat. Acad. Sci. USA* **79** (1982) 5582.
82. N. Vasantha, L. D. Thompson, *J. Bacteriol.* **165** (1986) 837.
83. T. Imanaka, T. Tanaka, H. Tsunekawa, S. Aiba, *J. Bacteriol.* **147** (1981) 776.
84. M. F. Tuite et al., *EMBO J.* **1** (1982) 603.
85. R. A. Hitzeman et al., *Nature (London)* **293** (1981) 717.
86. R. Derynk, A. Singh, D. V. Goeddel, *Nucleic Acids Res.* **11** (1983) 1819.
87. F. Watts, C. Castle, J. Beggs, *EMBO J.* **11** (1983) 2085.
88. R. A. Smith, M. J. Duncan, D. T. Moir, *Science* **229** (1985) 1219.
89. J. L. Bennetzen, B. D. Hall, *J. Biol. Chem.* **257** (1982) 3026.
90. J. R. Broach et al. (eds.): *The Molecular Biology of the Yeast Saccharomyces: Life Cycle and Inheritance,* Cold Spring Harbor Laboratory, New York, 1981, p. 445.
91. G. Tschumper, J. Carbon, *Gene* **10** (1980) 157.
92. G. Tschumper, J. Carbon, *J. Mol. Biol.* **156** (1982) 293.
93. D. T. Stinchcomb, C. Mann, E. Selker, R. W. Davis, *ICN-UCLA Symp. Molecular Cellular Biology* **22** (1981) 473.
94. D. T. Stinchcomb, K. Struhl, R. W. Davis, *Nature (London)* **282** (1979) 39.
95. A. W. Murray, J. W. Szostak, *Cell* **34** (1988) 961.
96. L. Clarke, J. Carbon, *Nature (London)* **287** (1980) 504.
97. K. S. Bloom, J. Carbon, *Cell* **29** (1982) 305.
98. T. L. Orr-Weaver, J. W. Szostak, R. J. Rothstein: *Methods in Enzymology,* Academic Press, New York 1983 pp. 101, 228.
99. T. L. Orr-Weaver, J. W. Szostak, *Curr. Genet.* **6** (1983) 19.
100. K. J. Shaw et al., *DNA* **7** (1988) 117 – 226.
101. J. W. Szostak, E. H. Blackburn, *Cell* **29** (1982) 245.
102. A. Miyanohara et al., *Proc. Nat. Acad. Sci. USA* **80** (1983) 1.
103. T. P. St. John, R. W. Davis, *J. Mol. Biol.* **152** (1981) 285.
104. G. A. Bitter, K. M. Egan, *Gene* **32** (1984) 263.
105. A. J. Brake et al., *Proc. Nat. Acad. Sci. USA* **81** (1984) 4642.
106. A. Miyajima et al., *EMBO J.* **5** (1986) 1193.
107. A. Miyajima et al., *Gene* **37** (1985) 155.
108. J. M. Gregg et al., Bio/Technol. **5** (1987) 479.
109. D. J. Ballance, G. Turner, *Gene* **36** (1985) 321.
110. D. Cullen et al., *Bio/Technol.* **5** (1987) 369.
111. E. Goldwasser: "Blood Cell Differentiation," *Fed. Proc.* **34** (1975) 2285.
112. G. Simons et al., *Gene* **28** (1984) 55.
113. T. Yokota et al., *Proc. Nat. Acad. Sci. USA* **82** (1985) 68.

114. M. Goto et al., *Bio/Technol.* **6** (1988) 67.
115. D. S. Anson, D. E. G. Austen, G. C. Brownlee, *Nature (London)* **315** (1985) 683.
116. P. J. Southern, P. Berg, *J. Mol. Appl. Genet.* **1** (1982) 327.
117. Y. Gluzman, Cell **23** (1981) 173.
118. M.-F. Law, D. R. Lowy, I. Dvoretzky, P. M. Howley, *Proc. Nat. Acad. Sci. USA* **78** (1981) 2727.
119. B. Binetruy, G. Meneguzzi, R. Breathnach, F. Cuzin, *EMBO J.* **5** (1982) 621.
120. N. Sarver et al., *Mol. Cell. Biol.* **1** (1981) 486.
121. D. DiMaio, V. Corbin, E. Sibley, T. Maniatis, *Mol. Cell. Biol.* **4** (1984) 340.
122. N. Hsiung et al., *J. Mol. Appl. Genet.* **2** (1984) 497.
123. R. Fukunaga, Y. Sokawa, S. Nagata, *Proc. Nat. Acad. Sci. USA* **81** (1984) 5086.
124. T. Kusano et al., *Proc. Nat. Acad. Sci. USA* **84** (1987) 1789.
125. E. Gilboa, S. W. Mitra, S. Goff, D. Baltimore, *J. Virol.* **41** (1979) 271.
126. J.-F. Nicolas, J. L. R. Rubenstein: *Vectors,* Butterworths, Stoneham, MA 02180, United States 1987, p. 498.
127. D. A. Williams, I. R. Lemischka, D. G. Nathan, R. C. Mulligan, *Nature* **310** (1984) 476.
128. H. Stuhlman, R. Cone, R. C. Mulligan, R. Jaenisch, *Proc. Nat. Acad. Sci. USA* **81** (1984) 7151.
129. K. Gordon et al., *Bio/Technol.* **5** (1987) 1183.
130. G. H. Swift, R. E. Hammer, R. J. MacDonald, R. L. Brinster, *Cell* **38** (1984) 639.
131. A. S. Perkins, P. T. Kirschmeier, S. Gattoni-Celli, I. B. Weinstein, *Mol. Cell. Biol.* **3** (1983) 1123.
132. J. L. R. Rubinstein, J. F. Nicolas, F. Jacobs, *Proc. Nat. Acad. Sci. USA* **81** (1984) 7137.
133. C. J. Tabin, J. W. Hoffman, S. P. Goff, R. A. Weinberg, *Mol. Cell Biol.* **2** (1982) 426.
134. T. Schenk: *Curr. Top. Microbiol. Immunol.* **93** (1981) 25.
135. S. D. Gillies, S. L. Morrison, V. T. Oi, S. Tonegowa, *Cell* **33** (1983) 717.
136. T. Wilson, *Bio/Technol.* **2** (1984) 753.
137. M. Bienz, H. R. B. Pelham, *Cell* **45** (1986) 753.
138. E. Serfling, A. Lubbe, K. Dorsch-Hasler, W. Schaffner, *EMBO J.* **4** (1985) 3851.
139. S. Goodbourn, K. Zinn, T. Mainatis, *Cell* **41** (1985) 509.
140. V. L. Chandler, B. A. Maler, K. R. Yamamoto, *Cell* **33** (1983) 489.
141. D. DeFranco, K. R. Yamamoto, *Mol. Cell. Biol.* **6** (1986) 993.
142. L. A. Chasin, G. Urlaub, *Proc. Nat. Acad. Sci. USA* **77** (1980) 4216.
143. J. S. Powell, K. L. Berkner, R. V. Lebo, J. W. Adamson, *Proc. Nat. Acad. Sci. USA* **83** (1986) 6465.
144. R. J. Kaufman et al., *J. Biol. Chem.* **261** (1986) 9622.
145. F. McCormick et al., *Mol. Cell. Biol.* **4** (1984) 166.
146. V. A. Luckow, M. D. Summers, *Bio/Technol.* **6** (1988) 47.
147. G. E. Smith, M. D. Summers, M. J. Fraser, *Mol. Cell. Biol.* **3** (1983) 2156.
148. K. Jeang, C. Giam, M. Nerenberg, G. Khoury, *J. Virol.* **61** (1987) 708.
149. Y. Matsuura, R. D. Possee, D. H. L. Bishop, *J. Gen. Virol.* **67** (1986) 1515.
150. G. E. Smith, M. D. Summers, "A Manual of Methods for Baculovirus Vectors and Insect Cell Culture Procedures," Texas Agricultural Experiment Station Bulletin no. **1555,** 1987.
151. S. Maeda et al., *Nature (London)* **315** (1985) 592.
152. G. E. Smith et al., *Proc. Nat. Acad. Sci. USA* **82** (1985) 8404.
153. S. Maeda et al., *Nature (London)* **315** (1985) 680.
154. J. St. Schell, *Science* **237** (1987) 1176.
155. P. J. J. Hooykaas, in [133] p. 517.
156. J. Schroeder et al., *FEBS Lett.* **129** (1981) 166.
157. M. DeCleene, J. DeLey, *Bot. Rev.* **42** (1976) 389.
158. J. Hille, I. Klasen, R. Schilperoort, *Plasmid* **2** (1982) 155.
159. M.-D. Chilton et al., *Cell* **11** (1977) 263.
160. A. J. M. Matzke, M.-D. Chilton, *J. Mol. Appl. Genet.* **1** (1981) 149.
161. S. G. Rogers, R. B. Horsch, R. T. Fraley, H. J. Klee in [126] p. 539.
162. A. J. De Framond, K. A. Barton, M. D. Chilton, *Bio/Technol.* **1** (1983) 262.
163. A. Hoekama, P. R. Hirsch, P. J. J. Hooykaas, R. A. Schilperoort, *Nature (London)* **303** (1983) 179.
164. R. Horsch, H. Klee, *Proc. Nat. Acad. Sci. USA* **83** (1986) 4428.
165. P. J. J. Hooykaas, M. Hofker, H. den Dulk-Ravs, R. A. Schilperoort, *Plasmid* **11** (1984) 195.
166. J. Paszkowski et al., *EMBO J.* **3** (1985) 2717.
167. R. D. Shilliot et al., *Bio/Technol.* **3** (1985) 1099.
168. A. Deshayer, L. Herrera-Estrella, M. Coboche, *EMBO J.* **4** (1985) 2731.

169. T. Reich, V. N. Iyer, B. Miki, *Proc. Congr. Plant Mol. Biol. 1st (Abstr.)* (1985) 28.
170. R. B. Horsch, J. E. Fry, N. L. Hoffman, M. Wallroth et al., *Science (Washington, D.C., 1883-)* **227** (1985) 1229.
171. E. C. Cocking, M. R. Davey, *Science* **237** (1987) 1259.
172. R. T. Fraley et al., *Bio/Technol.* **3** (1985) 629.
173. G. Morelli, F. Nagy, R. T. Fraley, S. G. Rogers et al., *Nature (London)* **315** (1985) 200.
174. C. Sengupta-Gopalan et al., *Proc. Nat. Acad. Sci. USA* **82** (1985) 3320.
175. D. V. Goeddel et al., *Proc. Nat. Acad. Sci. USA* **76** (1979) 106.
176. A. Isaacs, J. Lindenmann, *Proc. R. Soc. London Ser. B* **147** (1957) 258.
177. S. Nagata, N. Mantei, C. Weissman, *Nature (London)* **287** (1980) 401.
178. S. Nagata et al., *Nature (London)* **284** (1980) 316.
179. M. Streuli et al., *Proc. Nat. Acad. Sci. USA* **78** (1981) 2848.
180. J. J. Hooks, B. Detrick, *The Interferon System.* University of Texas Press, Austin 1987, p. 319.
181. D. Collen, *Haemostasis* **43** (1980) 77.
182. D. Pennica et al., *Nature (London)* **301** (1983) 214.
183. M. Verstraete et al., *Lancet I* (1985) 842.
184. M. Matsuo, *Haemostasis* **48** (1982) 242.
185. D. Lau et al., *Bio/Technol.* **5** (1987) 953.
186. D. V. Goeddel et al., *Nature (London)* **281** (1979) 544.
187. H. Masunaga, M. Goto, M. Veda, *JPN* **49** (1987) 807.
188. T. B. Haddy, J. M. Lusher, S. Hendricke, B. K. Trosker, *Scand. J. Haematol.* **22** (1979) 289.
189. D. Metcalf, *Science* **229** (1985) 16.
190. L. M. Souza et al., *Science* **232** (1986) 61.
191. F. Lee et al., *Proc. Nat. Acad. Sci. USA* **82** (1985) 4360.
192. S. Nagata et al., *Nature (London)* **319** (1986) 415.
193. Y.-C. Yang et al., *Cell* **47** (1986) 3.
194. T. Yokota et al., *Proc. Nat. Acad. Sci. USA* **83** (1986) 5894.
195. J. L. Marx, *Science* **239** (1988) 257.
196. A. Klausner, *Bio/Technol.* **3** (1985) 119.
197. P. C. Watkins, R. Eddy, A. K. Beck, V. Vellucci, *DNA* **6** (1987) 205.
198. W. W. Chin, S. D. Gharib, *Adv. Exp. Med. Biol.* **205** (1986) 245.
199. C. L. Funckes, C. D. Minth, R. Deschenes, M. Magazin, *Proc. Nat. Acad. Sci. USA* **258** (1983) 8781.
200. P. Valenzuela, A. Medina, W. J. Rutter, G. Ammerer et al., *Nature (London)* **298** (1982) 347.
201. E. J. Patzer et al., *Bio/Technol.* **4** (1986) 630.
202. H. Schellekens, A. de Reus, J. H. Peetermans, P. A. van Eerd, *Postgrad. Med. J. (England)* **63** (1987) 93.
203. A. Safary, F. Andre, *Postgrad. Med. J. (England)* **63** (1987) 105.
204. L. H. Miller et al., *Science* **234** (1986) 1349.
205. S. F. Josephs et al., *Science* **234** (1986) 601.
206. R. C. Kennedy et al., *Science* **231** (1986) 1556.
207. S.-L. Hu, S. G. Kosowski, K. F. Schaaf, *J. Virol.* **61** (1987) 3617.
208. D. H. Smith et al., *Nature (London)* **238** (1987) 1704.
209. J. R. Haynes et al., *Bio/Technol.* **4** (1986) 637.
210. L. A. Laskey et al.: *Modern Approaches to Vaccines: Molecular and Chemical Basis of Virus Virulence and Immunogenicity,* Cold Spring Harbor Laboratory, New York, 1984, p. 189.
211. D. Panicali, E. Paoletti, *Proc. Nat. Acad. Sci. USA* **79** (1982) 4927.
212. R. E. Hammer, R. L. Brinster, R. D. Palmiter, *Nature (London)* **311** (1984) 65.
213. A.-C. Andres et al., *Proc. Nat. Acad. Sci. USA* **84** (1987) 1299.
214. F. Constantini, K. Chada, J. Magram, *Science* **233** (1986) 1192.
215. N. Maclean, D. Penman, Z. Zhu, *Bio/Technol.* **5** (1987) 257.
216. R. E. Hammer et al., *Nature (London)* **315** (1985) 680.
217. R. E. Hammer, V. G. Pursel, C. E. Rexroad Jr., R. J. Wall, *J. Anim. Sci.* **63** (1986) 269.
218. K. Radke, D. Lagaris, *Bio/Technol.* **4** (1986) 14.
219. J. P. Simons, M. McCleaghan, A. J. Clark, *Science* **328** (1987) 530.
220. A. Klausner, *Bio/Technol.* **4** (1986) 403.
221. A. M. Dandekar, P. K. Gupta, D. J. Durzan, V. Knauf, *Bio/Technol.* **5** (1987) 587.
222. T. J. Parsons et al., *Bio/Technol.* **4** (1986) 533.
223. P. Umbeck, G. Johnson, K. Barton, W. Swain, *Bio/Technol.* **5** (1987) 263.
224. L. Comai, D. Facciotti, W. R. Hiatt, G. Thompson et al., *Nature (London)* **117** (1985) 741.
225. H. Bialy, A. Klausner, *Bio/Technol.* **4** (1986) 96.
226. D. C. Banlcombe et al., *Nature (London)* **321** (1986) 446.
227. V. A. Hilder, A. M. R. Gatehouse, S. E. Sheerman, R. F. Barker et al., *Nature (London)* **330** (1987) 160.

228. P. F. Lurquin, C. I. Kado, *Mol. Cell. Genet.* **154** (1977) 113.
229. M. Caboche, A. Deshayes, *C. R. Acad. Sci. (Paris)* **299** (1984) 663.
230. E. Neumann, M. Schaefer-Ridder, Y. Wang, P. H. Hofschneider, *EMBO J.* **1** (1982) 841.
231. J. A. Thompson, R. Abdullah, E. C. Cocking, *Plant Sci. (Lucknow)* **47** (1986) 123.
232. D. A. Williams, S. H. Orkins, *J. Clin. Invest.* **77** (1986) 1053.
233. W. F. Anderson, *Science* **226** (1984) 401.
234. G. Chu, H. Hayakawa, F. Berg, *Nucleic Acids Res.* **15** (1987) 1311.
235. R. G. Gregg, O. Smithies, *Cold Spring Harbor Symp. Quantitative Biology* **51** (1987) 1093.
236. A. D. Miller et al., *Science (Washington, D.C., 1883-)* **225** (1984) 630.
237. R. F. Selden et al., *Science (Washington, D.C., 1883-)* **236** (1987) 714.
238. E. A. Dzierzak, T. Papayannopoulou, R. C. Mulligan, *Nature (London)* **331** (1988) 35.
239. J. L. Fox, *Bio/Technol.* **6** (1988) 110.
240. R. Mulligan, paper presented at the Miami Winter Symposium–Advances in Gene Therapy, Miami, Florida, 1987.
241. J. van Brunt, A. Klausnen, *Bio/Technol.* **5** (1987) 211.
242. R. K. Saiki, N. Arnheim, H. A. Erlich, *Bio/Technol.* **3** (1985) 1008.
243. C. Y. Ou et al., *Science (Washington, D.C., 1883-)* **239** (1988) 295.
244. H. H. Kazazian et al., in *Recombinant DNA Applications to Human Disease,* 14th Banbury Report, Cold Spring Harbor Laboratory, New York 1983, p. 29.
245. G. I. Bell, J. H. Karam in [244] p. 317.
246. J. A. Phillips III in [244] p. 305.
247. C. T. Caskey, *Science* **237** (1987) 1223.
248. P. Berg et al., *Nature (London)* **255** (1975) 442.
249. Guidelines for Research Involving Recombinant DNA Molecules, Federal Register 1976.
250. W. Booth: "News & Comment," *Science* **239** (1988) 341.
251. P. P. Trotta, H. Van Le, B. Sharma, T. L. Nagabhushan, *Dev. Ind. Microbiol.* **27** (1987) 53.

Geriatric Drugs → Psychopharmacological Agents
Geriatric Drugs → Vitamins

Germanium and Germanium Compounds

JEAN SCOYER, Metallurgie Hoboken-Overpelt, Olen, Belgium (Chaps. 1 – 10)

HÉLIAN GUISLAIN, Metallurgie Hoboken-Overpelt, Olen, Belgium (Chaps. 1 – 10)

H. UWE WOLF, Abteilung Pharmakologie und Toxikologie, Universität Ulm, Federal Republic of Germany (Chap. 11)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
4. Resources and Raw Materials 3
5. Production 4
5.1. Production of Germanium Concentrate 5
5.1.1. Bleiberger Bergwerks Union 5
5.1.2. Jersey Minière Zinc 5
5.1.3. Musto Exploration 6
5.1.4. Peñarroya 6
5.1.5. Pertusola 6
5.1.6. Other Primary Producers 7
5.2. Production of Germanium from Germanium Concentrate 7
5.3. Purification 8
5.4. Transformation Processes 8
6. Environmental Protection 9
7. Quality Specifications 9
8. Chemical Analysis 10
9. Storage and Transportation 10
10. Uses and Economic Aspects 10
11. Toxicology and Occupational Health 12
12. References 13

1. Introduction

Germanium [*7440-56-4*], Ge, atomic number 32, atomic mass 72.59, is a semiconducting element found in group 14 of the periodic table.

Germanium has five natural isotopes: ^{70}Ge (20.53 %), ^{72}Ge (27.43 %), ^{73}Ge (7.76 %), ^{74}Ge (36.54 %), and ^{76}Ge (7.76 %).

Germanium also has some radioactive isotopes with greatly different half-lifes, the isotope with the longest half-life being ^{68}Ge (287 days).

In 1871, MENDELEEV predicted the existence and the properties of an unknown element, ekasilicon, to be placed between silicon and tin in his periodic system of elements. In 1886, WINKLER isolated a new element from argyrodite, $4Ag_2S \cdot GeS_2$. This element, named germanium, showed properties remarkably close to those predicted by MENDELEEV.

Germanium found its first industrial application in 1947 with the development of the transistor. Until the early 1970 s, it was used extensively in the solid-state electronics industry, although it has been replaced almost entirely by the cheaper silicon.

The main use of metallic germanium currently is for infrared optics. Germanium tetrachloride is used in optical fiber production. Germanium dioxide is used mainly as a catalyst in the production of polyester and synthetic textile fiber.

The total consumption of germanium was estimated to be ca. 70 t of the element in 1986. The main producers of germanium end products are located in the United States, Belgium, France, Federal Republic of Germany, and Japan.

Germanium is considered as a strategic material by the United States government. The erection of a stockpile with a capacity of 30 t of Ge was decided in 1984. The first delivery began in 1987 and the capacity may be increased to 140 t.

2. Physical Properties

Although high-pressure polymorphs are known, germanium normally crystallizes in the diamond structure with comparatively weak covalent atomic bonding, yielding the typical hardness and brittleness, as well as the semiconductor behavior of this gray amphoteric metal.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_351

The specific structure of the forbidden energy gap (band gap E_g), the binding forces between the germanium atoms, and their mass determine the major electron – photon and photon –phonon interactions in the crystal lattice, which result in a broad optical transmission range for infrared radiation. The fact that this range covers a main atmospheric transmittance window between 7 and 14 µm, coinciding with the spectral emittance peak of a blackbody at room temperature, has promoted the use of germanium as an infrared optical material. Germanium's high refractive index results not only in high optical power and little spherical aberration, but also in the need for antireflection coatings. These materials often occur as exotic contaminants in recycling operations. Germanium's thermal properties sensitize this brittle material to thermal shock; crucible design and seeding techniques must also cope with a 6.6 % volumetric expansion of germanium during solidification.

Due to the occurrence of a narrow band gap, most physical properties of germanium (e.g., resistivity) depend strongly on temperature as well as on minute concentrations of solutes. An updated and extensive compilation of the physical properties of germanium can be found in [1]; Table 1 summarizes typical room temperature values.

Table 1. Physical properties of germanium at room temperature

Property	Value
Mechanical	
Lattice parameter, nm	0.565790
mp, K	1210.4
Density, g/cm^3	5.3234
Microhardness (Vickers-ASTM E384), kg/mm^2	780
Tensile fracture strength, MPa	100
Electrooptical	
Band gap, eV	0.67
Intrinsic resistivity, Ωcm	47.6
Transmission wavelength range, µm	2 – 16
Absorption coefficient at $\lambda = 10.6$ µm, cm^{-1}	0.02
Refractive index at $\lambda = 10.6$ µm	4.0027
Thermal	
Thermal conductivity, W cm^{-1} K^{-1}	0.586
Specific heat, J kg^{-1} K^{-1}	310
Linear thermal expansion, K^{-1}	5.9×10^{-6}

3. Chemical Properties [2], [3]

Elemental Germanium. Metallic germanium is stable in air. At temperatures above 400 °C in oxygen, a passivating oxide layer is formed. This layer is destroyed by water vapor.

Germanium resists concentrated hydrochloric and hydrofluoric acids even at their boiling points, but reacts slowly with hot sulfuric acid. The dissolution of germanium in nitric acid proceeds faster than in sulfuric acid. In alkaline media in the presence of an oxidizing agent, e.g., hydrogen peroxide, germanium dissolves rapidly, even at room temperature. Germanium reacts readily with halogens to form tetrahalides.

Germanium and silicon are totally miscible and form a continuous range of alloys. The system with tin has a eutectic.

Germanium Compounds. Germanium can have valences of two or four. The divalent compounds are less stable and oxidize easily to the tetravalent compounds, which are similar to their silicon analogs.

Germanium dioxide [*1310-53-8*], GeO_2 M_r 101.59. Some properties of germanium dioxide are listed in Table 2. It exists in two crystalline forms and one amorphous form, all of which are white powders. The hexagonal form is obtained when germanium tetrachloride is hydrolyzed with water. This reaction leads to a voluminous gel that continuously releases its water during drying. The transformation to the tetragonal form occurs only on heating at 300 – 900 °C in the presence of a catalyst (e.g., an alkali metal halide).

The amorphous form is obtained by rapid solidification of molten GeO_2 or by reaction of $GeCl_4$ or GeO with oxygen in the gas phase. Moisture catalyzes the transformation from the amorphous to the hexagonal form.

Germanium monoxide [*20619-16-3*], GeO, M_r 88.59, is obtained by the reaction of GeO_2 with reducing agents like germanium metal, iron, carbon, carbon monoxide, or hydrogen above 600 °C. Germanium monoxide sublimes above 710 °C, and this property is used in some industrial processes to recover germanium from slags or residues (fuming of Ge). Germanium monoxide is insoluble in water.

Table 2. Properties of germanium dioxide

Property	Structure		
	Hexagonal	Tetragonal	Amorphous
Thermal stability, °C	1049 – 1116	<1049	
Density at 25 °C, g/cm^3	4.228	6.239	3.637
mp, °C	1116	1086	
Solubility in			
water at 25 °C, g/L	4.53	insoluble	5.18
water at 100 °C, g/L	13	insoluble	
HF, HCl, NaOH	soluble	insoluble	soluble

Table 3. Properties of germanium tetrahalides

Property	$GeCl_4$ [*10038-98-9*]	GeF_4 [*7783-58-6*]	$GeBr_4$ [*13450-92-5*]	GeI_4 [*13450-95-8*]
M_r	214.43	148.6	392.26	580.28
Form at 25 °C	liquid	gas	crystals	crystals
Color	colorless	colorless	gray white	red
Density at 25 °C	1.8755			
mp, °C	−49.5	−15	26.1	146
bp, °C	83.2	−36.5 (sublimes)	186.5	348

Germanium Halides. Of the germanium halides, only Germanium tetrachlorgermanium tetrachloride [*10038-98-8*] has practical applications. The properties of the tetravalent halides are listed in Table 3. ide is obtained by reaction of chlorine gas with germanium metal and of hydrochloric acid with GeO_2 or germanates. Germanium tetrachloride is soluble in carbon tetrachloride. Industrial practice takes advantage of the insolubility of $GeCl_4$ in concentrated HCl (>6M) to purify $GeCl_4$ from trace impurities such as As, Cu, and B by solvent extraction with HCl.

Germanium disulfide [*12025-34-2*], GeS_2, M_r 136.72, is obtained by the action of H_2S on acidic germanium solutions. Because GeS_2 is insoluble in acid media, this reaction is used industrially to obtain germanium concentrate.

Germanates are obtained in reactions at elevated temperature (800 – 1200 °C), in which glasses are produced from melts of GeO_2 and metal oxides or carbonates. The reactions also occur at room temperature in aqueous solution; e.g., sodium germanate [*12195-31-2*], $Na_3HGe_7O_{16}$, is formed by treatment of a NaOH solution of GeO_2 with sulfuric acid.

Germanides can be produced by melting other metals with germanium and freezing the melt, e.g., Mg_2Ge [*1310-52-7*]. Germanes (germanium hydrides) are evolved when germanides react with acids: monogermane [*7782-65-2*], GeH_4, digermane [*13818-89-8*], Ge_2H_6, and trigermane [*14691-44-2*], Ge_3H_8, can be separated by fractional distillation.

Organometallic Compounds. The organometallic compounds of germanium have few technical uses. Properties and methods of preparation have been extensively described [4].

Table 4. Germanium minerals

Name and composition	Ge content, wt %	Mining location
Stottite, $FeGe(OH)_6$	29	none
Schauerteite, $Ca_3Ge(SO_4)_2(OH)_6 \cdot 3\ H_2O$	14	none
Briarite, $Cu_2(Fe,Zn)GeS_4$	13 – 17	none
Germanite, $Cu_3(Ge,Fe)S_4$	5 – 10	Tsumeb, Namibia
Renierite, $(Cu,Fe,Ge,Zn,As)S$	6.3 – 7.7	Tsumeb, Namibia Kipushi, Zaire
Argyrodite, $4Ag_2S \cdot GeS_2$	1.8 – 6.9	Freiberg, Germany
Canfeldite, $4Ag_2S \cdot (Sn,Ge)S_2$	0.1	Bolivia

4. Resources and Raw Materials [5], [6]

The concentration of germanium in the earth's crust averages 6.7 ppm, which is comparable to the concentration of zinc or lead. Thus, germanium is not scarce, but it is widely dispersed so that only a few minerals have been isolated. The most important are listed in Table 4.

Germanium is present in the lattices of certain sulfides, such as sphalerite. The concentration of germanium in some germanium-rich deposits is given in Table 5.

Table 5. Germanium content of diverse deposits

Type	Location	Estimated Ge content, ppm
Zinc ∗	Gordonsville, Tennessee, United States	400
	Nanisivik, Baffin, Canada	230
	Saint Salvy, France	150
	Tsumeb, Namibia	250
	Lyon Lake, Ontario, Canada	200 – 250
	Salafosa, Italy	200
	Bleiberg, Austria	200
Copper	Apex, Utah, United States	640
Coal	Northumberland, United Kingdom	300

∗ Germanium concentration in zinc concentrate.

In the past, germanite (Tsumeb) and renierite (Kipushi) were the principal sources of germanium. Currently, germanium is recovered mainly from byproducts of the zinc industry.

The only mining operation specific for germanium (and Ga) is located at an old copper mine, the Apex, that has been reopened for this purpose.

At one time, germanium was recovered from the coal ash of electric power plants in England. This source of germanium vanished due to the decreased coal consumption for electric power production. The former Soviet Union, and probably China, still recover germanium from this source.

5. Production

The overall production processes can be divided into two steps: (1) production of germanium concentrate (primary production) and (2) production of germanium end products.

The first step varies depending on the raw materials processed and the chosen technology. Pyrometallurgical processes are based on the volatility of GeO and GeS. Hydrometallurgic processes are based on precipitation of germanium as its sulfide or hydroxide or with tannic acid [*1401-55-4*].

More recently, solvent extraction has been reported using hydroxyoximes (e.g., LIX63 [*6873-77-4*]), or oxime derivatives (e.g., Kelex 100 [*29171-27-5*], and LIX26 [*85746-33-4*]) [7–18]. Fixation of germanium on resin impregnated with amines has also been patented [19]. However, published information on all of these is scarce and not always reliable.

The second step relies on the volatility of $GeCl_4$ as the separation – purification step. Because the processes used by the diverse producers are quite similar, this step can be described with one flow sheet (see Fig. 6).

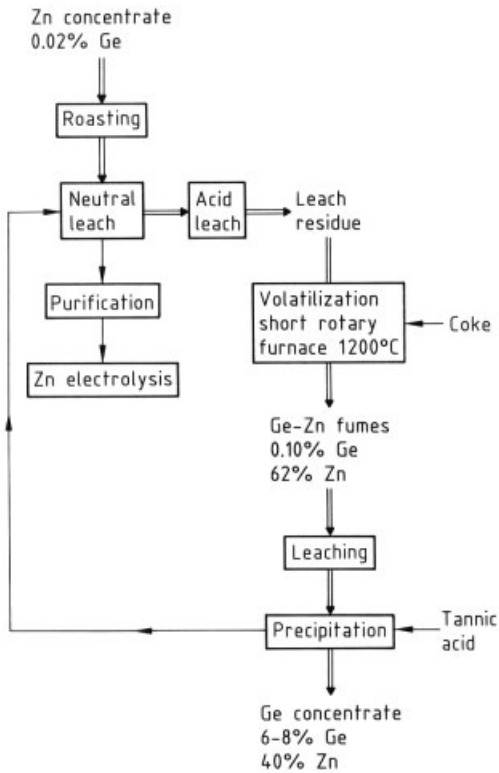

Figure 1. Production of a germanium concentrate: flow sheet of Bleiberger Bergwerks Union (BBU)
Location: Arnoldstein, Austria; Bleiberg mine. Estimated Ge capacity: 6 t/a.

5.1. Production of Germanium Concentrate

The flow sheets for the production of germanium concentrate from germanite and renierite as raw materials have been described [20].

5.1.1. Bleiberger Bergwerks Union

Bleiberger Bergwerks Union (BBU) processes the leach residues of its own zinc plant by a thermic process in a modified Waelz oven (Fig. 1). Germanium monoxide is sublimed at ca. 1200 °C. After the fumes are leached, germanium is precipitated with tannic acid [*1401-55-4*]. The germanium-free solution is recycled to the zinc plant and the germanium concentrate is sold as such.

The use of tannic acid leads to a relatively poor germanium concentrate. Care must be taken to avoid problems in the zinc electrolysis due to the recycling of unreacted tannic acid [21–23].

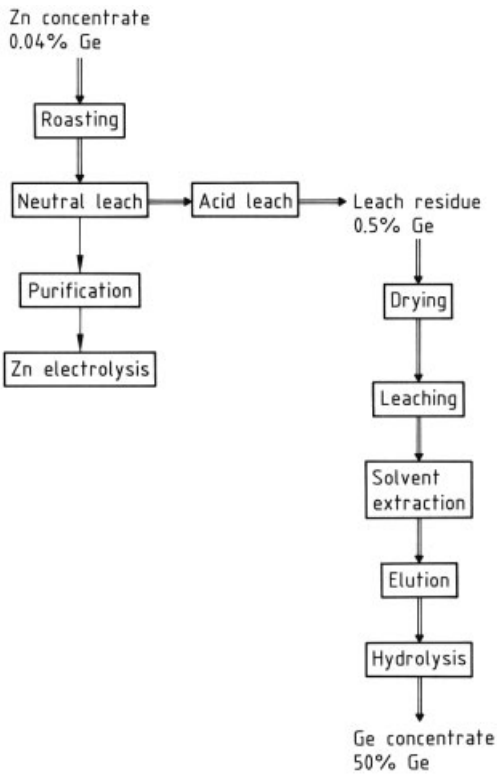

Figure 2. Production of a germanium concentrate: flow sheet of Jersey Minière Zinc (JMZ) and Metallurgie Hoboken-Overpelt (MHO)
Locations: JMZ at Clarksville, Tennessee, United States; Gordonsville mine. MHO at Olen, Belgium. Estimated Ge capacity: 65 t/a

5.1.2. Jersey Minière Zinc

Jersey Minière Zinc produces zinc from its mines in Gordonsville and Elmwood, Tennessee. Germanium is concentrated in the leach residue, which is dried and shipped to Metallurgie Hoboken-Overpelt (MHO), Belgium (Fig. 2) [24].

At the Olen plant in Belgium, the germanium residue is leached, and germanium is separated by solvent extraction. After elution in an alkaline medium, a germanium concentrate is precipitated by hydrolysis [25]. The use of solvent extraction technology enables MHO to produce very rich germanium concentrates, regardless of the concentration in the feed [8–13].

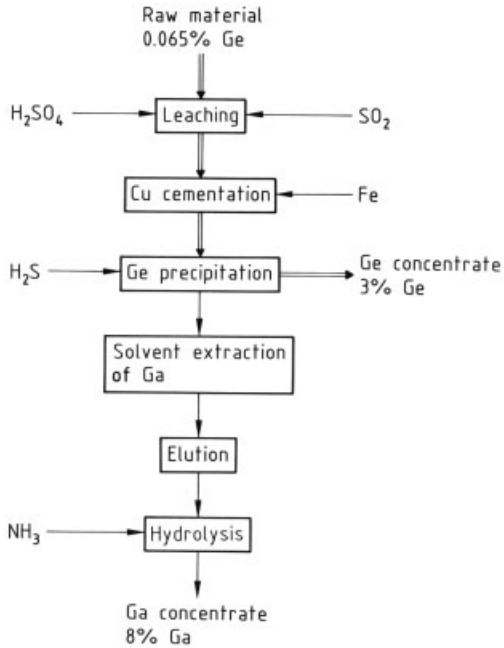

Figure 3. Production of a germanium concentrate: flow sheet of Musto Exploration.
Location: St. George, Utah, United States; Apex mine. Capacity: 18 t/a of Ge and 10 t/a of Ga.

5.1.3. Musto Exploration

The Ge – Ga concentrates from the company-owned Apex mine are leached in sulfuric acid with SO_2 to reduce trivalent iron (Fig. 3). After cementation with iron to remove copper, H_2S is added to give a first sulfide concentrate. Gallium is then separated by solvent extraction. The

germanium concentrate containing ca. 3 % germanium is upgraded to GeO_2 of 98 % purity before being sold to refiners. Exploitation started in 1985 [26].

5.1.4. Peñarroya

Concentrates of Zn – Pb and germanium matte from Pertusola (see Fig. 5) , a subsidiary of Peñarroya, are processed together by the Imperial Smelting Process. Germanium is collected in the zinc phase, which is further purified by distillation in New Jersey columns (Fig. 4). Germanium is then concentrated in a residue by volatilization of zinc. However, due to ecological problems, thermal processes for zinc are losing importance compared with hydrometallurgic processes.

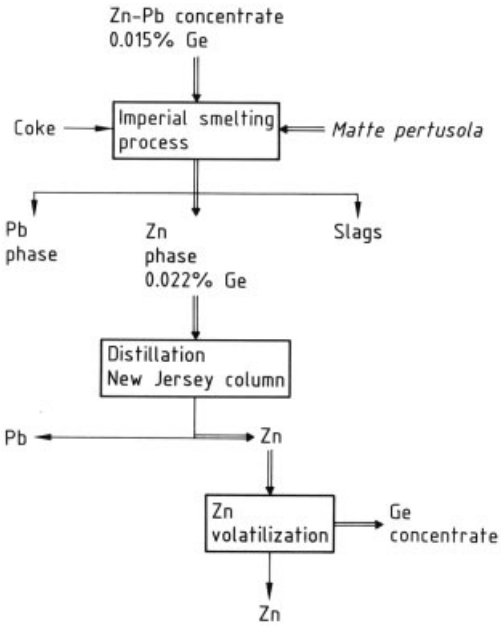

Figure 4. Production of a germanium concentrate: flow sheet of Peñarroya
Location: Noyelles-Godault, France; Saint-Salvy mine. Capacity: GeO_2 – $GeCl_4$, 60 t/a; Ge metal, 30 t/a

5.1.5. Pertusola

The zinc concentrates of Salafosa are treated at the refinery of Crotone (Fig. 5). The leach and purification residues of the hydrometallurgic plant are thermally decomposed in a cubilot. Germanium is recovered in the Pb – Zn fumes, a small portion going into the matte phase, which is processed further by Peñarroya (Fig. 4). After releaching the fumes, a germanium concentrate containing 7 – 35 % germanium is produced.

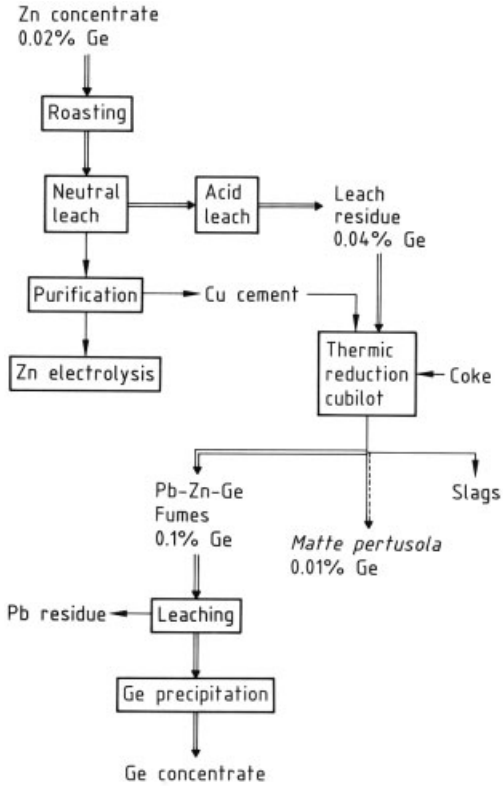

Figure 5. Production of a germanium concentrate: flow sheet of Pertusola
Location: Crotone, Italy; Salafosa mine.

5.1.6. Other Primary Producers

The Metallgesellschaft (Germany) recovers ca. 10 t/a of germanium from the hydrometallurgical plant of Datteln. China exports ca. 4 t/a of germanium. The production of the former Soviet Union is mainly used internally (estimated 40 t/a of germanium). Cominco (British Columbia, Canada) is developing a project for 10 – 20 t/a of germanium.

Eagle Picher (Quapaw, Oklahoma, United States) processed the residues of its zinc smelter by fuming in a Waelz oven at 1250 °C. Be-

cause this source is not economical at current prices, Eagle Picher now supplies its refining unit (25 – 40 t/a of germanium capacity) with purchased germanium concentrates [27]. Preussag (15 t/a) and Otavi (5 t/a) (both Federal Republic of Germany) processed germanite from Tsumeb before depletion, but they now buy concentrate [5].

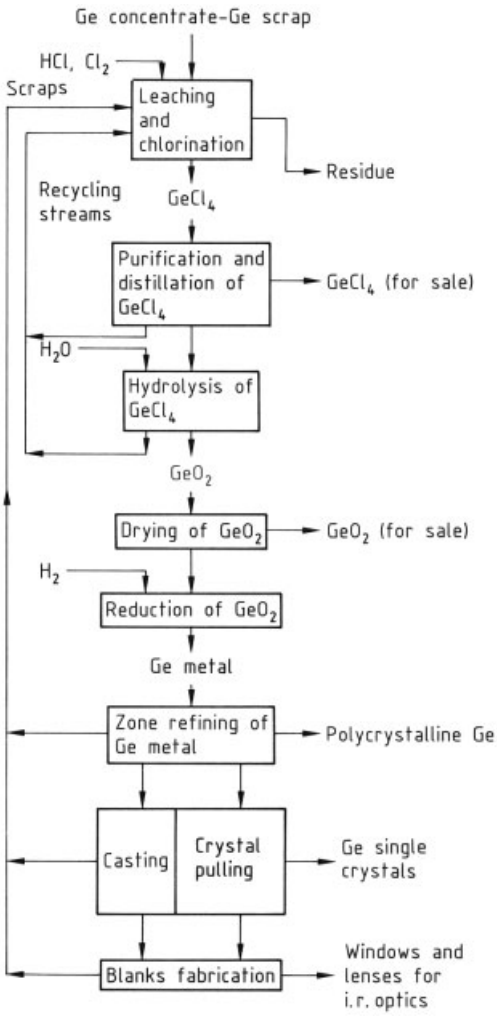

Figure 6. Flow sheet of the production of germanium metal from germanium concentrate

5.2. Production of Germanium from Germanium Concentrate

The transformation of a germanium concentrate (mostly GeO_2) or germanium metal scrap to germanium (Fig. 6) [28] begins with conversion to $GeCl_4$:

$$GeO_2 + 4\,HCl \rightarrow GeCl_4 + 2\,H_2O$$

$$Ge + 2\,Cl_2 \rightarrow GeCl_4$$

The reactions are run at the boiling point of $GeCl_4$ (83.2 °C) in concentrated HCl (6 M) to avoid hydrolysis. Even when only GeO_2 is processed, a chlorine atmosphere is used to oxidize the volatile $AsCl_3$ to nonvolatile H_3AsO_4.

The $GeCl_4$, containing As^{3+} as the major impurity, is purified by fractional distillation. Traces of As, Sb, Cu, and B are extracted by HCl (>6 M) saturated with chlorine. Other purification techniques have been reviewed [20].

The purified $GeCl_4$ is hydrolyzed with high-purity water,

$$GeCl_4 + 2\,H_2O \rightarrow GeO_2 + 4\,HCl$$

After filtration and thorough washing, the GeO_2 is dried.

The usual method of producing elemental germanium is hydrogen reduction [29] of germanium dioxide in electrically heated tubular furnaces at ca. 650 °C, yielding a gray metallic germanium sponge. In industry, graphite and/or quartz boats and furnace linings are used. Quartz boats must be sandblasted and/or coated to avoid sticking and breakage.

Care must be taken to keep the temperature below 700 °C during reduction to avoid generation of volatile germanium monoxide. Adequate processing limits the germanium loss to less than 0.1 wt %.

As a result of the very stringent purity requirements, reduction processes [2] using carbon, flux electrolysis, or the direct reaction of germanium tetrachloride with zinc have lost industrial importance. The reaction of $GeCl_4$ with H_2 can be used to grow epitaxial germanium layers [30], but bulk germanium production is hampered by low metal yield and overall energy yield.

After complete reduction, the temperature is raised above the melting point (937 °C), and a polycrystalline germanium bar is solidified from one end to the other. Due to the different solubilities of the impurities in the liquid and solid

phases, the impurities concentrate by segregation [31] in the first or the last melt fraction to solidify. Hydrogen is highly soluble in molten germanium; therefore, to avoid porosity and blow holes, the crystallization speed must be limited to a few centimeters per hour.

5.3. Purification

To obtain semiconductor-grade germanium, the material must be further purified by zone refining [32], [33], under hydrogen, nitrogen, or another controlled atmosphere. Due to the low surface tension and high density of molten germanium, a crucible-free float zone-refining process never matured [34] (→ Crystal Growth, Chap. 6.4.). High-purity germanium is obtained by repeated passage of crude germanium through the zone-refining crucible. The number of efficient zone-refining passes which can be carried out in a single crucible is limited by the partition coefficient of the impurities between germanium and the crucible material. Eventually, an equilibrium between the concentration of impurities in the melt and in the crucible is reached. For this reason, a cyclic batch process, in which the zone refining is continued in a new crucible, is unavoidable if high-purity germanium is to be obtained. Between the successive refining steps, the less pure metal parts are mechanically cut away and recycled. A suitable etching technique [35] is required to clean and prepare the load as well as the crucible for the next refining step. Grain boundaries and other crystallographic imperfections (dislocations, vacancies, etc.) act as precipitation sites for impurities, impeding the ultrapurification.

Ultimate purity levels can therefore be obtained (and measured) only in single crystals grown in the most aseptic, dust-free environment. The physics and chemistry of ultrapurification of germanium have been discussed [36–38].

5.4. Transformation Processes

For most applications, the germanium must be processed further into a suitable size and crystallographic perfection and doped within a narrow range (e.g., 1–8 ppb) with a definite impurity (usually a group 13 or group 15 element) to yield the specified electrooptical properties.

Two techniques are available for this transformation: (1) the *Bridgman process* [39] which entails horizontal or vertical, possibly seeded, directional solidification in a graphite crucible, often improperly called casting, and (2) the *Czochralski process* [40], by which a crystal seed is dipped into a germanium melt and allowed to grow by suitable control of pull speed and heat balance (→ Crystal Growth, Chap. 6.). The second process yields the most perfect material, e.g., dislocation-free single crystals. The fundamental physicochemical background of crystal-growth processes was reviewed [41–43]. Modern equipment handles up to 150 kg germanium loads and allows growth of single crystals up to 300 mm in diameter in a typical 48 h cycle.

Germanium crystals are ground, sliced, and milled with high precision microcomputer-controlled, diamond-tooled machines into the desired geometrical shape. This operation generates huge and costly amounts of scrap fines which have to be recycled. The brittleness of germanium and the need to preserve a near-perfect, strain-free crystal lattice do not allow germanium forging; segregation of the indispensable dopants hinders precision casting.

6. Environmental Protection

In the primary production of germanium, environmental problems are not specific to germanium, but are those of the zinc industry (e.g., SO_2 emission, cadmium pollution).

When solvent extraction units are used (see Section 5.1), special attention must be paid to fire protection and cleaning of the off-gas. The off-gas of $GeCl_4$ production must be scrubbed with alkaline solution to destroy HCl and chlorine. Monitoring the hydrogen content in the air avoids any risk of explosion in the reduction section.

7. Quality Specifications

Major impurities (<10 ppm) in $GeCl_4$ are O, C, and H, all detectable by infrared spectrometry.

The concentration of all metallic impurities can be depressed to or below their detection limits (typically <5 ppb). Process-induced impurities in GeO_2 are H_2O (<3000 ppm, ASTM F5-60) and Cl (<300 ppm, turbidimetry). Often, C, F, and S are detected at ppm levels (conductimetry); all other elements are at or below their detection limit (<10 ppb).

Density (ASTM B 417-76, B 527-70) and grain size (ASTM B 330-76) obviously influence reduction kinetics, but the contamination risks inherent to all high-temperature processes suggest the use of all-quartz furnaces and synthetic quartz boats with palladium-diffused H_2 for the assessment of electronic grade GeO_2 by reduction (ASTM F 27-83).

The most commonly used characterization techniques for germanium metal are those that determine the electrical properties of semiconductors: e.g., transport properties (ASTM F 43-78, F 76-73, F 42-77) and crystallographic perfection (ASTM F 389-76). For evaluation of the optical quality of germanium, a number of routine spectroscopic and specialized interferometric techniques exist [44]. All of these methods are, however, sensitive to the surface preparation (polishing) of the sample. Since many of the optical and electronic properties of germanium are linked to one another [45], [46], the electronic properties, which can be more accurately and reliably measured, can be used to define the optical properties. Thus, for example, standard optical grade germanium is specified as *n*-type germanium with resistivity 5 – 40 Ωcm.

To determine the chemical nature and concentration of electrically active impurities, three sensitive (10^{-5} ppb) techniques are used: deep level transient spectroscopy (DLTS) [47], far infrared transmission (FIRT) [48] (Fig. 7), and photothermal ionization spectroscopy (PTIS) [49].

Impurity, ppb	"As reduced"	Refined
P	1.5	0.36
As	0.77	0.005
Al	1.1	0.27
Ga	0.031	0.005

The major residual impurities in high purity germanium (H_2 < 100 ppb; C, Si, and O < 10 ppb) are process-induced and barely detectable by their lattice phonon absorption in the germanium matrix (ASTM F 120-75, F 122-74) or by autoradiospectrometry [50], [51]. The concentration of other impurities can be depressed below 10^{-3} ppb.

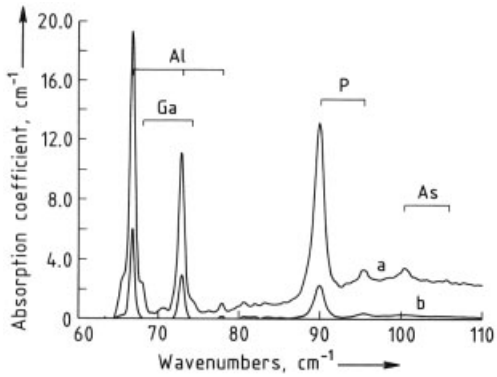

Figure 7. Purification by zone refining
a) 8K-FIRT spectra of germanium "as reduced" vs. b) after zone refining

8. Chemical Analysis

The analytical determination of germanium has been reviewed [52].

Germanium can be determined by photometry at 510 nm after complexation with phenylfluorone [*975-17-7*]. Atomic absorption spectroscopy at 261.1 nm is less sensitive than the previous method. X-Ray fluorescence enables determination down to 5×10^{-3} %. Polarographic methods such as differential pulse polarography and anodic stripping have been developed for the determination of germanium (and other impurities) in Zn electrolytes.

The determination of impurities in $GeCl_4$, GeO_2, and Ge can be achieved by graphite furnace atomic absorption spectroscopy (GFAAS) [53] or spark source mass spectroscopy [54].

9. Storage and Transportation

Germanium tetrachloride is classified as a dangerous corrosive liquid subject to international regulations (ADR/RID-IMDG-IATA) under UN 1760 Class 8 Group II. The combination of these with the very stringent purity requirements of the optical fiber manufacturers demands the use of expensive packaging (e.g., glass/teflon [*9002-84-0*], tantalum-coated steel [*7440-25-7*]).

Table 6. Germanium market in 1986 ∗

Application	Product	Consumption, t	Annual growth rate, %
Infrared optics	optical-grade Ge, shaped mono- and poly-crystalline blanks	35	15
Optical fibers	high-purity $GeCl_4$	13	10
Catalysis	GeO_2 (hexagonal or amorphous), $NaH_3Ge_2O_6$	13	5
Electronics, optoelectronics	doped Ge single-crystal slices	2	5
High-energy radiation detectors	high-purity Ge single crystal	1	5
Luminescence	GeO_2	1	
Medical, alloys, and glasses	GeO_2, Ge pellets, and powders	5	5
	Total	70	11

∗ Excluding Eastern-bloc countries

The packaging should preserve the purity of the GeO_2 during transportation and storage. For relatively short storage, a hermetically sealed high-density polyethylene bottle is satisfactory. The quality degradation during longer storage is linked to diffusion of impurities through or deterioration of the polyethylene. In this case the use of fluorinated plastic foil or coating is advisable.

Germanium metal oxidizes slowly in air; therefore, its packaging should primarily take its brittleness into account.

10. Uses and Economic Aspects

Germanium is used in manufacturing a wide range of high-technology products. Table 6 summarizes these applications for the market economies with respect to market volume and trends, as well as the germanium compound used.

The use of germanium in *optical materials*, by far the most important today, represents ca. 50 % of the world consumption. Germanium itself is the primary high-quality lens and window material for thermal imaging in the 8 – 12 µm wavelength range. The materials trade-off for this application was reviewed [55]. The replacement of silicon by germanium as a cation in optical glasses [56] extends their transmittance toward longer wavelengths and increases the refractive index. The second property is used to confine the signal within the core of telecommunication optical fibers. The chemistry of germanium in fiber manufacture is discussed in [57]. With As, Sb, and Se, germanium forms nonoxide glasses [58], also used in IR optics (Fig. 8), and whose photosensitivity yields high-resolution photoresists [59].

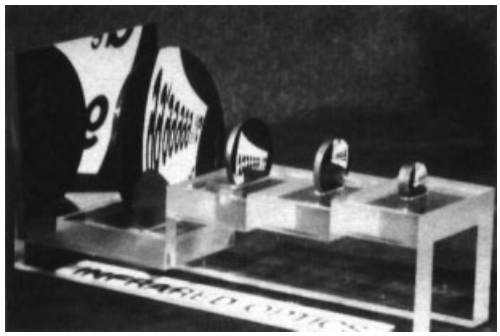

Figure 8. Germanium for infrared optics: typical set of polished Ge blanks

The second major use of germanium is as *catalyst or cocatalyst* [60] in the production of polyesters (e.g., poly(ethylene terephthalate)) and synthetic textile fibers. The more expensive germanium catalysts are used when the final product must be a colorless fiber or plastic, e.g., photographic films.

Some germanates crystallize as typical *phosphors,* which reemit absorbed energy as visible radiation. Magnesium germanates and fluorogermanates, e.g., $Mg_{28}Ge_{10}O_{48}$ and $Mg_{56}Ge_{15}O_{66}F_{20}$, in which magnesium acts as activator, are used in lamps as a constituent of the fluorescent coating [61]. This is preferred to the cheaper arsenic-based phosphors, for which safety regulations are severe. Bismuth germanate [*12233-56-6*], $Bi_4Ge_3O_{12}$, is an in-

trinsic phosphor that is used as a single crystal in high-energy gamma-ray scintillator detectors [62]. The stopping power is extremely high because of the high atomic number of bismuth.

The remaining interest for germanium in electronics is based on the advantageous mobility characteristics of charge carriers in this material. Typical applications are high-power devices with low energy loss and photodetectors [63]. The most demanding photodetector application involves *gamma-ray spectroscopy detectors* [64], for which germanium must be refined to an extremely low impurity level ($<10^{-3}$ ppb) and grown in nearly perfect single crystals. Recent progress in producing stable insulating layers on germanium [65], as well as its crystal lattice match ($\Delta a/a < 5\times10^{-3}$) with GaAs [*1303-00-0*] (epitaxy) offers new hope for germanium as an electronic material in integrated circuits.

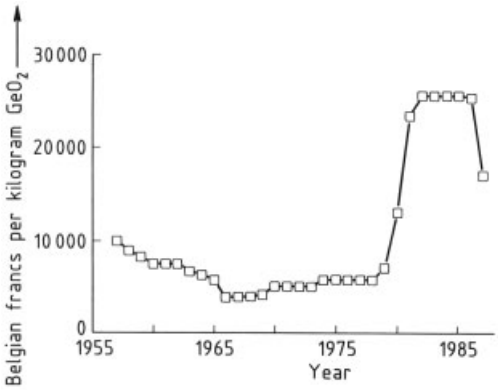

Figure 9. Belgian producer price for electronic-grade germanium dioxide (June 30 values from *Metal Bulletin;* the Belgian franc is the usual currency for the quotation of GeO_2)

Another use of germanium, although limited virtually to Japan, is the industrial production of organogermanium compounds such as bis(carbamoylethylgermanium) sesquioxide [*27031-32-9*], $(GeCH_2CH_2CONH_2)_2O_3$, for medical applications [66]. The use of germanium in alloys, e.g., high-precision resistors, is minor. The geographic production – consumption pattern for 1986 is summarized in Table 7.

Table 7. Production – consumption pattern for Ge in 1986

Country	Production, t	Consumption, t
North America	29	35
Western Europe	36	18
Asia	5	17

In consideration of the large internal scrap recycling coefficient of most production lines (see Section 5.2), care should be taken in evaluating the apparent refining overcapacity of many producers.

Figure 9 illustrates the producer price evolution of electronic-grade GeO_2, largely determined by the refining costs of low-grade ores. The price of germanium metal in its various commercialized forms reflects the stringent requirements for purity and crystallographic perfection and, therefore, depends strongly on the specifications.

11. Toxicology and Occupational Health

Germanium and most germanium compounds are comparatively low in toxicity because of pharmacological inertness, diffusibility, and rapid excretion. However, some exceptions exist, the most important being germanium hydride (germane). Surprisingly, soluble germanium compounds are more toxic by oral than by parenteral uptake. Intoxication is characterized by a lack of gross tissue changes in animals dying from oral uptake of germanium. Industrial exposure is due mainly to germanium fumes and dusts generated during production.

Absorption, Metabolism, Excretion, and General Effects. Gastrointestinal absorption of germanium oxides and cationic salts is poor; in contrast, the absorption of germanates from small oral doses is ca. 75 % in 4 h and 96 % in 8 h [67]. Inhaled germanates are quickly absorbed from the lungs into the blood [68]. Elemental germanium and germanium dioxide are excreted mainly via the urine in rats, rabbits, and dogs [69]. No reports of Ge accumulation in human or animal tissues exist. High exposure levels of germanium salts disturb water balance leading to dehydration, hemoconcentration, decrease in blood pressure, and hypothermia [70], without showing gross tissue damage.

Acute Toxicity. The symptoms of acute germanium toxicity in animals are listlessness, marked hypothermia, diarrhea, cyanosis, edema, hemorrhage of the lungs, petechial hemorrhage in the wall of the small intestine, peritoneal efflusion, and edematous changes in heart muscle and in parenchymal cells of the liver and kidneys [71]. Germanium tetrafluoride and tetrachloride vapors irritate the eyes and mucous membranes of the respiratory tract [72]. Organic compounds, such as dimethylgermanium oxide [*1112-60-3*], are only slightly toxic to rats [73], whereas triethylgermanium acetate [*5061-27-8*] shows considerable toxicity [74]. Germanium metal dust (particle size 1.7 µm, concentration unknown) and GeO_2 (particle size 0.45 µm, concentration unknown) had no adverse effects on rats after inhalation for 1 h [68]. In rats and mice, the acute oral toxicity of alkylgermanium compounds ranged between 300 and 2870 mg/kg [74], [75]. Triethylgermanium acetate has an LD_{50} value (rat, oral) of 250 mg/kg body weight. The LD_{100} for germane (mouse, inhalation) is 610 mg/m^3 [67].

Germanium hydride, GeH_4, is a colorless, highly inflammable, toxic gas of low stability with a characteristic unpleasant odor. By inhalation, gaseous germanium hydride is the most toxic of germanium compounds. Inhalation of GeH_4 (70 mg/m^3 , minimally effective concentration) for 2 – 15 d caused nonspecific and nonpersistent changes in the nervous system, kidneys, and blood composition. Like other metal hydrides such as AsH_3, it shows hemolytic action in animals [76]. The lethal concentration in air is 150 ppm [77].

Chronic Toxicity. Symptoms of chronic germanium intoxication in experimental animals are mainly fatty degeneration of the liver and inhibition of growth: 1000 ppm neutralized GeO_2 (pH adjusted to 7) given in the diet of rats for 14 weeks depressed growth and caused 50 % mortality; 5 ppm germanium administered as germanates in the drinking water of rats and mice resulted in decreased life-span and mild fatty degeneration of the liver; 100 ppm GeO_2 in the drinking water for 4 weeks caused 50 % mortality without hematological or gross pathological changes.

Exposure Limits. A TLV for germane of 0.2 ppm, equivalent to 0.64 mg/m^3, was fixed by the TLV Committee of the American Conference of Governmental Industrial Hygienists in 1973 as an average limit of permissible exposure during an 8-h working day. No other TLVs for germanium or its compounds have been set in industrialized countries. Neither germanium nor its compounds are considered or suspected to exert a carcinogenic effect on humans or animals [78].

12. References

1. *Landolt-Börnstein,* **GIII, 17 a** (1982) 87 – 126, 400 – 433.
2. *Gmelin,* System No. 45, Germanium Supplement Volume, 1958.
3. *Kirk Othmer,* 3rd ed. **11,** pp. 791 – 802.
4. M. Lesbre, P. Mazerolles, J. Satge: *The Organic Compounds of Germanium,* J. Wiley & Sons, London 1971.
5. Roskill: *The Economics of Germanium,* 4th ed., Roskill Information Service Ltd., London 1984.
6. U.S.B.M., Minerals Yearbooks, U.S. Dept. of the Interior, Washington.
7. G. Cote, D. Bauer, *Hydrometallurgy* **5** (1980) 149 – 160.
8. A. De Schepper, *Hydrometallurgy* **1** (1976) 291 – 298.
9. MHO, US 3 883 634, 1975 (A. De Schepper, A. Van Peteghem).
10. MHO, US 4 432 951, 1984 (A. De Schepper, M. Coussement, A. Van Peteghem).
11. MHO, US 4 432 952, 1984 (A. De Schepper, M. Coussement, A. Van Peteghem).
12. MHO, EP 68 540 B1, 1983 (A. De Schepper, M. Coussement, A. Van Peteghem).
13. MHO, EP 68 541 B1, 1983 (A. De Schepper, M. Coussement, A. Van Peteghem).
14. Peñarroya, US 4 389 379, 1983 (D. Bauer, G. Cote, P. Fossi, B. Marchon).
15. Peñarroya, US 4 568 526, 1986 (G. Cote, P. Fossi, B. Marchon, D. Rouillard).
16. Peñarroya, EP-A 46 437 A1, 1981 (D. Rouillard, G. Cote, P. Fossi, B. Marchon).
17. Peñarroya, EP-A 167 414 A1, 1986 (D. Bauer, G. Cote, B. Marchon).
18. Preussag, DE-OS 3 508 041 A1, 1986 (K. Hanusch, W. Krajewski).
19. Cominco, US 4 525 332, 1985 (D. L. Ball, D.A.D. Boateng, G. M. Swinkels).
20. *Ullmann,* 4th ed., **12,** 221 – 226.

21. P. Müllner, *Erzmetall* **30** (1977) no. 7/8, 326 – 329.
22. F. Hilbert, F. Stary, *Erzmetall* **35** (1982) no. 4, 184 – 189.
23. F. Stary, *Erzmetall* **35** (1982) no. 6, 311 – 313.
24. L. A. Painter, *Eng. Min. J.,* July1980, 65 – 88.
25. I. de Ruijter, *Photonics Spectra* **21** (1987) no. 7, 55 – 59.
26. J. L. Hopkins, *Can. Metall. Bul.* **78** (1985) no. 879, 86 – 87.
27. P. B. Queneau, P. Avotins, L. Farais, *Can. Metall. Bul.* **79** (1986) no. 886, 92 – 97.
28. J. E. Hoffmann, *J. Met.* **6** (1987) 42 – 45.
29. R. Hasegawa, *Trans. Nat. Res. Inst. Metals (Jpn.)* **20** (1978) no. 2, 111 – 127.
30. V. J. Silvestri, *J. Electrochem. Soc.* **116** (1969) no. 1, 81 – 87.
31. V. N. Vigdorovich: *Purification of Metals and Semiconductors by Crystallisation,* Freund Publ. House, Tel Aviv 1971.
32. W. G. Pfann: *Zone Melting,* J. Wiley & Sons, New York 1958.
33. K. Hein, E. Buhrig: *Kristallisation aus Schmelzen,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1983.
34. A. F. Bogenschütz: *Aetzpraxis für Halbleiter,* Hauser Verlag, München 1967.
35. P. J. Holmes: *The Electrochemistry of Semiconductors,* Academic Press, London 1962.
36. E. E. Haller, W. L. Hansen, F. S. Goulding, *Adv. Phys.* **30** (1981) no. 1, 93 – 138.
37. L. S. Darken, *J. Electrochem. Soc.* **126** (1979) no. 5, 827 – 833.
38. L. S. Darken, *IEEE Trans. Nucl. Sci.* **NS 26** (1979) no. 1, 324 – 333.
39. P. W. Bridgman, *Proc. Am. Acad. Arts Sci.* **60** (1925) 305.
40. J. Czochralski, *Z. Phys. Chem.* **92** (1917) 219 – 221.
41. F. Rosenberger: *Fundamentals of Crystal Growth,* (Solid State Science no. 5) Springer Verlag, Berlin –Heidelberg – New York 1979.
42. J. C. Brice: *Crystal Growth Processes,* Blackie, Glasgow 1986.
43. A. A. Chernov, *Modern Crystallography III, Crystal Growth* (Solid State Science no. 36) Springer Verlag, Berlin 1984.
44. J. A. Savage, *Proc. SPIE Int. Soc. Opt. Eng.* **274** (1981) 175 – 179.
45. C. J. Hutchinson, C. Lewis, J. Savage, A. Pitt., *Appl. Opt.* **21** (1982) no. 8, 1490 – 1495.
46. L. Van Goethem, L. P. Van Maele, M. Van Sande, *Proc. SPIE Int. Soc. Opt. Eng.* **683** (1986) 160 – 165.
47. E. Simoen, P. Clauws, J. Vennik, *J. Phys. D:* **18** (1985) 2041 – 2058.
48. E. Rotsaert, P. Clauws, J. Vennik, L. Van Goethem, *Proc. Int. Conf. Shallow Impurity Centers, 2nd,* Triest (1986).
49. G. Bambakidis, G. J. Brown, *Phys. Rev. B Condens. Matter* **33** (1986) no. 12, 8180.
50. E. Haller, W. Hansen, P. Luke, R. Murray, B. Jarrel, *IEEE Trans. Nucl. Sci.* **NS 29** (1982) no. 1, 738 – 744.
51. W. Hansen, E. Haller, P. Luke, *IEEE Trans. Nucl. Sci.* **NS 29** (1982) no. 1, 745 – 750.
52. G. Henze, *Fresenius Z. Anal. Chem.* **324** (1986) 105 – 110.
53. Y. A. Zolotov, M. Grosserbauer, C. H. Morrison, Y. Karpov, *Pure Appl. Chem.* **57** (1985) no. 8, 1133 – 1152.
54. C. Verlinden, R. Gybels, F. Adams, *J. Anal. Atom. Spectr.* **1** (1986) 411 – 419.
55. J. A. Savage: *Infrared Optical Materials and Their Antireflection Coatings,* Adam Hilger, Bristol 1985.
56. S. Musikant: *Optical Materials,* Dekker, New York 1985.
57. D. L. Wood, Kr. Walker, J. B. Macchesney, J. R. Simpson, R. Csencsits, *J. Lightwave Technol.* **LT 5** (1987) no. 2, 277 – 285.
58. D. Hameda, G. Tilloca, *Proc. SPIE Int. Soc. Opt. Eng.* **400** (1983) 44 – 49. G. Frischat, *Proc. SPIE Int. Soc. Opt. Eng.* **400** (1983) 73 – 77.
59. P. Hugget, H. Lehmann, *J. Electron. Mater.* **14** (1985) no. 3, 205 – 230.
60. K. Wolf, B. Küsters, H. Helinger, C. Tchang, E. Schollmeyer, *Angew. Macromol. Chem.* **68** (1978) 23 – 27.
61. K. Butler: *Fluorescent Lamp Phosphors,* The Pennsylvania State Univ. Press, Pennsylvania 1980.
62. A. Horowitz, G. Kramer, *J. Cryst. Growth* **78** (1986) 121 – 128; **79** (1986) 296 – 301.
63. J. I. Pankove: *Optical Processes in Semiconductors,* Prentice Hall, Englewood Cliffs, N.J., 1971.
64. M. Van Sande, L. Van Goethem, H. Guislain, *Erzmetall* **38** (1985) no. 12, 588 – 593.
65. O. J. Gregory, E. C. Tisman, *Symposium on Integrated Circuit Fabrication and Technology (Am. Chem. Soc.)* 1984.
66. K. Asai: *Organic Germanium, A Medical Godsend,* Kogakusha, Tokyo 1977.
67. B. Venugopal, T. D. Luckey: "Metal Toxicity in Mammals," *Chemical Toxicity of Metals and Metalloids,* vol. 2, Plenum Press, New York-London 1978.

68. H. C. Dudley, *AMA Arch. Ind. Hyg. Occup. Med.* **8** (1953) 528.
69. H. C. Dudley, E. J. Wallace, *AMA Arch. Ind. Hyg. Occup. Med.* **6** (1952) 263.
70. G. Rosenfeld, E. J. Wallace, *AMA Arch. Ind. Hyg. Occup. Med.* **8** (1953) 466.
71. G. H. Bailey, B. P. Davidson, C. H. Bunting, *JAMA J. Am. Med. Assoc.* **84** (1925) 1722.
72. G. C. Harold, S. F. Meek, *Ind. Med.* **13** (1944) 326.
73. E. G. Rochow, B. M. Sindler, *J. Am. Chem. Soc.* **72** (1950) 1218.
74. J. E. Cremer, W. N. Aldrigde, *Br. J. Ind. Med.* **21** (1964) 214.
75. NIOSH Registry of Toxic Effects of Chemical Substances, Washington, DC , 1976.
76. S. H. Webster, *J. Ind. Hyg. Toxicol.* **28** (1946) 167.
77. R. Kühn, K. Birett: *Merkblätter Gefährliche Arbeitsstoffe,* vol. 5, Blatt Nr. G 06, Verlag Moderne Industrie, München 1980.
78. Deutsche Forschungsgemeinschaft (ed.): *Maximale Arbeitsplatzkonzentration (MAK),* Liste A1, A2, B, 1987, VCH Verlagsgesellschaft, Weinheim 1987.

Ghatti Gum → Polysaccharides
Gibberellins → Plant Growth Regulators
Gilsonite → Asphalt and Bitumen
Gin → Spirits

Glass

Glass Ceramics and Metallic Glasses are separate keywords.

Bernard H. W. S. De Jong, Corning Incorporated, Corning, New York 14831, United States (Chaps. 1–4, and 7)

Ruud G. C. Beerkens, Eindhoven University of Technolgy, Eindhoven, The Netherlands (Chaps. 5, 6, and 8)

Peter A. van Nijnatten, TNO Institute of Applied Physics, Eindhoven, The Netherlands (Chaps. 5, 6, and 8)

1. Definition and Historical Summary . . . 3
2. Structure of Glass . . . 6
2.1. Structure of Special Melts and Glasses . . . 8
2.2. Composition of Glass . . . 10
2.2.1. Single-Component Glass . . . 10
2.2.2. Silicate Glasses with Two Components . . . 12
2.2.3. Borate, Phosphate, and Germanate Glasses . . . 15
2.2.4. Borosilicate and Aluminosilicate Glasses . . . 17
2.2.5. Lead Silicate Glasses . . . 18
2.2.6. Chalcogenide Glasses . . . 18
2.2.7. Halide Glasses . . . 19
2.2.8. Glasses with Complex Composition 19
3. Glass Formation, Crystallization, and Liquid Immiscibility . . . 19
4. Chemical, Mechanical, and Physical Properties of Industrially Important Melts and Glasses . . . 23
4.1. Rheological Properties of Glass Melts . . . 27
4.1.1. Temperature Dependence of Viscosity . . . 28
4.1.2. Time Dependence of Viscosity . . . 29
4.1.3. Composition Dependence of Viscosity . . . 29
4.2. Surface Tension . . . 31
4.3. Density . . . 31
4.4. Thermal Expansion . . . 32
4.5. Elastic Properties . . . 35
4.6. Mechanical Strength . . . 35
4.7. Hardness . . . 36
4.8. Thermal Properties . . . 36
4.9. Electrical Properties . . . 37
4.10. Gas Permeability . . . 39
4.11. Chemical Stability and Surface Properties . . . 39
4.12. Optical Properties . . . 41
4.12.1. Transmission, Absorption, and Reflection . . . 42
4.12.2. Refractive Index, Dispersion, and Optical State . . . 42
4.12.3. Temperature Dependence of Refractive Index . . . 44
4.12.4. Stress Optical Coefficients . . . 44
4.12.5. Radiation Effects . . . 45
4.13. Color of Glasses . . . 46
5. Production of Glass . . . 49
5.1. Raw Materials . . . 49
5.1.1. Chemical Impurities and Decoloration . . . 52
5.1.2. Waste Glass (Cullet) . . . 53
5.1.3. Preparation of Raw Materials . . . 54
5.2. Melting Units . . . 54
5.2.1. Flat Glass and Container Glass Tanks . . . 55
5.2.2. Unit Melters . . . 58
5.2.3. All-Oxygen-Fired Melters . . . 58
5.2.4. Electrical Melting Units . . . 59
5.2.5. New Melter Designs . . . 60
5.2.6. Pot Furnaces and Day Tanks . . . 60
5.3. Melting, Fining, and Homogenization . . . 61
5.4. Energy-Saving Technologies . . . 62
5.5. Glass Cooling . . . 63
5.6. Heating and Regulating Glass Melts . . . 64
5.7. Refractory Lining of Melting Units . . . 65
5.8. Vapor-Deposited Glasses . . . 66
5.9. Occupational Health . . . 66
6. Forming . . . 70
6.1. Hand Forming . . . 71
6.1.1. Blowing . . . 71
6.1.2. Pressing . . . 72
6.1.3. Blowing Combined with Pressing . 73
6.1.4. Drawing . . . 73
6.1.5. Redrawing . . . 75
6.1.6. Casting . . . 75
6.1.7. Fritting . . . 76
6.1.8. Flat Glass Forming . . . 76
6.1.8.1. Pilkington Float Process . . . 76
6.1.8.2. Rolling . . . 76

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_365

6.1.8.3. Overflow Process 77
6.1.8.4. Lamination 77
6.1.9. Fiberglass Forming 77
6.2. Annealing 78
6.3. Secondary or Finishing Operations 78
6.3.1. Cutting 78
6.3.2. Grinding 78
6.3.3. Polishing 79
6.3.4. Heat Processing 79
6.3.5. Sintering 79
6.3.6. Chemical Treatment 80
6.3.7. Surface Coating 80
7. Uses 81
7.1. Silica and Silica-like Glass 81
7.2. Flat Glass 82
7.2.1. Reflective Glass 82
7.2.2. Laminated Glass 82
7.2.3. Solar Glass 82
7.3. Laboratory Glassware 82
7.4. Light-Sensitive Glass 83
7.4.1. Photosensitive Glass 83
7.4.2. Photochromic Glass 83
7.5. Display Devices 83
7.6. Glass Fibers 84
7.7. Molded Optics 84
7.8. Glasses for Nonlinear Optical Devices 84
8. Economic Aspects 84
9. References 86

1. Definition and Historical Summary

Many molten materials do not crystallize to their parent crystalline phases once the thermodynamic melting temperature T_m is passed on cooling. Such melts easily supercool to a temperature far below T_m and congeal to solids without any attendant discontinuous changes in volume or entropy. These solids, which are isotropic in all their properties, are known as glass.

A common misconception is that all glass has the same composition. A variety of organic and inorganic materials can form glass, and most of those that do exhibit a moderately sharp transition into the glassy state from the liquid. Most common inorganic glass-forming systems are covalent compounds of the following elements:

Si, B, Ge, Sn, Pb;
P, As, Sb;
O, S, Se, Te.

Some elements (e.g., Se, Te, and S) also readily form glasses.

The canonical definition of the term glass is that by MOREY [54]: "Glass is an inorganic substance in a condition which is continuous with, and analogous to, the liquid state of that substance but which, as the result of a reversible change in viscosity during cooling, has attained so high a degree of viscosity as to be, for all practical purposes, rigid." Similarly, the ASTM defines glass as "an inorganic product of fusion that has cooled to a rigid condition without crystallizing."

Ease of glass formation is almost entirely a kinetic question. At no temperature is glass an equilibrium phase, although glasses hundreds of millions of years old are not uncommon in nature. The cooling rate required to form a glass varies with composition. Some melts (e.g., those of SiO_2 or B_2O_3) can form a glass at very slow rates of cooling, whereas mixed nitrates of potassium and calcium or sulfates of potassium and zinc, for instance, require rapid cooling from their normal melting temperature to form glass. Glassy metals have become quite common with the advent of rapid (splat) cooling techniques, the easiest to form having the approximate formula M_xL_y, where M is a transition metal; L is Si, P, C, B, or Ge; $x = 0.7 - 0.8$; and $y = 0.3 - 0.2$ (see → Metallic Glasses). Glasses can even be made from melts of metals such as iron and cobalt, but still higher quench rates are required than those for metal – metalloid glasses. Almost any inorganic melt can apparently be quenched into a glass with a sufficiently fast quenching rate.

Glass is not merely a supercooled liquid. The distinction between a supercooled liquid and a glass lies in the ability of structural elements in the former to rearrange themselves in accordance with the thermodynamic state of the system, whereas in the latter such rearrangement is not possible. This inhibition of rearrangement in glass is caused by the large increase in viscosity on cooling and gives rise to an endothermic effect on heating, which occurs at the glass transition temperature (i.e., the temperature above which structural elements in the glass are suffi-

ciently mobile to rearrange themselves according to their equilibrium configuration).

Literature on glass is abundant. The structure and content of this article are based on encyclopedia articles by GLIEMEROTH and MÜLLER [1] and by BOYD and THOMPSON [2]. Books on the scientific and manufacturing aspects of glass are the ones by BALTA and BALTA (1976) [3], DUNKEN (1981) [4], FRISCHAT (1975) [5], MCLELLAN and SHAND (1984) [6], PAUL (1982) [7], TRIER (1987) [8], TURKDOGAN (1983) [9], VOGEL (1985) [10], the Proceedings of the Conferences on Glass Problems edited by DRUMMOND (1986) [11], the Proceedings of the First International Conference on Advances in the Fusion of Glass edited by BICKFORD et al. [12] and the ones edited by BOYD and MCDOWELL (1986) [13], TOOLEY (1974) [14], and GIEGERICH and TRIER (1969) [15], [16]. The last book, though fairly old, is still the standard work on all glass-forming machines except for float glass; the primary difference between then and now being the optimization of economic performance of glass machines. A fairly extensive review of the spectroscopic properties of glasses is found in WONG and ANGELL (1976) [17]. A review of the spectroscopic techniques of choice for silicates can be found in the book edited by HAWTHORNE (1988) [18]. For the physics of silica the book edited by PANTELIDES (1978) is still the primary source [19]; for the chemistry of silicon the book by ILER (1979) [20] and the one edited by COREY et al. (1988) [21]. Among the literature on colors in glasses and minerals the books by WEYL (1961) [22], BURNS (1970) [23], BRAMFORD (1977) [24] and NASSAU (1983) [25] and the articles by FRITSCH and ROSSMAN (1987, 1988) [26] – [28] and SCHREURS and BRILL (1984) [29] should be consulted. The older literature on glasses is extensive and includes a review article by KAUZMANN (1948) [30] and UHLMANN (1983) [31], and books by UBBE-LOHDE (1965) [32] and RAWSON (1966) [33]. Pertinent classical references are the works by MOREY (1938) [34], TAMMAN (1925) [35], and WEYL and MARBOE (1962 –1964) [36]. Some sense of developments in the physics of amorphous materials, including spin glasses, and attempts to grasp their character in terms of localization and percolation can be gleaned from the thirty-first session on ill-condensed matter at Les Houches edited by BALIAN et al. (1978) [37]. Compilations of glass systems are found in the Phase Diagrams for Ceramists vol. I – V (1973 – 1985) [38] and in MAZURIN's books on phase diagrams (1973 – 1980) [39] and glass properties (1983 – 1987) [40]. Physical properties of glasses can be found in handbooks written by BANSAL and DOREMUS (1986) [41], CARMICHAEL (1982) [42], and CLARK (1966) [43]. BARIN and KNACKE (1973) [44], CHASE et al. (1985) [45], KUBACHEWSKI and EVANS (1958) [46], and ROBIE et al. (1979) [47] provide a wealth of thermodynamic data on silicate systems as does the definitive book on the structure of silicates by LIEBAU (1985) [48]. Phosphate mineralogy is covered in the book by NRIAGU and MOORE (1984) [49], its chemistry by VAN WAZER (1958) [50]. The Slag Atlas (1981) [51] gives a fairly extensive compilation of properties as well as an excellent synopsis of all property measurement techniques for glass systems of interest to the welding and metal industry. Two more books on glass are the ones edited by LEWIS (1981) [52] and MAZURIN (1989) [53].

The discoverer of glass manufacture will probably remain unknown. A source of inspiration may have been the abundant occurrence of glasses in nature. Obsidian, pumice (a natural foam glass), and tektites (glassy bodies probably of meteoric origin) are examples of naturally occurring glass. The earliest tektites worked by humans date from the Magdalenian period about 25 000 years ago. A synopsis of the developmental history of glass follows.

	b.c.
~2000	Glass was manufactured in the Caucasus.
~1500	Small artifacts were made in Atchana, Alalakh, and Tell al Rimah.
~1480	In what was probably the first occurrence of glass in Egypt (THUTMOSE III: Amarna period), alkali silicate glasses were colored with CaO, $Pb_2Sb_2O_7$, Al_2O_3; chromophores were Cu, Fe, and Mn. Asian glassmakers settled in Egypt.
~1450	Beautiful glass vases were made, inlaid with multicolored filigree glass.
~1200	Glass technology declined in Asia minor (Egypt, Mesopotamia).
~900	New glass centers started on the coast of Syria and in Mesopotamia.

~700 – 600 Glass making was established on Rhodes and Cyprus.
~630 Clay tablets in the Sardanapal library in Assyria describe use of tin as opacifier, production of alkali from plants, manufacture of frits and colorants.
~500 First Chinese glass artifacts date from this period.
~332 ALEXANDER THE GREAT supported expansion of Syrian technology.
~250 The blowpipe was discovered.

a.d.
Torso of an Aphrodite from Alexandria, cut from green glass, dates from this time.
First glass retorts were made for alchemists.
~70 Glassworks were constructed in Spain and France (glass manufacture was imported by the Romans from Egypt).
79 PLINY described the manufacture of glass.
100 Glass was manufactured in Köln.
100 Glass vessels were sold in Syrian shops in Sidon.
First vessels with relief were blown in molds.
300 Late Roman glass art from this time has been found in Christian catacombs.
~500 Syrian glass shops manufactured vessels with glass filament supports, head vessels with cut and etched patterns.
591 GREGORY OF TOURS mentioned glass church windows.
674 Glass windows were used in English churches.
~700 Window glass was manufactured in Syria by blowing, cutting, and sagging.
~700 Alkali content was lowered with respect to P_2O_5 (in bone ash).
~1000 Glass windows were made for Tegernsee cloister.
~1100 Small hand-held bend mirrors were introduced.
1180 Glass windows were first used in private English houses.
1202 Zenith of melt painting with copper containing colored glasses in the middle Rhine region.
~1300 Glass windows became more common in Europe.
1405 Silver engravings with glass layers were produced in Italy.
1453 Glass-making knowledge moved from Byzantium to Venice.
~1500 Clear glass was manufactured by glass shops in Murano, particularly wine glasses exported to Germany. Glass was decorated with pearls and buttons, painted, and gold plated.
~1500 Diamond was used to cut glass.
1507 ORLANDO GALLO improved Venetian glass technology. Mirrors were produced. Zenith of Venetian glass manufacture occurred.
~1550 Late blooming of Gothic glass painting appeared in Chartres and Paris.
1550 Glass thermometer was manufactured in Venice.
1590 Lenses were produced, presumably by ZACHARIAS JANSSEN.
1665 Large mirrors were made in France.
1680 JOHAN KUNKEL (1630 – 1702) manufactured ruby glass.
1688 Mirror glass was poured.
1697 Bohemian glass was produced in Venice.
1728 REAMUR observed devitrification.
1754 KLINGENSTJERNA disproved the thesis of NEWTON that refractive index is independent of wavelength.
1757 DELLOND constructed achromatic objective lenses.
1782 Industrial temperature measurement became possible with the Wedgwood pyrometer.
~1800 The first attempts were made to understand the constitution of glass.
1834 LENG interpreted devitrification as crystallization of silicic acid salts.
1845 Pocket watch industry started in a glass shop.
1857 The first glass oven with regenerative heating was developed.
1859 Semiautomatic bottle-blowing machine was used.
1867 Continuous tank was introduced.
1869 Silica was first drawn.
1879 The first blanks for incandescent lamps were made.
1900 Cylindrical blow machines were used for window manufacture.

1903 OWENS invented suction machine for bottle manufacture.
1904 The first neon light fixtures were made.
1906 FOURCAULT developed sheet-draw machine.
1909 A 1.52-m telescope mirror was manufactured for Mount Wilson observatory.
1915 Pyrex was introduced.
1917 Tank-drawn sheet glass appeared (Colburn machine).
1922 Single gob feeder was developed.
1925 Pittsburgh process for sheet glass was introduced.
1925 Draw process was used for tubes.
1926 Ribbon machine was developed for incandescent lamp bulb blanks.
1934 World's largest piece of glass was cast for 200-inch (5.08-m) telescope mirror on Mount Palomar, California.
1949 Centrifugal casting of television tubes began.
1950 First photosensitive glasses were made.
1967 Floating glass process was invented by PILKINGTON.
1970 Optical fibers were demonstrated with 20-dB/km loss.
1972 Vertical melting units were developed.
1978 Polychromatic glass was invented.
~1980 Development of fluoride-based glasses for optical fibers.
1983 Sol – gel extruded mullite fiber was invented.
1984 Pressing of elongated automobile headlights.

2. Structure of Glass

Methodology in condensed-matter physics and chemistry consists of identifying and relating the physical properties, structure, and constituent elements of a class of materials. For crystalline solids, the constituents and structure can be characterized readily. In glass, on the other hand, understanding of the structure consists almost exclusively of negative statements: no metric geometry, trivial space group, no Bloch states, no single ground state, no unique best structure. In addition, the topology of glass structure is almost amenable only to indirect experimental investigation.

It is not surprising in light of the undetermined nature of the amorphous state of glass that competing perceptions exist with regard to its character. The principal difficulty in distinguishing between models has been the small amount of energy associated with long-range ordering. For instance, the enthalpy of formation of quartz from silicon and oxygen at 298 K is − 860 kJ/mol; this is only 12 kJ/mol more negative than that for vitreous fused silica (− 848 kJ/mol). Only 1 % of the enthalpy of formation of quartz is therefore associated with long-range ordering. The two predominant structural models for glass are the microcrystallite and the random network hypotheses.

Microcrystallite Hypothesis. The microcrystallite "cybotactic group" hypothesis was constructed primarily to account for the discontinuous changes in the properties of glasses which can be correlated with similar discontinuities in the properties of associated crystalline phases. Proponents of this hypothesis were (with some rare exceptions such as STEWART and RANDALL [55], almost exclusively Russian, the first being FRANKENHEIM (1835)
and subsequently VON WEIMARN (1907)
who held that all matter in any state, be that gas, liquid, or solid is crystalline [56]. This work was followed up by LEBEDEFF (1921), who showed that vitreous amorphous silica varies considerably between 540 and 600° C in its double refraction, refractive index, and coefficient of expansion [57]. He inferred that this variation was due to a polymorphic transformation assumed to be connected with the low – high temperature transformation of quartz, which occurs at 575° C. He furthermore suggested that vitreous silica consists of an aggregate of very minute crystals with included quartz crystals that are probably not in a pure state, but in the form of a solid solution with other substances, hence accounting for the temperature range over which the polymorphic transition occurs. VALENKOV and PORAY-KOSHITZ (1936) analyzed the X-ray diffraction curves of glasses and concluded that the observed diffraction patterns [58] were produced by 0.75 – 2.5-nm crystallites that were connected to each other by an amorphous layer. These results made the microcrystallite postulate questionable because 2.5-nm crystallites do not represent long-range ordering in a crys-

tallographic sense. Refinement of diffraction experiments, particularly at small angles, has invalidated the original microcrystallite hypothesis. Recent electron microscopy results, on the other hand, indicate that multicomponent glasses can be chemically inhomogeneous, and contain microheterogeneities which could be taken to represent microcrystallites.

Random Network Hypothesis. The Zernike–Prins–Bernal–Fowler–Warren hypothesis [55] considers the liquid a system of atomic or ionic networks, designated as a random network. For glasses this implies a rigid system of continuous noncrystalline networks similar to those assumed to be present in the liquid. It remains open to debate whether there is one unique random network or perhaps more than ten thousand or what random really means in the context of liquid structures. The hypothesis was successful in providing qualitative explanations for the glass-forming tendency of simple glass-forming systems as being caused by a rapid increase in viscosity during cooling of the melt. In addition successful predictions could be made about potential glass-forming systems as well as rationalizations for most of property-composition relationships. Zachariasen (1933) postulated glass-forming tendencies for simple oxides based on random networks and the concomitant need for flexibility in linkages [59]. These postulates, in part already phrased by Goldschmidt (1926) for onide glasses [60], are the following:

1) Each oxygen atom can be linked to no more than two cations; the number of oxygen atoms around any one cation must be small, i.e., three or four; the oxygen polyhedra must share corners, not edges or faces, to form a three-dimensional network; at least three corners must be shared.
2) Network formers have a coordination number of 3 or 4 (Si, B, P, Ge, As, Be, etc.); network modifiers have coordination numbers greater than or equal to 6 (Na, K, Ca, Ba, etc.). Intermediate atoms with a variable coordination number between 4 and 6 (Li, Al, Mg, Zn, Pb, Nb, Ta, etc.) can function as network modifiers as well as network formers; the coordination change is a reflection of their amphoteric character, i.e., solubility in both inorganic acids and bases.

Additional refinements have been made to the random network theory. Dietzel (1942) introduced the concept of cationic field strength defined as formal cation charge divided by the cation–anion distance [61]. This concept enables a quantifiable distinction to be made between the three categories of atoms:

1) network formers with a field strength of 1.4–2 N/m,
2) network modifiers with a field strength of 0.1–0.4 N/m, and
3) amphoteric atoms with a field strength of 0.5–1.0 N/m.

Glass formation is favored if the difference in field strength between the cations is larger than 0.3. Smekal (1949) expanded on the cationic field strength concept by taking the bond polarity into account [62]. Glass formation is favored if mixed bonding with a homopolar and heteropolar component occurs. Weyl (1958–1959) postulated the screening theory used to explain the mechanical properties of glasses [63]. In this theory the polymerization of polyhedra to threedimensional random networks is caused by the polarization and deformation of residual valence forces. Stevels (1948) developed a theory for invert glasses, which contain polyhedra with less than three connected corners in contradiction to the Zachariassen postulates [64]. His structural parameter Y is a measure of the average number of bridging anions. The properties of the glass are mainly determined by the network modifier atoms when Y is smaller than 2. This is due to the fact that these network modifiers compete for the available anions, thus resulting in an increased degree of disorder.

Among the less fruitful extensions of the random network theory are those by Huggins and Tilton. Huggins (1955) attempted to explain known discontinuities in properties by assuming the presence of fixed atom groups, called structons [65]. A last refinement of the network hypothesis represents the work by Tilton (1957), who postulated the presence of vitrons: aggregates of silica tetrahedra primarily consisting of five-membered rings, which combined to form a pentagondodecahedron consisting of 20 tetrahedra [66]. These vitrons would be connected to one another by less ordered domains.

With the advent of NMR, laser Raman spectroscopy, and extended X-ray absorption fine structure (EXAFS), focus has been directed to dissecting glass structure in terms of local environments, avoiding any deterministic statements regarding extended structures. An area of research involves determining the extent to which long-range structural elements can be derived from such local environments.

2.1. Structure of Special Melts and Glasses

Almost all industrially manufactured glasses are silicate glasses. As a consequence, these systems are emphasized in the following discussion, the goal being to provide a terminus a quo rather than a terminus ad quem for the study of amorphous materials. The viscosity of various silica-containing melts is used as a reference against which structural concepts can be matched. The use of viscosity is somewhat ironic, however; although viscosity illustrates nicely the role of network-forming atoms in comparison with network-modifying atoms, it is the least understood property of melts or glasses. The structural description of glasses and melts is based on Q (quartz) distribution theory and Q designation. The use of this model enables description of silica species in both solids and aqueous solutions in terms of the distribution of local silicon environments, which are amenable to investigation by NMR. These designations are as follows (Fig. 1):

1) Each silicon atom is coordinated tetrahedrally to four oxygen atoms.
2) If all oxygen atoms in a tetrahedron are connected to two silicon atoms, the local environment around the silicon atom is designated Q_4. All four Si−O bonds of the tetrahedron are, therefore, bridging bonds, designated Si−O(br).
3) The local silicon environments are designated Q_3, Q_2, Q_1, and Q_0 if three, two, one, or zero oxygen atoms are connected to two silicon atoms. In Q_3, three Si−O(br) bonds and one Si−O nonbridging bond, designated Si−O(nbr) exist. For Q_0, all Si−O bonds are Si−O(nbr).
4) Designations Q_4 to Q_0 of local environments coincide with connectivity 4 to connectivity 0 of the extended environment.

Four reactions suffice to describe all possible rearrangements in local environments in those alkali oxide (R_2O) silicate glasses in which all oxygen atoms are connected to silicon atoms [69]:

1) $2\,Q_4 + R_2O \rightleftharpoons 2\,Q_3$ (depolymerization)
2) $2\,Q_3 \rightleftharpoons Q_2 + Q_4$
3) $Q_2 + Q_3 \rightleftharpoons Q_1 + Q_4$ (stepwise condensation)
4) $Q_1 + Q_3 \rightleftharpoons Q_0 + Q_4$

The Q distribution can in principle be determined experimentally by high-resolution solid-state NMR. This distribution does not, however, necessarily relate glass structure to glass properties. Examination of Figure 2 makes this clear. Two hypothetical two-dimensional networks, each containing 45 open tetrahedra and 11 black tetrahedra, are shown. The open tetrahedra represent silicon; the black tetrahedra, lithium in fourfold oxygen coordination. The difference in Q distribution between the two networks is small. Nevertheless, measurement of the transport properties of the two networks would give wildly different results; the arrangement in Figure 2B has very high electrical conductivity and diffusivity (and possibly low viscosity) because of a continuous pathway through the solid in contrast to that in Figure 2A. Clearly the Q distribution does not correlate glass structure with glass properties and additional information is required to make such a connection. This additional information is obtained from Monte Carlo type computer calculations.

Keefer [70] recognized that the question of the structure of a silica-rich alkali-metal silicate glass is analogous to that of electron spin correlation in a two-dimensional Ising model. This model attempts to explain magnetic properties of materials by constructing a lattice with two possible electron spin states, up or down, which allows the problem to be considered in terms of a statistical distribution. Alkali silicate and spin glasses share two characteristics: (1) slow relaxation and (2) a tendency to settle into any one of a large number of atomic configurations. In silica-rich alkali-metal silicate glasses, the two possible spin states correspond to bridging and nonbridging oxygen atoms; silica-rich glass is required because photoelectron spectroscopy has

demonstrated that only in such systems are nonbonding oxygen atoms not present. In the absence of nonbonding oxygen, the ratio between bridging and nonbridging oxygen atoms is fixed by the stoichiometry of the sample.

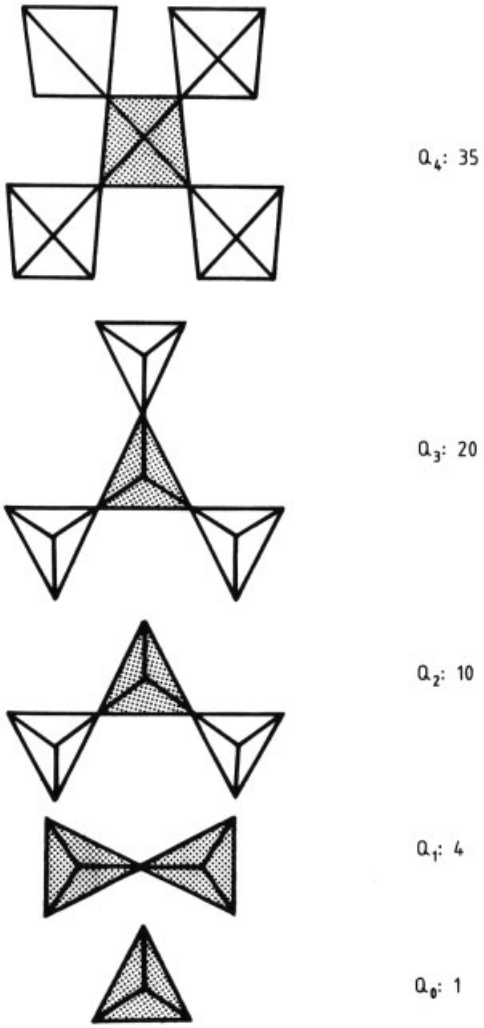

Figure 1. Structure and designation of silica species; possible number of local environments around a silicon atom up to the third-nearest neighbor [67], [68]

As stated before, connecting Q distribution with physical properties of a glass or melt requires Monte Carlo type calculations. The computational problem involves assignment of the enthalpies of formation of the four principal reactions between Q species; calculation of network patterns; derivation of the thermodynamic properties of the system from these patterns; and use of percolation theory to compute the rheological and transport properties of melts [72]. A rule of thumb is given by Ziman with respect to site percolation [73]: site percolation occurs in a regular three-dimensional assembly when favorable regions occupy ca. 15 % (in the two-dimensional case 45 %) of the total volume.

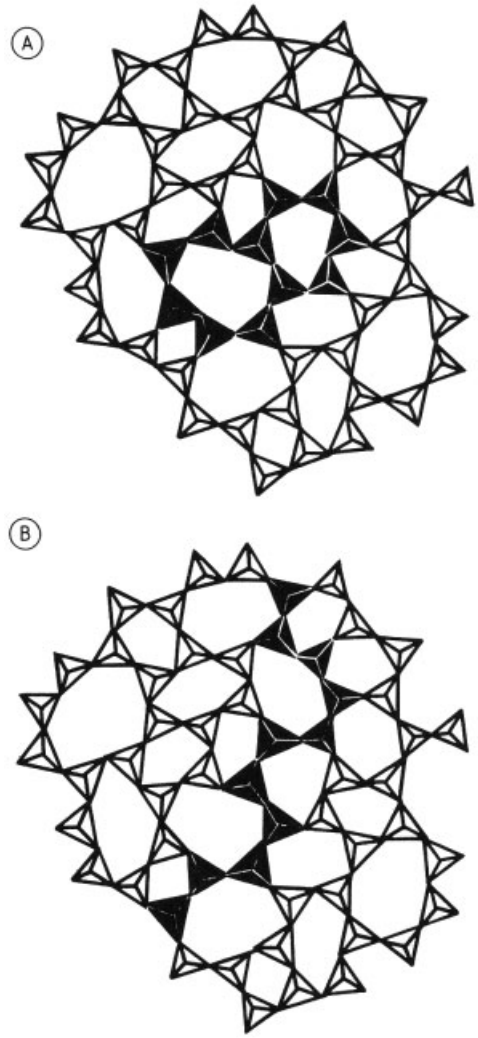

Figure 2. Two-dimensional nets of silica tetrahedra (△) interspersed with lithium tetrahedra (▲)
A) 36 Q_3, 9 Q_2; B) 33 Q_3, 13 Q_2 Courtesy of J. Dickinson, Corning Incorporated

Local silicon environments can be measured by ^{29}Si magic angle spinning (MAS) NMR. Two problems arise in using this technique: variations in chemical shift for a specific silicon species and, partly as a result of this, limited spectral

resolution. The first problem is illustrated in Figure 3, which shows the ^{29}Si MAS NMR spectrum of $Na_{24}Y_8[Si_{24}O_{72}]$, which is a single-chain silicate with 24 silica tetrahedra per unit cell (a 24er chain in the Liebau classification) and local environment Q_2. The ^{29}Si MAS NMR spectrum of this compound shows 12 peaks with a chemical shift range of 5.12 ppm. Thus, small perturbations in the local environment around one Q_2 species can cause a substantial chemical shift. In this case, the variations are due to chain twisting [71]. Such variations in chemical shift of a single silica species, combined with the large number of species possibly present (ten Q_2 species are possible) and the limited NMR spectral resolution in glass, render determination of the silica species distribution in glass tentative at best. Despite this, some assessment of the degree of intermediate-range order in lithium silicate glasses has been obtained by combining NMR and ESCA (electron spin chemical analysis) results together with the application of proper stoichiometric constraints on the system [69]. Doing this shows clearly that the ^{29}Si NMR spectra of such glasses indicate the presence of only a small number of local environments, in contrast to the total possible number enumerated in Figure 1.

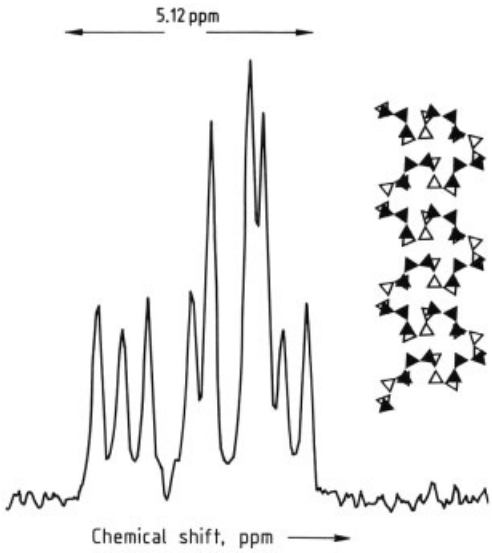

Figure 3. The ^{29}Si MAS NMR spectrum of $Na_{24}Y_8[Si_{24}O_{72}]$
δ (ppm) $= -26.793 - 0.4255\,\theta$ where δ is the chemical shift and θ the $\langle Si-O-Si \rangle$ angle.

2.2. Composition of Glass

Multicomponent glass-forming systems contain, according to Zachariasen, appreciable amounts of elements that form vitreous oxides or other elements that can replace the former isomorphously. Ternary glasses can consist of one network former and two modifiers, two network formers and one modifier, or three network formers. Most ternary oxide systems have been sampled, and investigation has been extended to quaternary and quinary systems.

The properties of complex industrial glasses tend to be rationalized in terms of simple systems. The following discussion, therefore, begins with vitreous silica as the canonical single-component glass, followed by simple two-component systems: the alkali-metal and alkaline-earth silicate, borate, phosphate, and germanate glasses. Boro-, alumino-, and lead silicate glass is then discussed, followed by two nonoxide systems, the chalcogenide and halide glasses. Finally, the chemical composition of industrially important glasses is tabulated.

2.2.1. Single-Component Glass

The most important single-component glass is vitreous silica. The structure consists of corner-sharing SiO_4 tetrahedra. Each oxygen atom is shared by two silicon atoms, forming Q_4 local silicon environments. Long-range order is, of course, absent. Vitreous silica is an excellent dielectric with a very low equilibrium solubility (on the order of 100 ppm at room temperature) in all acids except hydrofluoric. A synopsis of the properties of Corning 7940 fused silica and other amorphous silicas is given in Table 1.

The difference in structure between vitreous silica and two crystalline silicas is shown in Figure 4. Here the frequency of occurrence of average Si−O−Si angles is illustrated by using the ^{29}Si MAS NMR spectrum of vitreous silica [76] and the formula of Thomas and coworkers [77], which relates chemical shift to the Si−O−Si angle. Average Si−O−Si angles vary between 122° and 170° in vitreous silica, the most common value being 147° (or 151° if the spectrum of Gladden and coworkers is used [78]). On the other hand, quartz and cristobalite each have

Table 1. Properties of Corning 7940 fused silica and low-water (<100 ppm) vitreous silica [74], [75]

Softening point, K	1853
Annealing point, K	1357
Strain point, K	1229
Thermal expansion coefficient (273 – 573 K), K^{-1}	5.5×10^{-7}
Density, g/cm	2.20
Refractive index	1.459
Log volume resistivity	
523 K	11.8
623 K	10.2
Dielectric constant	3.8
Loss tangent (1 MHZ, 293 K), %	0.001
SiO vapor pressure over SiO_2, kPa	
1000 – 3000 K (vacuum)	$\log p\,(\mathrm{SiO}) = 101.3\left[11.37-\frac{2.7117\times10^4}{T}\right]$
1000 – 3000 K (20.3 kPa partial oxygen pressure)	$\log p\,(\mathrm{SiO}) = 101.3\left[12.92-\frac{4.066\times10^4}{T}\right]$
Maximum crystallization rate of cristobalite from vitreous	
silica (1948 K), cm/sec	2×10^{-7}
Surface tension (calculated), N/m	
298 K	5.200
2073 K	0.3
1149 – 2341 K	$\sigma = 0.236212 + 3.25\times10^{-5}T$
Thermal conductivity (298 K), W m^{-1} K^{-1}	1.37
Heat capacity (298 – 1500 K), J mol^{-1} K^{-1}	$74.369-\left(7.2594\times10^{-3}T\right)+\left(5.5704\times10^{-6}T^2\right)-\left(3.114\times10^{-6}T^{-2}\right)^*$
Heat of formation (298 K), kJ/mol	−903.2
Free energy of formation (298 K), kJ/mol	−850.559
Entropy of formation (298 K), J mol^{-1} K^{-1}	47.40
Heat capacity (298 K), J mol^{-1} K^{-1}	37.94
Viscosity (1873 – 2773 K), 0.1 Pa · s	$\log\eta = -6.24+\left(2.69\times10^4\right)/T$
Young's modulus, GPa	72.9
Shear modulus, GPa	31.3
Poisson ratio	0.165
pH at point of zero charge	2

∗ Data for vitreous silica.

one single Si−O−Si angle, with values of 143° and 146°, respectively.

Superimposed on the plot of Si−O−Si angle versus frequency in Figure 4 is the two-center energy, taken to reflect the bond energy, for a silicon – oxygen bridging bond as a function of Si−O−Si angle, calculated by using semiempirical molecular orbital calculations [79]. Inspection of this curve shows that a Si−O bond with a Si−O−Si angle of 180° is stronger (i.e., has a more negative two-center energy, ca. − 1760 kJ) than that at a Si−O−Si angle of 120° (two-center energy, ca. − 1530 kJ). Examination of Figure 4 suggests that although glasses are intrinsically less stable than their crystalline counterparts, a vitreous silica more stable than quartz or cristobalite is conceivable; the challenge is to make a glass with a larger percentage of Si−O−Si angles >150°. Only this angle must be considered because the energy barrier to rotation between connected silica tetrahedra is negligibly small.

Fused silica transmits light of wavelengths into the UV region and is the oxide glass most resistant to damage caused by radiation. It finds use in windows for space vehicles and wind tunnels, ultrasonic delay lines, crucibles for growing ultrapure silicon or germanium crystals, optical fibers, and optical systems in spectrophotometric equipment.

Vitreous silica can be produced by several methods. These processes and the impurities associated with them are listed in Table 2.

Fused quartz made by electrically fusing quartz crystals has very low moisture content and, hence, good IR transmission for wavelengths < 4500 nm. The disadvantage of quartz as a raw material is that even high-grade pure

quartz crystals contain 1 – 2 ppm of aluminum, alkali, manganese, and titanium. The presence of these elements increases the number of Si–O(nbr) bonds, which thereby reduces UV transmission for wavelength < 4500 nm. Flame fusion of quartz or flame hydrolysis of $SiCl_4$, on the other hand, gives glasses of very high purity except for large amounts of water, which decrease IR transmission. For some applications, impurities as low as 1 part in 10^{12} become important. An example is uranium; this problem is best solved by ensuring that the silica source is biogenic rather than associated with granitic rocks or pegmatites.

Table 2. Impurity levels in various grades of fused silica

Production method	Impurity, ppm	Trade name
Flame fusion of quartz	Al 10	
	B <0.5	OG, OH Vitreosil (Thermal Syndicate), Herosil, Homosil, Vitrasil (Heraeus)
	Ca 0.35	
	Fe 0.51	
	Na 0.06	
	OH 400	
Electric fusion	Al 50	IR Vitreosil (Thermal, Syndicate), Infrasil (Heraeus), Pursil (Quartz et Silice)
of quartz	Sb 0.23	
	B <0.5	
	Ca 0.4	
	Fe 0.74	
	Na 4.0	
	OH 3.0	
Vapor-phase hydrolysis	Al <0.02	Spectrosil (Thermal Syndicate), Suprasil (Heraeus), Tetrasil (Quartz et Silice)
in flame: $SiCl_4 \longrightarrow SiO_2$	Ca <0.1	
	Fe <0.1	
	Na <0.04	
	OH 1200	
High-temperature	Al <0.02	Spectrosil WF (Thermal Syndicate), Suprasil I (Heraeus)
vapor-phase oxidation: $SiCl_4 \longrightarrow SiO_2$	Ca <0.1	
	Fe <0.1	
	Na <0.04	
	OH 3	
	Cl 500	
Fusion of quartz	oxides 260 ppm (max.)	Ge 201
	Al_2O_3 180	
	water free	
Fusion of quartz	oxides 110 ppm (max.)	Ge 204 A
	Al_2O_3 52	
	CaO 23	
	water free	
Soot deposition on Pt mandrel by $SiCl_4$ flame hydrolysis;	Al <0.02	Corning 7943
	Ca <0.1	
	Fe <0.1	
fired on graphite mandrel in H_2	Na <0.04	
	OH 10	
Flame hydrolysis of $SiCl_4$	OH 1000	Corning 7940
Vacuum fusion of sintered Cab – O – Sil. powder	Al_2O_3 100	
	Na_2O <200	
	OH 2000	

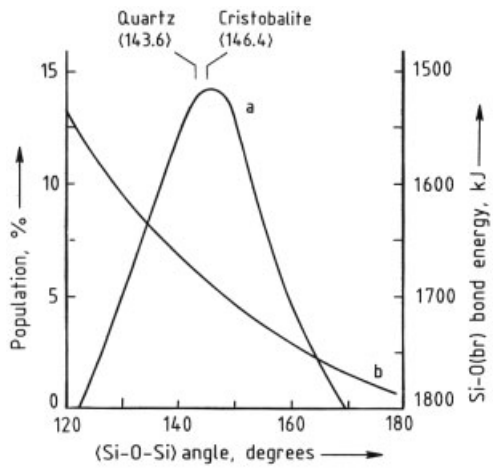

Figure 4. Difference in structure between vitreous silica and two crystalline silicas
a) Frequency of occurrence of average Si–O–Si angles in vitreous silica; b) Si–O bond energy as a function of Si–O–Si angle
Si – O – Si angle: Quartz 143.6°; Cristobalite 146.4°; Coesite 144.2°, 152.8°; Tridymite 146.7 – 157.2°.

2.2.2. Silicate Glasses with Two Components

Two-component silicate glasses are of particular interest in studying glass formation. Addition of alkali-metal or alkaline-earth oxides breaks Si–O–Si linkages, the alkali-metal or alkaline-earth nestling at nonbridging oxygen sites in the network, hence, the description network modifiers. This decrease in connectivity of the silica network manifests itself in a very large decrease

Table 3. Viscosity and other properties of alkali-metal (R_2O) and alkaline-earth (RO) silicates and silica [63], [80], [81]

Composition, mol % R_2O or RO	CAS registry no.	Cationic field force, 10^{15} N/m^2	Single bond energy, kJ/mol [82]	Viscosity, Pa · s [a]
0 (100 % SiO_2)	[*10097-28-6*]	1.57	463	1.47×10^4 [b]
				2.40×10^3 [c]
Li_2O	[*12057-24-8*]	0.23	150	
20.0				19.1
25.0				8.1
30.0				2.8
35.0				1.4
40.0				1.0
Na_2O	[*1313-59-3*]	0.19	84	
15.0				50.1
20.0				15.8
25.0				7.9
30.0				4.5
35.0				2.5
40.0				1.4
K_2O	[*12136-45-7*]	0.13	54	
16.9				35.5
20.0				27.0
22.3				22.9
25.7				11.5
37.0				6.0
Rb_2O	[*18088-11-4*]	0.12	50	
20.0				56.5
25.0				28.3
33.3				9.3
Cs_2O	[*20281-00-9*]	0.10	42	
20.0				100.0
25.0				39.9
33.3				10.0
MgO	[*1309-48-4*]	0.45	155	
44.3				0.46 [d]
45.8				0.26 [d]
51.3				0.15 [d]
CaO	[*1305-78-8*]	0.33	134	
43.7				0.24
48.7				0.15
52.7				0.10
SrO	[*1314-11-0*]	0.27	134	
40.5				0.48
44.3				0.28
50.9				0.20
BaO	[*1304-28-5*]	0.24	138	
40.2				0.54
42.1				0.33
49.8				0.19

[a] 1673 K for R_2O and 1973 K for RO compositions, respectively.
[b] 2197 K.
[c] 2333 K.
[d] Extrapolated.

in viscosity (see Table 3), melting point, and UV transmission. Modifiers also cause a decrease in resistivity, an increase in thermal expansion, and generally lower chemical durability.

Different alkali-metal atoms have different effects on the properties of silicate glasses. Thus, lithium silicate glasses are nonhygroscopic, whereas sodium, potassium, rubidium, or cesium silicate glasses show increasingly higher hygroscopicity. Another example is the volume contraction of the silicate network in the presence of lithium and its expansion in the presence of potassium. The almost monotonic variation in cation – oxygen distance, as measured by

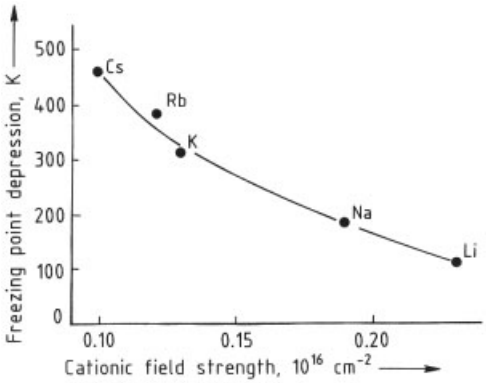

Figure 5. Freezing point depression for alkali-metal silicate melts ($10\,R_2O \cdot 90\,SiO_2$) as a function of cationic field strength

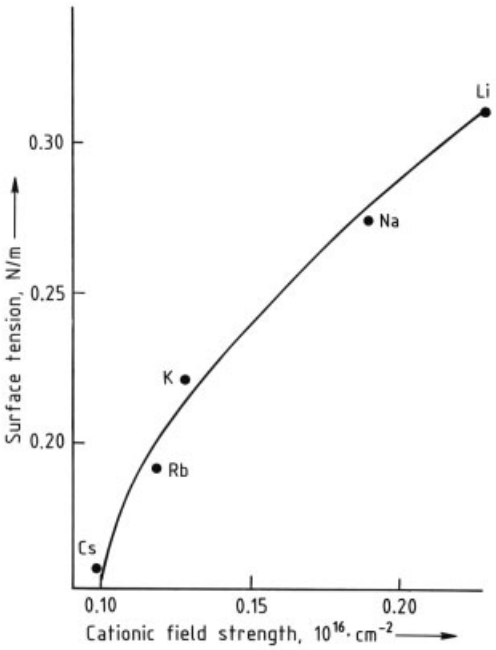

Figure 6. Surface tension of alkali-metal silicate melts ($20\,R_2O \cdot 80\,SiO_2$) at 1673 K as a function of cationic field strength

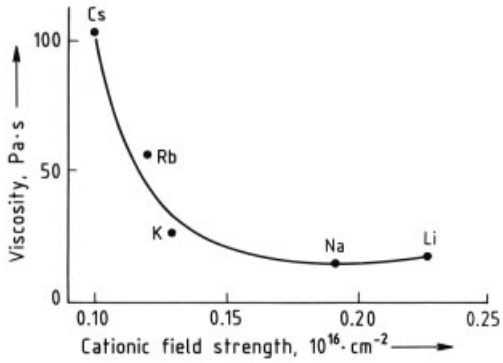

Figure 7. Viscosity of $20\,R_2O \cdot 80\,SiO_2$ melts at 1673 K as a function of cationic field strength

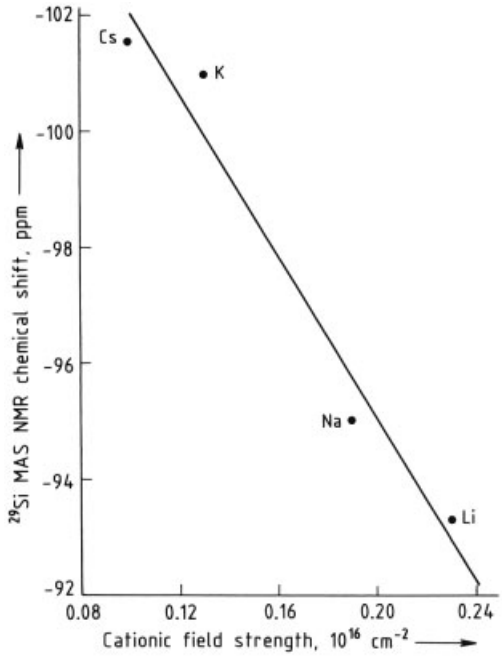

Figure 8. ^{29}Si MAS NMR chemical shifts of $33\,R_2O \cdot 67\,SiO_2$ (R = alkali metal) crystalline silicates as a function of cationic field strength of the alkali-metal atom
The observed chemical shifts relative to tetramethylsilane ($SiMe_4$) are −93.3, −95, −101, and −101.5 for crystalline lithium, sodium, potassium, and cesium disilicate respectively

X-ray diffraction, has led to a variety of schemes for the comprehensive compilation of the properties of alkali-metal or alkaline-earth silicate crystals or glasses. Among these are the cation field energy and field strength, which are the formal cation charge divided by the cation – anion distance and the square of the cation – anion distance, respectively.

Many physical properties of alkali-metal or alkaline-earth silicate glasses change almost monotonically as a function of cationic field strength or energy. For example, the critical

liquid immiscibility temperature decreases linearly as a function of cationic field strength from magnesium to barium silicates, and from lithium to cesium silicates. Similarly, the freezing point depression increases more or less monotonically from lithium to cesium silicates (see Fig. 5), whereas the surface tension decreases (see Fig. 6). Viscosity, refractive index, and density (see Fig. 7) show a minimum for sodium silicate melts.

Even the few ^{29}Si MAS NMR chemical shift data on crystalline alkali-metal silicates suggest a relation between chemical shift and cationic field strength, less negative (downfield) chemical shifts being associated with greater cationic field strength (Fig. 8). More sophisticated analysis of a variety of silica-containing systems has been carried out [83], [84]. In alkali-metal or alkaline-earth silicate glasses, most physical properties vary more or less linearly with cationic field strength, with the exception of viscosity, refractive index, and density. Although it is useful as a mnemonic device, no true cause and effect relations are associated with cationic field strength or energy.

Table 4. Viscosity of alkali borate glass (1173 K) [86]

R_2O, mol %	Viscosity, Pa · s
0 (100 % B_2O_3)	14.2
Li_2O	
2.5	3.1
13.9	0.8
25.6	0.46
Na_2O	
3.0	3.0
15.7	1.4
24.5	0.8
K_2O	
2.1	3.7
15.9	1.5
23.5	1.0

2.2.3. Borate, Phosphate, and Germanate Glasses

Borate glasses contain planar BO_3 groups as structural units, rather than tetrahedral SiO_4 groups. The oxygen atoms are, as in silica, again connected to two network-forming atoms, in this case boron. Radial distribution analysis describes the B_2O_3 glass structure as consisting of boroxyl rings, i.e., planar rings containing three boron atoms and three oxygen atoms. Molecular dynamics have shown that the radial distribution pattern is consistent with structures having a low concentration of such rings [85].

Although borate glass forms a three-dimensional network, its viscosity is substantially lower than that of silicate glass, as shown in Table 4. Again, addition of alkali lowers the viscosity of the melt, but the effect is by no means as dramatic as for silicate glass [86].

Introduction of alkali or moisture to alkali-metal borate glasses causes some of the three-coordinate boron atoms to become four-coordinate, as shown by the pioneering NMR work of Bray and O'Keefe [87], [88]. The ratio of three- to four-coordinate boron in alkali-metal borate glasses has an upper limit of about 1.2.

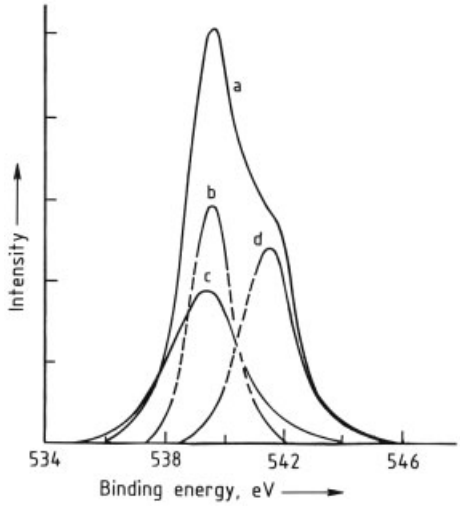

Figure 9. Oxygen 1 *s* photoelectron spectrum of 50 SrO · 50 P_2O_5 glass [88]
Solid lines indicate experimental data. Dashed lines indicate analytical data.
a) Total spectrum; b) Nonbridging oxygen; c) Doubly bonded oxygen (P = O); d) Bridging oxygen

Phosphate glasses contain tetrahedrally coordinated PO_4 building blocks; however, they do not show the same type of connectivity as silicate glasses. Oxygen 1 *s* photoelectron spectroscopy suggests that in phosphate glasses, three types of oxygen atoms occur: those bonded to two phosphorus atoms, the bridging oxygens O(br); those bonded to one phosphorus atom and one alkali-

metal atom, the nonbridging oxygens O(nbr); and those bonded only to phosphorus, the doubly bonded oxygens O(d). A characteristic oxygen 1 *s* photoelectron spectrum of a phosphate glass is shown in Figure 9. In contrast, silicate glasses never contain double-bonded oxygen atoms and only rarely nonbonded oxygen, i.e., oxygen atoms not connected to at least one silicon atom. An example of the latter is glassy Pb_2SiO_4.

Table 5. Oxygen 1 *s* binding energies and oxygen charges for crystalline and amorphous oxides (glasses)

Compound	Binding energy, eV	O charge
Crystalline oxides		
SiO_2 (vitreous silica, quartz, cristobalite)	532.4	−0.74
B_2O_3	532.3	−0.75
GeO_2	531.4	−0.78
Al_2O_3	531.0	−0.81
BeO	530.8	−0.82
ZnO	530.6	−0.83
Mg_2SiO_4	530.6	−0.83
Nb_2O_5	530.3	−0.85
SnO_2	530.2	−0.86
Cr_2O_3	530.2	−0.86
TiO_2	530.1	−0.87
As_2O_3	530.1	−0.87
ZrO_2	529.8	−0.88
CuO	529.7	−0.89
Fe_2O_3	529.7	−0.89
MgO	529.5	−0.90
PbO	529.4	−0.91
MnO	529.2	−0.93
NiO	529.1	−0.94
CaO	528.5	−0.96
$YBa_2Cu_3O_x$	530.2	−0.86
	528.2	−0.99
$Bi_2CaSr_2Cu_2O_{16-x}$	531.2	−0.79
	529.2	−0.93
Amorphous oxides (glasses)		
33 CuO – 17 Fe_2O_3 – 50 P_2O_5	532.05	−0.75
	531.55	−0.77
25 WO_3 – 75 TeO_2	530.90	−0.81
	530.65	−0.78
15 BaO – 35 ZnO – 50 TeO_2	530.30	−0.85
	529.65	−0.89
20 Na_2O – 80 SiO_2	531.90	−0.75
	530.50	−0.84
24 GeO_2 – 76 SiO_2	531.3	−0.79
34 PbO – 61 SiO_2 – 2 K_2O – 2 Na_2O – As_2O_3	532.6	−0.70
	530.4	−0.85

Oxygen 1 *s* photoelectron spectroscopy indicates both the concentration of different oxygen atoms in a material and the actual charges on these atoms. The relation between energy shift of the oxygen 1 *s* photoelectron spectral line and charge is shown in Figure 10. According to this method, quartz, cristobalite, and vitreous silica have an oxygen charge of −0.74. Oxygen atoms with higher negative charge are encountered, for example, in lead silicate glasses, which show oxygen charge variations comparable to those found for superconducting materials. Some oxygen 1 *s* derived charges for various compounds are compiled in Table 5 [90]. Assignment of charges is a partitioning problem without a unique solution.

Some special phosphate glasses show a connectivity of four, identical to that found in vitreous silica. Examples are ZnP_2O_6 and $AlPO_4$ glasses.

The structure of alkali phosphate glasses is well known because these glasses, in contrast to alkali silicate glasses, either do not, or only at very small rates, repolymerize on dissolution in aqueous solution. As a result, the length of the phosphate chains remains unchanged. The relative proportions of different chains in solution, analyzed chromatographically, give a true measure of the distribution of extended environments in phosphate glasses. The most common chain length is three to four phosphorus atoms, with a maximum chain length of seven. No analogies between the distribution of structural elements in phosphate and silicate glasses exist. This is due to the limited cross-linking and branching that occur in phosphate glasses in comparison to silicate glasses due to the presence of P=O bonds.

Phosphate glasses tend to have low durability. Despite this, important commercial applications exist. One of these is associated with the sharper absorption bands of iron oxide in the ultraviolet and infrared in phosphate glasses compared with silicate glasses. Iron-containing phosphate glasses are, therefore, nearly transparent to visible light, enabling the manufacture of virtually clear heat-absorbing glasses containing several percent iron oxide.

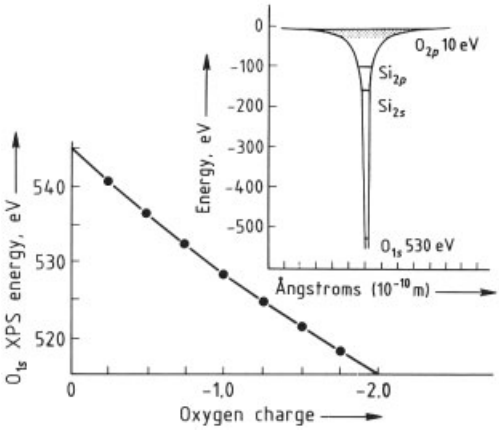

Figure 10. Oxygen charge derived from oxygen 1 *s* photoelectron spectra [89]

Phosphate-based glasses are more resistant than silicate glasses to hydrofluoric acid. Some optical glasses produced by Schott, Hoya, Owens-Illinois, and Corning France use phosphate as the primary glass former. Fluorophosphate glasses, designated FK-5 and FK-50 by Schott, have very low optical dispersion, with Abbe numbers of 70.4 and 81.5, respectively.

Germanate and Germanosilicate Glasses. The viscosity of molten GeO_2 (*mp* 1389 K) is 633 Pa · s at 1673 K. For comparison vitreous silica with < 10 ppm water has a viscosity of 8.17×10^8 Pa · s at the same temperature. Addition of alkali lowers the viscosity of molten germania substantially, as shown in Table 6. The GeO_2–SiO_2 glass system finds application in optical waveguides. The variation of viscosity of GeO_2–SiO_2 glass (42.2 mol % GeO_2) with temperature is shown below:

$10^{14.5}$ Pa · s (899 K)
10^{13} Pa · s (992 K)
10^{12} Pa · s (1054 K)

The initial increase in density of molten GeO_2 with the addition of alkali is attributed to changes in oxygen coordination around germanium from tetrahedral to octahedral. This coordination change is associated with the so-called germanate anomaly, which gives rise to observed maxima and minima in property – composition curves (e.g., molar volume, viscosity, acoustic velocity, and heat of solution).

Table 6. Viscosity and density of alkali germanate glasses

R_2O, mol %	Viscosity (1673 K), Pa · s	Density (1473 K), g/cm^3
0 (100 % GeO_2)	633.4	3.46
Li_2O		
1.39	55.0	
2.48	5.7	
5.0		3.57
8.02	0.63	
10.0		3.59
20.0		3.49
Na_2O		
1.58	11.0	
4.29	1.8	
5.0		3.61
7.79	0.51	
10.0		3.61
20.0		3.44
K_2O		
0.84	31.9	
4.58	1.4	
5.0		3.56
7.41	0.3	
10.0		3.49
20.0		3.18

2.2.4. Borosilicate and Aluminosilicate Glasses

Borosilicate Glasses. Network modifiers cause a change from trigonal to tetrahedral coordination in borosilicate glasses just as in borate glasses. Addition of boron to silica melts reduces the melt viscosity, as shown in Table 7; however, the effect is substantially less pronounced than that produced by addition of alkali to silica melts.

Table 7. Melt viscosity in the B_2O_3–SiO_2 system (1573 K)

B_2O_3, mol %	Viscosity, Pa · s
0 (100 % SiO_2)	1.622×10^{11}
42.4	752
62.4	78.6
71.9	25.4
82.5	10.2
90.0	5.8
93.9	4.4
100.0	3.0

Many technically important glasses have been developed on the basis of the *boron anomaly,* i.e., a minimum or maximum in certain physical properties of alkali borates at 15 – 30 mol % modifier. Replacement of alkali by boric acid decreases the thermal expansion of the

glass. The chemical resistance of these glasses is excellent. Phase separation of borosilicate glasses into silica- and borate-rich parts commonly occurs during heat treatment and is useful in certain applications. For example, Vycor glasses are based on liquid immiscibility in the $Na_2O-B_2O_3-SiO_2$ system. Most commercial borosilicate glasses, however, have compositions that are miscible and homogeneous. Borosilicate glasses are used in ovenware, laboratory equipment, piping, and sealed-beam headlights and in glass fibers for reinforcement, textile, or insulation fiber applications.

Aluminosilicate Glasses. The addition of small amounts of alumina (up to 4 mol %) to silica melts tends to increase the density of the resulting glass, presumably because of the formation of triclusters (i.e., oxygen atoms connected to one aluminum and two silicon atoms). The effect of large amounts of alumina on the viscosity of aluminosilicate melts at 2223 K is comparable to that of Li_2O at 1673 K (see Table 8).

Structural rules, based on X-ray diffraction analysis of crystalline network-modifier aluminosilicates, suggest that aluminosilicate glass has a silica-like structure in which all oxygen atoms are bridging when the molar ratio of alkali metal or alkaline earth to alumina is one. The high viscosity of alkali aluminosilicate melts is supposed to reflect this silica-like structure. For example, a melt of composition $Na_2O \cdot Al_2O_3 \cdot 6\,SiO_2$ (albite) has a viscosity of 230 Pa · s at 1973 K, 582 K above the melting point of the crystalline phase. On the other hand, a melt of composition $CaO \cdot Al_2O_3 \cdot 2\,SiO_2$ (anorthite) is fluid (0.96 Pa · s at 1973 K, 143 K above the melting point of the crystalline phase), indicating that silica-like behavior does not necessarily occur in alkaline-earth aluminosilicate melts. In contrast, however, the viscosity of fused silica at 1973 K is 2.48×10^6 Pa · s, so that neither of the two aluminosilicate melts seems to be particularly silica-like in its rheological properties.

Aluminosilicate glasses are used commercially because they can withstand elevated temperature and be chemically strengthened by alkali-metal or alkaline-earth exchange. Applications, therefore, include airplane and space vehicle windows, frangible containers, lamp envelopes, and cooking utensils.

Table 8. Viscosity of molten aluminosilicates (2223 K)

Al_2O_3, mol %	Viscosity, Pa · s
0	1.72×10^4
6.2	1.80×10^2
12.9	70.0
20.2	3.90
50.0	0.20
70.0	0.10
100	0.042 *

* 2327 K.

2.2.5. Lead Silicate Glasses

The $PbO-SiO_2$ system has the largest glass-forming region among oxide glasses. Characteristically, PbO also exhibits extended glass-forming regions with other network formers. The behavior of lead as a network modifier or former is thought to arise from its low coordination number with oxygen, which varies between 2 and 4. Note that although crystalline Pb_2SiO_4 has the orthosilicate composition commonly indicative of Q_0 silica species, its actual local silicon environment is Q_2, the silica tetrahedra in this compound forming Si_4O_{12} rings. In this compound, bridging (Si−O−Si), nonbridging (Si−O−Pb), and nonbonding (Pb−O−Pb) oxygen atoms are present. The last type is tetrahedrally coordinated to lead atoms, forming a quadricluster [91], [92].

Lead glasses can be melted easily; they have a long working range, low surface tension, and high refractive index. These properties make them useful for lead crystal, optical glass, and hand-formed art ware. Lead-containing glasses effectively absorb high-energy radiation and are, therefore, used for radiation windows, fluorescent lamp envelopes, and television tubes. Low-melting solder glasses and frit or decorative enamels are usually based on low-melting lead silicate compositions.

2.2.6. Chalcogenide Glasses

Glasses based on sulfur, selenium, or tellurium rather than oxygen are well known. Chalcogenide glasses are defined as noncrystalline solids from melts of the three chalcogens and

their combination with elements of groups 14, 15, and 17. Structurally such glasses are quite different; they resemble organic polymers rather than oxide glasses. These glasses, although often opaque to visible light, transmit further into the infrared than oxide glasses. Many of the chalcogenide glasses are semiconductors; the conductivity increasing with increasing atomic number. Chalcogenide glasses tend to be much more stable than fluoride glasses.

2.2.7. Halide Glasses

Spectacular developments have occurred in fluoride glass. Before about 1974, the only well-known fluoride glass was BeF_2, with tetrahedrally coordinated beryllium, a viscosity on the order of 10^5 Pa · s at its melting point (813 K), and a connectivity analogous to that observed for silica. Heavy-metal fluoride glasses containing zirconium, barium, and lanthanum (ZBL), with a characteristic viscosity at the melting point (813 K) of ca. 0.1 Pa · s, and the more stable ZBLA glasses (57 mol % ZrF_4, 34 mol % BaF_2, 5 mol % LaF_3, and 4 mol % AlF_3) have been discovered and examined [93]. Their use in optical fibers is discussed in [94]. They are commercially available.

Vitreous beryllium fluoride is a prime candidate in the production of optical waveguides because of its unique spectral properties, including transmission from < 160 to 5500 nm; very low refractive index, $n_D = 1.275$; and very low dispersion, $\nu^{-1} = 9.3 \times 10^{-3}$.

Drawbacks are that it is highly hygroscopic and contains beryllium, which is toxic. The former can be reduced by the addition of aluminum, alkali-metal or alkaline-earth fluorides.

Some difficulties must be overcome in the application of heavy-metal fluorides in waveguides: two of the more serious are poor mechanical strength and scattering loss due to transition-metal ion impurities and crystalline inclusions.

2.2.8. Glasses with Complex Composition

Most industrially important glasses contain a large number of components. The composition of these glasses is given in Table 9. The simple systems already mentioned give some guide about what to expect of the properties and structure in such systems. However, attempts are still being made to relate properties to melt or glass structure even for simple systems. The structure – property relationship in complex systems is still largely obscure, and substantial scope remains in glass production for classical heat and beat treatments.

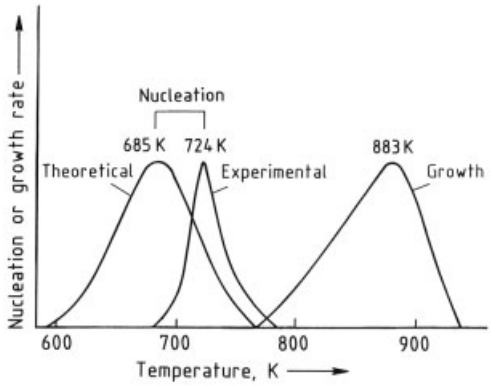

Figure 11. Observed and calculated nucleation and growth regions as a function of temperature for $Li_2O \cdot 2\,SiO_2$ glass

3. Glass Formation, Crystallization, and Liquid Immiscibility

Glasses are thermodynamically unstable, compared with isochemical crystals. The reason for glass formation can be inferred from the work of Tammann [95], among others, who showed that the crystallization process occurs in two steps: first crystal nuclei form, then these nuclei grow. Depending on the temperature and the substance, either step may determine the rate of spontaneous crystallization. Therefore, at least two free energy barriers are involved in the metastability of the supercooled liquid. The first of these is the free energy of crystal nucleus formation, arising primarily because the melting point of small crystals is lower than that of large crystals. The second barrier is due to impeded molecular motion at the crystal – liquid interface in going from a liquid to a crystal-like position. The two noncoinciding regions of nucleation and growth are shown for the Li_2O – $2\,SiO_2$ system in Figure 11. This figure also

Table 9. Composition of some industrially important glasses (wt %)

Use	Glass code	SiO_2	Al_2O_3	B_2O_3	Li_2O	Na_2O	K_2O	MgO	CaO	BaO	PbO	ZnO	Others
Lamp bulbs, sealing	0080	73	1			17		4	5				
Lamp tubing	0120	56	2			4	9				29		
Color TV neck	0137	52.5	1			0.5	13				28		SrO : 5
Color TV funnel	0138	54	2			6	8	2	3		23		
Lead glass for Steuben	0160	56	1		1	8	4				31		Sb_2O_3 : 1, As_2O_3 : 0.2
Microsheet	0211	65	2	9		7	7					7	TiO_2 : 3
Consumer wave	0281	74	1	1		14			9				Sb_2O_3 : 0.3, As_2O_3 : 0.1
Aircraft windows	0317	61	17			13	3	3	0.4				TiO_2 : 1, As_2O_3 : 1
Tape reel	0320	63	12	13	5	6							Cl : 0.4
Aircraft windows	0331	66	21		4	9	0.2	1					
Ferrite seal	1415	33	2	10					1	48		2	ZrO_2 : 5, Sb_2O_5 : 0.2, As_2O_3 : 0.1
Electron tubes	1723	57	16	4				7	10	6			
T/H lamp	1724	57	17	5				6	8	8			
Heat ABS	4605	23	14			1						4	FeO : 2, P_2O_5 : 57
Table ware	6720	60	10	1		8	2		5			10	F^- : 4
Series sealing	7050	68	2	24		6							
Kovar sealing	7052	64	8	19	1	2	3			3			F^-
Kovar sealing	7056	68	3	18	1	1	9						
Substrate	7059	49	10	15						25			As_2O_3 : 1
Low-loss electrical	7070	72	1	25	0.5	0.5	1						
Auto sealed beam lamps and bakeware, Pyrex	7251	82	2	12		4							
Solder sealing	7570	3	11	12							74		
Tungsten sealing	7720	74	1	15		4					6		
Dental restorations	7724	51	8	9						32			
Laboratory ware, Kimax, Pyrex	7740	81	2	13		4							
General	7760	78	2	15		3	1						As_2O_3 : 1
Pharmaceutical	7800	72	6	11		7	1		1	2			
Vycor high temperature	7913	96.5	0.5	3									
Optical	7940	99.9											
U.C.E. optical	7971	93											H_2O : 0.1

Table 9. (Continued)

Use	Glass code	SiO_2	Al_2O_3	B_2O_3	Li_2O	Na_2O	K_2O	MgO	CaO	BaO	PbO	ZnO	Others
Ophthalmics	8078	44				6			1	21	24	4	La_2O_3 : 2, TiO_2 : 3.0, ZrO_2 : 5, As_2O_3 : 0.1, Sb_2O_3 : 0.1
Ophthalmics	8092	62	3			8	9	3				12	TiO_2 : 0.7, Sb_2O_3 : 0.4, SO_3 : 0.2, As_2O_3 : 0.2, B_2O_3 : 1.0
Electron tubes, sealing	8160	56	2			3	10		1	5	23		F^- : 0.7
Electron tubes, sealing	8161	39					6			2	51		Rb_2O : 2
Ophthalmics	8361	68	2			8	9		4			4	TiO_2 : 0.4, Sb_2O_3 : 0.5, As_2O_3 : 0.10
CRT panel	9025	68	4	12	3	5	6						TiO_2 : 1, CeO_2 : 1
Color TV panel	9068	65	2			7	9	1	2	2	2		SrO : 10, F^-, TiO_2, CeO_2, As_2O_3, Sb_2O_3
UV-transmitting	9741	65	5	27	1	2							F^-
Optical	N16B	5		34					2			6	ZrO_2 : 6, La_2O_3 : 41, CdO : 6
Laboratory ware	G20	76	5	7		6	1		1	4			
Reinforcement fiber	E-Glass	54	14	<10		<0.5	<0.5	4.5	20				TiO_2
Insulation fiber	T-Glass	59	4.5	3.5		11	0.5	5.5	16				
Chemical-resistant fiber	C-Glass	65	4	5.5		8	0.5	3	14				
Insulation fiber	SF-Glass	59.5	5	7		14.5							ZrO_2 : 4, TiO_2 : 8, F^-
High strength	S-Glass	65	25					10					
Container-Glass *		71.9	1.7			12 – 13.9	0.4	3.6	7.9	0.3			SO_2 : 0.2, F_2 : 0.2, Fe_2O_3 : 0.04
Window sheet *		71 – 73	0.5 – 1.5			12 – 15		1.5 – 3.5	8 – 10				
Plate glass *		71 – 73	0.5 – 1.5			12 – 14		1 – 4	10 – 12				

* Composition changes depending on regional variations in the USA: these compositions are characteristic for the North East.

illustrates the discrepancy between the nucleation area derived from classical nucleation theory and that observed experimentally [96], [97].

The rate of crystallization of an undercooled liquid is specified by the rate of both crystal nucleation and crystal growth at the expense of the liquid. With glass, crystallization is called *devitrification,* which occurs when the glass is heated above the glass transition temperature T_g but below the liquidus temperature. Glass transition temperatures are given in Table 10.

Table 10. Glass transition temperatures and melting points for some silicate glasses [80]

Glass	T_g, K	T_m, K
SiO_2	1381	1996
$Na_2Si_3O_7$	790	1073
$Na_2Si_2O_5$	725	1147
$K_2Si_4O_9$	775	1043
$K_2Si_3O_7$	820	1198
$K_2Si_2O_5$	763	1318
$CaMgSi_2O_6$	1026	1665
$NaAlSi_3O_8$	1088	1391
	1028	1391
	1036	1391
$KAlSi_3O_8$	1178	1806
$CaAl_2Si_2O_8$	1086	1826
	1153	1826

In undercooled liquids, nucleation at interfaces (i.e., heterogeneous nucleation) dominates other nucleation processes. Homogeneous nucleation (i.e., nucleation in the interior of the liquid in the apparent absence of an interface) is known for only a few glass systems (e.g., $Li_2O \cdot 2\,SiO_2$, $Na_2O \cdot 2\,CaO \cdot 3\,SiO_2$, and $BaO \cdot 2\,SiO_2$ [98]). Locations in which heterogeneous nucleation occurs are at the air – liquid interface, the liquid container wall, particles or gas bubbles, and boundaries between immiscible regions. A well-known example of the predominance of heterogeneous nucleation is the formation of ice, which nucleates heterogeneously at 273 K and homogeneously in undercooled water at ca. 233 K.

Crystallization in undercooled liquids is controlled primarily by the rate of crystal growth. Nucleation studies by Tamman [95] of silicate melts indicated that the rate of crystal growth increases with increasing undercooling, reaches a maximum, and then decreases with further increase in undercooling. Theoretical studies in the early 1930s showed that the growth rate increases with increasing undercooling, but decreases with increasing viscosity, which accounts for the experimentally observed maximum rate of crystal growth at some intermediate temperature.

A measure of glass-forming tendency is indicated by the reciprocal of the growth rate. Thus, peak growth rates of 2×10^{-6}, 1×10^{-5}, and 1.5×10^{-2} cm/s for cristobalite (SiO_2), GeO_2, and anorthite ($CaAl_2Si_2O_8$), respectively, indicate the first glass to be substantially more stable than the last. Some additional examples are given in Table 11.

As a rule, technical glasses have peak growth rates below 10^{-3} cm/s. Some optical glasses have peak growth rates greater than 10^{-2} cm/s, thereby causing considerable production problems.

Growth rates are generally extremely dependent on composition. Alkali – silica systems are illustrated in Figure 12. Even minor components can substantially influence growth rate and, hence, the ability to form a glass. The role of minor components is usually thought to be one of poisoning the surface of the crystallite.

Nucleation and growth are diffusion-controlled. Glasses, therefore, become more stable with increasing number of components. Consideration of the Stokes – Einstein relation, which describes the inverse proportionality between diffusion and viscosity, suggests that lower diffusion and, therefore, higher viscosity at the melting temperature enhance the glass-forming tendency. Increasing the concentration of network formers, therefore, tends to enhance melt viscosity and decrease crystal growth for the same thermodynamic driving force of crystallization.

In addition to crystalline precipitation, phase separation can occur in specific glass systems. Such separation is usually manifested as changes in color, optical transparency, or chemical durability. Glass systems exhibiting phase separation possess, in a specific temperature region, a miscibility gap characterized at the upper and lower ends by an immiscibility temperature. Liquid immiscibility can be made visible by proper heat treatment, especially in glasses that tend to undergo phase separation, e.g., those found in the silica-rich part of the $Li_2O - 2\,SiO_2$ and alkaline earth – silica systems. The growth rate of the secondary glassy phase, separated from the orig-

Table 11. Measured growth rate of minerals and related substances [80]

Crystal	Composition	Liquidus temperature, K	ΔT at peak, K	Rate at peak, cm/s
Cristobalite	SiO_2	1996	50	2×10^{-6}
Anorthite	$CaAl_2Si_2O_8$	1830	300	1.5×10^{-2}
Diopside	$CaMg(SiO_3)_2$	1664	>100	2.2×10^{-2}
$Na_2Si_2O_5$		1147	60	1×10^{-3}
GeO_2		1389	100	1×10^{-5}
SrB_4O_7		1270	100	1.6×10^{-2}
PbB_4O_7		1048	120	2×10^{-4}
Alkali feldspar	$Ab_{70}Or_{30}\cdot$ 1.7 % H_2O	1243	160	3.2×10^{-6}

inal homogeneous glass, is diffusion-controlled in most cases.

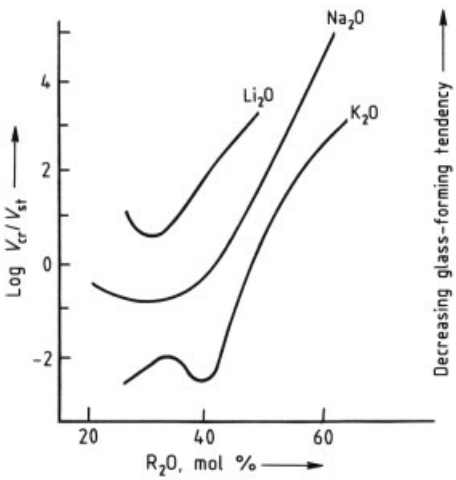

Figure 12. Critical cooling rate as a function of composition for alkali – silica melts (R_2O-SiO_2) [99]

For example, opal glasses may be either translucent or opaque, depending on the particle size and quantity of the second phase and on the difference in refractive index between the second phase and the matrix. Fluorine-containing glasses, which separate into phases containing NaF or CaF_2, are commonly made into pressed items such as tableware. On cooling, such glasses become spontaneously opaque, phase-separating into a silicate-rich matrix and fluoride droplets with dimensions of the order of 10^{-6} m. Other products in which phase separation is used are Vycor glass, photochromic glass, and luminescent and magnetic glasses.

4. Chemical, Mechanical, and Physical Properties of Industrially Important Melts and Glasses

Extensive research has been carried out on the relation between composition and physical properties of multicomponent glass systems. Partial molar coefficients for most physical properties (e.g., expansion, viscosity, heat capacity, and refractive index) have been determined since the pioneering work in this area by WINCKELMANN and SCHOTT [100], enabling calculation of properties for a given composition. Some properties of industrially important glasses are collected in Table 12, 13, 14, 15.

The lack of proper probes to analyze glass structure has prevented derivation of quantitative relationships between glass structure and properties. Qualitative rationalizations in terms of one model or another abound, and with such models, predictions can be made as to the influence of various components on the physical and structural nature of the glass.

The most important property of a glass or melt is its viscosity. Therefore this discussion starts with the rheological properties of glass melts and the three major variables which influence these properties: temperature, time, and glass composition. Next, surface tension of melts is discussed, followed by glass density, thermal expansion, mechanical properties such as elasticity and mechanical strength, hardness, thermal properties, electrical properties, gas permeability, chemical stability, surface properties and optical properties. This chapter concludes with a discussion of the nature of color in glasses.

Table 12. Properties of some industrially important glasses

Type of glass	Glass code [a]	Color	Forms usually available [b]	Corrosion resistance			
				Class	Weathering	Water	Acid
Soda lime	0080	clear	BMT	I	3	2	2
Potash soda lead	0120	clear	TM	I	2	2	2
Alkaline earth lead	0137	clear	T				
Potash soda lead	0138	clear	P	II	1		3
Lead crystal for Steuben	0160	clear					
Potash soda zinc	0211	clear	S	I	1		2
Sculptured ware	0281				3		2
Aircraft windows	0317	clear	S		1		3
Tape reel	0320						
Pipets and centrifuge tubes	0331		T				
Ferrite sealing glass	1415		Powder disks cane				
Aluminosilicate	1723	clear	BT	I	1	1	3
Halogen cycle lamps	1724	clear	T				2
Heat-absorbing, visible-transmitting	4605		gob strip pressing				
Soda zinc	6720	opal	P	I	[d]	1	2
Borosilicate	7050	clear	T	II	3 [c]	3 [c]	4 [c]
Borosilicate	7052	clear	BMPT	II	2 [c]	2 [c]	4 [c]
Borosilicate	7056	clear	BTP	II	2	2	2
Barium—bariumsilicate	7059	clear	S	II	1		4
Borosilicate	7070	clear	BMPT	I	2 [c]	2 [c]	2 [c]
Borosilicate	7251	clear	P	I	1 [c]	2 [c]	2 [c]
High lead	7570	clear		II	3	3	4
Borosilicate	7720	clear	BPT	I	1	2 [c]	2 [c]
Frit for dental restorations	7724	white	ground and screened cullet		2		4
Borosilicate	7740	clear	BPSTU	I	1 [c]	1 [c]	1 [c]
Borosilicate	7760	clear	BP	I	2	2	2
Soda barium borosilicate	7800	clear	T		1	1	1
96 % Silica	7913	clear	BPRST	I	1	1	1
Fused silica	7940	clear	U	I	1	1	1
Titanium silicate	7971	clear	U		1	1	1
Ophthalmic lens blanks	8078						
Ophthalmic lens blanks	8092						
Potash soda lead	8160	clear	PT	II	2	2	3
Potash lead	8161	clear	PT	I	2	1	4
Ophthalmic lens blanks	8361				2		2
Alkali—barium	9025	clear	P	II	3		2
Alkali—strontium	9068	clear	P	I	2		2
Borosilicate	9741	clear	BUT	II	3 [c]	3 [c]	4 [c]
Borosilicate	N16B	clear					
Aluminoborosilicate	G20	clear	BPT				
Lime—aluminoborosilicate	E-glass	clear	F				
Soda lime—aluminoborosilicate	T-glass	clear	F				
Soda lime—borosilicate	C-glass	clear	F				
Soda—titania—zirconia	SF-glass	clear	F				
Magnesium—aluminosilicate	S-glass	clear	F				

[a] Glasses 7905, 7910, 7911, 7912, 7913, and 7917 for special UV and IR applications. Glass 1720 is available with improved UV transmittance (designated glass 9730). Glass 7760 is also available with special transmission suitable for sun lamps. [b] B = blown ware; M = multiform; U = panels; P = pressed ware; R = rolled sheet; C = castings; S = plate glass; T = tubing and rod; F = fiber. [c] These borosilicate glasses may rate differently if subjected to excessive heat treatment. [d] Since weathering is determined primarily by clouding which changes transmission, a rating for the opal glasses is omitted.

Table 13. Properties of some industrially important glasses (continued)

Type of glass	Thermal expansion, $\times 10^{-7}$ °C [e]		Upper application temperatures [g] (mechanical consideration only)				Thermal shock resistance [h], plates 15×15 cm, annealed		
	0–300 °C	25 °C to setting point	Annealed		Tempered		3.2 mm thick, °C	6.4 mm thick, °C	12.7 mm thick, °C
			Normal service, °C	Extreme service, °C	Normal service, °C	Extreme service, °C			
Soda lime	93.5	105	110	460	220	250	65	50	35
Potash soda lead	89.5	97	110	380			65	50	35
Alkaline earth lead	97.0	102							
Potash soda lead	98.5	109							
Lead crystal for Steuben	93								
Potash soda zinc	74.0	84.0							
Sculptured ware	87								
Aircraft windows	87	98							
Tape reel	70.7								
Pipets and centrifuge tubes	75.5								
Ferrite sealing glass	79	96.7							
Aluminosilicate	46	54	200	650	400	450	125	100	70
Halogen cycle lamps	43	50							
Heat-absorbing, visible-transmitting	54								
Soda zinc	78.5	90	110	480	220	275	70	60	40
Borosilicate	46	51	200	440	235	235	125	100	70
Borosilicate	46	53	200	420	210	210	125	100	70
Borosilicate	51.5	56	200	460					
Barium–bariumsilicate	46	50.5							
Borosilicate	32	39	230	430	230	230	180	150	100
Borosilicate	36.5	41.4	230	460	260	260	160	130	90
High lead	84	92	100	300					
Borosilicate	36	43	230	460	260	260	160	130	90
Frit for dental restorations	52.6								
Borosilicate	32.5	35	230	490	260	290	160	130	90
Borosilicate	34	37	230	450	250	250	160	130	90
Soda barium borosilicate	50	53	200	460					
96 % Silica	7.5	5.5 [f]	900	1200					
Fused silica	5.6	3.5 [f]	900	1100					
Titanium silicate	0.5	−2	800	1100					
Ophthalmic lens blanks	91.4								
Ophthalmic lens blanks	94								
Potash soda lead	91	100	100	380			65	50	35
Potash lead	90	99	100	390					
Ophthalmic lens blanks	94								
Alkali–barium	90.0	98.0							
Alkali–strontium	99.0	111							
Borosilicate	39.5	50	200	390			150	120	80
Borosilicate	87								
Aluminoborosilicate									
Lime–aluminoborosilicate	60								
Soda lime–aluminoborosilicate	80								
Soda lime–borosilicate	72								
Soda–titania–zirconia	75								
Magnesium–aluminosilicate	34								

[e] Code 9608 may be produced in a range of expansion values depending upon intended application. [f] Extrapolated values. [g] Normal service: no breakage from excessive thermal shock is assumed. Extreme limits; glass will be very vulnerable to thermal shock. Recommendations in this range are based on mechanical stability considerations only. Tests should be made before adopting final designs. These data approximate only. [h] These data approximate only. Based on plunging cold sample into cold water after oven heating. Resistance of 100 °C means no breakage if heated to 110 °C and plunged into water at 10 °C. Tempered samples have over twice the resistance of annealed glass.

Table 14. Properties of some industrially important glasses (continued)

Type of glass	Thermal stress resistance, °C [i]	Viscosity data [j] Strain point, °C	Annealing point, °C	Softening point, °C	Working point, °C	Knoop hardness, HK_{200} [k]	Density, g/cm^3	Young's modulus, GPa	Poisson's ratio
Soda lime	16	473	514	696	1005	465	2.47	70	0.22
Potash soda lead	20	395	435	630	985	382	3.05	59	0.22
Alkaline earth lead		436	478	661	977		3.18		
Potash soda lead		450	490	670			3.02		
Lead crystal for Steuben		367	405	583			3.09		
Potash soda zinc	20	508	550	720	1008		2.57	74	0.21
Sculptured ware		500	540	719			2.48		0.22
Aircraft windows		574	624	871		440	2.45	73	0.22
Tape reel		463	493	638			2.38		
Pipets and centrifuge tubes		510	548				2.41		
Ferrite sealing glass		620	654	776			3.74		
Aluminosilicate	26	665	710	908	1168	514	2.64	86	0.24
Halogen cycle lamps		613	722	926	1200	572	2.64	85	0.24
Heat-absorbing, visible-transmitting		560	612	806			2.52		
Soda zinc	20	505	540	780	1023		2.58	70	0.21
Borosilicate	39	461	501	703	1027		2.24	60	0.22
Borosilicate	41	436	480	712	1128	375	2.27	57	0.22
Borosilicate	33	472	512	718	1058		2.29	64	0.21
Barium—bariumsilicate	32	593	639	844	1160		2.76	68	0.28
Borosilicate	66	456	496		1068		2.13	51	0.22
Borosilicate	48	521	565	808	1192	451	2.26	64	0.19
High lead	21	342	363	440	558		5.42	55	0.28
Borosilicate	49	484	523	755	1146		2.35	63	0.20
Frit for dental restoration		633	671	847			2.96		
Borosilicate	54	510	560	821	1252	418	2.23	63	0.20
Borosilicate	52	478	523	780	1198	442	2.24	62	0.20
Soda barium borosilicate	33	533	576	795	1189		2.36		
96 % Silica	220	890	1020	1530		487	2.18	68	0.19
Fused silica	286	956	1084	1580		489	2.20	73	0.16
Titanium silicate	3370		1000	1500			2.21	68	0.17
Ophthalmic lens blanks		506	544	686			4.02		
Ophthalmic lens blanks		500	540	730			2.62		
Potash soda lead	18	397	438	632	973		2.98		
Potash lead	22	400	435	600	862		3.99	54	0.24
Ophthalmic lens blanks		500	543	726	1032		2.54		0.23
Alkali—barium		421	460	647	988		2.60		
Alkali—strontium		462	503	688			2.68		
Borosilicate	54	408	450	705	1161		2.16	49	0.23
Borosilicate	19		540	720	1065		2.48		
Aluminoborosilicate		524	569	794	1190		2.39	72	
Lime-aluminoborosilicate		507	657	846			2.60	74	
Soda lime—aluminoborosilicate				715			2.54		
Soda lime—borosilicate		435	585	752			2.61	70	
Soda—titania zirconia				675			2.57		
Magnesium—alumino silicate		760	810	970				87	

[i] Resistance in °C is the temperature differential between the two surfaces of a tube or a constrained plate that will cause a tensile stress of 6.9 MPa (1000 psi) on the cooler surface. [j] These data subject to normal manufacturing variations. [k] Determined by revised ASTM standard: number of standard not yet assigned.

Table 15. Properties of some industrially important glasses (continued)

Type of glass	Log_{10} of volume resistivity, $\Omega \cdot cm$			Dielectric properties at 1 MHz, 20 °C			
	25 °C	250 °C	350 °C	Power factor, %	Dielectric constant	Loss factor, %	Refractive index [m]
Soda lime	12.4	6.4	5.1	0.9	7.2	6.5	1.512
Potash soda lead	>17	10.1	8.0	0.12	6.7	0.8	1.560
Alkaline earth lead		10.1	8.4				1.570
Potash soda lead			7.7				1.563
Lead crystal for Steuben							1.569
Potash soda zinc							1.52
Sculptured ware		6.5	5.2				1.515
Aircraft windows		5.5	4.3		8.94		1.506
Tape reel							
Pipets and centrifuge tubes							
Ferrite sealing glass							
Aluminosilicate	>17	13.5	11.3	0.16	6.3	1.0	1.547
Halogen cycle lamps		13.8	11.8				1.54
Heat-absorbing, visible-transmitting							1.51
Soda zinc							1.507
Borosilicate	16	8.4	6.8	0.33	4.9	1.6	1.479
Borosilicate	17	9.2	7.4	0.15	5.1	1.3	1.484
Borosilicate		10.2	8.3	0.27	5.7	1.5	1.487
Barium—bariumsilicate		13.1	11.0		5.9		1.53
Borosilicate	>17	11.2	9.1	0.06	4.1	0.25	1.469
Borosilicate	18	8.0	6.5	0.45	4.9	2.18	1.476
High lead	>17	10.6	8.7	0.22	15	3.3	1.86
Borosilicate	16	8.9	7.3	0.23	4.6	1.3	1.487
Frit for dental restoration							1.523
Borosilicate	15	8.1	6.6	0.39	4.6	2.6	1.474
Borosilicate	17	9.4	7.7	0.16	4.6	0.79	1.473
Soda barium borosilicate		7.0	5.7				1.491
96 % Silica	>17	9.7	8.1	0.04	3.8	0.15	1.458
Fused silica	>17	12.4	10.7	0.001	3.8	0.0038	1.459
Titanium silicate	20.3	12.2	10.1	<0.002 [1]	4.0 [1]	<0.008 [1]	1.484
Ophthalmic lens blanks							1.70
Ophthalmic lens blanks							1.523
Potash soda lead	>17	10.6	8.4	0.09	7.0	0.63	1.553
Potash lead	>17	12.0	9.9	0.06	8.3	0.50	1.659
Ophthalmic lens blanks		8.7	6.9				1.523
Alkali—barium		8.5	6.8		6.5		1.523
Alkali—strontium		9.3	7.4				1.529
Borosilicate	>17	9.4	7.6	0.32	4.7	1.5	1.468
Borosilicate		5.8	4.7		7.7	1.4	
Aluminoborosilicate		7.4	6.0	0.51	4.9	2	
Lime-aluminoborosilicate					6.4		1.548
Soda lime—aluminoborosilicate					7.3		1.541
Soda lime—borosilicate					7.8		1.549
Soda—titania zirconia					8.3		1.537
Magnesium—alumino silicate							1.523

[1] At 10 kHz. [m] Refractive index may be either the sodium yellow line (589.3 nm) or the helium yellow line (587.6 nm). Values at these wavelengths do not vary in the first three places beyond the decimal point.

4.1. Rheological Properties of Glass Melts

Viscosity determines the optimum melting, forming, and annealing procedures for a glass, as well as the limitations on its use at high temperature. The optimal viscosity regions for these processes are illustrated in Figure 13. A characteristic of glass is that rapid but smooth changes in viscosity occur in the entire region from 10 to 10^{13} Pa · s, provided that no devitrification or crystallization occurs. Molten silicates with vis-

cosity up to 10^3 Pa · s are Newtonian liquids, i.e., liquids in which viscosity is independent of shear stress.

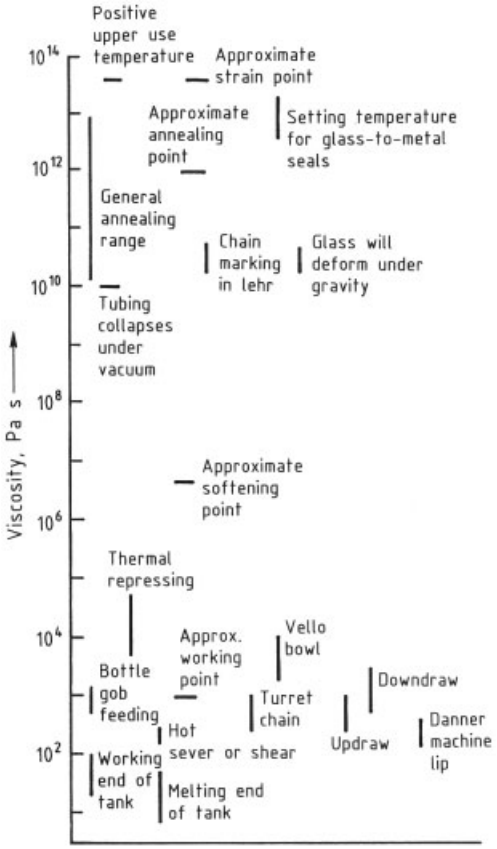

Figure 13. Optimum viscosity regions for forming and annealing glasses

Four types of viscosity measurement exist: (1) stationary methods, which include capillary and falling body; (2) rotation methods, including cone-and-plate and rotating cylinders; (3) oscillating methods; and (4) bending beam methods. No single instrument can measure viscosity over the entire range of interest for glasses. Below 10^6 Pa · s a cone-and-plate or rotation viscometer is used. In the range $10^6 - 10^{11}$ Pa · s, a wire pull method is used. The parallel-plate method covers the $10^4 - 10^8$ Pa · s range. The bending beam method is used for viscosities of $10^8 - 10^{14}$ Pa · s. A synopsis of methods for viscosity measurement is given in [101] and [102].

4.1.1. Temperature Dependence of Viscosity

The temperature dependence of viscosity is shown for some oxide glasses in Figure 14. The working, softening, annealing, and strain points are indicated as reference temperatures. Definition of these points are given by the ASTM.

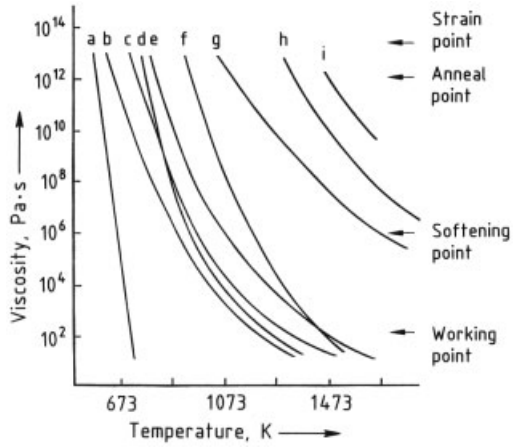

Figure 14. Temperature dependence of viscosity for some industrial glasses
a) Glass no. 8363; b) Glass no. 0010; c) Glass no. 7070; d) Glass no. 0080; e) Glass no. 7740; f) Glass no. 1720; g) Glass no. 7900; h) Glass no. 7940, fused silica; i) Nitrided 96 % SiO_2

The *working point* of a glass melt is the temperature at which its viscosity is exactly 10^3 Pa · s; the glass is then sufficiently fluid for forming by most methods. The *working range* is the range of surface temperature in which glass is formed into ware in a specific process. The *upper end* is the temperature at which glass is ready for working and corresponds generally to a viscosity of $10^2 - 10^3$ Pa · s. The *lower end* refers to the temperature at which glass is sufficiently viscous to hold its formed shape, corresponding to a viscosity $\geq 10^3$ Pa · s.

The *softening point* is defined, according to ASTM C 338 as the temperature at which a uniform fiber of glass with a diameter of 0.55 – 0.75 mm and a length of 23.5 cm increases its length by 1 mm/min when the upper 10 cm is heated at a rate of 5 °C/min. The viscosity at the softening point depends on the density and surface tension of the glass, and is $10^{6.6}$ Pa · s for a glass with a density of 2.5 g/cm^3 and a surface tension of 0.3 N/m.

The *annealing point* is the temperature corresponding to a viscosity of approximately 10^{12} Pa · s. At this point, internal stress is released to an acceptable value of 1.7 MPa (250 psi) in about 15 min.

Finally, the *strain point* corresponds to the temperature at which internal stresses are reduced to an acceptable level of 1.7 MPa in about 4 h. The viscosity at this temperature is approximately $10^{13.5}$ Pa · s. This point represents the temperature of transition from viscoelastic to brittle behavior. The *glass transition temperature* T_g lies in this viscosity range. The temperature region between the annealing and the strain points must be crossed with welldefined slow cooling rates if strain-free large glass bodies are to be produced. Both annealing and strain points may be determined by fiber elongation, ASTM C 336-71 (1981), or by beam bending, ASTM C 598-72 (1978).

In the parlance of glass science, glass with a high softening point is known as *hard glass*; that with a low softening point, as *soft*. If the temperature interval corresponding to 10^2 – 10^6 Pa · s is large, the glass is called *sweet* or *long* in contrast to *short* glasses with steep viscosity – temperature curves. Long glasses, therefore, solidify more slowly than short glasses on cooling. Combinations of these terms are also used; e.g., fused silica, Corning Code 7940, would be identified as a hard, long glass.

Viscosity decreases with increasing temperature; for liquid metals and simple ionic melts, the temperature dependence of the viscosity can be represented by the exponential equation

$$\eta = \eta_0 \exp\left(E_\eta / RT\right)$$

where η is the viscosity and E_η the activation energy for viscous flow. This Arrhenius-type relation rarely holds for polymeric melts because of depolymerization with increasing temperature. Such depolymerization results in a decrease in the activation energy for viscous flow.

For silicate melts, the temperature dependence of viscosity is adequately represented by the empirical Vogel – Fulcher – Tammann equation

$$\log\eta = A + \frac{B}{(T - T_0)}$$

in which A and B are constants, B being a measure of the activation energy for viscous flow; T_0 is added as a third adjustable constant. A Fulcher and two Arrhenius-type equations for three network-forming pure liquid oxides are given below:

For B_2O_3:

$$\log\left(\eta, \mathrm{Pa{\cdot}s}\right) = 0.198 + 145.1/\left(T - 443\right)$$

For GeO_2(1773 – 1973 K):

$$\log\left(\eta, \mathrm{Pa{\cdot}s}\right) = \frac{1.423 \times 10^3}{T} - 0.4704$$

For SiO_2 glass (1508 – 1608 K) and SiO_2 melt (1973 – 2273 K):

$$\log\left(\eta, \mathrm{Pa{\cdot}s}\right) = \frac{2.7115 \times 10^3}{T} - 0.5884$$

The presence of suspended solid matter or gas bubbles in the melt seriously affects measurement of viscosity. The effect is usually to increase the apparent viscosity. In addition, non-Newtonian behavior is commonly observed in crystal – melt slushes or melted gas bubble foams.

4.1.2. Time Dependence of Viscosity

In some highly viscous melts a dependence of viscosity on previous treatment of the melt and on time has been observed, particularly in the subliquidus area. The reason is that highly viscous undercooled liquids do not relax rapidly enough for the structural elements in the liquid to rearrange themselves in a state commensurate with the ambient temperature.

4.1.3. Composition Dependence of Viscosity

As shown previously, the addition of network modifiers to liquid silicon dioxide, boron trioxide, or germanium dioxide strongly affects melt viscosity. Also, among network modifiers, those with the highest cationic field strength are most effective in reducing viscosity.

Although the effect of network modifiers such as alkali-metal and alkaline-earth atoms is fairly well understood, the same cannot be said for amphoteric oxides (e.g., Al_2O_3, ZnO, or MgO). Their varying role is ascribed to changes

in coordination in melts with strong bases, giving rise to terms such as the germanium, aluminum, or boron anomaly.

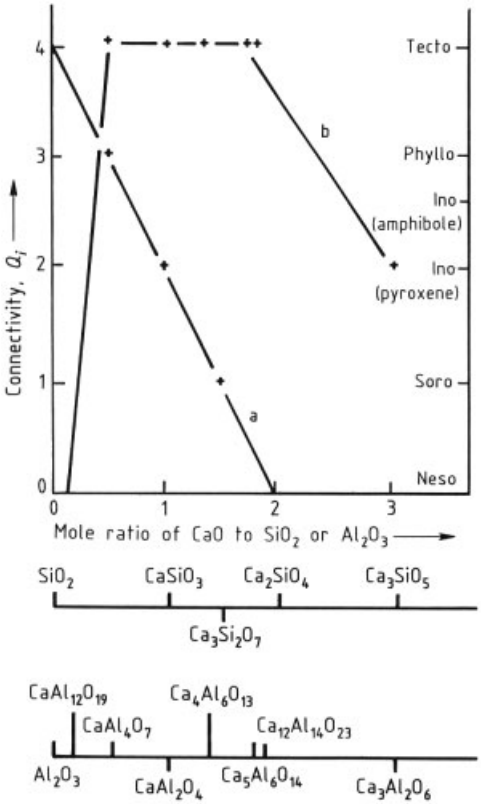

Figure 15. Variation in connectivity of crystalline calcium silicates and aluminates [103]
a) $CaO-SiO_2$ system; b) $CaO-Al_2O_3$ system

Consider the $CaO-Al_2O_3-SiO_2$ system as an illustration of the difficulty in deriving viscosity for ternary systems. The known crystal structures in the two binary systems $CaO-Al_2O_3$ and $CaO-SiO_2$ show that addition of calcium oxide to silica decreases the connectivity of the silica network in a regular fashion from Q_4 to Q_0; however, addition of calcium oxide to alumina results initially in a coordination change for aluminum from six- to fourfold, followed by the formation of network aluminates with Q_4 connectivity. These trends are illustrated in Figure 15. The two tendencies, depolymerization of silica and polymerization of alumina, must be taken into account if the viscosity of a $CaO \cdot 2\,SiO_2 \cdot Al_2O_3$ melt is to be predicted. Actual measurements indicate that the viscosity of a $CaO \cdot 2\,SiO_2$ melt at 1973 K is 0.78 Pa · s and for $CaO \cdot Al_2O_3$ 0.11 Pa · s, whereas a melt with composition $CaO \cdot 2\,SiO_2 \cdot Al_2O_3$ has a viscosity of 0.96 Pa · s at the same temperature, which suggests that the $CaO \cdot 2\,SiO_2$ system is viscosity-determining in this ternary system.

The viscosity of the $CaO-Al_2O_3-SiO_2$ system is still amenable to some rationalization. The melting behavior in $CaO-SiO_2$– metal oxide systems is more puzzling. As an example, the variation in the three-phase liquidus boundary between crystalline $CaSiO_3$– crystalline silica – liquid, as a function of a third metal oxide, is illustrated in Figure 16. The liquidus boundary suggests that Cr_2O_3 associates with CaO, which extends the silica field with an attendant increase in viscosity, in contrast to BaO, which decreases the silica field most. One might expect concomitant viscosity variations due to this extension or contraction of the silica field in the $CaO-SiO_2$ field. No compatible data are available to assess whether this is indeed the case. Preliminary ^{29}Si MAS NMR work has not shown variations in the melt structure among (BaO, ZrO_2, Al_2O_3, B_2O_3, TiO_2)$-CaO-SiO_2$ glass systems [106].

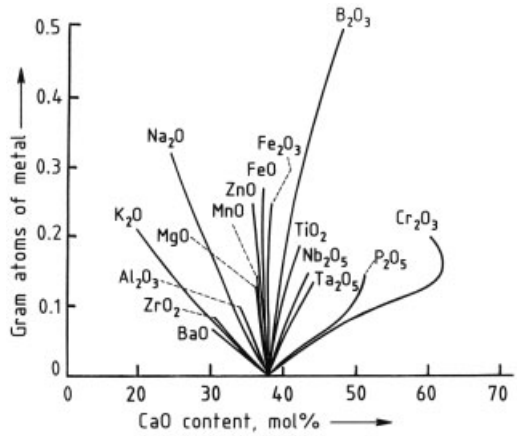

Figure 16. Variation in liquidus boundary in the $CaO-SiO_2$ system as a function of a third added component [105]

The above examples serve to illustrate that no acceptable mathematical formulation for the relation between viscosity and glass composition exists; the viscosity of a glass is determined by the superposition of several factors. The dependence of viscosity on composition has been

examined in several systematic studies, but the results are applicable only to special groups of glasses. Examples include the effect of various components on the viscosity of sodium silicate glasses [107] and the determination of partial molar viscosity coefficients for multicomponent naturally occurring melt systems [108].

4.2. Surface Tension

Surface tension is defined as the force needed to increase the surface area of a material by 1 cm^2. The numerical value of this measure of the strength of a surface is expressed in N/m. Surface tension is an important variable for the following reasons: it predicts the degree of wetting between different melts, melt – metal, and melt – tank lining surfaces; it predicts movement toward the surface of ions that decrease surface tension and away from the surface of ions that increase surface tension; and it suggests a means of removing gas bubbles in a melt.

The surface tension of silicate melts is not very sensitive to either composition (Fig. 17A) or temperature (Fig. 17B), compared to melt viscosity. Characteristic values for glasses and other materials are given in Table 16. See also Figure 17.

Table 16. Surface tension of various glasses and reference materials

	Surface tension, N/m	Temperature, K
Soda lime	0.33	1573
Silica	0.275	1673
PbO_2–SiO_2 melt	0.230	1675
B_2O_3 melt	0.081	1173
H_2O	0.073	293
Molten alloy	0.60	

Most of the available data on interfacial tension are for systems of interest in ferrous metallurgy. Frequently, these measurements are made in relation to the study of specific problems such as wetting characteristics of welding fluxes and dispersion of metal droplets in slags. The Slag Atlas contains a discussion of such systems [101].

4.3. Density

The density of most technical silicate glasses is 2.2 – 2.5 g/cm^3. For molten B_2O_3, the density varies between 1.67 g/cm^3 at 723 K and 1.49 g/cm^3 at 1673 K; for a GeO_2 melt at 1473 K, the density is 3.46 g/cm^3; for molten SiO_2, the density variation between 2208 and 2438 K is given by the equation

$$\varrho = 2.508 - 2.13 \times 10^{-4} t$$

For molten Al_2O_3, the density variation over the temperature range 2373 – 2623 K is given by

$$\varrho = 5.370 - 11.46 \times 10^{-4} t$$

Note that in both equations the temperature, *t,* is in degrees Celsius. Components with high molecular mass may increase the density to ca. 7 g/cm^3.

The molar volume derived from the density is related to the average molecular mass of the liquid

$$V = \Sigma x_i M_i / \varrho$$

where V is the molar volume, in cubic centimeters per mole; ϱ is the density, in grams per cubic centimeter; x_i is the mole fraction; and M_i is the gram molecular mass of pure component i. In terms of partial molar volumes v_i, the molar volume is given by the sum

$$V = \Sigma x_i v_i$$

The volumetric deviation from ideal mixing is defined by the difference

$$\Delta V = V - V_{\text{ideal}} = V - \Sigma x_i v_i^{\circ}$$

where v_i° = molar volume of the pure component i.

The density of oxide melts is commonly a linear function of temperature over a relatively wide range. Densities are often calculated by using additive relations such as the following:

$$\frac{1}{\varrho} = \frac{1}{100} \Sigma \left[p_i / \varrho_i \right]$$

where ϱ is the density, ϱ_i the calculated partial molar or mass density coefficient, and p_i the concentration of component i in the liquid, in mol %

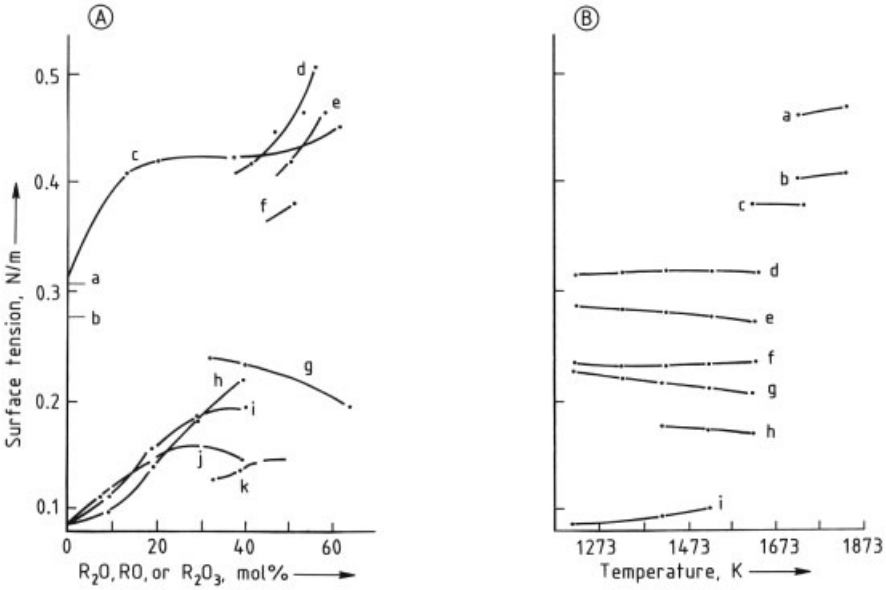

Figure 17. Variation of melt surface tension
A) With composition: a) SiO_2 (2341 K); b) SiO_2 (1149 K); c) Al_2O_3–SiO_2 (2300 K); d) CaO – SiO_2 (1823 K); e) MnO – SiO_2 (1773 K); f) MgO – SiO_2 (1843 K); g) PbO – SiO_2 (1573 K); h) Li_2O – B_2O_3 (1173 K); i) Na_2O – B_2O_3 (1173 K); j) K_2O – B_2O_3 (1173 K); k) Bi_2O_3–SiO_2 (1313 K)
B) With temperature: a) 39.2 CaO · 15.4 Al_2O_3 · 45.7 SiO_2; b) 36.5 CaO · 63.5 SiO_2; c) 40 FeO · 5 MgO · 55 SiO_2; d) 30.3 Li_2O · 69.7 SiO_2; e) 30.1 Na_2O · 69.9 SiO_2; f) 33.4 PbO · 66.6 SiO_2; g) 31.2 K_2O · 68.8 SiO_2; h) 32 Rb_2O · 68 SiO_2; i) B_2O_3

or wt %. Coefficients for calculating the density of soda lime – magnesia glasses in weight percent on the basis of p_i are given below [109]:

SiO_2	2.20
Na_2O	3.47
CaO	5.00
MgO	3.38
Al_2O_3	2.75

Density is somewhat dependent on the degree of cooling, slow rates of cooling giving a higher density. Density measurements are frequently used as quality control indicators in glass plants because this property is very easy to measure (sink – float technique) and deviations from established standards indicate changes in glass composition.

4.4. Thermal Expansion

The thermal expansion of glass is an important property in many applications. Thermally induced stress frequently imposes operating limits because the stress is directly proportional to the rate of expansion. The feasibility of glass – metal seals, processing, annealing, tempering, and strengthening depends on thermal expansion. Thermal expansion is also a factor in determining the ability of a glass article to survive thermal shock and thermal cycling.

Like most materials, glasses expand on heating and contract on cooling. Hysteresis is rarely observed if thermal cycling is sufficiently slow. The linear expansion coefficient α is defined by:

$$\alpha_{\Delta T} = (L/L_0)\,(\Delta L/\Delta T)$$

where L_0 is the initial length, L is the length measured at temperature T, and ΔL the change of length with a temperature change of ΔT. The linear expansion coefficient is used extensively in the glass industry. The variation of linear thermal expansion coefficient with temperature is shown in Figure 18. The average coefficient of linear expansion in the range 273 – 573 K is de-

termined on annealed samples, as specified in ASTM C 598-72 (1978).

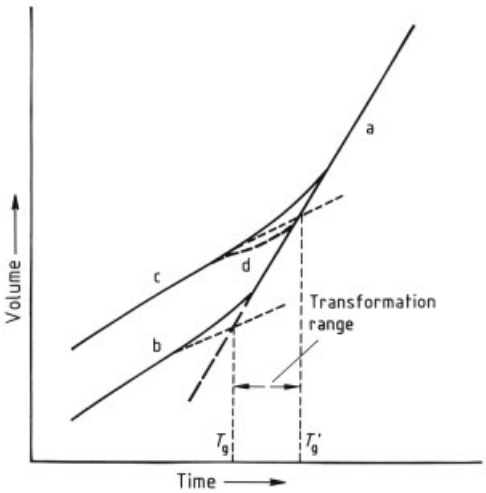

Figure 18. Variation in linear expansion as a function of time a) Undercooled fluid; b) Contraction curve on slow cooling; c) Contraction curve on fast cooling; d) Slow reheating of (c)T_g and T_g' are the glass transition temperatures upon slow and fast cooling, respectively.

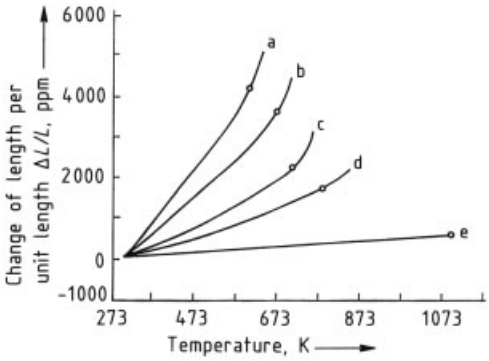

Figure 19. Linear thermal expansion coefficients for some industrial glasses
a) Glass no. 1990, Iron sealing; b) Glass no. 0120, Dumet sealing; c) Glass no. 7052, Kovar sealing; d) Glass no. 7740; e) Glass no. 7900, 96 % silica

Expansion curves are generally nearly linear up to the glass transition temperature. Above this, thermal expansion increases by a factor of ca. 3 – 4 because of variations in vibrational amplitudes. The point at which the extrapolated slopes of both segments of the curve intersect is defined as the *transformation temperature* of the glass, according to DIN 52 324.

Inspection of Figure 19 shows that the thermal expansion of glasses can differ considerably. Technical glasses are customarily classified according to their coefficients of expansion:

Low expansivity: $\alpha < 60\times10^{-7}\,K^{-1}$
High expansivity: $\alpha > 60\times10^{-7}\,K^{-1}$

Characteristic values are

Pyrex $32\times10^{-7}\,K^{-1}$
Soda lime glass $90\times10^{-7}\,K^{-1}$
Cesium borate glass $30\times10^{-7}\,K^{-1}$

Pure SiO_2 glass has a linear expansion coefficient α of $5.0\times10^{-7}\,K^{-1}$ at 293 – 573 K. Addition of other oxides to silica may increase this value to $125\times10^{-7}\,K^{-1}$. Most technical glasses have a thermal expansion between that of Pyrex ($30\times10^{-7}\,K^{-1}$) and commercial soda lime glasses [$(80-90)\times10^{-7}\,K^{-1}$]. Glasses with expansion coefficients below $60\times10^{-7}\,K^{-1}$ are called hard glasses, whereas those with higher expansion coefficients are designated as soft. This distinction between hard and soft coincides with high and low softening points as discussed previously (Section 4.1.1).

Thermal expansion coefficients may be rationalized in terms of the degree of connectivity of the glass network. Introduction of network modifiers increases the thermal expansion of glasses. In the alkali – silica system, lithium silicate glasses show the smallest thermal expansion and potassium silicate glasses the largest. This difference can be rationalized easily by inspection of thermal expansion and compressibility as a function of coordination number of selected metal –oxygen bonds (Table 17). The observed α/β ratio is approximately constant for many cation octahedra. Changes in bond length during thermal expansion, therefore, tend to be the inverse of those during compression. Lithium, sodium, and potassium, with assumed oxygen coordination variation between four and eight in silicate melts, show a difference in R–O expansion ($\Delta\alpha$) of $42\times10^{-7}\,K^{-1}$. Partial molar expansion coefficients have been calculated and compiled [111].

The transformation temperature has been defined as that point on an expansion – temperature graph at which straight lines drawn to the thermal expansion curves above and below the glass transition temperature intersect. Volume changes in this transformation region are of great practical use. A plot of expansion coefficient versus time is shown in Figure 18. The equilibrium

Table 17. Predicted values of thermal expansion and compressibility for element – oxygen bonds in oxygen-based minerals [110]

Element	Coordination number	$\alpha \times 10^{-6}$, K^{-1}	$-\beta \times 10^{-6}$, $kbar^{-1}$	α/β
Si(IV)	4		39	
Ge(IV)	4		50	
B(III)	4		39	
Al(III)	4		65	
Al(III)	6	8.2	83	0.099
Li(I)	4	16.4	274	0.060
Li(I)	6	19.5	353	0.055
Na(I)	6	19.5	499	0.063
Na(I)	8	20.6	606	0.034
K(I)	6	19.5	761	0.026
K(I)	8	20.6	875	0.024
K(I)	12	27.4	999	0.027
Rb(I)	6	19.5	884	0.022
Rb(I)	8	20.6	969	0.021
Cs(I)	8	20.6	1102	0.019
Mg(II)	6	13.8	167	0.071
Mg(II)	8	16.4	211	0.067
Ca(II)	6	13.8	243	0.057
Ca(II)	8	16.4	282	0.058
Sr(II)	6	13.8	303	0.046
Sr(II)	8	16.4	333	0.049
Ba(II)	6	13.8	368	0.038
Ba(II)	8	16.4	442	0.037
Co(II)	6	13.8	174	0.079
Co(III)	6	8.2	89	0.092
Cr(III)	6	8.2	94	0.087
Fe(II)	4	8.2	146	0.056
Fe(II)	6	13.8	181	0.076
Fe(III)	4		78	
Fe(III)	6	8.2	99	0.083
Ga(III)	4		76	
Mn(II)	6	13.8	194	0.071
Mn(II)	8	16.4	231	0.106
Mn(III)	6	8.2	100	0.082
Ni(II)	6	13.8	159	0.087
U(IV)	8	8.2	122	0.067
V(III)	6	8.2	99	0.083
W(IV)	6		76	
W(VI)	4		41	
W(VI)	6		46	
Zn(II)	4	8.2	139	0.059
Zn(II)	6	13.8	171	0.081
Zr(IV)	6	2.7	83	0.033

volume is reached almost spontaneously above the glass transition temperature. Lowering the temperature and hence increasing the viscosity result in a point being reached at which the volume relaxation time becomes at least as large as the observation time. The cooling glass body does not reach its equilibrium geometry, and a volume change is observed as a result. Thus care must be taken in cooling larger glass bodies to obtain a final product with acceptable low stress.

The feasibility of glass-to-glass or glass-to-metal seals depends, apart from interfacial tension, on the absence of glass reboil and degassing of the metal at the sealing temperature, and on the thermal expansion mismatch of the two materials. For glass-to-glass sealing with acceptable reduced strain, the difference in thermal expansion coefficient generally should not exceed ca. $3 \times 10^{-7}\,K^{-1}$.

The production of glass-to-metal seals is complicated by the substantially larger linear thermal expansion of steel, which varies between $130 \times 10^{-7}\,K^{-1}$ for unalloyed steel and $160 \times 10^{-7}\,K^{-1}$ for stainless steel. Conditions

are relatively simple if thin metal sections are used or if the seal is a so-called compression seal, where the higher expansion metal surrounds the glass and, after cooling, keeps the glass in compression.

Two properties must be controlled to ensure proper glass-to-metal seals. The first involves a complicated and rigorously controlled oxidation – reduction at the interface to form a good bond; the second involves matching the thermal expansion of the glass with that of the metal. The latter causes no problems as long as the glass adjusts to any thermal expansion mismatch by viscous flow. The system must be designed so that below the set point, thermal contractions of the glass and the metal are similar in magnitude and mismatches are taken up preferentially by the metal.

4.5. Elastic Properties

Two classes of variable suffice to describe the mechanical properties of glass: (1) elastic variables and (2) mechanical strength. An *elastic material* has the ability to recover size and shape after deformation. An ideal elastic material deforms instantaneously when subjected to stress, and the deformation recovers instantaneously when the stress is removed. Elasticity is measured as Young's modulus E, which is defined as the ratio between linear stress and collinear strain for a rod or bar, or as bulk modulus K, if a uniform compressive stress is applied to a body. The compressibility K is the inverse of the compressive stress. Measurement of the elastic modulus on annealed samples is described in ASTM C 623-71 (1981) and entails testing the deformation of a material in the direction of the applied force.

$$E = \sigma / \varepsilon$$

Here E is the proportionality constant between applied stress, σ, and the resulting strain, ε. The ratio of the resulting lateral stress to the applied longitudinal stress is the Poisson ratio σ. For shear stress, the proportionality constant is the shear modulus G. The relation of Young's modulus E, bulk modulus K, shear modulus G, and Poisson ratio σ is as follows:

$$E = 3K\,(1 - 2\sigma) = 2G\,(1 + \sigma)$$

Glasses are perfectly elastic: the deformation is directly proportional to the applied stress, independent of loading rate, until failure. A high-modulus glass is one in which the atoms are closer together or the bonds are stronger and hence more resistant to deformation. Vitreous silica has an elastic modulus of about 72.9 GPa, a shear modulus of 31.3 GPa, and a Poisson ratio of 0.165. Below ca. 1273 K it is close to an ideal elastic body under most loading conditions. It deforms instantaneously, does not creep (i.e., deform slowly under continuous load), and recovers instantaneously on removal of the load. Introduction of alkali in vitreous silica decreases the connectivity and hence the elastic modulus. Some elastic constants for alkali silicate and other glasses are collected in Table 18.

The modulus can again be increased by adding network formers (e.g., B_2O_3 or Al_2O_3) to alkali silicate glasses, as illustrated in Table 16 for the sodium – alumina – silica system. Similar results can be obtained by adding oxides (e.g., CaO, MgO, La_2O_3, or ZrO_2). Relatively high moduli are encountered in alkaline-earth aluminosilicate glasses, with values as high as 118 GPa. Elastic moduli normally tend to decrease with increasing temperature and increase with increasing pressure, approaching zero only above the glass transition temperature.

Vitreous silica, Pyrex, and some other glasses with low modifier concentrations, behave anomalously: the elastic modulus increases with increasing temperature. The rationalization for this is the uncommonly small thermal expansion coefficient of the Si–O and B–O bonds, which have thermal expansions of zero between 296 and 1273 K.

4.6. Mechanical Strength

The theoretical strength of a glass is extremely high ($> 10^4$ MPa) when the energy required to break chemical bonds is calculated. In practice, however, measured strengths are much lower and vary widely. Typical values are ca. 10^4 MPa for glass fibers, 10^2 MPa for flame-polished silica rods, and 7×10^3 MPa for thin-walled freshly drawn glass. The reason for the discrepancy between theory and practice, as well as the

Table 18. Elastic constants of silicate glasses [74]

Composition, mol %	E, GPa	G, GPa	K, GPa	σ
$20\,Li_2O \cdot 80\,SiO_2$	76.99	31.40	46.81	0.226
$20\,Na_2O \cdot 80\,SiO_2$	61.08	25.18	35.46	0.213
$20\,K_2O \cdot 80SiO_2$	49.04	19.75	31.62	0.242
$20\,Cs_2O \cdot 80\,SiO_2$	37.67	14.86	27.02	0.274
$20\,Na_2O \cdot 20\,Al_2O_3 \cdot 60\,SiO_2$	73.2	30.6	40.13	0.196
$30\,PbO \cdot 70\,SiO_2$	50.13	21.35	25.63	0.174
$25\,MgO \cdot 25\,CaO \cdot 50\,SiO_2$	99.14	38.81	74.11	0.277
$25\,MgO \cdot 25\,BaO \cdot 50\,SiO_2$	79.52	30.97	61.24	0.284

scatter in data, is the inclusion of many surface and internal flaws that act as stress concentrators or risers. Fracture is usually in the tensile, as opposed to the compressive, mode due to this type of stress concentration. The useful strength of glass articles is, therefore, only a fraction of the theoretical value.

The susceptibility of glass to stress corrosion additionally reduces the practical strength of a glass article by a further 30 % in long-term usage. In building construction, the maximum allowable tensile load on a soda lime glass is 7.0 Pa under sustained loading (at least 1000 h). The chemical composition of glass has a significant effect on its design strength. The allowable stress for glasses with greater resistance to stress corrosion (e.g., fused silica) can be twice as high as that for soda lime glass.

Plastic deformation of glass is prevented by its high viscosity at room temperature, combined with the absence of slip planes. Broken bonds resulting from excessive stress tend not to anneal but function as nucleation sites for cracks, unless enough time passes for rounding of the tips due to aging or zero stress corrosion. If the crack grows, the stress intensity concentration at the tip increases. Those bonds that were intact will break, causing glass fracture.

Glass can be made more load-resistant in several ways: The number of surface flaws can be reduced, by careful handling, annealing after forming, or acid etching. The highest-strength common glass articles are optical waveguide fibers that remain untouched until a coating which protects the fibers from mechanical abuse is applied. Secondary treatment, such as ion exchange, thermal tempering, glazing, differential crystallization, or layering, can induce compressive stresses on the surface of the glass which must be overcome before surface stress reaches the necessary tensile stress that leads to fracture.

4.7. Hardness

Glass hardness is tested in one of three ways: scratching, grinding or abrasion, and penetration. Results obtained with these methods, can be correlated to some extent. However, the hardness sequence for a series of glasses may depend on the method. The most commonly used is Knoop hardness, measured with a diamond pyramid indenter. Historically, Moh's scale of scratch hardness has been employed. On this scale, glass hardness lies between that of the calcium phosphate apatite, with a hardness of 5, and quartz, with a hardness of 7. Common materials that are hard enough to scratch glass include diamond, agate (amorphous SiO_2), sand, silicon carbide, hard steel, and emery. Glass itself can be used as an abrasive.

4.8. Thermal Properties

Heat transfer and transport in melts and glasses are important variables in the design and control of material processing at elevated temperature. They occur in two ways in rigid glass: conduction and radiation. In a glass melt convection also transports heat. In dielectric (i.e., polarizable) materials, two independent modes of heat transfer operate: thermal (phonon) conduction due to lattice wave propagation, and radiative (photon) conduction due to absorption and reradiation of electromagnetic energy. At low to moderate temperature, the transfer of thermal energy occurs primarily by phonons; at elevated temperature, heat transfer by photons predominates. Knowledge of heat transfer is essential

in growing crystals from semitransparent melts and in melting, annealing, and forming glasses. Heat-transfer values are also an important prerequisite for the construction of furnaces, molds, and energy balances in glass technology.

Thermal conductivity is measured in watts per meter per degree; the room temperature thermal conductivity of glass is 0.67 – 1.38 $W\,m^{-1}\,K^{-1}$. Common glass compositions have thermal conductivity values near the upper end of the range. Conductivity increases by ca. 25 % from room temperature to 473 K. Radiative heat transfer appears in clear glasses around 800 K and becomes the major source of heat transport at higher temperature. Vitreous silica has the highest thermal conductivity at room temperature (1.38 $W\,m^{-1}\,K^{-1}$), and high lead silicate glass has the lowest (0.71 $W\,m^{-1}\,K^{-1}$). For Corning Code 7740, a Pyrex brand glass, the thermal constants are as follows: thermal conductivity (298 K), 0.96 $W\,m^{-1}\,K^{-1}$; mean specific heat (298 – 448 K), 0.84 $J\,g^{-1}\,K^{-1}$; thermal diffusivity, 0.56 $mm^2\,s^{-1}$; and emissivity coefficient, 0.94.

The common heat capacity of glass at room temperature is ca. 0.837 $J\,g^{-1}\,K^{-1}$. Glasses that contain atoms of high atomic mass show somewhat lower specific heats (e.g., 0.5 $J\,g^{-1}\,K^{-1}$ for a 50 wt % PbO – silicate glass). The specific heat increases slightly with increasing temperature, e.g., for vitreous silica: 0.73 $J\,g^{-1}\,K^{-1}$ at 298 K; 1.12 $J\,g^{-1}\,K^{-1}$ at 773 K; and 1.21 $J\,g^{-1}\,K^{-1}$ at 1273 K. An average value of 1.05 $J\,g^{-1}\,K^{-1}$ in the region 273 – 873 K is typical.

4.9. Electrical Properties

Electrical conductivity in industrial glasses is associated with the transport of ions. The temperature dependence of electrical conductivity in glasses is positive; conductivity increases with increasing temperature. This property of glass is useful in electrical engineering, where low conductivity is required at ambient temperature, and in electric melting of glass, in which corrosion-resistant electrodes are submerged in the melt and the relatively low electrical resistance of the melt itself at high temperature is used. Electrical resistance R, is defined by the following equation:

$$R = \varrho l / a$$

where ϱ is the resistivity, a is the cross-sectional area perpendicular to the direction of current flow, and l is the length of the sample. Electrical conductivity is defined as the inverse of resistivity. Log ϱ is reported at 523 and 623 K for comparative purposes and to establish temperature dependence.

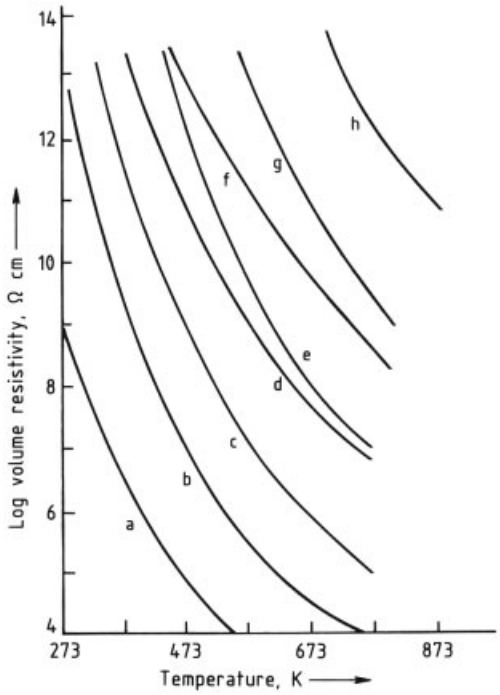

Figure 20. Volume resistivities for some industrial glasses a) Glass no. 8058; b)Glass no. 0080; c) Glass no. 7740; d) Glass no. 7900; e) Glass no. 7070; f) Glass no. 7040; g) 33 $CaO \cdot 67\,B_2O_3$; h) Glass no. 9753

The electrical conductivity at room temperature for common silicate glasses is 10^{-10} – $10^{-11}\,\Omega^{-1}\,cm^{-1}$, indicative of their good insulating properties. Glass can be used as an electrical insulating material (→ Insulation, Electric). The value $T_{K=100}$ is the temperature at which a glass has a specific conductivity of $10^{-8}\,\Omega^{-1}\,cm^{-1}$. For window glass, which is not a particularly good electrical insulator, this temperature is between 423 and 473 K. Water layers on the surface of glass at low temperature may increase surface electrical conductivity by a factor of 10^3. Some examples of volume resistivity are shown in Figure 20.

In highly connected liquid oxides such as B_2O_3, GeO_2, and SiO_2, electrical conductivities are typically $<10^{-5}\,\Omega^{-1}\,cm^{-1}$ at the melting point, consistent with the high viscosity. Melts of network-modifying oxides, on the other hand, have low viscosity ($<0.05\,Pa \cdot s$) and high specific conductivity ($>1\,\Omega^{-1}\,cm^{-1}$) at the melting point, indicating dissociation into M^{n+} and O^{2-} ions. Most of these molten oxides have higher conductivities than molten salts, indicating that both electron conduction and ion conduction occur. This is particularly true when the metal atoms have a variable valence (e.g., Fe^{2+} and Fe^{3+} in liquid iron oxide). A characteristic of ionically conducting melts is that their specific conductivity increases sharply on fusion, in contrast to melts that conduct by means of electrons.

Specific conductivity in melts depends on the type and amount of alkali ions present. Some examples are given in Table 19.

Table 19. Electrical conductivity of molten alkali borates and silicates

R_2O, mol %	Conductivity, $\Omega^{-1}\,cm^{-1}$
$10\,Li_2O \cdot 90\,B_2O_3$	0.051 (1273 K)
$10\,Na_2O \cdot 90\,B_2O_3$	0.045 (1273 K)
$10\,K_2O \cdot 90\,B_2O_3$	0.040 (1273 K)
$Li_2O - SiO_2$ [a]	1.7 – 6.3 (1773 K)
$Na_2O - SiO_2$ [b]	0.64 – 1.8 (1773 K)
$K_2O - SiO_2$ [c]	0.40 – 1.1 (1773 K)

[a] Li_2O 33 – 58 mol %.
[b] Na_2O 19 – 34 mol %.
[c] K_2O 18 – 34 mol %.

Mixtures of alkali-metal oxides in silicate, borate, or phosphate glasses give rise to anomalously low electrical conductivities. This effect, which is not well understood, is called the *mixed alkali effect.* An example is the specific conductivity of a $Na_2O \cdot K_2O \cdot 4\,SiO_2$ glass ($10^{-12}\,\Omega^{-1}\,cm^{-1}$) which is much smaller than that of either $Na_2O \cdot 2\,SiO_2$ glass ($10^{-8}\,\Omega^{-1}\,cm^{-1}$) or $K_2O \cdot 2\,SiO_2$ glass ($10^{-9}\,\Omega^{-1}\,cm^{-1}$). This effect is not limited to cations such as alkali metals but also occurs with anions (e.g., halides). In this case, it is called the *mixed halide effect.* This effect has been characterized in lead silicate glasses [112].

The temperature dependence of electrical conductivity is given by the relation:

$$\log(1/\varrho) = A - \frac{B}{T}$$

The activation energy for charge transport is lower above the glass transition temperature. The activation energy E for alkali-metal silicates is 35 – 50 kJ and is virtually independent of composition or type of cation present. For alkali-metal borates, the activation energy is about twice as large. The activation energy for alkaline-earth silicates is about 180 kJ at 20 mol % R_2O and decreases to about 85 kJ at 60 mol %, values similar to those found in borate melts. For iron, manganese, and lead silicates, E is about 40 kJ lower than for alkaline-earth silicates. For all these silicate melts, E is about 40 kJ lower than the activation energy for viscous flow. This activation energy is determined by measuring the temperature dependence of the melt viscosity.

The static *dielectric constant* represents the ratio of the energy of the electric charge stored between two metal plates with the material as insulator to the energy of the electric charge stored between the same plates with vacuum as insulator. The magnitude of the dielectric constant depends on the polarizability of the glass constituents. Pure vitreous silica, with its very low polarizability, has a dielectric constant of 3.7. High lead glasses can have a dielectric constant at ca. 15, and specialty glasses with high PbO, BaO, or TiO_2 content have a dielectric constant of ca. 60. The numerical value of dielectric constant can be estimated for most industrial glasses (except for vitreous silica) by using the empirical formula:

$$\text{Dielectric constant} = 2.2 \times \text{density}$$

where the density is expressed in grams per cubic centimeter.

The *dielectric loss,* a dissipative factor, is a measure of the deviation of a material from an ideal dielectric. Generating an electric field in an insulating material introduces a specific quantity of electrical energy. If this energy is partially recoverable, the material is called a dielectric. The irrecoverable part of the energy is lost as heat. The ratio between irrecoverable and recoverable energy is called the *dielectric loss.* At least three types of dielectric loss can be distinguished: resonance vibrational losses, migration losses caused by mobile ions, and deformation

losses related to defects in the silicon – oxygen network.

Dielectric loss depends strongly on frequency. Conductivity and relaxation losses dominate at low temperature. The influence of composition on energy loss in this temperature region is comparable to that observed for electrical conductivity. Dielectric losses have a minimum value at room temperature in the range of $10^6 - 10^8$ Hz. Losses for common glasses in this frequency range vary between 10^{-2} and 3×10^{-2}. Dielectric loss can be lowered to $<10^{-3}$ (10^6 Hz, 298 K) in specialty glasses, usually PbO – BaO rich or mixed alkali borosilicate glasses. Vitreous silica has still lower loss.

Dielectric failure may be due to discharge or arcing, thermal failure, or intrinsic breakdown. In *thermal failure,* electric losses within the dielectric raise its temperature, which in turn increases the losses and flow of current. Breakdown occurs when this cumulative action results in instability. *Intrinsic breakdown* is associated with defects and voids in the glass. The intrinsic dielectric strength of glass is high, e.g., for silica, borosilicate, soda lime, and lead – alkali, values are 5000, 4800, 4500, and 3100 kV /cm (d.c.), respectively, for thin sections.

4.10. Gas Permeability

The permeability of glass to gas is important in high-intensity tungsten – halogen lamps. The glass must be able to withstand leakage of pressurized halogen at elevated temperature. The volume of gas that permeates through glass is directly proportional to the time of exposure, the area of the glass, and the difference between ambient and gas pressure, and is inversely proportional to the thickness of the glass:

$$V = \frac{ktAp}{l}$$

where t is time, A is area, p is pressure, l is thickness, and k is a constant. The constant of proportionality k is thus given by the equation:

$$k = \frac{Vl}{tAp}$$

Gas permeation is also strongly dependent on temperature, as is shown by plotting k as a function of temperature (see Fig. 21).

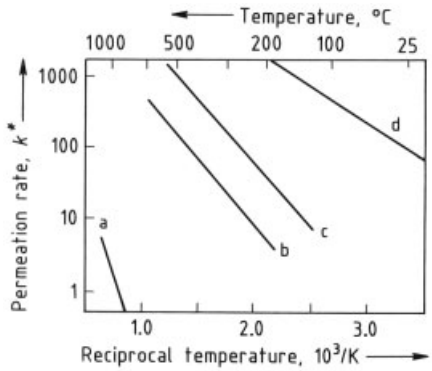

Figure 21. Dependence of gas permeation upon temperature for 96 wt % SiO_2 glass (Corning code 7900)
The constant of proportionality, k, is plotted as a function of the temperature.
a) Oxygen; b) Neon; c) Hydrogen; d) Helium∗ Value of k when V is in cubic meters, l in meters, t in days, A in square meters, p in 10^{15} Pa.

4.11. Chemical Stability and Surface Properties

The chemical stability of a glass depends on its chemical composition and on the medium with which it is in contact. The solubility of amorphous silica is illustrated in Figure 22 and of alumina, a typical amphoteric element soluble in both acids and bases, in Figure 23. Such amphoteric solubility behavior is commonly determined by the ability of an element to change its coordination state as a function of pH. Thus, in the case of alumina, the prevailing solution species at high pH has aluminum coordinated tetrahedrally to four hydroxyl groups $[Al(OH)_4]^-$. At low pH, the prevailing solution species is an aluminum atom surrounded octahedrally by water molecules, $Al(H_2O)_6^{3+}$. If these coordination changes are considered as a function of pH, the formation of zeolites containing only fourfold oxygen-coordinated aluminum atoms in neutral or alkaline environment is certainly no coincidence, whereas clay minerals with sixfold oxygen-coordinated aluminum atoms are always formed in an acidic environment.

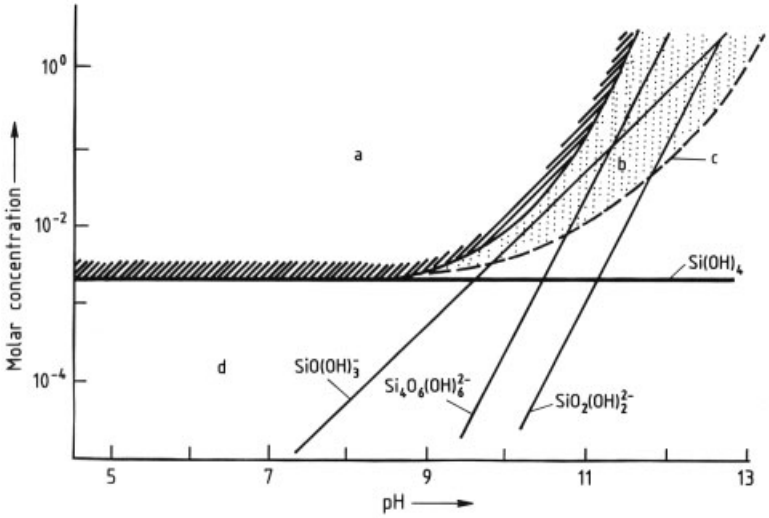

Figure 22. Solubility of amorphous SiO_2 as a function of pH [113]
a) Insolubility domain; b) Stable multimeric domain; c) Mononuclear wall; d) Mononuclear domainSolution NMR shows that silica speciation at high silica concentration is more extended than indicated here [67]. At low silica concentration and neutral pH silica dimers occur [114]. The mononuclear wall (dashed line) and mononuclear domain indicate the region in which mononuclear silicon species occur, as shown by potentiometry.

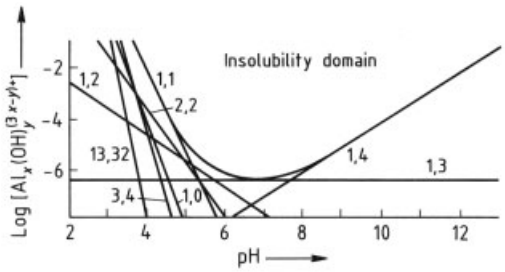

Figure 23. Solubility of amorphous Al_2O_3
The numbers represent the x and y values for a specific solution species. ^{27}Al solution NMR shows fewer species than indicated in this diagram [103], [104].

The solubility behavior of alkali-metal aluminosilicate glasses can be inferred qualitatively from that of silica and alumina. A review of glass – water interactions and solubility has been published [115].

Vitreous silica is soluble to about 100 ppm in acidic or neutral solution (Fig. 22). The corrosion of alkali silicate glasses in such solutions involves an exchange between alkali-metal atoms and protons in a diffusion-controlled process. The silica framework is thought to remain intact and is presumed to form a silica gel-like protective surface layer. Only in leaching with hydrofluoric acid or concentrated phosphoric acid is the silica framework itself attacked. In a diffusion controlled process such as the alkali-metal atom enchange, the rate of dissolution decreases as the square root of time during glass leaching.

The temperature coefficient of glass leaching reactions increases by a factor of ca. 1.5 for every 10° increase in temperature. Increasing the network-former content or doping with multivalent cations strengthens the network. Next to vitreous silica, borosilicate and aluminosilicate glasses have the best acid resistance. No protective gel coating is formed during acid leaching of phosphate or borate glasses; therefore, these glasses are not very resistant to acidic solution.

Alkaline attack involves dissolution of the glass framework and formation of anions of silicic acid. Glasses strengthened with CaO, MgO, Al_2O_3, or B_2O_3 increase corrosion resistance against bases. The most corrosion-resistant glasses are those with substantial ZrO_2 content.

Glass corrosion can be used to advantage. One industrially important class of materials consisted of the soluble silicates. These are solutions with a viscosity on the order of 1 mPa · s, manufactured by dissolution primarily of sodium or potassium silicate glasses in pressurized steam. A review of their industrial uses is given in [116].

Water dissolution can be considered a combination of acid and base attack. First, the pH of the water increases by inclusion of protons in the glass network at the expense of alkali metal; this is followed by attack of the network in the resultant basic environment. Base attack plays a particularly prominent role in glasses because of daily fluctuations in climate. Morning dew causes proton – alkali metal exchange, followed by an increase in alkali-metal concentration on evaporation, which enhances attack by alkali.

For industrial purposes, weathering is defined as corrosion by atmospheric-borne gases and vapors, such as water or carbon dioxide. The acid resistance of glasses is classified by their behavior in 5 % hydrochloric acid at 368 K (95 °C) for 24 h.

Metal or metal halide attack on silica glass at elevated temperature is especially important in light-bulb production. Reaction between vitreous silica and various metals or halides occurs at elevated temperature. Thus, although alkali metals stain glass only at relatively low temperature (<573 K), severe induced devitrification can occur at higher temperature (ca. 1173 K); with attendant glass failure. A 1-h exposure of silica tubes to sodium vapor below 1273 K causes failure due to devitrification. Other elements that induce devitrification at 1523 – 1623 K are Mn, Sn, Fe, and Zn. Metal halides that react with vitreous silica below 1473 K are LiCl, NaCl, $SnCl_2$, NaBr, and KBr. At 1473 – 1673 K, RbCl, CsCl, NaI, KI, $MgCl_2$, $CaCl_2$, $BaCl_2$, and $AlCl_3$ cause failure.

Surface Properties. The acid-base properties of glass surfaces in aqueous solution depend on the point of zero charge (pzc), i.e., that pH at which the surface is not charged. The surface is positively charged if the pH is lower than the pzc, and negatively charged if the pH is higher. The reason for this change in surface charge is that metal oxide surface chemistry in water is dominated by the following reactions [117].

$$\mathrm{MOH_2^+} \longrightarrow \mathrm{MOH} + \mathrm{H^+} \quad \mathrm{p}K_1$$
$$\mathrm{MOH} \longrightarrow \mathrm{MO^-} + \mathrm{H^+} \quad \mathrm{p}K_2$$

Here M represents a surface metal ion and OH is the surface hydroxyl exposed to the medium. The two reactions indicate that a positive surface charge is associated with Bronstedt acidity (proton donor) whereas a negative surface charge is associated with Bronstedt basiscity. The point of zero charge (pzc) is the average of the two pKs:

$$\mathrm{pzc} = \frac{\mathrm{p}K_1 + \mathrm{p}K_2}{2}$$

Some typical values are shown in Table 20.

Table 20. Values of pK_1, pK_2, and pzc for some oxides

Oxide	pK_1	pK_2	pzc
Silica	−2.77	6.77	2.0
Alumina	7.70	9.30	8.5
Titania	5.40	6.40	5.9
Goethite, FeO(OH)	6.99	8.40	7.7

As might be anticipated, a relation between the point of zero charge and cationic field energy can be constructed [118]. High field energy is associated with a pzc at low pH, and low field energy with a pzc at high pH.

The importance of the pzc's of oxides for glasses lies in the fact that pzc's for crystalline and amorphous materials are similar. The pzc's of glasses can therefore be estimated from the pzc's of their constituent oxides. For example, the pzc of quartz occurs at a pH of 2, and the pzc's of the high-silica glasses Pyrex, Vycor, and silica glass occur at a similar pH.

The minimum rate of dissolution of crystalline silicas occurs at their pzc (i.e., pH 2) [119]; this is of importance in adjusting glass compositions for maximum durability.

4.12. Optical Properties

Light transmission, absorption, reflection, and scattering describe the macroscopic aspects of light – glass interaction. The optical properties of glasses can be characterized simply and unequivocally by the spectral transmittance τ_λ as a function of wavelength λ of the irradiating light. Spectral transmittance is the ratio of the light flux Φ_d leaving the glass to the light flux Φ of monochromatic light radiation:

$$\tau_\lambda = \Phi_\mathrm{d}/\Phi$$

The incident light flux can be partitioned over three light – glass interactions:

$$\Phi = \Phi_\varrho + \Phi_\mathrm{a} + \Phi_\mathrm{d}$$

where Φ_d is the transmitted flux, Φ_a the absorbed flux, and Φ_ϱ the reflected flux.

4.12.1. Transmission, Absorption, and Reflection

Light is in part reflected, absorbed, and transmitted. *Absorption* of light requires the operation of an absorption mechanism in the glass. Visible light absorption usually occurs by excitation of electrons from the valence to the conduction band and often results in a change of color of the glass.

The total *transmission* of a glass is determined by reflection at the glass surfaces and optical absorption in the glass. Neglecting multiple reflections between the glass surfaces, the total transmission in air of a flat sample at a specific wavelength is approximately equal to $(1-R)^2 e^{-\beta t}$ where β is the absorption coefficient, t the thickness of the glass, and R the air–glass reflection coefficient. Transmission is a function of wavelength. The lower wavelength transmission limit in ultrapure silica is ca. 150 nm in the ultraviolet. The upper wavelength transmission limit occurs at 5000 nm in the far infrared due to excitation of vibrational states. The lower transmission limit occurs in practice at only ca. 2700 nm because of the ubiquitous presence of water in vitreous silica.

Transmission can be controlled by the type of glass (e.g., silicate, phosphate, or borate) or by the control of coloring additives, oxidizing or reducing melting atmosphere, melting temperature, and cooling schedules. Diffuse transmission occurs in glasses that contain suspended particles (e.g., opal glasses).

Reflectance depends strongly on the quality of the glass surface, the angle of incidence, the indices of refraction of the glass and the surrounding medium, and the wavelength. Reflectance can be specular, as in polished or precision-molded surfaces, or diffuse for ground or irregularly etched ones. The fraction of light of normal incidence reflected by a single plane surface is given by

$$R = [(n-1)/(n+1)]^2$$

in which n is the refractive index of the glass at the wavelength of interest (if the refractive index of air is assumed to be 1). Reflectance loss at normal incidence for simple soda lime glasses is about 4 % per surface, so that 92 % of the incident light is transmitted by a perfectly clear soda lime glass. Glass with a higher index of refraction transmits less light. Thus, glasses with a refractive index of 1.7 and 1.9 transmit 87 and 81 %, respectively, of the incident light. Low-index coating or special surface treatment of glass may significantly reduce losses due to reflection.

4.12.2. Refractive Index, Dispersion, and Optical State

The most important characterizations of an optical glass are refractive index and Abbe number. The refractive index of a glass is defined as the ratio of the velocity of light in a vacuum to that in the glass; larger refractive indices, therefore, indicate lower velocity of light in the material.

Refractive index depends on wavelength, so that values are reported with the wavelength employed. Wavelengths commonly used are the yellow helium D line at 587.6 nm, the green mercury line at 546.1 nm, and the yellow sodium D line at 589.29 nm. The refractive index is generally measured against air, although the refractive index of air is defined against a vacuum ($n_{air}/n_{vac} = 1.00023$). Refractive indices are usually between 1.46 (amorphous silica) and 1.97 (lead silicate glass). Some exotic glasses reach a refractive index exceeding 2.0. Comparison of refractive index variations of amorphous and crystalline silicas and alkali silicate glasses suggests that intra-orbital charge transfer on the oxygen atom, i.e., variation in bonded and nonbonded electron density on the oxygen atom (Fig. 24), are primarily responsible.

The refractive index is proportional to the square root of the number density of dipole oscillators in a material. Thus, the refractive index of amorphous and crystalline silica increases systematically with increasing density (see Fig. 25). Changes in the electronic structure of silica occur concomitantly with density variations, as can be seen from variations in the ^{29}Si MAS NMR chemical shifts. These chemical shift variations result from changes in s electron density at the silicon nucleus (Fig. 26). Thus, the charge on oxygen and silicon in silica varies as a function of density. On the other hand, oxygen 1 s photoelectron spectroscopy indicates that the overall electronic charge on oxygen is -0.74 for vitreous silica, quartz, and cristobalite. This similar-

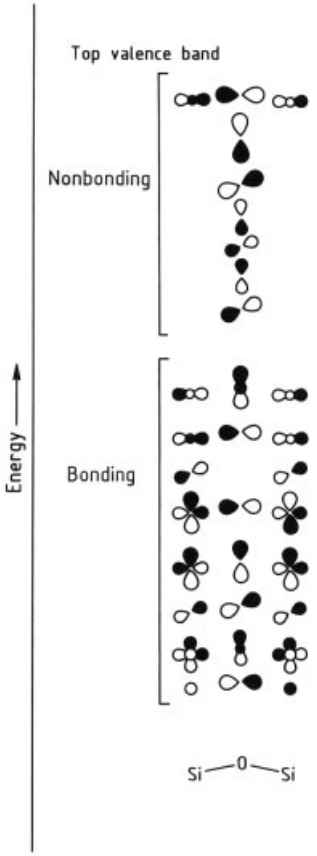

Figure 24. Change of molecular orbitals into atomic orbitals of silicon and oxygen in a Si—O—Si bond [79]

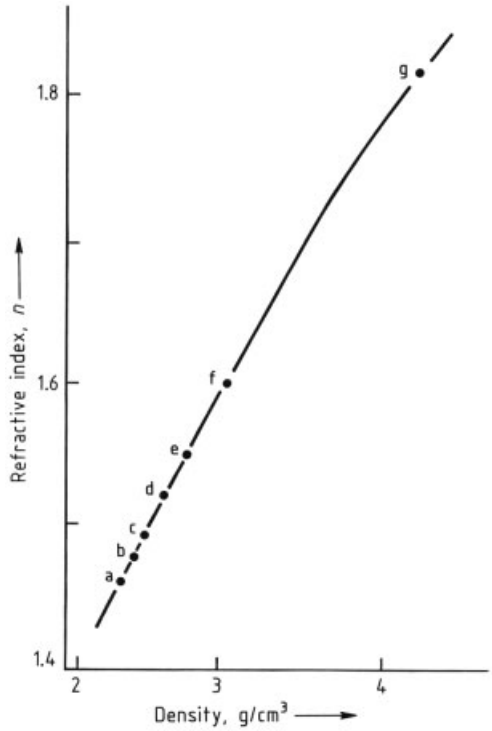

Figure 25. Variation in refractive index as a function of density for silica
a) Lechatelierite, a natural amorphous silica; b) Tridymite; c) Cristobalite; d) Keatite; e) Quartz; f) Coesite; g) Stishovite

ity in charge on oxygen and, therefore, silicon in crystalline and amorphous silicas suggests that a more subtle variation in electron density is responsible for chemical shift and polarizability variations in silicas.

One subtle variation in electron density in Si—O bonds is associated with shifts in the bonding and nonbonding electron density on the oxygen atom as a function of Si—O—Si angle. Hartree–Fock calculations on $H_6Si_2O_7$ show that although the overall charge on oxygen does not vary as a function of the Si—O—Si angle, the ratio of nonbonded to bonded electron density decreases with decreasing Si—O—Si angle from 1.44 (180°) to 1.07 (120°). This shift in electron density accounts for the increased polarizability of the oxygen atom and for the observed shift of the ^{29}Si MAS NMR resonances to lower field with smaller Si—O—Si angle.

The density of a glass and its refractive index are not necessarily coupled with one another. Plots of density versus refractive index of 20 mol % alkali silica glasses (Fig. 27) show that no clear cause and effect relation between these two variables exists. The reason for the absence of such a relation in the alkali silica glasses presumably lies in the increase in Si—O—Si angle with higher atomic mass alkali atoms, hence decreasing the number density of oscillators with increasing glass density. Alternatively it can be argued that larger Si—O—Si angles decrease the oxygen polarizability, hence accounting for the observed refractive index behavior. The crystal structures of the heavier alkali silicates have not yet been determined. Hence, such increase

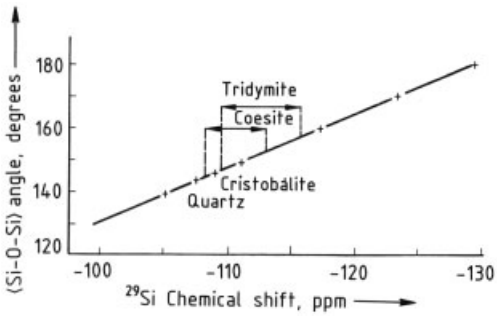

Figure 26. Variation in ^{29}Si chemical shift as a function of Si–O–Si angle [66]
The chemical shift is related to the Si–O–Si angle by the following formula: δ (ppm) = – 19.8215 – 0.608526 ⟨Si–O–Si⟩

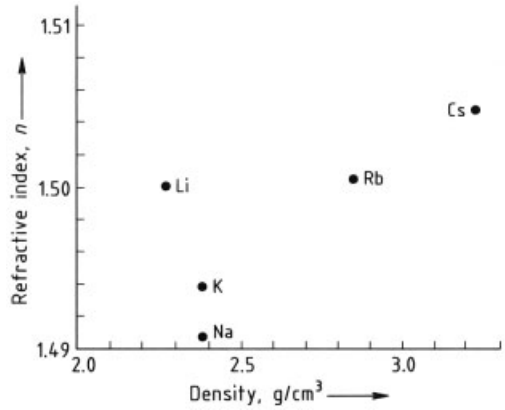

Figure 27. Variation in refractive index with density for $20\,R_2O - 80\,SiO_2$ silicate glasses

in Si–O–Si angle with heavier alkalis in crystalline alkali silicates is still a matter of conjecture [120].

Optical glasses are usually described in terms of refractive index at the sodium D line, n_D 589.3 nm, and ν value (Abbe number). The Abbe number is calculated from the relation

$$\nu = (n_D - 1) / (n_F - n_C)$$

in which n_F and n_C are the refractive indices at the hydrogen F (486.1 nm) and C (656.3 nm) lines, respectively. The difference $n_F - n_C$ is the average dispersion, and its reciprocal $1/\nu$ is the relative dispersion. Materials with high dispersion have Abbe values around 30; those with low dispersion, around 80.

The refractive index and the Abbe number are incorporated by Schott Glass into a six-digit numbering system used to identify optical glasses. Common spectacles are designated 523 – 586, i.e., $n_D = 1.523$ and $\nu = 58.6$.

Crown glasses have $n_D < 1.60$ and $\nu \geq 55$. They are usually alkali silicate glasses. Glasses such as alkali lead silicate glasses with $\nu < 50$ are designated flint glass. The Schott nomenclature for optical glass is used in Figure 28. Labeled areas represent commercially available glass.

When n_D and ν are identical for glasses with regular dispersion properties, the rule holds that the remaining dispersions are also identical. For standard glasses, the complete refraction behavior can, therefore, be determined from these two variables.

Refractive index and dispersion depend on the structure and constituents of the glass and on their interaction with light. Like most other glass properties, partial molar refractive index coefficients have been calculated.

Measurement of refractive index, although generally requiring some grinding and polishing of the sample, can be carried out with precision and with reasonable speed; it is, therefore, used in quality control to determine the homogeneity of the glass. Dispersion and Abbe number are used mainly to characterize optical and ophthalmic glasses.

4.12.3. Temperature Dependence of Refractive Index

The density of glass and the UV absorption edge vary with temperature. Both effects influence the refractive index, causing in some glasses a negative, and in others a positive, temperature coefficient. Athermal glasses, which do not change their refractive index in specific temperature ranges, are available.

4.12.4. Stress Optical Coefficients

Glass becomes birefringent when strained. Stress and strain are proportional to each other in

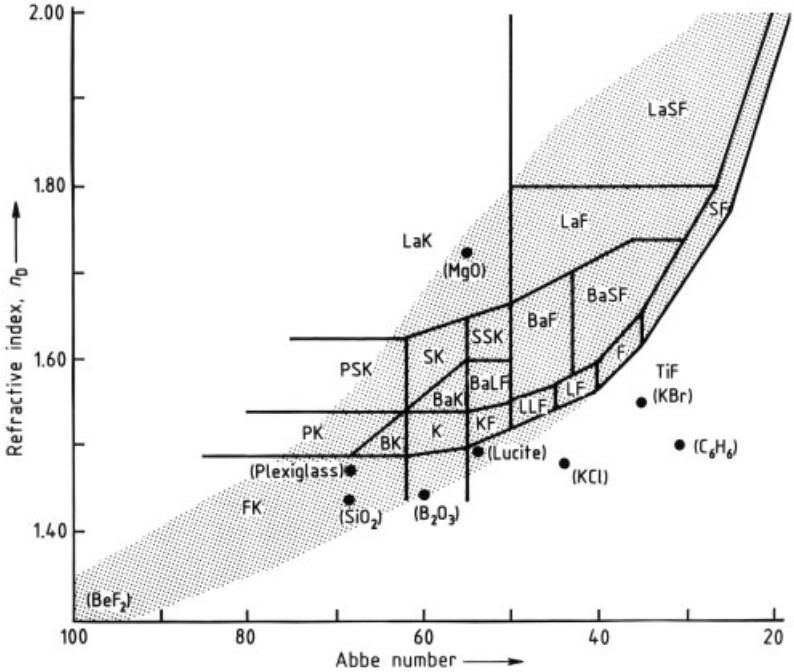

Figure 28. Refractive index vs. dispersion and optical classification of glasses
Shaded area indicates region of glass formation. Plexiglass is a methyl methacrylate resin, lucite is an acrylic resin. BaF = barium flint; BaK = barium crown; BaLF = light barium flint; BaSF = heavy barium flint; BK = borosilicate crown; F = flint; FK = fluorocrown; K = crown; KF = crown flint; LaF = lanthanum flint; LaSF = heavy lanthanum flint; LaK = lanthanum crown; LF = light flint; LLF = very light flint; PK = phosphate crown; PSK = heavy phosphate crown; SF = heavy flint; SK = heavy crown; SSK = very heavy crown; TiF = titanium flint

glass. The stress optical coefficient B, or Brewster's constant, expresses the relationship between optical retardation produced in a beam of plane-polarized light as it passes through a glass and the imbalance between principal stresses that exist along the path of this light beam. Brewster's constant is the ratio between retardation R (in nanometer path difference per centimeter path) and stress F (in kilograms per square millimeter)

$$B = R/F$$

The stress optical coefficients vary greatly with glass composition. The value of B ranges from ca. − 1 for extra-dense flint containing 80 % PbO to ca. 2.5 for soda lime glass, ca. 3.4 for 96 % silica glass, 7 for epoxy resin, and 2000 – 14 000 for gelatin.

4.12.5. Radiation Effects

The changes occurring in silicate glass under the action of ionizing radiation (e.g., high-energy electrons, gamma-rays, X-rays, or combinations of these) tend to be the same in both the nature of the excitation process involved and the degree of defect formation. This is true at saturation as long as the same amount of material absorbs the same amount of radiation energy (the energy approximation). Mass absorption coefficients for X-rays are, however, substantially higher than those for ^{137}Cs and ^{60}C gamma rays with energies of 0.66 and 1.17 MeV, respectively. Defect saturation will, therefore, proceed much faster with X than with gamma rays. Color changes induced by sunlight are commonly referred to as solarization.

The stability of glass in a flux of neutrons depends in particular on the amount of boron, which possesses a high absorption coefficient for neutrons. Boron-free glass can generally with-

stand a flux of 10^{19} neutrons/cm^2. Ion bombardment disrupts the glass structure, producing discoloration and free oxygen. The effective lifetime of electronic tubes can be affected considerably by reaction between free oxygen and the cathode.

Electron browning affects the performance of cathode-ray tubes containing oxides of lead or bismuth. These compounds are easily reduced by the incident electrons, producing a brown tint in the glass.

Relatively low doses of gamma rays tend to discolor dense lead silicate glasses. Among amorphous oxides, ultrapure vitreous silica is most resistant to discoloration. The drawbacks of vitreous silica are its low density and the expense of production. Alternatives are glasses with intermediate density to which oxides with variable cation valence are added: Ce_3O_4 (Ce^{3+}, Ce^{4+}) is commonly used as such an oxide because of its effectiveness in increasing color stability. Lead silicate glass combined with intermediate density, cerium-doped glass is used in composite windows with remarkable resistance to radiation damage. They are used, for instance, in environments with relatively high-intensity radiation fluxes, matching the absorption characteristics of heavy concrete used for shielding walls.

Cobalt-containing glasses have been developed for their radiation sensitivity. More sensitive silver phosphate glasses are used as dosimeters to measure exposure to X– or gamma rays.

4.13. Color of Glasses

White light is composed of a mixture of colors, the different colors representing a spectrum of frequencies and hence of energies. The relation between wavelength, frequency, and energy is given in Table 21. The human eye is sensitive to light of frequency $(3.9-7.5)\times 10^{14}$ Hz, corresponding to a wavelength of 400 – 760 nm and an energy of 1.6 – 3.1 eV or 157.4 – 299.1 kJ/mol.

Absorption of a particular component of white light by a transparent material causes the appearance of color. If the material absorbs in the red part of the spectrum, the color is blue; absorption in the yellow region yields a brown color. Thus, for it to have color, some light absorption must occur in the glass.

Limiting wavelengths (lower wavelength limit) of absorption for some oxide and fluoride glasses are given in Table 22. Clearly, in absence of impurities that absorb light in the visible region, none of these glasses shows coloration because no absorption extends beyond 400 nm.

Table 22. Limiting absorption wavelength for various oxide and fluoride glasses

Glass type	Absorption wavelength λT, nm [a]	Energy, eV
P_2O_5	<150 [b]	>8.3
SiO_2	162	7.7
B_2O_3	200	6.2
HPO_3	273	4.5
GeO_2	363	3.4
$P_2O_5-R_2O$ (R = alkali metal)	220	5.6
SiO_2-R_2O	270	4.6
$B_2O_3-R_2O$	360	3.4
BeF_2 [c]	90	13.8
ZBLA [c]	210	5.9

[a] Position of UV edge at 50 % transmission.
[b] Difficult to measure because of the ubiquitous presence of water.
[c] Data from [93], [122], [123].

Transmission curves for some simple uncolored glasses are shown in Figure 29. The UV absorption is of principal importance. This absorption arises from the energy absorbed by electrons excited from the valence to the conduction band as shown in Figure 30. The occupied states are the valence band and the empty states are the conduction band; the energy difference between the two is the band gap. More sophisticated analyses have been made [121].

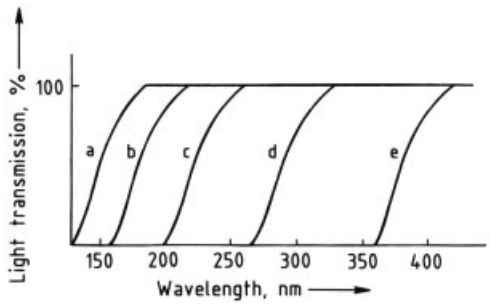

Figure 29. Transmission curves for simple uncolored glass a) P_2O_5 glass; b) SiO_2 glass; c) B_2O_3 glass; d) HPO_3 glass; e) GeO_2 glass

The widest band gap among the glasses in Table 22 is shown by BeF_2. For oxide glasses, the absorption shifts to longer wavelength in

Table 21. Conversion values for wavelength, energies, and frequencies of light ∗

Wave-number, cm^{-1}	Wave-length, nm	Energy eV	kJ/mol	kcal/mol	10^4 K	Frequency, 10^{14} Hz
400	25 000	0.050	4.824	1.143	576	0.1199
500	20 000	0.062	5.982	1.429	720	0.1499
1 000	10 000	0.124	11.964	2.859	1 439	0.2998
2 000	5 000	0.248	23.929	5.717	2 878	0.5997
3 000	3 333	0.372	35.893	8.576	4 317	0.8995
4 000	2 500	0.496	47.857	11.434	5 756	1.1990
5 000	2 000	0.620	59.822	14.293	7 195	1.4990
6 000	1 667	0.744	71.786	17.151	8 634	1.7990
7 000	1 429	0.868	83.750	20.010	10 073	2.0990
8 000	1 250	0.992	95.715	22.869	11 512	2.3990
9 000	1 111	1.116	107.679	25.727	12 951	2.6990
10 000	1 000	1.240	119.643	28.586	14 390	3.0000
11 000	909	1.364	131.608	31.444	15 829	3.2980
12 000	833	1.488	143.572	34.303	17 268	3.598
13 000	769	1.612	155.536	37.161	18 707	3.8980
14 000	**714**	**1.736**	**167.501**	**40.020**	**20 146**	**4.1980**
15 000	**667**	**1.860**	**179.465**	**42.879**	**21 585**	**4.4970**
16 000	**625**	**1.984**	**191.429**	**45.737**	**23 024**	**4.797**
17 000	**588**	**2.108**	**203.394**	**48.596**	**24 463**	**5.097**
18 000	**556**	**2.232**	**215.358**	**51.454**	**25 902**	**5.397**
19 000	**526**	**2.356**	**227.322**	**54.313**	**27 341**	**5.697**
20 000	**500**	**2.480**	**239.287**	**57.171**	**28 780**	**5.997**
25 000	**400**	**3.100**	**299.108**	**71.464**	**35 975**	**7.496**
30 000	333	3.720	358.930	85.757	43 170	8.995
35 000	286	4.340	418.752	100.050	50 365	10.494
40 000	250	4.960	478.574	114.343	57 560	10.871
45 000	222	5.580	538.395	128.636	64 755	13.492
50 000	200	6.120	590.498	141.084	71 022	14.798
55 000	182	6.820	658.039	157.222	79 145	16.495
60 000	167	7.440	717.860	171.514	86 340	17.990
65 000	154	8.060	777.682	185.870	93 535	19.489
70 000	143	8.680	837.504	200.100	100 730	20.988
75 000	133	9.299	897 229	214.370	107 914	22.485
80 000	125	9.919	957.051	228.663	115 109	23.984
85 000	118	10.539	1016.872	242.956	122 304	25.483
90 000	111	11.159	1076.694	257.248	129 499	26.982
95 000	105	11.779	1136.516	271.541	136 694	28.481
100 000	100	12.399	1196.337	285.834	143 889	29.980

∗Bold type indicates visible region of spectrum.

the increasing order P_2O_5, SiO_2, HPO_3, and GeO_2. A further shift to longer wavelength occurs when alkali-metal oxides are added to the pure network-forming glasses. The degree of shift can be correlated with cation field strength. Addition of Li_2O to a network-forming glass causes the smallest increase in wavelength; addition of Cs_2O, the largest.

Homogeneous glasses and crystals are colorless if the bandgap is sufficiently large. Thus, Al_2O_3, SiO_2, and alkali-metal or alkaline-earth oxides are colorless when pure. Coloring of glass occurs because of the presence of impurities, often *transition metals*, that absorb light in the visible region. The reasons for this absorption are the transition-metal *d* orbitals. The five *d* orbitals, which are degenerate in a free transition-metal atom, are split into two groups under the influence of the ligand electric field. The split in energy between the two sets of *d* orbitals is designated $10\,D_q$. Metal ions and ligand groups can be arranged in a *spectrochemical series*: the metal ions according to the order of increasing $10\,D_q$ developed with given ligand; the ligand groups according to the order of increasing $10\,D_q$ for any given metal ion.

The metal ions fall in the order $Mn^{2+} < Ni^{2+} < Co^{2+} < Fe^{2+} < V^{2+} < Fe^{3+} < Cr^{3+} < V^{3+} < Co^{3+} < Mn^{4+} < Mo^{3+} < Rh^{3+} < Ru^{3+} < Pd^{4+} < Ir^{3+} < Re^{4+} < Pt^{4+}$, and ligands in the order $I^- < F^- < H_2O < SCN^- < NH_3 < SO_3^{2-} < NO_2^- < CN^-$. Tri- and higher valent metal atoms tend to absorb in the visible region in oxide glasses. Electronic transitions between lower occupied and higher unoccupied states have an energy value of 2 – 3 eV.

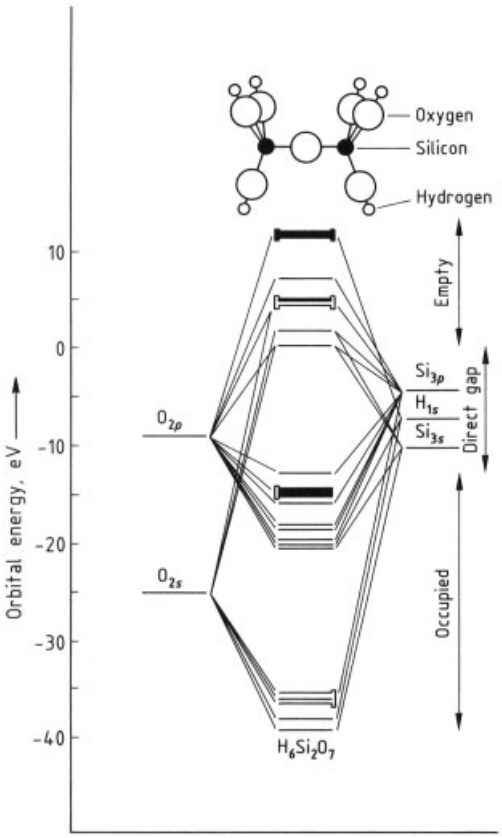

Figure 30. Calculated direct gap for the $H_6Si_2O_7$ molecule at a Si–O–Si angle of 180°
The direct gap represents the difference in energy between the highest occupied (HOMO) and lowest unoccupied molecular orbital (LUMO).

The degree of splitting of the metal ion *d* orbitals is affected by the coordination number of the metal, oxidation state of the metal, symmetry of the coordination sphere, type of ligand, and nature of the metal ion. Splitting values ($10\,D_q$) for complexes of the first transition series are 800 – 1333 nm for divalent ions and 400 – 714 nm for trivalent ions. Divalent ions, therefore, absorb light outside the visible region and tend to give only weak coloration arising from absorption tails that extend into the red part of the visible region.

Table 23. Transition-metal colors in glasses

Metal	Color
Titanium	no color in low concentration; reduction of Ti(IV) to Ti(III) in high concentration gives slightly violet color; brown in phosphate glasses
Vanadium	vanadium produces a variety of colors because of its variable valence: green in silica glass; colorless in alkali silicate glass; olive green in phosphate glass; brown in borate glass
Chromium	chromium(III) makes glasses green; oxidation leads to chromium(VI) which is yellow
Manganese	manganese(III) makes glasses violet; manganese(II) gives a faint yellow
Iron	iron(II) gives a green color to glasses molten in neutral or slightly reducing atmosphere; iron(III) gives yellow-brown glasses in combination with sulfide (S^{2-})
Cobalt	cobalt(III) gives intense blue color in silicate and phosphate glasses; pink in lowalkali borate glasses; green in glasses with very high alkali content
Nickel	nickel(III) gives olive green or brown color in silicate glasses; higher alkali oxides give yellow; lower content, a greenish brown; blue occurs in some instances
Copper	copper(II) gives blue in silicate glasses; reduction to copper(I) leads to colorless glasses; green in phosphate and borate glasses
Rare-earth metals	cerium gives yellow; praseodymium, weak green; neodymium, red to violet; erbium, pink

Producing a consistent, predictable color in a glass is not trivial, because the same transition metal may give rise to a variety of colors, depending on type of ligands and oxidation state of the metal. This color variability for some cobalt(III) complexes is shown below:

$[Co(NH_3)_6]^{3+}$	yellow
$[Co(NH_3)_5H_2O]^{3+}$	pink
$[Co(NH_3)_5Cl]^{2+}$	purple
$[Co(NH_3)_4Cl_2]^+$,	
cis	red violet
trans	green
$[Co(NH_3)_4(NO_2)_2]^+$,	
cis	yellow brown
trans	yellow red
Multinuclear Co complexes	black

The most common transition-metal oxides used in the glass industry and the expected colors are listed in Table 23. Concentrations are typically 0.5 – 5 wt %. Varying oxidation conditions

can yield results spectacularly different from those anticipated; for example, high-manganese welding fluxes tend to be black to dark purple; however, slightly reducing conditions make such fluxes yellowish green. The colors in Table 23 should, therefore, be regarded only as guidelines.

Striking glasses are the second largest family of colored glasses. Their color is caused by the controlled nucleation and growth of a crystal-like chalcogenide phase. These glasses form the basis of commercial yellow-to-red sharp-cut filters originally produced by Corning Glass Works. The composition of these filter glasses is given in Table 24.

Striking glasses are produced by casting and fairly rapid cooling, which results in a colorless glass. Rapid cooling causes a single burst of nuclei to form with similar dimensions, and reheating causes these nuclei to grow. Color develops (strikes) after heat treatment of such glass between 823 and 973 K. The absorption edge steepens during striking and moves toward longer wavelengths. By controlled reheating, isochemical glasses with different colors can be produced.

The most probable explanation for color generation in striking glass involves an electronic transition from the occupied valence band to the empty conduction band in semiconducting chalcogenide crystals in a silicate matrix. The striking effect is specifically caused by the variation in size of the gap as a function of crystallite size, hence, the name *quantum confinement* for this effect. Typical striking glass chromophores are CdS 2.42 eV (518 nm), CdSe 1.73 eV (717 nm), CdTe 1.44 eV (860 nm), and ZnS 3.53 eV (340 nm) [125]. Of these, only CdS and CdSe absorb in the visible region.

Striking glasses differ from those colored by transition-metal ions mainly by the nearly quantitative transmission of light from a specific wavelength to longer ones. The absorption and transmission regions are separated by a sharp edge, making pure saturated colors possible. These glasses make excellent cutoff filters for optical instruments, unlike filters based on ionic colors with their characteristically broad absorption bands.

5. Production of Glass

Glass-melting technology has passed through six stages:

1) Glass manufacture in open pits ca. 3000 B.C. until the invention of the blowpipe in ca. 250 B.C.
2) The mobile wood-fired melting pot furnace used until about the 17th century by travelling glass manufacturers.
3) The local pot furnace, fired by wood and coal (1600 – 1850).
4) The gas-heated melting pot and tank furnaces dating from 1860, followed by the electric furnace of 1910.
5) Tank furnaces with special designs for obtaining better glass quality (since 1950).
6) All-oxygen-fired glass-melting furnaces (since ca. 1985).

5.1. Raw Materials

Most glass requires heat to convert the raw materials such as minerals and synthetic inorganic chemical components to a homogeneous melt at fusion temperatures that turns to a rigid glass on cooling. Raw materials are selected according to purity, consistency, grain size, water content, availability, pollution potential, ease of mixing and melting, and cost. The most common ingredients in glass production are listed in Table 25.

Today, the most important raw materials are sand and recycled glass cullet. In Western Europe more than 7.5×10^6 t are recycled annually. The primary contaminants are ceramics, organic food residues, and paper. The metal contaminations are almost completely removed from the waste glass in modern post-consumer glass purification plants.

The primary contaminant in sands is Fe_2O_3, followed by minerals such as titania (rutile), zircon ($ZrSiO_4$), and chromia or iron chromites. Acceptable limits for Fe_2O_3 are 0.005 – 0.1 wt %, very much depending on the glass product. Acid-washed varieties of sand have the lowest iron concentrations. The price of sand reflects the quality: varying between \$ 15/t for common glass to \$ 800 – 1000/t for sand with

Table 24. Composition of Corning sharp-cut filter glass [124] ∗

Component			Composition, mol %		
SiO_2			73.5		
Na_2O			5.8		
B_2O_3			10.5		
Al_2O_3			0.6		
ZnO			8.9		
CdO			0.3		
Sb_2O_3			0.2		
Filter ∗∗	Cs 3-66	Cs 2-61	Cs 2-60	Cs 2-59	Cs 2-64
Se	0.06	0.11	0.11	0.13	0.19
S	0.09	0.07	0.05	0.06	0.04

∗ Base glass composition is determined by X-ray fluorescence, sulfur and selenium content by chemical analysis.
∗∗ The adsorption edges for the filters, as determined by photoluminescence, are as follows: Cs 3-66, 473; CS 2-61, 500; CS 2-60, 518; CS 2-59, 528; and CS 2-64, 546 nm.

Table 25. Raw materials used in glass production [a]

Raw material	Chemical composition	Glass-making oxide	Oxide, wt %
Sand	SiO_2 [b]	SiO_2	99.8
Soda ash	Na_2CO_3	Na_2O	58.5
Trona	$Na_2CO_3 \cdot NaHCO_3 \cdot 2\,H_2O$	Na_2O	41.9
Limestone	$CaCO_3$	CaO	56.0
Dolomite	$CaCO_3 \cdot MgCO_3$	CaO	30.5
		MgO	21.5
Feldspar	$(K,Na)_2O \cdot Al_2O_3 \cdot 6\,SiO_2$ [b]	SiO_2	68.0
		Al_2O_3	18.5
		$(K,Na)_2O$	12.8
Nepheline	$NaAlSiO_4$ [b]	SiO_2	60.6
Syenite		Al_2O_3	23.3
		$(K,Na)_2O$	14.8
Petalite	$LiAlSi_4O_{10}$	Li_2O	4.9
		Al_2O_3	16.7
		SiO_2	78.4
Borax	$Na_2B_4O_7 \cdot 5\,H_2O$	Na_2O	21.8
pentahydrate		B_2O_3	48.8
Boric acid	H_3BO_3	B_2O_3	56.3
Colemanite	$Ca_2B_6O_{11} \cdot 5\,H_2O$	CaO	27.2
		B_2O_3	51.0
Ulexite	$NaCaB_5O_9 \cdot 8\,H_2O$	Na_2O	9.5
		CaO	17.2
		B_2O_3	53.7
Litharge	PbO	PbO	99.9
Anhydrous potash	K_2CO_3	K_2O	68.0
Potassium nitrate	KNO_3	K_2O	46.2
Fluorspar	CaF_2	CaO	69.9
Zinc oxide	ZnO	ZnO	100.0
Barium carbonate	$BaCO_3$	BaO	76.9
Strontium carbonate	$SrCO_3$	SrO	70.1
Sodium antimonate	$NaSbO_3$	Na_2O	16
		Sb_2O_3	75.6
Blast furnace slags [c]	$x\,SiO_2 \cdot y\,CaO \cdot z\,Al_2O_3$ + sulfide		

[a] Sulfates, chlorides, and fluorides are only partly retained in the melt. Sodium, potassium, lead, antimony, arsenic, and boron oxides generally evaporate in small quantities from the melt; retention of these components in the glass is 90 – 99 %. Carbon dioxide and nitrogen oxides are almost completely released from the batch blanket or glass melt. [b] Refractory components. [c] Composition is source-dependent.

an iron content of only 15 – 20 mg/kg. For vitreous silica production, expensive high-quality sands are used.

The main sources of calcium are limestone and dolomite, which also supplies magnesium. About 5 % of the total amount of limestone quarried in the United States is consumed in glass manufacture. Limestone contains 56 wt % CaO, and dolomite 30.0 wt % CaO and 21.5 wt % MgO. High-quality limestone contains < 0.1 wt % Fe_2O_3 and ca. 1 wt % silica and alumina. A source of calcium carbonate for low-iron glasses (ca. 0.02 wt % iron) is aragonite, which is found in high purity in the Bahamas. A drawback of this carbonate is the high content of sulfates and organic material. Other raw materials that are sources of calcium include colemanite and fluorspar, often used in fiberglass production, and iron blast furnace slags, used for soda-lime glasses.

Trona, mined in Trona, Wyoming, is the primary source of soda in the United States. About 40 % of all soda ash produced in the United States is consumed by the glass industry. Synthetic soda produced from common salt (NaCl) is used in Europe and contains small amounts of residual sodium chloride (0.12 – 0.25 wt %). Caustic soda solutions are sometimes used in the wet batching process as a batch component and as a dusting suppressant for subsequent handling. When preparing agglomerated batch in the form of pellets, caustic soda is used both as a binder and as a sodium raw material. Worldwide about 50 % of the soda production is consumed by the glass industry.

Other raw materials for borosilicate glasses include borax and boric acid. Colemanite and ulexite are often preferred as sources of boron on economic grounds. For glass compositions with very low alkali concentrations such as E-glass, colemanite or boric acid is used to avoid an increased sodium content in the glass when using other boron raw materials.

Feldspar and nepheline syenite are common sources of alumina. In the container- and float-glass industries, purified iron blast furnace slags are used as an alumina-rich melting flux. These slags also contain sulfides which decrease the temperatures of primary fining (removal of bubbles, degassing). Litharge is used in the production of lead silicate and lead-crystal glasses. Powdered anthracite, corn starch, oil, raw potatoes, or sugar can be used as reducing agent. Typical oxidizing agents are nitrates such as potassium or sodium nitrate, cerium(IV)oxide, and sodium sulfates.

Most commercial glass compositions contain small amounts of other constituents which, in general, do not significantly affect the physical properties of the glass. *Fining agents* are added in small amounts to eliminate gas bubbles formed during melting. Common fining agents are sulfates, halides (e.g., NaCl), CeO_2, MnO_2, As_2O_5, and Sb_2O_5. These agents are thought to release oxygen, sulfur dioxide, or other gases, increasing the size of the gas bubbles and therefore their buoyancy in the melt. The melts dissolve gases such as carbon dioxide, nitrogen, water vapor, oxygen, and argon that originate from decomposition of the raw material or from the furnace atmosphere. Most of these gases can be removed from the melt by the growing and ascending bubbles during the fining process. Some fining agents, like sodium sulfate, cerium(IV) oxides, antimonates or arsenates also oxidize the glass melt and reduce the iron(II) concentration by oxidation reactions. Nitrates are often used to oxidize batch components such as As_2O_3 (to As_2O_5) and Sb_2O_3 (to Sb_2O_5) in order to make these components effective for primary fining.

Refractory batch components are materials with a high melting point which exceeds the maximum temperatures in the melting tank. These components do not simply melt but dissolve in the primary melts of the other raw materials. Alumina and sand grains, for example, react with and dissolve in sodium-, potassium-, and boron-rich melt phases. The dissolution rates depend on temperature, composition of the primary melts, and the size of the refractory batch grains.

Besides purity, the grain-size distribution of the refractory batch materials is an important factor in successful melting. Too coarse a fraction results in incomplete dissolution of the sand or alumina raw material in the primary melts, with attendant stones and knots in the glass; too fine a fraction results in dusting during handling and too rapid melting, which leads to very fine gaseous inclusions (seeds). The latter may also result in the formation of an insulating alumina or silica-rich scum on top of the glass melt. This scum, due to flotation of very fine silica or alumina grains and subsequent sintering on top of

the melt, can decrease the heat transfer into the melt and slow the rate of melting. In some critical cases, the scum may form silica- or alumina-rich glass defects or cords.

Table 26. Conversion table for micrometers, mesh, and inches

Micrometer	Inch	Mesh∗
4760	0.185	4
3360	0.131	6
2380	0.093	8
1680	0.065	12
1190	0.046	16
840	0.0328	20
590	0.0232	30
420	0.0164	40
297	0.0116	50
250	0.0097	60
210	0.0082	70
177	0.0069	80
149	0.0058	100
105	0.0041	140
74	0.0029	200
62	0.0024	230
53	0.0021	270
44	0.0017	325
37	0.0015	400
20	0.0008	625
10	0.0004	1250
5	0.0002	2500

∗ Tyler Standard Screens; note that five other standards are commonly employed: (1) British Standard Test Sieves, (2) Institute of Mining and Metallurgy, (3) the German DIN, (4) the French AFNOR, and (5) the U.S. Standard Sieve Series.

Nework modifiers increase the melting rate and generally decrease the viscosity of the melt. Their effect is to start the dissolution of refractory materials, and to decrease the viscosity and surface tension of the melt. They commonly have a low melting point and high vapor pressure. A low surface tension of the melt improves the wetting of the surface of the refractory batch grains by the aggressive primary melts. Some efficient melting fluxes apart from network modifiers are sodium sulfate, sodium chloride, and calcium fluoride (fluorspar), which have a low surface tension, cullet which needs no fusion energy, and slags containing sulfides and alumina in an easy-to-melt form.

Melting considerations tend to favor grain sizes of refractory batch components that are one to two screen sizes finer than those of commercially available fluxes. Container glass manufacturers generally use sand between 100 and 600 µm (140 – 30 mesh) as the best compromise between melting efficiency and the high cost of producing fine-grained sand. Fiberglass manufacturers, on the other hand, use sand with dimensions < 74 µm (200 mesh). The conversion between mesh size, micrometers, and inches is given in Table 26.

Batches with raw materials having different grain size distributions are sensitive to batch segregation. Most glass melting batches show this behavior during dry transport especially during vibration and free fall. Therefore, water or sometimes light oil is added to prevent demixing during transport and charging of the batch feed into the furnace. Wetting of the batch also reduces the entrainment of fine batch components in the combustion gases in the furnace. This entrainment or batch carryover can lead to batch losses but also to fouling of the flue-gas channels and equipment downstream of the flue-gas ducts.

5.1.1. Chemical Impurities and Decoloration

Pure minerals do not occur in nature. Contaminants such as transition metals, sulfates, halides, or organic matter are always present. Natural or upgraded ores are used, depending on the glass specifications. The most common impurity is iron, whose presence is usually undesirable for clear and flint glasses. Upper limits for this element in some glass products are listed below in percent by weight:

Optical waveguides	10^{-6}
UV transparent Schott glass 8337	0.004
Lead glass	0.018
Ophthalmic glass	0.020
Container flint glass	0.035
Container green glass	0.25 – 0.35
Container amber glass	0.30
Flat glass (clear)	0.07 – 0.1
Tinted flat glass	≤ 1.6
Tableware (clear)	0.02

Other undesirable impurities include primarily the colored oxides of transition metals (e.g., nickel, vanadium, copper, manganese, chromium, cerium, and cobalt). In addition to coloring agents, excess sulfate may cause sulfate gall formation on top of the melt and sodium sulfate inclusion in the glass. Too much sulfate may also lead to precipitation of SO_2 seeds in the glass. Sulfate-containing melts prepared under relatively strong reducing conditions may form SO_2 bubbles which, if not completely removed,

generate low-pressure blisters with sulfur and sulfate deposits in the final product [135]. The presence of chloride impurities in raw materials leads to chloride-containing melts, which may attack the forming molds. Ammonia, coke, sulfides, and organic materials affect the oxygen fugacity (equilibrium pressure of dissolved oxygen gas) of the melt, with attendant reduction of transition metals and subsequent color changes, e.g., reduction of Ti^{4+} (colorless) to Ti^{3+} (blue) or increased concentration of Fe^{2+} due to reduction of Fe^{3+}, or formation of S^{2-}, which forms an amber chromophor in combination with Fe^{3+}.

Impurities are not always undesirable. Small amounts of iron are sometimes necessary to develop a certain glass color. During melting, these impurities often influence heat radiation into the glass, and during forming, they affect the heat loss to the mold or environment. Iron(II) sulfide (FeS) and iron pyrites (FeS_2) produce amber-colored glass used in the container industry. The residual Fe^{3+} forms an iron(III) sulfide oxide complex with S^{2-} giving the glass the amber color. Green container glasses use sodium dichromate or iron chromates. Green and amber container glasses protect beverages, pharmaceuticals, and foods against UV radiation. Selenium, with iron and cobalt, yields a bronze color. In small quantites, selenium, added as elemental selenium, selenite, or selenate, often in combination with cobalt oxides, decolorizes glasses containing iron(II). Selenium plus cobalt oxide give the glass the complementary color of the iron(III) coloring. Cerium dioxide is used to increase UV absorption in optical or colorless soda-lime glass and protects against X-ray browning.

Mass-produced glasses tend to use cheap raw materials with fairly high concentrations of iron. Especially iron(II) gives the glass a green color. Physical or chemical decoloration of such glasses is therefore important. In chemical decoloration, the iron is oxidized to its weakly colored trivalent state by increasing the oxygen fugacity of the melt, usually by adding oxygen-producing compounds such as As_2O_5 (or arsenates), Sb_2O_5 (or antimonates), KNO_3, $NaNO_3$, Na_2SO_4, and CeO_2. Chemical decoloration is effective if the iron concentration does not exceed 0.02 – 0.1 wt %. Physical decoloration is used when the melt contains a higher concentration. Here, the addition of other coloring agents compensates for the yellow-green color of the glass. Common compensating agents are selenium and manganese dioxides with a purple to pink coloration, and the oxides of cobalt, nickel, or neodymium. A side effect of physical decoloration is a reduction in total light transmission.

The redox state, determined by the oxygen fugacity of the melt, determines the color of the glass, because dissolved oxygen has a strong influence on the redox state of polyvalent ions such as iron, chromium, and sulfur in the melt. The color of the glass very much depends on the oxidation state of these ions.

5.1.2. Waste Glass (Cullet)

Waste glass, known as cullet, is an almost universal addition to the batch melted in glass furnaces [136]. In Western Europe some countries, including Germany, the Netherlands, and Switzerland, recycle more than 80 % of the container glass brought onto the market. Recycled cullet saves the resources of other raw materials, requires up to 25 – 30 % less energy to melt than an isochemical batch of raw materials, and, provided it is not too fine, reduces dust during batching and carryover of dust from the batch blanket in the furnace. Optimized furnaces using large cullet ratios in the batches have lower specific nitrogen oxide and carbon dioxide emissions.

Typical amounts of cullet used in different glasses are listed below, in weight per cent:

Container glass	25 – 95
Flat glass	20 – 35
Light-bulb glass	40 – 70
Tableware	25 – 50
TV glass	20 – 45
Insulation glass wool	40 – 45
Commercial glass	50 – 70

In container and insulation-wool glass furnaces, large quantities of external cullet can be recycled. Flat glass, TV glass, tableware, and lighting-glass furnaces use their own cullet or cullet from well-identified sources.

Generally, about 10 – 30 % of the cullet comes from crushed recycled glass produced in the plant. Iron contamination introduced by the crushing process is removed magnetically. Cullet of unknown origin, often collected from bins

specially installed for household glass waste disposal, commonly contains impurities that may be incompatible with the desired quality of the manufactured article, giving unwanted color, stones, crystals, metallic inclusions, cords, or gas bubbles. Metallic impurities in the cullet form liquid metal melts that can react with the refractory and may also cause holes in the bottom of the melting unit (downward drilling). Recycled bottles constitute up to 80 % of the batch used in the container glass industry in some Western European and Asian/Pacific Rim countries. This amount is expected to increase substantially in other parts of the world with the increasing introduction of recycling laws in the United States, Eastern Europe, Canada, and probably South America.

The insulation glass wool industry uses flat glass cullet as one of the main raw materials.

Another waste material added as a raw material to glass melting batches is filter dust separated from the glass furnace flue gases by electrostatic precipitators or baghouse filters. This filter dust often contains sulfates, sodium, and lime, and it can replace part of the sodium sulfate used as a fining agent in the batch.

5.1.3. Preparation of Raw Materials

Accuracy and precision in weighing the various batch constituents are important, because of the required consistency in glass composition. Thorough mixing of ingredients and prevention of subsequent segregation is essential to obtain a homogeneous glass melt. Batch handling systems vary widely throughout the industry, from manual for small production sites to fully automatic (→ Solids Handling). Some glass factories purchase their batches in granulated form (granule size 8 – 20 mm). In large-volume production, each batch material is stored in its own bin, and the correct computer-controlled amount is weighed from the bin onto a conveyor belt that feeds directly into an intensive mixer.

The efficiency of the melting operation and the uniformity and quality of the glass are often determined by the conveyance of the raw materials and the quality of mixing. Mixing of small amounts of specific raw materials, such as fining agents, coloring/decoloring compounds or cokes, to the bulk of the batch is carried out just before delivery of the batch to the furnace to prevent segregation. Different types of batches use different mixers. For instance, high-silica glasses such as soda-lime, borosilicate, and aluminosilicate are mixed in pan-type mixers with vertical stirrers. The batch is first dry-blended, followed by addition of a small amount of liquid (up to 3 – 4 wt % water or oil) for wet blending. The whole operation takes 3 – 8 min; longer treatment may result in segregation. A typical batch size may be a few hundred up to 4000 kg of material.

Many speciality glasses, especially those with high-density components such as PbO, were once mixed either in muller-type mixers or in horizontally rotating drums, but these mixers lead to segregation due to free-fall movement. High-intensity mixers with vertical stirrers are now commonly used. Wet mixing and batch agglomeration-pelletizing, briquetting or compaction are becoming popular for several reasons, one of which is the avoidance of dust formation, especially during preparation and transport of raw materials containing hazardous components like lead and arsenic-containing batches. Other reasons are prevention of carry-over in the furnace, absence of segregation, and hence better homogeneity, which improves glass quality. Radiative heat penetrates pelletized material faster than batch blankets containing normal raw materials, the single grains in the pellets are bound to the other grains, increasing reactivity during the fusion processes. The use of pellets lowers the energy demand of the glass furnace or allows higher melting loads in a glass furnace.

Cullet is generally added to the mixed batch just before mixer discharge. Sometimes the cullet or the cullet plus normal batch is preheated by raw material preheaters using the flue gas heat contents.

5.2. Melting Units

Most glass is produced by continuous melting processes. Only glasses with special characteristics are manufactured in batch-type processes, which allow greater compositional flexibility. Small production sites manufacturing lead crystal or art glasses often use batch-type processes such as pot furnaces. The function of glass furnaces is the conversion of raw materials into a

homogeneous melt, free of inclusions. In the furnace chemical reactions between batch components leading to fusion and melting takes place. Complete melting is necessary to produce a glass without batch stones. The dissolved gases and gas bubbles should be removed from the melt and the melt homogenized, because small difference in chemical composition cause variations in refractive index or lead to thermal stresses in the final products.

5.2.1. Flat Glass and Container Glass Tanks

Volume production of technical glass is carried out exclusively in rectangular, relatively shallow glass tanks, capable of holding up to 1200 t of melt. The tank functions as a rectangular melting crucible, directly heated from the top by flames or in the melt by electrodes. Convection in the melt mixes the heated surface of the glass melt with the rest of the melt.

The tank is covered at the top by a vault (arch) of silica bricks, known as the cupola or crown, supported by sidewalls which are constructed from different materials: refractories for high temperatures at the inside and insulation bricks or mats at the outside. The total thickness can be 600 – 800 mm. The crowns are constructed mainly from silica or mullite ($Al_6Si_2O_{13}$), but in some cases fusion-cast refractories are used. For the inner layers of the sidewalls, fused-cast alumina – zirconia – silica (AZS) refractory materials and, for some furnaces, chromium oxide refractories are used. The outer layers provide thermal insulation to the furnace.

The melting tank is divided into two sections by a permanent bridge wall or by floating refractory baffles. The larger section into which the fill (i.e., batch plus cullet) is charged by means of the batch house – a small extension (called doghouse) of the furnace in the backwall – is called the *melting end*. The batch plus cullet floats on top of the melt as a blanket or as batch islands. The batch blanket is continuously fed by new raw materials and extends several meters into the melting end. The other separated section of the tank, in which the melt starts to cool and to condition toward its uniform working viscosity, is called the *working* or *refining end* (refiner). Partitioning of the tank into two parts allows continuous production of glass. The working and melting ends are connected by a neck (large float glass furnaces) or a throat. Most container-glass and tableware furnaces are designed with a throat, a covered channel through which the melt flows from the melting end into the bottom section of the working end. In furnaces with a neck, the glass melt flows into the working end without a contact with refractories at the surface. Some furnaces with a direct connection between melting end and refiner use a skimmer to prevent the flow of surface glass that differs from the bulk glass or to prevent the surface scum to flow into the refiner tank.

The melting tank depth is generally 50 – 150 cm, depending on the transmission of thermal radiation from the flames in the combustion chamber on top of the tank to the molten glass. The melting tank design depends also on the glass composition, melting rate, type of fuel and oxidant (air or oxygen), and production rate. Tank parts that come in contact with the melt are usually made of highly refractory fusion-cast AZS or chromium oxide blocks. A typical Siemens-type side port fired continuous tank is shown in Figure 31.

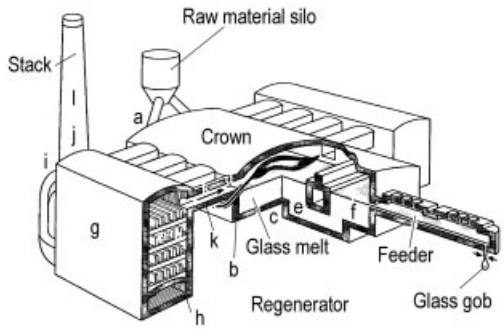

Figure 31. Fossil fuel fired Siemens-type regenerative furnace for container glass production
a) Doghouse; b) Fuel-injection lance; c) Melting end; d) Burner block; e) Throat; f) Working end; g) Regenerator; h) Checkerwork (an open structure of bricks); i) Air/fuel reversal mechanism; j) Chimney damper; k) Air inlet; l) Waste gas exit.

The conventional side-port regenerative furnace shown in Figure 31 is equipped with two regenerator sections on each side. Each regenerator is connected to the furnace by four to seven ports. During a period of 20 – 35 min ambient air is conducted to the left-hand regenerator for preheating. The hot air (1100 – 1300 °C) flows through the ports into the combustion chamber

of the furnace. In the combustion chamber air is mixed slowly with fuel oil or natural gas injected by one to three lances per port. The mixture produces a flame with a maximum temperature of about 2000 °C in this furnace type.

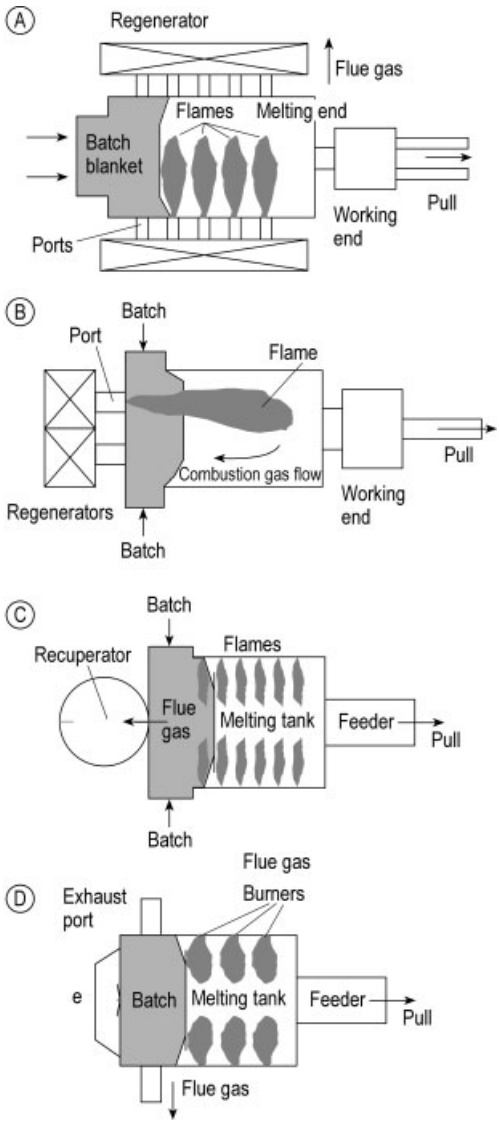

Figure 32. Top-view schematics of fossil-fuel-fired tank furnaces
A) Cross-fired regenerative furnace; B) End-port furnace with regenerators at the backwall; C) Recuperative furnace with a flue gas/air heat exchanger; D) All-oxygen-fired furnace (without preheating of oxygen)

After 8 – 12 s, the gases exit the furnace and heat the opposite regenerator chambers. After 20 – 35 min, the air and fuel flows are switched, and the preheated regenerator on the right-hand side is used to heat the combustion air. Thus air, flames, and flue gas flows are reversed every 20 – 35 min. By this method about 45 – 50 % of the energy content of the fuel is transferred to the melt for heating and fusion reactions. In direct-fired glass melters without air preheating, the combustion air energy efficiency is much lower.

Schematics of other furnaces in top view are shown in Figure 32.

In the recuperative furnace the flue gases preheat the air in heat exchanger type recuperators; the flue gases are cooled from about 1250 – 1300 to 800 – 900 °C, and air is preheated to about 750 °C. The furnace is often constructed as a unit tank (simple rectangular tank shape) furnace. Similar furnaces but without recuperators and special burner designs are used for all-oxygen-fired tanks, which have been replacing some of the larger air-fired regenerative and recuperative furnaces since 1991.

Some glass melting tanks are provided with electrodes to boost the energy supply to the melt. Forced bubbling improves mixing and heat transfer into the melt. Short cut melt flow along the bottom of the tank prior to its exiting through the throat or neck can be suppressed by a weir or dam, the refractory material of which must withstand high temperatures and glass-melt velocities. Molybdenum-reinforced AZS weirs and chromium oxide refractory walls are sometimes used.

Some tanks are equipped with a section called a deep refiner, which is much deeper than the first part of the melter. The deep refiner allows the bubbles to be more efficiently removed from the melt. The deep refiner is connected to the working end by a throat. In most cases the first section of the melting tank and the deep refiner are separated by a flow barrier. Special furnace designs are often patented by specialized glass furnace construction companies.

Optimum tank design is still being studied. Simulation models are based on three-dimensional finite-element numerical procedures (Computational Fluid Dynamic codes). Numerical models are based on conservation of mass, energy, momentum, and electric charge or potential. The models predict heat transfer, melting rates, temperatures and glass melt flow patterns in the tank. Some models also include

the combustion process and heat transfer in the combustion chamber. The results of such calculations for the temperature distribution in a rectangular glass tank with a combustion chamber on top are shown in Figure 33. A review of the status of computational techniques applied to tank design and a comparison of different models is given in [137].

Important in glass furnace design is a melting tank with extra means to improve heat transfer and which shows a good melting, fining, and homogenization performance. Short-cut flow patterns along the relatively cold bottom should be avoided, and separation of a first loop and a second convection loop in the melting tank may be essential for obtaining a sufficiently degassed and homogeneous glass melt. Numerical glass-tank models can be used to find the optimum furnace design or process conditions with respect to stability, glass quality, energy efficiency, and emissions [138].

Producers of flat glass use the largest tanks with surface areas of 400 – 800 m^2. The fill is distributed evenly over the full width of the tank by auger screws in barrels through multiple openings in the back wall, by pusher bars, or by vibratory plates. The last two methods often act in concert to form a blanket fill across the rear of the melting-end surface. A refractory extension of the melting end, called the *doghouse*, is frequently used to fuse the fill before it moves into the melting end itself. Container-glass tanks are smaller, with melting end surface areas up to 140 m^2, and often no blanket fill is used. Most container-glass furnaces are constructed as end-port-fired regenerative furnaces, and one or two screw feeders in small extensions of the sidewall(s) supply the batch to the top of the melt. The batch generally floats on top of the melt as small islands that together form a caterpillar-shaped batch line.

In sufficiently wide cross-fired tanks, cross-firing ports in the sidewalls are connected to regenerator chambers. In smaller tanks, the ports are usually in the backwall; these types of furnaces are called end-port-fired or U-flame furnaces. Regenerator chambers, which operate in pairs, are used to economize on fossil fuel by extracting heat from the flue gas before it is emitted through the stack. In U-flame furnaces, the strongly elongated flames make a U-turn, and the exhaust port and air inlet port are both in the backwall.

Temperatures up to 1400 – 1600 °C, may be required to melt the fill and to remove the gas bubbles from the melt, depending on the type and quality requirements of the glass. The temperature limit of a tank is determined by the quality of the superstructure material and the melt-line attack (i.e., increased wear of the refractory at the contact refractory – melt – furnace atmosphere). The weakest points that determine the lifetime of a glass-melting furnace are the throat, the refractories at the melt-line positions, and sometimes the crown and the regenerators. Carbon dioxide from carbonate-containing batch ingredients together with water and air trapped in the interstices of the fill, produces an initially foamy melt. Temperature profiles along the length of the tank are controlled to ensure that all batch material is dissolved in the fresh melt and that fining agents remove gaseous inclusions before the glass passes into the refining end of the tank. In some furnaces the surface of the melt (containing scum or foam) in the melting end is skimmed off by floating baffles. In most tableware, container-glass, and lighting-glass furnaces, the melt exits the tank or refining end through a throat in the solid bridge wall.

The principal working characteristics of glass melting tanks are melt efficiency and energy consumption, which depend on the glass type. For float glass a specific load of 1.5 $t\,m^{-2}\,d^{-1}$ of melt and 6500 kJ/kg of glass melt are typical; for container glass 2.5 – 3.5 $t\,m^{-2}\,d^{-1}$ of glass can be melted with a specific energy consumption, depending on cullet ratio in the batch, of 3600 – 5.000 kJ/kg. Higher energy requirements (8000 kJ/kg) and lower specific melting loads ($< 1.5\,t\,m^{-2}\,d^{-1}$) have to be accepted for TV panel glass because of the required high quality.

Enlarging and locally deepening the tank increases the allowable throughput, especially for colorless glasses. Additional heating by electrodes is required for deeper tanks. Applying an a.c. voltage between molybdenum or tin oxide electrodes in the melt generates heat in proportion to the square of the electric current and the glass melt resistance. Bubbling air or steam helps to achieve an even bath temperature and the extra mixing provides a better homogeneity.

The life span of a tank depends on the quality of the refractory bricks. Modern container-glass

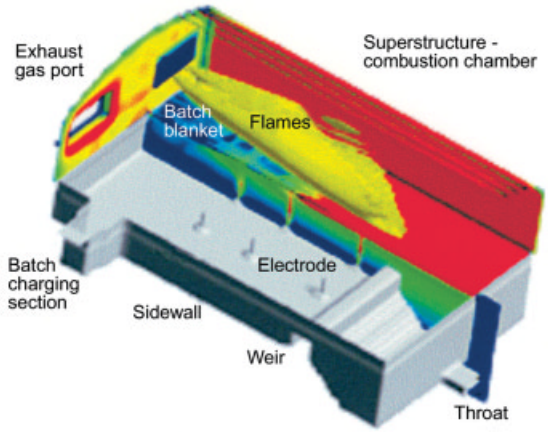

Figure 33. Temperature contours calculated for a end-port-fired regenerative container glass furnace. The surface temperatures and cross-section temperatures along the length axis are presented. Red: $> 1450\,^{\circ}\mathrm{C}$, blue: $< 1350\,^{\circ}\mathrm{C}$. The initial part of the flames calculated by the simulation model is also shown.

tanks are designed to be used for six to eight years before rebuilding is necessary. Float-glass furnaces producing the melt to be shaped as flat glass reach lifetimes up to 13 years before rebuilding.

5.2.2. Unit Melters

Recuperatively fired unit melters are used for continuous production of high-quality optical or ophthalmic glasses, soda-lime glasses for drinking bottles, and glass fibers or opal glasses. Often, typical tank dimensions are small: 2 – 5 m wide, 10 – 12 m long, containing 4 – 100 t of melt. Larger melters are generally designed as regeneratively fired furnaces or as oxy-fuel fired melters, but recuperatively fired unit melters with production rates of up to 150 t/d exist.

Some unit melters are directly fired without any air preheating, but they have a very poor energy efficiency and very low melting loads. Today, directly fired furnaces are only used in developing countries. In recuperatively fired and directly fired unit melters the combustion air and fuel are premixed in small burners and the flames cover the glass-melt surface. Firing temperatures of at least 1600 °C can be reached by positioning the burners opposite each another or in staggered arrangement. The specific melting loads barely exceed values of $1.5\,\mathrm{t\,m^{-2}\,d^{-1}}$. Low heating efficiency and higher operating costs offset the lower capital investments of unit melters compared with container- and flat-glass tanks. Unit melters are economical only in processes that require fairly small quantities of glass or rapid changes in glass composition.

5.2.3. All-Oxygen-Fired Melters

Improvements in the technology for separating oxygen from air by cryogenic or selective gas absorption techniques in the 1980s and 1990s and strict emission limits for nitrogen oxides justified the use of oxygen instead of air for the combustion of fossil fuels in glass furnaces in numerous glass-melting operations in the United States, Central America, South America, Korea, Japan, and Europe. Instead of preheated air, pure or almost pure oxygen is used as oxidant for fuel combustion. The furnace atmosphere is amost free of nitrogen, whereas air-fired furnaces have nitrogen concentrations of about 65 vol %. The volume flows of flue gases are only 15 – 25 % of those of similar air-fired furnaces. These much lower volume flows of hot flue gases result in much smaller energy losses than in air-fired furnaces. Due to the low volume flows and the risks of preheating oxygen, the flue gas heat is not recovered. Only in a few cases the flue gases are used to preheat part of the batch.

Most oxygen-fired melters are constructed similarly to unit melters. Generally, three to five burner blocks are installed in each sidewall. Typical oxygen burners have separate nozzles for

oxygen and for atomized fuel oil or natural gas. The luminescence of the flames can be improved by delayed oxygen – fuel mixing, which creates soot particles. The furnace atmosphere contains high concentrations of volatilized components from the batch blanket and glass-melt surface. Precautions must be taken to protect the combustion chamber refractories from the aggressive vapors.

Oxygen-fired furnaces show low energy consumption, although the energy requirement for the oxygen production also must be taken into account. The nitrogen oxide emissions and particulate emissions are generally lower compared to air-fired furnaces [139]. Specific melting loads of up to 2.5 t m^{-2} d^{-1} for oxygen-fired container-glass furnaces are feasible.

5.2.4. Electrical Melting Units

In some countries, all-electric melting in glass furnaces has become a leading objective for economic and environmental reasons. The method is based on the electrical conductivity (Joule heating) of a melt. Electrical conductivity increases as a function of temperature. The conductivity is caused by the presence of mobile alkali and alkaline earth metal oxides in the glass. Low-frequency (50 – 60 Hz) alternating current is used to minimize energy dissipation resulting from the activation energy required for electrical conduction.

The all-electric vertical melter is charged from the top, and the fresh batch layer covers the complete melt. Because of the relatively cold batch on top, these melters are also called cold-top electric melters. The crown or roof of the furnace is also relatively cold. Most vapors from the hot melt are entrapped in the batch layer, unless part of the melt is not covered, and therefore emissions of these melters are very low compared to fossil-fuel-fired melters with uncovered glass melts. For example, in an air-fired furnace up to 40 % of the fluorine evaporates and is lost, compared with only 2 % in an electric melter. Especially for glass or batch compositions containing volatile or environmentally hazardous volatile components an all-electric cold-top furnace offers an environmentally sound alternative. Production is usually limited to small charges of 4 t/d, although larger units with 120-t/d capacity also exist.

Molybdenum, tin oxide, sometimes graphite (under reducing conditions), or platinum electrodes extend from the bottom or well into the melt. A cold-top furnace is shown in Figure 34.

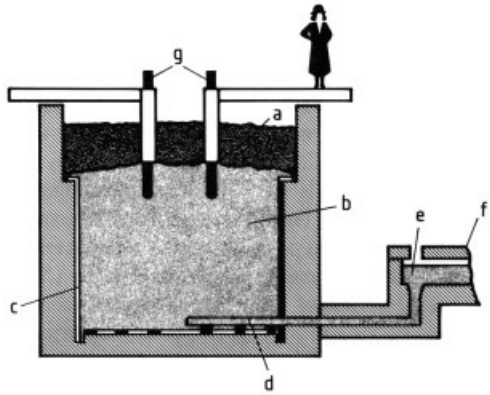

Figure 34. Cold-top electric glass melting tank
a) Blanket of batch; b) Melt; c) Molybdenum liner (not always necessary); d) Molybdenum pipe or refractory canal; e) Riser; f) Forehearth/working end; g) Electrodes

Electrodes can extend through the bottom as vertical rods, horizontally through the sidewalls, or can dip into the melt through the batch blanket. During the heating phase, burners are used to bring the cullet or batch in the furnace up to temperatures at which the electrical resistivity of the melt becomes low enough to conduct electric current. The burners are then removed, and the electrodes provide all the energy required for melting, fining, and homogenization. Controlled convection of the melt is important to obtain a homogeneous melt and to avoid short-cut flows.

An additional advantage of electric heating is the control of glass composition, especially when using volatile components, which is particularly useful in the production of certain optical glasses. Controlling the reaction between the batch blanket and molten glass in the reaction zone is of special importance. First a melt of the low-melting components is formed, and the refractory batch components dissolve in these primary melts. Just after melting, in the hottest zones, the fining process occurs. Bubbles are released from the melt and escape through the pores in the batch blanket. In cold-top electric melters gases released from the melt must find a way through the fill. Because of this and the

possibility of fresh cold dense melt flowing directly to the bottom of the tank, the fining process can sometimes become critical and unstable. The molten glass flows from the melting end through a throat or canal into a riser and into the working end.

Production of optical glass represents the epitome of criteria for glass homogeneity and freedom from flaws. Platinum crucibles are used to produce blister-free glass at a temperature up to 1550 – 1600 °C. The platinum-covered vaults of the melting section are usually made of zircon or silica bricks, depending on the type of glass being produced. Older burner systems have been replaced by environmentally safer platinum, tin oxide, or molybdenum electrodes.

Electrical melting has some disadvantages, including the higher energy cost, which is only partly offset by the lower amount of energy required (2500 – 4000 kJ per kilogram of glass), and the more rapid decrepitation of the refractory material compared to conventional units. Hence, the working life of a typical cold-top electric melter is only about 50 % of that of the lifetime of regeneratively fired glass melting tank furnaces. The advantages, such as small units, low capital investment, flexibility in types of batch materials, small vaporization loss of expensive or environmentally hazardous components, no nitrogen oxide emissions in case of raw material batches without nitrates, and particularly environmental compatibility, make resistance heating by electrodes sometimes an attractive proposition.

5.2.5. New Melter Designs

An overview of alternative melting technologies is given in [140]. Most alternative melting methods have separate sections for melting, degassing/fining, and conditioning or homogenization. The melting zone can even be divided into a batch preheating area, a primary melting stage, and a final melting and firing stage.

Some new concepts use subatmospheric-pressure fining to enhance bubble and gas removal from glass melts, and some melters use centrifugal forces to separate the gaseous inclusions from the molten glass.

A combination of batch preheating, firing at low temperatures to reduce NO_x formation, and separate zones for melting, fining, and refining and conditioning is presented by the LoNOx Melter developed in the mid-1980s in Germany (Fig. 35) [141].

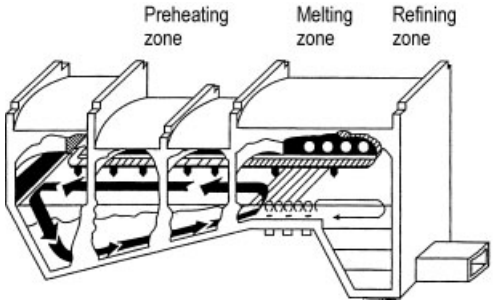

Figure 35. Design of the Sorg LoNOx melter for container-glass melting

5.2.6. Pot Furnaces and Day Tanks

Pot furnaces are scaled-up research furnaces. The name derives from the use of monolithic high-quality clay pots to contain the molten glass. Generally, the freshly prepared pots are tempered slowly at 900 °C by electric heating, prefired at 1200 °C, and fired to obtain a dense low-porous material at about 1400 °C in the pot furnace itself. Most pots are made of aluminum silicate materials. Open pots containing 150 – 1000 kg of melt are still used in Europe to manufacture colored and optical glasses as well as crystal glass. Typical areas in Europe using pot furnaces for art glass or crystal glass are the Venice area (Murano island) in Italy, the South of Sweden, the Czech Republic, the Bavarian forest in Germany, and some areas in Portugal, France, and in the vicinity of Birmingham in the United Kingdom. Closed pots surrounded by refractory brick walls were the preferred setup in the United States. The largest pot furnaces contain eight to ten pots, each with a diameter of about 0.7 – 1 m. Today, pot furnaces are hardly used in the United States because of their low energy efficiency and operational fluctuations, which can lead to variations in glass quality. Only for small production sites manufacturing handmade glass objects, pot furnaces are still applied.

The type of batch and glass specification determine the manner in which the pots are used. A typical melt cycle for crystal glass production in a pot furnace takes about 17 h and entails the

following: 6 h to melt the first fill, 2.5 h melting after addition of a second batch, 7 h melting and fining after adding a third batch at the highest temperature level close to 1350 – 1400 °C, and finally a lower temperature stage for 1.5 h to refine and condition the melt. The working temperature at which the glass melt is collected from the pot for forming processes depends on the glass composition and is in the range of 1100 – 1200 °C.

Production of specialty, high-value products (e.g., specially colored glass) or enamel glass frits is often carried out in day tanks. These tanks, usually built from refractory bricks, are typically heated rapidly by one to three pairs of oil or gas burners and produce 1 – 5 t/d of glass. The glass melt depth is only 10 – 20 cm. These day tank furnaces are particularly well suited for the manufacture of soft (i.e., low-viscosity) glass, such as crystal or soda-lime glass, which requires special production techniques. They are not as well suited for the manufacture of low thermal expansion borosilicate glasses because of the high boron loss during rapid heating. Such glasses can be produced if less rapid heating is used, with an increase in the average melting and fining time to 48 h. Even then, silica-rich stones, crystals, or cords, resulting from loss of sodium or boron at the surface of the melt, may still be found in the glass.

5.3. Melting, Fining, and Homogenization

The three steps involved in the production of a glass melt are melting, fining, and homogenization. The cold batch is charged to the tank or pot and melted at 1200 – 1650 °C, depending on the type of glass. Decarbonation and dehydration are the first chemical processes to take place in the batch, followed by partial melting and sintering as the eutectic temperature of some batch components is reached. Several new crystalline phases can be formed, and depending on the temperature ranges, can melt or dissolve in eutectic melts. Some of the sulfate, almost all carbonate, and water are removed from the melt. The primary melts containing the alkali metal oxides (sodium and potassium, sometimes lithium), alkaline earth metal oxides (mainly calcium and magnesium), or boron components act as fluxes that dissolve the refractory batch components. Gas solubility in the melt decreases with increased dissolution of the refractory components, leading to bubble trails. The bubbles contain CO_2, N_2, O_2, SO_2, Ar, water vapor, or combinations thereof.

Dissolution of refractory batch components, the most important being coarse sand grains and alumina raw material, slows down as the melt becomes nearly saturated, and a viscous melt with only a small fraction of nonmolten material appears at ca. 1150 – 1250 °C. The remaining refractory batch components will dissolve at further temperature increase. The batch slowly becomes transparent but remains chemically inhomogeneous. Blisters (bubbles) and cords (inhomogeneities in the melt) are particularly prominent at this point. Removal of dissolved gases and bubbles and finally resorption of residual seeds (small bubbles < 0.2 mm) characterize the next stage. The rate of melting depends on batch composition, particle-size distribution, temperature, and amount of cullet used.

The removal rate of gases from the melt depends on the kind and amount of active fining agent added to the batch and the redox state of the melt, which often determines the temperature of fining-gas release and the quantity of gas produced by the fining agent. These gases will diffuse into small seeds, which grow due to this gas infiltration. The bubble ascension rate in the melt is proportional to the square of the bubble diameter. The dilution of the gases in the bubble with fining gases cause not only the fining gases but also the other dissolved gases in the melt to diffuse into the growing bubbles. Fining gases include sulfur dioxide, oxygen, and sodium chloride vapor. During cooling of the melt the residual fining gases can be resorbed by the melt, since their solubility increases with decreasing temperature.

Blisters, seeds, and cords, already detectable by refractive index variations of around 5×10^{-4}, are removed in the fining and homogenization stage. Blisters and seeds are ≥ 0.2 and 0.2 – 0.05 mm in diameter, respectively. The use of fining agents is reviewed in [126]. Considerable attention is required in glass manufacture to make the melt blister-free by both expulsion of gas bubbles (fining) and subsequent resorption of residual small seeds during controlled cooling of the melt (refining). Blisters can arise

from several sources, including the type of raw material used, the tank atmosphere, reactions in the boundary layer between melt and refractory lining, electrochemical reactions in electrically heated furnaces, electrochemical reactions in refractories containing polyvalent ions such as iron and titanium, air seeping through the tank structure, or mechanical operations (e.g., stirring or gobbing). Methods for fining are:

- Addition of chemical fining agents to the batch
- Raising the tank temperature, with attendant decrease of melt viscosity and increase of bubble expulsion
- Lowering the temperature in the working end and thus increasing gas resorption

Homogenization is carried out by:

- Blowing air or water vapor into the bottom of the tank through bubbling pipes.
- Mechanical stirring or acoustic homogenization.
- Applying a certain temperature profile in the combustion chamber along the length of the furnace. A hot spot in the central part causes free convection flows in the melt and thus improves mixing and heat transfer. A well-developed spring zone in the central part of the tank may prevent short-cut glass-melt flows along the relatively cold melting tank bottom.

Sodium sulfate and arsenic and antimony oxides are generally used as fining agents. Only the pentavalent arsenic and antimony compounds produce oxygen as a fining gas at high temperatures. Therefore arsenates or antimonates are used in the batch, or nitrates are added to oxidize the trivalent arsenic or antimony compounds during melting. Nitrates produce oxygen between 500 and 900 °C [142]. Antimony oxide, in combination with arsenic trioxide and nitrates, is used in the production of optical glasses. Other, less important fining agents include fluorides and, more rarely, chlorides and bromides. Typically, ca. 0.3 – 0.6 wt % arsenic oxide or arsenates is added to the batch, although for some glass types up to 1.5 wt % arsenic trioxide may be employed. Antimony fining is used for TV panel glass compositions and for sulfate-free soda-lime glasses. Arsenic trioxide plus nitrates or arsenates is a widely used fining agent for lead-containing glasses that are melted under strongly oxidized conditions to avoid formation of lead metal.

Above about 1350 °C As_2O_5 in the melt decomposes into As_2O_3 and oxygen, which is the fining gas. During cooling below about 1300 °C the reaction is reversed. The fining mechanism with antimonates is similar, but the pentavalent antimony species dissolved in the melt decompose at lower temperatures than As_2O_5, and oxygen is released above about 1250 – 1300 °C, depending on the redox state (oxygen fugacity) of the melt. An additional advantage of arsenic or antimony is that oxygen gas bubbles which do not combine with ambient gas bubbles in the melt are resorbed at lower temperature because of the back reaction from arsenic trioxide to arsenates or antimony trioxide to antimonates.

The low price of sodium sulfate makes it attractive as a fining agent in commodity glasses. It is, however, soluble only to ca. 1 wt % in most melts and may lead to foaming or even sodium sulfate droplets (sulfate gall) in the glass product if used in excess. Sodium sulfate in oxiding melts decomposes at above 1350 – 1450 °C to give SO_2 and O_2 as fining gases. During cooling, bubbles containing both oxygen and sulfur dioxide can be resorbed. In reduced molten glasses, sulfates react with reducing agents to give SO_2 gas at temperatures just above 1000 – 1100 °C. This SO_2 gas is the major fining gas in these glass melts. Resorption of SO_2 bubbles in the absence of oxygen is limited, and only in strongly reduced melts can SO_2 or S_2 bubbles be resorbed due to formation of S^{2-} in the melt. The oxygen fugacity of the melt determines the temperature at which sulfates start to decompose, and the amount and kind of released gases also depend on the redox state of the batch and melt.

5.4. Energy-Saving Technologies

The melting process requires relatively large quantities of energy. The costs of energy contribute 12 – 15 % to the total glass production costs, depending on the country and type of glass. Only about 25 – 55 % of the energy added to the furnaces is used for heating the raw material and for the endothermic reactions involved in fusion of the raw materials into a melt. The most important energy savings to be achieved

in industrial glass melting and glass production focus on:

1) Improvement of the thermal insulation of glass furnaces by using high-quality refractory materials for the glass-melting tank and superstructure, allowing higer temperatures inside the furnace.
2) Combustion control to avoid excess air in the combustion chamber.
3) Reclaiming of energy from the waste gases by extra air preheating or batch or cullet preheating. In Western Europe some container glass furnaces are equipped with cullet preheaters or mixed cullet/raw material preheaters, which recover 10 – 20 % of the energy from the gases and return it into the furnace. There are different types of batch/cullet preheaters: direct contact preheaters (no physical separation between flue gas and raw materials); indirect contact preheaters (different channels for flue gas and raw materials), and cullet preheaters with electrostatic fields for capturing flue gas dust particles.
4) Using oxygen instead of air, which reduces the flue gas volume flow.
5) Improving control of glass melt viscosity in combination with new forming techniques to manufacture lightweight glass products.
6) Application of new furnaces, improving the heat transfer to and in the glass melt (e.g., Sorg LoNOx melter).

5.5. Glass Cooling

During cooling, a melt is transformed from a viscous fluid into a viscoelastic material and eventually into a brittle substance. These gradual transformations are accompanied by large changes in properties.

High residual stress can be generated during the transition from a melt to a cooled glass. The magnitude of such stresses depends largely on the cooling rate within the temperature range of the transformation from a liquid into a rigid state, the coefficient of linear expansion of the glass, the thickness of the glass, and the glass homogeneity. Annealing by controlled cooling between the so-called annealing point (viscosity $10^{12.4}$ Pa · s) and strain point ($10^{13.6}$ Pa · s) reduces residual stress to an acceptable level. The *upper stress release temperature* (Section 4.1.1) corresponds to a glass viscosity of ca. $10^{12.4}$ Pa · s or a stress release time of 15 min with a gradient of 40 °C/min. *The lower stress anneal* temperature (strain point) corresponds to a viscosity of $10^{13.5}$ Pa · s or a stress release time of 15 h.

Determination of the time required to release stress at a certain temperature requires measurement of the viscosity. This is not practical in the region of the stress release temperature. In practice, stress in the cold glass is measured optically as the retardation of polarized light in a stressed medium and expressed in units of nanometers of retardation per centimeter of material.

Glass is amorphous and therefore isotropic; its properties are often the same in all directions. However, under compressive or tensile stress anisotropic behavior appears, and the glass becomes birefringent. The applied stress deforms the polarizable (an)ions within the glass (prestressed electron cloud), which results in a different velocity of light transmission in one polarization direction as opposed to the other. A phase shift D appears, depending on the glass thickness and the stress present. This phase shift is is generally expressed per unit of glass thickness; usually 1 cm: $D/d = C \cdot s$, where D is the phase shift in nm, s the stress in Pa, C the stress-optical coefficient or Brewster's constant, and d the glass thickness in cm. Some values of C (in nm cm^{-1} Pa^{-1}) follow:

Soda-lime glass	0.27
Borosilicate glass	0.40
Quartz glass	0.34
$3\,PbO \cdot 5\,SiO_2$	0

Due to this birefringence phenomenon, stresses in glass can be measured by using a polarizing microscope. Monochromatic light is polarized by a polarizer. Because polarizer and analyzer are perpendicular to each other, no light is transmitted. When a birefringent material is placed between polarizer and analyzer at an angle of 45°, the two components of the beam of light – one parallel and the other perpendicular to the stress direction – are transmitted with different velocities through the material (retardation of polarized light in the direction of compressive stress), resulting in a difference in phase. Then

after passing through the sample, the two light beams interfere with each other. If the phase shift is equal to $1/2\lambda$, $(1 + 1/2\lambda)$, etc., the composite beam vibrates in the direction of the analyzer and therefore is transmitted. If the phase shift equals λ or a multiple thereof, the composite beam vibrates perpendicular to the analyzer and is extinguished. The stress release viscosity is determined from the required stress retardation time t by using the following formula:

$$D_s = D_0 + CA^{-1}t^{-1}$$

where D_0 is the original retardation caused by the stress, and D_s the retardation after time t, both expressed in nanometers per centimeter; A is the cooling constant; and C is the Brewster constant of the glass.

Residual stresses, expressed as maximum retardation of polarized light in the stressed glass, of the order of 100 – 200 nm/cm remain in the glass after cooling. The geometry of the article and glass composition affect the cooling process and hence the amount of residual stress. Tableware is cooled at such a rate that residual stress is less than 50 nm/cm; optical glass requires slow cooling to keep residual stress below 10 nm/cm. Figure 36 shows a recommended temperature – time curve for the cooling of a glass article.

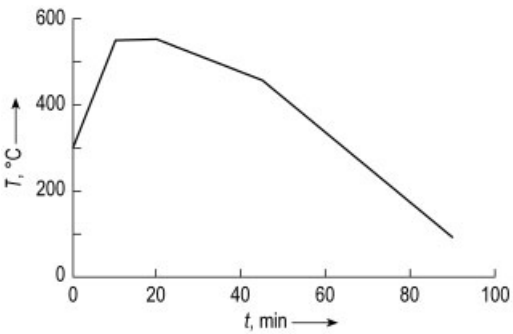

Figure 36. Annealing and cooling of a glass article (tube) with a diameter of about 10 cm and a wall thickness of 1 cm The annealing temperature is about 530 – 540 °C, and the strain point about 500 °C

5.6. Heating and Regulating Glass Melts

Natural gas, fuel oil, or electricity is used in most glass melting units. Radiation directly emanating from the flame and reflection from the crown into the melt are the principal processes of heat transfer in fossil-fuel-fired melters. Iron in the glass, especially in the iron(II) state, increases the IR absorption and the transparency of the melt for IR radiation decreases with increasing content of iron(II). Glass melts for products like tinted float glass and green or amber container glass show limited transmission of the IR radiation from the crown and flames. Extra convection or electric boosting in these glass melts is necessary to provide heat to the bottom section of the melting tank. Production of glass that absorbs strongly in the infrared requires a shallow bath, electrodes, or bubbling to maintain a sufficiently high temperature at the bottom of the pool.

Preheated air is injected into the combustion chamber in most furnaces. Both recuperative and regenerative preheating can be used. *Recuperative preheating* entails countercurrent or cocurrent heat exchange from the flue gases to the combustion air in a metal heat exchanger, the so-called recuperator. Often the flue gases flow through the core of the recuperator while air flows through the mantle of the twin shell recuperator or through tubes positioned in the flue gas channel of the recuperator. About 25 – 35 % of the flue gas heat can be transferred to the combustion air in such recuperators. Small glass tanks and tanks used for glass-fiber manufacture tend to operate with this system.

Regenerative preheating involves a discontinuous process that originated in the steel industry. Here, air is blown before combustion through a refractory brick chamber, which is itself heated by the flue gas. In this cyclic procedure, during a period of 20 – 35 min air is preheated to 1200 – 1350 °C in one regenerator, and the flue gases preheat the bricks (piled as a checkerwork) in another regenerator at the same time. The exhaust gases are cooled from about 1450 to 550 °C, transferring about 60 – 65 % of their sensible heat to the refractory bricks. After this period, the preheated regenerator is used for combustion air preheating, and the other regenerator is heated by flue gases.

The temperature in the melting tank and in the refractory walls is measured by Pt – Pt – Rh thermocouples embedded in zircon, AZS, or chrommium oxide refractory blocks and positioned about 5 cm deep in the vault of the tank. Some

thermocouples are positioned in the melt, for example, just before the melt enters the throat. Ideally, this temperature is constant and must be controlled; the set point for this thermocouple depends on melting load and glass color.

Thermocouples in the refractories of the crown or positioned in the combustion chamber just below the crown measure the combustion chamber temperatures, and temperature profiles in the combustion chamber are controlled to avoid damage to the crown and to achieve the required temperature profile in the melter.

Optical pyrometers are used to measure temperature along the walls of the tank or the temperature of the metal-line. Melt viscosity is not monitored but inferred from the estimated melt temperature, although some feeders are equipped with on-line viscosimeters to control glass melt viscosity prior to delivery to the forming process. Platinum-sheathed thermocouples are used in the melting end and before the melt enters the forming cycle. The temperature of the regenerator is also monitored with thermocouples. Tank pressure is kept within ± 0.04 kPa of ambient pressure and is measured by metal membrane manometers located below the furnace in the superstructure. Flue gas blowers or valves are used to control the pressure. To avoid leakage of flames or gases to the ambient air through the joints, a furnace pressure slightly lower than ambient pressure is preferred.

Sensor electrodes, optical laser reflection systems or nuclear radiation level devices measure the glass melt level. Changes in glass melt level are corrected by changing the batch charging rate.

The four principal variables that must be monitored during melting are the flows of natural gas or fuel oil and combustion air, batch charging rate, and temperatures in the melt and combustion chamber, all of which are kept constant for specific melt and fining temperatures. Standard technical gas analyzers or zirconia/platinum-based oxygen activity sensors are used to control the intake of fresh air and the composition of flue gases.

5.7. Refractory Lining of Melting Units

The three criteria for selecting refractory materials are: required lifetime, potential glass defect rate, and cost per amount of glass produced. To obtain an idea of the amount of refractory material required, consider a medium-size, 130-t/d tank with regenerative heating. Construction of such a tank requires 460 t of fire clay, 175 t of fused-cast tank bricks, 104 t of chrome magnesite ($MgO \cdot Cr_2O_3$) or magnesia – alumina spinel refractories, 112 t of magnesite or forsterite (Mg_2SiO_4) bricks, 100 t of silica bricks, 9 t of zircon ($ZrSiO_4$), and 2 t of unspecified lining material. Tank blocks poured from refractory melts are increasingly used as lining material in large glass tanks to improve contact between melt and lining material and to prevent problems with stones and cords in the final product.

The most widely applied refractory material for glass melting tanks is fused cast AZS produced from a melt of alumina, zircon (zirconia silicate), and small amounts of fluxing agents like soda. The AZS melt is poured in sand molds at temperatures exceeding 2000 °C. The sand molds containing the cast material are cooled very slowly to avoid thermal stress and formation of cracks. The refractory fused cast blocks must be post-treated to achieve the desired dimensions and to remove parts which contain voids caused by contraction during cooling.

A more recent development is the use of fusion-cast 96 % ZrO_2 blocks, a refractory material more expensive than conventional blocks. These are used only in those parts of the tank that are particularly susceptible to corrosion. The use of these blocks in melting tanks for the production of high-quality glasses like TV panel glass reduces the defect level in the final products.

Chemical considerations limit the choice of lining material. The most common materials used are silica, alumina, or zirconia, or combinations thereof. Linings containing sintered chromium oxide are used in the production of some specialty glasses like E glass for continuous filament fibers. The high price of tin oxide prevents its use in normal applications. Magnesium oxide dissolves in the glass melt and can therefore only be used in the regenerator or parts of the combustion chamber.

The crown of most glass furnaces is made of silica with very low amounts of residual quartz and with > 97 wt % SiO_2. For higher crown temperatures or aggressive gas atmospheres, the more expensive mullite, AZS (special quality

with a low fraction of silicates), or fusion-cast alumina refractories are used. The sidewalls of the combustion chamber are constructed from fusion-cast AZS, chromium oxide, and sillimanite or mullite for some kinds of glasses. Zircon refractories are applied between AZS and silica to avoid a direct reactive contact between silica and AZS. An overview of the refractory materials used is given in Table 27.

Glass melts tend to wet platinum (electrodes, stirrers, feeder elements, bushings) and molybdenum (electrodes). Caution is needed in using these materials in melts because some glass compositions and low or very high oxygen fugacity will corrode them. The vault of a tank is usually lined with silica, although its chemical resistance is not particularly good in specific temperature regions. The corrosion products of silica are soluble in the melt, however, and hence do not significantly affect melt composition. Melting products derived from silica, especially those resulting from reactions with sodium vapors from liquid sodium silicates, can react with the AZS sidewall refractories forming alumina-rich sodium silicate melts with some baddeleyite (ZrO_2) crystals or dissolved zirconia. These melts dripping into the melting tank can form knots and recrystallized components like secondary baddeleyite in the glass products.

The refractory lining of the tank and superstructure is 250 – 350 mm thick and covered by additional refractory layers. These layers provide thermal insulation, and the requirements are different to those of refractories in direct contact with the melt or furnace atmosphere. Insulating refractory bricks often show a high porosity. Typical insulating refractories are made of aluminum silicates like chamotte and sillimanite. Patch is used to seal the joints to avoid leakage and cold condensation spots. The tank refractories at the position of the metal line are not covered by insulation layers. The refractory blocks at melt-line level are often cooled with air from the outside to reduce the rate of attack by the melt at these positions. Melt-line corrosion is caused by microconvection of the glass melt at the surface due to surface tension gradients of the melt close to the slowly dissolving refractory. Microconvection removes the protecting diffusion barrier formed between the melt and the refractory material. This microconvection becomes slower at lower temperatures.

5.8. Vapor-Deposited Glasses

Vapor-deposited glasses are discussed with optical waveguides (see → Fiber Optics, Chap. 3.5.). The process is also of interest for specialty glasses (for instance for optical fibers) that require extremely high purity or are prone to liquid immiscibility in conventional melting. All vapor deposition processes use volatile halides (e.g., $SiCl_4$ or BCl_3) or metal chelate compounds. Rare earth organic chelates are used as additives in laser optics. These gases/precursors are processed to provide bulk high-purity oxide glasses. The pure halides are entrained in a carrier gas (oxygen) at < 323 K and react at > 1773 K. Reaction of the halide with oxygen results in condensation of a finely divided oxide powder on the substrate. The high-purity, inclusion-free powder has a surface area of about 20 m^2/g, which provides a sufficiently powerful drive for rapid, thermally activated (using flames, microwave, or plasma techniques), viscous sintering. Examples of glasses manufactured in this way are Corning Code 7943 and 7940.

5.9. Occupational Health

The principal health concerns in the glass industry have been reviewed [127–129]. The potential hazards specific to this industry include silica, lead, arsenic, antimony, asbestos, barium, chromium and other heavy metals. Heat stress, noise, and IR radiation must also be addressed. Other factors of concern are particulate and NO_x emissions from furnaces. Firing under reducing conditions may lead to excessive CO formation in the furnace. Most of the CO is oxidized to CO_2 in the flue gases, but those working in the vicinity of poorly sealed furnaces must be aware of health or poisoning risks. Three *control measures* are recommended to deal with the hazards associated with glass manufacture:

1) Engineering controls that include confinement, automation, and isolation in handling raw materials; batch room ventilation; and, whenever possible, substitution of questionable chemicals by harmless ones; batch handling is preferably carried out in closed

Table 27. Refractory materials used in glass furnaces

Refractory	Composition, wt %						Apparent porosity,	Density, g/cm^3	Maximum use temperature,	Linear thermal expansion, %		Use
	ZrO_2	Al_2O_3	SiO_2	MgO	Cr_2O_3	Other	vol%		°C	at 982 °C	at 1316 °C	
Fused cast												
ZB-X950	94	0.8	4.5			0.4 (Na_2O)	0.7	5.41		0.70 [a]	0.45 [b]	glass contact
AZS	33–41	46–50	12–16			0.8–1.8 (Na_2O)	1	3.8–4.1	1650	0.90	0.94	glass contact, superstructure
Corundum-Zircon	32.5–40	45–51	12–16			1.1–1.5 (Na_2O)		3.77–4.09		0.70 [a]		glass contact, superstructure
α-Alumina		99					1.1		1870	0.77	1.05	superstructure, crown
β-Alumina		94–95				5–6 (alkali)	4.1		>1650	0.65	0.85	breastwall, arches, crowns
α,β-Alumina		95	1			4 (alkali)	2		1870	0.73	1.05	fiberglass furnace throats
Alumina–chrome		60–61	1.5–2	6	27–28	4–5 (Fe_2O_3)	4.2		>1650	0.76	1.13	refiner or channels
Chrome–spinel		4–5	1–2	8–9	79–80	6–7 (Fe_2O_3)	4.9		>1650	0.71	0.97	textile fiberglass (high ware areas)
Bonded or sintered												
Dense zircon	64–65		33–34				0.5	4.22	1650	0.39	0.60	high-duty glass contact
Medium dense zircon	65–66		33–34				11	3.92	1650	0.48	0.73	channels, refiner, forehearth
Sintered zircon	65–66		33–34				16–28	3.39–3.76	1650	0.45	0.62	melting and refiner bottoms
Sintered zircon AZS	20	70	10				20–23	2.92–3.04	34–35	0.62	0.82	tubes, refiner paving, regenerator target walls

Table 27. Continued

Refractory	Composition, wt %						Apparent porosity,	Density, g/cm^3	Maximum use temperature,	Linear thermal expansion, %		Use
	ZrO_2	Al_2O_3	SiO_2	MgO	Cr_2O_3	Other	vol%		°C	at 982 °C	at 1316 °C	
Dense sintered alumina		94	4 – 5			0 – 1 (TiO_2)	1.6	3.44	1760	0.81		optical and lead glass tanks
Chrome – alumina		81			16	3				0.76	1.13	fiberglass tanks
90 % alumina brick		89 – 91	8 – 10				19 – 26	2.66 – 2.86	1750	0.69	1.04	superstructure
Chrome onide					96	4 (TiO_2)	18	4.06	> 1650	0.74	1.16	fiberglass tanks
Mul-lite/silimanite		60 – 80	18 – 37			1 – 6	14 – 24	2.24 – 2.59	1650 – 1700	0.58/0.62	0.81/0.93	backwalls, breastwalls, bottom checkers
Fireclay, bricks and flux		37 – 45	57 – 50				6 – 22	2.05 – 2.42	1400	0.38/0.56	0.62/0.75	day tanks or non-continuous furnaces, sidewall insulation
Superduty silica brick			96.3			2.9 (CaO)	20 – 24	1.84	1590 – 1650			crown
Fused silica brick			98 – 99			1.2	11 – 15	1.87 – 1.94	1650			burner blocks, thermocouple tubes, crown
Insulation brick (I.F.B.)		34 – 36	56 – 58			6 – 10	63 – 77	0.67 – 0.86	1090 – 1430	0.43	0.60	outer insulation

Table 27. (Continued)

Refractory	Composition, wt %						Apparent porosity,	Density, g/cm^3	Maximum use temperature,	Linear thermal expansion, %		Use
	ZrO_2	Al_2O_3	SiO_2	MgO	Cr_2O_3	Other	vol%		°C	at 982 °C	at 1316 °C	
Basic refractories												
High MgO brick				98		1–2 (CaO)	17–22	2.77–2.88	>1760	1.31	1.81	refractory checkers
Spinel-bonded magnesite		8.3		88.5		3–4	22	2.77	>1760	1.05	1.50	refractory checkers
Magnesite brick			4.1	93.6		2–3	20	2.78	1700	1.31	1.81	refractory checkers
Forsterite brick		10.9	29.5	50.3			22	2.62	>1650	1.10	1.39	refractory checkers
Dense chromic oxide					95	4 (TiO_2)	18	4.06	>1650	0.74	1.16	fiberglass furnace throats
Chrome brick		15–34	3–6	14–19	22–28	11–17 (Fe_2O_3)	18–25	2.88–3.36	1650	0.81	1.11	lower zone refractory checkers
Chrome–magnesia brick		16–27	4–8	27–53	12–28	8–14 (Fe_2O_3)	14–25	2.56–3.26	1540–1650	0.86	1.19	regenerator
Chrome–alumina		81			16	3				0.76	1.13	fiberglass tanks
Unichrome		59	29	1.0	4.2	2 (Fe_2O_3)						fiberglass tanks

[a] Linear thermal expansion at 1000 °C.
[b] Linear thermal expansion at 1500 °C.

spaces with baghouse filters to prevent dusting during filling or discharging of the silos.
2) Administrative controls, including job rotation, education, safety instructions and proper work practices.
3) Use of personal protective equipment, including respiratory protection, proper clothing, and hearing and eye protection.

In the United States, OSHA has established the right of employees to be informed of the potential hazards of materials in their workplace. This organization also requires that producers of hazardous materials determine the potential health and physical hazards of such materials and transmit this information to the consumer. Among the standard reference works on industrial hygiene is *Patty's Industrial Hygiene and Toxicology* [130]. For most toxic components limits have been set for levels in ambient air. Procedures and instructions for handling toxic materials or flammable and explosive components are often part of ISO 9002 or other standardized practices in modern glass factories.

6. Forming

The shape of a piece of hot glass (temperature above T_g) can be changed by an external force which causes viscous deformation. In the known forming techniques the external force is applied by:

- Drawing (fiberglass, flat glass, tube glass)
- Blowing (container glass, bulbs)
- Pressing (tableware, container glass, TV panels)
- Centrifuging (insulation glass wool)
- Rolling (wired glass, figure glass)
- Casting

In some cases, these processes are combined. The degree to which the glass is transformed depends on the magnitude of the force and the prevailing glass viscosity. A low viscosity imparts a high rate of deformation. After the forming process a high viscosity is required, so that the product retains its shape (viscosity of about 10^7 Pa · s, softening point). During forming, the glass should be cooled down precisely to achieve these conditions. The degree of cooling depends on the heat content of the glass product and the rate of production (contact time in the mold). Large products with a high heat content can not be formed as fast as smaller products. Most forming processes take place at a viscosity of 10^2 to about 10^4 Pa · s (working range). The working temperature is defined as the temperature at which the melt has a viscosity of 10^3 Pa · s. During forming, the glass melt cools and the viscosity increases. With increasing viscosity, the required transforming force increases, and the deformation rate will decrease. Viscosities and temperatures for some forming techniques are listed in Table 28.

Table 28. Deformation and forming methods for different glass products

Process	Viscosity, Pa · s	Temperature,∗ °C
Drawing, centrifuging	10^2	1200
Fast pressing	10^3	1050
Fast blowing	10^4	900
Slow pressing	10^5	800
Slow blowing	10^6	750
Form becomes rigid (or stable)	10^7	700

∗ For soda-lime glass.

Most of today's forming methods have their origin in the use of the blowpipe, invented by the Romans and reaching the pinnacle of expression with the Venetian glassblowers of Murano.

Molten glass is pressed, blown, drawn, rolled, poured or quenched, depending on the desired shape and intended use. Bottles, dishes, optical lenses, and television tubes are formed by blowing, pressing, casting, or spinning glass against a mold to cool it and set its final shape. Window glass is made by floating the glass melt on a molten tin bath; tubing rods and optical fibers are made by drawing the glass in air until it sets and can be cut to length. Frit glass, some of which is intended to be crushed into powder, is quenched between water-cooled rollers or ladled directly into water and then dried (dry-gauged). Tube glass is drawn from the melt by conducting the melt through a feeder opening on a sloped or vertical rotating mandrel (blowpipe). The glass distributes itself along the cylindrical surface of the mandrel and is drawn out into a tube. The mandrel consists of a hollow ceramic cylinder through which air is blown. This air flows through the resulting glass tube, which is released from the mandrel. Mandrel

size, air pressure and drawing speed and glass melt viscosity determine the dimensions of the glass tube. Art glass is often hand-formed by free blowing and shaping while soft. These forming operations all require a melt cooled to a characteristic viscosity of between 10^2 and about 10^4 to 10^5 Pa · s, generally at 800 – 1400 °C (depending on glass type), as shown in Figure 13.

6.1. Hand Forming

Some industrial forming methods derive from hand-forming techniques and are therefore discussed in some detail here. The major breakthrough in glass production was the invention of the blowpipe by the Romans. The flexibility of this technique still makes it popular in fabricating complex specialty items. The blowpipe is a hollow stainless steel pipe. A gob of glass is collected at the end of the pipe by dipping it into the melt. Blowing air through the pipe results in swelling of the gob, provided the viscosity of the glass is still sufficiently low, reheating is necessary if the viscosity becomes too high.

The *Rhenian technique* uses a blowpipe with a conically thickened end. The glassblower gathers the appropriate amount of glass melt from the working end of the furnace or the pot and collects the glass on the end of the blowpipe, and the gob is cooled by rotation in a bowl-shaped wooden form or on a metal plate until the viscosity required for blowing is reached. The gob is then placed in an iron or wooden mold and blown to the shape of the mold.

In the *Bohemian technique*, a straight blowpipe is used. A gatherer, who collects the gob on the pipe, and the glassblower work in concert. The gatherer collects a small amount of glass on the end of the pipe and blows a small sphere with dimensions prescribed by those of the final product. The glassblower takes the pipe from the gatherer, dips it again in the oven, collects as much glass as needed for the final product and blows it to the shape of the mold.

6.1.1. Blowing

The hand-forming operations described above immediately suggest a way of automating a glass-forming process by blowing gobs of the proper viscosity into molds. Many products are indeed made by variations on this theme, as suggested by the names of the various techniques: blow, blow-and-blow, puff-and-blow, and press-and-blow processes.

Deep items (e.g., bottles, jars, and light-bulb envelopes) are formed by *blowing* glass into molds with air pressure, as shown in Figure 37. Hinged molds, which can be opened to remove the ware, are generally used for the blowing operation. Articles with circular cross section are blown into hinged paste molds. These molds are lined with a water-absorbent coating or wetted cork, which develops a steam cushion between the coating and glass during blowing, allowing the glass to be rotated in the mold. A second type of mold is the hinged hot-iron mold used for the manufacture of ware with noncircular cross section. Here, the glass comes directly into contact with the mold surface on blowing. The surface quality of articles blown into hot-iron molds is inferior to that of paste molds. Hot-iron molds are used for ware of any shape, in particular for screw threads, multiple decoration, or raised lettering.

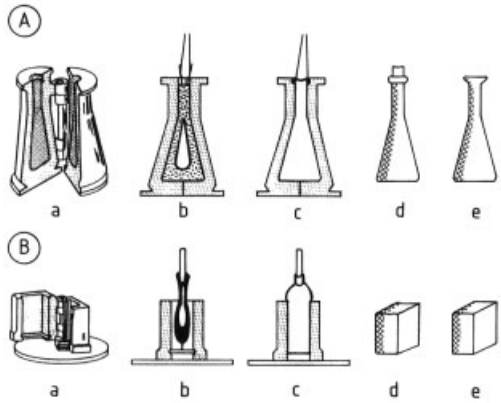

Figure 37. Forming by blowing
A) a) Empty mold; b) Blank ready for blowing; c) Blank blown out; d) Untrimmed piece; e) Finished piece
B) a) Empty mold; b) Blank ready for blowing; c) Blank blown out; d) Untrimmed piece; e) Finished piece

Narrow-mouth containers are manufactured in a two-stage, *blow-and-blow* process (Fig. 38). The gob is delivered to a blank mold, where a settle blow and a counter blow form it into a blank. This blank is then transferred to a blow mold, where it is blown into its final shape.

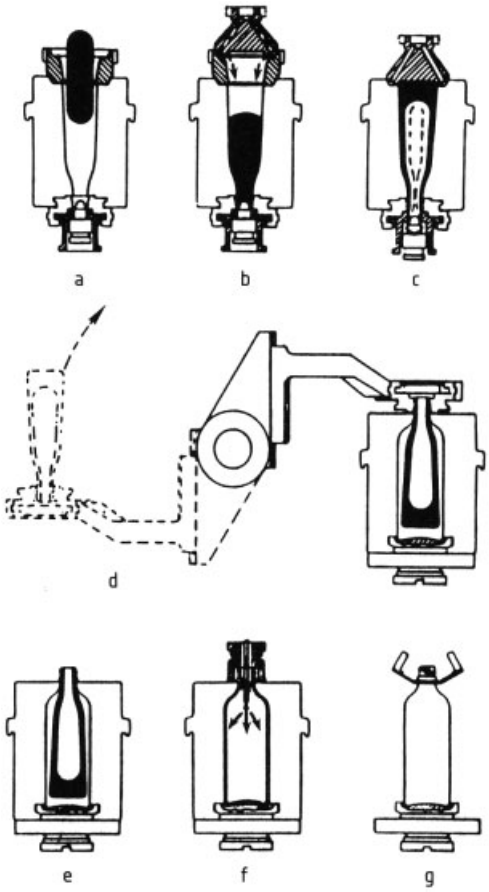

Figure 38. Forming by the blow-and-blow process
a) Delivery; b) Settle blow; c) Counter blow; d) Transfer from blank mold to blow mold; e) Reheat; f) Final blow with internal cooling; g) Takeout

The Corning ribbon machine is the classical example of mechanized glass-forming processes (Fig. 39). The machine is used to manufacture simple seamless products, such as light bulbs (1200/min), flash-bulb envelopes (2000/min), and Christmas tree ornaments. A *puff-and-blow* method is used in which a glass ribbon 50 mm thick is formed at about 1080 °C by two rollers. The lower roller has small indentations, which form small circular thicker regions in the glass ribbon. Transport of the ribbon brings the thickened regions above a hole. The glass sinks into the hole and is blown out by blow-heads that move synchronously with the ribbon. The turret chain machine is used for larger articles. This device also uses the puff-and-blow method but with individual glass gobs rather than a ribbon, thereby reducing the amount of cullet.

So-called gobs are delivered by the feeders of the glass furnace by shears cutting portions of a certain weight from a flow of glass melt. These gobs are conducted to the blowing machines.

Blowing machines for bottles in use today include the Lynch-10, capable of delivering 25 – 80 pieces per minute with bottle weights of 30 – 600 g; the Roirant-R-7, delivering 20 – 80 pieces per minute at weights of 100 – 1200 g; and the Hartford-IS-12, now manufactured by Emhart and Maul and used almost universally, which operates with the blow-and-blow process and produces about 10 pieces per minute. There are IS-6, IS-8, IS-10, IS-12 forming machines supplied by several machine manufactures; the number stands for the number of sections in series per machine. In each section 1, 2, or 3 gobs are simultaneously formed into a jar or bottle. A IS-8 machine with 3 gobs per section may produce more than 250 bottles per minute with a weight of about 200 g per bottle. Today, even larger machines are used for mass production.

6.1.2. Pressing

The role of air in blowing glass is taken over by a plunger (male half of the mold) in *pressing glass*. Blowing is used in the manufacture of deep items, whereas pressing is used particularly for relatively flat items such as dinnerware, optical and sealed-beam lenses, filter glass, or television tube panels. The press mold consists of three parts: the mold bottom, the plunger, and the ring; the latter functions as a seal between the other two parts. In hand pressing, molten glass is gathered on a solid steel rod *punty* (pontil) and drips into the mold bottom. A pair of shears severs the glass remaining on the punty from that in the mold. The plunger is then forced into the mold, causing the glass to flow into the space between plunger and mold. Movement of the plunger ceases when the space is filled. The plunger is withdrawn and the pressed glass removed from the mold. Three types of molds used for pressing are shown in Figure 40.

The principles of *machine pressing* are similar to those of hand pressing. For instance, a set of molds, commonly made of cast iron, bronze steel, or superalloys with galvanized surfaces,

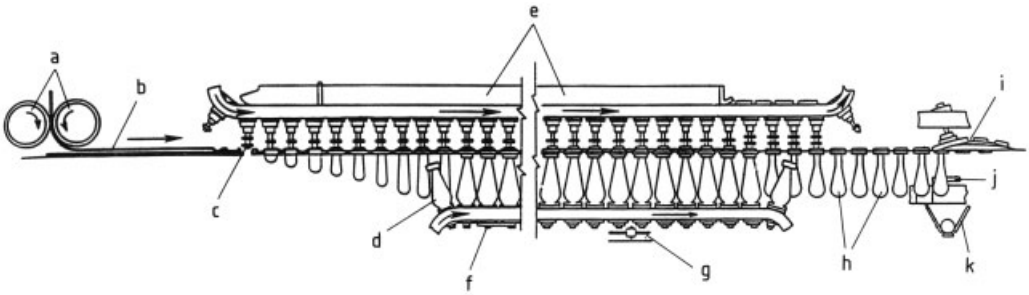

Figure 39. The Corning ribbon machine
a) Water cooled rollers; b) Glass ribbon; c) Blow heads; d) Rotating paste molds; e) Blow box; f) Mold-closing cam; g) Mold-opening cam; h) Glass bulbs; i) Stripper; j) Crack-off bar; k) Conveyor for removing product

is mounted on a circular, horizontally rotating press table. Typically 8 – 16 mold positions lie on the periphery of this table. Glass flows from the refining end or feeder into a refractory bowl with a circular hole (spout) in the bottom. A reciprocating plunger of refractory material above this hole causes the glass to be extruded in pulses. Extruded glass is separated into gobs of the proper volume by a pair of shears directly below the hole. The motions of reciprocating plunger, shears, and press table are synchronized, ensuring that the gobs fall into successive molds. The plunger of the pressing process descends and presses the gob into the desired shape. The glass is usually cooled by blowing air against it. The hollow plungers are cooled separately by circulating coolant.

Mold temperatures must be monitored; glass may adhere to a mold that is too hot or have a wrinkled surface if the mold is too cold. A rubbering compound like a graphite suspension is applied to the mold, plunger, and ring to prevent adhesion of glass to the mold.

The pressing speed is determined by thermal conductivity, temperature, and temperature distribution. Normal pressing temperature is 670 – 800 °C; the lower limit is determined by the shapeability of the glass; the upper, by the oxidation resistance of the mold.

6.1.3. Blowing Combined with Pressing

In blowing and pressing, the gob is first pressed in a blank mold followed by transfer of the blank or parison to the blow mold. The blank is blown to its final shape in this second mold by a blow-head set over the neck (Fig. 41). For modern mass production IS machines with 10 – 16 sections are used to make press-blow items with wide necks (jars) and narrow necks (narrow neck press blowing) at high production rates. In each section two or three jars or bottles are produced. Generally, press-and-blow forming processes give a more uniform and more controllable glass distribution during forming than blow-and-blow processes. For bottles and jars press-and-blow processes allow smaller weights per item, while still obtaining sufficient mechanical strengths.

6.1.4. Drawing

The three forming processes used to manufacture glass tubing or rods are the Danner, Vello, and Schuller process. The Vello process has gained favor over the Danner process because of greater flexibility.

In the *Vello process*, molten glass flows downward through the annular space between a hollow vertical mandrel and a refractory or (noble) metal ring set in the bottom of a special forehearth, the bowl. Air blown through the mandrel into the fresh glass tube maintains the dimensions of both bore and diameter. The soft tubing, hanging in the shape of a catenary between the end of the mandrel and the drawing machine, is pulled onto a horizontal conveyor to the cutting machine.

Diameter tolerance for drawn glass tubing is typically 5 – 7 % for 6-mm tubing and 2 – 3 % for 150-mm tubing. Wall thickness tolerance is 11 – 15 % for the smaller diameter tubing and 12 % for tubing of 100-, 125-, and 150-mm diameter.

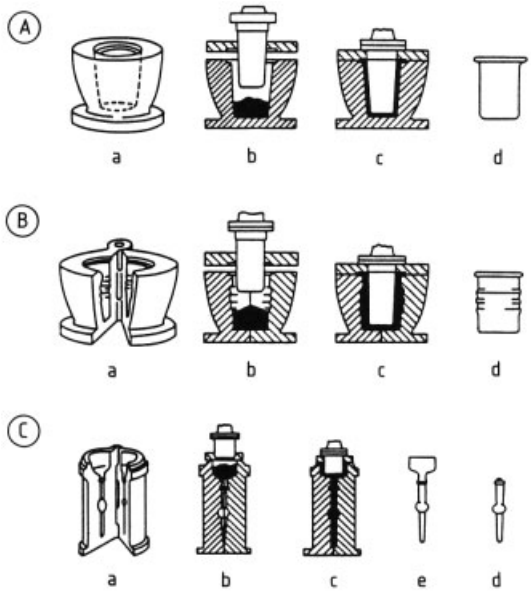

Figure 40. Types of molds used for pressing
A); B); C); a) Empty mold; b) Loaded mold; c) Glass pressed; d) Finished; e) Untrimmed

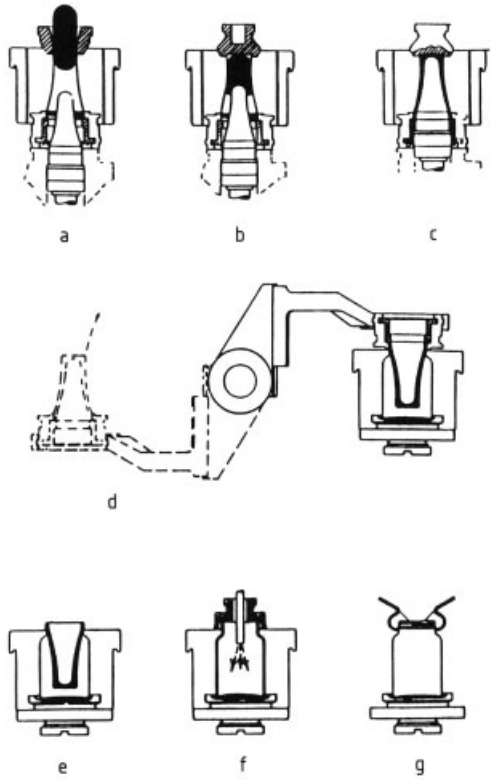

Figure 41. Blowing combined with pressing
a) Delivery; b) Start of plunger pressing stroke; c) Full pressing stroke; d) Transfer from blank mold to blow mold; e) Reheat; f) Final blow with internal cooling; g) Takeout

The Vello process can accommodate a wide range of glass compositions and production rates of >45 kg/min with good dimensional control for tube diameters up to 65 mm. With special care, as required for thermometer tubing, the standard deviation of the tubing diameter can be held to 0.3 – 0.5 % of the nominal dimension.

The change in direction of the drawn tube from vertical to horizontal in the Vello process causes kinking in larger diameter tubes. A modification of the Vello process, the *down-draw process* prevents this by drawing tubing up to 280 mm straight down from the bowl with good dimensional control, while maintaining the production rates and glass-composition flexibility of the Vello process.

Rods can be drawn by any of the tube-drawing processes by turning off the air flow through the mandrel. The glass tube collapses, and rods are then formed, rather than tubing. The first mechanical drawing process for tubing glass was developed by Danner (USA) in 1917 (Fig. 42). In this process the liquid glass flows through a feeder opening on top of a downward-sloping rotating (5 – 10 rpm) mandrel (blowpipe). The glass distributes itself along the cylindrical surface of the mandrel and is drawn out into a tube. The mandrel consists of a hollow ceramic cylinder through which air is blown. This air flows through the resulting glass tube, which is released from the mandrel. Depending on the desired diameter of the glass tubes (1 – 70 mm), mandrels are used with a diameter varying from 100 to 700 mm.

The tube dimensions can be adjusted by means of the pressure of the compressed air and the drawing speed. At constant drawing speed, a higher compressed air pressure leads to a larger

tube diameter with a thinner wall thickness; a larger drawing speed with constant compressed air pressure results in a smaller tube diameter and a thinner tube wall. By combining these control factors, tubes of different dimensions can be manufactured with the same mandrel.

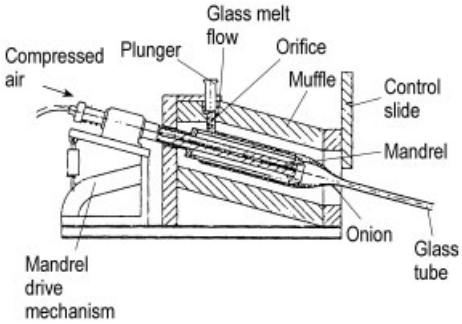

Figure 42. The Danner process

The consistency of the desired dimensions strongly depends on a constant discharge of the glass melt from the feeder and on a stable temperature profile of the glass during both the discharge of the melt from the feeder into the mandrel and drawing into a tube. The viscosity of the discharging glass melt from the feeder opening should be about 10^2 Pa · s and at the end of the mandrel the viscosity should be about $10^{3.5}$ Pa · s. The drawing speed amounts to about 400 m/min. for glass tubing of 2.5 mm diameter and about 100 m/min for 37 mm tube. The maximum production capacity of the Danner process is about 40 t/d.

In the Schuller process, developed in Germany in 1931, the glass tube is vertically drawn upwards. There are several modified versions of this process. The glass flows from the feeder into a rotating refractory dish with a blow pipe in the center at the bottom (cone). From this dish the glass tube is directly drawn upwards. Around the initial part of the glass tube a ceramic cylinder is located that rotates synchronously with the dish and which ensures a uniform temperature of the glass melt while it is drawn upwards. The tube dimensions are determined by the dimensions of the cone, the compressed air pressure, and the drawing speed. They can also be influenced by means of the cooling air, which is blown from holes in a cooling ring, located above the rotating cylinder. By using different blow pipes glass tubes can be manufactured with a diameter ranging from 3 to 50 mm. The production capacity per machine is up to about 3 t/d.

6.1.5. Redrawing

In the redrawing process, glass first is formed into the desired shape and subsequently reheated. The specific characteristic of this redrawing is that it preserves shape; e.g., a piece of glass with a square cross section of centimeter dimensions can be redrawn into a long rod with a square cross section with micron dimensions. The topology-preserving character of the redrawing process makes it ideally suited for miniaturization and is the basis for all optical waveguide technology.

6.1.6. Casting

Casting is a glass-forming process in which melt is poured into a mold. The most spectacular examples of cast products are telescope mirrors for the space observatories, often made from borosilicate glass; pieces larger than 5 m in diameter and weighing more than 36 t have been produced. Dimensional tolerances as low as 0.1 mm can be achieved in this process. Centrifugal casting is a variant of the casting process: a gob of glass drops into a mold, which then spins rapidly, spreading the glass by centrifugal force over the inside of the mold (Fig. 43). Funnel-shaped items, such as television tubes and missile radomes (nose cones), are formed by this process.

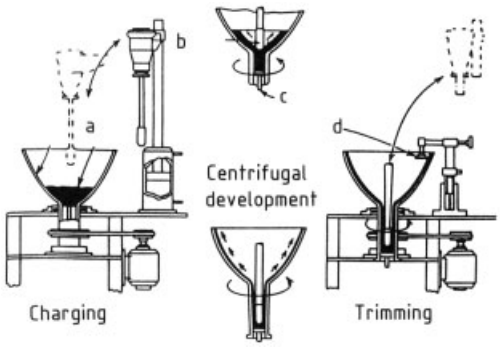

Figure 43. Centrifugal casting of television tubes
a) Mold glass; b) Distributor; c) Ejector valve; d) Grooving disk

6.1.7. Fritting

Fritting involves quenching a melt to create a highly thermally stressed glass that can be crushed easily in a ball mill. The glass is quenched between water-cooled rollers or poured into water. The former method is advantageous because it produces a thin friable ribbon that often gives more uniform mill feed.

6.1.8. Flat Glass Forming

6.1.8.1. Pilkington Float Process

Flat glass manufacture was revolutionized in the late 1950s by the invention of the Pilkington float process. This process has made other flat glass manufacturing processes (e.g., the Libbey – Owens, Fourcault, and Pittsburgh processes) obsolete. In the older processes, the glass sheet is pulled vertically through a slotted draw bar by the action of a pair of rolls contacting the glass melt sheet (Fourcault), or the glass is pulled over a bar. In the Pittsburgh system the glass is pulled upwards between pairs of rotating rolls after passing across a solid drawing bar. At the same time, the glass sheet is cooled by the draught created within the machine, which is about 13 m high. The maximum capacity of these machines is 50 t/d. Up to ten machines per furnace have been realized, resulting in a furnace capacity of about 300 t/d. This process still is being used in Eastern European countries and in the developing countries because of the relatively low investment costs.

In a similar system, developed by Libbey – Owens, the glass is also pulled vertically from the furnace. However the glass ribbon is bent across a bending roll at a height of about 1 m from the surface roof and subsequently pulled over a horizontal roller conveyor through a cooling zone. This process and the Fourcault process are still being used for thin glass (≤ 0.4 mm), for optical applications, and for LCD-displays.

In float glass production large cross-fired furnaces with regenerators are used for glass melting. Today, all-oxygen-fired furnaces are also used to melt the glass. The glass tank in which the molten glass is melted and conditioned prior to forming is typically 9 m wide, at least 45 m long, and usually contains more than 1200 t of glass (Fig. 44). Molten glass flows from the melting tank into a bath of molten tin, 3 – 4 m wide, 50 m long, and about 6 cm deep. The temperature of the molten tin bath (650 to 1100 °C) lies in the range of the working temperature of the glass. The glass surface flows and smooths itself while on the float bath, taking on the excellent surface quality of molten tin and thus requiring no further polishing. The glass body is rigid enough to be handled without damage by the time it leaves the bath at about 600 °C. Tin is used as flotation medium because of its surface tension (about 0.55 N/m), density (5.9 g/cm^3 compared with 2.2 – 2.5 g/cm^3 for glass), low melting point (232 °C), and relatively low vapor pressure. The thickness of the float glass ribbon is controlled by the pull rate and top rollers, which force the ribbon of still-molten glass to pile up or to spread. Sometimes an inorganic coating on the surface of the glass is directly applied just before the ribbon exits the tin bath section.

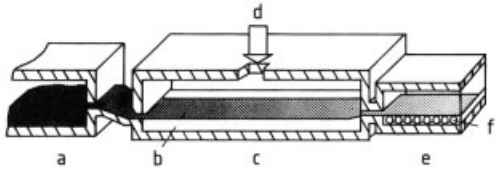

Figure 44. The Pilkington float glass process
a) Furnace; b) Molten tin; c) Float bath; d) Controlled nitrogen – hydrogen atmosphere to prevent oxidation of the tin; e) Lehr; f) Rollers

The advantage of Pilkington float glass, apart from the mirror quality of the glass surface, is the high production capacity relative to other drawing processes. The velocity of the emerging glass is about five to ten times higher than in the Libbey – Owens process. An additional economical advantage is that a float glass plant with a daily production of 400 – 800 t/d can be operated by four to six persons during an 8-h shift.

6.1.8.2. Rolling

Sheet glass with figured or embossed surfaces is formed by squeezing a stream of molten glass between rollers that have the embossing pattern on their surface. Wired glass is produced either by rolling a sheet of glass, covering it with wire, covering the wire with another sheet of glass,

and fusing both sheets together with the wire between them, or by making a sandwich of two sheets of glass with the wire between them which is then rolled.

6.1.8.3. Overflow Process

The overflow process is used by Corning Glass Works for specialty glasses such as ophthalmic glass, borosilicate glass, and glass that has to be chemically strengthened. The melt flows over two opposite sides of a trough, forming a combined sheet underneath the trough. The glass surfaces are untouched by a forming or handling tool until rigid, producing a surface quality comparable with that of glass manufactured by the float process.

6.1.8.4. Lamination

A stressed laminated glass with all surfaces in compression can be formed by laminating a low-expansion glass as a skin to a high-expansion glass core. The two glasses are fed into a laminating orifice, through which the composite sheet is drawn. This process is used primarily to produce food-service ware.

6.1.9. Fiberglass Forming

Four processes exist for manufacturing glass fibers: 1) rotary fiberizing, 2) steam or air blowing, 3) mechanical traction or drawing, and 4) flame blowing. The mechanical and rotary methods are the most important.

In *rotary fiberizing* a stream of molten glass is poured into a superalloy spinner rotating at 2000 – 3000 rpm. Centrifugal forces push the melt up the sides of the spinner and out through the perforated band around the periphery. A downward blast of air or hot gas stretches the loose fibers to a diameter of 3 – 6 µm. A phenolic binder in aqueous solution is used to bond the fibers together, and the fibers are collected as a mat on a conveyor belt and then cured in an oven and cut. A typical rotary fiberization operation is shown in Figure 45.

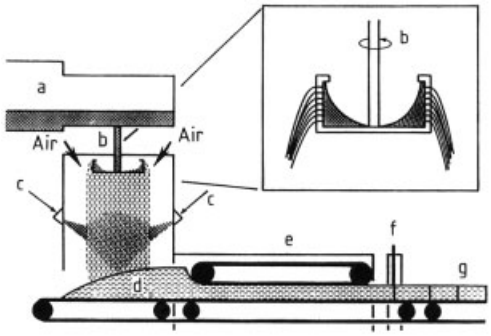

Figure 45. Rotary fiberization
a) Melting tank; b) Centrifuge with blowers; c) Binder spray; d) Uncured wool; e) Binder curing oven; f) Chopper; g) Products

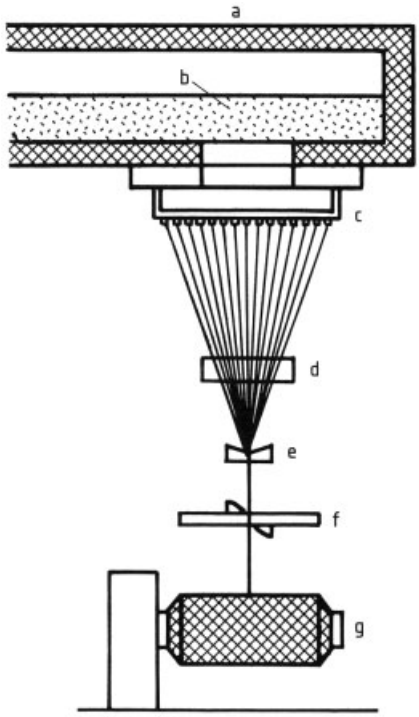

Figure 46. Continuous drawing of textile glass fiber
a) Forehearth; b) Glass; c) Bushing; d) Size application; e) Strand formation; f) Traversing; g) Winding

The tanks used in *mechanical drawing* of continuous fibers contain up to 200 t of molten glass. Glass furnaces with up to 100 t fiberglass production per day exist. The molten glass flows from the forehearths into platinum alloy bushings (Fig. 46). The bushing, provided with several hundred nozzles, is located below a slot in the bottom of a canal (forehearth) on top of the

spinning section where the single filaments are spun into a single fiber (strand) by the collet. A solution is sprayed on the filaments to establish a fixation of the filaments in the strand and to give good adhesion properties. At the lower end, a rotating drum (collet) is located, which takes care of pulling and winding of the bundle of glass filaments (strand) at a constant velocity. Due to a meticulous system of traversing the strand upon the cylinder shaped package, it is possible to unwind the strand without any damage during subsequent treatment. Strand-winding velocities vary from about 1000 to 5000 m/min. Each bushing can deliver from several hundred up to more than 1000 glass filaments to be processed as fiber strand. Diameter control and winding techniques are similar to those encountered in optical waveguide manufacture.

6.2. Annealing

Annealing of a glass product is generally necessary after any forming operation. The glass is heated uniformly to a temperature sufficiently high to relieve any internal stress without causing excessive deformation of the object under its own weight. The ware is subsequently cooled slowly to prevent formation of new stresses. Cooling may proceed as rapidly as desired once the glass is rigid, provided that temporary stress does not cause breakage. Glassware is normally annealed in long, continuous ovens called *lehrs*, which are usually heated electrically, but also gas fired lehrs are applied. They are designed with great care to ensure a uniform side-to-side temperature, especially in the hottest zone. Bottles produced at a rate of 100 t/d are annealed in 15 – 20 min. Lehr design improvements include shortening and widening of the temperature chamber and new control techniques to ensure precise temperature distribution.

6.3. Secondary or Finishing Operations

6.3.1. Cutting

After emerging from the lehr, flat glass is rough-cut by scoring with a cutting wheel or a diamond. After inspection, the glass is cut into standard market sizes. Tubing is also rough-cut after leaving the conveyor, inspected for proper dimensional tolerance, and then cut to the required length. Several methods are used for this final cutting, ranging from scoring the tube and applying mechanical stress, to local heating of the tube with an electric wire and sudden chilling (thermal shock).

After glass blowing, part of the glass often has to be cut off mechanically by scoring the surface and applying a stress. Then this blown ware commonly contains broken edges. But the moil, or waste glass, is also frequently separated from blown ware by directing pinpoint flames onto the glass at the cutting line and simultaneously rotating the ware until separation of the moil and ware edges is complete. The surface tension of the fused glass produces a smooth, well-rounded edge. This operation is called burn-off or flame cutoff. Hot glass is normally trimmed by cutting the form with a hollow plunger.

6.3.2. Grinding

The purpose of grinding is to remove the upper layer of the glass surface. An abrasive produces a small check or crack in the glass, the dimensions of the checks depending on the grain size of the abrasive: increasingly finer abrasive produces a relatively smooth surface. Natural abrasive grits (e.g., quartz, sandstone, corundum, garnet, and diamond) are used, as well as synthetic grits (e.g., silicon carbide, aluminum oxide, and B_2O_3). Water or a suitable cutting fluid is used to speed up the process and to prevent overheating of the glass and glazing of the abrasive surface in bonded grinding wheels. Continued grinding results in an interconnected network of checks on the surface that break off with the formation of new pressure points. The result of grinding is an irregularly broken surface on the glass. By decreasing the coarseness of the grits in steps, the glass surface becomes more and more polished. In using very fine polishing techniques, the irregularities merge to a smooth surface.

A variety of grinding machines are used for different purposes. Grinding mills are large horizontally revolving iron plates, fed with loose abrasive grains such as silicon carbide and water. Copper engraving wheels used to engrave

art glass also employ loose abrasive grains and water. The common feature of bonded grinding wheels, belt grinders, and diamond-impregnated wheels is that the abrasive is embedded in the moving part. Another method is sandblasting, used to produce a diffuse surface.

6.3.3. Polishing

Polishing is carried out either mechanically or chemically. *Mechanical polishing* requires finely powdered abrasives such as rouge (ferric oxide) or CeO_2. These abrasives operate on the same principle as those in grinding. The abrasive is applied on a revolving buff made of felt, leather, lead, copper, or pitch. Polishing removes virtually no glass from the surface. The effect is thought to arise from induced surface flow at the molecular level, which fills local pits and removes the material lying above the average surface level. An optical polish produces very fine and accurate surfaces and is an art mastered only by specialists.

Chemical polishing includes acid polishing and flame polishing. In acid polishing the glass is submerged in a mixture of hydrofluoric and concentrated sulfuric acids. The material lying above the average surface level is preferably dissolved. In flame polishing a flame is directed onto a glass surface, producing a fire polish or fire glaze. In plasma polishing, short-term, local high temperature (plasma temperature: 8000 – 10000 °C) is used to melt the glass surface.

6.3.4. Heat Processing

In some finishing operations, reheating of glass is required. These operations include fusion sealing, shrinking, and tempering.

Fusion sealing is used when glass parts must be sealed together, e.g., the spun cast television tube (funnel) and its faceplate (panel). Fusion sealing is often carried out electrically. The two ends to be mated, are positioned in a lathe and heated by gas burners until the glass becomes sufficiently electrically conducting. At this stage an electric arc is struck to the glass, raising the edges of the glass pieces to the fusion temperature. The arc continues for a few seconds after the parts are joined. The seal is worked by blowing and pulling after the arc is turned off. Often, a low melting sealing glass is used to seal two glass parts together.

Electrostatic sealing (anodic bonding) is a type of field-assisted glass-to-metal sealing, which has been used successfully on steels, alloys, and silicon substrates. The application of a large d.c. field to a closely-fitted glass-metal interface allows sealing at temperatures well below those required for viscous flow, wetting, and chemical bonding. Both strong electrostatic forces and ionic diffusion have been implicated in the mechanism.

Shrinking is used to produce tubing with an extremely close inside diameter tolerance, e.g., for fluid flowmeters. A mandrel is inserted into the tubing, and the assembly is heated until the glass softens. The tubing collapses onto the mandrel after the space between the two parts is evacuated. Tubing of any shape (e.g., square, rectangular, triangular) can be produced in this manner. Optical glass fibers are produced from massive cylindrical rods by heating until the glass softens and subsequently stretching the rod to the desired fiber diameter.

Thermal tempering, or quench strengthening, is used to strengthen glass by stressing to bring all surfaces into compression. This finishing operation is used for flat glass, piping, some cookware (e.g., pots and pans), tumblers, gauge and sight glass. The ware is heated close to the softening point at least above the strain point and then quenched by a blast of air, a fused salt bath, or some other means. The outer skin of the ware solidifies whereas the inner mass remains soft, resulting in compression of the outer skin on further cooling. An increase in strength by a factor of four or more can be obtained in this manner, although in practice most thermal tempering treatments result in a strength increase by a factor of 2 – 3. Glass thinner than about 0.5 mm, and glasses with very low thermal expansion coefficients such as fused silica, and 96 % silica cannot be tempered.

6.3.5. Sintering

Sintering is used whenever intricate shapes such as porous filters and tiny sealing beads are required. A variety of glass compositions from lead glass to fused silica can be used. The glass

is finely ground, made into a paste or slurry with a liquid, and cast or pressed to the green, intermediate shape desired. After drying, the green form is heated until the particles soften and fuse with concomitant shrinking of the ware. Sintering usually produces a translucent glass. which can, however, be rendered transparent with a density similar to that of the original glass.

6.3.6. Chemical Treatment

Hydrofluoric acid is used for *etching* (e.g., graduations on volumetric laboratory glass ware). The item is covered with acid-resistant wax in which the graduations are cut, and the exposed glass is then etched by immersion in a hydrofluoric acid bath. Light bulbs are frosted with hydrofluoric acid or by injection of ammonium fluoride solution into the bulb.

Staining involves ion exchange between alkali metal atoms in the glass and copper or silver in the stain, producing a red or amber color, respectively. Red stain is used for markings on volumetric glassware, and silver stain for ware with low near-UV transmittance.

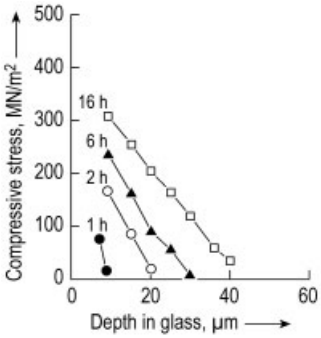

Figure 47. Na^+/K^+ exchange of flat glass in a KNO_3 melt at 440 °C; the compressive stress at the surface reaches a value of 430 – 480 MN/m^2 after a short time.

Strengthening also involves ion exchange: the glass is immersed in a hot salt bath, which causes small alkali metal ions near the surface of the ware to be replaced by the larger ions from the bath. The surface is under compression because of the presence of the larger ions. The degree of compression can be as high as 700 MPa, depending on the glass composition and the strengthening cycle employed. The compressive layer can be much thinner (100 μm) than that obtained by tempering. Examples include windows of supersonic aircraft, glazing, and ophthalmic lenses. Figure 47 shows the compressive stress in the surface layers of a glass treated in a potassium nitrate melt as a function of the duration of the treatment.

6.3.7. Surface Coating

Vitreous enamels are applied by spraying, silk screening, and decalomania, i.e., the transfer of pictures from one medium to another. They are fused to the ware at high temperature but below the softening point of the glass itself.

Metal oxide coatings are produced either by applying a liquid suspension of metal oxide to the surface and then firing it or by vapor deposition of metal oxides on the glass. These coatings are used on flat glass as a means of controlling transmission and reflection of light, as electrical semiconductors in tin oxide films, as resistors in electronic circuitry, as defrosting agents and on aircraft glazing. Metal oxide coatings (hot end coatings) are also applied on container glass articles to improve scratch resistance and adhesion to polymer coatings (cold end coatings).

Metallic coatings are produced by special coating installations that sputter metal layers onto glass surfaces at low pressures. They are used for thermal insulation of flat glass products. The coating material is bombarded with argon (positive) ions generated in a corona discharge at a high positive voltage and accelerated in an electric field. Since the deposition source is located at the cathode, the process is also called cathode atomization. The atoms or molecules released from the target by the impingement of the argon ions are deposited on the substrate. During the sputtering process, other interactions occur between the argon plasma particles, the target, and the substrate, which can influence the quality of the coating. The uniformity of the coatings is generally good; the growth rate is typically 0.5 – 3 nm/s for d.c. sputtering. However, this rate is too low for bulk production. The growth velocity can be increased by a factor of ten or more by using the magnetron (microwave) sputtering technique, in which the intensity of the argon plasma is increased by focusing with a magnetic field.

Apart from metals (Ag, Cu, Au, Ni, Cr, Ti, etc.) also oxides, nitrides, and carbides can be applied with the sputtering or PVD method. This is realized by bringing reactive gas mixtures (e.g., Ar/O_2) into the reaction chamber (reactive deposition/sputtering).

In physical vapor deposition (PVD) processes the glass substrate need not be heated. However, a chemical bond is then generally not formed between the coating and the glass substrate; there is only a weak physical adhesion. Therefore PVD coatings on glass also are called soft coatings. Spray or CVD coatings or mostly hard coatings. PVD methods are used on a large scale for the application of sunlight protection and heat reflecting (low E) coatings on window glass. Sun-protecting coatings partially transmit the visible part of the spectrum (0.4 – 0.8 μm) and reflect solar IR radiation (0.8 – 2.5 μm).

Heat-reflecting coatings must be transparent throughout the whole solar spectrum (0.4 – 2.5 μm) to obtain the heat from the solar radiation, and reflective beyond 2.5 μm, where the spectral range of the thermal radiation at room temperature begins. In this way the heat in the room is not lost by thermal radiation through the windows.

Lacquers are used for decorative purposes, for example, in Christmas ornaments and electric light bulbs.

Tin oxide and titanium oxide coatings are applied to container glass by spraying or by atmospheric CVD. In the latter case, the hot bottles pass through a reaction tunnel with an oxidizing atmosphere, in which the coating material (e.g., $SnCl_4$) is present as a vapor. In this case, deposition of SnO_2 will take place.

The CVD process is also used industrially in combination with the float-glass process. In the reducing atmosphere of the tin bath section, Si layers are formed on the glass surface by treatment with silane (SiH_4). The Si coatings have a thickness of about 30 nm and function as sun protection mainly by reflection for wavelengths up to 1 μm. They do not reflect heat because of the decreasing reflection of the Si layer at higher wavelengths ($\gg$ 1 μm).

In oxidizing atmospheres SnO_2 layers on the glass surface (Pilkington) can be formed from $SnCl_4$.

7. Uses

7.1. Silica and Silica-like Glass

Pure vitreous silica is the only industrially important single-component glass. The important characteristics of this material are its extremely low coefficient of thermal expansion, high operation temperature, good chemical durability, and excellent UV transparency (Section 2.2.1). The principal drawback is the high production cost resulting from the high melting temperature (>2300 K) required to obtain transparent material, which limits the number of economical applications. Nevertheless, several important uses exist. A more detailed discussion of applications of silica is given in [131].

Chemical purity, good chemical durability, and thermal shock resistance make vitreous silica the glass of choice in chemical laboratories. Thermal shock resistance is important when making sleeves around high-temperature thermocouples. Transparency to far UV radiation and low thermal expansion are the most important properties in optical applications: e.g., lenses, incandescent tungsten halogen cycle lamps, solar cell covers, high optical quality vitreous silica (Corning Code 1940), windows for space vehicles, and mirror blanks. Another important application is thin-film dielectric coating in integrated circuits. The refractory nature of silica (SiO vapor pressure <0.87 Pa in vacuum, and 3.9 mPa at 20.3 kPa partial oxygen pressure and 2000 K) and its inertness make it the laboratory ware of choice in the semiconductor industry.

Less critical components are made of silica-like material (Vycor), produced at ca. 1800 K; the process consists of melting 65 – 86 wt % SiO_2, 10 – 30 wt % B_2O_3, and 4 – 5 wt % Na_2O. Glasses with these compositions exhibit phase separation into a silica-rich and a sodium borate-rich phase at ca. 870 K. The latter is leached out in the presence of hot acid, resulting in a porous "thirsty" glass with about 96 wt % SiO_2 and 3 wt % B_2O_3. Heat treatment of this silica-rich skeleton results in a dense, fully transparent glass, used as laboratory ware.

Porous glasses have been made by (1) the viscous sintering of powdered glass, with or without soluble fillers, or (2) by heating and leaching certain phase-separable borosilicate

compositions. The latter method gives materials with controlled, uniform pore diameters of 4 – 300 nm. Applications include micro- and ultrafiltration, chromatography packing, desiccants, and substrates for immobilization of enzymes used in biotechnology. A typical 96 % SiO_2 Vycor-type porous glass has 4 nm diameter pores, a pore volume of 28 %, and a specific surface area of 250 m^2/g.

Specialty glasses, with thermal expansion even lower than that of fused silica, have been developed for use as astronomical mirror blanks. In addition to silica glass, they contain TiO_2, ZrO_2, and B_2O_3 or Al_2O_3. Inclusion of titania and zirconia (ZrO_2) results in a continuous decrease in the coefficient of thermal expansion.

7.2. Flat Glass

7.2.1. Reflective Glass

Reflective glass is covered with a metal or metallic oxide coating on one surface, furnishing high reflectivity in the desired wavelength range. A coating that provides maximum infrared reflectance reduces heat losses through windows. The coatings are produced by wet chemical, vacuum, or pyrolytic deposition. In *wet chemical deposition* the metal is precipitated from solution onto the surface of the glass. These coatings are fragile and must be protected after application. *Vacuum deposition* involves evaporation of the metal by electric heating in a vacuum chamber. Such coatings are not as fragile as chemically deposited coatings, but the glass cannot be tempered or heat strengthened after coating. *Pyrolytic deposition* involves spraying a solution containing the desired metal oxide onto the glass, which is held near red heat. The oxide is integrated into the surface and is extremely durable. Glass coated in this manner can be tempered and may be exposed to weather without a protective coating.

7.2.2. Laminated Glass

Laminated glass is a sandwich consisting of a plastic layer, usually a poly(vinyl butyral), between two or more glass sheets. Automobile windshields are made of laminated glass designed to be thin enough to flex slightly, thereby reducing the risk of injury in the event of collision. The plastic inner layer may be tinted to act as an IR filter. Laminated glass is more resistant to fracture than monolithic glass, and in the event of fracture, the plastic layer holds the glass fragments in place and absorbs part of the impact energy. Three-layer laminated glass is used for material requiring high impact resistance.

7.2.3. Solar Glass

Solar glass is low iron oxide soda lime glass. The IR transmittance of such glass may be as high as 93 %. This glass is used in solar collectors for heating water and in passive solar-heating systems for buildings.

7.3. Laboratory Glassware

Many laboratory items (e.g., beakers, Petri dishes, reaction flasks, volumetric equipment, ground joints, stopcocks, fritted ware, test tubes, and thermometers) are made from glass. The particular composition and thermal expansion of these glasses are given in Table 29. The corresponding properties of soda lime glass are included for comparison.

Borosilicate glass has the lowest coefficient of thermal expansion of all glass produced in bulk. Its resistance to thermal failure is high, and therefore, borosilicate objects can be heated and cooled more rapidly than high-expansion glasses. Aluminoborosilicate glass also has a high chemical durability with an appreciably larger expansion coefficient. Fused silica and 96 % silica glassware can be used at still higher temperature and will, of all the glasses mentioned here, best withstand severe change in temperature. Beakers of 96 % silica can be heated to redness and dropped into ice water without breaking. The four glasses listed above have few chemical constituents, have a low alkali-metal content, and are free from heavy-metal oxides, a distinct advantage in chemical analysis.

Table 29. Characteristic thermal properties of laboratory glassware

Glass type	Thermal coefficient expansion, 10^7 K^{-1}	Maximum operating temperature, °C	
		Continued use	Short-term use
Borosilicate, low expansion	32.5	510	
Aluminoborosilicate	55	520	
Soda lime	80 – 90	500	
96 % Silica	8	900	1200
Fused silica	5.5	1000	1350 – 1450

7.4. Light-Sensitive Glass

7.4.1. Photosensitive Glass

The sensitivity of some chemicals to light has been known since the time of DAGUERRE (1839). Characteristics of such chemicals have been reviewed [132]. An important process for glass applications involves cleavage of a molecule of silver halide under the influence of light, resulting in its reduction to metallic silver. Photolysis is used to produce photosensitive glass, the process consists principally of irreversible photoinduced homogeneous nucleation.

Photosensitive glasses contain a small amount of silver. Ultraviolet light or X-rays are of sufficiently high energy to reduce the silver ion and mobilize the homogeneously dispersed silver atoms, causing the silver to cluster in small particles. In Fotoform, lithium silicate is added to photosensitive glass. Subsequent heat treatment of Fotoform results in an opal glass due to the growth of lithium meta- or disilicate crystals that use the silver nuclei as substrates. Masking the glass before exposure produces, after heat treatment, a pattern which duplicates that of the mask. The exposed glass area dissolves, because of the presence of lithium silicate, approximately 14 times faster than the unexposed area; patterns can, therefore, be etched out, leaving void areas where the glass was exposed to radiation.

The principal use of photosensitive glasses lies in the manufacture of intricate and precise patterns. For example, as many as 250 000 holes, each 0.23 mm in diameter, have been etched on an area of 654 mm^2.

7.4.2. Photochromic Glass

Photochromism is the change in color of a solid upon irradiation with light of certain wavelength. This process can be reversible or irreversible. Most common practical applications use reversible photochromic materials [133]. The effect is based on the photolysis of a silver halide. In contrast to photosensitive glasses, however, after rupture of the halogen – silver bond, the halide ion is not sufficiently mobile to move away from the silver atom and the bond is reestablished during darkness. Thus the system is excited by light and, in darkness, reverts to its ground state.

The principal application of this effect is in photochromic lenses, with or without a corrective prescription. Corning photochromic filter (CPF) lenses have specially formulated compositions. These lenses are fired in hydrogen, turn yellow or red, maintain their photochromic character, and cut out blue light. They are made for people with diseases of the cornea or retina.

7.5. Display Devices

A cathode-ray tube (CRT) consists of three parts: a glass envelope or bulb, a phosphor screen, and an electron gun. The bulb also consists of three parts: the panel on which the phosphor screen is deposited, the funnel, and the neck in which the electron gun is sealed. These three parts are produced separately, and two of them are subsequently flame sealed with a solder glass (frit). The panel is usually pressed, the funnel is either pressed or centrifugally cast, and the neck is made from glass tubing.

Cathode-ray tubes have a number of built-in safety features. To prevent implosion, bulbs are designed to withstand an external pressure

of at least 3 atm (303 kPa); in addition, a tension band may be placed around the periphery of the panel skirt. To prevent exposure of the viewer to X-rays, the glass composition and wall thickness of the bulb are adjusted to maximize the 20.6-kV linear X-ray absorption coefficient of the glass; for Asahi – Corning color TV glass, this is currently 28.5 cm^{-1} for panels and 64.8 cm^{-1} for funnels (see also → Display Technology, Chap. 4.).

7.6. Glass Fibers

The basic products of the glass-fiber industry are unbonded and bonded wool, mats, staple fiber, cut strands, textile strands, sliver (i.e., parallel, overlapping, and untwisted bundles of noncontinuous or short-length fibers), and molded products such as pipe insulation or ceiling tile (see also → Fibers, 5. Synthetic Inorganic, Chap. 2.).

Glass-fiber wool insulation can reduce heat loss to one-half to one-fifth. Typical overall coefficients of heat transmission, or *U* values, for three wall structures are collected in Table 30.

Table 30. Typical *U* factors (heat conductivity) for wall structures

Type of construction	*U*, $W\,m^{-2}\,K^{-1}$	
	Uninsulated	Insulated
Wood frame	0.6 – 2.4	0.23 – 0.35 ∗
Masonry	1 – 4	0.4 – 0.4 ∗∗
Concrete block	1 – 3	0.32 – 0.45 ∗∗

∗ Based on 10-cm-thick insulation.
∗∗ Based on 7-cm-thick insulation.

7.7. Molded Optics

The impetus behind the fabrication of molded lenses is the production of cheap optical lenses for mass-produced cameras or optical memory disks. The challenge is to produce a bulk optic lens with a smooth surface and proper dimensions directly by molding only, bypassing the finishing stage. Four parameters are important in this process: (1) the materials from which the mold is made, (2) the variations in the mold, (3) the molding process itself, and (4) the size of the molded product. Glass compositions used in this process can vary over a wide range from fused silica to lead fluorophosphate. The Kodak company is most advanced in this technology, producing millions of aspheric lenses per year.

7.8. Glasses for Nonlinear Optical Devices

Glasses for nonlinear optical devices represent the frontier of research in glass chemistry and physics. The nonlinear effect in glass consists of a change in the refractive index of the glass as a function of the intensity of incident light. A necessary, but not sufficient, condition seems to be a high-index glass [134]. This nonlinear effect is small in standard glass but, sampled over a large volume as in optical waveguides, sufficiently large to function as a switching device. The aim is to increase photonic switching and signal processing by producing optical transistors that operate in the range of 10^{-12} s (ps) to 10^{-15} s (fs), compared with standard transistors, which operate in the range of 10^{-6} – 10^{-9} s (μs – ns).

8. Economic Aspects [143–147]

World production of glass is estimated at 80×10^6 t/a. The market shares (expressed in mass %) of major applications are shown in 48.

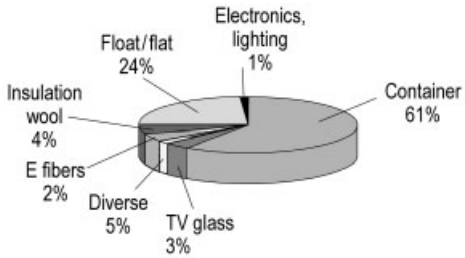

Figure 48. Market shares of major glass applications

The most important glass producing regions in the world are the EU with about 25.7×10^6 t of glass production in 1995 (turnover, including secondary glass manufacture, ca. $\$ 40 \times 10^9$); the United States with 13.5×10^6 t of glass in 1994; and Japan and Korea with about 8×10^6 t/a of glass. Glass production is growing in China, South America, and Turkey. The total turnover of primary glass products worldwide is estimated at about $\$ 130 \times 10^9$ annually.

In the United States the 1994 glass production value amounted to \$ 24×10^6.

The principal container glass producers are: Owens Brockway (USA), Consumer Glass (Canada), Ball-Foster (USA, owned by Saint Gobain), St. Gobain, Gerresheim (Germany), Heye Glas (Germany), BSN Emballage (France), PLM (Sweden), Yamamura Glas (Japan), Vetropack (Switzerland), Oberland Glas (Germany, owned by St. Gobain), Sise-Cam (Turkey), Nienburger Glass (Germany), ACI (Australia, owned by Owens Brockway), AVIR (Italy, owned by Owens Brockway).

Glass containers are used for food, beverages, liquor and wine, beer, perfume, pharmaceutical, and cosmetic products. The total container glass production in 12 countries (Belgium, Germany, France, Italy, The Netherlands, Ireland, Portugal, Spain, Denmark, Greece, Luxemburg, UK) of the European Union in 1994 was 16.4×10^6 t, in the United States about 13.5×10^6 t (1994); and in Japan 2.6×10^6 t (1990). Total annual world production is about $45-50 \times 10^6$ t container glass. An average 3.5 % production growth has been accomplished since 1986 in the EU container glass industry.

There are a very large number of domestic and tableware crystal glass producers worldwide, but the most important companies are: Durand-Verrerie Cristallerie d'Arques (France), Bormioli Rocco (Italy), Libbey Glass (USA), Waterford Crystal (Ireland), Kosta Boda (Sweden), Sise-Cam (Turkey), Nachtmann (Germany), Schott (Germany), Kedung (Indonesia), Crestalux (Argentina), Vitrocrisa (Mexico), Sasaki Glass (Japan), Anchor Hocking (USA). Annual production is estimated at 3.2×10^6 t/a. The US turnover for tableware glass products in 1996 was \$ 1.9 billion.

The major flat glass producers are: Pilkington (UK), Asahi (Japan), PPG Industries (USA), Glaverbel (Belgium), Saint Gobain (France), Ford (USA, since 1998: Visteon), Libbey Owens Ford (USA, owned by Pilkington), AFG (USA), Central Glass (Japan), Nippon Sheet Glass (Japan), Keumkang (South Korea), Hankuk (South Korea). The companies produce mainly float glass for automotive and architectural applications. China is an important producer of flat glass, with about 30 – 40 operating float furnaces. Float glass production in the United States in the mid-1990s was ca. 3.7×10^6 t/a. In the EU in 1995, 5.1×10^6 t of float glass was produced. Japanese float glass production in 1990 was about 2.7×10^6 t.

Mineral wool is produced by a few large companies and many small companies. The principal mineral wool producers worldwide are: Johns Manville (USA), Owens Corning Fiber Glass (USA), Certain Teed (USA, affiliated to Saint Gobain), Saint Gobain-Isover (France), Guardian Fiberglass (USA), Knauf Fiber Glass (USA), Rockwool (Denmark). The major markets for mineral wool are structural insulation, industrial equipment insulation, pipe insulation, and substitute for soils in agricultural applications.

Continuous fiberglass for textile or reinforcement applications is mainly produced by PPG FiberGlass (USA), Owens Corning Fiberglass (USA), Vetrotex (France, owned by St. Gobain), Johns Manville (USA), Hankuk Vetrotex (South Korea), Asahi Fiber Glass (Japan). The EU production of continuous fiberglass amounted to 462 000 t in 1995.

The advanced and specialty glass market includes optical waveguides for telecommunications, electronic devices such as thick-film dielectrics, resistors, thermistors, glasses for sealing ferrites in magnetic recording, electronic displays, glass nozzles for high speed printers, channel electron multipliers, high-performance composites such as alumina borosilicate glass for use in secondary structures in aircraft, medical or dental bioglasses (based on calcium phosphates), glass ceramics, and vitrified high-level nuclear waste.

Among the worldwide producers of specialty glasses and TV glasses are Schott (Germany), General Electric (USA), Corning (USA), Nippon Electronic Glass (Japan), Thomson (France), Philips Display Components (The Netherlands), Asahi (Japan), Samsung Corning (South Korea), and Hankuk Electric Glass (South Korea).

Glass for lighting applications, such as bulbs, tubes, and automotive lamps, are produced by Philips Lighting (The Netherlands), Osram Sylvania (Germany, USA), and General Electric (USA).

Sodium silicates, mainly intended for further chemical processing, are produced in Europe and United States by companies like Henkel (Germany), AKZO Philadelphia Quartz (The

Netherlands, USA), Rhône-Poulenc (France), and PPG (USA).

9. References

General References

1. *Ullmanns,* 4th ed. **12,** 317 – 366.
2. *Kirk-Othmer,* **11,** 807 – 880.
3. P. Balta, E. Balta: *Introduction to the Physical Chemistry of the Vitreous State,* Abacus Press, Tonbridge Wells, England, 1976.
4. H. H. Dunken: *Physikalische Chemie der Glasoberfläche,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1981.
5. G. H. Frischat: "Ionic Diffusion in Oxide Glasses," *Trans. Tech. Publications,* Bay Village, Ohio, 1975, p. 181.
6. G. W. McLellan, E. B. Sand: *Glass Engineering Handbook,* McGraw-Hill, New York 1984.
7. A. Paul: *Chemistry of Glasses,* Chapman and Hall, New York 1982.
8. W. Trier: *Glass Furnaces,* Society of Glass Technology, Sheffield 1985.
9. E. T. Turkdogan: *Physiochemical Properties of Molten Slags and Glasses,* The Metals Society, London 1983.
10. W. Vogel: *Chemistry of Glass,* The American Ceramic Soc. Inc., Columbus, Ohio, 1985.
11. C. H. Drummond, III, *47th Conference on Glass Problems,* The Ohio State University, Columbus, Ohio, 1986.
12. D. F. Bickord et al.: *Advances in the Fusion of Glass,* The American Ceramic Soc. Inc., Westerville, Ohio, 1988.
13. D. C. Boyd, J. F. McDowell: "Commercial Glasses," *Adv. Ceram.* **18** (1986) 229.
14. F. V. Tooley: *The Handbook of Glass Manufacture,* vol. I – II, Books for Industry Inc., New York 1974.
15. W. Giegerich, W. Trier: *Glass Machines,* Springer Verlag, New York 1969.
16. H. Scholze: *Glas – Natur, Struktur und Eigenschaften,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York 1977.
17. J. Wong, C. A. Angell: *Glass Structure by Spectroscopy,* Marcel Dekker, New York 1976.
18. F. C. Hawthorne: "Spectroscopic Methods in Mineralogy and Geology," *Rev. Mineral.* **18** (1988) 698.
19. S. T. Pantelides: *The Physics of SiO_2 and its Interfaces,* Pergamon Press, New York 1978.
20. R. K. Iler: *The Chemistry of Silica,* Wiley-Interscience, New York 1979.
21. E. R. Corey, J. Y. Corey, P. P. Gasper (eds.): *Silicon Chemistry,* Halsted Press, New York 1988.
22. W. A. Weyl: *Colored Glasses,* Society of Glass Technology, Sheffield, England, 1961.
23. R. G. Burns: "Mineralogical Applications of Crystal Field Theory," *Cambridge Earth Science Series,* Cambridge University Press, London 1970.
24. C. R. Bramford: "Colour Generation and Control in Glass," *Glass Sci. Technol.* **2** (1977) 224.
25. K. Nassau: *The Physics and Chemistry of Color,* Wiley-Interscience, New York 1983.
26. E. Fritsch, G. R. Rossman: "An Update on Color in Gems. Part 2. Colors Involving Multiple Atoms and Color Centers," *Gems Gemol.* **24** (1988) no. 1, 3 – 15.
27. E. Fritsch, G. R. Rossman: "An Update on Color in Gems. Part 3. Colors caused by bandgaps and physical phenomena," *Gems Gemol.* **24** (1988) no. 2, 81 – 93.
28. E. Fritsch, G. R. Rossman: "An Update on Color in Gems. Part 1. Introduction and Colors caused by Dispersed Metal Ions," *Gems Gemol.* **23** (1987) no. 3, 126 – 139.
29. J. W. H. Schreurs, R. H. Brill: "Iron and Sulfur related colors in ancient glasses," *Archaeometry* **26** (1984) 199 – 209.
30. W. Kauzmann: "The Nature of the Glassy State and the Behavior of Liquids at Low Temperature," *Chem. Rev.* **48** (1948) 219 – 256.
31. D. R. Uhlmann: "Glass Formation, a Contemporary View," *Am. Ceram. Soc.* **66** (1983) no. 2, 95 – 100.
32. A. R. Ubbelohde: *Melting and Crystal Structure,* Clarendon Press, Oxford 1965.
33. H. Rawson: *Inorganic Glass-Forming Systems,* Academic Press, New York 1967.
34. G. W. Morey: *The Properties of Glass,* American Chemical Society Monograph 77, Reinhold, New York 1938.
35. G. Tamman: *States of Aggregation,* van Nostrand, New York 1925.
36. W. A. Weyl, E. C. Marboe: *The Constitution of Glasses, A dynamic Interpretation,* **vols. 1,** 2, Interscience, New York 1962 – 1964.
37. R. Balian, R. Maynard, G. Toulouse: *Ill-Condensed Matter,* Les Houches, Session 31, North Holland, New York 1978.
38. *Phase Diagrams for Ceramists,* **vols. I – V,** American Ceramic Society Inc., Columbus, Ohio, 1973 –1985.

39. O. V. Mazurin: *Properties of Glasses and Glass-forming Melts,* **vols. I – IV,** Nauka Publishing House, Leningrad 1973 – 1980.
40. O. V. Mazurin, M. V. Streltsina, T. P. Shvaiko-Shvaikovskaya: *Handbook of Glass Data,* Part A, B, C, Elsevier, Amsterdam 1983 – 1987.
41. N. P. Bansal, R. H. Doremus: *Handbook of Glass Properties,* Academic Press, New York 1986, pp. 680.
42. R. S. Carmichael: *Handbook of Physical Properties of Rocks,* vols. I – II, CRC Press, Boca Raton, Florida. 1982.
43. S. P. Clark: *Handbook of Physical Constants,* Geology Society America, New York.
44. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, New York 1973.
45. M. W. Chase et al.: *JANAF Thermochemical Tables,* Part 1, Al – Co; Part 2, Cr – Zr; J. Chem. Ref. Data, vol. 14, Suppl. 1 (1985) 1856.
46. O. Kubachewski, E. L. Evans: *Metallurgical Thermochemistry,* Pergamon Press, New York 1958.
47. R. Robie, B. S. Hemingway, J. R. Fischer: "Thermodynamic properties of minerals and related substances at 298.15 K and 1 bar (10^5 Pa) pressure and at higher temperatures," *Seological Survey Bulletin 1452* , U. S. Government Printing Office, Washington 1979, pp. 456.
48. F. Liebau: *Structural Chemistry of Silicates,* Springer Verlag, New York 1985.
49. J. O. Nriagu, P. B. Moore: *Phosphate Minerals,* Springer Verlag, New York 1984.
50. J. R. Van Wazer: *Phosphorus and its Compounds,* vol. I, Interscience, New York 1958.
51. *Slag Atlas,* Verlag Stahleisen mbH, Düsseldorf 1981.
52. H. H. Lewis (ed.): *Glasses and Glass Ceramics,* Chapman and Hall, London-New York 1989.
53. O. V. Mazurin (ed.): *Glass '89. Survey Papers of the XVth International Conference on Glass,* Leningrad 1989.

Specific References

54. G. W. Morey: "The Properties of Glass," *ACS Monogr.* **77** (1938) 561.
55. J. T. Randall: *The diffraction of X-rays and Electrons by Amorphous Solids, Liquids, and Gases,* Wiley, New York 1934, pp. 290.
56. P. P. v. Weimarn: "Zur Lehre von den Zuständen der Materie," *Z. Chem. Ind. Kolloide* **2** (1907) 81, 218 – 233.
57. A. A. Lebedev: "On Polymorphism and Annealing of Glass," *Opt. Inst. Leningrad* **2** (1921) no. 10, 1 – 18.
58. N. Valenkov, E. Poray-Koshitz: "X-ray investigation of the glassy state," *Z. Kristallogr.* **95** (1936) 195 – 229.
59. W. H. Zachariasen: "The Atomic Arrangement in Glass," *JACS* **54** (1932) 3841 – 3851.
60. V. M. Goldschmidt: "Geochemical Distribution Laws of the Elements. VIII. Researches on the Structure and Properties of Crystals," *Skr. Nor. Vidensk.-Akad.* (Kl.) 1: Oslo. I. Matemat.-Naturv. Kl. 8 (1926) 7 –156.
61. A. Dietzel: "The cation field strengths and their relation to devitrifying processes, to compound formation and to melting points of silicates," *Z. Elektrochem.* **48** (1942) 9 – 23.
62. A. Smekal: "über die Natur der glasbildenden Stoffe," *Glastech. Ber.* **22** (1949) 278 – 289.
63. W. A. Weyl: "Une nouvelle méthode d'étude de la chimie de l'état solide et son application aux problèmes des industries de silicates," *Silic. Ind.* **23** (1959) part 4, 518 – 527.
64. J. M. Stevels: "Progress in the theory of the physical properties of glass' in *Monographs On the Progress of Research in Holland during the War,* vol. 20, Elsevier, New York, pp. 104.
65. M. L. Huggins: "The structure of glasses," *J. Am. Cer. Soc.* **38** (1955) 172.
66. L. W. Tilton: "Ionic model for silica glass," *J. Res. Nat. Bur. Stand.* **59** (1957) 139 – 154.
67. G. Engelhardt et al.: "Zur Abhängigkeit der Struktur der Silicatanionen in wässrigen Natrium-Silicatlösungen vom Na : Si Verhältnis," *Z. Anorg. Allg. Chem.* **418** (1975) 17 – 28.
68. B. H. W. S. De Jong: "Uneconomic Bonding in Glasses," *Nature* (London) 330 (1987) 422.
69. B. H. W. S. De Jong, C. M. Schramm, V. E. Parziale: "^{29}Si Magic Angle Spinning NMR on Local Silicon Environments in Amorphous and Crystalline Lithium Silicates," *J. Am. Chem. Soc.* **106** (1984) 4396 – 4402.
70. K. D. Keefer: "The Distribution of Non-Bridging Oxygens and the Structure of Silicate Melts," *14th Meeting of the International Mineralogical Association,* Stanford, Ca., 1986, p. 139.
71. W. Cuijpers et al.: "29-Si MAS NMR shows large chemical shift variation for Q-2 silica polymers in 24-er and 12-er crystalline Sodium

Yttrium silicates," Int. NMR Conference, 1989, Les Morsines (France), in press.
72. D. J. Thouless: "Percolation and Localization" in R. Balian, R. Maynard, G. Toulouse (eds.): *Ill-condensed Matter,* North Holland, Amsterdam 1978, pp. 1 – 62.
73. J. M. Ziman: *Models of Disorder,* Cambridge University Press, Cambridge 1979, p. 525.
74. N. P. Bansal, R. H. Doremus: *Handbook of Glass Properties,* Academic Press, New York 1986, p. 680.
75. R. Robie, B. S. Hemingway, J. R. Fisher: "Thermodynamic properties of minerals and related substances at 298.15 K and 1 bar (10^5 Pascals) pressure and at higher temperatures," *Geological Survey Bulletin 1452,* United States Government Printing Office, Washington 1979, p. 456.
76. J. B. Murdoch, J. F. Stebbins, I. S. E. Carmichael: "High resolution 29-Si NMR study of silicate and aluminosilicate glasses: the affect of network modifying cations," *Am. Mineral.* **70** (1985) 332 – 343.
77. J. M. Thomas et al.: "The evaluation of nonequivalent tetrahedral sites from silicon-29 NMR chemical shifts in zeolites and related aluminosilicates," *Chem. Phys. Lett.* **102** (1983) 158 – 162.
78. L. F. Gladden, T. A. Carpenter, S. R. Elliot: "^{29}Si MAS NMR studies of the spin-lattice relaxation time and bond-angle distribution in vitreous silica," *Philos. Mag. B* **35** (1986) L81 – L87.
79. B. H. W. S. De Jong, G. E. Brown: "Polymerization of silicate and aluminate tetrahedra in glasses, melts, and aqueous solutions – I. Electronic structure of $H_6Si_2O_7$, $H_6AlSiO_7^{1-}$, and $H_6Al_2O_7^{2-}$," *Geochim. Cosmochim. Acta* **44** (1980) 491 – 511.
80. E. T. Turkdogan: *Physicochemical Properties of Molten Slags and Glasses,* The Metals Society, London 1983, p. 516.
81. S. P. Clark: *Handbook of Physical Constants,* "The Geological Society of America Memoir 97," The Geological Society of America, Inc., New York 1966, p. 587.
82. P. Balta, E. Balta: *Introduction to the Physical Chemistry of the Vitreous State,* Abacus Press, Tonbridge Wells, England, 1976, p. 287.
83. B. L. Sherriff, H. D. Grundy: "Calculations of ^{29}Si MAS NMR chemical shift from silicate mineral structure," *Nature* (London) 332 (1988) 819 – 822.
84. N. James, E. Oldfield: "Prediction of Silicon-29 Nuclear Magnetic Resonance Chemical Shifts Using a Group Electronegativity Approach: Applications to Silicate and Aluminosilicate Structures," *JACS* **107** (1985) 6769 – 6775.
85. M. P. Teter: "Glass Structure using Quantum Molecular Dynamics, 231," *89th Annual Meeting American Ceramic Society,* Pittsburgh 1987.
86. L. Shartsis, S. Spinner, *J. Res. Nat. Bur. Stand* **46** (1951) 385 – 390.
87. P. J. Bray, J. G. O'Keefe: "Nuclear Magnetic Resonance Investigations of the Structure of Alkali Borate Glasses," *Phys. Chem. Glasses* **4** (1963) 37 – 46.
88. R. Brueckner, H. U. Chun, H. Goretzki, M. Sammet: "XPS Measurements and Structural Aspects of Silicate and Phosphate Glasses," *J. Non Cryst. Solids* **42** (1980) 49 – 60.
89. P. S. Bagus, C. W. Bauschlicher: "Core Binding Energy for Free Negative Ions of Oxygen: O^0 to O^{2-}," *J. Electron Spectrosc. Relat. Phenom.* **20** (1980) 183 – 190.
90. Y. Kaneko, Y. Suginohara: "Fundamental Studies on Quantitative Analysis of O^0, O^-, and O^{2-} Ions in Silicate by X-Ray Photoelectron Spectroscopy," *J. Nippon Metall. Soc.* **41** (1977) no. 1, 371 – 380.
91. L. S. Dent Glasser, R. A. Howie, R. M. Smart: "The structure of lead 'orthosilicate', 2 PbO · SiO2," *Acta Crystallogr. Sect.* **B 37** (1981) 303 – 306.
92. K. Kato: "Die OD Struktur von Bleisilicat Pb_2SiO_4 und Bleisilicat-Germanat-Mischkristall $Pb_2(Si,Ge)O_4$," *Acta Crystallogr. Sect.* **B 36** (1980) 2539 – 2545.
93. R. M. Almeida: "Halide Glasses for Infrared Fiberoptics," *NATO ASI Ser* . Ser. E 123 (1987) 412.
94. P. A. Tick, P. L. Bocko: "Optical Fiber Materials" in S. Musikant (ed.): *The Handbook of Optical Materials,* vol. 3, 1988.
95. G. Tamman: *States of Aggregation,* van Nostrand, New York 1925, p. 297.
96. G. F. Neilson, M. C. Weinberg: "A Test of Classical Nucleation Theory: Crystal Nucleation of Lithium Disilicate Glass," *J. Non Cryst. Solids* **34** (1979) 137 – 147.
97. K. Nakagawa, T. Izumitani: "Relationship between phase separation and crystallization in Li_2O, 2.5 SiO_2 glass and a lithium silicate containing a large amount of titanium oxide," *Phys. Chem. Glasses* **10** (1969) no. 5, 179 – 184.

98. P. F. James: "Kinetics of Crystal Nucleation in Silicate Glasses," *J. Non Cryst. Solids* **73** (1985) 517 – 540.
99. A. C. J. Havermans, H. N. Stein, J. M. Stevels: "Critical Cooling Rates in Alkali Silicate Systems," *J. Non Cryst. Solids* **5** (1970) 66 – 69.
100. A. Winckelmann, O. Schott: "Zug und Druckfestigkeit neuer Gläser: Chemische Zusammensetzung und Temperatur," *Ann. Phys. (Leipzig)* **51** (1894) 730 – 746.
101. *Slag Atlas,* Verlag Stahleisen mbH, Düsseldorf 1981, p. 282.
102. H. E. Hagy in L. D. Pye, H. J. Stevens, W. C. LaCourse (eds.): "Rheological behavior of glass," *Introduction to Glass Science,* Plenum, London 1972, pp. 343 – 371.
103. B. H. W. S. De Jong, C. M. Schramm, V. E. Parziale: "Polymerization of Silicate and Aluminate Tetrahedra in Glasses, Melts, and Aqueous Solutions–IV. Aluminum Coordination in Glasses and Aqueous Solutions and Comments on the Aluminum Avoidance Principle," *Geochim. Cosmochim. Acta* **47** (1983) 1223 – 1236.
104. C. F. Baes, R. E. Mesmer: *The Hydrolysis of Cations,* Wiley, New York 1976. C. F. Baes, R. E. Mesmer: "The Thermodynamics of Cation Hydrolysis," *Am. J. Sci.* **281** (1981) 935 – 962.
105. I. Kushiro: "On the Nature of Silicate Melt and its Significance in Magma Genesis: Regularities in the Shift of the Liquidus Boundaries Involving Olivine, Pyroxene, and Silica Minerals," *Am. J. Sci.* **275** (1975) 411 – 431.
106. B. H. W. S. De Jong, C. M. Schramm, V. E. Parziale: "A ^{29}Si MAS NMR Study on the Role of Additives on the Silica Polymerization in the CaO-SiO_2 System," *Geol. Soc. Am. Fall Meeting* **14** (1982) no. 7, 474.
107. G. Gehlhof, M. Thomas: "Die physikalischen Eigenschaften der Gläser in Abhängigkeit von der Zusammensetzung III. Die Viskosität der Gläser," *Z. Tech. Phys.* **7** (1926) 261 – 278.
108. Y. Bottinga, D. Weill: "The Viscosity of Magmatic Silicate Liquids: A Model Calculation," *Am. J. Sci.* **272** (1972) 438 – 475.
109. S. English, W. E. S. Turner: "The Density of Soda-Lime-Magnesia Glasses and some Observations on the Calculation of Density," *J. Soc. Glass Technol.* **6** (1922) 228 – 231.
110. R. M. Hazen, C. T. Prewitt: "Effects of Temperature and Pressure on Interatomic Distances in Oxygen-Based Minerals," *Am. Mineral.* **62** (1977) 309 – 315.
111. H. Scholze: *Glas,* Viehweg & Sohn, Braunschweig 1977, p. 370.
112. M. S. Mizzoni, W. H. Dumbaugh, B. P. Tyndell: "Mixed Halide Effect in Lead Silicate Glasses," *Am. Ceram. Soc. Bull.* **67** (1988) no. 9, 1494.
113. W. Stumm, J. J. Morgan: *Aquatic Chemistry,* Wiley-Interscience, New York 1981, p. 780.
114. L. W. Cary, B. H. W. S. De Jong, W. E. Dibble: "A ^{29}Si NMR Study of the Silica Species in Dilute Aqueous Solution," *Geochim. Cosmochim. Acta* **46** (1982) 1317 – 1320.
115. H. Scholze: "Glass-water interactions," *J. Non-Cryst. Solids* **102** (1988) 1 – 11.
116. J. S. Falcone: "Soluble Silicates," *ACS Symp. Ser.* **194** (1982) 364.
117. W. H. Morrison, *J. Coating Technol.* **57** (1985) 55 – 65.
118. G. Parkes, *Chem. Rev.* **65** (1965) 177 – 198.
119. H. Kamiya, K. Shimokata: "The Role of Salts in the Dissolution of Powdered Quartz," *Proceedings Intern. Symp. Water-Rock Interactions,* Prague 1974.
120. F. Liebau: *Structural Chemistry Silicates,* Springer Verlag, New York 1965, pp. 347.
121. D. C. Allan, M. P. Teter, *Phys. Rev. Lett.* **59** (1987) no. 10, 1136 – 1139.
122. K. L. Bedford et al.: "Electronic Structure of Berylium Fluoride," *Phys. Rev. B.* **27** (1983) 2446 – 2457.
123. American Conference of Governmental Industrial Hygienists (1987 – 1988): Threshold Limit Values and Biological Exposure Indices, ACGIH, Cincinnati, Ohio, p. 114.
124. N. F. Borrelli, D. W. Hall, H. J. Holland, D. W. Smith: "Quantum confinement effects of semiconducting microcrystallites in glass," *J. Appl. Phys.* **61** (1987) 5399 – 5409.
125. W. Vogel: *Chemistry of Glass,* American Ceramic Society Inc., Columbus, Ohio, USA, 1985, p. 325.
126. H. N. Mills in C. G. Bergeron (ed.): "Use of Fining Agents," *Proceedings of the 46th Conference on Glass Problems,* The American Ceramic Soc. Inc., Columbus, Ohio, 1986, pp. 76 – 81.
127. V. G. Ciaraldi, J. N. Boyd in L. V. Cralley, L. J. Cralley (eds.): "Specialty Glasses," *In-Plant Practices for Job-Related Health Hazards Control,* vols. I, II,Wiley-Interscience, New York 1989 (in press).
128. R. Drake in C. H. Drummond III (ed.): "A Regulatory Update for the Glass Industry,"

47th Conference on Glass Problems, 1986, pp. 150 – 157.
129. R. A. Drake: "Environmental Issues that have an Impact on Glass Manufacturers," *Glass Ind.* **69** (1988) no. 6, 14 – 16.
130. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., Wiley-Interscience, New York 1987.
131. *Kirk-Othmer,* **20,** 782 – 817.
132. R. J. Araujo in D. C. Boyd, J. F. McDowell (eds.): "Photosensitive Glasses," *Advances in Glass Ceramics,* vol. 18, The American Ceramic Society Inc., Columbus, Ohio, 1986, pp. 151 – 155, 229.
133. G. P. Smith: "Chameleon in the sun: photochromic glasses," *IEEE Spectrum* **3** (1966) 39 – 47.
134. D. W. Hall et al.: "Measurements of X_3 glasses," *Conference on Lasers and Electrooptics,* Anaheim, CA, 1988.
135. H. R. Golob, E. L. Swarts, "Disproportionation of SO_2 in Bubbles Within Soda-Containing Glasses", *J. Am. Ceram. Soc.* **67** (1984) 564 – 567.
136. H. A. Schaeffer, "Recycling of Cullet and Filter Dust–Challenges to the Melting of Glass″, Chimica Chronica, New Series, 23 (1994) pp. 21 – 29; Proceedings of the International Symposium on Glass Science and Technology, Athens 6 – 8 October 1993.
137. W. Muschick, E. Muijsenberg, "Round Robin for Glass Tank Models", *Glastechn. Ber. Glass Sci. Technol.* **71** (1998) no. 6, 153 – 156.
138. R. Beerkens, "Heutige Möglichkeiten zur Modellierung von Glasschmelzofen, Voraassagen zur Qualität des Glasschmelzprozesses", *Glastech. Ber. Glass Sci. Technol.* **71** (1998) N35 – N47.
139. H. Kobayashi, R. G. C. Beerkens, P. Ercole, R. Barbiero, "Emissions of Particulates and NO from Oxy-Fuel Fired Glass Furnaces" in *Fundamentals of Glass Science and Technology*, Proceedings of the Second Conference of the European Society of Glass Science and Technology, Venice, Italy, 21 – 24 June 1993, pp. 45 – 50.
140. J. L. Barton, "Innovation in Glass Melting",*Glass Technol.* **34**(1993) 170 – 177.
141. H. Pieper, "Der heutige Glasschmelzofenbau – Herausforderung zur Energieeinsparung und Umweltverträglichkeit, *Glastech. Ber. Glass Sci. Technol.* **70** (1997) no. 8, N117 – N124.
142. F. Krämer, "Gasprofilmessungen zur Bestimmung der Gasabgabe beim Glasschmelzprozeß", *Glastech. Ber.* **53** (1980) no.7, 177 – 188.
143. J. L. Barton, A. Merlin, "The Glass Industry in Europe", *Ceram. Eng. Sci. Proc.* **14** (1993) 45 – 55.
144. I.-K. Byun, "Overview of the Glass Industry in Asia", *Eng. Sci. Proc.* **14** (1993) 56 – 67.
145. C. P. Ross, "Energy Benchmarking: A tool for Continuing Process Improvement for the Glass Industry", *Ceram. Eng. Sci. Proc.* **18**(1997) 95 – 108.
146. G. Maeyaert, "Mixed Fortunes for Europe's Glassmakers", *Glass* **73** (1996) no. 8, 305 – 306.
147. G. Maeyaert: "A Contrasted Landscape – The EU Glass Industry 1995 – 1996", *International Glass Review*, Spring/Summer 1996, p. 15 – 16

Glass Ceramics

Bernard H. W. S. De Jong, Corning Research, Corning, New York 14831, United States
Jane W. Adams, Corning Research, Corning, New York 14831, United States
Bruce G. Aitken, Corning Research, Corning, New York 14831, United States
James E. Dickinson, Corning Research, Corning, New York 14831, United States
Gerald J. Fine, Corning Research, Corning, New York 14831, United States

1. Introduction 1
2. Nucleation and Crystallization of Glass Ceramics 2
2.1. Homogeneous Nucleation 2
2.2. Heterogeneous Nucleation 3
2.3. Crystal Growth 4
2.4. Heat Treatment 5
3. Nature of Crystalline Phases 6
3.1. Metastability 6
3.2. Physical Properties 7
3.2.1. Thermal Properties 7
3.2.2. Transparency 9
3.2.3. Electrical Properties 9
3.2.4. Mechanical Properties 10
3.2.5. Durability 11
3.3. Microstructure 11
4. Types of Glass Ceramics 12
4.1. Silicate-Based Glass Ceramics . . . 12
4.2. Aluminosilicate Glass Ceramics . . 14
4.3. Fluorosilicate Glass Ceramics 16
5. References 17

1. Introduction

The development of glass ceramics is a comparatively recent advance in materials science. The phenomenal growth in a wide variety of applications for these materials since their discovery by Stookey in the late 1950s attests to their great utility.

Glasses are formed when molten materials are cooled and form a solid without crystallization (→Glass). Glass ceramics, however, are uniformly grained, crystalline materials which are made by controlled nucleation and crystallization of homogeneous glasses. Crystallization is induced by a nucleating agent and heat treatment. The major advantage of glass ceramics over conventional ceramics is that fully densified bodies with complex shapes can be manufactured using standard hot glass forming techniques. Glass ceramics also offer a variety of unique, uniform physical properties, including great strength, machinability, transparency, and thermal shock resistance that cannot be attained in either glass or conventional ceramics.

Literature on glass ceramics is not as extensive as that for glasses. The excellent recent reviews of Beall and coworkers [1–4] and Smith [5] together with the books written by Vogel [6] and Strnad [7] and the one edited by Lewis [8] give a proper overview of glass ceramic technology.

2. Nucleation and Crystallization of Glass Ceramics

The specific properties of glass ceramics are controlled by the physical properties of the crystalline phases developed in the glass upon heat treatment and also by their microstructure, i.e., the textural relationship between the crystalline phases and the residual glass. The uniformity and reliability of these properties is a consequence of the uniform grain size of the crystalline phases. Therefore, an important prerequisite for the formation of glass ceramics is efficient nucleation of the crystalline phases. The heat treatment (ceramming) of glasses to produce a glass ceramic typically involves two or

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_433

more steps: a lower temperature step to induce nucleation (*the nucleation hold*), and one or more higher temperature treatments to promote crystal growth (*crystallization hold*).

2.1. Homogeneous Nucleation

Homogeneous nucleation occurs when crystals are formed within a supersaturated fluid without addition of foreign particles. The parameters involved in crystal nucleation are shown in Equation (1) which expresses the steady-state homogeneous nucleation rate in a one-component supercooled system [8], [9]:

$$I = A \exp\left[-\left(W + \Delta G_{\mathrm{D}}\right)/kT\right] \tag{1}$$

Here I is the number of nuclei formed per unit volume, W is the free energy barrier to formation of a nucleus of critical size, ΔG_{D} the kinetic barrier to nucleation, k is Boltzmann's constant, and T is absolute temperature. The preexponential factor A can be expressed to a first approximation by

$$A = n_{\mathrm{v}}\left(kT/h\right) \tag{2}$$

where n_{v} is the number of formula units of the nucleating phase per unit volume of liquid and h is Planck's constant.

The parameter ΔG_{D} is usually expressed as the activation energy of diffusion in the Arrhenius equation

$$D = D_0 \exp\left(-\Delta G_{\mathrm{D}}/kT\right) \tag{3}$$

where $D_0 = kT\lambda^2/h$, λ being the jump distance of the atoms in the direction of diffusion. Assuming validity of the Stokes–Einstein relation, D can be expressed in terms of the viscosity η as $D = kT/6\pi\lambda\eta$. Thus $\exp(-\Delta G_{\mathrm{D}}/kT) = h/6\pi\lambda^3\eta$, and Equation (1) becomes

$$I = \left(A_{\mathrm{c}}T/\eta\right) \exp\left(-W/kT\right) \tag{4}$$

where A_{c} is $n_{\mathrm{v}}k/6\pi\lambda^3$. The free energy barrier to formation of a critical spherical nucleus W is

$$W = 16\pi\sigma^3 V_{\mathrm{m}}^2/3\Delta G^2 \tag{5}$$

where σ is the interfacial energy per unit area between crystal and liquid, V_{m} the molar volume of the crystalline phase, and ΔG the change in free energy per mole of material crystallized. Substitution into Equation (4) gives the final expression for the steady-state homogeneous nucleation rate as

$$I = \left(A_{\mathrm{c}}T/\eta\right) \exp\left(-16\pi\sigma^3 V_{\mathrm{m}}^2/3\Delta G^2 kT\right) \tag{6}$$

in which time does not appear explicitly because I does not vary with time in steady-state homogeneous nucleation.

Although exact evaluation of Equation (6) is not possible due to the lack of relevant thermodynamic data, interfacial energy values, and n_{v} data, this equation is still a useful guide for making adjustments to produce an efficiently nucleated glass ceramic. For example, it indicates that lowering the interfacial energy σ should have the largest effect on increasing nucleation efficiency. This might be accomplished by slight variations in glass composition, in particular by addition of WO_3 and MoO_3, which effectively reduce surface tension.

Similarly the number of nuclei per unit volume I will be larger if the change in free energy between liquid and crystalline precipitate is large. Although this free energy change is largely beyond experimental control, introduction of components which preferentially enter into solid solution with the crystalline phase may lower its free energy relative to the liquid and hence increase the number of nuclei.

Inspection of Equation (6) also suggests that increasing viscosity will decrease the nucleation efficiency because of impeded transport of nutrients to the nucleus. Lowering the viscosity at the optimal nucleation temperature is therefore desirable, although this is limited because the article may sag during the nucleation hold.

The least understood parameter in Equation (6) is n_{v}, the number of formula units of the nucleating phase per unit volume. The reason for this is a general lack of understanding of the relationship between glass and crystalline structure. Intuition suggests that structural similarity between the glass and the crystalline structure would accelerate nucleation because of the increased probability of forming a critical nucleus, hence increasing n_{v}.

An expression for the time-dependent nucleation rate I' is given by Kaschiev [10]:

$$I' = I\left[1 + 2\Sigma\left(-1\right)^m \exp\left(-m^2 t/\tau\right)\right] \tag{7}$$

where the summation over the index m is from one to infinity, $\tau = 48\sigma\lambda^5\eta/\pi V^2 \Delta G_v^2$ is the induction time, I is the steady-state nucleation rate, and t is time.

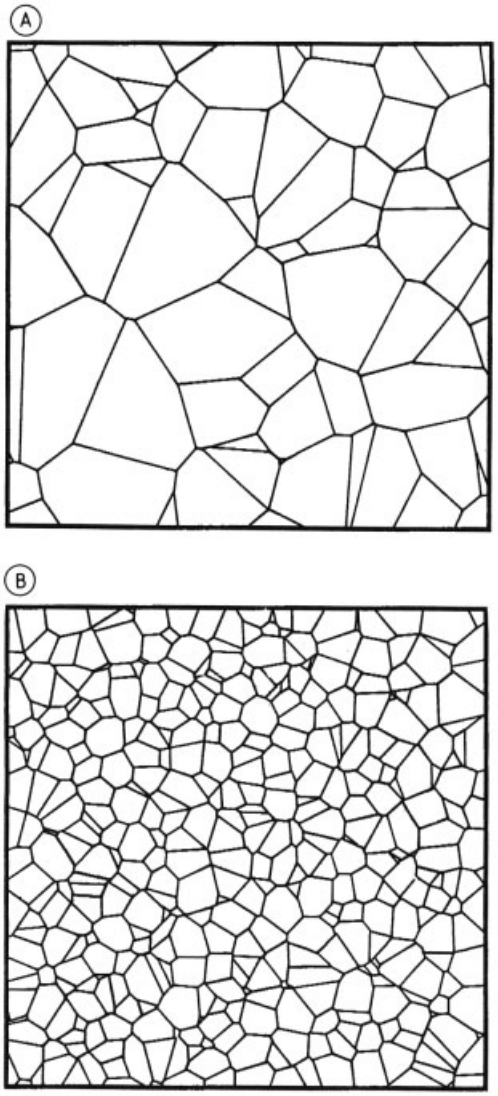

Figure 1. A two-dimensional slice through three-dimensional Voronoi cells calculated for a random distribution of nucleation sites separated by an average distance of 32 µm (A, 40 points) and 16 µm (B, 400 points) Courtesy of Leslie Button (Corning Incorporated).

2.2. Heterogeneous Nucleation

Homogeneous nucleation occurs in only a small number of glass systems, such as $Li_2O \cdot 2\,SiO_2$, $Na_2O \cdot 2\,CaO \cdot 3\,SiO_2$, $BaO \cdot 2\,SiO_2$, and $SiO_2 \cdot P_2O_5$. No industrially important glass ceramic systems nucleate homogeneously. In most glass ceramic systems, nucleation is heterogeneous, i.e., it is a consequence of epitaxial growth of the dominant crystalline phase upon nuclei of a foreign or earlier formed phase. The latter are usually crystallites of relatively refractory and/or insoluble phases (e.g., rutile, anatase, ZrO_2, $ZrTiO_4$, Cr_2O_3, and many of the noble metals), but can also be globules of an immiscible glass. As a result, "nucleating agents" or "nucleants" such as TiO_2, ZrO_2, P_2O_5, Pt, Pd, and Au are essential components in the bulk composition of most commercial glass ceramics. The nucleation and growth of the nucleating agent itself can be considered to be homogeneous.

The equation describing the steady-state heterogeneous nucleation process is analogous to that describing homogeneous nucleation:

$$I^* = n_s\,(kT/\eta)\exp\left[-\left(W^* + \Delta G_D\right)/kT\right] \qquad (8)$$

In this equation $W*$ is a function of the contact angle between the nucleating agent (substrate) and the epitaxial phase, and n_s is the number of formula units per unit volume (i.e., number density) of the latter in contact with the substrate. The contact angle is assumed to be a function of the degree of lattice mismatch between substrate and epitaxial phase. The effectiveness of heterogeneous nucleation therefore depends on the structural relationship between the substrate and the epitaxial phase and on the number density of heterogeneous nuclei in the supercooled liquid. The lattice constants of the nucleating agent and the desired crystalline phase should agree to within ca. 15 %. Examples are given in Table 1.

The number density of heterogeneous nuclei depends strongly on solubility; effective nucleating agents have low solubilities at the nucleation temperature. For any crystalline phase, solubility is controlled by the interactions among the melt components. Using TiO_2 and ZrO_2 as examples, their relative efficacy as nucleating agents is partly governed by the relative stability of titanate versus zirconate melt complexes. In alkali silicate glasses TiO_2 is not an effective nucleating agent because it forms stable melt complexes with alkali-metal ions; ZrO_2, on the other hand, is more effective because complex formation is not favored to the same degree [11].

Table 1. Lattice constants of crystalline phases used in glass ceramics and their mismatch relative to rutile (TiO_2 nucleant)

Crystalline phase	Lattice constant, nm			Deviations from TiO_2*	
	a	*b*	*c*		
Rutile (TiO_2, nucleant)	0.46		0.30		
β-Spodumene ($LiAlSi_2O_6$)	0.95	0.83	0.52	*a*:	3.7 % ($2a_r$)
				b:	−9.4 % ($2a_r$)
				c:	−11.2 % ($2c_r$)
β-Eucryptite ($LiAlSiO_4$)	0.52		1.12	*a*:	12.6 % ($2a_r$)
				c:	−5.3 % ($4c_r$)
Cordierite [$Mg_2Al_3(AlSi_5O_{18}$]	1.70	0.98	0.93	*a*:	−6.7 % ($4a_r$)
				b:	6.8 % ($2a_r$)
				c:	1.9 % ($2a_r$)

* Deviations are expressed as percentages relative to the quantity defined in parentheses. The subscript r denotes rutile.

The effect of the number density of heterogeneous nuclei on the grain size and, hence, the texture of a glass ceramic is illustrated in Figure 1 with the aid of three-dimensional Voronoi cells. Voronoi cells are geometric constructions in which polyhedral elements are calculated between a set of points such that the distance to each point is halved. In Figure 1, two cross sections of three-dimensional Voronoi cells have been calculated, the one with small crystallites represented by 400 points, the other with large crystallites represented by 40 points. As illustrated in Figure 2, a small number of nucleation sites results in large crystallites (A), a large number in small crystallites (B).

2.3. Crystal Growth

Once the critical radius of a crystallite is exceeded, the free energy of the system is lowered by further crystal growth rather than by crystal dissolution (Fig. 3). In both homogeneously and heterogeneously nucleating systems, the critical radius can be defined as

$$r_c = -2\sigma V/\Delta G \tag{9}$$

where V is the volume of the critical nucleus and σ is the surface free energy. The effect of the nucleating agent on the critical radius in a heterogeneously nucleating system is expressed through the surface energy term; this will be smaller the greater the structural congruence between the nucleating agent and the crystalline phase. The rate of crystal growth ζ is governed by parameters similar to those involved in nucleation. An equation has been derived by Jackson [13] to describe continuous growth by random attachment of atoms to surface sites:

$$\zeta = \lambda\nu\,(H\Delta T/kT_m T) \cdot \exp(-Q_f/kT)\exp(-H/kT_m) \tag{10}$$

The rate of crystal growth depends on the thermodynamic driving force H (the enthalpy of fusion), the jump distance λ, the vibrational frequency ν, the atomic mobility as expressed by an activation energy for diffusion Q_f, and ΔT, the temperature difference between the melting temperature T_m and the growth temperature T. As with nucleation, the mobility term can be expressed in terms of diffusivity and viscosity by the Stokes – Einstein relation. Equation (10) then becomes

$$\zeta = h\nu/6\pi\lambda^2\eta\,(H\Delta T/kT_m T)\exp(-H/kT_m) \tag{11}$$

2.4. Heat Treatment

Although the parameters governing Equations (8) and (11) are similar, the functional dependence of the nucleation rate and the growth rate on these parameters is not. Consequently, nucleation and growth rates generally attain maxima in separate temperature regimes (Fig. 4). As a result, heat treatment (ceramming) of a glass article to form a glass ceramic typically includes at least two steps. The first of these is the *nucleation hold:* a soak or gradual heating at relatively low

temperature, usually about 50 – 100 °C above the annealing point of the glass. Following the nucleation hold, the temperature is raised rapidly to that of the crystallization step. The article is then heated at constant temperature until crystal growth is complete and the desired microstructure attained. This is termed the *crystallization hold.*, Chap. 3.

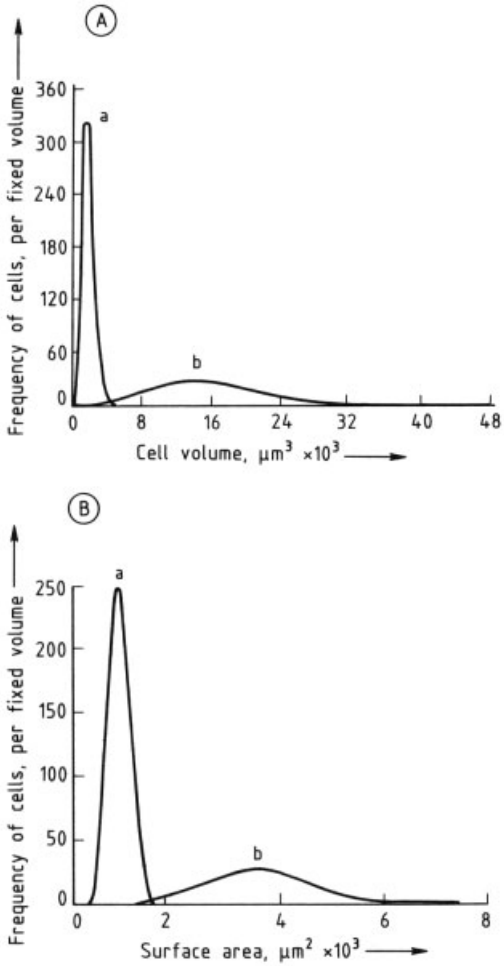

Figure 2. Calculated distributions for the volume (A) and surface area (B) of the Voronoi cells illustrated in Figure 1 Each figure shows a narrow distribution (a) with sites separated by an average distance of 16 μm and a broad distribution (b) with sites separated by an average distance of 32 μm. The areas under the two volume distributions are normalized with respect to one another. The distributions of the surface areas calculated from these volume distributions are therefore not normalized.Courtesy of Leslie Button (Corning Incorporated).

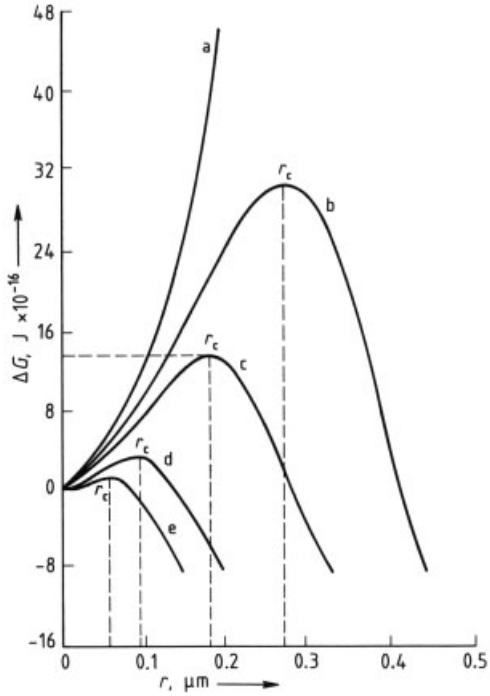

Figure 3. Calculated free energy of formation of a spherical crystallite of Mg_2SiO_4 in molten Mg_2SiO_4 as a function of crystallite radius, r, for different degrees of undercooling, ΔT, below the melting temperature, T_M [12]
ΔT, °C: a) 0; b) 10; c) 15; d) 30; e) 50 Crystallites smaller than the critical radius, r_c, redissolve in the melt.

During the crystallization hold, distortion can arise from gravitational sag of the glass or from thermal gradients in the furnace. The latter can cause one portion of the material to crystallize before the rest. Since most crystalline phases are denser than the precursor glasses, volume shrinkage of several percent is typical and must be accomplished uniformly throughout the article. Thermal gradients can arise from improper furnace design, shadow effects of refractory kiln furniture, or improper placement of glass articles. Also, the central portions of massive glass bodies are slow to heat because of their low thermal conductivity. Moreover with large articles, heat of crystallization effects can be severe: thermal waves due to exothermic crystallization of massive articles may cause the central portion to become up to 50 °C hotter than the other areas of the ceramming furnace.

To release stresses caused by differential crystallization, the apparent viscosity of the glass ceramic should be kept below 10^{12} P dur-

ing crystallization to allow viscous flow and prevent fracturing. To avoid gravitational sag, the viscosity should have a lower limit which depends on the shape and the size of the article, as well as on the support in the kiln, but is commonly around 10^{10} P.

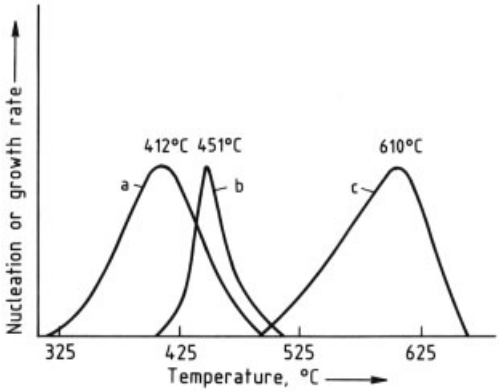

Figure 4. Temperature regime for nucleation (a, b) and growth (c) in the $Li_2O \cdot 2\ SiO_2$ system, showing the non-coinciding temperature domains in which nucleation (low temperature) and growth (high temperature) tend to occur
a) Theoretical nucleation; b) Experimental nucleation; c) GrowthThe observed nucleation domain is characteristically narrower than the growth domain. This figure also indicates that theoretical calculations do not support the observed nucleation behavior in this system (→Glass).

3. Nature of Crystalline Phases

3.1. Metastability

The ceramming (i.e., heating) of glass to produce glass ceramics can be described as subsolidus crystallization (i.e., crystallization at a temperature below which melting occurs). Accordingly, the expected crystalline phases in glass ceramics are those of an equilibrium subsolidus assemblage appropriate for the bulk composition of the precursor glass and the temperature of the crystal growth step. However, in many glass ceramic systems, the observed crystalline phases are metastable. This is not unexpected, because glass itself is in a metastable state. Thus, with glass rather than with a melt as the initial state, crystalline phases can be formed which do not occur under conditions of "equilibrium crystallization".

Perhaps the best example of this phenomenon is the precipitation of β-quartz solid solutions from lithium aluminosilicate glasses that are cerammed for a short time at about 900 °C. Not only is the β-quartz metastable at this temperature, it also appears to be thermodynamically unstable, i.e., this phase is not a member of any equilibrium assemblage at any temperature for these bulk compositions. Prolonged heating at 900 °C, or raising the ceram temperature, causes the β-quartz glass ceramic to invert to a stable assemblage containing β-spodumene solid solution (Fig. 5). However, the early formed β-quartz glass ceramics persist indefinitely when cooled to room temperature. Moreover, such glass ceramics can withstand indefinite thermal cycling between room temperature and 900 °C without undergoing inversion to a β-spodumene glass ceramic.

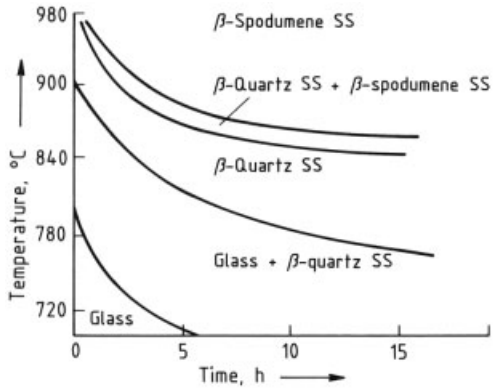

Figure 5. Phase transitions in a lithium aluminosilicate glass as a function of time and temperature
All phases in this system are solid solutions (SS) with varying silica content.

3.2. Physical Properties

The physical properties rather than the morphological properties of the crystalline phases tend to dominate the behavior of a glass-ceramic article.

3.2.1. Thermal Properties

Crystalline phases encompassing a wide range of *thermal expansion coefficients* have been precipitated from glass (Table 2). Glass ceramics containing high-expansion phases such as nepheline, $NaAlSiO_4$ (expansion coefficient

50×10^{-6} K^{-1}), are commercially important because they can be strengthened by the application of low expansion glazes. Phases with coefficients of thermal expansion as high as 12.5×10^{-6} K^{-1} (e.g., fluorcanasite) can easily be precipitated from glass. Crystalline phases with ultralow coefficients of thermal expansion are even more important, however, and are widely used in glass ceramic articles that are subjected to extensive thermal cycling or articles in which even minor distortion due to thermal fluctuation is undesirable.

The most important of the low-expansion phases are the lithium aluminosilicates, β-spodumene ($LiAlSi_2O_6$) and β-eucryptite ($LiAlSiO_4$), which can be regarded as "stuffed" derivatives of the two polymorphs of crystalline silica, keatite and β-quartz, respectively. Keatite consists of a three-dimensional network of five-membered tetrahedral rings, whereas β-quartz consists of a three-dimensional network of six-membered tetrahedral rings. The helical arrangement of silica tetrahedra in both phases results in unique behavior with increasing temperature; expansion along one crystal axis is accompanied by contraction along the perpendicular axes. As a result, the net thermal expansion of both crystalline silicas is exceptionally low. To capitalize on this property in a glass ceramic body, individual crystals must be small and unoriented. Solid solutions of β-spodumene (stuffed keatite, $Li_2O-Al_2O_3-nSiO_2$) have been crystallized from a variety of lithium aluminosilicate glasses. Important in most applications of β- spodumene glass ceramics is the fact that thermal expansion can be tailored to suit a wide range of uses. Volume thermal expansions and thermal expansion for the tetragonal axes are illustrated in Figure 6 A and 6 B, respectively. Whereas bulk expansion decreases with increasing silica content, the individual axes show a very large anisotropy in expansion behavior (Fig. 6 B).

Stuffed derivatives of β-quartz show a wide range of solid solutions between SiO_2 and $LiAlO_2$, $Mg(AlO_2)_2$, $Zn(AlO_2)_2$, Li_2BeO_2, $Al(AlO_2)_3$, Li_2ZnO_2, BPO_4, or $AlPO_4$. This allows crystallization of ultralow expansion phases from a wide variety of glass compositions. The thermal expansion coefficient of β-eucryptite (lithium β-quartz solid solution) is of the order of -2×10^{-6} K^{-1}. The thermal expansion coefficient can be increased gradually by partially replacing lithium with magnesium or zinc. By taking advantage of this effect, stuffed β-quartz glass ceramics with a wide range of thermal expansion coefficients, $(-2$ to $+5)\times10^{-6}$ K^{-1} can be manufactured.

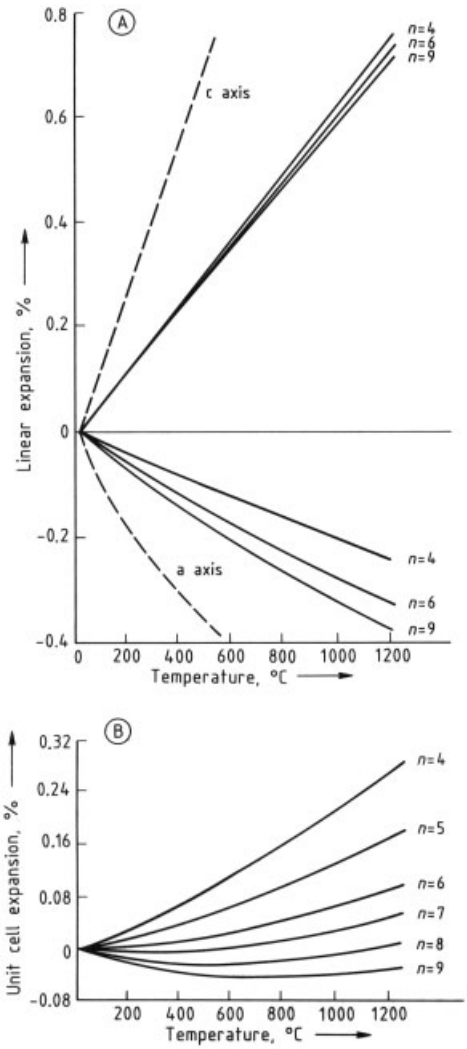

Figure 6. Thermal expansion in the lithium aluminosilicate system
A) Linear expansion for keatite (dashed lines, five-membered rings of silica tetrahedra) and synthetic β-spodumene (solid lines, 6-membered rings of silica tetrahedra;
B) Volume expansion for β-spodumene solid solutions. The n value represents the silica concentration in the $Li_2O-Al_2O_3-nSiO_2$ system.

Table 2. Bulk linear coefficients of thermal expansion of crystalline phases used in glass ceramics [14], [15]

Compound	Formula	Expansion coefficient $10^{-6}\,K^{-1}$	Temperature, °C **
Nepheline	$NaAlSiO_4$	50	RT – 500
Anorthite	$CaAl_2Si_2O_8$	19	RT – 850
Celsian	$BaAl_2Si_2O_8$	3	RT – 600
Mg cordierite	$Mg_2Al_3AlSi_5O_{18}$	7	RT – 800
Commercial cordierite		3	RT – 900
Fluorocanasite	$Ca_5Na_4K_2Si_{12}O_{30}(OH,F)_4$	12.5	RT – 800
β-Spodumene	$LiAlSi_2O_6$	2	RT – 1200
Keatite	SiO_2	−1	RT – 500
β-Eucryptite	$LiAlSiO_4$	1	RT – 450
α-Quartz	SiO_2	40	RT – 450
SiO_2 glass	SiO_2	50	−23 – 1000
α-GeO_2	GeO_2	30	RT – 850
β-Quartz	SiO_2	−6	600 – 900
Pyroxenes	$(Mg,Fe)_2Si_2O_6$	32 – 38	RT – 800
Andalusite	$Al_2Si_2O_5$	43	RT – 800
Sillimanite	$Al_2Si_2O_5$	26	RT – 800
Topaz	$Al_2SiO_4(F,OH)_2$	25	RT – 800
Zircon	$ZrSiO_4$	18	RT – 800
MgO		15.5	−150 – 1100
ZrO_2		10 – 11	RT – 1000
Al_2O_3*		9.4	−150 – 1100
TiO_2*		8.8	−150 – 1100
SiC		5.4	−150 – 1100
Diamond		4.1	−150 – 1100

* Polycrystalline samples.
** RT = room temperature.

The *specific heat* of glass ceramics varies only slightly with composition. A characteristic value in the temperature range between 20 and 400 °C is 1000 J $kg^{-1}\,K^{-1}$, and is similar to values for glass and conventional ceramics.

The *thermal conductivity* of glass ceramics is an important property in such applications as range tops and cookware. Glass ceramic materials tend to have higher thermal conductivity than glass, but lower than ceramics made from pure oxides. A characteristic value for lithium aluminosilicate glass ceramics is about 5.5 W $m^{-1}\,K^{-1}$.

3.2.2. Transparency

Most glass ceramic bodies are opaque. In order for a glass-ceramic body to be transparent either: (1) the dimension of the individual crystals should be about one third of the wavelength of visible light, or (2) the refractive index of each crystal should match the refractive index of the residual glass and the birefringence of each crystal should be negligible, (i.e., optical isotropy).

The first condition is exemplified by glass ceramics containing mullite (aluminum silicate). A fine dispersion of mullite can be formed from glasses in the Al_2O_3–SiO_2 system. The resulting glass ceramic is clear, even though mullite is birefringent and has a higher refractive index than the residual glass. Upon prolonged heat treatment, crystal size increases and is accompanied by loss of transparency.

The second condition has been achieved in a number of glass ceramics where the principal crystalline phase is the stuffed β-quartz solid solution. Pure β-quartz is an optically uniaxial positive mineral, but in solid solution its birefrigence is negligible at approximately 70 wt % SiO_2. Crystals of β-quartz containing less than 70 wt % SiO_2 are optically negative. The large number of components in the β-quartz solid solution also allow the composition of the residual glass to be tailored such that the refractive index of the crystals matches that of the residual glass, resulting in transparent glass ceramics. Compositions of commercial transparent glass ceramics are shown in Table 6.

3.2.3. Electrical Properties

Resistivity. In commercial glass ceramics, the dominant crystalline phases are silicates,

aluminosilicates, and phosphates; they can be broadly classified as electrical insulators. Electrical conduction occurs by diffusion of weakly bonded ions, primarily alkali-metal ions. At low temperatures, ionic mobility is negligible and, consequently, most of these phases are characterized by low electrical conductivity. This, coupled with the fact the crystalline phases in glass ceramics are commonly surrounded by a siliceous residual glass, is responsible for the high electrical resistivity of most glass ceramics.

Dielectric Constant. The dielectric constant of most glass ceramics is low, characteristic values being around 6 at 1 MHz and 20 °C. The structure of the crystalline phases in these materials is dominated by the rigid framework of network-forming components, such as silicate tetrahedra, which are only slightly polarized by an applied electric field. Moreover, glass ceramics in which the crystalline phases as well as the residual glass contain only network-forming components can have exceptionally low dielectric constants (as well as very high resistivity). Good examples of such materials are glass ceramics consisting of crystallites of cubic BPO_4 within a highly siliceous residual glass; they can have dielectric constants as low as 4.

Glass ceramics with relatively high electrical conductivity and/or dielectric constants can be formed from relatively depolymerized glasses, i.e., glasses with a high network modifier content. Crystallization of such glasses can result in the precipitation of oxide phases, such as titanates and niobates. For example, glass ceramics containing an interconnected network of crystallites of titanium suboxides can be formed from reduced titanium aluminosilicate glasses. These suboxides are themselves good electrical conductors and, as they are in grain-to-grain contact, make the glass ceramic itself conductive. Heat treatment of certain alkaline earth – lead aluminosilicate glasses can result in the precipitation of ferroelectric oxides with a perovskite structure. In this way, glass ceramics with dielectric constants of the order of 10^3 can be manufactured and utilized in the fabrication of capacitors.

Dielectric losses include losses through conductance, ion relaxation, deformation, and vibration. Conductance and relaxation losses are primarily determined by the mobility and concentration of the alkali-metal ions present and predominate at frequencies below 1 MHz. Deformation and vibration losses are important at frequencies of about 10 MHz. Characteristic values for glass ceramic loss factors are around 0.02 at 1 MHz and 20 °C.

Dielectric Strength. The dielectric strength is important in applications involving the use of glass ceramics as high-voltage insulators or condensers. The very fine-grained, nonporous microstructure of glass ceramics ensures that they have a substantially higher dielectric strength than conventional ceramics with characteristic values around 47 kV/mm.

3.2.4. Mechanical Properties

Glass ceramics can be considered as brittle glass – crystal composites. Their mechanical properties are strongly affected by the particle size and volume fraction of the crystalline phase, the interfacial bond strength, and differences in elastic modulus and thermal expansion between glass and crystallites.

Strength and Elasticity. Glass ceramics are stronger than the parent glasses. In glasses, theoretical strengths cannot be achieved due to the presence of intrinsic and/or extrinsic flaws. In atomic terms, flaws may be regarded as "sharp cracks" or radii that act as stress concentrators. The strength of the material is inversely proportional to the critical flaw dimension. A commonly used model developed by GRIFFITH states that

$$\sigma = (2E\gamma/\pi c)^{1/2} \tag{12}$$

where σ is the strength, E is the modulus of elasticity, γ is the fracture surface energy, and c is the length of the crack (flaw). Stress is proportional to the square root of the flaw size and cracks grow rapidly when the stress (due to mechanical or thermal loading) overcomes the atomic bonding forces that hold the material together.

For most brittle oxide solids, three types of flaw can be identified: surface cracks, internal

processing defects, (e.g., porosity), and environmentally induced damage. Strength and toughness (resistance to crack propagation) of crystalline brittle materials depend on grain boundary residual stress, the ratio of flaw dimensions to grain size, crystallite cleavage resistance, percent crystallinity, crystallite size, and relative elastic properties. Many of these issues are overcome in glass ceramics through the creation of highly crystalline microstructures which are specifically designed to resist crack propagation by increasing material toughness.

One method used to increase the toughness of glass ceramics is modification of crystallite size. If the length of the microcrack c is assumed to be proportional to the crystal diameter, then from the Griffith equation the strength σ is approximately inversely proportional to the square root of the crystal size d

$$\sigma \sim \frac{1}{\sqrt{d}} \tag{13}$$

Thus, small crystallites increase the strength of the glass ceramic.

Additional toughening occurs when the glass ceramic contains a crystalline phase that is stronger than the glass matrix. This is generally the case because, at an atomic level, the crystals which precipitate from the glass are tightly packed so that the number of bonds to be broken per unit volume is large and the modulus of elasticity is high. The size of the microcracks is then limited to the region between crystals and the critical parameter of mechanical strength will be the mean free path in the residual glass L_s:

$$L_s = d\,(1 - V_f)\,/V_f \tag{14}$$

where V_f is the volume fraction of the crystalline phase. A small mean free path indicates a major strength improvement due to deflection toughening of crack tips.

Hardness. The hardness of glass ceramics can vary significantly depending upon the composition of the crystalline phases. Knoop values in excess of 10 000 MPa can be obtained in glass ceramics where the dominant phase is magnetite (Fe_3O_4), spinel ($MgAl_2O_4$), or sapphirine ($Mg_2Al_4SiO_{10}$). This permits use of glass ceramics in applications where high abrasion resistance is required.

3.2.5. Durability

The chemical stability and durability of glass ceramics are affected by the composition of the crystalline phase, as well as the amount and composition of the residual glass. For example, a primary cause of poor chemical durability is the presence of alkali-metal ions in the residual glass. Many aluminosilicate glass ceramics contain phases that are extremely siliceous, such as β-spodumene and mullite; this results in exceptional resistance to chemical etching and staining.

3.3. Microstructure

The texture of the crystalline assemblage is one of the principal factors determining the physical properties of glass ceramics, ranking in importance only second to the physical properties of the dominant precipitating phase. Changes in the mechanism and rate of nucleation and growth affect crystal morphology, permitting the development of myriad textures, which are often distinct from normal textures characteristic of sintered powders or fusion-cast materials. Key microstructures of commercial glass ceramics are described as:

1) ultrafine-grained
2) relict
3) “coast and island”
4) “house of cards”
5) acicular interlocking
6) dendritic

Ultrafine-Grained Textures. The Voronoi patterns (Fig. 1) have shown how ultrafine-grained textures can be achieved: the greater the number of nucleation sites, the finer the precipitate and, as shown in Section 3.2, the better the mechanical strength. In the lithium aluminosilicate system, these textures can be obtained by adding roughly equivalent amounts (on a molar basis) of TiO_2 and ZrO_2 to the glass. Small ($<$10 nm) zirconium titanate crystallites precipitate when such glasses are heated to about 50 °C above the glass transition temperature. These crystallites form a uniform dispersion throughout the residual glass, providing a multitude of

nucleation sites for the epitaxial growth of the stuffed β-quartz solid solution. The efficient nucleation, coupled with sluggish grain growth, allows the crystals to grow only to about 50 nm in diameter before impinging upon one another.

Relict Textures. Some glass-ceramic microstructures mimic the texture of the precursor glass. This is, for instance, the case in mullite glass ceramics where the original glass separates into aluminous droplets within a siliceous matrix. When heated above the annealing point of the phase-separated glass, the alumina-rich droplets become more fluid than the matrix and crystallize rapidly to mullite. The mullite crystals are restricted by the shape of the original droplet and reflect its morphology. Due to the small size of the mullite "droplets" ($<0.1\,\mu$m), these glass ceramics are transparent.

Relict microstructures can also develop when a crystal phase retains the morphology of a metastable, crystalline precursor. For example, a relict dendritic texture (see below) is observed in photoetched lithium disilicate glass ceramics. The crystals reflect the dendritic texture inherited from metastable lithium metasilicate, the first crystalline phase formed upon ceramming.

"Coast and Island" Texture. A coast and island texture usually develops when an equilibrium crystal phase forms at the expense of a metastable assemblage of phases. An example is the growth of hexagonal cordierite in magnesium aluminosilicate glass ceramics at the expense of α-quartz, sapphirine, and spinel. Growth of hexagonal cordierite is initiated at boundaries between grains of the latter phases, and cordierite gradually becomes the dominant phase. Any excess silica crystallizes as cristobalite laths, which along with early-formed rutile, become partially or wholly enveloped by the cordierite. This results in an interlocking structure with good mechanical strength and toughness.

"House of Cards" Texture. The precision machinability of fluoromica glass ceramics such as Macor, with conventional metal-working tools is due to the "house of cards" texture, which is characterized by randomly oriented, flexible flakes. In the case of fluoromica glass ceramics, the plate-like crystallites have low shear strength due to their perfect basal cleavage and thus tend to either arrest fractures or cause deflection or branching of cracks. Another important property dependent on this microstructure is high dielectric strength. Such insulating qualities are possible because of the continuous interlocking of the mica sheets. Very low helium permeation rates, important in high vacuum applications, are also related to this microstructure.

Acicular Interlocking Texture. Interlocking rod- or blade-like crystals in glass ceramics can serve as important strengthening and toughening agents. Most chain silicate glass ceramics contain crystallites with this morphology. The high toughness is attributed to rod reinforcement by elongated amphibole crystals such as potassium fluororichterite.

An even tougher microstructure is developed in the highly crystalline monophase glass ceramics which consist of interpenetrating blades of fluorcanasite. Cleavage splintering can be observed, causing energy absorption through crack branching and deflection.

Dendritic Textures. The dendritic microstructure is characterized by the presence of crystallites with skeletal morphology, which is associated with very high surface energies. Snow crystals are familiar examples of dendritic crystals. Glass ceramics with dendritic textures typically contain a relatively high percentage of residual glass, due to inefficient crystal packing. The texture assumed by lithium metasilicate crystallites in the chemically etchable glass ceramic Fotoform is an example of the dendritic microstructure. The metasilicate dendrites form a continuous three-dimensional lacework. The glass ceramic can be etched with hydrofluoric acid, which attacks the low silica crystallites at a far higher rate than the durable aluminosilicate residual glass.

4. Types of Glass Ceramics

Useful glass ceramics are generally silica-containing glasses and can be classified in three groups: silicates, aluminosilicates, and fluorosilicates. Some of the more important systems currently used in the manufacture of commercial glass ceramics are shown in Table 3.

Table 3. Silicate glass ceramic systems

System	Advantages
Silicate	
Li_2O-SiO_2	photoetching capability
Aluminosilicate	
$Li_2O-Al_2O_3-SiO_2$	very low thermal expansion
$MgO-Al_2O_3-SiO_2$	good mechanical, thermal, and dielectric properties
$Na_2O-Al_2O_3-SiO_2$	high mechanical strength from compression glazing
$CaO-MgO-Al_2O_3-SiO_2$	low cost
Fluorosilicate	
$K_2O-MgO-Al_2O_3-SiO_2-F$	machinability

4.1. Silicate-Based Glass Ceramics

The compositions of some silicate-based glass ceramics are listed in Table 4.

Table 4. Composition of commercial silicate glass ceramics (mol %)

Component	Fotoform/ Fotoceram Corning 8603	"Slag-sital" white, Soviet Union	"Minelbite" gray, Hungary
SiO_2	75.8	56.3	63.8
Al_2O_3	2.2	5.0	8.8
Li_2O	17.8		
Na_2O	1.5	5.3	3.3
K_2O	2.5	0.4	1.3
MgO		3.3	8.9
CaO		27.0	10.1
ZnO [a]		1.0	
MnO [a]		0.8	1.8
Fe_2O_3 [a]		0.1	1.0
S [a]		0.8	1.2
CeO_2 [b]	0.0047		
SnO_2 [b]	0.0011		
Sb_2O_3 [c]	0.079		
Ag [a]	0.06		
Au [a]	0.00029		

[a] Nucleant.
[b] Sensitizing agent.
[c] Fining agent.

Lithium Silicates. Photosensitive lithium silicate glass ceramics (e.g., Fotoform/ Fotoceram, Corning) include a noble metal such as gold or, more commonly silver, and trivalent cerium ions (Table 4). Upon exposure to UV light, Ce^{3+} is thought to act as an electron donor, ultimately leading to the precipitation of metallic silver:

$$Ce^{3+} + h\nu \longrightarrow Ce^{4+} + e^-$$
$$e^- + T \longrightarrow T_e \quad (T = \text{trapping site})$$
$$T_e + Ag^+ \xrightarrow{\text{Heat}} Ag^0$$
$$Ag^0 + e^- + Ag^+ = (Ag^0)$$

In lithium silicate glasses, the photoinduced, thermally developed silver speck acts as a nucleation site for the epitaxial growth of dendritic metasilicate (Li_2SiO_3), which is more readily etched by hydrofluoric acid than the high-silica parent glass. Growth of lithium metasilicate occurs selectively only in those areas which have been exposed to UV radiation. Using masking techniques, complex patterns of lithium metasilicate can be formed in these glasses. The irradiation pattern can then be selectively etched out of the glass with hydrofluoric acid, forming the basis for Fotoform articles. (See also → Imaging Technology). The photoetched glass is flood-exposed to UV radiation and recerammed at higher temperature, resulting in the inversion of metastable lithium metasilicate to the stable disilicate, $Li_2Si_2O_5$. The resulting glass ceramic is strong, with an abraded modulus of rupture around 150 MPa, and negatively replicates the original irradiation pattern to a resolution of less than 0.1 mm.

Applications of such photosensitive glass ceramics include recording head pads, fluidic devices (e.g., hydraulic networks), cellular face plates for gas discharge displays, and charged plates for ink-jet printing.

Calcium and Calcium Magnesium Silicates. Blast furnace slags have been used since the 1970s in eastern Europe to manufacture inexpensive glass ceramics. Compositions of two glass ceramics of this type are given in Table 4. The furnace slags are rolled as sheets or cast as tiles and used for both interior and exterior use. The slag is remelted at around 1450 °C, formed into a glass and subsequently heat treated at about 1000 °C. The crystalline phases diopside [*14483-19-3*] ($CaMgSi_2O_6$) or wollastonite [*13983-17-0*] ($CaSiO_3$) nucleate heterogeneously on zinc, iron, or manganese sulfide particles. These alkaline-earth silicates form fine, 1 – 5 µm, equiaxial grains in a matrix of residual aluminosilicate glass, and confer good abrasion resistance and chemical durability on the glass ceramic. The glass ceramics have a va-

riety of applications depending on their composition. Their main uses are in the construction and chemical or petrochemical industries as abrasion-resistant or chemically resistant floor and wall coverings, chimneys, parts of chemical pumps, and funnels.

Magnesium Silicates. Glass ceramics based on enstatite [*14681-78-8*] ($MgSiO_3$) are of scientific interest because of the martensitic (i.e., dilatational) transformation from protoenstatite to clinoenstatite upon cooling. This transformation results in fine lamellar (i.e., parallel) twinning which produces a superior crack-tip blunting texture in the body. Dilution of enstatite glass with other components is required because pure magnesium metasilicate does not form a stable glass. Two representative compositions from the SiO_2–MgO–Al_2O_3–Li_2O–ZrO_2 and SiO_2–MgO–ZrO_2 systems are given in Table 5. These refractory, tough, fine-grained glass ceramics consist of 50–85 wt % enstatite with zircon ($ZrSiO_4$) and β-spodumene, ($LiAlSi_2O_6$), or tetragonal zirconia (ZrO_2) as additional phases.

Table 5. Compositions and properties of enstatite ($MgSiO_3$) glass ceramic

Parameter	E-1	E-2
Composition, mol %		
SiO_2	55.0	49.3
MgO	35.3	44.9
Al_2O_3	3.0	
Li_2O	1.7	
ZrO_2	4.9	5.8
Ceramming schedule	800 °C, 2 h 1200 °C, 4 h	800 °C, 2 h 1400 °C, 4 h
Abraded (modulus of rupture, MPa)	193 ± 15	200 ± 15
Fracture toughness, MPa $m^{1/2}$	3.5 ± 0.4	4.6 ± 0.6
Refractory temperature, °C	1250	1500
Coefficient of thermal expansion (0–1000 °C), K^{-1}	68×10^{-7}	80×10^{-7}

During the ceramming sequence, phase separation occurs between 800 and 900 °C and is followed by rapid crystallization of tetragonal zirconia. These zirconia crystallites form the nucleation centers on which enstatite precipitates. At temperatures above 1200 °C, zircon is formed by reaction of zirconia with unpolymerized silica present in the residual glass, thereby further stabilizing the glass ceramic.

The toughest (about 4 MPa $m^{1/2}$) and most refractory compositions are basically two-phase enstatite – zircon glass ceramics (type E-2). These have use temperatures approaching 1525 °C, the eutectic temperature in the SiO_2–MgO–ZrO_2 system. They are well-suited for high-temperature applications.

4.2. Aluminosilicate Glass Ceramics

Lithium and magnesium aluminosilicate glass ceramics are currently the principal commercial glass ceramic systems. Their utility lies in their resistance to thermal shock combined with excellent chemical durability.

The Li_2O–Al_2O_3–SiO_2 system has produced glass ceramics with exceptionally low thermal expansion coefficients based on either β-quartz or β-spodumene solid solution. These phases can be melted and quenched to a stable glass permitting the manufacture of virtually monophase β-quartz or β-spodumene glass ceramics which contain only minor amounts of nucleant phases and residual glass. Glass ceramics containing β-quartz or β-spodumene can be made from glasses of the same composition by modifying the heat treatment: β-quartz is formed by ceramming at or below 900 °C, and β-spodumene by ceramming above 1000 °C.

β-Quartz Solid Solution. The composition of the β-quartz solid solution is $(Li_2O, RO)\cdot Al_2O_3\cdot nSiO_2$, in which n varies between 2 and 10 and R is a divalent cation, commonly Mg^{2+} or Zn^{2+}; the stoichiometric end member, $LiAlSiO_4$, is called β- eucryptite. For commercial glass ceramics, n ranges between 6 and 8 because the high silica content of these compositions furnishes a melt that is sufficiently viscous to be rolled, pressed, blown, and vacuum formed. Representative compositions are listed in Table 6.

A phase with the fluorite structure, $ZrTiO_4$, has proven to be the most effective nucleating agent in precipitating β-quartz from lithium aluminosilicate glasses. Very fine crystallites with grain sizes of the order of 60 nm can be produced if the nucleant concentration is ca. 4 wt %.

The combination of transparency, low thermal expansion behavior, optical polishability,

and strength greater than glass has generated applications such as see-through cookware, telescope mirror blanks, woodstove windows, IR-transmitting range tops, optically stable platforms, and ring-laser gyros.

Table 6. Composition (mol %) of transparent glass ceramics based on β-quartz solid solution

Component	Visions (Corning)	Zerodur (Schott)	Narumi (Nippon Electric)
SiO_2	74.2	63.8	70.6
Al_2O_3	12.2	17.1	14.4
Li_2O	5.8	8.6	9.2
MgO	2.9	1.7	0.8
ZnO	0.8	1.2	
P_2O_5		3.8	0.6
F			0.3
Na_2O	0.2	0.6	0.6
K_2O	0.1		0.2
BaO	0.3		
TiO_2	2.2	2.0	1.6
ZrO_2	0.9	1.1	1.2
As_2O_3	0.3	0.2	0.4
Fe_2O_3	0.04	0.01	0.01

β-Spodumene (Keatite) Solid Solution. Two representative compositions of β-spodumene glass ceramics are given in Table 7. Comparison with the Vision composition in Table 6 indicates the similarity between β-spodumene and β-quartz glass ceramic materials.

Table 7. Composition and properties of glass ceramics based on keatite (β-spodumene) solid solution

Parameter	Corning Ware	Cercor Corning
Composition, mol %		
SiO_2	73.6	75.9
Al_2O_3	11.0	13.6
Li_2O	5.9	10.5
MgO	4.1	
ZnO	0.8	
Na_2O	0.4	
K_2O	0.1	
TiO_2	3.7	
ZrO_2	0.1	
Fe_2O_3	0.1	
As_2O_3	0.2	
Coefficient of thermal expansion, K^{-1}	1.2×10^{-6} (0–500 °C)	0.5×10^{-6} (0–1000 °C)
Abraded flexural strength, MPa	100	

Corningware is made from a multicomponent glass containing TiO_2 as nucleating agent. The ceram schedule attains a maximum near 1125 °C, about 100 °C below the liquidus temperature. At least 93 vol % of the material is crystalline, with β-spodumene being the dominant phase. Spinel, rutile, and residual glass are secondary phases. The glass ceramic is produced as a honeycomb product for a turbine engine heat exchanger. It is made from a powdered glass frit, which as slurry impregnates paper wound on a wheel as alternately corrugated sheets. After firing, a porous ceramic regenerator wheel is produced which allows energy to be transferred from hot exhaust gases to the cold intake air in the turbine engine.

Cordierite-Based Glass Ceramics. The hexagonal (high temperature) polymorph of cordierite [*1302-88-1*] ($Mg_2Al_4Si_5O_{18}$) is the principal crystalline phase in magnesium aluminosilicate glass ceramics. These materials are strong, have excellent dielectric properties, and good thermal stability and shock resistance. The thermal expansion of cordierite is low and anisotropic, the c axis expanding, and the a and b axes contracting, resulting in virtually zero volume expansion.

In these materials, hexagonal cordierite crystallizes together with cristobalite, rutile, and magnesium dititanate; a small amount of residual glass is isolated at the grain boundaries. This glass ceramic is used for missile radomes. A representative glass ceramic composition, together with selected physical properties, is shown in Table 8.

Table 8. Composition and properties of commercial cordierite glass ceramic (Corning 9606)

Parameter	Value	Phases
Composition, mol %		
SiO_2	58.1	
Al_2O_3	12.1	
MgO	22.6	cordierite
CaO	0.1	cristobalite
TiO_2	6.9	rutile
As_2O_3	0.1	Mg-dititanate
Fe_2O_3	0.1	
Coefficient of thermal expansion, K^{-1}	4.5×10^{-6}	
Fracture toughness, MPa $m^{1/2}$	2.2	
Thermal conductivity, $J\,s^{-1}\,cm^{-1}\,K^{-1}$	0.37	
Knoop hardness, MPa	7000	
Dielectric constant (8.6 GHz)	5.5	
Loss tangent	0.0003	

Sodium Aluminosilicates. Fine-grained glass ceramics, containing nepheline [*12251-27-3*] ($NaAlSiO_4$) can be formed from glasses in the $Na_2O-Al_2O_3-SiO_2-TiO_2$ system. Nepheline is a stuffed derivative of tridymite, a high-temperature polymorph of silica. Titanium dioxide acts as an internal nucleating agent and opacifier, and crystallizes as anatase in the cerammed material. These glass ceramics have thermal expansion coefficients around 5×10^{-6} K^{-1}, near that of nepheline, reflecting the expansion behavior of tridymite. A typical composition of this glass ceramic is given in Table 9.

Table 9. Nepheline glass ceramic tableware (Pyroceram 9609)

Oxide	Content, mol %	Phases
SiO_2	53.0	
Al_2O_3	21.4	
Na_2O	16.0	nepheline ($NaAlSiO_4$)
BaO	2.6	celsian ($BaAl_2Si_2O_8$)
TiO_2	6.0	anatase (TiO_2)
As_2O_3	0.4	glass

Barium is added to promote the formation of celsian ($BaAl_2Si_2O_8$), which has a lower thermal expansion coefficient (3×10^{-6} K^{-1}) than nepheline, thereby improving thermal shock resistance. The glass ceramics are used in a wide variety of products ranging from coverings for microwave antennas, large electric insulators, and miniature vacuum tubes to heat-resistant kitchenware and stove hotplates.

Other Aluminosilicate Systems. Glass ceramic systems based on mullite [*55964-99-3*] ($3Al_2O_3\cdot 2SiO_2$) can be produced from modified binary $Al_2O_3-SiO_2$ glasses. Transparent mullite glass ceramics, doped with ions such as Cr^{3+} have unique luminescent characteristics.

Calcium aluminosilicate glass ceramics based on anorthite [*1302-54-1*] ($CaO\cdot Al_2O_3\cdot 2SiO_2$) can be formed by surface crystallization sintering of glass frit. They form ideal glass ceramic matrices for silicon carbide fiber-reinforced composites. The good thermal stability of anorthite (up to 1500 °C), combined with its close match in thermal expansion with silicon carbide makes this composition particularly attractive.

Glass ceramics belonging to the *cesium aluminosilicate system* are highly refractory. Glass ceramics made from pollucite [*1308-53-8*] ($Cs_2O\cdot Al_2O_3\cdot 4SiO_2$) and mullite crystallized at 1600 °C have a thermal stability that is 350 °C higher than that of fused silica. Due to their refractory nature, these materials must be melted at 1900 °C and formed near 1800 °C.

4.3. Fluorosilicate Glass Ceramics

Two classes of glass ceramics fall in this category: (1) fluorosilicate micas consisting of sheets of silica tetrahedra and (2) fluoroinosilicates, whose structure is based on double or quadruple chains of silica tetrahedra.

Fluormica Glass Ceramics. Fluormica sheet silicates provide machinable glass ceramics which can be turned, drilled, milled, or threaded using conventional metal-working machines.

Macor is a commercial glass ceramic based on fluorophlogopite [*12003-38-2*] ($KMg_3AlSi_3O_{10}F_2$). Due to its high fluorine content, stable glasses with fluorophlogopite composition cannot be formed. Addition of B_2O_3 and SiO_2, as shown in Table 10, results in stable glass. The Macor composition is melted to an opal glass that is subsequently crystallized by heating to 950 °C. During ceramming, metastable chondrodite, $2\,Mg_2SiO_4\cdot MgF_2$, is the first phase to precipitate from the glass, crystallizing from the magnesium-rich matrix at the surface of the aluminosilicate droplets that constitute the rest of the parent opal. Chondrodite is subsequently transformed into norbergite, $Mg_2SiO_4\cdot MgF_2$, which in turn reacts with the residual glass to form fluorophlogopite and some mullite.

The mica structure is particularly well adapted for machinability due to the rapid propagation of fracture in the plane of the silica sheets. Furthermore, the "house of cards" texture of fluorophlogopite glass ceramics (see Section 3.3) increases the tortuosity of the fracture path and, hence, fracture energy is rapidly dissipated. The machinability of Macor also benefits from the relatively coarse size of the mica flakes. Glass ceramics with small crystallites are less easy to machine than those with large crystallites; the

difference is presumably due to the additional energy required to cleave many more crystallites.

Table 10. Commercial fluormica glass-ceramic compositions (mol %)

Oxide	Macor	Dicor
SiO_2	42.2	46–57
B_2O_3	6.6	
Al_2O_3	8.8	0–1
MgO	19.3	22–23
K_2O	5.4	8–9
F	17.8	13–21
ZrO_2		0–1
CeO_2		0.02
Mica type		
Macor	$K_{1-x}Mg_3Al_{1-x}Si_{3+x}O_{10}F_2$	
Dicor	$K_{1-x}Mg_{2.5+x}/2Si_4O_{10}F_2$	
	$x < 0.2$	

Macor has been employed in a wide range of applications, including precision electrical insulators, vacuum feedthroughs, windows for microwave tube parts, sample holders for field ion microscopes, seismograph bobbins, gamma-ray telescope frames, and boundary retainers on the space shuttle. The precision machinability of Macor with conventional metal-working tools, its high dielectric strength (40 kV/nm), and very low helium permeation rates are particularly important in high-vacuum applications.

Fluorosilicate glass ceramics based on tetrasilicic micas, such as $KMg_{2.5}Si_4O_{10}F_2$, have been developed for use in dental restoration. Tetrasilicic fluormica glass ceramics, marketed under the trade name Dicor, have improved chemical durability and translucency relative to Macor. Good strength ($\sim$150 MPa) is associated with the development of anisotropic flakes at relatively high temperature ($>$1000 °C). Translucency is achieved by nearly matching the refractive indices of the fluormica and residual glass, while simultaneously maintaining a fine ($\sim$1 µm) crystal size. Cerium is added to simulate the fluorescent character of natural teeth. The glass ceramic may be cast accurately using a lost wax technique and conventional dental laboratory molds. The high strength and low thermal conductivity of these materials provide advantages over conventional metal–ceramic systems.

Chain Fluorosilicates. Natural nephrite-jade is renowned for its toughness (i.e., resistance to fracture) due to crack tip blunting by randomly oriented acicular crystallites. Two chain fluorosilicates can produce glass ceramics with similar high toughness: potassium fluorrichterite and fluorcanasite.

Potassium Fluorrichterite. The crystal structure of potassium fluorrichterite $KNaCaMg_5Si_8O_{22}F_2$, consists of double chains of silica tetrahedra. An optimized fluorrichterite glass-ceramic composition in mole percent is:

SiO_2	59.2
Al_2O_3	1.1
MgO	18.5
CaO	4.4
Na_2O	2.5
K_2O	2.7
Li_2O	1.3
BaO	0.1
P_2O_5	0.4
F	9.7

In the early stages of the ceramming of a fluorrichterite glass, an amorphous emulsion is formed in which one of the two phases has a similar composition to tetrasilicic fluormica, $K(Mg_{2+x}Li_{1-2x})Si_4O_{10}F_2$. This phase crystallizes rapidly at 600 °C and incorporates most of the network modifiers in the system, thereby significantly increasing the viscosity of the residual glass. Precipitation of diopside ($CaMgSi_2O_6$) follows above 700 °C. Fluormica and diopside subsequently react above 850 °C with the residual glass to produce the stable assemblage potassium fluorrichterite and cristobalite, the latter phase growing above 900 °C.

The residual glass is enriched in silica, alumina, and P_2O_5, resulting in exceptional overall chemical durability. The high thermal expansion coefficient of this glass ceramic allows the development of compressive strength through conventional glazing techniques. Thus, the flexural strength of these glass ceramics (150 ± 15 MPa) can be increased to 200 ± 15 MPa. Glazed fluorrichterite glass ceramics have a translucency and gloss comparable to those of bone china. This glass ceramic is currently being manufactured as high-performance institutional tableware and mugs for the retail Corelle line.

Fluorcanasite Glass Ceramics. Fluorcanasite, $Ca_5Na_4K_2Si_{12}O_{30}(OH,F)_4$, has a crystal

structure that is characterized by four interconnected silica chains running parallel to the b axis, thereby forming a tube. Potassium and sodium ions are located centrally in the tubes, which are interconnected by a network of edge-sharing, calcium and sodium oxyfluoride octahedra.

Fluorcanasite has been crystallized from glasses. It is nucleated by crystallites of CaF_2 and grows in spherulitic fashion on these nuclei. Fluorcanasite glass ceramics can be highly crystalline and virtually monophase. They have flexural strengths near 300 MPa, a modulus of elasticity near 80 GPa, and a thermal expansion coefficient near $12.5 \times 10^{-6}\ K^{-1}$. The strain rupture is about 0.35, a high value for a silicate material. The measured fractured toughness is $5.0\ MPa\ m^{1/2}$, indicating a fracture surface energy of approximately $150\ J/m^2$, higher than that for natural nephrite-jade.

Glass with a fluorcanasite composition can be rolled, pressed, or cast. Ceramming can be completed below 900 °C in less than 1 h. Due to their low silica content, fluorcanasite glass ceramics are inferior in chemical durability to soda – lime glass, but superior to many other building materials. Potential applications for fluorcanasite glass ceramics include architectural cladding, interior partitioning, thin housewares, and magnetic memory disk substrates.

Fotoform, Fotoceram, Visions, Corning Ware, Cercor, Pyrocam, Macor, Dicor, Corelle, Zerodur, and Narumi are registered trade marks.

5. References

1. G. H. Beall: "Design of Glass Ceramics" in H. Jain, A. R. Cooper, K. J. Rao, D. Chakravorty (eds.): *Trends in Science and Technology of Glass*, World Scientific Press, Singapore 1989.
2. G. H. Beall, R. C. Doman: "Glass-ceramics," in: *Encyclopedia of Science and Technology,* vol. 6, Academic Press, New York 1987, pp. 21 – 35.
3. G. H. Beall, *Adv. Ceram.* **18** (1986) 157 – 173.
4. G. H. Beall, H. L. Rittler, *Adv. Ceram.* **4** (1982) 301 – 312.
5. W. Vogel: *Chemistry of Glass,* The American Ceramic Society, Columbus, Ohio, 1985, pp. 325.
6. G. P. Smith, *Mater. Des* . 10 (1989) no. 2, 44 – 63.
7. Z. Strnad, *Glass: Sci. Technol.* **8** (1986) 268.
8. M. H. Lewis (ed.): *Glasses and Glass Ceramics,* Chapman and Hall, London – New York 1989, p. 378.
9. P. F. James, *Adv. Ceram.* **4** (1982) 1 – 29.
10. D. Kashiev, *Surf. Sci.* **14** (1969) 209 – 220.
11. J. E. Dickinson, P. C. Hess, *Geochim. Cosmochim. Acta* **49** (1985) 2289 – 2296.
12. I. S. E. Carmichael, F. J. Turner, J. Verhoogen, *Igneous Petrology,* McGraw-Hill, New York 1974.
13. K. A. Jackson, *Prog. Solid State Chem.* **4** (1967) 53 – 80.
14. M. F. Hochella, G. E. Brown, *J. Am. Ceram. Soc.* **69** (1986) 13 – 18.
15. L. Cartz, J. D. Jorgensen in : D. C. Larson (ed.): *Thermal Expansion 7, Chicago, Illinois,* Plenum Press, New York 1982, pp. 147 – 154.

Glass Fibers → Fibers, 5.Synthetic Inorganic
Glass Wool → Fibers, 5.Synthetic Inorganic
Glassy Metals → Metallic Glasses
Globulins → Blood

Gluconic Acid

Helmut Hustede, Joh. A. Benckiser GmbH, Ladenburg, Federal Republic of Germany

Hans-Joachim Haberstroh, Biochemie Ladenburg GmbH, Ladenburg, Federal Republic of Germany

Elisabeth Schinzig, Biochemie Ladenburg GmbH, Ladenburg, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Chemical Oxidation 3
4.2. Electrochemical Oxidation 3
4.3. Catalytic Oxidation 3
4.4. Methods Involving Biotechnology . . . 4
5. Uses 6
6. Economic Aspects 7
7. Physiology, Product Specifications . . 7
8. References 7

1. Introduction

D-Gluconic acid [*526-95-4*], 1,2,3,4,5-pentahydroxy pentane-1-carboxylic acid, $C_6H_{12}O_7$, M_r 196.16, was discovered in 1870 by Hlasiwetz and Habermann during the oxidation of glucose with chlorine. The substance was isolated in the form of its barium and calcium salts. Several authors subsequently reported that gluconic acid could be obtained by treatment of various mono-, di-, and polysaccharides with oxidizing agents such as elemental halogen, copper(II) or hexacyanoferrate(III) salts, or mercury(II) oxide. Depending on the type of sugar and the oxidant employed, byproducts of the reaction include formic acid, glycolic acid, oxalic acid, and carbon dioxide.

As early as 1880 Boutroux recognized that gluconic acid was produced, together with acetic acid, by the oxidative action of *Acetobacter aceti* on glucose. This characteristic was also found to be associated with numerous other bacteria [1]. Molliard was the first to report the presence of gluconic acid in cultures of *Sterigmatocystis nigra*, now known as *Aspergillus niger* [2]. The currently preferred method for preparing gluconic acid and its derivatives with the aid of *Aspergillus* strains is based on the work of a number of authors [3]. The catalytic activity of the enzyme glucose oxidase was first described by Müller [4].

During the 1930s anodic oxidation was suggested for the preparation of calcium gluconate [5]; proposals for catalytic oxidation of glucose with the aid of air or oxygen followed somewhat later [6].

Sources. Gluconic acid is a naturally occurring substance in humans and other organisms (cf. Chap. 7); it is found in food products such as wine and honey.

2. Physical Properties

Free D-gluconic acid is difficult to prepare in crystalline form. According to Milsom and Meers [7], crystallization of the anhydrous substance is possible below 30 °C, and a monohydrate has been reported to crystallize at 0 – 3 °C with a characteristic crystalline structure, *mp* ca. 85 °C [8].

Anhydrous gluconic acid is a white, odorless, crystalline powder, specific rotation $[\alpha]_D^{20}$ $-6.7°$ [9], $[\alpha]_D^{25}$ $-5.4°$ [10].

Literature values for the melting point of gluconic acid fall in the range 120 – 131 °C. The spread in values is due to the formation of intramolecular anhydrides whose presence lowers the melting point.

Two *lactones* of gluconic acid are known. These exist both in the solid state and also in aqueous solution, where they are in equilibrium with each other and with the free acid.

This complex equilibrium is also influenced by the dissociation of the free acid, which has a

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_449

reported [10] dissociation constant K_A at 25 °C of 1.99×10^{-4} and a pK_A of 3.70 (3.62 according to [11]).

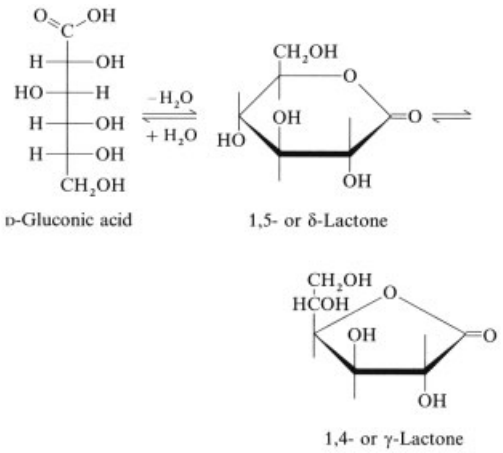

Lactone hydrolysis and formation are both first order with respect to hydrogen ion concentration, with rate constants (k and b, respectively) related as follows [12]:

$$k_{1,5}+b_{1,5}=c_{H^+}\cdot 5.5\times 10^{-2}\,s^{-1}$$
$$k_{1,4}+b_{1,4}=c_{H^+}\cdot 4.3\times 10^{-4}\,s^{-1}$$

Thus, equilibrium is established with the 1,5-lactone roughly 100 times faster than with the 1,4-lactone. The following empirical equation describing the kinetics of hydrolysis for the 1,5-lactone in the pH range 0.6 – 8.2 has been reported [12]:

$$k_{1,5}=c_{H^+}\cdot 4.7\times 10^{-2}+c_{OH^-}\cdot 4\times 10^{3}$$
$$+2.5\times 10^{-4}\,s^{-1}$$

This expression indicates that the OH^- ion strongly catalyzes the opening of the lactone ring, while the H^+ ion is considerably less effective. In the region of neutrality, hydrolysis is quite slow. The rate of hydrolysis over the narrower pH range 3 – 5 is given by

$$-dc_L/dt=kc_L$$

where c_L is the concentration of the 1,5-lactone and $k=2.3\times 10^{-4}\,s^{-1}$. The activation energy for the process is reported to be 62.8 kJ/mol [13].

Gluconic acid is quite soluble in water. At 20 °C, commercial 50 % gluconic acid solution has a pH of 1.82 and a density of 1.23 g/cm^3. In this concentration region, solution density increases almost linearly with increasing concentration. The acid is only slightly soluble in ethanol and virtually insoluble in nonpolar solvents.

Storing gluconic acid over a desiccant at room temperature or heating it above 50 °C leads to the formation of lactones. Pyrolysis occurs above 200 °C.

Gluconic acid 1,5-lactone [*90-80-2*] is a white, crystalline substance with a faintly sweet taste, *mp* 153 °C, $[\alpha]_D^{20}$ +66.2 ° [13]. On standing, both the pH and the specific rotation of a lactone solution decrease as a result of hydrolysis, becoming constant only after equilibrium is established. Approximately 90 g of 1,5-lactone dissolves in 100 mL of water at 20 °C; this increases to about 205 g at 50 °C. Only small amounts dissolve in organic solvents. The *1,4-lactone* [*1198-69-2*] crystallizes as fine needles, *mp* 134 – 136 °C, $[\alpha]_D^{20}$ +67.8 ° [13].

3. Chemical Properties

The action of oxidizing agents such as nitric acid or hydrogen peroxide on either the lactones or the calcium salt of gluconic acid leads under mild conditions to mixtures of oxogluconic acids, with the carbonyl groups at the 2- and the 5-positions. 2-Oxo-D-gluconic acid [*669-90-9*] is the principal product of anodic oxidation, oxidation with sodium chlorate in acid solution, or fermentation by species such as *Acetobacter suboxydans, A. xylinum,* or *A. gluconicum.* Treatment with concentrated nitric acid or with N_2O_4 results in formation of glucaric acid, $HOOC(HCOH)_4COOH$ [*87-73-0*].

Hydrogenation of gluconic acid in aqueous solution over a platinum oxide catalyst results in a modest yield of D-glucose, whereas the 1,5-lactone undergoes this reaction in high yield. Refluxing with concentrated hydriodic acid in the presence of red phosphorus causes reduction to hexanoic acid [*142-62-1*].

The six functional groups of gluconic acid can, in principle, react with a variety of reagents, such as alcohols, acids, etc. Nevertheless, the resulting derivatives tend to be stable only if reaction is complete. Partial reaction leads to nonuniform mixtures that are sensitive to hydrolysis; such reactions have little significance.

In contrast, considerable interest exists in the ability of gluconic acid and its alkali salts to form complexes with polyvalent cations; some of these complexes are very stable (see Chap. 5). Nuclear magnetic resonance spectra suggest that complex formation involves both the carboxyl and the hydroxyl groups.

4. Production

For commercial purposes, D-gluconic acid and its salts are prepared exclusively by the oxidation of glucose or glucose-containing raw materials. The oxidation method may be chemical, electrolytic, catalytic, or biochemical.

4.1. Chemical Oxidation

Chemical methods have the disadvantage of limited specificity even under carefully controlled and optimized reaction conditions, which results in unsatisfactory yields (60 – 80 %) and undesirable byproducts. Isolation and purification of the product are correspondingly difficult. Hydrogen peroxide [14], ozone [15], and oxygen [16] have been the oxidants of choice, presumably because of ecological problems associated with other oxidizing agents and their reduction products.

4.2. Electrochemical Oxidation

Similar problems accompany electrochemical procedures, most of which are based on the oxidation potential of halogens. The usual approach is to electrolyze a glucose solution containing bromide (ca. 10 mol % relative to glucose) at a current density of 1 – 20 A/dm^2. Carbonates (preferably calcium carbonate) or hydroxides are added to neutralize the resulting acid [17]. Electrical current yields are 80 – 86 %; the product yield based on glucose is 80 – 97 %. The rapid increase in the cost of electricity in recent years has made the electrolytic process noncompetitive.

4.3. Catalytic Oxidation

In contrast, recent developments have led to great improvements in the technique of oxidizing glucose with oxygen or air and serious consideration has been given to adopting this strategy on an industrial scale. Oxidation is carried out, with the aid of an appropriate catalyst, on a glucose solution whose concentration is 1 – 2 mol/L. The pH is held between 8 and 11 (preferably 9 – 10) by continuous addition of alkaline solution.

The catalysts originally employed were finely divided platinum-group metals suspended on activated charcoal, aluminum oxide, or some other carrier [18]. The corresponding reaction mechanism for a Pt – C catalyst has largely been elucidated [19].

The most important observations with respect to the catalytic reaction include a rapid decrease in activity of the catalyst from its initial levels (a decline that cannot be completely reversed despite efforts at reactivation), the need for highly purified glucose solutions, and the formation of various byproducts as a result of insufficient catalyst specificity and the tendency of the substrate to undergo side reactions under the prevailing reaction conditions.

The foregoing disadvantages have been circumvented through a series of developments with respect to the catalysts employed. Thus, platinum-group metals have been found to become more active and more selective if they are doped with lead [20], [21], selenium [22], thallium [21], or bismuth [21], [23]. The preferred carrier is activated charcoal.

The catalyst described in [23] permits conversion of a 2 M glucose solution to gluconic acid with oxygen at 50 ± 1 °C and pH 9.5 ± 0.2 (neutralization with 40 % sodium hydroxide solution) in >99.5 % yield within 1 h. The process is also highly selective: product purity (based on sodium gluconate) is >99.5 %. In addition, no special effort need be taken to purify the glucose solution. Loss of catalytic activity is minimal, permitting the catalyst to be recycled many times without an intervening reactivation step.

The economic viability of the process depends largely on the activity, selectivity, lifetime, and cost of the catalyst, as well as on the measures required for product purification and the energy demands.

Despite earlier suggestions to the contrary, glucose cannot be oxidized photochemically to gluconic acid [24].

4.4. Methods Involving Biotechnology

The principal organisms employed today for large-scale biological preparation of gluconic acid are *Aspergillus niger* and *Gluconobacter suboxydans.*

Aspergillus Niger Process. The biochemistry of gluconic acid formation by *Aspergillus niger* is illustrated in the following diagram [25], [26]:

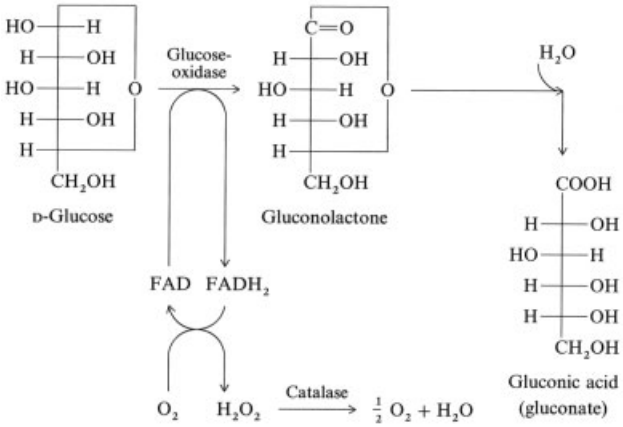

where FAD and $FADH_2$ are the oxidized and reduced forms, respectively, of flavine – adenine dinucleotide. β-D-Glucose is oxidized catalytically to glucono-1,5-lactone by the enzyme glucose oxidase (see also → Enzymes). Subsequent hydrolysis of the lactone to gluconic acid is to some extent spontaneous, but the enzyme gluconolactonase also plays a role.

Glucose oxidase (E.C. 1.1.3.4) [*9001-37-0*], M_r 186 000, contains two FAD moieties as prosthetic groups. These are responsible for the abstraction of hydrogen atoms from glucose (to form $FADH_2$), which ultimately combine with oxygen to produce hydrogen peroxide. This is in turn cleaved to water and oxygen by the enzyme catalase. Both glucose oxidase and catalase are present as endoenzymes in the fungal mycelium [27], [28].

In addition to glucose oxidase, *Aspergillus niger* may have other enzymes that also lead to gluconic acid formation, comparable to the glucose dehydrogenase and phosphomonoesterases found in *A. oryzae* [29], [30].

Metabolic investigation has shown fermentation to be a three-stage process [31]: a brief lag phase, followed by an acceleration phase during which glucose oxidase activity rises nearly logarithmically, and finally a stationary phase, which is characterized by a constant rate of gluconic acid formation. The optimum pH for the reaction lies near 5.6 [32].

Production Parameters. Fermentative synthesis of sodium gluconate requires a sterilized substrate; a typical formulation is as follows: glucose 250 – 350 g/L, $MgSO_4 \cdot 7H_2O$ 0.2 – 0.3 g/L, KH_2PO_4 0.2 – 0.3 g/L, and $(NH_4)_2HPO_4$ or urea 0.4 – 0.5 g/L. The substrate may be sterilized either in batches at 110 °C with a residence time of 45 min or continuously under conditions providing several minutes exposure to a temperature of 135 – 150 °C. The sterile medium is then transferred to the fermentation vessel, adjusted to a pH of 4.5 – 5.0, and inoculated with cultured microorganism.

The culture medium contains only about 100 g of glucose per liter, but as much as twice the above amounts of nutrient salts, an increased level of nitrogen compounds, and a growth-stimulating additive consisting of 0.2 – 0.4 g of corn steep powder (a residue obtained by evaporating the water in which maize has been soaked). A lyophilized permanent culture is used to grow the conidia. This culture is first activated with a specific growth medium in culture tubes. After several subcultures have been prepared, the organism is introduced into a culture flask

containing a special medium that encourages the formation of conidia. After an incubation period of 5 – 10 d the conidia are harvested, either in dry form or as an aqueous suspension, and used to inoculate the preculture.

During the production phase, the system is maintained at 30 – 32 °C and pH 5.5 – 6.5. The pH is controlled by continuous neutralization with 30 – 50 % sodium hydroxide solution, the consumption of which is an indication of reaction progress. Depending on the starting concentration of glucose, fermentation continues for a period of 40 – 100 h. Under conditions of maximum specific rate of product formation, 20 – 30 mmol of gluconic acid is generated per hour for each gram of biomass (dry weight) [33], [34].

After fermentation is complete, the microorganisms are removed by filtration and washed. The filtrate is decolorized with activated charcoal, subjected to fine filtration, and then either evaporated and crystallized or directly spray-dried.

The fungal mycelium residue may be used for the isolation of glucose oxidase or it may simply be burnt [35].

Various attempts have been made to use mathematical models to optimize the above process [33], [34], [36], [37]. These efforts have shown that both the specific rate of oxygen uptake and the specific rate of formation of gluconic acid increase approximately linearly with increasing concentration of dissolved oxygen or increasing oxygen partial pressure. If the effects of the brief growth phase and retentive metabolism are ignored, the productive phase follows simple Michaelis – Menten kinetics, with K_M 3.6×10^{-4} mol/L at 33 °C and pH 6.6 [34]. At 25 °C and pH 5.6, $K_M = 2.0 \times 10^{-4}$ mol/L [32]. Isolated glucose oxidase under these conditions has K_M 4.8×10^{-4} mol/L.

To ensure economically viable yields of more than 80 %, an adequate oxygen supply (0.1 L of oxygen per liter of solution per minute) must be maintained, and gas distribution within the fermentor must be optimized. The oxygen partial pressure may be increased by conducting fermentation at elevated pressure or by employing air enriched in oxygen [38].

Product concentration in the fermentation solution can be increased if glucose is introduced at a higher concentration and pH control is discontinued near the end of the fermentation [39]. These conditions result in formation of a mixture of sodium gluconate and free gluconic acid, which has a considerably higher solubility than the salt alone. Such a solution (e.g., Naglusol) is suitable for commercial use in technical applications as soon as the mycelium has been removed.

Gluconobacter Suboxydans Process. The *Gluconobacter suboxydans* process is rather different, in that conversion of β-D-glucose to glucono-1,5-lactone occurs through the action of two glucose dehydrogenases: a particulate quinoprotein glucose dehydrogenase and a soluble NADP-dependent glucose dehydrogenase [40], [41].

Gluconobacter suboxydans fermentation is distinguished by a high affinity for oxygen relative to the *Aspergillus* process: simple aeration suffices to ensure a high rate of glucose conversion. In addition, the process displays greater tolerance with respect to acidity, permitting the isolation of free gluconic acid directly from the fermentation solution [42].

Other Methods. Another approach that leads directly to gluconic acid uses acidophilic methylotropic bacteria (*Acetobacter methanolicus*) under nonsterile conditions [43]. Other procedures have been described that involve immobilized enzymes, microorganisms, or both [44–47]. To date, none of these methods is used for commercial production of gluconate, largely because of rapid deactivation of the enzymes or limitations with respect to oxygen diffusion.

Downstream Processing. Ordinary commercial grades of free gluconic acid may be prepared from the corresponding sodium salt either through cation exchange or electrodialysis. Passage of a mixture of sodium gluconate and free gluconic acid over a strong anion-exchange resin results in adsorption of gluconate [48].

Glucono-1,5-lactone crystallizes from a supersaturated aqueous gluconic acid solution between 36 and 57 °C [49]. A two-step continuous crystallization is said to result in increased productivity per unit volume under conditions that are favorable in terms of energy consumption [50]. In the first step, a crystalline suspension (3 – 10 % crystalline material) is produced in a supersaturated gluconic acid solution held at

65 – 75 °C. The second step involves controlled cooling to 40 – 45 °C, followed by a period of crystal maturation at constant temperature. A portion of the resulting crystalline suspension is then recycled to the first step.

According to [51], the 1,5-lactone may also be obtained from an aqueous gluconic acid solution by dehydration involving azeotropic distillation with alcohols, followed by crystallization from the alcohol-containing residue.

Crystallization above 70 °C leads to the 1,4-lactone.

5. Uses

The primary applications of gluconic acid follow from its most important characteristic: it is a weak acid capable of dissolving the oxides, hydroxides, and carbonates of polyvalent cations without attacking metallic or nonmetallic surfaces. The value of this property is further enhanced by the fact that the acid forms water-soluble complexes with such cations. Gluconic acid is thus exceptionally well suited for use in removing calcareous and rust deposits from metals or other surfaces, including beer and milk scale on galvanized iron, magnesium alloys, or stainless steel. The compound is also used in the textile industry together with magnesium salts as a stabilizer for peroxide bleach baths.

Because of its physiological properties, D-gluconic acid has also been used in both the food and beverage and the pharmaceutical industries. Thus, low concentrations (0.02 – 0.1 %) of the substance effect the inversion of sucrose without causing the resulting fructose to undergo further reaction. Trace elements are usually administered in the form of gluconate salts because such compounds are readily absorbed into the body and are well tolerated. Potassium gluconate also has certain pharmaceutical applications and may be dispensed either in anhydrous form or as the monohydrate.

In many applications, gluconic acid 1,5-lactone is a convenient substitute for the free acid. The lactone offers particular advantages in circumstances involving acidic conditions over a long period, e.g., in the preparation of pickled goods and frankfurter-type sausages, or in curing fresh sausage. Another example is its use as a leavening agent in baked goods.

The most common gluconate is sodium gluconate, M_r 218.14, $[\alpha]_D^{25}$ +12.0° [10], whose solubility behavior is as follows:

t, °C	0	20	50	80	100
c, g/100 g H_2O	43	60	85	133	160

Sodium gluconate, like the free acid, forms complexes with metal cations. The stability of such complexes often increases considerably with increasing pH [13]. This salt also displays cleansing characteristics with respect to surfaces of widely varying type and structure.

Combinations of sodium gluconate with sodium carbonate or sodium hydroxide solution are useful for removing grease and corrosion from aluminum, rust from steel, oxide coatings from copper and copper alloys, and for etching aluminum. Strongly alkaline gluconate baths permit the removal of zinc coatings from metallic objects, castings, and scrap metal. Such baths are characterized by a high zinc capacity and a correspondingly long life.

Alkaline sodium gluconate solutions at 95 – 100 °C are effective agents for the rapid removal of paint and varnish, without damaging underlying surfaces. Gluconate is also useful in the pretreatment of certain surfaces, e.g., in the galvanic deposition of nickel – cobalt brazing surfaces onto aluminum or in the baths required for preparing smooth, shiny surface platings of nickel, tin, and zinc. In the latter application, gluconate is increasingly used to replace the highly toxic cyanide ion.

Mixtures of gelatin and sodium gluconate are used as sizing agents in the paper industry because they result in a product with increased acid resistance. Textile manufacturers employ gluconate (sometimes in combination with polyphosphates) for desizing polyester or polyamide fabrics and for finishing natural cellulose fibers.

Sodium gluconate is a component of commercial cleansers because of its absolute stability to hydrolysis at high temperature and high pH, as well as its sequestering properties with respect to hardening agents in water. For example, sodium gluconate is used in bottle washing and in cleaning aluminum surfaces (e.g., building facades, aircraft, and containers).

Concrete manufacturers have found sodium gluconate to be a highly effective agent for retarding the curing process. Addition of 0.02 – 0.2 wt % of this substance (relative to cement) produces a very homogeneous concrete with high resistance to water, frost, and cracking. This material is also easily worked and shows increased stability upon setting. Other advantages accompanying the introduction of gluconate include better flow properties for the wet concrete mixture, increased wettability with respect to iron reinforcing structures, and maintenance of plasticity despite reduced water content.

Relative to other organic and inorganic complexing agents, gluconic acid, its lactone, and its salts offer one major advantage: they are all destroyed effectively and quickly by biological wastewater treatment, whether adapted or not. The same applies [52] to the soluble complexes of gluconate with aluminum, copper, iron, and zinc, although biodegradation of the chromium gluconate complex occurs more slowly. Metal ions released during gluconate complex degradation are largely removed from the purified wastewater, either by precipitation or by adsorption on the sludge. This circumstance reduces a major threat that might appear to accompany the use of gluconate as a complexing agent: mobilization or remobilization of heavy metals into surface water [53].

Destruction of heavy-metal gluconate complexes in wastewater may occur not only through biodegradation but also by simple hydroxide precipitation [54], in which iron(III) acts as a precipitant at a pH of 9 – 10; see also [55].

6. Economic Aspects

Annual worldwide production capacity for gluconic acid and its derivatives is estimated to be 60 000 t. The bulk of production (85 %) is in the form of sodium gluconate and other alkali gluconate salts. Present consumption levels are, however, significantly lower, resulting in only 60 – 70 % utilization of existing capacity. Principal manufacturers include: Akzo-Chemie, Amersfoort, Holland; Biochemie Ladenburg, Ladenburg, Federal Republic of Germany; Finnish Sugar Co. Ltd., Espoo, Finland; Fujisawa Pharmaceutical Co., Osaka, Japan; Pfizer, New York, New York; Roquette Frères, Lille, France.

7. Physiology, Product Specifications

D-Gluconic acid and its 1,5-lactone are important intermediates in carbohydrate metabolism. These compounds serve two important functions: (1) they contribute to the synthesis of reduced nicotinamide – adenine dinucleotide phosphate (NADPH), which is required in the biosynthesis of fatty acids and steroids, and (2) they lead to the formation of ribose 5-phosphate, which is used in nucleic acid synthesis [56].

The gluconate salts of sodium [*527-07-1*] calcium [*299-28-5*], copper [*13005-35-1*], iron(II) [*299-29-6*], and manganese [*6485-39-8*] are on the list of GRAS substances [57], [58]. The gluconates of potassium [*299-27-4*], magnesium [*3632-91-5*], and zinc [*4468-02-4*] have not been shown to have teratogenic, mutagenic, or carcinogenic properties [56].

Under the food and beverage law of the Federal Republic of Germany, D-gluconic acid is classified as a foodstuff and not regarded as an additive. Purity standards for gluconate salts are prescribed in the U.S. Food Chemical Codex (FCC III) of 1981, the U.S. Pharmacopeia (1985), and the European Pharmacopeia. Methods of analysis can be found in [32] and [59].

8. References

1. W. Henneberg: *Handbuch der Gärungsbakteriologie*, 2nd ed., vol. 2, Springer Verlag, Berlin 1926.
2. M. Molliard, *C.R. Hebd. Seances Acad. Sci.* **174** (1922) 881.
3. K. Bernhauer: *Die oxydativen Gärungen,* Springer Verlag, Berlin 1932.
4. D. Müller, *Biochem. Z.* **199** (1928) 156.
5. H. S. Isbell, L. Frush, F. J. Bates, *J. Res. Nat. Bur. Stand.* **6** (1931) 1148: **8** (1932) 572.
6. M. Busch, DE 702 729, 1941.
7. P. E. Milsom, J. L. Meers in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 3,** Pergamon Press, Oxford-New York-Toronto-Sydney-Frankfurt 1985, pp. 681 – 700.
8. Noury and van der Lande, GB 902 609, 1962 (D. W. van Gelder); *Chem. Abstr.* **57** (1962) 13020 g.

9. F. J. Prescott, J. K. Shaw, J. P. Bilello, G. O. Cragwall, *Ind. Eng. Chem.* **45** (1953) 338 – 342.
10. D. T. Sawyer, J. B. Bagger, *J. Am. Chem. Soc.* **81** (1959) 5302 – 5306.
11. L. H. Skibsted, G. Kilde, *Dan. Tidsskr. Farm.* **45** (1971) 320 – 324; *Chem. Abstr.* **76** (1972) 72 722 d.
12. R. E. Mitchell, F. R. Duke, *Ann. N.Y. Acad. Sci.* **172** (1970) 129.
13. D. T. Sawyer, *Chem. Rev.* **69** (1964) 633.
14. M. Boruch, T. Pierzgalski, *Zesz. Nauk. Politech. Lodz. Chem. Spozyw.* **32** (1977) 33; *Chem. Abstr.* **89** (1978) 131 450 g.
15. Rin Kagaku, Kogyo Co., Ltd., JP 7 330 618, 1973 (T. Nakagawa, Y. Murai, N. Kanai); *Chem. Abstr.* **80** (1974) 71 052 f.
16. Kao Soap Co., DE-AS 2 344 990, 1973 (Y. Nakagawa, S. Koshigaya, K. Sato, S. Hakozaki, C. Funabashi).
17. H. V. K. Udupa et al., *Proc. Semin. Elektrochem.* **14th** (1974) 79. R. Ohme, P. Kimmerl, DD 104 779, 1974.
18. Johnson, Matthey and Co., Ltd., GB 1 208 101, 1970 (G. J. K. Acres, A. E. R. Budd); *Chem. Abstr.* **74** (1971) 14 347 h. Kawaken Fine Chemicals Co., Ltd., JP 8 007 230, 1980 (H. Saito, M. Nozue); *Chem. Abstr.* **93** (1980) 72 211 n. Mitsui Toatsu Chemicals Inc., JP 7 652 121, 1976 (T. Kiyoura, T. Kimura, T. Sugiura); *Chem. Abstr.* **85** (1976) 160 467 r. Kawaken Fine Chemicals Co., Ltd./Kao Corp., JP 5 872 538, 1983; *Chem. Abstr.* **99** (1983) 88 543 g.
19. J. M. H. Dirkx, H. S. van der Baan, *J. Catal.* **67** (1981) 1.
20. Asahi Chemical Industry Co., Ltd., DE-OS 2 936 652, 1980 (J. Nishikido, N. Tamura, Y. Fukuoka); *Chem. Abstr.* **93** (1980) 150 595 g. Kao Corp./Kawaken Fine Chemicals Co., Ltd., JP 59/205 343, 1984; *Chem. Abstr.* **102** (1985) 149 721 t.
21. N. Xonoglou, G. Kokkinidis, *Bioelectrochem. Bioenerg.* **12** (1985) no. 5 – 6, 485 – 498; *Chem. Abstr.* **102** (1985) 149 654 y.
22. Kawaken Fine Chemicals Co. Ltd./Kao Corp., JP 60/92 240, 1985; *Chem. Abstr.* **103** (1985) 88 175 g.
23. Kawaken Fine Chemicals Co. Ltd./Kao Corp., EP 0 142 725, 1984 (H. Saito, S. Ohnaka, S. Fukuda).
24. P.-P. Degelmann, G. Ehmig, H. Hofmann, *Zuckerindustrie (Berlin)* **107** (1982) 117.
25. G. Gottschalk: *Bacterial Metabolism,* 2nd ed., Springer Verlag, New York-Berlin-Heidelberg-Tokyo 1986, p. 172.
26. H. G. Schlegel: *Allgemeine Mikrobiologie,* Thieme Verlag, Stuttgart-New York 1985, p. 327.
27. E. E. Bruchmann: *Angewandte Biochemie,* E. Ulmer, Stuttgart 1976, p. 64.
28. J. P. van Dijken, M. Veenhuis, *Eur. J. Appl. Microbiol. Biotechnol.* **9** (1980) no. 4, 275 – 283.
29. H. M. Müller, *Zentralbl. Bakteriol. Parasitenkd. Infektionskr. Hyg. Abt. 2 Naturwiss. Allg. Landwirtsch. Tech. Mikrobiol.* **132** (1977) 14 – 24; *Chem. Abstr.* **87** (1977) 1598 j.
30. T.-G. Bak, *Biochim. Biophys. Acta* **139** (1967) 277 – 293.
31. B. Drews, H. Smalla, *Branntweinwirtschaft* **109** (1969) 21 – 27; *Chem. Abstr.* **71** (1969) 79 730 k.
32. H. U. Bergmeyer: *Methoden der enzymatischen Analyse,* 3rd ed. vol. 1, Verlag Chemie, Weinheim 1974, p. 486.
33. T. Takamatsu, S. Shioya, T. Furuya, *J. Chem. Technol. Biotechnol.* **31** (1981) 697.
34. S. Fröhlich, *Inst. Ferment. u. Brauwesen,* Dipl.-Arbeit, TU Berlin 1981.
35. L. B. Lockwood in J. E. Smith, D. R. Berry (eds.): *Filamentous Fungi,* **vol. 1,** Wiley, New York 1975, pp. 140 – 157.
36. Y. Miura, K. Tsuchiya, H. Tsusho, K. J. Miyamoio, *Ferment. Technol.* **48** (1970) 795.
37. M. Reuss, S. Fröhlich, B. Kramer, K. Messerschmidt, H. Niebelschütz, *3rd Eur. Congr. Biotechnol.,* **vol. 2, 455 – 460,** Verlag Chemie, Weinheim 1984.
38. L. B. Lockwood in H. J. Peppler, D. Perlmann (eds.): "Production of Organic Acids by Fermentation," *Microbial Technology I* , Academic Press, London – New York 1979, pp. 355 – 387.
39. Pabst Brewing Co./Joh. A. Benckiser GmbH, DE 1 817 907, 1979 (J. Ziffer, A. S. Gaffney, S. Rothenberg, T. J. Cairney).
40. W. Olijve, J. J. Kok, *Arch. Microbiol.* **121** (1979) 283 – 297.
41. P. R. Levering, C. Lucas, J. v. Heijst, W. Olijve, *4th Eur. Congr. Biotechnol. Amsterdam,* **vol. 3, 477,** Verlag Chemie, Weinheim 1987.
42. J. B. M. Meiberg, H. A. Spa, *Antonie van Leeuwenhoek* **49** (1983) 89 – 90.
43. Akad. der Wissenschaften der DDR, DD 236 754 A1, 1986 (W. Babel, D. Miethe, V. Iske, K. Sattler, H.-P. Richter, J. Schmidt, R. Duresch); *Chem. Abstr.* **106** (1987) 31 387 t.

44. K. Buchholz in H. Dellweg (ed.): *Fifth Int. Ferment. Sympos. Berlin 1976,* "Abstracts of Papers," Verlag Versuchs- und Lehranstalt für Spiritusfabrikation und Fermentationstechnologie im Institut für Gärungsgewerbe und Biotechnologie zu Berlin 1976, p. 273.
45. G. Richter et al., *Starch Stärke* **31** (1979) no. 12, 418 – 422.
46. Boehringer Mannheim, DE-AS 2 214 442, 1972 (K. U. Bergmeyer, D. Jaworek).
47. T. Döppner, W. Hartmeier, *Starch Stärke* **36** (1984) no. 8, 283 – 287.
48. Roquette Frères, FR 2 054 829, 1971 (M. Huchette, F. Devos).
49. Daiichi Kogyo Seiyaku Co., Ltd., JP 75 123 674, 1975 (T. Yamauchi, K. Shimizu); *Chem. Abstr.* **85** (1976) 6008 k.
50. Roquette Frères, EP 0 203 000, 1986 (J. B. Lelen, P. Lemay).
51. Daicel Chemical Industries, Ltd., JP 60/45 570, 1985; *Chem. Abstr.* **103** (1985) 123 356 y.
52. G. Magureanu, *Z. Wasser Abwasser Forsch.* **14** (1981) 10.
53. G. Müller, *Umsch. Wiss. Tech.* **79** (1979) 778.
54. D. Ott, *Galvanotechnik* **73** (1982) no. 4, 339;no. 5, 453.
55. D. Ott, C. J. Raub, *Proc. Electrochem. Soc.* **83 – 12** (1983) (Proc. Symp. Electroplat. Eng. Waste Recycle New Dev. Trends 1982, 399 – 407).
56. Federation of American Societies for Experimental Biology (FASEB) PB 288 675 (1978).
57. Federal Register, vol. 47, no. 210, 49 028 (1982).
58. Federal Register, vol. 49, no. 114, 24 118 (1984).
59. R. Lenz, G. Zoll in I. Molnar (ed.): "Pract. Aspects Mod. High Perform. Liq. Chromatogr.," *Proc., Meeting Date 1981,* De Gruyter, Berlin 1983, pp. 355 – 361.

Glucose and Glucose-Containing Syrups

FRED W. SCHENCK, Consultant (retired), Sun City, Arizona 85373–1113, United States

1. Introduction and History 1
2. Properties of Glucose 3
2.1. Structure 3
2.2. Physical Properties 5
2.3. Chemical Properties 5
3. Raw Materials 5
4. General Principles of Starch Hydrolysis 6
4.1. Production and Properties of Acid Glucose Syrups 8
4.2. Production and Properties of Acid – Enzyme Glucose Syrups 10
4.3. Production and Properties of Enzyme – Enzyme Glucose Syrups . . 10
5. Production of Solid Glucose 12
5.1. Production of Total Sugars 12
5.2. Production of Pure Crystalline Products 12
6. Glucose – Fructose Syrups 14
6.1. Properties 14
6.2. Production 15
6.3. High-Fructose Syrups from Inulin . . 16
7. Analysis 17
8. Legal Aspects 18
9. Storage and Transportation 18
10. Uses 19
11. Economic Aspects 19
12. Physiological Properties and Toxicology of Glucose 20
13. References 21

1. Introduction and History

The saccharides are among the most common organic compounds found in nature and glucose is by far the most abundant monosaccharide. D-Glucose (D-glucopyranose) [*50-99-7*], $C_6H_{12}O_6$, M_r 180.159, is the primary six-carbon sugar (hexose) found in the juice of most fruits and berries. Starch (→ Starch) and cellulose (→ Cellulose) are composed of anhydroglucopyranose units connected through α- and β-linkages, respectively, to form long-chain polymers. Thus, a considerable portion of the human diet is composed of glucose in either the free or the polymerized state.

When α-linked polymers of glucose (starch) are ingested, acid and enzymatic processes convert these macromolecules to glucose in much the same way that it is made commercially. The digestive process begins in the mouth with the action of salivary amylase. Final digestion is accomplished by the action of numerous other enzymes in the small intestine.

D-Glucose was probably known to early humans because it appears as granules in honey due to crystallization. However, despite the abundance of glucose in nature, the disaccharide sucrose has historically been the primary sweetener for humans. Sucrose is sweeter than glucose, and very pure crystals can be recovered from syrups extracted from sugar cane and sugar beets (→ Sugar).

Most glucose, on the other hand, is bound up as starch, requiring complicated manufacturing processes to release its sweetness. The growth of the glucose sweetener industry has been based upon a classic response to market needs for improved purity and functionality.

Pure glucose was first obtained from grape juice by the German chemist A. S. MARGGRAF in 1747 [1]. PROUST was one of the first to precipitate α-D-glucose crystals from grape juice in 1801 [2]. The synonym for D-glucose, grape sugar, owes its origin to this process.

The first sweetener created from starch was the disaccharide maltose made in China about 1000 B.C. by the enzymatic hydrolysis of rice [3]. By 1781, it was known in Europe that starches could be treated with acids to produce a soluble, sweet-flavored substance [4]. This knowledge was little more than a laboratory curiosity until the Napoleonic Wars when the blockade of cane sugar imports from the West

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_457.pub2

Indies provided incentive for the development of alternative sweeteners.

In 1811, G. S. C. Kirchoff, who was looking for a substitute for gum arabic (used in the manufacture of porcelain), overcooked a mixture of potato starch and sulfuric acid [5]. The sweet, syrupy substance that resulted earned him a lifetime pension of 500 rubles per year from the Russian Emperor.

Kirchoff recognized the commercial potential of his discovery and proceeded to improve the process. He succeeded in preparing a syrup that crystallized partially on standing, and from which a solid product (impure D-glucose) could be obtained by pressing the mass in cloth sacks. This solid material contained large quantities of uncrystallized syrup and was quite impure. Nevertheless, serious attempts were made to manufacture the product on a large scale. However, the defeat of Napoleon at Waterloo made cane sugar available once again and the venture was abandoned.

The French chemist Saussure determined that the acid conversion of starch was achieved through a hydrolytic reaction that cleaved the linkages between the anhydroglucopyranose units in starch by adding one molecule of water to each unit to produce the basic building block, D-glucose. It was also recognized that candies manufactured with glucose syrups were not cloyingly sweet as were those made from sucrose, thus allowing delicate flavors to reach the palate. As a result, the "starch sugar" industry developed in Europe (especially in Germany) and in the United States [6].

Early efforts to improve the quality of "starch sugars" were attempts to control the nonspecific nature of the acid hydrolysis. The market needed purer glucose, but early commercial production (1880 – 1920) of crystalline glucose required either repeated crystallization steps or crystallization from nonaqueous solvents. The only solid form of glucose available in quantity at low price at this time was a cast product; starch was acid converted to as high a degree as possible, evaporated to a thick syrup, and poured into molds where it hardened into solid blocks. It was subsequently removed from the molds, chipped into small pieces, and sold.

In the United States in 1923, William B. Newkirk (Corn Products Refining Company) received a patent for developing a commercially feasible process for the production of pure crystalline glucose [7]. By careful control of the conditions during the crystallization step, large, easily separable crystals could be produced economically and an industry was born.

Although glucose and glucose-containing syrups can have functional advantages over sucrose (higher viscosity, greater solubility, bland flavor, etc.) in certain applications, the unmatched sweetness of sucrose prevented glucose-containing syrups from capturing more than about one-fifth of the caloric sweetener market in the United States. In many of its uses, sucrose in aqueous solution is treated with either acid or enzymes to produce invert syrup consisting of a 50 : 50 glucose – fructose mixture (→ Fructose). Thus, in a highly converted glucose syrup, isomerization of half of the glucose to fructose would make starch-based syrups about as sweet as those made from sucrose.

Glucose isomerization under alkaline conditions (see Section 2.3), which increases sweetness somewhat, results in a highly colored product with off-flavor. In 1957, a xylose isomerase enzyme was reported to convert glucose to fructose [8], and a US patent was issued to the Corn Products Refining Company in 1960 [9]. However, no industrial activity was immediately pursued in the United States, and glucose isomerase research shifted to Japan in the 1960s [10].

Once the economical enzymatic conversion of glucose to fructose had been demonstrated by the Japanese, the technology moved back to the United States. In 1965, the Clinton Corn Processing Company (United States) and the Japanese Agency of Industrial Science and Technology worked together to commercialize fructose-containing syrups made in the United States by enzymatic isomerization of glucose. At about the same time, the first batchwise production began in Japan.

The degree of enzymatic isomerization of glucose to fructose depends on a number of operating parameters. Because higher fructose content is desired, higher operating temperature and/or increased reaction time is required. However, both of these process conditions reduce the effective enzyme productivity. Although enzymatic isomerization of glucose to fructose in excess of 50 % db (dry basis) is possible, conversion to a 42 % db fructose content was chosen as the best balance between increased sweetness

and process economics. The first commercial shipment of 42 % fructose syrup in the United States was made in 1968.

This new sweetener began to replace sucrose in many liquid applications (see Table 1), but it was not sweet enough to satisfy the demands of the most important U.S. soft drink producers. By use of chromatographic separation techniques, the 42 % fructose syrup was enriched to a fructose content in excess of 90 % db. Blending this enriched syrup with 42 % fructose syrup, the sweetness level of total invert syrup was matched at 55 % fructose content. More than 50 % fructose was required because the various glucose polymers remaining in the syrup, due to incomplete starch hydrolysis have a lower sweetening power.

The first large-scale production of 55 % fructose syrup began in 1978; by 1984, fructose syrups derived from glucose had replaced sucrose in nearly 100 % of the soft drinks produced in the United States, substantially reducing sucrose consumption. Today, glucose-based sweeteners account for more than half of the total caloric sweeteners consumed in the United States (see Fig. 1) [12].

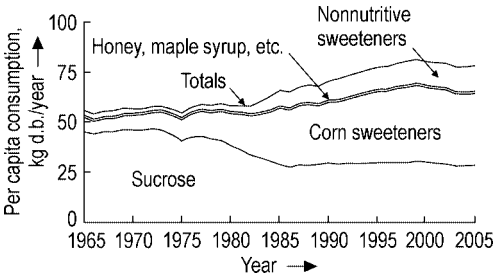

Figure 1. Per capita US sweetener consumption by type 1965 – 2005

North America, led by the United States, is the dominant leader in the production and consumption of glucose – fructose syrups, consuming 74.1 % of the world production of 11.8×10^6 t (dry basis) today, followed by Asia and Oceania at 14.8 %, with European consumers using 7.1 % and Latin America and Africa accounting for the remaining 4 % [13]. This market dominance is expected to continue for the forseeable future the major market growth (on a percentage basis) expected in Asia [14] has not occurred, because genetically modified organism (GMO) regulations have complicated maize procurement and low sugar prices have reduced substitution incentives [15]. Consumption of glucose – fructose syrups increased in the United States, on average, by 19 % annually between 1980 and 1985 and grew much faster than population (4 % annually) between 1990 and 1995. The per capita nutritive sweetener consumption in the United States peaked about 1999; since then, per capita consumption of glucose – fructose syrups has fallen about 1 % per year [16]. Imports of sugar-containing products may explain some of this reduction but reasons may also include reduced economic activity and a drop in the consumption of baked goods and cereal products in the United States [17].

Chromatographic enrichment to $> 90\,\%$ fructose syrups has allowed the economical production of crystalline fructose (→ Fructose) [18]. New drying and encapsulation technology claims to be able to produce solid glucose–fructose syrups [19], so that they can be used in certain dry applications as well.

2. Properties of Glucose

2.1. Structure

At the end of the 19th century, the general structure of glucose was elucidated by Emil Fischer. He found that the six carbon atoms of the glucose molecule are linked in a straight chain. At first glucose was thought to be an open polyhydroxy aldehyde. However, more than 99 % of the molecules exist as a six-membered cyclic hemiacetal which is formed by reaction of the aldehyde group with the hydroxy group at C 5. As a result, glucose does not undergo the specific addition reactions (e.g., with ammonia or sodium bisulfite) that are typical for aldehydes.

Glucose predominantly occurs in nature in the form of the D-enantiomer which is therefore important commercially. D-Glucose is generally believed to exist in three crystalline forms: anhydrous α and β-D-glucose crystals are both orthorhombic while α-D-glucose monohydrate crystals are monoclinic (see Fig. 2). However, a fourth form, thought to be a hydrated form of β-D-glucose, has been reported [22].

Table 1. Distribution by industry of crystalline glucose, glucose – fructose, and glucose-containing syrups in the United States (in 10^3 t, commercial basis) [11]

Year	Baking	Beverage	Canning	Confectionery	Dairy	Total
			Crystalline glucose			
1965	177	9	23	36	5	440
1970	172	9	23	50	7	515
1975	159	18	14	63	7	515
1980	50	68	5	54	2	430
1985	57	82	2	59	1	460
1990	75	100	4	75	1	500
1997	135	118	5	90	2	560
2002	145					520
			Glucose – fructose syrups			
1970	18	41	9	1	5	75
1975	127	281	63	5	36	700
1980	363	1093	236	14	141	2880
1985	424	4204	286	39	211	7180
1990	440	5250	480	40	260	8270
1997	520	5270	660	45	360	10990
2002	660	5320				12060
			Glucose-containing syrups			
1965	209	36	104	408	159	1155
1970	222	95	100	440	213	1380
1975	313	204	172	417	277	1890
1980	195	385	127	444	240	1840
1985	142	356	141	467	274	1940
1990	180	365	165	480	275	2125
1997	200	370	170	490	350	2955
2002	180					2780

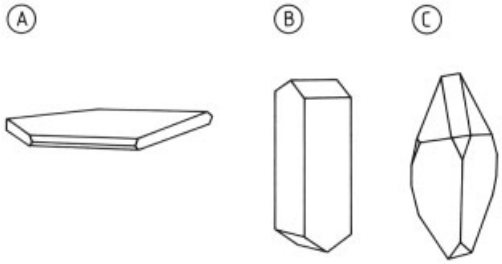

Figure 2. Crystal forms of glucose [20, 21]
A) α – Monohydrate; B) α – Anhydrous; C) β – Anhydrous

Conformation. In aqueous solution, D-glucose exists as a mixture of the α and β $^4C_1$ chair forms (approximately 62 % β and 38 % α when equilibrated at 31 °C [23]) that mutarotate from one anomeric form to the other through the open chain aldehyde form (see Fig. 3).

Chemical structure and conformation have been confirmed with the aid of high-field nuclear magnetic resonance spectrometry using superconducting magnets, advanced pulsars, and expanded computer capabilities to perform one and two-dimensional Fourier transform experiments [24].

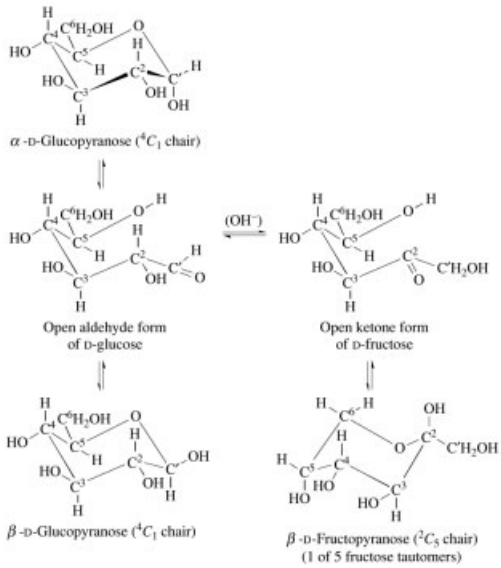

Figure 3. Mutarotation and alkali isomerization of glucose

Table 2. Physical properties of D-glucose

	α-D-Glucose	α-D-Glucose hydrate	β-D-Glucose
Formula	$C_6H_{12}O_6$	$C_6H_{12}O_6 \cdot H_2O$	$C_6H_{12}O_6$
Optical rotation $[\alpha]_D^{20}$	$112.2 \rightarrow 52.7°$*	$112.2 \rightarrow 52.7°$*	$18.7 \rightarrow 52.7°$*
Specific heat of solution (25 °C), kJ/kg	+ 59.4	+105.1	+ 26.0

* Rotation of an α/β equilibrium mixture = 52.7°.

2.2. Physical Properties

Some important physical properties of crystalline D-glucose are shown in Table 2. The large positive heat of solution of the crystalline monohydrate form imparts a unique cooling effect to the palate when incorporated into candy and icing in the crystalline state.

Solubility is an important property of glucose since it is crystallized commercially from aqueous solution and many of its commercial uses involve water systems. The solubility of glucose in water is shown in Figure 4 [25]. Below a temperature of ca. 100 °C, the stable, crystalline form of D-glucose is the α-form, which crystallizes as a monohydrate below ca. 50 °C; above ca. 100 °C, the β-anhydrous form is most stable.

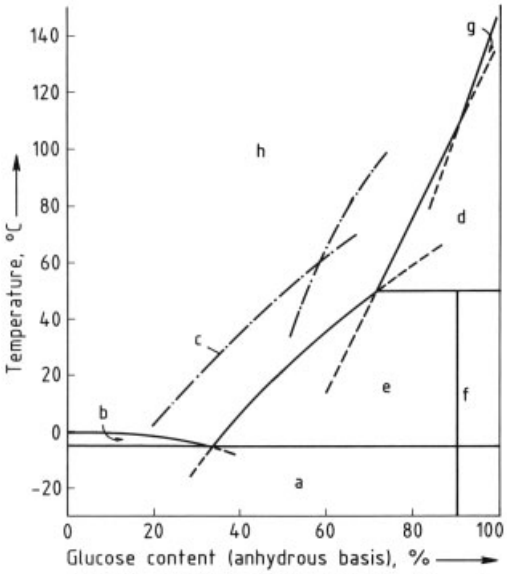

Figure 4. Phase diagram for glucose – water solutions
a) α-Glucose monohydrate and ice; b) Ice and solution; c) Solubility of glucose in water; before mutarotation to α-,β-equilibrium; d) Anhydrous α-glucose and solution (α-, β-equilibrium); e) α-Monohydrate and solution (α-,β-equilibrium); f) α-Glucose (monohydrate and anhydrous); g) β-Glucose and solution; h) Unsaturated solution

2.3. Chemical Properties

Above a pH of 6.3, the open aldehyde form of glucose rearranges to keto-D-fructose (Lobry de Bruyn–van Ekenstein rearrangement → Carbohydrates, see Fig. 3).

Glucose is a *reducing sugar*, i.e., it reacts with oxidizing agents such as cupric hydroxide. Cupric hydroxide oxidizes reducing sugars more rapidly if the copper is kept in solution by the formation of a complex salt with tartrate ion (Fehling's solution) or citrate ion (Benedict's solution) to form red cuprous oxide. Benedict's solution is the more stable and is not affected by uric acid; hence it is preferred for detection of glucose in urine.

The inherent reducing property of glucose inhibits oxidative degradation in food; glucose maintains the bright red color of tomato catsup, red berry preserves, and red meats without the need for artificial antioxidants.

Glucose is fermented rapidly and completely by yeast and other organisms. The degree of fermentation possible (or fermentable extract) of glucose and glucose-bearing syrups is shown in Table 3.

Table 3. Fermentable extract values of glucose and glucose-containing syrups *

Substrate	Fermentable extract, wt % db
36 DE	34.8 [27]
42 DE	41.5 [27]
54 DE	53.4 [27]
62 DE	70.0 [27]
64 DE	76.0 [28]
68 DE	79.0 [28]
Glucose – fructose syrup	95.0 [27]
Crystalline glucose	100.0 [27]

* DE = dextrose equivalents; db = dry basis.

3. Raw Materials

Starch (→ Starch) is, for all practical purposes, the only raw material used for commercial pro-

duction of glucose-containing syrups today, although efforts have been made to utilize the other natural glucose polymer, cellulose, as well (→ Ethanol, Chap. 5.4.3).

All starches are polymers consisting of α-linked anhydroglucopyranose units (Fig. 5 A). Two types must be distinguished: Amylose (Fig. 5 B) is an essentially linear molecule in which the anhydroglucopyranose units are linked almost exclusively via α-1,4-bonds. Amylopectin (see Fig. 5 C) is a much larger, highly branched molecule (the molecular mass of amylopectin is ca. 1000 times that of amylose [29]); α-1,4 linkages predominate but there are a significant number of α-1,6-linkages which lead to branching. The proportion and structure of these fractions in starch granules vary, depending on the botanical source of the starch (Table 4). In general, amylose from grain sources has a lower degree of polymerization (DP), ranging from 200 to 1800, the DP of amylose from root or tuber sources is higher (ranging from 1000 to 6200). The amylopectin fractions have a much greater DP, averaging about 2×10^6 with no major difference between starch sources.

Maize has become the leading source of starch for conversion to sweeteners [34], although starch from barley, sago, sorghum, wheat, rice, potato, tapioca, and other starch-bearing grains and roots serves as the raw material for the production of glucose-containing syrups around the world. About 70 % of the world production of maize starch is converted to glucose-containing sweeteners; nearly 100 % of the glucose-based sweeteners produced in the United States and 84 % of the production in Japan are from maize. In Japan it is required by law that one part potato starch is used for every 12 parts corn starch for fructose syrup production. Although the use of wheat starch (as a co-product of vital gluten production) is increasing in Europe, glucose-containing sweeteners produced there also originate from maize starch. In Australia, however, all glucose-bearing syrups are produced from wheat starch.

In Asia, factories often use sago, tapioca, or maize (corn) starch for glucose production, depending on the availability and price of each ingredient. The starch source and purity have a strong influence on manufacturing requirements, but a variety of starches can be used successfully.

In Thailand, for example, a syrup called "Chinese Maltose" is produced from wheat and glutinous rice. In Pakistan, broken rice is used to produce 42 % fructose and enzyme – enzyme glucose syrups having qualities comparable to products made from maize starch.

4. General Principles of Starch Hydrolysis

Hydrolysis of starch occurs when linkages between the anhydroglucopyranose units are broken catalytically (by using either acid or enzymatic catalysts in the presence of heat) with the addition of water. The formation of the monomer (D-glucose) and of dimers having a 1,4-(maltose) or 1,6-(isomaltose) configuration are shown in Figure 6.

Mutarotation, (see Section 2.1), isomerization reactions (see Section 2.3), and possible repolymerization reactions, coupled with the complex chemical structures of amylose and amylopectin (which offer a number of possible hydrolysis sites), provide the potential for the creation of molecules having a nearly infinite number of combinations of ketohexose units (→ Carbohydrates).

A simplified flow diagram illustrating the process steps used to manufacture sweeteners from starch is shown in Figure 7. Starch hydrolysis is catalyzed by both acids and enzymes. Irrespective of the catalyst used, all processes begin with a water slurry of starch granules. After hydrolysis, insoluble protein, lipid material, and some of the inorganic components present initially in the starch are separated by either centrifugation or filtration (or both). The presence of free starch due to incomplete hydrolysis (detected by blue or purple hue with iodine staining) leads to difficult separation and must be avoided.

Additional impurities are removed by refining with activated carbon, decolorizing resins, or ion-exchange resins, either singly or in combination, before evaporation to final product solids concentration. Because color-forming bodies are ionic in nature and are efficiently removed by resins, the total demineralization of syrups, with or without carbon treatment, provides a product with improved color removal and color stability [83]. And, the removal of ionic material, notably chloride ions, reduces the corrosives tendencies

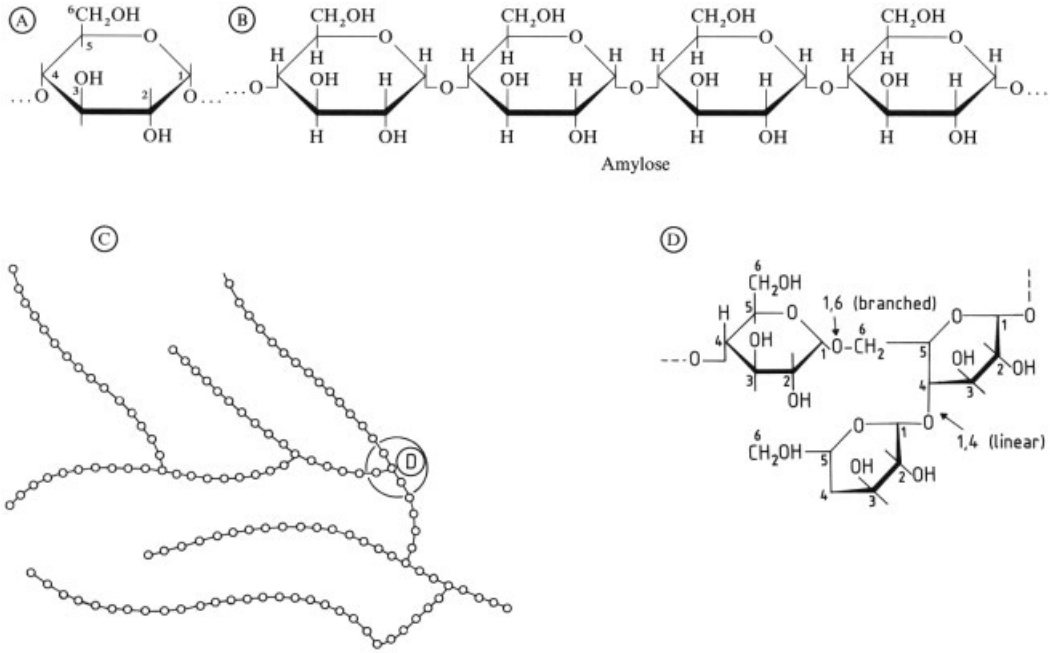

Figure 5. A) Anhydroglucopyranose unit. Shown as a 5-membered pyranose ring structure with each carbon identified by number
B) Linear molecular structure from Amylose; normally a molecule consists of 200 – 6200 units linked together with α-1,4 bonds
C) Amylopectin. Each "bead" represents an anhydroglucopyranose unit.
D) A branch point is detailed to show an α-1,6 bond

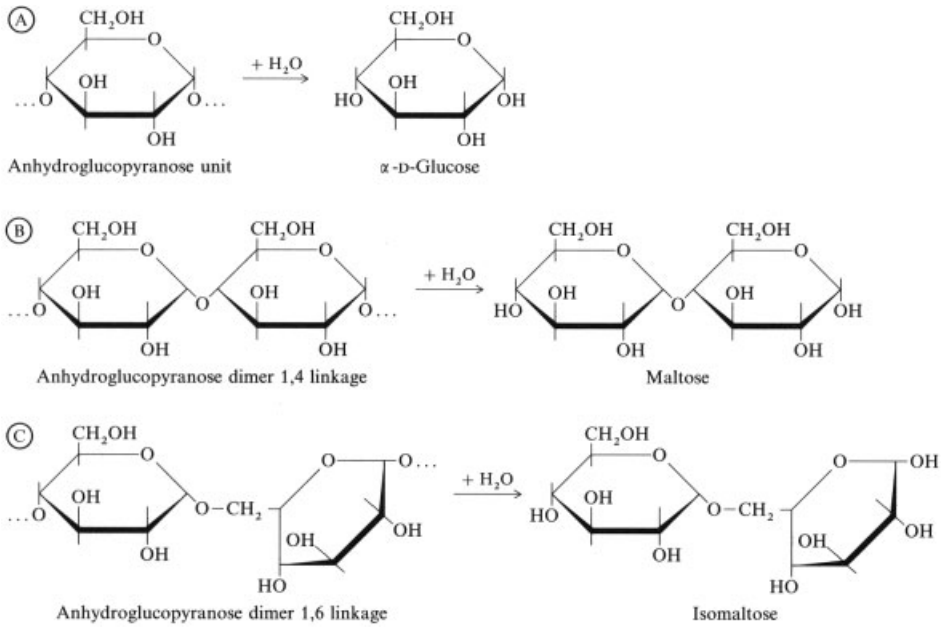

Figure 6. Hydrolysis of anhydroglucopyranose polymers

Table 4. Starch type, composition, and production [29 – 33]

Botanical starch source	Amylose content, wt % db *	Amylopectin content, wt % db *	World production, 10^6 t/a	Principal countries of production
Arrowroot	20	80		Brazil, China
Barley	19 – 23	77 – 81		Europe
Maize				
Normal	26 – 28	72 – 74	27.6	United States, Japan, Europe, former states of USSR
Waxy	0 – 2	98 – 100		United States
High-amylose	55 – 85	15 – 45		United States, Australia
Potato	20 – 24	76 – 80	2.7	The Netherlands, Japan, Poland, former states of USSR
Rice	14 – 26	74 – 86		Asia, South America
Sago	27	73	< 0.1	Indonesia, South America
Sorghum	22 – 27	73 – 78		Asia, United States
Tapioca	17 – 22	78 – 83	3.65	Thailand, Brazil
Wheat	19 – 27	73 – 81	2.92	Australia, Europe, United States

* db = dry basis

of syrups allowing the use of less expensive materials of construction in process equipment located after ion exchange.

Glucose-containing syrups are generally categorized by their degree of conversion as measured by the amount of reducing sugars present. Although it is becoming more common for customers to specify a range of saccharide distribution to obtain desired functionality, glucose syrups can be divided into groups on the basis of their *dextrose equivalent* (DE) as shown in Table 5.

Table 5. Classification of glucose-containing syrups according to dextrose equivalents (DE)

Degree of hydrolysis	Accepted terminology
< 20 DE	maltodextrins
20 – 38 DE	type I syrup
38 – 58 DE	type II syrup
58 – 73 DE	type III syrup
≥ 73 DE	type IV syrup

Most glucose-bearing syrups are sold in liquid form usually containing between 70 and 85 % dry substance, but they can also be dried to form granular, semicrystalline, or amorphous products. These "solid syrups" range in degree of conversion from low-DE maltodextrins to highly converted (> 96 % DE) "total sugars" [84].

4.1. Production and Properties of Acid Glucose Syrups

In the acid conversion process, a starch slurry usually having a dry substance content between 30 and 40 wt % is acidified to pH 2 or lower and subjected to a combination of time, temperature, and pressure to effect hydrolysis. Higher temperature, permitted by operation under pressure, reduces the reaction time considerably. KIRCHOFF's conversion at atmospheric pressure, for example, lasted 36 h, early starch converters that were limited to pressures of 35 kPa took 6 h to complete the reaction. At 415 – 620 kPa and up to 160 °C, modern converters accomplish the conversion within minutes. Acid conversion can be accomplished by using either batch or continuous equipment although continuous conversion is used almost universally today.

When the desired degree of conversion is achieved, temperature and pressure are reduced and the reaction is stopped by addition of a neutralizing agent (usually sodium carbonate), raising the pH to 4 – 5.5. Insoluble compounds are then removed, and the syrup is refined and evaporated to the desired final concentration.

While acid conversion is a random-action mechanism, it results in remarkably reproducible saccharide composition for any given degree of hydrolysis. Typical compositions of acid glucose syrups are shown in Table 6. This reproducible composition is actually a limitation of the acid process because the manufacturer

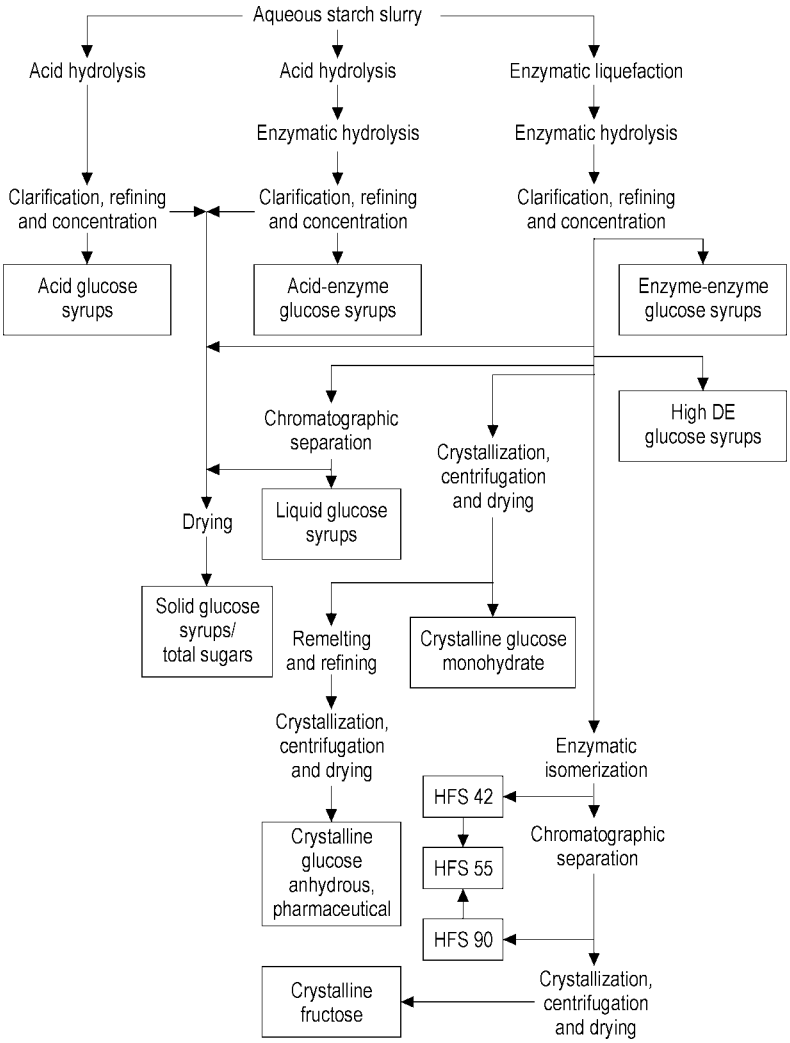

Figure 7. Simplified process flow for the production of sweeteners from starch
* HFS = High-fructose syrup

Table 6. Saccharide composition of acid-hydrolyzed glucose syrups

Degree of polymerization	Dextrose equivalent (DE) of glucose syrup			
	27 – 30	34 – 36	42 – 43	55
DP 1 (glucose)	9 – 10	13 – 14	19 – 20	31
DP 2 (maltose)	8 – 9	11 – 12	14	18
DP 3 (maltotriose)	9 – 10	10	12	12 – 13
DP 4 (maltotetrose)	7 – 8	8 – 9	9 – 10	10
DP 5 (maltopentose)	7	8	8	7
DP 6 (maltohexose)	6	6 – 7	7	5
DP 7 (higher sugars)	50 – 54	40 – 42	29 – 31	16 – 17

cannot influence the saccharide distribution for a given degree of conversion, which is important in some applications of glucose syrups.

Additionally, acid syrups having a DE value below ca. 30 will tend to cloud upon standing due to precipitation (retrogradation) of the longest chain linear polymers. Also, although acid hydrolysis to DE values in the high 80s is possible, the extreme conditions required to produce DE values greater than ca. 55 pro-

mote the formation of excessive amounts of glucose degradation products (e.g., 5-hydroxymethylfurfural see Section 2.3) which are difficult to remove during refining. These color precursors cause the product to yellow eventually. The most common acid glucose product has a DE value of about 40 – 45 % because of these upper and lower conversion limitations.

The limitations at both the low and the high ends of the conversion spectrum, coupled with the inflexible saccharide distribution resulting from acid hydrolysis, led manufacturers to investigate enzymatic hydrolysis as a means of making improved glucose-containing syrups.

4.2. Production and Properties of Acid – Enzyme Glucose Syrups

Coupling a secondary enzymatic hydrolysis to a preliminary acid cleavage allows expansion of the range of DE values both up and down. It also permits control over the saccharide distribution by enzyme selection.

After acid hydrolysis, the temperature is lowered and the pH is raised to levels compatible with the enzyme to be used; a pH between 4 and 6, and a temperature in the range of 60 – 70 °C are usually employed. Additionally, the degree of acid hydrolysis is usually (but not always) lower when making acid-enzyme syrups (see also → Enzymes, Chap. 5.1.3.2).

If the goal is a high glucose content, control of the acid hydrolysis step to produce a DE of 10 – 20 is important. Higher DE levels result in reduced glucose yield during subsequent enzymatic hydrolysis due to the presence of acid-polymerization products that resist enzymatic conversion; conversely, starch retrogradation can occur with a DE less than 10, resulting in problems during clarification.

If maximum glucose content is not the aim, a higher DE value is usually chosen and the syrup can be clarified immediately after acid hydrolysis or the insoluble substances can be allowed to remain during enzymatic conversion.

If the insoluble compounds are removed, the acid hydrolysate can be enzymatically converted continuously, by using immobilized enzymes; otherwise, a batch treatment (with single use of the enzymes) must be employed.

When the desired level of enzymatic conversion is achieved by *batch conversion*, residual enzymes are deactivated with heat or pH adjustment, or are removed by refining. With *immobilized enzyme treatment*, the degree of conversion is controlled by the flow rate through a bed of insoluble enzymes; because the enzymes remain in the converter, their removal from the syrup is unnecessary. Refining and evaporation follow as in acid glucose syrup manufacture.

Bacterial *α-amylases* (E.C. 3.2.1.1) [*9000-90-2*] cleave α-1,4-linkages in a random fashion whereas *β-amylases* (E.C. 3.2.1.2) [*9000-91-3*] (and to a lesser degree, fungal α-amylases) mainly produce maltose, again only cleaving α-1,4-linkages. *Glucoamylase* (E.C. 3.2.1.3) [*9032-08-0*], on the other hand, splits off one glucose molecule at a time from the non-reducing end of the polymer by hydrolyzing α-1,4-linkages; glucoamylase also has the capability of splitting α-1,6 linkages but at a much slower rate. *Pullulanase* (E.C. 3.2.1.41) [*9075-68-7*] can split α-1,6-linkages, and addition of this enzyme in combination with another enzyme(s) increases the capability of tailoring saccharide distributions [30].

At first, enzyme conversion was used to extend the degree of conversion upward, primarily to maximize glucose production for the manufacture of crystalline glucose as well as for the production of sweeter, less viscous glucose-containing syrups. However, it also enabled selective production of either more maltose or more glucose, depending upon the enzyme chosen. This is illustrated by the compositional data shown in Table 7.

4.3. Production and Properties of Enzyme – Enzyme Glucose Syrups

The development of α-amylase from bacterial sources capable of withstanding the high temperature needed to gelatinize starch completely allowed the development of totally enzymatic processes that further expanded control of saccharide range and distribution. A total enzymatic conversion process eliminates the need for preliminary acid hydrolysis.

Thermostable α-amylases are added to a starch slurry (containing calcium as an enzyme stabilizer) that has been adjusted to a pH of

Table 7. Saccharide composition of various acid – enzyme glucose syrups

	High maltose			High glucose	
DE	DE 42 – 43	DE 48 – 50	DE 60 – 63	DE 70	DE 95
DP1 (glucose)	6 – 8 %	9 – 10 %	31 – 36 %	43 – 47 %	92 %
DP2 (maltose)	40 – 45 %	50 – 53 %	31 – 36 %	27 – 30 %	4 %
DP3 (maltotriose)	12 – 15 %	14 – 15 %	10 – 13 %	5 – 7 %	1 %
DP4 (maltotetrose)	3 – 7 %	1 – 2 %	(DP4+ = 20 – 23 %)	4 – 5 %	1 %
DP5 (maltopentose)	2 %	2 %		3 – 4 %	(DP5+ = 2 %)
DP6 (maltohexose)	2 %	2 %		2 – 3 %	
DP7+ (higher polymers)	26 – 30 %	18 – 20 %		9 – 10 %	

DP = Degree of polymerization.

6 – 6.5. A pH lower than 6.3 is preferable to prevent the formation of maltulose (4–*O*-α-D-glucopyranosyl-D-fructose) [35]. Maltulose cannot be converted enzymatically to glucose and its formation reduces glucose yield.

The recent introduction of a novel α-amylase capable of liquefying starch at pH 4.5 without the need for calcium ion addition has the potential to reduce chemical and refining costs as well as possibly eliminate the formation of maltulose. This enzyme's action pattern reportedly can increase the production of saccharides with low DP (1 – 4) while practically eliminating the higher saccharides (DP 8+). This characteristic allows higher solids levels during liquefaction [36]. While presently employed for the manufacture of ethanol from starch and whole grains, the application in the manufacture of starch sugars holds interest.

The slurry is heated, usually by direct steam injection, to ca. 103 – 107 °C and held for 5 – 10 min to liquefy the starch completely. A subsequent holding period of 1 – 2 h at about 95 °C allows continued enzymatic activity to raise the DE to a value between 10 and 15. The temperature and pH are then adjusted to levels compatible with the enzyme used for the secondary conversion. From this point, the process is similar to the acid – enzyme process.

On the high end of the glucose conversion scale, a totally enzymatic conversion greatly increases the yield of glucose while decreasing the refining requirements. The composition of highly converted syrups produced by acid, acid – enzyme, and totally enzymatic hydrolysis with respect to glucose concentration and impurity content is shown in Table 8.

Maximum glucose content during glucoamylase saccharification is also determined by solids content in the slurry; at 30 wt %, a glucose level of 96 wt % db is usually achieved. Decreasing the solids content to 10 – 12 wt % allows glucose levels of 98 – 99 wt % db to be reached. This higher glucose yield is produced at the expense of increased evaporation cost and risk of microbial contamination, and is seldom used commercially because chromatographic separation technology has become available and allows the increase of the glucose content of syrups to be $>$ 99 wt % db. Nanofiltration has also been used to enrich glucose content [37].

Total enzymatic conversion also affects the characteristics (via saccharide distribution) of less highly converted glucose-bearing syrups, as shown in Table 9. Syrup viscosity is an important property that can be affected by shifting saccharide distribution.

The viscosity is increased mainly by the longer chain polysaccharides. Thus, at a constant DE value, solids content, and temperature, an acid-hydrolyzed syrup would have the highest viscosity, an acid – enzyme syrup would be intermediate, and an enzyme – enzyme syrup would have the lowest viscosity. Viscosity affects the handling and storage temperature as well as the application properties. Many desirable results occur from lowering the viscosity, including improved flavor release in sweets, lowering of boiling temperature in candy manufacture, faster operation of candy production lines, and better "set" in jellies and jams.

Enzymatic conversion also allows the production of syrups with reduced glucose content. In the presence of protein, glucose undergoes non-enzymatic browning due to the Maillard reaction more readily than polymers of glucose. Syrups with higher concentrations of glucose polymers are more color stable.

Table 8. Effect of conversion method on composition of filtered unrefined highly converted glucose-containing syrup

	Conversion method		
	Acid	Acid – enzyme	Enzyme – enzyme
Glucose content, %	86	93	97
Ash content, %	1.6	0.4	0.1
HMF content, ppm *	3000	80	30
Color (relative)	10	0.3	0.2

* HMF = 5-hydroxymethylfurfural.

Table 9. Saccharide distribution of 42 DE syrups produced by different conversion methods

Degree of	Conversion method		
polymerization	Acid	Acid – enzyme	Enzyme – enzyme
DP 1	19	6	2.5
DP 2	14	45	56.0
DP 3	12	12	16.0
DP 4	10	3	0.7
DP 5	8	2	0.4
DP 6	7	2	0.7
DP 7+	30	30	23.7

5. Production of Solid Glucose

Whereas most glucose is sold as glucose-containing syrups, a portion is sold as granular free-flowing solids. In this form it can be incorporated into dry mix products and shipped without the added cost of transporting water. Bagged product accommodates smaller-scale users. Solid products include total sugars and pure crystalline forms of glucose (dextrose). Total sugars are solidified hydrolysates that can either be in glassy (amorphous) or crystalline states. Pure glucose crystals are grown from aqueous solution to form a slurry of pure crystals in mother liquor (massecuite), the crystals are separated from the mother liquor by centrifugation and dried in hot air dryers.

5.1. Production of Total Sugars

When the entire starch hydrolysate is solidified, it is called a total sugar. A variety of processes have been developed to produce total sugars [84]. Dry maltodextrins have a DE of less than 20 while corn syrup solids have higher DE values, commonly between 24 and 48. Highly converted total sugars usually contain between 95 and 99 % db glucose although processes have been patented for the production of fructose-bearing total sugars and solid starch sugars that incorporate other ingredients such as intense sweeteners.

About 40 % of the crystalline glucose sold in Japan is a solidified hydrolysate (*total sugar* form) of monohydrate glucose distantly related to the cast product produced early this century (but of much higher purity due to enzymatic starch hydrolysis, modern refining methods, and evaporation under high vacuum).

More than half of the crystalline glucose sold in Korea is also a monohydrate total sugar that is used as a carrier in many dry foods. These total sugars are produced by using a modern, continuous solidification process that employs spray drying of a partially crystallized, refined hydrolysate [38, 39].

However, with the exception of a relatively small amount of spray-dried hydrolysate products (a monohydrate product with the trade name Cantab produced by Penford Products Products Co., United States and an anhydrous product manufactured by Cargill, United Kingdom/United States), total sugars have not been the large market successes in the West that they have in the Orient; in Europe and the Americas, pure glucose crystals (in either the monohydrate or the anhydrous form) are the more universally accepted product.

5.2. Production of Pure Crystalline Products

Initially, crystalline glucose was produced from acid-converted starch hydrolysates. Because of degradation products formed during conversion,

the maximum glucose content achievable in the hydrolysate was limited to about 88 % db. By 1938, enzymatic conversion processes were being developed that provided much purer sources from which crystalline glucose could be manufactured.

Either continuous or batch crystallization can be used to produce the desired crystals. In *continuous crystallization*, supply liquor is either mixed with recycled seed crystals or new crystals are formed by spontaneous nucleation. In *batch crystallization*, a portion (usually 15 – 30 %) of the finished batch is retained in the crystallizer and mixed with the supply liquor.

Supersaturation to promote crystal growth can be achieved by either cooling or evaporation [40, 41]. When continuous crystallization is employed for the production of monohydrate dextrose crystals that form below 50 °C, great care must be taken to prevent microorganisms, primarily yeast, from acting on the glucose solution. Yeast fermentation is only a problem at the lower temperature and solids levels, that occur at high crystal content at the final crystallization conditions.

When producing glucose monohydrate, batch crystallization has the advantage that the operating conditions at the beginning of the cycle (highest temperatures and solids content) are hostile to yeast. Thus, when hot supply liquor is mixed with a portion of the previous batch, any yeasts that might be present are killed.

Because a conventional batch crystallizer process retains a portion of the finished product to be mixed with supply liquor for the next batch, a more efficient process for the production of glucose monohydrate has been proposed [42]. By continuously producing a lean phase massecuite in a continuous crystallizer, that is operated at temperature and solids level hostile to yeast, and then supplying this material to a batch crystallizer to produce a high-crystal-content massecuite suitable for centrifuging, the full operating volume of the batch crystallizer is used for production; the batch crystallizer is completely emptied at the end of the crystallization cycle, and the processing time devoted to seed mixing is eliminated.

Since anhydrous glucose crystals form above 50 °C, yeast fermentation is not an issue and either batch or continuous crystallization can be used.

Production of Glucose Monohydrate Crystals. The α-D-glucose monohydrate crystals, usually produced by batch crystallization, have the largest share of the world market. Starting material for the manufacture of glucose crystals is a syrup with a glucose content of 95 – 96 wt % db produced by enzymatic hydrolysis of starch to the highest practical degree. This syrup is produced and refined by using the process described previously for the production of enzyme – enzyme glucose syrups.

In a typical batch crystallization process, the syrup is evaporated to a solids content of 74 – 79 wt %. Higher solids content results in increased crystallizer productivity [43]. The supply is cooled to about 46 – 50 °C, and mixed with seed crystals (left behind from the previous batch) in crystallizers (see Fig. 8). As mentioned previously, an alternative process having greater productivity would utilize continuous crystallization to produce a lean massecuite that would be supplied to the batch crystallizer, thus allowing complete emptying of the batch unit at the end of the cycle. Alternatively, continuous crystallization could be used entirely.

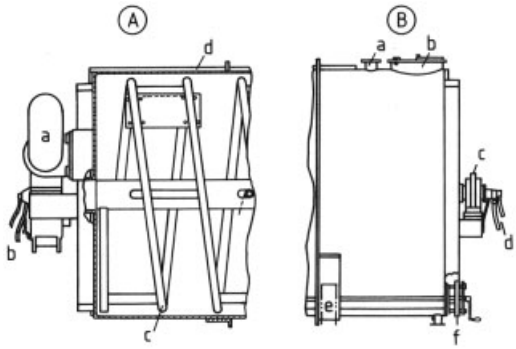

Figure 8. Batch monohydrate glucose crystallizer
A) Cutaway view showing internals a) Motor and Speed reducer which drives coil; b) Cooling water connections; c) Rotating coil – agitator; d) External vessel shell
B) External side view a) Supply liquor inlet; b) Inspection hatch; c) Bearing; d) Cooling water connections; e) Support saddle; f) Massecuite discharge valve

The resulting mixture is cooled slowly, usually by indirect heat transfer to a final temperature between 20 and 40 °C over 2 – 5 d during which about 60 % of the glucose is crystallized in the α-monohydrate form, yielding a viscous

magma of crystals and mother liquor. Exact values of the solids concentration, the initial and final temperatures and the rate of cooling depend on the glucose content of the crystallizer supply [44 – 46].

The resulting magma is routed to perforated screen centrifuges (usually batch machines) to separate most of the mother liquor from the crystals; most of the remaining mother liquor is then removed by washing the crystals with a spray of water.

The crystals are spun as dry as possible (about 14 wt % moisture content) and then transferred to continuous rotary dryers (→ Drying of Solid Materials) where they are dried in a stream of warm, filtered air to 8.5 – 9.0 wt % moisture content, slightly below the theoretical moisture content of the monohydrate (9.08 wt %).

Because crystallization of glucose monohydrate occurs satisfactorily despite the presence of large quantities of impurities [4, 47], the mother liquor can be reconcentrated and a second crop of crystals can be recovered. The presence of impurities alters the crystal size and shape; purer liquors produce longer, narrower crystals than those produced from less pure solutions [48].

Otherwise, a portion of the mother liquor can be recycled and combined with the incoming hydrolysate as supply to a single crystallization step. Either process increases the yield of glucose crystals from a given amount of hydrolysate. Recovery of as much as 87.5 % of the glucose from the incoming hydrolysate as monohydrate crystals has been reported [49]. By reconverting the mother liquor, yields approaching 100 % have been reported [85].

Production of Anhydrous Crystalline Glucose. Anhydrous crystalline glucose is produced in evaporative crystallizers operated in either a batch or a continuous mode. *Anhydrous* α-D-*glucose* is usually produced in batch vacuum pan evaporative crystallizers operated at 60 – 65 °C. The crystallization cycles are much shorter than those needed for monohydrate production, usually only 6 – 8 h [4]. The crystals are separated from the mother liquor, washed, and dried in a manner similar to monohydrate crystals, but the final product moisture is less than 0.1 wt %.

Production of anhydrous glucose crystals is more demanding than that of their monohydrate relatives and they can only be successfully crystallized from relatively pure solutions [4]. The purest grades (U.S.P.) are produced from very pure glucose syrups made by dissolving monohydrate glucose crystals in water.

Anhydrous β-D-*glucose* can be produced by operating the crystallization step slightly above 100 °C [4]. Another commercial method of preparing a crystalline glucose total sugar rich in the β-anomer consists of spraying a hot, concentrated glucose syrup onto a moving bed of previously made product at high temperatures (85 – 105 °C) in a rotary dryer to yield a granular, pearl-shaped crystal [50].

The β-anhydrous glucose is much more soluble than the α-form (72 wt % vs. 30 wt % at 25 °C, the solubility of the equilibrium α-β-mixture is 51 wt % at 25 °C).

6. Glucose – Fructose Syrups

6.1. Properties

Isomerization of glucose to fructose significantly increases the sweetness of starch-based sweeteners. This single property propelled the market share of starch sweeteners from 20 to over 50 % in the United States. More than half of the starch-based sweeteners sold in Japan and about two-thirds of those sold in the United States are 42 and 55 % fructose syrups.

Sweetening Power. Sweetening power cannot be measured quantitatively in absolute terms but can be determined subjectively by using trained taste panels. Many factors including stereochemical configuration influence the perception of sweetness: α-glucose is definitely sweeter than the β-form [51]. Stereochemistry can explain the wide variation in reported sweetness of fructose, whose crystalline form β-D-fructopyranose undergoes rapid tautomerization in solution to five forms: α and β-D-fructopyranose as well as α and β-D-fructofuranose and open chained keto-D-fructose. Thus, although crystalline fructose is reportedly 1.8 times as sweet as sucrose, solutions of fructose are much less sweet than crystalline fructose because of the tautomeric distribution in water [52].

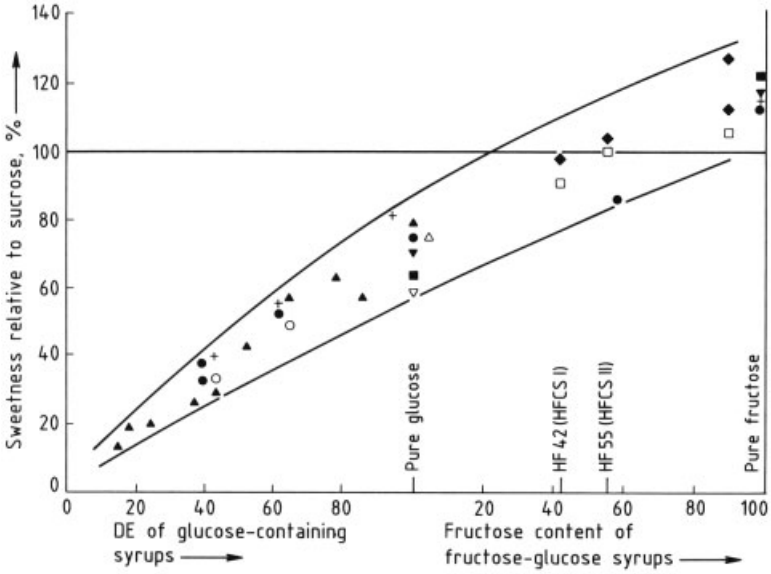

Figure 9. The relative sweetness of glucose-containing and fructose – glucose syrups vs. composition
All data at about 10 wt % solids water solutions, except [54] □ [55]; ● [56]; ○ [51]; ▲ [54]; ▼ [53]; ■ [57]; + [58]; △ [59]; ▽ [60]; ♦ [61];

Concentration [51], temperature, and pH [53] affect the perception of sweetness too. To eliminate the possibility of enzymatic hydrolysis in the taster's mouth, rapid and consistent evaluation procedures are required, especially when tasting low-DE syrups.

In general terms, however, as the glucose content of glucose-containing syrups increases, sweetness increases; as glucose is isomerized to fructose, sweetness increases further. The sweetness of glucose-containing syrups, relative to sucrose, in solutions of about 10 wt % solids concentration is shown in Figure 9.

Most glucose–fructose syrups are usually produced from enzymatically converted high glucose (95 – 96 % db) hydrolysates which are the fructose–glucose syrups shown in Figure 9. However, lower DE syrups that have been isomerized slightly to 10 wt % fructose content to increase sweetness are sold in the European Union. Syrups isomerized to this low level do not violate fructose production quotas that have been established to protect domestic sucrose production.

Chemical Properties. Fructose reacts with amino groups (Maillard reaction) much more readily than glucose does (→ Fructose) [62]. Therefore, glucose – fructose syrups must be refined to a greater degree and are inherently less stable. Fortunately the higher water solubility of fructose (→ Fructose) allows lower storage temperatures without crystallization (see Chap. 9).

6.2. Production

The clarified, refined, and ion-exchanged high-glucose-content syrup is evaporated to a 40 – 50 wt % solids level. Magnesium(II) ions (usually as $MgSO_4$) are added as a cofactor to activate and stabilize the isomerase enzyme, as well as to overcome the inhibitory effect of residual calcium not removed by ion exchange.

This feedstock is then passed through reactors containing immobilized glucose isomerase (E.C. 5.3.1.18). Process conditions vary depending upon the supplier of the enzyme but most systems operate at 55 – 65 °C and with pH of 7.5 – 8.5. The flow rate is adjusted to yield a product having a minimum of 42 % of the dry substance converted to fructose.

Enzymatic activity decreases during the effective life of the reactors due to thermal inactivation and the presence of trace impurities in the feed liquor. Fructose content is maintained at the desired 42 % level by reducing the flow rate through the columns. The effective life of

a reactor is several months; during this period, the nominal retention time in the reactor may increase typically from 0.5 to as much as 4 h.

Immediately after isomerization, the pH is lowered to 4 – 5, the syrup is refined and evaporated to the final concentration.

Chromatographic Enrichment. Although 42 % db fructose syrups are nearly as sweet as sucrose, the fructose content had to be increased to 55 % db to provide the exact sweetness match demanded by the soft-drink market. This is accomplished by employing chromatographic separation techniques to separate 42 % fructose syrups into fructose-enriched and glucose-enriched streams.

Separation is accomplished by passing the 42 % fructose syrup through an adsorbent resin containing calcium ions. Fructose is bound to the adsorbent to a greater extent than glucose (and polymers of glucose). Elution with water produces a discharge stream initially rich in the loosely bound glucose and glucose polymer fraction, followed by a fraction rich in the more tightly bound fructose. A continuous (simulated moving bed) process is used commercially, in which 42 % fructose supply liquor and elution water enter the columns at different points, while glucose-rich and fructose-rich product streams are withdrawn at others. The points of addition and withdrawal are moved along the bed periodically to correspond to changes in composition within the column [63, 64].

Typical operation uses 50 wt % solids feedstock (42 % fructose). The separation produces a glucose- and glucose-polymer-rich byproduct stream that is recycled either to the isomerization step (to convert additional glucose to fructose) or back to the hydrolysis step (to convert the glucose polymers to glucose). A fructose-rich stream containing 80 – 95 % db fructose is the primary product (the composition depends on the equipment and operating conditions). This fructose-rich stream is blended with 42 % fructose syrup to produce a 55 % fructose syrup that is refined and evaporated to the final concentration.

The fructose-rich stream can also be refined and evaporated to about 80 wt % solids for customers requiring a very high fructose-content product. At the highest fructose levels, it can supply a crystallization process to produce fructose crystals.

In 1987 more than 80 commercial factories on almost every continent were reported to produce more than 6.6×10^6 t (dry basis) per year of glucose–fructose syrups [65]. Worldwide production increased to 10.1×10^6 t (dry basis) per year in 1995 and is estimated to have grown to nearly 12.5×10^6 t (dry basis) in 2002 before dropping to 11.8×10^6 t (dry basis) per year in 2005 [14]. The phenomenal growth of glucose – fructose syrup (called high fructose corn syrup or HFCS in the United States) production is shown in Figure 10.

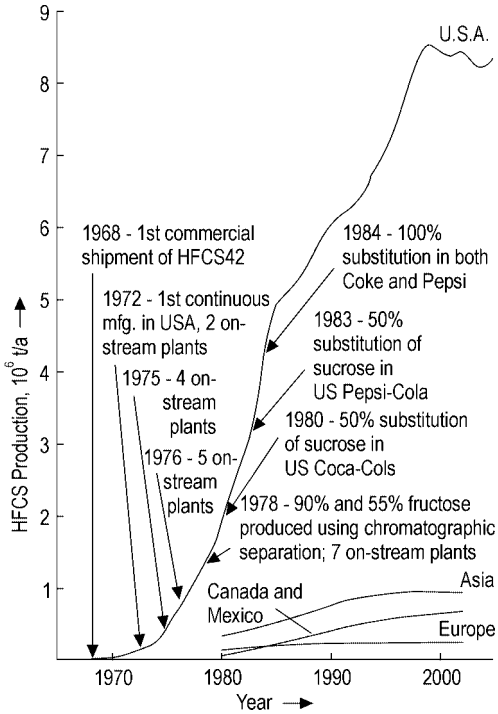

Figure 10. Growth of glucose – fructose syrup production, 1968 – 2005

6.3. High-Fructose Syrups from Inulin

Inulin is a naturally occurring polymer of fructose which replaces starch as the food reserve in many plants of the family Compositae (e.g., Jerusalem artichoke, chicory, salsify, cardoon elecampane, burdock, dahlia, and ti plants).

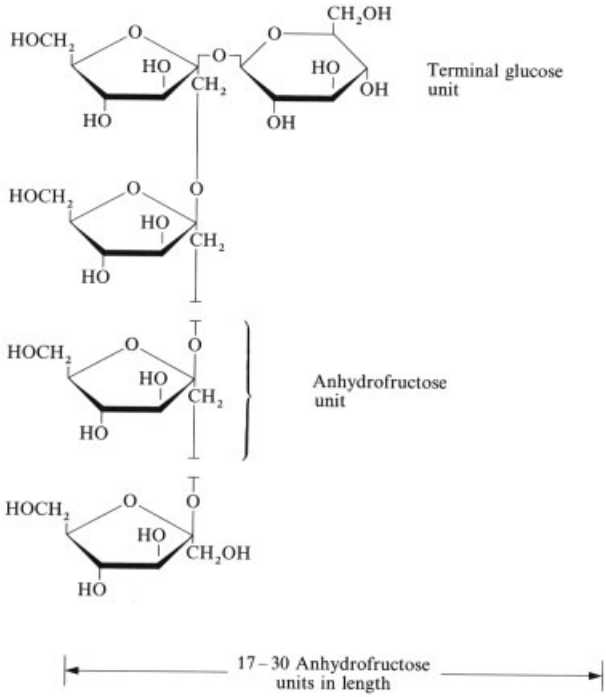

Figure 11. Part of an inulin chain

The mature inulin molecule has an M_r of 3000 – 5000 and consists of 17 – 30 anhydrofructose units and a terminal glucose unit (see Fig. 11). The chain length varies during the growing season, and consequently, the glucose – fructose ratio is not constant [66].

A process for the production of high-purity fructose syrup from inulin-containing roots has been described [67]. The pH of extracts is raised to > 11 using lime, to precipitate impurities. Under these conditions, inulin is not precipitated by high levels of calcium (as is fructose) and can be recovered by filtration [68].

The lack of commercial interest in high-fructose syrup production from inulin is probably due to the high quality of fructose syrups produced by isomerization of glucose followed by chromatographic separation, and to lack of the large crop production industry needed to produce inulin. While the Jerusalem artichoke is reported to produce tuber yields as high as 100t/ha [86], present commercial sources of inulin are primarily obtained from chicory [87]. While some fructose is being produced, much of the inulin is partially enzymatically hydrolyzed to make a range of nondigestible oligofructose ingredients that are finding use as fat mimetics. Only the European Union is a substantial producer of fructose-containing inulin syrup that originally was not limited by production controls as was glucose-derived fructose syrups. Inulin syrups have now been brought under control by the EU.

7. Analysis

Solids concentration of glucose-containing syrups is usually measured by refractive index or density. This measurement can be related to dry substance or the more traditional Baumé or specific gravity values by using established mathematical models and conversion factors that compensate for syrup composition. Glucose-containing syrups are sold at concentrations usually varying between 75 and 85 wt % solids. Glucose–fructose syrups are usually produced at 71 – 75 wt % solids for the 42 % fructose syrup and 75 – 77 wt % solids for the 55 % fructose product. The moisture content of dry products (crystalline glucose and solid glucose

syrups) is usually measured by weight loss under vacuum drying or by Karl Fischer titration.

The *carbohydrate composition* of glucose and glucose–fructose syrups is almost always measured by high performance liquid chromatography (HPLC). The carbohydrate composition can be related to the traditional dextrose equivalent (DE) by calculation [69]. However, direct measurement is still practiced by using copper reduction (the Lane and Eynon method) [70].

The *color of syrups* is generally measured by the amount of light transmitted through (or absorbed by) a fixed depth of syrup. The wavelength of the light beam is chosen to be most sensitive to soluble impurities such as ash. Most syrups are referred to as "water white"; their color content is so low that it is difficult to distinguish (visually) between syrups and water.

8. Legal Aspects

The legal aspects of glucose-containing products are governed by a number of regulations defined, in part, by such references as the U.S. Code of Federal Regulations, the Codex Alimentarius Commission, and the U.S. and British Pharmacopeia. In the United States, glucose syrups and monohydrate and anhydrous crystalline glucose are considered standardized foods by the U.S. Food and Drug Administration, while fructose syrups derived from glucose are considered as GRAS (generally regarded as safe) with specifications defined by the Food Chemicals Codex of the National Academy of Sciences.

Solid Concentration. The Codex Alimentarius Commission recommends that the solids content of *glucose-containing syrups* be no less than 70 wt %. The U.S. Code of Federal Regulations (CFR) specifies that *anhydrous crystalline glucose* contain no less than 98 wt % dry substance (in practice, the commercial product usually contains < 0.1 wt % water) and that *monohydrate crystalline glucose* contain no less than 90 wt % dry substance (in commercial practice, dry substance ranges between 91 and 92 wt %).

Carbohydrate Composition. The Codex Alimentarius recommends that glucose syrups have a DE value not lower than 20 (below this level they are defined as maltodextrins). The CFR requires that *crystalline glucose* have a DE value no less than 99.5 [66].

Impurity Content. Ash content limits of 1 wt % for *syrups* (Codex Alimentarius) and 0.25 wt % for *food-grade crystalline glucose* (CFR) are met easily by modern ion-exchange refining. The U.S. Pharmacopeia places much stricter specifications (including specific rotation, acidity, residue after ignition, and content of chloride, sulfate, arsenic or other heavy metals, and other components) on pharmaceutical-grade glucose [72].

9. Storage and Transportation

Storage temperatures of glucose-containing syrups are a compromise between low temperatures to minimize color development with time (see Fig. 12) and high temperatures either to reduce the viscosity of low-DE syrups below 15.0 Pa s (to allow handling with lobe or vane pumps) or to prevent crystallization of glucose in higher DE syrups. Suggested temperatures for various syrups are shown in Table 10.

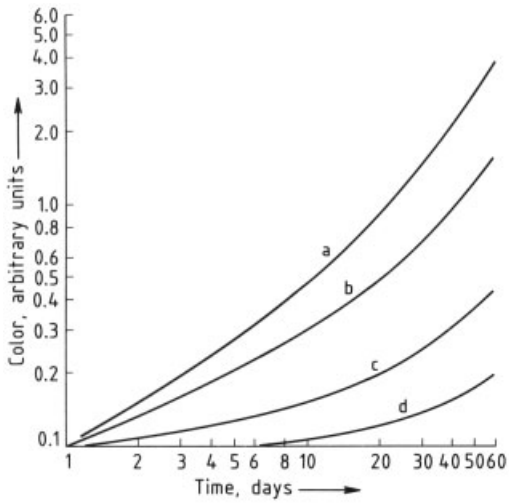

Figure 12. Glucose-containing syrup color as a function of time and temperature
a) 54 °C; b) 49 °C; c) 38 °C; d) 27 °C

Shipping containers for syrups range in size from 200 L drums to jumbo railcars holding 90 t.

Table 10. Storage conditions of glucose-containing syrups

Syrup type	Solids content, wt %	Temperature, °C	Limiting condition
42 DE	80	38	viscosity
	83	49	viscosity
60 DE	82	32	viscosity
	84	41	viscosity
95 DE	71	49	crystallization
42 HFCS	71	32	crystallization
55 HFCS	77	27	crystallization
90 HFCS	80	16	crystallization
Liquid glucose (99.5 DE)	71	52	crystallization

Trucks carrying 25 t are also used. Alternatives to drums include 1000 L disposable containers consisting of reinforced corrugated board outer containers holding a heavy-duty plastic bag.

Bags for storage of *crystalline glucose* and *solid syrups* must be protected against atmospheric moisture pickup by incorporation of a vapor barrier in the bag construction. Storage temperatures of solid products should be lower than 40 °C and rapid temperature changes should be avoided to minimize product caking.

Crystalline glucose is shipped in vapor-resistant bags usually containing 50 or 100 pounds (or the approximate metric equivalent sizes of 20, 25, and 40 kg) and in bulk trucks and railcars (jumbo railcars can hold as much as 82 t). Reusable containers holding about 1 t of crystalline material are used for customers who do not require bulk delivery and who do not want the bother and expense of bag opening and disposal.

10. Uses

Glucose-containing syrups and glucose are used in a variety of food and nonfood applications. The manufacturing flexibility provided by enzymatic conversion techniques, coupled with modern refining and concentrating technology, provides a wide range of properties useful to the food formulator. Table 1 illustrates the relative volumes of usage in major food industries in the United States. The impact of the introduction of glucose–fructose syrups on other glucose applications can also be seen from this table. The rapid and complete fermentation of glucose is a valuable property in the fermentation and baking industries. In countries where its use is permitted by law, glucose is used by the brewing industry. Glucose also serves as a raw material for the manufacture of food-grade acids (citric, lactic, acetic, etc.), enzymes, amino acids, vitamin C, and a variety of antibiotics.

The largest nonfood use of unrefined glucose-containing syrups is as a carbohydrate source for the production of fuel-grade ethanol (→ Ethanol, Chap. 5.1.2). In 1986, the annual production capability in the United States was reported to be ca. 2.7×10^9 L [73] and a 1988 article indicated a capacity of 3.9×10^9 L annually [74]. Annual U.S. production capacity in 2006 was reported as 1.68×10^{10} L [75]

In other applications, the dehydration of glucose to 5-hydroxymethylfurfural allows the incorporation of glucose-containing syrups into phenolic and urea – formaldehyde-based plastics [76]. Heating glucose and methanol in the presence of anhydrous hydrogen chloride yields α-methyl-D-glucoside which is used for insulating foams. Catalytic hydrogenation of glucose yields sorbitol.

Glucose-bearing syrups are also added to adhesives to improve stability and flow control, used as humectant in wallboard and tobacco manufacture, added to tiles and bricks to prevent cracking during manufacture, used to retard setting in concrete, are applied in resin and dye manufacture and leather tanning to name but a few of their nonfood uses.

11. Economic Aspects

The economics of starch-based sweeteners are related to the economics of sucrose (cane and beet sugar) except for uses in which production costs and unique functionality allow higher prices (such as high-purity crystalline glucose for intravenous injection) or significant competition causes price reduction (such as glucose syrups in the United States). In countries where

no protective tariff exists and sucrose is sold essentially at world prices, the domestic consumption volume of starch-based sweeteners depends on sucrose pricing. For non tariff nations, consumption and pricing figures can vary significantly from year to year. The production and pricing of starch-based sweeteners around the world are shown in Table 11.

Raw material (starch) pricing is also an important factor. Competition of markets for starch-based products, primarily used for the production of high-fructose syrups and fermentation to fuel-grade ethanol has become important in recent years. The impact of these two products on the growth of the corn wet milling industry in the United States is shown in Figure 13. In 2004, corn (maize) consumption for ethanol production amounted to more than half of that consumed by all food and industrial uses in the United States. While the food and industrial use of maize in the United States has grown substantially since the 1970s, farm production has more than kept up. The price of maize during this period has trended downward (see Fig. 14).

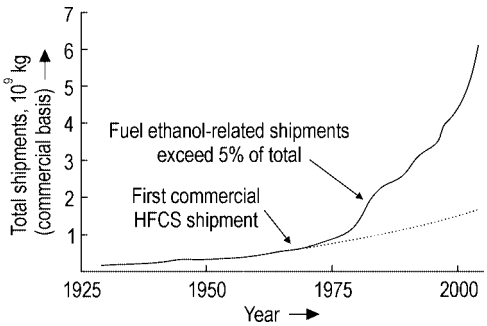

Figure 13. Growth of corn wet miller's shipments in the United States 1929-2005, comprising starch, refinery products, fuel ethanol and coproducts

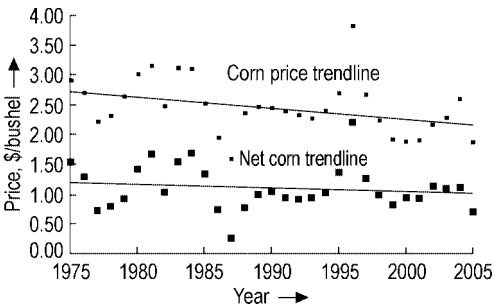

Figure 14. Average annual corn and net corn price 1 bushel corn = 25.4 kg

But the real cost of maize to starch-based sweetener manufacturers is the cost of maize (corn) minus the income from the sale of coproducts of the wet milling process (corn gluten feed, corn gluten meal, and corn oil), i.e., the net corn cost. Year-to-year, net corn cost has followed corn price in the USA since the 1970s but the trendlines for corn and net corn cost are converging, indicating an average reduction in coproduct credit of 37 cents per bushel over the past 30 years (see Fig. 14.). The rapid growth in starch use in non-sweetener applications and the impact on raw material cost and the price of glucose-bearing products has been analyzed [77, 78]. To date, it appears that agricultural productivity increases have more than kept up with expanding food and industrial use of maize in the United States, keeping maize cost stable. For wet millers in the USA, increased volumes of coproduct have negatively impacted their prices, primarily for corn oil and corn gluten feed. Since fuel ethanol production is expected to nearly double by 2012, net corn cost must be watched.

12. Physiological Properties and Toxicology of Glucose

Glucose is absorbed into the bloodstream and used by the body (via the glycolytic cycle) for energy production or converted into other metabolic intermediates required for growth or maintenance of body tissues. By polymerization into glycogen (a highly branched macromolecule referred to as body starch), it is also stored in the liver for future use.

Glucose (and other sugars) acts as a dietary chelating agent under certain conditions, binding iron and several other di- and trivalent metals [79]. Simultaneous ingestion of glucose and a calcium supplement is reported to boost the body's uptake of that mineral by 25 % [80]. Blood serum cholesterol and phospholipid levels are reportedly lowered within days of beginning a high-glucose, lipid-free diet, whereas single doses of glucose syrup may be responsible for improved intellectual performance [81].

The lethal dose of glucose has been reported as 20 g/kg of body weight when administered orally to rabbits (the lethal intravenous dose for the rabbit is 15 – 35 g/kg). The lethal oral dose for dogs is ca. 8 – 12 g/kg [82].

Table 11. Production and pricing of starch-based sweeteners by country (all production values in 10^3 t of dry substance per year, 1996 values used except where noted.)

Country	Glucose syrups	Maltose syrups	Crystalline glucose		Total sugar	Fructose-glucose syrup	
			Monohydrate	Anhydrous		42 %	55 %
United States							
Production	3200	645	25	68	2300		5000
Price, US$/kg d.s.	0.30		0.62	1.25	0.88	0.41	0.45
Europe (EU 15)							
Production	2200[ab]		320[ab]	45	330[ab]		
Price, US$/kg d.s.	0.73[c]			0.96[a]		0.88	0.69[c]
Canada							
Production	118[d]					300[e]	
People's Republic of China							
Production	250	100		250[f]	50		100[e]
Price, RMB[f]/kg d.s.	4	4		6		3	6[h]
Japan							
Production		440[ai]	35[d]d	39[d]d	44[d]d		845[ae]
Price, Yen/kg d.s.	34[a]	46[a]	95[a]	126[a]	113[d]	48[a]	54[a]
Latin Americad							
Production		285		9		36	155
Price, US$/kg d.s.	0.30		2.86		0.17	0.40	
Korea							
Production	200[j]		33[dk]	3[dk]	55[dk]		250[el]
Price, Won/kg d.s.	384l		420l	420l	390[m]		390[l]
South Africa							
Production			175[i]	3			
Price, US$/kg d.s.	0.43[n]		0.68				
Australia[o]							
Production	57						4.3
Malaysia[o]							
Production	3.8			2.4			
Price, US$/kg d.s.			0.57			0.59	
Bangladesh[p]							
Production	3.5			1.0			
Price, US$/kg d.s.			0.54			0.64	

[a] 1995 data.
[b] 90+% of the volume shown is sold within the EU15; less than 10% is sold in other European countries.
[c] 1983 data.
[d] 1986 data.
[e] Combined production of 42 % db and 55 % db fructose-bearing syrups shown.
[f] Most crystalline (pure) dextrose is monohydrate, some is anhydrous. Main uses are medical in nature.
[g] J. At the time of the preparation of this chart, US$1.00 = 8.5 RMB.
[h] Price for 42 % db fructose syrup shown; only one known producer of 55 % fructose syrup in the P.R.C.
[i] Combined production data for glucose and maltose syrups.
[j] 1992 data.
[k] Totals for the production of these three products in 1992 virtually unchanged from 1986, individual product data not available in 1992 so 1986 values are shown.
[l] 1994 data.
[m] 1991 data.
[n] 1987 data.
[o] 1996 list price for 43 DE glucose syrup.
[p] 2005 data.

13. References

1. E. Martin: *Dextrose Therapy in Everyday Practice*, Harper, New York 1937, p. 5.
2. H. Wichelhaus: *Der Stärkezucker*, Akademische Verlagsgesellschaft, Leipzig 1913.
3. Lie-Tien Chang: *Starch/Stärke* **44** 1992 no. 3, 117.
4. G. R. Dean, J. B. Gottfried: "The Commercial Production of Crystalline Dextrose," in C. S. Hudson and S. M. Cantor (eds.): *Advances in Carbohydrate Chemistry, Vol. 5*, Academic Press, New York 1950, pp. 127 – 143.
5. G. S. C. Kirchoff, *Academic Imperiale des Sciences de St. Petersbourg, Memoires*, **4** (1811) 27.

6. C. E. G. Porst, *La. Plant. Sugar Manuf.* **67** (1921, Jul. 2) 1921 14 – 15.
7. Corn Products Refining Co., US 1 471 347, 1923 (W. B. Newkirk).
8. R. O. Marshall, E. R. Kooi, *Science* **125** (1957) 468 – 469.
9. CPC International, US 2 950 228, 1960 (R. O. Marshall).
10. J. P. Casey, *Stärke* **29** (1977) no. 6, 196 – 204.
11. 1965 – 1980 individual industry data reproduced with permission from: *Cereal Foods World* **31** (1986) no. 12, 859 and updated with 1985, 1990 1997 and 2002 industry data estimates. Total volumes from USDA ERS data.
12. United States Dept. of Agriculture (USDA) Sugar and Sweetener Situation and Outlook Report SSRV 12 N 1, Mar., 1987.
13. "Annual Review of the World HFCS Industry" in Sweetener Analysis November, 2005, p. 2, LMC International Ltd, Oxford 2005.
14. S. Vuilleumier: *Int Sugar Jnl.* **98** (1996) no. 1173, 467–478.
15. J. Cropley: "The Global Market Outlook for Corn Refiners", 2002 Corn Utilization Technology Conference Proceedings, p. 128–130.
16. USDA-ERS: Sugar and Sweetener Situation and Outlook Yearbook, July, 2004, Table 50.
17. S. Haley: "Measuring the Effect of Imports of Sugar-Containing Products on U.S. Sugar Deliveries", USDA Publication SSS-237-01, September, 2003.
18. P. A. Meyer, *Milling & Baking News* **65** (1986, Aug. 26) no. 26, 58.
19. "Dry HFCS: Problem Solved," *Food Eng.* **59** (1987) no. 4, 112.
20. W. B. Newkirk, *Ind. Eng. Chem.* (1924, Nov.) 1173.
21. G. R. Dean, *Anal. Chem.* **45** (1973, Dec.) no. 14, 2441.
22. G. R. Dean, *Carbohydr. Res.* **34** (1974) 315 – 322.
23. W. Pigman, E. F. L. J. Anet in W. Pigman, D. Horton (eds): *The Carbohydrates: Chemistry and Biochemistry*, 2nd ed. IA, Academic Press, New York 1972, p. 170.
24. Paper presented by CPC International at Detmold (Germany) Conference, April 23 – 25, 1986.
25. see also F. E. Young, *J. Phys. Chem.* **61** (1957) 616 – 619.
26. J. C. Speck, *Adv. Carbohydr. Chem.* **13** (1958) 63 – 103.
27. Data source: American Maize Products Co. (1963).
28. Data source: A. E. Staley Mfg. Co. (undated).
29. J. J. M. Swinkels, *Starch/Stärke* **37** (1985) no. 1, 3.
30. G. M. A. van Beynum, J. A. Roels (eds.): *Starch Conversion Technology*, Marcel Dekker, New York – Basel 1985.
31. Private Communication Dr. H. F. Zobel (ABCV Starch Assoc., 1105 Belair Dr. Darien, Il. USA 60559).
32. C. A. Brautlecht: *Starch: Its Sources, Production and Uses*, Reinhold, New York 1953, p. 276.
33. Fachverband der Stärke-Industrie e.V., *Stärke Fortschritt durch Tradition 1946 – 1996*, p. 48 – 49.
34. D. Howling in G. C. Birch, K. J. Parker (eds.): *Sugar: Science and Technology* Applied Science, Publ., London 1979, p. 260.
35. R. C. W. Berkeley, G. W. Gooday, D. C. Ellwood (eds.): *Microbial Polysaccharides and Polysaccharases*, Academic Press 1979, p. 361.
36. J. M. Segraves, W. G. Allen, J. Harris, K. L. Pratt: "Valley "Ultra-Thin" – A novel alpha-amylase for true pH 4.5 liquefaction", *Int. Sugar J.* **107** (2005) no. 1283, 616–621.
37. Archer Daniels Midland Company, 5 869 297, 1999 (T. P. Binder, D. K. Hadden, L. J. Sievers).
38. Y. Komai, *Stärke* **17** (1965) no. 11, 346 – 350.
39. Nippon Shiryo Kogyo Co., Ltd., US 3 540 927, 1970 (N. Masahiro, T. Furukawa, H. Masada).
40. Corn Products Company, US 3 257 665, 1966 (L. R. Idaszak).
41. Struthers Scientific and International Corporation, US 3 506 486, 1970 (N. Geniaris).
42. CPC International, US 4 357 172, 1982 (L. W. Edwards).
43. A. F. M. Mahbubar Rahman, F. W. Schenck: "Supply solids level and monohydrate dextrose productivity", *Int. Sugar J.* **101** (1999) no. 1207, 379, 380.
44. Corn Products Refining Co., US 1 508 569, 1924 (W. B. Newkirk).
45. Corn Products Refining Co., US 1 890 913, 1932 (W. B. Newkirk).
46. Corn Products Refining Co., US 2 152 874, 1939 (C. J. Copland).
47. P. Elankovan et al, *AIChE J.* **33** (1987, Nov.) no. 11, 1844.
48. [4] pp. 133, 134.

49. R. V. MacAllister in R. S. Tipson, D. Horton (eds.): *Advances in Carbohydrate Chemistry and Biochemistry*, vol. **36**, Academic Press, New York 1979, p. 41.
50. Corn Products Refining Co., US 2 369 231, 1945 (A. T. Harding).
51. C. Nieman, *Manuf. Confect.* **XL** (1960) no. 8, 19 – 24, 43 – 46.
52. S. A. Watson, P. E. Ramstad (eds.): *Corn: Chemistry and Technology*, Am. Assoc. of Cereal Chem., St Paul, Minn. 1987, p. 523.
53. L. Hyvonen, et al., *Lebensm.-Wiss. Technol.* **11** (1978) no. 1, 11 – 14.
54. M. W. Kearsley et al., *Lebensm.-Wiss. Technol* **11** (1978) no. 1, 23 – 25.
55. L. M. Hanover in G. Charlambous, G. Inglett (eds.): *Chemistry of Foods and Beverages: Recent Developments*, Academic Press, New York 1982, pp. 211 – 233..
56. A. Rapaille et al., *Bios*, **8** (1977) no. 10, 19 – 22.
57. S. Yamaguchi et al., *Agric. Biol. Chem.* **34** (1970) no. 2, 181 – 186.
58. A. C. Dahlberg et al., New York State Ag. Exp. Station Bulletin, No. 258, Apr., 1941.
59. R. Whymper, *Manuf. Confect.* **35** (1955) no. 6, 15, 16, 19, 20, 22.
60. N. Deerr: *Int. Sugar J.* **24** (1922) 481.
61. *Kirk-Othmer*, 3rd ed., **22**, p. 508.
62. L. C. Maillard, *Compt. Rend.* **154** (1912) 66.
63. H. W. Keller et al., *Starch/Stärke* **33** (1981) no. 2, 55 – 57.
64. J. R. Teague et al., *Sugar/Azucar* **78** (1983) no. 8, 18 – 23.
65. M. Ruffolo: "Starch and Chemical Sweeteners in the Changing Sweetener World" 3rd Conference of World Sugar Farmers – World Association of Beet and Cane Growers, Gold Coast, Australia, July 21, 1987.
66. J. S. Bacon et al., *Biochem. J.* **51** (1952) 208.
67. M. Kierstan, *Process Biochem.* **15** (1980) no. 4, 4.
68. M. Kierstan, *Biotechnol. Bioeng.* **20** (1978) 447.
69. D. L. Kiser et al. in G. Charalambous (ed.): *Symposium on the Analysis of Foods and Beverages by HPLC*, Academic Press, New York 1979, pp. 363 – 378.
70. *Approved Methods of the American Association of Cereal Chemists (AACC)*, 8th ed., AACC, St. Paul, Minn., 1983.
71. Code of Federal Regulations effective April 1, 1987 Title 21 – Part 168 – Section 168.111.
72. U.S.P., 21st ed., 300.
73. A. S. Greditor: *Research Report, CPC International*, Drexel Burnham Lambert, New York Nov. 10 1986 13.
74. J. G. Joyce, *Investor News* **10** (1988) no. 3, 1, 4.
75. Renewable Fuels Association, From Niche to Nation: Ethanol Industry Outlook 2006 available at http://www.ethanolrfa.org.
76. H. Koch, J. Pein, *Polym. Bull.* **13** (1985) 525 – 532.
77. F. W.Schenck: "The impact of growth on the US corn wet milling industry", *Int. Sugar J.* **101** (1999) no. 1210, 510–516.
78. F. W.Schenck: "Factors influencing the price of head products of the corn wet milling industry", *Int. Sugar J.* **103** (2001) no. 1234, 462–466.
79. P. Saltman: "Role of Simple Carbohydrates in Metal Absorption," *Proc. Corn Refiners Association Scientific Conference*, 1988.
80. *Insight* **4** (1988) no. 9, 52.
81. G. G. Birch, M. W. Kearsley, *Stärke* **29** (1977) no. 10, 348 – 352.
82. W. S. Spector (ed.): *Handbook of Toxicology*, vol. **I**, Saunders, Philadelphia – London 1956, pp. 150 – 151.
83. F. W. Schenck, R. E. Hebeda: *Starch Hydolysis Products*, VCH Publ., 1992, p. 288.
84. F. W. Schenck, *Starch*/Stärke **48** (1996) no. 5, 188 – 190.
85. A. F. M. Mahbubar Rahman, F. W. Schenck, "Improving monohydrate dextrose economics with enzymatic reconversion", *Int. Sugar J.* **100** (1998) no. 1197, 457–461.
86. M. Parameswaran, *Food Australia* **46** (1994) no. 10, 473 – 475.
87. P. Coussement, *Food Ingredients and Analysis International* **18** (1996) no. 5, 25 – 27.

Glycerol

Ralf Christoph, Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany
Bernd Schmidt, Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany
Udo Steinberner, Cognis Deutschland GmbH, Düsseldorf, Federal Republic of Germany
Wolfgang Dilla, Solvay Soda Deutschland GmbH, Rheinberg, Federal Republic of Germany
Reetta Karinen, Helsinki University of Technology, Finland

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 4
4. Production 4
4.1. Glycerol from Fats and Oils 5
4.1.1. Fat Splitting 5
4.1.2. Pretreatment and Concentration of Crude Glycerol 5
4.1.3. Purification and Refining 6
4.2. Synthesis from Propene 8
4.2.1. Production from Allyl Chloride 8
4.2.2. Production from Acrolein 9
4.2.3. Production from Propylene Oxide . . 9
4.3. Other Processes 9
5. Environmental Protection 10
5.1. Production from Fats and Oils . . . 10
5.2. Production from Epichlorohydrin . 10
5.3. Environmentally Important Properties of Glycerol 11
6. Quality Specifications and Analysis 11
7. Storage and Transportation 12
8. Derivatives 12
8.1. Esters 12
8.2. Chlorohydrins 12
8.3. Oxidation Products 12
8.4. tert-Butyl Ethers 12
9. Use of Glycerol and its Derivatives . 13
9.1. Use in Foods, Cosmetics, and Pharmaceuticals 13
9.2. Use in Plastics, Resins, and Cellophane 14
9.3. Use in Tobacco Industry 14
9.4. Other Applications 14
9.5. Future Applications 14
10. Economic Aspects 14
11. Toxicology 15
12. References 15

Abstract

Glycerol is the simplest triol. It can be found in all natural fats and oils as fatty esters and is an important intermediate in the metabolism of living organisms. It is completely miscible with water and many alcohols and also with many heterocyclic compounds. A remarkable and growing source of glycerol is as a byproduct in the production of biodiesel. Glycerol is presently used in a large variety of applications because of its particular combination of chemical and physical properties and because it is physiologically innocuous.

1. Introduction

Glycerol [*56-81-5*], $C_3H_8O_3$, M_r 92.09, 1,2,3-propanetriol, commonly known as glycerin, is the simplest triol. It can be found in all natural fats and oils as fatty esters and is an important intermediate in the metabolism of living organisms.

Glycerol was discovered in 1779 by Scheele through saponification of olive oil with lead oxide. In 1813, Chevreul showed that fats are glycerol esters of fatty acids. He also gave glycerol its name, after the Greek word $\gamma\lambda\upsilon\kappa\epsilon\varrho o\varsigma$ which

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_477.pub2

means sweet. The first industrial use of glycerol was in 1866 when NOBEL produced dynamite, in which the trinitrate of glycerol — nitroglycerin — is stabilized by absorption on diatomaceous earth.

The most important industrial synthesis of glycerol, which uses propene as starting material, was developed in the late 1930s by I.G. Farben in Germany and by Shell in the United States.

Glycerol is now used in a large variety of applications because of its particular combination of chemical and physical properties and because it is physiologically innocuous. Total production is estimated (1998) to be ca. 750 000 t/a; about 90 % is produced by processing natural oils or fats and 10 % is synthesized from propene [1, 2].

2. Physical Properties

Glycerol is a sweet-tasting hygroscopic liquid; when pure, it is colorless and odorless. At room temperature it is viscous.

The boiling point of glycerol is 290 °C at atmospheric pressure (101.3 kPa). The corresponding values under reduced pressure are listed in Table 1. With some liquids, (e.g., biphenyl), glycerol forms azeotropes; with others (e.g., water), nonazeotropic mixtures are formed. The vapor pressure curves of glycerol – water solutions at various glycerol concentrations are shown in Figure 1. The liquid – vapor equilibria of glycerol – water solutions, important for distillation and fractionation, have been established experimentally [3 – 5]. The boiling- and condensing curves at 13.3 and 1.33 kPa are shown in Figure 2. Suitable theoretical methods of calculation are also available [6].

Table 1. Vapor pressure vs. temperature of glycerol

Temperature, °C	Pressure, kPa
290	101.3
266	53.3
222	13.3
204	6.67
175	2.00
152	0.67
130	0.18
100	0.03
20	<0.0001

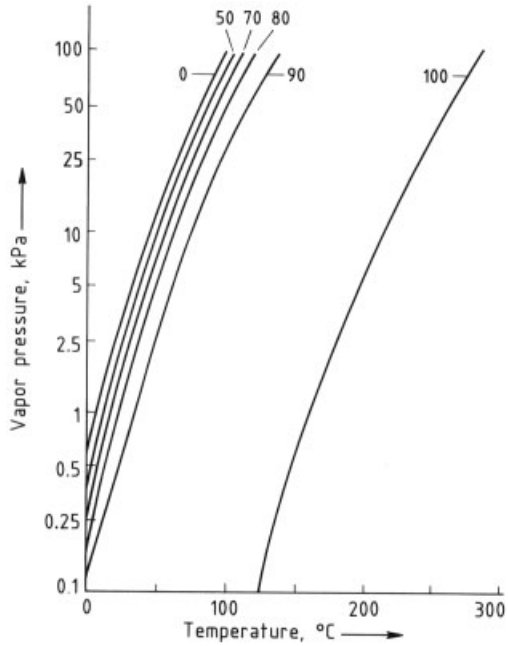

Figure 1. Vapor pressure of glycerol – water solutions (wt % glycerol) [7, 10]

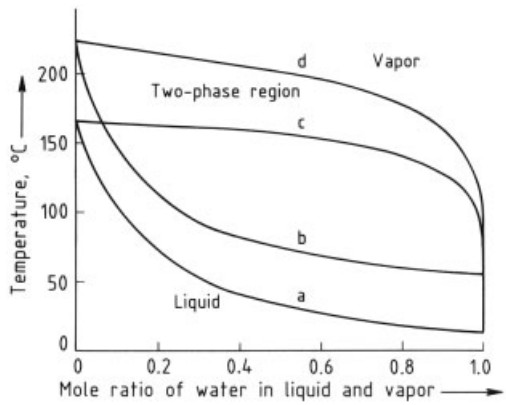

Figure 2. Liquid–vapor equilibria for glycerol–water solutions
a) Boiling curve, 1.33 kPa; b) Boiling curve, 13.3 kPa; c) Condensing curve, 1.33 kPa; d) Condensing curve, 13.3 kPa

The freezing point of pure glycerol is ca. 18 °C. The crystalline state is seldom reached, however, because of a strong tendency toward supercooling. A solid, glassy state can be observed from ca. − 70 to − 110 °C. Crystallization can be brought about at ca. 0 °C by seeding with glycerol crystals. The freezing point

of glycerol – water solutions is shown in Figure 3. At 66.7 wt % glycerol, an eutectic mixture is formed, *fp* − 46.5 °C.

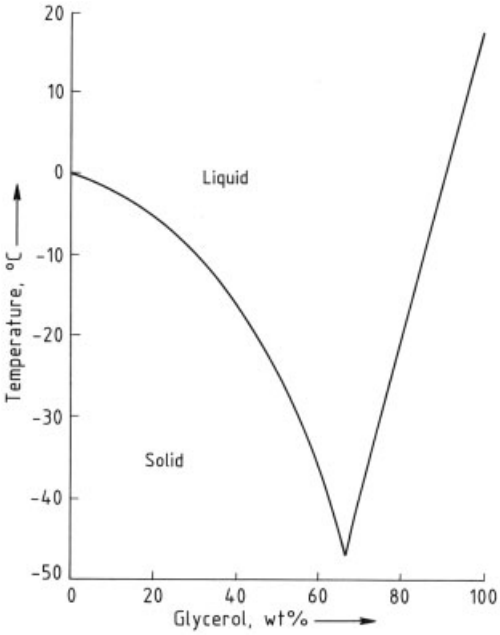

Figure 3. Freezing point of glycerol–water solutions

The density of glycerol is 1.261 g/cm^3 (20 °C). Tables of the density of glycerol – water solutions at various concentrations of glycerol are given in [7]. For rough estimations, a mean density change of 0.0007 g/cm^3 per unit change in temperature may be used in the range from 0 to 200 °C. The compressibility of glycerol is only about half that of water.

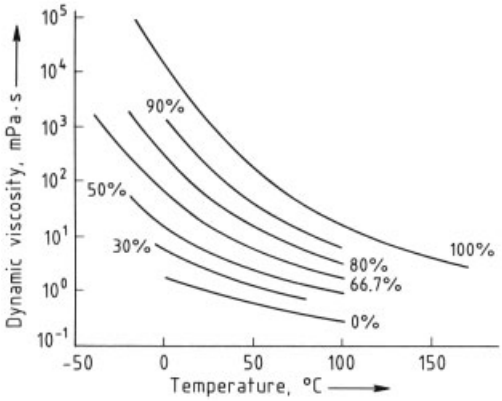

Figure 4. Dynamic viscosity of glycerol – water solutions

The refractive index of aqueous glycerol for the range 0 – 100 % glycerol is tabulated in 1 % steps in [7]. The viscosity of glycerol – water solutions (0 – 100 wt % glycerol) from − 40 to 100 °C has been determined [7]. Some curves are shown in Figure 4. Aqueous glycerol solutions, at any concentration, will supply moisture until equilibrium with water vapor in the atmosphere is reached [7]. The relative humidity over glycerol solutions is illustrated in Figure 5.

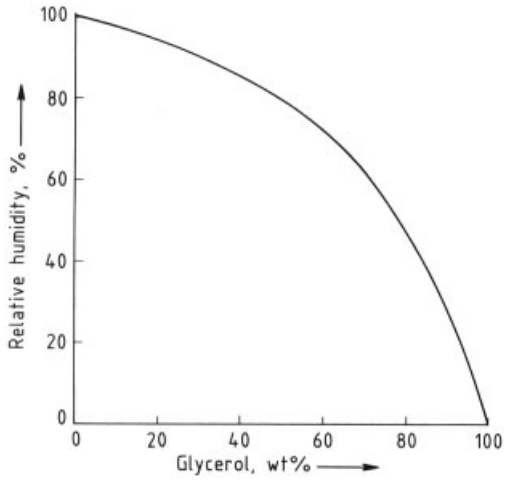

Figure 5. Relative humidity over aqueous glycerol (20 – 100 °C)

Glycerol has useful solvent properties, similar to those of water and simple aliphatic alcohols, because of its three hydroxyl groups. It is, accordingly, completely miscible with water, methanol, ethanol, and the isomers of propanol, butanol, and pentanol. It is also fully miscible with phenol, glycol, propanediols, amines, and heterocyclic compounds containing a nitrogen atom in the ring (e.g., pyridine, quinoline). Its solubility, however, is limited in acetone, diethyl ether, and dioxane. Glycerol is almost insoluble in hydrocarbons, long-chain aliphatic alcohols, fatty oils, and halogenated solvents such as chloroform. Ternary systems, such as glycerol – water – phenol and the nonaqueous system glycerol – ethanol – benzene, display significant temperature-dependent miscibility gaps. Glycerol also forms azeotropes with some other substances [8]. Glycerol is a useful solvent for many solids, both organic and

inorganic (e.g., see Table 2), which is particularly important for the preparation of pharmaceuticals. The solubility of gases in glycerol, like other liquids, is temperature- and pressure-dependent [9]. Some important physical properties of glycerol are presented in Table 2; a comprehensive collection of data is available [10].

Table 2. Physical properties of glycerol

M_r	92.09
mp	18.0 °C
bp (101.3 kPa)	290.0 °C
Density (20 °C)	1.261 g/cm^3
Refractive index n_D^{20}	1.4740
Dynamic viscosity (20 °C)	1.410 Pa · s
Compressibility (28.5 °C)	2.1×10^{-4} MPa^{-1}
Gravity coefficient of thermal expansion (15 – 20 °C)	0.000615 K^{-1}
Surface tension (20 °C)	63.4 mN/m
Heat of formation	−669 kJ/mol
Heat of combustion	−1662 kJ/mol
Heat of vaporization (55 °C)	88.2 kJ/mol
(195 °C)	76.1 kJ/mol
Heat of fusion (18 °C)	18.3 kJ/mol
Heat of solution (infinite dilution)	− 5.8 kJ/mol
Heat capacity (298 K $< T <$ 383 K)	$C_p(T) = 987.98 + 4.7074 \times T$ in J kg^{-1} K^{-1}
(− 80 °C)	1.91 kJ kg^{-1} K^{-1}
(− 108 °C)	0.91 kJ kg^{-1} K^{-1}
Thermal conductivity (0 °C)	0.29 W m^{-1} K^{-1}
Diffusion constant of water into glycerol (20 °C)	1.33×10^{-11} m^2/s
Specific electrical conductivity (20 °C)	0.1 µS/cm
Relative dielectric constant (25 °C)	42.48
Flash point	177 °C
Fire point	204 °C
Autoignition temperature on glass	429 °C
Calorific value	18 kJ/g
Solubility of KCl (99.5 % glycerol) (25 °C)	6.01 g/100 g
(90 °C)	8.78 g/100 g
Solubility of NaCl (99.5 % glycerol) (25 °C)	7.22 g/100 g
(90 °C)	7.31 g/100 g
Solubility of Na_2SO_4 (99.5 % glycerol) (25 °C)	0.20 g/100 g
(90 °C)	0.63 g/100 g
Solubility of N_2 (99.25 % glycerol) (15 °C)	0.553 vol %
Solubility of CO_2 (99.26 % glycerol) (15 °C)	43.8 vol %

3. Chemical Properties

Glycerol is a reactive molecule that undergoes all the usual reactions of alcohols (→ Alcohols, Aliphatic; → Alcohols, Polyhydric). The two terminal primary hydroxyl groups are slightly more reactive than the internal secondary hydroxyl group. In the presence of small amounts of strong mineral acid the odor of acrolein (odor threshold 0.2 – 0.4 ppm) is already quite perceptible at 160 °C. At 200 °C and more the formation of acrolein is vigorous. When heated under neutral or alkaline conditions, only traces of acrolein have been observed. At 180 °C, alkaline glycerol begins to dehydrate, forming ether-linked polyglycerols. At room temperature glycerol rapidly absorbs water; when dilute it is attacked by microorganisms. Glycerol is easily oxidized: the terminal carbon atoms to aldehyde or carboxyl groups and the central carbon atom to a carbonyl group [11, 12]. Biochemical oxidation of glycerol has also been described [13]. Catalytic hydrolysis [14] or microbial conversion [15] of glycerol to acrolein and 1,3-propanediol was also a topic of research.

Industrially important reaction products of glycerol include:

1) Mono-, di-, and triesters of inorganic and organic acids;
2) Mono- and diglycerides of fatty acids formed by transesterification of triglycerides (from fats);
3) Aliphatic and aromatic esters, formed by reaction with alkylating or arylating agents, respectively [16, 17];
4) Polyglycerols, formed by reaction of glycerol and epichlorohydrin with consecutive alkaline hydrolysis or by intermolecular elimination of water with alkaline catalysts;
5) Cyclic 1,2- or 1,3-acetals or ketals, formed by reaction with aldehydes or ketones, respectively;
6) Mono- or diglycerates, formed by the action of alkali or metal alcoholates.
7) Glycerol polyoxyalkylenes, formed by alkoxylation of glycerol(s) with ethylene- or propylene oxide under alkaline conditions.

In addition to these, there is a growing interest to oxidation products of glycerol and ethers, especially tert-butyl ethers (see Chapter 8).

4. Production

Glycerol is obtained as a byproduct in the conversion of fats and oils to fatty acids or fatty acid

methyl esters. This type of glycerol is known as natural or native glycerol, in contrast to synthetic glycerol from propene. Other methods of production (e.g., fermentation of sugar or hydrogenation of carbohydrates) are not industrially important.

The fastest growing source of glycerol is the production of fatty acid methyl esters for biodiesel (see → Automotive Fuels). In Europe alone the production in 2004 was 1.5×10^6 t and it is expected to increase to over 10×10^6 t by the year 2010. Per each tonne of biodiesel 100 kg of glycerol is formed.

In Europe the legislation requires the use of renewable components in fuels and another important factor enhancing the biodiesel introduction is the fact that for many countries it is produced from domestic raw material and thus it reduces the dependence from the imported oil. It is also an environmentally friendly source of energy.

4.1. Glycerol from Fats and Oils

Although a component of all vegetable and animal fats and oils, glycerol does not exist in the free form but as fatty acid esters; all three hydroxyl groups are usually esterified. Such esters are called triglycerides (→ Fats and Fatty Oils; → Fatty Acids). The glycerol content of fats and oils varies between 8 and 14 wt %, depending on the proportion of free acid and on the chain-length distribution of the fatty acid esters. To obtain glycerol, the oils and fats must be split hydrolytically.

4.1.1. Fat Splitting

(For a detailed description of processes used in the splitting of fats and oils see → Fatty Acids, Chap. 3)

Natural glycerol today is mainly produced by high-pressure splitting and transesterification. Glycerol from the neutral saponification of oils is encountered only in small quantities.

High-Pressure Splitting. Splitting under pressure has been known since 1854; continuous process reactors are now used. Water and fat are fed into a splitting column in countercurrent fashion at 2 – 6 MPa and 220 – 260 °C, leading to a ca. 15 % solution of glycerol in water, known as sweet water. This glycerol is marketed as 88 % saponification- or hydrolysis-crude glycerol. Such glycerol is extremely low in ash: a typical value is ca. 0.1 % or less of inorganic salts.

Transesterification. Natural crude glycerol of similar quality can be obtained from the transesterification (either high or low pressure regime [18]) of oils and fats to their methyl esters. Continuous processes are dominating. Glycerol from the low pressure transesterification process has a much higher salt content of 2 – 5 %. The crude glycerol is obtained directly at a concentration of ca. 90 – 92 %.

Saponification. The splitting of fats by saponification of neutral oils is a traditional method; caustic alkali or alkali carbonates are used, as in the production of soap (→ Soaps). The use of calcium hydroxide in the form of milk of lime is also possible.

4.1.2. Pretreatment and Concentration of Crude Glycerol

The pretreatment of glycerol water is an important step in the work-up, which is crucial for trouble-free production and for the quality of the final product. Glycerol water produced by high-pressure hydrolysis is slightly acidic. Dispersed fat and fatty acid components can be largely removed under these conditions simply by settling or by centrifugation. In a subsequent step the concentration is immediately increased to 70 – 90 % glycerol to avoid decomposition by fermentation (resulting in formation of 1,3-propanediol) and to render the crude product stable to storage. A multistage flash evaporator of stainless steel yields the greatest economic advantage. This main pretreatment takes place directly before final purification and refining.

An optimum treatment effect can be obtained in dilute solutions with a glycerol content of ca. 50 %. A two-step process is normally employed. Activated carbon and sodium hydroxide are used as auxiliaries, the former for bleaching and absorption of impurities, the latter to saponify any remaining fat components. A separation by filtration follows each step. The so-called soda-lime process has been valued for

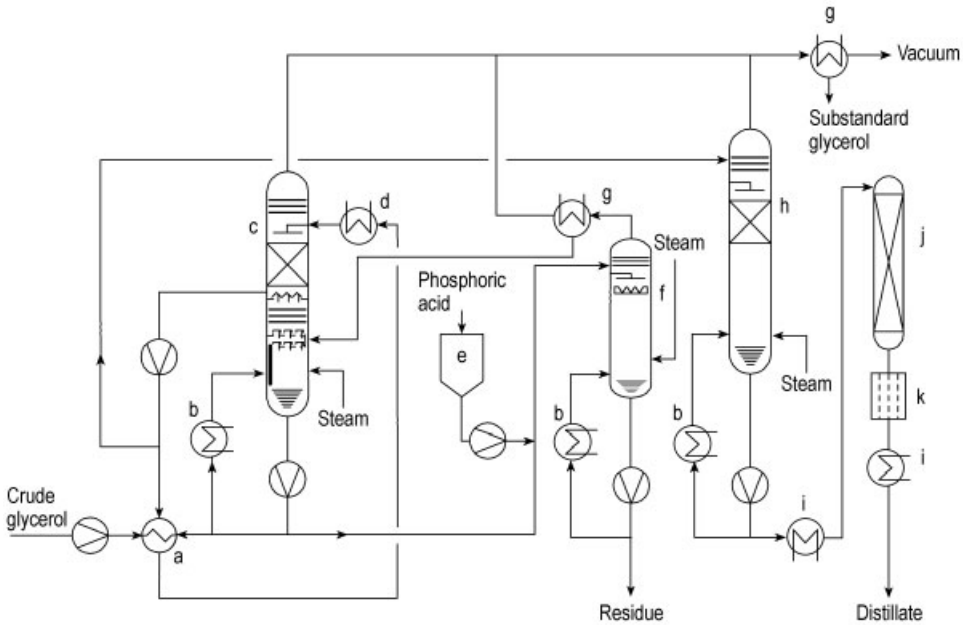

Figure 6. Continuous glycerol distillation (Crown Iron Works Company)
a) Economizer; b) Circulation heater (reboiler); c) Crude still; d) Reflux cooler; e) Dosing system; f) Pitch still; g) Condenser; h) Deodorizer column; i) Cooler; j) Bleaching column; k) Polishing filter

decades for the pretreatment of crude glycerol contaminated with larger amounts of organic impurities [7]. The treated crude glycerol is further concentrated by distillative work-up. Single- or multiple-step plants are used with a vacuum of 10 – 15 kPa. The solubility limits of the salts must be taken into consideration, and the relevant stages suitably equipped for removal of the salt.

Desalination of crude glycerol is also possible by ion exclusion chromatography, using a cationic, strongly acidic exchange resin. The technique was originally developed in the sugar industry. The resin column is first fed with a certain amount of glycerol, the ionic substances remain in the liquid volume between the resin particles (Donnan effect), whereas the nonionic components concentrate in the resin pores. When the resin column is rinsed with water in the second step, the ionic substances are eluted first, followed by the nonionic compounds [19, 20]. By measuring density and conductivity, the process steps can be run completely automatically.

4.1.3. Purification and Refining

Until the end of the 1950s the large-scale production of pharmaceutical-grade glycerol was possible only by *distillation.* Batches were distilled under vacuum with fractional condensation. The development of new technologies allowed conversion to continuous production processes, with correspondingly low losses.

A glycerol refining unit used for the thermal work-up of crude pretreated and evaporated glycerol to glycerol of 99.5 % purity is shown in Figure 6 [21, 22]. For work-up of glycerol with a high salt content, *ion exclusion chromatography* [23] (see Section 4.1.2) or *thin-film distillation* is employed [24, 25]. The glycerol is spread by a rotor in a thin film onto the inner wall of an externally heated column and vaporized. The residue flows or trickles to the bottom. This equipment, combined with a fractionating column, is shown in Figure thin-film distillation is the extremely short residence time; thermal loading of the glycerol is thereby minimized. Refinement of diluted (40 – 50 % glycerol) low-salt crude glycerol can also be carried out by *ion exchange* (see Fig. 8). This process became available with the development of exchange resins

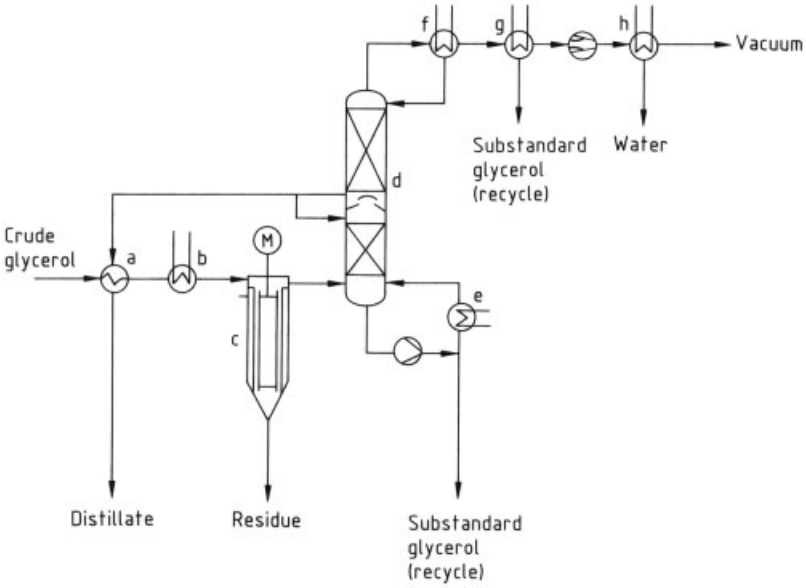

Figure 7. Continuous glycerol distillation (Cognis) [25]
a) Economizer; b) End heater; c) Thin-film distillation; d) Fractionating column; e) Reboiler; f) Reflux condenser; g) Glycerol condenser; h) Water condenser

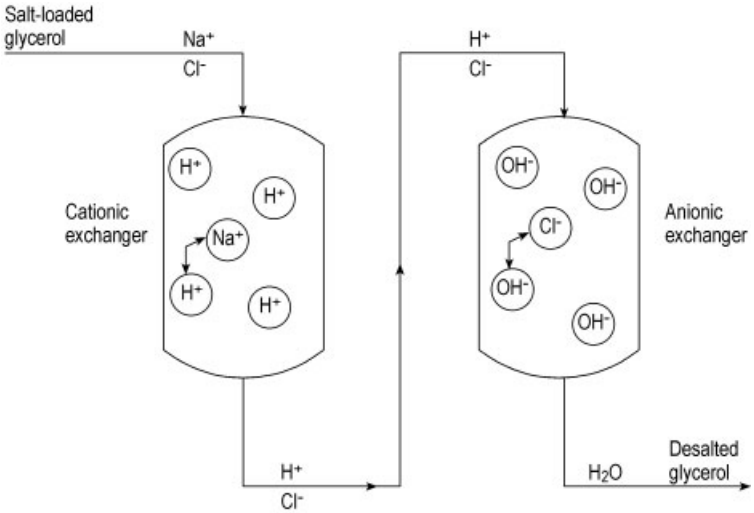

Figure 8. Ion exchanger (principle diagram)

suitable for glycerol [26]. Ion exchangers are used in pairs and consist of a cationic and an anionic exchanger. Charged impurities are thereby successively removed and ultimately exchanged for water. The process permits removal of inorganic salts, fat and soap components, colored matter, odor-causing substances, and other impurities. The so-called adsorption resins possess special decolorizing properties; nonionic components such as chromophores can be removed physically by adsorption [27, 28].

The subsequent concentration of pharmaceutical-grade glycerol solutions is carried out under especially mild conditions to avoid changes in color and odor: a three- or four-step stainless steel flash evaporator with natural circulation is mostly used (Fig. 9).

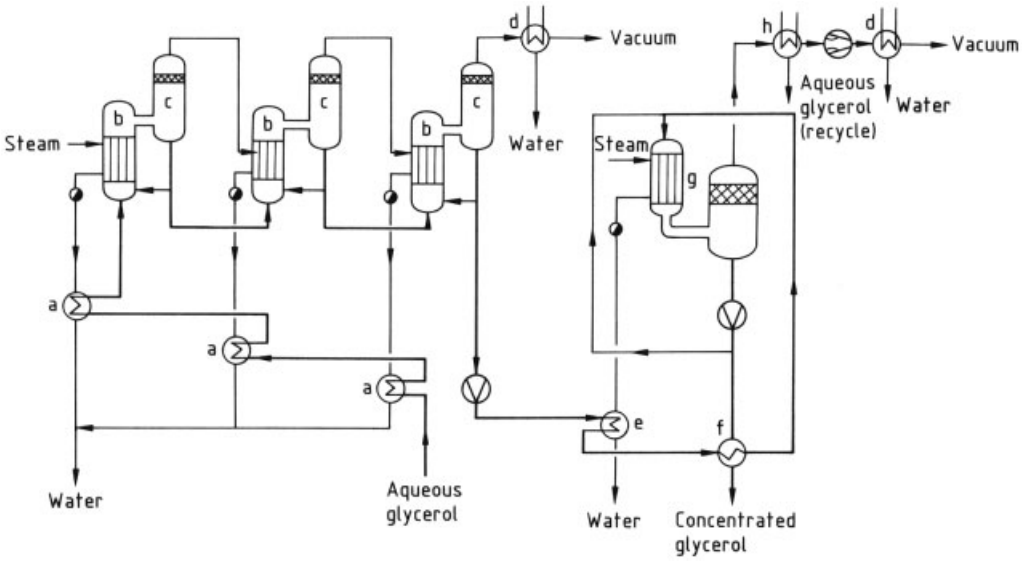

Figure 9. Continuous glycerol concentration
a) Feed heater; b) Evaporator; c) Separator with demister; d) Water condenser; e) Glycerol heater; f) Glycerol heater/final product cooler; g) Falling film evaporator; h) Glycerol condenser

At 10 – 15 kPa, a glycerol concentration of ca. 90 – 95 % is attained. The removal of residual water to give a glycerol concentration > 99.5 % is best carried out in a separate final step by forced circulation in a falling-film evaporator. The vacuum used is 0.5 – 1.5 kPa. The continuous processes described above are suitable for automation when equipped with the appropriate control equipment.

4.2. Synthesis from Propene

Three processes are known for the production of glycerol from propene [*115-07-1*], involving the following intermediate stages (Fig. 10):

1) Allyl chloride – epichlorohydrin
2) Acrolein – allyl alcohol – glycidol
3) Propylene oxide – allyl alcohol – glycidol

Only the first of these is important industrially.

4.2.1. Production from Allyl Chloride

The first synthetic glycerol was produced in 1943 by I.G. Farben in Oppau and Heydebreck and in 1948 by Shell in Houston, Texas. This method became available once the high-temperature chlorination of propene to allyl chloride [*107-05-1*] could be controlled properly (→ Allyl Compounds, Chap. 1.3.1). The allyl chloride produced is oxidized with hypochlorite to dichlorohydrin, which is converted without isolation to epichlorohydrin [*106-89-8*] by ring closure with calcium or sodium hydroxide (→ Epoxides, Chap. 4.1). Hydrolysis to glycerol is carried out with sodium hydroxide or sodium carbonate [29 – 31].

Epichlorohydrin is hydrolyzed to glycerol at 80 – 200 °C with a 10 – 15 % aqueous solution of sodium hydroxide or sodium carbonate at atmospheric or overpressure. The residence time in one or a series of several closed, continuously operating reactors amounts to several minutes or several hours depending on the plant concerned. The yield of dilute (10 – 25 %) glycerol solution is > 98 %. The solution contains 5 – 10 % sodium chloride and less than 2 % of other impurities (Fig. 11). This aqueous glycerol solution containing sodium chloride is evaporated in a multistage evaporation plant under vacuum to a glycerol concentration of > 75 %; precipitated sodium chloride is separated at the same time (Fig. 12). The glycerol solution is then distilled under high vacuum (ca. 0.5 – 1.0 kPa) as shown, e.g., in Section 4.1.3, Figs. 6 and 7; codistilled water is separated by fractional condensation. Residual inorganic salts and higher oligomers

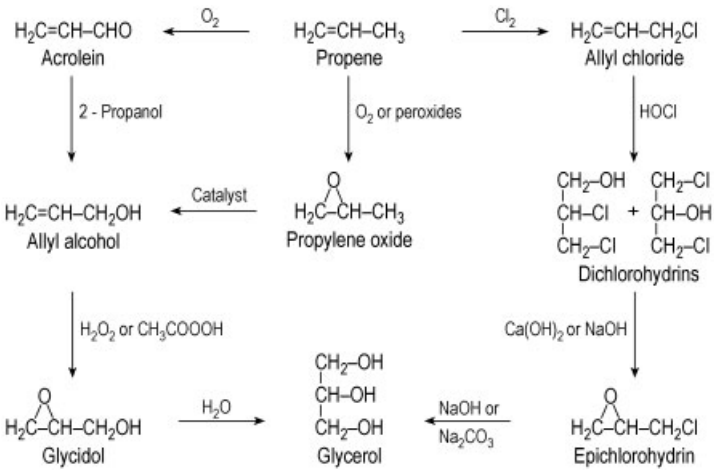

Figure 10. Routes for synthesis of glycerol from propene

of glycerol remaining after the evaporation must be worked up further or discarded. The glycerol, practically free of water, is treated further to remove colored impurities and odorous material; this can be performed, for example, with activated carbon.

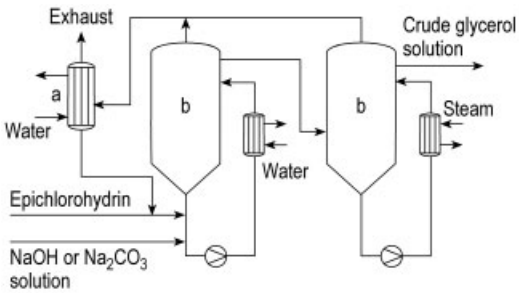

Figure 11. Synthesis of glycerol from epichlorohydrin
a) Exhaust gas cooler; b) Circulation reactor

4.2.2. Production from Acrolein

The acrolein process for the production of glycerol was developed by Shell and does not require the use of chlorine [32]; the first plant was built in 1958 in Norco, Louisiana. Propene is oxidized to acrolein [*107-02-8*], which is then reduced to allyl alcohol [*107-18-6*] (Meerwein–Ponndorf–Verley reduction). The allyl alcohol is epoxidized with hydrogen peroxide, and the resulting glycidol [*556-52-5*] is hydrolyzed to glycerol.

4.2.3. Production from Propylene Oxide

As with acrolein, preparation of glycerol from propylene oxide does not use chlorine. Propene is epoxidized to propylene oxide, which is then isomerized to allyl alcohol by the Progil process [33] (→ Allyl Compounds, Chap. 2.3.2). A second epoxidation is carried out with peracetic acid, and the resulting glycidol is hydrolyzed to glycerol.

4.3. Other Processes

Fermentation of Sugar. The formation of glycerol by fermentation of alcohol was discovered in 1858 by Pasteur. Industrial use became possible once the mechanism was understood; the fermentation could be interrupted at the glyceraldehyde 3-phosphate stage with sodium carbonate or with alkali or alkaline earth sulfites. After reduction to glycerol phosphate, glycerol is obtained in yields up to 25 % by hydrolysis [34 – 37]. The process is economically unimportant.

Hydrogenation of Carbohydrates. Hydrogenation of natural polyalcohols such as cellulose, starch, or sugar leads to mixtures of glycerol and glycols, which can be separated by distillation. Catalysts used in this high-temperature

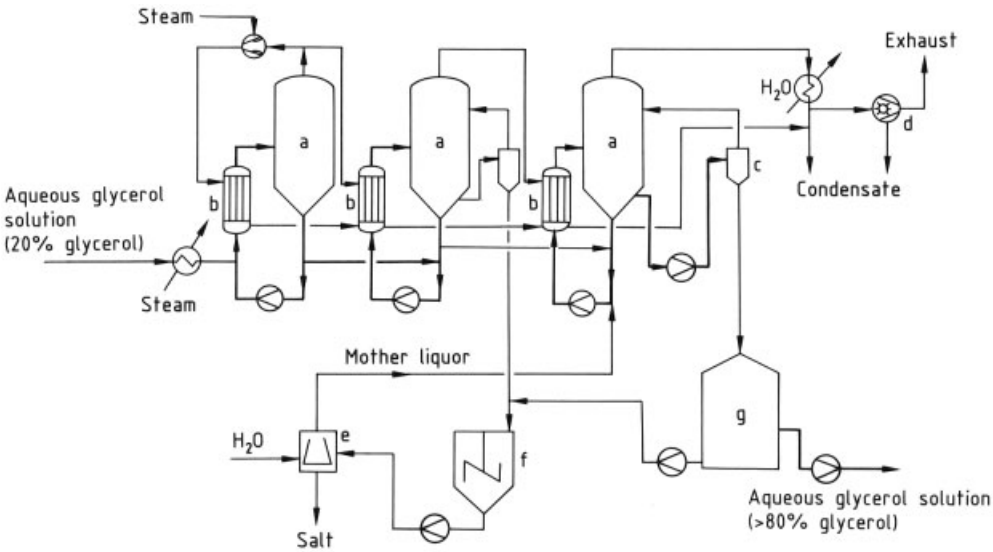

Figure 12. Three-stage evaporation plant for aqueous glycerol solutions: Buss glycerol evaporation process [24]
a) Evaporator vessel; b) Reboiler; c) Cyclone; d) Vacuum pump; e) Centrifuge; f) Stirred tank; g) Storage tank for glycerol solution

reaction include nickel, copper, cobalt, chromium, and tungsten, as well as oxides of some of the lanthanides [38].

Other Noncommercial Processes. Glycerol can be obtained in yields of ca. 7 % by enzymatic splitting of fats and oils with lipases in special reactors [39 – 41] (→ Fatty Acids, Chap. 3.2.3).

Photoproduction of glycerol along with other biomass is possible by means of solar energy and algae [42, 43]. Another synthetic process involves the catalytic hydrogenation of carbon monoxide [44].

5. Environmental Protection

Ecologically, glycerol presents no special problems. Discussion of environmental protection is limited to those industrial processes based on fats and oils and on epichlorohydrin, because only these processes are economically important.

5.1. Production from Fats and Oils

Production Plants. No special protective measures are necessary for the work-up of natural crude glycerol with the technology described above.

Gaseous Emissions. Certain measures are necessary to prevent atmospheric pollution by odors during storage and purification of native crude glycerol. Exhaust vapor from the respective tanks and process units should be drawn off and treated (e.g., burned) in a closed system (→ Fatty Acids, Chap. 7).

Residues. The residues of higher glycerol oligomers and inorganic salts remaining from evaporation and distillation are water soluble and must be treated further, recycled, or disposed of properly.

5.2. Production from Epichlorohydrin

Extensive protective measures must be taken to allow proper handling of epichlorohydrin, because of the hazardous properties of this material (→ Epoxides). Regulations exist in several countries concerning exhaust gas and wastewater purification as well as measures for the protection of plant personnel.

Production Plants. Production plants must be explosionproof to prevent ignition of epichlorohydrin. Manipulation of epichlorohydrin exclusively in closed systems guarantees protection of personnel.

Gaseous Emissions. Epichlorohydrin and accompanying organic material in exhaust gases from the hydrolysis of epichlorohydrin can be removed, for example, by adsorption on activated carbon or can be incinerated.

Wastewater. Wastewater contaminated with epichlorohydrin can be treated with alkali, thereby converting the epichlorohydrin to glycerol.

Residues. Residues from evaporation and distillation should be treated as described in Section 5.1.

5.3. Environmentally Important Properties of Glycerol

Gaseous Emissions. No problems are encountered with gaseous emissions because of the low vapor pressure of glycerol and the lack of odor of the pure material.

Wastewater. Glycerol poses no problems for wastewater because it is completely and readily biodegradable in sewage treatment plants according to OECD classifications. Glycerol is not regarded as dangerous to the water supply: the relevant values are: COD = 1217 mg of O_2 per gram; BOD_5= 780 mg of O_2 per gram. The fish toxicity is >100 mg/L (LC_{50} according to acute fish toxicity test OECD 203). Local regulations regarding wastewater must be taken into account.

6. Quality Specifications and Analysis

Glycerol can be detected qualitatively by heating a small sample with potassium hydrogen sulfate; the presence of glycerol is indicated by the characteristic pungent odor of acrolein.

The most accurate wet method for quantitative determination of glycerol is periodate oxidation. Glycerol is oxidized by sodium periodate to formaldehyde and formic acid:

$$\underset{}{\overset{OH}{\overset{|}{CH_2}}}-\overset{OH}{\overset{|}{CH}}-\overset{OH}{\overset{|}{CH_2}} + 2\ NaIO_4 \longrightarrow 2\ HCHO + HCOOH + H_2O + 2\ NaIO_3$$

Formic acid is determined by titration and is a measure of the amount of glycerol; polyols with three or more neighboring hydroxyl groups (e.g., sugars) interfere. For quantitative determination of impurities in glycerol gas chromatography is employed.

Natural and synthetic glycerols can be distinguished by site-specific natural isotope fractionation nuclear magnetic resonance spectroscopy (SNIF – NMR) and isotope ratio mass spectroscopy (IRMS) [45, 46]. Water in glycerol can be determined quantitatively by Karl Fischer titration or by near infrared spectroscopy (NIR). A rapid method for establishing the glycerol content of commercial glycerol–water mixtures is measurement of the refractive index at 20 °C or determination of the density with a digital instrument that electronically measures the period of oscillation of a swinging body suspended in the glycerol [47]. In contrast, measurement of the density with a pycnometer is, for the same degree of accuracy (0.1 %), considerably more time consuming. Color is given as Hazen or APHA color number; typical color values would be 5 – 20, depending on the glycerol quality. For rapid measurement, equipment is available for direct color comparison with standardized slides. Ash, heavy metals, chloride and organic chlorine compounds, pH, and acidic or alkaline substances can be determined by standard analytical methods or by methods given in various pharmacopoeias. Traces of reducing agents, such as aldehydes, sometimes still present in glycerol can be determined quantitatively by special tests, e.g., with silver nitrate.

Glycerol is sold in crude (natural glycerol) or pure form. In the latter, a distinction is made between various grades that are linked with given sets of specifications and analytical methods:

- Technical glycerol at various concentrations (e.g., Nobel test glycerol 99.5 %, Glycerol 99.8 %)
- Pharmaceutical grade (e.g., Ph. Eur., U.S.P., Japanese Pharmacopoeia): Glycerol Ph. Eur. 85 %, Glycerol Ph. Eur.
- Food grades [e.g. E 422 (EU-directives 98/86/EC, 95/2/EC), Food Chemical Codex grade (FCC); generally recognized as safe (GRAS § 182.1320) for human consumption by the U.S. FDA].

7. Storage and Transportation

Storage. Glycerol should preferably be stored at 40 – 60 °C under nitrogen blanketing. It is not corrosive and presents little risk of ignition because of its high flash point. Highly concentrated glycerol does not corrode steel, but storage tanks of carbon steel must be protected by surface coating to prevent rusting by residual moisture. Glycerol is therefore usually stored in tanks of stainless steel [48, 49] or aluminum. Regulations exist in some countri es to control the construction and arrangement of industrial storage tanks.

Transportation. Glycerol is shipped in tank trucks, containers, and drums. The tank trucks and containers are usually made of stainless steel. Galvanized or resin-coated steel as well as plastic is used for drums.

8. Derivatives

8.1. Esters

Glycerol forms esters with both inorganic and organic acids; depending on reaction conditions and degree of esterification, these can be mono-, di-, or triglycerides. Of most importance are esters produced from nitric, acetic, and fatty acids [7, 51].

Glycerol trinitrate [*55-63-0*], nitroglycerin, is produced in a mixture of nitric and sulfuric acids. It is used as an explosive and pharmaceutical.

The *acetins* are produced industrially by esterification of glycerol with acetic acid, acetic anhydride, or both. They are the most important esters of glycerol from short-chain carboxylic acids. The greatest use is of triacetin.

Glycerol fatty esters, the partial glycerides, are obtained by transesterification of fats or oils with glycerol to a mixture of mono- and diglycerides. The mono- and diesters of fatty acids are edible, as are the triglycerides.

8.2. Chlorohydrins

Chlorohydrins are treated in a separate keyword (→ Chlorohydrins); epichlorohydrin see → Epoxides, Chap. 4.1.

8.3. Oxidation Products

The oxidation products of glycerol are presented in Figure [13]. The components derived from glycerol are valuable intermediates for fine and specialty chemicals and some of them are useful as such. Their current and common production methods involve complex organic syntheses, fermentation and multiple step reactions and the production costs are high and amounts of produced waste are large. Selective catalysed reactions from glycerol offer benefits in reduced productions costs, increased selectivity and reduced amounts of waste and solvents. Studies are concentrated on heterogeneous catalysis and supported platinum and palladium are regarded as the best catalysts for glycerol oxidation e.g.,[52, 53]. Also supported gold catalysts [54] as well as some homogeneous catalysts, such as TEMPO (2,2,6,6-tetramethylpiperidine-1-oxyl) [55], are studied.

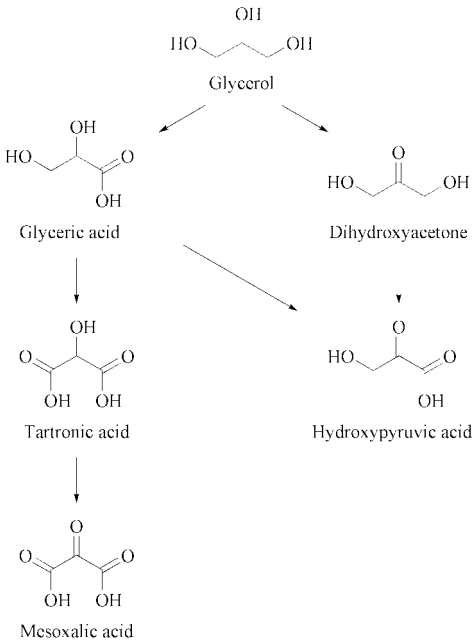

Figure 13. Oxidation products of glycerol

8.4. tert-Butyl Ethers

Glycerol can be etherified with alkenes and especially etherification with isobutene to produce

tert-butyl ethers of glycerol is regarded interesting [56 – 60]. Depending on the extent of the etherification reaction various products are formed. Altogether five isomers are formed as 1–3 hydroxyl groups in glycerol molecule can be etherified and also isomers exist (Fig. [14]). The product distribution can be controlled by process conditions, e.g., temperature and feed compo -sition.

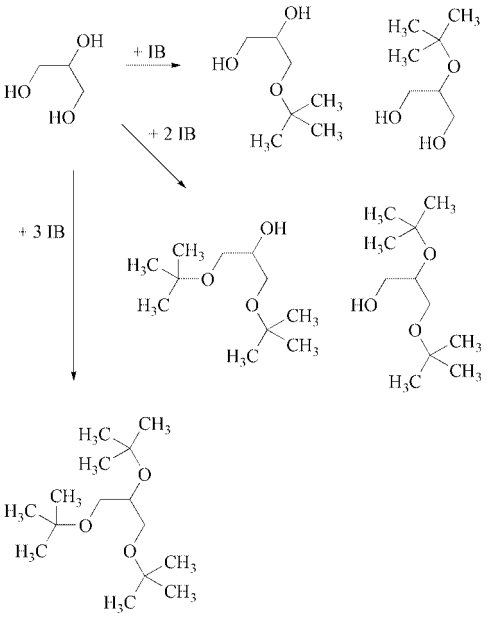

Figure 14. *tert*-Butyl ethers of glycerol

9. Use of Glycerol and its Derivatives

Today, glycerol has more than 2000 different applications that arise from its properties: glycerol is water-soluble, is odor- and colorless, has a certain sweetness, absorbs and retains water, has a capacity to dissolve flavors and dyes, has plasticizer properties, protects against fermentation processes, has antioxidant properties, is nontoxic and easily biodegradable and is a polyol. Information on areas of application is available in the literature [61]; over 800 references are cited in [62] alone. Consumption of glycerol according to major applications is shown in Figure 15.

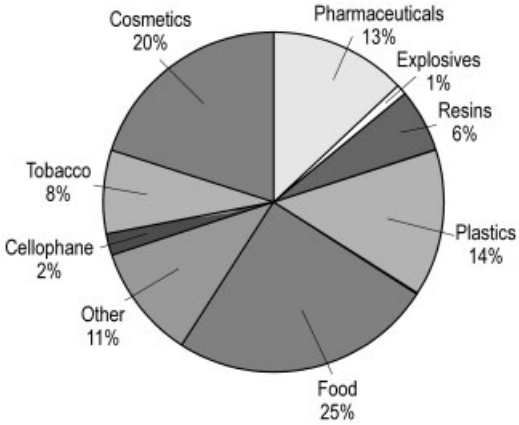

Figure 15. Applications of glycerol (worldwide)

9.1. Use in Foods, Cosmetics, and Pharmaceuticals

Foods. In beverages, syrups, and flavor extracts glycerol is applied as a solvent for the preparation of concentrated flavor extracts from natural products such as vanillin or spices. Furthermore, it is a carrier for many dyes and synthetic flavor compositions and has a good compatibility with sucrose for the preparations of concentrated syrups. Glycerol as such or in the form of its monoglycerides improve s the texture and extends the shelf life of baked goods. Glycerol is used as a conditioner for wine gums, fondants, soft creams, or pastilles. Fruits and vegetables are sprayed with glycerol to keep them in fresh conditions. Mono- and diglycerides are main emulsifiers in ice-cream production, they improve texture, stability and melting properties; in margarine, th ey improve texture and water retention. Glycerol esters of acetic and tartaric esters are used as emulsifiers in bread production; citric acid esters are used in sausages and meat products, and lactic acid esters serve, e.g., as emulsifier for desserts. Glycerol is applied as a moisturizer in sweets and dried fruits to prevent loss of water followed by shrinkage and is a preservative for fish, poultry, and meat [62].

Cosmetics. In cosmetics and toiletries glycerol is mainly used of its chemical properties (formation of esters that serve as emulsifiers or conditioners), moisturizing, and conditioning

properties. Glycerol finds application in creams, lotions, and preparations for skin and hair care as well as in tooth pastes and dental mouthwashes [63].

Pharmaceutical applications of glycerol arise from its conditioning, self-preserving, and dissolving properties in syrups, ointments, and parenteral solutions. Glycerol trinitrate is an important drug for the treatment of angina pectoris. Glycerol also is used as a mild laxative in suppositories. It is a softener in gelatin capsules and suppositories [64].

9.2. Use in Plastics, Resins, and Cellophane

The polyfunctional reactivity of glycerol is used in the production of phthalic and maleic alkyd resins (→ Alkyd Resins) and in cross-linked polyesters from mono- and dicarboxylic acids and polyols. Polyether polyols made of glycerol in combination with propylene- and ethylene oxide are polymerized with diisocyanates in the presence of highly volatile solvents to form flexible or rigid polyurethane foams. Glycerol also serves as a moisturizer and plasticiz er in cellulose films.

9.3. Use in Tobacco Industry

Glycerol is in use as a solvent for flavor compositions for tobacco. It also serves as a humidifier and plasticizer for the tobacco. Glycerol triacetate (triacetin) is used as a softener for cigarette filters.

9.4. Other Applications

Glycerol is used as a hydraulic agent and in hydraulic fluid formulations. Technical esters of glycerol with fatty acids are in use for synthetic lubricants in the automotive industry and as lubricants in the food industry. They are used for readily and totally biodegradable loss lubrication, e.g., in the agriculture and forest industries. Polyglycerols have a wide range of applications as emulsifiers. Inks and printing colors contain glycerol to prolong their servic life. The paper industry uses glycerol a s a humidifier, plasticizer, and lubricant.

9.5. Future Applications

tert-Butyl ethers of glycerol have mainly been studied for fuel applications. The octane numbers for the ether product mixture are 112-128 (BRON) and 91-99 (BMON) [65] and these values indicate that in this sense they are suitable to be used as gasoline components. However, these products are not yet in commercial production, but other ethers resembling their structure, such as MTBE (methyl-*tert*-butyl ether) and ETBE (ethyl-*tert*-butyl ether), are currently used in gasoline as high -octane components. The oxygen-containing components, i.e., oxygenates have been observed to enhance the burning of the fuel resulting in cleaner burning and less-harmful exhaust gases.

The ethers are mainly studied for the gasoline applications but it has also been suggested that they could be used as biodiesel additive [66]. The addition of the ethers in the biodiesel decreases the viscosity, cloud point, and pour point which enables the fuel to be used in lower temperature. Thus the properties of biodiesel are brought closer to conventional diesel fuel.

10. Economic Aspects

Estimated worldwide consumption of glycerol amounts to ca. 750 000 t/a; about 65 % of the total production is in Europe and the United States. Annual production (10^3 t) is shown in Figure 16.

The fraction of synthetic glycerol is ca. 10 %.

The price of glycerol is subject to fluctuation. In times of higher prices, glycerol is partly replaced in certain applications by products such as glycols or sorbitol. Nevertheless, the consumption of glycerol increased from 550 000 t to 750 000 t from 1986 to 1998 corresponding to an increase of ca. 3 % per year.

As glycerol is a byproduct in many processes its production rate is not dependent on the market demand. Since 1995 there has been oversupply of glycerol in the world market and the increasing production of biodiesel increases the oversupply further. It has been estimated that by

the year 2020 the production is six times more than the demand.

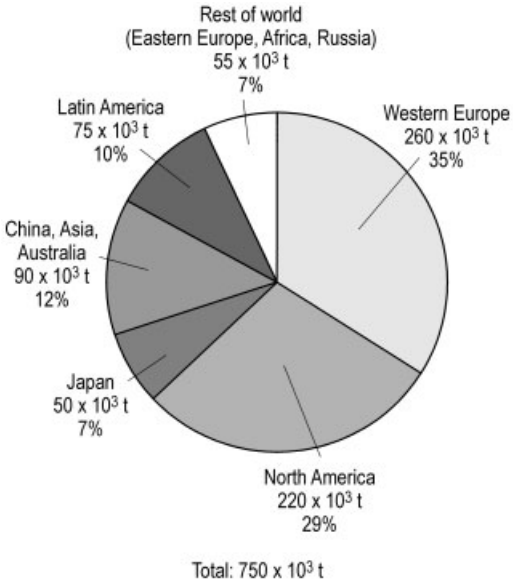

Figure 16. Annual production of glycerol (worldwide)

11. Toxicology

Glycerol is not harmful to health. Oral doses have a demulcent and mildly laxative effect. Ingestion of large amounts may produce headache, thirst, nausea, and hyperglycemia. Slight irritation of the skin or mucous membranes is possible on contact with undiluted glycerol because the strongly hygroscopic glycerol absorbs water from the skin.

There is no glycerol-specific aerosol limit. However, in UK and USA the recommended general long-term (8-hour TWA) exposure limit for aerosols is 10 mg/m^3 and has also to be observed for glycerol mists.

12. References

1. M. Heming, "Glycerin Market Report", January 1999, CTVO-Networkshop.
2. S. Claude, *Fett/Lipid* **101** (1999) 101 – 104.
3. D. F. Stedmann, *Trans. Faraday Soc.* **24** (1928) 289 – 298.
4. I. C. Chu, S. L. Wang, S. L. Levy, R. Paul: *Vapor Liquid Equilibrium Data*, I. W. Edwards, Michigan 1956.
5. N. M. Sokulov, L. N. Tsygankova, N. M. Zhavornokov, *Khim. Promst. (Moscow)* **48** (1972) 96; *Chem. Abstr.* **76** (1972) 154 052 z.
6. R. C. Reid, I. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids*, 3rd ed., McGraw-Hill, New York 1977; *VDI Wärmeatlas*, 8th ed., VDI Verlag GmbH, Düsseldorf 1997/98.
7. C. S. Miner, N. N. Dalton: *Glycerol*, Reinhold Pub. Corp., New York 1953.
8. L. H. Horsley, *Ind. Eng. Chem., Anal. Ed.*, **19** (1947) 508 – 609.
9. A. Hammel, *Z. Physik. Chem.* **90** (1915) 121 – 125.
10. *Physical Properties of Glycerine and its Solutions*, Glycerine Producers Association, New York 1963.
11. *Chemical Properties and Derivatives of Glycerine*, Glycerine Producers Association, New York 1965.
12. R. Garcia, M. Besson, P. Gallezot, *Appl. Catal.* **127** (1995) 165 – 167.
13. L. H. Ohrem, H. Vosss, *Process Biochem.* **31** (1996) 295.
14. E. Drent, W. W. Jäger, PCT Int. Appl. WO 99/05085, 4.2.99
15. A. P. Zeng, H. Biebl, W. D. Deckwer, *ACS Symp. Ser.* **64** (1997) 666.
16. K. Urata, N. Takaishi, *JAOCS* **73** (1996) 819.
17. K. Urata, N. Takaishi, *J. Surf. Det.* **1** (1998) 73.
18. B. Gutsche, *Fett/Lipid* **99** (1997) 418 – 427.
19. D. R. Asher, D. W. Simpson, *J. Phys. Chem.* **60** (1956) 518 – 521.
20. P. Deckert, W. Arlt, *Chem.-Ing.-Tech.* **66** (1994) 1334 – 1340.
21. Glycerine recovery, Crown Iron Works Company, Minneapolis (Minn.), U.S.A. (company brochure '96 and information '99).
22. Am. Oil Chem. Soc.: "Short Course on Fatty Acids", Kings Island, Ohio, Sept. 13 – 16, 1987.
23. H. W. Keller, A. W. Michalson, A. D. Payne, *JAOCS* **33** (1956) 435 – 437.
24. *Glycerine*, Buss SMS GmbH, Butzbach, Federal Republic of Germany; Company brochure.
25. EP 014 135 8531, 184 (R. Brockmann et al.).
26. *Lewatit/Lewasorb*, Bayer AG, Leverkusen (FRG); Company brochure.
27. F. Martinola, *Chem. Ing. Tech.* **51** (1979) 728 – 736.

28. K. Terelak et al., *Chem. Tech. (Leipzig)* **33** (1981) 315 – 316.
29. Shell Dev. Co., US 2 605 293, 1952 (F. T. Tymstra).
30. Shell Dev. Co., US 2 810 768, 1957 (K. B. Cofer).
31. Pittsburgh Plate Glass Co., GB 926 804, 1963 (W. B. Graybill, H. J. Vogt, D. E. Wiley).
32. Shell Dev. Co., US 2 779 801, 1957 (H. D. V. Finch, A. D. Benedictis).
33. Progil S. A., FR 1 271 563, 1962 (A. Thizy, M. E. Degeorges, E. Charles).
34. P. Vijaikishore, N. G. Karanth, *Process Biochem.* **21** (1986) no. 2, 54 – 57.
35. R. Fusheng, *Xiandai Huagong* **16** (1996) 36 – 40.
36. C. Compagno, F. Boschi, B. M. Ranzi, *Biocatal. Biotransform.* **16** (1998) 135.
37. N. A. Mostafa, Y. H. Magdy, *Energy Convers. Manage.* **39** (1998) 671.
38. Daicel Chem. Ind., JP 58 180 444, 1982 (J. Goto).
39. Kao Corp., DE 3 533 615, 1986 (M. Tanigaki et al.).
40. M. Bühler, Ch. Wandrey, *Fat Science Technology* **89** (1987) 156 – 164; 598 – 605.
41. J. Hirano, *Chem. Econ. Eng. Rev.* **18** (1986) 9 – 13.
42. M. Avron, *Adv. Photosynth. Res. Proc. Int. Congr. Photosynth. 6th* 1983, 745 – 753.
43. B. J. Chem., C. H. Chi, *Biotechnol. Bioeng.* **23** (1981) no. 6, 1267 – 1287.
44. R. Fonseca et al., *High-Pressure Sci. Technol. AIRAPT Conf. 6th 1977* **1** (1979) 733 – 738.
45. B. L. Zhang et al., *J. Agric. Food Chem.* **46** (1998) 1374 – 1380.
46. B. L. Zhang et al., *Anal. Chem.* **71** (1999) 2301 – 2306.
47. Rechnender Digitaler Dichtemesser für Flüssigkeiten und Gase (nach O. Kratky, H. Leopold und H. Stabinger, Graz); Fa. Paar K.G., A-8054 Graz–Austria.
48. DIN 17 440: Nichtrostender Stahl (July 1985).
49. American Society for Testing and Materials (ASTM), no. A 167, Type 304; no. A 167, Type 316.
50. A. A. Newman: *Glycerol*, Morgan-Grampian, London 1968.
51. E. Schlenker: *Das Glycerin*, Wissenschaftliche Verlagsgesellschaft, Stuttgart 1932.
52. H. Kimura, *Appl. Catal. A*, **105** (1993) 147.
53. R. Garcia, M. Besson, P. Gallezot, *Appl. Catal. A*, **127** (1995) 165–176.
54. S. Carrettin, P. McMorn, P. Johnston, K. Griffin, C. J. Kielyc and G. J. Hutchings, *Phys. Chem. Chem. Phys.* **5** (2003) 1329–1336.
55. R. Ciriminna, M. Pagliaro, *Adv. Synt. Catal.* **345** (2003) 383–388.
56. A. Behr, L. Obendorf, *Chemie Ingenieur Technik* **73** (2001) 1463–1467.
57. ARCO Chemical Technology L.P., US 5 476 971, 1995 (V. P. Gupta).
58. Fina Research S.A., EP 0649 829, 1995 (C. Dewattines, H. Hinnekens).
59. Henkel KGAA, DE 4 222 183, 1994 (A. Behr, H. Schmidke, C. Lohr, M. Schneider).
60. ARCO Chemical Technology, US 5 308 365, 1994 (H. S. Kesling, L. Karas, F. J. Liotta).
61. U. Steinbrenner, W. Preuss, *Fat Science Technology* **89** (1987) 297 – 303.
62. *75 Years Henkel Glycerine*, Henkel KGaA, Düsseldorf 1986, company brochure.
63. E. Jungermann, N. O. V. Sonntag: *Glycerine: A Key Cosmetic Ingredient*, Marcel Dekker Inc., New York 1991.
64. The United Kingdom Glycerine Producers' Assoc., London: *Glycerine in pharmacy, glycerine in cosmetics, glycerine-miscellaneous uses*.
65. R. Wessendorf, *Erdöl Kohle Erdgas Petrochem.* **48** (1995) no. 3, 138–143.
66. University of Nebraska, US 6 015 440, 2001 (H. Noureddini).

Glyoxal

Georges Mattioda, Société Française Hoechst, C.R.A., Stains, France

Alain Blanc, Société Française Hoechst, C.R.A., Stains, France

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 2
5. **Uses** 3
6. **Toxicology and Occupational Health** . . 4
7. **References** 4

1. Introduction

Glyoxal [*107-22-2*], ethanedial, diformyl, $C_2H_2O_2$, M_r 58.04, was first prepared in 1856 by Debus, through the controlled oxidation of ethanol with nitric acid [1].

2. Physical Properties

Anhydrous glyoxal is a liquid at ambient temperature; it crystallizes at 15 °C in the form of yellow prismatic crystals and boils at 50.4 °C (101.3 kPa), giving off green vapors with a pungent odor. The liquid has a specific gravity of 1.14 at 20 °C; $n_D^{20.5} = 1.3826$. Electron diffraction and infrared spectra show that the molecule is planar; the trans configuration is the more stable [2].

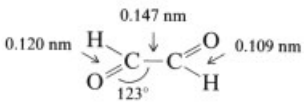

3. Chemical Properties

As the simplest dialdehyde, glyoxal undergoes reactions characteristic of aldehydes (→Aldehydes, Aliphatic and Araliphatic, Chap. 2.2.).

Anhydrous monomeric glyoxal can be obtained by heating polyglyoxal in the presence of phosphorus pentoxide. Anhydrous glyoxal polymerizes rapidly under the action of traces of water, forming a series of hydrated oligomers. Glyoxal is used only as an aqueous solution, at concentrations between 30 and 50 %. Evaporation of this solution produces a white, solid mass of polyglyoxal, which is infusible, depolymerizes upon heating, and decomposes above 150 °C.

In 40 % aqueous solution, glyoxal exists mainly in the form of the hydrated monomer (**1**), together with a dioxolane dimer (**2**) and two bis(dioxolane) trimers (**3**); these components represent ca. 80 % of its total composition [3].

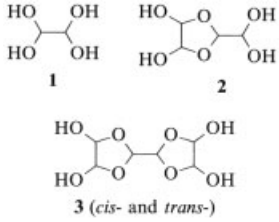

In concentrated aqueous solution the most abundant component is the hydrated trimer (**3**), which is responsible for the cloudiness or crystalline precipitates in aqueous glyoxal solutions.

More highly condensed oligomers are also present, as shown by ^1H and ^{13}C NMR during acetalization by low molecular mass primary alcohols [4]; the species present have been quantified as a function of concentration and temperature.

With alcohols, under acid catalysis, a mixture of oligomeric acetals is obtained. Glycolates are also formed; glycolate formation predominates in the presence of a strong acid [5]. Glyoxal acetals and thioacetals are used as complex-forming agents [6].

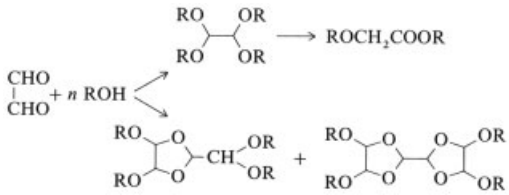

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_491

With primary amines glyoxal forms diimines, reduction of which leads to ethylenediamines, [7] which are used industrially to cross-link isocyanates in coatings.

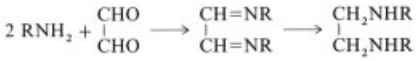

Glyoxal reacts with β-hydroxyalkylamines to give a heterodecalin:

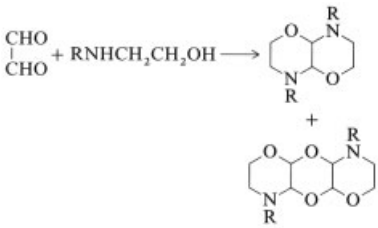

With diethanolamine [*111-42-2*], the final product is *N,N*-bis(2-hydroxyethyl)glycine [*150-25-4*] (R = $-CH_2CH_2OH$):

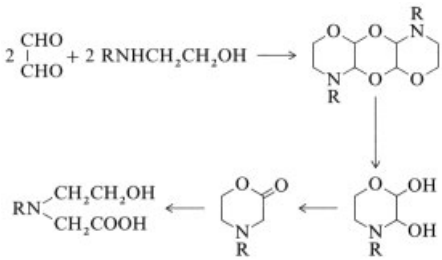

The latter is used to form complexes with polyvalent cations (see also → Ethylenediaminetetraacetic Acid and Related Chelating Agents) [8].

The corresponding aminoacetamides are formed with secondary amines [9], [10]:

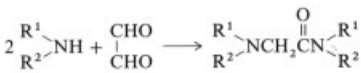

Controlled oxidation leads successively to glyoxylic acid [*298-12-4*] and oxalic acid [*144-62-7*] [11], [12].

With aldehydes and enolizable ketones, glyoxal leads to the expected aldol condensation products [13]; e.g., with acetone or isobutyraldehyde:

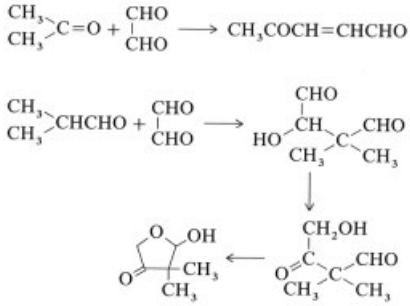

With amides and urethanes, di- and tetramides are obtained, depending on the pH [14], [15].

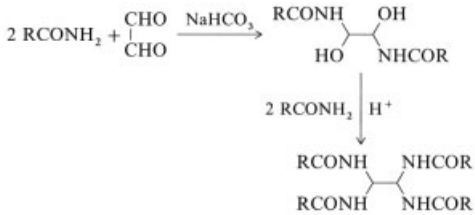

In alkaline medium, glyoxal disproportionates by an internal Cannizzaro reaction to the glycolate [16–18]:

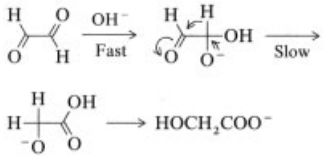

4. Production

Among the numerous processes for producing glyoxal, only those using *acetaldehyde* [*75-07-0*] and *ethylene glycol* [*107-21-1*] as starting materials have been developed commercially.

From Acetaldehyde. Oxidation with nitric acid was examined by LJUBOWIN as early as 1875 and patented in 1942 [19]. Reaction takes place at ca. 40 °C and is carried out industrially as a continuous process. Maximum yield is ca. 70 %; selectivity is a function of the relative concentrations of reagents. After the removal of ex-

cess acetaldehyde, the glyoxal formed, which is contaminated with acetic, formic, and glyoxylic acids, is purified by passage of the aqueous solution through an ion-exchange resin. The solution is then concentrated to a glyoxal content of about 40 %.

Selenium oxide [*7446-08-4*] is more selective than nitric acid, and the yield is ca. 84 %; the selenium can be recycled by oxidation with hydrogen peroxide [20]. This process has not been carried out on an industrial scale.

From Ethylene Glycol. The gas-phase oxidation of ethylene glycol by atmospheric oxygen in the presence of dehydrogenation catalysts (metallic copper or silver) represents the basis of the Laporte process [21] and has been used in several industrial production processes. Reaction occurs between 400 and 600 °C; the yield is 70 – 80 %. The main impurity is formaldehyde [*50-00-0*], whose subsequent separation is difficult. This reaction has also been carried out in the liquid phase and under irradiation.

Other Processes. Ethylene can be oxidized by aqueous nitric acid in the presence of palladium [22], by atmospheric oxygen, or by selenium oxide deposited on silica [23]. Glyoxal may also be formed by oxidation of acetylene [24] or benzene [25] with ozone. Ethylene oxide has been proposed as a substrate for oxidation. Although oxalic acid and its derivatives can be reduced to glyoxal, these processes have not been developed further.

5. Uses

Glyoxal is supplied mainly as a 40 % aqueous solution. Various polyhydroxy polymers (e.g., starch or poly(vinyl alcohol)) can be used to stabilize glyoxal.

The bifunctionality of glyoxal has been used to cross-link functionalized macromolecules such as cellulose, polyacrylamides, poly(vinyl alcohol), keratin, and other polycondensates. For example, glyoxal is used as a cross-linking agent for imparting wet strength to coated paper. With cellulose, unstable hemiacetals are obtained in the cold, which irreversibly form acetals when heated in the presence of acid catalysts [26].

The dual functionality and the ability of glyoxal to form heterocyclic compounds have been used in the production of resins for imparting crease resistance to textiles. The reaction proceeds through the intermediate formation of 4,5-dihydroxy-2-imidazolidinone (**4**) [*3720-97-6*] (DHEU) and its methylolated derivatives (DMDHEU).

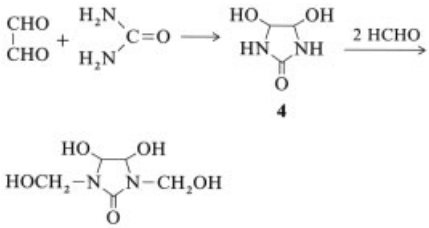

4

Ethers of DMDHEU are among the most effective resins for the crease-resistant treatment of cotton-based fabrics [27], [28].

The reducing properties of glyoxal are used in the photographic industry and in glassmaking for the production of silvered glass mirrors.

Glyoxal bisulfite (available as the monohydrate) is used as a resist agent in printing with reactive dyes and as a leveling agent in dyeing polyamide with acid dyes.

An important class of molecules used as cross-linking agents is obtained by nucleophilic substitution of glyoxal. Examples include the bisacrylamide (**5**) [*76843-24-8*] and the tetraallylacetal (**6**) [*16646-44-9*].

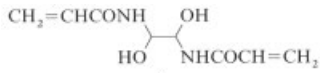

5

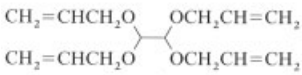

6

Glyoxalbisacrylamide is used for the functionalization of ion-exchange resins and in latices for treating textiles. Tetraallyloxyethane is used as a cross-linking agent for polyacrylates (→Polyacrylates) and for pressure-sensitive adhesives.

Glyoxal is used in the fine chemicals industry for the production of various heterocyclic compounds including tetraacetylgly-

coluril [*10543-60-9*] (**7**) (a perborate accelerator), imidazoles such as metronidazole [*443-48-1*] (**8**) that are effective against anaerobic bacteria, and the pyrazine derivatives sulfapyrazine [*116-44-9*] (**9**), thionazine [*297-97-2*] (**10**), and pyrazinamide [*98-96-4*] (**11**).

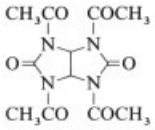

7

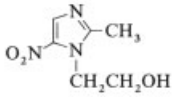

8

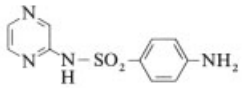

9

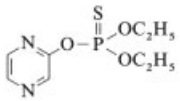

10

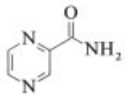

11

Glyoxal has bactericidal properties comparable with those of glutaraldehyde and is used as a disinfectant.

6. Toxicology and Occupational Health

Glyoxal has low toxicity. The LD_{50} is 3300 mg/kg (oral, rat). No mortality was observed in rats at a dosage of 1.3 mg/L inhaled in a saturated atmosphere or with aerosols. Glyoxal was well tolerated topically in a rabbit skin test but was irritating to the eyes. Glyoxal causes an allergic sensitization in guinea pigs.

Glyoxal gives a positive result in the Ames mutagenesis test, but it does not change cells in the cell transformation test and has no elastogenic effect in the micronucleus test; it is also not mutagenic in the Chinese hamster lung test. All of these tests are generally positive with substances known to be mutagenic.

7. References

1. H. Debus, *Ann. Chem. Pharm.* **100** (1856) 5; **102** (1857) 20 – 24.
2. U. Pincelli, B. Cadioli, D. J. David, *J. Mol. Struct.* **12** (1971) 171.
3. E. B. Whipple, *J. Am. Chem. Soc.* **92** (1970) 7183 – 7186.
4. F. Chastrette, C. Bracoud, M. Chastrette, G. Mattioda, Y. Christidis, *Bull. Soc. Chim. Fr.* 1983 33 – 40.
5. J. M. Kliegman, R. K. Barnes, *J. Org. Chem.* **38** (1973) 556 – 560; *J. Org. Chem.* **39** (1974) 1772 – 1776.
6. F. Chastrette, M. Hassamblay, M. Chastrette, *Bull. Soc. Chim. Fr.* **1976,** 601 – 606;607 – 612; 613 – 615.
7. Hoechst, DE 2 938 710, 1979 (H. Diery, W. Wagemann).
8. A. Le Rouzic, D. Raphalen, D. Papillon, M. Kerfanto, *Tetrahedron Lett.* **26** (1985) 1853 – 1856;Société Française Hoechst, FR 2 561 645, 1984 (A. Blanc); FR 2 561 648, 1984 (A. Blanc).
9. K. Maurer, E. H. Woltersdord, *Z. Physiol. Chem.* **254** (1938) 18 – 24.
10. P. Ferruti, A. Feré, L. Zetta, A. Bettelli, *J. Chem. Soc. C.* 1971, 2984 – 2985.
11. BASF, DE 932 369, 1955 (H. Spänig, G. Triem).
12. Nobel Bozel, FR 1 326 605, 1962 (L. Gandon).
13. BASF, DE 2 003 600, 1970 (F. Merger).
14. S. L. Vail, C. M. Moran, R. H. Barker, *J. Org. Chem.* **30** (1965) 1195 – 1199.
15. S. L. Vail, A. G. Pierce, *J. Org. Chem.* **37** (1972) 391 – 393.
16. P. Salomaa, *Acta. Chem. Scand.* **10** (1956) 311 – 319.
17. C. L. Arcus, B. A. Jackson, *Chem. Ind. (London)* 1964, 2022 – 2023.
18. A. R. Fratzke, P. J. Reilly, *Int. J. Chem. Kinet.* **18** (1986) no. 7, 757 – 773.
19. I. G. Farbenindustrie AG, BF 885 931 (1942).
20. Air Liquide, FR 2 038 575, 1969 (J. P. Zumbrunn).
21. Laporte Chemicals, GB 1 272 592, 1963 (B. K. Howe, F. R. Hary, D. A. Clarke).
22. BASF, DE 1 166 173, 1962 (R. Platz, W. Fuchs); DE 1 231 230, 1964 (R. Platz, H. Nohe).
23. E. Costa Novella, *Ann. Quim.* **68** (1972) no. 3, 325 – 332;Chem. Abstr., **77**, 113 760 f.
24. Imperial Chemical Industries, GB 1 071 902, 1965 (R. A. Rennie).
25. Inmont Corp., US 3 637 860, 1968 (W. P. Keaveney, J. J. Pappas).
26. F. S. H. Head, *J. Text. Inst.* **49** (1958) T 345 – T 356.
27. H. Petersen, *Text. Res. J.* (1958) 156 – 176.
28. S. L. Vail, G. B. Verburg, A. H. Young, *Text. Res. J.* (1969) 86 – 93.

Glyoxylic Acid

Georges Mattioda, Société Française Hoechst, C.R.A., Stains, France

Yani Christidis, Société Française Hoechst, C.R.A., Stains, France

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 3
5. **Toxicology** 3
6. **References** 3

1. Introduction

Glyoxylic acid [*298-12-4*], oxoacetic acid, glyoxalic acid, OHC−COOH, M_r 74.04, is the simplest α-oxocarboxylic acid. It is found in plants and is involved in the metabolic cycle of animals. Glyoxylic acid was discovered by Debus, who established its formula in 1856 [1]. This formula was described more precisely as that of the monohydrate by Perkin in 1868 [2].

In 1957, Kronberg and Krebs established the basis of the metabolic cycle of glyoxylic acid [3], in which isocitrate is converted to malate anaerobically. This cycle acts as a relay in the tricarboxylic acid cycle under specific conditions of plant life, such as germination or ineffective photosynthesis [4], [5].

2. Physical Properties

Glyoxylic acid crystallizes as the monohydrate upon controlled concentration of its aqueous solution. When dry, the monohydrate has *mp* 52 – 53 °C. The dissociation constant in aqueous solution is 4.7×10^{-4}, and the specific heat is 1.80 kJ kg^{-1} K^{-1}. The density of a 50 % aqueous solution of the acid, the only commercial form, is 1.34 g/cm^3 at 20 °C; the refractive index (20 °C) is 1.416.

The use of NMR spectroscopy (^{1}H [6] and ^{13}C [7]) has confirmed the existence of glyoxylic acid in aqueous solution as the dihydroxy acid (**1**) together with a small proportion of the linear dimer (**2**).

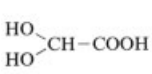

1

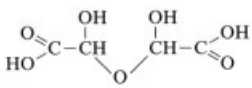

2

Glyoxylic acid is strongly hydrophilic; it is soluble in alcohols, with which it reacts, and in water-miscible solvents; it has a very low solubility in ether and other organic solvents.

Glyoxylic acid forms complexes with ions of alkali and alkaline-earth metals; it has a chelating power of 30 mg of calcium carbonate per gram of 50 % aqueous solution.

3. Chemical Properties

Glyoxylic acid contains two functional groups: the carbonyl group, which undergoes reactions characteristic of aldehydes (→Aldehydes, Aliphatic and Araliphatic, Chap. 2.2.), and the carboxylic acid group (→Carboxylic Acids, Aliphatic, Chap. 3.). The aldehyde group reacts readily with nucleophilic reagents; the hydrated aldehyde and the hemiacetal react similarly. With ambident nucleophiles, the carboxylic group may also react, leading to intramolecular ring formation. Various heterocyclic compounds can be obtained by coupling polynucleophiles with glyoxylic acid: *o*-phenylenediamine [*95-54-5*] gives 2-hydroxyquinoxaline [*1196-57-2*] (**3**); with urea [*57-13-6*] and an acid catalyst, allantoin [*97-59-6*] (**4**) is obtained in 60 % yield [8].

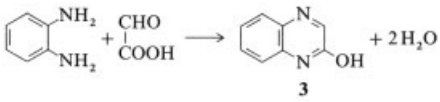

3

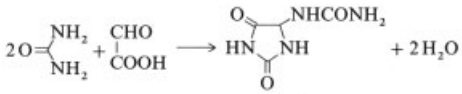

4

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_495

On heating, glyoxylic acid disproportionates to a mixture of glycolic acid [*79-14-1*] and oxalic acid [*144-62-7*]; glyoxylic acid is also easily oxidized to oxalic acid by nitric acid.

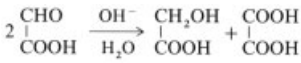

Reactions that have been used industrially include the Mannich reaction and amidoalkylation. Thus, with phenol and ethylenediamine in alkaline medium, glyoxylic acid leads to the sodium salt of *N,N′*-ethylenebis[2-(2-hydroxyphenyl)glycine] (EHPG, **5**), which forms complexes with iron (III) [9].

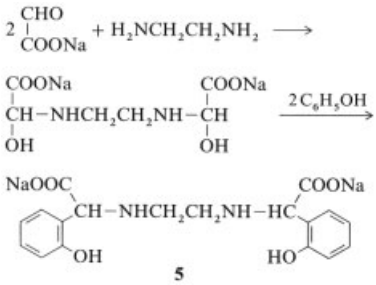

Glyoxylic acid, phenol, and ammonia react to give 4-hydroxyphenylglycine, an intermediate for the semisynthetic penicillin amoxicillin [10] (see →Antibiotics).

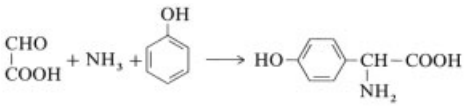

At slightly alkaline pH, glyoxylic acid reacts with amides such as acrylamide [*79-06-1*] to form acrylamidoglycolic acid (AGA), which is used as a copolymerizable cross-linking agent [11] (→Polyacrylamides and Poly(Acrylic Acids)).

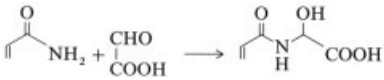

Several derivatives of AGA have found industrial application; e.g., the methyl ether ester is used in coatings and paint for automobiles.

Glyoxylic acid combines with phenol in alkaline medium to give 4-hydroxymandelic acid with good selectivity [12].

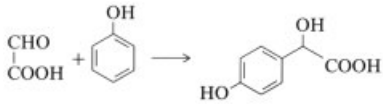

The reaction with phenols is used in several industrial syntheses of benzaldehydes. For example, guaiacol [*90-05-1*] is converted to vanillin [*121-33-5*] by oxidative decarboxylation of the corresponding mandelic acid [13].

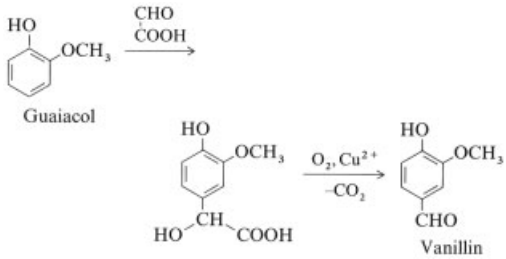

The mandelic acid intermediate may be modified further: reaction of phenol and glyoxylic acid in the presence of a reducing agent such as phosphorus – iodine leads directly to hydroxyphenylacetic acid [14], a building block for pharmaceuticals such as Atenolol.

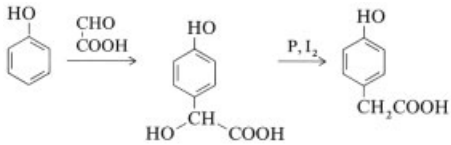

With thiophene, this reaction yields 2-thienylacetic acid, which is used to prepare the semisynthetic cephalosporins cephalothin and cefoxitin.

Glyoxylic acid and hydrochloric acid can be used to chloroalkylate aromatic compounds [15].

Glyoxylic acid undergoes the aldol reaction; the corresponding benzoylacrylic acids are obtained with acetophenones [16]:

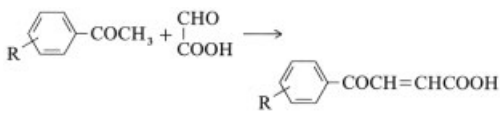

Glyoxylic acid is treated with Wittig ylides to form carboxyl functional polyenes of the vitamin A type [17].

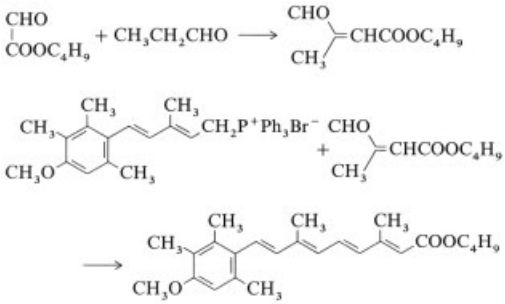

Esters of glyoxylic acid, obtained by dehydrative distillation of the hemiacetal esters in the presence of phosphorus pentoxide, can be polymerized under the action of a base. After saponification, the resulting polymers show useful chelating properties with the advantage of good biodegradability [18], [19]. Low molecular mass polyglyoxylates have been proposed as substitutes for sodium tripolyphosphates (TPP) in detergents to control water hardness.

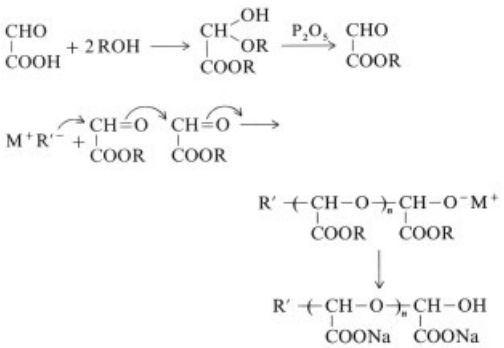

4. Production

Glyoxylic acid is produced industrially by oxidation of glyoxal in aqueous solution with 65 % nitric acid in mole ratios of 1 : 1 to 1 : 1.5 between 40 and 90 °C. The main byproduct of this process is oxalic acid, which is separated by low-temperature crystallization. The solution is then purified by passage through an anion-exchange resin or by electrodialysis, which removes the residual nitric acid.

Glyoxal can also be oxidized at the anode of a two-compartment electrolytic cell in the presence of chloride ion [20]. Glyoxylic acid may be synthesized by catalytic oxidation of ethylene or acetaldehyde, but the selectivity is low and these routes have not been used industrially.

The cathodic reduction of oxalic acid gives a very good chemical yield (85 %), but this technique encounters problems with passivation of the lead electrodes [21], [22].

Another method is oxidative cleavage of maleic acid or its esters by ozone. This process, has been adapted to the preparation of hemiacetal esters [23].

5. Toxicology

The LD_{50} is 2500 mg/kg (oral, rat). Glyoxylic acid is corrosive to the eye (rabbit), does not irritate the skin, but causes an allergic sensitization in guinea pigs. No mutagenic effect was found in the Ames mutagenicity test; glyoxylic acid is completely inactive in the so-called micronucleus mutagenicity test in mice.

Glyoxylic acid is a metabolite in mammalian biochemical pathways.

6. References

1. H. Debus, *Ann. Chem. Pharm.* **100** (1856) 1.
2. W. H. Perkin, B. F. Duffa, *Bull. Soc. Chim. Fr.* **10** (1868) 213.
3. H. L. Kornberg, H. A. Krebs, *Nature (London)* **179** (1957) 988 – 991.
4. H. Beevers, *Ann. N.Y. Acad. Sci.* **168** (1969) 313 – 324.
5. M. Cioni et al., *Comp. Biochem. Physiol.* **70 B** (1981) 1 – 26.
6. G. Ojelund, I. Wadso, *Acta Chem. Scand.* **21** (1967) no. 6, 1408 – 1414.
7. F. Chastrette, C. Bracoud, *Bull. Soc. Chim. Fr.* 1985 (II) 66 – 74.
8. BASF, DE-OS 1 939 924, 1969 (W. Mesch, O. A. Grosskinsky, N. Lösch).
9. Ciba-Geigy, US 4 130 582, 1977 (H. E. Petree, H. Myatt, A. M. Jelenevsky).
10. Beecham, GB 978 178, 1962 (J. H. C. Nayler, H. Smith).
11. Nobel Bozel, FR 1 411 715, 1964.
12. Soc. Française Hoechst, FR 2 440 350, 1978 (A. Schouteeten, Y. Christidis).
13. Haarmann & Reimer, DE-OS 2 115 551, 1971 (K. Bauer, W. Steuer).
14. Soc. Française Hoechst, FR 2 470 127, 1979 (J. C. Vallejos, Y. Christidis).

15. Soc. Française Hoechst, FR 2 376 112, 1976 (E. Herman, H. Diery, M. Soreau, Y. Christidis).
16. Roussel-Uclaf, FR 2 504 005, 1981 (Y. Christidis, R. Fournex, C. Tournemine).
17. Hoffmann La Roche, US 4 224 244, 1980 (W. Bollag, R. Ruegg, G. Ryser).
18. Monsanto, US 4 144 226, 1977 (M. M. Crutchfield, V. D. Papanu, C. B. Warren).
19. Procter & Gamble, US 4 284 524, 1980 (L. A. Gilbert).
20. Chlorine Engineers and Daicel Chem., FR 2 443 517, 1979 (H. Harada, K. Hirao, M. Ichino, T. Mitani).
21. I. V. Kudryashov et al., *Tr. Mosk. Khim. Tekhnol. Inst.* **49** (1965) 111 – 115; *Chem. Abstr.* **65** 3328 d.
22. K. Scott, *Chem. Eng. Res. Des.* **64** (1986) no. 4, 266 – 272.
23. Lentia, DE 3 224 795, 1982 (A. Sajtos, M. Wechsberg, E. Roithner).

Glyptals → Alkyd Resins

Gold, Gold Alloys, and Gold Compounds

HERMANN RENNER, Degussa AG (retired), Hanau, Federal Republic of Germany (Chaps. 1 – 3, Sections 4.1, 4.2, Chap. 11)

GÜNTHER SCHLAMP, Demetron GmbH (retired), Hanau, Federal Republic of Germany (Section 2.1, Chap. 8, Sections 10.3.1, 10.3.2, 10.4 – 10.10)

DIETER HOLLMANN, dmc2 Degussa Metals, Catalysts AG, Frankfurt, Federal Republic of Germany (Chaps. 5 – 7)

HANS MARTIN LÜSCHOW, Degussa-Hüls AG (retired), Hanau, Federal Republic of Germany (Chap. 9)

PETER TEWS, Allgem. Gold und Silberscheideanstalt (AGOSI), Pforzheim, Federal Republic of Germany (Sections 10.1, 10.2)

JOSEF ROTHAUT, Ducera Dental GmbH und Co. KG, Hanau, Federal Republic of Germany (Section 10.7)

KLAUS DERMANN, Ducera Dental GmbH und Co. KG, Hanau, Federal Republic of Germany (Section 10.7)

ALFONS KNÖDLER, Forschungsinstitut für Edelmetalle und Metallchemie (retired), Schwäbisch Gmünd, Federal Republic of Germany (Section 10.8.1)

CHRISTIAN HECHT, dmc2 Degussa Metals, Catalysts AG, Frankfurt, Federal Republic of Germany (Section 10.8.2)

MARTIN SCHLOTT, Leybold Materials, Hanau, Federal Republic of Germany (Section 10.8.3)

RALF DRIESELMANN, dmc2 Degussa Metals, Catalysts AG, Frankfurt, Federal Republic of Germany (Chap. 11)

CATRIN PETER, Klinikum der Friedrich-Schiller-Universität, Jena, Federal Republic of Germany (Chap. 12)

RAINER SCHIELE, Klinikum der Friedrich-Schiller-Universität, Jena, Federal Republic of Germany (Chap. 12)

1. History . . . 3
1.1. Centers of Gold Production . . . 3
1.2. Production . . . 5
1.3. Development of Production Processes . . . 5
2. Properties . . . 6
2.1. Physical Properties . . . 6
2.2. Chemical Properties . . . 7
3. Occurrence . . . 8
3.1. Abundance . . . 8
3.2. Gold Deposits . . . 8
3.3. Gold Reserves and Resources . . . 9
4. Production . . . 9
4.1. Ore Treatment . . . 9
4.2. Cyanidation . . . 10
4.3. Recovery of Gold with Carbon . . . 11
4.3.1. Adsorption of Gold by Carbon . . . 11
4.3.2. Carbon-in-Pulp Process . . . 12
4.3.3. Carbon-in-Leach Process . . . 14
5. Gold Refining . . . 14
5.1. Chemical Refining . . . 14
5.2. Miller Process . . . 15
5.3. Wohlwill Electrolysis . . . 16
5.4. Solvent Extraction . . . 17
6. Recovery of Gold from Secondary Materials . . . 17
6.1. Recovery from Gold Alloys . . . 17
6.2. Recovery from Sweeps . . . 18
6.3. Recovery from Surface-Coated Materials . . . 18
7. Gold Compounds . . . 19
7.1. Potassium Dicyanoaurate(I) . . . 19
7.2. Tetrachloroauric(III) Acid . . . 21
7.3. Sodium Disulfitoaurate(I) . . . 21
7.4. Miscellaneous Gold Compounds . . . 22
8. Gold Alloys . . . 22
8.1. Binary Alloys . . . 24
8.2. Ternary Alloys . . . 29
8.3. Higher Alloys . . . 31
8.4. Production and Processing . . . 31
9. Quality Specifications and Analysis 33
9.1. Quality Specifications . . . 33
9.2. Sampling . . . 33
9.3. Qualitative and Semiquantitative Analysis . . . 33
9.4. Quantitative Analysis . . . 34
9.5. Purity Analysis . . . 35
9.5.1. Direct Analysis of Metallic Gold . . . 35
9.5.2. Purity of Gold Solution . . . 36

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_499

9.6. Trace Analysis 36
10. Uses of Gold and Gold Alloys . . . 36
10.1. Coins, Medals, and Bars 36
10.2. Jewelry 37
10.3. Electronics and Electrical Engineering 39
10.3.1. Electronics 39
10.3.2. Electrical Engineering 42
10.4. Solders 42
10.5. Pen Nibs 43
10.6. Chemical Technology 43
10.7. Dental Materials 43
10.8. Coatings 43
10.8.1. Electroplating and Electroforming . 43
10.8.2. Bright Gold 44
10.8.3. Other Gold Coatings 46
10.9. Gold Leaf 46
10.10. Catalysts 46
11. Economic Aspects 47
12. Toxicology and Occupational Health 48
13. References 49

1. History [1–24]

1.1. Centers of Gold Production

Gold is the first element that humans recognized as a metal. Towards the end of the Middle Stone Age and the onset of the Neolithic Age (ca. 8000 B.C.), the world's climate changed greatly. Large areas became arid, necessitating the establishment of permanent settlements in river valleys such as the Euphrates, the Tigris, and the Nile. The earliest archeological findings that can be reliably dated were made in predynastic Egypt (ca. 4000 B.C.) and Mesopotamia, which the Sumerians settled in 3000 B.C. Outstanding finds (ca. 3000 B.C.) were made close to modern Varna, on the Bulgarian shores of the Black Sea. Gold was first mentioned in literature in the Indian Vedanta (before 1000 B.C.), the writings of Herodotus (484 – 425 B.C.), and the Old Testament (1000 B.C.).

The historical development of gold processing and important gold production centers are summarized in Figure 1. Egypt was the principal gold country in the pre-Christian era and maintained that status until ca. 1500 B.C. Gold production reached its peak about 1300 B.C. when the first legal foundations for the production of gold were laid, awarding the Pharaoh the absolute monopoly. In ca. 2700 B.C. gold rings were introduced as means of payment; the first gold coins appeared around 600 B.C.

The origin of the gold used by the Egyptians is unclear. Substantial portions appear to have come from Nubia in Upper Egypt (nub = gold), but considerable quantities were probably imported following the frequently mentioned expeditions to "Punt."

The special status enjoyed by gold in Egypt also exerted an influence in neighboring countries. The country of Ophir mentioned by Solomon as the origin of his gold may be identical with Punt, but India may also have been the supplier. The Egyptian gold trade expanded particularly under the impetus of the seafaring Phoenicians and Greeks.

Although Egypt was the principal gold country until about 1 B.C., gold was also found and utilized in other regions including India, Ireland, Bohemia, the Carpathian Mountains, Gaul, on the Iberian peninsula, and in the Caucasus.

Even in ancient times the ownership of gold shifted from one ruler to the other through conquests and the collection of tributes. Alexander the Great obtained possession of Indian gold as well as considerable portions of the Pharaohs' treasure. The Romans had little of the metal in their own regions but their military expeditions netted them major amounts in the form of booty; they also exploited the mineral wealth of the countries they had conquered, especially Spain, where up to 40 000 slaves were employed in mining. The state's accumulation of gold bars and coins was immense. Later, however, more and more gold was used in luxury goods and towards the end of the Roman empire a gold shortage was experienced.

In the region corresponding to modern-day Zimbabwe, organized gold production existed prior to 1000 A.D.

The advent of Christianity in the Middle Ages reduced the general striving for gold. Moreover, until the beginning of the Middle Ages no dominant political power existed for organizing large-scale gold production. In Europe, only the deposits in the Sudeten mountains, the Carpathians, and the Alps were of any significance. Out-

Year	Production center	Development	Period	
2000			Modern times	Technical revolution
	Indonesia (Exploration of Borneo)			
1990				
	Papua New Guinea (Ok Tedi)			
1985				
	Australia			
1980				
		Carbon-in-pulp process		
1970	Brazil	Free U. S. gold price		
		Microelectronics		
1960				
1940				
1920				
1900	Gold Coast Witwatersrand	Cyanidation		Industrial revolution
	Alaska California Australia	Wohlwill electrolysis Miller process		
1850				
1800				Scientic discoveries
1750	Brazil			
1700	Urals			
1600		Affination		
1400	Central Europe	Inquartation	Middle ages	Natural philosophy, alchemy
1200				
		Amalgamation		
1000				
	Zimbabwe*		Ancient times	Iron age
500				
A. D. 0	Iberian peninsula			
B. C. 500	Gaul	Minting of coins		
1000	S. E. Africa Egypt	Au/Ag separation by heating with NaCl		Bronze age
2000				
3000		Rings as means of payment		
4000		Soldering of jewelry	Prehistory	Stone age
	Balkan (Varna)			
6000				
8000	Fertile crescent			

Figure 1. Historical survey of important gold production centers and gold processing techniques

side Europe, gold was produced in India, Japan, and Siberia.

Following the discovery of America and the rapid conquest of the New World at the end of the fifteenth century, the Spaniards transferred considerable amounts of gold from the New World to Europe. Although the conquistadors found a highly developed mining industry in Central America, their efforts to increase gold production were largely unsuccessful; most of the finds consisted of silver. It was not until the discovery of deposits in Brazil that there was a noticeable

increase in world gold production. These deposits were exploited from 1725 to about 1800.

Since about 1750 gold has been mined on a major scale on the eastern slopes of the Ural mountains. In 1840, alluvial gold was discovered in Siberia. The Russian deposits were exploited by the Czars and the land owners, who had to pay their taxes in gold. Russia produced about one fifth of the world's gold production, a proportion maintained until the present day.

The discovery of gold in California in 1848 increased gold production greatly. The special laws issued in the Western parts of the United States allowed private mining thanks to the right to stake claims. This type of working continued when gold deposits were found in Eastern Australia (1851), Nevada (1859), Colorado (1875), Alaska (1886), New Zealand and Western Australia (1892), and Western Canada (1896). However, these deposits soon lost much of their importance.

The strongest impetus was given to gold production through the discovery of the goldfields of the Witwatersrand in South Africa in 1885. This extremely rich deposit appeared to guarantee steady exploitation far into the future. South African gold soon occupied a commanding position in the world market. Production grew continuously except for a short interruption by the Boer War (1899 – 1902). In the 1970s, gold production largely stabilized in South Africa and in the rest of the world. In South Africa, several hundred thousand people are now employed in the production of gold.

The discovery of large deposits of gold in Brazil in the 1970s stimulated prospecting activities. New production centers have been established in the Sierra Pelada of Brazil, Canada, Venezuela, New Guinea (Ok Tedi), and especially Western Australia, causing a pronounced shift in the geographic distribution of world gold production.

Gold mining in Ghana (Gold Coast) only began to play a role, if a modest one, in the twentieth century, although the deposits were already known in the Middle Ages. Gold production in Zimbabwe and the neighboring eastern part of the South African bush veld today form a moderate but not insignificant part in the overall production of southern Africa.

1.2. Production

Total gold production in antiquity can only be assessed approximately. Up to the fall of the Roman Empire, production may have amounted to 10 000 t.

A total production figure of 2000 – 3000 t has been quoted for the Middle Ages. By the time of the discovery of America, annual world production had reached about 5 t.

An annual production of ca. 10 t was reached in ca. 1700, rising to ca. 15 t in 1800 and to 40 t by 1848, the year the Californian deposits were discovered. As early as 1852, more than 200 t/a were mined, but production subsequently decreased until 1890. Thereafter, the increased output of the South African mines raised annual world production to 500 t in 1904, 700 t in 1907, and 1000 t in 1936. Since then, world annual production, which was at times subject to major fluctuations, has risen to over 2000 t. A further 300 t/a result from industrial recycling.

The total world production of gold in 1991 exceeds $1.09\ 10^5$ t [187]. More than a third of the present-day gold inventory is held by the central banks of the Western industrialized nations as currency backing. An even larger proportion is in private hands, much of it in the form of jewelry. Krugerrands alone accounted for 2000 t in 1967 – 1985, and smaller quantities are circulating in industry.

The shares of individual states in the overall production until now are divided among the Republic of South Africa (40 %), the United States (15 %), the ancient empires (ca. 10 %), Russia/CIS (ca. 10 %), Australia and the Pacific Region (ca. 10 %), and Canada (ca. 5 %).

1.3. Development of Production Processes [25–27]

The development of gold production processes is summarized in Figure 1. In antiquity, gold grains obtained by washing river sands were cold-worked into the desired objects. From ca. 3900 B.C., alluvial gold could be fused into larger lumps. The ancient Egyptians were the first to quarry gold-bearing rocks. Comminution of the rocks and washing were often preceded by heat treatment.

Analyses of archeological finds show that, in Egypt, the separation of gold from silver and copper was feasible, possibly as early as 2000 B.C. Silver was removed by annealing with common salt to give silver chloride [8]. The naturally occurring gold – silver alloy, electrum magicum, was separated into its constituents. The slagging of copper by adding lead followed by cupellation was also known.

In Spain, the Romans developed the technique of flush mining, by which vast masses of rock were dropped from a height, comminuted, and moved by currents of water. Amalgamation presumably originated at that time, although it was first mentioned in the literature in the 11th century A.D.

In the Middle Ages, fusion with lead and cupellation were developed further. Water-powered crushing machines were also introduced for pulverizing ore, and miners learned to process arsenic-containing gold ores by roasting.

The alchemists endeavored to manufacture gold by transmutation of base metals. It was not until the end of the 18th century that the entire concept was finally rejected as false. However, these endeavors led to a better understanding of chemical processes and to the birth of the true natural sciences.

The 17th century saw the discovery of separation by inquartation, i.e., the separation of gold and silver by nitric acid, and of affination, i.e., separation by sulfuric acid. With advancing industrialization in the 19th century new methods replaced the old, but some of the latter have retained some importance to this day. Production of gold as a byproduct of other metallurgical processes, (e.g., the refining of copper, zinc, and lead), played an increasingly important role in Germany (Rammelsberg/Harz). In 1863, Plattner's method of chlorination was introduced in the United States and shortly afterwards in Australia. In 1867, MILLER succeeded in refining gold with chlorine. Refining by electrolysis according to WOHLWILL was introduced in 1878 and is still used for all fine gold of 9995 and 9999 purity. Since 1888, cyanide leaching has permitted economical beneficiation of the Witwatersrand ores, which were less amenable to other methods due to the extremely fine distribution of the gold. Since 1970, conventional cyanide leaching has been superseded by the carbon-in-pulp process, which dispenses with filtration of the leached rock powder. More recently, the ecological problems caused by cyanide leaching have been overcome by treating the cyanide in the wastewater with hydrogen peroxide.

The recent expansion of gold production is also due to the mechanization of ore transportation and beneficiation. In South America, however, manual mining of ores is once again being resorted to. This politically motivated measure, taken with a view to creating jobs, made employment possible for 500 000 people.

Solvent extraction is being investigated as a relatively new method for the rapid and effective refining of fine gold.

2. Properties [1], [20], [22], [28–37]

The distinction between noble and base metals is in many ways arbitrary, and is generally determined by practical considerations and tradition. Gold is the classic noble metal and complies with all the criteria for this group of elements: resistance to air, humidity, and to normal wear. Gold is also remarkable among the metals in that it occurs in nature almost exclusively in its elementary state.

2.1. Physical Properties

Gold, atomic number 79, atomic mass 196.96654, has only one naturally occurring isotope, ^{197}Au. Its most important radioisotope, which is used in medicine, is ^{195}Au; it emits ε and γ rays and has a half-life of 183 d. The electronic configuration of gold is $[Xe]\,4f^{14}\,5d^{10}\,6s^{1}$. Its atomic radius is 0.1442 nm. The ionic radius for coordination number 6 is 0.1379 nm for Au^{+}, and 0.085 nm for Au^{3+}.

Some physical properties of gold are as follows:

mp	1064.43 °C
bp	2808 °C
Density at 20 °C	19.32 g/cm^3
at 900 °C	18.32 g/cm^3
at 1000 °C	18.32 g/cm^3
at 1065 °C	17.32 g/cm^3
at 1200 °C	17.12 g/cm^3
at 1300 °C	17.00 g/cm^3

Vapor pressure at 1064 °C	0.002 Pa
at 1319 °C	0.1 Pa
at 1616 °C	10 Pa
at 1810 °C	100 Pa
at 2360 °C	10 000 Pa
Atomic volume at 20 °C	10.21 cm^3/mol
Electrical resistivity at 0 °C	2.06×10^{-6} Ω cm
Thermal conductivity at 0 °C	3.14 W cm^{-1} K^{-1}
Specific heat	0.138 J g^{-1} K^{-1}
Enthalpy of fusion	12.77 kJ/mol
Enthalpy of vaporization	324.4 kJ/mol
Tensile strength	127.5 N/mm^2

The melting point of gold has been a fixed point on the temperature scale since 1968.

The unit cell of gold is face-centered cubic (type A1), with a lattice constant (a_0) of 0.40781 nm. Gold as it occurs in nature usually does not have a very crystalline appearance. It exhibits threadlike, leaf-shaped, and spherical forms, on which cubic, octahedral, and dodecahedral surfaces can sometimes be seen. When large amounts of molten gold solidify, a characteristic pattern of concentric rings appears on the surface.

Pure gold that has not been mechanically pretreated is very soft. Its hardness on the Mohs' scale is 2.5, and its Brinell hardness is 18 HB. Gold is the most ductile of all metals. It can be cold drawn to give wires of less than 10 µm diameter, and beaten into gold foil with a thickness of 0.2 µm (see Section 8.4). Because of its softness, gold can be highly polished; this, together with its noble characteristics and brilliant color, gives it its yellow luster. The color of utility gold is less rich and varies considerably according to its alloy composition (see Chap. 8). Very thin gold foil is translucent; transmitted light appears blue-green.

The physical properties of gold and its alloys have been thoroughly investigated because of their significance for modern technology. For detailed information see [22], [28], [36], [37].

2.2. Chemical Properties [38]

Gold does not react with water, dry or humid air, oxygen (even at high temperature), ozone, nitrogen, hydrogen, fluorine, iodine, sulfur, and hydrogen sulfide under normal conditions.

Sulfuric acid, hydrochloric acid, hydrofluoric acid, phosphoric acid, halide-free nitric acid (except in very high concentrations), and practically all organic acids have no effect on gold, either in concentrated or dilute solutions and at temperatures up to the boiling point. If a hydrohalic acid is combined with an oxidizing agent, such as nitric acid, a halogen, hydrogen peroxide, or chromic acid, gold will dissolve. Gold can also be dissolved in a combination of water and a halogen (the Plattner process) and in selenic acid. Figure 2 shows the dissolution rates of gold in the most important industrial agents used for its dissolution.

Aqueous solutions of alkali metal hydroxides, alkali metal salts of the mineral acids, and alkali metal sulfides do not attack gold. However, gold dissolves in solutions of alkali metal cyanides in the presence of oxygen (see Fig. 2 and Section 4.2) or other oxidizing agents, such as cyanogen bromide (the Diehl process), 4-nitrobenzoic acid (see Fig. 2) and 3-nitrobenzenesulfonic acid, provided they do not rapidly destroy the cyanide. Gold is also attacked by sodium thiosulfate solutions in the presence of oxygen, and by alkali metal polysulfide solutions.

Fused caustic alkalis do not attack gold, provided air and other oxidizing agents are excluded. Gold reacts vigorously with alkali metal peroxides to form aurates. It is inert to the alkali metal phosphates and borates, and to the alkali metal salts of the mineral acids, which can therefore be used as slagging agents for removing metallic impurities from gold.

Gold reacts readily with dry chlorine. The maximum reactivity occurs at 250 °C, and the minimum at 475 °C. Above 475 °C the reactivity increases with increasing temperature up to and beyond the melting point.

Gold can be recovered from solution by electrolytic deposition or by chemical reduction. If the tetrachloroaurate(III) complex is present, then iron(II) salts, tin(II) salts, sulfur dioxide, hydrazine, hydrazonium salts, oxalic acid, or ascorbic acid can be used as reducing agents. The very stable dicyanoaurate complex requires stronger reducing agents such as zinc. Anion exchangers, which are used for the recovery of gold from solutions, sometimes reduce this complex to metallic gold. Similar results are achieved with activated carbon.

The standard potential of Au/Au^{3+} is +1.498 V, of Au/Au^{+} +1.68 V, and of Au^{+}/Au^{3+} +1.29 V.

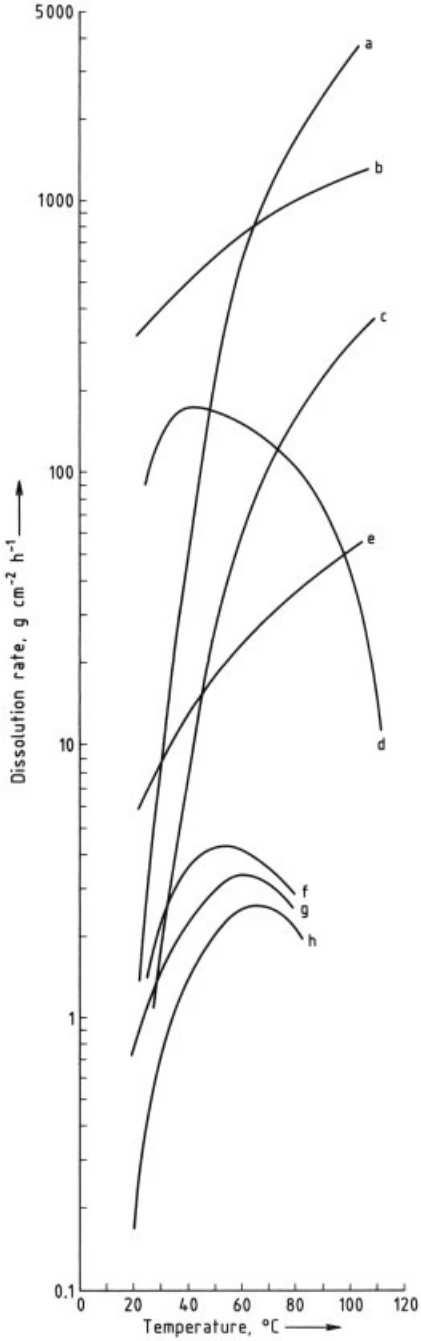

Figure 2. Rate of dissolution of fine gold sheet metal in various oxidizing agents
a) Aqua regia, 6 mol/L; b) HCl, 6 mol/L + Br_2, 0.2 mol/L; c) NaCN, 0.45 mol/L + 4-nitrobenzoic acid, 0.1 mol/L + NaOH, 0.2 mol/L; d) HCl, 6 mol/L + Cl_2 (saturated); e) HCl, 6 mol/L + H_2O_2, 0.22 mol/L; f) NaCN, 1 mol/L + air; g) NaCN, 0.45 mol/L + NaOH, 0.2 mol/L + air; h) NaCN, 0.006 mol/L + $Ca(OH)_2$, 0.004 mol/L + air

Gold can be alloyed with many other metals. In classic metallurgical processes (e.g., the lead blast furnace process and the reverberatory furnace process for copper ore), gold and silver follow the same route. Zinc, lead, and copper act as collecting agents for gold through the formation of alloys. Gold exhibits the greatest affinity for zinc, followed by lead, and then copper. Zinc is used to remove gold from molten lead in the Parkes process (→ Lead). The readiness with which gold takes up lead, tellurium, selenium, antimony, and bismuth is a disadvantage, particularly with regard to subsequent mechanical processing. Gold alloys readily with mercury at room temperature to form an amalgam. The mercury can be distilled out by heating. This property is utilized in the amalgamation process (see Section 4.1), and in fire gilding (see Section 10.8).

Colloidal gold forms hydrosols of an intense red or violet color, which are relatively resistant even without protective colloids.

3. Occurrence [1], [2], [6], [20], [40–45]

3.1. Abundance

Gold is distributed very unevenly in the Earth's crust, mainly due to enriching processes that have taken place near the surface. Its average abundance is very low and is estimated at ca. 0.005 ppm, although widely varying figures are given.

The gold content of ocean water also varies greatly, depending on the location. Gold contents of 0.008 – 4 mg/m^3 (ppb) have been reported.

3.2. Gold Deposits [7], [11], [44]

The gold deposits which are most easily recognized, and which were the earliest to be discovered, are enriched veins and deposits of gold particles. These particles were originally present in primary rock that was worn down by weathering. Enrichment then followed with partial consolidation due to flowing water. Such deposits are known as *placer* or *secondary deposits.* Typical examples are the relatively small

gold deposits in the Rhine Valley, California, and Alaska. The abundance of gold in placer deposits fluctuates greatly especially as the gold particles may be concentrated in very small areas, e.g., a stream bed. Under favorable conditions, placer deposits containing as little as 1 ppm gold may be successfully exploited.

Quartz veins containing gold are often found along the fault plane of rock fractures. As the gold particles have remained at their place of origin, these are termed *primary deposits.* In general, it can be assumed that this gold has been formed hydrothermally, i.e., it has been through an intermediate stage in aqueous solution. Such deposits are found in East Africa, Australia, Canada, and the Soviet Union. Their gold content varies greatly.

The Witwatersrand goldfields in the Republic of South Africa (Transvaal and Orange Free State) were also formed by sedimentation. These are sand and shingle deposits that have been compacted to form massive rock, in which the gold is distributed as very fine particles. This type of gold deposit is known as a *conglomerate deposit.* The average gold content of the ore when separated from the gangue is ca. 12 ppm. Mining reaches a depth of 4000 m where temperatures of 60 – 70 °C can prevail. Work is only possible here with cooling, which is performed with chilled air, cooled water, or ice [21]. To date about one-third of the total world gold production has been derived from the Witwatersrand conglomerate ore. The waste extracted sand which is found, for example, around Johannesburg, contains about 1500 t gold (0.5 ppm), mainly contained in pyrites (FeS_2) which is not dissolved by cyanide treatment. This material can also be processed economically.

Sulfidic copper ores may have gold inclusions which can become highly concentrated as a result of weathering. In the outer oxidation zone, hydrothermal reactions take place, such as

$$3\,Fe^{3+} + Au \xrightarrow{H_2O} Au^{3+} + 3\,Fe^{2+}$$

while in the underlying cementation zones, the corresponding back reaction occurs. Such deposits are found in Papua New Guinea (Ok Tedi) and in Brazil. The Ok Tedi deposit contains about 4 ppm of gold in the cementation zone.

Copper sulfide ores normally contain only a small proportion of gold (< 1 ppm); however, they can be a significant gold source. During smelting, gold accompanies silver and can be separated in the copper anode slimes. Practically all silver ores also contain some gold.

3.3. Gold Reserves and Resources

The term reserves denotes those resources whose existence has been established by prospecting and for which mining is economically viable.

World gold ore reserves are assessed at 75 000 t in 1991, or less than 40 times the world annual primary production [187]. In 1970, gold reserves were calculated to be one-fifth of this amount. At that time, the extensive Brazilian and southeast Asian deposits had not been discovered.

Of the reserves known today, 40 % are found in the Republic of South Africa, 35 % in Brazil, and 15 % in Russia/CIS. These are followed by the United States, Canada, Australia, China, Borneo, Papua New Guinea, Zimbabwe, and Ghana, with 1 – 5 % each.

4. Production [1], [2], [20], [27], [35], [39], [45–56]

4.1. Ore Treatment

In many places, gold is still mined by individuals and converted on the spot into marketable raw gold using simple manual and mechanical processes, such as panning (gravity separation), milling, and amalgamation. Amalgamation is carried out by allowing a slurry of ground gold-containing ore to flow over mercury-coated copper plates. The resulting gold amalgam is periodically removed by scraping. Very fine gold particles cannot be recovered by these methods, and in many cases, especially in Brazil, the use of cyanidation (see Sections 4.2 and 4.3) to extract the residual gold has been proposed.

Where gold is found in river sands covering a large area, the ore is often mined and processed in floating dredgers. This type of mining is found, for example, in Siberia, and in the north of the American continent.

In the conglomerate gold deposits in Witwatersrand, South Africa, most of the gold occurs as very fine particles. This means that mechanical enrichment and amalgamation are impossible, and the gold must be converted to a soluble form by reaction with sodium cyanide. For this purpose, the gold particles are first released from the rock material by means of breakers, wet ball mills, and classifiers. In newer plants, this milling process takes place underground.

Ground gold ore that contains large gold particles or sulfides may be unsuitable for cyanidation. Pretreatment, consisting of gravity concentration, generally followed by amalgamation, is therefore nearly always necessary; this also allows up to 50 % of the gold to be extracted faster than by the cyanidation process.

Gravity concentration

was formerly carried out using a cord cloth. The cloth was laid on a suitable support, and a water slurry of ground ore was passed over it, the grooves in the cloth being arranged at right angles to the direction of flow. The denser particles were retained in the grooves, while the lighter quartz particles flowed away with the water.

The cord cloth has now been replaced by corrugated rubber (thickness 10 mm, groove depth 3 mm, distance between grooves 6 mm). Modern mechanical equipment has endless belts (width 1.5 m, length 7.2 m), tilted at an angle of 11°. These advance at a speed of 0.4 m/min, against the direction of flow of the ore slurry. The concentrate is sprayed off with water and sent to the amalgamation plant. In place of endless belts, slowly rotating cylinders lined with corrugated rubber are sometimes used (length 3.6 m, diameter 0.9 m, inclination 3.75°).

Concentrates from gravity separation processes cannot be directly melted down into gold bars, because they contain considerable amounts of iron pyrites and metallic iron. Gold and silver are therefore generally separated from these components by *amalgamation.* The concentrate, which has a water content of about 70 %, is filled into a cast iron drum (length 0.9 m, diameter 0.6 m) containing steel balls (diameter 50 mm). The drum is rotated for 12 h, after which the gold particles are free from all impurities. Mercury is then added, and the drum is rotated for a further 2 h. The resulting amalgam is separated from the other components in a hydrocyclone (diameter 200 mm, inclination 20°); water and excess mercury are removed in a filter press. Remaining mercury is removed by distillation, leaving an impure mass of spongy gold, which is melted down into gold bars.

A *flotation process* is often used before gravity concentration in cases where the gold is closely associated with pyritiferous materials.

Roasting of ores in air is a secondary process which is sometimes used after gravity separation or flotation. The resulting oxides are then washed and treated by cyanidation. Gold ores containing sulfidic minerals can also be treated in a *bioleaching* process (see → Liquid – Solid Extraction, Chap. 3.3.3.), which dissolves the sulfides, exposing the gold particles for subsequent cyanidation.

4.2. Cyanidation [48]

The cyanidation process has been used in South Africa since 1890. In this process, the powdered mineral slurry, which contains ca. 10 ppm gold in the solid matter, is treated with an aerated 0.03 % sodium cyanide solution. Black cyanide ($Ca(CN)_2$ containing carbon and sodium chloride as impurities), a product of American Cyanimid, is often used instead of the more expensive sodium cyanide. The addition of calcium oxide ensures that the solution remains slightly alkaline. Dissolution takes place according to the following reaction:

$$4\,Au + 8\,NaCN + 2\,H_2O + O_2 \longrightarrow 4\,Na[Au(CN)_2] + 4\,NaOH$$

The dead powdered mineral is filtered off in large rotary vacuum filters. The filter cake contains less than a tenth of the original gold content of the ore. The filtrate is treated with zinc chips, which are preactivated in lead acetate solution, to precipitate the gold:

$$2\,Na[Au(CN)_2] + Zn \longrightarrow Na_2[Zn(CN)_4] + 2\,Au$$

The raw gold is treated with sulfuric acid to remove excess zinc, dried, and then roasted in air at 800 °C to oxidize lead, zinc, and iron. A flux, usually borax, is added, and the material is melted down to raw gold, with a gold content of 80 – 90 %.

An ecological problem is caused by the presence of sodium cyanide in the cake of dead rock

material and in the wastewater. However, when exposed to air and sunlight, the cyanide is converted to nontoxic cyanate, and subsequently carbonate.

4.3. Recovery of Gold with Carbon

[57–69]

The first mention of the ability of carbon to adsorb precious metals was made in 1847. In 1880, it was found that gold can be recovered from chlorinated leach liquors by wood charcoal. McArthur and the Forest brothers discovered that cyanide was a good lixiviant for gold in 1890 and, in 1894, charcoal was used to recover gold from cyanide solutions. The charcoal was prepared from wood and did not possess the high surface area and porosity of carbon today. As no elution procedure was known, the gold was recovered from the carbon by smelting. The use of carbon reached a high point of efficiency in Australia in 1917 when fine carbon was used to recover gold from pregnant cyanide solution, but, as the zinc cementation process advanced, so interest in the use of carbon dropped off.

In the 1940s, a carbon of higher activity and greater abrasion resistance was developed from fruit pips and, in 1952, an elution procedure involving the use of sodium hydroxide and cyanide (the caustic – cyanide procedure) was developed. In 1960, a plant using carbon was erected in Canada, and the first major carbon-in-pulp (CIP) plant to treat the fraction from which the coarse material has been removed (slimes) was built in the United States at Homestake in 1973 to treat 2200 t/d.

Major developments in CIP continued in South Africa, for treatment of the total cyanided pulp. By 1976, a small pilot plant was in operation and, by 1978, a plant processing 250 t/d was on line. The CIP process is now the preferred method worldwide for the recovery of gold from cyanided pulp. The only exception appears to be the Soviet Union, where the resin-in-pulp process is used. The CIP process is used for the treatment of a variety of feed materials ranging from run-of-mine ore to dump materials and roaster-bed products.

The advantages of the CIP process over zinc cementation are:

1) Capital costs are lower.
2) Operating costs are lower.
3) The ability of carbon to adsorb gold is not affected by any of the common constituents of leach liquors.
4) Carbon is added directly to the cyanided pulp, and therefore the need for the expensive filtration and clarification stage is avoided.
5) The losses of soluble gold are usually lower than in the zinc cementation process.
6) Ores that contain carbonaceous material can be processed without loss of gold to the carbonaceous fraction.
7) Materials that are difficult to filter or thicken can be treated successfully.

4.3.1. Adsorption of Gold by Carbon

Activated carbon has a porous structure (→ Carbon, Chap. 5.). The following theories have been proposed for the mechanism by which activated carbon loads gold cyanide:

1) complete reduction to metal
2) a chemical precipitation mechanism involving gold, carbon monoxide, and cyanide
3) physical adsorption of sodium dicyanoaurate(I)
4) adsorption of the dicyanoaurate(I) ion
5) ion-exchange adsorption of the dicyanoaurate(I) ion
6) adsorption of a neutral complex, whose nature is pH dependent
7) electrostatic interaction between the dicyanoaurate(I) ion and positively charged sites
8) a physisorption process
9) a two-step process, in which an ion pair is adsorbed onto carbon, and then reduced to an unidentified species.

The last-mentioned theory is now generally accepted.

The adsorption of gold cyanide onto activated carbon is reversible. Thus, an equilibrium exists between the gold in solution and the gold loaded on the carbon. Factors which affect the rate of gold adsorption and those which affect the equilibrium loading of gold are listed in Table 1.

Certain materials poison activated carbon for gold adsorption. Calcium carbonate can form in

the pores and is detrimental to adsorption, but is removed by acid-washing. Organic materials (e.g., machine lubricants, detergents, flotation reagents) also poison carbon to some extent, but are removed during reactivation. Lower adsorption efficiencies are attained when a pulp containing calcine, shale, or clay is used, since these finely divided minerals can block the pores. Copper can decrease efficiency of adsorption by competing with gold for adsorption sites, particularly at low concentrations of cyanide.

Table 1. Factors influencing the adsorption of gold by carbon

Factor	Effect of increasing the factor	
	on rate	on equilibrium loading of gold
pH	slight decrease	decrease
Ionic strength	slight decrease	increase
Free cyanide	slight decrease	decrease
Temperature	slight increase	decrease
Base metals	decrease	decrease
Carbon particle size	decrease	none
Mixing intensity	increase	none
Pulp density	decrease	none

4.3.2. Carbon-in-Pulp Process

Ores containing 0.25 – 100 g/t gold are processed in CIP plants at tonnages from 100 – 10^6 t per month. The density of the pulp varies from 1.3 to 1.45 g/cm^3, depending on its viscosity.

A schematic of a CIP plant is shown in Figure 3. The cyanided pulp is prescreened to remove coarse material that would otherwise move with the carbon granules, and might later block the screens in the CIP circuit. The pulp then flows through a series of six or eight flat-bottomed, cylindrical, agitated tanks. The residence time of the pulp in each stage is ca. 1 h. Reactivated carbon is added to the last stage, and is moved countercurrent to the flow of pulp. The carbon is held in each stage by interstage screens. The residence time of the carbon in each stage is 1 – 2 d. The carbon concentration in a CIP circuit is 15 – 30 g per liter of pulp. The barren pulp leaving the circuit is screened to remove fine carbon.

The loaded carbon, which contains 200 – 20 000 g of gold per ton, is removed periodically from the first stage and eluted with a caustic – cyanide solution. The carbon is washed with acid to remove calcium carbonate, and then reactivated at high temperature in a kiln. The gold in the eluate is generally recovered by electrowinning.

For reasons of economy, the CIP process is carried out underground in some mines.

Screening. The total pulp fraction has to be *prescreened* at a smaller aperture (0.6 mm, 28 mesh) than that of the interstage screens. Prescreening removes coarse material to avoid blocking of the screens further downstream. Two types of screen are used: (1) vibrating screens with woven wire or polyurethane mesh, and (2) linear moving-belt screens.

Wood chips constitute a small fraction of the incoming pulp, and cause problems further downstream when vibrating screens are used. Another prescreening device, such as a dummy tank, is then required. A dummy-tank is a closed circuit tank in which the pulp passes through submerged air-cleaned screens to remove wood chips.

Interstage screens are used to retain carbon while allowing the finer pulp to flow down through the train.

The two types of screen that are used are external and submerged. With *external screens,* the pulp is air lifted onto the screens (normally vibrating decks), and transferred to the next stage, while the carbon is returned to the stage from which it came. A number of *submerged screens* have been developed. Air-cleaned screens normally use square metal mesh, while mechanically cleaned screens use wedge-wire. Duties on these screens vary from 50 to 200 m^3 of pulp per square meter of screen per hour. The aperture of the screens is 0.8 mm (20 mesh). The size of the carbon particles used in CIP is 1 – 2 mm (8 to 16 mesh), although larger particles are used when problems are encountered with screening.

Abraded carbon resulting from breakage of the carbon in the CIP circuit is recovered from the barren pulp by a *final screen.* The units and apertures of this screen are similar to those used for prescreening.

Interstage movement of carbon is achieved using airlift pumps, vertical spindle pumps, centrifugal pumps, or peristaltic pumps.

Factors that affect the choice of pump are controllability, carbon wear, and operating costs. In most plants, carbon is transferred intermittently,

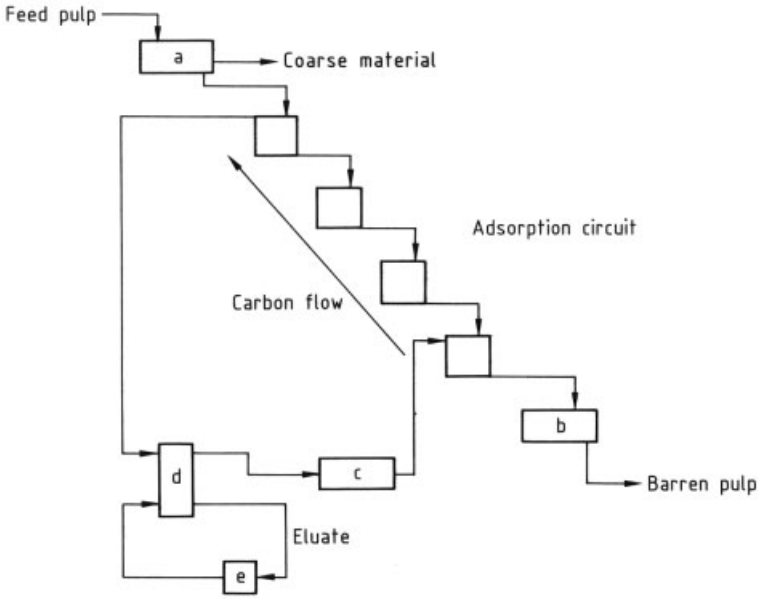

Figure 3. Schematic of the carbon-in-pulp circuit
a) Prescreening; b) Final screening; c) Regeneration; d) Elution; e) Electrowinning

but may also be transferred continuously. The following methods are used:

1) Loaded carbon is removed from stage one, the carbon in stage two is then transferred to stage one, and so on down the train.
2) Regenerated carbon is added to the last stage, the carbon in the last stage is then transferred to the second-last stage, and so on up the train.
3) All the pumps are turned on to transfer the carbon in each stage to the next stage up the train.

Agitation. The optimum agitation system provides a high pulp flow and low shear to avoid carbon breakage. Mechanical agitation is favored over air agitation because it consumes less power, provides better "live" vessel volume, and has acceptable carbon-attrition rates. Draft tubes and open-impeller agitators are used. Power inputs vary from 0.015 to 0.2 kW/m^3 of pulp.

Elution. The gold capacity of carbon decreases with increasing temperature. Increasing the concentration of cyanide and hydroxide ions also decreases the capacity of the carbon for gold.

These two facts were exploited by Zadra in 1952 to develop the first efficient elution technique to remove gold from carbon. Other elution procedures have since been developed.

The *Zadra process* involves the recycling of a solution containing 1 – 2 % sodium cyanide and 1 – 2 % sodium hydroxide through a bed of carbon. The temperature is over 85 °C, and the flowrate is about one bed volume per hour. The elution takes 48 h to complete. An electrowinning cell is used to recover the gold from the eluate. The advantages of this system are that the capital costs are low, the process is simple, and the reagent consumption is low. Its major disadvantage is that the rate of stripping is slow.

The *AARL process,* which was developed by the Anglo American Research Laboratories, has two stages. In the first stage, the carbon is presoaked under pressure at a temperature above 115 °C, in a solution containing 1 – 5 % sodium cyanide and 1 – 5 % sodium hydroxide for several hours. In the second stage, the carbon is eluted with five to seven bed volumes of water at 50 – 100 kPa and above 110 °C for 5 h at a flowrate of three bed volumes per hour. The advantages of this procedure are that elution is fast and the gold concentration in the eluate is high, which leads to efficient electrolysis. Its disad-

vantages include the need for high-quality water, more expensive equipment than that used by the Zadra process, and higher reagent consumption.

Elution with Organic Solvents. The addition of, for example, a 20 % solution of ethanol to a caustic – cyanide eluant increases the rate of elution, which takes ca. 6 h at atmospheric pressure and 80 °C. Disadvantages include higher reagent costs and handling problems associated with organic solvents.

The *high-pressure stripping process* is similar to the Zadra process, except that elution is carried out at up to 160 °C and at ca. 350 kPa. It has the advantages of lower reagent consumption, smaller elution inventory, and a smaller elution circuit. Disadvantages are more costly equipment, and that effluent solutions have to be cooled to avoid flashing.

Gold Recovery. The gold is recovered from the eluate either by zinc precipitation or by electrowinning. A number of electrowinning cells have been developed.

The *Zadra cylindrical cell* consists of a cylindrical core (polypropylene basket), which contains steel wool and acts as the cathode. The anode (stainless steel) surrounds the cathode in the cell, and the eluate is pumped into the center.

The most commonly used cell is the *Mintek cell.* It has a rectangular configuration with alternating anodes and cathodes in parallel, similar to those in copper-electrowinning cells. The cathodes are removable plastic baskets with perforated sides, packed with steel wool. The anodes are perforated stainless-steel plates. The flow of eluate is parallel to the flow of current.

Regeneration of carbon is carried out in two stages: (1) washing with acid to remove calcium carbonate and some base metals, and (2) thermal reactivation to remove organic materials and reexpose the pore structure.

The loaded carbon is washed with a 5 % solution of hydrochloric acid for 6 h at 75 °C in a column constructed of Hastelloy or rubber-lined mild steel. The carbon is then neutralized with sodium hydroxide. Acid-washing can be carried out before or after elution, but is generally done first because this increases the efficiency of elution.

In thermal reactivation, the degree of activation increases with increasing temperature and residence time, but micropores are destroyed if the conditions are too harsh. Reactivation is typically carried out at 700 °C for 10 min. Air should be excluded during regeneration. The steam generated from the carbon is normally sufficient to exclude air. The regenerated carbon is normally cooled by quenching in water.

Regeneration is usually carried out in a *rotary kiln,* which is heated externally by electrical resistive windings. In the *Rintoul furnace,* heating is achieved by passing an electric current through a bed of predried carbon in the presence of steam.

4.3.3. Carbon-in-Leach Process

Some ores contain carbonaceous material that is slightly active and can adsorb dissolved gold. These carbonaceous materials are termed preg-robbers. When a preg-robber is present in an ore, the carbon-in-leach (CIL) process is used, in which cyanidation and carbon adsorption occur in the same reactor.

The CIL circuit is similar to a CIP circuit, the main difference being the residence time of the pulp, since in CIL this is determined by the rate of gold dissolution and not by the rate of gold adsorption as in CIP. Hence, residence times of 4 h per stage are used. The presence of activated carbon prevents the preg-robber from adsorbing gold.

5. Gold Refining [1], [20], [27], [35], [39], [46], [47], [51], [53], [55]

5.1. Chemical Refining [2], [70]

The historical methods for separating gold from silver depended on dissolving silver and any accompanying base metals from solid solutions containing gold, using either nitric acid (inquartation) or sulfuric acid (affination) (Section 1.3). They were accepted up to the second half of the nineteenth century as the best way of manufacturing relatively pure gold. However, the exclusive use of these methods is insufficient today, and they must hence be combined with the other purification steps described below.

A method which allows recovery of relatively pure gold is to dissolve the raw gold in hydrochloric acid in the presence of an oxidizing agent. Silver chloride precipitates and can be removed, while gold is precipitated by reduction with oxalic acid or other reducing agents like $FeCl_2$, $FeSO_4$, SO_2, or Na_2SO_3. However, a single reductive precipitation process is usually insufficient, as part of any palladium present is entrained by the gold. If sulfur dioxide is used instead, the same applies to copper, nickel, zinc, and lead. Precipitation and separation normally have to be repeated several times before fine gold of sufficient purity is obtained. This method has therefore only found acceptance in cases where the commoner processes (see Sections 5.2 and 5.3) have not been established, either because the amounts involved are too small, or because the need arises only sporadically or in combination with other raffination processes like electrolysis. However, acid dissolution has several advantages: the metal is tied up for a short time only; a high yield of gold can be obtained in a single process step; and conventional equipment can be used.

5.2. Miller Process [71]

The Miller process has been of commercial significance since the 1870s. Probably more than two thirds of all gold produced to date has been through this process. It is extremely economical, and the gold quality obtained (99.5 – 99.6 %) meets the requirements of the gold trade. The largest plant in the world is the Rand Refinery in Gerniston, Republic of South Africa, which refines all the gold from South African mines, as well as that from neighboring countries. The Miller process is used not only by the mining companies of almost all gold-producing countries, but also in recycling processes.

The Miller process is based on the fact that gold chlorides are unstable above 400 °C, and thus do not form at a reaction temperature of ca. 1100 °C, whereas silver and base metals react with chlorine to form stable chlorides. At this temperature, AgCl (*bp* 1554 °C) and CuCl (*bp* 1490 °C) are molten and therefore collect in a slag layer, which is considerably less dense than molten gold. The chlorides of the other metals volatilize at the reaction temperature, for example, $PbCl_2$ (*bp* 954 °C), $FeCl_2$ (*bp* 1023 °C), $FeCl_3$ (*bp* 319 °C), $ZnCl_2$ (*bp* 732 °C), and $CuCl_2$ (*bp* 655 °C). As the process is effective when only a slight excess of chlorine is used, the amount of CuCl and $FeCl_2$ formed is greater than that of $CuCl_2$ and $FeCl_3$.

The raw gold delivered to the Rand refinery is weighed to an accuracy of 0.01 oz and then melted in an induction furnace. Four samples are taken and analyzed for gold content. The gold is cast into bars and reweighed. The gold content of the whole load is then determined. The bars are remelted in an induction furnace in batches of 500 kg, and chlorine gas is passed into the molten metal through quartz pipes, at a pressure of 0.2 kPa. First, the chlorine reacts selectively with iron, zinc, and lead (Fig. 4), whose chlorides are volatile at the reaction temperature and are thus removed as vapors. Finally, silver and copper chlorides form (Fig. 4). These are molten at the reaction temperature and thus form a slag which floats on the molten gold. Borax is added to the slag as a flux. The end point of the reaction is indicated by the appearance of red gold chloride vapor, and is confirmed by taking a sample for analysis. The silver content, which is determined by X-ray fluorescence analysis, must be below 0.4 %. After pouring off the slag, the gold is poured into tared molds.

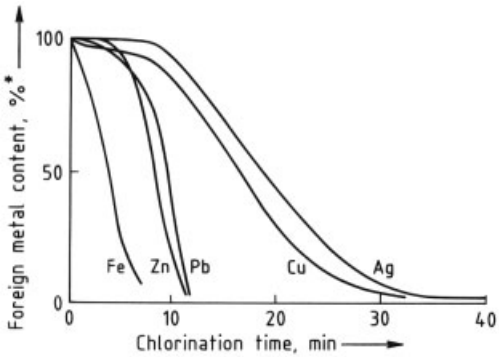

Figure 4. Decrease in the concentration of impurities in gold as a function of chlorination time
∗ Expressed as a percentage of the initial metal content. Typical starting concentrations, wt %: Ag 9.0, Cu 1.4, Pb 0.35, Fe 0.18, Zn 0.06

The chloride slag is ground and screened. Entrained gold is separated and returned to the process. Copper(I) chloride is oxidized to copper(II) chloride using sodium chlorate solution containing hydrochloric acid, and dissolved. The

solution is filtered and copper is cemented. Silver chloride is reduced with zinc in aqueous solution, and sent to a silver-refining electrolysis process.

Platinum-group metals cannot be separated by the Miller process (Fig. 5), as their chlorides, like those of gold, do not exist at the reaction temperature. However, this is not a disadvantage for gold from the Witwatersrand deposits, as this is free from platinum-group metals.

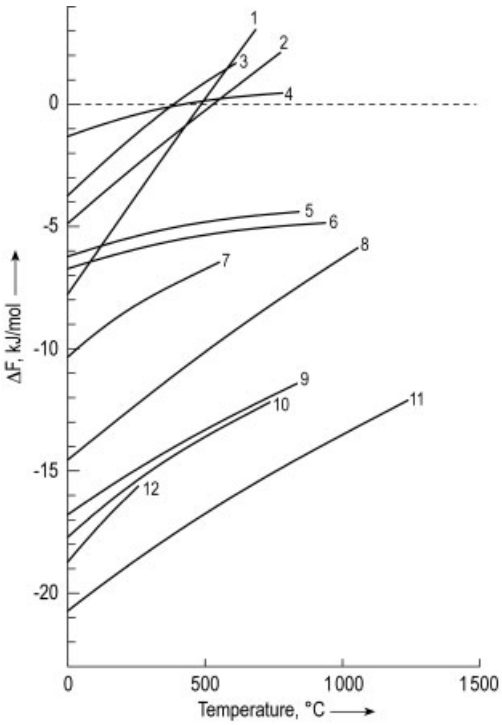

Figure 5. Thermodynamic of the Miller process [22, p. 39] (Courtesy Giesel Verlag).
1) $Pt + 2\,Cl_2 \longrightarrow PtCl_4$;
2) $Pt + 2\,Cl_2 \longrightarrow PtCl_2$;
3) $Au + 1\frac{1}{2}\,Cl_2 \longrightarrow AuCl_3$;
4) $Au + \frac{1}{2}\,Cl_2 \longrightarrow AuCl$;
5) $Ag + \frac{1}{2}\,Cl_2 \longrightarrow AgCl$;
6) $Cu + \frac{1}{2}\,Cl_2 \longrightarrow CuCl$;
7) $Cu + Cl_2 \longrightarrow CuCl_2$;
8) $Ni + Cl_2 \longrightarrow NiCl_2$;
9) $Fe + Cl_2 \longrightarrow FeCl_2$;
10) $Pb + Cl_2 \longrightarrow PbCl_2$;
11) $Zn + Cl_2 \longrightarrow ZnCl_2$;
12) $Fe + 1\frac{1}{2}\,Cl_2 \longrightarrow FeCl_3$

The great economic advantage of the Miller process, in addition to low production costs, lies in the fact that the gold leaves the refinery in a marketable form, thus minimizing the financial losses due to tying up of the metal. To meet requirements for higher purities, Wohlwill Electrolysis may follow in addition.

5.3. Wohlwill Electrolysis [27], [72], [73]

The Wohlwill electrolytic refining process for gold was developed in 1878 at the Norddeutsche Affinerie in Hamburg. It is still indispensable for the industrial production of fine gold of quality $\geq 99.9\,\%$ (in practice usually 99.95 % and 99.99 %).

The Wohlwill process uses an electrolyte containing 2.5 mol/L of hydrochloric acid and 2 mol/L of tetrachloroauric acid. Electrolysis is carried out with agitation at 65 – 75 °C. The raw gold is introduced as cast anode plates. The cathodes, on which the pure gold is deposited, were for many years made of fine gold of thickness 0.25 mm. These have now largely been replaced by sheet titanium or tantalum cathodes, from which the thick layer of fine gold can be peeled off. In a typical electrolysis cell (Fig. 6), gold anodes weighing 12 kg and having dimensions $280 \times 230 \times 12$ mm are used. Opposite them are conductively connected cathode plates, arranged two or three on a support rail. One cell normally contains five or six cathode units and four or five anodes. The maximum cell voltage is 1.5 V and the maximum anodic current density 1500 A/m^2. At the anode, the reaction

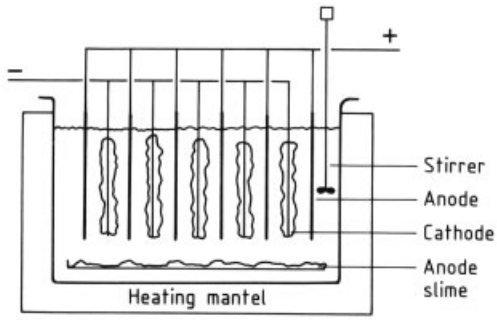

Figure 6. Wohlwill electrolysis cell

$$Au + 4\,Cl^- \longrightarrow [AuCl_4]^- + 3\,e^-$$

takes place, and at the cathode the reverse reaction. Some of the gold dissolves anodically as gold(I)

$$Au + 2\,Cl^- \longrightarrow [AuCl_2]^- + e^-$$

Anodes and cathodes are normally replaced every two days. About 10 % of the anode gold, especially parts located above the electrolyte, is remelted to form new anodes. The anode slime is collected in a trough in the bath. In addition to silver chloride, it contains rhodium, iridium, ruthenium, and osmium, which can be recovered, and approximately 10 % of the anode gold formed by the reaction

$$3\,AuCl + HCl \longrightarrow 2\,Au + H[AuCl_4]$$

Platinum and palladium can be recovered from the electrolyte, which also contains copper, iron, and nickel.

The accumulation of metallic impurities in the electrolyte, and of anode slime in the cell, means that the electrolyte can only be used for a limited period of time. Therefore, the material to be refined should have a high gold content (generally > 95 %) and it is often best to use gold which has already been through the Miller process. Especially the silver content should not be too high to avoid passivation of the anodes by formation of insoluble chloride films. The gold quality of the cathode deposits depends on the condition of the electrolyte, and on the amount of anode slime that has collected in the cell. A purity of 99.9 % is considered to be a minimum, which is the quality of Russian commercial gold. Today, 99.99 % can be considered as the norm, rather than 99.95 %. Sometimes anode bags are used to increase the purity of the gold.

The great advantage of the Wohlwill electrolysis process is the high purity of the gold produced. In addition, the byproducts, especially platinum-group metals, are also relatively easy to isolate. A disadvantage is the fairly long period during which the noble metal is tied up, leading to considerable financial losses.

World capacities for Wohlwill electrolysis are very large. In the Republic of South Africa, approximately a quarter of all gold from primary production, after it has gone through the Miller process, undergoes final purification to fine gold by Wohlwill electrolysis. In the Soviet Union, all gold is treated by electrolysis, as it normally contains platinum-group metals. Practically all noble-metal parting works which recover gold from secondary materials are dependent on electrolytic refining, again because platinum-group metals are normally present. Today, almost exclusively fine gold (99.99 %) is used, both for industrial purposes, and for investment; trading countries such as Switzerland therefore have large refining capacities, mainly producing gold of a quality between "good delivery" (99.5 %) and 99.99 %.

Chemical reactions (see Section 5.1) for the refining of gold have both advantages and disadvantages in comparison to electrochemical reactions. While Wohlwill electrolysis has low operating costs, it needs a high hold up of gold as electrolyte and residual anode. The chemical reactions can be used for different kinds of gold-containing input material and are carried out in multipurpose vessels. However, the demand for chemicals and the amount of off-gas produced are higher.

5.4. Solvent Extraction [74], [75]

Tetrachloroauric(III) acid can be extracted from aqueous solution by many organic solvents. Solvents which can form metal complexes (e.g., ethers and esters) are most effective. Total separation from the usual accompanying elements cannot, however, be achieved.

With the exception of dibutyl carbitol, no systems have yet been discovered which can be applied commercially. Yet a system with a high degree of separation efficiency and simple process management, including stripping the pure gold, could offer considerable advantages over currently used methods, even concerning the absence of a gold stock.

6. Recovery of Gold from Secondary Materials [1], [3], [20], [27], [35], [46], [47], [76–81]

6.1. Recovery from Gold Alloys

For the recovery of gold from alloys with a gold content of more than 30 %, the Miller process (see Section 5.2) is the most economical method. This also applies to the treatment of anode slimes from the electrolytic refining of silver (→ Silver, Silver Compounds and Silver Alloys), providing these have a high gold content and can be melted.

For small quantities of *gold-rich alloys,* a useful treatment is dissolution in hot aqua regia, or

in hydrochloric acid (usually at 20 – 50 °C) containing a halogen or hydrogen peroxide. Silver precipitates as sparingly soluble silver chloride. Relatively pure gold is recovered from the solution by selective reduction. If copper and lead ions are present, the best reducing agents are sulfur dioxide or hydrazine; if platinum-group metals are present, oxalic acid is used.

In the case of *gold – copper alloys,* which also may contain silver and nickel, treatment with aerated dilute sulfuric acid allows the separation of noble metals from base metals. Gold can only be separated from base metals in this way if the gold content of the alloy is less than ca. 65 %, and the material has been finely milled or powdered.

There are two historically important processes for treating *gold – silver* or *gold – silver – copper alloys*. The silver may be dissolved with nitric acid, preferably after realloying to a silver to gold ratio or a silver and copper to gold ratio of 3 : 1 (inquartation); or hot concentrated sulfuric acid can be used (affination). The latter process is rarely used today. Silver – copper alloys with a low gold content are not suited to the above process, but can be treated in noble-metal works by the lead cupellation process (see Section 6.2). Lead is removed as litharge, PbO, which also takes up copper as CuO, and other metal and metalloid oxides, leaving behind doré silver. This is subjected to electrolytic refining (→ Silver, Silver Compounds and Silver Alloys); gold collects in the anode slime.

6.2. Recovery from Sweeps

The noble-metal processing industry produces large amounts of waste dust and debris, known as sweeps (e.g., slag, ash, soot, precipitation and emulsion residues, and sweepings). The gold content of these materials is mostly in the range 0.5 – 10 %. Usually they are first carefully ground, screened, and, if necessary, dried or burned. The fine nonmetallic fraction and the coarse metallic fraction, which can usually be melted, are analyzed separately.

The usual method for treating gold-containing sweeps is to melt the material down, together with silver-containing sweeps, in a lead shaft furnace. Depending on the composition of the sweeps, calcium carbonate, silicic acid, lead oxide, carbon, and materials containing sulfur are added. The mixture of powdered components is normally briquetted, pelletized, or sintered before being fed into the shaft furnace. The metallic phase (alloy) produced in the furnace contains lead and almost all of the noble metal. This alloy is then converted to doré silver (a few percent Au; < 0.1 % Pb; the balance is Ag) in a cupel furnace (→ Silver, Silver Compounds and Silver Alloys) by oxidizing lead to lead(II) oxide. The silicate slag from the shaft furnace contains a negligible quantity of noble metal, and can usually be discarded. The slagging process may have to be repeated, for example if the noble-metal content exceeds 300 ppm. In the sulfidic phase, which forms between the metallic phase and the slag, almost all the copper collects with the nickel to form copper matte. The relatively high content of silver and valuable nonferrous heavy metals in the copper matte makes special treatment necessary; this is usually carried out at copper smelting works.

When raw gold is refined (see Chap. 5), the resulting residues contain more gold than silver. These substances are treated in a shaft furnace and then in a refining furnace, together with raw silver from the Miller process. Unlike other pyrometallurgical recycling methods for silver and gold, a copper matte phase is not formed in the shaft furnace. The Au : Ag ratio is normally 1 : 4.

6.3. Recovery from Surface-Coated Materials

Surface coatings containing gold come mainly from the electronics industry, plus a small proportion from the jewelry industry and electrical engineering. The base material is essentially metallic; the gold coatings are often only a few micrometers thick, but may account for a gold content between a few tenths percent and a few percent.

The most economic solution is usually to remove only the gold coating, and to send the underlying base material directly to a recycling process. Mechanical pretreatment is often required to expose the gold surface. Depending on the particular combination of materials and the degree and type of finishing employed, this can be carried out with a shredder, jaw crusher,

or edge runner. Frequently, this is followed by air separation, gravity separation, or magnetic separation to remove the ballast materials. It is normally not advisable to burn off plastic materials because the gold diffuses into the metallic base, and it is then impossible to remove it all. For this reason it is preferable to decompose the plastic material by pyrolysis.

To remove the gold from the metallic base, the material is usually agitated in alkaline cyanide solutions (10 – 20 g/L NaCN). Air is only rarely used as an oxidizing agent, because of the relatively low rate of dissolution. Instead, water-soluble aromatic nitro compounds (e.g., nitrobenzoic acid or nitrobenzenesulfonic acid) are normally employed. Potassium peroxodisulfate and hydrogen peroxide are also suitable, but they have a greater tendency to oxidize the cyanide ion. The base materials (e.g., iron, nickel, and cobalt) are attacked only to a limited extent during the dissolution process. When the base material is composed of copper, or a copper-rich alloy, zinc cyanide or lead salts are usually added as inhibitors to keep them from being attacked.

The gold is recovered from cyanide solution by adding zinc powder (see Section 4.3), after any excess oxidizing agent has been reduced with hydrazine or formaldehyde. The gold can also be recovered by electrolysis, using insoluble graphite or magnetite anodes (100 A/m^2 cathode, 3 – 4 V), or platinized titanium anodes. Excess oxidizing agent should also be reduced to obtain optimum current efficiency and speed of deposition. In addition to the reductive deposition of gold, thermal decomposition of the dicyanoaurate(I) complex is also possible.

Gold-plated metallic wastes can also be directly deplated by electrolysis, preferably in electroplating drums used as anodes. The gold is recovered in metallic form from the cathode in a single step. The method is, however, not very well suited to the recovery of gold from bulk material in anode baskets, because the electric field is screened off from the material at the center of the load, and as a result not all the gold is removed.

Gold can be extracted from solutions with a very low gold content (≤ 0.1 g/L) by anion-exchange resins (e.g., Lewatit M 500) [80] or activated charcoal [81]. However, elution is costly and usually incomplete; the gold-loaded carriers are therefore most frequently ashed, and the gold recovered from the ash. The solutions are not returned to the deplating process because they contain degradation products, especially from organic nitro compounds, which would interfere with the process.

The methods described above can also be used to treat other cyanide gold solutions, e.g., unusable electroplating baths.

Gold-plated materials are occasionally smelted together with copper ores in large copper-smelting works. In addition to silver, these ores always contain some gold. A disadvantage of this method is the long time the gold has to stay in the process. Gold collects in the anode slimes during electrolytic refining of the copper and can be recovered from these. All gold concentrates recovered using the processes described here still require refining (see Chap. 5).

7. Gold Compounds [1], [12], [20], [38], [45], [70], [82–86]

Almost all gold compounds occur in the oxidation states +1 and +3. The oxidation states +2 and +5 are also known. Bimetallic compounds of gold(II) are usually mixed valence compounds containing gold(I) and gold(III).

The generally very low stability of binary gold(I) and gold(III) compounds based on ionic bonding can be ascribed to the structure of the two outer electron shells. The ten electrons of the $5d$ level and the single electron of the $6s$ level differ only slightly in energy. Therefore, ionization of the $6s^1$ electron to leave a $5s^2p^6d^{10}$ valence shell, i.e., the reaction $Au \rightarrow Au^+$, is energetically not particularly favorable. Gold(III) compounds with the $5s^2p^6d^8$ valence shell exhibit a tendency to fill the $5d$ level. Thus, the electron-acceptor effect is strong, and the compounds are strong oxidizing agents. The simultaneous participation of both outer electron shells gives rise to properties characteristic of the transition metals, such as variable oxidation state, colored compounds, and the tendency to form complexes.

7.1. Potassium Dicyanoaurate(I)

Properties. Potassium dicyanoaurate(I) [*13967-50-5*], $K[Au(CN)_2]$, M_r 288.10, d

3.452 g/cm^3, forms colorless crystals which are readily soluble in water and alcohol, but insoluble in acetone and ether. The solubility in water is strongly dependent on the temperature (Fig. 7). The effect of potassium cyanide concentration on the solubility of potassium dicyanoaurate is shown in Figure 8. The complex is neither air nor light sensitive and is stable in aqueous solution above pH 3.

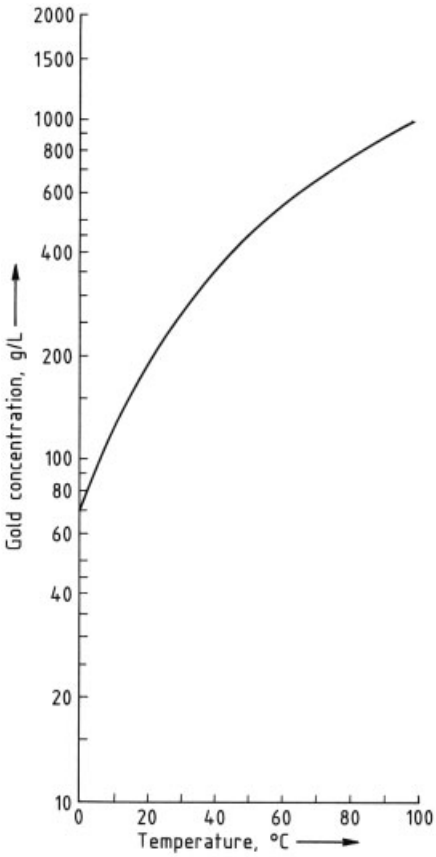

Figure 7. Solubility of potassium dicyanoaurate(I) in water as a function of temperature

Preparation [70], [87]**.** Potassium dicyanoaurate is prepared from fulminating gold, from gold(III) hydroxide, and by electrolysis. The third method is now the most common preparation from fulminating gold is potentially hazardous.

From Fulminating Gold. Fulminating gold is precipitated from a solution of tetrachloroauric(III) acid by addition of excess aqueous ammonia; the precise composition of the precipitate is not known. The chloride content of the precipitate decreases with increasing excess of ammonia.

Fulminating gold must not be allowed to dry, since it is highly explosive in this state. The fulminating gold must be thoroughly washed with deionized water to remove chloride, since chloride ions interfere in electroplating processes. After washing, the precipitate is dissolved in a small excess of potassium cyanide solution. This is followed by filtration, concentration, and crystallization. The mother liquor is repeatedly reused in the process.

Electrolysis. When a gold anode is dissolved in aqueous potassium cyanide, potassium dicyanoaurate(I) is formed with hydrogen being liberated at the cathode. To prevent the dicyanoaurate(I) ion from being electrolyzed at the cathode with deposition of gold, a diaphragm is used. Fluoropolymer-based ion-exchange membranes are used for this purpose. The mother liquor is reused or worked up following crystallization of the potassium dicyanoaurate(I). This can be carried out under particularly mild conditions by using the continuously-cooled crystallization process. Cyanide and chloride are the major impurities. Electrolysis produces potassium dicyanoaurate(I) with a particularly low chloride content, which depends solely on the quality of the potassium cyanide and of the water.

From Gold(III) Hydroxide. Instead of fulminating gold, gold(III) hydroxide may be reacted with potassium cyanide solution. Precipitation of gold(III) hydroxide from tetrachloroauric(III) acid by addition of an alkali metal hydroxide does not always go to completion. Aging of the gold(III) hydroxide considerably reduces its rate of dissolution in potassium cyanide solution.

Uses and Economic Aspects. Potassium dicyanoaurate(I) serves as the gold component in the baths commonly used for electroplating and to a lesser extent in electroless plating (→ Electrochemical and Chemical Deposition, Chap. 8.9.). Plating baths with higher acidity are based on tetracyanoaurate (III) [88].

With increasing utilization of gold-plated electronic components, the demand for gold plating baths, and hence for potassium dicyanoaurate(I), has grown considerably. In 1997

ca. 6 % of the 3890 t of gold processed annually in the world was used by the electronics industry. In the jewelry industry, too, that demand has increased as a result of rolled gold being superseded by electroplating with hard gold.

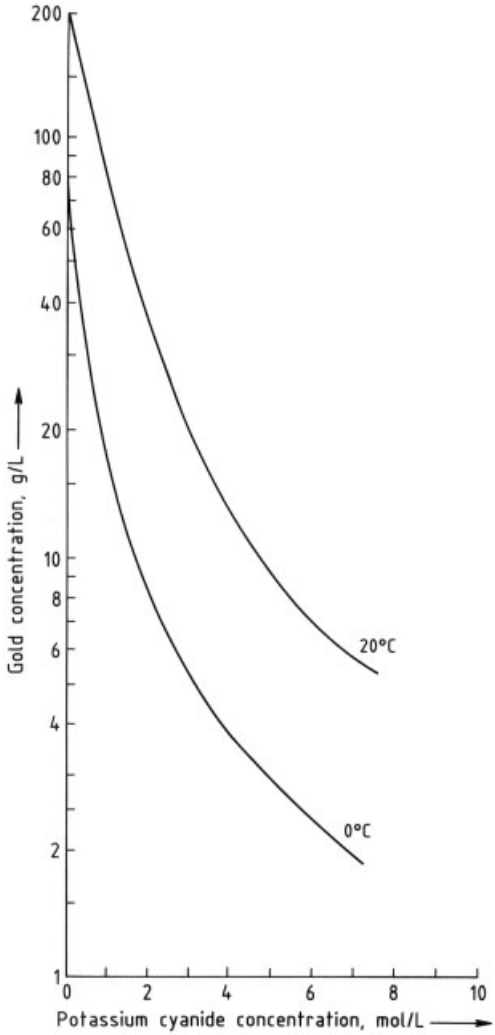

Figure 8. Dependence of the solubility of potassium dicyanoaurate(I) in water upon the concentration of potassium cyanide at 20 °C and at 0 °C

7.2. Tetrachloroauric(III) Acid

Properties. Tetrachloroauric(III) acid [*16903-35-8*], $H[AuCl_4]$, M_r 411.85, crystallizes as a tetrahydrate in the form of light yellow, deliquescent crystals. It is readily soluble in water, soluble in alcohol and ether, and is corrosive.

Preparation [70]**.** Tetrachloroauric acid is prepared by dissolving gold in warm aqua regia. In order to remove residual nitric acid, the solution is concentrated by evaporation with repeated addition of hydrochloric acid, and the resulting melt is poured into porcelain dishes. Moisture must be excluded while the melt cools, solidifies, and is finally powdered. Crystallization from the aqueous mother liquor produces an extremely hygroscopic material. Instead of nitric acid, other oxidants (e.g., chlorine) may be added to the hydrochloric acid. The high rate of dissolution of aqua regia is only rarely attained (see Fig. 2), but the pollution caused by it is lower and control of the reaction is simpler, especially under industrial conditions. Care must be taken to assure removal of any excess oxidants and their reaction products. Tetrachloroauric(III) acid can also be prepared by anodic dissolution of gold in hydrochloric acid, but this requires the use of a diaphragm.

Uses [1]**.** Tetrachloroauric(III) acid is used to prepare other gold compounds. It is also used to make gold ruby glass, gold purple (Purple of Cassius), and purple colorants for enameling of ceramics (→ Ceramic Colorants) and to prepare the electrolyte for the Wohlwill electrolysis.

7.3. Sodium Disulfitoaurate(I)

Properties. Sodium disulfitoaurate(I) [*39394-92-8*], $Na_3Au(SO_3)_2$ M_r 426.06, is relatively unstable in the solid state, and is therefore not isolated as crystals for industrial use. It is stable in weakly alkaline solutions above pH 8.5, even upon heating and when exposed to light.

Preparation. Sodium disulfitoaurate(I) is prepared from fulminating gold or from gold(III) hydroxide.

From Fulminating Gold. Fulminating gold (see Section 7.1) is dissolved with stirring in a dilute solution of sodium hydrogen carbonate; sodium sulfite is then added. The reaction mixture is stirred at 70 °C until it clears. The pH must be maintained above 9. This is achieved by addition of sodium hydrogen carbonate solution. In order to remove ammonia, the solution is heated to about 90 °C with the simultaneous introduction of air. The gold content of the solution is kept at ca. 100 g/L by evaporation or by dilution with deionized water. This concentration is desirable for electroplating purposes. Instead of sodium sulfite, gaseous sulfur dioxide and a correspondingly higher quantity of sodium hydroxide solution may be used; the pH is controlled as described above.

From Gold(III) Hydroxide. Instead of fulminating gold, gold(III) hydroxide may be used. The reaction takes place under approximately the same conditions as with fulminating gold. The disadvantages are the same as in the preparation of potassium dicyanoaurate(I) from gold(III) hydroxide (see Section 7.1).

Uses. Sodium disulfitoaurate(I) is sometimes used in preference to potassium dicyanoaurate(I) in electroplating baths. These baths are especially advantageous for the production of ductile and wear-resistant coatings and in white-gold and rose-gold electroplating. Their disadvantages are difficult handling, shorter service life, and higher price. Their low toxicity is counterbalanced by problems in waste water disposal, e.g., those caused by the ethylenediaminetetraacetic acid (EDTA) complexes of the alloying elements.

Sodium disulfitoaurate(I) is increasingly being replaced by the corresponding ammonium salt, which provides special advantages with Au – Pd – Cu baths, such as the facility to produce very dense coatings.

7.4. Miscellaneous Gold Compounds [1], [20], [70], [89]

Bright gold preparations (gold sulforesinates) are used for gilding ceramics and glass (→ Ceramic Colorants). Other compounds, some of which have applications in preparative and analytical chemistry, include gold(III) chloride $AuCl_3$, gold(I) chloride AuCl, gold(III) hydroxide $Au(OH)_3$ or AuO(OH), gold(III) oxide Au_2O_3, the gold sulfides Au_2S, AuS, and Au_2S_3, gold(I) cyanide AuCN (→ Cyano compounds, Inorganic, Chap. 2.3.4.), and gold(I) acetylide Au_2C_2. Sodium dithiosulfatoaurate(I), $Na_3[Au(S_2O_3)_2] \cdot 2\,H_2O$, is used as a photographic sensitizer, and gold(III) selenate, $Au_2(SeO_4)$, for staining glass.

([90], [91]): Gold compounds, particularly gold thiosulfate, gold mercaptides, and gold malate, have long been used for the treatment of rheumatic disease (→ Anti-inflammatory – Antirheumatic Drugs, Chap. 4.1.). The triethylphosphine gold complex of thioglucose tetraacetate, Auranofin [*34031-32-8*], is very effective against chronic inflammation of joints. Gold(I) complexes of diphosphines may be beneficial in tumor therapy. Gold(I) and gold(II) complexes have antimicrobial activity.

8. Gold Alloys [1], [12], [20], [28], [35], [45], [92–103]

Because of its mechanical softness, gold is, like most other noble metals, used in elemental, unalloyed form only in special functions in electronics, sensors, optical devices, and catalysts. For the majority of applications, gold must be alloyed to adjust its mechanical (strengthening), thermal, physical (electrical, color), and structural (grain refining) properties according to special requirements.

The practical uses of gold alloys and of fine gold are mainly determined by their color, hardness, corrosion resistance, melting point, and relative value. The degree of hardness can be controlled by heat treatment. The melting point plays a significant role during manufacture and processing.

Structure Types. Gold forms numerous alloys and intermetallic compounds with many metals and metalloids. The alloys have in most cases face-centered cubic (fcc), hexagonal close packed (hcp), or body centered cubic (bcc) crystal structures, but there are also a great variety of different structure types, especially for the intermetallic compounds. Gold forms continuous solid solutions with the fcc-crystallizing

metals silver (atomic radius 144.5 pm), copper (127.8 pm), platinum (137.3 pm), palladium (137.6 pm), iron (124.1 pm), cobalt (125.3 pm), and nickel (124.6 pm). The solid solutions are generally of the substitutional type, where the solute atoms replace atoms of the solvent atoms in their lattice positions. Substitutional solid solutions are only formed if the components have the same or similar crystal structures and if the atomic radii of solute and solvent do not differ by more than ca. 15 % ("size factor" after HUME-ROTHERY [104]).

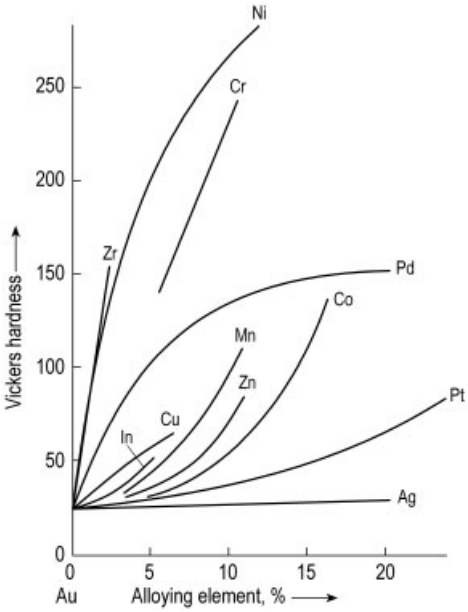

Figure 9. Solid solution hardening of gold [110]

Some systems show miscibility gaps, which in combination with temperature-dependent decreasing mutual solubility are used for precipitation hardening.

Alloying with elements with other lattice structures or larger differences in atomic radii like the hexagonal-crystallizing metals zinc (atomic radius 133.5 pm), cadmium (148.9 pm), or mercury (150.3 pm) results in the formation of more complex structures.

In some of these alloy systems, structural ordering reactions occur with the formation of superlattices of the same or similar lattice types. Such transformations are accompanied by changes in the mechanical and physical properties which find application in the hardening of the corresponding gold alloys. Superlattices, formed by disorder – order reactions are found in gold alloys with components which show marked differences to gold in their electrochemical properties or in their atomic sizes. Typical compositions have atomic ratios between 3 : 1 and 1 : 3, or 2 : 1 and 1 : 1 (e.g., Au3Cu, AuCu3, Au3Pd, Au3Cd, Au3Zn, Au2Mn, AuCu, AuMn) [105, p. 636]. The ordered phases form in certain temperature ranges superstructures with alternating regular linear arrangement of periods of definite lengths, which meet in the crystal at the so-called anti-phase boundaries (APBs). The number of cells per period depends on the valence-electron concentration and changes with alloying element and composition [106].

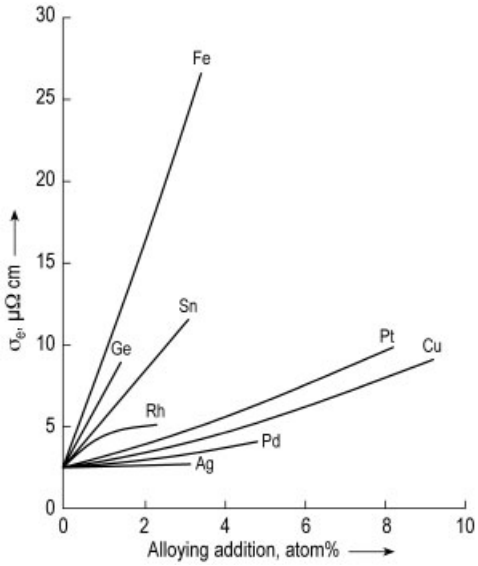

Figure 10. Influence of alloying elements on the electrical conductivity of gold [112]

Intermetallic compounds of gold with metals of the A group and with the B group elements of the periodic table form a variety of different structures in dependence on the alloying elements and the composition. Many of them are stable over an extended composition range and represent metallic phases rather than intermetallic compounds of definite stoichiometric composition [107], [108].

Intermetallic compounds with B metals form identical or similar crystal structures at electron – to – atom ratios of 3/2, 21/13, and 7/4

Table 2. Hume-Rothery phases of Cu, Ag, Au

Electron: atom ratio = 3 : 2			Electron: atom ratio = 21 : 13	Electron: atom ratio = 7 : 4
Body-centered cubic structure	Complex cubic ("β-manganese") structure	Close-packed hexagonal structure	"γ-brass" structure	Close-packed hexagonal structure
CuBe	Cu_5Si	Cu_3Ga	Cu_5Zn_8	$CuZn_3$
CuZn	AgHg	Cu_5Ge	Cu_5Cd_8	$CuCd_3$
Cu_3Al	Ag_3Al	AgZn	Cu_5Hg_8	Cu_3Sn
Cu_3Ga	Au_3Al	AgCd	Cu_9Al_4	Cu_3Ge
Cu_3In		Ag_3Al	Cu_9Ga_4	Cu_3Si
Cu_5Si		Ag_3Ga	Cu_9In_4	$AgZn_3$
Cu_5Sn		Ag_3In	$Cu_{31}Si_8$	$AgCd_3$
AgMg		Ag_5Sn	$Cu_{31}Sn_8$	Ag_3Sn
AgZn		Ag_7Sb	Ag_5Zn_8	Ag_5Al_3
AgCd		Au_3In	Ag_5Cd_8	$AuZn_3$
Ag_3Al		Au_5Sn	Ag_5Hg_8	$AuCd_3$
Ag_3In			Ag_9In_4	Au_3Sn
AuMg			Au_5Zn_8	Au_5Al_3
AuZn			Au_5Cd_8	
AuCd			Au_9In_4	

(Hume-Rothery phases) [97, p. 197]. Table 2 gives some examples.

Besides these, there are a huge number of other intermetallic phases with predominantly cubic, hexagonal, or tetragonal symmetry [36, p. 502].

Influence of Alloying. Strengthening is performed by alloying and by thermal treatment. Hardening is effected either by solid solution (causing elastic interactions between dislocations and solute atoms), by precipitation hardening using miscibility gaps, or by dispersion hardening [109]. Figure 9 gives some examples for the solution hardening effect of different alloying elements [110]. Precipitation hardening is of practical importance in jewelry, in dentistry, and in industrial apparatus and construction components. In binary and ternary gold – copper and gold – copper – silver alloys with higher copper contents, additional hardening can be obtained by means of ordering reactions [111]. Dispersion hardening is effected by incorporating in very fine distribution small amounts (0.1 – 2 vol %) of very small particles of high-melting compounds (e.g., oxides, carbides), which are insoluble or only become soluble near the melting temperature of the noble metal matrix. Dispersion strengthening is of importance if high electrical conductivity has to be combined with high-temperature mechanical strength and corrosion resistance. Dispersion-hardened alloys are produced by powder-metallurgical methods, supported by milling of the mixed powders (mechanical alloying), by internal oxidation of the base metal components in the alloy, or by precipitation of the gold powder in the presence of refractory oxides.

Fine grain structure is needed for reasons of higher strength and forming characteristic to give a smooth surface, especially in jewelry and dentistry. Effective grain refining for castings of gold and gold alloys is performed by adding small amounts of iridium, ruthenium, rhenium, zirconium, barium, tantalum, niobium (0.01 – 0.1 %), or combinations of Co/Ba, Co/Mo, Ni/Mo. Zirconium, barium, and cobalt increase the recrystallization temperature by about 150 – 200 °C and suppress grain growth on further annealing.

Figure 10 shows the influence of alloying additions on the electrical conductivity [112]. The increase in resistivity in the low concentration range is proportional to the atomic concentration of the solute.

8.1. Binary Alloys [28], [98–102], [113]

The phase diagrams of some important binary gold alloys are shown in Figures 11, 12, 13, 14, 15, 16, 17, 18, 19. The most frequently used binary alloys of gold are:

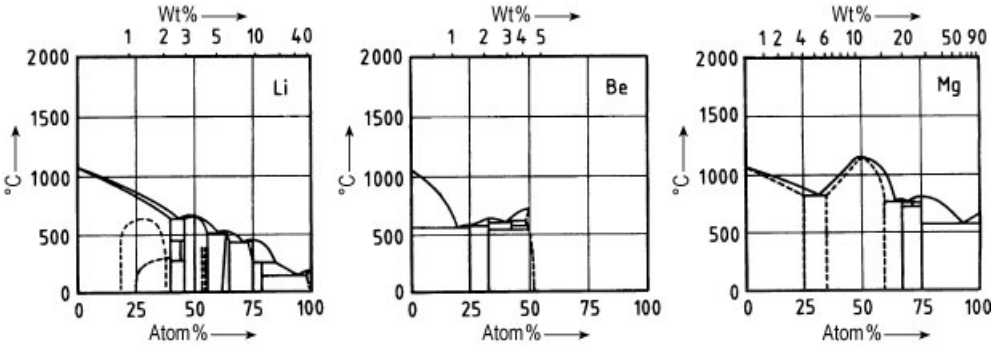

Figure 11. Phase diagrams of groups 1 and 2 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

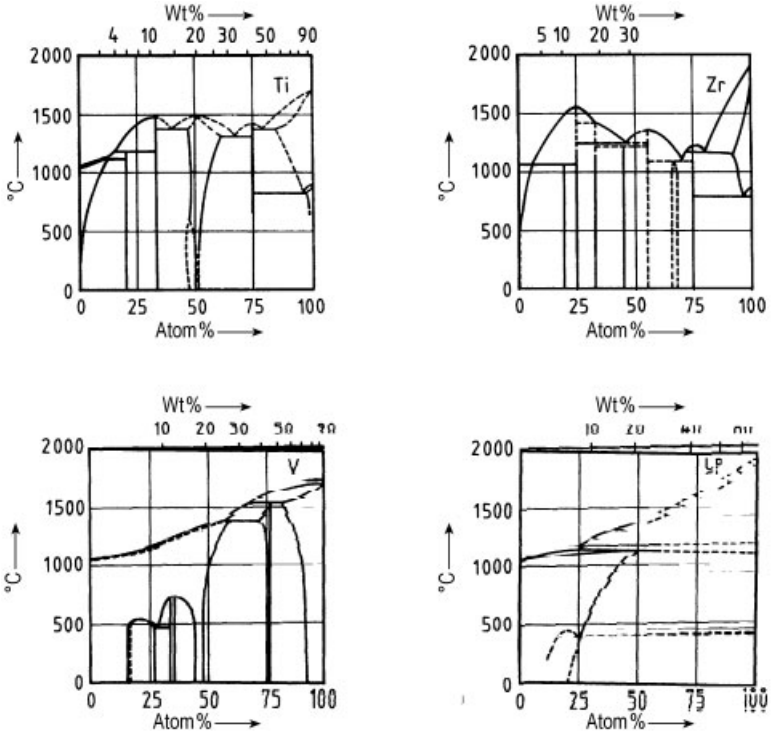

Figure 12. Phase diagrams of groups 4 to 6 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

Gold – Silver. Gold forms a continuous series of mixed crystals with silver, without any transition or separation in the solid state. Au – Ag 20 is used for low-voltage electrical contacts, e.g., rivets.

Gold – Copper. Gold – copper alloys solidify as a continuous series of mixed crystals, with a melting point minimum at 910 °C and ca. 20 wt % copper. Below 400 °C, the ordered phases Cu_3Au [*60748-60-9*], CuAu [*12006-51-8*], and $CuAu_3$ [*12044-96-1*] appear in the solid state. Au – Cu alloys are also used as highly oxidation resistant, high-temperature stability alloys in turbines, nuclear, and space technology, and jet engine components.

Gold – Nickel. This continuous series of mixed crystals, with a melting point minimum at 950 °C and ca. 18 wt % nickel, decomposes at temperatures below 800 °C into gold-rich and nickel-rich mixed crystals. These alloys are used

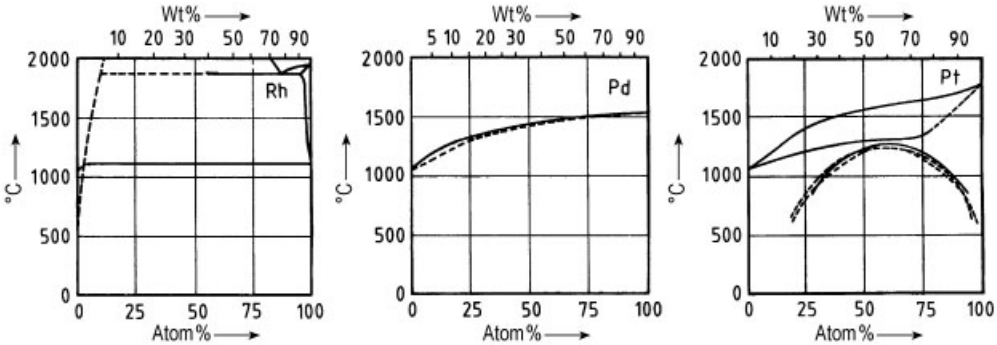

Figure 13. Phase diagrams of groups 9 and 10 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

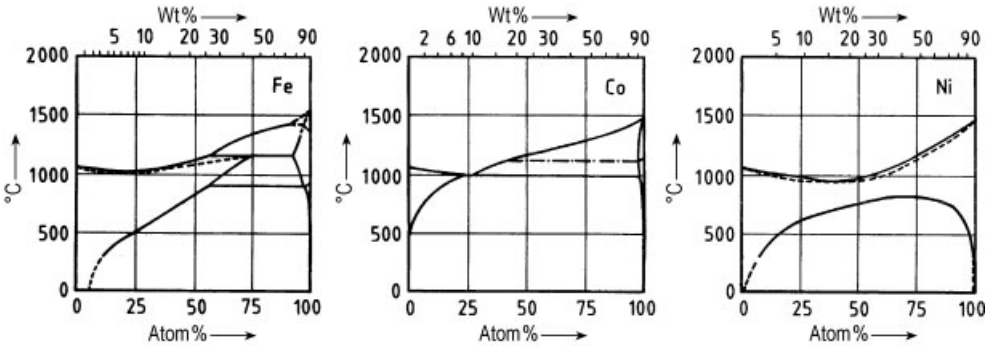

Figure 14. Phase diagrams of groups 8 to 10 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

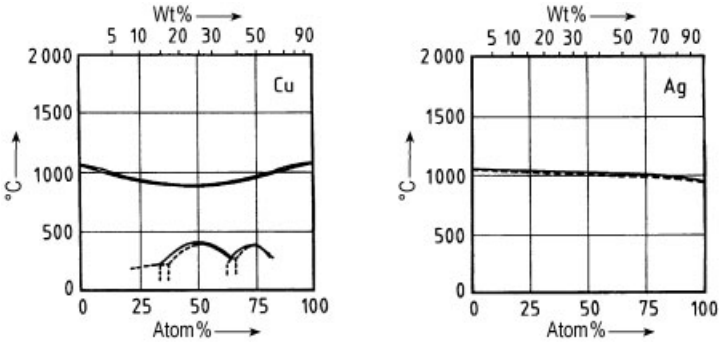

Figure 15. Phase diagrams of group II binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

as brazing alloys; the most common composition is Au – Ni18 wt %. It has high oxidation resistance and good thermal stability up to 800 °C. The production of turbines, jet engines, and components and devices in nuclear and space technology are fields of application. Au – Co5 and Au – Ag26 – Ni3 serve as contact materials which are resistant to silver migration.

Au alloys with small amounts (up to 1 – 3 wt %) of iron, cobalt, or nickel are electrolytically deposited as hard, wear-resistant surface coatings of electrical contacts.

Gold – Palladium. With palladium, gold forms a continuous series of mixed crystals, without any phase separation in the solid state.

Gold – Platinum. Gold mixes with platinum in all proportions. Unlike the narrow melting range in gold – palladium alloys, most gold – platinum alloys solidify within a very

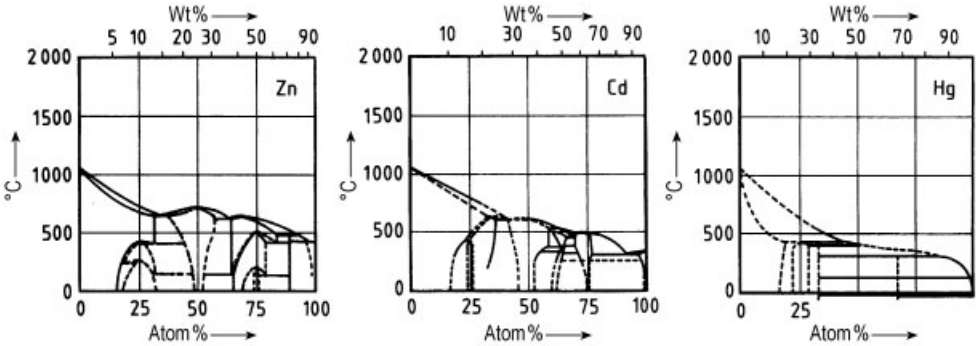

Figure 16. Phase diagrams of group 12 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

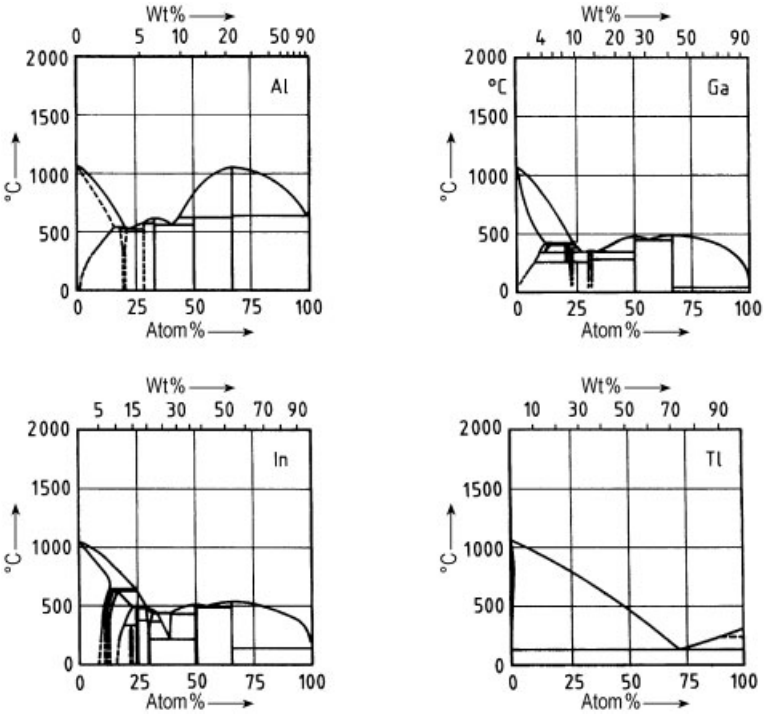

Figure 17. Phase diagrams of group 13 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

broad temperature range. In the solid state, a miscibility gap develops, which approaches the solidus curve in the middle of the phase diagram. Because of their high corrosion resistance, Au – Pt alloys are used for spinnerets in rayon production, and also serve as high-melting platinum solder. A typical composition is Au – Pt30 with a liquidus temperature of 1450 °C and a solidus temperature of 1228 °C. Alloys with 40 – 65 wt % Au can be effectively hardened by quenching from 1100 °C and annealing at ca. 500 °C to a tensile strength at room temperature up to ca. 1400 N/mm^2. Au – Pt10 and Au – Ag25 – Pt5 are used for electrical contacts operating under highly corrosive conditions. Au – Cu14 – Pt9 – Ag4 is used for oxide-free slip contacts of high reliability at very low voltage and current.

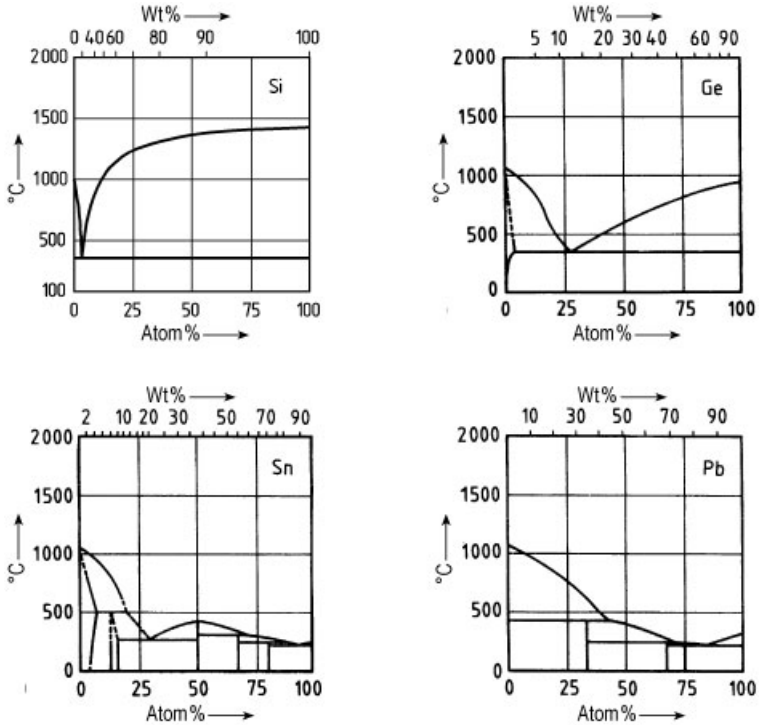

Figure 18. Phase diagrams of group 14 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

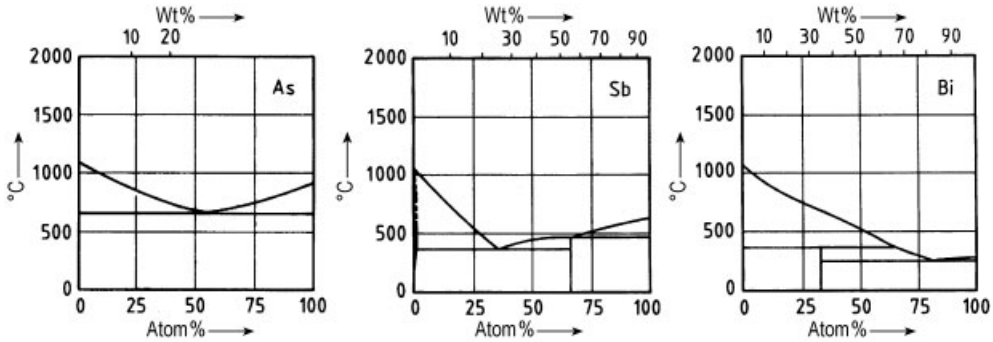

Figure 19. Phase diagrams of group 15 binary gold alloys (upper abscissas in wt %, lower abscissas in atom %)

Gold – Zinc. With zinc, gold forms mixed crystals to a limited extent. The phase diagram for this system is very complex, and shows a large number of intermetallic phases and transitions in the solid state.

Gold – Cadmium. The characteristics of this system are very similar to those of gold – zinc.

Gold – Titanium. The alloy Au – Ti1 wt % has become important in the field of bonding wire for the connection of bond pads of semiconductor chips of intergrated circuits to the outer leads. Thermomechanical treatment effects strengthening by forming the intermetallic compound Au4Ti. Additional hardening can be performed by dispersion hardening by prolonged heating at elevated temperatures in air, which effects the formation of finely dispersed titanium oxide particles by internal oxidation of the titanium containing component [114].

Au – Si, Au – Ge, Au – Sn. The alloy systems gold – silicon, gold – germanium, and gold – tin form low-melting eutectics (AuSi3, 1 wt %: *mp* 370 °C, AuGe 12 wt %: *mp* 356 °C, AuSn 20 wt %: *mp* 280 °C). They are of high strength and serve as corrosion resistant and mechanically stable solders in the fabrication of semiconductor devices.

Gold – Mercury. Up to a gold content of ca. 10 %, these alloys are liquid at room temperature. For higher gold contents, many intermetallic compounds occur. Room temperature is usually sufficient for the formation of gold amalgams, which are still used in the recovery of gold.

The phase diagrams for most of the remaining systems, including those which are not used in industry, have also been thoroughly investigated.

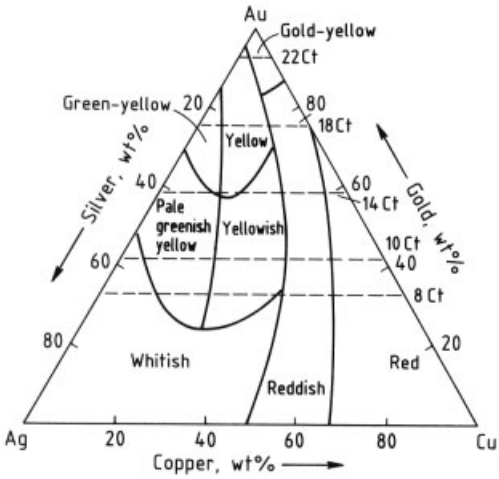

Figure 20. Gradation of color of gold – silver – copper alloys

8.2. Ternary Alloys [103]

For centuries, the color (Fig. 20), hardness (Fig. 21), and corrosion resistance (Fig. 22) of gold were modified by variations in the Au – Ag – Cu system. When white gold, with its resemblance to platinum, became fashionable at the beginning of this century, Ni and Pd were added; these were soon followed by Zn, and later by platinum. The gold alloys used for modern dental materials are complex because they must have specific properties.

The ternary alloys of gold are industrially more important than the binary alloys and have been thoroughly investigated and documented.

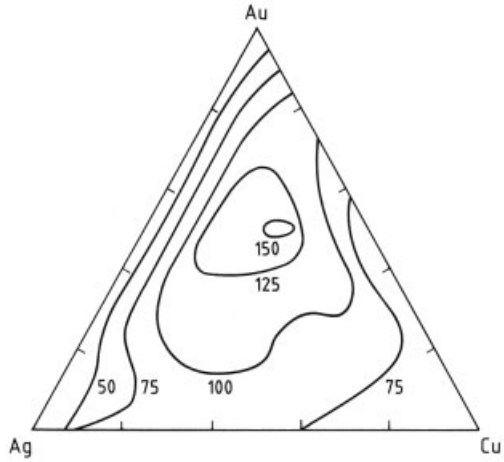

Figure 21. Brinell hardness of annealed and quenched gold – silver – copper alloys

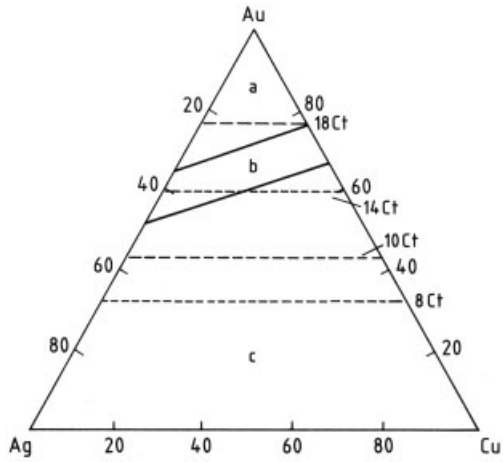

Figure 22. Corrosion resistance of gold – silver – copper alloys
a) Alloys are tarnish resistant and resist all acids except hydrochloric acid in the presence of an oxidizing agent; b) Acids may dissolve out copper and silver from the surface of the alloy; c) Alloys tarnish and are attacked by acids

Gold – Silver – Copper. Figure 23 shows the liquidus, and Figure 24 shows isothermal sections through the miscibility gap of the gold – silver – copper system. The melting surface is characterized by a gutter starting from the

silver – copper eutectic and continuing in the direction of the melting point minimum of the gold – copper system. As the temperature decreases, the miscibility gap at the silver – copper binary edge broadens out considerably towards the gold corner. Shortly after solidification, the heterogeneous phase only extends to alloys containing slightly more than 40 % gold; at 400 °C, it includes alloys with 75 % gold.

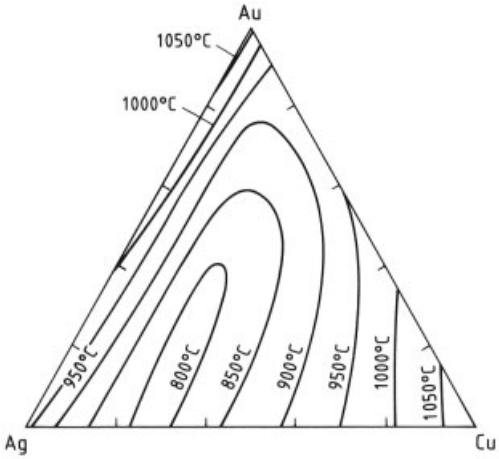

Figure 23. Liquidus of the gold – silver – copper system

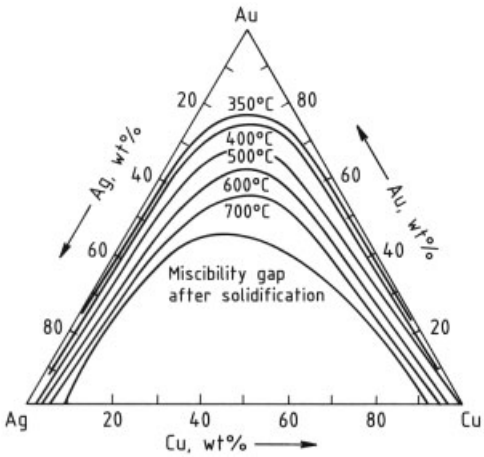

Figure 24. Isothermal sections through the miscibility gap of the Au – Ag – Cu system

The miscibility gap covers a large part of the gold – silver – copper system, and its extent is strongly dependent on the temperature. This is of great significance in the processing and application of gold alloys, and is the main reason why age-hardening is possible.

Figure 25 shows a section through the phase diagram of the gold – silver – copper system, running from the gold corner to the middle of the silver – copper edge. As the temperature decreases, the regions in which mixed crystals and melt occur are followed by regions in which gold-rich α mixed crystals, and the two mixed crystals α_1 (silver-rich) and α_2 (copper-rich) coexist. The saturation line of the α mixed crystal, which separates the two regions from each other, is highly temperature dependent. At 750 °C, the α mixed crystal can dissolve ca. 50 % silver and copper (in approximately the ratio 1 : 1); at ca. 300 °C, it can only take up about 10 % silver and 10 % copper. An alloy with 70 % gold, which at high temperatures consists only of α mixed crystals, reaches the saturation line during cooling at ca. 480 °C. With further cooling, it splits into the α_1 and α_2 mixed crystals. At tempering temperatures below 480 °C, the alloy therefore age-hardens.

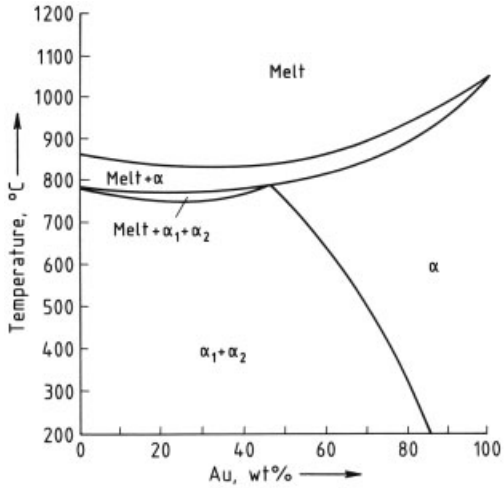

Figure 25. Section through the Au – Ag – Cu phase diagram running from the Au corner to the middle of the Ag – Cu edge

Figure 26 shows a section through the ternary system parallel to the silver – copper binary edge with a constant gold content of 70 %. The extensive region of the α mixed crystals adjoins the miscibility gap which ends just below 500 °C. In the miscibility gap region, the α_1 mixed crystal, which has a relatively high silver content, and the α_2 mixed crystal, which has a relatively high copper content, coexist.

Gold – Nickel – Copper. Figure 27 shows the liquidus isotherm of the gold – nickel – copper system, which produces white alloys over broad areas. Below 800 °C in the gold – nickel system (Fig. 14), a miscibility gap stretches far into the ternary system, as the temperature falls, in the direction of the copper corner. These alloys can therefore be age-hardened, but are not always easy to process, particularly as the nickel content makes them very sensitive to traces of sulfur. As little as 0.01 % sulfur is sufficient to make them brittle when hot. Phase separation in nickel-containing white gold alloys leads to the formation of a gold-rich mixed crystal, and a nickel-rich ferromagnetic phase, which tends to lower the chemical resistance of the alloys.

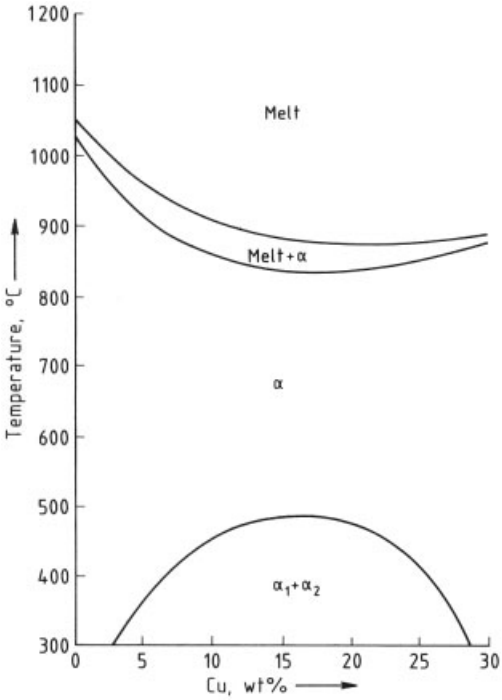

Figure 26. Section through the Au – Ag – Cu phase diagram parallel to the Ag – Cu edge with a constant gold content of 70 %

Gold – Silver – Palladium. Figure 28 shows the liquidus isotherms of this system, which produces white, high-melting alloys over a broad area. These are relatively soft and cannot be age-hardened.

8.3. Higher Alloys

In jewelry, the most important quaternary alloys are derived from the ternary systems Au – Ag – Cu, Au – Ni – Cu, and Au – Ag – Pd, most often through the addition of zinc. The Au – Pd – Ag – Cu and Au – Ni – Ag – Cu systems are also of practical importance. Colored gold solders may contain cadmium in addition to the above elements.

Dental gold alloys often have five or more components. These are formed from Au – Ag – Cu by addition of Pd and Pt.

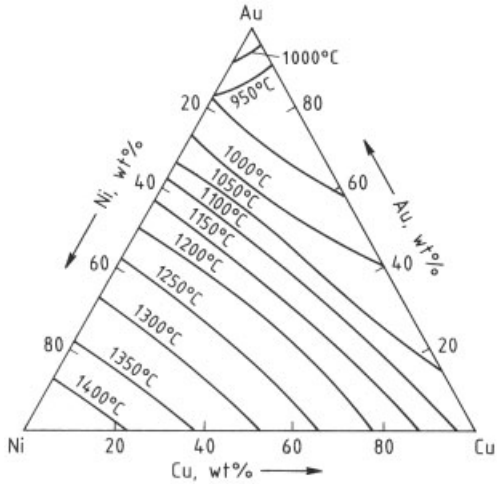

Figure 27. Liquidus isotherms of the gold – nickel – copper system

8.4. Production and Processing [115–117]

Manufacture. Bars, rods, foils, and wires are manufactured by classical metallurgical processes: melting, casting, rolling, drawing, and powder-metallurgical compaction. Pure gold is very malleable. It can be rolled to very thin foils and drawn to wires of very small diameters. Special procedures have been developed for the production of gold leaf (see Section 10.9) and for casting small jewelry items. Gold powders of different grain sizes and shapes are made by precipitation from preferably aqueous solutions of gold compounds with reducing agents. Thicker, compact surface layers are produced by cladding, especially for electrical contact pieces

and jewelry items (Doublé). Thin surface layers or films are produced by galvanic processes, by application of gold-powder-containing paste preparations, by evaporation, or by sputtering. Metallizing pastes contain additions of organic vehicles for flow properties. They are applied by painting, by brushing or by screen printing, in the latter case usually followed by drying and firing. Special preparations differing in gold powder grades and type of liquid components are in use for the production of electronic devices. Pastes for the production of so-called thick-film layers on ceramic substrates may contain glass powders which effect in the firing operation bonding on the ceramic surface.

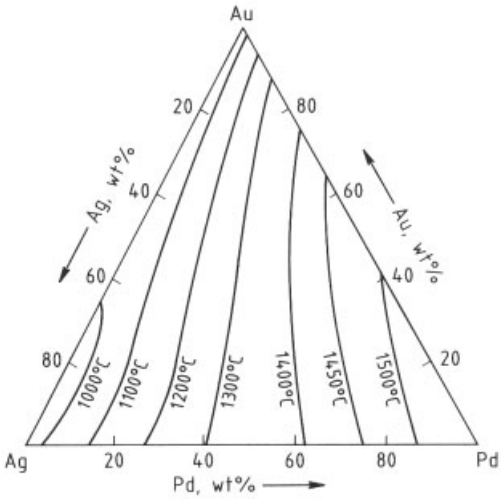

Figure 28. Liquidus isotherms of the gold – silver – palladium system

Melting. Gold alloys are made by melting fine metals in crucibles made of a clay – graphite mixture or fireclay. Gas-heated furnaces used to be common, but electrically heated smelting plants or induction furnaces are now used. In order to prevent oxidation of the base metal components, the melt is covered with lumps of charcoal and borax. The air above the crucible can also be replaced by carbon dioxide. Melting is sometimes carried out in vacuum. The molten metal is poured into preheated iron molds to obtain the required form.

Rolling and Drawing. Cast plates are flat rolled to form sheets; billets are first shaped on grooved rolls, and then drawn to form wires. Intermediate annealing processes are usually carried out in electrically heated furnaces. In order to prevent oxidation of gold alloys, a protective gas consisting of partially burned coal gas or propane is blown through the furnace.

Fabrication of Jewelry. In former times, gold jewelry was made by hand from sheets and wires, and the pieces soldered together. The mechanical properties of the alloys were not critical because the goldsmith was able to adapt his methods according to the material. With the increase in factory production, quality requirements became much more important. The correct choice of a gold alloy for a particular production process has thus become critical.

Beating, Deep-drawing, Spinning, and Pressing. Soft gold alloys are most suitable for these methods. Moderately hard or soft gold alloys are used to manufacture jewelry from sheet metal, such as brooches and rings. Chains and plaits are now almost exclusively machine-made; for this purpose, all soft alloys, and a very few moderately hard alloys can be used. Alloys for chains must have constant mechanical and processing properties, so that the machine can produce a well soldered strand. For pin shanks, spring catches, and other similar articles, the hardest possible alloys must be used, because good resistance to wear and high elasticity are required [117].

Punching. To make seamless wedding rings, a hole is punched out of a disk. The resulting piece then undergoes a heavy forming operation without removing any material, in a specially adapted production process.

Casting. The manufacture of jewelry by casting gold in molds made of metallic or inorganic materials, or sometimes organic materials (cuttlebone), has been used throughout history. The casting of jewelry using the lost wax process has become an important manufacturing method. Several wax models of the desired shape are made in a split mold, provided with a joint gate, and embedded in a paste mixed with water, consisting essentially of quartz and gypsum. When this paste has hardened, the wax is melted out, and the hollow mold fired at red heat. The molten gold alloy is then poured in. After cooling, the formed pieces are removed, smoothed, and polished.

Age-Hardening. The hardness of many gold – silver – copper alloys can be increased by heat treatment at 250 – 350 °C. However, in the fabrication of jewelry only limited use has been made of this possibility, because the hardening that takes place during forming is usually sufficient.

As age-hardening in gold – silver – copper alloys takes place at lower temperatures than recovery after cold forming, the superposition of age-hardening and forming operations can lead to a considerable increase in the hardness and strength of a workpiece. This method is useful in the fabrication of parts which are to serve as holding or fixing elements.

9. Quality Specifications and Analysis [20], [118–124]

9.1. Quality Specifications [39]

The gold content of fine gold and gold alloys is usually expressed in fineness (parts per thousand) or in carat, where 24 carat represent 1000/1000 (‰) fine, 18 carat thus being equal to 750/1000 (Table 3). For ores, intermediate industrial products, residues and waste materials, the gold content is usually given in gram per metric ton (g/t), µg/g, ounces per metric ton (oz/t) or ounces per short ton, where 1 oz = 1 troy ounce = 31.1035 g and 1 short ton = 907.184 kg. Other mass units are 1 dwt = 1 pennyweight = 0.05 oz = 1.555 g and 1 troy grain = 1/24 dwt = 0.0648 g. 1 tola = 11.664 g is an Indian mass unit for gold. The Chinese Gold and Silver Exchange Society uses the unit tael (1 tael = 37.429 g).

Gold bars are classified into two "qualities", depending on their method of manufacture: cast and struck. For the accredited refiners 38 different bar masses ranging from 12.5 kg to 1 g are recorded [125].

In international gold trading the content of gold is the decisive criterion. The concentrations of individual impurities are normally not taken into consideration if their sum does not exceed the specified upper limit. If agreed upon, the ASTM Standard Specification for Refined Gold (ASTM: B 562-95) can be used as a guideline for the chemical composition. Individual refiners may offer different specifications based on their own products. ISO 9202 lists the fineness of gold jewelry alloys, which constitute the mayor use of refined gold. Special industrial applications of gold may require particular specifications for individual metallic or nonmetallic trace elements.

9.2. Sampling [119], [126]

An essential part of any determination of the gold content of any material and of great economic significance is sampling before analysis. A number of partly automated standard sampling procedures are available for solid materials that can be homogenized by comminution and mixing (ores, concentrates, sweeps, and others), and for unrefined gold and gold alloys that can be homogenized by inductive melting. Where these approaches are not possible, homogenization by, for example, dilution melting with copper or other base metals may be necessary.

Grindable materials are usually screened after grinding to separate the fine from the (metallic) coarse fraction. These are then subjected separately to mechanical, electrothermal, or manual sampling processes.

In the case of metallic materials such as mined gold, doré silver, or industrial gold alloys, the material to be analyzed can be obtained by sawing sample bars or by drilling or sawing small dip samples (buttons).

For the sampling of jewelry alloys, an ISO procedure is being worked on at present [127].

Recycled electronic material contains mostly only a few parts per thousand gold together with plastic and base metal or ceramic components. In a multistage process, mechanical size reduction in a shredder in combination with fractionation by screening produces material from which samples can be taken [128].

9.3. Qualitative and Semiquantitative Analysis

For the determination of gold, a large number of spectroscopic methods, like FAAS, GFAAS, ICP-OES, and XRF is available. In emission and absorption spectroscopy the most sensitive lines at 242.8 nm and 267.6 nm are used. To detect gold in the µg/g range in powdered materials, fire assay with lead as collector (see Section 9.4) can

Table 3. Fine gold specifications

Designation	Gold content, wt %	Content of other metals, µg/g	
"Good delivery" gold	≥ 99.5	any metals (total)	≤ 5000
Fine gold	≥ 99.99	Ag	≤ 100
		Cu	≤ 20
		others	≤ 30
		total	≤ 100
Fine gold, chemically pure	≥ 99.995	Ag	≤ 25
		others	≤ 25
		total	≤ 50
Fine gold, high purity	≥ 99.999	Ag	≤ 3
		Fe	≤ 3
		Bi	≤ 2
		Al	≤ 0.5
		Cu	≤ 0.5
		Ni	≤ 0.5
		Pd + Pt	≤ 5.0
		total	≤ 10

be applied, in which silver nitrate, containing less than 1 ng/g gold, is added to the fluxing material, in order to improve the collector effect, and as protection against gold losses during cupellation. The resulting bead can be dissolved in nitric or sulfuric acid, gold traces collected on a membrane filter, dissolved in aqua regia, and semiquantitatively determined spectroanalytically. Occasionally, the formation of so-called purple of Cassius [124] is used as rapid industrial procedure for the semiquantitative detection of gold in tailing solutions from gold extraction.

9.4. Quantitative Analysis

Quantitative analysis forms the basis for calculating the gold content of all types of materials containing gold during its primary and secondary extraction. The first group includes, for instance, gold ores, concentrates, and mined gold, while the second represents recovery from scrap dental and jewelry alloys, doré silver, ion-exchange resins, spent galvanic baths, electronic scrap, and byproducts of the base metal industry, such as sludges from copper electrolysis, etc. Quantitative analysis is also used for quality control of the end products of refining ranging from commercial gold to ultrapure gold and all the products (e.g., alloys, and salts) obtained when gold products are processed.

The duality of the methods for analyzing gold for most of the materials mentioned is especially characteristic, that is, the parallel use of classical and modern instrumental methods. The most important analysis method for gold is still the centuries-old docimastic analysis or docimacy (Greek: testing, assay), more commonly known as the fire assay. In this process, precious metals are taken up by molten lead while base metals are removed by slagging. Two variations of the method are used to extract the gold from a mainly nonmetallic or base metal matrix. In the *crucible assay*, a mixture of lead oxide, a reducing agent (e.g. carbon), and a flux is melted at ca. 1300 °C. In the *scorification assay*, the sample is subjected to oxidative smelting with granulated lead and borax at ca. 1000 °C. In each case a lead button containing the gold and the other precious metals of the sample is formed.

The lead is then oxidized at ca. 900 °C to litharge (PbO) on a magnesia support known as a cupel (*cupellation process*). The PbO melt is absorbed by the porous magnesia (together with base-metal oxides possibly present) and a drop of precious metal(s), which later solidifies to a bead remains behind. The precious metal bead is treated with nitric acid. Depending on the silver/gold ratio in the sample material, it may be necessary to inquart the precious metal

bead with additional silver to facilitate quantitative separation of silver and gold during parting. If the precious metal bead consists of Au, Ag, Pt, and Pd in suitable concentration ranges, Ag, Pt, and Pd are dissolved with boiling nitric acid, while gold remains undissolved and can be weighed. Cupellation losses are compensated for by the use of proof samples of similar composition to the material being assayed.

If the precious metal bead contains rarer platinum group metals (Rh, Ir, Ru) in addition to gold then special decomposition techniques have to be applied.

An important alternative to the extraction of gold and other precious metals with lead is the use of copper as collector.

The cupellation of metallic samples, wrapped in lead with an addition of silver, is the most important method for the accurate determination of gold in alloys with ca. 30 to 99.9 % gold, e.g., dental alloys or mined gold [129]. The degree of accuracy of gold analysis of alloys by cupellation lies between about 0.1 ‰ and 1 ‰ depending on the gold concentration and the composition of the matrix.

Spectrometric methods with a precision much better than 1 % are acceptable as alternatives to fire assay for the accurate analysis of gold in the range from 0.1 % to the high-percentage region. Up to the late 1980s this was only true for wavelength and energy dispersive X-ray fluorescence analysis (XRF) of solid alloys or solutions, and to a limited extent for glow discharge optical emission spectrometry (GD-OES) of gold alloys [130], [131]. With the X-ray solution method as a multielement method for all precious metals except Ag, an accuracy of 1 ‰ can be achieved for the analysis of gold as main constituent [132], [133]. A small energy dispersive X-ray instrument [135] allows gold determinations in carat alloys with an accuracy of 2 to 3 ‰.

The basis for the precision analysis of precious metals by inductively coupled plasma optical emission spectrometry (ICP-OES) is internal standardization. The essential method characteristics for the determination of gold [134] are

- The use of a simultaneous spectrometer with high measurement stability
- The simultaneous measurement of the 267.6 nm or the 242.8 nm Au atomic line and the 371.0 nm ionic line of yttrium as internal standard
- The preparation of all solutions and the addition of the Y standard solution by exact weighing
- In special cases the use of a Na/Cu buffer, but in any case the use of a HCl concentration of at least 5 %.

Several wet chemical/gravimetric methods are used in special cases in conventional gold analysis, for instance, reduction of gold by SO_2 in hydrochloric acid solution.

9.5. Purity Analysis [121]

A large number of methods for purity analysis of gold are available; some of them can be applied to metallic gold, and others to the analysis of gold dissolved in acid.

9.5.1. Direct Analysis of Metallic Gold

1) Spark discharge emission spectrometry (SD-OES) with high-speed time resolution of individual sparks is well suited for the analysis of fine gold with total impurities in the range of 50 to 1000 µg/g, that is, for the most common grades of fine gold in routine operation; however, it is not suitable for 99.999 % gold.
2) DC arc spectrometry with simultaneous background correction offered by the application of a solid state detector (CID). The detection limits for all trace elements are lower than for spark excitation.
3) Spark ablation ICP-OES offers a further possibility of direct impurity analysis in which sample ablation is caused by high-voltage sparks and the resulting aerosol is transported into the ICP by an argon carrier gas stream.
4) Glow discharge emission (GD-OES) or ablation spectrometry (GDA-AAS). The detection limits for impurities in gold with classical DC-powered GD-OES are equal to or better than 1 µg/g.
5) Laser ablation ICP mass spectrometry can be applied to purity analysis in high purity gold with detection limits of better than 0.2 µg/g.

Table 4. Coinage alloys

Coin	Country	Mintage period	Fineness	Carat	Gross weight, g	Gold content
20 mark	German Reich	1871 – 1915	900	21.6	7.964	7.168 g
10 mark	German Reich	1871 – 1915	900	21.6	3.982	3.584 g
5 mark	German Reich	1871 – 1915	900	21.6	1.991	1.792 g
1 ducat	German Confederation	up to 1871	986.1	23.7	3.490	3.441 g
Krugerrand ∗	R. South Afr.	since 1967	916.6	22	33.931	1 oz
Maple Leaf ∗	Canada	since 1979	999.9	24	31.103	1 oz
American Eagle ∗	USA	since 1986	916.6	22	33.931	1 oz
Britannia ∗	Great Brit.	since 1987	916.6	22	33.931	1 oz
Nugget ∗	Australia	since 1987	999.9	24	31.103	1 oz
Tscherwonez	USSR	since 1975	900	21.6	8.60	7.74 g

∗ Also in 1/2, 1/4 and 1/10 oz gold.

It has also found an interesting application in the characterization of trace element composition of unrefined gold of different origins, akin to traditional fingerprinting. The method therefore facilitates identification of the provenance of stolen gold.

6) Glow discharge mass spectrometry GD-MS) is the queen among spectroscopic techniques for the analysis of gold of quality 99.995 % Au or better; it is used in the electronics industry.

9.5.2. Purity of Gold Solution

As an independent alternative method to direct metal analysis ICP-OES with a SIM/SEQU or sequence instrument with axial plasma can be used for production control or for the analysis of gold calibration standards. The use of an Au concentration of 50 g/L offers no difficulties. Especially advantageous is the use of ICP spectrometers with semiconductor detectors.

9.6. Trace Analysis

The determination of trace amounts of gold (ca. 0.1 – 100 µg/g) is important in ore prospecting and in the evaluation of residues and waste materials from the metallurgical industry.

The solids can be dissolved and the resulting solution analyzed by a suitable spectroscopic method:

- Flame atomic absorption (FAAS)
- Graphite furnace atomic absorption (GFAAS)
- X-ray fluorescence analysis with totally reflected exciting beam (TXRF)
- Direct current plasma optical emission spectrometry (DCP-OES)
- Inductively coupled plasma emission (ICP-OES)
- ICP quadrupole mass spectrometry (ICP-MS)

The gold traces may need to be enriched prior to determination to improve the detection limits, for instance, by solvent extraction. In the field of trace analysis, composite fire assay/atomic spectroscopy or mass spectrometric methods play an important role. They make use of the almost universal application of fire assay for the separation of gold and the possible of using large samples.

10. Uses of Gold and Gold Alloys [6], [12], [19], [20], [22], [35], [39], [45], [47], [136], [137]

10.1. Coins, Medals, and Bars [138]

In antiquity, gold was considered an object of value and used as a means of exchange, at first in the form of nuggets or flattened disks, and from 650 B.C. as minted coins. The composition of the metal used varied from alloys containing a high proportion of gold, to electrum which had a relatively high silver content. The confusing variety of gold coins, and the great variation in fine

gold content, was first limited in the 19th century, when legal regulations were introduced for coinage, which specified nominal values of gold coins, their weight, and gold content. After the First World War, gold coins were no longer legal tender. They are now traded as collectors' items and as a form of investment. With the liberalization of private gold trading in the 1950s, the demand for gold increased. A successful new enterprise in this market was the minting and marketing of the krugerrand (gold content of 1 oz troy) by the Republic of South Africa. Between 1967 and 1985, a total of 2000 t (ca. 65×10^6 oz.) were sold. Following the repressive measures which have been inflicted on the krugerrand, new coins were introduced by other countries (Table 4).

Gold was already stored in the form of bars in antiquity, especially in Rome. Today, gold bars weighing ca. 400 oz. (ca. 12 kg) of good delivery quality (99.5 %) are the main form used in public and institutional investment. Gross weight, gold content, the manufacturing firm and reference number are stamped on the bars, and provide direct and binding information as to their value. Smaller gold bars are manufactured for private investors.

The difference between the buying and selling price in banks (broker's commission) is normally ca. 3 – 4 % plus value added tax.

10.2. Jewelry [139]

Fine Content of Jewelry Gold. The gold value of jewelry alloys is determined by their gold content (fineness, cf. Section 9.1). In most countries, laws govern the terms used in designating the fineness of gold jewelry for manufacturers, processors, and dealers.

Only alloys with a minimum gold content of 585/000 are sufficiently tarnish resistant for jewelry. Better quality jewelry customarily has a gold content of 750/000. These alloys offer optimum color and mechanical properties. For less expensive jewelry, alloys with a low gold content are often used. In Germany, 333/000 fine is common, in other countries 375/000, and in the United States 417/000. These alloys tarnish under unfavorable conditions, and cracks may form as a result of stress corrosion. The highest grade alloys are used only occasionally in the manufacture of jewelry, due to their low strength. However, no alloy can match the fine color of pure gold.

Colored gold alloys used in the jewelry industry are mostly based on the ternary alloy system Au – Ag – Cu, Cu or the quaternary system Au – Ag – Cu – Zn allowing a wide variety of colors (see Fig. 20). The workability and resistance to wear of an alloy depend on its mechanical properties; these, its resistance to corrosion, and the color can be controlled by adding zinc. The classification ranges for colored gold alloys used in the jewelry sector are shown in Table 5. To designate the different qualities, special codes are used, most of which are specific to particular firms. The properties cover a broad range of values to meet all practical requirements (Tab. 6). In 1966 in the Federal Republic of Germany, an industrial standard, DIN 8238 "Gold colors" (including white gold), was created to standardize colors in gold alloys and to provide manufacturers with a better means of mutual understanding. This standard closely reflects similar specifications in Switzerland and France. A combination of values representing tone (T), saturation level (S), and darkness (D), measured using spectrophotometric methods, is attributed to each color (Tab. 7).

Table 5. Composition of gold alloys for jewelry

	Gold content, wt %	Content of alloy components, wt %
Colored gold		
22 carat	91.66	Ag 0 – 6.0, Cu 2 – 8.3
18 carat	75	Ag 0 – 20, Cu 5 – 25
14 carat	58.5	Ag 0 – 41.5, Cu 0 – 33.5, Zn 0 – 10
9 carat	37.5	Ag 0 – 56.0, Cu 4 – 60, Zn 0 – 15
8 carat	33.3	Ag 0 – 58.5, Cu 4 – 60, Zn 0 – 15
White gold		
18 carat	76	Cu 4 – 8, Ni 10 – 18, Zn 3 – 6
18 carat	75	Pd 10 – 20, Ag 0 – 1, rest Cu, Zn
14 carat	59	Cu 15 – 25, Ni 10 – 16, Zn 5 – 8
14 carat	59	Pd 10 – 20, Ag 10 – 30, rest Cu, Zn

Gold Solders. The most frequently used method of making joins in the manufacture of jewelry from gold alloys is hard soldering. In addition to fine gold, colored gold solders also contain silver, cadmium, copper, zinc, and indium. Their melting temperature is always lower than that of the material to be soldered. The additives are adjusted so that the solders have graduated

Table 6. Properties and applications of jewelry gold alloys (selection)

Carat	Destignation gold/silver, rest copper or (+Zn)	Color	Melting range, °C	Density g/cm^3	Vickers hardness (after soft annealing), HV 10	Applications								
						1	2	3	4	5	6	7	8	9
Colored gold														
18ct	750/40	red	880 – 900	15.1	160 hard	X			X				X	
	750/90	reddish	870 – 890	15.3	150 medium hard	X		X	X	X	X		X	X
	750/125(130)	yellow	850 – 890	15.4	140 medium hard	X	X	X	X	X	X		X	X
	750/150(160)	pale yellow	890 – 920	15.5	130 medium hard	X	X	X	X	X	X		X	X
	750/200	pale yellow	900 – 970	15.8	90 soft							X		
14ct	585/40	reddish	900 – 920	13.0	130	X	X	X	X	X	X		X	
	585/45 (+Zn)	yellow	835 – 860	13.0	100	X	X	X		X	X			X
	585/100	reddish yellow	820 – 870	13.3	140	X		X			X			X
	585/200	yellow	830 – 850	13.6	170	X								
	585/260	pale yellow	830 – 850	13.7	160	X		X		X			X	
	585/300	pale yellow	820 – 890	13.9	140	X	X	X	X	X	X			X
	585/340	green yellow	860 – 940	14.0	120	X	X	X	X	X	X			X
	585/415	green yellow	1015 – 1020	14.3	50							X		
9ct	375/90	reddish	870 – 930	11.2	120	X		X		X	X			
	375/60	yellow	870 – 950	11.0	95	X	X	X						X
	375/150	pale yellow	770 – 860	10.9	140	X		X		X	X			X
8ct	333/90	reddish	890 – 930	10.9	110	X		X		X	X			
	333/75	yellow	850 – 910	10.9	105	X	X	X						X
	333/120	pale yellow	800 – 860	10.5	100	X		X		X	X			X
White Gold														
18ct	760S (contain Pd)	whitish gray	1000 – 1160	16.0	100	X	X	X		X				X
	750M (contain Pd)	yellowish white	1040 – 1150	15.8	140	X	X	X		X			X	X
	760H1 (contain Ni)	white	920 – 960	14.9	210	X					X			X
14ct	590S (contain Pd)	whitish gray	1040 – 1130	14.5	90	X	X	X		X				X
	590M (contain Pd)	yellowish white	1060 – 1150	14.5	160	X	X	X		X			X	X
	590H1 (contain Ni)	white	910 – 990	12.7	150	X					X			X

1 Rings, brooches; 2 Deep pulling work; 3 Pressing work; 4 Enamelling; 5 Castings; 6 Chains, lattice work; 7 Blanks for minting; 8 Wedding rings; 9 Pipes, seamless.

Table 7. Gold colors according to DIN 8238

Symbol	German classification	Swiss and French classification	Colorimetric measures according to DIN 6164			Examples of approximately corresponding gold alloys
			T	S	D	
N	weiß		1.2	0.9	1.6	590 H
8 N	grüngelb		24.8	1.6	1.2	585/17 340
1 N	blassgelb	jaune pâle	1.7	1.7	1.2	585/260
2 N	hellgelb	jaune pâle	1.8	1.8	1.3	750/160 (750/150)
3 N	gelb	jaune	2.0	1.8	1.3	750/125 (750/130)
4 N	rosé	rosé	2.4	1.6	1.4	750/90
5 N	rot	rouge	2.6	1.5	1.4	750/40

working temperatures. Three solders are generally sufficient, with graduations of ca. 50 °C in their working temperatures. Some examples of colored gold solders are given in Tables 8 and 9.

White gold alloys were first developed in the early 1900s, in an effort to replace platinum by a cheaper material with identical properties. White gold also differs from colored gold in having a higher melting range (white gold with palladium) and is usually harder (white gold with nickel). Nickel and palladium are the only suitable additives to give gold a color approaching whitish gray. Because of the allergic reaction of skin to Ni the demand for white gold has rapidly fallen. Information about white gold alloys is given in Table 5.

Table 8. Working temperatures of Cd-containing colored gold solders

Gold content, carat	Solder designation (Degussa)	Working temperature, °C
18	750 L1	820
	750 L2	750
	750 L3	700
14	585 L1	780
	585 L2	720
	585 L3	670
9	375 L3	650
8	333 L1	700
	333 L3	640

Table 9. Working temperatures of Cd-free solders

Gold content, carat	Designation (Degussa)	Working temperature, °C
14	585 L1 cdf	800
	585 L3 cdf	730
18	750 L1 cdf	820
	750 L2 cdf	780
	750 L3 cdf	750

10.3. Electronics and Electrical Engineering [22, pp. 475 – 500], [140], [141]

10.3.1. Electronics

The great advantage of gold is its high resistance to oxidation and corrosion, and its high conductivity which give it excellent contact properties.

It is therefore primarily used if oxide-free surfaces are indispensable. Gold surfaces enable the formation of electrical connections with lowest ohmic resistances, forming perfect electrical connections for solder, bond, and pressure contacts where even thinnest oxide layers would cause interruptions, formation of voids, and leakages. It is used in active components (diodes, transistors, integrated circuits, semiconductor memories), assembly and connection engineering (packages, thick-film circuits, printed boards, and plugs) and, to a lesser extent, for passive components (capacitors and resistors).

Table 10 gives a survey of the main application fields.

Partly due to its high intrinsic cost, it is mostly employed in the form of thin layers on different substrate materials.

The gold used in electronics, with the exception of gold solders, is practically always fine gold of purity 99.99 or 99.999 %.

The surface coats are deposited by electroplating [142], sputtering, evaporation, or by means of paste preparations, which are made from special gold powder grades and organic binders and – for adhesion on ceramic substrates – blended with glass frits. These are applied by screen printing and firing. They are especially used for conductor layers of hybrid thick-film circuits and ceramic devices. Gold conductor paths, pure or alloyed with silver, palladium, and platinum in thick-film circuits and in submicron-structure units can also be made by thermal decomposition by electron-beam lithography.

Gold-plated surfaces serve for connections by pressure, by bond – or solder – contacts on semiconductor chips and for the joining of the semiconductor pieces with assembling parts and package components [143]. Gold conducting layers are specially of importance in highly reliable electronic circuits, where silver conductor layers may cause problems by current-induced silver-ion migration.

Gold powder of different grain sizes with spherical or platelike grain shapes are used in metallizing and thick-film pastes to produce bond pads and conductor layers. Gold powder, as needed for paste preparations is produced by the chemical reduction of tetrachloroauric acid $H[AuCl_4]$. Depending on the type of reducing agent and the use of protecting organic colloids, either spherical particles (with sizes between 1 µm and 10 µm) or platelike particles (with

Table 10. Application of gold and gold alloys in electronics , p. 557[36]

Bulk	Bimetal	Surface Coat			
		electroplated	sputtered or evaporated	pastes, lacquers	solder connection
bonding wires for integrated circuits transistors, diodes	plug-in contacts	PCBs, plug-in contacts, solderable and bond surfaces, switches, corrosion protection of semiconductor device packages	thin film hybrid circuits, conductor contact, bond surfaces	thick film hybrid circuits, solderable areas for contacts, die-attach, IC technology	special solders for die-attach, package soldering, and connections in semiconductor power devices

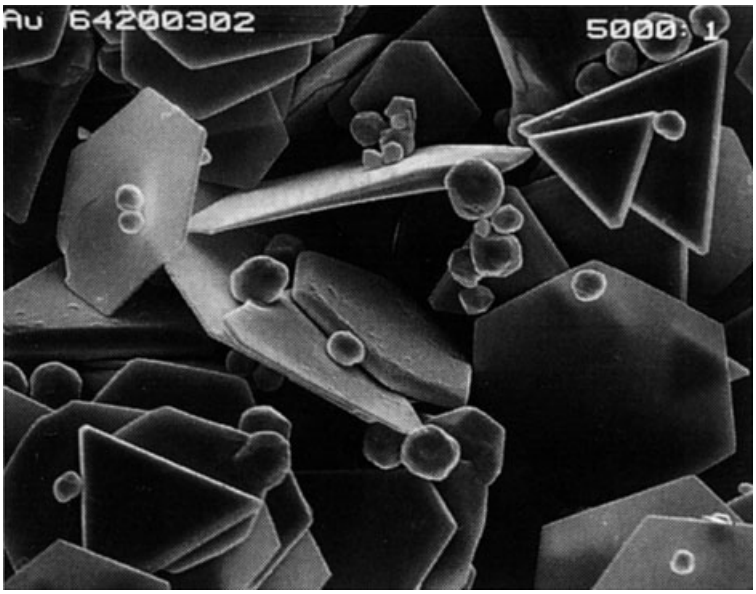

Figure 29. Gold powder with platelet-shaped particles (0.5 – 50 µm) made by chemical reduction of an aqueous solution with special additives [157, p. 57]

thickness of about 0.5 µm and flat dimensions of about 50 µm; see Fig. 29) are formed.

Thin wires ("bond wire", standard diameter 25 µm) made of pure gold, alloyed with small amounts (ca. 7 ppm) of beryllium serve for the connection of the majority of monolithic integrated semiconductor chips (I/Cs) with the inner leads of the device. Bonding is performed by special welding processes (thermocompression and thermosonic bonding), forming a bond loop between the connection pads and the outer leads of the device. The wire is fed through a heated capillary and the tip of the outcoming wire is melted to a small ball, which is pressed onto the semiconductor pad to form a weld contact (ball bonding or nailhead bonding, Fig. 30).

Bonding wire made of the alloy AuTi 1 has higher strength and higher mechanical stability at elevated temperatures. Strengthening is achieved by thermal treatment which causes the precipitation of the intermetallic compound $TiAu_4$. Additional dispersion hardening can be achieved by internal oxidation to form a finely dispersed titanium oxide phase in the alloy structure [114].

Gold-containing brazing and solder alloys are used for the joining of packaging and for the assembly of components. The eutectic alloys of gold with silicon, germanium, and tin (Table 11) are used for low-temperature solder connections with high mechanical strength. Gold and gold – silicon alloys are used to connect the semiconductor crystals with the base contact

Table 11. Low-melting gold solder alloys for electronic applications , p. 556[22]

Composition	Melting range, °C	Elongation, %	Thermal expansion, $10^{-6}\,K^{-1}$	Electrical resistivity, $\Omega\,mm^2\,m^{-1}$	Thermal conductivity, $W\,m^{-1}\,K^{-1}$
AuSn20	280	<1	16	>0.2	57.3
AuGe12	356	<1	13.4	0.14	44.4
AuSi2	363 – 740	0.5 – 3 ∗	12.6	0.031	50

∗ Hard-rolled condition.

supports (die-attach) on lead frames, packages, or hybrid circuits.

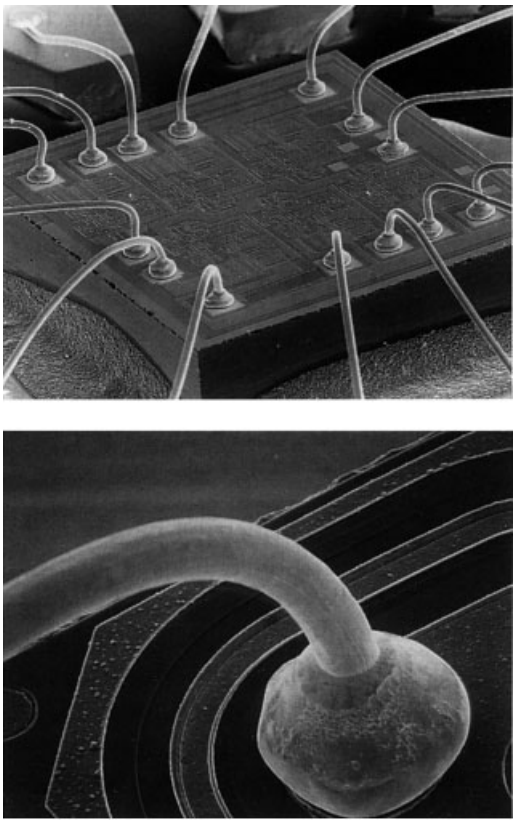

Figure 30. Ball bonding of semiconductor chips with gold bond wire [36, p. 561]

Gold-plated metallic lids and solder preforms of the alloy AuSn20 are used for the hermetic (vacuum-tight) sealing of semiconductor device packages (Fig. 31) and for soldering of laser crystals. AuSn alloy connections are also formed by thermal bonding of Sn-plated leads onto Au bumps, placed on the surface of the semiconductor crystal with the tape-automated-bonding (TAB) technology [144]. AuSi alloys give solder bonds with the highest stability against thermal and mechanical stress and are therefore preferably applied in high-reliability semiconductor devices submitted to thermal cycling. AuGe12 and Au70In20Sn10 can be soldered at 400 – 500 °C. They are used for the soldering of Mo supports in power devices, packages, and components and leads of Cu and FeCoNi alloys.

Ternary solder alloys of AuAgGe, AgCuSn and AgCuIn of different compositions cover the temperature range of 400 – 650 °C. Some of these are brittle and must be applied in powder or paste form.

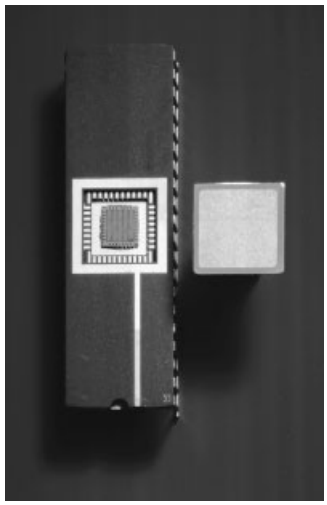

Figure 31. Flat pack ceramic package for integrated circuit devices with gold contact frame (with connections in dual inline array). Hermetic encapsulation by soldering of gold plated metal lid with AuSn20 solder frame.

Nanosized gold particles have potential applications in electronics, computing, sensors, and catalysis. On the nanometric scale, materials properties are strongly size dependent. Massive gold clusters, which consist of 1 – 2 nm sized gold cores encapsulated within a shell of tightly packed hydrocarbon chains have been prepared by decomposition of gold thiolate com-

pounds. The electromagnetic and conduction properties of these cluster molecules are extremely sensitive to charging and are of potential interest for use in microelectronic circuitry ("single electronics") [145–149]. Vapor-synthesis techniques have been used to prepare solvent-stabilized gold particles in the 1 – 3 nm regime. The clusters may provide active catalysts for the low-temperature oxidation of CO to CO_2 [150], [151]. Long "wires" of gold nano particles have been deposited along step edges of highly oriented pyrolytic graphite. The technique is supposed to perform in this way the connection between components in a nanoscale electrical circuit [152]. Spacing of nanoparticles of gold has been also performed by attaching single stranded DNA oligonucleotides of defined length and sequence to the nanoparticles [153], [154], [155].

10.3.2. Electrical Engineering [22, pp. 449 – 459]

Because of its high oxidation resistance, gold surfaces enable the formation of electrical connections with the lowest ohmic resistances. The surface layers are made by roll-bond cladding or electrodeposition. Gold and gold alloys are indispensable for conductor material and low-voltage contacts for communication and information transfer equipment, where even very thin oxide layers would cause interruptions or failures in signal transfer.

Gold – nickel and gold – silver alloys are used in weak-current engineering, as contact materials for very low voltage switches, and where the contact forces are low (relays, plugs, measuring instruments). Micromigration of these alloys is very low, and there is little tendency for insulating layers to form.

AuCr2 serves as resistor material in corrosive environments with stable temperature coefficient of resistance (TCR) between − 20 and + 40 °C.

Wires of multicomponents alloys based on gold and palladium serve as potentiometer windings in high precision techniques, where the lowest TCR values in particular temperature ranges are needed.

Gold-manganese alloys are used for wirewound resistance thermometers. The thermocouples Cu + Au 99,4/Co 0,6 and Pt 90/Ir 10 + Au 60/Pd 40 are suitable for the temperature ranges 0 to − 240 °C and 0 to 700 °C, respectively; the gold alloys form the negative leg. Au – Ag – Cu-alloys are used as surface material in sliding contacts.

10.4. Solders [22, pp. 492, 561 – 562], [156], [157, pp. 157 – 163]

A variety of gold-based solder alloys are used in the fields of apparatus technology, electronics, jewelry, and dentistry.

Certain types of apparatus have components made of iron and nickel alloys which have to withstand high vacuums and high temperatures. To join these materials, vacuum hard solders are used, made of either fine gold, or gold – copper, gold – silver – copper, gold – nickel, gold – copper – nickel, and gold – palladium alloys. Soldering is carried out in a vacuum furnace or in a protective gas (hydrogen, cracked gas).

Hard silver solder is normally used to join stainless steels. However, if these joints do not exhibit sufficient corrosion resistance, gold – nickel – zinc alloys similar to white gold, with ca. 80 % gold, 15 % nickel, and 5 % zinc, are occasionally used.

In the fabrication of electronic devices, gold based alloys of the composition gold – tin (25 wt % Sn, *mp* 280 °C), gold – silicon (3.1 wt % Si, *mp* 370 °C), and gold – germanium (12 wt % Ge, *mp* 356 °C) are used.

They allow soldering at relatively low temperatures and give connections of high mechanical strength, good corrosion resistance and thermal conductivity (see also Section 10.3.1).

In jewelry, individual chain links, setting for stones, and hollow-ware parts are soldered with solder alloys which must correspond in cartage, color, and mechanical properties to the materials to be joined, but have melting temperatures that are sufficiently lower to avoid melting of the noble metal pieces [158]. These solder and brazing alloys are principally of the same or similar composition to the metal parts to be joined. For

consecutive solder operation, alloys with different melting temperatures (first, second, third solders) are applied.

Gold-based solder alloys are used in dentistry to join prosthetic dental pieces. Alternative methods are microplasma welding and laser welding. The solder alloys are of similar compositions to the dental alloys. They must correspond to the base materials in their corrosion behavior and thermal expansion. Solders for crowns and bridges melt in the temperature range of around 760 – 840 °C. They are based on the Au – Ag – Cu system with additions of zinc and tin. Melting is usually done by gas torch. Solders for PFM alloys (see also → Dental Materials, Chap. 2.3.) are used before porcelain veneering. The melting temperatures and softening ranges must be sufficiently high so that subsequent baking does not influence the mechanical stability of the solder joint. These solders are based on the Au – Ag system with reduced palladium and platinum content. Color lightening is effected by palladium additions, and the melting temperatures are adjusted by additions of tin, indium, or zinc [159].

Joining of gold and gold alloys at considerably lower temperatures can be performed by using diffusion processes, for instance by using tin foil between components of pure gold and heating the assembly above the melting point of tin but below the melting temperature of gold [160].

10.5. Pen Nibs

Nibs for fountain pens are usually made from Au – Ag – Cu yellow gold alloys which are relatively hard; occasionally white gold alloys are also used. The alloys must be able to withstand the very corrosive ferro-gallic inks; only 14 or 18 carat alloys are suitable. The nib points must be made of hard metal alloys, which usually contain Ru, Os, Ir, W, or Co.

Nibs made of stainless, ink-resistant nickel chromium steel are sometimes coated with a thin layer of gold; however, this does not improve their ink resistance.

10.6. Chemical Technology [161]

Gold – platinum alloys containing 50 – 70 % gold, which can be age-hardened, are used to make spinnerets used in the production of man-made fibers. Their fine-grained structure is of great advantage in making the necessary fine holes (diameter 25 – 120 μm).

Gold alloys are sometimes used for seals and rupture disks that come into contact with corrosive substances. A gold – silver – palladium alloy (Pallacid) containing 30 % gold and 30 % palladium is resistant to strong mineral acids, is considerably cheaper than gold, and also has greater high-temperature strength.

A gold alloy containing 10 % platinum is used to make crucibles for analytical laboratories, e.g., for ash determination of flour and other phosphorus-containing foods. Unlike platinum crucibles, it is resistant to corrosion by phosphorus compounds when red-hot.

10.7. Dental Materials

Gold alloys are of great importance in prosthetic dentistry, for restorations such as gold fillings, crowns, bridges, cast dentures, clasps, anchorage pins, and metal frameworks for dental ceramics (→ Dental Materials). These materials have to meet a number of requirements. They must be resistant to normal conditions in the mouth, of a suitable color, of different strengths, and be easy to work. Today almost all alloys used are of a complex composition, containing high proportions of gold, palladium, and platinum. In the past few years the number of components is decreasing because of allergenic problems. The advantage of these alloys mainly lies in their very fine-grained and homogeneous structure.

10.8. Coatings [162–168]

10.8.1. Electroplating and Electroforming

The technical and decorative properties of gold can be combined with a variety of cheap substrates by applying a thin layer of gold to base metals, ceramics, glass, or plastics.

Electroplating is by far the most frequently used method (→ Electrochemical and Chemical Deposition, Chap. 8.9.). Modern gold electroplating is mainly gold alloy plating for the following reasons:

1) For the adjustment of color. Gold alloys can be deposited in a variety of shades. For example, co-deposition of copper gives a reddish color, nickel a white color, and cadmium a yellow color. Colors of gold alloys are standardized in DIN 5033 and DIN EN 28 654.
2) To improve hardness and wear resistance.
3) In gold alloy deposits part of the gold is replaced by alloying elements which in most cases have a lower density. Thus, substantial savings of gold may be achieved.

Gold coatings are deposited mainly from solutions of potassium dicyanoaurate(I), $K[Au(CN)_2]$. There are three groups of plating solutions:

1) Alkaline cyanide electrolytes, which enable the dissolution of Cu, Ag, Cd and other metal salts and their co-deposition with gold.
2) Neutral electrolytes have been developed for lower carat deposits and higher thicknesses. These electrolytes are also used for pure gold deposits with high thicknesses.
3) Acidic electrolytes (pH 3 – 5) for producing high-carat deposits with properties required by the electronics industry and also for decorative purposes. These electrolytes are based on the finding that solutions of $K[Au(CN)_2]$ are stable at a pH value as low as 3.2. Deposits can be hardened by co-deposition of elements such as Ni, Co, or Fe.

An electrolyte based on the cyanide complex of trivalent gold, $K[Au(CN)_4]$ has been disclosed. The electrolyte is operated at low pH (0.2 – 1.5) and is suitable for difficult-to-plate metals, e.g., stainless steels [164].

Non-cyanide electrolytes contain gold mostly as the sulfito complex $[Au(SO_3)_2]^{3-}$. These plating solutions have a very good throwing power and result in bright crack-free thick deposits with excellent ductility. However, wear resistance is poor. The main drawback of the sulfite electrolytes is the instability of the solution.

Properties of gold coatings which are relevant to their applications are: color, brightness, hardness, wear resistance, porosity, and corrosion resistance. Gold and gold alloy coatings are produced in a wide range of thicknesses (0,1 – 200 µm), depending upon their application.

Electroless or autocatalytic gold-plating solutions usually contain potassium dicyanoaurate(I) together with a reducing agent such as sodium hypophosphite, sodium borohydride, or dimethylamine borane (DMAB). Electroless gold plating is used only where electrolytic processes cannot be used. Typical fields of application are the electronics industry, e.g., in printed circuit boards and in semiconductor electronics [165].

Electroforming (→ Electrochemical and Chemical Deposition) has been defined as "the production or reproduction by electrodeposition of an article on a mandrel or mold, that is subsequently separated from the deposit". Gold up to a thickness of 0.2 mm is deposited on a mandrel, which is then removed. Wax mandrels can be melted out. A self-supporting gold layer is left which can be reinforced by a filler material. A process for depositing 8-, 9-, 14-, or 18-carat gold – silver alloy coatings has been developed [166], [167]. The process of electroforming is also used for manufacturing high-technology equipment [168].

Table 12. Classification of precious metal preparations

Appearance	Consistency	
	liquid/pasty	*solid*
glossy	bright gold bright lemon gold bright palladium bright platinum lusters	
matt	burnish gold burnish palladium burnish platinum	dusting gold dusting white gold dusting silver

10.8.2. Bright Gold

Ceramic materials, especially high-quality porcelain and glass, are often gilded by firing on preparations of bright gold and burnish gold (Table 12, see also → Ceramic Colorants). The essential components of these lacquer-like paints

are gold sulforesinates, mixed with natural oils and resins. The preparations are applied mainly either by hand or by screen printing processes. Firing is carried out at 500 – 1250 °C. Gold coatings for technological applications also may be applied by this method.

Precious metal (PM) preparations for the decoration of ceramic (CerDecOr, dmc^2) materials are classified by:

- Consistency: liquid/pasty or solid
- Appearance: glossy or matt
- Precious metal content: gold, platinum, palladium or silver preparations.

Bright PM preparations are varnishlike organic compounds containing precious metals, combined with flux components (based on organometallic compounds) as adhesion promoters. These preparations also contain resin solutions as film formers. After firing, a high-gloss precious metal film forms on a smooth surface. According to the color shade of the fired product, a distinction is made between:

- *Bright gold*, which develops a typical reddish yellow gold shade after firing and contains a small proportion of silver.
- *Bright lemon gold*, which fires to a lemon-yellow shade, achieved by formation of an gold – silver alloy.
- *Bright platinum* and *bright palladium*, which fire to a white-gold shade, achieved by formation of an Au – Pt or Au – Pd alloy.

Burnish PM preparations are varnishlike organic compounds containing precious metals combined with flux components based on organic metal compounds. The flux compounds promote adhesion, and resin solutions as film formers. A dull precious metal film is formed after firing, which develops a satin surface during subsequent burnishing with a glass-fiber brush, sand, or other compacting materials. Burnish PM preparations are subdivided into:

- *Burnish gold*, which often contains a small portion of silver in addition to gold, enabling the color shade to be varied from reddish yellow to yellow to greenish yellow, depending on the alloy.
- *Burnish platinum* and *burnish palladium*, which fire to a white gold shade, achieved by the formation of an Au – Pt or Au – Pd alloy.
- *Burnish silver*, which is mainly used for the decoration of gift articles and contain only silver as its only precious metal.

To meet the requirements of differing production techniques, two types of *dusting PM preparation* are available, which differ in their particle morphology: **leaf gold flakes** consist of precious metals or their alloys in form of very finely flaked particles. **Painter's gold** is a chemically precipitated, crystalline precious metal powder with mostly spherical particles:

- *Dusting gold* has a typical reddish yellow shade after firing. Preparations with yellow shades can also produced by alloying with silver.
- *Dusting white gold* achieves its color by alloying with platinum and/or palladium.
- *Dusting silver* contains silver which may be modified with a slight fraction of gold.

Dusting PM preparations are primarily used to produce high-grade decals. Decorations produced with dusting gold (leaf gold flakes), dusting white gold, and dusting silver have a permanent attractive metallic appearance without after-treatment. Decorations produced painter's gold (used to hand paint gold decorations) have to be burnished. Dusting silver trends to tarnish after some time.

Lusters are varnishlike preparations, based on metal- and/or precious-metal-containing organic compounds, with resin solutions as film formers. When fired on porcelain, earthenware, enamel, tiles or glass they produce very thin, intensively colored and lustrous layers, which give them their metallic iridescent finish. They are classified according to color shade, precious metal content, and special effects, as indicated by their classification as craquelling or marbling lusters, running lusters, and halo lusters.

Table 13 lists the characteristics of different preparation ranges.

The precious metal content influences mainly the color of the preparations. The differentiation within the range of bright preparations is summarized in Table 14.

Table 13. Characteristics of different preparation ranges

Type of preparation Parameter	Lustres	Bright PMP	Burnish PMP	Matt PMP	Dusting PMP
Surface	glossy, iridescent	glossy	matt after polishing	matt without polishing	semi matt
Condition of the components	liquid	liquid	partly solid	partly solid	solid
Stirring before use	no	no	yes	yes	no
Gold content, %	0 – 6	6 – 15	16 – 40	9 – 15	80 – 100

Table 14. Colors of PM Preparations according to precious metal content

Color	Au, %	Ag, %	Pd, %	Pt, %	Name
Reddish yellow	100				Bright Gold
Yellow	90	10			Bright Gold
Greenish yellow	70	30			Bright Lemon Gold
White	70	10	20		Bright Palladium
White	70	10	10	10	Bright Palladium

10.8.3. Other Gold Coatings

Rolled gold is still used for spectacle frames and gold-plated watches. However, it has lost its former significance in favor of electroplating. Rolled gold is fabricated by soldering gold sheets, usually 14 carat, to blocks made of copper alloy or stainless steel. These are then rolled or drawn to the desired shape, producing strips and wires with a largely non-porous gold coating, usually 5 – 10 µm thick.

Fire Gilding. In this process, the part to be gilded is painted with gold amalgam. The mercury is evaporated by heating, leaving a relatively thick gold layer. The process is very problematical with regard to industrial hygiene.

Vapor Deposition of Gold. To achieve a sufficiently abrasion resistant coating of true gold color, a base layer of TiN is applied which is topped with less than 1 µm of gold [169]. In microelectronics gold layers are deposited for contacts or leads on hybrid ICs by sputtering or evaporation. Sometimes architectural glass is coated with gold instead of silver (reference to low emissivity coatings with Ag in the Ag-chapter). This combines the solar control function with a spectacular decorative effect. Heat protection by thin gold layers also is used in space technology, e.g., in space suits.

10.9. Gold Leaf

Gold leaf is usually made of fine gold or Au – Ag – Cu alloys with a very high gold content. Thin foils are rolled in a package arrangement. Production starts with sheets of about 20 µm thickness. These are placed between sheets of vellum, beaten down, and finally hammered by hand, using separating membranes made of the cecum of cattle ("goldbeater skins") to a thickness of about 0.2 µm [170]. A modern method of making gold leaf is cathode sputtering.

The crystal plane (100) in beaten gold leaf lies in the plane of the leaf. Gold leaf is used to gild wooden statues, book edges, and fabric prints. In certain beverages (Danziger Goldwasser) foods, it is used to achieve visual effects.

In former times, small rolls of gold leaf were used to fill cavities in teeth.

10.10. Catalysts

Gold shows activity in both heterogeneous and homogeneous catalysis.

Gold in platinum group metal and in silver-based catalysts acts as promoter to effect higher yields and higher selectivity. Hydroforming of paraffins and cycloalkanes with gold-containing iridium catalysts gives higher yields of aromatic

compounds at temperatures above 500 °C. Another example is the acetoxylation of ethene to form vinyl acetate over Pd – Au catalysts [171].

Supported active gold catalysts are obtained by coprecipitation or deposition precipitation of gold from precursor compounds on to transition metal oxides such as Mn_2O_3 or Fe_2O_3. The optimum conditions for functionality are gold particle sizes of 2 – 5 nm, which may interact with oxide support that can itself in some way participate in the reaction [172]. They surpass the activity of mercury catalysts for the hydrochlorination of ethyne and are especially suited for the catalytic activated low-temperature oxidation of carbon monoxide [173]. The method of coprecipitation, followed by aging of the gold containing precipitate, which changes thereby from the amorphous to crystalline structure, is important for sustained high activity.

Gold catalysts are also promising in both the selective and nonselective oxidation of hydrocarbons, in the reduction of nitrogen oxides by carbon monoxide, propene, or hydrogen, in the hydrogenation of carbon oxides [172], and in the epoxidation of propane with O_2/H_2 mixtures [174].

Gold on carbon catalysts have been used for the selective liquid-phase oxidation of diols with dioxygen in alkaline solution under mild conditions (343 – 363 K). As with gold catalysts in the gas phase, the gold particle size is important. Particle size was 7 nm, the catalyst was prepared by deposition precipitation from $H[AuCl_4]$ and sodium carbonate, followed by reduction with formaldehyde [175].

Homogeneous catalytic action is shown by cationic gold(I) phosphine complexes in the addition of alcohols to alkynes. These gold catalysts are also significantly better than the previously used mercury catalysts and work under mild conditions (20 – 50 °C) in the presence of acidic co-catalysts [176], [177].

Platinum vapor, which forms from platinum – rhodium catalysts during the oxidation of ammonia to nitric acid can be retained by a gold or gold – palladium gauze. This process has long been in industrial use.

11. Economic Aspects [2], [4], [18–20], [39], [40], [47], [178–180]

Production. Gold represents a significant portion of the total world economy. In the Republic of South Africa, it occupies in terms of value the highest position among all industrial goods. About 500 000 people are employed worldwide in gold production 300 000 – 400 000 of them in the Republic of South Africa. The value of the gold from primary production in 1997 was about € 20×10^9. Thus it occupies third place in the metal trade, after pig iron (€ 100×10^9) and aluminum (€ 23×10^9), and before copper (€ 11×10^9) [178]. Global gold mine production grew from 390 t in 1900 to 2576 t in 1999 [178].

At present, almost half of the gold obtained from ores is produced in the Republic of South Africa, The United States, Canada, and Australia. Table 15 shows the primary production of gold in the various producer countries [179].

Table 15. Global mine production of Gold 1998 and 1999 in t [179]

Country	1998	1999
South Africa	464	450
United States	366	342
Australia	310	303
Canada	164	158
China	160	156
Indonesia	139	155
Russia	127	138
Peru	91	127
Uzbekistan	80	86
Ghana	73	78
PNG	63	65
Brazil	55	54
Chile	47	48
Argentina	20	35
Philippines	35	32
Zimbabwe	27	30
Mali	23	25
Mexico	26	23
Colombia	22	21
Kyrgyzstan	21	20
Rest of World	226	231
Total	2541	2576

Supply and Demand. The supply of gold on the market in 1997 was about 2700 t/a (Fig. 32). Most of this (73 %) came from total world mine production, 20 % came from recycling processes, and 7 % was sold by the former Soviet Union. Buying and selling on the part of government monetary authorities can have a strong effect on the market; in recent years official sector sales have increased significantly.

Figure 33 shows the demand for gold according to industrial applications.

Gold Trade. London is the most important market for gold in terms of the quantity traded. The London gold market has a fixing system, whereby the price of gold is determined twice daily, according to supply and demand, by representatives from five firms of brokers. This is then published as the "fixing price". The United States futures markets have a strong influence on the calculation of this price. The fixing price serves as an orientation point for gold traders all over the world.

Zurich is also an important center of the gold trade. A considerable part of the gold from South African primary production, which up to 1968 all went to London, is now sold to Zurich, as well as most of the gold from the Soviet Union.

Another important center is New York. Other metal exchanges, such as Hongkong, Frankfurt, Paris, Singapore, and Tokyo, are of more regional importance.

Gold is traded internationally in the form of gold bars of various weights, and with a gold content of 99.50 – 99.99 %. In London, gold bullion must contain gold weighing between 350 and 430 oz; the commonest figure is about 400 oz (ca. 12 kg).

Important factors which influence the price of gold are assessment of political circumstances, inflation rates, and the rate of interest of the US dollar as a key world currency. The huge gold holdings of central banks and several central bank disposals had been weighing on the price of gold in 1997 and 1998 together with increasing amounts of gold forward sales of producers. On August 28, 1998, the price of gold was fixed at $ 273.40 per ounce – at that time a fresh 19 year low.

12. Toxicology and Occupational Health

Gold and its soluble compounds generally only seldom cause acute or chronic problems in the workplace and in the environment. Tetrachloroauric acid $H[AuCl_4]$ [*16903-35-8*] as is a strong acid with corrosive effects on the skin. Gold(III) salts are more poisonous than gold(I) compounds. After intake of gold(III) chloride lethal poisonings with toxic effects on liver and kidneys are observed. Potassium dicyanoaurate $K[Au(CN)_2]$ [*14265-59-3*] liberates hydrocyanic acid, which can cause poisoning. Poisonings by high doses of other gold compounds are also known [181], [182].

The daily intake of gold in food is estimated at less than 7 µg. Usual concentrations in blood are 15 µg/L [183] or over less than 0.5 µg/L, with an average of 0.1 µg/L [184].

Gold Therapy. Gold compounds are used in the treatment of chronic rheumatoid arthritis and of pemphigus vulgaris in dermatology. A 30 % incidence of side effects is known [185]. Therapeutic effects and toxicity correspond with water solubility [186]. Medical treatment uses weakly soluble gold(I) thiolates such as Myocrysin (disodium gold(I) malate [*42722-04-3*]) or Solganol or Aureotan (gold(I) thioglucose [*12192-57-3*]). The drug is administered by intramuscular injection in relatively high doses. Toxic side effects include cardiovascular disorders, nerve dysfunction and adverse effects on liver, stomach, and bone marrow. Allergic effects occur on the skin, mucous membranes, and, rarely, in the lungs. The development of drug-induced lichen planus is also possible [185].

A possible consequence of chronic gold application is chrysiasis, a blue gray discoloration of the skin, similar to argyria [185].

Gold Allergy. Skin allergies to gold jewelry or dental gold have rarely been observed. In many cases they could be attributed to less precious alloying elements such as nickel, cobalt, or copper.

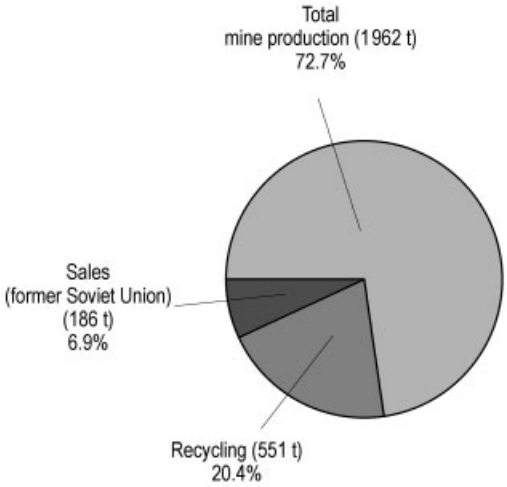

Figure 32. Gold supply in 1997 (ca. 2700 t)
(source: Bericht der Degussa Ffm Edelmetallmärkte 1997)

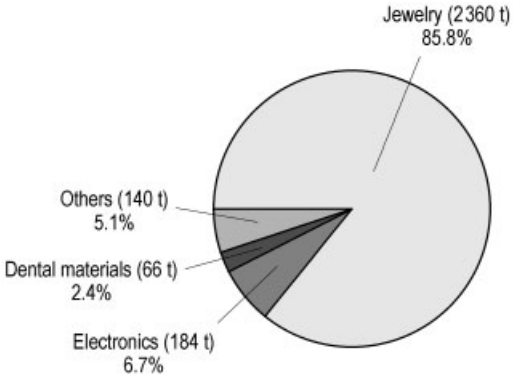

Figure 33. Gold industrial demand in the world in 1997 (ca. 2750 t)
(source: Bericht der Degussa Ffm Edelmetallmärkte 1997)

13. References

1. *Gmelin,* System no. 62.
2. H. Quiring: *Die metallischen Rohstoffe,* 2nd ed., vol. 3 "Gold," Enke Verlag, Stuttgart 1953.
3. Degussa, *Sol & Luna – auf den Spuren von Gold und Silber,* Frankfurt 1973.
4. B. Neumann: *Die Metalle – Geschichte, Vorkommen und Gewinnung nebst ausführlicher Produktionsstatistik,* Verlag Wilhelm Knapp, Halle 1904.
5. H. Moesta: *Erze und Metalle – ihre Kulturgeschichte im Experiment,* Springer Verlag, Berlin 1983.
6. H. Römpp: *Chemie der Metalle,* Franckh'sche Verlagshandlung, Stuttgart 1941.
7. H. W. A. Sommerlatte: *Gold und Ruinen in Zimbabwe – aus Tagebüchern und Berichten des Schwaben Karl Mauch* (1837 – 1875), Verlag Bertelsmann Fachzeitschriften GmbH, Gütersloh 1987.
8. A. Hartmann, E. Sangmeister: "Zur Erforschung urgeschichtlicher Metallurgie," *Angew. Chem.* **84** (1972) 668 – 678.
9. K. Löhberg: "Beiträge der Metallkunde zur Archäologie," *Jahrbuch 1979 der Berliner Wissenschaftlichen Gesellschaft e. V.,* p. 48.
10. G. Biegel (ed.): *Das erste Gold der Menschheit,* Verlag Karl Schillinger, Freiburg 1986.
11. F. Kirchheimer: "über das Rheingold," *Jahreshefte des geologischen Landesamts Baden Württemb.* **7** (1965) 77.
12. *World Gold Council,* Gold Bulletin – Quaterly review of research on gold and its applications

in industry, vol. 1 (1968) – vol. 31 (1998), Geneva.
13. A. Binz: *Edelmetalle – ihr Fluch und Segen,* Limpert-Verlag, Berlin 1943.
14. H. Pohl: *Gold – seine Macht und Magie in der Geschichte,* W. Heyne Verlag, München 1958.
15. C. H. V. Sutherland: *Gold – Macht, Schönheit und Magie,* Verlag A. Schroll & Co., Wien – München 1970.
16. G. Breitling et al.: *Das Buch vom Gold,* Verlag C. J. Bucher, Luzern – Frankfurt 1975.
17. H. G. Fuchs: *Gold,* Fritz Knapp Verlag, Frankfurt 1981.
18. P. Vilar: *Gold und Geld in der Geschichte,* Verlag H. C. Beck, München 1984.
19. H. J. Funk (ed.): *Gold,* Edition Deutsche Bank, Battenberg Verlag, München 1986.
20. Fonds der Chemischen Industrie, *Edelmetalle – Gewinnung, Verarbeitung, Anwendung,* Folienserie, Frankfurt 1989.
21. L. F. Trueb: *Gold – Bergbau, Verhüttung, Raffination und Verwendung,* Verlag Neue Zürcher Zeitung, Zürich 1992.
22. DEGUSSA AG (ed.): *Edelmetall-Taschenbuch,* Hüthig-Verlag, Heidelberg 1995.
23. L. F. Trueb: *Die chemischen Elemente,* S. Hirzel Verlag, Stuttgart – Leipzig 1996.
24. G. Biegel: *Das Erste Gold der Menschheit – die älteste Zivilisation in Europa,* Verlag Karl Schilliger, Freiburg 1986.
25. A. Neuburger: *Die Technik des Altertums,* Voigtländer Verlag, Leipzig 1920, Prisma-Verlag, Gütersloh 1983.
26. J. H. F. Notton, "Ancient Egyptian Gold Refining," *Gold Bulletin* **7** (1974) 50 – 56.
27. V. Tafel: *Lehrbuch der Metallhüttenkunde,* vol. 1, S. Hirzel Verlag, Leipzig 1951.
28. Degussa, *Edelmetall-Taschenbuch,* Frankfurt 1967.
29. Doduco, *Datenbuch,* Pforzheim 1974.
30. R. C. Weast: *Handbook of Chemistry and Physics,* 69th ed., CRC-Press, Cleveland 1989.
31. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* Springer Verlag, Berlin 1967.
32. D. G. Cooper: *Das Periodensystem der Elemente,* Taschentext Nr. 6, Verlag Chemie, Weinheim 1972.
33. G. H. Aylward, T. J. V. Findlay: *Datensammlung Chemie,* Taschentext Nr. 27, Verlag Chemie, Weinheim 1975.
34. E. Fluck, K. G. Heumann: *Periodic Table of the Elements,* VCH Verlagsgesellschaft, Weinheim 1986.
35. E. M. Wise: *Gold – Recovery, Properties and Application,* Van Nostrand, New York 1964.
36. G. Schlamp: "Noble Metals and their Alloys," in *Materials Science and Technology,* **vol. 8,** Chap. 9, VCH Verlagsgesellschaft, Weinheim 1996.
37. M. Fieberg, H. Renner, G. Schlamp: "Silver, Gold, and other Noble Metals," *Encyclopedia of Applied Physics,* **vol. 18,** VCH Verlagsgesellschaft, Weinheim 1997.
38. C. E. Wicks, F. E. Block: *Thermodynamic Properties of 65 Elements – their Oxides, Halides, Carbides, and Nitrides,* Bureau of Mines Rep. No. 605. Washington, D.C. 1963.
39. W. Gocht: *Handbuch der Metallmärkte,* 2nd ed., Springer Verlag, Berlin 1985.
40. R. Saager: *Metallische Rohstoffe von Antimon bis Zirkonium,* Bank Vontobel, Zürich 1984.
41. J. Lurie: *South African Geology for Mining, Metallurgical, Hydrological and Civil Engineering,* McGraw-Hill, Johannesburg 1981.
42. C. B. Coetzee: *Mineral Resources of the Republic of South Africa,* Pretoria, Department of Mines 1976.
43. H. Schneiderhöhn: *Erzlagerstätten,* G. Fischer Verlag, Stuttgart 1962.
44. R. Jackson: *Ok Tedi – The Pot of Gold,* World Publishing Co., Boroko, Papua – New Guinea 1982.
45. *Kirk-Othmer,* 3rd ed., **11,** pp. 972 – 995.
46. F. Pawlek: *Metallhüttenkunde,* de Gruyter, Berlin 1983.
47. *Winnacker-Küchler,* 3rd ed., **4,** 544 – 547.
48. *Ullmann,* 3rd ed., **8,** pp. 269 – 284.
49. G. Agricola, *Vom Berg- und Hüttenwesen,* DTV-Verlag, München 1980,
50. D. Meadows, E. Zahn, P. Milling: *Die Grenzen des Wachstums,* Deutsche Verlagsanstalt, Stuttgart 1972.
51. R. J. Adamson (ed.): *Gold Metallurgy in South Africa,* Chamber of Mines, Johannesburg 1972.
52. J. V. N. Dorr, F. L. Bosqui: *Cyanidation and Concentration of Gold and Silver Ores,* McGraw-Hill, New York 1950.
53. A. King: *Gold Metallurgy on the Witwatersrand,* Chamber of Mines, Johannesburg 1949.
54. D. M. Liddell: *Handbook of Nonferrous Metallurgy,* McGraw-Hill, New York 1945.
55. T. K. Rose, W. A. C. Newman: *The Metallurgy of Gold,* Griffin, London 1937.

56. G. G. Stanley (ed.): *The Extractive Metallurgy of Gold in South Africa,* The South African Institute of Mining and Metallurgy, Johannesburg 1987. E. Krone: "Probleme der Primärgoldlaugung", *Erzmetall* **42** (1989) 253 – 258.
57. E. A. Cho, S. N. Dixon, C. H. Pitt: "The Kinetics of Gold Cyanide Adsorption onto Activated Charcoal," *Metall. Trans. B* **10 B** (1979),185 – 189.
58. A. S. Dahya, D. J. King: "Developments in Carbon-in-Pulp Technology for Gold Recovery," *CJM Bull.* **76** (1983) no. 857, 55 – 61.
59. C. A. Fleming: "Some Aspects of the Chemistry of Carbon-in-Pulp and Resin-in-Pulp Processes," in *papers presented at The Australasian Institute of Mining and Metallurgy, Seminar on Carbon-in-Pulp Technology for the Extraction of Gold,* Kalgoorlie (Australia) July 1982.
60. C. A. Fleming in K. Osseo-Assare, J. D. Miller (eds.): *Hydrometallurgy: Research, Development and Plant Practice,* "Recent Developments in Carbon-in-Pulp Technology in South Africa," The Metallurgical Society of AIME, Warrendale (U.S.A.) 1983, pp. 839 – 858.
61. C. A. Fleming, M. J. Nicol: "The Adsorption of Gold Cyanide onto Activated Carbon. III. Factors Influencing the Rate of Loading and the Equilibrium Capacity," *J. S. Afr. Inst. Min. Metall.* **84** (1984) no. 4, 85 – 93.
62. C. A. Fleming, M. J. Nicol, D. I. Nicol: "The Optimization of a Carbon-in-Pulp Adsorption Circuit Based on the Kinetics of Extraction of Aurocyanide by Activated Carbon," *paper presented at the Symposium: Ion Exchange and Solvent Extraction in Mineral Processing held at Mintek,* Randburg Feb.1980.
63. P. A. Laxen, G. S. M. Becker, R. Rubin: "Developments in the Application of Carbon-in-Pulp for Gold Recovery from South African Ores," *J. S. Afr. Inst. Min. Metall.* **79** (1979) no. 11, 315 – 326.
64. P. A. Laxen, T. D. Brown in L. F. Haughton (ed.): *Mintek 50: Proceedings of the International Conference on Mineral Science and Technology,* "The Carbon-in-Pulp Plant at Rand Mines Milling and Mining Company Limited: Problems and Development," vol. 2, Council for Mineral Technology, Randburg 1985, pp. 695 – 706.
65. P. A. Laxen, C. A. Fleming, D. A. Holtum, R. Rubin in H. W. Glen (ed.): *Proceedings, Twelfth Congress of the Council of Mining and Metallurgical Institutions,* "A Review of Pilot-Plant Testwork Conducted on the Carbon-in-Pulp Process for the Recovery of Gold," vol. 2, The South African Institute of Mining and Metallurgy, Johannesburg 1982, pp. 551 – 561.
66. G. J. McDougall, R. D. Hancock: "Gold Complexes and Activated Carbon. A Literature Review," *Gold Bull.* **14** (1981) no. 4, 138 – 153.
67. M. J. Nicol, C. A. Fleming, G. Cromberge: "The Adsorption of Gold Cyanide onto Activated Carbon. I. The Kinetics of Adsorption from Pulps," *J. S. Afr. Inst. Min. Metall.* **84** (1984) no. 2, 50 – 54.
68. M. J. Nicol, C. A. Fleming, G. Cromberge: "The Adsorption of Gold Cyanide onto Activated Carbon. II. Application of the Kinetic Model to Multistage Absorption Circuits," *J. S. Afr. Inst. Min. Metall.* **84** (1984) no. 3, 70 – 78.
69. D. F. Williams, D. Glasser: "The Modelling and Simulation of Processes for the Absorption of Gold by Activated Charcoal," *J. S. Afr. Inst. Min. Metall.* **85** (1985) no. 8, 237 – 243.
70. G. Brauer: *Handbuch der Präparativen Anorganischen Chemie,* **vol. 2, 3rd ed.,** Enke Verlag, Stuttgart 1978.
71. Chamber of Mines (ed.): *Rand Refinery,* P.R.D. Series No. 149, Johannesburg 1970.
72. G. Eger: *Das Scheiden der Edelmetalle durch Elektrolyse,* W. Knapp Verlag, Halle 1929.
73. M. Grehl, H. Meyer: "Electrochemical versus chemical reactions in PM refining," *IPMI, PM* 1997, 13 – 28.
74. M. Fieberg, I. R. Edwards: *The Extraction of Gold from Chloride Solutions,* Report No. 1996, National Institute for Metallurgy, Randburg 1978.
75. M. Lin, S. Zhong, S. Sun, *Kuongye Gongcheng* **15** (1995) no. 2, 37 – 40.
76. W. W. Behning, *Metall (Berlin)* **24** (1970) 794.
77. E. Krone, W. Dähne, *Chem. Ztg.* **101** (1977) 421.
78. G. Cornelius, *Erzmetall* **32** (1979) 467.
79. *Recycling – Rohstoffsicherung und Umweltschutz,* Degussa, Frankfurt (M) 1981.
80. F. H. Burstall, P. J. Forrest, R. A. Wells, *Ind. Eng. Chem.* **45** (1953) 1648 – 1658.
81. K. B. Ball, *World Min.* **27** (1974) no. 11, 44 – 45.
82. R. J. Puddephatt: *The Chemistry of Gold,* Elsevier, Amsterdam 1978.

83. G. Wilkinson, R. D. Gillard, J. A. McCleverty: *Comprehensive Coordination Chemistry,* vol. 5, Pergamon Press, Oxford 1987.
84. G. Wilkinson, F. G. A. Stone, E. Abel: *Comprehensive Organometallic Chemistry,* vol. 2, Pergamon Press, Oxford 1982.
85. N. N. Greenwood, A. Earnshaw, *Chemistry of the Elements,* Pergamon Press, Oxford 1984; German: Chemie der Elemente, VCH-Verlagsgesellschaft, Weinheim 1988.
86. N. P. Finkelstein, R. D. Hancock: "A New Approach to the Chemistry of Gold," *Gold Bull.* **7** (1974) 72 – 77.
87. A. Prior: "Possibilities of Using Ion Exchange Membrane in the Recovery and Processing of Precious Metals," *Proceedings of the 11th International Precious Metals Institute Conference,* Brussels 1987.
88. F. Simon, *Galvanotechnik* **78** (1987) 2450 – 2457.
89. R. V. Parish: "Organogold Chemistry I: Structure and Synthesis," *Gold Bull.* **30** (1997) 3 – 12; "Organogold Chemistry II: Reactions," *Gold Bull.* **30** (1997) 55 – 62; "Organogold Chemistry III: Applications," *Gold Bull.* **31** (1998) 14 – 21.
90. S. P. Fricker, *Gold Bull.* **29** (1996) no. 29, 53 – 60.
91. S. L. Best, P. J. Sadler, *Gold Bull.* **29** (1996) no. 3, 87 – 93.
92. E. Raub: *Die Edelmetalle und ihre Legierungen,* Springer Verlag, Berlin 1940.
93. E. Jäneke: *Kurzgefaßtes Handbuch aller Legierungen,* C. Winter Verlag, Heidelberg 1949.
94. J. L. Haughton, A. Prince: *The Constitution Diagrams of Alloys. A Bibliography,* Institute of Metals, London 1956.
95. A. Prince: *Multicomponent Alloy Constitution, Bibliography 1955 – 1973,* Metals Society, London 1978.
96. A. Prince: *Multicomponent Alloy Constitution, Bibliography 1974 – 1977,* Metals Society, London 1981.
97. W. Hume-Rothery, G. V. Raynor: *The Structure of Metals and Alloys,* Institute of Metals, London 1956.
98. M. Hansen: *Der Aufbau der Zweistoff-Legierungen,* Springer Verlag, Berlin 1936.
99. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* 2nd ed., McGraw-Hill, New York 1958.
100. R. P. Eliott: *Constitution of Binary Alloys,* 1st Suppl., McGraw-Hill, New York 1965
101. F. A. Shunk: *Constitution of Binary Alloys,* 2nd Suppl., McGraw-Hill, New York 1969.
102. W. G. Moffatt: *The Handbook of Binary Phase Diagrams,* General Electric Company, Schenectady, N.Y. 1976/77 (halbjährliche Nachträge).
103. G. Petzow, G. Effenberg: *Ternary Alloys,* **vol. 1, 2,** VCH-Verlagsgesellschaft, Weinheim 1988.
104. W. Hume-Rothery, G. W. Mabbot, K. M. Channel-Evans, *Phil. Trans. R. Soc.* **A 233** (1934) 1.
105. L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (eds.): *Precious Metals Science and Technology,* Int. Prec. Met. Inst. (IPMI), Allentown 1991.
106. H. Sato, R. S. Toth in T. B. Massalski (ed.): *Alloy Behavior and Effects in Concentrated Solvent Solution,* Gordon & Breach, New York 1964, p. 295.
107. G. V. Raynor: "The Alloying Behaviour of Gold Part I: Solid Solutions", *Gold Bull.* **9** (1976) 12 – 19; "The Alloying Behaviour of Gold Part II: Compound Formation," *Gold Bull.* **9** (1976) 50 – 54.
108. W. S. Rapson: "Intermetallic Compounds of Gold," *Gold Bull.* **29** (1996) no. 4, 141 – 142.
109. D. M. Jacobson, M. R. Harrison, S. P. S. Sangha: "Stable Strengthening of 990-Gold," *Gold Bull.* **29** (1996) no. 3, 95 – 100.
110. E. M. Wise, R. F. Vines in: *Metals Handbook,* 9th ed., **vol. 2,** Metals Park, OH 1979, p. 699.
111. K. Yasuda, K. Hisatune, *Gold Bull.* **26** (1963) no. 2, 50.
112. Doduco: *Datenbuch* Pforzheim 1977 p. 50.
113. B. Kempf, S. Schmauder, *Gold Bull.* **31** (1998) no. 2, 51 – 57.
114. G. Humpston, D. M. Jacobson, *Gold Bull.* **25** (1992) no. 4, 132.
115. *Diebeners Handbuch des Goldschmieds,* vol. 1, Diebener, Stuttgart 1952.
116. E. Wagner, *Metall* **12** (1985) 628 – 630.
117. Degussa: "Sonne und Mond," *Degussa-Informationen für Juweliere, Gold- und Silberschmiede,* Frankfurt, since 1968.
118. F. Ensslin: *Edelmetall-Analyse,* Springer Verlag, Berlin 1964.
119. Chemikerausschuß der GDMB (ed.): *Analyse der Metalle,* 2nd ed., Springer Verlag, Berlin 1961.
120. H.-M. Lüschow: *Edelmetallanalytik, Eine Übersicht über bewährte klassische und moderne Methoden,* Heft 69 der Schriftenreihe der GDMB, Clausthal-Zellerfeld 1993.

121. H.-M. Lüschow: "Gold Analysis in a Precious Metal Refinery – State and Development," *Ertmetall* **50** (1997) 22 – 42.
122. P. Raw, C. W. Corti (ed.): *The Assaying and Refining of Gold,* World Gold Council, London 1997.
123. M. Brill: "Analysis of Carat Gold," in *Gold Technology, Quality Assurance,* World Gold Council, Issue No. 22, 1997, 10 – 25.
124. R. S. Young: "Analysis for Gold," *Gold Bull.* **13** (1980) 9 – 14.
125. N. Desebrock: *Gold Refiners and Bars Worldwide, An International Register,* Grendon International Research Pty Ltd, Fremantle, Australia 1991.
126. H. M. Lüschow: *Probenahme, Theorie und Praxis,* "Schriftenreihe d. GDMB" Verlag Chemie, Weinheim 1980. H.-M. Lüschow, "Zur Probenahme von Edelmetallen. Teil 1: Probenahme von metallischen Materialien", *Erzmetall* **42** (1989) 153 – 159.
127. ISO/Committee Draft No. 11546 (still under discussion)
128. P. Beckmans: "Gross Sampling of Shredded Circuit Boards," in G. Kraft (ed.): *Sampling in Non-Ferrous Metals Industry,* Trans Tech Publications, Clausthal-Zellerfeld 1993, pp. 105 – 109.
129. ISO Standard No. 11426, "Determination of Gold in Gold Jewellery Alloys – Cupellation Method (Fire Assay)".
130. M. Brill: "The Quality Control of Alloys from the Precious Metal Foundry with the Aid of the Glow Discharge Lamp Emission Spectrometer," Proc. 11th IPMI Conf., Brussels, Belgium 1989, pp. 369 – 400.
131. J. W. Stankiewicz et al.: Gold and Gold Alloy Reference Materials for XRF Analysis.
132. J. Suchomel, *Fresenius' Z. Anal. Chem.* **300** (1980) 257 – 266.
133. J. Suchomel, *Fresenius' Z. Anal. Chem.* **307** (1981) 14 – 18.
134. M. Brill, K.-H. Wiedemann: "Determination of Gold in Gold Jewellery Alloys by ICP-Spectrometry," *Gold Bull.* **25** (1992) 13 – 25.
135. "Spectro X-Test," Spectro Analytical Instruments Ltd., D-47533 Kleve.
136. A. Prince, G. V. Raynor, W. S. Evans: *Phase diagrams of Tertnary Gold Alloys,* The Institute of Materials, London 1990.
137. W. S. Rapson, T. Groenewald: *Gold Usage,* Academic Press, London 1978.
138. A. Luschin v. Ebengreuth: *Allgemeine Münzkunde und Goldgeschichte,* Oldenburg, München 1971.
139. E. Drost, J. Hausselt: "Uses of Gold in Jewelry," *Interdiscip. Sci. Rev.* **17** (1992) 271.
140. G. Humpston: "The Essential Role of Gold in the Fabrication of Microwave Electronic Systems," *Gold Bull.* **32** (1999) no. 3, 75 – 79.
141. A. Keil: *Werkstoffe für elektrische Kontakte,* Springer Verlag, Berlin 1960.
142. Y. Okinana: "Some recent Topics in Gold Plating for Electronics Applications," *Gold Bull.* **31** (1998) no. 1, 3 – 13.
143. P. Bowell: "Low Resistance Gold Contacts for Gallium Nitride," *Gold Bull.* **31** (1998) no. 4, 132 – 133.
144. H. Reichl, *DVS-Ber.* **129** (1990) 10.
145. R. L. Whetton et al., *Adv. Mater.* **8** (1996) 428.
146. W. D. Luedtke, U. Landman, *J. Phys. Chem.* **100** (1996) 13324.
147. I. Vezmar et al., *Z. Phys. D.* **40** (1997) 147 – 151.
148. C. L. Cleveland et al., *Z. Phys. D.* **40** (1997) 503 – 508.
149. M. Shirai et al., *Gold Bull.* **32** (1999) no. 3, 80 – 89.
150. R. Whyman, *Gold Bull.* **29** (1996) 11 – 15.
151. G. J. Hutchings, *Gold Bull.* **29** (1996) 123 – 130.
152. G. M. Francis, R. E. Palmer, *Gold Bull.* **29** (1996) 47 – 51.
153. C. Jones, *Financial Times* (1996) 15. Aug.
154. A. P. Alivisatos et al., *Nature* **382** (1996) 607.
155. C. A. Mirkin et al., *Nature* **382** (1996) 607.
156. Degussa: "Technik die verbindet," *Berichte aus Forschung und Praxis.*
157. Degussa: Metall Forschung und Entwicklung, Degussa, Ressort Metall, Frankfurt/Main 1991.
158. D. M. Jacobson, S. P. S. Sangha, *Gold Bull.* **29** (1996) no. 1, 3 – 9.
159. B. Kempf, J. Hausselt, *Interdiscip. Sci. Rev.* **17** (1992) 251.
160. G. Humpston, D. M. Jacobson, S. P. S. Sangha, *Gold Bull.* **26** (1993) no. 3, 90.
161. W. Funk, G. Reinacher: "Gold Alloy Spinnerets for the Production of Viscose Rayon," *Gold Bull.* **7** (1974) 2 – 9.
162. H.-D. Lachmann, H. G. Simanowski: *Gold als Oberfläche,* Leuze Verlag, Saulgau 1982.
163. F. H. Reid, W. Goldie: *Gold Plating Technology* Electrochemical Publications Ltd., Ayr 1974.
164. F. Simon: "Goldabscheidung aus stark sauren Electrolyten," *Galvanotechnik* **78** (1987) 2450.

165. H. O. Ali, I. R. A. Christie: "A Review of Electroless Gold Deposition Processes," *Gold Bull.* **17** (1984) 118.
166. E. Arnet: "Elektroformung von Hohlschmuck nach dem AUROFORM-Verfahren," *Galvanotechnik* **86** (1995) 3575.
167. F. Simon: "Recent Developments of Electroforming with Precious Metals for Jewellery," *Trans. Inst. Metal Finish.* **75** (1997) B56.
168. D. Withey: "Gold Electroforms and Heavy Electrodeposits," *Gold Bull.* **10** (1983) 70.
169. U. Kopacz, R. Riedl: "Abscheidung dekorativer, goldfarbiger Dünnschichten mit ionengestützten PVD-Verfahren," *Z. Metallkunde* **83** (1992) no. 7, 492 – 499.
170. E. D. Nicholson, *Gold Bull.* **12** (1979) no. 4, 161.
171. Degussa, US 5 808 136, 1998 (H. Mueller et al.).
172. G. C. Bond: "Gold: the Catalytic Metal for the 21st Century?" *Gold Bull.* **31** (1998) no. 4, 110.
173. D. Thompson: "New Advances in Gold Catalysis Part I," *Gold Bull.* **31** (1998) no. 4, 111 – 118; Part II, **32** (1999) no. 1, 12 – 19.
174. Laura Prati: "New Gold Catalysts for Liquid Phase Oxidation," *Gold Bull.* (1999) no, 3, 96 – 101.
175. A. Ueda, M. Haruta, *Gold Bull.* **32** (1999) no. 1, 3 – 11.
176. A. L. Prati, M. Rossi, *J. Catal.* **176** (1998) 552.
177. J. H. Teles, S. Brode, M. Chabanas, *Angew. Chem. Int. Ed.,* **37** (1998) 1415.
178. H. Renner, *Ullmann's ,* 5th ed, **A12,** p. 530.
179. Gold Fields Mineral Services Ltd. London: *Gold Survey 2000*.
180. Degussa: Edelmetallmärkte 1997.
181. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 6th ed., Georg Thieme Verlag, Stuttgart, 1986.
182. E. Merian (ed.): *Metalle in der Umwelt,* Verlag Chemie, Weinheim 1984.
183. H. G. Seiler, A. Sigel, H. Sigel (eds.): *Handbook on Metals in Clinical and Analytical Chemistry,* Marcel Dekker, New York 1994.
184. G. Perreli, G. Piolatto: "Tentative Reference Values for Gold, Silver, and Platinum: Literature Data Analysis," *Sci. Tot. Environ.* **120** (1992) 93.
185. A. Rütter: "Metalldermatosen II," *Hautarzt* **47** (1996) 400 – 409.
186. W. Wirth, Ch. Gloxhuber: *Toxikologie,* 4th ed., Georg Thieme Verlag, Stuttgart 1985.
187. J.M. Lucas, *Mineral Commodity Summaries* Bureau of Mintes, Washington, DC, 1992, p. 75

Gout Remedies

Larry J. Silverman, University of Michigan, Ann Arbor, Michigan 48109, United States

Thomas D. Palella, University of Michigan, Ann Arbor, Michigan 48109, United States

1. Introduction 1
2. Gout and Hyperuricemia 2
2.1. Pathophysiology 2
2.2. Clinical Manifestations 3
3. Treatment 4
3.1. Drugs Used to Treat Acute Gout . . 4
3.1.1. Colchicine 4
3.1.2. Nonsteroidal Anti-inflammatory Drugs 5
3.1.3. Corticosteroids 7
3.2. Antihyperuricemic Drugs 7
3.2.1. Uricosuric Drugs 7
3.2.2. Xanthine Oxidase Inhibitors 8
3.3. Dietary Management 9
4. References 9

1. Introduction

Gout is an exclusively human disease. Its cardinal manifestations are high serum urate levels (hyperuricemia) and recurrent episodes of acute arthritis. In its fully developed form, gout also results in deposition of uric acid in the kidneys (nephrolithiasis), interstitial renal disease, and deposition of monosodium urate monohydrate aggregates (tophi) in and around joints (chronic tophaceous gout).

History. The term gout was first introduced in the 13th century, although its clinical manifestations had been described by Hippocrates in the 5th century B.C. [1]. In the 17th century, Sydenham distinguished gout from other arthritidies [2]. The importance of uric acid in the pathogenesis of this disease was elucidated in the 19th century [3].

An understanding of the pathogenesis of gout has lead to the development of specific therapies. Although colchicine has been used for centuries to treat acute gouty arthritis, newer anti-inflammatory agents are now used to manage acute attacks [4]. In addition, effective antihyperuricemic therapy has been developed. These agents decrease serum urate, reduce the total body pool of uric acid, prevent urate deposition, and promote dissolution of preformed tophi and stones.

Biochemistry. Uric acid is the ultimate degradative product of purine metabolism in humans. It is derived primarily from the breakdown of nucleic acids, nucleotides, and ingested, preformed purines. The proximate steps in uric acid formation are dephosphorylation of ribonucleotides and deoxyribonucleotides to their nucleoside forms, adenosine and guanosine. Adenosine and guanosine are both converted to inosine. Purine nucleoside phosphorylase converts inosine to hypoxanthine in a freely reversible reaction. Hypoxanthine, thus generated, is a substrate for the iron-containing flavoenzyme, xanthine oxidase. Xanthine oxidase converts hypoxanthine to xanthine, and then xanthine to uric acid.

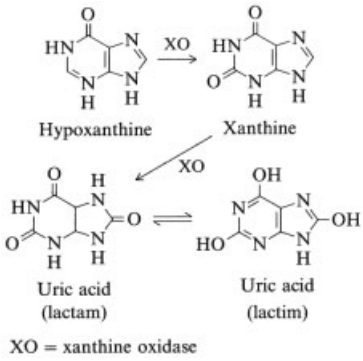

Thus, the production of uric acid is dependent upon the rate of purine nucleotide degradation and the subsequent oxidation of hypoxanthine. The serum concentration of uric acid reflects the difference between this production rate and the net excretion rate of uric acid. Under normal conditions, two-thirds of the daily production of uric

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_535

acid is excreted in the urine. The remaining fraction is excreted through the intestine where uric acid is cleaved by the action of bacterial uricases.

Epidemiology. Hyperuricemia denotes an elevation of serum urate concentration and may be defined on either a physicochemical or an epidemiologic basis. Physicochemically, serum is saturated with urate at concentrations exceeding 7 mg/dL as measured by the uricase method at 37 °C [4]. In sex- and age-matched populations, relative hyperuricemia is present when the serum urate concentration exceeds two standard deviations above the mean. In postpubertal men, this value is 7 mg/dL; in postpubertal females, 6 mg/dL [4]. Epidemiologically, an increased risk for development of gout and kidney stones exists when the serum urate concentration exceeds 7 mg/dL [4]. In the majority of subjects, however, hyperuricemia is asymptomatic. Only a small fraction of hyperuricemic individuals develop gout [5].

The serum urate concentration varies with age and sex [6–8]. Values range from 3 to 4 mg/dL in children. With the onset of puberty, serum urate concentration normally rises in males to 4 – 6 mg/dL and remains at that level throughout life. In females, the serum urate concentration typically remains unchanged until the menopause at which time it approaches values comparable to those in adult males. Factors such as blood pressure, serum creatinine, height, weight, age, and alcohol intake correlate with serum urate concentration [9–12].

Serum urate levels above 7 mg/dL are found in 6.5 – 9.3 % of adult males [4]. Among women, 6.3 % have serum urate values >6 mg/dL [4]. The development of gout depends on the degree and duration of serum urate elevation. Therefore, the prevalence of gout increases with age and with the degree of hyperuricemia [11, 12]. The annual incidence of gout is 0.1 % for urate levels less than 7 mg/dL and 4.9 % for levels greater than 9 mg/dL [13].

Symptoms and Treatment. Gout is manifested clinically by recurrent episodes of acute arthritis. Intraleukocytic crystals of monosodium urate monohydrate are found in synovial fluid from inflamed joints [14].

An acute attack of gout is generally self-limited and confined to a single joint (monoarticular). Recurrent episodes follow, which tend to become polyarticular. The frequency of attacks increases with time and leads to disabling chronic arthritis. Prompt diagnosis and implementation of therapy at the onset of an acute attack shorten the episode and relieve the disabling pain. Prophylactic therapy aimed at preventing recurrent arthritis significantly reduces the risk of chronic arthritis and joint deformity [15]. Antihyperuricemic agents lower the serum urate concentration and thus promote solubilization of tophi. Reduction of serum urate levels also decreases the risks of recurrent arthritis, uric acid nephrolithiasis, and urate nephropathy, as well as the development of chronic gouty arthritis.

2. Gout and Hyperuricemia

2.1. Pathophysiology

Hyperuricemia may result from insufficient excretion of uric acid by the kidneys, endogenous overproduction of uric acid, or a combination of both. *Primary hyperuricemia* develops independently of other diseases and is unrelated to drug use. *Secondary hyperuricemia* is either a manifestation of an underlying disease or the result of an ingested drug or toxin.

Impaired Renal Uric Acid Excretion. The most common cause of hyperuricemia is diminished uric acid excretion by the kidneys. Although this accounts for approximately 90 % of cases of primary hyperuricemia [16], the reason has not been established; impaired urate clearance has been documented in these patients [17–21].

Underlying renal disease, drugs, or toxins may also impair the ability of the kidney to excrete uric acid and result in secondary hyperuricemia. Intravascular volume contraction resulting from diuretic therapy, adrenal insufficiency, or liver disease leads to secondary hyperuricemia due to diminished glomerular filtration and increased tubular reabsorption of uric acid [22, 23]. Drugs such as ethanol, ethambutol, and nicotinic acid may diminish renal clearance of uric acid [24]. Organic acids inhibit uric acid secretion by the kidneys. Therefore, diseases resulting in metabolic acidosis may also lead to hyperuricemia [24].

Overproduction of Uric Acid. Endogenous overproduction of uric acid accounts for approximately 10 % of cases of primary hyperuricemia [16]. Overproduction is characterized by a high level of uric acid in the urine (hyperuricaciduria), which is the result of accelerated purine synthesis. The etiology of urate overproduction in primary hyperuricemia is unknown in the majority of cases, but it may be due to specific inborn errors of metabolism. Patients with 5-phosphoribosyl-1-pyrophosphate synthetase overactivity or partial hypoxanthine phosphoribosyltransferase (HPRT) deficiency demonstrate primary hyperuricemia [4].

Endogenous urate overproduction caused by the following factors may also lead to secondary hyperuricemia:

1) Inborn enzyme defects resulting in overproduction of uric acid, e.g., the Lesch–Nyhan syndrome, glucose 6-phosphatase deficiency, and the glycogen storage diseases types I, III, V, and VII [24]
2) Myeloproliferative disorders leading to accelerated nucleic acid turnover
3) Increased adenosine triphosphate (ATP) degradation caused by alcohol consumption or tissue hypoxia [24]

Urate Precipitation. The acute attack of gout is triggered by precipitation of monosodium urate monohydrate crystals in or near a joint [6,25,26]. The cause of this precipitation is unclear, and hyperuricemia is not essential. The inflammatory response to intra-articular urate crystals includes activation of the kallikrein–kinin and complement systems, formation of leukotrienes, platelet activation, prostaglandin synthesis [16], and leukocyte migration. The leukocytes phagocytose the urate crystals and release chemotactic factors. Intracellular urate crystals cause lysosomal disruption and cell injury. Release of lysosomal enzymes into the joint space leads to further inflammatory changes including increased capillary permeability and tissue damage. For a detailed description of inflammation, see → Anti-inflammatory – Anti-rheumatic Drugs, Chap. 1.

2.2. Clinical Manifestations

Articular Disease. The clinical spectrum of gout encompasses four stages:

1) asymptomatic hyperuricemia,
2) acute gouty arthritis,
3) intercritical gout, and
4) chronic tophaceous gout [16].

Asymptomatic hyperuricemia is characterized solely by elevation of the serum urate concentration. *Acute gouty arthritis* is usually the first clinical manifestation of gout and often occurs after years of asymptomatic hyperuricemia. The first attack is usually monoarticular and develops rapidly. The most commonly affected joint is the metatarsal–phalangeal joint of the big toe [27], but other joints may be affected [16]. Pain, swelling, and redness develop and intensify within a few hours. Untreated acute gout is generally self-limited; symptoms persist from several hours to weeks in severe attacks.

In *intercritical gout* the first attack of gouty arthritis is followed by a variable, asymptomatic period. The majority of patients experience the second attack within six months to two years, but the asymptomatic (intercritical) period may be ten years or longer [16]. During the phase of intercritical gout, the asymptomatic periods between acute attacks shorten. The arthritis becomes polyarticular, and if untreated, chronic persistent arthritis may develop [16,24].

Chronic tophaceous gout is characterized by chronic polyarticular arthritis and development of tophi in and around joints and in soft tissues [16]. Joint erosion and destruction occur, which may be accompanied by marked deformity and crippling of hands and feet.

Renal Disease. The clinical manifestations of hyperuricemia are not limited to articular disease; three types of renal disease are also found: (1) uric acid nephropathy, (2) urate nephropathy, and (3) nephrolithiases.

Uric acid nephropathy results from precipitation of uric acid within the collecting ducts of the kidneys [16]. It is typically associated with marked hyperuricemia and hyperuricaciduria. Uric acid nephropathy is seen most frequently in lymphoproliferative disorders following chemotherapy [28–30], when the large

nucleic acid pool liberated by tumor cell lysis leads to marked hyperuricemia. Acute renal failure may occur.

Urate nephropathy is caused by deposition of monosodium urate monohydrate in the renal interstitium [16]. Unlike uric acid nephropathy, it does not cause acute renal failure but may contribute to the chronic renal insufficiency of gouty patients.

Nephrolithiasis. Gouty patients have an increased prevalence of kidney stones [16, 31, 32]. Renal calculi occur in 10 – 25 % of patients with primary gout and in as many as 42 % of patients with secondary gout. The incidence of stone formation increases with increasing serum urate concentration and urinary uric acid excretion [31, 32].

Other Disorders. Several other disorders are associated with hyperuricemia and gout; these include hypertension, obesity, hyperlipidemia, and diabetes mellitus, as well as cerebrovascular and cardiovascular disease [16]. No direct cause and effect relationship has been established between these disorders and hyperuricemia.

3. Treatment

Medical management of gout is undertaken to:

1) promote rapid resolution of acute arthritis;
2) reduce the likelihood of recurrent gouty attacks;
3) lower the serum urate concentration, thereby inhibiting urate deposition and promoting dissolution of tophi; and
4) prevent uric acid nephrolithiasis.

3.1. Drugs Used to Treat Acute Gout

3.1.1. Colchicine

Colchicine [*64-86-8*], (*S*)-*N*-(5,6,7,9-tetrahydro-1,2-3,10-tetramethoxy-9-oxobenzo[α]heptalen-7-yl)acetamide, is an alkaloid of the plant *Colchicum autumnale*. The plant has been used in the treatment of articular disorders since the sixth century A.D. [33] and colchicine has been used to treat gout for over two centuries. This anti-inflammatory drug is uniquely effective against gouty arthritis (see also → Alkaloids; → Anti-inflammatory – Antirheumatic Drugs).

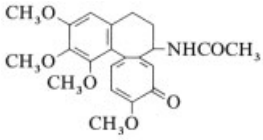

Colchicine

Colchicine impairs leukocyte migration, thereby inhibiting the inflammation responsible for acute gouty attacks [34]. Colchicine has also been shown to inhibit lysosomal degranulation, phagocytosis, and neutrophil adherence [34–39]. The drug binds to microtubular structures and arrests cell division [33].

Colchicine is administered orally or intravenously [33]. Peak plasma levels occur 0.5 to 2 h after an oral dose. The drug is eliminated principally through feces; ca. 10 – 20 % is eliminated in the urine.

Oral colchicine is used to treat acute gouty arthritis [16]. The drug is most effective when administered within the first few hours of an attack at a dose of 0.5 or 0.6 mg/h until symptoms subside or gastrointestinal side effects develop (maximum total dosage 6.0 mg). Symptoms usually improve within 12 h and often completely resolve within three days. As many as 50 – 80 % of patients develop gastrointestinal side effects before relief from arthritis occurs. Daily oral colchicine (0.5 – 1.0 mg) is useful in preventing recurrent episodes of acute gout [16].

Intravenous colchicine effectively terminates acute attacks of gout without causing gastrointestinal side effects [16]. It is administered intravenously as a single 2-mg dose in 20 mL of saline, followed by 1-mg dosages every 6 h up to a maximum of 4 – 5 mg. Complications include local extravasation with inflammation and skin necrosis, as well as bone marrow suppression.

3.1.2. Nonsteroidal Anti-inflammatory Drugs

Several nonsteroidal anti-inflammatory drugs have been employed in the treatment of acute gout. These drugs are thought to exert their anti-inflammatory effects by inhibiting prostaglandin synthesis, inhibiting leukotriene syn-

thesis, and preventing free-radical formation [40]. Prostaglandins are potent mediators of inflammation. The nonsteroidal anti-inflammatory agents reversibly inhibit cyclooxygenase, an enzyme that converts arachidonic acid to prostaglandin precursors [40]. Prostaglandin synthesis is thereby suppressed. Leukotrienes, also derived from arachidonic acid, have been implicated in leukocyte aggregation, degranulation, and free-radical formation [40]. Some nonsteroidal anti-inflammatory agents inhibit leukotriene synthesis and presumably modify the inflammatory response by interfering with these mechanisms.

Oxygen free radicals are formed during the synthesis of prostaglandins and can cause tissue injury. Inhibition of prostaglandin synthesis by nonsteroidal agents suppresses free-radical formation [40].

(For physical properties, syntheses, and trade names of the drugs treated in this section see → Anti-inflammatory – Antirheumatic Drugs).

Indomethacin [*53-86-1*], 1-(4-chlorobenzoyl)-5-methoxy-2-methyl-1*H*-indole-3-acetic acid (Indocin) is rapidly absorbed from the gastrointestinal tract. Peak plasma levels are obtained within 3 h of ingestion [33], and the drug is highly bound to plasma proteins. It is eliminated in both urine and feces, 10 – 20 % being excreted unchanged in the urine. Approximately 50 % of a given dose undergoes O-demethylation, and 10 % is conjugated with glucuronic acid.

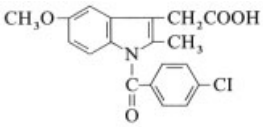

Indomethacin

For treatment of acute gout, indomethacin is administered at a 50 – 75-mg dosage, followed by 50 mg every 6 h for a maximum dose of 200 mg in the first 24 h [16]. The dosage is subsequently reduced [16].

Indomethacin may also be used for gout prophylaxis [16]. A daily dose of 25 – 50 mg is usually sufficient to prevent recurrence of acute attacks.

Phenylbutazone [*50-33-9*], 4-butyl-1,2-diphenyl-3,5-pyrazolidinedione (Butazolidin) has

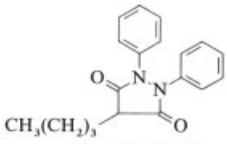

Phenylbutazone

been used for over 30 years to treat acute gout [41, 42]. After oral administration, peak plasma concentration is reached within 2 h [33]. The drug is completely absorbed from the gastrointestinal tract and 98 % is bound to plasma proteins. Plasma half-life ranges from 50 to 100 h. The drug undergoes glucuronidation and hydroxylation, and is excreted primarily in the urine. Oxyphenylbutazone, an active metabolite of phenylbutazone, has a long plasma half-life and contributes to the side effects of phenylbutazone. Oxyphenylbutazone is also excreted in the urine after glucuronidation [33].

For treatment of acute gout, phenylbutazone is administered at an initial dosage of 200 mg – 600 mg [16], followed by 100 – 200 mg every 6 – 8 h for 24 h (maximum dose = 600 mg/d). The dosage is then reduced and discontinued after two days.

Side effects of phenylbutazone include hypersensitivity reactions, gastric discomfort, peptic ulcer disease, and fluid retention [33]. Deaths from aplastic anemia have been reported [43].

Naproxen [*22204-53-1*], (*S*)-6-methoxy-α-methyl-2-naphthaleneacetic acid (Naprosyn) was effective in acute gout management in several clinical trials [44]. It is fully absorbed after oral administration [33]. Plasma half-life is 12 – 15 h. The drug is extensively bound to plasma protein and eliminated in the urine. Approximately 30 % undergoes 6-demethylation prior to elimination. To treat acute gout, naproxen is administered at an initial dose of 500 – 750 mg, followed by 250 mg every 8 h until the acute attack subsides [16].

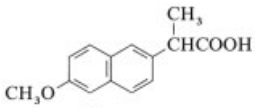

Naproxen

Ibuprofen [*15687-27-1*], α-methyl-4-(2-methylpropyl)benzeneacetic acid (Motrin) is absorbed efficiently from the gastrointestinal tract [33]. Maximum plasma levels are achieved within 1 – 2 h, and its half-life is approximately 2 h. Like naproxen, ibuprofen is almost completely bound to plasma proteins and is eliminated principally via the urine. It is metabolized prior to elimination in the urine. Ibuprofen has demonstrated efficacy in the treatment of acute gout [45]. A dosage of 800 mg every 8 h is given until symptoms are resolved [16]. The dose may then be reduced.

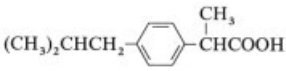

Ibuprofen

Fenoprofen [*31879-05-7*], (±)-α-methyl-3-phenoxybenzeneacetic acid (Nalfon) is rapidly absorbed, and peak plasma concentrations are achieved within 2 h of oral administration [33], [46, pp. 890 – 891]. The plasma half-life is approximately 3 h. Fenoprofen is highly bound to plasma protein, and its major urinary metabolites are fenoprofen glucuronide and 4′-hydroxyfenoprofen glucuronide. Over 90 % of the drug is metabolized prior to elimination in the urine. Successful termination of an acute gout attack may be achieved with a dosage of 800 mg every 6 h [16]. The dose is then tapered when symptomatic improvement occurs.

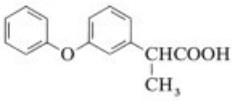

Fenoprofen

Sulindac [*38194-50-2*], (*Z*)-5-fluoro-2-methyl-1-{[4-(methylsulfinyl)phenyl]methylene}-1*H*-indene-3-acetic acid, (Clinoril) is approximately 90 % absorbed after oral administration [33]. The active sulfide metabolite is formed by oxidation to the sulfone and subsequent reduction. Sulindac has a plasma half-life of 7 h; the sulfide metabolite has a half-life of 18 h. Sulindac and its metabolites are extensively protein bound. Excretion of the drug occurs in the urine and feces. Approximately 50 % of the administered dose is eliminated in the urine, the major product being the conjugated sulfone. Sulindac administered at a dosage of 200 mg twice a day is effective in terminating an acute attack of gout [47].

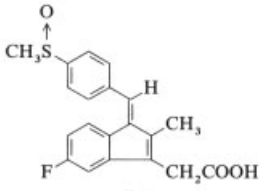

Sulindac

Piroxicam [*36322-90-4*], 4-hydroxy-2-methyl-*N*-2-pyridyl-2*H*-1,2-benzothiazine-3-carboxamide-1,1-dioxide (Feldene) is well absorbed after oral administration, and peak plasma concentrations are achieved within 3 – 5 h [46, pp. 1558 – 1559]. Mean plasma half-life is 50 h. The drug is excreted in both urine and feces, the majority being excreted as metabolites formed by hydroxylation and conjugation, cyclodehydration, and hydrolysis of the amide linkage, decarboxylation, ring contraction, or N-demethylation. Piroxicam administered at a daily dose of 20 mg, effectively suppresses or completely relieves symptoms of acute gouty arthritis within five days [48–50].

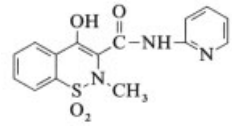

Piroxicam

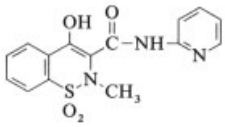

3.1.3. Corticosteroids

Corticosteroids have both anti-inflammatory and immunomodulatory properties [16]. They inhibit arachidonic acid release from phospholipids, thereby inhibiting prostaglandin synthesis. They are also purported to stabilize lysosomal membranes, inhibit cellular metabolism,

decrease serum immunoglobulins, and diminish synovial permeability. Intra-articular corticosteroids are particularly effective in suppressing acute gouty arthritis [51]. Commonly used intra-articular steroids include methylprednosolone acetate (Depo-medrol), triamcinolone acetonide (Kenalog), and triamcinolone hexacetonide (Aristopan). Relief of symptoms usually occurs within 12 h. This form of treatment may be employed when one or two joints are involved and the patient is unable to tolerate other forms of therapy. Many different corticosteroid preparations are available. For structure, properties, and preparation, see → Hormones.

3.2. Antihyperuricemic Drugs

The antihyperuricemic drugs consist of (1) uriscosuric drugs which increase urinary uric acid excretion and (2) xanthine oxidase inhibitors which inhibit uric acid synthesis. The aim of antihyperuricemic therapy is to lower the concentration of serum urate below 6.4 mg/dL [16].

The selection of an antihyperuricemic agent should be based on the cause of the hyperuricemia and the condition of the patient. Patients without kidney problems who excrete less than 700 mg of uric acid per day may be treated with either uricosuric agents or xanthine oxidase inhibitors [24]. Patients who excrete more than 700 mg of uric acid a day should be treated with xanthine oxidase inhibitors. Antihyperuricemic agents should not be given until an acute gouty attack has subsided because they may exacerbate such attacks [16]. In addition, they may also precipitate the onset of acute gouty arthritis. Patients should, therefore, be treated prophylactically with colchicine or indomethacin prior to the introduction of antihyperuricemic medication.

3.2.1. Uricosuric Drugs

Uricosuric drugs increase the rate of excretion of uric acid by interfering with its reabsorption by the renal tubules [52]. They reduce serum urate levels in the majority of patients but are generally ineffective in patients with marked renal insufficiency [16]. Concomitant administration of salicylate blocks their uricosuric effect [52].

The use of uricosuric drugs should be accompanied by liberal intake of fluids to prevent crystallization of uric acid within the urinary tract. Alkalinization of urine by the administration of sodium hydrogen carbonate or potassium citrate may prevent formation of kidney stones [24].

Probenecid [*57-66-9*], 4-[(dipropylamino)-sulfonyl]benzoic acid (Benemid) inhibits epithelial transport of organic acids [52]. It is completely absorbed after oral administration, with peak plasma levels achieved in 2–4 h. It is extensively bound by plasma proteins and has a plasma half-life of approximately 9 h. Probenecid is eliminated in the urine. Metabolism may occur by hydroxylation or conjugation to glucuronide [52].

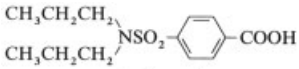

Probenecid

Probenecid is administered at a dose of 250 mg twice a day for one week; the dosage is then increased up to a maximum of 2 g per day [46, pp. 1246–1247].

Adverse side effects to probenecid include gastrointestinal irritation, hypersensitivity reactions, and central nervous system abnormalities [52].

Sulfinpyrazone [*57-96-5*], 1,2-diphenyl-4-[2-(phenylsulfinyl)ethyl]-3,5-pyrazolidinedione (Anturane) is a strong organic acid [52]. Although sulfinpyrazone inhibits reabsorption of uric acid, like probenecid it interferes with tubular secretion of other organic acids. It is absorbed efficiently from the gastrointestinal tract and is extensively bound to plasma proteins. This drug is excreted in the urine, with 10 % being the N1-*p*-hydroxyphenyl metabolite.

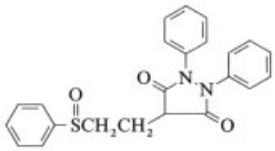

Sulfinpyrazone

Sulfinpyrazone is administered initially at a dose of 50 – 100 mg twice a day, which can be increased gradually to a daily dosage of up to 800 mg [46, pp. 833 – 834].

Side effects include gastrointestinal irritation, hypersensitivity reactions, and blood dyscrasias [52]. In addition, this drug may potentiate the action of certain sulfonamides such as sulfadiazine and sulfisoxazole [46, pp. 833 – 834].

Benzbromarone [*3562-84-3*], (3,5-dibromo-4-hydroxyphenyl)(2-ethyl-3-benzofuranyl)methanone, is a potent uricosuric agent with demonstrated efficacy in reducing serum urate levels [53–55]. Onset of uricosuric activity is rapid, with peak uric acid excretion occurring at approximately 5 h Benzbromarone does not affect tubular secretion of other organic acids. The usual maintenance dose is between 50 and 100 mg a day [53–56]. Side effects include diarrhea, allergy, and urolithiasis [53, 54, 56].

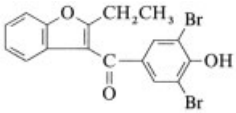

Benzbromarone

Benziodarone [*68-90-6*] (2-ethyl-3-benzofuranyl)(4-hydroxy-3,5-diiodophenyl)methanone, and **zoxazolamine** [*61-80-3*], 5-chlorobenzoxazolamine, have uricosuric activity [57, 58], but neither agent is used clinically in the United States to manage hyperuricemia.

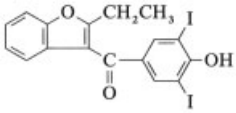

Benziodarone

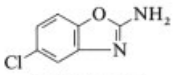

Zoxazolamine

3.2.2. Xanthine Oxidase Inhibitors

Xanthine oxidase catalyzes the oxidation of hypoxanthine to xanthine and of xanthine to uric acid [24]. Allopurinol and its oxidation product oxipurinol are structural analogues of hypoxanthine and xanthine, respectively. Both inhibit xanthine oxidase, thereby diminishing uric acid synthesis and lowering serum urate concentration.

Allopurinol [*315-30-0*], 1,5-dihydro-*H*-pyrazolo[3,4-*d*]pyrimidin-4-one (Zyloprim) is absorbed rapidly from the gastrointestinal tract [33]. Peak plasma levels occur within 30 – 60 min of an oral dose, and plasma half-life is 2 –3 h. Allopurinol is metabolized to oxipurinol [*2465-59-0*]. Oxipurinol, 1*H*-pyrazolo[3,4-*d*]-pyrimidine-4,6(5*H*,7*H*)-dione is cleared slowly from plasma and has a half-life of 15 – 20 h [46, pp. 818 – 820]. Both compounds are biologically active and neither is protein-bound [33]. Only 10 % of a single dose of allopurinol is excreted unchanged in the urine.

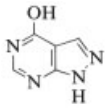

Allopurinol

The usual dosage of allopurinol is 300 mg administered once a day. The serum urate concentration falls within the first two days of therapy, and the minimum serum urate construction is achieved within two weeks [24]. The dose of allopurinol may gradually be increased to a maximum of 800 mg daily [46, pp. 818 – 820]. Dosages above 300 mg should be given twice daily. In patients with renal insufficiency, the half-life of oxipurinol is increased and the dose of allopurinol should be reduced.

Side effects of allopurinol therapy include acute gouty arthritis, hematologic abnormalities, gastrointestinal intolerance [33], and life-threatening hypersensitivity reactions.

Allopurinol may interfere with the hepatic metabolism of oral anticoagulant drugs and must be used with caution in this context [33]. Azathioprine and 6-mercaptopurine are potentiated when administered concomitantly with allopurinol. The combined administration of allopurinol and uricosuric drugs may improve serum uric acid control [33].

3.3. Dietary Management

The aims of dietary management in gout are to reduce body weight, alcohol intake, and ingestion of foods known to trigger acute attacks [16]. Patients with gout are often overweight [16]; weight reduction may lower the serum urate concentration [59].

Alcohol intake may lead to hyperuricemia and gout [60], therefore patients with gout should reduce or discontinue alcohol consumption.

Ingestion of certain, often purine-rich, foods may be associated with the development of acute gouty arthritis in some patients [16]. Specific foods known to precipitate acute attacks of gout should be avoided. Examples of purine-rich foods are herrings, sardines, liver, sweetbreads, kidneys, and brewer's yeast. Severe dietary restriction of purine is usually unnecessary, but patients should avoid excessive purine intake. High-protein diets should also be avoided because endogenous purine production is increased by excessive protein consumption [16].

Obesity, hyperlipidemia, and hypertension are associated with gout and hyperuricemia [16]. Management of these conditions with diet, exercise and appropriate medications is important in prolonging life and preventing the complications of these associated conditions.

This work is supported by National Institutes of Health Grant Number DK-19045. Dr. Palella is the recipient of an Arthritis Investigator Award from the Arthritis Foundation.

4. References

1. Hippocrates: *The Genuine Works of Hippocrates,* **vols. 1 and 2,** translated from the Greek with a preliminary discourse and annotations by Francis Adams, W. Wood & Co, New York 1886.
2. T. Sydenham: *Tractatus de Podagra et Hydrope,* G. Kettilby, London1683.
3. A. B. Garrod, *Trans. M-Chir. Soc. Edinburgh* **31** (1848) 83 – 97.
4. J. B. Wyngaarden, W. N. Kelley: *Gout and Hyperuricemia,* Grune & Stratton, New York 1976.
5. H. E. Paulus, A. Coutts, J. J. Calabro, J. R. Klinenberg, *J. Am. Med. Assoc.* **211** (1970) 277 – 281.
6. W. M. Mikkelsen, H. J. Dodge, H. Valkenburg, *Am. J. Med.* **39** (1965) 242 – 251.
7. S. Akizuki, *Ann. Rheum. Dis.* **41** (1982) 272 – 274.
8. W. R. Harlan, J. Cornoni-Huntley, P. E. Leaverton, *Pediatrics* **63** (1979) 569 – 575.
9. W. R. Harlan, A. L. Hull, R. P. Schmouder, F. E. Thompson et al.: "Dietary intake and cardiovascular Risk Factors," part 2, *Serum Urate, Serum Cholesterol and Correlates, in Vital and Health Statistics, United States, 1971 – 1975*, Series 11, no. 227, Dept. of Health & Human Services (DHHS) Pub. no. (PHS) 83 – 1677, Public Health Service, Washington, D.C., Mar. 1983.
10. D. Kuntz, J. M. Chretien, A. Ryckewart, *Sem. Hop. Paris* **55** (1979) 241 – 248.
11. J. Zalokar, J. Lellouch, J. R. Claude, *Sem. Hop. (Paris)* **57** (1981) 664 – 670.
12. K. Nishioka, K. Mikanagi, *Adv. Exp. Med. Biol.* **122 A** (1980) 155.
13. E. W. Campion, R. J. Glynn, L. O. DeLabry, *Am. J. Med.* **82** (1987) 421 – 426.
14. D. J. McCarty, J. L. Hollander, *Ann. Intern. Med.* **54** (1961) 452 – 460.
15. T-F. Yu, A. B. Gutman, *Ann. Intern. Med.* **55** (1961) 179 – 192.
16. W. N. Kelly, I. H. Fox in W. N. Kelley, E. D. Harris, S. Ruddy, C. B. Sledge (eds.): *Textbook of Rheumatology,* 2nd ed., W. B. Saunders Co., Philadelphia 1985, pp. 1359 – 1398.
17. C. A. Nugent, W. D. MacDiarmid, F. H. Tyler, *Arch. Intern. Med.* **113** (1964) 115 – 121.
18. J. T. Scott, M. L. Snaith, *Ann. Rheum. Dis.* **30** (1971) 285 – 289.
19. J. B. Haupt, M. A. Ogryzla, *Arthritis Rheum.* **7** (1964) 316.
20. J. E. Seegmiller, A. I. Grayzel, R. R. Howell, C. Plato, *J. Clin. Invest.* **41** (1962) 1094 – 1098.
21. W. Latham, G. P. Rodman, *J. Clin. Invest.* **41** (1962) 1955 – 1963.
22. I. H. Fox in J. A. Spittell (ed.): *Clinical Medicine,* Harper & Row, Philadelphia 1982.
23. E. I. Feinstein, V. H. Quion, E. M. Kaptein, S. G. Massry, *Am. J. Nephrol.* **4** (1984) 77.
24. T. D. Palella, I. H. Fox in C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle (eds.): *Methabolic Basis of Inherited Disease,* 6th ed., McGraw-Hill, New York 1988, in press.
25. N. J. Zvaifler, T. J. Pekin, *Arch. Intern. Med.* **111** (1963) 99 – 102.
26. J. E. Seegmiller, L. Laster, R. R. Howell, *N. Engl. J. Med.* **268** (1963) 712 – 716.

27. C. Scudamore: *A Treatise on the Nature and Cure of Gout and Rheumatism, including General Consideration on Morbid States of the Digestive Organs: Some Remarks on Regimen and Practical Observations on Gravel,* 2nd ed., Longman, London 1817.
28. E. Frei, C. J. Bentzel, R. R. Rieselbach, J. B. Block, *J. Chronic Dis.* **16** (1963) 757 – 776.
29. R. E. Rieselbach et al., *Am. J. Med.* **37** (1964) 872 – 884.
30. C. M. Kjellstrand, D. C. Campbell, B. von Hartitzsch, T. J. Buselmeier, *Arch. Intern. Med.* **133** (1974) 349 – 359.
31. A. B. Gutman, T.-F. Yu, *Am. J. Med.* **45** (1968) 757 – 779.
32. T-F. Yu, A. B. Gutman, *Ann. Intern. Med.* **67** (1967) 1133 – 1148.
33. R. J. Flower, S. Moncada, J. R. Vane in A. G. Gilman L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th ed., Macmillan Publ. Co., New York 1980, pp. 682 – 728.
34. S. E. Malawista, *Arthritis Rheum.* **18** (1975) 835 – 846.
35. R. B. Zurier, S. Holfstein, G. Heissman, *J. Cell Biol.* **58** (1973) 27 – 41.
36. Y. H. Cheng, *Arthritis Rheum.* **11** (1968) 473.
37. R. I. Lehrer, *J. Infect. Dis.* **127** (1973) 40 – 48.
38. R. Penney, D. A. G. Galton, J. T. Scott, V. Eifen, *Br. J. Haematol.* **12** (1966) 623 – 632.
39. E. Dellaverde. P. T. Fan, Y. H. Chang, *J. Pharmacol. Exp. Ther.* **223** (1982) 197 – 202.
40. H. E. Paulus, D. E. Furst in D. J. McCarty (ed.): *Arthritis and Allied Conditions,* 10th ed., Lea & Febiger, Philadelphia 1985, pp. 453 – 486.
41. R. H. Freyberg, *Arthritis Rheum.* **5** (1962) 624 – 632.
42. A. B. Gutman, *Arthritis Rheum.* **8** (1965) 911 – 920.
43. M. F. Cuthbert, *Curr. Med. Res. Opin.* **2** (1974) 600 – 610.
44. R. F. Willkins, J. B. Case, F. J. Huix, *J. Clin. Pharmacol.* **15** (1975) 363 – 366. R. A. Sturge et al., *Adv. Exp. Med. Biol.* **76 B** (1977) 290 – 296.
45. M. C. Schweitz, D. J. Nashel, F. P. Alepa, *J. Am. Med. Assoc.* **239** (1978) 34 – 35.
46. E. R. Barnhart (ed.): *Physicians Desk Reference,* 41st ed., Medical Economics Co., Oradell, N.J., 1987, pp. 890 – 891.
47. G. N. Karachalios, G. Donas, *Int. J. Tissue React.* **4** (1982) 297 – 299.
48. E. Murpy, *J. Int. Med. Res.* **7** (1979) 507 – 510.
49. G. N. Karachalios, *Clin. Exp. Rheum.* **3** (1985) 63 – 65.
50. R. H. Bluestone, *Am. J. Med.* **72** (1982) 66 – 70.
51. J. L. Hollander, J. Bone, *Joint Surg.* **35 A** (1953) 983 – 990.
52. G. H. Mudge in A. G. Gilman, L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th ed., Macmillan Publ. Co., New York 1980, pp. 929 – 934.
53. A. Masbernard, C. P. Giudicelli, *S. Afr. Med. J.* **59** (1981) 701 – 706.
54. W. A. M. Scott, *J. Int. Med. Res.* **9** (1981) 82 – 87.
55. G. W. Schepers, *J. Int. Med. Res.* **9** (1981) 511 – 515.
56. R. Bluestone, J. Klinenberg, I. K. Lee, *Adv. Exp. Med. Biol.* **122 A** (1980) 283 – 286.
57. G. Lemieux, P. Vinay, A. Gougoux, G. Michaud, *Am. J. Phys.* **224** (1973) 1440 – 1449.
58. J. J. Burns, T.-F. Yu, L. Berger, A. B. Gutman, *Am. J. Med.* **25** (1958) 401 – 408.
59. B. T. Emmerson, *Aust. N. Z. J. Med.* **3** (1973) 410.
60. J. Faller, I. H. Fox, *N. Engl. J. Med.* **307** (1982) 1598 – 602.

Graphite → Carbon

Guanidine and Derivatives

Thomas Güthner, Degussa AG, Trostberg, Germany
Bernd Mertschenk, Degussa AG, Trostberg, Germany
Bernd Schulz, Nigu Chemie GmbH, Waldkraiburg, Germany

1. Guanidine and Guanidine Salts . . . 1
1.1. Properties . . . 1
1.2. Production . . . 2
1.2.1. Guanidine Salts from Dicyandiamide 2
1.2.2. Guanidine Salts from Cyanamide . . . 4
1.2.3. Guanidine Salts from Urea . . . 4
1.2.4. Other Processes . . . 5
1.3. Environmental Protection . . . 5
1.4. Quality Specifications . . . 5
1.5. Analysis . . . 6
1.6. Storage and Transportation . . . 6
1.7. Legal Aspects . . . 6
1.8. Uses . . . 6
1.8.1. Pharmaceuticals . . . 6
1.8.2. Plant Protection . . . 7
1.8.3. Cosmetics . . . 7
1.8.4. Textile Impregnation, Flame Retardants, Paper and Resin Manufacture . 7
1.8.5. Propellants and Explosives . . . 7
1.8.6. Biotechnological Applications of Guanidine Salts . . . 7
1.9. Economic Aspects . . . 8
2. Derivatives . . . 8
2.1. Nitroguanidine . . . 8
2.2. Aminoguanidine . . . 8
2.3. Diamino- and Triaminoguanidine . 9
2.4. Organoguanidines . . . 10
2.4.1. Alkylguanidines . . . 10
2.4.2. Arylguanidines . . . 11
2.5. Biguanide and Derivatives . . . 13
3. Toxicology and Occupational Health 14
4. References . . . 14

Guanidine salts are made from cyanamide or dicyandiamide by addition of ammonium salts. The highly soluble guanidine hydrochloride is mainly used in pharmaceutical and biotech applications, while guanidine nitrate is used in propellants and explosives.Nitroguanidine, prepared by nitration of guanidine nitrate, is used for explosives and as a building block for neonicotinoid insecticides. Aminoguanidines, prepared from hydrazines and cyanamide or guanidines, are mainly used for agrochemicals and — due to their high nitrogen content — as a component for propellants and gas generators.Organoguanidines can be prepared from cyanamides or cyanogen chloride. They are used as building blocks for agrochemicals and pharmaceuticals, and as curing agents for plastics and rubbers. Biguanides can be obtained from dicyanamide or cyanoguanidines, they are mainly used as pharmaceuticals and biocides.

1. Guanidine and Guanidine Salts

Guanidine [*113-00-8*], CH_5N_3, M_r 59.08, was first isolated in 1861 by Strecker as a degradation product of guanine [*73-40-5*].

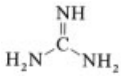

Although free guanidine or guanidine salts naturally occur only in trace amounts [1], many guanidine derivatives are important components of living organisms [2 – 4]. Guanidine and its derivatives have a wide field of application, e.g., as building blocks in the synthesis of pharmaceutical and agricultural chemicals, in the manufacture of textiles and plastics, and also in biochemistry [5 – 8].

1.1. Properties

Free guanidine can be isolated from guanidine salts by reaction with a strong base such as an

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_545.pub2

alkali metal hydroxide or methoxide. After removal of the precipitated salt the free guanidine base is obtained as colorless, waxy, very hygroscopic crystals (*mp* <50 °C).

Guanidine is a strong monobasic Brønsted base, which absorbs water and carbon dioxide from the air and forms strongly alkaline solutions with water and alcohols; the pH of a 20 wt % aqueous solution is ca. 13.5 (25 °C). In aqueous solution at elevated temperature guanidine is hydrolyzed to urea; airtight alcohol solutions are stable. Guanidine is infinitely soluble in water, ethanol, methanol, and dimethylformamide.

Protonation of guanidine gives the mesomeric, highly stabilized guanidinium cation:

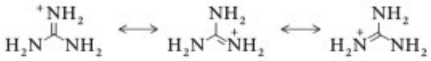

Guanidine forms stable, easily crystallizable salts with acids, even with a weak acid such as stearic, carbonic, or silicic acid. Some physical properties of various guanidine salts are listed in Table 1.

Salts of guanidine should, according to IUPAC nomenclature, be named as guanidinium compounds. In this article, however, common (although technically incorrect) names will be used, e.g., guanidine nitrate rather than guanidinium nitrate.

The very stable guanidinium cation is unreactive, so reactions of guanidine with organic compounds are usually carried out under strongly alkaline conditions. Like other amines, guanidine undergoes alkylation, acylation, methylolation, and condensation. Many guanidine derivatives are synthesized from guanylating agents such as cyanamide [*420-04-2*] (→ Cyanamides) or *O*-methylisourea [*2440-60-0*] (see Sect. 2.4) because of the relatively low reactivity of guanidine itself.

The most important reaction of guanidine is the formation of heterocyclic compounds such as pyrimidines, imidazoles and oxazoles. Guanidine condenses with bifunctional compounds such as diketones, acetoacetic acid esters, and malononitrile [*109-77-3*], eliminating ammonia or water to form 2-aminoheterocycles, for example:

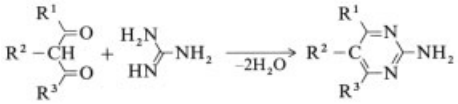

Nitration of guanidine leads to nitroguanidine (see Chap. 2.1), while chlorination with elemental chloride or sodium hypochlorite leads to *N,N″*-dichloro-diazenedicarboxamidine (azochloramide, [*502-98-7*]).

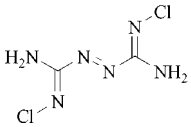

1.2. Production

No practical method exists for the direct synthesis of free guanidine. All major manufacturing processes therefore involve the production of guanidine salts by reaction of the appropriate ammonium salt with compounds containing the NCN group. At present, only *dicyandiamide* [*461-58-5*] (→ Cyanamides, Chap. 3), *cyanamide* (→ Cyanamides, Chap. 2), and *urea* [*57-13-6*] (→ Urea) are used for large-scale production of guanidine salts.

1.2.1. Guanidine Salts from Dicyandiamide

Dicyandiamide is by far the most important raw material for the production of guanidine salts. Earlier processes conducted in acidic aqueous solution with guanylurea as an intermediate [5] have been abandoned. At present, guanidine salts are manufactured by fusing dicyandiamide with ammonium salts, with biguanide [*56-03-1*] as an intermediate:

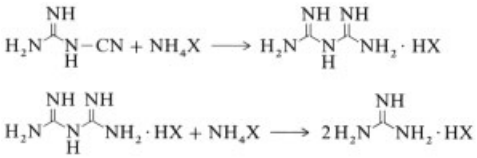

When carried out in batches in open stainless steel vessels at atmospheric pressure, reaction begins above 130 – 140 °C and becomes exothermic at 170 °C. The temperature is allowed to rise to 210 – 230 °C and is maintained

Table 1. Physical Properties of Guanidine Salts

	Chloride	Nitrate	Carbonate	Dihydrogen phosphate	Hydrogen phosphate	Sulfate	Stearate	Acetate	Thiocyanate	Aminosulfonate
CAS registry number	[*50-01-1*]	[*506-93-4*]	[*593-85-1*]	[*5423-22-3*]	[*5423-23-4*]	[*594-14-9*]	[*26739-53-7*]	[*593-87-3*]	[*593-84-0*]	[*51528-20-2*]
M_r	95.54	122.09	180.17	157.07	216.22	216.1	343.36	119.13	118.16	156.2
mp, °C	185	215	>190 [a]	130	246	>280	78	>227	117	127
d (25 °C)	1.344	1.436	1.251	1.684	1.481	1.29	1.026		1.45	1.68
pH of aqueous solution (20 wt %, 25 °C)	4.8	4.9	11.2	3.9	8.1	9.6		10	5	7
Solubility[b], g per 100 g of										
Water	228	20	50	230	23	ca. 270	[c]	115	160	120
90 °C	450	148	80	410	160	ca. 400		160	∞	
Ethanol	30	1.4	0.03	<0.1	<0.1	0.02	135	8	65	
70 °C	57	8.6	0.1	<0.1	<0.1	0.05		ca. 25	170	
Methanol 20 °C	76	5.5	0.55		0.1				90	
Acetone	0.05	0.1	0.02	<0.05	<0.05	<0.03	0.24	[c]	105	
Benzene	<0.05	0.05	0.01	<0.05	<0.05	<0.05	0.38	[c]	0.4	
DMF[d]	19.5	6.1	0.02			0.05	0.21	15		

[a] Decomposes.
[b] Unless otherwise stated, $t = 30$ °C.
[c] Trace.
[d] Dimethylformamide.

within this range until the reaction has reached completion. The molten mass is then dissolved in water or mother liquor; insoluble triazine byproducts are filtered off. The guanidine salt is crystallized by cooling or concentration of the filtrate, centrifuged, washed, and dried.

Large-scale production of guanidine salts such as guanidine nitrate is performed continuously [6] (see Fig. 1).

Dicyandiamide (85 parts) and ammonium nitrate [*6484-52-2*] (180 parts) are mixed and dosed by a screw into a heated stirring vessel where the mixture is melted down. The melt is then run into a cascade of two further vessels where the main reaction occurs. The hot molten mass is subsequently dissolved in mother liquor to give a warm salt solution, which is pumped through a filter (to remove solid triazines) and cooled in a continuously operating crystallizer. The crystallized guanidine nitrate is centrifuged, dried, and stored. This process is conducted on a technical scale for guanidine nitrate, hydrochloride, and sulfamate. The yield of pure salt is normally ca. 90 – 95 %. By solidifying the molten mass on a cooling conveyor belt immediately after it leaves the reaction cascade, a crude product can be obtained containing ca. 91 – 92 % guanidine salt and 3 – 5 % ammonium salt, together with small amounts of dicyandiamide and triazines.

Other processes involve the reaction of dicyandiamide, ammonium salts, and ammonia in aqueous solution or in liquid ammonia under pressure [6]. These processes are not industrially important.

1.2.2. Guanidine Salts from Cyanamide

Cyanamide is used mainly for the production of substituted guanidines (see Chap. 2). For simple guanidine salts, calcium cyanamide [*156-62-7*] is industrially important only in the production of guanidine nitrate [9].

Calcium cyanamide is mixed with mother liquor (from guanidine nitrate crystallization) and concentrated ammonium nitrate solution, and reacted at 120 °C. An aqueous solution of ammonium carbonate [*506-87-6*] is subsequently added to precipitate calcium carbonate. The solids are separated by concentration and filtration. The guanidine nitrate solution is neutralized with nitric acid, freed of precipitated sulfur by filtration, and of noxious gases such as hydrogen sulfide by stripping, and worked up as described above. Although this process gives good yields, it is technically very complicated and thus susceptible to breakdowns.

In modern cyanamide processes an aqueous solution of an ammonium salt and some free ammonia is heated in a stirred autoclave to 120 – 140 °C [10]. Aqueous cyanamide solution is then added slowly, and the reaction is carried out at 0.5 – 1.5 MPa. After being cooled to <100 °C, the solution can be worked up as usual. This method is used particularly to make guanidine salts that cannot be produced efficiently in a fusion process, e.g. guanidine carbonate [95, 96].

1.2.3. Guanidine Salts from Urea

Especially for guanidine nitrate, many attempts have been made to produce guanidines from urea feedstocks because of the cost advantage compared to calcium carbide/calcium cyanamide feedstocks. Melting urea with ammonium nitrate in presence of a silica catalyst at 120-200°C gives guanidine nitrate in high yield according to the following equation [11 – 14]

$$2H_2N{-}CO{-}NH_2 + NH_4NO_3 \rightarrow$$
$$H_2N{-}C(NH){-}NH_2 \cdot HNO_3 + CO_2 + 2NH_3$$

Reaction safety hazard problems, difficulties recycling off-gases and the ammonium nitrate excess, and problems with the re-usage of the spent catalyst have prevented to date from large scale realization of this process.

Guanidine carbonate is formed from urea as a byproduct of some melamine [*108-78-1*] production processes. It can be isolated from the melamine mother liquor by heating (to remove ammonia and carbon dioxide, and precipitate the remaining melamine) and filtering. After further concentration, the liquor is cooled and the crystallized guanidine carbonate is separated [15]. Although the removal of byproducts is rather expensive, significant amounts of guanidine carbonate are isolated in this way by certain melamine producers, mainly to minimize the waste from melamine production which would otherwise have to be burned or buried.

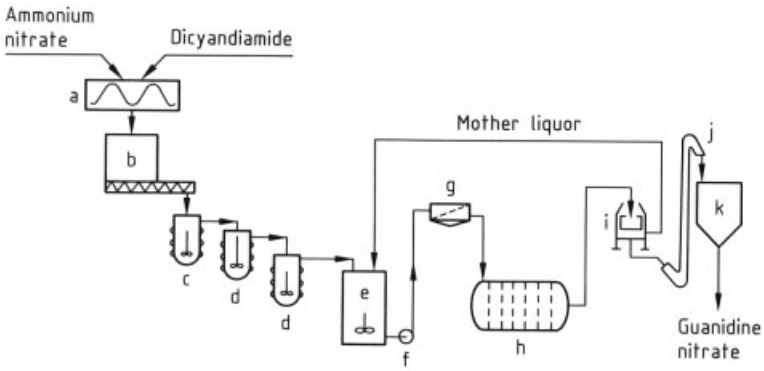

Figure 1. Production of guanidine nitrate
a) Mixer; b) Metering bunker with screw; c) Melting vessel; d) Reaction vessels; e) Dissolving tank; f) Pump; g) Filter press; h) Crystallizer; i) Centrifuge; j) Dryer; k) Storage bunker

1.2.4. Other Processes

Many processes for the production of guanidine salts are mentioned in the literature: e.g., from urea and carbon disulfide; from urea, sulfur dioxide, and ammonia; from carbon disulfide and ammonia; from carbon dioxide, ammonia, and sulfur dioxide; from ammonium thiocyanate; or from thiourea. None of these methods is used on an industrial scale. For details, see [1, 5 – 8].

Guanidine hydrochloride can be made from cyanogen chloride upon reaction with NH_3 in absence of a solvent at 140-220°C [16]. Guanidine salts of weak acids or anions that do not form stable ammonium salts can be manufactured via the free guanidine base. The latter is prepared by reaction of a suitable guanidine salt with a strong base, whose cation can be precipitated with the anion of the guanidine salt. Such reactions are usually carried out in alcohol solution [17]. To obtain alkali-free guanidine solution, guanidine carbonate or phosphate is reacted in water with calcium hydroxide. After removal of the precipitated salt, the free guanidine base can be isolated by careful vacuum distillation of the solvent. A further possibility is addition of acid or acid anhydride to give the desired guanidine salt. As an example, guanidine silicate is produced in this manner from guanidine carbonate, calcium hydroxide, and amorphous silica [18].

A good way to prepare various guanidine salts of strong or moderate acids, such as stearic acid, in smaller amounts is to react guanidine carbonate with acids or anhydrides. This reaction can be carried out in aqueous solution or by direct mixing and controlled heating of the pure compounds [19].

This method gives guanidine salts of very high purity (depending on the raw materials), because the only byproducts are carbon dioxide and water, which are lost in the vapor.

1.3. Environmental Protection

Apart from the ecological aspects of the production of raw material (e.g., calcium cyanamide), guanidine production itself causes few environmental problems. Ammonia, the only volatile byproduct, must be removed by proper ventilation and burned or recycled. Solid byproducts are mainly nontoxic melamine salts that can be worked up (for applications, see → Melamine and Guanamines), deposited, or burned.

At moderate concentration, guanidine and its common salts are nontoxic to microorganisms, but nitrification is rather slow [20]. Wastewater containing guanidine can therefore best be removed by slow addition to a sewage plant that does not receive too many other nitrogen-containing compounds.

1.4. Quality Specifications

Guanidine salts are normally available in pure (98.5 – 99.5 %) grade as white crystalline substances containing some ammonium salts (0.3 – 0.6 %), triazines (0.2 – 0.4 %), water (0.1 – 0.3 %), and possibly anticaking agent (SiO_2,

0.2 – 0.4 %). Technical grades are also marketed for some purposes as white to yellowish lumps, granules, or powder, with 91 – 96 % guanidine salt, 2 – 5 % ammonium salt, and 1 – 3 % triazines. Some salts are also available as aqueous solutions, e.g., guanidine hydrochloride (60 %), sulfamate, dihydrogen phosphate, or silicate (each 50 %).

High-purity grades for specific applications are available for guanidine thiocyanate and guanidine hydrochloride. For example, solutions of guanidine hydrochloride can be obtained as iron-free (Fe < 0.1 ppm) or low absorbance products ($\varepsilon_{225} < 0.1$).

1.5. Analysis

Nowadays guanidine salts are assayed routinely by cation chromatography with conductivity detection [21]. Besides other methods such as spectroscopic and colorimetric methods, guanidine salts were formerly analyzed gravimetrically by precipitation as the picrate [22].

1.6. Storage and Transportation

Guanidine salts are packed in polypropylene bags, drums with polyethylene lining, or polypropylene big bags. If dry storage is available, guanidine salts can be stored for an unlimited period. Caking of the guanidine salts however can be serious, leading to a single block of product within a few months. Anticaking agent can be added to reduce caking. Solutions must be stored and shipped in plastic or coated steel vessels to avoid corrosion. Cooling of aqueous solutions may cause precipitation of the guanidine salt.

Guanidine nitrate is a dangerous commodity according to transportation regulations and is classified, along with other nitrates, as a substance that has oxidizing and ignitable properties (class 5.1., 22c, UN 1467). For guanidine hydrochloride no declaration is necessary for transport.

1.7. Legal Aspects

Within the European Union, guanidine hydrochloride and guanidine nitrate must be labeled as hazardous substances as follows: guanidine hydrochloride labeled with the hazard symbol X_n (harmful), R 22 (harmful if swallowed), R 36 – 38 (irritating to eyes and skin), S 22 (do not breathe dust); guanidine nitrate is provisionally labeled by the manufacturer with the hazard symbol X_n, R 22, R 36, and S 26.

Guanidine perchlorate [*10308-84-6*] and picrate [*4336-48-5*] are classified in the Federal Republic of Germany as explosives according to the Explosives Act (Sprengstoffgesetz). Handling and storage of these salts are, therefore, subject to special safety regulations prescribed by this act [23] (→ Explosives).

Although the act does not specifically mention pure guanidine nitrate (or nitroguanidine), some mixtures of these substances with other materials (e.g., potassium nitrate, ammonium dichromate, or chlorinated hydrocarbons) are also classified as explosives [23]. The export of guanidine nitrate or nitroguanidine is controlled by special regulations depending on the country of destination.

1.8. Uses

1.8.1. Pharmaceuticals

Guanidine is a valuable intermediate in the synthesis of a variety of pharmaceuticals. Important industrial processes based on guanidine are, for example, the synthesis of sulfonamides such as Sulfaguanidine [*57-67-0*] [24], Sulfamethazine [*57-68-1*] [25], Sulfadiazine [*68-35-9*] [25], and related compounds like Pyrimethamine [*58-14-0*] [26] and Trimethoprim [*738-70-5*] [27]. In addition, purine derivatives such as Triamterene [*396-01-0*] [28], Methotrexate [*59-05-2*] [29], or Folic acid [*59-30-3*] are manufactured from guanidine. More recently, this list of purines, based on guanidine as a building block, has been extended to include commercially available antiviral agents, e.g., Abacavir [*188062-50-2*], Penciclovir [*145839-78-7*] and Retrovir [*256500-90-0*]. Many syntheses involve 2-aminopyrimidines as intermediates. Promising acylguanidine cardioprotectants are synthesized from guanidine salts. Monoacylguanidines, e.g., Cariporide [*159138-80-4*], are effective Na^+/H^+-exchange inhibitors [30]. The guanidine used is generally prepared in situ by reacting salts such as guani-

dine nitrate, hydrochloride, or carbonate with a strong base such as sodium methoxide.

1.8.2. Plant Protection

Guanidine salts are important building blocks for the syntheses of various classes of pesticides. The sulfonylurea herbicides, developed by DuPont, often contain a 2-amino-4,6-dimethoxypyrimidine fragment derived from guanidine. Prominent examples of these active ingredients which inhibit the syntheses of essential amino acids by blocking the enzyme acetolactate synthase are Bensulfuron-methyl [*83055-99-6*] (Dupont) and Halosulfuron-methyl [*100784-20-1*] (Monsanto, Nissan).

Another class of crop-protection agents which are derived from guanidine compounds are the neonicotinoids, which all contain a nitroguanidine fragment. Examples are Imidacloprid [*138261-41-3*] (Bayer), Thiamethoxam [*153719-23-4*] (Syngenta) and Clothianidine [*210880-92-5*] (Takeda, Bayer).

1.8.3. Cosmetics

Guanidine carbonate is used industrially as a nontoxic, highly soluble component of hair care products.

1.8.4. Textile Impregnation, Flame Retardants, Paper and Resin Manufacture

Guanidine hydrochloride is used for textile impregnation as an antistatic agent. Guanidine phosphates are used as flame retardants for cellulosic products like wood, paper, textiles, and laminates. They decompose above 300°C forming noncombustible gases together with polymetaphosphates [31]. The highly soluble primary guanidine phosphate is used for direct impregnation, while the lesser soluble secondary phosphate often is incorporated in, e.g., bonding resins for laminates. Guanidine sulfamate has large-scale application in the manufacture of flame-retardant papers, especially in Japan where paper walls are common. Guanidine stearate is used as an emulsifier and surface-modifying agent in the production of resin coatings and papers.

1.8.5. Propellants and Explosives

Guanidine nitrate is used nowadays in many non-azide formulations for gas generators in automobile airbag applications [32 – 34]. As a constituent of these pyrotechnic mixtures, guanidine nitrate is not regarded as an oxidizer, but as a gas-generating fuel due to its net oxygen demand upon combustion. Guanidine nitrate is used in large amounts to produce nitroguanidine, which is employed widely in the manufacture of propellants and explosives. This process involves the reaction of guanidine nitrate with concentrated sulfuric acid.

Guanidine perchlorate and picrate have been tested as explosives but are of no industrial importance.

1.8.6. Biotechnological Applications of Guanidine Salts

Guanidine hydrochloride and thiocyanate are among the most potent protein denaturants known [35], readily surpassing urea, the other widely employed denaturing agent, in effectiveness for most applications. The physicochemical mechanism involves strong hydrogen bonding to guanidine, modulating structure and solvent properties of water [36]. Their primary application appears to be in the isolation of proteins and enzymes, particularly from genetically modified microorganisms. After successful manipulation of the genetic constitution of a microorganism, the microbial cell begins to produce the protein of interest, which then accumulates in special cellular compartments known as inclusion bodies. To isolate these proteins, they must be released from the inclusion bodies and separated from other cellular constituents.

Traditionally, urea (6 – 9 M) has been used for this purpose, but guanidine salts, particularly the hydrochloride (2 – 6 M), are gaining acceptance as substitutes with distinct advantages [37]. For example, guanidine hydrochloride tends to cause less protein inactivation [37], partly because lower concentrations can be used, compared with urea. Furthermore, guanidine hydrochloride is removed from the extracted protein after denaturing simply by dilution [38]. The use of guanidine salts generally

results in higher yields of extracted protein than is the case with urea.

In addition no protein modification is observed with guanidine hydrochloride as is the case with cyanate present in urea solutions which leads to carbamoylated lysine in the protein of interest.

Furthermore, recombinant Protein A, bound to a chromatography support for immunoaffinity purification of humanized monoclonal antibodies, is effectively regenerated between purification batches (so-called column stripping).

Guanidine salts have attained an unrivaled position in the extraction and purification of biologically active RNA [39, 40], because of their strong denaturing properties. Guanidine thiocyanate is often the salt of choice and is used frequently in combination with, or in place of, phenol for the extraction of RNA [41, 42]. The above-mentioned advantages make guanidine salts particularly attractive for the isolation of substances such as interferons [37, 43], interleukins [44], and the blood factors tissue plasminogen activator (tPA) and urokinase, as well as humanized monoclonal antibodies like Herceptin or ReoPro.

1.9. Economic Aspects

The worldwide production capacity of guanidine salts is estimated to be several tens of thousands of tons per year. This figure includes significant spare capacity for the production of guanidine nitrate. The exact capacity and output of these plants is not known. In most cases, the amount produced is far less than capacity and is used exclusively as captive raw material for the production of nitroguanidine. Apart from this, worldwide annual consumption of guanidine salts is estimated to be several thousands of tons. The main producers are located in Germany and Japan. Smaller amounts are produced in Austria, India, and the People's Republic of China.

2. Derivatives

For Cyanoguanidines (Dicyandiamides) see → Cyanamides, Chap. 3.

2.1. Nitroguanidine [45 – 47]

Nitroguanidine [*556-88-7*], $CH_4N_4O_2$, M_r 104.07, is a very weak base that forms salts only with strong acids. Its solubility in water at 20 °C (100 °C) is 0.27 % (7.6 %). It decomposes at 220 °C.

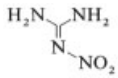

Nitroguanidine is made by treating guanidine nitrate with concentrated sulfuric acid, fuming nitric acid, or both. The reaction of nitroguanidine with amines gives substituted nitroguanidines [45].

Nitroguanidine (available from Nigu Chemie) is used as a component in triple-based propellants, impact-insensitive military ammunition and as a gas generator in automotive airbag applications [48]. Nitroguanidine is also used in anticorrosive phosphate coatings [49].

2.2. Aminoguanidine [50 – 53]

Aminoguanidine [*79-17-4*], CH_6N_4, M_r 74.09, is a base that forms salts with both inorganic and organic acids [50]. It is soluble in water and ethanol. Preparation of the nitrate [*10308-82-4*], $CH_6N_4 \cdot HNO_3$, involves the risk of explosion [54].

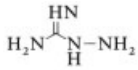

Free aminoguanidine is stable for only a short time; therefore, it is marketed in the form of its stable salts. Commercially, aminoguanidine is offered as the bicarbonate salt [*2582-30-1*], a white powder of usually 99 % assay. The water solubility at room temperature is low (< 5 g/L). It is offered by most world-scale hydrazine producers (→ Hydrazine). Other salts can be prepared from the bicarbonate by reaction with acid.

Aminoguanidine is prepared from hydrazine and aqueous cyanamide solution. Processes based on the reduction of nitroguanidine have no industrial significance.

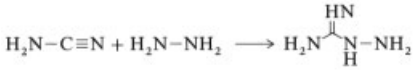

Uses. Aminoguanidine can act as both a hydrazine and a guanidine derivative; it is particularly suitable for the synthesis of various nitrogen-containing heterocyclic compounds such as pyrazoles, 1,2,4-triazoles, thiadiazoles, tetrazoles, pyrimidines, 1,2,4-triazines, thiadiazoles, and tetrazines [53].

Aminoguanidine salts can be used for removing undesired aldehydes by formation of hydrazones, e.g., in oxidation processes or in polymer syntheses with methacrylic acid.

Cyclization with formic acid produces 3-amino-1,2,4-triazole [*61-82-5*]:

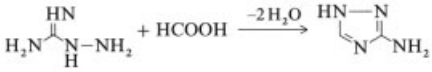

This compound is an important herbicide (Amitrole, ATA), which has a synergistic action with many nonselective herbicides. Some commercially available agrochemicals are based on aminoguanidine, e.g., the fungicide Meferimzone [*89269-64-7*], 4,6-dimethyl-2-[1-(2-methylphenyl)ethylidenehydrazino]-pyrimidinone. 3-Amino-1,2,4-triazole is also used for the preparation of red and yellow cationic triazole dyes such as C.I. Basic Red 22 and C.I. Basic Yellow 25, for poly-(acrylonitrile) fibers (→ Methine Dyes and Pigments). Pharmaceutical applications include Lamotrigine, 3,5-diamino-6-(2,3-dichlorophenyl)-1,2,4-triazine, an anticonvulsant and glutamate inhibitor (Glaxo-Wellcome), and the treatment of type II diabetes with aminoguanidine [56].

Aminoguanidine reacts with nitrous acid in acidic aqueous environment to form 5-amino-(1*H*)-tetrazole [*4418-61-5*] via guanylazide as an intermediate, the former being used a a gas generator, e.g., for automotive airbags [57, 58]. Under neutral pH conditions, diazotization of aminoguanidine leads to 4-(1*H*-tetrazol-5-yl)-1-tetrazene-1-carboxamidine [*539-57-1*], used as a component for shock sensitive explosive primers [59]. Diazotization in glacial acetic acid gives 1,3-di-(tetrazol-5-yl)-triazene [*56929-36-3*] [60].

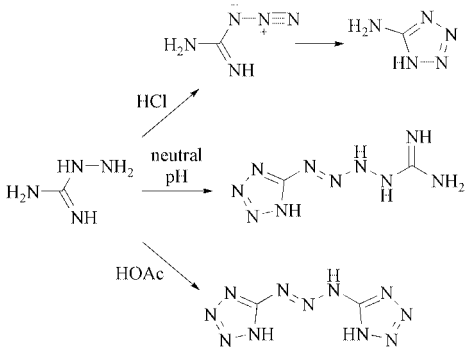

Salts and derivatives of aminoguanidine are recommended as stabilizers for poly(vinyl chloride), as flame retardants, for coatings of pharmaceutical microspheres and as textile auxiliaries.

2.3. Diamino- and Triaminoguanidine

[51, 52, 61]

Diaminoguanidine. Development of the chemistry of *N,N′*-diaminoguanidine [*4364-78-7*], CH_7N_5, M_r 89.10, has been hampered by the lack of convenient methods for synthesizing it in reasonable purity; a summary of the available methods is given in [52]. In preparing of the hydrochloride [*38360-74-6*] by reaction of hydrazine with cyanogen chloride in a molar ratio of 2 : 1, the formation of byproducts can be suppressed if the reaction is carried out in saturated aqueous sodium chloride solution: yield 71 %, *mp* 170 – 176 °C [62].

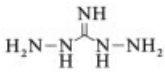

Diaminoguanidine is used in the preparation of robenidine [*25875-51-8*], the hydrochloride [*25875-50-7*] of which is an anticoccidial agent for poultry.

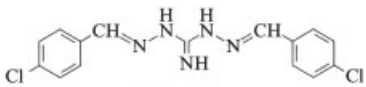

Robenidine

Triaminoguanidine. Salts of *N,N′N″*-triaminoguanidine [*2203-24-9*], CH_8N_6, M_r 104.11, are formed by heating the corresponding guanidine salts with hydrazine (molar ratio 1 :

3) in aqueous solution or in alcohol. The triaminoguanidine base can be liberated from its salts by treatment with sodium hydroxide and can be crystallized from aqueous dimethylformamide.

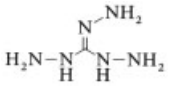

Triaminoguanidine nitrate (TAGN) [*4000-16-2*] is an effective explosive [64]. It is useful as a component of propellants.

2.4. Organoguanidines

A large number of organoguanidine derivatives are known. The many methods of preparation are described in detail in an excellent monograph on carbon dioxide derivatives [65, 66].

The addition of amines, polyamines, hydrazines, or hydroxylamines to cyanamides is one of the most common industrial processes for the synthesis of guanidines. Cyanamide [*420-04-2*] is available as a 50 % aqueous solution or in crystalline form; substituted cyanamides can be obtained from amines and cyanogen chloride (→ Cyanamides).

The classical preparation of guanidines by the action of amines on *S*-alkylisothioureas is versatile and suitable also for the preparation of cyclic compounds. The elimination of alkanethiol byproducts is, however, a disadvantage.

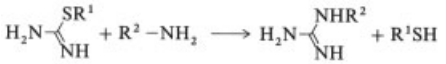

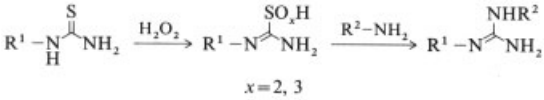

A method for synthesizing di- and trisubstituted guanidines is the reaction of oxidized thioureas with nucleophilic amines [67].

An advanced reagent to produce *N′*,*N″*-bisprotected guanidines from various amines is *N*,*N′*-di-BOC-*N″*-triflylguanidine ("Goodman-reagent", [*207857-15-6*]) [68].

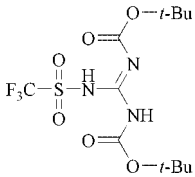

2.4.1. Alkylguanidines

Monosubstituted and *N,N*-disubstituted compounds, including guanidinocarboxylic acids (e.g., creatine [*57-00-1*]), are produced industrially by the reaction of cyanamide with the corresponding amines or aminocarboxylic acids.

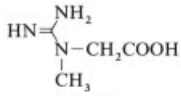

Creatine

However, *l*-arginine (→ Amino Acids), the most frequently found guanidino compound in nature, can be best isolated from proteins.

For convenient guanylation (amidination) of primary alkylamines and amino acids without formation of methanethiol, *S*-methylisothiourea sulfate (2 : 1) [*867-44-7*] can be replaced by commercially available *O*-methylisourea sulfate (2 : 1) [*52328-05-9*]. Thus, methylguanidine sulfate (2 : 1) [*598-12-9*] and ethylguanidine sulfate (2 : 1) [*3482-86-8*] are obtained in quantitative yield by the action of methylamine and ethylamine, respectively, on *O*-methylisourea sulfate in aqueous solution at 50 °C:

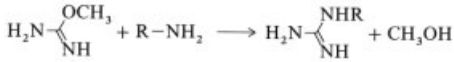

For steric reasons, *O*-methylisourea salts react preferentially with primary amines and are particularly suitable for selective guanylation of the terminal amino groups of polyamines. *O*-Methylisourea sulfate (available from Degussa AG) is a convenient guanylating agent and is also used on an industrial scale for the synthesis of heterocyclic compounds, e.g., benzimidazole anthelmintics and the antibacterial agents pipemidic acid [*51940-44-4*] and piromidic acid [*19562-30-2*].

Guanidine Bases. Free alkylguanidines are strong bases that are more stable than unsubstituted guanidine.

N,N,N′,N′-Tetramethylguanidine (TMG) [*80-70-6*], $C_5H_{13}N_3$, M_r 115.18, is a stable, unusually strong base, pK_a 13.6 (25 °C), with a free imino group. It is a colorless to light-yellow liquid, miscible with water and common organic solvents: 0.92 g/cm^3 (25 °C), *bp* 160 °C, vapor pressure 0.03 kPa (25 °C). Prepared from cyanogen chloride and dimethylamine in a 1 : 2 molar ratio [69], TMG is used as a polyurethane foam catalyst, in the preparation of pigments for color photography, and in steroid synthesis. It is available from Degussa AG and Lonza.

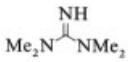

Undiluted *N,N,N′,N′*-tetramethylguanidine strongly irritates the eyes, skin, and mucous membranes. Its LD_{50} is ca. 850 mg/kg (rat, oral).

Salts of tetramethylguanidine [70] or salts of cyclic, peralkylated guanidines [71] have been proposed as ionic liquids for solvent use in "green" chemical reactions.

Short-Chain Alkylguanidines as Building Blocks in Synthesis. Some monosubstituted and *N,N*-disubstituted derivatives are used as intermediates for the preparation of crop-protecting agents.

Pyrimidines can be prepared by condensation of alkylguanidines with a 1,3-difunctional C_3 fragment. Cyclization of alkylguanidines with 2-alkylacetoacetic acid esters produces fungicidally active 5-alkyl-2-alkylamino-4-hydroxy-6-methylpyrimidines. The systemic fungicides bupirimate [*41483-43-6*] and dimethirimol [*5221-53-4*], developed by ICI, belong to this class of compounds.

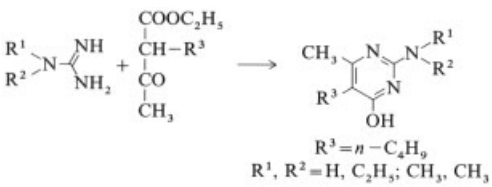

N,N-Diethylguanidine [*18240-93-2*] and *N,N*-dimethylguanidine [*6145-42-2*] are used as an N−C−N building block for the pyrimidine insecticides pirimiphos-methyl [*29232-93-7*], pirimiphos-ethyl [*23505-41-1*], and pirimicarb [*23103-98-2*]. Pirimiphos-methyl has proved particularly effective for the control of storage pests.

Biologically Active Compounds. Higher alkylguanidines and polyguanidines, because of their fungicidal and antibacterial action, are economically important as fungicides in agriculture or as antimicrobial agents in disinfectants and industrial biocides [72, 73].

n-Dodecylguanidine acetate [*2439-10-3*], Dodine, used primarily for the control of apple and pear scab, exhibits both protective and curative action. Dodine is also a bacteriostatic agent. Dodecylguanidine derivatives are recommended as effective biocides for controlling biological growth in industrial coolant systems.

Other broad-spectrum microbiocidal uses based on alkylguanidine salts include disinfection, sanitation, preservation, and control of *Desulfovibrio desulfuricans* corrosion [73].

Guazatine, prepared by the guanylation of aliphatic diamines and polyamines, is used for cereal seed dressing. It is also used as a postharvest dip in citrus growing. Guazatine is effective against wood-discoloring fungi and, together with quaternary ammonium compounds, is a constituent of wood preservatives.

Based also on guazatine, contact fungicides were developed which control, e.g., the growth of *Septoria nodorum* in wheat [74].

Iminoctadine triacetate is a broad-spectrum contact fungicide, developed by Dainippon Ink and Chemicals (DIC); the active ingredient is the triacetate salt of 1,1′-[iminobis(octamethylene)]diguanidine [*57520-17-9*]. The product is used in Japan primarily for the control of top fruit (tree fruit) and citrus diseases; in Western Europe, good action has been achieved against leaf spot and glume blotch in winter wheat [75].

Other Uses. Guanidinium salts of unsaturated fatty acids are used as corrosion inhibitors [76]. Alkylguanidines linked to a styrene polymer matrix have been proposed as phosphate scavengers for pharmacological treatment of hyperphosphatemia [77].

2.4.2. Arylguanidines

Monoarylguanidines can be prepared by heating aromatic amine hydrochlorides with

excess cyanamide in aqueous solution. Arylbiguanides are common byproducts. Phenylguanidines can be precipitated from the reaction mixture as a carbonate or bicarbonate salt [78].

Phenylguanidine has found application in the production of pyrimidine fungicides like Cyprodinil [*121552-61-2*] [79], Pyrimethanil [*53112-28-0*] [80], and Mepanipyrim [*110235-47-7*] [81].

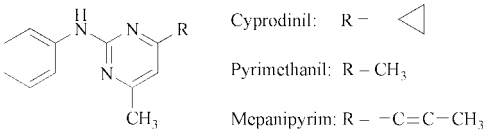

From the group of 2-guanidinoanilides with anthelmintic properties, febantel [*58306-30-2*] (Rintal), supplied by Bayer, has emerged as an effective broad-spectrum anthelmintic; it is used to treat nematode infections in domestic animals.

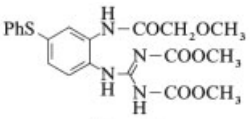

Febantel

The guanidine derivative linogliride fumarate (1 : 1) [*78782-47-5*] is a new oral antidiabetic, developed by McNeil. It can be prepared by the Maryanoff process [67].

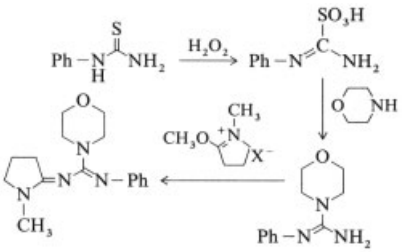

Diarylguanidines. *N,N′*-Diphenylguanidine (DPG) [*102-06-7*] and *N,N′*-di-*o*-tolylguanidine (DOTG) [*97-39-2*] are used as vulcanization accelerators. Because of their low activity they do not play a major role as primary accelerators; however, they are very important in combination with other accelerators. Diarylguanidines can be used as secondary accelerators for thiazoles, sulfenamides, thiurams, and dithiocarbamates; the most frequently used combination is with mercapto accelerators. These accelerators and the guanidine produce an extraordinary increase in the degree of cross-linking and the rate of vulcanization; high mechanical values (high tensile stress and elasticity) are achieved. *N,N′*-Diphenylguanidine is the most important of the three guanidines DPG, DOTG, and OTBG (*o*-tolylbiguanide; see Section 2.5).

Both DPG and DOTG are prepared by reaction of cyanogen chloride with aniline or *o*-toluidine, respectively, in a molar ratio of 1 : 2.

Cyclic Derivatives. *N,N′-o*-Phenyleneguanidine [*934-32-7*], 2-aminobenzimidazole, $C_7H_7N_3$, M_r 133.15, *mp* 231 – 232 °C (decomp.), obtained by cyclization of *o*-phenylenediamine with cyanogen chloride or cyanamide, is the nucleus for a number of compounds of economic interest; for a review of 2-aminobenzimidazoles, see [82].

Several 2-(methoxycarbonylamino)benzimidazoles exhibit biological activity. Carbendazim (R = R′ = H) [*10605-21-7*] is an important systemic fungicide, while its derivative Benomyl (R = H, R′R =R CO–NH–C_4H_9) [*17804-35-2*] has been withdrawn from the market recently.

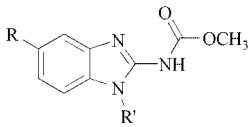

Effective anthelmintics, mainly for veterinary use, are:

Albendazole [*54965-21-8*]	(R = S–C_3H_7, R′ = H)
Fenbendazole [*43210-67-9*]	(R = S–Ph, R′ = H)
Flubendazole [*31430-15-6*]	(R = CO–C_4H_6–p–F, R′ = H)
Mebendazole [*31431-39-7*]	(R = CO–Ph, R′ = H)
Oxfendazole [*53716-50-0*]	(R = SO–Ph, R′ = H)
Oxibendazole [*20559-55-1*]	(R = O–C_3H_7, R′ = H)

Albendazole and flubendazole are also effective against worm infection in humans (→ Anthelmintics, Chap. 3.0.0.0.0.7). Substituted 2-(methoxycarbonylamino)benzimidazoles can be prepared by cycloguanylation of the corresponding *o*-phenylenediamine derivatives with *N*-methoxycarbonyl-*S*-methylisothiourea [*39259-32-0*] or *N*-methoxycarbonyl-*O*-methylisourea [*40943-37-1*].

2.5. Biguanide and Derivatives

Biguanide [*56-03-1*], $C_2H_7N_5$, M_r 101.11, *mp* 130 °C (which decomposes violently at 142 °C), is a moderately strong base (pK_b 2.9). It is readily soluble in water and ethanol, and insoluble in ether, benzene, and chloroform. The free base decomposes gradually in aqueous solution. Biguanide forms stable salts with most acids.

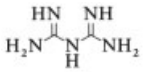

Reviews of methods for the preparation of biguanide can be found in [83, 84]. Biguanide salts are prepared by the fusion of dicyandiamide with ammonium salts. [Dibiguanide – copper(II)] sulfate, $[Cu(C_2H_7N_5)_2]SO_4$, is formed by heating dicyandiamide with ammoniacal copper sulfate solution to 90 – 110 °C in a pressure vessel [85]; it can be converted by decomposition with hot 10 % sulfuric acid or hydrogen sulfide to biguanide sulfate (1 : 1) [*6945-23-9*], $C_2H_7N_5 \cdot H_2SO_4$, or biguanide sulfate (2 : 1) [*49719-55-3*], 2 $C_2H_7N_5 \cdot H_2SO_4$ [85, 86].

Biguanide is used to prepare 6-substituted 2,4-diamino-1,3,5-triazines [84]. The reaction of biguanide with esters to form guanamines is catalyzed by sodium alkoxides.

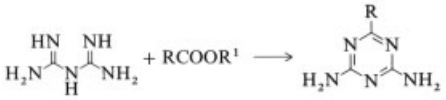

Biguanides substituted at the N-1 position are obtained by heating dicyandiamide with aliphatic or aromatic amine hydrochlorides in water, higher boiling alcohols, or in a melt reaction.

N-substituted biguanide hydrochlorides have found widespread application in pharmacology. Buformin [*692-13-7*], Phenformin [*114-86-3*] and Metformin [*657-24-9*] are used against type II diabetes, the latter being sold in quantities of more than 1000 tons per year worldwide (→ Antidiabetic Drugs, Chap. 3).

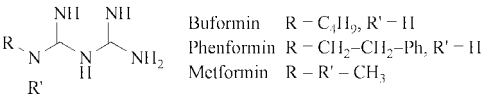

o-Tolylbiguanide (OTBG) [*93-69-6*], $C_9H_{13}N_5$, M_r 191.23, *mp* 144 °C, is used as a vulcanization accelerator in the same manner as the guanidines DPG and DOTG (see Section 2.4.2.). An important use is in the preparation of consumer articles that come in contact with food and drinking water (Recommendation XXI of the German Federal Department of Health). *o*-Tolylbiguanide is also employed as a curing agent with moderate reactivity for color-stable, highly glossy or transparent epoxy resin powder coatings.

The N-1,N-5-disubstituted biguanides can be prepared by addition of amines to substituted cyanoguanidines. A particularly convenient synthesis of N1-alkyl-N-5-aryl biguanides involves heating aliphatic cyanoguanidines with aromatic amine hydrochlorides [87, 88]: the antimalarial drug proguanil [*500-92-5*] is obtained as the hydrochloride [*637-32-1*] (Paludrine) by reaction of isopropyldicyandiamide with 4-chloroaniline hydrochloride in boiling water [87].

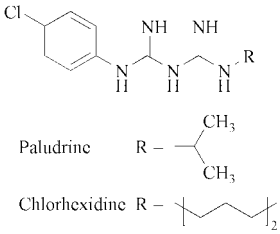

Chlorhexidine [*55-56-1*], 1,6-bis(4-chlorophenylbiguanido)hexane, $C_{22}H_{30}Cl_2N_{10}$, M_r 505.45, *mp* 134 °C, is a widely used antiseptic (→ Disinfectants). It can be prepared, for example, from 1,6-hexamethylenebis(dicyandiamide) and 4-chloroaniline hydrochloride [89]. Chlorhexidine is used primarily as its salts (e.g., the dihydrochloride [*3697-42-5*], diacetate [*56-95-1*], and digluconate [*18472-51-0*]) in disinfectants (disinfection of the skin and hands), cosmetics (additive to creams, toothpaste, deodorants, and antiperspirants), and pharmaceutical products (preservative in eyedrops, active substance in wound dressings and antiseptic mouthwashes) [73].

Poly(hexamethylene biguanide) hydrochloride (PHMBG, [*27083-26-7*]) is a broad-spectrum industrial biocide developed by ICI. The most important uses are in surface disinfection in hospitals and the food and beverage

industry, and in the preservation of aqueous systems (i.e., aqueous solutions of organic products, such as cutting oils and domestic cleaning agents) [73]. PHMBG also is used to make up spa water and private swimming pools, where it acts as a flocculant for organic biogenic colloids. Polymeric biguanide hydrochlorides with broad antimicrobial action, very low toxicity, and good skin tolerance are also contained in preservatives for cosmetic formulations.

3. Toxicology and Occupational Health

The guanidino group is low in acute toxicity and is not associated with severe health hazards. No mutagenic, carcinogenic, or teratogenic potential has been demonstrated for guanidine.

Free guanidine and its solutions exhibit strong alkaline reactions. They must therefore be handled like alkali metal hydroxides or alkoxides. Guanidine salts of weak acids (e.g., guanidine silicate) also have alkaline properties. Animal tests (rabbit) show that guanidine hydrochloride is a severe primary skin and eye irritant [7, 90]. In a similar test, guanidine nitrate produced mild to moderate eye irritation [91]. Guanidine hydrochloride and guanidine nitrate did not show sensitizing properties in the Buehler test [92, 93].

Toxicities of some guanidine salts (acute oral LD_{50}, rat) are guanidine hydrochloride 1120 mg/kg, guanidine nitrate 1260 mg/kg, guanidine carbonate 1020 mg/kg, guanidine sulfamate 2334 mg/kg, and guanidine stearate 7500 mg/kg. Further toxicological data for guanidine, its salts, and derivatives can be found in [93, 94].

The toxicological properties of individual guanidine derivatives depend primarily on the type of the substitutents. Commercially available guanidines generally have relatively low toxicity. The following are LD_{50} values (rat, oral) for some derivatives: aminoguanidine bicarbonate >5000 mg/kg; nitroguanidine 4650 mg/kg; *N,N,N′,N′*-tetramethylguanidine ca. 850 mg/kg; *n*-dodecylguanidine acetate 1000 mg/kg; *N,N′*-diphenylguanidine 507 – 850 mg/kg; *o*-tolylbiguanide 1800 mg/kg.

Guanidine derivatives may, however, contain amines as impurities, because amines are used as starting material in the production of guanidines.

With the exception of nitroguanidine, cyanoguanidines, and acylated guanidines, guanidine derivatives are basic; the free bases can cause irritation and corrosion of the skin, eyes, mucous membranes, and respiratory tract. Skin and eye contact with the strong free bases should be avoided. The salts are less irritating than the free bases. Exposure limits (MAK and TLV values) have not been established.

4. References

1. *Beilstein* System no. 207.
2. A. Mori, B. D. Cohen, A. Lowenthal (eds.): *Guanidines – Historical, Biological, Biochemical, and Clinical Aspects of the Naturally Occuring Guanidino Compounds*, Plenum Press, New York 1985.
3. R. G. S. Berlinck, *Natural Products Reports* **19** (2002) 617.
4. R. G. S. Berlinck, *Natural Products Reports* **13** (1996) 377.
5. *Ullmann*, 3rd ed., **8**, 328 – 337.
6. *Ullmann*, 4th ed., **12**, 441 – 420.
7. *Kirk-Othmer*, 3rd ed., suppl. vol., 514 – 521.
8. *Gmelin*, **14 D1**, 466 – 475.
9. G. Desseigne, A. Audiffren, *Meml. Poudres* **35** (1953) 15 – 38.
10. Südd. Kalkstickstoffwerke, DE-OS 1 593 472, 1966 (H. Michaud, H. Prietzel).
11. Pittsburgh Coke & Chem. Co., US 2 949 484, 1957 (J. S. Mackay).
12. Industrie Chemie Thoma GmbH, DE 3236221, 1984 (M. Thoma).
13. Projects and Development India Ltd., IN 157293, 1986, CA 105:210788.
14. SKW Trostberg AG, DE 3744358, 1989 (C. Grambow, W. Kristof, P. Reitsamer, K. Scheinost).
15. Lentia GmbH, DE-OS 2 234 732, 1974 (A. Schmidt, K.-H. Wegleitner, J. H. Hatzle, R. Sylpra, F. Weinrotter);
    DE-OS 2 435 167, 1976 (F. Weinrotter, A. Schmidt, K. Wegleitner, A. Garber, J. H. Hatzl, R. Sykora).
16. Degussa AG, DE 3007530, 1981 (P. Schalke, M. Langer, K. Huthmacher).
17. G. Desseigne, *Bull. Soc. Chim. Fr.* 1955, 1193 – 1194.
18. Du Pont, US 3 475 375, 1969 (P. C. Yates).
19. Goldschmidt, DE-OS 1 922 043, 1970 (E. Ruf).

20. H. Lees, J. H. Quastel, *Biochem. J.* **40** (1946) 825.
21. Degussa AG, SOP-Method 009–009/2, 1999 (M. Wildenauer).
22. P. Fainer, J. L. Myers, *Anal. Chem.* **24** (1952) no. 3, 515 – 517.
23. Sprengstoffgesetz, annex I, part 1, no. 33, 34 and part 2, no. 2.5, mixtures 4,5.
24. American Cyanamid, US 2 233 569, 1941 (P. S. Winnek).
25. Sharp & Dohme, US 2 407 966, 1946 (J. M. Sprague).
26. Burroughs Wellcome, US 2 576 939, 1951 (G. H. Hitchings, P. B. Russell, E. A. Falco).
27. Burroughs Wellcome, US 2 909 522, 1959 (G. H. Hitchings, B. Roth).
28. Smith Kline & French, US 3 081 230, 1963 (J. Weinstock, V. D. Wiebelhaus).
29. D. R. Seeger, D. B. Cosulich, J. M. Smith, M. E. Multquist, *J. Am. Chem. Soc.* **71** (1949) 1753 – 1758.
30. H. Strömer et al., *Eur. Heart J.* 1999, 20 (Suppl.);
A. Sugiyama et al., *J. Cardiovasc. Pharmacol.* **33** (1998) no. 1, 116.
31. K. Fukushima et al., Nippon Kagaku Kaishi (1996), 146, CA 124:231710.
32. TRW Airbag Systems GmbH, US 6004411, 1999 (S. Zeuner, A. Hofmann, R. Schropp, K.-H. Rodig).
33. Nippon Kayaku K.K., WO 2004/048296 (T. Kawakami, E. Sato, K. Wakisaka, K. Tamura).
34. SNPE Materiaux Energetiques, FR 2866022, 2005 (G. Chounet, E. Giraud, B. Thibieroz).
35. Y. Nozaki, C. Tanford, *J. Biol. Chem.* **245** (1970) 1648 – 1672.
36. R. D. Mountain, D. Thirumalai, *J Physical Chemistry B* **108** (2004) 19711.
37. Hoffmann-La Roche Inc., US 4 476 049, 1984 (Hsiang Fu Kung).
38. Hoechst, DE 3 511 011, 1987 (R. Obermeier).
39. J. R. Chirgwin, A. E. Przybyla, R. J. McDonald, W. J. Rutter, *Biochemistry* **18** (1979) 5294 – 5299.
40. J. Logemann, J. Schell, L. Willmitzer, *Anal. Biochem.* **163** (1987) 16 – 20.
41. P. Chomczynki, N. Sacchi, *Anal. Biochem.* **163** (1987) 156 – 159.
42. Merck Patent GmbH, DE 10358137, 2005 (K. Holschuh, U. Michelsen, A. Schwaemmle).
43. Takeda Chem. Ind., EP 150 066, 1985 (Kiyoshi Nara, Susumi Honda).
44. Hoffman-La Roche & Co., EP 200 986, 1986 (U. Gubler, P. T. Lomedico, S. B. Mizel).
45. T. Urbanski (ed.): *Chemistry and Technology of Explosives*, vol. **3**, Pergamon Press, Oxford 1967, pp. 22 –33.
46. T. Urbanski (ed.): *Chemistry and Technology of Explosives*, vol. **4**, Pergamon Press, Oxford 1984, pp. 365 –367.
47. R. Meyer: *Explosives*, 3rd ed., VCH Verlagsgesellschaft, Weinheim 1987, pp. 243 – 245.
48. Nigu Chemie GmbH, WO 02/51773 (E. Gast, B. Schmid, C. Recker, S. Walz, T. Mayr, P. Semmler).
49. Henkel KGaA, EP 1310578, 2003 (S. Hohagen, K. Lepa, J. Kroemer, P. Droniou).
50. *Beilstein*, 3, H 117, E I 57, E II 95, E III 230, E IV 236.
51. E. Lieber, G. B. L. Smith, *Chem. Rev.* **25** (1939) 213 – 271.
52. F. Kurzer, L. E. A. Godfrey, *Chem. Ind. (London)* 1962, 1584 – 1595.
53. F. Kurzer, L. E. A. Godfrey, *Angew. Chem.* **75** (1963) 1157 – 1175.
54. H. Koopman, *Chem. Weekbl.* **53** (1957) 97.
55. Atochem SA, EP 168296, 1986 (M. Alicot).
56. Smithkline-Beecham, WO 98 356 607, 1998 (N. C. Turner, V. Piercy).
57. Olin Corp., US 5424449, 1995 (E. F. Rothgery, K. O. Knollmueller).
58. Nippon Carbide Kogyo K K, EP 669325, 1995 (M. Murotani, H. Mura, M. Takeda, H. Shibafuchi).
59. S. H. Patinkin, J. P. Horwitz, E. Lieber, *J. Am. Chem. Soc.* **77** (155) 562.
60. J. Reilly, J. P. Teegan, M. F. Carey, *Sci. Proc. Roy. Dublin Soc.* **24** (1948) 349, CA 43:8364.
61. *Beilstein*, 3, H 122, E I 57, E II 97, E III 232, E IV 241/242.
62. American Cyanamid, US 2 721 217, 1955 (G. A. Peters, D. W. Kaiser).
63. Rockwell I. C., DE-OS 2 453 167, 1976 (V. E. Haury).
64. W. Sauermilch, *Explosivstoffe* **12** (1964) no. 9, 197.
65. *Houben-Weyl*, **E 4**, 608 – 624.
66. Houben-Weyl, *Science of Synthesis*, vol. 18, 2005, p. 1077 ff.
67. C. A. Maryanoff, R. C. Stanzione, J. N. Plampin, *Phosphorus Sulfur* **27** (1986) 221.
A. E. Miller, J. J. Bischoff, *Synthesis* 1986 no. 9, 777.
68. J. T. Baker, M. Tomioka, M. Goodman, *Organic Syntheses* **78** (2002) 91.
69. Mitsubishi Gas Chemical Co., EP 1072585, 2001 (T. Hashimoto, T. Motoi).
70. H. Gao et al., *Synth. Commun.* **34** (2004) 3083.

71. Merck Patent GmbH, WO 2004/106288 (U. Welz-Biermann, N. Ignatyev, H. Willner, G. Bissky).
72. H. R. Hudson, I. A. O. Ojo, M. Pianka, *Int. Pest Control* **28** (1986) no. 6, 148 – 155.
73. K. H. Wallhäußer: *Praxis der Sterilisation-Desinfektion-Konservierung-Keimidentifizierung-Betriebshygiene*, 4th ed., Thieme Verlag, Stuttgart – New York 1988, pp. 558 – 564.
74. W. Wimschneider, A. O. Klomp, K. E. Friedrich, *Gesunde Pflanz.* **39** (1987) no. 6, 263.
D. G. Cameron, I. Hylten-Cavallius, E. Jordow, A. O. Klomp, W. Wimschneider, *Proc. Br. Crop Prot. Conf. Pests Dis.* 1986, vol. **3**, p. 1201.
75. M. Masui, N. Yoshioka, R. I. Harris, H. Roos, A. Cornier, *Proc. Br. Crop Prot. Conf. Pests Dis.* 1986, vol. **1**, p. 63.
76. Henkel KGaA, DE 4436764, 1996 (J. Geke, H.-D. Speckmann, B. Stedry, A. Westfechtel).
77. British Technology Group Ltd., WO 96/25440 (R. C. Hider, B. L. Goodwin).
78. Ciba-Geigy AG, EP 560726, 1993 (R. Waditschatka).
79. Ciba-Geigy AG, EP 655441, 1995 (W. Baettig, R. G. Hanreich).
80. Hoechst Schering AgrEvo GmbH, EP 717038, 1996 (H.-J. Ressel, G. Schlegel).
81. T. Nagata, K. Masuda, S. Maeno, I. Miura, *Pest Management Science* **60** (2004) 399.
82. R. Rastogi, S. Sharma, *Synthesis* 1983, no. 11, 861 – 882.
83. P. Ray, *Chem. Rev.* **61** (1961) 313 – 359.
84. F. Kurzer, E. D. Pitchfork, *Fortschr. Chem. Forsch.* **10** (1968) no. 3, 375 – 472.
85. K. Rackmann, *Justus Liebigs Ann. Chem.* **376** (1910) 163, 170.
86. D. Karipides, W. C. Fernelius, *Inorg. Synth.* **7** (1963) 56.
87. F. H. S. Curd, J. H. Hendry, T. S. Kenny, A. G. Murray, F. L. Rose, *J. Chem. Soc.* 1948, 1630.
88. G. Rembarz, H. Brandner, H. Finger, *J. Prakt. Chem.* **26** (1964) 314.
89. Sogo Pharmaceutical Co., DE-OS 2 932 951, 1980 (S. Uchikuga, M. Itoh, K. Nagahama).
90. E. W. Morgan, L. Mullen, D. W. Korte, *Gov. Rep. Announce. Index U.S.*, **86** (1986) (15), Abstract no. 633,219;
Chem. Abstr. **106** (1987) 62 436.
91. E. W. Morgan, J. W. Bauserman, D. W. Korte, jr., *Gov. Rep. Announce. Index U.S.*, **86** (1986) (15), Abstract no. 633,223;
Chem. Abstr. **106** (1987) 45 362.
92. G. F. S. Hiatt, E. W. Morgan, D. W. Korte, *Gov. Rep. Announce. Index U.S.*, **86** (1986) (12), Abstract no. 625,904;
Chem. Abstr. **106** (1987) 151 112.
93. IUCLID, data sheets on CD-ROM, European Chemical Bureau (Ispra, Italy), 1999.
94. NIOSH Registry of Toxic Effects of Chemical Substances, Washington, D.C.
95. Nippon Carbide Industries, JP 51136623, 1976 (H. Horikiri, Y. Tobise).
96. Z. Zhou, J. Xie, X. Zhou, Y. Jiang, *Huaxue Shiji* **22** (2000) 314, CA 134:178267.

Hafnium and Hafnium Compounds

RALPH H. NIELSEN, Teledyne Wah Chang Albany, Albany, Oregon 97321, United States

1.	**Introduction**	1
2.	**Physical Properties**	1
3.	**Chemical Properties**	2
4.	**Occurrence and Raw Materials**	2
5.	**Production**	3
5.1.	**Ore Cracking**	3
5.2.	**Separation of Zirconium**	3
5.3.	**Winning of Metallic Hafnium**	6
5.4.	**Refining**	7
6.	**Environmental Protection**	7
7.	**Quality Specifications**	7
8.	**Analysis**	7
9.	**Storage and Transportation**	8
10.	**Uses**	8
11.	**Economic Aspects**	9
12.	**Hafnium Compounds**	9
12.1.	**Hafnium Borides**	9
12.2.	**Hafnium Tetrahydroborate**	9
12.3.	**Hafnium Carbide**	9
12.4.	**Hafnium Halides**	9
12.5.	**Hafnium Hydride**	10
12.6.	**Hafnium Nitride**	10
12.7.	**Hafnium Dioxide**	11
12.8.	**Other Derivatives**	11
13.	**Toxicology and Occupational Health**	11
14.	**References**	11

1. Introduction

The element hafnium [*7440-58-6*], Hf, is always found in zirconium ore and is available as a by-product of the production of zirconium metal. The average concentration of hafnium in the earth's crust is estimated at 2.8 – 4.5 ppm. Six isotopes occur naturally: ^{174}Hf (0.16 %), ^{176}Hf (5.2 %), ^{177}Hf (18.6 %), ^{178}Hf (27.1 %), ^{179}Hf (13.7 %), and ^{180}Hf (35.2 %). Hafnium is used principally as a minor alloying element in nickel-based superalloys and as a thermal neutron absorber in nuclear power reactors. Annual world consumption is about 50 t.

2. Physical Properties

Hafnium is a heavy, hard, ductile metal similar in appearance to stainless steel. Although chemically very similar to zirconium, it has several physical differences: hafnium has twice the density of zirconium, a higher phase transition temperature, and a higher melting point. Hafnium also has a high thermal neutron absorption coefficient (1.04×10^{-26} m^2), whereas that for zirconium is very low (1.8×10^{-29} m^2). Some physical properties of hafnium are listed below:

Atomic number	72
Relative atomic mass A_r	178.49
mp	2500 K (2227 °C)
bp	4875 K (4602 °C)
Density	13.31 g/cm^3
Thermal conductivity (298 K)	23.0 W m^{-1} K^{-1}
Coefficient of linear expansion (273 – 1273 K)	5.9×10^{-6} K^{-1}
Specific heat (298 K)	117 J kg^{-1} K^{-1}
Vapor pressure	
(2040 K)	10^{-5} Pa
(2280 K)	10^{-4} Pa
Electrical resistivity (298 K)	3.51×10^{-7} Ω · m
Thermal neutron absorption cross section	1.04×10^{-26} m^2
Crystal structure;	
α-form	hexagonal close-packed (HCP)
β-form	body-centered cubic (BCC)
Temperature of $\alpha-\beta$ transformation	2033 K (1760 °C)

3. Chemical Properties

The ionic radii of hafnium and zirconium are almost identical because of the lanthanide contraction. Both elements exhibit a valence of four. Therefore, the chemistry of hafnium is similar to that of zirconium, and the elements are always found together in nature (see also →Zirconium and Zirconium Compounds).

Elemental hafnium reacts with hydrogen (>250 °C), carbon (>500 °C), and nitrogen

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_559

(>900 °C) to form brittle, nonstoichiometric interstitial compounds with metal-like conductivity. In molten salts, hafnium is normally quadrivalent, but in anhydrous molten halide salts, it can be reduced to hafnium(III) and hafnium(II).

In aqueous solution, hafnium is always quadrivalent, with a high coordination number (6, 7, or 8). In dilute acid, hafnium slowly hydrolyzes and polymerizes. Hafnium hydrous oxide precipitates from aqueous solution at ca. pH 2. The only inorganic compounds with significant solubility in neutral or slightly basic aqueous solution are the ammonium – hafnium carbonate and potassium – hafnium carbonate complexes. The tendency for hafnium to form inorganic complexes with anions decreases in the following order: $OH^- > F^- > PO_4^{3-} > CO_3^{2-} > SO_4^{2-} > NO_3^- \approx Cl^- > ClO_4^-$.

Reviews of zirconium chemistry [1,2] are frequently useful for indications of the chemical behavior of hafnium, because of the similarity in chemical properties.

4. Occurrence and Raw Materials

Mineralogically, hafnium is always found with zirconium. Although about 40 known minerals contain these elements, the main commercial sources of zirconium and hafnium are *zircon* [*1490-68-2*] and *baddeleyite*, since they are available as byproducts in the recovery of other minerals. Zircon sand is obtained during the processing of alluvial heavy mineral sands to recover the titanium minerals rutile and ilmenite. Commercially recoverable deposits of these heavy mineral sands are found in China, Malaysia, Thailand, India, Sri Lanka, Australia, South Africa, Madagascar, and the United States. Clean (high quality) zircon contains 64 % zirconium oxide, 34 % silicon dioxide, 1.2 % hafnium dioxide, with the balance including aluminum, iron, phosphorus, rare earths, titanium, uranium, and thorium.

Worldwide, the hafnium content of zircon is usually ca. 2 % of the zirconium content. A notable exception is a Nigerian zircon, available in commercial quantities during 1957 – 1965, that contained 6 % hafnium (based on zirconium). Two minerals, a Norwegian thortveitite and a Mozambiquan hafron, have been reported to contain more hafnium than zirconium. Baddeleyite usually contains hafnium at <2 % of the zirconium content. At the Phalaborwa complex in the Republic of South Africa, baddeleyite is recovered from process tailings in the extraction of copper and the extraction of apatite for phosphate fertilizers.

Although hafnium reserves are significant, demand is low and easily met as a byproduct of the production of reactor-grade zirconium. Known hafnium reserves are listed in Table 1.

5. Production

Processing steps and alternative processes for the production of hafnium metal from zircon sand are shown in Figure 1.

5.1. Ore Cracking

Zircon is a refractory mineral, whose decomposition requires the use of aggressive chemicals and high temperature. Today, the commonly used techniques are caustic fusion, direct carbochlorination, or thermal dissociation; however, other procedures such as fusion with

Table 1. Known hafnium reserves, in 10^3 t

Country	Annual zircon production	Zircon reserve *	Hafnium content of zircon reserve
Australia	480	32 000	390
United States	135	15 000	145
South Africa	135	15 000	145
Brazil	18	4000	38
India	15	9000	85
Sri Lanka	5	3000	29
Malaysia	3	3000	29

* Estimated.

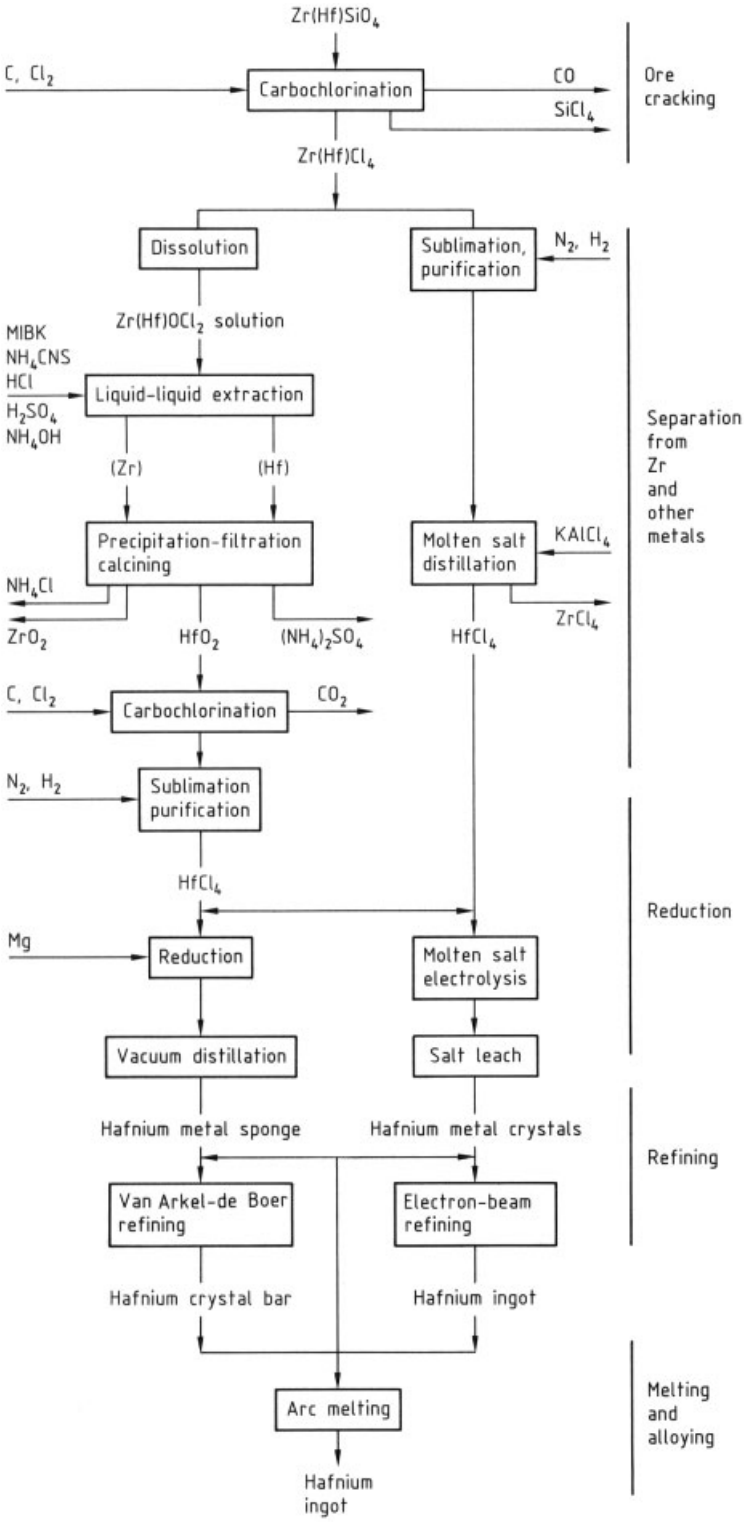

Figure 1. Flow diagram for production of hafnium

dolomite, fusion with potassium hexafluorosilicate, and a two-step process involving carbiding and exothermic chlorination of the crude carbide have all been used extensively to obtain zirconium and hafnium in more chemically active form.

The caustic fusion of zircon by using a slight excess of sodium hydroxide at 600 °C is the most common zircon ore-cracking process [3]:

$$ZrSiO_4 + 4\,NaOH \rightarrow Na_2ZrO_3 + Na_2SiO_3 + 2\,H_2O$$

The cooled reaction mass is broken up and slurried in water, which dissolves sodium silicate and hydrolyzes sodium zirconate to soluble sodium hydroxide and insoluble hydrous zirconia; the latter is recovered by filtering the slurry. Hydrous zirconia is soluble in strong mineral acid; hafnium can be recovered from the primary zirconium stream by ion exchange or liquid – liquid extraction.

Variations of this fusion process involve the use of sodium carbonate at 900 °C or the use of only half the quantity of sodium hydroxide to produce sodium zirconium silicate; the latter is subsequently treated with acid to form a soluble zirconium compound. These fusion processes are used worldwide to produce zirconium oxide, aqueous solutious of zirconium, and water-soluble zirconium compounds.

Fluidized-bed carbochlorination of zircon sand is the ore-decomposition process used by all three hafnium producers in the Western hemisphere: Teledyne Wah Chang Albany, CEZUS Division of Pechiney, and Western Zirconium Division of Westinghouse:

$$ZrSiO_4 + 4\,C + 4\,Cl_2 \xrightarrow{1100\,^\circ C} ZrCl_4 + SiCl_4 + 4\,CO$$

All the reaction products are gases when they leave the chlorinator. Zirconium and hafnium tetrachlorides are recovered as a powder by cooling the gas mixture to 200 °C in a large-volume space condenser. Silicon tetrachloride is then condensed by direct countercurrent contact of the remaining gas mixture with refrigerated (−40 °C) recycled liquid silicon tetrachloride.

Thermal dissociation of zircon in the high temperature of a plasma torch followed by rapid quenching is used to produce intimately mixed crystals of zirconium and silicon oxides. The oxides can be leached countercurrently with sodium hydroxide to yield zirconium oxide and a solution of sodium silicate, or leached with hot strong sulfuric acid to give a zirconium sulfate solution and silica. Both procedures are used by the Z-Tech Corporation Division of ICI.

5.2. Separation of Zirconium

Many well-known methods exist for separating hafnium from zirconium: *ion exchange* with a dilute sulfuric acid eluent [4]; *liquid – liquid extraction* of hafnium from hydrochloric acid solution by a methyl isobutyl ketone – thiocyanic acid solution [5–8]; *distillation* of the mixed binary tetrachlorides at a pressure above the triple point [9]; and *extractive distillation* of hafnium tetrachloride from the mixed tetrachlorides dissolved in a molten halide solvent [10, 11].

Ion exchange is the easiest laboratory procedure. A dilute solution of $Zr(Hf)OCl_2$ is introduced into a bed of strong cation-exchange resin and eluted with 0.25 M sulfuric acid. The eluate contains three components: (1) a small amount of unseparated metal, apparently polymerized, with minimum retention time; (2) a major fraction containing zirconium; and (3) a smaller fraction containing hafnium. A continuously fed rotating cylindrical bed system has been proposed for commercial separation of hafnium and zirconium [4].

Liquid – Liquid Extraction. The methyl isobutyl ketone (MIBK) liquid – liquid extraction is based on an analytical method for determining hafnium [12]. The process was developed at Oak Ridge National Laboratory to obtain hafnium-free zirconium to contain uranium fuel for nuclear-powered submarines. Recovery of hafnium as a byproduct was initiated when uses were found for the hafnium.

The zirconium – hafnium dichloride oxide solution is introduced at the midpoint of a series of countercurrent contact stages in which the organic solvent is MIBK containing thiocyanic acid and the countercurrent aqueous solution is dilute hydrochloric acid. Hafnium is extracted preferentially into the organic solution as a hafnium oxide – thiocyanate complex. The separation factor per stage is 4 – 5. After extrac-

tion, hafnium is recovered by treating the organic phase with dilute sulfuric acid, the water-soluble hafnium sulfate being more stable than the thiocyanate complex. The sulfuric acid is then neutralized to precipitate hydrous hafnium oxide, which is calcined to hafnium oxide. This procedure is used by both U.S. producers of hafnium.

Fractional Distillation. Separation by fractional distillation would be feasible if the tetrachlorides could be handled as liquids; however, the tetrachlorides are solid or gaseous unless kept under pressure while heated to the triple point of zirconium tetrachloride (435 °C, 2.0 MPa). The operating range of a binary distillation system is limited because the critical temperature for zirconium tetrachloride is ca. 505 °C.

Extractive distillation of the mixed tetrachlorides from molten potassium chloroaluminate at atmospheric pressure was developed and used commercially by CEZUS [10, 11]. The crude tetrachloride product from carbochlorination is first purified by sublimation in an atmosphere of nitrogen with 1 – 5 % hydrogen. The purified tetrachloride is then revaporized, and the vapor is introduced continuously above the midpoint of the distillation column. The potassium chloroaluminate solvent, equilibrated with hafnium-rich tetrachloride, is fed into the top of the distillation column, which is maintained at 350 °C. As the solvent cascades down the column it is gradually depleted of the more volatile hafnium tetrachloride, so that zirconium tetrachloride in the solvent reaching the bottom of the column contains <50 ppm of the hafnium compound. The separation factor per stage is 1.4.

The solvent is fed to a boiler where much of the zirconium tetrachloride is vaporized and passed back up the column to become enriched in hafnium. Solvent from the boiler is fed into a stripper where it is treated with nitrogen to remove the remaining zirconium tetrachloride.

The stripped solvent is pumped to an absorber – condenser above the column where the solvent is equilibrated with the hafnium-rich tetrachloride and returned to the top of the distillation column. Unabsorbed hafnium-rich tetrachloride vapor ($HfCl_4$ content 30 – 50 %) is collected in the hafnium condenser. This hafnium-enriched tetrachloride accumulates until it can be reprocessed in the distillation column under different conditions to produce pure hafnium tetrachloride (<1 % $ZrCl_4$).

The degree of separation attained by these various processes is limited only by the number of stages provided in the separation equipment. Zirconium containing <20 ppm hafnium and hafnium containing <50 ppm zirconium have been produced.

5.3. Winning of Metallic Hafnium

Hafnium dioxide from the separation process is converted first into hafnium tetrachloride by fluidized-bed carbochlorination:

$$HfO_2 + C + 2\,Cl_2 \xrightarrow{950\,^\circ C} HfCl_4 + CO_2$$

The recovered hafnium tetrachloride is purified by sublimation in a nitrogen – hydrogen atmosphere to reduce the levels of aluminum, iron, and uranium.

Kroll Process. Hafnium tetrachloride powder is placed in a vertical cylindrical steel retort welded onto a stainless steel-lined steel crucible containing cast ingots of magnesium. The two chambers are connected by a pipe rising through the center of the hafnium tetrachloride charge. The assembled retort is sealed, evacuated, and filled with argon several times to remove traces of air. The retort is lowered into a furnace; the lower section is then heated to 850 °C. Molten magnesium reduces the hafnium tetrachloride vapor as it sublimes out of the upper chamber. After the reaction is complete, the lower crucible contains hafnium metal, magnesium chloride, and excess magnesium. The retort is cooled and unloaded. The stainless steel liner is separated from the reduction mass, and the upper layer of magnesium chloride is physically removed from the hafnium – magnesium metallic regulus. The latter is loaded into a vacuum distillation furnace for removal of magnesium and residual magnesium chloride. As the temperature gradually increases, magnesium sublimes, leaving a skeletal structure of hafnium sponge. Magnesium chloride melts and drains into the cool lower portion of the retort. After the furnace is cooled and the sponge conditioned by slow, controlled exposure

to air, the sponge is removed and placed in an argon atmosphere, where it is broken and chopped into pieces finer than 20 mm.

Electrowinning. Electrowinning of hafnium and zirconium has long been investigated as an alternative to metallothermic reduction. CEZUS produces hafnium by electrolysis of hafnium tetrachloride dissolved in a molten equimolar bath of potassium and sodium chlorides [13]. Hafnium tetrachloride vapor is introduced periodically below the liquid level of the bath to replenish hafnium. The bath contains just enough sodium fluoride to convert dissolved hafnium into the hexafluorohafnate complex.

5.4. Refining

Applications of metallic hafnium require lower oxygen levels than commonly achieved by the Kroll process. Purification is accomplished by the *Van Arkel – de Boer* process, also known as the iodide-bar or crystal-bar process [14, 15]. Hafnium sponge is treated with iodine vapor to form volatile hafnium iodide, which diffuses to an electrically heated hafnium filament. Hafnium iodide dissociates at the filament temperature of 1400 °C, depositing hafnium on the filament and releasing iodine to repeat the diffusion cycle. The hafnium filament slowly grows from a 3-mm wire to a 40-mm crystal-faceted bar. Purification is effective in leaving behind interstitial impurities such as oxygen, carbon, and nitrogen. Metallic impurities generally transport poorly, depending on the behavior of individual iodides, so the hafnium filament is much purer than the sponge.

Electron-beam melting can also be used to purify hafnium. Hafnium is slowly drip-melted into a superheated molten pool under extremely high vacuum. The more volatile metallic impurities and suboxides are boiled off at a much faster rate, relative to the vaporization of hafnium. Two slow melts are standard for hafnium.

Typical impurity levels for Kroll-process hafnium sponge and electrowon hafnium crystals are shown in Table 2, together with the effectiveness of refining sponge by electron-beam melting or the Van Arkel – de Boer process.

6. Environmental Protection

Zircon contains traces of uranium and thorium in solid solution substitutionally and isomorphically in its structure. The retention of these elements and their radioactive decay daughter products are so complete that geological dating of rocks can be conducted by determining the radiological age of the zircon. Conversely, when the zircon structure is destroyed chemically, uranium, thorium, and their daughter products are released. If the attack on zircon is similar to carbochlorination, in which both zirconia and silica are converted to volatile species, the reaction residue may have sufficient concentrations of less volatile radioactive materials for special disposal procedures to be necessary.

Recovery of hafnia and zirconia from the acid solutions of liquid – liquid extraction results

Table 2. Typical impurity levels ∗ for hafnium metal

	Kroll process sponge	Electrowon crystals	Electron-beam melting ingot	Van Arkel–de Boer crystal bar
Oxygen	875	670	320	<50
Nitrogen	35	15	40	<5
Carbon	<30	40	<30	<30
Chlorine	100	50	<5	<5
Aluminum	200	10	<25	<25
Chromium	40	30	<20	<20
Iron	530	100	<50	<50
Magnesium	440	<10	<10	<10
Manganese	15	10	<10	<10
Nickel	<25	40	<25	<25
Silicon	25	<25	<25	<25
Titanium	<25	30	<25	<25

∗ In parts per million.

in process streams containing ammonium sulfate and ammonium chloride. Ammonia must be recovered for recycling or sale to avoid a heavy biological oxygen demand on effluent-receiving streams. Extreme care must be taken to ensure that hypochlorite scrubbing solutions from chlorination do not come in contact with effluent streams containing traces of thiocyanate because the mixture produces cyanides.

7. Quality Specifications

For wrought hafnium the following U.S. specifications apply: ASTM B 737-84 for Hot-Rolled or Cold-Finished Rod and Wire; ASTM B 776-87 for Hafnium and Hafnium Alloy Strip, Sheet, and Plate; ASTM G 2-88 and G 2 M-88 for Corrosion Testing of Products of Zirconium, Hafnium, and Their Alloys in Water at 680 °F (360 °C, 633 K) or in Steam at 750 °F (400 °C, 673 K); and ASTM C 1076-87 for Nuclear-Grade Hafnium Oxide Pellets.

In the Federal Republic of Germany, VdTüV Material Sheet 463, Hafnium, 05.83, covers wrought hafnium sheet.

Hafnium for alloying purposes is sold in chunk or particulate form, at the particular chemical purity specified by the customer; the zirconium content is particularly important.

8. Analysis

Hafnium is chemically so similar to zirconium that most gravimetric, volumetric, or colorimetric methods of analysis result in a combined hafnium – zirconium determination. Methods that do not require prior separation of hafnium include atomic absorption spectroscopy, emission spectroscopy, plasma emission spectroscopy, mass spectroscopy, X-ray spectroscopy, and neutron activation [16]. Ores and compounds are frequently analyzed by precipitating hafnium and zirconium with mandelic acid [17], igniting the precipitate to give the combined oxides, and determining the relative proportion of each oxide by X-ray or emission spectroscopy.

Many analytical methods for determining impurities in zirconium can also be applied to hafnium [16, 18]. Carbon and sulfur in hafnium are measured by combustion, followed by chromatographic or IR determination of the oxides [16]. Hydrogen is determined by thermal conductivity after hot-vacuum extraction or by fusion of hafnium with a transition metal in an inert atmosphere and subsequent separation by gas chromatography [19, 20].

Simultaneous determination of nitrogen and oxygen is accomplished by a technique similar to that for hydrogen [21, 22]. Nitrogen can also be determined by the Kjeldahl technique [16]. Phosphorus may be determined by visible spectrophotometry using molybdenum blue [16] but can be determined more rapidly by phosphine evolution and flame emission spectroscopy.

Chloride at low level is determined indirectly by precipitation with silver nitrate and determination of the silver by X-ray or atomic absorption spectroscopy. Fluoride and higher levels of chloride can be measured by selective ion electrode techniques [23, 24].

Other impurities are usually determined by emission spectroscopy, either arc – spark [25, 26] or plasma excitation [27, 28]. Many impurities can be determined rapidly and inexpensively. Atomic absorption is frequently used to determine a variety of impurities in hafnium metal and sponge [29].

9. Storage and Transportation

Hafnium metal powder is a flammable solid and should be kept dry, cool, and separate from potential oxidants. Dry hafnium powder, chemically produced, finer than 8.40×10^{-4} m (840 µm, 20 mesh), or mechanically produced, finer than 5.3×10^{-5} m (53 µm, 270 mesh), must have the following warning labels for shipping:

UN no. 2545
United States DOT: Flammable Solid
IATA Dangerous Goods Regulations: Combustible Solid IMDG Code: Combustible Solid

The same hafnium powder, wet, containing at least 25 wt % water, must be labeled as a flammable solid for all shipping and hazard communication purposes (UN no. 1326).

Hafnium powder of the above descriptions, wet or dry, is forbidden on U.S. passenger

aircraft by DOT. Dry hafnium metal powder, mechanically produced, finer than 3×10^{-6} m, or chemically produced, finer than 10×10^{-6} m, is forbidden from all air transportation by IATA.

Although not specifically required, hafnium hydride or hafnium carbide powder should be labeled, handled, and stored in similar fashion.

Hafnium tetrachloride must be shipped as a corrosive solid (UN no. 1759).

10. Uses

The major uses of hafnium involve the metal. The largest use is as an alloying additive (1–2 %) in nickel-based superalloys. These alloys are used in turbine vanes in the combustion zone of jet aircraft engines. Addition of hafnium to some present-generation alloys has raised the allowable service temperature by 50 °C, improving engine efficiency. Hafnium forms stable precipitates at grain boundaries, improving high-temperature creep strength.

The second major use of hafnium is as control-rod material in nuclear reactors. In early reactors, bare hafnium metal in long cruciform shape was used because of the excellent hot-water corrosion resistance, good ductility, and machinability of hafnium, as well as its high thermal neutron absorption cross section. Hafnium clad in stainless steel is now replacing stainless steel-clad silver–indium–cadmium and stainless steel-clad boron carbide in some commercial nuclear power plants.

Hafnium is also used as an alloying element in niobium, tantalum, titanium, molybdenum, and tungsten alloys, as well as some new nickel aluminides.

Hafnium and hafnium–zirconium alloys have been suggested for use in spent nuclear fuel storage racks, and for tanks and piping in spent fuel reprocessing plants because of the neutron absorption capability of hafnium and its resistance to nitric acid. For a while, shredded hafnium foil was used instead of zirconium foil as fuel in some photographic flashbulbs because hafnium provided a higher intrinsic color temperature and a greater light output. Pure hafnium has also been used as the active tip for plasma arc cutting tools.

Hafnium carbide–niobium carbide solid solutions have been used as a replacement for tantalum carbide in steel-cutting grades of cemented carbide tool bits.

Hafnium oxide has been used as a specialized refractory, including thermocouple insulation for short-term use above 2000 °C. Hafnium and hafnium oxide sputtering targets are used for coatings and specialized electronic applications.

Hafnium tetrachloride has been used to prepare hafnium metallocene Ziegler–Natta-type catalysts, which were the first catalysts to provide high yields of high molecular mass isostatic polypropylene [30]. Hafnium tetrafluoride is used in some heavy-metal fluoride glass cladding [31].

11. Economic Aspects

Production of hafnium oxide and hafnium tetrachloride has been steady since 1980, as a result of the leveling off of zirconium metal production for the nuclear energy program. The availability of hafnium still exceeds market demand. Annual Western hemisphere availability is estimated at 80 t, with consumption estimated at 50 t. Hafnium is produced commercially in the United States (Teledyne Wah Chang Albany and Western Zirconium Division of Westinghouse), France (CEZUS Division of Pechiney), and the Soviet Union.

In 1988, hafnium oxide was available at \$ 100/kg; hafnium crystal bar, at \$ 200/kg; and hafnium wrought metal products (plate, sheet, wire, foil), from \$ 250 to \$ 500/kg.

12. Hafnium Compounds

Compounds of hafnium have been studied mostly to develop methods of separating hafnium from zirconium, producing the metal, and comparing properties with similar zirconium compounds. Although the properties of inorganic hafnium and zirconium compounds are generally comparable, differences in chemical behavior become more evident in organometallic compounds.

12.1. Hafnium Borides

Hafnium boride [*12007-23-7*], HfB_2, M_r 200.11, is prepared by carbon reduction of

hafnium oxide and boron carbide, by coreduction of hafnium and boron chlorides with magnesium or hydrogen, or from the elements. It is of interest because it is very refractory. Hafnium dodecaboride, HfB_{12}, has been synthesized by compressing the elements at 1660 °C and 6.5×10^3 MPa.

12.2. Hafnium Tetrahydroborate

Hafnium tetrahydroborate [*25869-93-6*], $Hf(BH_4)_4$, M_r 237.86, is prepared by reaction of hafnium tetrachloride with lithium borohydride in diethyl ether, followed by double sublimation under vacuum at 20 °C [32]. afnium tetrahydroborate is one of the most volatile hafnium compounds known: *mp* 29 °C, *bp* 118 °C, vapor pressure 2 kPa at 25 °C. Its decomposition at ca. 250 °C forms the basis for a low-temperature method of depositing amorphous hafnium diboride films [33].

12.3. Hafnium Carbide

Hafnium carbide [*12069-85-1*], HfC, M_r 190.60, is usually produced by carbothermic reduction of hafnium oxide in an induction-heated graphite-lined vacuum furnace. Hafnium carbide film can be deposited by reaction of hafnium tetrachloride and methane in a hydrogen atmosphere at 900 – 1400 °C. Hafnium carbide is a nonstoichiometric interstitial carbide with carbon atoms at octahedral interstices of the hafnium metal lattice. The hafnium – carbon phase diagram shows that hafnium carbide is single phase within the range 39.5 – 49.5 atom % carbon at 1400 °C [34]. Hafnium carbide is a hard refractory material with metallic conductivity.

12.4. Hafnium Halides

Hafnium tetrafluoride [*13709-52-9*], HfF_4, M_r 254.48, can be made by addition of a stoichiometric amount of hydrofluoric acid to hafnium oxide chloride dissolved in 8 M nitric acid, to precipitate hafnium tetrafluoride monohydrate. The filtered crystals are dried in a stream of anhydrous hydrogen fluoride gas. Alternatively, anhydrous hydrogen fluoride can undergo an exchange reaction with hafnium tetrachloride powder or vapor, or ammonium fluorohafnate can be heated to drive off ammonium fluoride, leaving hafnium tetrafluoride as residue.

Potassium hexafluorohafnate [*16871-86-6*], K_2HfF_6, M_r 370.68, can be crystallized from hafnium fluoride solution by adding a stoichiometric amount of potassium fluoride. The solution should be acidic to minimize hydrolysis. Other alkali metal hexafluorohafnates (Na_2HfF_6 [*16921-79-2*], M_r 338.46; Rb_2HfF_6 [*16962-20-2*], M_r 463.42; and Cs_2HfF_6 [*16919-32-7*], M_r 558.29) and the corresponding ammonium salts can be prepared similarly. With different proportions of alkali fluoride, other salts can be precipitated, including $MHfF_5$, M_3HfF_7, and $M_5Hf_2F_{13}$.

Hafnium tetrachloride [*13499-05-3*], $HfCl_4$, M_r 320.30, is prepared commercially by carbochlorination of hafnium oxide at 900 °C. Chlorination can be carried out with phosgene or carbon tetrachloride at 450 °C. Hafnium tetrachloride is used as starting material for the preparation of organic derivatives of hafnium, alkoxides, and alkali chlorohafnates. The partitioning of mixed tetrachloride vapor by alkali-metal salts, with preferential formation of alkali chlorohafnate and elimination of unreacted zirconium tetrachloride, has been proposed as a means of separating hafnium and zirconium [35, 36]. Hafnium tetrachloride vapor reacts with steam to form finely divided hafnium oxide. Solid hafnium tetrachloride reacts with water to form hafnium oxide chloride and hydrochloric acid. Hafnium dichloride oxide, $HfOCl_2$, is soluble in water and can be crystallized from hydrochloric acid solution as the octahydrate, $HfOCl_2 \cdot 8H_2O$, [*14456-34-9*], M_r 297.52. The solubility of hafnium dichloride oxide increases significantly with temperature and decreases with increasing acidity up to 8.5 M hydrochloric acid. This crystallization procedure is the preferred method for obtaining pure hafnium dichloride oxide prior to the preparation of other hydrated hafnium compounds. Hafnium dichloride oxide is converted to the hydrous oxide $HfO_2 \cdot xH_2O$ on neutralization, at pH 2. On heating, hafnium dichloride oxide first loses water, then slowly releases water and hydrogen chloride, and is converted into granular hafnium oxide.

Hafnium tetrabromide [*13777-22-5*], $HfBr_4$, M_r 498.11 and hafnium tetraiodide [*13777-23-6*], HfI_4, M_r 686.11 are produced by reaction of the respective halogen with hafnium metal in the absence of air above 300 °C. Thermal dissociation of the tetraiodide is employed in the iodide-bar refining process.

Lower valence hafnium halides have been formed by reduction of tetrahalides with aluminum or hafnium in molten aluminum chloride or molten alkali halide baths. Zirconium is more easily reduced; this is the basis of several proposals for the separation of hafnium and zirconium [37,38].

12.5. Hafnium Hydride

Hafnium hydride, HfH_x, is formed as hafnium absorbs hydrogen. The proportion of hydrogen absorbed depends on temperature and hydrogen pressure. Above 500 °C, as hydrogen is absorbed, hafnium changes from hexagonal close-packed metal, to face-centered cubic hydride, and then to face-centered tetragonal hydride as the composition approaches the limiting solubility, with $x = 2$ [39]. At room temperature, hafnium hydride exists as the face-centered cubic form when $x = 1.7 - 1.8$, and as the face-centered tetragonal form when $x \geq 1.87$.

Hafnium hydride is brittle and easily crushed. Crushing must be conducted in an argon atmosphere to avoid ignition of the fine powder. Hydrogen absorption is reversible; the process of hydriding, crushing, and dehydriding is therefore used to convert hafnium metal pieces into powder with little contamination. Hydrogen is removed by heating the hydride under vacuum.

12.6. Hafnium Nitride

Hafnium nitride [*25817-87-2*], HfN, M_r 192.5, is a nonstoichiometric interstitial nitride, which is single phase within the range 42 – 50 atom % nitrogen [40]. Hafnium nitride film can be deposited by reacting hafnium tetrachloride vapor with ammonia or nitrogen in a hydrogen atmosphere above 1000 °C. Above 1400 °C, nitride powder is produced. This film deposition process is used to coat cemented carbide tool bits. Hafnium nitride film is more effective than titanium nitride in reducing frictional forces and wear when machining steel [41].

12.7. Hafnium Dioxide

Hafnium dioxide [*12055-23-1*], hafnia, HfO_2, M_r 210.49, is the only stable oxide of hafnium. It melts at 2900 °C and exists in three solid phases: a monoclinic phase stable up to 1475 – 1600 °C, above which hafnia changes into the tetragonal phase, the stable form up to ca. 2700 °C. Between 2700 and 2900 °C, hafnia has a cubic fluorite structure. Like zirconia, the tetragonal – monoclinic transition is a martensitic transformation that exhibits considerable hysteresis. The cubic structure can be stabilized by addition of calcium oxide, yttrium oxide, or some of the rare-earth oxides, including those of cerium and europium. The use of stabilized hafnia for high-temperature engineered ceramics avoids the catastrophic 7 % shrinkage that accompanies transformation from monoclinic to tetragonal form.

Hafnium dioxide is very stable and, except for strong hydrofluoric acid, is attacked only at elevated temperature. Hafnia reacts with carbon tetrachloride or phosgene above 400 °C, and with chlorine and carbon above 700 °C, to form hafnium tetrachloride. At high temperature, hafnia reacts with many metal oxides to form solid solution oxides or hafnate compounds such as $CaHfO_3$.

12.8. Other Derivatives

Several excellent publications that cover the preparation and properties of hafnium alkoxides, hydrides, and organic derivatives are available [42–44].

13. Toxicology and Occupational Health

Hafnium metal with a high surface area, such as thin machining chips, powder, sponge, or grinding dust, is extremely flammable, even pyrophoric. If the metal is slightly wet (5 – 15 wt %

H_2O) ignition can occur with explosive violence. Fires can be extinguished with a layer of dry salt or with a blanket of argon if the metal is in a container.

Water can be hazardous for fire fighting because it is a source of oxygen for the burning metal. The resulting hydrogen and steam may blow the burning chips throughout the area. Several reviews of methods for safe handling of hafnium and zirconium are available [45,46].

Hafnium is essentially nontoxic. No health problems have been attributed to hafnium in more than 30 years of industrial experience with the metal and its compounds. Toxicological studies of ingestion indicate that the toxicities of hafnium and zirconium are similar. Hafnium was found to accumulate in the liver when hafnium compounds were injected intravenously or intraperitoneally into rats [47]. Chronic feeding of 1 % hafnium tetrachloride to rats for 90 d caused liver damage [48]. The ACGIH TLV–TWA for hafnium is 0.5 mg/m^3.

Hafnium tetrachloride and tetrabromide hydrolyze immediately on contact with water, releasing hydrochloric or hydrobromic acid fumes. These compounds should be handled with proper ventilation and personal protection.

14. References

1. A. Clearfield, *Rev. Pure Appl. Chem.* **14** (1964) 91.
2. R. Clark, D. Bradley, P. Thornton: *The Chemistry of Titanium, Zirconium, and Hafnium,* Pergamon Press, Oxford 1973.
3. H. S. Choi, *Can. Min. Metall. Bull.* **67** (1965) 65.
4. J. M. Begovich, W. G. Sisson, *Hydrometallurgy* **10** (1983) 11 – 20.
5. W. A. Stickney: "Zirconium – Hafnium Separation," USBM RI 5499, 1959.
6. J. M. Googin in F. R. Bruce, J. M. Fletcher, H. H. Hyman (eds.): *Progress in Nuclear Energy,* Series III: Process Chemistry, vol. 2. Pergamon Press, New York 1958.
7. J. H. McClain, S. M. Shelton in C. R. Tipton, Jr., (ed.): *Reactor Handbook,* vol. 1, Materials, 2nd ed., Interscience Publishers, New York 1960.
8. W. Fischer et al., *Angew. Chem. Int. Ed. Engl.* **5** (1966) 15 – 23.
9. E. I. DuPont de Nemours, US 2 852 446, 1956 (M. L. Bromberg).
10. Pechiney Ugine Kuhlman, US 4 021 531, 1976 (P. Besson, J. Guerin, P. Brun, M. Bakes).
11. L. Moulin, P. Thouvenin, P. Brun in D. G. Franklin (ed.): *Zirconium in the Nuclear Industry,* 6th Int. Symposium, ASTM STP 824, American Society for Testing and Materials, Philadelphia 1984, pp. 37 – 44.
12. W. Fischer, W. Chalybaeus, *Z. Anorg. Allg. Chem.* **255** (1948) 277.
13. Pechiney, US 4 657 643, 1986 (M. Armand, J.-P. Garnier).
14. E. M. Sherwood, I. E. Campbell in D. E. Thomas, E. T. Hayes (eds.): *The Metallurgy of Hafnium,* US Government Printing Office, Washington, DC 1960, 108 – 118.
15. R. F. Rolston: *Iodide Metals and Metal Iodides,* The Electrochemistry Society, New York 1961.
16. R. VanSanten, J. Schlewitz, G. Beck, R. Walsh in F. D. Snell, L. S. Ettre (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 14,** Wiley-Interscience, New York 1971, pp. 103 – 148.
17. R. B. Hahn, E. S. Baginski, *Anal. Chim. Acta* **14** (1956) 45 – 47.
18. R. B. Hahn in I. M. Kolthoff, P. J. Elving, E. B. Sandell (eds.): *Treatise on Analytical Chemistry,* Part II, vol. 5, Interscience, New York 1961, pp. 61 – 138.
19. R. K. McGeary in *A Symposium on Zirconium and Zirconium Alloys,* American Society for Metals, Cleveland 1953, pp. 168 – 175.
20. W. G. Guldner, *Talanta* **8** (1961) 191 – 202.
21. W. G. Smiley, *Anal. Chem.* **27** (1955) 1098 – 1102.
22. P. Elbling, G. W. Goward, *Anal. Chem.* **32** (1960) 1610 – 1613.
23. J. Surak, D. Fisher, C. Burros, L. Bate, *Anal. Chem.* **32** (1960) 117 – 119.
24. J. Lingane, *Anal. Chem.* **39** (1967) 881.
25. L. Carpenter, J. M. Nishi: "Analysis of High-Purity Columbium by Optical Emission Spectrography," USBM RI 6384, 1964.
26. R. Brayer, R. O'Connell, A. Powell, R. H. Gale, *Appl. Spectrosc.* **15** (1961) 10–13.
27. G. L. Beck, O. T. Farmer, *J. Anal. Atomic Spectrometry* **3** (1988) 771 – 773.
28. G. L. Beck, J. P. Fraley. *Teledyne Wah Chang procedure ASP-OES-1,* Rev. O, 1988.
29. J. Schlewitz, M. Shields, *At. Absorpt. Newsl.* **10** (1971) 39 – 43.
30. J. A. Ewen, L. Haspeslagh, J. L. Atwood, H. Zhang, *J. Am. Chem. Soc.* **109** (1987) 6544 – 6545.

31. D. C. Tran et al. in J. Lucas, C. T. Moynihan (eds.): *Halide Glasses I, Material Science Forum,* **vol. 5,** Trans Tech Pub., Switzerland 1985, 339 – 352.
32. B. D. James, B. E. Smith, *Synth. React. Inorg. Met-Org. Chem.* **4** (1974) 461 – 465.
33. J. A. Jansen, J. E. Gozum, D. M. Pollina, G. S. Girolami, *J. Am. Chem. Soc.* **110** (1988) 1643 – 1644.
34. E. Rudy: "Ternary Phase Equilibria in Transition Metal – Boron – Carbon – Silicon Systems," Part V, *Compendium of Phase Diagram Data,* AFMLTR-65-2, Air Force Materials Laboratory, Wright-Patterson AFB, Ohio 1969, pp. 166 – 167.
35. S. Mazumdar, H. S. Ray, *J. Appl. Chem. Biotechnol.* **4** (1972) 565 – 576.
36. S. N. Flengas, J. E. Dutrizac, *Metall. Trans. B* **8 B** (1977) 377 – 385.
37. E. M. Larsen, J. W. Moyer, F. Gil-Arnao, M. J. Camp, *Inorg. Chem* . 13 (1974) 574 – 581.
38. Teledyne Industries, US 4 072 506, 1975 (J. Megy).
39. W. M. Mueller, J. P. Blackledge, G. G. Libowitz: *Metal Hydrides,* Academic Press, New York 1968, pp. 321 – 330.
40. [34] p. 671.
41. E. Rudy, B. F. Kieffer, E. Baroch, *Planseeber. Pulvermetall.* **26** (1978) 105 – 115.
42. D. C. Bradley, R. C. Mehrotra, D. P. Gaur: *Metal Alkoxides,* Academic Press, New York 1978.
43. D. J. Cardin, M. F. Lappert, C. L. Raston: *Chemistry of Organo-Zirconium and -Hafnium Compounds,* Ellis Horwood Ltd., Chichester 1986.
44. P. C. Wailes, R. S. P. Coutts, H. Weigold: *Organometallic Chemistry of Titanium, Zirconium, and Hafnium,* Academic Press, New York 1974.
45. J. Schemel: *ASTM Manual on Zirconium and Hafnium,* ASTM STP 639, Philadelphia 1977.
46. *National Fire Codes,* **vol. 7,** NFPA No. 482 – 1987, National Fire Protection Association, Boston 1988.
47. C. F. Kittle, E. R. King, C. T. Bahner, M. Bruler, *Proc. Soc. Exp. Biol. Med.* **76** (1951) 278 – 282.
48. T. J. Haley, R. Raymond, N. Komesu, H. C. Upham, *Toxicol. Appl. Pharmacol.* **4** (1962) 238 – 246.

Hahnium → Transuranium Elements

Hair Preparations

Thomas Clausen, Wella Service GmbH, Darmstadt, Germany (Chap. 2, 3, 4, 5, 6, 8 and 10)
Annette Schwan-Jonczyk, Wella Service GmbH, Darmstadt, Germany (Chap. 1, 2 and 9)
Günther Lang, Wella Service GmbH, Darmstadt, Germany (Chap. 3, 4, 6, 7 and 8)
Werner Schuh, Wella Service GmbH, Darmstadt, Germany (Chap. 10.1)
Klaus Dieter Liebscher, Wella Service GmbH, Darmstadt, Germany (Chap. 10.2)
Christian Springob, Wella Service GmbH, Darmstadt, Germany
Michael Franzke, Wella Service GmbH, Darmstadt, Germany
Wolfgang Balzer, Wella Service GmbH, Darmstadt, Germany
Sonja Imhoff, Wella Service GmbH, Darmstadt, Germany (Chap. 6)
Gerhard Maresch, Wella Service GmbH, Darmstadt, Germany (Chap. 6)
Rudolf Bimczok, Wella Service GmbH, Darmstadt, Germany

1. Introduction 2
2. The Scalp and Hair 2
2.1. Structure of the Skin and Hair Growth 2
2.2. Structure of Hair 3
2.2.1. Morphology 3
2.2.2. Chemical and Physical Properties 4
2.2.3. Natural Hair Color 5
3. Hair Cleansing and Care Preparations 6
3.1. Shampoos 6
3.1.1. Composition 6
3.1.1.1. Surfactants 6
3.1.1.2. Other Constituents 8
3.1.2. Types of Shampoo 8
3.1.3. Dry Shampoos 9
3.2. Conditioning Agents and Treatments 10
4. Hairstyling Preparations 11
4.1. Setting Lotions, Mousses, and Restructurants 11
4.2. Hair Sprays 13
4.3. Other Hairstyling Preparations 14
4.3.1. Gels 14
4.3.2. Waxes 15
4.3.3. Emulsion-Based Styling Products 15
4.3.3.1. Creams and Pomades 15
4.3.3.2. Gums/Puttys 15
4.3.3.3. Serums 16
5. Hair Coloring Preparations . . . 16
5.1. Bleaching 16
5.2. Dyeing 17
5.2.1. Dyeing with Dye Precursors 17
5.2.2. Dyeing with Inorganic Compounds 20
5.2.3. Dyeing with Direct Dyes 20
5.2.4. Product Forms 21
5.2.5. Testing Hair Dyes 22
5.2.6. Dye-Removal Preparations 23
6. Waving 23
6.1. Principles and Methods 23
6.2. Permanent-Wave Preparations . 24
6.2.1. Waving Preparations 24
6.2.2. Neutralizers 26
6.2.3. Notes on Use 26
7. Hair Straightening Preparations 26
8. Hair Removal 27
8.1. Epilation and Permanent Hair Removal 27
8.2. Depilation 27
8.3. Shaving Aids 27
9. Testing Methods 28
10. Toxicology and Legal Aspects . . 29
10.1. Toxicology 29
10.1.1. Introduction 29
10.1.2. Strategy for Safety Assessment of Cosmetics 29
10.1.3. Test Methods 30
10.1.3.1. Basic Principles 30
10.1.3.2. Systemic Toxicity Studies 31
10.1.3.3. Local Compatibility 32
10.1.3.4. Sensitizing (Contact-Allergenic) Potential 32
10.1.3.5. Mutagenic/Genotoxic Potential . . 33
10.1.3.6. Reproductive Toxicology 34

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_571.pub2

10.1.3.7. Tumorigenic Potential 34
10.1.3.8. Toxicokinetics 35
10.1.4. Description of the Safety Assessment Process 35
10.2. Legal Aspects 37
10.2.1. General 37
10.2.1.1. Legal Aspects in the European Union 37
10.2.1.2. Legal Aspects in Other Countries 37
10.2.2. Individual Hair Preparations 38
11. References 39

1. Introduction

The term *hair preparations* covers all preparations used on the scalp or scalp hair. The most important of these are cleansing and conditioning products, colorants, hairstyling preparations (setting lotions, hair sprays, hairdressing aids), and permanent-wave preparations. Recent developments in this field should be viewed from several standpoints. Fashion trends have spurred the development of new products such as high-hold hair lacquers and styling gels, as well as dyes with brilliant fashion colors. Toxicological considerations have also become increasingly important.

Economic Aspects. The total worldwide market for personal care and cosmetics in 1996 has been estimated at 114×10^9 €; hair-care products accounted for 29×10^9 €. The geographic structure of the retail hair-care market is given in Table 1. In the hair-care sector, the most important products were care (55 %), styling (23 %), color (18 %), and others, including hair tonics and home perms (4 %). In recent years hair color products showed the strongest growth on a worldwide basis, while styling products grew only moderately.

Table 1. The geographic structure of the retail hair-care market

Europe	38 %
North America	21 %
Latin America	12 %
Africa/Asia/Pacific	29 %

2. The Scalp and Hair

2.1. Structure of the Skin and Hair Growth

Hair is an appendage of the skin that grows in varying density everywhere on the body. In contrast to the fine, barely visible hair (vellus) of the body, the approximately 100 000 – 150 000 scalp and beard hairs are 40 – 100 µm thick and grow roughly 1/3 mm per day [1]. The structure of the skin and a hair follicle are shown in Figure 1. The hair follicle may be anchored up to 3 mm deep in the scalp and has a high rate of cell division. Scalp hair is one of the fastest growing body tissues, and reflects the life history of its owner. Hair analysis (trichogram) is therefore used as a diagnostic tool in medicine and also in forensic investigations [2, 3].

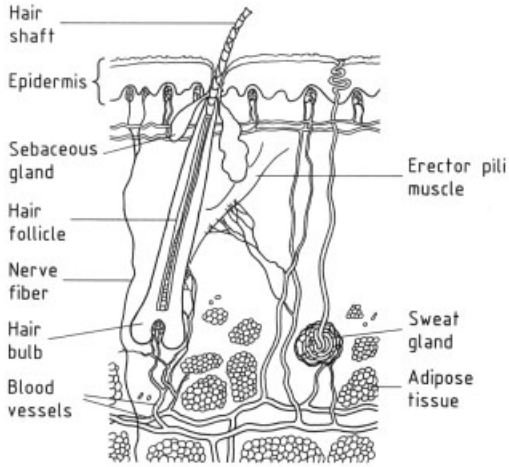

Figure 1. Vertical section through the skin showing a hair follicle

The growth cycle of human hair has three main phases. In the growth (anagen) phase, lasting two to six years, new hair keratin is formed continuously. The death of some cells marks the beginning of a roughly two-week transitional (catagen) phase, followed by a resting (telogen) phase of three to six months, after which the hair falls out. This accounts for a daily loss of some 100 hairs.

Because hairs are linked to the circulatory and nervous systems through their roots, their

growth can be influenced by hormones, nutrition, and drugs [5]. Balding mostly is a clear indication of a hormonally induced disturbance of hair growth. Hair "cures" are often claimed to promote hair growth, but only minoxidil, finasterid, and antiandrogens have been shown to have a temporary hair-restoring effect after topical application [6].

The scalp contains sweat and sebaceous glands, which are located next to the hair roots in the dermis. The secretions of these glands spread over the scalp surface and onto the hair, where they produce the impression of "oily" or "greasy" hair. The cosmetic problem of oily hair probably results not from the overproduction of sebum by the sebaceous glands, but from a change in its rheological properties [7]. The causes of dandruff (pityriasis simplex capillitii) are also unclear, but this condition is often attributed to a disturbance of scalp microflora [8]. For this reason, antidandruff preparations often contain antimicrobial substances such as selenium disulfide or zinc pyrithione.

2.2. Structure of Hair

2.2.1. Morphology

Cells constitute the principal components of a hair (Fig. 2); they undergo two changes in the hair follicle on their way from the root to the scalp surface. Firstly the cell proteins differentiate by chemically changing from an "uncross-linked" state into a conformation cross-linked by the amino acid cystine. Typically, fibrillar keratin proteins are expressed with a partially α-helical structure. These microfibrils, also called intermediate filaments (IF) [10] reside in amorphous matrix proteins (intermediate filament associated proteins, IFAP). Secondly all the cells change their shape. After differentiation into several cell types, the spherical hair-forming cells at the root periphery (*keratinocytes*) become thin, flat *cuticle cells*, whereas the central cells form lines of elongated, spindle-shaped *cortex cells*. In grown-out hair, the cuticle cells are arranged around the cortex cells in the manner of roof tiles. Cortex cells form the bulk of the hair (ca. 80 %). They have a fibrillar structure and are oriented predominantly parallel to the hair axis. The cells occur in two different morphological forms — ortho- and paracortical cells — which differ in their cystine content and solubility [11]. The natural curliness of a hair is discussed to be determined either by the arrangement of these cells or due to the bent shape of the hair follicle. A concentric arrangement of cells was found in straight hair, whereas segmental arrangement gives rise to curly hair [12]. Thick Caucasoid, Mongoloid, and Negroid hair has a central row of vacuole-containing cells termed the *medulla.*

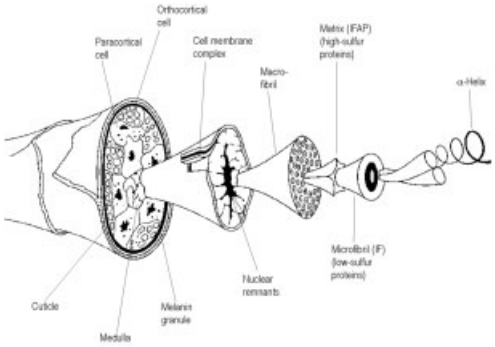

Figure 2. Structure of a hair

All the hair cells are joined by an *intercellular cement.* This cement consists of a continuous phase of proteins and lipoproteins [13] which, together with adjacent membrane lipids, forms the *cell membrane complex.* The cement provides cohesion and gives the hair the appearance of a strong fiber [11, 14]. The composite structure of the hair [15] is continued within the cortex cells by the macrofibrils, which are an association of microfibrils. The latter form a crystalline phase embedded in an amorphous, highly cross-linked protein matrix. The physical behavior of this natural fiber (see also → Wool) can be described in similar terms to other fibrous polymers or composites [16].

The structural components of the hair dictate its cosmetically important properties:

- Luster: cuticle, medulla, melanin, shape
- Color: melanin
- Combability: cuticle
- Feel: cuticle surface, cortex
- Natural wave: cortex, ratio and arrangement of ortho and para cells

- Deformability (temporary set, permanent wave): cortex, microfibrils
- Elasticity, strength: cortex
- Friction: cuticle scales, hair surface geometry
- Hygroscopicity, swelling: cortex
- Electrostatic charging: cuticle surface
- Volume, body: hair thickness and geometry, cortex curliness

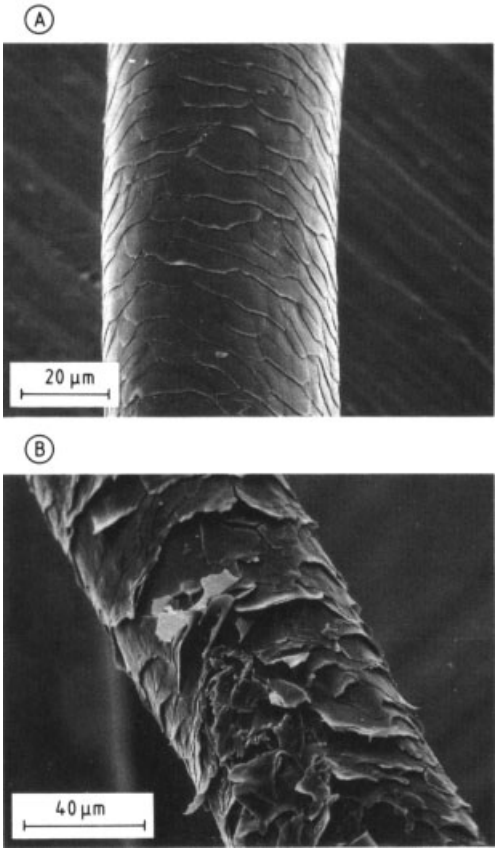

Figure 3. Scanning electron micrographs of a normal, smooth hair (A) and a damaged hair (B) with detached cuticle scales

As hair grows (i.e., ages), it experiences increasing wear. It is subject to "weathering" from light, abrasion during washing and brushing (especially when wet), and heating due to the sun or to hair dryers. Weathering results in damage to hair structure in the form of detached cuticle scales, greater surface friction, increased porosity, and decreased elasticity. The cohesiveness of the cortex cells can be weakened so much that the hair end splits [17]. Thus, the hair structure varies from the root to the tip; this fact must be considered in the use of hair preparations. A normal hair is shown in Figure 3 A and a damaged hair in Figure 3 B.

2.2.2. Chemical and Physical Properties

Hair consists of *keratin*; this term comes from the Greek word for horn and was originally used to denote a cornified tissue such as hair, nails, hooves, horn, feathers, or cornea. However, keratin has become more and more difficult to define because the term is used both to denote a structural component of hair (e.g., α-keratinic microfibrils) and to characterize the sulfur-containing cross-linked matrix proteins [18]. All structures composed of keratin are, however, distinguished by insolubility in water and common organic solvents.

Hairs consist of a mixture of proteins, some of which have a characteristic amino acid composition and can be classified into protein families. The protein content of a normal hair at 20 °C and 65 % R.H. is ca. 80 wt %. Further components are ca. 5 wt % internal hair lipid, <1 wt % trace elements and metals, and ca. 14 wt % water [19].

Of the 22 amino acids found in hair protein, the most important is L-cystine [*56-89-3*], a sulfur-containing amino acid (ca. 750 µmol per gram of hair). This compound primarily contributes to the cross-linking and thus the mechanical strength of the fibers.

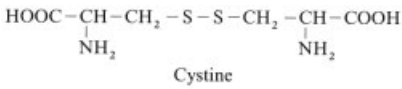

Hair proteins may be classified as high- or low-sulfur; their characteristics are listed in Table 2 [11, 20], see however [21].

The distribution of individual proteins in hair is not homogeneous. High-sulfur proteins are concentrated in the outer layer of cuticle cells, the exocuticle. These cells thus display greater insolubility and chemical resistance than the cortex; they afford protection against ultraviolet radiation [22] and can also serve as a diffusion barrier [19, p. 231]. In cortex cells, the low-sulfur microfibril proteins (= intermediate filaments) are distributed in a matrix of high-sulfur proteins

Table 2. Characteristics of hair proteins

Property	High-sulfur proteins	Low-sulfur proteins	
M_r	10 000 – 30 000	40 000 – 58 000	
Fraction of total protein	ca. 45 %	ca. 55 %	
State	amorphous	amorphous	α-helical, ca. 50 % crystalline microfibrils of cortex
Location	matrix and outer cuticle	nuclear remnants of cuticle and cortex; cell membrane complex; intermacro-fibrillar material	

(keratin associated proteins), giving a composite structure similar to that of a fiber-reinforced plastic.

More than half of the amino acids in hair have hydrophilic or ionizable side groups. Interactions between these groups lead to the formation of hydrogen bonds and salt bridges, which contribute to the stability of the hair fiber.

The variety of reactive acidic and basic amino acid substituents allows chemical reactions similar to those used in the treatment of wool and animal hair [23]. The number of raw materials available for hair preparations is, however, limited by the following requirements: they must not cause skin reactions; they must be effective below 60 °C and within 30 min; and their reactivity should be so well controlled that irreversible structural damage and hair dissolution are avoided. Hair can be regarded as an ion exchanger (e.g., for components of shampoos, rinses, and conditioners [19, p. 249]) or as a biopolymer whose cross-linking can be partially cleaved by controlled reduction of cystine and then restored. This principle forms the basis of permanent waving — fortunately, very little reaction takes place at the skin surface because skin epidermal keratins contain relatively little cystine.

Additional cross-linking reactions on the hair help to stabilize its structure. Cross-linking agents such as formaldehyde are ruled out for toxicological reasons[23].

Though its fraction is rather low, hair lipids gain more and more attention. Especially an 18-methyleicosanoic acid is analyzed as being responsible for the hydrophobic nature of the hair surface [9, 19].

Of particular importance are the interactions of hair with water [19]. Depending on the atmospheric humidity, hair can absorb up to 30 % of its own weight of water. In an aqueous medium, however, it can take up another 100 %, bound primarily by capillary forces [24, p. 96]. The water content of hair dictates its chemical reactivity as well as most of its physical (especially mechanical) properties. The diameter of hair swells by 16 %, but its length increases by only 1.2 %. The longitudinal mechanical stability, expressed as the modulus of elasticity, decreases by a factor of 2.7, whereas the extensibility increases by roughly one-third. The torsional and flexural strengths in the wet state are low, so that the fiber on the whole is softened. This softening allows hair to be shaped and set with water or by permanent waving; however, it provides poor hold in humid air after a temporary set.

Dry hair holds its set best when it is dried under the atmospheric conditions in which it relaxes after setting. Drying then brings the hair into thermal and moisture equilibrium with its surroundings ([19, p. 133]). Above 150 °C (which is seldom reached in practice), hair becomes overdried; its sorption capacity is lowered, and combing causes increased static charge.

The friction of the hair surface is also influenced strongly by its moisture content. Wet hair has a higher friction coefficient than dry hair. Because of the correlation between hair's surface friction and its combability [19, p. 437ff.], wet hair is more difficult to comb than dry hair.

2.2.3. Natural Hair Color

In addition to keratinocytes (hair-forming cells), the follicle contains a second type of cell, the *melanocyte,* which produces tiny grains of pig-

ment. Initially, these granules (melanosomes) consist primarily of a protein called *melanoprotein.* As the granules mature, the protein binds increasing amounts of the pigment *melanin.* The pigment is formed from the phenolic amino acid tyrosine. Tyrosine is converted to a polymeric structure with indole quinone subunits by enzymatic catalysis (tyrosinases) via several intermediates (3,4-dihydroxyphenylalanine, DOPA). A variety of mesomeric pigment structures with high optical absorption appear to be formed. Melanin granules are transferred from the dendrites (armlike outgrowths) of the melanocytes to the adjacent keratinocytes, probably by phagocytosis. The granules are then transported into the mature hair inside the keratinocytes. Here they are stored in the colorless protein mass of the cortex cells. Melanin granules give hair its natural color.

Nature provides an abundance of hair colors and shades, which result not just from the number, size, and distribution pattern of melanin granules in the hair, but also from the type of melanin [12, 25]. At least two distinct species occur: — brown — black *eumelanin* and yellow – red *phaeomelanin.* Eumelanin is found chiefly in Mongoloid, Negroid, and dark Caucasoid hair; red and blond Caucasoid hair contains chiefly phaeomelanin, with only small amounts of eumelanin.

Phaeomelanin, whose structure is largely unknown, is distinguished from eumelanin [26] by its synthesis via cysteine, its disputed high sulfur and iron contents [26], low molecular mass, size, appearance, and solubility. A polymeric, graphitelike structure has been postulated for eumelanin [27], the characteristics of which could best be determined after enzymatic isolation from hair[28]. The exact correlation of melanin types with naturally occurring hair colors and shades is, however, not understood completely [29]; other "trichochromes" have been proposed as constituents of the melanin in red hair. Attempts have been made to color hair with melanin precursors [30].

Gray hair, a mixture of white and pigmented hair, has a low content of melanin pigments. The reasons for pigment loss in old age include lowered melanin production or a decrease in the number of melanocytes [24, p. 158].

3. Hair Cleansing and Care Preparations

3.1. Shampoos

Shampoos are cosmetic products for cleaning hair and scalp. The word "shampoo" is derived from a Hindi word meaning massage.

Dirt adhering to the hair may consist of unpleasantly smelling skin flakes, sebum, perspiration residues, and dust, but residues of hair sprays, setting lotions, and conditioning agents may also be present. A shampoo should leave hair pliable and lustrous, with a pleasant odor. The hair should not "fly away" and should be clean but not too strongly defatted, easy to untangle and comb, but at the same time easy to manage and style. Decisive factors for the development of modern shampoos are *grooming* (combability and conditioning of the hair) and *mildness* (compatibility with skin, eyes, and mucosa; no burning on the scalp or in the eyes).

Shampoos are the cosmetic products most often applied to the hair. Nearly all modern shampoos are aqueous surfactant preparations. Dry shampoos are of secondary importance. In the developed Western economies, the trend toward daily hair washing is increasing. Accordingly, shampoos are among the most important products in the area of personal grooming (see Table 1).

3.1.1. Composition

3.1.1.1. Surfactants

The principal constituents and most important raw materials of shampoos are surfactants (→ Laundry Detergents; → Surfactants). Their task is to break the physical bonds between dirt and substrate, and to transport dirt into the aqueous medium. As in other fields, the selection of a surfactant or combination of surfactants is dictated by the consumer's association of lathering ability with cleansing power. Surfactants for shampoo must, therefore, satisfy the following requirements: they must have high cleansing power; they must lather well, even in the presence of dirt, sebum, and sweat; and they must be compatible with body tissues as well as environmentally harmless (biodegradable).

These standards are best met by *anionic surfactants. Amphoteric and nonionic surfactants* are also employed in modern shampoos. *Cationic surfactants* play only a secondary role in special-purpose formulations. For comprehensive reviews of surfactants employed in shampoos, see [31 – 33].

The first surfactants used for hair cleansing were *fat soaps.* In hard water, however, these products create a deposit of lime soap on the hair, so their use must be followed by treatment with an organic acid, such as citric or acetic acid. Another drawback is the alkaline pH of soap, which may increase the swelling of hair.

Anionic Surfactants. The most important anionic surfactants for shampoos are alkyl sulfates, alkyl and alkylaryl sulfonates, olefinsulfonates, secondary alkyl sulfonates, alkylaryl sulfates, alkyl ether sulfates, sulfosuccinates, and protein – fatty acid condensates.

Alkyl sulfates are far superior to soaps because they have much lower hardness sensitivity. *Lauryl sulfates* are employed most often, they form a good, stable foam but have a strong defatting action on hair and skin.

The skin compatibility of the alkali salts (sodium lauryl sulfate [*151-21-3*]) is not optimal; for this reason and because of solubility, the ammonium salt [*2235-54-3*], the alkanolammonium salts, or mixtures of lauryl sulfates with gentler surfactants, such as sulfosuccinates, protein – fatty acid condensates, or amphoteric surfactants, are used.

Alkyl and alkylaryl sulfonates are the most important detergents used in household cleansers worldwide. They have a very strong defatting action, and their skin tolerance is no better than that of the alkyl sulfates. The latter can, however, be manufactured in higher purity.

Olefinsulfonates and secondary alkyl sulfonates have not yet found much use. With respect to skin compatibility, they lie between the alkyl sulfates and the alkyl ether sulfates. α-Olefinsulfonates have good foaming properties, even in hard water.

Alkylaryl sulfates are no longer important primarily because of their poor biodegradability.

Alkyl ether sulfates are the most important surfactants, especially in Europe, but also in the United States. They display virtually all the positive qualities sought in a surfactant for use as a shampoo. Alkyl ether sulfates with one to three ethylene oxide units are easy to thicken; they are also easy to use and have good skin compatibility. The magnesium salts of alkyl ether sulfates with an ethoxylation value higher than 3 are especially gentle.

Sulfosuccinate esters (e.g., disodium laureth sulfosuccinate [*39354-45-5*]) and *protein – fatty acid condensates,* such as potassium coco-hydrolyzed animal protein [*68920-65-0*] (i.e., the protein is hydrolyzed and reacted with coconut oil fatty acid chlorides), have been known for a long time and are particularly gentle to the skin. Some drawbacks are their relatively high price and the sensitivity of the ester and amide bonds to hydrolysis. They are used in shampoo only as mixtures with the surfactants already cited.

Amphoteric surfactants used in shampoo can be divided into two classes: (1) the betaines, and (2) the alkyl amphoglycinates and alkyl amphopropionates. The *betaines* (e.g., alkylbetaines and especially alkylamidopropylbetaines such as cocoamidopropylbetaine [*61789-40-0*]) generally have good foaming qualities. Because of their normally more gentle properties, the amidopropyl derivatives are preferred. Betaines are compatible with anionic, cationic, and nonionic surfactants.

The *alkyl amphoglycinates* and *amphopropionates* especially the coco amphocarboxyglycinates [*68650-39-5*] and coco amphocarboxypropionates [*68919-41-5*], are in some cases even more gentle than betaines in terms of compatibility with skin and eye mucosa, but most of them have poorer lathering qualities.

Cationic surfactants have a positive charge and are strongly absorbed by the hair. Because of the problem of residue buildup (especially at acidic pH values), they are not generally suitable for use in shampoo alone and are used in mixtures with anionic surfactants. Only a few special-purpose shampoos, especially those for the treatment of severely damaged hair, contain these compounds.

Nonionic surfactants, with a few exceptions, have poor lathering power. Alkylpolyglucosides such as lauryl polyglucose (lauryl glucoside [*110615-47-9*]) are a new generation of nonionic surfactants used in shampoos. Based

on glucose and fatty alcohols, they have good cleaning and foaming characteristics, are stable over a wide pH range, and have positive ecological and dermatological properties. Other products in use are polysorbates (Polysorbate 20 [*9002-64-5*] and Polysorbate 80 [*9005-65-6*]), and fatty alcohol ethoxylates and polyglycerides, which have especially good lathering ability.

3.1.1.2. Other Constituents

Foam builders are used to improve qualities desired by the consumer, such as creaminess and quantity of lather, and foam stability. The main products are fatty acid mono- and dialkanolamides (e.g., Cocoamide MEA [*68140-00-1*] and Cocoamide DEA [*61791-3.21-9*]).

Refatting agents either reduce the defatting of hair caused by shampoo or are deposited on the hair to refat it. This function is carried out chiefly by the fatty acid mono- and dialkanolamides used as foam builders.

Conditioning additives prevent tangling during combing and brushing of wet hair, and impart a pleasant, smooth feel to the hair after drying. They also reduce static charge, which causes "flyaway hair" on combing and brushing. Products employed for this purpose include cationic polymers (e.g., quaternary hydroxyalkyl celluloses [*53568-66-4*]), silicones, and silicone derivatives. The admixture of amphoteric surfactants, fatty amides, and fatty acid esters to the basic anionic shampoo surfactant can also bring about conditioning effects.

Thickeners must be added to surfactant solutions to prevent the shampoo from running down the face and neck and into the eyes. Salts such as sodium chloride or ammonium chloride are used in combination with amphoteric surfactants, particularly betaines, poly(ethylene glycol) – fatty acid esters, alkanolamides, glycerol – fatty acid monoesters, and polymers such as hydroxyalkyl celluloses and acrylates.

Opacifiers give shampoo a rich, pleasing appearance. Relatively insoluble, higher melting, waxlike substances impart opacity or pearlescence to the shampoo, depending on the crystallization properties. The primary agents are fatty acid alkanolamides, chiefly in mixtures with ethylene glycol monostearate [*111-60-4*] and distearate [*627-83-8*] or cetyl alcohol [*36653-82-4*] and stearyl alcohol [*112-92-5*].

Coloring agents are added to improve the appearance of a shampoo. Only certain agents are allowed by law, see Section 10.2; care must be taken that they do not color the hair.

Ultraviolet absorbers are often required to stabilize dyes against the action of light, especially when the packaging is transparent. Use of these materials is also regulated in many countries.

Fragrances in shampoos mask the characteristic odor of the ingredients and impart a pleasant aroma to washed hair after rinsing. Because these substances often do not form a clear solution in the shampoo base, solubilizers such as highly ethoxylated fatty alcohols must be added. These oils and solubilizers can both greatly affect the viscosity of the product.

Buffers (pH stabilizers) such as citric, tartaric, adipic, and phosphoric acid and their salts are added to adjust the pH of shampoo to a value that is gentle to the skin. They also protect surfactants such as the alkyl ether sulfates against hydrolysis.

Preservatives are required in shampoo because surfactants normally represent a food source for a variety of microorganisms and are thus biodegradable. The use of preservatives is also regulated by law. Both the Federal Republic of Germany and the EC list permissible preservatives, see Section 10.1.

The many possible interactions between shampoo components and preservatives mean that the proper preservative for a given shampoo formulation must be determined by testing.

Other additives are described in Section 3.1.2.

3.1.2. Types of Shampoo

Shampoos can be classified according to the form in which they are supplied, as well as the effects produced by special additives.

Form.

Clear shampoos are used most frequently worldwide. The viscosity of these formulations is adjusted on the basis of practical experience. They are easy and inexpensive to manufacture, and completely satisfy consumer requirements.

Gel shampoos resemble clear shampoos but have a higher viscosity and are normally packaged in tubes. They may, however, be harder to distribute and to rinse out.

Opacified and pearlescent shampoos are more difficult to formulate than clear or gel shampoo. However, they have a richer appearance and can be formulated to contain more special additives because a clear solution is not necessary.

Aerosol shampoos in pressurized packages make even thin liquid formulations convenient to use. However, the packaging is more expensive, so this type of shampoo is seldom seen.

For numerous sample shampoo formulations, see [33 – 36].

Effects of Special Additives (see also Section 3.1.1)**.**

Shampoos for frequent use designed to meet the consumer trend toward daily hair washing and *baby shampoos* are based on especially mild anionic surfactants (e.g., sulfosuccinate esters, magnesium alkyl ether sulfates) or mixtures of anionic, amphoteric, and sometimes nonionic surfactants. With surfactant mixtures, the gentle action on the scalp and hair is due not to the concentration of the surfactants, but to the interaction between surfactants and other ingredients. Baby shampoo should further be formulated so that a burning sensation does not occur when the product comes in contact with the eyes.

Conditioning shampoos must render the hair easy to comb and tangle-free in the wet state, as well as glossy and soft when dry. These shampoos are especially important for dyed or bleached hair and long hair, which is difficult to untangle and comb after washing with simple shampoo. Conditioning effects are provided chiefly through the addition of cationic polymers. The deposition of insoluble polymer – surfactant complexes upon rinsing produces a thin film on the hair. In the wet state this film acts as a lubricant, and on dry hair it prevents static charge and stops hair from "flying" [35, 37, 38]. Addition of amphoteric surfactants or refatting agents can also improve the conditioning qualities of anionic surfactant formulations.

Two-in-one shampoos are a class of conditioning shampoos, containing silicones (dimethicones [*9006-65-9*]) or silicone derivatives (e.g., quaternium 80 [*134737-05-6*]) as additional conditioning agents.

Antidandruff shampoos contain agents that reduce excessive scalp flaking to a normal level. Because all such ingredients have been found to have a fungicidal action against *Phytosporum ovale,* this fungus is probably involved in the formation of dandruff [39, 40].

The most important antidandruff ingredients are zinc pyrithione [*13463-41-7*], Octopirox [*68890-66-4*] (1-hydroxy-4-methyl-6-(2,4,4-trimethylpentyl)-2(1*H*)-pyridone), and Climbazole (Baypival; 1-(4-chlorophenoxy)-1-(1-imidazolyl)-3,3-dimethyl-2-butanone [*38083-17-9*]). Selenium disulfide [*7488-56-4*] is also effective, but in some countries it is allowed only in medicinal products.

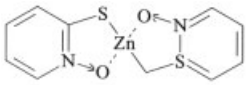

Zinc pyrithione

Shampoos for oily hair are used to diminish the excessive production of sebum or limit its distribution on the hair. This can be achieved through an appropriate choice of surfactant or the addition of special ingredients [36].

Special shampoos, lotions, and powders for the treatment of hair lice contain *insecticides,* e.g., 1,2,3,4,5,6-hexachlorocyclohexane [*608-73-1*], pyretrine I [*121-21-1*], pyretrine II [*121-29-1*], malathione [*121-75-5*], piperonyl butoxide [*51-03-6*], and mercury.

Head and body shampoos do not significantly differ from hair shampoos. *Tinting shampoos* are discussed in Section 5.2.4.

3.1.3. Dry Shampoos

Dry shampoos do not contain surfactants; they are powders that absorb sebum and dirt on the hair surface and remove it mechanically. Examples of ingredients are rice starch [*977000-08-0*], cornstarch, mechanical additives such as kieselguhr, [*61790-53-2*], and possibly alkaline substances such as borax or sodium carbonate. Dry shampoos have the advantage of cleaning without disturbing the hairstyle. However, drawbacks such as incomplete cleansing of the hair, soiling of clothes, and leaving residues on the hair have severely limited their importance in

the marketplace. For examples of dry shampoo formulations, see [35, 36].

3.2. Conditioning Agents and Treatments

Hair experiences a variety of influences in the course of its life: climatic effects such as humidity and temperature differences (weathering); exposure to sunlight; washing with products containing surfactants; cosmetic treatment such as bleaching, dyeing, permanent waving, or straightening; and combing or brushing. These processes alter the physical, chemical, and morphological properties of hair [41] as well as its reaction to cosmetic treatments such as dyeing and permanent waving. They lead to perceptible roughening of the hair surface, difficulty in combing, tangling, increased static charge, formation of split ends, and loss of natural luster.

Hair conditioning agents prevent, retard, or mask such changes. Their action is restricted largely to modifying the surface qualities of the hair [42, 43] and making it glossy. Nearly all modern shampoos, permanent wave lotions, setting lotions, and dyes contain conditioning additives that prevent excessive mechanical damage to the hair. This simple type of conditioning is not, however, sufficient for chemically treated or long hair; in such cases a special conditioning treatment is also required.

Conditioning treatments are sold as, e.g., conditioners, rinses, intensive treatments, mousse conditioners, and balms. The earliest conditioning treatments were waxes and oils, which provided lubrication and enhanced luster, as well as acid rinses (vinegar), which counteracted the undesirable deposition of lime soap and the swelling of hair. These early products were then combined with anionic emulsions in rinsable "packs."

An improvement in performance was achieved by the addition of cationic surfactants as emulsifiers. The lubricant action of emulsified fats and waxes is aided by charge neutralization when the cationic surfactant is adsorbed on the anionic surface of the washed hair [19, 43, 44]. In this way, an adherent lubricant coating is produced that consists of the nonpolar portions of the surfactant.

The oil-like substances used are primarily fatty alcohols such as cetylstearyl alcohol [*8005-44-5*]; cationic surfactants include cetyltrimethylammonium chloride [*112-02-7*] (cetrimonium chloride) or alkylamidoamines (e.g., stearamidopropyl dimethylamine [*7651-02-7*]). Cationic polymers [45, 46] (see page 11) can also be used in conditioning treatment, along with special ingredients such as antidandruff additives. Polydimethylsiloxanes and organofunctional silicones [47] are widely used in conditioning agents because of their excellent conditioning properties. Numerous formulations can be found in [34 – 36].

Conditioning treatments are generally made by emulsifying the molten wax phase in the aqueous phase containing the surfactant. When the emulsion is cooled, part of the wax phase crystallizes; thus the viscosity increases and the system becomes more stable [48]. The fatty alcohols and the surfactants form gel systems in which the wax and fat constituents are dispersed. Dilution with water breaks the gel structure and liberates the wax [49 – 51].

The addition of special wax components yields pearlescent preparations. If water-insoluble components are omitted and cationic polymers are added, the formulas obtained are transparent and clear.

Besides the usual highly viscous systems (packs, conditioners) and the less viscous ones (rinses), other product forms such as mousses and spray conditioners (aerosol or pump-spray) are possible [7].

Conditioning treatments are generally rinsed out after they have been allowed to work for a defined time. However, some products with a lower content of active ingredients can be left on the hair without rinsing, so called leave-in treatments/conditioner [52]. Leave-in treatments have become more important because specific market needs/niches are convered by that product type. Examples are products for volumizing hair containing water-soluble polymers to give volume, hair ends fluids for splissed hair containing cationic actives and/or combinations of oils in order to increase the cohesion of split ends.

Beside the raw materials needed for base formulations (cationic compounds, fatty alcohols, etc.) additional actives to achieve specific effects on hair are increasingly important. Examples are

panthenol [53] for moisturizing and strengthening of hair, hydrolyzed proteins [54 – 56] for reducing hair damage or creatine [57] and ascorbic acid [58] for hair strengthening.

Other Hair Conditioning Products. Changes in fashion have meant that *hair creams,* especially for men, have lost much of their commercial importance. These products improve the combability and luster, and help to hold the hairstyle. Creams include oil-in-water and water-in-oil types.

Hair tonics refresh the scalp and serve as simple styling aids. They may also contain disinfectants; soothing, cooling, or hair-growth-promoting ingredients, and anti-oil or antidandruff agents [35].

4. Hairstyling Preparations

Styling preparations help to obtain a desired hairstyle and stabilize it during and after its creation. Tools to achieve the hairstyle are comb, brush, flatting irons, curlers (with and without heat) or simply the hands. Base for styling is usually achieved by a temporary set (see Chap. 6). Styling products may also make hair easier to manage, for example by improving its wet combing or brushing qualities. These products include hair sprays, setting lotions, setting mousses, and restructurants. Newer products developed in response to technology and fashion changes are "styling gels", "wet gels", "gel sprays", "hair waxes", shine products like "shine sprays" or "serums", and different types of emulsion-based products with and without polymers.

Styling preparations generally remain in the hair and are not rinsed out. Their active ingredients are usually dissolved polymers, known as film-forming agents, which are deposited on the hair after the evaporation of a suitable solvent. The stabilization of the hairstyle achieved with setting lotions and mousses is due to stiffening and increased friction between hairs [59]; with hair sprays, to cross-linking in the outermost layer of hair; and with gels and waxes, to adhesion between hairs. These preparations also protect the hairstyle to some degree against the action of moisture; i.e., they counteract the reversibility of the setting process.

4.1. Setting Lotions, Mousses, and Restructurants

Setting lotions and mousses make a hairstyle more durable. They must prevent hair from "flying" away, reduce the amount of charge when it is combed or brushed, improve wet or dry combing, and improve its feel and luster. They must not separate from the hair or form flakes, become tacky or sticky at high humidity, or lather during wet combing (see also [7, 60]). *Setting lotions/mousses for temporary sets* (wave sets) are normally distributed in the hair after it has been washed and toweled dry. After that the hair is brushed and combed while using a blow dryer. In the classic way of use of these lotions/mousses hair is wound on rollers and dried.

Restructurants are setting products with special care qualities; they are especially suitable for badly damaged hair.

Composition. The first preparations comparable to modern setting lotions were aqueous or aqueous – alcoholic solutions or gels of natural substances. Examples are egg white or sugar solutions, plant mucilage (pectins, alginates, carrageenan, karaya gum, tragacanth) [61], and beer. These products simply "glued" the hairs together. They had the drawback of forming opaque, brittle residues on the hair, which created dust during combing or brushing and became sticky in the presence of moisture [62].

Modern setting mousses contain polymers (film-forming agents) [62], solvents surfactants, combability improvers, plasticizers, fragrances, coloring agents, UV stabilizers, preservatives (if necessary), and other special ingredients. For sample formulations, see [35, 36].

Film formers may be nonionic, anionic, or cationic polymers dissolved in water or water – alcohol mixtures.

Nonionic film-forming polymers include polyvinylpyrrolidone [*9003-39-8*], the first polymer employed for this purpose. Its high solubility in a variety of solvents is offset by its hydroscopicity, which causes the film to become tacky when air humidity is high. A substantial improvement was made with vinylpyrrolidone – vinyl acetate copolymers [*25086-89-9*], which form much less tacky films. The moisture sensitivity and hardness can be controlled via the ratio of monomer units.

Other nonionic polymers are copolymers of vinylpyrrolidone and dimethylaminoethyl methacrylate monomers [*30581-59-0*] (CTFA vinylpyrrolidone dimethylaminoethyl methacrylate copolymers), polyvinylcaprolactam [*25189-83-7*] (CTFA polyvinylcaprolactam), or vinylcaprolactam dimethylaminopropyl methacrylate copolymer. Vinylcaprolactam offers a better humidity resistance than vinylpyrrolidone [63 – 66].

Anionic film formers are available in a large variety. Normally they are neutralized with organic amines such as 2-amino-2-methyl-1,3-propanediol (AMPD), 3-amino-2-methylpropanol (AMP), or triisopropanolamine (TIPA). Neutralization makes the polymers more soluble and easier to wash out, and the films become softer. Copolymers of vinylacetate-crotonic acid [*25609-89-6*], of methyl vinyl ether and maleic acid hemiesters [*67724-93-0*], acrylates copolymers, sodium polystyrene sulfonate, terpolymers of vinyl acetate, crotonic acid and vinylneodecanoate [*55353-21-4*], acrylic acid are examples of this polymer class.

Cationic polymers have an affinity for keratin, thus making the hair easier to untangle; they also prevent the accumulation of static charge during combing and brushing. However, because cationic polymers become tacky at high atmospheric humidity, they are usually combined with nonionic or anionic film-forming agents. The most important cationic film-forming polymers are quaternary vinylpyrrolidone – dimethylaminoethyl methacrylate copolymers [*37348-62-2*] (CTFA Polyquaternium 11); vinylpyrrolidone – vinylimidazole copolymers [*29297-55-0*] (CTFA Polyquaternium 16) [67]; cationic hydroxyethyl cellulose [*54351-50-7*] (CTFA Polyquaternium 10); poly(dimethyldiallylammonium chloride) [*26062-79-3*] (CTFA Polyquaternium 6); poly(dimethyldiallylammonium chloride) copolymers [*26590-05-6*] (CTFA Polyquaternium 7); copolymers of hydroxyethyl cellulose and polydimethylallylammonium chloride [*92183-41-0*] (CTFA-Polyquaternium-4), vinylpyrrolidone-vinylimidazolium-methacrylamide terpolymers [*827346-45-2*] (CTFA Polyquaternium 68), vinylcaprolactam vinylpyrrolidone and methylvinylimidazolium methosulfate [*174761-16-1*] (CTFA Polyquaternium 46), vinylpyrrolidone dimethylaminopropyl methacrylamide methacroylaminopropyl lauryldimonium chloride copolymer [*306769-73-3*] (CTFA Polyquaternium 55) and chitosan (obtained from chitin by alkali treatment) or its derivatives [68 – 72].

The *solvents* used are water, ethanol, 2-propanol, and their mixtures [70].

Combability improvers include the cationic film-forming polymers listed above and cationic surfactants such as cetyltrimethylammonium chloride, bromide, or phosphate. Cationic silicones known as amodimethicone are also often used.

Plasticizers enhance the flexibility of films formed on the hair. They include esters such as triethyl citrate, adipates, silicones, and polyglycols.

Other ingredients are *fragrances, coloring agents,* and *UV stabilizers.* The regulations in force in each country must be taken into account in the selection of these substances, see 10.1. The effects of their interactions on product stability are determined by storage tests.

Preservatives are generally required at alcohol contents of <15 wt %.

Antidandruff and anti-oil additives, skin-soothing and moisturizing substances, and plant extracts may also be included.

Foaming agents are cationic surfactants listed above other surfactants that improve foaming are alkylpolyglucosides or highly ethoxylated fatty alcohols. These nonionic surfactants are compatible with all types of film-forming agents and are important for solubilizing fragrances and other water-insoluble additives in purely aqueous systems.

Propellants are hydrocarbons such as propane, butane, isobutane and their mixtures, dimethylether or hydrofluorocarbons as HFC 152a [*75-37-6*].

Organosilicones are a special class of additives. Depending on their structure, they can improve combability, shine, and feel of the hair. Several types of organosilicones are used, e.g., cyclomethicone [*69430-24-6*], amodimethicone, dimethicone, dimethicone copolyols, and phenyltrimethicone [*63148-58-3*].

Formulation. Mousses are usually thin aqueous or aqueous-alcoholic solutions in alumina cans with the abovementioned propel-

lants. The solution/propellant ratio varies between 95:5 and 75:25. The amount of polymers or polymer mixtures varies from 0.5 % – 8 %.

Three types of mousses are the most common: Cationic mousses: Here PQ 11 or PQ 4 are the most often used polymer types. Nonionic/cationic mousses: Mixtures of VP/VA Copolymers combined with cationic polymers such as PQ 16, PQ 4, or chitosan are frequently employed. Anion-based mousses: Here combinations of acrylic acid hemiesters or acrylates copolymers with cationic polymers are often found. Also combinations of VA/crotonic acid copolymer with PVP are in use [68, 72].

Due to legal requirements mousses for the US market are limited in their VOC (volatile organic compounds) content up to 6 % (i.e., hydrocarbons, dimethyl ether, or ethanol are classified as VOCs). For mousses with a higher content of propellants fluorohydrocarbons F 152a is used in combination with hydrocarbons [73].

Nonaerosol mousses have a similar composition as aerosol mousses. Here, the mousse is generated through use of squeeze or pump dispenser with a riddle in the head of the pump. The setting liquid is mixed with air in the head of the pump and dispensed as mousse.

Setting lotions are also marketed as thin, aqueous – alcoholic solutions in single- or multiple-application packages. Depending on intended use, they contain 0.5 – 4 % of polymers or polymer mixtures and up to 50 % alcohol. Setting lotions for waving have a higher polymer content than those for blow-drying. Liquid lotions can also be packaged as aerosol sprays (propellant or pump spray, see page 11). Addition of adsorptive or penetrating dyes to setting lotions/mousses allows simultaneous hair coloring. Tinting and setting lotions or mousses are discussed in Section 5.2.4.

4.2. Hair Sprays

Hair sprays are used to stabilize the hairstyle against wind and humidity, for example. The hair must nonetheless appear natural and must not be sticky even at high humidity. The hair spray must dry quickly; it should also be removed easily by brushing or washing because otherwise, the hair becomes heavy and readily soiled. Removal of the spray must not generate visible dust; i.e., the film must be clear, colorless, and invisible, even on dark hair. After application, the hair must have a pleasant smell, be glossy, and not feel dull. The pressure in the hair spray container must not be so high that it destroys the hair style, but must be high enough to give an acceptably small droplet size, so that the hair spray does not appear "wet."

For general information on cosmetic aerosols, see [61].

Aerosol Sprays. Hair sprays are usually two-phase aerosols; both a liquid and a gaseous phase are present inside the container.

Hair sprays are supposed to have originated from the insecticide bombs developed in the United States during World War II. The propellant was dichlorodifluoromethane, but this gas has a rather high pressure (0.5 MPa at 20 °C). Hair sprays did not become practical until "liquefied" gases with a lower vapor pressure became available. In the container, these gases are present in the liquid state. By virtue of the liquid – gas equilibrium, the pressure remains constant through spraying. As early as 1926 – 1929, however, patent applications in Norway and Germany disclosed methods and devices corresponding to those found in modern hair sprays. Dimethyl ether, methyl chloride, isobutane, and other propellant gases were used [74].

The first hair spray resin was shellac [*9000-59-3*], which was difficult to comb or wash out of the hair. In some cases, hairdressers had to treat the hair several times with alcohol – shampoo – water mixtures.

The most important ingredients in hair spray today are the *concentrate* (film formers, solvents, plasticizers, agents for luster, and fragrances) and the *propellant.* For lists of hair sprays and their formulations, see [7, 34 – 36, 61, 75].

Film-Forming Agents. The film-forming agent shellac was superseded in the early 1950s by polyvinylpyrrolidone, which is easier to remove from the hair. Better results are attained with vinylpyrrolidone – vinyl acetate copolymers (see Section 4.1), the harder, more hydrophobic types being used for hair sprays. Still harder films are possible with the vinyl acetate – crotonic acid copolymers discussed in Section 4.1. The other anionic film formers listed in Section 4.1 can also be employed in hair sprays.

Copolymers of methyl vinyl ether and ethyl maleate [*67724-93-0*] or butyl maleate are used most often in the United States. More recently, acrylates, such as *tert*-butylacrylamide – ethyl acrylate – acrylic acid copolymers [*26062-56-6*], [*83457-27-6*]; acrylate – acrylamide copolymers [*977062-93-3*]; and vinylcaprolactam – vinylpyrrolidone – dimethylaminoethyl methacrylate copolymers have been used [76].

Solvents. Ethanol and 2-propanol are the most important solvents in hair sprays. Acetone is still occasionally used. Methylene chloride (dichloromethane) used to be popular, but it is no longer employed in some countries due to controversy concerning its toxicological properties. The maximum water content in hair spray is limited to about 10 – 15 %, otherwise the hairstyle is impaired by destruction of the temporary set. Other components in the concentrate are fragrances [77], plasticizers, and ingredients used to add luster. The same products are used as in setting lotions (see Section 4.1).

Propellants. For a comprehensive list of propellants and diluents, see [61]. Formerly, the most-common propellants were chlorofluorocarbons (CFCs). These products are nonflammable, chemically inert, and good solvents for virtually all polymers. However, they have a high ozone-depletion potential (ODP).

Therefore, by 1979, controversy over the possible action of chlorofluorocarbons on the earth's ozone layer [78 – 83] led the United States to ban their use. International agreements and national legislation have led to a gradual abandonment of CFCs as propellants worldwide. At present, only negligible quantities of CFCs are still used.

Alternatives are the hydrocarbons propane, butane, isobutane, pentane, and their mixtures. These hydrocarbons are highly flammable and poor solvents [84] for most polymers; special polymers (polyacrylates) that are more readily soluble in organic solvents or cosolvents (e.g. dimethyl ether) must be employed. Dimethyl ether is on a par with hydrocarbons with respect to flammability, but it combines the properties of a solvent and a propellant [85, 86].

Packaging. In contrast to other hair preparations, packaging and packaging technology play an important role for aerosols. Containers are usually made of metal, such as tinplate (sheet steel plated with tin) or aluminum, generally with a protective coat of lacquer on the inside. The other components are the valve and the spray nozzle. The selection of nozzle is crucial for the turbulence and particle size of the spray.

Pump sprays ("nonaerosol" sprays) do not employ a propellant gas; thus the selection of resins is less critical. The energy needed to swirl and atomize the concentrate must be supplied by a pump, which is operated with a finger or through a lever. The ingredients of nonaerosol hair sprays are the same as those of aerosol products, but the propellant content is replaced by additional solvent.

4.3. Other Hairstyling Preparations

4.3.1. Gels

Gels became more and more popular since the 1990s. They consist of the following ingredients:

Thickeners: Synthetic gel builders are poly(acrylic acid) cross-linked with alkyl ethers of pentaerythritol [*9003-01-4*] (CTFA Carbomer), Copolymer of C10-C30 alkyl acrylate and/or acrylic acid methacrylic acid or one of their esters cross-linked with alkyl ethers of sucrose (CTFA Acrylate C10-C30 Alkyl Acrylate Crosspolymer), copolymers of half esters of itaconic acid and Ceteth-20 and one or more monomers of acrylic or methacrylic acid or one of their esters (CTFA: Acrylates/Ceteth-20 Itaconate Copolymer) or copolymer of Beheneth-25 methacrylic ester with acrylic/methacrylic acids or their esters (Acrylates/Beheneth-25 Methacrylate Copolymer). Natural-based gel builders are, e.g., cellulose with hydroxyethyl side chains [*9004-62-6*] (CTFA Hydroxyethylcellulose) or the propylene glycol ether of guar [*39421-75-5*] (CTFA Hydroxypropylguar) [87 – 91].

Solvents, formulation aids and polymers are the same as described in Sections 4.1 and 4.2

Formulation. Thickeners are used in concentration from 0.15 % up to 3 %. The concentration of setting polymers varies between 0.5 % and 10 %. For wet gels moisturizing agents like glycerol, propylglycol or sorbitol are used up to 25 %. Common packaging are polyethylene tubes, dispensers, jars or pump spray bottles for sprayable gels.[94, 95]

4.3.2. Waxes (→ Water)

Waxes are used to form the hairstyle in part or in total. Major components are hydrocarbon waxes such as paraffin wax [*8002-74-2*] (CTFA Paraffin), ceresin [*8001-75-0*] (CTFA Ceresin), ozokerite [*12198-93-5*] (CTFA Ozokerite) but also natural waxes as beeswax [*8006-40-4*] (CTFA Beeswax), carnaúba wax, orange peel wax (CTFA Citrus Aurantium Dulcis (Orange) peel wax) or apple wax (CTFA Pyrus Malus (Apple) peel wax) [92].

Oils of synthetic or natural origin are often used as additives (i.e., paraffin oils, avocado oil, jojoba oil or silicone oils) [90, 93].

Emulsifiers such as the Ceteth-4 triester of phosphoric acid [*119415-05-3*] (CTFA Tricetearetn-4 phosphate), ethoxylated fatty alcohols with ethylene oxide chain length between 15 and 40, i.e., Ceteth-20 [*9004-95-9*] (CTFA Ceteth-20), or Steareth-30 [*9005-00-9*] (CTFA Steareth-30) are added to improve the washout ability of the hydrophobic waxes [60].

Another type of wax components are poly(ethylene glycols) [*25322-68-3*] with higher molecular mass (M_r 400–10 000), for example CTFA PEG-8 – PEG-220. Due the good washout ability of the PEGs no additional emulsifiers are needed. PEGs as main wax components impart mixtures with hydrophobic waxes such as paraffin wax or petrolatum (up to 25 %) still a good washout ability [60, 93].

Formulation. Mixtures of waxes are melted and mixed with other components.

Beside the composition of the wax components also filling temperature and cooling time and temperature can influence hardness of the wax, because these parameters can induce different types of crystallization.

Aerosol-wax, a special application type of waxes, can be achieved by mixing the wax components with volatile hydrocarbons such as pentane, isopentane, isododecane and hydrocarbon propellants. This type of wax can be sprayed on the hair or applied into the hands.

4.3.3. Emulsion-Based Styling Products

There are several product types that are based on emulsions. Since the mid-1990s these products became more and more important especially for professional hairstyling.

4.3.3.1. Creams and Pomades

A special form of cream-type products are gel waxes, that mainly consist of emulsifiers together with oil and wax components.

Polyethoxylated fatty alcohols, short- and long-chain ethoxylated oleyl alcohols such as Oleth-5 [*5353-27-5*] (CTFA Oleth-5), Oleth-20 [*9004-98-2*] (CTFA Oleth-20), and ethoxylated hydrogenated castor oils, i.e., hydrogenated castor oil with 40 ethylene oxide units [*61788-85-0*] (CTFA PEG-40 hydrogenated castor oil) are some of the most common types.

The most commonly used oils are hydrocarbon-based like paraffin oil [*8012-95-1*], [*8020-83-5*, [*8042-47-5*] (CTFA Mineral Oil) and paraffin replacers like the mixed triester of capric and caprylic acid [*65381-09-1*], [*73398-61-5*] (CTFA Caprylic/Capric Triglyceride) or branched higher alcohol such as octyldodecanol [*5333-42-6*] (CTFA Octyldodecanol).

Common preservatives are systems as 1,3-dimethylol-5,5-dimethylhydantoin [*6444-58-0*] (CTFA DMDM Hydantoin), 2-phenoxyethanol [*122-99-6*] (CTFA Phenoxyethanol), benzylic alcohol [*100-51-6*] (CTFA Benzyl Alcohol), or the esters of hydrobenzoic acid with methanol, ethanol, *n*-propanol or *n*-butanol: [*99-76-3*] (CTFA Methylparaben), [*120-47-8*] (CTFA Ethylparaben), [*94-25-7*] (CTFA Butylparaben) [60].

Formulation. Emulsifiers are used in a concentration between 15% and 30 %, oil concentration varies between 10% – 20 %. Emulsifying temperature ideally is above 90 °C to achieve a clear emulsion. Optional raw materials are polymers which to improve level of setting. Type of polymer depends on type of emulsifier used, it always has to be checked which emulsifier/oil system is compatible with the type of polymer. Using lower amounts of emulsifier would lead to classic hairstyling creams.

4.3.3.2. Gums/Puttys

This product type encompasses emulsions from fatty alcohols, waxes and/or oils, emulsifiers and

polymers. The consistency of these products can vary from a soft thin cream to hard rubberlike materials.

Formulation. Major wax components are fatty alcohols, carnaúba wax, ozokerite or ceresin. There is a broad variety of emulsifiers in use. Often used polymers are nonionic types like VP/VA Copolymers, PVP but also amphoteric or anionic polymers are applied.

Often a thickener is incorporated to stabilize the emulsion and/or to improve the texture of the product [60].

4.3.3.3. Serums

Serums are used to improve the shine of the hair but can be also used for preparation of curls or for straightening the hair. Most important components are silicones. Dimethicones of different molecular mass are one major component to influence the viscosity of the products. Phenyltrimethicones improve the shine and improve definition properties. Cyclomethicones but also volatile hydrocarbons and ethanol are used as "solvent" of serum products. Depending to their viscosity the products can be applied via a dispenser pump or as a non aerosol spray to improve shine of the hairstyle.

5. Hair Coloring Preparations

This chapter deals with cosmetic agents used to modify hair color. However, changes due to intervention in biochemical processes (e.g., side effects of medication) are also known [98]. For example, the antimalarial drug chloroquine [*54-05-7*] and the spasmolytic mephenesin [*59-47-2*] brighten hair color. After discontinuation of drug treatment, the hair grows back in its original color. Chemicals such as hydroquinone [*123-31-9*] and its ether derivatives, also inhibit melanin formation [99]. These ways of altering color should not, however, be regarded as cosmetic treatment because of their mechanism of action and potential side effects (allergy, irritation, hair loss).

5.1. Bleaching

Bleaching, blonding, and lightening of human hair involve the irreversible destruction of melanin pigments by oxidation; partial or complete degradation is possible. The degree of lightening is controlled by the type and concentration of oxidant, the temperature, and the duration of treatment. Because melanin degradation also entails modification of the keratin in hair, the maximum possible extent of lightening is limited by hair quality.

Chemistry of Bleaching. The structure and formation of melanin are discussed in Section 2.2.3. Research has shown that the melanin in hair is relatively resistant to reduction and, in acid solution, to oxidation as well. In alkaline hydrogen peroxide, however, the melanin granules are attacked quickly. Evidently the polymer structure is partly destroyed by oxidation, with the formation of carboxyl groups that facilitate dissolution of the granule under alkaline conditions [19, pp. 67 – 79], [100]. If bleaching is complete, a void is left at the site of the melanin granule [101]. Persulfates are generally added to hydrogen peroxide to intensify the bleaching action. Hydrogen peroxide can also be replaced by its addition compound with urea, carbamine peroxohydrate [*124-43-6*]. Other oxidants, such as sodium carbonate peroxohydrate ($2Na_2CO_3 \cdot 3\ H_2O_2$) or sodium peroxoborate tetrahydrate ($NaBO_3 \cdot 4\ H_2O$) have not met with acceptanc; they are only used in speciality products. For a detailed description of peroxide bleaching, see → Bleaching, Chap. 2.2.4.

The undesirable reaction between bleach and the keratin in hair or wool has been well studied [102]. The principal point of attack is the disulfide bond of cystine, which is oxidatively cleaved via a series of intermediates [103]. The main degradation product is cysteic acid, which is responsible for the higher alkali solubility of bleached hair. In addition, hydrogen bonds and ionic bonds can be destroyed by hydrolysis. The hydroxyl and amino side chains of various keratin amino acids offer points of attack for oxidative degradation [7, pp. 216 – 221]. When a strong bleach is used, alteration of the chemical properties of hair also results in modified physical properties: a higher extensibility and thus lower mechanical strength; a rough, strawlike feel when dry; and a spongy feeling when wet.

Because of its higher porosity, hair swells more rapidly in water, a permanent wave can be produced with more dilute solution, and the dyeing behavior is modified.

Photochemical bleaching of hair by sunlight or UV light has also been investigated with regard to bleaching of natural hair color [104, 105] and the damaging properties of light on hair keratin [106, 107].

Bleaches. Bleach consists of an oxidant and a vehicle to prevent the product from running out of the hair. For sample formulations, see [108].

The most intensive bleaching action is achieved with *paste bleaches,* which are prepared before use by mixing hydrogen peroxide solution (6 – 12 vol %) with a bleach powder. The powder consists of a peroxodisulfate, an alkalinizing agent, stabilizers, thickeners, and other additives. Sodium, potassium, or ammonium peroxodisulfate is used. The ammonium salt is most effective; when combined with an alkalinizing component such as sodium carbonate or silicate, ammonia is formed. The base readily penetrates the hair and promotes bleaching. The addition of stabilizers such as sodium pyrophosphate or sodium oxalate [109, 110] retards the decomposition of hydrogen peroxide in the alkaline preparation and thus enhances the bleaching action. The same holds for complexing agents (sequestrants) such as ethylenediaminetetraacetic acid, which hinder decomposition due to traces of heavy metals. Thickening additives include carboxymethyl celluloses, xanthine derivatives, and synthetic polymers. Certain dyes can also be added.

Dusting of bleaching powders has to be prevented to avoid inhalation of peroxosulfates. This can be achieved by :

- Addition of oils to bind fine particles [111]
- Granulation to enlarge particle size [112]
- Formulation of water free oil containing pastes [113]

Other formulations are *cream and gel bleaches*. These consist of an alkaline vehicle in cream or gel form, which is combined with hydrogen peroxide solution. Peroxosulfate is not contained in these bleaches, but can be added as a bleaching powder to reinforce the action. Bleaching powders formulated on the basis of peroxosulfates and water-free hydrogen peroxide addition compounds are activated by mixing with water instead of hydrogen peroxide solution. In view of their market success, this advantage does not compensate for their poor stability.*Setting lotion* and *shampoo* forms are less important; in acidic solution, these have low hydrogen peroxide concentrations and give a weak lightening effect.

In spite of the importance of bleaches in oxidation dyeing, very few technical improvements apart from formulation aspects have occurred in this field. The selective adsorption of metal ions [114], especially of iron(II) salts [115], on melanin has been proposed for gentler bleaching of human hair. This process has achieved no more acceptance than the use of peroxocarboxylic acids or their precursors, which are important as bleaching intensifiers in textile bleaching.

Application. The highly effective paste bleaches are normally prepared just before use and applied with a brush; skin contact is minimized. The products are left to work for up to 60 min; this time can be shortened by heating or by increasing the concentration of hydrogen peroxide. The problem of hair damage must always be considered. After the bleach has been rinsed out, the hair is generally treated with a conditioning product (see Section 3.2). The formation of acid groups during oxidation promotes the adsorption of cationic compounds, which improve the feel and combability of the hair. To prevent "delayed" oxidation, which may result from residual hydrogen peroxide in the hair, antioxidants such as ascorbic or glyoxylic acid may be added to the conditioning agent [116]. If the conditioner is adjusted to an acidic pH, alkali residues are neutralized and the hair swelling is reversed.

5.2. Dyeing

Hair dyeing includes the use of permanent, semipermanent, and temporary dyes. A *permanent dye* lasts through any number of washings as well as permanent waving. A *semipermanent dye* is removed after two to ten washings, and a *temporary dye* is largely eliminated after one washing.

5.2.1. Dyeing with Dye Precursors

The best method of achieving a permanent hair color is the use of oxidation hair dyes. Only with

these dyes can all the requirements, such as adequate color range, masking of white hair, and permanence, be satisfied. The principle of oxidation dyes is over 100 years old [117] and is used in a similar form for dyeing furs (→ Furs, Chap. 4.3). In hair dyeing, however, toxicological considerations are more important. The dyes are produced inside the hairs from colorless precursors by oxidation with hydrogen peroxide in alkaline solution [118]. Given a suitable choice of intermediates, the hair is uniformly colored. The dyes themselves and the kinetics of their formation have been studied intensively [119].

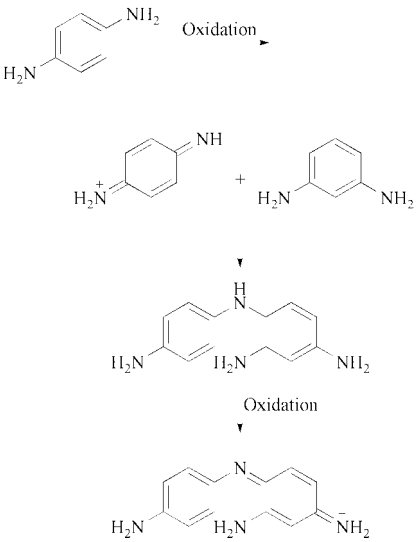

Figure 4. Oxidative formation of hair dyes

Oxidation. A *primary intermediate,* an aromatic para compound, such as a derivative of 1,4-diaminobenzene or of 4-aminophenol, is oxidized to a quinonediimine or quinonemonoimine, respectively. The imine then reacts with a *secondary intermediate* (coupler), which is a meta compound such as a derivative of 1,3-diaminobenzene, 3-aminophenol, or resorcinol (see Fig. 4). Another oxidation step yields an indo, phenazine, or oxazine dye (→ Azine Dyes, Chap. 2.2.7) which is three times the size of the precursors. This size enlargement is largely responsible for fixing the dye in the hair. With the proper choice of substituents, any color from yellow to blue can be obtained [7, pp. 279 – 283], [120]. If one or two primary intermediates are mixed with various (usually three to five) couplers, a range of colors is produced in the hair, giving the desired shade as an overall impression (Fig. 5).

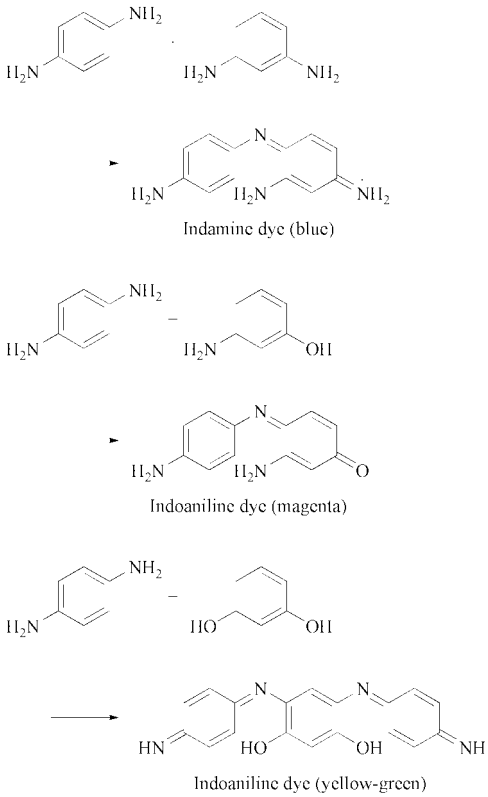

Figure 5. Formation of color using different couplers

Oxidants include hydrogen peroxide or its addition compounds; in special cases, atmospheric oxygen can also be used. Enzyme systems have also been proposed [121]. The alkali used to accelerate dyeing and swelling of the hair is ammonia or, less often, monoethanolamine. Under these conditions, lightening of the natural melanin pigment depends on the concentration of the oxidant and the primary intermediates. Lightening is desirable because it produces a more even background color; the synthetic dye would otherwise give rise to different shades when applied to gray hair as a mixture of naturally pigmented and white hair.

The simultaneous bleaching process also allows hair to be dyed a lighter color; only small amounts of dye intermediates are needed to off-

set the shade produced by bleaching. The degree of lightening is limited; the color corresponds to that obtained by cream bleaching without peroxodisulfate (Section 5.1).

Advances in the field of hair dyeing by oxidation have related mainly to the areas of toxicology and application techniques.

Primary Intermediates. 1,4-Diaminobenzene [*106-50-3*] (*p*-phenylenediamine) has been the main primary intermediate used worldwide for more than a 100 years but has now been largely replaced by 2,5-diaminotoluene [*615-50-9*]. Other primary intermediates are tetraaminopyrimidine and N,N-dialkylated, ring-alkylated, or ring-alkoxylated 1,4-diaminobenzene derivatives [e.g., 2,5-diamino(hydroxyethylbenzene) [*93841-24-8*]]. 4-Aminophenol and its derivatives, such as 4-amino-3-methylphenol [*2835-99-6*], yield red-violet and orange indoanilines and indophenols with the usual couplers; they are therefore important for red shades, whereas diaminopyrazoles and their derivatives [e.g., 1-(2-hydroxyethyl)-4,5-diaminopyrazole [*155601-17-5*]] yield bright red, red violett and orange colors with a broad range of different couplers [122].

Couplers can be classified into three groups according to the color obtained with the aromatic 1,4-diamines: blue, red, and yellow-green couplers.

The usual *blue couplers* are 1,3-diaminobenzene derivatives with the ability to couple at the 2,4-positions relative to the amino groups. Examples in use include *N*- [123] and *O*-hydroxyethyl derivatives [124] dimeric couplers [125], and heterocyclic compounds [126].

The principal *red couplers* are 3-aminophenol [*591-27-5*], 5-amino-2-methylphenol [*2835-95-2*], and 1-naphthol [*90-15-3*]. Among *yellow-green couplers,* resorcinol [*108-46-3*], 4-chlororesorcinol [*95-88-5*], and benzodioxoles should be mentioned, as well as (with restrictions) methylresorcinol derivatives. The importance of the yellow-green couplers lies in the broad-band absorption of the dyes produced, which makes natural-looking hair shades possible. A overview on Combinations of primary intermediates and couplers used in hair colorants is given in Table 3.

Table 3. Combinations of primary intermediates and couplers used in hair colorants

Primary intermediate	Coupler	Color on hair
2,5-Diaminotoluene	*m*-phenylenediamine	blue
2,5-Diaminotoluene	2,4-diaminophenoxyethanol	violet-blue
2,5-Diaminotoluene	*m*-aminophenol	magenta-brown
2,5-Diaminotoluene	5-amino-2-methylphenol	magenta
2,5-Diaminotoluene	1-naphthol	purple
2,5-Diaminotoluene	resorcinol	greenish brown
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	*m*-phenylenediamine	green-blue
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	2,4-diaminophenoxyethanol	green-blue
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	*m*-aminophenol	green-blue
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	5-amino-2-methylphenol	violet-blue
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	1-naphthol	blue
N,*N*-Bis-(2-hydroxyethyl)-p-phenylenediamine	resorcinol	greenish brown
4-Amino-3-methylphenol	*m*-phenylenediamine	magenta
4-Amino-3-methylphenol	2,4-diaminophenoxyethanol	magenta
4-Amino-3-methylphenol	*m*-aminophenol	pale orange red
4-Amino-3-methylphenol	5-amino-2-methylphenol	orange red
4-Amino-3-methylphenol	1-naphthol	red
4-Amino-3-methylphenol	resorcinol	yellow-gray
4,5-Diamino-1-hydroxyethyl pyrazole	*m*-phenylenediamine	magenta brown
4,5-Diamino-1-hydroxyethyl pyrazole	2,4-diaminophenoxyethanol	purple-brown
4,5-Diamino-1-hydroxyethyl pyrazole	*m*-aminophenol	bright orange red
4,5-Diamino-1-hydroxyethyl pyrazole	5-amino-2-methylphenol	intense red
4,5-Diamino-1-hydroxyethyl pyrazole	1-naphthol	purple
4,5-Diamino-1-hydroxyethyl pyrazole	resorcinol	red

Other dye intermediates are derivatives of 2-aminophenol [*95-55-6*]. Trihydroxybenzenes, dihydroxyanilines, or diphenylamines are used chiefly in "autooxidative" systems, i.e., systems using oxidation with atmospheric oxygen without an added oxidant. The diphenylamines are oxidized to indo dyes (an overall term including indamine, indoaniline, and indophenol dyes). The oxidation of 1,4-diaminobenzene alone leads to Bandrowski's base, a skin irritant [127]; (→ Azine Dyes, Chap. 2.2.7). Intensive work has been carried out since the first patent appeared [128] to produce indole precursors that mimic the biological melanogene-

sis process. Until now, the performance of these products has been low. For reviews of the patent literature on new intermediates as well as general progress in the field of hair dyeing generally, see [129, 130], [131].

Some nitro dyes (see below) are used in oxidation hair dyes as well, especially for the shading of brilliant (mainly red) fashion colors. However, they need to be stable to the oxidative conditions during use, as well to the oxidants required to stabilize the respective formulation.

Recently some classes of cationic azo-dyes have been found to be also stable under these conditions. These dyes impart brilliance and shine and are mainly used in fashion shades achieving simultaneously high lift [132].

5.2.2. Dyeing with Inorganic Compounds

Permanent hair dyes also include metal salt dyes, which produce a finely dispersed metal deposit on the hair. Metal oxides or sulfides may also be formed; this process can be promoted by the addition of sulfur compounds [7, p. 240]. A metallic gray, brown, or black coloration is obtained depending on the concentration and type of metal ion used.

Metal salts can be used in two ways. First, pretreatment with a trihydroxybenzene compound such as pyrogallol [*87-66-1*] is followed by treatment with an ammoniacal silver salt solution. This allows rapid dyeing due to the formation of metallic silver and oxidation products of the trihydroxybenzene derivative. If pretreatment is performed with thiosulfate instead of a benzene derivative, the process yields unstable silver thiosulfate and finally black silver sulfide.

In the second method, a metal salt solution (silver, lead, or bismuth; less often nickel, cobalt, or manganese) is applied; colloidal sulfur may also be added. "Progressive" coloration is obtained; i.e., the color intensity increases gradually with each treatment. Dyeing is based on (1) the reaction of the metal salts with the added sulfur and the sulfur in the hair keratin, which yields metal sulfides, and (2) the deposition of finely divided metals or metal oxides.

Metal salt dyes are no longer important in professional hairdressing, for three reasons:

1) Their use involves toxicological problems.
2) In permanent-wave neutralization, bleaching, or oxidation dyeing, metal salts catalyze the decomposition of hydrogen peroxide used as oxidant. When the temperature rises, the liberated oxygen can make the hair brittle.
3) The selection of colors is limited, and the shades look metallic and unnatural.

In some countries, certain oxidation dyes may not be used on the lashes and eyebrows, and the above disadvantages of metal salt dyes are of no consequence. In this area, dyes based on silver nitrate are still used [133]. Dyes based on lead and bismuth salts are sold in small quantities as "color restorers."

5.2.3. Dyeing with Direct Dyes

With direct dyes, as opposed to oxidation dyes, the actual dye (and not its precursor) is applied to the hair. It imparts a semipermanent or temporary color that lasts for a variable time. Lightening is not possible.

The first historically known hair dyes were semipermanent ones. The ancient Egyptians and Romans dyed hair and fingernails with henna, which contains lawsone [*83-72-7*] (2-hydroxy-1,4-naphthoquinone), a red-orange dye; walnut shells, containing juglone [*481-39-0*] (5-hydroxy-1,4-naphthoquinone), which gives a yellow-brown color; and indigo [*482-89-3*]. They also employed combinations or mordants with metal salts. To a slight extent, hair is still dyed with henna, juglone, indigo, and extracts such as chamomile (containing apigenin, 4′,5,7-trihydroxyflavone [*520-36-5*], which gives a yellow color). The use of these products is limited, however, because of poor selection of shades, uneven coloring, and laborious method of application [7, pp. 236 – 241].

Nearly all the chromophore systems common in dye chemistry (nitro, azo, anthraquinone, triphenylmethane, and azomethine) are currently used [35, pp. 526 – 533], [131]. Disperse dyes, cationic (basic) dyes, and anionic (acidic) dyes are employed.

Nitro Dyes. Nitro dyes are the most important class of direct hair dyes; they are substituted derivatives of nitrobenzene or nitrodiphenylamine (→ Nitro and Nitroso Dyes). By proper selection of donor groups and substitution site on the benzene ring, a spectrum of dyes

from yellow to blue violet can be prepared [7, pp. 247 – 250], [134] (see Fig. 6).

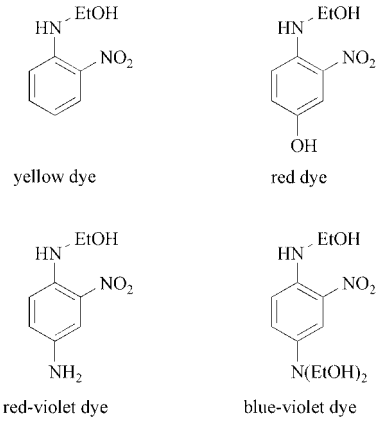

Figure 6. Influence of substituents on the color of nitro dyes

Nitro dyes are special because the small dye molecules penetrate the hair and impart color throughout the hair or in an annular pattern [135]. In this way, intense colors can be obtained even in hair with a large proportion of white. The dyes are not, however, fixed in the hair by size enlargement or ionic bonds (as in oxidation dyes), and so they slowly wash out.

The physical and chemical properties of dyes are very important for even coloring. First a hair dye should have comparable affinities for the roots, damaged areas, and tips. Second, the combined yellow-to-blue dyes used for shading must have similar properties so that color shifts will not take place when the hair is washed, for example. The importance of these problems is illustrated by the number of patent applications disclosing new, "custom-tailored" dyes [129] and suitable dye mixtures [136]. The relatively good colorfastness and stability of some nitro dyes allow them to be used in oxidation hair dyes as well, especially for the shading of brilliant (mainly red) fashion colors.

Cationic (basic) dyes such as methyl violet [*8004-87-3*] or methylene blue [*61-73-4*] are used in rinses; cationic azine or azo dyes are also used in tints. Their molecular size and charge prevent them from penetrating the hair, so that the color pattern in cross section is annular [118]. The charge on the dyes, however, leads to a relatively stable ionic bonding of the dye to acid groups of the hair, which lasts through several washings.

Anionic (acidic) dyes, usually azo dyes, are employed only in special cases. Because the skin is more basic than the hair, these dyes are adsorbed preferentially by the skin, so that contact with the dyeing product leads to severe scalp staining. For this reason, the mainly anionic food colorings are not employed as hair dyes. Attempts to mitigate these disadvantages by using appropriate vehicles (see Section 5.2.4.) have not been successful.

Disperse dyes of the azo and anthraquinone types, on the other hand, are used in hair tints. They also give annular color patterns, which are relatively durable because of the poor solubility of the dyes in aqueous systems. They can be combined with other classes of dyes.

Other Dyes. Metal complex dyes are of minor importance (→ Metal-Complex Dyes). They can be used as such or can be formed on the hair, although the metal salt treatment entails problems (see Section 5.2.2). Brilliant fashion colors (pink, green, etc.) are mostly obtained with anionic dyes, that are often food colorants. Other methods for dye formation in situ [137] or with reactive dyes have not been accepted because of toxicological concerns.

5.2.4. Product Forms

The dyes discussed in Sections 5.2.1, 5.2.2, and 5.2.3 can be used in a variety of coloring products [138] and forms. Many proposed formulations for the preparations discussed below can be found in [108, pp. 526 – 577, 649 – 676] and a more application-related review is given in [139].

Color rinses represent the simplest way of altering hair color. The hair is rinsed with a dilute aqueous or aqueous – alcoholic dye solution. The dyes are generally cationic and are adsorbed by the hair surface; they can mask the yellow color that results from bleaching or, given the proper selection of shade, can impart an attractive accent to the hair. Because of the short contact time with the hair, only slight color changes are possible.

The same holds for *colored or tint setting lotions* (see Section 4.1) that contain added dyes

(usually cationic, disperse, and/or nitro dyes). When the setting lotion is applied, the hair is tinted slightly. The colored film-forming polymers remaining on the hair also contribute to color impression. Changes are limited to refreshing the color, masking individual white hairs, and producing color effects as well as silver or gray shades in white or graying hair.

More pronounced color changes are possible with *tints.* Tints are formulated with direct dyes; intense colors can be obtained, especially with nitro dyes. Contents of 10 – 50 % white hair can be masked, depending on the type of dye used. Appropriate shading can produce fashionable colors or extreme color effects, especially with bleached streaks.

Depending on the desired coloring effect, tints are left on the head for 3 – 20 min. Therefore, thickened products are generally employed. In *foam tints,* a surfactant solution is dispensed as a foam from an aerosol. The foam is distributed easily and does not run off the head. Tints can also be thickened with cellulose derivatives, natural mucilage, or synthetic polymers. Concentrated solutions with intense coloring action can be obtained by using cosolvents (e.g., alcohols and ethylene glycol ethers) and vehicles (e.g., urea derivatives or benzyl alcohol). In *emulsion tints,* the dye base consists of an emulsion.

Permanent hair colors can be achieved with *tint shampoos.* The shampoo base is adjusted to an alkaline pH and contains oxidation dye intermediates. Before application, it is mixed with hydrogen peroxide or a hydrogen peroxide addition compound. In comparison with oxidation hair dyes, tint shampoos employ lower concentrations of base and oxidant. This suppresses the simultaneous bleaching process that occurs during dyeing (Section 5.2.1). As a result, damage to the keratin in hair is diminished, but the uniform coloring action is lost.

Early *oxidation hair dyes* were used in solution form; these have been replaced by cream- or gel-based formulas. The oil-in-water emulsions commonly used can be supplemented with auxiliary ingredients, such as polymers to improve combing ability, as well as other conditioning additives. Extensive patent literature is available on this point [129]. Gel formulations may be based on alcoholic solutions of nonionic surfactants or fatty acid alkanolamide solutions, which form a gel when mixed with the oxidant. The type (emulsion or gel) and the basic composition of the preparation strongly influence dyeing [140]. Different base formulations with the same dye content yield varying color depths and shading due to the distribution of the dye between the different phases of the product, interaction with surfactants, and diffusion from the product into the hair.

For use, the preparation is mixed with the oxidant and applied with a brush or squeeze applicator to the hair, first to the roots and then to the shafts. After a total contact time of ca. 30 min at room temperature, the hair is rinsed and washed, then treated as in bleaching.

In addition to cream and gel formulations, *powder hair dyes* are also marketed but to a lesser extent. Besides the dye intermediates, they contain thickeners and oxidants in powder form. They are stirred with water before use.

5.2.5. Testing Hair Dyes

New dyes and formulations must be tested for toxicological acceptability (see Chap. 10), and hair dyes must also meet a number of conditions related to their end use.

Color can be assessed by colorimetry [141]. The limits of precision are set by the substrate on which the measurement is performed. Studies on test subjects are difficult because of the uneven natural hair color and the background color (i.e., of the scalp). Tresses are hard to prepare at a constant quality level. Measurements on wool cloth give reproducible results, but for oxidation dyes the shades are not identical to those produced on hair. Colorimetric methods are therefore useful only for comparative measurements on the same object, for example, in lightfastness tests. Because hair must be redyed after four to six weeks due to growth, the fastness required of hair dyes is generally less than that needed for textiles. However, stability is still a problem with many indo dyes (see Section 5.2.1). Some of them "fade" even in the absence of light; others exhibit poor light- [142], sweat-, or acidfastness [143]. Results on the lightfastness of nitro dyes appear in [144]. The washfastness of dyes is also important [130, 145], as is wear resistance, especially for colored tints and setting lotions.

5.2.6. Dye-Removal Preparations

The complete removal of oxidation hair colors is a problem that has not yet been solved. Older, more intense dyes, in particular, cannot be removed entirely. Special reductive dye-removal products, mostly based on sulfites, reduce indo dyes to the leuco bases (diphenylamines), which are easier to remove by washing. The process has two deficiencies: it is not complete and the remaining diphenylamine is reoxidized. Oxidation treatment (bleaching) is, therefore, often used for the removal of dyes.

Recently mixtures of reducing agents (e.g., ascorbic acid, sulfites and/or cysteic acid) together with organic acids have been described [146], which may be used for the complete removal of a wide range of oxidation dyes.

Metal salt dyes cannot be removed with these preparations because of the catalytic decomposition of hydrogen peroxide.

Direct dye tints can be removed by washing with shampoo (the rate of removal depending on the type of dye) or, in some cases, by rubbing the hair with ethanol.

6. Waving

Besides cutting and coloring, perming is still one of the most important services at a hairdressing salon. Home permanent waving is not very frequent. Usership shows great geographical variance: the straighter the prevailing hair type, the higher the percentage of women using perms. While in the Latin countries in Europe this percentage is relatively low at ca. 40 %, it is significantly higher in central and northern Europe ($>$ 50 %), and in Japan, the biggest hairdressing market after the United States, it reaches almost 80 %. In 2004 the worldwide market volume totaled $811 \cdot 10^6$ €, in Germany, in comparison, it reached $72 \cdot 10^6$ €.

6.1. Principles and Methods

The principle of hair waving is closely related to the chemical structure of the hair. In waving, both intermolecular and intramolecular bonds are broken and then reformed after shaping. This process involves hydrogen bonds, and salt bridges, disulfide bonds, and hydrophobic and van der Waals interactions between the individual amino acids [7], [19][147 – 149].

Hair was originally waved with curling irons. Here, plastic deformation (heat, water) is based on hydrolytic cleavage of the disulfide bridges in the hair keratin. The London hairdresser K. L. Nessler first produced a "hot permanent wave" in 1906, using heated tongs, spiral curlers pointing away from the head, and an aqueous alkali solution. Lanthionine groups were created that permanently stabilized the shape of the hair keratin [150] (in lanthionine, the – S – S – bridge of cystine is replaced by an – S – bridge). In 1932, Clark and Speakman introduced sulfitolytic cleavage of the cystine bridges with alkaline sulfite solutions. When the solution is washed out, most of the cystine bridges are automatically restored. The mercaptan-based cold-wave preparations now in use were first described in a 1941 patent application by the New York firm of Sales Affiliates [151].

For *temporary waving* (setting) of hair without chemical agents, the keratin is softened by water absorption; the hair loses its original shape, chiefly through the dissolution of hydrogen bonds. If the wet, softened hair is put into a new shape (e.g., on rollers), this shape is retained after drying until the hair is again destabilized by moisture.

Permanent waving can only be achieved through the reduction of cystine; ca. 25 % of the cystine bridges are reduced to cysteine. After the hair has been reshaped, this reaction is reversed with an oxidant [152].

Waving preparations cause changes in the chemical and physical properties of hair. For example, some mercaptans used as reducing agents are bond to hair by thioacetylation [153]. The lanthionine formed during the reduction (softening) phase and the methionine of the hair are partly oxidized to sulfoxides and sulfones during the oxidation (hardening) step [150]. The ability of the hair to be reshaped is markedly diminished by repeated treatment.

The conventional cold-wave process proceeds as follows. The hair is washed and dried with a towel. It is then premoistened treated with a prelotion to equalize different hair structures; wound on curlers ca. 5 – 15 mm in diameter; thoroughly moistened with ca. 50 – 100 mL of the reducing *waving lotion;* left for ca. 10 –

30 min; and then thoroughly rinsed with water. Next, the hair is treated with a *neutralizer* (ca. 50 – 100 mL), which has an oxidizing action; this agent is allowed to act for ca. 5 – 10 min and the curlers are removed. If necessary, neutralization is continued for ca. 5 min more, with 20 – 50 mL of neutralizer. After the neutralizer has been rinsed out, a conditioning treatment (see Section 3.2) can be applied before the hair is finally styled [154].

6.2. Permanent-Wave Preparations

6.2.1. Waving Preparations

Reducing Agents. Since the late 1950s, most waving preparations have used salts of thioglycolic (mercaptoacetic) acid as the principal reducing agents. Glycerol monothioglycolate (GMT) and cysteamine (2-mercaptoethylamine) are still used in several countries in spite of certain toxicological problems. Reducing agents used less commonly are cysteine, 2-mercaptopropionic acid (thiolactic acid), and sulfite. 3-Mercaptopropionic acid, thioglycolic acid hydrazide, and thioglycerol are no longer used because of physiological problems.

Quality criteria for reducing agents are

1) Physiological tolerance (toxicity, sensitization, etc.),
2) Reducing action,
3) Shelf life,
4) Odor,
5) Purity of commercial grades, and
6) Availability and price

Thioglycolate [*68-11-1*] rates highly in all these areas and is, therefore, widely used as a reducing agent. To avoid hazard to skin and hair, the concentration of active ingredients is generally limited by law (see page 38).

Glycerol monothioglycolate [*30618-84-9*] is highly effective in all pH ranges, because of its low pK_a (SH) value. Its drawbacks are that the ester is readily saponifiable (special packaging is required), it has a somewhat greater risk of sensitization.

Cysteine [*52-90-4*] is a reducing agent with a weak waving action; a comparatively high pH (ca. 9 – 9.5) is needed to activate it. Cysteine is highly sensitive to oxidation, and cystine is produced during the waving process. Cystine is poorly soluble in water and forms a white precipitate that is difficult to eliminate. For ways of avoiding this, see [155].

2-Mercaptopropionic acid [*79-42-5*] is slightly less effective in waving than thioglycolic acid; it has a particularly unpleasant odor. *2-Mercaptoethylamine* [*60-23-1*] has similar waving properties to those of thioglycolic acid, but is more sensitive to oxidation and can cause an empyreumatic smell on the hair. Like cysteine, 2-mercaptoethylamine is expensive and is, therefore often employed merely as an additive [156, 157].

Sulfite [*7757-83-7*] (sodium or ammonium salts) does not have such an unpleasant odor as the mercaptans. In the useful pH range 6 – 8, however, added carriers are needed. Sulfite makes hair markedly harder to comb. It also lightens the hair slightly and, with oxidation-dyed hair, causes a relatively great loss of color. Various partial remedies have been proposed to counteract these drawbacks [157, 158].

Alkalinizing Additives. The content of reducing agent (ca. 5 – 11 wt % for thioglycolic acid) and the pH control the waving efficiency of a preparation. However, effectiveness is also strongly influenced by both the type and the quantity of base used for alkalinization [153]. Ammonia and monoethanolamine are most effective in promoting waving action; a combination of ammonium hydrogen carbonate and ammonium carbonate is also effective. Basic amino acids (e.g., lysine) counteract the waving action significantly. With a combination of urea and the enzyme urease, alkalinization increases during processing due to ammonia formation [159].

Other Additives.

Surfactants are included in waving preparations to ensure wettability, but they are also combined with cationic resins in waving lotions or neutralizers to improve wet combing [160]. In permanent-wave foams, the type and concentration of surfactant are especially critical for skin compatibility.

Fragrances are needed to mask the characteristic mercaptan odor. They are usually solubilized with nonionic surfactants (emulsifiers). The fragrances must be inert to the reducing agent used and must also be stable against alkali. A variety of substances have been suggested

for reducing the characteristic mercaptan odor [161].

Combability improvers are primarily cationic polymers, e.g., poly(dimethyldiallylammonium chloride) [*26062-79-3*], or cationic surfactants, e.g., cetyltrimethylammonium chloride [*112-02-7*].

Thickeners. Cellulose derivatives, polyacrylate salts, and other products are used for thickening to a gel consistency. Thickening generally lowers the availability of active ingredients and makes the preparation more difficult to process (oiliness, poor rinsability). Some degree of thickening is, however, often desirable to allow reliable measuring in use and to prevent dripping [162].

Thickened products in cream form are difficult to handle and have technical disadvantages similar to those of gels. They are used only in special preparations [163].

Opacifiers include polymer dispersions, e.g., styrene – vinylpyrollidone copolymers [*25086-29-7*].

Coloring agents give a pleasant appearance but can also serve as an identifying feature. They must be permitted by law and must be stable against reducing agents.

Carriers enhance the effectiveness of the product. Urea is used frequently, as are ethanol and 2-propanol. Other substances, most of them protected by patents, are known [164].

Complexing agents are added to products containing thioglycolates to prevent intensive red-violet coloring due to ferrous and ferric ions.

Finished Products. Waving preparations are sold as aqueous, clear, or opaque solutions, as formulations thickened to gels or creams, and also as aerosol foam permanent waves. For sample formulations, see [150, 165, 166]. The preparations differ chiefly in pH.

Acidic and neutral permanent-wave preparations (pH < 8) contain glycerol monothioglycolate as their most common active ingredient. Since the early 1970s, these products have become more important in Europe and especially in the United States. Because hair swells relatively little at about pH 7, these preparations are particularly gentle. They make hair easy to comb and leave it with a pleasant feel. The danger of overwaving is slight. Drawbacks are the slightly longer application time, an odor somewhat more unpleasant than that of mildly alkaline permanent-wave processes, a possible lack of durability of the wave, and the more expensive packaging (GMT must be packaged separately). However, because of an increasing number of allergic responses toward these acidic permanent-wave preparations since 1990, the consumption of these preparation dropped in the EU.

Mildly alkaline permanent-wave preparations (pH 8.0 – 9.0) are the most commonly used products. Alkalinization is achieved with bases such as ammonia and ammonium hydrogencarbonate or ammonium carbonate, or monoethanolamine.

The preparations are distinguished by good skin compatibility, strong and durable waving, simple packaging, and short application times. They are markedly less suitable for split hair than acidic preparations and, if allowed to act for too long, can easily produce overwaving.

Alkaline permanent-wave preparations are the oldest form of cold-wave preparations. They are alkalinized with ammonia (ca. 0.5 – 1.5 wt %) to a pH of ca. 9 – 9.5. Because of their strong odor (ammonia) and the frequent occurrence of skin inflammation, these products were largely superseded by mildly alkaline formulations in the mid-1960s.

Special preparations include thermally controlled, two-phase, enzyme-based, and foam permanent-wave preparations [154] [166].

Manufacture. Demineralized water is used in the manufacture of permanent-wave preparations. The production plant should be made largely of plastic or glass to prevent contamination of the product with iron. To prevent the buildup of nickel, the use of stainless steel should be restricted to the filling unit, because nickel allergies represent an increasing problem in the hairdressing trade.

Packaging. Preparations should be packaged in plastic materials with low oxygen permeabilities, such as poly(vinyl chloride) (PVC) or poly(ethylene terephthalate) (PET). The free air volume should be as small as possible because the vacuum created by the reaction of atmospheric oxygen can easily deform the packaging.

6.2.2. Neutralizers

The following commercially available oxidizing agents are used as neutralizers:

1) *hydrogen peroxide:* the concentration of commercial aqueous solutions is roughly 0.5 –3.0 wt % hydrogen peroxide, pH 2.0 – 4.5;
2) *potassium and sodium bromates:* the concentration of commercial aqueous solutions is roughly 6 – 12 wt %, pH 6.0 – 8.5;

The most widely used preparations are those containing hydrogen peroxide. Hydrogen peroxide has the advantages of physiological safety (it is degraded by catalases), excellent environmental compatibility, low solution concentrations, and good availability. However, metal salt impurities (iron, copper, and manganese ions) catalyze the decomposition of hydrogen peroxide. Even when such ions are absent or are complexed, hydrogen peroxide solutions still must be stabilized. Inorganic phosphates, salicylic acid, and 4-acetaminophenol are used for this purpose.

Bromate-based neutralizers are employed mainly in Asia because they do not greatly lighten dark hair. At present, aerosol foam neutralizers are also based on bromate because no suitable coating is yet available for protecting aerosol cans against hydrogen peroxide formulations. Bromate has the advantage of being insensitive to metal salt impurities.

Auxiliary ingredients in neutralizers must often be stable not only to oxidizing agents but also to acids. This fact is the primary restriction on the selection of coloring agents and fragrances.

Application Forms. The most common form is the "foam neutralizer," which is applied to the curlers with a sponge and lathered into a foam. Thickening to a stable gel is much more difficult with peroxide-based preparations than with the waving products, because the cellulose derivatives and other polymers are decomposed by oxidation. For ways of overcoming this problem, see [167].

Packaging. Bromate-based products can be packaged in a number of plastics such as poly(vinyl chloride), polyethylene, and polypropylene, whereas only polyethylene is suitable for products containing hydrogen peroxide.

6.2.3. Notes on Use

The prerequisite for a good wave is proper assessment of hair quality; this provides the basis for choosing a preparation and selecting the application time. However, assessment is not simple because even hair that has not been chemically treated exhibits progressive environmental damage from root to tip (see Section 2.2.1). Furthermore, hair on top of the head is subject to more severe weathering (sunlight, humidity) than neck hair; its porosity is thus higher. Chemical treatment (permanent waving, dyeing, or bleaching) also renders the hair more porous [168].

As damage to hair increases, the strength of the preparation used must be reduced and the application time shortened. By supplying heat (drying hood, 50 – 55 °C), the application time can be shortened by 30 – 60 %. The desired wave is controlled by the number and diameter of rollers used. The durability of the wave is generally better in strong, unfrayed hair than in frayed or thin hair.

7. Hair Straightening Preparations

Preparations similar to those employed for permanent waving can be used to straighten naturally wavy or curly hair. However, hair straightening products are usually gels or creams rather than lotions. They contain strong bases such as alkali hydroxides (ca. 1.5 – 4.0 wt % LiOH, NaOH) and other active ingredients such as guanidinium hydroxide [169] or tetraalkylammonium hydroxides [170].

Lye-based products have a stronger action and do not require oxidative posttreatment. However, improper use can easily make the hair brittle, and skin burns may even occur. To reduce these dangers, formulations have a high content of mineral oil.

Thickened preparations are applied to the hair with a brush and distributed through the hair with a comb. The hair is straightened by further combing. The preparations are allowed to work for a predetermined time (ca. 2 – 15 min) and then thoroughly rinsed out. Products containing reducing agents are subsequently treated with a neutralizer, which is completely rinsed out after

ca. 10 min. For sample formulations, see [36, pp. 590 – 591], [171, pp. 581 – 586], [172].

8. Hair Removal

The removal of hair has been practiced for several millenia. Quicklime, gum resin, hot ash, and scrapers were formerly used to eliminate unwanted hair from the arms, legs, armpits, and face. A wider variety of more pleasant methods is now available.

8.1. Epilation and Permanent Hair Removal

Epilation denotes hair removal by physical methods. The following methods are common:

1) Cutting the hairs with a razor, knife, or scissors: hair is quickly eliminated for a period of about two days.
2) Abrading the hairs with pumice or emery paper: this rather painful method is often preceded by an oxidation treatment to weaken the hairs.
3) Pulling the hairs with tweezers or an adhesive: only small amounts of hair can be removed with tweezers. For larger areas, adhesives such as warm or cold wax preparations are employed. The hairs are entangled in an adhesive mass (e.g., resin-wax mixtures) that is attached to a flexible backing. A quick pull tears out the hairs. However, hair grows back after this treatment.

Permanent hair removal is possible by electrolytic destruction of the hair papillae with a d.c. epilating needle or by thermal coagulation of the protein in the hair follicle (diathermy).

8.2. Depilation

Depilatories are chemically active preparations for hair removal. They usually contain a reducing agent and a strong base; both reductive and hydrolytic cleavage of the hair keratin take place between pH 11 and 12.5. Because of the high alkalinity, damage to the horny layer of the skin may occur. The principal chemicals used formerly were alkali and alkaline-earth sulfides, strontium sulfide being the most suitable. Because hydrolysis of sulfides yields the unpleasantly smelling hydrogen sulfide, these agents have been replaced almost completely by thiol salts. The lithium, calcium, or strontium salts of thioglycolic acid or thiolactic acid, in combination with calcium or strontium hydroxide as an alkalinizing agent, are currently found in almost every commercial product. Preparations containing stannite, titanate, and keratinase are often mentioned in the literature [171, pp. 142 – 150] but have no practical importance .

Thickening and moisturizing ingredients include alginates, cellulose derivatives, poly(vinyl alcohol), casein, sorbitol, glycerol, and propylene glycol. Carbonates, zinc oxide, clay, mineral oil, talc, and other substances are employed as fillers. The most widely used form is cream; aerosol foams are also commercially available. Suitable cream bases include stearyl alcohol, cetostearyl alcohol and oleocetyl alcohol ethoxylate. α-Bisabolol, allantoin, and lanolin derivatives are added to protect the skin. A layer of cream about 1 mm thick is applied to the area to be depilated with a spatula. After ca. 5 – 10 min, the spatula is again used to remove the cream and the hair from the skin surface. After the skin has been rinsed thoroughly, posttreatment with vaseline or a slightly acidic cream is recommended.

For further information and sample formulations, see [36, pp. 591 – 597], [153, pp. 449 – 450], [171, pp. 142 – 155], [172].

8.3. Shaving Aids

In wet shaving, *preshave preparations* carry out the following functions:

1) lubrication to allow easier sliding of the razor blade;
2) swelling of the keratin to lower the resistance of hairs to cutting; and
3) desensitization of oversensitive skin.

Many forms are available, ranging from simple shaving creams (superfatted oil-in-water emulsion) to shaving foams, soaps, and products that dispense heated shaving foam from a two-chamber pressurized package. For methods used to test effectiveness, see [153, pp. 447 – 448].

In dry or electric shaving, swelling of the hairs is not desired; hence, for example, high-volume-percent alcohol solutions are applied to

dewater the skin and stiffen the hairs. Isopropyl myristate is a suitable lubricant additive.

Aftershave preparations often contain 40 – 60 vol % alcohol and have pH values between 5 and 7. Lactic or citric acid is used to adjust the pH, and astringent aluminum salts are often added. Skin-care components include polyglycols, ureas, and polyalcohols.

For sample formulations and further information, see [36, pp. 613 – 626], [153, pp. 447 – 449], [165, pp. 381 – 384, 771 – 781], [171, pp. 156 – 189], [172].

9. Testing Methods

The interactions and effects produced by the application of various hair preparations are determined by *half-head testing* on the hair of experimental subjects. Preparation A is applied to one side of the head and preparation B to the other. Specially trained testers, usually hairdressers, then assess which of the two preparations is more effective in producing the desired hair-care effect(s).

Testing criteria include combability, feel, color, luster, manageability during hairstyling, curliness, durability (hold) of waving, setting ability, buildup of product residues, formation of dust on combing, and static charging of dry hair. The testing of hair colorings is described in Section 5.2.5. Long-term tests are used to determine whether the quality of the hair structure is maintained, improves, or deteriorates after repeated use. If the hair has been damaged, it resists styling; has a dull, rough feel; and lacks luster. Naturally, the judgment of the testers is subjective, but it can be rendered more objective by using appropriate statistical methods [173]. Uncertainties in the ranking of a product occur when hair quality changes gradually, climatic conditions vary, and only few numbers of subjects are involved in the tests.

Several of the above criteria can be evaluated reliably by instrumental measurements. These have the advantage that mixed, uniform batches of hair (in the form of tresses) can be used and several hair preparations can thus be compared directly. Testing conditions are defined, and numerical values obtained, which yield statistically significant results more readily than subjective testing. Unfortunately, internationally valid testing standards have not yet been adopted for quantifying the action and efficacy of hair-care products (as has been done for wool in the International Wool Textile Organization, IWTO, standard, (see also → Wool).

Combability testing usually measures the force required to untangle a randomly tangled hair bundle. The result depends strongly on the friction between the hairs and between the hair and the comb. The force can be measured with a "dynamometric comb" [7], a tensile testing apparatus [174], or a combing machine with rotating combs where the ambient conditions are set automatically [175].

Alternatively, the sound produced by combing hair can be recorded and compared against a scale; this method is used in testing shampoos, setting lotions, and waving products [7, p. 314]. All methods give relative values, which depend heavily on the type and preparation of the hair and on how closely test conditions imitate reality.

The *hold of temporarily set curls* after the use of setting lotions or hair spray is tested with the *curl-retention test* [19, p. 465]. In this test a weight is attached to the curled hair and allowed to hang for a predetermined time. The lengths of the hair before and after weighting are then compared. Differences in the setting effect can best be evaluated for well-defined geometries of the hair and roller at all values of atmospheric humidity [176]. Setting "factors" can thus be calculated [177].

The *elasticity and hold of permanent-wave curls* depend highly on the degree to which neutralization restores the hair to its initial chemical state and its physicochemical equilibrium. Hold is impaired by frequent contact with water and by mechanical deformation in brushing. In the curl-retention test for permanent-waves, curls are best weighted over a long period (24 h) at a high relative humidity (90 % R.H.) or in water containing surfactants. A dynamic elasticity test also simulates practical conditions [7, 9]. Chemical testing with 5,5′-dithiobis-2-nitrobenzoic acid (Ellman's reagent) indicates whether additional thiol groups remain on the waved hair [178] and whether the permanent wave was neutralized sufficiently.

Most hair preparations are designed to reduce the static charge of dry hair and impart tactile and visual "body," natural luster, and a smooth

feel. *Static charging* can be produced by repeatedly combing hair strands with a hard-rubber comb [179]. The discharge current measured in a Faraday cage indicates whether the cationic polymers present in most setting lotions, conditioning treatments, and shampoos adhere too tightly to the hair, causing it to acquire a charge.

A compression test on tresses can show how much *body* fine hair gains, for example, after permanent waving or treatment with a hair-care product [180]. The test measures the force required to pull treated hair through a series of increasingly narrow apertures.

Objective measurement of the luster or feel of hair strands is difficult. Very few methods yield any correlation with practical experience. Image analysis represents a new approach to the measurement of reflected light, i.e., luster [7]. More recently, shine measurements in good correlation with consumer assessment have been reported for evaluations that use hair tresses rather than single fibers [181]. The feel of hair may possibly be determined by the Kawabata system, which is used for testing textiles [182].

Study of the *chemical changes* in hair proteins produced by oxidative or reductive treatment is reviewed in [41]. Cysteic acid determinations [183], streaming potential measurements [184], spectroscopic methods,[41] and scanning force microscopy [185] provide information about the condition of the hair surface, whereas strength testing of hair bundles, swelling – sorption measurements, and solubility experiments yield information on the condition of the entire hair sample [19]. Optical methods, such as scanning electron microscopy of the hair surface [186] or the determination of hair birefringence under the optical microscope [187], provide visual evidence of the degree of damage.

10. Toxicology and Legal Aspects

10.1. Toxicology

10.1.1. Introduction

Cosmetic products are used extensively by most humans, at least in the economically developed countries. Many hair cosmetics are used daily or frequently over long periods of time. Consequently, consumer exposure to cosmetic ingredients may be considerable, so that current legislation, which defines far reaching safety requirements for cosmetic products, appears to be justified. Safety requirements from a regulatory point of view and consumer expectations are adequately summarized by the corresponding text within the EU Cosmetics Directive: "A cosmetic product put on the market within the community must not cause damage to human health when applied under normal or reasonably foreseeable conditions of use ..." [188].

The history of cosmetology, past experience with the safety of cosmetic products, and in particular statistics published by intoxication advisory centers show that the cosmetic industry in general has met these criteria [189]. This is achieved by continuously evolving safety assessment strategies which take into account the actual state of the art in safety or toxicity testing.

10.1.2. Strategy for Safety Assessment of Cosmetics

A complete safety assessment of cosmetic products has to cover the aspects microbiology, physicochemical characteristics, and human safety with regard to occupational and consumer exposure. This section exclusively deals with safety with regard to occupational and/or consumer exposure.

Safety assessment of hair preparations is conducted in two highly integrated steps at the ingredient and finished-product levels. In general, the potential of a hair preparation for sensitisation, genotoxicity and all other systemic toxic effects is evaluated on the basis of the ingredients' characteristics. Adequate consideration of human exposure is already of fundamental importance for data interpretation at the ingredient level. Safety assessment for each individual ingredient has to cover the different partial aspects of toxicology, such as acute and subchronic systemic toxicity, local effects, sensitizing potential, mutagenicity, tumorigenicity and teratogenicity [190]. Whether experimental data on all these aspects are necessary depends on the expected external and systemic exposure that will result from use of the hair preparation under consideration. Knowledge about the conditions of

human use enables the safety assessor to investigate and judge the possible route of entry into the human organism, e.g., by percutaneous absorption, oral ingestion or absorption via the respiratory tract. The main source of toxicological data on ingredients are the manufacturers' documents (raw material suppliers). Because these companies generally have to comply with national or supranational legislation on chemicals, they are obliged to generate toxicological data [191]. Additional sources are publications in the scientific literature, monographs released by health authorities or other official bodies, expert judgement based on analogy to chemically related substances, and in-house experience from the marketing of similar products containing the particular ingredient. In almost all cases, only a small range of experimental studies will remain to be done on the finished product. These include skin compatibility studies in humans under relevant test conditions [192]. The question of possible interactions between different ingredients will usually be evaluated on the basis of experience, and only in rare cases will additional experiments be required. Depending on the nature of unresolved problems after completion of this assessment, additional experiments may be initiated, in vitro or, very rarely, on animals or human volunteers.

An adequate safety assessment of hair preparations can not be performed by means of a formalized procedure such as working through different steps on a checklist. Instead, flexibility and expertise cosmetic chemistry and toxicology are necessary. The four main elements of the health related in part of safety assessment are

1) Exposure assessment
2) The toxicological profile of ingredients
3) The test results obtained with the finished product
4) Post-marketing (human) experience with analogous products

For each type of hair preparation key areas which are of major importance during safety assessment can be defined, e.g., the question of inhalative toxicity and respiratory sensitization in the evaluation of aerosol products [193]. In particular, permanent-wave and hair-color preparations usually contain chemicals that are more reactive than the majority of basic cosmetic ingredients. These active ingredients can be expected to interact with biological systems, and thus complete toxicological characterization is an essential part of the early screening phase. In contrast, rinse-off products such as shampoos and conditioners are usually not very challenging for a safety assessor in view of the nature of the ingredients and the conditions of use (exposure). Slight adverse local effects which can be observed in particularly susceptible individuals are mainly caused by surfactants [194].

As regards exposed humans, the safety assessor has to differentiate between the two main user profiles, consumer and professional, whereby the latter may differ in exposure route and duration. Finally the available toxicological or safety data must also be used to assess possible risks of foreseeable excess use, realistically conceivable misuse, and use by consumers suffering from certain diseases.

10.1.3. Test Methods

10.1.3.1. Basic Principles

In contrast to legislation on safety testing of chemical substances such as pesticides or pharmaceuticals, no detailed guidelines on the test programme and individual test methods have been issued for cosmetic ingredients. However, the fact that the toxicological profile of the ingredients has to be considered for safety assessment of cosmetics is clearly stated in the 6th Amendment to the EU Cosmetics Directive [188]. Moreover, a draft guideline prepared by the Scientific Committee on Cosmetology (advisory panel to the European Commission), intended for publication in 1997, clearly defines the range of toxicological aspects to be covered [195]. Because toxicology is a universal science, requirements for cosmetic ingredients do not fundamentally differ from those for chemicals intended for other areas of use. The toxicological profile of a chemical substance or a mixture of chemicals consists of information about [190, 195]:

– Acute systemic toxicity
– Local compatibility or local irritation (ski, mucous membranes, eye)
– Sensitizing potential (allergenicity)
– Mutagenic/genotoxic potential

- Phototoxicity (only under certain conditions)
- Absorption, distribution, metabolism and excretion (toxicokinetics)
- Subchronic and chronic systemic toxicity
- Reproductive and developmental toxicity
- Tumorigenic potential

Depending on the use of a chemical (exposure characteristics), emphasis will be placed on only a few of the above areas. It is, however, not acceptable to disregard completely one or more of these aspects. This does not mean that all questions must be answered by using standardized toxicological test methods. Instead, all scientifically relevant information is valuable, including theoretical considerations, structure – activity relationships, and results from unconventional experiments, provided they are interpreted on the basis of adequate scientific expertise. With regard to international regulatory acceptance of test data, the study protocols should adhere as closely as possible to standardized test guidelines, as published by the OECD [196] or comparable official bodies.

The cosmetics legislation in the European Union (see Section 10.2. Legal Aspects) is putting particular emphasis on the use of non-animal (alternative) test methods for the purpose of toxicological characterization of cosmetic products and cosmetic ingredients. Scientists working in this field should therefore be informed about the progress achieved in development of alternative methods. Depending on the status of scientific and regulatory acceptance of a given alternative method, its use should be carefully and seriously considered.

10.1.3.2. Systemic Toxicity Studies

Systemic toxicity studies determine whether a test compound can damage the individual organs and tissues with respect to structure or function. All possible routes of entry into the organism have to be considered, and these depend on the physicochemical properties of the test compound and the use pattern of the finished product. Apart from the dermal (percutaneous) absorption, ingestion and inhalation must be considered. A basic test usually performed at the beginning of the experimental program is the acute toxicity study. This experiment should enable the intoxication pattern after a single exposure to very high amounts of the test compound to be assessed. Traditionally, the results of such studies are summarized as LD_{50} values (→ Toxicology). The LD_{50} value is used for classification and labeling. The toxicologist, however, makes use of all details of the study results, which include the type and severity of intoxication symptoms, the time of onset and duration of clinical symptoms, and type and severity of organ damage [197 – 199]. Consideration of this complete range of study results allows an improved occupational and consumer risk assessment to be made.

There are various reasons for considering the possible consequences of a repeated, continuous, and long-term exposure of humans to cosmetic ingredients. Hair preparations used for cleansing, conditioning, and styling are usually applied frequently (daily) and for long periods of time. If such products contain biologically reactive ingredients that can penetrate the skin, then the question of systemic toxicity must be further investigated. These conditions are mimicked by subchronic and chronic toxicity studies, in which the test compound is administered in graded dose levels over an adequate period of time to a group of animals [200 – 202] (→ Toxicology). Test compounds that appear unproblematic following extremely high single doses in acute toxicity studies can exert massive biological effects after repeated administration of relatively low dosages. Such profiles are known for natural biologically active substances and synthetic chemicals such as pharmaceuticals. In addition, the possibility of accumulation in the body might lead to toxic effects even after minimal exposure over long periods of time. Thus, subchronic and chronic toxicity studies provide information about the threshold of exposure beyond which the induction of toxic effects must be expected, the target organs of toxicity, and the nature of adverse health effects. The results are used to decide whether the estimated exposure of humans to the tested ingredient resulting from use of a given hair preparation is safe.

Replacement of these rabbit primary irritation test methods by non-animal alternatives has been the top priority for the expert scientists in the cosmetic industry. Although there is still no validated and officially recognized stand-alone method available, experienced toxicologists can make use of a large variety of different in vitro

tools for assessment of local irritation of chemicals and mixtures of chemicals. Whereas artificial skin models seem to be the most promising approach for skin compatibility testing, ex vivo assays such as the bovine corneal opacity test, the chicken enucleated eye test, the isolated rabbit eye test or the hens egg test on the Chorio-Allantois membrane are currently considered to give the most reliable results. Such tools can be particularly useful for comparative testing of chemicals or complex preparations in case that there is already broad experience with irritation data from structurally related chemicals or similar preparations. An adequate use of these in vitro methods does definitely require a large in-house database and specific experimental expertise and cannot be recommended as a standard approach.

10.1.3.3. Local Compatibility

Local compatibility studies (also known as local irritation studies) assess whether a test compound can induce damage directly at the sites of contact with the human body. For hair preparations these are mainly the skin, eyes, and in certain cases also the upper respiratory tract mucosa. Local compatibility results obtained with undiluted compounds are of limited relevance to the prediction of local compatibility of finished hair preparations. Therefore, dermatological testing of final formulations is indispensable. Studies with undiluted compounds are nevertheless helpful in selecting of suitable ingredients during the early phase of product development. For ingredients known to be strong irritants as pure compounds, the irritation threshold should also be determined experimentally.

Workplace related dermatoses are among the most frequent occupational diseases in industrialized countries. The prevalence of dermatoses is also increasing in the general population, an observation which is associated by many experts with the increased exposure to a growing number of chemicals. Of the available experimental methods, the dermal compatibility study in the rabbit (primary irritation, skin) is the test method with the broadest scientific and regulatory acceptance [203]. The rabbit is considered particularly sensitive towards chemically induced skin irritation.

Depending on the type of hair preparation, contact with mucous membranes (predominantly the eye) is either unavoidable or must be expected from inadvertent contamination. The classical test method for evaluating the consequences of such contact is the (primary) eye irritation study in the rabbit (Draize Test, → Toxicology, Chap. 4.4) [204]. Again, this test model is known to be particularly sensitive towards chemical injury and offers additional tissues such as the cornea and iris as objects of examination apart from the conjunctiva (mucous membrane) [205]. The determination of the concentration – response relationship for local effects is of essential importance because all cosmetic ingredients are used in the form of a complex formulation composed of different substances.

10.1.3.4. Sensitizing (Contact-Allergenic) Potential

Allergy is a disease characterized by a massive (over)reaction of the organism to the contact with a chemical. Usually, the allergenic response occurs only after a repeated contact. The process which changes the reaction of the organism from normal to hypersensitive is termed sensitization. There are vast differences in potency of sensitizing effects between different chemicals, as is true for the individual susceptibility of humans towards the sensitizing effects of a given chemical. Therefore, the allergenic effects of a chemical usually affect only few individuals from a given large group of humans. The severity of the allergenic reaction is not linearly correlated with the dose of the allergen. Often only traces of a chemical are sufficient to induce an allergic reaction [206].

Cosmetic ingredients (and chemicals in general) cannot simply be categorized as sensitizers or nonsensitizers. In principle, the statement is justified that for any chemical at least one human could be found who develops an allergy. Animal studies are used to differentiate between cosmetic ingredients without relevant sensitizing potential and those whose sensitizing potential warrants a warning text on the label or use restrictions. Cosmetic ingredients are usually tested on the guinea pig to assess their sensitizing potential following dermal contact (→

Toxicology, Chap. 4.5). The guinea pig is the test species of choice for this type of experiment because it is highly susceptible towards the allergenic effects of chemicals. The Magnuson – Kligman maximization test is one of the most reliable methods of the various test protocols. The maximization test is able to also identify chemicals which are regarded as weak human sensitizers. In contrast, the Bühler test has a lower sensitivity but is increasingly used for characterizing cosmetic ingredients. It uses exclusively dermal administration of the test compound, which mimics the use conditions of cosmetic products better than the intradermal injection of the maximization test [207].

In the 1990s, the local lymph node assay (LLNA) in the mouse has been qualified as a stand-alone alternative to the guinea pig tests. Following rigorous validation efforts, the LLNA has ultimately found broad scientific and regulatory acceptance as a test method providing both, improvements in scientific value and animal welfare (lower number of animals; less suffering). The scientific basis for the LLNA is the measurement of the incorporation of ^{3}H-methyl thymidine into lymphocytes in draining lymph nodes of mice topically exposed to the test substance. The endpoint of interest is a stimulation index giving the ratio of thymidine incorporation in lymph nodes from dosed animals compared to control animals. The test is considered to indicate a relevant sensitizing potential when the stimulation index exceeds 3 ($SI > 3$). In contrast to the classical guinea pig tests, the LLNA results are providing a direct measurement of sensitizing potency of the tested chemical.

Research on a number of in vitro approaches for assessment of the sensitizing potential is currently being conducted. So far, the most advanced methods under consideration have reached the prevalidation stage and thus are not sufficiently mature for stand-alone use in human safety assessment.

10.1.3.5. Mutagenic/Genotoxic Potential

Mutations are alterations of the genetic information inside living cells which occur spontaneously or are induced by radiation, chemicals or viruses. Depending on the type and severity of alterations in the genetic material three types of mutations are distinguished: gene, chromosome, and genome mutations [208, 209].

Gene mutations (point mutations) are molecular changes of DNA not extending beyond one gene and therefore not affecting the microscopic aspect of the chromosomes. Chromosome mutations are microscopically detectable structural changes to chromosomes. The term genome mutations is used for mutagenic effects which induce numerical changes in the genome, such as missing or supernumerary single chromosomes or doubled, tripled, or multiplied sets of chromosomes.

Mutations may affect the genetic information of gonadal cells within the testes or ovaries or that of somatic cells (cells of any other tissue or organ of the body). Mutations in gonadal cells may lead to impaired fertility, to embryonic or perinatal death. Nonlethal gonadal mutations usually lead to inherited and heritable (genetically fixed) diseases. Although the consequences of somatic mutations are limited to a small part of the entire organism, their occurrence deserves particular attention because they are associated with the development of tumors and ageing phenomena. The currently discussed model of tumor development assumes definitely that mutations in somatic cells contribute fundamentally to the process of tumorigenicity in mammals [210].

The following test systems are superior in terms of practicability, standardization and validation and are therefore preferentially used for mutagenicity testing of all types of chemicals. They are all in vivo and in vitro tests in somatic cells. Test methods for the detection of mutations in gonadal cells are scientifically and technically demanding, time consuming, and expensive and are therefore mainly used in special cases.

The Salmonella microsome mutagenicity test, also known as Ames Test (→ Toxicology, Chap. 4.9.1.1) is the most widely used mutagenicity assay [211].

The Ames test is suitable for the detection of point mutagenic effects of chemicals with relatively low expense in terms of time and resources. However, a positive result only means that the test compound is a mutagen in bacteria under test conditions. This also holds true for other in vitro test systems, so that a positive result will lead to further investigations.

For the detection of gene mutations in vitro in mammalian cells cell lines originating from the Chinese hamster or the mouse are generally used (→ Toxicology, Chap. 4.9.2.2) [212]. In addition to a number of in vitro assays [213], the mouse micronucleus test (→ Toxicology, Chap. 4.9.5.3) is routinely used for the indirect detection of a clastogenic (chromosome-breaking) effect. The micronucleus test is also suitable for the detection of spindle poisons, which can cause genome mutations. The micronucleus test is widely accepted as a rapid in vivo mutagenicity test [214].

In human risk assessment, the extrapolation of the results from a complete mutagenicity test battery is problematic. If there is convincing evidence from the test results that the test compound had no mutagenic effects, it can be concluded that the probability of a mutagenic risk in humans is acceptably low. If the results clearly indicate interaction of the test compound with genetic mechanisms, they must be interpreted as evidence for an increased genotoxic and tumorigenic risk in humans. Animal studies that allow quantitative assessment of human genotoxic risk are expensive and require large numbers of animals. Their use is therefore only justified under particular circumstances, e.g., in the case of chemicals for which no suitable alternatives are available.

10.1.3.6. Reproductive Toxicology

Reproductive toxicology deals with possible effects of chemicals on reproduction, which can be manifested as impairment of fertility or of the development of the progeny. These studies are not mandatory for all cosmetic ingredients and are conducted only if a relevant systemic exposure is expected. Reasons for increased concern in terms of reproductive toxicology include:

- Structural similarity with known embryotoxic or fertility impairing compounds
- Indications of a very high biological/toxicological activity from other studies
- Extensive exposure, extended contact time and/or high percutaneous absorption rates

Throughout the course of the reproductive cycle there are many different targets for the adverse effects of chemicals. Considering the life span of one generation, these targets could be maturation of sperm or ova female sexual cycle, fertilization, implantation, prenatal development, delivery, and early postnatal development of the offspring. Reproductive toxicology is investigated with the aid of animal experiments. The rat is the predominantly used species, but mice and rabbits are also frequently employed in this type of experiment [215].

Comprehensive characterization of the reproductive toxicological potential of a test compound can be achieved by conducting multigeneration studies (→ Toxicology, Chap. 4.7.4) [216].

Embryotoxicity studies assess impairment of prenatal development. For this purpose pregnant female animals (usually rats) are dosed with a test compound throughout the most sensitive phase of prenatal development [217 – 219] (→ Toxicology, Chap. 4.7.2).

All subsequent phases of the reproductive cycle are covered by peri- and postnatal studies (→ Toxicology).

Developmental and reproductive effects are regarded as threshold phenomena. Therefore, human risk assessment is performed by comparing the expected exposure to the compound with the maximum dose tested in the various animal experiments that did not induce adverse effects (no observed effect level, NOEL). If there is a relevant risk of reproductive or developmental effects under human use conditions, then the test compound is not suitable as a cosmetic ingredient because contra-indications or use restrictions are not practicable for hair-care products.

10.1.3.7. Tumorigenic Potential

Chemicals which induce or enhance the growth of tumors are termed tumorigens or carcinogens. Although only malignant tumors are called cancer, the term cancerogenicity is often used as a synonym for the correct term tumorigenicity.

For evaluation of the tumorigenic potential of a cosmetic ingredient, the complete range of toxicological information is considered. The results of mutagenicity studies are of particular importance because positive results from such experiments are also indications of a tumorigenic potential. This suspicion might be further supported by results from subchronic or chronic

systemic toxicity studies (e.g., proliferation of growth of certain tissues) or by structural analogy to known tumorigenic compounds. If there is some evidence of tumorigenicity at this stage, special tumorigenicity studies in animals might be considered. Because these studies are very time consuming, expensive, and require large numbers of animals, they will only be conducted on very important ingredients for which no alternatives are available. The animal experiments for detection of a tumorigenic potential are usually performed in rats and/or mice and include the continuous daily administration of the test compound over a substantial part of the life span of the test animals, which is usually set at two years for laboratory rodents. In addition to the continuously performed in-life observations, which include palpation to detect tumors at the earliest possible time, a comprehensive program of macroscopic and histopathological (microscopic) examinations for the post-mortem evaluation of organs and tissues is an essential part of standard tumorigenicity studies [220].

For tumorigenic risk assessment, human experience is of particular importance in addition to experimental data. Depending on the quality of such epidemiological data, problematic results from animal toxicity studies might even be outweighed. There is general agreement among cosmetic toxicologists from industry and public health agencies that chemicals with a relevant tumorigenic potential are not suitable as cosmetic ingredients.

10.1.3.8. Toxicokinetics

Toxicokinetics describes the time curve of a chemical's concentration within different compartments of the organism. Toxicokinetics includes all processes which determine the fate of a chemical in a living organism, such as absorption, distribution, biotransformation (metabolism), and excretion.

For hair preparations the most important toxicokinetic information concerns the question of percutaneous absorption. Information about this parameter enables the externally applied amount of a hair preparation (dermal dose) to be transformed into a value representing the systemically available amount of a given ingredient of this preparation (internal dose). The internal dose is the relevant basis for all risk related considerations of systemic toxicity. Percutaneous absorption of ingredients is investigated with the aid of in vitro tests [221, 222]. The most widely accepted method utilizes skin from pigs, rats, or even humans. The skin membranes are mounted on cylindrical chambers with a cross-sectional area of ca. 10 cm^2. The test formulation is applied to the outer surface of the skin membrane, and the portion of the test compound that penetrates is collected in a receptor fluid contacting the inner surface. Provided that the application conditions (type of formulation, concentration of the test compound, and area of dose, and exposure time) mimic the human use situation adequately, the results of such in vitro experiments give a reliable estimate of the absorbed portion which has to be expected for human use of the hair preparation. Whereas the penetration rates through rat skin usually overestimate absorption through human skin, the results obtained with pig skin and human skin in vitro predict reliably the human in vivo situation [221 – 223].

For spray products which deliver inhalable aerosols, inhalative exposure must be considered additionally. This parameter can be determined by animal experiments, e.g., inhalative toxicity studies (→ Toxicology, Chap. 4.2.3) [224]. However, if the physical characteristics of the aerosol, in particular the concentration of toxicologically relevant chemicals and particle or droplet size distribution are well-known, then a theoretical estimation of the systemic exposure can give reliable values because a very large portion of the inhaled material is usually absorbed in the alveolar parts of the lung due to the large contact area.

10.1.4. Description of the Safety Assessment Process

In principle, safety assessment already starts with the first steps of product development. The cosmetic chemist selects the ingredients from a pool of chemicals generally suitable for the intended cosmetic use. In this respect, suitability means for example that the ingredients are compatible with the cosmetics legislation in the countries intended for marketing (positive/negative lists appended to the EU cosmetics directive) and are nonmutagenic and non-

tumorigenic. Product innovations are often be achieved by introducing new chemicals which must be characterized toxicologically according to the principles described in Section 10.1.3.

In the succeeding phase, the safety assessor considers human exposure. Apart from the exact qualitative and quantitative composition of the product, detailed knowledge of the anticipated human use conditions is required, such as the amount used per application, frequency of use, contact time with the skin (rinse-off or nonrinse product), the size of the contact area, possible routes of exposure (dermal, oral, inhalation), the probability of eye contamination (viscosity of the preparation) and the target consumer group (males, females, children). In addition, special consideration will be given to the exposure assessment for professional users of hair preparations in salons.

The information about toxicological profiles of ingredients and exposure scenario are integrated to reveal possible areas of concern. In principle, the contribution of each individual ingredient to local and systemic tolerance will be considered. Many routinely used cosmetic ingredients can induce dermal irritation if applied undiluted. Because chemical skin irritation is a concentration-dependent phenomenon, an irritation threshold can be defined. Comparison of the irritation threshold with the use concentration may already adequately answer the safety question. However, the issue is complicated by the fact that many cosmetic ingredients interact with each other resulting in increased or decreased skin irritation. Final evaluation therefore requires adequate experience with products of similar composition or experimental results obtained with mixtures of ingredients. The risk of contact allergenic effects is initially judged by considering the sensitizing potential of each individual ingredient at its concentration in the formulation. If the available experimental data and human experience appear to be irrelevant because exposure conditions differed from those anticipated for the product under review, then additional investigations are required.

Interactions among ingredients which enhance or reduce dermal penetration could also influence the sensitizing potential. Likewise, the percutaneously absorbed amount of each individual ingredient is required to assess possible systemic (toxic) effects. The integration of information about external exposure (dermal dose), corrected by an absorption factor to yield the internal or systemic exposure, possible influences of other components of the preparation and the toxicological profile of the ingredients as observed in animal studies including threshold values (NOELs), dose – response curves, and target organs of toxicity, are usually included in the calculation of a safety factor. The arithmetic procedure which leads to a safety factor or safety margin is common to toxicological risk assessments for all types of chemicals. The basic principle underlying following formula has also been recommended by the Scientific Committee on Cosmetology [195] for quantitative risk assessment with regard to systemic effects of cosmetic ingredients (e.g., applicable for hair preparations of the nonaerosol type):

$$D_{ext} \cdot C_{ingr} \cdot A_{syst} \cdot F_{CT} \cdot \frac{1}{60\text{kg} \cdot \text{NOEL}} = \text{safety margin}$$

D_{ext} total amount used per day (mg) in the case of hair preparations corrected by the "distribution factor scalp/hair" of 0.1

C_{ingr} concentration of the particular ingredient in the hair preparation

A_{syst} absorption factor (dermal)

F_{CT} correction factor for short contact time (e.g., 0.1 for rinse-off products)

NOEL highest dose tested in subchronic or chronic systemic toxicity studies or in reproduction toxicity studies without any indication of adverse effects

For the protection of consumers from "classical" systemic toxic effects such as organ toxicity and impairment of physiological functions (cardiovascular system, kidneys, intestinal tract) or the endocrine system, a safety factor of 100 is usually considered to be the minimum requirement. Depending on the nature of toxic effects observed in animal experiments beyond the NOEL, larger or smaller safety margins might be necessary or acceptable. Regulators have to act in a conservative manner and therefore they tend to request safety margins beyond 100 if the first toxicological manifestations concern prenatal development (embryotoxicity), the nervous system, or the immune system.

In addition, possible chemical reactions such as neutralization, hydrolytic cleavage, complex formation, and others, which could be expected

due to the reactivity of the ingredients must be carefully evaluated with regard to their secondary biological consequences. For the majority of newly developed hair preparations, possible interactions among ingredients can be predicted theoretically and from experience with analogous mixtures of compounds and products. However, additional experiments must be conducted in those cases where information gaps are left. Commonly, the predicted safety of new hair preparations is confirmed by appropriate tests on human skin, and pre-market controlled studies for other purposes such as product performance can also yield valuable safety information.

The question of consumer safety is not only a function of product composition but depends in addition on the use conditions. Thus, the use instructions must be taken into account by the safety assessor and in certain cases they may have to be supplemented by special warning texts such as "avoid contamination of the eye because the product might cause irritation".

Following the integration of available human experience as the last step in the safety assessment process, the final safety clearance will enable the progress of the new hair preparation to the marketplace.

10.2. Legal Aspects

10.2.1. General

10.2.1.1. Legal Aspects in the European Union

In the European Union (EU) hair preparations are classified as cosmetics and subject to legal provisions stated in the EC Cosmetics Directive 76/768/EEC [225]. (The implementation of the Directive into national cosmetic regulations by each member state is being updated continuously.) In principle, the responsibility for cosmetic products lies with the manufacturer. Postmarket surveys as well as occasional inspections at the manufacturers' facilities by the competent authorities — instead of premarket product registration — ensure that the products comply with the relevant regulations. As a basic requirement cosmetics ". . . must not cause damage to human health when applied under normal or reasonably foreseeable conditions of use . . ." (Article 2 of the EC Cosmetics Directive). A safety assessment by a qualified person must be available for each finished product.

For cosmetic colorants, preservatives and UV filters no substances other than those listed in Annexes IV, VI, and VII of the Directive may be used. Any new substance with one of these functions has to be registered with the Scientific Committee Cosmetology of the European Commission before being used in cosmetics. Other ingredients which are neither banned (Annex II) nor restricted (Annex III) may be used without limitations provided Article 2 is followed.

Each cosmetic product introduced to the EU market has to be provided with an ingredients label that uses the common names listed in the respective EU inventory [226] based on the International Nomenclature of Cosmetic Ingredients (INCI). New raw materials are subject to the requirements of the Dangerous Substances Directive [227] in addition to the Cosmetics Directive.

Some EU member states, however, depart from the EC Cosmetics Directive regarding restrictions on certain raw materials (Denmark [228], Italy [229]), and labeling requirements (France [230]). These deviations from the EC Cosmetics Directive complicate free trade within the EU. France [231], Germany [232, 233], and Spain [234] require notification to the poison information centers before a product is marketed. A common system based on framework formulations is foreseen for all EU member states.

10.2.1.2. Legal Aspects in Other Countries

In the *United States*, certain cosmetics are classified as over-the-counter (OTC) drugs (i.e., for sale without prescription) and are subject to registration [235]. The list of cosmetic color additives is strictly limited, and each production batch has to be certified by FDA. In preparations sold directly to the consumer all ingredients have to be listed on the label [236].

In *Japan*, cosmetics are classified as quasi-drugs or cosmetics. New raw materials have to be registered and extensive toxicological data submitted. All raw materials have to comply with official standard specifications [237 – 245]. Human patch tests from other countries are recog-

nized only if data are obtained from Asian test persons. Until 1986, all preparations had to be registered. Since then, Comprehensive Licensing Standards by Category (CLS) have gradually been introduced for cosmetics [246]. Notification to the health authorities is sufficient for products that comply with the CLS. Other cosmetics and all quasi-drugs have to be registered with the health authorities. As announced by the Japanese Ministry of Health and Welfare in spring 1997, a deregulation of the Pharmaceutical Law is planned to simplify the procedures for cosmetics in the near future.

The cosmetic regulations of *other countries* are often influenced by those of the United States or the EU. For example, provisions in Norway [247] and Turkey [248] are closely modeled on the EC Cosmetics Directive. The countries of Central and Eastern Europe are in the process of harmonizing their cosmetic regulations with the EC Cosmetics Directive. Premarket product registration is required for cosmetics in Turkey, most countries of Central and Eastern Europe, Latin America, and Southeast Asia.

10.2.2. Individual Hair Preparations

Shampoos The general provisions stated in Section 10.2.1 apply in the EU countries. Not all substances allowed in the EU may be used in other countries; provisions in specific countries must be complied with. Antidandruff shampoos are classified as OTC drugs in the United States [249] and as quasi-drugs in Japan [239, 242].

Hair-Bleaching Preparations In the *EU*, peroxosulfates can be used without restrictions. Some countries (e.g., *Thailand* [250]) limit their concentration. The *United States* does not have any specific regulations. In *Japan*, hair-bleaching preparations are classified as hair dyes of quasi-drugs [242, 251].

Hair Dyes

European Union. The substances *o*-phenylenediamine, 4-methyl-*m*-phenylenediamine, 1-methoxy-2,4-diaminobenzene, 1-methoxy-2,5-diaminobenzene, 2-amino-4-nitrophenol, 2-amino-5-nitrophenol, 4-ethoxy-*m*-phenylenediamine, catechol, pyrogallol, 4-amino-2-nitrophenol and 2-methyl-*m*-phenylenediamine are banned in the EU (EC Cosmetics Directive , Annex II) [225].

Other phenylenediamines and methylphenylenediamines, *N*-substituted derivatives of phenylenediamines and methylphenylenediamines, diaminophenols, hydroquinone, α-naphthol and resorcinol are permitted, with restrictions on concentration. Except for α-naphthol, these substances may be employed for dyeing eyelashes and eyebrows by professionals only. Warning labels are also prescribed (EC Cosmetics Directive, Annex III, Part 1) [225]. Other appropriate substances may be used as long as they are neither banned (Annex II) nor restricted (Annex III) and the finished product complies with Article 2 of the EC Cosmetics Directive [225]. In addition to the EU provisions, 2-nitro-*p*-phenylenediamine and 4-nitro-*o*-phenylenediamine are banned in Italy [229], and 2-nitro-*p*-phenylenediamine, 4-nitro-*o*-phenylenediamine and *m*-phenylenediamine in Denmark [228].

United States. The United States requires a special warning label. With this label, any hair dye can be used without restriction [252].

Japan. In Japan oxidation hair dyes are classified as quasi-drugs [242]. A list of permissible hair dyes has been issued [251]. New dyes have to be registered with the Ministry of Health and Welfare (MHW) before being used in products [239].

Permanent Wave, Hair-Straightening and Hair-Removal Preparations. The *EC Cosmetics Directive*, Annex III, Part 1, restricts the pH range of permanent wave preparations and hair straighteners based on thioglycolic acid, its salts, and esters (pH 7 – 9.5 for the acid and salts, 6 – 9.5 for esters). The maximum concentration (calculated as thioglycolic acid) is 8 wt % for general use and 11 wt % for professional use. A special warning label is required. Hair-removal preparations are also subject to restrictions and require special warning labels; thioglycolate esters are not, however, allowed [225].

In the *United States* no legal restrictions exist. However, the Cosmetic Ingredient Review (CIR) Expert Panel considers 15.4 wt % max. of thioglycolic acid as safe in permanent-wave products [253]. The CIR recommendation is respected by the manufacturers of permanent waves. The FDA recommendations for hair straighteners are given in [254].

In *Japan*, permanent wave products are classified as quasi-drugs [242]. Thioglycolic acid and its salts are permitted, as is cysteine. Restrictions on concentration and pH must be taken into account. Thioglycolate esters are not allowed. Concentration limits also apply to neutralizers [255].

Aerosols. Here the term aerosol refers to hair cosmetics filled in pressurized containers, such as hair sprays. The use of chlorofluorocarbons is banned in all countries that are following the Montreal protocol on the protection of the ozone layer in any kind of hair preparations.

In the *EU*, methylene chloride is permitted as a solvent up to a concentration of 35 wt %. (EC Cosmetics Directive, Annex III, Part 1) [225]. Special warning labels are required by Commission Directive 94/1/EC [256]. In *Austria*, pressurized containers with > 45 wt % or > 250 g of flammable ingredients require specific storage conditions [257]. General requirements for aerosols in the *United States* are given in [258]. Based on the Federal Clean Air Act the content of volatile organic compounds (VOC) is restricted in several states, such as California, New York, etc. [259]. In *Japan* the internal pressure of aerosols is restricted to 8 kg/cm^2 at 30 °C. [260]. For specific labeling requirements, see [261].

Other Hair Care Products. In the EU, some preservatives may be added in high concentration provided they do not act as preservatives but have another function (e.g., conditioning or antidandruff agent) which must be apparent from the presentation of the preparation (EC Cosmetics Directive, Annex VI, Preamble) [225]. When products containing alcohol (ethanol) are exported, the alcohol-denaturation provisions of the country of destination have to be followed. In the United States antidandruff preparations are regarded as OTC drugs [249]. Specific restrictions on volatile organic compounds (VOC) for certain product groups are stated in [259]. In Japan all antidandruff preparations are classified as quasi-drugs [239, 242].

11. References

1. C. E. Orfanos (ed.): *Haar und Haarkrankheiten*, G. Fischer Verlag, Stuttgart 1979.
2. A. Chatt, S. A. Katz: *Hair Analysis*, VCH, Weinheim 1988.
3. R. C. Marshall, J. M. Gillespie, *J. Forensic Sci. Soc.* **25** (1985) 377.
4. P. Kintz: *Drug Testing in Hair*, CRC Press, Boca Raton 1996.
5. F. Noser, *ärztl. Kosmetol.* **13** (1983) 415 – 422.
6. F. Noser, M. Rothe: *Hair Growth*, Wella AG, Darmstadt 2004.
7. C. Bouillon, J. Wilkinson: *The Science of Hair Care*, 2nd Ed., CRC Press, Taylor & Francis Group; Boca Raton, London 2005.
8. R. Röthlisberger: *Kopfschuppen, ein kosmetisches Problem*, Wella AG, Darmstadt 1985.
9. P. Jollès, H. Zahn, H. Höcker (eds.): *Formation and Structure of Human Hair*, Birkhäuser Verlag, Basel 1997.
10. B. C. Powell, G. E. Rogers in [9], pp. 59 – 148.
11. R. D. B. Fraser, T. P. MacRae, T. E. Rogers in I. N. Kugelmass (ed.): *Keratins*, C. C. Thomas Verlag, Springfield, Illinois 1972.
12. J. A. Swift: Fundamentals of Human Hair Science, *Cosm. Sci. Monogr. I*, Micelle Press, Weymouth, Dorset, England 1997.
13. P. W. Wertz in [9], pp. 227 – 238.
14. A. Schwan-Jonczyk, P. Kassenbeck, H. Zahn: *Hair Structure*, Wella AG, Darmstadt 2000.
15. H. Zahn, *Lenzinger Ber.* **42** (1977) 19 – 34.
16. S. A. Wainwright, W. D. Briggs, J. D. Currey, J. M. Gosline: *Mechanical Design in Organisms*, Princeton University Press, Princeton, New Jersey 1982.
17. H. Meichelbeck, *Riechst. Aromen Kosmet.* **31** (1981) 6 – 11.
18. A. G. Matoltsy, *Adv. Biol. Skin* **9** (1967) 559 – 569.
19. C. R. Robbins: *Chemical and Physical Behaviour of Human Hair*, 4th ed., Springer Verlag, Berlin-Heidelberg 2002.
20. G. Wortmann et al., Proceed. 13th Int. Hair Symposium HAIRS 05, Fulda 2005.
21. D. A. D. Parry in [9], p. 177.
22. A. S. Inglis, I. H. Leaver, F. G. Lennox, 3rd *Int. Wool Text. Res. Conf. Proc.*, (1965) 121.
23. J. A. Maclaren, B. Milligan: *Wool Science*, Science Press, Marrickville, Australia 1981.
24. C. E. Orfanos, W. Montagna, G. Stüttgen (eds.): *Hair Research*, Springer Verlag, Berlin 1981.

25. K. Jimbow, T. Fitzpatrick, W. C. Quevedo in J. Bereiter-Hahn, A. G. Matoltsy, K. S. Richards (eds.): *Biology of the Integument II*, Springer Verlag, Berlin (1986).
26. G. Prota: *Melanins and Melanogenesis*, Academic Press, New York 1992.
27. R. A. Nicolaus: *Melanins*, Verlag Hermann, Paris 1968.
28. M. Giesen: *Beiträge zur Kenntnis des Melanins in Keratinfasern*, Dissertation, Technische Hochschule Aachen, 1981.
29. I. Castanet, J. P. Ortonne in [9], pp. 209 – 226.
30. K. Brown, A. Mayer, B. Murphy, T. Schultz, L. Wolfram, *Proc. Ann. Meet. Am. Chem. Soc.* (1987).
31. C. Fox, *Cosmet. Toiletries* **100** (1985) no. 3, 31, 32, 34, 36, 38, 41, 42, 44 – 46.
32. R. Matissek: *MvP-Ber,* vol. 3, Max von Pettenkofer Inst., Bundesgesundheitsamt, Berlin 1979.
33. A. L. L. Hunting: *Encyclopedia of Shampoo Ingredients*, Micelle Press, Cranford, London, 1983, 1985.
34. M. S. Balsam, E. Sagarin (eds.): *Cosmetics. Science and Technology,* 2nd ed., vol. 2, Wiley-Interscience, New York 1972.
35. J. B. Wilkinson, R. J. Moore: *Harry's Cosmeticology*, 7th ed., George Godwin, London 1982.
36. K. Schrader: *Grundlagen und Rezepturen der Kosmetika*, Hüthig Verlag, Heidelberg 1979.
37. R. Y. Lochhead et al.: "A Brief Review of Polymer/Surfactant Interaction," *Cosmet. Toiletries* **119** (2004) no. 2, 37 – 46.
38. R. Y. Lochhead et al.: "Advances in Polymers for Hair Conditioning Shampoos", *Cosmet. Toiletries* **120** (2005) 569 – 578.
39. A. M. Kligman et al., *J. Soc. Cosmet. Chem.* **27** (1976) 111 – 139.
40. F. C. RoiaJr., R. W. Vanderwyk, *J. Soc. Cosmet. Chem.* **20** (1969) 113 – 134.
41. J. Jachowicz, *J. Soc. Cosmet. Chem.* **38** (1987) 263 – 286.
42. M. M. Breuer, G. X. Gikas, I. T. Smith, *Cosmet. Toiletries* **94** (1979) no. 4, 29 – 34.
43. H. Edelstein, *Cosmet. Toiletries* **100** (1985) no. 4, 31 – 36.
44. M. M. Breuer, *J. Soc. Cosmet. Chem.* **30** (1979) 41 – 64.
45. B. Idson, W. Lee, *Cosmet. Toiletries* **98** (1983) no. 10, 41 – 46.
46. A. L. L. Hunting, *Cosmet. Toiletries* **99** (1984) no. 6, 57 – 60.
47. J. Newton et al.: "Silicone Technology offers Novel Methods for Delivering Active Ingredients, *SÖFW-Journal* **130** (2004) no. 5, 8 – 13.
48. G. Allen, *Int. J. Cosmet. Sci.* **6** (1984) no. 2, 61 – 69.
49. C. Führer, H. Junginger, S. E. Friberg, *J. Soc. Cosmet. Chem.* **29** (1978) 703.
50. C. Müller-Goymann, *Seifen öle Fette Wachse* **110** (1984) no. 14, 395 – 400.
51. W. J. Benton, C. A. Miller, R. L. Wells, *JAOCS J. Am. Oil Chem. Soc.* **64** (1987) no. 3, 424.
52. P. Gross et al., E-PA 0 208 951, 1986.
53. Lebensmittelchemische Gesellschaft, Fachgruppe in der Gesellschaft Deutscher Chemiker, Arbeitsgruppe "kosmetische Mittel": *Datenblätter als Hilfsmittel*, Wirksamkeitsbewertung kosmetischer Wirkstoffe – Panthenol, Parfümerie und Kosmetik **80** (1999) no. 3, 42.
54. J. Vollhardt: "Native hydrophobic wheat proteins with intelligent hair care properties", *SÖFW-Journal* **12**5 (1999) no. 2/3, 2 – 9.
55. I. Hütter et al.: "Energie, Struktur und Hair Repair durch intelligente Proteine", *SÖFW-Journal* **127** (2001) no. 9, 53 – 58.
56. J. A. Swift et al.: "Investigations on the penetration of hydrolyzed wheat proteins into human hair by confocal laser-scanning fluorescence microscopy", *J. Cosmet. Sci.* **51** (2000), 193 – 203.
57. Wella, JP 3590628, 2002 (R. Bimczok, T. Kripp, B. Grasser, C. Springob).
58. Wella, US 6482808, 2002 (C. Springob et al.).
59. H. Gottschalck, G. Hohm, H. Kaminski, *Proc. – Int. Wolltextil Forschungskonf. 5th* (1975) 349 – 358.
60. W. Umbach: *Kosmetik und Hygiene*, 3rd ed., Wiley-VCH, Weinheim 2004, pp. 279 – 293.
61. J. S. Jellinek: *Kosmetologie, Zweck und Aufbau kosmetischer Präparate*, 3rd ed., Hüthig Verlag, Heidelberg 1976.
62. R. Lochhead, *Cosmet. Toiletries* **103** (1988) 23 – 26, 27 – 30, 32 – 34, 36 – 61.
63. Seifen Öle Fette Wachse **130** (2004) 33 – 35.
64. C. Wood, S. Nguyen-Kinu, P. Hoessel, *Cosmet. Toiletries* **119** (2004) 59 – 62, 64, 66.
65. S. Martiny, *Int. J. Cosmet. Sci.* **24** (2002) 125 – 134.
66. V. Schehlmann, *Seifen Öle Fette Wachse* **123** (1994) 649 – 650, 652, 655 – 656.
67. K. Seib, F. Vogel, *Seifen Öle Fette Wachse* **110** (1984) no. 17, 515 – 516.
68. D. Howard, S. D. Mole, *Seifen Öle Fette Wachse* **129** (2003) 12 – 14, 16 – 17.

69. J. Dallal, *Cosmet. Toiletries* **117** (2002) 33 – 34, 36 – 38, 40, 42.
70. G. Wypych: *Handbook of Solvents*, Chem. Tec. Publishing, Toronto 2001, pp. 881 – 882.
71. V. Signori, *Cosmetic Science and Technology Series* 28 (2005) 142 – 146.
72. E. Poppe, M. Baumschieper, T. Schroeder, *Seifen Öle Fette Wachse* **131** (2005) 38, 40, 42 – 43.
73. S. A. Vona *Jr., HAPPI* **39** (2002) 127 – 133.
74. DE 471 705, 1927;
DE 527 137, 1928.
75. T. Kitazawa, M. Eguchi, T. Ozawa, J. Yamada, *Cosmet. Perfum.* **89** (1974) no. 1, 33 – 42.
76. P. J. Petter, S. C. Johnson, *Seifen Öle Fette Wachse* **113** (1987) no. 13, 427 – 430.
77. P. D. Stagg, *Aerosol Rep.* **13** (1974) no. 2, 56 – 64.
78. M. J. Molina, F. S. Towland, *Nature (London)* **249** (1974) 810.
79. R. J. Murgatroyd, R. G. Derwent: *The Impact of CFC and Other Man-Made Pollutants on Stratospheric Ozone: A Resume*, Environment and Medical Sciences Division, AERE Harwell R10248 England, July1981, pp. 1 – 30.
80. *Causes and Effects of Changes in Stratospheric Ozone, Update* 1983, National Academy Press, Washington, D.C., 1984.
81. "Studies Negate Ozone Depletion Theory," *Aerosol Age*, **29** (1984) no. 4, 20 – 22.
82. Frigen Information, "Fluorkohlenwasserstoffe in der Atmosphäre, G.T. 03/84," *Aerosol Rep.* **24** (1985) no. 2, 82 – 85.
83. J. G. v. Schweinichen, *Aerosol Rep.* **23** (1984) no. 1, 25 – 31.
84. P. J. Petter, *Aerosol Rep.* **21** (1982) no. 9, 369, 372 – 375.
85. L. Bohnenn, *Aerosol Rep.* **23** (1984) no. 2, 66 – 73.
86. L. Bohnenn, *Aerosol Rep.* **22** (1983) no. 12, 594, 596 – 603.
87. A. Keenan, *Research Disclosure* **478** (2004) 118 – 119.
88. H. Fischman, *HAPPI* **38** (2001) 38.
89. R. Lochhead, *HAPPI* **27** (1990), 58, 60, 62, 64
90. Wella AG, EP 1197 201, 2002 (B. Stein, M. Franzke).
91. Wella AG, EP 02012 961, 2002 (H. Wendel, S. Birkel, S. Niesig, M. Franzke).
92. Wella AG, WO 93/17083, 1993 (G. Lang, D. Hoch, E. Konrad, W. Geibel, H. Wendel, T. Kripp).
93. Wella AG, EP 1302 193, 2002 (M. Hannich, M. Franzke).
94. Wella AG, EP 1321 12981, 2001 (T. Krause, A. Duchscherer, H. Henze).
95. Wella AG, EP 96 104 847, 1996 (J. Schmenger, B. Stein, W. Keil).
96. J. v. Gruber, *Cosmet. Sci. Technol. Series* **22** (1999), 217 – 274.
97. P. Hoessel, M. Pierobon, C. Wood, *Seifen Öle Fette Wachse* **131** (2005), 19 – 29.
98. F. Herrmann, H. Ippen, H. Schaefer, G. Stüttgen: *Biochemie der Haut*, Thieme Verlag, Stuttgart 1973, pp. 100 – 104.
99. S. S. Bleehen, M. A. Pathak, Y. Hori, T. B. Fitzpatrick, *J. Invest. Dermatol.* **50** (1968) 103.
100. C. R. Robbins, *J. Soc. Cosmet. Chem.* **22** (1971) 339 – 348.
101. I. J. Kaplin, A. Schwan, H. Zahn, *Cosmet. Toiletries* **97** (1982) 22 – 26.
102. J. Cegarra, J. Gacén, "The Bleaching of Wool with Hydrogen Peroxide," *Wool Sci. Rev.* **59** (1982) 1 – 44.
I. J. Kaplin, J. M. Marzinkowski, *J. Text. Inst* **74** (1983) 155 – 160.
H. Zahn, S. Hilterhaus, A. Strüßmann, *J. Soc. Cosmet. Chem.* **37** (1986) 159 – 175.
103. H. Zahn, *J. Soc. Cosmet. Chem.* **17** (1966) 687 – 701.
104. Clairol, US 4 792 341, 1986 (S. D. Kozikowski, J. Menkart, L. J. Wolfram).
105. Wella, EP 451 261, 1990 (R. E. Godfrey, T. Clausen, W. R. Balzer, G. Mahal, R. Rau).
106. C. Dubief, *Cosmet. Toiletries* **107** (1992) 95 – 102.
107. S. Kanetaka, K. Tomizawa, H. Iyo, Y. Nakamura in: *Preprints Platform Pres.*, Vol. 3, IFSCC Intern. Congress Yokohama 1992, Yokohama 1992, pp. 1059 – 1072.
108. K. Schrader: *Grundlagen und Rezepturen der Kosmetika*, Hüthig Verlag, Heidelberg 1979, pp. 575 – 577.
J. S. Jellinek: *Kosmetologie, Zweck und Aufbau kosmetischer Präparate*, 3rd ed., Hüthig Verlag, Heidelberg 1976, pp. 674 – 675.
109. J. Cegarra, J. Ribé, J. Gacén, *J. Soc. Dyers Colour.* **80** (1964) 123 – 129.
110. V. Böllert, L. Eckert, *J. Soc. Cosmet. Chem.* **19** (1968) 275 – 288.
111. Goldwell, EP-A 560 088, 1993 (H. Lorenz, F. Kufner).
Goldwell, EP 684 036, 1994 (H. Lorenz, W. Eberling).

112. L'Oréal, DE 2 023 922, 1970 (F. A. Vorsatz, A. F. Risch).
Wella, US 5 279 313, 1991 (T. Clausen, W. R. Balzer).
Wella, EP-A 650 719, 1994 (T. Clausen, W. R. Balzer, V. Port, J. Kujawa).
L'Oréal, EP 619 114, 1994 (J.-M. Millequand, C. Tricaud, A. Gaboriaud).
113. Scwarzkopf, DE 3 844 956, 1988 (T. Oelschläger, W. Wolff).
114. A. Bereck, H. Zahn, S. Schwarz, TPI *Text. Prax. Int.* **37** (1982) 621 – 629.
115. Shiseido KK, JP-Kokai 5 4129 – 134, 1978, (S. Kubo, I. Yamada).
Deutsches Wollforschungsinstitut an der RWTH Aachen, DE 3 149 978, 1981 (A. Bereck).
116. Wella, DE 1 158 213, 1962 (H. Wilmsmann, E. Konrad).
117. H. Monnet, FR 158 558, 1883.
118. H. Wilmsmann, *J. Soc. Cosmet. Chem.* **12** (1961) 490–500.
119. J. F. Corbett, *J. Soc. Cosmet Chem.* **30** (1979) 191 – 211 and references cited therein.
H. Husemeyer, *J. Soc. Cosmet. Chem.* **25** (1974) 131 – 138.
120. J. F. Corbett in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes*, vol. **5**, Academic Press, New York, London 1971, pp. 475 – 534.
121. Kyowa Hakko Kogyo Kabushika Kaisha, EP 310 675, 1988 (Y. Tsujino, Y. Yokoo, K. Sakato, H. Hagino).
122. Wella, DE 38 43 892, 1990 (T. Clausen, U. Kern, H. Neunhoeffer); Wella, DE 42 34 887, 1994 (H. Neunhoeffer, S. Gerstung, T. Clausen, W. R. Balzer); L'Oréal, FR 9 505 422, 1995 (L. Vidal, A. Burande, G. Malle, M. Hocquaux).
123. I. G. Farben, DE 559 725, 1930 (E. Lehmann).
124. L'Oreal, FR 2 362 116, 1976 (A. Bugaut, J.-J. Vandebossche).
125. Henkel, DE 2 852 156, 1978 (D. Rose, P. Busch, E. Lieske); DE-OS 3 235 615, 1982 (D. Rose, E. Lieske).
126. Henkel, EP 106 987, 1983 (N. Maak, P. Flemming, D. Schrader).
127. J. F. Corbett, *J. Soc. Dyers Colour.* **85** (1969) 71 – 73.
128. L'Oréal, US 2 934 396, 1958 (R. Charle, C. Pigerol).
129. J. F. Corbett, *Rev. Prog. Color. Relat. Top.* **4** (1973) 3 – 7;
15 (1985) 52 – 65.
130. J. C. Johnson: *Hair Dyes*, Noyes Data Corp., Park Ridge, N.J. 1973.
131. J. F. Corbett, *Cosmet. Toiletries* **106** (1991) 53 – 57.
132. Ciba-Geigy, EP 0 681 464, 1994 (P. Möckli); L'Oréal,
EP 0 810 851, 1995 (H. Samin).
133. F. E. Wall in M. S. Balsam, E. Sagarin (eds.): *Cosmetics, Science and Technology*, 2nd ed., vol. **2**, Wiley-Interscience, New York 1972, p. 328.
134. J. F. Corbett, *J. Soc. Dyers Colour.* **83** (1967) 273 – 276.
135. S. K. Han, Y. K. Kamath, H.-D. Weigmann, *J. Soc. Cosmet. Chem.* **36** (1985) 1 – 16.
G. Blankenburg, H. Philippen, *J. Soc. Cosmet. Chem.* **37** (1986) 59 – 71.
136. L'Oreal, DE 3 131 366, 1981 (R. de la Mettrie, P. Canivet).
137. Ciba-Geigy, DE 2 807 780, 1978 (A. Bühler, A. Fasciati, W. Hungerbühler).
138. P. Greß, D. Hoch, M. Schmock, D. Wanke: *Das Färben des Haares*, Wella AG, Darmstadt 1984.
139. D. F. Williams, W. H. Schmitt (eds.): *Chemistry and Technology of the Cosmetics and Toiletries Industry*, 2nd ed., Blackie Academic & Professional (Chapman and Hall), London 1996.
140. R. L. Goldemberg, H. H. Tucker, *J. Soc. Cosmet. Chem.* **19** (1968) 423 – 445.
K. Schrader, *Parfuem. Kosmet.* **63** (1982) 649 – 658.
141. M. den Beste, A. Moyer, *J. Soc. Cosmet. Chem.* **19** (1968) 595 – 609.
R. Feinland, W. Vaniotis, *Cosmet. Toiletries* **101** (1986) 63 – 66.
142. H. H. Tucker, *J. Soc. Cosmet. Chem.* **18** (1967) 609 – 628;
H. H. Tucker, *J. Soc. Cosmet. Chem.* **22** (1971) 379 – 398.
143. I. Schwartz, J. Kravitz, A. D'Angelo, *Cosmet. Toiletries* **94** (1979) 47 – 50.
144. J. F. Corbett, *J. Soc. Cosmet. Chem.* **35** (1984) 297 – 310.
145. M. Y. M. Wong, *J. Soc. Cosmet. Chem.* **23** (1972) 165 – 170.
146. Wella, DE 196 47 493 1998 (M. Kunz, D. LeCruer);
Wella, DE 196 47 494, 1998 (M. Kunz, D. LeCruer);
Clairol, US 5 982 933, 1998 (G. Wis-Surel, A. Mayer, I. Tsivim).
147. M. Feughelman: "Morphology and properties of hair," *Cosmet. Sci. Technol. Series* (1997) no. 17 (Hair and Hair Care), 1 – 12.

148. J. Yu, D. W. Yu, D. M. Checkla, I. M. Freedberg, A. P. Bertolino: "Human hair keratins," *J. Investigative Dermatol.* (1993), Jul. **101** (1 Suppl.), 56S – 59S.
149. S. Hilterhaus-Bong, H. Zahn: "Contributions to the chemistry of human hair: III. Protein chemical aspects of permanent waving treatments," *Int. J. Cosmet. Sci.* **11** (1989), no. 5, 221 – 31.
150. C. R. Robbins: *Chemical and Physical Behavior of Human Hair*, 4th ed., Springer Verlag N.Y., Berlin, Heidelberg 2002, p. 123 – 159.
151. Sales Affiliates New York, US 2 577 710, 1941 (E. McDonough).
E. J. Klemm, *Am. Perfum. Cosmet.* **83** (1968) 83 –87.
152. E., T. Borish: "Hair Waving," *Cosmet. Sci. Technol. Series* (1997), **17** (Hair and Hair Care), 167 – 190.
153. *Ullmann's*, 5th ed., **A12**, pp. 588 – 591.
154. P. Gress, H. Lehmberg, *Körperpflege*, vol. **2**, Dümmler Vlg., Bonn 1984, pp. 67 – 98.
155. Iwasakia: "Cysteine Waving Lotions: Past Improvements and Future Prospects," *Cosmet. Toiletries* **109** (1994) no. 3, 69 – 78.
156. KAO SOAP, JP 57 062 217, 1982 (T. Nemoto, Y. Ikeda, Y. Honma).
157. Wella, DE-OS 3 620 849, 1987 (Y. Greiche, P. Hartmann, J. Köhler, K. Rönsch).
158. C. R. Robbins: *Chemical and Physical Behavior of Human Hair*, 4th ed., Springer Verlag N.Y., Berlin, Heidelberg 2002, p. 122.
159. Wella, DE 19816072 A1, 1999 (C. Weitzel, G. Maresch, G. Lang, S. Glaser, B. Schmidt).
160. C. Fox, *Cosmet. Toiletries* **100** (1985) no. 4, 72 – 76.
161. Zofos Int., EP 610892 A1, 1994 (A. Nandaguri, B. H. Solka, J. A. Kocis).
162. Wella, GB 2 116 218, 1985 (T. Wajaroff, P. Hartmann).
163. Goldwell, DE-OS 3 401 221, 1985 (D. Heinz et al.).
164. Toporiina OY, WO 2000027349 A1, 2000 (H. Overmark-Lehto).
165. G. A. Nowak, *Die kosmetischen Präparate*, vol. **2**, Ziolkowsky Vlg., Augsburg 1984, pp. 544 – 576.
166. Wella, DE19860239, 2000 (G. Lang, T. Kripp, B. Grasser, G. Maresch).
167. Wella, WO 2005067874 A1, 2005 (T. Cassier, M. Lede).
168. Wella, EP1302124, 2003 (T. Cassier, B. Schreiber, A. Kalbfleisch, J. Allwohn).
169. Carson Prod., CA 1 117 423, 1982 (M. J. de la Guardia).
170. L'Oreal, US 2003/003033677 A1 2003 (N. V. Nguyen, D. W. Cannell).
171. J. B. Wilkinson, R. J. Moore: *Harry's Cosmeticology*, 7th ed., George Godwin, London 1982, pp. 555 – 580.
172. W. Umbach (ed.) *Kosmetische Mittel-Entwicklung, Herstellung und Anwendung kosmetischer Mittel,* Thieme Verlag, Stuttgart 1988.
173. *IFSCC Monograph*: "Principles of Product Evaluation: Objective Sensory Methods," vol. 1, 1987.
174. M. L. Garcia, J. Diaz, *J. Soc. Cosmet. Chem.* **27** (1976) 379 – 398.
175. P. Busch, *Seifen Öle Fette Wachse* **113** (1987) 627 – 632.
176. A. Schwan, personal communication.
177. P. Diaz, M. Y. M. Wong, *J. Soc. Cosmet. Chem.* **34** (1983) 205 – 212.
178. H. Meichelbeck, A. G. Hack, C. Sentler, *Z. Gesamte Textilind.* **70** (1968) 242.
179. G. Blankenburg, V. Heß, D. Teasdale, *Parfuem. Kosmet.* **62** (1981) 169 – 172.
180. C. R. Robbins, R. J. Crawford, *J. Soc. Cosmet. Chem.* **35** (1984) 369 – 377.
181. C. Reich, C. R. Robbins, *J. Soc. Cosmet. Chem.* **44** (1993) 221 – 234.
182. S. Kawabata in J. Hearle (ed.): *The Mechanics of Flexible Fibre Assemblies*, Sijthoff and Noordhof, Alphen and. Rhijn, Netherlands 1980, pp. 405 – 417.
183. C. M. Pande, *J. Soc. Cosmet. Chem.* **44** (1993) 257 – 268.
184. H. D. Cook, J. T. Smith, *Appl. Polym. Symp.* **18** (1971) 663 – 672.
185. P. Hössel, D. I. R. Sander, W. Schrepp, *Cosm. Toilletr.* **111** (1996) 57.
186. I. A. Swift, *Int. J. Cosmet. Sci.* **13** (1991) 143 – 159.
187. R. K. Curtis, D. R. Tyson, *J. Soc. Cosmet. Chem.* **27** (1976) 411 – 431.
188. Council Directive of 27 July 1976 on the approximation of the laws of the Member States relating to cosmetic products (76/768/EEC)
189. A. Hahn, H. Michalak, K. Noack, G. Heinemeyer, Ärztliche Mitteilungen bei Vergiftungen nach § 16e Chemikaliengesetz (Zeitraum 1990 – 1995) Zweiter Bericht der "Dokumentations- und Bewertungsstelle für Vergiftungen" im Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin

190. N. Loprieno, G. Loprieno, *In Vitro Toxicol.* **8** (1995) 3, 291 – 322.
191. Council Directive 92/32 of 30 April 1992 7th Amendment of Directive 67/548 EEC
192. Industrieverband Körperpflege- und Waschmittel (ed.): Guideline for Safety Testing and the Recording of Unwanted Effects of Cosmetic Products on Human Health, October 1995.
193. I. H. Draize et al., *Soap Chem. Spec.* **35** (1959) 91 – 94, 149 – 151.
194. F. Klaschka, *Ärztl. Kosmetol.* **16** (1986) 337 – 346.
195. Commission of the European Communities, 803/5/90, Consumer Policy Service Notes of guidance for testing of cosmetic ingredients for their safety evaluation.
196. OECD Guidelines for Testing of Chemicals Section 4: Health Effects.
197. OECD Guideline for Testing of Chemicals No. 401 "Acute oral toxicity".
198. Animals in Research Committee of The Society of Toxicology and SOT Council, "Comments on the LD_{50} and Acute Eye and Skin Irritation Tests," *Fundamental and Applied Toxicology* **13** (1989) 621 – 623.
199. J. Yam et al., "Comparasion of the up-and-down method and the fixed-dose procedure for acute oral toxicity testing," *FD Chem Toxic* **29** (1991) 4, 259 – 263.
200. OECD Guideline for Testing of Chemicals No. 408 "Subchronic Oral Toxicity — Rodent: 90-day Study".
201. WHO Publication: Environmental Health Criteria No. 6, "Principles and Methods for Evaluating the Toxicity of Chemicals," Part I, Geneva 1978.
202. OECD Guideline for Testing of Chemicals No. 407, "Repeated Dose Toxicity — Rodent: 28-day or 14-day Study".
203. OECD Guideline for Testing of Chemicals No. 404, "Acute Dermal Irritation/Corrosion".
204. OECD Guideline for Testing of Chemicals No. 405, "Acute Eye Irritation/Corrosion".
205. ECETOC Monograph No. 11, "Eye Irritation Testing," European Chemical Industry Ecology and Toxicology Centre, Brussels 1988.
206. ECETOC Monograph No. 14, "Skin Sensitization Testing," European Chemical Industry Ecology and Toxicology Centre, Brussels.
207. OECD Guideline for Testing of Chemicals No. 406, "Skin Sensitization".
208. J. Ashby, "The prospects of a simplified and internationally harmonized approach to the detection of possible human carcinogens and mutagens," *Mutagenesis* (1986) 1, 3 – 16.
209. F. H. Sobels, "Approaches to assessing genetic risks from exposure to chemicals," *Env. Health Perspectives* **Suppl. 101** (1993) 3, 327 – 332.
210. R. W. Tennant et al., "Prediction of chemical carcinogenicity in rodents from in vitro genetic toxicity assay," *Science* **236** (1987) 933 – 941.
211. B. N. Ames et al., *Mutation Res.* **31** (1975) 347 – 364.
212. OECD Guideline for Testing of Chemicals No. 476, "Genetic Toxicology: In vitro Mammalian Cell Gene Mutation Tests".
213. OECD Guideline for Testing of Chemicals No. 473, "Genetic Toxicology: In vitro Mammalian Cytogenetic Test".
214. OECD Guideline for Testing of Chemicals No. 474, "Genetic Toxicology: Micronucleus Test".
215. H.-A. Poggel, "Reproduktionstoxikologische Untersuchungen," *Klin. Pharmakol. 4. Auflage 1984*.
216. OECD Guideline for Testing of Chemicals No. 416, "Two-Generation Reproduction Toxicity Study".
217. J. G. Wilson: *Environment and Birth Defects*, Academic Press, New York and London, 1973.
218. J. G. Wilson, "The evolution of teratological testing," *Teratology* **20** (1979) 205 – 212.
219. OECD Guideline for Testing of Chemicals No. 414, "Teratogenicity".
220. OECD Guideline for Testing of Chemicals No. 451, "Carcinogenicity Studies".
221. ECETOC Monograph No. 20, "Percutaneous Absorption," 1993.
222. M. Balls et al., "Practical aspects of the validation of toxicity test procedures. The report and recommendations of ECVAM Workshop 5," *ATLA* **23** (1995) 129 – 147.
223. W. Diembeck et al., "Test Guidelines for In Vitro Assessment of Percutaneous Absorption and Penetration of Cosmetic Ingredients," *Regulatory (Food and Chemical Toxicology)*, in press.
224. OECD Guideline for Testing of Chemicals No. 413, "Subchronic Inhalation Toxicity: 90-day Study".
225. *Council Directive 76/768/EEC of 27 July 1976 on the Approximation of the Laws of the Member States Relating to Cosmetic Products*, Official Journal of the European Communities No. L 262, September 27, 1976, incl. *Twentyfirst Commission Directive 97/45/EC of*

14 June 1997 adapting to technical progress Annexes II, III, VI and VII to Council Directive 76/768/EEC on the Approximation of the Laws of the Member States Relating to Cosmetic Products, Official Journal of the European Communities No. L 196, July 24, 1997.
226. *96/335/EC* of May 8, 1996, Official Journal of the European Communities No. L 132, June 1, 1996.
227. *67/548/EEC* of June 27, 1967, Official Journal of the European Communities No. L 196, August 16, 1967, incl. *8th Amendment* 96/56/EC of September 9, 1996, Official Journal of the European Communities No. L 236, September 18, 1996
228. *Bekendtgørelse om Kosmetiske produkter*, Miljøministeriets bekendtgørelse no. 456, June 2, 1995.
229. *Decreto Ministeriale*, Nov. 24 1987, no. 530, Gazzetta Ufficiale no. 303, Supplemento Ordinario no. 3, December 30, 1987.
230. *Décret no. 77 – 469 du 28 avril 1977, relatif à la présentation et à la publicité des produits cosmetiques et produits d'hygiène corporelle*, Journal Officiel, May 5, 1977.
231. *Décret no. 77–1558 du 28 décembre 1977 relatif à la constitution du dossier et aux transmissions préalables à la mise sur le marché d'un produit cosmétique ou d'un produit d'hygiène corporelle*, Journal Officiel, January 25, 1978.
232. *Zweites Gesetz zur Änderung des Lebensmittel- und Bedarfsgegenständegesetzes*, November 25, 1994, BGBl. I, pp. 3538 – 3544.
233. *Fünfundzwanzigste Verordnung zur Änderung der Kosmetik-Verordnung*, December 23, 1996, BGBl. I, p. 2186 – 2192.
234. *Real Decreto 1415/1995* of August 4, 1995, Boletín Oficial del Estado, September 14, 1995.
235. FDA, *Federal Food, Drug and Cosmetic Act (FDC Act)*.
236. FDA, *Fair Packaging and Labeling Act (FPLA)*.
CTFA Labeling Manual, CTFA, Washington 1997.
237. *Pharmaceutical Administration in Japan*, 7th ed., Yakuji Nippo, Tokyo 1996.
238. *Principles of Cosmetic Licensing in Japan*, 2nd ed., Yakuji Nippo, Tokyo 1989.
239. *Drug Approval and Licensing Procedures in Japan 1995*, Yakugyo Jiho Co. Ltd., Tokyo 1996.
240. *The Japanese Standards of Cosmetic Ingredients*, 2nd ed., Yakuji Nippo, Tokyo 1985.
Supplement I, Yakuji Nippo, Tokyo 1986.
Supplement II, Yakuji Nippo, Tokyo 1992.
241. *Nippon Keshohin Genryoshu (Japanese Cosmetic Ingredient Dictionary)*, 3rd ed., Yakuji Nippo, Tokyo 1994.
242. *Guide to Quasi-drug and Cosmetic Regulations in Japan*, Yakuji Nippo, Tokyo 1992.
243. *Quasi-drug Registration Requirements in Japan*, Yakuji Nippo, Tokyo 1990.
244. *The Japanese Standards of Quasi-drug Ingredients*, Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Tokyo 1991.
Supplement I, PAB Notification No. 115 of February 10, 1993, Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Tokyo.
Supplement II, PAB Notification No. 243 of March 15, 1994, Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Tokyo.
245. *The Japanese Cosmetics Ingredient Codex 1993*, Yakuji Nippo, Tokyo 1994.
The Supplement to the Japanese Cosmetics Ingredient Codex 1993, Yakuji Nippo, Tokyo 1995.
246. *The Comprehensive Licensing Standards of Cosmetics by Category 1994*, Yakuji Nippo, Tokyo 1994.
247. *Generell Forskrift for Produksjon, Import og Frambud m.v. av Kosmetiske Produkter og Kroppspleieprodukter*, 26.10.1995, no. 871.
248. *Kozmetik Yönetmeligi* (Cosmetics Regulations), Resmi Gazete (Official Gazette), issue no. 21889, April 8, 1994, p. 3 – 66.
249. 21 CFR 358, Federal Register vol. 51, no. 146, July 30, 1986.
250. *Thai Cosmetics Act B.E. 2535* (1992) and Regulations, Cosmetic Control Division, Ministry of Public Health, Bangkok October 1994.
251. *Approval Standards for the Manufacture and Import of Hair Dyes*, PAB Notification No. 533 of May 14, 1991, Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Tokyo.
Guidebook for the Application of Hair Dyes Manufacturing (Import) Approval, 5th revised ed., January 1997, Ministry of Health and Welfare, Tokyo (in Japanese).
252. FDA, *Federal Food, Drug, and Cosmetic Act*, Section 601 (a).

253. *J. Am. Coll. Toxicol.* **10** (1991) 135 – 192.
254. FDA Plant Inspections Operation Manual (FDA recommendation, undated).
255. *Approval Standards for the Manufacture and Import of Permanent Waving Agents*, PAB Notification No. 111 of February 10, 1993, Pharmaceutical Affairs Bureau, Ministry of Health and Welfare, Tokyo;
Guidebook for the Application of Permanent Waving Agents, 5th revised ed., October 1996, Ministry of Health and Welfare, Tokyo (Japanese).
256. *Commission Directive 94/1/EC of 6 January 1994 adapting some technicalities of Council Directive 75/324/EEC on the approximation of the laws of the relating Member States to aerosol dispensers*, Official Journal of the European Communities No. L 23, 28.01.1994, pp. 28–29.
257. *Verordnung des Bundesministers für wirtschaftliche Angelegenheiten und des Bundesministers für Arbeit und Soziales über die Lagerung von Druckgaspackungen in gewerblichen Betriebsanlagen*, BGBl. Nr. 221/1995, October 6, 1995.
258. *Federal Food, Drug, and Cosmetic Act*, **Section 740.11**.
Consumer Product Safety Commission (CPSC).
259. *State Regulation of the Volatile Organic Compound Content of Personal Care Products*, CTFA, Washington, November 30, 1995.
260. *Standard Technique for Test and Examination of Aerosols in Japan (53 Security No. 117)*, Ministry of International Trade and Industry, Tokyo, September 9, 1991.
261. *Ordinance No. 56* of August 21, 1989, Ministry of International Trade and Industry, Official Gazette, August 21, 1989.
Notification No. 419 of August 21, 1989, Ministry of International Trade and Industry, Official Gazette, August 21, 1989.
262. G. Stüttgen, *Soc. Cosmet. Chem.* **32** (1981) 231 – 245
263. W. Wirth, G. Hecht, Chr. Gloxhuber: *Toxikologie Fibel*, Thieme Verlag, Stuttgart 1967.
264. A. K. Palmer, A. M. Bottomley, I. A. Edwards, R. Clark, *Toxicology* **26** (1983) 313 – 315.
265. G. Schuster, A. Domsch, *Seifen öle Fette Wachse* **108** (1982) 177 – 184, 248 – 250.
266. M. M. Breuer, G. X. Gikas, I. T. Smith, *Cosmet. Toiletries* **94** (1979) 29 – 34.
267. V. Vítez, *Riechst. Aromen Körperpflegem.* **10** (1970) 392 – 396.
268. H. Ippen, *ärztl. Kosmetol.* **16** (1986) 432 – 435.

Hard Materials

HEINZ ESCHNAUER, Plasma-Technik AG, Wohlen, Switzerland (Chap. 1)

ERICH KNY, Metallwerk Plansee GmbH, Reutte, Austria (Chap. 2)

1. Introduction 1
2. Cemented Carbides 3
2.1. Production 5
2.1.1. Raw Materials 5
2.1.2. Powder Batch Manufacturing 6
2.1.3. Pressing and Forming 6
2.1.4. Sintering 6
2.1.5. Machining 7
2.1.6. Coating 7
2.1.7. Testing and Quality Control 7
2.2. Commercial Grades 7
2.3. Properties 10
2.4. Uses 10
2.5. Economic Aspects 11
2.6. Toxicology 12
3. References 13

1. Introduction

Hard materials are solid substances with high intrinsic hardness values: between 8 and 10 on the Mohs' scale, but occasionally lower. The hardest of all known substances is diamond (→Carbon, Chap. 2.) with a Mohs' hardness of 10 (ca. 10 000 on the Vickers scale) [1].

The technology of hard materials forms a part of materials science and technology that can be called "hard materiology." The manufacture and use of these special substances involves the sciences of chemistry, physics, metallurgy, and powder metallurgy (sintered metals).

The classification of hard materials may be based on their occurrence and manufacture, or on their properties and uses. Their commercial importance is their use in the cemented carbide and abrasives industries (→Abrasives) and as wear-resistant materials (tribology).

Hard materials occur naturally or may be synthetic. In nature they usually are found as elements or oxides, some of them precious stones such as diamonds, rubies, or sapphires. The classification of natural and synthetic hard materials is summarized in Figure 1.

Metallic and nonmetallic hard materials may be defined according to their properties and uses. Of the metallic materials, the most commonly used are the hard metals. Physical properties, such as electrical and thermal conductivity, luster, and color, are taken into account. This classification depends on the bond character, which can vary from metallic or covalent to ionic (heteropolar), corresponding to the metallic and nonmetallic hard materials (Table 1) [2].

Table 1. Classification of synthetic hard materials

Metallic	Nonmetallic
Binary compounds	Oxide ceramics
Carbides∗	Aluminum oxide (Al_2O_3)
Nitrides∗	Chromium oxide (Cr_2O_3)
Borides∗	Titanium oxide (TiO_2)
Silicides∗	Zirconium oxide (ZrO_2)
Multicomponent systems	Other oxides
Mixed carbides	Mixed oxides
Mixed nitrides	
Mixed borides	Nonoxide ceramics
Mixed silicides	Carbides (B_4C, SiC)
Carbonitrides	Nitrides (AlN, CBN, Si_3N_4)
Boronitrides	Borides (α-AlB_{12}, SiB_6)
Borocarbides	
Silicoborides	Elements
Silicocarbides	Boron
Siliconitrides	Carbon
Double carbides, complex carbides	Silicon
Intermetallic compounds (Laves phases)	Cermets (composites)
Mixed hard materials without binder	Nonmetallic hard material and metal
Hard material alloys that can be hardened	

∗ of the refractory metals of groups 4 – 6 of the periodic table: Ti; Zr, Hf; V, Nb, Ta; Cr, Mo, W.

Metallic Hard Materials. The most important metallic hard materials are the binary compounds of the refractory metals (transition metals) of groups 4 (titanium, zirconium, hafnium), 5 (vanadium, niobium, tantalum), and 6 (chromium, molybdenum, tungsten) of the pe-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_603

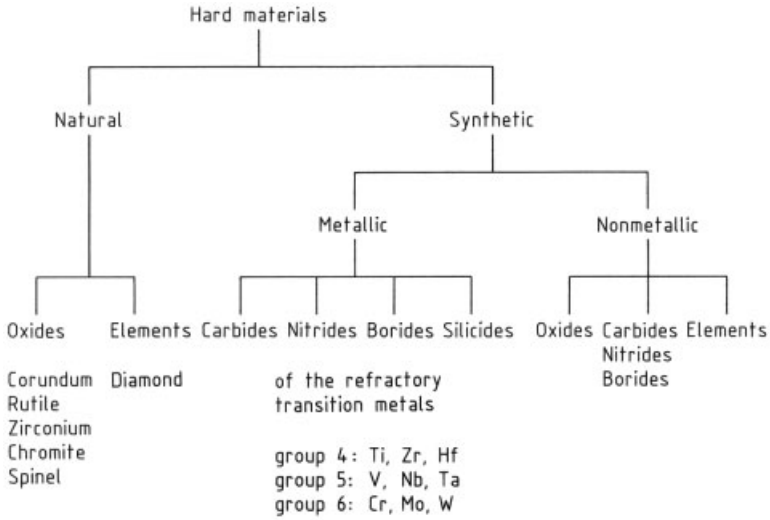

Figure 1. Classification of natural and synthetic hard materials

riodic table with carbon (carbides, →Carbides), nitrogen (nitrides [3], see →Nitrides), boron (borides), and silicon (silicides). The metallic hard materials also include multicomponent systems, double carbids, complex carbides, intermetallic compounds (Laves phases), and mixed hard materials without binder metals.

Hard material alloys that can be hardened consist of titanium carbide bonded in a matrix of hardenable steel (Ferrotitanite). The titanium carbide content can be as high as 50 %, the remainder being mostly iron, with small amounts of other alloying components and, in some cases, other hard materials.

When a metallic hard material is combined with a metal, usually of the iron group, the result is either a *cemented carbide* (hard metal) [4], [5] or a *hard alloy* [6], [7] depending on the ratio of hard material to metal (Fig. 2).

The raw materials for cemented carbides are manufactured on a large scale and include: tungsten, titanium, tantalum, and niobium carbides; tungsten – titanium and tantalum – niobium carbides of various proportions; and recently titanium nitride and titanium carbonitrides, also of various proportions (→Carbides). Cemented carbides are made by powder metallurgy (sintering) and consist of a matrix of hard material (usually tungsten carbide) with a small proportion of metal (usually cobalt). Hard alloys are multicomponent (usually ternary or quaternary) alloys that are made by melting techniques, usually atomization. They generally have an intercalated structure comprising a metallic matrix (usually iron, cobalt, or nickel) with an appropriate proportion of hard materials (carbides, borides, or silicides) for wear resistance.

Nonmetallic Hard Materials. The nonmetallic hard materials include:

1) The oxide ceramics, such as the oxides of aluminum, chromium, titanium, and zirconium, as well as their mixtures (e.g., of aluminum oxide and titanium dioxide) and compounds (e.g., aluminum oxide – magnesium oxide).
2) The nonmetallic carbides, nitrides, and borides, as well as the now very important nonoxide ceramics silicon carbide and silicon nitride (→Ceramics, General Survey, Chap. 1.2.1.).
3) The nonmetallic elements boron, carbon, and silicon.

Cermets (CERamic + METal) (→Ceramics, Ceramic – Metal Systems, Chap. 6.1.), are composite systems consisting of nonmetallic (usually oxide ceramic) hard materials in a metal matrix; they are normally manufactured by powder metallurgy.

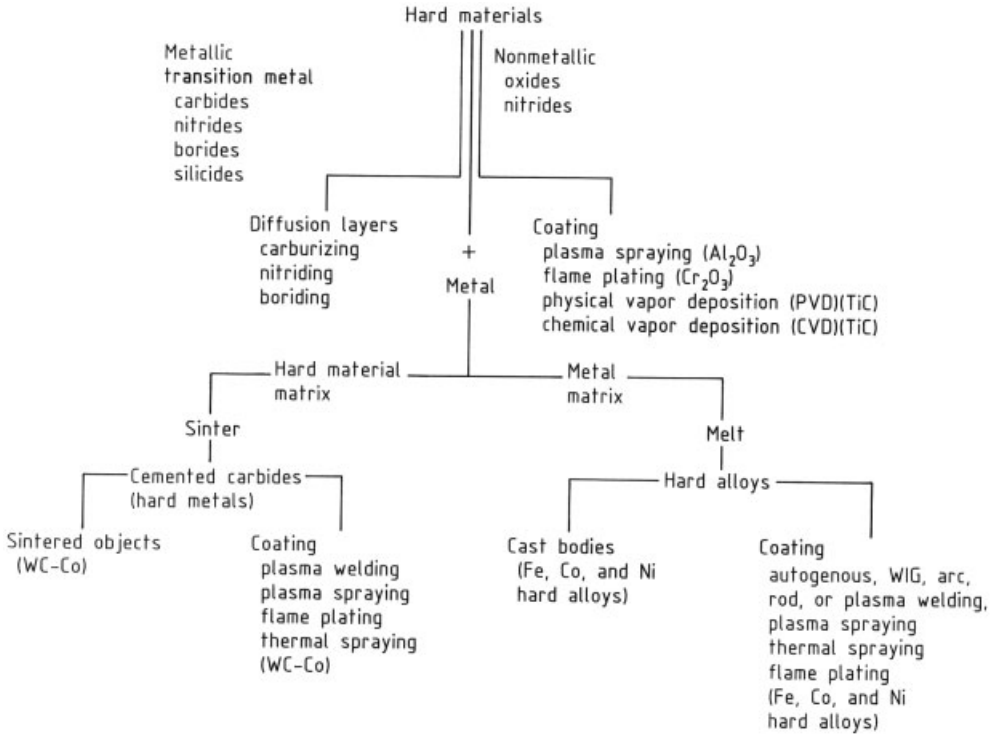

Figure 2. Hard material – metal systems

2. Cemented Carbides [8–16]

Cemented carbides or "hard metals" are materials that combine great hardness with metallic properties. They differ from the nonmetallic hard materials in possessing a metallic luster and good electrical and thermal conductivity. They consist of particles of brittle metallic hard materials (usually carbides) in a matrix of ductile binder metals belonging to the iron group (iron, nickel, and cobalt).

The cementation of an extremely hard component with a binder metal of great ductility is a composite material referred to as a cemented carbide. The hard material and the binder metal occur as separate phases. In the typical cemented carbide structure shown in Figure 3, the dark angular grains are the hard material tungsten carbide (WC), which are bound together by thin, lighter colored layers of cobalt. Increasing the hard material content gives a harder product, whereas increasing the binder metal content gives increased toughness.

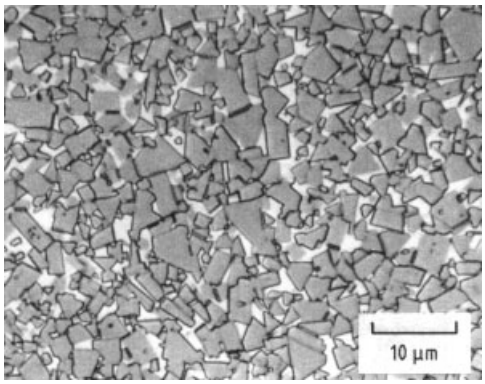

Figure 3. Microstructure of WC – 9 % Co cemented carbide

History. The manufacture of cemented carbides was made possible by the invention of the electric furnace at the turn of the century. Moissan [17] used this furnace in his attempt to make diamonds. He obtained extremely hard crystals, but analysis revealed that these were metallic carbides and not diamonds. They were not as hard as diamonds but harder than any other known substance. Many attempts were

made to exploit these hard carbides to make tools and wear-resistant components. In 1914, VOIGTLÄNDER and LOHMANN succeeded in pulverizing, pressing, and sintering tungsten carbide to produce wear-resistant articles such as wire-drawing dies [18]. The toughness of these articles was improved by alloying with metals of the iron group. Hot pressing was also successfully employed in addition to melting and sintering techniques.

The decisive step in the development of sintered carbides was taken in 1923 by K. SCHRÖTER and F. SKAUPY of the Osram Study Association [19]. They combined tungsten carbide powder with cobalt powder, formed this mixture into shapes with a press, and heated the shapes above the melting point of cobalt. The tungsten carbide partially dissolved in the cobalt to give a molten eutectic that wetted the tungsten carbide grains. Shrinkage then took place, giving a very dense product. Because their strength values were greater than those of pure carbides, sintered carbides were subsequently used for metal cutting. SCHRÖTER's patents were acquired by the Krupp organization, which in 1925 introduced the first lathes fitted with cutting tools of tungsten carbide–cobalt under the name "Widia" (from the German "*wie Diamant*," i.e., diamond-like). Cast alloys could now be machined at two or three times the speeds achievable with the usual high-speed steel tools. Results on steel were, however, disappointing because of the occurrence of crater wear in the tool material when it was used for turning steel. Diffusion between the hot turnings and the tool caused this deterioration of the cemented carbide.

The vacuum sintering process and the use of titanium carbide (TiC) subsequently allowed great increases in steel-cutting speeds [20].

A tungsten-free cemented carbide (87 % TaC and 13 % Ni) was marketed under the name " Ramet" by Fansteel in the United States. General Electric, in cooperation with Krupp, developed tungsten carbide–cobalt alloys with the trade name " Carboloy." In the mid 1930s ternary cemented carbides (WC, TiC, and TaC) were also produced in the United States [21]. Kennametal also started to manufacture cemented carbides during this period.

In the mid 1970s low-tungsten cemented carbide tools were developed with a fine-grained structure obtained by spinodal decomposition of the components in the system TiCN – MoW – Ni [22]. During the same period WC – Co cemented carbide tools with a grain diamter <1 µm were also produced. These submicron cemented carbide tools combine improved abrasion resistance with satisfactory toughness [23], [24].

Around one fifth of the WC – Co cemented carbides tools produced currently are used in the mining industry. Sweden is an important producer; Sandvik started production there during World War II.

Products containing little or no tungsten carbide have also become important for fine cutting. These are of the type Ti(C, N) with additions of Cr, Mo, W, Fe, Co, and Ni. Much of the initiative in the development and marketing of these materials has come from Japan.

The coating of materials with layers of hard material by means of chemical vapor deposition (CVD) was first described in 1953 [25], [26]. However, this technique was first put to practical use in the mid 1960s for the machining of steel and iron castings [27]. Multilayer coatings (TiC, TiN, and Al_2O_3) were developed in the early 1970s [28], [29].

2.1. Production

Cemented carbides are usually produced by powder metallurgy rather than on melting and casting (→Powder Metallurgy and Sintered Materials). The structure shown in Figure 3 is most often achieved by a sintering technique in which pressed shapes made from the mixed hard material and binder are heated to temperatures well below the melting point of the hard material, but close to the melting point of the binder. The molten binder wets the particles of hard material and results in densification. A short description of the technology of hard metal fabrication can be found in [30–32].

2.1.1. Raw Materials

With the exception of Ti and C, the elements found in cemented carbides rarely occur in the earth's crust. Furthermore, the manufacturing procedure is complex and a high degree of purity is required, so that the overall raw material cost is very high.

Tungsten Carbide (WC) (→Carbides, Chap. 2.1.). Tungsten carbide [*12070-2-1*] can be made from blue tungsten oxide (W_2O_5), ammonium paratungstate (APT) or tungstic acid (H_2WO_4), which are reduced with hydrogen at 800 – 1100 °C to tungsten powder. The particle size of this product is primarily determined by the temperature and moisture content of the hydrogen. Depending on the type of cemented carbide required, particle size can be between 0.3 and 12 µm. Carburization leaves the particle size and size distribution of the tungsten powder largely unchanged. The powder is first intensively mixed with very pure carbon black or graphite. The mixture is then heated to 1500 – 2000 °C. Next the resulting solid cakes of carbide are broken down to a particle size <0.5 mm by means of jaw crushers, hammer mills, blower – beater mills, pin mills, vibratory mills, or ball mills. Vibratory sieves are used to obtain a product of the desired particle size range.

Titanium Carbide (TiC) (→Carbides, Chap. 2.2.). Titanium carbide [*12070-08-5*] is usually obtained from TiO_2 produced by the pigment industry. Carburization with carbon black is carried out at 2100 – 2300 °C, and the product is powdered as described for tungsten carbide. Titanium carbide is mainly used to reduce crater wear resistance of cemented carbides.

Tantalum – Niobium Carbides (→Carbides, Chap. 3.2.). Either an expensive grade of tantalum carbide with a maximum of 1 % niobium carbide is used, or a cheaper grade with a higher niobium content (Ta : Nb = 6 : 1 to 1 : 1). The most common raw material for the manufacture of tantalum – niobium carbide is pure $(Ta, Nb)_2O_5$, which is mixed with carbon, usually in the form of carbon black, and reacted at 1600 –1800 °C. Tantalum carbide improves the toughness and resistance against thermal fatigue of cemented carbides. If the carbides are only to be used for turning operations, the tantalum carbide can be replaced with a higher percentage of niobium carbide. The metal-cutting properties of some milling cutting tools are, however, impaired by higher niobium carbide contents.

Mixed Carbides. The mixed carbides (Ta, Nb, Ti, W) C are often used. They are usually manufactured in the same way as the pure carbides. However, they can also be produced by the "menstruum" process, developed by Kennametal Inc. [33] (USA). This special method involves carburization in molten nickel, which promotes diffusion. Very pure, rather large carbide crystals are then isolated from the reaction products by treatment with acid.

Cobalt and Other Binder Metals. Cobalt powder suitable for the manufacture of cemented carbides is produced by reducing cobalt oxide, carbonate, oxalate, or formate with hydrogen at 600 – 800 °C. The powder must be very fine (1 – 5 µm) and free from impurities. Coarser cobalt powder causes local cobalt enrichment in the cemented carbides.

Nickel, iron, and chromium are rarely used as binder metals. However, nickel and chromium give corrosion-resistant products (see Section 2.3). Many attempts have been made to find a substitute for cobalt, but without success. All other metals are difficult to process and confer inferior properties in the cemented carbide.

2.1.2. Powder Batch Manufacturing

The powdered carbides, binder metals, and carbon are weighed. The mixture is then wet ground and agents are added to aid pressing. Subsequent drying and granulation give a powder batch ready for pressing and forming.

Grinding. Wet grinding converts the powder mixture into a very fine, intimate dispersion. An organic liquid milling medium (e.g., a hydrocarbon, alcohol, or acetone) prevents both reagglomeration and oxidation of the powder.

Early ball mills had a capacity of 5 – 10 L but have been replaced by mills with over 100 L capacity, and by vibration mills and attrition mills (→, Chap. 3.2.2.Size Reduction).

Drying. The slurry obtained after grinding is sieved and the liquid medium usually is removed by distillation. Thus, the liquid can be recovered in a pure state and reused. The powder must have good flow properties to allow the automatic mold-filling equipment to operate smoothly. To achieve this, the powder is formed into granules

with the aid of paraffin wax or poly(ethylene glycol). Simultaneous drying and granulation in a spray dryer are very economic.

2.1.3. Pressing and Forming

Presintered shapes are formed by various pressing methods followed, if necessary, by machining.

Direct Pressing. About two thirds of all cemented carbide products are directly pressed into the required shapes using a suitable die in a mechanically or hydraulically operated automatic press.

Extrusion. Rods, bars, and tubes can be formed with the extrusion presses commonly used in the ceramic industry. Further additives such as paraffin wax or poly(ethylene glycol) are required to produce a formable plastic mass that can flow and be extruded through suitably shaped cemented carbide nozzles.

Cold Isostatic Pressing. Large cemented carbide components of simple geometry are often cold isostatically pressed, presintered, and then machined. The carbide powder is compressed in a sealed leakproof tube on all sides by a liquid (usually water) at high pressure (ca. 200 MPa).

Indirect Forming. A large demand exists for components that do not justify the high cost of pressing equipment either because of their complicated shape or the small number required. Such components must be formed by machining pressings of a suitable shape or pieces cut from large, cold isostatically pressed bodies. The articles are first presintered to increase their strength by heating to 700 – 900 °C in hydrogen or vacuum. The pieces are then machined with diamond tools. The powder produced during machining is carefully recovered, if possible sorted into grades, and reused.

2.1.4. Sintering

The final sintering causes linear shrinkage of about 18 – 25 % in each direction resulting in a 40 – 50 % decrease in volume. The "green" pressed shapes are placed on graphite plates for sintering which is usually performed in a vacuum, less often in a hydrogen atmosphere. In modern furnaces, the pressed shapes can be directly sintered without a presintering stage. This vaporizes the wax additives used for pressing, which are then trapped in a wax separator. Despite technical improvements, the sintering step has a great influence on the quality of the end product. The carbon balance, which is so important for the properties of the product (see Section 2.3), depends not only on the carbon concentration of the powder but also on the sintering conditions (temperature, time, residual gas composition, nature of the supporting surface, and degree of filling of the furnace).

2.1.5. Machining

Although the sintered part is already fairly accurately shaped, final shaping with diamond tools or spark erosion is often necessary. The latter process allows the production of complex shapes (e.g., with holes or complex contours) that cannot be obtained directly by pressing. Surface roughness is removed by grinding, and closer dimensional tolerances can be achieved by this means in comparison with dimensional tolerances obtained after sintering.

2.1.6. Coating

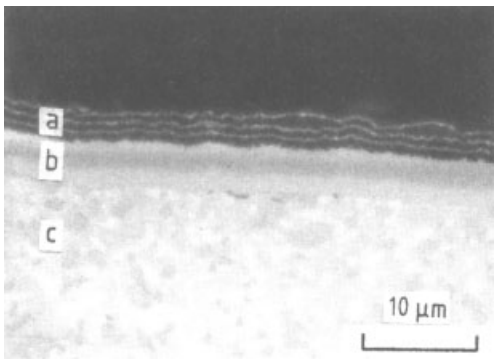

Figure 4. Multilayer wear-resistant coating (total thickness ca. 8 µm) on cemented carbide substrate
a) Oxide layers; b) Ti(C,N)-containing intermediate layers; c) TiC – Ti(C,N) underlayer

Table 2. Commonly used combinations of wear-resistant coatings with cemented carbide substrates [34]

Substrate	Wear-resistant coating							
(cemented carbide) *	TiC	TiN	TiC – Ti(C,N) – TiN	TiC Al_2O_3	TiC – Ti(C,N) – TiN Al_2O_3	Al_2O_3	HfN	Al – O – N
M 15	+		+	+	+	+	+	+
P 25		+	+				+	
P 40	+		+					
K 10	+	+	+			+		+

* For explanation of designations, see Table 3.

The wear resistance of cemented carbide articles is increased by applying thin, hard coatings of TiC, TiN, Ti(CN), Al_2O_3, or other hard materials. The usual coating methods are chemical vapor deposition (CVD) or physical vapor deposition (PVD) (→Thin Films). Deposition temperatures for CVD are ca. 1000 °C, but for PVD they can be under 400 °C. Multiple coatings of layers of various hard materials are now common (Fig. 4). The total coating thickness seldom exceeds 10 µm. The particular combination of coating materials used depends on the nature of the cemented carbide substrate (Table 2) and the particular application.

2.1.7. Testing and Quality Control

Throughout its manufacture, the cemented carbide is subjected to metallographic, dimensional, physical, and chemical tests. Many of these are standardized (in Europe according to ISO and DIN standards and, in the United States, according to ASTM). Users of cemented carbides measure the following important parameters (ISO 4498):

Compressive strength:	ISO 3326
Density:	ISO 3369, ASTM B 311-58
Rockwell hardness:	ISO 3738, ASTM B 294-76
Vickers hardness:	ISO 3878.

Other determinations are sometimes carried out, e.g., of microstructure (ISO 4499, ASTM B 657-79) and of porosity and free carbon (ISO 4505, ASTM B 276-79).

The coercive force is a sensitive indicator of the structure of the material and is strongly influenced by changes in the grain size of the WC, an important factor in determining the cutting performance. The metal-removing capacity is an important parameter for the user and is usually determined by means of short duration cutting tests simulating practical conditions. The various performance characteristics of wear-resistant parts for other applications can usually only be assessed while they are in service.

2.2. Commercial Grades

The existence of many types of cemented carbides differing in composition and properties has led to a considerable broadening of their field of application, but also to a confusing multitude of products. In order to simplify the selection of a product for a particular application, various classification systems have been devised.

DIN 4990. The first standardization system was established in Germany, leading in May 1942 to the publication of DIN 4990. A color code was used to indicate different types of brazed tools.

United States. Here the Industry Code is used, and sometimes also the Buick Code, developed by the Buick Motor Division of General Motors [30].

ISO. The International Standardization Organization (ISO) began to concern itself with cemented carbides in the mid 1950s. By this time the DIN 4990 was not adequate for the classification of the newly developed universal cemented carbides for cutting cast iron and steel. The color system had also become increasingly confusing. The identifying colors blue, yellow, and red were then allotted to the metal-cutting groups P, M,

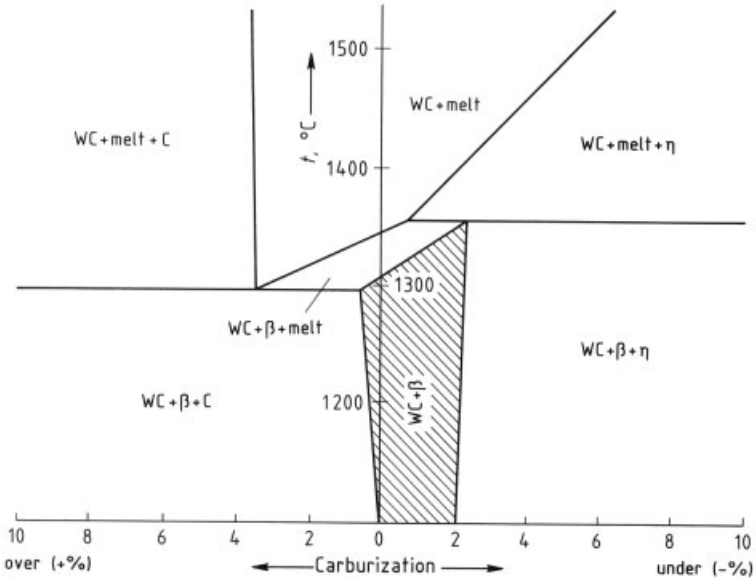

Figure 5. Vertical section of W – C – Co ternary-phase diagram at 16 % Co [35]
β = solid solution of W and C in Co; η = embrittling η-phase (W_6Co_6C); C = carbon

and K [30]. These main groups were subdivided into application groups by means of two-digit numbers whose value indicated the toughness of the carbide material. Higher values indicated higher toughness to suit the requirements both of the metal to be cut and the working conditions, with higher feed rates and lower cutting speeds. Lower values correspond to better wear resistance at the expense of toughness and higher cutting speeds. Thus, high numbers indicate high toughness and low index numbers indicate harder, more wear-resistant materials (Table 3).

The *World Directory and Handbook of Hardmetals* gives collected data on cemented carbides in tabular form [30]. The various products available under each ISO classification are listed for comparison to facilitate product choice.

Great Britain. In 1967 the British Hardmetal Association (BHMA) brought out a classification system based on hardness, transverse rupture strength, and titanium and tantalum carbide contents as a measure of crater wear resistance. This system has not been used outside Great Britain and has found little acceptance even there [30].

2.3. Properties

The important application properties and the manufacturing techniques for WC – Co cemented carbides are dictated by the ternary-phase diagram for W – C – Co (Fig. 5) [34]. The important region for both production and properties is that marked WC + β (solid solution). Cemented carbides for industrial use must have a composition lying within this region, which is ca. 0.1 wt % C "wide." Carbon deficiency (undercarburization) causes the appearance of the η-phase and hence embrittlement. An excess of carbon (overcarburization) separates out as finely divided graphite.

The transverse rupture strength is an important parameter in practice. It reaches the maximum in the binary-phase region and is reduced by the appearance of the η-phase or by the precipitation of a graphite phase. However, hardness increases with undercarburization (Fig. 6).

The "art" of manufacturing cemented carbides depends on the control of the carbon balance. Other important properties of cemented carbides in comparison with tool steels and other hard materials are given in Table 4. Cemented carbides with a cobalt binder are generally not resistant to corrosion. Even aqueous solutions

Table 3. ISO classification of cemented carbides with typical compositions and properties

Designation ISO/TC 29/GT 9			Color code	Composition[b], wt %			Density, g/cm³	Vickers hardness HV 30	Transverse rupture strength, N/mm²	Compressive strength, N/mm²	Modulus of elasticity, kN/mm²	Thermal expansion coefficient, 10^{-6}/K	Thermal conductivity, $W m^{-1} K^{-1}$
Main groups for machining	Application groups	Distinguishing properties		WC, %	TiC + TaC, %	Co, %							
P	P01.2	↑ [a]	blue	30	64	6	7.2	1800	750	3500			
	P01.3		(RAL 5012	51	43	6	8.5	1750	900	4200	460	7.5	17
steel,	P01.4		Munsell	62	33	5	10.1	1750	1000				
cast steel,	P05		2.5 PB 5/10)	77	18	5	12.2	1700	1100	4300		6	
malleable cast	P10	1	2	63	28	9	10.7	1600	1300	4600	530	6.5	29
iron	P20			76	14	10	11.9	1500	1500	4800	540	6	33
	P25			71	20	9	12.4	1450	1750	4800		6	
(long chipping)	P30			82	8	10	13.1	1450	1750	5000	560	5.5	59
	P40			75	12	13	12.7	1400	1950	4900	560	5.5	59
	P50	↓		68	15	17	12.5	1300	2200	4000	520		
M steel, cast steel manganese	M10	↑ [a]	yellow	84	10	6	13.1	1700	1350	5000	580	5.5	50
hardened steel, alloyed cast	M20		(RAL 1007	82	10	8	13.4	1550	1600	5000	570	5.5	63
iron, austenitic steel, malle-	M30	1	2 Munsell	81	10	9	14.4	1450	1800	4800			
able cast iron, nodular iron	M40		7.5 YR 7/10)	79	6	15	13.6	1300	2100	4400	540		
pearlitic, free cutting steel		↓											
K cast iron, chilled iron,	K01	↑ [a]	red	92	4[c]	4	15.0	1800	1200				
malleable cast iron (short	K05		(RAL 3001	91	3[c]	6	14.5	1750	1350	5900	630	5	
chipping) hardened steel,	K10	1	2 Munsell	92	2[c]	6	14.8	1650	1500	5700	630	5	80
non-ferrous metals,	K20		5 R 3/10)	92	2	6	14.8	1550	1700	5000	620	5	80
plastics, wood	K30			89	2	9	14.4	1400	1900	4700	580		71
	K40	↓		88		12	14.3	1300	2100	4500	570	5.5	70

[a] 1) Upward direction: increasing hardness, wear resistance and cutting speed; 2) Downward direction: decreasing hardness, wear resistance, increasing toughness and feed rate.
[b] Average composition of European cemented carbides. [c] Including VC.

can attack the binder phase by the formation of the local galvanic elements. The relative corrosion resistance of cemented carbides containing cobalt, nickel, and nickel – chromium binders is shown in Table 5. A nickel – chromium binder gives acceptable corrosion resistance [37]. Resistance can be further improved by using a high content of Ta – Nb carbides, which are highly corrosion-resistant ceramic carbides [38].

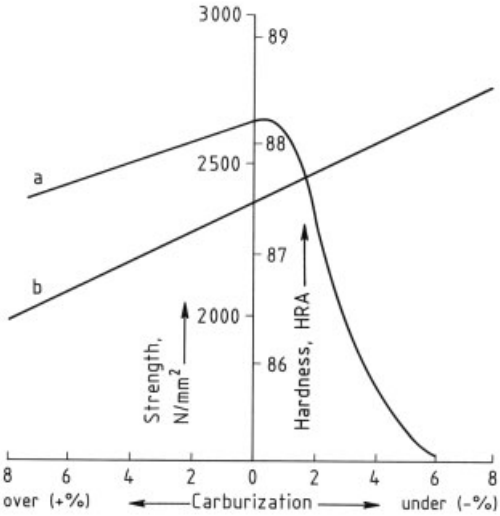

Figure 6. Rockwell hardness (R_A) and transverse rupture strength of WC – Co cemented carbides as a function of carbon content of the cemented carbide [36]
a) Transverse rupture strength; b) Rockwell hardness

2.4. Uses

About 50 – 60 % of cemented carbides has always been used for cutting. Only 8 – 10 % is employed in forming equipment (wire-drawing dies, press dies, rolls, etc.) although this proportion is increasing. The mining industry, building industry, and road and tunnel construction account for a further 20–25 %. The remainder (10 – 12 %) is used for components subjected to wear, such as grinding balls, sealing rings, plungers, grinding tools in size reduction equipment, spikes, and equipment for manufacture of synthetic diamonds.

Wear-resistant components used in the chemical industry must be constructed from corrosion-resistant cemented carbides. The corrosion-susceptible cobalt binder is replaced by Ni, Ni – Cr alloys [37], or noble metals (Pt) in these products, such as the mass-produced cemented carbide seal rings shown in Figure 7.

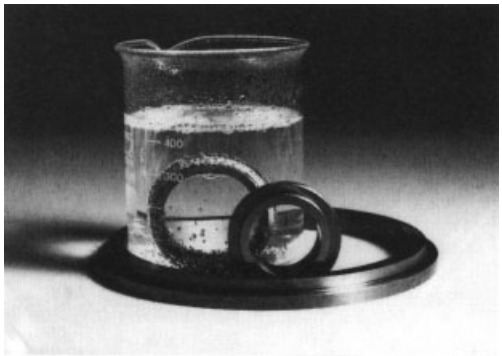

Figure 7. Cemented carbide seal rings for corrosive chemical environments

2.5. Economic Aspects

Although the production of cemented carbides is small in comparison to that of steel (Table 6), these products have achieved great economic importance because they have improved the productivity of cutting and forming processes. This is illustrated in Figure 8, which shows the enormous increase in cutting speeds that followed the introduction of cemented carbides. Results for other cutting materials are shown for comparison. The newer materials (e.g., ceramics, polycrystalline diamond, and cubic boron nitride) have produced further increases in cutting speeds.

Table 6. Estimated annual world production of hard metal and steel [31]

Year	Estimated production,	
	Cemented carbides, 10^3 t/a	Steel, 10^6 t/a
1950	1.1	190
1960	2.3	330
1965	4.5	447
1970	12.0	597
1975	18.0	680
1980	24.0	710
1985	30.0	750

World production of cemented carbides is divided about equally among the United States, Europe (excluding the former Soviet Union),

Table 4. Properties of important cutting materials [31]

Property	High-speed steel	Cemented carbides	TiC based cemented carbides	Ceramics (oxides, nitrides)	Polycrystalline cubic boron nitride	Polycrystalline diamond
Density, g/cm^3	8.0 – 9.0	9.0 – 15.0	5.0 – 9.0	3.2 – 5.0	3.45	3.5
Rockwell hardness, HRA	750 – 800	800 – 1900	1500 – 2200	1500 – 2500	4000	8000
Transverse rupture strength, N/mm^2	3000 – 4000	1000 – 3000	900 – 1800	400 – 600	500 – 800	600 – 1100
Hot transverse rupture strength (1000 °C), N/mm^2	(500)	900 – 1500	600 – 1500	400 – 600	500 – 700	600 – 1000
Compressive strength, N/mm^2	3500 – 4000	4000 – 7000	3000 – 5000	3000 – 4000	6000	7600
Modulus of elasticity, kN/mm^2	210	400 – 680	300 – 450	300 – 400	680	850
Thermal expansion coefficient, 10^{-6}/K	12	5.0 – 7.0	6.0 – 8.0	2.5 – 7.5	3.5	1.2

Table 5. Corrosion resistance of cemented carbides with Co, Ni, and NiCr binders [37]

Medium	Temperature, °C [a]	Concentration, mol/L	Corrosion resistance [b]				
			WC – 6Co	WC – 12Co	WC – 6Ni [c]	WC – 9Ni [c]	TiC – 16Mo_2C – 18Ni
Acetic acid	RT	1	2	2	1	1	1
	100	1	3	3	2	2	2
Hydrochloric acid	RT	6	2	3	2	2	2
	100	6	4	4	4	4	4
Nitric acid	RT	6	2	3	1	1	4
	100	6	4	4	4	4	4
Phosphoric acid	RT	0.67	3	3	2	2	2
	100	0.67	4	4	3	3	2
Sodium chloride	RT	0.5	2	2	1	1	1
Sodium hydroxide	RT	1	1	1	1	1	1
Seawater	RT		2	2	1	1	1
Sulphuric acid	RT	1	3	3	2	2	2
	100	1	4	4	3	4	3
Water, tap	RT		1	2	1	1	1
Water, pure	RT		1	1	1	1	1
	100		1	1	1	1	1

[a] RT = room temperature.
[b] 1 = negligible attack; 2 = light attack; 3 = medium attack; 4 = strong attack.
[c] Ni – Cr alloy with 5 % Cr.

and the rest of the world. The twenty largest producers in the Western World plus Japan are as follows (estimated outputs for 1986 in tonnes per year):

AB Sandvik (Sweden)	3000
Kennametal Latrobe (USA)	1500
Carboloy Systems of Gen. Electric (USA)	1500
Seco Tools AB, Sweden, Italy, Brasil	800
Teledyne Firth Sterling (USA)	800
Krupp Widia (FRG)	700
Mitsubishi Metal (Japan)	700
Hertel AG, Werkzeuge und Hartstoffe (FRG)	650
Céramétal S.A.R.L. (Luxemburg)	600
Toshiba Tungaloy (Japan)	500
Sumitomo Electric (Japan)	450
Boart International Springs	450
GTE Walmet, Pleasant Ridge (USA)	450
Tungsten Carbide Manufacturing (USA)	450
Plansee TIZIT (Austria)	400
Newcomer Products (USA)	400
Fansteel VR/Wesson (USA)	300
Reed Tool Company (USA)	250
Hughes MPD (USA)	250
Atrax Div., Atrax Newcarb (USA)	250

2.6. Toxicology

Losses must be avoided in the processing of very fine dust consisting of hard material (carbide), cobalt, and cemented carbide. Loss of dust represents not only an economic loss but also a danger to health, and must be prevented by means of dust extraction equipment, filters, and masks. The adverse effects of the dust have been the subject of numerous investigations [39–42].

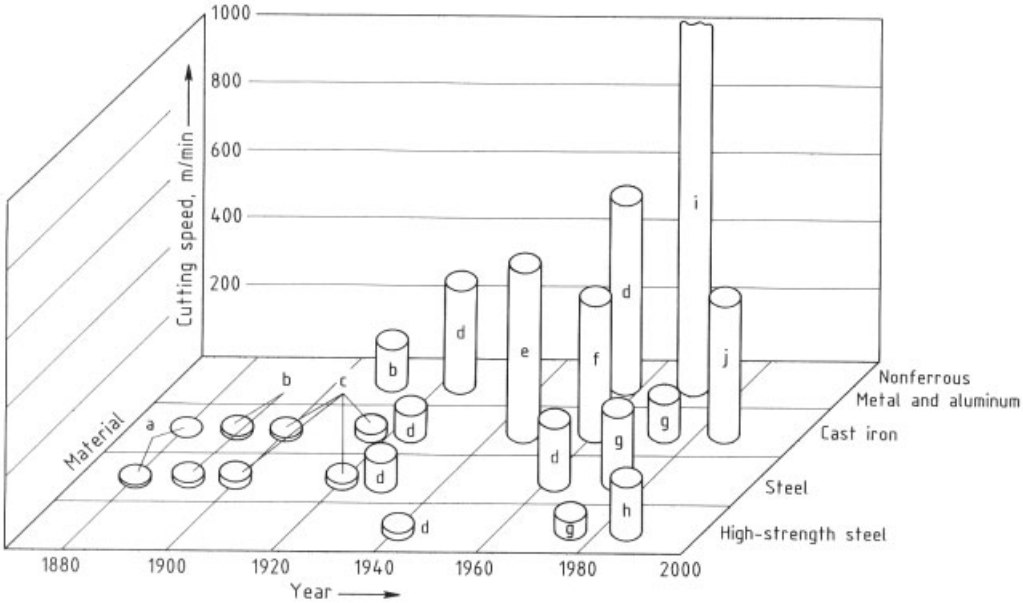

Figure 8. Increase of cutting speeds for various materials as a result of the progress in cutting materials [31]
a) Carbon steel; b) Alloyed tool steel; c) High-strength steel; d) Cemented carbide; e) Oxide ceramics; f) Mixed ceramics; g) Coated cemented carbide; h) Cubic boron nitride; i) Polycrystalline diamond; j) Silicon nitride

The toxicity of cemented carbides is primarily a function of the cobalt content, but the carbides can also form chemically inert deposits in the lungs that are not degraded or eliminated by the body. Diseases caused by emented carbides are associated with the presence of cobalt in the blood and other organs. However, the predominant illness is fibrosis of the lungs, which reduces their performance. If exposure to the dust ceases during the early stages of the disease (e.g., by a change of work environment), rapid abatement of the symptoms usually results. The TLV value for cobalt is 0.1 mg/m^3 at present, although this may be reduced to 0.05 mg/m^3. For tungsten, tantalum, and their insoluble compounds, the corresponding value is 5 mg/m^3. Regular medical checkups are required for workers in contact with cemented carbide dust [43].

Toxicity problems can also arise where cemented carbides are processed by the user [44]. Wet grinding may produce dangerous concentrations of cemented carbide material in the form of aerosols.

3. References

1. R. Binder, *Radex Rundsch.* **4** (1975) 531 – 557.
2. H. Holleck: "Material Selection for Hard Coatings", *Int. Conf. Metallurgical Coatings,* San Diego, April 7 –11, 1986.
3. H. Holleck: *Binäre und ternäre Carbid- und Nitridsysteme der übergangsmetalle, Materialkundlich-Technische Reihe 6,* Gebrüder Borntraeger Berlin Stuttgart 1984.
4. R. Kieffer, F. Benesovsky: *Hartmetalle,* Springer Verlag, Wien 1965.
5. St. W. H. Yih, Ch. T. Wang: *Tungsten Sources, Metallurgy, Properties, and Applications,* Plenum Press, New York London 1979.
6. O. Knotek, E. Lugscheider, H. Eschnauer: *Hartlegierungen zum Verschleißschutz,* Verlag Stahleisen, Düsseldorf 1975.
7. H. Eschnauer, *Thin Solid Films* **73** (1980) 1 – 17.
8. H. E. Exner, J. Gurland, *Powder Metall. Int.* **2** (1970) 59 – 63, 104 – 105.
9. V. I. Tumanov: *Eigenschaften von WC – Co-Legierungen,* Metallurgija, Moscow 1971.
10. H. E. Exner, *Int. Met. Rev.* **24** (1979) 149 – 173.

11. W. Schedler: *Hartmetalle für den Praktiker,* VDI-Verlag, Düsseldorf 1988.
12. H. Fischmeister: *Int. Conf. on the Science of Hard Materials,* Jackson Lake Lodge, Wyoming, August1981.
13. W. König: *Fertigungsverfahren Drehen, Fräsen, Bohren,* **vol. 1, 2nd ed.,** VDI-Verlag, Düsseldorf 1987.
14. *Milling Handbook of High Efficiency Metal Cutting,* Carboloy Inc., USA, 1980.
15. R. Storf: *Drehen–Grundlagen und Anwendungstechnik,* VDI-Verlag, Düsseldorf 1987.
16. *Properties and Proven Uses of Kennametal Hard Carbide Alloys,* Kennametal Inc., Latrobe 1972.
17. H. Moissan: *Der elektrische Ofen,* Berlin 1900.
18. Metall-Fabrikationsgesellschaft, DE 286 184, 1914; DE 289 066, 1914.
19. Patent-Treuhand-Ges. f. Elektrische Glühlampen, DE 420 689, 1923; DE 434 527, 1925.
20. Metallwerk Plansee, DE 720 502, 1929; AT 160 172, 1931.
21. Firth Sterling Steel Corp., US 1 973 428, 1932.
22. E. Rudy, S. Worcester, W. Elkington, *High Temp. High Pressures* **6** (1974) 447 – 454.
23. A. Hara et al., *Sumitomo Denki* **16** (1973) 70 – 77.
24. G. Schumacher, G. Ostermann, *Kobalt* **4** (1973) 69 –81.
25. A. Münster, W. Ruppert, K. Sagel: *Z. Elektrochem.* **57** (1953) 564 – 571. 571 – 579.
26. Metallgesellschaft, DE 1 056 449, 1954.
27. W. F. G. Rix, G. E. Dix, DE 1 521 166, 1966.
28. W. Schedler, *Maschinenmarkt* **80** (1974) 1320 – 1322.
29. W. Schintlmeister, O. Pacher, *Metall (Berlin)* **28** (1974) 690 – 695.
30. K. J. A. Brookes: *World Directory and Handbook of Hardmetals,* 4th ed., Int. Carb. Data, East Barnet, Herts EN48DN, UK, 1987.
31. W. Schedler: *Hartmetalle für den Praktiker,* VDI-Verlag, Düsseldorf 1988.
32. E. Lardner, *Powder Metall.* **8** (1976) 7.
33. P. M. McKenna, *Met. Prog.* **36** (1939) 152 – 155.
34. O. Rüdiger, H. E. Exner, *Powder Metall. Int.* **36** (1939) 152 – 155.
35. P. Rautala, T. Norton: *Proc. 1st Plansee Seminar,* Reutte 1952, pp. 303 – 316.
36. R. Bernhard: *Neue Hütte* **2** (1957) 757 – 761.
37. E. Kny et al., *Werkst. Korros.* **37** (1986) 230 – 235.
38. E. Kny, L. Schmid, *Int. J. of Refract. Hard Met.,* **6** (1978) no. 3, 145 – 148.
39. M. Scherrer, et al., *Schweiz. Med. Wochenschr.* **100** (1970) 2251 – 2255.
40. E. O. Coates, J. H. L. Watson, *Ann. Intern. Med.* **75** (1971) 709.
41. M. Ratzenhofer: *Sitzungsber. Heidelb. Akad. Wiss. Mat. Naturwiss. Kl.* 1975, 24 – 26, 1. Abh.
42. R. Alexandersson: *Arbete och Hälsa,* 1979, 10.
43. *Int. J. Refract. Hard Met.,* June1978, 76 – 77.
44. M. Lichtenstein, F. Bartl, R. Pierce, *Am. Ind. Hyg. Assoc. J.,* Dec.1975, 879 – 885.

Hard Metals → Hard Materials
Hardboard → Wood
Hazard Analysis → Chemical Products: Safety Regulations

Heating Oil

Christian Küchen, Institut für Wirtschaftliche Oelheizung, Hamburg, Germany (Chap. 1)
Knut Spitzmüller, Hamburg, Germany (Chap. 2)

Related Articles → Aviation Turbine Fuels, → Liquefied Petroleum Gas, → Automotive Fuels, and → Oil Refining are separate keywords.

1. Domestic Heating Oil 1
1.1. Composition 2
1.2. Quality Aspects 2
1.3. Production 4
1.3.1. Product and Processing 4
1.3.2. Upgrading with Additives 4
1.4. Uses and Heating Efficiency 6
1.5. Product Specifications 7
1.6. Storage and Transportation 7
1.7. Environmental Protection 12
1.8. Economic Aspects 13
2. Residual Fuel Oil 13
2.1. Composition and Properties 13
2.2. Production 14
2.3. Uses 15
2.4. Quality Specifications, Testing, and Analysis 15
2.5. Storage, Transportation, and Handling 17
2.6. Environmental Protection 18
3. References 18

Domestic heating oil (domestic fuel oil, industrial gas oil) is an extra-light liquid distillate fuel produced from crude oil in refineries. In some countries a certain percentage of fatty acid methyl ester (FAME) can be included. This fuel is used for commercial and industrial processes, and mainly for heating purposes in working areas and domestic accommodation. The fuel is burned in atomizing or vaporizing burners.Domestic heating oil is a very common energy carrier in the domestic heating market of many European countries, in Northern America, and in Japan. As a result of emission legislation and technical requirements different products are available in different countries. Basically heating oils can be divided in three classes:1. Kerosene, 2. Standard heating oil, 3. Low-sulfur heating oil.The large volume and share of industrial gas oil in the total refinery stream (20 to 30 %) makes this product economically significant.

1. Domestic Heating Oil

Domestic heating oil (domestic fuel oil, industrial gas oil, Heizöl EL) is an extra-light liquid distillate fuel produced from crude oil in refineries (see → Oil Refining). In some countries the product can contain a certain percentage of fatty acid methyl ester (FAME). This fuel is used for commercial and industrial processes, and mainly for heating purposes in working areas and domestic accommodations. The fuel is burned in atomizing or vaporizing burners.

The first industrial applications of heating oil were reported at the beginning of the 20th century. In the United States, industrial gas oil was being used to produce hot water and steam for residential heating as early as the 1930s. The major expansion of domestic heating oil and oil firing in Europe has occurred since the end of the 1950s. Today domestic heating oil and natural gas (in some countries electricity) are the most common energy carriers in the domestic heating market of many European countries, in Northern America, and in Japan. Due to emission legislation and technical requirements different products are available in different countries. Basically heating oils can be divided into three classes:

1) Kerosene
2) Standard heating oil
3) Low-sulfur heating oil

The large volume and share of industrial gas oil in the total refinery stream (20 to 30 %) makes this product economically significant.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_617.pub2

1.1. Composition

Domestic heating oil is an extra-light liquid fuel composed almost entirely of paraffinic, naphthenic, and aromatic hydrocarbons. In addition to these hydrocarbons, it contains small quantities of sulfur compounds and still smaller (ppm) amounts of nitrogen and oxygen compounds. Since the beginning of the 21st century special heating oil qualities have been developed, which are almost free of sulfur and nitrogen. The chemical characterization of the fuel is limited to group analysis by classes of compound. The product structure depends on the origin of the crude oil and the processing employed. The elemental composition, however, does not vary significantly. The boiling range is between 160 and 390 °C (kerosene: 300 °C). A typical domestic heating oil has the elemental composition given in Table 1.

Table 1. Elemental composition of a typical domestic heating oil

Element	Concentration
Carbon	86.5 mass %
Hydrogen	13.3 mass %
Sulfur	< 0.2 mass %
Nitrogen	50 – 400 ppm
Oxygen	traces

1.2. Quality Aspects

Quality aspects are mostly described by physical parameters or special test methods for mineral oil products. Basically for all three different qualities (kerosene, standard heating oil, low-sulfur heating oil) the requirements can be classified as follows.

1) Requirements due to national legislation (tax, etc.)
2) Requirements because of the climatic situation
3) Performance parameters with similar test methods and limits in different countries
4) Performance parameters, where national limits vary in a wide range
5) Performance parameters, where test methods and limits have to be developed and/or harmonized

Density. The density depends on temperature and for heating oil is specified at 15 °C. Density alone is of little significance as an indication of the burning characteristics of heating oil in modern atomizing type burners. However, in vaporizing-type burners soot and CO emissions frequently increase when fuels are used whose density and viscosity are close to the specified limits. That is the reason why in some countries (UK and Japan) kerosene with its lower density and boiling range is used as heating fuel for vaporizing-type burners. Density limits vary in a wide range from country to country. When used in conjunction with other properties the density is of value in mass – volume relationships and in calculating the specific energy (heating value, net calorific value) of an oil.

Net Calorific Value. The net or inferior calorific value H_i (MJ/kg; synonymous with lower heating value, see → Gas Production, Chap. 8.2.1.1) does not vary significantly between different heating oils. H_i can be calculated from the density at 15 °C (d_{15}, kg/m^3) and the sulfur content w(S) (mass %):

$$H_i = 52.92 - 11.93 \times \frac{d_{15}}{1000} - 0.29 \times w\,(\mathrm{S})$$

Viscosity. Viscosity limits are specified to help maintain uniform fuel flow in appliances with gravity flow, and to provide satisfactory atomization and constant flow rate through the small nozzles of domestic burners. A viscosity of 2.5 to 3 mm^2/s is needed for good spray formation in conventional atomizing burners. This can be achieved by preheating of the oil to a temperature of 60 °C maximum. At the same time, the oil throughput in a conventional nozzle is reduced by preheating because of thinner film formation in the nozzle itself. The effect of temperature on viscosity is shown in Figure 1.

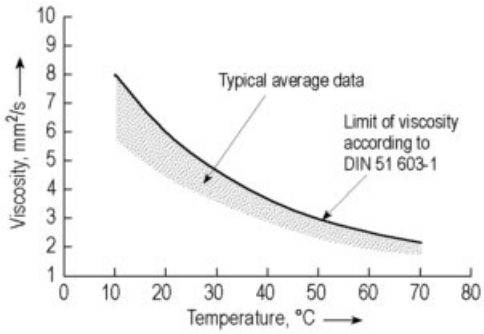

Figure 1. Effect of temperature on viscosity

Distillation. The distillation test shows the volatility of a fuel and the ease with which it can be vaporized. It is of greater significance for oils that are to be burnt in vaporizing type burners than for those burnt in the atomizing type. For example, a low 10 % value assures easy starting of a vaporizing type burner. Nevertheless, a homogeneous distillation curve is necessary to reduce emissions of CO and soot. A typical distillation curve of a domestic heating oil is depicted in Figure 2. Kerosene type fuels however vaporize at significantly lower temperatures.

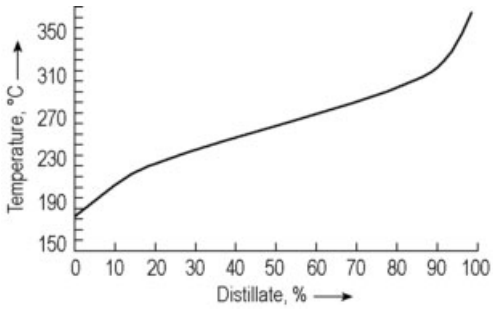

Figure 2. Typical distillation curve of domestic heating oil

Flash Point. The flash point of a fuel is an indication of the maximum temperature at which it can be stored and handled without serious fire hazard. The minimum permissible flash point is usually regulated by law and can therefore differ from country to country.

Sulfur Content. The sulfur content is usually limited by national legislations. The sulfur content in domestic heating oil has been reduced significantly over the years. In Europe an EC directive allows sulfur content to be as high as 0.2 %. This value is reduced to 0.1 % starting from 2008. The amount of low-temperature corrosion products formed in modern energy efficient boilers can be influenced by the sulfur content of the oil. In condensing boilers which is the most efficient burner technology but not very common for domestic heating oil in contrast to natural gas equipment — the pH value of the condensate depends on the sulfur content. To promote these energy saving appliances in many countries a second quality with significantly lower sulfur content has been introduced to the market and standardized.

Low-Temperature Behavior. In heating oil crystallization of high molecular mass *n*-paraffins (waxes) can occur at low ambient temperature. The wax crystals can block fuel pipes and especially the fuel filter. Three test methods are used to assess the low-temperature behavior.

Pour Point. The pour point gives the lowest temperature at which a fuel oil can be stored and still is capable to flow under very weak forces. The determination of the pour point alone is not sufficient to assess the cold flow behavior in practical application.

Cloud Point. The cloud point is the temperature at which heating oil begins to become turbid at a defined cooling rate because of the formation of small wax crystals.

Cold Filter Plugging Point (CFPP). The CFPP is the temperature at which, under defined conditions, a heating oil sample can no longer pass a defined filter within a specified time. The CFPP can be improved by additives (see Section 1.3.2). Neither cloud point nor cold filter plugging point alone can sufficiently describe the cold flow behavior in real installations but the combination of both does.

Carbon Residue. The carbon residue of a fuel is a measure of the carbonaceous material that is left after all volatile fuel components are vaporized in the absence of air. It is a rough approximation of the tendency of a fuel to form deposits in vaporizing burners where the fuel is vaporized in an air-deficient atmosphere. To obtain measurable values of carbon residue in the lighter distillate heating oils, first 90 % of the oil must be removed by distillation, then the carbon residue concentrated in the remaining 10 % bottoms is determined.

Ash. The amount of ash is the quantity of noncombustible material in an oil. The amount of ash in domestic heating oil is almost zero if heating oil directly from the refinery is investigated. However, traces of metals can be found in heating oil, as steel equipment is mainly used for transportation and storage [1].

Water and Sediment. If a domestic heating oil contains appreciable amounts of water and sediment this will lead to fouling of facilities for handling the oil, and difficulties in burner mechanisms will arise. Sediment may accumulate in storage tanks and on filter screens or burner parts, resulting in obstruction to flow of

oil from the tank to the burner. Water in distillate fuels can cause corrosion of tanks and other metal equipment.

Storage Stability. The aging of domestic heating oil is caused by oxidation and radical polymerization reactions, which cause darkening and then sediment formation. The phenomenon of natural aging of middle distillates has been treated in many publications. However, a comprehensive, detailed and consistent theory of the aging mechanism was not found so far. A detailed summary of more than 330 publications dealing with stability of middle distillates is given in [2].

The following items can deteriorate the stability of domestic heating oil:

Heat
Light
Oxygen
Catalytic effects of metals (copper)

Thermal Stability. If the thermal stability of a heating oil is not sufficient, deposits may be formed in hot spots in the burner — usually between the oil preheater and the nozzle. In modern burners with reduced NO_x emissions and low throughput, temperatures significantly higher than 100 °C have been measured after shut-down of the burner. This is due to back radiation from the heating chamber. It is known from many laboratory tests that the aging process (sediment formation) of the oil increases significantly between 100 and 120 °C. To circumvent this, burner manufacturers tend to develop burners with maximum temperature of 100 °C especially after shut-down of the burner. If the oil contains copper ions, the thermal stability of the fuel is reduced by approximately 20 K. A new laboratory test to assess the thermal stability of a fuel has recently been developed in Germany [3].

Lubricity. As sulfur components contribute to the lubricity of the fuel, problems may occur with desulfurized heating oil qualities. Therefore lubricity requirements have been implemented in some standards for low-sulfur heating oil.

1.3. Production

1.3.1. Product and Processing

Since the early 1970s, production of petroleum-based products has been influenced by an increasing range of crude oils used and by a restructuring of refinery production facilities.

There is a persistent trend towards minimizing the amount of residual products and maximizing the quantity of middle distillates and gasolines. This change has become possible as the following equipment has been incorporated in the refinery process:

1) Vacuum distillation equipment,
2) Thermal gas-oil units,
3) Catalytic crackers, and alternatively
4) Hydrocrackers.

To reduce the sulfur content of domestic heating oil, hydrodesulfurization equipment has been installed in many refineries. A simplified flow diagram of the production of domestic heating oil is given in Figure 3. The following distillation cuts are the main components for the blending of heating oil:

1) Kerosene
2) Light gas oil (LGO)
3) Heavy gas oil (HGO)
4) Hydrodesulfurized gas oil (HDS-GO)

The necessity and extent of blending domestic heating oil from different components depends on the crude oil or the mixture of crude oils employed, on the refinery facilities and the product mix that has to be achieved, and on the specifications which have to be met.

1.3.2. Upgrading with Additives

Product quality and application requirements can quite often only be met by employing special purpose additives, which are added to the final product in the required concentrations. Some types of additive are usually introduced into the product in the refinery. The use of additives for quality improvement of heating oil is approved according to many heating oil standards. These additives may be applied in suitable concentrations, provided they are not known to exhibit any harmful side effects. Recommendations, which should help to select suited additives, are published in [4].

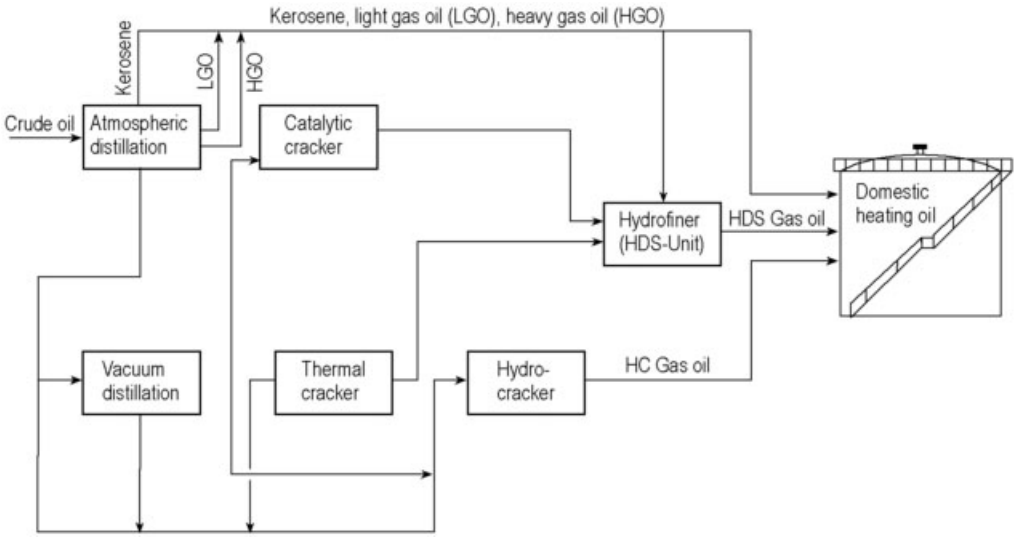

Figure 3. Production of domestic heating oil (simplified flow diagram)

Flow Improvers. Flow improver additives improve the cold flow properties of domestic heating oil. These additives are added to the product stream in many refineries. Ash-free polymers are used as flow improvers. At a temperature slightly higher than the cloud point, the additive forms crystallization nuclei, to which the wax crystals become absorbed. Further growth of the wax crystals is reduced by adsorption of the additive on the crystal surface. Thus, both the crystal size and the tendency to agglomerate are lowered. The small crystals formed pass through the filters of the heating system. Therefore, the CFPP can be lowered significantly by adding flow improvers. The dosage depends on the base fuel and on the CFPP value which is to be achieved. Typical concentrations are 50 – 500 ppm.

Flow improvers have to be added to the fuel at temperatures above the cloud point; addition after crystal formation is not effective.

Metal Deactivators. In special cases, metal deactivators are also added to industrial gas oil to attain long-term storage stability of more than three years. The negative influence of copper [5, 6] on storage stability can be reduced by these additives.

Other types of additives are sometimes introduced into the customer's storage tank but not into the fuel in the refinery.

Combustion Improvers. The purpose of these additives is to suppress carbon formation during combustion. Combustion improvers are predominantly organometallic compounds, ferrocenes have proved particularly effective in domestic heating oil. Although burner technology has advanced considerably with respect to near-stoichiometric combustion with low carbon formation, combustion improvers can still effect and stabilize optimal combustion over long periods. The application of combustion improver additives is not necessary in modern blue flame burners with reduced NO_x emissions.

Additives for Increased Storage Stability. The application of metal deactivators and antioxidants can improve the storage stability of the fuel. The importance of these types of additives is increasing because the average storage time of heating oil in domestic storage tanks increases due to reduced oil consumption (higher efficiency equipment, better insulation of houses) while the size of the storage tanks has been kept constant.

Reodorants. Some additive packages contain reodorants to improve the odor of heating oil. The unpleasant smell is caused particularly by certain sulfur-containing and unsaturated components. If there is no leakage in the oil heating system, the application of reodorants is not necessary.

Biocides. In some cases water can be found in the bottom of storage tanks. Conditions at the interface between water and oil allow the growth of microorganisms. The result is very often sediment formation. The addition of suitable biocides stops the growth of microorganisms but the best way is to remove water regularly from the bottom of the tank.

1.4. Uses and Heating Efficiency

Domestic heating oil is mainly burnt for heating purposes and for the heating of drinking water in burner/boiler systems. The substantial rise in energy prices, the growing awareness of air pollution and the international discussion on greenhouse gases have changed the assessment of heating systems. It is possible to save a lot of energy in the domestic heating sector if modern heating systems replace the old-fashioned systems.

A modern heating system has to fulfill the following demands:

1) Minimal pollutant load on the environment
2) Optimal matching of capacity to demand
3) High efficiency on an annual basis
4) Reliable operation
5) Comparatively attractive product prices
6) Compliance with all standards
7) Long-term availability of the energy source employed

The pollution of the environment could be reduced considerably by development and installation of more efficient heating systems (reduced oil consumption) and of low-emission burners. Generally, the fuel is evaporated to a spray before combustion. In modern low-emission burners, up to 30 % of the flue gas is recycled to the fuel spray. The partial pressure of oxygen and the maximum temperature in the flame is reduced by this technology which allows to meet strict NO_x emission limits. As the oil spray is completely vaporized before it reaches the combustion zone, the formation of soot and particulate matter emission (PM10) can be avoided. Thus, similar advantages as in the combustion of natural gas can be achieved. A good summary of basic knowledge for air-fuel mixing, evaporation and combustion technology with special regard to liquid fuels is given in [7].

Since investments like improved thermal insulation in existing buildings and especially the better insulation standards of new buildings gave the expected energy saving, oil burners designed for lower throughput have been developed. In new buildings the heating system is often dimensioned according to the demand of hot drinking water rather than by the demand of space heating. The minimum power for conventional pressure jet burners is 15 kW (ca. 1.5 L/h) because of the dimension of the nozzle. Other spray formation systems have been developed to allow the construction of burners with reduced throughput. However, not many of these systems have been installed yet.

The boiler efficiency on an annual basis has been improved significantly during the last 25 years by better insulation of the boiler walls, reduced internal cooling loss and by reduction of the flue gas loss. The annual efficiency of a modern low-temperature burner – boiler unit is more than 93 % (basis: net calorific value). The radiation loss and the internal cooling loss account for less than 1 % each, while the flue gas loss is influenced by the flue gas temperature and the amount of excess air.

To maximize the efficiency condensing boilers have been developed. The annual efficiency of condensing boilers is between 100 and 103 % (basis: net calorific value). The market share of condensing boilers differs from country to country but it is steadily increasing. The construction and materials used in condensing boilers are very much dependent on the fuel specification, especially the sulfur content. If the sulfur content is below 50 ppm, it is possible to use similar heat-exchanger materials as in gas condensing boilers. If the sulfur content is higher, the materials used in the condensing areas of the boiler and the flue gas pipe have to be specially resistant to corrosion by acidic condensate. The pH value of typical condensate varies from 2 to 3 depending on the construction of the burner and boiler, the sulfur content of the fuel and the amount of condensate [8]. For low sulfur heating oil the pH value is between 2.8 and 4, which is similar to gas condensing boilers [9].

In some countries the basis for the calculation of the flue gas loss is the gross calorific value of the heating oil, in other countries the calculation is based on the net calorific value. In the latter case, the flue gas loss can be calculated from the

difference in temperature between flue gas and ambient air (ΔT) and the concentration of CO_2 or O_2 in the flue gas:

$$\text{flue gas loss} = \Delta T \left(\frac{0.5}{[CO_2]} + 0.007 \right)$$

or

$$\text{flue gas loss} = \Delta T \left(\frac{0.68}{(21 - [O_2])} + 0.007 \right)$$

These formulae cannot be used for condensing boiler technologies because the latent heat of condensation of the water (produced in combustion) is utilized in addition to the sensible heat. Therefore, the flue gas loss and the efficiency of a boiler should be calculated with the gross or superior calorific value, H_s (synonymous with higher heating value, see → Gas Production, Chap. 8.2.1.1). The flue gas loss can be taken from Figures 4 and 5.

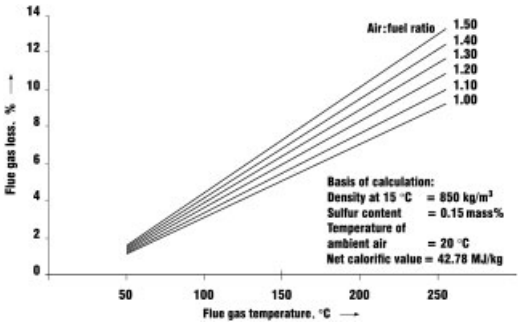

Figure 4. Flue gas loss (basis: net calorific value)

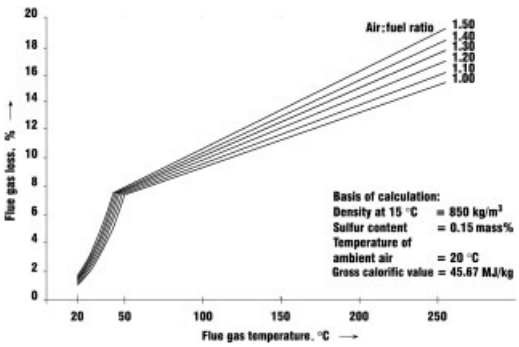

Figure 5. Flue gas loss (basis: gross calorific value)

Domestic heating oil is used in *small power stations* (with Diesel engines) for simultaneous production of electricity and hot water or steam for heating purposes [10]. In research projects and pilot applications domestic heating oil has been used to produce hydrogen and carbon monoxide as feedstock for high temperature fuel cells [11].

1.5. Product Specifications

There is no international or European standard for domestic heating oil. The minimum requirements of domestic heating oil in some national standards are given in Table 2 and 3. Some countries have introduced standards which differ mainly in sulfur content. Other countries have set different limits for cold flow properties in winter and summer. Generally the lighter qualities (e. g. kerosene type fuels, Grade 1 in USA) are specially suited for vaporizing burners.

In 2003 a new European Standard (EN 14213) has been published. This standard specifies requirements and test methods for marketed and delivered fatty acid methyl esters (FAME) to be used either as a heating fuel at 100 % concentration, or as a blending component for the production of heating fuel. However, mixtures have not been standardized yet. An increasing demand for bio-fuels as components of automotive fuels can be observed especially in Europe. As in some countries the logistics of diesel and heating oil are not completely separated, it has to be specified to what extent FAME is allowed to be a component of heating oil.

1.6. Storage and Transportation

A particular characteristic of domestic heating oil is the flash point, established at > 55 °C. This value leads to classification of the fuel in Hazard Class III A (GefStoffV). There are no special safety restrictions concerning storage or transportation of the product.

During transportation and storage domestic heating oil has to be kept at such temperatures that the sedimentation of wax crystals is avoided. Therefore the cold flow behavior of the fuel has to be considered for storage and transportation. This can be a problem especially in underground storage tanks, when the pipes to the burner are not sufficiently insulated. In Germany the storage temperature must be above 0 °C. Lower tem-

Table 2. Product specifications of domestic heating oil in various EU countries (minimum requirements)

	Austria		Belgium		France	Germany		Great Britain		Italy	Sweden
	ÖNORM C1109		NBN T 52-716	NBN EN 590	M 15-008	DIN 51603-1		BS 2869-2		UNI 6579	SS 15 54 10
Issue	2003		2000	2005	1997	2003		1998		1998	1998
Remarks	Low sulfur	Standard		Low sulfur		Low sulfur	Standard	Class C1	Class D		Summer/Winter
Remarks				Diesel EN 590				Kerosene			
Property	Requirements		Requirements		Requirements	Requirements	Requirements	Requirements	Requirements	Requirements	Requirements
Density at 15°C. g/cm^3	max. 0.860		0.820–0.880	0.820–0.845	0.830 – 0.880*	max. 0.860			—	0.815–0.875	stated
Gross calorific value, MJ/kg	—		min. 44.4		—	—			—	—	—
Net calorific value, MJ/kg	min. 42.0		—		—	min. 42.6			—	—	—
Flash point Pensky-Martens, °C	> 55		min. 55	>55	min.55	> 55		min. 43.0	min.56	min. 55	min. 56
Viscosity kinematic at 20 °C, mm^2/s	2.8 – 6.0		max .6.0		max 9.5	max. 6.00			—	—	—
Viscosity kinematic at 40 °C, mm^2/s	—		—	2.0 – 4.5	—	—			1.50 – 5.50	2.0 – 7.4	2.0–4.0/1.5–3.0
Viscosity kinematic at 38 °C, mm^2/s	—		—		—	—			—	—	—
Distillation recovery at 150°C, vol%	—		—		—	—			—	max. 2	—
Distillation recovery at 200°C, vol%	—		—		—	—		min. 15, max. 60	—	—	—
Distillation recovery at 230°C, vol%	—		—		—	—			—	—	min. 10
Distillation recovery at 250°C, vol%	—		max. 65	max. 65	< 65	< 65			max. 65	< 65	—
Distillation recovery at 300°C, vol%	—		—		—	—			—	—	—

Table 2. (continued)

	Austria		Belgium		France	Germany		Great Britain		Italy	Sweden
	ÖNORM C1109		NBN T 52-716	NBN EN 590	M 15-008	DIN 51603-1		BS 2869-2		UNI 6579	SS 15 54 10
Issue	2003		2000	2005	1997	2003		1998		1998	1998
Remarks	Low sulfur	Standard		Low sulfur		Low sulfur	Standard	Class C1	Class D		Summer/Winter
Remarks				Diesel EN 590				Kerosene			
Property	Requirements		Requirements		Requirements	Requirements	Requirements	Requirements	Requirements	Requirements	Requirements
Distillation recovery at 350°C, vol%	min. 90		min. 85	min. 85	min. 85	min. 85			min. 85	min. 85	—
Distillation recovery at 360°C, vol%	—		—	min. 95	—	—			—	—	—
Distillation recovery at 370°C, vol%	—		—		—	—			—	—	min. 95
Distillation temperature, 10 % recovered, °C											
Distillation temperature, 90 % recovered, °C											
Distillation temperature, 95 % recovered, °C											
Pour point, °C	—		—		max. −8	—			—	max. −6	—
Cloud point, °C	—		—		max. 2	max. 3			—	—	max. 0/−22
Cold filter plugging point (CFPP), °C	max. −8		max. −10	max. 0/−15	—	max. −12 (CP 3)/−11 (CP 2)/−10 (CP 1)			max. −4 / −12	max. 0	max. −10/−32
Carbon residue (on 10% residue), %(m/m)	max. 0.15		max. 0.10	max. 0.30		max. 0.3			max. 0.30		max. 0.2
Sulfur content, %(m/m)	—		max. 0.2		max. 0.20		0.0050–0.20	max. 0.04	max. 0.20	max. 0.20	< 0.10

Table 2. (continued)

	Austria		Belgium		France	Germany		Great Britain		Italy	Sweden
	ÖNORM C1109		NBN T 52-716	NBN EN 590	M 15-008	DIN 51603-1		BS 2869-2		UNI 6579	SS 15 54 10
Issue	2003		2000	2005	1997	2003		1998		1998	1998
Remarks	Low sulfur	Standard		Low sulfur		Low sulfur	Standard	Class C1	Class D		Summer/Winter
Remarks				Diesel EN 590				Kerosene			
Property	Requirements		Requirements		Requirements	Requirements	Requirements	Requirements	Requirements	Requirements	Requirements
Sulfur content, mg/kg	max. 50	50 – 1000		max. 50		max. 50					
Water content, mg/kg	max. 200		max. 200	max. 200	max. 200	max. 200			max. 200	—	max. 200
Particulate contamination, mg/L	—		—		—	—			—	—	—
Total contamination, mg/kg	max. 24		max. 24	max. 24	—	max. 24			max. 24	—	max. 24
Ash, %(m/m)	max. 0.01		—	max. 0.01	—	max. 0.01			max. 0.01	max.0.01	max. 0.01
Neutralisation number, mg KOH/100 mL	—		—		—	—			—	—	
Copper corrosion	—		max. 1	class 1	—	—		class 1	class 1	—	max. 1
Sediment, %(m/m)	—		—		—	—			max. 0.01	—	max. 0.05
Mechanical contamination, % (m/m)	—		—		—	—			—		—
Water and sediment, vol%	—		—		max. 0.10	—			—	max. 0.05	—
Color	red		—		red	—			—	—	—
Color ASTM D 1500	—		—		—	—			—		—
Oxidation stability, mg/100 mL	—		—	max. 2.5	—	—			—	—	—
Cetane index	—		—		—	—			—	—	—
Organic chlorine, mg/kg	—		—		—	—			—	—	max. 6
PCB, mg/kg	—		—		—	—			—	—	max. 2
Nitrogen, mg/kg	—		—		—	—			—	—	stated

* See

Table 3. Product specifications of domestic heating oil in some non-Eu countries

	Japan	Norway	Switzerland		USA	
Issue	JIS	Gasoil No.3	SN 181160-2		ASTM D396	
Remarks		2002	1999			
Remarks	Kerosene	Production Standard	CH-Quality	Euro-Quality	Grade 1	Grade 2
Property	Requirements	Requirements	Requirements	Requirements	Requirements	Requirements
Density at 15°C. g/cm^3		0.820–0.860	0.820–0.855	0.820–0.860	max. 0.850	max. 0.876
Gross calorific value, MJ/kg		—	min. 45	—		
Net calorific value, MJ/kg		—	min. 42.6	min. 42.6		
Flash point Pensky-Martens, °C	min. 40	min. 60	min. 55	min. 55	min. 38	min. 38
Viscosity kinematic at 20 °C, mm^2/s		—	—	—		
Viscosity kinematic at 40 °C, mm^2/s		2.0 – 4.0	2 – 4	max. 4	1.3–2.1	1.9–3.4
Viscosity kinematic at 38 °C, mm^2/s		—	—	—		
Distillation recovery at 150°C, vol%		—	—	—		
Distillation recovery at 200°C, vol%		—	—	—		
Distillation recovery at 230°C, vol%		—	—	—		
Distillation recovery at 250°C, vol%		max.65	—	max. 65		
Distillation recovery at 300°C, vol%		—	—	—		
Distillation recovery at 350°C, vol%		min. 85	—	min. 85		
Distillation recovery at 360°C, vol%		—	min. 90	—		
Distillation recovery at 370°C, vol%		min. 95	—	—		
Distillation temperature, 10 % recovered, °C					max. 215	
Distillation temperature, 90 % recovered, °C					max. 288	282 – 338
Distillation temperature, 95 % recovered, °C	max. 270					
Pour point, °C		—	—	—	max. −18	max. −6
Cloud point, °C		max. 0	max. −2	max. 3		
Cold filter plugging point (CFPP), °C		max. −11	max. −10	−12/−11/−10		
Carbon residue (on 10% residue), %(m/m)		max. 0.15	max. 0.05	max. 0.05		
Sulfur content, %(m/m)		max. 0.0499	max. 0.05	max. 0.2		

Table 3. (continued)

	Japan	Norway	Switzerland		USA	
Issue	JIS	Gasoil No.3	SN 181160-2		ASTM D396	
Remarks		2002	1999			
Remarks	Kerosene	Production Standard	CH-Quality	Euro-Quality	Grade 1	Grade 2
Property	Requirements	Requirements	Requirements	Requirements	Requirements	Requirements
Sulfur content, mg/kg	max. 80					
Water content, mg/kg		max. 100	max. 150	max. 200		
Particulate contamination, mg/L		max.8.0	—	—		
Total contamination, mg/kg		—	max. 12	max. 24		
Ash, %(m/m)		max.0.01	max. 0.01	max. 0.01		
Neutralisation number, mg KOH/100 mL		max.0.5	—	—		
Copper corrosion		1	—	—		
Sediment, %(m/m)		—	—	—		
Mechanical contamination, % (m/m)		—	—	—		
Water and sediment, vol%		—	—	—		
Color		—	—	—		
Color ASTM D 1500		<=1.5	—	—		
Oxidation stability, mg/100 mL		max.2.5	—	—		
Cetane index		min. 45	—	—		
Organic chlorine, mg/kg		—	—	—		
PCB, mg/kg		—	—	—		
Nitrogen, mg/kg		—	ÖKO-oil max. 100			

peratures are allowed if a special fuel with satisfactory low temperature-properties is used.

1.7. Environmental Protection

The importance of the *sulfur content* of domestic heating oil for the consumer lies in the emission of sulfur dioxide and thus the problem of environmental pollution. In the EU, the sulfur content of domestic heating oil is now limited to 0.2 mass % (Austria 0.1 mass %). The EU-sulfur directives (1999/32/EC) requires to reduce that limit to 0.1 mass % by 2008. In some countries or regions there are special fuels available with sulfur content < 0.05 mass % or even < 50 mg/kg (0.005 mass %, low-sulfur heating oil). In general the average sulfur content has been decreasing continuously.

Further reduction of the sulfur content of the fuel and thus reduced SO_2 emissions when burning the fuel can be accomplished with desulfurization processes, if these facilities are available in the refinery. However, removal of sulfur requires hydrogen and is performed at pressures of 2.5 to 7 MPa with the result of increased energy consumption and CO_2 emissions. On the other hand this is more than compensated by higher annual efficiencies of low temperature or condensing boilers, because fouling of the heat exchanger surfaces is reduced.

The *nitrogen content* of domestic heating oil is not specified in national standards. It depends on the crude oil and the processing employed. Some desulfurization processes may also reduce the nitrogen content of the fuel. Typically, the nitrogen content is between 50 and 200 mg/kg, occasionally it can be as high as 400 mg/kg. For

low-sulfur heating oil or kerosene the nitrogen content often is below 50 mg/kg. The fuel nitrogen contributes to the emissions of nitrogen oxides (NO_x) during combustion. 50 to 100 % of the fuel nitrogen are converted to NO_x depending on the type of burner [12].

Typical emission levels for oil heating (Table 4) are dependent on the technology employed [13]. As could be expected, modern boiler and burner technology results in lower emission values of CO, NO_x, and particulate matter (PM10) than older technologies.

The emissions of components that are traditionally not determined have been investigated [14]. The study encompasses emissions of N_2O, methane, benzene, toluene, xylene, and PCDD/PCDF (polychlorinated dibenzo-*p*-dioxin/dibenzofuran). The emission levels of these components are very low. Oil-fired systems do not contribute significantly to total emissions of these components.

The emissions of NO_x and SO_2 could be reduced by secondary treatment of the flue gas. The technology for that flue gas treatment is available on the market but is usually not used for domestic installations, because the emission limits can be met by modern combustion systems alone.

Recently particulate matter emissions have become more important. It is well accepted that with modern combustion technology in which the fuel is evaporated before being combusted the particulate matter emissions are similar to those of gas appliances.

1.8. Economic Aspects

The consumption of heating oil in selected countries is given in Table 5 [17]. The sizes of the US market and the Japanese market are similar to the market in Germany. Market data are available on www.euro-fuel.net. Data for the US market can be taken from [15].

The consumption of heating oil has been decreasing in many countries. This is caused by better insulation of houses and more efficient boiler systems and by the fact that in some countries oil heating systems have been switched to other systems. In many countries natural gas is the main competitor, in other countries electrical heat-pumps are gaining market share. In contrast, in some countries the number of oil installations has increased by replacing direct electrical heating.

2. Residual Fuel Oil

Residual fuel oil (RFO) is a refinery product derived from petroleum crudes. Since the introduction of oil conversion technologies in the mid-1960s most RFO has been produced by blending vacuum distillation residues with residues from thermal crackers (e.g., visbreakers). Material with a marketable viscosity range results from further addition of catcracked components as well as gas oil fractions of varying density derived from thermal processing. Heavy oils derived from lignite and coal will not be discussed here because these are only of local interest.

2.1. Composition and Properties

Residual fuel oil is an extremely complex material as a consequence of its origin in the cracking process. In addition to paraffinic, naphthenic, and aromatic hydrocarbons (some with molecular masses approaching 20 000), RFO contains nitrogen and oxygen compounds as well as organometallic substances that contribute to its ash content. A typical elemental analysis shows a distribution (in wt %) of 87.5 % C, 10.6 % H, 1.8 % S, and 0.1 % ash (C : H ratio 8.25); these values may be assumed valid for performing combustion calculations.

RFOs are colloidal systems of highly dispersed macromolecular components (asphaltenes) in solution equilibrium with a continuous oil phase (the maltenes). Just like the maltenes, asphaltenes may be regarded as natural fuel components, although with a much higher C : H ratio. Most of the metallic ash components are present in oil-soluble form, either as porphyrins or as part of the asphaltene phase. Metal atoms are in most cases bonded to the heteroelements sulfur, oxygen, and nitrogen.

Crude oil is separated into its various fractions by fractional distillation, so RFO tends to be enriched in metals. The most important ash components are vanadium, sodium, and nickel, as well as calcium and magnesium (all in the

Table 4. Emissions of oil heating systems (atomizing burners)

Emission	Average data	Modern technology	Best available technology
Carbon monoxide, mg/MJ[a]	7 – 76 [13]	7	1.4
Nitrogen oxides[b], mg/MJ[a]	29 – 50 [13]	30	17
Unburnt hydrocarbons, mg/MJ[a]	1.4	< 1	< 1
PM10, mg/MJ [a]	0.05 – 1.2	< 0.3	< 0.1
Sulfur oxides[c], mg/MJ[a]	70	70	70
Carbon dioxide, kg/MJ[a]	0.074	0.074	0.074

[a] Net calorific value of fuel.
[b] Nitrogen content of reference fuel = 140 mg/kg.
[c] Depending on sulfur content of fuel (here 0.15 mass %).

Table 5. Consumption of domestic heating oil in selected countries (10^3 t)

Country	1996	1997	1998	1999	2000	2001	2002	2003	2004
Germany	38 420	36 292	34 631	29 633	27 875	31 888	28 544	28 064	25 424
Austria	1956	1907	1897	1852	1592	1961	1761	1979	1690
Belgium	6479	6490	6295	5825	5418	5736	4999	5538	5497
Denmark	2021	1880	1806	1778	1669	1727	1453	1462	1397
Spain	8130	7587	8310	9468	9207	9628	9534	10 235	10 843
Finland	2446	2361	2485	2435	2150	2367	2285	2372	2180
France	17 262	16 725	17 303	19 960	15 583	17 257	15 597	16 022	16 042
Greece	3283	2784	3581	3638	3961	4379	4528	5057	4788
Ireland	1324	1383	1556	1710	1965	1850	1565	1400	1368
Italy	7027	6820	6452	6627	6289	5797	5223	5559	5333
Luxemburg	366	368	376	355	301	335	311	303	319
Holland	1675	1660	1602	1604	1246	1136	1153	1209	1191
Portugal	783	1074	1299	1048	1266	1223	1070	984	1177
UK	7631	7326	7245	6682	6583	6598	6129	6142	4927
Sweden	3105	2214	2175	2313	2969	1950	3129	631	1588
EU (15)	101 928	96 871	96 993	91 928	88 074	94 032	87 271	86 957	83 764
Iceland	331	317	345	363	408	324	331	326	404
Norway	2073	2003	2026	2114	1791	1810	1812	2062	1870
Switzerland	4939	5388	5315	4734	4299	5274	4856	4562	4454
Turkey	2912	3172	3054	3475	3723	3258	3258	3293	3436
Total	112 183	107 751	107 735	102 614	98 295	104 698	97 528	97 200	93 928

Table 6. Metal content of residual fuel oil (in mg per kilogram atmospheric residue) as a function of source

Crude oil source	Vanadium	Sodium	Nickel
Nigeria/Forcados	1	8	10
North Sea/Brent	6	1	3
Libya/Brega	6	2	10
Middle East/Oman	30	5	10
Middle East/Arab Heavy	75	15	27
Venezuela/Tia Juana Pesado	426	22	56

ppm range). The metal content depends on the geographic origin of the crude oil (see Table 6).

Numerous other metals are present in trace amounts (µg/kg), but may be ignored for commercial applications. Most metal analyses today result from X-ray fluorescence determinations performed directly on the RFO.

The principal distinction between RFO and domestic heating oil is that the former must be preheated prior to use. Compared to coal, RFO offers the advantages of relatively straightforward handling, favorable burning characteristics, and a high specific heat content. Furthermore, the relatively low ash content of RFO largely eliminates problems related to the disposal of ash and filter dust.

2.2. Production

Depending on the proposed application, RFO components as described in Section 2.1 are blended in the refinery to meet appropriate product specifications. The characteristic that most

clearly distinguishes one grade of RFO from another is nominal viscosity at 100 °C (or 80 °C in English-speaking countries). Rheologic factors preclude the use of 50 °C viscosity data.

Standard grades of RFO for steam boilers and industrial furnaces have viscosities of 20 – 50 mm^2/s, whereas the marine market prefers material in the 10 – 55 mm^2/s range (all at 100 °C). Viscosities between 100 and 3000 mm^2/s are typical for oil gasification units and for internal use within a refinery.

Economic considerations are responsible for the fact that suppliers seldom employ additives in RFO. Additive addition remains the responsibility of the end user, and is limited to metalcontaining, oil-soluble substances that improve combustion characteristics (e.g., particulate control by suppressing soot formation) and minimize low-temperature corrosion at cold junctions (through dew-point reduction). Compounds of magnesium and other alkaline earth metals are occasionally added to raise the melting point of the ash.

2.3. Uses

The most important uses of RFO have already been noted in Section 2.2. On a world basis, most of the product is used in steam generators and industrial furnaces (ca. 400×10^6 t/a) or marine applications (85×10^6 t/a). Currently, RFO represents ca. 20 – 22 % of the "cut of the barrel" (based on data from the Western world), though demand in the industrialized countries is diminishing. Interest in RFO as a tuyere-injection "reducing agent" for blast furnaces varies according to the coal – residual oil cost ratio, because from a technological standpoint the two materials can be regarded as equivalent. Increasing application of oil conversion technologies coupled with environmental factors suggests that consumption of RFO in the industrialized nations will continue to decline relative to so-called "white products" (middle distillates and gasoline). The following market share figures (RFO as a fraction of total oil consumption) for 1988 provide additional insight: United States, 9 %; Federal Republic of Germany, 10 %; France, 13 %; United Kingdom, 29 %; Japan 31 %.

Usually RFO is combusted in furnaces with any of several types of burners that are distinguished by the method employed for atomization. The most important types are:

1) Jet atomizers of the nozzle or lance type (operating pressures up to 4.0 MPa, atomizing viscosities of ca. 10 mm^2/s, preheating to 140 °C)
2) Rotary cup burners (0.5 MPa, ca. 15 – 25 mm^2/s, 110 °C)
3) Air or steam atomizers (operating parameters analogous to rotary cup burners, air or steam flows of 5 – 7 wt % relative to fuel throughput)

The choice of burner type depends on the design of the combustion facility. Key factors include adequate residence time for combustion of the asphaltenes and low specific heat load, together with near-stoichiometric combustion to minimize NO_x emission.

Some marine users of RFO continue to employ steam boiler – turbine combinations, although diesel engines now account for 96 % of the market worldwide, and their use continues to increase. Diesel engines are clearly more economical to operate than turbines, especially under partial load conditions. With the aid of turbocharging and waste heat recovery, a maximum thermal efficiency of ca. 50 % can now be achieved at a specific fuel consumption as low as 150 g/kWh. Today's RFO can be utilized without difficulty in both 4-stroke trunk piston engines (minimum output: 160 kW per cylinder at 750 min^{-1}) and 2-stroke cross head engines (maximum output: 3500 kW/cylinder at 80 min^{-1}). Pretreatment of RFO for use as a motor fuel entails centrifugal separation (purifier/clarifier) and fine filtering (10 – 20 µm). In addition, effective preheating is necessary (to 160 °C with an injection viscosity of 9 – 12 mm^2/s and an injection pressure ahead of the jet of as much as 140 MPa). Taken together, these measures assure optimal combustion as well as adequate protection against abrasive wear.

2.4. Quality Specifications, Testing, and Analysis

For technological and economic reasons reliable standards must be available for RFO classification, and product specifications must be clear and acceptable to all parties involved. To date,

standards have been developed for residual fuels for *industrial uses* only on a national basis. The most important distinctions between various product grades relate to viscosity, density, flash point, and sulfur and ash content. Standards applicable within the United States are covered by ASTM-D 396–84; for the United Kingdom, BS 2869; for the Federal Republic of Germany, DIN 51063 (part 3 and, in the future, part 5, which specifies a 1 % sulfur content); for Switzerland, SN 181160/2–86; for Austria, öNORM C 1108–84; and for Japan, JIS K K 2205.

Standards for RFO intended for marine use are given in → Marine Fuels.

Analytical values for RFO vary widely depending on applicable comprehensive standards and local requirements. They are a function not only of refinery processing but also of the origin of the crude oil. Considerable progress toward uniformity has been achieved in recent years, largely because of increasing awareness and knowledge with respect to environmental protection.

Density. Density depends on temperature, thus RFO density data are uniformly reported at 15 °C to facilitate conversion of volume into the commercial measure mass (in tonnes). Industrial residual fuels are not normally subject to a density limit (typical values: ca. 995 – 1020 kg/m^3). However, density is important for marine applications (maximum 925 – 1010 kg/m^3) because it is a limiting factor in the effective use of centrifugal separators. Density is also related to aromatic content, which in turn affects ignition characteristics in a combustion or diesel engine. Further refinement of ISO 8217 is expected to result in incorporation of a density limit, particularly for material with a viscosity $<$80 mm^2/s.

Kinematic Viscosity. This parameter also depends on temperature. The reported "nominal viscosity" should be determined at a temperature of at least 80 °C, preferably 100 °C. Kinematic viscosity is important in both the industrial and marine sectors because it largely determines the level of preheating required for RFO transfer and for its input into a burner or diesel engine. Reliable and economic operation requires the availability of appropriate viscosity measurement and control systems.

Flash Point (Closed Cup). Flash point information is important in ensuring safe transport and storage. In the case of residual fuels, flash point determinations above 65 °C are normally carried out according to ASTM-D 93 (Pensky –Martens). Flash point specifications in the various norms range from a minimum of 60 °C (marine fuels) to 70 – 100 °C (inland applications).

Sulfur Content. Sulfur content is largely a function of the origin of the corresponding crude oil. Residual fuel oil desulfurization at the refinery presents serious technological problems, and is currently not economical. The sulfur restrictions imposed by various countries must usually be met by proper choice of crude oil. Low-sulfur crude is available only from a few sources (North Sea, North Africa), and for this reason has a higher market value than ordinary crude. The most important commercial RFO gradations with respect to sulfur content are 1.0, 1.5, 2.0, and 2.5 % for the industrial market, and up to 5.0 % for marine applications.

The principal problems posed by sulfur arise from its combustion products, sulfur dioxide (an environmental hazard) and sulfur trioxide (conversion rate to sulfur trioxide is 1 – 3 % because of combustion with surplus oxygen). Sulfur trioxide increases the sulfuric acid dew point in the off-gases. The resulting condensate, together with water, causes low-temperature corrosion of metallic surfaces.

Conradson Carbon Residue (CCR): Asphaltene Content. The CCR test provides insight into the tendency of a fuel to coke. In the case of RFO, this test is expected to be replaced in the near future by the Micro Carbon Residue Test (ASTM D 4530) because of the relatively wide variation in CCR test results (14 – 20 %). However, even with the expected improvement in precision and a considerable reduction in effort, the results are likely to remain of marginal value. This is because the test methods themselves are based on low-temperature *carbonization*, whereas the process that occurs in a motor or a burner is an *overstoichiometric combustion*, an entirely different phenomenon.

Asphaltene content is more significant, although it is not the subject of any standards. By definition, asphaltenes are substances that are precipitated from an oil sample by addition of a 30-fold excess of *n*-heptane, and that are soluble in pure toluene. Asphaltenes have a molecular mass in the range of 5000 – 20 000, with

particle sizes of 1 – 10 nm. Asphaltenes often agglomerate because of polar forces.

Ash Content: Catalyst Fines. The ash from a residual fuel is defined as the oxide residue remaining after complete conversion of the fuel (ashing temperature 775 ± 25 °C). The ash content of commercial RFO is typically between 0.02 and 0.1 wt %. The most important ash components from a technical standpoint are those containing vanadium, sodium, nickel, aluminum, and silicon. Sodium reacts with sulfur to produce sodium sulfate, Na_2SO_4 (*mp* 880 °C). Vanadium oxidizes to vanadium pentoxide, V_2O_5 (*mp* 670 °C). A eutectic mixture of these two compounds melts at 565 °C. In addition, provided the conditions are appropriate (e.g., a critical Na : V molar ratio in the fuel of 1 : 3), Na_2O and V_2O_5 can combine to give vanadyl vanadates. The most commonly observed forms are vanadate I ($Na_2O \cdot V_2O_4 \cdot 5\ V_2O_5$, *mp* 559 °C) and vanadate II ($5\ Na_2O \cdot V_2O_4 \cdot 11\ V_2O_5$, *mp* 577 °C). The combination of the aforementioned materials leads to deposits on the heat transfer sections of boilers and in the combustion chambers of diesel engines or turbo-chargers. Vanadate melts cause high-temperature corrosion of such thermally stressed components as superheater surfaces and exhaust valves. Countermeasures include careful choice of construction materials and appropriate component design (e.g., maximum steam temperatures in RFO-fired boilers of 540 °C, exhaust valve cooling).

Small amounts of aluminum and silicon are introduced into RFO in the form of fine zeolite residues ("cat fines") during catalytic cracking. With a particle-size distribution up to ca. 20 µm coupled with distinct oxidizing characteristics these materials are responsible for significant amounts of abrasive wear, especially in diesel motors. Nickel residues are converted by combustion into both water-soluble (nickel sulfite, nickel sulfate) and water-insoluble substances (nickel oxide, nickel carbonyl).

Limits with respect to the metals vanadium and aluminum have now been added to the norms for RFO. ISO 8217 specifies a maximum vanadium content of 150 – 600 mg/kg depending on the product classification, and a maximum aluminum content (regarded as an indicator of the level of cat fines) of 30 mg/kg.

Net and Gross Calorific Value Net calorific values of most residual fuels are in the range of 39.82 – 41.53 MJ/kg (11.06 – 11.54 kWh/kg). Systematic investigations by Marder have shown a direct relationship between the net calorific value of a petroleum fuel and its density and sulfur content. Net calorific values can be calculated with sufficient accuracy by the use of the following equations (modified Marder equations, neglecting the contribution of ash and based on anhydrous material):

$$\begin{aligned} \mathrm{LHV} &= 52.919 - 11.930 \times 10^{-3} \\ &\quad \times d^{15} - 0.293 \times \mathrm{S}\ (\mathrm{MJ/kg}) \\ &= 14.700 - 3.314 \times 10^{-3} \\ &\quad \times d^{15} - 0.081 \times \mathrm{S}\ (\mathrm{kWh/kg}) \end{aligned}$$

where d^{15} is the density at 15 °C (kg/m^3) and S is sulfur content (wt %)

Sediment Content. Sediments consist of inorganic or organic foreign matter remaining after extraction of RFO with boiling toluene (ISO 3735), *Sediment by Extraction*). Apart from inorganic materials, sediment arises through aging or from polymerization or condensation of various oil components. Sediment content is a subject of numerous standards; maximum observed values for RFO are ca. 0.25 wt %. Just like the asphaltenes, sediments can be the cause of filter clogging.

Water Content. RFO is essentially anhydrous when it leaves the refinery. Nevertheless, multiple transfers at sea and the formation of condensate during storage may result in product with a water content of 0.1 – 0.5 %. This can cause foaming and spattering when the material is heated above 100 °C and can also lead to burner problems.

2.5. Storage, Transportation, and Handling

From a legal standpoint and with respect to the risk of fire or water contamination, RFO is not a dangerous substance (flash point 70 –100 °C, ignition temperature > 220 °C, explosive limits in air 0.5 – 6.5 vol %). Thus, RFO can be stored in single-walled tanks with a capacity of 50 – 50 000 m^3 without special precautions (e.g., fire extinguishing equipment, protective dikes). For safety reasons the storage temperature should be

kept 10 – 15 °C below the flash point. Hydrogen sulfide may accumulate in the vapor space of an RFO storage tank, so vapor inhalation and skin contact should be avoided. Cleanup should be effected with ordinary cleansing agents; solvents should not be employed.

Water transport (ocean, inland waterways) may utilize bulk carrier or barge (500 – 1000 t); land transport is normally by railroad tank car (30 – 50 t) or tank truck (24 – 28 t). Transport temperature must be maintained high enough so that the maximum pumping viscosity of 600 –800 mm^2/s is not exceeded (e.g., at least 45 °C with a nominal viscosity at 100 °C of 50 mm^2/s). Unloading normally entails a centrifugal or screw-type pump, some RFO tank trucks are equipped to employ pressure supplied from exhaust gas either with or without an auxiliary compressor. On-site transfer generally involves either permanently installed pipelines or oil-tight hoses with appropriate couplings.

2.6. Environmental Protection

Residual fuel oil itself is not a source of environmental problems, but its combustion products are. The relationship between RFO components and air contamination contribution may be deduced from Table 7, which presents a schematic overview of the combustion process in the absence of flue gas treatment.

Table 7. Environmental consequences of fuel oil combustion

Starting materials	Reaction products	Product concentration in flue gas (based on 3 % O_2)
Metals, solid inorganic foreign matter (0.1 %)	ash	particulates, 50 – 200 mg/m^3
Carbon (ca. 87.5 %)	soot	particulates, 50 – 200 mg/m^3
Carbon – combustion air	CO	50 – 100 mg/m^3
Carbon – combustion air	CO_2	inert, though contributing to the "greenhouse effect"
Hydrogen – combustion air (ca. 10.6 %)	H_2O	inert, though contributing to the "greenhouse effect"
Nitrogen – combustion air (0.2 – 0.5 %)	95 % NO	
	5 % NO_2	NO_x; Σ 700 – 1100 mg/m^3
Sulfur – combustion air	97 % SO_2	
	3 % SO_3	SO_x; Σ 1700 – 8500 mg/m^3

Environmental considerations have led to worldwide efforts to enact legislation for drastically reducing the emission of harmful materials. The resulting emission limits can be met by consumers of large amounts of RFO only with the aid of flue gas treatment facilities, installation of which is proceeding rapidly. The most common secondary treatment measures at steam generating plants include:

1) Electrotatic precipitators (particulate removal)
2) Desulfurization plants (DESOX: wet gypsum method of SO_x removal)
3) Denitrification plants (DENOX: selected catalytic design for NO_x removal)

These secondary treatment methods may be supplemented by primary treatment directly at the burner. This is intended to minimize the formation of hazardous materials—especially CO and NO_x—directly at their source. Staged air/fuel combustion plays an important role in primary treatment, as does flue gas recycling ("low NO_x burners").

The EEC recommendation of November 24, 1988 sets the following limits with respect to liquid RFOs as a function of facility size:

SO_x = 500 – 1700 mg/m^3
Particulates = 50 mg/m^3
NO_x = 450 mg/m^3

Desulfurization is mandatory above 400 MW thermal output (new construction), EEC member states are expected to adopt these guidelines by July 1, 1990. The guidelines are intended to apply to existing as well as new facilities, and should assist the states in meeting their obligation to reduce total emissions. The United States, Japan, and the Scandinavian countries have adopted pollution control legislation in which targets are based upon local emission data

3. References

General References

1. H. M. Spiers, J. W. Rose, J. R. Cooper: *Technical Data on Fuel*, 7th ed. British National Committee, Scottish Academic Press, Edinburgh 1977.
2. W. Hansen: *Ölfeuerungen-Brennstoff, Technische Einrichtungen, Anwendungen*, 2nd ed., Springer Verlag, Berlin, Heidelberg, New York 1970.

3. Deutsche Babcock AG, H. P. Niepenberg (ed.): *Industrie-Ölfeuerungen*, Oberhausen, Federal Republic of Germany, 1968.
4. G. H. Clark: *Industrial and Marine Fuels*, Butterworths, London 1988.
5. S. H. Kägler (ed.): *Neue Mineralölanalyse*, vol. 1, Spektroskopie, 2nd ed., Hüthig Verlag, Heidelberg 1987.
6. C. Zerbe: *Mineralöle und verwandte Produkte*, 2nd ed., Springer Verlag, Berlin, Heidelberg 1969.
7. Shell International Trading Co., *Shell Marine Service, Know Your Fuels, Shell Marine Fuel Specifications*, London 1987
8. *Brennstoff, Wärme, Kraft*, VDI Verlag, Düsseldorf.
9. *Die Industriefeuerung*, H. P. Niepenberg (ed.), Vulkan Verlag, Essen.
10. Reports of the VGB in: *VGB Kraftwerkstechnik* Technische Vereinigung der Großkraftwerksbetreiber, Essen.
11. Institut für Wirtschaftliche Oelheizung e. V. (IWO): "Heizöl EL Produkt und Anwendung,"Hamburg 2003. www.iwo.de/sites/iwo/iwo_fachbroschueren/pdf/HEL_Broschuere_06_11_2003.pdf
12. P. Derst: "Alterungsmechanismen bei Mitteldestillaten. Eine Literaturrecherche, "DGMK Forschungsbericht 484, Hamburg 1995.
13. DIN 51371 (in preparation), Bestimmung der thermischen Stabilität von Heizöl EL
14. DGMK-Forschungsbericht 646, Kriterienkatalog für Heizöl EL Additive, DGMK 2005.
15. J. Lohmar, S. Laurs, O. Akin, P. Beiss, *Sanitär und Heizungstechnik* **60** (1995) no. 12, 55 – 63.
16. J. Lohmar, S. Laurs, O. Akin, P. Beiss, *Sanitär und Heizungstechnik* **61** (1996) no. 1, 54 – 74.
17. K. Lucka: "Beschreibung der Reaktionsphänomene in Brennstoff-Luft-Gemischen unterhalb der Selbstzündung," Habilitationsschrift (in preparation), RWTH Aachen 2006.
18. C. Küchen, *Wärmetechnik* **39** (1994) 590 – 601.
19. DGMK-Forschungsbericht 601, Ermittlung eines technisch begründeten Grenzwertes für Schwefelgehalte im Heizül EL zur Förderung des Einsatzes von Brennwerttechnologie in Ölheizungsanlagen, DGMK 2002.
20. *Handbuch "Kraft – Wärme-Kopplung*," VDI – GET.
21. L. Hartmann, K. Lucka, H. Köhne: "Mixture preparation by cool flames for diesel reforming technologies," *J. Power Sources* **118** (2003) 286–297.
22. U. Feisst, *Wärmetechnik* **35** (1990) 107 – 122.
23. C. Hübner et al., *Erdöl Erdgas Kohle Petrochem.* **112** (1996) 170 – 174.
24. F. Pfeiffer, M. Struschka, G. Baumbach, *Erdöl Erdgas Kohle Petrochem.* **113** (1997) 137 – 140.
25. Energy Information Administration, Official Energy Statistics of the U.S. Government http://tonto.eia.doe.gov/dnav/pet/pet_cons_821use_dcu_nus_a.htm.

Hexamethylenediamine

ROBERT A. SMILEY, E. I. du Pont de Nemours & Company, Wilmington, Delaware 19898, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 2
5. **Quality Specifications and Analysis** . . . 3
6. **Storage and Transportation** 3
7. **Economic Aspects** 4
8. **Toxicology and Occupational Health** . . 4
9. **References** 4

1. Introduction

CURTIUS published the first synthesis of hexamethylenediamine [*124-09-4*], 1,6-hexanediamine, 1,6-diaminohexane, $H_2N(CH_2)_6NH_2$, M_r 116.21, in 1900 [1]. The compound did not become important industrially until the late 1930s after the pioneering work at Du Pont on synthetic polymers [2]. One of these polymers was *nylon* (→ Polyamides), a polyamide produced from adipic acid [*124-04-9*] (→ Adipic Acid) and hexamethylenediamine. Du Pont subsequently began to produce hexamethylenediamine for manufacture of a synthetic fiber which they named nylon 66 (to distinguish it from nylon 6, a polyamide based on caprolactam). Nylon 66 rapidly gained acceptance for use in hosiery, lingerie, and other types of clothing, for military applications, and later for molded plastic parts. Du Pont was quickly joined by ICI in the production of nylon 66.

The commercial success of nylon led to expanded production of hexamethylenediamine in the postwar years, but it remained a captive product for nylon manufacture and was not generally for sale. This situation changed gradually, particularly after the development of light-stable polyurethanes based on aliphatic diisocyanates and the introduction by Bayer of a polyisocyanate (Desmodur N) made from hexamethylenediamine [3]. Hexamethylenediamine is now a commercial product available as an aqueous solution in several strengths, as anhydrous material in drums, or as a melt in insulated tank cars.

2. Physical Properties [4,5]

Hexamethylenediamine is a colorless solid with a typical fishlike amine odor. It is very soluble in water, soluble in alcohols and aromatic solvents, and poorly soluble in aliphatic hydrocarbons. The following shows its vapor pressure as a function of temperature [6]:

t, °C	200.6	181.2	154.3	135.1	117.7
p, kPa	101.3	60.0	26.7	13.3	6.7

Some physical properties are:

mp	41 °C
Density (25 °C)	0.854 g/cm^3
Flash point (open cup)	94 °C
Dissociation constant (water, 20 °C)	
pK_1	11.11
pK_2	10.01
Heat of combustion (25°C)	40 208 kJ/kg
Specific heat capacity, c_p (solid)	2.85 kJ kg^{-1} K^{-1}
mp of water eutectic (53 % H_2O)	− 20 °C
Solubility in 100 g water (30 °C)	960 g
Heat of solution in water, J/g	
90 % solution	56
85 % solution	71
70 % solution	112
Infinite dilution	245

Properties of an aqueous solution [7] are as follows:

	Concentration, wt %		
	90	85	70
mp, °C	30	24	6
bp, °C	129.5	122.3	111.5
Density (25 °C), g/cm^3	0.889	0.900	0.929
Flash point (open cup), °C	102	107	116

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_629

3. Chemical Properties

Hexamethylenediamine is a strong organic base, e.g., a 10 % aqueous solution has a pH of 12.3 (20 °C), and forms stable salts with both organic and inorganic acids. The most important salt is that produced by neutralization with adipic acid (salt strike): the so-called nylon salt or AH salt. This compound is the raw material for the preparation of nylon by thermal dehydration under vacuum.

Other reactions of hexamethylenediamine are typical of aliphatic amines (→ Amines, Aliphatic). The preparation of hexamethylene diisocyanate [*822-06-0*] by reaction with phosgene is industrially important:

$$H_2N(CH_2)_6NH_2 + 2\,COCl_2 \longrightarrow OCN(CH_2)_6NCO + 4\,HCl$$

The reaction takes place in a chlorinated aromatic solvent with a yield of 95 % or more [8]. Hexamethylene diisocyanate can also be produced by converting hexamethylenediamine to an aliphatic urethane, followed by thermal cleavage [9] (→ Isocyanates, Organic).

4. Production

Virtually all hexamethylenediamine is now produced by the *catalytic hydrogenation of adiponitrile* [*111-69-3*] (→ Adipic Acid, Chap. 8.1.):

$$NC(CH_2)_4CN + 4\,H_2 \longrightarrow H_2N(CH_2)_6NH_2$$

This process is carried out in the liquid phase in the presence of ammonia, which serves as a heat-transfer medium and which, with some catalysts, improves yields by decreasing the extent of side reactions [10, 11]. A variety of catalysts can be used, most of which are based on promoted (i.e., activated or doped) cobalt [12–15], but iron oxide is also effective [16, 17]. Typical reaction temperature is 100 – 200 °C, and operating pressure is usually high, e.g., 28 – 41 MPa. Purification is by distillation, with a yield of ca. 85 % [18]. Byproducts are 1,2-diaminocyclohexane [*694-83-7*], hexamethyleneimine [*111-49-9*], and bis(hexamethylenetriamine) [*143-23-7*], all of which are recovered for sale.

Adiponitrile itself is produced commercially by at least three processes, each based on a different hydrocarbon. The original process commercialized by Du Pont and still used in a number of adiponitrile plants starts with adipic acid made from cyclohexane:

$$C_6H_{12} \xrightarrow{\text{Air}} C_6H_{11}OH + C_6H_{10}O$$
$$C_6H_{11}OH + C_6H_{10}O + HNO_3 \longrightarrow HOOC(CH_2)_4COOH$$

Adipic acid reacts with ammonia at 300 – 400 °C and atmospheric pressure in the presence of a catalyst to give adiponitrile in yields as high as 95 % [19–21]:

$$HOOC(CH_2)_4COOH \xrightarrow{NH_3} NC(CH_2)_4CN$$

Du Pont abandoned this method after developing a process for the catalytic addition of hydrogen cyanide to butadiene [22, 23]:

$$CH_2{=}CHCH{=}CH_2 \xrightarrow{HCN} NC(CH_2)_4CN$$

This reaction is carried out at ca. 100 °C and sufficient pressure to maintain the reaction in the liquid phase. Adiponitrile is separated from byproducts by distillation.

A third route to adiponitrile is by electrolytic dimerization of acrylonitrile [*107-13-1*], produced by the ammoxidation of propene [*115-07-1*] [24]. This process, developed by Monsanto, was commercialized in 1965. The following reaction takes place at the cathode in an electrolytic cell:

$$2\,CH_2{=}CHCN + 2\,H^+ + 2\,e \longrightarrow NC(CH_2)_4CN$$

Hydrogen ions are formed from water at the anode and pass to the cathode through a membrane. The catholyte consists of a mixture of acrylonitrile, water, and a tetraalkylammonium salt. The anolyte is aqueous sulfuric acid. The yield of adiponitrile based on acrylonitrile is 92 – 95 %. Small amounts of propionitrile and bis(cyanoethyl) ether are formed as byproducts.

A direct route to hexamethylenediamine, not based on adiponitrile, was used by Celanese but discontinued in 1984: hydrogenation of adipic acid esters to 1,6-hexanediol [*629-11-8*],

which then reacted with ammonia at 219 °C and 22.8 MPa.

In Western Europe, hexamethylenediamine is made from adipic acid via adiponitrile by ICI, BASF, and Rhône-Poulenc; BASF also makes adiponitrile by electrolytic dimerization of acrylonitrile. Butachimie, a joint venture of Du Pont and Rhône-Poulenc, uses adiponitrile produced by the Du Pont hydrocyanation route. In Japan, the only hexamethylenediamine producer is Asahi Chemical, which uses electrolytic dimerization of acrylonitrile to manufacture the adiponitrile intermediate. Rhodia, the South American producer, also uses acrylonitrile as raw material.

5. Quality Specifications and Analysis

The color of aqueous hexamethylenediamine solutions and nylon salt solutions is considered an indication of purity. When these solutions are prepared from pure hexamethylenediamine, they are essentially water-white at first but slowly develop color on storage in contact with air. Solutions prepared from diamine containing impurities such as 1,2-diaminocyclohexane, hexamethyleneimine, and tetrahydroazepine develop color more rapidly, depending on the level of impurities and the storage temperature. The color of the polyamide produced is related to the color of the hexamethylenediamine and the nylon salt solution used.

Although no impurity specifications are given for commercial hexamethylenediamine, typical specifications for polymer-grade hexamethylenediamine (maximum content, parts per million) are

1,2-Diaminocyclohexane	50
6-Aminocapronitrile	10
2-Aminomethylcyclopentylamine	100
Hexamethyleneimine	25
Tetrahydroazepine	100
Ammonia	50
Color, Pt – Co scale (APHA)	10

1,2-Diaminocyclohexane and 2-aminomethylcyclopentylamine [*21544-02-5*] are determined by gas chromatography with a capillary column. Ammonia [*7664-41-7*], 6-aminocapronitrile [*2432-74-8*], and hexamethyleneimine are analyzed by a combination of distillation and titration. Ammonia is removed first by simple distillation with methanol as a carrier; the distillate is collected in boric acid solution and titrated with standard acid. After removal of ammonia, the sample is made alkaline with aqueous sodium hydroxide and refluxed to hydrolyze the aminocapronitrile to sodium aminocaproate. The ammonia liberated in this step is distilled with methanol and titrated with acid. The distillation is then continued into an aqueous fraction that carries over the hexamethyleneimine, which forms a binary azeotrope (*bp* 95.5 °C; 50.5 wt % imine) with water. This fraction is collected in neutral water and titrated with standard acid. Tetrahydroazepine can be determined by pulse polarography. Because azepine is a Schiff base, standards may be prepared by adding an aliphatic aldehyde to the hexamethylenediamine sample to give another Schiff base. The color is determined by a standard comparison colorimeter.

6. Storage and Transportation

Hexamethylenediamine should be stored in tightly closed containers in well-ventilated areas.

Shipping Information [5], [7].

United States, other than by air (DOT)

Shipping name	hexamethylenediamine, solid hexamethylenediamine, anhydrous hexamethylenediamine, solution
Hazard class	corrosive material
UN/NA no. (solid)	UN 2280
UN/NA no. (liquid)	UN 1783
DOT hazard classification	corrosive
Freight class	chemicals NOI

International, by water or air (IMO/ICA)

Hazard class	IMCO 8
UN no. (solid)	UN 2280
UN no. (liquid)	UN 1783
IMO/ICAO label(s) (solid)	corrosive
IMO/ICAO label(s) (solution)	corrosive/poison
Subsidiary risk	6.1
Packaging group	II

7. Economic Aspects

Hexamethylenediamine production capacities (thousand tonnes) in 1986 are given below:

Canada	45
United States	630
South America	13
Western Europe	331
Japan	30
People's Republic of China	21
	1071

The United States and Western Europe account for approximately 90 % of the estimated world capacity of hexamethylenediamine. Because more than 95 % of production is used for nylon 66 fibers and plastics, the future of hexamethylenediamine remains dependent on the growth of this particular nylon. Production of Hexamethylenediamine in 1995 is 658 x 10^3 t in the USA, 370 x 10^3 t in Western Europe, and 46 x 10^3 t in Japan [28].

8. Toxicology and Occupational Health

Exposure of humans to hexamethylenediamine by inhalation, ingestion, or skin or eye contact may cause eye irritation or blurring of vision, upper respiratory tract irritation, or skin irritation [25]. Higher exposures lead to abnormal liver function, skin burns, or eye corrosion with corneal or conjunctival ulceration. Skin sensitization has been reported in some individuals, and some evidence exists that hexamethylenediamine permeates the skin [26].

The following animal toxicities have been determined [5]: (ALC), LC_{50} 1580 ppm (mouse, inhalation, 10 min); LC_{50} 1100 mg/kg (rabbit, skin absorption); and LD_{50} 792 – 1127 mg/kg (rat, oral). Hexamethylenediamine is corrosive to the skin and eyes of animals but no evidence of skin sensitization exists. On exposure by inhalation, animals have exhibited upper and lower respiratory tract irritation [27]. Other reported effects on animals by ingestion, inhalation, and skin contact include weight loss and liver or kidney damage. The compound is not embryotoxic in animal tests. No mutagenic activity has been observed in bacterial and mammalian cell cultures.

9. References

1. T. Curtius et al., *J. Prakt. Chem.* **62** (1900) 189.
2. H. Mark, G. S. Whitby (eds.): *High Polymers,* vol. 1, Interscience, New York 1940.
3. Products for Surface Coatings, Desmodur/Desmophen, ed. 1.6, Farbenfabriken Bayer AG, 14, 1986.
4. *Merck Index,* 10th ed., Merck & Co., Rahway, NJ, 1983.
5. *Hexamethylenediamine (Anhydrous)* Material Safety Data Sheet, Petrochemicals Dept., Du Pont, Wilmington, DE 19803, 1985.
6. L. S. Scott, *Personal Communication,* Du Pont.
7. *Hexamethylenediamine (Solution)* Material Safety Data Sheet, Petrochemicals Dept., Du Pont, Wilmington, DE 19803, 1985.
8. Du Pont, US 2 374 340, 1945 (M. W. Farlow).
9. BASF, US 4 613 466, 1986 (F. Meyer, G. Nestler, F. Towae, H. Hellbach).
10. Du Pont, US 3 461 167, 1969 (D. R. Buehler, G. P. Keister, I. F. Long).
11. BASF, US 3 972 938, 1976 (D. Voges, L. Hupfer, S. Wederl, K. W. Leonhard, H. Hoffman).
12. BASF, US 3 232 888, 1966 (K. Adam).
13. Du Pont, US 3 773 832, 1973 (L. D. Brake).
14. Montedison Fibre, US 3 821 305, 1974 (G. Bartalini, M. Guiggioli).
15. BASF, US 4 598 058, 1986 (G. Frank, G. Neubauer, P. Duffman, H. J. Wilfinger).
16. Du Pont, US 3 696 153, 1972 (B. J. Kershaw, M. G. Pounder, K. R. Wilkins).
17. BASF, US 4 587 228, 1986 (G. Frank, G. Neubauer).
18. Chemstrand Corp., US 3 017 331, 1962 (C. R. Campbell et al.).
19. ICI, US 3 325 532, 1967 (J. D. Rushtan, R. A. Williams).
20. BASF, US 3 671 566, 1972 (M. Decker, J. Schmidt, H. Hoffman, H. J. Pistor).
21. ICI, US 3 325 531, 1967 (J. G. Mather, R. A. Williams).
22. Du Pont, US 3 496 215, 1970 (W. C. Drinkard Jr., R. V. Lindsey Jr.).
23. Du Pont, US 3 536 748, 1970 (W. C. Drinkard Jr., R. V. Lindsey Jr.).
24. M. M. Baizer, *J. Electrochem. Soc.* **111** (1964) 215.
25. G. Gallo, L. Ghiringhelli, *Med. Lav.* **49** (1958) 683.

26. C. Ceresa, *Med. Lav.* **39** (1948) 162.
27. J. C. Gage, *Br. J. Ind. Med.* **27** (1970) no. 1, 1.
28. K. Weissermel, H.-J. Arpe *Industrielle Organische Chemie,* Wiley-VCH, Weinheim, (1998)

Hexanes → Hydrocarbons
Hexanols → Alcohols, Aliphatic
High-Density Polyethylene → Polyolefins
High Explosives → Explosives
High-Impact Polystyrene → Polystyrene and Styrene Copolymers

Heat Storage Media

ERICH HAHNE, Universität Stuttgart, Institut für Thermodynamik und Wärmetechnik, Stuttgart, Federal Republic of Germany

1. **Introduction** 1
2. **Practical Principles for Heat Storage Media** 1
3. **Storage of Sensible Heat** 2
3.1. **Solid Storage Materials** 2
3.2. **Liquid Storage Materials** 4
3.3. **Gaseous Storage Materials** 6
4. **Storage of Latent Heat** 7
4.1. **Salt Hydrates** 8
4.2. **Paraffins** 9
5. **Storage of Chemical Heat** 10
5.1. **Thermochemical Transformer and Chemical Heat Pump** 12
5.2. **Adsorption Processes** 14
5.3. **Effective Properties** 15
6. **Summary** 16
7. **References** 16

1. Introduction

Heat – in the physical sense – is a form of energy and can be stored in various ways and for many different applications. *Low-temperature heat* is stored for domestic hot water supply, *high-temperature heat* for power processes; short-term and long-term storage can be applied in house heating; storage in stationary constructions has been experienced in ground stores or buildings, while storage in mobile equipment is commonly used in fuel for automotive engines. Each storage concept has its best suited materials and these may occur in different physical phases: as solids, liquids, or gases.

A special kind of heat storage media are fossil or nuclear fuels, such as wood, coal, oil, or uranium. These are characterized by a very high energy density (e.g., 42.7 MJ/kg in diesel oil) and an irreversible application. Since storage is understood in this article as a reversible process which can be repeated more or less indefinitely, the irreversible storage processes and their materials will not be treated here. The materials considered in what follows are characteristic for the storage of (1) sensible heat, (2) latent heat, and (3) chemical heat.

2. Practical Principles for Heat Storage Media

The following requirements should be met by heat storage materials:

1) Long service life, noncorrosive, nontoxic, nonflammable
2) Large heat storage capacity
3) Proper thermal diffusivity (temperature diffusivity) $a=\lambda/\varrho c\ \left[\mathrm{m^2/s}\right]$ and high heat diffusivity $b=\sqrt{\lambda\varrho c}\ \left[\mathrm{Jm^{-2}K^{-1}s^{-1/2}}\right]$ (where λ is the thermal conductivity, ϱ the density, and c the specific heat capacity
4) Ability to undergo charging – discharging cycles without loss in performance and storage capacity, or change in structure
5) Easy availability, simple in handling, and storable in simple containment
6) Inexpensiveness

A decisive criterion of a heat storage medium is its price and the costs that arise upon its utilization. Long life and a high cycling stability are prerequisites for economic application, i.e., at a price competitive with existing storage facilities. High thermal diffusivity of the heat storage material provides a quick response to temperature differences, i.e., quick charging and discharging; high heat diffusivity yields a high amount of heat being stored.

Heat transfer processes have to be considered: the heat may be either transferred directly to the storage material as, for example, in a dry pebble bed with air flow, or a heat exchanger may be required as in a solar domestic hot water store where the water – antifreeze mixture flowing through a solar collector has to be separated from the hot water for consumption. Other aspects of selecting a heat storage material may be

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a12_b30

Table 1. Generalized survey of heat storage media

Materials for the storage of		
Sensible heat	Latent heat	Chemical heat (in reversible processes)
Solids	phase change materials	dissociating compounds: solids or liquids releasing gas (thermochemical transfer, chemical heat pump)
Liquids	solid ↔ liquid	gaseous compounds for catalytic reactions (thermochemical pipeline)
Solid – liquid combinations	liquid ↔ gas	adsorbents – adsorbates
Water-flooded pebbles	salt hydrates	
Oil – cast iron	eutectica	

operational advantages in energy supply systems or a larger flexibility in application. A survey of heat storage media is given in Table 1.

3. Storage of Sensible Heat

Sensible heat always results in an increase or decrease of the material temperature. All materials have a capability of absorbing and storing heat due to the fact that they have a mass m and a specific heat capacity c_{p} at constant pressure. For a temperature difference $\Delta T = T_2 - T_1$ this heat (or enthalpy) amounts to

$$Q_{\mathrm{sensible}} = m c_{\mathrm{p}} \, (T_2 - T_1) \tag{1}$$

where the unit of Q is, for example, J. T_2 denotes the material temperature at the end of the heat absorbing (charging) process and T_1 at the beginning of this process. This heat is released in the respective discharging process. For pure solids (especially heavy elements), the specific heat per mole of a substance is ca. $3R$ (Dulong – Petit rule), with R being the molar gas constant ($R = 8.31441$ J mol^{-1}K^{-1}). Thus, the molar thermal energy q_{mol} stored in solids can be approximated by

$$q_{\mathrm{mol}} \approx 3R \cdot \Delta T \tag{2}$$

and q_{mol} is measured in, for example, J/mol. Thus, approximately 25 J/mol can be stored with a temperature difference of $\Delta T = 1$ K. With the molar mass M (kg/mol), the thermal energy q stored per mass (storage capacity in J/kg) is obtained

$$q = c \Delta T = q_{\mathrm{mol}} / M = 3R \cdot \Delta T / M \tag{3}$$

In Table 2 some characteristic materials are listed together with their thermophysical properties.

The metals and graphite are best suited for quick charging and discharging (high thermal diffusivity a) and for a large amount of heat stored (high heat diffusivity b). Other solid materials such as stones are much less advantageous, their respective values are smaller by an order of magnitude. Liquids and gases – despite their inattractive thermophysical properties – offer the advantage of a possible use as both storage medium and heat carrier. The most widely applied media, in this respect, are water and air.

3.1. Solid Storage Materials

Solid materials can be heated to very high temperature (e.g., magnesia bricks in Cowper regenerators to 1000 °C). They are usually chemically inert and, in the form of rocks and pebbles, abundant and cheap.

Pebble bed and rock pile stores consist of loosely packed material in a container. Heat has to be carried to or from such stores by a heat carrier such as air or water. These beds provide a large surface area for heat transfer in their dispersed materials; the internal heat losses are only small because the particles have little surface contact to surrounding particles and thus, little conduction losses. Reduced insulation around the storage container is sufficient when air – with a small thermal conductivity – is used as a heat carrier and fills the gaps between the particles. Such particle beds, however, cause a large pressure drop in the fluid flow. This has to be compensated by pumps or compressors. Another drawback is the large storage volume required. Conveniently, this is provided in underground installations.

Table 2. Thermophysical properties of media for sensible heat storage

Material	θ, °C	ϱ, kg/m^3	c, kJ/kg · K	λ, W/m K	$10^6 \times a$, m^2/s	$10^{-3} \times$b, J/m^2 ks$^{1/2}$
Solids						
Aluminum 99.99 %	20	2700	0.945	238.4	93.3	24.66
Copper (commercial)	20	8300	0.419	372	107	35.97
Iron	20	7850	0.465	59.3	16.3	14.7
Lead	20	11340	0.131	35.25	23.6	7.24
Brick (dry)	20	1800	0.84	0.50	0.33	0.87
Concrete (gravelly)	20	2200	0.879	1.279	0.66	1.57
Granite	20	2750	0.892	2.9	1.18	2.67
Graphite (solid)	20	2200	0.609	155	120	14.41
Limestone	20	2500	0.741	2.2	1.19	2.02
Sandstone	20	2200	0.71	1.8	1.15	1.68
Slag	20	2700	0.836	0.57	0.25	1.13
Soil (clay)	20	1450	0.88	1.28	1.0	1.28
Soil (gravelly)	20	2040	1.84	0.59	0.16	1.49
Liquids ($c = c_p$)						
Paraffin (liquid)	20	900	2.13	0,26	0.14	0.71
Silicone oil (AK 250)	25	970	1.465	0.168	0.118	0.49
Sodium	100	927	1.385	85.84	66.85	10.50
Transformer oil	60	842	2.09	0.122	0.069	0.46
Water	20	998	4.183	0.598	0.142	1.58
Gases ($c = c_p$)						
Air (1 bar)	20	1.205	1.005	0.026	21.6	0.006
Hydrogen	50	0.075	14.36	0.202	191	0.015
Water vapor	100	0.589	2.135	0.029	9.61	0.006

In ground stores, heat is stored in soil, stones, or solid rocks, especially for low-temperature storage. *Cast iron* is an inexpensive material with a high volumetric heat capacity (energy density) $q \cdot \varrho$. It is used for high-temperature storage together with oil as the heat carrier. As solid materials for storage always have to be combined with fluid heat carriers, properties of the solids have to be taken into account which affect fluid flow and heat transfer, e.g., particle and container size, mechanical durability against abrasion and thermal cycling, hygroscopy, etc. In Figure 1, a storage bin for solid materials with gas as a heat carrier is shown. If the packing density is high, e.g., for small solid particles, the storage capacity is high, but the pressure drop is also high. A proper particle size which should be rather uniform has to be found.

The thermal stratification (see Section 3.2) is well maintained in a solid-material heat store, because natural convection in the heat carrier is suppressed by the granulate and internal heat conduction is low, because of the aforementioned reasons. Containers as shown in Figure 1– in form of pillars – are used as architectural components in buildings. The combination of hollow bricks and air has been known as hypocaust-heating systems since antiquity. Water-flooded pebble beds are known as artificial aquifers.

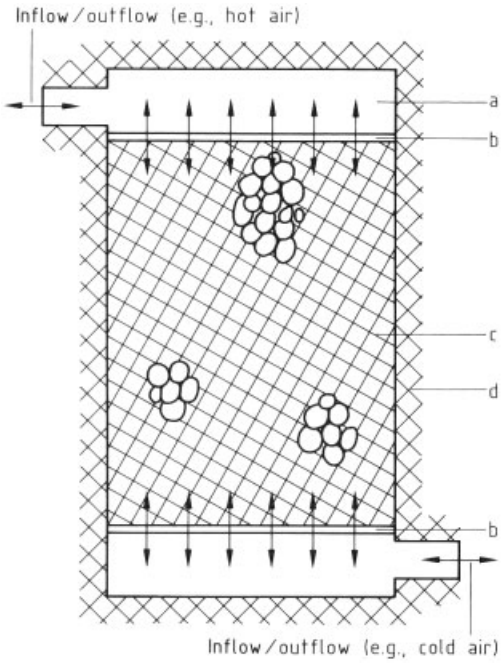

Figure 1. Storage bin for solid materials
a) Air-distribution chamber; b) Grating; c) Rocks; d) Insulation

3.2. Liquid Storage Materials

Water. The most widely used liquid for thermal storage is water. Water has the following advantages:

1) It is abundant and inexpensive
2) It has a relatively high heat diffusivity ($b = 1.58 \times 10^3\ \mathrm{Jm^{-2}\,K^{-1}s^{-1/2}}$)
3) A relatively low thermal (temperature) diffusivity ($a = 0.142 \times 10^{-6}\ \mathrm{m^2/s}$); an advantage for thermal stratification within a hot-water storage tank
4) It can be easily stored in all kinds of containers
5) The control of water flow systems is highly flexible and is often state-of-the-art
6) Water can be used without heat exchangers
7) Is is easy to handle, nontoxic, noncombustible, experience with water is common
8) It is easily mixable with additives (antifreeze, anticorrosive)

Thermal stratification in heat storage vessels is desired, because the value of high temperature heat is maintained in the upper part of the vessel, while low-temperature fluid (as backflow from a heat consumer) can still be stored in the lower part of the vessel. The temperature distribution within a storage vessel is shown in Figure 2 for various lengths of time. Due to transient internal heat conduction, the stratification is destroyed in the course of time, even in very well insulated vessels (time 3 represents a longer time duration than time 1).

Water has the following disadvantages:

1) It only provides a limited operation range between 0 °C (freezing) and 100 °C (boiling)
2) It is corrosive
3) It has high vapor pressures ($p > 5$ bar) at temperatures > 155 °C

The vapor pressure can be reduced when mixtures of water with other chemicals are applied; e.g., a mixture of 50 % H_2O and 50 % NaOH can be used up to 140 °C without a pressure vessel.

Other Liquid Storage Materials. Other liquids, oils, molten salts, and metals (Na) are used for high-temperature storage as their vapor pressures are low. Freezing points of such materials, however, have to be taken into consideration (e.g., for Na $\theta_m = 98$ °C). Oils and paraffins may be used as heat carriers in combination with solids.

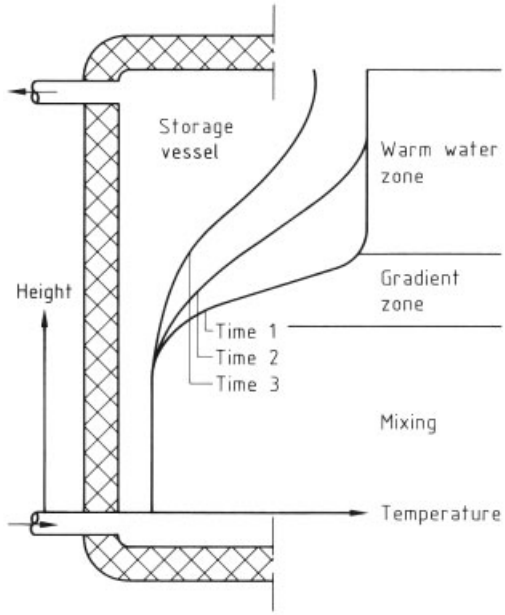

Figure 2. Thermal stratification in a hot-water storage vessel (warm water zone, gradient zone, and mixing are valid for time 1)

Applications Storage tanks for hot water are used in industry and dwellings. They come in sizes of 0.1 m^3 (domestic hot water storage) to 12 000 m^3 (solar assisted district heating). The most recent development are "solar combination stores" of about 0.5 m^3 which serve for both domestic hot water supply and house heating. In Figure 3 such a metal storage tank is shown with its heat exchangers for a good thermal stratification.

Water is commonly used also in large thermal stores, both as storage medium and heat transfer fluid. In large stores (e.g., seasonal stores for solar energy) the storage capacity can be turned over only once or – at the most – three times a year. Such a small energy turnover makes the stored heat expensive unless the storage container is inexpensive. A rough comparison of energy cost for various storage concepts is given in Table 3. Here, "energy cost" is the cost per kilowatt-hour. The energy considered is the heat stored in one annular cycle. The cost is based on investments for a 15 years' payoff time and 6 % interest.

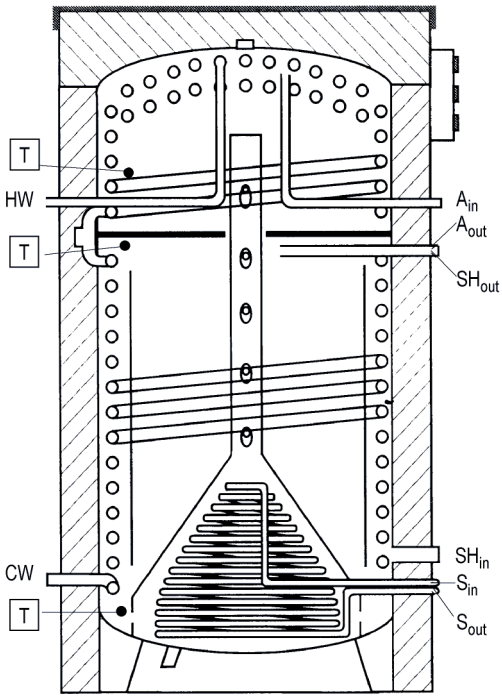

Figure 3. Solar combination storeS_{in} = Solar loop from collector; S_{out} = Solar loop to collector; A_{in} = Auxiliary heating loop inlet; A_{out} = Auxiliary heating loop outlet; SH_{in} = Space heating loop inlet; SH_{out} = Space heating loop outlet; CW = cold water inlet; HW = Hot water outlet; T = Temperature sensor

Table 3. Energy cost for various storage concepts

	Cost, DM/kWh	Volume, m^3	Max. temp., °C	Max. efficency, %
Steel tank	0.044 – 0.078	10 000	85	78
Ground container	0.030 – 0.070	15 000	85	80
Rock cavern	0.021 – 0.050	105 000	75	85
Rock boreholes	0.009 – 0.027	11 000	50	85
Ground boreholes	0.009 – 0.034	30 000	50	50
Aquifer	0.003 – 0.026	800 000	35	80

A technical/economical boundary for *steel tanks* is a size of around 100 000 m^3. For *ground containers,* the waterproof liner may create problems (e.g., tightness after welding, perforation by stones): plastic- and stainless steel-liners are used. These large containers are built of concrete and they are fully or partly buried in the ground. Their capacity is around 80 kWh/m^3. A schematic view of such a container is shown in Figure 4. In summer, hot water is fed to the water layers on top and cold water withdrawn from the bottom, in winter, hot water is withdrawn from the top layers and cold water re-fed at the bottom.

In so-called *ground stores* low temperature heat is stored in soil or pebbles. The storage medium in such cases is a mixture of water and solid particles. Density ϱ and specific heat capacity c of these mixtures can be obtained from the mixing laws. The thermal conductivity λ depends strongly on the composition of the mixture. Although some mathematical equations exist in the literature, reliable data for λ depend on measurements. Heat is usually exchanged with the ground by tubes containing a circulating carrier fluid, mostly water. These tubes may be placed close to the surface (0.5 – 1 m below) in a meandering arrangement or spirally wound in horizontal ditches. For large stores (more than 10 000 m^3 water equivalent storage capacity), vertical holes are drilled for the heat exchanger tubes (see Fig. 5, next page).

In some areas, geology allows for a so-called *aquifer store.* This requires a permeable layer of soil (pebbles or sand) enclosed between horizontal layers of impermeable material (clay). Under such advantageous conditions, two fountains can be drilled some distance apart from each other and hot water can be inserted in summer and withdrawn in winter (see Fig. 6). Prob-

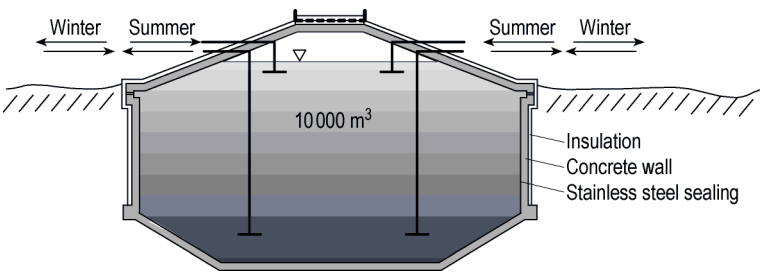

Figure 4. Schematic view of a large hot water storage vessel

lems can arise from minerals dissolving in the water and precipitating on tubes and heat exchangers.

So-called *man-made aquifers* are built in sizes of 10 000 m^3; the ground excavations are filled with pebbles and water and sealed by plastic sheets [14]. In all cases where mixtures of water and solid material are used for heat storage, the storage volume has to be larger than in the case of pure water. Rough data are given in Table 4.

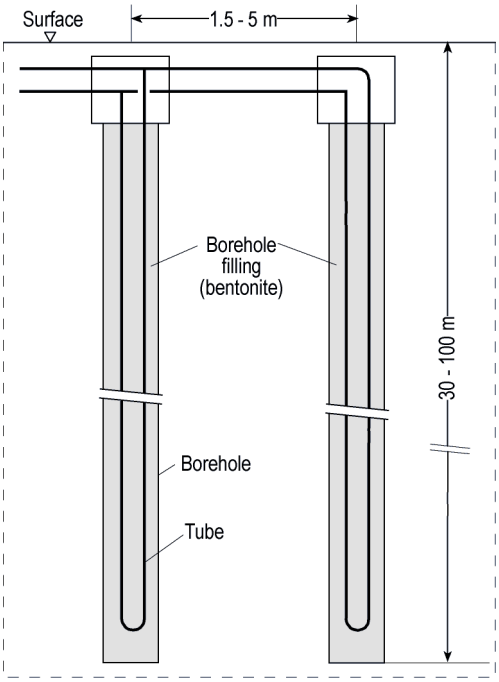

Figure 5. Schematic view of a borehole heat store

Table 4. Storage volume of different storage media compared to pure water for 1 m^3 water equivalent and 60 kWh capacity at $\Delta T = 50$ K

Storage media	Storage volume
Hot water storage	1 m^3
Pebble – water storage	1.3 – 2 m^3
Aquifer storage	2 – 3 m^3
Borehole storage	4 – 5 m^3

3.3. Gaseous Storage Materials

Gases have extremely low heat diffusivities and are particularly unsuitable for *storing heat.* Moreover, their volumetric specific heat capacity $\varrho\ c_p$ is extremely low (ca. 1 $kJm^{-3}\,K^{-1}$) while that of solids or liquids is between 1000 and 4200 $kJm^{-3}\,K^{-1}$ (e.g., for water $\varrho\ c_p = 4200\ kJm^{-3}\,K^{-1}$).

Gases are, however, highly important *heat carriers,* especially at elevated and very high temperature. *Air* and *flue-gases* are the heat carriers in Cowper regenerators. *Air* is used as a quickly responding carrier with little thermal inertia in air heating systems with pebble-bed storage. *Hydrogen* is stored in metal hydrides either as a later fuel or as a reactant for chemical heat storage. *Water vapor* is created when the pressure from a steam accumulator (Ruths store) is released. Ruths stores are used in peak-load storage plants of power stations or for process heat in industries of chemical, food, and metallurgical branches. The heat is actually stored in the saturated steam or pressurized water. Gases require large volumes or pressure vessels for storage.

4. Storage of Latent Heat [6–12]

Materials for the storage of latent heat are also named *phase change materials* (PCM), because of the fact that they change their physical phase

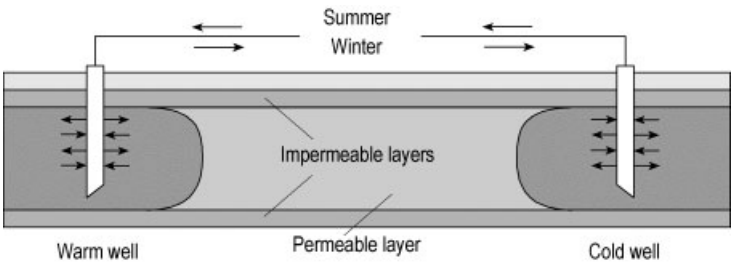

Figure 6. Schematic view of an aquifer heat store

from solid to liquid or liquid to gas and vice versa. The phase change is always coupled with the absorption of heat when the solid melts and the liquid evaporates – heat of melting or heat of evaporation – and a heat release when the liquid solidifies or the gas condenses. The phase change occurs at the melting or evaporation temperature T_{m} or T_v. At that temperature, these materials melt or evaporate when heat is added, but do not exhibit a rise in temperature. Thus, the heat added cannot be sensed and appears to be latent.

A change of one crystalline form into another without a physical phase change may also be considered as storage of latent heat. The phase change of salt hydrates is considered a solid – liquid transformation. The solid hydrate crystals release their water of crystallization at a certain temperature (or a temperature range) and the anhydrous salt dissolves in this water. This process is reversed when heat is withdrawn. The heat (or enthalpy at p = const.) stored with a phase change amounts to

$$q_{\mathrm{latent}} = \Delta h_{\mathrm{m}} \text{ or } \Delta h_{\mathrm{v}} \tag{4}$$

where Δh_{m} is the latent heat (enthalpy) of melting and Δh_{v} the latent heat (enthalpy) of evaporization.

When a material is heated from an initial temperature T_1 to the melting temperature T_{m} and the melt is further heated to T_2, the total heat stored is

$$Q_{\mathrm{total}} = Q_{\mathrm{sensible,solid}} + Q_{\mathrm{latent}} + Q_{\mathrm{sensible,liquid}} \quad q_{\mathrm{total}} \cdot m \tag{5}$$

or using Equations (1) and (4):

$$Q_{\mathrm{total}} = m \cdot c_{\mathrm{solid}} (T_{\mathrm{m}} - T_1) + m \Delta h_{\mathrm{m}} + m \cdot c_{\mathrm{liquid}} (T_2 - T_{\mathrm{m}}) \tag{6}$$

The specific heat capacity of the solid phase is c_{solid} and of the liquid phase c_{liquid}.

For a *liquid – gaseous phase change*, the total heat stored is accordingly

$$Q_{\mathrm{total}} = m \cdot c_{\mathrm{liquid}} (T_{\mathrm{v}} - T_1) + m \Delta h_{\mathrm{v}} + m \cdot c_{p\mathrm{gas}} (T_2 - T_{\mathrm{v}}) \tag{7}$$

with T_{v} as the temperature of evaporation. The heat of fusion (i.e., the heat of melting or evaporation) is much greater than the sensible heat. Therefore, phase change materials have a larger volumetric energy storage capacity than sensible storage materials. A further advantage of PCMs is the absorption and the release of heat at a constant temperature; this melting or evaporation temperature is material-specific and the material can be selected accordingly.

A rough estimate of the heat of fusion of the various phase change materials can be obtained from the following formulas. For most pure metals, the *molar heat of melting* $\Delta h_{\mathrm{m,\,mol}}$ is within the range (T_{m} in K)

$$RT_{\mathrm{m}} < \Delta h_{\mathrm{m,mol}} < 1.5RT_{\mathrm{m}} \tag{8}$$

For semiconductors, eutectic, and inorganic compounds, the range is

$$2.5RT_{\mathrm{m}} < \Delta h_{\mathrm{m,mol}} < 5RT_{\mathrm{m}} \tag{9}$$

The following approximation can be used for organic substances

$$\Delta h_{\mathrm{m}} \approx 24T_{\mathrm{m}}/M \tag{10}$$

giving Δh_{m} in J/g for the molar mass M in g/mol.

The *latent heat of evaporation* can roughly be approximated (at 1 bar) by Trouton's rule

$$\Delta h_{\mathrm{v,mol}} \approx 9T_{\mathrm{v}}R \tag{11}$$

Comparison with data in literature, however, indicates a range of

$$6RT_{\mathrm{v}} < \Delta h_{\mathrm{v,mol}} < 15RT_{\mathrm{v}} \tag{12}$$

Latent heat storage materials should be inexpensive and be characterized by

1) A large phase-change enthalpy and a high density
2) A large thermal diffusivity in the solid and liquid phase
3) Thermal stability and cycle stability
4) Little or no subcooling during freezing and little or no supersaturation during melting
5) A low thermal expansion, i.e., low volume change during melting

They should neither exhibit segregation (e.g., like Glauber's salt), nor be corrosive, or react with casing materials, should not be poisonous, inflammable, or form explosive phases.

Various phase change materials are listed together with their thermophysical properties in Tables 5 and 6.

Table 5. Thermophysical properties of solid – liquid phase-change materials

Material	T_m, K	Δh_m, J/g	Δh_m*,[a] J/cm^3	c_{solid}, J/g K	$c_{p,\,liqu.}$, J/g K	$10^6 a_{solid}$, m^2/s
Elements						
Na	371	113	110	1.23	1.39	111
Li	453	433	231			
Pb	601	23	260	0.13	0.15	24
Zn	693	113	806	0.38		41
Inorganic substances						
H_2O (Ice)	273	336	333	2.09	4.19	1.2
$LiNO_3$	525	530	941	2.03	2.04	0.29
LiOH	744	1080	1500	3.3	3.9	0.28
LiF	1121	1050	2770	2.64		
MgF_2	1540	925				
Organic compounds						
$C_3H_5(OH)_3$) (Glycerol)	291	192	250			
$(C_{17}H_{35}CO_2)_3C_3H_5$ (Tristearin) (Glycerol tristearate)	329	191	164			
Salt hydrates						
$KF4H_2O$	292	231	336	1.84	2.39	
$CaCl_2 6\,H_2O$	302	172	258	1.47	1.47	
$Na_2SO_4 10\,H_2O$	305	251	354	1.76	3.31	
$Na_2HPO_4 12\,H_2O$	308	280	403	1.55	3.18	
Eutectic mixtures (in % mol)						
12 NaF/40 KF/44 LIF/4 MgF_2	722	690	1490 liquid			
46 LiF/44 NaF/10 MgF_2	905	864	2255			
67 LiF/33 MgF_2	1019	904	2378			

[a] $\Delta h*_m$ = volumetric energy density (volumetric heat of melting).

Table 6. Evaporation temperature and heat of evaporation of liquid – gas phase-change materials

Material	T_v, K	Δh_v, J/g	Δh_v*, J/m$^3_{vap}$.
NH_3	241	1758	1.34
C_2H_5OH	351	947	
H_2O	373	2229	1.34
Hg	630	304	
K	1052	2031	
Na	1187	4257	
NaCl	1680	13159	

Thermophysical properties are not always available and their values may differ in literature. Impurities in the substances can change the properties considerably. The tables only present characteristic materials, many more are found in literature. For heat storage in the low-temperature range (<50 °C; e.g., for solar energy) salt hydrates and paraffins are mainly considered.

4.1. Salt Hydrates

Many salt hydrates are sufficiently inexpensive; they are only little corrosive and slightly toxic. They exhibit only small changes in volume during phase change and have a relatively high heat of fusion (compared to organic compounds) at low melting temperature. Salt hydrates are of the chemical form $X(Y)_i \cdot n\,H_2O$ with $X(Y)_i$ representing an inorganic compound. When the salt hydrate is heated to its transition temperature, one of the following reactions will take place

$$X(Y)_i \cdot n\,H_2O \longrightarrow X(Y)_i + n\,H_2O \quad (13)$$

$$X(Y)_i \cdot n\,H_2O \longrightarrow X(Y)_i \cdot o\,H_2O + (n-o)\,H_2O) \quad (14)$$

At the melting point, the hydrate crystals split up into either an anhydrous salt and water (13) or into a hydrate with a lower number of water molecules and water (14).

Salt hydrates have two drawbacks for thermal storage: they suffer from incongruent melting and supercooling.

Incongruent melting occurs when not enough water of crystallization is released during thawing and the anhydrous salt cannot completely dissolve. The solid residual part settles at the bottom of the container due to its higher density. In the reverse process (during recrystallization) some fraction of the precipitated solid is unable to come into contact with the water necessary for its recrystallization. Thus, the useful heat of fusion reduces after several charge – discharge cycles. Incongruent melting can be prevented by:

1) Addition of a thickening agent which holds the anhydrous solid part in suspension and prevents its settling
2) Stirring, vibrating, or rotating the material
3) Addition of excess water
4) Encapsulation of the material (e.g., in plastics, as metallic containers may be corroded)

Supercooling occurs during cooling (heat discharge) when no sufficient nucleation sites are available. The crystallization does not take place at the melting point; the liquid is cooled below that point and the stored heat is not discharged at the expected melting temperature, but at some arbitrary temperature. In some cases it might happen that the stored heat cannot be withdrawn, because the temperature difference between the subcooled melt and the required load temperature, is not high enough.

One of the following means can promote the initial crystallization and prevent supercooling:

1) Addition of small quantities of a nucleating agent with a crystal structure similar to that of the salt
2) Use of a container with rough walls (the peaks and pins form nucleation sites)
3) Preservation of a "cold spot" in the container, where no melting takes place and crystals survive

If the rate of crystal growth is high and there is good heat transfer from the walls, then the temperature of the supercooled melt can quickly rise to the melting point and is maintained for the recrystallization process. In this case, subcooling may be tolerated.

4.2. Paraffins

Paraffins are waxy solids at room temperature. Their melting points and heat of fusion increase with molar mass. Compounds in the series $C_{14}H_{30} - C_{40}H_{82}$ have melting points in the range 6 – 80 °C. They are advantageous storage materials for the following reasons:

1) They are widely available and inexpensive
2) Noncorrosive and nontoxic
3) Chemically stable, with a low density
4) Phase transformation is rapid
5) Supercooling is negligible

But they show the following drawbacks:

1) Low volumetric energy density (compared to salt hydrates)
2) Change in volume during melting by 10 %
3) Low thermal conductivity and diffusivity
4) No distinct melting point, because cheap technical paraffins used for thermal storage are mixtures of many compounds
5) Not compatible with plastics

5. Storage of Chemical Heat [12, 13]

Reversible thermochemical reactions can be used for storing thermal energy. The energy is stored in the form of chemical compounds created by an endothermic reaction and it is recovered again by recombining the compounds in an exothermic reaction. The heat stored and released is equivalent to the heat (enthalpy) of reaction.

The enthalpy of reaction, generally, is much larger than the enthalpy of transition (in latent heat storage) or the sensible heat stored over a reasonable temperature span. The storage density (based on solid mass or volume) is much larger for thermochemical storage materials than for latent or sensible heat storage materials.

For thermochemical processes, the molar heat of reaction $\Delta h_{\mathrm{mol,\,r}}$ (at 1 bar) is approximately within the range

$$11RT_{\mathrm{r}} < \Delta h_{\mathrm{mol,r,gas}} < 19RT_{\mathrm{r}} \tag{15}$$

where T_{r} is the reaction temperature in K. The

Table 7. Temperature and heat of reaction of materials for chemical heat storage

Material	T_r (at 1 bar), K	$\Delta h_{\mathrm{r\,gas}}$, J/g (gas)	$\Delta h_{\mathrm{r\,comp}}$ [a], J/g (comp)[a]	$\Delta h*_{\mathrm{r\,comp}}$ [a], J/cm^3 (comp)[a]
Hydroxides	$A(OH) \rightleftarrows B + H_2O$			
$Mg(OH)_2$	544	4235	1306	3134
$Ca(OH)_2$	793	5865	1414	3167
$Ba(OH)_2$	923	7383	771	3470
Ammoniated salts	$A \cdot n(NH_3) \rightleftarrows B \cdot o(NH_3) + (n-o)NH_3$			
$CaCl_2 \cdot 8\,NH_3$	313	2733	745	410
$FeCl_2 \cdot 6\,NH_3$	373	3079		
$ZnCl_2(NH_3)_2$	613	3201		
Carbonates	$A(CO_3) \rightleftarrows B + CO_2$			
$CaCO_3$	1133	3520	1546	4189
$MgCO_3$	673	2474	1296	3836
Metal hydrides	$\mathrm{Metal} \cdot nH_2 \rightleftarrows \mathrm{Metal} \cdot oH_2 + (n-o)H_2$			
FeTiH	266	14040	134	744
$LaNiH_6$	277	15120	197	1285
MgH_2	533	38520	2922	4237
$MgNiH_4$	603	32040	1147	2948
Catalytic reactions				
$CH_4(g) + H_2O(g)$	1273			
$\rightleftarrows CO(g) + 3\,H_2(g)$	(100 bar)	6053	6053	91
$SO_3(g) \rightleftarrows SO_2(g)$	1273			
$+1/2\,O_2(g)$			1235	

[a] comp = compound.

specific heat of reaction (in J/g) can be roughly calculated from

$$\Delta h_{\mathrm{r}} \approx (100 - 150)\, T_{\mathrm{r}}/M \tag{16}$$

where M denotes the molar mass in g/mol. Materials for the storage of chemical heat, their temperature of reaction, and heat of reaction are given in Table 7.

The reversible reactions suitable for heat storage can be classified into thermal dissociation reactions and catalytic reactions.

Dissociation Reactions. Certain solid or liquid compounds can undergo dissociation reactions when they are heated. A gas is released while the depleted solid or liquid remains in the reactor (endothermic reaction, i.e., *charging of the store*). The reverse reaction will occur spontaneously if the equilibrium is changed by a temperature decrease or a pressure increase. Therefore, the dissociation products have to be separated and stored individually. For *discharge*, in the exothermic reaction, the gas is recombined with the solid or liquid. The reactions follow the scheme:

$$\mathrm{AB\ (solid\ or\ liquid)} \pm \mathrm{heat} \rightleftarrows \mathrm{C\ (solid\ or\ liquid)} + \mathrm{D\ (gas)}$$

A characteristic example is the dissociation of $Ca(OH)_2$:

$$Ca(OH)_2 \rightleftarrows CaO + H_2O$$

The respective storage arrangement is shown in Figure 7, see next page. The $Ca(OH)_2$ container serves as both the reactor and the CaO store. The water vapor evolved in the charging reaction is condensed and stored as liquid. For the discharge reaction, the liquid water is evaporated and returned to the CaO store where $Ca(OH)_2$ is formed again.

Catalytic Reactions. These are usually gas – gas reactions. A catalyst is required to obtain high reaction rates. The reaction products are transported in a pipeline from the charging source to the consumer (discharging sink). This system is called chemical heat pipeline or chemical heat pipe. The reactions follow the schemes

$$\mathrm{A\ (gas)} + \mathrm{B\ (gas)} \pm \mathrm{heat} \rightleftarrows \mathrm{C\ (gas)} + \mathrm{D\ (gas)}$$

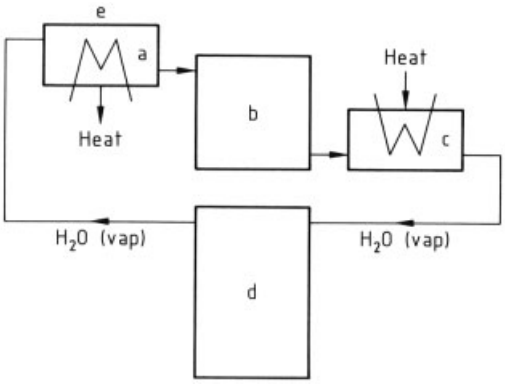

Figure 7. Thermochemical energy storage using the reaction $Ca(OH)_2 \rightleftharpoons CaO + 2\,H_2O$
a) Condenser; b) Water store; c) Evaporator; d) $Ca(OH)_2$ reactor, CaO store

or

$$A\,(gas) \pm heat \rightleftarrows C\,(gas) + D\,(gas)$$

A characteristic example is the reaction:

$$CH_4 + H_2O \rightleftarrows CO + 3\,H_2$$

(Adam – Eva system, Jülich, Germany; see → Gas Production). The pipeline system is shown schematically in Figure 8.

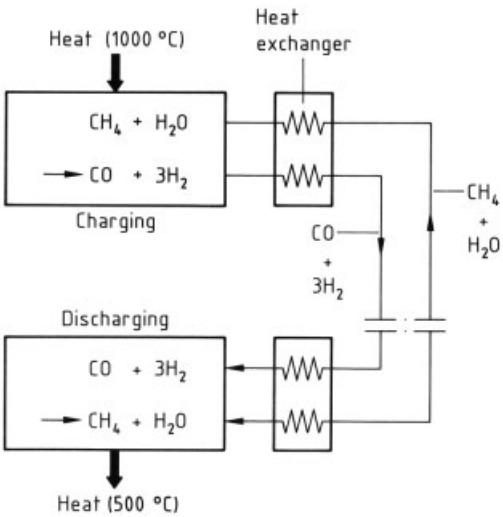

Figure 8. Thermochemical pipeline

In thermochemical heat storage, the chemical compounds and the reaction products can be stored at ambient temperature for infinite time without heat losses. Thus, a thermal insulation of the storage containers is not required. The chemicals, however, have to be heated and cooled before reactions. This must be utilized in a carefully designed heat exchange equipment, in order to maintain a high overall system efficiency for the operational cycle.

In selecting the proper chemical materials, the following criteria have to be considered.

Reversibility. The reaction must go to completion, no side reactions or changes in reaction rate with time should occur.

Reaction Rates. Both forward and reverse reactions must be rapid enough to absorb all the available energy or release it promptly. The reaction rates must not decrease with cycling. Such a decrease is observed with the crushing of particles into powder when high heating rates and high temperatures are applied.

Controllability. The reactions must be controllable because they have to be turned on and off when required. Controllability is achieved by product separation, by control of temperature and pressure, or by catalysts.

Ease of Storage. This affects the cost considerably. The reaction products have to be easily separable prior to storage. Reactions must not occur at storage temperature.

Safety. Toxicity, inflammability, and corrosiveness of the reaction products may pose unacceptable safety hazards.

Cost. Cost should be low; this requires materials which are readily available.

For catalytic reactions the products need not be separated before storage which reduces the cost. In dissociation processes, when the reaction product is in a gaseous phase, separation is easy, but storage poses problems. The gas requires large storage volumes and the volumetric energy storage density is small: e.g., at 1 bar in the range of 0.5 to 1 kWh/m^3. When the gas is compressed, expensive pressure vessels and extra electrical energy for the compressor are needed. Various possibilities of reactions and storage are shown in Figure 9. The gas can be *condensed* and stored as a liquid in a small volume (Fig. 9B). Vapor of water and methanol can easily be condensed. For gases such as SO_2 and NH_3 this is more difficult and the liquification of H_2 requires a considerable amount of extra energy. A disadvantage of liquid storage is that the condensed phase has to be again evaporated for

the reverse reaction. This latent heat of evaporation must be considered a thermal loss, unless it is utilized in the condensation process. When the latent heat is not used, the obtainable thermal efficiency will be less than 50 %. Under favorable conditions (depending on the pressure – temperature relation), the evaporation occurs within the ambient temperature level. Such arrangements are utilized in intermittently working absorption refrigerators.

The gaseous product of a primary reaction can also be *absorbed* by another chemical in a secondary reaction (Fig. 9C). The secondary-store absorbent should be capable of absorbing the gas near ambient temperature. When energy is required, some heat is supplied to the secondary-storage container, the gas is released, and flows back to the primary-storage container. Here high-temperature heat is produced in a reversible reaction (thermochemical transformer, chemical heat pump).

An interesting arrangement of a *thermochemical power plant* is shown in Figure 9D. The exhaust steam of a turbine reacts in the store with some chemical (e.g., NaOH) exothermally. Water vapor is created within the store by the heat released and this vapor is used for the operation of the turbine. Recharging of the store may be accomplished by heating the chemical from a high-temperature heat source and returning the desorbed and condensed water vapor to a boiler reservoir. This principle was applied in the "Honigmann locomotive" (1882) with steam of 5.5 bar; the NaOH concentration was reduced from 82 % to ca. 65 %.

5.1. Thermochemical Transformer and Chemical Heat Pump

The pressure – temperature interrelation of such processes can be obtained from the equilibrium equation

$$\ln p = -(\Delta h_{\mathrm{r}}/\mathrm{R}) \cdot 1/T + \Delta s/R \qquad (17)$$

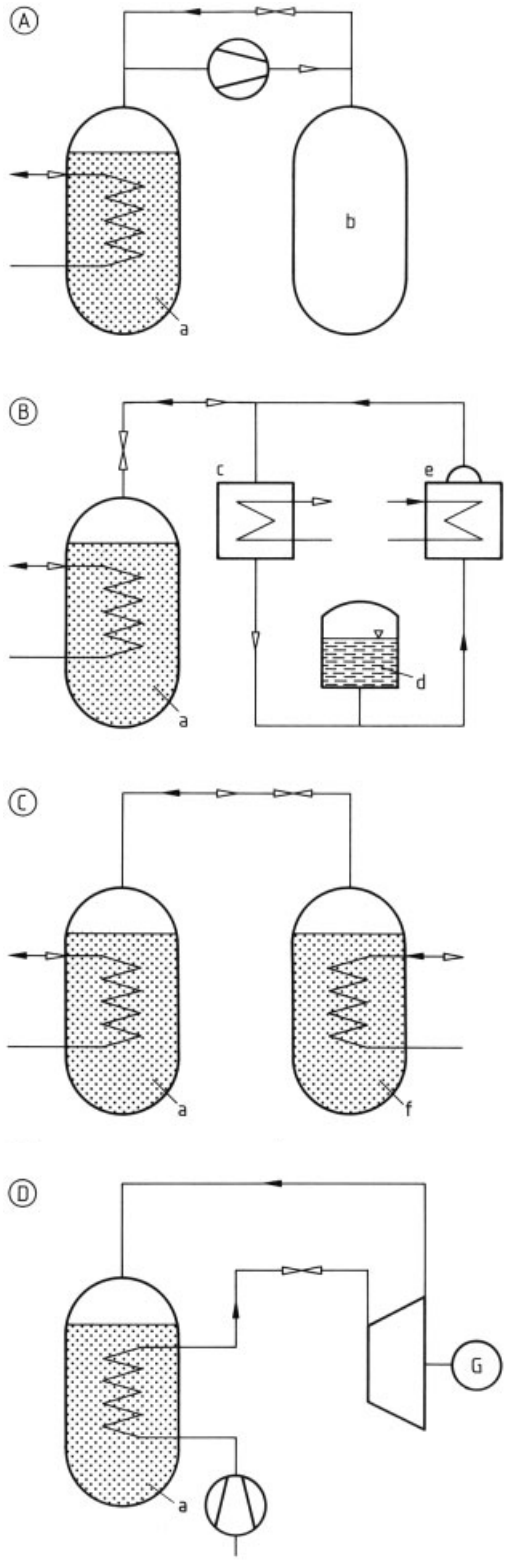

Figure 9. Various concepts for heat storage in dissociation processes
A) Gas storage; B) Compression of gas and subsequent liquid storage; C) Absorption in secondary reaction; D) Thermochemical power plant a) Absorbent; b) Gas store; c) Condenser, d) Liquid store; e) Evaporator; f) Secondary-store absorbent ▷ = charge; ▶ = discharge

where Δh_r is the (specific) heat of reaction, R the molar gas constant, and Δs the change of (specific) entropy due to the reaction. With Δh_r and Δs being considered constant over the temperature range used, the value of $\ln p$ depends linearly on $1/T$ and this dependence is plotted to provide necessary information on the required states of materials. The cycle of a *thermochemical transformer* is presented in Figure 10.

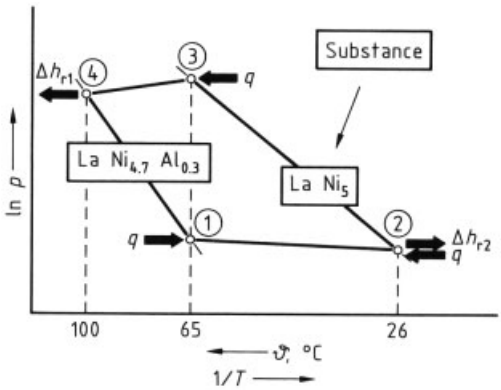

Figure 10. Pressure – temperature interrelation for a metal hydride heat transformer cycle

Two metal compounds are used which form hydrides according to the scheme

$$\text{Metal } nH_2 \pm \text{heat} \rightleftarrows \text{Metal } oH_2 + (n-o)H_2$$

with $o < n$ or $o = 0$.

The two compounds, $LaNi_{4.7}Al_{0.3}$, as the primary substance and $LaNi_5$ as the secondary one, are stored in two containers as in Figure 9C. Heat q is added to the hydride of $LaNi_{4.7}Al_{0.3}$ at 65 °C (1), hydrogen is released at a pressure of about 2.4 bar and absorbed in the second vessel by $LaNi_5$ at 26 °C (2) and a somewhat lower pressure. Low-temperature heat of reaction Δh_{r_2} has to be dissipated to the surroundings. For mere heat storage, the secondary store would be heated to 26 °C, in order to release hydrogen back to the primary store, where the heat of reaction is obtained at a higher temperature than 26 °C [the direction (1) to (2) is reversed, but with a pressure drop from (2) to (1)]. For transformer operation, however, the secondary store is heated up to 65 °C and 7.8 bar (3). At that state, hydrogen is released from the secondary store and reabsorbed in the primary store with an exothermic reaction which provides heat of reaction at 100 °C and 7 bar (4). When the heat delivered from the store reaches 65 °C (1), the cycle is closed, the store is discharged. The counterclockwise transformer cycle becomes a *chemical heat pump cycle* when the direction of the process is reversed (into a clockwise cycle, i.e., arrow directions being reversed). In this case, the flow of gas requires $p_4 > p_3$ and $p_2 > p_1$. The *transformer process* yields heat of a higher temperature than the heat delivered to the cycle, with a conversion efficiency (upgraded heat to delivered heat) of 0.5 to 1. The *heat pump process* utilizes heat of high temperature and of low temperature to provide heat of an intermediate temperature. Usually, the low-temperature heat input (at ambient temperature) is neglected. The conversion efficiency (provided heat to delivered heat) is in the range of 1 to 2.

Temperature levels required for such cycles depend on the compounds and the conversion efficiencies. The reactions are either of the liquid – gas (wet) or solid – gas type (dry). *Liquid – gas processes*, e.g., ammonia – water, water – lithium bromide, sulfuric acid – water are in use in absorption refrigerators and absorption heat pumps. They operate at varying input and output temperature, since concentration, besides pressure, affects the reaction. Heat and mass transfer problems are usually less than with solid – gas processes. *Solid – gas processes* have high rates of reaction which allow for compact devices; they operate at constant output temperature. Their drawback is a poor heat and mass transfer behavior, with low thermal diffusivity and high contact resistance. Solid materials exhibit sagging, swelling, and decomposition in long-term cycling. Side reactions may occur.

5.2. Adsorption Processes

Adsorption of a gas (adsorbate) on the surface of a solid (adsorbent) is a spontaneous exothermic process. By heating the solid, the gas is desorbed. This provides a means of storing heat: the absorption material is being charged when the gas is desorbed by heating (the adsorbent is free of gas), and discharged when the gas is adsorbed (the adsorbent becomes saturated with gas). The energy evolved during the process is the heat of adsorption.

A high heat-storage density can be achieved when a large amount of gas is adsorbed, and the heat of adsorption is large. A large quantity of gas can be adsorbed on a large surface: good adsorbents are porous solids with a very large internal surface area, originating from pore sizes of ca. 3 – 500 pm. The adsorbance rate is large when the temperature is low (below the equilibrium temperature). The heat of adsorption increases with the polarity of the adsorbate molecules; H_2O, NH_3, and CO_2 are suitable compounds. Water proves to be the best, because of its high dipole moment and latent heat. An adsorbent – water pair would not require a store for water. Since water vapor is readily available in the air, such a system could work in an open cycle.

Adsorbent materials have to be easily available and cheap. They should be chemically inert and should not undergo side reactions. Hygroscopic materials are proper adsorbents for water and widely used in industry. A number of such materials are listed with their relevant properties in Table 8

5.3. Effective Properties

Sorption processes are either exo- or endothermic. In either case, heat has to be transported within the storage material. In order to provide a large surface area for reaction in solid – gas processes, the storage material is in granular or powdery form. In such material, heat is transported simultaneously by conduction, convection, and radiation. These transport regimes cannot be separated, therefore effective properties such as effective thermal conductivity λ_{eff} and effective thermal diffusivity a_{eff} have to be measured so that layout calculations can be performed. Powdery materials of any kind, metal or ceramic, are very poor thermal conductors; especially when they are evacuated, they are often used as good insulators.

The *effective thermal conductivity* of the aforementioned $LaNi_{4.7}Al_{0.3}$ is shown as a function of pressure for various filling gases in Figure 11 [15]. The S-shaped curves are characteristic for powdery materials. The effective thermal conductivities differ according to the thermal conductivities of the filling gas λ_{gas}. The effect of the solid material can be observed by the fact that λ_{eff} is about 10 times higher than λ_{gas}. But, on the other hand, the thermal conductivity of solid $LaNi_{4.7}Al_{0.3}$ would be another ten times higher ($\lambda_{\mathrm{solid}} = 12.5\ \mathrm{Wm^{-1}\,K^{-1}}$). At high Knudsen numbers (i.e., mean free path/characteristic length) $Kn \geq 10$, at pressures below 10^{-4} bar, in the region of free molecular flow, when molecules move without collisions across the pores, the effective thermal conductivity is only about $0.004\ \mathrm{Wm^{-1}\,K^{-1}}$. This value is roughly independent of pressure and filling gas. Here, heat is transported by heat conduction in the grid.

Table 8. Thermophysical properties of adsorbent – adsorbate pairs

Material	Maximum adsorbate capacity, $\frac{\mathrm{kg\,adsorbate}}{\mathrm{kg\,adsorbent}}$	Heat of adsorption (aver.) Δh_{ads}, $\frac{\mathrm{kJ}}{\mathrm{kg\,adsorbate}}$	Energy capacity $q_{\mathrm{adsorbent}}$, $\frac{\mathrm{kJ}}{\mathrm{kg\,adsorbent}}$	Adsorbent specific heat $c_{\mathrm{adsorbent}}$, $\frac{\mathrm{kJ}}{\mathrm{kg\,K}}$	Adsorbent density, $\frac{\mathrm{kg}}{\mathrm{m^3}}$
Water with:					
Charcoal	0.4	2320	930	1.09	480
Silica gel	0.37	2560	950	0.88	670
Activated alumina	0.19	2480	470	1.00	980
zeolites					
4 A	0.22	4410	970	1.05	780
5 A	0.22	4180	920	1.05	680
13 X	0.30	4410	1320	0.92	600
Methanol with:					
Active carbon	0.32	1400	450	0.9	
Zeolites					
4 A	0.16	2300	370	1.05	780
5 A	0.17	2300	390	1.05	680
13 X	0.20	2400	480	0.92	600

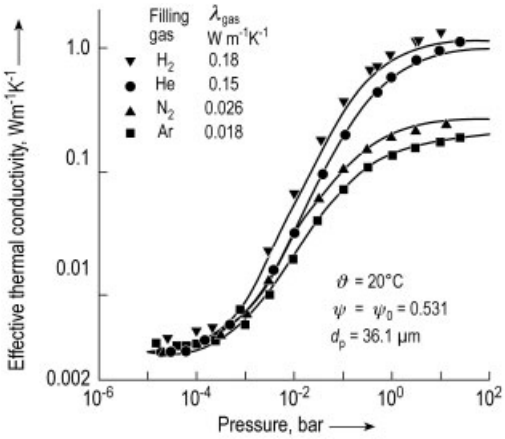

Figure 11. Effective thermal conductivity of $LaNi_{4.7}Al_{0.3}$ (ψ_0 initial porosity, d_p initial particle diameter)

A similar, roughly S-shaped, distribution is obtained for the *effective thermal diffusivity.* In Figure 12 this diffusivity is shown for a zeolite and various filling gases [16].

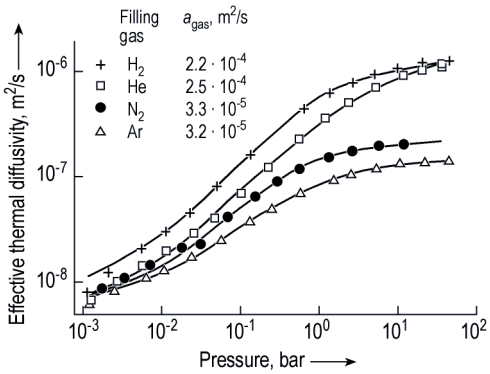

Figure 12. Effective thermal diffusivity of zeolite NaA

The temperature effect on λ_{eff} is linear with about 2.2×10^{-3} K^{-1}. These low effective properties of sorbent materials are a drawback for their technical application; for the improvement a well conducting matrix (Al or Cu) is inserted into the storage material.

6. Summary

A comprehensive presentation of heat storage materials and their volumetric energy density is given in Figure 13 (see next page).

7. References

General References

1. H. P. Garg, S. C. Mullick, A. K. Bhargava: *Solar Thermal Energy Storage,* D. Reidel Publ. Co., Dordrecht 1985.
2. J. A. Duffie, W. A. Beckman: *Solar Engineering of Thermal Processes,* Wiley & Sons, New York 1991.
3. G. Beghi (ed.): *Thermal Energy Storage (Ispra Courses),* D. Reidel Publ. Co., Dordrecht 1981.
4. C. den Ouden (ed.): *Thermal Storage of Solar Energy,* Martinus Nijhoff Publ., The Hague 1981.
5. G. Beckman, P. V. Gilli: *Thermal Energy Storage,* Springer Verlag, Wien 1984
6. M. Goldstein: "Some Physical Chemical Aspects of Heat Storage," *U.N. Conference on New Sources of Energy,* Rome 1961, no. 35, 5 – 7.
7. M. Telkes: "Solar Energy Storage," *ASHRAE J.* **16** (1974) 38 – 44.
8. M. Altmann, H. Yeh, H. G. Lorsch: *Conservation and Better Utilization of Electric Power by Means of Thermal Energy Storage and Solar Heating–Final Summary Report,* NSF/RANN/SE/G/27976/PR 73/5, Univ. of Penns, July 1973.
9. H. G. Lorsch: *Thermal Energy Storage – Final Report,* NSF/RANN/74-021C, 1974.
10. B. Carlson, H. Stymme, G. Wettermark: "Storage of Low-Temperature Heat in Salt-Hydrate Melts – Calcium chloride hexahydrate," *Swed. Council F. Buildg. Res.,* **D 12,** 1978.
11. T. Ozawa et al.: "Screening of Latent Heat Thermal Energy Storage Materials by Using Evaluated Thermodynamic Data," *7th Codata Int. Conf.,* Kyoto 1980.
12. R. W. Mar: "Material Science Issues Encountered During the Development of Thermochemical Concepts" in L. E. Murr (ed.): *Solar Mater. Sci.* Academic Press, London 1980.
13. W. E. Wenthworth, F. Chen: "Simple Thermal Decomposition Reactions for Storage of Solar Thermal Energy," *Sol. Energy* **18** (1976) 205 – 214.
14. E. Hahne, M. F. Hornberger, "Combined Heat- and Cold-Storage in a Gravel/Water Pit, Proc. Calorstock '94," *6th Int. Conf. on Thermal Energy Storage,* vol. 1, Espoo, 1994, pp. 365 – 372.

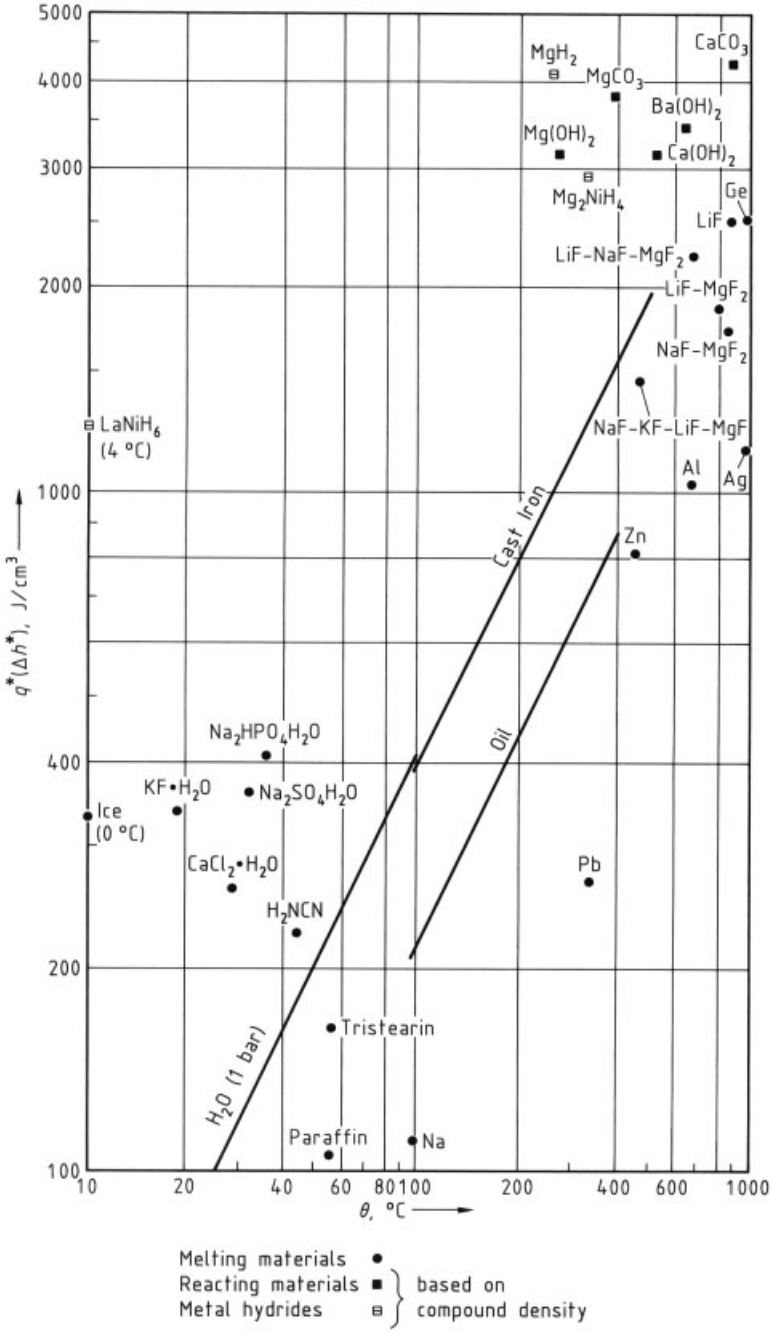

Figure 13. Volumetric energy density of various storage media for sensible heat: $q* = \varrho\, c\, (\theta - 0)$, no voids or matrix considered

15. E. Hahne, J. Kallweit, "Thermal Conductivity of Metal Hydride Materials for Storage of Hydrogen: Experimental Investigation," *Int. Journal Hydrogen Energy* **23** (1998) 107 – 114.

16. A. Griesinger: *Experimentelle und theoretische Untersuchung des Wärmetransports in Zeolithschüttungen,* Doctoral Thesis, Universität Stuttgart, ITW, 1998.

Heat Transfer → Heat Exchange
Heavy Water → Isotopes, Natural
Helium → Noble Gases
Hepatotherapeutic Agents → Gallbladder and Liver Therapy
Heptanols → Alcohols, Aliphatic
Herbicides → Weed Control

High-Performance Fibers

Hiroshi Mera, Teijin Limited, Osaka, Japan

Tadahiko Takata, Teijin Limited, Osaka, Japan

1. Definition and Classification . . . 1
2. Heat-Resistant Fibers 4
2.1. Introduction 4
2.2. Solution-Spun Heat-Resistant Fibers 5
2.2.1. Aramid Fibers 5
2.2.2. Polyazole and Polyimide Fibers . . 9
2.3. Melt-Spun Heat-Resistant Fibers 11
2.3.1. Novoloid Fibers 11
2.3.2. Fibers from Engineering Plastics . . 11
3. High-Strength and High-Modulus (HS – HM) Fibers 12
3.1. Introduction 12
3.2. Solution-Spun HS – HM Fibers . . 14
3.2.1. Para-Oriented Aramid Fibers 14
3.2.1.1. Production 14
3.2.1.2. Properties 17
3.2.1.3. Uses 18
3.2.2. HS – HM Fibers from Some Aramid-Related Polymers 19
3.2.3. Polyazole and Polyimide Fibers . . 19
3.2.3.1. Polyazole Fibers 20
3.2.3.2. Polyimide Fibers 21
3.3. Melt-Spun HS – HM Fibers 21
4. References 23

Abbreviations used in this article:

BBB	poly(bisbenzimidazobenzophenanthroline)
FF	fiber – fiber
HM	high modulus
HS	high strength
LOI	limiting oxygen index
PAI	poly(amide – imide)
PBI	polybenzimidazole
PBO	poly(*p*-phenylenebenzobisoxazole)
PBT	poly(*p*-phenylenebenzobisthiazole)
PEEK	polyetherether ketone
PEI	polyetherimide
PI	polyimide
PMIA	poly(*m*-phenyleneisophthalamide)
PMTA	poly(*m*-phenyleneterephthalamide)
PPIA	poly(*p*-phenyleneisophthalamide)
PPPI	poly(*p*-phenylenepyromellitimide)
PPP	poly(*p*-phenylene)
PPS	poly(phenylene sulfide)
PPTA	poly(*p*-phenyleneterephthalamide)

1. Definition and Classification

Although a strict definition of high-performance fibers does not yet exist, the term generally denotes fibers that give higher values in use in a range of applications. It commonly refers to fibers with some unique characteristics that differentiate them from commodity fibers such as nylon, polyester, and acrylic fibers. Synonyms are *specialty fibers* and, in some cases, *high-functional fibers.* High-performance fibers can be classified broadly into three categories according to their applications:

1) heat-resistant fibers, including flame-retardant ones;
2) high-modulus and high-strength fibers; and
3) other fibers

Tractable or fusible polymers can be spun from a polymer melt (*melt spinning*). Polymers that can be dissolved in a suitable solvent can be spun from a solution (*solution spinning*). The dissolved polymer is spun into a hot gas where the solvent evaporates (*dry spinning*) or into a liquid coagulating bath (*wet spinning*) (→ Fibers, 3. General Production Technology).

The concept of high-performance fibers covers not only organic but also inorganic fibers of carbon, alumina, and boron, which often include various whiskers (→ Fibers, 5. Synthetic Inorganic; → Whiskers). A variety of high-performance fibers can be found in composite materials (→ Composite Materials). Moreover, highmodulus fibers from aliphatic polymers, such as polyethylene or polyoxymethylene, also belong to the class of high-performance fibers

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_001

(→ Fibers, 4. Synthetic Organic, Chap. 4.). The classification of high-performance fibers is summarized in Figure 1.

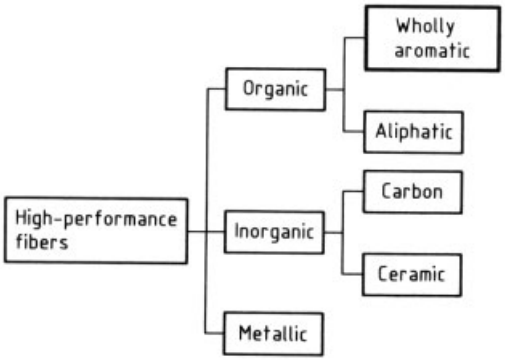

Figure 1. Classification of high-performance fibers

The term high performance may also be applied to other properties such as radiation resistance [11], [12] or electrical conductivity [13] (→ Polymers, Electrically Conducting), but a discussion of these fibers is beyond the scope of this article. Here, discussion is restricted to the polymerization, fiber production, properties, and applications of fibers obtained from wholly aromatic polymers.

Table 1. Aromatic residues of meta- and para-oriented aromatic polymers [14]

Polymer	Aromatic residues
Meta-oriented	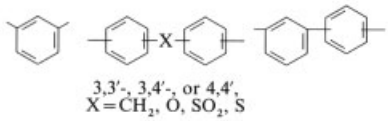 3,3′-, 3,4′-, or 4,4′, X = CH_2, O, SO_2, S
Para-oriented	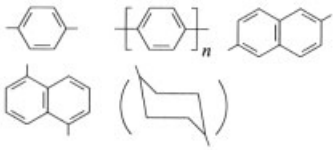

Wholly aromatic polymers may be classified as *meta-* or *para-oriented,* depending on their chemical structure and properties. Fibers from meta-oriented polymers are useful as heat-resistant fibers. Para-oriented polymers are expected to be useful not only as heat-resistant fibers but also as high-strength (HS) or high-tenacity and high-modulus (HM) fibers. Typical examples of aromatic residues of the two classes are given in Table 1 [14].

Figure 2 shows the melting (softening) points of aromatic polyamides (aramids) with four components having *m-* or *p*-phenylene residues [15–18]. These four components are poly(*m*-phenyleneisophthalamide [*24938-60-1*] (PMIA),

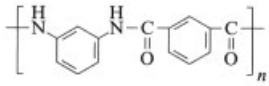

poly(*m*-phenyleneterephthalamide) [*24938-63-4*] (PMTA),

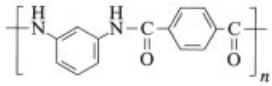

poly(*p*-phenyleneisophthalamide) [*24938-61-2*] (PPIA), and

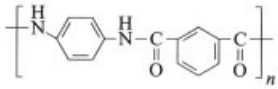

poly(*p*-phenyleneterephthalamide) [*24938-64-5*] (PPTA).

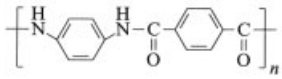

Fibers from PMIA, PMTA, or PPIA are heat-resistant and are classified in this article as meta-oriented according to their fiber properties.

Meta-oriented aramid fibers have a higher heat resistance than commodity fibers such as polyester or polyacrylonitrile fibers, as well as favorable physical and mechanical properties for clothing or textile use. The PMTA fibers have a higher heat resistance than PMIA fibers, but their physical and mechanical properties are similar; thus, they are defined as meta-oriented.

Heat-resistant, high-modulus PPTA fibers are defined as para-oriented. Some ambiguities inevitably arise in copolymers containing more than 50 mol % of para-substituted residues. Among the aramids shown in Figure 2, optical anisotropy in concentrated sulfuric acid is shown by polymers in which more than ca. 75 mol % of

the total phenylene residues are para-oriented. Such polymers can give high-modulus fibers [18].

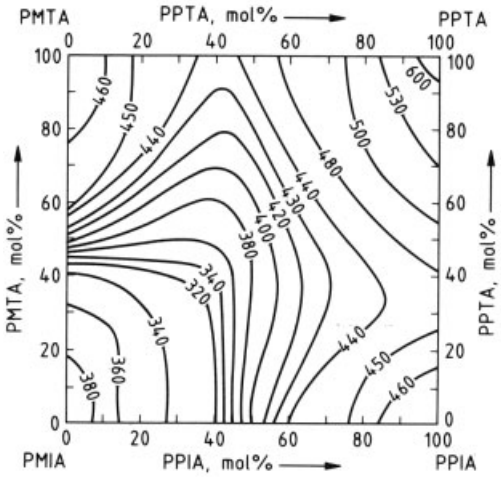

Figure 2. Melting points (°C) of aramids [16]
PMIA = poly(*m*-phenyleneisophthalamide);
PMTA = poly(*m*-phenyleneterepthalamide);
PPIA = poly(*p*-phenyleneisophthalamide);
PPTA = poly(*p*-phenyleneterephthalamide)

Liquid crystalline polymers are divided into three classes based on their chemical structures: backbone, side-chain, and mixed-configuration types. Predictions and explanations of the structural characteristics of low molecular mass compounds expressed as liquid crystallinity have been investigated in detail. These predictions are also applicable to liquid crystalline polymers. Important factors affecting liquid crystallinity are pressure, density, temperature, and chemical structure (degree of polymerization and properties of the repeating units, i.e., length, shape, intrachain rotational energy, dipole moment, site – site polarizability) [19]. Polymer concentration is an important factor in lyotropic liquid crystalline polymers.

Most meta-oriented fibers are heat-resistant; they are regarded as the first generation of high-performance fibers. Research and development on these fibers was spurred by the strong demand for heat-resistant and flame-retardant materials for space programs or industrial use after World War II. Many attempts were made to commercialize heat-resistant fibers, but only a few were successful at that time.

Para-oriented fibers are considered the second generation of high-performance fibers; they are composed mainly of para-substituted residues, instead of the meta-substituted residues of the first generation.

Table 2. Examples of high-performance fibers

Polymer	Trade name or developmental name	
	Meta-oriented fibers	Para-oriented fibers
Aramid	Nomex (HT-1; Du Pont)	Kevlar (Fiber-B, PRD-49, HT-4; Du Pont)
	Teijinconex (Teijin)	Technora (HM-50; Teijin)
	Fenilon (former Soviet Union)	Terlon (former Soviet Union)
		Twaron (Arenka; Akzo)
	Apyeil (Unitika)	Vniivlon (former Soviet Union)
	KM-21 (Kuraray – Mitsui Toatsu)	
Engineering polymers		
Poly(phenylene sulfide)	PPS (Phillips)	
Polyetherether ketone	PEEK (ICI)	
Polyether-imide	ULTEM (General Electric)	
	PI 2080/P8,4 (Dow – Lenzing)	
Novoloid (cross-linked phenolic resin)	Kynol (Gun-ei Chem. – Nippon and American Kynols)	
Polyamidehydrazide		X-500 (Monsanto)
		X-715 (Monsanto)
Polyarylate		Vectran (Kuraray – Celanese)
		Econol (Sumitomo Chemical – Unitika)
Polyazole	Celazole (Celanese)	PBT, PBO (Stanford Research Institute International – Dow)
Poly(amide – imide)	Kermel (Rhône – Poulenc)	
Polyimide	Arimide (former Soviet Union)	
Ladder polymer	BBB (J.E.Mark – Celanese)	
Chelate polymer	Enkatherm (Akzo)	

Compared to meta-oriented fibers, highly sophisticated polymerization and production techniques are needed for the para-oriented type to overcome difficulties caused by their rigid molecular structure. Du Pont initiated the second generation with Kevlar, a successor to the

first-generation Nomex. A tremendous amount of competitive research and development work has been undertaken in the industrialized countries. Despite these efforts, some fibers have had to be withdrawn at the developmental stage. Nevertheless, commercial exploitation has continued; HS and HM fibers have been introduced in conjunction with the development of composite materials (→ Composite Materials, Chap. 4.1.2.; → Reinforced Plastics).

Fibers from engineering polymers such as poly(phenylene sulfide), polyetherether ketone, and polyetherimide are attracting attention as a way to fill the gap between commodity fibers and first-generation heat-resistant fibers. The improved ratio of performance to cost is also attractive, as are the unique characteristics of this class of fibers; although they are melt-spinnable, they still have good thermal, mechanical, heat-adhesive, and chemical resistance properties.

Some typical high-performance fibers, including those that have not been commercialized, are listed in Table 2.

2. Heat-Resistant Fibers

2.1. Introduction

Polymers used for heat-resistant fibers are based on a comparatively simple concept of molecular design. The melting point (T_m) of polymers can be described by the formula

$$T_m = \Delta H / \Delta S$$

where ΔH and ΔS are the enthalpy and entropy of fusion, respectively.

The enthalpy of fusion can usually be increased by introducing symmetrical, rigid, or planar structures into the molecular backbone. The entropy of fusion can be decreased by introducing planar structures or polar linkages, such as amide bonds, into the molecular chain. For example, aramids have high melting points and high glass transition temperatures owing to strong intermolecular interactions between the aromatic residues and the amide linkages in their chain structures. In addition, the chemical structure of most wholly aromatic polymers makes them stable against moisture and oxidation. This conforms with the properties of cross-linked polymers (e.g., novoloid polymers) because the polymers are expected to have a rather small entropy of fusion in the softening stage even though they do not exhibit a well-defined melting point.

New technologies are required for the polymerization and spinning of heat-resistant fibers because the processing qualities of heat-resistant polymers are usually not as good as those of commodity fibers. Most of the problems are caused by a relatively high melting point and poor solubility. These problems can be overcome by using the solution polymerization–spinning method with polar solvents. Alternative approaches are cross-linking or additional solid-phase polymerization after melt spinning. For melt-spun fibers, sophisticated fiber production technologies have had to be developed, for example, increasing the degree of polymerization or improving the stabilizer, spinning equipment, and spinning conditions.

Solution spinning, melt spinning, and other aspects of fiber production discussed in this article are described in detail elsewhere (see → Fibers, 3. General Production Technology).

2.2. Solution-Spun Heat-Resistant Fibers

2.2.1. Aramid Fibers

Polymerization and Spinning. Combinations of aromatic diamines and aromatic dicarboxylic acids yield a variety of aramids, many of which are processed into fibers. Excellent review articles are available on the thermal and physical properties of these fibers [1], [20]. This section describes mainly meta-oriented aramid fibers of industrial importance.

Since Du Pont commercialized Nomex in 1967, various combinations of processes have been proposed for the production of aramid fibers based on poly(*m*-phenyleneisophthalamide) (PMIA) (Fig. 3). The polymer is prepared by low-temperature solution polymerization or interfacial polymerization according to the following reaction [21]:

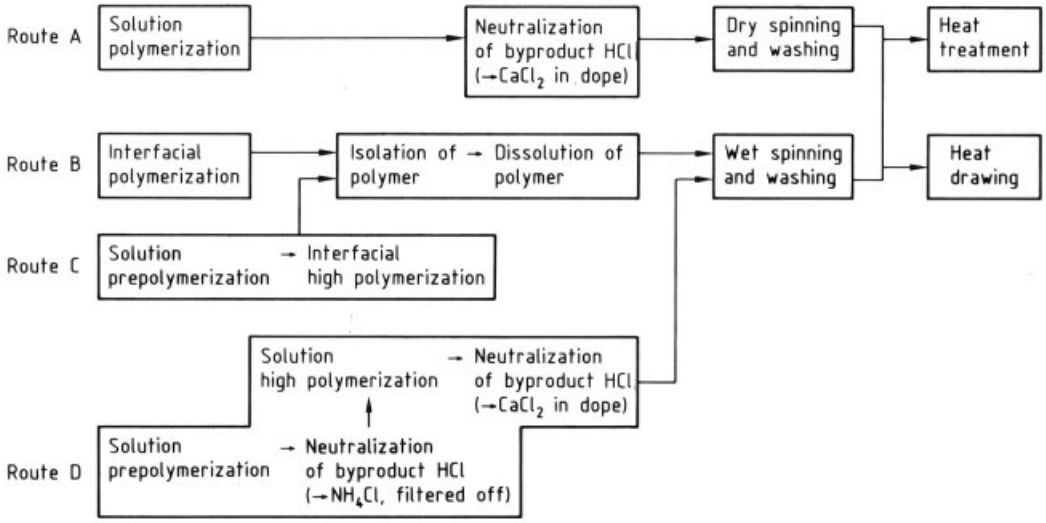

Figure 3. Production processes for PMIA fibers

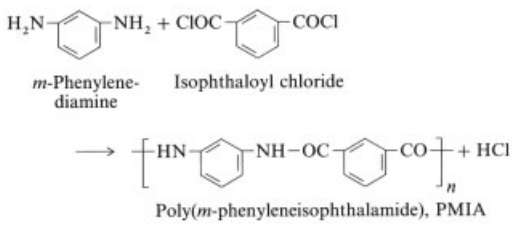

In both the solution and the interfacial methods, the following factors are important for the preparation of a high molecular mass polymer: use of high-purity monomers, stoichiometric balance of the two parent monomers, and minimization of the water content of the polymerization system. A small amount of a monofunctional compound such as aniline is generally used to control the degree of polymerization (Fig. 4) [22].

Polar amide solvents such as *N,N*-dimethylacetamide [*127-19-5*] are used in *low-temperature solution polymerization* [23]. These solvents are not only superior with regard to solubility but are also effective in activating electrophilic reagents. They act as acceptors of byproduct HCl, as well. Spin dope is obtained by neutralizing byproduct HCl with calcium hydroxide (Fig. 3, route A). This solution coagulates poorly in wet spinning because of the high content of calcium chloride and water. Accordingly, dry spinning is adopted in the Nomex process; residual calcium chloride and solvent are removed subsequently by water and heat treatment, respectively [24].

Interfacial polymerization is carried out at the interface between a cyclic ether solvent, such as tetrahydrofuran [*109-99-9*], and an aqueous alkaline carbonate solution as an acid acceptor (Fig. 3, route B) [25]. Polymers produced by the interfacial method contain a larger proportion of low molecular mass oligomers than those obtained by the solution method; this decreases the thermal stability of the fibers. In an improved interfacial method used by Teijin [26], preliminary oligomerization is carried out in tetrahydrofuran, and the resultant stoichiometric mixtures are subjected to interfacial polymerization (Fig. 3, route C) [27]. Subsequently, the powdery polymer is dissolved in *N*-methylpyrrolidone at a concentration of 20 – 23 wt % [28], and the resultant spin dope is wet-spun in aqueous calcium chloride solution [29]. The as-spun fiber is preoriented and subsequently hot-drawn at 200 – 350 °C.

Two modified processes have been proposed to overcome the poor coagulation of spin dope in the wet-spinning method. One utilizes a coagulant that gives void-free, as-spun fiber with high transparency [30]; The other involves reducing the calcium chloride content in the spin dope by removing hydrogen chloride in the form of ammonium chloride, a byproduct of the pre-

polymerization step (Fig. 3, route D) [23], [31], [32].

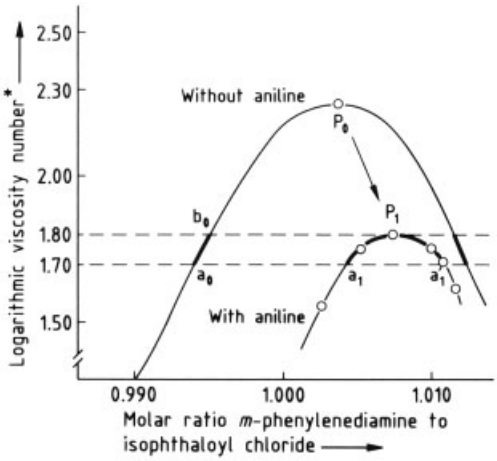

Figure 4. Control of the molecular mass of PMIA by addition of aniline [22]
Addition of aniline gives a wider molar ratio range for controlling η (i.e., molecular mass) within the limits 1.7 – 1.8 (compare $a_0 \rightarrow b_0$ and $a_1 \rightarrow a_1'$). P_0 and P_1 denote the maximum molecular mass with and without aniline, respectively.
∗ In aramid technology, molecular mass is usually expressed as a logarithmic viscosity number $\ln(\eta/\eta_0)/c$ (η = viscosity coefficient of solution, η_0 = viscosity coefficient of solvent, c = polymer concentration, dL/g).

Poly(*m*-phenyleneisophthalamide) is fusible in an inert atmosphere, although decomposition proceeds simultaneously in the molten state. Teijin has proposed a melt process for producing filaments by using a special spinneret, which has only its exit surface heated to avoid thermal decomposition [33].

The Mitsui Toatsu and Kuraray group has developed a new aramid fiber KM-21. Because the fiber is produced by polymerizing an aromatic isocyanate and aromatic carboxylic monomers, it differs from the foregoing PMIA fibers with respect to its chemical structure, production, and properties. However, its exact chemical structure and production method have not been made public [34].

$$\mathrm{OCN{-}Ar^1{-}NCO + HOOC{-}Ar^2{-}COOH} \longrightarrow \mathrm{\left[HN{-}Ar^1{-}NH{-}OC{-}Ar^2{-}CO\right]_n + CO_2}$$

where Ar = aryl

Fiber Properties. Important properties of meta-oriented aramid fibers are summarized in Table 3. The heat resistance and flame retardance of these fibers are better than those of commodity textile fibers. The physical properties of Teijinconex, a typical PMIA fiber, follow [35]:

Degree of crystallinity	35 – 39 %
Degree of orientation	90 – 93 %
Crystallite size	3.5 – 3.9 nm
Birefringence index	0.14 – 0.15
Density	1.37 – 1.38, g/cm^3

The moisture regain of this fiber is about 5 % under standard conditions (21 °C, 65 % R.H.) and 9 % at 95 % R.H., which is between equivalent values for cotton and nylon fibers.

Mechanical Properties. Most of the mechanical properties of PMIA fibers are about the same as those of commodity fibers (see Table 3). However, their wear resistance, both flexural and frictional, is lower than that of nylon or polyester fibers.

Weatherability and Dyeability. The weatherability of PMIA fibers is not inferior to that of nylon fiber, as far as strength retention is concerned (Fig. 5) [35]. However, the whiteness of PMIA deteriorates because the light-induced Fries rearrangement occurs readily.

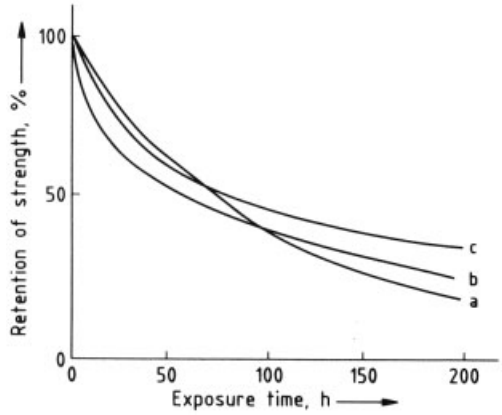

Figure 5. Weatherability of PMIA fiber [35]
a) Nylon; b) Teijinconex (PMIA); c) Polyester

Fibers from wholly aromatic polymers are, in general, highly sensitive to light exposure and poorly dyeable. Intrinsic color is a defect of wholly aromatic polymers, particularly when they are used for textile materials. Only PMIA

Table 3. Physical properties of meta-oriented heat-resistant aramid fibers (PMIA)

Property	Teijinconex, (Teijin) [35]		Nomex, (Du Pont) [36]		Fenilon (former Soviet Union) [37]	Apyeil, (Unitika) [38]	KM-21 ∗ (Mitsui Toatsu) [39]
	Regular staple fiber	HT type staple fiber	Staple fiber	Filament	Staple fiber	Staple fiber	Staple fiber
Density, g/cm^3	1.37 – 1.38	1.37 – 1.38	1.38	1.38	1.33 – 1.36	1.38	1.32
Tensile strength, GPa	0.55 – 0.67	0.85	0.49	0.65	0.38 – 0.43	0.55 – 0.61	0.61
Tensile modulus, GPa	6.95 – 9.99	12.16	8.53	17.06			9.75
Elongation at break, %	35 – 50	29	31	22	24.9 – 35.6	35 – 45	22
Moisture regain, %	5.0 – 5.2	5.5	5.5	5.5		5.5	
LOI	30 – 32	30	30	30		30 – 32	33
Heat shrinkage, %	4 – 6 (300 °C, 15 min)	5 (300 °C, 15 min)			2.6 – 4.6	>1 (300 °C, 15 min)	8 (400 °C, 10 min)

∗ Composition unknown; made by polymerizing aromatic isocyanate and aromatic carboxylate monomers.

is colorless and can thus be employed in dyed textiles.

Methods proposed for dyeing PMIA fibers include the use of carriers, cationic dyes, or dope dyeing. Attempts to improve dyeability have involved modifying the fiber structure [40] and blending PMIA with other polymers [41]. Most fiber manufacturers have their own proprietary products featuring improved dyeability or lightfastness.

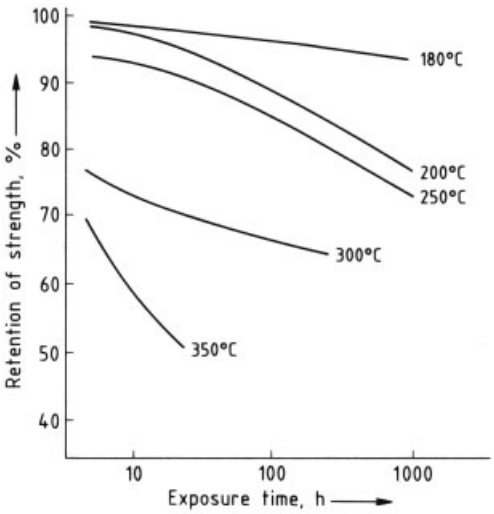

Figure 6. Strength retention in PMIA fiber exposed to dry heat [35]

The limiting oxygen index (LOI) is an important measure of flame retardance and represents the percent concentration of oxygen needed for self-supporting combustion. The LOI values of meta-oriented aramid fibers are given in Table 3.

Heat Resistance and Flame Retardance. Because PMIA exhibits high crystallinity and strong intermolecular cohesion due to hydrogen bonding, it has a high melting point and a high decomposition temperature. Accordingly, PMIA fibers have better thermal properties than commodity fibers. At elevated temperature, PMIA fibers offer better long-term retention of mechanical properties than commodity fibers; they also have good dimensional stability.

In the thermogravimetric analysis curve of a PMIA fiber, the decomposition temperature is 400 – 430 °C. Mechanical properties are almost unchanged down to −35 °C, and the fiber is still flexible with slight hardening.

Teijinconex (PMIA) staple fiber shrinks only 1 % at 250 °C (dry heat) and 5 – 6 % up to 300 °C. Long-term heat resistance of the fiber at various temperatures is shown in Figure 6.

Unmodified PMIA fibers shrink in a high-temperature flame. Improvement in dimensional stability and flame retardance has been achieved by treating PMIA fibers or fabrics with halogen, sulfur, and phosphorus compounds (e.g., Cl_2, Br_2, molten sulfur, $SOCl_2$, SO_2Cl, $POCl_3$, $POBr_3$, PCl_3, PCl_5) [42] and by blending with fibers that shrink less, such as Kevlar [43]; some products are available in flame-retardant grades. Thermal properties can be improved markedly by replacing one of the meta monomers by a para monomer as in PMTA (see Fig. 2). A self-

extinguishing PMTA fiber has even been developed [44]. The KM-21 fiber is also self-extinguishing and has a higher heat resistance than PMIA fibers (see Table 3).

Uses. Because of the excellent properties of PMIA fibers (e.g., high thermal and chemical resistance, as well as radiation resistance), their end uses are growing. Typical applications follow.

Clothing. Meta-oriented aramid fibers do not ignite, flare, or melt and stick to the skin. This makes them suitable for heat-resistant clothing material in the following areas:

1) heated furnaces: work uniforms, aluminized coats and pants, capes and sleeves, gloves and mitts, leggings, and spats;
2) emergency services: aluminized proximity suits, turnout coats and jumpsuits, station uniforms, rescue uniforms, fire-fighting and aviation garments, riot police uniforms, ranger uniforms, gloves, underwear, leggings, and spats;
3) fuel handling: work uniforms, rubber coats, gloves, socks, underwear, etc.

Interior Fittings. Materials from PMIA are used in aircraft interiors (for increased safety and enhanced flame retardance).

Industrial Materials. Uses here include filtration fabrics (especially filter bags for hot stack gases); high-temperature heat insulants (especially replacing asbestos); reinforcement in fire hoses, V-belts, and conveyor belts; threads for high-speed sewing; and cut-fiber reinforcement for rubber composites.

Electrical Insulation. High-temperature paper insulation for electric motors, dynamos, transformers, and cables; braided tubing for wire insulation; and dryer belts for papermaking are among the uses of PMIA fibers.

Miscellaneous Uses. Home ironing-board covers and kitchen gloves are also made from PMIA fibers.

2.2.2. Polyazole and Polyimide Fibers

Three types of wet-spun polyazole and polyimide fibers important for industrial and military uses are discussed:

1) polybenzimidazole fibers (e.g., Celazole),
2) poly(amide – imide) fibers (e.g., Kermel), and
3) polyimide fibers (e.g., Arimide).

Fiber Production. The polymer structure of the *polybenzimidazole (PBI) fiber* Celazole [*25734-65-0*] manufactured by Celanese is:

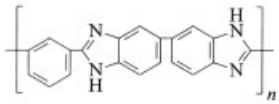

Production is described in detail in [45]. Polymerization is carried out in two steps. The first step involves conventional melt polymerization to yield a low molecular mass polymer with an intrinsic viscosity of 0.2 – 0.3 dL/g. The product is porous due to the formation of gaseous byproducts such as phenol and water. In the second step, this porous prepolymer is crushed to a fine powder and heated gradually to 380 °C to produce a fiber-grade polymer with an intrinsic viscosity $\geqq$0.75 dL/g. Both steps are carried out under a nitrogen atmosphere.

The polymer is powdered and dissolved in *N,N*-dimethylacetamide at a concentration of ca. 23 wt %. Lithium chloride in *N,N*-dimethylacetamide is reported to form a better solvent system for PBI than the acetamide alone.

A PBI fiber is obtained by dry spinning from a filtered dope, followed by drawing the as-spun fiber under a nitrogen atmosphere. The fiber obtained has excellent flame retardance, although heat shrinkage is appreciable. The sulfonation of imidazole residues in the PBI greatly improves its dimensional stability in flame-retardance tests.

Kermel (Rhône – Poulenc) is an example of a *poly(amide – imide) (PAI) fiber* [46].

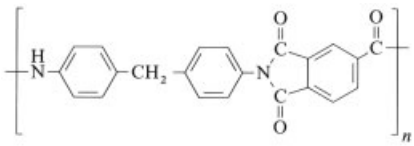

Polymerization is carried out in a polar amide solvent to form the precursor polymers.

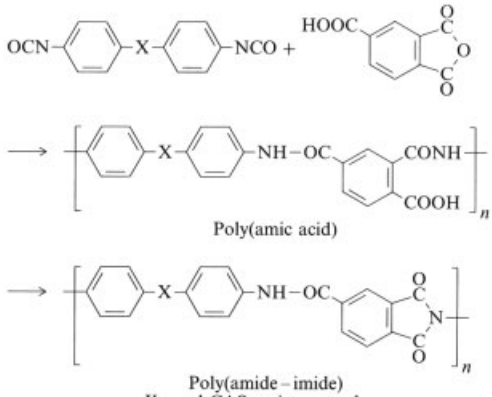

Kermel CAS registry numbers
X = O [*27084-44-2*], [*25038-81-7*]
X = CH_2 [*25821-35-6*]

A spin dope of PAI is obtained by in situ imidation of the precursor polymer [poly(amide – amic acid)] in solution by heating to ca. 100 – 120 °C; *N*-methylpyrrolidone is used as a solvent. The PAI fiber is spun from the dope and subjected to heat treatment.

In the case of *polyimide (PI) fiber* [47], spin dope containing poly(amic acid) is obtained. A precursor fiber is spun from the dope and then converted to PI fiber by heating, because PI is insoluble in amide solvents whereas its precursor polymer is readily soluble.

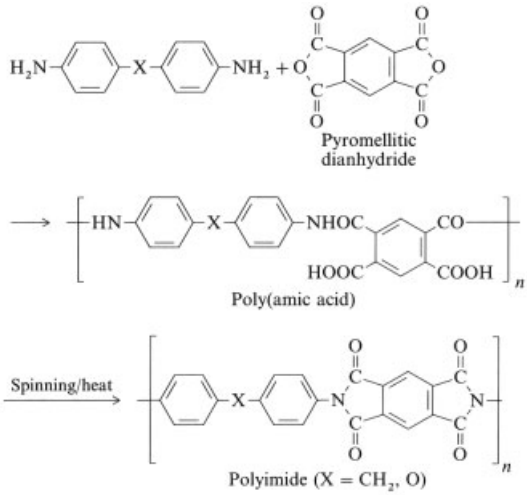

In Arimide fiber [*25036-53-7*] and [*25038-81-7*], X is oxygen.

Poly(bisbenzimidazobenzophenanthroline) [*39319-28-3*] (BBB) fiber attracted the attention of many polymer chemists and fiber scientists when it was reported by MARVEL et al. [48].

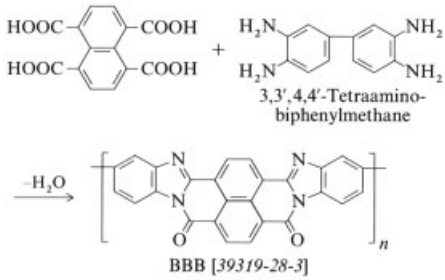

The fiber has excellent heat resistance, but no commercial product is known.

The BBB polymer is a double-stranded *ladder polymer* with a very high heat resistance. Ladder polymers have relatively high melting points and excellent weight retention as a function of temperature. However, ladder polymers are difficult to spin into fibers because of their low tractability.

Properties and Uses [49], [50]. Properties of typical PBI, PAI, and PI fibers are listed in Table 4. Polybenzimidazole fiber was developed in response to the needs of the U.S. Air Force to satisfy special requirements for the NASA space program. Its moisture regain is similar to that of wool, and its modulus is lower. Of fibers known at present, PBI has the highest LOI value; it tends to compete against PMIA fibers in its applications but is more expensive. Applications should exploit its resistance to chemicals and high temperature; they include

1) thermal-protection clothing: garments for fire fighters, industrial work uniforms, space and military flight suits, underwear and escape suits, and other aerospace materials;
2) high-temperature filtration: high-temperature filters (e.g., flue-gas filters); and
3) asbestos substitute: protective gloves for high-temperature uses.

The uses of meta-oriented PAI and PI fibers are similar to those of the meta-oriented aramid fibers. Kermel (PAI fiber) is used in underwear for pilots or operators of armed tanks because the resulting fabric is more supple than aramid fabrics.

Table 4. Properties of heat-resistant polybenzimidazole (PBI), poly(amide – imide) (PAI), and polyimide (PI), staple fibers

Property	PBI fiber (Celazole, Celanese) [49]	PAI fiber (Kermel, Rhône – Poulenc) [50]		PI fiber (Arimide, former Soviet Union) [47]
		234 AGF	235 AGF	
Density, g/cm^3	1.4	1.34	1.34	1.41 – 1.43
Tensile strength, GPa	0.33	0.47	0.2	0.63 – 0.7
Elongation at break, %	28.5	23	32	6 – 8
Tensile modulus, GPa	3.96	6.57	3.29	9.15 – 10.03
Moisture regain, %	15	3 – 5	3 – 5	1.0 – 1.5
LOI	>41	30 – 32	30 – 32	
Heat shrinkage, %	0.5	0 – 0.1	0 – 0.5	1.0 – 1.5

2.3. Melt-Spun Heat-Resistant Fibers

2.3.1. Novoloid Fibers (→ Phenolic Resins)

Fiber Production. The melt-spinning method has the advantage of eliminating the need for solvent recovery; however, melt-spun fibers are less heat-resistant than solution-spun fibers. Kynol [*9003-55-8*], a commercial novoloid fiber, exhibits exceptionally high heat resistance because cross-linking occurs after melt spinning. Kynol is produced from a novolac polymer by melt spinning [52] and subsequent cross-linking with formaldehyde [53]. The as-spun novolak reacts with formaldehyde in the presence of an acid catalyst to form cross-linked networks of methylene and dimethylene ether bonds.

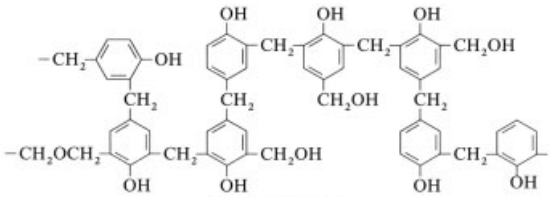

Kynol [*9003-55-8*]

Novoloid fiber was first developed by Carborundum as a carbon-fiber precursor and commercialized in the 1970s as a heat-resistant fiber [51]. It is produced by Gun-ei Chemicals and distributed by Nippon Kynol and American Kynol [54], [55].

Properties. One of the most remarkable features of novoloid fiber is its flame retardance. It generates virtually no smoke or gas. It has high LOI values and can be converted to carbon fibers without melting or shrinkage by heating in an inert atmosphere.

Kynol shows good thermal stability at 150 °C in air and 200 – 250 °C in the absence of air. Its chemical stability is also excellent. Kynol has excellent heat-insulation and acoustic properties. Its density is relatively low compared to other heat-resistant fibers.

The weatherability and light stability of Kynol are moderate. Dyeability is limited because of its intrinsic golden-brown color. Textile properties of Kynol fibers follow [56], [57]:

Diameter	14 – 33 µm
Density	1.27, g/cm^3
Tensile strength	0.15 – 0.20 GPa
Elongation at break	30 – 60 %
Tensile modulus	3.3 – 4.4 GPa
Loop strength	0.2 – 0.35 GPa
Knot strength	0.12 – 0.17 GPa
Elastic recovery (3 % elongation)	92 – 96 %
Moisture regain (20 °C, 65 % R.H.)	6 %
LOI	30 – 34

Uses. Novoloid fiber is used as reinforcement in resins and rubber to improve properties such as heat resistance, electrical insulation, and dimensional stability. Further applications are in cable material, brake material, various industrial sheets (e.g., welders' protective sheet, workplace shielding sheet), and safety products for accident prevention.

2.3.2. Fibers from Engineering Plastics

Fiber Production. Progress in both polymerization and fiber production technologies in the field of engineering plastics has stimu-

lated melt spinning of fibers from such plastics as poly(phenylene sulfide) [*9016-75-5*] (PPS) [57], polyetherether ketone [*31694-16-3*] (PEEK) [58], and polyetherimide (PEI) [58]; P84 [*65328-60-1*] fibers have also been reported [59]. Polymers for fiber production must have higher molecular masses than those used for plastics. They are spun into fibers in processes similar to conventional melt spinning (→ Fibers, 3. General Production Technology, Chap. 3.1.2.).

Some important technical problems are as follows:

1) The spinning apparatus must be improved to prevent gel formation at the surface of the spin die.
2) Spinning should be performed with a relatively viscous polymer melt.
3) The moisture content of the polymer chips should be reduced.

Properties and Uses. Properties of heat-resistant PEEK, PEI, and PPS fibers are listed in Table 5. The fibers normally have a better heat resistance than commodity fibers. They also offer an attractive performance – cost ratio.

Thermoplastic engineering polymers that can be utilized in fibers are also useful for composite materials. For example, PEEK fiber can be hybridized with carbon fibers for aircraft materials (usually called FF or fiber – fiber composite) [60]. Fiber – fiber composite has several advantages over conventional composites because its production does not involve prepregnation. Rationalization of the process and the better shapability of the matrix components (e.g., PEEK fiber) are definite advantages of FF composite technology.

Fiber making from high-performance thermoplastics is still in the developmental stage. Nevertheless, prospects for the further development and early commercialization of PEEK, PPS, and PEI fibers are promising.

Several reports on PPS fibers have cited their good dimensional stability, heat resistance, nonflammability, chemical resistance, and electrical properties [61]. Filtration fabric for baghouses of coal boilers is one promising application of PPS. Because of its high nonflammability, it could also be used for interiors and other parts of aircraft.

3. High-Strength and High-Modulus (HS – HM) Fibers

3.1. Introduction

Du Pont was successful in developing Nomex as a heat-resistant fiber in 1967, and Kevlar as a high-strength and high-modulus (HS – HM), poly(*p*-phenyleneterephthalamide) (PPTA) fiber in 1973. As a new member of the PPTA class, Twaron (Akzo) became available in 1986. Terlon [*63428-84-2*] (former Soviet Union) is also considered a member of the PPTA family, as is Technora (Teijin) which was commercialized in 1987.

The PMIA fiber-making technologies are based mainly on those of conventional solution-spun fibers such as rayon and polyacrylonitrile. The outstanding mechanical properties of Kevlar are achieved by using liquid crystalline (i.e., anisotropic) dope for fiber spinning. Spinning of anisotropic dope opened up a new era for HS – HM fibers. This concept has also been applied to making fibers from other liquid crystalline polymers, both lyotropic and thermotropic. Polyamidehydrazides [62] and polyazoles [63] belong to the lyotropic class, whereas polyarylates [64], polyazomethines [65], poly(amide – ethers), and poly(amide – esters) [66] are thermotropic polymers (Fig. 7).

As Flory et al. predicted [67] in 1956, a polymer with an essentially linear structure gives a nematic liquid crystalline solution in concentrated sulfuric acid when the polymer concentration exceeds a certain critical value. As shown in Figure 8, the viscosity of the solution increases with concentration in the optically isotropic region and reaches a maximum. At concentrations above this maximum, the solution starts to become optically anisotropic (liquid crystalline) and its viscosity passes through a minimum before finally increasing again [68].

The use of liquid crystalline dopes in the spinning process results in several advantages. A highly oriented fiber is obtained without the need for hot drawing. Anisotropic spinning can be carried out with a more highly concentrated dope than isotropic spinning. Panar et al. have reviewed solution spinning with regard to the

Table 5. Properties of heat-resistant PEEK, PEI, and PPS fibers

Property	PEEK (Teijin [58])	PEI (Teijin [58])	PPS (Phillips [57])
Density, g/cm^3	1.27 – 1.28	1.27	1.335
Tensile strength, GPa	0.28 – 0.34	0.27 – 0.29	0.36 – 0.42
Tensile modulus, GPa	5.5	4	3.63 – 4.84
Elongation at break, %	50 – 60	60 – 70	25 – 35
Moisture regain, %	0.20	0.25	0.6
Melting temperature, °C	343		285
Glass transition temperature (T_g), °C	143 – 145	217	93
Thermal decomposition temperature, °C	(500)	370	
LOI	35	47	34 – 35
Boiling water shrinkage			
Low shrink			3 – 5
High shrink			12 – 15
Heat shrinkage for 30 min, %			
At 180 °C	12 – 14	1.3	
At 200 °C	13 – 15	4.0	
At 220 °C	13 – 15	60	
At 240 °C	15 – 18	80	

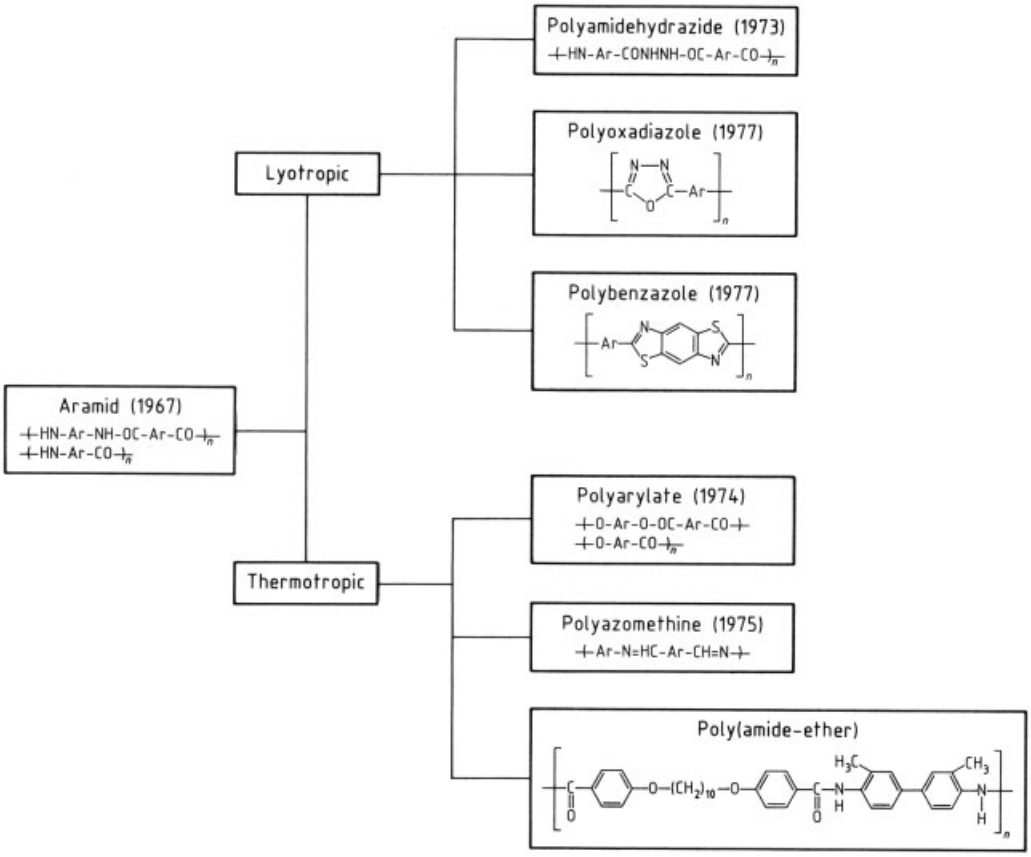

Figure 7. Classification and chemical structure of liquid crystalline polymers
Years in which patents were first published or applied for are indicated in parentheses; Ar = aryl.

structure and properties of anisotropic aramid solutions [69].

Research and development work on HS – HM fibers has continued. In particular, an extraordinarily HM fiber is obtained by spinning rod-shaped polymers, such as para-oriented polyazole [70], [71].

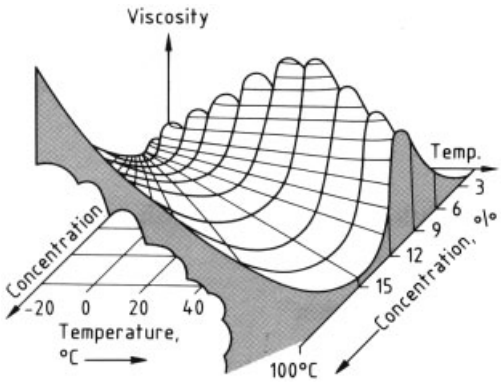

Figure 8. Viscosity as a function of polymer concentration and temperature [68]

Figure 9 shows the specific tensile strengths and moduli of organic and inorganic HS – HM fibers [72].

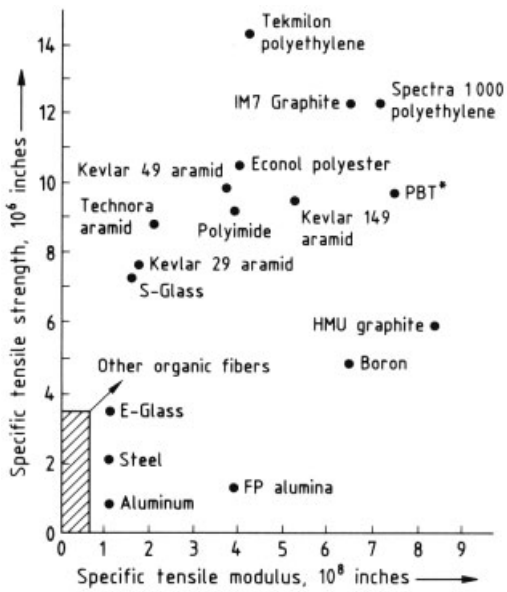

Figure 9. Tensile properties of high-performance organic and inorganic fibers [72]
∗ PBT = poly(*p*-phenylenebenzobisthiazole).

3.2. Solution-Spun HS – HM Fibers

Solution-spun HS–HM fibers that are already available or being developed are listed in Table 6 together with their chemical structures.

3.2.1. Para-Oriented Aramid Fibers

3.2.1.1. Production

Poly(*p*-phenyleneterephthalamide) [*24938-64-5*] (PPTA), a typical para-oriented aramid, is formed by the reaction

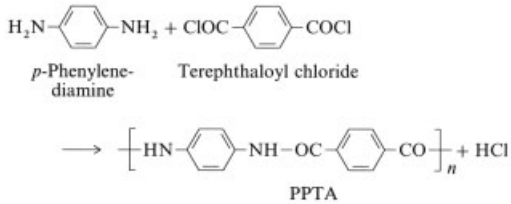

Hexamethylphosphoric triamide [*680-31-9*] was formerly used as a solvent in PPTA production. Since the toxicity of the amide was announced by Du Pont [73], several alternative solvent systems have been developed; these include polar amide solvents, which are particularly effective at higher concentrations of metal chlorides such as calcium chloride [74–76]. International legal conflict occurred between Du Pont and Akzo because Du Pont owned several patents relating to dope formation and the spinning of PPTA fibers, whereas Akzo owned the patents regarding polymerization. However, in 1988, both firms agreed to cross-license each other's patents [77].

The polymerization feed is an isotropic monomer solution; the mixture goes through an anisotropic state, and solidification occurs toward the end of the polymerization. Continuous stirring – shearing is important in order to obtain high molecular mass polymers, because the reaction mixture behaves like a Bingham fluid and exhibits plastic flow with a yield value [77]. During chain growth, mechanical shear stress applied to the polymerization mixture produces short, pulplike fibers without the need for a spinning step [78–80].

The production of PPTA fibers is summarized in Figure 10A. At a concentration of about

Table 6. Polymer structure of solution-spun HS – HM fibers

Polymer structure	Commercial name (manufacturer)
Anisotropic Polymers	
Aramid	
PPTA [*24938-64-5*]	Kevlar (Du Pont – Toray) Twaron (Akzo), Terlon (Soviet Union)
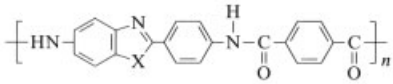 [*25068-38-6*] [*29471-34-9*]	Vniivlon (Soviet Union)
Polyazole	PBT, PBO (Stanford Research Institute International – Dow)
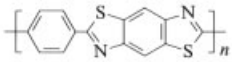 PBT [*69794-31-6*] [*75855-83-3*]	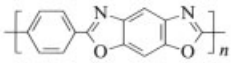 PBO [*60871-72-9*]
Isotropic Polymers	
Aramid	Technora (Teijin)
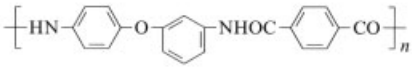	
Polyamidehydrazide	X-500 (Monsanto)
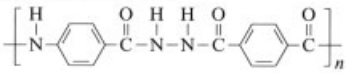 [*27102-68-7*]	
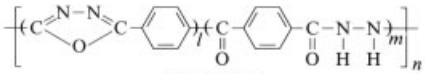 [*73714-39-3*]	X-715 (Monsanto)
Polyimide	Upilex (Ube Industries)
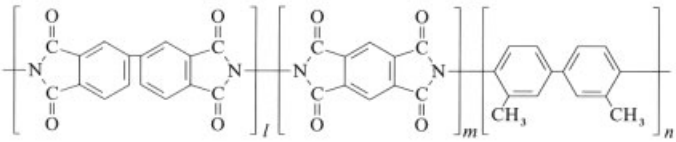	

20 wt %, PPTA gives an anisotropic spin dope suitable for dry-jet wet spinning at about 80 – 90 °C (Kevlar 29 fibers). In conventional wet spinning, the spin die (or cap) is located in the coagulation bath. However, in dry-jet wet spinning an air gap is provided between the cap surface and coagulation bath. The dry-jet wet spinning method offers the following advantages over conventional wet spinning:

1) At equivalent polymer concentrations a lower solution viscosity can be used because the dope temperature can be higher than that of the coagulation bath.

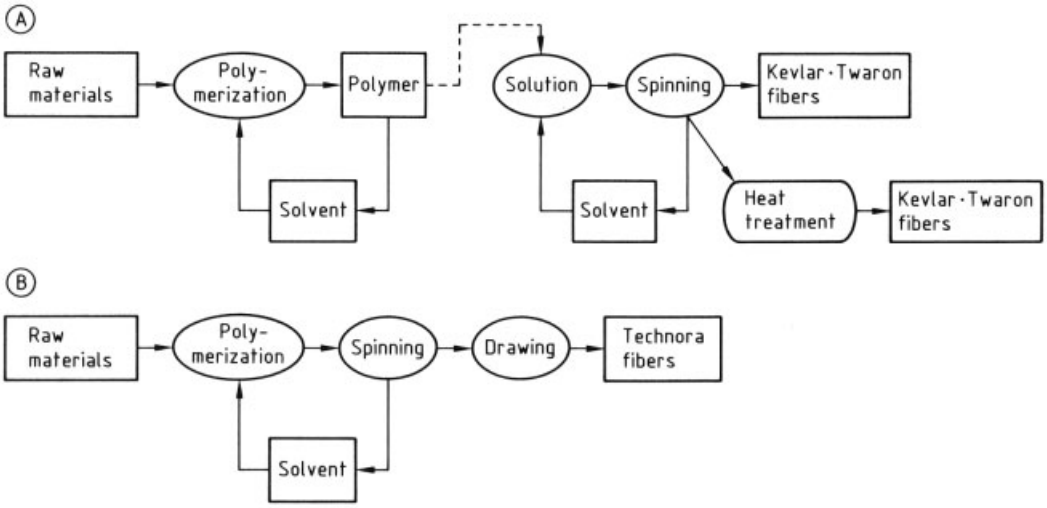

Figure 10. Production of HS – HM aramid fibers [82], [83], [88]
A) Kevlar and Twaron PPTA fibers; B) Technora fibers produced from a copolymer of PPTA
Polymer structures of the fibers are given in Table 6.

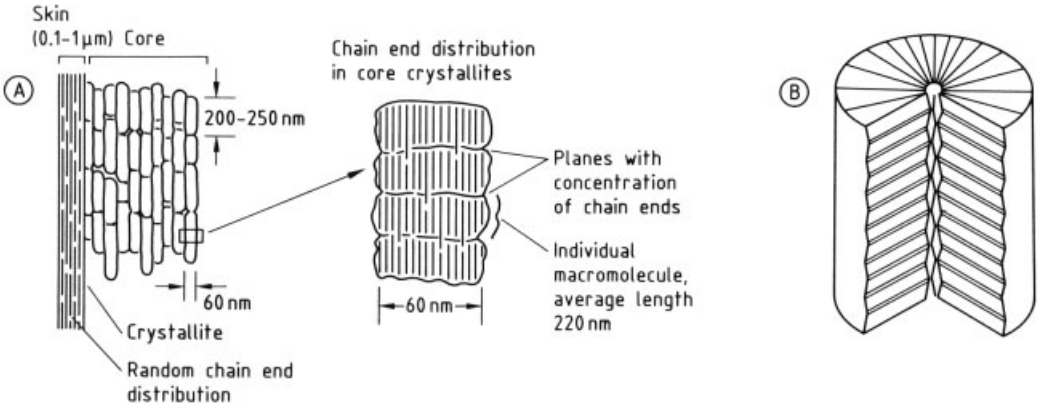

Figure 11. Structural models for PPTA fibers
A) "Skin – core" model [93], B) "Pleated sheet" model [95]

2) At equivalent solution viscosities a higher polymer concentration can be used, which improves the coagulation properties of the spun fiber.
3) The method favorably affects the coagulation properties of the spun fiber because the fiber surface releases a small amount of solvent as it passes through the air gap, and its temperature decreases.

The spin dope solidifies at room temperature and shows a sharp melting point at 79 – 80 °C. On the basis of X-ray diffraction data [81], the solidified dope is regarded as a crystal solvate or crystalline polymer complex. A crystal solvate is also observed in meta-oriented PMIA as a reflection of its polar structure [82].

"Anisotropic" spinning produces HS – HM fibers that have adequate mechanical properties even in the as-spun state [83]. The modulus can be improved by heat treatment of the as-spun fiber (Kevlar or Kevlar 29 fibers) [84].

Theoretical fiber moduli have been estimated for a variety of polymers [85], [86]. The tensile fiber modulus of Kevlar 149, announced by Du Pont [87], agreed with the value predicted for PPTA.

However, HS – HM fibers from semirigid aramid copolymers (e.g., Technora fibers from a copolymer of terephthalic acid and a diamine mixture consisting of 3,4′-diaminodiphenyl ether and *p*-phenylenediamine) can be produced by using the isotropic dope adopted in the meta-oriented aramid. The reaction mixture obtained by solution polymerization is then neutralized to give a spin dope, which is subjected directly to dry-jet wet spinning. The as-spun fiber is superdrawn (i.e., a very high draw ratio) at high temperature after the residual solvent is driven off [88]. Because only a single solvent is used throughout and the polymer does not have to be isolated, fiber production becomes very simple (Fig. 10B).

The most characteristic feature of the fiber is its exceptionally high drawability, which is significantly higher than that of nylon and polyester [89]. This property may be due partly to flow orientation from the thermotropic liquid crystal, but this has not been verified because the polymer starts to decompose before forming a clear melt.

3.2.1.2. Properties

Para-oriented aramid fibers all have high tenacity, high tensile modulus, and high heat resistance, which reflect their highly oriented rigid molecular structure. The properties vary slightly, however, with the chemical structure of the polymer and the microfine structure of the fiber. Several structural models have been proposed for PPTA fibers; Figure 11 shows the *skin – core* [69], [90–93] and *pleated sheet* models [94], [95].

Properties of Kevlar [87], [96]. The regular products are Kevlar, Kevlar 29, and Kevlar 49.

Mechanical Properties. Kevlar 149 has a higher modulus, and Kevlar HT [97] a higher tenacity, than the early Kevlar fibers (see Table 7). New brands of Kevlar such as Kevlar 68 and colored Kevlar have been announced in 1988 [96].

Kevlar 149 exhibits the highest tensile modulus of all commercial organic fibers, nearly twice that of E-glass but lower than that of carbon fiber from polyacrylonitrile. In contrast, its compressive properties are inferior to those of glass and carbon fibers. Kevlar displays structural defects ("kink bands") and buckling under compressive stress (see Fig. 12). This phenomenon is reported to correspond to the transition of carbon – nitrogen bonds in the amide group from the usual trans to the cis conformation [99]. The fiber will have further applications if its compressive strength is improved. This may be realized by reinforcement design improvements (→ Composite Materials).

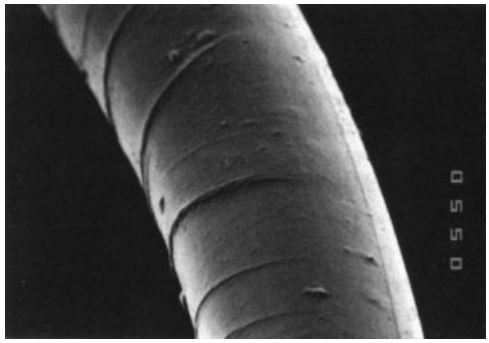

Figure 12. "Kink bands" in compressed PPTA fiber (×4500)

Thermal Properties. Kevlar shows no heat shrinkage up to 160 °C and remains stable below 180 °C for long periods. Although some scientists classify PPTA as a nonmelting polymer (i.e., the melting point exceeds the decomposition point), it is reported to have a melting point measured by differential thermal analysis [100]. The polymer carbonizes at approximately 430 °C. It is self-extinguishing and has an LOI value of 29.

Chemical Resistance. Kevlar has comparatively high chemical resistance to organic solvents (e.g., acetone, methyl ethyl ketone), sodium chloride solution, and hot water. However, its resistance to strong acid and strong alkali is inadequate (cf. Fig. 13).

Weatherability. Because the weatherability of Kevlar is not as good as that of other synthetic organic fibers, a protective coating is needed when the fiber is exposed to ultraviolet light.

Fatigue Properties. Kevlar has a tensile fatigue resistance almost comparable to that of nylon and polyester [101], but its flexural fatigue resistance is poorer [102]. Wear fatigue resistance is also poor because of the tendency toward fibril formation [103].

Table 7. Physical properties of para-oriented aramid fibers (PPTA)

Property	Kevlar 29 [96]	Kevlar 49 [96]	Kevlar 149 [87]	Kevlar HT [98]	Twaron (regular) [104]	Twaron (high-modulus) [104]	Technora (PPTA copolymer) [105]
Density, g/cm^3	1.44	1.45	1.48		1.44	1.45	1.39
Tensile strength, GPa	2.8	2.8	2.4	3.31	2.8	2.8	3.07
Tensile modulus, GPa	58	120	165	92.82	80	125	69.96
Elongation at break, %	4.0	2.5	1.3	3.3	3.3	2.0	4.4
LOI	29	29	29		29	29	25
Heat resistance at 200 °C, %	75 [a]	75 [a]			90 [b]	90 [b]	75 [c]
Alkali resistance, %							84
Acid resistance, %	10	10					89
Moisture regain, %	7 [d]	4 [d]	1		7	3.5	3

[a] 100 h.
[b] 48 h.
[c] 100 h.
[d] R.H. 55 %, 22 °C.

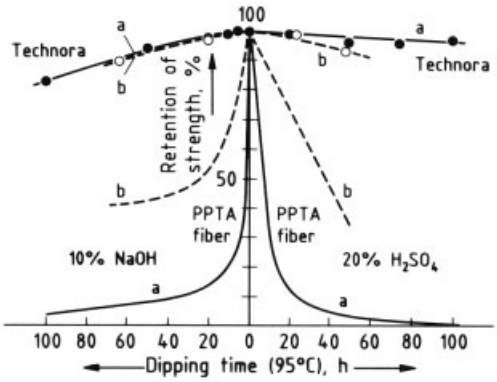

Figure 13. Resistance of para-oriented HM – HS aramid fibers (Technora and PPTA) to sodium hydroxide and sulfuric acid [88]
a) Tenacity;b) Intrinsic viscosity

Properties of Twaron Fiber. Being similar in chemical structure, Twaron is almost identical to Kevlar in many of its properties; see Table 7 [104].

Properties of Technora Fiber. Technora fiber (a PPTA copolymer) has excellent *mechanical and thermal properties* (Table 7) [105]. Its tensile strength is higher than that of PPTA fibers, whereas its tensile modulus is lower than that of high-modulus PPTA fibers.

Technora is an excellent heat-resistant organic fiber, although its thermal decomposition temperature is somewhat lower than that of PPTA fibers. Strength retention in dry heat is excellent. The LOI value of 25 is somewhat lower than that of PPTA fibers.

Chemical Resistance. Technora is highly resistant to hypochlorous acid and most organic chemicals. One characteristic feature is its extremely high chemical resistance to strong acid and alkali, which exceeds that of PPTA fibers [88]. The fiber deteriorates and its strength declines because its inherent viscosity decreases (Fig. 13).

The excellent chemical resistance of Technora presumably stems from its molecular structure of ether linkages, its microfine structure that prevents fibril formation, and its fine surface structure.

Fatigue Properties. Technora's comparatively high fatigue resistance may derive from the flexibility of its copolymer chain and the rather loose crystal structure, which minimizes fibril formation [102].

Para-oriented aramid fibers have a variety of outstanding properties, but some effort is still needed to improve their poor dyeability and poor adhesion to fiber-reinforced plastic matrices.

3.2.1.3. Uses

Because of their excellent mechanical and thermal properties, para-oriented aramid fibers are used widely as industrial materials (rope, cable, lightweight tension members); asbestos substitutes; antiballistic materials; and reinforcement

fibers in rubber (tires, belts, hoses, etc.), fiber-reinforced plastics including thermoplastics, fiber-reinforced metals, and fiber-reinforced cement or concrete. Typical applications are listed below.

Antiballistic Materials. Para-oriented aramid fibers are used in bulletproof jackets and boards, as well as in military helmets.

Tire Cord. Para-oriented aramid fibers can replace steel in tire cord. Consumption is low, however, primarily because of their high cost (a definite drawback of para-oriented aramid fibers), but also because of their inadequate adhesion to rubber and poor compressive fatigue properties. To increase their use, the price must be lowered and the properties improved.

Belts. Para-oriented aramid fibers are being considered for use as core cord for rubber belts; applications in synchronous power transmission belts and variable-speed belts in automobiles are especially promising. The fibers are also used as core cord for belts in continuously variable transmissions [106]. In 1986, Bridgestone Cycle began selling "Bellrex" bicycles, in which a belt reinforced with para-oriented aramid fiber replaces the metal chain [107]. Use of these fibers in conveyor belts has also been studied.

Hoses. Para-oriented aramid fibers are promising as reinforcement material for high-pressure hoses, heat-resistant hoses, and the heat-resistant radiator hoses of snowmobiles.

Composites. Composite materials are regarded as a use of aramid fibers (→ Composite Materials, → Reinforced Plastics).

Rope. Ropes utilizing the high specific strength of para-oriented aramid fibers are promising in the field of undersea development (e.g., ropes for offshore drilling rigs). Because of their light weight and high corrosion resistance, they are expected to replace metal chains and wire ropes. However, para-oriented aramid fibers are inadequate with regard to weatherability and wear resistance, and require a suitable coating (e.g., polyethylene).

Cable. The fibers are promising for the tension members of optical-fiber cables and are already being used as such.

Fiber-Reinforced Cement and Concrete. Asbestos and steel have conventionally been used to overcome the brittleness of cement and concrete, but health and corrosion problems create a need for new substitutes. Reinforcing concrete with chopped HS – HM fibers prevents cracking, and enhances both flexural and tensile strength [108], [109]. Fiber-reinforced concrete and cement are thus expected to be useful construction materials, for example, in curtain walls, floors, and ceilings.

Plastic rod with continuous-filament reinforcement is being developed. Its primary intended use is to replace steel in prestressed concrete, thereby eliminating corrosion problems [110].

Protective Clothing. The excellent cutting and piercing resistance of para-oriented aramid fibers is being exploited in protective materials used, for example, in safety gloves and work clothes for chain-saw operators.

Protective Materials. Para-oriented aramid fibers have excellent antiballistic properties, presumably due to their high tensile strength, high modulus, heat resistance, and easy fibril formation [111]. The typical use is in bulletproof vests and jackets.

3.2.2. HS – HM Fibers from Some Aramid-Related Polymers

A series of HS – HM fibers from aramid related polymers have followed the same pattern of development of para-oriented aramid fibers. The fibers (polyamidehydrazide, polyhydrazide, polyoxadiazole, polyamidehydrazide – oxadiazole) have analogous chemical structures and similar thermal and mechanical properties to PPTA fibers.

Monsanto has announced the development of X-715 (polyazolehydrazide) [112] and X-500 (polyamidehydrazide) fibers [113]; for structures, see Table 6. No commercial use of the fibers has been reported.

3.2.3. Polyazole and Polyimide Fibers

Fiber scientists in pursuit of an ultrahigh-modulus fiber have been concerned with perfectly linear polymers of rodlike molecular geometry. Although such fibers are expected to offer an extremely high modulus, the preparation of such high molecular mass polymers and the design of appropriate fabrication systems involve many difficulties.

3.2.3.1. Polyazole Fibers

Polyazoles used for fiber production are shown below:

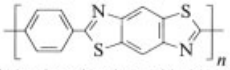

Poly(*p*-phenylenebenzobisthiazole)
[*69794-31-6*], [*75855-83-3*] (PBT)

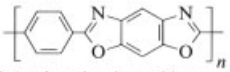

Poly(*p*-phenylenebenzobisoxazole)
[*60871-72-9*] (PBO)

Production. Rodlike polyazoles such as poly(*p*-phenylenebenzobisthiazole) (PBT) and poly(*p*-phenylenebenzobisoxazole) (PBO) can be made into ultrahigh-modulus fibers [71]. Iwakura et al. first proposed synthesizing polyazoles by using polyphosphoric acid as solvent [114]:

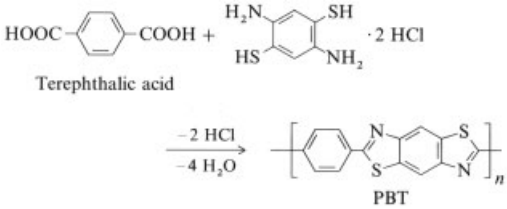

Wolfe et al. developed a sophisticated method for obtaining polymers with an intrinsic viscosity of 20 – 42 dL/g [63], [71]. Figure 14 shows the relationship between the intrinsic viscosity η and the molecular mass of the polymer [115], [116].

The liquid crystalline reaction mixture is used as the spin dope. Ultrahigh-modulus fibers are made by dry-jet wet spinning followed by heat treatment, in a process similar to that used for PPTA fibers (see Section 3.2.1.1).

As an extension of the rodlike polymers, fibers made of a molecular composite system have been proposed by Takayanagi et al. [117]. In this system, rodlike polymers are used as reinforcement and flexible polymers as the matrix component. An optically isotropic spin dope is used to make molecular composite fibers of polyazole [105], [118], or polyimide [119]. The reinforcing component in the molecular composite fiber system is reportedly well-dispersed in the matrix [120]. Such systems are attracting attention as new materials with excellent characteristics. Prospects for future development are promising because these fibers may exhibit unique characteristics not displayed by any known fiber. Production of a novel composite with a stress – strain curve totally different from those of its parent materials is conceivable.

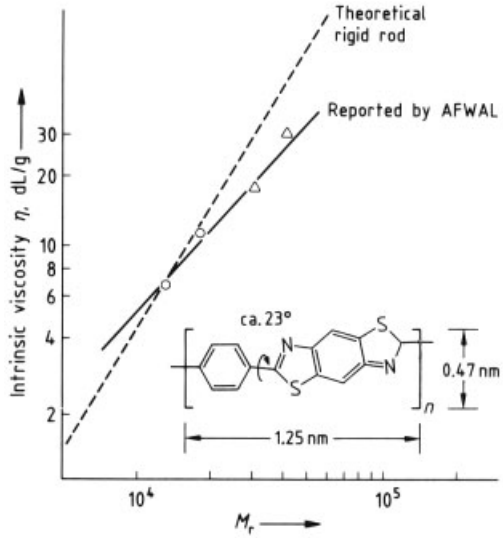

Figure 14. Relationship between molecular mass and intrinsic viscosity of poly(*p*-phenylenebenzobisthiazole) [115], [116]
∗ AFWAL = Air Force Wright Aeronautical Laboratories (United States).

Properties and Possible Uses. The fiber structure of PBT is similar to that of PPTA; it is composed of well-grown fibrils. The fibers readily form kink bands under applied forces; the polymer crystals probably glide along the crystalline plane of the aromatic rings [121]. The approximate tensile strength of PBT is 3.1 GPa; its tensile modulus is 340 GPa [71]. The approximate tensile strength of PBO is 3.4 GPa; its tensile modulus, 410 GPa [71]. Although paraoriented polyazole PBT and PBO fibers have the best thermal and mechanical properties of all known organic fibers, no practical uses have been developed because of their high production cost.

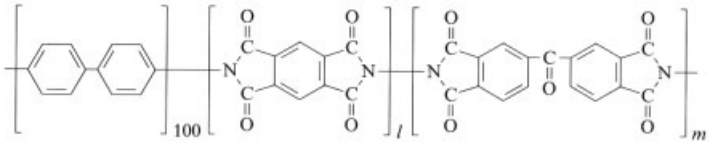

3.2.3.2. Polyimide Fibers

Production. Fibers from rigid polyimides are expected to form HS – HM fibers with a reasonable performance – cost ratio, because highly linear molecular structures can be synthesized from less expensive monomers than in the case of para-oriented polyazoles. However, studies on rigid polyimide fibers have not advanced far because the polymers are infusible and mostly insoluble. Copolymerization makes the polymers tractable.

With regard to para-oriented polyimides, several research groups have obtained high-modulus polyimide fibers by incorporating a third component to make the polymer tractable. Kotton et al. have made a significant contribution to the area of high-modulus PI fibers [122].

Ube Industries has announced HS – HM fibers (Upilex) obtained from a spin dope that contains polyimides based on biphenyltetracarboxylic dianhydride in 4-chlorophenol [123]. The chemical structure of Upilex is given in Table 6.

Because poly(*p*-phenylenepyromellitimide) (PPPI) and poly(*p*-phenylene) (PPP) are neither fusible nor soluble, fibers cannot be produced directly from these polymers [124], [125]. If tractable precursor polymers can be found, they can be converted to PPPI or PPP fibers by thermal or chemical treatment. As of now, only brittle PPPI and PPP fibers can be produced.

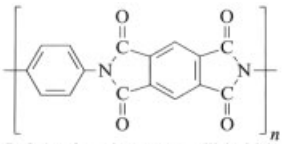

Poly(*p*-phenylenepyromellitimide) [*26099-65-0*] (PPPI)

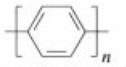

Poly(*p*-phenylene) [*25190-62-9*] (PPP)

Jinda et al. reported polyimide fibers from copolyimides based on pyromellitic dianhydride and substituted amines. They prepared the fibers by wet spinning the precursor poly(amic acid) solution and then imidating by thermal or chemical means [126]. Research on HS – HM polyimide fibers continues [127].

Properties and Possible Uses. Wholly aromatic polyimides rank among the most heat-resistant organic polymers. Despite a number of patents and research papers, PI fibers are not yet used commercially.

Some PI fibers have mechanical properties almost equivalent to those of para-oriented aramid fibers. Upilex fibers have a tensile strength of 2.9 GPa and a tensile modulus of 197 GPa [123]. The following polymer has a tensile strength of 2.6 GPa and a modulus of 151 GPa [126] (see top of this page).

The thermal properties of para-oriented PI fibers are better than those of PPTA fibers [123]. Their strength retention is better than that of PPTA fibers in a humid, saturated atmosphere at 200 °C. Weatherability is also better. Polyimide films are stable against radiation; this characteristic is also expected in PI fibers.

Para-oriented PI fibers, when available commercially, may be used in applications such as heat-resistant insulating materials, flameproof clothing, filter cloth, and atomic reactor materials.

3.3. Melt-Spun HS – HM Fibers

The success of Kevlar, based on the spinning of liquid crystalline polymers, triggered investigations of thermotropic para-oriented polyarylates [64]. Initially, use of these polymers was limited to engineering plastics, but it has been extended to fibers. Later, polyazomethine was found to form HS – HM fibers in the thermotropic state. Highly symmetrical polymers are infusible and, of course, insoluble because of their low affinity for polar solvents. Copolymerization or substitution of aromatic residues makes the polymers tractable. Typical fibers from thermotropic liquid crystalline polymers are shown below [128–130]:

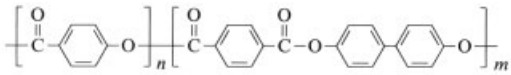

Econol [*59763-21-2*] (Sumitomo chemical)

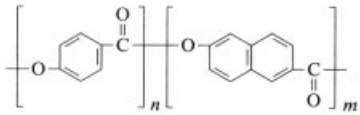

Vectran [*81843-52-9*] (Kuraray)

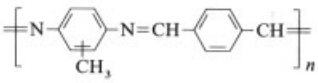

Polyazomethine [*61674-11-1*], [*61594-45-4*] (Du Pont)

Production. The rigid polyarylates are synthesized from derivatives of diphenols, dicarboxylic acids, and hydroxycarboxylic acids.

$$CH_3COO-Ar^1-OCOCH_3 + HOOC-Ar^2-COOH$$
$$+ CH_3COO-Ar^3-COOH \longrightarrow$$
$$\left[O-Ar^1-O-OC-Ar^2-CO\right]_n\left[O-Ar^3-CO\right]_m$$
$$+ CH_3COOH$$

where Ar = aryl

Melt polymerization is employed in most cases. The process is stopped when the melt reaches a viscosity suitable for spinning. The as-spun fibers are usually brittle. Subsequent heat treatment improves both the strength and the elongation because polymerization then continues in the solid phase, resulting in a further increase in molecular mass. This solid-phase polymerization takes rather long, and shortening the time is a pressing technical problem. Polyazomethine is produced by the following reaction:

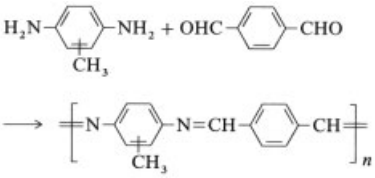

It must be spun in the form of liquid crystalline oligomers [130] because (1) the melt viscosity is suitable for spinning and (2) the liquid crystalline form of the oligomers means that the fiber can be highly oriented.

JACKSON et al. have proposed the term "self-reinforced polymers" for injection-moldable arylate block copolymers that consist of a liquid crystalline polyarylate reinforcing component and a flexible polyester matrix component [131]. Melt-spun fibers from self-reinforced polyarylates are analagous to solution-spun fibers from molecular composites. Other fiber types based on self-reinforced polymers are known [132].

Properties. Econol [128] and Vectran [129] are being developed. Table 8 lists some of their physical properties.

Econol fiber, announced in 1985 by Sumitomo Chemical, is a three-component copolymer. Its strength and modulus are higher and its elongation at break lower than those of Vectran (Table 8). The fiber seems promising for use in lightweight materials. It has extremely low moisture regain.

Vectran, announced by the Kuraray–Celanese group, is a two-component copolymer. The molar ratio of the components has not been disclosed. Its chemical resistance is excellent; its tensile strength and modulus are somewhat inferior to those of Econol but better than those of PPTA fibers.

Polyazomethine fiber properties have been reported [130]. As-spun fibers have a tensile strength of 0.97 GPa and a modulus of 121 GPa, whereas the values after heat treatment are 5.83 GPa and 148 GPa, respectively. The fibers are in the research stage. They resemble polyarylate fibers because they are spun from a relatively low molecular mass polymer melt. The as-spun fibers are usually brittle, and subsequent heat treatment improves strength and elongation due to solid-phase polymerization in the fiber shape.

Possible Uses. Econol and Vectran have several desirable properties. At present, both fibers are still at the market-development stage. A variety of test specimens (e.g., multifilament, chopped fiber, yarn, monofilament) are available for research applications. Tire cords, rubber reinforcement in belts and hoses, ropes, cables, tension members for optical-fiber cables, friction materials, and reinforcement fibers for various composites are examples of applications.

Vectran may also be produced as a pulp that can replace asbestos.

Table 8. Physical properties of polyarylate and polyazomethine fibers

Property	Econol (Sumitomo Chemical) [128]	Vectran (Kuraray –Celanese) [129]	Polyazomethine (Du Pont) [130]
Density, g/cm^3	1.40	1.40	1.50
Tensile strength, GPa	3.4 – 4.1	2.9	5.83
Tensile modulus, GPa	129 – 139	64.4	148.1
Elongation at break, %	2.4 – 3.1	3.7	
Moisture regain, %	0.01	0	
LOI		37	

The physical properties of both Vectran and Econol are similar to those of para-oriented aramid fibers (see Section 3.2.1.2), and competition between these two classes and high-performance polyethylene fibers appears unavoidable. Therefore, suitable applications must be found in which the low moisture regain of the fiber is exploited. This approach offers a way of overcoming the high cost of this type of fiber.

Problems still to be solved for these fibers include poor adhesion to rubber and poor chemical stability. The fibers are attacked by components present in rubber (i.e., amines, decomposition products of vulcanization accelerators, and antioxidants) and moisture.

4. References

General References

1. *Ullmann* 4th ed., **11,** 291 – 358.
2. P. W. Morgan, *J. Polym. Sci. Polym. Symp.* **65** (1978) 1.
3. J. Preston, J. Economy (eds.): *High-temperature and Flame Resistant Fibers,* Wiley-Interscience, New York, 1973.
4. G. C. Berry, C. E. Sroog (eds.): "Rigid Chain Polymers: Synthesis and Properties," *J. Polym. Sci. Polym. Symp.* **65** (1978) 222.
5. A. H. Frazer (ed.): "High Temperature Resistant Fibers," *J. Polym. Sci. Polym. Symp.* **19** (1967) 3 – 101.
6. M. Lewin, J. Preston (eds.): *High Technology Fibers,* Part A, Marcel Dekker, New York 1985.
7. W. B. Black, J. Preston (eds.): *Wholly Aromatic High-modulus Fibers,* Marcel Dekker, New York 1973.
8. C. B. Carter, V. T. J. Schenk in I. M. Ward (ed.): *Ultra-High Modulus Organic Fibers,* Halsted Press, New York 1975.
9. H. Kanbe (ed.): *Heat Resistance of Polymers (in Japanese),* Baifu-Kan, Tokyo 1970; *Thermal Decomposition and Heat Resistance of Polymers (in Japanese),* Baifu-Kan, Tokyo 1974.
10. L. B. Sokolov: *Thermally Stable Aromatic Polyamides (in Russian),* Izdatelstov, Ximiya, Moscow (1975).

Specific References

11. D. W. van Krevelen, *Angew. Macromol. Chem.,* **22** (1972) 133.
12. O. Yoda, *Polym. Commun.* **25** (1984) 238.
13. T. Inabe, J. F. Lomax, T. J. Marks, J. W. Lyding et al., *Macromolecules* **17** (1984) 260.
14. Du Pont, US 3 671 542, 1972 (S. L. Kwolek); 3 991 037, 1976 (P. Eckhardt).
15. H. F. Mark, S. M. Atlas, N. Ogata, *J. Polym. Sci.* **61** (1962) no. 172, 49.
16. G. A. Kuznetsov, V. M. Savinov, L. B. Sololov, V. K. Belyakov et al., *Vysokomol. Soedin. Ser. A* **11** (1969) 1491.
17. K. Kuze, S. Miwa, *Kogyo Kagaku Zasshi* **71** (1968) 448.
18. H. Mera, Y. Nakagawa, unpublished observations (1973).
19. F. Dowell, NTIS 1 298 748 (1987).
20. *Kirk-Othmer,* **3,** 213.
21. P. W. Morgan: *Condensation Polymers: By Interfacial and Solution Methods,* Interscience Publisher, New York 1965.
22. H. Mera, K. Shimada, unpublished observations (1979).
23. Du Pont, US 3 063 966, 1962 (S. L. Kwolek, P. W. Morgan).
24. Du Pont, US 3 006 899, 1961 (H.-W. Hill, Jr., S. L. Kwolek).
25. Du Pont, US 3 360 598, 1967 (C. R. Earnhart).
26. S. Ozawa, Preprint for Annual Meeting, Soc. Polymer Sci., Japan, 340 1 – 85 (1982).
27. Teijin, US 3 640 970, 1974 (H. Hawkins).
28. Teijin, US 3 963 664, 1976 (A. Aoki, H. Fujie, S. Ozawa).

29. Teijin, JP 48-17551, 1973 (K. Kozai, Y. Tabe, H. Hond, K. Matsuda et al.).
30. Unitika, JP-Kokai 51-564, 1976 (S. Inokai, K. Kimura, Y. Tanaka, G. Shoji).
31. E. P. Krasnov et al., *Khim. Volokna* **13** (1971) 48.
32. L. V. Prozorov, B. I. Shizdyuk, A. S. Chegolya, Deposited Doc., SPSTL 45 khp-D81. 86 pp (Avail.SPSTL).
33. Teijin, US 4 526 735, 1985 (S. Norota, Y. Segawa, T. Yamanchi).
34. Mitsui-Toatsu, JP-Kokai 61-201009, 1986 (M. Ohsawa, M. Tinno, K. Shizu, T. Toyota et al.).
35. Teijin, Teijinconex Technical Information, 1973, Osaka, Japan.
36. Du Pont, Nomex Technical Information.
37. A. Aoki, K. Shimada, unpublished observation (1973).
38. Unitika, Apyeil Technical Information, 1986, Osaka, Japan.
39. Mitsui-Toatsu, KM-21 Technical Information.
40. Du Pont, US 3 506 990, 1970 (J. A. Richardson, R. E. Walcke Jr.).
41. Teijin, US 4 278 779, 1981 (K. Matsuda, Y. Nakagawa, T. Nakamura, K. Shimada).
42. C. E. Hatherway, C. L. Early, *Appl. Polym. Symp.* **21** (1973) 101. Monsanto, US 3 576 769, 1971 (S. S. Hirsch, J. R. Holsten).
43. H. Fujie, *Nikkakyo Geppo* **8** (1987) 8 – 15.
44. Du Pont, US 4 198 494, 1980 (W. G. Burckel).
45. A. B. Conciatori, B. Alan, D. E. Stuets: J. Preston (eds.): Polybenzoimidazole Fibers, *High Technology Fibers,* **Part A,**Chap. 6 Marcel Dekker, New York 1985, pp. 221 – 267.
46. J. Chambion, *Chim. Ind. Genie Chim.* **106** 1973, no. 7 453.
47. G. I. Kudryavtsev, L. V. Balakleitseva, *Khim. Volokna* **12** (1970) 42; **11** (1969) 66.
48. W. H. Gloor, *Appl. Polym. Symp.* **6** (1967) 151; **9** (1969) 159.
49. Celanese, Celazole Technical Information.
50. Rhône-Poulenc, Kermel Technical Information.
51. J. Economy, *Polym. News* **2** (1975) no. 7 – 8, 13 – 19.
52. Nippon Kynol, US 3 972 959, 1976 (H. Koyama).
53. Nippon Kynol, US 4 021 410, 1977 (I. Kimura, H. Koyama).
54. F. Ogata, *Gosei Jushi* **32** (1981) no. 12, 86.
55. *Kirk-Othmer,* **16** 125.
56. Nippon Kynol, Kynol Technical Information.
57. G. H. Hagler, J. G. Scruggs, *Int. Fiber J.* **2** (1987) no. 2, 40.
58. Teijin, The Brochure of Teijin Innovative Thermoplastic Fibers, 1988, Osaka, Japan.
59. Lenzing AG, EP 119 185, 1984 (Klaus Weinrotten et al.).
60. S. R. Clemans et al., Technical Information of BASF Structural Materials, *Hybrid Yarns for High Performance Thermoplastic Composites.*
61. S. Takeda, *Kogyo Zairyo* **36** (1988) no. 6, 26.
62. Teijin, JP-Kokai 49-132 150, 1974 (T. Shinoki, H. Mera, S. Ozawa). Du Pont, US 3 526 611, 1970 (G. F. D'Alelio).
63. SRI Int., US 4 225 700, 1980 (B. H. Loo, J. F. Wolfe).
64. Du Pont, US 3 991 013, 1976 (T. C. Plecher); US 3 991 014, 1976 (J. J. Kleinschuster).
65. P. W. Morgan, T. C. Pletcher, S. L. Kwolek, *Polym. Prepr. Am. Chem. Soc., Div. Polym. Chem.* **24** (1983) no. 2, 470.
66. W. J. Jackson Jr., H. F. Kuhfuss, *J. Appl. Polym. Sci* **25** (1980) 1685. A. C. Griffin, *Mol. Cryst. Liq. Cryst.* **82** (1982) 145.
67. P. Flory, *Proc. Soc. London Ser. A* **234** (1956) 73.
68. T. Kikuchi, *Hyomen* **20** (1982) 270.
69. M. Panar et al., *J. Polym. Sci., Polym.-Phys. Ed.* **21** (1983) 1955.
70. S. R. Allen et al., *Macromolecules* **14** (1981) 1135.
71. SRI Int., WO 84/01160, 1984 (J. F. Wolfe, P. D. Sybert; WO 84/01162, 1984 (J. F. Wolfe, P. D. Sybert).
72. J. R. Schaefgen, *Internatl. Fiber J.* **2** (1987) no. 2.
73. *Chem. Eng. News,* (1975) Sept. 19, 17.
74. AKZO, US 4 308 374 (NL 157 327), 1981 (T. J. Veerman, L. Vollbracht).
75. Teijin, US 4 389 521, 1983 (A. Aoki, H. Mera, T. Nakamura, K. Shimoda).
76. L. B. Sokolov, GB 1 542 550 (NL 76-07 863), 1979.
77. P. L. Layman, M S. Reisch, *Chem. Eng. News* Aug. 22 (1988) 10. Du Pont, US 3 884 881, 1975 (A. R. Bice, J. A. Fitzgerald, A. E. Hoover).
78. Du Pont, US 3 817 941, 1974 (J. Bain, P. Morgan).
79. Tejin, JP 61-42004, 1986 (H. Mera, Y. Nakagawa, K. Matsuda).
80. KAIST, US 4 511 623, 1985 (Y. W. Son, H. S. Yoon).
81. M. M. Iovleva et al., *Vysokomol. Soedin. Ser. B* **24** (1982) 42.
82. J. Nakauchi, H. Kakida, S. Minami, T. Yoshihara, *J. Polym. Sci., Polym. Lett. Ed.* **15** (1977) 527;Mitsubishi Rayon, JP-Kokai 54-84285, 1979 (K. Sagawa).

83. Du Pont, US 3 767 756, 1973 (H. Blades); USP 3 869 429, 1975 (H. Blades).
84. Du Pont, US 3 869 430, 1975 (H. Blades).
85. H. Tadokoro, *Sen'i Gakkaishi* **31** (1975) no. 9, 278.
86. T. Ii, K. Tashiro, M. Kobayashi, H. Tadokoro, *Macromolecules* **20** (1987) 347; **19** (1986) 1809. K. Tashiro, M. Kobayashi, *Sen'i Gakkaishi* **43** (1987) no. 2, 78.
87. J. E. van Trump, J. Lahijani, *Proceeding of the 32nd Int. SAMPE Symp.,* April 6 – 9, 1987, pp. 917 – 927.
88. S. Ozawa, *Polym. J. (Tokyo)* **19** (1987) 119. Teijin, US 4 075 172, 1978 (K. Matsuda et al.).
89. S. Ozawa, K. Matsuda in M. Leving, J. Preston (eds.): *Aramid Copolymer Fibers,* "High Technology Fibers," Part B, Marcel Dekker, New York, in press.
90. L.-S. Li, L. F. Allard, W. C. Begelow, *J. Macromol. Sci. Phys.* **B21** (1983) 1955.
91. H. Manabe et al., *Sen'i Gakkaishi* **33** (1980) T-101.
92. R. J. Morgan, C. O. Pruneda, W. J. Steele, *J. Polym. Sci. Polym. Phys. Ed.* **15** (1977) 1757.
93. K. Yabuki, H. Itoh, T. Ohta, *Sen'i Gakkaishi* **31** (1975) T-524.
94. M. G. Dobb et al., *J. Polym. Sci., Polym. Phys. Ed.* **15** (1977) 2201.
95. K. Matsuda, T. Nishihara, T. Watanabe, *The 1984 Int. Chem. Congress of Pacific Basin,* 09p02, Honolulu, Hawaii, Dec. 1984.
96. Du Pont, Kevlar Aramid Technical Information, 1986; Du Pont: *High Tech Materials Exhibition 1988*, Tokyo 1988.
97. Du Pont: *Special Report: Organic Fibers,* Performance Materials, May 1987, pp. 1 – 4.
98. S. Tada: "Kevlar as Rubber Reinforcements," *Meeting of the Japan Rubber Society,* Kansai, July 5, 1988.
99. D. Tanner, J. A. Fitzgerald, W. F. Knoff, J. J. Pigliacampi, *Proceedings of the International Symposium on Fiber Science and Technology* **31** ISF-85, Hakone, Japan 1985.
100. A. R. Brown, B. C. Ennis, *Text. Res. J.* **42** (1977) 62.
101. A. R. Bunsell, *J. Mater. Sci.* **10** (1975) 1300.
102. T. Takata: "Material Sience," *Zairyo Kagaku* **21** (1985) 348.
103. S. Kazama, *Preprints for Plastics Engineering Forum,* The Society of Polymer Science, Japan, 1981, p. 3.
104. Enka Glanzstoff, Twaron Technical Information, 1987, Wuppertal, Federal Republic of Germany.
105. Teijin, Technora Technical Information, 1987, Osaka, Japan.
106. *Rubber & Plastics News* June 17 (1985) 27 *Rubber Post,* No. 1130, August 26 (1985) 5.
107. J. Uemae, DIME, 6 Nov. (1986).
108. P. L. Walton, A. J. Majumdar, Building Research Establishment, CP 57/78, Juli (1978), p. 11.
109. S. Akihama, H. Nakagawa, T. Takata, M. Yamaguchi, *Developments in Fibre Reinforce Cement and Concrete, RILEM Symp. FRC 86,* Sheffield, UK, 13 – 17 July, 1986.
110. K. Kobayashi, R. Cho, *Seisan Kenkyu* **36** (1984) no. 8, 24.
111. R. C. Laible: *Ballistic Materials & Penetration Mechanics,* Elsevier, Amsterdam 1980.
112. Monsanto, GB 1 407 439, 1975 (H. C. Bach); 1 417 568, 1977 (H. C. Bach).
113. Monsanto, US 3 796 693, 1974 (W. Black, H. Morgan, J. Preston).
114. Y. Imai, K. Uno, Y. Iwakura, *Macromol. Chem.* **83** (1965) 179. Teijin, JP 42-19273, 1965 (Y. Iwakura, K. Uno, T. Imai, I. Tanaka).
115. W. F. Hwang, D. R. Wiff, C. L. Benner, *J. Macromol. Sci., Phys.* **B22** (1983) 231.
116. W.- F. Hwang, D. R. Wiff, *Am. Chem. Soc. Div. Org. Coat. Plast. Prep.* **44** (1981) 32.
117. M. Takayanagi, T. Ogata, *J. Macromol. Sci. Phys.* **B17** (1980) no. 4, 591.
118. T. E. Helminiak, W. W. Adams, *Polym. Eng. Sci.* **23** (1983) no. 14, 784;USA WO 79/01029, 1979 (T. E. Helminiak, F. E. Arnold, G. E. Husman).
119. R. Yokota et al., *Polym. Prepr. Japan* **35** (1986) no. 10, 3304.
120. T. Nishihara, H. Mera, K. Matsuda, *Am. Chem. Soc. Polym. Matr. Sci. Eng.* **55** (1986) 821.
121. E. W. Choe, S. W. Khim, *Macromolecules* **14** (1981) 920.
122. M. M. Kotton, *Vysokomol. Soedin. Ser. A* **21** (1979) no. 11, 2496.
123. T. Kaneda, T. Katsura, K. Nakagawa, H. Makins, *J. Appl. Polym. Sci.* **32** (1986) 3133, 3151.
124. J. G. Speight, P. Kovacic, F. W. Koch, *J. Macromol. Sci. Rev. Macromol. Chem.* **C5** (1971) no. 2, 295.
125. D. G. H. Ballard, A. Courtis, I. M. Shirley, S. C. Taylor, *J. Chem. Soc. Chem. Commun.* 1983 954.
126. T. Jinda, T. Matsuda, M. Sakamoto, *Sen'i Gakkaishi* **42** (1986) no. 10, T-554.
127. D. Yu Likhachev et al., *Vysokomol. Soedin. Ser. B* **27** (1985) 723.
128. Sumitomo Chem., Econol Technical Information.

129. K. Ueda, *Sen'i Gakkaishi (in Japanese)* **43** (1987) no. 4, 135.
130. P. W. Morgan, *J. Polym. Sci. Polym. Symp.* **65** (1978) 1.
131. W. J. Jackson Jr., *J. Polym. Sci. Polym Chem. Ed.* **14** (1976) 2043.
132. Phillips, NL 86/00307, 1986 (D. J. Broer, G. N. Mol, C. H. Jochem). JP-Kokai 62-191506, 1987 (D. J. Broer, G. N. Mol, C. H. Jochem).

High-Temperature Materials

High-Performance Fibers and Polymers, High-Temperature are separate keywords.

G. W. MEETHAM, Rolls-Royce plc, Derby, United Kingdom

1. Introduction . . . 2
2. Factors Influencing Temperature Capability . . . 5
2.1. Melting Temperature . . . 5
2.2. Volatility . . . 6
2.3. Crystal Structure . . . 6
2.4. Zero-Time Deformation . . . 6
2.5. Creep . . . 7
2.6. Mechanical Fatigue . . . 8
2.7. Thermal Fatigue . . . 9
2.8. Oxidation . . . 9
2.9. Hot Corrosion . . . 10
2.10. Carburization . . . 11
2.11. Erosion . . . 11
2.12. Wear . . . 11
3. Mechanisms for Increasing Temperature Capability . . . 12
3.1. Metallic Materials . . . 12
3.1.1. Solid-Solution Strengthening . . . 12
3.1.2. Precipitation Hardening . . . 12
3.1.3. Dispersion Strengthening . . . 15
3.1.4. Grain Boundary Effects . . . 15
3.2. Nonmetallic Materials . . . 16
3.3. Corrosion-Resistant Coatings . . . 17
3.3.1. Aluminum . . . 17
3.3.2. Steel . . . 17
3.3.3. Nickel . . . 17
3.4. Insulating Coatings . . . 19
3.5. Wear-Resistant Surfaces . . . 20
4. Aluminum Alloys . . . 21
4.1. Conventional Alloys . . . 21
4.1.1. Wrought Alloys . . . 21
4.1.2. Casting Alloys . . . 23
4.2. Powder Alloys and Rapidly Solidified Alloys . . . 23
4.2.1. Sintered Aluminum Powder . . . 23
4.2.2. Recent Powder Alloys . . . 23
4.3. Metal-Matrix Composites . . . 24
5. Titanium Alloys . . . 25
5.1. Metallurgy . . . 25
5.2. Uses . . . 28
6. Steels . . . 29
6.1. Low-Alloy Steels . . . 29
6.2. Martensitic 12 % Chromium Steels 30
6.3. Ferritic Heat-Resistant Materials 31
6.4. Corrosion-Resistant Austenitic Steels . . . 31
6.5. Controlled-Transformation Stainless Steels . . . 32
6.6. High-Strength Austenitic Steels . . 32
7. Nickel Alloys . . . 33
7.1. Metallurgy . . . 33
7.2. Oxidation-Resistant Nickel Alloys 35
7.3. Nickel Superalloys . . . 36
7.3.1. Turbine Rotor Blade Materials . . . 36
7.3.2. Turbine Stator Blade Materials . . . 39
7.3.3. Turbine Disk Materials . . . 41
7.3.4. Sheet Materials for Fabricated Structures . . . 43
7.3.5. Nonaerospace Applications . . . 44
8. Refractory Metals . . . 44
9. Engineering Ceramics . . . 46
9.1. Refractories . . . 46
9.2. Early Engineering Ceramics . . . 46
9.3. Advanced Engineering Ceramics . 47
9.3.1. Ceramic Processing . . . 47
9.3.2. Thermal Shock Resistance . . . 47
9.3.3. Component Design and Life . . . 48
9.3.4. Defect Tolerance . . . 48
9.3.5. Alumina . . . 49
9.3.6. Zirconia . . . 49
9.3.7. Silicon Nitride . . . 50
9.3.8. Sialons . . . 50
9.3.9. Silicon Carbide . . . 51
9.3.10. Glass Ceramics . . . 51
10. Carbon and Carbon–Carbon Composites . . . 52
11. Ceramic Composite Materials . . . 54
12. Intermetallic Materials . . . 55
13. Thermal Ablative Materials . . . 56
14. Thermal Insulating Materials . . . 57
15. References . . . 58

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_025

1. Introduction

Definitions. High-temperature materials are those materials that are used specifically for their heat-resisting capabilities for which the temperature criterion (*capability*) is expressed as the ratio of the maximum use temperature to the melting temperature (T_m) rather than simply as the maximum use temperature. Various material systems are compared on this basis in Table 1.

Mechanical properties and environmental resistance are major factors in the selection of such materials, but physical properties can also have a significant influence. Density is particularly important in rotating machinery. Thermal expansion coefficients are important in designs involving different materials in close contact, and thermal stress is controlled by the thermal expansion coefficient, thermal conductivity, and Young's modulus.

In economic terms, the most cost-effective material in any application is the material that satisfies the total operational requirement at minimum cost. The cost element must include consistency of component manufacture in relation to its specification.

Applications Requiring High-Temperature Materials. Thermal energy is used in three ways: (1) for conversion to mechanical or electrical power, (2) in processing materials and controlling the environment, and (3) as a sink for dissipating other energy forms by design or because of the inefficiency associated with another process. High-temperature technology is thus of major importance in many industries including primary metal and nonmetal production, material processing, chemical engineering, transportation, power generation, and domestic appliances. For many of these industries, the price of fuel is a major component of overall operating cost. Materials that allow operation at high temperature are essential for industrial competitiveness because the efficiency of fuel conversion and use is related to operating temperature. Current operating temperatures in various industries are shown in Figure 1 together with the melting temperatures of selected materials. Up to ca. 4000 °C more than one material option may exist, depending on operational requirements. However, no material is capable of operating above 4000 °C; here, thermal ablation must be used.

Many industries could benefit from higher operating temperatures than those currently in use. The temperature of superheater tubes for steam-raising boilers in power plants is currently 560 °C; if this could be increased to 650 °C, efficiency would increase from 35 to 36.5 %. Although the higher temperature produces a more aggressive environment and requires new materials, it would also result in saving ca. 3.5×10^6 t of coal per year, i.e., 4 % of the total coal used in the United Kingdom for power generation [8]. The use of ceramics in the development of diesel engines in the United States has allowed higher operating temperatures, thus improving efficiency (an increase from 38 to 65 % appears possible) and reducing exhaust pollution [9]. For a given thrust, the thermal efficiency and reduction in size of a jet engine can be improved by simultaneously increasing turbine entry temperature and avoiding the need to air-cool the turbine blades internally (Fig. 2) [10].

Table 1. Temperature capability of high-temperature material systems

Material	Approximate maximum use temperature, °C	Ratio of maximum use temperature (°C) to melting temperature (°C)	Application
Aluminum (RR 350)	300	0.45	stressed
Titanium (IMI 685)	550	0.33	stressed
Low-alloy steel	580	0.37	stressed
Austentitic steel (A 286)	600	0.4	stressed
Nickel superalloy (MarM 002)	1000	0.77	stressed
Corrosion-resistant austenitic steel	1150	0.75	unstressed
Nickel (Brightray H)	1250	0.9	unstressed
Ferritic corrosion-resistant steel (FeCrAl)	1350	0.9	unstressed
Alumina	1850	0.9	unstressed
Tungsten	2500	0.74	unstressed

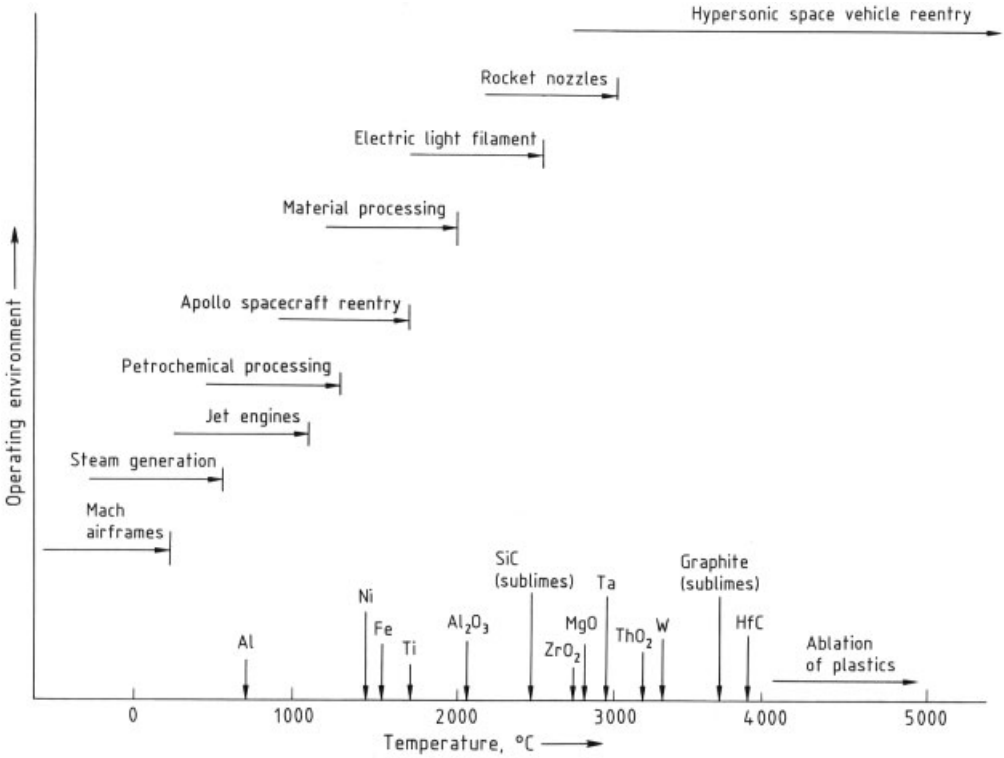

Figure 1. Typical operating temperatures in high-temperature environments

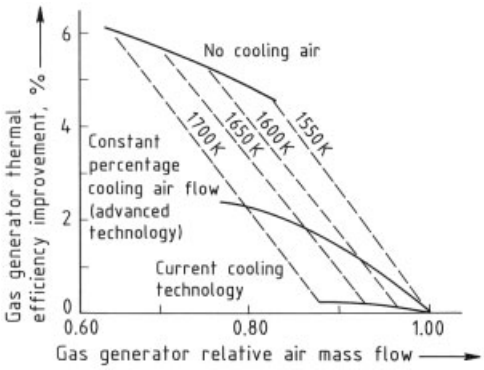

Figure 2. Effect of turbine entry temperature and component cooling on thermal efficiency of a gas turbine
Temperatures are the takeoff turbine entry temperatures.

High-temperature materials also conserve energy by minimizing heat loss. Compared with conventional insulating firebrick furnace linings, for example, lightweight fibrous insulating materials provide more effective insulation, and their low thermal mass allows shorter cycle times and improved productivity. Such materials are widely used in heat treatment furnaces and ceramic firing kilns. Compared with earlier linings, they provide fuel savings of 25 – 50 % and halve process cycle times. A similar type of material, combined with an external radiating coating, is used to resist the high temperature developed by the space shuttle orbiter vehicle on reentry.

Operating Environment. High-temperature materials operate in environments with a wide spectrum of mechanical and chemical conditions. The temperature may be relatively constant (e.g., in insulating a continuously operated heat treatment furnace) or subject to severe cycling (e.g., in jet engines where the thin edges of turbine blades can experience temperature changes $> 1000\,°C$ in seconds). The operating stress can vary from high combined centrifugal and thermal stress in a gas turbine disk to very low stress levels in furnace heating elements. Erosion is another important mechanical environmental factor (e.g., in fluidized-bed coal combustion systems).

The fundamental chemical reaction occurring at high temperature is *oxidation.* Thus, some measure of inherent oxidation resistance is an essential characteristic of high-temperature materials; however, many applications involve significantly more complex corrosive environments than simple oxidation. In certain types of coal gasification, the atmosphere may be reducing to some extent: thus, alloys that rely on chromium oxide scales for protection do not perform as well as aluminum oxide-forming alloys, which remain stable even in severely reducing environments. Oxidation is further modified by atmospheric and fuel contaminants. An atmosphere with low oxygen activity that contains significant amounts of carbon-containing gases can cause carbon absorption, with consequent reduction in corrosion resistance of materials that depend on chromium oxide scales for protection. Atmospheric contamination with sulfates and chlorides (such as which may be experienced by gas turbine engines operating in coastal areas) can result in an accelerated form of corrosion known as *hot corrosion*. Another problem arises from the desire to use lower grade fuel (such as residual fuel oil) instead of higher fractions (such as kerosene) in industrial turbines and boilers. The lower grade fuels contain larger amounts of impurities such as vanadium, sulfur, and sodium, which accelerate corrosion.

The other major operating requirement, material lifetime, depends on the type of machinery, component, application, and environment. It can range from minutes for a rocket motor, through 50 000 – 100 000 h for industrial steam and gas turbines, to more than 20 years for large steam turbine installations (100 – 1000 MW). For many component applications, no single material system is able to meet all operational requirements; coatings are then required to provide corrosion resistance, insulation, and wear resistance.

Effect of Stress. Operational stress has a major influence on temperature capability, with the limit being $0.7 - 0.9\,T_m$ in unstressed applications and $0.3 - 0.7\,T_m$ in stressed applications (Table 1). Nickel superalloys are particularly useful with a stressed capability in excess of $0.7\,T_m$. The broad stress – temperature operational regimes of various materials in stressed applications are illustrated in Figure 3. The strength is based on the 0.2 % proof stress (PS) for metals and an appropriate proportion of the flexure strength for ceramics and composites. The position of carbon – carbon composite material depends on the volume fraction and number of directions of reinforcement. Specific strength (strength divided by density) is the prime consideration in rotating machinery (Fig. 4), and the superiority of titanium relative to steel and nickel has provided the major motivation for its development in aerospace applications. The potential benefits of silicon nitride, silicon carbide, and carbon – carbon composites over nickel superalloys are considerable if the current major problems–lack of defect tolerance and oxidation resistance–can be resolved. The relative capabilities of materials used in unstressed applications such as furnace heating elements and low-density fibrous insulation are shown in Figure 5.

Applications at temperatures > 1350 °C can be satisfied only by refractory metals and nonmetallic materials (Figs. 3, 4, 5). The high den-

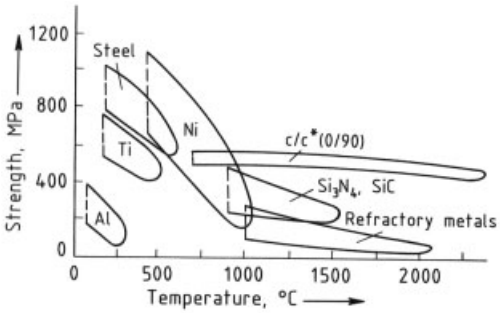

Figure 3. Strength – temperature regimes of high-temperature materials
∗ c/c: Carbon – carbon composite, 0/90: fibers run in two orthogonal directions

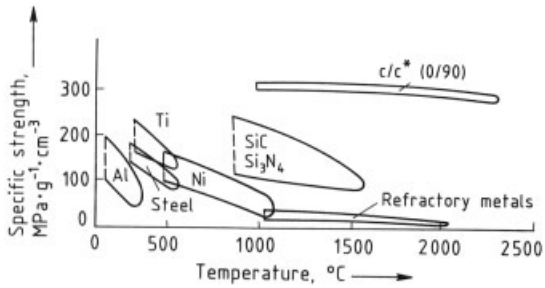

Figure 4. Specific strength – temperature capability
∗ c/c: Carbon – carbon composites

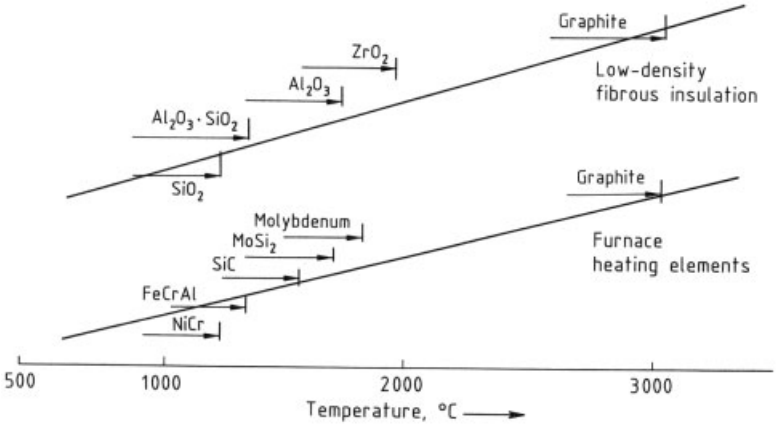

Figure 5. Material temperature capabilities in unstressed applications

sity and fundamental oxidation problem of refractory metals seriously limit their wider application and further justify the recent increased interest in the use of high-temperature ceramic materials.

2. Factors Influencing Temperature Capability

2.1. Melting Temperature

The *melting temperature* of a material is a major indicator in its usefulness at high temperature.

The theory of atomic structure is well documented [11]. *Ionic bonds* can be strong, particularly with multivalent ions as in MgO. *Covalently bonded* diamond is the hardest material in nature and melts at ca. 3700 °C. *Metallic bonds* can also give rise to high melting points (e.g., 3410 °C for tungsten). Mixed bond types are common; refractory metal carbides and nitrides possess a combination of covalent and metallic bonds.

All bonds assume the most stable electron configuration attainable when the two bonding atoms are brought together. The interatomic bonding forces are influenced by many factors including crystal structure, number of close neighbors (coordination number), bond type, and number of electrons involved per atom. The stronger the attraction or bonding force between two atoms, the smaller is the interatomic distance and the higher the melting point. The variation of melting point with interatomic distance is depicted in Figure 6. Refractory metals, including tungsten, tantalum, molybdenum, and niobium, display the highest metallic melting points. The melting points of the carbides and oxides do not follow the same order as that of the metals themselves because of the modified bonding involved in compound formation (Table 2).

Table 2. Melting points of high-temperature, nonmetallic materials

Material	*mp*, °C
HfC	3890
TaC	3880
ZrC	3530
NbC	3500
WC	2870
SiC	2400 *
ThO_2	3300
HfO_2	2900
MgO	2800
ZrO_2	2680
Al_2O_3	2050
Si_3N_4	1850 *

* Sublimes.

2.2. Volatility

Some materials with a high melting point have a significant vapor pressure (volatility) at elevated temperature, which can limit their high-temperature capability. For example, MgO melts at 2800 °C but exhibits appreciable volatility at 1600 °C; tungsten melts at 3410 °C but suffers ca. 1 % loss in weight, due to evaporation, in

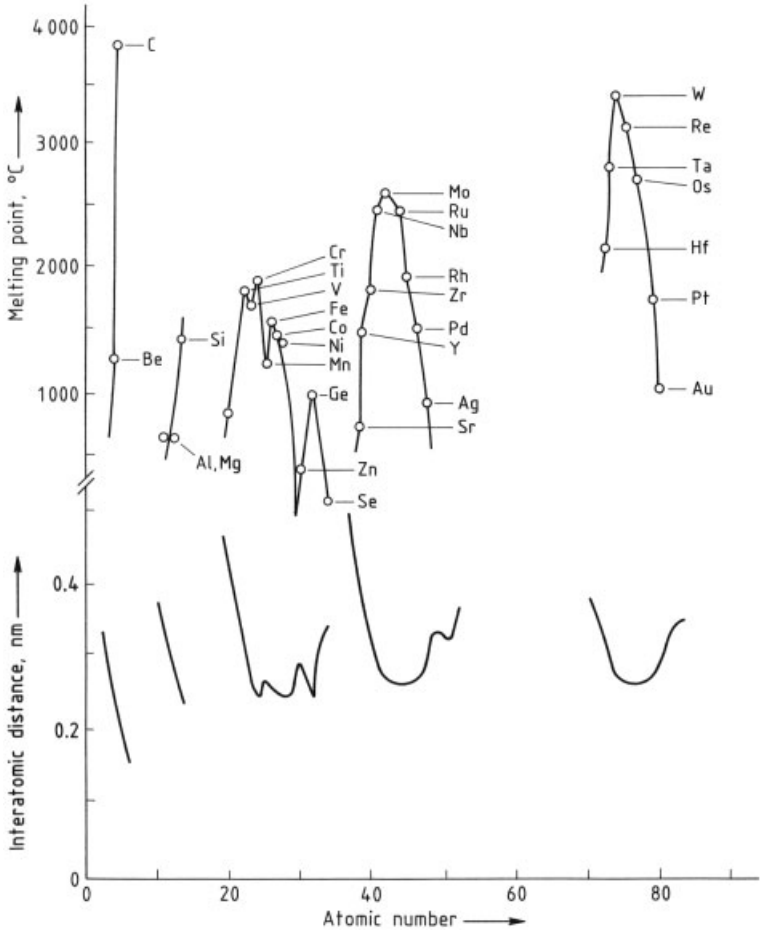

Figure 6. Interatomic distance and melting points of elements as a function of atomic number

100 h at 2560 °C; chromium melts at 1890 °C but experiences 1 % loss in weight at 895 °C [12]. For most materials other than oxides in an oxidizing environment, oxidation or other forms of corrosion impose a lower temperature limit than does volatility.

2.3. Crystal Structure

The majority of useful engineering materials have a cubic crystal system. Examples include most of the common metals and some of the simpler ceramics such as MgO and TiC. Most ceramic materials have a relatively complex crystal structure. The type of crystal structure has a major influence on plastic deformation and diffusion characteristics, which in turn influence zero-time and real-time deformation behavior and, consequently, temperature capability. Zero-time deformation deals with instantaneous behavior when a load is applied; real-time deformation is concerned with time-dependent processes such as creep and fatigue.

2.4. Zero-Time Deformation

The major mechanism involved in zero-time deformation of metals is slip, the relative movement of one crystal plane over another by a process involving the movement of crystal imperfections known as dislocations [13]. Slip occurs more readily along some crystal directions and planes than along others, and the number of slip

systems depends on the crystal structure. Slip is beneficial because it allows metals to be forged into shape and allows stress redistribution to occur in regions of local stress concentration during service. In metallic systems, however, mechanisms that make slip more difficult must be found in order to increase strength and temperature capability.

2.5. Creep

Creep is essentially a slow, continuous real-time deformation at high temperature under stress. It normally occurs at a temperature that is high relative to the melting point of the material, generally at 0.3 – 0.4 T_m for metals and at 0.4 – 0.5 T_m for ceramics, where T_m is the absolute melting point. Mechanisms that strengthen zero-time deformation can also increase the temperature at which creep becomes a limiting factor so that some aluminum alloys can be used under stresses up to 0.5 T_m and nickel superalloys up to 0.7 T_m (see Sections 3.1.1, 3.1.2, and 3.1.4). The precise relationship between creep strain and time depends on temperature, stress, and other factors but generally takes the form shown in Figure 7.

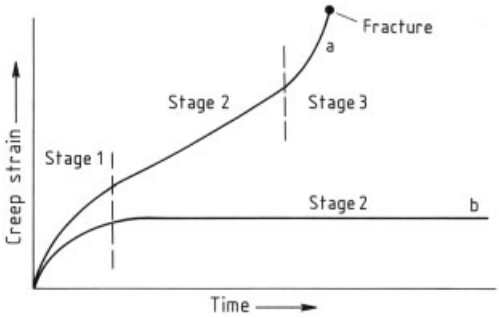

Figure 7. Generalized effects of stress and temperature on creep behavior
a) High stress or high temperature; b) Low stress or low temperature

The creep rate first decreases for a relatively short period of time (stage 1). Steady-state creep then occurs over an extended time period (stage 2) until an acceleration of creep rate (stage 3) is followed by fracture. At low temperature (curve b), the creep strain is restricted because microstructural features (e.g., grain boundaries, precipitates) impede dislocation movement. At higher temperature (curve a), the dislocations can climb out of their blocked slip plane and continue the creep process. The dislocation climb takes place by diffusion, and for many crystalline solids the activation energies for creep and self-diffusion are similar (Fig. 8) [14].

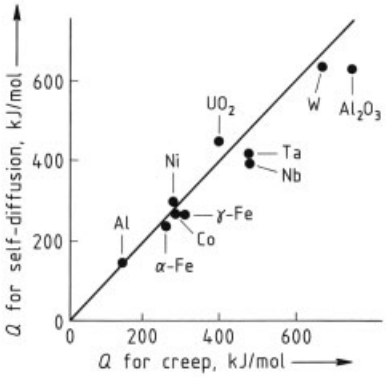

Figure 8. Relationship between activation energies Q for creep and self-diffusion

The influence of diffusivity is illustrated by creep tests on body-centered cubic α iron just below 910 °C and face-centered cubic γ iron just above 910 °C. The α iron creeps at ca. 200 times the rate of the γ iron, reflecting their relative rates of self-diffusion [15]. The mechanisms that increase creep strength are discussed in Section 3.1, and typical creep rupture data for various metallic systems are shown in Figure 9.

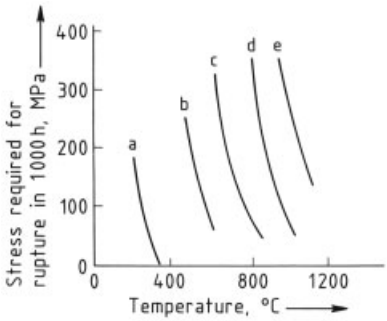

Figure 9. Creep strength relationships for metal systems
a) Al (RR 58); b) Ti (Ti – 5 Al – 2.5 Sn); c) Fe (A 286); d) Ni (Inco 713 C); e) Mo (Mo – 0.5 Ti)

Creep is experienced by components that operate at high temperature under load. Turbine

rotor blades experience temperature and stress conditions in the aerofoil such that creep is likely to determine blade life. Another aspect of creep is evident in components such as bolts in turbine casings. Here creep produces a stress relaxation with time and the bolts must be retightened at intervals.

2.6. Mechanical Fatigue

Mechanical fatigue is a common failure mechanism that can occur when a component or structure is subjected to repeated stress or strain cycles, with failure occurring at a stress well below the tensile strength of the material. Fatigue is often classified as high-cycle or low-cycle fatigue. In *high-cycle fatigue*, the operating stresses are below the general yield stress of the material, and failure, often caused by local stress concentration, occurs after $>10^4$ cycles. In *low-cycle fatigue,* the stress levels are above the general yield, and fracture generally occurs in $<10^4$ cycles [16]. Operation often involves an extended period under load in the normal fatigue cycle (dwell on load) as, for example, in an airframe structure that is pressurized and depressurized during each flight. Temperature reduces material strength and, consequently, fatigue strength, but an additional effect can occur in low-cycle fatigue. Dwell on load at temperature can further reduce life by a creep – fatigue interaction that may result in accelerated growth of fatigue cracks.

2.7. Thermal Fatigue

In thermal fatigue, stresses are generated by temperature cycling. In the extreme case referred to as *thermal shock*, a ceramic may be subjected to a temperature change of sufficient magnitude for the resultant thermally induced stress to cause failure in one cycle. Although ductile metals are more resistant to thermal fatigue than other materials, failure can occur solely from thermally generated stress. More commonly, components operating in high-temperature machinery, such as the rotor blades in gas turbines, experience a combination of thermally induced and direct stresses. Such creep fatigue interaction is referred to as thermo-mechanical fatigue. Component life under such conditions can be predicted from testing in which direct stress and temperature cycling appropriate to component operation is applied to test samples[17]. Thermo-mechanical fatigue life can be maximized by selecting materials with the most appropriate property combination, including physical properties that minimize thermally generated stress. This last requirement is satisfied, for example, (1) by the glass ceramic lithium aluminum silicate, which has a low coefficient of thermal expansion, and (2) by directionally solidified nickel superalloy turbine blades, which have a Young's modulus lower than that of conventionally cast blades.

2.8. Oxidation

The tendency of a metal to react with oxygen either directly or indirectly is indicated by the *thermodynamic free energy change* ΔG accompanying oxide formation. Free energy changes for metal oxides are shown in Figure 10. The more reactive metals, and thus the more stable oxides, are found toward the bottom of the diagram. The rate of oxide formation is determined by *reaction kinetics*, which in turn depend largely on the nature and properties of the oxide. After initial chemisorption of oxygen atoms, the nucleated oxide grows until an oxide film composed of multiple grains covers the surface. The oxide film acts as a barrier to further oxidation. Some oxides, such as MoO_3 and WO_3, are volatile and thus offer no protection against oxidation (Fig. 11, curve a). If the oxide film is porous, oxygen can pass through it and react at the metal – oxide interface so that oxidation continues (Fig. 11, curve b). Usually, however, the oxide is not porous and further oxidation can occur only because metal ions diffuse through the oxide to the oxide surface or because oxygen ions diffuse through to the metal. The rate of diffusion and consequent rate of oxidation depend on temperature and defects in the oxide lattice. The extent of oxidation normally decreases with time (Fig. 11, curves c); good protection is provided by Al_2O_3 on aluminum and Cr_2O_3 on stainless steel. Initially protective oxides may become ineffective due to factors such as cracking or spallation (Fig. 11, curve d) which are caused by an inability to accommodate the stress from the growing volume of oxide relative

to the volume of metal from which the oxide is formed.

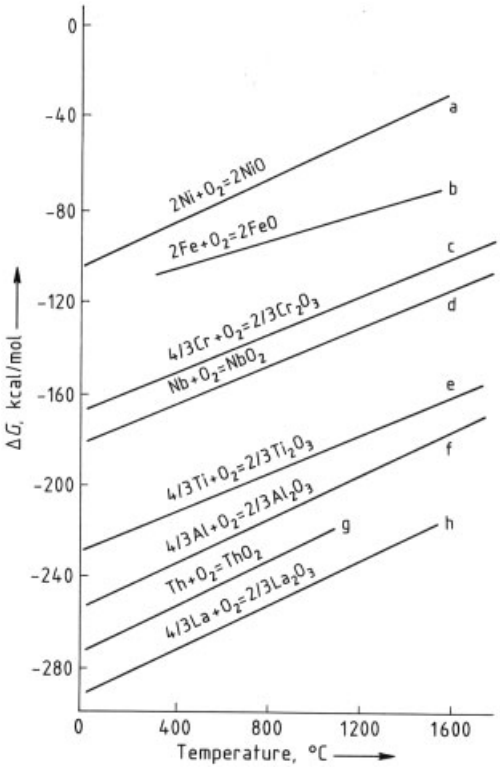

Figure 10. Free energy change ΔG for oxide formation

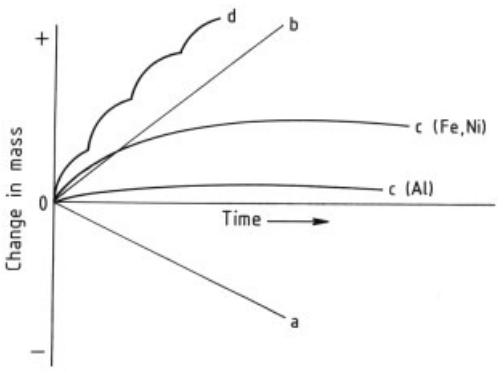

Figure 11. Schematic representation of oxidation behavior

Nonoxide ceramics are attacked in oxidizing atmospheres but the oxide formed may provide some protection. Silicon carbide and molybdenum disilicide are protected by the SiO_2 surface layer and are used extensively for heating elements. Secondary phases arising from the metallic oxides added to silicon nitride to aid densification during sintering can reduce the protection afforded by the SiO_2 because of the formation of complex surface glasses. Most oxide ceramics are unaffected by oxidizing atmospheres; however, the use of silicon-based ceramics under high-temperature reducing conditions is limited by the formation of volatile SiO.

In practice, the solid – gas reaction is more complex than the simple metal – oxygen reaction and involves gaseous mixtures (e.g., combustion products from jet engines or rocket nozzles). Elements such as sulfur, sodium, vanadium, and carbon that are present in the atmosphere or the fuel can participate in corrosion.

2.9. Hot Corrosion

Hot corrosion denotes the accelerated oxidation of nickel or cobalt superalloys that occurs when the alloys are exposed to high temperature and the surface becomes coated with foreign deposits such as Na_2SO_4, V_2O_5, or NaCl. These deposits may come from the environment, or they may be present as contaminants in the fuel. Gas turbines operating close to the sea ingest sea salt, which reacts with sulfur in the fuel to form Na_2SO_4. Sulfides can take part in the hot corrosion process, whose rate reaches a maximum at 800 – 900 °C, depending on alloy composition; this process can be much more severe than basic oxidation at the same temperature (Fig. 12). Formation of chromium sulfides is common with nickel superalloys and results in localized chromium depletion with a detrimental influence on oxidation resistance. The sulfides are then oxidized, and the liberated sulfur penetrates further into the alloy to produce sulfides, which are reoxidized. The process continues with a progressive increase in corrosion rate because the chromium is unable to reform a protective surface film.

A type of corrosion known as *sulfidation* can occur in the low-oxygen atmospheres found in gasification plants operating on coal or oil. Such atmospheres contain appreciable amounts of sulfur, and conditions can be such that protective oxide scales are not formed. Sulfide scales are not protective because they are mechanically less stable than oxides; they allow diffusion rates

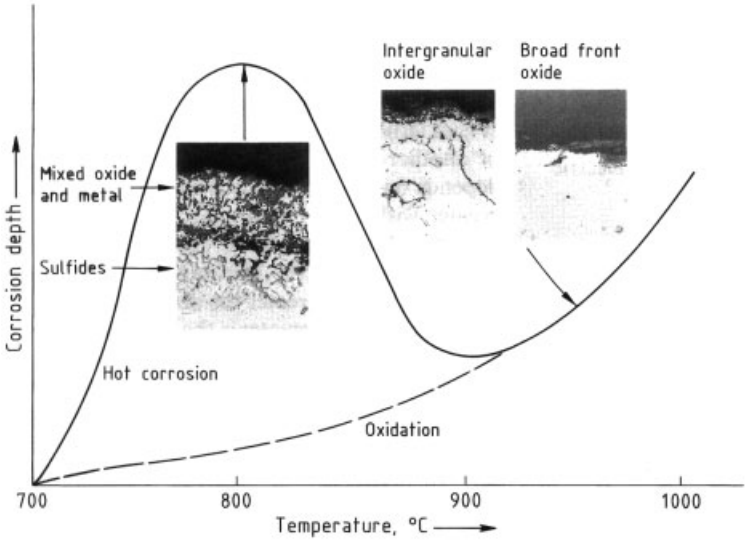

Figure 12. Oxidation and hot corrosion of nickel superalloys

several orders of magnitude faster than oxidation, probably because of the relatively low melting points of sulfides.

Accelerated oxidation can occur from a flux in the Fe_2O_3 film formed on the superheater tubes of coal-fired power plant boilers; sulfides are not involved in the corrosion process. Ash from the combustion is deposited on the superheater tubes. A protective oxide forms under the ash, but sulfates gradually accumulate at the base of the porous ash deposit. The mixed sulfate remains molten at tube surface temperatures $>600\,°C$ and forms a thin layer of extremely corrosive liquid adjacent to the protective oxide scale. This layer contains free SO_3, which reacts with the oxide to form more sulfate, ensuring continued corrosion [18].

2.10. Carburization

Carburization may occur in nickel alloy furnace equipment in low-oxygen atmospheres containing significant concentrations of carbon monoxide, methane, and other hydrocarbons (e.g., in the carburizing process used to case-harden steel). Carbon is absorbed onto the surface from the gaseous phase and diffuses into the base metal to produce carbides (usually of chromium). These carbides reduce ductility, thereby increasing the possibility of cracking under thermal cycling conditions, as well as the possibility of oxidation or sulfidation by reducing the level of available chromium in the base material.

2.11. Erosion

Erosion is the mechanical removal of material by a moving fluid, usually containing solid particles. An example is the severe damage to compressor blades experienced by turbines in helicopters operating in dusty terrain. Erosion at high temperature can be more or less severe than at room temperature, depending on the relative erosion resistance of the oxide scales and base material. Particulate erosion and chemical corrosion often occur simultaneously (e.g., in fluidized-bed coal combustion systems for electric power generation).

2.12. Wear

Wear takes place during the relative motion between two surfaces and has been a consistent cause of component damage in many applications. *Abrasive wear* is caused by hard particles such as metallic debris or oxides. *Adhesive wear* (or galling) is caused by shearing of the microwelds formed at mating surface asperities; wear

debris is removed from the weaker of the two mating surfaces. The type and severity of wear are determined by applied load, relative movement of contacting surfaces, impact effects, temperature, etc. In discussing the effects of temperature, a distinction must be made between the temperature of the environment and the localized high temperatures of short duration that are generated by friction. Wear tends to be less severe at high temperature because of the beneficial effects of the oxides formed, provided they adhere to the surface. Poor oxide adherence can give rise to *oxidative wear.*

3. Mechanisms for Increasing Temperature Capability

3.1. Metallic Materials

Several mechanisms have been used to strengthen metallic materials by impeding the movement of dislocations, thus reducing deformation under load. Solid-solution strengthening is applicable in different degrees to all base metals. Strengthening by precipitation of second phases is effective but not possible in all alloys. Incorporation of dispersions of stable particles can increase strength considerably at very high temperature. Grain boundaries play a significant role in determining strength; a fine grain size is generally beneficial in the lower temperature creep region, and a coarser grain size tends to improve strength in the higher temperature creep region.

3.1.1. Solid-Solution Strengthening

A solid solution forms when atoms in the crystal lattice of one metal are replaced by atoms of another metal. The extent of solid solution depends on the atoms involved and is normally greater with atoms of similar size and electron structure. If the diameters of the two atoms differ by more than ca. 15 %, the amount of solid solution is small. The greater the discrepancy in electron structure, the greater is the tendency to form compounds rather than extensive solid solutions. *Strengthening* is due to internal strains generated by the presence of solute atoms. Atoms with similar lattice parameters may have high mutual solubility but provide relatively little strengthening. Atoms with different lattice parameters may have limited solubility, but low concentrations could produce significant strengthening. For similar lattice strains, the larger the valence difference between solute and solvent, the greater the strengthening.

Very few elements have sufficient solubility in aluminum to develop effective solid-solution hardening, and the effect is small compared with precipitation hardening (see Section 3.1.2). Nickel has a much greater scope for solid solution strengthening. Aluminum, tungsten, and molybdenum are effective solid-solution strengtheners. Strengthening persists to high temperature, but in the creep regime above $0.6\,T_{\mathrm{m}}$, diffusion effects become more important; molybdenum and tungsten are effective in lowering diffusion rates.

3.1.2. Precipitation Hardening

Precipitation of a second phase occurs in many alloy systems as a result of a decrease in the solid solubility of alloying elements with decreasing temperature. In the reverse process, the phase dissolves as the temperature rises above the solvus temperature for the precipitate. If the material is then cooled rapidly (e.g., quenched), precipitation is suppressed and the alloy exists at low temperature as a supersaturated solid solution. If material in this condition is allowed to age at ambient or higher temperature (depending on the alloy system), the second phase precipitates. This can provide a major strengthening mechanism in high-temperature materials such as aluminum alloys, nickel superalloys, and steels. The mechanism is illustrated in Figure 13.

The strengthening ability of the precipitate depends on its compatibility with the matrix (e.g., crystal structure and lattice parameter) and also on the volume fraction of the precipitate and its size. Relatively small differences in lattice parameters allow coherence between precipitate and matrix of similar crystal structure; i.e., the two structures match at the interface, with the degree of distortion depending on the amount of mismatch in lattice parameters. Precipitates that do not show coherence with the matrix provide less strengthening. Nickel super-

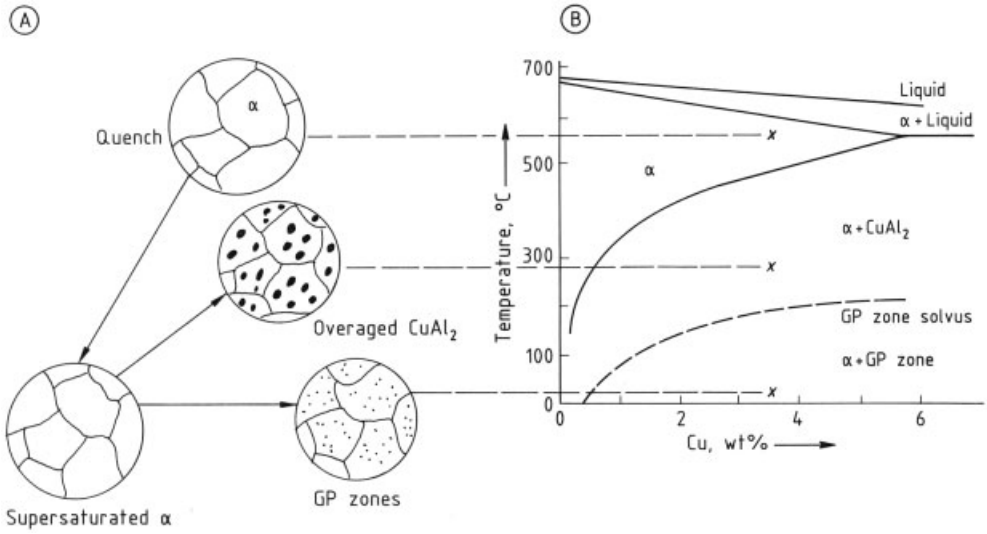

Figure 13. Schematic representation of precipitation hardening
A) Typical microstructures: the black dots represent $CuAl_2$ precipitate; B) Phase diagram

alloys can be strengthened by precipitates based on Ni_3Al, usually referred to as γ'. The γ nickel matrix and the γ' Ni_3Al precipitate have similar crystal structures, and if the mismatch in lattice parameters does not exceed ca. 1 %, the precipitate is an effective strengthener. Besides providing different degrees of solid-solution strengthening, alloying additions can modify coherency strains by modifying the lattice parameter of the phase in which they are partitioned. The effect of $\gamma - \gamma'$ mismatch on hardness for different alloys is shown in Figure 14 [19]. Precipitates with higher mismatches tend to be less stable and thus less effective in increasing temperature capability than lower mismatch precipitates.

The precipitates strengthen by impeding the progress of dislocations through the material. When they reach the precipitates, dislocations either must bypass them by climbing or looping or must cut through them. Both mechanisms operate and there is an optimum precip-

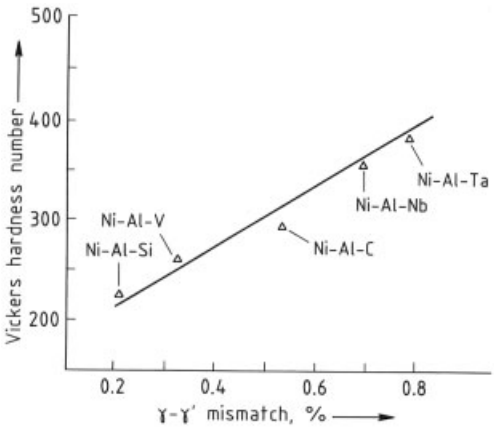

Figure 14. Effect of $\gamma - \gamma'$ mismatch on hardness of nickel – aluminum ternary alloys after aging at 870 °C for 50 h

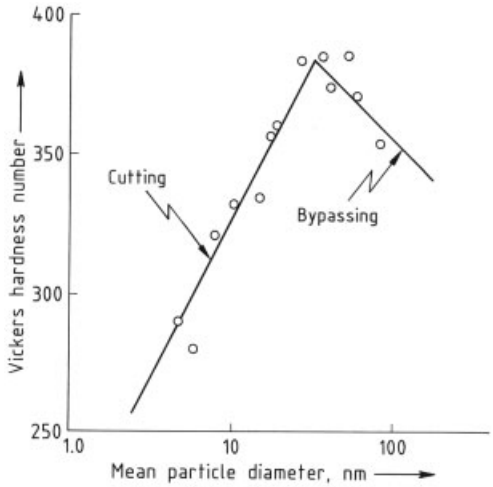

Figure 15. Relationship between γ' particle diameter and hardness of Ni – 22 Cr – 2.8 Ti – 3.1 Al alloy after age hardening at 750 °C

itate size (Fig. 15) [20]. Strength also increases with increasing volume fraction of precipitate (Fig. 16) [21]. Some alloying elements influence the solubility of the precipitating phases. For example, the addition of cobalt increases the solvus temperature of γ' and thereby increases the creep strength of Nimonic 90 over the cobalt-free Nimonic 80 (Fig. 17).

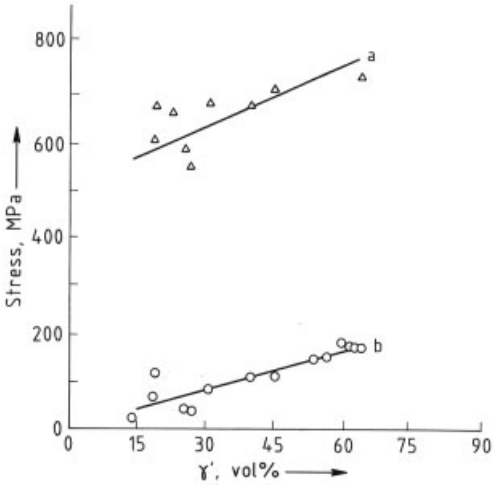

Figure 16. Effect of γ' volume fraction on 100-h creep strength of a nickel superalloy
a) 705 °C; b) 980 °C

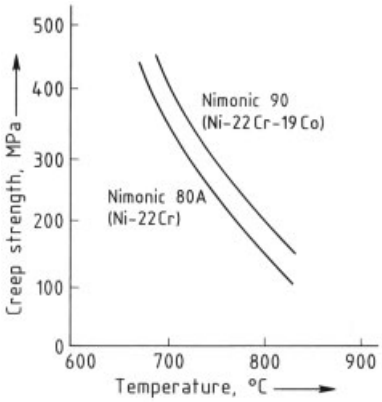

Figure 17. Creep strength of Nimonic 80 A and Nimonic 90: rupture in 300 h

In aluminum alloys, changes in the early stages of the precipitation process produce peak hardening (→Aluminum Alloys, Chap. 2.3.3.). Large numbers of ordered solute-rich clusters of atoms, possibly only one or two atom layers thick, are formed in zones known as *Guinier–Preston* (*GP*) *zones*. These zones retain the structure and coherence of the matrix and usually produce appreciable elastic strain. An intermediate precipitate (e.g., tetragonal $CuAl_2$ in Al–Cu, hexagonal Mg_2Si in Al–Mg–Si) is subsequently formed, which is much larger than a GP zone and has only partial coherence with the matrix so that strengthening is somewhat lower [22]. Formation of the final precipitate (body-centered tetragonal $CuAl_2$ in Al–Cu, face-centered cubic Mg_2Si in Al–Mg–Si) involves complete loss of coherence with consequent reduction in strength. The precipitation reactions can be influenced greatly by the presence of small amounts of certain elements that can modify phase behavior or promote precipitate formation [23].

The solvus temperature of a GP zone broadly defines the upper temperature limit of stability. In some alloys, GP zones act as nuclei for the subsequent stage of the aging process, usually the intermediate precipitate [23,24]; this is exploited in two-stage (duplex) aging. In other alloys, the intermediate precipitate nucleates heterogeneously at lattice defects such as dislocations.

Precipitation reactions occur in low-alloy and 12 % chromium martensitic steels during the tempering of martensite. In low-alloy steels, Mo_2C, V_4C_3, and TiC lead to precipitation hardening. As with aluminum alloys, the sequence of reactions in the precipitation hardening of steels (also known as secondary hardening) is usually the formation of zones followed by intermediate precipitates and equilibrium precipitates [25]. The intermediate precipitates generally form at 450–600 °C with a coherent matrix interface. The equilibrium carbide phases do not usually form below ca. 650 °C and have incoherent interfaces with the matrix [26].

In 12 % chromium steels, secondary hardening is due to the precipitation of a chromium-rich carbide of the M_2X type. Alloying elements such as molybdenum and vanadium can stabilize this carbide, and increase the temperature range over which it forms, the associated coherency strain, and thus secondary hardening [27]. The effect of molybdenum and vanadium on the temper resistance of a 0.1 C–12 Cr–2 Ni steel is shown in Figure 18.

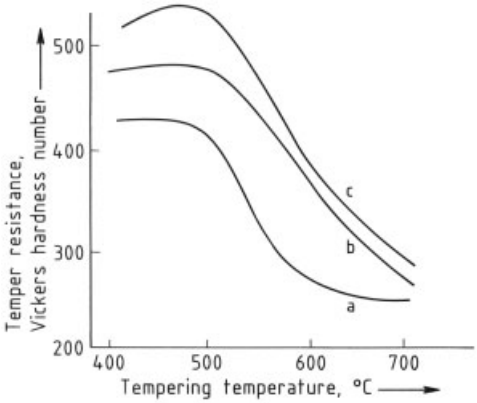

Figure 18. Influence of alloying elements on the temper resistance of 12 % chromium steels
a) Base steel 0.1 C – 12 Cr – 2 Ni; b) Base steel containing 1 wt % vanadium; c) Base steel containing 3 wt % molybdenum

3.1.3. Dispersion Strengthening

In precipitation-hardening systems, the precipitated phases redissolve above their solvus temperature, thus placing a limit on their temperature capability. The incorporation of phases that are insoluble in the base alloy, such as Al_2O_3 in aluminum and ThO_2 in nickel, allows strengthening at a temperature above that where precipitation hardening is effective. Strengthening is somewhat less than in precipitation-hardened material because the dispersed phase is not coherent with the matrix. Strengthening can, however, still be significant provided the dispersed phase has a fine, uniform distribution with the necessary interparticle spacing. Thus dispersion-strengthened alloys have a relatively flat strength – temperature curve compared to precipitation-hardened alloys (Fig. 19). The dispersed phases impede dislocation movement, and strengthening is increased by controlled mechanical working to produce the optimum microstructure that is stabilized by the dispersion.

The ability of dispersed phases to stabilize grain structure was applied at one time to tungsten lamp filaments which operate at ca. 2500 °C (0.8 T_m) and must be substantially free of distortion. Incorporation of ca. 1 % ThO_2 inhibited the recrystallization and grain growth that led to failure in pure tungsten filaments. The thorium dioxide dispersion has since been replaced by an alternative technology (Chap. 8).

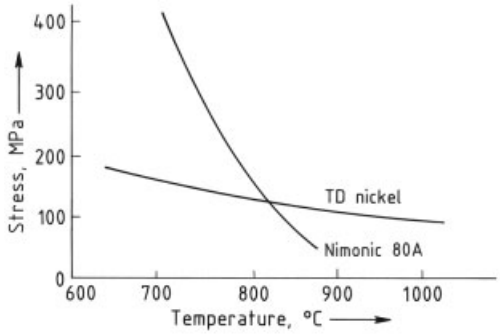

Figure 19. Comparison of 100-h creep rupture of precipitation-strengthened (Nimonic 80 A) and dispersion-strengthened (TD nickel) nickel superalloys

3.1.4. Grain Boundary Effects

A *grain boundary* is a region of mismatch between the atomic lattices of adjacent grains, in which the mismatch produces less efficient local atomic packing. This boundary impedes the progress of dislocation movement and thus provides a strengthening mechanism at temperatures up to ca. 0.5 T_m. Below this limit, the influence of grain size on yield strength is given by the *Hall – Petch relation*:

$$\sigma_y = \sigma_i + K_y d^{1/2}$$

where σ_y is the yield strength, σ_i is a lattice friction term, K_y is the dislocation locking term and $2d$ the grain diameter. Controlled thermomechanical processing has been used to reduce grain size and to improve tensile and fatigue strength in nickel superalloy disks for gas turbine engines. Data for Rene 95 are given in Table 3 [28].

Table 3. Effect of grain size on tensile strength of Rene 95

Heat treatment, °C	Grain size, µm	Tensile strength	
		0.2 % YS [b], MPa	UTS [c], MPa
1120 AC [a]	5	1130	1500
760 AC	5	1130	1500
1220 AC	150	920	1250
1120 AC	150	920	1250
760 AC	150	920	1250

[a] AC = air-cooled.
[b] YS = yield strength at 650 °C.
[c] UTS = ultimate tensile strength at 650 °C.

Above ca. 0.5 T_m, diffusion becomes significant and the grain boundaries can act as sinks for dislocations; the rate of diffusion in the grain boundary region can be several orders of magnitude higher than in the grains themselves. Thus, at high temperatures, grain boundaries have a weakening effect, and an increase in grain size is normally accompanied by an increase in creep life. The effect of grain size on the creep life of a wrought nickel superalloy is shown in Figure 20 [29]. Test piece dimensions relative to the number of grains in the section under load are also important. The extent of grain boundary weakening can be reduced by microstructure control. For example, some grain boundary carbides in nickel superalloys inhibit grain boundary sliding and thus increase creep strength. Very low carbon content can give rise to inadequate carbide levels, which result in low tensile strength and ductility with intergranular failure [30]. Excessive levels of carbon can result in continuous grain boundary carbides and consequent brittleness. Most nickel superalloys contain 0.05–0.15 % carbon. The carbide–matrix interface can be a source of weakness, although a grain boundary film of Ni_3Al (see Fig. 44) can protect this interface. Further grain boundary strengthening can be achieved by the addition of small amounts of boron and zirconium, which segregate to grain boundaries and reduce diffusion rates (Table 4) [31].

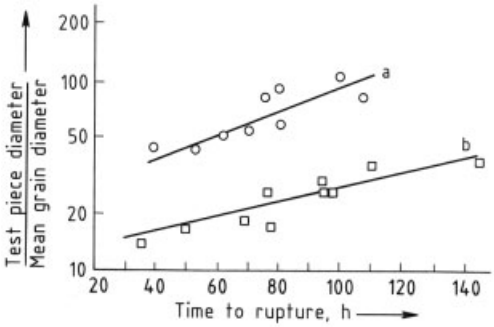

Figure 20. Effect of grain size and test piece diameter on creep rupture life of a wrought nickel superalloy at 150 MPa and 870 °C
Grain size: a) 8 µm; b) 23 µm

One way of avoiding the high-temperature weakening due to grain boundaries is to eliminate them. Directional solidification produces castings with aligned columnar grains. If the boundaries between the grains are in the direction of the principal applied stress, their effect is minimized and a significant increase in creep life and ductility is achieved [32]. Elimination of grain boundaries by the production of single-crystal components further increases life (Fig. 21). The lower secondary creep rate shown in Figure 21 (curve c) is essentially due to the more effective use of the γ' precipitate in single-crystal materials (see Section 7.3.1).

Table 4. Effects of boron and zirconium on creep properties of Udimet 500

Alloy	Creep life at 175 MPa and 870 °C, h	Rupture elongation, %
Udimet 500 base alloy	50	2
Base alloy + 1900 ppm Zr	140	6
Base alloy + 90 ppm B	400	8
Base alloy + 1900 ppm Zr and 90 ppm B	647	14

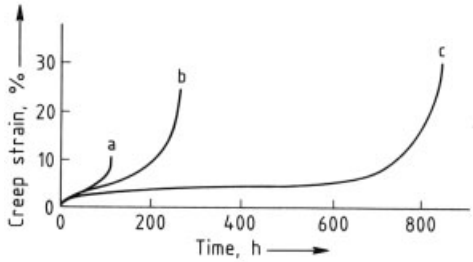

Figure 21. Creep life at 150 MPa and 1000 °C
a) Conventionally cast Mar M 002; b) Directionally solidified Mar M 002; c) Single-crystal SRR 99 nickel superalloys

3.2. Nonmetallic Materials

In contrast to metallic materials, which deform by dislocation movement and are strengthened by making movement more difficult, few high-temperature, nonmetallic materials are able to deform plastically in zero time. Room-temperature tests on polycrystalline MgO produced fracture at <1 % plastic strain, and the behavior was essentially brittle. Dislocation movement led to crack nucleation, growth, and cleavage failure.

Most high-temperature ceramics have complex crystal structures and do not deform plastically. Therefore, they cannot relieve local stress concentrations and are sensitive to the presence of defects such as pores and small cracks. The

problem in such materials is to find ways of developing consistent, useful strength levels. Elimination of all defects is impractical. Even if a defect-free material could be produced, some stress concentration features and defects would certainly develop during use. The practical approach is to develop defect-tolerant ceramic materials for use in applications involving significant stress levels (see Section 9.3.4).

Most manufacturing processes reduce the temperature capability of the ceramic below that theoretically possible for the pure ceramic. For example, the sintering aids added to Al_2O_3 lower strength and softening temperature compared with pure Al_2O_3. Temperature capability can be maximized by selecting the production process and material composition so that the minimum temperature decrease is achieved relative to the pure ceramic.

3.3. Corrosion-Resistant Coatings

(→Corrosion; →Metals, Surface Treatment)

3.3.1. Aluminum

Aluminum oxidizes to form a protective film of Al_2O_3. The oxide that forms on freshly exposed aluminum is thin, and thicker oxide films with better protection capability can be formed by chemical or electrochemical treatment. Film thicknesses of ca. 1 – 2 µm can be formed by chemical treatment and can incorporate substances such as chromates to improve corrosion resistance. Electrolytic anodizing produces films 10 to 20 µm thick, but in the most severe environments (e.g., jet engine compressor blades in salt-containing environments), corrosion resistance may be inadequate and the anodized surface can be treated with a phenolic or an epoxy resin. Cladding with a more corrosion-resistant aluminum alloy is also used to give protection in certain applications (see Section 4.1.1).

3.3.2. Steel

Low-alloy steels have poor high-temperature oxidation resistance because the FeO or higher oxide formed provides inadequate protection. Several types of protection are used for steels [33]. Chromized coatings are applied by a pack cementation process, the pack consisting of chromium or ferrochromium powder, a halide activator, and an inert powder diluent such as Al_2O_3 or Cr_2O_3. Chromium diffuses into the steel surface at ca. 1000 °C, creating a protective layer of Cr_2O_3 that forms on high-temperature oxidation rather than the less protective iron oxides.

Phenolic or epoxy resin coatings physically separate the steel surface from the oxidative or corrosive atmosphere. They can be used up to 200 °C; their temperature capability increases to ca. 500 °C when they are filled with aluminum powder. Such coatings do not provide cathodic protection. This can be developed when aluminum-filled inorganic paints based on silicate or phosphate binders are made electrically conducting by peening or low-temperature heat treatment. Such coatings are necessary to avoid the severe corrosion found on jet engine compressor components that operate in sea salt ingestion conditions. Electroplated coatings based on nickel and cadmium have also been used in such applications.

3.3.3. Nickel

The types of corrosion against which protection is required are discussed in Sections 2.8 and 2.9; corrosion resistance is provided by the stable protective oxides Cr_2O_3 and Al_2O_3. The base nickel alloy may have the necessary inherent corrosion resistance in some applications, but the highest temperature oxidation and most severe hot corrosion environments require protective coatings containing higher amounts of aluminum and chromium. The protective effect of Cr_2O_3 is limited to a maximum temperature of ca. 1000 °C due to the formation of volatile CrO_3; Cr_2O_3, provides the best protection against hot corrosion, whereas Al_2O_3 provides the best protection against oxidation.

Coatings on nickel superalloys are either diffusion or overlay types. *Diffusion coatings* are applied by pack cementation processes in which the part is surrounded by powdered aluminum- or chromium-rich material, a halide activator, and an inert powder diluent [34]. A chemical vapor deposition mechanism transfers aluminum at ca. 700 – 1100 °C and chromium at ca. 1000 °C to the part surface. These diffuse

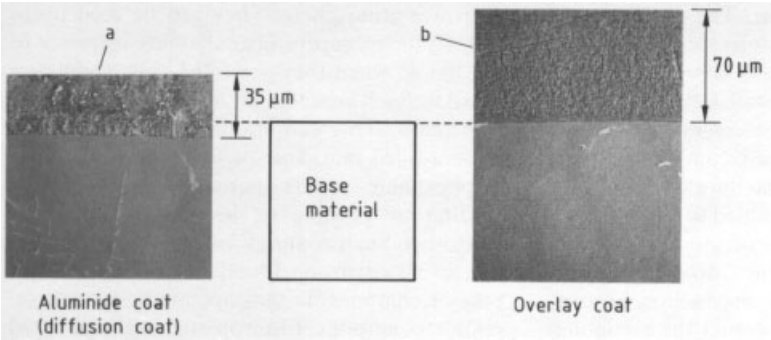

Figure 22. Microstructures of aluminide and overlay coats
a) NiAl; b) CoAl or NiAl dispersed in nickel or cobalt solid solution

from the surface into the base material to form a nickel aluminide or nickel – chromium solid solution, which may contain α-chromium particles. The composition of the coating is influenced significantly by the base material because of diffusion. Coating composition and thickness are also determined by the pack composition, process temperature, and any heat treatment after coating. A typical aluminide coating is 50 µm thick and contains >30 % aluminum. A typical microstructure is shown in Figure 22. The protection provided by aluminide coats is illustrated in Figure 23. Aluminide coats were introduced in the early 1960s and are still used on gas turbine rotor and stator blades and on other nickel superalloy components.

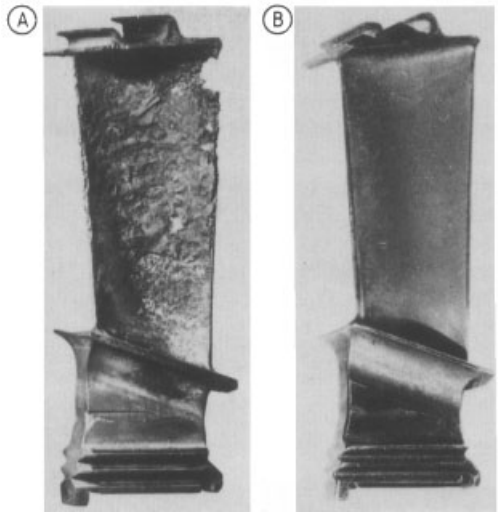

Figure 23. Uncoated (A) and aluminized (B) gas turbine blades after hot corrosion operation
The uncoated turbine blade (A) shows the degree of hot corrosion after 2500 h of service between airfields with low sea approaches and relatively short sector times. The aluminized blade (B) shows virtually no corrosion at a similar lifetime [33].

Improved coatings are, however, required for the highest temperature oxidation and the most severe hot corrosion environments. This need has resulted in the development of improved aluminides and the MCrAlY overlay coats. One of the most important modified aluminide coats is *platinum aluminide* [34]. A thin layer of platinum is applied to the component surface and aluminized. Platinum improves protection by introducing a platinum aluminide phase into the coating microstructure and by stabilizing the nickel aluminide phase.

Overlay coats [35] differ from aluminide coats in that the material is deposited on the component surface, usually by electron-beam physical vapor deposition, or a low-pressure or argon-shrouded plasma spray. Coating composition is independent of base material composition, although significant interdiffusion can occur during subsequent high-temperature service. A wide range of MCrAlY compositions is available in which the base coating element M can be nickel, cobalt, or iron. Chromium (15 – 30 %) and aluminum (6 – 12 %) provide corrosion resistance. Yttrium improves oxide scale adherence, which is particularly important during thermal cycling; similar effects have been found with cerium (Fig. 24) [36]. Several models have been proposed to explain the improved adhesion resulting from the addition of rare-earth elements [37].

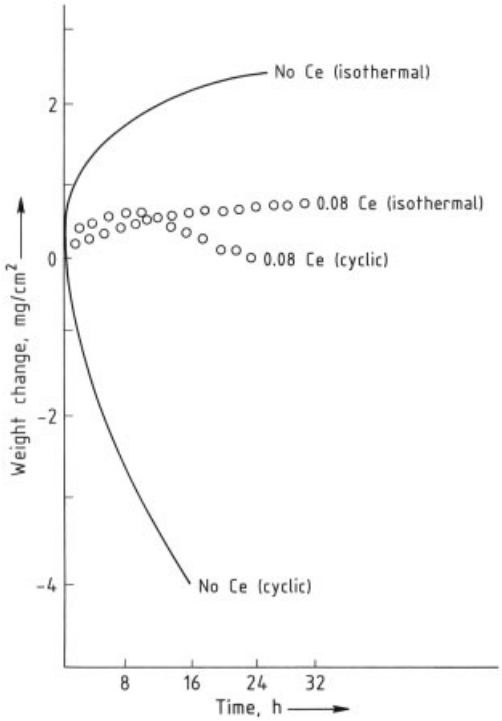

Figure 24. Effect of cerium on the isothermal and cyclic oxidation of Ni – 50 Cr at 1100 °C

Coating thickness is normally 100 – 150 µm, and composition depends on the environment against which protection is required. A CoCrAlY coating with high chromium and low aluminum that forms Cr_2O_3 may be better in hot corrosion conditions, whereas a CoNiCrAlY coating with high aluminum and low chromium that forms Al_2O_3 may be better against oxidation. (A typical MCrAlY microstructure is shown in Figure 22.) Addition of small amounts of silicon, hafnium, and tantalum is reported to improve oxidation resistance. Coatings may be degraded by several mechanisms during high-temperature operation. The oxide-scale-forming elements chromium and aluminium are reduced by progressive oxide formation and spallation and also by diffusion from the coating into the base material, particularly at the highest temperatures. Depending on operating conditions, coatings may also experience mechanical or thermal strains. These strains can result in cracking in the coating, particularly in highly stressed components such as turbine rotor blades in military jet engines. Improved life to cracking was achieved initially by increasing coating ductility. Although the ductility of aluminide coats depends on composition, overlay coating compositions can be selected to provide higher levels of ductility than are possible with aluminide coats (Fig. 25). This was found to be significant in extending life to cracking in severe thermal cycling. More recently, coating life has been predicted from thermo-mechanical fatigue tests.

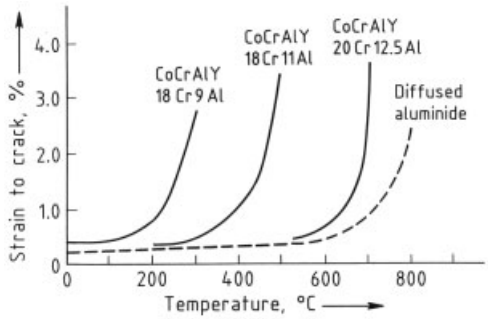

Figure 25. Effect of coating composition on ductile – brittle transition temperature

3.4. Insulating Coatings

Thermal barrier coats (TBCs) based on ZrO_2 have been used on combustion chamber walls of jet engines since the mid-1970s and are being evaluated in diesel engines. In jet engine combustors, TBCs have produced local temperature reductions of 50 °C or more, and combustor life has more than doubled [38]. In diesel engines, components such as pistons, valves, and liners have high surface temperature. More than 55 % of the energy produced by combustion is lost through the cooling system and exhaust gas, reducing efficiency [39]. Use of TBC cuts the heat loss from combustion chambers by 50 – 80 %, depending on coating thickness [40], and thermal efficiency is improved by > 30 %.

A thermal barrier coat must have (1) low thermal conductivity, (2) resistance to the corrosive environment, (3) thermal shock resistance, and (4) a coefficient of thermal expansion compatible with the metal substrate. Zirconium dioxide meets these requirements, with a thermal conductivity that is about one-tenth that of nickel superalloys and cast iron. The ZrO_2 must be stabilized because a phase change occurs at 1170 °C, with the tetragonal form transforming to monoclinic on cooling. This phase change generates

a volume expansion of ca. 4 %, creating large shear strains that can lead to cracking. A MgO-stabilized ZrO_2 was initially used in jet engine combustors, but some destabilization occurs on long-term soaking above 1000 °C. Full stabilization with 20 % Y_2O_3 is possible, but partial stabilization with 8 % Y_2O_3 to a nonequilibrium tetragonal phase with high-temperature capability gives the best performance in jet engine applications [38]. An oxidation-resistant bond coat is included because ZrO_2 does not protect the substrate metal from oxidation. This also accommodates the interface strains arising from thermal expansion differences. Early bond coats were based on the Ni – Cr system, with more recent coats being of the MCrAlY type. The microstructure of a typical TBC is shown in Figure 26.

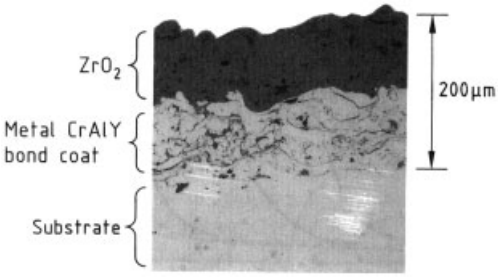

Figure 26. Typical microstructure of plasma-sprayed thermal barrier coat (TBC)

Bond coats are normally applied by plasma spray processes, and spraying under low pressure is necessary for the more oxidation resistant compositions. Plasma spraying in air has been the standard process for applying the zirconia coatings, but physical vapor deposition (PVD) has subsequently been used for applying TBCs in the more severe turbine rotor blade applications [41]. The columnar structure produced by the PVD process minimizes stress development in the ceramic coating.

The prime cause of failure in TBCs is the stress developed due to the differential thermal expansion between ceramic and metal during thermal cycling. Failure occurs at the interface between the ceramic and the bond coat. The major factors controlling life are the maximum temperature and the severity of the thermal cycling, and life-prediction methodology has been developed [42].

3.5. Wear-Resistant Surfaces

Lubrication is probably the most common technique for reducing wear by lowering friction and separating the contacting surfaces. Liquid lubricants are limited to ca. 250 °C by their thermal stability. Solid lubricants are used at higher temperature, with $MoSi_2$-based materials being used up to 300 °C and graphitic lubricants to 400 °C.

Many situations require the use of wear-resistant surfaces. Such surfaces can be produced on steels by thermochemical diffusion processes such as carburizing or nitriding. Nitriding has the higher temperature capability because nitriding steels are normally more temper resistant than carburizing steels. Nitrided surfaces can be used up to 500 – 600 °C, and one of the earliest applications was to improve the life of high-speed steel machining tools. The temperature of the cutting edge can exceed 700 °C under some conditions, and nitriding increases tool life up to 500 % [43]. Hot extrusion dies and pistons in the Stirling engine are other applications. Nitriding can be carried out by gas, salt bath, or plasma techniques [44–46].

General requirements for high-temperature, wear-resistant coatings include hardness, toughness, strength, and corrosion resistance. Specific requirements depend on the particular wear environment; for example, hot hardness is more important in high-impact loading. Most coatings contain stable carbide-forming elements (e.g., chromium, tungsten, and vanadium), with hardness increasing and toughness decreasing as the carbide volume fraction increases. The binder phase is commonly cobalt, cobalt – chromium, or nickel – chromium, depending on the corrosion resistance and toughness required. One factor contributing to the performance of cobalt-based coatings is the beneficial effect of cobalt oxides formed during high-temperature wear. Coatings are applied by weld deposition, electroplating, flame spraying, or a detonation gun [47]; examples of the potential improvement relative to uncoated materials are shown in Figure 27. Applications include Cr_3C_2 – NiCr and hard cobalt-based, heat-resisting alloys on turbine blade shroud faces, WC – Co on fan blade midspan stiffeners in gas turbine engines, and

WC or Cr_3C_2 in cobalt-free matrices in the nuclear power generation industry.

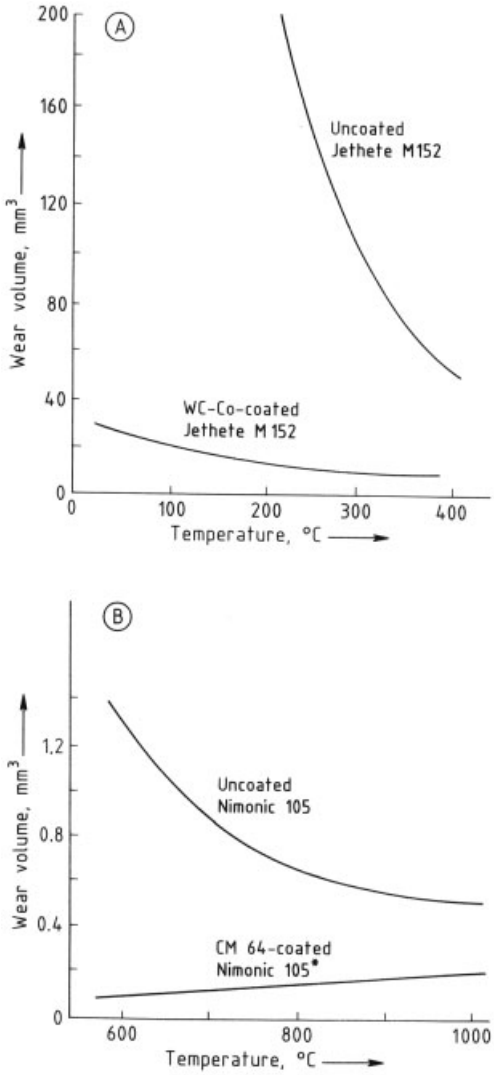

Figure 27. Effect of coating on wear performance
A) Under hammer wear conditions; B) Under reciprocating sliding wear conditions
∗ CM 64: Co – 28 Cr – 19 W – 1 V – 5 Ni – 0.8 C

4. Aluminum Alloys

The first official alloy designation was drawn up for commercially pure aluminum in 1888. Many different alloys have since been produced. During the past 20 years, substantial developmental work has been carried out in the fields of powder metallurgy, rapidly solidified alloys, and aluminum matrix composite materials.

4.1. Conventional Alloys

The basis for high-strength alloys was the work carried out in Berlin in 1906 by WILMS, in which the existence of age hardening in Al – Cu alloys was discovered. The development of high-temperature aluminum alloys depends upon precipitation hardening (see Section 3.1.2), and in the highest temperature alloys this is supplemented by dispersion hardening (see Section 3.1.3) with various high-melting intermetallics. The major precipitation-hardening systems are Al – Cu, Al – Mg – Si, and Al – Zn – Mg. The materials are produced as wrought and cast alloys. The production processes for wrought and cast aluminum and the Aluminum Association designations for alloys and tempers are discussed elsewhere (→Aluminum Alloys). Compositions of conventional aluminum alloys are listed in Table 5.

4.1.1. Wrought Alloys

Al – Cu – Mg Alloys (2XXX Series). The early work of WILMS led to the production of the alloy Duralamin, which was used for the rigid framework of Zeppelin airships; subsequent developments (e.g., 2014) have been used widely for aircraft construction. The relatively high silicon content in 2014 increases the hardening response on aging. Alloy 2024 is a development of 2014 with increased magnesium and with silicon reduced to impurity level. Both copper and magnesium give rise to precipitation hardening.

The early emphasis in aircraft construction materials was on maximizing tensile properties. The subsequent need to improve defect tolerance led to a greater importance of properties that govern behavior in the presence of cracks and other flaws, especially fracture toughness. As with most metals, toughness varies inversely with strength, and proof strength must be limited in high-strength alloys where fracture toughness requirements are essential. Crack growth is encouraged by the presence of insoluble intermetallic particles that contain silicon and iron (as an impurity); toughness can be improved by

Table 5. Composition (wt %) of aluminum alloys ∗

Alloy	Cu	Mg	Mn	Si	Fe	Ni	Ti	Zn	Sb	Zr
Duralumin	3.5	0.5	0.5							
2014	4.4	0.5	0.8	0.9	0.7					
2024	4.4	1.5	0.6	0.5	0.5					
2124	4.4	1.5	0.6	0.2	0.3					
2618	2.3	1.5		0.2	1	1	0.05			
7075	1.6	2.4		0.4	0.5			5.5		
7475	1.6	2.4		0.1	0.1			5.5		
A332	1	1		12.0		1				
Y alloy	4	1.5				2				
RR 50	1			2.5	1	1				
RR 350	5			0.2	0.2	1.5	0.2		0.25	0.2

∗ The balance is aluminum.

maintaining the combined level of iron and silicon below 0.5 % (Fig. 28) [48]. Alloy 2124 is essentially 2024 with a lower iron and silicon content. The requirement for a material with long-term creep capability at ca. 120 °C (e.g., for the skin of the supersonic aircraft Concorde) led to the development of 2618 which, in its earlier forged form as RR 58, was used for pistons, impellors, and compressor blades in aircraft gas turbine engines. In addition to the precipitation hardening from copper and magnesium, the nickel and iron give rise to dispersion strengthening by the intermetallic compound $FeNiAl_9$.

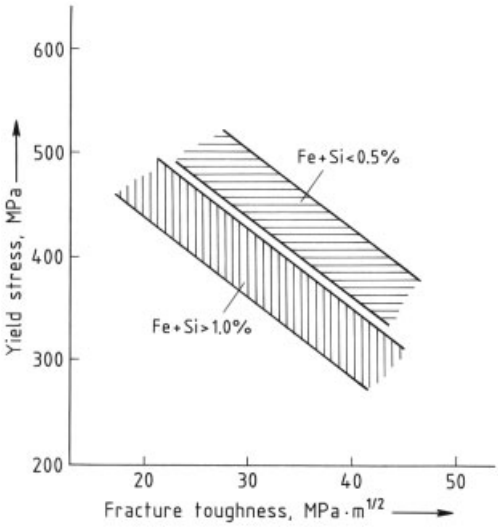

Figure 28. Effect of iron and silicon content on fracture toughness of Al – Cu – Mg alloys

Al – Zn – Mg – Cu Alloys (7XXX Series). World War II was largely responsible for the development of the Al – Zn – Mg – Cu alloy series, which has the greatest response to age hardening of all aluminum alloys; however, susceptibility to stress corrosion restricted the use of the strongest alloys. Control of composition (particularly the zinc – magnesium ratio) and of residual stress was found to be beneficial. Use of a duplex aging treatment to nucleate a fine dispersion of $MgZn_2$ from preexisting GP zones increased the stress corrosion resistance in alloy 7075. This alloy has been used widely in the construction of civilian and military aircraft [49]. A tougher alloy 7475 with reduced silicon and iron content is used in modern aircraft. Alloy 7075 is somewhat stronger than 2024 and RR 58 at lower temperature but has inferior temperature capability (Fig. 29). When used in sheets for aircraft skin, the 2XXX and 7XXX series of alloys are roll clad with aluminum or Al – 1 Zn to provide corrosion protection. Cladding lowers the tensile and fatigue strengths, which must be allowed for in the design.

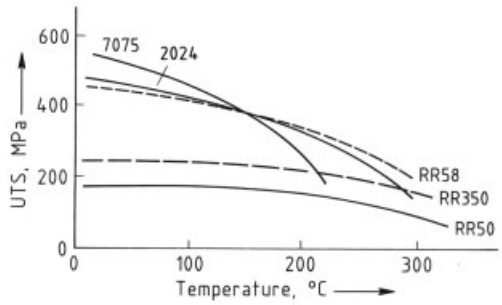

Figure 29. Ultimate tensile strength (UTS) of conventional aluminum alloys

4.1.2. Casting Alloys

High-silicon alloys are important casting alloys because of the high fluidity caused by the Al–Si eutectic. Alloy A 332 is used for pistons in internal combustion engines, but Al–Si alloys generally do not have high-temperature capability. Alloys based on Al–Cu are better, but casting is more difficult, with potential problems in hot tearing and shrinkage porosity. Alloys have been developed with improved high-temperature properties and adequate castability; they utilize precipitation hardening and dispersion strengthening by intermetallic compounds. Y alloy has been used for diesel engine pistons; RR 50 was used for air-cooled cylinder heads of aircraft piston engines and continues to be used in gas turbines for intake and compressor casings.

Developments to further increase temperature capability resulted in RR 350. Grain refinement is achieved by the presence of titanium intermetallics. Dispersions of nickel and zirconium intermetallics increase creep strength, and addition of a small amount of antimony improves castability. The alloy is effective up to 300 °C in jet engine components (e.g., compressor casings).

4.2. Powder Alloys and Rapidly Solidified Alloys

4.2.1. Sintered Aluminum Powder

The first aluminum alloy produced by powder metallurgy techniques was probably *sintered aluminum powder* (SAP) [50]. It was produced by ball-milling aluminum powder in an 8 % oxygen atmosphere. The surface oxide layer that formed on the powder was crushed and dispersed in the aluminum matrix by extrusion, forging, or rolling after the initial pressing and sintering. The amount of oxide present in the final material depends on milling conditions. With ca. 6 % Al_2O_3, a significant level of tensile strength is maintained at a temperature close to the melting point. With up to 20 % Al_2O_3, some strength is retained above the melting point of aluminum, which implies a continuous oxide network. Despite attractive high-temperature strength (Fig. 30), the ductility at elevated temperature is poor and SAP has found relatively few applications. Its major use has been as cans for fuel elements in nuclear reactors.

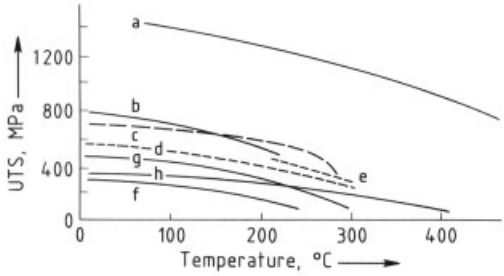

Figure 30. Ultimate tensile strength (UTS) of novel and composite aluminum alloys
a) Al – 50 % longitudinal boron fibers; b) Al – 20 % SiC particles; c) RST 3(Al – 7.5 Cr – 1.2 Fe); d) RST 1(Al – 8 Fe – 3.4 Ce); e) RST 2 (Al – 8 Fe – 1.0 Zr); f) Al – 50 % transverse boron fibers; g) Alloy 2024; h) SAP

4.2.2. Recent Powder Alloys

Recent interest in powder alloys was motivated by the need for increased strength. The major advantages are that dispersed particles (e.g., the Al_2O_3 in SAP) can be incorporated and rapid solidification rates can be achieved because of the small volume of material that solidifies to produce the individual powder particles. This results in metastable alloys that cannot be produced by conventional cast ingot methods. In principle, alloying limits can be increased by enhanced supersaturation, thereby providing greater precipitation hardening and introducing nonequilibrium metastable phases [51]. Modified versions of alloy 7075 have been produced in which the zinc content was increased and cobalt, usually an insoluble element, was added [52]. Some mechanical properties improved, but fatigue crack growth rates were higher; this had also been observed with some alloys of conventional composition made by powder metallurgy techniques. The result has been a cautious approach to the application of powder aluminum technology. Another consideration is that metal working temperatures must be restricted if the advantages of the metastable alloy compositions are to be retained.

The mechanical alloying process has been applied to aluminum alloys [53]. Oxide and carbide dispersions were introduced in small amounts

(2 – 4 vol %), but because of the fine dispersion, the effect on strength was somewhat greater than with the coarser particles typically found in SAP. Mechanical working of the material is important in property development. The alloy IN 9052 (Al – 4 Mg – 1.1 C – 0.8 O_2) is reported to combine the corrosion resistance of the Al – Mg series with the strength of the 2XXX series.

Techniques other than powder technology have been used to achieve the rapid solidification that allows the use of novel alloy compositions (e.g., Al – Fe and Al – Cr alloys in Fig. 30). These include splat cooling (RST 2 alloy), centrifugal atomization (RST 1), and vapor deposition (RST 3) followed by consolidation and mechanical working. Strengthening is achieved by supersaturating the aluminum lattice with iron and chromium and incorporating a dispersion of 20 – 30 vol % fine intermetallic particles such as Al_6Fe and Al_7Cr. Compared with conventional alloying elements (e.g., copper and magnesium), transition elements (e.g., iron and chromium) have low solubilities and low diffusion coefficients, which are beneficial (Table 6) [54]. The creep strength data in Figure 31 [55] show the potential of this technology to improve temperature capability over conventional alloys.

Table 6. Solubility and diffusion coefficients of elements in aluminum

Solute	Maximum solid solubility, wt %	Diffusion coefficient D_0, m^2/s
Copper	5.7	1.5×10^{-5}
Magnesium	17.4	1.2×10^{-5}
Iron	0.05	4.1×10^{-13}
Chromium	0.72	3.0×10^{-11}

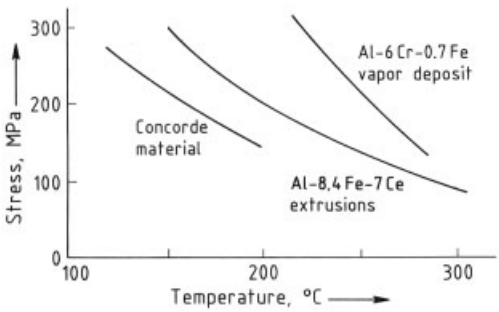

Figure 31. Creep strength (0.2 % strain in 100 h) of rapidly solidified aluminum alloys

4.3. Metal-Matrix Composites

(→Metal-Matrix Composites)

Aluminum matrix composites have been developed considerably since the 1960s. The aluminum matrix can be reinforced with particulate materials (e.g., silicon carbide grit), whiskers, or continuous fibers (e.g., alumina, silicon carbide, or boron coated with silicon carbide). Manufacturing techniques depend on the type of reinforcement: squeeze casting [56] and powder technology are used for discontinuous reinforcements of ca. 20 – 30 vol %, and plasma spraying, liquid metal infiltration, and diffusion bonding techniques using metal foil are employed for continuous reinforcements.

The superior properties of the reinforcement result in benefits such as a considerable increase in strength and stiffness, together with the ability to control thermal expansion (Table 7). The highest strength and stiffness are achieved by using continuous reinforcement such as boron or silicon carbide fibers (Fig. 32 area a; see next page), whereas lesser improvements are achieved by using particle reinforcement (Fig. 32 area b) [51]. The reinforced particulate material has nearly isotropic properties, whereas continuous reinforcement produces highly anisotropic properties.

Table 7. Comparison of properties of potential reinforcements with a typical aluminum alloy

Material	Modulus of elasticity, GPa	Ultimate tensile strength (UTS), MPa
Aluminum alloy 2024	74	450
SiC monofilament	400	3500
Nicalon fiber	200	2500

The upper temperature limit for aluminum matrix composites depends on reinforcement type and orientation. Particulate reinforcement can increase temperature capability to some extent, but the best results are obtained with continuous fibers aligned in the direction of reinforcement (Fig. 30). Much of this benefit is lost in the direction perpendicular to the fiber; the use of two-directional reinforcement to reduce anisotropy reduces the strength relative to the undirectional value.

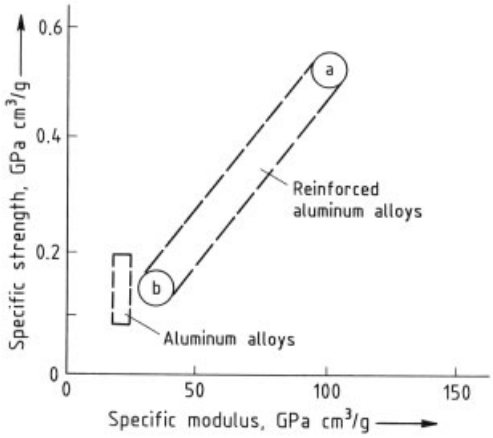

Figure 32. Improvement in properties due to fiber (a) and particulate (b) reinforcement of aluminum alloys

Property improvements are expensive due to the high cost of most reinforcements, but high-volume production could result in significant cost reduction. The cost is likely to result in the use of aluminum matrix composites where performance is critical and economics are secondary. Typical applications could include satellite and space structures; missile, aircraft, gas, and turbine compressor components; and reciprocating engine components (e.g., pistons and connecting rods).

5. Titanium Alloys (→Titanium, Titanium Alloys, and Titanium Compounds)

Titanium alloys date back to the early 1950s. The extraction and melting of titanium present unique problems caused by the high reactivity of molten titanium and its ability to reduce the refractory oxides used for traditional melting crucibles. Consumable electrode melting of titanium in the form of compacted sponge or granules blended with the alloying additions is used to produce ingots [58]. Conventional casting is impossible, and skull melting into rammed graphite molds has been used.

The two properties that motivated the development of titanium alloys are their low density and high corrosion resistance. The density of titanium falls between those of aluminum and iron, and the high specific strength of titanium alloys has resulted in rapid growth in their use in aerospace. Their excellent corrosion resistance is due to the formation of a thin protective layer of TiO_2, which has led to their use in general engineering applications where resistance to salt water and other liquids is necessary.

5.1. Metallurgy

Unalloyed titanium exists in two forms: a close-packed hexagonal α phase stable to up to 882 °C, and a body-centered cubic β phase stable from 882 °C to the melting point (1668 °C). The α phase has a higher inherent temperature capability. The use of heat treatment to develop the properties of titanium alloys is based on the controlled transformation of the β to the α phase. The extent to which the transformation can be controlled is governed by the cooling rate after heat treatment and by alloy composition. Alloying elements are of three types:

1) oxygen, nitrogen, carbon, and aluminum are strong α stabilizers,
2) molybdenum, vanadium, and niobium stabilize the β phase; and
3) tin and zirconium are largely neutral in terms of phase stabilization and very soluble in both α and β phases.

Compound formation occurs with addition of copper and silicon. The Ti – Cu system exhibits age hardening with precipitation of the Ti_2Cu phase, whereas silicon forms the compound Ti_5Si_3 and increases creep resistance significantly [59].

Three basic types of titanium alloys exist: α, $\alpha + \beta$, and β (Fig. 33). A fourth type, near-α alloys is usually included in alloy classifications.

α Alloys. The α alloys contain α-stabilizing alloying additions, the most important are the commercially pure (CP) titanium grades. Oxygen is used as a strengthening element to provide several grades of CP titanium with various combinations of strength and ease of fabrication (Table 8). The alloy Ti 115 is soft and ductile, and is used when maximum formability is required, particularly after welding. The Ti – 2.5 % Cu precipitation-hardening alloy (IMI 230) is as formable and weldable as CP titanium but has better mechanical properties and can be used up to 350 °C. Normal heat treatment is by solution

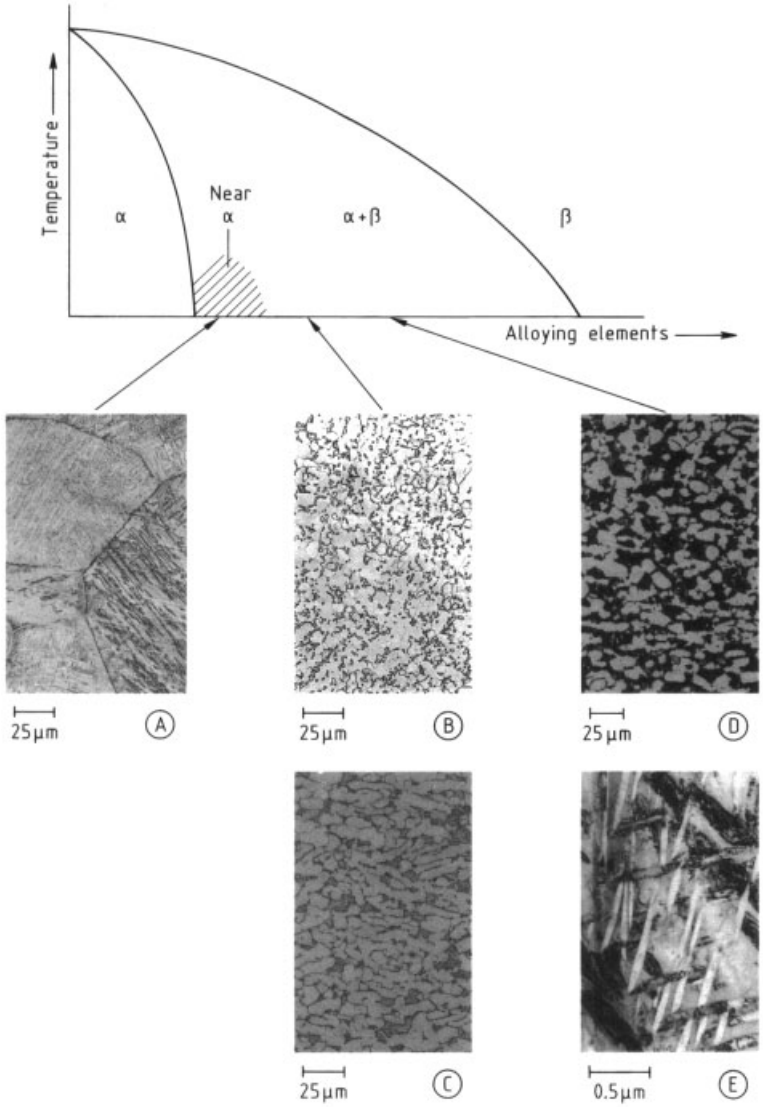

Figure 33. Typical titanium alloy microstructures
A) IMI 685; B, C) Ti 6 – 4; D, E) IMI 550

at 800 °C, followed by age hardening at 400 – 415 °C.

Table 8. Properties of commercially pure titanium

Alloy	Oxygen, %	Minimum 0.2 % PS, MPa	Elongation, %
ICI titanium 115	0.07	200	30
130	0.20	310	20
150	0.30	430	18

($\alpha + \beta$) Alloys. The early ($\alpha + \beta$) alloys (e.g., IMI 315) were forged and annealed at a temperature in the ($\alpha + \beta$) phase field to produce a microstructure of fine, equiaxial α grains with discrete β particles in the grain boundaries; Ti 6 – 4 was first used in this condition (Fig. 33B). However, fracture toughness was subsequently found to improve when two-stage heat treatment was used to produce a microstructure with a significant proportion of coarse transformed β phase (Fig. 33C) [60].

This structure is used in Ti 6 – 4 in compressor disks. The strongly β-stabilized alloys (e.g., IMI 550, Fig. 33D) retain more than the equilibrium amount of β phase on cooling from heat treatment in the $(\alpha + \beta)$ phase field. Subsequent heat treatment at a lower temperature precipitates a fine acicular form of α phase within the β phase (Fig. 33E) and increases strength. The low-pressure compressor rotor of the Olympus 593 engine, which powers the Concorde, uses IMI 550 disks and TI 6 – 4 blades.

β Alloys. Stabilization of the β phase down to ambient temperature requires comparatively large amounts of the dense β-stabilizing elements. These alloys are, consequently, ca. 20 % denser than the α and $(\alpha + \beta)$ alloys. They have inherently poor creep strength because they possess a body-centered cubic structure, and are thus not considered high-temperature alloys.

Near-α Alloys. The need to increase creep resistance led to the development of near-α alloys. The $(\alpha + \beta)$ alloys Ti 6 – 4 and IMI 550 contain too much β phase to have adequate creep strength in the highest temperature applications. The near-α alloys are strengthened mainly by α-stabilizing elements and neutral elements, with minor amounts of β stabilizers. Early alloys such as Ti 8 – 1 – 1 and Ti 6 – 2 – 4 – 2 were $(\alpha + \beta)$ forged and heat treated, and some contained small additions of silicon to increase creep strength. The discovery that significant improvement in creep resistance could be achieved by heat treating in the β phase field and quenching to produce a microstructure consisting of the basket-weave α (Fig. 33A) led to the alloys IMI 685 and IMI 829 [58].

Compositions of the alloys discussed are given in Table 9; proof strength and creep strength are shown in Figures 34 and 35, respectively.

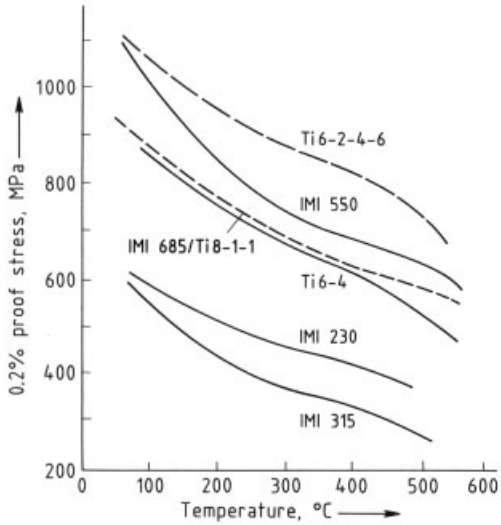

Figure 34. Typical 0.2 % proof strength of titanium alloys

Much activity in recent years has been devoted to the development of alloys based on titanium aluminides. Conventional titanium alloys would require some form of surface protection to prevent the formation of a brittle oxygen-containing surface α layer at a temperature higher than the current maximum capability of ca. 600 °C. Compositions based on the intermetallic compounds TiAl and Ti_3Al have good oxidation resistance and low density. The binary compositions, however, are unacceptably brittle, and research has been directed toward improving the mechanical properties by compositional modifications [61].

Table 9. Composition (wt %) of titanium alloys ∗

Alloy	Cu	Mn	Al	V	Sn	Mo	Si	Zr	Nb
IMI 230	2.5								
IMI 315		2	2						
Ti 6 – 4			6	4					
IMI 550			4		2	4			
Ti 8 – 1 – 1			8	1		1			
Ti 6 – 2 – 4 – 2			6		2	2		4	
Ti 6 – 2 – 4 – 6			6		2	6		4	
IMI 685			6			0.5	0.2	5	
IMI 829			5.5		3.5	0.25	0.3	3	1

∗ The balance is titanium.

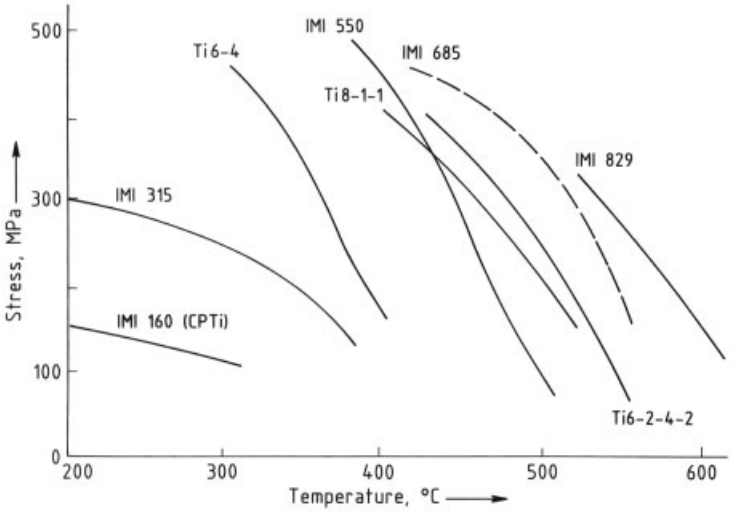

Figure 35. Typical creep strength (0.1 % strain in 100 h) for various titanium alloys

5.2. Uses

The aerospace industry has provided the prime motivation for the development of titanium alloys. In the United States, all titanium alloys produced in 1955 were used in aerospace, with over 60 % being used in jet engine compressors. In 1973, the bulk of production was still directed toward aerospace, but ca. 20 % was used by the chemical and general engineering industries [62].

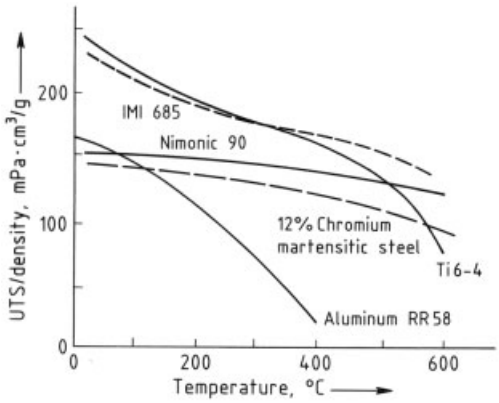

Figure 36. Specific strength comparison of typical jet engine materials

The benefits of titanium derive from its superior strength – density ratio relative to high-temperature steels. The newer high-temperature titanium alloys have similar advantages over nickel superalloys at the compressor temperatures of modern jet engines (Fig. 36). Replacement of steel with Ti 6 – 4 and design changes (made possible in large measure by the use of titanium, including an additional two-stage front fan), resulted in 42 % increase in takeoff thrust, 13 % reduction in fuel consumption, and 18 % reduction in weight in one jet engine development in 1960 [63].

Table 10. Temperature capability of titanium alloys

Alloy	Year available	Approximate temperature capability, °C
Ti 6 – 4	1950	350
IMI 550	1956	400
Ti 8 – 1 – 1	1958	420
Ti 6 – 2 – 4 – 2	1966	475
IMI 685	1967	550
IMI 829	1977	600

Titanium use now accounts for ca. 30 % of the dry weight of current engines. Temperature capability has increased ca. 250 °C relative to the first Ti 6 – 4 alloy (Table 10) [64]. Typical applications include compressor disks, blades, and casings. Some of these applications require weldability, but not all titanium alloys can be welded. The Ti 6 – 4 alloy and some of the more recent near-α alloys can be welded, but the stronger ($\alpha + \beta$) alloys such as Ti 6 – 2 – 4 – 6 cannot. Component reliability is vital in many applications (e.g., compressor disks), and such properties as fracture toughness and crack prop-

agation rate are important design factors. To optimize these properties, the overall forging and heat treatment process is controlled to develop the microstructure [65].

Many airframe components, (e.g., wings and fuselage) are subject to bending loads so that a considerable proportion of the structure is designed in compression. Titanium alloys have a higher compression panel efficiency than other materials except at low compression load intensities (Fig. 37) [66]. Modern military aircraft contain more titanium than earlier aircraft did. The F 14 airframe contains 25 % titanium and the F 15 up to 34 %, compared with 9 % in the earlier F 4. The Boeing 747 airframe uses ca. 14 t of mill-shape titanium [62].

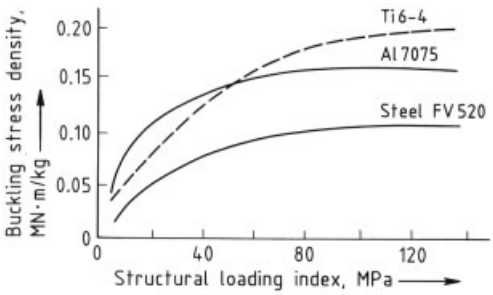

Figure 37. Panel buckling characteristics for steel, aluminum, and titanium alloys

The chemical and general engineering industries use mainly commercially pure titanium because of its excellent corrosion resistance. This property is increasingly important in the chemical, petrochemical, and process engineering industries where titanium is used in tubing for condensers, steam turbines for nuclear- and fossil fuel-fired power plants, and heat exchangers in onshore oil refineries where the corrosion potential of poor-quality brackish or estuary cooling water must be accommodated with temperatures up to 350 °C [62].

6. Steels (→Steel)

High-temperature steels are defined as those which contain alloying elements to improve properties for high-temperature. These elements include chromium, nickel, and aluminum for corrosion and oxidation resistance and relatively small amounts of molybdenum, vanadium, titanium, and niobium for strength. The two basic types of steel are ferritic – martensitic and austenitic, both have high-strength and high-corrosion resistance versions. *Austenitic steels* have a face-centered cubic (fcc) structure and thus an inherent advantage in temperature capability over the body-centered cubic (bcc) *ferritic/martensitic steels,* although the latter have higher proof and tensile strength at low temperature.

6.1. Low-Alloy Steels

Carbon steels are used widely for superheater tubes and steam pipes operating up to ca. 450 °C, but above this temperature low-alloy steels are required. Solid-solution strengthening is developed essentially by addition of small amounts of molybdenum [67]. The major strengthening effect in low-alloy steels, however, is precipitation hardening with coherent carbide precipitates. The intermediate form of precipitate is usually coherent with the matrix, whereas the equilibrium precipitate, which forms above 650 °C, is not. The cubic carbides V_4C_3 and NbC produce the highest creep strength; Mo_2C is more effective than Cr_7C_3.

Several types of commercial low-alloy steel can be found.

Molybdenum Steel. Molybdenum steel is the simplest steel in high-temperature service. It contains 0.5 % molybdenum and is not normally used above 480 °C. It is used in the normalized condition, e.g., with a ferrite – pearlite microstructure that contains an M_3C carbide. This carbide has a low solubility for molybdenum; thus, solid-solution strengthening is the major factor in developing creep resistance. This steel has been used extensively for superheater tubing and other boiler components for steam turbines.

Chromium – Molybdenum Steel. Molybdenum steels suffer from low creep ductility, and a 1 Cr – 0.5 Mo steel has superior creep strength with improved ductility. It is also used in the normalized condition. The addition of chromium eliminates the danger of graphitization that can occur in molybdenum steel, and thus Cr – Mo steels can be used up to 520 °C.

Vanadium-Containing Steels. Further improvement in creep strength can be obtained by addition of a small amount of vanadium (Fig. 38). A 1 Cr – 0.5 Mo – 0.2 V steel is normalized and then tempered at 650 °C to develop the V_4C_3 precipitate. This precipitate does not coarsen rapidly below 600 °C, and the steel can be used up to 580 °C. It has been employed in steam turbine rotors and also used for turbine disks in some early jet engines [68]. Large section disks require steel with better hardenability and a 3 Cr – 0.5 Mo – 0.5 W – 0.7 V steel has been used in such applications, as well as in jet engine disks operating at higher stress than was possible with the 1 Cr – Mo – V material.

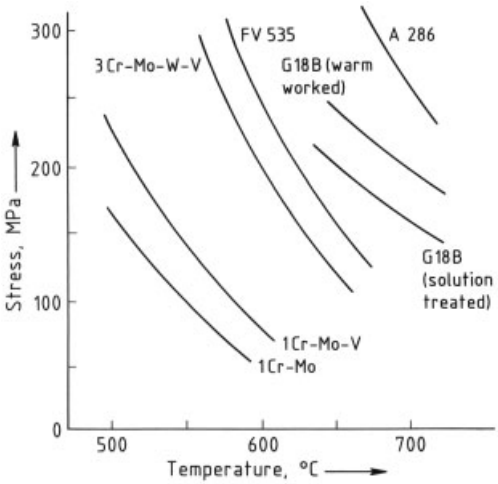

Figure 38. Creep properties (rupture in 1000 h) for various high-temperature steels

6.2. Martensitic 12 % Chromium Steels

The need to increase jet engine component life at the temperature at which the strongest low-alloy steels were operating led to the development of the high-strength martensitic 12 % chromium steels. Steels containing 12 % chromium, sometimes with small additions of molybdenum and vanadium had been used for many years for steam turbine blades because of their high corrosion resistance, but their strength was not equal to that of the low-alloy steels. Addition of further chromium to the 3 Cr – Mo – W – V steel causes an unacceptable loss of creep resistance (Fig. 39) because it results in the replacement of V_4C_3 with the less effective strengthener Cr_7C_3 [69]. The 12 % chromium steels can be strengthened by an M_2X precipitate, however, where X is a combination of carbon and nitrogen [27]; M_2X is stabilized by molybdenum, vanadium, niobium, and nitrogen.

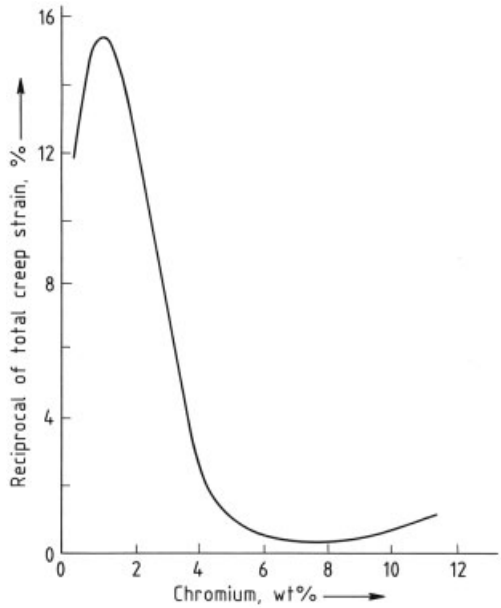

Figure 39. Effect of chromium content on creep strength (strain at 125 MPa, 600 °C) of a low-alloy steel base composition

Molybdenum, vanadium, niobium are ferrite-stabilizing elements and must be balanced by austenite stabilizing elements (e.g., cobalt, nickel, and nitrogen) to maintain the required properties. High-temperature heat treatment at ca. 1150 °C is needed to fully dissolve the alloying elements that form the M_2X precipitate on subsequent tempering at 600 – 650 °C. The creep properties of FV 535 are illustrated in Figure 38. Materials of this type were used for disks in early jet engines although they were superseded by nickel superalloys in subsequent engine designs.

The efficiency of steam plant in electric power generation is increased by operation at higher steam temperatures and pressures. Early subcritical boilers used low-alloy steels, but supercritical plant required higher temperature steels. FV535 was found to be unsuitable for

thick-section components because of chemical segregation. Developments based on 9 % chromium steels used in the petrochemical industry have resulted in steels such as P91 [70], which have allowed operation at temperatures around 600 °C, some 50 °C higher than is possible with low-alloy steels.

6.3. Ferritic Heat-Resistant Materials

According to the definition of steels as iron–carbon alloys with other alloying additions, ferritic heat-resistant materials are not strictly steels and are sometimes referred to as *stainless irons*. Alloys of iron and chromium with <0.05 % carbon are ferritic and cannot be strengthened by heat treatment. They have been used in applications requiring oxidation resistance, but little strength. The oxidation resistance increases with chromium content (Fig. 40) [71] and can be increased further by addition of aluminum and small amounts of rare-earth elements (e.g., yttrium), which reduce oxide spallation by an oxide pegging mechanism. Materials of this type are used as resistance heating elements for furnaces from laboratory- to production-scale size in the ceramic, glass, engineering, and metal-processing industries. Maximum temperature capability depends on composition as illustrated in Table 11.

Table 11. Temperature capability of ferritic, heat-resistant materials

Temperature limit for continuous use, °C	Composition, wt %		
	Cr	Al	Y
1050	15	4.5	
1300	16	5	0.3
1350	22	5	

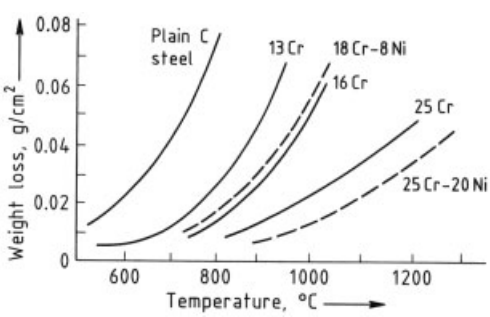

Figure 40. Influence of chromium and nickel on cyclic oxidation resistance of steel

6.4. Corrosion-Resistant Austenitic Steels

The ferritic high-chromium steels have low strength at high temperature, but the addition of sufficient nickel to stabilize the face-centered cubic austenitic structure significantly increases the strength (Fig. 41) with little effect on oxidation resistance (Fig. 40). Steels with higher chromium and nickel content are more suitable than the 18 – 8 types for high-temperature service because of their superior high-temperature strength, oxidation, and corrosion resistance (Table 12).

Chromium carbides precipitate in the grain boundaries when 18 – 8 steels are heated at 500 – 850 °C, producing local chromium depletion that leads to intergranular corrosion when the alloy is subsequently subjected to corrosive environments. Addition of small amounts of titanium or niobium, in quantities related to the carbon content, prevents this problem by the for-

Table 12. Temperature capability and typical uses of corrosion-resistant austenitic steels

Steel	Maximum temperature in oxidizing environment, °C	Typical applications
18 Cr – 8 Ni	850	heat exchanger tubes, exhaust manifolds
24 Cr – 12 Ni	1050	furnace components
25 Cr – 20 Ni	1150	petrochemical equipment, nuclear fuel element cans

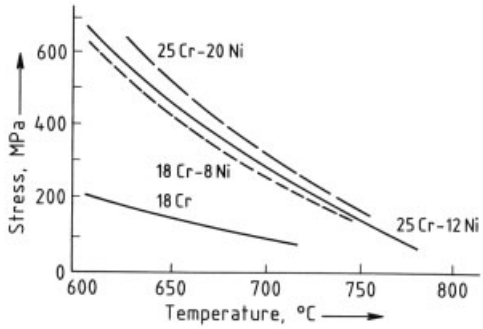

Figure 41. Creep strength (1 % strain in 10 000 h) of corrosion-resistant austenitic steels

mation of stable TiC or NbC. To improve general corrosion resistance, ca. 2 % molybdenum is often added to austenitic steels; this type of steel has been used for many years in superheater tubes for land boilers and reactor heat exchangers. Embrittlement due to precipitation of the σ phase, an iron – chromium-based composition, can be experienced in some steels on long-term exposure at 750 – 850 °C; σ phase formation is accentuated by ferrite-stabilizing elements such as molybdenum and silicon. Selection of steels with composition appropriate for the particular service environment is, therefore, important.

Although nickel has little effect on oxidation resistance, it provides a major improvement in corrosion resistance, particularly in relatively low oxidation potential environments. Steels with high nickel content are required for resistance to carburization (Fig. 42) (e.g., in petrochemical processing to produce ethylene by thermal cracking of hydrocarbon – steam mixtures at 1000 – 1100 °C). Low carbon absorption is essential to minimize the stress gradient, which can result in failure when added to the operating stress [72].

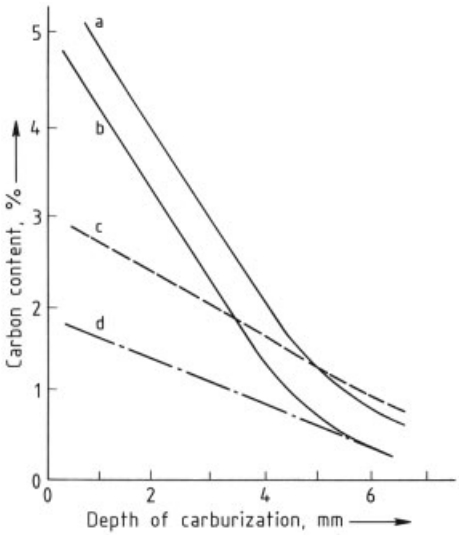

Figure 42. Effect of nickel content on carburization characteristics after 1000 h at 1100 °C in an atmosphere of 65 % N_2, 25 % H_2, and 10 % CH_4
a) 25 Cr – 12 Ni; b) 25 Cr – 20 Ni; c) 18 Cr – 37 Ni; d) 17 Cr – 66 Ni

6.5. Controlled-Transformation Stainless Steels

Some of the 12 % chromium martensitic steels are available in sheet form (e.g., Jethete M 152 and FV 607), but their formability is somewhat limited. The controlled-transformation stainless steels AM 350, FV 520, and 17 – 7 PH combine the excellent formability of the austenitic steels with strength approaching that of the 12 % chromium martensitic steels. Compositions are such that the transformation of austenite to martensite occurs below room temperature. Thus the steel in the solution-treated condition is austenitic, and formability and weldability are good. Strengthening is then achieved by transforming the structure to martensite either by refrigeration close to the martensite formation temperature (M_f) or by heat treatment at 450 – 700 °C. This heat treatment precipitates carbides and raises the martensite transformation range above room temperature. Secondary hardening can be achieved by small amounts of copper and aluminum. Use includes applications where strength and corrosion resistance are required, for example, in jet engine fabrications and in certain parts of the skin of high-speed missiles and aircraft operating up to 550 °C.

6.6. High-Strength Austenitic Steels

High-strength austenitic steels are strengthened either by complex carbides or by intermetallic compounds. The material used by WHITTLE for the turbine blades and disk of his WI engine in 1937 was Stayblade, a simple austenitic stainless steel developed for steam turbine uses. This was found to have inadequate creep strength in the disk rim and was replaced by G 18 B, a carbide-strengthened austenitic steel that had been used for the exhaust valves of reciprocating aircraft engines. Although G 18 B was found to have inadequate proof strength in the disk bore, warm working at 650 – 900 °C significantly increased proof and creep strength (Fig. 38) [73]. A similar concept was used in the United States with 16 – 25 – 6, a 16 Cr – 25 Ni – 6 Mo steel. The materials strengthened by intermetallic compounds are based on Tinidur, which is precipitation hardened with an $(Fe,Ni)_3Ti$ precipitate; A 286 steel was developed from Tinidur. In addition to the

major alloying elements, A 286 contains a small amount of molybdenum to control the morphology of the precipitate and thus improve notch ductility at elevated temperature; a small amount of aluminum and boron improves hot workability. Together with a similar steel Discaloy, A 286 was used in early jet engine disks and is still employed for turbine disks in industrial gas turbines.

Compositions of the steels discussed in this section are given in Table 13.

7. Nickel Alloys (→Nickel Alloys)

Nickel – chromium alloys date back to the early 1900s when their high-temperature oxidation resistance was first recognized. The development of nickel alloys has had two objectives: (1) increased oxidation and corrosion resistance, with strength improvement being of secondary importance, and (2) increase in strength and temperature capability, which has been motivated almost entirely by the gas turbine. This has been so successful that nickel superalloys can be used in stressed applications up to $0.8\,T_{\mathrm{m}}$, a greater proportion of their melting temperature than any other alloy system.

7.1. Metallurgy

The alloying additions that may be present in high-temperature nickel alloys are shown in Figure 43. The heights of the major alloying element blocks give some indication of the amount that may be present, and their main functions are indicated [74]. Trace elements can have a significant influence on properties and behavior; beneficial trace elements are cross-hatched, and detrimental trace elements are shaded horizontally.

The range of microstructures that can exist in nickel alloys is shown in Figure 44. Chromium and aluminum are sufficiently soluble in the nickel matrix to allow protective scales of either Cr_2O_3 or Al_2O_3 to be formed, depending on composition. The two oxide formers have a powerful synergistic effect; in ternary alloys, chromium significantly reduces the amount of aluminum required to form Al_2O_3 compared with that required in binary Ni – Al alloys (Fig. 45) [75]. To some extent, temperature and time influence the oxides that are formed, and the boundaries are thus shown as dashed lines.

Solid-solution, precipitation, and dispersion strengthening are all important strengthening mechanisms. Precipitation strengthening involves the γ' phase based on Ni_3Al. Other alloying elements can be substituted for nickel and aluminum, modify the characteristics of γ', and thus modify alloy behavior. Cobalt can replace nickel; titanium, niobium, tantalum, and hafnium can be substituted for aluminum; and molybdenum, chromium, and iron can replace both nickel and aluminum. Too little aluminum results in a metastable γ', which transforms to a weaker platelike η phase (Ni_3Ti) if the titanium –aluminum ratio exceeds a critical value, or to Ni_3Nb if the alloy contains more than 4 % niobium. The influence of the $\gamma - \gamma'$ mismatch, size, and volume fraction of γ' precipitate is discussed in Section 3.1.2.

Table 13. Composition (wt %) of some steels ∗

	C	Cr	Ni	Mo	W	Nb	V	Cu	Al	Ti	Co	N	B
FV 535	0.08	10.5	0.6	0.7		0.3	0.2				6	0.02	
Jethete M 152	0.10	11.7	2.5	1.7			0.3						
FV 607	0.15	11	0.7	1.0			0.25						
AM 350	0.10	17	4	2.75									
FV 520	0.05	16	5.5	1.8				1.8					
Armco 17 – 7 PH	0.07	17	7						1.1				
Stayblade	0.2	20	8.5							1.2			
G 18 B	0.4	13	13	1.7	2.5	3					10		
Tinidur	0.1	15	30							2			
A 286	0.05	15	26	1.3					0.2	2			0.015
P91	0.1	9		1		0.06	0.2					0.05	

∗ The balance is iron.

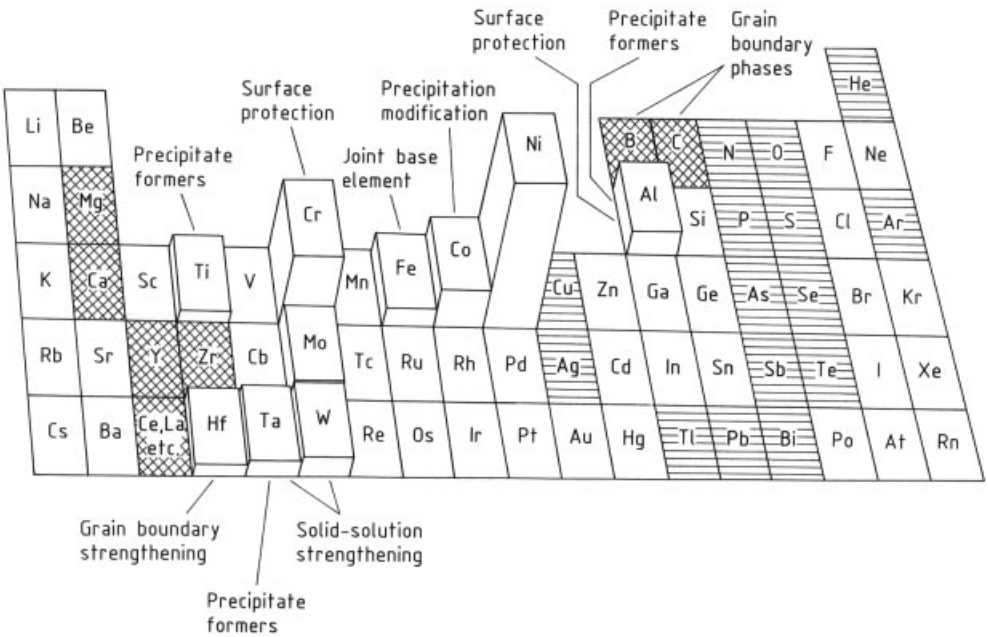

Figure 43. Elements in nickel superalloys and their principal effects (see text for further description)

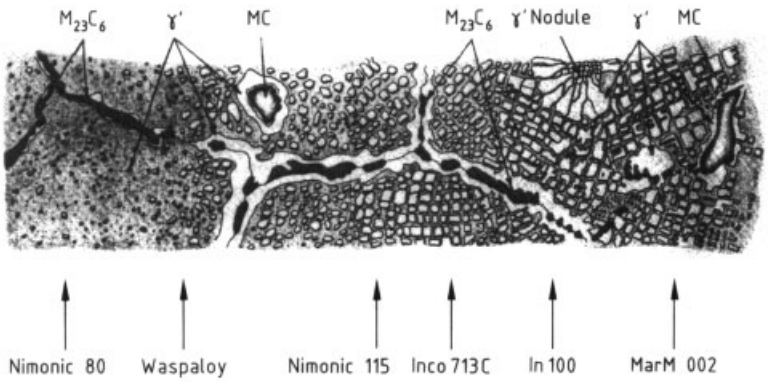

Figure 44. Schematic microstructure of nickel superalloys

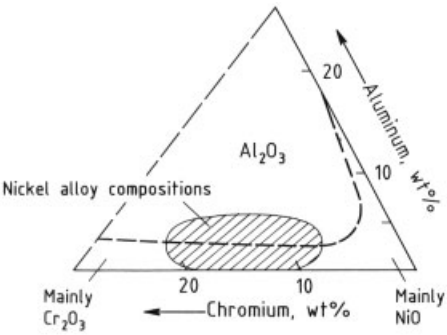

Figure 45. Effects of chromium and aluminum on oxides formed on nickel superalloys

The main types of carbide are $M_{23}C_6$ and MC. The $M_{23}C_6$ carbides tend to predominate in high-chromium alloys. They play a significant role in inhibiting grain boundary sliding when they exist as discrete particles in the boundaries. The MC carbides are very stable and usually form during solidification. They form in alloys containing significant amounts of hafnium, tantalum, and niobium, and can react with the γ matrix to produce $M_{23}C_6$ and γ'. This is an important reaction because it surrounds the grain boundary carbides with a relatively ductile, creep-resistant layer.

Table 14. Composition, temperature capability, and typical uses of electric resistance heating materials

Alloy	Composition, wt % *				Maximum operating temperature, °C	Main application
	Cr	Fe	Si	Al		
Brightray B	16	24			950	low duty domestic heating
Brightray S	20				1150	larger section furnace resistors
Brightray C	20		1.5		1150	frequent and wide temperature variations e.g., domestic cookers, radiant fires
Brightray H	20			3.5	1250	furnace resistors

* The balance is nickel.

In addition to the γ' and carbide phases, which are controlled to develop the beneficial properties of nickel alloys, other phases form in some alloys on long-term exposure at 650–925 °C and have serious adverse effects on toughness and creep strength. The most common of these is the σ phase. It has a platelike morphology and its general formula is $(CrMo)_x(NiCo)_y$.

7.2. Oxidation-Resistant Nickel Alloys

After the excellent oxidation resistance of Ni–Cr alloys had been recognized, the Ni–20 % Cr alloy was found to be particularly suitable for electric resistance heating elements. This type of alloy remains one of the most important materials used in this application, although significant improvements have been made. One modification incorporates small amounts of reactive elements such as calcium, cerium, zirconium, and silicon in Brightray C. These are effective in increasing lifetime in thermal cycling by improving the adhesion of the protective oxide. The compositions and primary applications of some alloys are listed in Table 14.

Since the early development of the electric resistor alloys, a parallel development of general corrosion-resistant nickel alloys has occurred (Table 15). The material that has become the traditional alloy for general industrial furnace parts (e.g., retorts and heat treatment equipment) is Inconel 600. It is resistant to attack by N_2–H_2 mixtures and NH_3, and is a standard material for ammonia cracking plants and nitriding equipment. Incoloy DS is a higher iron version containing silicon. It is particularly resistant to carburization and alternating carburizing–oxidizing environments, and is used extensively in industry under such conditions. It was originally developed for wire-woven furnace conveyor belts but has been used increasingly in other equipment. Inconel 601 is a higher chromium alloy containing aluminum with improved oxidation resistance and strength (Fig. 46). In addition to general furnace equipment, it is also

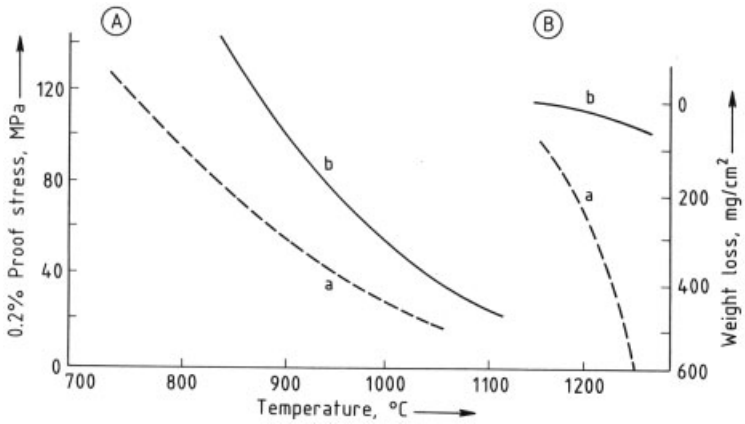

Figure 46. Strength (A) and oxidation resistance (B) of nickel alloys
a) Inconel 600; b) Inconel 601

used in combustion chambers in solid-waste incinerators and in superheater tube supports and ash-handling equipment in power generation.

Table 15. Composition (wt %) of corrosion-resistant nickel alloys

Alloy	Cr	Fe	Si	Al
Inconel 600	15	8		
Inconel 601	23			1.2
Incoloy DS	18	42	2	

7.3. Nickel Superalloys

Nickel superalloys are alloys developed for elevated temperature service in which severe mechanical stress is encountered and high surface stability is often required. They have been developed to meet gas turbine requirements, with specific applications in mind; thus, the major gas turbine components are used as the basis for this discussion. The development of nickel superalloys and their associated manufacturing processes can be divided into four categories: *turbine rotor blade, turbine stator blade, turbine disk,* and *sheet materials*. When first introduced into turbine components, their properties were so superior to those of their predecessors that alloy development alone was sufficient to meet the design needs, with relatively little contribution from the manufacturing process other than to produce the required shape. At a later stage in component development, the manufacturing process became a key part of the material technology because processing made a major contribution to formation of the required microstructure. In more recent developments (e.g., directional solidification), the manufacturing process makes the primary contribution, with material composition being of secondary importance. These trends are evident in the four major gas turbine components discussed in the following sections.

7.3.1. Turbine Rotor Blade Materials

The high-pressure turbine rotor blade operates under the most arduous conditions of temperature and stress of any component in the jet engine; it experiences peak temperatures in excess of 1000 °C, rapid temperature transients at various points in the flight cycle, high direct stress from rotational speeds exceeding 10 000 rpm, and a highly oxidizing environment. The first blades were die forged and superalloys used for blade applications have progressively developed in a process-controlled sequence (Fig. 47) [76]. The volume fraction of γ' precipitate was progressively increased from that in Nimonic 80 to increase creep strength. Cobalt was added, initially in Nimonic 90, to raise the γ' solvus temperature, and solid-solution strengthening was increased by the addition of molybdenum in Waspaloy, Nimonic 105, and later alloys. Alloy content higher than that of Nimonic 115 resulted in serious forging difficulties and therefore Nimonic 115 represents the strongest wrought blade alloy.

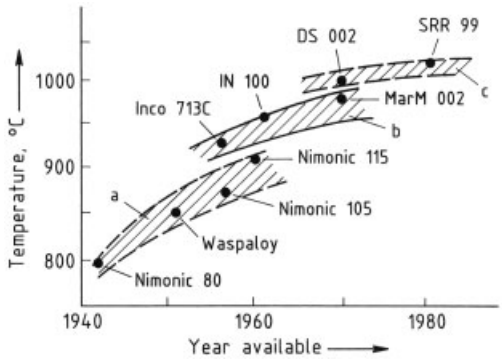

Figure 47. Temperature capability of gas turbine blade alloys based on rupture in 1000 h under stress of 150 MPa
a) Wrought; b) Conventionally cast; c) Directionally solidified and single crystal

Investment or lost-wax casting has existed for ca. 2000 years but was adopted for commercial applications only in the early 1900s. It was first applied in the manufacture of supercharger buckets using the cobalt-base alloy HS 21 in the piston engines of U.S. military aircraft during the 1940s. The reactivity of the aluminum and titanium in nickel superalloys caused problems in air melting and casting; the use of casting for the manufacture of blades in the strongest superalloys was made possible by the development of vacuum melting and casting techniques in the late 1950s [77]. Inco 713 C was one of the first alloys to be vacuum cast and is widely used. As with wrought alloys, the demand for stronger cast alloys was initially met

by increasing the aluminum and titanium content. This led to the development of IN 100, whose microstructure contains 60 – 70 vol % of γ'. Early compositions of IN 100 were found to be metallurgically unstable, with σ phase forming on long-term exposure at intermediate temperatures (650 – 900 °C). This problem was avoided by modifying the composition. Because further increase in the amount of γ' phase to increase the strength was impractical, later alloys such as B 1900 tended to have (1) higher levels of molybdenum and tantalum to increase solid-solution strengthening, and (2) reduced titanium, what was found to improve castability. Stronger alloys (e.g., Mar M 002) were subsequently developed in which molybdenum was replaced by tungsten and hafnium was added to ensure adequate creep ductility at intermediate temperature [78]. Trace elements in quantities ranging from <1 ppm up to ca. 2000 ppm, depending on the element, can have a major effect on creep properties and tensile ductility. As shown in Figure 43, elements such as sulfur, bismuth, lead, and silver can be detrimental, whereas carbon, boron, zirconium, and magnesium have specific beneficial effects [74]. The changes in composition necessary to achieve higher creep strength have relatively little effect on high-temperature oxidation resistance but produce a marked deterioration in hot corrosion resistance (Fig. 48). Developments [79] include alloys with strengths equivalent to the earlier cast alloys such as Inco 713 C, but with improved corrosion resistance. Both IN 738 and IN 939 are such alloys; they have a higher chromium content, higher titanium – aluminum ratio, and modified refractory metal content compared to the less corrosion-resistant alloys. They are employed in industrial turbines.

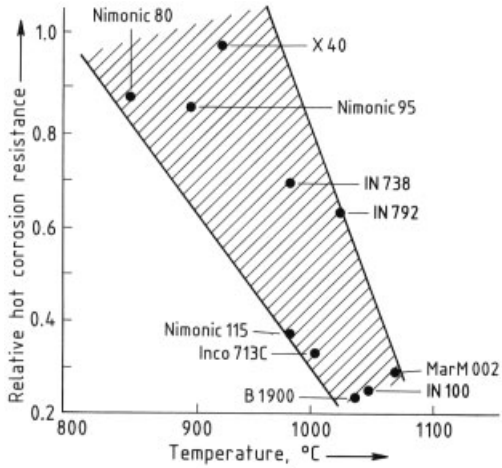

Figure 48. Relation between temperature capability and hot corrosion resistance for various nickel superalloys
The material temperature capability is based on rupture in 100 h under stress of 110 MPa.

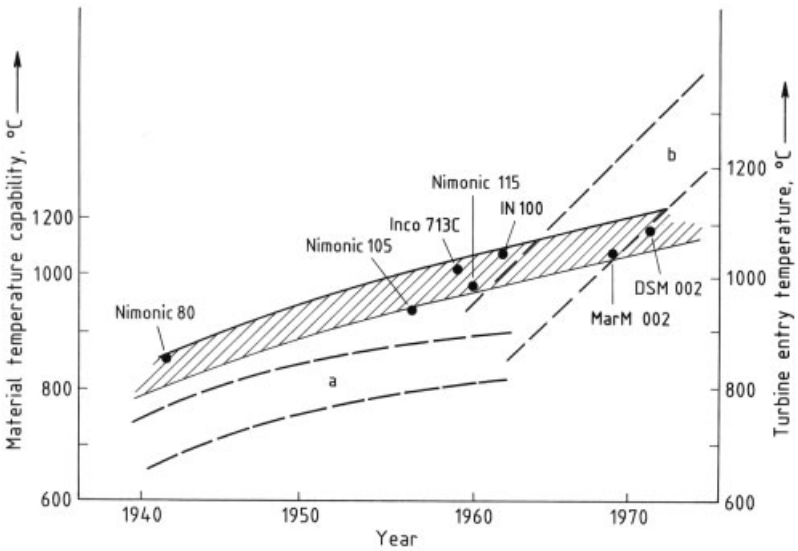

Figure 49. Contributions of material development and blade cooling to turbine entry temperature in aircraft gas turbines
a) Uncooled blades; b) Cooled blades

Since the inception of the jet engine, the need for higher turbine operating temperatures to increase the thrust developed by the engine has continued. In the early years, material development was responsible for the increase in turbine entry temperature (TET) (Fig. 49). In the 1960s, internally air-cooled turbine blades were introduced and blade cooling has played a major role in allowing increased operating temperature. The first cooled blade was forged from a drilled preform to produce a relatively simple three-hole blade (Fig. 50A). More complex and efficient cooling hole geometries can be produced in cast blades (Fig. 50B), and the latest designs feature film cooling; cooling air is exhausted through a multiplicity of small holes to form a film of relatively cool air on the surface of the blade, thereby shielding it to some extent from the hot gas stream.

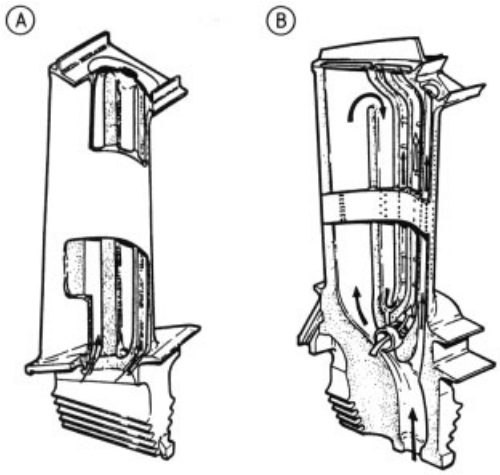

Figure 50. Cooling configurations in Conway (A) and RB 211 (B) high-pressure turbine blades

The technique of directional solidification (DS) was applied to the production of turbine blades in the 1970s in alloys such as Mar M 002 and M 200 DS. Creep failure in conventionally cast blades is initiated at grain boundaries normal to the major direct stress, so that aligning grains from blade root to tip (Fig. 51) eliminates transverse grain boundaries and thus increases the creep life of the blade. To produce the directional grain structure, a macroscopically planar liquid – solid interface must be maintained in the mold perpendicular to the solidification direction; this is achieved by removing heat from the molten metal in a direction parallel to that of the desired growth. Current production processes (Fig. 52) use controlled mold withdrawal from furnaces in which the mold and metal within it have been heated to a temperature in excess of the liquidus [80]. Castings with cooling passages of similar complexity to those of conventional castings can be produced. The absence of transverse grain boundaries is a factor in the increased life of DS blades, but the major benefit arises because the [100] crystal growth direction is also the direction of lowest Young's modulus. This produces major reductions in the thermal stress from temperature gradients in the blade during service. The DS technology has been extended by relatively simple process and alloy composition modifications to produce castings consisting of a single crystal (e.g., in alloys such as SRR 99 and PWA 1480). The major process modification is the incorporation of either a geometric constriction crystal selection device, which eliminates all but one of the grains growing into it, or a seed crystal, which nucleates one grain of crystal orientation like that of the seed.

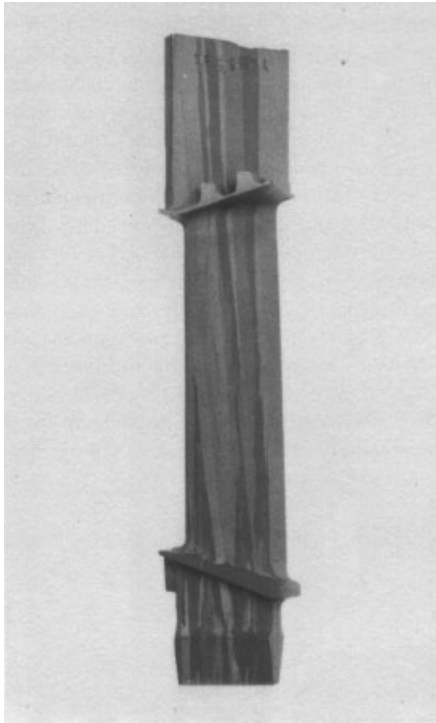

Figure 51. Directionally solidified turbine blade

Alloy composition modification involves the elimination of elements such as carbon, boron, zirconium, and hafnium, which normally provide grain boundary strengthening in polycrystalline castings. Elimination of these elements increases the solidus temperature of the alloy, allowing the use of higher temperature solution heat treatment. This leads to reduced chemical segregation, a more uniform distribution of the γ' precipitate, and consequent increased temperature capability. Such blades were introduced into commercial jet engines in the 1980s. It is essential to control the crystallographic orientation of the blades produced, and computerized real-time techniques are used [81]. The latest alloys used for single-crystal blades, such as CMSX 4, have significantly higher temperature capability than the first-generation alloys. They contain additions of rhenium, which contributes to the improved strength [82].

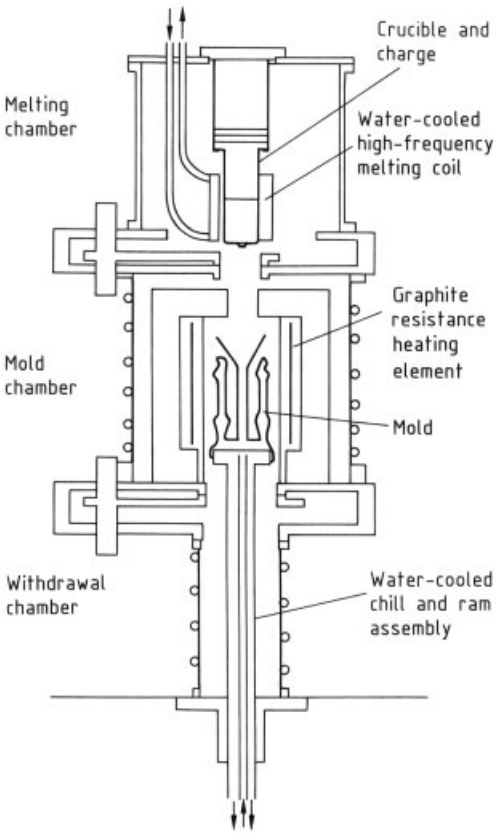

Figure 52. Rolls-Royce directional solidification furnace

Other developmental materials that use directional solidification techniques have not been commercialized. Directionally solidified eutectics (DSE) have been referred to as in situ composites because the reinforcing phase is aligned during the solidification process. Several different DSE materials have been developed [83], and the reinforcing phases are of two types: carbide in NiTaC materials and intermetallic in $\gamma-\gamma'-\alpha$ materials. Reinforcement is inevitably undirectional, which leads to a pronounced strength anisotropy, with transverse strength being much lower than longitudinal strength (Fig. 53). This problem has been addressed (e.g., Nitac 116 A), but cost has prevented the commercialization of DSE. The cost is high because the solidification rates necessary to develop the aligned microstructure are some two orders of magnitude slower than for DS superalloys.

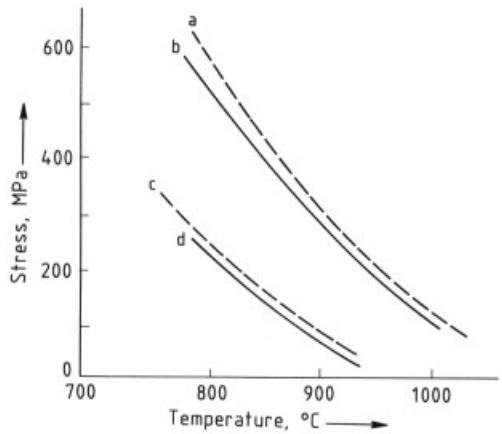

Figure 53. Creep strength anisotropy in directionally solidified eutectics (DSE) based on rupture in 1000 h
a) Longitudinal $\gamma-\gamma'-\alpha$; b) Mar M 002; c) Transverse $\gamma-\gamma'-\alpha$; d) Nimonic 105

Turbine rotor blade material compositions are given in Table 16.

7.3.2. Turbine Stator Blade Materials

Turbine stator blades operate at higher temperatures than rotor blades but, being stationary components, experience essentially only thermal stress. Strength requirements are thus less severe than for rotor blades. Early stator blade materials were the cobalt-based alloys HS 21

Table 16. Composition of superalloys ∗

Alloy	Composition, wt % ∗∗										
	Ni	Cr	Co	Mo	W	Ta	Nb	Al	Ti	Fe	Other
Alloys containing nickel as the principal component											
Nimonic 75	bal	20							0.4		
Nimonic 80 A	bal	20						1.4	2.3		
Nimonic 90	bal	20	17					1.4	2.3		
Nimonic 105	bal	15	20	5				4.7	1.2		
Nimonic 115	bal	15	13	3.5				4.9	3.7		
Waspaloy	bal	20	14	4				1.3	3		
Udimet 500	bal	18	18	4				3	3		
Inconel X 750	bal	15					1	0.7	2.5	7	
Inco 713 C	bal	12		4			2	6	0.8		
IN 100	bal	10	15	3				5.5	4.7		1.0 V
B 1900	bal	8	10	6		4		6	1		
Mar M 002	bal	9	10		10	2.5		5.5	1.5		1.5 Hf
IN 738	bal	16	9	1.7	2.6	1.7	1	3.5	3.5		
IN 939	bal	22	19		2	1.4	1	2	3.5		
Mar M 200 DS	bal	9	10		12		1	5	2		1.8 Hf
SRR 99	bal	9	5		10	3		5.5	1.5		
PWA 1480	bal	10	5		4	12		5	1.5		
$\gamma-\gamma'-\alpha$ DSE	bal							6			
Nitac 116 A	bal	2	3.7	32		8.2		6.5			0.24 C,6.3 Re,4 V
Incoloy 901	bal	12		5.7				0.2	2.8	36	
IN 718	bal	19		3			5	0.5	1	19	
Astroloy	bal	15	17	5				4	3.5		
Rene 95	bal	14	8	3.5	3.5		3.5	3.5	2.5		
Hastelloy X	bal	22		9						18	
C 263	bal	20	20	6				0.5	2		
Rene 41	bal	19	11	10				1.5	3		
MA 754	bal	20						0.3	0.5		0.6 Y_2O_3
Nimonic 86	bal	25		10							0.03 Ce
Alloys containing a metal other than nickel as the principal component											
Haynes 188	22	22	bal		14						0.05 La
MA 956		20						4.5	0.5	bal	0.5 Y_2O_3
HS 21	2	28	bal	6							0.3 C
HS 31 (×40)	10	25	bal		7.5						0.5 C
Mar M 509	10	23	bal		7	3.5					0.6 C

∗ All alloys contain carbon, boron, and zirconium, except the two single-crystal alloys SRR 99 and PWA 1480. Wrought alloys usually contain 0.05 – 0.10 % carbon, 0.005 % boron, and 0.05 – 0.10 % zirconium. Cast alloys generally contain 0.10 – 0.15 % carbon, 0.01 – 0.015 % boron, and 0.05 – 0.10 % zirconium.
∗∗ bal = balance.

and HS 31 used for cast supercharger buckets in piston engines. These materials have a high chromium content (28 and 25 %, respectively) to develop good oxidation resistance. Cobalt alloys do not have a precipitation-hardening mechanism comparable to that of nickel superalloys. They have higher melting temperatures than nickel superalloys, however, and superior creep strength and thermal fatigue resistance at very high temperature. The major strengthening mechanisms are solid-solution strengthening by molybdenum and tungsten, and strengthening by carbide precipitation [84]. Nickel superalloys have been used increasingly in recent years as strength requirements in highly cooled stator blades become more important.

The dispersion-strengthened material MA 754 was used in one recent engine type because of the benefit of strength at very high temperature (Fig. 54). The first practical application of dispersion strengthening in superalloys was TD Nickel (nickel with 2 % ThO_2) in the early 1960s for some combustion chamber components, where the relatively high thermal conductivity of TD Nickel was its major advantage. Poor oxidation resistance was addressed by the addition of 20 % chromium in TD Nichrome. These TD materials were produced by selective hydrogen reduction processes [85]. Mixing of thorium dioxide sols with nickel-containing solutions ensured good dispersion, and subsequent reduction of the metallic oxides after drying pro-

duced fine composite Ni – ThO_2 powder, which was then compacted, extruded, and rolled to develop its strength. The γ'-forming elements aluminum and titanium could not be incorporated because their oxides are not reduced by hydrogen. Mechanical alloying was developed to provide this capability [86]. It is essentially a dry, high-energy, ball-milling process in which particles of elemental or prealloyed powder are continually welded together and broken up until a homogeneous mixture of the constituent powder particles and dispersoid is produced. Consolidation and thermomechanical processing yield the final product. This process has produced mechanical alloy (MA) 754, which contains 20 % chromium and 0.6 % Y_2O_3; MA 6000, a γ'-containing alloy with the potential for turbine blade applications; and MA 956, a potential combuster material discussed in Section 7.3.4.

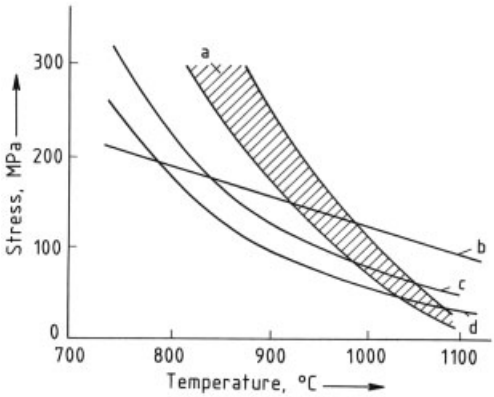

Figure 54. Creep properties of stator blade alloys (1000 h to rupture)
a) Cast nickel superalloys; b) MA 754; c) Mar M 509; d) X 40

Compositions of stator blade materials are listed in Table 16.

7.3.3. Turbine Disk Materials

Superalloy disks were initially produced by the ingot – forged billet – die forged disk method. *Vacuum induction melting* (VIM) was used because of the reactive elements present. The chemical segregation that occurred on solidification of the relatively large ingots led to forging difficulties and the need for a *duplex melting sequence*, i.e., the remelting of VIM cast electrodes by either vacuum arc remelting (VAR) or electroslag remelting (ESR) [87] to control solidification, improve the structure, minimize segregation, and improve cleanliness. This duplex melting provided the basis for microstructural control of the disk superalloys until the late 1960s.

The materials in this period can be divided into two groups: nickel – iron-based alloys such as Incoloy 901 and IN 718, which developed from the precipitation-strengthened austenitic steel A 286, and nickel-based alloys such as Waspaloy and Astroloy, which were developed initially for turbine blade applications and were used subsequently, with modified heat treatment, for disks.

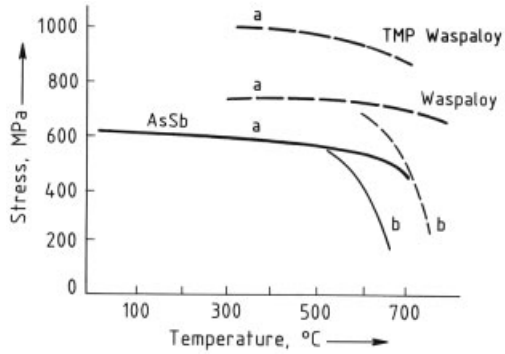

Figure 55. Properties of superalloy disk material compared to earlier austenitic steel
a) 0.1 % proof strength; b) 0.1 % creep in 100 h

When they were introduced, the superalloys exhibited such strength superiority over austenitic steels (Fig. 55) in the fully recrystallized and age-hardened condition that only relatively elementary control of microstructure was required in forging. The disks were die-forged using hammers or presses [88]. Recognition in the late 1960s that low cycle fatigue (LCF) was the major life-limiting factor led to the development of *thermomechanical processing* (TMP), an elementary form of which had been used in the earlier G 18 B and 16 – 25 – 6 austenitic steel disks. In thermomechanical processing, forging conditions and subsequent heat treatment are selected to produce a predetermined microstructure to enhance specific properties. The design of the forging sequence, including the deformation rate, the temperature at which the deformation is carried out in relation to the γ' carbide, and other

Table 17. Effect of grain size on properties of Incoloy 901

	Grain size, ASTM no.	Heat treatment * ST, °C	AGE, °C	0.2 % PS, MPa 20 °C	575 °C	LCF (0 – 850 MPa, 500 °C), cycles
Conventional forging	2	1090	775 + 705	866	752	$>10^4$
Modified forging	6	980	720 + 650	968	839	$>3\times10^4$

* ST = solution treatment; AGE = age hardening

solvus temperatures, is of crucial importance. The increase achieved in 0.1 % proof stress (PS) of Waspaloy is shown in Figure 55. The low cycle fatigue properties were initially improved by reducing grain size; the improvement in Incoloy 901 is shown in Table 17. The necessary structural control extends to the dislocation level [89], which plays a major role in determining component behavior. Homogeneous dislocation distributions (Fig. 56A) are satisfactory, whereas heterogeneous dislocation distributions with local high-density planar arrays (Fig. 56B) can result in premature component failure.

Attempts to develop higher fatigue strength by forging materials more highly alloyed than Waspaloy achieved little success because chemical segregation in the ingot made forging more difficult. Techniques such as hot die forging at 650 – 900 °C and isothermal forging with molybdenum TZM dies at 1000 – 1150 °C, compared to a conventional die temperature of 200 – 350 °C, were developed to provide improved deformation control during forging, but adequate microstructural control was not achieved.

This problem led to the introduction of a prealloyed powder technology for disk manufacture in the mid-1970s; a billet made by the compaction of atomized powder particles is much more chemically homogeneous than a cast ingot. The powder is produced in an inert atmosphere to avoid contamination, and argon atomization and the soluble hydrogen process are most commonly used [85]. The sieved powder, typically <50 µm in diameter, is placed in evacuated containers and consolidated either by extrusion to billet or by *hot isostatic pressing* (HIP) of canned powder to billet or rectilinear ultrasonic inspection shape. Final forging is via conventional press forging, hot die forging, or isothermal forging; superplastic processing has also been used. Materials such as Astroloy and Rene 95 are produced by this powder technology, and their strength shows significant improvement over earlier materials (Fig. 57, see next page). A better understanding of the optimum microstructure for improvements in low-cycle fatigue has accompanied their development. The fatigue performance of such powder superalloys (Fig. 58, see next page) is limited by small oxide inclusions that result from the melting process and are too small to be detected by nondestructive testing. Current activity is thus directed to fundamentally cleaner manufacturing process development and to materials that have even better defect tolerance. Compositions of disk materials are given in Table 16.

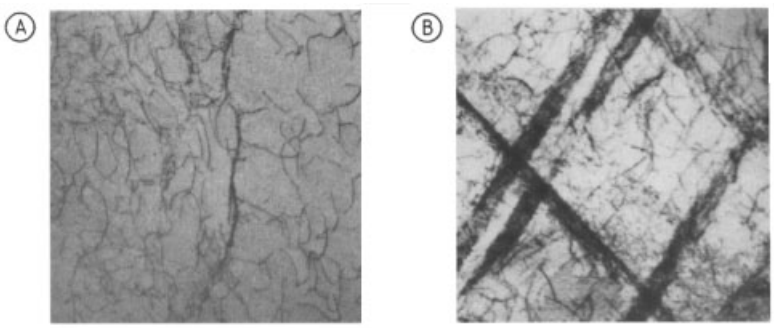

Figure 56. Homogeneous (A) and heterogeneous (B) dislocation distributions in Incoloy 901 forgings

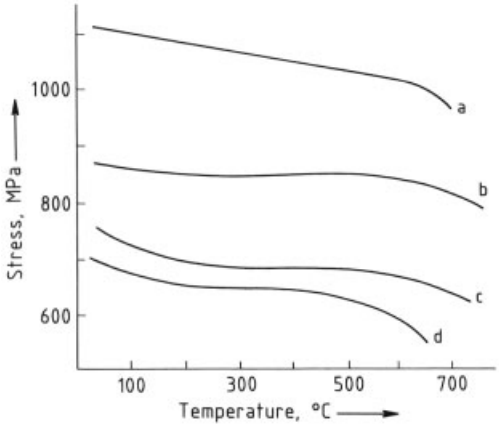

Figure 57. Proof strength (0.1 % PS) of conventional and powder disk superalloys
a) PM Rene 95 (PM = powder metallurgy); b) PM Astroloy; c) Waspaloy; d) Incoloy 901

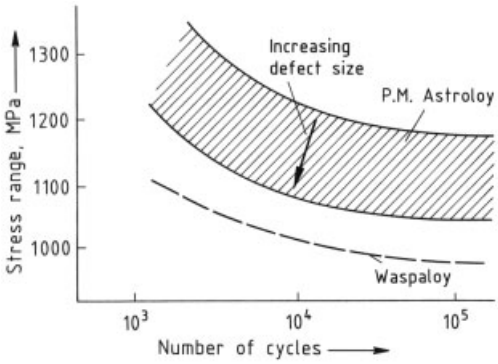

Figure 58. Low cycle fatigue properties of disk superalloys

7.3.4. Sheet Materials for Fabricated Structures

Combustion chambers, high-temperature exhaust units, and jet pipes are examples of components fabricated primarily from sheet material. Weldability and formability are thus prime requirements, and the strength limitation thereby imposed is compatible with the strength and oxidation resistance necessary in these components.

Early gas turbines used several separate combustion chambers, whereas more recent designs are of annular construction. Nimonic 75, one of the first commercial nickel-based alloys, was used in early combustors and is still in use; Hastelloy X has similar capabilities. The need for higher strength combined with weldability has been addressed by the incorporation of aluminum and titanium in C 263 and Rene 41 (Fig. 59) in amounts sufficient to cause precipitation hardening but not high enough to prevent welding. The rate of precipitation in welded regions must be controlled to avoid cracking, and addition of molybdenum controls this in alloys such as C 263 and Rene 41. Increased strength is obtained at the expense of oxidation resistance, however, and the most recent alloys, Nimonic 86 and cobalt-based Haynes 188, are strengthened only by solid-solution strengthening; small amounts of cerium and lanthanum, respectively, are added to improve oxidation resistance.

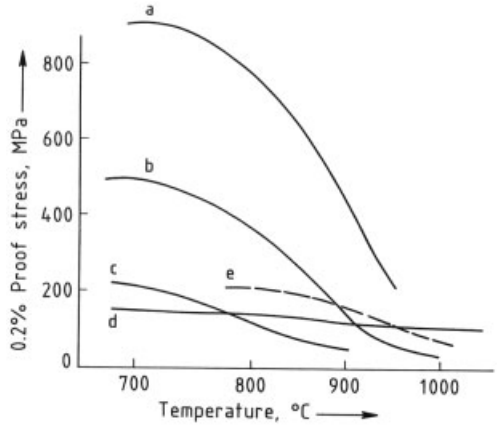

Figure 59. Proof strength of superalloys in sheet form
a) Rene 41; b) C 263; c) Nimonic 75; d) MA 956; e) Nimonic 86

The nickel-based mechanical alloys are not available in sheet form, but an iron-based alloy, MA 956 is available as sheet. It has excellent oxidation resistance (Fig. 60) but, like other dispersion-strengthened materials, cannot be welded. It is not yet used in gas turbine engines but has been subjected to field evaluation in oil-fired boiler hardware and flame stabilizers [90].

Nickel superalloys are sometimes necessary in airframe structures; the precipitation-hardened alloy Inconel X 750 was used for a large part of the skin of the experimental X 15 rocket plane, and Inconel 600 and Rene 41 have been used for the skin of space capsules. Compositions of sheet superalloys are given in Table 16.

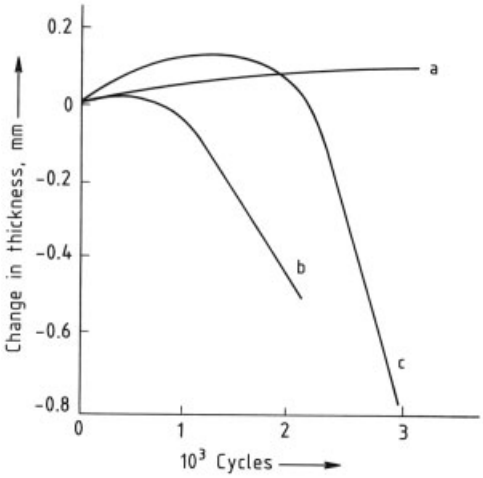

Figure 60. Cyclic oxidation behavior of sheet superalloys (one cycle lasts 30 min)
a) MA 956 (1150 °C); b) Nimonic 86 (1150 °C); c) Inconel 617 (1100 °C)

7.3.5. Nonaerospace Applications

The development of nickel superalloys was motivated by the gas turbine engine, but they have been used in a variety of nonaerospace applications where their unique combination of high-temperature strength and corrosion resistance are required.

Materials such as Nimonic 80, Inconel X 750, and Nimonic 90 have been used for internal combustion engine exhaust valves, which are subject to high temperature, impact, gas erosion, and corrosion. They have also been used for precombustion chambers in indirect injection diesel engines. The use of Inconel 600 and Nimonic 75 in exhaust manifolds for aircraft piston engines has been extended to the exhaust manifolds of high-performance sports cars.

Nimonic 90 has been used for high-temperature springs where conventional spring steels are limited to ca. 400 °C. Materials such as Nimonic 80 have been used for high-temperature bolts (e.g., in steam turbine flanges) because they have much greater resistance to stress relaxation than low-alloy steel.

Nickel superalloys are used widely in the metal-working industry for components such as hot forging dies, extrusion dies, and forging press tools. Nimonic 90 is reported to be three times more effective than conventional die steels in the extrusion of copper because of its higher thermal shock resistance and better wear properties [91].

8. Refractory Metals

Refractory metals vary according to the individual defining them, but all definitions include molybdenum, niobium, tantalum, and tungsten. These metals are all body-centered cubic in structure, and their density increases with increasing melting point. Molybdenum and tungsten are susceptible to loss of ductility at low temperature; the ductile – brittle transition temperature (DBTT) for recrystallized material is given in Table 18. The major strengthening mechanisms are solid-solution strengthening by other refractory metals and by interstitial elements at intermediate temperature, and dispersion strengthening at high temperature. In molybdenum, relatively small amounts of titanium and zirconium increase the strength by both solid-solution strengthening and the formation of small coherent precipitates of TiC and ZrC, which also modify the recrystallization behavior (Table 19). A fine dispersion of 0.5 – 2.0 % ThO_2 increases the strength of tungsten, whereas most niobium and tantalum alloys are strengthened by other refractory metals. The tensile properties of the most common refractory metal alloys are shown in Figure 61 (see next page).

Table 18. Properties of refractory metals

Metal	Melting point, °C	Density, g/cm³	DBTT, °C
Tungsten	3410	19.3	300
Tantalum	2996	16.6	<−196
Molybdenum	2610	10.2	0
Niobium	2415	8.56	−200

Table 19. Effect of alloying additions on recrystallization temperature of molybdenum

Alloy	Recrystallization temperature, °C
Mo – 0.02 C	1200
Mo – 0.5 Ti – 0.02 C	1350
TZM Mo – 0.5 Ti – 0.07 Zr – 0.02 C	1500

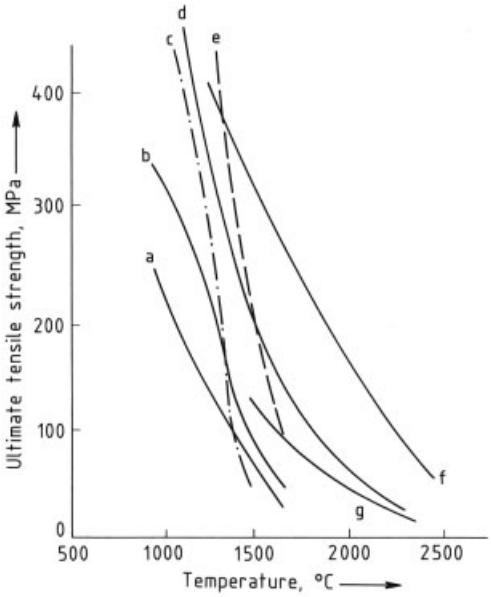

Figure 61. Tensile strength of typical refractory metal alloys
a) C 103 (Nb – 10 Hf – 1 Ti); b) Mo; c) Nb – 15 W – 5 Mo – 1 Zr; d) W; e) TZM (Mo – 0.5 Ti – 0.07 Zr – 0.02 C); f) W – 1 ThO_2; g) Ta – 10 W

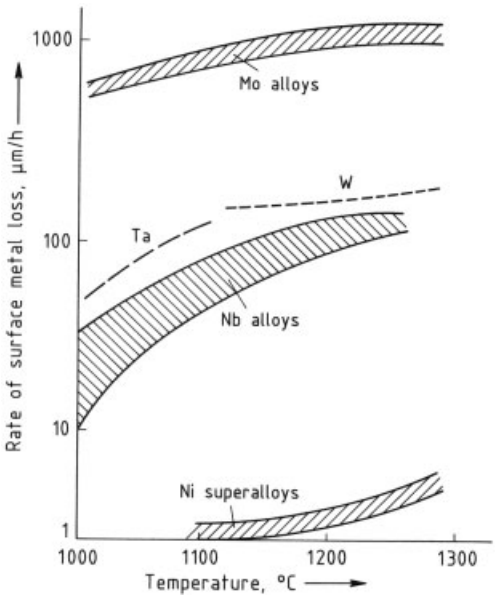

Figure 62. Comparison of oxidation of refractory metals with that of typical nickel superalloys

Refractory metals and their alloys are readily oxidized at high temperature, as shown in Figure 62. Considerable effort was made in the 1950s and 1960s to improve oxidation resistance because of the perceived potential for increasing operating temperatures in gas turbine engines. However, refractory alloys require coating protection in high-temperature oxidizing environments; most useful coatings are based on silicides [92]. Many commercial applications involve nonoxidizing atmospheres or vacuum, and coatings are then unnecessary.

Tungsten has been used in the filaments of incandescent electric lamps since the early 1900s. Its low vapor pressure and its strength at high temperature are essential properties. The filaments consist of wire 0.015 – 0.075 mm in diameter wound in the form of a fine helix with a diameter two to six times that of the wire. The filament operates at ca. 2500 °C in an atmosphere typically consisting of 93 % argon and 7 % nitrogen. To ensure high efficiency throughout the life of the lamp, the filament must remain substantially free from distortion; tight dimensional tolerances and metallurgical control are necessary to give a stable grain structure. Two techniques have been used to achieve this: (1) the production of a fine-grained structure in which further growth is inhibited by the presence of up to 1 % ThO_2 and (2) the production of a stable large-grained structure. Slow grain growth occurs in material containing ThO_2 and its use is limited to special applications where resistance to vibration is essential. The stable coarse-grained approach predominates in most applications. Addition of potassium and other dopants, together with controlled processing, results in wire with large axially aligned grains. Such a structure is resistant to further grain growth and has good high-temperature creep strength, which is essential for ensuring stability of the filament throughout its burning life [93]. Other high-temperature uses of tungsten include welding electrodes, springs, and electrical contacts. Molybdenum is used for heating elements and shields in vacuum furnaces, incandescent light bulbs to support tungsten filaments, X-ray tubes, and electronic components. The TZM alloy is used for cores and dies in die casting brass, aluminum, and steel, where its excellent thermal diffusivity and resistance to heat checking are required [94]. The elevated strength of TZM has led to

its use in extrusion and forging dies, e.g., in the isothermal forging of nickel superalloy disks for turbines.

Because of their high cost and limited availability, tantalum alloys are not likely to be used in large quantities for high-temperature applications. They are employed in specialized applications, for example, as heaters and shields in furnace hardware for the manufacture of tantalum capacitors, an application in which tungsten would lead to unacceptable contamination. Niobium alloys such as C 103 have been used for rocket components operating at 1100 – 1350 °C and for flaps in reheat systems in military jet engines. A Nb – 1 Zr alloy is used extensively in nuclear fusion reactors containing liquid metals because of its high-temperature strength and low absorption cross section for thermal neutrons.

9. Engineering Ceramics

9.1. Refractories (→Refractory Ceramics)

The earliest types of ceramics were clay-based refractories. The most common type, fireclay in the form of bricks, has been widely used in the metallurgical industry (e.g., to line blast furnaces, cupolas, heat treatment furnaces). Compositions range from 20 to 45 % Al_2O_3 and from 50 – 80 % SiO_2, with several clays being used along with other additions to reduce firing shrinkage and to increase stability. The firing temperature is ca. 1500 °C, and fired material typically contains 50 % glass, 35 % mullite, and 15 % cristobalite. The load-bearing capability of fireclay is directly related to the amount of glassy phase and its viscosity. Higher Al_2O_3 content gives better service under severe conditions. Softening temperatures usually range from 1750 to 1850 °C, depending on composition; typical creep properties of firebrick are shown in Figure 63 [95].

Magnesium oxide is strongly basic and is used in the processing of other basic materials (e.g., open hearth steel-making furnaces, Bessemer converters, and cement clinker processing). Mechanical weakness has been a limitation, but the hot strength of magnesite bricks has been increased significantly by reducing the silicate content and increasing the firing temperature.

The need to conserve heat led to the development of insulating firebricks, which reduced heat loss and allowed the use of thinner walls. The clay for insulating firebricks was made porous, commonly by mixing wood flour with the clay and firing to burn out the wood. Thermal conductivities ca. 25 % those of dense firebrick were obtained. More sophisticated techniques for producing porosity have subsequently been used.

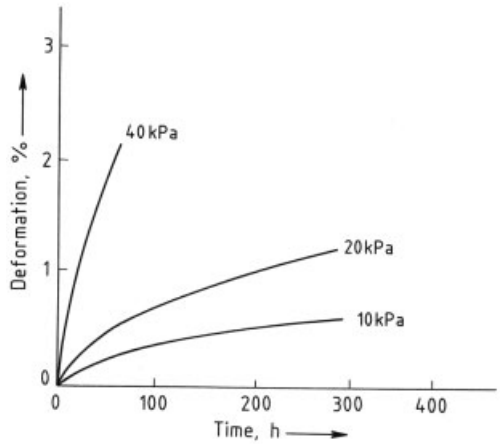

Figure 63. Creep properties of firebrick at 1250 °C

9.2. Early Engineering Ceramics

Aluminum oxide is probably the first engineering ceramic; it has been used for spark plug insulators since the 1930s. For economic and technical reasons, 5 – 15 % SiO_2 is added to the Al_2O_3 in this application, where temperatures up to 850 °C and pressures up to 70 atm (7.0 MPa) can occur with rapid thermal fluctuations. Electric high-temperature furnaces with tungsten or molybdenum heating elements frequently use sintered Al_2O_3 electric insulators, and in these applications, Al_2O_3 can withstand temperatures up to 1850 °C for long periods. Silicon carbide, with a 1600 °C temperature capability in air, has long been used for kiln furniture and electric resistance heating elements. These elements are made from high purity α-SiC sintered at >2300 °C. Molybdenum disilicide ($MoSi_2$) is also used in resistance heating elements for operation up to 1700 °C; its oxidation resistance has led to its use as a protective coating on SiC elements.

9.3. Advanced Engineering Ceramics

The applications discussed in Sections 9.1 and 9.2 involve relatively low stress and moderate to high temperature. Rudimentary understanding of material behavior has allowed the improvement of properties such as strength and temperature capability.

During the last 25 years the use of ceramics in high-temperature applications involving significant tensile stress has increased largely because of an improved understanding of the relationship of processing, microstructure, and properties, and the consequent improved materials and manufacturing technology. Ceramics with better strength and temperature capability are now available. More significantly, progress has been made in reducing brittleness, which had been a major restriction on stress capability.

Several terms (e.g., fine ceramics and advanced ceramics) have been used to describe these new ceramics, but the most appropriate description would probably be *engineering ceramics*. Such ceramics require much greater control over the quality of the raw materials, as well as more sophisticated processing, controlled chemistry, and regulated microstructure. These materials include alumina, silicon carbide, silicon nitride, sialon, zirconia, and glass ceramics (see also →Ceramics, Advanced Structural Products; →Glass Ceramics).

9.3.1. Ceramic Processing

Many of the defects in ceramic components originate as defects in the green unsintered material; others are introduced during processing, shaping, densification, or final machining. Total process control from starting powders to finish grinding is, thus, crucial in the fabrication of satisfactory ceramic components.

Powder purity, size distribution, and mixing are important aspects in the early process stages. Additives include sintering aids; binders to provide green body strength during handling and machining; and plasticizers, wetting agents, and deflocculants to provide the necessary rheological properties during shaping. Important shaping processes include isostatic pressing, extrusion, and hot pressing for relatively simple shapes, with slip casting, injection molding, and green machining used for more complex shapes. In each case, the powder characteristics and additives must be optimized to the particular shaping and subsequent densification processes. The latter involves some form of sintering; it may be reaction sintering as in reaction-bonded silicon nitride (RBSN, see Section 9.3.7) or reaction-bonded silicon carbide (RBSC) and can occur at pressures ranging from atmospheric (pressureless sintering) to >1000 atm (100 MPa) in hot isostatic pressing. Final machining of densified ceramics is difficult and should be minimized as much as possible by the production of net shapes. Full understanding of the mechanisms involved in material removal is essential in minimizing machining damage. The various process stages are discussed in detail in [96].

9.3.2. Thermal Shock Resistance

The ability to withstand rapid temperature changes is a requirement in some ceramic applications. The physical properties of ceramics determine the stress generated and, consequently, the thermal shock resistance because, unlike metals, ceramics cannot deform plastically to redistribute stress. A broad indication of relative thermal shock resistance can be obtained from the equation

$$\Delta T = \frac{\sigma_f\,(1-\nu)}{E\alpha}$$

where ΔT is the temperature change that the material can withstand without exceeding the fracture stress σ_f, E is Young's modulus, α is the coefficient of thermal expansion, and ν is Poissons ratio. Typical values of ΔT calculated according to this equation are given in Table 20. Only a limited number of ceramics are likely to be used in environments involving severe thermal cycling, such as the gas turbine engine, because of these considerations. Silicon nitride and silicon carbide are the ceramics with the earliest potential application.

Table 20. Thermal fatigue resistance of ceramic materials

Material	ΔT, °K
ZrO_2, fully stabilized	80
Al_2O_3, 99 %	100
Sintered SiC	215
Reaction-bonded Si_3N_4	250
Sintered Si_3N_4	625

9.3.3. Component Design and Life

The inability of most ceramics to redistribute stress by plastic deformation increases the importance of accurate prediction of both steady-state and transient stress in component design. This prediction requires advanced three-dimensional thermal and stress analyses including finite element techniques, which permit accurate analysis of the whole component. The traditional design techniques used for ceramic components have been probabilistic in approach. Brittle materials such as ceramics are very sensitive to flaws, and the Griffith equation provides the basis for understanding the relationship between short-term strength, critical crack size, and fracture toughness:

$$\sigma_{\mathrm{f}} = \left(\frac{2E\gamma}{\pi c}\right)^{1/2}$$

Where σ_{f} is the fracture strength, E the elastic modulus, γ the fracture surface energy, and c the flaw size which will propagate at local stress σ_{f}.

One approach to the preparation of stronger ceramics is to reduce defect size [97]. Powder processing, fabrication, and surface finishing of components were identified as key factors in this approach.

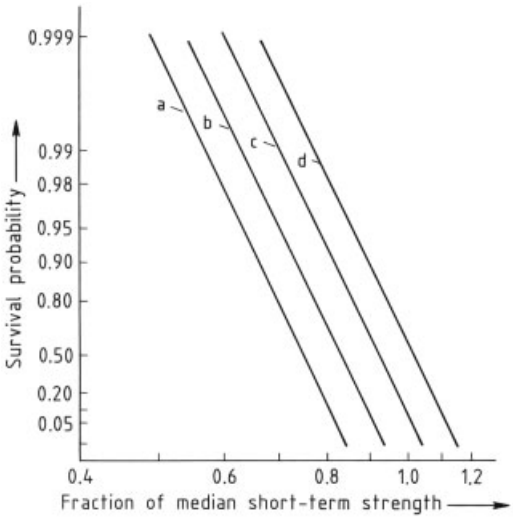

Figure 64. Typical stress – probability – time (SPT) diagram Time, s: a) 10^7; b) 10^5; c) 10^3; d) 10

Life prediction requires knowledge of the size of defects present in the material. Weibull statistics have been used in life prediction of ceramic components. The technique involves the construction of strength – probability – time (SPT) diagrams of the type illustrated in Figure 64 [98]. These are produced from test data generated under relevant test conditions including time-dependent tests.

The Weibull approach was intended to allow the prediction of stress levels with the required survival probability. Some success has been claimed in low-temperature applications [99]. Accurate specification of flaw size by statistical strength characterization is difficult, however.

At high temperature, additional effects produce microstructural or chemical changes in the material, including localized plastic flow, which can induce crack healing or generate creep voids; oxidation, which can heal cracks or introduce new defects; and microstructural changes such as grain growth, phase changes, or surface evaporation. Thus several high-temperature effects can influence strength, and the original defect population present at ambient temperature is not always an accurate reflection of how a material will respond mechanically to stress at high temperature. Another problem is that the small size of critical flaws, possibly in the range of 1 – 100 µm, limits the applicability of nondestructive testing in quality assurance.

A complementary approach to ensuring the specified component strength is proof testing. This has been applied but can introduce additional problems, because the proof test itself can modify the flaw population [100].

The basic problem with the traditional probabilistic approach, however, is that new flaws can be created during service operation and this would invalidate any performance predictions.

9.3.4. Defect Tolerance

Appreciation of the limitations of the traditional lifing approach has led to the realization that for successful application in environments with significant tensile stress, the defect tolerance of ceramics must be increased. This can be achieved in several ways. The fracture toughness of silicon nitride containing Al_2O_3 and Y_2O_3 sintering aids can be doubled by controlling the

sintering parameters to introduce high aspect-ratio grains of β-Si_3N_4 into the microstructure [101]. The expansion accompanying transformation from the tetragonal to the monoclinic form of ZrO_2 has been used to produce toughening in several ceramics. Increasing interest has been shown during the last decade in whisker- and fiber-reinforced ceramics. Sintered Al_2O_3 has a fracture toughness of ca. 2 MPa $\cdot$ $m^{1/2}$, and the incorporation of 20 vol % of SiC whiskers increases this to 8 MPa $\cdot$ $m^{1/2}$. Since 1990 there has been major activity to develop ceramic composite materials in which the incorporation of fibers introduces significant toughening [102] (see Chap. 13).

9.3.5. Alumina

Alumina is currently the most highly developed engineering ceramic. A wide range of high Al_2O_3 ceramics is produced under different trade names, with the Al_2O_3 content varying from 85 to >99 %. Strength and temperature capability increase with Al_2O_3 content, as shown in Table 21 [103], but shaping capability decreases with increasing Al_2O_3. The balance of the composition involves SiO_2, MgO, and other additives to aid shaping and to control sintering temperature. The microstructure of high-Al_2O_3 ceramic contains some glassy phase arising from the sintering additives. Poor thermal shock resistance (see Table 20) and the difficulty of machining sintered material to obtain high dimensional accuracy of complex-shaped components limit the applications of alumina. Indeed, the hardness of α-Al_2O_3 (corundum) has long been utilized in abrasives, either in the form of loose grains or bonded material in grinding wheels. The use of Al_2O_3 as a cutting tool was first considered in the late 1930s, but not until 20 years later could strong, fine – grained, high-quality alumina tool bits be made. Hardness retention at high temperature is an important characteristic because tool temperature can reach 1000 °C during the machining of ductile materials such as steel. A recent development has been to incorporate silicon carbide whiskers in Al_2O_3, which increase the thermal conductivity and, consequently, the thermal shock resistance. They also significantly increase toughness (Section 9.3.4), which has led to improved tool performance.

Table 21. Effect of composition on properties of alumina materials

Property	Al_2O_3 content		
	90 %	95 %	99 %
Modulus of rupture, MPa	200	350	550
Young's modulus *E*, GPa	200	300	400
Softening temperature, °C	1600	1750	1900

Sintered Al_2O_3 is also used for missile radomes (external containers for radar and other equipment). The radome must be transparent to radar signals which requires particular dielectric properties; as an external surface, it is also exposed to aerodynamic and thermal effects during flight.

9.3.6. Zirconia

Zirconia is a highly refractory material, *mp* 2680 °C, which is polymorphic, with a monoclinic phase stable to ca. 1170 °C, a tetragonal phase stable from 1170 to 2370 °C, and a cubic phase stable to 2680 °C. The monoclinic – tetragonal phase change is accompanied by a volume change of ca. 4 %, which is large enough to cause spontaneous failure on cooling large pieces of pure ZrO_2. Phase transformation must, therefore, be prevented in applications such as metal melting crucibles, atomizing nozzles, and extrusion dies. This can be achieved by the addition of CaO, which stabilizes the cubic form of ZrO_2 from room temperature to its melting point.

The volume expansion during the tetragonal to monoclinic phase change can be used to advantage to improve both the strength and the toughness of ceramics. This concept was first proposed in 1975 [104] for a ZrO_2 material and has since been applied to other ceramics. Zirconia particles within a critical size range are introduced into the ceramic, and the constraining pressure exerted by the surrounding matrix retains the tetragonal structure in the ZrO_2. If a crack grows in such a ceramic, the tensile stress around the advancing crack tip allows the tetragonal – monoclinic transformation to take place. The resulting expansion introduces a compressive stress around the crack tip, thereby reducing its tendency to propagate [105]. The addition of 15 vol % ZrO_2 to Al_2O_3 improves fracture toughness to 12 MPa $\cdot$ $m^{1/2}$ [106], and such material is used commercially in modern

Al_2O_3 cutting tools. Transformation toughening in ZrO_2 is achieved by isothermally aging a partially stabilized ZrO_2 (ZrO_2 – 3 % MgO) at 1400 °C to develop small tetragonal precipitates in a cubic matrix (Fig. 65) [106]. This material is used for cutting and slitting industrial materials; the insulating potential in automotive engine parts, where the low thermal conductivity is advantageous, is being evaluated. The low thermal conductivity of ZrO_2 has also resulted in its use in coating jet engine components (Section 3.4).

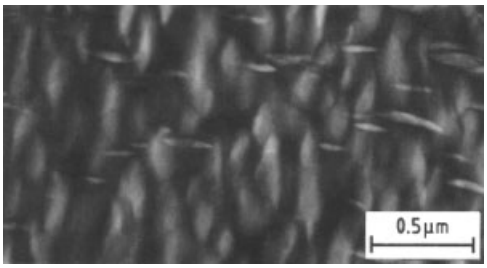

Figure 65. Tetragonal precipitates in transformation-toughened zirconia

9.3.7. Silicon Nitride

Silicon nitride exists in two basic forms: porous and dense. The porous form, *reaction-bonded silicon nitride (RBSN)*, is produced by controlled nitridation of a silicon powder compact in a nitrogen atmosphere at 1400 °C. *Dense silicon nitride* is made by several processes including hot pressing and pressureless sintering.

Being porous, RBSN has the lowest strength over most of the temperature range. The strength depends on bulk density, which can be controlled in the range 2.0 – 2.8 g/cm^3, compared with the theoretical density of ca. 3.2 g/cm^3. Pore morphology is an additional strength-controlling factor. Flexure strength is relatively low but is maintained to >1200 °C (Fig. 66) because the absence of sintering aids means that no glassy phases are present in the microstructure. One manufacturing advantage is that little shrinkage occurs on nitridation. Thus, the pressed and machined silicon powder shape can often be made close to the final dimensions. The porous surface can be sealed by chemical vapor deposition (CVD) of a layer of dense silicon nitride, significantly improving oxidation resistance. Reaction-bonded silicon nitride is used in light alloy foundries for pouring spouts, continuous casting dies, etc., and in heat treatment and brazing fixtures.

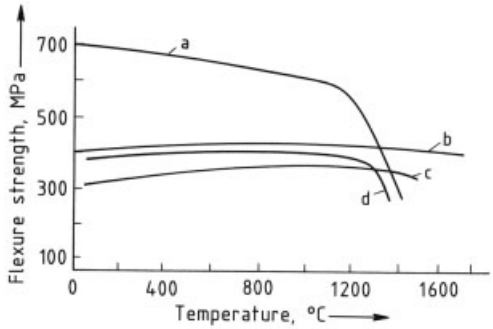

Figure 66. Flexural strength of silicon nitride and silicon carbide materials
a) Hot-pressed SiN; b) Sintered SiC; c) Reaction-bonded SiN; d) Reaction-bonded SiC

The production of dense silicon nitride requires the incorporation of sintering aids. Sintering may be carried out at atmospheric or slight overpressure to produce sintered material (sintered silicon nitride, SSN). Sintering takes place during hot pressing (hot pressed silicon nitride, HPSN) or hot isostatic pressing; the amount of sintering aid required generally decreases with applied pressure. Oxides such as MgO, Y_2O_3, and Al_2O_3 are incorporated as sintering aids, and densification occurs at 1600 – 1800 °C by solution and precipitation of β-Si_3N_4 from the liquid produced by the sintering aids. The resultant microstructure contains two or more phases depending on the aids used. The first widely used additive was MgO, which reacts with SiO_2 to create a low-viscosity glassy phase that limits temperature capability. A more refractory glass is produced by addition of Y_2O_3; Y_2O_3-containing materials, therefore, possess higher temperature capability (Fig. 67, see next page).

In the past decade, efforts to demonstrate the potential of ceramic components in gas turbine engines have increased. Dense silicon nitride has been used for various static and rotating turbine components and has been built into automobile turbocharger rotors.

9.3.8. Sialons

Sialons are based on a solid solution of Al_2O_3 in Si_3N_4, where nitrogen is partially replaced

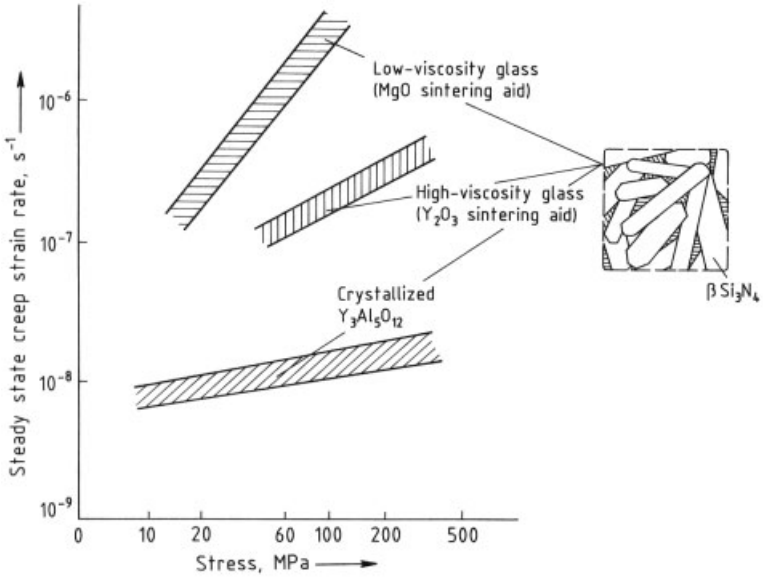

Figure 67. Influence of grain boundary phase on creep properties of silicon nitride and sialon materials at 1350 °C

by oxygen and silicon is partially replaced by aluminum. They also contain additives such as Y_2O_3, because they are made by hot pressing or sintering. The second phase in such materials can be crystallized to $Y_3Al_5O_{12}$ [107]. The result is a higher temperature capability because this is not a glassy phase (see Fig. 67). Sialon containing a glassy second phase has been used for machining nickel superalloy turbine disks at a surface speed of 5 m/s, compared to a maximum of 0.25 m/s for tungsten carbide inserts [108].

9.3.9. Silicon Carbide

Silicon carbide is manufactured in three major ways: reaction bonding, sintering, and hot pressing. Reaction-bonded silicon carbide (RBSC) is produced by infiltrating molten silicon into a powder compact of SiC and carbon. Silicon reacts exothermically with carbon to form silicon carbide and bonds the structure together. Not all of the silicon reacts, however, and the final material contains free silicon, usually more than 8 vol %. A major limitation on temperature capability exists because silicon melts at ca. 1400 °C (see Fig. 66). Although RBSC has been used in developmental gas turbine engines, the major applications (e.g., pump seals) make use of its excellent wear resistance.

Higher operating temperatures are possible with sintered and hot-pressed material. Hot pressing with the addition of Al_2O_3 can give full densification, whereas sintering is achieved by addition of boron and carbon. Cubic β-SiC was used initially, but the growth of large transformed grains of the high-temperature α phase in the β matrix made processing difficult. Subsequently, α-SiC was successfully sintered. The usual manufacturing route is to press blanks, machine in the green state, and then sinter.

Silicon carbide and silicon nitride have been used for major high-temperature components in ceramic gas turbine demonstrator programs during the past decade.

9.3.10. Glass Ceramics (→Glass Ceramics)

The concept of manufacturing polycrystalline materials from glass is ca. 200 years old and has undergone major development in recent years. The use of glass-shaping technology allows automated production of relatively complex shapes with close dimensional control. The molten stage leads to excellent property consistency to which the fine grain size (typically

1 – 5 μm) contributes. Bend strengths of 75 – 400 MPa are possible, with temperature capabilities approaching 1400 °C, depending on composition.

In the manufacturing process, a glass of selected composition, incorporating nucleating agents for crystallization, is melted and shaped. It is then crystallized by a duplex heat treatment sequence with a lower temperature nucleation stage followed by a higher temperature crystallization growth stage. Early nucleants included gold, silver, copper, and platinum; more recently, P_2O_5 and TiO_2 have been used. Typically up to 95 % crystallization is obtained with some residual glass. Expansion coefficient, dielectric properties, and temperature capability depend on glass composition. Cordierite ($2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$) has been used for missile radomes, and a lithium aluminum silicate with low coefficient of expansion is used for domestic cooker hobs. Coefficients of expansion can be tailored to match many metals, and glass – metal joints can easily be made by using a controlled oxide on the metal to provide a good bond with the glass ceramic.

10. Carbon and Carbon – Carbon Composites (→Carbon)

Carbon occurs naturally as graphite and was used for alloy-melting crucibles as early as the fifteenth century [109]. Carbon and graphite have been used in various manufactured forms in high-temperature applications that require high mechanical strength and good thermal shock resistance since the beginning of the twentieth century. Low elastic modulus, low coefficient of thermal expansion, and high thermal conductivity contribute to the good thermal shock resistance. Unlike other materials, the strength of graphite increases with temperature: at 2400 °C, it is approximately twice as strong as at 0 °C. Oxidation resistance above 500 °C in air, however, is poor.

Carbon and graphite can be made in various forms from the almost perfect crystalline structure of pyrolytic graphite to the amorphous structure of glassy carbon and carbon black. Shaped carbon products are made in the following way: a mixture of crushed petroleum coke, coal, or lampblack with a binder, such as pitch or tar, is extruded or pressure-molded to shape and then baked in gas-fired furnaces with a controlled temperature rise from 700 to 1200 °C to drive off volatiles and carbonize the binder. Furnace time varies from a few days to several weeks depending on size. Graphite is made from petroleum coke in a similar way, except that after baking at 950 °C, it is graphitized at 2600 – 2800 °C in resistance furnaces for a total cycle time of 350 – 700 h. All manufactured graphites consist of microcrystalline graphite particles in an amorphous carbon matrix.

The primary use of carbon and graphite (in terms of tonnage) is in electrodes for electric arc furnaces, where the higher electrical conductivity of graphite enables the production of an electrode with half the cross-sectional area of a carbon electrode having the same current-carrying capacity [110]. The other major use in metal processing is in carbon blocks for lining furnaces, including blast furnace hearths. Because they are not wetted by molten metals, graphite molds are used in casting, and graphite jigs are used in brazing.

Graphite is also used extensively in resistance and induction-heated furnaces. The low vapor pressure of graphite allows it to be used as a heating element up to 3000 °C in inert or reducing atmospheres and up to 2500 °C in vacuum. Complex machined shapes are often employed to increase the electrical resistance of the element. Graphite is also used for susceptors in indirect induction heating furnaces. Pyrolytic graphite is a highly anisotropic form of graphite with thermal and mechanical properties that depend strongly on crystal direction. It is produced by chemical vapor deposition from methane; carbon is deposited onto a suitable substrate held at ca. 2000 °C. The carbon grows as a highly oriented graphite structure with the *c* axis of the crystallite normal to the substrate surface. The resultant material has thermal conductivity in the *a* direction ca. 100 times that in the *c* direction (Fig. 68). Thus in the direction perpendicular to the basal plane, pyrolytic graphite has one of the lowest thermal conductivities of all solid insulating materials at high temperature. For this reason, it has been used for rocket nozzle inserts and nose tips in reentry vehicles. Brittle failure problems were encountered with some of these components, and the need for a tougher material was met by carbon – carbon composites, which

were produced fortuitously in the laboratory in 1958 [111].

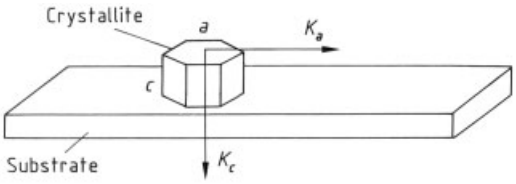

Figure 68. Relationship between thermal conductivity K and crystal direction in pyrolytic graphite ($K_a \approx 100\,K_c$)

Carbon – carbon composites are a logical evolution of carbon fiber-reinforced resin composite systems [112]. Carbon fibers are usually made by heat treating polymer fibers (e.g., polyacrylonitrile) at ca. 250 °C under tension in air to oxidize and stabilize the fibers (see also →Fibers, 5. Synthetic Inorganic, Chap. 5.). They are then heated to 2500 °C under tension to orient the graphite planes along the fiber axis. Fibers thus produced are ca. 10 µm in diameter and are available as tows of large numbers of individual fibers. Process parameter selection allows fiber properties to be varied (Table 22).

Table 22. Properties of various carbon materials

Carbon material	Flexure strength, MPa	Young's modulus, GPa
Fine-grained monolithic graphite	15	7
High-modulus carbon fiber	2000 ∗	400
High-strength carbon fiber	3500 ∗	200
Unidirectional carbon – carbon composite, longitudinal	1250	160
Unidirectional carbon – carbon composite, transverse	10	

∗ Strength in tension.

A typical component is fabricated from carbon fiber and a suitable resin by a technique such as molding or filament winding, then carbonized, and densified. The resin should have a high char yield, resulting in the selection of phenolic or other resins rather than epoxides or polyimides. The carbonization is carried out in an inert atmosphere at 1000 – 1100 °C to pyrolyze the resin and form a carbon residue. The resultant weak and porous carbonized component is then densified either by reimpregnating with a resin and recarbonizing or by chemical vapor deposition at 800 – 1500 °C with a hydrocarbon gas. Several resin impregnation – carbonization cycles are necessary to achieve the densification level required for the specified strength. Further heat treatment may be necessary depending on application, e.g., at 1500 °C to improve oxidation resistance to some extent or at 2500 °C (graphitization) to increase thermal stability and thermal conductivity.

The toughness of carbon – carbon composites is illustrated by the work of fracture of 60 – 125 kJ/m^2 [113], depending on final heat treatment. The elastic properties are similar to those of carbon fiber-reinforced plastic analogues, but the strength is much lower because of the low failure strain of the carbon matrix (typically 0.2. % compared with 2 % for a resin and 0.5 – 1 % for fibers) and the high porosity inherent in carbon – carbon composites. The strength increases with temperature to ca. 2000 °C, and significant strength is retained as high as 2800 °C. The composite material is no more resistant to oxidation than monolithic graphite. This is not a major problem in single-mission applications (e.g., rocket nozzles or cruise missile engines) or when the material experiences only brief expossure to high temperature and design allows a measure of oxygen exclusion (e.g., aircraft brakes).

The major reason for using carbon – carbon brakes is the saving in weight relative to conventional metal brakes. The high specific heat of carbon allows heat developed from the kinetic energy to be absorbed, and the excellent high-temperature strength is needed to cope with the 1200 – 1500 °C temperature developed by extreme braking during an aborted takeoff in advanced military aircraft. Carbon also possesses the friction and wear characteristics required for brake applications.

For more long-term, high-temperature applications, some form of oxidation protection is required [114]. The external surface of the space shuttle orbiter vehicle reaches a temperature of ca. 1500 °C in the nose cone and wing leading edges during reentry [115]; carbon – carbon composite protected with an external coating of silicon carbide is used. The coating is applied by encapsulating the part in silicon carbide, silicon, and alumina powder and heating to ca. 1650 °C. The silicon vapor that forms reacts with carbon to form the SiC coating. The SiC cracks during the coating cycle, because of the large difference

in expansion coefficient between the SiC coating and the substrate, and the final step is to seal the cracks by impregnation with a silicon compound such as tetraethoxysilane [111].

For potential jet engine applications involving a greater degree of thermal cycling and extended dwell times at ca. 1000 °C (in addition to high-temperature excursions >1500 °C), improved oxidation resistance is required, and techniques such as matrix inhibition are being developed. Glass-forming elements are incorporated in the matrix material so that glass forms as cracks are generated, thereby protecting the base material from further oxidation.

11. Ceramic Composite Materials

The initial developmental motivation for ceramic matrix composite materials was similar to that for carbon – carbon composites: the need to increase the toughness of the unreinforced material. Work on carbon – carbon composites in the early 1960s demonstrated that significant toughening can be achieved in brittle matrix materials by incorporating suitable fibers. The strength of the interface between fiber and matrix must be controlled to ensure that sufficient energy is absorbed by debonding at the fiber – matrix interface, by post-debonding friction as the fibers are progressively pulled out, and by unbroken fibers that were circumvented by the crack in the matrix (Fig. 69). Too strong an interface will not deflect cracks and will result in brittle failure.

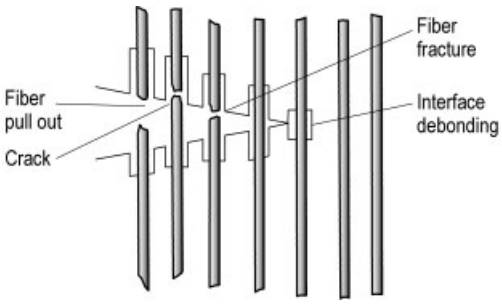

Figure 69. Schematic illustration of energy-absorbing phenomena in a brittle-matrix composite material

In the late 1960s carbon fibers were incorporated in Pyrex glass matrices, and high-strength, tough composites were produced (Fig. 70) [116].

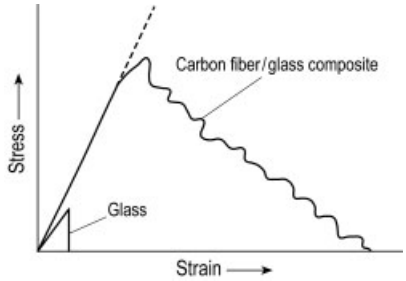

Figure 70. Stress – strain behaviour of unreinforced and reinforced glass

Subsequently carbon fiber reinforced glass ceramics were produced, but the temperature capability was limited by the oxidative instability of the carbon fibers. In the early 1980s, the availability of the organometallic-derived Nicalon SiC fiber changed the situation. The properties of carbon and Nicalon fibers are compared in Table 23. Nicalon fiber has far superior oxidation resistance to carbon fibers, and during the 1980s major programmes were carried out to develop a SiC-reinforced glass ceramic composite capable of operation at 1000 °C and above [117]. The simplest manufacturing route involved impregnating monolayer tapes with a slurry containing glass ceramic powder, a resin binder, and water; stacking the tapes, burning out the resin binder; and densifying the compact by hot pressing. The importance of controlling the fiber – matrix interface was confirmed. A carbon-rich layer some 25 µm thick was formed at the interface during processing. This layer forms a bond of sufficient strength to allow load transfer but allows fiber debonding and pull-out during crack propagation and thus develops good toughness [118]. Combustor components made of Nicalon – glass ceramic composite were run in a military engine programme in 1993.

Table 23. Properties of reinforcing fiber materials

Fiber	Strength, GPa	*E*, GPa	Diameter, µm
Carbon, high-modulus	2.5	490	8
Carbon, high-strength	5	290	8
Nicalon	2.8	190	15

Glass ceramic matrix composites did not have the temperature capability necessary for projected engine designs, however, and materials with higher temperature capability were

necessary. This led to the development of Nicalon – silicon carbide ceramic matrix composites [119]. In the manufacture of these composites, a low-density Nicalon preform is produced. The preform can be created by a variety of processes including filament winding, weaving, and braiding [120]. The fibers are pre-coated with a thin layer of carbon to develop the appropriate interface properties. Chemical vapor infiltration (CVI) of the preform with a gaseous precursor such as a chlorosilane forms the matrix of the composite [121]. The rate of deposition is low, and several infiltration cycles are likely to be required, depending on component sizes, to produce the required composite density. Liquid infiltration with a polymeric precursor followed by pyrolyzation is an alternative to CVI, but again, multiple re-infiltration may be required. Nicalon – SiC composites have good toughness but the temperature capability of the composite is lower than that of monolithic silicon carbide (Fig. 71). A major factor in this is the nonstoichiometry of the Nicalon fibers, which contain oxygen and excess carbon.

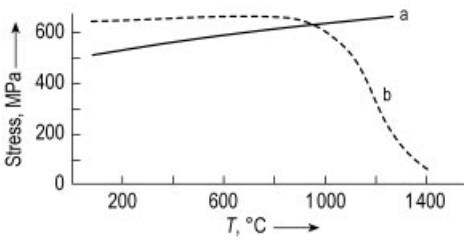

Figure 71. Temperature dependence of strength for silicon carbide materials
a) Sintered silicon carbide; b) Nicalon – silicon carbide composite

Toughness is reduced after high temperature exposure because of the oxidative instability of the carbon layer which controls interface properties. Nevertheless, exhaust nozzle flaps were flight engine tested on military aircraft in 1989 and have been standardized in some applications.

Despite the progress made in the development of ceramic matrix composite technologies, further work is needed in several areas in order to achieve the full potential of these technologies [122, 123]. These have been defined as:

- Fibers with higher temperature capability
- Cheaper manufacturing technology
- Optimization of fiber – matrix interface behavior
- Development of relevant design methodology and life-prediction capability

Considerable work is in progress to develop improved fibers, including oxide-based fibers, single-crystal oxide fibers, and mixed-oxide fibers. The cost of manufacturing processes is being addressed [124]. The Lanxide process involves reaction between a vapor phase and a molten metal alloy, with the reaction product growing through a fiber preform. Large complex shapes have been manufactured in a variety of ceramic systems, including automotive and jet engine components. Work is also in progress to develop stable interfacial coatings for composite systems of various types, including oxide – oxide composites.

12. Intermetallic Materials

Intermetallic materials (see also Intermetallics) consist of phases formed by two or more metals combining in integral ratios. Their crystal structure is usually ordered with each constituent metal occupying a particular sub-lattice. The ordering tends to increase high-temperature strength by inhibiting diffusion and creep deformation. Intermetallic materials have significant potential for high-temperature operation but are usually brittle at ambient temperature because slip systems are limited. Major development activities have hence aimed at improving toughness while retaining or increasing high-temperature capability. The lower density intermetallics such as titanium aluminides are of particular relevance in aerospace applications where minimum weight is essential. Nickel aluminides have considerable potential at temperatures beyond the capability of titanium aluminides, and iron aluminides have been developed as low-cost materials with good corrosion resistance. The various manufacturing techniques appropriate for intermetallic materials are discussed in [130].

Titanium Aluminides. Materials in the Ti – Al system have been the subject of major research since the early 1960s, particularly Ti_3Al and TiAl materials. Development of Ti_3Al has led to the Super-α_2 material which contains

15 % Al, 2 % Nb, 3 % V, and 2 % Mo. Its strength is comparable with that of conventional titanium alloys. It has higher temperature capability, currently limited to 600 – 700 °C by embrittlement due to oxygen penetration [131]. Demonstration parts have been run in jet engines with weight savings of around 40 % compared with nickel superalloy components.

TiAl materials have higher temperature capability than Ti_3Al although protective coatings may be required at temperatures above about 800 °C. A duplex-structure material containing some Ti_3Al phase with additions of chromium and niobium has a useful level of toughness, and components have been produced by casting and by powder manufacturing techniques. Automotive turbocharger rotors and exhaust valves have been trialled, and low-pressure turbine blades have been run in a jet engine and demonstrated significant weight saving relative to conventional nickel superalloy blades [132].

Nickel Aluminides. Ni_3Al was the first material studied in the Ni – Al system, largely because it is the strengthening phase in nickel superalloys and it has inherent oxidation resistance. Chromium additions have increased the high-temperature oxidation resistance, and ambient-temperature toughness and ductility have been improved by the addition of small quantities of boron. Optimum properties are developed in alloys containing some nickel solid-solution phase [133], and materials of this type have been used in diesel engine components.

NiAl materials have lower density and higher oxidation resistance than Ni_3Al and are thus of more potential interest in aerospace applications. High-temperature strength is improved to a useful level by the presence of an Ni_2AlTi second phase, but methods to improve the ambient-temperature brittleness have not yet been developed, and further work is required.

Iron Aluminides. The principal advantages of iron aluminides are oxidation and corrosion resistance that are superior to those of commercial ferritic heat-resistant iron-base alloys together with reduced materials costs. Early aluminides were brittle and had low strength at high temperature. Improved understanding of material behaviour and of the effects of alloying led to the development of strong, ductile Fe_3Al and FeAl materials [134]. Early applications could be in coal-combustion technologies because of the excellent oxidation and corrosion resistance of these materials.

Other Intermetallics. The intermetallic materials discussed above have mostly found some commercial application or have been subject to demonstration programmes which have indicated directions for further work. There are other intermetallics which are more speculative but may have the potential for even higher temperature capability. Molybdenum disilicide has seen commercial usage in heating element applications. An SiO_2 scale provides protection at high temperatures but oxidation is very rapid at lower temperatures around 450 °C. This type of behavior occurs in a range of intermetallic materials and is not fully understood [135]. Further work is in progress to address the ambient-temperature brittleness of $MoSi_2$ and to increase high-temperature strength. Reviews on a wide range of intermetallic compounds with the highest melting points have been carried out [136] in order to provide guidelines for future R & D programmes.

13. Thermal Ablative Materials

Some environments generate temperatures capable of vaporizing most known engineering materials, e.g., hypersonic reentry vehicles can approach 8000 °C. Materials with specific characteristics are required to withstand such temperatures by using *thermal ablation*.

The use of materials as a heat sink is a long established technique that employs the heat content of the material in the solid state. This is typically ca. 10 % of the total heat-absorbing capacity of a material, including melting and vaporization. The total heat capacity is employed in thermal ablation, which is defined as the consumption of solid material by using the total heat capacity and an air flow to continuously remove the products formed. Materials for ablative application should possess a high melting point, high specific heat, high latent heats of fusion and vaporization, and low thermal conductivity. The latter is probably the most important requirement because of the need to confine the high temperature to a thin surface region and minimize

the temperature of the structure being protected. Full ablative potential is achieved only if the material is retained at the surface long enough for it to vaporize; uniform surface removal is essential. These requirements led to the concept of a composite material with the ablative component held in place by fibers that are more refractory. Reinforced plastics are excellent ablative materials and have been used in many aerospace applications, including the Apollo reentry heat shield and Titan 3 C rocket nozzles [125].

Ablation takes place in the following sequence: the surface temperature increases rapidly because of the low thermal conductivity until thermal degradation starts; the gas produced by the volatile components in the plastic forms a boundary layer on the component and provides some insulation; the remaining carbon char is oxidized, sublimed, or eroded. The process is illustrated schematically in Figure 72.

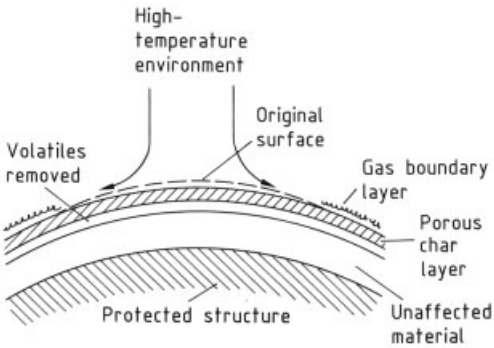

Figure 72. Schematic representation of the ablation process

Phenolic resins have been used most commonly because of the high proportion of char formed. The char yield of a phenolic resin in a nitrogen atmosphere at 900 °C is ca. 47 %, compared with <5 % for an epoxy resin [126]. Glass fibers are often used as reinforcement; constituents other than SiO_2 are removed during ablation so the resultant surface film of viscous SiO_2 can play a useful role if the force of the gas stream is relatively low. A reinforcement level of 40 % is typical; the optimum level depends on the operating environment. Variable ablation rates of reinforced phenolic systems in some rocket nozzle applications have been reported [127]. The more thermally stable carbon – carbon composites may offer improvement because their ablation behavior is superior to that of polycrystalline graphite.

14. Thermal Insulating Materials

Insulating materials of low thermal conductivity restrict the flow of heat from one region to another. They are of two basic types, solid and fibrous.

Solid Insulators. Solid insulators are produced in the form of bricks, blocks etc. They are nonhomogeneous and contain voids, either incidentally or intentionally, that reduce conductivity. Heat transmission in fully dense materials occurs by conduction and is thereby dependent on thermal conductivity; convection through the gas contained in the pores of porous materials is an additional factor. Most solid insulators have approximately isotropic properties, but the thermal conductivity of pyrolytic graphite is very dependent on crystal direction (see Chap. 10).

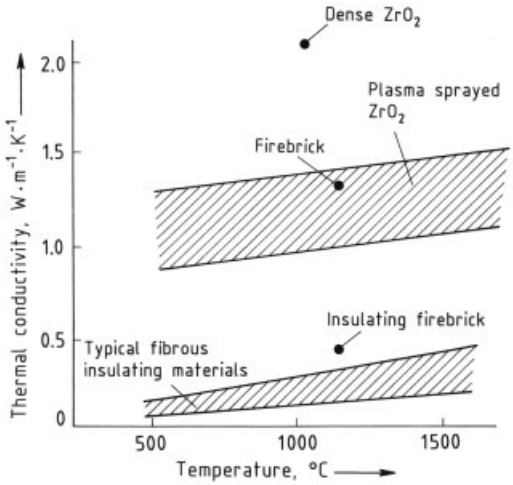

Figure 73. Thermal conductivity of fibrous insulating materials and bulk insulating materials

Fibrous Insulators. Fibrous insulators are made from natural or synthetic fibers with or without chemical binding agents. They are available as batting, cloth, paper, and in a loose form for use as filler. Heat transmission occurs by conduction along the fibers and across regions of fiber contact, convection in the voids, and radiation between fibers. Conductivity is influenced

strongly by bulk density and fiber diameter; the net result is minimum thermal conductivity at a characteristic packed density [128]. Although actual conductivity depends on the factors mentioned, fibrous insulators, as a family, have significantly lower conductivity than solid insulators (Fig. 73, see above). Strength generally increases with bulk density and any chemical fiber bonding that may be used (Fig. 74) [129]. The actual fiber material makes a relatively minor contribution to the overall thermal conductivity but determines the temperature capability of the insulator (Table 24).

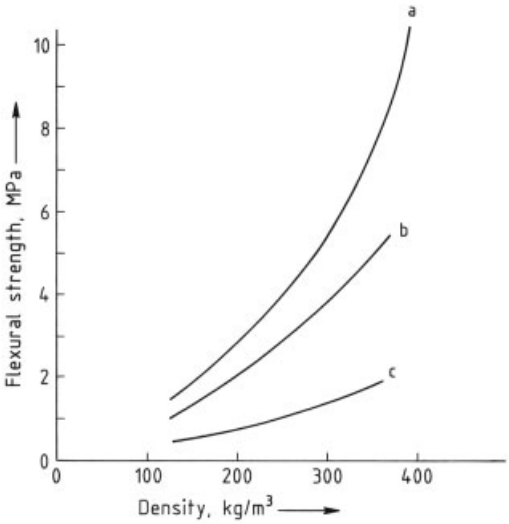

Figure 74. Effect of bulk density on strength of fibrous insulating materials
a) $SiO_2 - Al_2O_3$ fibers bonded with BN; b) SiO_2 – aluminum borosilicate fiber system; c) Rigidized fibrous SiO_2

Table 24. Temperature capability of fibrous insulating materials

Fiber material	Maximum use temperature, °C
SiO_2	1250
$Al_2O_3 \cdot SiO_2$	1350
Al_2O_3	1750
ZrO_2	2200
Carbon	2700

Fibrous insulating materials are used as hot-face linings in heat treatment furnaces, where their lower weight and thermal mass result in faster heating and cooling rates and, consequently, reduced cycle time and fuel savings over conventional refractories. Other applications include combustion chamber and pipeline insulation in the petrochemical industry, and burner components and imitation coals in domestic coal-effect gas fires. Rigidized silica fiber tiles are used on the space shuttle orbiter *Columbia* in which the silica fiber structure supports a glass-coated outer skin that radiates a high proportion of the heat energy back to the atmosphere.

15. References

General References

1. A. H. Cottrell: *The Mechanical Properties of Matter,* Kreiger, Melbourne-Florida 1964.
2. E. Lang: *Coatings for High Temp. Applications,* Applied Science Publ., Barking, United Kingdom, 1983.
3. G. W. Meetham: *The Development of Gas Turbine Materials,* Applied Science Publ., Barking, United Kingdom, 1981.
4. C. T. Sims et al.: *Superalloys II,* J. Wiley & Sons, New York 1987.
5. D. W. Richerson: *Modern Ceramic Engineering,* Marcel Dekker, New York 1982.
6. W. Bettridge: *The Nimonic Alloys,* Edward Arnold, London 1974.
7. I. J. Polmear: *Light Alloys-Metallurgy of the Light Metals,* Edward Arnold, London 1981.

Specific References

8. P. Hancock, J. R. Nichols: *High Temp. Technol.,* Aug. 1982, p. 3.
9. R. Kamo et al.: *NTIS Conference (790747)* 1979.
10. G. W. Meetham: *Proceeding of COST Conference,* Liege Belgium, D. Reidel Publ., Dordrecht 1986, p. 1.
11. W. Hume-Rothery, G. V. Raynor: *Structure of Metals and Alloys,* Institute of Metals, London 1950.
12. E. M. Sherwood in I. E. Campbell (ed.): *High Temperature Technology,* J. Wiley & Sons, New York 1956, p. 17.
13. A. H. Cottrell: *The Mechanical Properties of Matter,* Kreiger, Melbourne-Florida 1964.
14. J. Weertman, *Trans. Am. Soc. Met.* **61** (1968) 681.
15. O. D. Sherby, J. L. Lytton, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **206** (1956) 928.
16. M. F. Ashby, D. R. Jones, *Engineering Materials,* Pergamon Press, Oxford 1985, p. 135.

17. H. L. Bernstein et al. in H. Shitoglu (ed.): *Thermo-Mechanical Fatigue Behaviour of Materials,* ASTM, Philadelphia 1993, p. 212.
18. A. J. B. Cutler et al.: *The Metallurgist & Materials Technologist* **13** (1981) 69.
19. R. F. Decker, J. R. Mihalisin, *Trans. Am. Soc. Met.* **62** (1969) 481.
20. W. I. Mitchell, *Z. Metallkd.* **57** (1966) 586.
21. C. H. White in G. W. Meetham (ed): *The Development of Gas Turbine Materials,* Applied Science Publishers Ltd., Barking, United Kingdom, 1981, p. 89.
22. J. M. Silcock et al., *J. Inst. Met.* **82** (1953 – 1954) 239.
23. I. J. Polmear: *Light Alloys,* Edward Arnold, London 1981.
24. G. W. Lorimer, R. Nicholson: *The Mechanism of Phase Transformations in Crystalline Solids,* Institute of Metals, London 1969.
25. J. Nutting, J. M. Arrowsmith, *Iron Steel Inst. London* **70** (1961) 147.
26. J. Woodhead, A. G. Quarrell, *J. Iron Steel Inst. London* **203** (1965) 605.
27. K. J. Irvine et al., *J. Iron Steel Inst. London* **195** (1960) 386.
28. R. V. Miner, J. Gayda, *Int. J. Fatigue* **6** (1984) no. 3, 189.
29. E. G. Richards, *J. Inst. Met.* **96** (1968) 365.
30. G. W. Meetham, *Met. Technol. (London)* **11** (1984) 414.
31. R. F. Decker, J. W. Freeman, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **218** (1961) 277.
32. J. S. Erickson in P. R. Sahm, M. O. Speidel (eds.): *High Temp. Materials in Gas Turbines,* Elsevier, Amsterdam 1974.
33. J. E. Restall in G. W. Meetham (ed.): *The Development of Gas Turbine Materials,* Applied Science Publ., Barking, United Kingdom, 1981, p. 266.
34. C. Duret, R. Pichoir in E. Lang (ed.): *Coatings for High Temp Applications,* Applied Science Publ., Barking, United Kingdom, 1983, p. 33.
35. F. S. Pettit, G. W. Goward in [34], p. 348.
36. G. M. Acer, G. H. Meier, *Oxid. Met.* **13** (1979) 159.
37. D. P. Whittle, J. Stringer, *Philos. Trans. R. Soc. London Ser. A* **295** (1980) 305.
38. G. W. Meetham, *J. Vac. Sci. Technol. A* **3** (1985) no. 6, 2509.
39. I. Kvernes in E. Lang (ed.): *Coatings for High Temp Applications,* Applied Science Publ., Barking, United Kingdom, 1983, p. 382.
40. S. Robb, *Am. Ceram. Soc. Bull.* **62** (1983) 755.
41. S. M. Meier, D. K. Gupta, *Trans. ASME* **116** (1994) 250.
42. J. T. De Masi-Marcin et al., *J. Eng. Gas Turbines Power* **112** (1990) 521.
43. H. C. Child in E. Lang (ed.): *Coatings for High Temp Applications,* Applied Science Publ., Barking, United Kingdom, 1983, p. 415.
44. K. E. Thelning: *Steel and its Treatment,* Butterworths, London 1975.
45. P. Astley, *Heat Treat. Met.* **2** (1975) 51.
46. B. Edenhofer, *Heat Treat. Met.* **1** (1974) 23.
47. E. C. Kedward et al., *Tribol. Int.* **7** (1974) 221.
48. M. O. Speidel: *Proc 6th Int. Con. on Light Metals Leoben,* Austria Aluminium-Verlag, Düsseldorf 1975. See Polnear, General References, p. 37.
49. H. Y. Hunsicker: *Rosenhain Centenary Conf. on Contribution of Phys. Met to Eng Practice.* The Royal Society, London 1976.
50. R. Irmann, *Metallurgia* **46** (1952) 125.
51. J. R. Pickens, *J. Mat. Sci.* **16** (1981) 1437.
52. W. S. Cebulak, D. J. Truax, US Army Contract DAAA-25-70-Co 358, 1971.
53. P. J. Bridges, Institution of Metallurgists 1601-83-Y, 1983.
54. I. B. MacCormack, *Met. Mater.* **2** (1986) 131.
55. R. W. Gardiner, M. C. McConnell, *Met. Mater.* **3** (1987) 254.
56. M. A. H. Howes, *J. Met.* **38** (1986) 28.
57. E. A. Feest, *Mat. Des.* **7** (1986) 58.
58. R. M. Duncan et al. in G. W. Meetham (ed.): *The Development of Gas Turbine Materials,* Applied Science Publ., Barking, United Kingdom, 1981, p. 63.
59. M. R. Winstone et al., *J. Less Common Met.* **39** (1975) 205.
60. P. H. Morton, *Phil. Trans. R. Soc.* **282** (1976) 401.
61. P. A. Martin et al., Titanium 80 Science &Technology TMS-AIME 1980, p. 1245.
62. T. W. Farthing, *Proc. Inst. Mech. Eng.* **191** (1977) 59.
63. E. F. Bradley, M. J. Donachie, *J. Met.* **22** (1970) 25.
64. P. Blenkinsop: "Materials in Aerospace," *Proc. RAe. Conf.,* London, April 1986.
65. B. Borradaile, R. H. Jeal, *Proc. 4th Int. Conf. on Titanium,* Kyoto 1980.
66. N. F. Harpur, *Metall. Mater. Technol.* **13** (1981) 259.
67. C. R. Austin et al., *Trans. Am. Inst. Min. Metal. Pet. Eng.* **162** (1945) 84.
68. E. W. Colbeck, J. R. Rait, *Iron Steel Inst. London Spec. Report* no. 43, 1952.

69. J. O. Ward, J. R. Rait, *J. West Scotl. Iron Steel Inst.* Jan. 1953, p. 3.
70. R. Blum *in Materials for Advanced Power Engineering,* Kluwer Academic Publishers, Dordrecht1994p. 15.
71. D. A. Oliver et al. in W. E. Benbow: *Steels in Modern Industry,* Iliffe & Sons, London 1951.
72. W. Herda, A. J. Rickard, *Metall. Mater. Technol.* July (1973) 342.
73. G. T. Harris, W. H. Bailey, *Iron Steel Inst. London Spec. Report* no. 43, 1952.
74. G. W. Meetham, *Met. Technol. (London)* **11** (1984) 414.
75. J. L. Smialek, G. H. Meier in C. T. Sims et al. (eds.): *Superalloys II,* J. Wiley & Sons, New York 1987, p. 293.
76. C. Wells in G. W. Meetham (ed.): *The Development of Gas Turbine Materials,* Applied Science Publ. Barking, United Kingdom, 1981, p. 207.
77. D. A. Ford, in [76], p. 147.
78. P. S. Kotval et al., *Metall. Trans.* **3** (1972) 453.
79. S. K. W. Shaw, *Met. Prog.* **115** (1979) 47.
80. G. J. S. Higginbotham, *Mater. Sci. Technol.* **2** (1986) 442.
81. M. J. Goulette et al., "Superalloys 1984", *Trans. Metall. Soc. AIME* (1984) 117.
82. G. L. Erickson, *JOM* **47** (1995) no. 4, 36.
83. M. McLean: *Directionally Solidified Materials for High Temp Service,* The Metals Society, London 1983.
84. A. M. Beltran in C. T. Sims et al. (eds.): *Superalloys II,* J. Wiley & Sons, New York 1987, p. 135.
85. G. H. Gessinger, M. J. Bomford, *Int. Met. Rev.* **19** (1974) no. 181, 51.
86. J. S. Benjamin, *Metall. Trans.* **1** (1970) 2943.
87. L. W. Lherbier in C. T. Sims et al. (eds.): *Superalloys II,* J. Wiley & Sons, New York 1987, p. 387.
88. F. Turner in G. W. Meetham (ed.): *The Development of Gas Turbine Materials,* Applied Science Publ., Barking, United Kingdom, 1981, p. 177.
89. G. W. Meetham in D. Coutsouradis et al. (eds.): *High Temperature Alloys for Gas Turbines,* Applied Science Publ., Barking, United Kingdom, 1978, p. 837.
90. D. M. Macdonald: "Frontiers of High Temperature Materials," *IncoMAP Conf.,* New York 1981.
91. P. B. Wallis in W. Betteridge, J. Heslop (eds.): *The Nimonic Alloys,* Edward Arnold, London 1974, p. 459.
92. J. E. Restall, *Met. Mater.* **1** (1967) 241.
93. D. J. Jones, *Metall. Mater. Technol.* **5** (1973) 503.
94. W. Fairhurst, *Metall. Mater. Technol.* **6** (1974) 68.
95. F. H. Norton: *Refractories,* McGraw-Hill, New York 1968, p. 237.
96. D. W. Richerson: *Modern Ceramic Engineering,* Marcel Dekker, New York 1982.
97. J. D. Birchall et al. in R. W. Davidge (ed.): *Engineering with Ceramics,* British Ceramic Society, Stoke-on-Trent 1982.
98. J. R. Maclaren et al., *Proc. Br. Ceram. Soc.* **26** (1998) 67.
99. P. Stanley et al., *J. Strain. Anal. Eng. Des.* **April** (1978) 43/88/78.
100. J. E. Ritter et al., *J. Mat. Sci.* **15** (1980) 2275.
101. F. Lange, *Am. Ceram. Soc. Bull.* **62** (1983) 1369.
102. K. Upadhya, *JOM* (1992) 15.
103. B. E. Waye: *Intro. to Tech. Ceramics,* Maclaren & Sons Ltd., London 1967, p. 105.
104. R. C. Garvie et al., *Nature (London)* **258** (1975) 703.
105. N. Claussen, M. Rühle in A. H. Heuer, L. W. Hobbs (eds.): *Advances in Ceramics,* **vol. 3,** The Am. Ceram. Soc. Columbus, OH 1981, p. 137.
106. R. Stevens: *An Introduction to Zirconia,* Magnesium Elektron Ltd. Twickenham, United Kingdom 1981.
107. M. H. Lewis et al.: *Pub. Conf. on Non-Oxide Tech. Ceramics (Limerik),* Elsevier, Barking, United Kingdom, 1985.
108. M. H. Lewis, *Spectrum* **179** (1982) 10.
109. *Kirk-Othmer,* **3,** 84.
110. M. S. Wright in I. E. Campbell (ed.): *High Temp Technology,* J. Wiley & Sons, New York 1956, p. 2.
111. J. H. Brahney, *Aerosp. Eng.* **7** (1987) 11.
112. J. D. Buckley, *Am. Ceram. Soc. Bull.* **67** (1988) no. 2, 364.
113. C. R. Thomas, E. J. Walker: "Materials in Aero-space," *RAe Soc. Conf.* , London 1986.
114. J. R. Strife, J. E. Sheehan, *Am. Ceram. Bull.* **67** (1988) no. 2, 369.
115. D. M. Curry et al., *Am. Inst. Aeronaut. Astronaut.* 83-0483.
116. R. A. J. Sambell, *J. Mater. Sci.* **7** (1972) 676.
117. K. M. Prewo et al., *Ceram. Bull.* **65** (1986) no. 2, 305.
118. R. J. Kerans et al., *Ceram. Bull.* **68** (1989) no. 2429.
119. P. J. Lamicq et al., *Ceram. Bull.* **65** (1986) no. 2336.

120. F. K. Ko, *Ceram. Bull.* **68** (1989) no. 2401.
121. W. J. Lackey, *Ceram. Eng. Sci. Proc.* **10** (1989) no. 7/8, 577.
122. E. L. Courtright, *Ceram. Eng. Sci. Proc.* **12** (1991) no. 9/10, 1725.
123. N. M. Tallon, *Ceram. Eng. Sci. Proc.* **12** (1991) no. 7/8, 975.
124. J. R. Strife, *Ceram. Eng. Sci. Proc.* **11** (1990) no. 7/8, 871.
125. M. A. Schwartz, *Space/Aeronaut* **52** (1969) 95.
126. G. P. Peterson, D. L. Schmidt in I. E. Campbell, E. M. Sherwood (eds.): *High Temp Materials & Technology,* J. Wiley & Sons, New York 1969, p. 526.
127. J. J. Choury, *L'Aeronautique et L'Astronautique* **68** (1981) 30.
128. G. H. Kesler in I. E. Campbell, E. M. Sherwood (eds.): *High Temperature Materials & Technology,* J. Wiley & Sons, New York 1969, p. 693.
129. J. F. Creedon et al., *Mat. Eng. Cleveland* **101** (1985) 57.
130. E. A. Feest, J. H. Tweed, *Mat. Sci. Tech.* **8** (1992) 308.
131. S. J. Balsone: in T. Grobstein, I. Doychak (eds.): *Oxidation of High Temperature Intermetallics,* TMS 1989, p. 219.
132. P. Bertolotta et al., *JOM* **49** (1997) no. 5, 48.
133. C. T. Liu, K. S. Kumar, *JOM* **45** (1993) no. 5, 38.
134. C. G. McKamey et al., *J. Mater. Res.* **6** (1991) 1779.
135. G. H. Meier, F. S. Petit, *Mat. Sci. Eng.* **A153** (1992) 548.
136. P. R. Subramanian et al., *JOM* **48** (1996) no. 1.

High-Temperature Plastics → Polymers, High-Temperature
High-Tenacity Fibers → High-Performance Fibers

Holography

NICHOLAS GEORGE, University of Rochester, Institute of Optics, Rochester, NY 14627, United States

THOMAS STONE, University of Rochester, Institute of Optics, Rochester, NY 14627, United States

1. Introduction 1
2. Operating Principles of Holography 1
2.1. Recording Technique 4
2.2. Diffraction Efficiency 5
3. Materials Used in Holography . . . 6
3.1. Photographic Materials 6
3.2. Dichromated Gelatin 6
3.3. Photopolymers 7
3.4. Photoresist 7
3.5. Embossed Films 7
3.6. Other Materials 8
4. Uses in Chemical and Related Industries 8
4.1. Holography in Chemistry 8
4.1.1. Study of Chemical Reactions 8
4.1.2. Holographic Microscopy 8
4.1.3. Interferometry and Non-Destructive Testing 9
4.1.4. Other Uses in Chemistry 9
4.2. Some Other Uses of Holography . . 9
4.2.1. Holographic Optical Elements 9
4.2.2. Information Processing 9
5. References 10

1. Introduction [1–6]

Holography is the field of optical science in which *wavefronts* are recorded, stored, and recovered. Holograms are often used for virtually perfect three-dimensional lensless photographs and they also have a number of important technical applications such as diffractive optical elements, holographic filters for optical processing, and optical memories.

Holography can be described, in general, as the technique of using diffraction to transform one wavefront into a second in a prescribed manner. These transformations include, for example, sunlight into a three-dimensional scene (an example of pictorial holography), or an expanding wave into a converging wave and image (an example of a holographic optical element). A *hologram* is generally a two-dimensional mask which contains structures with details on the order of the wavelength of the wavefront being transformed. The technique of holography may be applied in the entire spectrum of *electromagnetic waves* including visible, infrared, ultraviolet, and X-ray regions. *Nonelectromagnetic waves* may also be transformed by holography, e.g., acoustic waves and matter waves (such as electrons). The hologram effects desired wavefront transformations by imposing spatially varying amplitude and/or phase modulations on the incident wavefront. The modulated incident wavefront is thereby diffracted into the desired wavefront as it propagates through and from the hologram region.

2. Operating Principles of Holography

The recording and viewing of a *two-beam transmission* hologram is perhaps the most direct way to give an understanding of holography. In Figure 1 (see top of next page) the reflected wavefront (W) from the pyramidal object (P) is recorded at the hologram plane (H) using a reference beam (R) from the same laser (L). The recording film medium (H) is developed and returned to its original position. Viewing of the reconstructed wavefront (W) is accomplished by illuminating the developed medium (H) using the same illumination beam (R) as used in the recording; however now, the object (P) is removed. The observer (E) looking toward (H) as in Figure 2 sees a perfect three-dimensional reconstruction of the pyramid (P) by virtue of the reconstructed wavefront.

Several important types of holograms have been devised, including Gabor's single beam, Leith and Upatnlek's two-beam, Denisyuk's reflection type, image plane, Fourier, and Fraunhofer holograms [1–3]. The "holographer's rule"

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_079

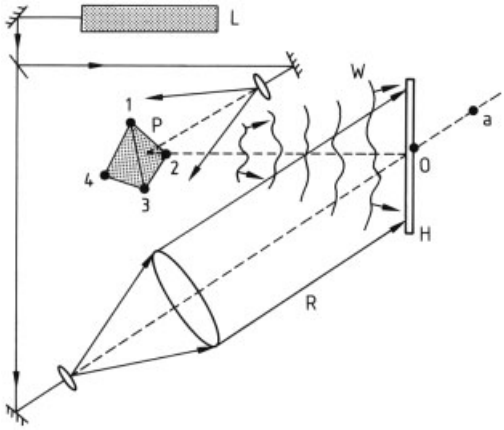

Figure 1. Layout for recording a hologram

is a valuable aid to gain insight into the reconstruction of any of these types. Consider a large number of plane waves represented by rays (1 – 4), incident on the holographic medium (H) in Figure 3. After processing, the plate is returned to its original position and any one of the *original beams* can be used for the reconstruction. The rule is that if any beam, e.g., ray (3), is used as the reference beam (3-R in Fig. 4), all other beams present in the "taking" will be reconstructed as shown, together with some spurious rays (see Section 2.2) which can often be minimized in practice. These reconstructed primary beams will be traveling in the same direction as they were in the original scene.

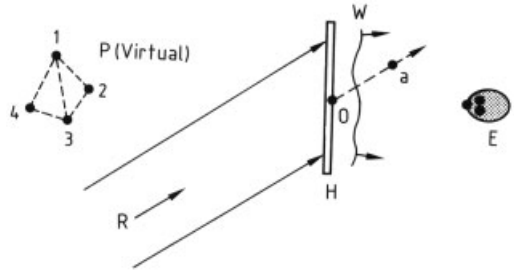

Figure 2. Hologram (H) is illuminated by a reference wave (R) for viewing by observer (E) who sees a perfect three-dimensional reconstruction of the pyramid in its original position (virtual)

A hologram can also be illuminated with a *reversed reference beam* (3-R$^-$) as shown in Figure 5 (see next page). In general this gives rise to what is termed the *conjugate wavefront* W$*$ which is the recorded wavefront but traveling back on itself. The holographer's rule for this reconstruction is shown in Figure 5.

If the same hologram as was made in Figure 1 is illuminated with a reversed reference beam, ray (R$^-$) (see Fig. 6) the conjugate wave W$*$ is set up propagating in space back toward the former position of the pyramidal object. An observer (E_2), looking into H, sees points of the pyramid in their true positions relative to the original, however the visual perspective is reversed. This reversal causes a false stereo perspective which is termed pseudoscopic; and the observer located at E_2 in Figure 6 perceives a

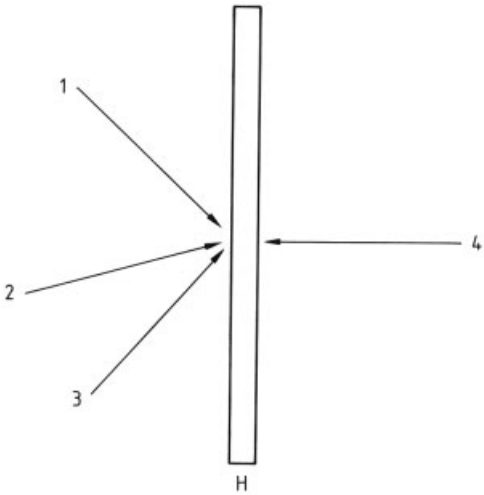

Figure 3. Recording of separate beams (1) through (4)

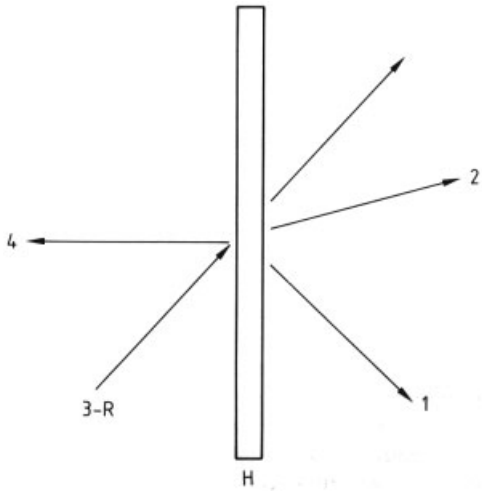

Figure 4. Viewing of the hologram from Figure 3 with beam (3) as the illumination (3-R)

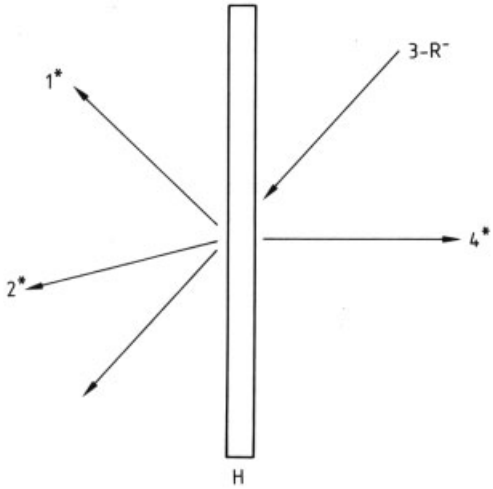

Figure 5. Viewing of the conjugate reconstruction of the hologram from Figure 3 with the reference illumination provided by using a reversed beam 3 (3-R^-)

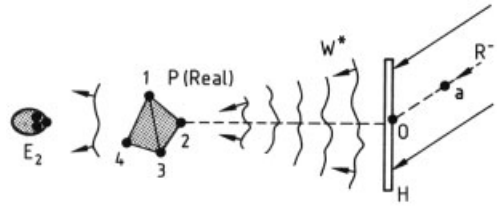

Figure 6. Viewing of the pseudoscopic image from the hologram (H) of Figure 1 using a reversed illumination R^-

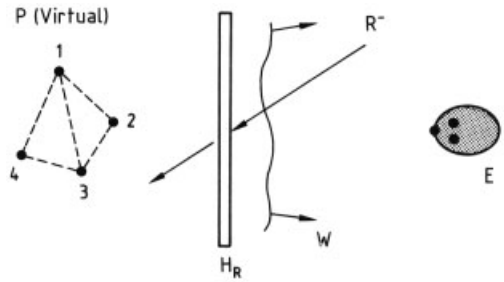

Figure 7. Viewing of a reflection type hologram (H_R) in which the illumination (R^-) and the reflected reconstructed wavefront (W) are on the same side of the hologram as the observer (E)

left-to-right reversal, e.g., of points (2) and (3), as well as a depth inversion, i.e., (4) is closer to the observer than (2). If the subject had been a person's face, this pseudoscopic real image has the perspective of a "death mask."

A *reflection-type hologram* is made by bringing the recording reference beam in from the opposite side of the plate, e.g., along ray (R^-) in Figure 7. Reconstruction is as shown in Figure 7 with the observer viewing the wavefront W reflected from the hologram (H_R). For this type of hologram, the basic interference structure is a multilayer, and it is highly efficient in reflecting W only over a relatively narrow band of wavelengths. Hence, in practice, the reflection-type hologram can be viewed with incident white-light illumination, because only a narrow band is reflected and the other wavelengths are transmitted. Viewing is often improved somewhat if the hologram H_R is backed with a black, light absorbing paper or paint.

The diffracting structure for a hologram can be formed interferometrically, as described above, or it can be calculated and written fringe-by-fringe using modern digital techniques. These computer-generated holograms are made typically in two or more step: (1) theoretical analysis of the diffracting structure corresponding to a given object, (2) printout of the initial hologram pattern to a scale set by the plotter, and (3) photoreduction to a convenient scale for illumination by a visible optical beam. Theory and performance for this important type of hologram are given [2–4, 6–8].

2.1. Recording Technique

Consideration of a few practical matters and of tradeoffs often encountered in holography is useful in understanding holograms and their characteristics. Perhaps the most common method of making a hologram is to record the interference of two or more coherent waves in a photographic film of high resolution. This is illustrated schematically in Figure 1, where the coherence between the reference and object waves is obtained by splitting them from the same laser beam. For light from a laser (→Lasers), the coherence requirement which insures good contrast in the interference fringes is that the maximum difference in path lengths between the reference and the object beams after they are split should be a small fraction of the coherence length of the light. The coherence length is given by $c/\Delta\nu$ where c is the speed of light and $\Delta\nu$ is the spectral bandwidth of the light. The coherence length ranges from the order of a

centimeter for a multi-longitudinal mode argon-ion laser to about half a meter for a helium-neon laser and to thousands of meters for a gas laser running in a single longitudinal mode, which is commonly induced by an intracavity etalon in the laser.

Exposure times vary widely with the type of laser used and the recording material, but are often in the range from a fraction of a second to several minutes. Care must be taken to assure that the fringe pattern is stationary for the duration of the exposure. Because a fluctuation as small as half a wavelength in one of the beam paths, e.g., as caused by a vibrating mirror, can wash out the recorded fringe pattern, sources such as building vibrations, acoustic noise, and thermal variations must be controlled. It is important in making holograms to have a very stable table, commonly pneumatically isolated tables are used. Another very useful technique to minimize deleterious effects of such motions is to use a fringe-locking system which may include, for example, an interferometer and electronics which drive a galvonometer or piezo-electric mirror mount to counter fringe shifts detected by the interferometer.

In many technical applications of holography, the required diffracting structure can be computed; or it is otherwise known digitally, and relief forms of holograms are appropriate. Photolithographic methods employing photoresists (→Imaging Technology, Chap. 3.4.1.) are used with combinations of direct exposure, electron-beam writing, or ion-beam milling. The techniques used are similar to the manufacture of binary optics or of compact video and audio discs (→Information Storage Materials). A relief image is prepared on a master photolithographically; and thereafter, embossed holograms can be mass produced using pressure-sensitive molding.

2.2. Diffraction Efficiency [1–6, 9, 10]

The efficiency with which a given type of hologram diffracts an incident wavefront into a desired one depends on many parameters and tradeoffs. Holograms that cause diffraction by a spatially varying *absorption* are generally less efficient than those utilizing *lossless phase variations* to diffract the incident field. Whether a hologram is lossy or lossless depends on the choice of the recording medium and the processing (see Chap. 3). In addition to possible absorption in the holographic medium, much energy may be lost to spurious diffracted orders and diffuse scattering in the recording material. Spurious diffraction refers to wavefronts generated by the hologram above and beyond the idealized original recording wavefronts described in Chapter 2. Further, holograms may be grouped into *thin* and *thick* categories depending roughly on whether the modulation spatial period is much less than or much greater than the actual thickness of the diffracting structure. Thin sinusoidally modulated holograms tend to exhibit diffraction efficiencies below 6.25 % and 33.9 % for absorption and lossless media, respectively [2]. Shaping or "blazing" the diffracting structure can greatly increase the diffraction efficiency of thin lossless holograms, to levels approaching unity.

The diffraction efficiency of lossless holograms can also be enhanced to nearly 100 % by utilizing *volume diffraction*, i.e., the thick hologram regime. In volume holograms, the energy diffracted into spurious orders may be dephased with the result that essentially all the incident energy may be diffracted into a desired order. This ideal limiting efficiency may be somewhat reduced for the case of a complicated wavefront as originating, for example, from a diffuse extended object. Finally, there are many tradeoffs that arise in attempting to optimize the efficiency of a hologram. For the volume hologram example, 100 % peak diffraction efficiency may be approached for a given combination of spatial period, modulation strength, fringe tilt angle, emulsion thickness, incident angle and wavelength. However, as any of these parameters are varied from this optimum the efficiency usually decreases. Generally, the thicker the element is, the more rapidly the efficiency drops with any such parameter detunings. In many applications (e.g., a holographic lens), it would be desirable to maintain high efficiency over large wavelength shifts and/or incident angle shifts. Decreasing the element thickness results in broader spectral and angular bandwidths, but also in lower peak efficiencies. Compromises of this type can result in elements that exhibit peak efficiencies >98 % and spectral bandwidths on the order of 400 nm centered in the visible.

3. Materials Used in Holography
[1–6, 11]

Recording materials often limit the quality and performance of holograms, and are therefore of central importance. Several basic requirements of a photosensitive material for the interferometric recording of holograms include (1) a mechanism for mapping the interference pattern during exposure into an amplitude and/or phase variation in the reconstructing waves; (2) spectral sensitivity to the wavelength(s) used in recording; and (3) spatial resolution capable of following variations in the interference pattern, which are typically on the order of the wavelength used. As the angle between the reference and object beam increases, the spatial period of the fringe structure decreases (see Section 4.2.1). For 30°, 60°, and 180° beam separations, the fringe spacings setting the resolutions are 2 λ, λ, and $\lambda/2$, respectively. Hence, a resolution of >1000 lines per millimeter is often required for optical holography. Phase modulations are imparted to the reconstructing wavefront by varying the refractive index of the material, the thickness or height of the material, or more commonly, both simultaneously. It is also desirable to minimize diffuse scatter in the material both during exposure and reconstruction. Brief descriptions of a number of holographic recording materials are given below.

3.1. Photographic Materials
[12–16](→Photography)

Photographic emulsions have been useful as a recording medium since the beginning of holography. When used as an *amplitude-modulating material*, the silver-halide emulsion may be developed conventionally giving a lossy silver structure with optical density which is a mapping of the original exposure. Thick lossy holograms are limited theoretically in diffraction efficiency to 3.7 % [2], however, this is adequate for many applications. If more efficiency is useful, bleached holograms are appropriate. One means for obtaining bleached holograms is simply to use a reversal bleaching method in which the scattering centers are silver bromide instead of lossy silver [6, 16]. Earlier in bleaching methods it was common to convert the lossy silver into silver bromide. Efficiencies for the bleached hologram can reach 100 % theoretically. In practice, effects such as tanning of the gelatin during processing tend to cause surface height modulations which add a phase modulation component to both the lossy and refractive index mechanisms. Bleached holograms can exhibit higher diffraction efficiencies than their lossy counterparts, but noise (diffuse scatter) is also generally increased. To achieve the high resolutions required, the silver-halide microcrystals suspended in the gelatin of holographic emulsions must be made much smaller than in common consumer pictorial films. This results in a photographic speed of holographic films which is typically hundreds of times slower than in the latter case. Still, the speed of photographic emulsions used in holography is many times faster than in many other holographic recording materials.

3.2. Dichromated Gelatin [17–22]

The observation that the dichromate ion renders hardness and solubility of gelatin light sensitive was known in the 1800s, when dichromated gelatin was used as a resist. Dichromated gelatin is a good volume holographic recording material and many processing schedules are available [20, 21]. Whereas the theory for diffraction by phase modulating structures is well understood, the exact mechanism of volume refractive index modulation in dichromated gelatin is not fully understood. The recording mechanism is apparently dependent on the production of trivalent chromium ions from optical excitation, and the subsequent tanning (cross-linking) effect of these ions. Volume holograms recorded in dichromated gelatin exhibit diffraction efficiencies approaching the theoretical limit (which can attain 100 % in idealized cases) with low scatter noise. With dichromated gelatin, the relative humidity must be controlled during exposure and processing. The finished hologram should also be sealed to prevent deterioration. Dichromated gelatin is particularly sensitive in the blue and ultraviolet spectral regions, with essentially no sensitivity in the red. The spatial resolution of dichromated gelatin exceeds 4000 lines per millimeter. A hybrid technique is sometimes used

to increase the speed and extend the spectral sensitivity, wherein a photographic emulsion is exposed but processed so as to produce a pure gelatin medium [22].

3.3. Photopolymers [23–29] (→Imaging Technology)

Photosensitive mechanisms have been used to initiate the polymerization of a monomer, or to further link (or destroy links in) existing polymer chains. Many such processes form the basis of very useful volume phase holographic recording materials. In a broad sense dichromated gelatin (see Section 3.2) and photoresists (see Section 3.4) fall into the photopolymer category because polymeric bonds are formed or broken with exposure, giving rise to changes in solubility and/or refractive index.

An example of a photopolymer system for holographic recording is metal acrylate monomer which undergoes free radical polymerization initiated by 4-toluene sulfinic acid, which is sensitized to red light by methylene blue [25]. Advances by groups including Polaroid and DuPont have given rise to photopolymeric recording materials for holography that are capable of achieving near theoretical limits of diffraction efficiency, as well as low scatter and insensitivity to moisture. Photopolymers have been made sensitive throughout the visible range and some have exhibited spatial resolutions well in excess of $5000\,\mathrm{mm}^{-1}$. In the Polaroid material DMP-128, the effective modulation in the refractive index can be as high as 0.2 and the scattering level in the visible range is very small because the grain size in the recording is $<0.1\,\mu\mathrm{m}$. Further, these materials are roughly an order of magnitude more sensitive than dichromated gelatin.

3.4. Photoresist [29–31] (→Imaging Technology, Chap. 4.1.)

Photoresists are materials, typically organic polymers, which exhibit a marked variation in solubility with actinic exposure. Examples include PMMA, poly(methyl methacrylate), for electron exposure and Shipley AZ-1300 series resists for short-wave visible and near ultraviolet exposure. When exposure causes a decrease in solubility, e.g., by formation of additional polymeric bonds or cross-links, the photoresist is termed negative. Resists whose solubility in an appropriate solvent or developer is increased after exposure are termed positive.

In normal application, these materials are exposed and then the soluble regions removed to leave a surface-height variation which is related to the exposure. This surface relief may be used directly as a phase modulator, or it can be overcoated or etched (using chemicals, reactive ions, etc.) into the substrate. Photoresists readily reproduce sharp submicron details, and the technology of this material has been extensively developed by the electronic photolithography industry. Photoresists are important materials in holography because of their ability to map holographic exposures into surface relief. This is valuable for creating embossing masters (see Section 3.5) as well as in fabricating blazed reflection and transmission gratings and computer-generated holograms.

3.5. Embossed Films [32–35]

Mass replication of holograms is commonly accomplished through the embossing of clear or reflective films. In a typical embossing process, a master hologram is recorded in photoresist which, upon development, exhibits a surface relief mapped from the intensity distribution of the holographic interference pattern. The photoresist master is then copied onto a metal stamping master. This can be accomplished by coating the photoresist with a conductive film and then electroforming additional metal (e.g., nickel). This stamping master is then heated and pressed in contact with a thermoplastic material (e.g., polyvinyl or polyester) which is then separated from the stamping master and contains a replication of the original hologram structure. Injection molding is a common alternative. Thousands of impressions may be made from a given stamping master, and the resultant embossed plastic holograms may be reconstructed in transmission or reflection. In the latter case, the plastic is usually aluminized to enhance diffraction efficiency. Advances include embossed copies of computer-generated holographic elements with

shaped fringe structure resulting in high diffraction efficiencies.

3.6. Other Materials

In a general sense, any photosensitive medium or process capable of providing a spatially modulated refractive index or absorption may be considered for use as a holographic recording material. A few of the important materials not mentioned above are briefly discussed here.

Thermoplastic recording materials [36,37] utilize the reversible heat-softening cycle to impart thickness variations in the plastic film. For example, a photoconductor may be used to convert the optical holographic exposure into a spatially modulated electric field. Electrostatic forces from this field are then used to modulate the thickness of the softened film which, upon cooling, provides a stable lossless hologram. Spatial resolutions exceeding 1000 mm^{-1} have been obtained.

Photorefractive materials [38–40] such as bismuth silicon oxide, barium titanate, and PLZT ($Pb(Zr,Ti)O_3$) ceramics have been extensively studied as holographic materials, and also in the closely related field of four-wave mixing. One of the many possible mechanisms of refractive index modulation in these materials is the photo-induced release of electrons which drift (or are driven with an applied electric field) to a neighboring trap, where the resultant space-charge induces a local change in refractive index.

4. Uses in Chemical and Related Industries

Holography is a useful tool in a broad range of disciplines including optical elements, metrology, physics, art, medicine, and chemistry, with many of these applications given in general references [1–6,41,42]. Several applications related to the field of chemistry are described below, followed by a coarse sampling of other application areas of holography.

4.1. Holography in Chemistry

4.1.1. Study of Chemical Reactions

Holography has been used extensively to study reaction rates and diffusion parameters in chemical reactions [43–47]. In many such techniques, a volume holographic grating is formed in a test medium utilizing photosensitive chemical reactions, photochromic tracers, or other light-sensitive mechanisms. The holographic illumination usually consists of a sinusoidal intensity pattern, with a variable spatial period typically on the order of a micron, created by the superposition of two plane waves. As the photosensitive reaction proceeds, the refractive index, thickness, and/or absorption of the test medium is modulated and a probe beam is used to measure the diffraction efficiency of the induced grating, thus providing a measure of the rate of reaction. Many variations on this general technique exist including " holographic relaxation spectroscopy," where a volume grating is induced in mobile species (which are either photosensitive or tagged with a photosensitive tracer) in the test medium. When the driving interference pattern is removed, the induced grating decays as the mobile species diffuse; and the decay of the diffraction efficiency, as measured with a probe beam, is used to determine the diffusion rates of the given species.

4.1.2. Holographic Microscopy

Holography was invented by Gabor in the context of improving electron microscopy [48], and has since proven useful in a range of microscopic applications of interest to chemists [49–52]. By recording wavefronts in a microscope holographically, volumes spanning many depths of focus may be investigated without changing conditions, or alternatively, wavefronts may be superposed revealing interferometrically the effects of changes in the observed scene. In-line holograms of chemical fibers have been formed by recording the interference pattern of X-rays with wavelengths of 830 pm and reconstructing the hologram with visible light. As sources of X-rays achieve greater coherence, the ability to examine molecular structures holographically will be greatly enhanced. In the visible spectral region, the usefulness of holographic

interference microscopy for chemical investigations was demonstrated with interferometric investigations of a variety of organic and inorganic crystals [52].

4.1.3. Interferometry and Non-Destructive Testing [42,53–56].

The ability to store and reconstruct a perfect three-dimensional image makes holography very useful in interferometry and testing. For example, a hologram of a test object may be used to reconstruct an image which is exactly coincident with the object itself. Then, under laser illumination, any deformations in the object that are induced by change in temperature, pressure, stress, etc., result in interference fringes which can reveal the magnitude and direction of the displacement. This technique is sensitive to displacements on the order of the wavelength used, and may also be implemented with greatly reduced sensitivity to displacement if desired. Typical applications include the study of curing adhesives, measuring Poisson's ratio for materials, and examination of chemical bonding between concrete and smooth steel reinforcements.

4.1.4. Other Uses in Chemistry [57,58]

Holographic techniques may be applied in a broad variety of chemical studies. The combustion in engines and flames may be investigated using holography to record transient phenomena or as an interferometric tool. The versatility of holography as an analytic tool for molecular systems is demonstrated by techniques where holographic gratings are rapidly formed using interfering plane waves of 10^{-12} s duration. These transient gratings excite phonons because of electrostriction or absorptive heating, permitting the measurement of optoelastic constants, absorption, and acoustic attenuation. This type of " coherent optoacoustic spectroscopy" is used to measure the absorption spectrum of the fifth overtone transition of a C-H stretch in benzene [58].

4.2. Some Other Uses of Holography

4.2.1. Holographic Optical Elements

Holograms can perform the functions of basic optical elements, e.g., holographic diffraction gratings, holographic multilayer mirrors, diffractive lenses, chirped holograms as scanners, etc. [3,6,59]. Each type of holographic optical element has unique characteristics which often distinguish it from the more conventional counterpart. For an example of holographic optical elements, holographic diffraction gratings are discussed below.

Holography has had a strong impact on the production of diffraction gratings. First, the conventional "ruling engine" approach to scribing diffraction gratings is often replaced by a holographic exposure of photoresist and subsequent etching of the grating profile into the substrate (or direct metallic overcoating of the photoresist). The holographic exposure may consist of the interference pattern produced between two plane waves, where the period Λ along the surface of the grating is controlled by the exposing wavelength λ and the angles, θ_1 and θ_2, which the plane waves subtend with respect to the surface normal. These parameters are governed by the relation:

$$L = \frac{\lambda}{\sin(\theta_1) - \sin(\theta_2)}.$$

Diffraction gratings produced in this manner are free from "ghosts" caused by systematic ruling errors, and they can be fabricated rapidly with a large aperture. Further, holographic diffraction gratings can be recorded throughout a volume recording material rather than as a surface modulation. Such volume holograms can be recorded as transmission or multilayer reflection gratings. In the former case, they can exhibit peak diffraction efficiencies near unity and spectral bandwidths of 400 nm [60]. Holographic multilayers can be used as notch filters with bandwidths on the order of 10 nm and efficiencies of 99.99 %.

4.2.2. Information Processing

Special holographic filters are used in the field of optical information processing. The basic reason for their usefulness is that holograms can be made in the form of transmission masks that

have controllable phase delay and absorption. One of the more commonly used such masks is the holographic matched filter [2,3,6,61].

Another important technical application of holography is the field of information storage. Motion picture films, books, journals, and census data have all been successfully converted to a holographic format that provides a highly compact storage medium of archival quality [2,61].

5. References

General References

1. H. M. Smith: *Principles of Holography,* Wiley, New York 1969.
2. R. J. Collier, C. B. Burckhardt, L. H. Lin: *Optical Holography,* Academic Press, New York 1971.
3. H. J. Caulfield (ed.): *Handbook of Optical Holography,* Academic Press, New York 1979.
4. L. M. Soroko: *Holography and Coherent Optics,* transl. by A. Tybulewicz, Plenum, New York 1980.
5. N. Abramson: *The Making and Evaluation of Holograms,* Academic Press, London 1981.
6. P. Hariharan: *Optical Holography,* Cambridge University Press, Cambridge 1987.

Specific References

7. L. P. Yaroslavskii, N. S. Merzlyakov: *Methods of Digital Holography,* transl. by D. Parsons, Consultants Bureau Plenum, New York 1980.
8. W. H. Lee, "Computer Generated Holograms: Techniques and Applications" in E. Wolf, (ed.): *Progress in Optics,* vol. XVI, North Holland, Amsterdam 1978.
9. L. Solymar, D. J. Cooke: *Volume Holography and Volume Gratings,* Academic Press, London 1981.
10. R. Petit: *Electromagnetic Theory of Gratings,* Springer Verlag, Berlin 1980.
11. H. M. Smith: *Holographic Recording Materials,* Springer Verlag, Berlin 1977.
12. Eastman Kodak Company, "Characteristics of Kodak Plates for Scientific and Technical Applications," Publication P-140, Rochester 1981.
13. Agfa-Gevaert Inc., "Photographic Materials for Holography," Mortsel 1972.
14. Illford, Inc., "Holographic Film," Technical information, London 1986.
15. R. L. Lamberts, "Characterization of a Bleached Photographic Material," *Appl. Opt.* **11** (1972) no. 1, 33.
16. M. Chang, N. George, "Holographic Dielectric Grating: Theory and Practice," *Appl. Opt.* **9** (1970) no. 3, 713.
17. L. H. Lin, "Hologram Formation in Hardened Dichromated Gelatin Films," *Appl. Opt.* **8** (1969) no. 5, 963.
18. R. K. Curran, T. A. Shankoff, "The Mechanism of Hologram Formation in Dichromated Gelatin," *Appl. Opt.* **9** (1970) no. 7, 1651.
19. S. K. Case, R. Alferness, "Index Modulation and Spatial Harmonic Generation in Dichromated Gelatin Films," *Appl. Phys.* **10** (1976) no. 1, 41.
20. C. D. Leonard, B. D. Guenther, "A Cookbook for Dichromated Gelatin Holograms," Technical Report T-79-17, U.S. Army Missile Research and Development Command (Redstone Arsenal, 1979).
21. B. J. Chang, C. D. Leonard, "Dichromated Gelatin for the Fabrication of Holographic Optical Elements," *Appl. Opt.* **18** (1979) no. 14, 2407.
22. W. R. Graver, J. W. Gladden, J. W. Eastes, "Phase Holograms Formed by Silver Halide (Sensitized) Gelatin," *Appl. Opt.* **19** (1980) no. 9, 1529.
23. J. A. Jenney, "Holographic Recording with Photopolymers," *J. Opt. Soc. Am.* **60** (1970) no. 9, 1155.
24. B. L. Booth, "Photopolymer Material for Holography," *Appl. Opt.* **14** (1975) no. 3, 593.
25. T. Todorov et al., "Photopolymers-Holographic Investigations, Mechanism of Recording and Applications," *Opt. Quant. Electron.* **16** (1984) no. 6, 471.
26. R. T. Ingwall, H. L. Fielding, "Hologram Recording with a New Photopolymer System," *Opt. Eng.* **24** (1985) no. 5, 808.
27. R. T. Ingwall, M. Troll, "The Mechanism of Hologram Formation in DMP-128 Photopolymer," *Proc. Soc. Photo-Opt. Instrum. Eng.* **883** (1988) 94.
28. T. Stone, N. George, B. D. Guenther, "Index Variation and Scattering in a Holographic Medium," *Proc. Soc. Photo.-Opt. Instrum. Eng.* **1136** (1989) 36.
29. F. P. Laming, "Holographic Grating Formation in Photopolymers-Polymethylmethacrylate," *Polymer Eng. Sci.,* **11** (1971) no. 5, 421.
30. R. A. Bartolini, "Characteristics of Relief Phase Holograms Recorded in Photoresist," *Appl. Opt.* **13** (1974) no. 1, 129.

31. J. A. MacAndrew, "Sensitivity and Non-Linearity of Shipley AZ-1375 Photoresist at 441.6 and 457.9 nm," *Opt. Acta* **25** (1978) no. 8, 751.
32. R. Bartolini, W. Hannan, D. Karlsons, M. Lurie, "Embossed Hologram Motion Pictures for Television Playback," *Appl. Opt.* **9** (1970) no. 10, 2283.
33. F. Iwata, J. Tsujiuchi, "Characteristics of a Photoresist Hologram and its Replica," *Appl. Opt.* **13** (1974) no. 6, 1327.
34. J. Cowan, W. D. Slafer, "Holographic Embossing at Polarold: The Polaform Process," *Proc. Soc. Photo-Opt. Instrum. Eng.* **600** (1985) 49.
35. J. R. Burns, "Large-Format Embossed Holograms," *Proc. Soc. Photo-Opt. Instrum. Eng.* **523** (1985) 7.
36. J. C. Urbach, R. W. Meier, "Thermoplastic Xerographic Holography," *Appl. Opt.* **5** (1966) no. 4, 666.
37. W. S. Colburn, E. N. Tompkins, "Improved Thermoplastic Photoconductor Devices for Holographic Recording," *Appl. Opt.* **13** (1974) no. 12, 2934.
38. J. Felnberg, D. Heiman, A. R. Tanguay, Jr., R. W. Hellwarth, "Photorefractive Effects and Light-Induced Charge Migration in Barium Titanate," *J. Appl. Phys.* **51** (1980) no. 3, 1297.
39. J. P. Herriau, D. Rojas, J. P. Huignard, J. M. Bassat, J. C. Launay, "Highly Efficient Diffraction in Photorefractive BSO-BGO Crystals at Large Applied Fields," *Ferroelectrics* **75** (1987) no. 1 – 2, 271.
40. J. W. Burgess, R. J. Hurditch, C. J. Kirkby, G. E. Scrivener, "Holographic Storage and Photoconductivity in PLZT Ceramic Materials," *Appl. Opt.* **15** (1976) no. 6, 1550.
41. E. S. Barrekette, et al.: *Applications of Holography,* Plenum, New York 1971.
42. G. von Bally (ed.): *Holography in Medicine and Biology,* Springer Verlag, Berlin 1979.
43. H. Hervet, W. Urbach, F. Rondelez, "Mass Diffusion Measurements in Liquid Crystals by a Novel Optical Method," *J. Chem. Phys.* **68** (1978) no. 6, 2725.
44. D. M. Burland, Chr. Bräuchle, "The Use of Holography to Investigate Complex Photochemical Reactions," *J. Chem. Phys.* **76** (1982) no. 9, 4502.
45. F. W. Deeg, J. Pinsel, Chr. Bräuchle, J. Voltlander, "The Evaluation of Photochemical Quantum Yields by Holography," *J. Chem. Phys.* **79** (1983) no. 3, 1229.
46. H. Kim, T. Chang, H. Yu, "Electrophoretic Mobility by Electric Field Modulated Forced Rayleigh Scattering," *J. Phys, Chem.* **88** (1984) no. 18, 3946.
47. C. S. Johnson, Jr., "Chemical Exchange Effects in Holographic Relaxation Spectroscopy," *J. Chem. Phys.* **81** (1984) no. 12, 5384.
48. D. Gabor, "Microscopy by Reconstructed Wave-Fronts," *Proc. Royal Soc.* **A 197** (1949), 454.
49. A. V. Baez, "A Study in Diffraction Microscopy with Special Reference to X-Rays," *J. Opt. Soc. Am.* **42** (1952) no. 10, 756.
50. M. E. Haine, T. Mulvey, "The Formation of the Diffraction Image with Electrons in the Gabor Diffraction Microscope," *J. Opt. Soc. Am.* **42** (1952) no. 10, 763.
51. S. Aoki, S. Kikuta, "X-Ray Holographic Microscopy," *Jpn. J. Appl. Phys.* **13** (1974) no. 9, 1385.
52. R. F. Cournoyer, M. B. Rhodes, S. Siggia, "Holographic Interferometric Microscopy," *Analyt. Chem.* **48** (1976) no. 14, 2253.
53. W. Schumann, M. Dubas: *Holographic Interferometry,* Springer Verlag, Berlin 1979.
54. Y. I. Ostrovsky, M. M. Butusov, G. V. Ostrovskaya: *Interferometry by Holography,* Springer Verlag, Berlin, 1980.
55. I. Yamaguchi, H. Salto, "Application of Holographic Interferometry to the Measurement of Poisson's Ratio," *Jpn. J. Appl. Phys.* **8** (1969) no. 6, 768.
56. S. S. Nathan, S. D. Mehta, A. Selvarajan, "Holographic Interferometry Study of Chemical Bond of Concrete to Smooth Steel Rods," *J. Test. Eval.* **6** (1978) no. 4, 284.
57. A. G. Gaydon, H. G. Wolfhard: *Flames. Their Structure, Radiation, and Temperature,* Chapman and Hall, London 1970.
58. R. J. D. Miller, R. Casalegno, K. A. Nelson, M. D. Fayer, "Laser-Induced Ultrasonics: A Dynamic Holographic Approach to the Measurement of Weak Absorptions, Optoelastic Constants, and Acoustic Attenuation," *Chem. Phys.* **72** (1982) 371.
59. L. Beiser: *Holographic Scanning,* Wiley, New York 1988.
60. T. Stone, N. George, "Wavelength Performance of Holographic Optical Elements," *Appl. Opt.* **24** (1985) no. 22, 3797.
61. F. T. S. Yu, *Optical Information Processing,* Wiley, New York 1983.

Hormones

Peptide and Protein Hormones is a separate keyword. For insect hormones, see →Biological Control, Chap. 5., →Insect Control. For plant hormones, see →Plant Growth Regulators.

Jürgen Sandow, Hoechst AG, Frankfurt, Federal Republic of Germany (Chap. 1, Section 2.2)

Ekkehard Scheiffele, Henning Berlin GmbH, Berlin, Federal Republic of Germany (Section 2.1)

Michael Haring, Henning Berlin GmbH, Berlin, Federal Republic of Germany (Section 2.1)

Günter Neef, Schering AG, Berlin, Federal Republic of Germany (Chaps. 3–5)

Klaus Prezewowsky, Schering AG, Berlin, Federal Republic of Germany (Chaps. 3–5)

Ulrich Stache, Hoechst AG, Frankfurt, Federal Republic of Germany (Chaps. 3–5)

1. Introduction 3
1.1. Basic Principles 3
1.1.1. Definitions and Classification 3
1.1.2. Hormone-Producing Systems 5
1.1.3. Structural Analysis and Synthesis . . 7
1.1.4. Receptors and Mechanisms of Action 8
1.1.5. Therapeutic Use of Hormones 8
1.2. Peptide and Protein Hormones . . 8
1.2.1. Hypothalamic Hormones 8
1.2.2. Proopiomelanocortin Hormones . . . 9
1.2.3. Glycoprotein Hormones 10
1.2.4. Growth Hormone, Prolactin, and Placental Lactogen 11
1.2.5. Placental Hormones 11
1.2.6. Calcium-Regulating Hormones . . . 11
1.2.7. Cardiovascular Hormones 12
1.2.8. Glucose-Regulating Hormones . . . 12
1.2.9. Gastrointestinal Hormones 13
1.2.10. Growth Factors and Related Peptides 13
1.2.11. New Hormonal Peptides and Proteins 14
1.3. Amino Acid Derivatives 14
1.3.1. Catecholamines and Indolamines . . 14
1.3.2. Thyroid Hormones 14
1.4. Steroid Hormones 15
1.4.1. Gonadal Steroids 15
1.4.2. Adrenal Steroids 15
1.4.3. Calciferols 16
1.4.4. New Synthetic Steroid Hormones . . 16
1.5. Prostaglandins 16
1.6. Hormone Antagonists and Analogues 17
2. Amino Acid Hormones 17
2.1. Thyroid Hormones 17
2.1.1. Physiology and Therapeutic Uses . . 17
2.1.2. Physical Properties 19
2.1.3. Production 19
2.2. Hormones of the Adrenal Medulla 20
3. Steroid Sex Hormones 23
3.1. Introduction 23
3.2. Partial Chemical Synthesis of Steroid Intermediates 24
3.2.1. Raw Materials 24
3.2.2. Partial Synthesis 27
3.3. Microbiological Methods for the Synthesis of Steroid Intermediates 32
3.4. Total Synthesis of Sex Hormones . 33
3.5. Estrogens 35
3.5.1. General Aspects 35
3.5.2. Natural Estrogens and Estradiol Derivatives 36
3.5.3. Nonsteroidal Estrogens 37
3.5.4. Antiestrogens 38
3.6. Androgens and Anabolic Agents . 39
3.6.1. General Aspects 39
3.6.2. Androgens 40
3.6.3. Antiandrogens 41
3.6.4. Anabolic Agents 42
3.7. Gestagens 43
3.7.1. General Aspects 43
3.7.2. Gestagens of the Pregnane and 19-Norpregnane Series 45
3.7.3. Gestagens of the Androstane Series 45
3.7.4. Gestagens of the 19-Norandrostane Series 45
3.7.5. Antigestagens 50
4. Adrenal Steroid Hormones 50
4.1. Introduction 50
4.2. Historical Aspects 51
4.3. Physiology and Mode of Action . . 51
4.4. Therapeutic Uses of Glucocorticoids 52
4.5. Therapy and Side Effects 52
4.6. Biosynthesis and Metabolism . . . 53
4.7. Detection Methods 54
4.8. Naturally Occurring Glucocorticoids 54

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_089

4.8.1. Properties and Uses 54
4.8.2. Partial Synthesis of Cortisone 55
4.8.3. Partial Synthesis of Cortisol 57
4.9. Naturally Occurring Mineralocorticoids and Antagonists 58
4.9.1. Cortexone 58
4.9.2. Aldosterone 59
4.9.3. Spironolactone 59
4.10. Chemically Modified Adrenal Steroid Hormones 60
4.10.1. 1,2-Dehydrocorticosteroids 60
4.10.2. 9α-Halocorticosteroids 61
4.10.3. 16-Hydroxycorticosteroids 62
4.10.4. 6α-Methylcorticosteroids 63
4.10.5. 16-Methylcorticosteroids 65
4.10.6. 16-Methylenecorticoids 67
4.10.7. 6α-Fluoro-, 6α-Chloro-, and 6α,9α-Difluorocorticoids 67
4.10.8. 17-Deoxycorticosteroids 69
4.10.9. 9α,11β-Dihalocorticosteroids 71
4.11. Other Developments 71
5. Cholecalciferol 72
5.1. Introduction 72
5.2. Biosynthesis and Metabolism . . . 72
5.3. Biological Activity 73
5.4. Chemical Synthesis 74
5.5. Synthetic Analogues 75
6. References 75

Abbreviations used in this article:

ACTH adrenocorticotropic hormone, corticotropin
ACE angiotensin converting enzyme
ADH, AVP antidiuretic hormone, vasopressin
ANP atrionatriuretic peptide
CCK cholecystokinin
CRH corticotropin releasing hormon, corticocoliberin
CT calcitonin
FSH follicle stimulating hormone, follitropin
GH growth hormone (also designated STH, somatotropic hormone)
GnRH gonadotropin releasing hormone, gonadorelin (also designated LHRH, luteinizing hormone releasing hormone)
GHRH,GRH growth hormone releasing hormone
HCG human chorionic gonadotropin = choriotropin
HCS human chorionic somatotropin (also designated HPL, human placental lactogen)
HMG human postmenopausal gonadotropin = menotropin
HPL human placental lactogen (also designated HCS, human chorionic somatotropin)
LH luteinizing hormone, lutropin (also designated ICSH, interstitial cell stimulating hormone)
LHRH luteinizing hormone releasing hormone
MSH melanophore stimulating hormone, melanotropin
PIF prolactin inhibiting factor
PMS, PMSG pregnant mare serum gonadotropin, equitropin
PRL prolactin
PTH parathormone, parathyrin
STH somatotropic hormone, somatotropin
T_3 triiodothyronine
T_4 thyroxine
TRF thyrotropin releasing factor (also designated TRH)
TRH thyrotropin releasing hormone, thyroliberin
TSH thyroid stimulating hormone, thyrotropin

1. Introduction

1.1. Basic Principles

1.1.1. Definitions and Classification

Definitions. Hormones are discrete substances which regulate biological function, either directly by acting on target organs or indirectly by controlling the secretion and synthesis of secondary or tertiary hormone systems [1,8,21]. They appear early in evolution. Hormonal substances have been identified in bacteria and plants which later in phylogenesis acquired functional roles in animal species. Plant hormones are treated elsewhere (→Plant Growth Regulators). The highly selective action of hormones is due to their specific interaction with cell receptors. Adaptation of hormones and

receptors is a long evolutionary process, during which the hormone receptors may change and the chemical nature of the hormone remains the same. For example, insulin controls gonadal function in certain invertebrates but in mammals it is a glucose-regulating hormone that also controls the metabolism of fatty acids and amino acids. Precise definition of new hormonal substances is difficult due to rapid progress in hormone biochemistry, physiology, and pharmacology. In classical endocrinology, hormones are defined as products that are synthesized by a group of cells (a hormone-producing organ), secreted into the bloodstream, and transported to their site of action (a hormone-responsive organ or tissue) where they exert specific effects. The hormone-producing organ is identified by the deficiencies observed after surgical removal of the source of hormone synthesis (endocrine gland). The hormone is characterized by isolation, structural analysis, and chemical synthesis; the deficiencies observed after removal of the endocrine gland are corrected by substitution. In this experimental approach, the biological effects of the hormone are used for its initial identification (bioassay) during purification from biological materials. Another source of information about hormones is clinical endocrinology [17]. The disease symptoms resulting from hormone deficiencies or hypersecretion provide a concept for the biological function of a particular hormone, and suggest methods for its bioassay. If the hormone is produced in cells disseminated throughout a nonremovable organ or at multiple sites in the body, a biochemical approach must be employed for identification of the hormone's physicochemical characteristics that can be exploited for its extraction and purification from normal tissue or endocrine tumors. Some tumors produce hormones in much larger quantities [23] than the physiological concentrations found in normal, nonneoplastic tissue.

The definition of hormones and regulatory factors will certainly change, classical hormones will no longer be defined as products associated only with specific endocrine glands, but will turn up in new unexpected locations and functions. Hormonal regulatory factors will acquire a more comprehensive role as endogenous substances secreted by many endocrine and nonendocrine cells. Their definition may then depend on regulation of cell activity rather than on their site of origin. The future classification of hormones may be based on assignment to chemical families, on interaction with receptors and intracellular effector systems, physiological and pharmacological effects, and localization in different tissues.

Classification of Hormones. Hormones can be classified according to their structural and biochemical similarities. This procedure is particularly useful because the wide distribution of hormonal substances in nonendocrine tissues makes it difficult to assign them to specific endocrine organs. Progress in structural analysis and isolation of natural compounds is rapid, especially as regards peptide and protein biosynthesis from genes expressed in different tissues and posttranslational processing to mature hormones [24]. Peptide and protein hormones and their precursors may be secreted directly into the bloodstream (*endocrine secretion*), released from nerve endings into specialized local vascular systems (*neuroendocrine secretion*), or secreted into the surrounding tissue (*paracrine secretion*).

A contemporary definition of hormones must include the peptides and amino acid derivatives (neurotransmitters) produced and secreted by specialized neuroendocrine and endocrine cells in the endocrine hypothalamus [14], the gastrointestinal tract [25], the right atrium of the heart [25], the endocrine pancreas [26], the thymus [27], and other organs not exclusively devoted to hormone secretion and synthesis. Neuroendocrine cells with a dual role of hormone secretion and neurotransmitter release are found in the adrenal medulla and sympathetic ganglia. Surprisingly, many hormone products that resemble the peptides of the central nervous system and the gastrointestinal hormones are found in the skin of amphibia.

Advances in structural elucidation, chemical synthesis, and specific detection methods (radioimmunoassay, immunohistochemistry, high-performance liquid chromatography) have disclosed a surprisingly wide dissemination of hormone-producing cells in tissues that were previously inaccessible to investigation by surgical removal and chemical substitution. The structural analysis of peptides, proteins, and their precursors by gene sequencing is progressing much more rapidly than the biological evalu-

ation of new hormone candidates. It is therefore helpful to discuss hormones according to their chemical and functional families rather than according to hormone-producing glands.

The most important chemical groups are the *steroid hormones* (Chaps. 3, 4, 5), [20] closely related to their common precursor, cholesterol, and the *peptide and protein hormones* [7] composed of amino acids which form chains of different length and conformation, frequently containing discrete subunits (→Peptide and Protein Hormones). The hormone group of *amino acid derivatives* (Chap. 2) comprises the *thyroid hormones*, and the hormonally active *catecholamines* and *indolamines*. The biogenic amines have multiple functions as hormones and neurotransmitters in the central and autonomous peripheral nervous systems. Hormone candidates are the *kinins* [12,28] and *prostaglandins* [29,30], which are a group of widely distributed endogenous substances with marked hormone-like activities.

1.1.2. Hormone-Producing Systems

Many hormones are integrated in complex regulatory systems. Their secretion may be activated or suppressed by hypothalamic and/or pituitary control, and the central regulation is in turn under the influence of feedback signals from peripheral hormone-responsive organs and tissues. Hormones with a homeostatic function maintain a constant internal environment by adaptation to changing external conditions (thyroid hormones and adrenal steroids). Hormones may be secreted throughout life if they are indispensable for metabolic activities (e.g., glucose- and calcium-regulating hormones), or at certain times of ontogenesis such as sexual maturation (gonadotropins and gonadal steroids).

Classical hormone-producing organs are the anterior and posterior pituitary gland, the thyroid gland, the parathyroid glands (epithelial bodies), the endocrine pancreas, the adrenal glands (adrenal cortex and adrenal medulla with chromaffin cells), the gonads, and the pineal gland. Destruction of these glands is associated with characteristic deficiency syndromes. Transient hormone-secreting organs also exist, e.g., the corpus luteum of the ovary and the placenta.

Site of Origin and Biosynthesis. To identify a hormone's site of origin, an endocrine gland is removed to study the resulting functional deficiencies and the effect of hormone substitution. More recent techniques to identify hormone-producing cells in non endocrine organs are immunohistochemistry and active immunization of animals against a hormone to neutralize its effects and define its function. Such studies sometimes disclose characteristic deficiency symptoms, but often fail due to compensatory mechanisms which replace the defective link [15].

Cells of common embryological ancestry may produce similar functional compounds even after their integration into nonendocrine tissues during ontogenesis. One hypothesis assigns a specific role to neuroendocrine Amine Precursor Uptake and/or Decarboxylation cells (APUD cells) which contain peptides and neurotransmitter-like amino acid derivatives. Such cells are widely found in the endocrine hypothalamus, anterior pituitary, parathyroid, and the gastrointestinal tract [31]. Biosynthesis of steroids is limited to the adrenal cortex. testes, and ovaries; the liver and kidneys metabolize steroid precursors to active calciferol derivatives. The thyroid hormones (thyroxine, triiodothyronine) are synthesized from monoiodotyrosine and diiodotyrosine. In the pineal gland, the hormone melatonin is synthesized from the precursor tryptophan. Remarkably, the placenta as a transient hormone-producing structure can synthesize peptides (e.g., LHRH, also called gonadorelin), proteohormones (e.g., HCG, human chorionic gonadotrophin), and steroid hormones (e.g., estrogens and progestins). Similarly, the corpus luteum of the ovary (a transient structure formed after ovulation) can secrete steroid hormones (estrogens and progestins), peptides (oxytocin), proteohormones (relaxin), and tissue hormones derived from arachidonic acid (prostaglandins).

Mode of Secretion. Many hormones are released directly into the bloodstream and reach their target organs over considerable distances. However, special modes of secretion also occur. Tissue hormones with direct (paracrine) action on neighbouring cells are secreted directly into the surrounding tissue and regulate cell-to-cell interaction. Some neurosecretory hormones (oxytocin, vasopressin) are transported to their

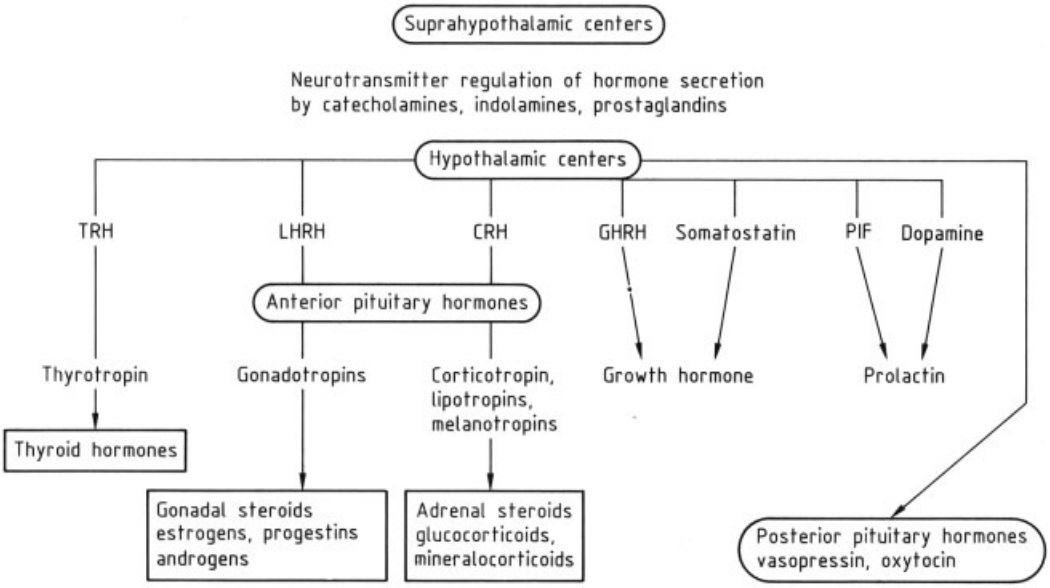

Figure 1. Hierarchy of hormone systems

site of storage in vesicles by axonal transport and released upon appropriate stimuli. The hypothalamic regulatory hormones reach the anterior pituitary gland (their target organ) by a special vascular pathway, the portal vessel system [14]. The pituitary gland is a composite, complex organ that consists of the anterior lobe (adenohypophysis), and a posterior lobe (neurohypophysis). The adenohypophysis contains several distinct cell types that secrete TSH, LH, FSH, growth hormone, prolactin, and corticotropin as well as proteohormones with presently undefined functions (endorphins, lipotropin, melanotropin). The posterior pituitary gland is a storage organ for two oligopeptide hormones, vasopressin and oxytocin that are synthesized in hypothalamic centers and transported to the posterior pituitary gland by axonal transport in neurosecretory cells, together with their binding proteins, the neurophysis.

The hierarchy of hormonal systems is summarized in Figure 1. Hierarchic hormone systems with two or three levels of control are regulated by hypothalamic hormones in the hypophysiotropic zone of the hypothalamus. These regulatory peptides activate or inhibit secretion of the anterior pituitary hormones. The *pituitary glandotropic hormones* (TSH, ACTH, FSH, LH) regulate their target tissues indirectly by hormone secretion of peripheral endocrine glands (thyroid and adrenal glands, gonads). The *pituitary effector hormones* (growth hormone, prolactin) act directly on target organs (Fig. 2), regulating proliferation and metabolism of peripheral

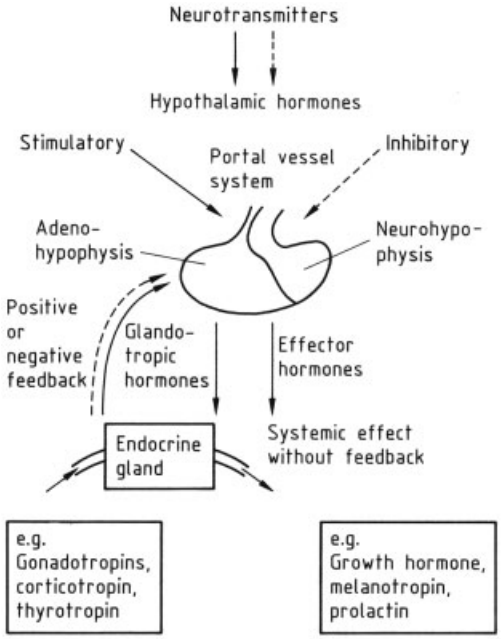

Figure 2. Secretion of pituitary hormones

tissues; in this respect they are similar to the calcium-regulating, glucose-regulating, and gastrointestinal hormones. Steroid hormones secreted by the gonads and adrenals act directly on target cells to induce specific proteins which regulate gene expression.

Control of Hormone Secretion. An important principle in the homeostatic regulation of body functions by hormones is their organization in hierarchic *feedback systems*, which may operate independently or in interdependence. Secretion of effector hormones is regulated by the feedback of their respective substrate (e.g., insulin by glucose, or parathormone by calcium). Secretion of glandotropic hormones is controlled by the feedback of their products (e.g., steroids) synthesized by the target organ. The serum concentration is monitored by hypothalamic and/or pituitary receptors. In the endocrine hypothalamus (hypophysiotropic zone), the secretion of regulatory factors (peptides and neurotransmitters) drives the endogenous rhythms of pituitary hormones [32]. The corticotropin releasing hormone (CRH) induces the secretion of ACTH (corticotropin), which has a marked diurnal rhythm with an early morning peak [33]. The gonadotropin releasing hormone GnRH, also known as gonadorelin or luteinizing hormone, LHRH) has a pulsatile circhoral rhythm which is reflected by pulsatile gonadotropin secretion, the amplitude and frequency of gonadotropin pulses being regulated by gonadal steroids [34]. The secretion of gonadotropin in a multiloop feedback system is summarized in Figure 3. The growth hormone releasing hormone (GHRH) activates the secretion of growth hormone (somatotropin), which has a marked endogenous sleep-associated rhythm closely coupled to rapid eye movement activity of the central nervous system [35]. Hormone secretion can also be controlled by neural regulation (e.g., prolactin by the suckling reflex of nursing), and is often modulated by the autonomous nervous system.

1.1.3. Structural Analysis and Synthesis

In hormone chemistry and biochemistry, the constraints of classical organ-related endocrinology are replaced by new concepts based on chemical, biochemical, and functional similarities, which take full advantage of modern methodology for identification of minute amounts of hormones, their precursors, and their metabolites in various tissues. New natural compounds (hormone candidates) can be identified by classical extraction, purification, and isolation followed by structural analysis and chemical synthesis, or by indirect structural analysis via gene sequencing, and biosynthesis by recombinant DNA technology. Their physiological and pharmacological functions must be established in biological studies.

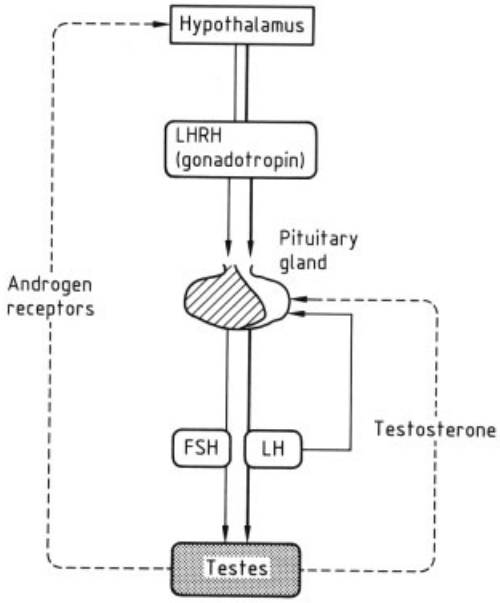

Figure 3. Feedback regulation of gonadotropin secretion
The pituitary secretion of LH can act directly (short feedback) on the secretion rate. The testicular steroid secretion can also modulate gonadotropin secretion (long feedback). In the hypothalamic centers the LHRH pulsatility is controlled by the binding of gonadal steroids to receptive cell areas.

1.1.4. Receptors and Mechanisms of Action

Peptide and protein hormones act on specific receptors located in the cell membrane, whereas steroid hormones act on receptors located in the cytoplasm and cell nucleus. The formation of a hormone – receptor complex activates specific intracellular processes [13].

Membrane receptors for peptides and protein hormones are coupled to adenyl cyclase [36] or phospholipase as the intracellular effector system; they activate protein kinases, which regulate protein synthesis and the calcium permeability of the cell membrane and intracellular structures [37].

The binding capacity of receptors for peptides and protein hormones depends on the concentration of their ligand in the endocrine cells. The receptors are expressed upon physiological stimulation, their number increases with the ligand concentration (up-regulation), and declines when the physiological level of stimulation is exceeded. During supraphysiological stimulation, the receptor response is desensitized, and binding capacity is down-regulated (i.e., the number of receptors is reduced). The binding of hormone analogues with enhanced resistance to degradative enzymes often reduces the concentration of available receptors. Receptors are removed by enhanced coupling between receptors (microaggregation), followed by internalization of hormone – receptor complexes, and intracellular receptor degradation. Steroid hormones bind to intracellular (cytoplasmic or nuclear) receptors. The activated steroid-receptor complexes are translocated to the cell nucleus, where they regulate the synthesis of enzymes or functional proteins by means of gene expression [18, 38]. The binding capacity of the steroid hormone receptors is regulated in a dose-dependent manner. Steroid receptors are generally induced in the presence of their ligands and decrease with declining secretion of the relevant steroid.

1.1.5. Therapeutic Use of Hormones

Natural hormones provide valuable leads for structure – activity studies. Numerous steroid hormone analogues have been introduced into therapy, and synthetic analogues of peptides have also become valuable diagnostic and therapeutic aids. The larger proteins previously obtained by extracting biological material are now produced more economically by recombinant DNA technology [2] (→Genetic Engineering). The enhancement of specific biological activities or modification of the activity spectrum toward more desirable effects are the general aim of structure – activity studies.

1.2. Peptide and Protein Hormones

(→ Peptide and Protein Hormones)

1.2.1. Hypothalamic Hormones

The hypothalamic hormones comprise peptides which regulate anterior pituitary function (*neurohumoral mechanism*), and others (vasopressin and oxytocin) which are stored in the posterior pituitary and released upon demand (*neurosecretory mechanism*). Paracrine secretion of hormones into the surrounding tissue is the most ancient phylogenetic transmission principle. The neuroendocrine secretion of vasopressin and oxytocin is analogous to the more specific and effective transmission between nerve cells that involves the synaptic release of amine or peptide neurotransmitters. Vasopressin and oxytocin are synthesized in the hypothalamic neurosecretory cells and are transported along the cell axons to their storage vesicles in the posterior pituitary gland. The hormones are structurally similar nonapeptides that are stabilized by intramolecular sulfhydryl bridges. *Vasopressin* controls water reabsorption by the kidney and osmolality in the intravascular space. At pharmacological concentrations, vasopressin has marked cardiovascular effects (vasoconstriction and hypertension). Numerous vasopressin analogues have been synthesized to enhance antidiuresis, and to reduce hypertensive activity [39]. A second group of vasopressin agonists is used for the local control of bleeding by vasoconstriction. A central role for hormonally inactive vasopressin fragments has been postulated in the control of memory and learning. *Oxytocin* acts primarily on the contractility of smooth muscle of the uterus during parturition, and on the myoepithelium of the lactating mammary gland. Surprisingly, oxytocin has been identified in the ovary, but its physiological role in this organ is unknown. Detection of other established hormones in tissues where the occurrence was previously unknown is increasing due to improved microanalytical techniques.

1.2.2. Proopiomelanocortin Hormones

Peptides and proteins are generally synthesized as large prehormone molecules and processed to smaller active substances. A complex

hormone precursor found both in hypothalamic and anterior pituitary tissue gives rise to a family of opioid, corticotropic, and melanotropic peptides and proteohormones with diverse actions and functions [40–42]. The precursor, preproopiomelanocortin is processed to corticotropin (ACTH), melanotropin (MSH), lipotropins (LPH), enkephalins and endorphins.

Corticotropin (ACTH) is a linear polypeptide of 39 amino acids. This classical pituitary hormone is an interesting example of the post-translational processing of a precursor to form a highly active product. Structure – activity relations of ACTH analogues have been studied extensively [43]. The full corticotropic activity is expressed in the N-terminal amino acid sequence 1 – 24; the C-terminal sequence 25 – 39 does not contribute to biological activity. Corticotropin analogues with protected N- and C-termini significantly increase and prolong cortisol and aldosterone secretion in the adrenal cortex. They consist of 17 or 18 amino acids with enhanced enzyme resistance and a prolonged half-life for plasma elimination. ACTH (1 – 39) is secreted after hypothalamic stimulation of the corticotropin-producing cells by the corticotropin releasing hormone (CRH), a hypothalamic polypeptide of 41 amino acids [33]. This type of regulation of a small pituitary hormone by a slightly larger regulatory hormone is a surprising exception to the rule that hypothalamic hormones are small molecules which release several multiples of their own weight from the anterior pituitary gland (amplifier mechanism). Secretion of ACTH in the human has a characteristic peak in the early morning hours and can be reproduced by pulsatile injection of human CRH. There are clear indications for a dual control of ACTH by CRH and arginine – vasopressin. A similar redundant dual control system is found for mineralocorticoid secretion, aldosterone being secreted by the adrenal cortex after stimulation by both ACTH and angiotensin II. The latter is formed from angiotensinogen by the successive action of the enzymes renin and angiotensin converting enzyme (ACE). ACTH is a potent stimulator of the adrenal cortex, its hypersecretion by corticotropic pituitary adenomas can override the adrenal feedback completely. ACTH regulates biosynthesis of adrenal steroids from cholesterol via the intermediate Δ-5-pregnenolone [7]. It also has extraadrenal effects on pigmentary cells (melanotropic activity), and on mobilization of fatty acids from fat cells (lipolytic activity). When ACTH secretion is abolished by hypothalamic or pituitary lesions, the adrenal cortex rapidly involutes resulting in a characteristic hormone deficiency syndrome (Addison's disease) that is also observed after destruction of the adrenal cortex by tumors or infectious diseases. Excessive ACTH secretion results in marked melanotropin activity, inducing intense pigmentation of the skin.

Other polypeptides with significant sequence homologies to ACTH are also expressed on the proopiomelanocortin gene, namely α-, β-, *and* γ-*MSH.* These hormones have intense pigmentary activity in animals and humans. α-MSH is an N-terminal derivative that is acetylated by posttranslational processing. β-MSH is found as a circulating hormone in the human, but no physiological role has been established. The *lipotropins* [44] are of intermediate molecular mass (58 – 92 amino acids) and are also derived from the common ACTH – MSH precursor. Their physiological role is not fully established, they release free fatty acids from adipose tissue in vitro and in vivo. The *enkephalins* and *endorphins* (neuropeptides or peptide neurotransmitters) are regulators of central nervous activities, and of peripheral sensory functions such as perception of pain.

1.2.3. Glycoprotein Hormones

A group of pituitary hormones closely related by common subunits (an α-subunit and a hormone-specific β-subunit) comprises thyroid stimulating hormone (thyrotropin, TSH) and the gonadotropins follicle stimulating hormone (follitropin, FSH) and luteinizing hormone (lutropin, LH). The biological effects of these polypeptides are different despite extensive sequence homologies of their subunits. They undergo post-translational glycosylation after assembly of the α and β-subunits [45–48].

Thyrotropin. The secretion of thyrotropin (TSH) is under hypothalamic control of thyrotropin releasing hormone (TRH, protirelin), a tripeptide with protected N- and C-terminus. To a limited extent, protirelin also stimulates the release of prolactin (PRL) and is involved in other hypothalamic control systems for the release and

inhibition of prolactin secretion [14]. Numerous analogues of TRH have been synthesized with enhanced TSH releasing activity, or with significant neuropharmacological and cardiovascular effects but without endocrine activity [11]. TRH is useful in the diagnosis of thyroid function, but has not acquired a therapeutic role. Excess stimulation of TSH secretion by TRH can induce thyroid hyperplasia in animals, whereas human hyperthyroidism is generally caused by receptor-stimulating antibodies. Goiter formation in the human may result from disorders of thyroid hormone synthesis, which cause hypersecretion of TSH and proliferation of thyroid tissue. Human TSH has not found a therapeutic role, because thyroid hormone secretion is more easily substituted by synthetic thyroxin preparations.

Gonadotropins. The secretion of the gonadotropins FSH and LH is regulated by the hypothalamic luteinizing hormone releasing hormone (LHRH), also designated as the gonadotropin releasing hormone (GnRH). The decapeptide LHRH is used for diagnosis and therapy of infertility. With an intact pituitary gland, pulsatile infusion of LHRH stimulates gonadotropin secretion. This provides a method for treating human hypogonadotropic hypogonadism and is an alternative to gonadotropin preparations. Numerous agonists and antagonists of LHRH have been synthesized [9,49] to control gonadal steroid secretion in diseases such as endometriosis, uterine leiomyoma, and steroid-dependent tumors (prostate and mammary carcinoma). The LHRH agonists induce persistent receptor desensitization and selective inhibition of gonadotropin secretion during daily high dose injections or continuous release from implants. The LHRH antagonists have similar inhibitory pharmacological effects on gonadal steroid secretion although their mechanisms of action are clearly different.

The gonadotropins are glycoproteins with about 180 amino acids, and multiple glycosylation sites [46,50]. Follitropin (FSH) stimulates estrogen secretion from ovarian follicles and prepares them for subsequent induction of ovulation by lutropin (LH). LH has ovulatory activity and enhances ovarian steroidogenesis in synergy with FSH, in particular the secretion of progesterone by the corpus luteum after ovulation. In men, spermatogenesis is dependent on FSH, which acts on the germinal epithelium in synergy with LH and testosterone. LH stimulates the interstitial cells of the testes to secrete testosterone and small quantities of estrogens.

In animals and humans, FSH and LH are excreted in the urine; similar gonadotropins have been extracted from the serum of pregnant animals (pregnant mare's serum gonadotropin, PMSG). These urinary gonadotropins have been extensively studied, and are used for therapy. Urinary gonadotropins with high FSH- and LH-activities can be extracted from the urine of postmenopausal women for the stimulation of follicular maturation in infertile women (human menopausal gonadotropin, hMG).

1.2.4. Growth Hormone, Prolactin, and Placental Lactogen

Extensive structural homology is found between pituitary growth hormone (GH, somatotropin), pituitary prolactin (PRL), and human placental lactogen (hPL, see under Section 1.2.5). GH and prolactin contain 191 and 198 amino acids, respectively [51]. Both hormones have been extracted from human and animal pituitary glands, but only growth hormone has become a valuable therapeutic agent. It is produced by recombinant DNA biosynthesis to avoid contamination with the neurotropic virus found in human pituitary tissue [52].

The amino acid sequence of *growth hormone* has marked species specificity [53,54], therefore substitution therapy in children should only be performed with human growth hormone (hGH) to avoid antibody formation and loss of efficacy. The plasma elimination half-life of injected growth hormone is about 20 min. Growth hormone has many effects on metabolism (e.g., increased glucose utilization and lipid mobilization) and a marked anabolic effect on amino acid transport into cells and subsequent protein biosynthesis. Treatment with hGH can correct growth deficiency and short stature in children provided that epiphyseal closure of the bones has not occurred. Appropriate species-specific GH can also induce growth in farm animals. Many of the metabolic effects of GH are mediated by inducing the synthesis of growth factors [4], particularly the somatomedins. The GH-dependent somatomedins stimulate the formation of new cartilage; increased incorporation of sulfate into

newly formed cartilage (sulfation factor) is a sensitive bioassay. During lactation, growth hormone acts in synergy with sexual steroids and prolactin on milk production. Hypersecretion of GH in acromegaly is manifested by pathological growth of the hands and feet, facial skeleton, enlargement of the skull, and heart, frequently followed by cardiac insufficiency.

Despite marked sequence homology with growth hormone, the growth-inducing activities of prolactin (PRL) and human placental lactogen (HPL) are negligible. *Prolactin* contributes to the maintenance of corpus luteum function in animals, and stimulates milk production during lactation, in synergy with GH and sexual steroids. Prolactin secretion appears to be under dual hypothalamic inhibitory control effected by release of dopamine [55] and specific peptide factors (prolactin inhibiting factor, PIF) into the pituitary portal vessels. Hypersecretion of PRL is found in pituitary micro- and macroadenoma. The insufficient endogenous inhibition of PRL secretion by endogenous dopamine can often be remedied by using potent dopamine agonists [56].

1.2.5. Placental Hormones

The placenta secretes structural homologues of TSH, FSH, LH, GH, and prolactin. A placental gonadotropin, human chorionic gonadotropin (choriotropin, hCG), is extracted from urine of pregnant women in large quantities for therapy of infertility; it is a glycoprotein that is similar to FSH and LH [57]. Other placental proteins (human placental lactogen and a TSH-like glycoprotein) have not found therapeutic applications [51]. Numerous oligopeptides occur in placental tissue, but are apparently not secreted during pregnancy.

1.2.6. Calcium-Regulating Hormones

There are two groups of calcium regulating hormones, the rapidly acting polypeptides calcitonin and parathormone (PTH), and the slowly acting steroids (calciferols, see Chap. 5). The polypeptides allow rapid adjustment of the serum calcium concentration to physiological reguirements. *Calcitonin* is a species-specific polypeptide of 32 amino acids (M_r 3700); salmon and chicken calcitonin are among the most potent natural calcitonins [58]. The hormone is synthesized in the parafollicular C cells of the human thyroid gland, and in the ultimobranchial gland of several species of fish. Calcitonin lowers elevated serum calcium and phosphate levels, inhibits the activity of osteoclasts (cells inducing the resorption of bone), decreases renal reabsorption of calcium and phosphate ions, and decreases absorption of calcium from the gastrointestinal tract. Calcitonin acts on specific peptide receptors in the bone, repeated administration causes progressive receptor desensitization. In the brain, a calcitonin-gene related peptide is found with unidentified central regulatory effects [24]. Calcitonin has analgesic effects, which may be mediated by brain and spinal cord receptors.

The physiological antagonist of calcitonin, *parathormone* (PTH) is a species-specific polypeptide of 84 amino acids; the N-terminal 34 amino acids are required for full biological activity [59]. Parathormone is synthesized in the parathyroid cells (epithelial bodies), which in the human are found close to the thyroid gland. The effects of PTH include rapid mobilization of calcium from the bone matrix by activation of osteoclasts, activation of phosphate excretion by the kidney via inhibition of reabsorption, concomitantly increased renal excretion of calcium, and enhanced absorption of calcium from the gut. Endocrine-active tumors may secrete parathormone and induce hypercalcemia, which responds rapidly to calcitonin injections.

1.2.7. Cardiovascular Hormones

Three groups of peptides contribute to the regulation of blood pressure and intravasal volume: the angiotensins, the atrionatriuretic peptide (ANP), and the kinins [12]. Marked vasoconstriction and blood pressure increase is found after the stimulation of angiotensin I production. Angiotensin I is a nonapeptide that is produced from the precursor angiotensinogen by the enzyme renin. Renin is secreted by the juxtaglomerular kidney cells upon reduction of the glomerular filtration rate and hypoxia. Angiotensin I is converted to the octapeptide angiotensin II by the angiotensin convert-

ing enzyme (ACE). Numerous angiotensin antagonists have been synthesized as antihypertensive (blood pressure-lowering) peptides, but all have intrinsic vasoconstrictor activity. Angiotensin antagonists can be used to diagnose angiotensin-induced hypertension because they competitively inhibit the increased angiotensin concentrations found after renal artery occlusion. The di- and tripeptide ACE inhibitors captopril and enalapril, are used for the long-term therapy of hypertension; they have found great importance due to their high efficacy after oral absorption.

The *atrionatriuretic peptide* (ANP) is a functional antagonist of angiotensin II. It occurs in the right atrium of the heart and is released upon acute dilatation of the right atrium [60, 61]. ANP is synthesized as a preprohormone complex of 152 amino acids, and post-translationally cleaved to an active peptide containing 28 amino acids. ANP agonists consisting of peptides containing 22 or 23 amino acids are available; they have markedly enhanced vasodilatory and blood pressure-lowering activity, as well as acute and sustained stimulatory effects on sodium excretion and urinary flow rate.

The *kinins* are also physiological antagonists of angiotensin and are found in many tissues. They are nona-, deca- or undecapeptides that are released by the enzyme trypsin from the α-globulin fraction of serum [12]. All kinins are potent vasodilators; they lower blood pressure acutely, increase capillary permeability, and rapidly induce local tissue reactions characterized by edema, pain, and acute inflammation. The kinin peptides cause contraction of smooth muscle in the intestinal tract and the uterus. They can be secreted in significant quantities by endocrine-active tumors of the gastrointestinal tract. Many bradykinin-derived peptides have been studied without confirmation of therapeutic usefulness.

1.2.8. Glucose-Regulating Hormones

The protein hormones insulin and glucagon have vital effects on energy metabolism, cell permeability, and the homeostasis of cell function. Insulin and glucagon are the most important physiological factors in the control of cellular glucose concentrations; however, other hormones (e.g., growth hormone, glucocorticoids and somatostatin) also have significant effects. Insulin is used for substitution therapy of diabetes mellitus; glucagon is employed very infrequently.

Insulin is a polypeptide of 51 amino acids. It contains an α- and a β-chain that are synthesized as part of a larger precursor molecule of 84 amino acids (proinsulin). After post-translational removal of the connecting peptide from proinsulin, the two chains are joined by disulfide bonds [3]. Insulin is stored in the β-cells of the pancreas as a hexameric zinc-containing complex. The main action is to regulate cell permeability to glucose and potassium. The primary sequence of insulin is species-specific, but the biological action of human and animal insulins are similar. Porcine insulin is most closely related to human insulin. Human insulin can be obtained from porcine insulin by removing the eight N-terminal amino acids and replacing them with the corresponding human sequence. Human insulin has a plasma elimination half-life of about 20 min after intravenous injection. Numerous derivatives and pharmaceutical formulations have been developed to prolong the action of insulin by delayed absorption and sustained release from the subcutaneous injection site. Insulin pumps can deliver insulin according to a preset daily program; nasal administration of insulin is also being studied. The chemical synthesis of insulin has been described [6], but human insulin produced by recombinant DNA technology is now gradually replacing the previous porcine and bovine insulin preparations. Numerous structure activity studies have been performed with insulin analogues but they are not used in clinical medicine [6].

The physiological antagonist of insulin is *glucagon* [22], a linear polypeptide of 29 amino acids (M_r 3500). The biosynthesis of glucagon proceeds via a large preprohormone. The pancreatic β-cells contain biologically active glucagon and two glucagon-related peptides of unknown physiological significance. The gut contains an inactive precursor form of glucagon (glicentin). Glucagon is released from the pancreas upon a decrease in the blood-glucose concentration and stimulates gluconeogenesis. Its plasma elimination half-life is short (ca. 5 min). Glucagon has a strong inotropic action (i.e., enhances myocardial contractility). In contrast to insulin, glucagon has not acquired an important

therapeutic role, although some of its pharmacological effects at higher doses have been widely studied.

1.2.9. Gastrointestinal Hormones

Numerous peptides and proteins with established or possible hormonal function are found in the salivary glands, stomach, duodenum, ileum, and colon. These gastrointestinal hormones influence acid and bile secretion, digestive enzyme production, and gastrointestinal motility [5,10,25,62]. *Tissue hormones* such as kinins and neurotensin have been isolated from the salivary glands, but their physiological role is doubtful [12]. In the stomach, *gastrin* is found in three different molecular forms with 13, 17, and 34 amino acids [16]. It is an N-terminally processed peptide whose biological activity (stimulation of acid secretion in the stomach) is associated with its C-terminal 14 – 34 sequence. Gastrin has extensive sequence homology with the C-terminal active sequence of *cholecystokinin* (CCK), which is synthesized in the cells of the duodenal mucosa. Cholecystokinin stimulates contraction of the gall bladder and release of bile into the duodenum; it also enhances the secretion of pancreatic enzymes. The biological activity of CCK is attributed to a linear sequence of 33 amino acids (M_r 3883). Several smaller C-terminal forms of CCK are found in the duodenum and in central nervous tissue [63]. The central role of CCK fragments may be in the control of appetite as a satiety signal. Another duodenal hormone is *secretin*, a linear polypeptide with 28 amino acids (M_r 3000). Secretin stimulates the release of bicarbonate-rich pancreatic fluid which neutralizes the acid gastric juice when it reaches the duodenum. Secretin has significant sequence homologies with glucagon and other gastrointestinal peptides such as the vasoactive intestinal polypeptide and gastrointestinal polypeptide. *Motilin* is a further gastrointestinal polypeptide (22 amino acids) which stimulates gastric motility. Tumors of the gastrointestinal tract may contain and secrete a large variety of hormonally active peptides, e.g., growth hormone releasing hormone (GHRH) and its fragments, somatostatin, vasoactive intestinal peptide, and gastrin [5].

1.2.10. Growth Factors and Related Peptides

The growth requirements of cultured cells have been studied extensively. Specific growth factors appear to be involved in cell replication, physiological tissue repair, and pathological tumor growth [4]. Their secretion may be activated by human oncogenes and viruses [64].

The growth factors are produced by the tissues in response to growth hormone and other hormones. For example, the *somatomedins* are produced in response to the action of growth hormone; and *the estromedins* in response to the action of estrogens [65]. Numerous closely related groups of factors have been described such as the nerve growth factors [4], the epidermal growth factor and transforming growth factor alpha, the basic and acidic fibroblast growth factor [66], the endothelial cell growth factor [67], the platelet-derived growth factor, and the transforming growth factor beta [68]. The hormonal roles of these regulatory polypeptides are not clearly established because they have many pharmacological effects in vitro, and definition of their physiological functions is difficult.

1.2.11. New Hormonal Peptides and Proteins

The rapid progress of indirect structural analysis by gene sequencing and recombinant DNA technology has resulted in characterization and biosynthesis of many new hormonal factors. An example is *erythropoietin*, a glycoprotein found in the kidney which stimulates erythrocyte production [69] and hemoglobin synthesis. Local ovarian hormones have been identified such as relaxin [70,71], and the Muellerian-inhibiting substance (MIS). The MIS is a glycoprotein of 300 – 350 amino acids that is synthesized as a preprohormone of 560 amino acids but its hormonal role is difficult to assess [72]. It controls the development of the male phenotype by suppressing development of the (female) Muellerian duct during embryonal differentiation. The MIS may have a pharmacological use to suppress proliferation of ovarian and endometrial tumors. Another ovarian protein regulates FSH secretion in vitro and in vivo. It consists of two subunits, which may associate to form either an FSH-inhibiting heterodimeric protein (*inhibin*) or an FSH-releasing homodimeric protein (*activin*) [73,74].

1.3. Amino Acid Derivatives

The catecholamines, indolamines, and the more complex thyroid hormones are amino acid derivatives.

1.3.1. Catecholamines and Indolamines

Catecholamines have a dual role as cardiovascular and stress hormones. They are found in the chromaffin cells of the adrenal medulla and in sympathetic nerve endings; they also act as neurotransmitters in the central nervous system. *Noradrenaline* (norepinephrine) and *adrenaline* (epinephrine) are neurotransmitters that control hormone secretion by modulating the endocrine activity of the hypothalamus and pituitary gland. They are also involved in the central and peripheral regulation of blood pressure, and in the modulation of metabolic functions by the autonomous nervous system [75]. The most important indolamine is *melatonin*, a substance found in the pineal gland of animals and humans [76–78]. Melatonin inhibits sexual development in young animals before puberty (antigonadotropic hormone), its function in adults is not known.

1.3.2. Thyroid Hormones

The thyroid hormones thyroxine (3,5,3′,5′-tetraiodothyronine, T_4, and 3,5,3′-triiodothyronine, T_3) are iodinated derivatives of thyronine and are formed by oxidative coupling of the precursors 3-monoiodotyrosine and 3,5-diiodotyrosine. Thyroid-specific enzymes are responsible for the iodination of tyrosine. Numerous thyroid hormone analogues have been synthesized but do not have practical benefits for therapy [79]. Thyroxine (T_4) is stored in the thyroid follicles in the form of thyroglobulin (a glycoprotein with iodinated tyrosyl residues). The thyroid stimulating hormone (TSH) from the anterior pituitary stimulates release of thyroxine into the blood stream. Thyroxine serves both as a direct stimulator of metabolic activity, and as a prohormone for its more active metabolite, thyronine (T_3). The thyroid hormones regulate the adaptation of body temperature and metabolic rate to exogenous conditions. The pituitary – thyroid system is largely independent of hypothalamic activation and can function in animals with brain lesions that eliminate endogenous TRH secretion. Several synthetic compounds that inhibit the biosynthesis of thyroid hormones have been found, they act primarily by inhibiting thyroid hormone iodination, or by restricting the uptake of iodine into the thyroid gland [75].

1.4. Steroid Hormones

Steroid hormones share cholesterol as their common precursor [18, 20] and can be divided into three main physiological groups: the *gonadal steroids* produced in the ovary (estrogens and progestins) and testis (androgens), the *adrenal steroids* produced in the adrenal cortex (glucocorticoids and mineralocorticoids), and the *calcium-regulating sterols* (calciferols). The calcium-regulating calciferols [80], derivatives of 7-dehydrocholesterol and ergosterol, are activated by hydroxylation in the liver (25-hydroxycalciferol) and in the kidneys (1,25-dihydroxycalciferol and 24,25-dihydroxycalciferol).

There are three structural groups, the C-21, C-19, and C-18 steroids.

1.4.1. Gonadal Steroids

The most important C-21 gonadal steroid is *progesterone*, which is secreted in large amounts by the corpus luteum and the placenta. The most important C-19 gonadal steroid is *testosterone*, which is responsible for the secondary sex characteristics in the male and supports spermatogenesis in synergy with follicle stimulating hormone (FSH). *Estrogens* are C-18 gonadal steroids that are secreted in large quantities by the developing ovarian follicles, and in small quantities by the Leydig cells of the testis. Natural estrogens, progestins, and androgens are of little use for therapy because of poor gastrointestinal absorption and insufficient duration of action. Since 1930 many groups of workers have synthesized steroid hormone analogues to obtain synthetic hormones with improved therapeutic properties for the control of pituitary function (steroid inhibition of estrogen or androgen secre-

tion by gonadotropin suppression), for substitution therapy in deficiency syndromes such as hypogonadism or after the menopause, and particularly for the control of ovulation (contraceptive steroids). In general, the pharmacological activity of synthetic gonadal steroids is never limited to interaction with a particular type of receptor (i.e., estrogenic, androgenic, progestagenic, glucocorticoid, or mineralocorticoid); instead, the spectrum of biological and pharmacological activities is dose-dependent. For example, some synthetic C-19 steroids have a high progestagenic potency, but frequently retain androgenic effects. The synthetic estrogen 17-ethinylestradiol displays high oral absorption due to its low metabolic inactivation and has become a major component of oral contraceptives. Stilbene compounds are highly active nonsteroidal estrogenic substances with high oral absorption rates. Steroidal and nonsteroidal receptor antagonists have also been developed which competitively inhibit the receptor binding and the biological effects of natural androgens, estrogens, and progestins [9]. Of particular importance are the steroidal antiandrogen cyproterone acetate and the nonsteroidal antiandrogen flutamide. In treatment of postmenopausal mammary carcinoma, the partial antiestrogen tamoxifen is most widely used [9]; nonsteroidal antiestrogens are in development. The pharmacological activities of synthetic gonadal steroids have been reviewed extensively [19]. Work is aimed at developing new specific receptor antagonists (antiprogestins), and steroids with novel therapeutic activities (e.g., inhibitors of aromatase and 5-α-reductase).

1.4.2. Adrenal Steroids

The adrenal cortex contains glucocorticoids (cortisol and cortisone) and mineralocorticoids (aldosterone and desoxicorticosterone). These C-21 steroids [7] are of particular importance because of their life-supporting activity in substitution therapy of adrenal insufficiency.

The *glucocorticoids* regulate gluconeogenesis from amino acids, and serve as stress-adaption hormones. Their medical importance is due to the strong antiinflammatory and immunosuppressive potential. The glucocorticoids are synthesized in the zona fasciculata of the adrenal cortex and their diurnal pattern of secretion is closely linked to that of ACTH (corticotropin). Feedback between the adrenal steroids and the pituitary secretion of ACTH is very effective. During long-term therapy with synthetic glucocorticoids, endogenous ACTH secretion is often suppressed, and may result in reversible adrenal involution or irreversible adrenal atrophy. At higher doses, the glucocorticoids suppress the tissue reaction to inflammation, reduce cell division in the bone marrow and lymphopoetic tissues (thymus gland and lymph nodes), and acutely inhibit exudation by reducing capillary permeability. This pharmacological activity is of importance in the treatment of acute kidney disease because it suppresses pathological permeability of the glomerular filtration system for proteins, and prevents progressive loss of albumin. Glucocorticoids have catabolic and antianabolic activities, they enhance protein catabolism and gluconeogenesis, and mobilize calcium by lowering the anabolic stimulation of bone formation anc bone mineralization. Under glucocorticoid therapy, calcium is lost from bones resulting in steroid-induced osteoporosis.

The *mineralocorticoids* regulate water and electrolyte metabolism by enhancing the renal reabsorption of sodium and water, and the excretion of potassium and hydrogen ions. The most important mineralocorticoid in the human is aldosterone (sodium-retaining hormone), which increases sodium reabsorption in the distal tubule of the kidney. Together with vasopressin it also regulates the osmolarity and ionic composition of the intravascular space. Enhanced secretion of aldosterone by adrenal tumors may lead to hypertension. In cardiovascular disease and hepatic failure, circulating aldosterone may be elevated (secondary aldosteronism) stimulating the formation of edema due to pathological sodium retention. Specific receptor antagonists of aldosterone (spironolactones) are used in therapy to counteract the sodium-retaining effect [9].

1.4.3. Calciferols

The calciferols (calcium-regulating sterols) are secosteroids with an opened ring B (between C 9 and C 10) derived from 7-dehydrocholesterol or ergosterol. They have a significantly delayed

onset of action because they induce intestinal calcium-binding proteins [80] by interaction with the genome, they also metabolize calcium and phosphate from the bone to provide a higher concentration in the extracellular fluid for osteoid mineralization.

1.4.4. New Synthetic Steroid Hormones

Progress in the development of synthetic steroid hormone analogues has resulted in new possibilities for the specific control of reproductive function. Receptor antagonists of progesterone have been developed for clinical studies on postcoital contraception by reducing or blocking corpus luteum formation [9]. These compounds also inhibit the action of cortisol and may stimulate the immune system by neutralizing the immunosuppressive effects of glucocorticoids.

1.5. Prostaglandins

The prostaglandins are endogenous compounds with hormone-like activities that all have the prostanoic acid skeleton and share arachidonic acid as their common precursor [29,30]. They influence the regulation of blood flow, kidney function, and hormone secretion. For a detailed description, see →Prostaglandins.

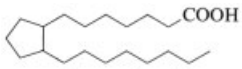

Prostanoic acid

Prostaglandin analogues of the E- and F-series have marked biological activities on reproductive functions, they lower progesterone secretion from the corpus luteum, and increase uterine contractility during pregnancy, and at parturition. Their multiple effects and sites of action make the clear assignment of classical hormonal functions very difficult, they should therefore be viewed as hormone candidates. In contrast to many peptides, proteins, and steroids which can be administered systemically without major tolerance problems, the prostaglandins are more difficult to handle because of their wide spectrum of activities that often includes severe side effects.

1.6. Hormone Antagonists and Analogues

Functional or competitive hormone antagonists (antihormones) can be divided into several groups with different modes of action [9]. The competitive receptor antagonists displace hormones from their receptors, and prevent physiological receptor activation by endogenous hormones. Such antagonists may be structurally related to the natural ligand (e.g., LHRH antagonists), or structurally unrelated, but of similar three-dimensional conformation (e.g., nonsteroidal antiandrogens). The noncompetitive inhibitors of hormone secretion often interfere with biosynthesis. They can act at a multitude of sites, as shown by the antithyroid drugs [75], the inhibitors of adrenal steroid secretion, aminoglutethimid and mitotan [9], and the inhibitor of gonadal and adrenal steroid biosynthesis, ketoconazole [9]. Future developments in drug design will be based on computer modeling of the active sites of hormone receptors, and of the three-dimensional conformation of synthetic hormone analogues capable of interacting with these receptors. Examples for nonsteroidal compounds interacting with steroid receptors are numerous, whereas similar developments are rare in the peptide and protein field, with the exception of structural similarities between opiates and endogenous opioid peptides [81], the encephalins and endorphins. There are until now no examples of nonpeptide drugs capable of blocking peptide receptor binding.

The concept of using natural hormones as lead compounds for synthetic modification by structure – activity studies has been particularly successful. Gonadal and adrenal steroids have been modified by chemical synthesis to obtain more potent compounds with enhanced metabolic resistance, specific receptor antagonists with selective inhibitory potential, and nonsteroidal compounds with simplified structures. The development of new steroidal and nonsteroidal molecules with improved biological and therapeutic characteristics is still in progress; there are indications for selective antiprogestins and antiglucocorticoids [9]. In the field of peptides and proteins, significant progress is being made by the widespread application of molecular biology and genetic engineering. Structural

analysis of polypeptides and proteins by gene sequencing has afforded many new lead compounds which can be studied by recombinant DNA synthesis in bacteria, yeast, or mammalian cell lines [2]. The availability of these biosynthetic molecular probes may initiate structure–activity studies on highly selective and specific synthetic peptide hormones. The receptor binding of new synthetic hormone derivatives, and an understanding of the mechanisms of action of natural hormones has been crucial for a more effective approach to drug design. Homogeneous receptor proteins for peptides and steroids are sequenced and synthesized in genetic engineering. The understanding of the molecular basis of cellular activation is increasing, and the mechanisms of action of peptide, protein, and steroid hormones are becoming more transparent.

2. Amino Acid Hormones

2.1. Thyroid Hormones

The thyroid hormones are produced in the thyroid gland, an endocrine gland located in front of the trachea and below the larynx, which consists of two lobes, connected by an isthmus. The thyroid gland stores ingested iodine in the follicular cells and secretes the thyroid hormones thyroxine (L-tetraiodothyronine) and triiodothyronine into the blood, as required. These two iodinated amino acid derivatives affect metabolism in a variety of ways. They not only regulate energy turnover, water balance, and the metabolism of carbohydrates, proteins, fats, and minerals, but also affect growth and maturation.

In addition, the parafollicular C cells of the thyroid produce another type of hormone, calcitonin, a peptide hormone, which participates in the regulation of calcium metabolism. However, the exact physiological function of this hormone is not yet known. Another peptide hormone, parathormone, which also influences calcium metabolism, is secreted by the parathyroid glands.

2.1.1. Physiology and Therapeutic Uses

Biosynthesis of Thyroid Hormones. The healthy thyroid gland takes up approximately 75 µg of iodine per day, provided enough dietary iodine is available (150 – 300 µg/d). Iodine is used to iodinate the tyrosine residues of a thyroid-specific protein *thyroglobulin*. The resulting 3,5-diiodotyrosine groups in the peptide chain of thyroglobulin react with each other or 3-monoiodotyrosine to form thyroxine (3,5,3′,5′-tetraiodothyronine) or triiodothyronine, respectively. This thyroxine- and triiodothyronine-containing thyroglobulin is stored in the colloid within the thyroid follicle. The normal human thyroid contains approximately 10 mg of iodine in this form. During secretion, thyroglobulin leaves the colloid and is hydrolyzed by the surrounding follicular cells. The liberated hormones, thyroxine (T_4) and triiodothyronine (T_3), are released into the blood at rates of ca. 100 and 8 µg/d, respectively.

The thyrotropic hormone of the pituitary hypophysis (thyrotropin or thyroid-stimulating hormone, TSH) stimulates the secretory function of the thyroid gland. The hypophyseal secretion of TSH, in turn, is stimulated by the hypothalamic hormone TRH (thyrotropin-releasing hormone; see also →Peptide and Protein Hormones) and inhibited by elevated levels of blood T_4 or T_3 in a feedback mechanism. In the blood, T_4 and T_3 are bound to serum proteins, principally thyroxine-binding globulin. Only 0.03 % of T_4 and 0.3 % of T_3 are present in the free (physiologically active) form. The biological half-life of T_3 is only ca. 19 h, whereas that of T_4 is ca. 190 h. Further formation of T_3 (approximately 30 µg/d) occurs in the peripheral tissues as a result of the deiodination of T_4. Because T_3 has approximately ten times the biological activity of T_4, thyroxine is considered to be the prohormone of the biologically more active triiodothyronine [87]. In addition to T_3, a second isomeric iodoamino acid, reverse T_3, (L-3,3′,5′-triiodothyronine, rT_3) is formed from T_4 in the periphery but has no apparent physiological activity [88].

Effect of Thyroid Hormones. Free thyroxine and especially free triiodothyronine probably bind to the nuclear and mitochondrial receptors, thereby activating protein synthesis or adenosine triphosphate production, respectively. The primary effects of the thyroid hormones are

1) stimulation of carbohydrate turnover, growth, and maturation in the central nervous system, skeletal and muscular systems, and genital organs; stimulation of heat production, oxygen consumption, cholesterol metabolism, turnover of free fatty acids, muscle contraction, heart rate, and cardiac output; and
2) inhibition of glycogen and protein synthesis (only at nonphysiological concentrations).

The amount of thyroid hormones required by the human organism varies between 50 and 300 µg/d, depending partly on the individual metabolism.

Thyroid Diseases. Thyroid diseases can affect the morphology (goiter, nodules, carcinoma) or the function (hyperthyroidism, hypothyroidism) of the thyroid. In many cases, thyroidal autoantibodies or iodine deficiency is the cause of the thyroid disorder. Hyperthyroidism is characterized by tachycardia, sweating, weight loss, diarrhea, etc. Hypothyroid children show abnormal physical and mental development. Hypothyroid adults tend to react slowly or to be depressed; additional symptoms are dry skin, obstipation, and sensitivity to cold.

Diagnosis. Diagnosis of thyroid disease is based not only on clinical symptoms but also on the measurement of serum concentrations of T_4, T_3, and TSH (basal or stimulated by the addition of TRH, i.e., the TRH test) and, if necessary, on thyroid uptake, iodine turnover, and the functioning of the hypophysis – thyroid feedback control system.

Therapy with Thyroid Hormones. (See also →Thyrotherapeutic Agents). Diseases of the thyroid gland that involve a thyroid hormone deficiency can be treated easily and successfully by oral administration of thyroid hormones. Treatment with levothyroxine is preferred because of the long half-life of this hormone. The sodium salt of levothyroxine in tablet form is absorbed to the extent of 40 – 60 % [89, 90], or 70 – 85 %, depending on the galenic preparation [91]. The daily dose of levothyroxine varies between 50 and 300 µg, depending on the severity of thyroid deficiency. The daily dose of liothyronine is correspondingly lower (10 – 100 µg). Injectable thyroid hormone solutions are used in cases of severe hypothyroidism (myxedema coma). Preparations containing both thyroid hormones are also available. Treatment with thyroid gland powder is no longer recommended because the standardization and stability of these preparations are problematic.

Therapy with thyroid hormones is sometimes recommended in certain types of obesity. Both dextrothyroxine (4 – 10 mg/d) and dextrotriiodothyronine (0.5 – 2 mg/d) are used in the treatment of disorders of fat metabolism, particularly hypercholesterolemia. L-Diiodotyrosine is also administered clinically; however, this preparation functions primarily as an iodine carrier because it is rapidly metabolized.

A number of analogues and derivatives of thyroid hormones have been synthesized and studied. Although substances with many times the thyromimetic activity of levothyroxine have been produced [92], none has found practical use. Triiodothyroacetic acid [*51-24-1*] (TRIAC), a degradation product of the thyroid hormones, occurs only in small amounts in the human body and exhibits a TSH-suppressing effect without the peripheral effects of levothyroxine or liothyronine. This compound has been successfully used in the treatment of hyperthyroidism caused by hypophyseal TSH-producing tumors [93].

2.1.2. Physical Properties

Levothyroxine [*51-48-9*] (L- 3,5,3′,5′-tetraiodothyronine, T_4), $C_{15}H_{11}O_4I_4N$, M_r 776.87, *mp* 235 – 236 °C (decomp.), $[\alpha]_D$ +20.4° (2 % solution in a 1 : 4 mixture of HCl and ethanol), $[\alpha]_D$ −5.8° (3 % solution in a 1 : 2 mixture of 1 mol/L NaOH and ethanol), forms colorless, light-sensitive crystals. It is sparingly soluble in water, slightly soluble in methanol, and readily soluble in alkali or dilute mineral acids.

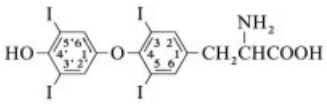

The sodium salt of levothyroxine [*25416-65-3*] crystallizes from a hot solution of the free acid in sodium carbonate: $C_{15}H_{10}O_4I_4NNa \cdot 5\,H_2O$, M_r 888.94 [94, p. 476], [95, p. 642].

Liothyronine [*6893-02-3*] (L- 3,5,3′-triiodothyronine, T_3), $C_{15}H_{12}O_4I_3N$, M_r 650.98, *mp* 233–234 °C, $[\alpha]_D^{24}$ +23.6° (5 % solution in a 1 : 2 mixture of 1 mol/L HCl and ethanol), forms colorless crystals. It is slightly more water-soluble than thyroxine and readily soluble in dilute alkali [94, p. 268], [95, p. 643].

Reverse T_3 [*5817-39-0*] (L-3,3′,5′-triiodothyronine, rT_3) $C_{15}H_{12}O_4I_3N$, *mp* 206 °C (decomp.), $[\alpha]_D^{21}$ +16.7° (1 % solution in a 1 : 1 mixture of 1 mol/L HCl and ethanol), forms colorless needles [96].

l-3,5-Diiodothyronine [*1041-01-6*] (T_2), $C_{15}H_{13}O_4I_2N$, M_r 525.08, has the following properties: *mp* 254 °C (decomp.), $[\alpha]_D^{25}$ +25.2° (5 % solution in a 1 : 2 mixture of 1 mol/L HCl and 95 % ethanol).

l-3,5-Diiodotyrosine [*66-02-4*] (DIT, iodogorgoic acid), $C_9H_9O_3I_2N \cdot 2\,H_2O$, M_r 469.02, *mp* 204° (decomp.), $[\alpha]_D$ +2.75° (4.8 % solution in 1 mol/L HCl [97]), has the following structure:

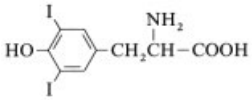

l-3-Iodotyrosine [*70-78-0*] (monoiodotyrosine, MIT), $C_9H_{10}O_3IN$, M_r 307.09, has the following properties: *mp* 204–206 °C, $[\alpha]_D^{22}$ −8.8° (4 % solution in 1 mol/L HCl).

The corresponding D-isomers of the above iodoamino acids are also known. Dextrothyroxine is used therapeutically to lower cholesterol levels.

2.1.3. Production

In the thyroxin synthesis, first performed in 1927, 4-(4-methoxyphenoxy)-3,5-diiodonitrobenzene was converted into the corresponding aldehyde and then into D,L-diiodothyronine, which was iodinated to yield thyroxine [98]. This method was used commercially for a long time and produced a racemate. Resolution of the racemates of thyroxine has not yet been achieved.

The natural hormones levothyroxine and liothyronine could be produced on a large scale from protected L-3,5-dinitrotyrosine by inserting the iodine substituents after formation of the ether bond [99]. This synthesis also led to the production of D-T_4 and D-T_3. However, optically pure preparations could not be obtained easily.

A method adapted from the biological synthesis was used for some time to produce levothyroxine. Two molecules of L-diiodotyrosine are coupled at pH 9.5 in the presence of manganese dioxide at 60 °C. Levothyroxine is produced with a yield of 2.8 %. The reaction of diiodotyrosine with 4-hydroxy-3,5-diiodo-phenylpyruvic acid proceeds similarly except that yields up to 50 % can be obtained [100, 101].

A total synthesis of T_4 and T_3 was developed by Hoechst in 1963 and involves racemate resolution at the level of diiodothyronine [102, 103]. The reaction steps are shown below:

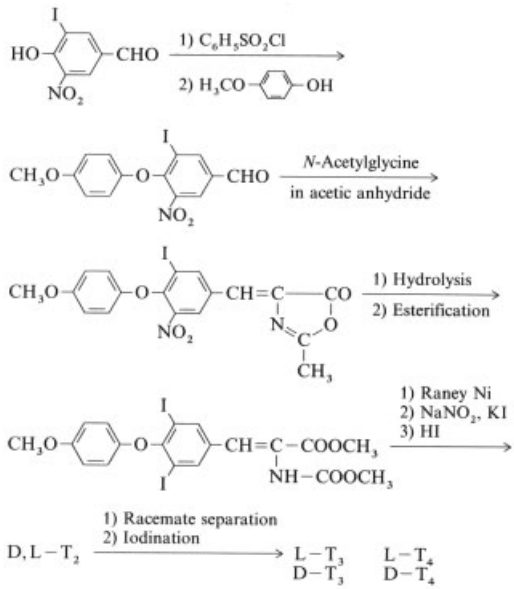

Most other methods in the large-scale production of thyroid hormones are modifications of the procedure described by HILLMANN in 1956 [104, 105]. The ether bond in thyronine is synthesized by using diaryliodonium salts and the only slightly reactive diiodotyrosine (see top of next page) [106, 107].

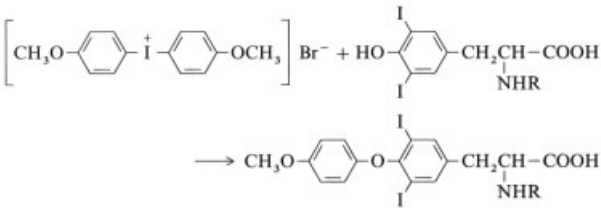

L-Thyronine is made from L-tyrosine, obtained by protein hydrolysis (e.g., of wool). A number of methods are available for obtaining T_3 and T_4 by subsequent iodination [102, 104, 108]. The arylation of diiodotyrosine with iodonium salts permits the recovery of optically pure levo- and dextrothyroxine preparations. Especially in the case of dextrothyroxine, which is used therapeutically to lower cholesterol levels, the amount of L-isomer should not exceed 0.2 % [109]. Chemically pure T_4 and T_3 preparations are obtained by preparative column chromatography [110]. Analogues of the thyroid hormones and ^{14}C-labeled preparations are also produced by the above methods [111–113].

Analysis and Testing. The following methods are used to test the identity and purity of iodoamino acids: iodine determination; UV absorption [114]; paper [115], thin-layer [116], and gas chromatography [117], polarography [118]; enzymatic analysis [109]; and radioimmunoassays (→Diagnostic Reagents, Chap. 2.5.→Diagnostic Reagents, Chap. 2.6.→Diagnostic Reagents, Chap. 3.2.).

Producers of thyroid hormones in Germany and Austria are Henning (Berlin), Hoechst (Frankfurt), and Sanabo (Vienna).

Trade names are listed in Table 1, see right.

Table 1. Selected trade names of thyroid hormone preparations

Trade name	Producer	Country
Sodium levothyroxine		
Eferox	Efeka	FRG
Eltroxin	Glaxo	Denmark, Great Britain, Switzerland
Euthyrox	Merck	Austria, Belgium, FRG
L-Thyroxin Henning	Henning Berlin	FRG
L-Thyroxine Roche	Roche	France
Levoroxine	Bariatric	USA
Levothroid	USV	USA
Levothyrox	Merck-Clevenot	France
Levothyroxine Sodium	Armour-Yamanouchi	Japan
Synthroid	Flint	USA
Thevier	Glaxo	FRG
Thyrex	Sanabo	Austria
Sodium liothyronine		
Cynomel	Uhlmann-Eyrand	Switzerland
Cynomel	Merrell	France
Tertroxin	Glaxo	Denmark, Great Britain
Thybon	Hoechst	FRG
Combination preparations of sodium levothyroxine and sodium liothyronine		
Novothyral	Merck	FRG
Prothyrid	Henning Berlin	FRG
Thyroxin-T_3	Henning Berlin	FRG
Combination preparation of sodium levothyroxine and iodine		
Jodthyrox	Merck	FRG
Injection solutions (sodium levothyroxine)		
L-Thyroxin inject Henning	Henning Berlin	FRG
Synthroid	Flint	USA
Injection solutions (sodium liothyronine)		
Tertroxin	Glaxo	Denmark
Thyrotardin inject	Henning Berlin	FRG

2.2. Hormones of the Adrenal Medulla

The hormones of the adrenal medulla are the catecholamines *adrenaline* (epinephrine) and *noradrenaline* (norepinephrine). The catecholamines act on α and β-adrenoreceptors in the central nervous system, in the autonomous nervous system, and in other tissues (Figure see right) [119].

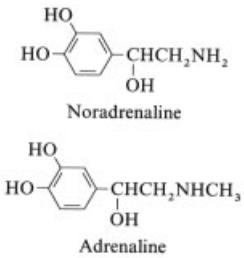

They also modify the metabolism of many cells [120]. Both hormones are synthesized in the adrenal medulla. After transsection of the human adrenal gland, the central adrenal medulla is clearly visible because it is darker than the surrounding adrenal cortex. The adrenal medulla and the adrenal cortex are regulated by different mechanisms, the adrenal cortex being under hormonal control, whereas the adrenal medulla is under nervous control. The blood-pressure increasing effect of an adrenal extract was described in 1985 by OLIVER and SCHÄFER [121]. The first chemical synthesis of adrenaline was performed in 1904 by STOLZ, the definitive elucidation of its structure by FRIEDMANN was in 1906.

Normally only noradrenaline is secreted; the medulla secretion of adrenaline is restricted to acute stressful situations which require brief, maximal activation of cadiovascular function, and enhanced central nervous tone. Noradrenaline administered intravenously or subcutaneously has a predominantly peripheral action on blood pressure and cardiovascular function. After subcutaneous injection, absorption is much delayed due to potent vasoconstriction. The adrenal medulla also produces *dopamine*, a precursor of noradrenaline.

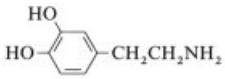

Dopamine

Low concentrations of dopamine have therapeutic effects on specific dopamine receptors in the kidney, increasing the glomerular filtration rate in shock syndromes [122]. At increasing concentrations, dopamine stimulates α-adrenoreceptors, it is therefore administered by infusion at a controlled rate to avoid undesirable vasoconstriction. A lack of dopamine is also responsible for Parkinson's disease (→Parkinsonism Treatment).

The catecholamines have major endogenous functions as synaptic transmitters in the central regulation of blood pressure [122]. Noradrenaline stored in synaptic vesicles is released from the presynaptic nerve endings and acts on receptors in the postsynaptic membrane. Its local concentration in the synapse is controlled by reuptake mechanisms. Noradrenaline also acts as a neurotransmitter in the autonomous nervous system.

Noradrenaline [*51-41-2*], levarterenol, norepinephrine, R (−)- or D (−)-2-amino-1- (3,4-dihydroxyphenyl)ethanol; $C_8H_{11}NO_3$, M_r 169.2, *mp* 217 – 218 °C (decomp.), sparingly soluble in water and ethanol;

Hydrochloride: $C_8H_{11}NO_3 \cdot HCl$, M_r 205.6, *mp* 145 – 146 °C, readily soluble in water, α_D^{25} − 40° (6 % aqueous solution);

Hydrogen tartrate: $C_8H_{11}NO_3 \cdot C_4H_6O_6$, M_r 319.3, *mp* 158 – 159 °C, α_D^{25} − 10.7° (1.6 % aqueous solution).

Adrenaline [*329-65-7*], suprarenin, epinephrine, R (−)- or D (−)-1- (3,4-dihydroxyphenyl)-2-methylaminoethanol $C_9H_{13}NO_3$, M_r 183.2, *mp* 211 – 212 °C, α_D^{25} − 50 – 53° (5 % in 0.5 mol/L HCl);

Hydrogen tartrate: $C_9H_{13}NO_3 \cdot C_4H_6O_6$, M_r 333.3, *mp* 149 – 152 °C, α_D^{25} − 14.5°

Hydrochloride: $C_9H_{13}NO_3 \cdot HCl$, M_r 219.7, *mp* 206 °C.

The levorotatory isomers of both noradrenaline and adrenaline are 15 – 40 times more potent in several test systems. This is also a general characteristic of many other catecholamines.

Synthesis. Both compounds are obtained from 4-chloroacetylcatechol which is prepared from catechol by reaction with chloroacetic acid and $POCl_3$.

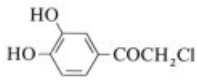

For the preparation of adrenaline, 4-chloroacetylcatechol is reacted with methylamine, the keto group is converted to the hydroxy group by catalytic hydration and the racemic adrenaline base thus obtained is cleaved with tartaric acid. For noradrenaline, the 4-chloroacetylcatechol is reacted with ammonia.

Therapeutic Uses. Adrenaline and noradrenaline have limited therapeutic importance,

because their duration of action is short and therapy of hypotension would require administration by infusion. Their rapid rate of disappearance after presynaptic release is a physiological prerequisite for rapid synaptic transmission. However, it has promoted the development of numerous synthetic catecholamine derivatives and indirect sympathomimetics [123] that act on the stored neurotransmitters in synaptic vesicles (→Antihypotensives; →Sympatholytics and Sympathomimetics).

Many catecholamine derivatives and sympathomimetic compounds have been developed in an attempt to obtain specific stimulation of α-adrenoreceptors and β(2)-adrenoreceptors involved in smooth muscle relaxation, e.g. as bronchodilatory compounds [123]. Research on nonselective β-adrenoreceptor blockers and selective β-blocking agents has been very successful [124] (see also → β-Receptor Blocking Agents). The indirect sympathomimetics displace catecholamines from their storage vesicles and thus temporarily enhance postsynaptic stimulation [125]. This effect is abolished if the storage vesicles are first depleted of catecholamines by administration of reserpine [126].

Clinical Uses. Adrenaline has important clinical applications in internal medicine, surgery, and dental medicine. It provides rapid relief in hypersensitivity reactions caused by anaphylactic and allergic reactions (by injection or infusion), in cardiogenic shock caused by myocardial infarction, and relieves respiratory distress due to bronchospasm when administered by oral inhalation. In surgery and dental medicine, adrenaline is used as a hemostatic agent; it is administered by infiltration or locally application on bleeding surfaces. In combination with local anesthetics it prolongs an anesthesia by preventing diffusion of the anesthetic from the site of injection. Adrenaline acts as a bronchodilatory and antiallergic compound, because it prevents the release of histamine and other mediators of inflammation [126]. Noradrenaline and related synthetic sympathomimetic amines that act predominantly on α-adrenoreceptors have similar clinical uses to adrenaline.

3. Steroid Sex Hormones (→ Steroids)

3.1. Introduction

The sex hormones, like the hormones of the adrenal cortex, are steroids with the following general formula:

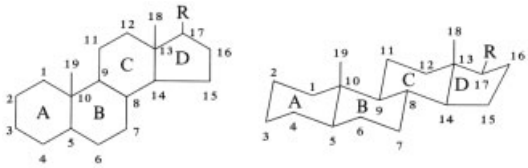

All substituents below the plane of the ring system are designated as α and those above the plane of the molecule as β. These orientations can be indicated with a dashed or solid line, respectively. The substituent R may be an alkyl residue or an oxygen function (for further details on steroid structures, see [127, 128]). These compounds are called "steroids" because they bear a great structural resemblance to the naturally occurring sterols [127]. The sex hormones are classified as

1) *androgens* (male sex hormones, e.g., testosterone);
2) *estrogens* (female sex hormones, e.g., estrone, estradiol, estriol); and
3) *gestagens* (female sex hormones, e.g., progesterone).

Naturally occurring steroids have several asymmetric carbon atoms (centers of chirality) and are optically active. They are defined as belonging to the "normal series" [127, 128]. Synthetic steroids are considered to belong to the normal series when C-10 or, if this position is not asymmetric (as in the estrogens), C-13 has the same configuration as in cholesterol. If all the centers of asymmetry in the steroid molecule have the opposite configuration to normal (i.e., mirror image of the normal form), the compound carries the prefix "ent" (enantiomer). A racemate is an equimolar mixture of the normal and the enantiomeric forms of a steroid molecule and is indicated by the prefix "rac" [128].

Of the large number of possible stereoisomers of sex hormones, only the isomer with the normal (natural) configuration is active (in this respect, estrogen is the simplest sex hormone;

it has four centers of asymmetry, 16 stereoisomers, and 8 racemates). All other stereoisomers either are less active, are inactive, or exert other effects. Cholesterol is the principal mammalian sterol and is biosynthesized from 18 molecules of acetyl coenzyme A (acetyl-CoA) by a complex route involving consecutive formation of mevalonic acid, farnesyl pyrophosphate, squalene, lanosterol, and cholesterol. For further details, see [133, pp. 20 – 70].

Cholesterol is the biochemical precursor of both the sexual and the adrenal steroid hormones. Figure 4 briefly summarizes the most important steps of cholesterol transformation, as well as the principal enzymes involved.

Commercial methods used in the production of sex hormones or their precursors can be divided into four groups:

1) *Partial chemical synthesis* of steroid hormones from natural materials, that have the

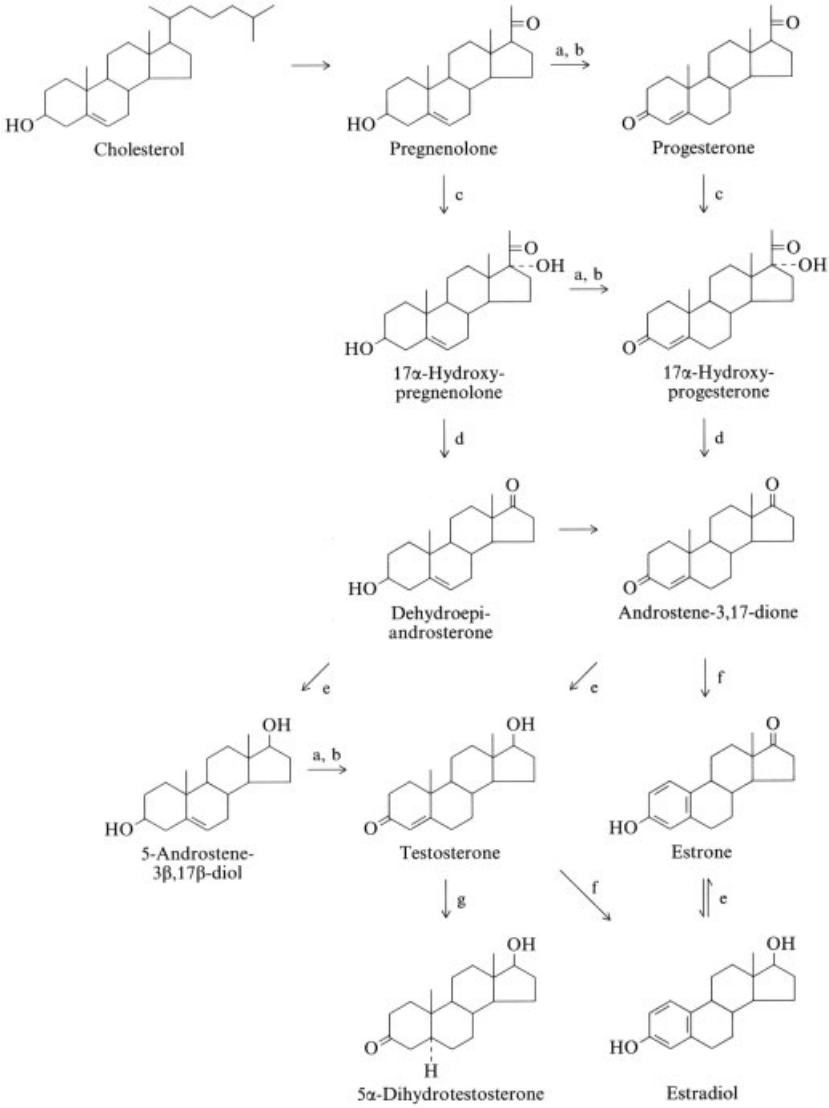

Figure 4. Biosynthesis of steroid sex hormones [130, 133]
Enzymes: a) 3β-Hydroxy-Δ^5-steroid dehydrogenase; b) Steroid-Δ-isomerase; c) Steroid-17α-monooxygenase; d) C-17,20-Lyase; e) 17β-Hydroxysteroid dehydrogenase; f) Aromatase; g) 5α-Reductase

steroid nucleus (e.g., sterols, saposterols), was the first method used for commercial production of sex hormones; it still has considerable economic importance.

2) *Microbiological degradation* of sterols, such as cholesterol, stigmasterol, ergosterol, and β-sitosterol, which cannot otherwise be efficiently converted into steroid intermediates, has rapidly become commercially important [141–144].
3) The *total synthesis* of steroid hormones involves synthesizing the sterically complex steroid skeleton from inexpensive starting substances. This method is only important for steroids that do not occur in nature (see Section 3.4) [134–136].
4) The *extraction* of sex hormones or their conjugates from organic tissue or urine is the oldest method for isolating therapeutically important sex hormones. However, it is also the most unproductive because the hormones occur in very low concentration (10 mg of estrone can be isolated from 1 L of mare urine and 1 mg of progesterone from 35 g of pig corpus luteum). Today, this method is only used to recover a mixture of estrogen conjugates from the urine of pregnant mares, which is required for the menopause preparation Presomen. However, hormone extraction is still a very important analytical tool, e.g., in the identification and isolation of metabolites of sex hormones [132].

3.2. Partial Chemical Synthesis of Steroid Intermediates

3.2.1. Raw Materials

Raw materials for the partial chemical synthesis of steroid intermediates must fulfill the following requirements:

1) they must contain the steroid nucleus,
2) they must belong to the normal series,
3) they must occur in nature in sufficient quantities, and
4) they must be easily convertible into sex hormones by simple chemical reactions.

The most important raw materials are listed in Table 2 [146]. All have a hydroxyl group at position 3 and a double bond between C-5 and C-6. These common structural elements readily permit the introduction of an unsaturated keto group in ring A (e.g., by Oppenauer oxidation). These four compounds differ from each other with respect to the side chain at C-17; cholesterol and β-sitosterol possess a saturated side chain that is not easily accessible to chemical attack and subsequent modification. Stigmasterol has a double bond in its side chain and is thus more susceptible to chemical modification. Diosgenin contains a spiroketal side chain which not only makes it easy to modify, but also permits insertion of a double bond between C-16 and C-17.

Table 2. Starting substances for the partial synthesis of steroid intermediates

Name and natural source	Formula
Diosgenin [*512-04-9*] (**1**): Dioscoreaceae	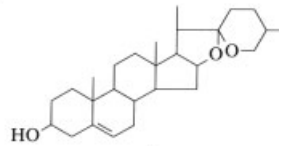 **1**
Stigmasterol [*83-48-7*] (**2**): soybean oil	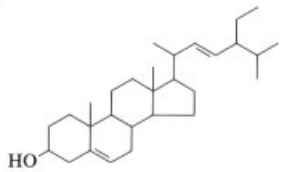 **2**
β-Sitosterol [*83-46-5*] (**3**): wheat germ oil, soybean oil, sugarcane wax	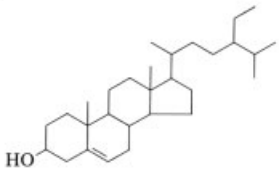 **3**
Cholesterol [*57-88-5*] (**4**): wool grease, spinal cord, brain	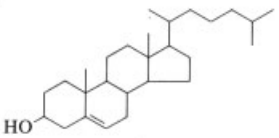 **4**

The steroid starting materials and their derivatives obtained by chemical or microbiological modification are shown in Figures 5 and

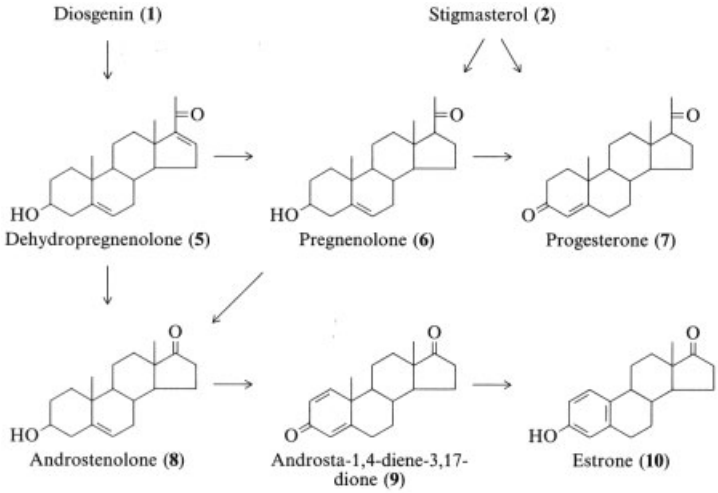

Figure 5. Steroid starting materials and their chemically synthesized derivatives

6, respectively. The derivatives, in turn, are starting materials for the synthesis of different series of sex hormones (see Table 3). Almost all steroid hormones belonging to the pregnane, (21 carbon atoms in the steroid skeleton), androstane (19 carbon atoms), and estrane (18 carbon atoms) series can be obtained by chemical modification of diosgenin (**1**) and stigmasterol (**2**). Microbiological modification of cholesterol (**4**), β-sitosterol (**3**), and stigmasterol (**2**) and subsequent chemical conversions have only yielded androstanes and estranes [141–144].

Table 3. Steroid derivatives for the synthesis of steroid hormones

Steroid derivative	Hormone product
Dehydropregnenolone (**5**), progesterone (**7**)	1) Gestagens of the 17α-hydroxyprogesterone series 2) Adrenal cortex hormones (corticoids)
Adrostenolone (**8**)	1) Androgens 2) Gestagens of the 17α-ethynyltestosterone series 3) Anabolic androstane steroids
Estrone (**10**)	1) Estrogens 2) Gestagens of the 17α-ethynyl-19-nortestosterone series 3) Anabolic 19-nortestosterone steroids

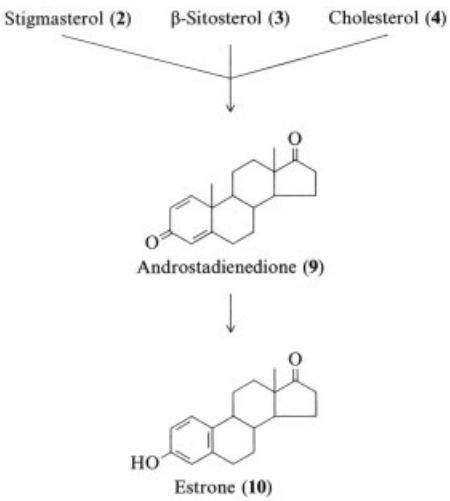

Figure 6. Steroid starting materials and their microbiologically synthesized derivatives

Cholesterol [*57-88-5*], $C_{27}H_{46}O$, M_r 386.6, *mp* 148 °C (for structure see Table 2), was an important starting material in the early industrial production of sex hormones. Oxidation of cholesterol or its 5,6-dibromide derivative with chromic acid gave androstenolone acetate [127]; maximum yields were 7 – 8 %. Today, both cholesterol and β-sitosterol have little importance as starting materials for chemical syntheses. However, they are all more important as raw materials for syntheses involving microbiological degradation of the side chain (see Section 3.3) [141–144].

Cholesterol occurs in animal tissue at the following concentrations, in weight percent:

Skeletal muscle	0.09 – 0.12
Liver	0.32 – 0.61
Brain (calf)	2.11 – 2.36
Kidney	0.28
Egg yolk	2 – 3.9

Spinal cord (calf)	4
Wool fat	15

It can be recovered from these sources by extraction with chlorinated hydrocarbons, if necessary, after saponification of cholesterol esters. Purification is difficult because cholesterol is often isolated with lanosterol and other triterpenes [145].

Stigmasterol [*83-48-7*], $C_{29}H_{48}O$, M_r 412.7, *mp* 170 °C, and **β-sitosterol** [*83-46-5*], $C_{29}H_{50}O$, M_r 414.7, *mp* 140 °C (for structures see Table 2), are two additional steroid starting materials. After diosgenin, stigmasterol is the most important raw material for steroid hormone synthesis. As a result of the double bond between C-22 and C-23 in the side chain, chemical modification of stigmasterol yields either pregnenolone [145], via bisnorcholic acid, or progesterone [127], via bisnorcholic aldehyde. Both compounds are key substances in the production of pregnane, androstane, and estrane hormones. Microbiological degradation of stigmasterol and β-sitosterol yields androstadienedione (see Fig. 6).

Stigmasterol occurs with β-sitosterol at a concentration of 12 – 25 % in the nonsaponifiable fraction of soybean oil. The sterols are extracted as a mixture, but their similarity makes the purification and isolation of stigmasterol extremely difficult. Thus, stigmasterol is not a starting material for chemical syntheses. However, microbiological degradation can be carried out conveniently with this sterol mixture.

Diosgenin [*512-04-9*], $C_{27}H_{42}O_3$, M_r414.6, *mp* 208 °C (for structure see Table 2), is a steroid sapogenin. When compared with the sterols mentioned above, diosgenin is the most suitable candidate as a starting material for the production of steroids and sex hormones [127]. Not until the discovery of this compound and the elaboration of a commercially feasible method for degradation of the spiroketal side chain, to give dehydropregnenolone (**5**) (see Fig. 5), was the large-scale industrial production of sex hormones possible. Almost all steroid hormones, including those of the adrenal cortex, can be derived from dehydropregnenolone (see Fig. 5). Since the early 1950s diosgenin has become the most important steroid raw material. Its discovery gave a tremendous technological boost to the production of sex hormones. The price of progesterone dropped within three years from $ 80 to $ 3 per gram, and sufficient amounts of sex hormones became available for medical use.

Diosgenin occurs as a 3-glycoside (dioscin) [*19057-60-4*] in numerous *Liliaceae* and *Dioscoreaceae* species and is extracted from the air-dried rhizomes of *D. mexicana, D. floribunda,* and *D. composita* (also called barbasco) with ethanol [127, 145]. The extracts are concentrated by evaporation, glycoside cleavage is achieved by heating with dilute hydrochloric or sulfuric acid, and crude diosgenin is isolated by filtration. This raw material can be purified by crystallization from alcohol; however, it is usually employed without previous purification in the production of dehydropregnenolone (see Section 3.2.2).

3.2.2. Partial Synthesis

This section describes only the synthesis of intermediates having special importance for the commercial production of sex hormones used as active ingredients in drugs.

Batch synthesis of the active substances is still prevalent; however, both automation by programmed monitoring and semicontinuous or continuous methods are becoming increasingly important.

Synthesis of 3β-Acetoxydehydropregnenolone. 3β-Acetoxydehydropregnenolone [*979-02-2*], also called dehydropregnenolone acetate(**5b**), is synthesized from diosgenin [127, 145] by a method first developed by Marker and coworkers and later improved upon by several teams. Diosgenin acetate is heated to 190 – 200 °C with acetic anhydride or a homologous compound and is converted to pseudodiosgenin diacetate which, in turn, on mild oxidation with chromic acid in glacial acetic acid yields the ketone diosone. Hydrolysis with dilute alkali gives dehydropregnenolone or its acetate (**5b**), depending on reaction conditions, with an overall yield of ca. 45 % (see first figure on next page).

Dehydropregnenolone is the key substance in the synthesis of all the steroid and sex hormones derived from diosgenin (see Fig. 5).

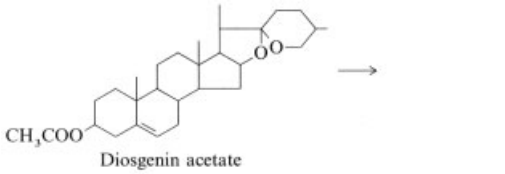

Diosgenin acetate

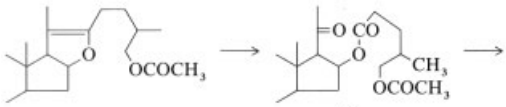

Diosone

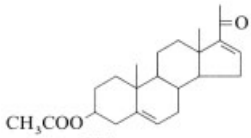

3β-Acetoxy-dehydropregnenolone (**5b**)

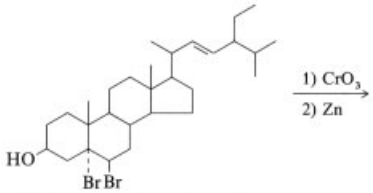

5,6-Dibromostigmasterol (**11**)

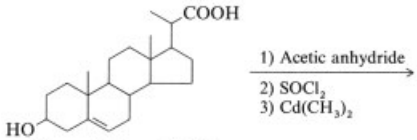

Bisnorcholic acid (**12**)

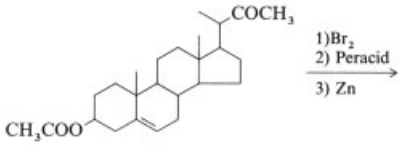

13

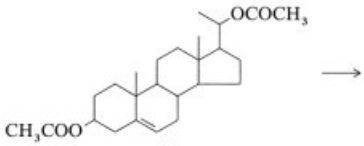

14

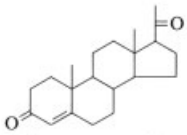

Progesterone (**7**)

Synthesis of Progesterone. The natural gestagen progesterone [*57-83-0*] (**7**) is used as an active ingredient in gestagen preparations (see Section 3.7.2), and is also an important intermediate in the production of adrenal cortical hormones. Two methods are available for the preparation of progesterone:

Selective hydrogenation of dehydropregnenolone acetate (**5b**) with palladium charcoal results in the retention of the Δ^5 double bond and the formation of pregnenolone acetate (**6b**). This compound is saponified and then oxidized (Oppenauer) in benzene or toluene to progesterone [127, 145] by refluxing with cyclohexanone and aluminum isopropylate. Yields of ca. 85 % are obtained.

The *conversion of stigmasterol* (**2**) to progesterone can be carried out in several ways. The Ciba method uses 5,6-dibromostigmasterol (**11**) as the starting compound [148] (see lower figure in left column).

Oxidation of the dibromide with chromic acid followed by zinc reduction gives bisnorcholic acid (**12**), which is then converted to the methyl ketone (**13**). The 5,6-double bond of compound **13** is protected by bromination before the compound is oxidized with peracid to give the 20-acetate derivative (**14**). In the final step, saponification, Oppenauer oxidation, and oxidation with chromic acid give progesterone with a total yield of 40 % (based on **12**).

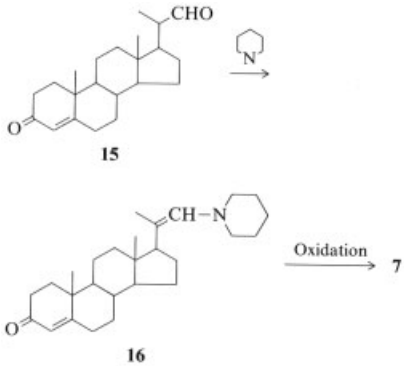

In a method developed at Upjohn [127], the hydroxyl group of stigmasterol (**2**) is first oxidized (Oppenauer) to the ketone. If this ketone is ozonized under specific conditions, only the double bond in the side chain is attacked, yielding the aldehyde (**15**). Compound (**15**) is con-

verted first to the enamine (**16**) by azeotropic distillation with piperidine and, finally, to progesterone (**7**) by oxidation. The overall yield is 60 %.

Synthesis of 3β-Acetoxy-5-androsten-17-one. 3β-Acetoxy-5-androsten-17-one [*53-43-0*] (**8**) (formerly known as androstenolone) and its acetate (**8b**) serve as starting materials for the synthesis of testosterone, estrone, and other androgens and estrogens including 19-norandrostane steroids derived from nortestosterone. Two of the most important production methods are described below.

1) 3β-Acetoxy-5,16-pregnadien-20-one (**5**) (16-dehydropregnenolone acetate), derived from diosgenin (**1**), is heated with hydroxylamine hydrochloride and sodium acetate or pyridine in ethanol to give the oxime (**17**) [149]. Beckmann rearrangement of **17** yields the enamide (**18**), which is then subjected to mild acid hydrolysis to give androstenolone acetate (**8b**).

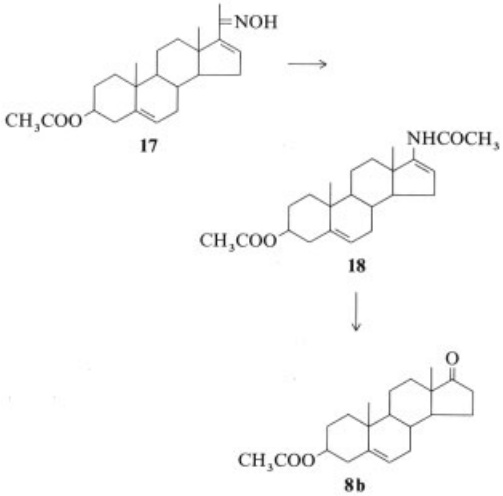

2) The second method uses pregnenolone acetate (**6b**), synthesized as described in page 28, as the starting material. Compound **6b** is first converted to the oxime and then subjected to a Beckmann rearrangement to give the acetamide (**19**), which in turn undergoes alkaline hydrolysis to yield the corresponding amine [150]. Reaction of the amine with hypochloric acid gives the chloramine (**20**), which is then treated with alkoxide (ketimine), followed by acetylation and acid hydrolysis to produce androstenolone acetate [151].

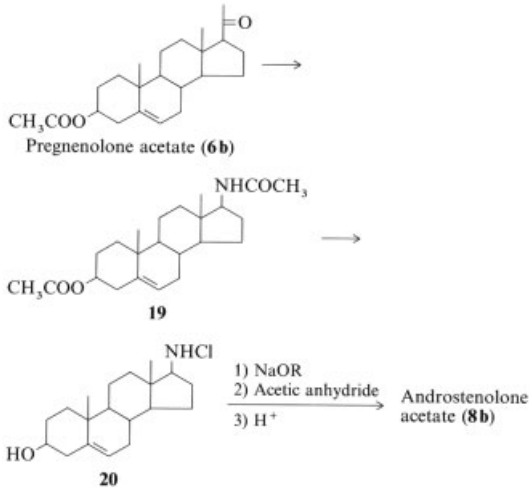

Synthesis of Estrone. Estrone [*53-16-7*] (**10**) is an active ingredient of commercial hormone preparations; it is also an important starting substance in the production of estradiol, ethinyl estradiol, other estrogenic steroid hormones, and 19-norandrostane compounds.

Estrone is synthesized from androstenolone (**8**), which is first hydrogenated with palladium charcoal to give epiandrosterone (**21**). Subsequent oxidation with chromic acid yields the 3,17-diketone, which is then treated with bromine in acetic acid to give the 2,4-dibromide. The dibromide, in turn, is converted to androstadienedione (**9**) by elimination of two molecules of hydrogen bromide. Dissolution of **9** in tetralin or mineral oil and pyrolysis (i.e., brief exposure to high temperature, ca. 600 °C) in the presence of quartz beads yields estrone (**10**, see next page) [152].

The "aromatization" of androstadienedione to give estrone has long been the subject of intensive research [127, 145]. Chemical synthesis of compound **9** from compound **8** requires four steps, whereas microbiological synthesis proceeds directly [153].

Synthesis of Testosterone. Testosterone [*58-22-0*] (**23**) is used therapeutically as the free steroid or as an ester. Testosterone is synthesized from androstenolone acetate (**8b**)

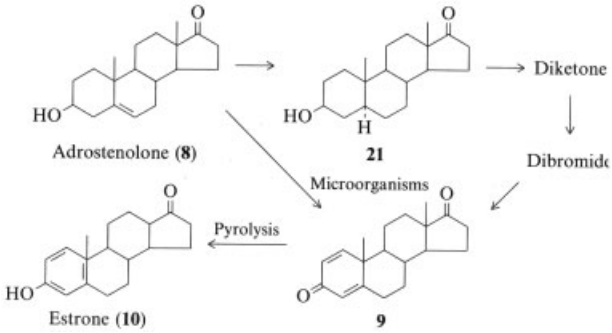

which is reduced to the 17β-alcohol with Raney nickel and then esterified with benzoyl chloride in pyridine. This protecting ester group permits partial saponification of the 3-acetate with methanolic sodium hydroxide solution to yield the 3-hydroxy compound (**22**). Subsequent Oppenauer oxidation produces the testosterone benzoate (**23**), which is then subjected to alkaline hydrolysis to give testosterone (**23**) [145]. Testosterone 17-esters can be obtained by normal esterification procedures.

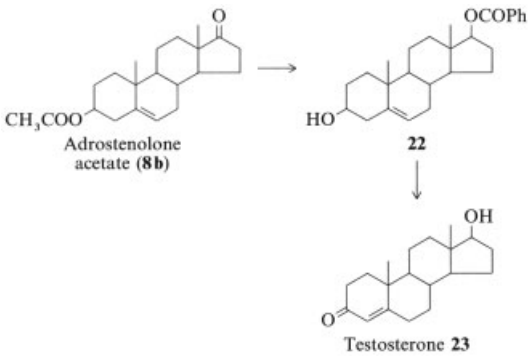

Synthesis of 19-Nortestosterone. 19-Nortestosterone [*434-22-0*] (**26**), which differs from testosterone (**23**) in that it does not have a methyl group at position 19, is the parent compound of the anabolic 19-norandrostane steroids and of the highly gestagenic hormones belonging to the norethisterone series.

Birch reduction of the methyl ether of estrone (**24**) is one of the most important methods used for the partial synthesis of 19-norandrostane compounds [154, 155].

The methyl ether is obtained by methylation of estrone with dimethyl sulfate and methanolic potassium hydroxide solution and then reduced to the alcohol with a complex hydride (e.g., $LiAlH_4$ or $NaBH_4$). The aromatic ring is subsequently reduced with lithium in liquid ammonia and diethyl ether in the presence of a small amount of ethanol to give the enolether (**25**). Treatment with methanolic hydrochloric acid yields 19-nortestosterone (**26**).

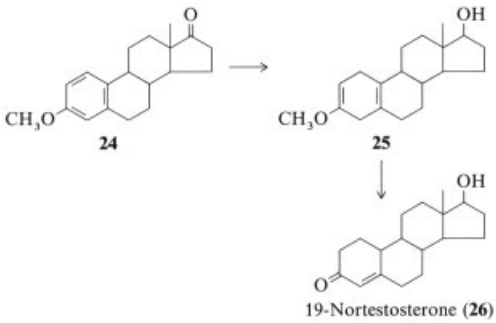

Synthesis of Norethisterone [127, 145, 156]. 17α-Ethynyl compounds may be obtained by ethynylation of 17-keto steroids. Most of these ethynyl compounds not only have interesting biological properties but also have a high oral activity. Norethisterone [*68-22-4*] (**29**) is a highly potent, orally active gestagen; many of its derivatives have characteristic activities.

An exceptionally short synthesis starts with the enolether (**25**) prepared by Birch reduction of estrone methyl ether (**24**). Oppenauer oxidation [157] of **25** yields the 17-keto compound, which can be ethynylated to compound **28** by one of the usual methods [156]. Subsequent acid hydrolysis gives norethisterone.

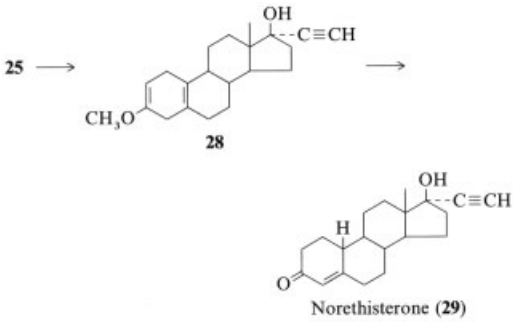

Another method used in the production of 19-norsteroids entails the functionalization and subsequent elimination of the 19-methyl group of 19-methyl steroids to produce the corresponding 19-nor compounds [158]. Both the aromatization of ring A and the subsequent Birch reduction are circumvented. Thus, androstenolone acetate (**8b**) can be converted to norethisterone (**29**), and pregnenolone acetate (**6b**) to gestonorone caproate (**66**, see below) [178].

The synthesis of norethisterone (**29**) and norethynodrel (**71**) proceeds as follows. Addition of hypochloric acid to androstenolone acetate (**8b**) gives the corresponding chlorohydrin. On reaction with lead tetraacetate and iodine, the chlorohydrin is converted to a cyclic ether, which in turn is saponified and oxidized to give a diketone. Reductive opening of the cyclic ether ring with zinc in glacial acetic acid produces a 19-hydroxy derivative, and subsequent oxidation with chromic acid gives the 10β-carboxylic acid, which is decarboxylated to the diketone by heating with pyridine. Partial ketalization yields the acetal, and ethynylation with sodium acetylide followed by treatment with either a weak acid (malonic acid) or a mineral acid finally yields norethynodrel (**71**) or norethisterone (**29**), respectively.

Synthesis of 17-Acetoxyprogesterone. Introduction of an acetoxy group at position 17 of progesterone (**7**) considerably increases the oral activity of this gestagen. 17-Acetoxyprogesterone [*302-23-8*] (**32**) is the active ingredient of a large number of commercial gestagen preparations and the parent compound of many gestagens with excellent oral activity [127, 145].

One of the syntheses available for the preparation of 17-acetoxyprogesterone starts with dehydropregnenolone acetate (**5b**), obtained from diosgenin. Compound **5b** is first converted to the 16α,17-epoxide and then to the bromohydrin (**30**). Subsequent debromination gives

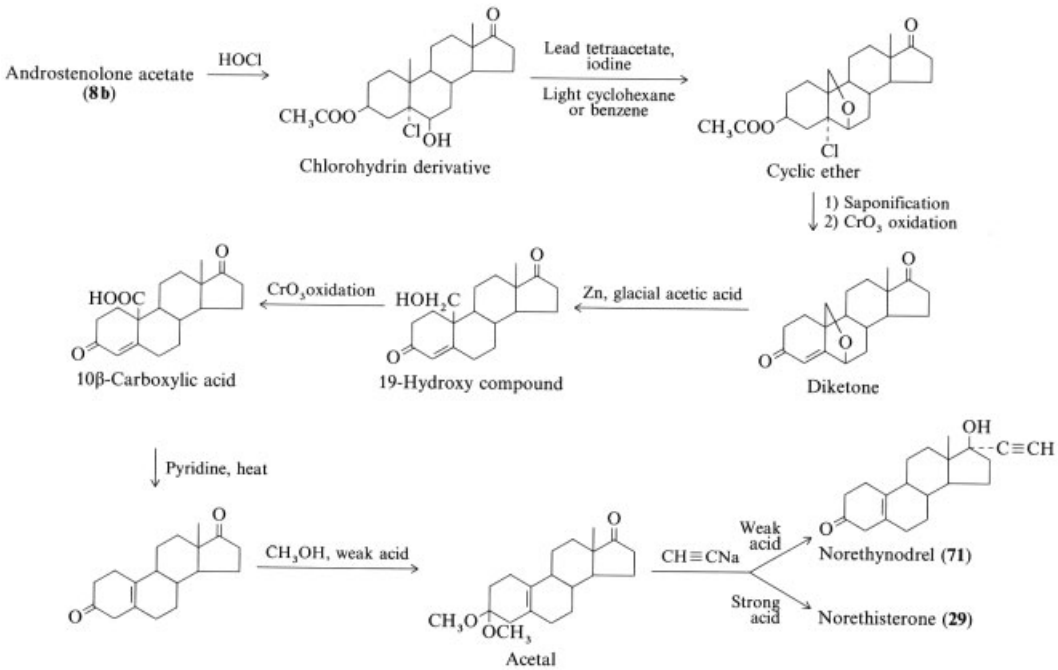

the 3-formate (**31**), which is subjected to acid-catalyzed acetylation and Oppenauer oxidation to give compound **32**.

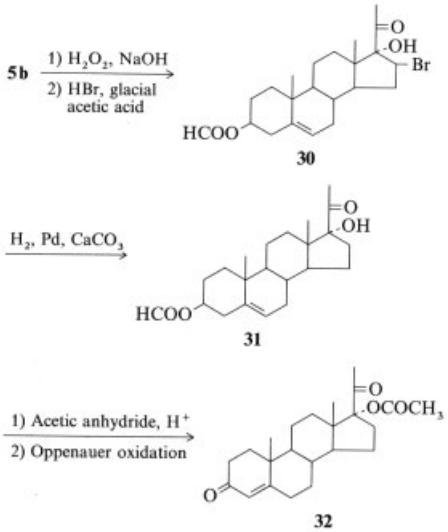

3.3. Microbiological Methods for the Synthesis of Steroid Intermediates

Since the beginning of the 1950s, microbiological conversion has become increasingly important in the commercial production of steroid hormones. Advantages of this technique include stereospecific reactions at positions that are not accessible to chemical reaction, mild reaction conditions, and reaction coupling (e.g., dehydrogenation and isomerization). These advantages more than compensate for the disadvantages (low substrate concentrations and, therefore, large reaction volumes; higher personnel and energy costs; more sophisticated equipment; and sterility of all media and materials involved in the reaction).

Microbiological reduction has become the key reaction in the total synthesis of sex hormones. This is the only economic way to produce pure enantiomers of the natural, optically active form of these hormones and to circumvent

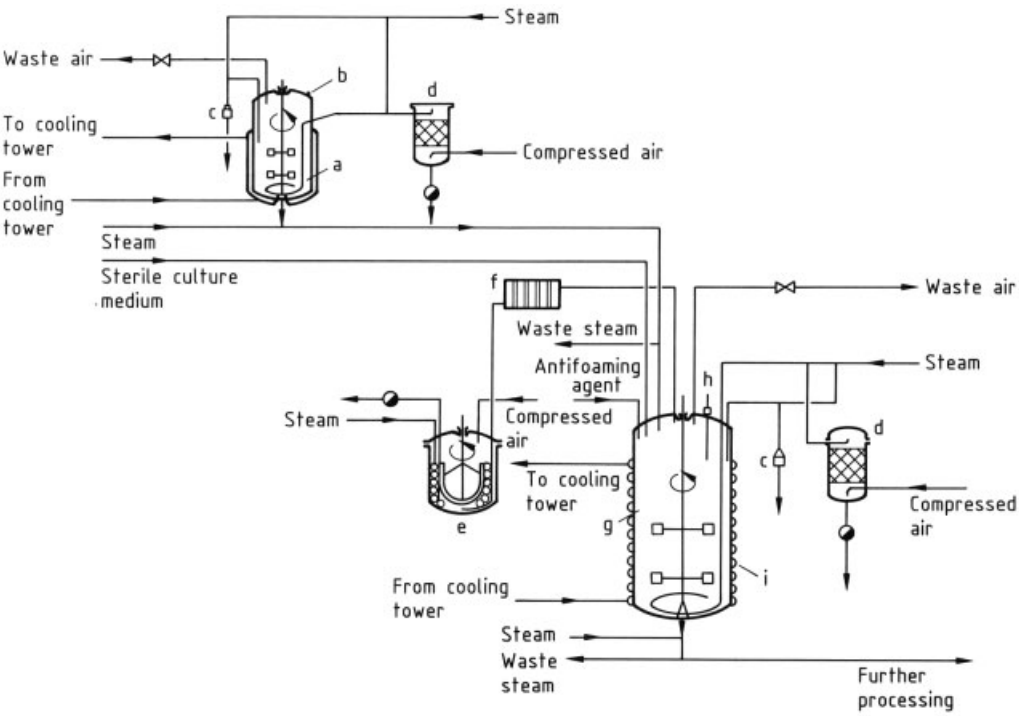

Figure 7. Schematic of a fermentation plant [159]
a) Prefermenter; b) Inoculation port; c) Sampling device; d) Filter; e) Substrate tank; f) Sterile filter; g) Main fermenter; h) Electrode for foam control; i) Cooling coil

the tedious and inefficient resolution of racemates (see Section 3.4).

Microbiological degradation can be used to convert naturally occurring sterols, such as cholesterol, β-sitosterol, stigmasterol, and ergosterol, to androstadienedione (**9**). Chemical degradation of these abundantly occurring compounds is inefficient and unprofitable. The technology used in the microbiological conversion of steroids is similar to that used in the production of antibiotics (→Antibiotics, Chap. 5.4.). The required quantity of the desired microorganisms is obtained by growing the cells in one or two prefermentation steps. Then, reaction with the relevant substrate occurs in the main fermenter. All steps must be carried out under sterile conditions to avoid contamination. The flow diagram of a fermentation plant is shown in Figure 7. The main fermenter can have a volume up to 50 000 L [159].

Degradation to Androstadienedione. The complete fermentative degradation of sterols to carbon dioxide and water by microbial enzymes has been known for a long time. Methods have been developed to prevent complete breakdown of the sterol nucleus after degradation of the side chain, e.g., at the stage of androstadienedione (**9**) [141, 142, 144]. In all these methods, inexpensive sterols (cholesterol, sitosterol, stigmasterol, and ergosterol) serve as starting materials, a variety of microorganisms are used, and the aerobic fermentation must be carried out in the presence of inhibitors of the enzymes that continue degradation to carbon dioxide and water (e.g., inorganic heavy-metal salts, such as nickel sulfate, and organic chelating agents). An overview of the microbiological degradation methods available and their yields of androstadienedione is presented in Table 4.

3.4. Total Synthesis of Sex Hormones

The consumption of sex hormones as oral contraceptive agents as well as the demand for steroids to produce adrenal cortical hormones has increased dramatically since 1960 so that demands could not be met by available raw material sources [147]. For this reason, the total synthesis of steroid substances is very important. In addition, total synthesis of steroids permits the production of structural elements that are impossible or extremely difficult to obtain with partial syntheses (e.g., production of homologous 13-alkyl steroids).

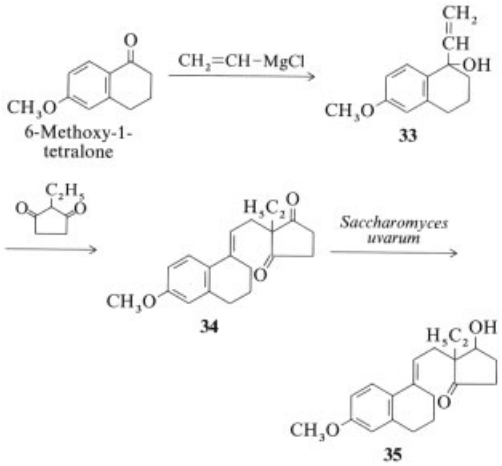

Synthesis of d-Norgestrel. D-Norgestrel [*6533-00-2*] (**38**), the gestagenic component of the contraceptives Neogynon and Stedirild, is produced by total synthesis [159–161]. 6-Methoxy-1-tetralone is converted to the vinyl alcohol (**33**) with vinylmagnesium chloride. Condensation of **33** with 2-ethylcyclopentane-1,3-dione gives the secodione (**34**). Stereospecific

Table 4. Microbiological degradation used for the synthesis of androstadienedione

Sterol used as starting material	Microorganism	Inhibitor	Yield of androstadienedione, %	Reference
Cholesterol (**4**)	*Mycobacterium phlei*	8-hydroxyquinoline	38	[144]
		nickel(II) sulfate	50	[141]
	Nocardia	cobalt(II) chloride	13	[142]
	Arthrobacter	nickel(II) sulfate	25	
β-Sitosterol (**3**)	*Mycobacterium phlei*	8-hydroxyquinoline	18	[144]
	Arthrobacter	nickel(II) sulfate	13	[142]
Stigmasterol (**2**)	*Mycobacterium phlei*		27	[144]

microbiological reduction of **34** yields the β-hydroxy ketone (**35**), which has two asymmetric carbon atoms (C-13 and C-17). Compound **35** is optically active and belongs to the normal series; i.e., C-13 and C-17 have the same configuration as in natural steroid compounds. Because subsequent reactions are stereospecific, the final product (D-norgestrel) retains the optical activity and normal configuration introduced by microbiological reduction [161] (see figure on previous page).

Acidic cyclization of **35** produces **36**, which has a double bond between C-14 and C-15. In the subsequent Birch reduction, the $\Delta^{8,9}$ double bond and the aromatic ring are reduced. Oppenauer oxidation of the 17β-hydroxyl group of the resulting dienolether leads to the corresponding ketone. Subsequent ethynylation of the ketone and cleavage of the enolether yield D-norgestrel (**38**).

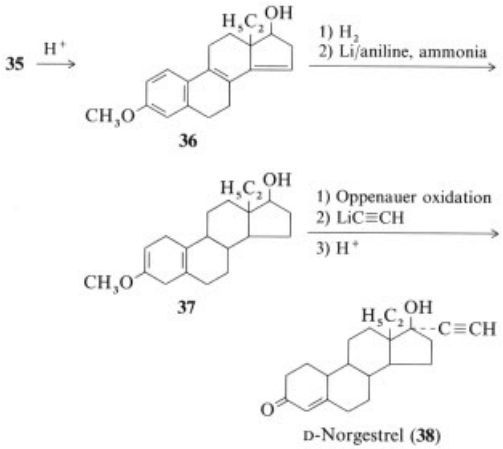

Synthesis of Norgestrienone. Norgestrienone [*848-21-5*] (**46**) is the gestagenic component of the oral contraceptive Planor, registered only in France [162]. Michael addition of methyl 5-oxo-6-heptenoate to 2-methylcyclopentane-1,3-dione produces the methyl ester of the triketocarboxylic acid (**39**). As in the synthesis of D-norgestrel, **39** is then subjected to partial, stereospecific reduction with the help of microorganisms. The resulting optically active β-hydroxy ketone (**40**) belongs to the normal series [163]. Again, subsequent reactions proceed stereospecifically, and the final product retains the correct steric configuration attained in one of the early steps.

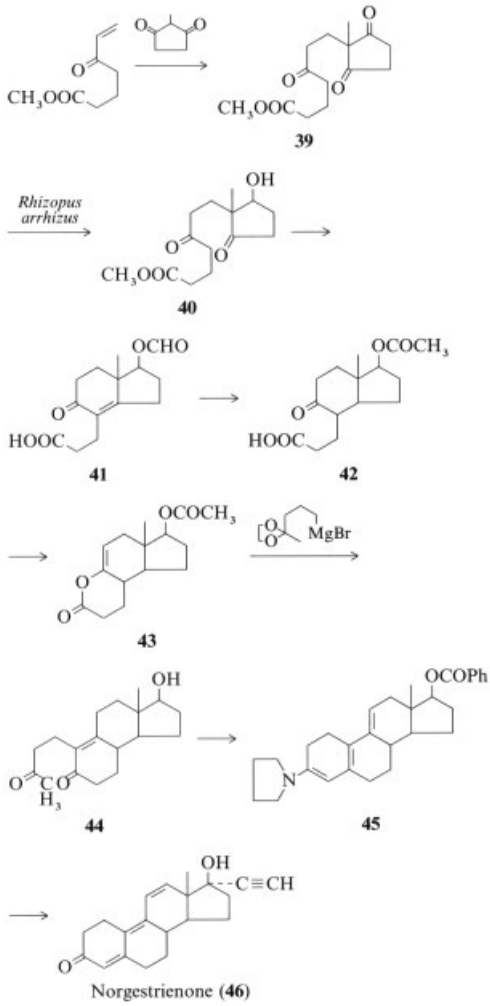

Compound **40** is converted to the formate (**41**) by aldol condensation, acid-catalyzed hydrolysis, and formylation. Compound **41** is converted to the saturated acetate (**42**) by stereoselective hydrogenation, saponification, and acetylation. Treatment with acetic anhydride gives the enol lactone (**43**). Subsequent Grignard reaction yields **44**, which already possesses rings B, C, and D of the steroid skeleton. Compound **44** is benzoylated at position 17 and then converted to compound **45** by aldol condensation and enamine formation with pyrrolidine. Subsequent dehydrogenation with 2,3-dichloro-5,6-dicyano-1,4-benzoquinone, protection of the resulting 3-keto group by oxime formation, oxidation to the 17-ketone, ethynylation, and oxime cleavage finally yield norgestrienone (**46**) [163].

Both syntheses use inexpensive chemicals as starting materials and apply microbiological reduction (production of an asymmetric carbon atom) to skillfully circumvent the formation of racemates and their inefficient resolution. Older versions of both methods produced racemates. Consequently, the contraceptives Eugynon (Schering) and Stediril (Wyeth) contain racemic norgestrel as gestagen.

3.5. Estrogens

3.5.1. General Aspects

Physiology and Clinical Applications. The first sex hormone to be isolated was estrone (**10**). It was recovered from the urine of pregnant humans by Doisy and Butenandt in 1929; later, the urine of pregnant mares was used as a richer source of estrogens. The structure was elucidated in 1932 by Butenandt, Marrian, and Haslewood.

The naturally occurring estrogens estradiol (**47**), estrone (**10**), and estriol (**48**), together with the gestagen, progesterone (**7**), regulate the menstrual cycle in humans and several animal species [164]. In addition, estrogenic hormones influence the growth and blood circulation of the uterus, vagina, and mammary glands. They are used in substitution therapy to control estrogen deficiency caused by ovarian hypofunction, castration, or menopause. They are also used in suppression therapy to suppress gonadotropin-induced ovulation and the endogenous production of estrogens [164]. Because of the latter, estrogens are used together with gestagens as combination contraceptive agents (→Contraception, Chap. 4.). A defined dose ratio of estrogen to gestagen together with a set mode of administration prevents ovulation and, consequently, fertilization with minimal side effects. The hormones are also administered in osteoporosis senilis, prostate carcinoma, mastocarcinoma, and for suppression of lactation [164].

Lists of the most important commercially available estrogenic sex hormones are given in Tables 5 (natural estrogens) and 6 (estradiol derivatives). They are available as single substances or as combination preparations.

Bioassays. Four types of bioassay are used for the quantitative determination of estrogenic hormones:

1) *Vaginal Smear Test.* The vaginal smear test, developed by Allen and Doisy [130, 131], is specific for active estrogens and permits the isolation of estrogenic sex hormones. Subcutaneous or oral administration of these hormones leads to a marked proliferation of the vaginal epithelium and keratinization of the surface cell layers (Schollen stage) in small rodents. These characteristic changes in cell structure are visible on microscopic examination of vaginal smears. When equal hormone doses are used, the degree of change depends on the duration of the effect and the potency of the estrogenic substances. The reference substances used are estradiol (**47**) for subcutaneous application, ethinyl estradiol (**51**) for oral application, and estradiol heptanoate or undecenoate for long-acting substances with depot activity. Castrated animals are used to avoid the effects of the natural sexual cycle.
2) *Sialic Acid Content of Mouse Vagina.* Administration of estrogen to castrated mice leads to a dose-dependent reduction of the sialic acid content of the vagina [165]. After subcutaneous or oral treatment, the animals are sacrificed, the vaginas are removed and processed, and the sialic acid content is determined colorimetrically. This test is specific for estrogenic substances; the sialic acid content is a measure of the potency of the estrogen. Reference substances are estradiol for subcutaneous application and ethinyl estradiol for oral administration.
3) *Vaginal Opening Test.* The vagina of rodents remains closed until sexual maturity. Active estrogens cause premature opening of the vagina. This test, however, is not estrogen-specific because androgens and anabolic substances produce a similar effect.
4) *Uterine Growth Test.* Administration of estrogens to rodents causes increased growth of the endometrium and myometrium of the uterus. The increase in uterine weight is proportional to the dose and potency of the estrogen administered.

Table 5. Natural estrogens

Formula	Names	CAS registry no.	Trade name (manufacturer)
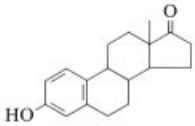 **10**	estrone, 3-hydroxy-1,3,5(10)-estratrien-17-one [a], [b]	[53-16-7]	Estrone (Abbott, Lilly, Wyeth); Glandubolin (Gedeon Richter); Kolpon (Organon); Ovex (Leo Danmark); Theelin (Parke Davis)
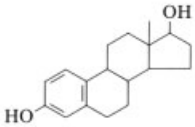 **47**	estradiol, 1,3,5(10)-estratriene-3,17β-diol [a], [b]	[50-28-2]	Dimenformon (Organon); Diogyn (Pfizer); Gynoestryl (Roussel); Ovocyclin (Ciba); Profoliol (Schering Corp.); Progynon-Salbe (Schering AG)
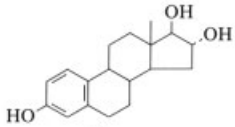 **48**	estriol, 1,3,5(10)-estratriene-3,16α,17β-triol [a], [b]	[50-27-1]	Gynäsan (Bastian); Hormomed (Merckle); Ovestin (Organon); Theelol (Parke Davis); Triovex (Leo)
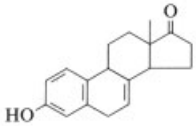 **49**	equilin	[474-86-2]	
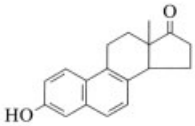 **50**	equilenin	[517-09-9]	
	mixture of estrone, equilin, equilenin, and the sodium 3-sulfate salts of the corresponding 17α-dihydro compounds [c]		Amniotin (Squibb); Conestron (Wyeth); Konogen (Lilly); Premarin (Ayerst); Presomen (Kali-Chemie)

[a] [130], [b] [131], [c] [127].

3.5.2. Natural Estrogens and Estradiol Derivatives

Natural Estrogens. The naturally occurring estrogens, estrone (**10**), estradiol (**47**), and estriol (**48**), as well as the estrogens from mare urine, equilin (**49**), equilenin (**50**), and their 17-dihydro compounds, are of considerable importance in the treatment of menopausal symptoms (for structures, see Table 5) [161]. However, very high oral doses are required to elicit responses because these compounds are rapidly metabolized and inactivated in the liver. Injection preparations for parenteral administration can only be produced in the form of crystal suspensions because of the low solubility.

Estradiol Esters. Esterification of naturally occurring estrogens with longer chain carboxylic acids is an important method for increasing their duration of action. In contrast with free estrogens, oily solutions of esterified estrogens can be produced for depot injection preparations.

17-Alkylated Estradiol Derivatives. Alkylation of estradiol derivatives at position 17 increases their oral activity many times. The 17-alkyl group imparts stability to biodegradation which, in natural estrogens, begins with oxidation of the secondary 17-hydroxyl group to a ketone [131, 133]. As a result of the high oral activity of these estrogenic compounds,

Table 6. Estradiol derivatives

Formula	Names	CAS registry no.	Trade name (manufacturer)
Estradiol esters			
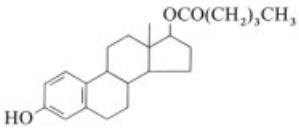	estradiol valeriate, 17β-valeryloxy-1,3,5(10)-estratrien-3-ol	[*979-32-8*]	Cyclo-Progynova (Schering AG); Delestrogen (Squibb); Östrogynal sine (Asche); Progynova, Primogyn-Depot (Schering AG)
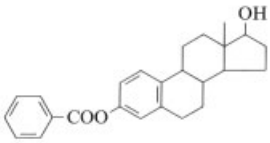	estradiol benzoate, 3-benzoyloxy-1,3,5(10)-estratrien-17β-ol	[*50-50-0*]	Benzo-Gynoestryl (Roussel); Diogyn (Pfizer); Follicyclin (Ciba); Östradiol "Vitis" (Vitis); Ostroform (British Drug Houses); Ovocyclin M (Ciba)
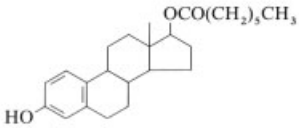	estradiol enanthate, 17β-heptanoyloxy-1,3,5(10)-estratrien-3-ol	[*4956-37-0*]	Used only in the depot contraceptive Deladroxat (Squibb)
17-Alkylated estradiol derivatives			
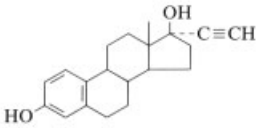 **51**	ethinyl estradiol, 17α-ethynyl-1,3,5(10)-estratriene-3,17β-diol	[*57-63-6*]	Diogyn-E (Pfizer); Eticyclin (Ciba); Feminone (Upjohn); Gynolett-Tabletten (Labopharma); Kolpolyn (Organon); Progynon C, Progynon M (Schering AG); Ylestrol (Ferndale), used in oral contraceptives
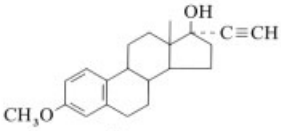 **52**	mestranol, 17α-ethynyl-3-methoxy-1,3,5(10)-estratrien-17β-ol	[*72-33-3*]	Ovastol (Rendell), used in oral contraceptives
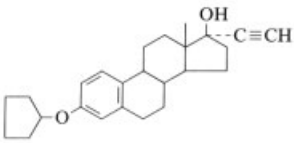	quinestrol, 17α-ethynyl-3-cyclopentyloxy-1,3,5(10)-estratrien-17β-ol *	[*152-43-2*]	Estrovis, Estrovis 4000 (Gödecke), used in depot contraceptives
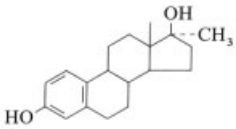	methylestradiol, 17α-methyl-1,3,5(10)-estratriene-3,17β-diol	[*302-76-1*]	Klimanosid "R" (Boehringer Mannheim)

* [166].

[e.g., ethinyl estradiol (**51**) and its methyl ether, mestranol (**52**)], they are used widely in combination preparations for oral contraception (→Contraception, Chap. 4.1.).

3.5.3. Nonsteroidal Estrogens

Estrogenic activity is not restricted to the estrogenic sex hormones. Certain plant constituents,

bituminous components of the Seefeld oil shale, easily synthesized stilbene derivatives, and their dihydro compounds have estrogenic properties [130]. The therapeutic application of stilbestrol derivatives in clinical medicine is limited. They are of greater importance in veterinary medicine (e.g., in the artificial caponizing of chickens).

The parent compounds of the nonsteroidal estrogens are diethylstilbestrol [*56-53-1*] (**53**), hexestrol [*84-16-2*] (**54**), and dienestrol [*84-17-3*] (**55**).

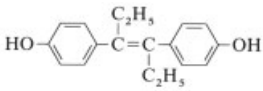

Diethylstilbestrol (**53**)

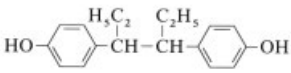

Hexestrol (**54**)

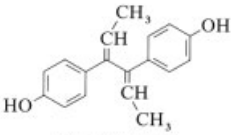

Dienestrol (**55**)

Compounds **53**–**55** are used in the free form, as their diethers or diesters, or with substituted aromatic rings or double bonds. The most important commercial preparations are listed in Table 7.

3.5.4. Antiestrogens [130, 131]

The term antiestrogen designates a very heterogeneous group of substances that counteract the effects of estrogen. The use of antiestrogens is indicated primarily in the palliative treatment of metastasizing mastocarcinoma. In addition, they can be used to treat benign prostate hyperplasia and endometriosis.

Antiestrogens can be divided into three groups on the basis of their mode of action:

1) *Competitive Antagonists.* These active antiestrogenic substances act by binding competitively to the estradiol receptor. The nonsteroidal stilbene derivatives tamoxifen [*10540-29-1*], cyclofenil [*2624-43-3*], and clomiphene [*911-45-5*] have practical importance. Chemical modification of estradiol has also produced derivatives with antiestrogenic properties. However, these compounds usually exert a mixture of dosage-dependent, antagonistic and agonistic effects (see upper figure in next column).
2) *Aromatase Inhibitors.* Transformation of androstenedione to estrone is catalyzed by the enzyme aromatase (see Fig. 4). A series of steroidal and nonsteroidal substances reversibly or irreversibly inhibit this enzyme system and are thus suitable agents for lowering the biological level of estrogen. In addition to androstane derivatives, imidazole

Table 7. Stilbestrol derivatives

R*	Generic name	CAS registry no.	Trade name (manufacturer)
H	diethylstilbestrol, *trans*-3,4-bis(4-hydroxyphenyl)-3-hexene	[*56-53-1*]	Cyren A (Bayer); Klimax Taeschner (Taeschner); Oekolb (Dr. Kade); Östrogen-Holzinger (Holzinger)
H_2PO_3	diethylstilbestrol phosphate	[*522-40-7*]	Cytonal (VEB Berlin-Chemie); Honvan (Asta)
CH_3CH_2CO	diethylstilbestrol dipropionate	[*130-80-3*]	Cyren B (Bayer); Syntoestron (Eros, Kuessnacht)
CH_3	dimethoxydiethylstilbene, *trans*-3,4-bis(4-methoxyphenyl)-3-hexene	[*130-79-0*]	Depot Östromon (Merck AG); Dimethylöstrogen-Holzinger (Holzinger)

* All derivatives have the following structure:

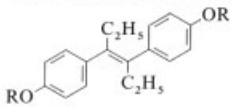

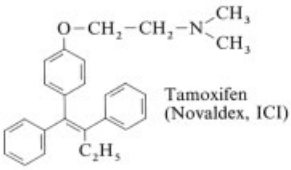

Tamoxifen
(Novaldex, ICI)

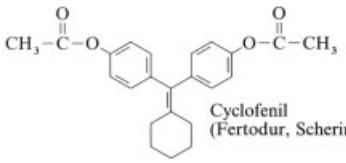

Cyclofenil
(Fertodur, Schering)

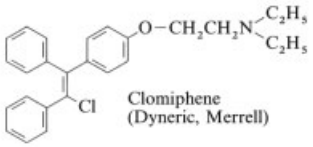

Clomiphene
(Dyneric, Merrell)

derivatives also have remarkable inhibitory properties. Examples of aromatase inhibitors follow:

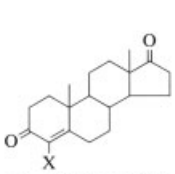

X = OH [*566-48-3*]
NH_2 [*105051-81-8*]
F [*98102-30-8*]
CH_3COO [*61630-32-8*]

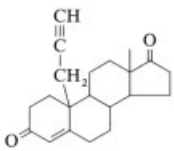

PED [*77016-85-4*]
(Merrel-Dow)

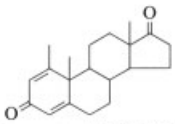

Atamestan [*96301-34-7*]
(Schering)

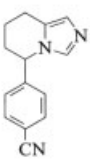

CGS 16949 A
[*102676-47-1*]
(Ciba-Geigy)

3) *Antigonadotropic Substances.* Another way of reducing ovarian estrogen production is by inhibiting hypophyseal gonadotropin secretion. Although Danazol [*17230-88-5*] has a weak androgenic, anabolic component, it primarily exhibits a central gonadotropin-inhibiting effect. Danazol is especially useful for the treatment of endometriosis.

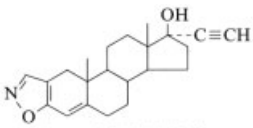

Danazol [*17230-88-5*]
(Sterling Winthrop)

3.6. Androgens and Anabolic Agents [140]

3.6.1. General Aspects

Physiology and Clinical Applications. Testosterone (**23**), the principal male sex hormone (androgen), was isolated in 1935 from bull testes (100 kg of bull testis gave approximately 10 mg of testosterone), and its structure was subsequently elucidated. A weaker androgenic substance androsterone (**56**) was isolated in 1931 from human urine and shown to be a breakdown product of testosterone.

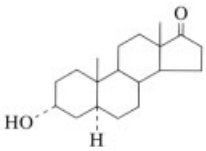

Androsterone (**56**)

Testosterone synthesis occurs in the testis and is regulated by the hypophyseal luteinizing hormone (LH). However, small amounts of androgens are also produced in the ovaries, the adrenal glands, and the placenta. The chief functions of testosterone are to ensure normal development of the seminal vesicles, prostate, epididymis, and Cowper glands and to maintain secondary male characteristics (facial hair, deep voice). Spermatogenesis, on the other hand, is regulated by the follicle-stimulating hormone (FSH).

Clinical use of androgens is indicated in cases of androgen deficiency, such as climacterium virile, hypogenitalism, prostate hypertrophy, and other potency disorders (substitution therapy). The administration of testosterone also has a suppressive effect on the hypophysis; it inhibits the release of LH and FSH and, consequently, spermatogenesis. A "rebound effect" is observed after treatment stops; i.e., the production of LH and spermatogenesis are intensified. This effect may be desirable in cases of oligozoospermia. Androgens are prescribed for carcinoma of the breast, carcinoma of the uterus, or endometriosis, as well as to inhibit lactation.

Testosterone and androsterone, like all androgens, exert an anabolic influence and stimulate protein synthesis in muscle and bone. They

are used therapeutically in the treatment of consumptive disease, anorexia, old-age infirmity, osteoporosis, etc. Chemical modification of androgens has produced substances with a high anabolic-to-androgenic potency, which can also be administered to women. However, all anabolic agents do have a certain residual androgenic effect [167].

Bioassays. Two types of bioassay are used for the quantitative determination of androgens and anabolic agents:

Assay of Androgenic and Anabolic Activity [168]. The function and size of the accessory sex glands (seminal vesicles and prostate) and of the musculus levator ani (a muscle of the perineal complex) in castrated male rats depend on the presence of androgens. The fresh weights of the seminal vesicles (androgenic activity) and levator ani (anabolic effect) are measured. The anabolic and androgenic effects are then determined from a dose – response curve. The organ weights are compared with values given by reference substances (testosterone propionate for subcutaneous application and 17α-methyltesterone for oral application). The relative potencies are expressed as fractions or multiples of the activity of the reference substance.

Chicken Comb Test. The comb of a young cockerel grows on administration of androgens. After standard treatment, the comb and body weights are measured, and the ratio thus obtained is compared with that produced by reference substances. Again, testosterone propionate is used as reference for subcutaneous application and 17α-methyltestosterone for oral administration.

3.6.2. Androgens

Testosterone and Testosterone Esters. A large number of preparations contain testosterone (**23**) in the free form or as one of its esters; examples are listed in Table 8. The rapid breakdown of testosterone (oxidation at position 17, reduction at position 3, and reduction of the double bond) and its subsequent inactivation can be slowed down by esterification with longer chain fatty acids; consequently, the duration of its action can be effectively prolonged (depot effect).

Other Androgens. Commercial preparations available for androgen therapy are listed in Table 9. They are androgens with good oral activity, sufficient duration of action, and a low inhibitory effect on spermatogenesis [140]. Intro-

Table 8. Testosterone and testosterone esters

R *	Names	CAS registry no.	Trade name (manufacturer)
H	testosterone; 17β-hydroxy-4-androsten-3-one	[*58-22-0*]	Androlin (Lincoln); Malestrone (Kirk); Perandren (Ciba); Synandrol F (Pfizer); Testoviron T (Schering AG); Testryl (Squibb)
CH_3CH_2CO	testosterone propionate, 17β-propionyloxy-4-androsten-3-one	[*57-85-2*]	Anertan (Boehringer Mannheim); Masenate (Schiefelin); Sterandryl (Roussel); Synandrol (Pfizer); Testosid (Boehringer Mannheim); Testoviron (Schering AG)
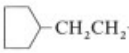	testosterone cyclopentylpropionate, 17β-cyclopentylpropionyloxy-4-androsten-3-one	[*58-20-8*]	Depo-Testadiol (Upjohn); Depovirin (Hoechst); Telipex-Retard (Sanabo)
$CH_3(CH_2)_5CO$	testosterone enanthate, 17β-heptanoyeoxy-4-androsten-3-one	[*315-37-7*]	Delatestryl (Squibb); Sterandryl Retard (Roussel); Testoviron Depot (Schering)

* Compounds have the following structure:

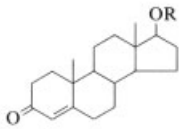

l**Table 9.** Commercially available androgens for therapeutic use

Formula	Name	CAS registry no.	Trade name (manufacturer)
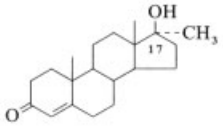	methyltestosterone, 17β-hydroxy-17α-methyl-4-androsten-3-one	[*58-18-4*]	Andrometh (Central Pharm.); Metandren (Ciba); Neo Hombreol M (Organon); Testosid (Boehringer Mannheim), Testosteron "Berco" (Berco)
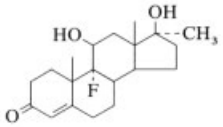	fluoxymestrone, 9-fluoro-11β,17β-dihydroxy-17α-methyl-4-androsten-3-one [a]	[*76-43-7*]	Afluteston (Arcana); Halotestin (Upjohn); Ora Testryl (Squibb); Ultranden (Ciba)
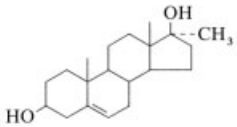	methandriol, 17α-methyl-5-androstene-3β,17β-diol [b]	[*521-10-8*]	Metandiol (Roussel); Methostan (Schering Corp.); Neosteron (Organon)
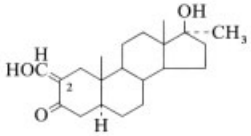	oxymetholone, 17β-hydroxy-17α-methyl-2-hydroxymethylen-5α-androstan-3-one [c]	[*434-07-1*]	Adroyd (Parke Davis); Anadrol (Syntex); Nastenon (Cassenne)
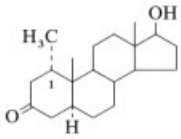	Mesterolone, 17β-hydroxy-1α-methyl-5α-androstan-3-one [d]	[*1424-00-6*]	Mestoran (Schering Denmark); Proviron (Schering AG)

l

l[a] [169], [b] [170], [c] [171], [d] [172].

duction of a methyl group at positions 17 and 1 of the steroid nucleus, a hydroxymethyl group at C-2, and the production of 9α-fluoro-11β-hydroxy compounds have proved particularly interesting. However, 17-methyl derivatives tax the liver to a greater extent and are contraindicated in cases of liver malfunction.

3.6.3. Antiandrogens [139, 140]

The term antiandrogen designates any substance that inhibits the action of the male sex hormone testosterone. Examples are cyproterone acetate [*427-51-0*] (for synthesis, see [181]) and flutamide [*13311-84-7*], which represent different classes of substances.

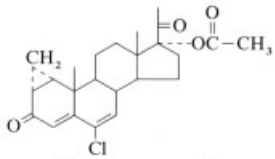

Cyproterone acetate

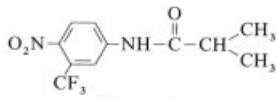

Flutamide

Both antiandrogens function by competitively displacing testosterone from the androgen receptor. Whereas flutamide has no other endocrinal side effects, cyproterone acetate possesses a strong gestagenic activity and, consequently, can also be used in combination preparations as an oral contraceptive. Antiandrogens

Table 10. Anabolic norandrostane derivatives

Formula	Name	CAS registry no.	Trade name (manufacturer)
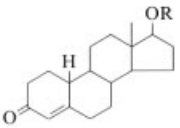 R = H (**26**)	nandrolone, nortestosterone, 17β-hydroxy-4-estren-3-one	[*434-22-0*]	Nortestonat (Upjohn)
R = 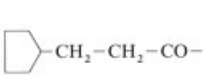	nortestosterone cyclopentylpropionate, 17β-cyclopentylpropionyloxy-4-estren-3-one	[*601-63-8*]	Depo-Nortestonat (Upjohn)
R = $CH_3-(CH_2)_8-CO-$	nortestosterone decanoate, 17β-decanoyloxy-4-estren-3-one	[*360-70-3*]	Deca-Durabolin (Organon)
R = 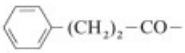	nortestosterone phenylpropionate, 17β-phenylpropionyloxy-4-estren-3-one	[*62-90-8*]	Durabolin (Organon); Norandrol (Osta); Norstenol (Ravizza)
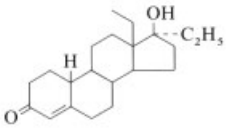	norbolethone, 17α-methyl-18-methyl-17β-hydroxy-4-estren-3-one [a], [b]	[*1235-15-0*]	Genabol (Wyeth)
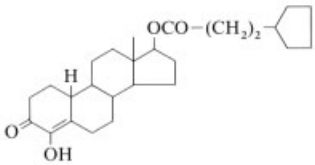	oxabolone cypionate, 17β-cyclopentylpropionyloxy-4-hydroxy-4-estren-3-one [c]	[*1254-35-9*]	Steranabol ritardo (Farmitalia)
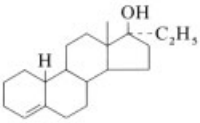	ethylestrenol, 17α-ethyl-4-estren-17β-ol	[*965-90-2*]	Duraboral, Durabolin O, Maxibolin, Orabolin (Organon)

[a] [160], [b] [173], [c] [174].

are applied primarily in the treatment of prostate tumors, hypersexuality in men, pubertas praecox in young men, androgen-dependent acne, seborrhea, and alopecia in women.

Biological Testing of Antiandrogenic Activity. Administration of an antiandrogen to pregnant rodents at the fetal sexual differentiation stage produces an androgen deficit; consequently, normal development of the exterior sex characteristics of the male fetus does not occur. A mixture of male and female characteristics is obtained, depending on the potency of the administered antiandrogen .

3.6.4. Anabolic Agents [167]

In many cases, chemical modification of the steroid nucleus of androgens of the 19-norandrostane, androstane, and androstanolone series increases their anabolic effects at the expense of their androgenic activity. Generally, the production of long-chain esters prolongs the action of subcutaneous preparations (depot effect) and methylation at C-17 increases oral potency.

19-Norandrostane Compounds. 19-Nortestosterone (**26**) is the parent compound of the series of therapeutically important anabolic agents listed in Table 10. Apart from esterification with long-chain acids, the introduction of

Table 11. Anabolic testosterone derivatives

Formula	Name	CAS registry no.	Trade name (manufacturer)
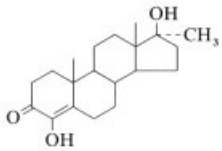	oxymesterone, 4,17β-dihydroxy-17α-methyl-4-androsten-3-one	[*145-12-0*]	Oranabol (Farmitalia) Theranabol (Theraplix)
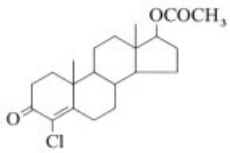	chlorotestosterone acetate, 17β-acetoxy-4-chloro-4-androsten-3-one	[*855-19-6*]	Steranabol (Farmitalia); Turinabol (VEB Jenapharm)
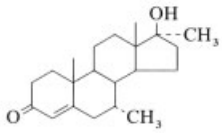	bolasterone, 7α,17α-dimethyl-17β-hydroxy-4-androsten-3-one	[*1605-89-6*]	Myagen (Upjohn)
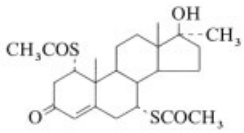	thiomesterone, 17β-hydroxy-17α-methyl-1α,7α-bis(thioacetyl)-4-androsten-3-one	[*2205-73-4*]	Emdabol (Merck)
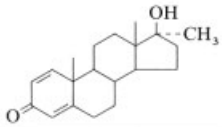	metandienone, 17β-hydroxy-17α-methyl-1,4-androstadien-3-one	[*72-63-9*]	Dianabol (Ciba)

an 18-methyl group or a 4-hydroxyl group, the reduction or removal of the 3-keto group, and alkylation at C-17 give rise to potent anabolic agents.

Testosterone Compounds. Some anabolic testosterone derivatives are listed in Table 11. Introduction of a hydroxyl group or a chlorine atom at position 4, methylation at C-7, introduction of a thioacetyl group or an additional double bond between C-1 and C-2 favorably influence activity, and 17-methylation produces high oral potency.

Androstanolone Compounds. Androstanolone (**57**) is the parent compound of another series of anabolic sex hormones [175]; see Table 12. The remaining androgenic activity of androstanolone is further reduced by methylation at C-1 or C-2, or by replacing C-2 with oxygen or the A ring with a heterocyclic ring. Again, methylation at C-17 increases oral activity, but also increases stress on the liver.

3.7. Gestagens

3.7.1. General Aspects

Physiology and Clinical Applications. Progesterone (**7**) is the most important representative of a group of female sex hormones known as gestagens. Ludwig Frenkel discovered progesterone in 1903, but not until 1934 was progesterone isolated and its structure analyzed.

Progesterone acts with another group of female sex hormones, the estrogens (see Section 3.5), to regulate the sexual cycle. After ovulation, progesterone is continuously formed by

Table 12. Anabolic androstanolone derivatives

Formula	Name	CAS registry no.	Trade name (manufacturer)
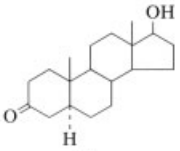 **57**	androstanolone, 17β-hydroxy-5α-androstan-3-one	[*521-18-6*]	Anaboleen (Uni-Chemie); Anabol-Tablinen (Sanorania); Stanaprol (Pfizer)
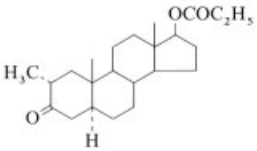	androstanolone propionate, 2α-methyl-17β-propionyloxy-5α-androstan-3-one	[*521-12-0*]	Drolban (Lilly); Masterid (Grünenthal)
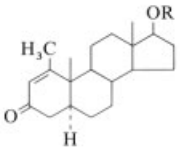	R = CH_3CO: methenolone acetate, 17β-acetoxy-1-methyl-5α-androst-1-en-3-one	[*434-05-9*]	Primobolan (Schering AG)
	R = $CH_3(CH_2)_5CO$: methenolone enanthate	[*303-42-4*]	Nibal (Squibb); Primobolan Depot (Schering AG)
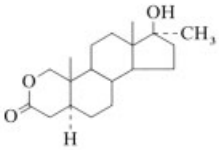	oxandrolone, 17β-hydroxy-17α-methyl-2-oxa-5α-androstan-3-one	[*53-39-4*]	Anavar (Searle); Protivar (Byla)
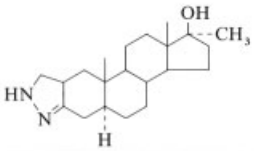	stanazolol, 17α-methyl-5α-androstano-[3,2-*c*]pyrazol-17β-ol	[*10418-03-8*]	Stromba, Winstrol (Winthrop)

the corpus luteum and prepares the endometrium of the uterus for implantation of the fertilized egg (transition from proliferation to secretion) [130, 131]. It prevents renewed ovulation and fertilization during pregnancy, decreases the tonus of the uterus, and thus maintains pregnancy. The inhibition of ovulation is the basis for the use of gestagens as contraceptive agents. The administration of gestagens "imitates" pregnancy–the secretion of FSH by the hypophysis diminishes, thus preventing maturation and release of the ovum [176]. The most important application of gestagens, alone or in combination with estrogens, is in the inhibition of ovulation or contraception (suppression therapy). In addition, gestagens are important in the treatment of deficiency symptoms such as imminent abortion, sterility, and irregularities of the menstrual cycle (substitution therapy). The strong inhibition of gonadotropin secretion exerted by some gestagens is also exploited in the treatment of prostate adenoma.

Hormonal Contraception [130, 131]. Hormonal contraception is based on the suppression of ovulation in the middle of the menstrual cycle. The secretion of LH, which is reponsible for ovulation, does not occur under the influence of a gestagen – estrogen combination preparation. An additional contraceptive effect is achieved by the fact that inhibitors of ovulation also influence the development of the endometrium and the consistency of the cervical secretion. The hormonal contraceptives presently available contain different types of gestagenic components

and varying doses of individual components. Although older combination preparations had a high proportional estrogen, the tendency is to reduce estrogen content because this is primarily responsible for side effects. The successful search for new gestagens has produced a generation of low-dosage ovulation inhibitors. New administration regimes (sequential preparations, two- and three-phase preparations) have resulted in further dose reduction; the doses of estrogen and gestagen are adapted to the course of the menstrual cycle (see also →Contraception, Chap. 4. [130, 131].

Bioassays. Two types of bioassay are used for the quantitative determinations of gestagens:

Clauberg Test [130]. The synergistic action of estrogens and gestagens induces characteristic changes of the endometrium. The endometrium proliferates under the action of estrogens and undergoes characteristic morphological changes under the subsequent influence of active gestagens. These changes are most apparent in rabbits; they are evaluated histologically and compared with those produced by reference substances. This test is used routinely in the search for novel gestagenic sex hormones and also permits determination of relative potencies.

Maintenance of Pregnancy. In pregnant rodents, removal of the corpus luteum or ovaries results in the termination of pregnancy. Active gestagens can maintain pregnancy in castrated, pregnant animals. The test is performed on pregnant rats and is used for qualitative characterization of the activity spectrum of a compound. However, it is not suitable for the determination of relative potencies.

3.7.2. Gestagens of the Pregnane and 19-Norpregnane Series

Progesterone (**7**), the naturally occurring gestagen, is administered subcutaneously as an injection solution; very high oral doses are required to elicit responses because of rapid degradation in the liver. For this reason, progesterone is not used as an oral contraceptive.

The extensive search for gestagens with high oral activity and low side effects started with three compounds that were more orally active than progesterone [127]: hydroxyprogesterone acetate (**58**), ethisterone (**68**), and 19-norethisterone (**29**), with approximately 30, 20, and 100 times the potency of progesterone, respectively. Modification of hydroxyprogesterone acetate (**58**), in particular substitution at C-6 and C-17, produced a large number of highly active oral gestagens. Some of these preparations and their trade names are listed in Table 13 (see next pages). Characteristic modifications include introduction of the Δ^6 double bond (**61**–**65**), substitution at C-6 with chlorine or a methyl group (**60**–**62**, **64**), changing the ring A – B coupling (**65**), and transition to the 19-nor series (**66**). The gestagens listed in Table 13 differ in their effectiveness; those that exert a strong antiovulatory effect (central inhibition) are of considerable importance as contraceptives (**60**–**62**, **67**). Other gestagens exert a stronger proliferative effect on the endometrium [176]. Gestonorone caproate (**66**) is also used in the treatment of prostate adenoma.

3.7.3. Gestagens of the Androstane Series

Insertion of a 17α-ethynyl group into testosterone yields ethisterone (**68**) which, surprisingly, is only slightly androgenic but is a highly potent oral gestagen [127]. The synthesis of ethisterone was the first successful step in the search for orally active gestagens. Dimethisterone (**69**) is even more potent than ethisterone; these are the only two products of this series on the market (Table 14, see next pages).

3.7.4. Gestagens of the 19-Norandrostane Series

Compounds of the 19-norandrostane (or estrone) series can be prepared easily by Birch reduction of estrone methyl ether. Examples commercially available as contraceptive agents are listed in Table 15 (see next pages) [176]. Transition from the androstane to the 19-norandrostane series further increases the biological potency of a large number of these sex hormones. 19-Norethisterone (**29**) was synthesized as an analogue of ethisterone (**68**); like ethisterone, it has only slight androgenic activity, but it is a fivefold stronger oral gestagen. A number of very

Table 13. Pregnane and 19-norpregnane gestagens

Formula	Name	CAS registry no.	Trade name (manufacturer)
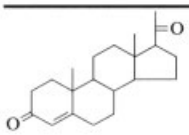 **7**	progesterone, 4-pregnene-3,20-dione [a, b]	[*57-83-0*]	Colprosteron (Ayerst); Geston (Paines and Byrne); Luteogan (Henning); Luteosid (Boehringer Mannheim); Progestin (Organon); Proluton (Schering AG)
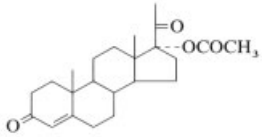 **58**	hydroxyprogesterone acetate, 17-acetoxy-4-pregnene-3,20-dione [a, b]	[*302-23-8*]	Prodox (Upjohn)
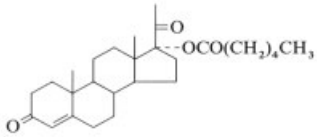 **59**	17α-hydroxyprogesterone caproate, 17-hexanoyloxy-4-pregnene-3,20-dione [a]	[*630-56-8*]	Delalutin (Squibb); Idrogesten (Farmila); Neolutin forte (Spofa); Proluton-Depot (Schering AG)
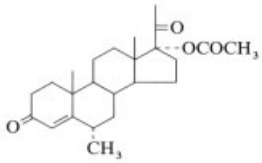 **60**	medroxyprogesterone acetate, 17-acetoxy-6α-methyl-4-pregnene-3,20-dione [a]	[*71-58-9*]	Clinovir (Upjohn); Gesinal (Hoechst); Provera (Upjohn); combination preparation with ethinyl estradiol: Provest (Upjohn)
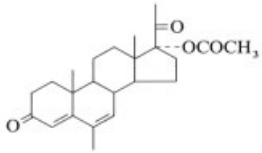 **61**	megestrol acetate, 17-acetoxy-6-methyl-4,6-pregnadiene-3,20-dione [a, b]	[*595-33-5*]	Megage (Mead Johnson); Ovarid (Glaxo); combination preparation with ethinyl estradiol or mestranol: Delpregnin (Novo); Volidan (British Drug Houses)
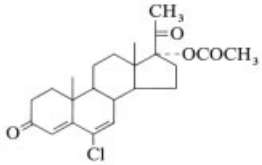 **62**	chlormadione acetate, 17-acetoxy-6-chloro-4,6-pregnadiene-3,20-dione [a, b]	[*302-22-7*]	Gestafortin (Merck); Lormin (Lilly); Lutoral (Syntex); combination preparation with mestranol: Ovosiston (VEB Jenapharm)
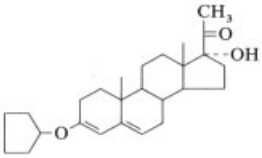 **63**	pentagestrone, 17α-hydroxy-3-cyclopentyloxy-3,5-pregnadien-20-one	[*7001-56-1*]	Gestovis, Gestovister (Vister)

[a] [130], [b] [131], [c] [177], [d] [178], [e] [179].

Table 13. Continued.

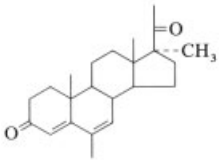 64	medrogestone, 6,17-dimethyl-4,6-pregnadiene-3,20-dione [c]	[*977-79-7*]	Colpro, Colprone (Ayerst); Prothil (Kali-Chemie)
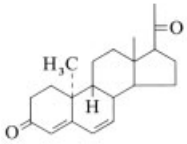 65	dydrogesterone, 9β,10α-pregna-4,6-diene-3,20-dione	[*152-62-5*]	Duphaston (Philips Duphar); Gestatron (Leo); Gynorest (Mead Johnson), Terolut (Ferrosan)
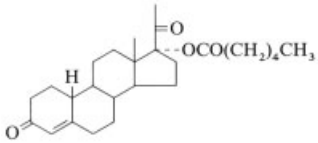 66	gestonorone caproate, 17-hexanoyloxy-19-nor-4-pregnene-3,20-dione [d]	[*1253-28-7*]	Depostat (Schering AG)
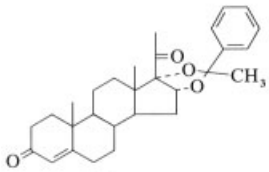 67	dihydroxyprogesterone acetophenonide, 16α,17-(1′-methyl-1′-phenylmethylenedioxy)-4-pregnene-3,20-dione [e]	[*1179-87-9*]	Deladroxone, Droxone (Squibb); combination preparation with estradiol enanthate: Deladroxat (Squibb)

[a] [130], [b] [131], [c] [177], [d] [178], [e] [179].

Table 14. Androstane gestagens

Formula	Name	CAS registry no.	Trade name (manufacturer)	Combination preparations with ethinyl estradiol (manufacturer)
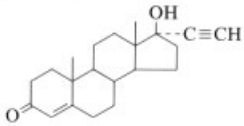 68	ethisterone, 17α-ethynyl-17β-hydroxy-4-androsten-3-one	[*434-03-7*]	Lutocyclin (Ciba); Lutoral (Schieffelin); Pranone (Schering Corp.)	Amenorone (Roussel); Menstrogen (Organon); Norma-sterin (Norma)
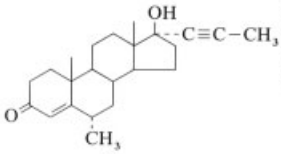 69	dimethisterone, 17β-hydroxy-6α-methyl-17α-(1-propinyl)-4-androsten-3-one	[*79-64-1*]	Secrosteron (British Drug Houses)	Oracon, Ovin (Mead Johnson); Secrodyl (British Drug Houses)

Table 15. 19-Norandrostane gestagens

Formula	Name	CAS registry no.	Trade name (manufacturer)
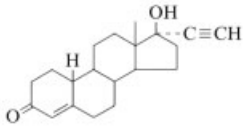 **29**	norethisterone, 17α-ethynyl-17β-hydroxy-4-estren-3-one [a], [b]	[*68-22-4*]	Micronett (Ortho); Micronor (Cilag); Norlutin (Parke Davis); Proluteasi (Tiber); combination preparation with mestranol: Norinyl (Syntex); Ortho-Novum (Cilag)
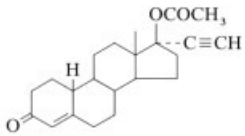 **70**	norethisterone acetate, 17α-ethynyl-17β-acetoxy-4-estren-3-one [a], [b]	[*51-98-9*]	Norlutate, Norlutin A (Parke Davis); Primolut Nor (Schering AG); combination preparation with ethinyl estradiol: Anovlar (Schering AG); Etalontin, Orlest (Parke Davis)
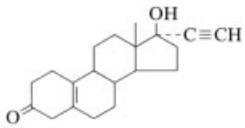 **71**	norethynodrel, 17α-ethynyl-17β-hydroxy-5(10)-estren-3-one [a], [b]	[*68-23-5*]	combination preparation with mestranol: Enavid, Enovid (Searle)
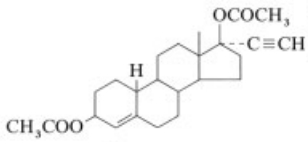 **72**	ethynodiol diacetate, 17α-ethynyl-3β,17β-diacetoxy-4-estrene	[*297-76-7*]	Demulen (Searle); combination preparation with mestranol: Ovulen (Searle)
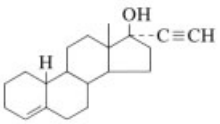 **73**	lynestrenol, 17α-ethynyl-4-estren-17β-ol	[*52-6-6*]	Orgametril (Organon); combination preparation with mestranol: Lyndiol (Organon); Noracyclin (Ciba)

[a] [130], [b] [131], [c] [162], [d] [163], [e] [160], [f] [161], [g] [158], [h] [180], [i] [182], [j] [183], [k] [184].

Table 15. Continued.

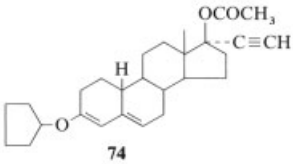 74	quingestanol acetate, 17α-ethynyl-17β-acetoxy-3-cyclopentyloxy-3,5-estradiene	[*3000-39-3*]	combination preparation with ethinyl estradiol: Reglovis (Vister); combination preparation with ethinyl estradiol-3-cyclopentyl-ether: Unovis (Warner-Chilcott)
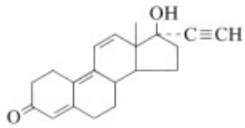 46	norgestrienone, 17α-ethynyl-17β-hydroxy-4,9,11-estratrien-3-one [c, d]	[*848-21-5*]	combination preparation with mestranol: Planor (Roussel-Uclaf)
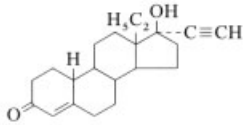 38	norgestrel (racemate), rac. 17α-ethynyl-13-ethyl-17β-hydroxy-4-gonen-3-one [e]		combination preparation with ethinyl estradiol: Duoluton, Eugynon (Schering AG); Stediril (Wyeth); combination preparation with estradiol valerate: Cyclo-Progynova (Schering AG)
	D-norgestrel, 17α-ethynyl-13-ethyl-17β-hydroxy-4-gonen-3-one [f]	[*6533-00-2*]	Microlut (Schering AG); combination preparation with ethinyl estradiol: Microgynon, Neogynon (Schering AG); Stediril-d (Wyeth)
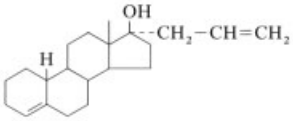	allylestrenol, 17α-allyl-17β-hydroxy-4-estrene [g]	[*432-60-0*]	Gestanin, Gestanon (Organon); Gestanyn (Pharmacia)
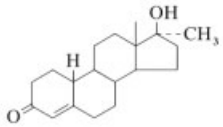	methylestrenolone, 17β-hydroxy-17α-methyl-4-estren-3-one [h]	[*514-61-4*]	Methalutin (Syntex); Orgasteron (Organon); combination preparation with methylestradiol: Gynaekosid (Boehringer Mannheim)
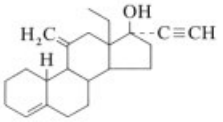 75	desogestrel [i]	[*54024-22-5*]	combination preparation with ethinyl estradiol: Marvelon (Organon)
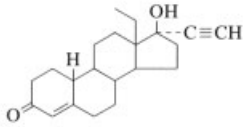 76	gestoden [j]	[*60282-87-3*]	combination preparation with ethinyl estradiol: Femovan (Schering AG)
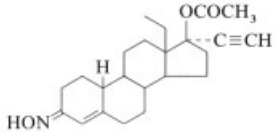 77	norgestimat [k]	[*35189-28-7*]	combination preparation with ethinyl estradiol: Cilest (Cilag)

[a] [130], [b] [131], [c] [162], [d] [163], [e] [160], [f] [161], [g] [158], [h] [180], [i] [182], [j] [183], [k] [184].

potent oral gestagens with strong antiovulatory activity has been derived from 19-norethisterone by methods such as acetylation (**70**, **72**, **74**), isomerization of the double bond (**71**), reduction (**72**), elimination of the 3-keto group (**73**), or enolether formation (**74**) [130, 131]. The totally synthetic gestagens–norgestrienone (**46**), norgestrel (**38**, racemate) and D-norgestrel (**38**), desogestrel (**75**), gestoden (**76**), and norgestimat (**77**)–all belong to this series.

3.7.5. Antigestagens [185]

Antigestagens may be either progesterone antagonists or inhibitors of progesterone synthesis.

Competitive progesterone antagonists are substances that inhibit gestagen function by blocking the progesterone receptor. The first compound found to have this property was mifepristone [*84371-65-3*] Roussel-Uclaf [RU 486, 11β- (4-dimethylaminophenyl)-17β-hydroxy-17α-(1-propynyl)estra-4,9-dien-3-one]. This substance can be administered in fertility control (nidation inhibition, menstruation commencement, abortion induction), obstetrics, and the treatment of hormone-dependent tumors.

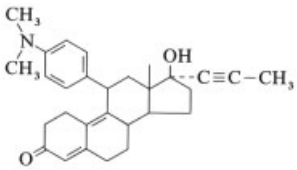

Mifepristone (Roussel-Uclaf)

Inhibitors of progesterone synthesis block the action of progesterone by preventing its formation [139]. At present, attempts are being made to inhibit the conversion of pregnenolone to progesterone by reversibly or irreversibly blocking the enzyme Δ^5-3β-hydroxysteroid dehydrogenase. Epostane [*80471-63-2*] (4α,5α-epoxy-17β-hydroxy-4β,17α-dimethyl-3-oxo-androstane-2α-carbonitrile) is being tested clinically [186]. Some nonsteroidal compounds (e.g., aminogluthetimide and antimycotics of the ketoconazole family) can also suppress steroid biosynthesis by enzyme inhibition.

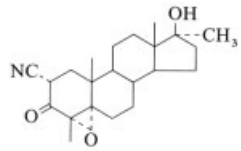

Epostane (Sterling-Winthrop)

4. Adrenal Steroid Hormones

(→ Steroids)

4.1. Introduction

Very few classes of substances exert so many physiologically and therapeutically useful effects as the steroid hormones of the adrenal cortex. These hormones, also called corticosteroids or corticoids, are divided into two groups depending on their physiological mode of action: the *glucocorticoids* promote glycogen formation from carbohydrates and gluconeogenesis from proteins, whereas the *mineralocorticoids* act primarily on electrolyte and water metabolism. The corticoids are thus responsible for the maintenance of important metabolic functions and are essential for life. The life of adrenalectomized animals is maintained by continuous corticoid administration (survival test) [191]. The "liver glycogen test" was developed especially for glucocorticoids [187].

Administration of glucocorticoids results in glycogen accumulation in the liver of normal or adrenalectomized, fed or fasting animals. The liver glycogen test measures the glycogen content of liver homogenate after the treated animals have been sacrificed.

The simultaneous increase of blood glucose and certain gluconeogenetic enzymes in the liver is a further useful parameter for determining the glucocorticoid potency of a test compound.

Both groups of substances are C_{21} steroids and are structurally related to the pregnane series.

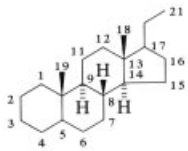

Pregnane skeleton

The important structural characteristics of these compounds are the α,β-unsaturated 3-keto-$\Delta^{4(5)}$ system and an α-hydroxyketone group in the 17β side chain. An 11β-hydroxyl group or 11-keto group is important for a predominantly glucocorticoid effect. In the mineralocorticoids, this group is absent (cortexone) or physiologically inactivated (e.g., lactol formation as in aldosterone). An additional hydroxyl group at the 17α-position increases glucocorticoid activity significantly.

Approximately 50 steroids of the pregnane (steroids with 21 carbon atoms), androstane (19 carbon atoms), and estrane series (18 carbon atoms) have been obtained from the adrenal cortex since the 1930s. These steroids have all been isolated and their structures analyzed, but only seven compounds are considered to be typical corticosteroids:

Glucocorticoids

1) Cortisol (**78**), also called hydrocortisone, 17α-hydroxycorticosterone, $11\beta,17\alpha$,21-trihydroxypregn-4-ene-3,20-dione
2) Cortisone, 17α,21-dihydroxypregn-4-ene-3,11,20-trione
3) Corticosterone, 11β,21-dihydroxypregn-4-ene-3,20-dione
4) 11-Dehydrocorticosterone, 21-hydroxypregn-4-ene-3,11,20-trione

Mineralocorticoids

1) Cortexone, deoxycorticosterone, 21-hydroxypregn-4-ene-3,20-dione
2) Cortexolone, 17α-hydroxydeoxycorticosterone, 17α,21-dihydroxypregn-4-ene-3,20-dione
3) Aldosterone (**79**), 18,21-dihydroxy-11β,18-oxidopregn-4-ene-3,20-dione

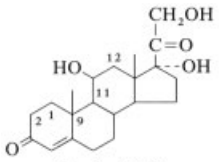

Cortisol (**78**)

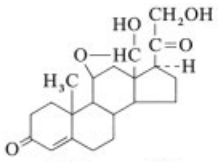

Aldosterone (**79**)

Of the steroids listed above, only cortisol, corticosterone, and aldosterone exhibit hormonal characteristics. The other corticoids, despite their often high physiological activity, are considered precursors or metabolites of the true adrenal cortical hormones.

4.2. Historical Aspects

The isolation and structural analyses of naturally occurring corticoids were carried out by Wintersteiner, Pfiffner, Kendall, and Reichstein. Crystalline aldosterone was first obtained in 1953 from adrenal cortical tissue by Reichstein and coworkers, who then proceeded to elucidate the structure and configuration of this steroid. The mineralocorticoid cortexone (commercially available since 1938) was the first corticoid to be used in the treatment of Addison's disease, a corticosteroid deficiency caused by damage or destruction of the adrenal glands. Cortexone can be synthesized easily. At first, cortisone and cortisol were employed primarily in the treatment of rheumatic disease, but after the mid-1950s, cortisone, cortisol, and especially the newly developed, chemically modified glucocorticoids achieved tremendous therapeutic and industrial importance.

The former industrial extraction of naturally occurring corticoids, especially cortisone and cortisol, from animal adrenal glands [196] is no longer important. It has been completely replaced by partially synthetic production methods.

4.3. Physiology and Mode of Action

The adrenal cortex is a peripheral endocrine gland whose hormone production (with the exception of aldosterone) is regulated by the anterior pituitary hormone, adrenocorticotropic hormone (ACTH). The formation of aldosterone is independent of ACTH; it is controlled primarily by the angiotensin – renin system and by the functional coupling of adrenal cortex, hypophysis, and hypothalamus.

The adrenal corticosteroid hormones play an important part in the metabolism of carbohydrates, proteins, and minerals. Their importance is exemplified by the fact that even slight stress to the organism causes a rapid increase in the activity of the adrenal cortex.

The adrenal steroid hormones cannot be strictly divided into mineralocorticoids and glucocorticoids because they all produce the effects of both groups. Thus, aldosterone exerts its prime action as a mineralocorticoid but also

has some glucocorticoid activity; for example, in the liver glycogen test [191], it is only slightly less potent than cortisone.

The *mineralocorticoids* are responsible for maintaining the correct distribution of sodium and potassium between the cell contents and the extracellular fluid and, consequently, for the maintenance of osmotic equilibrium. Malfunction of the adrenal cortex results in increased or diminished excretion of sodium and chlorine, with correspondingly diminished or increased excretion of potassium. Severe disturbances in water metabolism occur. Aldosterone exerts a much stronger effect on mineral metabolism than the other hormones. It has approximately 25 – 30 times the sodium-retaining activity and 5 times the potassium excretory activity of deoxycorticosterone. In humans, ca. 100 – 200 µg of aldosterone is synthesized daily.

In the case of natural or chemically modified *glucocorticoids,* a distinction must be made between their physiological action and their subsequently discovered therapeutically important pharmacodynamic effects. The primary physiological functions of glucocorticoids are to increase glycogen formation in the liver, raise the blood sugar level, and augment gluconeogenesis from amino acids by protein breakdown (catabolic effect). Moreover, glucocorticoids enable the organism to withstand stress (e.g., physical or psychological demands) and to maintain a normal internal state (homeostasis). In humans, 15 – 40 mg of cortisol is produced daily, but ten times this amount can be secreted under stress conditions. The amount of corticosterone synthesized daily is 1.3 – 4 mg.

Androgenic hormones, such as 11-hydroxyandrostenedione and androsterone, and to a small extent estrogenic hormones are also produced in the adrenal cortex [194].

4.4. Therapeutic Uses of Glucocorticoids

The therapeutic effects of glucocorticoids, particularly of chemically modified corticoids, are based primarily on the inhibition of proliferative reactions of mesenchymal tissue, especially allergic and inflammatory tissue reactions (e.g., polyarthritis, pleuritis). However, glucocorticoids are used to treat the symptoms but not the cause of these diseases. Glucocorticoids reduce the number of lymphocytes and eosinophilic granulocytes in the blood. Because their main effects are anti-inflammatory, antiexudative, cytotoxic, immunosuppressive, antitoxic, and antipruretic, they may be applied in the treatment of the following syndromes:

Substitution Therapy. Glucocorticoids can be used in acute and chronic adrenal cortical insufficiency, adrenogenital syndrome, thyroidal crises, Addison's disease (in combination with mineralocorticoids), and inhibition therapy of hyperadrenal cortical states.

Pharmacodynamic Therapy. The most important indications for glucocorticoid treatment are acute and chronic polyarthritis; rheumatic fever; allergic reactions such as asthma and obstructive bronchitis (→ Antiasthmatic Agents, Chap. 3.); cardiogenic, anaphylactic, and traumatic shock; stress situations; collagen disease; autoimmune disease; shock caused by burns and acid burns; poisoning; acute attack of gout; malignant tumors (especially leukemia and lymphogranulomatosis); and local application in a number of skin afflictions and dermatoses (especially eczema and exanthema, → Dermatotherapeutic Agents). In bacterial diseases, glucocorticoids can be administered with antibiotics or chemotherapeutic agents.

4.5. Therapy and Side Effects

Types of Therapy. Three types of corticoid therapy are possible: systemic, topical, and local. *Systemic therapy* can in turn be divided into oral application (administration of corticoids in the free 21-alcohol form or in the ester form as tablets, capsules, dragées, etc.); intravenous application (administration as a water-soluble 21-ester, e.g., 21-sodium hemisuccinate, 21-disodium phosphate, or 21-dialkylamino carboxylate); and intramuscular application (administration of corticoids as a free 21-alcohol or a short- or long-chain 21-ester). In *topical corticoid therapy,* free 21-alcohols or their esters are injected as crystal suspensions into a joint or tissue cavity.

Local corticoid therapy is used commonly to treat skin disease. Corticoids are again applied as the free 21-alcohols or as specially structured 21- or 17-esters in ointments, creams, lotions, or sprays. A broad-spectrum antibiotic is usually

added to local corticoid preparations to prevent bacterial infection (e.g., 0.5 % neomycin sulfate or 0.1 % gentamicin).

Side Effects. Medium- or long-term systemic application of glucocorticoids can cause the following side effects: hypercorticism or Cushing's syndrome (adiposity of the trunk, moon-shaped face, steroid diabetes, hypertension); tumor formation in the gastrointestinal tract; retardation of granulation and scar formation; osteoporosis; hypophyseal inhibition; atrophy of the adrenal cortex; psychological effects (e.g., euphoria); increased risk of infection; endocrine disturbance; morphological changes of the skeletal muscle and muscular atrophy; or adynamia and fatigue.

The central mineralocorticoid side effects of cortisone and cortisol have been eliminated to a large extent by the development of chemically modified glucocorticoids. To reduce the risk of side effects in systemic anti-inflammatory therapy, glucocorticoids are often combined with nonsteroidal anti-inflammatory agents (→ Analgesics and Antipyretics, Chap. 3.).

Dosage. The dose equivalence is used to compare the effects of systemically administered glucocorticoids; it is the glucocorticoid or anti-inflammatory activity relative to that produced by cortisol (cortisol equivalence = 1). The average maintenance dose in milligrams per day is also important in establishing the dosage. Both values are listed in Table 16 for the most important glucocorticoids used systemically in medicine today; the Cushing threshold dosage is also given.

4.6. Biosynthesis and Metabolism

Biosynthesis of the adrenal corticosteroid hormones starts with the conversion of cholesterol to pregnenolone, which is subsequently oxidized to progesterone (see Fig. 4). This compound is then converted to the glucocorticoids by primary hydroxylation at C-17 and subsequent hydroxylation at C-21 (cortisone) and C-11 (cortisol). In contrast, biosynthesis of the mineralocorticoids is initiated by primary hydroxylation at C-21. The C-17 hydroxylase is unable to attack the C-21 hydroxylated product. Glucocorticoids are produced in the central zona fasciculata and the mineralocorticoids in the outer zona glomerulosa of the adrenal cortex.

Corticoid metabolism occurs primarily in the liver, the major enzymatic degradation reactions being [190]

1) hydrogenation of the double bonds at the 4,5-positions and, in the case of $\Delta^{1,4}$ corticoids, the 1,2-positions to give products with a saturated ring A;
2) reduction of the 3-keto, 20-keto and, in the case of 11-oxocorticoids, 11-keto groups to the corresponding hydroxyl groups; and
3) oxidative cleavage of the side chains at the 17β-position to give 17-keto steroids.

The metabolites of adrenal cortical steroids are excreted in the urine as 3-glucuronides or 3-sulfates, e.g., as glucuronides of type **80** (tetrahydrocortisol derivatives) or type **81** (tetrahydroandrostane derivatives).

Table 16. Cortisol equivalence, average maintenance dose, and Cushing threshold dose of the most important glucocorticoids used in systemic human therapy listed in order of increasing effectiveness [193, 197]

Glucocorticoid	Cortisol equivalence (cortisol = 1)	Average maintenance dose, mg/d	Cushing thresholddose, mg/d
Cortisone	0.8	10 – 50	50
Cortisol	1.0	10 – 50	40
Prednylidene	2	12 – 15	18 – 24
Prednison, prednisolone	4	5 – 15	10
Fluocortolone	4	2.5 – 12.5	10 – 15
6α-Methylprednisolone	5	6 – 10	8
Triamcinolone	5 – 10	6 – 8	8
Paramethasone	10	2 – 6	4 – 6
Fludrocortisone	15		
Dexamethasone	20 – 25	1 – 2	2
Betamethasone	25 – 50	1 – 1.5	2

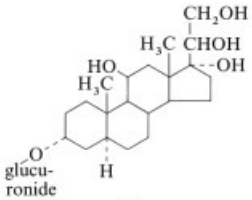

80

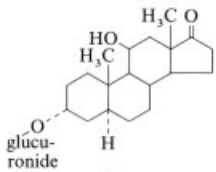

81

4.7. Detection Methods

The special detection methods used in the analysis of corticoids are almost all based on chemical reactions specific for the side chain at position 17. Important reactions are degradation with periodic acid and chromic acid [187], and oxidation of the C-17 side chain with sodium bismuthate to give the 17-keto derivative (e.g., androsterone) or 17-ethio acid (17β-carboxylic acid) [198]. Another method for identification of the dihydroxyacetone side chain is the Porter–Silber reaction [199], which is based on the formation of a yellow dye on heating with phenylhydrazine and sulfuric acid. This substance has a characteristic UV absorption band at 410 nm, and its molar extinction can be measured easily. For further details on chemical, spectral, and radioactive detection methods for corticoids, see [194].

4.8. Naturally Occurring Glucocorticoids

4.8.1. Properties and Uses

Cortisone [*53-06-5*] (**82**), $C_{21}H_{28}O_5$, M_r 360.46, *mp* 220–224 °C, $[\alpha]_D^{20}$ +209° (95 % ethanol), and **cortisol** [*50-23-7*] (**78**),

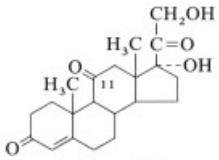

Cortisone (**82**)

$C_{21}H_{30}O_5$, M_r 362.47, *mp* 217–220 °C, $[\alpha]_D^{20}$ +167° (ethanol), are glucocorticoid hormones whose importance has diminished considerably because of their relatively low potency and, above all, their pronounced side effects (especially the strong mineralocorticoid action). They have been replaced by chemically modified corticoids and are used only in exceptional cases, namely, in substitution therapy and in special preparations for local application (e.g., in ear, nose, and throat conditions and in ophthalmology). However, cortisone and cortisol are even more important today as intermediates in the synthesis of chemically modified corticoids. Because totally synthetic methods are not economically feasible, the industrial production of cortisone and cortisol is still based on a series of efficient, partially synthetic methods. These methods use inexpensive, simple plant steroids as raw materials.

Total synthesis of cortisone was achieved in 1951 by Woodward and coworkers [200] and a year later by Sarett and coworkers [201]. A total synthesis of cortisone had already been published in 1950 by Wendler and coworkers [202]. These methods are solely of scientific interest and have no industrial importance. Although the large-scale corticoid total syntheses developed by Roussel–Uclaf have been improved considerably, they are still not as economical as the partially synthetic processes [203–205].

Trade Names. Despite their minor therapeutic importance, a large number of cortisone and cortisol preparations are still available on the market. Only a few typical representatives can be listed here: Cortisone: Cortisone "Ciba" (Ciba, Wehr, FRG); Scheroson (Schering).

Cortisol: Scheroson F (Schering); Ficortril (Pfizer, FRG); Hydrocort (Ferring, FRG); Hydrocortisone "Hoechst" (Hoechst).

Corticosterone [*50-22-6*], $C_{21}H_{30}O_4$, M_r 346.45, *mp* 180–182 °C, $[\alpha]_D^{25}$ +223° (alcohol), and **11-dehydrocorticosterone** [*72-23-1*],

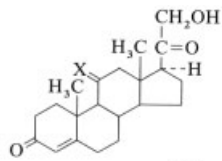

Corticosterone, X = HO, H
11-Dehydrocorticosterone, X = O

$C_{21}H_{28}O_4$, M_r 344.43, *mp* 178–180 °C, $[\alpha]_D^{25}$ +258 ° (alcohol), are two other glucocor-

ticoids. They do not have a hydroxyl group at position 17; therefore, they can be considered as 17-deoxycortisol and 17-deoxycortisone, respectively. In comparison with cortisol and cortisone, these compounds have a lower glucocorticoid activity but exert a stronger mineralocorticoid effect. For this reason, these corticoids have achieved neither industrial nor therapeutic importance. For partial synthesis, see [189].

4.8.2. Partial Synthesis of Cortisone [191]

The most important methods for the industrial production of cortisone (**82**) from steroid raw materials are

1) production from progesterone,
2) production from 16-dehydropregnenolone via cortexolone, and
3) production from bile acid.

The first two methods are more important because they use an elegant microbiological process to insert the oxygen function at C-11.

Production from Progesterone. Production of cortisone from progesterone was developed in 1952 and is important because progesterone can be produced cheaply from diosgenin (see Section 3.2.2). Progesterone is converted to 11α-hydroxyprogesterone (**83**) by a high-yield (90 %) microbiological oxidation with the mold *Rhizopus arrhizus* or *R. nigricans* [206, 207]. The method used by Upjohn proceeds as shown in Figure 8 [208]. 11α-hydroxyprogesterone (**83**) is oxidized with chromium trioxide to 11-oxoprogesterone, which in turn is converted to the sodium salt of 21-(ethoxalyl)-11-oxoprogesterone (**84**) with diethyl oxalate and sodium ethoxide [209]. The next step involves treatment with 2 molar equivalents of bromine. The resulting 21-dibromo derivative is not further isolated but is subjected to a Favorskii rearrangement in the presence of sodium methoxide, to yield compound **85**. The 3-oxo-Δ^4 group of this compound is protected by formation of the enamine (**86**). Reduction of compound **86** with lithium aluminum hydride leads to the formation of compound **87**, in which both the 21-carbomethoxy and the 11-keto groups are reduced to alcohol groups. After hydrolytic cleavage of the enamine group

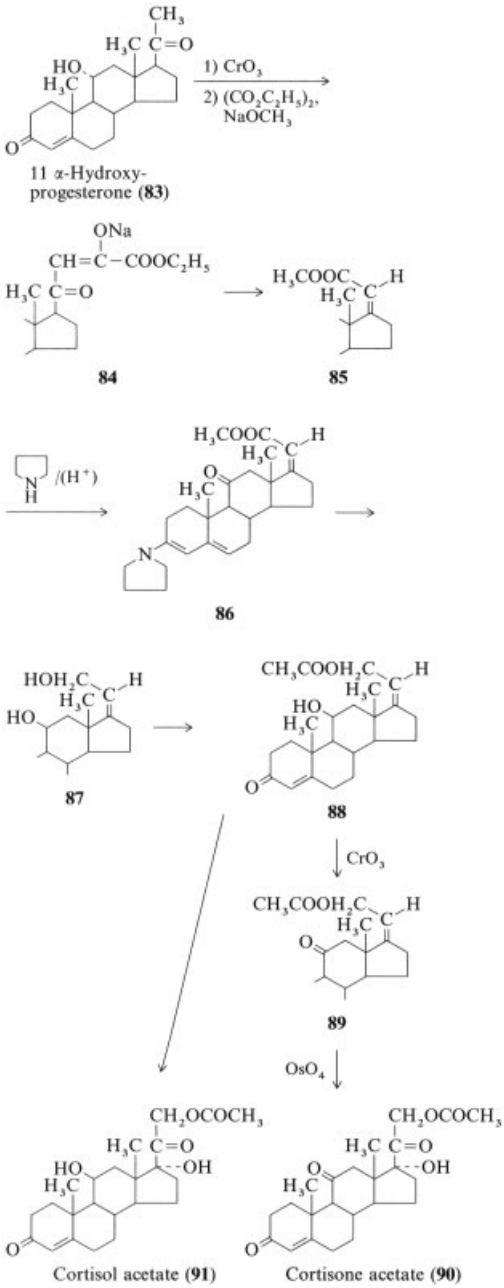

Figure 8. Synthesis of cortisone from 11α-hydroxyprogesterone

and esterification of the 21-hydroxyl group with acetic anhydride – pyridine, compound **88** is obtained (the 11β-hydroxyl group is not esterified for steric reasons). Compound **88** is subjected to chromic acid oxidation and then ef-

fectively converted to cortisone acetate (**90**), by using a method developed at Upjohn, with *N*-methylmorpholine *N*-oxide and osmium tetroxide [210]. The total yield is 20 – 25 %, based on 11α-hydroxyprogesterone (**83**).

Production from 16-Dehydropregnenolone. 16-Dehydropregnenolone (**92**) is prepared easily on a large scale from diosgenin (see Section 3.2.2). A method developed by JULIAN [211] and modified by Syntex [212] requires several steps to convert this compound to cortexolone (Fig. 9; **98**).

16-Dehydropregnenolone (**92**), obtained by hydrolysis of 16-dehydropregnenolone acetate, is converted with alkaline hydrogen peroxide to 16α,17α-oxidopregnenolone (**93**), which in turn reacts with hydrobromic acid to give the bromohydrin (**94**). After removal of the 16β-bromine atom by hydrogenation with palladium – hydrogen and partial esterification of the 3-hydroxyl group with formic acid, 3β-formoxy-17α-hydroxypregnene (**95**) is formed. This compound can also be prepared by using the method described in [213]. Compound **95** is converted to the tribromo derivative (**96**) with 2 molar equivalents of bromine; then the Δ^5 double bond is regenerated with sodium iodide, and the 21-bromine group is replaced by the 21-acetoxy group with potassium acetate. Esterification of the 17α-hydroxyl group with acetic anhydride and toluene-*p*-sulfonic acid gives the diester (**97**). Direct Oppenauer oxidation (during which the 3-formoxy group near the less vulnerable acetate groups at C-17 and C-21 is partially saponified) and subsequent alkaline saponification of the 17- and 21-acetate groups yield cortexolone (**98**). The total yield of **98** based on **95** is 45 – 50 %. Again, the hydroxyl group at C-11 is introduced microbiologically. Compound **98** is reacted with the fungus *Cunninghamella blakesleeana* or *Curvularia lunata* to yield up to 77 % cortisol and 22 % cortisone [214, 215]. Microbiological hydroxylation with *Absisia glauca* converts compound **98** to a mixture of equal parts cortisol (**78**) and 11α-cortisol (**99**), the 11α-epimer of cortisol, with a total yield of 90 % [216, 217]. Compound **99** is separated from cortisol by means of its acetate and then oxidized to cortisone (**82**). Another Upjohn method permits the specific hydroxylation of compound **98** to 11α-cortisol **99** and gives a yield of ca. 85 %. 11α-Cortisol can subsequently be oxidized to cortisone with chromium trioxide [218, 219].

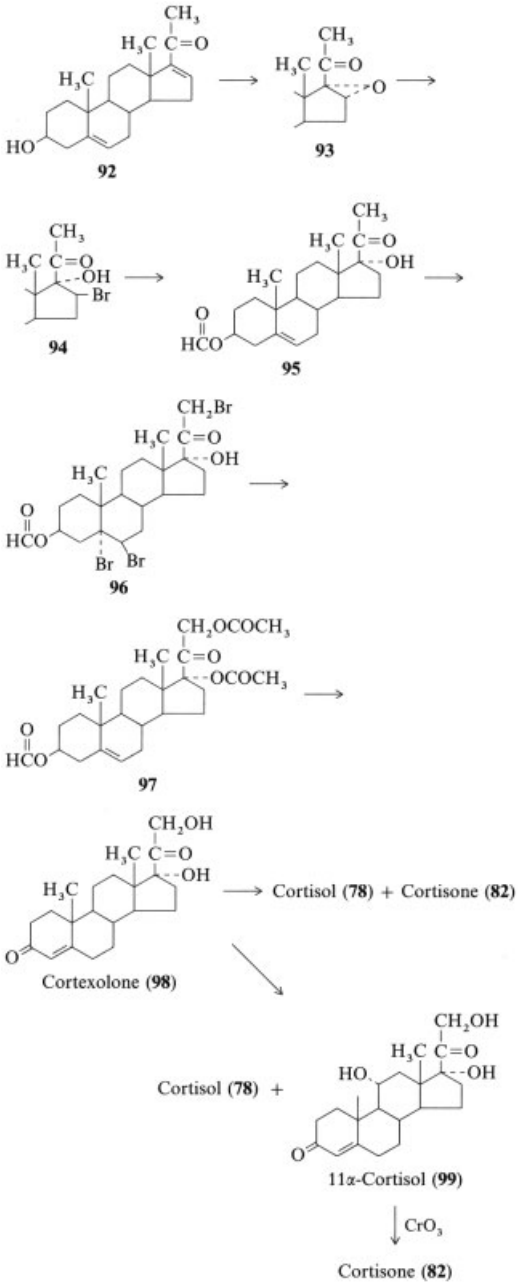

Figure 9. Synthesis of cortisone from 16-dehydropregnenolone

Production from Bile Acid. Cortisone was first produced from bile acid by SARETT [220] and by KENDALL of the Merck Corporation [221]. Bile acid (i.e., cholic acid [*81-25-4*], **100**) is isolated from the bile of slaughtered animals and has a hydroxyl group at position 12. Although the synthesis of cortisone from cholic acid originally entailed 36 reaction steps, compared with the considerably shorter syntheses from progesterone and 16-dehydropregnenolone, it is still used today.

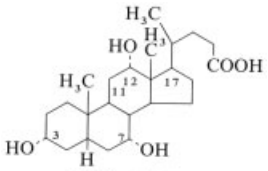

Cholic acid (**100**)

A large number of improvements and changes have been made (reviewed in [188]), and recommendation of any one variation is difficult (cf. [187–189, 191, 195]). However, all variations share the following basic concepts:

1) Elimination of the 7α-hydroxyl group of cholic acid to give 7-deoxycholic acid [209]
2) Transfer of the hydroxyl function at position 12 of deoxycholic acid to position 11
3) Partial degradation of the side chain at C-17 of the bile acid and conversion to the 17-dihydroxyacetone side chain characteristic of corticoids
4) Introduction of the 3-keto-4,5-dehydro structure into the corticoid molecule

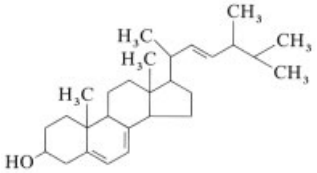

Ergosterol (**101**)

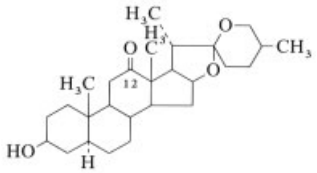

Hecogenin (**102**)

Miscellaneous Methods. Other commercial methods of cortisone production start with ergosterol (**101**) and stigmasterol [222], phytosterols of yeast and soybean oil, or with the sapogenin hecogenin (**102**), obtained as a waste product in the recovery of sisal [223, 224] (see left figure below). However, these methods no longer have any importance.

4.8.3. Partial Synthesis of Cortisol

Pregnane derivatives are generally used as starting materials in the partially synthetic commercial production of cortisol. These derivatives already have an 11-keto or 11α-hydroxyl group and are obtained in the production of cortisone from progesterone and 16-dehydropregnenolone described in Section 4.8.2. The Upjohn process (Fig. 8) allows conversion of 11α-hydroxyprogesterone (**83**), obtained microbiologically from progesterone, to cortisol [196, 225].

In the American Cyanamide method, cortisol (**78**) is produced from 11α-cortisol (**99**) or cortisone (**82**) with a yield of 72 % [196, 226]. Reaction with ethylene glycol and subsequent oxidation with chromium trioxide yield the 3,20-bisketal (**103**).

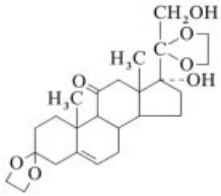

103

Treatment with sodium borohydride stereospecifically reduces the 11-keto group to an 11β-hydroxyl group, which is finally hydrolyzed to cortisol (**78**).

Another cortisol production process, developed by Upjohn [227], starts with pregnane-3α,17α-diol-11,20-dione (**104**) obtained from progesterone [228]. Here, the 3α-hydroxyl group is oxidized to a 3-keto group with *tert*-butyl hypochlorite with simultaneous chlorine substitution at position 4. The bisketal (**105**) is subsequently obtained by reaction with ethylene

glycol. Conversion of the bisketal to cortisol requires six additional steps and proceeds with a yield of 15 % [196].

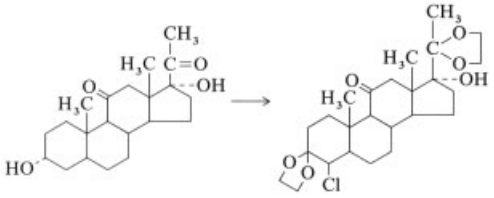

104 **105**

Other commercially important biochemical methods are available in which cortexolone (**98**) can be stereoselectively hydroxylated at the 11β-position to give cortisol (**78**) with yields of approximately 40 %. This is achieved microbiologically by using *Cunninghamella blakesleeana* [229] or *Curvularia lunata* [230] (see also Fig. 9 and [231]).

The production of cortisol (**78**) from cortisone acetate (**90**) [232] involves formation of the 3,20-bissemicarbazone. The 11-keto function does not react with semicarbazide for steric reasons. Subsequently, the 11-keto group is reduced with potassium borohydride, whereby the 11-hydroxyl group formed acquires the desired β-configuration and the 21-acetate group is simultaneously hydrolyzed. Cortisol is finally obtained by hydrolytic cleavage of the semicarbazone group with nitric acid.

4.9. Naturally Occurring Mineralocorticoids and Antagonists

The most important mineralocorticoids are deoxycorticosterone (**112**), also called *cortexone,* and the more powerful, physiologically more important *aldosterone* [*52-39-1*] (**79**). *Spironolactone* (**115**) is an important synthetic aldosterone antagonist. In comparison with the glucocorticoids, the mineralocorticoids are seldom used as therapeutic agents because specific mineralocorticoid activity is required in only a few conditions, e.g., in the treatment of Addison's disease, a rare condition that requires substitution therapy, and (with glucocorticoids) in stress situations. Cortexone is commercially important as an intermediate or starting material in the synthesis of other steroids.

Although a number of interesting studies have been carried out on the synthesis of aldosterone (**79**), it has little importance as a therapeutic agent and intermediate. Aldosterone has 20 – 30 times the mineralocorticoid potency of cortexone 21-acetate and one-third the glucocorticoid activity of cortisone, whereas cortexone acetate has no significant glucocorticoid activity.

Both mineralocorticoids are administered systemically. Cortexone is generally given parenterally as an oily solution of a 21-ester (e.g., acetate, cyclopentylpropionate, trimethylacetate, or heptanoate). Substitution therapy is indicated in chronic adrenal cortical insufficiency (Addison's disease) and in adrenogenital salt-loss syndrome. Hypertension and advanced cerebral sclerosis are contraindications to the use of mineralocorticoids.

4.9.1. Cortexone

Cortexone [*64-85-7*], deoxycorticosterone, $C_{21}H_{30}O_3$, M_r 330.45, *mp* 141 – 142 °C, $[\alpha]_D^{22}$ + 178 ° (ethanol), is one of the most important mineralocorticoids.

Production. Three methods are used for the production of cortexone:

1) *Production from Pregn-5-en-3β-ol-20-one.* Pregnenolone (**106**) is subjected to Ruschig condensation with diethyl oxalate to give the sodium salt of the corresponding enol (**107**) [209], which is then converted with iodine to the iodine ester (**108**). Subsequent ketonic cleavage with sodium methoxide gives compound **109**, which yields the precursor (**110**) of cortexone on treatment with potassium acetate in aqueous acetone. A 71 % yield is obtained, based on compound **106** as starting material. Oppenauer oxidation of the 3β-hydroxyl group yields the 3-keto-Δ^4 group, and saponification of the 21-acetate group of **111** with 1 mol equivalent of sodium methoxide produces almost quantitative yields of cortexone (**112**).
   This synthesis can be carried out easily on a large scale because only two of the intermediates, compounds **108** and **110**, must be isolated.

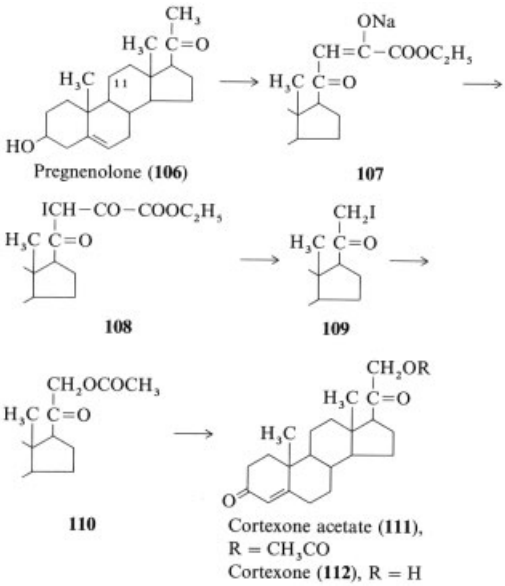

2) *Microbiological Production from Progesterone.* Progesterone is specifically hydroxylated at position 21 by a microbiological method developed by Ciba, which uses fungal strains of scoleco spores, e.g., *Ophiobolus herbotrichus.* Cortexone (**112**) is formed with a yield of 60 % [233].
3) *Chemical Production from Progesterone.* In the commercially important process developed by Syntex, iodine reacts directly with progesterone in tetrahydrofuran – methanol in the presence of calcium carbonate to give 21-iodoprogesterone. The iodoprogesterone is then converted to cortexone 21-acetate (**111**) with potassium acetate in a mixture of acetic acid and acetone [234].

Trade Names. Cortexone is available under the following names: Cortenil depot (cyclopentylpropionate, Hoechst); Cortiron (acetate, Schering); Cortiron depot (heptanoate, Schering); Percoten (Ciba, Wehr).

4.9.2. Aldosterone

Aldosterone [*52-39-1*] (**79**), $C_{21}H_{28}O_5$, M_r 360.44, has the following properties: *mp* 108 – 112 °C (contains water of crystallization), *mp* 164 °C (anhydrous), $[\alpha]_D^{23}$ + 152.2 ° (anhydrous acetone).

Production. After the structure of aldosterone had been established, its total synthesis was achieved by WETTSTEIN, SZPILFOGEL, PRELOG, REICHSTEIN, and their coworkers. For a review of these syntheses, see [191].

The partial syntheses of aldosterone are of greater preparative importance [191]; however, only the most significant methods will be mentioned here. One method converts corticosterone acetate (corticosterone, see page 54) in four steps to aldosterone acetate [235]. Another method developed by Ciba starts with 11α-acetoxypregnenolone (cf. **106**), which is converted to compound **113** with lead tetraacetate and iodine and, after another 16 reaction steps, to aldosterone (**79**) [236].

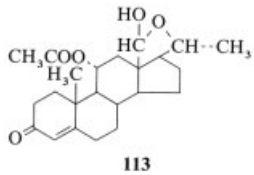

The synthesis of aldosterone developed by VELLUZ and coworkers at Roussel-Uclaf is described in [237].
Trade Name. Aldosterone is marketed unter the trade name Aldocorten (Ciba, Wehr).

4.9.3. Spironolactone

Spironolactone [*52-01-7*] (**115**), $C_{24}H_{32}O_4S$, M_r 416.69, *mp* (double melting point) 134 – 135 °C and 201 – 202 °C, $[\alpha]_D^2$ – 33.5 ° (chloroform), the synthetic aldosterone antagonist, is of considerable therapeutic importance. It increases excretion of sodium and water, and is used as an oral saluretic. Treatment with aldosterone antagonists is indicated in cirrhosis of the liver with ascites and edema, water retention of cardiac and renal origin, hypertension resistant to antihypertonic agents containing saluretics, and in myasthenic and hypokalemic crises. Therapy is contraindicated in kidney insufficiency with anuria and hyperkalemia.

Production. The production of spironolactone from 17α-ethynylandrost-5-ene-3β,17β-diol (**114**) [238, 239] involves the following steps (Searle) [240]:

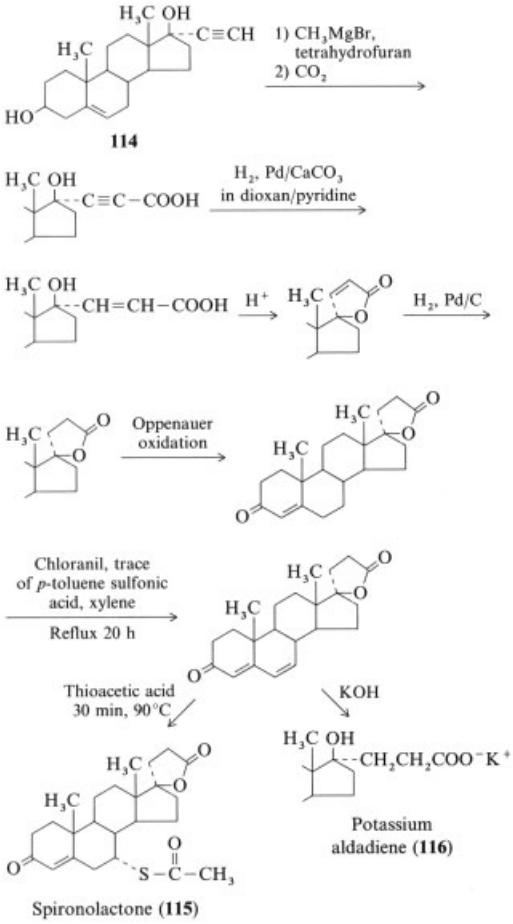

Trade Names. Spironolactone for oral administration is sold under the names Aldactone (Searle, Boehringer Mannheim) and Osyrol (Hoechst). Potassium aldadiene [*2181-04-6*] (**116**) is used for intravenous administration: Aldactone for injection (Searle, Boehringer Mannheim) and Osyrol for injection (Hoechst).

4.10. Chemically Modified Adrenal Steroid Hormones

A large number of chemically modified, highly potent corticosteroids have been prepared for therapeutic purposes by substitution or dehydrogenation reactions at positions 1, 2, 6, 9, 12, 16, and 21 of the corticosteroid skeleton. These compounds have greatly diminished the importance of naturally occurring cortisone derivatives with their severe mineralocorticoid side effects. However, despite intensive efforts, adequate reduction of the severe side effects that accompany systemic glucocorticoid therapy (e.g., Cushing's syndrome or osteoporosis) has not yet been achieved. The emphasis in corticoid therapy has changed from systemic administration to local or topical application because the associated side effects are then minor. A large number of corticoids have been developed for local or topical administration.

4.10.1. 1,2-Dehydrocorticosteroids

Introduction of a second double bond at the 1,2-positions of the cortisol molecule by HERSHBERG and coworkers of the Schering Corporation in 1955 represented a breakthrough in the synthesis of modified corticoids [241]. 1,2-Dehydrogenation yields derivatives with glucocorticoid, anti-inflammatory, and antiallergenic effects that are four to five times higher than those of cortisone. At the same time, these substances have one-third the mineralocorticoid activity of cortisone [242]. The 1,2-dehydro derivatives of cortisone and cortisol, *prednisone* (**117**) and *prednisolone* (**118**), respectively, do not have significant mineralocorticoid activity and are, therefore, widely used in general, systemic pharmacodynamic therapy, particularly as anti-inflammatory and antirheumatic agents.

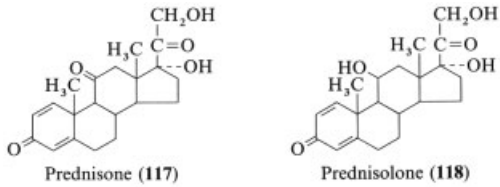

Prednisone [*53-03-2*] (**117**), 1,2-dehydrocortisone, $C_{21}H_{26}O_5$, M_r 368.44, *mp* 233 – 235 °C, $[\alpha]_D^{25}$ + 172 ° (dioxane), and **prednisolone** [*50-24-8*] (**118**), 1,2-dehydrocortisol, $C_{21}H_{28}O_5$, M_r 360.44, *mp* 240 – 241 °C, $[\alpha]_D^{25}$ + 102 ° (dioxane), are the two most important 1,2-dehydrocorticosteroids.

Production. Insertion of the 1,2-double bond is achieved by chemical or microbiological methods, the latter being more economical. Prednisone can be produced chemically by

a process developed by the Schering Corporation from pregnane-17α,21-diol-3,11,20-trione 21-acetate (**119**) involving a 1,4-dibromide and subsequent hydrogen bromide elimination with pyridine [242]. Other methods use the 1,2-dehydrogenation of cortisone with selenium dioxide [243, 244].

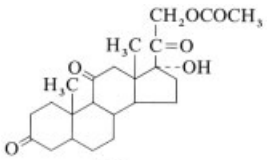

119

The industrially important biological methods permit high-yield production of prednisone and prednisolone by microbiological dehydrogenation at the 1,2-positions of cortisone (**82**) and cortisol (**78**), respectively [245], with *Corynebacterium simplex* [246], *Rhizopus nigricans,* or *Bacillus subtilis* [247].

The chemical conversion of prednisone to prednisolone has also attained some commercial significance. This reaction proceeds via the 3,20-semicarbazone of prednisone and is very similar to the analogous conversion of cortisone to cortisol (see Section 4.8.3) [242].

Effects and Uses. Prednisone and prednisolone have four to five times the anti-inflammatory activity of cortisone when administered orally or parenterally, but only two to three times its activity when applied locally. In contrast, their mineralocorticoid potency is only one-third that of cortisone. The indications listed in Section 4.4 apply to both compounds, but prednisolone has become the "standard" corticoid in general practice. In comparison with cortisone and cortisol, the mineralocorticoid side effects are very slight.

Trade Names. A large number of commercial preparations are derived from prednisolone and prednisone; only a few examples are given here: Prednisolone: Decortin-H (Merck); Deltacortril (Pfizer); Hostacortin-H (Hoechst); Scherisolon (Schering); and Ultracorten-H (Ciba, Wehr).

Prednisone: Decortin (Merck); Hostacortin (Hoechst); Keteocort (Desitin, Hamburg); and Ultracorten (Ciba, Wehr).

4.10.2. 9α-Halocorticosteroids

In 1953 scientists at Syntex succeeded in inserting halogen atoms at the 9α-position of the cortisone molecule. The resulting modified corticoids had many times the glucocorticoid or anti-inflammatory activity of cortisone [215]. The glucocorticoid activity of hydrocortisones and cortisones with a halogen substituent at position 9α increases dramatically in the order: iodine, bromine, chlorine, fluorine. In comparison with cortisone acetate (activity = 1), the following activities are obtained for the 9α-halocorticoids in the liver glycogen test: I = 0.1, Br = 0.3, Cl = 5, F = 11. The value for 9α-fluorocortisone is 9. The mineralocorticoid effect also increases with the glucocorticoid activity; for example, 9α-fluorocortisol (**123**) has two to four times the mineralocorticoid activity of cortexone. Consequently, the halogen preparations have a very limited therapeutic applicability, despite their high glucocorticoid activities.

9α-Fluorocortisol [*127-31-1*] (**45**) is also known as fluorocortisone, $C_{21}H_{29}FO_5$, M_r 380.46, *mp* 260 – 262 °C, $[\alpha]_D^{23}$ + 139° (95 % ethanol).

Production. A method developed by Syntex uses 11α-cortisol 21-acetate (**120**) as the starting material [248]. This compound is converted to the 11α-tosylate, and subsequent elimination of the tosylate group with sodium acetate yields compound **121**. Stereospecific addition of hypochloric or hypobromic acid at the 9,11-double bond yields 9α-chloro- or 9α-bromocortisol 21-acetate, respectively. The 9,11β-epoxide (**122**) is obtained by treating the chloro- or bromohydrin with alkali. Introduction of fluorine or iodine at the 9α-position is achieved by cleavage of **122** with hydrofluoric or hydriodic acid (see upper one of next two figures).

Preparation of the corresponding 1,2-dehydrocorticoid, 9α-fluoroprednisolone (**124**), is described in [196]. The industrial and therapeutic significance of this compound has diminished (see lower one of next two figures).

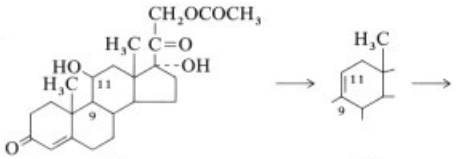

120 **121**

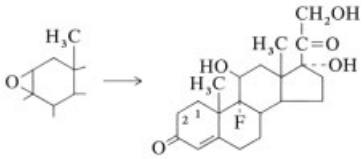

122 9α-Fluorocortisol (**123**)

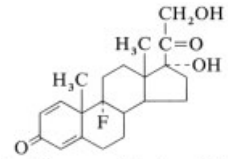

9α-Fluoroprednisolone (**124**)

Effects and Uses. Animal experiments have shown that 9α-fluorocortisol has almost 10 times the glucocorticoid or anti-inflammatory activity of cortisol, but its mineralocorticoid potency is approximately 15 – 20 times that of cortisone. 9α-Fluoroprednisolone has approximately 20 times the glucocorticoid potency of cortisol; its mineralocorticoid activity is the same as that of 9α-fluorocortisol.

9α-Fluorocortisol and its derivatives are primarily used locally in dermatology and ophthalmology in the form of ointments, cream, or drops. They are also used topically in rheumatology in the form of crystal suspensions for intraarticular injections. Oral substitution therapy in tablet form (0.1 mg per tablet) is indicated in cases of adrenogenital salt-loss syndrome and for raising blood pressure in constitutional hypotension (→ Antihypotensives, Chap. 3.2.).

Trade Names. The following trade names are used:

9α-Fluorocortisol: Astonin (Merck); Scherofluron (21-acetate, Schering); Florinef (Squibb).

9α-Fluoroprednisolone: Predef for veterinary medicine (21-acetate, Upjohn).

4.10.3. 16-Hydroxycorticosteroids

Scientists at Lederle succeeded in almost completely eliminating the mineralocorticoid side effects of 9α-fluoroprednisolone by introducing a 16α-hydroxyl group into the steroid nucleus [249]. Although the glucocorticoid and anti-inflammatory activities of 16α-hydroxycorticoids are approximately one-fifth those of 9α-fluoroprednisolone, they are still of tremendous therapeutic significance because they have practically no mineralocorticoid effects. A large number of derivatives have been obtained from the best known 16α-hydroxycorticoid, triamcinolone (**129**). The 16α,17α-acetonides have a much higher anti-inflammatory effect than the parent compounds. Surprisingly, they even exhibit a diuretic effect when pharmacologically tested for mineralocorticoid activity [250,251]. For this reason, the 16α,17α-acetonides are now indispensable in local corticoid therapy.

Triamcinolone [*124-94-7*] (**129**), $C_{22}H_{27}FO_6$, M_r 394.45, *mp* 269 – 271 °C, $[\alpha]_D^{25}$ + 75° (acetone) and **triamcinolone acetonide** [*76-26-5*] (**130**), also called triamcinolone 21, $C_{24}H_{31}FO_6$, M_r 434.49, *mp* 292 – 294 °C, $[\alpha]_D^{23}$ + 109° (chloroform), are the most important hydroxycorticosteroids.

Production. Triamcinolone (**129**) is prepared commercially via a method developed by Bernstein and coworkers [252]. By starting with cortisol 21-acetate (**91**), reaction with ethylene glycol gives the 3,20-bisethylene ketal (**125**). The 11β- and 17α-hydroxyl groups of this compound are then dehydrated with thionyl chloride in pyridine, followed by alkaline saponification of the 21-acetate group, acid deketalization of the cyclic acetal groups at positions 3 and 20, and reacetylation of the 21-hydroxyl group to yield compound **126**. Selective hydroxylation of the double bond at position 16 with osmium tetroxide in pyridine – benzene gives compound **127**. The 9,11-double bond of this compound is converted to a 9α-fluoro-11β-hydroxyl group (**128**) by the method described for 9α-fluorocortisol (**123**) in Section 4.10.2. Triamcinolone (**129**) is obtained after alkaline saponification of the 21-acetate group and microbiological introduction of the

Δ^1 double bond with *Corynebacterium simplex*. The total yield is not given in the literature.

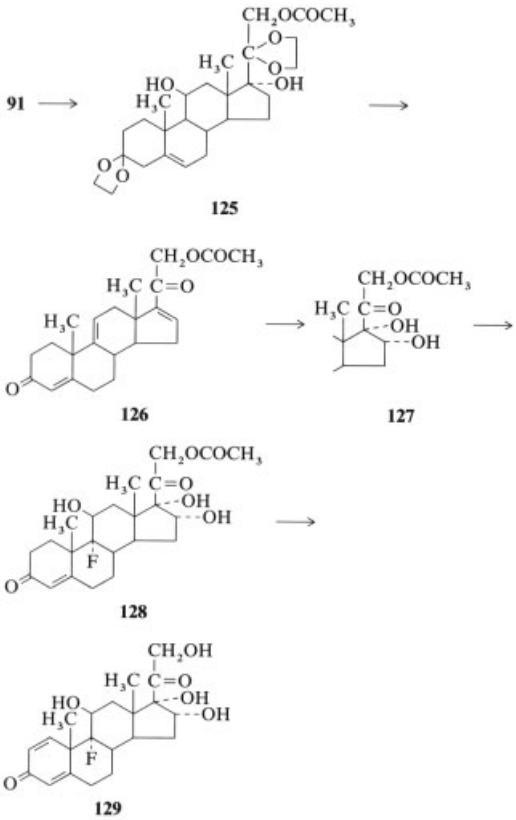

The 16α-hydroxyl group can also be introduced directly into 9α-fluoroprednisolone (**124**) by using a microbiological method originally developed by Squibb [253, 254]. Triamcinolone 16α,17α-acetonide (**130**) is prepared by stirring a suspension of triamcinolone (**129**) in acetone in the presence of catalytic amounts of perchloric acid at approximately 20 °C until the steroid is completely dissolved. Yields of 95 % are obtained [255].

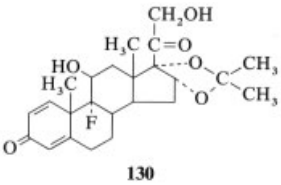

Effects and Uses. Systemic administration of 4 mg of triamcinolone produces the same effect as 5 mg of prednisolone or 25 mg of cortisol. Triamcinolone acetonide has 40 times the local anti-inflammatory activity of cortisol. Triamcinolone and its derivatives cause very low sodium and water retention, exert only a slight appetite-stimulating effect, but occasionally produce myopathy. Administration of these preparations is indicated particularly in patients who have a tendency to edema and who are over-weight.

Triamcinolone and its derivatives are applied systemically (oral, intravenous, or intramuscular), topically (intraarticular as a crystal suspension), and locally (ointment, cream, or lotion).

Trade Names. The following trade names are used:

Triamcinolone (chiefly for systemic application): Delphicort, Aristocort (Lederle); Volon (Heyden); Kenacort (Squibb). Triamcinolone 21-(3,3-dimethylbutyrate), also called triamcinolone hexacetonide (for topical long-term therapy): Aristophan (Lederle). Triamcinolone 16,17-diacetate (for topical therapy): Delphicort (Lederle).

Triamcinolone acetonide (for topical therapy): Volon-A (Heyden); Delphicort (Lederle). Triamcinolone acetonide is used as its 21-disodium phosphate in Aristosol and as its 21-hemisuccinate in Solutedarol for systemic intravenous administration.

4.10.4. 6α-Methylcorticosteroids

Insertion of a 6α-methyl group into the corticosteroid nucleus by scientists at Upjohn [256] and British Drug Houses [257] gave rise to derivatives with enhanced glucocorticoid activity and low mineralocorticoid activity. The most potent orally, parenterally, and topically active compounds are 6α-methylprednisolone (**131**), which has approximately 10 times the anti-inflammatory activity of cortisol in animal experiments, and 6α-methyl-9α-fluoroprednisolone (**132**), which has relatively low mineralocorticoid activity and almost 100 times the glucocorticoid potency of cortisol in animal experiments. 21-Deoxy-6α-methyl-9α-fluoroprednisolone (**133**) is important in the local treatment of inflammatory skin diseases.

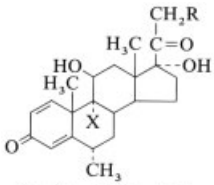

131: X = H, R = OH
132: X = F, R = OH
133: X = F, R = H

6α-Methylprednisolone [*83-43-2*] (**131**), $C_{22}H_{30}O_5$, M_r 374.46, *mp* 228–237 °C, $[\alpha]_D^{20}$ +83° (dioxane), and **21-deoxy-9α-fluoro-6α-methylprednisolone** [*426-13-1*] (**133**), fluorometholone, $C_{22}H_{29}FO_4$, M_r 376.47, *mp* 292–303 °C, are the most important 6α-methylcorticosteroids.

Production. 6α-Methylcorticoids are prepared commercially from 11α-hydroxyprogesterone (**83**), as shown below [256]:

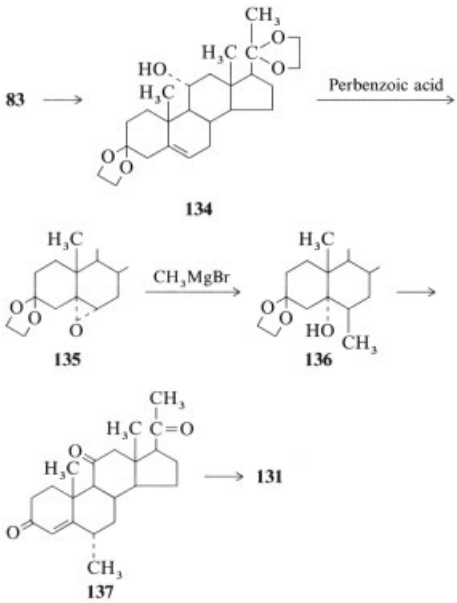

Compound **83** is treated with ethylene glycol to give the 3,20-bis(ethylene ketal). The 6-methyl group is then introduced into the epoxide (**135**) by the Grignard reaction. The 11α-hydroxyl group of compound **136** is subsequently oxidized with chromic acid to give the 11-keto function, and compound **137** is finally obtained after deketalization, dehydration, and isomerization. The 17-acetyl group of compound **137** is then converted to the 17-dihydroxyacetone side chain by using the progesterone method (see Section 4.8.2). The 11-keto group is reduced to an 11β-hydroxyl group, and introduction of the Δ^1 double bond finally leads to 6α-methylprednisolone (**131**).

Another method developed by Merck, Sharp, and Dohme uses cortisone as the starting material. Cortisone is first converted to the 6-keto derivative via suitable intermediates. Subsequent Grignard reaction with methylmagnesium bromide and further reaction steps then yield 6α-methylprednisolone [258].

In the production of 6α-methyl-21-deoxycorticosteroids, the corresponding 11-deoxy-21-hydroxylated compounds are used as starting materials because they can easily be hydroxylated microbiologically at position 11. After 11-hydroxylation, the 21-hydroxyl group is first replaced with an iodine moiety, via the tosylate or mesylate, and then reduced with hydrogen [259]. Preparation of 21-deoxy-6α-methyl-9α-fluoroprednisolone (**133**) is described in [260].

Effects and Uses. 6α-Methylprednisolone (4 mg of which is equivalent to 5 mg of prednisolone) is well tolerated by the stomach, has no mineralocorticoid side effects, and is the corticoid of choice for systemic long-term therapy. It increases appetite slightly and is mildly stimulating psychologically.

6α-Methylprednisolone is administered orally, intravenously (as a water-soluble solution of the 21-sodium hemisuccinate), or topically (as a crystal suspension). Its 21-cyclopentylpropionate derivative is most suitable for intramuscular application and has an extremely long duration of action. Fluorometholone (**133**) is a strong anti-inflammatory substance and is used exclusively in the form of ointments, creams, and lotions in the local treatment of dermatosis.

Trade Names. The following trade names are employed:

6α-Methylprednisolone (for oral application): Urbason (Hoechst); Medrate (Upjohn); Medrol (Upjohn).

6α-Methylprednisolone 21-acetate (for topical application): Urbason crystal suspension (Hoechst); Depo-Medrol (Upjohn).

6α-Methylprednisolone 21-sodium hemisuccinate (for intravenous application): Urbason soluble (Hoechst); Medrate soluble (Upjohn); Solu-Medrol (Upjohn).

6α-Methylprednisolone 21-cyclopentylpropionate (for intramuscular application with long-term effects): Urbason Depot (Hoechst).

21-Deoxy-6α-methyl-9α-fluoroprednisolone (for local application): Delmeson (Hoechst); Oxylone (Upjohn).

4.10.5. 16-Methylcorticosteroids

Introduction of a methyl group at the 16α-position of the corticoid molecule greatly increases glucocorticoid and anti-inflammatory activity, and decreases mineralocorticoid activity [261]. 16α-Methylcorticosteroids include 16α-methylhydrocortisone, 16α-methylprednisolone and 16α-methyl-9α-fluoroprednisolone. The last mentioned, also known as *dexamethasone* (**141**), has very low mineralocorticoid activity and approximately 30 – 40 times the anti-inflammatory potency of hydrocortison. It is, therefore, highly effective when given orally, parenterally, or locally in the treatment of inflammatory, rheumatic, and allergic diseases.

The corresponding 16β-*methylcorticosteroids* also possess high glucocorticoid activity and exert only a slight effect on mineral metabolism. Their anti-inflammatory activity is equal to that of the 16α-methylcorticoids. *Betamethasone* (**68**) [262–264], 16β-methyl-9α-fluoroprednisolone, was developed by the Schering Corporation and, like dexamethasone, is of great importance in oral, parenteral, and local therapy.

Dexamethasone [*50-02-2*] (**141**), 9α-fluoro-16α-methylprednisolone, $C_{22}H_{29}FO_5$, M_r 392.45, *mp* 262 – 264 °C, $[\alpha]_D^2$ +77.5°, and **betamethasone** [*378-44-9*] (**146**), 9α-fluoro-16β-methylprednisolone, $C_{22}H_{29}FO_5$, M_r 392.45, *mp* 231 – 234 °C, $[\alpha]_D$ +108° (acetone), are representative of the 16-methylcorticosteroids.

Production. A number of multistage procedures for the synthesis of *dexamethasone* have been developed, mainly by Merck [261], Schering [265], Roussel-Uclaf [266], and Upjohn [269]. The Schering method uses 16-methylpregna-4,16-dien-3β-ol-20-one (**138**) as the starting material (Fig. 10). This compound is obtained by copper(II)-catalyzed 1,4-addition of methylmagnesium bromide to the Δ^{16}-20-keto system of 16-dehydropregnenolone (**92**) [267].

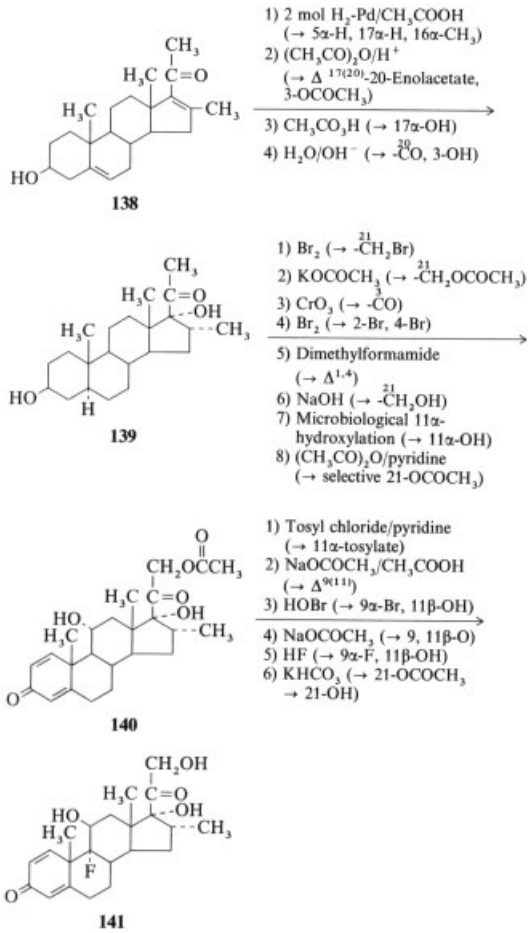

Figure 10. Production of dexamethasone (**141**) from 16-methylpregna-4,16-dien-3β-ol-20-one (**138**)

Stereospecific catalytic hydrogenation of the double bonds at the 3,4- and 16,17-positions of compound **138** yields 16α-methyl-3β-hydroxy-5α-pregnan-20-one. Compound **139** is obtained by the stereospecific introduction of a 17α-hydroxyl group into the $\Delta^{17,20}$-20-enolacetate with peracetic acid. The 21-acetate group is then inserted via the 21-bromide. The next steps involve oxidation of the 3β-hydroxyl group with

chromium trioxide to the 3-keto group, dibromination in the α,α'-position to give the α,α'-dibromo derivative, elimination of the bromine groups to yield the $\Delta^{1,4}$-dien-3-one system, and alkaline saponification of the 21-acetate group. Compound **140** is then obtained microbiologically by introduction of the required oxygen function at the 11α-position.

The 11α-hydroxyl compound is converted to the 9α-fluoro-11β-hydroxyl derivative as described in Section 4.10.2. Alkaline saponification of the 21-acetate group finally produces dexamethasone (**141**).

The Merck [261] method involves the preparation of 16-pregnen-3α-ol-11,20-dione 3-acetate (**142**), as described in [268]. This compound already has the essential oxygen function at position 11. Conversion to dexamethasone (**141**) requires 14 further reaction steps.

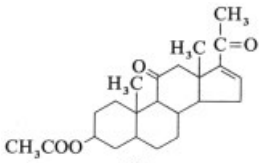

142

Similar concepts have been developed for the industrial production of *betamethasone* by Merck [270], Schering [271], and Roussel-Uclaf [272].

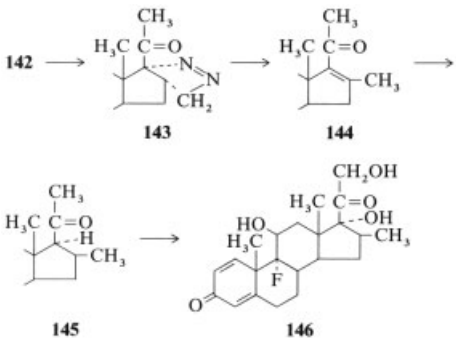

The 16β-methyl group is inserted into the appropriate precursor. The Merck method uses 3α-acetoxy-16-pregnene-11,20-dione (**142**) as the starting material [273]. A 1,3-dipolar addition reaction with diazomethane at the double bond between positions 16 and 17 gives compound **143**. Alkaline cleavage or pyrolysis at ca. 200 °C then leads to the homologous 16-methyl-16-dehydro-20-keto compound (**144**). Stereospecific, catalytic hydrogenation of the Δ^{16} double bond with palladium and calcium carbonate yields compound **145**, which has a β-methyl group at position 16. The other essential groups (the 17α- and 21-hydroxyl groups, and the 9α-fluoro group) and the dienone structure in ring A are then introduced in a series of steps to give betamethasone (**146**).

Effects and Uses. A dose of 1 mg of dexamethasone or betamethasone is equivalent to 5 mg of prednisolone or 20 mg of cortisol.

Of all the corticoids used in medicine, dexamethasone and betamethasone exert the strongest glucocorticoid and anti-inflammatory effects. However, these compounds not only have a high glucocorticoid potency, they also have a number of unpleasant side effects that must be considered, especially in systemic therapy. Both compounds are very psychologically stimulating (euphoria) and increase appetite. Other unfavorable side effects include high catabolic activity, strong hypophyseal inhibition, and considerable calcium and phosphate excretion leading to osteoporosis. For these reasons, both of these highly potent corticoids are more suitable for limited, pharmacodynamic therapy than for long-term treatment. Betamethasone is slightly better tolerated by the stomach than dexamethasone.

Trade Names. The following are some trade names for these two compounds:

Dexamethasone and derivatives (21-acetate, 21-trimethylacetate, 21-disodium phosphate, 21-sodium hemisuccinate, 21-isonicotinate): Decadron (Sharp and Dohme); Fortecortin (Merck); Millicorten (Ciba); Auxiloson (Thomae); Dexa-Scheroson (Schering); Oradexon (Organon); Dexa-Cortidelt (Roussel-Pharma); Dexamethason "Ferring" (Ferring); Dexonil (Schering).

Betamethasone and derivatives (as above) for systemic, topical, and local therapy: Betnesol (Glaxo); Celstone (Schering); Celestan (Byk-Essex) Derivatives for local therapy include betamethasone 17-valerate [Bentnesol V (Glaxo); Celestan V (Byk-Essex) and betamethasone 17,21-dipropionate [Dipcosone (Schering)]. 9α-Chloro-16β-methylprednisolone 17,21-dipropionate [*4419-39-0*], also

known as beclomethasone, is especially suitable for local therapy: Propaderm, production is described in [274,275].

4.10.6. 16-Methylenecorticoids

Animal experiments have shown that insertion of a 16-methylene group into the prednisolone nucleus enhances glucocorticoid activity. 16-Methyleneprednisolone, prednylidene (**147**), has an almost total absence of mineralocorticoid activity. Substitution at the 9α-position with fluorine leads to the local corticoid fluprednylidene [*2193-87-5*] (**148**).

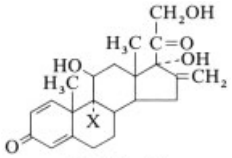

147: X = H
148: X = F

16-Methyleneprednisolone [*599-33-7*] (**147**), prednylidene, $C_{22}H_{28}O_5$, M_r 372.44, *mp* 233–235 °C, $[\alpha]_D^{23}$ +31 ° (dioxane), is produced as described below.

Production. The 16-methylene-17-dihydroxyacetone group of the 16-methylenecorticoids is achieved by using the following reaction sequence [276–278]:

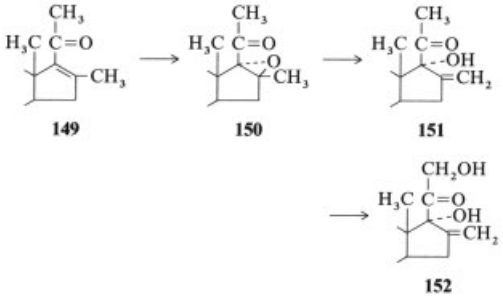

149 **150** **151** **152**

The addition of alkaline hydroperoxide to the double bond at the 16,17-positions of the 16-methyl-16-dehydro-20-keto system (**149**) of a suitable pregnane derivative stereospecifically produces the 16α,17α-oxide (**150**). Treatment with toluene-*p*-sulfonic acid in acetone then leads to opening and isomerization of compound **150**, giving the 17α-hydroxyl and 16-methylene groups (**151**). Insertion of a 21-hydroxyl group by conventional means gives the complete 16-methylene-17-dihydroxyacetone structure (**152**). Both prednylidene (**147**) and fluprednylidene (**148**) are synthesized in this way.

Effects and Uses. A dose of 9 mg of 16-methyleneprednisolone is equivalent to 7.5 mg of prednisolone or 30 mg of cortisol. Prednylidene, which is slightly less potent than prednisolone in humans, is similar to 6α-methylprednisolone in that it is very mildly, psychologically stimulating, only slightly stimulating to the appetite, well tolerated by the stomach, and suitable for long-term therapy. Fluprednylidene is more active when applied locally and is used exclusively in local corticoid therapy because of its strong mineralocorticoid activity.
Trade Names. Prednylidene is used primarily in systemic therapy as such or as the water-soluble 21-diethylaminoacetate hydrochloride: Decortilen (Merck).

Fluprednylidene is used exclusively in local corticoid therapy as the 21-acetate: Decoderm (Merck).

4.10.7. 6α-Fluoro-, 6α-Chloro-, and 6α,9α-Difluorocorticoids

Insertion of fluorine or chlorine into the 6α-position of the steroid skeleton increases glucocorticoid potency and eliminates mineralocorticoid activity [279]. Here again, the 6α-fluoro derivatives have a stronger anti-inflammatory effect than the homologous chlorine compounds.

6α-Fluoroprednisolone (**153**), also called fluprednisolone, has approximately the same glucocorticoid activity as triamcinolone. Insertion of other activating groups into the 6α-fluoroprednisolone molecule (primarily a methyl or a hydroxyl group at position 16 or another fluoro group at position 9α) strikingly enhances glucocorticoid activity. However, the undesirable mineralocorticoid side effects also increase. Insertion of a 16α-methyl group into the fluprednisolone molecule yields systemically active paramethasone (**154**). 6α,9α-difluoro-16α-methylprednisolone (**155**), flumethasone, has not only the 6α-fluoro group, but also the

strongly activating 9α-fluoro and 16α-methyl groups. The 16,17-acetonides (**156**) are also widely used. The following are some therapeutically important compounds:

6α-Fluoroprednisolone [*53-34-9*] (**153**), fluprednisolone, $C_{21}H_{27}FO_5$, M_r 378,45, *mp* 208 – 213 °C

Paramethasone [*53-33-8*] (**154**), $C_{22}H_{29}FO_5$, M_r 392.45, 21-acetate: *mp* 228 – 241 °C

Flumethasone [*2135-17-3*] (**155**), $C_{22}H_{28}F_2O_5$, M_r 410.46

Fluocinolone acetonide [*67-73-2*] (**156**), $C_{24}H_{30}F_2O_6$, M_r 452.50, *mp* 265 – 266 °C, $[\alpha]_D$ +95 ° (chloroform)

6α-Fluoro-16α-hydroxycortisol (**158**), fluandrenolone, $C_{21}H_{29}FO_6$, M_r 396.46, *mp* 234 – 236 °C, $[\alpha]_D$ +95 ° (dioxane)

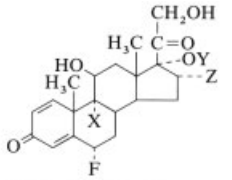

153: X = Y = Z = H
154: X = Y = H; Z = CH_3
155: X = F; Y = H; Z = CH_3
156: X = F; Y and Z = 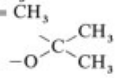

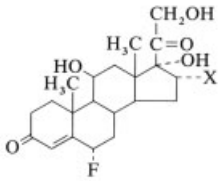

157: X = H
158: X = OH

Production. 6α-*Fluorocortisol* (**157**) is prepared from 3β,17α,21-trihydroxypregn-5-en-20-one 17,21-diacetate [280] with a method developed by Syntex [281]. This compound has no therapeutic importance but typifies insertion of the 6α-fluoro group.

A method developed by Syntex [282] is used to synthesize the therapeutically important compound *paramethasone* (**154**). The double bond at the 5,6-positions of the starting compound, 16α-methylpregnenolone 3-acetate (**159**), is converted stereospecifically with perbenzoic acid to the 5,6α-oxide (**160**). Compound **160** is then cleaved with boron trifluoride to give the 6β-fluoro-5α-hydrin (**161**). After insertion of the 17α- and 21-hydroxyl groups, oxidation with chromium trioxide, and dehydration to the 3-keto-4,5-dehydro system (**162**), the 6β-fluoro group is isomerized to the 6α-fluoro group by treatment with acid. The 11β-hydroxyl group is then inserted enzymatically by incubation with bovine adrenal gland [283], to give a yield of 50 – 60 %. Product **163** is dehydrogenated with selenium dioxide in *tert*-butanol and pyridine to yield paramethasone (**154**).

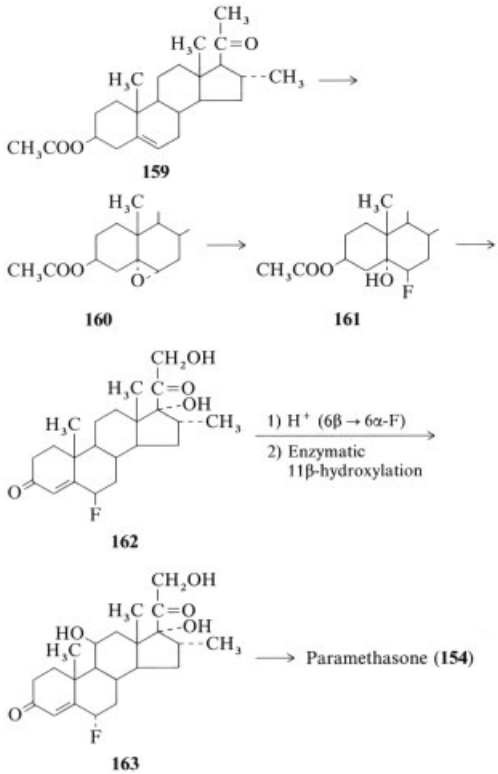

Compound **163** also serves as an intermediate in the commercial synthesis of the local corticoid *flumethasone* (**155**) as follows:

1) dehydration of **163** with $SOCl_2$ and pyridine to give the $\Delta^{9,11}$ derivative;
2) formation of the 9,11β-oxide, as described in Section 4.10.2;

3) cleavage with hydrogen fluoride to give the 9α-fluoro,11β-hydroxy derivative; and
4) dehydrogenation with selenium dioxide to $\Delta^{1,2}$.

Fluprednisolone (**153**) is synthesized with methods developed by Upjohn [284], Syntex [285], and Bayer [286]. *Fluocinolone acetonide* (**156**) is important in local therapy and can be prepared commercially from 16α,17α-oxidopregn-5-ene-3β,21-diol-20-one 21-acetate (made according to [287]) via fluandrenolone (**158**) in a multistage process (more than 20 reaction steps) [288].

In addition to the cleavage of 5,6α-oxy groups with hydrofluoric acid, other methods have been elaborated to insert the 6α-fluoro group into the corticoid nucleus. In a method developed by Syntex, FBr (from hydrofluoric acid and *N*-bromosuccinimide) is added to the Δ^5 double bond of compound **164**, forming the corresponding 5α-bromo-6β-fluoro compound (**165**). Subsequent oxidation of the 3-hydroxyl group with chromium trioxide gives the 3-oxo group, and elimination of hydrogen bromide produces 6β-fluoro-Δ^4-3-keto steroids (**166**). Acid isomerization of the 6β-fluoro group yields 6α-fluoro-Δ^4-3-keto steroids [279, 289].

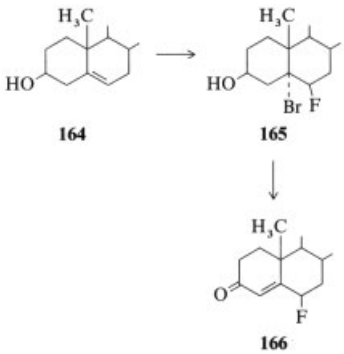

The 6-fluoro group can also be inserted by converting the $\Delta^{3,5}$-3-enolacetates of suitable pregnane precursors with perchloryl fluoride ($FClO_3$) in dioxane as solvent [290].

Similarly, the 6α-chlorine atom can be inserted into the steroid skeleton by reacting the $\Delta^{3,5}$-3-enolethers or Δ^5-3-ethylene dioxysteroids with *N*-chloroacetamide [291].

Effects and Uses. A systemic dose of 2.5 mg of fluoroprednisolone or 2.0 mg of paramethasone is equivalent to 5.0 mg of prednisolone. Both corticoids are used in systemic therapy; they are well tolerated by the stomach and suitable for short- or long-term treatment. However, the relatively strong hypophyseal inhibition exerted by both corticoids must be taken into consideration.

The highly active local corticoids flumethasone and fluocinolone acetonide are widely applied as the free 21-alcohols or their esters in local corticoid therapy. They are not administered systemically.

Trade Names. The following trade names are used:

Fluprednisolone (systemic administration): Alphadrol (Upjohn).

Paramethasone (systemic administration as the 21-acetate or 21-disodium phosphate): Monocortin (Grünenthal); Stemex (Syntex); Haldrone (Lilly).

Flumethasone (applied locally as the 21-pivalate): Locacorten (Ciba and Syntex).

Fluocinolone acetonide (for local application): Jellin (Grünenthal); Synalar (Syntex); Lidex (21-acetate, Syntex).

4.10.8. 17-Deoxycorticosteroids

In 1964 Raspé and coworkers found that 16α-methylcorticosteroids lacking the 17α-hydroxyl group are still relatively strong anti-inflammatory substances with very low mineralocorticoid activity [292]. For instance, 16α-methyl-1-dehydrocorticosterone has approximately the same anti-inflammatory activity as prednisolone when applied locally, subcutaneously, or orally. 6α-Fluoro-16α-methyl-1-dehydrocorticosterone, fluocortolone (**169**) is used in systemic corticoid therapy [293]. The similarly structured 9α-fluoro-16α-methyl-1-dehydrocorticosterone, desoximetasone (**173**), is commercially available as a highly active local corticoid [294].

Fluocortolone [*152-97-6*] (**169**), $C_{22}H_{29}FO_4$, M_r 376.47, *mp* 188 – 190.5 °C, $[\alpha]_D^{20}$ +100° (dioxane), and **17-desoximetasone** (**173**), $C_{22}H_{29}FO_4$, M_r 376.47, *mp* 217 °C,

$[\alpha]_D^{20} + 109°$ (chloroform), are important 17-deoxycorticosteroids.

Production. *Fluocortolone* (**169**) is prepared from 16α-methylpregn-5-ene-3β,21-diol-20-one 21-acetate (**167**) (obtained as described in [295]) in a method developed by Schering [293, 296]. 6α-Fluoro-16α-methyldeoxycorticosterone (**168**) is formed in a reaction sequence analogous to **164** → **166**, with subsequent isomerization. The 11β-hydroxyl group is introduced microbiologically by using *Curvularia lunata* [297], followed by stereospecific dehydrogenation at the 1,2-positions with *Corynebacterium simplex* [298] or *Bacillus lentus* [299] and saponification of the 21-acetate group to yield fluocortolone (**169**).

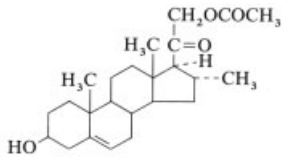

167

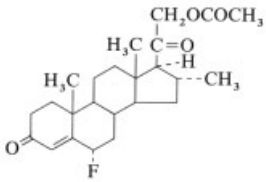

168

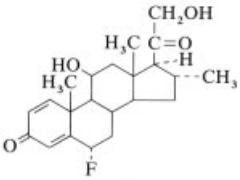

169

Desoximetasone (**173**) is prepared from compounds **170** or **171** according to two different methods developed by Roussel – Uclaf [294, 300]. These processes require six and eight reaction steps, respectively, to produce the important intermediate **172**, which is then converted to desoximetasone (**173**) in four further steps:

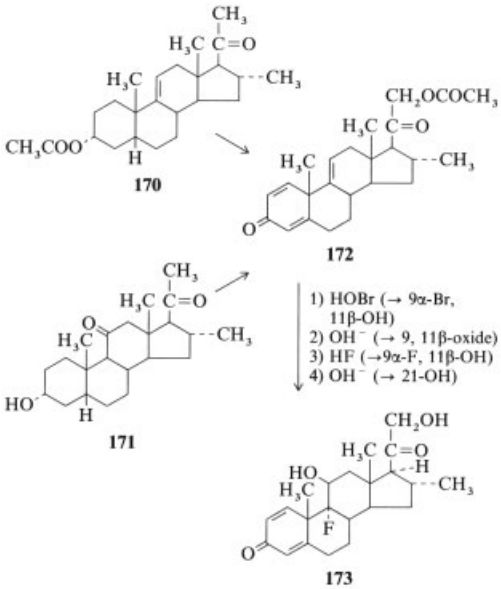

Effects and Uses. A systemic dose of 7.5 mg of fluocortolone is equivalent to 7.5 mg of prednisolone. Although fluocortolone is a corticosterone derivative, it exerts the same systemic glucocorticoid and anti-inflammatory effects in humans as prednisolone. It is well tolerated by the stomach, exhibits only slight catabolic activity, and is thus highly suitable for long-term treatment. In addition, fluocortolone is a powerful local anti-inflammatory substance and, consequently, is used in local corticoid therapy. Because 17-desoximetasone has an unusually high local anti-inflammatory activity, it is an excellent local corticoid.

Trade Names. The following trade names are used:

Fluocortolone (systemic, local, and topical): Ultralan (Schering). It is used locally as such, as the 21-trimethylacetate, as the 21-capronate, or as a mixture of these compounds.

17-Desoximetasone (local): Topisolon (Hoechst); Topicort (Roussel).

Chlorocortolone [*4828-27-7*] (**174**) is another 17-deoxycorticoid that has a chlorine atom at position 9α and a fluorine atom at position 6α. This compound is prepared by a method patented by Schering and is used exclusively in local corticoid therapy [301].

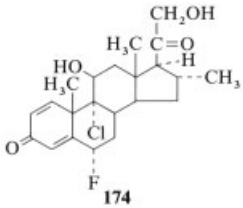

Trade Name. Chlorocortolone is applied locally as the 21-trimethylacetate or 21-capronate: Kaban (Asche, Hamburg).

4.10.9. 9α,11β-Dihalocorticosteroids

Scientists at Schering found that the oxygen at position 11 in corticosteroids can be replaced by a halogen without loss of glucocorticoid or anti-inflammatory activity. Thus 9α,11β-dichloroprednisolone (**175**), also called dichlorisone, not only has a very low mineralocorticoid activity but also has ten times the anti-inflammatory activity of prednisolone [302]. It was launched on the market as the 21-acetate.

Dichlorisone [*7008-26-6*] (**175**), $C_{21}H_{26}Cl_2O_4$, M_r 413.35, has the following properties: *mp* 238 – 241 °C, $[\alpha]_D^{20}$ + 134 ° (pyridine).

Production. The synthesis of dichlorisone involves selective addition of equimolar amounts of the desired halogens to steroids with a double bond at position $\Delta^{9,11}$ [302, 303].

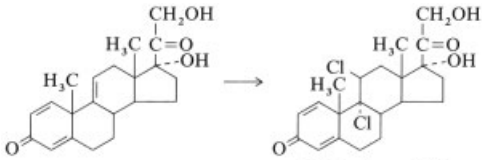

Dichlorisone (**175**)

Effects and Uses. In animal experiments, dichlorisone has ten times the anti-inflammatory potency of prednisolone. The 9α,11β-dichlorocorticoids have not been used in systemic therapy because they exert exceptionally strong (specific) glucocorticoid effects (Cushing's syndrome). However, they also have a strong local and topical anti-inflammatory effect. Dichlorisone is especially useful in local antipruritic therapy.

Trade Names. Dichlorisone is used for topical and local therapy only: Diloderm, Disoderm (Schering).

4.11. Other Developments

Extensive efforts to obtain better separation of systemic and topical effects have produced a series of new preparations with little or no systemic activity. Therefore, these compounds are especially suitable for treating dermatosis in children and large cutaneous areas. These preparations are derivatives of known, active corticoids.

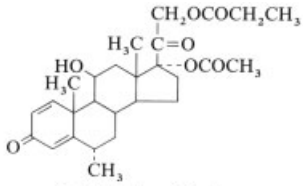

6α-Methylprednisolone aceponate [*83-43-2*] (Schering AG) [4.122]

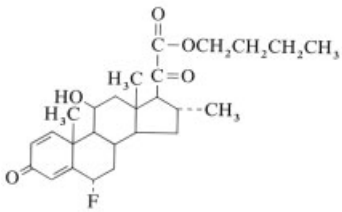

Fluocortin butyl [*41767-29-7*] (Vaspit, Schering AG) [4.119]

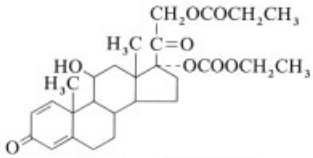

Prednicarbate [*73771-04-7*] (Hoe 777, Hoechst) [4.121]

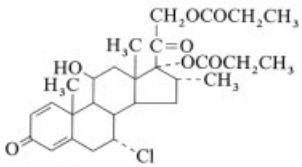

Alclometasone dipropionate [*66734-13-2*] (Delanol, Modrasone, Schering Corp.) [4.123]

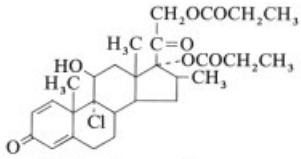

Beclomethasone dipropionate [*5534-09-8*] (Sanasthmyl, Glaxo) [4.118]

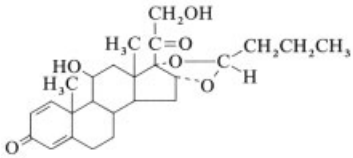

Budesonide [*51333-22-3*] (Pulmicort, Astra) [4.120]

Fundamentally new preparations or replacement of corticoids with another class of substances is not in sight. Improvements are to be found in application protocols (the circadian rhythm is taken into account) and in the wide range of galenical preparations.

5. Cholecalciferol

5.1. Introduction

Cholecalciferol [*67-97-0*] (vitamin D_3) is formed by UV irradiation of 7-dehydrocholesterol in the skin [310]. In 1971, DeLuca and coworkers succeeded in identifying 1α,25-dihydroxyvitamin D_3 [*32222-06-3*] (calcitriol) as the basic active principle formed by consecutive hydroxylation of vitamin D_3 in the liver and kidney [311]; it functions as a regulator of calcium and phosphate transport.

Cholecalciferol and its biologically active metabolite calcitriol must be classified as hormones rather than as vitamins (i.e., substances in food required for normal development). However, certain vitamin D deficiencies can be treated by diets containing vitamin D_3 or D_2.

Historical Aspects. The first description of rickets, a vitamin D deficiency disease, dates to antiquity. During the Industrial Revolution the incidence of rickets increased to epidemic proportions, especially in Europe and North America. The systematic search for an antirachitic factor in food began around 1920, stimulated by the observation that the disease could be cured by either UV irradiation or the intake of fish-liver oil [312]. A major breakthrough was achieved when vitamin D_2 [*50-14-6*] (ergocalciferol) was identified as a potent antirachitic principle that could be obtained by irradiating natural ergosterol [313,314]:

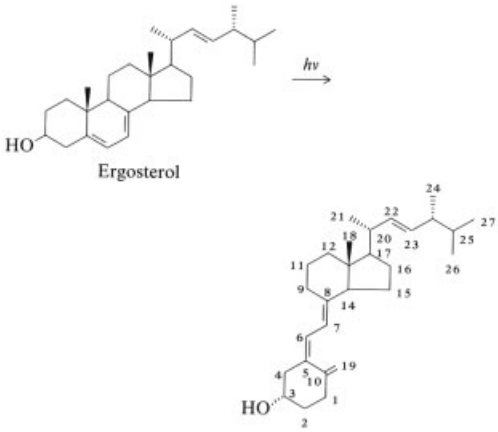

Ergosterol

Vitamin D_2

Although vitamin D_2 occurs naturally (in fish-liver oil), it proved nonidentical with the factor produced by photochemical conversion in the skin of mammals. Windaus and coworkers succeeded in identifying and synthesizing vitamin D_3 as the natural principle formed by photolysis of 7-dehydrocholesterol [315,316].

5.2. Biosynthesis and Metabolism

The metabolism of vitamin D_3 is summarized in Figure 11. In a reaction catalyzed by UV light (wavelength 250 – 300 nm), 7-dehydrocholesterol accumulates in the skin and is transformed into previtamin D_3 which undergoes isomerization to vitamin D_3 by a thermal 1,7-hydrogen shift [317].

Alternatively, dietary vitamin D_3 is taken into the bloodstream by the intestinal route. A specific vitamin D_3 binding protein transports vitamin D_3 to the liver where a cytochrome P450 enzyme hydroxylates the compound at position C-25 to yield 25-hydroxyvitamin D_3. In the kidney, another cytochrome P450 steroid hydroxylase (1α-hydroxylase) introduces a hydroxyl group at position C-1 to form the active

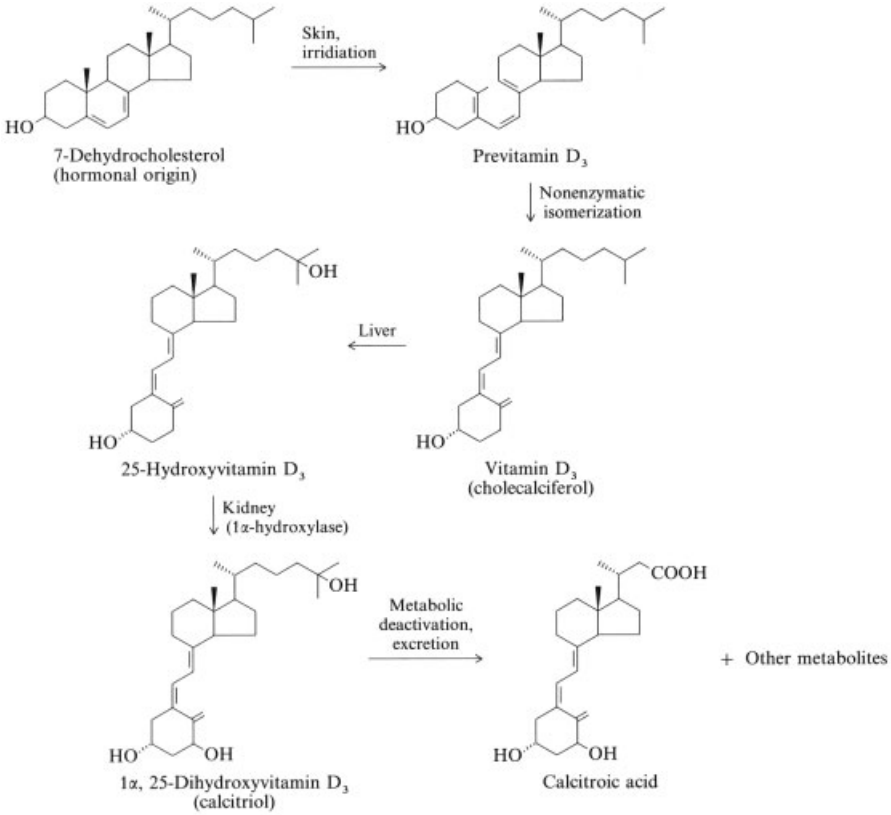

Figure 11. Metabolism of vitamin D_3

metabolite 1α,25-dihydroxyvitamin D_3 (calcitriol). This step appears to be modulated by parathyroid hormone [310, 318]. Calcitriol is bound with high affinity to a specific protein receptor (1α,25-dihydroxyvitamin D_3 receptor, M_r 50 000 – 60 000), which binds to DNA. 1α,25-Dihydroxyvitamin D_3 receptors have been detected in many tissues (intestine, bone, pancreas, thyroid, muscle, gonads).

Calcitriol is inactivated by further enzymatic hydroxylation and oxidation. Approximately 20 metabolic derivatives of calcitriol have been found in human and animal tissue, all of them considerably less active than 1α,25-dihydroxyvitamin D_3.

5.3. Biological Activity

DeLuca and coworkers demonstrated that the biological activity previously ascribed to vitamine D_3 was actually produced by 1α,25-dihydroxyvitamin D_3 (calcitriol).

Calcitriol stimulates bone growth and mineralization by increasing the intestinal absorption of calcium and phosphate, thereby maintaining the necessary serum levels. The discovery of 1α,25-dihydroxyvitamin D_3 receptors in bone indicates a direct action of calcitriol on calcium transport in bone tissue. Bone undergoes constant remodeling; both synthesis and resorption are influenced by calcitriol in a dose-dependent manner. Calcitriol can be used to treat rickets and osteomalacia (softening of the bones), and is indispensable in the treatment of chronic renal failure. Controversy still exists as to whether osteoporosis in postmenopausal women is amenable to calcitriol therapy.

In addition to its effects on mineral homeostasis, 1α,25-dihydroxyvitamin D_3 influences cell proliferation and differentiation in tissues not related to the classical sites of action (intestine, bone, kidney). The irregular growth and maturation of epidermal cells (a symptom of psoriasis) can be inhibited by topical adminis-

tration of calcitriol. Inhibitory effects have been demonstrated for various cancer cell lines, so that calcitriol and its derivatives may find new therapeutic uses [310, 318].

5.4. Chemical Synthesis

Following the pioneering work of WINDAUS and coworkers [314, 315], numerous research groups have been engaged in the synthesis of vitamin D_3, its hydroxylated metabolites, and analogues [319].

Classical photolysis of a steroidal 5,7-diene precursor remains the basis for several recent strategies, although other nonphotochemical concepts are being developed. The principle of convergence (i.e., separate construction of major molecular portions that are combined in a final step to form the structural entity) appears to be a promising approach to an economical, large-scale production of vitamin D_3 and derivatives.

Examples are given for two calcitriol syntheses based on the classical concept [320] and the convergent strategy [321].

Classical Concept. BARTON et al. started their synthesis of calcitriol with 25-hydroxycholesterol (**176**) [320]. Dehydrogenation with dichlorodicyano-*p*-benzoquinone gave the trienone **177**, which was selectively epoxidized to form the $1\alpha,2\alpha$-epoxide **178**. Lithium–ammonia reduction of **178** followed by acetylation yielded the triacetate **179**, which was transformed to the key intermediate **180** by allylic bromination–dehydrobromination. Electrocyclic ring opening by irradiation via the intermediate **181**, thermal isomerization, and saponification resulted in the formation of 1α,25-dihydroxyvitamin D_3 (calcitriol; see right figure).

Convergent Strategy. USKOKOVIC and coworkers achieved a highly stereoselective, total synthesis of calcitriol by combining synthons **182** and **183**, prepared from easily accessible chiral starting materials, in a Wittig–Horner reaction as the final step [321] (see figure on top of next page).

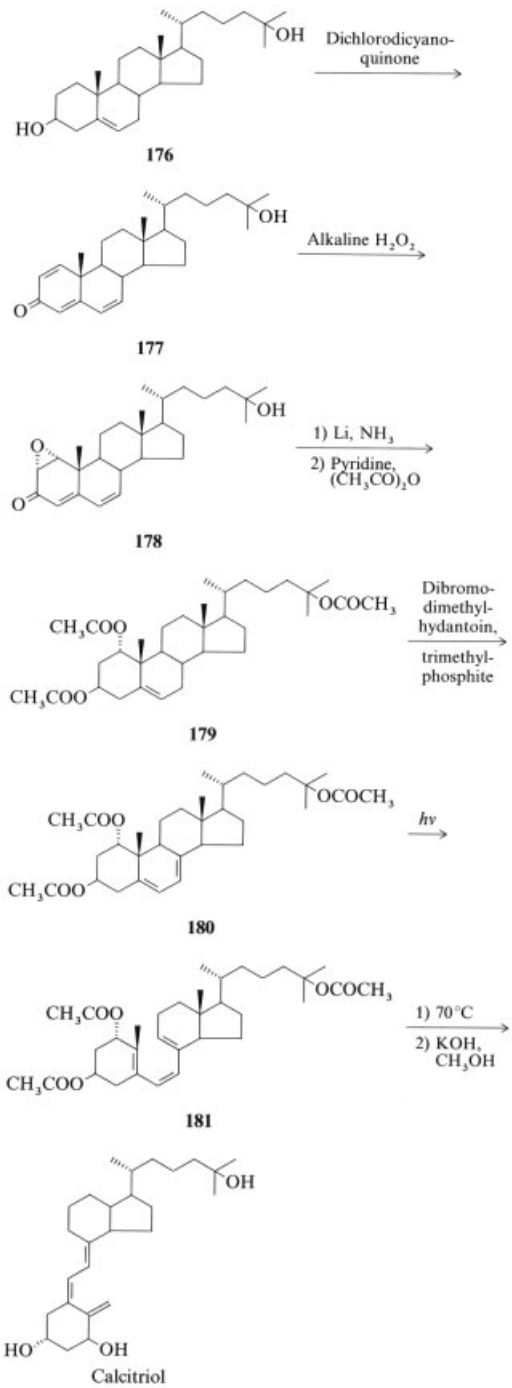

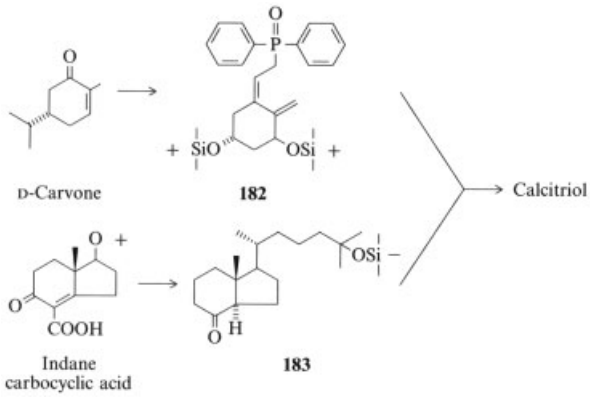

5.5. Synthetic Analogues

Numerous attempts have been made to find derivatives of vitamin D_3 and 1α,25-dihydroxyvitamin D_3 with better separation of the systemic effect on calcium metabolism and the antiproliferative, cell-differentiating activities [322]. Fluorination at position C-24, C-26, or C-27 substantially increases potency, compared with fluorine-deficient compounds. Other chemical modifications have led to derivatives with much lower activity, so that synthetic efforts in this field continue [322].

6. References

General References

1. "Endocrinology," *Handbook of Physiology,* **vol. 1 – 7,**Sect. 7, American Physiological Society, Washington 1974.
2. *Cold Spring Harbor Symposia on Quantitative Biology,* vol. LI/1, LI/2, Cold Spring Harbor Laboratory, Cold Spring Harbor 1986.
3. S. Baba, T. Kaneko, N. Yanaihara: *Proinsulin, Insulin, C-Peptide.* Excerpta Medica, Amsterdam 1979.
4. R. Baserga: *Handbook of Experimental Pharmacology,* "Tissue Growth Factors," vol. 57, Springer Verlag, Berlin 1981.
5. S. R. Bloom: *Gut Hormones,* Churchill Livingstone, Edinburgh 1978.
6. D. Brandenburg, A. Wollmer: *Insulin. Chemistry, Structure and Function of Insulin and Related Hormones,* Walter de Gruyter, Berlin 1980.
7. I. Chester Jones, I. W. Henderson: *General, Comparative and Clinical Endocrinology of the Adrenal Cortex,* **vol. 2,** Academic Press, London1978.
8. L. DeGroot et al.: *Endocrinology,* **vol. 1 – 3,** Grune & Statton, New York 1979.
9. B. J. A. Furr, A. E. Wakeling: *Pharmacology and Clinical Uses of Inhibitors of Hormone Secretion and Action,* Baillière Tindall, Eastbourne 1987.
10. G. B. J. Glass: *Gastrointestinal Hormones,* Raven Press, New York 1980.
11. E. C. Griffiths, G. W. Bennett: *Thyrotropin-Releasing Hormone,* Raven Press, New York 1983.
12. F. Gross, H. G. Vogel: *Enzymatic Release of Vasoactive Peptides,* Raven Press, New York 1980.
13. G. Litwack: *Biochemical Actions of Hormones,* **vol. 13,** Academic Press, London 1986.
14. J. B. Martin, S. Reichlin, G. M. Brown: *Clinical Neuroendocrinology,* F. A. Davis Co., Philadelphia 1977.
15. E. Nieschlag: *Immunization with Hormones in Reproduction Research,* North-Holland/American Elsevier, Amsterdam 1975.
16. J. A. Parsons: *Peptide Hormones,* MacMillan Press, London 1976.
17. D. Rabin, T. J. McKenna: *Clinical Endocrinology, Principles and Practice,* Grune & Stratton, New York 1982.
18. D. Schulster, S. Burnstein, B. A. Cooke: *Molecular Endocrinology of the Steroid Hormones,* John Wiley & Sons, London 1976.
19. M. Tausk: *Pharmacology of the Endocrine System and Related Drugs,* "Progesterone, Progestational Drugs and Antifertility Agents," vol. 1 – 2, Pergamon Press, Oxford 1972.
20. L. Träger: "Biosynthese, Stoffwechsel, Wirkung," *Steroidhormone,* Springer Verlag, Berlin 1977.

21. J. D. Wilson, D. W. Foster: *Textbook of Endocrinology,* 7th Ed., W. B. Saunders Company, Philadelphia 1985.
22. R. H. Unger, L. Orci: "Physiology, Pathophysiology, and Morphology of the Pancreatic A-Cells," *Glucagon,* Elsevier, New York 1981.

Specific References
23. S. B. Baylin, G. Mendelsohn: "Ectopic (inappropriate) hormone production by tumors: Mechanisms involved and the biological and clinical implications," *Endocr. Rev.* **1** (1980) 45 – 77.
24. J. F. Habener: "Genetic control of hormone formation," in [21], pp. 9 – 32.
25. G. F. Grossman: "Gastrointestinal hormones," in [16]pp. 105 – 117.
26. D. H. Coy: "Somatostatins: Pharmacology and potential clinical uses," in [9]pp. 449 – 460.
27. A. L. Goldstein et al.: "Current status of thymosin and other hormones of the thymus gland," *Recent Prog. Horm. Res.* **37** (1981) 369 – 416.
28. S. E. Leeman, N. Aronin, C. Ferris: "Substance P and neurotensin," *Recent Prog. Horm. Res.* **38** (1982) 93 – 132.
29. M. P. L. Caton, in G. P. Ellis, G. B. West (eds.): *Progress in Medicinal Chemistry,* "The prostaglandins," vol. 8, Butterworths, London 1971, pp. 317 – 376.
30. M. P. L. Caton, K. Crowshaw in G. P. Ellis, G. B. West (eds.): *Progress in Medicinal Chemistry,* "Prostaglandins and thromboxanes," vol. 15, Elsevier/North Holland Biomedical Press, Amsterdam 1978, pp. 356 – 423.
31. A. G. E. Pearse: "Evolutionary and developmental relationships among the cells producing peptide hormones," in [16], pp. 33 – 47.
32. D. W. Lincol et al: "Hypothalamic pulse generators," *Recent Prog. Horm. Res.* **41** (1985) 369 –420.
33. F. A. Antoni: "Hypothalamic control of adrenocorticotropin secretion: Advances since the discovery of 41-residue corticotropin-releasing factor," *Endocr. Rev.* **7** (1986) 351 – 378.
34. E. Knobil: "The neuroendocrine control of the menstrual cycle," *Recent Prog. Horm. Res.* **36** (1980) 53 – 88.
35. L. A. Frohman, J.-O. Jansson: "Growth-hormone releasing hormone," *Endocr. Rev.* **7** (1986) 223 – 253.
36. J. Roth, C. Grunfeld: "Mechanism of action of peptide hormones and catecholamines," in [21], pp. 76 – 122.
37. R. V. Farese: "Phosphoinositide metabolism and hormone action," *Endocr. Rev.* **4** (1983) 78 – 96.
38. J. H. Clark, W. T. Schrader, B. W. O'Malley: "Mechanisms of steroid action," in [21], pp 33 – 75.
39. V. J. Hruby: "Structure-activity of the neurohypophyseal hormones and analogs and implications for hormone-receptor interactions," in [16]pp. 192 – 241.
40. B. A. Eipper, R. E. Mains: "Structure and biosynthesis of pro-ACTH/endorphin and related peptides," *Endocr. Rev.* **1** (1980) 1 – 27.
41. D. T. Krieger, A. S. Liotta, M. J. Brownstein, E. A. Zimmermann: "ACTH, β-lipotropin, and related peptides in brain, pituitary, and blood," *Recent Prog. Horm. Res.* **36** (1980) 277 – 335.
42. J. L. Roberts et al.: "Glucocorticoid regulation of proopiomelanocortin gene expression in rodent pituitary," *Recent Prog. Horm. Res.* **38** (1982) 227 – 256.
43. K. Hofmann: "Chemistry of adenohypophysial hormones," vol. 4, The Pituitary Gland and Its Neuroendocrine Control, Part 2 (1974), in [1] pp. 28 – 131.
44. M. Chrétien in S. A. Berson, R. S. Yalow (eds.): "Lipotropins (LPH)," *Methods in Investigative and Diagnostic Endocrinology,* vol. 2 A: Peptide Hormones, North Holland Publ. Co., Amsterdam 1973, pp. 617 – 632.
45. W. H. Bishop, A. Nureddin, R. J. Ryan in J. A. Parsons (ed.): *Peptide Hormones,* "Pituitary luteinising and follicle-stimulating hormones," The Macmillan Press Ltd., London 1976, pp. 273 – 298.
46. S. C. Chappel, A. Ulloa-Aguirre, C. Coutifaris: "Biosynthesis and secretion of follicle-stimulating hormone," *Endocr. Rev.* **4** (1983) 179 – 212.
47. J. C. Fiddes, K. Talmadge: "Structure, expression, and evolution of the genes for the human glycoprotein hormones," *Recent Prog. Horm. Res.* **40** (1984) 43 – 78.
48. B. D. Weintraub et al.: "Glycosilation and posttranslational processing of thyroid-stimulating hormone: Clinical implications," *Recent Prog. Horm. Res.* **41** (1985) 577 – 606.
49. M. J. Karten, J. E. Rivier: "Gonadotropin-releasing hormone analog design. Structure-function studies toward the

development of agonists and antagonists: Rationale and perspectives," *Endocr. Rev.* **7** (1986) 44 – 66.
50. P. Licht et al.: "Evolution of gonadotropin structure and function," *Recent Prog. Horm. Res.* **33** (1977) 169 – 248.
51. A. G. Frantz: "Prolactin, growth hormone and human placental lactogen," in [16] pp. 199 – 231.
52. J. T. Potts, G. D. Auerbach: "Chemistry of the calcitonins," Parathyroid Gland, vol. 7 (Endocrinology, Sect. 7) (1976) in [1], pp. 423 – 430.
53. W. L. Miller, N. L. Eberhardt: "Structure and evolution of the growth hormone gene family," *Endocr. Rev.* **4** (1983) 97 – 130.
54. C. C. Nicoll, G. L. Mayer, S. M. Russell: "Structural features of prolactins and growth hormones that can be related to their biological properties," *Endocr. Rev.* **7** (1986) 169 – 203.
55. N. Ben-Jonathan: "Dopamine: A prolactin-inhibiting hormone," *Endocr. Rev.* **6** (1985) 564 – 589.
56. R. Markstein: "Pharmacology of dopamine receptor agonists," in [9], pp. 461 – 498.
57. E. G. Armstrong, S. Birken, W. R. Moyle, R. E. Canfield (1986): "Immunochemistry of human chorionic gonadotropin," in [13], pp. 91 – 128.
58. P. H. Seeburg: "Human growth hormone: From clone to clinic," in [2], pp. 669 – 678.
59. D. V. Cohn, R. R. MacGregor: "The biosynthesis, intracellular processing, and secretion of parathormone," *Endocr. Rev.* **2** (1981) 1 – 26.
60. S. A. Atlas: "Atrial natriuretic factor: A new hormone of cardiac origin," *Recent Prog. Horm. Res.* **42** (1986) 207 – 242.
61. M. Cantin, J. Genest: "The heart and the atrial natriuretic factor, *Endocr. Rev.* **6** (1985) 107 – 127.
62. J. D. Gardner, R. T. Jensen: "Gastrointestinal peptides: The basis of action at the cellular level," *Recent Prog. Horm. Res.* **39** (1983) 211 – 244.
63. E. Straus, S. W. Ryder, J. Eng, R. Yalow: "Immunochemical studies relating to cholecystokinin in brain and gut," *Recent Prog. Horm. Res.* **37** (1981) 447 – 476.
64. M. Ellis, K. Sikora: "Genes and cancer: A clinical perspective," *J. R. Coll. Physicians London* **21** (1987) 122 – 128.
65. R. B. Dickson, M. E. Lippman: "Estrogenic regulation of growth and polypeptide growth factor secretion in human breast carcinoma," *Endocr. Rev.* **8** (1987) 29 – 43.
66. D. Gospodarowicz, G. Neufeld, L. Schweigerer: "Fibroblast growth factor," *Mol. Cell. Endocrinol.* **46** (1986) 187 – 204.
67. M. Jaye et al.: "Human endothelial cell growth factor: Cloning, nucleotide sequence, and chromosome localization," *Science* **233** (1986) 541 – 545.
68. M. B. Sporn, A. B. Roberts, L. M. Wakefield, R. K. Assolan: "Transforming growth factor-β: Biological function and chemical structure," *Science* **233** (1986) 532 – 534.
69. J. K. Browne et al. (1986): "Erythropoietin: Gene cloning, protein structure, and biological properties," in [2], vol. LI/1, pp. 693 – 702.
70. G. D. Bryant-Greenwood: "Relaxin as a new hormone," *Endocr. Rev.* **3** (1982) 62 – 90.
71. C. Schwabe et al.: "Relaxin," *Recent Prog. Horm. Res.* **34** (1978) 123 – 212.
72. R. L. Cate et al. (1986): "Development of Mullerian inhibiting substance as an anti-cancer drug," in [2], vol. LI/1, pp. 641 – 648.
73. A. J. Mason et al., "Complementary DNA sequences of ovarian follicular fluid inhibin show precursor structure and homology with transforming growth factor-β," *Nature (London)* **318** (1985) 659 – 663.
74. W. Vale et al.: "Purification and characterization of an FSH releasing protein from porcine ovarian follicular fluid," *Nature (London)* **321** (1986) 776 – 779.
75. F. Murad, R. C. Haynes in A. Goodman Gilman, L. S. Goodman, A. Gilman (eds.): "Hormones and hormone antagonists," *The Pharmacological Basis of Therapeutics,* 6th ed., **Sect. 15,** Macmillan Publ. Co., New York 1980, pp. 1367 – 1550.
76. D. P. Cardinali: "Melatonin. A mammalian pineal hormone," *Endocr. Rev.* **2** (1981) 327 – 346.
77. J. P. Preslock: The pineal gland: Basic implications and clinical correlations," *Endocr. Rev.* **5** (1984) 282 – 307.
78. R. J. Reiter: "The pineal and its hormones in the control of reproduction in mammals," *Endocr. Rev.* **1** (1980) 109 – 131.
79. C. S. Pittman, J. A. Pittman: "Relation of chemical structure to the action and metabolism of thyroactive substances," Thyroid, vol. 3 (Endocrinology, Sect. 7) 1974, in [1]pp. 233 – 253.
80. A. W. Norman, J. Roth, L. Orci: "The vitamin D endocrine system: Steroid metabolism, hormone receptors, and biological response

(calcium binding proteins)," *Endocr. Rev.* **3** (1982) 331 – 366.
81. T. J. Cicero: "Basic endocrine pharmacology of opioid agonists and antagonists," in [9], pp. 518 – 537.

General References
82. K. Oberdisse, E. Klein, D. Reinwein: *Die Krankheiten der Schilddrüse,* 2nd ed., Thieme Verlag, Stuttgart 1980.
83. S. H. Ingbar, L. E. Braverman: *Werner's The Thyroid,* 5th ed., Lippincott, Philadelphia 1986.
84. P. Pfannenstiel: *Schilddrüsenkrankheiten – Diagnose und Therapie,* Henning GmbH, Berlin 1985.
85. A. G. Gilman, L. S. Goodman, T. W. Rall, F. Murad (eds.): *The Pharmacological Basis of Therapeutics,* 7th ed., Macmillan, New York 1985, p. 1397.
86. R. O. Greep, E. B. Astwood, M. A. Greer, D. H. Solomon (eds.): Thyroid; *Handbook of Physiology,* **vol. 3,** Section 7, Endocrinology,Am. Physiol. Soc., Washington 1974.

Specific References
87. I. J. Chopra, D. H. Solomon, G. N. Chua Teco, *J. Clin. Endocrinol. Metab.* **36** (1973) 1050.
88. I. J. Chopra, *J. Clin. Invest.* **54** (1974) 583 – 592.
89. M. T. Hays, *J. Clin. Endocrinol. Metab.* **28** (1968) 749.
90. M. I. Surks et al., *J. Clin. Invest.* **52** (1973) 805.
91. K.-W. Wenzel, H. Kirschsieper, *Metabolism.* **26** (1977) 1.
92. C. M. Greenberg et al., *Am. J. Physiol.* **205** (1963) 821.
93. P. Lind, O. Eber, *Acta Med. Austriaca* **13** (1985) no. 1, 13.
94. British Pharmacopeia 1973.
95. US Pharmacopeia 18 (1970).
96. J. S. Varcoe, W. K. Warburton, *J. Chem. Soc.* 1960, 2711.
97. B. Block Jr., G. Powell, *J. Am. Chem. Soc.* **65** (1943) 1430.
98. *Biochem. J.* **21** (1949) 3424.
99. *J. Chem. Soc.,* 1949, 3424.
100. R. I. Meltzer, R. J. Stanaback, *J. Org. Chem.* **26** (1961) 1977.
101. S. Lissitzky, C. Cheftel, *C. R. Hebd. Séances Acad. Sci.* **256** (1963) 3898.
102. H. Nahm, W. Siedel, *Chem. Ber.* **96** (1963) 1.
103. Hoechst, DE 1 077 673, 1958 (W. Siedel, H. Nahm, J. König).
104. G. Hillmann, *Z. Naturforsch.* **11 b** (1956) 319.
105. G. Hillmann, DE 1 065 855, 1956. G. Hillmann DE 1 064 529, 1956.
106. Hoffmann-La Roche, US 2 895 927, 1957.
107. Sandoz, DE 1 221 646, 1963.
108. R. Wegner, K. Rudnick, DL 65 933, 1968.
109. H. Neudecker, E. Scheiffele, *Arzneim. Forsch.* **21** (1971) 432.
110. A. D. Williams, D. E. Freeman, W. H. Florsheim, *J. Chromatogr.* **45** (1969) 371.
111. A. Dibbo et al., *J. Chem. Soc.* 1961, 2645.
112. Chemie Grünenthal, DE-OS 1 493 533, 1964; 1 793 776, 1965.
113. T. Shiba, H. J. Cahnmann, *J. Org. Chem.* **27** (1962) 1773. P. Block Jr., *J. Med. Chem,.* **19** (1976) no. 8, 1067.
114. E. Wachholz, S. Pfeifer, *Pharmazie* **24** (1969) 459.
115. C. A. Johnson et al., *Analyst London* **92** (1967) 328.
116. R. H. Osborn, T. H. Simpson, *J. Chromatogr.* **40** (1969) 219.
117. B. M. R. Heinl, H. M. Ortner, H. Spitzy, *J. Chromatogr.* **60** (1971) 51.
118. E. Wachholz, S. Pfeifer, *Pharmazie* **27** (1972) 43, 97.
119. R. F. Furchgott: "The Classification of Adrenoceptors (Adrenergic Receptors). An Evaluation from the Standpoint of Receptor Theory," in H. Blaschko, E. Muscholl (eds.): *Catecholamines. Handbuch der Experimentellen Pharmakologie,* vol. 33, Springer Verlag, Berlin 1972, pp. 283 – 335.
120. J. Himms – Hagen: "Sympathic Regulation of Metabolism," *Pharmacol. Rev.* **19** (1967) 367 – 461.
121. A. Burger: "Adrenergic Hormones and Drugs," in *Medicinal Chemistry, Part II,* Chap. 46, 3rd. ed., J. Wiley & Sons, New York 1970.
122. D. Palm, D. Hellenbrecht, K. Quiring: "Pharmakologie des noradrenergen und adrenergen Systems Katecholamine, Andrenozeptor-Agonisten und -Antagonisten Antisympathotonika und andere Antihypertensiva," in W. Forth, D. Henschler, W. Rummel (eds.): *Allgemeine und spezielle Pharmakologie und Toxikologie,* 5th ed., BI Wissenschaftsverlag, Mannheim 1987, pp. 124 – 168.
123. A. Weiner: "Norepinephrine, Epinephrine, and the Sympathomimetic Drugs," in A. Goodman Gilman, L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th

ed., Macmillan Publ. Co., New York 1980, pp. 138 – 175.
124. A. Weiner: "Drugs that Inhibit Adrenergic Nerves and Block Adrenergic Receptors," in [121], p. 176 – 210.
125. J. H. Burn, M. J. Rand: "The Action of Sympathomimetic Amines in Animals Treated with Reserpine," *J. Physiol. (London)* **144** (1958) 314 – 336.
126. L. I. Goldberg: "Cardiovascular and Renal Actions of Dopamine: Potential Clinical Applications," *Pharmacol. Rev.* **24** (1972) 1 – 29.

General References
127. L. Fieser, M. Fieser: *Steroids,* Reinhold Publishing Corp. New York 1959; *Steroide,* Verlag Chemie, Weinheim 1961.
128. IUPAC-IUB 1967: "Tentative Rules for Steroid Nomenclature," *Biochim. Biophys. Acta* **164** (1968) 453 – 486.
129. J. Fried, J. A. Edwards (eds.): *Organic Reactions in Steroid Chemistry,* vol. 1 and 2, Van Nostrand Reinhold Comp., New York, Cincinatti, Toronto, Melbourne 1972.
130. W. Forth, D. Hentschler, W. Rummel (eds.): *Allgemeine und Spezielle Pharmakologie und Toxikologie,* 5th ed., Wissenschaftsverlag, Mannheim, Wien, Zürich 1987, pp. 397 – 435.
131. Goodman and Gilman's: *The Pharmacological Basis of Therapeutics,* 7th ed., Macmillan Publ. Co., New York 1985, pp. 1360 – 1459.
132. B. Green, R. E. Leake: *Steroid Hormones, a practical approach,* IRL Press, Oxford, Washington DC 1987.
133. H. L. J. Makin (ed.): *Biochemistry of Steroid Hormones,* 2nd ed., Blackwell Scientific Publications, Oxford, London, Edinburgh, Boston, Palo Alto, Melbourne 1984.
134. R. T. Blickenstaff, A. C. Ghosh, G. C. Wolf: *Total Synthesis of Steroids,* Academic Press, New York, London 1974.
135. A. A. Akhrem, Y. A. Titov: *Total Steroid Synthesis,* Plenum Press, New York, London 1970.
136. M. B. Groen, F. J. Zeelen: "Steroid Total Synthesis," *Recl. J. R. Neth. Chem. Soc.* **105** (1986) 465 – 487.
137. Bundesverband der Pharmazeutischen Industrie e. V.: Rote Liste 1987, Editio Cantor, Aulendorf (Württ.).
138. E. R. Barnhart (publisher): *Physicians' Desk Reference,* 4th ed., Medical Economies Company Inc., Oradell (New Jersey) 1987.
139. M. K. Agarwal (ed.): *Adrenal Steroid Antagonism,* Walter de Gruyter, Berlin, New York 1984.
140. L. Martini, M. Motto (eds.): *Androgens and Antiandrogens,* Raven Press, New York 1977.

Specific References
141. Kon. Nederland, Gist-en Spiritusfabr., Delft,DE-AS 1 568 932, 1966.
142. Kon. Nederland. Gist-en Spiritusfabr., DE-OS 1 768 215, 1968.
143. Noda Inst. for Scient. Res., DE-AS 1 543 269, 1965 (K. Arima, G. Tamura, M. Bae, M. Nagasawa).
144. Richter Gedeon Vegyeszeti Gyar Rt., DE-AS 1 593 327, 1966 (G. Wix, K. Büki, E. Tömörkeny, G. Ambrus).
145. N. Applezweig: *Steroid Drugs,* McGraw-Hill, New York 1962.
146. K. S. Markley: *Soybean and Soybean Products,* **vol. 2,** Interscience Publ., New York 1951, p. 837.
147. N. Applezweig, *Chem. Week* **115** (1974) no. 2, 31.
148. C. Meystre, K. Miescher, *Helv. Chim. Acta* **32** (1949) 1758, 1764. P. Wieland, K. Miescher, ibid, 1922.
149. Parke Davies, US 2 335 616, 1941 (F. H. Tendick, E. J. Lawson); G. Rosenkranz, O. Mancera, F. Sondheimer, C. Djerassi, *J. Org. Chem.* **21** (1956) 520.
150. J. Schmidt-Thomé, *Chem. Ber.* **88** (1955) 895.
151. G. Ehrhard, H. Ruschig, W. Aumüller, *Angew. Chem.* **52** (1939) 363.
152. H. H. Inhoffen, *Naturwissenschaften* **25** (1937) 125. E. B. Hersberg, M. Rubin, E. Schwenk, *J. Org. Chem.* **15** (1950) 292. Schering Corp., US 2 361 847, 1941 (H. H. Inhoffen).
153. Schering AG, DE-AS 1 135 899, 1960 (G. Raspé, K. Kieslich, E. Olivar, R. Müller, B. Wagner). Eli Lilly, US 3 128 238, 1962 (G. E. Mallet). Schering Corp., US 3 134 718, 1963 (A. Nobile).
154. A. J. Birch, H. Smith, *Q. Rev. Chem. Soc.* **12** (1958) 17.
155. A. L. Wilds, N. A. Nelson, *J. Am. Chem. Soc.* **75** (1953) 5366. C. Djerassi, L. Miramontes, G. Rosenkranz, F. Sondheimer, *ibid* **76** (1954) 4092.
156. D. Onken, D. Heublein, *Pharmazie* **25** (1970) 3.
157. F. B. Colton, L. N. Nysted, B. Riegel, A. L. Raymond, *J. Am. Chem. Soc.* **79** (1957) 1123. Searle, US 2 655 518, 1952 (F. B. Coltan).

158. T. B. Windholz et al., *Angew. Chem.* **76** (1964) 249. H. Ueberwasser et al., *Helv. Chim. Acta* **46** (1963) 344. A. Popper et al., *Arzneim. Forsch.* **19** (1969) 352.
159. R. Müller, K. Kieslich: "Technologie der Darstellung organischer Substanzen mit Mikroorganismen," *Chem. Ing. Tech.* **38** (1966) 813. K. Kieslich, *Synthesis* 1969, 120, 147.
160. H. Smith et al., *J. Chem. Soc.* 1964, 4472.
161. C. Rufer, H. Kosmol, E. Schröder, K. Kieslich, H. Gibian, *Justus Liebigs Ann. Chem.* **702** (1967) 141.
162. L. Velluz et al., *C. R. Hebd. Seances Acad. Sci.* **257** (1963) 3086.
163. P. Bellet, G. Nominé, J. Matthieu, *C. R. Hebd. Seances Acad. Sci. Ser. C* **263** (1966) 88.
164. J. Ufer: *Hormontherapie in der Frauenheilkunde,* 4th ed., De Gruyter, Berlin 1972.
165. L. Svennerholm, *Biochem. Biophys. Acta* **24** (1957) 604.
166. A. Ercoli, E. Gardi, *Chem. Ind. (London)* 1961, 1037. Warner-Lambert Pharmac. Corp., DE-AS 1 157 610, 1961 (A. Ercoli).
167. H. L. Krüskemper: *Anabole Steroide,* 2nd ed., Thieme Verlag, Stuttgart 1965.
168. L. G. Hershberger, E. G. Shipley, R. K. Meyer, *Proc. Soc. Exp. Biol. Med.* **83** (1963) 175.
169. Upjohn, US 3 029 263, 1959 (A. Campbell, J. C. Babcock). Ciba, De-AS 1 037 447, 1955 (A. Wettstein, G. Anner, Ch. Mestre).
170. K. R. Bharucha, *Experientia* **14** (1958) 5. L. Ruzicka, M. W. Goldberg, H. R. Rosenberg, *Helv. Chim. Acta* **18** (1935) 1487.
171. Syntex, DE-AS 1 159 943, 1957 (H. W. Ringold, G. Rosenkranz). R. O. Clinton et al., *J. Am. Chem. Soc.* **83** (1961) 1478.
172. F. Neumann, R. Wiechert, *Arzneim. Forsch.* **15** (1965) 1168.
173. H. Smith, *Experientia* **19** (1963) 394.
174. Soc. Farmaceutici Italia, DE-AS 1 146 491, 1960 (B. Camerino, B, Patelli).
175. A. Butenandt, K. Tscherning, G. Hanisch, *Ber. Dtsch. Chem. Ges.* **68** (1935) 2097.
176. J. Haller: *Ovulationshemmung durch Hormone,* 2nd ed., Thieme Verlag, Stuttgart 1968.
177. R. Deghenghi et al., *Tetrahedron* **19** (1963) 289; *J. Med. Chem.* **6** (1963) 301.
178. A. Popper et al., *Arzneim. Forsch.* **19** (1969) 352.
179. Olin Mathieson, DE-AS 1 125 423, 1959 (J. Fried).
180. Syntex, GB 748 824, 1952.
181. R. Wiechert, H. Steinbeck, W. Elger, F. Neumann, *Arzneim. Forsch.* **17** (1967) 1103.
182. A. J. van den Broek, C. van Bokhoven, P. M. J. Hobbelein, J. Lemhuis, *Recl. Trav. Chim. Pays Bas* **94** (1975) 35.
183. H. Hofmeister et al., *Arzneim. Forsch.* **36** (1986) 781.
184. Ortho Pharmaceut, DE-AS 1 620 102, 1976 (A. P. Schroff).
185. E.-E. Baulieu, S. J. Segal: *The Antiprogestin Steroid RU 486 and Human Fertility Control,* Plenum Press, New York, London 1985.
186. K. Hiller, *Drugs of the Future* **7** (1982) 661; **ibid 9** (1984) 714; **ibid 10** (1985) 793.

General References

187. G. Erhart, H. Ruschig (eds.): *Arzneimittel,* 2nd ed., Verlag Chemie, Weiheim 1972. W. Forth, D. Hentschler, W. Rummel (eds.): *Allgemeine und Spezielle Pharmakologie und Toxikologie,* 5th ed., Wissenschaftsverlag, Mannheim, Wien, Zürich 1987. Goodman and Gilman's: *The Pharmacological Basis of Therapeutics,* 7th ed., Macmillan Publishing Company, New York 1985. L. Fieser, M. Fieser: *Steroids,* Reinhold Publishing Corp., New York 1959; *Steroide,* Verlag Chemie, Weinheim 1961. W. Stalmans, M. Laloux in J. D. Baxter, G. G. Rousseau (eds.): *Glucocorticoid Hormone Action,* Springer Verlag, Berlin 1979, pp. 517 – 533.
188. N. Applezweig, *Chem. Week* **115** (1974) no. 2, 31.
189. Z. Buděssinský, M. Protiva: "Corticoide,"*Synthetische Arzneimittel,* Chapt. 19, Akademie-Verlag, Berlin, GDR, 1961.
190. A. Burger: "Steroids, I. The Adrenal Cortex Hormones," *Medical Chemistry,* 3rd. ed., Chapt. 34, J. Wiley, New York, London, Sydney, Toronto 1970.
191. L. F. Fieser, M. Fieser: *Steroids,* Reinhold Publishing Corp., New York 1959; *Steroide,* Verlag Chemie, Weinheim 1961.
192. B. Helwig: *Moderne Arzneimittel,* "Nebennierenrindenhormone," Chapt. 5 A, 4th ed., Wissenschaftl. Verlags GmbH, Stuttgart 1972.
193. H. Kaiser: *Cortisonderivate in Klinik und Praxis,* 5th ed., Thieme Verlag, Stuttgart 1968.
194. R. T. Curnow, J. Larner in J. Litwack (ed.): *Biochemical Actions of Hormones,* Academic Press, New York 1979, pp. 77 – 119. N. R. Slaunwhite, A. A. Sandberg in E. Diezfalusy

(ed.): *Steroid Assay by Protein Binding,* Karolinska Symposium on Research Methods in Reproductive Endocrinology, Stockholm 1970, pp. 144 – 154 (Acta Endocrinologica, Suppl. 147).
195. H. Lettré, H. H. Inhoffen, R. Tschesche: *über Sterine, Gallensäuren und verwandte Naturstoffe,* 2nd ed., Enke Verlag, Stuttgart 1959.

Specific References
196. *Ullmann,* 3rd ed., **8,** 613 ff.
197. R. Schüppel, *Pharm. Unserer Zeit* **2** (1973) 161.
198. C. J. Brooks, J. N. Norymberski, *Biochem. J.* **55** (1953) 371.
199. C. C. Porter, R. M. Silber, *J. Biol. Chem.* **185** (1952) 201.
200. R. B. Woodward, F. Sondheimer, D. Taub, *J. Am. Chem. Soc.* **73** (1951) 3547.
201. L. H. Sarett et al., *J. Am. Chem. Soc.* **74** (1952) 4974.
202. N. L. Wendler, R. Richard, P. Traber, R. E. Jones et al., *J. Am. Chem. Soc.* **72** (1950) 5793.
203. L. Velluz et al., *Angew. Chem.* **72** (1960) 725; **77** (1965) 185.
204. R. Bucourt, G. Nominé, *Bull. Soc. Chim. Fr.* 1966, 1537.
205. J. C. Gase, L. Nédélec, *Tetrahedron Lett.* 1971, 2005.
206. D. H. Peterson et al., *J. Am. Chem. Soc.* **74** (1952) 5933.
207. Upjohn, US 2 602 769, 1952. (H. C. Murray, D. H. Peterson).
208. J. A. Hogg et al., *J. Am. Chem. Soc.* **77** (1955) 4436.
209. H. Ruschig, *Angew. Chem.* **60** (1948) 247; *Chem. Ber.* **88** (1955) 878.
210. Upjohn, US 2 769 823, 1954 (W. P. Schneider, A. R. Hanze).
211. P. L. Julian, E. W. Meyer, W. J. Karpel, J. R. Waller, *J. Am. Chem. Soc.* **72** (1950) 5145.
212. H. J. Ringold, B. Löken, G. Rosenkranz, F. Sondheimer, *J. Am. Chem. Soc.* **78** (1956) 816.
213. J. Romo, G. Rosenkranz, F. Sondheimer, *J. Am. Chem. Soc.* **79** (1957) 5034.
214. F. R. Hanson et al., *J. Am. Chem. Soc.* **75** (1953) 5369.
215. J. Fried et al., *Recent Prog. Horm. Res.* **11** (1955) 149.
216. J. Schmidt-Thomé, *Angew. Chem.* **69** (1957) 238.
217. Hoechst, DE 1 009 627, 1955 (F. Lindner et al.).
218. Upjohn, US 2 602 769, 1952 (H. C. Murray, D. H. Peterson).
219. D. H. Peterson, *Research (London)* **6** (1953) 309/319.
220. L. H. Sarett, *J. Biol. Chem.* **162** (1946) 601.
221. B. F. McKenzie, V. R. Mattox, L. L. Engel, E. C. Kendall, *J. Biol. Chem.* **173** (1948) 271.
222. G. Rosenkranz, F. Sondheimer, *Fortschr. Chem. Org. Naturst.* **10** (1953) 274.
223. C. Djerassi, H. J. Ringock, G. Rosenkranz, *J. Am. Chem. Soc.* **73** (1951) 5513.
224. J. H. Chapman, J. Elks, L. J. Wyman, *Chem. Ind. (London)* 1955, 603.
225. J. A. Hogg et al., *J. Am. Chem. Soc.* **77** (1955) 4436.
226. W. S. Allen, S. Bernstein, R. Littell, *J. Am. Chem. Soc.* **76** (1954) 6116. Amer. Cyanamid, US 2 622 081, 1951, 2 666 069, 1951 (S. Bernstein, R. M. Antonucci, M. D. Heller); . 2 700 666, 1953 (S. Bernstein, R. Littel).
227. R. H. Levin et al., *J. Am. Chem. Soc.* **75** (1953) 502.
228. L. H. Sarett, *J. Am. Chem. Soc.* **70** (1948) 1454.
229. F. R. Hanson et al., *J. Am. Chem. Soc.* **75** (1953) 5369.
230. G. M. Shull, D. A. Kita, *J. Am. Chem. Soc.* **77** (1955) 763.
231. A. Wettstein, *Experientia* **11** (1955) 465.
232. E. Oliveto et al., *J. Am. Chem. Soc.* **78** (1956) 1736.
233. C. Meystre, E. Visher, A. Wettstein, *Helv. Chim. Acta* **37** (1954) 1548.
234. H. J. Ringold, G. Storck, *J. Am. Chem. Soc.* **80** (1958) 250.
235. D. H. R. Barton et al., *J. Am. Chem. Soc.* **82** (1960) 2640, 2641.
236. K. Heusler et al., *Helv. Chim. Acta* **44** (1961) 502; **45** (1962) 1317.
237. L. Velluz, *C. R. Hebd. Seances Acad. Sci.* **250** (1960) 725.
238. L. Ruzicka, K. Hofmann, *Helv. Chim. Acta* **20** (1937) 1280.
239. H. H. Inhoffen, W. Logemann, W. Hohlweg, A. Serini, *Ber. Dtsch. Chem. Ges.* **71** (1938) 1024.
240. J. A. Cella et al., *J. Org. Chem.* **24** (1959) 743, 1109. Searle, US 3 013 012, 1960 (J. A. Cella, R. C. Tweit).
241. H. L. Herzog et al., *Science (Washington D.C.)* **121** (1955) 176.
242. H. L. Herzog et al., *J. Am. Chem. Soc.* **77** (1955) 4781.
243. S. A. Szpilfogel et al., *Rec. Trav. Chim. Pays Bas* **75** (1956) 475.

244. C. Meystre, H. Frey, W. Voser, A. Wettstein, *Helv. Chim. Acta* **39** (1956) 734.
245. E. Vischer, C. Meystre, A. Wettstein, *Helv. Chim. Acta* **38** (1955) 835, 1502.
246. A. Nobile et al., *J. Am. Chem. Soc.* **77** (1955) 4184.
247. F. Lindner et al., *Naturwissenschaften* **43** (1956) 39.
248. J. H. Fried, E. F. Sabo, *J. Am. Chem. Soc.* **75** (1953) 2273; **76** (1954) 1455.
249. S. Bernstein et al., *J. Am. Chem. Soc.* **78** (1956) 5693.
250. J. Fried, W. B. Kessler, P. Grabowich, E. F. Sabo, *J. Am. Chem. Soc.* **80** (1958) 2338.
251. S. Bernstein et al., *J. Am. Chem. Soc.* **81** (1959) 1689.
252. S. Bernstein et al., *J. Am. Chem. Soc.* **77** (1955) 1028; **78** (1956) 5693; **81** (1959) 1689.
253. R. W. Thoma, J. Fried, S. Bonanno, P. Grabowich, *J. Am. Chem. Soc.* **79** (1957) 4818.
254. L. L. Smith, H. Mendelsohn, T. Foell, J. J. Goodman, *J. Org. Chem.* **26** (1961) 2859.
255. J. Fried et al., *J. Am. Chem. Soc.* **80** (1958) 2338.
256. G. B. Spero et al., *J. Am. Chem. Soc.* **78** (1956) 6213; **79** (1957) 1515.
257. G. Cooley, B. Ellis, D. N. Kirk, V. Petrow, *J. Chem. Soc.* 1957, 4112.
258. J. H. Fried et al., *J. Am. Chem. Soc.* **81** (1959) 1235.
259. Upjohn, DE 1 056 605, 1957 (F. H. Lincoln, W. P. Schneider, G. P. Spero); 1 123 321, 1958 (B. J. Magerlein, W. P. Schneider, G. B. Spero, J. A. Hogg).
260. Upjohn, US 2 867 638, 1957 (F. H. Lincoln, W. P. Schneider, G. B. Spero); DE-AS 1 082 261, 1959 (F. H. Lincoln, G. B. Spero, W. P. Schneider).
261. G. E. Arth et al., *J. Am. Chem. Soc.* **80** (1958) 3161.
262. E. P. Oliveto et al. *J. Am. Chem. Soc.* **80** (1958) 6687.
263. D. Taub et al., *J. Am. Chem. Soc.* **82** (1960) 4012.
264. Schering Corp., US 3 164 618, 1958 (R. Rausser, E. P. Oliveto).
265. E. P. Oliveto, *J. Am. Chem. Soc.* **80** (1958) 4431.
266. Lab. Français de Chimiothérapie, US 3 007 923, 1961 (G. Müller, R. Bardoneschi, J. Jolli).
267. R. E. Marker, H. M. Crooks, *J. Am. Chem. Soc.* **64** (1942) 1280.
268. W. R. Nes, H. L. Mason, *J. Am. Chem. Soc.* **73** (1951) 4765.
269. Upjohn, GB 869 511, 1959.
270. D. Taub et al., *J. Am. Chem. Soc.* **80** (1958) 4435; **82** (1960) 4012.
271. E. P. Oliveto et al., *J. Am. Chem. Soc.* **80** (1958) 4428, 6678.
272. Roussel-Uclaf, US 3 104 246, 1963 (G. Amiard, V. Torelli, J. Cérède).
273. H. L. Slates, N. L. Wendler, *J. Org. Chem.* **22** (1957) 498.
274. Merck Corp., GB 912 378, 1959.
275. Schevico, GB 901 093, 1958.
276. Merck AG, DE 1 134 074, 1959 (H.-J. Mannhardt, K.-H. Bork, K. Brückner, H. Metz).
277. H. J. Mannhardt et al., *Tetrahedron Lett.* 1960, no. 16, 21.
278. D. Taub et al., *J. Org. Chem.* **25** (1960) 2258.
279. J. A. Hogg et al., *Chem. Ind. (London)* 1958, 1002.
280. H. J. Ringold, G. Rosenkranz, F. Sondheimer, *J. Am. Chem. Soc.* **78** (1956) 820.
281. A. Bowers, H. J. Ringold, *J. Am. Chem. Soc.* **80** (1958) 4423.
282. J. A. Edwards, H. J. Ringold, C. Djerassi, *J. Am. Chem. Soc.* **82** (1960) 2318.
283. A. Zaffaroni et al., *J. Am. Chem. Soc.* **80** (1958) 6110. Syntex, US 2 671 752, 1951 (A. Zaffaroni).
284. Upjohn, US 2 841 600, 1957 (J. A. Hogg, G. B. Spero).
285. Syntex, DE 1 079 042, 1958 (E. Batres et al.).
286. Bayer, DE 1 088 953, 1958 (H. Lettré, D. Hotz).
287. Sterling Drug Inc., US 2 678 932, 1951 (J. S. Buck, R. O. Clinton).
288. J. S. Mills, A. Bowers, C. Djerassi, H. J. Ringold, *J. Am. Chem. Soc.* **82** (1960) 3399. Syntex, US 3 014 938, 1959(J. S. Mills, A. Bowers).
289. A. Bowers et al., *J. Am. Chem. Soc.* **81** (1959) 4107, 5991.
290. B. M. Bloom, V. V. Bogert, R. Pinson Jr., *Chem. Ind. (London)* 1959, 1317.
291. H. L. Ringold et al., *J. Am. Chem. Soc.* **80** (1958) 6464.
292. G. Raspé, K. Kieslich, U. Kerb, *Arzneim. Forsch.* **14** (1964) 450.
293. A. Doménico et al., *Arzneim. Forsch.* **15** (1956) 46.
294. R. Joly et al., *Arzneim. Forsch.* **24** (1974) 1.
295. V. Petrow, D. M. Williamson, *J. Chem. Soc.* 1959, 3595.
296. Schering AG, DE-AS 1 169 444, 1961 (K. Kieslich, U. Kerb, G. Raspé).

297. G. N. Shull, A. Kita, *J. Am. Chem. Soc.* **77** (1955) 763.
298. A. Nobile et al., *J. Am. Chem. Soc.* **77** (1955) 4184.
299. Schering AG, DE 1 135 899, 1960 (G. Raspé et al.).
300. Roussel-Uclaf, DE 1 159 441, 1961 (R. Joly, J. Warnant, B. Goffinet). DE 1 205 096, 1961 (R. Joly, J. Warnant, J. Jully, B. Goffinet).
301. Schering AG, DE-AS 1 249 270, 1963 (K. Kieslich, U. Kerb, G. Raspé).
302. C. H. Robinson, L. Finckenor, E. P. Oliveto, D. Gould, *J. Am. Chem. Soc.* **81** (1959) 2191.
303. Schering Corp., US 2 894 963, 1959 (D. H. Gould, H. Reimann, L. E. Finckenor).
304. Glaxo, US 3 312 590, 1967 (J. Elks, P. J. May, N. G. Weir).
305. Schering AG, US 3 824 260, 1974 (H. Laurent et al.).
306. A. Thalen, R. Brattsand, E. Gruvstad, *Acta Pharm. Suec.* **21** (1984) 109.
307. Schering Corp., US 3 444 217, 1969 (E. L. Shapiro, L. E. Finckenor).
308. S. Sugai et al., *Chem. Pharm. Bull* **33** (1985) 1889.
309. H.-J. Shue, M. J. Green, *J. Med. Chem.* **23** (1980) 430.
310. H. F. DeLuca in H. L. J. Makin (ed.): *The Metabolism and Function of Vitamin D in Biochemistry of Steroid Hormones,* 2nd ed., Blackwell Scientific Publ., Oxford 1984, pp. 71 – 116.
311. M. F. Holick, H. F. DeLuca, J. Suda, R. J. Cousins, *Biochemistry* **10** (1971) 2799.
312. K. Huldshinsky, *Dtsch. Med. Wochenschr.* **45** (1919) 712. H. Chick, E. J. Palzell, E. M. Hume, Studies of rickets in Vienna 1919 – 1922. *Med. Res. Counc. (G.B.) Spec. Rep. Ser.* **77** (1923).
313. F. A. Askew et al., *Proc. R. Soc. London, B* **107** (1931) 76.
314. A. Windaus, O. Linsert, A. Lüttringhaus, G. Weidlich, *Justus Liebigs Ann. Chem.* **492** (1932) 226.
315. A. Windaus, F. Schenck, F. von Werder, *Hoppe-Seylers Z. Physiol. Chem.* **241** (1936) 100.
316. L. F. Fieser, M. Fieser: *Steroids,* Reinhold Publ. Co., New York 1959.
317. R. B. Woodward, R. Hoffmann: *The Conservation of Orbital Symmetry,* Verlag Chemie, Weinheim and Academic Press, New York 1970.
318. H. L. Henry, A. W. Norman, Vitamin D: Metabolism and Biological Actions, *Annu. Rev. Nutr.* **4** (1984) 493. H. F. DeLuca, H. K. Schnoes, Vitamin D: Recent Advances, *Annu. Rev. Biochem.* **52** (1983) 411.
319. T. Kametani, H. Furuyama: *Synthesis of Vitamin D_3 and Related Compounds, Medicinal Research Reviews,* **vol. 7,** no. 2, J. Wiley & Sons, New York 1987, pp. 147 – 171.
320. D. H. R. Barton, R. H. Hesse, M. M. Pechet, E. Rizzardo, *J. Chem. Soc. Chem. Commun.* 1974, 203.
321. E. G. Baggiolini, J. A. Iacobelli, B. M. Hennessy, M. R. Uskokovic, *J. Am. Chem. Soc.* **104** (1982) 2945.
322. S.-J. Shiuey, J. J. Partridge, M. R. Uskokovic, *J. Org. Chem.* **53** (1988) 1040. A. Kutner et al., *J. Org. Chem.* **53** (1988) 3540. H. S. Gill et al., *J. Steroid Biochem.* **31** (1988) 147. M. J. Calverley, *Tetrahedron* **43** (1987) 4609.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Hydraulic Fluids

Standard Article

Uwe Jens Möller[1] and David G. Young[2]

[1]The Burmah Oil (Deutschland) GmbH, Hamburg, Federal Republic of Germany
[2]Castrol Technology Centre, Pangbourne, United Kingdom

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a13_165.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Functions of Hydraulic Fluids
3. Parameters to Describe the Properties
4. Additives to Hydraulic Fluids
5. Types of Hydraulic Fluids
5.1. Mineral Oils for Hydrostatic Machines
5.2. Mineral Oils for Hydrokinetic Systems
5.3. Mineral Oils as Multipurpose Oils
5.4. Aircraft and Marine Hydraulic Fluids
5.5. Fire-Resistant Hydraulic Fluids
5.6. Brake Fluids
5.7. Environmentally-Friendly Hydraulic Fluids
6. Various Aspects of the Use of Hydraulic Fluids

[Top of Page]

1. Introduction

Hydraulic fluids transmit energy in fluid-filled, pressurized systems. The hydraulic energy is transferred from the generating section, (the hydraulic pumps) to the motor section (the hydraulic motors, actuating cylinders, or turbines). The quantity, pressure, and direction of flow of the fluid are controlled by valves. Basically, two types of hydraulic principles can be distinguished: hydrostatic and hydrokinetic drives.

Hydrostatic drives utilize the fact that the pressure in a fluid is distributed evenly in all directions and thus displaces the fluid into closed but changing spaces. The pump displaces a certain volume of hydraulic fluid per stroke or revolution against the pressure in the system (hydraulic motor or actuating cylinder). The main areas of application of hydrostatic drives are (1) construction machinery, (2) machine tools, (3) tractors and agricultural equipment, (4) materials-handling technology, (5) aviation, (6) processing machinery, and (7) auxiliary drives and braking equipment in vehicles.

Hydrokinetic drives work with the kinetic energy of the flowing hydraulic fluid. Usually a generator (pump) and a motor (turbine) are positively connected to one another by the fluid stream in the most confined space possible. If there is a difference in the rotational speeds of the pump and the turbine an impulse exchange takes place and energy is transmitted to the slower impeller.

Hydrokinetic drives are found in road and rail vehicles and in marine applications. The fluids used in these cases are known as *power transmission fluids, automatic transmission fluids* (ATF), or *hydraulic gear oils.*

[Top of Page]

2. Functions of Hydraulic Fluids

The most important function of a hydraulic fluid is, of course, energy transfer. However, to ensure trouble-free operation hydraulic fluids have to perform a series of other functions.

Energy Transfer. It is important for pressure transmission that the volume (and, thus, density) of a hydraulic fluid does not change under pressure, i.e., the compressibility must be low. The flow properties (viscosity) should also be largely independent of pressure.

Lubrication. A good hydraulic fluid lowers the friction between moving sliding parts and reduces wear. Especially for reciprocating parts it is important that the fluid has a good load-carrying capacity at the reversal points.

Cooling. Unavoidable power dissipation in a hydraulic drive leads to an increase in temperature of the system. Effective heat removal improves the efficiency of a machine. A hydraulic fluid with a high specific heat and a low viscosity (low internal friction) makes a particular contribution here.

Corrosion Protection. Steel or nonferrous metal surfaces must be protected against corrosion, including periods of fairly long downtime. Changes in the hydraulic fluid during use, that are caused by oxygen, heat, water, etc., must not reduce the corrosion protection.

System Cleanliness. Precision hydraulic machines, e.g., servo valves contain control systems with gap widths of less than 1 µm. Oil-insoluble products, such as dust, abrasive particles, or high-molecular oxidation products, must not produce any deposits in the control equipment. The equipment functions at its best if the foreign matter is held in suspension and carried through to the filter system.

[Top of Page]

3. Parameters to Describe the Properties (→ Lubricants and Lubrication)

In addition to the "conventional parameters," (color, flash point, total acid number, TAN, etc.) the "operating properties" must also be known to achieve trouble-free operation of a hydraulic machine. These include those properties (expressed by parameters) which are important for the projected use of the hydraulic fluid and properties which may change during use.

Viscosity (ASTM D445, IP71, DIN 1342, 51 550, 51 562, 5 015). A suitable viscosity is the most important criterion when selecting a hydraulic fluid. This is determined by the design of the hydraulic system and is governed to a great extent by the requirements of the hydraulic pumps or motors [1]. In order to achieve low flow losses and very high precision during control operations a hydraulic liquid should have the minimum viscosity accepted by the equipment manufacturer. Table 1 lists some guide values for maximum permissible initial and minimum permissible operating viscosities for various types of hydraulic pump [2]. At the *highest operating temperature* the kinematic viscosity should not fall below 5 mm^2/s. Exceptions are water-based hydraulic systems so-called pressurized-water systems. At low viscosities the wear increases quickly and the efficiency of the hydraulic plant decreases rapidly. At *low temperatures* hydraulic pumps will only draw in hydraulic fluid up to viscosities of approximately 1000 mm^2/s.

Table 1. Reference values for maximum permissible initial and minimum permissible operating viscosities for various types of pump

Type of pump	Maximum permissible initial viscosity, mm^2/s	Minimum permissible operating viscosity, mm^2/s
Gear pumps		
Without hydraulic gap compensation		
	1000–2000	25
With hydraulic gap compensation		
	1000–2000	15
Rotary vane pumps	850	20*
Fixed vane pumps	850	25*
Radial and axial piston pumps		
With control plane or edges	500 – 850	10 – 15
With sprung-loaded valves	200 – 300	8 – 12

Viscosity–Temperature (V–T) Behavior (ISO 2909, DIN 51 563). The viscosity of virtually all liquids decreases with increasing temperature. The viscosity index (VI) is an empirical measure of the V-T behavior. Fluids having a large viscosity–temperature dependence are commonly referred to as 'low-viscosity index' fluids. Those having a small viscosity–temperature dependence are known as high-viscosity index fluids. In the case of the normal demands on hydraulic oils a VI of ca. 100 is generally used. For sensitive control systems a VI of 100 is not adequate. Examples are hydraulic motors in machine tools or military equipment; and "mobile hydraulics" in aircraft, railways, motor vehicles, and construction equipment. In such cases *high VI hydraulic fluids* (HVI) are used. They are composed of mineral oils with polymer additives as VI improvers, (e.g., polymethacrylates) or occasionally synthetic oils.

Viscosity–Pressure (V–p) Behavior. At constant temperature the viscosity of hydraulic fluids increases exponentially with increasing pressure. The viscosity–pressure coefficient is larger at low temperatures than at high temperatures and the viscosity–pressure dependence is more pronounced in high-viscosity oils than in low-viscosity oils. Thus, the viscosity–pressure curve for a polyoxyalkylene–water mixture is significantly flatter than for mineral oils [3]. There are no standardized methods for determining the pressure dependence of viscosity [4].

Viscosity–Temperature–Pressure Behavior. The opposing effects on viscosity caused by an increase in pressure and temperature to some extent compensate one another. A three-dimensional form of diagram is used to show the combined effect [5].

Density–Temperature–Pressure Behavior (ISO 91 Pt. 1, DIN 51 757). Pressure losses in pipes are directly proportional to the density of the operating fluid. The density decreases with increasing temperature and increases with increasing pressure. For mineral oils the *volumetric expansion coefficient* is ca. 70×10^{-5} °C^{-1}. Up to ca. 300 bar the average *compressibility* of air-free mineral oils is 70×10^{-6} bar^{-1} [1]. At increased pressure liquids have a greater resistance to reduction in volume, i.e., with increasing pressure the fluids have a "harder" action [6].

Air-Dissolving and Air-Separating Ability (ASTM D 2779, ASTM D 3427, ASTM D 3827, DIN 51 381, IP 313). Under normal conditions hydraulic fluids can dissolve significant quantities of air: mineral oils and phosphate esters dissolve 7–9 vol %; aqueous solutions up to 5 vol % [7]. A certain time is required to dissolve air bubbles at increased pressure: ca. 10 s for 10 mm^2/s mineral oils at 2.0 MPa and 40 °C, and ca. 60 s for 40 mm^2/s oils [8]. Dispersions of air bubbles in hydraulic oils are undesirable because they increase the compressibility, promote aging of the fluid, reduce the thermal conductivity and can lead to cavitation damage in the pumps [9]. The ability of a fluid to expel dispersed air is known as its air release property. Under the conditions of the test standard DIN 51 381 the time is determined in which a dispersed quantity of air is released from the oil down to a residual air content of 0.2 vol %. The air release approaches this value asymptotically [10]. There is still no well-founded experience available to determine the quantity of dispersed air which leads to serious damage in hydraulic systems.

Foaming Behavior (ASTM D 892). If stable air bubbles build up on the surface of a fluid and are enclosed in extremely thin films of oil then unwanted surface foam has formed. This can lead to overflowing of the hydraulic oil container, or in the worst cases, to foam instead of oil being drawn into the system. Defoamers containing silicone oil remedy the situation but have an adverse effect on the air release properties [1]. More recent formulations contain defoamers based on polydimethylsiloxanes. The laboratory test as specified in ASTM D 892 in which a foam is produced in the liquid with the aid of a spherical gas diffuser stone has only limited practical value.

Behavior in the Mixed-Friction Zone (ASTM D 2271, ASTM D 2783, ASTM D 2882, DIN 51 350 Pt. 2, DIN 51 354 Pt. 2, Normungsaufgabe 0 051 589 Pt. 2). In a hydraulic system friction parts move (at least for short periods) under conditions of boundary friction. In this situation the normal force components which are not hydrodynamically supported cause partial elastic and plastic deformation of the sliding components. Physically-acting *antiwear additives* based on phosphorus, compounds and fatty acids and their derivatives, reduce wear and friction.

The FZG gear rig is used to determine the maximum load-carrying ability with stepwise increase in loading. The method is suitable for assessing the load-carrying ability of hydraulic oils and experience has shown that it gives very good agreement with axial piston pumps.

The Vickers 35VQ25 vane pump provides a practical test of the antiwear properties of hydraulic oils, the loss in mass of the vane and ring being measured before and after test run.

Seal Compatibility (ISO 1817, ISO 6072, ASTM D 3604, DIN 53 538 Pt. 1, Pt. 2 and Pt. 3, CETOP (Comité Européen des Transmissions Oléohydrauliques et Pneumatiques) RP 81H. There must be very little swelling, hardening, shrinking or softening of rubber-elastic sealing materials in contact with the hydraulic fluid. Apart from the type of sealing material and the operating conditions, the basic oil and additives (and any decomposition

products) have the greatest effect on the interaction between the sealing material and the hydraulic oil [12]. Due to the difficulty of applying static storage tests to dynamic operating processes there is as yet no satisfactory means of predicting seal compatibility. In the ISO 1817 method, disks of the sealing material are stored with all sides exposed in the hydraulic oil. Appropriate standard reference elastomers are specified in ISO 6072, DIN 53 538 Pt. 1 to Pt. 3, and in the CETOP R 81H regulations [13]. At the end of the test period the changes in mass, volume, hardness and tensile strength of the test piece are determined.

Fire-Resistance (ASTM D 3119, ASTM E 659, DIN 51 719, FTM 791b-6053) Fire-resistant hydraulic fluids are used in cases where the escaping fluid can come into contact with the sources of ignition. The method specified in DIN 51 794 is used for mineral oil products. Here a small sample quantity is placed in a heated, open Erlenmeyer flask. The lowest value at which combustion occurs in a series of tests is recorded. In the United States manifold ignition test (FTM 791b-6053) the fluid is dropped onto a hot discharge pipe and the temperature of combustion is recorded. The *mining industry* determines the flammability of a fluid sprayed under pressure [14]. There are differences between the continental and British tests [15]. In both cases the flame propagation is measured in a mixture of pulverized coal and fluid. The American industry applies two types of test: The United States Mining Safety and Health Administration Schedule 30 and the Factory Mutual-Less Hazardless Hydraulic Fluid Approval Standard 6930. Each test consists of two sets of tests with the pure fluid. The term “fire-resistant” has not yet been defined; it is not identical with “non-combustible.”

Cooling Ability. In a typical hydraulic system the proportion of power which is converted into heat is 15–25 % [16]. The specific heat and thermal conductivity of the hydraulic fluid used is necessary to calculate the power loss, dimensions of heat exchangers and amount of preheating (see Table 2).

Table 2. Calorific data of hydraulic fluids

	ISO/DIS	Water	Mineral oil 6074	HFA* 6071	HFB* 6071	HFC* 6071	HFD* 6071
Specific heat	W kg^{-1} K^{-1}	1.163	0.523	1.163	0.93	0.92	0.34
Thermal conductivity	W m^{-1} K^{-1}	0.628	0.129	0.628	0.523	0.309	0.114

Corrosion Protection (ASTM D 4636, DIN 51 585, FTM 7916–5308, SAE J 1703). The simultaneous action of oxygen and water on a metal surface forms galvanic cells which can lead to damage of the surface [1].

In the method as specified in DIN 51 585 a steel rod is immersed in a mixture of oil and water which is stirred for 24 h at 60 °C. However, this *dynamic test* is not sufficient for identifying the degree of corrosion protection when plants are shut down and parts covered with a thin film of oil are permanently in contact with moist air.

So-called immersion tests in which various metals and metal pairs are immersed in the fluid are used as *static tests*. Examples are magnesium, aluminum, cadmium, steel and copper as specified in FTM 791b-5308; or tin plate, steel, aluminum, cast iron, brass, and copper as specified in SAE J 1703.

System Cleanliness. The addition of dispersants as cleaning agents has the effect of dispersing solid foreign matter and suspending small quantities of water in the hydraulic oil. Any excess of water is rapidly precipitated. Suitable additives can significantly prolong the service life of hydraulic oils [17]. To test the cleaning ability solid precipitates are formed by hydrolyzing suitable additives, particularly zinc dialkyldithiophosphates. Subsequent filter tests are then used to test the filterability of the particles held in suspension [18]. Hydraulic oils with these properties are generally designated as “HLPD” without it being described more precisely in a specification standard. Further tests on hydraulic fluids comprise the determination of the hydrolytic stability (ASTM D 2619), the thermal stability (ASTM D 2160), and the deposition tendencies (ASTM D 3711). The standard practice ASTM D 4174 is intended to aid operators towards maintaining clean petroleum fluid hydraulic systems.

[Top of Page]

4. Additives to Hydraulic Fluids

Lubricating additives (⟶ Lubricants and Lubrication), are highly important components of modern hydraulic fluids. Whereas some properties of hydraulic oils cannot in practice be influenced by additives (e.g., *V–p* behavior), improvements of other properties are possible (e.g., anticorrosion, foaming) or the additives even give the hydraulic oils properties which are completely lacking in the base fluid (e.g., antiwear behavior). Additives in hydraulic fluids can have three different functions: (1) protection of the metal surface, (2) protection of the pressurized fluid, and (3) increasing the efficiency [19]. The most important additives for hydraulic fluids are listed in Table 3.

Table 3. Breakdown of additives for hydraulic fluids

Additive type	Function	Type of material	Mode of operation
Protection of the metal surface			
EP additives	prevention of surface damage under severe loading	organic compounds with sulfur, phosphorus or chlorine	chemical reactions with metal surface; formation of boundary layers
Antiwear additives	prevention of surface damage under light loading	fatty acids and derivatives	chemisorption at the metal surface; formation of boundary layers
Corrosion inhibitors	prevention of ingress of oxygen and water	fatty acids, sulfonates, salts of fatty acids	formation of coating layers, impermeable to water or air
Protection of the hydraulic fluid			
Oxidation inhibitors	inhibition of the reaction between air and fluid	phenols, amines, sulfides	interruption of the oxidation radical chain mechanism
Defoamers	prevention of the formation of stable foam	silicone oils	destruction of the oil skin surrounding the air bubbles
Increasing the efficiency			
VI improvers	reduction in the temperature dependence of viscosity	polyolefins, polymethacrylates, polyalkylstyrenes	greater increase in viscosity at higher rather than at lower temperatures
Pour-point depressants	lowering of the pour point	polymethacrylates, condensation products	coating of paraffin crystals and prevention of their growth
Demulsifiers	separation of oil and water	ionogenic and nonionogenic polar compounds	interaction with the oil–water interface
Dispersants	prevention of deposits	sulfonates, amides, surfactants	peptization

Additives which are suited to the type of fluid concerned are used alone or in multiple combinations to fulfil given specifications. Other additives are used for special purposes, e.g., emulsifiers, thickeners, friction modifiers.

[Top of Page]

5. Types of Hydraulic Fluids

Basically, any liquid is suitable for transmitting pressure. Water is therefore the cheapest medium and is universally available. Pure water, however, is too corrosive, can be used only between 0 and 100 °C, and does not provide sufficient wear protection for moving parts because of its low viscosity. These systems also require a significant amount of maintenance (e.g., treatment with biocides). Thus, the majority of hydraulic oils today have mineral oils as the base fluid. The required properties are achieved by choosing suitable mineral oil fractions and adding selected additives. Other base liquids are used for special power transmission duties.

5.1. Mineral Oils for Hydrostatic Machines

Hydrostatic machines, e.g., machine tools, presses, and construction machinery, normally require kinematic viscosities between 32 and 100 mm^2/s at 40 °C (ISO viscosity grades, VG, 32–100). Control plants, tippers and small loaders also use hydraulic fluids with lower viscosities of 10 to 22 mm^2/s at 40 °C (ISO-VG 10 to 22). Mineral-oil-based hydraulic oils for hydrostatic machines are therefore available in grades ISO VG 10 to 100 [1], [20].

Depending on the requirements resulting from the structural specification of the hydraulic machine and on the operating conditions there are five different qualities of hydraulic fluids available in the range ISO VG 10–100 (see Table 4).

Table 4. Mineral-oil-based hydraulic oils

German designation	International designation	Properties	Additives	Application	Quality specifications
Hydraulic oil H	HH	not oxidation inhibited	none	for lightly loaded systems	DIN 51 517 Pt. 1
Hydraulic oil HL	HL	oxidation-inhibited rust-preventing	ca. 0.5 %	for moderately loaded systems	DIN 51 524 Pt. 1 ISO 6074 SEB 181 222 VDMA 24 318
Hydraulic oil HLP	HM	oxidation-inhibited rust-preventing	ca. 1.0 %	for high pressure systems	DIN 51 524 Pt. 2 ISO 6074 SEB 181 222 VDMA 24 318
Hydraulic oil HVLP	HV	as for HM, shallow VT curve, high VI	ca. 8 %	for high pressure plants with fluctuating temperatures	DIN 51 524 Pt. 3 ISO 6074
Hydraulic oil HLPD		as for HM, dirt-suspending, limited water-compatibility, foam controlled	ca. 2.0 %	for systems with ingress of water	DIN 51 524 Pt. 4 SEB 181 222 VDMA 24 31

The characteristic data for the specification standards as specified in DIN 51 517 Pt. 1, DIN 51 524 Pt. 1, DIN 51 524 Pt. 2, DIN 51 524 Pt. 3, SEB (Stahl-Eisen-Betriebsblätter) 181 222 and CETOP RP 91H for ISO-VG 46 are listed in Table 5. The NATO Code H-540 (TL 9150–0035) designates a hydraulic oil for an operating range from – 30 to + 90 °C with a base oil comprising a mixture of refined mineral oil and polyalphaolefins.

Table 5. Selected quality data for mineral-oil-based ISO VG 46 hydraulic oils

Property	Test method	Units	DIN 51 517 Pt. 1	DIN 51 524 Pt. 1	DIN 51 524 Pt. 2	SEB 181 222	CETOP RP 91H	DIN 51 524 Pt. 3
Classification	ISO/DIS 6074		HH	HL	HM	HM	HV	HV
Viscosity at 0 °C	DIN 51 562	mm^2/s		<780	<780	<700		<420
Viscosity at + 40 °C	DIN 51 562	mm^2/s	41.4–50.6	41.4–50.6	41.4–50.6	46	41.4–50.6	41.4–50.6
Viscosity at + 100 °C	DIN 51 562	mm^2/s		>6.1	>6.1	6.2		>7.9
Viscosity index	ISO 2909		a	a	a	a	>130	a
Pour point	ISO 3016	°C	<–15	<–15	<–15	<– 15	<– 36	<–33
Flash point	ISO 2592	°C	>175	>185	>185	>190	>150	>180
Corrosion protection properties	DIN 51 585			0–A	0–A	0–A	0–A	0–A
Copper corrosion	DIN 51 759			2–100A3	2–100A3	1–125A3	2	2–100A3
Aging behavior	DIN							

	51 587							
Increase in the neutralization number		mgKOH/g		<2	<2	<2	b	<2
Elastomer behavior	DIN 53 521							
Rel. volume change		%	c	0–12	0–12	c	b	0–12
Change in Shore hardness				0–7	0–7	c	b	
Air release property	DIN 51 381	min		<10	<10	<10	b	<10
Demulsifying behavior	DIN 51 599	min	c	<40	<40	<20	<40	<40
Foaming behavior	ASTM D 892	mL		150/75/150	150/75/150	100/50/100	300/100/300	150/75/150
Load-bearing ability	FZG DIN 51 354	load stage			>10	>10	>10	>10
Wear, vane-cell pump	DIN 51 389	mg			<150	<150	<150	<120
Wear, Timken test	SEB 181 232	mg	a	a	a	<6	a	a

[a] Not required.
[b] Will be included later in the standard.
[c] Report.

5.2. Mineral Oils for Hydrokinetic Systems

In hydrokinetic systems (e.g., hydrodynamic couplings; turbo gears systems with power transmission oils; automatic gear systems in vehicles) refined mineral oils are used which have a low viscosity and low viscosity-temperature dependence in order to achieve a high efficiency through low flow losses. The viscosity should not exceed 40 mm^2/s at 40 °C and the VI may reach 150.

Good oxidation protection is required for oils in *hydrodynamic couplings* and converters as the slip can cause severe heating. Higher demands are placed on *power transmission oils*. These are (1) corrosion protection in the oil and vapor phases; (2) good air release properties; (3) low tendency to foam; and (4) high load-bearing ability for lubrication of the heavily-loaded toothed gear systems. The oils used for *lubricating* and *controlling* motor vehicles with automatic transmission systems are called *automatic transmission fluids* (ATF). Because of their many functions and the high demands placed on them ATFs contain more additives (friction modifiers, oxidation inhibitors, and antiwear additives). The two commercially available types of fluid differ principally in their coefficients of friction: the coefficient of friction of Dexron II-ATF from General Motors (specification GM 3032M or GM 6137M) is considerably lower when static than when in motion; Ford type F fluids (specification M2C-33F or M2C-9007A) follow the natural friction pattern of mineral oils with high static and lower kinetic coefficients of friction [21].

5.3. Mineral Oils as Multipurpose Oils

Universal tractor transmission oils (UTTOs) are used for hydraulic performance in tractor applications, where manufacturers will have their own specifications that need to be met. Low viscosity motor oils, SAE 10W or SAE 5W, are also used as the hydraulic medium for the steering hydraulics in motor vehicles [22].

Since the end of the 1970s universal oils have been on the market for lubricating the motor and for the entire power transmission system of tractors. They are known as STOU (super tractor oil universal). The STOUs correspond principally to the viscosity grades SAE 10W-30, 10W-40, or 15W-30, fulfil the requirements of API CF-4/CE/SF and satisfy the gear oil specification MIL-L-2105 (API GL4) and industrial specification DIN 51524 Part 3 [23]. Tractor manufacturers require that their own specifications should be met for wet hydraulic braking systems, couplings and hydraulic equipment [24]. Dual-purpose, water containing metal-working and hydraulic fluids are on the market (MH-fluids) [25].

5.4. Aircraft and Marine Hydraulic Fluids

In aircraft, the ailerons, vertical rudder, elevators, tailplane, spoilers, landing flaps, equipment for lowering the landing gear, steering and braking equipment are actuated by hydrostatic hydraulic systems. The particular requirements for hydraulic oils used in aircraft are: (1) extremely low wear in pumps and valves, (2) very good low-temperature behavior (down to – 54 °C), and (3) very low tendency to fire propagation [26]. At present there are four groups of aviation hydraulic oils in use: (1) mineral oils (specification MIL-H-5606, NATO Code H-515); (2) polyalphaolefin oils (specification MIL-H-83 282) [27]; (3) oils made from alkyl or aryl phosphoric acid esters (Boeing company specification BMS 3-11E); and (4) the silicate ester-based oils (used in Concorde supersonic aircraft). Novel silahydrocarbon-based fluids have been under test in the 1980s [28], [29].

A comparison of selected requirements for hydraulic oils for aircraft is given in Table 6. Chlorofluoroalkane oils, which are substantially more fire-resistant than previous hydraulic oils, are under trial (Table 7) [26]. Hydraulic oils used in marine applications must have outstanding corrosion protection properties, e.g., marine hydraulic oil, NATO Code H-570. Consequently they must pass the corrosion test according to ASTM D 665 with sea water.

Table 6. Selected requirements for aviation hydraulic oils

	MIL-H-5 606 G	MIL-H-83 282 D
Base oil	mineral oil	polyalphaolefins
Viscosity (– 54 °C) mm^2/s	<2500	*
Viscosity (– 40 °C) mm^2/s	<600	<2200
Viscosity (+ 100 °C) mm^2/s	>4.9	>3.45
Pour point, °C	<–60	<–55
Elastomer swelling, %	19–30	18–30
Wear scar in the four-ball test, mm	<1.0	< 0.65

Sludge volume, mL	<65	<65

* Not required.

Table 7. Ignition and combustion properties of aviation hydraulic oils

	Flash point, °C (COC)	Fire point °C	Spontaneous ignition temperature, °C	Ignition on hot surface, °C
Test method (ASTM)	D 92	D 2155	E 659	D 1241
Type of aviation hydraulic oil				
Mineral oil (MIL-H-5606)	>82	113	232	593
Polyalphaolefin (MIL-H-83 282)	>204	250	344	538
Phosphoric acid ester (BMS 3–11E)	171	186	524	760
Silicate ester	200	250	380	
Chlorofluoroalkane	none	none	643	none

5.5. Fire-Resistant Hydraulic Fluids

The flammability of mineral oils and their subsequent combustion can be dangerous in many industrial plants. Examples of this type of plant are: hydraulic systems in underground mining, in blast furnaces, in foundries, for welding equipment, and for injection molding machines. Fire-retardant hydraulic fluids have been developed for these applications since 1960 [30]. In accordance with a recommendation in the 7th Luxembourg Report [14] the fire-resistant hydraulic fluids (HF) were divided into four classes as specified in ISO/DIS 6071 (see Table 8).

Table 8. Types of fire-resistant industrial hydraulic fluids

Designation	Description	Water content, %	Viscosity at 40 °C, mm^2/s	Range of application, °C
HFA	oil-in-water emulsion	> 80	ca. 1	+ 5–+ 55
HFAE	HFA with 80 % water [a, d]			
HFAS	solution of chemical substances [a, e]			
HFB	water-in-oil emulsion	> 40	46–100	+ 5–+ 60
HFC	aqueous polymer solutions	> 35	22–100	– 20–+ 60
HFD	water-free synthetic liquids		15–100	– 20–+ 150
HFDR	phosphoric acid ester [c]			
HFDS	chlorinated hydrocarbons [c]			
HFDT	mixtures of HFDR and HFDS [c]			
HFDU	other compositions [c, f]			

[a] ISO 6743/4.
[b] ISO/DIS 6071.
[c] ISO 6743/4; ISO/DIS 6071; VDMA 24 317; CETOP RP 77H.
[d] DIN 24 320.
[e] Mainly polyalkylen glycols.
[f] Not specified.

HFA fluids (pressurized water) with water contents of 95–98 % are the only fire-resistant hydraulic fluids that are effectively incombustible. Due to the low viscosity of the fluid (practically that of water) valves with gap-sealing cannot be used at high pressures and the antiwear properties are very poor [31], [32]. Attempts are being made to overcome these disadvantages by the use of microemulsions [33] (colloidal oil droplets of 1–5 µm in diameter) and by the addition of water thickeners (raising the viscosity to 150 mm^2/s at 40 °C) [34], [35].

HFB fluids are in widespread use in the United Kingdom and the United States [36]. However, they do not play any role in.

The majority of *HFC fluids* are aqueous *polyoxyalkylene solutions* (⟶ Polyoxyalkylenes) containing about 10 % additives. Their antiwear properties are unsatisfactory, especially with roller bearings.

Phosphate ester based oils (*HFDR* fluids) have good fire resistance and excellent antiwear/anticorrosion properties. However they are expensive and also require careful maintenance—for example, to prevent water contamination and in turn build up of acidic components.

Polyol ester based fluids (*HFDU*) also claim similar performance benefits and are more cost-effective. Although they are not considered to be as fire resistant as phosphate ester based fluids, they are more commonly used in many industrial applications where a synthetic fluid is preferred.

5.6. Brake Fluids

Specifications. Brake fluids are specified according to SAE standards, of which SAE J1703 is the best known (see Table 9). Equally well known are the FMVSS (Federal Motor Vehicle Safety Standard) specifications issued by the United States Department of Transportation (DOT). These standards are recognized worldwide by vehicle manufacturers. The ISO standard 4925 corresponds with these specifications. The reader is referred to SAE Specification J1709 for a description and classification of the different types of brake fluid.

Table 9. Specifications for brake fluids with the required minimum boiling point

Specification	Designation	Minimum boiling point, °C

SAE J 1703	motor brake fluid	205
SAE J 1704	borate ester based brake fluid	230
FMVSS 116 DOT 3	motor vehicle brake fluid	205
FMVSS 116 DOT 4	motor vehicle brake fluid (borate)	230
FMVSS 116 DOT 5.1	motor vehicle brake fluid (borate)	260
FMVSS 116 DOT 5	silicone based brake fluid	260
ISO 4925	motor vehicle brake fluid	205

Glycols as Brake Fluids. In motor vehicles the pressure developed in a master brake cylinder is transmitted by the brake fluid to the wheel brakes where a friction surface is pressed against a rotating drum or disk. The main requirements of such fluids are a high boiling point and compatibility with the brake tubing and seals. In addition to glycols of the general formula HO —R —OH, use is also made of glycol ethers (HO —R —O —R′). These materials are hygroscopic. Components are added to enhance lubricity, antiwear and antioxidation properties and for corrosion inhibition. Brake fluids are classified by three main physical characteristics: (a) dry equilibrium boiling point as defined by specification (boiling point), (b) wet equilibrium boiling point—relating to simulated years of (hygroscopic) fluid in service, and (c) kinematic viscosity at – 40 °C.

Newer Brake Fluid Types.
Borate Esters. Developments in the 1970s led to borate esters (glycol ether borates) which have higher boiling points and higher performance characteristics than glycol ethers, but are still hygroscopic.

Silicon Esters. In the late 1990s, silicon esters, (glycol ether silicates) with very high performance were developed. These fluids are expensive, and used mainly in racing applications.

Silicone Oils. These are used in some applications but are expensive and more compressible than current glycol ether derivatives.

Mineral Oils. In some specific vehicle applications, specialized mineral oil based brake fluids are used. These are not compatible with glycol ether-based fluids and must never be mixed. ISO specification 7308 relates to these fluids.

5.7. Environmentally-Friendly Hydraulic Fluids (⟶ Chlorinated Hydrocarbons – Chlorinated Biphenyls)

Environmentally-friendly hydraulic fluids are slowly gaining popularity with growing environmental awareness and, in particular, increasing legislation stipulating use of ‘bio’ products in sensitive outdoor applications. Natural oil-based products (HETG type) are typically based on rapeseed oils and have good biodegradability and lubricity. However, they have poor oxidative stability (which can lead to deposits in severe applications), reduced service life and variable low temperature properties. Products based on synthetic esters (HEES type) also show good biodegradability, improved oxidative stability, good low temperature properties and longer service life. However, cost is an issue as they are six to ten times as expensive as natural oil-based products. Compatibility with seal materials is a concern for both types of fluid.

[Top of Page]

6. Various Aspects of the Use of Hydraulic Fluids

Filtration Problems. Impurities in hydraulic systems lead to premature wear of the parts and to unforeseen malfunctions. Continuous cleaning of the operating fluid (usually with filters or electrostatic liquid cleaners [40]) is therefore necessary. The choice of filter is governed by the operating conditions and requirements. The filter mesh size, the dirt-retaining capacity of the filter, the permissible pressure drop and the operating viscosity of the fluid [41] are the most important parameters to be considered. The *filter fineness* is required to achieve a given degree of purity [42], [43]. This depends mainly on the hydraulic system and its gap widths [44], [45].

The *cleanliness of a liquid* can be given by two numbers as specified in ISO/DIS 4406: the first number refers to the number of particles per milliliter with a diameter of >5 µm and the second figure to particles above 15 µm diameter; e.g., 15/12 = 160–320 particles per milliliter of diameter greater than 5 µm of which 20–40 particles per milliliter have a diameter greater than 15 µm. The relation between number of particles and the corresponding code is given in ISO/DIS 4406. The *efficiency of a filter* is expressed by the β-value [46]. This is determined in the ISO 4572 Multipass Test as the quotient of the number of particles before and after the filtration. Table 10 shows the purity classifications for the fluid and the β-values for the filter for selected areas of application [47].

Table 10. Areas of application, purity classification of the fluids and β-values of the filter

Area of application	Purity classification ISO/DIS 4406	β-Value
Machine tools	8/4–14/10	75 *
Hydraulic presses	8/4–14/10	75
Plastic injection machines	9/5–13/9	75
Construction machines	10/4–12/10	75
Excavators	13/6–20/13	75
Control hydraulics	10/6–13/8	75
Mobile hydraulics	13/6–18/13	75
Hydrostatic drives	13/7–17/10	75

* 75 Corresponds to a separating efficiency of 98.67 %.

Change of Type of Hydraulic Oil. If a change is to be made from a mineral oil to an aqueous fluid or a synthetic oil, or the reverse, a number of points must be taken into account. The questions regarding the disposal of the used fluid, material, tubing, paint, sliding bearings and filters, tolerances, auxiliary cleaning aids, etc. are listed in the VDMA (Verband deutscher Maschinenbau-Anstalten) 24 314 guidelines. The cost of equipment alterations may also need to be taken into account (for example, if switching from a mineral oil to HFA type fluid). (see Table 11) [48].

Table 11. Cost comparison of hydraulic equipment for a transfer line

	Operation with mineral oil	Operation with HFA fluid
Plant costs, %	100	187

Energy costs, % 100	190

Condition Monitoring. Where fairly large quantities of fluid are involved, the serviceability of a working fluid in operation is determined by laboratory testing of used oil samples, sometimes in combination with wear metals analysis [49]. There are standardized guidelines (ISO 4021; CETOP RP 91H) and practical advice for taking representative samples [50], [51]. In addition to systematic laboratory tests [1], [52] it is sometimes also advantageous to use rapid analysis kits [53], [54].

Toxicology and Occupational Health. Mineral oils are relatively nontoxic substances. Their proper use is not associated with any increased health risk if the usual working precautions are observed for avoiding skin contact or inhalation of vapors and mists [1].

Biocides, which often operate by releasing formaldehyde, are used with the *HFA* and *HFB* hydraulic fluids to combat infection by microorganisms. The concentration of biocide is generally so low that neither the MAK value (0.5 ppm in 2002) nor the TLV (1.5 mg/m^3) for formaldehyde is reached. However, specialist advice should be sought if there are any doubts regarding health issues if using these fluids.

The *HFDS fluids* were composed of polychlorinated biphenyls (PCBs, trademarks: Avoclor, Askarel, Clophen, Pyroclor, etc.). It is no longer permissible to manufacture and use PCBs because of their persistent, and presumed carcinogenic, properties.

The main product group of the *HFDR fluids* are the triarylphosphates which have a series of isomers showing pronounced toxic properties. However, starting from 3,4- and 3,5-xylenols it is possible to prepare products which have met the strict toxicity requirements of the 7th Luxembourg Report [14].

[Top of Page]

References

1. U. J. Möller, U. Boor: *Schmierstoffe im Betrieb*, VDI-Verlag, Düsseldorf 1987.
2. H. Dietterle: “Druckflüssigkeiten,” in: *ölhydraulik und Pneumatik*, **vol. 1**, *Grundlagen der ölhydraulik*, Krauskopf-Verlag, Mainz 1970.
3. M. Weber: *Druckölviskositätsmessung*, FVA/DGMK-Studie, DGMK-Forsch.-Ber. 381. DGMK, Hamburg 1987.
4. H. W. Thoenes,*VDI-Ber.* **138** (1969) 23–31.
5. H. Peeken, M. Spilker, *ölhydr. Pneum.* **25** (1981) no. 12, 903–907; **26** (1982) no. 1, 24–27.
6. H. Peeken, J. Blume, *Konstr.* **35** (1983) no. 12, 473–479.
7. D. Staack, *Schmierungstech. Tribol.* **34** (1987) no. 4, 201–207.
8. O. Menzel, W. Hamann: “Gase in Hydrauliköl en,” in W. J. Bartz (ed.): *Hydraulikflüssigkeiten—Reihe Tribotechnik*, vol. **1**. Vincentz-Verlag, Hannover 1982, pp. 87–101.
9. B. Basu, S. Chand, K. C. Jayaprakash, S. P. Srivastava *et al.*, *ASLE Trans.* **28** (1985) no. 3, 313–324.
10. T. Mang, H. Jünemann, *Mineralöltechnik* **18** (1973) no. 2, 1–24.
11. H. Jünemann, *Erdöl Kohle Compend.* **75/76** (1976) 484–497.
12. K. Nagdi: “Wechselbeziehungen zwischen Hydraulikflüssigkeiten,” in W. J. Bartz (ed.): *Reihe Tribotechnik*, vol. **1**. Vincentz-Verlag, Hannover 1982, 125–152.
13. K. Nagdi, *ölhydr. Pneum.* **23** (1979) no. 2, 88–91.
14. Commission of the European Communities: Seventh Report on Specifications and Testing Conditions Relating to Fire-Resistant Hydraulic Fluids Used for Power Transmission in Mines. Luxembourg 2002.
15. W. Guse: “Wasserhaltige und synthetische Hydraulik-Flüssigkeiten,” in W. J. Bartz (ed.): *Hydraulikflüssigkeiten—Reihe Tribotechnik*, vol. **1**, Vincentz-Verlag, Hannover 1982, 223–242.
16. U. W. Krämer, *ölhydr. Pneum.* **28** (1984) no. 1, 15–25.
17. M. Okada, M. Yamashita, *Lubri. Eng.* **43** (1987) no. 6, 459–466.
18. R. Benda, *Mineralöltechnik* **30** (1985) no. 2, 1–12.
19. H. Ebertshäuser, *ölhydr. Pneum.* **28** (1984) no. 7, 448–454.
20. B. R. Chadha, L. R. Gupta, D. P. Dobhal, *Ind. Lubr. Tribol.* **38** (1986) no. 4, 150–154.
21. H. E. Deen, J. Ryer, SAE Paper 841 214 (1984).
22. H. W. Thoenes, H.-J. Hoffmann: “Hydraulikflüssigkeiten,” in W. J. Bartz (ed.): *Handbuch der Betriebsstoffe für Kraftfahrzeuge*, Part 1, expert verlag, Grafenau 1982, 101–149.
23. Ein Öl für alle Fälle?? *top agrar* (1979) no. 11, 68–71.
24. A. Grünwald, A. A. Hubmann, A. Marhold: “Motorenöle,” in W. J. Bartz (ed.): *Handbuch der Betriebsstoffe für Kraftfahrzeuge*, Part 2, expert verlag, Grafenau 1983, 145–260.
25. R. Lemar, *Lubri. Eng.* **41** (1985) no. 8, 494–498.
26. E. Jantzen, *ölhydr. Pneum.* **28** (1984) no. 5, 320–323.
27. C. E. Snyder, L. J. Gschwender, *Lubri. Eng.* **42** (1986) no. 9, 547–557.
28. L. J. Gschwender, C. E. Snyder, *Lubri. Eng.* **40** (1984) no. 11, 659–663.
29. L. J. Gschwender, C. E. Snyder, *Lubri. Eng.* **42** (1986) no. 8, 485–490.
30. B. J. Wiggins, H. L. Blackford, *Lubri. Eng.* **43** 1987) no. 6, 467–472.
31. H. Ebertshäuser, *ölhydr. Pneum.* **28** (1984) no. 6, 387–392.
32. H. A. Spikes, *J. Synth. Lubr.* **4** (1987) no. 2, 115–135.
33. R. Taylor, Y. Y. Wang, *Lubri. Eng.* **40** (1985) no. 1, 44–50.
34. K. Bauer, W. Guse, P. Hermann, *Schmiertech. Tribol.* **30** (1983) no. 6, 322–325.
35. H. Singer, A. Kaljee, *S. Afr. Mech. Eng.* **36** (1986) no. 4, 106–107, 110–111.
36. J. T. Schmiede, S. Simandri, A. J. Clark, *Lubri. Eng.* **41** (1985) no. 8, 463–469.
37. R. Stein, *VGB Kraftwerkstechnik* **65** (1985) no. 5, 507–513.
38. W. D. Phillips, *Lubri. Eng.* **44** (1988) no. 9, 758–767.
39. K. Goebels, C. J. Harrington, *Mineralöltechnik* **28** (1983) no. 7, 1–33.
40. A. Sasaki, M. Sasaoka, T. Tobisu, S. Uchiyana *et al.*, *Lubri. Eng.* **44** (1988) no. 3, 251–256.
41. *Ind. Lubr. Tribol.* **38** (1986) no. 5, 180–187.
42. R. M. Verdegan, T. Jaroszczyk, J. A. Stinson, *Lubri. Eng.* **44** (1988) no. 5, 424–430.
43. J. P. Cossette, R. Oliver, J. P. Crine, *Lubri. Eng.* **44** (1988) no. 7, 617–621.
44. P. Wüsthoff, E. Eisenbacher: “Verunreinigungen in den Hydraulikflüssigkeiten und ihr Einfluß auf die Konstruktion,” in W. J. Bartz (ed.): *Hydraulikflüssigkeiten—Reihe Tribotechnik*, vol. **1**. Vincentz-Verlag, Hannover 1982, 59–69.
45. B. Sinclair, *Hydraul. and Air. Engng.* **15** (1987) no. 2, 4–5.

46. L. P. Cote, *Lubri. Eng.* **42** (1986) no. 11, 650–657.
47. W. Ruf, *ölhydr. Pneum.* **31** (1987) no. 4, 305–308.
48. H. Nikolaus, *ölhydr. Pneum.* **26** (1982) no. 1, 22–23.
49. L. O. Leugner,*Lubri. Eng.* **43** (1987) no. 5, 365–369.
50. F. Eckhardt, *ölhydr. Pneum.* **24** (1980) no. 5, 358–364.
51. H. Sötje, *ölhydr. Pneum.* **30** (1986) no. 7, 532–535.
52. Vereinigung Deutscher Elektrizitätswerke—VDEW e.V. (ed.): *Ölbuch Part 1—Schmierstoffe und Steuerflüssigkeiten im Kraftwerksbetrieb.* Verlags- u. Wirtsch.-Ges. d. Elektrizitätswerke, Frankfurt 1982, pp. 33–54.
53. A. Hänssler, *ölhydr. Pneum.* **26** (1982) no. 6, 468–469.
54. J. Görlich, *ölhydr. Pneum.* **31** (1987) no. 3, 206–213.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Hydrazine

Jean-Pierre Schirmann, Paris, France

Paul Bourdauducq, ATOFINA, Pierre-Bénite, France

1.	**Introduction**	1
2.	**Physical Properties**	1
3.	**Chemical Properties**	2
4.	**Production**	3
4.1.	**Raschig Process**	5
4.2.	**Olin Raschig Process**	5
4.3.	**Urea Process**	6
4.4.	**Bayer Ketazine Process**	6
4.5.	**Fisons Process**	7
4.6.	**Peroxide Process**	7
5.	**Environmental Protection**	8
6.	**Quality Specifications**	8
7.	**Analysis**	8
8.	**Handling, Storage, and Transportation**	9
9.	**Uses**	10
10.	**Derivatives**	14
11.	**Economic Aspects**	16
12.	**Toxicology and Occupational Health**	17
13.	**References**	17

1. Introduction

The existence of hydrazine [*302-01-2*], $H_2N{-}NH_2$, M_r 32.05, was predicted by Emil Fischer in 1875 [12], and it was first isolated in 1887 by Curtius [13]. Anhydrous hydrazine was isolated in 1893 by de Bruyn [14]. The first commercial production process was invented by Raschig in 1907 [15]; it is still in use in Japan, Russia, China, and Korea. Following the rapid increase in the use of hydrazine and its derivatives as blowing agents for plastic foams came other industrial applications: boiler water treatment, polymerization initiators, pesticides, pharmaceuticals, photographic chemicals, and dyes.

A century after its discovery, hydrazine is still difficult to synthesize, mainly for thermodynamic reasons. Most hydrazine is produced by variations of the Raschig process, the oxidation of ammonia by hypochlorite. However, the new plants built since 1980 are based on the PCUK process, which uses hydrogen peroxide as oxidant.

Most hydrazine is sold as an aqueous solution of up to 64 % concentration, corresponding to hydrazine hydrate [*7803-57-8*], $N_2H_4 \cdot H_2O$.

2. Physical Properties

Hydrazine is a colorless liquid with an ammoniacal odor. It is miscible with water in all proportions, and its aqueous solutions are highly alkaline. Some physical properties of hydrazine and its aqueous solutions are listed in Table 1. Certain physical properties of the aqueous solutions, e.g., viscosity and density, display a maximum value at the 64 % composition (corresponding to the monohydrate), suggesting that the hydrate, $N_2H_4 \cdot H_2O$, exists in both the solid and the liquid phase (Figure 1). Hydrazine forms an azeotrope (*bp* 120.5 °C) with water, containing 58.5 mol % hydrazine.

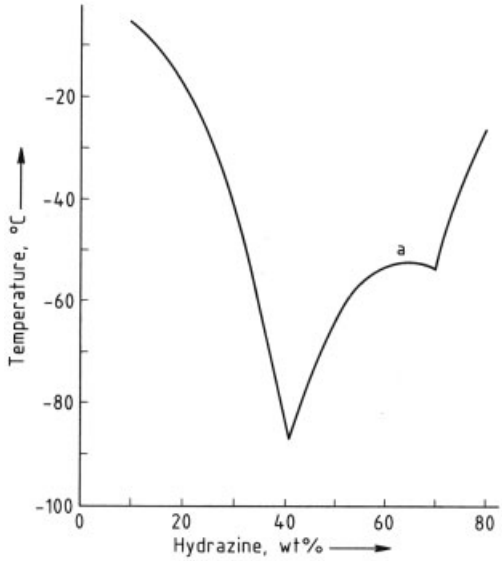

Figure 1. Freezing point of aqueous hydrazine solutions
a) Monohydrate, $NH_2NH_2 \cdot H_2O$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_177

Table 1. Physical properties of hydrazine and its aqueous solutions

Property	Hydrazine concentration, wt %					
	100	64	51.2	35.2	22.4	15.4
fp, °C	2.0	−51.7	−59.8	−64.6	−26	−14
bp (101.3 kPa), °C	113.5	120.5	117.2	108	107	103
ϱ (25 °C), g/mL	1.0045	1.0320	1.0281	1.0209	1.0132	1.0083
n_D^{25}	1.4644	1.4284	1.4120	1.3888	1.3690	1.3575
Viscosity (20 °C), μPa · s	0.974	1.50	1.44	1.10	1.08	1.04
pH		12.75	12.10			10.5
surface tension (25 °C), mN/m	66.7	74.0				
dielectric constant (25 °C)	51.7					

Some thermodynamic properties of anhydrous hydrazine are listed in Table 2 [16–20].

Table 2. Thermodynamic properties of hydrazine [16–20]

Property	Value
Critical constants	
P_c, MPa	14.69
T_c, °C	380
d_c, g/mL	0.231
Heat of vaporization, kJ/mol	45.27
Heat of fusion, kJ/mol	12.66
Heat capacity (25 °C), J mol^{-1} K^{-1}	98.87
Heat of combustion, kJ/mol	−622.1
Heat of formation, kJ/mol	50.63
Free energy of formation, kJ/mol	149.2
Entropy of formation, J mol^{-1} K^{-1}	121
Flash point (COC), ∗ °C	52

∗ Cleveland open cup.

Hydrazine is an endothermic compound with a heat of formation of +50.6 kJ/mol. The explosion limits in air are 4.7 – 100 %. The upper value indicates that anhydrous hydrazine is self-explosive. Dilution with an inert gas such as nitrogen or water significantly reduces the flammable domain by raising the lower explosion limit [21]. Hydrazine hydrate (30.9 vol% hydrazine) can therefore be handled without danger at atmospheric pressure at 120 °C in the absence of air.

3. Chemical Properties

The chemical properties of hydrazine are strongly influenced by the following characteristics: the compound is endothermic, a base, and a reducing agent.

Thermal Decomposition. A relatively high temperature (250 °C) is required, in the absence of catalysts, for significant decomposition to occur [17, 22]:

$$3\,N_2H_4 \longrightarrow N_2 + 4\,NH_3 \quad \Delta H = -157\ \text{kJ/mol}$$
$$N_2H_4 \longrightarrow N_2 + 2\,H_2 \quad \Delta H = -95.4\ \text{kJ/mol}$$
$$N_2H_4 + H_2 \longrightarrow 2\,NH_3 \quad \Delta H = -187.8\ \text{kJ/mol}$$

The decomposition temperature is lowered by several catalysts (e.g., copper, cobalt, molybdenum, and their oxides) [17]. Hence, hydrazine should be handled carefully.

Acid – Base Reactions. Hydrazine is a weak base that reacts with water:

$$N_2H_4 + H_2O \longrightarrow N_2H_5^+ + OH^- \quad K_b = 8.5 \times 10^{-7}\ (25\,°C)$$
$$N_2H_5^+ + H_2O \longrightarrow N_2H_6^{2+} + OH^- \quad K_b = 8.9 \times 10^{-16}\ (25\,°C)$$

The cation $N_2H_6^{2+}$ occurs only in strongly acidic solutions or in the solid state [23].

Hydrazine forms salts with acids [17], some of which are explosive, e.g., the nitrate, perchlorate, and azide. Other salts, such as the hydrochloride, hydrobromide, or sulfate, are commercially available and can be handled in the same way as hydrazine hydrate.

Reducing Agent. Hydrazine is a strong reducing agent which reacts exothermically with oxygen:

$$N_2H_4 + O_2 \longrightarrow N_2 + 2\,H_2O \quad \Delta H = -621\ \text{kJ/mol}$$

Many of the uses of hydrazine are based on this reaction (see Chap. 9). Several metals catalyze the oxidation of hydrazine by air in alkaline solution. For this reason, copper and poly-

valent metals or their salts must be absent or deactivated when hydrazine solutions are distilled [24,25]. The oxides of cadmium, magnesium, zinc, and aluminum stabilize hydrazine solutions against aerial oxidation [26,27].

In acid solution hydrazine reacts with halogens [28,29]:

$$N_2H_4 + 2\,X_2 \longrightarrow N_2 + 4\,HX$$

where X = F, Cl, Br, I

These reactions are used to determine N_2H_4 (with iodine), to purify crude hydrogen halides, and to remove traces of halogens in wastewater. Traces of hydrazine may be removed by the same procedure. For waste or spills it is more convenient to use sodium hypochlorite:

$$N_2H_4 + 2\,NaClO \longrightarrow N_2 + 2\,NaCl + 2\,H_2O$$

or hydrogen peroxide in the presence of iron(III) or copper(II) salts:

$$N_2H_4 + 2\,H_2O_2 \longrightarrow N_2 + 4\,H_2O$$

Various metal ions or oxides, such as those of copper, silver, gold, mercury, nickel, and platinum, can also be reduced to pulverulent metals by hydrazine [30,31].

Ketones and aldehydes are reduced by hydrazine (the Wolff – Kishner reaction) [32]:

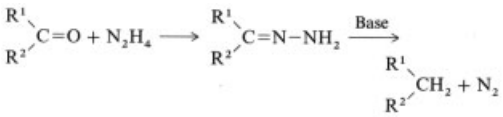

In the presence of a hydrogenation catalyst, such as Raney nickel, aromatic nitro compounds are reduced to the corresponding amines [33]:

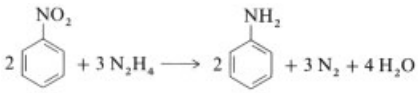

In the presence of hydrogen peroxide, hydrazine is oxidized to diimide [*3618-05-1*], which reduces acetylenes to *cis*-alkenes:

$$N_2H_4 + H_2O_2 \longrightarrow NH{=}NH + 2\,H_2O$$

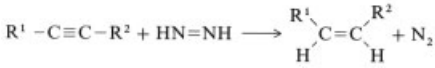

and hydrogenates residual double bonds in acrylonitrile – butadiene rubber [34].

Diamine Reactions. Hydrazine is widely used in the synthesis and production of numerous open-chain and heterocyclic nitrogen compounds, including hydrazo and azo compounds, pyrazoles, triazoles, urazoles, tetrazoles, pyridazines, and triazines [31].

4. Production

Availability of raw materials and production costs rule out most of the possible routes to hydrazine; nitrogen and ammonia are the only obvious starting materials for a reasonably direct process.

Consideration of the variation of standard free energy ΔF°(g) (298 K) for the gaseous system H_2–N_2–NH_3–N_2H_4 (Fig. 2, see next page) indicates that the direct synthesis of hydrazine from nitrogen and hydrogen is energetically unfavorable. The free energy of formation is clearly much more favorable for production of ammonia.

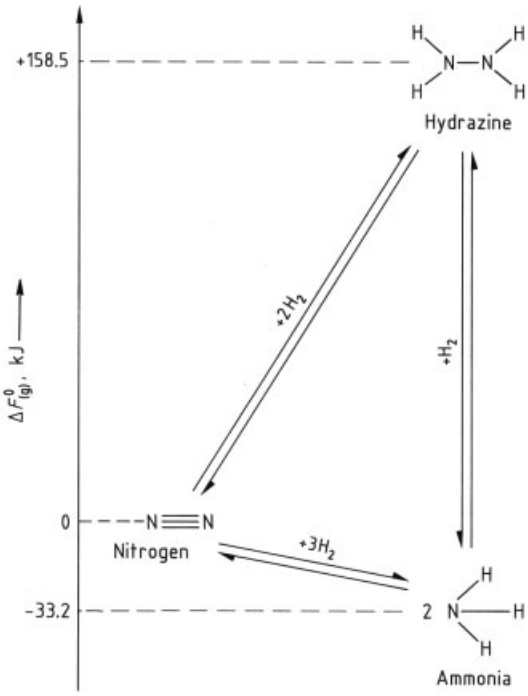

Figure 2. Variation of standard free energy (298 K) in the gaseous system H_2 – N_2 – NH_3 – N_2H_4

The search for selective reduction of nitrogen has not yet found a practical or economic solution. Therefore, ammonia remains the only valuable nitrogen starting material for production of hydrazine.

Coupling of two molecules of NH_3 with coproduction of hydrogen also appears, on paper, to be an attractive process. Such a reaction is, however, endothermic and highly inefficient. For example, decomposition of ammonia by an electric discharge, photolysis, or radiolysis gives only low yields of hydrazine. An alternative method is to oxidize the hydrogen atoms removed from the ammonia:

$$2\,NH_3 \xrightarrow{[o]} N_2H_4 + H_2O$$

Only three oxidants are relevant to an industrial process: chlorine, oxygen, and hydrogen peroxide.

A further difficulty is that hydrazine, which is a much more powerful reducing agent than ammonia, may also react with the oxidizing agent.

Chlorine has been widely used in the *Raschig process*, which is still operated. To avoid further oxidation of hydrazine by chlorine, very dilute conditions have to be employed. Yields are no higher than 60 %.

The use of air or oxygen as a clean oxidizing agent is hardly feasible. This process, discovered in the 1950s by Meyer et al. [35] and later studied extensively by Hayashi [39] can only be applied to a few aromatic imines that lead to aromatic azines, from which hydrazine can only be obtained as the sulfate:

$$(C_6H_5)_2C{=}O + NH_3 \xrightarrow[400\,^\circ C]{ThO_2} (C_6H_5)_2C{=}NH + H_2O$$

$$2\,(C_6H_5)_2C{=}NH + \tfrac{1}{2}O_2 \xrightarrow[Py]{Cu(I)} (C_6H_5)_2C{=}N{-}N{=}C(C_6H_5)_2 + H_2O$$

The only modern method consists of obtaining hydrazine as a derivative from which it may be easily and efficiently released. Azines of low molecular mass are suitable for such a purpose [17, 31]:

$$N_2H_4 + 2\,R_2C{=}O \rightleftharpoons R_2C{=}N{-}N{=}CR_2 + 2\,H_2O$$

The azines of acetone and methyl ethyl ketone are easily hydrolyzed under pressure forming hydrazine and regenerating the ketone [36–38]. Bayer has considerably improved yields by introducing acetone into the Raschig process.

In the 1970s, PCUK, (now Atochem) developed a new, efficient, and clean process based on the oxidation of ammonia by hydrogen peroxide in the presence of a ketone. Most hydrazine is now produced by the *ketazine process*, with oxidation of ammonia by chlorine or hydrogen peroxide.

4.1. Raschig Process

In the Raschig process [14, 40–42] sodium hypochlorite (obtained by reaction of chlorine with sodium hydroxide) is used to oxidize ammonia.

Two steps are involved in the oxidation (Fig. 3). In the first, carried out at ca. 5 °C, chloramine [*10599-90-3*] is formed by a fast reaction:

$$NaClO + NH_3 \longrightarrow NH_2Cl + NaOH$$

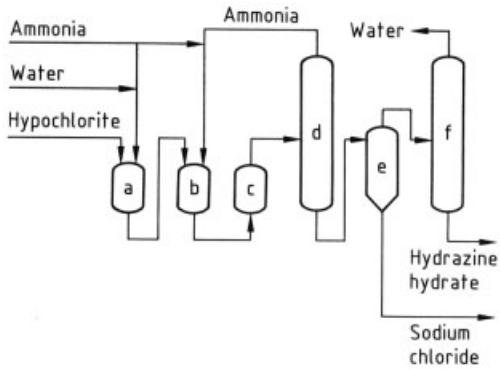

Figure 3. Raschig process for the production of hydrazine
a) Chloramine reactor; b), c) Hydrazine reactors; d) Ammonia evaporator; e) Hydrazine – sodium chloride separator; f) Hydrazine hydrate concentrator

The reaction mixture is then mixed with a large molar excess of ammonia (40 : 1) and the slow reaction leading to hydrazine is carried out at 130 – 150 °C and 3.0 MPa:

$$NH_2Cl + NH_3 + NaOH \longrightarrow N_2H_4 + NaCl + H_2O$$

The kinetics and mechanism [43–46] and the influence of various parameters such as temperature, pressure, concentration, and molar ratios have been examined [31].

The most important side reactions are the following:

$$2\,NH_2Cl + NaClO + 2\,NaOH \longrightarrow N_2 + 3\,NaCl + 3\,H_2O$$
$$2\,NH_2Cl + N_2H_4 \longrightarrow N_2 + 2\,NH_4Cl$$

At the outlet of the reactor, the reaction liquor contains 1 % hydrazine hydrate and ca. 4 % sodium chloride; the pressure is reduced to atmospheric in a battery of evaporators. Ammonia is condensed, concentrated, and recycled. The liquor from the bottom of the stripping columns is freed from salt in a conventional forced–circulation salting evaporator. The distillate is then concentrated to 100 % hydrazine hydrate.

For such a process, dilute solutions and a very high ammonia/hypochlorite ratio are essential in order to obtain reasonable yields. Evaporation costs are therefore high even when steam savings are realized.

4.2. Olin Raschig Process

In the Olin Raschig process (Fig. 4), which is used by Olin to produce anhydrous hydrazine for aerospace applications, the production of sodium hypochlorite is carefully controlled. A low temperature is used to prevent decomposition and chlorate formation, and the excess of sodium hydroxide is kept at a low level.

The sodium hypochlorite solution is mixed with a threefold excess of ammonia at 5 °C to form chloramine, which is then rapidly added to a 30-fold molar excess of anhydrous ammonia under pressure (20 – 30 MPa) and heated to 130 °C [47–49].

The reaction liquor, containing 1 – 2 % hydrazine hydrate, is treated as in the conventional Raschig process to give hydrazine hydrate. Anhydrous hydrazine is obtained by removing the water by azeotropic distillation with aniline in a column at atmospheric pressure. Condensation of the vapor yields a two-layer distillate; the aqueous phase is removed and the aniline phase refluxed to the top of the column. Anhydrous hydrazine is recovered as a mixture with aniline, from which it is separated by distillation [50].

4.3. Urea Process

Urea may be used as a source of ammonia in the Raschig process [51, 52]. Although not currently used for the production of commercial hydrazine hydrate, this process has been operated commercially and it is described in [53]. Compared with the standard Raschig process, it was the most economical method for low production levels, but with the rapid growth in plant size it became obsolete.

However, since 1990, large quantities of hydrazodicarbonamide are produced in Asia, by using the crude reaction mixture of urea and sodium hypochlorite as the source of hydrazine (see page 10). Because this process involves extensive formation of byproducts and expensive effluent treatment, it is likely to become obsolete.

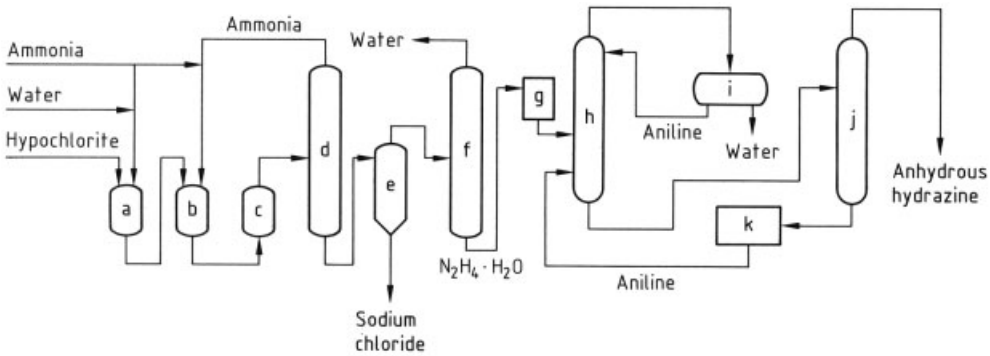

Figure 4. Olin Raschig process
a) Chloramine reactor; b), c) Hydrazine reactors; d) Ammonia evaporator; e) Hydrazine – sodium chloride separator; f) Hydrazine hydrate concentration; g) Hydrazine hydrate storage; h) Hydrazine hydrate dehydration; i) Aniline – water decantation; j) Anhydrous hydrazine distillation; k) Aniline storage

4.4. Bayer Ketazine Process

The Bayer process (Fig. 5) is a variation of the Raschig process and is based on the reaction of chloramine with ammonia in the presence of acetone at pH 12 – 14 [54, 55]:

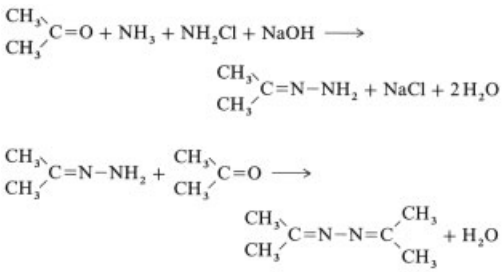

Sodium hypochlorite, acetone, and a 20 % aqueous solution of ammonia (molar ratio 1: 2: 20, respectively) are fed simultaneously and continuously into a reactor at ca. 35 °C and 200 kPa. Excess ammonia is removed from the reaction mixture by stripping, quenched with water, and recycled to the reactor as an aqueous solution. The aqueous dimethyl ketazine solution, freed from ammonia but containing unconverted acetone, sodium chloride, and organic impurities, is fed into a distillation column where the dimethyl ketazine is recovered as an aqueous azeotrope (containing 55 % dimethyl ketazine; *bp* 95 °C at 101.3 kPa) at atmospheric pressure. The injection of acetone into the distillation column is claimed to prevent premature hydrolysis of the ketazine. The byproduct from the still is a solution of sodium chloride containing traces of hydrazine and organic compounds. The solution must be treated before disposal or recycling to electrolysis.

The dimethyl ketazine is then hydrolyzed in a distillation column under pressure (0.8 – 1.2 MPa), giving acetone, which is recycled to the reactor, and a 10 % aqueous solution of hydrazine. The latter is then concentrated to a hydrazine content of 64 %.

4.5. Fisons Process

This process, originally operated by Whiffen and Sons, was developed by Fisons and went on stream in the 1960s. It used methyl ethyl ketone instead of acetone to trap the hydrazine; the hydrolysis was carried out with sulfuric acid.

This process is no longer operated.

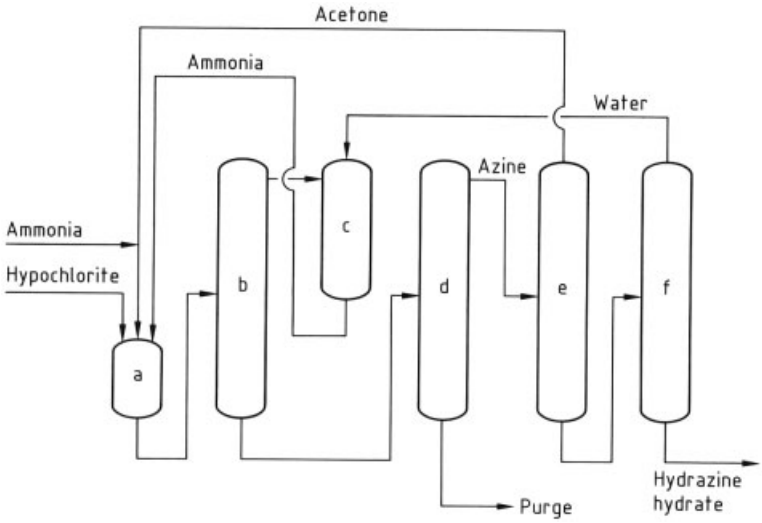

Figure 5. Bayer hydrazine process
a) Azine reactor; b) Ammonia stripping; c) Ammonia quenching; d) Azine distillation; e) Azine hydrolysis; f) Hydrazine hydrate concentration

4.6. Peroxide Process

The most recent hydrazine process, invented by PCUK, is operated by ATOFINA in France and by Mitsubishi Gas in Japan. Hydrogen peroxide is the oxidizing agent. The reaction is carried out in the presence of methyl ethyl ketone (MEK) at atmospheric pressure and 50 °C. The ratio of H_2O_2: MEK: NH_3 used is 1: 2: 4.

The hydrogen peroxide is activated by acetamide [*60-35-5*] and disodium hydrogen phosphate [*7558-79-4*] (ATOFINA) or by an arsenic compound (Mitsubishi Gas) [56]. The overall reaction results in formation of methyl ethyl ketazine [*5921-54-0*] in high yield [57] (Fig. 6):

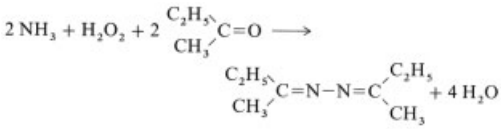

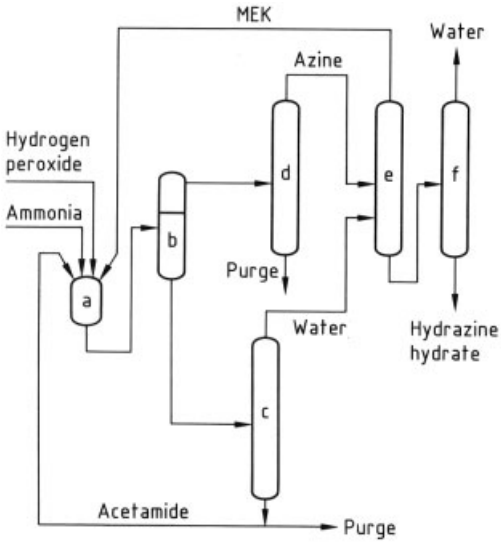

Figure 6. Peroxide process
a) Azine reactor; b) Phase separator; c) Aqueous phase concentration; d) Azine purification; e) Azine hydrolysis; f) Hydrazine hydrate concentration

The mechanism requires the activation of ammonia and hydrogen peroxide as these two reactants, unlike ammonia and hypochlorite in the Bayer process, do not react together [58–62]. The reaction pathway involves the formation of an oxaziridine intermediate that is able to oxidize ammonia to a hydrazine derivative.

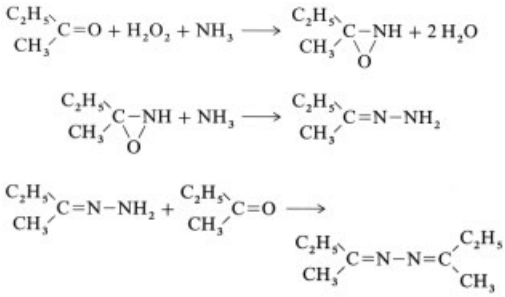

Since methyl ethyl ketazine is insoluble in the reaction mixture, it is easily separated by decantation; it is then purified by distillation. The purified ketazine is hydrolyzed under pressure (0.8 – 10 MPa) to give concentrated aqueous hydrazine and overhead methyl ethyl ketone, which is recycled [37].

The aqueous layer containing the activator is concentrated to remove water and recycled to the reactors after a purge of water-soluble impurities.

The peroxide process has many advantages compared with other processes: no salt byproduct, high yields, low energy consumption, low molar excess, no aqueous effluent treatment [63].

5. Environmental Protection

Hydrazine has a noxious effect on bacteria, algae, and aquatic wildlife; therefore, emission of hydrazine-containing wastewater is not permitted.

Wastewater and spills that contain hydrazine must be collected, analyzed, and treated (e.g., by oxidation with NaClO or H_2O_2). The ammonia content should also be examined because hydrazine may also decompose into ammonia.

6. Quality Specifications

Hydrazine is commercially available as anhydrous hydrazine, as an aqueous solution, and as solid dihydrazinium sulfate. Some typical specifications are summarized in Tables 3 and 4 (both see next page).

Table 3. Specifications of anhydrous hydrazine (monopropellant grade)

Component	Specification
Hydrazine, wt %	≥98.5
Water, wt %	≤1.0
Chloride, ppm	≤5.0
Iron, ppm	4
Aniline, wt %	≤0.5
Nonvolatile residue, wt %	≤0.005
Carbon dioxide, wt %	≤0.003
Carbon volatile, wt %	≤0.02
Density (25 °C), g/mL	1.008 – 1.002

Table 4. Specifications of hydrazine hydrate

Component	Specification
Hydrazine hydrate, wt %	>100±0.5
Hydrazine, wt %	>64±0.4
Ammonia, wt %	<0.1
Chloride, ppm	<1
Iron, ppm	<1
Nonvolatile residue, ppm	<20

Although most hydrazine hydrate is manufactured by the ketazine process, there are no specifications relating to organic impurities.

7. Analysis

Analytical determination of hydrazine in solutions or mixtures is based on its properties as a weak base and strong reducing agent, and on the formation of highly colored derivatives.

A large number of methods have been described for hydrazine derivatives [64]; in this chapter, only analytical methods for hydrazine and hydrazine hydrate will be discussed.

Hydrazine is a weak base ($pK_b = 6.07$) and, in the absence of other bases, can be easily titrated with acid with methyl purple as the indicator. In the absence of other reducing agents, hydrazine can be determined quantitatively by oxidation, most commonly with iodine [65]:

$$N_2H_4 + 2\,I_2 \longrightarrow N_2 + 4\,HI$$

To obtain a stable end point, the pH must be maintained near neutrality, usually by adding solid sodium bicarbonate or ammonium acetate as a buffer. Starch is used as an indicator.

Traces of hydrazine (as low as 1 ppb) can be detected with an iodide-selective electrode [66].

Dilute solutions of hydrazine can be determined by colorimetry. The best method (ASTM D 1385 – 78) involves converting the hydrazine into highly colored azines, such as those of aromatic aldehydes; *p*-dimethylaminobenzaldehyde is particularly suitable for this purpose.

8. Handling, Storage, and Transportation

Hydrazine is volatile and toxic (see Chap. 12) and must be handled with care. Observation of elementary safety rules allows hydrazine to be used with little risk. In the absence of decomposition catalysts, anhydrous hydrazine may be heated up to 250 °C without appreciable decomposition. Aqueous and anhydrous hydrazine can be stored for long periods without problems; clean storage vessels of suitable material are required, together with an inert gas blanket, e.g., nitrogen.

Hydrazine vapors do, however, present hazards in gas mixtures, as shown in Table 5. Working in the vapor phase requires operation below the lower explosion limit.

Table 5. Lower explosion limits of gaseous hydrazine mixtures

Gas mixture	Hydrazine, %	Temperature, °C	Pressure, kPa
Hydrazine	100	30	2
Hydrazine – air	54.7	95	100
Hydrazine – nitrogen	38	110	100
Hydrazine – water	38	100	30

Hydrazine can be stored in drums of 304 stainless steel at room temperature, but stainless steel 316 must be avoided because of its high content of molybdenum, which is an efficient catalyst for the decomposition of hydrazine. Cold-rolled steel is satisfactory for hydrazine concentrations below 10 %. Other materials convenient for storage of anhydrous hydrazine include aluminum, tantalum, and titanium. Polytetrafluoroethylene, polyethylene, and polypropylene are also suitable, but poly(vinyl chloride) is not recommended.

Hydrazine must generally be stored under conditions that preclude oxidants. Anhydrous hydrazine may cause ignition of wood, rags, paper, or other common organic material. Fires can be extinguished easily, however, as hydrazine is completely miscible with water. Contact with

carbon dioxide must also be prevented, as this gas reacts readily to give a hydrazinium carbazic acid salt, which is quite unstable. Storage in direct sunlight should also be avoided.

Hydrazine is easily adsorbed orally, percutaneously, and by inhalation. Proper protection must be provided. This includes well-aired premises, gloves, glasses, appropriate masks, and boots. Skin or eye contact must be treated by washing copiously with water; a physician must be consulted. Premises where hydrazine is handled must be equipped with continuously operated analytical devices to ensure that the concentration in air does not exceed 0.1 ppm in Europe and 0.01 ppm in the United States.

Anhydrous hydrazine and its aqueous solutions are subject to national and international regulations relating to the transportation of hazardous materials, e.g., for transportation by rail (RID) and by road (ADR) in Europe, and the international regulations for transportation by sea (IMCO code) and air (IATA).

Anhydrous hydrazine (UN no. 2029) and hydrazine hydrate (UN no. 2030) aree classified (Class 8) as flammable liquids and poisous. Dilute aqueous solutions of hydrazine hydrate belong only to the Class 6 but are still considered as corrosive and toxic.

9. Uses

Most hydrazine is sold as an aqueous solution. The only outlet for anhydrous hydrazine is as a rocket fuel or as a mono- or bipropellant for satellites and spacecraft.

About 80 – 90 % of hydrazine production is converted into organic derivatives. All other applications are based on its use as a reducing agent, as an energy-rich compound, or on its hydrogen storage capacity.

The most important uses of hydrazine and its derivatives are as polymerization initiators and blowing agents for foamed plastics, and for the production of pesticides. Other uses include: synthetic building block, pharmaceuticals, propellants, and airbags for cars.

Blowing Agents. Many hydrazine-based blowing agents are produced industrially; world hydrazine hydrate consumption in 1998 for this application amounted to 50 000 t/a. They are all hydrazo or azo derivatives; the latter are obtained by oxidation of the former with chlorine or hydrogen peroxide. Blowing agents decompose on heating into nitrogen and mixtures of other gases thereby producing a foaming action in polymers to form pores or cells (→ Foamed Plastics). Some commercially produced hydrazine-based blowing agents are listed in Table 6.

The decomposition temperature of blowing agents depends on the particle size, the pH, and the presence of salts of barium, cadmium, or zinc, which act as activating agents.

One of the earliest commercial blowing agents was azobis(isobutyronitrile) (AIBN), which is used as a porophore for sponge rubber products and PVC foams. It is also em-

Table 6. Some hydrazine-based blowing agents

Compound	CAS registry number	Structure	Decomposition temperature in air, °C	Gas yield, mL/g
Azobis(isobutyronitrile)	[*78-67-1*]	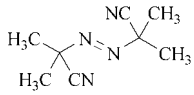	115	130
Azodicarbonamide	[*123-77-3*]	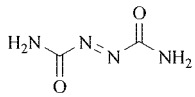	195 – 200	220
Benzenesulfonylhydrazide	[*80-17-1*]	$C_6H_5SO_2NHNH_2$	95	130
4-Methylbenzenesulfonylhydrazide	[*1576-35-8*]	p-$CH_3C_6H_4SO_2NHNH_2$	103	120
4,4′-Oxybis(benzenesulfonylhydrazide)	[*80-51-3*]	$O<-(>C_6H_4SO_2NHNH_2)_2$	150	125

ployed as a free radical source in the initiation of polymerization. Azobis(isobutyronitrile) is produced from acetone cyanohydrin [*75-86-5*] and hydrazine:

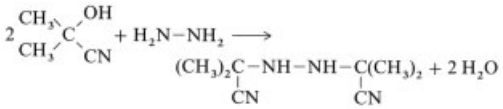

Oxidation to the azo compound is effected with chlorine [67, 68]:

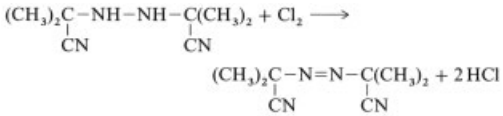

Azodicarbonamide is readily produced from urea and hydrazine:

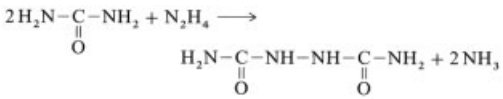

The intermediate hydrazodicarbonamide, which can also be produced directly in situ from urea and sodium hypochlorite [69], is oxidized either by chlorine or hydrogen peroxide in the presence of a catalytic amount of bromide ions in a strongly acidic medium:

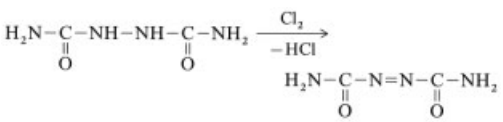

Azodicarbonamide is, in terms of production volume, the most important blowing agent. Its success is due to the large volume of gas evolved (220 mL/g at STP). Decomposition at 190 °C yields mainly nitrogen (60 vol %) and carbon monoxide (35 vol %), as well as small amounts of ammonia and carbon dioxide. The solid residues are colorless, odorless, nonstaining, and do not support combustion.

The need for blowing agents with a higher decomposition temperature for the manufacture of new rubber or porous plastic materials led to the appearance of sulfonic acid hydrazides in the 1950s. In contrast to the inorganic blowing agents previously used (e.g., ammonium nitrite, ammonium bicarbonate, sodium bicarbonate) these sulfonic acid hydrazides are more easily dispersed, can be safely handled at a higher temperature, and give improved foam cell structure. They are also colorless, odorless, nonstaining, and the decomposition products are safe. The following explanation of decomposition has been proposed:

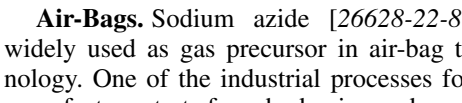

The sulfonic acid mono- and dihydrazides are produced from hydrazine hydrate and the appropriate sulfonic acid chloride:

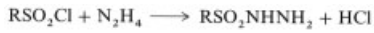

The handling and technology of blowing agents have been reviewed [70, 71].

Air-Bags. Sodium azide [*26628-22-8*] is widely used as gas precursor in air-bag technology. One of the industrial processes for its manufacture starts from hydrazine and an alkyl nitrite [72].

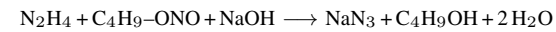

An other hydrazine derivative is under development for the same application: 5-aminotetrazole [*4418-61-5*] obtained from aminoguanidine salts [73].

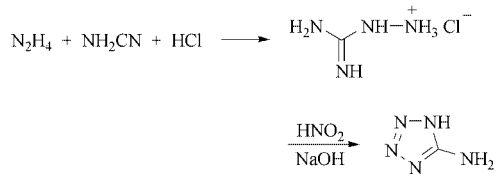

Free-Radical Polymerization Initiators. Azo compounds are widely used as free radical polymerization initiators. The most important (Table 7) are symmetrical azodinitriles which are synthetized from hydrazine, a ketone and HCN

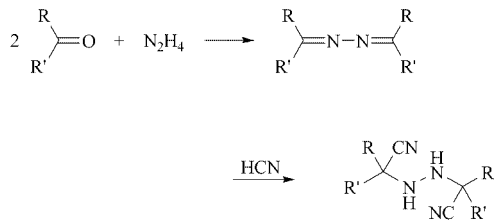

Table 7. Some azo polymerization initiators

Compound	CAS registry no.	Structure	*mp* °C	10-h half-life decomposition temperature in toluene, °C
2,2′-azobisisobutyronitrile	[*78-67-1*]	R=R′=CH_3	ca. 100	65
2,2′-azobis(2-methylbutyronitrile)	[*13472-08-7*]	R=CH_3,R′=C_2H_5	ca. 50	67
2,2′-azobis(2,4-dimethylvaleronitrile)	[*4419-11-8*]	R=CH_3,R′=CH_2–CH–$(CH_3)_2$	45 – 70	51

The hydrazo derivative is oxidized by chlorine or hydrogen peroxide in the presence of a bromide as catalyst:

Water-soluble compounds like 2,2′-azobis(2-aminopropane) dihydrochloride [*2997-92-4*] as well as liquid azo compounds like diethyl 2,2′-azobisisobutyrate [*3879-07-0*], are under development.

Pesticides. Hydrazine-based pesticides represent the second major outlet for consumption of hydrazine. Maleic hydrazide was the first such pesticide used. It is produced by reaction of maleic anhydride with hydrazine:

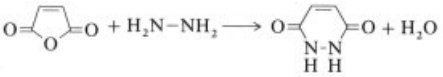

3-Amino-1,2,4-triazole, obtained from cyanamide, hydrazine hydrate, and formic acid, is another general purpose herbicide:

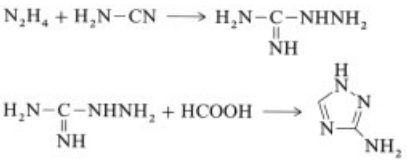

This compound is used as a selective herbicide in vineyards and orchards; consumption is now several thousand tons per year.

Hundreds of patents have been filed for pesticide applications of hydrazine-based compounds; about fifty compounds are produced commercially [74], mostly heterocyclic compounds such as triazines, oxadiazoles, pyrazoles, pyridazines, thiadiazoles. Some representative examples are listed in Table 8.

During the period 1970 – 1980, triazines became important, examples being metribuzin and metamitron (see Table 8). A new family of fungicides based on the triazole ring is now being developed.

Information concerning structures, production, and use of these pesticides is compiled in [75–78].

Pharmaceuticals. Although representing only a small percentage of the total production, consumption of hydrazine in pharmaceuticals is very important. For example, isoniazid [*54-85-3*], the hydrazide of isonicotinic acid, was first used in the 1950s against tuberculosis.

In the 1980s and 1990s, other hydrazine-based pharmaceuticals were introduced, many of these containing the 1,2,4-triazole group. They have been used as antidepressants, antihypertensives, and as antibacterial or antifungal agents.

More recently, new pharmaceuticals, containing the 4-amino-1,2,4-triazole group and showing better efficiency [80, 81] (Table 9), have been developed.

Water Treatment. Sodium sulfite has long been used for scavenging oxygen dissolved in water in order to avoid corrosion in boilers or hot water heating systems:

$$2\,Na_2SO_3 + O_2 \longrightarrow 2\,Na_2SO_4$$

At higher temperatures, sodium sulfite may decompose to corrosive sulfide products. It can also form deposits, which affect heat transfer and may even cause pipes to rupture.

Hydrazine is used in water treatment for corrosion protection of steel in boilers. The only reaction products are nitrogen and water:

$$N_2H_4 + O_2 \longrightarrow N_2 + 2\,H_2O$$

Table 8. Some important hydrazine-based pesticides

Compound	CAS registry no.	Structure	Producer	Use
Triazoles				
3-Amino-1,2,4-triazole	[*61-82-5*]	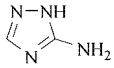	Atochem	herbicide
Triadimefon	[*4321-43-3*]	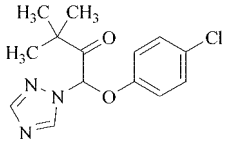	Bayer	fungicide
Propiconazole	[*60207-90-1*]	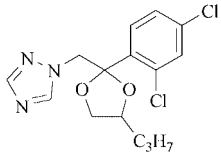	Novartis	fungicide
Diclobutrazole	[*75736-33-3*]	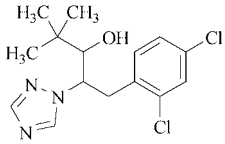	Zeneca	fungicide
Paclobutrazol	[*76738-62-0*]	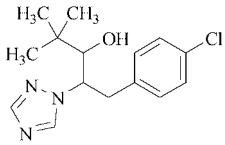	Zeneca	plant growth regulator
Triazines				
Metribuzin	[*21087-64-9*]	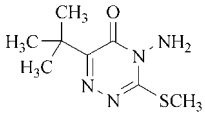	Bayer	herbicide
Metamitron	[*41394-05-2*]	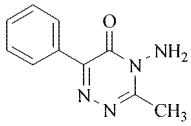	Bayer	herbicide
Pyrazole				
Metazachlor	[*67129-08-2*]	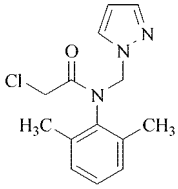	BASF	herbicide
Pyridazine				
Maleic hydrazide	[*123-33-1*]	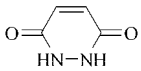	Uniroyal	plant growth regulator

Table 9. Some important hydrazine-based pharmaceuticals

Compound	CAS registry no.	Structure	Producer	Use
Fluconazole	[*86386-73-4*]	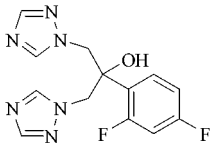	Pfizer	antifungal
Anastrozole	[*120511-73-1*]	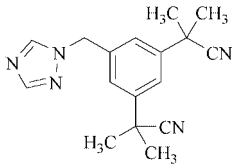	Zeneca	anticancer
Rizatriptan	[*144034-80-0*]	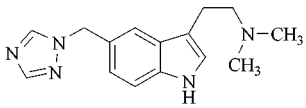	Merck	antimigraine
Cefazolin	[*25953-19-9*]	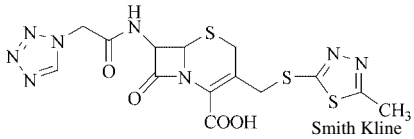	Smith Kline	antibacterial
Isoniazid	[*54-85-3*]	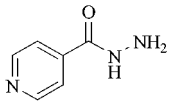	Pfizer	tuberculostat

Moreover hydrazine reacts with iron(III) oxide to form magnetite, which protects the metal surface against corrosion by water and oxygen [82–84]:

$$6\,Fe_2O_3 + N_2H_4 \longrightarrow N_2 + 4\,Fe_3O_4 + 2\,H_2O$$

Full corrosion protection is afforded by residual hydrazine concentrations lower than 0.1 ppm. Catalyzed hydrazine hydrate formulations (activated hydrazine) are commercially available and are efficient even at room temperature.

Propellants. The first large-scale use of hydrazine was as a rocket fuel. Anhydrous hydrazine is an excellent propellant; only hydrogen has a greater specific impulse, i.e., kilograms of thrust developed per kilogram of fuel consumed per second. Rocket propellants now in use include anhydrous hydrazine, monomethylhydrazine and unsymmetrical dimethylhydrazine (see Chap. 10). They are used mainly as bipropellant fuels in rockets such as Titan or Ariane.

Anhydrous hydrazine is also used as a monopropellant for satellites and spacecraft [85]. The hydrazine decomposes in a complex reaction over a catalyst to give a mixture of gases:

$$3\,N_2H_4 \longrightarrow 4(1-x)\,NH_3 + (2x+1)\,N_2 + 6x\,H_2$$

Typical catalysts are manufactured by Shell Chemical and Rocket Research. They are often based on iridium or ruthenium, deposited on alumina. Catalysts are described as spontaneous (working at room temperature) or nonspontaneous (working above 100 °C) [86].

A review on space applications and potential rocket propulsion systems has been published [87].

Fuel Cells. Fuel cells based on oxidation of hydrazine either with oxygen or hydrogen peroxide have been extensively studied. Their use is restricted to military applications because of the cost of hydrazine hydrate.

10. Derivatives

Alkylhydrazines. Monomethylhydrazine [*60-34-4*] MMH, and unsymmetrical dimethylhydrazine, [*57-14-7*], UDMH, are the most important derivatives; their major use is rocket fuels.

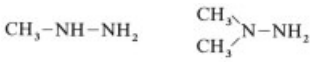

MMH UDMH

Monomethylhydrazine is produced commercially in the Raschig process by reaction of chloramine with methylamine:

whereas UDMH is now produced either by the chloramine route:

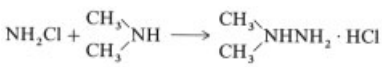

or by the reductive catalytic alkylation of acetic hydrazide [*1068-57-1*] with formaldehyde and hydrogen:

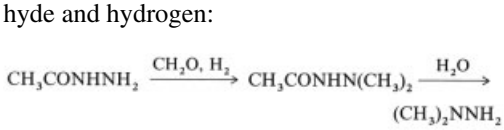

The preparation and properties of alkylhydrazines have been reviewed [88, 89]. Hydroxyethylhydrazine [*109-84-2*], $HOCH_2CH_2NHNH_2$ is obtained by reaction of hydrazine hydrate with ethylene oxide [90]:

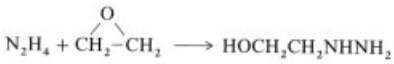

tert-Butylhydrazine hydrochloride [*7400-27-3*] is used in the synthesis of pesticides.

Arylhydrazines. Aromatic hydrazines are not produced from hydrazine, except for 2,4-dinitrophenylhydrazine, which is obtained by reaction of 2,4-dinitrochlorobenzene with hydrazine hydrate. It is used as an analytical reagent for the determination of carbonyl compounds.

Phenyl hydrazines are obtained by diazotization of the corresponding anilines and reduction of the diazonium salts:

Characteristics of some of these compounds are listed in Table 10 [91, 92].

Table 10. Some aromatic hydrazines

Hydrazine	CAS registry no.	*mp*, °C	Appearance
Phenylhydrazine	[*100-63-0*]	19.6	colorless crystals
4-Tolylhydrazine	[*539-44-6*]	61	colorless crystals
3-Nitrophenyl-hydrazine	[*619-27-2*]	153	yellow crystals
2,4-Dinitrophenyl-hydrazine	[*119-26-6*]	200	red crystals
N,N′-Diphenyl-hydrazine	[*122-66-7*]	128	colorless crystals

Hydrazides. Some common hydrazides are listed in Table 11.

Table 11. Some hydrazides

Hydrazide	CAS registry number	Structure	M_r	*mp*, °C
Formic hydrazide	[*624-84-0*]	$HCONHNH_2$	60.06	54
Acetic hydrazide	[*1068-57-1*]	$CH_3CONHNH_2$	74.08	67
Propionic hydrazide	[*5818-15-5*]	$C_2H_5CONHNH_2$	88.12	40
Oxalic dihydrazide	[*996-98-5*]	$(H_2NNHCO)_2$	118.10	240 ∗
Semicarbazide	[*57-56-7*]	$H_2NCONHNH_2$	75.07	96
Thiosemicarbazide	[*79-19-6*]	$H_2NCSNHNH_2$	91.14	180
Carbohydrazide	[*497-18-7*]	$H_2NNHCONHNH_2$	90.09	157
Thiocarbohydrazide	[*2231-57-4*]	$H_2NNHCSNHNH_2$	106.15	171 ∗
Benzenesulfonylhydrazide	[*80-17-1*]	$C_6H_5SO_2NHNH_2$	172.21	103
Adipicdihydrazide	[*1071-93-8*]	$H_2N–NH–CO–(CH_2)_4CONH–NH_2$	174	18

∗ Decomposes

Carboxylic acid hydrazides are generally prepared by reactions of hydrazine hydrate with the corresponding methyl or ethyl ester:

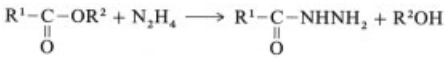

Diacids can lead to polyhydrazides which dehydrate on heating to give polyoxadiazoles, which exhibit high thermal stability. Hydrazine carboxamide and carbohydrazide are produced from hydrazine hydrate and urea, depending on the excess of urea used:

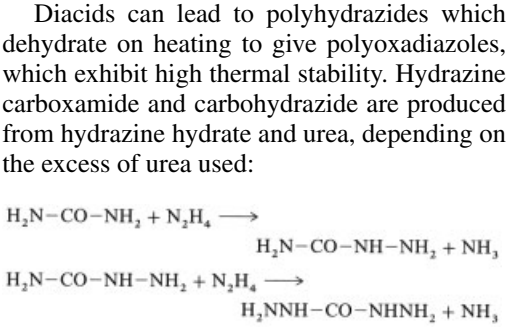

Hydrazine carbothioamide (thiosemicarbazide) is made from CS_2 and hydrazine, and thiocarbohydrazide from CS_2, ammonia, and hydrazine [93–95].

Heterocycles. A few simple heterocycles are used as synthetic building blocks [96]. The most important heterocycle industrially is 1,2,4-triazole, which is produced from hydrazine and formamide [97, 98].

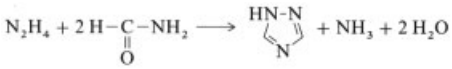

1,2,4-Triazole is a white solid, *mp* 120 °C. It is a valuable building block in the production of many fungicides.

Miscellaneous Derivatives. Different types of derivatives have found application in various industrial fields. For example, in photography, heterocycles such as triazoles, tetrazoles, 2,5-dimercapto-1,3,4-thiadiazole are used as stabilizers, fog inhibitors, or sensitizers.

A number of hydrazine derivatives are useful explosives, including the simple salts such as the nitrate [*37836-27-4*], perchlorate [*13762-80-6*], azide [*14662-04-5*], and aminoguanidine derivatives [85].

Aminoguanidine bicarbonate [*2582-30-1*], manufactured industrially from cyanamide [*420-04-2*] hydrazine, and carbon dioxide [99], is now largely used for the synthesis of various heterocycles useful in the production of pharmaceuticals, agrochemicals, and lube oil additives.

11. Economic Aspects

Considerable growth in hydrazine production capacity took place during the period 1960–2000 (Fig. 7). Estimated capacities for major producers, together with the production process used, are listed in Table 12. Total capacity (western world, 2000) is ca. 90 000 t of hydrazine hydrate. However, to be accurate, one must add to

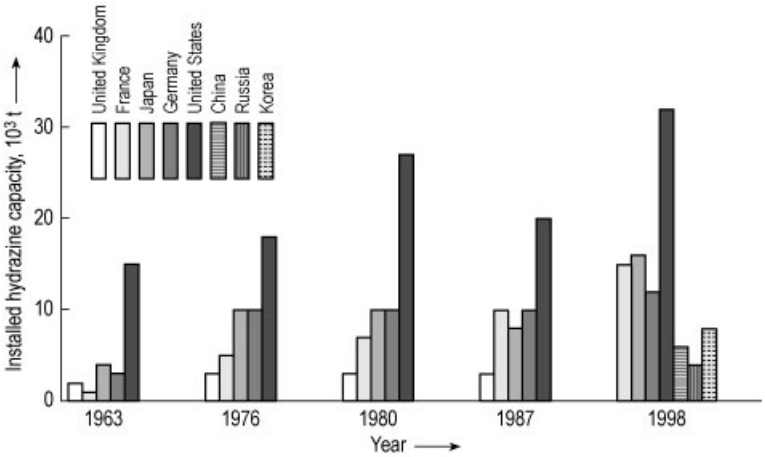

Figure 7. Installed hydrazine capacity for various countries
UK, United Kingdom; US, United States; J, Japan; F, France; G, Germany; C, China; R, Russia; K, Korea

this capacity, the captive production of crude hydrazine hydrate (30 000 t/a) associated with the direct azodicarbonamide manufacture in Asia. Thus the total world capacity of hydrazine hydrate is ca. 120 000 t/a. Approximate consumption of hydrazine according to application is shown in Figure 8.

Table 12. Producers of hydrazine hydrate, estimated capacities, and process used (1998)

Country	Producer	Capacity, 10^3 t	Process
United States	Arch	5	Raschig
		15	Ketazine-Raschig
	Bayer US	12	Ketazine-Raschig
Federal Republic of Germany	Bayer	12	Ketazine-Raschig
France	ATOFINA	15	Peroxide
Japan	Nippon Carbide	3	Raschig
	Otsuka	4	Raschig
	MiGas	9	Peroxide
Korea	Otsuka	8	Raschig
China	Various	6	Raschig
Russia	Kuybichev	4	Raschig

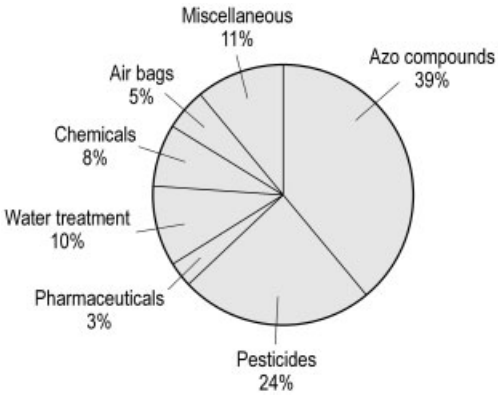

Figure 8. Hydrazine consumption

12. Toxicology and Occupational Health

Hydrazine is highly toxic, even in dilute solution, and must be handled with care. No antidote to hydrazine is known; appropriate treatment must be based on the symptoms observed.

Contact of the liquid with the skin and the eyes should be avoided. Irritation, damage due to alkalinity, dermatitis, or allergy may result. Inhalation may cause irritation or damage to the nose, throat, and lungs. Ingestion may induce convulsions and depression of the central nervous system. Repeated exposure to hydrazine may result in damage to the lungs, liver, or kidneys.

The acute and chronic toxicity of hydrazine and its derivatives have been examined in detail; reviews are available [100–103].

The results of experiments on animals are listed in Table 13. Long-term studies with laboratory animals indicate that hydrazine is mutagenic and carcinogenic; this has not, however, been demonstrated for humans.

Table 13. Acute toxicity of hydrazine

Animal	Nature	LD_{50}, mg/kg
Mouse	intravenous	58
	oral	58
Rat	intravenous	55
	oral	60
Guinea pig	cutaneous	190
Rabbit	intravenous	25
Monkey	intraperitoneal	20

The threshold limit value (TLV) has been fixed at 0.1 ppm (0.13 mg/mm^3). The olfactory detection limit is between 3 and 5 ppm.

13. References

General References

1. L. F. Audrieth and B. A. Ogg: *The Chemistry of Hydrazine,* Wiley, New York 1951.
2. C. C. Clark: *Hydrazine,* Mathieson Chemical Corp., Baltimore, MD 1953.
3. R. A. Reed, *Hydrazine and its Derivatives,* Lectures, Monographs and Reports, no. 5, The Royal Institute of Chemistry, London 1957, 1 – 49.
4. E. W. Schmidt: *Hydrazine and its Derivatives,* J. Wiley Sons, New York 1984.
5. *Houben-Weyl,* **X/2,** 71 – 122.
6. *Kirk-Othmer,* **13,** 560 – 606.
7. *Ullmann,* 5th ed. **13,** 177 – 191.
8. R. Powell: "Hydrazine Manufacturing Processes," *Chem. Process. Review* no. 28, Noyes Development (1968).
9. *Gmelin,* System no. 4, "Stickstoff," 307 – 320.
10. S. Pataï: *The Chemistry of the Hydrazo, Azo, and Azoxy Groups,* Wiley-Interscience, New York 1975.
11. *Beilstein,* **4,** 546; **4 (1),** 650; **4 (2),** 957; **3,** 116; **3 (1),** 56; **3 (2),** 95; **15,** 67, 123; **15 (1),** 23, 28; **15 (2),** 44, 52; **11,** 52; **11 (2),** 29.

Specific References

12. E. Fischer, *Ber. Dtsch. Chem. Ges.* **8** (1875) 589 – 594.
13. Th. Curtius, *Ber. Dtsch. Chem. Ges.* **20** (1887) 1632.
14. L. de Bruyn, *Recl. Trav. Chim. Pays Bas* **13** (1894) 433.
15. F. Raschig, *Ber. Dtsch. Chem. Ges.* **40** (1907) 4587.
16. C. C. Clark: *Hydrazine,* Mathieson Chemical Corp., Baltimore, MD 1953.
17. L. F. Audrieth, B. A. Ogg: *The Chemistry of Hydrazine,* Wiley, New York 1951.
18. G. H. Hudson et al.; *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol 8, Suppl. 2, Wiley, New York 1967, pp. 69 – 113.
19. C. L. Yaws et al.: *Chem. Eng. NY* **81** (1974) 91.
20. R. W. Gallant, *Hydrocarbon Process,* **48** (1969) 117.
21. F. E. Scott et al.: *Explosive Properties of Hydrazine,* Report of Investigation 4460, U.S. Dept. of the Interior, Bureau of Mines, Pittsburgh, PA 1949.
22. M. L. Davis et al., *Kinetics, Thermodynamics, Physico-Chemical Properties and Manufacture of Hydrazine,* Dept. of Chem., Ohio State University, 1952.
23. H. Bock, *Z. Naturforsch. B Anorg. Chem. Org. Chem.* **178** (1962) 426.
24. E. J. Cuy, W. C. Bray, *J. Am. Chem. Soc.* **46** (1924) 1786.
25. W. R. Hodgkinson, *J. Soc. Chem. Ind. London* **33** (1914) 815.
26. Olin Mathieson Corp., GB 738 441, 1952 (F. Haller).
27. Olin Mathieson Corp., GB 778 347, 1954 (F. Haller).
28. U. Roberto, F. Roncali, *Chim. Ind. (Turin)* **6** (1904) 93.
29. A. W. Browne, F. F. Shetterley, *J. Am. Chem. Soc.* **30** (1908) 60.
30. R. A. Reed, *Chem. Prod. Chem. News* **20** (1957) 271.
31. R. A. Reed, *Hydrazine and its Derivatives,* Lectures, Monographs and Reports no. 5, The Royal Institute of Chemistry, London, 1957, 1 – 49.
32. P. S. Wharton, D. H. Bohlen, *J. Org. Chem.* **26** (1961) 3615.
33. D. Balcon, A. Furst, *J. Am. Chem. Soc.* **75** (1953) 4334.
34. D. K. Parker et al., *Rubber Chem. and Tech.* **65** (1992) 245.
35. Rhône-Poulenc, FR 1 162 413, 1958 (R. Meyer, D. Pillon).
36. Bayer, FR, 1 315 346, 1962 (H. Kohnen et al).
37. Produits Chimiques Ugine Kuhlmann, FR, 2 323 635, 1975 (J. P. Schirmann et al.).
38. Gilbert, *J. Am. Chem. Soc.* **51** (1929) 3394.
39. H. Hayashi, *Catal. Rev. Sci. Eng.* **32** (1990) 229.
40. F. Raschig, DE 198 307, 1907.
41. F. Raschig, US 910 858, 1909.
42. F. Raschig: *Schwefel- und Stickstoffstudien,* Verlag Chemie, Berlin 1924.
43. S. R. M. Ellis et al., *Ind. Eng. Chem. Proc. Des. Dev.* **3** (1964) no. 1, 18.
44. G. V. Jeffreys, J. T. Wharton, *Ind. Eng. Chem. Proc. Des. Dev.* **4** (1965) no. 1, 71.
45. J. W. Kahn, R. E. Powell, *J. Am. Chem. Soc.* **76** (1954) 2565.
46. G. Yagil, M. Ambar, *J. Am. Chem. Soc.* **84** (1962) 1797.
47. K. A. Kobe, J. J. McKetta, Adv. Pet. Chem. Refin. **2** (1959) 531.
48. Olin Mathieson Corp., US 2 715 061, 1955 (J. Felger, B. Nicolaidsen).
49. Olin Mathieson Corp., US 2 935 451, 1955 (J. Troyan).
50. Olin, USP 2 773 814, 1956 (B. H. Nicolaisen).
51. P. Schestakoff, DE 164 755, 1903.
52. O. Seuffert, E. Ihwe, DE 578 486, 1931.
53. *Ullmann,* 4th ed. **13,** 95 – 107.
54. R. Mundil, US 3 077 383, 1963.
55. R. Schliebs: "The Chemistry of the Bayer Hydrazine Process," *193rd ACS Meeting Abstracts,* Denver, 1987.
56. Mitsubishi Gas, EP 070 155, 1982 (K. Yasuhisa).
57. Produits Chimiques Ugine Kuhlmann, US 3 972 878, 1976 (J. P. Schirmann et al.).
58. *Chem. Eng. News* **52** (1974), no. 37, 18.
59. J. P. Schirmann, S. Y. Delavarenne: *Hydrogen Peroxide in Organic Chemistry,* Edisciences, Paris 1978, p. 81.
60. J. P. Schirmann, F. Weiss, *Tetrahedron Lett.* 1972 635.
61. E. G. E. Hawkins, *J. Chem. Soc. C* 1969 2663.
62. E. Schmitz et al., *Chem. Ber.* **97** (1964) 2521.
63. *Chem. Eng. News* **59** (1981) no. 44, 32.
64. H. E. Malone: *The Determination of HydrazinoHydrazide Groups,* Pergamon Press, London 1970.
65. I. M. Kolthoff, R. Belcher: *Volumetric Analysis,* vol 3, Wiley-Interscience, New York 1957.
66. R. Christova et al., *Anal. Chim. Acta* **85** (1976) 301.

67. *Houben-Weyl,* Bd XIV/1, p. 220 (1967).
68. *Houben-Weyl,* **X 2,** 764.
69. Szeres, HU 34 723, 1983 (T. Wein et al.).
70. H. R. Lasman: "Blowing Agents" *in Encyclopedia of Polymer Science and Technology,* **vol. 2,** Wiley, New York 1965, pp. 532 – 565.
71. B. A. Hunter, J. Frados: *Plastics Engineering Handbook of the Society of the Plastics Industry,* Reinhold Publ. Co., New York 1976, p. 502 – 510.
72. Olin, USP 5 098 597, 1992 (E. F. Roth).
73. Nippon Carbide, EP 669 325, 1994 (M. Masha).
74. H. Martin, C. R. Worthing: "Pesticide Manual," 5th ed., *British Crop Protection Council,* Nottingham 1977.
75. E. Y. Spencer: *Guide to the Chemicals used in Crop. Protection,* Pulication 1093, 6th ed., Information Canada, Ottawa 1973.
76. W. T. Thomson: *Insecticides, Herbicides, Fungicides,* Thompson Publications, Fresno, CA 1977.
77. R. P. Oulette, J. A. King: *Chemical Week Pesticides Register,* McGraw-Hill, New York 1977.
78. *Farm. Chemicals Handbook,* Meister Publishing Co., 1977.
79. P. A. Worthington, *Pestic. Sci.* **31** (1991) 457.
80. *Info. Chemie* **353** (1993) 106.
81. Elf Atochem, EP 841 328, 1996 (P. Bourdauducq).
82. G. Bohnsack, *Sonderheft VGB-Speisewassertagung* (1972) 2 – 8.
83. V. K. Gouda, S. M. Sayed, *Corros. Sci.* **13** (1973) 647.
84. G. Bohnsack, *Vom Wasser* **53** (1979) 147.
85. T. Urbanski: *Chemistry and Technology of Explosives,* vol. 3, Pergamon Press, New York 1967.
86. G. Schulz-Ekloff, *Catal. Lett.* **6** (1990) 383.
87. E. W. Schmidt: *Hydrazine and its Derivatives,* Wiley and Sons, New York 1984.
88. A. N. Kost, R. S. Sagitullin, *Russ. Chem. Rev. (Engl. Translation)* **33** (1964) 159.
89. R. Ohme, A. Zubek, *Z. Chem.* **8** (1968) 41.
90. US 2 660 607, 1953 (C. Gever, C. J. O'Keefe).
91. E. V. Brown et al., *Method. Chim.* **6** (1975) 73.
92. *Houben-Weyl,* **X,** 2.
93. Mobay, US 4 172 092, 1978 (J. A. Malone).
94. Akzo, FR 1 252 335, 1974 (A. Casalonga).
95. F. Kurzer, M. Wilkinson, *Chem. Rev.* **70** (1970) 111.
96. E. Hafez, *Heterocycles* **22** (1984) 1827.
97. Chemie Linz EP 44 438, 1980 (H. Beer).
98. Bayer, EP 3550, 1978 (R. Kaiser).
99. VEB, DDR 249 009, 1986.
100. C. C. Haun, E. R. Kinkead, *Chronic Inhalation Toxicity of Hydrazine,* University of California, Irvine, Toxic Hazards Research Unit, Dayton, Ohio, Jan. 1975.
101. *Guide for Short Term Exposure of the Public to Air Pollutants,* V. Guide for Hydrazine, Monomethylhydrazine and 1,1-Dimethylhydrazine, Committee on Toxicology of the National Academy of Science, National Research Council, Washington D.C., Jan. 1975.
102. W. W. Melvin, W. S. Johnson: *A Survey of Information Relevant to Occupational Health Standards for Hydrazines,* Environmental Health Labs., NTIS, U.S. Dept. of Commerce, Springfield, VA., Mar. 1976.
103. *Occupational Exposure to Hydrazines,* Criteria for a Recommended Standard, U.S. Dept of Health, Education and Welfare, NIOSH, June 1978.

Hydrazoic Acid and Azides

Horst H. Jobelius, Dynamit-Nobel AG, Troisdorf, Federal Republic of Germany

Hans-Dieter Scharff, Dynamit-Nobel AG, Troisdorf, Federal Republic of Germany

1. Introduction 1
2. Physical Properties of Sodium Azide . 1
3. Chemical Properties of Sodium Azide and Hydrazoic Acid 1
4. Production 2
4.1. Sodium Azide 2
4.2. Hydrazoic Acid 3
5. Quality Specifications and Analysis . 3
6. Storage and Transportation 3
7. Uses 3
7.1. Salts 3
7.2. Organic Synthesis 4
8. Safety Precautions and Environmental Protection 5
9. Economic Aspects 5
10. Toxicology and Occupational Health . 5
11. References 6

1. Introduction

Hydrazoic acid itself has little industrial significance; the most important salts are sodium azide and lead azide.

The oldest process for preparing both the acid and the salts is based on diazotization of hydrazine and its salts [1, 2]. The method still most widely used today for production of sodium azide is the reaction of sodium amide with dinitrogen monoxide. This process was introduced industrially in 1924, but not patented until 1935 [3].

The original use of sodium azide was the safe production of lead azide, a fulminating (detonating) agent. Lead azide now constitutes only a small part of the total consumption of the sodium salt. By far the major proportion is used in the organic chemical industry. Increasing importance is ascribed to the use as a gas generator fuel that produces nontoxic nitrogen.

The toxicity and the unsettled question of potential carcinogenicity [4] make special workplace safety measures imperative.

2. Physical Properties of Sodium Azide

Sodium azide [*26628-22-8*], NaN_3, M_r 65.01, is a white, odorless, crystalline substance. Some physical properties of sodium azide are as follows:

mp	275 °C (decomp.)
Density (20 °C)	1.846 g/cm^3
Vapor pressure (20 °C)	1 Pa
Solubility in 100 mL water (15 °C)	41.0 g

Thermal decomposition occurs at 300 °C with formation of sodium metal and nitrogen.

3. Chemical Properties of Sodium Azide and Hydrazoic Acid

Heavy-metal azides, which are sensitive to friction and shock, precipitate readily from aqueous solution. Acidification of sodium azide produces explosive and toxic hydrazoic acid.

Free hydrazoic acid reacts with solutions or slurries of alkali-metal or alkaline-earth hydroxides, carbonates, and hydrogen carbonates to give solutions of the corresponding azides, which can be obtained in pure form by evaporation under vacuum.

Alkyl and acyl halides react with sodium azide to form alkyl and acyl azides, respectively. Sodium azide is also used in the Curtius reaction for the production of amines. (For further details of organic syntheses with HN_3 and NaN_3, see Section 7.2.)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_193

4. Production

4.1. Sodium Azide

The most important production process has been developed from the work of WISLICENUS, first published in 1892 [5]; it consists of two steps. In the first step, ammonia is fed into molten sodium metal, which is contained in a closed steel vessel, at 350 °C:

$$2\,Na + 2\,NH_3 \longrightarrow 2\,NaNH_2 + H_2$$

In the second step the resulting liquid sodium amide is converted to sodium azide, sodium hydroxide, and ammonia by reaction with dinitrogen monoxide at ca. 230 °C in a horizontal nickel reactor:

$$2\,NaNH_2 + N_2O \longrightarrow NaN_3 + NaOH + NH_3$$

The solid mixture is known as crude azide. Part of the crude azide is left in the reactor for the next charge; this gives better distribution of the added sodium amide and also improves the conversion. Dinitrogen monoxide is generated by decomposition of a 50 % ammonium nitrate solution at 230 °C. The latter is prepared by absorption of the ammonia liberated in the second stage and that remaining from the first stage in 50 % nitric acid. An additional absorption medium is spent acid from nitration processes that has been recycled and worked-up. Nitric acid oxidation of the waste from adipic acid production also gives dinitrogen monoxide [6].

To isolate sodium azide, the crude azide is dissolved in water and the resulting solution is clarified and then evaporated in vacuum. The separated residual crystalline material is dried at 110 °C and may be used or shipped after the agglomerates have been broken up.

The mother liquor must be evaporated at least twice. While this is being done, any residual sodium azide is decomposed by the addition of sodium nitrite and dilute nitric acid (cf. Chap. 8):

$$NaN_3 + NaNO_2 + 2\,HNO_3 \longrightarrow N_2 + N_2O + H_2O + 2\,NaNO_3$$

The yield based on sodium is ca. 90 %, when the formation of an equimolar amount of sodium hydroxide is taken into account.

In addition to the toxicity of the final product, the risk of explosion of the gas mixture is a potential hazard. Sodium amide reacts vigorously with water.

A variation of the Wislicenus process uses sodium monoxide and consists of a single step [7]:

$$2\,Na_2O + NH_3 + N_2O \longrightarrow NaN_3 + 3\,NaOH$$

Ammonia can be replaced, wholly or in part, by sodium amide:

$$Na_2O + NaNH_2 + N_2O \longrightarrow NaN_3 + 2\,NaOH$$

No gas is liberated in this process, which may thus be run at elevated pressure.

Formation of sodium amide in liquid ammonia [8–10] and azide reaction in liquid ammonia or high-boiling hydrocarbons [11, 12] have also been reported.

When dinitrogen monoxide is added before amide formation is complete, nitrogen is given off in the one-step reaction rather than ammonia. The yield of sodium azide with respect to sodium hydroxide is less favorable [13]:

$$4\,Na + 3\,N_2O + NH_3 \longrightarrow NaN_3 + 3\,NaOH + 2\,N_2$$

Liquid ammonia [14, 15] or ethanol [11] can be used to separate the sodium hydroxide.

By following the basic experiments of CURTIUS and THIELE, production by diazotization of hydrazine has been developed further [16–18]:

$$N_2H_4 + C_2H_5ONO + C_2H_5ONa \longrightarrow NaN_3 + 2\,C_2H_5OH + H_2O$$

The ethyl nitrite required for this reaction is produced from ethanol and sodium nitrite. A serious drawback to the use of this process is the carcinogenicity of hydrazine and its salts (→Hydrazine).

4.2. Hydrazoic Acid

Hydrazoic acid is liberated as a gas by decomposition of sodium azide solution with dilute sulfuric acid. A nonexplosive gas mixture is formed by simultaneous dilution with air or nitrogen to a concentration of less than 15 %. Aqueous solutions of ca. 10 % that are safe to handle can be obtained by concentration of dilute solutions. Solutions with a hydrazoic acid content in excess of 20 % are liable to decompose explosively.

5. Quality Specifications and Analysis

In accordance with manufacturers specifications, sodium azide must fulfill the following requirements:

Color, form	White, crystalline
Azide content	99.0 % (min.)
Alkalinity	0.1 % (max.)
Water	0.1 % (max.)
Insolubles	0.1 % (max.)
Chloride (as NaCl)	0.05 % (max.)
Sulfate (as $BaSO_4$)	0.05 % (max.)
Carbonate	0.05 % (max.)
Nitrate	0.05 % (max.)
Iron	0.05 % (max.)

The azide content is determined titrimetrically by addition of an excess of cerium (IV) salt and back titration with oxalic acid.

On testing for nitrate, acidic diphenylamine solution must not turn blue when sodium azide is added.

Azides may be detected with strips of filter paper soaked in 10 % iron(III) chloride solution. This test is highly sensitive because of the formation of the intensely colored complex ion $[Fe(N_3)_6]^{3-}$.

6. Storage and Transportation

Dry sodium azide is stable for years if stored in sealed containers at room temperature. Sodium azide is packed in sheet-iron drums lined with polyethylene sacking.

In accordance with regulations concerning dangerous substances, sodium azide must be marked as a poison. The MAK value is 0.07 ppm or 0.2 mg/m^3.

Transportation codes:

IMDG (GGVSee):	Class 6.1 D E-F 6165 UN no. 1687
RID (GGVE):	Class 6.1 Rn 601, Section 42b
ADR (GGVS):	Class 6.1 Rn 2601, Section 42b
ADNR:	Class 6.1 Rn 6601, Section 42b
EEC Directives/ GefStoffV:	011-004-00-7
United Kingdom Blue Book:	Poison and IMDG Code E 6165
United States CFR 49:	Paragraph 172.101 Poison B
ICAO/IATA-DGR:	Class 6.1 UN no. 1687

7. Uses

7.1. Salts

Sodium azide is the starting material for other metal azides. Nitrogen is released during thermal decomposition of sodium azide. The reaction is exothermic and is initiated by application of heat. In gas generators, the sodium metal formed is oxidized and reacts with silicon oxide or boron oxide to form sodium silicate or borate, respectively.

Sodium azide is used in automobile safety devices, e.g., steering-wheel air bags, where instantaneous delivery of cold, dust-free, and nontoxic gas is of utmost importance in preventing injury to the driver.

Lead azide is used as a detonating (fulminating) agent (→Explosives, Chap. 4.) [19–21]. The actual content of lead azide depends on the purpose for which it is required; the range is from 80 % to almost 100 %. The product must not contain needle-shaped crystals. Crystal growth to produce spheroids is controlled by addition of dextrin; gelatin, carboxymethyl cellulose, and polyhydroxypolycarboxylic acids or their soluble salts are also effective.

Silver azide [*13863-88-2*], AgN_3, has some advantages in certain electrical ignition systems over the less expensive lead azide.

Copper azides CuN_3 and $Cu(N_3)_2$ are hazardous products that are liable to form when sodium azide powder and hydrazoic acid are being handled. Copper azides form on objects made of copper or brass and tend to decompose explosively under mechanical load. Sieves, and dust or oil filters of pumps must be free of copper whenever a risk of contact with azides is present.

Barium azide [*18810-58-7*], $Ba(N_3)_2$, differs from other alkali-metal and rare-earth azides in that it can be precipitated directly from aqueous solutions of sodium azide and barium bromide. Barium azide is used in fluorescent light tubes to improve the light yield.

7.2. Organic Synthesis

Sodium azide is used in organic synthesis under very strict safety precautions. Reactions with sodium azide or its solutions may be carried out in plastic, glass, or enameled vessels that are free of exposed metal parts. Solid sodium azide is best stored in plastic containers. Before experiments begin, all apparatus must be checked for bare metal parts. Overheating of individual parts, (e.g., agitator shafts) should be avoided. Care must be taken to remove the poisonous and highly explosive gases resulting from formation of hydrazoic acid, and effective protection against explosion must be provided.

Reactions with sodium azide or hydrazoic acid enable a series of organic azides and derivatives to be obtained.

Substitution. The S_N2 reaction of alkyl halides with sodium azide is generally applicable in the aliphatic series. Dimethylformamide is particularly suitable as solvent. Acid chlorides react rapidly, producting the acid azides necessary for the Curtius reaction:

$$RC(=O)-Cl + NaN_3 \longrightarrow RC(=O)-N_3 + NaCl$$

Acid azides are also used in peptide synthesis because the reaction with amino acids proceeds under such mild conditions that the optically active configuration is retained.

Particularly active esters and tertiary alcohols can also react with metal azides to yield the corresponding azides. Aromatic diazonium salts react with azide ions in a type of extended Sandmeyer reaction:

$$R-N\equiv N^+ + N_3^- \longrightarrow R-N_3 + N_2$$

Addition. Olefins react with hydrazoic acid to form alkyl azides:

$$RC(H)=CR^1(H) + HN_3 \longrightarrow RCH_2-CH(R^1)(N_3)$$

Isocyanates are transformed to carbamic acid azides. Ketenes form acid azides. *Trans*-α-Azidoalcohols are formed from epoxides.

Salt Formation. Amines and phosphines form salts with hydrazoic acid. For example, triphenylphosphine reacts with 2 mol of hydrazoic acid to produce the triphenylphosphine–iminium salt.

Miscellaneous Reactions. Organic azides are used as starting materials for the preparation of heterocyclic compounds such as imidazoles, oxazoles, 1,2,3-triazoles, and 5- or 1-substituted tetrazoles:

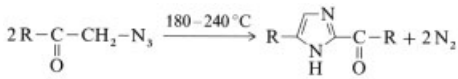

Thermal or photochemical decomposition of organic azides occurs with evolution of nitrogen to give the corresponding nitrenes. The latter contain monovalent nitrogen and, like carbenes, undergo isomerization, insertion reactions, or addition of nucleophiles. Many azide reactions, such as their frequently explosive decomposition, ring closure, Curtius or Schmidt reactions, photochemical cross-linking in polymer systems, and ring expansion, involve nitrene intermediates.

8. Safety Precautions and Environmental Protection

In the production of sodium azide, sodium amide represents a dangerous material, because it reacts violently with water; contact with water or air during storage must therefore be prevented. As the azide reaction continues, the evolving hydrogen must be burned off outside the building or purified for further use.

All equipment and gas piping must be fitted with explosion valves or rupture membranes because all the evolved gases are explosive in a wide range of mixtures. The hazard is increased by airborne dust particles.

When handling sodium azide, its solutions, and highly volatile hydrazoic acid, impervious protective clothing must be worn. To protect the respiratory organs from dust or gases, dustproof masks must be used. Before the masks are put on, their blank metal parts must be lacquered, care being taken that the lacquer is free of pores. A supply of fresh air to the masks is advantageous.

Working materials should preferably be plastics, but nonsparking metals may also be used. Sodium azide moistened with water reacts violently when ignited, giving off a highly caustic smoke.

In the event of fire, water-containing extinguishers must not be used; dry sand and anhydrous soda (soda ash) are suitable.

Small amounts of azide-containing waste and azide dust can be decomposed harmlessly with ceric ammonium nitrate. Larger quantities must be destroyed by addition of a sodium nitrite solution followed by a mineral acid solution. Azides react with acids to give a gas cloud capable of spontaneous detonation.

9. Economic Aspects

The main producers of azides are Canadian Industries (Canada), Masuda (Japan), and Dynamit-Nobel AG (Federal Republic of Germany). Total annual production is ca. 250 t. New production plants are expected to be built in the near future because of the increasing demand from producers of automobile safety devices.

10. Toxicology and Occupational Health

Dry sodium azide, its solutions, and hydrazoic acid are extremely toxic and can be absorbed by inhalation, ingestion, and skin absorption [22, 23]. In the event of contact of the skin or mucous membranes with azides, the affected parts must be rinsed thoroughly with water.

The MAK value for sodium azide is 0.07 ppm; for hydrazoic acid, 0.1 ppm.

Even minute amounts of azides cause headache, irritation of the eyes and mucous membranes, heightened irritability and excitation, impairment of vision, shortness of breath, reduction of blood pressure, risk of collapse, and fainting. The maximum nonfatal limit is given as 0.01 g; nevertheless, 0.005 – 0.01 g can have harmful effects.

Minor complaints can be eased quickly by fresh air or by administration of oxygen. No chronic poisoning has been observed. Azides appear to be similar to cyanides and prussic acid, in their ability to affect the cytochrome oxidase system and probably become bound to methemoglobin; *p*-aminopropiophenone has been suggested as an antidote. Amyl nitrite, which is effective in prussic acid poisoning, increases the risk of collapse.

11. References

1. T. Curtius, *Ber. Dtsch. Chem. Ges.* **23** (1890) 3023.
2. J. Thiele, *Ber. Dtsch. Chem. Ges.* **41** (1908) 2681.
3. Dynamit AG, DE 619 753, 1935.
4. MAK-Wertliste 1987.
5. W. Wislicenus, *Ber. Dtsch. Chem. Ges.* **25** (1892) 2084.
6. H. J. Zimmer Verf.tech., DE 1 144 243, 1959 (F. J. Carduck).
7. Degussa, DE 1 206 405, 1964 (G. Bretschneider, G. Düsing, H. Klöpfer, F. Sperr, H. Schmidt).
8. Canadian Industries LTD, CA 192 862, 1974.
9. Asahi Glass Co. LTD, JP 7 395, 1960 (S. Abe, M. Funaoka).
10. Hercules Powder Co., US 2 994 583, 1961 (D. Lewering).
11. JP 52 565, 1974 (T. Suwa).
12. WASAG, DE 619 017, 1935.
13. S. Abe, T. Kawakami, *Sci. Rep. Res. Inst. Tohoku Univ. Ser.* **A 4** (1952) 105.
14. PPG Industries Inc., US 2 994 583, 1976 (D. Tunison).
15. T. Kawakami, S. Moroto, A. Watanabe, *Chem. Abstr.* **54** (1960) 4226 h.
16. Atlas Powder Co., US 1 628 380, 1927 (F. Wilcoxon, B. Grotta).
17. C. Nägeli, E. Vogt-Marcus, *Helv. Chim. Acta* **15** (1932) 60 – 75.
18. YU 33 175, 1968 (M. Bozovic).
19. B. T. Fedoroff: *Encyclopedia of Explosives,* **vol. 1,** Picatinny Arsenal, Dover, N.J. USA, 1960, pp. A 521–644.
20. T. Urbanski: *Chemistry and Technology of Explosives,* **vol. 3,** Pergamon Press, Oxford 1967, pp. 161 – 200; **vol. 4,** 1984, pp. 469 – 490.
21. H. D. Fair, R. F. Walker: *Energetic Materials,* vol. 1 + 2, Plenum Press, New York 1977.
22. DIN-Sicherheitsdatenblatt und Hinweise für das sichere Arbeiten mit Natriumazid, Dynamit Nobel AG, D 5210 Troisdorf, Kaiserstr.
23. G. Hommel: *Handbuch der gefährlichen Güter,* Merkblatt 894, Springer-Verlag, Berlin 1985.

Hydrides

PETER RITTMEYER, Chemetall GmbH, Frankfurt, Federal Republic of Germany

ULRICH WIETELMANN, Chemetall GmbH, Frankfurt, Federal Republic of Germany

1. Introduction 2
2. Hydrides of the Alkali and Alkaline Earth Metals 3
2.1. General 3
2.2. Lithium Hydride 5
2.3. Sodium Hydride 6
2.4. Potassium Hydride 7
2.5. Beryllium Hydride 8
2.6. Magnesium Hydride 9
2.7. Calcium Hydride 9
3. Binary and Complex Hydrides of the Boron Group 10
3.1. General 10
3.2. Lithium Tetrahydridoborate 12
3.3. Lithium Triethylhydridoborate . . . 12
3.4. Sodium Tetrahydridoborate 13
3.5. Sodium Cyanotrihydridoborate . . 16
3.6. Potassium Tetrahydridoborate . . . 17
3.7. Aluminum Hydride 17
3.8. Lithium Tetrahydridoaluminate . . 17
3.9. Lithium (tri-tert-butoxy)hydridoaluminate 20
3.10. Sodium Tetrahydridoaluminate . . . 20
3.11. Sodium bis(2-methoxyethoxy)dihydridoaluminate 20
3.12. Sodium Diethyldihydridoaluminate 21
4. Recommended Techniques for Working with Hydrides 21
5. Hydrides of the Transition Metals . 22
5.1. Binary Hydrides 22
5.1.1. Titanium Hydride 24
5.1.2. Zirconium Hydride 25
5.2. Intermetallic Hydrides 25
5.3. Complex Hydrides of Transition Metals 26
5.3.1. Di(η^5-cyclopentadienyl)chlorohydridozirconium(IV) 26
5.3.2. Carbonylhydrido-tris(triphenylphosphine)rhodium(I) 27
6. Handling, Storage, and Transportation 27
7. Toxicology and Occupational Health 28
8. References 28

1. Introduction [1–8]

The hydrides are a large group of compounds with diverse structures and bonding types. They may be divided into four classes according to their bonding character:

1) Ionic hydrides
2) Polymeric covalent hydrides
3) Volatile covalent hydrides
4) Metallic hydrides.

The demarcation between the classes is not sharp; they merge into each other according to the electronegativities of the elements concerned.

Arranging the binary hydrides as in the periodic table (see Fig. 1, next page), the *ionic hydrides* are seen to be formed by the strongly electropositive alkali and alkaline earth metals. Beryllium hydride, which is partly covalent in character, is an exception.

The elements that form *polymeric covalent hydrides*, apart from beryllium, belong to Groups 12 (zinc) and 13 (boron). Boron occupies a special position. It forms numerous volatile hydrides with unusual structures (→ Boron Compounds, Chap. 5.1.). The hydrides of Group 13 are electron deficient; saturation is achieved in the complex hydride ions such as BH_4^- and AlH_4^-. The covalent hydrides formed by the elements of groups 14 – 17 (carbon – f luorine) are characterized by high volatility. The polarity of the element – hydrogen bond changes with increasing electronegativity of the elements. Whereas the hydrides of the metallic elements and of silicon contain hydridic (negatively polarized) hydrogen, those of the nonmetals contain acidic hydrogen, sometimes strongly so. The latter are not regarded as hydrides in the usual sense

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_199

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
LiH 0.97	BeH_2 1.47											BH_3 2.01	CH_4 2.50	NH_3 3.07	H_2O 3.50	HF 4.10
NaH 1.01	MgH_2 1.23											AlH_3 1.47	SiH_4 1.74	PH_3 2.06	H_2S 2.44	HCl 2.83
KH 0.91	CaH_2 1.04	ScH_2 1.20	TiH_2 1.32	VH VH_2 1.45	CrH (CrH_2) 1.56				$NiH_{<1}$ 1.75	CuH 1.75	ZnH_2 1.66	(GaH_3) 1.82	GeH_4 2.02	AsH_3 2.20	H_2Se 2.48	HBr 2.74
RbH 0.89	SrH_2 0.99	YH_2 YH_3 1.11	ZrH_2 1.22	(NbH_2) 1.23					$PdH_{<1}$ 1.35		(CdH_2) 1.46	(InH_3) 1.49	SnH_4 1.72	SbH_3 1.82	H_2Te 2.01	HI 2.21
CsH 0.86	BaH_2 0.97	LaH_2 LaH_3 1.08	HfH_2 1.23							(AuH_3) 1.42	(HgH_2) 1.44	(TlH_3) 1.44	PbH_4 1.55	BiH_3 1.67	H_2Po 1.76	HAt 1.96
Lanthanides		LaH_2 LaH_3 1.08	CeH_2 1.06	PrH_2 PrH_3 1.07	NdH_2 NdH_3 1.07		SmH_2 SmH_3 1.07	EuH_2 1.01	GdH_2 GdH_3 1.11	TbH_2 TbH_3 1.10	DyH_2 DyH_3 1.10	HoH_2 HoH_3 1.10	ErH_2 ErH_3 1.11	TmH_2 TmH_3 1.11	(YbH_2) YbH_3 1.06	LuH_2 LuH_3 1.14
Actinides		AcH_2 1.00	ThH_2 1.11	PaH_2 1.14	UH_3 1.22	NpH_2 NpH_3 1.22	PuH_2 PuH_3 1.22	AmH_2 AmH_3 1.2								

Ionic hydrides
Covalent polymeric hydrides
Covalent hydrides
Metallic hydrides

Figure 1. Periodic system of binary hydrides (the figures in the table are the Allred – Rochow electronegativities) Hydrides in parentheses are unstable at ambient temperature.

and are not discussed in the present article.

Metallic hydrides are formed by the transition metals, the lanthanides, and the actinides. The structures of these hydrides are fundamentally similar to those of metals. The compositions of the metallic hydrides can vary and are frequently nonstoichiometric. Another class of hydrogen compounds, the transition metal hydride complexes, is becoming important. In these molecular complexes the hydrogen atoms are covalently linked to a transition metal. Economically, the most important hydrides are those of the alkali and alkaline earth metals and the complex hydrides of boron and aluminum. These are mostly produced in tonnage quantities and have a wide range of applications.

2. Hydrides of the Alkali and Alkaline Earth Metals

2.1. General

The alkali and alkaline earth metals react exothermically with hydrogen at elevated temperature to form ionic hydrides:

$$2\,Na + H_2 \longrightarrow 2\,Na^+H^-$$
$$Ca + H_2 \longrightarrow Ca^{2+}H_2^-$$

The typical course of a metal – hydrogen reaction is shown in Figure 2. The rate of hydrogen uptake is shown as a function of H_2 pressure at different temperatures. At constant temperature the metal initially takes up hydrogen by dissolving it (solution phase). As more hydrogen is taken up hydride formation starts, resulting in a mixture of two phases: hydride and solution.

The pressure remains constant (plateau region) until the combined hydrogen reaches a limiting stoichiometry, after which the pressure again increases sharply until the ideal stoichiometric composition MH_2 is reached (hydride phase). In the case of the salt-like hydrides, the hydride phase is very small, i.e., the limiting stoichiometry is near to the ideal.

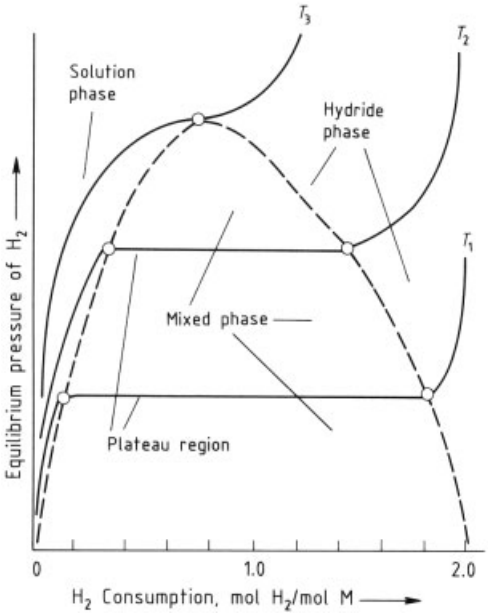

Figure 2. Isotherms of a hypothetical dihydride MH_2 ($T_3 > T_2 > T_1$).

For the metallic hydrides the hydride phase can become very large. It can be seen from Figure 2 that although the rate of hydride formation can be increased by increasing the temperature and the hydrogen pressure there is a penalty: heating salt-like or metallic hydrides causes reversible decomposition into the elements. The dissociation pressure in the plateau region is given over a wide range of temperature by the following relationship:

$$\log_{10} p_{\mathrm{H2}} = -\frac{A}{T} + \mathrm{B}$$

Values for the constants A and B are given in the literature [2,4]. Plateau dissociation pressures of some alkali and alkaline earth metal hydrides are shown in Figure 3. The noticeable instability of MgH_2 may be attributed to a transition towards the behavior of BeH_2 and other covalent polymeric hydrides.

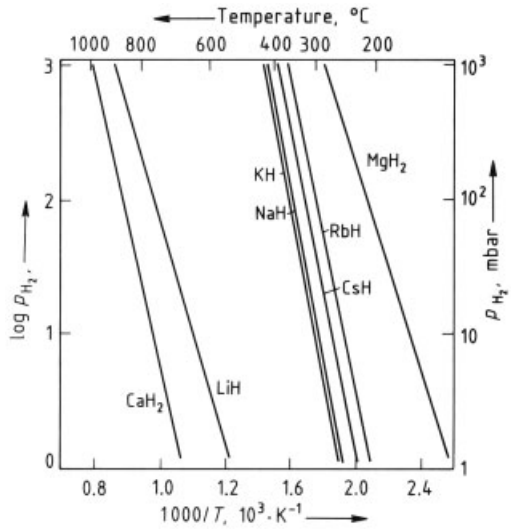

Figure 3. Plateau dissociation pressures of alkali and alkaline earth metal hydrides as a function of temperature.

The other hydrides of the alkali and alkaline earth metals possess an ionic structure with negatively charged hydrogen ions, as illustrated by the following:

1) Electrolysis of a molten salt mixture of the hydrides with alkali metal halides and hydroxides causes liberation of hydrogen at the anode.
2) Lattice energy calculations based on an ionic model give results in good agreement with experimental values.
3) They are insoluble in inert organic solvents.

The crystal structures of the alkali metal hydrides are of the rock salt (NaCl) type, the alkaline earth metal hydrides have structures of the $PbCl_2$ type, only MgH_2 having the rutile (TiO_2) structure [9]. The more important physical properties are given in Table 1, see next page. It should be noted that the density of the hydrogen (grams of hydrogen per cubic centimeter of hydride) in some of these hydrides is sometimes appreciably higher than that of liquid hydrogen itself ($\varrho = 0.0710$ g/cm^3 at 20.4 K).

The *chemical properties* of the aforementioned hydrides are determined by their high basicity and reducing power. The hydride ion

Table 1. Physical data for alkali and alkaline earth metal hydrides

Hydride	M_r	CAS registry number	Hydrogen content, wt %	Density, g/cm^3	ΔH^0_B, [a] kJ/mol	ΔF^0_B, [b] kJ/mol	S^0, [c] J K^{-1} mol^{-1}
LiH	7.95	[*7580-67-8*]	12.68	0.775	− 90.7	− 70.0	25
NaH	24.00	[*7646-69-7*]	4.20	1.363	− 56.5	− 37.7	48
KH	40.11	[*7693-26-7*]	2.51	1.430	− 57.9	− 37.3	61
RbH	86.48	[*13446-75-8*]	1.17	2.595	− 47.4	− 30.6	71
CsH	133.91	[*13772-47-9*]	0.75	3.410	− 49.9	− 30.6	81
BeH_2	11.03	[*13597-97-2*]	18.28	0.65	− 19.3		
MgH_2	26.32	[*7693-27-8*]	7.66	1.45	− 75.4	− 36.0	
CaH_2	42.10	[*7789-78-8*]	4.79	1.902	−186.3	−147.4	42
SrH_2	89.64	[*13598-33-9*]	2.25	3.269	−180.5	−138.6	54
BaH_2	139.36	[*13477-09-3*]	1.45	4.156	−171.2	−132.3	67

[a] ΔH^0_B = Enthalpy of formation.
[b] ΔF^0_B = Free enthalpy of formation (298 K).
[c] S^0 = Entropy.

reacts irreversibly with compounds containing acidic protons, liberating hydrogen:

$$M^+H^- + H^{\delta+}X^{\delta-} \longrightarrow M^+X^- + H_2$$

X = halogen, OH, OR, HNR, SR

The reactivity of the hydrides increases from lithium to cesium and from magnesium to barium; it is also strongly dependent on particle size. Superactive alkali metal hydrides are obtained by the hydrogenation of alkali metal alkyls in hexane in the presence of a strong base (e.g., tetramethylethylenediamine) [10, 11]. These superactive hydrides can be used to metallate compounds that do not react with commercially available hydrides. The high reducing power of the ionic hydrides ($H^- \rightarrow 1/2\,H_2 + e^-$; $E^0 = -2.23$ V) is used in metallurgical processes (see Sections 2.3 and 2.7). The reduction of organic compounds is better carried out with the complex hydrides obtained by reaction of alkali or alkaline earth metal hydrides with Lewis acids (see Chap. 3).

Analysis of these hydrides is carried out using the exothermic reaction with water, e.g.,

$$LiH + H_2O \longrightarrow LiOH + H_2$$

The hydrogen liberated is determined volumetrically and the alkaline solution titrated with acid.

2.2. Lithium Hydride

Properties (see also Table 1). Pure lithium hydride forms colorless, cubic crystals. The commercial product contains traces of impurities, e.g., unreacted lithium metal, and is consequently light gray or blue. Physical and thermodynamic data on lithium hydride and deuteride have been determined in detail [12, 13]. Lithium hydride is thermally very stable (see Fig. 3), being the only ionic hydride that melts without decomposition at atmospheric pressure (*mp* 688 °C). In contrast to the other alkali metal hydrides, lithium hydride is slightly soluble in inert polar organic solvents such as ethers. It forms eutectic mixtures with a large number of salts [2].

Lithium hydride is stable in dry air but ignites at increased temperature. In moist air it is hydrolyzed exothermically; finely divided material can ignite spontaneously. At elevated temperature, it reacts with oxygen to form lithium oxide, with nitrogen to form lithium nitride and hydrogen, and with carbon dioxide to form lithium formate. With compounds containing acidic hydrogen atoms (e.g., alcohols) it reacts with liberation of hydrogen:

$$R{-}OH + LiH \longrightarrow R{-}OLi + H_2$$

Reaction of lithium hydride with water is extremely violent. Above 180 °C it reacts with silicates (as does metallic lithium) so that apparatus made of glass, silica, porcelain, or enamelled materials should not be used. The nucleophilic properties of lithium hydride enable it to be used in condensation reactions, e.g., of aldehydes and ketones, and reductions, e.g., the conversion of acyl chlorides to aldehydes. Nevertheless, sodium hydride is more effective for

condensations and complex hydrides for reductions. Another characteristic reaction is the addition of the hydride anion to the coordinatively unsaturated Lewis acids (see Chap. 3).

Production. Lithium hydride is manufactured on a large scale by combination of the elements at 700 – 900 °C:

$$2\,Li + H_2 \longrightarrow 2\,LiH$$

The lithium metal is held in an iron crucible of low-carbon content enclosed in a vessel of highly alloyed scale-resistant steel. After being melted under vacuum, the metal is reacted with hydrogen at 700 °C and ca. 0.1 MPa. The strongly exothermic reaction is controlled by the hydrogen feed rate. The denser lithium hydride sinks to the bottom of the crucible, leaving the reactive metal at the surface. On completion of the reaction, the melt is slowly cooled and the blue-gray crystalline block is broken down to the required particle size. The product has a purity of 98 – 99 % and is supplied in the form of lumps or powder with a bulk density of ca. 0.29 kg/L. The commercial producers are Chemetall GmbH, Lithium Corporation of America, and Ventron Corporation. The catalytic hydrogenation of lithium at atmospheric pressure and 25 – 50 °C in tetrahydrofuran (THF) in the presence of $TiCl_4$ – naphthalene has been reported [14]. This process has not yet been used on an industrial scale.

Uses. Most lithium hydride, which is produced in tonnage quantities, is used as a raw material in the production of lithium aluminum hydride (see Section 3.8) and lithium amide (→ Lithium and Lithium Compounds):

$$4\,LiH + AlCl_3 \longrightarrow LiAlH_4 + 3\,LiCl$$
$$LiH + NH_3 \longrightarrow LiNH_2 + H_2$$

Another important use is in the production of silane by the Sundermeyer process. Lithium hydride is treated with tetrachlorosilane, the reaction medium being a molten LiCl – KCl eutectic:

$$4\,LiH + SiCl_4 \longrightarrow SiH_4 + 4\,LiCl$$

In the industrial process this reaction is linked with the regeneration of lithium hydride by electrolysis of the molten salt and hydrogenation of the lithium metal in situ [1, 15].

The reaction can also be carried out in high-boiling solvents by the addition of small amounts of lithium aluminum hydride, e.g., in the production of disilane [16]:

$$2\,Si_2Cl_6 + 3\,LiAlH_4 \longrightarrow 2\,Si_2H_6 + 3\,LiCl + 3\,AlCl_3$$
$$4\,LiH + AlCl_3 \longrightarrow LiAlH_4 + 3\,LiCl$$

The reaction of lithium hydride with boron trialkyls gives trialkylborohydrides, which are soluble in THF.

$$LiH + R_3B \longrightarrow LiR_3BH$$

When R is an ethyl group the product is known as Super-Hydride (see Section 3.3). Trialkylborohydrides are used in organic chemistry as regio- and stereoselective reducing agents [17].

As a result of its high hydrogen content (hydrolysis of 1 g LiH produces 2.5 L of H_2 at STP), lithium hydride is used as a transportable hydrogen source, e.g., for inflatable rubber dinghies and balloons. Lithium deuteride [*13587-16-1*], LiD, is commercially available (Chemetall) and is used for labeling organic compounds with deuterium.

2.3. Sodium Hydride [18]

Properties (see also Table 1). Pure sodium hydride forms colorless cubic crystals; however, the commercial product contains traces of sodium metal, giving it a light gray color.

At atmospheric pressure, sodium hydride slowly evolves hydrogen above 300 °C. At 420 °C decomposition is rapid but melting does not take place. Sodium hydride is a salt and therefore insoluble in inert organic solvents. It dissolves in molten sodium hydroxide, in sodium – potassium alloys and in molten LiCl – KCl eutectic mixtures (352 °C). Sodium hydride is stable in dry air but ignites above 230 °C, burning to form sodium oxide (for fire precautions see Chap. 6). It is hydrolyzed rapidly in moist air and as a dry powder it is spontaneously flammable. Sodium hydride reacts extremely violently with water, the heat of hydrolysis being sufficient to ignite the liberated hydrogen. It reacts with carbon dioxide to form sodium formate.

The most important chemical property of sodium hydride is its basicity. Sodium hydride reacts irreversibly with a large number of compounds containing acidic hydrogen atoms, including alcohols, phenols, thiols, and amines, but also with weak carbon acids such as ketones, carboxylic esters, and acetylenic hydrocarbons. The carbanions formed play an important role in synthetic organic chemistry [18].

Sodium hydride reacts with Lewis acids to give addition complexes, e.g., with esters of boric acid it yields sodium borohydride (see Section 3.4). The practicability of these reactions is strongly dependent on the particle size and surface area of the sodium hydride because of the low solubility. Activators such as trialkylboranes and trialkylalanes can be used to accelerate the reaction, e.g., in the production of sodium aluminum hydride (see Section 3.10).

Sodium hydride is a poor reducing agent because of its high basicity and its insolubility in inert organic solvents. If the sodium hydride is complexed with alcoholates and metallic salts, complex reducing agents are formed that can reduce halides, alkenes, and ketones. The use of specific metal salts makes it possible to reduce the carbonyl group or the carbon – carbon double bond of α,β-unsaturated ketones regioselectively [19]. At high temperature, e.g., in molten sodium hydroxide, sodium hydride is strongly reducing and is used in the production of finely powdered metals and in metal surface treatment.

Production. The best method of synthesis is to react molten sodium metal with hydrogen in mineral oil. A satisfactory rate of reaction is obtained by finely dispersing the sodium with a high-shear turbine mixer (ca. 8000 rpm). The reaction is very rapid at 250 – 300 °C and atmospheric pressure and can be operated continuously [20]. A large part of the NaH – oil dispersion produced is used directly for production of sodium borohyride.

Sodium hydride is available commercially as 80 % and 60 % dispersions in mineral oil. The light-gray pourable material has an average particle size of 10 µm (Degussa, Chemetall, Ventron). The oil protects the surface of the highly reactive sodium hydride, which easily becomes passivated by the slightest atmospheric exposure or, worse, can ignite on longer contact with air.

Handling is facilitated by packing the material in bags made of plastic sheet soluble in organic solvents. For use in metal-descaling baths, the sodium hydride is produced directly in molten sodium hydroxide.

Uses. In the United States, a high proportion of the sodium hydride is used in the production of sodium borohydride (see Section 3.4). In the metal industry it is used as a solution (0.2 – 2.5 %) in molten sodium hydroxide for the descaling of metal surfaces [21]. The high reducing power of the hydride results in the conversion of the surface oxide to metal:

$$Fe_3O_4 + 4\,NaH \longrightarrow 3\,Fe + 4\,NaOH$$

Its ability to reduce metal salts is used in the production of powdered metals, often with very small particle size and large surface area [22].

The main application of sodium hydride is in organic synthesis [18, 23]. Its strong basicity is used in the production of ethers (Williamson synthesis), alkylation of amines, and alkylation of the α-carbon atoms of ketones, esters, carboxylic acids, and nitriles. The carbanions produced with sodium hydride are used in many

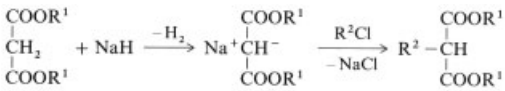

condensation reactions, (e.g., Claisen, Stobbe, and Dieckmann). These reactions are important in the production of pharmaceutical intermediates [24]. It should be noted that uncontrollable reactions have been reported when dimethylformamide was used as solvent [25]. Sodium hydride has several advantages compared with other reagents such as sodium alkoxides, sodium amide, or sodium metal: it reacts very quickly under mild conditions and apart from hydrogen no other byproducts (e.g., alcohols or water) are formed.

2.4. Potassium Hydride

Properties (see also Table 1). Potassium hydride is a considerably stronger base than lithium hydride or sodium hydride. It is able to remove protons from tertiary alcohols and ketones, a reaction that either does not occur or is very slow when sodium hydride is used [26]. Potassium hydride also reacts with weak Lewis acids,

converting sterically hindered boron trialkyls to the corresponding sterically hindered complex borohydrides:

$$KH + R_3B \xrightarrow{THF} KBHR_3$$

Potassium tri-*sec*-butylborohydride [*54575-49-4*] solution in tetrahydrofuran (THF) is sold as K-Selectride (Aldrich). These compounds may be used for highly stereoselective reductions, e.g., of ketones [26]. Chemical rearrangements induced by potassium hydride are important in organic chemistry. Furthermore, in combination with other reagents potassium hydride acts as a catalyst in the polymerization of alkenes and dienes [26].

Production. The process already described for sodium hydride manufacture can be used, i.e., hydrogenation of a metal dispersion in a high-boiling mineral oil. In addition, a process has been described in which phenanthrene acts as catalyst in the hydrogenation of potassium at atmospheric pressure and at 90 – 200 °C in hydrocarbon solvents such as nonane [27–29]. This catalytic reaction is specific for potassium and cannot be used with the other alkali metals. Potassium hydride is available in laboratory quantities only as a 20 – 35 % dispersion in oil.

2.5. Beryllium Hydride

Properties (see also Table 1). Beryllium hydride is a white, involatile, amorphous powder. The pure compound decomposes with evolution of hydrogen at about 220 °C. It is fairly stable to air at room temperature and is rapidly hydrolyzed by water if this is acidified. Beryllium hydride containing ether decomposes at ca. 130 °C and reacts very violently with water. Crystalline beryllium hydride ($\varrho = 0.78$ g/cm^3) is obtained when the LiH-doped amorphous material is densified by compression at 620 MPa and 200 °C [29]. The hydrogen density in this material (0.1425 g/cm^3) is twice that of liquid hydrogen. All forms of the covalent hydride possess a structure with Be–H–Be 3-center bonds. It is insoluble in inert organic solvents but reacts with tertiary amines at ca. 160 °C to form stable adducts. Beryllium hydride and its derivatives are extremely poisonous (→ Beryllium and Beryllium Compounds, Chap. 15.).

Production. The only known method of producing high purity beryllium hydride (95 – 98 wt % BeH_2) is the pyrolysis of di-*tert*-butylberyllium diethyl etherate in a high boiling hydrocarbon at 200 °C [30]:

$$((CH_3)_3C)_2Be \cdot O(C_2H_5)_2 \longrightarrow BeH_2 + 2\,(CH_3)_2C{=}CH_2 + O(C_2H_5)_2$$

The many attempts to react beryllium compounds (e.g., BeR_2, $BeCl_2$) with hydrides (e.g., $LiAlH_4$, $HAlR_2$) have given disappointing results, and direct synthesis from the elements has not proved possible.

Uses. Beryllium hydride has attracted considerable interest as a rocket fuel on account of its high heat of combustion. It has also been considered as a moderator for nuclear reactors. Methods of preparing and densifying the pure

Table 2. Physical data for alkali metal borohydrides and aluminum hydrides

Hydride	CAS registry number	M_r	Hydrogen content, wt %	Density, g/cm^3	*mp*, °C	ΔH^0_B, * kJ/mol
$LiBH_4$	[*16949-15-8*]	21.78	18.5	0.66	280	− 194
$LiBH(C_2H_5)_3$	[*22560-16-3*]	105.94	0.95			
$NaBH_4$	[*16940-66-2*]	37.83	10.6	1.07	505 **	− 191
$NaBH_3CN$	[*25895-60-7*]	62.84	6.4	1.20	240 dec.	
KBH_4	[*13762-51-1*]	53.94	7.4	1.17	585 **	− 229
$LiAlH_4$	[*16853-85-3*]	37.95	10.6	0.91	>125 dec.	− 119
$LiAlH(OC(CH_3)_3)_3$	[*38884-28-5*]	254.26	0.4	0.97	319 dec.	
$NaAlH_4$	[*13770-96-2*]	54.00	7.4	1.27	178	− 113
$NaAlH_2(OC_2H_4OCH_3)_2$	[*22722-98-1*]	202.16	1.0	1.12	<0	−1312
$NaAlH_2(C_2H_5)_2$	[*17836-88-3*]	110.11	1.8	0.88	85	

* ΔH^0_B = Enthalpy of formation.
** Under H_2 atmosphere.

compound are described in a large number of patents. However, the toxicity has always been a serious barrier to its commercial exploitation.

2.6. Magnesium Hydride

Properties (see also Table 1). Pure magnesium hydride is a white, nonvolatile powder. Although its properties are predominantly salt-like, it shows signs of a transition towards the covalent polymeric hydride structure of compounds such as beryllium and aluminum hydride. Its enthalpy of formation (see Table 1) and thermal stability (see Fig. 3) are considerably lower than those of the homologous calcium hydride. Magnesium hydride decomposes without melting at 280 °C. Its reactivity depends on the method of preparation. The product obtained by direct synthesis from the elements is stable in air. Nevertheless, if it is prepared by pyrolysis of dialkyl magnesium or by reaction of lithium aluminum hydride with dimethyl magnesium, the product is very pure and finely divided and ignites spontaneously on contact with air.

Production. Magnesium hydride is synthesized industrially from the elements at 300–400 °C and 10–15 MPa hydrogen pressure in a rotary autoclave provided with steel grinding balls:

$$Mg + H_2 \longrightarrow MgH_2$$

The commercial product is gray due to the presence of impurities, being about 97 % pure (Ventron). Less drastic reaction conditions are required when a homogeneous catalyst is used [31–33]. The hydrogenation can be carried out at 60–70 °C and 8 MPa in THF in the presence of halides of chromium, titanium, or iron and magnesium–anthracene complex. In contrast to the product from an uncatalyzed synthesis, this has a high specific surface area (90 m^2/g compared with 0.5 m^2/g), and is very reactive and pyrophoric.

Uses. Magnesium hydride contains a high proportion of hydrogen (7.65 wt %), which can be liberated by heat. The catalytically produced compound displays rapid rates of hydrogenation and dehydrogenation and it is therefore of interest as a hydrogen reservoir [34]. The high activity of catalytically produced magnesium hydride also enables it to take part in an addition reaction with 1-alkenes to form organomagnesium compounds [35]:

$$MgH_2 + 2\,CH_2{=}CHR \longrightarrow Mg(CH_2CH_2R)_2$$

These dialkylmagnesium compounds have become industrially important in the production of very active Ziegler catalysts.

The Mg–H_2 system can also be used for the simultaneous removal and purification of pure hydrogen from gas mixtures [36]. Magnesium hydride is also used as a drying agent for organic solvents and gases and in chemical syntheses.

2.7. Calcium Hydride

Properties (see also Table 1). Pure calcium hydride crystallizes in the form of colorless, hexagonal prisms, but the industrial product contains some calcium metal and is consequently gray. On heating, calcium hydride decomposes without melting. In keeping with its salt-like character it is insoluble in inert solvents but dissolves in molten LiCl–KCl eutectic mixtures (352 °C). Calcium hydride can be safely handled in air, but is slowly attacked and crumbles to powder if the air is moist. It reacts with dry oxygen or nitrogen if heated above 500 °C, forming the oxide or nitride, respectively. Once ignited, it burns with a strongly exothermic reaction:

$$CaH_2 + O_2 \longrightarrow CaO + H_2O$$

The reaction of calcium hydride with water is also very exothermic. Heating is necessary to initiate the reaction with alcohols to form alcoholates. The reaction with ammonia takes place at 700 °C with formation of calcium amide. At high temperature (600–1000 °C) calcium hydride is a powerful reducing agent.

Production. Calcium hydride is produced industrially by synthesis from the elements at 400 °C and atmospheric pressure:

$$Ca + H_2 \longrightarrow CaH_2$$

The reaction is strongly exothermic and is controlled by means of the hydrogen feed rate.

The commercial product contains 90–96 % CaH_2 (Degussa, Ventron). The main impurities are calcium metal (≤ 6 %) and calcium oxide (≤ 1 %). It is available as powder, granules, or lumps.

Uses. Calcium hydride is used to obtain metallic elements (e.g., Ti, Zr, V, Nb, U, and Th) from the oxides because of its very powerful reducing properties at elevated temperature (600–1000 °C):

$$2\,CaH_2 + TiO_2 \longrightarrow Ti + 2\,CaO + 2\,H_2$$

This process, the Hydrimet process, yields finely divided metal, the calcium oxide and excess calcium hydride being removed by washing with dilute acids. Powdered oxide mixtures are converted to powdered alloys.

The actual reducing agent in this process may not be the calcium hydride, but metallic calcium produced by thermal decomposition. Metal halides are also reduced to metal in a similar process. Another application of calcium hydride is in drying solvents and gases. As it does not strongly promote condensation reactions, it may be used to dry ketones, esters, higher alcohols, phenols, amines, and halogenated hydrocarbons in addition to the inert solvents; a water content of 1 ppm is attainable. An interesting application is the determination of the water content of diesel fuel and kerosene by reaction with calcium hydride and determination of the volume of hydrogen evolved. This water content is important in adjusting the optimum compression ratio of an engine.

3. Binary and Complex Hydrides of the Boron Group

3.1. General

All the elements of Group 13 (Boron group) form polymeric hydrides $(EH_3)_x$. The monomers EH_3 are strong Lewis acids and are unstable. Borane (BH_3) achieves electronic saturation by dimerization to form diborane (→ Boron Compounds, Chap. 5.1.). All other hydrides in this group attain closed electron shells by polymerization. Aluminum hydride, alane, $(AlH_3)_x$, has been extensively investigated, the hydrides of gallium indium and thallium much less so [3].

In contrast to the ionic and metallic hydrides, the hydrides of the boron group cannot be synthesized directly from the elements. A general method of preparation is to react a halide of the element with lithium hydride in an ether:

$$4\,LiH + ECl_3 \longrightarrow LiEH_4 + 3\,LiCl$$
$$3\,LiEH_4 + ECl_3 \longrightarrow 4\,EH_3 + 3\,LiCl$$

The complex hydride $LiEH_4$ is formed as an intermediate. The hydride EH_3 is then produced as a soluble etherate that rapidly polymerizes.

The binary hydrides can be stabilized by forming solvates or other addition compounds. Their stability decreases from boron to thallium. Borane and alane form a large number of donor–acceptor complexes whose properties are strongly dependent on the nature of the donor molecule. The main donor compounds are ethers, amines, phosphines, and hydride ions. Of particular interest are the hydride complexes of borane, the tetrahydridoborates $M(BH_4)_n$, and of alane, the tetrahydridoaluminates $M(AlH_4)_n$. The most industrially important complex hydrides are sodium tetrahydridoborate $NaBH_4$ and lithium tetrahydridoaluminate $LiAlH_4$. These are produced in tonnage quantities and used mainly as reducing agents in organic chemistry. Many other hydridoborates and hydridoaluminates are produced from these starting materials [37, 38]. The bonding characteristics and hence the properties of the complex hydrides are largely determined by the electronegativity difference between the metal atom and boron or aluminum. If this difference is large, e.g., in $NaBH_4$ or KBH_4, the compound is highly ionic and has a high melting point. If it is small, as with $Al(BH_4)_3$, the BH_4 group is covalently bonded and the compound is volatile (*bp* 44.5 °C). The reactivity of the complex hydrides corresponds to the stability, i.e., to the decomposition temperature. Thus, $NaBH_4$ (decomposition temperature > 400 °C) reacts relatively slowly with water (see Section 3.4), whereas $Al(BH_4)_3$ (decomposition temperature 100 °C) is instantly hydrolyzed and reacts explosively with moist air. The hydridoaluminates are very much less stable than the hydridoborates and therefore considerably more reactive, e.g.,

$NaBH_4$ is a much milder reducing agent than $NaAlH_4$ (decomposition temperature 210 °C).

The properties of the complex hydrides can also be varied by the introduction of substituents on the boron or aluminum atom. An example is lithium tri-*tert*-butoxyaluminum hydride (see Section 3.9). This is a much milder reducing agent than $LiAlH_4$ because of the steric hindrance affecting the single remaining hydride ligand. Solubility is also affected by substitution. Thus, whereas $NaAlH_4$ is completely insoluble in hydrocarbons, it becomes soluble in toluene when alkyl groups are introduced (see Section 3.12).

The physical properties of the important commercially available complex hydrides are listed in Table 2. Their reducing properties towards various organic functional groups are given in Table 3. The variation in reducing properties and solubility (see Tables 4 and 5) represents a wide choice of reagents with which many difficult reductions can be achieved. Reductions with complex hydrides are rapid under mild conditions, give good yields of pure products, and are often remarkably selective; this accounts for their wide field of application in the chemical industry. Without these highly reactive and selective reagents the multi-step syntheses of natural products or pharmaceuticals would be inconceivable. The literature on the use of complex hydrides in preparative organic chemistry is unusually extensive, and there are good review articles and monographs on the subject [5, 39].

Table 3. Reducing properties of various hydrides towards functional groups

	Aldehydes	Ketones	Epoxides	Acyl chlorides	Esters	Anhydrides	Carboxylic acids	Acyl amides	Nitriles	Nitro compounds	Alkyl halides	Tosylhydrazones	Alkenes	Alkynes
$LiBH_4$	+	+	+	+	+	+	(+)	(+)	−	−	−		−	−
$LiBH(C_2H_5)_3$	+	+	+	+	+			+	+	+	+	+	−	−
$NaBH_4$	+	+	−	+	−	−	−	−	−	−	(+)	−	−	−
$NaBH_3CN$	(+) ∗	(+) ∗	−		−			−	−	−	+	+	+	
$LiAlH_4$	+	+	+	+	+	+	+	+	+	+	+	+	(+)	(+)
$LiAlH(OC(CH_3)_3)_3$	+	+	−	+	(+)	−	−	−	−	−	−	−	−	−
$NaAlH_2(OC_2H_4OCH_3)_2$	+	+	+	+	+	+	+	+	(+)	(+)	+		−	−
$NaAlH_2(C_2H_5)_2$	+	+	+	+	+	+	+	+	+	(+)	(+)		−	−
Cp_2ZrHCl∗∗											+		+	+

+ Usually reducible; (+) sometimes reducible; − usually not reducible
∗ At pH 3–4 in protic solvents. ∗∗ Cp = Cyclopentadienyl, C_5H_5.

Table 4. Solubility of complex borohydrides in various solvents (g/100 g solvent at 25 °C unless otherwise stated)

Solvent	*bp* of solvent, °C	$LiBH_4$	$NaBH_4$	KBH_4	$NaBH_3CN$	$LiHBEt_3$
Water	100.0	20.9 [a]	55	19.3	217	dec.
Methanol	67.7	dec.	16.4 [b]	0.7	v. sol.	dec.
Ethanol	78.5	44	4.0 [b]	0.25	sol.	dec.
Isopropylamine	34.0	8 [c]	6.0 [d]		fairly sol.	dec.
Diethyl ether	36.0	4.3	insol.	insol.	insol.	
THF	65.0	22.5	0.1 [e]	insol.	36	v. sol.
Diglyme	162		5.5	insol.	17.6	
Toluene	111	insol.	insol.	insol.	insol.	
Ammonia	−33.3	77	104	20		
DMF	153		18.0 [f]	15 [f]		

[a] 0 °C, slow decomposition.
[b] 20 °C, decomposition.
[c] 10 °C.
[d] 28 °C.
[e] 20 °C.
[f] Dangerous decomposition possible at higher temperature.

Table 5. Solubility of complex aluminum hydrides in various solvents (g/100 g)

Solvent	*bp* of solvent, °C	$LiAlH_4$ *	$LiAlH(OC(CH_3)_3)_3$ *	$NaAlH_4$ **
Diethyl ether	36	39	2	insol.
THF	65	15	43	16
Ethylene glycol dimethyl ether	85	7	4	28
Diethylene glycol dimethyl ether	162	5	41	13
Triethylene glycol dimethyl ether	216	3		15

* At 25 °C.
** At 20 °C.

3.2. Lithium Tetrahydridoborate

Properties (see also Table 2). Lithium tetrahydridoborate (known industrially as lithium borohydride or lithium boranate), $LiBH_4$, is a white, microcrystalline powder that decomposes at the melting point (280 °C) with considerable evolution of hydrogen. It is stable in vacuum to about 200 °C. It is extremely hygroscopic, being rapidly decomposed by moist air, sometimes with spontaneous ignition. It is therefore stored and handled in an atmosphere of nitrogen or argon. It is hydrolyzed only slowly by water at 0 °C:

$$LiBH_4 + 2\,H_2O \longrightarrow LiBO_2 + 4\,H_2$$

By acidification of the solution or by addition of certain metal salts as catalysts, the hydrolysis is rapid and large volumes of hydrogen are liberated (4.11 L/g at STP). Lithium borohydride is a more powerful reducing agent than sodium borohydride although much less reactive than lithium aluminum hydride.

Lithium borohydride is used for the selective reduction of esters, carboxylic acids, amides, and epoxides in preference to other functional groups such as nitriles, nitro compounds, and halides [5]. The effect of solvents has been investigated [40]; diethyl ether and tetrahydrofuran are preferred. The addition of a stoichiometric amount of methanol leads to a very large increase in reducing power without significantly affecting the selectivity. Among the reactions of lithium borohydride with inorganic compounds [1,5,41], special mention should be made of reactions yielding borohydrides or hydrides of other metals, e.g.,

$$4\,LiBH_4 + ZrCl_4 \longrightarrow Zr(BH_4)4 + 4\,LiCl$$

Production. Industrial production of lithium borohydride makes use of the heterogeneous metathetical reaction between LiCl or LiBr and sodium borohydride in ethers (diethyl ether, tetrahydrofuran) or amines:

$$NaBH_4 + LiX \longrightarrow LiBH_4 + NaX$$

X = Cl, Br

The mixture is ball-milled while the reaction takes place in order to ensure a satisfactory rate of reaction and a good yield (up to 90 %) [42]. Synthesis from the elements is not used industrially.

The commercial product (Chemetall, Ventron) has a bulk density of 0.6 – 0.7 g/cm^3 and a purity of 94 %.

Uses. Lithium borohydride is used to a limited extent as a selective reducing agent in organic chemistry. It has been investigated as a means of storing hydrogen, but has not been used on an industrial scale for this purpose [43].

3.3. Lithium Triethylhydridoborate

Lithium triethylborohydride, Super-Hydride, $LiBH(C_2H_5)_3$, is the most frequently used alkyl substituted complex borohydride. It is produced by reaction of lithium hydride and triethyl borane in THF solution at room temperature:

$$LiH + B(C_2H_5)_3 \longrightarrow LiBH(C_2H_5)_3$$

In the pure state it is a colorless, viscous liquid that is oxidized and hydrolyzed in moist air; it may ignite on contact with water or alcohols.

It is commercially available in limited quantities as a 12 % solution in THF under the name Super-Hydride (Aldrich). This solution is quite stable provided that air and moisture are excluded [17]. This reagent possesses considerably greater nucleophilicity than, say, lithium

aluminum hydride, and is therefore used for the reductive substitution of electronegative functional groups, enabling the rapid and quantitative reduction of alkyl halides to the corresponding hydrocarbons to be carried out [23]. Another reaction with excellent yields is the reductive deoxygenation of tosylates (toluenesulfonates) [44].

The steric hindrance of the hydrogen atom linked to boron by the ethyl groups leads to reactions with marked regio- and stereoselectivity.

3.4. Sodium Tetrahydridoborate

Properties (see also Table 2). Sodium tetrahydridoborate, sodium borohydride, sodium boranate, $NaBH_4$, is a white, slightly hygroscopic salt that is stable in dry air up to 300 °C. It decomposes slowly in a vacuum at 400 °C, quickly above 550 °C, with liberation of hydrogen; *mp* (under 1 MPa H_2) 505 °C. If ignited with a naked flame it burns quietly. (For fire precautions see Chapter 6). It can be stored in an airtight container for many years without decomposition. Sodium borohydride absorbs water from moist air to form the dihydrate, $NaBH_4 \cdot 2\,H_2O$, which slowly decomposes to form sodium metaborate and hydrogen. In acidic media rapid hydrolysis takes place:

$$BH_4^- + H^+ + 3H_2O \longrightarrow H_3BO_3 + 4\,H_2;$$

$\Delta H = -372$ kJ/mol.

Sodium borohydride, in keeping with its ionic character, is readily soluble in polar solvents, especially protic compounds such as water, lower alcohols, and amines (see Table 4).

With a few solvents it forms isolable though not very stable solvates [37]. In protic solvents, solvolysis occurs, and hydrogen is evolved:

$$NaBH_4 + 4\,ROH \longrightarrow NaB(OR)_4 + 4\,H_2$$

Thus, sodium borohydride is 80 % decomposed after 1 h in methanol at 0 °C, 6 % decomposed after 1 h in ethanol at 60 °C, but is stable in 2-propanol and *tert*-butanol and at room temperature. The solubility and stability in water, the most frequently used solvent, have been extensively investigated [45] (see Fig. 4). Protonolysis takes place in aqueous solution, the rate increasing rapidly with increasing temperature and decreasing pH [1]. These solutions become more stable with increasing concentration of sodium borohydride. At 24 °C, a 10^{-2} M solution (pH = 9.56) decomposes to the extent of 15 % in 1 h, whereas a 0.10 M solution (pH = 10.05) undergoes only 5 % decomposition in the same time. On boiling or on acidification, however, rapid hydrolysis occurs and all the combined hydrogen is released (2.36 L/g at STP). A number of transition metal salts and finely divided metals such as nickel, cobalt, iron, and copper [1,46] catalyze the hydrolysis, but the addition of strong bases such as NaOH can stabilize solutions of $NaBH_4$ (Fig. 5).

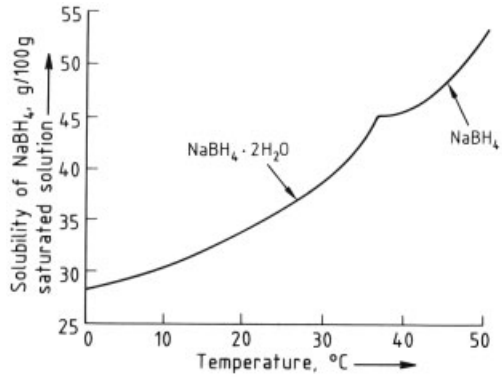

Figure 4. Solubility of sodium borohydride in water as a function of temperature
The transition temperature is 36.4 °C.

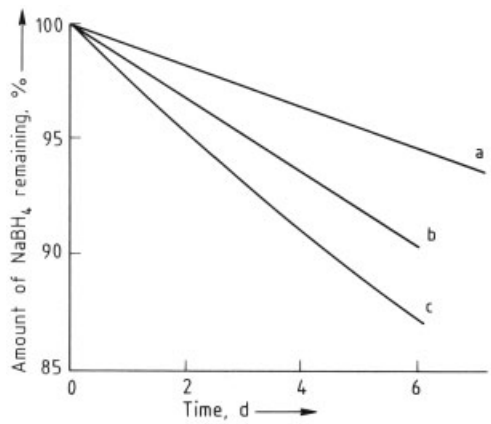

Figure 5. Stability of sodium borohydride in alkaline solution (0.1 M $NaBH_4$, 24 °C)
a) 1.0 M NaOH; b) 0.25 M NaOH; c) 0.10 M NaOH.

A solution of sodium borohydride in concentrated sodium hydroxide can be stored virtually without decomposition, although concentrated solutions of sodium borohydride in dimethylformamide must not be heated because of the danger of explosion caused by autocatalytic decompositions; such solutions should not even be stored for long periods at room temperature [47]. Contact with strong acid (e.g., concentrated sulfuric or phosphoric acid, or methanesulfonic acid) or Lewis acids must be avoided as these substances can liberate diborane, which is highly toxic and spontaneously flammable.

Sodium borohydride is a powerful reducing agent with wide applications in inorganic and particularly organic chemistry. The redox potential E^0 of the half reaction

$$BH_4^- + 8\,OH^- \longrightarrow B(OH)_4^- + 4\,H_2O + 8\,e^-$$

at pH 14 is − 1.24 V.

An aqueous solution of sodium borohydride will reduce a number of cations and anions, the product obtained depending on the nature of the ion and the reaction conditions. A metal ion of lower valency can be produced (e.g., iron(II) from iron(III)), a metal (e.g., all the platinum metals, cadmium, mercury, lead, silver, and gold), a metallic boride (CrB_2, Co_2B, Ni_2B), or a hydride (GeH_4, SnH_4, AsH_3, BiH_3). Many of these reductions are used for the recovery of noble metals, the precipitation of toxic metals from water, the preparation of finely divided metals, and the electroless deposition of metals and metallic borides onto metallic and nonmetallic substrates [48, 49].

Another reaction of industrial importance is the reduction of sulfur dioxide or sodium hydrogen sulfite to sodium dithionite:

$$NaBH_4 + 8\,NaOH + 8\,SO_2 \longrightarrow 4\,Na_2S_2O_4 + NaBO_2 + 6\,H_2O$$

The reducing power of sodium borohydride towards some important functional groups is shown in Table 3. This reveals that $NaBH_4$ is a milder but also more selective reducing agent than $NaAlH_4$, allowing the reduction of aldehydes, ketones, acid chlorides, and imines in the presence of esters, epoxides, amides, nitriles, or nitro groups. Much research has been undertaken with the object of varying the reducing power over a wide range [50], e.g., by mixing sodium borohydride with metal salts ($Pd(NO_3)_2$, $AlCl_3$, $CoCl_2$, $NiCl_2$), thiols (e.g., ethanedithiol), or carboxylic acids (e.g., acetic acid, trifluoroacetic acid), by using different solvents (e.g., aprotic solvents or solvent mixtures) [51], or by using phase transfer catalysts [52].

A solution of $NaBH_4 - TiCl_4$ (ca. 3 : 1) in 1,2-dimethoxyethane reduces carboxylic acids to alcohols and nitro compounds to amines. The $NaBH_4 - CoCl_2$ system dissolved in lower alcohols or dioxane is also effective for these reductions, as well as the reduction of sulfoxides to sulfides and the hydrogenation of alkenes. Lithium chloride and sodium borohydride dissolved in diethylene glycol dimethyl ether form the intermediate $LiBH_4$ which, unlike $NaBH_4$, can reduce esters and epoxides. Sodium triacetoxyborohydride, $NaBH(OAc)_3$, produced in situ from $NaBH_4$ and three equivalents of acetic acid, exhibits such a weakening of reducing power that aldehydes are selectively reduced without ketones being affected [53]. Sodium borohydride and boron trifluoride in diglyme or THF yield diborane:

$$3\,NaBH_4 + 4\,BF_3 \cdot OEt_2 \longrightarrow 2\,B_2H_6 + 3\,NaBF_4 + 4\,Et_2O$$

Diborane is used mainly for hydroboration, but also acts as a strong electrophilic reducing agent, reducing carboxylic acids and nitriles but not acyl chlorides or nitro groups [54].

Sodium borohydride is also an important starting material for the production of tetrahydridoborates (borohydrides) of other metals, and of boranes and organoboron compounds, their derivatives and adducts [1]. Some examples of practical importance are:

$$MX_n + n\,NaBH_4 \longrightarrow M(BH_4)_n + n\,NaX$$

MX_n = LiCl, KOH, NR_4OH, $CaCl_2$, $ZnCl_2$, etc.

$$ML_mX_n \xrightarrow{NaBH_4} ML_mH_yX_{n-y}$$

M = Cr, Fe, Co, Ni, Rh, Pt, etc.; L = PPh_3, X = Cl, Br

$$NaBH_4 + (R_2NH_2)Cl \longrightarrow R_2NH \cdot BH_3 + NaCl + H_2$$

Production. Since the discovery of sodium borohydride in 1942 by H. J. Schlesinger, H. C. Brown, H. R. Hoeckstra, and L. R. Rapp [55], over 100 methods of preparation have been described, but few of these have achieved any practical significance. Industrial production in the United States started

in 1954 using the Schlesinger process [56], in which trimethyl borate reacts with an extremly fine dispersion of sodium hydride in a high boiling hydrocarbon oil at 250 – 280 °C and atmospheric pressure:

$$4\,NaH + B(OCH_3)_3 \longrightarrow NaBH_4 + 3\,NaOCH_3$$

Water is then added, causing separation of the hydrocarbon oil, which is recirculated. The aqueous phase contains the $NaBH_4$ together with caustic soda and methanol formed by hydrolysis of $NaOCH_3$. The methanol is recovered by distillation and used again for production of trimethyl borate. The product solution contains 12 % $NaBH_4$ and 42 % NaOH ($\varrho = 1.4\,g/cm^3$; $\eta_{23} = 79\,mPa \cdot s$) and is sometimes marketed as Sodium Borohydride – SWS by Ventron or alkaline aqueous solution by Nokia. Solid sodium borohydride is obtained from the alkaline aqueous solution by extraction with isopropylamine; the yield is better than 90 %.

A process developed by Bayer [57] (see Fig. 6) uses finely ground borosilicate glass, sodium, and hydrogen:

$$Na_2B_4O_7 \cdot 7\,SiO_2 + 16\,Na + 8\,H_2 \longrightarrow 4\,NaBH_4 + 7\,Na_2SiO_3$$

The borosilicate $Na_2B_4O_7 \cdot 7\,SiO_2$ is produced by fusion of borax and silica. The borosilicate is cooled, ground, and then reacted with sodium in an atmosphere of hydrogen at 300 kPa and 400 – 500 °C in a partly heterogeneous reaction. The sodium borohydride is extracted from the borohydride – silicate mixture with liquid ammonia under pressure. The yield is in excess of 90 %. Other processes are described in Refs. [1,5], and [57].

Commercial sodium borohydride is in powder – lump form (bulk density 0.4 – 0.5 kg/L), the purity being typically 98 % (Bayer, Chemetall, Ventron).

Analysis. Two main methods are used for the analysis of sodium borohydride: one gas volumetric and the other titrimetric. In the first of these, the sodium borohydride is reacted with acid and the evolved hydrogen measured. The second method is equally straightforward and requires no special analytical equipment. The sodium borohydride is converted to metaborate by reaction with excess oxidizing agent such as sodium hypochlorite or potassium iodate:

$$NaBH_4 + 4\,NaOCl \longrightarrow NaBO_2 + 4\,NaCl + 2\,H_2O$$

The excess oxidizing agent is determined by iodometric back-titration [48,49].

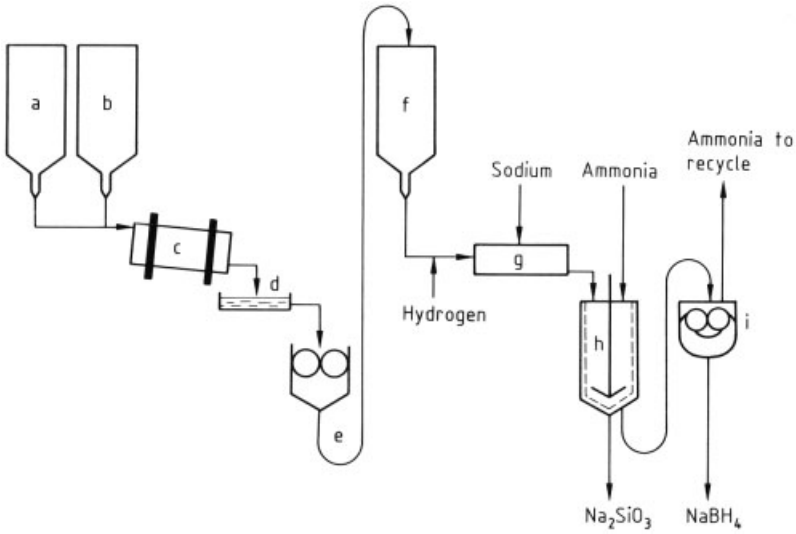

Figure 6. Flow diagram for production of sodium borohydride (Bayer process)
a) Borax reservoir; b) Quartz sand container; c) Oven; d) Quenching trough; e) Grinder; f) Borosilicate glass container; g) Reactor; h) Extractor; i) Drier.

Uses. The present annual consumption of sodium borohydride is estimated to be 2000–3000 t, mainly in the form of the 12 % solution. This means that it is by far the most important commercially available complex hydride. The largest amount is used in the paper and pulp industries, mainly in the United States and Scandinavia. The 12 % solution (Ventron, Nokia) is used for the in situ production of sodium dithionite, which is used to bleach wood pulp. Both sodium dithionite (made from sodium borohydride and sulfur dioxide) and sodium borohydride itself are used for reducing vat dyes to the leuco bases [58]. A patent claims that the reaction with sulfur dioxide can be used for the purification of exhaust gases from power stations burning fossil fuel [59].

Solid sodium borohydride is used as a reducing agent in the organic chemical and pharmaceutical industries, and although the quantities used for this purpose are considerably smaller than the demands of the paper industry, the selective reduction of aldehydes and ketones has become of major importance [24]. One of the biggest uses is in the stereoselective reduction of steroid ketones. Another is the synthesis of antibiotics such as chloramphenicol, dihydrostreptomycin, and thiophenicol, and also vitamin A. Other examples are reduction in the synthesis of prostaglandins, atropine, scopolamine, and flavorings and aromas.

Sodium borohydride is also used for cleaning and stabilization of process streams of many organic products. It chemically reduces harmful impurities such as aldehydes, ketones and peroxides, thereby improving color, odor, and the stability of the products towards heat, light, and oxidation. The technique is used in the production and subsequent treatment of epoxides, alcohols, amines, ethers, esters, carboxylic acids, alkenes, polymers [58], and paper [60]. The reduction of metal cations already mentioned is used on a large scale for the electroless deposition of metals and metal borides onto metals and other materials such as glass, ceramics, and plastics (“ Nibodur 1098” – Bayer [61], for the recovery of mercury from wastewater from the chlor-alkali electrolytic process (Ventron process, Koertrol) of lead from lead tetraalkyl production, and of silver and cadmium from spent photographic developing solutions [62, 63]. Finely disperse metals of the iron group for high quality magnetic tapes are produced by sodium borohydride reduction [58, 64]. Nickel boride and cobalt boride obtained by the reduction of the appropriate salt solutions are used as catalysts in gas- and liquid-phase reduction [58]. One final application for sodium borohydride depends on its ability to generate hydrogen gas: it is used as a foaming agent for organic polymers and inorganic construction materials [58, 65].

3.5. Sodium Cyanotrihydridoborate

Properties (see also Table 2). Sodium cyanotrihydridoborate, sodium cyanoborohydride, sodium cyanoboranate, $NaBH_3CN$, is a white, amorphous powder, melting with decomposition at 240–242 °C. Contact with the atmosphere must be avoided as it is extremely hygroscopic. It is very soluble in polar, especially protic, solvents such as water, alcohols, amines, and THF, but not in diethyl ether or hydrocarbons [66].

Production. Sodium cyanoborohydride can be prepared by reaction of sodium borohydride with anhydrous hydrogen cyanide with liberation of hydrogen. The use of free hydrogen cyanide can be avoided by reaction of $BH_3 \cdot THF$ (formed in situ from BF_3 and $NaBH_4$) with sodium cyanide in suspension in THF [67]. The yield is 59–71 %:

$$BH_3 \cdot THF + NaCN \longrightarrow NaBH_3CN + THF$$

Another method is the reaction of $NaBH_4$ with $Hg(CN)_2$ [68]. The commercial product contains 95 % $NaBH_3CN$ (Aldrich, Ventron).

Uses. Sodium cyanoborohydride is used as a selective reducing agent in the synthesis of pharmaceuticals. Its special reducing properties (see Table 3) arise from its remarkable resistance to hydrolysis. It is stable even at pH 3 in water and methanol, its rate of hydrolysis being slower than that of $NaBH_4$ by a factor of 10^8 [66]. Reactivity and stereoselectivity are considerably affected by the nature and acidity of the solvent. In polar aprotic solvents such as hexamethylphosphoric triamide (HMPT) or dimethyl sulfoxide (DMSO) it reduces aliphatic halides and toluenesulfonates to the corresponding hydrocarbons. Under these conditions other

reducible functional groups such as epoxides, ketones, and aldehydes are unaffected. In combination with ZnI_2, sodium cyanoborohydride deoxygenates certain aldehydes and ketones under very mild conditions (20 h at room temperature in dichloromethane) [69].

In protic solvents, preferably at pH 3 – 4, carbonyl compounds and derivatives such as oximes and imines can be reduced. At pH 6 the reaction with carbonyl groups practically ceases, whereas imines are reduced smoothly. This is a good method for the reductive amination of a ketone or aldehyde in the presence of an amine [70]. Because of its resistance to hydrolysis, sodium cyanoborohydride is used for the chemical precipitation of metals from neutral or even acidic solutions of their salts (e.g., those of Ni, Co, Cr, Cd, Cu, and Ag) [71].

3.6. Potassium Tetrahydridoborate

Properties (see also Table 2). Potassium tetrahydroborate, potassium borohydride, potassium boranate, KBH_4, is a white salt that is completely stable in dry air. It is not hygroscopic and hydrolyzes only extremely slowly in moist air (1.7 % in 48 d). Potassium borohydride is thermally more stable than the sodium salt, but is generally much less soluble (see Table 4).

Potassium borohydride ignites in air above 300 °C. It decomposes slowly at 500 °C in vacuum . In a hydrogen atmosphere it melts at ca. 585 °C. Aqueous solutions are relatively stable, but decompose completely on boiling, liberating hydrogen (1 g of KBH_4 liberates 1.66 L of H_2 at STP). The addition of alkali renders the solution stable up to 100 °C [1].

Production. Potassium borohydride is produced on an industrial scale by a metathetical reaction of sodium borohydride:

$$NaBH_4 + KOH \longrightarrow KBH_4 + NaOH$$

A methanol – water solution of sodium borohydride is treated with a stoichiometric amount of aqueous potassium hydroxide. The precipitated potassium borohydride is centrifuged, washed, and dried under vacuum [72]. The commercial product is a crystalline powder with a purity of 98 % (Goerrig-Chemie, Ventron).

Uses. Potassium borohydride is used to a limited extent as a selective reducing agent in preparative organic chemistry. Its reducing power is only slightly weaker than that of sodium borohydride.

3.7. Aluminum Hydride

Aluminum hydride [*7784-21-6*], alane, AlH_3, is a colorless, nonvolatile substance that exists as a polymer with Al – H – Al 3-center, 2-electron bonds. It is produced by reaction of lithium aluminum hydride with aluminum chloride or acid in diethyl ether [3]:

$$3\,LiAlH_4 + AlCl_3 \longrightarrow 4\,AlH_3 + 3\,LiCl$$

$$LiAlH_4 + HX \longrightarrow AlH_3 + LiX + H_2$$

$$HX = HCl,\ H_2SO_4$$

Monomeric alane etherate solution is first formed, which soon polymerizes, but the polyalane produced still contains ether, which cannot be removed even by heating under high vacuum. A solvent-free product may be obtained by reaction with excess lithium aluminum hydride [73]. Ether-containing products are decomposed by contact with water or by heating to 80 – 90 °C, but the ether-free products are more stable.

Aluminum hydride is a rather milder reducing agent than lithium aluminum hydride. It has been proposed for a large number of applications, including rocket fuel and for coating objects with a thin layer of aluminum; nevertheless, it is not commercially available. Many adducts and derivatives of aluminum hydride are known of which two have been employed industrially: polyiminoalanes are used as cocatalysts in the stereospecific polymerization of isoprene [74], and diisobutylaluminum hydride [*1191-15-7*] is used as a reducing agent [75] (Schering AG, Ethyl Corporation, Texas Alkyls) (→ Aluminum Compounds, Organic). The production of amine – alane complexes in hydrocarbons has been reported [76].

3.8. Lithium Tetrahydridoaluminate

Properties (see also Table 2). Lithium tetrahydridoaluminate, lithium aluminum hy-

dride, lithium alanate, $LiAlH_4$, is a fine crystalline powder. It is stable in the absence of moisture, but after some months at room temperature it becomes gray due to traces of elemental aluminum. This extremely slow decomposition hardly affects the reactivity of the product. On heating in the absence of air, decomposition starts at 125 °C with liberation of hydrogen. The first product is Li_3AlH_6, which itself decomposes above 200 °C into LiH, Ae, and H_2; the ultimate decomposition products are LiAl and H_2 (430 °C) [77]. If $LiAlH_4$ is heated in air, the liberated hydrogen ignites and burns quietly, and the metallic residue then burns with incandescence, whereby nitrogen also takes part in the reaction (see Chapter 6 for fire precautions). Moist air causes $LiAlH_4$ to lose reactivity as a result of hydrolysis, but under normal conditions it is not spontaneously flammable. It reacts extremely violently with water. The liberated hydrogen (2.3 L/g at STP) usually ignites spontaneously because of the large heat of hydrolysis ($\Delta H^0_{298} = -714$ kJ/mol).

Lithium aluminum hydride is soluble in various ethers (see Table 5), but not in hydrocarbons. It forms very stable solvates with some ethers and with tertiary amines; these solvates can very often be isolated. Some solvates are soluble in hydrocarbons. The solvate containing 1.6 mol of THF per mole of lithium aluminum hydride dissolves in toluene to form a 1.1 M lithium aluminum hydride solution [78]. The chelate complexes with tertiary polyamines are also soluble in hydrocarbons [79]. Solutions of $LiAlH_4$ are usually cloudy due to the presence of traces of insoluble oxygen-containing compounds and other byproducts that arise during production. Nevertheless, quality is not affected.

Concentrated solutions of $LiAlH_4$ are most conveniently obtained by the concentration of dilute solutions rather than by dissolving the solid compound, as the rate of dissolution is low. All operations with $LiAlH_4$ should be carried out in an inert atmosphere.

Lithium aluminum hydride is a powerful nucleophilic reducing agent. It reduces almost all functional groups (see Table 3), although isolated double and triple bonds in alkenes and alkynes are generally not attacked. There is a very extensive literature on the subject of $LiAlH_4$ reductions; for monographs and reviews, see [5,23,39,80,81]. A typical reaction is the reduction of a carboxylic acid:

$$4\,R{-}COOH + 3\,LiAlH_4 \longrightarrow (R{-}CH_2O)_4AlLi + 2\,LiAlO_2 + 4\,H_2$$

$$(R{-}CH_2O)_4AlLi + 4\,H_2O \longrightarrow 4\,R{-}CH_2OH + LiAl(OH)_4$$

One hydride equivalent is required for reaction with the acidic hydrogen atom and two further equivalents for the reduction of the carbonyl group. Essentially three methods are available for the hydrolysis of the aluminate complex and of the excess hydride [78]:

1) Hydrolysis with the theoretical amount of water (4 mol water per mole of $LiAlH_4$). The byproduct $LiAl(OH)_4$ precipitates and can be removed by centrifugation or filtration.
2) Hydrolysis with an excess of water (10–20 mol H_2O per mole $LiAlH_4$) and addition of sodium hydroxide or potassium carbonate solution [82]; the hydroxides form a granular or gelatinous precipitate. The product solution can then be decanted or siphoned off.
3) Hydrolysis with dilute mineral acid. The hydroxides dissolve and form an aqueous phase that can easily be separated from the organic phase.

The optimum hydrolysis conditions depend on the substrate and should be established in the laboratory for each reduction.

The powerful reducing action and consequent poor selectivity of $LiAlH_4$ can be modified by means of various additives. The addition of different amounts of $AlCl_3$ leads to the formation of the electrophilic chloroalanes and AlH_3, e.g.,

$$LiAlH_4 + AlCl_3 \longrightarrow 2\,AlH_2Cl + LiCl$$

These are milder and therefore more selective reducing agents than pure $LiAlH_4$ [5,81]. Nevertheless, chloroalanes encourage hydrogenolysis because of their electrophilic character. Thus, under certain conditions, ketones, carboxylic acids, esters, and alcohols can be converted into hydrocarbons. Reduction with mixtures of $LiAlH_4$ and $AlCl_3$ must be carried out in diethyl ether because THF is cleaved by this reagent.

The reducing power of $LiAlH_4$ can also be modified by the introduction of alkoxy substituents [5,81,83]:

$$LiAlH_4 + n\,ROH \longrightarrow LiAlH_{4-n}(OR)_n + n\,H_2$$

The alkoxyalanates so obtained are considerably more selective than $LiAlH_4$, and are able, for example, to reduce keto groups without reducing ester groups also present in the molecule. Alkoxyalanates can be prepared in situ; lithium (tri-*tert*-butoxy)hydridoaluminate (Section 3.9) is commercially available. Amines can also be used to modify the reducing properties of $LiAlH_4$. One example is the reduction of esters to aldehydes in the presence of diethylamine [84].

In the presence of chiral alcohols, amines, or aminoalcohols, prochiral ketones and aldehydes are reduced to optically active alcohols, often with very good selectivity [85]. The capacity for asymmetric reduction is mainly of interest for the pharmaceutical industry [86]. In the presence of catalytic quantities of $TiCl_4$, alkenes can be smoothly converted by $LiAlH_4$ to the corresponding alkanes [87, 88]. Hydroalumination is an alternative to hydroboration.

The reaction of $LiAlH_4$ with metal halides yields hydrides and the corresponding tetrahydroaluminates of the metals concerned [3, 38]. The preparation of transition metal tetrahydridoaluminate complexes using $LiAlH_4$ is described in a review [89].

Production. Lithium aluminum hydride was first synthesized in 1945 by FINHOLD, BOND, and SCHLESINGER [90]:

$$4\,LiH + AlCl_3 \longrightarrow LiAlH_4 + 3\,LiCl$$

The present industrial process uses this method. A solution of aluminum chloride in ether is slowly added to a suspension of finely divided lithium hydride containing some $LiAlH_4$. The heat of reaction keeps the mixture at its boiling point. After completion of the reaction, the lithium chloride and excess lithium hydride are filtered off and the ether is removed under reduced pressure. Only 25 % of the lithium is used to form the desired product. The economics can be improved by electrolysis of the lithium chloride byproduct to give lithium metal, which can be hydrogenated to lithium hydride, or used for other purposes (→ Lithium and Lithium Compounds).

Other preparative methods not used in industry are direct synthesis from lithium hydride, aluminum, and hydrogen under pressure in THF:

$$2\,LiH + 2\,Al + 3\,H_2 \longrightarrow 2\,LiAlH_4$$

and the metathetical reaction between sodium aluminum hydride and lithium chloride in diethyl ether [38]:

$$NaAlH_4 + LiCl \longrightarrow LiAlH_4 + NaCl$$

The usual commercial product (Chemetall, Ventron) is typically 97 % $LiAlH_4$. For ease of handling it is also supplied in tablet form or in bags that are soluble in diethyl ether, THF, or toluene [78]. Ready-to-use $LiAlH_4$ solutions are also available for industrial scale reductions, e.g., 10 % in THF, 20 % in diethyl ether, and 15 % as the THF complex in toluene [78]. The solutions are supplied in containers with dip tubes so that they can be transferred to reaction vessels without coming into contact with the atmosphere.

Analysis [78]. The analysis of solid $LiAlH_4$ is best carried out by the gas volumetric method. A weighed sample is cautiously hydrolyzed in a mixture of dioxane and water and the volume of hydrogen released is measured. Solutions of $LiAlH_4$ are conveniently analyzed by titration (Felkin's method [78]). The $LiAlH_4$ is reacted with an excess of iodine which is then back-titrated with sodium thiosulfate.

$$LiAlH_4 + 2\,I_2 \longrightarrow 2\,H_2 + LiAlI_4$$

Uses. Bulk lithium aluminum hydride is used almost exclusively for the reduction of organic functional groups. The principal consumers are the organic chemical, pharmaceutical, and the flavoring and perfumery industries. Lithium aluminum hydride is more expensive than sodium borohydride and is mainly used when the reducing power of the latter is inadequate.

Lithium aluminum hydride is used in industry mainly for the reduction of carboxylic acids and esters to alcohols, and of amides, imides, and nitriles to amines. Its powerful reducing properties are also used in the reduction of sterically hindered compounds. It has the advantage over some other reducing agents of being very clean to use as the precipitated sludges from the hydrolysis of the aluminate complex are easily separated and there are no toxic or gaseous byproducts. Furthermore, it has the highest hydrogen

content per unit mass of the commercially available complex hydrides. Lithium aluminum deuteride, $LiAlD_4$ [*14128-54-2*] (Chemetall), is also commercially available and is used for isotopic labelling of organic compounds with deuterium.

3.9. Lithium (tri-tert-butoxy)hydridoaluminate

Properties (see also Table 2). Lithium (tri-*tert*-butoxy)hydridoaluminate, $LiAlH[OC(CH_3)_3]_3$, is a white powder that sublimes at 280 °C and 200 Pa, and melts with decomposition at 319 °C. Its high thermal stability is consistent with its milder reducing properties compared with $LiAlH_4$ [83]. It can be handled in air without danger but is slowly hydrolyzed, thereby losing activity. Solubility data are given in Table 5. Lithium (tri-*tert*-butoxy)hydridoaluminate is produced by the reaction of $LiAlH_4$ with *tert*-butanol in ether solution:

$$LiAlH_4 + 3\,(CH_3)_3COH \longrightarrow LiAlH[OC(CH_3)_3]_3 + 3\,H_2$$

The commercial product has a typical purity of 98 % (Chemetall).

It is very suitable for selective reductions on account of its mild reducing power (see Table 3), the most important of these being the reduction of steroid ketones [23]. In addition to chemical selectivity, a certain stereoselectivity is shown in the reduction of cyclic ketones.

3.10. Sodium Tetrahydridoaluminate

Properties (see also Table 2). Sodium tetrahydridoaluminate, sodium aluminum hydride, sodium alanate, $NaAlH_4$, is a white, crystalline substance, *mp* 178 °C; it decomposes at 210 °C, liberating hydrogen and initially forming Na_3AlH_6 [77]. Sodium tetrahydridoaluminate is insoluble in hydrocarbons and diethyl ether. Its chemical properties resemble those of lithium aluminum hydride. However, the strong polar character of sodium aluminum hydride promotes side reactions during some reductions [5].

Production. Sodium aluminum hydride can be produced by reaction of sodium hydride with aluminum chloride in THF [91]:

$$4\,NaH + AlCl_3 \longrightarrow NaAlH_4 + 3\,NaCl$$

A more economical method is the direct synthesis from the elements:

$$Na + Al + 2\,H_2 \longrightarrow NaAlH_4$$

The reaction is carried out at elevated temperature (100 – 200 °C) and pressure (10 – 20 MPa) preferably in hydrocarbon solvent. The presence of a catalyst such as triethyl aluminum is necessary [92–94]. Sodium aluminum hydride is supplied as a laboratory reagent by Ethyl Corporation; however, it has not yet been produced on an industrial scale.

3.11. Sodium bis(2-methoxyethoxy)dihydridoaluminate

Properties [95] (see also Table 2). Sodium [bis(2-methoxyethoxy)]dihydridoaluminate (SDMA), $NaAlH_2(OCH_2CH_2OCH_3)_2$, is a colorless, viscous liquid at room temperature that solidifies to a glass below 0 °C; there is no sharp melting point. It is stable up to 170 °C and decomposes above 214 °C. At higher temperature decomposition is sudden and spontaneous.

The presence of the two methoxyethoxy groups confers solubility not only in ethers but also in aromatic hydrocarbons. It is completely miscible with toluene at concentrations below 6 % and above 42 %. Viscous solutions of SDMA can be handled in contact with air, although atmospheric moisture causes hydrolysis.

The reducing properties of SDMA are comparable with those of $LiAlH_4$ (see Table 3). For example, the reaction with aldehydes takes place as follows:

$$2\,R-CHO + NaAlH_2(OCH_2CH_2OCH_3)_2 \longrightarrow NaAl(R-CH_2O)_2(OCH_2CH_2OCH_3)_2 \xrightarrow{H_2O} 2\,R-CH_2OH + 2\,HOCH_2CH_2OCH_3 + NaAl(OH)_4$$

The complex aluminate intermediate is often soluble, so that the reduction takes place in a homogeneous phase and usually gives good yields. Hydrolysis can be carried out with acid or base. Most of the 2-methoxyethanol

[*109-86-4*] formed dissolves in the aqueous phase. 2-Methoxyethanol has a harmful effect on the central nervous system (MAK 5 ppm). In comparison to reductions with $LiAlH_4$, twice the amount of unwanted precipitate (sodium aluminum hydroxide) is formed, which can trap some of the reduced product and make work-up more difficult.

Production. One method of preparation is the controlled alcoholysis of $NaAlH_4$ with 2-methoxyethanol:

$$NaAlH_4 + 2\,CH_3OCH_2CH_2OH \longrightarrow NaAlH_2(OCH_2CH_2OCH_3)_2 + H_2$$

On an industrial scale, however, it is produced by direct synthesis from sodium, aluminum, and 2-methoxyethanol at 140 – 155 °C in an atmosphere of hydrogen at at least 7 MPa:

$$Na + Al + 2\,CH_3OCH_2CH_2OH \longrightarrow NaAlH_2(OCH_2CH_2OCH_3)_2$$

The quantity of hydrogen required is generated in situ by the reaction of sodium and aluminum with the alcohol. SDMA is marketed as a 70 % solution in toluene ($\varrho = 1.036$ g/cm^3) under the name Vitride (Hexcel).

Uses. The most important application of SDMA is as a reducing agent in the production of fine chemicals, but it is also used as a catalyst for the polymerization of lactams and alkenes, and for the trimerization of organic isocyanates. The use of SDMA as an activator in the preparation of Grignard reagents has been described [96].

3.12. Sodium Diethyldihydridoaluminate

Properties (see also Table 3). Sodium diethyldihydridoaluminate, $NaAlH_2(C_2H_5)_2$, is a white, crystalline substance, *mp* 85 – 87 °C, which ignites spontaneously in air. It is insoluble in aliphatic but fairly soluble in aromatic hydrocarbon solvents. The solubility of sodium diethyldihydridoaluminate in toluene can be considerably increased by addition of a small amount of THF. Solutions of this reagent must be kept under an inert atmosphere, since they lose their activity by hydrolysis and oxidation if exposed to air. The reducing power of this hydride is comparable with that of $LiAlH_4$ (see Table 3). The ethyl groups generally do not take part in the reduction, but give rise to large volumes of ethane when the intermediate aluminate complex is hydrolyzed. This ethane should be safely disposed of, preferably by burning.

Production. Sodium diethyldihydridoaluminate is produced by reaction of sodium, aluminum, and triethyl aluminum in toluene at ca. 150 °C and 7 MPa hydrogen pressure.

$$3\,Na + Al + 2\,Al(C_2H_5)_3 + 3\,H_2 \longrightarrow 3\,NaAlH_2(C_2H_5)_3$$

Improved yields and reduced reaction times are obtained by the addition of small amounts of the final product as a catalyst, and by splitting the process up into several stages [97]. Until 1987 the product was marketed in industrial quantities as a 25 % solution under the name OMH-1 (Ethyl Corp.).

The solubility of alkali metal dialkyldihydridoaluminates in aromatic hydrocarbons can be markedly increased if the two alkyl groups are different [98]. Thus, a 40 % solution of sodium butylethyldihydridoaluminate, $NaAlH_2(C_2H_5)_{1.5}(n\text{-}C_4H_9)_{0.5}$, is available in pilot-plant quantities (Schering).

4. Recommended Techniques for Working with Hydrides

It is quite possible to work with hydrides on an industrial scale without any problems provided that certain safety aspects of plant design are borne in mind [78, 99].

Almost all the binary and complex hydrides described here are highly reactive towards water, and therefore any possibility of accidental contact, e.g., by failure of a water-cooled reflux condenser, must be strictly avoided. If it is not possible to use inert coolants (such as mineral oil or decalin), pressure-tested metal reflux condensers must be installed, and the use of automatic water alarm systems is recommended.

All hydrides are sensitive to air and moisture to a greater or lesser extent, and they must therefore be handled in an inert atmosphere such as nitrogen or argon. By using such an atmosphere throughout the whole production plant,

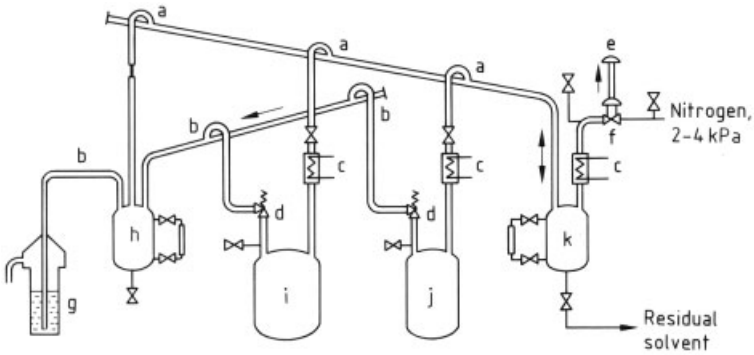

Figure 7. Inert gas breathing system recommended for working with metal hydrides
a) Breathing system, 2 – 4 kPa; b) Safety system; c) Condenser; d) Safety valve; e) Flashback arrestor; f) Respiration pressure regulator; g) Oil-immersion valve; h) Safety buffer vessel; i) Tank 1; j) Tank 2; k) Buffer breathing vessel, 2 – 4 kPa.

the risk of formation of explosive hydrogen – air mixtures, e.g., from the hydrolysis of reducing agents, can be largely avoided. A scheme for an inert gas breathing system in which all parts of the equipment are linked together is shown in Figure 7. Each vessel is connected by a pipeline to the breather manifold (a), which is connected via a buffer vessel and an efficient reflux condenser (c) to the breather regulator (f). The reflux condenser should be operated at as low a temperature as possible, to recover as much solvent vapor as possible from the inert gas, thereby minimizing emissions. In addition to the inert gas breathing system, a safety vent system (b) is provided. This is constructed in a similar way. All breathing and safety pipelines enter the common lines from above to prevent solvent carryover or possible transfer of product into other vessels by foaming.

Since flammable solvents are usually used in reduction and metallation reactions, the entire production area must be well ventilated to prevent the accumulation of solvent vapor – air mixtures; electrical installations must be explosion-proof.

A typical flow diagram for reductions with complex hydrides is shown in Figure 8 (see next page). The arrows show the route for a normal reduction procedure, i.e., that in which the substrate solution is passed into a solution of the reducing agent. If inverse addition is required, the reactor (b) can be used for dissolving the hydride. The hydrolysis of the intermediate complex need not be carried out in the hydrolyzer (c); it can be done in the reactor. However, the latter must be carefully dried before each new addition of reactants.

5. Hydrides of the Transition Metals

5.1. Binary Hydrides

Hydrogen reacts at elevated temperature with many transition metals and their alloys to form hydrides. The electropositive elements are the most reactive, i.e., scandium, yttrium, the lanthanides, the actinides, and the members of the titanium and vanadium groups. In contrast to the ionic hydrides, the binary hydrides of the transition metals are predominantly metallic in character and are usually referred to as metallic hydrides. They are good conductors of electricity, possess a metallic or graphite-like appearance, and can often be wetted by mercury. Many of these compounds (MH_n) show large deviations from ideal stoichiometry ($n = 1, 2, 3, \ldots$) and can exist as multi-phase systems. The lattice structure is that of a typical metal with atoms of hydrogen embedded in it; for this reason they are also called interstitial hydrides.

This type of structure has the limiting compositions MH, MH_2, and MH_3; the hydrogen atoms fit into octahedral or tetrahedral holes in the metal lattice, or a combination of the two types. The hydrogen carries a partial negative charge, depending on the metal (an exception is, for example, $PdH_{0.7}$) [100].

Only a small number of the transition metals are without known stable hydrides. The elements of group 3 (scandium, yttrium, actinium),

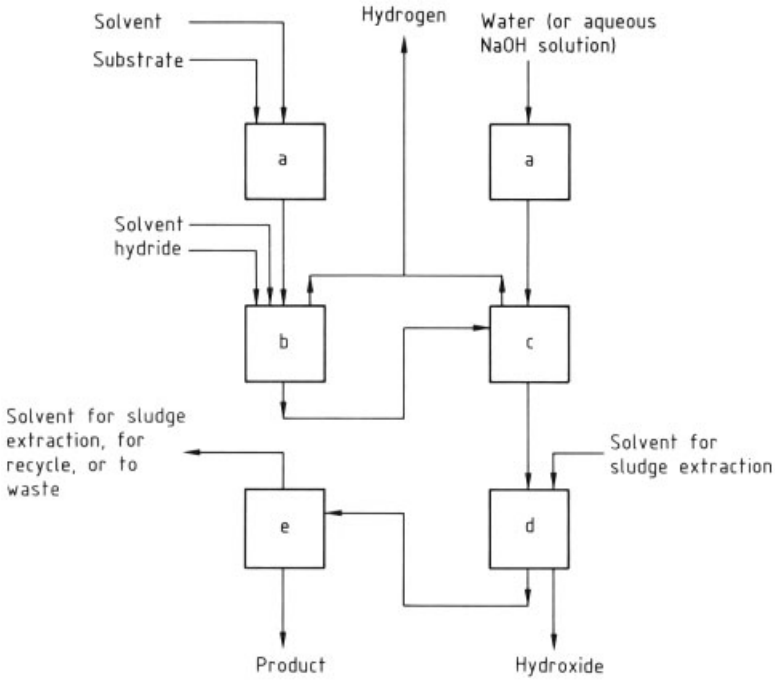

Figure 8. Simplified flow diagram for reductions with metal hydrides
a) Metering vessel; b) Reactor; c) Hydrolysis vessel; d) Evaporator; e) Filter or centrifuge.

the lanthanides, and the actinides have electronegativities comparable with those of the alkaline earths, and all react exothermically with hydrogen to form stable (defect) "dihydrides," which can mostly be converted into "trihydrides." The metals of group 4 (titanium, zirconium, hafnium) all form stoichiometric dihydrides whose enthalpies of formation are as high as those of the alkaline earth metal hydrides. Of the metals of group 5, vanadium and niobium are known to form mono- and dihydrides, while tantalum forms only a hydride of the limiting formula $TaH_{0.9}$. A considerable "hydride gap" exists in the periodic table, beginning at group 6, in which the only hydrides are palladium hydride ($PdH_{0.7}$), the very unstable nickel hydride ($NiH_{0.7}$), and the poorly defined hydrides of chromium (CrH, $CrH_{1.8}$) and copper (CuH). Nevertheless, there are many known complex-stabilized hydrides of these elements (see Section 5.3). In palladium hydride, the hydrogen has high mobility and probably a very low charge density. In the finely disperse state, platinum and ruthenium are able to adsorb considerable quantities of hydrogen, which thereby becomes activated. These two elements, together with palladium and nickel, are extremely good hydrogenation catalysts, although they do not form hydrides [4].

One of the most interesting features of the metallic hydrides is the extremely high density of the hydrogen atoms present. This is higher than it is in liquid hydrogen, e.g., in ZrH_2 by 70 %, in UH_3 by 93 %, and in TiH_2 by as much as 113 %. These compounds (e.g., TiH_2 and ZrH_2) have attracted great interest in the nuclear industry because hydrogen is a very powerful neutron moderator and because some metallic hydrides are both chemically stable and mechanically strong. According to the neutron absorption cross section of the metals, they can be suitable for use as moderators, reflectors, or shielding materials [4]. In the metallurgical industry, hydrogenation of the metals of groups 3, 4, and 5, and of thorium and uranium is used in metal powder production and in processing scrap metal. Hydrogen treatment reduces the density, ductility, and toughness, facilitating mechanical size reduction. After grinding, the product is degasified under vacuum, yielding a

powdered metal that is easily sintered and generally suitable for powder metallurgy [101].

5.1.1. Titanium Hydride

Properties (see also Table 6). The stoichiometric dihydride $TiH_{2.00}$ has a metallic appearance and crystallizes in a face-centered cubic structure of the fluorite type, which changes to a tetragonal structure below room temperature. The commercial product is a gray, finely divided powder with the composition $TiH_{1.95}$. The titanium – hydrogen system has been described in detail [102, 103]. One phase of the formula $TiH_{1.38}$ represents the most stable composition. The eutectic between α-Ti and TiH_2 (γ-phase) melts at 300 °C and contains ca. 38 atom% hydrogen. At atmospheric pressure, TiH_2 slowly loses hydrogen at ca. 350 °C. About 8 % of the theoretical hydrogen content is lost at 500 °C, 15 % at 600 °C, and 62 % at 800 °C. Only at the melting point of titanium (1668 °C) is the dissociation virtually complete. Titanium hydride possesses chemical resistance similar to that of the metal. At room temperature titanium hydride is completely unaffected by air and water. It is slowly attacked by most strong acids, but is readily soluble in hydrofluoric and hot sulfuric acids.

Table 6. Physical data for transition metal hydrides

Formula	CAS registry number	M_r	Density, g/mL	$H_{B\,(298)}$, kJ/mol
TiH_2	[*7704-98-5*]	49.92	3.76	− 124
ZrH_2	[*7704-99-6*]	93.24	5.61	− 163
Cp_2ZrHCl	[*37342-97-5*]	257.87		
$HRhCO(PPh_3)_3$	[*17185-29-4*]	918.81		

Titanium hydride reacts with oxidizing agents extremely vigorously, especially in a finely disperse state at elevated temperature; it ignites readily, and is therefore used in pyrotechnic mixtures. Dust explosions are possible. Powdered titanium hydride can be handled in air at room temperature and atmospheric pressure without danger; is is less sensitive than titanium powder of similar particle size. Titanium hydride ignites when heated in air above 400 °C. First, the liberated hydrogen burns quietly, and then the titanium reacts to form the oxide and nitride with a bright incandescence. To prevent ignition by friction, titanium hydride should only be ground when wet or under an inert atmosphere.

Production. Titanium hydride is produced industrially by hydrogenation of titanium sponge at atmospheric pressure and 300 – 500 °C. The exothermic absorption of hydrogen begins above 200 °C, is very rapid at 400 – 500 °C, and is completed during cooling in the hydrogen atmosphere. The hydrogenated sponge is brittle and easily ground to a fine powder. The commercial product has a particle size of 5 µm and a bulk density of ca. 1.4 g/cm^3 (Degussa, Ventron). The product contains 94.5 – 95.5 % titanium, 3.8 – 4.0 % hydrogen, and 0.1 – 1.0 % oxygen. An even finer product is produced by the Hydrimet process, in which titanium dioxide is reduced with calcium hydride at 600 – 700 °C to form titanium metal. This is then hydrogenated in situ at 300 – 400 °C and the byproduct calcium oxide is removed by dissolution in dilute hydrochloric acid [104].

Uses. Titanium hydride is mainly used in the powder metallurgical industry to promote good pressing and sintering. It is added to powdered metals because during sintering it generates a protective hydrogen atmosphere, which reduces the oxide film on the metal surface, combines with and removes carbon, and increases the density of the sintered product. It is used as an additive (2 – 5 %) in the production of Alnico and Ticonal sintered magnets for high performance loudspeakers and for sintered boride materials [105]. Titanium metal powder is produced industrially by heating ground titanium hydride to 600 °C and applying a vacuum. When this reaches 10^{-3} Pa the hydrogen is completely removed. Titanium nitride in extremely finely divided form is produced by reaction with ammonia or ammonium halides at atmospheric pressure and 400 – 700 °C. Titanium nitride is used as an ultra-hard, heat-resistant material [106].

Titanium hydride can also be used to assist ceramic materials such as glass and porcelain to bond with each other or with metals [107]. On heating with the hydride, a strongly bonded coating of titanium is formed on the nonmetallic surface. This can then be bonded to metals using a solder (e.g., silver – copper).

Titanium hydride is also used for making metallic foams [108] and as a surface coating on titanium alloys for corrosion protection against

sea water or acidic solutions [109]. Investigations are continuing into the possible use of shielding materials containing titanium hydride for fusion reactors of the Tokamak type [110].

5.1.2. Zirconium Hydride

Properties (see also Table 6). At 300 °C, pure zirconium metal reversibly absorbs hydrogen until the stoichiometric dihydride $ZrH_{2.00}$ (ε-phase) is formed. This compound crystallizes in a body-centered tetragonal structure. The reaction takes place in several stages, forming the metastable γ-phase (ZrH), then the stable δ-phase ($ZrH_{1.59-1.67}$), and finally the ε-phase with the highest hydrogen content ($ZrH_{1.67-2.00}$). The eutectic is at 550 °C and 33 atom% H [4, 102]. Zirconium hydride is thermally more stable than titanium hydride. At atmospheric pressure, it dissociates at 500 °C. At 600 °C it has lost ca. 6 % of its theoretical hydrogen content, at 800 °C 19 %, and at 900 °C 50 %. Hydrogen-free zirconium powder is produced by the dehydrogenation of powdered ZrH_2 in vacuum (10^{-3} Pa) at 800 °C. The chemical properties of zirconium hydride and titanium hydride are very similar, although zirconium hydride is much more sensitive towards oxidizing agents. It is inert to air and water at room temperature. The usual commercial product, a grayish black powder with particle size 2 – 6 µm, ignites in air at ca. 250 – 300 °C. The hydride should therefore only be ground in an inert gas atmosphere (argon) or when wetted with water.

Production. Zirconium hydride that conforms to various quality requirements is produced either by direct hydrogenation of metallic zirconium or by the reduction of zirconium dioxide with calcium hydride, calcium, or magnesium in a hydrogen atmosphere (Degussa, Ventron).

Zirconium sponge is hydrogenated at 300 °C mainly for the production of the pure hydride with a low hafnium content for the nuclear industry (96.9 – 97.9 % zirconium, 1.9 – 2.1 % hydrogen, <0.01 % hafnium, ≤0.14 % other metals), and also for the production of zirconium metal powder for the powder metallurgical industry.

The reduction of purified ZrO_2 with calcium hydride yields a very fine, pure hydride powder (96.5 – 97.5 % zirconium, 1.9 % hydrogen; particle size 2 – 3 µm), which is quite suitable for most applications:

$$ZrO_2 + 2\,CaH_2 \longrightarrow ZrH_2 + 2\,CaO + H_2$$

Technical quality zirconium dioxide is reduced by calcium/magnesium in a hydrogen atmosphere to give a relatively cheap hydride powder quite suitable for use in pyrotechnics (91.5 – 93.5 % zirconium, 1.3 – 1.5 % hydrogen, particle size 2 – 3 µm).

Uses. Zirconium hydride is used in the nuclear industry as a moderator for thermal neutrons, especially in light water reactors and fast breeder reactors. It can be formed by warm pressing and is characterized by low neutron absorption and high thermal stability and mechanical strength [111]. Very effective shielding materials contain a combination of a neutron absorber (e.g., boron carbide or zirconium) with a moderator such as zirconium hydride [112]. Zirconium hydride is also widely used as a getter in vacuum and ultra-high vacuum technology, e.g., for pipework and electric lights; it is able to bind up to 40 % oxygen or up to 20 % nitrogen [113]. Like titanium hydride, zirconium hydride can provide a bond between nonmetallic materials (e.g., SiN and SiC) and metals. Zirconium hydride is added at a concentration of 0.01 – 5 % to a matrix of silver, copper, or nickel alloys to form reactive brazing compounds [114].

Zirconium hydride is a component of pyrotechnic formulations (with potassium chlorate as the oxygen source) for signal flares, tracer munition, or primer compositions. The burning properties are determined by grain size, specific surface, and particularly hydride content. Increasing hydrogen content leads to slower burning rates [115].

5.2. Intermetallic Hydrides

There is increasing interest in the broad field of the ternary system $M^1 - M^2 - H$ of intermetallic compounds or alloys of transition metals and hydrogen. These compounds can sometimes contain more hydrogen than would be expected from the absorption properties of the individual metal

components. The alloys with the greatest capacity for hydrogen are of the AB (e.g., FeTi or EuRu) or AB_5 type (e.g., $LaNi_5$ or $CaNi_5$) [116].

The thermochemical processes associated with absorption and desorption can be used for heat pumps and refrigeration systems (→ Hydrogen).

5.3. Complex Hydrides of Transition Metals

Transition metals form a large group of hydrides stabilized by complex formation: $H_xM_nX_yL_z$ (L = neutral ligand, X = anionic ligand, M = central metal atom). In these mono- or polynuclear complexes the hydrogen is linked by covalent bonds either to one metal atom, or it acts as a bridge between two centers (e.g., in many polynuclear complexes and in tetrahydroborate complexes). Further bonds with neutral or anionic ligands (less frequently with cationic ligands such as NO^+) enable most complexes to attain the electronic configuration of the next inert gas (the 18 electron rule). Rhenium and technetium are the only transition metals known to form hydrido complexes containing only hydride ligands, i.e., $[ReH_9]^{2-}$ and $[TcH_9]^{2-}$. Most transition-metal hydride complexes are stabilized by π-bonding ligands, e.g., tertiary phosphines and arsines (PR_3, AsR_3, respectively), the cyclopentadienyl anion ($C_5H_5^-$), and carbon monoxide.

The first hydrido complexes were discovered by Hieber [117] in the early 1930s; these were the unstable carbonyl hydrides of iron, $[H_2Fe(CO)_4]$, and cobalt, $[HCo(CO)_4]$. In the 1950s, research activity increased greatly and led to the discovery of hydrido complexes of all the d-elements [6].

Hydrido complexes of cobalt, nickel, ruthenium, rhodium, and iridium are industrially important homogeneous catalysts for many reactions of alkenes and alkynes. These catalysts are usually prepared in situ and are rarely used in the form of isolated complexes. For example, a number of rhodium complexes can be used in the hydrogenation of alkenes, especially $RhCl(PPh_3)_3$ (Wilkinson's catalyst). Hydrogen oxidatively adds to this complex giving a dihydrido complex which then hydrogenates the alkene:

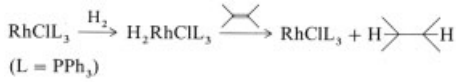

The reaction mechanism has been studied in detail [118].

The first industrial application of a hydrido complex as an intermediate dates from 1940 when Ruhrchemie operated a cobalt-catalyzed hydroformylation process (the Oxo process) on an industrial scale [6, 118] (→ Oxo Synthesis; see also → Carbonylation). The cobalt catalyst, introduced as $Co_2(CO)_8$, is converted under the reaction conditions into tetracarbonylcobalt hydride, which then catalyzes the reaction between an alkene and CO to form aldehydes:

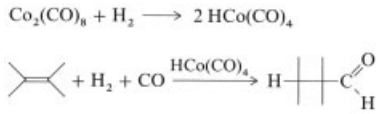

World output of C_3-C_{18} aldehydes by this method is ca. 100 000 t/a. Rhodium-catalyzed hydroformylation was developed in the mid 1970s, mainly by Union Carbide. The technical advantages include greater selectivity and conversion rate. The complex $HRh(CO)(PPh_3)_2$ is often used as catalyst [118, 119].

Complex hydrides are also important in isomerization, dimerization, and polymerization of dienes.

Stoichiometric reactions of metal hydrido complexes have also attained some importance. Hydrozirconation of carbon – carbon double and triple bonds with Cp_2ZrHCl (see Section 5.3.1) is chemo- and stereoselective and is a useful route for the functionalization of alkenes and alkynes [118].

A comprehensive literature exists on the extensive subject of hydrido complexes [6, 7, 118–121].

5.3.1. Di(η^5-cyclopentadienyl)chlorohydridozirconium(IV)

Properties (see also Table 6). Dicyclopentadienylzirconium hydride chloride, zirconocene hydride chloride, Schwartz's reagent, Cp_2ZrHCl ($Cp = C_5H_5$), is a white microcrystalline powder, *mp* >300 °C (dec.). Under the influence of light or heat the color changes to pink, then gradually to dark red. The compound

is sensitive to moisture; it is virtually insoluble in the usual organic solvents such as ethers, consistent with its polymeric character. It dissolves in halogenated solvents with evolution of heat, forming Cp_2ZrClX, where X = Cl, Br, or I.

Production. Schwartz's reagent is produced by reaction of zirconocene dichloride (Cp_2ZrCl_2) with $LiAlH_4$ or $LiAlH[OC(CH_3)_3]_3$ in THF solution [122]. The commercial product (Alfa Ventron, Aldrich) is slightly pink.

Uses. Schwartz's reagent is used in limited amounts in organic synthesis for the selective hydrozirconation of alkenes and alkynes. The reaction takes place rapidly and quantitatively at room temperature, forming zirconium alkyls or alkenyls that may be converted into useful functionalized derivatives. For example, reaction with carbon monoxide followed by removal of the zirconium-containing part of the molecule yields a variety of carbonyl compounds [118, 23].

5.3.2. Carbonylhydrido-tris(triphenylphosphine)rhodium(I)

Properties (see also Table 6).
Carbonylhydrido-tris(triphenylphosphine)rhodium(I), carbonyl-tris(triphenylphosphine)rhodium(I) hydride, forms yellow crystals that are stable in air, *mp* 172 – 174 °C in a closed nitrogen-filled capillary. It is readily soluble in benzene, chloroform, and dichloromethane. In solution, one phosphine ligand is readily lost by dissociation to give a 16-electron complex that is an active catalyst for the hydrogenation of alkenes and for hydroformylation [118].

Production. This complex is obtained by the reaction of a solution of $RhCl_3 \cdot 3\,H_2O$ in alcoholic alkali with aqueous formaldehyde (the carbon monoxide source) and an excess of triphenyl phosphine [123]. The commercial product is a microcrystalline powder with a purity of 97 % (Degussa).

Uses. Carbonylhydrido-tris(triphenylphosphine)rhodium(I) is used in industry for the hydroformylation of alkenes by the low pressure oxo (LPO) process developed by Union Carbide, Davy Power Gas, and Johnson Matthey. The reaction takes place rapidly at 25 °C giving a product in which the ratio of *n*- to *iso*-aldehydes is very favorable (ca. 20 : 1), a much higher ratio than can be obtained with cobalt catalysts [119].

6. Handling, Storage, and Transportation

Handling. Most of the alkali and alkaline earth metal hydrides and the complex hydrides of the boron group are unusually reactive and moisture sensitive. They should therefore only be handled by well trained and qualified personnel. On an industrial scale, the use of tight goggles, helmets with face protection, leather gloves, and special protective clothing is essential. Protective impregnated leather or aluminum-covered outer clothing is especially recommended. Sparks caused by static discharge should be prevented by the use of shoes with conducting soles. In the event of skin contact, the affected part should be treated with copious quantities of running water until pH paper gives a neutral reaction. If the eyes are affected, they should immediately be rinsed thoroughly with water; medical attention should be sought. Accidental ingestion should be treated by a physician.

The greatest potential source of danger from hydrides derives from their high hydrogen content. This gas is released in large quantities in a violent exothermic reaction when the easily hydrolyzable hydrides come into contact with water or other protic substances. In a closed vessel, dangerously high pressures can suddenly develop. In open air, gas mixtures are produced that can explode with great violence if ignited. Some hydrides are even more dangerous, being spontaneously flammable when finely divided material comes into contact with moist air, and thus provide a potential source of ignition for any explosive mixture of hydrogen, solvent, and air. Some commercially available hydrides tend to create dust on account of their small particle size and, if they are not sufficiently protected by an inert gas, can cause dangerous dust – air explosions. Fire-fighting equipment including dry powder extinguishers (based on sodium chloride) and breathing apparatus should be kept in readiness.

Smaller, localized fires can be extinguished with powdered calcium carbonate or by dry sand applied to the base of the flames with a shovel. Easily reachable dry powder sprinklers and emergency water basins for washing away alkaline spills are recommended for personal emergency use. Under no circumstances must water, carbon dioxide, or halogenated materials (such as carbon tetrachloride) be used.

Storage. The hydrides described here may be stored indefinitely in the absence of air and moisture, but see Sections 3.8 and 5.3.1.1 for conditions causing the decomposition of lithium aluminum hydride and Cp_2ZrClH, respectively. These chemicals should be stored in a cool, dry, well-ventilated area. Hydrides should not usually be kept in glass containers (with the exception of the transition metal hydrides) owing to the danger of accidental hydrogen pressure build-up leading to bursting of the container.

Transportation. Metal hydrides are classed as dangerous substances to which special transportation regulations apply (RID, ADR, IMDG). The alkali and alkaline earth hydrides and the complex boron and aluminum hydrides are in the Dangerous Goods Class 4.3 (materials that evolve flammable gases in the presence of water) and which includes Subgroups 2 b (solids) and 2 e (solutions). The 12 % alkali-stabilized solution of sodium borohydride is an exception, and is classified in Class 8 (caustic alkali solutions). Lastly, the transition metal hydrides of titanium and zirconium, which are stable to water, fall under Class 4.1 (flammable solids). Every country sets a limit to the amount of hydride per container, frequently 10 kg, that may be carried. Further details should be obtained from the producers. These hydrides may not be sent by mail.

7. Toxicology and Occupational Health

The alkali and alkaline earth metal hydrides, and the complex hydrides of group 13 are extremely reactive substances, which on contact with water or moist air give off hydrogen and are converted to the corresponding hydroxide. Considerable quantities of heat are often released, and in conjunction with the caustic effect of the alkaline hydroxides this leads to severe irritation of the skin and mucous membranes. Dust produced will attack the eyes and the respiratory tract. If ingested, these materials cause severe damage to the mouth and throat. If they were to reach the stomach the acid there would react rapidly even with those hydrides that are relatively resistant to hydrolysis, e.g., $NaBH_4$ and KBH_4, releasing large volumes of hydrogen, which can cause embolism.

Compared with the harmful effects resulting from the alkaline nature of these materials, the toxic properties of the associated metals (e.g., B, Li, Al) are of relatively small concern. An exception is beryllium hydride, since beryllium compounds are extremely poisonous and can also cause dermatitis and cancer [124]. Further information regarding these elements may be found in reference works [124–128] and the material safety data sheets issued by the hydride producers.

Special safety precautions must be observed when working with *sodium cyanoborohydride* as extremely toxic hydrogen cyanide (MAK 10 ppm) can be generated, especially in acidic solutions. Toxicological investigations have shown that aqueous $NaBH_3CN$ solutions do not cause skin or eye irritation, whereas solutions in THF are easily absorbed through the skin and have severe toxic effects. The oral LD_{50} lies between 10 and 100 mg/kg [129].

8. References

General References

1. Gmelin
2. A. F. Trotman-Dickenson (ed.): *Comprehensive Inorganic Chemistry,* Pergamon Press, Oxford 1973.
3. E. Wiberg, E. Amberger: *Hydrides of the Elements of Main Groups I – IV,* Elsevier, Amsterdam 1971.
4. W. M. Mueller, J. R. Blackledge, G. G. Libowitz (eds.): *Metal Hydrides,* Academic Press, New York 1968.
5. A. Hajos: *Complex Hydrides,* Elsevier, Amsterdam 1979.
6. G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* Pergamon Press, Oxford 1982.

7. R. Bau (ed.): "Transition Metal Hydrides," *Adv. Chem. Ser.* **167** (1978) 1 – 429.
8. E. L. Muetterties (ed.): *Transition Metal Hydrides,* Dekker, New York 1971.

Specific References
9. A. F. Wells: *Structural Inorganic Chemistry,* 5th ed., Oxford University Press, Oxford 1984.
10. P. A. A. Klusener, L. Brandsama, H. D. Verkruijsse, P. R. Schleyer et al., *Angew. Chemie* **98** (1986) 458 – 459.
11. R. Pi, T. Friedl, P. R. Schleyer, *J. Org. Chem.* **52** (1987) 4299 – 4303.
12. E. E. Shpilrain, K. A. Yakimovich: *Lithiumhydrid, seine physikalischen und thermophysikalischen Eigenschaften,* Staatsverlag für Normung, Moskau 1971.
13. C. D. Montgomery, *Nucl. Eng. Des.* **25** (1973) 309 – 314.
14. S. Liao, W. Kong, H. Liu, Y. Xu, Hydrogen Syst. Pap., Int. Symp. 1985 (Pub. 1986) 475 – 486; *Chem. Abstr.* **105** (1986) 193811 h.
15. Rhône-Poulenc Spécialités chimiques, EP 0 063 987-A1, 1981 (J. Grosbois, J. Y. Dumousseau).
16. Central Glass Co., DE-OS 35 06 071-A1, 1984 (K. Aono, T. K. Saito, C. K. Okade).
17. S. Krishnamurthy, *Aldrichimica Acta* **7** (1974) 55 – 60.
18. J. Plešek, S. Heřmánek: *Sodium Hydride,* Iliffe Books Ltd., London 1968.
19. P. Caubère, *Angew. Chemie* **95** (1983) 597 – 611.
20. W. F. Fedor, M. D. Banus, D. P. Ingalls, *Ind. Eng. Chem.* **49** (1957) 1664 – 1672.
21. N. L. Evans, *Corros. Technol.* **3** (1956) 45 – 51.
22. Tokuyama Soda, JP 60 238 406, 1984 (H. Yamada).
23. L. F. Fieser, M. Fieser: *Reagents for Organic Synthesis,* **vol. 1 – 13,** J. Wiley and Sons, New York 1967 – 1988.
24. A. Kleemann, J. Engel: *Pharmazeutische Wirkstoffe,* Thieme Verlag, Stuttgart 1982; Ergänzungsband, Thieme Verlag, Stuttgart 1987.
25. J. Buckley, R. L. Webb, T. Laird, R. J. Ward, *Chem. Eng. News,* **60** (July 12, 1982) 5, and G. Dewall, *Chem. Eng. News* **60** (Sept. 13, 1982) 5.
26. H. W. Pinnick, *Org. Prep. Proced. Int.* **15** (1983) 199 – 223.
27. Mine Safety, Appliances Co., US 4 327 071, 1982 (K. Chui, J. R. Strickler).
28. Ph. Eliardis, *J. Less Common Met.* **128** (1987) L7 – L8.
29. G. J. Brendel, E. M. Marlett, L. M. Niebylski, *Inorg. Chem.* **17** (1978) 3589 – 3592.
30. R. W. Baker, G. J. Brendel, B. R. Lowrance, J. R. Mangham et al., *J. Organomet. Chem.* **159** (1978) 123 – 130.
31. B. Bogdanovic, S. Lia, M. Schwickardi, P. Sikorsky, *Angew. Chem.* **92** (1980) 845 – 846.
32. Max Planck Institut für Kohleforschung, EP 0 003 564, 1978 (B. Bogdanovic).
33. Max Planck Institut für Kohleforschung, DE-OS 3 410 640-A1, 1984 (B. Bogdanovic).
34. B. Bogdanovic, B. Spliethoff, *Int. J. Hydrogen Energy* **12** (1987) 863 – 873.
35. B. Bogdanovic, *Angew. Chem.* **97** (1985) 253 – 264.
36. Max Planck Institut für Kohleforschung, DE-OS 3 247 361, 1982 (B. Bogdanovic).
37. B. D. James, M. G. H. Wallbridge: "Metal Tetrahydroborates," *Prog. Inorg. Chem.* **11** (1970) 99 – 231.
38. E. C. Ashby: "The Chemistry of Complex Aluminohydrides," *Adv. Inorg. Chem. Radiochem.* **8** (1966) 283 – 335.
39. *Houben-Weyl,* **4/1 d,** 1 – 487.
40. K. Soai, A. Ookawa, *J. Org. Chem.* **51** (1986) 4000 – 4005.
41. T. J. Marks, J. R. Kolb, *Chem. Rev.* **77** (1977) 265 – 293.
42. H. C. Brown, Y. Moon Choi, S. Narasimhan, *Inorg. Chem.* **20** (1981) 4456 – 4457.
43. Compagnie Francaise de Raffinage S.A., DE-OS 2 728 109, 1976 (A. Muller, F. Mathey, J. Bensoam).
44. R. W. Holder, M. G. Matturro, *J. Org. Chem.* **42** (1977) 2166 – 2168.
45. E. H. Jensen: *A Study on Sodium Borohydride, Nyt Nordisk Forlag,* Arnold Busck, Copenhagen 1954.
46. A. Levy, J. B. Brown, C. J. Lyons, *Ind. Eng. Chem.* **52** (1960) 211.
47. D. A. Yeowell, R. L. Seaman, J. Mentha, *Chem. Eng. News* **57** (Sept. 24, 1979) 4.
48. Bayer AG, Firmenschrift, *Natriumboranat,* 1975.
49. Ventron Corp., Firmenschrift *Sodium Borohydride,* 1979.
50. R. C. Wade, *J. Mol. Catal.* **18** (1983) 273 – 297.
51. R. S. Varma, G. W. Kabalka, *Synth. Commun.* **15** (1985) 985 – 990.
52. F. Rolla, *J. Org. Chem.* **46** (1981) 3909 – 3911.
53. G. W. Gribble, C. F. Nutaitis, *Org. Prep. Proced. Int.* **17** (1985) 317 – 384.

54. M. Follet, *Chem. Ind. (London)* 1986, 123 – 128.
55. H. C. Brown: *Boranes in Organic Chemistry,* Cornell Univ. Press, Ithaka 1972.
56. H. J. Schlesinger, H. C. Brown, *J. Am. Chem. Soc.* **75** (1953) 186 – 205; H. J. Schlesinger, H. C. Brown, US 2 534 533, 1945.
57. W. Büchner, H. Niederprüm, *Pure Appl. Chem.* **49** (1977) 733 – 743.
58. R. C. Wade, *Spec. Publ. R. Soc. Chem.* **40** (1981) 25 – 58.
59. Gea Wiegand GmbH, DE-OS 3 525 377, 1985 (U. Hochberg, G. Klinke, P. Sauglet).
60. L. C. Tang, *Adv. Chem. Ser.* **212** (1986) 427 – 441.
61. R. N. Duncan, T. L. Arney, *Plat. Surf. Finish.* **71** (1984) 49 – 54.
62. J. A. Ulman, *Spec. Publ. R. Soc. Chem.* **61** (1986) 173 – 196.
63. M. J. Lindsay, M. E. Hackmann, R. I. Tremblay, *Proc. Ind. Waste Conf.* **40** (1985) 477 – 482.
64. Mitsubishi Denki KK, DE-OS 3 704 075, 1986 (H. Ogama, Y. Fukotomi, M. Okutani).
65. Mosc. Eng. – Cons. Inst., SU 126 167, 1984 (A. P. Merkin, B. M. Runyantsev, L. E. Vitels).
66. C. F. Lane, *Aldrichimica Acta* **8** (1975) 3 – 10.
67. Ventron Corp., US 4 301 129, 1981 (R. C. Wade, B. C. Hui).
68. E. Jozsef, G. Bela, *J. Chem. Soc. Chem. Commun.* 1983, 1303 – 1304.
69. C. K. Lau et al., *J. Org. Chem.* **51** (1986) 3038 – 3043.
70. C. F. Lane, *Synthesis* 1975, 135 – 146.
71. Ventron Corp., US 3 637 472, 1972 (R. C. Wade).
72. W. S. Fedor, M. D. Banus, D. P. Ingalls, *Ind. Eng. Chem.* **49** (1957) 1664 – 1672.
73. F. M. Brower, N. E. Matzek, P. F. Reigler, H. W. Rinn et al., *J. Am. Chem. Soc.* **98** (1976) 2450 – 2453.
74. S. Cucinella, A. Mazzei, W. Marconi, *Inorg. Chim. Acta Rev.* **4** (1970) 51 – 71.
75. E. Winterfeldt, *Synthesis* 1975, 617 – 630.
76. Ethyl Corp., US 4 665 207, 1987 (E. M. Marlett).
77. J. A. Dilts, E. C. Ashby, *Inorganic Chem.* **11** (1972) 1230 – 1236.
78. Chemetall GmbH, Technical Information Brochure, *Lithium Aluminium Hydride. Industrial Use,* 1987.
79. T. A. Whitney, A. W. Langer, *Adv. Chem. Ser.* **130** (1974) 270 – 280.
80. J. S. Pizey: "Lithium Aluminium Hydride," *Synthetic Reagents* **vol. 1**, Ellis Horwood, Chichester 1977.
81. Chemetall GmbH, Technical Information Brochure, *Lithium Aluminium Hydride. Properties, Handling, Application,* 1984.
82. V. M. Micovic, M. L. Mihailovic, *J. Org. Chem.* **18** (1953) 1190 – 1200.
83. J. Malek, M. Cesny, *Synthesis* 1972, 217 – 234. J. Malek, *Org. React.* **34** (1985) 1 – 318. J. Malek, *Org. React.* **36** (1988) 249 – 590.
84. J. S. Cha, S. S. Kuon, *J. Org. Chem.* **52** (1987) 5486 – 5487.
85. E. R. Grandbois, S. I. Howard, J. D. Morrison, *Asymmetric Synth.* **2** (1983) 71 – 90.
86. Sumitomo Chemical, EP 3 535 5-B, 1984 (S. Terashima, K. Koga, N. Tanno).
87. F. Sato: "Hydroalumination Reactions with Lithium Aluminium Hydride," *IV. Hydrid-Symposium,* Chemetall GmbH, Frankfurt 1987, pp. 5 – 26.
88. Sumimoto Chemical, JP 580 351 29-A, 1983 (T. Suzukamo, T. Iguma, Y. Sakito, S. Tanaka).
89. A. R. Barron, G. Wilkinson, *Polyhedron* **5** (1986) 1897 – 1915.
90. A. E. Finhold, A. C. Bond, H. I. Schlesinger, *J. Am. Chem. Soc.* **69** (1947) 1199 – 1203.
91. V. V. Gavrilenko, M. I. Vinnikova, L. I. Zakharkin, *Zh. Obshch. Khim.* **49** (1979) 982 – 987.
92. Ethyl Corp., US 4 045 545, 1977 (E. C. Ashby).
93. Ethyl Corp., US 4 512 966, 1985 (G. E. Nelson).
94. Ethyl Corp., EP 0 157 887-A1, 1984 (G. E. Nelson).
95. O. Strauf, B. Casensky, V. Kubanek: *Sodium Dihydrido-bis(2-methoxyethoxo) aluminate (SDMA),* Elsevier, Amsterdam 1985.
96. J. Vit, US 3 758 620, 1969.
97. Ethyl Corp., DE-OS 25 44 120, 1975 (G. E. Nelson).
98. Schering AG, EP 0 199 089-A2, 1986 (R. Schrader).
99. W. Huter: "Verfahrenstechnische Aspekte für das Arbeiten mit Hydriden und metallorganischen Verbindungen," *IV. Hydrid-Symposium,* Chemetall GmbH, Frankfurt 1987, pp. 109 – 134.
100. G. R. Pearson, *Chem. Rev.* **85** (1985) 41 – 49.
101. R. Kiefer, G. Janegg, P. Ettmayer: *Sondermetalle,* Springer Verlag, Wien 1971.
102. D. G. Ivey, D. O. Northwood, *Mater. Sci. Technol.* **1** (1985) 600 – 602.
103. V. A. Livanov, A. A. Bukhanova, B. A. Kolachev: *Hydrogen in Titanium,* Israel Program for Scientific Translations, Jerusalem 1965.

104. Metal Hydrides Inc., US 2 427 338, 1945 (P. P. Alexander).
105. Toyo Kohan KK, JP-Kokai 6 0103148, 1983 (T. Takahashi, T. Tanaka, Y. Kondo).
106. Daicel Chem. Ind. KK, JP-Kokai 6 202 27312, 1985 (M. Hattori, S. Yamanaka, M. Ohashi).
107. Hydride Brazing: *Bulletin No. 37-A der Ventron Corp.,* 1971; Tokyo Inst. Technol., JP-Kokai 60 200 868, 1984 (T. Izeki).
108. Lor Corp., DE 1 408 468, 1972 (W. S. Fiedler).
109. Kobe Steel KK, JP-Kokai 61 019 785, 1984 (H. Sato, F. Kamikubo).
110. L. A. El-Guebaly, M. E. Sawan, *Fusion Technol.* **8** (1985) 1037 – 1042.
111. P. H. Horton, W. J. Kurzeka, *Nucl. Sci. Abstr.* **29** (1974) 7235. *Chem. Abstr.* **80** (1974) 140 032 p.
112. Hitachi KK, JP-Kokai 62095495, 1985 (K. Azekura, K. Kurihara, M. Bando).
113. W. Schreiter: *Seltene Metalle,* vol. 3, Verlag für Grundstoff industrie, Leipzig 1962.
114. Mitsui Eng. and Shipbuilding, JP-Kokai 59 141 465, 1983 (T. Degawa).
115. H. Baumgart: "Treib-Explosivst.," *Int. Jahrestag.* Fraunhofer-Inst., 1984, pp. 553 – 564; *Chem. Abstr.* **103** (1985) 73 412 r.
116. D. G. Ivey, D. O. Northwood, *Z. Phys. Chem. (Wiesbaden)* **147** (1986) 191 – 209; M. J. Benham, D. K. Ross, *Z. Phys. Chem. (Wiesbaden)* **147** (1986) 1219 – 229.
117. W. Hieber, F. Leutert, *Naturwissenschaften* **19** (1931) 360 – 361.
118. J. P. Collman, L. S. Hegedus: *Principles and Applications of Organotransition Metal Chemistry,* University Science Books, Mill Valley 1980.
119. F. J. Smith, *Platinum Met. Rev.* **19** (1975) 93 – 95.
120. T. J. Marks, J. R. Kolb, *Chem. Rev.* **77** (1977) 263 – 293.
121. D. S. Moore, S. D. Robinson, *Chem. Soc. Rev.* **12** (1983) 415 – 452.
122. P. C. Wailes, H. Weigold, *J. Organomet. Chem.* **24** (1970) 405 – 411; S. L. Buchwald et al., *Tetrahedron Lett.* **28** (1987) 3895.
123. N. Akmach, J. J. Levison, S. D. Robinson, M. F. Uttley, *Inorg. Synth.* **15** (1974) 59 – 60.
124. L. Friberg, G. F. Nordberg, V. B. Vouk (eds.): *Handbook on the Toxicology of Metals,* Elsevier, North Holland 1979.
125. N. I. Sax, R. J. Lewis: *Dangerous Properties of Industrial Materials,* 7th ed., V.N.R., New York 1989.
126. G. Weiss (ed.): *Hazardous Chemical Data Book,* Noyes Data Corp., Park Ridge, N.J. 1986.
127. R. Kühn, K. Briett: *Merkblätter Gefährlicher Arbeitsstoffe,* Verlag Moderne Industrie, Dummer, München 1974.
128. G. Hommel: *Handbuch der Gefährlichen Güter,* Springer Verlag, Berlin 1973, 1974.
129. Ventron Corp., *Sodium Cyanoborohydride,* Bull. No. 4 A, 1971.

Hydroboranation → Boron Compounds
Hydrobromic Acid → Bromine Compounds
Hydrocarbon Oxidation → Oxidation
Hydrocarbon Resins → Resins, Synthetic

Hydrocarbons

Acetylene, Anthracene, Benzene, Butadiene, Butenes, Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene, Cyclohexane, Cyclopentadiene and Cyclopentene, Ethylbenzene, Ethylene, Isoprene, Liquefied Petroleum Gas, Methane, Naphthalene and Hydronaphthalenes, Natural Gas, Propene, Styrene, Terpenes, Toluene, Waxes, Xylenes are separate keywords. Propyne → Acetylene

Karl Griesbaum, Universität Karlsruhe (TH), Karlsruhe, Federal Republic of Germany (Chap. 1)

Arno Behr, Henkel KGaA, Düsseldorf, Federal Republic of Germany(Chap. 2)

Dieter Biedenkapp, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 3.1– 3.7, Section 6.3 in part)

Heinz-Werner Voges, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Section 3.7– 3.10, Section 6.3 in part)

Dorothea Garbe, Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany (Section 3.11)

Christian Paetz, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 4, Section 6.4 in part)

Gerd Collin, Rüttgerswerke AG, Duisburg, Federal Republic of Germany(Chap. 5)

Dieter Mayer, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 6)

Hartmut Höke, Rüttgerswerke AG, Castrop-Rauxel, Federal Republic of Germany (Section6.5)

1. Saturated Hydrocarbons 3
1.1. Physical Properties 3
1.2. Chemical Properties 3
1.3. Production 4
1.3.1. From Natural Gas and Petroleum . 5
1.3.2. From Coal and Coal-Derived Products 8
1.3.3. By Synthesis and by Conversion of other Hydrocarbons 8
1.4. Uses 9
1.5. Individual Saturated Hydrocarbons 12
2. Olefins 14
2.1. Monoolefins 14
2.1.1. Properties 14
2.1.2. Production of Higher Olefins 16
2.1.2.1. Production from Paraffins 16
2.1.2.2. Oligomerization of Lower Olefins . 18
2.1.3. Uses of Higher Olefins 24
2.1.4. Economic Aspects of Higher Olefins 27
2.2. Dienes and Polyenes 27
2.2.1. Low Molecular Mass 1,3-Dienes . . 27
2.2.2. Synthesis of Dienes and Polyenes by Oligomerization 28
2.2.3. Synthesis of Dienes and Polyenes by Metathesis 29
3. Alkylbenzenes 29
3.1. Trimethylbenzenes 30
3.2. Tetramethylbenzenes 31
3.3. Penta- and Hexamethylbenzene . 31
3.4. Diethylbenzenes 32
3.5. Triethylbenzenes and more highly Ethylated Benzenes 32
3.6. Ethylmethylbenzenes (Ethyltoluenes) 33
3.7. Cumene 33
3.8. Diisopropylbenzenes 34
3.9. Cymenes; C_4- and C_5-Alkylaromatic Compounds 35
3.10. Monoalkylbenzenes with Alkyl Groups $>C_{10}$ 36
3.11. Diphenylmethane 37
4. Biphenyls and Polyphenyls 38
4.1. Biphenyl 38
4.2. Terphenyls 41
4.3. Polyphenyls. 42
5. Hydrocarbons from Coal Tar . . . 42
5.1. Acenaphthene 42
5.2. Indene 44
5.3. Fluorene 45
5.4. Fluoranthene 46
5.5. Phenanthrene 47
5.6. Pyrene 48
6. Toxicology and Occupational Health 49
6.1. Alkanes 49
6.2. Alkenes 50
6.3. Alkylbenzenes 50
6.4. Biphenyls and Polyphenyls 51
6.5. Hydrocarbons from Coal Tar . . . 51

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_227

6.5.1. Biological Effects 51
6.5.1.1. Carcinogenicity and Mutagenicity . 51
6.5.1.2. Mammalian Toxicity and Toxicokinetics 51
6.5.1.3. Ecotoxicology 52
6.5.2. Safety Regulations 52
7. References 52

1. Saturated Hydrocarbons

The class of saturated hydrocarbons comprises a myriad of individual compounds. A large number of the theoretically possible saturated hydrocarbons is known [1], although only a limited number of individual saturated hydrocarbons are used as raw materials in the chemical industry.

1.1. Physical Properties

Saturated hydrocarbons are colorless nonpolar substances, immiscible with polar solvents, but miscible with many nonpolar organic solvents. Some physical properties are given in Table 1. The boiling points and, starting with C_3, melting points of *n*-alkanes increase with increasing molecular mass. At 20 °C and atmospheric pressure, C_1 to C_4 *n*-alkanes are gases, C_5 to C_{16} are liquid, and from C_{17} solid. *i*-Alkanes do not display a definite correlation between the number of carbon atoms and their boiling- and melting points. The boiling points of *i*-alkanes all are lower than those of the corresponding *n*-alkanes; this is also true for many of the melting points. Furthermore, the boiling points of isomeric alkanes decrease with increasing degree of branching. The boiling points and melting points of cycloalkanes are generally higher than those of *n*-alkanes having the same number of carbon atoms (Table 1).

The density of the saturated hydrocarbons in the liquid state at 20 °C is <1 g/cm^3; it varies from 0.6 g/cm^3 for compounds with low carbon numbers to 0.8 g/cm^3 for compounds with high carbon numbers (Table 1). Additional physical properties are compiled in Tables 2, 3, 4, 5.

1.2. Chemical Properties

Alkanes and cycloalkanes are saturated, nonpolar, and lack functional groups; such hydrocarbons can undergo reaction only after cleavage of C−H or C−C bonds. Consequently, the scope of primary reaction steps is essentially limited to dehydrogenation, substitution, and chain- or ring cleavage. Saturated hydrocarbons cannot undergo addition reactions, which represent a versatile synthetic tool for unsaturated hydrocarbons (see Chap. 2). Most industrial reactions involving saturated hydrocarbons are radical reactions, e.g., thermal cracking, oxidation, sulfoxidation, halogenation, sulfochlorination, and nitration. Industrial ionic reactions of saturated hydrocarbons are restricted to acid-catalyzed processes with strong acids. Such reactions are employed mainly in the processing of petroleum by catalytic cracking, isomerization, and alkylation; they yield complex mixtures of products. Other ionic reactions of saturated hydrocarbons can be carried out with super acids (e.g., FSO_3H-SbF_5, $HF-SbF_5$): the alkylation of alkanes by alkanes [3], alkylation of benzene by alkanes [3], and ionic chlorination [5] and bromination [6] of alkanes have been explored, but are not used industrially (see also →Acylation and Alkylation.

Reactions of saturated hydrocarbons, because of the lack of functional groups, are nonselective, with respect to both the region of attack (regioselectivity) and the number of reaction sites (degree of substitution), unless the molecule possesses specific structural features, e.g., tertiary hydrogen atoms. Such reactions frequently afford mixtures of isomeric or structurally analogous compounds which can only be separated with difficulty or not at all.

On the basis of free enthalpy of formation (see Table 2) most of the saturated hydrocarbons are thermodynamically unstable with respect to the elements carbon and hydrogen. They are, however, kinetically stable at ambient temperature. Thermal decomposition of saturated hydrocarbons proceeds stepwise by loss of hydrogen or hydrocarbon fragments with concomitant formation of industrially useful unsaturated cracking products, such as acetylene, olefins, or aromatic hydrocarbons.

Table 1. Physical properties of saturated hydrocarbons [2]

Compound	Empirical formula	CAS registry number	*mp*, °C	*bp*, °C	η (20 °C), mPa · s	ϱ (20 °C), g/cm^3	n_D^{20}
Methane	CH_4	[*74-82-8*]	−182.5 [a]	−161.5			
Ethane	C_2H_6	[*74-84-0*]	−183.27 [a]	− 88.6			
Propane	C_3H_8	[*74-98-6*]	−187.69 [a]	− 42.1		0.5005 [b]	
Cyclopropane	C_3H_6	[*75-19-4*]	−127.4	− 32.8			
n-Butane	C_4H_{10}	[*106-97-8*]	−138.4	− 0.5		0.5788 [b]	1.3326 [b]
Cyclobutane	C_4H_8	[*287-23-0*]	− 90.7	12.5		0.6943 [b]	1.365 [a]
2-Methylpropane	C_4H_{10}	[*75-28-5*]	−159.6	− 11.7		0.5572 [b]	
n-Pentane	C_5H_{12}	[*109-66-0*]	−129.7	36.1	0.234	0.62624	1.35748
Cyclopentane	C_5H_{10}	[*287-92-3*]	− 93.9	49.3	0.438	0.74538	1.40645
2-Methylbutane	C_5H_{12}	[*78-78-4*]	−159.9	27.9	0.224	0.61967	1.35373
2,2-Dimethylpropane	C_5H_{12}	[*463-82-1*]	− 16.6	9.5		0.5910 [b]	1.342 [b]
n-Hexane	C_6H_{14}	[*110-54-3*]	− 95.3	68.7	0.3117	0.65937	1.37486
2-Methylpentane	C_6H_{14}	[*107-83-5*]	−153.7	60.3		0.65315	1.37145
3-Methylpentane	C_6H_{14}	[*96-14-0*]		63.3		0.66431	1.37652
2,2-Dimethylbutane	C_6H_{14}	[*75-83-2*]	− 99.9	49.7		0.64916	1.36876
2,3-Dimethylbutane	C_6H_{14}	[*79-29-8*]	−128.5	58.0		0.66164	1.37495
n-Heptane	C_7H_{16}	[*142-82-5*]	− 90.6	98.4	0.4169	0.68376	1.38764
2-Methylhexane	C_7H_{16}	[*591-76-4*]	−118.3	90.1		0.67859	1.38485
3-Methylhexane	C_7H_{16}	[*589-34-4*]		91.9		0.68713	1.38864
3-Ethylpentane	C_7H_{16}	[*617-78-7*]	−118.6	93.5		0.69816	1.39339
2,2-Dimethylpentane	C_7H_{16}	[*590-35-2*]	−123.8	79.2		0.67385	1.38215
2,3-Dimethylpentane	C_7H_{16}	[*565-59-3*]		89.8		0.69508	1.39196
2,4-Dimethylpentane	C_7H_{16}	[*108-08-7*]	−119.2	80.5		0.67270	1.38145
3,3-Dimethylpentane	C_7H_{16}	[*562-49-2*]	−134.46 [d]	86.1		0.69327	1.39092
2,2,3-Trimethylbutane	C_7H_{16}	[*464-06-2*]	− 24.9	80.9		0.69011	1.38944
n-Octane	C_8H_{18}	[*111-65-9*]	− 56.8	125.7	0.5450	0.70252	1.39743
2-Methylheptane	C_8H_{18}	[*592-27-8*]	−109.0	117.6		0.69792	1.39494
3-Methylheptane	C_8H_{18}	[*589-81-1*]	−120.5	118.9		0.70582	1.39848
4-Methylheptane	C_8H_{18}	[*589-53-7*]	−121.0	117.7		0.70463	1.39792
2,2,3-Trimethylpentane	C_8H_{18}	[*564-02-3*]	−112.3	109.8		0.71602	1.40295
2,2,4-Trimethylpentane	C_8H_{18}	[*540-84-1*]	−107.4	99.2	0.69193	1.39145	
2,3,3-Trimethylpentane	C_8H_{18}	[*564-02-3*]	−100.7	114.8		0.72619	1.40750
2,3,4-Trimethylpentane	C_8H_{18}	[*565-75-3*]	−109.2	113.5		0.71906	1.40422
2,2,3,3-Tetramethylbutane	C_8H_{18}	[*594-82-1*]	−100.7	106.5			
n-Nonane	C_9H_{20}	[*111-84-2*]	− 53.5	150.8	0.7139	0.71763	1.40542
n-Decane	$C_{10}H_{22}$	[*124-18-5*]	− 29.7	174.1	0.9256	0.73005	1.41189
n-Undecane	$C_{11}H_{24}$	[*1120-21-4*]	− 25.6	195.9	1.185	0.74024	1.41725
n-Dodecane	$C_{12}H_{26}$	[*112-40-3*]	− 9.6	216.3	1.503	0.74869	1.42160
n-Tridecane	$C_{13}H_{28}$	[*629-50-5*]	− 5.4	235.4	1.880	0.75622	1.42560
n-Tetradecane	$C_{14}H_{30}$	[*629-59-4*]	5.9	253.5	2.335	0.76275	1.42892
n-Pentadecane	$C_{15}H_{32}$	[*629-62-9*]	9.9	270.6	2.863	0.76830	1.43188
n-Hexadecane	$C_{16}H_{34}$	[*544-76-3*]	18.2	286.8	3.474	0.77344	1.43453
n-Heptadecane	$C_{17}H_{36}$	[*629-78-7*]	21.98	302.2	4.196 [c]	0.7779 [c]	1.4368 [c]
n-Octadecane	$C_{18}H_{38}$	[*593-45-3*]	28.2	316.7		0.7818 [c]	1.4389 [c]
n-Nonadecane	$C_{19}H_{40}$	[*629-92-5*]	31.9	330.6		0.7854 [c]	1.4408 [c]
n-Icosane	$C_{20}H_{42}$	[*112-95-8*]	36.4	343.8		0.7886 [c]	1.4425 [c]
n-Triacontane	$C_{30}H_{62}$	[*638-68-6*]	65.8	449.7		0.8096 [c]	1.4535 [c]
n-Tetracontane	$C_{40}H_{82}$	[*4181-95-7*]	81.5	525		0.8204 [c]	1.4592 [c]

[a] For saturation pressure (triple point).
[b] At saturation pressure.
[c] For the super-cooled liquid below the normal melting point.
[d] Value for the melting point of the metastable crystalline form.

On ignition, mixtures of saturated hydrocarbons and oxygen or air may lead to combustion or explosion, depending on the ratio of hydrocarbon and oxygen. Such reactions can be initiated either by imposed ignition or, if the ignition temperature of the mixture is exceeded, by self-ignition. These reactions are the basis for the use of hydrocarbons as heating- and engine fuel. Pertinent fuel properties of individual saturated hydrocarbons are summarized in Table 4.

Table 2. Molar enthalpies of fusion, vaporization and combustion [2]

	Molar enthalpy, kJ/mol		
	Fusion	Vaporization (25 °C)	Combustion [a]
Methane	0.942		802.861
Ethane	2.86	5.02	1 428.787
Propane	3.53	15.1	2 045.377
Cyclopropane	5.44		1 960.637
n-Butane	4.66	21.1	2 658.827
Cyclobutane	5.78	23.7	2 569.523
2-Methylpropane	4.54	19.1	2 650.454
n-Pentane	8.41	26.8	3 274.287
Cyclopentane	4.89	28.5	3 101.707
2-Methylbutane	5.15	24.8	3 266.248
2,2-Dimethylpropane	3.15	21.8	3 254.735
n-Hexane	13.1	31.7	3 886.81
2-Methylpentane	6.27	30.0	3 879.07
3-Methylpentane		30.4	3 881.71
2,2-Dimethylbutane	0.579	27.8	3 869.78
2,3-Dimethylbutane	0.80	29.2	3 877.86
n-Heptane	14.1	36.6	4 501.44
n-Octane	20.8	41.5	5 115.57
n-Dodecane	35.9	61.3	7 580.076
n-Hexadecane	51.9	81.1	10 040.616
n-Icosane	69.9	100.9 [b]	12 501.198

[a] For the reaction: Hydrocarbon (g) $\rightarrow CO_2$ (g) + H_2O (g).
[b] For the supercooled liquid, below the normal melting point.

Table 3. Enthalpy of formation, entropy, and free energy of formation for saturated hydrocarbons as gases (25 °C) [53]

	ΔH^0, kJ/mol	S^0, J mol^{-1} K^{-1}	ΔG^0, kJ/mol
Methane	− 74.898	186.313	−50.828
Ethane	− 84.724	229.646	−32.908
Propane	−103.916	270.090	−23.505
Cyclopropane	+ 53.382	237.810	104.503
n-Butane	−126.232	310.326	−17.166
Cyclobutane	+ 27.256	265.569	110.741
2-Methylpropane	−134.606	294.834	−20.934
n-Pentane	−146.538	349.179	− 8.374
Cyclopentane	− 77.079	293.076	38.895
2-Methylbutane	−154.577	343.820	−14.654
2,2-Dimethylpropane	−166.090	306.599	−15.240
n-Hexane	−167.305	388.661	− 0.293
2-Methylpentane	−174.422	380.789	− 5.024
3-Methylpentane	−171.743	380.036	− 2.135
2,2-Dimethylbutane	−185.685	358.474	− 9.923
2,3-Dimethylbutane	−177.897	366.010	− 4.103
n-Heptane	−187.945	428.058	+ 8.039
n-Octane	−208.586	467.038	16.412
n-Dodecane	−291.066	622.912	50.074
n-Hexadecane	−373.588	778.829	83.820
n-Icosane	−456.068	934.745	117.398

1.3. Production

By far the largest amount of saturated hydrocarbons is obtained from the natural sources *natural gas* and *petroleum*, either by isolation or by suitable conversion reactions. Additional sources include various products derived from coal processing. A number of saturated hydrocarbons, unavailable from natural sources, are produced by special synthesis or by conversion processes.

1.3.1. From Natural Gas and Petroleum

Natural Gas contains methane as the single major component (see also →Methane;

Table 4. Critical data and molar heat capacities of saturated hydrocarbons [2]

Compound	T_c, °C	p_c, MPa	c_p, * J mol^{-1} K^{-1}
Methane	−82.6	4.60	35.74
Ethane	32.28	4.88	52.67
Propane	96.7	4.25	73.56
Cyclopropane	125.1	5.57	56.27
n-Butane	152.0	3.80	97.51
Cyclobutane	187	4.99	72.26
2-Methylpropane	135.0	3.65	96.88
n-Pentane	196.5	3.37	120.3
Cyclopentane	238.5	4.50	82.98
2-Methylbutane	187.2	3.38	118.9
2,2-Dimethylpropane	160.6	3.20	121.7
n-Hexane	234.2	3.01	143.2
2-Methylpentane	224.3	3.01	144.3
3-Methylpentane	231.2	3.12	143.2
2,2-Dimethylbutane	215.6	3.08	142.0

Table 4. Continued.

Compound	T_c, °C	p_c, MPa	c_p, * J mol^{-1} K^{-1}
2,3-Dimethylbutane	226.8	3.13	140.6
n-Heptane	267.0	2.74	166.1
n-Octane	295.6	2.49	189.0
n-Nonane	321.4	2.29	211.8
n-Decane	344.4	2.10	234.7
n-Undecane	365.6	1.97	257.6
n-Dodecane	385.1	1.82	280.5
n-Tridecane	402.6	1.72	303.4
n-Tetradecane	418.7	1.62	326.3
n-Pentadecane	433.6	1.52	349.2
n-Hexadecane	447.4	1.42	372.0
n-Heptadecane	460.2	1.32	394.9
n-Octadecane	472.1	1.22	417.8
n-Nonadecane	483	1.12	440.7
n-Icosane	494	1.12	463.6

* Based on 25 °C and ideal gas state.

Table 5. Fuel properties of saturated hydrocarbons

	Flash point,[a] °C	Explosion limits (20 °C),[b] lower	upper	Ignition temperature,[a] °C	Octane number,[c] RON
Methane		5.00	15.00	595	
Ethane		3.00	12.50	515	
Propane		2.12	9.35	470	
Cyclopropane		2.40	10.40	495	
n-Butane		1.86	8.41	365	93.8
Cyclobutane		1.8 [a]			
2-Methylpropane		1.8 [a]	8.5 [a]	(460) [d]	
n-Pentane	<−20	1.40	7.80	285	61.7
Cyclopentane	<−20			380	102.5
2-Methylbutane	<−20	1.32	7.6 [a]	420	92.3
2,2-Dimethylpropane		1.38	7.50	(450) [d]	85.5
n-Hexane	<−20	1.18	7.40	240	24.8
i-Hexane	<−20	ca. 1 [a]	ca. 7.4 [a]	ca. 260	
2,2-Dimethylbutane	<−20	1.2 [a]	7.0 [a]	435	91.8
2,3-Dimethylbutane	<−20	1.2 [a]	7.0 [a]	415	103.5
n-Heptane	−4	1.10	6.70	215	0
i-Heptane	<−4	ca. 1 [c]	ca. 7 [c]	ca. 220	
2,3-Dimethylpentane	<0	1.12	6.75	330	91.9
2,2,3-Trimethylbutane	<0			450	112.1
2,3-Dimethylhexane	<12	0.95	6.5 [a]	210	
n-Octane	12	0.95	6.5 [a]	71.3	
i-Octane	−12	1.0	6.0	410	100
2,2,3-Trimethylpentane	<21			430	109.5
2,3,3-Trimethylpentane	<21			425	106
n-Nonane	31	0.83 [d]	5.6 [a, d]	205	
n-Decane	46	0.77 [d]	5.35 [d]	205	
n-Dodecane	74	0.6 [a, d]		200	
n-Hexadecane	>100			205	
n-Icosane	>100				

[a] Ref. [7]
[b] Ref. [8]
[c] Ref. [9]
[d] Approximate values.
[e] Values are valid for $t > 20$ °C.

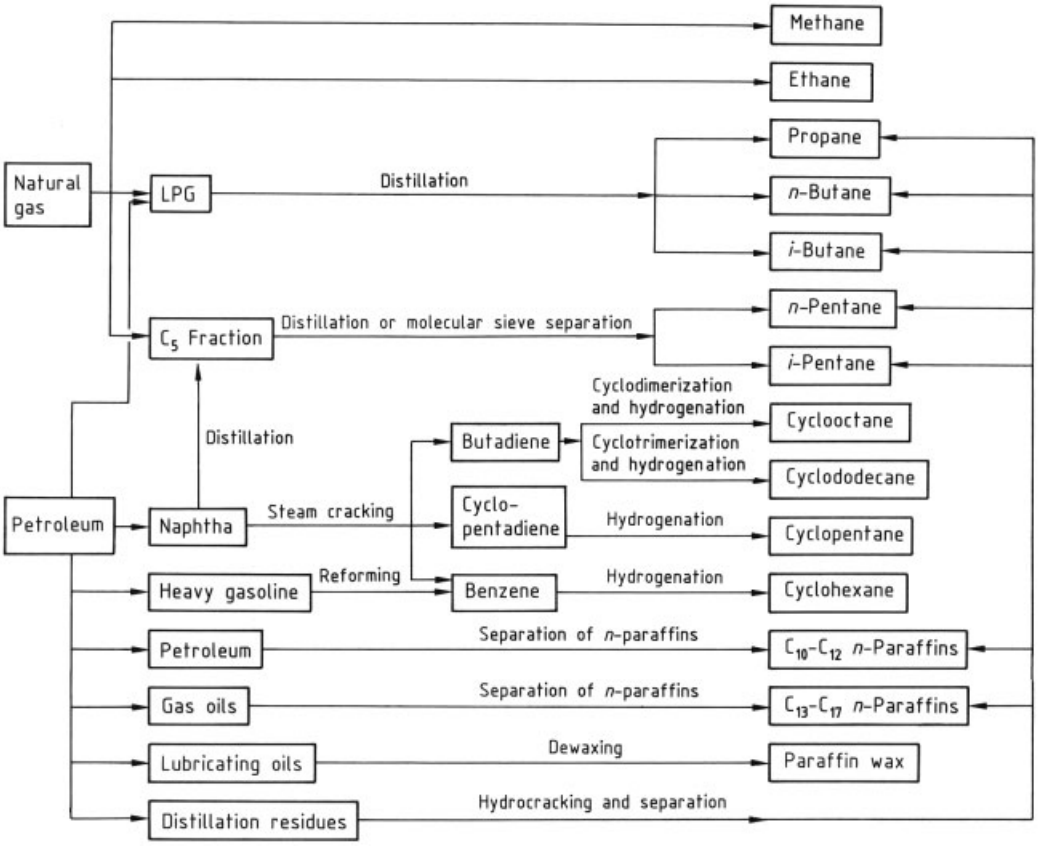

Figure 1. Saturated hydrocarbons from natural gas and from petroleum

→Natural Gas). Depending on the particular source, natural gas may also contain acyclic saturated hydrocarbons up to C_5 in proportions that permit their isolation [10]. The isolation of individual compounds from natural gas (see Figure 1) can be performed either by absorption or by partial condensation at low temperature, followed by distillation.

Petroleum is the most abundant source of saturated hydrocarbons, both with respect to the absolute amount and to the variety of individual compounds. Petroleum is separated into individual fractions by distillation in a refinery (see also →Oil Refining). Such fractions contain several (liquefied petroleum gas, LPG; →Liquefied Petroleum Gas) or many individual hydrocarbons (all other fractions); these fractions can be used as such or processed further for use as heating- or engine fuels, as lubricants, as raw materials for the production of petrochemicals (synthesis gas →Gas Production, acetylene, olefins, aromatics), or for the recovery of industrially important alkanes and cycloalkanes.

From the liquefied petroleum gas fraction, propane, *n*-butane, and *i*-butane are isolated by distillation. *n*-Pentane and *i*-pentane can be isolated from the liquid components of natural gas or from light gasoline (naphtha) either by molecular sieve separation or by superfractionation [11]. Steam-cracking of light gasoline provides, among other products, cyclopentadiene (→Cyclopentadiene and Cyclopentene), benzene (→Benzene), and 1,3-butadiene (→Butadiene); these can, in turn, be converted into the C_5-, C_6-, C_8-, and C_{12}-cycloalkanes. Catalytic reforming of heavy gasoline provides C_6- to C_8-aromatics, which can subsequently be converted into cyclohexane (→Cyclohexane), e.g., the C_7- and C_8-aromatics are hydrodealkylated, and the ensuing benzene subsequently hydrogenated. From the higher boiling petroleum, gas-oil, and wax distillate fractions, mixtures

of homologous *n*-alkanes are isolated by techniques, such as molecular sieve separation or urea extractive crystallization. Paraffin waxes can also be isolated by precipitation with suitable solvents.

Higher boiling petroleum distillates and distillation residues can be converted into mixtures of lower molecular mass hydrocarbons by hydrocracking [12, 13]. From such mixtures, saturated C_3–C_5-hydrocarbons and a broad spectrum of higher *n*-alkanes can be recovered.

Although petroleum and the various petroleum distillates contain a large variety of other saturated hydrocarbons in addition to those already mentioned, the recovery and subsequent industrial use of individual compounds is largely confined to those depicted in Figure 1; the increasing number of structural isomers with increasing molecular mass poses severe limitations on the economical separation of individual compounds from such distillate mixtures on an industrial scale.

1.3.2. From Coal and Coal-Derived Products

For a long time the most important sources of saturated hydrocarbons were coal and the products derived from the liquefaction, coking, and gasification of coal (→Coal; Coal Liquefaction; →Coal Pyrolysis). These sources became less important when natural gas and petroleum became essential raw materials for organic chemicals [14, 15].

The *liquefaction of coal* provided the greatest variety of saturated hydrocarbons. The Fischer–Tropsch Synthesis can produce alkanes in the range C_1 to C_{30} or higher, depending on the process variant (→Coal Liquefaction, Chap. 2.2.) [16–18]: the fluidized bed synthesis affords predominantly liquid hydrocarbons in the gasoline range, along with gases from C_1 to C_4. The liquid hydrocarbons contain considerable proportions of branched and olefinic compounds. The fixed bed synthesis provides higher molecular mass hydrocarbons in the range of diesel oil or paraffin wax. These products are rich in *n*-alkanes and are, therefore, suitable raw materials for detergents and for wax products. Suitable fractions may also serve as raw materials for cracking and reforming processes.

The various processes for the *hydrogenation of coal* [19–21] or of coal tar afford predominantly liquid products in the boiling range of gasoline, which are rich in aromatics, and minor amounts of C_1- to C_4-alkanes. From the gasoline fractions, benzene can be recovered and hydrogenated to cyclohexane. After the recovery of benzene, the remaining mixtures can be used as feed materials for steam crackers for the production of olefins.

Coking of coal produces methane. In addition, high temperature processes produce benzene, which can be hydrogenated to cyclohexane [22, 23]. Low temperature coking of soft coal provides a tar, from which paraffin wax may be recovered.

Gasification of coal [24] affords mainly carbon monoxide and hydrogen, which can be converted into methane by nickel-catalyzed reactions or into a range of alkanes by the Fischer–Tropsch process. Up to 10 % of methane and higher hydrocarbons such as naphtha and benzene are formed in addition to carbon monoxide and hydrogen if the gasification is carried out at elevated pressure [25].

1.3.3. By Synthesis and by Conversion of other Hydrocarbons

The major processing techniques are *cyclization, hydrogenation,* and *isomerization,* forming the basis of the industrial production of most cycloalkanes. Other techniques include *alkylation* of alkenes with alkanes to provide *i*-alkanes, and hydrocracking to give predominantly low molecular mass *i*- and *n*-alkanes, including methane.

Cyclization and Hydrogenation. Among the cycloalkanes only those containing a 5- or a 6-membered ring are directly available from natural sources, i.e., from petroleum or coal-derived products. Steamcracking of naphtha or gas oil produces cyclopentadiene as a byproduct, which can be converted into cyclopentane by catalytic hydrogenation. Similarly, cyclohexane is produced by catalytic hydrogenation of benzene, which in turn is formed by steam cracking of liquid hydrocarbons, by catalytic reforming of heavy gasoline, or by coking of coal.

The skeletons of all other industrially produced cycloalkanes are synthesized from acyclic precursors. Although there is a wide variety of synthetic procedures for such reactions, only a limited number are used industrially. Of prime importance are cycloaddition reactions of acyclic unsaturated compounds followed by hydrogenation of the resulting unsaturated cyclic products: Thus, [4 + 2]-cycloaddition (Diels – Alder reaction) of conjugated dienes and monoolefins yields saturated 6-membered ring compounds via cyclohexene intermediates, e.g., the synthesis of cyclohexane from 1,3-butadiene and ethylene [26]:

the catalyzed [4 + 4] -cycloaddition of 1,3-butadiene to give cyclooctane via 1,5-cyclooctadiene;

and the catalyzed [4 + 4 + 4] -cyclotrimerization of 1,3-butadiene to form cyclododecane via 1,5,9-cyclododecatriene:

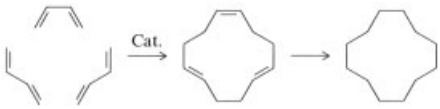

Cyclizations that lead to small (C_3-, C_4-) rings are not industrially important, although various laboratory methods are available [27]. Examples include the coupling of terminal carbon atoms by elimination of suitable α,ω-substituents from acyclic compounds, e.g., α,ω-dihalides, ω-halogenated ketones, esters, or nitriles (Eq. 1), addition of carbenes, to olefins to give cyclopropanes (Eq. 2) and [2 + 2] -cycloadditions of suitable unsaturated compounds, such as olefins, allenes, or acetylenes, to give cyclobutanes (Eq. 3) [28] (see figure at top of next column).

Most of these synthetic methods yield substituted or unsaturated cyclic compounds in the cyclization step. These primary products can be subsequently converted into the corresponding cycloalkanes by suitable measures.

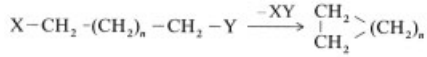

(1)

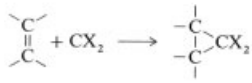

(2)

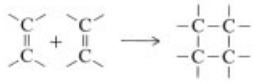

(3)

Isomerization and Alkylation. On account of their specific properties (antiknock activity, possibility for selective reactions), some *i*-alkanes are in higher demand than can be met by natural sources. *i*-Alkanes are produced either by acid-catalyzed isomerization of *n*-alkanes or by acid-catalyzed alkylation of olefins with *i*-alkanes. The most important isomerizations are the conversion of *n*-butane to *i*-butane, *n*-pentane to *i*-pentane, and *n*-hexane to *i*-hexane [29, 30]. Industrially important alkylations include the reaction of *i*-butane with propylene or with butenes to give highly branched C_7- or C_8-alkanes, respectively (see also →Oil Refining). The latter exhibit high octane numbers [30, 31].

Hydrocracking. A range of acyclic saturated hydrocarbons having low carbon numbers is accessible by catalyzed hydrocracking of heavy distillates or distillation residues. Hydrocracking of distillates can be optimized by adjusting the reaction conditions, so that the products are predominantly in the range of liquefied petroleum gases, including *i*-butane [13], or in the range of middle distillates [32]. The cracking of petroleum distillates with steam over a nickel catalyst produces methane [33, 34].

1.4. Uses

The major use of saturated hydrocarbons is as a mixture, sometimes with unsaturated or aromatic compounds, as heating fuels, motor fuels and lubricants. Saturated hydrocarbons also serve as raw materials for the production of carbon black. In the chemical industry, the following processes are important (Fig. 2, see next page).

Thermal Cracking. Thermal cracking of individual or mixed saturated hydrocarbons is

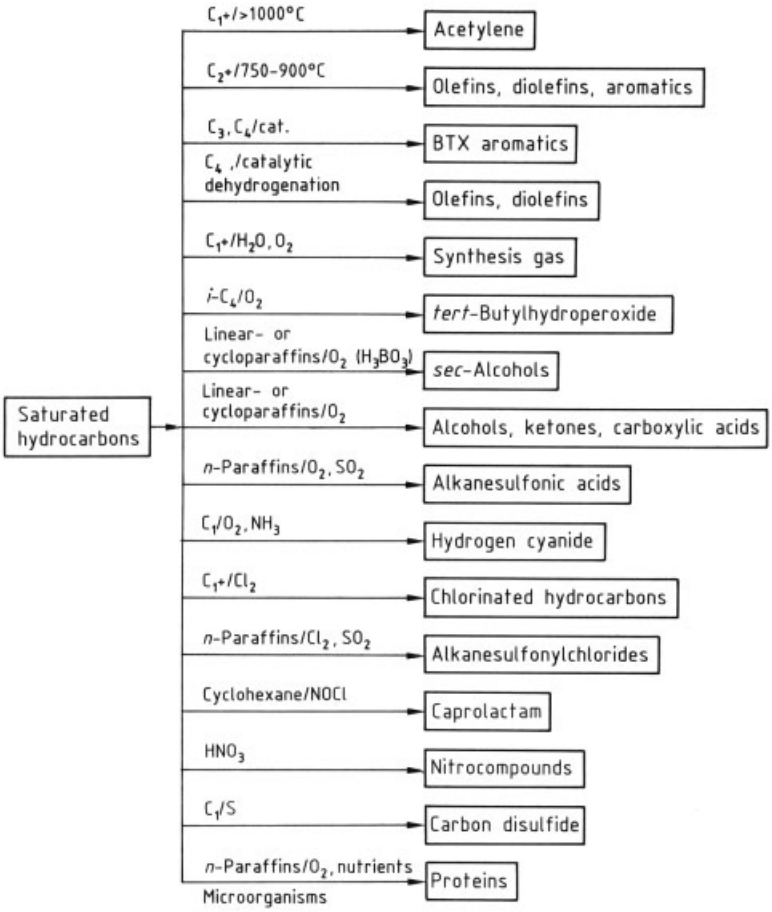

Figure 2. Industrial use of saturated hydrocarbons

the basis for the petrochemical production of acetylene, olefins, diolefins, and to some extent BTX aromatics: high temperature pyrolysis at >1000 °C affords acetylene (→Acetylene) and, depending on the hydrocarbon feed and the particular process applied, ethylene. The range of hydrocarbon feed materials can vary from methane to crude oil. Medium temperature pyrolysis at 750 – 900 °C in the presence of inert diluents such as steam (steam-cracking) or hydrogen (hydropyrolysis) yields ethylene (→ Ethylene). A wide range of hydrocarbons, e.g., ethane, liquefied petroleum gas, naphtha, and gas oil may serve as raw material, and increasing proportions of C_3- to C_5-olefins, C_4- and C_5-dienes, and C_6- to C_8-aromatics are coproduced with increasing molecular mass of the feed material. Pyrolysis of petroleum waxes at 500 – 600 °C affords mixtures of homologous olefins with predominantly terminal double bonds.

Catalytic Dehydrogenation. Catalytic dehydrogenation of acyclic saturated hydrocarbon atoms yields mono- and/or diolefins with the same number of carbon atoms. Industrial processes include the production of butenes and butadiene from *n*-butane, of isoprene from *i*-pentane, and C_6- to C_{19}-monoolefins, mainly with internal double bonds, from the corresponding *n*-alkanes [35]. More recently, improved processes for the production of ethylene, propene, and *i*-butene by catalytic dehydrogenation of the corresponding saturated hydrocarbons have been developed [36]. Catalytic dehydrogenation of cyclohexane and methyl cyclohexanes yields benzene and methylbenzenes, respectively.

Catalytic dehydrocyclooligomerization of propane and *n*-butane over modified zeolite catalysts represents a new method for the production of BTX-aromatics [37]. The process is apparently ready for industrial application.

Partial Oxidation. Partial oxidation of saturated hydrocarbons either by catalyzed reaction with steam (steam reforming) or by noncatalyzed reaction with deficient amounts of oxygen affords a mixture of carbon monoxide and hydrogen, which is used as town gas or as synthesis gas (→Gas Production). Desulfurized saturated hydrocarbons from methane to compounds with *bp* ca. 200 °C are suitable feed materials for catalytic processes; for noncatalytic processes all hydrocarbons may be used from methane to distillation residues.

Selective Oxidation. Selective oxidation represents the most important tool for the introduction of oxygen-containing functional groups into saturated hydrocarbons: Liquid-phase autoxidation of *i*-butane with air yields *tert*-butylhydroperoxide; the latter is used for the selective oxidation of propene to propylene oxide [38]. Liquid-phase oxidation of *n*-alkanes or cycloalkanes in the presence of boric acid affords the corresponding secondary alcohols. The reaction is used industrially for the oxidation of C_{10}- to C_{20}-*n*-alkanes, providing raw materials for detergents and for the oxidation of cyclododecane to cyclododecanol as an intermediate for the production of nylon 12. Catalyzed liquid-phase oxidation of *n*-alkanes and cycloalkanes in the absence of boric acid leads to mixtures of secondary alcohols and ketones or to carboxylic acids, depending on the reaction conditions: oxidation of cyclohexane leads to a mixture of cyclohexanol and cyclohexanone (→Cyclohexanol and Cyclohexanone, Chap. 3.2.); oxidation of *n*-butane gives acetic acid (→Acetic Acid, Chap. 4.2.); oxidation of longer chain *n*-alkanes produces mixtures of fatty acids; and oxidation of cyclohexane gives adipic acid. (→Adipic Acid, Chap. 4.). Liquid phase oxidation of bicyclo [4.4.0] decane yields the corresponding tertiary hydroperoxide, which is eventually converted into sebacic acid by a sequence of reaction steps [39].

Heterogeneously catalyzed gas-phase oxidation of *n*-butane has been used industrially as an attractive alternative for the production of maleic anhydride [40]. Selective oxidation of methane to methanol [41] or formaldehyde [42] and of *i*-butane to methacrylic acid [43] has been examined, but is not yet used industrially.

Oxidative Couplings. Oxidative coupling of methane either with itself to give ethane and ethylene [44] (Eq. 4), or with compounds having activated methyl groups (Eq. 5) to give terminal olefinic compounds [45] has been investigated, but is not yet industrially used.

$$CH_4 \xrightarrow[-H_2O]{O_2} CH_3CH_3 + CH_2=CH_2 \quad (4)$$

$$CH_4 + CH_3-X \xrightarrow[-H_2O]{O_2} CH_2=CH-X \quad (5)$$

$$X = -CH=CH_2,\ -C(CH_3)=CH_2,\ -C_6H_5,\ -C\equiv N$$

Three-Component Oxidation. Three-component oxidation allows introduction of hetero atoms or groups of hetero atoms other than oxygen into saturated hydrocarbons. The radical-induced action of both oxygen and sulfur dioxide (sulfoxidation) on alkanes yields the corresponding alkanesulfonic acids. The latter, produced from C_{14}– to C_{18}–*n*-alkanes, are useful as raw materials for detergents. The heterogeneously catalyzed reaction of oxygen and ammonia with methane (ammoxidation) affords hydrogen cyanide. The latter reaction can be also carried out in the absence of oxygen as a two-component codehydrogenation. The reaction of oxygen and ammonia with propane to give acrylonitrile has recently been explored.

Chlorinations. Chlorination of saturated hydrocarbons yields mono- or polychlorinated hydrocarbons, depending on conditions. The following reactions are used industrially: Chlorination of methane to give the four chloromethanes (→Chlorinated Hydrocarbons, Chap. 1.3.1.); the chlorination of *n*-alkanes to give mono- or polychlorinated alkanes having the same number of carbon atoms and chlorinolysis of low molecular mass alkanes or chloroalkanes to give tetrachloroethene and tetrachloromethane. Reaction of chlorine and sulfur dioxide together with alkanes yields alkanesulfonyl chlorides,

which in turn can be converted into the corresponding sulfonic acids, sulfonates, sulfonamides, or sulfonic esters.

Photonitrosation. Photonitrosation of saturated alkanes with nitrosyl chloride yields oximes via labile secondary nitroso compounds. The method has been used for the production of caprolactam from cyclohexane (→Caprolactam, Chap. 4.2.).

Nitrosation. Nitrosation of saturated hydrocarbons with nitric acid has been carried out with low molecular mass alkanes, particularly with propane in the gas phase. The reaction gave mixtures of nitro compounds with the same number of and fewer carbon atoms than those of the starting alkanes. Such mixtures are useful as specialty solvents, e.g., for polymers.

Reaction with Sulfur. Reaction of methane with sulfur gives carbon disulfide in high yields with hydrogen sulfide as byproduct. The reaction can be performed by thermal or catalytic methods (→Carbon Disulfide, Chap. 4.).

Fermentation. Fermentation of certain saturated hydrocarbons by microorganisms (yeast, bacteria) yields single cell proteins [46–50]. The microorganisms oxidize the hydrocarbons in the presence of oxygen and nutrients with concomitant formation of protein-containing cells. Certain microorganisms are specific for the fermentation of *n*-alkanes. Industrial processes have therefore been developed both for the fermentation of isolated *n*-alkanes and for *n*-alkanes in admixture with branched or cyclic hydrocarbons (e.g., diesel oil). The products obtained are used as animal feed.

1.5. Individual Saturated Hydrocarbons

Methane, Cyclohexane, and *Waxes* are separate keywords. Of the other saturated hydrocarbons only those used industrially will be discussed.

Ethane. Ethane is present in many natural gas sources and in refinery gases; it can be recovered for industrial use. Furthermore, ethane is formed in thermal- and hydrocracking of hydrocarbons (→ Oil Refining) and in the liquefaction of coal (→ Coal Liquefaction). The single most important industrial use for ethane is the production of ethylene by steam-cracking (→ Ethylene, Chap. 5.1.3.). Thermally-induced chlorination of ethane yields predominantly monochloro-, 1,1-dichloro-, or 1,1,1-trichloroethane, depending on the conditions used. Combined chlorination and oxichlorination of ethane in a salt melt affords vinyl chloride [51]. Reaction of ethane with nitric acid in the gas phase yields nitroethane and nitromethane (→ Nitro Compounds, Aliphatic).

Propane. Propane is a component of liquefied petroleum gas (LPG), which is derived from natural gas or petroleum. Propane can be recovered from LPG by distillation. The major industrial use of propane is the production of ethylene and propene by the steam cracking process (→ Ethylene, Chap. 5.1.3.). Chlorinolysis of propane at elevated temperature yields a mixture of tetrachloroethene and tetrachloromethane. The conversion of propane into BTX aromatics by zeolite-catalyzed dehydrocyclooligomerization has been developed recently and may find industrial application. Liquefied propane is used for the deasphalting of petroleum residues.

***n*-Butane.** *n*-Butane can be recovered from LPG by distillation; it has a variety of industrial uses: steamcracking yields ethylene and propene, catalytic dehydrogenation yields butadiene (→Butadiene), and acid-catalyzed isomerization provides *i*-butane. Oxidation of *n*-butane in the gas phase with heterogeneous catalysis is a modern process for the production of maleic anhydride [40]. Noncatalyzed oxidation of *n*-butane in the liquid or gas phase has been used industrially to produce acetic acid [52] (→Acetic Acid, Chap. 4.2.).

***i*-Butane.** *i*-Butane can be obtained either by recovery from LPG or by isomerization of *n*-butane. Traditional industrial uses include the production of high octane engine fuel by the alkylation of olefins [30,31] and liquid-phase autoxidation of *i*-butane to give *tert*-butylhydroperoxide. Recent developments comprise improved processes for the catalytic dehydrogenation of *i*-butane [53] to satisfy the in-

creasing demand for *i*-butene as a pecursor for methyl *tert*-butyl ether (MTBE) (→Methyl Tert-Butyl Ether), and the production of methacrylic acid from *i*-butane by oxidation [43].

***n*-Pentane and *i*-Pentane.** Both isomers are present in light gasoline fractions and can be recovered individually by superfractionation [11,54], possibly in combination with molecular sieve separation. *n*-Pentane is used as a solvent and for the production of *i*-pentane by acid-catalyzed isomerization [26]. *i*-Pentane is used as a blending component for high octane gasoline and for the production of isoprene by catalytic dehydrogenation [53,54].

***n*-Hexane.** *n*-Hexane can be isolated from suitable sources (e.g., light gasoline or BTX raffinates) by superfractionation, or by molecular sieve separation. It is used for the extraction of vegetable oils (e.g., from soybeans), as a solvent in chemical reactions (e.g., for coordination complex catalyzed polymerization of olefins) and in adhesive formulations.

Higher *n*-Alkanes. *n*-Alkanes having more than six carbon atoms are not isolated individually but as mixtures of several homologues. Suitable sources are the appropriate petroleum distillate fractions, from which the *n*-paraffins can be isolated in high isomeric purity (≥95 % linearity) by selective separation (e.g., molecular sieve separation or urea extractive crystallization). They are used mainly in applications for which *i*-alkanes are not acceptable for biological reasons, e.g., the production of detergents or proteins. Catalyzed gas-phase dehydrogenation of *n*-alkanes over noble metal catalysts and at low conversion (ca. 10 %) yields the corresponding *n*-alkenes with predominantly internal double bonds. The alkenes can be isolated in high purity by selective molecular sieve processes [35, 56]. *n*-Alkenes in the range C_7 to C_{10} are used as raw materials for the production of flexibilizers, by oxo-synthesis with hydrogen and carbon monoxide. *n*-Alkenes in the range C_{10} to C_{19} are used for the production of biodegradable detergents: acid-catalyzed reaction of C_{10}- to C_{13}-*n*-alkenes with benzene, followed by sulfonation and neutralization of the ensuing alkylbenzene sulfonic acids, affords linear alkylbenzenesulfonate (LABS) detergents. Oxo-reaction of C_{10}- to C_{19}-*n*-alkenes yields C_{11}- to C_{20}-alcohols, which are converted into nonionic detergents by reaction with ethylene oxide, or into anionic detergents by reaction with sulfur trioxide or chlorosulfonic acid. Liquid-phase oxidation of *n*-alkanes with oxygen in the presence of boric acid yields predominantly secondary alcohols. The latter can be dehydrated to the corresponding *n*-alkenes. The C_{12}- to C_{19}-alcohols are also used for the production of both ionic and nonionic detergents.

Chlorination of *n*-alkanes yields mono- or polychlorinated alkanes, depending on the conditions. Mixtures of C_{10}- to C_{13}-monochloroalkanes are used for the production of linear alkylbenzenes as raw materials for LABS-detergents. For this, the chloroalkanes are either directly reacted with benzene or they are dehydrochlorinated to the corresponding *n*-alkenes, which are then used to alkylate benzene. Polychlorinated *n*-alkanes are used in a number of special applications. *n*-Alkanes in the range of C_{10} to C_{21} are suitable substrates for the production of single cell proteins by fermentation. The microorganisms used selectively attack *n*-alkanes, so that industrial processes have been developed in which either isolated *n*-alkanes or the appropriate petroleum distillates containing *n*-alkanes (petroleum-, gas oil-, or lube oil fractions) are used as starting materials.

Solid *n*-alkanes (paraffin waxes) are used in a variety of applications, both as such and as feed materials in cracking, oxidation and chlorination.

Cycloalkanes. Of the various accessible cycloalkanes, only cyclohexane and cyclododecane are used industrially on a large scale; cyclooctane is used on a relatively small scale. The strained C_3- and C_4-cycloalkanes are not produced industrially. *Cyclopentane* is present in natural hydrocarbon sources and is formed as a byproduct in the processing of cracking products, e.g., by hydrogenation of cyclopentadiene or cyclopentene in the C_5-fraction of steamcracked products. No important industrial use exists for cyclopentane.

Cyclooctane

is accessible by catalytic hydrogenation of 1,5-cyclooctadiene. Liquid-phase oxidation of the former with air yields 1,8-octanedicarbox-

ylic acid, which is used on a moderate scale for the production of nylon 68.

Cyclododecane

is produced by liquid-phase hydrogenation of 1,5,9-cyclododecatriene over nickel catalysts. Liquid-phase oxidation of cyclododecane with air in the presence of boric acid yields mixtures of cyclododecanol and cyclododecanone in high selectivity. Further oxidation of such alcohol – ketone mixtures with nitric acid gives 1,12-dodecanedicarboxylic acid, used for the production of polyamides, polyesters, and synthetic lubricating oils. Catalyzed dehydrogenation of alcohol-ketone mixtures in the liquid phase yields cyclododecanone, which is then converted into laurinlactam and nylon 12.

Bicyclo [4.4.0] decane (decalin) is obtained by catalytic hydrogenation of naphthalene. Liquid-phase oxidation of decalin with oxygen gives the corresponding tertiary hydroperoxide, which is eventually converted into 1,10-decanedicarboxylic acid (sebacic acid) by a four-step sequence [39].

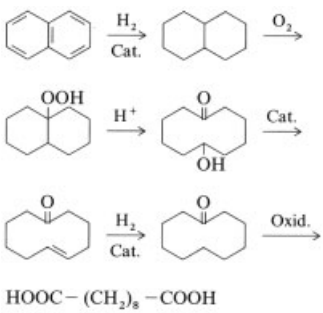

2. Olefins

Olefins are aliphatic hydrocarbons containing at least one carbon–carbon double bond. Monoolefins (alkenes) contain a single C=C double bond and form a homologous series with the empirical formula C_nH_{2n}. Hydrocarbons with two double bonds are differentiated according to the relative positions of the double bonds: in cumulated dienes (cumulenes, allenes) they are immediately adjacent to each other; in conjugated dienes they are separated by a single bond; and in other diolefins, the double bonds are isolated from each other.

More highly unsaturated polyenes are known, e.g., trienes, tetraenes. The name olefin is derived from the property of these compounds to form oily liquids on reaction with halogens (*gaz oléfiant,* oil-forming gas).

The industrial importance of olefins started in the 1950s when the lower olefins became widely available from thermal cracking of wet natural gas and petroleum fractions, displacing acetylene (ethyne) as the dominant commodity chemical (→Acetylene). In the time since then, ethene (ethylene), propene, 1,3-butadiene, and isoprene (2-methyl-1,3-butadiene) have found wide application, especially for synthetic polymers (→Butadiene ; →Ethylene ; →Isoprene ; →Polyolefins ; →Propene).

2.1. Monoolefins

The low molecular mass monoolefins ethylene, propene, styrene, cyclopentene, and the butenes are separate keywords: (→Butenes; →Cyclopentadiene and Cyclopentene; →Ethylene). Among the higher olefins ($C_6 - C_{18}$), the linear α-olefins are of particular industrial interest.

2.1.1. Properties

Physical Properties. Under normal conditions, the olefins from ethylene to the butenes are gases; from the pentenes to 1-octadecene liquids; and from 1-icosene onwards solids. These thermal phase transitions do not differ significantly from those of the corresponding saturated hydrocarbons (Chap. 1). The density of olefins ranges from 0.63 to 0.79 g/cm^3, only a few percent higher than those of the corresponding alkanes.

The heats of combustion of alkenes and the corresponding alkanes are also nearly identical. Olefins have low solubility in water or are insoluble; they dissolve well in most organic solvents, e.g., alcohols, ethers, and aromatic hydrocarbons. Some physical properties of important monoolefins are listed in Table 6 (see next page).

Chemical Properties. The chemistry of monoolefins is dominated by the reactive double bond. Some important reactions of olefins are summarized in Table 7 (see page after next).

Table 6. Physical properties of monoolefins

Olefin	CAS registry number	*mp*, °C	*bp*, [a] °C	d^{20}	n^{20}
Ethylene	[*74-85-1*]	−169.5	−103.8	0.5699 (−103.8 °C)	1.363 (−100 °C)
Propene	[*115-07-1*]	−185.3	− 47.7	0.6095 (−47.4 °C) 0.5139	1.3640 (−50 °C)
1-Butene	[*106-98-9*]	−185.3	−6.26	0.6356 (−15.1 °C)	1.3791 (−25 °C)
cis-2-Butene	[*590-18-1*]	−138.9	3.7	0.6447 (0 °C)	1.3783 (0 °C)
trans-2-Butene	[*624-64-6*]	−105.5	0.88	0.6269 (0 °C)	1.3696 (−20 °C)
Isobutene	[*115-11-7*]	−140.4	−6.9	0.6398 (−20 °C)	1.3776
1-Pentene	[*109-67-1*]	−167.2	29.97	0.6405	1.3718
cis-2-Pentene	[*627-20-3*]	−151.3	37.1	0.6554	1.3828
trans-2-Pentene	[*646-04-8*]	−140.2	36.4	0.6482	1.3798
2-Methyl-1-butene	[*58718-78-8*]	−137.5	31.2	0.6504	1.3778
3-Methyl-1-butene		−168.5	20.1	0.6332	1.3640
2-Methyl-2-butene	[*26760-64-5*]	−133.6	38.6	0.6625	1.3874
1-Hexene	[*592-41-6*]	−140.0	63.5	0.6732	1.3878
cis-2-Hexene	[*7688-21-3*]	−141.8	68.8	0.6969	1.3976
trans-2-Hexene	[*4050-45-7*]	−133.5	67.5	0.6780	1.3935
cis-3-Hexene	[*7642-09-3*]	−137.8	66.5	0.6794	1.3951
trans-3-Hexene	[*13269-52-8*]	−113.4	68.0	0.6778	1.3937
1-Heptene	[*592-76-7*]	−119.6	93.8	0.6993	1.3998
cis,trans-2-Heptene		−109.5	98.2	0.7034	1.4041
cis-3-Heptene	[*7642-10-6*]	136.6	95.8	0.7030	1.4059
trans-3-Heptene	[*14686-14-7*]	−136.3	95.7	0.6981	1.4043
1-Octene	[*111-66-0*]	−102.7	121.3	0.7160	1.4086
cis-2-Octene	[*7642-04-8*]	−100.5	125.6	0.7243	1.4146
trans-2-Octene	[*13389-42-9*]	− 87.8	124.5	0.7184	1.4132
cis-3-Octene	[*14850-22-7*]	−126	122.7	0.7223	1.4140
trans-3-Octene	[*14919-01-8*]	−110.0	123.3	0.7149	1.4129
cis-4-Octene	[*7642-15-1*]	−118.1	122.5	0.7212	1.4150
trans-4-Octene	[*14850-23-8*]	− 93.8	122.3	0.7136	1.4120
1-Nonene	[*124-11-8*]	− 81.6	146.8	0.7292	1.4157
2-Nonene	[*2216-38-8*]		149.4	0.7407	1.4201
3-Nonene	[*20063-77-8*]		148.2 (102 kPa)	0.7334	1.4191
4-Nonene	[*2198-23-4*]		148.0	0.7313	1.4200
1-Decene	[*872-05-9*]	− 66.6	170.6	0.7408	1.4215
2-Decene	[*6816-17-7*]				
3-Decene	[*19348-37-9*]		173 – 173.5	0.7447	1.4221
4-Decene	[*19689-18-0*]		170.6	0.7404	1.4243
cis-5-Decene	[*7433-78-5*]	−112	169.5 (98.5 kPa)	0.7445	1.4252
trans-5-Decene	[*7433-56-9*]	− 73	170.2	0.7412	1.4235
1-Undecene	[*821-95-4*]	− 49.1	192.67	0.75033	1.4261
5-Undecene	[*4941-53-1*]		192.2	0.7511	1.4289
1-Dodecene	[*25378-22-7*]	− 33.6	213.3	0.7583	1.4300
6-Dodecene	[*7206-29-3*]		212.9 (99.3 kPa)	0.7597	1.4328
1-Tridecene	[*2437-56-1*]	− 22.2	104 (1.33 kPa)	0.7654	1.4335
1-Tetradecene	[*1120-36-1*]	− 12	125 (1.85 kPa)	0.7726	1.4367
1-Pentadecene	[*13360-61-7*]	− 3.8	133.5 (1.33 kPa)	0.7763	1.4389
1-Hexadecene	[*629-73-2*]	4.0	155 (2.0 kPa)	0.7811	1.4412
1-Heptadecene	[*6765-39-5*]	10.7	159 (1.33 kPa)	0.7860	1.4436
1-Octadecene	[*112-88-9*]	18.0	176 – 178 (1.85 kPa)	0.7888	1.4450
1-Icosene	[*3452-07-1*]	28.5	151 (0.2 kPa)		1.4440 (30 °C)
1-Heneicosene	[*1599-68-4*]	35.5	134 (0.005 kPa)	0.7985	1.4510
1-Triacontene	[*18435-53-5*]	62 – 63	218 (0.07 kPa)	0.7620 (100 °C)	

[a] At 101.3 kPa, unless otherwise noted.

Most *addition reactions* proceed by an electrophilic mechanism via carbonium ions. Chlorination and bromination yields the trans adducts. Hydrohalogenation generally follows Markovnikov's rule to give a product containing halogen linked to secondary carbon. Addition of sulfuric acid to olefins affords alkyl sulfates, which are readily hydrolyzed to alcohols; the reaction corresponds to the acid-catalyzed addition of water. Ethanol, isopropylalcohol,

and *tert*-butyl alcohol are produced industrially by this method from ethylene, propene, and isobutene, respectively (→Alcohols, Aliphatic, Chap. 2.3.7.).

Table 7. Important reactions of olefins

Reaction	Reagents	Products
Halogenation	Cl_2, Br_2	dihaloalkanes
Hydrohalogenation	HCl, HBr	alkylhalides
Sulfation	H_2SO_4	alkyl sulfates
Hydration	H_2O; catalyst: H^+	alcohols
Hydrogenation	H_2; catalyst: Ni, Pt	alkanes
Epoxidation	a) O_2 catalyst: Ag	ethylene oxide
	b) RCOOOH, ROOH	epoxides
Hydroxylation	H_2O_2; catalyst: H^+	glycols
Wacker oxidation	H_2O; catalyst: Pd – Cu	acetaldehyde, ketones
Ozonolysis	O_3	aldehydes, ketones
Hydroformylation	CO, H_2; catalyst: Co, Rh	aldehydes
Hydrocarboxylation	CO, H_2O; catalyst: Ni, Co	carboxylic acids
Sulfonation	SO_3, NaOH	alkenesulfonates
Hydroboration	B_2H_6	trialkylboranes
Alkylation	a) of aromatics, with a Friedel – Crafts catalyst	alkylaromatics
	b) of isoalkanes, catalyst: H^+	branched alkanes
Polymerization	Catalyst: radicals, metal complexes	polymers
Oligomerization	Catalyst: metal complexes, BF_3	oligomers

Hydrogenation of olefins to alkanes requires a catalyst: Raney nickel is most frequently used industrially, although platinum-group metals are also used (→Hydrogenation and Dehydrogenation).

Epoxidation of ethylene may be carried out with oxygen in the presence of a silver catalyst. Higher olefins are converted to epoxides using hydroperoxides or peracids (→Epoxides). A related reaction is the hydroxylation of olefins with hydrogen peroxide in glacial acetic acid: an epoxide is initially formed, which subsequently reacts with water in the acidic medium to form a glycol.

Another olefin oxidation is the *Wacker oxidation* of ethylene to acetaldehyde (→Acetaldehyde, Chap. 4.3.1.). The stoichiometric reaction of ethylene with water and palladium(II) chloride yields acetaldehyde and palladium metal. Addition of copper(II) chloride converts the stoichiometric reaction into a catalytic process, wherein the copper(II) chloride reoxidizes the palladium to palladium(II) chloride the resulting copper(I) chloride is reoxidized to copper(II) chloride by atmospheric oxygen in hydrochloric acid solution.

Ozonolysis is used mainly for structure determination of olefins, but may also be used preparatively. Reaction of olefins with carbon monoxide and additional reagents (e.g., hydrogen, water, alcohols) leads to a plethora of valuable products (e.g., aldehydes, alcohols, carboxylic acids, esters). These reactions proceed under catalysis by transition metals, e.g., hydroformylations (Oxo process) with rhodium- or cobalt carbonyls.

Sulfonation of olefins leads to alkenesulfonic acids, which are used as detergents (→ Laundry Detergents). Hydroboration is used to generate primary alcohols via trialkylboranes. In contrast to the direct addition of water, this pathway is regioselective to yield the anti-Markovnikov products.

Olefins may undergo a number of reactions resulting in new C-C bonds. With aromatic compounds alkylaromatics are formed by the Friedel – Crafts reaction; with isoalkanes chain-lengthened, branched hydrocarbons, which are used in motor fuels, are formed (→Automotive Fuels). Olefins may also react with themselves or each other to form oligomers and polymers.

The industrial use of the higher monoolefins is discussed in Section 2.1.3.

2.1.2. Production of Higher Olefins

Olefins are synthesized on a laboratory scale by a variety of routes: e.g., selective hydrogenation of alkynes; dehydration of alcohols; pyrolysis of alkyl lithium compounds; the Wittig reaction; the Peterson reaction; and deoxygenation of vicinal diols. Only processes starting with cheap raw materials have gained industrial importance for the production of higher olefins. The following discussion considers processes that start either with long-chain alkanes (Section 2.1.2.1) or those that build up longer chain olefins from their lower molecular mass homologues (Section 2.1.2.2).

Table 8. Composition of fractions (wt %) obtained in a paraffin cracking process (Chevron process)

	C_6-C_7	C_8-C_{10}	$C_{11}-C_{12}$	$C_{13}-C_{14}$	$C_{15}-C_{18}$	$C_{18}-C_{20}$
α-Olefins	83	83	88	88	89	86
Internal olefins	12	11	5	5	3	1
Diolefins	3	4	5	5	6	4
Paraffins	2	2	2	2	2	9
Aromatics	0.4	0.2	0	0	0	0

2.1.2.1. Production from Paraffins

Thermal Cracking of Waxes. Thermal cracking of long chain alkanes yields mainly α-olefins.

$$R^1-CH_2CH_2CH_2CH_2-R^2 \longrightarrow R^1-CH{=}CH_2 + R^2-CH{=}CH_2 + H_2$$

Thermal cracking involves a radical mechanism. Cleavage of a C-C bond leads to carbon radicals, which are converted into olefins by loss of a hydrogen atom. The latter predominantly combine to form hydrogen; a small fraction reacts with carbon radicals to form alkanes.

To generate linear α-olefins, linear alkanes (waxes) must be used. Suitable sources are the petroleum, diesel oil, and lubricant fractions of paraffin-based crude oils. One possibility of enriching the linear paraffin content of such fractions is to dilute with a suitable solvent, followed by chilling to crystallize a mixture of paraffins. Linear paraffins can be separated from their branched and cyclic analogs by subsequent treatment with molecular sieves. Another method for the separation of *n*-paraffins, especially suitable for long-chain ($>C_{20}$) species, is *extractive crystallization with urea*. This process uses the property of urea to form inclusion compounds with linear paraffins, but not with branched or cyclic alkanes, or aromatics. During crystallization the urea molecules assemble to form cavities having a diameter of 0.53 nm, in which only linear alkanes can be accommodated. The alkanes are then liberated by subsequent decomposition of the complex, e.g., with water at 75 °C. Industrial processes use solid urea (Nurex process) or its saturated aqueous solution (Edeleanu process).

The long-chain paraffins are then cracked by first heating the vapor to 400 °C within a few seconds, followed by thermolysis at 500 – 600 °C in the presence of steam. Rapid quenching of the reaction mixture reduces side reactions such as isomerization or cyclization.

Using $C_{18}-C_{36}$ alkanes as starting material, α-olefins with even and odd numbers of carbon atoms in the range C_6-C_{20} are obtained; only small amounts of paraffins are present. Byproducts include nonterminal olefins, conjugated and nonconjugated dienes, and traces of aromatics. Typical composition of the olefin fractions from paraffin cracking are given in Table 8.

The dependence of product quality on the starting material and process parameters is discussed in a review [66]; catalytic cracking and cracking in the presence of oxygen are also reviewed. The latter processes have not gained acceptance because of the presence of numerous byproducts, e.g., branched olefins and oxidation products.

The first industrial paraffin cracking plant was brought into service in 1941 by Shell at Stanlow (United Kingdom). Other plants were built at Pernis (The Netherlands) and Berre (France). A U.S. paraffin cracking plant owned by Chevron is located in Richmond, California. Future expansion of paraffin cracking processes is unlikely because of the limited flexibility of paraffin cracking plants with respect to carbon number distribution, the limited supply of paraffin-containing crudes, the high percentage ($\geq$10 %) of byproducts in the α-olefin cut, and because of more efficient competitive processes.

Catalytic Dehydrogenation. Catalytic dehydrogenation of paraffins leads to olefins with the same number of carbon atoms and with random location of the double bond along the chain. Thermally, this reaction cannot be carried out economically because the energy required to cleave a C-H bond (365 kJ/mol) significantly exceeds that for breaking a C-C bond (245 kJ/mol). Cracking of longer chain hydrocarbons is, therefore, favored over dehydrogenation; however, the use of a catalyst facilitates the dehydrogenation of linear paraffins to linear olefins having predominantly the same number

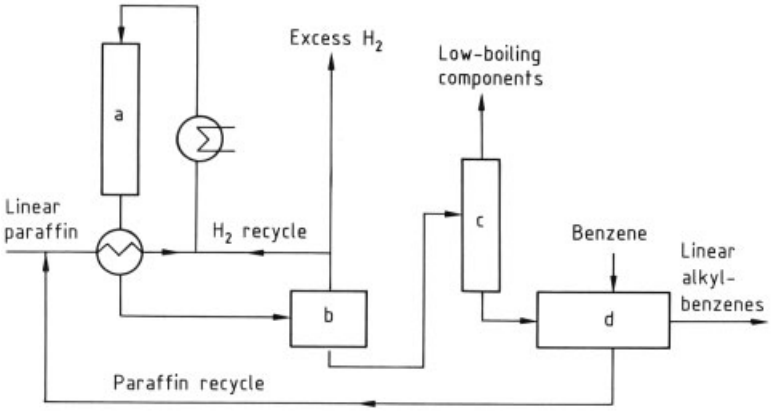

Figure 3. Simplified flow diagram for a combination of the UOP PACOL process and benzene alkylation
a) Pacol reactor; b) Separator; c) Stripper; d) Alkylation reactor

of carbon atoms (see also →Hydrogenation and Dehydrogenation).

The most important industrial process is the Pacol (paraffin catalyst olefin) process of Universal Oil Products (UOP). The reaction is carried out at 450 – 510 °C and 0.3 MPa using a hydrogen : carbon ratio of 9 : 1. A supported platinum catalyst, ca. 0.8 wt % Pt, on alumina, activated by addition of lithium and arsenic or germanium, is used. To minimize the formation of byproducts, especially aromatic hydrocarbons, which reduce catalyst activity, relatively low conversion rates of 10 – 15 % are used. This results in paraffin – olefin mixtures that are separated by means of the UOP Olex (olefin extraction) process. The Olex process employs molecular sieves, which absorb olefins more strongly than alkanes. The combination of both processes is known as the Pacol – Olex process [67]. In 1983, there were four Pacol – Olex plants worldwide, as well as 18 other plants where the Pacol process was combined with the alkylation of aromatics [68]. A simplified flow diagram of a Pacol benzene alkylation unit is shown in Figure 3.

Fresh *n*-alkane is added to the circulating paraffin, heated via heat exchangers, and dehydrogenated in the Pacol fixed-bed reactor (a) in the presence of hydrogen and at high throughput velocity. The products are separated into hydrogen and hydrocarbons in the separator (b) and volatile cracking products are removed in the stripper (c); these are then used chiefly as fuel. The stripper sump contains the higher linear monoolefins, which are reacted with benzene under acid catalysis to yield alkylbenzenes. The unreacted *n*-alkanes are returned to the Pacol reactor [69].

Chlorination – Dehydrochlorination. Another method for generating linear olefins from linear paraffins consists of chlorinating the paraffins to chloroalkanes, followed by a catalyzed second step in which hydrogen chloride is eliminated. Chlorination is carried out continuously in the liquid phase at 120 °C; the conversion is limited to a maximum of 30 %, to avoid excessive formation of dichlorides. A random mixture of linear olefins results from the dehydrohalogenation because the chlorine substituents are almost randomly distributed along the alkyl chain. The dehydrochlorination of chlorinated paraffins occurs in the presence of iron or iron alloy catalysts at 250 °C. The lower boiling olefins and paraffins are drawn off at the still head; the unreacted, higher boiling chlorinated paraffins remain in the reactor. Separating products in this manner requires that rather narrow cuts be taken, because the boiling points of the *n*-olefins and the chloroparaffins are quite close together [70, 71].

2.1.2.2. Oligomerization of Lower Olefins

All the methods described so far for the production of olefins from paraffins suffer from the disadvantage that crude oil containing linear paraffins with the required distribution of carbon atoms is in short supply. Several companies have, therefore, developed alternative routes to

Table 9. Composition of fractions (wt %) obtained from the Gulf process *

	C_6	C_8	C_{10}	C_{12}	C_{14}	C_{16}	C_{18}
Linear α-olefins	97.0	96.0	95.0	94.0	93.0	92.0	91.0
β-Branched α-olefins **	1.4	2.5	3.4	4.6	5.6	6.6	7.8

* *n*-Alkane content 1.4 %.
** Including traces of olefins with internal double bonds.

olefins that are based on oligomerization of olefins, especially of ethylene. Such starting materials are readily available from the pyrolysis of liquefied natural gas, naphtha, and gas oil fractions. The worldwide capacity for ethylene alone was ca. 50×10^6 t in 1983. Ethylene supplies appear assured for the forseeable future because it can be generated by the Mobil process from methanol, which is derived from coal-based synthesis gas.

The methods for buildup of longer chain olefins vary greatly. In the following sections syntheses via organoaluminum compounds, transition metal catalyzed oligomerization, and acid-catalyzed processes will be discussed.

Ethylene Oligomerization in the Presence of Organoaluminum Compounds. The first process for the industrial oligomerization of ethylene was discovered by ZIEGLER in the early 1950s [72]. The reaction proceeds in two stages: first a growth step followed by an elimination step.

$$AlEt_3 + 3nCH_2{=}CH_2 \longrightarrow Al[(CH_2CH_2)_nEt]_3$$

$$Al[(CH_2CH_2)_nEt]_3 + 3\,CH_2{=}CH_2 \longrightarrow AlEt_3 + 3\,CH_2{=}CH(CH_2CH_2)_{n-1}Et$$

The growth step occurs at ca. 100 °C under 10 MPa ethylene pressure. In the second stage, the high temperature displacement reaction, the alpha-olefins are displaced by ethylene at ca. 300 °C and 1 MPa. The composition of the product mixture corresponds to a Poisson distribution. This process suffers from the disadvantage of requiring stoichiometric quantities of reaction components, i.e., industrial plants would require large quantities of aluminum alkyls. Consequently, Gulf Oil and Ethyl Corporation developed two process variants that require less triethylaluminum.

The Gulf Oil Process. In the Gulf process, only catalytic quantities of triethylaluminum are used and growth step and elimination reactions occur simultaneously in the same reactor [73]. The reaction is carried out at <200 °C and ca. 25 MPa. The product mixture consists of a Schulz–Flory distribution of linear olefins. A typical composition of individual fractions from C_6-C_{18} is listed in Table 9.

The Gulf process produces markedly purer α-olefins than can be obtained with paraffin cracking (see Table 8). The major impurities in the α-olefins (trade name: Gulftenes) obtained by the Gulf process are ca. 1.4 % paraffins; with increasing carbon number, an increasing amount of β-branched α-olefin is also formed. The occurrence of these vinylidene and 2-ethyl compounds may be rationalized by consecutive reactions of α-olefins with the aluminum alkyls, as follows:

$$Al(CH_2CH_2R^1)_3 + 3\,CH_2{=}CH{-}R^2 \longrightarrow Al(CH_2\overset{\overset{R^2}{|}}{C}HCH_2CH_2R^1)_3 \xrightarrow{3\,CH_2{=}CH_2} AlEt_3 + 3\,CH_2{=}\overset{\overset{R^2}{|}}{C}{-}CH_2CH_2R^1$$

The Gulf Oil Co. uses the Gulf process to produce linear α-olefins in the C_4-C_{30} range at their plant in Cedar Bajou, Texas. Mitsubishi Chemical Industries uses the same process at Mizushima.

Ethyl Process. The Ethyl process combines a stoichiometric and a catalytic stage [74]. This modified Ziegler process is capable of controlling the chain length of the resulting α-olefins much more precisely, because the shorter chain olefins may be subjected to further chain growth.

The Ethyl process is illustrated schematically in Figure 4. Ethylene is first oligomerized using catalytic quantities of triethylaluminum in analogy to the Gulf process. The resulting products are fractionated (b) in Fig. 4) yielding fractions containing C_4, C_6-C_{10}, and $C_{12}-C_{18}$. The higher boiling fractions may be used directly,

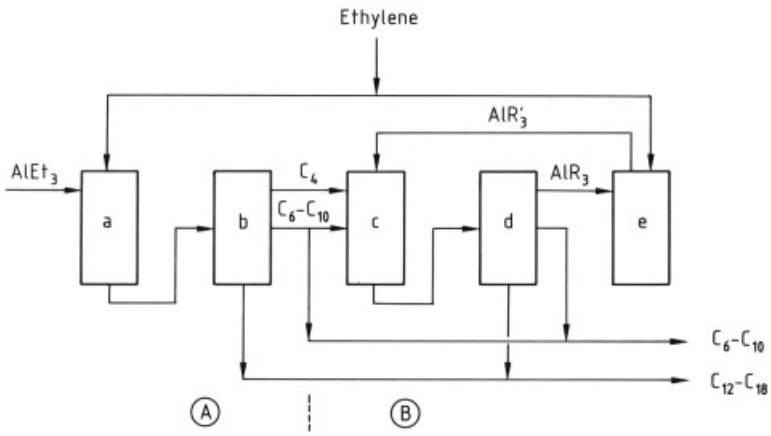

Figure 4. Typical flow diagram of Ethyl process
A) Catalytic step; B) Stoichiometric step;
a) Catalytic reactor; b) First distillation; c) Transalkylation reactor; d) Second distillation; e) Stoichiometric reactor

Table 10. Composition of fractions (wt %) obtained from the Ethyl process

	C_6	C_8	C_{10}	C_{12}	C_{12}–C_{14}	C_{14}–C_{16}	C_{16}–C_{18}
Linear α-olefins	97.5	96.5	96.2	93.5	87.0	76.0	62.7
Internal olefins	0.6	1.2	1.6	1.5	4.2	5.0	8.2
β-Branched α-olefins	1.9	2.3	2.2	5.0	8.8	19.0	29.1
Alkanes	0.1	0.6	0.3	0.4	0.4	0.4	0.8

whereas the shorter α-olefins, especially C_4, are subjected to transalkylation with long chain aluminum alkyls (c) in Fig. 4). This reaction releases the desired long-chain α-olefins; short-chain trialkylaluminum compounds are simultaneously formed. The aluminum trialkyls are separated in the second distillation step (Fig. 4, (d)) and transformed in another reactor (e) in Fig. 4) into longer chain alkyls by stoichiometric reaction with ethylene. The latter are then recycled to the transalkylation.

The Ethyl Corporation has operated a plant in Pasadena, Texas using this process since 1970; the initial capacity was 110 000 t/a. The process permits 95 % conversion of ethylene into higher α-olefins, a conversion efficiency not achievable by the single-step catalytic Gulf Oil process. This advantage is however, counterbalanced by a much lower product quality of the long-chain α-olefins. The C_{16}–C_{18} cut consists of only 63 % linear α-olefins (Table 10), the remainder being mainly β-branched α-olefins and internal olefins.

Ethylene Oligomerization with Transition-Metal Catalysis. In the presence of nickel, cobalt, titanium, or zirconium catalysts, ethylene may be converted into oligomers [75]. Commercial importance has been attained exclusively by the *Shell higher olefin process* (*SHOP*).

Shell Higher Olefin Process. The SHOP process is the most recent development in α-olefin synthesis. It works in the liquid phase using a nickel catalyst and yields α-olefins of unusually high purity [76–78]. Monoolefins are formed almost exclusively; analytical data indicate only trace amounts of dienes, aromatics, and alkanes. The content of α-olefin for the carbon numbers is in the range 96 – 97 %. A typical product composition is given in Table 11 (see next page).

The SHOP process involves a combination of three different reactions: *oligomerization, isomerization,* and *metathesis*. This integration of chain lengthening and shortening steps makes the SHOP process very flexible. Whereas all other α-olefin processes permit only limited control over product distribution, with the SHOP process it is possible to achieve almost any olefin cut desired. The SHOP process is shown schematically in Figure 5 [78–81].

Table 11. Composition of α-olefin fractions (wt %) obtained from the SHOP process

	C_6	C_8	C_{10}	C_{12}	C_{14}	C_{16}	C_{18}
All monoolefins	99.9	99.9	99.9	99.9	99.9	99.9	99.9
Linear α-olefins	97.0	96.5	97.5	96.5	96.0	96.0	96.0
Branched olefins	1.0	1.0	1.0	2.0	2.5	2.5	2.5
Internal olefins	2.0	2.4	1.0	1.5	1.5	1.5	1.5
Dienes	<0.1	<0.1	<0.1	<0.05	<0.05	<0.05	<0.05
Aromatics	<0.1	<0.1	<0.1	<0.05	<0.05	<0.05	<0.05
Alkanes	<0.1	<0.1	<0.1	<0.05	<0.05	<0.05	<0.05

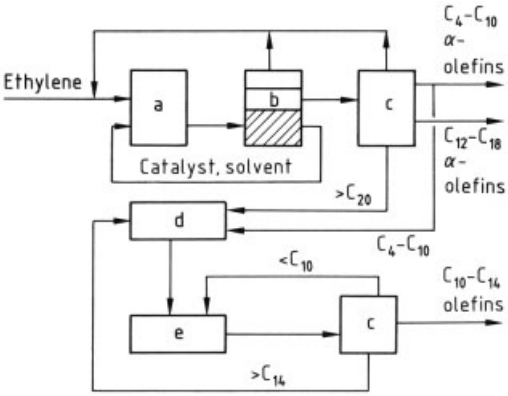

Figure 5. Typical flow diagram of SHOP process
a) Oligomerization reactor; b) Phase separator; c) Distillation; d) Isomerization reactor; e) Metathesis reactor

Oligomerization is carried out in a polar solvent such as ethylene glycol or 1,4-butanediol. The catalyst is produced in situ from a nickel salt, e.g., nickel chloride, sodium borohydride, and a chelating ligand. Suitable ligands are compounds of the general formula $RR^1P{-}CH_2{-}COR^2$, e.g., diphenyl phosphinoacetic acid, dicyclohexylphosphinoacetic acid, or 9-(carboxymethyl)-9-phosphabicyclo [3.3.1]- nonane. Ethylene is oligomerized to α-olefins with a Schulz – Flory distribution at 80 – 120 °C and 7 – 14 MPa. The olefins formed are immiscible with the polar solvent; product and catalyst phases are thereby readily separated so that the catalyst can be recycled repeatedly. Oligomerization is accomplished in a series of reactors interspersed with heat exchangers to remove the heat of reaction. The reaction rate can be regulated by the amount of catalyst. A high partial pressure of ethylene is required for high product linearity and a suitable rate of reaction. The chain length of the α-olefins is determined by the geometric factor K of molar growth:

$$K = \frac{n\,(C_{n+2}{-}\text{olefins})}{n\,(C_n{-}\text{olefins})}$$

where n is the number of moles.

The mass distribution of various α-olefins relative to the growth factor K is shown in Figure 6.

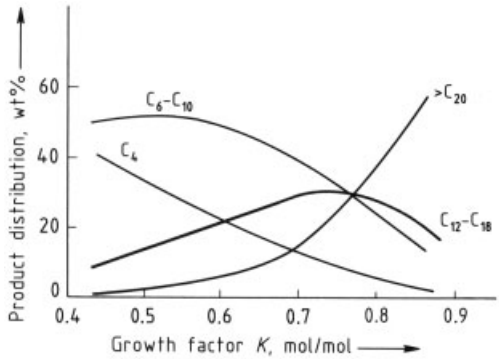

Figure 6. Product distribution as a function of growth factor K

Controlling the growth factor is extremely important for the overall process because K determines both the product distribution and the median chain length of the olefins obtained. It is, therefore, advantageous to be able to vary the K-factor by the composition of the catalyst.

After separation of the catalyst and product phase, the latter is washed with fresh solvent to remove the last traces of catalyst. In a subsequent distillation (see Fig. 5) the mixture is separated into the desired $C_{12} - C_{18}$ α-olefins and into low ($C_4 - C_{10}$) and high-boiling ($>C_{20}$) fractions. The $C_4 - C_{10}$ fraction can then be further fractionated, if desired, into individual compounds, for example, those used as comonomers for the manufacture of linear low density polyethylene (LLDPE). However, the $C_4 - C_{10}$ fraction and the $C \geq 20$ fractions are generally combined for isomerization and subsequent metathesis. These

two steps require only moderate reaction conditions: 80 – 140 °C, 0.3 – 2 MPa. The heat of reaction of both steps is quite low; because in isomerization only intramolecular migration of double bonds occurs, in metathesis carbon – carbon bonds are simultaneously cleaved and reformed. *Isomerization* takes place in the liquid phase in the presence of magnesium oxide catalyst [82]. About 90 % of the terminal olefins are converted to internal olefins in this reaction.

$$R-CH=CH_2 \xrightarrow{MgO} R^1-CH=CH-R^2$$

The subsequent joint *metathesis* of the internal, low, and high molecular mass olefins, mainly over heterogeneous rhenium or molybdenum catalysts, yields a mixture of internal olefins having a new chain length distribution, i.e., a random mixture of olefins of even and odd numbers of carbon atoms.

$$\begin{array}{c} R^1-CH=CH-R^2 \\ + \\ R^3-CH=CH-R^4 \end{array} \xrightarrow{\text{Re or Mo}} \begin{array}{c} R^1-CH \\ \| \\ R^3-CH \end{array} + \begin{array}{c} CH-R^2 \\ \| \\ CH-R^4 \end{array}$$

The products of metathesis are distilled and the $C_{11}-C_{14}$ fraction, representing ca. 10 – 15 % of the mixture, is separated for production of alcohols used in the detergent industry (see Section 2.1.3).

The composition of the internal $C_{11}-C_{14}$ olefins produced by the SHOP process is given in Table 12; more than 99.5 % is monoolefins, with only small amounts of dienes, aromatics, and paraffins.

Table 12. Composition of internal olefins (wt %) obtained from the SHOP process

	$C_{11}-C_{12}$	$C_{13}-C_{14}$
All monoolefins	>99.5	>99.5
Linear olefins	>96	>96
Branched olefins	3	3
Dienes	<0.1	<0.1
Aromatics	<0.1	<0.1
Paraffins	<0.5	<0.5

Although the double bonds of the olefins at the beginning of metathesis are distributed almost at random over the carbon backbone, they are shifted in the end product toward the chain end, a result of the high concentration of short-chain olefins (2-butene, 2-hexene, 3-hexene) in the metathesis feed.

Distillation of the metathesis products also yields low- and high-boiling components in addition to the $C_{11}-C_{14}$ fraction. One of the great advantages of the SHOP process is that these fractions do not have to be discarded but can be recycled. The fraction consisting of compounds with less than 11 carbon atoms is returned directly to the metathesis reactor; the fraction containing olefins with more than 14 carbon atoms is again subjected to isomerization before it reaches the metathesis stage. By combining oligomerization, isomerization, and metathesis, the SHOP process is capable of converting almost all the ethylene feed into olefins having the desired number of carbon atoms.

An interesting alternative of the SHOP process described above is the metathesis of the isomerized $C \geq 20$ fraction together with a ten-fold molar excess of ethylene [82]. Olefins, predominantly α-olefins, within the desired carbon range may be prepared in the presence of a rhenium catalyst (20 % Re_2O_7 on aluminum oxide).

$$\begin{array}{c} C_{10}H_{21}-CH=CH-C_{10}H_{21} \\ + \\ CH_2=CH_2 \end{array} \xrightarrow{Re_2O_7-Al_2O_3} 2\ C_{10}H_{21}CH=CH_2$$

The first SHOP processing plant was built in 1977 by Shell Oil Corporation at Geismar, Louisiana (see Section 2.1.4); annual capacity was 200 000 t.

The mechanism of the highly selective ethylene oligomerization with nickel chelate catalysts has been largely elucidated by the research group of W. Keim [76, 83–87]. A chelated nickel hydride complex is formed initially. This complex reacts with ethylene (Fig. 7, see next page) to form alkyl nickel complexes. Linear α-olefins are eliminated from the alkyl nickel complexes; the nickel hydride species is formed simultaneously to reenter the catalytic cycle.

The SHOP process is by far the most important method for oligomerization of ethylene involving transition-metal catalysts. To complete the picture, two other processes will be described below, although they have not yet reached commercial importance.

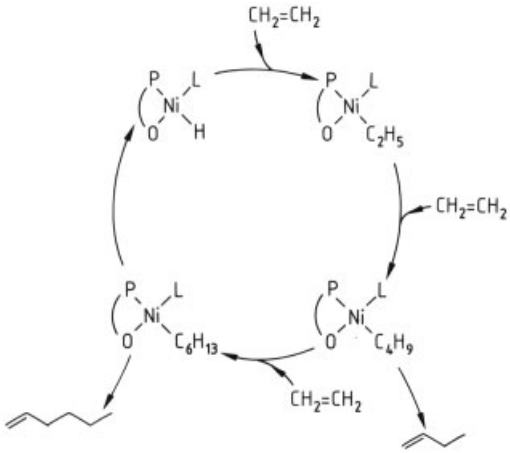

Figure 7. Mechanism of ethylene oligomerization in the presence of nickel P–O-chelate complex catalyst (for P–O ligand see text)

Exxon Process. The Exxon Corporation has developed a process in which ethylene is converted to high molecular mass olefins in the presence of a catalyst consisting of a transition-metal component and a soluble alkyl aluminum chloride [88]. Titanium chloride or an alkoxy titanium chloride are favored as the transition metal components. The catalyst is soluble in saturated hydrocarbons and active at 25 – 70 °C. This process yields linear α-olefins in the range C_4 – C_{100} or higher; it appears to be especially useful for the production of high-melting waxes, i.e., olefins higher than C_{20}.

Alphabutol Process. In the Alphabutol process, developed by the Institut Français du Pétrole (IFP), ethylene is selectively dimerized to 1-butene [89] using a homogeneous titanium catalyst. In contrast to the SHOP process, the mechanism does not involve metal hydrides but a titanium(IV) cyclopentane species that decomposes under the reaction conditions (50 – 60 °C, slight positive pressure) with β-hydrogen transfer to 1-butene:

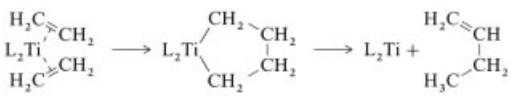

The 1-butene, which contains only traces of impurities (hexenes, 2-butene, butanes) may be used as comonomer in the production of LLDPE.

Oligomerization of Propene and Butenes with Transition-Metal Catalysis. Ziegler catalysts consisting of nickel compounds and aluminum trialkyls can be used to dimerize or codimerize propene and butene to give predominantly branched olefins. This process was also developed by the Institut Français du Pétrole and is referred to as the Dimersol process [90–95]. IFP research began as early as the mid 1960s. The first commercial Dimersol-G plant was brought into production in 1977 by Total in Alma, Michigan. In the *Dimersol G process*, propene (in admixture with propane) is dimerized mainly (80 %) to isohexenes; small amounts of propene trimers (18 %) and tetramers (2 %) are also formed:

$$2\ C_3H_6 \longrightarrow$$

n-Hexenes + 2-methylpentenes + 2,3-dimethylbutenes

The reaction takes place in the liquid phase without solvent. The pressure in the reactor is adjusted so that the C_3 compounds remain liquid at the operating temperature. The product mixture leaving the reactor head is fed to a vessel where the catalyst is neutralized with alkali. The product (trade name: Dimate) is then washed with water and freed of any remaining propane and propene on a stabilizing column.

The *Dimersol X process* is similar, although the starting materials are butenes or a mixture of propene and butenes. The products formed are, accordingly, isooctenes or mixtures of C_6 – C_8 olefins. The first Dimersol X plant was started in 1980 by Nissan Chem. in Kashima, Japan. A typical starting material for the Dimersol X process is the raffinate II obtained from the C_4 cut of the steam cracker, after removal of butadiene and isobutene. The resulting C_8 olefins can be converted by Oxo-synthesis into nonanols, which are used for the synthesis of plasticizers (dinonyl phthalate).

A more recent development is the *Dimersol E* process, which employs the exhaust gases of fluid catalytic crackers (FCC) that contain ethylene and propene. A gasoline fraction can be obtained from this product after hydrogenation. The first experiments with cracker exhaust gases from the Elf cracker near Feyzin (France) began in 1984.

By 1982, IFP had licensed 16 Dimersol plants, eight of which were in production. In addition to IFP, other industrial and academic groups have investigated the transition metal-catalyzed oligomerization of lower olefins [96–104]. The observation that in the propene dimerization with nickel-based Ziegler catalysts the addition of phosphines may significantly affect the isohexene product distribution is of interest. With the catalyst $[NiBr(\eta^3\text{-}C_3H_5)(PCy_3)] - EtAlCl_2$ (Cy = cyclohexyl) a remarkable turnover of 800 000 mol propene per mol nickel h^{-1} could be achieved. A turnover of 60 000 s^{-1} was calculated for operation at room temperature. This value is of an order of magnitude usually achieved only by enzyme reactions.

Oligomerization of Propene and Butene by Acid Catalysis. Mineral acids such as phosphoric or sulfuric acid can also be used as catalysts for the oligomerization of olefins; the major products are branched olefins. Alkylbenzenes were once prepared from tetramers of propenes (tetrapropylene). This reaction is now rarely used in Europe because of the low biodegradability of the alkylbenzenesulfonates obtained.

Oligomerization of isobutene leads to the formation of diisobutene, triisobutene, and tetraisobutene; after hydrogenation, these may be added to gasoline to improve the octane rating (polymer gasoline, oligomer gasoline). In the *cold acid process*, isobutene-containing C_4 fractions are treated with 60 – 70 % sulfuric acid at room temperature. Isobutene dissolves in the sulfuric acid as *tert*-butyl sulfate. This solution is then oligomerized at 100 °C. The products, mainly diisobutenes, separate as a second phase. In the *hot acid process*, the *n*-butenes of the C_4 fraction are also converted, forming $i\text{-}C_4 - n\text{-}C_4$ cooligomers.

The oligomerization can also be performed with poly(styrene sulfonic acid) resins as catalysts. This process is used by Erdölchemie – Dormagen using the raffinate I (C_4 fraction after extraction of butadiene). The reaction is performed in loop reactors at 100 °C and 1.5 – 2.0 MPa with residence times of 20 – 60 min. The product is composed of 36 % diisobutene, 22 % C_8 codimers, 38 % triisobutene, and 4 % tetraisobutene. The product distribution may be changed by recycling specific fractions.

2.1.3. Uses of Higher Olefins

The most important uses of linear olefins are described below.

Synthesis of Oxo-Alcohols. Linear α-olefins can be converted to linear primary alcohols, which are used in the synthesis of plasticizers ($<C_{11}$) or detergents ($>C_{11}$) by hydroformylation (oxo synthesis). The reaction can also be used to give aldehydes as the major products.

$$R-CH=CH_2 \xrightarrow[\text{Cat.}]{CO,\ H_2} R-CH_2CH_2CH_2OH \text{ or } R-CH_2CH_2CHO$$

The yield of linear products can be increased to >90 % if phosphine-modified cobalt- or rhodium catalysts are used.

The described synthesis route for long-chain alcohols has strong competition, however, from the natural fatty alcohols and from the Alfoles, generated from ethylene by an organo aluminum synthesis with subsequent oxidation.

In the United States Monsanto has produced oxo alcohols from $C_6 - C_{10}$ α-olefins since 1971. Exxon has used the oxo process since 1983 to produce plasticizer alcohols in the $C_7 - C_{11}$ range at Baton Rouge, Louisiana. Shell also converts olefins to alcohols. In recent years, olefins derived from paraffin crackers as starting materials have been replaced by internal olefins from the SHOP process.

The possibility of converting olefin-derived aldehydes to short-chain synthetic fatty acids by oxidation is also notable:

$$R-CH_2CH_2CHO \xrightarrow{\text{Oxid.}} R-CH_2CH_2COOH$$

In 1980 Celanese (now Hoechst-Celanese) built a plant in Bay City, Texas with an annual capacity of 18 000 t to produce heptanoic and nonanoic acids from 1-hexene and 1-octene, respectively, which are used in the production of lubricants. Again, there is strong competition from the natural products, which are generally less expensive. In addition, lubricant production requires highly linear acids. Fatty acids from

oxo aldehydes have a linearity of >95 %; natural fatty acids have a linearity of almost 100 % (→Fatty Acids).

Synthesis of Linear Alkylbenzenes. Currently the most important class of detergents for domestic use is still the linear alkylbenzenesulfonates (LAS or LABS). They are prepared by Friedel–Crafts reaction of benzene with linear olefins followed by sulfonation:

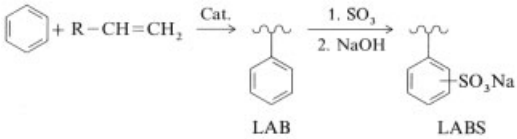

Total estimated production of LAS in Western Europe, Japan, and the United States in 1982 was ca. 1.1×10^6 t. In Europe, Asia, South Africa, and South America linear alkylbenzenes are produced from $C_{10}-C_{15}$ *alpha*-olefins. In 1982, 420 000 t of LAB were synthesized in Western Europe from ca. 80 000 t of linear olefins.

Synthesis of α-Alkenesulfonates. α-Alkene sulfonates (α-olefin sulfonates, AOS or OS) are obtained by direct sulfonation of $C_{14}-C_{18}$ α-olefins (→Detergents). An ionic mechanism results in a product mixture of 40 % isomeric alkenesulfonic acids (with the double bond predominantly in the β- or the γ-position) and 60 % 1,3- or 1,4-alkanesulfones. Following hydrolysis with sodium hydroxide, a mixture of alkenesulfonates and hydroxyalkanesulfonates is obtained:

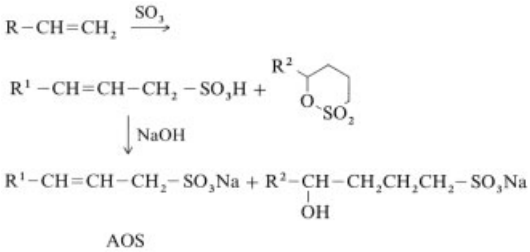

The double bonds and the hydroxyl groups contribute to the excellent water solubility of the AOS compounds. They are effective laundry agents even in hard water and at low concentration and are used increasingly in liquid soaps and cosmetics. In order to be used in such products the AOS must be of high quality, especially with regard to color; this is only possible with α-olefins that contain no more than traces of dienes.

Although AOS compounds have been known since the 1930s, they gained commercial importance only in the late 1960s. In 1982, ca. 7000 t of AOS were produced in the United States from α-olefins, with the Stepan Company and Witco Chemical Corporation being the largest U.S. producers. AOS represents an interesting alternative to alkyl sulfates and alkyl ether sulfates in cosmetic products.

Synthesis of Bromoalkanes and Derived Products. An important reaction of α-olefins is radical hydrobromation to give primary bromoalkanes. These compounds are valuable intermediates for, e.g., thiols, amines, amine oxides, and ammonium compounds:

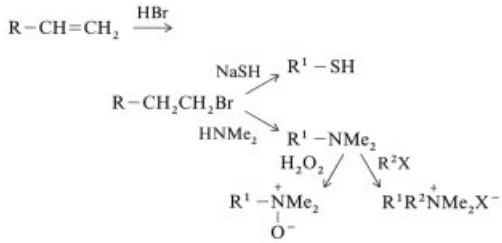

Thiols are produced by this route in Europe and in the United States by Pennwalt Corp. and Phillips Petroleum Co. They are used in the synthesis of herbicides, pesticides, pharmaceuticals, in the textile industry, or as polymer regulators.

In 1982, ca. 15 000 t of alkyldimethylamine (ADA) was produced from linear α-olefins in the United States by Ethyl Corporation and by Procter & Gamble. Most of these amines were directly converted to fatty amine oxides (FAO) or quaternary ammonium salts (Quats). Fatty amine oxides are nonionic surfactants that, because they are biodegradable and do not irritate the skin, are incorporated into numerous products, e.g., into rinsing agents. Quats are produced by alkylation of amines, e.g., with benzylchloride. Benzyl-Quats possess biocidal and antimicrobial properties and are incorporated,

for example, into disinfectants and deodorants (→ Laundry Detergents).

Production of Synthetic Lubricants (→Lubricants and Related Products). Oligomerization of α-olefins in the middle C-number range, in particular 1-decene, to isoparaffins (mostly trimers with a few tetramers and pentamers) leads to lubricants known as poly(α-olefins) (PAO) or synthetic hydrocarbons (SHC). They are used industrially, and in the automobile and aircraft sectors [105, 106]. They are in many instances superior to natural mineral oils, especially at lower temperature, because of their favorable viscosity, low volatility, high flash point, and high thermal stability. Oligomerization of the α-olefins can be performed with, e.g., BF_3, to which protonic cocatalysts, e.g., water or alcohols, are added.

In 1982, ca. 25 000 t of PAO were produced in the United States. The major producers are Mobil Oil, Gulf Oil, National Distillers, Burmah-Castrol, and Ethyl Corporation.

Production of Copolymers. Linear α-olefins, in particular 1-butene, 1-hexene, and 1-octene, are used as comonomers in the production of high density polyethylene (HDPE) and linear low density polyethylene (LLDPE). By adding α-olefins the density and other properties of the polymers may be significantly modified. HDPE copolymers contain only relatively small quantities of α-olefins, generally 0.5 – 3 %. The density of the HDPE homopolymer (0.965 – 0.955 g/cm^3) is reduced to 0.959 – 0.938 g/cm^3 for the HDPE copolymer. The most frequently employed comonomer is 1-hexene, followed by 1-butene.

On the other hand, 4 – 12 % α-olefin is generally added to LLDPE to reduce the density to 0.935 – 0.915. The more α-olefins added to the polymer, the greater the number of short-chain branches and the lower the density. Addition of 1-butene is preferred in many of the gas-phase processes, e.g., in the Unipol process introduced by Union Carbide in 1977 [107], whereas 1-octene and 1-hexene are preferred in liquid-phase processes.

Major United States producers are Dow Chemical, DuPont, Exxon, Gulf Oil, Union Carbide, Soltex Polymer Corporation, Phillips Petroleum Company, and Allied Corporation. An estimated 85 000 t of α-olefins were copolymerized in HDPE and LLDPE in the United States in 1982.

Because Shell began marketing SHOP olefins in Western Europe, the use of HDPE and LLDPE copolymers has increased in these countries. Olefins derived from paraffin cracking plants are unsuitable as comonomers for polyethylene because of the impurities they contain.

Additional Reactions of α-Olefins. The reaction of α-olefins with peracid to form epoxides is an interesting route to bifunctional derivatives [108]. Some examples are given below (see also →Epoxides, Chap. 2.1.).

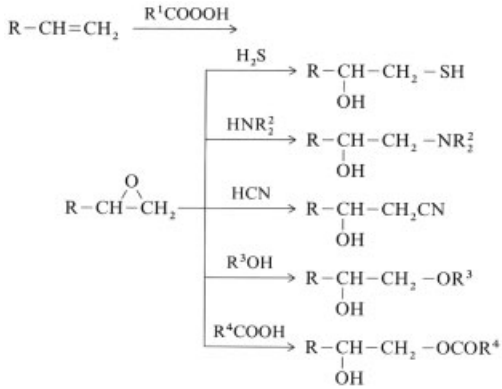

α-Olefin epoxides have also been used in polymer chemistry, e.g., for the modification of epoxy resins. United States producers of α-olefin epoxides include: Union Carbide, Viking, and Dow Chemical.

Another use of α-olefins is conversion into secondary monoalkylsulfates, which are readily degradable surfactants useful in wetting agents and household detergents. They are synthesized by reaction with concentrated sulfuric acid at low temperature (10 – 20 °C) and with short reaction times (5 min):

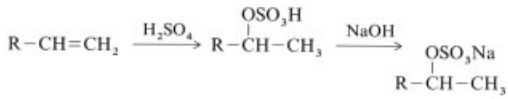

Hydrocarboxylation of α-olefins leads to odd-numbered carboxylic acids. This reaction is carried out with carbon monoxide and water at

Table 13. Linear α-olefin producers (January 1984)

Manufacturer	Location	Process	Annual capacity (10^3 t)
Gulf Oil	Cedar Bayou, Texas	Oligomerization (Gulf process)	90
Mitsubishi	Mizushima, Okayama	Oligomerization (Gulf process)	30
Ethyl Corp.	Pasadena, Texas	Oligomerization (Ethyl process)	450
Chevron	Richmond, California	Paraffin cracking	40
Shell	Berre (France)	Paraffin cracking	90
	Pernis (The Netherlands)	Paraffin cracking	170
	Geismar, Louisiana	SHOP	270
	Stanlow (UK)	SHOP	170
		Total	1310

200 °C and 20 MPa. The formation of linear carboxylic acids is favored in the presence of cobalt carbonyl – pyridine catalysts. Both the free acids and their esters are used as lubricant additives.

$$R-CH=CH_2 \xrightarrow[Co]{CO,\ H_2O} R-\underset{\underset{COOH}{|}}{CH}-CH_3 + R-CH_2CH_2COOH$$

The alkenylsuccinic anhydrides (ASA) should also be mentioned. They are obtained by heating maleic anhydride with α- or internal olefins, and are used as lubricant additives, detergents, and in leather and paper production. Olefins with more than 30 carbon atoms are waxy and can be used as paraffin substitutes. In 1984 ca. 8000 t of α-olefins were used in the United States for the production of candles, crayons, and coatings.

2.1.4. Economic Aspects of Higher Olefins

Production facilities for α-olefins and their capacities (1984) are listed in Table 13.

The α-olefin market has been subject to fluctuations [109]. Ethyl Corporation, the largest United States α-olefin producer, announced at the end of 1987 the intention to build a large α-olefin plant in Europe with an annual capacity of 250 000 t [110]. This plant would make Ethyl the largest α-olefin producer in Europe.

Shell has also been planning for some time to expand its SHOP processing capacity. The SHOP plant in Geismar, United States, which has been operating since 1977, was expanded in 1982 to an annual capacity of 270 000 t. Construction of a second SHOP unit in Geismar had been announced for 1984 [111]. According to [112], this second plant should come on stream in 1989 with an annual capacity of 243 000 t. Additional expansion is to bring the total annual capacity at Geismar to 590 000 t.

Shell has also announced changes in Europe [113]. The plant at Stanlow, United Kingdom, operating since 1982, is to be expanded to an annual capacity of 220 000 t; once this plant is fully operational, the plant at Pernis, The Netherlands which for some time has produced only 60 000 – 70 000 t/a, is to be closed down.

Mitsubishi Petrochemical has expressed an interest in constructing a SHOP plant in Kashina, Japan in a joint venture with Shell [114]. Other projects under discussion are SHOP plants in New Zealand and in Canada [115].

Uses of linear olefins in the United States and in Western Europe are given in Table 14.

Table 14. Uses of higher linear olefins (1982) (estimate by the Stanford Research Institute)

Final products	Olefin consumption (10^3 t)	
	United States *	Western Europe
Oxo alcohols	67	145 – 160
Amines and derivatives	13 – 17	**
α-Olefinsulfonates (AOS)	7	3
Linear alkylbenzenes (LAB)	1 – 2	75 – 80
Copolymers (HDPE, LLDPE)	85	**
Synthetic lubricants (SHC)	25	**
Lubricant additives	15	20

* Figures refer exclusively to α-olefins.
** Unknown.

2.2. Dienes and Polyenes

2.2.1. Low Molecular Mass 1,3-Dienes

Industrially, the most important 1,3-dienes are butadiene, isoprene, and cyclopentadiene

(→Butadiene; →Cyclopentadiene and Cyclopentene; →Isoprene). The conjugated dienes listed below are of lesser importance.

Piperylene (1,3-pentadiene) is present at a level of >10 % in the C_5 steam cracker fraction; it is expensive to separate from the C_5 cut. Polymers and copolymers with, for example, butadiene have been produced but have not yet found a technical application.

2,3-Dimethylbutadiene was the first monomer to be converted on an industrial scale to a synthetic rubber, referred to as methyl rubber. It was developed by HOFFMANN in 1910 and produced during 1914 – 1918. It suffered, however, from several disadvantages (excessive hardness, sensitivity to oxidation) and later lost its importance. The early synthesis of the monomer involved conversion of acetone to pinacol, which is then dehydrogenated to the diene in the presence of aluminum oxide as catalyst. Modern synthesis starts with propene, which is dimerized to 2,3-dimethylbutene [103, 104] and subsequently dehydrogenated:

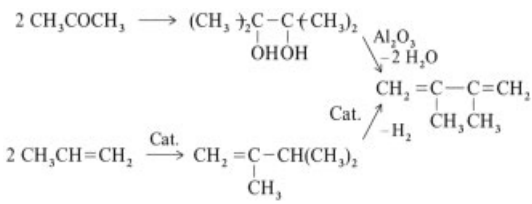

Chloroprene (→Chlorinated Hydrocarbons, Chap. 6.4.), $CH_2{=}CCl{-}CH{=}CH_2$, 2-chloro-1,3-butadiene, is used to form polychloroprene (neoprene), which is extremely resistant to oil, abrasion, and ageing. As with all polymers containing a large proportion of halogen, these products are resistant to ignition. Vulcanized polychloroprene is used in, for example, conveyer belts, cable shielding, and protective clothing. Nonvulcanized polymers are used as contact adhesives in plastics manufacturing.

Chloroprene can be produced by two routes. In the older process, acetylene is first dimerized to vinyl acetylene and hydrogen chloride then added to the triple bond; both steps take place in the presence of CuCl as catalyst. The second process involves chlorination of butadiene, at 300 °C with chlorine to yield a mixture of dichlorobutenes. The 1,2-adduct is subsequently dehydrochlorinated to yield chloroprene. The 1,4-adduct can be isomerized to the 1,2-adduct:

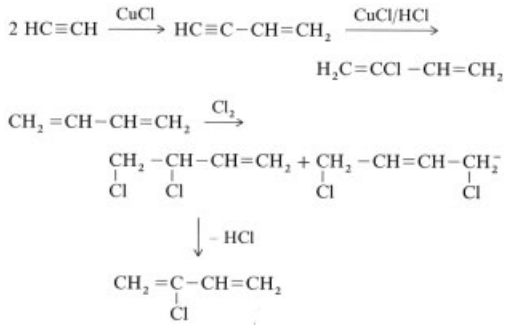

2.2.2. Synthesis of Dienes and Polyenes by Oligomerization

A multitude of linear and cyclic dienes and polyenes can be synthesized by transition metal-catalyzed oligomerization or by cooligomerization [75]. Only a few examples will be described below, some of which are used commercially. Butadiene can be converted into six cyclooligomers in a nickel-catalyzed reaction [116]. The synthesis can be steered in the desired direction by a suitable choice of catalyst.

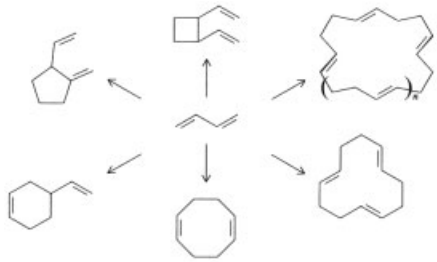

1,5-Cyclooctadiene is produced by Shell and by Hüls and is used, for example, as starting material for suberic acid and poly(octenamer) (→Cyclododecatriene, Cyclooctadiene, and 4-Vinylcyclohexene).

1,5,9-Cyclododecatriene is produced by DuPont, Shell, and Hüls using nickel or titanium as catalyst; it is used in the synthesis of nylon 12, dodecandioic acid, and hexabromocyclododecane.

4-Vinylcyclohex-1-ene is also accessible from butadiene but is, as yet, of little commercial importance. The possibility of using it as a precursor in styrene synthesis has been discussed.

An example of catalytic cooligomerization is the reaction between butadiene and ethylene [117].

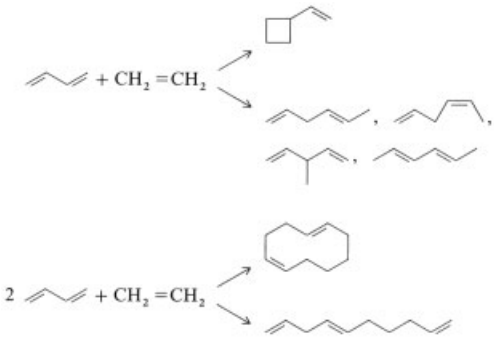

Whereas the cyclic 1 : 1 product vinylcyclobutane and the C_{10} oligomers cyclodecadiene and 1,4,9-decatriene are not yet used commercially, the linear oligomer, 1,4-hexadiene, is produced as a monomer for the production of ethylene – propene – diene (EPDM) elastomers. Homogeneous catalysis with rhodium or palladium complexes leads to *trans*-1,4-hexadiene; cobalt and iron catalysts yield predominantly the *cis*-1,4-product. Rhodium catalysts are used in commercial production.

Ethylidenenorbornene, also a comonomer for EPDM polymers, is also accessible via cooligomerization. One of the routes employs the Diels – Alder reaction starting with butadiene and cyclopentadiene, the other couples norbornadiene and ethylene in the presence of a nickel catalyst. Vinylnorbornene is formed as an intermediate in both routes.

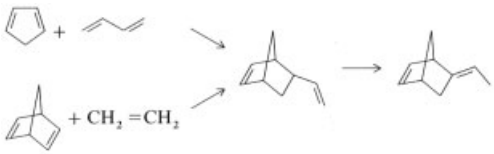

2.2.3. Synthesis of Dienes and Polyenes by Metathesis

Metathesis offers a simple path to α,ω-diolefins that are not easily obtained by other routes [118]. Starting materials are cycloolefins, which can be prepared, for example, by cyclooligomerization of butadiene (see Section 2.2.2) and subsequent partial hydrogenation. Metathetic conversion with ethylene opens the cycloolefin to an α,ω-diene. In this manner, 1,5-cyclooctadiene gives two molecules of 1,5-hexadiene; cyclooctene forms 1,9-decadiene; and cyclododecene leads to 1,13-tetradecadiene.

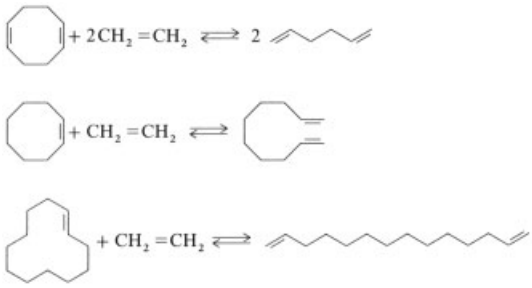

Heterogeneous rhenium catalysts are preferred because they allow the reaction to proceed under very mild conditions (5 – 20 °C, 0.1 – 0.2 MPa ethylene). A plant with an annual capacity of 3000 t using this FEAST (Further Exploitation of Advanced Shell Technology) process was brought on stream by Shell in Berre (France) in 1987. The products are suitable for the synthesis of numerous fine chemicals, e.g., production of long-chain diketones, for dibromides, diepoxides, dialdehydes, dithiols, and diamines.

4-Vinylcyclo-1-hexene undergoes self metathesis. Elimination of ethylene yields 1,2-bis(3-cyclohexenyl)ethylene, which has been used as a starting material for fire retardants.

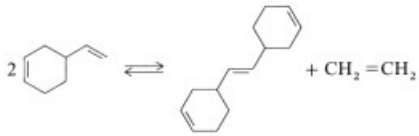

3. Alkylbenzenes

Toluene, Xylenes, and Ethylbenzene are separate keywords.

Table 15. Physical properties of alkylbenzenes [119]

Compound	CAS registry number	M_r	*mp*, °C	*bp*, °C	d_4^{20}, g/cm^3	n_D^{20}
1,2,3-Trimethylbenzene, hemimellitene	[*526-73-8*]	120.19	−25.4	176.1	0.8944	1.5139
1,2,4-Trimethylbenzene, pseudocumene	[*95-63-6*]	120.19	−43.8	169.3	0.8758	1.5048
1,3,5-Trimethylbenzene, mesitylene	[*108-67-8*]	120.19	−44.7	164.7	0.8652	1.4994
1,2,3,4-Tetramethylbenzene, prehnitene	[*488-23-3*]	134.22	− 6.2	205	0.9052	1.5203
1,2,3,5-Tetramethylbenzene, isodurene	[*527-53-7*]	134.22	−23.7	198	0.8903	1.5130
1,2,4,5-Tetramethylbenzene, durene	[*95-93-2*]	134.22	79.2	196 – 198	0.8380 [a]	1.4790 [a]
Pentamethylbenzene	[*700-12-9*]	148.25	54.5	232	0.917	1.527
Hexamethylbenzene, mellitene	[*87-85-4*]	162.28	166 – 167	265	1.0630 [b]	
1,2-Diethylbenzene	[*135-01-3*]	134.22	−31.2	183.5	0.8800	1.5035
1,3-Diethylbenzene	[*141-93-5*]	134.22	−83.9	181.1	0.8602	1.4955
1,4-Diethylbenzene	[*105-05-5*]	134.22	−42.8	183.8	0.8620	1.4967
1,2,3-Triethylbenzene	[*42205-08-3*]	162.27				
1,2,4-Triethylbenzene	[*877-44-1*]	162.27		217.5 [c]	0.8738	1.5024
1,3,5-Triethylbenzene	[*102-25-0*]	162.27	−66.5	215.9	0.8631	1.4969
1,2,3,4-Tetraethylbenzene	[*642-32-0*]	190.33	11.8	251 [d]	0.8875	1.5125
1,2,3,5-Tetraethylbenzene	[*38842-05-6*]	190.33	−21	247 [d]	0.8799	1.5059
1,2,4,5-Tetraethylbenzene	[*635-81-4*]	190.33	10	250	0.8788	1.5054
Pentaethylbenzene	[*605-01-6*]	218.38	<−20	277	0.8985 [e]	1.5127
Hexaethylbenzene	[*604-88-6*]	246.44	129	298	0.8305 [f]	1.4736 [f]
1-Ethyl-2-methylbenzene, 2-ethyltoluene	[*611-14-3*]	120.19	−80.8	165.2	0.8807	1.5046
1-Ethyl-3-methylbenzene, 3-ethyltoluene	[*620-14-4*]	120.19	−95.5	161.3	0.8645	1.4966
1-Ethyl-4-methylbenzene, 4-ethyltoluene	[*622-96-8*]	120.19	−62.3	162	0.8614	1.4959

[a] At 81 °C.
[b] At 25 °C.
[c] 100.4 kPa.
[d] 97.6 kPa.
[e] At 19 °C.
[f] At 130 °C.

3.1. Trimethylbenzenes

Properties. The three isomers of trimethylbenzene are colorless liquids. Some of their physical properties are listed in Table 15.

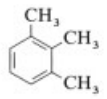

1,2,3-Trimethylbenzene
Hemimellitene

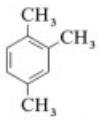

1,2,4-Trimethylbenzene
Pseudocumene

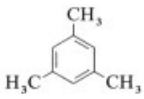

1,3,5-Trimethylbenzene
Mesitylene

The methyl groups of trimethylbenzenes are converted to carboxyl groups on oxidation. Dilute nitric acid oxidizes mesitylene to the corresponding tricarboxylic acid, trimesic acid [*554-95-0*]. In pseudocumene the preferred site of reaction for halogenation, sulfonation, and nitration is the 5-position; disubstitution leads to 3,5-derivatives. Mild oxidation of pseudocumene or mesitylene, with manganese- or lead dioxide, or electrolytically, leads to 2,4- or 3,5-dimethylbenzaldehyde, respectively.

Production. Hemimellitene, pseudocumene, and mesitylene are found in coal tar oil. The compounds can be isolated by distillation and purified by the differential hydrolysis of their sulfonic acids [120]. More important industrially is the occurrence of trimethylbenzenes in some mineral oils, and their formation during processing of crude oil, especially catalytic cracking and reforming. Pseudocumene is obtained in 99 % purity by superfractionation from the C_9 cut of crude oil refining [121]. Pseudocumene and hemimellitene can be obtained by extractive crystallization of the C_9 cut with urea [122]. Trimethylbenzenes can be produced by Friedel – Crafts methylation of toluene or xylene with chloromethane in the presence of aluminum chloride. Trimethylbenzenes, especially pseudocumene and mesitylene, are also formed by liquid-phase disproportionation of xylene in the presence of aluminum chloride. These reactions can be carried out in the gas phase at high temperature on aluminum silicate catalysts.

A more recent method for the methylation of xylene with methanol or dimethyl ether on fluorine-containing crystalline borosilicate, aluminosilicate, or boroaluminosilicate (ZSM-5) at 300 °C and atmospheric pressure yields pseudocumene and durene almost exclusively, both of which may be isolated at >99 % purity [123]. Trimethylbenzenes are also generated in the reaction of toluene with synthesis gas over a dual catalyst consisting of a metal oxide component and an aluminum silicate, at 200 – 400 °C and up to 20 MPa [124]. Pseudocumene is also generated from methanol as a byproduct of the Mobil MTG (methanol to gasoline) process (see Section 3.2). Fusel oils and other fermentation products may be reacted on the zeolite catalyst HZSM-5 to form aromatic compounds, including trimethylbenzenes [125].

Uses. Mesitylene is used in the synthesis of dyes and antioxidants, and as a solvent. Pseudocumene is a starting material for dyes, pharmaceuticals and, especially, for trimellitic anhydride [*552-30-7*], which is then converted to heat-resistant polyamideimides and polyesterimides (see also Carboxylic Acids, Aromatic, Chap. 4.3.). Hemimellitene is an intermediate in the production of fragrances (→ Flavors and Fragrances).

3.2. Tetramethylbenzenes

Properties. Some physical properties of the three tetramethylbenzene isomers are listed in Table 15. Prehnitene and isodurene are colorless liquids; durene forms colorless, monoclinic crystals that sublime slowly at room temperature and that have an odor reminiscent of camphor.

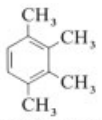

1,2,3,4-Tetramethylbenzene
Prehnitene

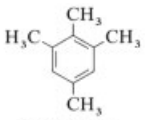

1,2,3,5-Tetramethylbenzene
Isodurene

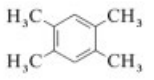

1,2,4,5-Tetramethylbenzene
Durene

Tetramethylbenzenes are readily soluble in aromatic hydrocarbons, acetone, ether, and alcohol. Durene is highly reactive; nitration, chlorination, sulfonation, and chloromethylation occur almost as readily as with phenol. The methyl groups may be successively oxidized to carboxylic acids; the presence of the corresponding mono-, di-, tri-, and tetracarboxylic acids can be demonstrated during liquid-phase oxidation of durene in the presence of cobalt catalyst [126].

Production. The tetramethylbenzenes also occur in coal tar from anthracite and lignite, and in the reformed fraction of oil refineries. The isolation of durene by crystallization is aided by its relatively high melting point; isodurene becomes enriched in the mother liquor. Methylation of xylene leads to an isomeric mixture, from which durene, prehnitene, and isodurene can be obtained [120]. Durene is produced by liquid-phase methylation of pseudocumene with chloromethane in the presence of aluminum chloride [127]. In more recent processes durene is produced by reaction of xylene or pseudocumene with methanol in the gas phase over aluminum silicates; the most successful process uses the zeolites ZSM-5 and TSZ [128, 129]. In the MTG process, commercialized in New Zealand since 1986, the production of automotive fuels from methanol over ZSM-5 yields an unacceptable quantity of durene, which has to be removed from the fuel to avoid carburetor clogging by the high melting point impurity. The durene is currently removed by catalytic isomerization, disproportionation, and reductive demethylation. The possibility of isolating durene from the MTG process is being investigated [130].

Uses. Oxidation of durene leads to the commercially important pyromellitic dianhydride [*89-32-7*] used as a starting material for heat-resistant polyimides (→Polyimides) and as a hardener for epoxy resins (→Epoxy Resins). Pyromellitic esters serve as migration-resistant plasticizers for poly(vinyl chloride) (→Carboxylic Acids, Aromatic, Chap. 4.2.).

3.3. Penta- and Hexamethylbenzene

Properties. Some physical properties are listed in Table 15. Both compounds are crys-

talline at room temperature; they dissolve easily in benzene and ethanol. The methyl groups are readily oxidized to carboxyl groups. Irradiation with light converts hexamethylbenzene into its valence isomer hexamethyl Dewar benzene (1,2,3,4,5,6-hexamethylbicyclo [2.2.0] hexa-2,5-diene, [*7641-77-2*]). Electrophilic substitution on the free ring site of pentamethylbenzene proceeds readily.

Production. Penta- and hexamethylbenzene are generated by liquid-phase Friedel – Crafts alkylation of xylene or pseudocumene with chloromethane in the presence of aluminum chloride [127]. If the reaction is carried out at 140 – 180 °C in *o*-dichlorobenzene, hexamethylbenzene with a purity of 98 % can be crystallized directly from the reaction mixture [131]. In a more recent method, penta- and hexamethylbenzene can be produced simultaneously by reaction of phenol and methanol in toluene in the presence of a zinc aluminate catalyst at 330 – 360 °C [132]. The products are then separated by fractional crystallization. Hexamethylbenzene is formed by thermal rearrangement of hexamethyl Dewar benzene, which is readily obtained from 2-butyne [133].

Uses. Pentamethylbenzene is of little commercial importance. Hexamethylbenzene is an intermediate in the fragrance industry. It is of scientific interest in studies of charge transfer complexes and exciplexes, in which, like other methylbenzenes, it may function as donor [134].

3.4. Diethylbenzenes

Properties. For some physical properties see Table 15.

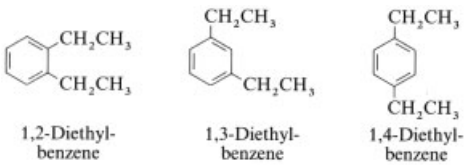

Production. The three isomeric diethylbenzenes (DEB) occur as byproducts in the production of ethylbenzene from benzene and ethylene but are, generally, recycled to undergo transalkylation (→Ethylbenzene). In the liquid-phase Friedel – Crafts process with aluminum chloride an increase in the feedstock ratio of ethylene to benzene may raise the DEB concentration in the output to 35 %. The DEB fraction consists of ca. 65 % 1,3-DEB, 30 % 1,4-DEB, and only 5 % 1,2-DEB. A similar isomer distribution is found in the product mixture from gas-phase alkylation of ethylbenzene with ethylene on an unmodified HZSM-5 zeolite catalyst. By modifying the catalyst, the zeolite pore may be narrowed to such an extent that, based on geometry alone, only 1,4-diethylbenzene is formed, rather than the bulkier 1,2- and 1,3-isomers. The Mobil gas-phase alkylation, which is based on such a shape-selective zeolite, converts ethylbenzene to DEB at 400 °C and 0.7 MPa with 88 % selectivity; more than 99 % of the DEB is the 1,4-isomer [135]. Nearly quantitative formation of 1,3-DEB occurs at low temperature in the liquid phase by disproportionation of ethylbenzene in the presence of at least 10 vol % HF and 0.3 – 0.5 mol BF_3 per mol ethylbenzene.

Uses. Diethylbenzenes may be converted to divinyl benzenes by catalytic dehydrogenation (→Hydrogenation and Dehydrogenation). The latter are useful in the production of cross-linked polystyrenes. DEB is contained in the product of the MTG process (see Section 3.2) and improves the properties of the gasoline.

3.5. Triethylbenzenes and more highly Ethylated Benzenes

Properties. For some physical properties see Table 15.

Production. The more highly ethylated benzenes are produced in the liquid phase in a similar manner as that described for diethylbenzene. 1,3,5-Triethylbenzene may be generated in excellent yield by disproportionation of diethylbenzene with BF_3 and HF at 75 °C. The same result is obtained with $AlCl_3$ – HCl as catalyst. 1,3,5-Triethylbenzene is also formed by cyclocondensation of 2-butanone in the presence of sulfuric acid. Hexaethylbenzene is formed in nearly quantitative yield when ethylene is passed into the $AlCl_3$ complex of aromatics until a mass

of crystals is formed, which is then worked up with water.

3.6. Ethylmethylbenzenes (Ethyltoluenes)

Properties. Some physical properties are collected in Table 15.

Production. Ethyltoluene is produced in similar fashion to ethylbenzene, i.e., by ethylation of toluene. Alkylation of toluene with ethylene in the presence of $AlCl_3$ and HCl gives all three isomers of ethyltoluene and more highly ethylated derivatives. A typical reaction mixture contains 48 % unreacted toluene, about 12 % 4-ethyltoluene, 19 % 3-ethyltoluene, 4 % 2-ethyltoluene, 14 % higher aromatics, and barely 1 % tar [135]. According to a Dow process, subsequent distillation gives a mixture of 3- and 4-ethyltoluenes [136]. 2-Ethyltoluene, toluene, and the more highly alkylated toluenes are then recycled. The gas-phase ethylation of toluene on unmodified HZSM-5 zeolite catalyst at 350 °C gives 23 % 4-ethyltoluene, 53 % 3-ethyltoluene, 8 % 2-ethyltoluene, and 10 % higher aromatics [137]. Formation of the 2- and 3-isomers in the ethylation of toluene can be almost completely suppressed by the use of a shape-selective ZSM-5 zeolite that has been modified by addition of phosphorus, boron, or other elements, and by suitable physical treatment. In a Mobil process, 4-ethyltoluene of 97 % purity is generated in this manner at 350–400 °C [137]. A patent granted to Universal Oil Products claims a process for the production of an alkylbenzene mixture, which also contains ethyltoluene, from toluene and synthesis gas [124].

Uses. The mixture of 3- and 4-ethyltoluene obtained by the Dow process is catalytically dehydrogenated to the corresponding mixture of methylstyrenes, known commercially as vinyltoluene [*25013-15-4*], and then polymerized to "poly(vinyltoluene)." Dehydrogenation of the practically pure 4-ethyltoluene from the Mobil process to 4-methylstyrene [*622-97-9*] occurs analogously on the route to "poly(*p*-methylstyrene)". A plant for the alkylation of toluene to 4-ethyltoluene by the Mobil process was opened in 1982 by Hoechst (now Hoechst-Celanese) in Baton Rouge, Louisiana [138]; the plant has an annual capacity of 16 000 t of 4-methylstyrene and has been purchased by Deltech in 1989. The output of the two products vinyltoluene and *p*-methylstyrene is estimated at 25 000 t/a [139] (→Polystyrene and Styrene Copolymers).

3.7. Cumene

Properties. Cumene [*98-82-8*], 2-phenylpropane, isopropylbenzene, C_9H_{12}, M_r120.19, is a colorless, mobile liquid with a characteristic, aromatic odor.

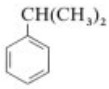

Cumene

The effect of cumene in increasing the octane rating of motor fuels (→Octane Enhancers) has long been known. Large plants for production of cumene, which was used as a component of aviation fuel, were constructed in the United States during the early 1940s. Some physical properties of cumene are listed below [140, 141]:

mp, °C	−96.04
bp, °C	152.39
d^{20}. g/cm^3	0.86179
n_D^{20}	1.49145
Specific heat (25 °C), J g^{-1} K^{-1}	1.96
Heat of vaporization, kJ/mol	
at 25 °C	44.23
at 152.4 °C	37.74
T_{crit}, K	631
p_{crit}, MPa	3.21
V_{crit}, cm^3/mol	428
Flash point, °C	43
Ignition temperature, °C	425

The variation of vapor pressure with temperature is given below:

t, °C	2.9	38.3	66.1	88.1	129.2
Vapor pressure, kPa	0.13	1.33	5.33	15.33	53.33

For more thermodynamic data, see [142]. Explosion limits for the system cumene vapor – oxygen – nitrogen at standard pressure and at 0.56 MPa are given in [143]. For cumene – air mixtures the explosion limits are 0.8 – 6.0 vol % cumene, corresponding to a cumene

concentration of 40 – 300 g/m^3 [144]. Cumene forms azeotropes with water, aliphatic carboxylic acids, ethylene glycol ethers, cyclohexanol, and cyclohexanone.

Production. Cumene is produced from benzene [*71-43-2*] and propene [*115-07-1*] in the presence of an acidic catalyst:

$$C_6H_6 + C_3H_6 \longrightarrow C_6H_5CH(CH_3)_2 \quad \Delta H = -113\ \text{kJ/mol}$$

Two processes are employed industrially (see also→ Acylation and Alkylation). In the *UOP process* a fixed bed consisting of polyphosphoric acid on a silica support is used as the catalyst [145]. The reaction takes place mainly in the gas phase at ≥200 °C and ≥2 MPa. Pure propene or propene with a propane content ≥5 % is used as raw material, together with benzene (min. purity 99.5 %) in up to 12-fold molar excess; this leads to high distillation costs in the product work-up. The yield of cumene is 97 % based on benzene and 90 % based on propene [146]. The losses are caused by overalkylation and oligomerization of the propene. The high-boiling residue contains 50 – 60 % diisopropylbenzenes (*o*- : *m*- : *p*- ratio 1 : 2 : 2 – 3) and numerous other byproducts. It is not possible to recycle these products to the alkylation reactor, because polyphosphoric acid has no transalkylation properties.

The *second industrial process* works in the liquid phase at 50 – 80 °C; aluminum chloride or the liquid double compound formed from aluminum chloride, (isopropyl-)benzene, and hydrogen chloride is employed as the catalyst. This complex is separated from the product stream, as a heavy phase and is returned to the alkylation reactor. Except for a small purge stream higher alkylated byproducts are recycled to the alkylation step, whereby essentially complete use of both benzene and propene bound in di- and triisopropylbenzenes is achieved by transalkylation. The yield of cumene, referred to benzene and propene, is 99 % and 98 %, respectively. The UOP process is used predominantly. Higher selectivity and the lower energy requirements, however, are now making the aluminum chloride process more attractive [146].

Uses. The production capacity for cumene in Western Europe, the United States, and Japan amounts to 2×10^6, 2×10^6, and 4×10^5 t/a, respectively. Cumene is used almost exclusively for the production of phenol [*108-95-2*] and acetone [*67-64-1*] (→Acetone, Chap. 4.1.; →Phenol). This process is based on the discovery in 1944 that cumene hydroperoxide [*80-15-9*], obtained by oxidation of cumene with oxygen, can be cleaved under the catalytic action of strong acids:

$$C_6H_5CH(CH_3)_2 + O_2 \longrightarrow C_6H_5C(CH_3)_2OOH$$

$$C_6H_5C(CH_3)_2OOH \longrightarrow C_6H_5OH + CH_3COCH_3$$

Both reactions are exothermic; the enthalpy of reaction amounts to 121 and 255 kJ/mol, respectively. (For cumene quality specifications, see Acetone, Chap. 4.)

Pure α-methylstyrene can be obtained as a byproduct of the phenol process. Cumene hydroperoxide is used as a radical initiator for the copolymerization of styrene with butadiene and acrylates, and also for the radical cross-linking of unsaturated polyester resins. Addition of cumene hydroperoxide to α-methylstyrene gives dicumyl peroxide, used for radical cross-linking of polyolefins. Hydrogenation of the aromatic nucleus in cumene gives isopropylcyclohexane (hydrocumene), a cycloaliphatic solvent with a high boiling point (154.5 °C) and a low freezing point (−90 °C).

3.8. Diisopropylbenzenes

1,4-Diisopropylbenzene [*100-18-5*] *p*-diisopropylbenzene, *p*-DIPB, $C_{12}H_{18}$, M_r162.26, 1,3-diisopropylbenzene [*99-62-7*], *m*-diisopropylbenzene, *m*-DIPB, and mixtures of these isomers, are industrially important. Some physical properties of diisopropylbenzenes are given in Table 16.

Table 16. Physical properties of diisopropylbenzene isomers [140]

Property	*m*-DIPB	*o*-DIPB	*p*-DIPB
CAS registry number	[*99-62-7*]	[*577-55-9*]	[*100-18-5*]
mp, °C	−63.13	−56.68	−17.07
bp, °C	203.18	203.75	210.35
d^{20}, g/cm^3	0.8559	0.8779	0.8568
η_D^{20}	1.4883	1.4960	1.4893

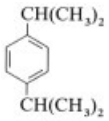

p-DIPB

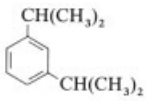

m-DIPB

The flash point of DIPB mixtures is 77 °C; the ignition temperature is 449 °C. Inhalation of DIPB vapor can lead to the appearance of narcotic effects.

Production. Diisopropylbenzenes are present in the high boiling fractions, in which they may account for 50 – 70 wt %, of cumene plants, together with numerous other byproducts. Processes for the production of *m*- and *p*-DIPB from these fractions have been suggested [147, 148]. The high-boiling fraction of the aluminum chloride process frequently contains 1,1,3-trimethylindan [*2613-76-5*], *bp* 204.8 – 204.9 °C, which cannot be separated from the *m*- and *p*-DIPB at a justifiable cost.

For production from benzene and propene alkylation catalysts that possess high isomerization and transalkylation activities are employed. Examples include aluminum chloride – hydrogen chloride [149, 150], silica – alumina [151, 152], boron trifluoride-treated alumina [153], and zeolithes [154]. The process is controlled in such a way that cumene formed at a later stage is alkylated with propene. The desired DIPB isomer is isolated by distillation and the undesired DIPB isomer, together with triisopropylbenzenes, are converted to cumene by reaction with benzene in a transalkylation step. The aim in both the alkylation and the transalkylation steps is to reach thermodynamic equilibrium, at which the *o*-diisopropylbenzene is present at only ca. 0.1 %.

Uses. 1,3-Diisopropyl- and 1,4-diisopropylbenzene are the starting materials for new synthetic routes to resorcinol (→Resorcinol) and hydroquinone (→Hydroquinone), respectively [155], which are used industrially in Japan and the United States [152]. Furthermore, a process for the production of phloroglucin, 1,3,5-trihydroxybenzene, from 1,3,5-triisopropylbenzene has been developed [156]. The monohydroperoxides of 1,3- and 1,4-diisopropylbenzene are used as initiators for the thermal cross-linking of unsaturated polyester resins, and for the emulsion polymerization of dienes, styrene, and acrylates. Oxidation gives *a,a,a′,a′*-tetramethylphenylenebiscarbinols. Dehydrogenation gives diisopropenylbenzenes, which by reaction with phenol or aniline give bisphenols or bisanilines, respectively, which are used as building blocks for polycarbonates, polyamides, polyesteramides, and polyurethanes (→Polyamides; →Polycarbonates; →Polyurethanes). Addition of isocyanic acid to diisopropenylbenzenes yields diisocyanates, from which light-stable polyurethanes can be produced [157]. Bis(*tert*-butylperoxyisopropyl)benzene, preferably prepared from the mixture of the corresponding biscarbinols of 1,3- and 1,4-DIPB, is used as a cross-linking agent for polyolefins.

3.9. Cymenes; C_4- and C_5-Alkylaromatic Compounds

The compounds listed in Table 17 are industrially important. Mixtures of 3-isopropyltoluene (*m*-cymene) and 4-isopropyltoluene (*p*-cymene) can be obtained by propylation of toluene in the presence of an isomerization-active Friedel – Crafts catalyst. These two isomers can be separated by distillation only with difficulty. The mixture is used for the production of *m*- and *p*-cresol, (→Cresols and Xylenols, Chap. 1.4.3.), [158]. Treatment of (methyl-)aromatic compounds with isobutene yields *tert*-butyl-substituted compounds, which are used as solvents (*tert*-butylbenzene) and for the production of fragrances, pharmaceuticals, and herbicides.

Table 17. Physical properties of some alkylbenzenes

Compound	CAS registry number	*mp*, °C	*bp*, °C	η_D^{20}
3-Isopropyltoluene	[*535-77-3*]	−63.75	175.05	1.4929
4-Isopropyltoluene	[*99-87-6*]	−67.94	177.10	1.4909
2-Isopropyltoluene	[*527-84-4*]	−71.54	178.15	1.5005
Isobutylbenzene	[*538-93-2*]	−51.48	172.76	1.4864
tert-Butylbenzene	[*98-06-6*]	−57.85	169.12	1.4926
4-*tert*-Butyltoluene	[*98-51-1*]	−53	192.7	1.4919
5-*tert*-Butyl-*m*-xylene	[*98-19-1*]	−18.3	209	1.4956
tert-Pentylbenzene	[*2049-95-8*]		192.38	1.4958

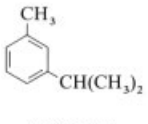

m-Cymene

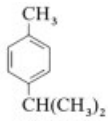

p-Cymene

The introduction of a *tert*-pentyl group can be achieved by the use of *tert*-pentyl chloride as the alkylating agent and aluminum chloride, iron(III) chloride, or zirconium tetrachloride as the catalyst. 2-*tert*-pentylanthraquinone, obtained from *tert*-pentylbenzene and phthalic anhydride, is employed in the industrial production of hydrogen peroxide. Isobutylbenzene, required for the synthesis of pharmaceuticals and fragrances, is produced in industrial quantities by treatment of toluene with propene over catalysts containing alkali metals, such as potassium – graphite inclusion compounds or sodium – potassium on a support [159].

3.10. Monoalkylbenzenes with Alkyl Groups $>C_{10}$

Monoalkylbenzenes with secondary alkyl groups C_{10} to ca. C_{14} are used as starting materials for the production of alkylbenzenesulfonates (ABS) (→Surfactants). Only secondary alkylbenzenesulfonates with unbranched (linear) side chains (LABS) are of practical importance as raw materials for detergents, as they are extensively degraded (>90 %) by microorganisms in sewage plants after a relatively short period of time.

Before the introduction of LABS, which took place in the Federal Republic of Germany in 1962, alkylbenzenesulfonates with highly branched C_{12} side chains were used. These compounds, tetrapropylene benzenesulfonates, were produced from the products of alkylation of benzene with propene tetramer, itself obtained by the tetramerization of propene by contact with phosphoric acid. These sulfonates possess excellent detergent properties, but their biological degradeability is insufficient.

Production. Linear secondary alkylbenzenes are obtained industrially by reaction of benzene with secondary alkyl chlorides (chloroparaffins) or with olefins under the influence of a Friedel – Craft catalyst such as aluminum chloride, hydrofluoric acid, or sulfuric acid (→Acylation and Alkylation, Chap. 2.1., →Acylation and Alkylation, Chap. 2.1.1.). The following methods have been described for the production of the required alkyl chlorides and olefins [160]:

1) Dehydrogenation of $C_{10}-C_{13}$-*n*-alkanes, according to the Pacol process of UOP (see Section 2.1.2.1), on a fixed bed at 400 – 500 °C, with a conversion of 10 – 15 %; the ensuing mixture of internal olefins and unreacted alkanes is used for the alkylation of benzene. The *n*-$C_{10}-C_{13}$-alkanes used in the dehydrogenation are obtained from the kerosene fraction of crude oil by adsorption on molecular sieves, according to well known processes such as IsoSiv of Union Carbide, or Molex of UOP.
2) By chlorination of $C_{10}-C_{13}$-alkanes at ca. 100 °C to a conversion of ca. 30 %, and use of the resulting mixture of alkanes and secondary monochloroalkanes for alkylation of benzene.
3) By chlorination of $C_{10}-C_{13}$-*n*-paraffins as in 2) above, followed by dehydrochlorination (elimination of hydrogen chloride) of the monochloroalkanes at 200 – 300 °C over an iron catalyst to give linear alkenes with internal double bonds; these alkenes, in admixture with unreacted paraffins, are then used to alkylate benzene.
4) Metathesis of higher ($>C_{18}$) with lower ($<C_{10}$) internal olefins from the Shell SHOP process (see Section 2.1.2.2) to give olefins of intermediate molecular mass.

A $C_{10}-C_{13}$-α-olefin fraction suitable for use in detergents can also be obtained by thermal cleavage of $C_{20}-C_{40}$-*n*-paraffins at 500 – 600 °C, according to the older wax-cracking process of Shell (see Section 2.1.2.1).

The alkylation of benzene with chloroparaffin – paraffin mixtures takes place in the presence of aluminum chloride at ca. 50 °C. A relatively high proportion of 2-phenylalkanes is characteristic of secondary alkylbenzenes obtained in this manner [161, 162]. Dichloroalkanes, always present in the mixture of chloroparaffins even at low levels of chlorination, can lead to the formation of diphenylalkanes, 1,4-dialkyltetralins, and 1,3-dialkylindans. For formation and analytical detection see

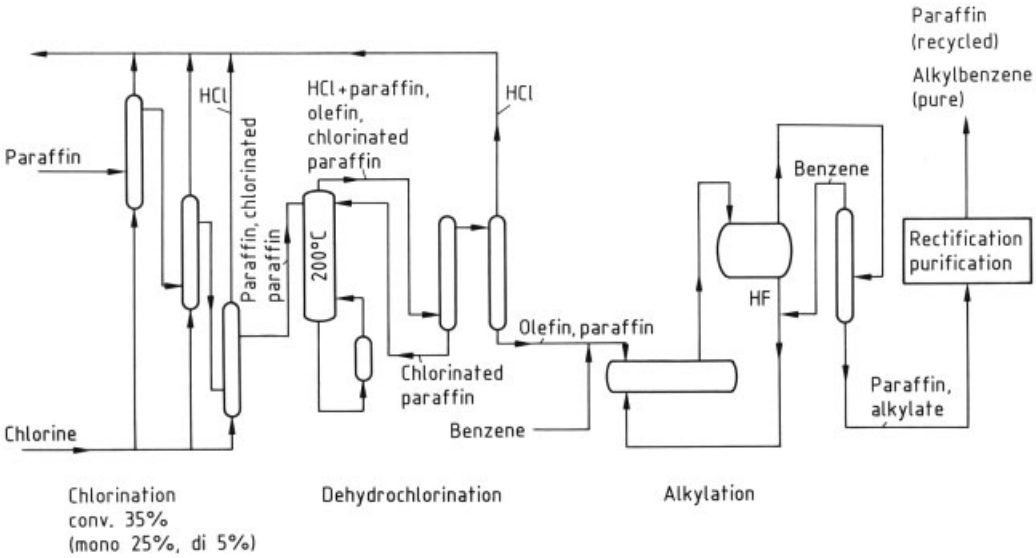

Figure 8. Production of alkylbenzenes, process of Hüls AG [165]

[163]. The removal of these compounds from monoalkylbenzenes is difficult. Formation of these byproducts can be largely avoided by carrying out the alkylation of benzene with olefins from the dehydrochlorination, where dichloroalkanes are dehydrochlorinated to give tar-like products that are easily separated. In industrial production the reaction of $C_{10}-C_{13}$-olefins with benezene is carried out with a 1 : 10 molar ratio in the presence of anhydrous hydrofluoric acid at 10 °C, a process that dates back to 1949 [164]. The industrial process using the steps paraffin chlorination – dehydrochlorination – alkylation in the presence of hydrofluoric acid was first realized by Chemische Werk Hüls in 1962 [162]. This process is shown in Figure 8. In this way a quality standard for alkylbenzenes for use in detergents was created; in ecological terms, this standard is still valid.

3.11. Diphenylmethane

Diphenylmethane [*101-81-5*] benzylbenzene, $C_{13}H_{12}$, M_r 168.24, was first prepared in 1871 by Zincke from benzyl chloride and benzene in the presence of zinc dust or copper (I) oxide.

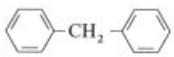

Physical Properties. Diphenylmethane forms colorless prismatic needles with a harsh herbaceous odor reminiscent of geranium leaves. It is insoluble in water, soluble in alcohol, ether, chloroform, and benzene.

mp	26 – 27 °C
bp	
101.3 kPa	261 – 262 °C
1.33 kPa	125.5 °C
d_4^{20}	1.006
Flash point ca.	130 °C
Enthalpy of formation (room temperature)	109.6 kJ/mol
Enthalpy of combustion (20 °C)	41.2×10^3 kJ/kg

Chemical Properties. As an araliphatic hydrocarbon, diphenylmethane displays the chemical properties of both aromatic and aliphatic compounds. It is difficult to limit substitution in the phenyl radicals to one of the rings; therefore, controlled syntheses of particular unsymmetrically substituted diphenylmethanes must be carried out by different paths. Under suitable reaction conditions the methylene group can also be substituted, e.g., by halogen or nitro groups. Benzophenone can be produced by oxidation, for example, with oxygen over various catalysts, chromic acid or dilute nitric acid. Hydrogenation in ethanol over a nickel catalyst leads to dicyclohexylmethane. When diphenylmethane is passed over platinum – charcoal at 300 °C, cyclization to fluorene takes place.

Production. Of the many known methods to synthesize diphenylmethane only a few are used commercially.

The Friedel – Craft reaction of benzyl chloride with benzene in the presence of aluminum chloride is an important method for production of diphenylmethane:

$$C_6H_5CH_2Cl + C_6H_6 \xrightarrow{AlCl_3} C_6H_5CH_2C_6H_5$$

Other known catalysts of this process are elemental aluminum, iron(III) chloride, antimony pentachloride, zinc chloride – hydrogen chloride and boron fluoride dihydrate. When sulfuric acid is used as a condensation agent diphenylmethane yields are 83 % [166] and with polyphosphoric acid yields are over 90 % [167].

Condensation of benzene with formaldehyde in presence of 85 % sulfuric acid yields 79 % diphenylmethane [168].

$$2\ C_6H_6 + HCHO \xrightarrow{H_2SO_4} C_6H_5CH_2C_6H_5$$

Quality Specifications. Commercial diphenylmethane should be at least 97 % pure (GC). It must be free of halogens and have a minimum melting point of 24 °C.

Uses. Diphenylmethane is used in the fragrance industry both as a fixative and in scenting soaps. It can serve as a synergist for pyrethrin in pesticides and insecticides [169, 170]. Diphenylmethane is recommended as a plasticizer to improve the dyeing properties [171], as a solvent for dyes [172], and as a dye carrier for printing with disperse dyes [173]. The addition of diphenylmethane to saturated, linear polyesters improves their thermal stability [174], and its addition to jet fuels increases their stability and lubricating properties [175].

Substituted diphenylmethanes are used as solvents for pressure-sensitive recording materials [176].

4. Biphenyls and Polyphenyls

In the series of aromatic hydrocarbons with polyphenyls or polyaryls of the general formula $C_6H_5(C_6H_4)_nC_6H_5$, only the lowest members ($n = 0$, 1) are economically important. Higher polyphenyls ($n > 2$) are also known and may be isolated from coal tar or from aryl halides by the Ullmann reaction.

4.1. Biphenyl

Biphenyl [*92-52-4*], phenylbenzene, diphenyl, $C_{12}H_{10}$, M_r154.2, was discovered in 1862 by Fittig by the reaction of bromobenzene with sodium metal. Berthelot obtained the compound in 1867 by passing benzene vapor through a heated tube. In 1875 Büchner demonstrated the presence of biphenyl in coal tar [177].

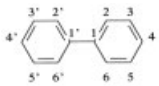

Physical Properties. Pure biphenyl is a white solid; when only slightly impure it has a yellow tint. It crystallizes from solution results as glossy plates or monoclinic prisms. Biphenyl is almost insoluble in water, but is easily soluble in organic solvents such as ethanol, diethyl ether, and benzene. The dipole moment and X-ray measurements demonstrate that the two benzene rings are coplanar in the solid state. In the melt, in solution, and in the vapor free rotation about the central C-C bond can take place; in substituted biphenyls, however, this rotation may be severely restricted, e.g., in dinitrodiphenic acid, leading to the existence of optically active isomers. The most important physical properties are listed below:

mp	69.2 °C [178]
bp (101.3 kPa)	255.2 °C [178]
d_4^{20} (solid)	1.041
d_4^{77} (liquid)	0.9896 [179]
n_D^{77}	1.5873
T_{crit}	515.7 °C [180]
p_{crit}	4.05 mPa
Flash point	113 °C [181]
Ignition temperature	570 °C
Heat of combustion	6243.2 kJ/mol
Heat of vaporization	53.9 kJ/mol
Heat of fusion	18.60 kJ/mol

The variation of some physical properties with temperature is given in Table 18 (see next page).

Table 18. Temperature dependence of some physical properties of biphenyl [179]

t, °C	100	200	300	350
Vapor pressure, kPa		25.43	246.77	558.06
Liquid density, g/cm^3	0.97	0.889	0.801	0.751
Heat capacity, J/g	1.786	2.129	2.468	2.640

Chemical Properties. Biphenyl is a very stable organic compound; in an inert atmosphere it remains unchanged even at a temperature >300 °C. The compound sublimes, distills with steam, and undergoes various substitution reactions; the reactivity is similar to, but less than that of benzene. Substitution reactions, e.g., chlorination, nitration, and sulfonation, occur at the 2- and 4-positions, the latter being the preferred site of attack in Friedel – Crafts reactions.

Production. *Hydrodealkylation of Benzene.* Biphenyl is currently obtained mainly as a byproduct in the production of benzene by thermal or catalytic hydrodealkylation of toluene (→Benzene, Chap. 5.3.1.) [182–185]. The reaction is carried out in a hot tube reactor at 700 °C with a hydrogen pressure of 4 MPa; the ratio of hydrogen to toluene is 4 : 1. Toluene conversion ranges from 35 to 85 %, with methane being formed at 2.5 mol % above the stoichiometric.

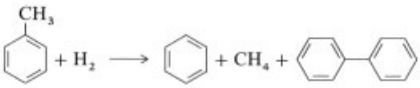

Biphenyl is obtained as the pot residue, following removal of gas and distillation of benzene and toluene; about 1 kg of biphenyl is obtained per 100 kg benzene. The biphenyl can be enriched to a purity of 93 – 97 % by distillation.

Thermal Dehydrocondensation of Benzene. Until the early 1970s, biphenyl was produced exclusively by specific thermal dehydrocondensation of benzene [186]. This process yields a higher quality product than the toluene route, e.g., distilled material with a purity >99.5 %; varying quantities of terphenyl isomers and several higher homologues are also obtained.

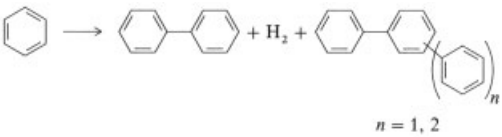

The slightly endothermic dehydrocondensation ($\Delta H = +8$ kJ/mol [187]) of benzene to biphenyl and terphenyl has been investigated intensively [188, 189]. The formation of biphenyl from benzene starts at ca. 480 °C and increases with increasing temperature and residence time.

Industrially, the reaction is carried out in a multitube reactor, heated electrically [190] or by heat exchanger, at 700 – 850 °C using residence times of the order of seconds. An optimal tube length of 45 – 75 m and diameter of 100 – 120 mm have been reported [191–194]. None of the published procedures work in equilibrium because carbon black formation increases with increasing temperature and longer residence time.

The preferred residence time is ca. 10 s, which results in a benzene conversion of ca. 10 % [192, 195]. By recycling the benzene, 1 kg of starting material may yield 0.8 – 0.85 kg biphenyl, 0.07 – 0.12 kg terphenyl, and some higher polyphenyls. The major technical problem in the production of biphenyl by dehydrocondensation of benzene is carbonization at the high temperature required [196]. The carbon deposit on the reactor tube walls and the release of large carbon flakes at bends or narrow passages within the tubes may lead to blockages. Countermeasures include addition of about 0.1 % oxygen or sulfur compounds, such as methanol, ethanol, acetone, or carbon disulfide [197–199] to the benzene feedstock, the use of special alloys [200, 201], uniform heating of the reactor by means of circulating gases, and high feedstock velocity, and the use of turbulent flow [192]. A proposal for further improving the process is to carry out the reaction in fluidized beds. Quartz sand [202] or sintered corundum [203] is used as energy carrier in this process; these materials are recycled after the carbon deposits have been removed by burning them outside of the reactor.

Other Synthetic Routes. In addition to these two industrial processes for the production of biphenyl, other synthetic routes have been described:

1) Ullmann reaction: heating halogenobenzenes with copper powder [204].
2) Pyrolysis of coal tar in the presence of molybdenum oxide – alumina catalyst [205, 206].
3) Reduction of phenyldiazonium chloride solution with copper and zinc [207].

4) Heating phenylmagnesium halides with copper(II) salts [208].
5) Oxidative dehydrocondensation of benzene [209–214]. Benzene is oxidized to biphenyl by platinum- or palladium salts in glacial acetic acid at 120 °C; the salts are reduced to the pure metals, and stoichiometric quantities of the salts are required.
6) Dehydrocondensation of benzene in the presence of catalytic quantities of noble metal salts requires reoxidation of the metals during the course of the reaction. This may be done with pure oxygen, but a partial pressure of 6 MPa is required [215]. Additives such as β-diketones [216], heteropolyacids [217], or organic copper salts [218] are incapable of significantly lowering the required oxygen pressure, so that a potentially explosive regime cannot be avoided [217]. Yields of biphenyl are ca. 260 mol %, calculated on the palladium salt charge.
7) Hydrodimerization of benzene to phenylcyclohexane, followed by dehydrogenation to biphenyl [219–223]:

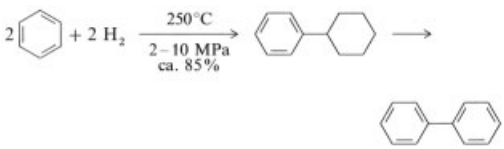

Suitable catalysts for this dimerization are noble metals on molecular sieve, alkali metals on alumina or carbon, and transition metals such as nickel, tungsten, or zinc on the zeolites faujasite or mordenite. A selectivity of up to 87 % has been claimed. Subsequent dehydrogenation is carried out at about 400 °C over $Pt-Al_2O_3$ [223], $Cr_2O_3-Na_2O-Al_2O_3$ [222, 224], $Cr_2O_3-K_2CO_3-Fe_2O_3$ [225], or $Cr_2O_3-MgO-Al_2O_3$ [226] giving yields of up to 99 %. Data concerning residence times over the catalyst are not available, either for the hydrodimerization or the dehydrogenation processes.

Analysis. Gas chromatography is used to analyze biphenyl and its derivatives.

Storage and Transportation. Biphenyl is marketed as a solid in the form of plates, flakes, or pellets, containing <10 ppm benzene and with a biphenyl content of >99.5 %; it is transported in bags or fiber drums; in the molten state (ca. 120 °C) in tank cars. It is not a regulated material for transportation, on account of its toxicity; there are no restrictions for transportation by sea, land, or air. The greatest hazards in handling biphenyl are the risk of dust explosions and the ignition of biphenyl vapor – air mixtures over the molten product.

Uses. Biphenyl is an important heat transfer agent, because of its high thermal stability. It is used as a composite with diphenyl ether, especially as a melting point depressant. The eutectic mixture of 26.5 % biphenyl and 73.5 % diphenyl ether is marketed under the trade names Diphyl (Bayer), Dowtherm A (Dow), Thermic (ICI), Gilotherm (Rhône-Poulenc), Therm S 300 (Nippon Steel), Santotherm VP and Therminol VP-1 (both Monsanto) for use at up to 400 °C. The eutectic begins to boil at 256 °C; the vapor pressure at the maximum operating temperature (400 °C) is ca. 1.1 MPa.

A further large quantity of biphenyl is used as a dye carrier [227, 228]. Until a few years ago, polychlorinated biphenyls (PCBs), generated by chlorination of biphenyl, were produced in large quantities for use as nonflammable hydraulic fluids and as transformer dielectrics. Production has largely ceased on ecological considerations (→Chlorinated Hydrocarbons). Among the hydroxy derivatives, *o*- and *p*-hydroxybiphenyl and *p,p′*-dihydroxybiphenyl are used industrially. *o*-Hydroxybiphenyl is used as a preservative and fungicide. It is synthesized mainly from cyclohexanone (→Phenol Derivatives). *p*-Hydroxybiphenyl and *p,p′*-dihydroxybiphenyl are generated by sulfonation of biphenyl, followed by fusion with alkali.

Economic Aspects. According to a 1976 estimate, 67 % of the annual United States production of 40 000 t biphenyl was produced by dealkylation of toluene and 33 % by thermal dehydrocondensation of benzene [229].

The annual production for 1984 is estimated at 15 000 t, with an increasing proportion from dealkylation of toluene. The production capacities of the major producers, Monsanto, (United States), Bayer (Federal Republic of Germany), and Monsanto (United Kingdom) have not been

Table 19. Physical properties of terphenyls

Property	*o*-Terphenyl [*84-15-1*]	*m*-Terphenyl [*92-06-8*]	*p*-Terphenyl [*92-94-4*]	Reference
mp, °C	59	87	212	[233]
bp (101.3 kPa), °C	332	365	376	[234]
Flash point, °C	171	206	210	[236]
T_{crit}, °C	613	603	621	[237]
P_{crit}, MPa	3.903	3.503	3.330	[238]
Heat of vaporization (at *bp*), kJ/mol	58.3	64.2	62.6	[234]
Vapor pressure, kPa				
93 °C	0.01172	0.00165		[233, 239]
204 °C	2.834	0.8274	0.789	
315.6 °C	64.4	27.3	24.56	
426.7 °C	439.9	240.6	174.4	

disclosed. Based on the legal restrictions imposed on the use of PCBs, capacity should far exceed current demand.

4.2. Terphenyls

Physical Properties. The three isomeric terphenyls, $C_{18}H_{14}$, M_r230.29, are colorless to pale yellow crystalline solids [230–235]. Some physical properties are listed in Table 19.

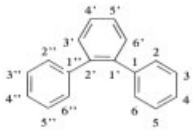

o-Terphenyl
1,1′:2′, 1″-Terphenyl

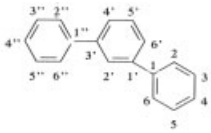

m-Terphenyl

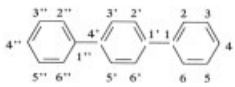

p-Terphenyl

Table 20. Terphenyl substitution pattern

	Position for substitution			
	First	Second	Third	Fourth
o-Terphenyl	4-	4″-	4′-	5′-
m-Terphenyl	4′-	4-	4″-	
p-Terphenyl	4-	4″-	2′-	

Chemical Properties. Terphenyls, like biphenyls, are thermally extremely stable organic compounds, again offering a potential for use as heat transfer media. Terphenyls are sublimable, steam volatile, and show the typical substitution reactions of aromatic hydrocarbons, such as bromination and nitration. The substitution pattern for the preferred positions of the three terphenyls is shown in Table 20.

Production. Terphenyls are byproducts in the production of biphenyl by dehydrocondensation of benzene; they are found in the high-boiling fraction of the pyrolysis products. The pot residue has the following approximate composition [240]:

3 – 8 %*o*-Terphenyl
44 %*m* – Terphenyl
24 %*p* – Terphenyl
1.5 %Triphenylene
22 – 27 %Higher polyphenylenes and tar.

When, in the early 1960s, terphenyl was proposed to be used as coolant and moderator in nuclear reactors [241], numerous reports were published of attempts to enhance the yield of terphenyl in dehydrocondensation reactions. Increasing the temperature or residence time was ruled out because of increased carbonization. These procedures also enhanced the biphenyl fraction, so that the latter had to be recycled to the reactor [193, 197, 242–244]. Recycling a concentration of up to 30 % biphenyl has been described [193]. A practical level appears to be using benzene with 10 % biphenyl [210]. Attempts have also been made to recycle the extremely high melting *p*-terphenyl, after isolation from the reaction mixture, to obtain a lower melting terphenyl mixture [242]. The *o*-derivative is readily separated from the terphenyl isomer mixture by distillation; *m*- and *p*-terphenyl distill together and the pure isomers can be ob-

tained by zone refining [245]. Terphenyl isomers, including hydrogenated terphenyl, can be analyzed by gas chromatography. The sole producer of terphenyl in the United States is Monsanto. The market amounts to several thousand tons per year. Terphenyl is available in Europe from Bayer.

Storage and Transportation. Solid terphenyl is shipped as flakes in laminated bags or in fiber drums, liquid, hydrogenated terphenyl in tank cars, steel drums, or barrels. Terphenyls and hydrogenated terphenyls possess a low toxicity, and are therefore not regarded as dangerous goods for the purpose of transportation.

Uses. Terphenyl is used as a heat transfer agent because of its excellent thermal stability; a mixture with a composition of 2 – 10 % *o*-terphenyl, 45 – 49 % *m*-terphenyl, 25 – 35 % *p*-terphenyl, and 2 – 18 % higher polyphenyls is used. This composition approximates the pot residue from the distillation of biphenyl. This mixture has *bp* ca. 360 °C; a disadvantage is the high *mp*, 145 °C. The largest fraction is partially hydrogenated to yield a clear oil, miscible with hydrocarbons and chlorinated hydrocarbons. Hydrogenation lowers the setting point and the viscosity but unfortunately also reduces the thermal stability. Partially hydrogenated terphenyl [246] is used mainly as a dye carrier for pressure sensitive recording materials and copy paper, and as a heat transfer oil. Heat transfer oils consisting of partially hydrogenated terphenyl isomers are marketed under the following trade names: Santotherm 66 or 88, Therminol (Monsanto, United States), Gilotherm (Rhône-Poulenc, France), and Therm S 900 (Nippon Steel, Japan). Their usage range extends to 340 °C.

4.3. Polyphenyls.

Polyphenyls with four or more phenyl residues have generally been synthesized as *p*-linked isomers and are used as scintillators [247]. They are economically unimportant.

A symmetrical hydrocarbon within this class is 1,3,5-triphenylbenzene, $C_{24}H_{18}$, *mp* 174 °C, prepared by condensation of three molecules of acetophenone.

5. Hydrocarbons from Coal Tar

Anthracene and Naphthalene and Hydronaphthalenes are separate keywords.

Hydrocarbons from coal tar are generally recovered in large-scale operations by distillation and crystallization, after the polar, co-boiling components, phenols and nitrogen bases, have been separated by extraction (→Tar and Pitch). Supplies of coal tar as a raw material have been sufficient to date to meet the demand for these hydrocarbons, so that no other production processes, such as syntheses, have yet been developed to the stage of commerciality. Hydrocarbons from coal tar are used predominantly in the synthesis of dyes, plastics, and pharmaceuticals.

5.1. Acenaphthene

Acenaphthene [*83-32-9*], $C_{12}H_{10}$, M_r 154.21, was discovered in coal tar in 1867 by BERTHELOT.

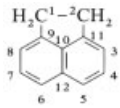

Physical Properties. Acenaphthene forms colorless needles; readily soluble in chloroform, toluene, and hot ethanol; soluble in cold ethanol, hot diethyl ether, and benzene; insoluble in water.

mp	95.4 °C
bp (101.3 kPa)	277.5 °C
ϱ (25 °C)	1.2195 g/cm^3
n_D^{100}	1.6048
T_{crit}	530 °C
Heat of fusion	1.344×10^5 J/kg
Heat of vaporization (40 °C)	3.48×10^5 J/kg
Heat of combustion (25 °C)	4.033×10^7 J/kg

Chemical Properties. Acenaphthene reacts with halogens, preferentially at 3-, 5-, and 6-positions, or at the 1-position when irradiated with visible light. Nitration and sulfonation occur at the 3-, 5-, or 6-positions. Catalytic hydrogenation leads to tetra- and decahydroacenaphthene. Decahydroacenaphthene can

be isomerized in the presence of Lewis acids to 1,3-dimethyladamantane. Oxidation of acenaphthene gives naphthalic acid or its anhydride, acenaphthenequinone, acenaphthenol, and acenaphthenone.

Production. High-temperature coal tar contains, on average, 0.3 % of acenaphthene; in addition, it is formed under the conditions of coal tar distillation by hydrogenation of acenaphthylene, which is found in crude tar at a level of 2 %. Acenaphthene is concentrated to ca. 25 % in the coal tar fraction boiling between 230 and 290 °C (wash oil), which is recovered at ca. 7 % in continuous tar distillation. The readily crystallizable acenaphthene fraction, boiling at 270 – 275 °C, is obtained from the wash oil by redistillation, or by removing it directly during primary tar distillation. From this, technical acenaphthene (95 – 99 % pure) is produced by crystallization in agitated coolers and centrifuges, or by continuous counter-current crystallization [256, 259]. The pure compound is obtained by further distillation and recrystallization.

Uses. Acenaphthene is used on a large-scale for synthesis of naphthalic anhydride by gas-phase with air oxidation at 300 – 400 °C, in the presence of vanadium containing catalysts [258, 259]. Liquid-phase oxidation with chromate or air in the presence of cobalt- or manganese acetate at 200 °C [260], or cobalt resinate at 120 °C [261] produces naphthalic acid (→Carboxylic Acids, Aromatic, Chap. 4.4.). Production of naphthalene-1,4,5,8-tetracarboxylic acid from acenaphthene by condensation with malonic acid dinitrile [262] is at present not industrially important. A technically feasible procedure has been developed using 5-methyl-*peri*-acenaphthindene-7-one [*42937-13-3*] (by reaction of acenaphthene with diketene in HF) (→Carboxylic Acids, Aromatic, Chap. 4.7.), [263]. Naphthalic anhydride and naphthalene-1,4,5,8-tetracarboxylic acid are used as intermediates for production of perylene and perinone pigments (→Pigments, Organic) [262]. Other valuable intermediates, particularly for the production of dyes and optical brighteners, and also for pharmaceuticals and pesticides, are produced from, e.g., 4-bromacenaphthene [264] (for 4-bromonaphthalic anhydride or naphthalic anhydride-4-sulfonic acid), from 5,6-dichloracenaphthene (for 4,5-dichloronaphthalic anhydride) [265] and from 3,5,6-trichloracenaphthene [266]. Nitration and oxidation of acenaphthene produce 4-nitronaphthalic anhydride, from which the fluorescent pigment Solvent Yellow 44 can be made [267]. Reaction of acenaphthalene with sodium cyanate in hydrofluoric acid gives acenaphthene-5,6-dicarboxylic acid imide, which can be used as an intermediate for dyes [268].

Bromination of acenaphthene in the presence of, e.g., iron(III)- or aluminum chloride, followed by further bromination by the addition of a radical initiator, such as azobisisobutyronitrile (AIBN), and subsequent dehydrobromination produces condensed bromacenaphthene, which is particularly well-suited for rendering plastics nonflammable and radiation-resistant [269]. Thermally stable plastics can be obtained from acenaphthene in the following ways: 1) by condensation with formaldehyde; 2) oxidation and reaction with an aromatic polyamine [270]; by cocondensation with phenol and formaldehyde [271]; and by condensation with terephthaloyl chloride [272]. Acenaphthene also serves as a feedstock for the production of acenaphthenequinone, acenaphthene, and dimethyladamantane [273].

Derivatives.

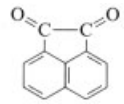

Acenaphthenequinone

Acenaphthylene

Acenaphthenequinone [*82-86-0*], $C_{12}H_6O_2$, M_r 182.17; *mp* 261 °C; yellow needles; slightly soluble in ethanol; soluble in hot benzene and toluene. It is formed by liquid-phase oxidation of acenaphthene with hydrogen peroxide or dichromate in acetic acid, by chlorination of dibromacenaphthylene followed by acid hydrolysis, or by condensation of acenaphthene with ethyl nitrite and separation of the dioxime. Acenaphthenequinone is suitable for use as an insecticide and fungicide and can be converted into vat dyes.

Acenaphthylene [*208-96-8*], $C_{12}H_8$, M_r 152.18; *mp* 92 – 93 °C; yellow prisms; soluble in ethanol, diethyl ether, and benzene; insoluble in water. It is found in coal tar at a concentra-

tion of 2 % and can be produced industrially by catalytic dehydrogenation of acenaphthene in the gas phase. Thermal, ionic, radical, or radiochemically induced polymerization produces polyacenaphthylene [274]. Polyacenaphthylene and copolymers with other olefins are characterized by high thermal and mechanical stability together with good electrical insulation properties. The thermal stability of polystyrene can be significantly increased by copolymerisation with acenapthylene [274]. Polymerisation of acenaphthylene with acetylene in the presence of a Lewis acid catalyst gives electrically conductive polymers; they are used for electronic engineering and for the antistatic finishing of plastics [275]. Thermosetting resins with good resistance to heat and chemicals can be obtained by cocondensation with phenol and formaldehyde [271]. Acenaphthylene possesses excellent properties as an antioxidant in cross-linked polyethylene and ethylene – propylene rubber [274]. Thermal trimerization of acenaphthylene leads to decacyclene, which can be further processed to sulfur dyes [276].

5.2. Indene

Indene [*95-13-6*], C_9H_8, M_r 116.16, was discovered in coal tar in 1890 by KRAEMER and SPILKER.

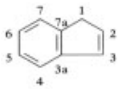

Physical Properties. Indene is a colorless liquid; insoluble in water; soluble in ethanol; it is miscible with diethyl ether, naphtha, pyridine, carbon tetrachloride, chloroform, and carbon disulfide.

mp	−1.5 °C
bp (101.3 kPa)	183 °C
ϱ (20 °C)	0.9966 g/cm^3
n_D^{20}	1.5763
Heat capacity (25 °C)	1.610×10^3 J/kg
Heat of fusion	8.33×10^4 J/kg
Heat of combustion (25 °C)	4.132×10^7 J/kg
Dipole moment (in CCl_4)	0.85 D

Chemical Properties. Indene polymerizes readily at ambient temperature and in the dark. Polymerization is accelerated by heat, acid, and Friedel – Crafts catalysts. Depending on reaction conditions, oxidation leads to dihydroxyindan, homophthalic acid, or phthalic acid. Hydrogenation produces indan. Halogenation occurs at the 2,3-double bond. Diene adducts are formed with maleic anhydride and hexachlorocyclopentadiene.

Production. Indene is found in high-temperature coal tar at an average concentration of 1 %. Pyrolysis residual oils from olefin production contain varying amounts of indene [277], which can be separated by extractive distillation with *N*-methyl-2-pyrrolidone [278] or by crystallization [279]. Indene is recovered industrially from dephenolated and debased tar light oil by rectifying distillation followed by crystallization. When the phenol extraction is omitted a highly concentrated indene fraction can also be obtained by azeotropic distillation of an indene – phenol fraction with water, whereby the phenol is separated as a bottom product [280].

Uses. Indene is used as a comonomer in coal-tar derived indene – coumarone and petroleum-derived aromatic hydrocarbon resins (→Resins, Synthetic). Technical pure indene is used for the production of indan and its derivatives. Esters of indene-1-carboxylic acid, which are effective as acaricides, can be synthesized from indene for use in, e.g., tick collars [281].

Derivatives.

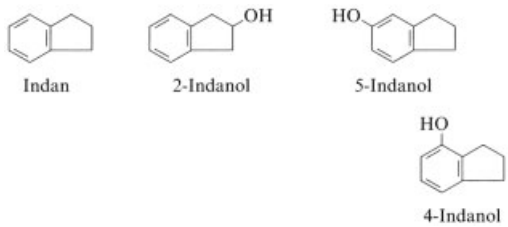

Indan [*496-11-7*], hydrindene, C_9H_{10}, M_r 118.18;

mp	−51.4 °C
bp (1013 kPa)	176 °C
ϱ (20 °C)	0.9644 g/cm^3

Indan is a colorless liquid; insoluble in water; soluble in ethanol and diethyl ether. It is found at a concentration of 0.1 % in coal tar, but is mainly produced by catalytic hydrogenation of indene.

Alkylated indans are used in the production of synthetic lubricants (→Lubricants and Related Products). Indan is used for the synthesis of indanols (→Phenol Derivatives).

2-Indanol [*4254-29-9*], $C_9H_{10}O$, M_r 134.18; *mp* 69 °C; colorless crystals. It is obtained by hydrogenation of 2-indanone prepared by reaction of indene with hydrogen peroxide and boiling in sulfuric acid over a platinum catalyst in ethanol. 2-indanol is used in the synthesis of the coronary therapeutic, Aprindine [*37640-71-4*], *N*-(3-diethylaminopropyl)-*N*-indan-2-ylaniline; (→Antiarrhythmic Drugs), [282] and can be employed as an intermediate for other pharmaceuticals, e.g., for antihypertonics [283].

5-Indanol [*1470-94-6*], $C_9H_{10}O$, M_r 134.18; *mp* 55 °C; colorless crystals. It is produced by fusion of indan-5-sulfonic acid with potassium hydroxide [284]. 5-Indanol is used in the synthesis of the multi-purpose antibiotic, Carindacillin [*26605-69-6*] (6-[2-phenyl-2-(5-indanyloxycarbonyl)acetamido] penicillanic acid; →Antibiotics), [285]. 5-Indanylesters of substituted picolinic acids can be used as antihypertonic [286]. 5-Indanol itself may be used as an antidandruff additive in hair shampoos (→Hair Preparations) [287].

4-Indanol [*1641-41-4*], $C_9H_{10}O$, M_r 134.18; *mp* 49 – 50 °C; colorless crystals. Analogous to 5-indanol it is synthesized by potassium hydroxide fusion of indan-4-sulfonic acid, which is produced as a byproduct in indan sulfonation. Catalytic gas-phase amination of 4-indanol produces 4-aminoindan [288], which is required as an intermediate, e.g., for pharmaceuticals. Clorindanol (7-chloro-4-indanol) is used as an antiseptic.

5.3. Fluorene

Fluorene [*244-36-0*], $C_{13}H_{10}$, M_r166.22, was discovered in coal tar in 1867 by BERTHELOT.

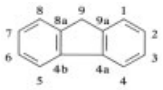

Physical Properties. Fluorene forms colorless flakes with slight violet fluorescence; sublimes readily; insoluble in water; soluble in benzene, diethyl ether, carbon disulfide, and hot ethanol; soluble with difficulty in cold ethanol.

mp	115 °C
bp (101.3 kPa)	298 °C
ϱ (20 °C)	1.181 g/cm^3
Heat of fusion	1.21×10^5 J/kg
Heat of combustion (25 °C)	3.994×10^7 J/kg

Chemical Properties. Fluorene is capable of numerous chemical reactions, through both the aromatic rings and especially the reactive methylene group. Oxidation produces fluorenone. Reaction with dialkyl carbonate and alkali metal hydride or alcoholate, followed by neutralization and saponification of the resulting ester produces fluorene-9-carboxylic acid; the latter is also produced by metalation with butyl lithium followed by carboxylation. Nitration leads predominantly to 2-nitrofluorene, chloromethylation to 2-chlormethylfluorene. Halogenation, depending on the reaction conditions, gives the mono-, di-, or tri-haloderivatives, substitution taking place at the 2-, 2,7-, or 2,4,7-positions; in the presence of catalysts waxy products containing ca. 50 wt % chlorine are formed.

Production. High-temperature coal tar contains on average, 2 % fluorene. The latter is recovered industrially by redistillation of coal tar wash oil or by direct removal of a fluorene fraction during primary distillation of coal tar, followed by recrystallization of the fluorene fraction (dephenolated and debased if necessary) e.g., from solvent naphtha or naphtha [289]. Fluorene can be isolated from a ≥60 % fluorene fraction by continuous counter-current crystallization [290]. It is possible to refine the technical product further via the sodium compound (by reaction with sodium or sodium amide), or by sulfuric acid purification and crystallization from methanol.

Uses. Fluorene is the starting material for production of fluorenone and fluorene-9-carboxylic acid. Fluorene-9-carboxylic acid is an intermediate in the production of antidiabetic and antiarrhythmic drugs, and plant growth regulators [291]. Plant growth regulators are also obtained by reaction of fluorene with phthalic anhydride [292]. Fluorene-2-acetic acid (by acylation of fluorene) can be used as a precursor in

the synthesis of anti-inflammatory drugs [293]. Reaction of fluorene with sulfur produces electrically conductive polymers [294]. The azine dye Sirius Light Violet FRL is produced from 2-aminofluorene (→Azine Dyes, Chap. 5.2.). The fluorene analog of malachite green dye is obtainable from 3,6-bis-(dimethylaminofluorene) [295]. 2,7-Diiodofluorene can be employed in the production of styryl dyes for optical brighteners [296]; heptabromofluorene can be used as a fire-retardant for plastics [297]. Fluorene itself can be used in admixture with biphenyl and phenyltoluene as a carrier for dyeing polyester fibers [298].

Derivatives.

9-Fluorenone [*486-25-9*], $C_{13}H_8O$, M_r 180.21;

mp	85 – 86 °C
bp 101.3 kPa)	341.5 °C
ϱ (99.4 °C)	1.130 g/cm^3
n_D (99.4 °C)	1.6369

9-Fluorenone forms yellow crystals; insoluble in water; soluble in hot sulfuric acid; very readily soluble in ethanol and diethylether; volatile in steam. Fluorenone can be produced by catalytic oxidation of fluorene [261, 299–302], or of fluorene fractions in the presence of a quaternary ammonium salt [303], or by catalytic oxidative cracking (oxicracking) of suitable aromatic [304].

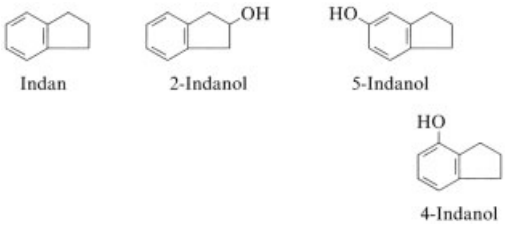

Di- and trinitrofluorenones are synthesized from fluorenone for use as electron acceptors in electrophotography, e.g., in copying systems [305]. 2,4,7-Trinitrofluorenone is used as an analytical reagent for polynuclear aromatic hydrocarbons. Fluorenone can be used as starting material for the synthesis of imidazolylfluorene salts with antimycotic activity [306]. 9,9-Bis(4-hydroxylphenyl)fluorene (from fluorenone by reaction with phenol [307]) can be used as starting material for thermally stable plastics [308]. 9,9-Bis(4-methylaminophenyl)fluorene (from fluorenone by reaction with phenyl methyl amine) has been suggested as a hardener for epoxy resins [309]. Fluorenone is also suitable as an oxidizing agent in the Oppenauer reaction.

5.4. Fluoranthene

Fluoranthene [*206-44-0*], $C_{16}H_{10}$, M_r 206.26, was isolated from coal tar by Fittig and Gebhardt in 1878.

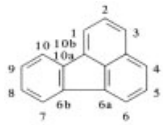

Physical Properties. Colorless crystals with light blue fluorescence; insoluble in water; readily soluble in diethyl ether, boiling ethanol, chloroform, carbon disulfide, and glacial acetic acid.

mp	110 °C
bp (101.3 kPa)	384 °C
ϱ (20 °C)	1.236 g/cm^3
Heat of vaporization	3.29×10^5 J/kg
Heat of combustion (25 °C)	3.913×10^7 J/kg

Chemical Properties. Oxidation of fluoranthene with chromic acid leads to fluorenone-1-carboxylic acid via fluoranthene-2,3-quinone; hydrogenation leads to 1,2,3,10 b-tetrahydrofluoranthene, and to perhydrofluoranthene via decahydrofluoranthene. 2,3-Dihydrofluoranthene is obtained by metalation with sodium. Halogenation, nitration, and sulfonation take place predominantly at the 4-position. Condensation with phthalic anhydride in the presence of aluminum chloride gives a mixture of 7,8- and 8,9-phthaloyl fluoranthene.

Production. Fluoranthene belongs to the main constituents of high-temperature coal tar, which contains, on average, 3.3 % of fluoranthene. It is recovered from the fluoranthene fraction, which boils at 373 – 385 °C, and is obtained from the high-boiling anthracene oil II and from pitch distillate by redistillation. Subsequent recrystallization of the fluoranthene fraction from

solvent naphtha gives 95 % pure fluoranthene. Further refining is carried out by recrystallization from xylene with simultaneous partial sulfonation of the impurities with ca. 1 % of concentrated sulfuric acid, or by recrystallization from a pyridine – water mixture.

Uses. Fluoranthene is used in the production of fluorescent dyes (→Fluorescent Dyes). Yellow vat dyes are obtained by condensation with phthalic anhydride, via phthaloyl fluoranthenes, whereas bromination [310] and condensation with 1-amino-4-benzoyl-amino anthraquinone produces olive dyes [311]. 2-Benzylfluoranthene (from 2,3-dihydrofluoranthene [312] and benzaldehyde) can also be used for the production of dye intermediates [313]. Fluoranthene and alkyl fluoranthenes can be used as stabilizers in epoxy resin adhesives [314], as additives in electrically-insulating epoxy resin compositions [315], and in electrically-insulating oils [316]. 2,3-Dihydrofluoranthene and 1,2,3,10 b-tetrahydrofluoranthene [317] are used for numerous derivatives, some of which (e.g., their derivatives of phenylethyl amine) are of pharmaceutical interest. 7-Fluoranthenyl aminoalcohols (for cytostatics, anthelmintics, or bactericides [318]) and bis-aminoketones, e.g., 3,8-bis(4-piperidinobutyryl)fluoranthene (as antiviral agents [319]) can be synthesized from fluoranthene.

5.5. Phenanthrene

Phenanthrene [*85-01-8*], $C_{14}H_{10}$, M_r178.24 was discovered in coal tar by FITTIG and OSTERMAYER in 1872.

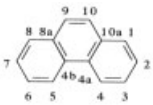

Physical Properties. Colorless flakes with weak blue-violet fluorescence; readily sublimable; insoluble in water; slightly soluble in ethanol; readily soluble in diethyl ether, benzene, carbon disulfide, and glacial acetic acid.

mp	101 °C
bp (101.3 kPa)	338.5 °C
ϱ (20 °C)	1.172 g/cm^3
Heat capacity (25 °C)	1.361×10^3 J/kg
Heat of fusion	1.04×10^5 J/kg
Heat of combustion (25 °C)	3.982×10^7 J/kg

Chemical Properties. Phenanthrene can be oxidized to phenanthrenequinone, diphenic acid, or to phthalic acid, depending on the oxidant. Hydrogenation produces 9,10-di-, tetra-, octa-, or perhydrophenanthrene. Phenanthrene is halogenated predominantly in the 9,10-positions, and nitrated in the 9-position. Sulfonation leads to mixtures of 2-, 3-, and 9-phenanthrene sulfonic acids. For acylation see →Acylation and Alkylation, Chap. 2.2.2.

Production. Phenanthrene, at a concentration of 5 %, is the second most important coal tar constituent in terms of quantity after naphthalene. During primary distillation of coal tar, it is concentrated in the anthracene oil fraction. After crystallization of the anthracene residues the phenanthrene is recovered as a fraction from the filtrate of this crystallization, or from the top fraction of crude anthracene distillation, by redistillation. Technically pure grades of phenanthrene are obtained by sulfuric acid refining and recrystallization from methanol, or by repeated rectification of the phenanthrene fraction. The accompanying substances can be separated either by partial sulfonation, or by partial condensation with formaldehyde and hydrogen chloride. The most persistent accompanying substance, diphenylene sulfide, can be completely removed by treating the melt with sodium and maleic anhydride.

Uses. Phenanthrene forms the basis for production of 9,10-phenanthrenequinone and 2,2′-diphenic acid (see Chap. 4) [320]. It can be used to synthesize anthracene, via the isomerization product of *sym*-octahydrophenanthrene (→Anthracene). Electrically conductive substances, e.g., for use in batteries and solar cells, can be produced by the electrochemical conversion of phenanthrene diazonium salts in a solvent containing a conductive salt, and subsequent doping with various ions (e.g., Na^+, Ba^{2+}, H^+, etc.) [321]. Liquid-crystalline 7-*n*-alkyl-9,10-dihydrophenanthrene-2-carboxylic acid ester, used for optical-electronic ap-

plications, can be synthesized from 9,10-dihydrophenanthrene [322]. By cross-linking with *p*-xylylene glycol and 4-toluenesulfonic acid, polycondensed thermosetting resins are obtained for composites or temperature-resistant, electrically insulating coatings [323]. A polyamide – polyimide resin can be produced by oxidation of phenanthrene to phenanthrene-9,10-quinone and -9,10-diol, condensation with formaldehyde, oxidation to the polycarboxylic acid, formation of the anhydride and finally reaction with an aromatic diamine. This resin is suitable for use in high-temperature insulators, printed circuit boards, and laminates [324]. Phenanthrene has been proposed as a plasticizer for plastics and molding compounds [325], phenanthrene and alkylphenanthrenes have been suggested as stabilizers for mineral oil products [326,327]. Deep khaki vat dyes are prepared by heating a mixture of phenanthrene and anthracene with sulfur [328]. [(Phenanthrylmethyl)amino]-propandiol, which possesses antitumor activity, can be synthesized via phenanthrene-9-carboxaldehyde [329].

Derivatives.

Phenanthrene-9-10-quinone [*84-11-7*], $C_{14}H_8O_2$, M_r 208.91; *mp* 217 °C; orange needles; poorly soluble in water; slightly soluble in ethanol; soluble in boiling water and in diethyl ether; readily soluble in hot glacial acetic acid. Phenanthrene-9,10-quinone is produced by liquid-phase oxidation of phenanthrene with chromates [330], catalytically with oxygen [331], with *tert*-butylhydroperoxide [332,333] or via phenanthrene-9,10-oxide with nitric acid or hypochlorite [334,335]. Further oxidation transforms the quinone into 9-hydroxyfluorene-9-carboxylic acid, which is used in the form of its salts or esters (in combination with the herbicide methyl chlorophenoxyacetic acid (MCPA), trade name: Aniten) or as the 2-chloro derivative (trade name: Maintain) as plant growth regulators [336,337]. Phenanthrene-9,10-quinone has been suggested as an additive for photographic or electrophotographic applications [308,338, 339], in UV-curable coatings and adhesives [340,341] and in production of cellulose by wood pulping [342]. An intermediate for azo pigments can be obtained by condensation with hydrazinobenzoic acid [343]. Condensation with aromatic amines gives intermediates for pharmaceuticals (e.g., immuno-suppressives) or fungicides [344,345].

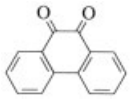

5.6. Pyrene

Pyrene [*129-00-0*], $C_{16}H_{10}$, M_r202.26, was found in coal tar by Graebe in 1871.

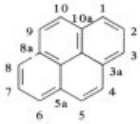

Physical Properties. Colorless crystals with blue fluorescence; insoluble in water; very readily soluble in diethyl ether, benzene, and carbon disulfide; poorly soluble in ethanol.

mp	150 °C
bp (101.3 kPa)	393.5 °C
ϱ (23 °C)	1.271 g/cm^3
Heat capacity (18 °C)	1.126×10^7 J/kg
Heat of vaporization	3.21×10^5 J/kg
Heat of combustion (25 °C)	3.878×10^7 J/kg
Dielectric constant (17 °C)	3.21

Chemical Properties. Pyrene is oxidized by mild oxidants such as chromic acid to pyrene-1,6- and 1,8-quinone; further oxidation leads to perinaphthenone [*548-39-0*] and napthalene-1,4,5,8-tetracarboxylic acid. Hydrogenation produces, depending on the reaction conditions, ditetra-, hexa- or decahydropyrenes; halogenation products substituted in the 1-, 1,6-, 1,8-, 1,3,6- and 1,3,6,8-positions. The substitution pattern for halogenation can be changed in favor of the 4-position by Diels – Alder reaction with hexachlorocyclopentadiene [*77-47-4*] in the presence of iron powder or thallium acetate, followed by retro-Diels – Alder reaction. With nitration and sulfonation the 1-, 1,6-, and 1,3,6,8-substitution products are formed. Pyrene can be condensed with phthalic anhydride by Friedel – Crafts reaction to give diphthaloyl pyrenes, and with benzoyl chloride to the dye pyranthrone [*128-70-1*].

Production. High-temperature coal tar contains an average of ca. 2 % of pyrene. It is recovered from a fraction crystallizing above 110 °C, which is obtained by redistillation of the high-boiling anthracene oil II or pitch distillate. Pure pyrene is produced by recrystallization, e.g., from solvent naphtha [346] or by fractional crystallization from the melt [256], dephenolation and debasing, and by refining with 80 % sulfuric acid. As an alternative to sulfuric acid refining, the pyrene-accompanying brasane (2,3-benzodiphenylene oxide) can also be separated by recrystallization from xylene by adding iron(III) chloride. Traces of tetracene are removed by reaction with maleic anhydride.

Uses. The dye intermediate naphthalene-1,4,5,8-tetracarboxylic acid [*128-97-2*] can be produced by halogenation of pyrene [347, 348] (halogenation of the pyrene fraction is also possible [349]) followed by oxidation (→Carboxylic Acids, Aromatic, Chap. 4.7.), [350]. Other pyrene-based dyes are Sirius Light Blue F 3 GL from 3-aminopyrene (→Azine Dyes, Chap. 5.2.) and the green fluorescent dye, pyranin (Solvent Green 7), from pyrene-1,3,6,8-sulfonic acid [351]. Anthraquinone dyes (e.g., pyranthrone) may be obtained from dibenzoyl pyrene or diphthaloyl pyrene. Optical brighteners may be synthesized by reaction of pyrene with a complex of cyanuric chloride and aluminum chloride (→Optical Brighteners) [352]. By analogy to fluoranthene, pyrene and alkylpyrenes can be used as additives in electro-insulating oils [316], and in epoxy resins for electrical-insulation [315]. In a similar manner to phenanthrene, thermosetting resins may be formed with *p*-xylene glycol and 4-toluenesulfonic acid [323]. Condensation products of 1-bromopyrene or nitropyrene with formaldehyde are used as photoconductive substances in electrophotography [305, 353, 354]. 1,2,3,6,7,8-Hexahydropyrene is effective as a synergist for dialkylsulfide antioxidants in lubricants [355]. On the basis of pyrene, pyrenyl aminoalcohols can be synthesized as cytostatics, anthelmintics, or bactericides [356]. Pyrene itself can serve as an electron donor to increase the blackness in pencil leads [357].

6. Toxicology and Occupational Health

6.1. Alkanes

Methane is toxicologically virtually inert. In very high concentrations (80 – 90 vol %) it causes respiratory arrest [358]. The bulk of an inhaled dose is exhaled unchanged [359].

Ethane easily causes dyspnea in laboratory animals, beginning at concentrations of approximately 2 – 5 vol % in the air inhaled [360]. At high concentration respiratory arrest occurs [361]. A slight sensitization of the myocard to catecholamines has been described at a concentration of 15 – 19 vol % [362].

Propane has narcotic properties [361]. It is neither a skin nor a mucosal irritant. At very high concentration it also causes respiratory arrest [363]. In dogs adverse effects on circulatory function were found with inspiratory concentrations of 1 % or greater. A negative inotropic effect, a drop in the mean aortic pressure, and an increase in pulmonary vascular resistance were found [364].

MAK (1987): 1000 ppm.

Butane causes slight drowsiness beginning at a concentration of 10 000 ppm in the air inhaled [365]. In the mouse the inhalatory LC_{50} is 680 g/m^3 [366]. It has been proven in the dog that butane also sensitizes the myocardium to the effects of catecholamines [364].

MAK (1987): 1000 ppm.

n-*Pentane* has a less pronounced narcotic effect than the C_1–C_4 hydrocarbons. In humans the lowest lethal dose after acute inhalatory exposure is apparently 130 000 ppm, the lowest toxic dose 90 000 ppm [367]. Pentane apparently possesses a neurotoxic effect, although it is not very pronounced [368, 369].

n-*Hexane* possesses a marked neurotoxic effect that distinguishes it from the other alkanes. The causal factor for peripheral neuropathies is the oxidation product (cytochrome P 450-dependent oxidases) 2,5-hexadione. Toxicity manifests itself clinically in polyneuropathies [370, 371]. Hexane is also a skin irritant. The LC_{50} (inhalation mice) is 120 g/m^3 [372]. In rats the acute oral LD_{50} is 24 – 49 mL/kg [373]. Hexane is absorbed through the skin. Dermal exposure can lead to poisoning (LD dermal in rabbits ca. 5 mL/kg) [374]. In rats subchronic expo-

sure to a concentration of 400 – 600 ppm leads to neuropathies [375], which are characterized by degeneration of both myelin sheaths and axons [376–378].

MAK (1987): 50 ppm.

n-*Heptane* has narcotic properties. In humans 0.1 % in the inhaled air leads to dizziness and 0.5 % to equilibrial disturbances with loss of motor coordination [365]. Within 3 min 4.8 vol % in the air inhaled leads to asphyxia [379]. Like hexane, heptane is a skin and mucosal irritant [365]. The biotransformation of *n*-heptane, as with hexane, is by oxidation. Heptane apparently also possesses neurotoxic properties [380]. Myocardial sensitization to catecholamines has also been found [381].

MAK (1987): 500 ppm.

n-*Octane* is similar to *n*-heptane in regard to its narcotic effect; however, it does not seem to cause any other effects on the nervous system [372]. Octane is also metabolized by oxidation [382].

MAK (1987): 500 ppm

Only limited toxicological data are available on the higher molecular mass alkanes.

6.2. Alkenes

The alkenes under consideration are toxicologically not very active. Higher molecular mass compounds possess narcotic properties. The alkenes are not neurotoxic. The α-olefins generally seem to be more reactive and toxic than the β-isomers.

6.3. Alkylbenzenes

Trimethylbenzenes. The toxicological data available on trimethylbenzenes is mainly rather old. The acute inhalatory toxic dose for the 1,2,4- and the 1,3,5-isomers is in the range 7000 – 9000 ppm [383, 384].

As a result of their occurrence in automotive fuel and heating oil, low concentrations of trimethylbenzenes can be demonstrated in air and water. For example, measurements in 1974 in a tunnel in Rotterdam indicated a level of 0.015 ppm pseudocumene; see [385] for more details. An analysis of the drinking water of Cincinnati, Ohio, in 1980, revealed a concentration of 45 ng/L hemimellitene, 127 ng/L pseudocumene, and 36 ng/L mesitylene [386]. Trimethylbenzenes are toxic to aquatic organisms; LC_{50} values between 10 and 100 ppm have been reported [387].

The TLV (TWA) level for trimethylbenzene is 25 ppm [388]. The MAK for Trimethylbenzenes (all isomers (1,2,3–;1,2,4–;1,3,5–) has been set at 20 ppm. [389]. The United States toxicity assigment ranges from low to moderately toxic for pseudocumene, to highly toxic for the isomer mixture; disturbances in the central nervous system and abnormal blood pictures have been reported [387].

Tetramethylbenzenes. The oral toxicity of tetramethylbenzenes is low; LD_{50} >5000 mg/kg (rat, oral). Durene is additionally classified as intravenously highly toxic; LD_{50} 180 mg/kg (i.v., mice). Tetramethylbenzenes cause mild skin reactions [387]. Two values are cited for the odor threshold concentration of durene: 0.083 and 0.087 mg/m^3 [385].

Hexamethylbenzene. Hexamethylbenzene has low oral toxicity; LDLo 5000 mg/kg (rat). It is, however, suspected of causing neoplastic effects. Addition of nitromethane to hexamethylbenzene can lead to an explosive reaction [387].

Diethylbenzenes. Diethylbenzene is more toxic than monoethylbenzene (→Ethylbenzene) [390]. DEB is emitted into the environment from engine fuel and heating oil [386]. Diethylbenzenes are toxic to aquatic organisms; the LC_{50} is between 10 and 100 mg/L [387]. In the cell multiplication inhibition test with protozoa, the toxicity threshold is <10 mg/L [385]. In the Federal Republic of Germany DEB is classified as an aquatic hazard; aquatic hazard class 2. DEB causes mild to moderate irritation of the eye and mucous membranes. The oral toxicity is low: LDLo 5000 mg/kg (rat) [387]. The vapor possesses an anesthetic action and may cause headache, vertigo, or vomiting; the odor threshold concentration is <10 ppm [391].

Triethylbenzene. The LC_{50} of triethylbenzene (aquatic organisms) is between 100 and 1000 mg/L. The oral toxicity is low: LDLo 5000 mg/kg (rat). Implant experiments with hexaethylbenzene in mice have raised the suspicion that the compound may be carcinogenic [387].

Ethyltoluene. Ethyltoluene can, as with aromatic hydrocarbons, be detected in the environment [385]. The oral toxicity of 2- and 4-ethyltoluene is low: LDLo 5000 mg/kg (rat) [387].

Cumene. Cumene appears to be orally as nontoxic as propylbenzene, it is even less so by inhalation [383,392,393]. Cumene has a prolonged depressant effect on the central nervous system [394].

The MAK for cumene has been set at 250 mg/m^3, corresponding to a cumene vapor concentration in air of 50 mL/m^3. The vapor pressure of cumene at 20 °C is 0.5 kPa. There is a risk of cumene being absorbed through the skin [395]. Inhalation of cumene vapor leads to the delayed appearance of a long lasting narcotic effect [396].

6.4. Biphenyls and Polyphenyls

Biphenyls. Biphenyl dust or vapor is irritating to the eye and mucous membrane at a concentration as low as 3 – 4 ppm [390], and to the skin after extended exposure. A concentration of >5 mg/m^3 for long periods is considered a health hazard; systemic toxic effects were elicited in humans by a concentration with inhalatory exposure maxima of 128 ppm [400,401]. The olfactory threshold is ca. 0.06 – 0.3 mg/m^3. Some toxicity data are listed below [397–399]:

LD_{50} (rat, oral)	3280 mg/kg
LD_{50} (rabbit, oral)	2400 mg/kg
LD_{50} (rabbit, dermal)	2500 mg/kg
TL_m (fathead minnow, 96 h)	1.5 mg/L

Triphenyls are only minimally toxic [402].

Terphenyls. Toxicologically, terphenyl isomers should be treated like biphenyl. The oral LD_{50} is ca. 4.6 – 4.7 g/kg [403]. The toxicity of partially (40 %) hydrogenated terphenyl (LD_{50} 17.5 g/kg) is much less than that of the fully aromatic terphenyls [404–407].

6.5. Hydrocarbons from Coal Tar

The aromatic hydrocarbons under consideration are usually found in complex mixtures of various other polycyclic aromatic hydrocarbons (PAH), which originate from combustion emissions and industrial processes. For this reason toxic phenomena possibly caused by exposure to humans are often unlikely to be attributed to one single aromatic compound. Incorporation of these compounds into the body occurs predominantly by way of the respiratory tract. Skin contact is important with occupational exposure.

6.5.1. Biological Effects

6.5.1.1. Carcinogenicity and Mutagenicity

The aromatic compounds under discussion are considered as noncarcinogens, but pyrene and fluoranthene are able to enhance the carcinogenic potential of benzo [*a*] pyrene [*50-32-8*] when applied simultaneously to mouse skin (cocarcinogenic effect) (→Carcinogenic Agents) [408]. Phenanthrene resulted in just one positive tumor-initiation test in a series of carcinogenicity studies [408,409]. Acenaphthene, known as an inhibitor of mitosis in plant cells [410], induced the initial phase of bronchocarcinoma in rats after an inhalation period (4 months) of 4 hours per day with 0.5 – 1.25 mg/m^3 [411], but was negative in repeated topical application to mouse skin [412]. With their low activity in in vivo and in vitro assays, there is limited evidence that phenanthrene and pyrene are mutagenic [408,409,413–415]. As for fluoranthene, several more recent positive results indicate a mutagenic potential [408,416–419].

So far, tests on fluorene have given negative results [408].

6.5.1.2. Mammalian Toxicity and Toxicokinetics

The acute toxicity of the aromatic compounds under consideration is low: LD_{50} (oral and dermal) is 700 – 2700 mg/kg body weight in rodents [408,409,411,413]; LD_{50} (rat, oral) for acenaphthene is 10 000 mg/kg [421]. The LC_{50} (inhalation) of pyrene is reported to be 170 mg/m^3 [413]. No fatalities occurred in rats exposed to indene vapor (800 – 900 ppm) for six inhalation periods of 7.5 h each; however, systemic pathological changes in the vascular system and several organs were observed [412].

Longterm inhalation (4 months) of acenaphthene (0.5 – 1.25 mg/m^3) [411,420] or pyrene (0.3 – 3.6 mg/m^3) [408,409,413] caused local irritation and systemic pathological effects in several organs and the blood of rodents. In direct contact at high concentration, the compounds under consideration are irritating to the skin and mucous membranes; the tri- and polycyclic aromatics especially are photosensitizing [408–410,412,413]. The compounds under consideration are absorbed on inhalation or oral exposure, and partly leave the body unchanged in the feces and by exhalation [408,409,412,413]. After metabolic conversion to more water-soluble intermediates (hydroxylation and conjugation), they are excreted into the bile or the urine. The half-life of pyrene is estimated to be 24 – 48 h in rats and swine [413].

6.5.1.3. Ecotoxicology

Several studies have shown that the aromatic compounds, in question are microbially decomposed [422–424], although the microbial conversion of indene has not yet been described. These compounds tend to be enriched in sediments, sludges, and aquatic organisms because of their low water solubility. As a result of microbial degradation, biotransformation, and photochemical decomposition, however, accumulation is limited and checked. Nonetheless, a concentration in water of the order of their low solubility (<0.2 – 4.0 mg/L) may have adverse effects on aquatic life, as shown for acenaphthene [425], phenanthrene, and pyrene [426,427].

6.5.2. Safety Regulations

A threshold limit value (TLV) at the work place exists only for indene: 10 ppm (45 mg/m^3) (ACGIH, United States 1980). No other TLVs have so far been established. The United States standard for exposure to coal tar volatiles (0.2 mg/m^3) was recommended to be used in case of pyrene-containing PAH mixtures, but a maximum level of 0.1 mg/m^3 should be applied in the case of exposure to pure pyrene [413]. With the exception of indene all aromatics are included in the recommended United States list of priority pollutants (United States EPA, 1977) [428].

7. References

1. F. D. Rossini, B. J. Mair, *Adv. Chem. Ser.* **5** (1952) 341.
2. *Selected values of properties of hydrocarbons and related compounds,* Am. Petrol. Inst. Research Project **44,** Supplementary vol. A-72, 1975.
3. G. A. Olah et al., *J. Am. Chem. Soc.* **91** (1969) 3261.
4. G. A. Olah et al., *J. Am. Chem. Soc.* **97** (1975) 6807.
5. G. A. Olah et al., *J. Am. Chem. Soc.* **95** (1973) 7686.
6. G. A. Olah et al., *J. Am. Chem. Soc.* **95** (1973) 7680.
7. K. Nabert, G. Schön, *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* Deutscher Eichverlag, Berlin 1965.
8. R. C. Weast, ed., *Handbook of Chemistry and Physics,* 55th ed., CRC Press, Cleveland 1974.
9. C. Zerbe, *Mineralöle und Verwandte Produkte,* Springer, Berlin 1969.
10. Deutsche Shell AG, *Erdöl-Nachrichten,* March 1976, p. 4.
11. F. Schmerling, H. Heneka, *Erdöl Kohle Erdgas Petrochem.* **21** (1968) 705.
12. K. Hedden, J. Weitkamp, *Chem. Ing. Tech.* **47** (1975) 505.
13. J. W. Ward, *Hydrocarbon Process.* **54** (1975) no. 9, 101.
14. R. Oberkobusch, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **28** (1975) 558.
15. J. Schulze, *Chem. Ing. Tech.* **46** (1974) 925, 976.
16. H. Pichler, H. Schulz, *Chem. Ing. Tech.* **42** (1970) 1162.
17. C. D. Frohning, B. Cornils, *Hydrocarbon Process.* **53** (1974) no. 11, 143.
18. J. C. Hoogendorn, *Gas Wärme Int.* **25** (1976) 283.
19. K. C. Hellwig, S. B. Alpert, E. S. Johanson, R. H. Wolk, *Brennst. Chem.* **50** (1969) 263.
20. F. W. Richardson: *Oil from Coal,* Noyes Data Corp., Park Ridge, N.J., 1975.
21. K. F. Schlupp, H. Wien, *Angew. Chem.* **88** (1976) 348.
22. G. Nashan, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **28** (1975) 562.
23. G. Collin, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **29** (1976) 159.

24. H.-W. v. Gratkowski, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **28** (1975) 81.
25. H. Hiller, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **28** (1975) 74.
26. P. D. Bartlett, K. E. Schueller, *J. Am. Chem. Soc.* **90** (1968) 6071.
27. *Houben-Weyl,* 4th ed., **vol. IV,** parts 3 and 4.
28. J. H. Lucas, F. Baardman, A. P. Kouwenhoven, *Angew. Chem.* **87** (1975) 740, **88** (1976) 412.
29. W. C. van Ziill Langhout, *Proc. World Pet. Congr.* 1975, vol. 5, 197.
30. W. L. Lafferty, Jr., W. R. Stokeld, *Adv. Chem. Ser.* **103** (1971) no. 7, 130.
31. L. F. Albright, *Chem. Eng.* **73** (1966) no. 14, 119; **73** (1966) no. 19, 205; **73** (1966) no. 21, 209.
32. L. Alcock, R. Martin, K. H. Bourne, W. A. Peet, J. Winsor, *Oil Gas J.* **72** (1974) no. 27, 102.
33. D. P. Thornton, D. J. Ward, R. A. Erickson, *Hydrocarbon Process.* **51** (1972) no. 8, 81.
34. H. Jockel, B. E. Triebskorn, *Hydrocarbon Process.* **52** (1973) no. 1, 93.
35. *Hydrocarbon Process.* **52** (1973) no. 11, 145.
36. P. R. Pujado, B. V. Vora, *Energy Prog.* **4** (1984) no. 3, 186.
37. M. S. Scurrell, *Appl. Catal.* **32** (1987) 1.
38. P. Leprince, A. Chauvel, L. Castex: *Procédés de Pétrochimie,* Editions Technip, Paris 1971, p. 269.
39. K. Ito, I. Dogane, K. Tanaka, *Hydrocarbon Process.* **64** (1985) no. 10, 83.
40. S. C. Arnold, G. D. Sucui, L. Verde, A. Neri, *Hydrocarbon Process.* **64** (1985) no. 9, 123.
41. Japan. Chemical Reporter, 1981, October, p. 2.
42. F. Solymosi, I. Tombácz, G. Kutsan, *J. Chem. Soc. Chem. Commun.* 1985, 1455.
43. R. V. Porcelli, B. Juran, *Hydrocarbon Process.* **65** (1986) no. 3, 37.
44. W. Hinsen, M. Baerns, *Chem. Zt.* **107** (1982) 9.
45. Kh. E. Khcheyan, O. M. Revenko, A. N. Shatalova: "Oxidative Methylation, a new route to petrochemicals through alternative feedstock," Special Paper No. 16, World Petroleum Congress 11th, 1983 (London).
46. G. de Pontanel: *Proteins from Hydrocarbons,* Academic Press, London 1972.
47. S. Gutcho: *Proteins from Hydrocarbons,* Noyes Data Corp., Park Ridge, N.J., 1973.
48. P. J. Rockwell: *Single Cell Proteins from Cellulose and Hydrocarbons,* Noyes Data Corp., Park Ridge, N.J., 1976.
49. G. Voss, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **26** (1973) 249.
50. W. Dimmling, R. Seipenbusch, *Hydrocarbon Process.* **54** (1975) no. 9, 169.
51. H. Riegel, H. D. Schindler, M. C. Sze, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **26** (1973) 704.
52. R. P. Lowry, A. Aguilo, *Hydrocarbon Process.* **53** (1974) no. 11, 103.
53. B. V. Vora, T. Imai, *Hydrocarbon Process.* **61** (1982) no. 4, 171.
54. L. B. Phillips, *Hydrocarbon Process.* **42** (1963) no. 6, 159.
55. A. A. Di Giacomo, J. B. Maerker, J. W. Schall, *Chem. Eng. Prog.* **57** (1961) no. 5, 35.
56. Universal Oil Products, US 3 649 177, 1972 (D. H. Rosback).

General References

57. F. Asinger: *Die Petrolchemische Industrie,* vol. 1 (Gewinnung der Olefine), vol. 2 (Verwendung der Olefine) Akademie-Verlag, Berlin 1971.
58. J. Falbe (ed.): *Methodium Chimicum,* **vol. 4,** Thieme Verlag, Stuttgart 1980, pp. 35, 111.
59. *Houben-Weyl,* 4th ed., **vol. 5/1 b,** Thieme Verlag, Stuttgart 1972.
60. S. Patai, J. Zabicky: *The chemistry of alkenes,* **vol. 1 (1964),** vol. 2 (1970), Interscience Publishers, New York.
61. E. S. Strauss, G. Hollis, O. Kamatari: *Chemical Economics Handbook Marketing Research Report* "Linear alpha-olefins," Stanford Research Institute, Menlo Park, CA, 1984.
62. B. Fell, *Tenside Deterg.* **12** (1975) 3.
63. H. Isa, *CEER Chem. Econ. Eng. Rev.,* **9** (1977) 26.
64. A. H. Turner, *J. Am. Oil Chem. Soc.* **60** (1983) 623.
65. A. Behr, W. Keim, *Arabian J. Sci. Eng.* **10** (1985) 377.

Specific References

66. K. G. Mittal, *J. Chem. Technol. Biotechnol.* **36** (1986) 291 – 299.
67. D. B. Broughton, R. C. Berg, *Chem. Eng.* **77** (1970) 86.
68. B. V. Vora, P. R. Pujado, J. B. Spinner, T. Imai, Proc. World Tenside Congr. **1984** (Munich) vol. II, 16 – 26.
69. R. C. Berg, B. V. Vora in McKetta, Cunningham (ed.): *Encyclopedia of Chemical Processing and Design,* **vol. 15,** Marcel Dekker, New York 1982, 266 – 284.
70. H. D. Wulff, *Fette Seifen Anstrichm.* **69** (1967) 32.

71. R. Stöbele, *Chem. Ing. Tech.* **36** (1964) 858.
72. K. Ziegler et al., *Angew. Chem.* **64** (1952) 323, 67 (1955) 542.
73. Gulf Res. Dev. Co., DE 1 443 927, 1961.
74. Ethyl Corp., US 3 906 053, 1973.
75. W. Keim, A. Behr, M. Röper in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 5,** Pergamon Press, Oxford 1982, 371.
76. W. Keim, *Chem. Ing. Tech.* **56** (1984) 850.
77. Shell Development Comp., US 3 635 937, 1972; 3 637 636, 1972; 3 644 564, 1972; 3 644 563, 1972; 3 647 914, 1972; 3 647 915, 1972; 3 661 803, 1972; 3 686 159, 1972 (W. Keim et al.).
78. E. R. Freitas, R. C. Gum, *Chem. Eng. Progress.* **75** (1979) no. 1, 73.
79. E. L. T. M. Spitzer, *Seifen öle Fette Wachse* **107** (1981) 141 – 144.
80. M. Sherwood, Chem. Ind. (London) (1982) 994 – 995.
81. W. Keim, F. H. Kowaldt, *Erdöl Kohle Erdgas Petrochem.* Compendium (1978/79) 453 – 462.
82. Shell Oil Comp., US 3 647 906, 1972 (F. F. Farley).
83. M. Peuckert, W. Keim, *Organometallics* **2** (1983) 594.
84. W. Keim, F. H. Kowaldt, R. Goddard, C. Krüger, *Angew. Chem.* **90** (1978) 493.
85. W. Keim, A. Behr, B. Limbäcker, C. Krüger, *Angew. Chem.* **95** (1983) 505.
86. W. Keim et al., *J. Mol. Catal.* **6** (1979) 79.
87. W. Keim et al., *Organometallics* **5** (1986) 2356.
88. *Chem. Eng. News* (1972), April 10, 16 – 17.
89. C. Commereuc, Y. Chauvin, J. Gaillard, J. Léonard, J. Andrews, *Hydrocarbon Process.* **63** (1984) no. 11, 118 – 120.
90. W. J. Benedek, J.-L. Mauleon, *Hydrocarbon Process.* **59** (1980) no. 5, 143.
91. J. Léonard, J. F. Gaillard, *Hydrocarbon Process.* **60** (1981) no. 3, 99.
92. J. F. Boucher, G. Follain, D. Fulop, J. Gaillard, *Oil Gas J.* **80** (1982) 84 – 86.
93. Y. Chauvin, J. Gaillard, J. Léonard, P. Bonnifay, J. W. Andrews, *Hydrocarbon Process.* **61** (1982) no. 5, 110 – 112.
94. D. Commereux, Y. Chauvin, G. Leger, J. Gaillard, *Rev. Inst. Fr. Pet.* **37** (1982) 639 – 649.
95. J. Gaillard, *Pet. Tech.* **314** (1985) 20 – 27.
96. B. Bogdanović, *Adv. Organomet. Chem.* **17** (1979) 105.
97. W. Keim, A. Behr, G. Kraus, *J. Organomet. Chem.* **251** (1983) 377.
98. A. Behr, V. Falbe, U. Freudenberg, W. Keim, *Isr. J. Chem.* **27** (1986) 277 – 279.
99. W. Keim, *Ann. N.Y. Acad. Sci.* **415** (1983) 191 – 200.
100. B. Bogdanović, B. Spliethoff, G. Wilke, *Angew. Chem. Int. Ed. Engl.* **19** (1980) 622.
101. G. Speier, *J. Organomet. Chem.* **97** (1975) 109.
102. M. G. Barlow, M. J. Bryant, R. N. Haszeldine, A. G. Mackie, *J. Organomet. Chem.* **21** (1970) 215.
103. S. Datta, M. B. Fisher, S. S. Wreford, *J. Organomet. Chem.* **188** (1980) 353.
104. R. Schrock, S. McLain, J. Sancho, *Pure Appl. Chem.* **52** (1980) 729.
105. J. A. Brennan, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 2 – 6.
106. R. L. Shubkin, M. S. Baylerlan, A. R. Maler, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 15 – 19.
107. F. J. Karol, *CHEMTECH* 1983 (April) 222.
108. B. W. Werdelmann, *Fette Seifen Anstrichm.* **76** (1974) 1.
109. *Chem. Eng. News,* (1988) May 30, 10.
110. *Eur. Chem. News,* (1987) Dec. 21/28, 21.
111. *Eur. Chem. News,* (1981) July 6, 23.
112. *Eur. Chem. News,* (1987) Jan. 5/12, 19.
113. *Eur. Chem. News,* (1987) Jan. 19, 8.
114. *Eur. Chem. News,* (1981) Jan. 26, 27.
115. *Eur. Chem. News,* (1982) July 5, 26.
116. G. Wilke, *Angew. Chem. Int. Ed. Engl.* **2** (1963) 105.
117. A. C. L. Su, *Adv. Organomet. Chem.* **17** (1979) 269.
118. S. Warwel, *Erdöl Erdgas Kohle* **103** (1987) 238.
119. R. C. Weast, M. J. Astle (eds.): *CRC Handbook of Data on Organic Compounds,* CRC Press, Boca Raton, Florida, 1985.
120. S. Coffey (ed.): *Rodd's Chemistry of Carbon Compounds,* III A, Elsevier, Amsterdam 1971, pp. 228 – 231.
121. A. D. Rudkovskii et al., *Chem. Technol. Fuels Oils (Engl. Transl.)* **19** (1983) 387 – 391.
122. F. P. McCandless, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 612 – 616; **21** (1982) 483 – 488.
123. Idemitsu Kosan Co. Ltd., EP 0 104 607, 1983 (H. Sato).
124. Universal Oil Products, US 4 487 984, 1984 (T. Imai).
125. O. A. Anunziata et al., *Appl. Catal.* **15** (1985) 235 – 245.

126. J. Hanotier, M. Hanotier-Bridoux, *J. Mol. Catal.* **12** (1981) 133 – 147.
127. L. I. Smith, *Org. Synth.,* coll. vol. 2 (1948) 248 – 253.
128. Toyo Soda Manufacturing, DE-OS 3 330 912, 1983 (T. Sakamoto, T. Hironaka, K. Sekizawa, Y. Tsutsumi).
129. Toa Nenryo, EP 0 101 651, 1983 (T. Maejima, N. Tagaya, S. Sakurada).
130. J. Haggin, *Chem. Eng. News,* (1987) June 1, 22 – 28; (1987) June 22, 22 – 25.
131. International Flavors and Fragrances, US 3 869 524, 1974 (K. K. Light, J. B. Anglim).
132. Phillips Petroleum, US 4 568 784, 1985 (R. L. Cobb).
133. S. A. Shama, C. C. Wamser, *Org. Synth.* **61** (1983) 62 – 64.
134. H. F. Davis, S. K. Chattopadhyay, P. K. Das, *J. Phys. Chem.* **88** (1984) 2798–2803.
135. W. W. Kaeding, G. C. Barile, M. M. Wu, *Catal. Rev. Sci. Eng.* **26** (1984) 597 – 612.
136. Dow Chemical, US 2 763 702, 1956 (J. L. Amos, K. E. Coulter).
137. W. W. Kaeding, L. B. Young, C.-C. Chu, *J. Catal.* **89** (1984) 267 – 273.
138. W. W. Kaeding, G. C. Barile, *Polym. Sci. Technol. (Plenum)* **25** (1984) 223 – 241.
139. *Eur. Chem. News,* Oct. 17 (1988) p. 4.
140. Physical Constants of Hydrocarbons C_1 to C_{10}, *"ASTM Spec. Techn. Publ. STP"* **106** (1963).
141. Handbook of Chemistry and Physics, R. C. Weast (ed.), 63rd ed., CRC Press, Boca Raton 1982 – 1983, p. D-213.
142. D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds,* J. Wiley & Sons, New York 1969, p. 370.
143. J. C. Butler, W. P. Webb, *Chem. Eng. Data Series* **2** (1957) no. 1, 42 – 46.
144. G. Sorbe: *Sicherheitstechnische Kenndaten chemischer Stoffe,* **vol. 2,** ecomed-Verlagsgesellschaft mbH, Landsberg 1983, pp. 15 – 16.
145. P. R. Pujado, J. R. Salazar, C. V. Berger, *Hydrocarbon Process.* **55** (1976) no. 3, 91.
146. R. C. Canfield, T. C. Unruh, *Chem. Eng. NY* **90** (1983) no. 3, 32.
147. Toyo Soda Manufacturing Co., DE-AS 2 418 038, 1974 (T. Yanagihera et al.).
148. Hitachi Chemical Co., JP-Kokai 77 139 030, 1974 (E. Kudo et al.); *Chem. Abstr.* **88** (1978) 120757 r.
149. BASF, DE-OS 2 141 491, 1971 (F. Neumayr).
150. Mitsui Petrochemical Ind., EP 0 046 678, 1981 (M. Hisaya).
151. J. P. Fortuin, M. J. Waale, R. P. Oosten, *Pet. Refiner* **38** (1959) no. 6, 189.
152. A. H. Olzinger, *Chem. Eng. NY* **82** (1975) no. 6, 50.
153. Universal Oil Products, US 3 763 259, 1973 (G. L. Hervert).
154. Mobil Oil Corp., EP-A 0 149 508, 1985 (W. W. Kaeding).
155. J. Ewers, H. W. Voges, G. Maleck, *Erdöl Kohle Erdgas Petrochem. Compendium* **74/75** (1975) 487.
156. K. Ito, I. Dogane, K. Tanaka, *Hydrocarbon Process.* **65** (1986) no. 11, 89.
157. H. A. Colvin, J. Muse, *CHEMTECH* 1986, no. 500.
158. K. Ito, *Hydrocarbon Process.* **52** (1973) no. 8, 89.
159. Phillips Petroleum Co., US 3 291 847, 1966 (P. F. Warner); US 3 316 315, 1967 (W. A. Jones).
160. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1987, pp. 72 – 79.
161. A. C. Olson, *Ind. Eng. Chem.* **52** (1960) no. 10, 833.
162. R. Ströbele, *Chem. Ing. Tech.* **36** (1964) 858.
163. H.-D. Wulf, T. Böhm-Gößl, L. Rohrschneider, *Fette Seifen Anstrichm.* **69** (1967) 32.
164. California Research Corp., US 2 477 382, 1949 (A. H. Lewis, A. C. Ettling); *Chem. Abstr.* **44** (1950) 1137 e.
165. H. Hartig, *Chem. Ztg.* **99** (1975) 175.
166. K. Ichikawa, H. Shingu, *Kogyo Kagaku Zasshi* **54** (1951) 183 – 184; *Chem. Abstr.* **47** (1953) 8662 b.
167. S. Murohashi, *Nagaoka Kogyo Koto Senmon Gakko Kenkyu Kiyo* **9** (1973) 103 – 106; *Chem. Abstr.* **80** (1974) 132917 r.
168. J. G. Matveev, D. A. Drapkina, R. L. Globus, *Tr. IREA* **21** (1956) 83 – 89, *Chem. Abstr.* **52** (1958) 15474 e.
169. A. D. Harford, H. W. Vernon, US 2 897 112, 1956.
170. British Petroleum Co., Ltd., DE-AS 1 094 035, 1957 (A. Douglas, H. W. Vernon).
171. N. K. Moshinskaya, V. S. Olifer, L. I. Zhuoiev, SU 137 993, 1960.
172. E. Siggel, US 2 710 849, 1952.
173. BASF, DAS 1 142 338, 1960 (W. Braun, E. Hartwig, R. Krallmann).
174. FMC Corp., US 3 539 640, 1967 (M. I. Stewart, O. K. Carlson).

175. Ashland Oil & Refining Co., US 3 529 944, 1967 (A. M. Leas).
176. Japan Petrochemicals Co., Ltd., DE-OS 2 210 133, 1972 (A. Sato, Y. Aida, I. Shimizi).
177. H. Winkler: *Der Steinkohlenteer,* Verlag Glückauf, Essen 1951, p. 143.
178. J. Chipman, S. B. Peltier, *Ind. Eng. Chem.* **21** (1929), 1106.
179. M. McEwen: *Organic Coolant Handbook,* sponsored by U.S. Atomic Energy Comm., publ. by Monsanto Co., St. Louis, Mo., 1958.
180. J. M. Cork, H. Mandel, N. Ewbank (North American Aviation Inc.), U.S. At. Energy Comm. Sci. Rept. 5129 (1960) 1 – 28. Recalculated by A. N. Syverud, The Dow Chemical Co. Thermal Laboratory, Midland, Mich.
181. M. McEwen, E. Wiederhold, *Nucl. Eng.* Part IV Symp. Ser. 5 (1959) no. 22, 9 – 15.
182. M. J. Foole, *Chem. Eng. Prog.* **58** (1962) no. 4, 37 – 40.
183. *Hydrocarbon Process.* **46** (1967) no. 11, 187.
184. S. Feigelman, C. B. O'Connor, *Hydrocarbon Process.* **45** (1966) no. 5, 140 – 144.
185. *Hydrocarbon Process. Pet. Refin.* **42** (1963) no. 11, 204.
186. Bayer, BIOS Final Rep. Nr. 1787.
187. R. Egbert, Trans. AICHE vol. 34 (1938) 435 – 488.
188. R. Dasgupta, B. R. Maiti, *Ind. Eng. Chem.* **25** (1986) 381 – 386.
189. Otkrytiya/Itobret, SU 1 165 677, 1985.
190. BIOS Final Rep. Nr. 893.
191. Soc. Progil, US 3 227 525, 1964 (M. E. Degeorges, M. Jaymond).
192. Soc. Progil, US 3 112 349, 1961 (M. E. Degeorges, M. Jaymond).
193. Soc. Progil, DE-AS 1 249 244, 1961 (M. Jaymond).
194. Dow Chemical Co., US 1 938 609, 1933 (J. H. Reilly).
195. Monsanto Co., US 3 228 994, 1962 (W. D. Robinson, W. Groves).
196. Hawker Siddeley Nuclear Power Co., FR 1 272 068, 1961.
197. Montecatini, BE 586 826, GB 865 302, IT Anm. 1144/59, 1960, IT Prior. 1959 (C. Ferry, B. Ponzini).
198. Monsanto Co., US 2 143 509, 1939 (C. Conover, A. E. Huff).
199. Montecatini, IT 586 113, 1962.
200. Kambaror YU G; SU 6 79 565, 1977, *Chem. Abstr.* **92** (1980) 58407 d.
201. Monsanto Co., US 3 228 994, 1962 (W. D. Robinson).
202. N. M. Indyukov, R. I. Gasanova, *Khim. prom-st. (Moscow)* 1975, no. 7, 555.
203. Farbenfabr. Bayer AG, DE-AS 1 012 595, 1955 (H. Holzrichter, O. Tegtmeyer).
204. F. Ullmann, *Justus Liebigs Ann. Chem.* **332** (1904) 48.
205. E. I. El'bert, SU 215 911, 1968.
206. R. G. Ismailow, D. A. Aliev, *Azerb. Neft. Khoz.* **44** (1965) no. 1, 30.
207. L. Gattermann, R. Ehrhardt, *Ber. Dtsch. Chem. Ges.* **23** (1890) 1226.
208. I. H. Gardner, P. Bergstrom, *J. Am. Chem. Soc.* **51** (1962) 3376.
209. Asahi Kasei Kogyo K. K., DE-AS 2 039 904, 1970.
210. Shell Oil Co., US 3 145 237, 1962 (R. van Helden, B. Balder).
211. R. van Helden, G. Verberg, *Rec. Trav. Chim. Pays-Bas* **84** (1965) 1263.
212. Gulf Research & Developm. Co., US 3 401 207, 1966 (C. M. Selwitz).
213. E. S. Rudakov, A. I. Lutsyk, R. I. Rudakova, *Kinet. Katal.* **18** (1977) 525; *Kinet. Catal. (Engl. Transl.)* **18** (1977) 441.
214. M. K. Starchevskii, M. N. Vargaftik, I. I. Moiseev, *Izv. Akad. Nauk SSSR* 1979, 242;(Engl. Transl.) 227.
215. Teijin Ltd., US 3 963 787, 1973 (Y. Ichikawa, T. Yamaji).
216. Ube Industries, JP 55 079 324, 1978.
217. L. N. Rachkovskaya et al., *Kinet. Katal.* **18** (1977) 792; *Kinet. Catal. (Engl. Transl.)* **18** (1977) 660.
218. Ube Industries, JP 80 141 417, 1979, *Chem. Abstr.* **94** (1981) 156 561 k.
219. British Petroleum, DE-OS 1 909 045, 1969 (D. Fiolka, C. R. Jentsch).
220. L. H. Slaugh, *Tetrahedron* **24** (1968) 4523.
221. L. H. Slaugh, J. A. Leonard, *J. Catal.* **13** (1969) 385.
222. G. N. Koshel et al., *Dokl. Akad. Nauk SSSR* **237** (1977) no. 1, 164 (Chem. Tech.); *Chem. Abstr.* **88** (1978) 50 394 u.
223. Phillips Petroleum, US 4 123 470, 1977 (T. P. Murths).
224. Farberov M I, SU 416 339, 1972; *Chem. Abstr.* **88** (1978) 62 124 u.
225. Standard Oil, US 4 218 572, 1978 (S. R. Dolhyj, L. J. Velenyi).
226. Y. Yamazaki, T. Kawai, A. Yagishita, *Sekiyu Gakkaishi* **21** (1978) no. 4, 229.
227. Koppers Co., US 3 059 987 1962 (H. P. Baumann).

228. N. Hendrich, *Melliand Textilber.* **43** (1962) 258.
229. W. M. Meylan, P. H. Howard: Chemical Market Input/Output Analysis of Biphenyl and Diphenyl Oxide to Assess Sources of Environmental Contamination, Syracuse NY, EPA Contract No. 68-01-3224, 1976.
230. W. E. Bachmann, H. T. Carke, *J. Am. Chem. Soc.* **49** (1927) 2089 – 98.
231. L. W. Pickett, *Proc. R. Soc. London A* **142** (1933) 333 – 346.
232. H. G. Römer, E. Hertel, *Z. Phys. Chem. Abt. B* **21** (1933) 292 – 296.
233. R. J. Good, E. E. Hardy, A. M. Ellenburg, H. B. Richards, Jr., *J. Am. Chem. Soc.* **75** (1953) 436.
234. J. A. Ellard, W. H. Yanko: *Thermodynamic Properties of Biphenyl and the Isomeric Terphenyls,* Final Report IDO-11,008, Monsanto Research Corp., U.S. At. Energy Comm. Contract AT (10-1)-1088,Oct. 31, 1963.
235. R. O. Bolt, R. O. Johnson, R. S. Kirk: *Organic Materials for Use as Reactor Coolant Moderators,* California Research Corp., June 13, 1956.
236. P. L. Geiringer: *Handbook of Heat Transfer Media,* Reinhold Publ., New York 1962, p. 173.
237. W. H. Hedley, M. V. Mines, *J. Chem. Eng. Data* **15** (1970) 122.
238. H. Mandel, N. Eubank: *Critical Constants of Diphenyl and Terphenyl,* Report No. NAA-SR 5129, Atomics International, Dec. 1960, p. 18.
239. N. Andrews, R. Ubbelohde, *Proc. R. Soc. London* **228 A** (1955) 435 – 37.
240. E. Pajda, J. Rusin, *Chemik Wkladka* 1972, 422.
241. R. Balent, *Proc. Am. Power Conf.* **21** (1959) 120.
242. Soc. Progil, DE-AS 1 272 281, 1962 (M. Jaymond).
243. Farbenfabr. Bayer AG, GB 957 065 = FR 1 297 408, 1961.
244. Hawker Siddeley Nuclear Power Co., FR 1 272 068, 1961.
245. D. B. Orechkin, SU 570 584, 1977.
246. Monsanto Co., DE-OS 2 306 454, 1973 (J. K. Sears).
247. M. He, *Huaxue Shiji* **8** (1986) no. 1, 60 – 61; *Chem. Abstr.* **105** (1986) 208 532.

General References

248. H.-G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Berlin – Heidelberg – New York 1968.
249. H.-G. Franck, J. W. Stadelhofer: *Industrielle Aromatenchemie,* Springer Verlag, Berlin – Heidelberg – New York – London – Paris – Tokyo 1987.
250. *Beilstein,* **5,** 586; **5 (1),** 274; **5 (2),** 494; **5 (3),** 1776; **5 (4),** 1834.
251. *Beilstein,* **5,** 515; **5 (1),** 248; **5 (2),** 410; **5 (3),** 1355; **5 (4),** 1532.
252. *Beilstein,* **5,** 625; **5 (1),** 300; **5 (2),** 531; **5 (3),** 1936; **5 (4),** 2142.
253. *Beilstein,* **5,** 685; **5 (1),** 344; **5 (2),** 609; **5 (3),** 2276; **5 (4),** 2463.
254. *Beilstein,* **5,** 667; **5 (1),** 327; **5 (2),** 579; **5 (3),** 2136; **5 (4),** 2293.
255. *Beilstein,* **5,** 693; **5 (1),** 343; **5 (2),** 609; **5 (3),** 2279; **5 (4),** 2467.

Specific References

256. Rütgerswerke, DE 3 203 818, 1982 (K. Stolzenberg, K. H. Koch, R. Marrett).
257. Nippon Steel, JP 82 98 219, 1980.
258. IG Farbenindustrie, DE 428 088, 1921.
259. Selden, GB 318 617, 1929.
260. Standard Oil, US 2 966 513, 1957 (E. K. Fields).
261. Rütgerswerke, DE 1 262 268, 1959 (E. Marx, H.-G. Franck, E. A. Dierk, K. Fiebach).
262. *Ullmann,* 4th ed., **18,** 685.
263. Farbwerke Hoechst, DE 2 209 692, 1972 (K. Eiglmeier); DE 2 343 964, 1973 (F. Röhrscheid).
264. Institute of Chemical Sciences, Academy of Sciences/Kazakh S.S.R., SU 1 084 275, 1982 (R. U. Umarova, A. K. Ergalieva, D. K. Sembaev, B. V. Suborov)
265. BASF, DE 2 721 640, 1977 (A. Hettche, M. Patsch); DE 2 653 346, 1976 (W. Eisfeld, W. Disteldorf, A. Hettche).
266. BASF, DE 2 721 639, 1977 (A. Hettche, M. Patsch).
267. *Ullmann,* 4th ed., **18,** 693.
268. Farbwerke Hoechst, DE 2 304 873, 1973 (K. Eiglmeier, H. Lübbers).
269. Toyo Soda Mfg., JP 85 94 928, JP 85 139 630, 1983 (M. Kubo, H. Satsuka, Y. Tsutsumi); EP 217 316, 1986 (M. Kubo, K. Kawabata, Y. Tsutsumi).
270. Coal Industry, EP 47 069, 1981 (J. G. Robinson, S. A. Brain).
271. Nippon Steel, JP 83 176 210, 1982.
272. Institute of Heteroorganic Compounds, Academy of Sciences/USSR, DE 3 526 155, 1985 (V. V. Korshak et al.).

273. Sun Research and Development, DE 1593 398, 1966 (A. Schneider, E. J. Janoski, R. N. McGinnis).
274. G. Collin, M. Zander, *Erdöl Kohle Erdgas Petrochem.* **33** (1980) 560.
275. BASF, DE 3 044 111, 1980 (H. Naarmann, K. Penzien, D. Nägele, J. Schlag).
276. *Ullmann,* 4th ed., **21,** 70.
277. R. Heering, P. Lobitz, K. Gehrke, *Plaste Kautsch.* **31** (1984) 412 – 413; J. Stein, G. Liebisch, S. Helling, *Chem. Tech. (Leipzig)* **31** (1979) 302 – 305.
278. Gulf Research and Development, US 4 280 881, 1977 (J. C. Montagna, R. D. Galli).
279. Dow Chemical, US 4 093 672, 1977 (Y. Ch. Sun).
280. Rütgerswerke, DE 2 443 114, 1974 (K. Rühl, G. Storch, H. Steinke, B. Charpey).
281. American Cyanamid, US 4 145 409, 1976 (N. R. Pasarela).
282. Manufacture de Produits Pharmaceutique A. Christiaens, DE 2 060 721, 1970 (P. Vanhoff, P. Clarebout).
283. Takeda Chemical Industries, JP 82 123 157, 1981.
284. Rütgerswerke, DE 2 208 253, 1972 (E. Pastorek, H. Miels, W. Orth).
285. Pfizer, DE 2 244 556, 1973 (D. P. O'Shea, R. C. Adams, S. Nakanishi).
286. Meiji Seika Kaisha, DE 2 911 492, 1979 (Y. Sekizawa et al).
287. Wella, DE 3 215 341, 1982 (E. Konrad, H. Mager).
288. Rütgerswerke, DE 2 635 401, 1976 (P. Stäglich, D. Hausigk).
289. H. J. V. Winkler: *Der Steinkohlenteer und seine Bedeutung,* Verlag Glückauf, Essen 1951, p. 156.
290. Nippon Steel, JP 82 98 220, 1980.
291. Rütgerswerke, DE 3 243 981, 1982 (W. Orth, E. Pastorek, W. Fickert).
292. Celamerck, DE 2 816 388, 1979 (J. Gante et al.).
293. Squibb, US 3 856 977, 1974 (E. T. Stiller, S. D. Levine, P. A. Principe, P. A. Diassi).
294. Bayer, DE 3 530 819, 1985 (J. Hocker, L. Rottmeier).
295. *Ullmann,* 4th ed., **23,** 415.
296. Nippon Kayaku, DE 2 325 302, 1973 (D. Matsunaga, M. Sumitani).
297. Chemische Fabrik Kalk, DE 2 725 857, 1977 (H. Jenkner).
298. Sandoz, DE 2 225 565, 1972 (H. P. Baumann).
299. Rütgerswerke, DE 2 704 648, 1977 (C. Finger).
300. Agency of Industrial Sciences and Technology, JP 81 82 430, 1979.
301. Magyar Tudomanyos Akademia, HU 36 764, 1983 (L. Simandi et al.).
302. Nippon Shokubai Kagaku Kogyo, JP 85 233 028, 1984 (T. Yoneda, S. Nakahara).
303. Dow Chemical, US 4 297 514, 1980 (W. M. King).
304. Hydrocarbon Research, DE 3 141 645, 1981. (D. Huibers, R. Barclay, R. C. Shah).
305. Fuji-Xerox, JP 82 144 567, 1981; Ricoh, JP 79 109 438, 1978 (E. Akutsu, A. Kojima); Oce-van der Grinten, DE 2 525 381, 1975 (J. Hagendorn, H. Knibbe); EP 24 762, 1981 (M. T. J. Peters, J. A. Verbunt).
306. Bayer, DE 2 430 476, 1974 (W. Draber, K. H. Büchel, M. Plempel).
307. Röhm, DE 3 439 484, 1984 (A. Riemann, W. Ude).
308. Isovolta, EP 41 496, 1981 (W. Rieder, M. Fehrle).
309. Minnesota Mining and Mfg., EP 203 828, 1986 (W. J. Schultz, G. B. Portelli, J. P. Tane).
310. BASF, DE 3 209 426, 1982 (M. Patsch, H. Flohr).
311. BASF, DE 3 536 259, 1985 (M. Patsch et al.).
312. Rütgerswerke, DE 2 114 219, 1971 (C. Finger, H.-G. Franck, M. Zander).
313. Rütgerswerke, DE 2 166 862, 1971 (C. Finger, H.-G. Franck, M. Zander).
314. Kureha Chemical Industry, JP 75 140 535, 1974 (M. Takahashi, A. Ito, A. Nobuo).
315. Kureha Chemical Industry, DE 2 447 894, 1973 (Y. Shisakaiso, A. Ito, M. Takahashi).
316. Kureha Chemical Industry, DE 2 446 591, 1974 (M. Takahashi et al.).
317. Rütgerswerke, DE 1 931 946, 1969 (C. Finger, H.-G. Franck, M. Zander).
318. Wellcome Foundation, EP 125 702, 1984 (K. W. Bair).
319. Richardson Merell, US 3 907 791, 1975 (W. L. Albrecht, R. W. Fleming).
320. Rütgerswerke, DE 1 126 372, 1958 (J. Altpeter, H. Sauer).
321. BASF, DE 3 218 229, 1982 (H. Naarmann, J. Schlag, G. Koehler).
322. VEB Kombinat Mikroelektronik, DE 3 131 625, 1981 (H.-J. Deutscher et al.).
323. Fuji Standard Research, EP 196 169, 1986 (S. Otani, Y. Nagai).
324. Coal Industry, EP 36 710, 1981 (J. G. Robinson, D. I. Barnes).
325. Hercules, US 3 926 897, 1972 (C. L. Cessna).
326. Licentia, DE 2 406 540, 1974 (M. Saure, H. Wagner, J. Wartusch).

327. Union Carbide, US 3 834 166, 1974 (R. A. Cupper, G. S. Somckh).
328. Bayer, DE 3 232 063, 1982 (H. Herzog, W. Hohmann).
329. Burroughs Wellcome, US 4 511 582, 1983 (K. W. Bair).
330. Rütgerswerke, DE 1 240 065, 1965 (E. Egerer, W. Fickert, L. Rappen, G. Grigoleit); DE 3 305 528, 1983 (D. Tonnius, W. Orth, H. Miele, E. Pastorek).
331. US Steel, US 2 956 065, 1958 (C. W. De Walt, K. A. Schowalter).
332. Mitsubishi Metal, JP 60 238 432, 1984 (H. Yoshida, M. Morikawa, T. Iwamura).
333. Nihon Joryu Kogyo, Nippon Shokubai Kagaku Kogyo, JP 85 255 747, 1984 (Y. Saito et al.).
334. Nippon Shokubai Kagaku Kogyo, JP 85 233 029, 1984 (I. Maeda, H. Mitsui, T. Nakamura).
335. Nippon Shokubai Kagaku Kogyo, JP 8 112 335, 1979.
336. Merck, US 3 506 434, 1970 (E. Jacobi et al.).
337. *Ullmann,* 4th ed. **12,** 614.
338. Agfa-Gevaert, EP 2 546, 1979 (F. C. Heugebaert, W. Janssens).
339. Fuji Photo Film, GB 2 009 955, 1979 (M. Kitajima, H. Tachikawa, F. Shinozaki, T. Ikeda).
340. PPG Industries, US 4 054 683, 1976 (G. W. Gruber).
341. National Starch and Chemical, US 4 069 123, 1972 (M. M. Skoultchi, I. J. Davis).
342. Canadian Industries, US 4 012 280, 1976 (H. H. Holton).
343. Ciba-Geigy, CH 580 673, 1972 (W. Müller).
344. Eli Lilly, DE 2 631 888, 1977 (J. B. Campbell).
345. EGIS, DE 3 618 715, 1986 (P. Benko et al.).
346. H. J. V. Winkler: *Der Steinkohlenteer und seine Bedeutung,* Verlag Glückauf, Essen 1951, p. 220.
347. Rütgerswerke, DE 2 914 670, 1979 (G.-P. Blümer, M. Zander).
348. Hoechst, DE 3 532 882, 1985 (S. Schießler, E. Spietschka).
349. Rütgerswerke, DE 2 709 015, 1977 (J. Omran, M. Zander).
350. Hoechst, DE 3 134 131, 1981; DE 3 324 881, 1983 (S. Schießler, E. Spietschka).
351. *Ullmann,* 4th ed., **17,** 471 – 472.
352. Bayer, EP 20 817, 1979 (H. Harnisch).
353. Ricoh, JP 7 908 527, 1977 (A. Kojima, M. Hashimoto, M. Ohta).
354. Hoechst, DE 2 734 288, 1977 (W. Wiedemann); DE 2 726 116, 1977 (H. J. Schlosser, E. Lind).
355. British Petroleum, GB 1 399 814, 1971 (A. J. Burn, T. A. Lighting).
356. Wellcome Foundation, EP 125 701, 1984 (K. W. Bair).
357. Pentel, JP 8 632 352, 1976 (K. Takemura).
358. W. R. von Oettingen: "Toxicity and Potential Dangers of Aliphatic and Aromatic Hydrocarbons," *Public Health Bull.* (1940) no. 255.
359. E. A. Wahrenbrock, E. I. Eger, R. G. Laravuso, G. Maruschak, *Anesthesiology* **40** (1974) no. 1, 19.
360. A. H. Nuckolls, *Underwriters Laboratory Report* (Nov. 13, 1933) no. 2375.
361. Y. Henderson, H. W. Haggard: *Noxious Gases,* 2nd ed., Reinhold Publ. Co., New York 1943.
362. J. C. Krantz, Jr., C. J. Carr, J. F. Vitcha, *J. Pharmacol. Exp. Therap.* **94** (1948) 315.
363. T. Ikoma, *Nichidai Igaku Zasshi* **31** (1972) no. 2, 71.
364. D. M. Aviado, S. Zakhari, T. Watanabe: *Non-fluorinated Propellants and Solvents for Aerosols,* CRC Press, Cleveland, Ohio 1977, pp. 49 – 81.
365. F. A. Patty, W. P. Yant: "Odor Intensity and Symptoms Produced by Commercial Propane, Butane, Pentane, Hexane and Heptane Vapor," U.S. Bur. Mines, Rep. Invest. **1929** no. 2979.
366. B. B. Shugaev, *Arch. Environ. Health* **18** (1969) 878.
367. *Documentation of the Threshold Limit Values for Substances in Workroom Air,* 3rd ed., American Conference of Governmental Industrial Hygiene, Cincinnati 1971.
368. M. G. Rumsby, J. B. Finean, *J. Neurochem.* **13** (1966) 1513.
369. D. A. Haydon, B. M. Hendry, S. R. Levinson, J. Requena, *Biochim. Biophys. Acta* **470** (1977) 17.
370. N. Battistini et al., *Riv. Patol. Nerv. Ment.* **95** (1974) 871.
371. G. W. Paulson, G. W. Waylonis, *Arch. Intern. Med.* **136** (1976) 880.
372. N. W. Lazarew, *Arch. Exp. Pathol. Pharmakol.* **143** (1929) 223.
373. E. T. Kimura, D. M. Elery, P. W. Dodge, *Toxicol. Appl. Pharmacol.* **19** (1971) 699.
374. C. H. Hine, H. H. Zuidema, *Ind. Med.* **39** (1970) no. 5, 215.
375. H. H. Schaumburg, P. S. Spencer, *Brain* **99** (1976) 182.
376. K. Kurita, *Jpn. J. Ind. Health Exp.* **9** (1974) 672.
377. A. K. Ashbury, S. L. Nielsen, R. Telfer, *J. Neuropathol. Exp. Neurol.* **33** (1974) 191.

378. E. G. Gonzales, J. A. Downey, *Arch. Phys. Med. Rehabil.* **53** (1972) 333.
379. H. E. Swann, Jr., B. K. Kwon, G. H. Hogan, W. M. Snellings, *Am. Ind. Hyg. Assoc. J.* **35** (1974) no. 9, 311.
380. W. R. F. Notten, P. W. Henderson, *Biochem. Pharm.* **24** (1975) 1093.
381. C. F. Reinhardt et al., *Arch. Environ. Health* **22** (1971) 265.
382. A. Y. H. Lu, H. W. Strobel, M. J. Coon, *Mol. Pharamcol.* **6** (1970) 213.
383. N. W. Lazarew, *Arch. Exp. Pathol. Pharmakol.* **143** (1929) 223.
384. *Criteria for a Recommended Standard – Occupational Exposure to Xylene,* U.S. Dept. Health, Education and Welfare, Natl. Inst. Occup. Safety and Health, Washington, D.C. (1975).
385. O. A. Anunziata et al., *Appl. Catal.* **15** (1985) 235 – 245. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., Van Nostrand Reinhold, New York 1983.
386. W. E. Coleman et al., *Arch. Environ. Contam. Toxicol.* **13** (1984) 171 – 178.
387. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., Van Nostrand Reinhold, New York 1984.
388. *Threshold Limit Values and Biological Exposure Indices for 1987 – 1988,* American Conference of Governmental Industrial Hygienists. Cincinnati, Ohio, 1987.
389. *DFG, MAK- und BAT-Werte Liste 2001, Mitteilung 37,* Wiley-VCH, Weinheim, 2001.
390. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons,* Elsevier Publishing Co., London 1960.
391. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin Heidelberg 1987.
392. H. W. Gerarde, *Am. Med. Assoc. Arch. Ind. Health* **19** (1959) 403.
393. M. A. Wolf et al., *Am. Med. Assoc. Arch. Ind. Health* **14** (1956) 387.
394. H. W. Werner, R. C. Dunn, W. F. von Oettingen, *Ind. Health Toxicol.* **26** (1974) 264.
395. Maximale Arbeitsplatzkonzentrationen und biologische Arbeitsstofftoleranzwerte, Dt. Forschungsgemeinschaft, Verlag Chemie, Weinheim 1984.
396. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., van Nostrand Reinhold. New York 1984, p. 819.
397. H. E. Christensen, T. T. Luginbyhl: *Registry of Toxic Effects of Chemical Substances,* U.S. Dept. of H. E. W., U.S. Government Printing Office, Washington, D.C., 1975.
398. J. Opdyke, *Food Cosmet. Toxicol.* **12** (1974) 707.
399. J. M. Haas, H. W. Earhart, A. S. Todd: *Book of Papers,* 1974 AATCC National Technical Conference, American Association of Textile Chemists and Colorists, Research Triangle Park, N. C., p. 427.
400. I. Hakkinen et al., *Arch. Environ. Health* **26** (1973) 70.
401. A. M. Seppalainen, I. Hakkinen, *J. Neurol. Neurosurg. Psychiatry* **38** (1975) 248.
402. *Patty's Industrial Hygiene and Toxicology,* vol. 2 B, Toxicology, 3rd ed.
403. Z. F. Khromenko, *Gig. Tr. Prof. Zabol.* 1976, 44.
404. Y. R. Adamson, J. L. Weeks, Arch. Environ. Health **27** (1973) no. 8, 68.
405. Y. R. Adamson, J. M. Furlong, *Arch. Environ. Health* **28** (1974) no. 3, 155.
406. R. J. Hawkins, J. L. Weeks, *At. Energy Can. Ltd. (Rep.)* **4439** (1973); *Chem. Abstr.* **80** (1974) 65 963.
407. J. S. Henderson, J. L. Weeks, *Ind. Med. Surg.* **40** (1973) 10.
408. International Agency for Research on Cancer (ed.): *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* vol. 32, IARC, Lyon 1983.
409. *Phenanthrene, Dangerous Prop. Ind. Mat. Rep.* **6** (1986) 68 – 69.
410. E. E. Sandmeyer in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 B,** J. Wiley & Sons, New York 1981, p. 3353.
411. Y. S. Rotenberg, F. D. Mashbits, *Gig. Tr. Prof. Zabol.* **9** (1965) 53 – 54.
412. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons,* Elsevier Publishing, Amsterdam – London – New York 1960.
413. S. D. Keimig, R. W. Pattillo, Brookhaven Natl. Lab., Rep. BNL-51885, Upton 1984.
414. Z. Matijasevic, E. Zeiger, *Mutat. Res.* **142** (1985) 149 – 152.
415. M. Sakai, D. Yoshida, S. Mizusaki, *Mutat. Res.* **156** (1985) 61 – 63.
416. J. E. Rice, T. J. Hosted, E. J. Lavoie, *Canc. Lett. (Shannon Irel.)* **24** (1984) 327 – 333.
417. F. Palitti et al., *Mutat. Res.* **174** (1986) 125 – 130.
418. J. R. Babson et al., *Toxicol. Appl. Pharmacol.* **85** (1986) 355 – 366.
419. R. P. Bos et al., *Mutat. Res.* **187** (1987) 119 – 125.

420. A. L. Reshetyuk, E. J. Talakina, P. A. En'yakova, *Gig. Tr. Prof. Zabol.* **14** (1970) 46 – 47.
421. K. Knobloch, S. Szendzikowski, A. Slusarczyk-Zalobna, *Med. Pr.* **20** (1969) 210 – 222.
422. D. T. Gibson, V. Subramanian in D. T. Gibson (ed.): *Microbial Degradation of Organic Compounds,* Marcel Dekker, New York – Basel 1984, pp. 181 – 252.
423. H. H. Tabak, S. A. Quave, C. J. Mashni, E. F. Barth, *J. Water Pollut. Control Fed.* **53** (1981) 1503 –1518.
424. W. S. Hall, T. J. Leslie, K. L. Dickson, *Environ. Contam. Toxicol.* **36** (1986) 286 – 293.
425. G. W. Holcombe, G. L. Phipps, J. T. Fiandt, *Ecotoxicol. Environ. Saf.* **7** (1983) 400 – 409.
426. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., Van Nostrand, New York – London 1983.
427. B. R. Parkhurst in R. A. Conway (ed.): *Environmental Risk Analysis for Chemicals,* Van Nostrand, New York – London,pp. 399 – 411.
428. M. Sittig: *Priority Toxic Pollutants,* Noyes Data Corp., Park Ridge, N.J., 1980.

Hydrochloric Acid

SEVERIN AUSTIN, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany

ARNDT GLOWACKI, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 4
2.3. The HCl–H_2O System 5
3. Production 6
3.1. Methods of Preparation 7
3.2. Industrial Production 7
3.2.1. From Hydrogen and Chlorine 7
3.2.2. Reaction of Metal Chlorides with Sulfuric Acid 8
3.2.3. Recovery as a Chlorination Byproduct 9
3.2.4. Recovery from Waste Incineration . . 11
3.3. Purification 12
4. Quality Specifications 12
5. Storage and Transportation 13
6. Uses and Economic Aspects 13
7. Toxicology and Occupational Health 14
8. References 15

1. Introduction

Hydrochloric acid, also known as muriatic acid, is a solution of hydrogen chloride (HCl) [*7647-01-0*] in water. Both hydrochloric acid and hydrogen chloride are discussed in this article.

History. In the 15th century, the German alchemist VALENTIN heated so-called green vitriol [iron (II) sulfate, $FeSO_4 \cdot 7\,H_2O$] with common salt. He thereby obtained what then was called spirit of salt. In the 17th century, GLAUBER prepared hydrochloric acid from common salt and sulfuric acid. In 1790, DAVY established the composition of hydrogen chloride by synthesizing it from hydrogen and chlorine. In the same year, LEBLANC discovered the process named after him for the production of soda, the first stage of which is to react common salt with sulfuric acid, liberating hydrogen chloride. This was at first considered an undesirable byproduct and simply released into the atmosphere in large amounts. In 1863, English soda producers were compelled by the Alkali Act to absorb the hydrogen chloride in water, which quickly led to large-scale industrial use of the acid produced. Excess hydrogen chloride that could not be used as hydrochloric acid was oxidized to chlorine.

The industrial synthesis of hydrogen chloride followed the development of the chlor-alkali electrolytic process early in the 20th century. This route was used in addition to that based on the reaction between chlorides and sulfuric acid or sodium hydrogen sulfate and resulted in a product of higher purity.

These processes are becoming less important because of the large amounts of hydrogen chloride arising as byproducts of chlorination processes such as the production of vinyl chloride from ethylene. Hydrochloric acid may also be recovered from gases produced by the incineration of chlorine-containing waste, a consequence of the ever-increasing emphasis on environmental protection.

2. Properties

2.1. Physical Properties

Under standard conditions, hydrogen chloride, HCl, M_r 36.461, is a colorless gas with a pungent odor and an extremely mordant effect on the mucous membranes of the respiratory system. Some physical properties of hydrogen chloride are given below.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_283

mp	−114.4 °C (158.8 K)
bp (101.3 kPa)	−85.0 °C (188.1 K)
Vapor density ϱ (0 °C, 101.3 kPa)	1.639 g/L
Liquid density ϱ (−85.0 °C)	1.187 g/cm^3
Critical temperature T_{crit}	51.6 °C (324.7 K)
Critical pressure P_{crit}	8.3 MPa (83.0 bar)
Critical density ϱ_{crit}	0.41 g/cm^3
Specific heat c_p [1]	
100 K	40 J mol^{-1} K^{-1}
153 K	49 J mol^{-1} K^{-1}
Thermal expansion coefficient	
100 K	2.5×10^{-4} K^{-1}
153 K	5.7×10^{-4} K^{-1}
Isothermal compressibility κ_T [2]	
100 K	2×10^{-5} bar^{-1}
153 K	1×10^{-4} bar^{-1}
Enthalpy of fusion ΔH_{fus}	1970 J/mol 54.0 J/g
Entropy of fusion ΔS_{fus}	12.54 J mol^{-1} K^{-1} 0.344 J g^{-1} K^{-1}
Enthalpy of vaporization ΔH_{vap} (−85.0 °C)	16.150 kJ/mol 0.443 kJ/g

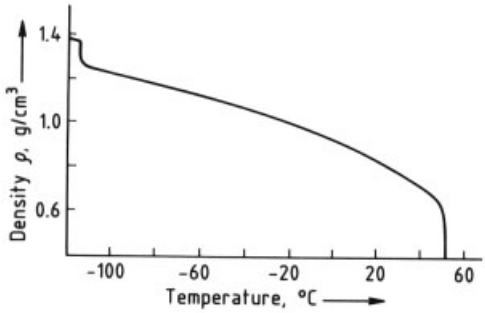

Figure 1. Density ϱ of liquid hydrogen chloride

Solid hydrogen chloride undergoes a polymorphic transition at 98 K from an orthorhombic face-centered low-temperature phase II to a cubic face-centered high-temperature phase I: enthalpy of transformation $\Delta H = 1190$ J/mol, entropy of transformation $\Delta S = 12.1$ J mol^{-1} K^{-1} [3]. The vapor pressure of solid hydrogen chloride, in millibars, is given [1] by

$$\log p = -911.31T^{-1} + 7.1875 + 4.313 \times 10^{-3}T$$

On melting the volume increases by 14 %.

The density, specific heat, viscosity and surface tension, thermal conductivity, and enthalpy of vaporization of *liquid hydrogen chloride*, each as a function of temperature, are shown in Figures 1, 2, 3, 4, 5, respectively. The vapor pressure of liquid hydrogen chloride, in millibars, is given by

$$\log p = -8.555 \times 10^2 T^{-1} + 7.553$$

which is fairly accurate over the entire liquid range.

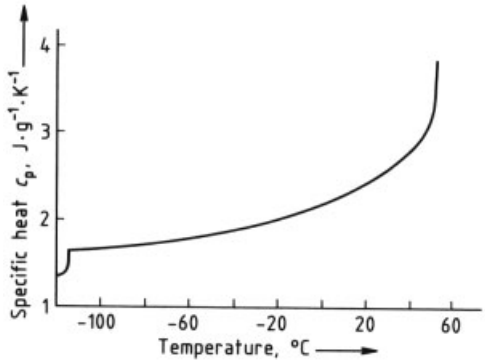

Figure 2. Specific heat c_p of liquid hydrogen chloride

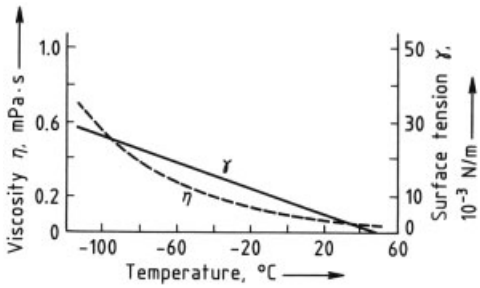

Figure 3. Surface tension (——) and dynamic viscosity (- - - -) of liquid hydrogen chloride

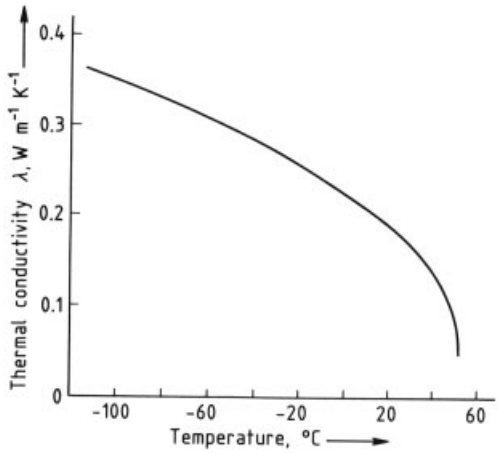

Figure 4. Thermal conductivity λ of liquid hydrogen chloride

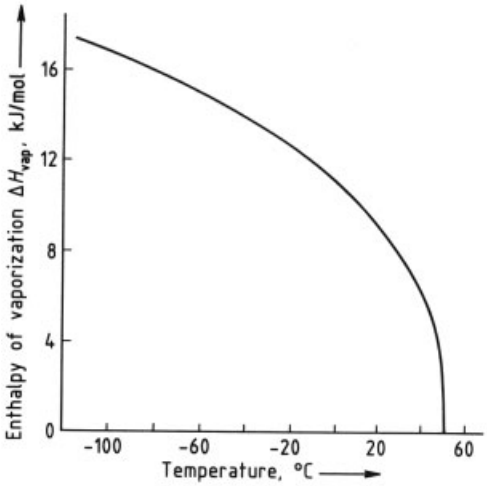

Figure 5. Enthalpy of vaporization ΔH_{vap} of hydrogen chloride

Table 1. Solubility of hydrogen chloride in various organic liquids at an HCl partial pressure of 101.3 kPa (1.013 bar) ∗

Solvent	Temperature, °C	Dissolved HCl, g/kg solvent
n-Hexane		7.1
n-Hexadecane	27	3.9
	102	2.2
	177	1.6
Benzene	20	20
	40	12.5
Toluene	20	21.3
Methanol	0	1092
	20	877
	40	688
Ethanol	0	838
	10	756
	20	681
	30	610
Diisopropyl ether	10	349
Dibutyl ether	10	250
Dioxane	10	433
Tetrahydrofuran	10	584
Chloroform	15	8.5
	25	6.9
Carbon tetrachloride	20	6.0
	50	3.9

∗ Hydrogen chloride forms addition compounds with many organic compounds; in such cases, its solubility often either is unknown or cannot be defined exactly.

The vapor density of *gaseous hydrogen chloride* shows that it consists of single HCl molecules. The specific heat, viscosity, and thermal conductivity of *gaseous hydrogen chloride,* each as a function of temperature, are shown in Figures 678, respectively [4]. The solubility of hydrogen chloride in various organic solvents is listed in Table 1 [5–7].

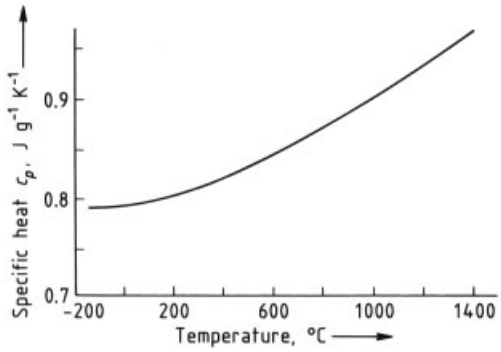

Figure 6. Specific heat c_p of gaseous hydrogen chloride

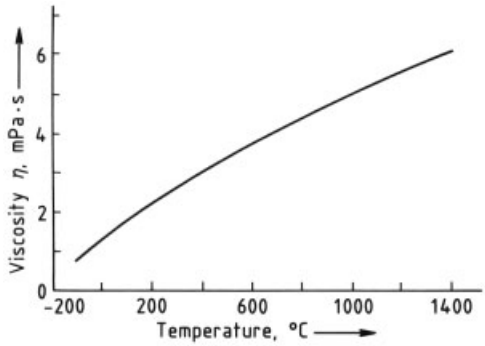

Figure 7. Viscosity η of gaseous hydrogen chloride at 101.3 kPa

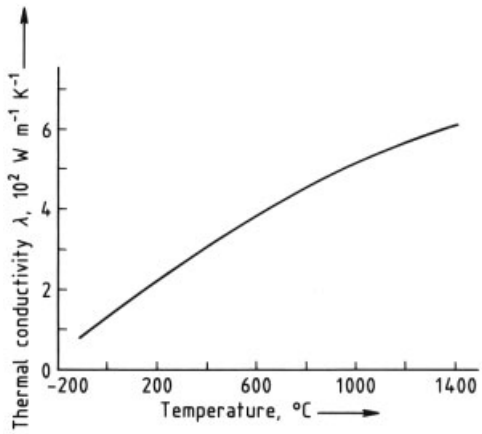

Figure 8. Thermal conductivity λ of gaseous hydrogen chloride at 101.3 kPa

The most important thermodynamic data are as follows [8]:

Enthalpy of formation	
ΔH_f^0 (298.15 K)	−92.31 kJ/mol
Entropy S^0 (298.15 K)	186.78 J K^{-1} mol^{-1}
Enthalpy H^0 (298.15 K) − H^0 (0)	8.64 kJ/mol
Specific heat c_p (298.15 K)	29.12 J mol^{-1} K^{-1}
Free energy of formation	
ΔG^0 (293.15 K)	−95.25 kJ/mol

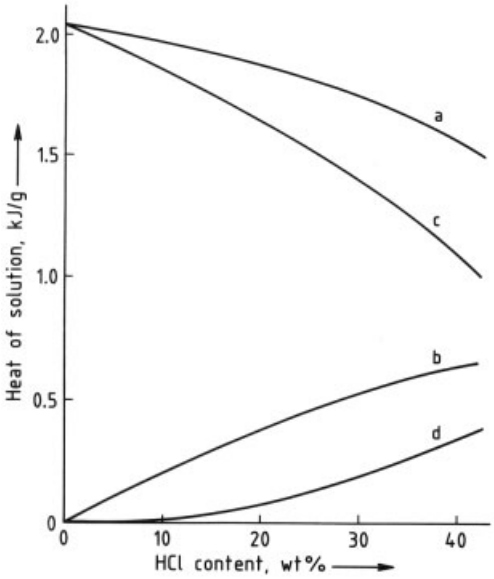

Figure 9. Heats of solution (25 °C) in the HCl–H_2O system a) Total enthalpy of solution $\Delta_{sol} H_{HCl}$, in kilojoules per gram of hydrogen chloride; b) Total enthalpy of solution in kilojoules per gram of solution; c) Partial molar enthalpy of solution $\overline{\Delta_{sol} H_{HCl}}$; d) Enthalpy of dilution $\overline{\Delta_{dil} H_{HCl}}$

2.2. Chemical Properties

Hydrogen chloride is thermally stable up to approximately 1500 °C, but appreciable dissociation occurs above this temperature. Completely dry hydrogen chloride is not very reactive: mild steel is not measurably attacked. Moreover, reactions involving dry hydrogen chloride often take place only in the presence of catalysts. In contrast, a solution of hydrogen chloride in a polar solvent is a strong acid and, therefore, an aggressive reagent.

Hydrogen chloride dissolves readily in water, liberating much heat. The enthalpy of solution, (i.e., the total heat of solution $\Delta_{sol} H_{HCl}$), the partial molar enthalpy of solution

$$\overline{\Delta_{sol} H_{HCl}} = \left(\frac{\partial \Delta_{sol} H_{HCl}}{\partial n_{HCl}} \right)_{p,T,n_{H_2O}}$$

and the partial molar enthalpy of dilution

$$\overline{\Delta_{dil} H_{HCl}} = \left(\frac{\partial \Delta_{sol} H_{HCl}}{\partial n_{H_2O}} \right)_{p,T,n_{HCl}}$$

are shown in Figure 9 [9]. The hydrochloric acid formed is a strong and extremely aggressive acid, attacking almost all industrially available metals and alloys to some extent. Metals must, therefore, be protected by an appropriate nonmetallic lining, or nonmetallic materials of construction must be used. Tantalum or highly resistant nickel alloys such as Hastelloy B may be used under precisely defined conditions. Less noble metals dissolve with liberation of hydrogen.

2.3. The HCl–H_2O System

In the HCl–H_2O system at low temperature, four congruently melting solid phases are formed. The corresponding compositions are given by the following formulas: $HCl \cdot H_2O$ (*mp* −15.4 °C), $HCl \cdot 2\,H_2O$ (*mp* −18 °C), $HCl \cdot 3\,H_2O$ (*mp* −25 °C), and $HCl \cdot 6\,H_2O$ (*mp* −70 °C). At higher temperature an important physical property is the solubility of gaseous hydrogen chloride in water and the partial pressure of hydrogen chloride and water over hydrochloric acid at various concentrations and temperatures.

These relationships are shown in Figure 10; they have great practical significance for the absorption and desorption of hydrogen chloride. At lower temperature (below about 100 °C) and pressure (less than about 0.2 MPa), the curves are based on data given in [10], [11]. After careful verification of the original experimental data, these remain the best values available; they are consistent with the partial pressures of hydrogen chloride and water derived from e.m.f. measurements [12]. Values for the high-pressure regions have been determined more recently [13–16] and these can be fitted smoothly to the older values.

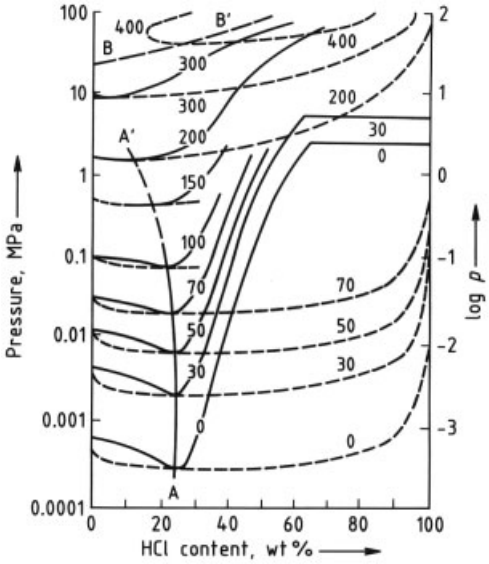

Figure 10. Vapor–liquid phase diagram in the $HCl–H_2O$ system
Temperature (°C) is shown for each isotherm. The two-phase region is delimited from the gas phase by the broken lines (- - - -) and from the liquid phase by the continuous lines (——). Curve A–A′ connects the azeotropic points; B–B′ is the critical segregation curve above the critical point of water (point B: 374.1 °C, 21.43 MPa)

The system has a vapor pressure minimum (azeotrope) whose position shifts toward lower hydrogen chloride concentration with increasing pressure and temperature. At 101.3 kPa, the azeotrope, *bp* 108.6 °C, contains 20.4 wt % hydrogen chloride. When the hydrogen chloride concentration is lower than that, the vapor has a higher water content than the liquid with which it is in equilibrium. In this case it is not possible to produce pure anhydrous hydrogen chloride by simple distillation. If the concentration of the acid exceeds that of the azeotropic mixture, however, dry hydrogen chloride can be isolated by distillation. The expense of distillation decreases with increasing hydrogen chloride concentration of the acid.

The hydrogen chloride gas phase coexisting with concentrated hydrochloric acid of more than 35 wt % has a very low water content. A corollary of this is that the moisture content of hydrogen chloride gas must be kept very low to prevent condensation of aqueous hydrochloric acid, which would cause corrosion problems in steel pipes and equipment. At higher pressure and lower temperature, the permissible moisture content must be even lower. Limiting concentrations are given in Table 2. For example, hydrogen chloride gas in steel pipes at 0.5 MPa and 30 °C may contain up to 800 ppm (by volume) of water, but at 0 °C the tolerable water content is < 100 ppm.

At higher pressure and below the critical temperature of hydrogen chloride, the $HCl–H_2O$ system has a miscibility gap, i.e., highly concentrated hydrochloric acid coexists with liquid hydrogen chloride. For example, at 0 °C and > 2.4 MPa, 64 % hydrochloric acid is in equilibrium with anhydrous liquid hydrogen chloride. At lower pressure, hydrogen chloride becomes a gas, and the equilibrium concentration of hydrochloric acid decreases with decreasing pressure. With increasing temperature, the width of the miscibility gap increases, and above the critical temperature of hydrogen chloride, the liquid hydrogen chloride phase is no longer present. Hydrochloric acid remains as the only homogeneous liquid phase, whose maximum concentration decreases as the temperature rises: the heterogeneous region simultaneously becomes wider. As soon as the critical temperature of water is exceeded, the liquid phase disappears. Nevertheless, initially a heterogeneous region exists that becomes smaller with increasing temperature. This region consists of two immiscible gas phases in equilibrium: at 50 MPa and 400 °C, gaseous water containing approximately 13 % hydrogen chloride is in equilibrium with hydrogen chloride containing approximately 35 % water.

The *density* of aqueous hydrochloric acid for the temperature range 0 – 100 °C is given in Table 3.

The *specific heat* c_p decreases with increasing concentration. Some values (at 20 °C) follow:

HCl content, wt %	16	20	24	28	32	36	40
c_p, $J\,g^{-1}\,K^{-1}$	3.11	2.93	2.78	2.65	2.55	2.47	2.40

The temperature dependence of c_p for the acid is considerably greater than that for water. In the above concentration range and at 0 – 100 °C, it is given by the following equation:

$$\left(\frac{\Delta c_p}{\Delta T}\right) = 4.2 \times 10^{-3}\,J\,g^{-1}\,K^{-2}$$

Table 2. Mole fraction ($\times 10^6$) of water in hydrogen chloride gas in equilibrium with concentrated (saturated) hydrochloric acid (based on the results of Kao [13])

p, MPa	Temperature, °C							
	−10	0	10	20	30	40	50	60
0.1	60	440	870	2400	4500	9200	15 500	30 000
0.2	35	230	430	720	3200	6100	10 300	19 000
0.5	18	96	180	350	800	2000	3 900	7 500
1.0	10	60	120	240	430	700	2 200	4 400

Table 3. Density of hydrochloric acid, in grams per cubic centimeter, as a function of temperature

HCl concentration, wt %	Temperature, °C						
	0	10	20	30	50	75	100
5	1.0266	1.0253	1.0230	1.0200	1.0120	0.9993	0.9841
10	1.0521	1.0501	1.0473	1.0440	1.0359	1.0234	1.0090
20	1.1065	1.1022	1.0979	1.0933	1.0839	1.0711	1.0575
30	1.1611	1.1551	1.1493	1.1433	1.1318	1.1175	1.1030
35	1.1875	1.1805	1.1738	1.1671	1.1540	1.1379	1.1220

The *viscosity* η is considerably higher than that of pure water. For 10 wt % acid, the viscosity varies with temperature as follows:

t, °C	0	10	20
η, mPa · s	1.89	1.45	1.16

The variation of viscosity with acid concentration is given below (20 °C):

HCl, wt %	10	20	30
η, mPa · s	1.16	1.36	1.70

The *thermal conductivity* λ is considerably lower than that of water. For 30 % hydrochloric acid, it is 4.6×10^{-5} W m^{-1} K^{-1} (30 °C), compared with 6.18×10^{-5} W m^{-1} K^{-1} for water.

3. Production

Free hydrogen chloride occurs naturally only in low concentration; it has been detected in the stratosphere at levels of ca. 10^{-11} vol % [17]. Inorganic chlorides, however, are widely distributed, chlorine being the eleventh chemical element in order of abundance. Alkali-metal and alkaline-earth chlorides, in particular rock salt, are the primary raw materials for all routes to hydrogen chloride or hydrochloric acid production including those involving intermediate products such as chlorine or chlorinated organic compounds.

In nature, hydrogen chloride is liberated in volcanic eruptions or other releases of subterranean gas. The amount of hydrogen chloride produced can be of the same order as that of sulfur dioxide. Hydrochloric acid occurs in the gastric juice of mammals; in the human stomach, the concentration of acid is about 0.1 mol/L.

3.1. Methods of Preparation

Hydrogen chloride can be formed according to the following reactions:

1) Synthesis from elements:

$$H_2 + Cl_2 \longrightarrow 2\,HCl \quad (1)$$

2) Reaction of metallic chlorides, particularly sodium chloride, with sulfuric acid or a hydrogen sulfate:

$$NaCl + H_2SO_4 \longrightarrow NaHSO_4 + HCl \quad (2)$$

$$NaCl + NaHSO_4 \longrightarrow Na_2SO_4 + HCl \quad (3)$$

$$2\,NaCl + H_2SO_4 \longrightarrow Na_2SO_4 + 2\,HCl \quad (4)$$

3) As a byproduct of chlorination, e.g., in the production of dichloromethane, trichloroethylene, perchloroethylene, or vinyl chloride:

$$C_2H_4 + Cl_2 \longrightarrow C_2H_4Cl_2 \quad (5)$$

$$C_2H_4Cl_2 \longrightarrow C_2H_3Cl + HCl \quad (6)$$

4) From spent pickle liquor in metal treatment, by thermal decomposition of the hydrated heavy-metal chlorides:

$$2\,FeCl_3 \cdot 6\,H_2O \xrightarrow{\Delta H} Fe_2O_3 + 3\,H_2O + 6\,HCl \quad (7)$$

5) From incineration of chlorinated organic waste:

$$C_4H_6Cl_2 + 5\,O_2 \longrightarrow 4\,CO_2 + 2\,H_2O + 2\,HCl \quad (8)$$

By far the greatest proportion of hydrogen chloride is now obtained as a byproduct of chlorination (method 3, Eqs. 5 and 6). The degree of purification required depends on the end use. At present, recovery from waste material according to methods 4 and 5 is becoming increasingly important, whereas the amount of hydrogen chloride generated by method 2 is decreasing. Recovery of hydrochloric acid from pickle liquor is described in [18]. Finally, hydrogen chloride can be formed in undesirable reactions such as the hydrolysis of Friedel – Crafts catalysts:

$$AlCl_3 + 3\,H_2O \longrightarrow Al(OH)_3 + 3\,HCl \quad (9)$$

3.2. Industrial Production

3.2.1. From Hydrogen and Chlorine

The simplest method for producing hydrogen chloride is direct synthesis from the elements, which yields a very pure product. The reaction is strongly exothermic; the standard enthalpy of reaction of Equation (1) is $\Delta H_R^0 = -184$ kJ. Mixtures of hydrogen and chlorine are extremely explosive over a wide range of composition, and explosion can often be initiated by light. Care must, therefore, be taken to avoid the occurrence of reactive mixtures of chlorine and hydrogen in industrial processes.

Industrial production involves the use of a burner; chlorine and hydrogen are fed by separate concentric tubes into the combustion chamber. After ignition, the chlorine burns in the hydrogen with a quiet, very hot flame at $> 2000\,°C$. Silica has proved to be a good material of construction for the burner. Cast iron or steel, sometimes water-cooled, or graphite has also been used. The burner is usually located on the bottom of a cylindrical combustion chamber, with the flame directed vertically upward. The combustion chamber is made of steel lined with refractory brick. A cooling section of appropriate size and shape is connected to the combustion chamber.

If the reactants are very moist or otherwise contaminated, corrosion-resistant construction materials such as silica or graphite must be used. Manufacturing plants still in use today, however, are constructed of carbon steel and the hydrogen chloride produced is water-cooled only to a temperature at which aqueous hydrochloric acid cannot condense. The most probable contaminants are moisture in the hydrogen and oxygen in the chlorine, but oxygen compounds such as dichlorine oxide (Cl_2O) or carbon dioxide can also cause problems by the formation of water.

The mixing ratio of the raw material is adjusted according to the intended use of the hydrogen chloride produced. A slight excess of hydrogen or chlorine may be used; it is inadvisable to work with exactly equimolar mixtures because the reaction is not quite quantitative and it is more difficult to avoid explosive mixtures. Control deviations can easily lead to alternation between excess hydrogen and excess chlorine, with explosive mixtures appearing from time to time. Hydrogen in hydrogen chloride is usually less troublesome than chlorine, so that an excess of 1 – 2 % hydrogen is usually employed. Depending on the purity of the raw materials, extremely pure chlorine-free hydrogen chloride can be obtained, with excess hydrogen being the only impurity. This hydrogen chloride can be used to produce very pure hydrochloric acid, or it may be liquefied or supplied directly to the user in steel pipes as a dry gas.

The use of steel for the synthesis furnaces and coolers is a particularly attractive possibility. They can then be operated under elevated pressure, the magnitude is determined only by the pressure of chlorine and hydrogen and the

temperature and pressure in the pipeline. Here the pressure must be low enough, or the temperature high enough, to prevent condensation of aqueous hydrochloric acid (see Fig. 10 and Table 2). The steel furnaces operated at Hüls can each produce 30 – 40 t/d hydrogen chloride. The hydrogen chloride gas is generated at a pressure up to 0.7 MPa, the water content being < 50 ppm (by volume); steel pipes are, therefore, used to convey it from the plant.

Chemically related processes [19] include the reaction of chlorine with carbon (in the form of coke) and water vapor

$$C + 2\,H_2O + 2\,Cl_2 \longrightarrow 4\,HCl + CO_2 \qquad (10)$$

or the reaction of chlorine with sulfur dioxide and water

$$SO_2 + 2\,H_2O + Cl_2 \longrightarrow 2\,HCl + H_2SO_4 \qquad (11)$$

3.2.2. Reaction of Metal Chlorides with Sulfuric Acid

The sulfate process has a high energy consumption. It is, therefore, gradually decreasing in importance compared with other processes, especially organic chlorination, which inevitably produces large amounts of hydrogen chloride as a byproduct. Moreover, the old argument that sulfate–hydrochloric acid provides a product with higher purity no longer holds; byproduct acid can now be purified sufficiently to satisfy even requirements for high quality. Hence, only the most important sulfate–hydrochloric acid process is described here. Further details are given in [20].

The reaction of common salt with concentrated sulfuric acid according to Equation (2) occurs at comparatively low temperature (150 – 300 °C); hydrogen chloride and sodium hydrogen sulfate are formed. The latter reacts with excess sodium chloride at a minimum temperature of 550 – 600 °C, according to Equation (3), forming neutral sodium sulfate. The use of sodium hydrogen sulfate as a starting material (Eq. 3) is quite feasible but has been discontinued because the high temperature (600 – 800 °C) required and the high energy consumption render the process uneconomic.

Sulfuric acid is now the exclusive starting material. In the Mannheim process, sodium sulfate is produced; in the Berlin hydrochloric acid process, sodium hydrogen sulfate.

In the *Mannheim process*, brick-lined externally fired muffle furnaces are used. These contain a stirrer with a scraping action which prevents agglomeration of the pasty mass and also removes the sodium sulfate produced. Hydrogen chloride passes through the side of the muffle and is propelled by a fan to the next stage, the furnace being maintained at a slightly reduced pressure. The mixture of gases exhausted from the furnace contains up to 85 % hydrogen chloride, as well as air, sulfuric acid mist, and fine particles of salt. As a rule, the gas must undergo several purification stages such as filtration through coke and activated carbon, wet scrubbing, or chemical reaction.

The *Berlin hydrochloric acid process* yields considerably purer hydrogen chloride gas than the Mannheim process. The reaction of common salt with sulfuric acid takes place in molten sodium hydrogen sulfate at about 300 °C in cast-iron retorts. Alternatively, potassium chloride can be used, to give hydrogen chloride and potassium hydrogen sulfate. This process has retained some commercial importance because the principal products are the potassium salts formed (i.e., $KHSO_4$ and $K_2S_2O_7$) [21], whereas hydrogen chloride is considered as a byproduct.

Gas from the retorts is more concentrated because the equipment is more leak proof, and less sulfuric acid mist is present as a result of the lower working temperature.

3.2.3. Recovery as a Chlorination Byproduct

By far the largest amounts of hydrogen chloride and hydrochloric acid are produced as byproducts of chlorination. Equation (6) is merely an example of the varied range of industrial chlorination processes. Nevertheless, the major source of hydrogen chloride is the cracking of 1,2-dichloroethane to give vinyl chloride (→ Chlorinated Hydrocarbons). The quantity of hydrogen chloride produced and consumed can be extremely large, in many installations reaching the order of 10 000 m^3/h (16 t/h or 140 000 t/a) and more. Consequently, hydrogen chloride is often fed directly to a chemical plant

which uses it as a raw material; installed storage capacities for hydrogen chloride are usually comparatively small. It is, therefore, essential to operate the plants at a balanced rate of production and consumption. If production problems or market variations affect one of the plants or products, the other plant can only continue to operate if reserve production capacity (normally not fully used) is installed. Buffer storage for liquid hydrogen chloride is used to overcome short-term inbalances in production versus consumption.

Hydrogen chloride produced by organic chlorination can be treated in various ways:

1) simple isolation of the gas by condensing the chlorinated hydrocarbons;
2) isolation and purification of hydrogen chloride by distillation;
3) aqueous absorption of hydrogen chloride.

Condensation of the chlorinated hydrocarbons is used when the presence of residual chlorinated hydrocarbons or other residues from the chlorination process does not cause any problems in the hydrogen chloride, which is fed as a gas. The degree of purification attainable depends on condensation pressure and temperature. One example is the chlorination of chloromethane to dichloromethane; the hydrogen chloride formed is used only to convert methanol to chloromethane, which in turn is used to produce dichloromethane:

$$CH_3Cl + Cl_2 \longrightarrow CH_2Cl_2 + HCl \tag{12}$$

$$CH_3OH + HCl \longrightarrow CH_3Cl + H_2O \tag{13}$$

The chloromethane and dichloromethane content of hydrogen chloride produced according to Equation (12) does not have any adverse effect on the reaction of methanol with hydrogen chloride (Eq. 13).

Fractional distillation of the liquefied gas may be employed if a purer product is required. For example, the gaseous products from the cracking of 1,2-dichloroethane are liquefied and then separated by distillation into vinyl chloride and hydrogen chloride. Liquefaction of these gases by reducing the temperature to extremely low values is uneconomical, so condensation and distillation are carried out at elevated pressure. An advantage of this procedure is that hydrogen chloride gas is produced at a high enough pressure (usually 1 – 2 MPa) for subsequent processing, thereby avoiding the necessity for subsequent mechanical compression of the gas.

Aqueous absorption of hydrogen chloride gives hydrochloric acid as an intermediate product and leads to inevitable problems with construction materials. These problems have become less severe since the introduction of fluoropolymers. In many installations, however, aqueous treatment has been replaced by nonaqueous methods (e.g., fractional distillation). The absorption of hydrogen chloride can be effected with water or 20 wt % hydrochloric acid. The use of water as the absorption medium is appropriate if the primary aim is the production of 30 – 35 wt % hydrochloric acid as the end product. If demand for the latter is insufficient, the 20 % azeotrope is used for the absorption of hydrogen chloride, thereby concentrating the acid to 30 – 35 wt %. The concentrated hydrochloric acid is first purified, if necessary. It is then distilled to give gaseous hydrogen chloride and azeotropic hydrochloric acid. The hydrogen chloride is dried and delivered to the consumer; the azeotrope is recycled to the absorption stage. If the presence of chlorinated hydrocarbons in hydrochloric acid is undesirable, they can be removed by introducing a stripping step. A simplified scheme is shown in Figure 11. A small amount of water is fed to the absorption column to render the gas stream free of hydrogen chloride. A corresponding amount of acid must be separated for other use.

A number of process and equipment parameters are important for assessing the economics of the aqueous absorption of hydrogen chloride from a production gas stream:

1) The concentration of hydrogen chloride in the gas stream determines the pressure and temperature that must be maintained in the absorption column.
2) The energy requirement for liberating pure hydrogen chloride gas decreases with increasing concentration of the concentrated acid formed.
3) Production of concentrated hydrochloric acid from a gas stream with a low hydrogen chloride concentration requires comparatively great effort because either the pressure must be increased, the temperature reduced, or both.

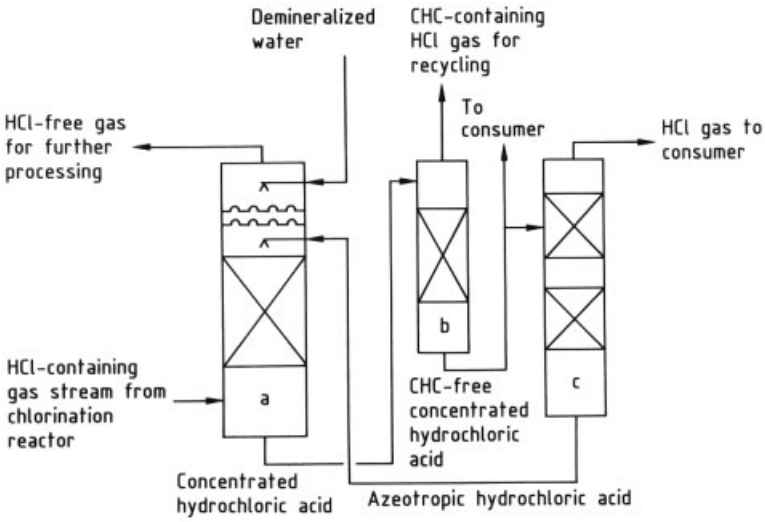

Figure 11. Flow diagram showing recovery of hydrogen chloride as a byproduct of chlorination in aqueous absorption (CHC = chlorinated hydrocarbons)
a) Hydrogen chloride absorption column; b) Chlorinated hydrocarbon stripping column; c) Hydrogen chloride desorption column

4) The presence of a high concentration of water vapor in the production gas stream can, under certain circumstances, render the production of concentrated hydrochloric acid impossible.

Industrial processes have been reviewed by SCHMIDT [10] and VIEWEG [22]. The latter proposes the use of magnesium chloride solution to absorb low levels of hydrogen chloride from gas containing a high level of water vapor.

The choice of *construction materials* for hydrogen chloride absorption equipment is restricted because of the possibility of attack both by concentrated hydrochloric acid and by chlorinated hydrocarbons. Graphite heat exchangers are used to cool the acid. For a long time a special ceramic was the only economically viable material for the pumps. The distillation column is lined with acid-proof bricks or graphite blocks, and the bricks are cemented with phenolic resin. Frequently, a special acid-resistant rubber lining is applied between the steel casing and the brick wall to prevent attack by acid that might penetrate pores and cracks. Vessels and equipment made entirely of phenolic resin are chemically very resistant, but brittle and mechanically fragile, and require support (e.g., by an outer shell of glass-fiber-reinforced polyester resin). This type of equipment is suitable only for operation at atmospheric or possibly slightly higher pressure.

The same is true for the chlorinated hydrocarbon stripping column. If the concentration of chlorinated hydrocarbons is low, or if they have been effectively removed from the acid, other materials such as poly(vinyl chloride) can be used for storage vessels. The desorption column, however, is made of graphite because of the high operating temperature. Alternatively, rubber-lined columns with inner brick linings can be used. Steel columns lined with perfluoropolymer, especially polytetrafluoroethylene (PTFE), have been used with generally excellent results, although differences exist among the products from different manufacturers with regard to microporosity, tear resistance, and flow properties. Compared with graphite columns, coated steel columns have the advantage that they can operate at elevated pressure, but failure of a PTFE lining would lead to rapid destruction of the steel casing. However, these columns can be constructed of fairly small individual elements that may be replaced quickly, reducing maintenance downtime. Heat exchangers are made of graphite in all cases. Another possibility is to use tantalum for lining and evaporators, although special precautions must be taken to avoid embrittlement due to absorption of hy-

drogen. Furthermore, hydrogen fluoride must be absent.

Completely dry hydrogen chloride will not attack steel pipes or compressors. Hydrogen chloride gas was formerly dried by using concentrated sulfuric acid, but another approach is to make use of the properties of the $HCl-H_2O$ system (see Fig. 10 and Table 2). At low temperature, the partial vapor pressure of water over highly concentrated hydrochloric acid is very low; hydrogen chloride can therefore be dried by cooling to sufficiently low subzero temperature. At slightly elevated pressure ($p_{abs} \approx 0.3$ MPa), it is sufficient to attain $-10\,°C$ to give satisfactory results. Hydrochloric acid mist must, however, be removed completely, or severe corrosion can result.

3.2.4. Recovery from Waste Incineration

All chlorination processes produce undesirable chlorinated hydrocarbons in addition to the product desired. These unusable and often noxious substances must be disposed of for environmental reasons. One possibility is incineration to hydrogen chloride, water, and carbon dioxide; the hydrogen chloride should be recovered for both economic and environmental reasons.

Although the methods used vary in detail [23–25], the general principle remains the same. Combustion takes place in a refractory-lined furnace at $> 1000\,°C$. Waste chlorinated hydrocarbons containing up to 70 wt % chlorine normally possess good combustion properties, but any further increase in chlorine content reduces the calorific value to such an extent that fuel oil or gas must be added to support combustion.

Optimum reaction conditions lie within the following limits:

1) Excess oxygen must always be present to guarantee complete decomposition of chlorinated hydrocarbons. The hydrogen chloride gas formed is consequently able to react with this oxygen to form chlorine and water in accordance with the Deacon equation:

$$2\,HCl + 1/2\,O_2 \rightleftarrows Cl_2 + H_2O \qquad (14)$$

The reaction has a negative enthalpy of reaction ($\Delta H_r^0 = -57.1$ kJ/mol) and entropy of reaction ($\Delta S_r^0 = -64.7\,J\,K^{-1}\,mol^{-1}$). With increasing temperature, the equilibrium shifts from right to left, thus favoring formation of hydrogen chloride. The temperature is maintained as high as construction materials permit, to prevent formation of chlorine; even if the partial pressure of water is kept high and the oxygen excess low, a temperature $> 1000\,°C$ is necessary to shift the equilibrium to the left. This keeps the chlorine at an acceptable level and also ensures complete combustion.

2) Reaction temperature must not exceed $1200\,°C$ to prevent formation of nitrogen oxides, which would contaminate the hydrochloric acid product and, which is worse, cause waste gas purification problems. A possible solution is to use pure oxygen instead of air [16].

For these reasons, incineration of waste chlorinated hydrocarbons is carried out between 1000 and $1200\,°C$. Cooling the flue gases combined with steam generation can cause problems with construction materials. The absence of corrosive combustion products must, therefore, be assured. The treated wastes should have a well-known composition (down to trace contaminants) for proper prediction of the expected combustion products, their concentration, and behavior. Slow cooling can cause chlorine to reform, so rapid cooling is preferred.

Hydrogen chloride is recovered from the combustion gases by absorption, i.e., the process is similar to that used in recovery from chlorination processes, except for the following:

1) A dry process is impossible because water vapor is always present.
2) The gases should be condensed in such a way as to form the most concentrated hydrochloric acid possible so that gaseous hydrogen chloride or concentrated hydrochloric acid can be obtained. Dilute hydrochloric acid has little commercial value.
3) Special corrosion problems must be overcome at places of possible, but not necessarily regular, acid condensation.
4) If the gaseous combustion products are used directly, the operating temperature must always be kept above the condensation temperature of hydrochloric acid.

One of the oldest plants producing commercial quantities of 33 wt % hydrochloric acid by incinerating chlorinated hydrocarbons is that of Atochem at St. Auban in the south of France [25]. This plant opened in 1975 and consumes approximately 16 000 t of chlorinated hydrocarbons material per year. Many additional chlorinated hydrocarbon incineration plants have since been built, and the hydrochloric acid produced is finding a wider range of uses (e.g., addition to ethylene) [27].

3.3. Purification

Many impurities in hydrogen chloride gas (e.g., SO_2, As, and Cl_2) can be removed by adsorption on activated carbon [28]. As the importance of the sulfate process decreases, the removal of chlorinated hydrocarbons from hydrogen chloride gas or hydrochloric acid is of greater practical relevance. Gaseous hydrogen chloride can be purified by low-temperature scrubbing with a high-boiling solvent, which may be another chlorinated hydrocarbon (e.g., hexachlorobutadiene or tetrachloroethane), or special oil fractions. The use of activated carbon is often unnecessary after such treatment. Chlorine may be removed with carbon tetrachloride, in which it is much more soluble than is hydrogen chloride gas.

Hydrochloric acid, used as such or for generation of hydrogen chloride, contains mainly volatile impurities such as chlorinated hydrocarbons. In these cases the impurities may be removed from the acid by stripping (see Fig. 11). Stripping with an inert gas stream, which results in lower energy consumption, is possible, although in most cases heating of the gas is preferable for environmental reasons. Inorganic impurities, in particular iron, are removed by ion exchange.

4. Quality Specifications

The quality specifications for *gaseous hydrogen chloride* depend on intended use. There are no general quality standards. Only a trace (a few parts per million) of water is usually permitted, particularly if the gas is to be compressed and liquefied. Dry hydrogen chloride does not attack iron; thus, steel equipment and pipelines can be used. Hydrogen chloride gas is regarded as dry if no danger exists of hydrochloric acid condensation at operating temperature and pressure.

Hydrochloric acid has no generally accepted quality specification. In the Federal Republic of Germany, the only official standard is DIN 19 610 (November 1975), which gives the following specifications for hydrochloric acid for boiler feed water treatment:

Hydrogen chloride content ≥ 30 wt %
Sulfate (SO_4^{2-}) content ≤ 0.5 wt %
Iron (Fe^{3+}) content ≤ 0.002 wt %
Content of organic chlorine compounds (expressed as
Cl^-) ≤ 0.02 g/L

Quality requirements for medicinal use are given in DAB 9 [29] for both concentrated (36 wt %) and dilute (10 wt %) hydrochloric acid. Prescribed purity tests include free chlorine (4 ppm), arsenic (2 ppm), heavy metals (2 ppm), sulfate (20 ppm), and residue on evaporation (0.01 %). Other countries have their own regulations [30].

5. Storage and Transportation

Wrought- and cast-iron, steel, and steel alloys have good resistance toward dry hydrogen chloride gas or liquid. Steel is, therefore, suitable for equipment and containers up to 280 °C, including pressure vessels used for storage and transportation of dry hydrogen chloride.

Even traces of moisture significantly increase the corrosion rate, and alloyed austenitic steels become liable to stress crack corrosion. If the working temperature cannot be raised above the condensation point, either nonmetallic materials or more exotic metals must be used, such as nickel–molybdenum alloys with high molybdenum content, zirconium, niobium, tantalum, or noble metals such as gold or the platinum-group metals. In practice, it is imperative to specify precisely the operating limits for metals to be used with hydrochloric acid or moist hydrogen chloride. In all cases, limits are given for acid concentration or temperature. Contamination with oxidizing or complexing agents must be avoided. Traces of fluoride ions in hydrochloric acid, for example, can lead to rapid corrosion

of zirconium or tantalum, although tantalum has proved suitable for hydrochloric acid distillation columns in the absence of damaging impurities.

A wide choice of nonmetallic construction materials is available. For high temperature, ceramics, graphite, or polytetrafluoroethylene may be suitable. Phenolic resins are used as cement for brick lining. At lower acid concentration (below 25 %), enameled or glass apparatus can be used. At lower temperature, (e.g., for storage and transportation), vessels may be made of rubber-lined steel, polyolefin, or poly(vinyl chloride), the latter usually with glass-fiber reinforcement. If the acid contains traces of chlorinated hydrocarbons such as dichloromethane, vessels fabricated in, or lined with, phenolic resin are preferred. Plastics such as poly(vinyl chloride) and polyethylene can withstand use for several years.

For transportation of hydrochloric acid and liquefied hydrogen chloride, special labels must be used that comply with the international regulations for transportation of hazardous goods:

Hydrochloric acid:	Hazard code number 80 UN no. 1789 Warning sign 8 (causes burns) CFR 49: 172.101 Cor. M RID/ADR, class 8, no. 5 b
Hydrogen chloride:	Hazard code number 286 UN no. 1050 Warning signs 6.1 (poisonous) and 8 (causes burns)

6. Uses and Economic Aspects

Hydrochloric acid and hydrogen chloride are two of the most important basic industrial chemicals.

The largest proportion of *hydrogen chloride* is used immediately by the producer himself, because the hydrogen chloride byproduct of organic chlorination processes has to be used as completely as possible. This is done by means of oxychlorination or hydrochlorination, or to a small extent, recovery of chlorine (e.g., by the KEL process, a modified Deacon process [31] or by electrolysis of hydrochloric acid [32]).

Aqueous hydrochloric acid has many uses as a strong inorganic acid: manufacture of chlorides, dissolution of minerals, pickling and etching of metals, regeneration of ion-exchange resins for water treatment, neutralization of alkaline products or waste materials, acidification of brine in chlor-alkali electrolysis, and many others.

Hydrochloric acid and hydrogen chloride production figures for various countries are given in Table 4.

7. Toxicology and Occupational Health

Hydrogen chloride causes severe irritation of the eye and respiratory tract and, to a lesser extent, the skin. The vapor can produce keratoconjunctivitis. Inhalation causes irritation and damage to mucous membranes. Hydrogen chloride is usually detectable by odor at 1 – 5 ppm and becomes objectionable at 5 – 10 ppm. Prolonged occupational exposure to excessive hydrogen chloride concentration causes increased incidence of chronic bronchitis or stomach and intestinal disorders due to abnormal acid levels. Increased incidence of a characteristic dental decay has also been observed with prolonged exposure, even at apparently tolerable levels. A concentration > 10 ppm is strongly irritating, even when some acclimatization has occurred [33].

Suitable protective goggles should be worn by personnel working with hydrochloric acid; the use of protective clothing, rubber gloves, and rubber boots is recommended. Inhalation of the vapor should be avoided. The respirator filter for inorganic gases is suitable (in the Federal Republic of Germany, code letter B, code color gray), as is the hydrogen chloride and sulfur dioxide filter (code letter E, code color yellow [34]).

In the Federal Republic of Germany, hydrogen chloride and hydrochloric acid (> 10 wt %) are controlled by the Verordnung über gefährliche Stoffe (GefStoffV, regulations for the control of dangerous substances), which have been adjusted to the corresponding directives of the EEC [35]:

1) Anhydrous hydrogen chloride: hazard symbol C (corrosive). Hazard information ("R" notices): 35 (causes serious burns); 37 (causes irritation of respiratory tract). Safety Instructions ("S" notices): 7/9 (keep containers tightly closed and store in a well-ventilated place); 26 (on contact with the eye,

Table 4. Production of hydrogen chloride and hydrochloric acid, 1979 – 1993 (based on 100 % HCl)

Production,	Country				
10^3 t	Federal Republic of Germany	France	United Kingdom	United States	Japan
1980	891	676	143	2551	570
1981	888	702	128	2269	565
1982	848	688	118	2252	551
1983	900	673	130	2468	511
1984	956	765	156	2732	521
1985	945	656	153	2807	549
1986	931	626	157	2413	549
1987	987	647	174	2495	
1989 *	958	238	167	3268	
1991 *	839	216	153	3381	
1992 *	878	199	148	3610	1786
1993 *	820	190		3492	

* [40]

thoroughly wash with water and seek medical advice).

2) Hydrochloric acid containing over 25 wt % hydrogen chloride: hazard symbol C (corrosive, causes burns).“R” notices: 34 (causes burns); 37 (see above).“S” notices: 2 (keep away from children), 26 (see above).
3) Hydrochloric acid with 10 – 25 wt % hydrogen chloride: hazard symbol X_i, irritant.“R” notices: 36/38 (causes eye and skin irritation)“S” notices: 2 (see above), 28 (on contact with the skin, immediately wash with copious amounts of water).

In the event of skin or eye contact, the affected area should be rinsed immediately and thoroughly with a large amount of water. Contaminated clothing should be treated similarly, and alkaline cleansing solution may also be used. For work with hydrogen chloride, especially at elevated pressure, breathing apparatus with an independent air supply and full protective clothing must be readily available in case of accident.

In the Federal Republic of Germany, the MAK value for hydrogen chloride is 5 ppm by volume, corresponding to 7 mg/m^3 [36]. The same values are valid as TLV in the United States. In the Soviet Union 3 ppm (5 mg/m^3) has been established.

In clean air regulations (TA Luft, 1986), concentration limits for hydrogen chloride in the atmosphere are 0.1 mg/m^3 as an annual average and 0.2 mg/m^3 for short periods. Phytotoxicological investigation has shown that plants have differing tolerances [37–39]. According to guidelines issued by the VDI, 2310 sheet 4, sensitive vegetation such as lettuce, strawberries, spruce, birch, and apple trees should not be exposed to a hydrogen chloride concentration higher than 0.8 mg/m^3 as a daily average or 0.1 mg/m^3 as a monthly average.

The concentration of hydrogen chloride in industrial waste gases must (according to regulations in the Federal Republic of Germany) normally not exceed 30 mg/m^3, although the limit is 50 mg/m^3 for waste incineration plants; the cost of gas cleaning can, therefore, vary considerably. A simple water scrubbing operation is often sufficient, but more than one stage may be necessary, sometimes including a final scrubbing with an alkaline solution. In many cases it is not possible to produce a commercial hydrochloric acid. The dilute acid is then neutralized before being discarded in the wastewater. Neutralization with caustic soda occasionally results in a salt solution that is pure enough to be recycled to a chlorine plant.

8. References

1. H. Chihara, A. Inaba, *J. Chem. Thermodyn.* **8** (1976) 915 – 934.
2. S. Miskiewicz, K. Rieser, T. Dorfmüller, *Ber. Bunsenges. Phys. Chem.* **80** (1976) 395 – 405.
3. K. Clusius, G. Wolf, *Z. Naturforsch.* **A** (1947) 495.
4. L. S. Adler, C. L. Yaws, *Solid State Technol.* **19** (1976) 35 – 38.

5. K. K. Tremper, J. M. Pransnitz, *J. Chem. Eng. Data* **21** (1976) 295 – 299.
6. W. Gerrard, E. D. Macklen, *Chem. Ind. (London)* 1959, 1070 – 1071.
7. S. M. Khodeeva, M. B. Rozovskii, *Russ. J. Phys. Chem. (Engl. Trans.)* **49** (1975) 824 – 827.
8. Codata Recommended Key Values, *J. Chem. Thermodyn.* **10** (1978) 904.
9. D. D. Wagman, W. H. Evans, I. Halow, V. B. Parker, S. M. Bailey, R. H. Schumm, *NBS Tech. Note (U.S.)* **270** (1965) no. 1, 25 – 27.
10. A. Schmidt, *Chem. Ing. Tech.* **25** (1953) 455 – 466.
11. *Ullmann,* 3rd ed., **15,** 70.
12. J. J. Fritz, C. R. Fuget, *Ind. Eng. Chem. Data* **1** (1956) 110 – 112.
13. J. F. F. Kao, *J. Chem. Eng. Data* **15** (1970) 362 –367.
14. B. G. Staples, J. M. Procopio, G. J. Su, *Chem. Eng. (N.Y.)* **77** (1970) 113 – 115.
15. W. Kindler, G. Wuster, H. Rau, *Ber. Bunsenges. Phys. Chem.* **82** (1978) 543 – 545.
16. R. W. Bach, H. A. Friedrichs, H. Rau, *High Temp. High Pressures* **9** (1977) 305 – 312.
17. O. F. Raper, C. B. Farmer, R. A. Toth, B. D. Robbins, *Geophys. Res. Lett.* **4** (1977) 531 – 534.
18. *Ullmann,* 4th ed., **10,** 431 – 433.
19. *Ullmann,* 3rd ed., **15,** 75.
20. *Ullmann,* 3rd ed., **15,** 71 – 73.
21. *Ullmann,* 4th ed., **13,** 482.
22. R. Vieweg, *Chem. Tech. (Leipzig)* **12** (1960) no. 4, 188 – 195.
23. J. C. Zimmer, R. Guaitella, *Hydrocarbon Process.* (1976) 117 – 118.
24. Hoechst, US 4 073 871, 1977 (W. Opitz, H. Hennen); DT 2 611 671, 1976.
25. Rhône-Poulenc, Company brochure, *Combustion of Chlorinated Residues* (1978).
26. Stauffer Chemical O., *Hydrocarbon Process.* (1981) 170.
27. Uhde GmbH, EP 0 137 160, 1985, 1987 (U. Rieker, G. Link).
28. *Ullmann,* 3rd ed., **15,** 80.
29. DAB, vol. 9.
30. Europäisches Arzneibuch, vol. 2, 137 – 138, Deutscher Apotheker-Verlag, Stuttgart 1975.
31. *Ullmann,* 4th ed., **9,** 357.
32. *Ullmann,* 4th ed., **9,** 355.
33. D. Henschler (ed.): *Gesundheitsschädliche Arbeitsstoffe,* Verlag Chemie GmbH, Weinheim 1972 – 1979.
34. Datenbank für wassergefährdende Stoffe (DABAWAS) vom 23. 03. 1987, 2989.
35. Verordnung über gefährliche Arbeitsstoffe vom 26. 08. 1986 (BGBL 1, S. 1470).
36. Mitteilung 23 der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, VCH Verlagsgesellschaft mbH, Weinheim 1987.
37. H. J. Grecelius, W. Forwer, R. Zahn, *Staub, Reinhalt. Luft* **36** (1976) 201 – 205.
38. H. van Haut, *Schriftenr. LJB Landesanst. Immissions Bodennutzungsschutz Landes Nordrhein Westfalen* **33** (1974) 39 – 43.
39. C. Holsenberg, *Schriftenr. Ver. Wasser Boden Lufthyg.* **41** (1974) 267 – 284.
40. K. H. Büchel, H.-H. Moretto, P. Woditsch: *Industrial Inorganic Chemistry,* Wiley-VCH, Weinheim 2000, p. 163.

My Profile L

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Hydrogen

Standard Article

Peter Häussinger[1], Reiner Lohmüller[2], Allan M. Watson[3]

[1]Linde AG Werksgruppe TVT München, Höllriegelskreuth, Federal Republic of Germany
[2]Fachhochschule Ostfriesland, Emden, Federal Republic of Germany
[3]Linde AG München, Höllriegelskreuth, Federal Republic of Germany

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a13_297
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. History
2. Properties
2.1. Physical Properties
2.1.1. General Physical Data
2.1.2. Solubilities, Diffusion Properties and Evaporation Equilibria
2.2. Chemical Properties
2.3. Ignition and Detonation Performance
2.4. Isotopes
3. Occurence
4. Production
4.1. Production from Coal and Hydrocarbons
4.1.1. Coke Oven Gas
4.1.2. Gasification of Coal and Hydrocarbons
4.1.2.1. Coal Gasification
4.1.2.2. Gasification of Liquid and Gaseous Hydrocarbons
4.1.2.3. Examples of Industrial Scale Hydrogen Production (by Partial Oxidation of Hydrocarbons)
4.1.3. Catalytic Reforming of Hydrocarbons
4.1.3.1. Feedstocks
4.1.3.2. Tubular Reformer
4.1.3.3. Steam Reforming Plant
4.1.3.4. Special Designs
4.1.3.5. Autothermal Reactors
4.1.4. Refinery Processes
4.1.5. Petrochemical Processes
4.2. Electrolysis

4.2.1. Principles
4.2.2. Conventional Water Electrolysis
4.2.3. New Developments in Electrolytic Processes
4.2.4. Hydrogen as Byproduct from other Electrochemical Processes
4.3. Thermochemical Water Cleavage
4.3.1. Thermodynamics of Closed Cycles
4.3.2. Further Criteria of Thermochemical Cycles
4.3.3. Thermochemical Cyclic Processes
4.3.4. Hybrid Cycle Processes
4.4. Other Methods for the Cleavage of Water
4.4.1. Thermolytic and Radiolytic Processes
4.4.2. Photochemical or Photoelectrical Water Cleavage
4.4.3. Hydrogen Formation in Biological Systems
4.5. Other Chemical Processes
4.5.1. Hydrogen from Conversion of Metals
4.5.2. Hydrogen from Ammonia
4.5.3. Hydrogen from Methanol
4.5.4. Hydrogen from Hydrogen Sulfide
4.5.5. Hydrogen as Byproduct
4.6. Economic Aspects
4.6.1. Analysis of the Cost Structure
4.6.2. Prognosis of the Hydrogen Costs
5. Purification of Hydrogen
5.1. Low-Temperature Processes
5.1.1. Condensation Processes
5.1.2. Condensation and Rectification
5.1.3. Low-Temperature Absorption Processes
5.2. Adsorption Processes
5.2.1. Principles
5.2.2. Thermally Regenerated Adsorbers
5.2.3. Pressure-Swing Adsorption (PSA)
5.3. Catalytic Gas Purification
5.3.1. Carbon Monoxide Removal
5.3.2. Methanation
5.3.3. Selective Oxidation
5.3.4. Catalytic Removal of Sulfur Compounds
5.3.5. Removal of Oxygen
5.3.6. Removal of Nitrogen Oxides
5.4. Scrubbing Processes
5.4.1. Principles
5.4.2. Physical Scrubbing (Rectisol Process)
5.4.3. Chemical Scrubbing Processes
5.5. Permeation Processes
5.5.1. Metal Membrane Techniques
5.5.2. Polymer Membrane Processes
6. Liquefaction of Hydrogen

6.1. Principles of Hydrogen Liquefaction
6.2. Liquefaction Plants
6.3. Further Developments
7. Handling of Hydrogen
7.1. Quality Specifications
7.2. Compression and Expansion
7.3. Transportation and Distribution
7.4. Storage
7.5. Materials
7.6. Removal of Hydrogen
7.7. Safety Aspects
7.8. Toxicology
7.9. Analysis
8. Conventional Uses
8.1. Hydrogen in the Chemical Industry
8.1.1. Ammonia Synthesis
8.1.2. Hydrogen in Refinery Processes
8.1.3. Hydrogen in Coal Refinement
8.1.4. Synthesis Gas
8.1.5. Hydrogen in Organic Synthesis
8.1.6. Hydrogen in Inorganic Synthesis
8.2. Hydrogen in Metallurgy
8.3. Other Uses
9. Hydrogen Energy
9.1. Aims of the Hydrogen Energy Economy
9.2. Production Systems for Hydrogen Energy
9.2.1. Energy Sources
9.2.2. Production Methods for the Energy Carrier Hydrogen
9.3. Hydrogen-Energy Conversion Systems
9.3.1. New Developments for Hydrogen Storage
9.3.1.1. Cryoadsorption
9.3.1.2. Hydride Technology
9.3.1.3. Liquid Organic Hydrogen Carriers
9.3.2. Combustion and Heating
9.3.3. Hydrogen Fuel Systems
9.3.3.1. Hydrogen Propulsion Systems for Space
9.3.3.2. Hydrogen as Aviation Fuel
9.3.3.3. Hydrogen for Automotive Vehicle Transport
9.3.4. Electricity Generation from Hydrogen
9.3.4.1. Heat - Power Processes
9.3.4.2. Electrochemical Energy Conversion
9.4. Future Developments
9.4.1. Production Schemes
9.4.2. Hydrogen Energy Economics
9.4.3. International Research

[Top of Page]

1. History

In the middle ages PARACELSUS is reported to have noted that a gas is yielded when iron is dissolved in the “spirit of vitriole”. TURQUET DE MAYERNE (1573 – 1655) noted that this gas was inflammable. However, hydrogen was first separated and identified in the second half of the 18th century; BOYLE produced “facticious air” from diluted sulfuric acid and iron.

CAVENDISH (1731 – 1810) proved that there were different types of air, one of which was “inflammable air” and that a number of metals, when dissolved in acid, produced various amounts of this gas. In 1766, he published precise values for the specific weight and density. During the late 1770s, he made experiments with electrical discharges in a hydrogen – oxygen mixture thereby producing water.

In 1785, LAVOISIER, demonstrated the splitting of water into hydrogen and oxygen in a heated copper tube. He also gave the “inflammable air” the name hydrogen.

Chronology of Further Significant Events

1800. First production of hydrogen and oxygen by electrolysis (NICHOLSON and CARLYLE)
1898. Liquefaction of hydrogen using the Linde process (J. DEWAR)
1902. First commercial electrolysis installation by OERLIKON
1929. Production of pure para-hydrogen (K. F. BONHOEFFER, P. HARTECK)
1931. Discovery of the hydrogen isotope deuterium (H. C. UREY)
1935. Synthesis of “superheavy hydrogen”, tritium, through neutron bombardment of deuterated phosphoric acid (P. HARTECK, OLIFANT, E. RUTHERFORD)
1954. Ignition of the first hydrogen bomb on the Bikini Atoll (USA)
1955. Description of the utilization of hydrogen as energy carrier medium (JUSTI)
1969. Development of an overall hydrogen energy concept using the favorable properties of hydrogen and including non-conventional energy systems (J. O. M. BOCKRIS, D. P. GREGORY, C. MARCHETTI, T. N. VEZIROGLU, and others)
1986. The total hydrogen production worldwide amounts to 500×10^6 m^3 (STP)/a. U.S. hydrogen production and consumption is about 8×10^6 t.

Hydrogen in Transportation. The buoyancy of hydrogen quickly lead to the use of this gas in balloons (C. CHARLES, 1783), but the use for aeronautics was ended by the HINDENBURG disaster at Lakehurst in 1937. Experiments using hydrogen as an *engine fuel* had already been carried out in 1820 by W. CECIL. These experiments were followed by investigations in car engines up until the Second World War and partly afterwards (RICARDO, BURSTAL, ERREN). Various transport systems using hydrogen were tested, e.g., in the United States by R. BILLINGS or in the Federal Republic of Germany (Daimler Benz, BMW) in the last decade.

The potential of hydrogen as an *aviation fuel* was established 1957 with the lift-off of a hydrogen-powered B-57 twin-engine jet bomber. From 1963 on liquid hydrogen –liquid oxygen propelled rockets were launched. For the Apollo moon flights 12×10^6 L of liquid hydrogen were necessary to tank the Saturn carrier rockets.

[Top of Page]

2. Properties

Hydrogen (hydrogenium, *ὑ δ ω ῏* = the water, *γ ε ν ν ε ι ν* = to give birth) [12 385-13-6] H, A_r 1.00 797, atomic number 1, is the first element in the periodic table, having the electron configuration 1 s^1. It has usually the oxidation state +1, but in salt-type hydrides –1 is also possible. Three isotopes with A_r 1, 2 and 3 are known; the isotope with A_r 3 is unstable. The differences in the relative isotopic masses are large so that physical and kinetic properties can differ considerably. The isotope with the relative mass 2 is named *deuterium* [16873-17-9] (symbol D), that with the mass 3 *tritium* [15 086-10-9] (symbol T). The atomic nuclei, all with a single positive charge, are given the names *proton, deuteron*, and *triton.*

With the expected increasing significance of hydrogen as a universal chemical and as an energy carrier, its physical and thermodynamic properties have undergone further extensive investigations. The compilation given here is an excerpt; comprehensive data can be found in various summaries and other literature, e.g., in [1-8].

2.1. Physical Properties

2.1.1. *General Physical Data*

The *atomic radius* of the free hydrogen atom is 53 pm, the *covalent atomic radius* of hydrogen in crystal structures is between 30 and 35 pm. The length of the H – C bond in saturated hydrocarbons is ca. 112 pm, in water vapor the H – O bond has a length of 96 pm. The isotopes H, D and T form diatomic molecules, the combinations being H_2, [1333-74-0], D_2 [7782-39-0], T_2 [10028-17-8], HD [13983-20-5], HT [14 885-60-0] and DT [14885-61-1]. Atomic and molecular data are given in Table 1.

Table 1. Atomic and molecular data of hydrogen and hydrogen isotopes

	Protium $^{1}_{1}H$	Deuterium $^{2}_{1}H$	Tritium $^{3}_{1}H$
A_r	1.007825	2.0140	3.01605
Natural abundance, %	99.985	0.015	$<10^{-12}$
Half-life, a			12.26
Particle energies, MeV			0.01861 (β-emission)
Thermal neutron capture cross section, barn (10^{-24} cm^2)	0.322	$0.51 \cdot 10^{-3}$	$<6 \cdot 10^{-6}$
Nuclear spin, $h/2\pi$	1/2	1	1/2
Nuclear magnetic moment, nuclear magnetons	+2.79278	+0.85742	

	n-H_2	n-D_2	n-T_2	HD	HT	DT
M_r	2.016	4.029	6.034	3.022	4.025	5.022
Bond length, pm	74.6	74.2		74.1		
Dissociation energy, kJ/mol	431.6	439.2		435.1		
Triple point						
temperature, K	13.96	18.73	20.62	16.60	17.63	19.71
pressure, kPa	7.3	17.1	21.6	12.8	17.7	19.4
Critical point						
temperature, K	32.98	38.35	40.44	35.91	37.13	39.42
pressure, MPa	1.31	1.67	1.85	1.48	1.57	1.77
Normal boiling point, K (atmospheric pressure)	20.39	23.67	25.04	22.13	22.92	24.38
Density at normal boiling point						
ϱ_L, kg/m^3	70.811	162.50	260.17	114.80	158.62	211.54
ϱ_V, kg/m^3	1.316	2.230	3.136	1.802	2.310	2.694
Heat of vaporisation (25 K), J/mol	825	1175	1400	1000	1100	1290

Ortho- and Para-Hydrogen. The atomic nuclei of hydrogen and its isotopes possess a spin, the value of which is 1/2 for H and T, and 1 for D. In the diatomic molecules both atomic nuclei can either rotate in the same or in opposing direction, i.e., parallel or anti-parallel. The molecules H_2, D_2 and T_2 can therefore exist in the form of two isomers with different nuclear spin, namely in the ortho-form (parallel nuclear spin) and in the para-form (anti-parallel nuclear spin); the two forms are interconvertible and in equilibrium with each other. This effect can be observed in all homonuclear diatomic molecules in which the nuclei have spins, but in the case of hydrogen the effects on properties are most pronounced. The following forms can be distinguished:

o-H_2 ortho-hydrogen

p-H_2 para-hydrogen

e-H_2 equilibrium hydrogen

n-H_2 normal hydrogen

e-H_2, e-D_2 and e-T_2 all correspond to mixtures of the equilibrium concentration of *definite* temperatures. n-H_2 (n-D_2, n-T_2) is a mixture of the equilibrium concentration at *ambient* temperature, i.e., 25 mol % p-H_2 or p-T_2 and 75 % o-H_2 or o-T_2. n-D_2 contains 33.33 % p-D_2 and 66.67 % o-D_2.

The conversions o-H_2 to p-H_2 and o-T_2 to p-T_2 are exothermic; both equilibrium composition and heat of conversion are a function of temperature, but independent of pressure (see Figs. 1 and 2).

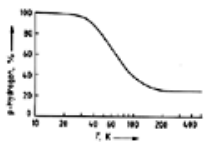

Figure 1. Equilibrium composition for the conversion of o-hydrogen – p-hydrogen

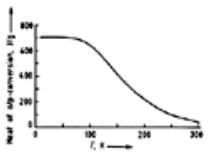

Figure 2. Heat of conversion of o-hydrogen – p-hydrogen

For D_2 the ortho : para relation is different to that of H_2 and T_2. The conversion *p*-D_2 to o-D_2 is exothermic so that at low temperatures the ortho – form is prevalent, e.g., at 20 K 99.52 %.

The ortho – para equilibrium is attained very slowly in the gaseous phase (see Fig. 3), but more rapidly in the condensed state (see Chap. Liquefaction of Hydrogen). The conversion of deuterium is slower than that of hydrogen, but tritium reacts more rapidly than H_2. Solid tritium, for example has a conversion rate 210 times faster than solid hydrogen. The conversion rate can be increased in the presence of catalysts or magnetic fields. The hydroxides of Fe (III), Co (III), Ni (II), Cr (III), Mn (IV) or active charcoal are effective as catalysts [9], [10]. Apart from strong homogeneous magnetic fields, the inhomogeneous magnetic fields of paramagnetic substances, electrical discharges as well as thermo and photochemical activation increase the conversion rate.

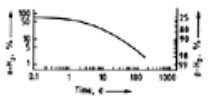

Figure 3. Self conversion of o-hydrogen to *p*-hydrogen in absence of catalysts as a function of time

Thermodynamic Properties. The differences in the masses of the isotopes are relatively large, so that thermodynamic properties differ considerably. Small differences are also found among the spin-spin isomers so that for certain purposes attention must be paid to the actual form present. At low temperature *p*-H_2, *o*-D_2 or *p*-T_2 can be present in a virtually pure state. The properties of *n*-H_2 (o – p isomerization prevented by rapid cooling) are also known at low temperatures.

Hydrogen gas is colorless, non-poisonous, odorless and tasteless. At 0 °C and 100 kPa 1 L of H_2 has a mass of 0.0898 g. The most important properties of n-H_2 and p-H_2 are given in Tables 2-4, and Figure 4.

Table 2. Selection of physical properties of n-hydrogen and p-hydrogen at fixed points

	p-Hydrogen	n-Hydrogen
Triple point		
Temperature, K	13.803	13.957
Pressure, kPa	7.04	7.2
Density (solid), kg/m^3	86.48	86.71
Density (liquid), kg/m^3	77.03	77.21

Density (vapor), kg/m^3	0.126	0.130
Boiling point (101.3 kPa), K	20.268	20.39
Heat of vaporization, J/mol	898.30	899.1
Liquid phase		
Density, kg/m^3	70.78	70.96
Specific heat capacity		
C_p, J mol^{-1} K^{-1}	19.70	19.7
C_v, J mol^{-1} K^{-1}	11.60	11.6
Enthalpy,* J/mol	–516.6	548.3
Entropy, J mol^{-1} K^{-1}	16.08	34.92
Viscosity, mPa s	13.2×10^{-3}	13.3×10^{-3}
Velocity of sound, m/s	1089	1101
Thermal conductivity, W m^{-1} K^{-1}	98.92×10^{-3}	100×10^{-3}
Compressibility factor	0.01712	0.01698
Vapor phase		
Density, kg/m^3	1.338	1.331
Specific heat capacity		
C_p, J mol^{-1} K^{-1}	24.49	24.60
C_v, J mol^{-1} K^{-1}	13.10	13.2
Enthalpy,* J/mol	381.61	1447.4
Entropy, J mol^{-1} K^{-1}	60.41	78.94
Viscosity, mPa s	1.13×10^{-3}	1.11×10^{-3}
Velocity of sound, m/s	355	357
Thermal conductivity, W m^{-1} K^{-1}	16.94×10^{-3}	16.5×10^{-3}
Compressibility factor	0.906	0.906
Critical point		
Temperature, K	32.976	33.19
Pressure, MPa	1.29	1.325
Density, kg/m^3	31.43	30.12

Properties at STP (273.15 K, 101.3 kPa)		
Density, kg/m^3	0.0899	0.0899
Specific heat capacity		
C_p, J mol^{-1} K^{-1}	30.35	28.59
C_v, J mol^{-1} K^{-1}	21.87	20.3
Viscosity, mPa s	8.34×10^{-3}	8.34×10^{-3}
Velocity of sound, m/s	1246	1246
Thermal conductivity, W m^{-1} K^{-1}	182.6×10^{-3}	173.9×10^{-3}
Dielectric constant	1.00027	1.000271
Compressibility factor	1.0005	1.00042
Prandtl number	0.6873	0.680

* The reference state for enthalpy is zero for the ideal gas at 0 K.

Table 3. Some physical and thermodynamic properties of gaseous n-hydrogen

T, °C	Density, kg/m^3	C_p, J mol^{-1} K^{-1}	C_v, J mol^{-1} K^{-1}	Enthalpy, kJ/mol	Entropy, J mol^{-1} K^{-1}
–150	0.1994	23.97	15.65	3.7086	191.38
–150	0.1418	26.42	18.08	4.9434	126.98
– 50	0.1100	27.85	19.55	6.3342	133.87
0	0.0899	28.61	20.32	7.7492	139.59
+ 50	0.0760	29.03	20.68	9.1914	144.44
+100	0.0658	29.20	20.84	10.6461	148.62
+150	0.0580	29.20	20.91	12.1062	152.30

Table 4. Vapor pressure and density of p-hydrogen at low temperatures

Temperature, K	Vapor pressure, kPa	Density, kg/m^3		
		ϱ_S	ϱ_L	ϱ_G
1	11.0×10^{-37}	89.024		
5	4.76×10^{-3}	88.965		
10	2.556×10^{2}	88.136		0.006

12	1.837×10^3	87.532		0.037
13.803 [a]	7.0	86.503	77.019	0.126
20	93.5		71.086	1.247
20.268 [b]	101.3		70.779	1.338
30	822.5		53.930	10.887
32.976 [c]	1.293×10^3			31.43

[a] Triple point.
[b] 101.3 kPa.
[c] Critical point.

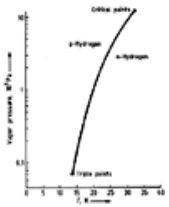

Figure 4. Vapor pressure of n-hydrogen and p-hydrogen

Contrary to most other gases, the *inversion temperature* of hydrogen lies below ambient temperature. The temperature, at which the negative Joule – Thomson coefficient μ (heating by expansion) changes to a positive μ (cooling by expansion), can be seen from the Joule – Thomson inversion curve (Fig. 5). A cooling effect is obtained only when the decompression takes place from a state which is characterized by a point below the inversion curve.

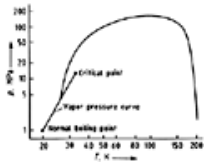

Figure 5. Joule – Thomson inversion curve for p-hydrogen with vapor pressure curve

Liquid hydrogen is a colorless, very mobile liquid with low viscosity and surface tension. Some properties are given in Tables 2 and 4.

Solid hydrogen is colorless and crystallizes in the hexagonal closest-packed structure. Data for p-H_2 are: a = 375 pm, c/a = 1.633, molar volume 22.56 cm^3/mol [11]. At higher pressure several phase transitions occur [12]. Under extremely high pressure (ca. 2 – 3×10^{11} Pa) a metallic, electrically conductive hydrogen phase with a density of >1.000 kg/m^3 has been reported [13-15]. Some properties of solid p-H_2 at the triple point are presented in Table 5.

Table 5. Properties of solid p-hydrogen at triple point (7.04 kPa, 13.803 K)

Density, kg/m^3	86.50
Heat of fusion, J/mol	117.5
Heat of sublimation, J/mol	1022.9
Molar heat, J mol^{-1} K^{-1}	5.73
Enthalpy, J/mol	–740.2
Entropy, J mol^{-1} K^{-1}	1.49
Thermal conductivity, W m^{-1} K^{-1}	0.9

Dielectric constant	1.286

2.1.2. *Solubilities, Diffusion Properties and Evaporation Equilibria*

Solubility of Hydrogen in Liquids. Hydrogen is only slightly soluble in liquids. In contrast to the highly-soluble gases, however, the solubility generally increases with increasing temperature. Examples for the solubility in *nonaqueous solvents* are given in Table 6 and Figure 6.

Table 6. Solubility of hydrogen in nonaqueous solvents

Solvent	*t*, °C	*p*, [a] MPa	Solubility, mol %
Hexane	0	0.1	5.82×10^{-2}
Benzene	20	0.1	2.47×10^{-2}
Diethylether	0	0.1	5.29×10^{-2}
Ethanol	0	0.1	1.8×10^{-2}
Carbon tetrachloride	0	0.1	2.60×10^{-2}
Freon 113	25	0.1	6.55×10^{-2}
Naphtha [b]	25	1	0.031 [d]
Naphtha	25	5	0.151 [d]
Naphtha	25	10	0.295 [d]
Gas oil [c]	25	1	0.027 [d]
Gas oil	25	5	0.137 [d]
Gas oil	25	10	0.274 [d]

[a] Partial pressure of hydrogen.
[b] Density 0.8003 g/cm^3, vapor pressure 10.7 kPa (25 °C).
[c] Density 0.8319 g/cm^3, vapor pressure 0.3 kPa (25 °C).
[d] Moles per kilogram solvent

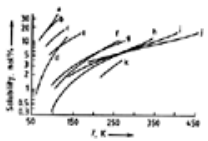

Figure 6. Solubility of hydrogen in various liquids (hydrogen pressure 10 MPa)

a) N_2; b) CO; c) Ar; d) O_2; e) CH_4; f) C_2H_4; g) C_2H_6; h) C_3H_6; i) C_4H_{10}; j) C_6H_{14}; k) CO_2

The water solubility has a minimum at ca. 50 °C and then increases again with increasing temperatures (see Table 7). Electrolytes lower the water solubility because of the salt effect.

Table 7. Solubility of hydrogen in water at a hydrogen pressure of 101.3 kPa [17]

T, K	Solubility, 10^3 mol %	Ostwald – coefficient $\times10^2$
273.15	1.755	2.184
283.15	1.576	2.032

293.15	1.455	1.940
303.15	1.377	1.893
313.15	1.330	1.883
323.15	1.310	1.905
333.15	1.312	1.957
343.15	1.333	2.037
353.15	1.371	2.144

Because of the reduced volatility, deuterium, tritium, and their various interisotopic compounds have a somewhat greater solubility than the comparable hydrogen compounds. The relative differences in solubility amount to a few percent only (close to the accuracy of measurement) and diminish with increasing temperature.

Solubility of Hydrogen in Metals. Metals, with not completely occupied inner electron levels, have a particularly high solubility potential for hydrogen. Hydrogen is dissolved in metals not in the molecular but in the atomic form, whereby *non-stoichiometrical* (alloy-type) or *stoichiometrical* (partly hydride-type) compounds may be formed.

The solubility of hydrogen in metals at small hydrogen concentrations is proportional to the root of the hydrogen partial pressure (Sievert's law). The hydrogen solubility decreases with increasing temperature. Hydrides can be decomposed thermally.

Palladium and palladium – silver alloys with a silver content <40 wt % have a particularly high solubility potential for hydrogen. The hydrogen atoms give their electrons to the electron gas of the palladium metal and occupy interstitial sites of the metal crystal structure. Palladium can absorb 2800 times its own volume of hydrogen. Figure 7 shows isotherms for hydrogen in palladium. They are similar to the Van der Waals isotherms of real gases. For an explanation of the so-called plateau dissociation pressure see Section Hydrogen-Energy Conversion Systems.

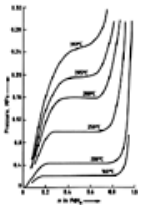

Figure 7. Pressure – composition isotherms for hydrogen in palladium [19]

The solubility of the hydrogen isotopes in palladium decreases in the order H > D > T; with group 5 metals (V, Nb, Ta), however, the solubility increases in this order. This effect is explained by differences in the vibration energies of the metal structure (in palladium: octaeder holes are occupied; in group 5 metals: tetraeder holes are occupied).

Impurities in the metal and in the hydrogen can lower the solubility or cancel it completely. Especially carbon monoxide, hydrogen sulfide and oxygen are poisonous with this respect.

Dissolved hydrogen in metals influences their mechanical properties (hydrogen embrittlement, see Section Materials), magnetic properties (reduction of the paramagnetism of Cr and Ti) and can cause superconductivity. For example, nonsuperconductive palladium becomes superconducting if hydrogen is added, to the extent of $PdH_{0.8}$. The transition temperature of $PdH_{1.0}$ is 8.8 K, that of $PdD_{1.0}$ 10.7 K, whereas $Pd_{0.55}Cu_{0.45}H_{0.7}$ even has a transition temperature of 17 K.

The absorption of hydrogen in metals and metal alloys has become important for the storage of hydrogen in hydride storage systems. An overview of hydrogen in metals is given in [20] and [21] (⟶ Hydrides, see Section Hydrogen-Energy Conversion Systems).

Solubility of Gases in Liquid Hydrogen. With the exception of helium all gases are solid at the temperature of liquid hydrogen. In liquid hydrogen they are soluble only in traces: oxygen (32 – 26 K) <0.5 ppm; nitrogen (32 K) ca. 15 ppm, (20.4 K) <0.1 ppm.

The solubility of helium in liquid hydrogen at 1 MPa is shown in Figure 8: In the range of 16 – 26 K it increases from 0.5 to 3 mol % and then decreases at higher temperatures.

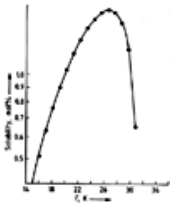

Figure 8. Solubility of helium in liquid hydrogen (pressure 1 MPa)

• liquid concentration He

Vapor – Liquid Equilibria. Various vapor – liquid equlibrium systems exist at low temperature. However, only the isotope mixtures, e.g., H_2 – HD, or H_2 – D_2 show simple boiling point diagrams with only little deviation from the ideal behavior. The vapor pressures of the mixture H_2 – HD, e.g., deviate 4 % at the most from that of an ideal mixture [22], [23].

Two- and multicomponent mixtures with nitrogen, carbon monoxide and methane are important in low-temperature processes for separation or purification of hydrogen. In these mixtures the hydrogen in the hydrogen-poor liquid phase is supercritical. Low hydrogen contents in the coexistent liquid can only be reached in the neighborhood of the melting point of the non-hydrogen components.

Figure 9 shows the hydrogen equilibrium concentrations in the liquid and the gas phase in the systems H_2 – CO and H_2 – N_2 [17], [18].

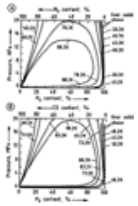

Figure 9. Vapor – liquid equilibria of the systems H_2 – N_2 (A) and H_2 – CO (B)

Isotherms show (clockwise from left to right) the composition of the liquid phase, the critical pressure at the maximum of the isotherm, and the composition of the gas phase as a function of pressure (concentrations in mol %)

The system H_2 – Ne shows a complicated behavior. Figure 10 shows a *p*-*x*-diagram for the composition range in which a miscibility gap in the liquid phase occurs: If a mixture with the composition x_1, is compressed isothermally, first two two-phase regions, each with a different liquid-phase composition are found. Then a one-phase region (a hydrogen-rich liquid phase) is reached. At the composition x_2 only one two-phase liquid system exists. At the composition x_3 an azeotropic mixture is formed, the liquid phase has the same composition as the gas phase. Equilibrium compositions of the system H_2 – Ne are given in [16].

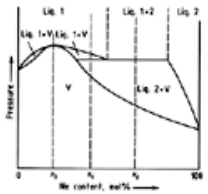

Figure 10. Sketch of a typical isotherm in the range of existence of a miscibility gap of the system n-H_2 and Ne [16]

(V = vapor phase, liq. 1 = hydrogen-rich liquid phase; liq. 2 = neon-rich liquid phase)

Diffusion Properties. Hydrogen is the element with the highest diffusion capacity. Some diffusion coefficients of hydrogen in gases and liquids are given in Table 8. For further values see [1], [4], [24], [25]. In metals with a high hydrogen solubility usually a high hydrogen diffusion coefficient is also found, e.g., in palladium 5×10^{-7} cm^2/s and in vanadium 5×10^{-5} cm^2/s (at 25 °C). These values are in the same range as the diffusion coefficient of hydrogen ions in water (ca. 10^{-4} cm^2/s).

Table 8. Diffusion coefficients of hydrogen in gases (pressure 100 kPa) and liquids

Diffusion in	*t*, °C	*D*, cm^2/s

N_2	0	0.674
O_2	0	0.701
H_2 (selfdiffusion)	0	1.285
H_2O, vapor	0	0.759
H_2O, liquid	25	4.8×10^{-5}
Iron, smelting	1600	5.64×10^{-3}
Aluminum, smelting	960	1.28×10^{-5}

2.2. Chemical Properties

The first electron shell can be filled with a maximum of two electrons. Therefore, the chemistry of hydrogen depends mainly on three processes, (1) loss of the valency electron to yield the hydrogen ion, H^+; (2) gain of an electron to form the hydride ion, H^-, and (3) formation of an electron pair bond. The degree of covalent: ionic character of this bond depends on the electronegativity of the element to which hydrogen is attached. The electronegativity of hydrogen according to PAULING is 2.1. Except in the hydrogen molecule itself, where the bond is homopolar, all H – X bonds will possess some polar character.

Hydrogen is not exceptionally reactive, although hydrogen atoms react with one another and with all other elements with the exception of the noble gases (only with helium short-lived instable forms such as HeH^+ and HeH_2^{2+} are known). Hydrogen oxidizes less electronegative elements (e.g., alkali and alkaline earth metals), and reduces more electronegative ones (e.g., halogens, oxygen, nitrogen, carbon).

The strength of the H – X bond in covalent hydrides depends on the electronegativity and size of the element X. The bond strength decreases in a group with increasing atomic number and generally increases across any period. The most stable covalent bonds are those formed between two hydrogen atoms (or isotopes), or with halogens, oxygen, carbon, and nitrogen. The average values for the bond energy are given in Table 9.

Table 9. Average bond-dissociation energies of representative covalent hydrogen bonds at 298 K

Bond	kJ/mol	Bond	kJ/mol
H–H	428.2	H–N	314
H–D	440	H–N=	435.4
D–D	443.4	H–O	428.2
T–T	446	H–OH	498.1
$H–CH_3$	435.3	H–S	344.5
H–C=	452	H–Cl	431.6
H–C≡	536	H–F	568.9

Halogens react with hydrogen to yield hydrogen halides, with increasing reactivity in the sequence iodine, bromine, chlorine, and fluorine. At –210 °C liquid fluorine ignites immediately in hydrogen; solid fluorine explodes violently in the presence of liquid hydrogen; similarly, combined hydrogen reacts more or less violently with fluorine. Chlorine and bromine also react explosively with hydrogen on excitation by heat or ultraviolet radiation (chlorine hydrogen reaction). Fluorine, chlorine, and bromine react with hydrogen in a radical chain reaction:

initiation reaction

$$\mathrm{Hal_2} \rightarrow 2\,\mathrm{Hal}\cdot$$

propagation reactions

$$\mathrm{Hal}\cdot + \mathrm{H_2} \rightarrow \mathrm{H}\cdot + \mathrm{HalH}$$

$$Hal_2 + H\cdot \rightarrow HHal + Hal\cdot$$

Oxygen reacts with hydrogen according to

$$H_2 + 1/2\,O_2 \rightarrow H_2O \quad \Delta H^0_{298} = -286.02\,kJ/mol$$

Above 550 °C the reaction occurs with flame propagation, explosion or detonation (oxyhydrogen reaction). The flame temperature is limited by the thermal dissociation of water vapor and attains a maximum of 2700 °C. For the explosion limits in air or oxygen see Section Ignition and Detonation Performance.

At elevated temperatures in presence of suitable catalysts, hydrogen will react with *nitrogen* to form ammonia. Industrially this is one of the most important reactions (⟶ Ammonia).

Hydrogen reacts with *carbon* at high temperatures to yield methane. The equilibrium of the reaction is on the methane side at low temperatures, and on the hydrogen and carbon side at high temperatures. The equilibrium constant, using graphite as a basis, is 2×10^5 bar at 150 °C and 9.4×10^{-3} bar at 1000 °C. Equilibria related to amorphous carbon and other values are given in [26], [27]. Carbon monoxide and carbon dioxide can react with hydrogen in the presence of a catalyst. Depending on the reaction conditions, catalyst and CO : H_2 ratio a variety of products can be formed (⟶ Gas Production, see Section Synthesis Gas).

Unsaturated hydrocarbons are converted to saturated or partially saturated hydrocarbons by hydrogenation.

Further reactions of hydrogen include reductions of functional organic groups ($–NO_2$ ⇒ NH_2, –COOH ⇒ CH_2OH, etc). In the petroleum industry (⟶ Oil Refining) hydrogenation and hydrocracking reactions are of importance as well as desulfurization and selective hydrogenation (dienes, monoolefins). For details of reduction and hydrogenation with hydrogen of other compounds see the appropriate keywords.

A great number of *metals* react with hydrogen to form hydrides. Depending on the nature of the hydrogen bond, hydrides are classified into three principal categories: saline, covalent and metallic (see surveys in [20], [21]), (⟶ Hydrides).

The *saline hydrides* or ionic hydrides are formed when the strongly electropositive alkali metals and the alkaline earth metals (calcium, strontium, barium), all usually in the elemental form, react with hydrogen at high temperatures. Salt-like hydrides contain the hydride ion, H^-; they are crystalline, have high heats of formation and high melting points. They are reactive, powerful reducing agents, all of them reacting with water to liberate hydrogen.

Covalent hydrides may be either solid, liquid or gaseous; typical are the hydrides of boron, aluminum, silicon, germanium and tin. Beryllium and magnesium hydride seem to represent a transition between ionic and covalent hydrides. The bond between hydrogen and the atom is of the nonpolar electron-sharing type. Molecular stability often is provided by means of three-centered two-electron bridge bonds (e.g., in diborane, B_2H_6, an electron-deficient type molecule), by hydride anion formation (BH_4^-, AlH_4^-), or formation of polymeric structures [polymeric $(AlH_3)_x$].

Metallic hydrides are formed by the transition metals. Transformation of a pure metal into a hydride occurs through continuous solution of hydrogen in the metal with subsequent abrupt phase transition at defined stoichiometric hydride phases (ZrH_2, PdH, VH, VH_2).

The Hydrogen Bond. Substances containing hydrogen coupled to the most electronegative elements exhibit properties that are best explained by assuming that the hydrogen atom of an H – X bond still has a small but significant affinity for other electronegative atoms although it remains strongly bonded to the original atom. This relatively weak secondary bond is called a hydrogen bond normally depicted by the notation

$$X-H\cdots\cdots Y$$

Hydrogen bonds result in molecular association. NH_3, H_2O, and HF have much higher boiling points and heats of vaporization than the corresponding homologues PH_3, H_2S, and HCl. Carboxylic acids such as acetic or benzoic acids are present as dimers with the configuration

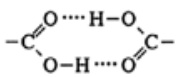

The presence of hydrogen bonds also causes formation of crystalline hydrates such as $NH_3 \cdot H_2O$, $SO_2 \cdot H_2O$ hydrocarbon hydrates (e.g., with natural gas components) as well as the zig-zag structure of the $(HF)_n$ polymer.

The existence of the hydrogen bond can be shown by infrared spectroscopy. When the hydrogen of the X – H group forms a hydrogen bond with the atom Y, the X – H stretching frequency is lowered and the absorption band characteristically broadened. In NMR spectroscopy a proton, which participates in a hydrogen bond, produces a definite paramagnetic shift.

The energies of hydrogen bonds are in the range of 4 – 40 kJ/mol, which is small as compared to 200 – 400 kJ/mol for ordinary bonds. The average bond length is 200 – 300 pm (O – H . . . O in ice 276 pm).

Atomic Hydrogen. Atomic hydrogen is much more reactive than the molecular form. The dissociation reaction

$$H_2 \rightarrow 2\,H;\;\; \Delta H^0 = 432.2\,kJ/mol$$

is highly endothermic. Atomic hydrogen can be produced thermally by supplying sufficient energy, e.g., in electric arcs with a high current density, electrical discharges at low pressures, irradiation with UV light as well as electron bombardment (10 – 20 eV). The hydrogen atom has a short half-life (ca. 0.3 s) and reacts even at room temperature with nonmetallic elements such as halogens, oxygen, sulfur, and phosphor to form the corresponding hydrogen compounds. At room temperature it reduces many oxides to the elements. The heat of recombination of hydrogen molecules from hydrogen atoms leads to very high temperatures (Langmuir flare, 4000 K).

Atomic hydrogen can be absorbed in the metallic structure of some elements of the Group 8 – 10 (Fe, Co, Ni) of the periodic system. The suitability of many catalysts for reactions involving hydrogen is based on the dissociation and solubility of hydrogen in the atomic form.

Hydrogen formed upon dissolution of metals in acids is very reactive at the moment of formation (nascent hydrogen), much more than normal hydrogen. This strong reduction capability is often explained by a shortterm formation of atomic hydrogen. Another explanation is the intermediate formation of a hydride-type compound which is very similar to Grignard compounds (⟶ Magnesium Compounds).

The Hydrogen Ion. The ionization potential of the reaction

$$H(g) \rightarrow H^+(g) + e^-;\;\; \Delta H = 1310\,kJ/mol$$

is 13.595 eV, which is even higher than the first ionization potential of the noble gas xenon.

The hydrogen ion can be formed only in a medium which solvates the protons. The solvation process provides the energy required for bond rupture, the order of magnitude can be seen in the solvation reaction with water:

$$H^+(g) + xH_2O \rightarrow H^+(aq);\;\; \Delta H = 1070\,kJ/mol$$

In aqueous systems H_3O^+ is the form implied by the customary designation hydrogen ion. The structure of the H_3O^+ ion is that of a rather flat triangular pyramid with a HOH bond angle of ca. 110^0, O – H bonds distances of 102 pm and a H – H distance of 172 pm. Normally, it is coordinated with four water molecules not all of which are equivalent, so that the aquated hydrogen ion is best represented as $H_9O_4 \cdot H_2O$ [28]. The lifetime of an individual H_3O^+ ion is exceedingly short, 10^{-13} s, because the protons are rapidly exchanged among the water molecules.

Hydrogen Electrode. In aqueous solutions the hydrogen ion activity is given in terms of the pH value, defined as log 1/[H^+], where [H^+] is the hydrogen ion activity. By definition, the potential of the redox system

$$1/2\,H_2(g) \rightarrow H^+(aq) + e^-$$

is zero (*E* = 0.000 V) at all temperatures, when an inert metallic electrode is immersed in a solution of unit activity (i.e., pH = 0) in equilibrium with hydrogen gas at 100 kPa pressure. The electrode material is a small platinum foil or a wire, coated with finely dispersed platinum black and is placed partly in the solution and partly in the hydrogen atmosphere. This standard hydrogen electrode is the reference electrode of all other oxidation-reduction systems.

For pH measurement the potential difference of a hydrogen electrode to any other suitable reference electrode can be used. The calomel electrode, or half-cell, is usually employed. The pH value is given by

$$pH = \frac{(\Delta\varepsilon - \varepsilon_v)}{0.059} - \frac{1}{2}\log p_{H_2}$$

where: $\Delta \in$ is the potential difference between the hydrogen and the reference electrode (in Volt), $\in_v$ the potential of the reference electrode (in Volt), and p_{H2} the hydrogen partial pressure (in bar) [29], [30].

2.3. Ignition and Detonation Performance

Given sufficient thermal activation (ignition energy), hydrogen reacts more or less violently with oxidizing agents such as oxygen (air), fluorine or chlorine, and N_2O. Depending on conditions, combustion, deflagration or detonation may occur.

Ignition and detonation properties of *hydrogen – air mixtures* are particularly important from the safety aspect (see also Section Safety Aspects). The flammability limits (i.e., the difference between the rich and lean limit concentration) are exeptionally wide for hydrogen. Table 10 shows a comparison of safety-relevant thermo-physical and combustion properties of hydrogen with those of methane, propane

and gasoline [31]. The flammability limits are affected by temperature as shown in Figure 11, so that a preheated mixture has considerably wider limits for coherent flames [32]. An increase in pressures up to 10 kPa has only a small effect. Water vapor has a strongly inhibiting influence on the oxyhydrogen reaction (see ternary diagram, Figure 12). The flammability limits of hydrogen in *pure oxygen* are 4.65 – 93.9 vol % hydrogen, those of deuterium in pure oxygen 5.0 – 95 vol %. Ignition and detonation performance based on ignition range, ignition energy, flame propagation as well as transition from deflagration to detonation, are discussed in detail in Section Safety Aspects.

Table 10. Combustion and explosion properties of hydrogen, methane, propane and gasoline

	Hydrogen	Methane	Propane	Gasoline
Density of gas at standard conditions, kg/m^3 (STP)	0.084	0.65	2.42	4.4 [a]
Heat of vaporisation, J/g	445.6	509.9		250 – 400
Lower heating value, kJ/g	119.93	50.02	46.35	44.5
Higher heating value, kJ/g	141.8	55.3	50.41	48
Thermal conductivity of gas at standard conditions, mW cm^{-1} K^{-1}	1.897	0.33	0.18	0.112
Diffusion coefficient in air at standard conditions, cm^2/s	0.61	0.16	0.12	0.05
Flammability limits in air, vol %	4.0 – 75	5.3 – 15	2.1 – 9.5	1 – 7.6
Detonability limits in air, vol %	18.3 – 59	6.3 – 13.5		1.1 – 3.3
Limiting oxygen index, vol %	5	12.1		11.6 [b]
Stoichiometric composition in air, vol %	29.53	9.48	4.03	1.76
Minimum energy for ignition in air, mJ	0.02	0.29	0.26	0.24
Autoignition temperature, K	858	813	760	500 – 744
Flame temperature in air, K	2318	2148	2385	2470
Maximum burning velocity in air at standard conditions, m/s	3.46	0.45	0.47	1.76
Detonation velocity in air at standard conditions, km/s	1.48 – 2.15	1.4 – 1.64	1.85	1.4 – 1.7 [c]
Energy [d] of explosion, mass-related, gTNT/g	24	11	10	10
Energy [d] of explosion, volume-related, gTNT/m^3 (STP)	2.02	7.03	20.5	44.2

[a] 100kPa and 15.5 °C.
[b] Average value for a mixture of C_1– C_4 and higher hydrocarbons including benzene.
[c] Based on the properties of *n*-pentane and benzene.
[d] Theoretical explosive yields.

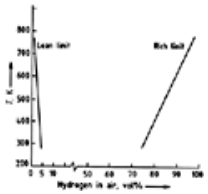

Figure 11. Effect of temperature on flammability limits of hydrogen in air (pressure 100 kPa)

Figure 12. Flammability and detonability limits of the three component system hydrogen – air – water vapor [33]

a) 42 °C, 100 kPa; b) 167 °C, 100 kPa, c) 167 °C, 800 kPa

Hydrogen reacts with chlorine after excitation with radiation, heat or in the presence of catalysts. The ignition and explosion behavior of this reaction is of interest for the chlor-alkali electrolysis, because chlorine is usually liquefied after formation. The residual gas contains hydrogen (among other components) and might reach the ignition or detonation range because of the decreasing chlorine concentration in the gas phase during chlorine liquefaction.

The lower ignition limit of hydrogen-chlorine air mixtures is ca. 4 – 6.5 vol % H_2, the upper ignition limit lies between 73 vol % H_2 (Cl_2 concentration = 0 vol %) and 89 vol % H_2 (air concentration = 0 vol %). The lower detonation limit is between ca. 20 vol % H_2 (Cl_2 concentration = 0) and ca. 25 vol % H_2 (Cl_2 concentration = 30 – 40 vol %). An overview of the ignition and detonation properties of hydrogen – chlorine mixtures, and the influence of oxygen and inert gas is given in [34].

2.4. Isotopes

Hydrogen and deuterium are stable isotopes; tritium emits a low-energy β-radiation (disintegration into ${}^{3}_{2}He$). The half-life is 12.26 a. The natural isotope distribution on earth is 99.985 atom % H and 0.0145 atom % D. Natural tritium occurs in the upper atmosphere and is caused by interaction of cosmic rays, mainly according to:

$$ {}^{14}_{7}N + {}^{1}_{0}n \rightarrow {}^{3}_{1}T + {}^{12}_{6}C $$

Other tritium sources are nuclear reactors, where tritium is formed from lithium or D_2O by a neutron capture mechanism:

$$ {}^{6}_{3}Li + {}^{1}_{0}n \rightarrow {}^{3}_{1}T + {}^{4}_{2}He $$

$$ {}^{2}_{1}D + {}^{1}_{0}n \rightarrow {}^{3}_{1}T $$

Tritium is also formed by fission processes, e.g., in nuclear experiments. Tritium disintegration and varying formation rates lead to changes in the tritium content. It is estimated to be ca. 10^{-13} – 10^{-12} atom % of the total hydrogen [35]. An overview is given in [36] and [37].

Physical Properties. The most important atomic and molecular data of hydrogen, deuterium, tritium and their diatomic combinations are given in Table 1. Further thermodynamic data are tabulated in [38] for deuterium and [39] for tritium. Hydrogen, deuterium, and tritium show nuclear spin isomerism (see Section General Physical Data).

Figure 13 shows the vapor pressures of hydrogen molecules with a different isotopic composition. In [40] the system D_2 – DT – T_2 is described in the range 17 – 22 K. The compounds are totally miscible in the liquid and in the solid state. A 50 : 50 mol % mixture of D_2 – T_2 has a T content of 42 % in the gas phase and of 52 % in the solid phase at 19.7 K.

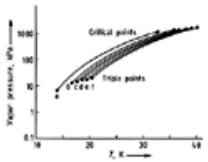

Figure 13. Vapor pressure of hydrogen isotopes a) H_2; b) HD; c) HT; d) D_2; e) DT; f) T_2

Various chemical and physical processes are used for the enrichment and separation of D_2 or D_2O from the natural isotopic mixture. All chemical processes are based on isotopic exchange reactions, e.g., the dual temperature H_2O – H_2S process, according to

$$ H_2O + HDS \rightleftharpoons HDO + H_2S $$

in the ammonia exchange process

$$NH_3 + HD \rightleftharpoons NH_2D + H_2$$

and variations of this process such as the methylamine hydrogen process from Aecl-Sulzer according to

$$CH_2NH_2 + HD \rightleftharpoons H_2 + CH_3NHD$$

The processes are described in [36] as well as in ⟶ Isotopes, Natural.

On a small scale the electrolysis of enriched H_2O – D_2O is used, because this gives an isotopic separation factor of 7 – 8.

Low-temperature destillation of H_2 – D_2 is an example of the physical separation of the isotopes [41]. The heat generated by the ortho-para conversion must be taken into consideration. Other processes use *pressure swing adsorption* techniques [42]; the various permeation rates of H_2 and D_2 through *metal membranes* (Pd, Ta); or differences in the hydride or deuteride formation with metals [43], (see Section Hydrogen-Energy Conversion Systems).

In general, the *chemical properties* of H, D and T are essentially identical. Small differences exist in the rates and the equilibrium constants of reactions because of the kinetic isotope and equilibrium effects. A greater amount of energy is required for the dissociation of a D – X bond and the activation into the transition state than for the comparable H – X bond (for numerable investigations see [44]). In biological systems substitution of deuterium for hydrogen atom can alter the delicately balanced processes substantially.

Deuterium has excellent properties as a moderator (i.e., a high ratio of neutron deceleration to absorption characteristics). The separation of the hydrogen isotopes is technically important for the production of heavy water (D_2O) as a moderator for nuclear reactors (⟶ Nuclear Technology).

Tritium also is used to mark organic compounds and for dating specimens. A detailed description of tritium and its compounds is found in [45].

[Top of Page]

3. Occurence

Hydrogen is the most abundant element in the universe. 70 – 80 wt % of the observable universe consists of hydrogen, the rest is mostly helium. Especially the younger stars (the novas) inter-stellar dust, and gas clouds contain hydrogen in large amounts. More than 50 wt % of the sun consists of hydrogen.

The neutral hydrogen atom which emits electromagnetic waves with a frequency of 1420 MHz, yields information on the structure of the universe. Large regions of interstellar space are filled with protons, which emit a radio continuum [46], [47].

Of the elements in the earth's crust, hydrogen takes the 9th place with 0.88 wt %. Based on the atomic concentration of 15.4 atom % it is the third most common element after oxygen and silicon. In the free state, hydrogen can be found in some vulcanic gases (up to ca. 30 vol %), in natural gas (from traces up to 40 vol % [48]); small quantities are trapped in some minerals, rocks and meterorites. In the upper atmosphere, the troposphere and the stratosphere, hydrogen is present only in traces of ca. 0.5 ppm (mL/m^3). Because of photolytic water cleavage the hydrogen concentration increases with increasing height. The exosphere (2000 – 20 000 km) contains atomic hydrogen.

Chemically bound hydrogen is present all over the earth: as part of the Earth's water mass (11.2 wt %), as part of organic matter (including fossil substances such as natural gas, oil, and coal). About 10 wt % of the human body consists of hydrogen.

[Top of Page]

4. Production

Hydrocarbons are the main source for the production of hydrogen on an industrial scale. An even larger source, however is water. When hydrogen is used as fuel, water is produced again. Thus, a cycle comparable to the carbon cycle (CO_2 – biomass – CO_2) —but much simpler—is involved, so that water can be regarded as an unlimited source of hydrogen. With the help of an external energy source, hydrogen can be produced from either of these sources alone or from a mixture of both.

Figure 14 shows the production of hydrogen using different sources of energy. The many production routes for hydrogen, however, are by no means of equal economic importance. Most of the hydrogen for industrial uses is produced from natural gas and oil, either as a main product or as a byproduct, a process involving a chemical conversion. A smaller percentage is produced electrolytically or occurs as a byproduct of electrochemical processes. Here the energy applied is in the form of electricity, which itself may be produced using hydroelectric, nuclear or, still predominant, the energy stored in fossil fuels.

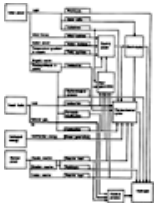

Figure 14. Primary energy, raw materials, production methods, paths of production, transformation, and generation of and for hydrogen, adapted from [49]

Thermal, thermochemical, biochemical and photochemical processes have so far not found industrial applications. The preference of producing hydrogen from hydrocarbons is due to the fact that the energy demand using hydrocarbons is much lower than that for electrolysis of water (see Table 11).

Table 11. Theoretical energy consumption for the production of hydrogen from various hydrocarbons, coal and water

	Natural gas as CH_4	LPG as $CH_{2.6}$	Naphtha as $CH_{2.2}$	Heavy oil as $CH_{1.4}$	Coal as $CH_{0.7}$	Water H_2O
Higher heating value, *						
kJ/kg	50 050	46 130	44 300	41 300	38 100	
kJ/mol	890	2480				
Byproducts	CO_2	CO_2	CO_2	CO_2/S	CO_2/S	O_2
Specific production of byproducts, kmol/kmol H_2	0.25	0.30	0.32	0.37/0.003 – 0.015	0.43/0.002 – 0.01	0.5
Theoretical energy consumption, * kJ/kmol H_2	**41 280**	**37 500**	**38 350**	**50 300**	**57 150**	**242 000**

* Calculated from the heat of formation, oxidizing carbon completely to carbon dioxide (at 298 K and 0.1 MPa) [73], [78].

Considering the strong correlation between energy prices and the prices for hydrocarbons, it is unlikely that the high percentage of the hydrocarbon processes for producing industrial hydrogen will change substantially in the near future. This last statement does not take into consideration environmental pollution by emission of carbon dioxide, sulfur compounds, nitrogen oxides and heat, so that this (environmental) aspect may reverse the situation one day.

4.1. Production from Coal and Hydrocarbons

The yield of hydrogen from hydrocarbons is summarized in Figure 15. Hydrocarbons and coal are characterized by their molar H : C ratio. Chemically pure compounds such as methane and its homologues can be found on line (A) and give maximum hydrogen yield. Raw materials such as hard coal (a), lignite (b) and natural gas (h) give a lower hydrogen yield because they contain inert components (sulfur, nitrogen, oxygen or minerals etc.). Other important sources for the production of hydrogen such as vacuum residue (c), naphtha (d) or hydrogen-rich off-gases (e.g. coke oven gas (j), are also shown in Figure 15.

Figure 15. Hydrogen production from hydrocarbons and coal

A) Maximum theoretical yield from (a) hard coal, (b) bituminous coal, (c) vacuum residue, (d) naphtha, (e) butane, (f) propane, (g) ethane, (h) natural gas, (i) methane, (j) coke oven gas

B) Commercial production by steam-reforming, including fuel, no credit for export steamLower curve: small plants, ca. 500 m^3 (STP)/hUpper curve: large plants

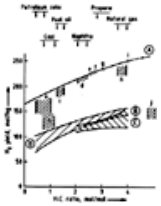

C) Commercial production by partial oxidationlarge plants only

Upper curve: without oxygen production

The yields actually achieved in industrial hydrogen production are quite different from the maximum possible theoretical yields. This, however, is not due to a low degree of conversion but to the fact, that the energy needed to produce hydrogen is usually supplied by the raw material itself. Technical and economic aspects of the production are dealt with in this chapter. For the latter not only the yield but also the prices of the feedstock and the investment costs for the plant are important. Because of the fact that (1) there are only small differences between the cost of hydrocarbons on the basis of heating value, (2) the investment costs for the production of hydrogen from light hydrocarbons are lower, and (3) yields from light hydrocarbons are higher, as shown in Figure 15, hydrogen is produced preferably from light hydrocarbons, if available.

4.1.1. *Coke Oven Gas (see also* ⟶ Coal Pyrolysis*)*

Coke oven gas (COG) was formerly a major source of hydrogen. Coke oven gas provided, in conjunction with water gas, almost all of the hydrogen for ammonia synthesis and for coal liquefaction (⟶ Coal Liquefaction) in the first half of this century. While the demand for hydrogen has risen sharply since 1950, the production of coke has been reduced. New production techniques for steel as well as reduced steel production in most industrialized countries have led to a decreasing market for coke and COG. Producing iron by direct reduction of iron oxide pellets with synthesis gas, for example, requires no coke at all. Nevertheless COG is, where available, an important and convenient source of hydrogen, but on the whole, production from coke oven gas has now been replaced by more effective hydrocarbon-based processes and by coal gasification.

Modern coke oven plants produce ca. 350 m^3 (STP) of coke oven gas per tonne of coal (15.6 mol/kg). 45 % of the raw COG produced is utilized to fire the coke oven batteries and supply other energy requirements of the coke oven plant. The rest can be used as a source of hydrogen or exported as a fuel gas. It can be expected that further developments and better heat management within the coke oven plant will raise the amount of excess gas up to 70 % of the total gas produced [50].

Raw Gas Treatment. Coke oven gas leaves the coke oven batteries at 750 – 850 °C and approximately atmospheric pressure. The main impurities which have to be eliminated are tar, several heavy to medium oil fractions with a high aromatic content, hydrogen cyanide, ammonia, organic and inorganic sulfur compounds, naphthalene, nitrogen oxides and water. The concentrations of the impurities vary with the kind of coal and the type and duration of the coking process. Several proprietary techniques for cleaning of raw COG are available [51], [52]. Typical cleaning steps are:

1. Quench cooling to ambient temperature (15 – 30 °C) and separation of the oil and naphthalene fractions,
2. First raw gas compression,
3. Absorption of hydrogen cyanide, ammonia, and sulfur compounds,
4. Second raw gas compression

Pretreated COG contains 50 – 60 % hydrogen (see Table 12).

Table 12. Range of composition for precleaned coke oven gases and production of hydrogen from typical gas by pressure swing adsorption

Components	Range of composition, vol %	Typical composition, vol %	Hydrogen product, vol %	Fuel gas, vol %
H_2	50 – 64	57	99.9*	22
N_2	3 – 12	3	0.09	5.5
CO	4 – 8	6.5	<10 ppm	12
CO_2	1 – 5	3	< 1 ppm	5.5
CH_4	24 – 27	26	0.01	47
C_2H_4	1 – 3	2		4
C_2H_6	0.7 – 2	1		2
C_3H_6	0.3 – 0.6	0.5		1

C_6H_6	0.5 – 1	0.8		1
O_2	0.1 – 0.5	0.1	<10 ppm	
H_2S + COS	0.1 – 0.8	0.1	<10 ppm	
Organic sulfur, g/kmol	0.5 – 10	2		
Total		100	100	100
LHV, kJ/mol	370 – 460	440	245	588
HHV, kJ/mol	420 – 520	490	286	645
Typical flows with pressure swing adsorption, m^3 (STP)/h		100	45	55

* Possible up to 99.999 %.

Separation of Hydrogen from Pretreated COG. The separation of hydrogen from pretreated COG has been accomplished in two different ways.

Low-Temperature Condensation. Conventionally, hydrogen is separated by a low-temperature condensation process. Condensation of methane, ethylene and other compounds with a higher boiling point is carried out in regenerators (large vessels filled with ceramic materials) at elevated pressures (0.5 – 1 MPa) and low temperatures (60 – 70 K).

The hydrogen purity attainable is 96.5 %. Utilizing the low temperatures, hydrogen can be further cleaned in a liquid nitrogen scrubber (see Section Low-Temperature Absorption Processes), where the residual methane, carbon monoxide and oxygen content can be reduced down to low levels. In the nitrogen scrubber, the gas can be conditioned for ammonia synthesis gas production (25 % N_2, 75 % H_2).

By fractionating the primary condensate from the regenerators ethane, ethylene, propane, and fuel gas can be obtained as byproducts along with hydrogen or ammonia synthesis gas.

Plants with capacities up to 3×10^5 m^3 (STP)/h have been built using this type of technology.

Pressure Swing Adsorption. Pure hydrogen can be recovered at much lower investment costs in modern pressure-swing adsorption (PSA) plants (⟶ Adsorption – Regeneration by Pressure Swing). A first industrial-scale plant with 10 000 m^3 (STP)/h capacity was put into operation in the German Ruhrgebiet in 1982. Figure 16 shows the principal steps of the plant.

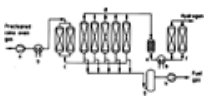

Figure 16. Hydrogen production from coke oven gas by pressure swing adsorption (Linde, Federal Republic of Germany)

a) Raw gas compressor; b) Cooler; c) Preadsorption (regeneration not shown); d) Pressure swing adsorption; e) Oxygen removal; f) Water removal; g) Rest gas equalizer; h) Fuel gas compressor

After compression, higher hydrocarbons and sulfur compounds are separated from the raw gas by a preadsorption stage. Separation of hydrogen from the other gas components occurs in an adsorption process with four or more beds (see Section Adsorption Processes).

Special precautions must be taken to avoid accumulation of oxygen on the adsorbent. Part of the oxygen remains in the product gas and is converted to water in a special reactor (see Section Selective Oxidation for Deoxo reaction). Hydrogen is then dried in another adsorption process yielding a purity >99.9 %, before it is delivered to a hydrogen pipeline system (see Chap. Handling of Hydrogen).

About 80 % of the hydrogen is recovered by this process. Thus, depending on the initial composition, the gas volume is reduced by 40 – 50 % on producing pure hydrogen. However, the residual gas, which is usually used for firing the coke oven batteries or otherwise as a fuel gas, has a lower heating value (LHV) of ca. 587 kJ/kmol so that recovery of the valuable hydrogen product stream leads only to a 20 – 30 % drop in the total heat content of the remaining fuel gas (see Table 12).

Proposed Technology for Production of Hydrogen from COG by Chemical Processes. A higher hydrogen yield could be obtained, if the hydrocarbons in the raw or pretreated COG were converted to hydrogen.

Catalytic processes for this purpose are *steam-reforming* and *autothermic* processes (see Section Catalytic Reforming of Hydrocarbons). Both processes run best when the sulfur content is low and unsaturated and cyclic hydrocarbons are absent. This requires multi-stage precleaning processes, which are technically feasible and often used for other applications (⟶ Gas Production). For COG, however, they have not yet been commercially applied so far though they have been frequently proposed [53], [54].

Partial oxidation of COG (see Section Gasification of Coal and Hydrocarbons) has also been proposed [52]. But, because of difficulties with the low initial pressure (approx. atmospheric) and the periodical changes in amount and composition of COG, no commercial application has been reported.

The above refers to COG from hard coal only. Similar aspects apply when gas from *coking operations* in refineries (see Section Refinery Processes) or from modern *pyrolysis* processes of hard coal, lignite, organic materials (straw, refuse, etc.) is considered as a source of hydrogen.

4.1.2. *Gasification of Coal and Hydrocarbons*

4.1.2.1. *Coal Gasification*

For the production of hydrogen or hydrogen-rich gas streams coal gasification is usually performed with oxygen (purity > 95 %) at high temperature and elevated pressure. Coal gasification is described in detail elsewhere (⟶ Gas Production). Here, only aspects relevant to the production of hydrogen and hydrogen-rich gases are dealt with.

Coal gasification on an industrial scale produces:

1. Synthesis gases, rich in carbon monoxide and hydrogen, e.g., gas for Fischer-Tropsch synthesis of gasoline at SASOL, South Africa (⟶ Coal Liquefaction), [55];
2. Synthesis gas for the production of acetic anhydride at Tennessee Eastman, United States [56];
3. Ammonia synthesis gas: 3 H_2 + N_2, e.g., Modderfontein, South Africa [57];
4. Substitute natural gas >90 % CH_4, e.g., North Dakota, United States [58];
5. Medium- and low-Btu fuel gas for combined-cycle processes [59] consisting of hydrogen, carbon monoxide, methane, carbon dioxide. The gas may also contain considerable amounts of inerts (nitrogen, argon), if gasification is performed with air.
6. Pure hydrogen (hydrogen content > 95 %) from a side stream of a synthesis gas unit [60].

All coal gasification processes can be used to manufacture hydrogen. The highest yield in the primary gasification step, however, is achieved, if pulverized coal is gasified with oxygen at high temperatures. This is done in the entrained-bed gasification reactor, commercialized especially in the Koppers – Totzek and the Texaco process.

The most important process steps are shown in Figure 17 for the Texaco coal gasification process.

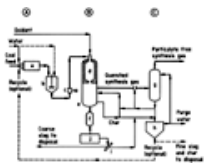

Figure 17. Texaco coal gasification process; direct quench mode of operation

A) Coal grinding and slurry preparation; B) Gasification and gas cooling; C) Gas scrubbing

a) Coal grinding mill, b) Slurry tank; c) Slurry pump; d) Texaco gasifier; e) Quench chamber; f) Lockhopper; g) Particulate scrubber; h) Clarifier; i) Slag separator; j) Slag sump

Gasification temperatures > 1300 °C result in a synthesis gas which contains only minor amounts of methane and other hydrocarbons. The chemical equilibrium at these temperatures leads to a gas containing mainly hydrogen and carbon monoxide (see Table 13). The latter can be shift converted to hydrogen (see below).

Table 13. Hydrogen production by coal gasification. Effect of gasification pressure, temperature and operation mode

	Low temperature gasification	Medium temperature gasification	High temperature gasification
Feedstock			
Coal type	North Dakota lignite	Bituminous coal	Bituminous coal
Volatile matter, wt %	26.9	34.3	33.5
Fixed carbon, wt %	30.6	50.2	41.7
Moisture, wt %	36.2	5.0	13.0
Ash, wt %	6.3	10.5	11.8

Carbon, wt % (waf*)	73.9	83.6	76.7
Hydrogen, wt %	5.2	5.7	5.9
Oxygen, wt %	18.9	6.8	12.6
Nitrogen, wt %	1.0	1.6	1.1
Sulfur, wt %	1.0	2.0	3.7
Gasification conditions			
Operating pressure, MPa	3.1	3.0	2.0 – 8.0
Operating temperature, °C	700 – 800	950	1250 – 1400
Ratio: steam: oxygen, kg/m^3 (STP)	6.9	1	1
Oxygen consumption per crude gas, m^3/m^3 (STP)	0.13	0.2	0.32
Crude gas yield, m^3 (STP)/t waf*	2000	2200	2200
Raw gas composition			
H_2, vol %	39.3	28.0	35.8
CO, vol %	15.8	56.4	44.6
CO_2, vol %	31.5	3.0	18.0
CH_4, vol %	11.8	7.1	0.0
C_{2+}, vol %	0.8	0.4	0.0
N_2/Ar, vol %	0.4	4.3	0.5
H_2S/COS, vol %	0.4	0.8	1.1
Other	ammonia, benzene, phenolics, tar, organic acids		
Commercial processes	fixed bed/slow moving bed gasification processes: Lurgi, Wellmann – Galusha	fixed bed/slow moving bed: British Gas – Lurgi, fluidized bed: Winkler Hygas,	entrained bed gasification: Texaco, Shell, Koppers – Totzek

* waf = water and ash free.

Gasification pressure affects the hydrogen yield and the economics of the process. Whereas the yield of hydrogen is slightly reduced, the economics are improved at high pressure because hydrogen is usually utilized at higher pressure. Modern gasification processes, therefore, are operated at least at 2 MPa. The negative effect on the hydrogen yield can be offset by a slight increase in temperature.

Table 13 shows the operating conditions of several commercial gasification processes. Since different coals have been gasified, the hydrogen yields cannot be compared directly. However, the influence of gasification temperature on hydrogen yield and the formation of byproducts can be clearly seen.

Principally, all types of coal can be gasified to produce hydrogen, although certain constraints associated with the gasification process or the characteristics of the coal have to be considered. The gasification of *petroleum coke* in a fluidized bed gasifier has been proposed [61] and the gasification of coke from *tar sands* has been successfully tested in an entrained bed gasifier [62].

Gasification of other solid fuels such as peat, wood, refuse, or sludge from wastewater treatment plants leads to a low-Btu gas, which is less suitable for the production of hydrogen. Because of the high

moisture and oxygen content of the raw material, these gases are generally used on the plant site for heat or electricity production. In Finland one plant is in operation since 1988 which produces ammonia synthesis gas from peat, employing a high temperature Winkler gasifier (⟶ Gas Production) [63].

Research and Development. Commercial conditions, namely the low crude oil prices in the 1980s, have prevented further development of new coal gasification processes developed in the 1970s from bench-scale experiments to commercial-scale processes. Instead research has aimed at making the established processes more reliable. Further progress has been made in improving the purification steps from the raw synthesis gas to pure hydrogen (see Chap. Purification of Hydrogen).

4.1.2.2. *Gasification of Liquid and Gaseous Hydrocarbons*

Partial oxidation of hydrocarbons in general is described in detail elsewhere (⟶ Gas Production – Partial Oxidation of Hydrocarbons). For the production of hydrogen, the Texaco and the Shell partial oxidation (PO) processes are applied commercially [64-67].

The principal steps of an oil gasification process (Texaco process) are similar to the coal gasification depicted in Figure 17. The Shell process differs in details such as the burner construction, the reactor cooling, the waste heat boiler and the soot recycle. Some more details of both processes are given below:

Feedstock	Problems
High boiling point fractions	Cracking and carbon formation on heating, high pressure drops at high viscosity
Solid residues	Slurrying hydrophobic residues, high demands on the burner
High metal content	Corrosion and metal deposition in down-stream equipment
Ash and slag content	Mechanical equipment for solid removal, melting point of the ash, eutectic formation with brick work, wear by erosion
Treatment of carbon deposits	No recycle: carbon loss, disposal problems Recycle: metal content build-up
Heat recovery	Quench: reliability, direct steam production for shift conversion applied with: H_2 production
	Waste-heat boiler: high steam production rate; applied with: synthesis gas production

Heavy residues from petrochemical processes are the preferred feedstocks for the production of hydrogen and carbon monoxide-rich gases. Oil fractions with high sulfur and/or high heavy metal contents are difficult to handle in upgrading processes such as hydrogenation or coking. For environmental reasons they should not be used as fuels. In principle, hydrogen can be produced by PO processes from all hydrocarbons. The production, however, is only economically viable if heavy residues are utilized. (see Section Economic Aspects for an economic comparison of hydrogen-producing processes).

Gaseous hydrocarbons are also occasionally converted to hydrogen and carbon monoxide in PO processes. This is done, if the gases cannot be processed by catalytic steam reforming (see Section Catalytic Reforming of Hydrocarbons) because of high concentrations of certain impurities such as olefins and H_2S or if the process must handle various feedstocks from natural gas up to fuel oils. Another incentive to use PO for the conversion of gaseous fuel is the possibility to obtain synthesis gas with a high carbon monoxide content.

The overall reactions during partial oxidation are as follows (heat of reaction see Table 14):

Oxidation reactions

$$C_mH_nS_o + m/2\,O_2 \rightarrow mCO + (n/2 - o)H_2 + oH_2S\,(\text{exothermic}) \quad (1)$$

$$2\,C + O_2 \rightarrow 2\,CO(\text{exothermic}) \quad (2)$$

Crack reaction

$$C_mH_nS_o \rightarrow (n/4 - o/2)\,CH_4 + (m - n/4 + o/2)C + oH_2S\,(\text{endothermic}) \quad (3)$$

Water gas reaction

$$C + H_2O \rightarrow CO + H_2\,(\text{endothermic}) \quad (4)$$

Shift reaction

$$CO + H_2O \rightarrow H_2 + CO_2\,(\text{exothermic}) \quad (5)$$

Methane reforming

$$CH_4 + H_2O \rightarrow \leftarrow CO + 3\,H_2 (\text{endothermic}) \quad (6)$$

Table 14. Equations and reaction enthalpies for the most frequent reactions occurring during catalytic steam reforming and their technical conditions [78]

	ΔH at 101.3 kPa, 298 K, kJ/mol	Technical reaction conditions			
		Start of reaction at t, °C	End of reaction at t, ° C	Reaction pressure, MPa	Heat of reaction ΔH, kJ/mol
Reforming reactions					
$CH_4 + H_2O \Rightarrow CO + 3\,H_2$ (7)	206.36	500	870	2.0	234.66
$CH_4 + 2\,H_2O \Rightarrow CO_2 + 4\,H_2$ (8)	165.13				
$C_2H_6 + H_2O \Rightarrow CO + CH_4 + 2\,H_2$ (9)	141.22	500	870	2.0	166.56
$C_2H_6 + 2\,H_2O \Rightarrow 2\,CO + 5H_2$ (10)	347.50	500	870	2.0	131.68
$C_3H_8 + H_2O \Rightarrow C_2H_6 + CO + 2\,H_2$ (11)	150.68	500	870	2.0	179.24
$C_3H_8 + 2\,H_2O \Rightarrow CH_4 + 2\,CO + 4\,H_2$ (12)	291.99	500	870	2.0	335.10
$C_3H_8 + 3\,H_2O \Rightarrow 3CO + 7H_2$ (13)	498.36	500	870	2.0	563.64
Shift reaction					
$CO + H_2O \Rightarrow CO_2 + H_2$ (14)	–41.16	220	250	2.0	–37.19
		350	420	2.0	–36.48
Oxidation reaction					
$C + O_2 \Rightarrow CO_2$ (15)	–196.89				
$2\,CO + O_2 \Rightarrow 2\,CO_2$ (16)	–566.37	800	1000	2.0	–542.58
$2\,H_2 + O_2 \Rightarrow 2\,H_2O$ (17)	–484.17	800	1000	2.0	–480.06
$CH_4 + 2\,O_2 \Rightarrow 2\,H_2O + CO_2$ (18)	–803.04	300	1400	0.1	–646.83
$2\,C_2H_6 + 7\,O_2 \Rightarrow 6\,H_2O + 4\,CO_2$ (19)	–2857.6	300	1400	0.1	–2329.2
$C_3H_8 + 5\,O_2 \Rightarrow 4\,H_2O + 3\,CO_2$ (20)	–2045.4	300	1400	0.1	–1673.9
Soot formation reactions					
$CH_4 \Rightarrow C + 2\,H_2$ (21)	74.9				
$2\,CO \Rightarrow C + CO_2$ (22)	–172.54				
Carbonyl formation					
$Ni + 4\,CO \Rightarrow Ni(CO)_4$ (23)	–159.50				
$Fe + 5\,CO \Rightarrow Fe(CO)_5$ (24)	–233.20				

The reactions of the hydrocarbon feedstock with substoichiometric amounts of pure oxygen (oxygen content > 95 %) follow Equation (1). Other reactions, which take place simultaneously or immediately following, are the water gas reaction (Eq. 4), the CO shift (Eq. 5) and the methane formation or decomposition reaction (Eq. 6).

The reactions take place primarily in the flame and in the empty brick-lined reactor. The residence time is usually <5 s, the temperature 1250 – 1500 °C. Pressure is often adjusted to accomodate subsequent processes, 3 – 12 MPa being the usual range.

The subsequent formation of hydrogen by the water – gas shift reaction (Eq. 5, see also Section Catalytic Gas Purification), is of equal importance for hydrogen production as is the immediate formation of hydrogen (Eq. 1). Carbon monoxide is thereby shifted to hydrogen to the highest degree possible. The shift reaction occurs simultaneously to the oxidation-gasification reactions at temperatures > 1200 °C without the presence of catalysts. Higher hydrogen yields, however, can be achieved by attaining equilibrium at 200 – 500 °C in the presence of steam and suitable catalysts.

Two process routes have been applied to perform the cooling down, gas purification and shift reaction:

1. The classical sequence of the three aforementioned process steps calls for a two-stage physical scrubbing process (e.g., the Rectisol wash, see Section Scrubbing Processes). The process gas from the partial oxidation reactor is cooled in a waste-heat boiler thus producing high pressure steam. Soot and ash particles, as well as water-soluble components, are removed with a water quench at temperatures between 300 °C and ambient. Hydrogen sulfide is removed (e.g., at – 50 °C) in the first stage of a Rectisol absorption unit. The synthesis gas is then heated again to 300 °C and steam is added. The shift reaction occurs in a series of adiabatic reactors with intermediate cooling and decreasing inlet temperatures at each stage. After subsequent carbon dioxide removal in the second stage of the absorber the carbon monoxide content can be as low as 0.5 vol %. The catalyst, especially the one used at the lowest temperature, is extermely susceptible to sulfur poisoning (see Section Adsorption Processes).
2. Since sulfur-tolerant shift catalysts have been developed (see Section Adsorption Processes), an alternative process has become possible. Here the high water content required in the gas for the CO-shift reaction is achieved by direct quenching of water into the raw gas. The shift reaction takes place in a series of adiabatic reactors without prior removal of sulfur. It has been proposed to replace them by one temperature-controlled, quasi-isothermic reactor [68]. The sulfur compounds hydrogen sulfide, carbonyl sulfide, and carbon dioxide are removed in a single stage Rectisol wash, which in its regeneration section can produce two separate fractions, namely a carbon dioxide and a hydrogen sulfide – carbon dioxide mixture. The CO content of the purified gas is <0.5 vol %.

4.1.2.3. *Examples of Industrial Scale Hydrogen Production (by Partial Oxidation of Hydrocarbons)*

Production of Hydrogen-Rich Ammonia Synthesis Gas. One example is the production of hydrogen-rich ammonia synthesis gas by partial oxidation of high-sulfur-content vacuum residue (⟶ Ammonia).

Starting from 44 t/h vacuum residue with a sulfur content of up to 4.5 wt %, 116 000 m^3 (STP)/h of 96 vol % pure hydrogen are generated. Hydrogen is further purified using a liquid nitrogen scrubber (⟶ Gas Production – Liquid Nitrogen Wash Process), whereby 148 900 m^3 (STP)/h of ammonia synthesis gas with a composition H_2 = 75 mol % and N_2 = 25 mol %, is obtained. 56.25 t/h of ammonia can be produced from this synthesis gas.

Hydrogen from Residues of Upgrading Operations in Refineries [69], [70]. After optimization of many technical, environmental and economic parameters the process sequence in Figure 18 consists of the following steps:

1) *Air Separation*	Molecular sieve unit (because of the high air pollution at the site)
	External compression of oxygen to 7.5 MPa
2) *Gasification*	Texaco combined reactor in two parallel trains (special arrangements to cope with high ash concentrations in the feedstock)
3) *Soot extraction*	100 % soot recycling, using the Texaco naphtha soot extraction system No special extraction oil required
	Vertical decanter
4) *Carbon monoxide converter*	3 beds with cobalt – molybdenum sulfide catalyst with a residual carbon monoxide concentration of 0.5 vol %. Waste-heat recovery for steam generation and other uses within the plant
5) *Sour gas absorber*	Single-stage selective Rectisol absorber, two parallel absorbing columns for converted and carbon monoxide-rich feed gas
	Purified streams of hydrogen, carbon monoxide, methanol synthesis gas and Claus fraction containing 70 % hydrogen sulfide
6) *Methanation*	Standard version with nickel catalyst and integrated zinc oxide bed.
7) *Carbon monoxide cold-box*	Single column apparatus with internal carbon monoxide recycle (cold-box = containment for low-temperature process)

Figure 18. Gasification of refinery residue for the production of hydrogen, methanol synthesis gas, and carbon monoxide (Linde, Federal Republic of Germany)

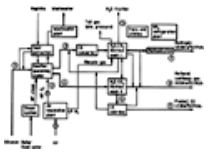

Material balance: The numbers denote the process streams in the above Figure

1 *Feedstock* : Bitumen, *m* = 33.3 t/h, *p* = 0.4 MPa, *t* = 150 °C, density (25 °C) = 1.056, viscosity (150 °C) = 8 000 Pa s*Composition:* 83.02 wt % carbon, 10.19 wt % hydrogen, 0.41 wt % nitrogen, 6.31 wt % sulfur, 45 ppm nickel (max 75), 150 ppm vanadium (max 250), 1000 ppm ash (max 1500)

Gas balance:

Mol, %	2	3	4	5	6	7	8	9
H_2		19.52	38.40	0.02	48.22	98.43	67.86	0.03
N_2	0.1	0.05	0.10	4.26	0.12	0.12	0.10	0.20
CO		21.61	41.61		51.23		26.91	98.74
Ar	0.4	0.06	0.11		0.14	0.14	0.11	0.30
CH_4		0.13	0.25		0.29	1.17	0.22	0.73
CO_2		2.12	3.98	21.12			4.80	
H_2S		0.64	1.25	73.30				
COS		0.03	0.06	1.3				
O_2	99.5							
H_2O		55.84	14.24			0.14		
total m^3(STP)/h	24 740	159 900	29 420	1 954	11 756	59 300	26 500	4 530
p, MPa	7.3	6.0	5.9	3.0	5.7	5.0	5.0	3.4
t, °C	120	240	170	40	–50	40	40	38

Figure 18 shows the overall process layout.

4.1.3. *Catalytic Reforming of Hydrocarbons*

A comprehensive treatment of catalytic reforming is given elsewhere (see ⟶ Gas Production – Steam Reforming of Natural Gas and Other Hydrocarbons). Here a brief outline of the most important industrial routes for hydrogen production is given.

4.1.3.1. *Feedstocks*

Hydrocarbons react with steam in an endothermic reaction to form carbon monoxide and hydrogen (Eq. 7, Table 14). The most important feedstock for the catalytic steam reforming process is natural gas. Other feedstocks are associated gas, propane, butane, liquefied petroleum gas and some naphtha fractions. The choice is usually made on the availability and the price of the raw material. Preference is given to lighter feedstock, if the price on a heating value basis is comparable. The reason for this is the better operability of plants using lighter feedstocks as well as the easier pretreatment operations, and thus the lower investment costs for comparable plants.

The chemical process is described by

$$C_nH_m + nH_2O \rightleftharpoons nCO + (n + m/2)H_2 \tag{25}$$

The shift reaction, where carbon monoxide, formed according to (25), reacts with excess water, occurs concurrently to this reaction. Additional hydrogen is formed by this exothermic process:

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{14}$$

Standard enthalpies of some important reactions occurring during steam reforming are given in Table 14. The heat of reaction of the overall endothermic process is provided by indirect heating. In *autothermal reactors* (e.g., secondary reformers) the oxidation reaction (see Table 14) supplies the necessary energy. This reaction takes place simultaneously or in advance of the reforming reaction.

4.1.3.2. *Tubular Reformer*

The equilibrium of the hydrocarbon steam reaction (7) and the water-gas shift reaction (14) dictates the conditions that must be met to achieve maximum hydrogen yields:

high temperature at the end of the reactor
high excess of steam
low pressure

These conditions can be realized best, if the endothermic reforming reaction occurs in a fired tubular reactor at high temperature (800 – 900 °C) and low pressure. The process pressure varies between 1.5 and 3.0 MPa. The exact choice of the process conditions depends on reaction kinetics, mechanical and economical terms.

The amount of steam added to the hydrocarbon feedstock is quantified by the molar steam-to-carbon ratio (S/C = ratio of S moles of H_2O to C moles of carbon contained in the hydrocarbon feedstock. If present, carbon monoxide and sometimes also carbon dioxide, is calculated with the C moles). The S/C ratio has values from 2.5 to 6, depending on the feedstock and subject to process optimization considerations. The principals of the optimization of a steam reformer plant for the production of hydrogen are discussed in more detail in Section Steam Reforming Plant.

The next step is the shift reaction at ca. 340 – 450 °C (high-temperature shift). In certain cases this is followed by a second shift reaction at 200 – 250 °C (low-temperature shift).

Catalyst. In commercial plants steam reforming and the concurrent shift reaction are catalyzed by supported nickel catalysts. The catalyst contains 15 – 25 wt % nickel oxide on a mineral carrier. Carrier materials are α-alumina, aluminosilicates, cement, and magnesia. Before start-up, nickel oxide must be reduced to metallic nickel. This is done preferably with hydrogen but also with natural gas or even with the feedgas itself.

With certain types of catalyst uranium oxide and chromium oxide are used as a promoter. This is reported to give a higher resistance to catalyst poisoning by sulfur components and a lower tendency to form carbon deposits (according to Eq. 21 in Table 14). For special applications such as low-temperature reforming, noble metal catalysts have been developed [71]. They have not been used commercially so far.

For the feedstocks *LPG* and *naphtha*, nickel catalysts with alkaline carriers [72], [73] or alkaline-free catalysts with magnesium oxide as additive [74] can be used. Both types of catalyst are less active than the conventional nickel catalyst. Therefore, a less rapid decomposition of the hydrocarbons is achieved. At the same time, the reaction of water with any carbon formed (Eq. 4) is catalyzed. Catalysts with alkaline carriers are mostly used for the upper part of the reactor tube only. The second part of the reactor is filled with more active nickel catalyst. Whereas the first catalyst gradually decomposes the higher hydrocarbons, the active catalyst is more suitable for the methane-rich fractions formed in the upper part. The aforementioned combination of catalyst can also be employed with natural gas feedstock, thereby offering the opportunity to switch to LPG or naphtha in case of natural gas shortages.

Required properties of the catalyst carriers are high specific area, low pressure drop and high mechanical resistance at temperatures up to 1000 °C. The catalysts are usually in the form of rings (e.g., outer diameter 16 mm, height 16 mm, inner diameter 8 mm). Other forms, such as saddles, stars, and spoked wheels are also commercially available.

The main catalyst poison in steam reforming plants is sulfur, which is present in most feedstocks. Concentrations as low as 0.1 ppm form a deactivating layer on the catalyst [74]. To some extent activity loss of a poisoned catalyst can be offset by raising the reaction temperature. This helps to reconvert the inactive nickel sulfide to active nickel sites.

Worse than small constant sulfur concentrations are sulfur peaks in the feedstock. They lead to permanent deactivation of the catalyst and to mechanical problems because of carbon deposits.

Other catalyst poisons are chlorine and other halogen compounds as well as heavy metals, especially lead, arsenic, and vanadium.

Nickel-free catalysts have been proposed for feedstocks heavier than naphtha [75], [76]. These catalysts consist mostly of strontium, aluminum and calcium oxides. They seem to be satisfactorily resistant to coking and work with high-sulfur feedstocks. However, with these catalysts even at reaction temperatures as high as 1000 °C the methane concentration in the exit gas is so high, that a secondary reformer has to be installed [77].

Equilibrium Conditions and Kinetics. The reactions in commercial reformers lead to a product composition which is usually very close to the equilibrium of the shift reaction (Eq. 14 in Table 14) and of the methane reforming reaction (Eq. 7) for design purposes. The resulting gas composition corresponds to an equilibrium temperature, which is usually only 5 – 20 K below the measured actual outlet temperature.

Higher hydrocarbons than methane are not usually found in the outlet products of a steam reformer. The equilibrium composition can be calculated from equilibrium data given in [73] or [78] as a function of the temperature, pressure and S/C ratio. For the dependence of the equilibrium composition on pressure and temperature at one (undisclosed) S/C ratio see ⟶ Gas Production.

A review of the many aspects of intrinsic kinetics of methane steam reforming is presented [79]. For most practical conditions, ranges of S/C and pressures, first-order kinetics can be employed for methane reforming. For low conversions of methane, which is not typical for industrial use, the concurrent carbon monoxide shift reaction deviates from equilibrium.

For the design of the reforming reactor not only the reaction kinetics within the reformer tube are important, but also the pattern of heat input required for the endothermic hydrocarbon decomposition reactions.

Heat distribution in conventional reformers is influenced (1) by the geometry of burner arrangements, (2) the pitch of the tubes, (3) the type and length of the flame, (4) the radiation of the flames, flue

gases and refractory walls, (5) size, shape and thickness of the tubes, (6) materials, and (7) whether the tube surfaces are machined. The calculation procedure for a standard type of hydrogen steam reformer is indicated in Figure 19.

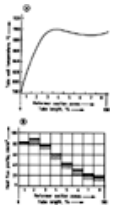

Figure 19. Tube wall temperature and heat flux in a reformer tube

A) Outer tube wall temperature vs. heat flux profile divided into, e.g., eight reformer sections

B) Hatched area: inclined to the right = heat requirements of the reaction; inclined to the left = heat provided by firing

Iteration for illustration purposes not completed

The radiation conditions in the fire box depend on the temperature of the flue gas, the reformer tubes and the refractory walls, as well as their emissivities. A zonal computer program calculates the interaction among these three factors for the desired heat duty at a given tube skin temperature and results in the firing requirements. The effects of combustion, air or fuel preheat are also considered in this calculation.

The given tube wall temperature is shown in Figure 19 A. The heat duty on the fire box side is shown for eight reformer section zones as one of the lines in Figure 19 B. The program which calculates the composition inside the reformer tube allows the heat requirements per length of tube to be calculated at the same time. This heat is supplied through the reformer tube wall and the given tube wall temperature determines the heat transferred to the inside of the tube. This is shown by the second line in the section zones of Figure 19 B. The aim is thus to find a tube skin temperature profile which will supply the heat to the process as required, or, in other words, to match the heat flux profiles. When this is accomplished, the temperature profile along the tube and the firing rate have been determined. The former is important for the mechanical design of the tube, the latter for the utility consumption.

Other factors, such as heat losses, flame lengths, staged burning, flue gas recirculation, position of the flue gas outlet duct, circumferential and other tube temperature asymmetries, also play a role. Therefore it is essential that results be corroborated by field data and the programs adjusted accordingly to reflect the influence of the above mentioned factors.

The behavior of the radiant section can thus be predicted by the results of these calculations with sufficient accuracy. Literature has been published in this field [80-82]. However, some problems in heat and mass transfer, influence of pressure and the decomposition of higher hydrocarbons are still not completely resolved.

The utilization of the flue gas leaving the reformer is also of prime importance. An exact prediction of the temperature of the flue gas leaving the radiant zone (i.e., bridgewall temperature) is essential because it affects the firing duty (i.e. the amount of heat introduced to the reformer by firing), and thus overall performance of the reformer.

Numerous variations in the order of the heat exachanger coils are possible. Normally the excess heat in the flue gas is utilized to produce export steam. Whatever arrangement chosen, the overall thermal efficiency of reformer and convection zone can normally be expected to be ca. 92 %.

Carbon Deposits. Formation of carbon deposits occurs predominantly because of the decomposition of methane according to Equation (21) in Table 14.

$$CH_4 \rightarrow C + 2\,H_2$$

Other reactions leading to soot formation or carbon-coke deposits are cracking of higher hydrocarbons, especially aromatic compounds, and of olefins.

Another reaction leading to carbon deposition, the Boudouard reaction (22) in Table 14

$$2\,CO \rightarrow CO_2 + C$$

is favored by low temperature and high pressure typically found at the inlet portion of radiant tubes in a steam reformer.

Theoretically at least, carbon laydown can be predicted from the equilibria of Equations (21) and (22) if other carbon monoxide forming reactions such as the steam reforming (4) and the shift reaction (14), are also taken into account (see Table 14).

Since the early investigations of DENT [83] much has been published on the kinetics and conditions of carbon laydown on commercial and experimental catalysts in plants and in the laboratory [84-86]. As a result, several factors influencing soot formation in commercial steam reforming processes have been identified:

Choice of Catalyst. Acid sites on the catalyst and its support, e.g., nickel on silica, catalyze carbon formation. Alkaline supports and a spinel structure of the catalyst permit operation close to the soot limits of equilibrium conditions.

Feedstock Selection and Pretreatment. Olefins and aromatics lead to carbon deposits and catalyst deactivation even in regions which are carbon-free according to equilibrium conditions. Therefore, if the

presence of these compounds cannot be avoided, a recycle of product hydrogen is often used to hydrogenate the unsaturated hydrocarbons. The hydrogen recycle ratio depends on the amount of unsaturated hydrocarbons expected and will often be ca. 3 – 5 mol % of the carbon content of the feedstock. Hydrogenation is usually performed simultaneously to the hydrogenation of sulfur containing compounds and the same type of catalyst (a sulfided cobalt or molybdenum catalyst) can be employed.

Kinetic Effects. High space velocities in critical regions also help to avoid soot formation.

Steam : Carbon (S/C) Ratio. Proper choice of the S/C ratio in steam reforming is imporant. Carbon formation is avoided if high S/C ratios are chosen because the equilibrium of the feed gas is shifted from the carbon regions. The S/C ratio also influences the hydrogen : carbon monoxide ratio of the product gas. This is important, if synthesis gas, not hydrogen, is to be produced. The shift reaction, which follows the reforming stage in hydrogen production plants and in many other applications, is also influenced by the S/C ratio. A high S/C ratio favors hydrogen production. At low S/C ratios, the danger of hydrocarbon formation exists with subsequent carbon deposits in the high-temperature shift reactor. In the worst case reduction of the iron oxide catalyst may even occur.

Finally, economic evaluations indicate that the optimum S/C ratio tends to be low. Figure 20 shows that the feedstock consumption is slightly less at low S/C ratios. Because investment cost also declines with smaller gas flows, lower values of S/C are thus favored, although the aforementioned restrictions must be taken into account.

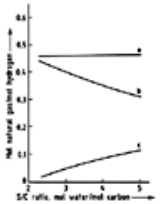

Figure 20. Consumption of natural gas in a modern hydrogen plant with PSA unit as a function of steam – carbon ratio

a) Total consumption of natural gas; b) Portion of natural gas used as process gas; c) Portion of natural gas used as fuel gas in addition to PSA off-gas [89]

Common values for natural gas operation are S : C 2.5 – 3, and for LPG and naphtha feedstocks S : C 3 – 5. If the desired product is not hydrogen, but carbon monoxide, the S/C ratio should be substantially lower.

Reformer Types. The aim of a reformer design for the production of hydrogen is to obtain a maximum temperature at the exit of the reformer tube and thus the maximum possible degree of reforming, using a minimum amount of fuel.

The fuel utilization is optimized, if the heat delivered to each section of the reformer tube corresponds to the heat requirements of the reaction in the tube. Various types of furnace and the geometry of the tubes and the burners in the furnace are shown in Figure 21 [87], [88].

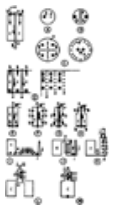

Figure 21. Types of steam reformer for hydrogen production; furnace and convection zone configurations

Cylindrical Furnaces: A) Small plants, pilot plants; B) Standard configuration; C) Tube/burner configuration for plants with higher capacity (up to 5 000 m^3 (STP)/h = 60 mol/s H_2); Process gas and flames (flue gas) usually cocurrent

Box Furnaces: D) Standard configuration; E) Bottom fired reformer; F) Side fired reformer; G) Terrace wall reformer; H) Combined configuration;

Convection Zone: I) Standard configuration, horizontal at grade; usually cocurrent flow of process gas and flames; J) Vertical configuration, downflow; K) Vertical configuration, upflow; L) Asymmetrical configuration; M) Central configuration;

a) Catalyst filled reformer tube; b) Burner; c) Radiation zone; d) Convection section; e) Fluegas duct; f) Inlet gas distribution system; g) Outlet gas tube (pig tail); h) Collector; i) Flue gas blower; j) Stack

For small reformer plants with a hydrogen capacity of up to 1000 m^3 (STP)/h (12.5 mol/s) cylindrical furnaces are often used. These furnaces require the least expenditure with regard to steel construction and insulation materials and can be prefabricated in the workshop. The plant size is then limited by the maximum transport dimensions.

Figure 21 C shows an arrangement, which accomodates a large number of tubes and allows a hydrogen production of up to 5000 m^3 (STP)/h (60 mol/s). Because of the varying distances of the tubes to wall and burner (see A and B) and because of the varying circumferential heat flux of each tube (C) an even heat flux distribution cannot be attained. This leads to designs with lower overall heat flux and shorter tube lengths for this type of cylindrical furnace.

For larger hydrogen plants box furnaces are well-proven. Often they can be prefabricated, so that a certain degree of preassembly is possible. The various types of box furnaces mainly differ in the arrangement of the burners. For hydrogen plants the usual configuration is the *top-fired* furnace (D), but *bottom-fired* (E), *sidewall-fired* (F) and *combined types* (G, H) are also in use. The box furnace allows a more even heat load distribution to each tube. The heat flux can therefore be increased.

Because the radiation is evenly distributed around the tube, the mechanical stress is minimized and larger tube diameters (up to 0.18 m outer diameter) can be used. The total number of tubes per

quantity of hydrogen produced can thus be decreased. This reduces furnace volume, piping and the number of burners.

In almost all designs the process gas flows through the tubes from the top to the bottom. The hot outlet gas is usually gathered in a manifold system, which may or may not be lined, and led to the process gas boiler. This equipment is best placed at ground level. Three different arrangements are possible for the combination of process gas – flue gas flow.

Most modern reformers are *top-fired reformers* with cocurrent flow. Figure 22 A shows the temperature profile of the outer tube wall and the gas temperature in the tube interior. The high temperature attained rapidly at the beginning of the tube allows a high reaction rate over a large portion of the tube.

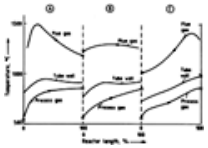

Figure 22. Temperature profiles in different steam reformer designs

A) Cocurrent flow; B) Sidewall-fired; C) Countercurrent flow

The upper curve shows the approximate flue gas temperature, the lower curve the temperature of the process gas, with an adjusted inlet temperature of 550 °C and exit temperature of 800 °C. The curve between shows the skin temperature of the tube wall, important for the mechanical design of the reformer tubes

Sidewall-fired furnaces exhibit a very even flue gas temperature profile and allow a certain amount of freedom to adjust the tube wall temperature (see Figure 22 B).

The flue gas exit temperature with a *countercurrent* system is particularly low (see Figure 22 C), which indicates that a large fraction of the heat content of the flue gas has been utilized. This type is used preferentially whenever a furnace with the convection zone at the top, using natural draft, and with a low degree of further utilization of the heat content of the flue gas is envisaged.

Further critical items of steam reformers are:

1. The support of the tubes to minimize the load on the distribution system and to avoid bowing of the tubes.
2. The preheated gas mixture (ca. 500 °C) enters the reformer via the inlet distribution system. The distribution system must be designed to take up the vertical expansion of the reformer tubes (for example, a 14 m tube may expand by as much as 0.25 m under operating conditions).
3. The horizontal thermal expansion of the outlet collector must be compensated for by the connecting pipework, the so-called pigtails.
4. Equal gas distribution over the tubes is attained by proper hydraulic design of the inlet and outlet collectors coupled with uniform loading of the catalyst in the reformer tubes.
5. The ratio of tubes : burners in large top-fired reformers is ca. 4. Smaller numbers of burners have the disadvantage of uneven heat distribution to the tubes but the advantage of lower investments because of the reduction in piping and ducting.
6. Design of burner, insulation, flue gas duct and reformer geometry. For details see [90].

Only a fraction of the heat supplied to the reformer furnace is used for reforming. A large amount of the heat in the flue gas is carried into the convection zone, where it is used to preheat the reformer feed, to produce and to superheat steam and, depending on circumstances at the site, maybe to preheat the combustion air.

The convection section can be installed as shown in Figure 21 (I – M). In most large hydrogen plants a flue gas fan must be installed, to induce the slight vacuum necessary for the draft burners and to compensate for the pressure drop over the heat exchanger coils in the convection section. For small plants, the convection section is sometimes installed on top of the reformer. This reduces the power requirements of the fan or allows design for natural draft.

The design of the convection section itself depends on the overall energy balance of the plant (see below) and the way the energy is to be utilized within the plant. This is influenced by the use of steam or electric drives or of gas turbines, etc. [91].

Energy Balance. The energy required for reforming in the furnace is (1) utilized as reforming energy and (2) in the convection section for preheating and steam production. In addition, energy losses such as insulation losses and the heat content of the flue gas lost through the stack must be considered. The efficiency factor for the steam reformer can be defined in several ways, for example:

Firebox Efficiency. The firebox efficiency is the ratio of energy necessary for reforming (Q_R) to the energy of fuel combustion (Q_{LHV}) (usually expressed in terms of the lower heating value of the fuel). If air or fuel are preheated it is necessary to consider the amount of energy brought in by preheated air Q_A, and/or preheated fuel Q_F to give:

$$\eta\,(\text{firebox}) = \frac{Q_R}{(Q_{LHV} + Q_A + Q_F)}$$

Firebox efficiencies lie in the range 0.45 – 0.55. This value only gives the relationship of the distribution of the energy between radiant and convection section, not, however, the degree of energy utilization.

Furnace Efficiency. The energy utilization in firebox and convection section is characterized by the overall furnace efficiency. This is the ratio of the sum of the utilized energy for reforming Q_R, for steam production Q_S and for any further energy utilization in the convection zone (e.g., feed preheating and superheating, export steam superheating, economizer Q_C) to the total energy supplied to the firebox

$Q_{LHV} + Q_A + Q_F$:

$$\eta(\text{furnace}) = \text{energy utilized}/\text{energy available}$$
$$= \frac{(Q_R + Q_S + Q_C)}{Q_{LHV} + Q_A + Q_F}$$
$$= (1 - \text{energy losses})/(\text{energy available})$$

The efficiency thus defined usually lies between 0.85 and 0.93 for hydrogen plants. The insulation losses are ca. 300 – 400 W/m^2 in the furnace section and ca. 600 – 700 W/m^2 in the convection zone which, depending on the size of the plant, amounts to ca. 1 – 5 % of the fired energy. The flue gas losses also vary with the size of the plant and the energy utilization and amount to 6 – 10 %.

Energy management plays a vital role in steam reformer design. The efficiency and the economy of the process are determined by the design variables, e.g., feedstock and fuel costs, bonus for export steam, use of electric or steam drives, S/C ratio, product gas pressure, etc. Suggestions for optimizing steam reformer plants can be found in the literature [92-94]. One typical example is discussed below.

4.1.3.3. *Steam Reforming Plant*

Figure 23 shows a simplified process flow diagram of a typical hydrogen plant. As in the majority of such plants, the feedstock is natural gas. Table 15 summarizes the process data. For the production of pure hydrogen, the following process stages are necessary:

Table 15. Mass balance of a typical steam reformer plant for the production of 5000 m^3 (STP)/h hydrogen. Feedstock: natural gas; byproduct: steam (3.0 MPa, 255 °C) 4700 kg/h (the numbers in Figure 23 denote the balance points)

	Feed, 1	Before desulfurization, 2	Reformer inlet, 3	Reformer outlet, 4	Pressure swing adsorption, 5	H_2 product, 6
H_2, vol %	0	4.8	1.1	51.9	74.6	99.99
N_2, vol %	1	1	0.3	0.2	0.25	0.01
CO, vol %	0	0	0	10.7	5.3	0
CO_2, vol %	0.5	0.5	0.1	5.1	14.8	0
CH_4, vol %	95	90.4	26.1	3.8	4.9	0
C_{2+}, vol %	3.5	3.3	1.0	0	0	0
H_2O, vol %	0	0	71.4	28.3	0.15	0
Quantity, m^3 (STP/h)	2160	2070	7160	10450	8180	5000.5
Temperature, °C	20	390	520	850	20	20
Pressure, MPa	2.0	1.9	1.9	1.7	1.6	1.5

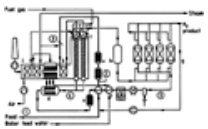

Figure 23. Hydrogen production by steam reforming, gas purification by pressure swing adsorption (Design: German Linde)

a) Desulfurization; b) Feed preheater/superheater; c) Reformer; d) Waste-heat boiler; e) CO shift reactor (HT shift); f) Cooling of raw gas; g) Pressure swing adsorption; h) Off-gas puffer for fuel; i) Convection zone with steam production, steam superheating, and air preheating

Feed Preheating and Desulfurization. Usually, natural gas contains only small amounts of sulfur compounds, predominantly in the form of hydrogen sulfide. Because of the high sulfur sensitivity of the catalyst in the reformer (and in the LT shift reactor, if installed) hydrogen sulfide and other sulfur compounds must be removed on a zinc oxide bed (see Section Catalytic Gas Purification). If higher organic sulfur compounds (mercaptans, thiophene) or carbonyl sulfide in concentrations in the ppm-range are present, the zinc oxide bed alone is not sufficient and the sulfur compounds are converted to hydrogen sulfide in a hydrogenation stage. The hydrogen necessary is taken from the product stream and amounts to ca. 5 % for natural gas (whereas for naphtha and LPG 10 % and 25 % H_2, respectively, is used). Hydrogen recycle and hydrogenation also lowers the amount of olefins in the feed. The excess hydrogen in the feed helps to keep the reforming catalyst at the top of tube in a reduced state.

For hydrogenation and desulfurization the inlet gas is preheated to 350 – 400 °C. The gas leaves the purification section with a sulfur content of <1 ppm at essentially the same temperature.

Steam Addition, Reaction and Cooling. Steam is added to the inlet feed to give a ratio of 2.5 moles of steam per mole of carbon. This mixture can be further preheated in the convection section to 520 °C without cracking occurring. Reforming takes place in the reformer tube, the outlet temperature being 850 °C. After the reaction the gas is cooled quickly, thereby producing steam in the process gas boiler.

Water Gas Shift Reaction and Gas Purification.
High-Temperature Shift Conversion. The gas cooled down to 350 °C in the process gas boiler still contains a fair amount of carbon monoxide which can be converted into hydrogen by the water-gas shift reaction (Eq. 14, Table 14). An iron-oxide based catalyst is used for this. The shift reaction is usually carried out in a one-stage adiabatic reactor. Since the reaction is exothermic, the process gas temperature rises to 400 – 450 °C so the conversion is incomplete.

Pressure-Swing Adsorption (PSA). After high-temperature shift conversion, the gas is cooled down to ambient temperature. Water is condensed and removed before the gas is fed to at least three (but up to twelve) parallel adsorbers, which run in a predetermined cyclic operation. The various purge and decompression gases of the adsorbers are led into a large buffer vessel to compensate for variations in pressure and gas concentrations and are subsequently fired in the reformer along with a small amount of fuel gas for trimming purposes.

Classical Process. In an alternative route after the HT shift, the so-called classical process, the gas is cooled to ca. 220 °C and fed into a low-temperature shift reactor; followed by a carbon dioxide scrubber (monoethanolamine or hot potash, see Section Scrubbing Processes). The remaining impurities in the hydrogen, traces of carbon monoxide and carbon dioxide, are removed by a catalytic methanation step. Using this process, a hydrogen purity of 97 – 99 % can be attained (not taking the inert gases into account). Both process routes have been compared in the literature [92], [95], [96]. For pure hydrogen plants the PSA version has become standard. Plants are built according to the classical route only in special cases in which (1) the feedstock (natural gas, LPG, naphtha) is expensive, (2) a cheap fuel is available to fire the reformer, (3) carbon dioxide is a desired product, or (4) a hydrogen purity of <99.9 % can be tolerated. For the coproduction of carbon dioxide combinations of both processes have also been built.

Heating, Steam Production, Heat Management. In Figure 23 the fuel gas is essentially the purge gas of the PSA plant. In addition, small amounts of the feed gas are fired. The combustion air is preheated in the convection section up to 500 °C. The process gas boiler and the flue gas boiler produce the required process steam. Depending on the design conditions of the plant, export steam is available for other users. From the point of view of the heat balance, a hydrogen plant can be optimized to be self-supporting, for maximum production of export steam, for minimum hydrocarbon consumption, etc.

4.1.3.4. *Special Designs*

The general design of tubular steam reformers has changed little in the past 20 years. In particular new materials for reformer tubes, heat exchangers and insulation as well as better and fewer burners and a better heat management have led to a highly sophisticated type of plant with an excellent efficiency and reliability.

Nevertheless, many new designs have been developed although their economic importance is so far very limited. The main areas for improvement of the reformer are:

improvement of the heat transfer for the endothermic reaction,
reduction of the plant size,
reduction of the maximum temperature (at the burner and in the flue gas).

Pressurized Fireboxes. The pressure in the fireboxes of conventional tubular reformers is atmospheric or slightly below. The reasons for this are:

1. Simple construction and safe operation of the reformer furnace,
2. Simple sealing of large numbers of reformer tubes,
3. Purge gas (e.g., from the PSA) can be fired at the low pressure,
4. Induced draft burners or burners using combustion air at low pressures can be used.

The use of higher pressure in the firebox leads to higher heat transfer coefficients from the flue gas to the reformer tube by convection. Figure 24 shows the construction of such a reformer and a similar heat exchanger concept [97], [98].

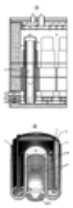

Figure 24. Pressurized and heat exchange type reformers

A) KTI (Kinetics Technology International BV), United Technology Corp. version [99]; B) Haldor Topsøe AS, Denmark [97], [98]

a) Gas inlet; b) Reaction zone (circular zone[s]); c) Reformed gas outlet; d) Heating gas inlet; e) Combustion air; f) Burner; g) Radiant section; h) Section with forced convection; i) Flue gas outlet; j) Insulation

The catalyst is contained in the concentric space between the tubes so that the thermal expansion caused by start-up and shut-down of the reformer tube does not lead to sealing problems. Pressure on

the flue gas side is 300 – 500 kPa [99]. One disadvantage is that both fuel gas and combustion air must be compressed. Plants with a hydrogen capacity of 120 – 5000 m^3 (STP)/h (1.5 – 60 mol/s) have been built.

Helium-Heated Reformers. To utilize the heat of the high-temperature nuclear reactor for the production of hydrogen from methane, a reformer with helium heating has been suggested. Since the helium temperature at the present stage of development is 900 °C, only a small temperature difference is available for the heat transfer. The reformer is similar to a conventional tube bundle reactor [100], however, the construction is very complex for safety and other considerations. Principally speaking, the reactor and heat exchanger could be used for other endothermic reactions. Up to now, only two prototypes with a capacity of 0.8 and 10 MW have been in operation. The helium is heated electrically in these plants.

Electrically Heated Reformers. For the production of small quantities of hydrogen electrically heated reformers have been suggested [101]. The hot product gas is used to heat the feed gas and produce part of the steam required. The energy requirement of ca. 300 kJ per mole of hydrogen is less than for a conventional steam reformer or for the electrolysis, whereby 80 kJ are brought in electrically and 220 kJ correspond to the lower heating value of the hydrocarbon used. However, it should be pointed out, that fossil and electrical energy can be compared on an equivalent basis only under special conditions.

Steam Reforming with Integrated Hydrogen Separation. High-temperature-resistant membranes have been integrated in a tubular reformer. Thus, hydrogen is withdrawn from the reaction mixture as it is formed. The chemical equilibrium is permanently shifted towards a higher feedstock conversion. For this reason either lower temperatures can be used or higher yields can be obtained. Palladium foil (100 µm) on a porous cylinder inserted into the reformer tube has been suggested as a suitable membrane. A small laboratory-scale plant has been built and successfully tested [102].

4.1.3.5. *Autothermal Reactors*

In autothermal reactors (secondary reformers) the energy for the production of carbon monoxide and hydrogen is produced by catalytic partial oxidation of part of the feedstock.

This type of reactor is employed:

1. to achieve high equilibrium temperatures (up to 1000 °C) thus minimizing methane concentration,
2. to enable use of elevated pressures (up to 5 MPa) together with high equilibrium temperatures (900 – 1000 °C) not feasible in fired steam reformer tubes,
3. to introduce nitrogen for the production of a stoichiometric syngas mixture for the ammonia synthesis.

In the first two cases the reactor is operated with pure oxygen (and steam). In the third case air or turbine off-gases are used. Figure 25 depicts three different types of autothermal reactor.

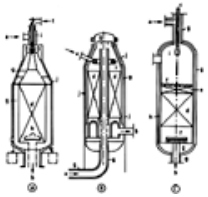

Figure 25. Designs of secondary (autothermal) reformers

A) Water-cooled jacket reactor with burner for methane-rich gases; B) Secondary reformer with toric oxygen or air supply device [90]; C) Secondary reformer with water-cooled tip without water jacket [103]

a) Raw gas inlet; b) Reformed gas outlet; c) Air, oxygen, and steam inlet; d) Catalyst; e) High-temperature catalyst; f) Inert material; g) Internal insulation; h) Multiple layers insulation; i) Burner; j) Water jacket

With the reactors shown in Figure 25 A and B no reaction occurs before the gas enters the catalyst bed. Reactor C features a burner, thus thermal oxidation takes place in the empty space above the catalyst bed. In all cases the highest temperatures occur at the top of the reactor because of the catalytic (A, B) or thermal oxidation (C) of hydrogen and carbon monoxide. The endothermic reforming of methane and other hydrocarbons according to the steam – methane equilibrium (Eq. 7 Table 14) is carried out in the lower part of the reactors. Nickel catalysts are used, often in the form of impregnated alumina lumps, which have a high resistance to fusion and thermal shock.

The waste-heat boiler is mounted immediately after the reactor, usually without a transfer line. Synthesis gas cools down rapidly in the boiler, thus avoiding recombination of carbon monoxide and hydrogen.

The direct application of the catalytic autothermal reactor without preceding decomposition of at least part of the hydrocarbons in a primary reformer has been studied in experimental plants [104]. Several layers of catalyst with increasing activities are used. Soot formation and high-peak temperatures are avoided by using a suitable catalyst at high space velocities, even though benzene feedstock, which is particularly liable to soot formation at a low S/C ratio is used. The experimental results are depicted in Figure 26. Because of the high reaction temperature the secondary reformers can stand a higher sulfur concentration in the feedstock. This fact has led to several tests with high-sulfur feedstock. However, commercial application has not yet been realized [76], [77].

Figure 26. Autothermal reforming of benzene on Ni-catalysts of increasing activity [104]

Pressure 101.3 kPa; inlet temperature 565 °C; molar air/carbon ratio 1.8; molar steam/carbon ratio 0.6

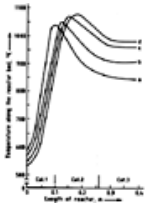

a) Space velocity = 5512 m^3 (STP) h^{-1} m^{-3}, carbon formation; b) Space velocity = 11 023 m^3 (STP) h^{-1} m^{-3}, carbon formation; c) Space velocity = 16 535 m^3 (STP) h^{-1} m^{-3}, carbon formation; d) Space velocity = 22 047 m^3 (STP) h^{-1} m^{-3}, no carbon formation

4.1.4. *Refinery Processes*

In refineries some processes require hydrogen, while others are a potential source of hydrogen. With the increasing quality demands for refinery products and with the increasing use of bottom-of-the-barrel products, the transfer of hydrogen within the refinery is increasing. Hydrogen is produced by steam reforming of hydrocarbons and by recovery from hydrogen-rich gases and refinery offgases, formerly used as heating gases, (Chap. Purification of Hydrogen).

Production and recovery of hydrogen from the byproducts of the following types of refinery processes are possible:

cokers and visbreakers
(thermal cracking processes)
catalytic crackers
catalytic reformers

For a detailed description of the processes see ⟶ Oil Refining. Here only facts relating to the hydrogen production are discussed.

Cokers and Visbreakers. In cokers and visbreakers heavy crude and residues are converted to petroleum coke, oil, light hydrocarbons (benzene, naphtha, LPG) and gas. Depending on the process, hydrogen is present in a wide range of concentrations. Since (like the coal coking process) petroleum coking processes need gas for heating purposes, adsorption processes (PSA) are best suited to recover the hydrogen because they feature a very clean hydrogen product and an off-gas suitable as fuel (see Section Coke Oven Gas).

If more hydrogen is required in the refinery, it can be derived from the fuel gas and from the LPG fraction by steam reforming. In times of oil shortage processes have been suggested to convert the excess coke into synthesis gas and hydrogen by coke gasification [105], [106].

Catalytic Crackers. Catalytic cracking is the most important process step for the production of light products from gas oil and increasingly from vacuum gas oil and residues.

In catalytic cracking the molecular mass of the main fraction of the feed is lowered, while another part is converted to coke. The coke is deposited on the hot catalyst. The catalyst is regenerated in one or two stages by burning the coke off with air, which also provides the energy for the endothermic cracking process. The cracking processes occurring can be categorized as follows [107]:

paraffins and naphthenes are cracked to olefins and to alkanes with shorter chain length,
monoaromatic compounds are dealkylated without ring cleavage,
diaromatics and polyaromatics are dealkylated and converted to coke.

Hydrogen is *formed* only in the last type of reaction, whereas the first two reactions produce light hydrocarbons and therefore *need* a certain amount of hydrogen. Figure 27 shows for various feedstocks, that an increase in coke production goes hand in hand with an increase in hydrogen production. Thus, a catalytic cracker can be operated in such a manner that enough hydrogen for subsequent processes is formed.

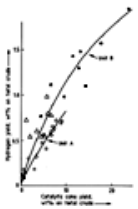

Figure 27. Hydrogen production vs. catalytic coke yield using amorphous catalyst for different feedstocks [108]; ○ West Texas Permian; △ Light Iranian; ▪ Kuwait; ● Residuum

Figure 28 shows how the hydrogen production from catalytic cracking can be linked to other hydrogen consuming processes. After desulfurization by absorption processes (see Section Scrubbing Processes) the off gas is usually purified and fractionated in a cryogenic unit (see Section Low-Temperature Processes).

Figure 28. Hydrogen production by catalytic cracking with subsequent use in hydro-desulfurization processes

——H_2-rich gases; —— H_2 purified

Catalytic Reformers. Naphtha fractions are reformed to improve the quality of gasoline. The most important reactions occurring during this process are the dehydrogenation of naphthenes to aromatics. For the dehydrogenation of cyclohexane, e.g., the hydrogen formation is as follows:

$$C_6H_{12} \rightarrow C_6H_6 + 3\,H_2 \quad \Delta H = +206.20\,\mathrm{kJ/mol}$$

This reaction is endothermic and is favored by low pressures. However, because hydrocracking occurs as a concurrent reaction the pressures are chosen between 1 and 3 MPa and the reaction temperature lies in the range of 300 – 450 °C. The reaction is performed on platinum catalysts, with other metals, e.g., rhenium, as promoters. The feedstock for the reforming process must be carefully purified from sulfur and nitrogen compounds (both <1 ppm). This uses up a great deal of the hydrogen produced. Table 16 summarizes the results of some reforming processes. The aromatic content of the feedstock is substantially increased and the hydrogen set free raises the content in the off-gas to 85 %.

Table 16. Hydrogen as a byproduct of catalytic reforming processes. Range of production and composition of the H_2-rich gas fraction

	Range	
Feedstock: naphtha		
Boiling range, °C	80 – 200	
Paraffins, vol %	75 – 50 – (25)	
Naphthenes, vol %	30 – 50	
Aromatics, vol %	5 – 20 – (40)	
Liquid products		
Octane number	(95) – 100 – 105	
Reformate, wt %	75 – 88	
Paraffins, vol %	30 – 20	
Naphthenes, vol %	0.5 – 2.5	
Aromatics, vol %	65 – 85	
Hydrocarbons C_2– C_4, wt %	10 – 15	
Hydrogen, wt %	2 – 3	
Composition of H_2-rich gas		typical
H_2,vol %	75 – 85	82
CH_4, vol %	5 – 9	7
C_2, vol %	4 – 7	5
C_3, vol %	2.5 – 5.5	4
C_4, vol %	0.5 – 2.5	2

4.1.5. *Petrochemical Processes*

Steam Cracking for Olefin Production. During the production of ethylene (⟶ Ethylene) and propene (⟶ Propene) large amounts of hydrogen are formed as byproduct. The composition of some crude gases from steam cracking of ethane, naphtha and hydrogenated vacuum gas oil is given in Table 17.

Table 17. Hydrogen as a byproduct in the production of ethylene in modern short residence time cracking furnaces. Cracking conditions standardized to an ethylene yield of 32 %

Raw gas	Ethane	Naphtha	Gas oil
After cracker and dryer			
H_2, mol %	33	16	10
CH_4, mol %	5	32	30
C_2H_4, mol %	32	32	32
C_2H_6, mol %	28	5	6
C_3H_6, mol %	1	11	16
C_3H_8, mol %		1	1
Other: C_{4+}, CO, H_2S, mol %	1	3	5
H_2-fraction of demethanizer			
H_2, mol %		80 – 95	
CO, mol %		1 – 0.5	
CH_4, mol %		19 – 4.5	
C_2, mol %		0.1	

Since propane, propene, ethane, and ethylene are fractionated at low temperature, hydrogen is usually separated from methane by condensation of the latter (see Section Low-Temperature Processes). The hydrogen recovered has a purity of 90 – 96 %.

Acetylene Production. Hydrogen is also produced as byproduct in the production of acetylene by thermal cracking (⟶ Acetylene, ⟶ Acetylene – Production from Liquid Hydrocarbons (Plasma Arc Process)). Because of the extreme cracking conditions the hydrogen content in the crack gas is particularly high. The production of 120 000 t/a acetylene and 50 000 t/a ethylene using the Hüls electric arc process yields 360×10^6 m^3 (STP)/a hydrogen as byproduct. In Table 18 the crude gases of several processes are compared. The acetylenes are recovered from the raw gas by absorption. Pure hydrogen is recovered from the remaining gas by cryogenic processes, sometimes in combination with adsorption processes.

Table 18. Hydrogen as a byproduct in acetylene production (raw gas after quench and condensation, dry basis)

	BASF Process based on			Hüls Plasma arc process[b]
	Methane [a]	LPG	Naphtha	Methane-rich gas

Gas composition				
H_2, vol %	56.5	46.4	42.7	57.6
C_2H_2, vol %	7.5	8.2	8.8	15.5
CH_4, vol %	5.2	5.0	4.9	13.8
Other hydrocarbons, vol %	0.8	1.0	1.2	2.5
CO, vol %	25.8	35.0	37.9	0.6
CO_2, O_2 inerts, vol %	4.2	4.4	4.5	10.0

[a] 1460 m^3 (STP) H_2 per tonne of feed.

[b] 3300 m^3 (STP) H_2 per ton of C_2H_2, 1830 m^3 (STP) H_2 per tonne of feed

In a further development of the electric arc process for the production of hydrogen, reduction gas, acetylene and ethylene, a plasma process has been suggested. A great number of feedstocks can be used ranging from natural gas to coal [109], [110]. The hydrocarbons are not cracked directly in the arc but a hydrogen plasma is used as energy carrier. In this way the acetylene yield can be increased while the hydrogen yield (and the amount of carbon deposits) decrease.

4.2. Electrolysis

Electrolysis has been used for approximately 100 years for hydrogen production. The first large installation was by Norsk Hydro in 1927 in Norway. Further plants were erected by Cominco in Trail, British Columbia, Canada in 1940 and, from 1945 onwards, some other plants with capacities up to 33 000 m^3 (STP)/h of hydrogen. As explained in Section Economic Aspects, the erection of large electrolysis plants depends on the availability of cheap electricity from hydropower stations. Plants for smaller amounts of hydrogen (50 – 500 m^3 (STP)/h) are often used by the industry because they are simple to operate. Moderate power costs are an additional incentive. A total of ca. 5 % of the world hydrogen production is by means of electrolysis.

4.2.1. *Principles*

For general principles see ⟶ Electrochemistry.

If a potential is applied to the electrodes of an electrolysis cell filled with a suitable electrolyte, the following reaction occurs at the electrodes:

1. Cathode

$$2\,H_2O(l) + 2\,e^- \rightarrow H_2(g) + 2\,OH^-(aq) \tag{26}$$

2. Anode

$$2\,OH^-(aq) \rightarrow 1/2\,O_2(g) + H_2O(l) + 2\,e^- \tag{27}$$

3. Cell Reaction

$$H_2O(l) \rightarrow H_2(g) + 1/2\,O_2(g) \tag{28}$$

where l is liquid, g is gaseous, and aq is dissolved in water.

Pure water is not suitable as an electrolyte because of its very low conductivity. Therefore, aqueous solutions of potassium or sodium hydroxide, sodium chloride, hydrochloric acid, etc. are used. Depending on the electrolyte and the material of the electrodes the reaction at the anode may lead to other products, in particular to sodium hydroxide and chlorine (see Section Hydrogen as Byproduct from other Electrochemical Processes).

The change in enthalpy for the electrochemical decomposition of water is

$$\Delta H = \Delta G + T\Delta S = -nFE + nFT\cdot\left(\frac{\partial E}{\partial T}\right)_{\mathrm{p}} \tag{29}$$

where *H* is the enthalpy, *G* is Gibbs free energy, *S* entropy, *T* temperature, *n* number of electrons transferred per formula conversion, *F* the Faraday constant, and *E* the electrochemical potential.

The thermodynamic standard data at *p* = 101.3 kPa and *T* = 298.15 K are:

Reaction enthalpy = $\Delta H^0_{298\,\mathrm{K}} = 286\,\mathrm{kJ/mol}$

Reaction entropy = $\Delta S^0_{298\,\mathrm{K}} = 163\,\mathrm{Jmol^{-1}K^{-1}}$

Gibbs free energy of reaction = $\Delta G^0_{298\,\mathrm{K}} = 237\,\mathrm{kJ/mol}$

The value of the Faraday constant is

$$F = 96495\,\mathrm{As/mol}$$

Using these data the required electrical energy becomes:

$$W = \Delta G = F \times n \times E_{\mathrm{rev}} \tag{30}$$

where E_{rev} is the ideal (reversible) decomposition potential. Under standard conditions it amounts to:

$$E_{\mathrm{rev}} = \frac{\Delta G}{n\cdot F} = 1.228\,\mathrm{V} \tag{31}$$

Because of irreversible processes on the electrodes and because of the resistance of the electrolyte the actual decomposition potential is always higher than the ideal value.

According to Equation (29) the free energy, ΔG, is smaller than the reaction enthalpy, ΔH, by the amount $T\,\Delta S$. In the ideal case operation of the cell requires an amount of electrical energy, $W = \Delta G$, and the addition of heat, $Q = T\,\Delta S$.

If both amounts of energy are to be supplied in the electrical form, the potential under standard conditions is increased by the thermal potential ΔE_Q.

$$\Delta E_{\mathrm{Q}} = \frac{T\Delta S}{n\cdot F} = 0.252\,\mathrm{V} \tag{31a}$$

Thus, the theoretical minimum decomposition potential at standard conditions is:

$$E_{\mathrm{th}} = 1.480\,\mathrm{V}$$

Figure 29 shows the temperature dependence of the energy or the decomposition potential, respectively at a constant pressure of 100 kPa. An increase in temperature lowers the reversible decomposition potential, while at the same time the thermal fraction of the total energy required increases. This means, that at higher temperatures heat energy instead of electrical energy can be partially used. Technical aspects of this behavior in regions above the boiling point of water (373 K) will also be considered in more detail in Section New Developments in Electrolytic Processes .

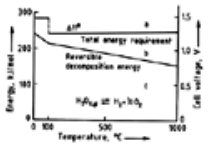

Figure 29. Temperature dependence of hydrogen formation by electrolysis at 101.3 kPa. The energy scale (kJ/mol) shows the standard free energy ΔG^0 and the enthalpy ΔH^0 of water cleavage. The voltage scale (V) allows determination of the corresponding theoretical cell voltage

a) Formation with H_2 heat output; b) H_2 formation with heat input necessary; c) No H_2 formation by electrolysis possible

The decomposition potential also depends on pressure. From the pressure dependence of the free energy

$$\left(\frac{\partial \Delta G}{\partial p}\right)_T = \Delta \mathrm{v}$$

the following expression can be derived for the pressure dependence of the decomposition potential of an ideal gas, i.e., neglecting the activity coefficients:

$$\Delta E_p = \Delta E^0 + \frac{RT}{2F} \ln \left\{ \frac{p_{(H_2)}}{p^0} \cdot \left[\frac{p_{(O_2)}}{p^0} \right]^{1/2} \right\}$$

For water at 298 K, taking the activity coefficients into account this equation becomes

$$\Delta E_p = \Delta E_{(p=p^0)} + 0.0189 \ \ln \frac{p}{p^0}$$

so that, on increasing the pressure from 100 to 1000 kPa, an increase of 0.0435 (= 0.0189 · ln 10) V would be expected. However, the fact that the gas produced at the electrodes has a smaller effective volume leads to a reduced overpotential at the electrodes, so that for electrolysis under increased pressure no increase in the total potential is in fact found.

The potential actually required is the sum of the following contributions:

$$U = E_1 + E_2 + E_3 + E_4$$

where U is the cell voltage (V), E_1 the theoretical cell voltage (E_{rev}) (V), E_2, E_3 the anodic and cathodic overvoltage (V) at the phase boundary electrolyte-electrode, and E_4 is the voltage drop (V), due to electrical resistance of the electrolyte system.

The overvoltages E_2 and E_3 can be influenced by suitable choice of the electrode materials and their surface conditions. The voltage drop in the electrolyte, E_4, depends (1) on the conductivity of the electrolyte, (2) on the permeability of cell diaphragms, (3) on the distance of the electrodes from one another, and (4) on the current density. Figure 30 shows the influence of the current density on the cell voltage during water electrolysis using potassium hydroxide electrolyte.

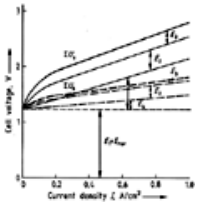

Figure 30. Relationships between current density and cell voltage for alkaline water electrolysis

—— conventional electrolysis, 30 wt % KOH, 90 °C; – – – prognosis for advanced conventional electrolysis according to [111], [112]

E_1 theoretical reversible cell voltage; E_2 anodic overvoltage, difficult to minimize; E_3 cathodic overvoltage, can be reduced by suitable catalysts; E_4 resistance of the electrolysis system, can be reduced by the use of thin-walled diaphragms and zero-gap constructions

The efficiency of the electrolysis is related to the minimum voltage according to Equation (29):

$$\eta = \frac{E + \frac{T \Delta S}{n F}}{U}$$

Under standard conditions:

$$\eta = \frac{1.48}{U} \frac{\mathrm{V}}{\mathrm{V}}$$

Principle Construction of a Water-Electrolysis Cell. Figure 31 shows the principal construction of a cell. Anode and cathode regions are separated by a diaphragm which allows the current to flow, but is otherwise gas-tight. The cell is filled with electrolyte. Oxygen is formed during the electrolysis at the anode, hydrogen at the cathode. Depending on the arrangement of the electrodes and the diaphragm, the cells may be designated as unipolar or bipolar cells (see Fig. 32).

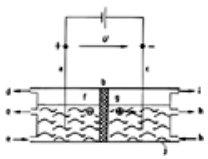

Figure 31. Schematic of a water electrolysis cell

a) Anode; b) Diaphragm; c) Cathode; d) Oxygen outlet; e) Anodic electrolyte cycle; f) Anion; g) Cation; h) Cathodic electrolyte cycle; i) Hydrogen outlet; j) Cell wall

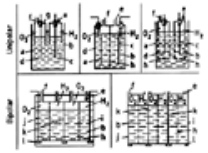

Figure 32. Different modes of water electrolysis cell according to [117]

a) Cell walls; b) Electrolyte; c) Cathode; d) Anode; e) Hydrogen outlet; f) Oxygen outlet; g) Gas collector; h) Diaphragm; i) Outer cathode; j) Outer anode; k) Bipolar electrode; l) Insulation

The *unipolar cells* are further differentiated in bell-type (without diaphragms), the usual diaphragm cells, and a bell-diaphragm type. A characteristic of the unipolar cell is that cathode and anode each have their own separate cell region.

Bipolar cells are characterized by the fact that the metallic separator between two cells connected in series serves as cathode in one of the cells and as anode in the next cell. The arrangement of the bipolar electrodes is similar to the construction of a filter press, so that they are often named accordingly.

Electrolyte. To minimize the resistance losses of the electrolytes usually 25 – 36 wt % potassium hydroxide solution is used. Other electrolytes are aqueous sodium hydroxide or, for the chlor-alkali electrolysis, solutions of hydrochloric acid, sodium chloride, etc. Electrolytes which are immobilized in polymers can also be used.

Since the conductivity of conventional electrolytes increases with temperature, electrolysis units usually operate at 70 – 90 °C, just below the boiling point of the aqueous solution. The conductivity of sodium hydroxide and potassium hydroxide as a function of temperature and electrolyte concentration is shown in Figure 33.

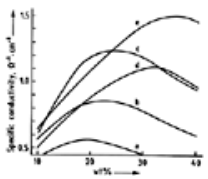

Figure 33. Conductivity of electrolytes used in conventional water electrolysis

a) NaOH, 40 °C; b) NaOH, 60 °C; c) NaOH, 80 °C; d) KOH, 60 °C; e) KOH, 80 °C

Materials. Electrochemical processes place particularly high requirements on corrosion protection. The conventional water electrolysis with potassium hydroxide as electrolyte can be carried out using carbon steel as a material of construction. Areas particularly subject to attack are coated with plastics or ceramics or are nickel-plated.

The cathode is usually made of steel. To reduce the hydrogen overvoltage E_3 the surface texture may be activated and coated with various catalysts. The anode and the electrodes of bipolar cells are usually made of nickel or nickel-coated steel. Next to platinum, which cannot be used for economic reasons, this material has the lowest overvoltage.

The diaphragm in the classical electrolysis units was initially made of asbestos. The asbestos mats, which separate the electrodes from one another, were reinforced with nickel nets. Because of the health hazards involved in the use of asbestos, substitute materials are being looked for. Ceramics and hydrophiliated polymers may be the solution to this problem.

4.2.2. *Conventional Water Electrolysis*

A conventional electrolysis system uses an alkali solution, usually potassium hydroxide, as electrolyte. The electrodes are in unipolar or bipolar arrangements. Hydrogen production takes place either at ambient or at increased pressure (Commercial plants up to 3 MPa).

Normal-Pressure Electrolysis Units. The advantages of a process at almost ambient pressure are:

1. Low mechanical strains on piping fittings, seals, and instruments,
2. Simple operation, because feed water is not required under pressure, quick start-up, simple maintenance,
3. Safer operation because of avoidance of high-pressure hydrogen, small pressure differences between the evolved gases, and simple blanketing of the electrolysis units and the piping network with nitrogen under low pressure

Unipolar Stuart cells are offered by a single producer (Electrolyser Inc., Toronto). The particular advantage of this type of cell is its robust nature. Since the cells are arranged in parallel they can be easily maintained or renewed, if necessary, without interruption of the process.

All other producers use *bipolar cells* which are mounted in series (filter press). This allows the use of higher voltages (up to 600 V) per unit. A schematic of bipolar electrolysis is shown in Figure 34. Pre-electrodes, which are placed close to the diaphragm to give the smallest possible resistance between the electrodes, can be clearly seen. The electrodes themselves are perforated or made from nets, so

that the gas bubbles evolve behind the electrodes and do not increase the resistance between the electrodes.

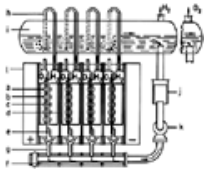

Figure 34. Cross section of a bipolar, atmospheric pressure electrolysis unit with alkaline electrolyte (Brown Boveri)

a) First anode with pre-electrode; b) Diaphragm; c) Pre-electrode of the cathode; d) Cathode for the first cell compartment and anode for the second cell compartment-bipolar middle electrode; e) Cell frame with insulation for the middle electrode; f) Distribution pipe for electrolyte; g) Inlet pipe for electrolyte; h) Take-off system for electrolyte and dissolved gases; i) Overflow pipe and collecting vessel for electrolyte and H_2 (an analogous system is used for O_2); j) Electrolyte filter; k) Circulation pump; l) End plate

The operating conditions for electrolysis units of various producers are given in Table 19 (see also ⟶ Electrochemistry). Typical operating conditions are electrolyte temperatures of 70 – 90 °C, a cell voltage of 1.85 – 2.05 V and a current density of 2 – 3 kA/m^2. The specific power consumption is between 4 and 5 kWh/m^3 (STP).

Table 19. Producer of medium and large sized commercial electrolysis units in alphabetical order. Information from producers and Achema publications in 1988

Producer, country	Type of electrolysis	Electrolyte conc., wt %	Pressure, MPa Temperature, °C	Current density, A/m^2 Cell Voltage, V Power consumption, kWh/m^3 (STP) H_2	Commercial status 1988 size of basic units, m^3 (STP) H_2/h [a]
BBC AG	bipolar	KOH	ambient	2000	5 – 300; 4×300
Switzerland	filterpress	25	80	2.05/4.9	
Davy-Bamag.	bipolar	KOH	ambient		3 – 330
Fed. Rep. of Germany	filterpress			4.2 – 4.5	
Electrolyser	unipolar	KOH	ambient	1340	0.5 – 100
Canada	tank	28	70	1.8/4.3 – 5.0 [c]	
				2500	larger plants
				1.9/4.4	
Krebs-Kosmo	bipolar	KOH	ambient	1000 – 9000	20 – 200
Fed. Rep. of Germany	filterpress	28	75	1.6 – 2/3.9 – 4.8	
Lurgi (Zdanski-Lonza)	bipolar	KOH	30	2000	110 – 750
Fed. Rep. of Germany	filterpress	25	90	1.86/4.3 – 4.6	
Norsk Hydro [b]	bipolar	KOH	ambient	1750	
Norway	filterpress	25	80	1.75/4.1	
Oronzio de Nora	bipolar	KOH	ambient	1500	5 – 1000
Italy	filterpress		80	1.85 – 1.95/4.6	
Teledyne	bipolar	KOH	0.7	3000	1 – 42
United States	filterpress	25	80	2.1 – 2.2/5.5 – 6.1	

[a] Much larger plants have been built by combining the appropriate number of units.
[b] No up-to-date information available.
[c] Higher value: standard electrode; lower value: activated.

It must be taken into account whether the energy consumption is only for the electrolysis itself, or whether further installed units, such as electrolyte pumps, water treatment plant, gas purification, are included.

The gas purity after compression and water removal is in general > 99.8 % for hydrogen and > 99.6 % for oxygen. Higher purities are possible by means of catalytic conversion and/or adsorptive drying

units.

Pressure Electrolysis. Pressure electrolysis units in the range of 0.6 – 20 MPa are available. Since hydrogen is usually required at high pressure, and because the power consumption for the pressure electrolysis does not increase significantly, these units save both equipment and energy for hydrogen compression. The construction and the sealing of the electrolysis cells, however, is complicated (see Fig. 35). The operating data reported by producers of pressure electrolysis are given in Tables 19 and 20. To compare the data with those of low pressure electrolysis the energy consumption for hydrogen compression to the end pressure has to be considered.

Table 20. Examples for small electrolysis units, projects, novelties, small series and advanced technologies

Producer, country	Type of electroysis	Electrolyte (membrane)	Pressure, MPa Temperature, °C	Current density, A/m^2 Cell Voltage, V Power consumption, kWh/m^3 (STP) H_2	Development status
Advanced Electrolysis [113],	bipolar	KOH,	0.6	10 000	project
Federal Republic of Germany	filterpress	conventional	140	1.8/4.3	
Asea Brown Boveri,	bipolar	water	2.0	10 000	small plants
Switzerland		(polysulfone)	80	1.74/4.2	1 – 20 m^3 (STP)/h
High Temp. Electrolysis [127],	bipolar	water	0.3	6000	1 ^{3} (STP)/h
Federal Republic of Germany		Y_2O_3+ ZrO_2	1000	1.32/3.2	pilot plant projected
Hydrogen Appliances,	bipolar	water	2.5	4.2 – 5	small plants
Belgium		(polysulfone)	50		0.04 – 1 m^3 (STP)/h
					0.1 – 5 m^3 (STP)/h
Sema Metra Conseil, France	with additional compressing fluid	conventional	20.0	5.5 – 6	20 – 100 m^3 (STP)/h
Sunshine-Project,	bipolar	KOH (30 wt %)	2.0	4000	20 m^3 (STP)/h
Japan [120], [126]	filterpress	conventional	120	1.65/4.0	pilot plant projected
Vandenborre,	bipolar	KOH (30 wt %)	1.0		small plants
Belgium			100	/3.9	

Figure 35. Cross section and exploded view of cells of a pressure electrolysis unit (Lurgi, Zdansky-Lonza pressure electrolysis)

a) Bipolar electrodes, dimple plate cell partititon; b) Pre-electrodes in the form of nets on both sides of the asbestos diaphragm (c); d) Cell frame; e) Sealing ring, sealing the electrolyte against the surroundings and insulating the bipolar sections against each other; f) Electrolyte duct; g) Bore in the electrolyte duct for supplying the electrolyte solution to both sides of the electrolyte; h) Parallel H_2 and O_2 ducts; i) Bore in the gas duct; j) Electrolyte compartment

Advanced Conventional Electrolysis. The conventional potassium hydroxide electrolysis according to [113], [114], can be improved considerably. Some of the suggestions made have been incorporated in single commercial electrolysis units. Other components necessary are still in the development stage. Areas of development are:

Zero-Gap Cell Geometry. By placing the electrodes directly on the surface of the diaphragm the space between the electrodes can be reduced and the voltage drop in the electrolyte, and thus the heat loss caused by the current, can be minimized. Figure 36 shows perforated foil electrodes in a zero-gap construction.

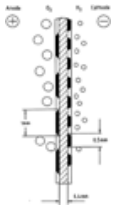

Figure 36. Zero gap cell geometry. Perforated cathodes and anodes, similar to the nets in Figure 35, in direct contact with the diaphragm and the membrane, respectively. The conical apertures with different diameters for O_2 and H_2, respectively, facilitate the detatchment of the developing gas bubbles [113]

In particular the construction with pre-electrodes has already been used in commercial plants. The disadvantage of this type is that in these electrolysis units, with high current density and high operating temperatures, the corrosion tendency is greatly increased. Products of corrosion, which may be formed at the anode, can grow through the diaphragm and lead to dangerous short-circuits.

New Materials for Diaphragms. A replacement for asbestos diaphragms are polysulfones, with hydrophobic character decreased by means of additives (e.g., polyantimonic acid) [115].

Another line of development is the use of oxide-ceramic diaphragms based on barium or calcium titanate or nickel oxide [116]. On a metallic net a hydrophilic gas-tight oxide layer is deposited. The thickness of the layer is <500 µm, which leads to a low resistance of ca. $2\times10^{-5}\ \Omega/m^2$. The stability against electrolytes at high temperatures is also favorable.

Electro-Catalysts. The kinetic inhibition of the electrode processes can be reduced by use of catalysts. This is already partly done in conventional plants. To assist the oxygen formation at the anode, lanthanum-containing perowskites, nickel-cobalt oxides and Raney nickel has been suggested.

The overvoltage at the cathode during the evolution of hydrogen can be reduced by the use of noble metals (platinum black). Commercial catalysts are, however, free of noble metals. Sintered, foamed, and alkali-leached nickel compounds, which are stabilized by the addition of TiO_2, ZrO_2 and MoO_3, are used. The long-term stability and the corrosion resistance against hot alkali solutions expecially in the depolarized state are the main problems encountered in the use of electro catalysts.

The increase in capacity of the new types of electrolysis cells makes further improvements necessary. The increased corrosion caused by the higher operating temperatures and greater current densities requires special materials of construction. In large-capacity plants nickel-plated steel, in low-capacity plants (≤ 20 m^3 [STP]) polysulfone, is employed. In particular, the gas collection system for the fine hydrogen bubbles necessitates additional equipment.

If all the theoretical possibilities to improve the electrolysis were employed, the water electrolysis could be greatly improved. Region b in Figure 37 shows the boundaries of the improvement possible in alkaline water electrolysis [113], [114].

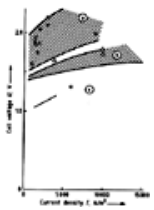

Figure 37. Illustration of the cell voltage and current density of commercially available or projected electrolysis units

The cell voltage U (V) can be converted by the equation $E = 2.392 \cdot U$ to the specific energy in kWh/m^3 (STP) H_2.

a) Range of commercially available electrolysis units [113], [116]; b) Range of advanced and membrane electrolysis [111], [114]; c) Steam electrolysis

○ suppliers data (see Table 19); △ present results; □ pilot plant project

Electrolysis Plants (see Fig. 38). Apart from the electrolysis unit an electrolysis plant consists of a number of different installations, such as electricity supply, water treatment, electrolyte circulation, the knock-out system (phase separator) to separate electrolyte from the product gases, gas purification, and the instrument and control system.

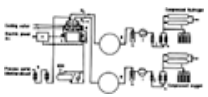

Figure 38. Process diagram of an electrolysis unit with the most important components

a) Rectifier unit; b) Process water demineralizer ion exchange unit; c) Electrolytic cell's water electrolyzer; d) Gas separator and cooler; e) Gas scrubber; f) Electrolyte tank and transfer pump; g) Gas holder; h) Filter; i) Compressor; j) Gas purifier; k) Drying; l) High-pressure storage tank and cylinder filling station

The water consumed during electrolysis has to be continuously replaced; the consumption is 0.805 L per cubic meter of hydrogen plus evaporation losses. The water make-up must meet stringent purity

requirements to avoid a build-up of substances in the electrolyte solution which poison the electrodes or promote corrosion. The required water purity (conductivity 1 µS/cm) is obtained with the aid of ion exchangers.

The power, which in large plants is supplied in the form of three-phase current at a high voltage, must be transformed and rectified down to the necessary d.c. voltage. In addition, power for the various auxiliary equipment, such as the electrolyte circulation pump and the instrument and control system is necessary. Considerable amounts of potassium hydroxide are only necessary for the first filling. Since the energy losses in the electrolyte are converted to heat, a large amount of cooling water is necessary to operate the plant. This is also true for the gas purification unit.

The *gas purification* depends on the purity requirements placed on the gas and may be a simple condensation to remove moisture present. For higher purity requirements the catalytic conversion of the oxygen present in the hydrogen (see Section Removal of Oxygen) and, after that, an adsorptive drying stage (Section Thermally Regenerated Adsorbers) is used.

4.2.3. *New Developments in Electrolytic Processes [117], [118]*

Electrolysis with Proton-Conducting Ion-Exchange Membranes. In this type of electrolysis, better known as solid polymer electrolysis (SPE, General Electric), the diaphragm is replaced by a proton-conducting ion-exchange membrane (⟶ Electrochemistry, ⟶ Fuel Cells – Proton Exchange Membrane Fuel Cells (PEMFC)). The water to be electrolyzed is introduced only to the anode side, and no additives to alter the conductivity are necessary.

The solid electrolyte consists of a Nafion membrane with the approximate structure:

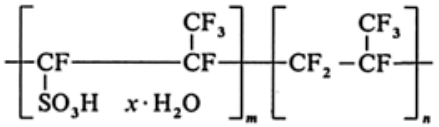

A cell consists of a ca. 0.25 mm thick film of this polymer. The following substances have been reported as electrode and catalyst material: platinum metal (cathode), and ruthenium oxide hydrate (anode).

The electrode layers have a high specific surface. The membrane becomes saturated with the water added at the anode side, which is electrolyzed and serves as a cooling medium at the same time. The electrical charge is transported by means of the hydrated H^+ ions of the sulfonic acid groups. Current densities of up to 20 000 A/m^2 can be reached. The principal construction of the electrolysis cell is shown in Figure 39 using the Membrel electrolysis (Asea Brown Boveri) as an example [119].

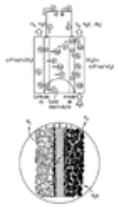

Figure 39. Principles and construction of an SPE electrolysis unit, Type: Membrel, Asea Brown Boveri [119].

At too high voltages ozone is produced at the anode

a) Porous cathodic current collector; b) Membrane, solid state electrolyte; c) Porous anodic current collector; d) Catalyst

The electrodes are in contact over the whole surface with a porous conductor; this is graphite on the cathode side and sintered nickel or titanium on the anodic side. The electrodes are in a bipolar arrangement.

Electrolysis units with proton-conducting membranes are offered in a capacity up to about 5 m^3/h of hydrogen from various producers. The characteristic operating data of such a module and of a conventional electrolysis unit are compared in Table 21. The advantages of these SPE units are not so much in the savings in power, but especially in the reduced volume of the unit, due to the much higher current density and the size reduction attainable because of the absence of the potassium hydroxide electrolyte system.

Table 21. Comparison of conventional potassium hydroxide electrolysis with electrolysis using polysulphone electrolyte and membranes

	BBC KOH electrolysis	ABB Membrel electrolysis	Hydrogen appliances SPE electrolysis
Electrolyte	25 wt % KOH	polysulfone Nafion	polysulfone Nafion
Temperature, °C	80	80	50
System pressure, MPa	0.1	≤2.0	2.5
Current density, A/m^2	2000	10 000	

Cell voltage, V	2.05	1.74	
Efficiency (η = 1.48/*U*), %	72	85	
Energy consumption (3.54/ η), kWh/m^3 (STP)	4.9	4.2	4.2 (at 20 % efficiency) 5 kWh/m^3 (STP) (at 100 % efficiency)
Block volume	30		

Further developments are proceeding in this direction. For example, in test plant modules used in the Japanese Sunshine Project [120], membranes and catalysts have been operated at current densities up to 20 000 A/m^2 and temperatures up to 150 °C.

Solid polymer electrolytic cells can also be used advantageously, for the coproduction of ozone (O_3) if the appropriate overvoltage is applied (see also Fig. 39).

Water Vapor Electrolysis. According to Equation (29) the total energy requirement for electrolysis is only slightly temperature-dependent. However, the reversible part of the energy i.e., ΔG, which must be supplied in the form of electrical energy, decreases with increasing temperature (see. Eq. 31 a and Fig. 29). This means that an increasing amount of the total energy can be supplied by thermal energy. Overvoltages, which result from kinetic inhibition, also decrease with increasing temperature.

In the conventional alkali electrolysis the reaction temperature is limited because of the vapor pressure of the electrolyte and corrosion problems. Thus, for the asbestos-diaphragm cell the upper temperature limit of operation is 120 °C. For the polysulfone membrane the upper temperature limit is ca. 150 °C.

In high-temperature electrolysis doped zirconium dioxide and some other rare earth oxides are used as electrolyte. At temperatures >800 °C ZrO_2 and some other rare earth oxides are oxygen ion conductors (lattice defect conductors). In Figure 40 the specific resistance of some mixed oxides is given. These mixed oxides are sintered into ring-shaped membranes, which are kept as thin as possible (0.5 mm, goal is 20 – 30 µm [121]) to reduce the electrical resistance of the cell. The rings are built up into tubes and the tubes to bundles of cells (Fig. 41).

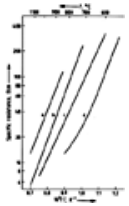

Figure 40. Specific resistance *r* (Ω cm) of various solid state electrolytes with oxygen-ion conductivity as a function of temperature [122]

a) $(ThO_2)_{0.93}(Y_2O_3)_{0.07}$; b) $(ZrO_2)_{0.88}(CaO)_{0.12}$; c) $(ZrO_2)_{0.90}(Y_2O_3)_{0.10}$; d) $(ZrO_2)_{0.92}(Yb_2O_3)_{0.08}$

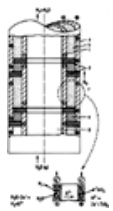

Figure 41. Principles and construction of a bipolar high-temperature electrolysis cell with an oxygen-ion conductive ZrO_2 electrolyte

a) Anode; b) Electrolyte (ZrO_2); c) Cathode; d) Insulation; e) Conductive cell sections; f) O_2 space; g) H_2, H_2O duct

Porous electrodes are positioned on the surface of the solid electrolyte. Mixed oxides with perovskite structure are used ($LaNiO_3$, $LaMnO_3$) for the anode. The cathode is a ZrO_2/Ni cermet, which prevents the active nickel metal of the cathode from agglomerating [123]. Water vapor is led to the *cathode*, whereby the cathode reaction for hydrogen formation takes place:

$$H_2O + 2e^- \rightarrow O^{2-} + H_2$$

The oxygen ion is oxidized at the anode:

$$O^{2-} \rightarrow 2e^- + 1/2\,O_2$$

Typical operating conditions are:

Potential	1.3 V
Current density	0.4 A/cm^2
Operating temperature	900 °C
Electrical energy consumption	ca. 3 kWh per cubic meter (STP) H_2

The performance characteristics of the high-temperature electrolysis is also given in Figure 37.

The process has been tested on a laboratory scale [124-126]. A pilot plant for the production of 1 m^3 (STP) hydrogen per hour is under construction. A 3.5 MW pilot plant is planned [127] but a number of material problems must be solved before realization.

The reversal of this process is expected to have a potential in fuel cell applications. This applies to the aforementioned solid polymer electrolysis as well (⟶ Fuel Cells – Proton Exchange Membrane Fuel Cells (PEMFC)). The direct production of electric power from hydrogen, carbon monoxide or hydrocarbons with air or pure oxygen in fuel cells of this type is still in the first stages of development.

4.2.4. *Hydrogen as Byproduct from other Electrochemical Processes*

See also ⟶ Electrochemistry; ⟶ Chlorine.

Electrochemical processes for the production of sodium hydroxide, chlorine and chlorine compounds yield hydrogen as a byproduct. About 3 % of the world hydrogen production, but a higher percentage of the merchant hydrogen, is produced in this way. The economically most important electrolysis is the chlor – alkali process. At present there are three different commercially important technologies to produce sodium hydroxide and chlorine.

In the *mercury process*, hydrogen is generated by decomposition of sodium or potassium amalgam, which is itself formed in a concurrent reaction to the kinetically strongly inhibited direct hydrogen evolution (32) on the mercury cathode (33).

$$2\,H^+ + 2\,e^- \rightarrow 2\,H \rightarrow H_2 \tag{32}$$

$$Hg_x + Na^+ + e^- \rightarrow NaHg_x \tag{33}$$

$$2\,NaHg_x + 2\,H_2O \rightarrow 2\,NaOH + H_2 + 2\,Hg_x \tag{34}$$

The decomposition reaction (34) is carried out in either horizontal or vertical decomposers. Because of the hydrogen overvoltage, a catalytically active decomposition material (technically graphite) is used and high temperatures are necessary. Hydrogen from the amalgam decomposers must be freed from mercury [to <0.002 mg Hg/m^3 (STP)] by cooling, scrubbing with brine and use of iodine-activated charcoal or a catalytic process.

In the *diaphragm process*, in which the overall reaction (35) occurs, hydrogen is formed on a hollow cathode made of iron, which is separated from the anode region by a diaphragm.

$$NaCl + H_2O \rightarrow 1/2\,H_2 + 1/2\,Cl_2 + NaOH \tag{35}$$

Both processes are, at this time, being replaced by the membrane process because of environmental and economic reasons.

The *membrane process* accounted for 10 % of the world capacity in 1987. New plants projected are based almost exclusively on the membrane process. Figure 42 shows the principal reactions and products of a membrane cell.

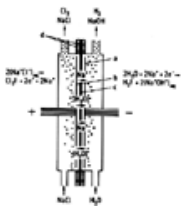

Figure 42. Schematic construction of an electrolysis cell for the production of chloride, sodium hydroxide, and hydrogen using the membrane process (usually bipolar construction)

a) Cathode; b) Anode; c) Membrane; d) Insulation

In the membrane process perfluorinated organic polymers with sulfonyl or carboxyl groups are used exclusively as membranes. The anodes are made of titanium, the cathodes of stainless steel or nickel.

Membrane cells can be in either unipolar or bipolar arrangement [128]. In addition to the main products chlorine and sodium hydroxide, the process yields high-purity hydrogen (> 99.9 %). Drying and removal of oxygen traces can be carried out, if required.

Hydrochloric Acid Electrolysis. Apart from the chlor – alkali electrolysis the electrolysis of hydrochloric acid (in particular for treatment of weak aqueous solutions to enable their disposal) has some economic importance. The cell products (chlorine and hydrogen) are saturated with water vapor and hydrogen chloride corresponding to the partial pressure of 20 % hydrochloric acid at 75 – 85 °C; both product streams are cooled and the remaining chlorine and hydrogen chloride removed from the hydrogen stream by scrubbing with sodium hydroxide. Hydrogen purity is 99.9 %. For further electrolysis processes of some economic importance, see ⟶ Electrochemistry.

4.3. Thermochemical Water Cleavage

In the range T < 2000 K and p = 100 kPa the cleavage of water is described by:

$$H_2O(l) \rightarrow H_2 + 1/2\,O_2 \tag{36}$$

The reaction is endothermic, the thermodynamic standard values are given below [129]:

Reaction enthalpy	$\Delta H^0_{298} = +286.01\ \mathrm{kJ/mol}$
Reaction entropy	$\Delta S^0_{298} = +163.35\ \mathrm{J\,mol^{-1}\,K^{-1}}$
Gibbs free energy	$\Delta G^0_{298} = +237.31\ \mathrm{kJ/mol}$

The equation for the equilibrium constant is as follows:

$$k_{p_{298K}} = \frac{p_{H_2} \times p_{O_2}^{1/2}}{p_{H_2O}} = e^{\frac{\Delta G}{RT}} = 4.2 \times 10^{-78}\sqrt{\mathrm{bar}}$$

The reaction enthalpy, ΔH^0, is fairly temperature independent, as is ΔS^0. As a result, the Gibbs free energy, ΔG^0, decreases substantially with increasing temperature (see Fig. 43). This behavior is typical for highly endothermic reactions.

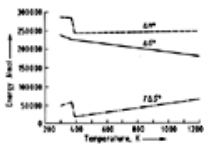

Figure 43. Gibb's energy, enthalpy, and entropy term of enthalpy ($T\ \Delta S^0$) for cleavage of water in a temperature range of 300 – 1200 K

The energy required for the production of hydrogen can be delivered in the form of chemical energy (see Section Production from Coal and Hydrocarbons), electrical energy (see Section Electrolysis), heat, light or radiation (see Section Other Methods for the Cleavage of Water), or in some combination of these various energy forms. In many cases, the total process consists of a number of steps using chemical, electrochemical, biochemical or photochemical systems. Purely *thermal cleavage* of water occurs only at >2500 K (see Fig. 48). A commercial realization of this reaction is presently not feasible.

In *thermochemical cyclic* processes reaction (36) is carried out in steps with the aid of chemical redox reactions utilizing energy at lower temperature levels. The substances involved in the reaction sequences are recycled. The great advantage compared to electrolysis is, that the low efficiency, resulting from production of electricity, can be avoided and thermal energy is directly converted into chemical energy. If one of the steps involved is an electrolysis step, the process is named a *hybrid cyclic process.*

The possibility of producing hydrogen via multistage cyclic processes was first pointed out by FUNK and REINSTROM [130]. In particular, the possibility of utilizing the heat of high-temperature reactors led to intensive investigations, and a number of processes for laboratory and pilot plants were selected.

4.3.1. *Thermodynamics of Closed Cycles*

Efficiency. A thermochemical cyclic process is subject to many thermodynamical constraints. The maximum attainable efficiency can be calculated according to [130].

$$\eta_{th} = \frac{\Delta H^0}{\Delta G^0} \cdot \frac{T_1 - T_2}{T_1}$$

T_1 is the upper process temperature of the heat source, T_2 the lower process temperature of the heat sink. The factor $\Delta H^0 / \Delta G^0$ is the ratio of the enthalpy to Gibbs free energy for the water cleavage reaction at 298 K and 100 kPa. The factor has a value of 1.2. Table 22 shows calculated efficiencies for T_2 = 298 K.

Table 22. Theoretical thermal efficiency of the thermochemical water cleavage (T_2= 298 K)

T_1, K	η, %	T_1, K	η, %
400	30.7	1000	84.6
600	60.6	1200	90.5
800	75.6	1400	94.8

Thermodynamical considerations show that in the range 300 – 1000 K at least three process stages are necessary to dissociate water thermochemically. At the same time, calculations show that the theoretical efficiency decreases as the number of stages increase [131].

If the heat losses and irreversible chemical reactions that are involved in practical processes are considered, the maximum efficiency attainable amounts to only 50 – 80 % of the theoretical value. However, the overall efficiency reached should still be clearly better than the total overall efficiency of water electrolysis (ca. 30 %).

Entropy-Temperature Diagram. A thermochemical cycle can be displayed graphically in an entropy-temperature diagram. Figure 44 shows this for the MARK 9 process (see Table 23). For simplification it is assumed that the reactions occur at equilibrium conditions (ΔG^0 = 0), that heat losses can be avoided and that no energy is necessary for the transport of reactants. Since the enclosed area corresponds to usable energy, in this case a constant equal to – ΔG^0 = 237.3 kJ/mol, the efficiency of the cyclic process can be determined using other thermodynamic data, e.g., the amount of heat absorbed by the process [132]. The thermodynamics of this process are summarized in [133], [134].

Table 23. Selection of reaction sequences of thermochemical cycles for water cleavage

Designation, Reaction sequence	Number of reactions	Maximum temperature, °C
Metal – bromine cycles		
MARK 1, (de Beni, Euratom)	4	730
$CaBr_2 + 2\ H_2O \Rightarrow Ca(OH)_2 + 2\ HBr$		
$2\ HBr + Hg \Rightarrow HgBr_2 + H_2$		
$HgBr_2 + Ca(OH)_2 \Rightarrow CaBr_2 + HgO + H_2O$		
MARK-1 B, MARK-1 C, MARK-1 S, MARK 5 use also bromine		
Iron – chlorine cycles		
MARK 9 (C. Hardy, Euratom)	3	650
$6\ FeCl_2 + 8\ H_2O \Rightarrow 2\ Fe_3O_4 + 12\ HCl + 2\ H_2$		
$2\ Fe_3O_4 + 12\ HCl + 3\ Cl_2 \Rightarrow 6\ FeCl_3 + 6\ H_2O + O_2$		
$6\ FeCl_3 \Rightarrow 6\ FeCl_2 + 3\ Cl_2$		

MARK 7, MARK 7 A/B, A 2 (Institute of Gas Technology),		
AGNES (General Electric) are further cycles of the Fe – Cl family		
Vanadium – chlorine cycles		
MARK 3 (Euratom), Allison div. of General Motors,		
RWTH Aachen (Fed. Rep of Germany)	4	800
$2\ VCl_2 + 2\ HCl \Rightarrow 2\ VCl_3 + H_2$		
$2\ VCl_3 \Rightarrow VCl_2 + VCl_4$		
$2\ VCl_4 \Rightarrow 2\ VCl_3 + Cl_2$		
$Cl_2 + H_2O \Rightarrow 2\ HCl + H_2$		
Metal – alkalimetal cycles		
Wentorf "C" (Li, K, Ni, J_2)	6	700
MARK-2 C (De Beni, Euratom, Na, Mn, C, Hg)	6	850
Methane cycles		
Methane – Methanol – Arsenous Oxide Cycle	5	700
$CH_4 + H_2O \Rightarrow CO + 3\ H_2$		
$CO + 2\ H_2 \Rightarrow CH_3OH$		
$CH_3OH + As_2O_4 \Rightarrow CH_4 + As_2O_5$		
$\frac{1}{2}\ As_2O_5 \Rightarrow \frac{1}{2}\ As_2O_3 + \frac{1}{2}\ O_2$		
$\frac{1}{2}\ As_2O_5 + \frac{1}{2}\ As_2O_3 \Rightarrow As_2O_4$		

Figure 44. Entropy – temperature diagram for MARK 9 cycle [131]

a) Hydrolysis of iron(II) chloride and hydrogen formation at 1100 K; b) Decomposition of iron(III) chloride releasing chlorine; c) Fusion of iron(III) chloride at 497 K; d) Chlorination reaction of magnetite at 497 K

The entropy changes at 298 K and 375 K include entropy of formation and vaporization of water. Without those values η is 0.477

4.3.2. *Further Criteria of Thermochemical Cycles*

Further criteria must be considered in order to judge the practicability of theoretical cycles. For example, the reactions should be in the temperature range of 300 – 1000 K; the reaction rate should be sufficiently large; chemistry and kinetics should be known. Furthermore, internal heat compensation should be possible; separation of the reaction products should be feasable using a minimum of energy with a minimum movement of solids. Finally, process and material know-how should be available. A further important condition for the development of a process is the nature of the high-temperature source, e.g., solar concentrator or nuclear high-temperature reactor, and the way this energy is integrated [135], [136].

With the help of computer programs and taking the aforementioned criteria into consideration, several thousand cyclic processes have been suggested. A summary of more than 200 processes can be found in the literature [137-139]. After the feasibility from a technical and process point of view has been investigated, 20 to 30 of these processes remain. These, finally, have been studied in greater detail.

Thermochemical cyclic processes can be categorized according to various aspects, e.g., the number of reaction stages, or the maximum temperature. However, it has become accepted to group them in

families according to the chemicals involved, e.g., iron-chlorine family (Fe, Cl_2, $FeCl_2$, $FeCl_3$), sulfur-halogen family (SO_2, halogen, X_2), copper-chlorine family (Cu, Cl_2, CuCl, $CuCl_2$). In addition, the basic reaction types for the thermochemical cyclic processes can be classified as follows:

Hydrogen formation usually occurs at low or medium temperatures (ambient to 400 °C). Examples are the reactions of metals (Eq. 37) or metal halides (Eq. 38) with acids or the decomposition of hydrides (Eq. 39).

$$2\,Cu + 2\,HCl \rightarrow 2\,CuCl + H_2 \tag{37}$$

$$2\,VCl_2 + 2\,HCl \rightarrow 2\,VCl_3 + H_2 \tag{38}$$

$$2\,HI \rightarrow I_2 + H_2 \tag{39}$$

Oxygen formation takes place at high temperatures (700 – 1000 °C) by oxygen elimination from oxides or oxygen-containing acids. The thermal cleavage of sulfuric acid (Eq. 40), for example, is one of the best-investigated high-temperature process steps:

$$H_2SO_4 \rightarrow H_2O + SO_2 + 1/2\,O_2 \tag{40}$$

Water is introduced into the cycles and allowed to react at various temperature levels. For example, the Deacon reaction (41) is of great interest at *high temperatures*, while at *low temperatures* the reaction of water with bromine or iodine in the presence of sulfur dioxide [Bunsen reaction (42) known since 1853] is particularly important:

$$Cl_2 + H_2O \rightarrow 2\,HCl + 1/2\,O_2 \tag{41}$$

$$X_2 + 2\,H_2O + SO_2 \rightarrow 2\,HX + H_2SO_4 \tag{42}$$

4.3.3. *Thermochemical Cyclic Processes*

Table 23 shows a selection of the various “purely thermochemical” reaction systems described in the literature. The Mark I cycle (Mark cycles, nomenclature JRC, i.e., Joint Research Center, Ispra, Italy) with a four-step mechanism was developed by De Beni in 1969. At the moment the processes belonging to the halogen-sulfur family are invoking the most interest. As an example, the sulfur – iodine process (S – I process) developed by General Atomic Technologies [140] will be described in more detail.

The main reactions are sulfuric acid cleavage (43) at high temperature and the Bunsen reaction (42), a low-temperature step, in which the water cleavage takes place. Hydrogen is converted to a hydrogen halide and oxygen is incorporated in the sulfuric acid

$$2\,H_2O + SO_2 + I_2 \longrightarrow H_2SO_4 + 2\,HI \tag{42}$$

$$H_2SO_4 \longrightarrow H_2O + SO_2 + 1/2\,O_2 \tag{43}$$

$$2\,HI \longrightarrow H_2 + I_2 \tag{44}$$

$$H_2O \longrightarrow H_2 + 1/2\,O_2$$

Using the molar amounts suggested by General Atomic, the Bunsen reaction can be written as:

$$16\,H_2O + 9\,I_2 + SO_2 \longrightarrow \begin{matrix} (H_2SO_4 + 4\,H_2O)^1 \\ (2\,HJ + 10\,H_2O + 8\,I_2)^2 \end{matrix}$$

The excess of iodine leads to a separation of two liquid phases, a lighter phase of aqueous sulfuric acid (1) and a heavier one consisting of hydrogen iodide, water and iodine (2). The process is subdivided into 4 stages as demanded by the necessary basic operations (Fig. 45 A) [140].

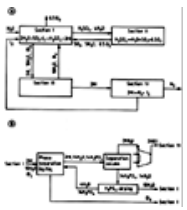

Figure 45. A) Material flows in the S – I process according to General Atomic

B) Section III within the S – I process according to General Atomic

In *stage 1* utilizing the Bunsen reaction, the separation of acidic products H_2SO_4 and HI and oxygen occurs. The oxygen is obtained in a mixture with sulfur dioxide from the sulfuric acid decomposition. This gas mixture is passed through the liquid mixture of water, sulfuric acid, iodine and hydrogen iodide. Here sulfur dioxide reacts according to the Bunsen reaction and oxygen remains as the sole gaseous component. The exothermic Bunsen reaction is performed at 120 °C, to make certain, that the products remain in the liquid phase.

Stage 2 represents the drying of sulfuric acid and the sulfur trioxide cleavage. In this stage pure sulfuric acid is produced in a six-stage drying process at medium temperature. In an ensuing evaporation the dry sulfuric acid decomposes into water and sulfur trioxide. After this, sulfur trioxide is catalytically decomposed to sulfur dioxide and oxygen by using high-temperature energy.

In *stage 3* pure hydrogen iodide is extracted from the heavy product phase of the Bunsen reaction and can then be decomposed economically in stage 4 using low-temperature energy. To obtain pure hydrogen iodide for the decomposition in *stage 4*, the heavy product phase of the Bunsen reaction is separated into three pure fractions. Figure 45 shows a simplified process scheme of this stage. The addition of highly concentrated phosphoric acid to the system causes a new phase separation, whereby the heavy phase contains iodine and the light phase hydrogen iodide, water and phosphoric acid. The light phase can be separated off at moderate temperatures. The gaseous product at the head of the separation column is almost pure hydrogen iodide. This is cooled, compressed to 5 MPa and fed to stage 4 in the liquid state.

The bottom product of the column consists of aqueous phosphoric acid. To be able to recycle the phosphoric acid, it must be freed from the excess water. This drying is achieved in a multistage process with vapor compression and consumes a large amount of electrical energy.

The following suggestions have been made to improve stages 2 and 3: power generation by decompression during sulfuric acid drying [141] and optimization of the hydrogen iodide separation with integrated hydrogen formation [142]. This process has been operated on a laboratory scale with a thermal efficiency of 47 %; the aforementioned improvements and optimization of the heat system is said to make an efficiency of 55 % possible.

Several variations of the S – I process are being investigated, e.g., the NIS process with the system Ni – $NiSO_4$ [143], the Mg – S – I cycle with the system MgO – $MgSO_4$ and the C – I – S process using methanol as a reactant as suggested by Japan Atomic Research Institute [144].

4.3.4. *Hybrid Cycle Processes*

Thermochemical – electrochemical hybrid cycle processes are a special form of thermochemical cyclic processes. Here an electrochemical reaction is involved. Hybrid cycle processes can, therefore, usually be designed as two-stage processes.

The *electrochemical* reaction stage is normally run at low temperature, and replaces either the hydrogen producing or the oxygen producing reaction. The necessary energy for the electrochemical reaction must be less than that required for direct electrolysis of water. Table 24 summarizes the thermochemical – electrochemical hybrid cyclic processes.

Table 24. Thermochemical – electrochemical cycles (hybrid cycles)

Designation, reaction sequence	Number of maximum reactions	Temperature, °C
Sulfur – halogen hybrid cycles		
MARK 13 (JRC Ispra)	3	650 – 850
$Br_2 + SO_2 + 2\,H_2O \Rightarrow 2\,HBr + H_2SO_4$		
$H_2SO_4 \Rightarrow H_2O + SO_2 + ½\,O_2$		
$2\,HBr \Rightarrow H_2Br_2$ (electrolytic step)		
Sulfur hybrid cycles		
MARK 11 (Westinghouse Electric Corp., JRC Ispra, General Atomic Technologies)	2	900
$H_2SO_4 \Rightarrow SO_2 + H_2O + ½\,O_2$		
$SO_2 + H_2O \Rightarrow H_2SO_4 + H_2$ (electrolytic step)		
Alkali metal hydride – hybrid cycles		
Me/MeH – hybrid process (F. Behr)	2	700

$H_2O + 2\ Me \Rightarrow 2\ MeH + ½\ O_2$ (electrolytic step)		
$2\ MeH \Rightarrow 2\ Me + H_2$		
Hydrocarbon – hybrid cycles		
Methane – methanol – methanal process (KFA Jülich)	4	950
$CH_4 + H_2O \Rightarrow 3\ H_2 + CO$		
$CO + 2\ H_2 \Rightarrow CH_3OH$		
$CH_3OH \Rightarrow CH_2O + H_2$		
$CH_2O + H_2O \Rightarrow CH_4 + O_2$ (electrolytic step)		

The most important hybrid cyclic process is the so-called Westinghouse cycle (MARK 11, according to the Ispra nomenclature). A large amount of work on this sulfur-family process and some of its steps have been carried out by Los Alamos Scientific Laboratories, Westinghouse Electric Corp., JRC Ispra, KFA Jülich, General Electric and others.

The sulfur hybrid cycle is a simple two-stage cyclic process. The first step is the electrochemical *low-temperature reaction* consisting of the steps (1) cathodic hydrogen formation (Eq. 45) and (2) anodic oxidation of sulfur dioxide to sulfuric acid (Eq. 46).

$$2\,H^+ + 2\,e^- \rightarrow H_2 \tag{45}$$

$$SO_2 + 2\,H_2O \rightarrow H_2SO_4 + 2\,H^+ + 2\,e^- \tag{46}$$

In the *high-temperature reaction* (Eq. 47) the electrically produced sulfuric acid is cleaved into oxygen, water, and sulfur dioxide, the latter is recycled to the anodic process.

$$H_2SO_4 \rightarrow SO_2 + H_2O + 1/2\,O_2 \tag{47}$$

This reaction can be accelerated in the temperature range of 500 – 1000 °C with catalysts (Fe_2O_3, CuO, Pt – TiO_2, Pt – SiO_2). Figure 46 depicts the flow sheet of the MARK 11 process. The stages of developement for the key steps of the process are as follows:

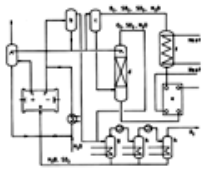

Figure 46. Schematic of the Westinghouse (Mark 11) hybrid sulfuric acid cycle, simplified [145]

a) Electrolysis; b) H_2SO_4 surge tank; c) Acid separator; d) Acid concentrator; e) Acid vaporizer; f) SO_3 decomposition reactor; g) Steam condensor; h) SO_2 condensor

Electrolysis. The desired conditions to be attained for the anodic reactions are:

50 wt % H_2SO_4
SO_2 pressure 100 – 200 kPa
temperature 80 °C

The high anodic overvoltage makes potentials of 680 – 800 mV necessary to obtain current densities of 200 mA/cm^2. To reduce this overvoltage catalysts are being tested. Diffusion of sulfur dioxide from the anode to the cathode and its reduction to sulfur or hydrogen sulfide there must be prevented. This is being attempted by use of H^+-permeable separator membranes (SPE, diaphragms) [146], gas diffusion electrodes as anodes [147], or three-compartment electrolysis cells [148].

Concentration and Cleavage of Sulfuric Acid. The ca. 50 wt % sulfuric acid produced at the anode is first concentrated to maximum 98 wt % in several concentration, temperature and pressure stages, then evaporated at 450 °C and 20 MPa and subsequently decomposed (at 400 – 800 °C into SO_3 – H_2O, at > 800 °C into SO_2 – O_2). Material corrosion problems occur. Silicon carbide and silicon nitride

materials are suitable for the evaporation equipment. For the decomposition of sulfuric acid special alloys with aluminum oxide coating (SS 444, Incolloy 800, Hastelloy G) have been proposed as reactor materials. The sulfur trioxide decomposition has been investigated on a pilot scale at the JRC Ispra in the so-called Christina process. This process has the following special features:

heat provision by hot air
adiabatically-operated cracking reactor with a special construction (see Fig. 47)
downstream heat recovery section.

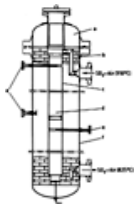

Figure 47. Design of the reactor for SO_3 decomposition for the Christina process

a) Quartz wool; b) Refractory lining; c) Catalyst container (Incoloy 800); d) Catalyst (iron oxide); e) Thermocouples; f) Pressure vessel (500 °C)

Figure 47 shows the sulfur trioxide cracker which is constructed for a pressure of 2.5 MPa and, with 11 s residence time, gives 95 % conversion with a crack rating of 12 kW. The results of the investigation show that sulfur trioxide cracking (a step of various cycles of the sulfur family), is technically feasible [149].

For the MARK 11 and related processes the theoretical thermal efficiency is 40 – 45 %. At present, an efficiency of 30 % is attainable.

4.4. Other Methods for the Cleavage of Water

All processes mentioned in the following are either still in the laboratory or in the development stage. A review is given by BOCKRIS in [150], where further methods such as *plasmolysis*, i.e., water splitting by high-temperature plasmas, *magnetolysis*, i.e., electrolysis using a potential difference created by magnetic induction, and *magmalysis*, i.e., interaction of steam with fresh basaltic lava in volcanic tubes, are presented.

4.4.1. *Thermolytic and Radiolytic Processes*

Water Thermolysis. The direct thermal cleavage of water, theoretically the simplest method and a priori an ideal process, requires very high temperatures. Figure 48 shows the equilibrium composition of dissociating steam at atmospheric pressure as a function of temperature [151]. The six components present to a significant extent are: H_2O, H_2, O_2, H, O and OH. The degree of water dissociation as a function of temperature is:

2000 K	0.69 mol %
2300 K	2.64 mol %
2700 K	10.35 mol %
3000 K	22.4 mol %
3500 K	57.43 mol %

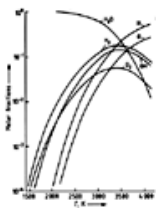

Figure 48. Molar fraction *a* of species for the dissociation of water at thermodynamic equilibrium as a function of temperature starting from pure steam at a pressure of 10^5 Pa

Appreciable quantities of hydrogen can only be obtained at temperatures significantly above 2000 K.

An economic utilization of this process is difficult, because of the recombination of the reaction products at the high temperatures. Other problems are associated with the recovery or separation of hydrogen from the hot mixture, conservation and recovery of the high-temperature energy and unsolved material problems.

The three different approaches generally considered are:

1. Catalytic thermal decomposition of water on metallic wires at 1600 K, the small degree of decomposition being compensated for by high reaction rates [152].
2. Decomposition of water in the vicinity of a high-temperature-resistant membrane which allows the removal of one of the products with accompanying displacement of thermodynamic equilibrium. Stabilized zirconia or porous ceramic membranes look promising [153].
3. Decomposition of water, followed by rapid quenching of product gases. A quenching speed of 10^5 – 10^6 K/s is required. Special quenching arrangements using cold steam or inert gas jets indicate that up to 90 % of the hydrogen formed can be recovered [154].

The research in this area is aimed at the utilization of high-temperature solar or nuclear energy.

Water Radiolysis. Water can also be cleaved by high-energy radiation. The overall formula for the water radiolysis is:

$$H_2O \rightarrow H, OH, H_2, H_2O_2, OH^-(aq), H^+(aq)$$

The principles of water radiolysis and radiation chemistry in general are described in [155].

Energy from *nuclear sources* is released as high-quantum energy photons (γ-radiation), high-speed particles, either charged (α-particles, protons, and β-radiation), or uncharged (neutrons). The rate of interaction of radiation with a medium, i.e., the rate of energy transfer, is termed linear energy transfer (LET). Pure γ-radiation sources have the smallest, uranium fission fragments the highest value (in water). The radiochemical yield (*G* value) depends on the LET value of the radiation. The *G* value is defined as the number of molecules converted per 100 eV absorbed energy. The $G(H_2)$ value is greatly influenced by the physical state and turbulence in the medium as well as the chemical composition.

The irradiation of water in the liquid state with γ-rays (^{60}Co) gives a $G(H_2)$ yield of 0.45, in the vapor state at 110 °C $G(H_2)$ is 5.2 [156]. With ^{210}Po a $G(H_2)$ value of 1.8 can be attained.

A well-known example of hydrogen formation is the undesired dissociation of the reactor coolant water in nuclear reactors (For safety measures see Section Removal of Hydrogen). Hydrogen production coupled with heat production in nuclear reactors has often been proposed or is the subject of experimental research. The most suitable configuration is seen in the so-called homogeneous reactor [157]. Here normal water serves as a solvent for the nuclear fuel (94 % enriched uranyl sulfate) and as a moderator. Using a $G(H_2)$ value of 1.6 a hydrogen yield of 28.8 kg H_2 per day and Megawatt can be calculated. Only a fraction (ca. 35 %) of the energy is used for the formation of hydrogen, the rest is converted to heat.

Hydrogen, along with higher hydrocarbons, is formed during the radiolysis of methane with a $G(H_2) = 6.4$. The radiolysis of higher aliphatic hydrocarbons yields, along with hydrogen ($G(H_2)$ ethane = 6.6, $G(H_2)$ propane = 8.2), a large number of other products [158].

Another indirect process for the production of hydrogen is considered to be of interest. Carbon monoxide, derived from the radiolysis of carbon dioxide, is converted to hydrogen and carbon dioxide in a subsequent shift reaction; carbon dioxide is then recycled. With a *G* (CO) value of 10 a higher hydrogen yield than obtainable with direct water radiolysis has been calculated [156].

4.4.2. *Photochemical or Photoelectrical Water Cleavage*

Direct photolytic cleavage of a water molecule leads initially to H and OH radicals

$$H_2O \rightarrow H + OH; \quad \Delta H = -457\,kJ/mol$$

The energy necessary corresponds to 1 mol of light quanta with an energy of at least 4.73 eV ($\lambda \leq 262$ nm). This process, however, only occurs in the vacuum UV range at a wavelength of $\lambda \leq 123$ nm, corresponding to a quantum energy of 10 eV [159]. Because of the unavoidable recombination reactions only a small fraction of hydrogen is found in the reaction mixture.

The photolytic cleavage of water with solar radiation is thus not possible; the process must be carried out in stages using several light quanta and with separation of the reactive intermediates. For the development of a process various possible routes can be followed.

Photochemical Systems. It has been known for a long time, that acidic water solutions of transition metal ions such as Cu^{2+}, Fe^{2+}, Eu^{2+}, evolve hydrogen upon irradiation with UV light. The metal ions provide an electron and are thereby oxidized. The quantum yield is relatively high. A cyclic water cleavage with recirculation of the oxidized metal ion back to the initial (reduced) stage is not possible except with Fe^{3+} and Ce^{4+}.

Efforts are presently concentrated on cyclic photo-redox processes with assistance of photosensitizers. The energy terms of such a system for the cleavage of water are shown in Figure 49. Photon energy, after exciting the sensitizer S to S*, is converted into chemical potential by creating the product ions S^+ and R^- (R is referred to as electron relay compound). This light-induced redox reaction is followed by catalytic steps causing water decomposition [160].

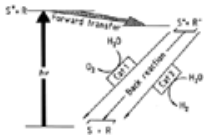

Figure 49. Energy term scheme of a light-induced redox reaction coupled with water cleavage [160]

S = sensitizer; R = electron relay compound; Cat^1, Cat^2 = catalysts

Examples for *sensitizers* are: $Ru(bipy)_3^{2+}$, porphyrine derivatives and acridine dyes such as proflavine. As *electron relay compounds* a viologen derivative, e.g., *N*-methyl-*N'*-tetradecylbipyridine, MV^{2+}, salicylate complexes of Eu^{3+}, V^{3+}, and macrocyclic cobalt complexes are used. Suitable catalysts are finely dispersed particles of platinum for formation of hydrogen and colloidal or macrodispersed RuO_2, which assists in the oxygen generation. A generalized overall reaction scheme is given in Figure 50. Difficulties to be overcome are the prevention of side-reactions and attaining a sufficient rate reaction. The former is achieved by using microheterogeneous systems (instead of homogeneous systems), the latter by using bifunctional catalysts, e.g., TiO_2 particles doped with RuO_2, which serve at the same time as support materials for platinum.

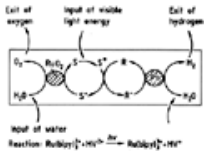

Figure 50. Schematic for cyclic water decomposition in a combined catalytic system [161]

A solution containing 10^{-4} mol/L $Ru(bipy)_3^{2+}$ and 2×10^{-3} mol/L MV^{2+} gives a quantum yield in hydrogen (and oxygen) formation of 2 %, with hydrogen yields being 400 mL/h per liter of substrate. This remains constant over at least 40 h [161].

Photoelectrochemical Systems. The photoelectrical water cleavage is based on the absorption of photons in a semiconductor electrode, which is in contact with an electrolyte. One part of the redox process, either oxidation or reduction, takes place on it. The opposing process takes place at the second electrode, thus separating hydrogen and oxygen evolution locally. Figure 51 depicts the energy term scheme of such a photoelectrochemical cell with an n-type semiconductor anode [162]. The steps of the cyclic process can be described as:

1. Absorption of photons creating charge separation by transportation of electrons from the valence band to the conduction band (electron – hole pairs);
2. Abstraction of electrons from the valence band of water, recombination with defect electrons (holes) and oxygen formation;
   Anodic reaction

   $$H_2O - 2e^- \rightarrow 1/2\,O_2 + 2H^+$$

3. Migration of hydrated protons to the cathode;
4. Transfer of electrons from the Fermi niveau of the metal to the protons followed by hydrogen formation.
   Cathodic reaction

   $$2H^+ + 2e^- \rightarrow H_2$$

The choice of the semiconductor is determined by the height of the band gap and its relative energy level. To cleave water, the energy difference between valence band and conduction band must be at least 1.23 eV; in actual fact, it must be much larger. The edge of the conduction band must be sufficiently above the redox Fermi level for water oxidation.

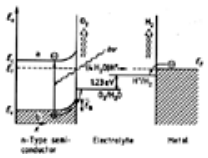

Figure 51. Energy scheme of the photoelectrolysis of water with a $SrTiO_3$ semiconductor electrode

a) Conduction band; b) Valence band

E_F = Fermi energy; E_G = energy gap between valence and conduction band; E_B = “buckling” of the energy level at the boundary layer electrode/electrolyte

With $SrTiO_3$ semiconductor anodes the band levels are at the right height. The band gap, however, is 3.2 eV, so that only the absorption of UV light is effective for charge separation. In case of the TiO_2 semiconductor electrode the conduction band has to be raised by 0.5 V above the hydrogen potential by application of an external voltage. A further complication is the photolytic decomposition of photoactive semiconductor surfaces (photocorrosion) of suitable electrodes [163].

In recent years the *light-induced cleavage* of water in a microheterogeneous semiconductor dispersion has been investigated (see [164]). As semiconductors titanium oxide, titanates, vanadium oxides and cadmium sulfide can be used. The principal mechanism is shown in Figure 52 A. Titanium oxide particles with a diameter of 100 – 500 nm form a transparent dispersion. Locally separated catalyst deposits of platinum and RuO_2 promote the water splitting reactions by reducing protons and forming oxygen; in the case of TiO_2, oxygen normally remains attached to the particle surface in the form of a peroxide.

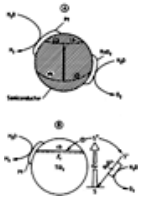

Figure 52. A) Catalytic water cleavage through band gap excitation of colloidal semiconductors

B) Complete water photolysis scheme through sensitization of a semiconductor particle

(vb = valence band; cb = conduction band, E_F = Fermi energy)

For particles of semiconductors with a large band gap a sensitization mechanism is required. Through light excitation, the sensitizer acquires the chemical potential necessary to promote an electron to the conducting band of the particle from where it is transferred to the platinum catalyst deposited on part of the surface. Following this, the redox reaction leading to water decomposition takes place as described above.

Figure 52 B shows a schematic diagram of the complete process. The sensitization of titanium oxide by forming surface derivatives with transition metal complexes shifts the photochemical excitation range down to 600 nm with a maximum at 480 nm, whereas pure titanium oxide exhibits a band-gap transition absorption only <400 nm [165]. In principle, it should be possible to use such an absorber system and to attain conversion efficiencies of the solar water cleavage up to 15 %.

4.4.3. *Hydrogen Formation in Biological Systems*

The capability of producing molecular hydrogen by biological action is inherent to some microorganisms including *chemotrophs*, i.e., organisms obtaining energy from chemical substrates, and *phototrophs*, i.e., those using light as a source of energy.

The essential step for the biochemical hydrogen formation is the reduction of two protons by the transfer of two electrons. This reaction is brought about by the interaction of appropriate electron-donating enzyme systems. The energy is supplied by any suitable substrate or by a photosystem using visible radiation. The process can take place in free organisms or in immobilized biological systems.

Hydrogen Production from Organic Substrates by Fermentative Bacteria. Some bacteria can be caused to ferment carbohydrates with the development of hydrogen and byproducts such as carbon dioxide. Obligate and facultative anaerobic microorganisms, e.g., various types of *Clostridium*, types of *Ruminococcus, Serratia marcescens*, and *Bacillus polymyxa*, are a means of carrying out this reaction.

In most cases, hydrogen is evolved as the result of anaerobic oxidation of pyruvate, an intermediate in carbohydrate metabolism. Starch, cellulose, and glucose can serve as substrates. The growth conditions of the microorganisms, the pH value of the substrate and the hydrogen partial pressure have a great influence on hydrogen production. Yields range from 0.1 – 2.5 mol H_2 per mole substrate ([166], [167]). The fermentation of carbohydrates is not of any importance from the point of view of hydrogen production. The energy conversion corresponds to ca. 20 % of the energy of combustion of the organic substrates.

Photobiological Production of Hydrogen. The ability to set hydrogen free with the aid of solar energy is found in a number of phototropic organisms such as purple bacteria, green bacteria, cyanobacteria (procaryotes) and several algae (eucaryotes). The majority are aquatic species. An overview is presented in [167].

The mechanisms vary from organism to organism, the main steps, however, are similar to those of the photosynthesis using electron-transporting photosystems connected in series. For example, with green algae, such as the Chlamydomonas types, the mechanism of the photobiological water decomposition follows the scheme depicted in Figure 53.

Figure 53. Photosynthetic electron transport system of hydrogen production [168]

PSI, PSII = photosystems I and II; Y, X = primary electron acceptors of photosystems I and II; Chl = chlorophyll; Q = plastoquinone; Cyt B, Cyt C = Cytochromes B and C; PC = plastocyanine; [Mn] = manganese carrying water oxidizing enzyme; Fd = Ferredoxin

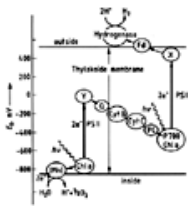

The active reaction centers, together with light-harvesting antenna pigments, are located in the so-called thylakoide membrane of chloroplasts. After irradiation, photosystem II (PS II) extracts electrons from water at a redox potential of + 0.85 V with the assistance of a manganese-containing enzyme. The electron acceptor of PS II delivers electrons to photosystem I through a down-hill electron chain. Photosystem I elevates the electrons to an energy level which allows the reduction of protons to hydrogen (E_0 = – 0.42 V). For the *normal* photosynthetic carbon dioxide fixation the electrons are transferred by means of the NADP – NADPH system at E_0 = – 320 mV, and carbohydrates are generated from carbon dioxide. In the case of *hydrogen formation*, however, the electrons are transferred to the protons by means of the low-molecular mass protein ferredoxin, a powerful reductant, and the enzyme hydrogenase.

Oxygen evolves on the inside and the ferredoxin reduction and hydrogen formation takes place on the outside of the thylakoide membrane (see Fig. 53). Since oxygen diffuses freely through the membrane and deactivates the extremely sensitive hydrogenase, such a system is of little interest for the commercial biophotolysis of water.

Purple bacteria lead to photosynthesis with sulfide, thiosulfate, sulfur or organic compounds as electron donors. Some strains (like *Rhodopseudomonas capsulata*) produce hydrogen and carbon dioxide during irradiation in the presence of organic compounds such as maleate, succinate or especially lactate [169]. Some mutants of *Rh. sphaeroides* may completely oxidize glucose to carbon dioxide and hydrogen. The necessary conditions (low concentrations of N-containing compounds and the absence of molecular nitrogen) indicate that the hydrogen photoproduction is catalyzed by the enzyme nitrogenase. The energy supply is most probably by photosystems I and II and the reduced electron carrier ferredoxin. At the moment, work is proceeding to develop an economical process for the photoproduction of hydrogen from highly polluted waste water using purple bacteria [170].

Cyanobacteria (blue-green algae) represent the most numerous group of phototropic procaryontes. Some of them are able to produce hydrogen. Many filamentous cyanobacteria (e.g., *Anabaena cylindrica* or *Nostoc muscorum*) have special types of cells (heterocysts) in their filaments of vegetative cells. These heterocysts contain the enzymes nitrogenase and hydrogenase, but are deficient in the photosystem II, and are thus not able to cleave water and to produce oxygen. The exceedingly oxygen-sensitive nitrogenase remains active because evolution of oxygen by photosystems I and II is confined to the vegetative cells. The carbohydrate of the vegetative cells provides the reductant for the nitrogen reduction to ammonia, respectively for hydrogen formation. Figure 54 shows schematically the interaction of the metabolism of vegetative cells and heterocysts [171]. The rate of hydrogen production of cyanobacteria is low, yet the feasibility with such systems is considered to be favorable, because hydrogen evolution takes place at ambient conditions and it is expected that suitable mutants can be selected.

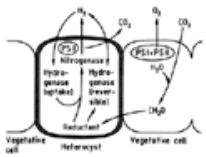

Figure 54. Model of hydrogen metabolism by heterocystous blue-green algae

PSI, PSII = Photosystems I and II

Cell-free systems containing chloroplasts, hydrogenases and an electron carrier such as ferredoxin, sometimes immobilized, have been investigated. In a comparatively short period of time a considerable increase in the rate of hydrogen evolution and stability of such systems has been achieved. Latest data show, that this process can go on for ca. 6 h with a rate of hydrogen production amounting to 100 mmol per hour and gram of chlorophyll [172]. Reasons for the stop of the reaction are inhibition of hydrogenases by oxygen, loss of activity of the chloroplasts and autooxidation of electron carriers.

Table 25 shows hydrogen evolution rates by phototropic microorganisms and cell-free systems [167]. The ultimate aim is the development of hydrogen-producing photoreactors containing artificial enzyme systems, but results are not promising at present.

Table 25. Maximum rate of hydrogen evolution by biological systems [167]

Phototropic microorganisms	H_2 evolution per hour and gram dry biomass	
	mL	mmol

Purple bacteria		
Rhodospirillum rubrum	146	6.5
Rhodopseudomonas palustris A	54	2.4
Rh. capsulata SL	130 – 150	5.8 – 6.7
Rh. capsulata B 10	300 – 500	13.4
Rh. capsulata LB 2	178	7.9
Thiocapsa roseopersicina	13	0.6
Cyanobacteria		
Anabaena cylindrica B 629	5 – 40	0.2 – 2.0
An. variabilis	32	1.3
Spirulina platensis	9	0.4
Green algae		
Chlorella vulgaris	4	0.2
Chlamydomonas reinhardii 137 C	45	2.0
Chl. moevussii ICC 97	8	0.4

Cell-free systems*	H_2 evolution per hour and gram chlorophyll		Functioning
	mL	mmol	time, h
Chp + Fd + H_2-ase (*C. pasteur*)	200 – 350	9.0 – 15.5	3.0 – 6.0
Chp + Fd + H_2-ase (*Th. roseopersicina*)	110 – 220	5.0 – 10	3.0 – 6.0
Chp + Fd + H_2-ase (from various sources)	450 – 900	20 – 40	4.0

* Selected basic components: chloroplasts of spinach (Chp), ferredoxin (Fd), hydrogenases of various sources; electron donor H_2O; the experiments were carried out in the presence of O_2-scavengers

4.5. Other Chemical Processes

4.5.1. *Hydrogen from Conversion of Metals*

Alkali and alkaline earth metals react with water or steam according to Equation (48) to yield hydrogen. For hydrogen production in the laboratory the reaction of metals with acids or bases is still used (Eq. 49 and 50).

$$\mathrm{Me + H_2O \rightarrow 1/2\,H_2 + MeOH} \tag{48}$$

$$\mathrm{Zn + 2\,HCl \rightarrow H_2 + ZnCl_2} \tag{49}$$

$$\mathrm{Al + NaOH + 2\,H_2O \rightarrow 3/2\,H_2 + NaAlO(OH)_2} \tag{50}$$

Aluminum, when the surface has been activated by amalgamation, reacts with *hot water* with hydrogen evolution [173]. Zinc, that has been activated with copper, can split *water vapor.*

Hydrogen can be obtained from metal hydrides by heating them as well as by subjecting them to the influence of water (⟶ Hydrides, see Section Hydride Technology). For certain purposes, especially in out-of-door duties (balloons) hydrogen can be produced from lithium hydride, which is light and easily transported (2.8 L H_2 per gram of LiH), from calcium hydride (1.06 L per gram of CaH_2) or from complex hydrides, such as lithium alanate (2.36 L per gram of $LiAlH_4$).

The steam – iron process, which is one of the oldest ways of producing hydrogen, has been subject to reawakened interest. In principle, it is a cyclic process for water cleavage, whereby coal is consumed. The principles of the steam – iron process is shown in Figure 55. In the first step, coal is gasified with steam and air to produce a reducing gas containing carbon monoxide and hydrogen (lean reducing gas). This gas is reacted with iron oxides in the second stage to produce reduced forms of iron oxides. The reduced species are then reoxidized in the third step of the process with water to form the original oxides at the same time producing hydrogen.

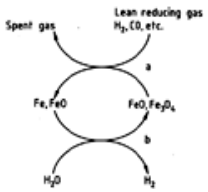

Figure 55. Principle of the steam – iron process

a) Iron oxide reduction reaction; b) Water cleavage reaction

Characteristic data for hydrogen production by the steam – iron process in generators, as they were previously used, are:

Table 26.

Periodical operation,	
cycle time	15 – 30 min
reduction time : oxidation time	2 : 1
mean reaction temperature	700 – 900 °C
hydrogen production per generator	250 m^3/h

Continuous production of hydrogen can be achieved if reduction and oxidation of the solids occur in two separate reaction zones and the solid is transported between them. New process suggestions [174] incorporate separated fluidized beds, as can be seen from Figure 56. The reduction gas is produced externally in a coal – steam – air gasification stage (a). The main disadvantage of the steam – iron process is that only 60 % of the reduction potential of the gas is utilized. Power production by both gas and steam turbines and production of process steam may be used to compensate for this.

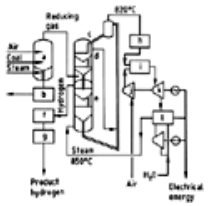

Figure 56. Process diagram of the continuous hydrogen production by the steam – iron process

a) Reducing gas producer; b) Ash disposal; c) Steam – iron system; d) Reducter; e) Oxidizer; f) Waste heat recovery; g) Desulfurization; h) Dust removal; i) Combustor; j) Air compression; k) Gas turbine; l) Steam generation with steam turbine

Other redox systems for water cleavage which, according to thermodynamics are more efficient than iron, have been suggested, e.g., Zn – ZnO dissolved in lead, 300 – 500 °C, 10 – 20 MPa [175], and Sn – SnO_2 dissolved in tin, ca. 800 °C, 2 MPa [176].

4.5.2. *Hydrogen from Ammonia*

Ammonia is easily thermally decomposed according to the reaction:

$$2\,NH_3 \rightarrow N_2 + 3\,H_2 \quad \Delta H_{298} = +91.86\ \mathrm{kJ/mol}$$

The endothermic reaction is favored by high temperature and low pressure and is accelerated by the presence of nickel or iron catalysts. From one kilogram of ammonia 1.97 m^3 (STP) of hydrogen and 0.66 m^3 (STP) of nitrogen are obtained with an energy requirement amounting to 2.7 MJ without consideration of preheating and evaporation. The residual ammonia concentration is usually lower than 500 ppm (mL/m^3) depending on the equilibrium conditions. Figure 57 shows the corresponding residual ammonia concentrations at 100 kPa with the temperature as parameter. At higher pressure, a higher ammonia content remains in the product gas (e.g., 1000 °C, 3 MPa: NH_3 residual concentration 0.2 vol %).

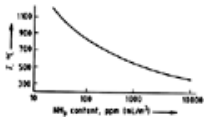

Figure 57. Residual ammonia concentration corresponding to equilibrium of ammonia formation reaction at 100 kPa

The reaction is conducted commercially at 800 – 900 °C. Standard plants for 1 – 100 m^3 (STP)/h are heated by a separate fuel gas and require water to cool down the product gas. The process can also be conducted autothermally, either by burning a part of the product hydrogen or by preigniting ammonia before the catalyst reactor. This results, however, in a mixture containing less hydrogen (ca. 50 vol %) and containing the product of combustion, water vapor. Usually carbon-free combustion air is used to avoid traces of carbon dioxide and carbon monoxide in the product gas [177].

The hydrogen – nitrogen mixture produced by cracking can be used without further treatment in various areas, e.g., as a protective atmosphere in metal processing (see Section Hydrogen in Metallurgy), or for catalytic oxygen removal from gases. For the recovery of pure hydrogen on a small scale, suitable process steps can be used, e.g., pressure-swing adsorption, diffusion cells (see Section Permeation Processes) or hydride technique (see Section Hydride Technology).

The possibility of hydrogen generation from ammonia leads to consideration of the use of anhydrous liquid ammonia as medium for hydrogen distribution and storage (ammonia as energy carrier, see Section Future Developments). Liquid ammonia contains 17.8 wt % hydrogen, which is 50 % more hydrogen per volume than liquid hydrogen [178].

4.5.3. *Hydrogen from Methanol*

Hydrogen and carbon monoxide can be obtained from the decomposition of methanol by reversal of the synthesis reaction. The endothermic decomposition occurs at temperatures >700 °C without a catalyst, or at 300 – 450 °C on catalysts based on copper – nickel or zinc –chromium alloys. Higher hydrogen yields are obtained by catalytic steam reforming of methanol. The process proceeds via initial methanol decomposition (Eq. 51) followed by a water-gas shift reaction (Eq. 52) as shown in the overall equation (Eq. 53).

$$CH_3OH \longrightarrow CO + 2\,H_2 \qquad \Delta H_{298} = +\,91.7\ \text{kJ/mol} \quad (51)$$

$$CO + H_2O \longrightarrow CO_2 + H_2 \qquad \Delta H_{298} = -\,41.0\ \text{kJ/mol} \quad (52)$$

$$CH_3OH + H_2O \longrightarrow CO_2 + 3\,H_2 \qquad \Delta H_{298} = +\,50.7\ \text{kJ/mol} \quad (53)$$

Information on thermodynamic and kinetic investigations, suitable catalysts and a summary of literature, is given in [179-181].

The reaction is carried out technically at pressures of 0.7 – 3 MPa on CuO – ZnO or CuO –Cr_2O_5 catalysts. The molar ratio of water to methanol will vary from 0.67 to 1.5; excess steam lowers the risk of carbon formation.

Hydrogen production from methanol is used in small to medium-sized plants (some of which are mobile) for the production of 1 – 2 000 m^3 (STP)/h. Purification of the product gas is performed by carbon dioxide scrubbing, pressure-swing adsorption or (especially for small units and high purity), by diffusion through a palladium-silver membrane. Figure 58 shows the schematic of a methanol cracker for 50 m^3 (STP)/h of hydrogen at 1.4 MPa using a PSA purification unit to give 99.9999 % purity. 1.5 m^3 (STP) of hydrogen are produced per kilogram of methanol.

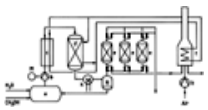

Figure 58. Methanol reformer with pressure swing adsorption purification for the production of 50 m^3 (STP) hydrogen per hour, purity 99.9999 vol % (German Linde)

a) Methanol/water vessel; b) Pump; c) Preheater; d) Reformer; e) Adsorber; f) Heater for heat transfer medium; g) Condensate feedback; h) Cooling water

In Berlin, peak requirements for town gas are covered by methanol cracking. At 1.5 – 20 MPa and ca. 500 °C endothermic methanol reforming reaction and exothermic methanation of the carbon monoxide takes place and a mixture with town gas quality (ca. 50 % H_2, 25 % CH_4, 4 % CO, 20 % CO_2) is obtained. The investment costs are ca. 40 % lower than for a naphtha reformer and the operating costs are also favorable, so that the gas production costs are relatively low, for the small number of days with peak loads [182].

4.5.4. *Hydrogen from Hydrogen Sulfide*

A potential source of hydrogen is hydrogen sulfide which can be obtained in large amounts from scrubbing processes for synthesis and natural gas as well as from desulfurizing processes in refineries. Normally, hydrogen sulfide is converted to elemental sulfur by means of the Claus process or by other oxidizing processes. The recovery of hydrogen and sulfur from waste hydrogen sulfide is an interesting alternative to the manufacture of sulfur alone.

The decomposition reaction is endothermic. At temperatures <1800 K the equilibrium position is thermodynamically unfavorable with respect to hydrogen formation (see Table 27) [183].

Table 27. Calculated equilibria concentration of hydrogen in the decomposition of hydrogen sulfide at 100 kPa [191]

Temperature, K	Equilibrium concentration, mol %
400	1.3×10^{-3}
500	6.1×10^{-3}
600	2.0×10^{-2}
700	4.8×10^{-2}
900	1.9
1073*	7.1
1200	13.1
1400	25.6
1600	37.7

* $K = 1.45\times10^{-2}$.

$$2\,H_2S \rightarrow 2\,H_2 + 1/4\,S_8 \quad \Delta H_{298} = 84.90\,kJ/mol$$

The reaction proceeds readily in the presence of catalysts, e.g., platinum – cobalt (at 1000 °C) [184], disulfides of molybdenum or tungsten (at 800 °C) [185], or other transition metal sulfides, preferably supported on alumina (at 500 – 800 °C) [186].

Development work to increase hydrogen yield and thus to establish commercial use of hydrogen sulfide decomposition, suggests several approaches to avoid thermodynamic restrictions.

Cyclic Processes. Thermodynamically favorable cycles have been suggested, e.g.:

$$H_2S + I_2 \rightarrow 2\,HI + S \tag{54a}$$

(30.0 MPa, 300 K)

$$2\,HI \rightarrow H_2 + I_2 \tag{54b}$$

(600 K, 70 MPa, activated carbon) [187]

$$H_2S + CO \rightarrow H_2 + COS \tag{55a}$$

(470 K, 100 kPa, over COS, NiS)

$$COS \rightarrow CO + S \tag{55b}$$

(1000 K, 100 kPa) [188]

$$aH_2S + H_xC_y \rightarrow yCS_2 + (a+x)/2\,H_2 \tag{56a}$$

$$yCS_2 + ySO_2 \rightarrow yCO_2 + 3\,y/2\,S_2 \tag{56b}$$

$$y/2\,S_2 + yO_2 \rightarrow ySO_2 \tag{56c}$$

(suggested [189])

Equilibrium Displacement. The hydrogen yield can be considerably enhanced by removing either sulfur or hydrogen from the reaction system, thereby shifting the equilibrium. A variety of methods have been considered for the separation of hydrogen from hydrogen sulfide in hot gas streams. Examples are the use of zeolitic, polymeric, metallic, glass and ceramic oxide membranes, pressure-swing

adsorption and thermal diffusion [190]. The hydrogen yields can be increased to twice the equilibrium value by diffusion through Vycor-type glass membranes or a microporous alumina membrane at 1000 °C without degradation of the latter [191]. Laboratory experiments using thermal diffusion columns at moderate to high temperature showed effective separation of hydrogen sulfide decomposition mixtures yielding hydrogen concentration > 90 % [190].

Economic realization of hydrogen sulfide cleavage would provide commercial products and offer the possibility of replacing Claus processes and of recycling hydrogen in refinery hydrodesulfurization processes. However, energy has to be provided at relatively high temperatures. Research is focused on (1) thermal decomposition processes using concentrated solar radiation or (2) photoelectrochemical hydrogen sulfide splitting by irradiation with visible light using colloidal semiconductor electrodes (see Section Photochemical or Photoelectrical Water Cleavage, [192]).

4.5.5. *Hydrogen as Byproduct*

In industrial organic chemistry reasonable amounts of hydrogen are often generated. The amounts are particularly large in the petroleum refineries within the area of catalytic reforming. The most important reactions here are thermal or catalytic cracking, dehydrogenation and oxidation (see Table 28 and Sections Refinery Processes and Petrochemical Processes).

Table 28. Reactions in industrial organic chemistry providing hydrogen

Reaction	Product	Process/Others
Thermal and catalytic cracking		
Hydrocarbons, e.g., $2CH_4 \xrightarrow{\Delta T} C_2H_2 + 3H_2$	acetylene	generally not used anymore. 57% H_2 in crackgas
Hydrocarbons, e.g., $C_2H_6 \xrightarrow{\Delta T} C_2H_4 + H_2$	ethylene	ethane cracking in ethylene plants
Reforming processes in refineries		250 m^3 H_2 per tonne of naphtha
Dehydrogenations		
C_4-fraction: 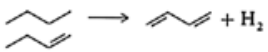	butadiene	Dow process, Houdry process
n-Paraffins $CH_3-(CH_2)_nCH_3 \longrightarrow R-CH=CH_2$	higher olefins	UOP process
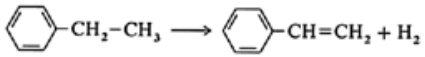	styrene	
cyclohexanol–cyclohexanone mixtures	phenol	partly commercially used
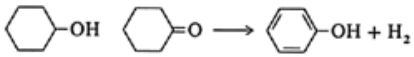		
Oxidations		
$CH_3-CH_2-OH \longrightarrow CH_3-C(=O)H + H_2$	acetaldehyde	Carbide and Carbon Corp.
2-Propanol *sec*-Butanol $\longrightarrow R-C(=O)-CH_3 + H_2$	acetone methylethyl ketone	Standard Oil process IFP process
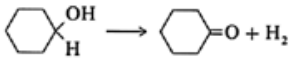	cyclohexanone	step in caprolactam production
Miscellaneous reactions		
$2\,CH_3-CH_2-OH \longrightarrow$ $+ H_2O + H_2$	butadiene	Lebedew.process, obsolete, of interest for developing countries
$CH_2(OH)-CH_2-CH_2-CH_2(OH) \longrightarrow$ $+ H_2$	butyrolactone	BASF process, use as solvent
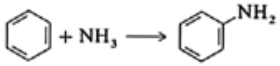	aniline	Du Pont, development state

4.6. Economic Aspects

At present 77 % of the hydrogen produced is from petrochemicals, 18 % from coal, 4 % by electrolysis of aqueous solutions and at the most 1 % from other sources. A review of the cost structure of hydrogen production will, therefore, pay particular attention to the production methods described in Section Production from Coal and Hydrocarbons. Electrolytic processes are also used under certain economic circumstances and are, therefore, included here and in a prognosis regarding the future supply of industrial hydrogen. For predictions of hydrogen production costs for use as an energy carrier see Chapter Hydrogen Energy.

4.6.1. *Analysis of the Cost Structure*

Hydrogen and hydrogen-rich synthesis gases (in particular H_2 – CO and H_2 – N_2) are either produced directly in plants especially built for that purpose or are recovered as a byproduct. In areas with highly concentrated industrial sites, such as mid-Europe, the ratio of directly to indirectly produced hydrogen is almost 1 : 1 [193]. In the developing countries and in countries with industrial sites widely separated from one another, the fraction of directly produced hydrogen is much larger than the fraction recovered in combined plants.

According to the broad range of feedstocks and the highly developed production processes from each of these feedstocks, an analysis of the cost structure depends on the local conditions, i.e., on the availability and price of the feedstocks. In many countries feedstock prices, such as those used for the direct production of hydrogen, are subject to intervention by the government. Heavy taxation is the rule, but subsidies also occur. Further uncertainties in the cost comparison are the various safety and environmental standards in each country and region which to a lesser extent, however, lead to different hydrogen costs.

For the conditions in the Federal Republic of Germany in 1983 costs for direct production of hydrogen have been calculated [194]. The feedstock prices (Fig. 59) include costs for the feedstock itself and, because in the cases considered the feedstock is also used as fuel, the fuel costs. The rest are capital, material and other energy costs (without the feedstock), personnel and administration costs. They lead to a typical situation, which is also found for other countries in which the feedstock is available at world market prices: Steam reforming of natural gas exhibits the lowest investment costs.

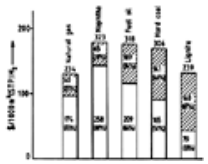

Figure 59. Feedstock and production costs for hydrogen in the Federal Republic of Germany in 1983 in DM/1000 m^3 (STP) H_2; numbers in columns converted to $ by the factor 1.8 (scale on the left)

plain area = costs of feedstock; hatched area = costs of production

The costs of hydrogen production from lignite are almost similarily low, whereby the low feedstock prices for surface-mined coal offsets the high cost of the plant. Whereas the much higher cost of hydrogen from naphtha or heavy fuel oil is a situation found worldwide, the high cost for hydrogen from hard coal is a special case limited to the Federal Republic of Germany, and reflects the special coalmining conditions there. On the world market the cost for coal is usually only slightly higher than the cost for lignite, so that the cost for hydrogen from hard coal should be much less.

The influence of the feedstock price is shown in Figure 60. The dashed lines correspond to the feedstock prices and the resulting hydrogen production costs in the Federal Republic of Germany in 1983 and correspond with the data given in Figure 59.

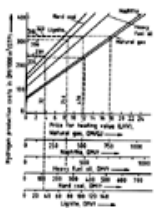

Figure 60. Production costs for H_2 as a function of feedstock price (Federal Republic of Germany, 1983)

As long as local conditions do not lead to large changes in the capital-dependent costs (ordinate of Fig. 59) the given presentation can be used to give a rough estimate of hydrogen costs and comparisons between various feedstocks by substituting local and actual prices.

The results in Figure 59 and 60 are for a capacity of 100 000 m^3 (STP)/h, which is a very large plant, although not unusual. If the dependence of the hydrogen production cost on plant capacity is required, then the simplified assumption may be made that the *utility costs* are independent of the capacity and that the *investment costs* become relatively less important as long as a large one-train plant can be built. If a modular construction is used, the investment costs do not decrease with increasing plant size. The upper limits for one-train units are ca. 500 m^3 (STP)/h for electrolysis units, ca. 80 000 m^3 (STP)/h for partial oxidation and ca. 100 000 m^3 (STP)/h for steam reforming plants.

Gas purification plants, in particular pressure-swing adsorption plants, are built in modules from ca. 50 000 m^3 (STP)/h upwards, so that a reduction in costs on capacity increase for larger plants is not to

be expected.

For comparison of hydrogen plants with a capacity of ca. 10 m^3 (STP)/h, it makes sense to choose plant types that would not be considered for very large plants, e.g., electrolysis plants or methanol reformers. Figure 61 shows the hydrogen costs as a function of the capacity for Europe in 1983. In addition to these production processes solely for hydrogen, the costs of recovery of hydrogen by means of pressure-swing adsorption from a hydrogen-rich refinery off gas are also given. In this case only the difference in heating value of the refinery gas before and after the hydrogen removal is used for calculation of the feedstock price. For the investment cost the cost of a pressure-swing adsorption unit without any further prepurification can be taken. The range of costs is caused by the additional expenditure for feedstock compression to adsorption pressure. Plants of this type up to a capacity of ca. 100 000 m^3 (STP)/h feed gas have been built.

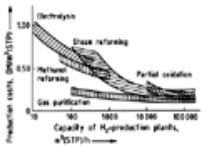

Figure 61. Production costs for H_2 as a function of the production process and plant capacity [195]

Basis of the cost determination:Investment costs: 1983, Federal Republic of Germany

Natural gas: 8.5 DM/GJ, HHV

Fuel gas: 8.5 DM/GJ, LHV

Power: 0.09 – 0.15 DM/kWh

Naphtha: 650 DM/t

Electrolysis: boundary limits 0.15 – 0.09 DM/kWh

Methanol reforming: boundary limits 525 – 400 DM/t

Steam reforming: boundary limits naphtha – natural gas

Partial oxidation: boundary limits heavy fuel oil 300 DM/t,residues 100 DM/t

A similar curve for U.S. prices of hydrogen was developed in 1978 with forecasts until 1980 [197]. Figure 61 also contains average commercial prices for 5580 m^3 (STP)/h merchant hydrogen which exceed prices for electrolytically produced hydrogen from a plant size of 0.1×10^6 scf/d (see Section Future Developments) on up.

The consequences of part-load operation of the plant also determines the choice of the hydrogen-production process. For a low-capacity rating, processes with low investment costs are preferred, because in this case the product cost depends more or less directly on the necessary feedstock costs and is not influenced by writing off an expensive high-capacity plant. For hydrogen plants in the lower capacity range (ca. 100 – 500 m^3 [STP]/h) electrolysis or methanol reforming is the best choice. For large plants this factor leads to the choice of steam reforming plants using natural gas.

4.6.2. *Prognosis of the Hydrogen Costs*

A short-term prognosis of the hydrogen cost development for industrial use (see Chap. Hydrogen Energy for hydrogen as energy carrier), should be oriented on the present stable supply situation and technically well-proven processes. With the aid of Figure 60 such a prognosis can be made. It shows that in the future hydrogen will still be preferentially produced on a large scale from hydrocarbons. The prognosis can therefore be reduced to investigating price and availability of hydrocarbons in the coming years.

Should it become possible to decouple the price of electrical energy from the price of fossil fuels by using nuclear energy or alternative energy sources and to make electrical energy relatively cheap, the electrolysis process could play an increasingly important role. Figure 62 shows what price changes must occur to assure electrolysis a greater market fraction on a purely economic basis.

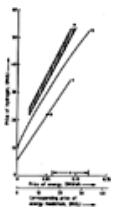

Figure 62. Cost prospects for electrolysis as compared to the main process for hydrogen production, steam reforming of natural gas [196]

a) Conventional electrolysis; b) Advanced electrolysis; c) Natural gas; d) Natural gas price (Mid Europe); e) Price range for electrical energy (Mid Europe)

[Top of Page]

5. Purification of Hydrogen

In many cases it is necessary to purify technical grade hydrogen by using a variety of processes. These must be chosen in accordance with the final use of the hydrogen to give an economically

satisfactory result. Chapter Purification of Hydrogen gives an overview of the technically important purification steps. They are classified according to the physical and chemical operations involved in the separation processes.

Frequently in industrial-scale plants combinations of the various processes are used. In addition, several alternative processes may be available and may be used depending on the economic situation. In Chapter Production and Chapter Handling of Hydrogen some examples of typical combinations of various gas production and purification processes are given.

5.1. Low-Temperature Processes

The large temperature difference between the boiling point of hydrogen and that of the most frequent impurities is the basis of the low-temperature process (LT process) for the purification of hydrogen. Table 29 shows important properties of hydrogen and the most frequent impurities present in the hydrogen to be treated. The impurities are condensed or sublimated. Low-temperature rectification processes are used, if the impurities are also to be separated from one another. The use of *scrubbing* and *adsorption processes* at low temperatures leads to both high purities and high yields.

Table 29. Boiling points at atmospheric pressure, triple point and critical data of hydrogen and some important components which can be present in raw hydrogen streams

Component	Boiling point at 101.3 kPa		Triple point data		Critical data	
	Temperature, K	°C	Temperature, K	Pressure, kPa	Temperature, K	Pressure, kPa
He	4.23	–268.9	2.17[b]	5.1	5.3	230
H_2[a]	20.4	–252.8	13.9	7.0	33.3	1 320
Ne	27.1	–246.1	24.6	43.3	44.5	2 760
Ar	87.3	–185.9	83.8	68.8	150.7	4 870
N_2	77.3	–195.9	63.2	12.6	126.1	3 390
O_2	90.2	–183.0	54.4	0.2	154.3	5 040
CO	81.6	–191.5	68.1	15.4	133.0	3 500
NO	121.2	–151.9	109.9	2.3	179.2	6 590
H_2O	373.2	100.0	273.2	0.6	647.4	22 140
H_2S	212.9	– 60.3	187.6	23.2	373.6	9 020
SO_2	263.2	– 10.0	197.8	1.7	430.4	7 870
CO_2	194.7[c]	– 78.5	216.6	18.0	304.2	7 390
NH_3	239.8	– 33.4	195.4	6.1	405.6	11 300
CH_4	111.7	–161.5	90.7	11.7	190.7	4 620
C_2H_6	184.6	– 88.6	89.9	<0.1	305.3	4 890
C_2H_4	169.5	–103.7	104.0	0.1	282.0	6 280
C_3H_8	231.1	– 42.1	85.4	<0.1	370.0	4 260
C_3H_6	225.5	– 47.7	88.0	<0.1	365.3	4 570

[a] In more detail in Section Physical Properties.
[b] Upper λ point.
[c] Sublimation

An overview of the basic physical processes which are used for the liquefaction of gases as well as the most important parts of the plants, equipment and machines is given in [198] and [199]. Properties of two- and multi-component mixtures with hydrogen, which are the basis for the calculation of the separation processes given in this chapter, can be seen in Figures 8, 9 and 10 (see Chap. Properties).

Such property data are available in computer data banks of engineering companies. Some data otherwise available is given in [200], [201].

5.1.1. *Condensation Processes*

Simple condensation processes are used, if the difference between the boiling point of hydrogen and the condensing phase, usually hydrocarbons, is sufficiently high. Hydrogen is frequently separated from refinery and dealkylation off-gases, as well as ammonia and methanol synthesis purge gases by the condensation process.

The gas from which hydrogen is to be recovered, usually contains components which are solid at the temperature of the final condensation step. Therefore, a *prepurification* stage is usually required to prevent blockage of heat exchangers, knock-out drums, and piping in the low-temperature section of the plant.

Prepurification and drying remove in particular impurities with a high boiling point such as water, carbon dioxide, and hydrocarbons such as benzene and toluene. These processes (often carried out in stages) are *scrubbing* (e.g., glycol scrubbing for water removal), *adsorption* (e.g., on activated carbon or molecular sieve beds), or *condensation* and *sublimation* of the impurities on reversible regenerator beds or reversing heat exchangers (Revex).

The residual gas mixtures, usually hydrogen – methane, hydrogen – carbon monoxide, or hydrogen – nitrogen mixtures, are cooled down below the condensation temperature. The lower the temperature chosen, the purer is the hydrogen product. In the case of the mixture hydrogen – methane the temperature in the condensator can be lowered below the solidification point of pure methane by the addition of ethane or propane and the hydrogen purity can thereby be increased.

The hydrogen yield depends on the solubility of the hydrogen in the condensate and on the amount of condensate. The yield of hydrogen increases, if the condensate is depressurized in a series of stages and the hydrogen-rich depressurization gas is recycled to the first stage of the compressor. Figure 63 shows the most important ways of generating refrigeration for low-temperature processes in industrial plants.

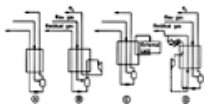

Figure 63. Principle ways of producing refrigeration for low-temperature hydrogen purification

A) Joule – Thomson expansion of residual gas; B) Hydrogen expansion in a turbine; C) Cold from external refrigeration sources (NH_3, freon, ethylene, etc.); D) Refrigeration cycle with residual gas

If amount and pressure of the impurities, such as methane, carbon monoxide or nitrogen, are large enough (> 3 – 5 vol %), the refrigeration requirements of the plant can be covered by *depressurizing the condensate* in a throttle (Joule – Thomson effect) (Fig. 63 A). Hydrogen is then recovered at high pressure. If the nitrogen or methane content is too low or variable, so that the refrigeration requirements cannot be covered reliably, then hydrogen itself can be expanded, e.g., by producing work in a turbine (Fig. 63 B).

If hydrogen, however, is required under high pressure, *external* refrigeration must be used. This can be done by a conventional external refrigeration plant using ammonia or freon as refrigerant (Fig. 63 C), or using other agents that are suitable for the required temperature level of condensation, in this case, e.g., nitrogen, carbon monoxide or methane (Fig. 63 D). The latter possibility requires the most sophisticated equipment. However, it leads to the greatest plant flexibility with regard to throughput and concentration of the impurities, because the refrigeration load is controlled directly by means of the external refrigeration cycle.

Some examples of processes using condensation alone are:

1. Refinery-gas purification plants, in which refrigeration is by a methane cycle or by decompressing the condensate (Fig. 63 A) with partial depressurization of the product [202].
2. Purification of the off-gas of the toluene dealkylation. Here a two-stage condensation is used for the recovery of pure hydrogen, a C_2, and a methane fraction. The refrigeration is supplied by an ammonia or freon cycle if necessary [203].
3. Recovery of hydrogen from the purge gas of ammonia plants. Here, a one-stage condensation process is sufficient. This process is particularly economic, when argon is recovered as an additional product by fractionation. A nitrogen cycle is used as refrigerant [204], [205].

Further common variations of the condensation process are as follows:

1. Staged decompression of the condensate and recycle of the decompressed gas to increase the hydrogen yield.
2. Partial condensation to fractionate the condensate and, eventually, to avoid the precipitation of solids. In this case the refrigeration can be supplied at a higher temperature, thus saving energy, by using refrigeration cycles which are optimized for the various condensation temperatures (see Fig. 63).

5.1.2. *Condensation and Rectification*

In large-scale plants higher yields and greater product purity are obtained by adding a rectification stage to the condensation step. This is also an advantage, if the impurities must be separated from one another. For the refrigeration many tailored variations exist.

Ethylene Plants. Pure hydrogen is recovered as a byproduct after low-temperature separation of ethane and ethylene from the methane – hydrogen fraction. The demethanizer plays an important part in the economics of the plant because the refrigeration required is at the lowest temperature of the process (⟶ Ethylene – Ethylene Fractionation.).

Coal Liquefaction. Hydrogen is recovered from the off-gas of the Fischer – Tropsch synthesis (operated on a large scale in South Africa by SASOL) by means of low-temperature condensation and rectification processes.

Figure 64 shows the LT separation of the purification unit of the SASOL II plant. In this unit 8675 mol/s (700 000 m^3 [STP]/h) crude gas is separated into 7 fractions, 5 of them in a LT plant. The hydrogen fraction is ca. 3700 mol/s (300 000 m^3 [STP]/h) of 80 % pure H_2 [206]. The crude gas, freed from carbon dioxide, is cooled down close to the hydrate point (i.e., the temperature at which precipitation of solid gas hydrates occurs) [207]. The water condensate is removed and separated into a water and a C_{3+} fraction. The gas phase is dried in an adsorber and cooled down to –35 °C. The condensate produced in this stage is fed to the rectification column g and separated into a C_2 and a C_{3+} fraction. In the following condensation stage a C_2 fraction is recovered by cooling down to ca. –95 °C, in the final stage at –150 °C liquid methane is produced. Methane, the C_2 fraction and the C_3 product are separated in two columns. The gaseous phase in the final knock-out drum e_5, is ca. 80 % pure hydrogen.

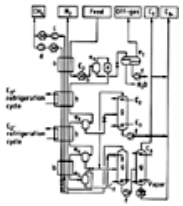

Figure 64. Production of H_2 and hydrocarbon fractions from Fischer – Tropsch synthesis off-gas by staged condensation. Separation of the hydrocarbon fractions by low-temperature rectification [206]

a) Adsorber for drying and elimination of traces of long-chain hydrocarbons; b) Heat exchanger; c) Compressor; d) CH_4 recycle compressor; e_1 – e_5) Knock-out drums; f) Pump; g) Column

This purity is suitable for the Fischer – Tropsch synthesis for which the main portion of the hydrogen-rich gas is required. A smaller part of the gas is used for hydrogenation purposes and is purified to 99.5 % using a pressure-swing adsorption process (see Section Pressure-Swing Adsorption (PSA)).

The refrigeration for the LT plant is supplied by an ethylene-propylene cascade and a methane cycle (see Figs. 63 C and 64, compressor d). The refrigeration must compensate for losses to the surroundings, in the heat exchanger and in the liquid product.

The refrigeration is supplied in 7 stages (including methane cycle) between + 9 °C and – 150 °C. The ethylene-propylene cycle has a power requirement of ca. 23 MW.

Low-temperature processes have been proposed for purification of the flash gas which is recovered on the decompression of the coal oil in plants for coal liquefaction by means of hydrogenation. Compared to other alternatives such as scrubbing at low temperature or diffusion separators [208], this concept has shown to be superior with regard to yield, energy requirements, economics and feasibility.

Hydrogen – Carbon Monoxide Recovery. Figure 65 shows a low-temperature process, which simultaneously delivers hydrogen and carbon monoxide of high purity from converted gases and gases from partial oxidation. After leaving a carbon dioxide scrubber, the converted and non-converted crude gas is (1) freed from traces of carbon dioxide and water in an adsorber, (2) cooled in two heat exchangers and in the sump of the carbon monoxide column down to the condensation temperature of carbon monoxide, i.e., ca. 80 K.

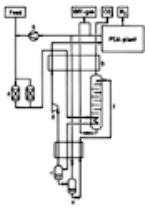

Figure 65. H_2 production from synthesis gas by low-temperature condensation and pressure-swing adsorption and production of pure carbon monoxide by rectification (German Linde)

a) Preadsorber; b) Heat exchanger; c) Condensation of CO and CH_4; d) Expansion turbine; e) Joule – Thomson expansion for H_2 recycle; f) CO/CH_4 rectification; g) Compression of expanded gas and PSA gas

Energy and refrigeration is produced by expanding hydrogen in a turbine, thereby expanding down to the inlet pressure of a PSA plant in which hydrogen of 99.9 % end purity can be recovered. The carbon monoxide content in the hydrogen is <0.5 ppm. Liquid carbon monoxide is decompressed stagewise, whereby the major part of the refrigeration requirements of the plant is covered. After warming, the carbon monoxide is fed into a rectification column and purified to about 99 %, the maximum hydrogen impurity thereby being 0.5 %. A plant corresponding to Fig. 65 for a gas throughput of 833 mol/s (67 250 m^3 [STP]/h) synthesis gas has been constructed in two trains. The purge gas of the PSA plant can be recycled to obtain a higher yield of hydrogen and carbon monoxide [209].

Hydrogen – Nitrogen Rectification. Whereas in low-temperature hydrogen purification processes the rectification is employed principally to purify the byproducts, large-scale industrial processes also exist in which hydrogen itself is directly involved in the rectification. In the Braun purifier (part of an Ammonia synthesis process, using steam reforming of light hydrocarbons) (⟶ Ammonia) [210] ammonia synthesis gas is produced by condensing excess nitrogen from the raw gas which contains hydrogen, carbon monoxide, methane and nitrogen. Figure 66 shows the rectification of the hydrogen-nitrogen mixture. The refrigeration is provided by the decompression under production of energy in turbine (b) and by the Joule – Thomson decompression of the remaining gas in a throttle. A similar

process for the production of ammonia synthesis gas from nitrogen rich blast-furnace gas or pit gases has been described [211].

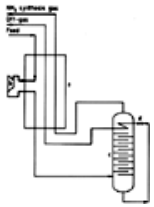

Figure 66. Rectification of an H_2/N_2 mixture and production of NH_3 synthesis gas (molar ratio $H_2 : N_2 = 3 : 1$) in the Braun purifier [210]

a) Heat exchanger; b) Expansion turbine; c) Rectification column; d) Joule – Thomson expansion

Ammonia Purge Gas Separation. A low temperature process for ammonia plant purge gas purification features especially high yields in ammonia recovery and in recovery of hydrogen as a product and for recycle. The application of this process is especially advantageous if the low temperatures occurring in the initial condensation step are utilized in further stages to recover other valuable components such as noble gases and methane [212], [213]. Figure 67 shows the integration of the low-temperature process into the ammonia plant. Several plants have been built for the recovery of liquid argon. If only hydrogen is to be recycled, adsorption is a competing process (see Section Adsorption Processes). Combinations of both processes have also been proposed [209], [214].

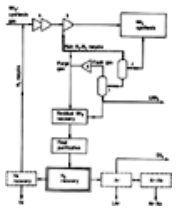

Figure 67. Integration of the low-temperature process into an ammonia plant [215]

a) Compressor; b) Recycle compressor; c) NH_3 condensate separator; d) Expansion turbine

5.1.3. *Low-Temperature Absorption Processes*

Absorption processes can yield greater purities than the condensation processes. Absorption processes are operated at higher temperature and thus with lower refrigeration requirements. The low-temperature processes described in this section are physical absorption processes, in which the impurities are condensed and either completely or partially utilized as scrubbing agent (for absorption processes with other scrubbing agents see Section Scrubbing Processes).

Nitrogen Wash. A high-purity hydrogen – nitrogen mixture for ammonia synthesis is obtained by using a liquid nitrogen scrubber (⟶ Ammonia) [216]. This scrubber can not only reduce the catalyst poison carbon monoxide down to a few ppm but the inert gases such as methane and argon can also be removed. The methanation step can thus be avoided.

For modern ammonia processes, in which the synthesis gas is produced at high pressure, the nitrogen scrubber is particularly interesting (see Section Gasification of Coal and Hydrocarbons) At low temperatures around 80 K the system hydrogen – nitrogen is still in the subcritical range, although pure nitrogen under these conditions is already supercritical. The necessary refrigeration for the nitrogen scrubber is produced by decompression of the nitrogen into the hydrogen stream.

Methane Scrubber. For the recovery of pure hydrogen from hydrogen – carbon monoxide mixtures (e.g., from synthesis gas) the use of a methane scrubber is frequently an alternative to the condensation of carbon monoxide. Methane scrubbers attain much higher hydrogen purities, especially with regard to the carbon monoxide content, than would be possible with the carbon monoxide condensation process. Simultaneously, pure carbon monoxide can be produced. Figure 68 shows the most important process stages of a methane scrubber [217].

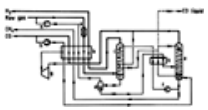

Figure 68. Plant for the recovery of hydrogen and carbon monoxide from H_2/CO-rich gases using a CH_4 scrubbing process

a) Adsorber for removal of H_2O and CO_2; b) Heat exchanger; c) Scrubbing column for CO with subcooled liquid methane; d) Knock-out drum; e) Separation of loaded wash methane to liquid lean methane (bottom) and pure carbon monoxide (top); f) Pump; g) Expansion turbine; h) Compressor

Crude gas between 1.5 and 4.5 MPa is dried in adsorber (a) and cooled down to ca. 90 K in a heat exchanger (b). The carbon monoxide remaining in the gas phase after condensation is removed down to a few ppm in the scrubber (c) with subcooled liquid methane. The liquid phase from the knock-out drum (d) and from the column bottom after decompression and partial evaporation is fed into the carbon monoxide – methane separation column (e). Here the sump product is pure methane, which is used for scrubbing, whereas a carbon monoxide fraction is obtained as the head product. A carbon monoxide cycle provides the heating and the reflux for the separation column as well as the cooling for the scrubbing column. By decompressing a fraction of the carbon monoxide cycle and generating energy the refrigeration for the heat exchanger is obtained.

In a similar fashion, however without a carbon monoxide – methane separation column, methane scrubbers are used for the purification of refinery off-gas or for the production of pure hydrogen from the purge gas of hydrogenation plants (coal hydrogenation, heavy oil hydrogenation). Similarly, as shown in Figure 68, methane scrubbers combined with adsorption processes can be used for the recovery of pure hydrogen, oxo-synthesis gas and a methane-rich heating gas fraction [213].

Other Scrubbing Agents. Absorption processes at low temperatures are not confined to the aforementioned scrubbing agents. Higher hydrocarbons such as ethylene or propane have also been used. Hydrogen itself can also be used as a scrubbing agent if the product desired is helium. Processes of this kind have been proposed and plants built for the recovery of helium and neon from air separation units and from ammonia purge gas streams [218]. In both cases the noble gases are already preconcentrated by several orders of magnitude compared to their natural abundance.

5.2. Adsorption Processes

5.2.1. *Principles*

Because of the only slight interaction of hydrogen with the usual adsorbing agents in technically important pressure and temperature ranges, adsorption processes are particularly suitable not only for the separation of hydrogen from other gases but also for the separation of impurity traces from hydrogen. For the principles of adsorption see ⟶ Adsorption [219-221].

The adsorbents for hydrogen purification are chosen according to the impurities to be removed. The most common adsorbents are given in Table 30. The gas mixtures are separated by various effects. *Steric effects* prevail if the adsorbent has a uniform and narrow micropore distribution, as is found particularly with the zeolites. *Equilibrium effects* result from different adsorption forces of the gases on the solids, and *kinetic effects* are caused by the different diffusion mobilities of the gases into the pores of the adsorbing agents.

Table 30. Adsorbents for the purification of raw hydrogen

Type	Main adsorption duty
Aluminum oxides	water
Silica gel	water, CO_2, C_{4+}
Activated carbon	CO_2, CH_4, C_{2+}, N_2
Molecular sieve (zeolite)	CH_4, CO, N_2
Carbon molecular sieve	O_2

Although each of these effects is present to a various degree in the adsorbent-gas combinations given in Table 30, the strength of the adsorption forces of the various impurities in the raw hydrogen mixtures can be used to put them in the order shown in Table 31. The sequence shown there may change slightly depending on adsorption agent, partial pressure of the impurity and total gas pressure of the system.

Table 31. Adsorption forces for various components in hydrogen purification

Component	Adsorption forces
Hydrogen	weak
Helium	↓
Oxygen	↓
Argon	↓
Nitrogen	↓
Carbon monoxide	↓
Methane	↓
Carbon dioxide	↓
Ethane	↓
Ethylene	↓
Propane	↓
Isobutane	↓
Propylene	↓
Hydrogen sulfide	↓
Mercaptans	↓
Benzene	↓
Toluene	↓
Ethylbenzene	↓
Styrene	↓
Water	strong

The static equilibrium loading capacity of the adsorbent for a particular component is depicted in the form of an *adsorption isotherm* (loading capacity depending on partial pressure at constant temperature). Adsorption isotherms have typical values for the adsorbents and the components in question and must be determined experimentally. For a number of typical components and some adsorbents these values have been compiled in handbooks and in specification sheets of the various producers.

Figure 69 shows the principal adsorption isotherms of various gases present during hydrogen purification on carbon molecular sieve (active charcoal obtained from coal). There is a characteristic affinity between each particular gas and the solid.

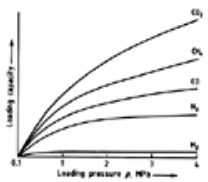

Figure 69. Principle form of equilibrium isotherms of mixtures of hydrogen with nitrogen, carbon monoxide, and methane [222]

Adsorbent material: carbon molecular sieves (activated carbon obtained from coal)

Adsorption is an exothermic process. Desorption, subsequently, is endothermic. In *thermally regenerated adsorbers* the energy necessary for the desorption is obtained by indirect heating of the bed or by using hot purge gas. In *pressure-swing adsorption* the desorption energy is obtained from the energy stored in the adsorption bed (thereby causing a temperature drop) if the pressure or the partial pressure of the components to be desorbed is decreased.

Hydrogen purification takes place in adsorption vessels with fixed beds of adsorbents. Other types, such as trickle beds, adsorption columns with solid circulation, moving beds, or adsorption in ebullating beds, are not used for hydrogen purification on a technical scale.

5.2.2. *Thermally Regenerated Adsorbers*

Thermally regenerated adsorbers are used to remove traces of impurities from hydrogen streams or for selective removal of certain substances which disturb ensuing process steps. The process is often designated as temperature-swing adsorption (TSA).

The adsorbers are usually switched alternately. While one adsorber is in the adsorption stage, the other is regenerated or is held in readiness. Switching usually takes place in a defined sequence, whereby the time for the adsorption step must be chosen shorter than the time calculated for the breakthrough of the impurities. Typical cycle times for these processes are in the range of some hours up to several days.

The process is illustrated in Figures 70 and 71 for the adsorption of carbon dioxide in a hydrogen-rich synthesis gas on a molecular sieve (5 A). As shown in Figure 71, the adsorber becomes loaded to give a local concentration distribution which is indicated by curve a. The maximum adsorbed amount of carbon dioxide in the zone of the equilibrium concentration EZ is given by the equilibrium isotherm for 25 °C at point a in Figure 70.

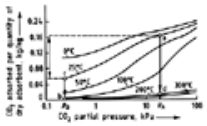

Figure 70. Adsorption and regeneration of CO_2 and molecular sieve 5 A, pellets

a) Adsorption loading; b) Desorption loading after decomposition and heating up; c) Theoretical desorption loading after decompression; d) Desorption loading after heating up without decompression

p_A = adsorption pressure; p_D = desorption pressure

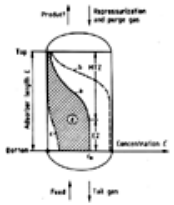

Figure 71. Concentration profiles in an adsorber (MTZ = mass transfer zone; EZ = equilibrium loading zone)

a) End of adsorption; b) End of cocurrent expansion; c) End of regeneration (residual loading); d) Useful cyclic capacity

In most cases regeneration is initiated by depressurizing. This procedure causes a desorption along the 25 °C isotherm to point c. However, because the depressurization is not isothermal, a cooling effect occurs, so that besides the kinetics of the desorption the adsorption conditions at the lower resulting temperature must also be considered. For a so-called cocurrent depressurization the distribution of the concentration in the adsorber is given by curve b of Fig. 71. The equilibrium along the isotherm is attained very slowly. The desorption energy requirements caused by the depressurization lead to a strong cooling of the adsorbent, which in the presence of water vapor can lead to formation of a liquid phase and thus to an irreversible damage of the adsorbent.

The main desorption takes place after the pressure has been lowered (point c in Fig. 70) mainly by increasing the temperature, in this case to 200 °C (point b or, if no previous depressurization took place, point d). The required heat is provided either indirectly by means of inserted heat exchanger coils or by direct heating with a purge gas at higher temperature. After regeneration, the adsorber is cooled to normal operating temperature and pressurized back to normal adsorption pressure.

Applications. Before low-temperature purification and separation can be carried out (see Section Low-Temperature Processes), gas components, which would at low temperatures form solid deposits in the heat exchangers and knock-out drums, must be removed. In synthesis gases from partial oxidation and steam reforming these are in particular water and carbon dioxide, whereas in refinery gases these are usually higher hydrocarbons. In dealkylation and coke oven gases benzene, toluene, and other aromatics must be removed, whereas in sulfur-containing raw gases hydrogen sulfide and carbonyl sulfide are the disturbing components.

If these impurities are present in higher concentrations, the major portion may first be removed by condensation (water and hydrocarbons) and/or scrubbing (CO_2, H_2S and COS). Final purification is then by TSA. A process description of the removal of water and carbon dioxide from a water-rich synthesis gas by TSA is given elsewhere (⟶ Gas Production – The “Classic” Method).

Further applications for TSA in hydrogen purification processes are:

1. Removal of water and ammonia from ammonia purge gases before low-temperature separation.
2. Purification of hydrogen from electrolysis plants. The main impurities are water and oxygen. The latter is converted by means of catalytic oxidation (see Section Catalytic Gas Purification) to water and then removed by condensation and TSA. Hydrogen from the chlor-alkali electrolysis contains also traces of mercury. This can be removed by TSA on a special molecular sieve [223].
3. Removal of various impurities from pure hydrogen before liquefaction. Here the adsorption is carried out at liquid nitrogen temperature. The regeneration is done with pure hydrogen from the recycle gas at ambient temperature (see Chap. Liquefaction of Hydrogen).

5.2.3. *Pressure-Swing Adsorption (PSA)*

Adsorption Mechanisms. Contrary to adsorption with thermal regeneration, pressure-swing adsorption operates at almost constant temperature using partial-pressure differences. However, the small temperature gradients occurring with ad- and desorption contribute considerably to the mechanism of the process. The components to be removed are adsorbed at high system pressure and high partial pressure and are desorbed at low system pressure in accordance with the appropriate equilibrium isotherms of the various impurities (see Fig. 69).

Figure 72 shows, for the example of a hydrogen – carbon monoxide mixture, how carbon monoxide, which is easily adsorbed, reaches the equilibrium loading, a, on the molecular sieve, while pure hydrogen exits the adsorber. When the adsorber capacity is exhausted, the adsorbent must be regenerated. The carbon monoxide loading on the absorbent is reduced by depressurization and hydrogen purging. The difference c is the *available working capacity* of the adsorbent. However, as in the case of TSA, the *actual working capacity* of the total adsorbent is influenced by the presence of mass transfer zones (MTZ) (see Fig. 71). The length of a MTZ is basically determined by (1) the concentration of the component to be removed in the raw gas, (2) by the required product purity, and (3) by kinetic effects.

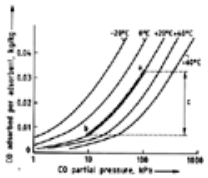

Figure 72. Adsorption isotherms for CO on molecular sieve

a) Adsorption; b) Desorption; c) Adsorption capacity

The PSA Process. Like the TSA process, the PSA process is cyclic, however, contrary to temperature swings, pressure changes can be carried out more rapidly and allow a shorter adsorption time. Typical cycle times are in the range of 3 to 10 min. Thus, removal of impurities present in high concentrations can be accomplished using moderately sized adsorber vessels. The essential steps carried out in a PSA unit are shown in Figure 73 [224]. Each adsorber is performing a different task, depending on which phase of the complete PSA cycle is regarded. The chronological sequence of all steps constitutes the cycle of one adsorber:

1. Adsorption at the pressure of the raw gas (highest system pressure), hydrogen production (adsorber A);
2. Cocurrent depressurization, provision of gas for pressure equalization and purging (adsorbers F and E);
3. Countercurrent depressurization, production of part of the tail gas (adsorber C);
4. Countercurrent purging (lowest system pressure), production of another part of the tail gas (adsorber D);
5. Countercurrent repressurization (adsorber B).

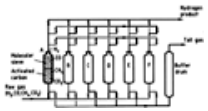

Figure 73. Six-bed PSA unit. Main flows for adsorption, repressurization, counterflow expansion (dump), counterflow purge, etc. for H_2 production by adsorber A. After termination of this adsorption step, next on-stream adsorber is B, etc.

Adsorber	A	B	C	D	E	F
Function	Adsorption	Repres-surization	Dump	Purge	Depres-surization	Depres-surization
Start pressure, (MPa)	2.8	1.6	0.4	0.2	0.8	2.8
Final pressure, (MPa)	2.8	2.2	0.2	0.2	0.4	2.2

Figure 74 [225] shows the different process stages with respect to the equilibrium loading on the adsorbent. The possibility of continuing the decompression stage into vacuum is also shown. This latter mode of operation is indicated for the enrichment and recovery of the adsorbed component. However, production of concentrated byproducts during the production of pure hydrogen is usually carried out by low-temperature processes and not by the adsorption process.

Figure 74. Theoretical pattern of the status of PSA adsorbers in respect to loading during one adsorption cycle [225]

a) Adsorption; b) Cocurrent depressurization; c) Countercurrent depressurization; d) Purge; e) Depressurization to vacuum; f) Repressurization; g) Equilibrium adsorption isothermal

p_A = adsorption pressure, p = purge gas pressure; p_v = final vacuum obtained

The mode of operation of a PSA can be summarized as follows:

1. Pure hydrogen is produced continuously at adsorption pressure.
2. Tail gas is produced continuously, the pressure, composition, and amount of which fluctuate periodically. These fluctuations are smoothed out by buffer vessels.
3. The efficiency of a PSA is defined by the amount of pure hydrogen produced in relation to the total hydrogen amount present in the crude gas. Depending on the crude gas composition and the degree of technical sophistication utilized, the PSA efficiency ranges between 70 and 90 %.

Technical PSA Plants. The simplest PSA system requires three adsorbers to enable continuous operation. Hydrogen yields of > 80 % can only be attained in systems with more than 4 adsorbers.

Since the size of the adsorber is usually limited because of manufacturing or transport considerations, larger quantities of crude gas can only be treated if several adsorbers are used in the adsorption

phase simultaneously (multi-bed process) [226], [227]. With such schemes PSA plants using 3 to 12 adsorbers can process amounts of crude gas ranging from 100 to more than 100 000 m^3 (STP)/h.

The maximum attainable hydrogen yield is influenced by a number of factors, the most important are:

1. *Composition of the crude gas.*
   For the economical operation of a plant the hydrogen content of the crude gas should be at least 50 %. The maximum hydrogen content is 95 – 99 %. Depending on the total amount of gas, TSA plants may be more suitable for such high-purity raw gases.
   The concentration and nature of impurities with respect to their ad- and desorption behavior determines basically the size of the adsorber which in turn results in higher or lower loss of hydrogen.
2. *Pressure ratio feed gas: purge gas.*
   This influences the size of the adsorber and the extent of the effort necessary for regeneration. At high pressure ratios the yields can be increased by using several pressure equalization steps (plants with more than 4 adsorbers).
3. *Purity requirements and nature of impurities.*
   The hydrogen purity in the PSA plants is usually >99.9 %. Higher purities with respect to those compounds which are poorly adsorbed, such as the inert gases nitrogen, argon, etc., are attainable at the expense of the yield. In spite of this, purities of 99.9999 %, i.e., hydrogen with <1 ppm impurities, can be realized on an industrial scale.
4. *Temperature of the feed gas.*
   High temperatures (>40 °C) result in a lower capacity of the plant, or in a poorer yield. Lower temperatures are therefore favored. Care must be taken, however, to avoid condensation of certain impurities in the regeneration step. For example, liquid or freezing water can damage the adsorbent.

Important developments in the design of large PSA plants were first possible after the introduction of microprocessor control systems. The main problems facing the control system are caused by the speed and the precision required by the short-time cycles. For this reason, modern plants are controlled by microprocessors using a 1-s sequence achieving “real time” control. Beyond this, it is now possible, not only to control the sequence and valves, but also to have a real process control system. This allows a high degree of flexibility for the changing of various process parameters and the indication of disturbances. In particular, the automatic initiation of disturbance strategies in the case of process, pressure and valve failures, e.g., shut-down of an upset adsorber without interrupting hydrogen production, is possible. Very high requirements are placed on the vessels and valves with up to 50 000 cycles (i.e., valve closures) per year. The operational practice has shown, however, that these plants are very reliable, so that they have become standard.

In the past years purification of hydrogen by means of PSA has increased more than any other gas purification process. Figure 75 gives an impression of the breakthrough of this new technology. As can be seen from the number of plants corresponding to the plant capacity, not only the number of plants per year put on stream has increased, but also the amount of hydrogen purified per plant. Plants > 100 000 m^3 (STP)/h inlet gas are not uncommon nowadays.

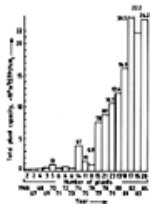

Figure 75. Worldwide capacity of H_2 gas purification plants with pressure-swing adsorption technology (put into operation from 1966 to 1984)

Applications [224].
Gas Purification in Steam-Reforming Plants (see Section Catalytic Reforming of Hydrocarbons). Other gas purification processes have been more or less completely replaced in this area [228]. For the production of hydrogen and carbon monoxide by steam reforming, PSA in combination with a low-temperature plant can also be economically used. The amount of gas handled by the plants built ranges from 500 – 87 000 m^3 (STP)/h of product gas. These plants use the tail gas produced to fire the reformer. Specially designed control loops and buffer systems lead to a very homogeneous tail-gas flow which is discharged with only small fluctuation in composition and flow, so that a reliable operation of the burners is possible.

Refinery Gases (see Section Refinery Processes). Because of the increasing demand for hydrogen for hydrogenation in refineries the PSA process is used there for hydrogen recovery and production. Refinery gases have a hydrogen content of 60 – 80 % the remainder being higher boiling hydrocarbons (C_6 – C_{10}). These are adsorbed well, but are difficult to desorb which can cause damage to the adsorbent. By choice of suitable adsorbents these difficulties can be avoided. Plants with feed gas capacities up to 100 000 m^3 (STP)/h have already been built. Figure 76 shows the hydrogen purification section in a large refinery in Kuwait.

Figure 76. Production of hydrogen from refinery off-gas. Plant with 12 adsorbers and 2 buffer drums (in the center) for flow and composition equalization

Plant capacity: 100 000 m^3 (STP)/h refinery gas; H_2 content: 80 %; product purity: >99 %; recovery rate: 89.5 %. Erected by German Linde in Kuwait in 1986

Coke Oven Gas (see Section Coke Oven Gas). Coke oven gas purification places the highest demands on the adsorption technology because of the inlet gas composition. Special preadsorbers have been used to remove impurities such as tar, naphthalene and higher sulfur compounds, which have a poor desorption capacity.

Hydrogen-Rich Gases from Various Synthesis Processes. Apart from numerous small plants which operate for example with ammonia or methanol plant purge gas, the coal gasification at SASOL is to be mentioned. 160 000 m^3 (STP)/h of off-gas from the coal gasification, respectively the Fischer – Tropsch synthesis is purified in five PSA plants.

5.3. Catalytic Gas Purification

Four distinct cases for the use of catalytic processes for gas purification can be distinguished:

1. A component of the crude gas is to be catalytically converted to hydrogen, i.e., a secondary hydrogen production process.
2. Impurities in the crude hydrogen are to be catalytically converted into a form which allows them to be more easily removed from the gas by other purification processes (condensation, adsorption).
3. Gas components, which even in traces could be detrimental to the use of the hydrogen, are converted into harmless components with only minor changes in the gas purity.
4. Adsorption processes, in which certain impurities are chemically bound to the adsorbent so strongly that a regeneration during normal operation does not take place.

5.3.1. *Carbon Monoxide Removal*

Hydrogen from synthesis gas may contain up to 50 % carbon monoxide (partial oxidation). Carbon monoxide can be recovered for CO production by a methane scrubber (see Section Low-Temperature Absorption Processes) and used for synthesis, or it can be reacted with water to produce hydrogen and carbon dioxide (water – gas reaction, shift conversion):

$$CO + H_2O \rightarrow H_2 + CO_2$$

The reaction is exothermal. Because of the equilibrium conditions the carbon monoxide conversion is incomplete in adiabatic reactors. For increased carbon monoxide conversion a greater amount of water ($c_{H2O}/c_{CO} > 3$) is necessary. At high carbon monoxide concentration a series of adiabatic reactors with intercooling is used. Isothermal reactors have also been suggested [229]. Depending on the type of catalyst used, three types of conversion may be identified.

Conversion of Sulfur-Containing Gas. Crude gas from partial oxidation or coal gasification, which contains at least 50 ppm hydrogen sulfide, can be converted on metals of group 6, 8, 9, or 10 of the periodic table (exceptions: Cr and Fe) or on mixtures of these metals. The carrier is aluminum oxide, sometimes alkaline compounds are present as a promoter. The metals are totally or partially sulfided, the sulfided form being the most active form. The usual catalysts contain cobalt, molybdenum, and nickel as active components. They can be used at 10 MPa and 250 – 550 °C.

An industrial process with these catalysts was described in Section Gasification of Coal and Hydrocarbons. A detailed description of the kinetics of crude-gas conversion is found in [230], patents in [231].

High-Temperature Shift (HT Conversion). For essentially sulfur-free gases, e.g., the product gas of the steam reformer or the secondary reformer, the well-known carrier-free conversion catalysts on iron oxide – chromium oxide basis can be used, e.g., 74 % Fe_2O_3, 10 % Cr_2O_3, 0.2 % MgO (rest: volatile components). Sulfur up to several hundred ppm can be tolerated although the activity then drops; catalyst poisons are phosphorus, silicon and unsaturated hydrocarbons in the presence of NO_x.

The shift reaction runs almost to equilibrium at space velocities between 3000 and 10 000 m^3 (STP) h^{-1} m^{-3} and pressures up to 7 MPa. The reaction temperatures are in the range 300 – 530 °C [232].

In steam reforming plants usually only one HT conversion stage is installed. To decrease the carbon monoxide concentration further, low-temperature conversion catalysts are used. If the purified synthesis gas in coal gasification plants is to be converted, a series of adiabatic reactors can be used. The last stage of this cascade usually consists of a low-temperature conversion.

Low-Temperature (LT) Conversion. The catalysts for LT conversion are supplied in their oxidized form, the major component of the unreduced catalyst is CuO, usually in a mixture with ZnO. Other components may be Cr_2O_3 and Fe_2O_3 (stabilizers) and Al_2O_3 (carrier) [230], [233]. The temperature range used is 190 – 260 °C; the lower temperature is limited by the dew point of the gas, the upper temperature because of the sensitivity of the copper catalyst to thermal deactivation.

These catalysts are extreme sensitive to sulfur, contents of 0.1 ppm leading to their deactivation, so that a suitably large reactor volume must be used to ensure that any slow deactivation of the catalyst will not cause any premature shut-down of the unit. Very often guard beds are placed before the actual catalyst [234]. For more details on the use of HT and LT shift conversion in steam reforming plants see Section Catalytic Reforming of Hydrocarbons and ⟶ Gas Production.

The thermal stability of copper catalysts can be raised considerably if they are used in a silica-supported form. These supported catalysts are prepared by precipitation from a homogeneous solution and are found to be very active for the water – gas shift reaction [235]. The catalyst can be reduced at a temperature of 450 °C and can be operated between 275 and 400 °C with initial carbon monoxide concentrations up to 10 %. A Fe_2O_3 – Cr_2O_3 catalyst has been prepared in a similar way and also shows improved activity at lower temperature, at the same time the resistance to traces of sulfur and chlorine compounds [236] still being preserved.

Because of considerations of equilibrium, reaction temperature and efficiency of the catalyst, the shift reaction can remove carbon monoxide only down to a remaining concentration of ca. 0.2 – 0.5 vol % in the dry gas. For refinery gases, however, purities <0.1 %, for hydrogenation of fats and oils <10 ppm (mL/m^3) and for ammonia synthesis <1 ppm (mL/m^3) are required (see Chap. Conventional Uses). For these and other purposes a further purification stage is necessary. In many cases this is the methanation step.

5.3.2. *Methanation*

Here, methanation is described as a process for the removal of traces of carbon monoxide or carbon dioxide. For the conversion of large carbon monoxide concentrations to yield substitute natural gas see ⟶ Gas Production – Methanation of Rich Gas.

Especially for ammonia plants the carbon monoxide concentration in the gas must be decreased (see also Fig. 77). This is done by using the reaction

$$CO + 3\,H_2 \rightarrow CH_4 + H_2O$$

The methanation converts carbon monoxide, which is a poison for the ammonia synthesis catalyst, to the inert gas methane. However, losses in hydrogen content (3 mol H_2 per mole of converted carbon monoxide) must be accepted. The design of methanation reactions is described [237].

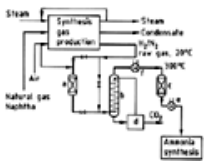

Figure 77. Removal of CO and CO_2 from ammonia synthesis gas ($3H_2 + N_2$) by selective oxidation (SELECTOXO), CO_2 scrubbing, and methanation

a) Selective oxidation; b) CO_2 scrubber; c) Methanation; d) Regeneration of the scrubbing agent; e) Gas cooling; f) Gas heating

Under typical methanation conditions (high hydrogen partial pressure, low methane and water content) the equilibrium is shifted far to the methane side, so the resulting carbon monoxide concentrations are usually below the detection limit. Nickel oxide, sometimes together with chromium oxide, is used as a catalyst. However, it must be reduced before the start-up. Aluminum oxide or diatomaceous earth is used as carrier. The metallic content of the catalyst is very high (30 – 60 %) compared to the nickel catalyst used for steam reforming.

The following undesired side reactions may occur:

$$CO_2 + 4\,H_2 \rightarrow CH_4 + 2\,H_2O$$

$$2\,CO \rightarrow C + CO_2$$

$$Ni + 4\,CO \rightarrow Ni(CO)_4$$

5.3.3. *Selective Oxidation*

The disadvantage of hydrogen losses and increased inert content in the aforementioned methanation process can be avoided, if carbon monoxide is oxidized to carbon dioxide, because the latter can be readily removed from hydrogen. However, as a side reaction oxidation of hydrogen, which is present in much greater concentration, may occur:

$$2\,CO + O_2 \rightarrow 2\,CO_2$$

$$2\,H_2 + O_2 \rightarrow 2\,H_2O$$

In spite of this, selective oxidation of carbon monoxide can be achieved on a special noble metal catalyst [238-240]. The incorporation of this process stage into the synthesis gas production for ammonia is shown in Figure 77.

The selective oxidation is placed behind the last shift reactor after removal of excess water by condensation. Air, which is taken from the compressed air for the secondary reformer, is used as oxidizing agent. Oxidation takes place at 50 – 70 °C in an adiabatic bed reactor. The temperature increase resulting from the heat of reaction amounts to 10 K per 0.1 mol % carbon monoxide. The carbon monoxide content can be reduced to ca. 100 ppm (mL/m^3) by this method. If the gas is to be used for ammonia synthesis, an additional methanation stage must still be incorporated for final clean-up.

5.3.4. *Catalytic Removal of Sulfur Compounds*

Sulfur compounds are catalyst poisons for hydrogen production and the various syntheses using hydrogen. Desulfurization is achieved in two stages: (1) hydrogenation and (2) hydrogen sulfide removal. Depending on the feed gas, these stages may be carried out separately or simultaneously. In the first step, organic sulfur compounds and carbonyl sulfide are hydrogenated to hydrogen sulfide

$$RCH_2SH + H_2 \rightarrow H_2S + RCH_3$$

$$COS + H_2 \rightarrow H_2S + CO$$

a cobalt – molybdenum contact is normally used as hydrogenation catalyst. If the gas contains a great amount of carbon monoxide, e.g., coke oven gases, the cobalt – molybdenum catalyst must be sulfided in order to avoid methanation of carbon monoxide (see Section Methanation).

Hydrogenation of sulfur compounds is favored by high hydrogen partial pressure and temperatures around 350 °C. After the hydrogenation of the sulfur compounds, the hydrogen sulfide formed can be removed by chemical or physical scrubbing (see Section Scrubbing Processes), if its concentration is > 1 %

For lower concentrations, hydrogen sulfide is absorbed on zinc oxide.

$$ZnO + H_2S \rightarrow ZnS + H_2O$$

Zinc oxide is available in the form of spheres, tablets or extrusions. The average sulfur pick-up until hydrogen sulfide breaks through the bed is ca. 25 % of the stoichiometric amount. The reaction temperatures are 100 – 350 °C (max. 400 °C). Organic sulfur compounds and carbonyl sulfide are partially cracked and absorbed in the zinc oxide bed, so that for natural gases with a low sulfur content the hydrogenation may not be necessary. The final purity after a zinc oxide bed is <1 ppm (mL/m^3) sulfur. The zinc oxide cannot be regenerated in situ but must be replaced.

Other materials used for desulfurization are, e.g., iron hydroxide (a byproduct of aluminum production), iron ore, and other purification agents containing iron oxide or iron hydroxide activated with alkaline compounds [241]. These absorption agents are mainly used for purification of coke oven gas before it is used as town gas. The achievable purity is <2 mg sulfur/m^3 (STP) (ca. 10 ppm [mL/m^3] S as H_2S). Iron- or copper-activated carbon can also be used. The regeneration of these agents with superheated steam is, however, only partially possible.

Chemisorption:

$$2\,Fe(OH)_3 + 3\,H_2S \rightarrow Fe_2S_3 + 6\,H_2O$$

Regeneration:

$$Fe_2S_3 + 3\,H_2O + 3/2\,O_2 \rightarrow 2\,Fe(OH)_3 + 3\,S$$

5.3.5. *Removal of Oxygen*

Hydrogen from coke oven gas and from an electrolysis (contrary to the hydrogen produced from other reducing processes) very often contains oxygen in concentrations up to 1 %. This oxygen is removed in *deoxo* reactors.

$$2\,H_2 + O_2 \rightarrow 2\,H_2O$$

Small levels of platinum and palladium on typical carriers such as α-Al_2O_3, active charcoal, or aluminosilicates provide highly active catalysts for the conversion (e.g., 0.3 % palladium on alumina). The catalyst is active even at ambient temperature, although usually higher temperatures are used.

Because of the strongly exothermic reaction a heat exchanger is installed after the deoxo reactor. The reaction product water is removed by adsorption. An example for the use of this process with prepurified coke oven gas is given in Figure 16.

Another method for oxygen removal is the use of catalytically active non-noble metals, usually copper or nickel, on carriers, e.g., silicon dioxide. These contacts operate at higher temperatures (300 °C) and are less sensitive to impurities such as hydrogen sulfide than are the noble metal catalysts. Both catalytic processes for oxygen removal are very selective, yielding purities of 1 ppm (mL/m^3) oxygen.

5.3.6. *Removal of Nitrogen Oxides*

In gases from coke ovens and coal gasification plants, in particular from entrained-bed gasifiers, traces of nitrogen oxides are present. These must be removed before gas separation in a cryogenic process is possible. Principally, nitrogen oxides can be removed from hydrogen streams by either oxidation or reduction. In the *reduction process* the nitrogen oxides are hydrogenated on sulfided cobalt – molybdenum catalysts, and water and nitrogen are produced. The reaction can be accelerated if small amounts of ammonia are present in the gas. Water and ammonia are removed in downstream scrubbers. The nitrogen produced can usually be neglected for the hydrogen purity.

Oxidative removal of NO takes place with ozone or, particularly selective, with ClO_2. The NO_2 produced is removed by use of scrubbers with alkaline-reducing agents (reduction to nitrogen). Final purities of ca. 0.01 ppm (mL/m^3) NO can be attained.

5.4. Scrubbing Processes

Scrubbing processes are used to purify large quantities of hydrogen. The compounds to be removed are absorbed in the scrubbing liquid. Since this can be done by utilizing either their different solubility in the scrubbing agent or by reaction with the scrubbing agent, the scrubbing processes are subdivided into physical and chemical processes.

5.4.1. *Principles*

For a detailed description of the physical and chemical principles of gas scrubbing see ⟶ Absorption and [242-244].

For selection, design and economics of a physical absorption process the most important factor is the solubility coefficient of a component given as λ (m^3 [STP] bar^{-1} t^{-1}) or λ' (kmol bar^{-1} t^{-1}) [200]. The capacity of an absorbent can be represented by the solubility curve as a function of pressure and temperature. These are shown in Figure 78 for carbon dioxide.

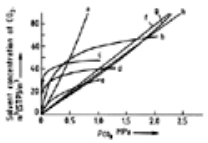

Figure 78. Absorption equilibria of carbon dioxide in various solvents [244]

a) Methanol, –15 °C; b) MDEA (methyl diethanolamine), 75 °C, 3.5 molar; c) MEA (monoethanolamine) 75 °C, 3 molar; d) Hot potassium carbonate, 110 °C; e) TEA (triethanolamine), 75 °C, 2.5 molar; f) NMP (*N*-methylpyrrolidone), 20 °C; g) Propylenecarbonate, 25 °C; h) Selexol (polyethyleneglycol dimethyl ethers)

For physical absorption the capacity at first follows Henry's law (i.e., a linear relationship, curves a, f, g, h), while for chemical absorbents typically a curve, which approaches a saturation value (curves b, c, d, e) is found. Apart from solely physical or solely chemical scrubbing, combined process can be used. Here the scrubbing stages may be separate and operated in series (e.g., crude and fine scrubbing) or a mixture of a physical and a chemical scrubbing agent may be used.

Chemical scrubbing is only suitable for acid gas components such as CO_2, H_2S, COS, and HCN. Physical scrubbing, on the other hand, can also be used for other components such as benzene, and higher hydrocarbons.

The raw gases for hydrogen production, depending on their source, contain several impurities in various concentrations; apart from that, they are at various pressures. For the choice of the optimum scrubbing process, the following factors are important:

1. Type, amount and partial pressure of the component to be removed;
2. Type, amount and total pressure of the gas to be purified;
3. Required final purity, utilization or subsequent process steps for the product gas and the impurities removed;
4. Consumption figures

Physical scrubbing is preferred for large gas quantities at high pressure with high concentrations of impurities to be removed; *chemical scrubbing* is particularly suitable when the absolute amount of the component to be removed is small. It can be used both at atmospheric and at higher pressure. With chemical scrubbing it must also be certain that there is no component in the raw gas, that can react irreversibly with the scrubbing agent. Carbonyl sulfide, for example reacts with MEA (monoethanolamine) to form stable salts. Here the absorbent cannot be regenerated by heat alone, and this leads to a loss of scrubbing agent.

Figure 79 shows how the choice of a suitable scrubbing process can be made for certain limiting parameters (e.g., partial pressure of the substance to be removed in the inlet and the exit gas). An overview of the processes used at present to remove acidic components from gases (synthesis and natural gas), is given in [246].

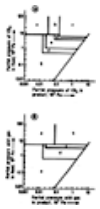

Figure 79. Examples for the selection of a suitable scrubbing process for CO_2 or CO_2 – H_2S removal with respect to partial pressures in feedgas and product gas, [245]

A) Selection of solvent for CO_2-removal, no H_2S; a) Physical solvent plus amine; b) Physical solvent, physical solvent plus amine or activated hot potassium carbonate; c) Physical solvents, d) Physical solvents or activated hot potassium carbonate; e) Activated hot potassium carbonate or concentrated amine; f) Activated hot potassium carbonate or amine; g) Amine

B) Selection of solvent for simultaneous removal of H_2S and CO_2; a) High loading amines (DEA), Selexol, Rectisol; b) Physical solvents; c) Physical solvents or diglycolamine; d) Activated hot potassium carbonate, Sulfinol, amines; e) Amines, Sulfinol

The most important processes used for purification in hydrogen production are:

1. Carbon dioxide removal from converted steam reformer gases for the production of hydrogen or ammonia synthesis gas. Here the hot potash scrubbers and modifications thereof, often combined with an amine wash, are used (see ⟶ Absorption). In recent times activated MDEA (methyl diethanolamine) wash processes and certain mixed amine processes have been found to have some

economic advantages, however, there still is a tendency to use physical scrubbing processes (Selexol, Sepasolv, etc.) for crude gas purification.

2. For the simultaneous and separate removal of carbon dioxide and hydrogen sulfide from converted or unconverted raw gas from partial oxidation of heavy oil or coal, the Rectisol process has been successful and is used almost exclusively for coal gasification plants.

5.4.2. *Physical Scrubbing (Rectisol Process)*

The Rectisol process (⟶ Gas Production) (jointly developed by Linde and Lurgi) is a physical scrubbing process with methanol as the preferred scrubbing agent. Newer variations use other organic solvents or solvent mixtures (e.g., Rectisol II, [248]). Since the absorption capability of organic solvents increases at lower temperatures and since, in addition, losses of scrubbing agent are smaller, Rectisol processes are usually operated at 200 – 260 K. The solubility of various gases in methanol is given in [200] and [248].

Depending on the process chosen, carbon dioxide and hydrogen sulfide can be recovered together or separately. Usually the simpler plants have a common scrubbing stage. Proper configuration of the main scrubbing and the purification section allows a very high attainable purity (20 – 50 ppm [mL/m^3] CO_2). For special purposes (e.g., methanol synthesis gas), a certain amount of CO_2 can be left in the gas and the fine scrubbing can be used only for the sulfur containing compounds. The disadvantage of the simple process is that a tail gas is obtained which, apart from carbon dioxide, also contains a certain amount of sulfur compounds and other traces (HCN, CO). Such a gas must, therefore, be treated by special processes.

For the production of a sulfur-free carbon dioxide fraction and a hydrogen sulfide fraction suitable for the Claus process, selective Rectisol scrubbing processes are used (⟶ Gas Production). Figure 80 shows a simplified process scheme with process description. Rectisol is very flexible with respect to the attainable purity of the product gas, the pressure range and any possible process variations. In addition, there are no corrosion problems. For ammonia synthesis gas the Rectisol process can be combined advantageously with a low-temperature liquid nitrogen wash (see Section Low-Temperature Processes). Other physical gas scrubbing processes are summarized in Table 32.

Table 32. Physical gas-scrubbing processes

Name	Solvent	Temperature, °C	Licensor
Fluor-Solvent	propylene carbonate	ambient	Fluor
Purisol	*N*-methyl pyrrolidone	ambient	Lurgi
Selexol	polyethyleneglycol dimethylether	0 – 40	Norton
Sepasolv	polyethyleneglycol diisopropylether	0 – 40	BASF

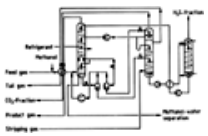

Figure 80. Selective Rectisol scrubbing process for production of a sulfur-free carbon dioxide fraction and a hydrogen sulfide fraction suitable as Claus plant feed [249]

5.4.3. *Chemical Scrubbing Processes*

An example of chemical scrubbing in hydrogen production is the combination Benfield – DEA scrubber (so-called High-Pure Benfield). For the SASOL II and III plants carbon dioxide is removed from the recycle gas of the Fischer –Tropsch synthesis, which in total amounts to 7×10^5 m^3 (STP). After the ensuing low-temperature separation, a gas with 80 % hydrogen is obtained which is fed either to the Synthol synthesis or to a PSA plant. A process scheme is shown in Figure 81.

Figure 81. Flowsheet of a combined Benfield – DEA scrubber (High Pure Benfield) for CO_2 removal from Fischer – Tropsch synthesis recycle gas (SASOL II and III) [250]

a) Absorption column; b) Regeneration column; c) Steam generator; d) Air cooler; e) Water cooler; f) DEA scrubber; g) Heat exchanger; h) Pump

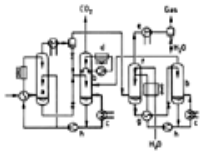

High-Pure Benfield Process. Both scrubbers contain an absorption and a regeneration tower, the latter operating at almost atmospheric pressure. The temperature ranges between 70 and 108 °C in (a) and between 50 and 60 °C in (f). The absorbents are regenerated by indirect heating with steam. The final cooling of the regenerated absorbents takes place in air coolers, the cooling of the purified gas in water coolers. The hot potash scrubber (first stage) lowers the carbon dioxide level below 0.8 vol %, the final DEA scrubber below 80 ppm (mL/m^3). The relatively small carbon dioxide amount from the DEA regeneration column (b_2) is fed to the hot potash regeneration column (b_1) and there, together with the main amount of carbon dioxide, vented into the atmosphere or recovered and used for other purposes.

Table 33 lists some other chemical scrubbers.

Table 33. Industrially important chemical scrubbing processes

Classification	Scrubbing agent	Additives	Trade name, licensor
Amine washes	MEA		
	MEA	not specified	Aminguard (UCC)
	DEA		
	HDEA		Linde
	MDEA		
	DGA		Fluor
	DIPA		Adip
	activated MDEA,	various	Shell BASF
	t-amines	various	UCARSOL (UCC)
Physical – chemical washes	sulfolane, DIPA		Sulfinol (Shell)
	methanol, amine		Amisol (Lurgi)
Hot potash washes	K_2CO_3– H_2O	vanadium, DEA	Benfield
		amineborate	Catacarb
		borax	Lurgi
		arsenate	Giammarco
Caustic scrubber	NaOH – H_2O	only used for fuel purification, no regeneration	

Oxidative scrubbing processes are sometimes used to remove hydrogen sulfide, e.g., for the purification of coke oven gas or other gases from hydrogen purification plants. Hydrogen sulfide is absorbed and oxidized in the scrubbing agent to form elemental sulfur (⟶ Gas Production). Such processes are:

Stretford process (British Gas Corp.)
Locat process (ARI)
Sulfolin process (Linde)

An overview of the literature on hydrogen sulfide removal from coke oven gas, low-Btu gas, rest gases and natural gas is given in [251], on newer oxidative scrubbing processes in [252]. The scrubbing processes described, however, are unable to remove the so-called “inert components” (noble gases, nitrogen, carbon monoxide, methane). In general, further purification steps on the basis of adsorption and catalysis (see Section Catalytic Gas Purification) as well as low-temperature processes including cryogenic scrubbing with liquid methane (see Section Low-Temperature Processes) are necessary.

5.5. Permeation Processes

For purification of hydrogen and for separation of hydrogen from gas mixtures processes based on the principle of diffusion and permeation through membranes are also suitable. The physical fundamentals of diffusion processes are given in [253] together with a detailed description of diffusion separation methods in [254].

The driving force for gas transport through membranes is the partial pressure of the gases. The permeating flux Q_i (per unit area of surface and time) is given by Equation (57):

$$Q_i = \frac{P_i}{d} \cdot (p_{i2} - p_{i1}) \tag{57}$$

where P_i, the permeation coefficient, is the product of solubility and diffusion coefficients of a gas *i*, p_{i2} and p_{i1} are partial pressures on the high- and the low-pressure sides of the membrane and *d* is the membrane thickness.

To separate various components it is essential to have different permeation coefficients. The separation capability of the membrane is defined by the selectivity α; this is the ratio of the permeation coefficients of the two gases *i* and *j*

$$\alpha = P_i / P_j$$

The fact that the permeation coefficient of hydrogen is much higher than that of other gases is utilized to separate hydrogen by diffusion through metal or polymer membranes.

5.5.1. *Metal Membrane Techniques*

Hydrogen diffuses well and with a high specific permeation rate through membranes consisting of palladium, palladium – silver or palladium – rare earths. Silver – palladium alloys with 23 wt % silver are most suitable. The diffusion rate of hydrogen is highest at a composition close to the maximum solubility of silver in palladium (i.e., ca. 23 wt %); furthermore, material distortion by α – β phase transition in the palladium-hydrogen system is substantially decreased. Optimum operating conditions, i.e., maximum hydrogen throughput and high durability are found at 350 °C and 2 MPa hydrogen pressure. Unsaturated hydrocarbons, halogens and sulfur compounds can lead to rupture of the membrane. The hydrogen suffers a great pressure drop after permeation through the membrane, however, it is particularly pure (impurities and moisture <0.5 ppm [mL/m^3]). Palladium membrane diffusion is discussed in [255].

Portable small-sized hydrogen generators or containerized medium-sized units (the latter are constructed from modular units) can be used for the final purification of hydrogen or for separation from raw gas mixtures, e.g., from methanol or ammonia reforming (see Section Other Chemical Processes). Main parts of a module are the diffusion elements which consist of a number of silver – palladium tubes contained within a cylindrical cell.

Figure 82 shows the cell performance data of commercial pure-hydrogen plants on the basis of diffusion. The raw hydrogen must be fed to the unit at 2.1 MPa, the operating temperature is 300 °C and the average specific energy consumption amounts to ca. 100 W per cubic meter of hydrogen [256]. The process is considered suitable for small- to medium-scale plants. Metal membrane processes have not been adopted for industrial-scale hydrogen purification.

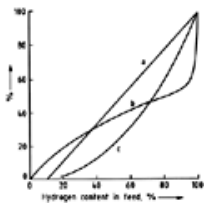

Figure 82. Operating data of hydrogen diffusion cells with Pd-Ag membranes

a) Yields (%, related to input hydrogen amount); b) Hydrogen content in off-gas (vol %); c) Cell output) (% of nominal capacity)

5.5.2. *Polymer Membrane Processes*

New developments in the area of polymer membrane materials also allow their utilization in gas separation technology. The main criteria for suitable separation membranes are high permeation rates together with high selectivity, pressure-resistant membrane structures and chemical as well as thermal stability. State-of-the-art membranes have asymmetric composite layer structures on a woven material support or a composite hollow-fiber membrane structure. The membrane separation stage is built as module; for gas separation module constructions with high packing densities are used, namely spiral-wound modules (> 900 m^2/m^3) and hollowfiber modules (9000 m^2/m^3). An overview is given in [257][258][259]. From the process point of view various module arrangements such as cascades or membrane columns [260] may be used.

Asymmetric composite membranes based on cellulose acetate or polysulfones have a particularly high specific permeability for hydrogen, as do aromatic polyimide membranes; they are used for the industrial scale separation of hydrogen.

The product of Permea, a Monsanto company, is leading in this field. The Perma Prism process [261] uses asymmetrical hollow-fiber membranes of polysulfone with an active siloxane layer (300 – 500 µm outer diameter, 100 – 250 µm inner diameter, effective siloxane separating layer 0.1 µm). Several 10 000 of hollow fibers are connected together in a module, as shown in Figure 83. The fiber bundle is sealed at one end (top) and enclosed in an epoxide mantle at the other end. The modules are installed vertically and have a diameter of 100 or 200 mm and a maximum length of 3 or 6 m. The feed gas, which should be freed of disturbing components (NH_3, H_2S, CH_3OH) by a suitable scrubbing system, is fed into the module near the base. The permeate is removed from the base of the module while the remaining gas is removed from the top. The fibers can be operated with a pressure of up to 15 MPa, whereby the pressure difference may amount to 7 MPa. The operating temperature can be up to 100 °C.

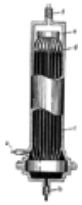

Figure 83. The Prism separator module of Permea

a) Feed stream of mixed gases; b) Permeate gas outlet; c) ASME-coded carbon steel shell; d) Hollow fiber membranes; e) Fiber bundle plug; f) Nonpermeate gas outlet

The stability of the membranes towards impurities is given in % of the corresponding saturation value at 20 °C:

Hydrogen sulfide	up to 10 % of the saturation value
Ammonia	up to 2 % of the saturation value
Water	up to 100 % of the saturation value
Hydrocarbons (C_2 – C_6)	up to 80 % of the saturation value

The degree of purity and the hydrogen recovery rate are determined by the partial pressure difference between the inside and the outside of the hollow fiber, the remaining gas components (i.e., their permeation coefficient) and their partial pressure.

Figure 84 A shows the rate of hydrogen recovery and the obtainable purity as a function of various hydrogen contents in the raw gas. Plots for two feeds (70 % and 40 % H_2) in Figure 84 B show that the ratio of total pressures dramatically affects permeate purity, whereas the effect of further increase in membrane selectivity leads to a plateau above a selectivity factor of 40.

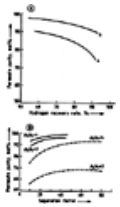

Figure 84. A) Obtainable hydrogen purity vs. hydrogen recovery rate

a) Partial pressure difference 3.52 MPa, H_2 in feed 63 vol %; b) Partial pressure difference 3.52 MPa, H_2 in feed 86 vol %B) Permeate purity with two feeds at different pressure ratios [257] ------- 40 % H_2, 60 % C_1; —— 70 % H_2, 30 % C_1

p_1 = permeate pressure, p_2 = feed gas pressure

Hollow-fiber separators were adopted quickly for commercial applications, more than 160 membrane separation systems have been installed for hydrogen recovery since 1979. Preferred applications are hydrogen recovery from purge gases of ammonia or methanol synthesis loops or from other hydrogen containing gases in the refinery or gas industries. Table 34 gives a survey. Figure 85 shows the

process scheme of a plant for hydrogen recovery from the purge gas of a 545 t/d ammonia plant. 3000 m^3 (STP)/h gas is led through two rows of modules (containing 8 and 4 modules, respectively) at various pressure levels, 90 % of the hydrogen thereby being recovered. This system is especially useful, because a part of the hydrogen can be led back to the synthesis stage at 7 MPa. In a 1000 t/d ammonia plant the recovered hydrogen leads to an additional ammonia production of 40 t/d (ca. 5 %) [262]. However, other valuable components of the gas, such as argon, cannot be recovered in a pure form.

Table 34. Hydrogen containing gases membrane modules can handle

Source of feedgas	Components		Pressure, MPa	Product composition, % H_2	Pretreatment, remarks
	% H_2	others			
Ammonia synthesis purge gas	60	N_2, CH_4, Ar, NH_3	13.0	H_2, 85	water wash, NH_3– conc. <200 ppm (mL/m^3)
Methanol synthesis purge gas	58	CO, CO_2, N_2, CH_4, CH_3OH, H_2O	5.0	H_2, CO_2, 80	water wash, CH_3OH conc. <100 ppm (mL/m^3)
Catalytic reformer off-gas	70 – 80	CH_4, alkanes, aromatics	3.0 – 5.0	H_2, 95	oil separation by refrigeration or oil scrubber, C_2– C_6 <80 % satur. aromatics <50 ppm (mL/m^3)
Catalytic cracker purge gas	10 – 30	N_2, O_2, CH_4, CO, CO_2, H_2S, alkenes, higher and lower olefins, aromatics	0.7 – 1.4	H_2, 60 – 80	compression, water wash
Hydrocracker purge gas	75	CH_4, H_2S, H_2O, NH_4, alkenes, aromatics	7.0 – 11.0	H_2, 95	aromatics <50 ppm (mL/m^3)
Hydrotreater purge gas	60	CH_4, H_2O, alkenes, aromatics	35 – 50	H_2, 97	feed <30 mol % H_2, uneconomic, aromatics <50 ppm (mL/m^3)
Toluene hydrodealkylation		CH_4, H_2O, alkenes, aromatics	30	H_2	oil separation by refrigeration or scrubber, aromatics <50 ppm, (mL/m^3)
Steam – methane reformer gas		CH_4, CO, CO_2, H_2O	30 – 80	H_2, H_2/CO_2	
Oxo-synthesis gas	70	CO	variable	H_2/CO 1 : 1	H_2/CO adjustment

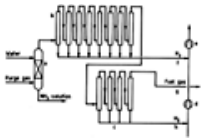

Figure 85. Flowsheet of a hydrogen recovery unit with Prism separator modules in an ammonia plant

a) Water scrubber; b) First bank Prism separators; c) Second bank Prism separators; d) Syngas compressor first stage (2.4 MPa suction pressure); e) Syngas compressor second stage (6.8 MPa suction pressure); f) H_2 recycle to second stage suction syngas compressor; g) Fuel gas to NO_x abatement or primary reformer; h) H_2 recycle to first stage suction syngas compressor

[Top of Page]

6. Liquefaction of Hydrogen

The boiling point of hydrogen is 20.39 K at 101.3 kPa (see Table 2) for n-H_2 and 20.26 K for p-H_2. For more details on the physical properties of n-H_2 and p-H_2 see Chapter Properties.

Like other gases, hydrogen is liquefied because storage and transportation is easier in the condensed state. Further applications of liquid hydrogen are rocket propulsion and scientific experiments, e.g., bubble chambers which are used to produce visible trails of various particles involved in reactions of nuclei and elementary particles. A wide application for liquid hydrogen is predicted in the future hydrogen energy concept (see Chap. Hydrogen Energy).

6.1. Principles of Hydrogen Liquefaction

For the liquefication of gases [263] energy must be removed until the condensation point is reached. This is achieved either by cooling down to low temperatures with external means, such as liquid nitrogen or the magneto-caloric effect, or internally by irreversible or isentropic decompression.

The decompression of hydrogen in a *throttle valve* can only lead to a temperature decrease in a certain range of the Joule – Thomson curve (see Fig. 5). The range which is used technically is from 77 K, the boiling point of liquid nitrogen at 0.1 MPa, down to 25 K. In this range (see Fig. 86) the isenthalpic lines (H = constant) in the T – S diagram are particularly steep. The decompression in the throttle leads to a decrease in temperature or, at the boiling point, to a high degree of liquefaction at a constant temperature. The whole range of the isenthalpic lines, isobars and isochores in the T – S diagram, especially in the two-phase areas and around the triple point is given in [264].

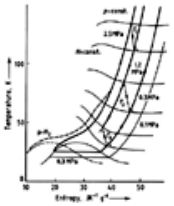

Figure 86. Temperature entropy diagram of n-hydrogen. Illustrated are the liquefaction process and the cooling cycle for a modern liquefaction process, described in Figure 88

Isentropic decompression in turbines or other machines is, generally speaking, possible in all temperature ranges. However, control of the two-phase streams in the neighborhood of the condensation point leads to technical problems. Apart from this, the efficiency of a decompression machine decreases for technical reasons at low temperatures. Therefore, this process is only used for hydrogen above 25 K.

Applying a *magnetic field* to a magnetic substance causes an alteration in the inner structure of the material which in turn leads to a temperature change. In particular, demagnetization of certain metals in the neighborhood of the Curie temperature leads to a maximum temperature decrease. (The Curie temperature is the temperature above which ferro-magnetic material reverts to the paramagnetic state.) Under favorable conditions an isentropic temperature decrease of 1 K/T flux density may be attained.

To obtain greater temperature differences respectively entropy changes, metals or alloys with different Curie temperatures of the components are necessary. For the liquefaction of hydrogen the rare earths (in particular gadolinium), their compounds and alloys are suitable. For a detailed description of the magneto-caloric effect in cryogenic refrigeration processes see [265] and [266].

Hydrogen liquefaction is technically achieved by a series of cyclic processes. The theoretical lowest energy consumption (minimum work) in a reversible Carnot process for the production of liquid p-H_2 (0.1 MPa) from gaseous n – H_2 (25 °C, 0.1 MPa) amounts to 14.23 kJ/g = 3.95 kWh/kg.

In natural hydrogen (n-H_2) at 25 °C the ratio of the nuclear spin isomers o-H_2 and p-H_2 is ca. 0.75 : 0.25. At the boiling point of hydrogen the equilibrium lies almost completely on the side of p-H_2 (see Section Physical Properties). The reaction enthalpy ΔH of the exothermal o – p conversion is in the same range as the heat of evaporation

$$\Delta H \text{ o – p (25 K)} = 1.448 \text{ kJ/mol}$$
$$\Delta H \text{ n – p (25 K)} = 1.086 \text{ kJ/mol}$$
$$\Delta H_{evap} \text{ for p-H}_2 \text{ (25 K)} = 0.811 \text{ kJ/mol}$$
$$\Delta H_{evap} \text{ for n-H}_2 \text{ (25 K)} = 0.828 \text{ kJ/mol}$$

Thus, if the reaction rate of the o – p-conversion is large in comparison to the retention time in storage and transport vessels, a major part of stored liquid n-hydrogen is evaporated even after a few days

[267]. In addition, heat losses through the insulation occur. For this reason, technical hydrogen liquefaction units have o – p conversion reactors incorporated into the process. The o – p conversion takes place on iron oxide gel, iron hydroxide and chromium oxide on alumina catalysts.

The heat of reaction of the o – p conversion must be compensated for by additional refrigeration capacity of the liquefaction unit. In the ideal case the reaction should take place along the equilibrium line (see Fig. 1), because then the lowest reaction enthalpy is involved and the necessary refrigeration is thermodynamically favorable, because it can be delivered at the highest temperature level. In technical plants this ideal case can be simulated by using a cascade of adiabatic or isothermal reactors.

Particular requirements are placed on the purity of the hydrogen to be liquefied, because at liquid hydrogen temperatures all impurities except helium become solid and this would lead to a blockage of the refrigeration unit. Before liquefaction hydrogen of > 99.5 % purity from electrolysis or of 99.99 % purity from a PSA plant is subjected to the following purification steps:

1. Oxygen removal by catalytic conversion in an adiabatic reactor at ambient inlet temperature (electrolysis);
2. Adsorptive drying (electrolysis);
3. Adsorptive removal of remaining impurities nitrogen, methane, argon, carbon monoxide near their condensation temperatures, technically usually at LN_2 temperature, (the adsorbers are regenerated with a fraction of heated hydrogen product) (electrolysis and PSA)

The various purification processes are described in more detail in Chapter Purification of Hydrogen.

6.2. Liquefaction Plants

Up to now, there are only a few hydrogen liquefaction plants operational in the world. The largest, with capacities up to 60 t/d, are in the United States and distribute liquid hydrogen throughout the whole country for general industry purposes and, in particular, as a rocket fuel to the space industry.

Some smaller plants are in operation in Europe, serving also space-transport related purposes, such as in Southern France, or scientific purposes as in the Federal Republic of Germany. The latest plant with a capacity of 5 t/d was erected in the Netherlands in 1988. For the first time European gas customers are to be delivered with liquid hydrogen from this plant. Very large plants have been described. Examples are those envisioned for the hydrogen energy concept with 10 modules, each of a capacity of 250 t/d [268].

Small plants use the Linde – Hampson process for liquefaction of hydrogen [269], [270]. The refrigeration is produced by expansion through a throttle at low temperatures. Precooling with liquid nitrogen (LN_2) at reduced pressure (*bp* 65 K) and relatively high hydrogen pressure (e.g., 12 MPa) is necessary. If liquid hydrogen is used directly, without storage, e.g., for bubble chambers, the o – p conversion becomes unnecessary. The liquefaction unit becomes simpler, although the specific energy consumption of 100 kJ/g (including the LN_2 consumption) still remains high.

The process can also be conducted in two steps. Thereby, the hydrogen is expanded to an intermediate pressure after LN_2 cooling and part is recycled to the second stage of the compressor. In Figure 87 a two-stage Linde-Hampson process is shown.

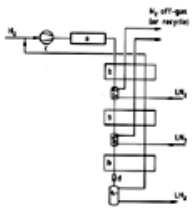

Figure 87. Hydrogen liquefaction for a two-stage Linde – Hampson process

a) Prepurification, adsorption; b) Heat exchanger; c) Hydrogen compressor; d) Expansion valve

Large-scale plants for hydrogen liquefaction not only utilize Joule – Thomson decompression for the production of the necessary refrigeration but also the decompression of hydrogen in turbines.

Figure 88 shows the process scheme of a hydrogen liquefaction unit without nitrogen precooling [271]. Purified hydrogen is cooled down to ca. 85 K in three heat exchangers against *decompressed* recycle hydrogen. A parallel stream of *pressurized* recycle hydrogen is also cooled down and, in the first turbine, provides part of the necessary refrigeration. In two further decompression stages the refrigeration necessary to cool down hydrogen further and to compensate for the o-p conversion is produced. The liquefaction takes place to ca. 85 % by means of Joule – Thomson decompression of the product hydrogen before entering the last reactor heat exchanger. Completion of the liquefaction and subcooling is accomplished by heat exchange with a small fraction of the recycle hydrogen cooled by Joule – Thomson decompression to lower pressures. The recycle hydrogen quantity is ca. 12 times that of the liquefied hydrogen.

Figure 88. Process flow diagram of a commercial H_2 liquefier without LN_2 precooling (German Linde) [271]

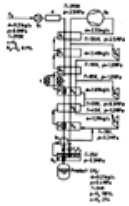

a) Adsorptive and catalytic precleaning; b_1 – b_8) Heat exchangers; c) Adsorbers at LN_2 temperatures; d) o-p conversion reactor; e_1 – e_3) Expansion turbines; f_1 – f_2) Expansion values; g_1) Compressor; g_2) Recycle compressor for H_2 coolant cycle

The *T* – *S* diagram (see Fig. 86) shows the thermodynamic cycle of this hydrogen refrigeration process. In particular, it illustrates the enthalpy differences obtained in the expansion turbines in three temperature ranges. The efficiency and losses during the liquefaction of hydrogen are shown in the energy flux diagram of Figure 89. The specific energy consumption amounts to 60.6 kJ per gram of liquid p-H_2.

Figure 89. Sankey diagram of a commercial H_2 liquefier, revealing the losses of the process (Process according to Figure 88)

b_1 – b_8) Heat exchangers; e_1 – e_3) Expansion turbines

Utilizing the energy produced in the turbines (ϒ = 0.8) a thermodynamic efficiency of 23.7 %, corresponding to a specific consumption of 56.3 kJ per gram of p-H_2 (98 % pure, rest o-H_2), is obtained. The power consumption of the hydrogen recycle compressor can be reduced considerably, if the cooling down to the temperature of liquid nitrogen is carried out with imported LN_2, by a nitrogen refrigeration cascade or by a multi-stage cascade with various refrigerants, e.g., propene, ethylene, methane and nitrogen (see Section Production from Coal and Hydrocarbons). The attainable saving in power must be balanced against the higher cost of the equipment and utilities (LN_2).

In large-scale plants [272] for LH_2 production, the specific energy consumption can be reduced by the following means:

1. Use of LN_2 for precooling,
2. Increased efficiency of compressors and turbines,
3. Decreased insulation losses, and
4. Decreased losses in heat exchangers and reactors.

6.3. Further Developments

Besides the Linde – Hampson process (refrigeration by isenthalpic decompression) and the Claude process (refrigeration by isentropic decompression) a number of other processes, such as the Stirling process, have been suggested for the liquefaction of hydrogen. These have been partially tested in small pilot-scale plants.

In pilot plants for low temperatures between 1 and 4 K, magnetic refrigeration has already been experimentally verified [266], [273], [274]. *Magneto-caloric refrigeration* is achieved in a similar way to thermodynamic processes, such as the isentropic expansion of gas. In the latter process the gas to be liquefied is used as a working medium, whereas in the magneto-caloric process the working medium is a ferromagnetic material (e.g., in the form of a bed through which the hydrogen is flowing).

Carnot, Stirling, Ericson and the Brayton process, which operates between two isentropes and two curves of constant magnetic field strength, can be used for the technical realization of the magneto-caloric process. The strong magnetic fields required (≥10 T) can be produced with the help of superconducting magnets.

The discovery of superconductors with a transition temperature around the boiling point of liquid nitrogen [275] is expected to lead to new developments in this area. Figure 90 shows the concept of a magneto-caloric liquefier operating with helium as a cooling fluid. In the four regenerators (d_1 – d_4) helium is cooled by the magneto-caloric effect. After each stage a partial stream is drawn off for indirect cooling of hydrogen. For the regenerators a common magnetic field coming from a superconducting magnet (8 – 10 T) is provided. The magnet here is still cooled with liquid helium, which must be produced in an additional stage. The projected technical data of such a hydrogen liquefaction unit are listed below:

Duration of magnetocaloric cycle, s	ca. 10
Magnetic flux density, T	ca. 10
Cooling power (at 20.4 K), W	20
Liquefaction rate, L/h	0.7
Liquefaction work, kWh/kg	7.26 (54 % of Carnot)
Power consumption, W	557
Heat losses (at 300 K), W	621
Volume, L	ca. 76
Mass, kg	ca. 166

The low power consumption and the efficiency of 54 % of the Carnot cycle is remarkable [265], [270].

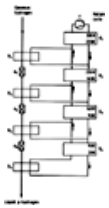

Figure 90. Principle flow diagram of a refrigerator using a magneto-caloric refrigeration cycle. The magnetic regenerators work in a cyclic mode

a) o-p Conversion reactors; b) Heat exchangers; c) Helium cycle compressor; d) Magneto-caloric regenerators

Hydrogen Slush. Liquid – solid hydrogen mixtures (slush) have an increased heat absorption capacity and a higher density than normal boiling liquids. This is an advantageous characteristic in space application, where longer storage times and higher storage densities are important.

Hydrogen slush is produced by a freeze – thaw method, where vapor is continuously withdrawn from a stirred mixture of liquid and solid hydrogen, liquefied and reexpanded into the storage vessel. The latent heat of vaporization is used to freeze the hydrogen. The mixture is then at its triple point (13.8 K, p-H_2, 7 kPa) and has to be protected against air leakage into the system by keeping it under a helium atmosphere.

Solid hydrogen can also be produced by freezing liquid hydrogen at higher pressure on a surface cooled by liquid or gaseous helium. In this case the higher pressure can only be maintained if a temperature stratification can be generated at the surface of the liquid phase [276]. For properties of hydrogen slush see [277].

[Top of Page]

7. Handling of Hydrogen

7.1. Quality Specifications

The hydrogen content of technical hydrogen produced from hydrocarbons is between 97 and 99.5 vol %. The impurities are mainly methane and nitrogen; traces of oxygen, carbon monoxide, carbon dioxide and, depending on the production and purification stages, hydrogen sulfide or ammonia may be present. Hydrogen produced electrolytically is typically >99.5 vol % pure. Impurities in this case are mainly nitrogen, oxygen and water. The mercury content must be limited to max. 0.15×10^{-3} mg/m^3 (for protection of the aluminum material in the downstream purification units [278]).

Hydrogen is available in high and very high purities. The impurities are removed by means of catalytic combustion (oxygen), drying systems (water) as well as by adsorption and diffusion systems. Hydrogen of the utmost purity is needed, for example, in the semiconductor industry. It is purified by diffusion through palladium silver membranes or, more recently, by adsorption and purification on hydride material (purity 7.0, i.e., total impurities <0.1 ppm [mL/m^3]) [279]. Table 35 gives an overview of typical hydrogen purities commercially available.

Table 35. Commercial hydrogen qualities

	Hydrogen “pure”	Liquid

	2.5	3.0	3.8	5.0	5.3	5.6	6.0	hydrogen 6.0
Purity, vol %	**99.5**	**99.9**	**99.98**	**99.999**	**99.9993**	**99.9996**	**99.9999**	**99.9999**
Impurities, ppm (mL/m^3)								
O_2	*	≤ 50	≤ 10	≤2	≤2	≤1	**	≤0.2
N_2		≤500	≤200	≤3	≤3	≤2		≤0.2
H_2O		≤100	≤ 20	≤5	≤2	≤1		≤0.5
CO								≤0.1
CO_2								≤0.1
C_nH_m				≤0.5	≤0.5	≤0.1		≤0.1

* Remainder N_2, O_2 and H_2O.

** Total amount of impurities ≤1 ppm (mL/m^3) thereof O_2 ≤0.5 ppm (mL/m^3) and C_nH_m ≤0.1 ppm (mL/m^3).

7.2. Compression and Expansion

Compression. Hydrogen is often needed at high and very high pressures (e.g., for hydrogenations: 5 – 70 MPa, for ammonia synthesis: 8 – 30 MPa). Because of its low molecular mass, however, it is difficult to compress in conventional machines. Special types of compressors have therefore been developed.

For very large amounts of hydrogen, (>50 000 m^3 [STP]), such as are necessary, for example, in the ammonia synthesis, *radial-type* high-pressure *turbo compressors* have become standard. Because the pressure ratio attainable with the hydrogen – nitrogen mixture is relatively low, multistage compressors are used. Leakage problems can be handled by suitable constructions. Extra stages of the synthesis gas compressor are used to recycle hydrogen. In smallercapacity plants (<500 t/d of ammonia) the synthesis gas is compressed in so-called mole pumps. Compressor and motor are thereby enclosed in a steel mantle.

For smaller quantities of gas and high end pressures *piston compressors* are chosen. Depending on the purity requirements placed on the compressed hydrogen three types may be used:

1. *Oil-lubricated* piston compressors for the compression of hydrogen for hydrogenation of oil products.
2. If the small quantities of oil in the hydrogen stream after the oil separator or adsorber cannot be tolerated, *water-lubricated* piston compressors can be used.
3. For the compression of high-purity hydrogen *dry-running* piston compressors are used. The pistons are sealed with specially pivoted and guided PTFE (polytetrafluoroethylene) piston rings or, to avoid friction losses, with labyrinth seals. The leakage gas is fed back to the suction side of the compressor via an active carbon filter in both cases.

Smaller quantities of low-pressure gas requiring moderate compression ratios can be handled with *screw compressors.* This type of compressor is used especially in hydrogen liquefaction plants [280].

Apart from these possibilities, compression can also be achieved by pumping liquid hydrogen to the high pressure desired and then evaporating the liquid. This is used for the LH_2 injection in some hydrogen-fueled vehicle engines. Pressures up to 10 MPa may be obtained [281].

Because of the problems encountered in the compression of hydrogen, methods for the indirect compression have been considered.

In the Donor Solvent process (⟶ Coal Liquefaction – Exxon Donor Solvent Process) [282] for coal hydrogenation, aromatic compounds at a low pressure of 5.0 MPa are first hydrogenated to naphthenic compounds. When the hydrogenated donor liquid is subsequently subjected to the conditions prevailing in coal hydrogenation processes (400 °C, 25 MPa), the hydrogen is set free and used for coal hydrogenation. In this way the amount of hydrogen to be compressed can be reduced considerably.

Decompression. The decompression of hydrogen-rich streams is used to produce refrigeration (see Chap. Liquefaction of Hydrogen) and to recover energy. For high pressure differences and small amounts of gas, piston expansion machines are used; large amounts of gas are handled by turbines. For high pressure drops and large amounts of gas a multistage decompression in turbines is applied. Depending on the power and the utilization of the energy the decompressors drive either generators, brake compressors, recycle compressors or oil brakes.

The power thereby generated can be between 1 and 25×10^3 kW. Since the efficiency of turbines increases with increasing speed, high revolutions are desirable. Turbines with revolutions up to 1×10^5 rpm are built. Depending on their power rating, these turbines may have oil, gas or magnetic bearings.

7.3. Transportation and Distribution

The consumers of hydrogen in the chemical industry (ammonia plants, synthesis gas plants, petrochemical plants) usually produce hydrogen on site. Small- and medium-size consumers can also be supplied with gaseous or liquid hydrogen. The principal possibilities for hydrogen transport within the structure of regional hydrogen logistics are shown in Figure 91 [283]. In any future hydrogen energy concept, transportation of large quantities of hydrogen between regions will be necessary and for this reason, transport in pipelines or liquid hydrogen transport in tankers is under consideration.

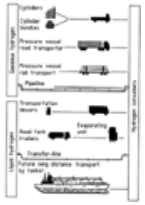

Figure 91. Possibilities for hydrogen transport

High-Pressure Hydrogen Transport. Small amounts of hydrogen can be transported in aluminum pressure cans (contents ca. 5 L H_2) and in steel cylinders of various sizes (up to 70 L with H_2 at 20 MPa). Greater amounts of hydrogen are transported in bundles of steel cylinders or in large sized pressure containers, transport being by road or rail. Modern high-pressure transports have a capacity of >4000 m^3 of hydrogen (see Fig. 92) and can be filled in less than 2 h. Oil-lubricated piston compressors are used to compress the gas to 20 MPa; oil and water adsorbers or, for higher purities, special low-temperature adsorbers assure the necessary product purity.

Figure 92. Road transporter for hydrogen with 9 seamless pressure vessels, max. operating pressure 20 MPa (courtesy of Mannesmann Röhren-Werke, Düsseldorf)

Pipelines play a particular role in the transport of large quantities of hydrogen. Table 36 gives information on existing hydrogen pipeline nets. Such nets are run, e.g., in the USA by Air Products, in France by L'Air Liquide and in the Federal Republic of Germany by Chemische Werke Hüls.

Table 36. Existing hydrogen pipelines

		Length, km	Diameter, mm	Operating pressure, MPa	Flow rate capacity
L'Air Liquide, France, Belgium, Netherlands	GH_2	total 300	100	<10	
Sweden	GH_2	6 transfer lines, 6 km	50 – 250	0.01 – 2.8	10 000 m^3/h
Air Products, Houston, USA	GH_2	62	80 – 30	3.5 – 4.0	270×10^6 m^3/a
Chemische Werke Hüls, Rhein-Ruhr-Gebiet, FRG	GH_2	208	100 – 300	2.2	250×10^6 m^3/a
NASA, Florida, USA	LH_2	0.4	150	0.15	

The last-named system was built in 1938. It has a length of 210 km and can transport 250×10^6 m^3 (STP)/a; four suppliers and fourteen users are tied into the system. The experience gained with this pipeline has been used to provide specifications for planning, constructing, operating and maintaining such a system [284]. If natural gas pipelines are to be converted for the transport of hydrogen, it is necessary to consider the suitability of the steel material incorporated under the expected conditions in the presence of hydrogen (pressure level up to 10 MPa; cyclic loads; see Section Materials).

LH_2 Transport. For laboratory uses liquid hydrogen is supplied in transport dewars. Larger quantities are transported in tanks (5000 L) or in tank trailers (3 – 5×10^4 L); in the United States liquid hydrogen is also transported by rail (10^5 L containers). Oversea containers, containing up to 20 000 L of liquid hydrogen are used to supply the European space project Ariane in French Guayana. During the Apollo program, barges with dewars containing up to 900 m^3 of liquid hydrogen were used.

To keep evaporation losses small, the transport vessels have either perlite vacuum insulation or superinsulation (see Section Storage). Figure 93 shows a transport dewar with an additional liquid nitrogen cooling system (evaporation loss <1 %/d) [285].

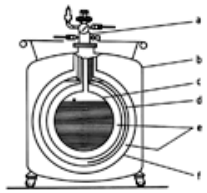

Figure 93. Laboratory transportation dewar with liquid nitrogen cooling [285]

a) Filling manifold and safety valve; b) Protective shell; c) Liquid hydrogen 21 K; d) Liquid nitrogen 77 K; e) High vacuum; f) Outer shell

Test rigs and transfer lines for liquid hydrogen in the space industry are vacuum-insulated. Piston pumps are suitable to transport small quantities of liquid hydrogen at high pressure, whereas larger quantities require multi-stage radial or axial pumps.

Transportation Costs. The total costs of the transport chain from the hydrogen production to the user site have often been assessed, e.g., in [286], [287]. A number of parameters have to be considered, e.g., to transport the same effective energy through transfer lines hydrogen requires 3.5 – 4 times the amount of compression energy compared with natural gas; the total costs are higher by a factor of 1.3 – 1.6. During the distribution of LH_2 an evaporation loss of 7 – 11 % of the handled quantity must be calculated [288]. A summary of the transport and distribution costs for hydrogen is given [289].

Pipelines are by far the most cost-effective means of transporting gaseous hydrogen over both short as well as long distances. The transport of liquid hydrogen via rail or road is ca. 5 times more expensive, the use of pressure gas cylinders is by far the most expensive (transportation costs are ca. 50 times higher, including the costs of idle equipment, such as gas cylinders and vehicles).

7.4. Storage

Conventional methods for storing hydrogen are well-proven and in service. Depending on the end use, (1) stationary, large-size storage systems, (2) stationary, small-size storage systems, and (3) mobile storage systems can be defined. The latter may be used for transport, for distribution or as a fuel reservoir for prime movers. In all these types of storage systems, hydrogen can be handled in the gaseous (GH_2) or liquid (LH_2) state.

Here methods, such as gas storage under pressure or storage in liquid form, are described. Newer concepts, such as absorption in metals (hydride storage systems) and cryoadsorber storage systems are being developed with a view to future use and are described in Section Hydrogen-Energy Conversion Systems. Table 37 shows various concepts for the storage of hydrogen [289]. For stationary, large-volume systems the optimum storage type is determined by the specific storage capacity *T*. This is the amount of energy storage capacity per unit of the rated charging power (kWh/kW). Specific storage cost estimations including further parameters, e.g., utilization, storage capability, annual cost of facility and subsystems are presented in [290].

Table 37. Overview of processes and possibilities for hydrogen storage

Designation	Compressed hydrogen gas storage			Absorption by metal hydrides		Cryo-adsorption	Liquid hydrogen storage	
	below ground	low pressure	high pressure	e.g., FeTi	e.g., FeTi + MgNi		large volume tank	small volume tank
Operating conditions								
Pressure, MPa	10	1.2	15	0.3 – 5	0.3 – 5	4.2	0.15	0.15
Temperature	ambient	ambient	ambient	0 – 80 °C	80 – 300 °C	77 K	21 K	21 K
Developmental status	commercial	commercial	commercial	development	development	development	commercial	commercial
Range of application	site-dependent seasonal storage	medium capacity	short term or mobile	low to medium capacity	mobile storage	medium capacity	long term high capacity	low capacity

Specific energy storage capacity *T*, kWh/kW	>1000	<30	<30	<3		30 – 45	>50	
Problem areas	availability, diffusion losses		safety	costs, energy transfer	capacity, weight	refrigera-tion source	safety	evaporation losses, safety

Compressed Hydrogen Storage. Pressure storage systems of various sizes and pressure ranges are considered state of the art. The same applies for temporary storage in pipelines ("line packing"). For hydrogen, steel cylinders (up to 70 L at 20 MPa), stationary high pressure vessels (20 m^3 at 20 MPa) or low-pressure spherical vessels (15 000 m^3 at 1.2 – 1.6 MPa) are available. Low-pressure and high-pressure systems are suitable for the storage of day or week supplies.

Underground storage in caverns or aquifers is expected to be possible; storage pressure is up to 16 MPa. Gas losses up to 1 – 3 % of the storage capacity are foreseen [291]. Specific problems due to hydrogen are not anticipated, however, the availability of suitable salt and aquifer formations is limited.

Liquid Hydrogen Storage. The technique for liquid hydrogen storage is well-developed; it is used for the storage, transport, distribution of hydrogen. The vessels are insulated with double-walls (dewars), the space between the inside and outside wall is evacuated and filled with perlite (large vessels) or super-insulation. The rate of loss because of evaporation diminishes with increasing size of the vessel. The losses are defined by the amount of hydrogen evaporated daily. Laboratory vessels have 2 – 3 %, mobile LH_2 transport tanks 1 %, perlite-insulated large-volume vessels <0.1 % losses. For additional reduction of heat leakage the refrigeration available in the boil-off gas can be utilized to cool the insulation. This is achieved by radiation shields thermally connected to the vent tube [292]. Liquid hydrogen vessels are attainable from 100 L cans to large tanks with up to 5000 m^3 volume. Tanks have a high specific storage capacity (ca. 100 kWh/kW) and are suitable as long-term storage systems. Figure 94 shows a tank with a capacity of 2000 m^3 LH_2 installed at the Kennedy Space Flight Center in Florida.

Figure 94. Liquid hydrogen tank with a capacity of 3800 m^3 LH_2 at the Kennedy Space Flight Center in Florida (Photo NASA)

Super-Insulation [293]. A large number of layers of insulation foil is wound around the inside vessel (10 – 60 layers per centimeter). A layer consists of a plastic film, both sides of which have been sputtered with metal, or it is produced from a very fine aluminum foil, which is held apart by spacers of a fabric made from glass filaments or similar material. Energy transport by radiation and conduction of the insulation material is more or less completely interrupted. Values of 5×10^{-6} $Wm^{-1}K^{-1}$ for the heat conductivity can be attained.

Liquid hydrogen storage costs have been estimated [286], [294] as has also been the cost of liquefaction [295]. The energy consumption to liquefy the hydrogen amounts to ca. 1/3 of its lower heating value.

Cost Comparison of Large-Volume Hydrogen Storage. The energy stored in a certain volume of hydrogen, relative to that stored in the same volume of natural gas, is determined by the ratio of the volumetric heating values. Therefore, for otherwise equal systems, the hydrogen storage is more expensive by the factor of ca. 3 than natural gas storage. Studies of the cost situation for various hydrogen storage systems allow the following conclusions [289]:

1. Underground hydrogen storage is the most economic method for all areas of application.
2. Storage in pressure vessels is the most economic above-ground storage system when t <30 h. It is suitable for all short-term storage systems.
3. LH_2 storage is the most economic storage system above ground for t > 30 h.
4. It is not clear, whether the recent concepts for hydrogen storage (metal hydrides or cryo-adsorber systems) (see Section Hydrogen-Energy Conversion Systems) are technically and economically more favorable options.

7.5. Materials

Metallic materials are used during production, transport and storage of hydrogen. The choice of material and its stability in the presence of gaseous, liquid or electrolytically produced hydrogen is therefore of great importance and essential for safe handling of hydrogen in a future hydrogen energy carrier concept [296]. Literature on hydrogen and corrosion is given in [297-299].

Mechanism of Hydrogen Corrosion. The destruction caused by hydrogen in various materials occurs via different mechanisms. The starting point is always the formation of diffusable *atomic* hydrogen. This can be formed by: hydrogen ions, thermal dissociation, or chemisorption and dissociation on active metal surfaces.

The embrittlement induced by formation of atomic hydrogen is well-known in electrolytical processes with gaseous hydrogen. Such defects are mainly encountered if the material surface is activated by plastic deformation, i.e., cyclic stress loading makes a metal more susceptible to hydrogen attack.

Atomic hydrogen is dissolved in the material, whereby it either occupies interstitial sites or accumulates in so-called traps. The reactions which lead to damage are described.

Types of Reaction Leading to Damage. The dissolved diffusable atomic hydrogen can interact with lattice dislocations, lattice defects, internal surfaces, or fields of tensile strain caused by mechanical distortion of the material. At these locations, segregation of the material occurs if the concentration of accumulated hydrogen atoms reaches a critical value. The formation of cracks is thereby initiated or, if microcracks or intergranular fractures are present, crack growth is enhanced (HIC, hydrogen-induced cracking [300]).

Hydrogen can react with the material and form hydrides, e.g., with titanium, niobium, tantalum, molybdenum and tungsten. Because of the fact that enhanced hydride formation takes place in the stress field of cracks, and that hydrides are extremely brittle materials, such elements in alloys can lead to hydrogen embrittlement.

A reaction of hydrogen with carbon contained in ferritic steels and alloys at higher temperature (t > 200 °C) is methane formation. This interaction is called decarburization.

$$Fe_3C + 4\,H \rightarrow 3\,Fe + CH_4$$

Recombination of atomic hydrogen to form molecular hydrogen can also be observed. Both reactions, formation of molecular hydrogen and methane, cause high internal pressures because a gas is trapped inside the material. This can lead to a weakening at the grain boundary, followed by formation of cavities, internal cracks and blistering (fish eyes). Macrocracks can form, if the material is placed under strain.

Effects of Hydrogen Pressure Below Temperatures of 200 °C. At constant pressure no critical reaction is observed. Carbon steel equipment may be used up to a recommended hydrogen pressure of 90.0 MPa. At higher pressure austenitic stainless steel should be used.

In areas where the strain changes periodically, e.g., in transport and storage vessels, changes in the ductility and rate of crack formation have been observed. These changes are influenced by the hydrogen pressure, the temperature, and the frequency of the cycling. The rate of the fissure formation was found to increase by a factor of 3 – 100, depending on the cycle frequency and hydrogen pressure as compared to the same treatment in a hydrogen-free atmosphere [301].

Effects of Hydrogen Pressure at Temperatures Above 200 °C. Hydrogen attack with accompanying decarburization by methane formation occurs at >200 °C and increases rapidly with increasing temperature. It depends on the hydrogen partial pressure and on the type of alloyed steel. Metals such as chromium, molybdenum, tungsten, vanadium, titanium, and niobium reduce the number of nucleation sites by lowering the amount of carbon available for methane formation (carbide stabilizers). The behavior of steel at elevated temperature and hydrogen pressure is summarized in a diagram (see Fig. 95). The original curves “Operating Limits for Steels in Hydrogen Service” were first published by G. A. NELSON in 1949 [302]. They were developed as a result of data from a variety of commercial processes and laboratory experiments. Periodic revisions have been made over the years, the latest being published in [303].

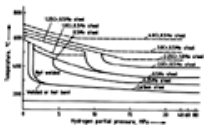

Figure 95. Operating limits for steels in hydrogen service (Nelson diagram)

------- surface decarburization; —— internal decarburization (hydrogen attack)

Below and to the left of the curves for each alloy satisfactory service has been experienced at periods of exposure up to 35 a. The length of time before hydrogen attack can be detected is termed the “incubation period”. It is maintained, that during the incubation period methane pressure builds up in submicroscopic voids; when a void reaches a critical size, microscopically visible fissures appear and the growth rate thereof increases [304]. At temperatures >540 °C the curves are not verified. In this temperature range, high-alloy austenitic steels, which resist decarburization, must be used.

Importance of Foreign Substances. Impurities in hydrogen can inhibit hydrogen attack more or less. Carbon monoxide and water reduce the corrosion activity, oxygen and sulfur dioxide are particularly inhibitive.

Ammonia causes the formation of nitride layers on steel which, when subject to deformation, are ruptured. For this reason, steels used for ammonia synthesis plants must be suitable for both hydrogen and nitride attack. Above 600 °C carbon monoxide causes a carburization of the material. Below 450 °C carbonyl formation can occur which simply removes layers from the iron surface (maximum at P_{CO} 12.5 MPa, 250 °C). Carbon dioxide and hydrogen sulfide are strongly oxidizing at elevated temperatures and can accelerate hydrogen attack. A knowledge of the various forms of hydrogen attack on metallic materials together with the effect of other gas components is especially necessary for high-pressure processes. The choice of material is of particular importance. A summary is given in [305].

Other Hydrogen-Resistant Materials. Other hydrogen-resistant materials are:

Copper and copper alloys up to 400 °C,
Nickel up to 250 °C,
Monel up to 250 °C, and
Inconell up to 540 °C.

For temperatures > 1400 °C the following materials are suitable:

Aluminum oxide
Zirconium oxide
Silicon carbide
Silicon nitride

In general, plastics are not subject to attack by hydrogen at ambient temperatures. At higher temperature, the following materials are suitable:

Poly(vinyl chloride) and polyethylene (up to ca. 60 °C)
Polychloroprene and polyisobutylene (up to ca. 100 °C)
Poly(tetrafluoroethylene) (up to ca. 250 °C)

Materials for Liquid Hydrogen (LH_2). At the low temperatures of liquid hydrogen none of the corrosion inducing mechanisms can take place. Material problems are reduced to the effect on mechanical properties.

Mechanical design with metallic and non-metallic materials at low temperatures as well as construction and assembly methods are compiled [306]. The alteration of material properties, such as elastic moduli, fracture behavior, plasticity and fatigue behavior, is handled in detail.

Only a few metals and alloys commercially available are used in cryogenic engineering. Aluminum alloys are widely used, because of their relative cheapness, good weldability and high toughness. Austenitic steels (300 series) are also widely used, because they retain high strength and good ductility down to low temperature. For cryogenic work, low carbon grades are preferred to prevent the precipitation of chromium carbides during welding. Other materials used are nickel-iron, titanium, and copper alloys.

Selection and use of organic polymer materials at cryogenic temperatures is described [307]. Modification of epoxy or polyester resins by incorporation of inert fillers or the use of fiber-reinforced composite materials and laminated structures using glass, Kevlar, carbon, etc., give materials satisfactory for cryogenic service.

7.6. Removal of Hydrogen

The removal of hydrogen may become necessary, if it develops in an uncontrolled manner and causes a safety risk or if a gas must be purified from traces of hydrogen. Normally, hydrogen, which cannot be utilized, is burned, often together with other disturbing compounds and combustible waste gases. If the hydrogen content is not sufficient to support a flame, the so-called flameless combustion with oxygen at 700 – 800 °C with external heating can be used. This technique is applied particularly often for fume incinerators.

More common is the catalytic combustion on noble metal catalysts (Deoxo) or mixed-oxide catalysts, e.g., CuO – MnO_2 (Hopcalite or similar) [308]; catalytic recombinators are also used in maintenance-free lead accumulators.

Further, hydrogen can be removed by absorption and oxidation, e.g., by aqueous solutions of silver or palladium salts or sodium chlorate. As a solid “absorbent” copper oxide or nickel oxide is suitable.

$$CuO + H_2 \rightarrow Cu + H_2O$$

$$NiO + H_2 \rightarrow Ni + H_2O$$

For the removal of 1 m^3 (STP) of hydrogen ca. 3 kg of copper in the form of copper oxide are necessary; the reaction takes place at 200 °C, the catalyst can be regenerated with air. For production of high-purity inert gases hydrogen traces can also be removed by means of hydride material (see Section New Developments for Hydrogen Storage).

Hydrogen Mitigation in the Nuclear Industry. During normal operation in water-moderated nuclear reactors small amounts of hydrogen are permanently formed by radiolytic splitting of water. In the waste-gas system of nuclear plants recombinators are installed. These are usually commercially available catalytic or thermal units.

In case of a nuclear core uncovery accident, hydrogen may be produced by several mechanisms. (1) The exothermic zircalloy – steam reaction occurs once the fuel cladding reaches temperatures of 900 – 1200 °C. (2) A considerable degree of radiolytic water cleavage inevitably accompanies an emergency core cooling procedure (with a great amount of water) for a period of time after the loss of the coolant has occurred. Hazard analysis studies provide values of 5000 m^3 of hydrogen formed within 50 d [309]. It is believed that at Three Miles Island hydrogen amounting to 8 vol % (370 kg) accumulated in the atmosphere of the containment building. (3) The interaction of a hypothetical core melt with concrete or other materials present within a containment system is considered to be an additional source of hydrogen.

Various specific countermeasures to cope with the hydrogen problem have been studied; some of them have been proposed for installation or have been implemented [310]. First of all, hydrogen distribution is to be prevented by subcompartmented containments, and devices to reduce hydrogen concentration should be provided (fans, recombinators, etc.). In case of massive hydrogen release in small containments, blanketing with nitrogen is foreseen, whereas in intermediate containments the use of water fogs or Halon 1301 injection is recommended to mitigate hydrogen deflagration risk and inhibit combustion. Finally, deliberate ignition and controlled burning systems are reported to be effective hydrogen controlling measures, whereby catalytic and spark hydrogen igniters can be used [311], [312].

7.7. Safety Aspects

General. Hydrogen forms inflammable and explosive mixtures with oxygen and other elements. Basic ignition and explosion data are given in Section Ignition and Detonation Performance. For comprehensive compilations of safety characteristics and hazards on handling hydrogen see [313-315].

The envisaged large-scale introduction of hydrogen as an energy carrier requires a comparative evaluation of safety aspects with regard to other fuels. Comparative data for methane (representing natural gas), propane and gasoline are shown in Table 10. Volumetric leakage of hydrogen gas will be 1.3 – 2.8 times as large as gaseous methane leakage and approximately four times that of air under the same conditions (rule: “air-proof is not hydrogen-proof”).

Released hydrogen disperses rapidly by turbulent convection, drift and buoyancy, thus shortening the hazard duration. The prompt dispersion, however, favors the formation of gas mixtures within the wide flammability and detonability limits; the lower limit is the vital one in most applications and is comparable to that of other fuels.

The minimum energy for ignition of hydrogen – air mixtures is extremely low. However, the ignition energy inherent in virtually every source is more than sufficient for ignition of any other fuel – air mixture as well. Ignition by catalytic action is also possible.

Fire Hazards. Hydrogen flames are nearly invisible in daylight. Fire-detection technology is summarized in [316]. Hydrogen fires last only one fifth to one tenth of the time of hydrocarbon fires, and the fire damage is less severe because of several characteristics:

1. high burning rate resulting from rapid mixing and high propagation velocity,
2. high buoyant velocity,
3. high rate of vapor generation of liquid hydrogen

Although the maximum flame temperature is not much different from that of other fuels, the thermal energy radiated from the flame is only a part of that of a natural gas flame. Smoke inhalation, one of the major causes of injury and, therefore, a main parameter of fire damage, is considered less serious in the case of hydrogen because the sole combustion product is water vapor.

Explosive Hazards. High laminar burning velocity as well as the high laminar flame speed of hydrogen makes the transition to turbulent flame speeds exceeding 800 m/s up to several km/s easy. Hence, hydrogen is more sensitive to *deflagration to detonation transition* (DDT) than hydrocarbons. Studies on DDT mechanisms are referred to in [317]. Hydrogen has by far the widest limits of detonability; detonation of stoichiometric hydrogen – air mixtures in confined spaces will produce a static pressure rise of ca. 15 : 1. If a DDT process precedes, i.e., unburnt hydrogen – air mixtures are precompressed, a pressure rise ratio of 120 : 1 could result.

Explosions are rated in terms of the amount of energy released, commonly expressed as an equivalent quantity of TNT. The theoretical maximum values of explosive potentials for fuels (TNT equivalents) are recorded in Table 20. Experimental data indicate, that real yield factors of 10 % are considered reasonable. Note that hydrogen is most potent on a mass basis and least potent on a volumetric basis. It has also the least theoretical explosive potential, when equivalent energy storage is taken into account.

Preventive Measures. After analyzing accidents involving hydrogen and hydrogen-carrying equipment, preventive safety aspects are discussed in [318]. Risk avoidance by general preventive design measures, as well as reactive fire and explosion measures are presented [315], [318], [319]. In safety concepts, distinctions are made between primary, secondary and tertiary measures. *Primary safety precautions* aim at the exclusion of causative risks such as leakage, formation of explosive mixtures by proper conceptual design (inertization, open-air installation, flame arrestors, etc.). *Secondary measures* consist mainly in the avoidance of ignition sources of any kind (electrostatically or mechanically generated sparks). *Tertiary measures* should minimize dangerous results in case fire or explosion occurs. This is achieved by installation of explosion-proof or explosion relief systems, hydrogen process shut-down systems and suitable fire extinguishing systems.

Safety Regulations. Regulations (mandatory) and standards (mandatory or nonmandatory) apply for the safe production, storage and handling of hydrogen. They are mostly concerned with transportation; other operations are covered by more general regulations. Excellent listings are given for the United States [320] and for Canada [315].

United States. Examples of *nonmandatory standards* are those issued by ASME, ANSI, NFPA, etc., that may be adopted by regulatory bodies. Current *mandatory regulations* are Title 49 of CFR (Code of Federal Regulations) and the requirements of DOT (Department of Transport) and CGA (Compressor Gas Association). General industrial safety matters including the production and handling of inflammable compressed and liquefied gases are regulated by OSHA [321] (Occupational Safety and Health Administration).

Federal Republic of Germany. Mandatory regulations are Druckbehälterverordnungen (TRB, TRG, i.e., regulations for installations and examination of pressure vessels and filling thereof), Unfallverhütungsvorschrift Gase (UVV, i.e., instructions for the prevention of accidents) and GGVS and GGVE, (i.e., instructions for the transport of dangerous goods).

EEC Regulations. For the EEC the association of the European Gas Industry (IGC) is making efforts to introduce common and uniform safety standards.

Future Outlook. Considering the possible use of hydrogen in a future hydrogen energy concept (see Chap. Hydrogen Energy), no major safety problems are anticipated. Hydrogen has been handled successfully for decades in the process industries. Handling as a liquid is also routine. In comparison with other inflammable gases or liquids, the safety risk of hydrogen as a fuel is not significantly different.

7.8. Toxicology

Hydrogen does not show any physiological effect. It is nonpoisonous. Inhalation of the gas leads to sleepiness and a high-pitched voice. A danger of asphyxiation exists if the oxygen content sinks below

18 vol % because of hydrogen accumulation in the air. Direct skin contact with cold gaseous or liquid hydrogen leads to numbness and a whitish coloring of the skin and to frost bite. The risk is higher than, e.g., with liquid nitrogen because of the greater temperature differences and the higher thermal conductivity of hydrogen.

7.9. Analysis

The physical properties of hydrogen, some chemical reactions, and its nuclear spin characteristics are used for hydrogen analysis.

Hydrogen in Gas Mixtures. For rapid determination or for continuous measurement in process gases the absolute determination (or the comparison to a reference gas) of properties such as *gas density, conductivity* or *sonar velocity* can be used. The most frequently used method for analysis of hydrogen in all types of gas mixtures is *gas chromatography.* Apart from mixtures also containing helium, hydrogen is eluated first. In the case of detection with a hot-wire detector using helium as carrier gas an unusual nonlinearity exists in the conductivity of hydrogen – helium mixtures, which leads to a minimum at 8 % hydrogen. This must be taken into consideration, or alternately argon can be used as carrier gas. Hydrogen contents in the percent range can be determined by separation and integral detection according to JANAK (CO_2 as carrier gas, absorption of the carrier gas in KOH, volumetric determination of the gas in an azometer).

Gas-chromatographic determination of the six isotopic species and their nuclear spin isomers is possible. Alumina or silica gel at liquid nitrogen temperature [322] or complex polymer molecular sieves at room temperature [323] are suitable separation materials. If these are impregnated with paramagnetic substances such as Fe_2O_3 or $MnCl_2$, rapid o – p isomerization is achieved and the spin-isomer mixture corresponding to the equilibrium at the separation temperature is eluated [324].

Trace impurities in hydrogen can be determined qualitatively by low-temperature condensation or by enrichment on silica gel at low temperatures and, after desorption, they can be separated by gas chromatography. A method for the determination of traces of carbon monoxide or carbon dioxide is methanation and analyzed with hydrogen followed by methane detection with a sensitive flame ionization detector.

Chemically Bound Hydrogen. In *organic* molecules hydrogen is determined using classical elemental analysis. After the sample is weighed (usually in the micro- or submicro-range, i.e., 0.2 – 10 mg and 0.02 – 0.2 mg, respectively) it is combusted and the amount of water produced, as determined by weighing, is related to the hydrogen content. Nowadays automatic C, H, N analyzers with simultaneous detection systems are used. The content of hydrogen in *metals*, an important value for steel production, is determined firstly by vacuum extraction of the hydrogen from the heated or molten sample and, finally, by volumetric (vacuum pipette) or heat conductivity measurement [325]. In a new method, a sample is taken by inserting a porous ceramic bell and purging out the hydrogen with a nitrogen stream [326].

“Active hydrogen”, e.g., hydrogen in OH groups or in enolizable C-H-bonds, reacts with $LiAlH_4$ or metal organic methyl compounds and can be determined as hydrogen or methane volumetrically.

Modern instrumental methods are mass spectrometry and nuclear spin resonance. By means of the first method hydrogen and its isotopes in complex gas mixtures can be analyzed over a large concentration range. Nuclear magnetic resonance (^{1}H-NMR) is an invaluable instrument for the analysis and structure determination in organic and inorganic chemistry. ^{1}H-NMR spectroscopy is also applied for quantitative determination of the hydrogen content in various materials such as coal or hydrocarbons, to measure oil and water contents, e.g., in the food industry, or to pursue intricate problems in reaction kinetics [325].

[Top of Page]

8. Conventional Uses

8.1. Hydrogen in the Chemical Industry

Hydrogen is an important raw material for the chemical industry. It takes part in reactions either by addition (hydrogenation) or by means of its reduction potential. Most of the hydrogen is used for these hydrogenation and reduction processes. An overview on production and consumption structures for the Federal Republic of Germany (1978) and the United States (1984) is given in Figure 96 [328]. Changes in the production and utilization sector have been relatively little since that time. On the *production side* the fraction of the hydrogen produced using natural gas has increased at the expense of the LPG and naphtha-based plants. Some plants based on heavy oil have even been closed down and replaced by natural gas or lignite gasification plants. In the *utilization sector*, hydrogen production is stagnating. Apart from ammonia synthesis, synthesis with hydrogen – carbon monoxide gas mixtures to produce methanol, hydrocarbons and oxo-synthesis products are of particular importance. Large quantities of hydrogen are needed in refineries for hydrotreating, whereas its use in coal processing is just beginning. The use of hydrogen in metallurgy is based in particular on its reducing properties.

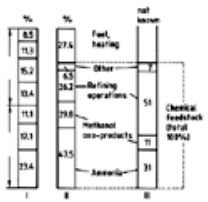

Figure 96. Hydrogen production and consumption structures

I Production, Federal Republic of Germany, 1978

II Consumption, Federal Republic of Germany, 1978

III Consumption, United States, 1984

At present, pure hydrogen is not used directly to produce energy (apart from space technology). Mixtures with hydrogen, however, have been used for a long time for combustion purposes. Even today the

amount of hydrogen distributed in urban or in industrial heating gas systems is very large.

In comparing the hydrogen production or usage data, a distinction must be made between the so-called captive hydrogen and the merchant hydrogen: *Captive hydrogen* uses, such as for ammonia and methanol synthesis or in various hydrotreating processes, relate to onsite hydrogen production and immediate consumption of the hydrogen. The designation *merchant hydrogen* is given to those quantities which are sold on the market. Merchant sales of hydrogen account for 10 % or less of the total hydrogen production. Import and export amounts are normally negligible. Table 38 gives a summary of a literature survey on hydrogen quantities produced as well as some forecasts for the future.

Table 38. Consumption of hydrogen and forecast consumption amounts [329-338]

	Consumption of hydrogen, $m^3 \times 10^9$		
Federal Republic of Germany	1978: 17 [329]	1984: 19.1 [330]	2000: 23 [331]
Japan		1985: 20 [335]	
United States	1978: 80 [332]	1985: 90 [333]	2000: 146 – 207 [334]
Worldwide	1980: 300 [336]	1986: 500 [337]	2025: 1500 – 2000 [338]

8.1.1. *Ammonia Synthesis*

Hydrogen reacts with nitrogen to yield ammonia. The reaction is exothermic.

$$3\,H_2 + N_2 \rightleftharpoons 2\,NH_3;\quad \Delta H_{298} = -91.22\,kJ/mol$$

For details on the properties, production and usage of ammonia see ⟶ Ammonia.

Technical ammonia syntheses are operated at 15 – 25 MPa. On a catalyst the reaction starts at ca. 350 °C, maximum reaction rates are reached in the range of 470 – 530 °C.

1 965 m^3 (STP) of hydrogen and 668 m^3 (STP) of nitrogen are necessary for each tonne of ammonia produced. For hydrogen production, the usual processes are suitable (see Section Production from Coal and Hydrocarbons). The production of the ammonia synthesis gas is usually followed directly by the ammonia synthesis. The necessary nitrogen content in the synthesis gas is introduced by the addition of air in a secondary reformer or via a liquid nitrogen scrubber. The conversion achieved per pass is ca. 25 – 35 %, so that a recycle loop is installed. Condensed liquid ammonia is removed continuously. Impurities containing oxygen, such as carbon monoxide, carbon dioxide, oxygen and water, are catalyst poisons and must be completely removed from the synthesis gas.

The make-up gas to the synthesis loop contains certain amounts of inert gases such as methane (from methanation), argon (from the process air) or helium (from natural gas). To prevent accumulation of the inert gases in the synthesis reactor, some recycle gas must be continuously removed from the loop, thus causing a loss of hydrogen. Most plants nowadays have a recovery unit:

- cryogenic hydrogen recovery units, sometimes combined with argon production (see Section Low-Temperature Processes)
- PSA hydrogen recovery units (see Section Adsorption Processes)
- membrane separation hydrogen recovery units (see Section Permeation Processes).

Hydrogen recovery by reversible hydride formation has also been suggested [339].

Modern synthesis plants have a capacity of 1000 – 2000 t/d; this corresponds to an hourly hydrogen requirement of 80 000 – 160 000 m^3 (STP). At present 75 – 80 % of the synthesis hydrogen is produced from natural gas, naphtha or LPG. 10 – 15 % are made from heavy fuel oil by partial oxidation and ca. 5 % from coal. Electrolysis hydrogen is used for ammonia production only in special cases, e.g., if cheap electric power is available. Small units are situated in Norway, Peru and Egypt [340], producing ca. 5 % of the world production of hydrogen.

An ammonia molecule contains three hydrogen atoms. For hydrogen utilization in future energy systems (see Chap. Hydrogen Energy), ammonia has often been suggested as a carrier. Hydrogen could be stored and transported more economically in the form of ammonia [341], but the latter is untenable as a direct fuel because of its toxicity and the formation of NO_x. However, on reconverting the ammonia to hydrogen the stored energy can be made available at any desired site. An overview of the possibilities of an ammonia-hydrogen concept in energy technology is given in [342].

8.1.2. *Hydrogen in Refinery Processes (⟶ Oil Refining)*

In refineries hydrogenation processes are included to increase the hydrogen content of the heavy crude oil fractions and to produce lighter fractions by reduction of the molecular mass. At the same time, undesirable elements, such as sulfur, nitrogen and metals, can be removed. For this group of related processes the designation hydroprocessing is used.

Process Fundamentals. Hydrogenation processes are characterized according to the feedstock, the partial pressure of the hydrogen used and the reaction temperature (see Fig. 97) [343]. In

hydrotreating processes addition reactions and hydrogenating cleavage (hydrogenolysis) occur. In *hydrocracking processes* mainly dehydrogenation, cracking, isomerization and dealkylation reactions take place. For the catalysts used in the individual processes, promoter and carrier materials, selectivity, life-times, coke and metal deposition see [344]; a compilation of available hydroprocessing catalysts is given in [345]. Newly developed bi- or multifunctional catalyst systems based on zeolites combine cracking functions on acid centers of the carrier materials with hydrogenation and dehydrogenation functions of the active metal components.

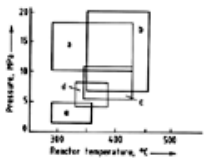

Figure 97. Process temperatures and pressures of refinery hydrogenating processes

a) Hydrocracking; b) Residue desulfurization; c) Heavy gas oil catalytic cracker feed hydrofining; d) Middle distillate; e) Naphtha hydrofining

The processes are carried out in one or two stages in the gaseous or liquid phase. Table 39 shows some examples of the most important hydroprocessing routes, typical process conditions and product ranges. Overviews are given in [346] and [347].

Table 39. Feed and product characteristics and process parameters of various hydrogenation processes during hydroprocessing

	Unionfining (UNOCAL)	Go-Fining (Exxon)	Hydroconverting (Linde)	Hydrocracking (BASF/IFP)
Feed	virgin kerosine	arabian heavy VGO [a]	VGO [a]	arabian light VGO [a]
Density, g/cm^3	0.806	0.933	0.910	0.909
Boiling range, °C	160 – 240	370 – 560	340 – 550	370 – 475
Sulfur content, wt %	0.4	2.96	2.7	2.55
Processing conditions				
Reactor stages	1	1	1	2 (for max. gasoline production)
Pressure, MPa	2.5 – 6.0		5.8	moderate
Temperature, °C	340		395	400°/400 °C
Hydrogen consumption, m^3/t	10	87	150	390
Products				
Yields, wt %	naphtha, 1 %	LPG, 0.4 %	LPG + gasoline, 39 %	C_4-, 15.6 %
	destillate, 99 %	gasoline, 0.8 %	middle distillate, 28 %	light gasoline, 23.8 % heavy gasoline, 61.3 %
		middle distillate, 100.6 %	residues, 33 %	H_2S + NH_3, 2.8 %
Main product				
Density, g/cm^3		0.890	0.813	0.664/0.758
Boiling range, °C	163/240	>200		
Sulfur content, wt %	0.01	0.1	0.2	
RON [b] of gasoline fraction			>88	86

[a] VGO = vacuum gas oil.
[b] RON = research octane number.

The use of hydrogen has also been suggested in other refinery areas; addition of hydrogen to the crude oil distillation is reported to make this more economic (*hydrostripping*, [348]). Subjecting residual oils or heavy crudes to hydrogen for short periods at high temperatures and pressures of up to 50 MPa leads to feed upgrading (*hydropyrolysis*, [349]).

Hydrogen Requirements. The hydrogen necessary for the hydroprocessing is produced in the refinery. The main source is the catalytic reforming unit where ca. 45 m^3 of hydrogen per ton of processed crude oil is produced. Additional hydrogen is produced by steam reforming of methane, LPG, and naphtha, and by partial oxidation. In the United States 60 – 70 % is produced by catalytic reforming, 30 % by steam reforming, and 2 % by partial oxidation.

The connections between processes utilizing and processes producing hydrogen in oil refining operations are shown schematically in Figure 98.

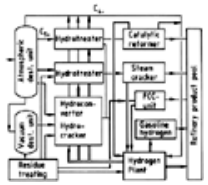

Figure 98. Integration of H_2-producing and H_2-consuming process units into crude oil processing

⇒ H_2 units from H_2-producing process

⇒ H_2 units to H_2-consuming process

The carbon:hydrogen ratio in the feedstocks (heavy and residual crudes, tar sands, oil shale) as well as their sulfur, nitrogen and metal components will become increasingly unfavorable in the future. To cover the expected increasing hydrogen demand it is necessary to improve the hydrogen recovery in refineries and, eventually, to utilize other potential sources of byproduct hydrogen.

8.1.3. *Hydrogen in Coal Refinement*

Coal refinement processes which are performed in the presence of hydrogen are coal hydrogenation, hydropyrolysis and hydrogenating coal gasification.

Coal Hydrogenation. (⟶ Coal Liquefaction – Direct Coal Liquefaction, the first three Sections). By thermal or thermocatalytical treatment in the presence of hydrogen under pressure, coal is almost completely converted into liquid and gaseous products (direct coal liquefaction). Details of the chemical fundamentals, types of process and reaction techniques are given in [350]; the use of catalysts is described in [351]. The reaction is catalyzed by iron compounds (sulfides or oxides), molybdenum, cobalt and tin or even mineral clays. Catalysts which can be cheaply disposed of are especially favored, but also highly-active cobalt – molybdenum carrier catalysts are being investigated.

Hydrogen can either react directly (in its molecular form) with coal (direct hydrogenation) or it can be transferred by means of hydrogen-carrying solvents (donor solvent coal liquefaction or hydrogenating extration) (⟶ Coal Liquefaction – Exxon Donor Solvent Process). This type of process with chemically bound hydrogen can be carried out in the presence of an additional amount of molecular hydrogen and is known as short hydrogenation, the product being primarily bitumen or solvent-refined coal (SRC).

Hydrogen is required for the important process stages, such as hydrogenative cracking in the slurry phase (feedstock slurries of hard coal, lignite or coal extract), cracking and refining hydrogenation in mixed phases (feedstock hydrogenation residues, tars) and for cracking and refining hydrogenation in the gas phase (feedstock middle distillates and bottom cuts). Donor solvent processes employ external, catalytical or noncatalytical hydrogenation of the recycled solvent.

Hydrogen Consumption and Product Range. The products obtained range from small quantities of coal residue over heavy and middle distillates and gasoline to gases. The boiling range of the hydrocarbon products depends on the quantity of hydrogen. This is shown in Figure 99 for hydrogenation in the slurry phase. Hydrogen consumption per tonne of dry coal for the catalytic direct coal hydrogenation amounts to ca. 1000 m^3.

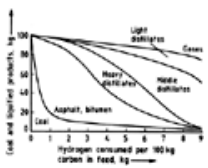

Figure 99. Relation of hydrogen consumption and product distribution during slurry-phase hydrogenation of coal

Hydrogen Production, Recycle Gas Separation. In all coal hydrogenation processes nonreacted hydrogen must be recovered and recycled back to the reaction together with fresh make-up hydrogen. The economics of the projected large-scale processes for direct hydrogenation depend not only on the further improvement of the hydrogenation stage (milder hydrogenating conditions, improved liquid product yields) but in particular on:

improved separation of the gaseous hydrogenation products with complete recycle of hydrogen with an acceptable purity (inert components decrease the partial pressure of the hydrogen) and low-cost production of the make-up hydrogen using improved and integrated processes for steam reforming and partial oxidation of byproducts of the coal hydrogenation, which are otherwise

difficult to utilize.

Process Developments. After the oil embargo of 1973, many developments were revived. The developments being furthered in the Federal Republic of Germany are based to a great extent on the *modified IG process.* Two plants on a pilot scale (6 t/d and 200 t/d) have been erected. In the United States, on the other hand, the process concepts being pursued involve new process steps, e.g., SRC 1, SRC 2 (Gulf Oil), Exxon Donor Solvent Process, H-coal process (Hydrocarbon Research). An overview is given in [352]. Improved two-stage processes are finding increasing attention, for example in Japan [353].

Hydropyrolysis. In this process coal is subjected to high temperatures (700 – 1000 °C) for short residence times in the presence of hydrogen and thereby thermally cleaved and partially hydrogenated (flash pyrolysis). The hydrogen consumption of 200 – 300 m^3 per tonne of coal is relatively low and 25 – 45 % liquid hydrocarbons are obtained; about half of the feedstock coal is recovered in the form of a low-sulfur residue coke [354].

Hydrogasification of Coal. In hydrogasification processes, coal is converted in the presence of hydrogen and/or steam in one stage or in a series of stages directly to methane. The aim of many process developments is the production of SNG.

The basis is the strongly exothermic reaction of hydrogen with carbon according to (greatly simplified):

$$C_{(coal)} + 2\,H_2 \rightarrow CH_4;\;\; \Delta H_0 = -87.4\,kJ/mol$$

Equilibrium data are given in [350]. The equilibrium yield of methane is increased at high pressure and low temperature. Table 40 shows calculated methane concentrations at equilibrium. The reaction sequence is a combination of the rapid hydrogenating gasification of the organic components in the coal and the slow heterogeneous gas – solid reaction between semicoke or coke residue and molecular hydrogen. The reaction occurs relatively slowly in the temperature range of 700 – 1300 °C; the technically feasible operating conditions are nowadays considered to be between 800 and 900 °C and 4.0 – 10.0 MPa. The multiphase reaction must be conducted in a fluidized bed and, because of the short residence times resulting, gas recycle processes must be used.

Table 40. Methane concentrations of the carbon hydrogenation reaction at equilibrium

Temperature, °C	Pressure, MPa	CH_4 content, vol %
800	1	26
	10	63
	100	87
1000	1	8
	10	37
	100	72

Most important for the economics are the hydrogen production costs. The overall methane yield per kilogram of carbon depends in particular on the process used to produce hydrogen:

hydrogen from the coal feedstock, autothermal	0.6 m^3 (STP) CH_4 per kilogram of carbon
hydrogen from recycle methane, allothermal	1 m^3 (STP) CH_4 per kilogram of carbon
hydrogen from external sources, e.g., water electrolysis	1.8 m^3 (STP) CH_4 per kilogram of carbon

Examples. The process developed by *Union Rheinische Braunkohlenkraftstoff AG* operates in a fluidized bed at 800 – 1000 °C and 5.5 – 9.5 MPa (data from the pilot plant operation). The crude gas obtained contains up to 40 % methane; the degree of gasification is ca. 82 %. For hydrogen production the residue coke is gasified in a High-Temperature Winkler (HTW) gasifier [355].

Hygas process (Institute of Gas Technology, United States): Hydrogen is produced from the residue coke in the lower part of a multistage reactor.

The *Hydrane process* (Bureau of Mines, United States) uses a two-stage gasifier with external hydrogen production from residue cokes, oxygen and steam.

8.1.4. *Synthesis Gas*

The process for the production of hydrogen from hydrocarbons yields gas mixtures containing mainly hydrogen and carbon monoxide (see Section Production from Coal and Hydrocarbons). Such gases are named synthesis gases since they are used for the synthesis of special chemical products. Some hydrogen – carbon monoxide combinations (because of the production method or the final use) have the designation water gas, cracked gas, methanol synthesis gas or oxo-synthesis gas. The fundamental reactions of synthesis gas chemistry are methanol synthesis, Fischer – Tropsch synthesis, oxo synthesis (hydroformylation), and methane synthesis; further synthesis gas reactions are being developed. Excellent overviews are given in [356][357][358].

Methanol Synthesis. Methanol can be produced from hydrogen – carbon oxide mixtures by means of the catalytic reaction of carbon monoxide and some carbon dioxide with hydrogen. Methanol may be considered virtually as synthesis gas at maximum compression.

$$CO + 2\,H_2 \rightleftharpoons CH_3OH \quad \Delta H^0 = -90.84\,kJ/mol$$

$$CO_2 + 3\,H_2 \rightleftharpoons CH_3OH + H_2O \quad \Delta H^0 = -49.6\,kJ/mol$$

The presence of a certain amount of carbon dioxide in the percentage range is necessary to optimize the reaction. Side reactions, also strongly exothermic, can lead to formation of byproducts such as methane, higher alcohols, or dimethylether.

2520 m^3 (STP) of synthesis gas (70 % H_2, 21 % CO, 7 % CO_2) are necessary to produce a tonne of methanol. The purity of the latter depends on the catalyst used. Copper – zinc catalysts used in the low-pressure process require a sulfur-free gas (H_2S <1 ppm [mL/m^3]). Catalysts for the high- and medium-pressure processes (ZnO activated with chromic acid) can accept 30 ppm (mL/m^3) of hydrogen sulfide. For details on the process schemes, catalysts, process techniques, economics, etc. see, for example, [360].

Fischer – Tropsch Synthesis. Hydrogen – carbon monoxide mixtures react to hydrocarbons according to reaction equations (58) – (62), (60) is the main secondary reaction. An example for one of the general equations for alkane synthesis is (63), for olefin synthesis (64). The overall reaction balance including synthesis gas formation corresponds to an indirect coal hydrogenation.

$$CO + 2\,H_2 \rightarrow -CH_2- + H_2O \quad \Delta H^0 = -165.0\,kJ/mol \tag{58}$$

$$2\,CO + H_2 \rightarrow -CH_2- + CO_2 \quad \Delta H^0 = -204.7\,kJ/mol \tag{59}$$

$$CO + H_2O \rightarrow H_2 + CO_2 \quad \Delta H^0 = -39.8\,kJ/mol \tag{60}$$

$$3\,CO + H_2O \rightarrow -CH_2- + 2\,CO_2 \quad \Delta H^0 = -244.5\,kJ/mol \tag{61}$$

$$CO_2 + 3\,H_2 \rightarrow -CH_2- + 2\,H_2O \quad \Delta H^0 = -125.2\,kJ/mol \tag{62}$$

$$nCO + (2n+1)H_2 \rightarrow C_nH_{2n+2} + nH_2O \tag{63}$$

$$nCO + 2\,nH_2 \rightarrow C_nH_{2n} + nH_2O \tag{64}$$

The Fischer – Tropsch synthesis is strongly exothermic. The total heat of reaction amounts to ca. 25 % of the heat of combustion of the synthesis gas. Undesirable side reactions include methanation, the Boudouard reaction, coke deposition, oxidation of the catalyst, or carbide formation. Other competing reactions such as alcohol, aldehyde, ketone, carbonic acid, and ester formation, also occur.

For development of the process, kinetics, thermodynamics and mechanism of the reactions, distribution of the products, etc., see ⟶ Coal Liquefaction and [359-362]. Up to 1954 4 000 publications and 4 020 patents regarding Fischer – Tropsch catalysts had already been reported [363], but virtually only those based on cobalt and iron have been successful.

The yield of the Fischer – Tropsch synthesis is determined by the initial hydrogen – carbon monoxide ratio and the stoichiometric requirements of the desired reaction products. The maximum yield is obtained when the aforementioned are nearly equal and thus yield 208.5 g hydrocarbon from 1 m^3 (STP) of synthesis gas; technically ratios of 1.7 : 1 to 3 : 1 are used.

The Fischer – Tropsch synthesis is used on a technical scale nowadays only in South Africa at SASOL. At SASOL I, five units of the ARGE reactor of the Ruhrchemie – Lurgi (gas phase, fixed bed reactors, 2.3 – 2.5 MPa, 220 – 250 °C, 40 m^3 Fe – Cu – K_2O – SiO_2 catalyst, 9 – 12 months running time) and three reactors using the Synthol circulating fluidized bed of Kellogg – SASOL (gas phase, 2.0 – 2.3 MPa, 300 –340 °C, 30 – 140 t Fe catalyst produced by reduction of magnetite; running time 45 d) are in operation. At SASOL II (start-up 1980) and SASOL III (start-up 1982) only the more flexible Synthol process is used. The enormous gas amounts (2×1 650 000 m^3 [STP]/h raw gas, 2×1 000 000 m^3 [STP]/h feed gas) make the recovery of byproducts (ethylene, alcohols esters, carbonic acids, etc.) economically worthwhile [364].

Variations of the Fischer – Tropsch synthesis with other selective catalysts can lead to the following products:

low-molecular-mass olefins (C_2 – C_4, mostly ethylene) [365], [366]
medium- and long-chain olefins [366]
polyethylene [367]

Polyethylene is obtained at high pressures (100 – 200 MPa), low temperature (100 – 120 °C), and using a ruthenium-based catalyst.

Methane Synthesis. The catalytic hydrogenation of carbon monoxide to methane is strongly exothermic

$$CO + 3\,H_2 \rightarrow CH_4 + H_2O \quad \Delta H = -206.4\ \mathrm{kJ/mol}$$

Consecutive and side reactions (shift conversion, Boudouard equilibrium, hydrogenation of carbon) make the calculation of equilibrium conditions very complex. The influence of the feed gas composition and operating parameters on the equilibrium composition of the product gas and the kinetic data, which are somewhat contradictory, are summarized in [368].

Classical methanation catalysts are magnesium-promoted nickel catalysts with diatomaceous earth as carrier. Catalyst developments are aimed at increasing the thermal and mechanical stability and yielding catalysts which are simultaneously suitable for the cracking of naphtha with subsequent methanation. Methanation is technically important as a catalytic purification step for the removal of carbon monoxide from gases (see Section Catalytic Gas Purification).

Hydroformylation of Olefins. Hydroformylation, also called oxo synthesis or Roelen reaction (O. Roelen, 1938), is a commercial-scale process for the production of aldehydes. By catalytic addition of hydrogen and carbon monoxide to an olefin an aldehyde is obtained under chain elongation

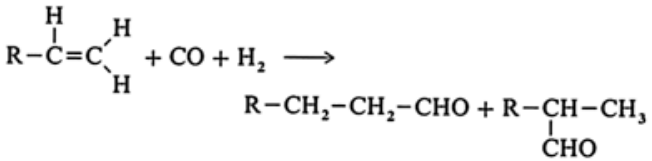

Hydroformylation is generally an exothermic, homogeneously catalyzed liquid-phase reaction of the olefin with hydrogen and carbon monoxide. As catalyst cobalt carbonyl hydride or, in some variations of the oxo-synthesis, cobalt- or rhodium-phosphin complexes are used.

Propen is the olefin mostly used. The oxoproducts are converted to alcohols, carboxylic acids, aldol-condensation products, and primary amines. About 20 commercial processes are state of the art. An excellent review is given in [369] (see ⟶ Oxo Synthesis).

Homologation. Under the reaction conditions of the hydroformylation alcohols and aldehydes react with carbon monoxide – hydrogen under elongation of the chain by one CH_2- unit

$$R-CH_2-OH + CO/H_2 \rightarrow R-CH_2-CH_2-OH + H_2O$$

Homologation has been performed with a number of alcohols, the production of ethanol from methanol has been most intensively investigated. The homologation is not used industrially because of the many side reactions which take place [370].

Synthesis Gas as Chemical Feedstock. Hydrogen – carbon monoxide mixtures, hydrogen alone, and their primary product methanol are important feedstocks for the chemical industry. Nowadays, ethylene which is produced from propane, ethane, naphtha, or gas oil is the most important feedstock for the production of industrial organic chemicals in the chemical industry. Bascially, it is, however, possible to obtain these compounds from synthesis gas thus changing the feedstock basis to coal (see Fig. 100).

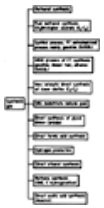

Figure 100. Synthesis gas as feedstock in the chemical industry

8.1.5. ***Hydrogen in Organic Synthesis***

Hydrogen is required for the production of chemicals and intermediates in organic chemistry. A large number of hydrogenations or reductions are carried out on a technical scale (⟶ Hydrogenation and

Dehydrogenation).

Activated and nonactivated double and triple bonds in olefins and acetylenes can be easily partially or totally hydrogenated, whereas the hydrogenation of aromatic and heterocyclic bonds requires more energetic conditions. Functional groups, such as carbonyl, nitro, nitroso, and nitrile groups, can also be hydrogenated.

The reaction conditions are dictated by equilibrium (65). The reactions are exothermal and run in the presence of a catalyst.

$$A + nH_2 \rightarrow BH_{2n} \tag{65}$$

Directions for carrying out catalytic hydrogenations on a laboratory or industrial scale are given in [371]. Summaries of hydrogenation reactions are given in [372] and [373]. Hydrogenation catalysts are metals of groups 8 – 10 of the periodic system (see front matter of this volume), e.g., Raney nickel, as well as copper and molybdenum. In particular the noble metals (Pt, Pd), are highly-active catalysts [374]. Homogeneous systems with molecularly dispersed catalysts in the solution, can be used for special synthesis problems (selective hydrogenation, asymmetric synthesis) but are at present of no great importance in commercial areas because of the frequently encountered difficulty to remove the catalyst from the reaction mixture. Table 41 gives an overview of commercially used hydrogenation reactions.

Table 41. Selection of important industrial hydrogenation reactions

Reaction	Product	Process features	Uses
Hydrogenation of unsaturated hydrocarbons and aromatics			
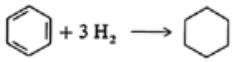 (analogous: cyclohexanol from phenol, cyclohexane carboxylic acid from benzoic acid)	cyclohexane	middle pressure hydrogenation over Ni-/Pt-Li-Al_2O_3 catalyst in liquid phase (IFP, Mitsubishi) gas phase (UOP, DSM, Thoray, Houdry)	starting material for nylon production (cyclohexanone/ol, adipic acid, caprolactam), solvent
$CN-CH_2-CH=CH-CH_2-CN + H_2 \longrightarrow CN-(CH_2)_4-CN$	adiponitrile	Du Pont (300 °C), liquid phase hydrogenation 25–30 MPa/70–100 °C over Raney nickel	hexanediamine
$OH-CH_2-C\equiv C-CH_2OH + 2H_2 \longrightarrow OH-(CH_2)_4-OH$	butanediol	trickle bed, 20 MPa, 180–200 °C Ni catalyst with Cu-, Cr-promotors	polyesters, polyurethane-plasticizer component, solvent
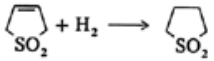 Selective hydrogenation as purification step during production of ethylene, propene, and butadiene	sulfolane	Shell process, 11 000 t/a	aromatics extraction solvents, sour gas, scrubbing agent
Further processing of oxo-products			
$R-C(=O)H + H_2 \longrightarrow R-CH_2OH$ (analogous: ethylhexanol from ethylhexanal)	oxo-alcohols	gas phase hydrogenation at 2–0.3 MPa/115 °C, Ni catalyst sump phase hydrogenation, 8 MPa/115 °C Ni catalyst	components for solvents, plasticizers, detergents
$R-C(=O)H + NH_3 + H_2 \longrightarrow R-CH_2-NH_2$	primary amines	hydrogenation (up to 30 MPa, 25–130 °C), Raney nickel catalyst	
Hydrogenation of other ketones and aldehydes			
Maleic acid $\longrightarrow$ butyrolactone $\longrightarrow OH-(CH_2)_4-OH$	butanediol	Mitsubishi, Kao Ind. (Japan) process	
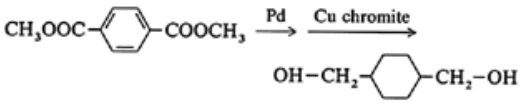	bis-(hydroxymethyl)-cyclohexane	two step Eastman-Kodak process	

Reaction	Product	Process features	Uses
$CH_2{=}CH{-}C(=O)H + H_2 \longrightarrow CH_2{=}CH{-}CH_2OH$	allylalcohol	Degussa gas phase process (heterogenous catalysis)	starting material for glycerol, glycidol
2 Acetone $\longrightarrow$ mesityl oxide $\xrightarrow{\text{Pd zeolite}}$ $i\text{-Bu}{-}C(=O){-}CH_3$	MIBK	hydrogenation at Pt–zeolites (modified one step processes)	(extraction) solvent
Fats, oils + $H_2 \longrightarrow R{-}CH_2{-}OH$	fatty alcohols	20–40 MPa/200–400 °C, catalysts of Adkins-type	
$R{-}O{-}CH_2CH_2{-}OH + HCHO + H_2 \longrightarrow R{-}O{-}CH_2CH_2{-}OCH_3$	polyethylene-glycol ethers	new Hoechst process	sour gas scrubbing agent (Selexol)
Hydrogenation of N-compounds			
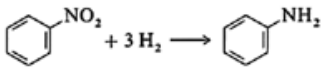	aniline	fixed bed hydrogenation over NiS/CuS, 300–475 °C (Bayer, Allied, Lonza), fluidized bed hydrogenation, 5 MPa/100 °C over Cu catalyst (BASF, Cyanamid)	starting material for dyes, pharmaceuticals, isocyanate polymers, solvents
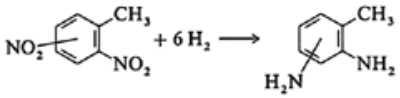	diaminotoluenes	analogous to nitrobenzene reduction	
Nitriles + $H_2 \rightarrow$ primary amines	hexanediamine	5–15 MPa/60–130 °C over Raney Ni or Raney Co	starting material for fibers, sour gas scrubbing agent
Miscellaneous reactions			
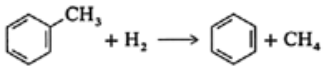	benzene	3–5 MPa/500–650 °C over Cr_2O_3–Mo_2O_3–CoO catalyst (Houdry, UOP, Shell, BASF) 10–25 MPa/400–500 °C over Pt–Al_2O_3–SiO_2	50 % of toluene production further processed to benzene
Xylene isomerization under H_2 partial pressure	o-/p-xylene	catalysts	terephthalic acid, phthalic acid

High-purity hydrogen is necessary for the partial or total hydrogenation of fats and oils (for the production of edible fats or for technical purposes). In fat hydrogenation the polyene, triene, and diene fatty acids in their glyceride form are selectively hydrogenated to the corresponding monoene acids.

The industrial production of sugar alcohols, such as sorbitol, xylitol or mannitol from the corresponding sugars is carried out by catalytic hydrogenation. Batch suspension processes using Raney nickel catalysts are mainly employed under reaction conditions of 120 – 150 °C and 3 – 7 MPa [375].

8.1.6. *Hydrogen in Inorganic Synthesis*

The catalytic hydrogenation of anthraquinone and its derivatives followed by their auto-oxidation to yield hydrogen peroxide is the basis of the commercially important process for hydrogen peroxide production (⟶ Hydrogen Peroxide – Anthraquinone Process (AO Process)). Further important reactions in inorganic chemistry are the production of hydrochloric acid from hydrogen and chlorine (⟶) and the hydroxylamine synthesis (⟶ Hydroxylamine).

8.2. Hydrogen in Metallurgy

Iron Metallurgy. To reduce iron ore, apart from coke (classical blast furnace process), other reducing agents can be used. For reduction a gas containing hydrogen, carbon monoxide, or mixtures of these is suitable. The reduction gas is produced by steam reforming or partial oxidation of fossil fuels. These *"direct reduction"* processes (⟶ Iron – Production) yield sponge iron, which can be melted to give crude iron which is further processed to steel.

The leading direct reduction technologies are the Midrex, the HyL I, and the HyL III process with 90 % of the total capacity [376]. The hydrogen content of the reducing gas is ca. 40 – 65 vol % (Midrex, shaft furnace) and 75 vol % (HyL III, retorts). To fully utilize the reduction potential of the gas, carbon dioxide and water vapor are removed and the gas is recycled.

The use of *pure hydrogen* has advantages with respect to the reaction time, the degree of reduction and the texture of the reduced pellets [377], but the carbidizing reaction necessary for steel production cannot take place, so that reduction with pure hydrogen has not been able to establish itself.

Nonferrous Metallurgy. Hydrogen is employed as reducing agent and as utility in some powder metallurgy production processes. Table 42 shows the use of hydrogen during the production and handling

of various nonferrous metals.

Table 42. Use of hydrogen in the nonferrous metallurgy

Metal	Unit operation	Product
Copper	reduction of copper salt solutions under pressure	Cu powder
Nickel	selective reduction during cobalt production	Ni powder
Cobalt	reduction of aqueous cobalt salt solutions under pressure (4 MPa, 175 °C)	Co powder
Molybdenum, tungsten	reduction of the oxides or molybdates and tungstenates	Mo, W powder
Tantalum	reduction of tantalum chloride, $TaCl_5$, in hydrogen plasma	Ta hydride, Ta powder
Germanium	reduction of germanium tetroxide, GeO_4, at 650 °C	Ge powder for further processing in zone melting
Uranium	reduction of the higher uranium oxides at 650 °C	UO_2

For recovery of copper from its sulfidic ores reduction with hydrogen in the presence of calcium oxide has been suggested [378]. The thermodynamically unfavorable position of the hydrogen reduction reaction on metal sulfides is improved by the removal of the developing hydrogen sulfide (as CaS) from the equilibrium mixture.

8.3. Other Uses

Use of the High Temperature of the Oxyhydrogen Flame. The combustion of a stoichiometrical hydrogen – oxygen mixture leads to flame temperatures in the range of 3000 – 3500 K. Such flames can be used for:

Cutting and autogenous welding in the metal industry.
Growing synthetic crystals (saphires, rubies) by melting aluminum oxide doped with the appropiate metal oxides.
Production and handling of quartz glass (melting of rock crystals).

A process for the production of olefins from hydrocarbons is described [379]. The hydrocarbon to be cracked is atomized and fed into a 1000 – 2000 °C hot gas stream of hydrogen and steam, which is produced in a prereactor by combustion of excess hydrogen in oxygen. High olefin yields are reported to be possible even from heavy hydrocarbons.

Hydrogen Plasma as a Heating Agent. A further temperature increase by using the free energy of the hydrogen combustion reaction is hindered by strongly endothermic processes, such as the onset of molecule dissociation and ionization of the atoms. Higher temperatures can be reached in a plasma gas produced by means of an electric arc or a high-frequency field. Hydrogen dissociates around 5000 K into hydrogen-atoms, which start to ionize to H^+ at 9000 K; the recombination of ions to atoms or of atoms to molecules produces a great amount of heat. The heat content of hydrogen plasma is used in some processes for the production of acetylene from hydrocarbons (see Section Production from Coal and Hydrocarbons; ⟶ Acetylene – Production from Coal (Arc Coal Process)). Hydrogen plasma has advantages compared to ordinary electric arc pyrolysis, for example, much lower carbon formation.

Hydrogen in Metal Processing. The high temperature of the plasma stream is used for plasma welding and cutting. The addition of hydrogen to plasma argon (2 – 15 vol %) increases the temperature and the cutting velocity. This process is used principally for cutting or welding high-alloy steels. Forming gas, a gas mixture of 8 – 20 % hydrogen and nitrogen, prevents the oxidation at the far side of the weld (root cut), e.g., when pipes are being welded.

In the thermal atomization technique for the production of surface layers, a plasma jet is used to spray high-temperature melting metals, their oxides, borides, etc. As plasma gas hydrogen, hydrogen – nitrogen or hydrogen – argon mixtures can be used.

Heat treatment of metals (annealing, sintering, melting) requires the presence of an inert atmosphere. A mixture of nitrogen with hydrogen is usually used. Hydrogen alone can decrease the annealing time by 30 % because of its higher heat conductivity and the shorter reaction times for the reduction of oxide layers [380]. Mixtures of nitrogen, hydrocarbons, and hydrogen are used for nitriding and carbonizing annealing.

During the production of plate glass by the float glass process, a protective nitrogen – hydrogen atmosphere is employed to keep the metal melt surface free of oxides.

Semiconductor Technology. For the production of doped semiconductors, traces of the elements in the form of the corresponding hydrides (SiH_4, AsH_3, GeH_4) are mixed with a high-purity hydrogen stream in the required concentration and deposited by thermal decomposition on a silicon carrier.

Water Treatment. In potable water, unduly high nitrate contents can be lowered by denitrification with hydrogen-oxidizing bacteria in bioreactors. Hydrogen, dissolved at 0.4 – 0.6 MPa, is used as an

energy supply for the biomass [381]. For the treatment of boiler feed and process water, dissolved oxygen can be removed by conversion with hydrogen in the presence of palladium-doped exchange resins.

Other Uses. Hydrogen is also used for the following purposes: as alloying element in metals to produce required chemical properties, and to increase the transition temperature of superconducting alloys, as well as carrier and fuel gas in gas chromatography (flame ionization detector). Liquid hydrogen is used as a refrigerant, e.g., in bubble chambers [382] and for cooling of superconducting metals below the transition temperatures.

[Top of Page]

9. Hydrogen Energy

Rev. WILLIAM CECIL of Cambridge, England, is the first person reported to have suggested using hydrogen in an engine in 1820. In 1874, JULES VERNE described in one of his last books “the fuel to be used, once coal, ran out . . . – Water . . . but decomposed into its primitive elements . . . and decomposed, doubtless, by electricity.” Unfortunately, he did not disclose any details on the nature of the primary energy source to split the water.

In the 1920s and the 1930s, a great number of scientists, especially J.B.S. HALDANE in Scotland and R. ERREN in Germany, described virtually every method of production, storage and application mentioned so far in today's hydrogen energy concept, e.g., electricity generation using wind turbines, electrolysis, hydrogen storage in liquid form, and utilization in fuel cells and aircrafts. More on the story of hydrogen energy in general can be found in [383].

9.1. Aims of the Hydrogen Energy Economy

Hydrogen is an energy carrier, an indirect source of energy, not a resource itself. In spite of this, hydrogen was at first foreseen as a substitute energy form, in particular to substitute coal, natural gas, oil, and any products derived from them.

The reserves of fossil energy, in particular oil, are limited. This led to the belief that hydrogen could soon be used economically as a substitute energy form. The economic viability of the energy carrier hydrogen is enhanced by price advantage in transport over large distances (see Section Hydrogen Energy Economics) and the possibility of energy storage (see Section Hydrogen-Energy Conversion Systems).

According to Figure 14, hydrogen for use as a universal energy carrier can be generated by using all possible primary energy sources. Coal, nuclear energy including fusion (although still undeveloped), and solar energy have been named.

In addition to economic reasons for the introduction of new energy sources and new energy carriers, ecological reasons are becoming increasingly important, for example:

The increasing carbon dioxide content of the atmosphere, caused by the use of fossil energy sources (Fig. 101) and other human activities such as destruction of the tropical rain forests are predicted to lead to extensive climatic changes, e.g., a slow warming up of the atmosphere and the oceans, and melting of pole caps [384], [386]. Because of the natural fluctuations in the climate this theory has not been proven beyond doubt. In particular, little is known about compensation effects.

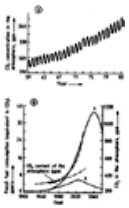

Figure 101. Actual and projected increase of the CO_2 content of the atmosphere, caused by human action, especially burning of hydrocarbons, coal, and wood

A) Measurements in Hawaii [385] from 1959 to 1983; B) Some projections extrapolating two scenarios of energy use

a) Maximum use; b) Optimum conservation of natural resources [384]

Well-known scientists have been demanding for years (without success) that a “low-risk strategy” for the use of fossil resources should be followed [387]. The utilization of the energy carriers coal, oil (in the form of gasoline, diesel, kerosene, etc.) and natural gas leads to emissions of sulfur dioxide, nitrogen oxides and hydrocarbons. In Europe, this has already led to the destruction of extensive wooded areas and to damage of the rest of the forests. The use of hydrogen as an energy carrier coupled with suitable utilization techniques could become particularly important for environmental protection.

Hydrogen competes with the conventional energy carriers hydrocarbons (methane, LPG, gasoline, etc.), coal, and electric power.

Further energy carriers are those recovered from regenerable sources and from refuse: biogas, alcohols, and vegetable oils.

For the suitability of an energy carrier, the following aspects are important:

- conversion efficiency of the primary energy carrier to the final use,

availability,
ease of storage,
safety, and
ecological and economic evaluation

The evaluation of the energy carriers for specific applications is given in Section Hydrogen-Energy Conversion Systems, the safety aspect is treated in Section Safety Aspects.

9.2. Production Systems for Hydrogen Energy

For use as a chemical reactant hydrogen is produced almost exclusively from hydrocarbons (see Table 11), because integration into the production processes and resulting overall plant economics are determining. A future hydrogen energy concept should be based on forms of energy which influence the environment as little as possible. It should offer overall economic advantages and should be socially acceptable. This is the case with solar energy, used (1) directly in photovoltaic or solarthermic plants or in direct hydrogen production, (2) climatically induced energy gradients, e.g., wind, waves, heat, (3) indirectly via production of biomass (carbohydrates, plant oil, ethanol, etc.). As an intermediate solution for a limited transition period a combined utilization of fossil primary energy, nuclear energy and regenerable energy forms have been suggested.

Figure 14 gives a review of the principal possibilities of generating hydrogen from the primary energy sources solar energy, fossil fuels, nuclear and geothermal energy. In hydrogen energy scenarios, the economical and ecological aspects of the conversion of primary energy to the required energy form heat, power, light, etc., using hydrogen as energy carrier is investigated from the point of view of present technologies.

9.2.1. *Energy Sources*

Direct Solar Energy. Solar radiation can be directly used to produce hydrogen (see Section Other Methods for the Cleavage of Water) by means of photochemical, photobiological or photoelectrochemical effects.

These methods, however, have up to now been only able to produce small amounts on a laboratory scale. The catalysts and the biomass tested have a short lifespan. Developments, especially in the area of basic research, are being followed up.

Thus, the short-term realization of projected hydrogen production systems is based on the indirect utilization of solar energy using photovoltaic cells (⟶ Photovoltaic Cells) with which electricity or, more exactly, electric voltage is produced. In addition, solar-thermal power stations, in which utilizable energy in the form of heat is developed, will become operable. In particular, the photovoltaic technology is already intensively utilized. Not only plants for power production, but also pilot hydrogen producing systems are already in operation or under construction.

Utilization of solar energy for photovoltaic power production is possible all over the world. The energy utilization factor in countries with high sunshine intensity (e.g., Sahara) is ca. 2.5 times higher than in the temperate climate zones (mid Europe) [388].

Those areas which are considered to be very good sites have a radiation energy > 2300 kWh/m^2 per year. If not only the high sunshine intensity, but also the geological suitability for setting up solar cell modules is considered, then ca. 6×10^5 km^2 are available worldwide; this is less than 0.5 % of the surface area of the globe. This area alone would be sufficient at a hydrogen production equivalent to 7000 tce/km^2 (tce = tons of coal equivalent, 1 tce $\hat{=}$ 29.31×10^9 J) per year, to cover the world energy consumption [389]. But also other concepts which do not concentrate as much on the areas with high radiation intensity, but on installation of solar cells close to the user on free surfaces, e.g., the roofs of houses, are said to be sufficient for hydrogen production [390].

Indirect Solar Energy. Solar radiation leads indirectly to energy potentials on the earth; these have been used by mankind for thousands of years to produce energy.

Water power is utilized at sites distant from private industrial consumers to produce hydrogen for ammonia production (Norway, Egypt, India, etc.). Hydrogen is produced by means of alkaline electrolysis (see Section Principles). An overview of the worldwide utilization of water power is given in [391] and, with respect to the utilization for hydrogen production, in [392]. According to this, the potential hydroenergy of 14 700 TWh/a worldwide is utilized at present to ca. 15 %. However, the greatest unused potential lies in zones far away from population and industry, such as the Arctic, Greenland, Alaska, Sibiria, or in the equator zones (Congo) [393]. About 30 % of the additional potential could be used for hydrogen production [392]. On the other hand, too great a utilization of the hydroenergy potential meets with ecological disapproval because of the destruction of the landscape by dam-building, flooding of large areas, destruction of forrests, etc.

Wind energy is used on islands to satisfy local energy requirements and to feed power into the grid. Pilot plants for the production of hydrogen already exist (see Section Future Developments). The coastal regions are particularly suitable for the generation of wind energy. Areas with particularly high wind speeds and low energy requirements which could be used for overregional production of hydrogen energy are Patagonia (South America), Somalia (Africa), and the Southwest Australian coast.

Biomass, Refuse, Wastewater. Hydrogen can be recovered from these substances by various methods.

The direct production of hydrogen by means of bacteria is possible; yields of up to 20 % have been attained (see Section Hydrogen Formation in Biological Systems). Other possibilities are the biological production of alcohols or methane from biomass or refuse by fermentation or other anaerobic processes. These substances can be converted by the methods described in Section Production from Coal and Hydrocarbons into hydrogen. Synthesis gases or hydrogen can be produced from biomass and refuse by pyrolysis or gasification processes analogous to those described for fossil feedstocks (see Section Production from Coal and Hydrocarbons).

The hydrogen production potential from biomass, refuse, or wastewater is not considered to be very high for various reasons. Firstly, other valuable products and energy carriers, such as edible products, alcohol and methane, can be produced directly from the substances. Secondly, these sources are usually widely distributed, so that for a central industrial hydrogen economy the costs for transport and logistics would be very high.

Nuclear Energy and Other Energy Forms. Nuclear energy has also been suggested for the production of hydrogen [394], [395]. Various types of conventional and breeding reactors can supply power for hydrogen production, whereas the high-temperature reactor supplies heat for chemical reactions. The direct production of hydrogen by means of radiolysis (see Section Thermolytic and Radiolytic Processes) is an undesired accompanying phenomenon in pressure and boiling water reactors.

Other energy forms, such as *geothermal* or *tidal energy*, play a marginal role in the energy production. They need not be considered for the hydrogen energy concept.

A combined utilization of nuclear energy and fossil, nonconventional resources has been sugggested for the tar sands in Canada, for oil shale deposits in the United States, and for the heavy oil found in Venezuela. Hydrogen and steam are thereby cogenerated. Steam is used for the recovery of hydrocarbons and for processing, hydrogen is used for the hydrogenation of heavy oil.

9.2.2. *Production Methods for the Energy Carrier Hydrogen*

The conventional methods described in Sections Production from Coal and Hydrocarbons and Electrolysis as well as Chapter Purification of Hydrogen are available for the purification and production of hydrogen for the hydrogen energy scenario.

Purification Methods. The purification requirements for the hydrogen energy economy depend not so much on the planned utilization as on the hydrogen storage technique and transport (vectorization). The highest purity is necessary if hydrogen is liquefied, because all impurities, apart from helium, are deposited as solids and could block the liquefication equipment.

Storage and transportation in metal hydrides also place high requirements on the hydrogen purity, because most impurities lead to irreversible destruction of the storage material. Purity requirements are less for the direct combustion of hydrogen and the reaction with oxygen or air in fuel cells. All purity requirements can be met using conventional technologies, particularly adsorption (see Section Adsorption Processes).

Electrolysis. The most important hydrogen production method for the hydrogen energy scenario is electrolysis (see Section Electrolysis). Here, the energy carrier electricity is converted to the energy carrier hydrogen. Conventional electrolyses have a conversion efficiency of 80 – 85 %. Improved electrolyses are expected to have an efficiency of 88 – 94 % [396], which can hardly be increased further. In the high-temperature electrolyses at present under development, however, part of the energy necessary for the electrolysis can be supplied as thermal energy and it can therefore produce hydrogen with a higher efficiency than with solely an electrical supply.

The combination of regenerable energy forms, such as photovoltaic or wind energy with electrolysis units has undergone a long-term test in pilot projects (see Section Future Developments). In addition, detailed system analyses have been made [397]:

electrolysis and thermal solar power plants (power generator can be driven by steam or a Stirling motor),
electrolysis and thermal solar power plants, coupled with high-temperature electrolysis [398],
electrolysis and wind energy,
electrolysis and photovoltaic plants.

Particular problems with the electrolysis are caused by the fluctuating operation of the solar plant cycle at various times of the day as well as weather-dependent fluctuations of solar and wind energy.

Typical problems caused by *rapid* load changes:

degradation and destruction of diaphragm because of local overload,
corrosion caused by current leakages, and
deactivation of the electrode catalysts

Problems caused by *slower* (daily) changes:

cooling of the electrolysis module,
mixing of gas and electrolyte,
pressure equalization between cathode and anode, and
start-up and shut-down procedures including eventual prepolarization or blanketing.

Additional Methods of Hydrogen Production. Additional methods are:

1. Hydrogen production by means of thermochemical cycle processes using thermal energy (solar or nuclear energy)
2. Hydrogen production by *gasification* of refuse or biomass (for process principles see coal or heavy oil gasification, Section Gasification of Coal and Hydrocarbons)

3. Hydrogen production by *pyrolysis* of refuse or biomass (for gas purification see coke oven gas treatment, Section Coke Oven Gas).

For technical and economic reasons the aforementioned methods play only a small part in the production of hydrogen for energy systems.

9.3. Hydrogen-Energy Conversion Systems

For utilizing the energy contained in the secondary energy carrier hydrogen, numerous techniques are available, in the area of energy technology in industry and household, in power stations and in the transportation sector. Most of these are based on principles which have been known for a long time. Section Hydrogen-Energy Conversion Systems gives an overview of the state of the art and of present developments world-wide.

9.3.1. *New Developments for Hydrogen Storage*

The possibility to store hydrogen amounts to the possibility to store energy; this is the particular advantage of a secondary energy carrier. In addition to storage in the gaseous or liquid state (Section Storage) there are numerous investigations or experimental applications of new processes for hydrogen storage in solids.

9.3.1.1. *Cryoadsorption*

Great quantities of gaseous hydrogen can be adsorbed on suitable adsorbents at medium pressure and low temperature. If the adsorber is cooled to 60 – 100 K and pressurized to ca. 4 MPa, the *average* hydrogen density on the active surface increases to values comparable to the density of liquid hydrogen. The *effective* storage density is, however, much smaller because part of the volume is taken up by the adsorbent and the pressure vessel; however, liquefaction can be avoided.

The process has been investigated experimentally [399], [400]. Only materials which are light and very porous, with a high specific surface or with a specifically activated surface can be used as adsorbing agents. Operating data for a cryoadsorption storage system on the basis of activated carbon bulk density 0.35 g/cm^3 are listed below:

Temperature range for isothermal operation	65 – 70 K
Charge-discharge pressures	0.2 – 4.2 MPa
Storage density of hydrogen (75 K and 4.2 MPa)	68 g H_2 per kilogram adsorbent
Maximum hydrogen storage density (65 K and 4.2 MPa)	82 g H_2 per kilogram adsorbent
Average mass related storage capacity	50 g H_2/kg
Average volume related storage capacity	15 g H_2/L

The economic range for the technical application will presumably be limited to medium-scale storage with a specific storage capacity of 20 – 30 kWh/kW.

9.3.1.2. *Hydride Technology*

The capability of metals or metal alloys to reversibly store hydrogen in the form of hydrides has lead to a great number of investigations with respect to technical applications. The main areas are seen in the storage of hydrogen but also in the separation of hydrogen from gas mixtures, the purification of gases, and in storage and separation of hydrogen isotopes. An overview is given [401][402][403][404].

Physical Fundamentals. Ionic and metallic hydrides are formed (exothermic process) and decomposed (endothermic process) according to

$$y\mathrm{Me} + x\mathrm{H_2} \rightleftharpoons \mathrm{Me}_y\mathrm{H}_{2x}$$

The hydrogen absorption and hydride formation behavior can be described by pressure – composition isotherms (see Fig. 102 A) as follows. After catalytic dissociation of the hydrogen molecule some of the atoms lose their electron and occupy an interstitial position in the crystal structure of the host metal (solid solution, α-phase). As the hydrogen concentration is increased, local H – H interaction causes nucleation which induces growth of the hydride phase (β-phase). The two phases ($\alpha + \beta$) coexist until, through further hydrogen uptake, the stoichiometric limit is reached. Here, the isotherms reach a plateau level (constant pressure) which is given the name ‘dissociation plateau' or ‘plateau pressure'. After the pure β-phase is reached, any further increase of the hydrogen : metal ratio requires a steep increase in hydrogen pressure. At higher hydrogen pressure further plateaus may be found because of the existence of other hydride phases (γ, δ, . . .).

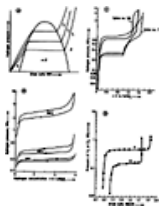

Figure 102. A) Ideal pressure-composition isotherms for a hydrogen-metal system

B) Absorption and dissociation isotherms of $LaNi_5$ hydride with hysteresis

C) Hysteresis loops of test cycles of FeTi at 25 °C. The plateau pressure of the monohydride stabilizes after ca. 20 cycles, whereas that of the dihydride formation is still rising

D) Pressure-composition isotherms of hydrogen and deuterium in zone-refined vanadium at 40 °C

On reducing the hydrogen pressure the process is run through in the reverse direction (desorption isotherm). Along the desorption plateau, the hydride phase decomposes into hydrogen-saturated metal and hydrogen gas. The decomposition process is endothermic. Absorption and desorption isotherms show a hysteresis effect. This is shown in Figure 102 B and C for $LaNi_5$ and FeTi [405], [406].

The plateau pressure *p*, which varies strongly with temperature, is connected with the formation or decomposition enthalpy by the Van't Hoff equation. The entropy term, ΔS^0, consists mainly of the standard entropy of gaseous hydrogen ($S^0[H_2^g]$ = 130.6 J mol^{-1} K^{-1}), because the entropy of the hydrogen dissolved in the metal can be neglected.

$$-RT \ln p_{H_2} = \Delta H - T\Delta S$$

$$\log p_{H_2} = \frac{\Delta H}{19.14\,(T/1000)} + 6.82$$

This relationship can be illustrated graphically for all hydrides and for the corresponding hydride transition steps (hydride isochores) (see Fig. 103). Applications can be found in a temperature range from –30 – 300 °C and at pressures up to 10 MPa.

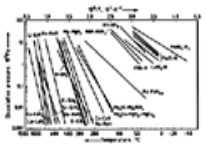

Figure 103. Hydride isochores for various metals and intermetallics. H_2-plateau dissociation pressure vs. 1/T

Mm denotes mischmetal, an inexpensive mixture of rare earths

The heat evolved by the hydrogen – metal reaction depends on the type of hydride. It can amount to ca. 30 % of the heat of combustion of the absorbed hydrogen. To liberate hydrogen, at least the same amount of energy is required (see Table 43).

Table 43. Comparison of the hydrogen absorption data of some metal – metal hydride systems

Hydride system	Heat of reaction, kJ per mole H_2	Plateau region [a]		Hydrogen content		Energy density,[c]
		pH_2, MPa	*T*, °C	wt %	g/mL [b]	kJ/g
Elemental metals						
Mg – MgH_2	77.4	0.1	286	7.7	0.107	9.9
Ti – TiH_2	124	0.1	650	4.0	0.15	5.7
V – VH_2		0.4	40	2.1	0.19	2.95
Alloys						
Mg_2Ni – Mg_2NiH_4	64.5	1.0	350	3.16	0.081	4.5
FeTi – $FeTiH_{1.95}$	31.4	0.5	30	1.72	0.096	2.5

$LaNi_5$– $LaNi_2H_{5.9}$	30.2	0.5	60	1.37	0.089	1.95
$Ti_{0.98}Zr_{0.02}V_{0.43}Fe_{0.09}Cr_{0.05}Mn_{1.5}$		1.0	24	1.75	0.09	2.5
Liquid hydrogen for comparison						
LH_2				100	0.077	142

[a] At selected conditions, dissociation pressure at 50 % of H/Me ratio.
[b] Refers to volume of compact hydride material without voids.
[c] Refers to HHV of absorbed hydrogen

Intrinsic Properties of the Hydride-Forming Materials. To analyze the utilization possibilities of hydride-forming metals, some further properties, some of which are crucial, must be known. The overall kinetics of hydride formation, e.g., is governed by the intrinsic kinetics (dissociation, surface penetration, diffusion, phase transformation), heat flow, and mass flow. In particular, surfaces of *intermetallic compounds* containing transition metals accelerate the dissociation reaction, whereas thin *oxide layers* on the surfaces of elemental metals often block this. Easy hydrogen sorption is facilitated by surface segregation with continuous formation of reactive subsurfaces (self-activation). Addition of small quantities of transition elements, e.g., nickel, will always favor activation. Impurities in the hydrogen, such as carbon monoxide, hydrogen sulfide and sulfur dioxide slow down the *sorption rate* because of surface deactivation, whereas oxidizing impurities, such as oxygen and water, gradually *decrease the capacity* by reaction with the host compound. Such “poisoning” of the hydride material can in many cases by reversed by hydrogenation – dehydrogenation cycles with pure hydrogen at higher temperature.

Diffusion is normally not the rate-determining step because most of the hydride-forming materials disintegrate into powder with a grain size down to 1 µm.

Incorporation of the hydrogen into the lattice and the α/ β-phase transition causes a deformation of the metal structure. The accompanying increase in volume can amount to 25 % and may lead to undesired distortion of the containing vessel.

Detailed analyses of the kinetics of hydride formation of $LaNi_5$ and MgNi are available [407]. For most hydride-forming systems using intermetallic compounds, heat transfer is rate-controlling. Sufficient heat transfer can be provided by adding inert metals (Cu, Al) or inert ceramic beads or by pressing the material into pellets to fit into heat exchanger tubes.

Hydride-Forming Materials. The criteria for selection of a material for a specific application are:

- high adsorption capacity
- suitable temperature and pressure range
- sufficient exchange rate
- long-term stability
- long cycle life

In the temperature range from –20 – 400 °C the following groups of alloys are known:

- – 20 – 100 °C
  - AB_2 alloys, e.g., $ZrMn_2$
  - AB_5 alloys, e.g., $LaNi_5$
  - FeTi-based alloys
- 300 – 400 °C
  - Mg and Mg-based alloys

Some metal – metal hydride systems and their important properties are summarized in Table 44. In recent years development has concentrated on cheaper multicomponent alloys and modeling materials to optimize uptake and storage parameters. About 10 different storage alloys are commercially available from producers in Europe, Japan and the United States at prices between $ 5 and 30 per kilogram.

Table 44. Hydrogen content of fuel gases and corresponding energy content

Fuel gases	H_2 content, %	HHV, kJ/m^3 (STP)	LHV, kJ/m^3 (STP)
Natural gas, rich gas	up to 12.5	25 000 – 39 800	35 800 (CH_4)

Town gas	40 – 60	16 700 – 20 000	
Coke oven gas	50 – 60	20 000 – 23 000	16 500 – 20 500
Water gas	50	10 500 – 12 200	10 700 – 11 700
Producer gas, blast furnace gas	2 – 5	4 300 – 4600	3 900 – 4700
Carbon monoxide		12 650	12 650
Hydrogen	100	12 760	10 800

Stationary Hydride Storage. For the stationary storage of hydrogen by using hydrides, FeTi or TiVFeMn alloys can be employed. A representative example for the many developments in this field is the hydride storage vessel of Gesellschaft für Hydrid- und Wasserstofftechnik (HWT). Figure 104 shows a module which consists of a bundle of pressure vessels containing the powdered metal alloy between fins. A sintered metal filter prevents the loss of powder. The gas bundle is surrounded by the heat transfer medium, water. The essential operating data are:

hydrogen capacity	1700 – 2000 m^3 (STP)
active storage material	10 t
operating pressure	5 MPa
operating temperature	80 °C

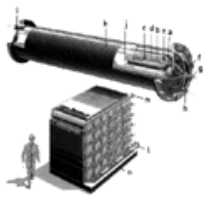

Figure 104. Hydride storage tank

a) Storage tube; b) H_2-guide tube; c) Lamella; d) Hydride; e) Filter (sintered metal); f) Gas-collecting tube; g) H_2 connection; h) Water inlet (module); i) Water outlet (module); j) Filling body; k) Shell; l) Water inlet (centr.); m) H_2 inlet/outlet (centr.); n) Container

Hydrogen used to fill the vessel should have at least 3.0 (99.9 %) purity. Hydrogen with a purity of 6.0 (max 1 ppm [mL/m^3] impurity) can be drawn off [408].

Mobile Hydride Storage. Possible applications for the hydride storage technology are mobile storage vessels for the laboratory or for hydrogen-fueled cars (see Section Hydrogen Fuel Systems). Figure 105 shows a hydrogen laboratory storage vessel and the corresponding ab- and desorption curves. These vessels are available for hydrogen quantities of 1 – 5 m^3 [409]. In addition to the small space requirements, they offer further advantages such as safe handling, purification effect and availability at higher pressure levels.

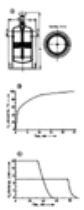

Figure 105. Hydrogen laboratory storage vessel and the pertaining adsorption and desorption curves

A) Schematic of a storage vessel; B) Loading at 0.5 MPa, water cooling at 20 °C; C) Unloading against 0.1 MPa, heat exchange medium is water at 80 °C

The energy density of hydride storage systems, including the vessel material, is in the range of 300 Wh/kg for the so-called low-temperature hydrides.

Separation of Hydrogen Isotopes via Hydride Formation. Some metals and alloys (TiNi, Ti_2Ni, V) show considerable differences in the plateau pressure of their hydrides and deuterides (see Fig. 102 D). This can be used for the enrichment of deuterium in deuterium-containing hydrogen streams [410]. The recombination of H and D atoms can lead to the formation of HD. The separation of tritium over vanadium monohydride in a temperature-swing process has been described [411].

Purification of Gases. Hydride-forming materials can be used to purify hydrogen as well as noble gases. Various mechanisms are thereby involved: oxygen, carbon dioxide, and water are removed by

the hydride material because of its getter properties. Carbon monoxide is converted to methane, which is more easily removed, whereas inert gas components, such as the noble gases, can be removed by a short purging step. The hydrogen, which can then be recovered from the hydride bed, has a purity of >7.0, i.e., impurities content <0.1 ppm (mL/m^3) (see Section Quality Specifications).

Hydrogen Recovery. The selective removal of hydrogen by means of hydride-forming metals can be used to separate hydrogen from a gas mixture. The feasibility of this method for recovery of hydrogen from the purge gas stream of an ammonia synthesis loop was demonstrated in a pilot plant. The unit operated with three absorber columns and produced hydrogen with a purity of 99 % and a recovery rate of 90 – 93 % of the hydrogen contained in the purge gas [412]. The process is adiabatic, the heat evolved during absorption is stored in the $LaNi_5$ hydride pellets, which contain a certain amount of ballast material. This stored heat is used for hydrogen desorption. The hydrogen uptake and liberation is pressure-controlled (adiabatic pressure-swing process). The complex dependency of the hydride bed lifetime on alloy type, trace components in the gas, poisons, and various combinations thereof has prevented commercial utilization until now [413].

Thermochemical Machines. Metal hydrides are promising materials for use in thermochemical machines such as heat pumps, heat transformers, refrigerators or hydrogen compressors. The principle is based on thermodynamic cycles in a closed system; a heat pump cycle is shown as an example in Figure 106. The equilibrium lines in the ln p – 1/T diagram are based on measurements of pressure – composition isotherms. The required heat output is assumed to be at 40 °C. Taking into account the different material imperfections (i.e., hysteresis, plateau slope) and the pressure drop necessary to drive the reaction, the level of the low-temperature heat source changes from 2.5 °C (ideal case) to 16.5 °C, while the level of the high-temperature source rises from 82 °C to 103 °C. An advantage is that hydrogen is handled in a closed loop without the risk of contamination. More than 10 developments are being followed up worldwide (United States, Federal Republic of Germany, Japan). The planned applications are in household heating, air-conditioning, refrigeration and power generation. The investigations concern reaction bed designs, flow of the heating or cooling media, heat transfer, etc., using fast-reacting alloys of $LaNi_5$, $LaNi_{4.9}Al_{0.1}$, FeTi for the low-temperature range and $CaNi_5$, Mg_2Ni, $LaNi_{4.7}Al_{0.3}$ for the high-temperature range [414].

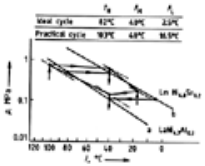

Figure 106. Thermodynamic cycle of a heat pump

a, b) Mean values of equilibrium lines for absorption and desorption

Dashed lines: pressures at end of plateau

Vertical arrows: inflow or outflow of heat

Horizontal arrows: transfer of hydrogen

($LmNi_{4.8}$ is a lanthanum-rich mischmetal)

Compression of Hydrogen. Hydrogen can be compressed thermally by absorption at low and desorption at higher temperature. A machine working on this principle as an open system with several stages is commercially available [415].

9.3.1.3. *Liquid Organic Hydrogen Carriers*

With the help of hydrogenation – dehydrogenation reactions it is possible to store hydrogen reversibly in a chemical compound. The combinations benzene – cyclohexane and toluene –methylcyclohexane have been investigated using noble metal – aluminum oxide carrier catalysts for the reversible reaction. Transportable hydrogenation – dehydrogenation systems operating at 380 – 500 °C and 1.1 MPa have been designed for a truck using hydrogen fuel; the necessary energy for the exothermic dehydrogenation reaction is taken from the exhaust gas of the engine. A specific energy density of 56 g hydrogen per liter liquid hydrocarbon is attained, but the weight of the necessary equipment must be taken into account [416].

The reversible uptake of hydrogen in liquid organic substances has also been suggested for transcontinental hydrogen transport. Existing facilities for gasoline could be used [417].

9.3.2. *Combustion and Heating*

The energy contained in the hydrogen can be released most simply by combustion with oxygen or with air. Characteristic combustion data are given in Chapter Properties.

Hydrogen Content of Fuel Gases. Hydrogen mixtures have long been used in the form of town gas for heating purposes. Table 44 shows the hydrogen content of some fuel gases. The difference between LHV and HHV (condensation enthalpy of water formed during combustion) is greatest for pure hydrogen, because water is formed as the only product.

Emissions. Since only water is obtained as a product of the combustion, emissions, such as carbon monoxide, sulfur dioxide, incompletely combusted hydrocarbons, particles, soot, ash, which are obtained by combustion of fossil fuels, are avoided (see Section Production Systems for Hydrogen Energy for a discussion of the carbon dioxide problem). If hydrogen is burned in the presence of air, the high temperatures favor the formation of nitrogen oxides. The NO_x formation depends on the flame temperature and can be influenced by the value of excess air, λ. Figure 107 shows typical NO concentrations for a hydrogen – air flame. Maximum NO formation occurs in the range of the stoichiometric hydrogen – oxygen ratio, whereas under lean conditions (excess air) low nitric oxide concentrations may be attained [418].

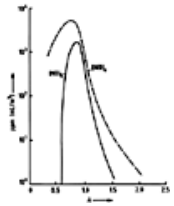

Figure 107. NO concentration resulting from hydrogen – air flames as a function of excess air value λ

($[NO]_e$ = equilibrium concentration; $[NO]_5$ = concentration behind flame front within 5 ms; p = 3.2 MPa)

Hydrogen Burners. Differences in combustion properties, i.e., flame speed and flame stability as well as density and heating value, make it necessary to use a different burner geometry when hydrogen or hydrogen-rich gases are to be burned. The flame speed of hydrogen – air mixtures is seven times higher than that of methane; the particular problem is therefore the flash-back over a wide range of hydrogen – primary air mixtures.

The most important criteria for household burners are summarized, e.g., in [419]. The ratio of secondary to primary air must be increased, if hydrogen is used. Forced draft burners for heating systems can be retrofitted by removal of the baffle plate and by shortening the ignition electrodes [420]. Industrial hydrogen burners are used for heat, steam, and electrical power production. The state of the art are premix burners or nozzle-mix burners with high mixture velocities. For supplying heat in certain industries (semiconductors, glass) the simultaneous presence of a reducing atmosphere is advantageous.

Catalytic Combustion. The thermal energy of hydrogen can also be released by means of flameless catalytic combustion. Suitable catalysts, apart from platinum and palladium which initiate reaction at room temperatures, are metal oxides, such as MnO_2 – CuO – Ag_2O (Hopcalite) or Co_3O_4. The surface can be enlarged accordingly by a suitable carrier material in the form of porous plates or honeycombs. Catalytic burners operate either on the diffusion principle or by premixing hydrogen with air. Figure 108 shows a catalytic diffusion burner. An overview of Japanese investigations is given [421].

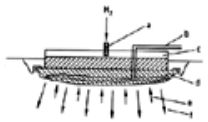

Figure 108. Catalytic diffusion burner

a) H_2 entrance port; b) Electric ignition wire; c) Distribution manifold; d) Catalytic combustion zone; e) Oxygen from air; f) Heat radiation

Catalytic heaters react very flexibly; the temperature can be adjusted and limited by means of the throughput; the heat is produced in the infrared range. The combustion is free of emissions. Since no stacks are required, almost 100 % of the heat set free can be utilized. In addition, the moisture content of the air can be regulated by the reaction product, water.

Transport to Small Consumers. Hydrogen could be distributed via present nets to industrial and household consumers and could substitute natural gas. The energy transport capacity of such a low-pressure network (ca. 10 kPa) leads to about the same heat flux for hydrogen as for natural gas.

9.3.3. *Hydrogen Fuel Systems*

In all sectors of transportation (space and aircraft propulsion, earthbound traffic systems including individual transport) the fuel systems hydrogen – air or hydrogen – oxygen can be used for the prime mover.

9.3.3.1. *Hydrogen Propulsion Systems for Space*

Because of the large specific enthalpy of combustion LH_2, combined with liquid oxygen (LOX), produces (apart from the system LH_2 – LF_2) the greatest specific impulse. Jet velocities up to 4500 m/s can be reached using combustion chamber pressures of 5 to 22 MPa.

The hydrogen – oxygen motor NASA Centaur (Pratt & Whitney) was used for the third stage of the Atlas Centaur and the fourth stage of the Titan III rocket. For the Apollo project Rocketdyne motors were used for the second and third stage of the Saturn V [422]. The peak of the US development is the Space Shuttle main engine, a high-pressure motor with a combustion chamber pressure of 21.5 MPa.

The European Ariane carrier system uses LH_2 – LOX in its third stage. Using 10 t of this fuel combination, the HM7 motor develops a thrust of 60 kN over a period of 700 s. A further development for 800 kN thrust and a fuel requirement of 120 t with the designation HM60 is being developed [423].

9.3.3.2. *Hydrogen as Aviation Fuel*

In the future the following alternative aviation fuels will be available:

- jet fuel “aviation grade” from coal (Synjet)

liquid methane (LCH_4)
liquid hydrogen (LH_2)

Liquid hydrogen has already been tested in 1957 in a B 57 Canberra by the NACA Lewis Research Center (Cleveland, Ohio) and its suitability as an aviation fuel was proven. Since 1973, especially in the United States, investigation of the use of LH_2 as an aviation fuel has been follwed up (overviews [424-426]). In 1988, test flights were made in the Soviet Union with a civil aircraft using LH_2 as fuel.

Hydrogen is suitable for turbines in the sub- and supersonic regions. Engine designs for turbines converted to hydrogen fuel have certain advantages: They are smaller, easier to control and have a longer lifetime than normal turbine fuels. The cryogenic fuel is used to cool engine oil and turbine cooling air, whereas the turbine core exhaust heats up the fuel before injection into the combustor. The savings in the specific fuel consumption amount to 8 %. Ecological benefits are the lower noise level, smaller emission of NO_x, and absence of CO, CO_2, and hydrocarbons.

The energy content related to volume and mass of LH_2 is the determining factor for the construction of a hydrogen aircraft. The four times greater fuel volume as compared to hydrocarbons for the same energy content and the cryogenic properties require a completely new concept for the aircraft fuselage and tanks. To minimize surface – volume ratio, the insulated tanks are placed both fore and aft of the passenger compartment. Weight advantages are gained by tank configurations which are integrated in the aircraft fuselage structure. For the complicated aircraft fuel system three absolutely equivalent fuel booster pumps per tank compartment are foreseen [427].

The 2.8 times higher energy content of LH_2 compared to kerosene (related to mass) makes it possible to have a much smaller lift-off weight for the same range and capacity. A comparison of design data and various performance characteristics of subsonic aircrafts for the fuels listed is shown in Table 45.

Table 45. Comparison of design and performance data of alternate fueled subsonic aircraft (basis: 400 passengers, 44 000 kg payload, 5500 NMi range, 0.85 Mach cruising speed)

	Liquid hydrogen (LH_2)	Liquid methane (LCH_4)	Synjet
Gross mass, kg	168 830	225 570	232 060
Block fuel, kg	21 620	58 980	72 530
Operating empty mass, kg	103 300	116 170	107 370
Wing area, m^2	296.8	385	380.3
Span, m	51.8	58.9	58.5
Fuselage length, m	65.7	61.4	60
Gross mass, kg	156 130		
Block fuel weight*, kg	15 270		
Specific fuel consumption (cruise), kg h^{-1} N^{-1}	0.206	0.504	0.615
Lift/drag (cruise)	17.4	19.21	19.13
Thrust per engine, N	135 000	177 000	185 000
Energy utilization, kJ $seat^{-1}$ km^{-1}	636	723	759

* with respect to laminar flow control.

Similar studies have been presented for LH_2-fueled supersonic aircraft. Here the immense fuel requirements are even more in favor of LH_2. A comparison shows a reduction in the lift-off weight by ca. 50 %.

Laminar Flow Control (LFC). The cryo-fuel LH_2 makes it possible to use the so-called *cryogenic wall cooling* technology. By cooling major aerodynamic surfaces of the aircraft to ca. 150 K, laminar flow in boundary layers is stabilized with a consequent reduction of skin-friction drag. The necessary refrigeration can be taken from the evaporating LH_2 directly or via a secondary refrigeration system (e.g.,

LN_2). Overall drag reduction in the order of 20 % leads to a fuel saving of about the same amount at cruising speeds [428].

Airport requirements for LH_2 as fuel have been investigated in different studies. The facilities include the liquefaction plant, LH_2 storage capacity and LH_2 fuel handling systems [429], [430]. Because of planning and development issues, the use of LH_2 as an alternative fuel for aviation will not occur before the next two aircraft generations, i.e., about 2030.

9.3.3.3. *Hydrogen for Automotive Vehicle Transport*

Mobile Otto and Diesel internal-combustion engines for vehicles can be operated on hydrogen or hydrogen – gasoline mixtures. RICARDO (1924) and BURSTALL (1927) carried out some of the first investigations. In 1930, ERREN made intensive studies on hydrogen – air and hydrogen – oxygen motors. The first internal-combustion engine car in the United States fueled by hydrogen was presented by BILLINGS in 1966 in Utah. At the moment further developments are in the Federal Republic of Germany, Japan, Australia and the Soviet Union. The literature on this topic is abundant. Overviews are given in [431-433].

Modification of the Engine. Some properties of hydrogen make engine modifications necessary; low ignition energy and high flame speeds can cause self-ignition during the mixture preparation or flame flash-back. The octane number of hydrogen is much lower than that of gasoline. However, the wide ignition range allows burning of lean mixtures and gives a large control range.

Various techniques are used to prepare the feed mixture to the combustion chamber. With *external mixing* a homogeneous hydrogen – air mixture is introduced. Uncontrolled preignition or flash-back into the intake manifold is prevented by adding ballast, preferentially water, and by timed individual port injection of hydrogen close to the cylinder intake. *Internal mixture* formation is achieved by direct injection into the combustion chamber. The necessary supply pressure must be at least 1.5 MPa, or even, depending on the time of injection, more than 10.0 MPa. This is only possible by using piston pumps for liquid hydrogen. Figure 109 shows the principle layout of internal and external mixture formation.

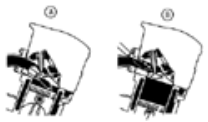

Figure 109. Internal mixture formation (A) using cryogenic hydrogen (direct injection) and external mixture formation (B) using ambient-temperature hydrogen

Attainable specific charge heating values (energy content per cylinder loading) are compared in Table 46 under various operating conditions. Revolutions per minute (rpm) and effective efficiencies are comparable to the gasoline motor. Because of the lower air quantities possible and the lower heating value of the mixture, *external mixture* formation with water injection gives power ratings comparable to that of Diesel motors, whereas with internal mixtures ca. 40 kW/L cylinder volumes can be attained [435].

Table 46. Specific charge heating values of hydrogen – air mixtures in combustion engines compared to gasoline

Operating conditions	Maximum of spec. charge HV $H_{spec.\ max}$		Maximum of efficiency $\eta_{e\ max}$		Maximum of excess air φ_{min}	
	φ	kJ/L	φ	kJ/L	φ	kJ/L
Gasoline	1	37.7	1.1	34.4	0.83	31.6
Hydrogen (external mixture formation)	1	31.9	0.4	15.5	0.2	8.3
Hydrogen (internal mixture formation)	1	45.3	0.4	18.1	0.2	9

Hydrogen-powered engines emit only water and NO_x. Numerous investigations concerning the calculation and measurement of the nitric oxide emission have been carried out, e.g., [436], [437]. In Figure 110 typical NO_x values for various modes of operation with hydrogen and gasoline are shown as a function of the equivalent ratio ($\varphi = l/\lambda$). Especially in lean modes of operation, i.e., with excess air, hydrogen motors operate more or less emission-free.

Figure 110. Typical NO_x values for various modes of operation of hydrogen (a – d) and gasoline engines

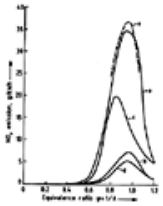

a) External mixture formation, load governing; b) External mixture formation, quality governing; c) External mixture formation, optimum attainable values; d) Internal mixture formation; e) Typical values for Otto and Diesel engines

On-Board Hydrogen Storage. At present three storage possibilities can be considered:

in the gaseous state in pressure vessels,
in the liquid state in vacuum-insulated tanks, or
chemically bound in metal hydride storage tanks.

Table 47 compares various mobile storage methods for vehicles. It is clear that none of the hydrogen storage systems have reached the undoubted advantages of the fossil energy carrier, gasoline.

Table 47. Comparison of mobile storage systems for hydrogen (basis: 45 m^3 H_2 corresponding to 15 L gasoline)

	Liquid H_2	Gaseous H_2	Metal hydride-H_2
Hydrogen content	57 L	45 m^3	4 kg
Gross storage mass, kg	20	2×60	320
Hydride material, kg			220
Maximum operating pressure, MPa	0.4	30.0	5.0
Operating temperature, °C	–253	–20 – +50	–20 – 100
Energy densities,			
kWh/kg*	6 – 7	1.1	0.4
kWh/L*	0.95	0.55	0.8

* Energy density of a gasoline tank: 8 – 8 kWh/kg (8 – 9 kWh/L).

Pressure gas storage vessels are disadvantageous because of their large weight and volume. The high storage pressure of ≥ 20.0 MPa raises safety and supply questions.

Liquid hydrogen tanks for vehicles are larger than gasoline tanks and require a complicated construction because of the low storage temperature. Characteristic data for a tank in the boot of a car (BMW-DFVLR experimental vehicle) are [434]:

double-walled cryo-vessel for 45 L of hydrogen,
vacuum superinsulation cooled by means of evaporating hydrogen,
mass (filled) of 45 kg,
operating pressure 0.2 MPa, lowest operation temperature –253 °C,
evaporation losses <2 %,
hydrogen delivery pressure via LH_2 piston pumps 1.5 – 2.0 MPa.

Problems exist concerning the evaporation rate, safety questions and tanking techniques. Tanking of the test vehicles occurred at a computer-controlled, half-automatic LH_2 station, so that maloperation was avoided. The tanking time was ca. 5 min.

Mobile hydrogen *hydride storage vessels* are used for example in Daimler Benz test vehicles. For details of the hydride technology see Section Hydride Technology. Suitable hydride-forming metals are those, in which reversible hydrogen absorption occurs in the temperature range of –25 – 100 °C. These are *low-temperature* hydride materials, such as FeTi. *High-temperature* hydride materials, such as Mg_2Ni, have a higher capacity, but more energy at 350 °C is required to drive off the stored hydrogen. Figure 111 shows a hydride tank using a titanium – vanadium – manganese alloy for a vehicle.

Addition or removal of heat is performed by a heat carrier, such as a water – defrosting agent mixture connected externally during tanking or internally during operation of the vehicle. With a full storage vessel the minimum pressure of 0.2 MPa required to start the motor even at –20 °C is available [438]. High-purity hydrogen (99.999 %) must be used to avoid a passivation of the active material. A quick tanking operation with gaseous hydrogen at 5.0 MPa takes ca. 10 min.

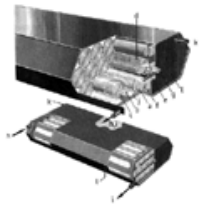

Figure 111. Hydride storage tank for motor vehicle drives of HWT

a) Storage tube; b) H_2-guide tube; c) Lamella; d) Hydride; e) Filter (sintered metal); f) Gas-collecting tube; g) H_2 connection; h) Water inlet; i) Water outlet; j) Filling body; k) Vent valve connection; l) Shell

The motor concept chosen for a vehicle depends on the method of storing the hydrogen on board. Low-pressure gaseous hydrogen allows only external mixture formation, whereas liquid hydrogen, because of its cryogenic properties, is more flexible (liquid compression, cooling, injection).

State of the Art. Representative for the many investigations being carried out in several countries, two development projects in the Federal Republic of Germany are described:

Between 1984 and 1986 Daimler Benz tested a fleet of 10 hydrogen vehicles under normal operating conditions in Berlin. Delivery vans (Fig. 112) were equipped with a 2.3 liter spark-ignition engine with external mixture formation and water injection and a power of ca. 77 kW. Hydrogen storage was by four low-temperature, liquid-heated hydride storage units with a total mass of 560 kg and a hydrogen capacity of 66 m^3. This was sufficient for a range of 120 km in city traffic. The total mileage for the entire fleet amounted to 160 000 km [439].
Deutsche Forschungsanstalt für Luft- und Raumfahrt (DFVLR) has been working with BMW since 1979 on vehicles with LH_2 tanks and different approaches for the mixture formation. The present state of development is the 745i Sedan with a 3.5 liter spark ignition engine with a so-called “late internal mixture formation,” i.e., hydrogen injection during the compression stroke. The power output is claimed to be ca. 150 kW [440].

Figure 112. Daimler-Benz delivery van with modified Otto engine for hydrogen as fuel and onboard hydride hydrogen storage

a) Refueling connectors; b) Hydride hydrogen storage; c) Filter; d) Pressure-reducing device; e) Magnetic shut-off valve; f) Hydrogen metering device; g) Differential pressure controller; h) Hydrogen distributor; i) Inlet manifold; j) Hydrogen injection nozzles; k) Recirculating pump; l) Exhaust gas – water heat exchanger; m) Exhaust flaps; n) Storage for engine-water injection system; o) Water feed pump; p) Filter; q) Water metering device; r) Water injection nozzles; s) Electronic control unit

9.3.4. *Electricity Generation from Hydrogen*

Apart from the direct utilization of hydrogen for heating purposes (see Section Combustion and Heating) the generation of electricity is important in the hydrogen energy concept. This can be done indirectly via combined heat – power processes or directly by electrochemical means.

9.3.4.1. *Heat – Power Processes*

Suitable heat – power processes for power generation from hydrogen are [441]:

1. conventional steam turbine processes with hydrogen-fueled boilers,
2. open hydrogen – air gas turbine processes, and
3. hydrogen – oxygen steam turbine processes

The thermal efficiency of various hydrogen-fired power station processes with up to 100 MW power is given in Figure 113 as a function of the range of turbine-entrance temperature t_E. A hydrogen – oxygen power station with condensation attains an efficiency of 52 % at full load. This type of power station (line a) is most suitable for peak-load power production. For immediate increase of power output of fossil-fired ground-load or medium-load power stations a hydrogen – oxygen steam production unit can be integrated (see Fig. 113, line b instantaneous reserve plant).

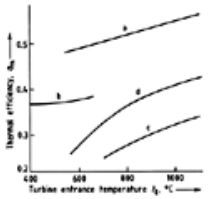

Figure 113. Thermal efficiency of various hydrogen-fired power station processes

a) Hydrogen-oxygen power station; b) Conventional hydrogen power station; c) Hydrogen power station with gas turbine; d) Hydrogen power station with gas turbine, improved design

A new component suggested for power stations is the hydrogen – oxygen steam producer. An experimental version for a thermal capacity of 15 – 40 MW exists at the DVFLR (Fed. Rep. of Germany). A schematic is shown in Figure 114. Hydrogen and oxygen are introduced through a suitable lance and equally distributed in the combustion chamber before burning. The combustion product, the temperature of which is 3000 °C, is cooled down to the required steam temperature by water quenching and fed to the turbine. It is particularly important to maintain the stoichiometric ratio of the fuel mixture in order to avoid noncondensable gas components (excess oxygen or hydrogen). This is achieved by means of a sophisticated control strategy. Starting from zero, full load can be reached within 1 s (here 40 MW); steam with a pressure of 4.0 – 9.0 MPa and a temperature of 560 – 950 °C can be produced [442].

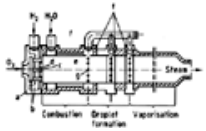

Figure 114. Schematic of the DFVLR H_2 – O_2 steam producer

a) Injection head; b) Gas distribution; c, d) Injection manifold; e) Combustion chamber; f) Injection rings; g) Radial water injection

9.3.4.2. *Electrochemical Energy Conversion*

In a reversal of the electrolysis process, the chemical energy of hydrogen can be converted directly into electrical energy. This takes place in fuel cells by allowing the “cold” combustion of hydrogen and oxygen to occur electrochemically (⟶ Fuel Cells) [443].

At room temperature the maximum possible voltage of the hydrogen – oxygen system is 1.23 V. In practice, cell voltages of 0.7 V are reached. Since heat – power conversions are avoided, efficiencies of 40 – 60 % can be attained. For details of the mode of operation, electrodes, catalyst materials and construction see [444], [445]. The most developed versions are the *low-temperature fuel cells*, which are usually identified by the electrolyte employed, as alkaline fuel cells (AFC), phosphoric acid fuel cells (PAFC), and solid polymer electrolyte fuel cells (SPEFC).

Alkaline fuel cells are used for power production in space as well as for special applications (relay stations, emergency power supply, military purposes). Without exception, they use pure hydrogen and carbon dioxide-free oxygen and have a low specific mass (7.2 kg/kW) and a long lifetime. They are, however, very expensive (1500 $/kW) [446].

Phosophoric acid fuel cells operate at 200 °C and use hydrogen or hydrogen – carbon monoxide mixtures produced from natural gas and air. The complete unit contains not only the cell modules but also various components, such as utility supplies, systems for temperature control, means to remove product water, and controls for the electrolyte pumping.

High-temperature fuel cells are still in the development stage; *molten carbonate fuel cells* (MCFC) operate at 600 °C and use an alkali-metal carbonate melt as electrolyte. *Solid-oxide fuel cells* (SOFC) contain ceramic ZrO_2 membranes, (solid electrolytes); on these electrolytes hydrogen reacts with air at 800 – 1000 °C. The hydrogen production from methane and steam can be integrated into the cell [447].

Fuel Cell Power Plants. Hydrogen, hydrogen-rich gases, or gases which can be converted to hydrogen can serve as fuel for fuel cell power plants. As depicted in Figure 115, such a facility for district electricity and heat generation comprises three subsystems; a fuel processor, a power section made up of a number of modules and a power conditioner. The suitability of the phosphoric acid fuel cell for this purpose has been proven in the United States in a field test (46 40-kW PAFC units) [449]. In Japan two 1 MW blocks have been built as an alternative energy production to thermal power generation [450].

Figure 115. Fuel cell power plant [448]

a) Fuel processor; b) Fuel cell power section; c) Power conditioner

9.4. Future Developments

Several books and monographs dealing with the perspectives and planning of the utilization of hydrogen as a major universal energy carrier have been published [386], [389], [451-453].

9.4.1. *Production Schemes*

Possible Schedule. The introduction of hydrogen as an energy carrier can be assumed to be governed mainly by economic aspects. The following developments can be expected [454]:

1. *First stage of utilization.* Future hydrogen utilization will be based on current uses of hydrogen, which are mainly nonenergetic. Hydrogen produced from non-hydrocarbon sources (normally from electrolysis) will be introduced to serve similar purposes.
   *Primary energy*will be supplied by nuclear or hydropower stations with energy reserves available during off-peak periods. The economic conditions for electrolytic production of hydrogen are especially favorable in France (nuclear power) [455] and Canada (hydropower) [456].
2. *Special Applications.* The combined production and utilization of hydrogen and oxygen by electrolysis could provide an additional economic impetus. Possible fields of utilization are:
   heavy oil upgrading (hydrogen for hydrogenation, oxygen for residue gasification) [457], [458]
   wood gasification combined with methanol production [459]
   all other kinds of coal, residue, or waste gasification with H_2, CH_4, CH_3OH, etc. as products
   biological process such as aerobic wastewater treatment (oxygen required). Hydrogen together with any methane produced in anaerobic stages, could be used as fuel for power production with gas motors or for a denitrification process.
3. *Introduction as a Direct Energy Supply.* Use of hydrogen will start as a supplement to actual energy markets, addition of up to 10 % hydrogen in natural gas nets or use as a fuel in transportation (city buses, underground/underwater devices, aircraft).
   The *primary energy supply* for the uses in both last sections would be the same as for the utilization as energy off peak power. In addition, wind, water, and solar units, which use electrolytically produced hydrogen for long-term storage would also be used.
4. *Advanced Stage of Utilization.* In countries with high solar radiation intensity but little industry, hydrogen could be used as a means of storing solar energy [460]. The solar radiation is converted into electric power using solar collectors, photovoltaic units, or Stirling motors driven by solar heat.
   First decentralized applications for hydrogen as an energy carrier could be for refrigerating large-scale units and for supplying heat in agricultural and industrial operations.
   Schemes for *supplying* the *total energy* requirements to remote areas by hydrogen have also been suggested, e.g., energy requirements of an island in Alaska supplied via wind energy and hydrogen [461].
5. *Hydrogen-Energy World.* The final goal of a world selfsufficient in energy (that does not depend on fossil resources) is the erection of large solar farms in the sunflooded desert areas of Africa, Asia, and Australia. Countries in these regions could export hydrogen (or electric power) to industrial regions with less sun. For more details see [454, pp. 297 – 304].
   According to this study, an area of 0.14×10^6 km^2 (0.1 % of the land surface of the globe), will be required for the production of 1000×10^6 tce/a. The total surface of suitable areas in desert regions is 0.6×10^6 km^2 which is more than sufficient.
   The material requirements for the construction of a solar farm are also within the capabilities of an industrialized country. This is shown in Table 48 [454, p. 321] for the Federal Republic of Germany. Nickel and silicon are the exceptions, the first because of the limited resources available and the latter because appropriate production technologies still have to be developed. Supply of other key items of solar farms such as solar cell panels, heliostat mirrors, Stirling motors, power generators, and electrolysis units, is also adequate (see Fig. 116) [454, p. 329].

Table 48. Comparison of material requirements for hydrogen plants with the consumption of primary raw materials in the Federal Republic of Germany (1983). Projected annual production of hydrogen plants ca. 155 PJ/a (equivalent to the energy produced by 100 100-MW-units per year, usable ca. 4000 h/a) [389]

Raw material	Annual consumption 1983*, 10^3 t/a	Consumption for production of solar plants, % of annual consumption (1983)			
		Photovoltaic solar cell	Solar tower	Paraboloid mirrors, solarthermic	Wind converter
Steel	26 600	12.7	32.0	24.0	11.8
Aluminum	1085	11.5	16.6	18.4	1.8
Nickel	63	95.2	206.3	135.0	47.6
Glass	1240	126.6	209.0	178.2	6.5
Concrete	210 000	10.6	18.3	21.0	5.0
Silicon	0.36	16 667			

* Annual primary energy use in 1980: World 300×10^3 PJ/a; Western Europe 76×10^3 PJ/a ; Fed. Rep. of Germany 7.6×10^3 PJ/a ; North America 90×10^3 PJ/a.

Figure 116. Comparison of annual production of certain industrial products in the Federal Republic of Germany and the required comparable items for hydrogen plants for the hydrogen energy concept [460]

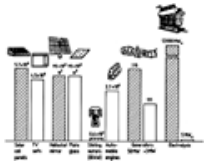

Projects. Already existing hydrogen projects serve and will serve the following purposes:

1. Demonstration of feasability
2. Study of technical problems occurring during operation under realistic conditions
3. Optimization and coordination of the various systems involved (e.g., mechanical – photovoltaic systems, electronic – photovoltaic systems)
4. Moving from the “tailor-made” stage of the components to a prototype production stage
5. Demonstration of self-sufficiency in terms of energy supply without fossil fuel or nuclear power

Although some individual components of a future hydrogen economy have already been tested (see Section Hydrogen-Energy Conversion Systems), few projects have realized the concept in its entirety from hydrogen production to its utilization. In 1979, R. E. BILLINGS built a house in Independence, Missouri, in which the energy supply, including that for car and tractor, was based on hydrogen. The primary energy was by hydropower, storage by means of metal hydrides and solar energy was used as subsidiary energy for heating purposes.

In Harnøsand, Sweden, O. TEGSTRØM has converted the energy supply of his car and his house, to hydrogen. Wind is used to supply the primary energy via and turbines with 55 kW. Metal hydrides are used for hydrogen storage.

Hysolar [462]. Founded in 1985, Hysolar (a joint venture between the Federal Republic of Germany and Saudi Arabia) foresees the installation of three photovoltaic electrolysis plants: a 100-kW (electric power) plant is to be built in Saudi-Arabia; two plants with 10 and 2 kW at a German and a Saudi Arabian university respectively serve research and demonstration purposes [462].

Solar Hydrogen Bavaria (SWB) is scheduled to go on stream in the Federal Republic of Germany in 1989 [463]. Figure 117 shows a diagram of the plant. Project costs amount to more than 50×10^6 DM (33×10^6 DM for the plant, 12×10^6 DM for personnel, 5×10^6 DM for infrastructure).

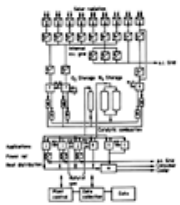

Figure 117. Block diagram of the Bavaria Solar Hydrogen Plant (planning status)

a) Photovoltaic cell; b) d.c. Converter; c) Power net; d) a.c./d.c. Converter; e) Electrolysis power supply; f) Electrolysis unit; g) Oxygen purification; h) Hydrogen purification; i) Alkaline fuel cell; j) Phosphoric acid fuel cell; k) Gas heater; l) Gas heater; m) Heat storage

Plant data:

Solar generator – 500 kW_{el} from various modules

Photovoltaic cells: total surface area 7000 m^2

Electrolysis unit:

100 kW_{el} conventional KOH electrolysis

100 kW_{el} zero gap KOH/NaOH electrolysis

80 kW_{el} solid polymer membrane electrolysis

Gas storage:

5000 m^3 (STP) of H_2 at 3 MPa

500 m^3 (STP) of O_2 at 3 MPa

Fuel cells:

6 – 8 kW_{el} alkaline type
80 kW_{el} phosphoric acid type
Catalytic combustion: 4 kW_{th}
Gas boiler: 20 kW_{th}

Euro-Quebec Project [464]. In this study initiated in the Federal Republic of Germany, the transport of hydroelectric power from Canada to Europe by means of hydrogen is being investigated. The Canadian hydropower is converted into hydrogen by electrolysis. The hydrogen is transported in one of three forms: (1) as liquid hydrogen shipped in special tankers (or eventually by aircraft); (2) as liquid ammonia shipped in special tankers; or (3) as methylcyclohexane shipped in oil tankers. The hydrogen is used in Europe for electricity, heat generation, vehicle and aviation propulsion, and hydrogen as an additive to natural gas for domestic or industrial use.

Figure 118 A shows the mass flow diagram developed for Phase I of the study; the energy flow diagram is shown in Fig. 18 B.

Figure 118. Flow diagrams for the transportation of hydrogen from Canada to Europe (Euro – Quebec project). It is planned to transport hydrogen in liquid form or chemically bound as methylcyclohexane (MCH) or ammonia by sea or air

A) Mass flow diagrams (transportation energy is neglected)

B) Energy diagram (in the energy diagram for chemically bound hydrogen only ammonia is shown because MCH is very similar and slightly less effective, η = 0.5 instead of 0.514)

Other Projects. Apart from these direct hydrogen projects, a great number of solar energy stations and wind farms (particularly common in California) can be seen as precursor of the hydrogen energy economy.

All hydrogen projects to date convert the universal but expensive energy carrier electricity with notable losses to hydrogen. Sometimes the cycle is completed and the hydrogen is converted back to electricity.

Direct production of hydrogen from solar radiation (e.g., by photochemical or biological methods) is still under research.

9.4.2. *Hydrogen Energy Economics*

Various factors can accelerate the introduction of the hydrogen technology, but do not exclude other solutions for the energy supply such as drastic savings or introduction of new energy forms. For time schedules for the introduction of hydrogen technology and other prognoses see [451], [454], [465]. The political decision to use hydrogen as an energy carrier reflects the following:

1. *Technological and Ecological Aspects.* The utilization of fossil and nuclear energy can be reduced in favor of energy savings, renewable resources, and solar energy. This could be forced by climate changes and environmental pollution.
2. *Economic Aspects.* Increased energy costs and storage costs for conventional primary energy and energy carriers could accelerate the introduction of the hydrogen economy.
3. *Political Aspects.* National hydrogen energy economics could be favored by the effect of primary energy import on the balance of trade, the creation of jobs by national energy programs, and the improvement of the structure of the economy.
4. *Security of Supply.* Energy produced from a local primary energy source can be stored in the form of hydrogen.
5. *Social Compatibility.* Energy production is coupled with an environmentally acceptable energy carrier.

Price estimates for hydrogen based on present-day technologies (see Fig. 119) [454, p. 341] depend on many factors. Comparisons of hydrogen production by means of solar cell plants, solar collector plants, parabolic mirror plants, or wind farms cannot consider the innovative potential of each technology over the projected time span.

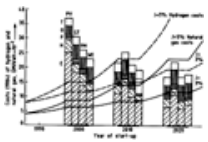

Figure 119. Hydrogen production costs for four hydrogen plants and three different start-up dates (interest rate: 8 %, depreciation: 30 a, construction: 4 a, inflation: 5 %). Comparison with real escalating hydrogen (– – –) and natural gas (——) costs (J = annual real rate of escalation, cost of electricity production = E, of hydrogen production = H, of piping = P, of operation = O, and T of transport for 3300 km)

The plants are: PV = photovoltaic solar cell plants, ST = solar thermic, solar power plants, PM = parabolic mirror solar thermic plants, WE = wind energy farms [454]

Another method to assess the role of hydrogen technology in the future energy supply is to compare the efficiency of energy supply systems. In Figure 120 the efficiencies of room heating using electric power produced by photovoltaic means are compared in the Federal Republic of Germany. The power is generated inside the country and then distributed either in a hydrogen pipeline (A) or as such (B) to the user. The much more efficient power production in the Sahara is also considered. Energy can be transported directly as current (D) or as hydrogen (C).

Figure 120. Comparison of photovoltaic energy generation in the Federal Republic of Germany (A, B) and in arid regions in North Africa (Sahara) (C, D), transportation in the form of hydrogen (A, C) or in the electric grid (B, D). End use: room heating in the Federal Republic of Germany [466]

9.4.3. *International Research*

A particular advantage of hydrogen technology is that all necessary systems and components are in an advanced state of development. Some technologies, such as hydrogen transport and storage, have been used industrially for decades. In spite of this, intensive research is still required in important areas. These areas are being investigated by several countries which cooperate under the auspices of the International Energy Agency, IEA (Table 49) [467].

Table 49. Hydrogen research activities of some member countries of the International Energy Agency (IEA) [467]

Subtask	Subject	Federal Republic of Germany	Japan	Sweden	Switzerland	United States
Storage	Metal hydride storage	×	×	×	×	×
	Liquefaction of hydrogen	×				
	Hydrogen storage on carbon					×
Conversion	Engine with inert gas recirculation		×			
	LH_2-Pump for internal combustion engines	×				
	Reversible hydrogen storage electrode		×			
Safety	Alternation of material properties by hydrogen embrittlement	×	×			
	Safety aspects of hydrogen combustion	×				
	Man years 1986:	13	10.5	1.5	3.2	4.5
	Man years 1987:	11.8	10.5	1.5	4	4.5
	Man years 1988:	11	10.5	1.5	4	4.5

Canada. Canada plans to generate hydrogen by hydropower and to test applications in pilot projects. Some activities are listed in Table 49. Systems analysis studies regarding special applications for hydrogen have also been carried out [468]. Canadian activities involve basic research in universities, industry, (especially in the field of electrolysis), and national research centers (sponsored by the National Research Council, NRC).

Federal Republic of Germany [452], [466]. Industry, research institutions, and universities are continuing the development of hydrogen technology with the help of government funds. Pilot projects are described in Section Hydrogen Energy Economics. In the 1990s, work will concentrate on following areas:

1. Energy generation (photovoltaic and solar thermal techniques in cooperation with countries with high solar radiation intensities)
2. Hydrogen production (electrolysis, advanced alkaline electrolysis, high-temperature electrolysis, basic research on biological hydrogen generation)
3. Applications (fuel cells, stationary power applications, peak power, power – heat combined cycle)
4. Materials and safety

The Federal Republic of Germany takes the leading role in the international funding of hydrogen technology. State funding in the energy and renewable resources areas amounts to 200×10^6 DM/a; 100×10^6 DM of this sum are to be spent in the 1990s on hydrogen technology.

Japan. Japanese activities are concentrated in the Sunshine Project of the Ministry of International Trade and Industry (MITI). According to annual reports on this topic [469], the following areas are being investigated (1986):

1. Hydrogen production by SPE water electrolysis (see Section Electrolysis)
2. Storage and transportation of hydrogen by using metal hydrides
3. Utilization of hydrogen (combustion, reciprocating engines, cars, etc.)
4. Safety (explosion prevention).
5. Photochemical processes, thermochemical cycles (Br^- system cycle), high-temperature – high-pressure water electrolysis, and hydrogen from biomass

Funding has decreased continuously from a maximum of 950×10^6 Yen in 1980 to 190×10^6 Yen in 1987. However, funds for solar energy (which amounted to 7560×10^6 Yen in 1987) and overhead costs are not included in these figures. Japan is also a member of the IEA (see Table 49).

United States. Research in the United States deals principally with energy generation via solar radiation and thus only indirectly with hydrogen technology, although a remarkable amount of solar energy is generated, in the South and the West of the United States in particular. This energy is fed directly into the grid. Hydrogen is at present not under consideration as an energy carrier.

Basic research at various universities takes place within the IEA framework and is sponsored by industry and government.

Soviet Union. Proof of the high standard of research in hydrogen technologies was the 1988 report of the flight of a Tupolev aircraft powered by liquid hydrogen (see Section Hydrogen-Energy Conversion Systems).

[Top of Page]

References

1. *Gmelin* System no. **2**, 8.
2. *Ullmann*, 3rd ed., **18**, 501.
3. *Ullmann*, 4th ed., **24**, 243.
4. R. D. McCarthy: *Hydrogen Technology Survey – Thermophysical Properties*, NASA SP-3089, Washington D.C. 1975.
5. JANAF: *Thermochemical Tables*, NSRD NBS 37, Washington D.C. 1975.
6. H. M. Roder in R. D. McCarthy, J. Hord (eds.): *Selected Properties of Hydrogen*, Dept. of Commerce NBS Report, Boulder 1981.
7. K. E. Cox, K. D. Williamson, Jr. (eds.): *Hydrogen: Its Technology and Implications*, **2nd vol. 1 – 5**, Boca Raton 1979.
8. L'Air Liquide: *Encyclopedie des Gaz*, Elsevier, Amsterdam 1976.
9. C. L. Newton, *Chem. Process. Eng. (London)* **48** (1967), no. 12, 51 – 58.
10. European Space Agency Contract Report, Contract 5977/84/NL/PB(SC), Chilton, United Kingdom, Jan. 1986.
11. K. F. Mucker, D. White *et al.*, *Phys. Rev. Lett.* **15** (1965) 586. Links
12. N. Sullivan *et al.*, *Recherche* **11 (111)** (1980) 587, see also *Bull. Inst. Int. Froid* **2** (1984) 484.
13. G. A. Neece *et al.*, *J. Comput. Phys.* **7** (1971) 621. Links
14. F. V. Grigorev, S. B. Kormer, O. L. Mikhailov, *JETP Lett. (Engl. Transl.)* **16** (1972) no. 5, 201.
15. R. Hawke, *Phys. Rev. Lett.* **41** (1978) 994. Links
16. *Landolt-Börnstein* **IV 2c**, Springer Verlag, Berlin-Heidelberg-New York 1980.
17. C. L. Young (ed.): *IUPAC Solubility Data Series*, vol. 5/6; Hydrogen and Deuterium, Pergamon Press, Oxford 1981.
18. D. Behrens, R. Eckermann (eds.): *DECHEMA Chemistry Data Series* vol. 6; “Vapor-Liquid Equilibria for Mixtures of Low Boiling Substances,” Dechema, Frankfurt 1982.
19. J. R. Lacher, *Proc. R. Soc. London, Ser. A* **161** (1937) 526.
20. W. Mueller, J. Blackledge, G. Libowitz (eds.): *Metal Hydrides*, Academic Press, New York – London 1968.
21. G. Alefeld, J. Völkl (eds.): *Hydrogen in Metals*, vol **1, 2**, Springer – Berlin – Heidelberg – New York 1978.
22. H. J. Hoge, R. D. Arnold, *J. Res. Natl. Bur. Stand. (U.S.)* **47** (1951) 63.
23. H. U. Karwat, *Chem. Ing. Tech.* **32** (1960) 605.
24. P. Protopapas *et al.*, *Hydrogene Met. Congr. Int.* 1972, 414.
25. G. Holleck, *J. Phys. Chem.* **74** (1970) 503.
26. D. R. Stull *et al.* in *JANAF Thermochemical Tables*, 2nd ed., NSRD NBS, Washington D.C. 1971.
27. J. W. Colton, *Hydrocarbon Process. 60* (1981) 177.
28. P. A. Giguere, *J. Chem. Educ.* **56** (1979) 571. Links
29. D. J. G. Ives, G. J. Janz: *Reference Electrodes*, "Chap. 2", Academic Press, New York 1961.

30. N. N., *Pure Appl. Chem.* **57** (1985) 531. Links
31. J. Hord, *Int. J. Hydrogen Energy* **3** (1978) 157.
32. I. L. Drell, F. E. Belles: *Survey of Hydrogen Combustion Properties*, NACA Report 1383, Washington D.C. 1952.
33. M. Fischer, H. Eichert, *DFVLR-Nachr.* **34** (1981) 33.
34. W. C. Eichelberger *et al.*, *Chem. Eng. Prog.* **57** (1961) no. 8, 94.
35. *Tritium in the Physical and Biological Sciences*, vol. **1**, Int. Atomic Energy Agency, Vienna 1962, p. 56.
36. *Kirk Othmer*, **7**, 539.
37. G. Vasaru: *Deuterium and Heavy Water – A Selected Bibliography*, Elsevier, Amsterdam – Oxford – New York 1975.
38. R. Prydz: *The Thermodynamic Properties of Deuterium*, NBS Report 9276, Boulder (Col.) 1966.
39. D. H. Liebenberg, R. L. Mills, J. C. Bronson, LA-7071-MS Informal Report, Los Alamos 1978.
40. P. C. Souers et al.: Three-Phase Region of D2-DT-T2, Lawrence Livermore Laboratory Report UCRL – 79 036, 1977.
41. H. Gutowski, *Kerntechnik* **12** (1970) no. 5/6, 233.
42. Y. W. Wong *et al.*, *Sep. Sci. Technol* **15** (1980) 423.
43. G. D. Sandrock, E. L. Huston, *Chem. Ing. Tech.* **11** (1981) 754.
44. K. B. Wiberg, *Chem. Rev.* **55** (1955) 713. Links
45. E. A. Evans: *Tritium and its Compounds*, 2nd ed., J. Wiley & Sons, New York 1974.
46. A. Roberts, *Sci. Am.* **206** (1963) no. 6, 94.
47. G. Winnewisser: *Cosmochemistry*, Springer Verlag, Berlin 1973.
48. E. Goebel *et al.*, *Oil Gas J.* **81** (1984, May 7) 215.
49. N. Getoff: *Wasserstoff als Energieträger*, Springer, Wien – New York 1977, p. 5. NASA/ASEE Systems Design Study “Hydrogen Energy Carrier,” I, II (1973)
50. H. Bertling, G. Nashan, *Erdöl Kohle Erdgas Petrochem.* **34** (1981) 397 – 401.
51. G. Dalluge, P. Polster, *Chem. Tech. (Leipzig)* **33** (1981) 636 – 640.
52. G. Bernard, J. Vidal, *Pétrole et techniques* **307** (May 1984) 28 – 38.
53. J. F. Meckel, C. Flockenhaus, *Chem. Anlagen + Verfahren* 1981, May, 122 – 138.
54. Y. Osumi, *CEER Chem. Econ. Eng. Rev.* **17** (June 1985) 5 – 20.
55. J. G. Kronseder, M. J. P. Bogard, *Encycl. Chem. Process* **9** (1979) 299 – 328.
56. G. G. Mayfield, V. H. Agreda, *Energy Prog.* **6** (1986) no. 4, 214 – 218.
57. J. F. Clausen, C. A. Zee: *TRW-Systems and Energy*, EPA contract 68–02–2635.
58. A. K. Kuhn, *Chem. Eng. Prog.* (1982) no. 4, 64 – 68.
59. W. N. Clark, R. L. Litzinger, I. R. Straughan, *Energy Prog.* **7** (1987) no. 2, 99 – 104.
60. B. Brunke *et al.*, *Erdöl, Erdgas* **103** (1987) 289 – 293.
61. D. H. Cortez, C. J. La Delfa, *Hydrocarbon Process.* **61** (1981, Feb.) 111.
62. F. A. Clauson, C. P. Marion, *AIChE – IMIQ Energy Meeting*, Acapulco, Oct. 1980.
63. Uhde GmbH, Dortmund, private communication (1988).
64. H. Jungfer, R. Lohmüller, *Erdöl Kohle Erdgas Petrochem.* **39** (1986) no. 4, 188 – 1892.
65. C. P. Marion, J. R. Muenger, AiChE, (Houston, Texas, USA, 1981) 4.
66. H.-J. Madsack, *Hydrocarbon Process.* **61** (1982) no. 7, 169 – 172.
67. Y. Shimizu, *CEER Chem. Econ. Eng. Rev.* **10** (1982) no. 7, 9 – 14.
68. Linde AG, DE 34 14 716.0, 1984 (R. Lohmüller, M. Heisel, P. Hesse, U. Lahne).
69. H. Jungfer, *Linde Reports on Science and Technology* **40** (1985) 14 – 20.
70. R. M. Eccles, A. M. Gray, W. B. Livingston, *Oil Gas J.* **79** (1982, Apr. 12) 123 – 129.
71. Osaka Gas Corp. JP 108 1726, 30, 34.
72. G. W. Bridger, *Chem. Process Eng. (London)* **53** (1972) 38.
73. *Catalyst Handbook* ICI, Wolfe Scientific Books, London 1972.
74. J. Rostrup-Nielsen, *Chem. Eng. Prog.* **73** (1977) no. 9, 87 – 92. Links
75. Toyo Engineering, US 4 025 457, 1977 (T. Tomita, K. Kikuchi, T. Sakamoto).

76. R. G. Minet et al., EPRI-Report EM 1487 (1980), EM 2096 (1981)
77. *Nitrogen*, **144** (1983) 33.
78. D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds*, Wiley, New York 1969, p. 631 ff. W. D. Seider, R. Gautam, C. W. White: “Computation of Phase and Chemical Equilibrium. A Review,” *ACS-Symp. Ser.* **124** (1980) 115 – 134.
79. J. P. van Hook, *Catal. Rev. Sci. Eng.* **21** (1980) no. 1, 1 – 51.
80. J. C. de Deken, E. F. Devos, G. F. Froment, *ACS Symp. Ser.* **196** (1982) 181 – 197.
81. A. P. Murray, T. S. Snyder, *Ind. Eng. Chem. Process Des. Dev.* **24** (1985) 286 – 294.
82. C. P. P. Singh, D. N. Saraf, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) no. 1, 1 – 7.
83. F. J. Dent *et al.*, *Trans. Inst. Gas Eng.* **95** (1945/46) 602 – 709.
84. S. Paloumbis, E. E. Petersen: “Chemical Reaction Engineering,” *ACS Symp. Ser.* **196** (1982) 489 – 497.
85. J. W. Colton, *Hydrocarbon Process.* **60** (1981) no. 1, 177 – 184.
86. D. L. Trimm, *Catal. Rev. Sci. Eng.* **16** (1977) no. 2, 155 – 190.
87. K. H. Brachthäuser, R. R. J. Jungerhans, H. Tielen, *Gas Wärme Int.* **22** (1973) no. 6, 209 – 214.
88. M. R. Agarwal, T. Bajpai, *Chem. Age India* (1977) no. 11, 995 – 1010.
89. A. Watson, *Hydrocarbon Process.* **62** (1983) no. 3 91 – 95.
90. H. D. Marsch, J. J. Herbort, *Hydrocarbon Process.* **61** (1982) no. 6, 101 – 105.
91. *Ann., Nitrogen* **166** (1987) no. 2, 24 – 31.
92. W. Balthasar, D. J. Hambleton, *Int. J. Hydrogen Energy* **5** (1980) 21 – 31.
93. W. F. van Weenen, J. Tielrooy, Proc., 4th World Hydrogen Energy Conf., Pasadena, California, USA, 1982.
94. S. P. Goff, S. I. Wang, *Chem. Eng. Prog.* (1987) no. 8, 46 – 53.
95. J. L. Heck, T. Johansen, *Hydrocarbon Process.* **57** (1978) no. 1, 175 – 177.
96. S.-J. Wang, *Oil Gas J.* **80** (1982) no. 11, 55 – 62. Links
97. J. R. Rostrup-Nielsen, H. Stahl, N. R. Udengaard, Fuel Cell Seminar, Tucson, Arizona, USA, 1985.
98. Endurance Testing of a High-Efficiency Steam Reformer for Fuel Cell Power Plants, EPRI AP-6071, Project 2192–1 Final Report, October 1988.
99. United Technologies Corp., US 4 071 330, 1978 (R. A. Sederquist).
100. J. Singh, H. F. Nielsen, R. Streicher, *Brennst. Wärme Kraft* **33** (1981) 412 – 416.
101. L. Silberring, Proc. of British Sulphur Corp., 4th Conf., paper 5, London, 1981, pp. 93 – 112.
102. M. Oertel *et al. Chem. Eng. Technol.* **10** (1987) 248 – 255.
103. E. J. Nobles in A. V. Sluck, G. Russel James (eds.): *Ammonia*, vol. **I**, Marcel Dekker, New York 1973, pp. 275 – 290.
104. M. Fytzani-Stephanopoulos, G. E. Voecks, *Energy Progress* **1** (1981) nos. 1 – 4, 52 – 58.
105. Alsands Projekt, *Oil Gas J.* **78** (1980) no. 24, 61.
106. G. R. Charamonte, R. Sharma, *Hydrocarbon Process.* **62** (1982) no. 9, 255 – 257.
107. R. R. Dean, J. L. Mauleon, W. S. Letzsch, *Oil Gas J.* **80** (1982) no. 40, 75 – 80; no. 41, 168 – 176.
108. Atlantic Richfield Comp., G. P. Masologites, L. H. Beckberger, R. R. Edison, company leaflet, 1974, Harvey, Illinois, United States.
109. R. Müller, Proc. World Hydrogen Energy Conf., 4th, Pasadena, California USA, 1982.
110. J. G. Wragg, M. A. Kaleel, C. S. Kim, *Coal Processing Technology VI, Am. Inst. Chem. Eng.* (1980) 186 – 192.
111. H. Wendt in C. J. Winter, J. Nitsch (eds.): *Wasserstoff als Energieträger*, Springer Verlag, Berlin 1986, p. 179.
112. C. Carpetis, *Dtsch. Forsch. Versuchsanst. Luft Raumfahrt Nachr.* **34** (1981) 15.
113. H. Wendt, *Chem. Ing. Tech.* **56** (1984) no. 4, 265 – 272.
114. J. Divisek, *Brennst. Wärme Kraft* **38** (1986) no. 11, 512 – 515.
115. H. Vandenborre, R. Leysen, *Electrochim. Acta* **23** (1978) 803 – 804. Links
116. J. Divisek, P. Malinowski, *J. Electrochem. Soc.* **133** (1986) no. 5, 915 – 920. Links
117. M. S. Casper, *Chemical Technol*, Review no. 102, Noyes Data Corp. (Park Ridge, N.J., 1978).
118. G. Sandstede, *Chem. Ing. Tech.* **61** (1989) 349 – 361.
119. Brown Boveri, R. Oberlin, M. Fischer, *Technik 8 – 86*, 1986, pp. 445 – 450.
120. Japan's Sunshine Project, *Annual Summary, 1983 – 1986 of Hydrogen Energy*, R&D, Pub. Japan Ind. Tech. Ass., Tokyo 1987.

121. G. Imarisio, A. A. Strub (eds.): *Hydrogen as an Energy Carrier Proc., 3rd Int. Seminar*, Lyon 1983, Dordrecht, Netherlands 1983, pp. 237 – 243.
122. W. Baukal, *Chem. Ing. Tech.* **41** (1969) 791.
123. W. Dönitz, R. Streicher, *Chem. Ing. Tech.* **52** (1980) 436.
124. E. Erdle, W. Dönitz, *VDI-Ber.* **602** (1987) 51 – 51.
125. A. O. Isenber, *Solid State Ionics* (1981) no. 3/4, 431 – 437.
126. K. Iwahara, *Kagaku Sochi* **28** (1986) no. 1, 90 – 91.
127. K. H. Quandt, R. Streicher, *Int. J. Hydrogen Energy* **11** (1986) no. 5, 309 – 315.
128. D. Bergner, M. Hartmann, R. Staab, *Chem. Ing. Tech.* **59** (1987) no. 4, 271 – 280.
129. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances*, Springer Verlag, Berlin – Heidelberg – New York – Düsseldorf 1973.
130. J. E. Funk, R. M. Reinstrom, *Ind. Eng. Chem. Proc. Res. Dev.* **5** (1966) 366.
131. B. M. Abraham, F. Schreiner, *Ind. Eng. Chem. Fundam.* **13** (1974) 30.
132. H. Hofmann, *Chem. Ing. Tech.* **48** (1976) 87.
133. G. E. Beghi, *Int. J. Hydrogen Energy* **10** (1985) 431.
134. H. Engels, J. E. Funk, K. Hesselmann, K. F. Knoche, *Int. J. Hydrogen Energy* **12** (1987) 291.
135. R. Schulten: *Nuclear Energy as a Primary Energy Source for Hydrogen Production*, 2nd World Hydrogen Conference, Zürich 1978.
136. M. G. Bowman: *Interfacing Primary Heat Sources and Cycles for Thermochemical Hydrogen Production*, 3rd World Hydrogen Energy Conference, Tokyo 1980.
137. C. E. Bamberger, D. M. Richardson, *Cryogenics* **16** (1976) 197. Links
138. C. E. Bamberger, D. M. Richardson, *Cryogenics* **18** (1978) 170. Links
139. G. DeBeni, *J. Electrochem. Soc.* **128** (1982) 1.
140. G. E. Besenbruch *et al.*: *Hydrogen Production by GENERAL ATOMIC Sulfur Iodine Process. A Progress Report*, 3rd World Hydrogen Energy Conference, Tokyo 1980.
141. K. F. Knoche, H. Schepers, K. Hesselmann: Second Law and Cost Analysis of the Oxygen Generation Step of the GENERAL ATOMIC Sulfur-Iodine Cycle. Proceedings of the 5th World Hydrogen Energy Conference, Toronto 1984
142. H. Engels, K. F. Knoche, M. Roth: *Direct Dissociation of Hydrogen Iodide—An Alternative to the GENERAL ATOMIC Process*. Proceedings of the 6th World Hydrogen Energy Conference, Vienna 1986.
143. S. Sato *et al.* in : T. N. Veziroglu, J. B. Taylor (eds.): “Studies on the Nickel-Iodine-Process for Hydrogen Production III,” *Hydrogen Energy Progress V*, vol **2**, Pergamon Press, New York 1984, pp. 457 – 465.
144. S. Shimizu, *Int. J. Hydrogen Energy* **12** (1987) 687.
145. K. F. Knoche, J. E. Funk, *Int. J. Hydrogen Energy* **2** (1977) 377, 387.
146. J. M. Sedlak *et al.*, *Hydrogen production using solid-polymer-electrolyte technology for water electrolysis and hybrid sulfur cycle*, General Electric Company, Wilmington, Mass. and Electric Power Research Institute, Palo Alto, CA. (ed.), EM-1185, Research Project 1086–3, 1979, pp. 6/1–8-1.
147. K. Petrov, I. Nikilov, T. Vitanov, *Int. J. Hydrogen Energy* **10** (1985) 245.
148. B. D. Struck, *et al.* in G. Imarisio, A. S. Strub (eds.): *Hydrogen as an energy carrier*, "Development and scaling up of an electrolytic cell for the anodic oxidation of sulphur dioxide and cathodic production of hydrogen for the sulphuric acid hybrid cycle: EUR 8651, EUR 8300" D. Reidel Publishing Company, Dordrecht, 1983, pp. 35 – 45.
149. E. Bilgen, C. Bilgen, *Int. J. Hydrogen Energy* **11** (1986) 241.
150. O. M. Bockris, B. Danapani, D. Cocke, J. Ghoro-ghchian, *Int. J. Hydrogen Energy* **10** (1985) 179.
151. J. Lede, F. Lapique, J. Villermaux, *Int. J. Hydrogen Energy* **8** (1983) 675.
152. H. H. G. Jellinek, H. Kachi, *Int. J. Hydrogen Energy* **9** (1984) 677.
153. J. Lede *et al.*, *Int. Chem. Engineering* **26** (1986) 647.
154. J. Lede, J. Villermaux, R. Ouzane, M. A. Hossain, R. Ouahes, *Int. J. Hydrogen Energy* **12** (1987) 3.
155. A. Farhataziz, M. A. Rodgers: *Radiation Chemistry*, VCH Verlagsgesellschaft, Weinheim, Germany 1987.
156. A. Danno, *CEER Chem*, Econ. Eng. Rev. (1974) 7.
157. W. Kerr, D. P. Majumdar in T. N. Veziroglu (ed.): *Hydrogen Energy*, Part A, Plenum Press, New York 1975, pp. 167.
158. J. W. T. Spinks, R. J. Woods: *An Introduction to Radiochemistry*, J. Wiley & Sons, New York, 1964.
159. I. Tanaka, T. Carrington, H. P. Broida, *J. Chem. Phys.* **35** (1961) Links
160. K. Kalyanasundaram, M. Grätzel, *Angew. Chem. Int. Ed. Engl.* **18** (1979) 70.
161. M. Grätzel, *Ber. Bunsenges. Phys. Chem.* **84** (1980) 981.
162. M. Tomkiewicz, H. Fay, *Appl. Phys.* **18** (1979) 1.

163. H. Gerischer, W. Mindt, *Electrochim. Acta* **13** (1968) 1329. Links
164. K. Kalyanasundaram, E. Pelizetti, M. Grätzel, *Coord. Chem. Rev.* (1989) in print.
165. D. Duhong, N. Serpone, M. Grätzel, *Helv. Chim. Acta* **67** (1984) 1012.
166. P. Candon, *Nature (London)* **262** (1976) 715.
167. E. N. Kondratieva, I. N. Gogotov in A. Fiechetr (ed.) *Adv. Biochem. Eng. Biotechnol.* **28** (1983) 139 – 191.
168. J. M. Olson, J. D. Bernstein, *Ind. Eng. Chem. Prod. Res. Dev.* **21** (1982) 640.
169. H. Gest, M. D. Kamen, *Science (Washington, D.C.)* **109** (1949) 558.
170. R. Bollinger, H. Zürrer, R. Bachofen, *Appl. Microbiol. Biotechnol.* **23** (1985) 147.
171. A. Daday, A. H. Mac Kerras, G. D. Smith, *J. Gen. Microbiol* **131** (1985) 231.
172. K. K. Rao, L. Rosa, D. O. Hall, *Biochem. Biophys. Res. Commun.* **68** (1976) 21. Links
173. *Techno Japan* **19** (1986) no. 7, 104.
174. P. B. Tarman, R. Biljetina: *The Steam-Iron-process for Hydrogen Production*, 85th National Meeting of the AIChE, Philadelphia 1978.
175. W. H. Seitzer, US 3 928 550, 1975.
176. D. C. Erickson: *Hydrogen from Coal via Tin Redox Process*, Final Report; Energy Concepts Company, Annapolis, USA 1981.
177. G. Kurz, *Proc*, Eng. (1976, Oct.) 97.
178. G. Strickland, *Int. J. Hydrogen Energy* **9** (1984) 759.
179. T. Yamamoto, *CEER Chem. Econ. Eng. Rev.* **13** (1981) no. 1 – 2, 11.
180. J. C. Amphlett *et al.*, *Can. J. Chem. Eng.* **59** (1981) 720. Links
181. J. C. Amphlett, M. J. Evans, R. F. Mann, R. D. Weir, *Can. J. Chem. Eng.* **63** (1985) 605. Links
182. K. Restin, P. Jurkat, H. Hiller, *GWF Gas Wasserfach Gas Erdgas* **123** (1982) 105.
183. R. W. Bartlett: *Preliminary Evaluation of Processes for Recovering Hydrogen from Hydrogen Sulfide*, "File No. 79–447", SRI International, July 1979, p. 55.
184. M. E. D. Raymond, *Can. Sulfur Symp. Calgary*, Canada 1974.
185. Y. Kotera, N. Todo, K. Fukuda, US 3 962 409, 1976.
186. T. Chivers, J. B. Hyne, C. Lau, *Int. J. Hydrogen Energy* **5** (1980) 499.
187. W. C. Chen, US 4 066 739, 1978.
188. M. Dokiya, H. Fukuda, T. Kameyama, *Bull. Chem. Soc. Jpn.* **51** (1978) 150. Links
189. I. G. Nixon, *Chem. Br.* **May 5**, (1974) 179.
190. T. Chivers, C. Lau, *Int. J. Hydrogen Energy* **12** (1987) 561.
191. T. Kameyama, *Int. J. Hydrogen Energy* **8** (1983) 5.
192. S. A. Naman, S. M. Aliwi, K. Al-Emara, *Int. J. Hydrogen Energy* **11** (1986) 33.
193. H. Kalenda, G. Ruckelhaus, BMFT-Forschungsbericht T 81–1012 (1981).
194. J. Schulze, H. Gaensslen, *Chem. Ind.* **36** (1984) no. 3, 135 – 140;no. 4, 202 – 208.
195. R. Lohmüller, *Chem. Ind.* **36** (1984) no. 3, 147 – 151.
196. G. Bauer in [126], p. 195
197. C. R. Baker, NASA Contractor Report 159 163 (1979).
198. A. L. Kohl, F. C. Riesenfeld: *Gas Purification*, 4th ed., Gulf Publishing, Houston 1985.
199. H. Baldus, K. Baumgärtner, H. Knapp, M. Streich, “Verflüssigung von Gasen,” in *Winnacker, Küchler*, 4th ed., vol. **2**, Hanser Verlag, Wien – München 1983.
200. *Landolt-Börnstein*, **IV, 2 c** , Springer Verlag, Berlin – Heidelberg – New York 1980.
201. M. G. Zellner, J. M. Claitor, J. M. Prausnitz, *Ind. Eng. Chem. Fundam.* (1970) no. 9, 549 – 564.
202. H. C. Rowles, D. M. Nicholas, H. L. Vines, M. F. Hilton, *Energy Prog.* **61** (1986) no. 3, 25 – 32.
203. W. Förg, *Chem. Process Eng. (London)* **52** (1971) no. 2, 57 – 68.
204. A. Haslam, P. Brook, H. Isalski, *Hydrocarbon Process.* **55** (1976) no. 1, 103 – 106.
205. T. Matsuoka, *Chem. Age India* **30** (1979) no. 3, 119 – 128.
206. W. Scholz, *Linde-Ber. Tech. Wiss.* **45** (1979) 9 – 17 (LINDE Reports on Science and Technology).
207. *Technical Data Book of Petroleum Refining*, vol. **2**, American Petr. Inst. 1982, pp. 9 – 17.
208. H. Becker *et al.*, *Ger. Chem. Eng.* **7** (1984) no. 2, 91 – 101.

209. Petrocarbon, GB 10 189.71, 1971 (K. S. Doherty).
210. B. J. Grotz, *Hydrocarbon Process.* **46** (1967) no. 4, 197 – 202.
211. E. Karwat, *Erdöl Kohle Erdgas Petrochem.* **17** (1964) 439 – 447.
212. R. Fabian, D. Tillmann, *CPP Ed. Europe* (1988) no. 3, 35 – 37.
213. J. S. Davis, J. R. Martin, *Chem. Eng. Prog.* **76** (1980) , no. 2, 72 – 79.
214. S. I. Wang, D. M. Nicholas, S. P. DiMartino, *Oil Gas J.* **82** (1984) no. 5, 111 – 117. Links
215. With permission of W. Schmid, LINDE AG Haus der Technik Vortrag 1985.
216. W. Scholz, *DECHEMA Monogr.* **58** (1968) 31.
217. W. Förg, *GWF Gas Wasserfach: Gas/Erdgas* **113** (1972) 538 – 540.
218. W. Förg, *Chem. Anlagen + Verfahren* 1970, no. 3, 33 – 44.
219. D. W. Breck: *Zeolite, Molecular Sieves*, J. Wiley & Sons, New York 1974.
220. D. M. Ruthven: *Principles of Adsorption and Adsorption Processes*, J. Wiley & Sons, New York 1974.
221. A. Mersmann, U. Münstermann, J. Schadl, *Chem. Ing. Tech.* **55** (1983) 446 – 458.
222. K. Knoblauch, H. Jüntgen, W. Peters, *Erdöl Kohle Erdgas Petrochem.* **32** (1979) no. 12, 551 – 556.
223. J. H. Collins, W. C. Miller, J. E. Philcox, *Union Carbide Corp., Air Poll. Contr. Assoc., 65th Meeting*, Miami Beach 1972.
224. F. G. Wießner, *Gas Separation and Purification* **2** (1980) no. 3, 115 – 119.
225. E. Richter, K.-B. Harder, K. Knoblauch, H. Jüntgen, *Chem. Ing. Tech.* **56** (1984) no. 9, 684 – 691.
226. Union Carbide, US 3 986 849, 1975 (A. Fuderer, E. Rudelstorfer).
227. D. H. Werner, G. A. Schlichtharle, *Ammonia Plant Saf.* **24** (1980) 12 – 18
228. A. Watson, *Hydrocarbon Process.* **62** (1983) no. 3, 91 – 95.
229. F. J. Brykowski (ed.): *Ammonia and Synthesis Gas, Recent Energy-Saving Processes*, Noyes Data Corp., Park Ridge, New Jersey 1981, pp. 130 – 147.
230. D. S. Newsome, *Catal. Rev. Sci. Eng.* **21** (1980) no. 2, 275 – 318.
231. F. J. Brykowski (ed.), [229], pp. 147 – 156.
232. J. M. Moe, *Chem. Eng. Prog.* **58** (1962) no. 3, 33 – 36 Links
233. F. J. Brykowski (ed.), [229], 156 – 162.
234. P. N. Hawker, *Hydrocarbon Process.* **61** (1982) no. 4, 183 – 187.
235. E. Kuijpers, *Appl. Catal.* **25** (1986) 139 – 147.
236. W. van der Wal, E. Kuijpers, P. Berben, J. W. Gens, *Proc. Int. Gas Res. Conf.*, 1986, Toronto, pp. 1134 – 1143.
237. D. W. Allen, W. H. Yen, *Chem. Eng. Prog.* **69** (1973) no. 1, 75 – 79. Links
238. Engelhard Min. & Chem., US 4 238 468, 1980.
239. J. H. Cramer, L. E. Thompson, T. H. Hsiumg, W. T. McShea, 2nd Latin American Petrochemical Congress, Mexico City, Nov. 1978.
240. J. C. Bonacci, T. G. Otchy, *AIChE Symp*. (Denver, Co., Aug. 1977).
241. R. Assmann, *GWF Gas Wasserfach: Gas/Erdgas* **121** (1980) no. 1, 23 – 27.
242. A. L. Kohl, F. C. Riesenfeld (eds.): *Gas Purification*, 4th ed., Gulf Publishing Company Book Division, Houston – London – Paris – Tokyo 1985.
243. S. A. Newman (ed.): *Acid and Sour Gas Treating Processes*, Gulf Publishing Company Book Division, Houston – London – Paris – Tokyo 1985.
244. H. W. Schmidt, H. J. Henrici, *Chem. Ztg.* **96** (1972) 154.
245. R. N. Tennyson, R. P. Schaaf, *Oil Gas J.* **75** (1978) 78.
246. Proc. Symp. Ammonia from Coal, TVA Muscle Shoals, May 1979.
247. LINDE AG, US 4 332 596, 1982 (G. Ranke).
248. K. Takeuchi, K. Matsumura, K. Yaginuma, *Fluid Phase Equilib.* **14** (1983) 255.
249. G. Ranke, *Linde Reports on Science and Technology* **18** (1973) 7.
250. W. Scholz, *Linde Reports on Science and Technology* **30** (1979) 9.
251. G. E. Pack: *Investigation of Sulfur Removal Using Direct Oxidation Processes*, 78th Annual Meeting of the Air Pollution Control Association, Detroit June 1985. F. Rosendahl, *Erdöl Kohle Erdgas Petrochem.* **34** (1981) 132.
252. M. Heisel, F. J. Marold, *Hydrocarbon Process.* **66** (1987) 35.
253. *Ullmann*, 4th ed., **1**, 134.

254. *Kirk Othmer*, **7**, 639.
255. G. J. Grashoff, C. E. Pilkington, C. W. Corti, *Platinum Met. Rev.* **27** (1983) 157.
256. Matthey Metals Ltd, Company Brochure, Diffusion Units.
257. “Special Report: Permeation – Membrane Separation,” *Hydrocarbon Process.* **62** (1983) no. 8, 43 – 62.
258. S. Egli, A. Ruf, A. Buck, *Swiss Chem.* **6** (1984) 90.
259. A. K. Fritsche, R. S. Narayan, *CEER Chem. Econ. Eng. Rev.* **19** (1987) nos. 1, 2, 3 (19).
260. R. Rautenbach, W. Dahm, *Chem. Eng. Process.* **21** (1987) 141. Links
261. *Hydrogen Separation Systems, PRISM Separators*, Monsanto, Company Brochure, 1981.
262. D. L. Mac Lean, C. E. Prince, Y. C. Chae, *Chem. Eng. Prog.* **76** (1980) March, no. 98.
263. R. Barron: *Cryogenic Systems*, McGraw-Hill, New York 1966.
264. H. W. Wooley, R. B. Scott, F. G. Brickwedde, *Journal Res. NBS* **41** (1948) 379, RP 1932.
265. J. A. Barclay, W. A. Steyert, *Magnetic Refrigeration for Space Applications, Report on a Design Study, Los Alamos Sci. Lab. Rep.* **LA-8134** (1980).
266. T. Hasimoto, T. Numasawa, M. Shino, T. Okada, *Cryogenics* 21 (1981) 647.
267. H. Blechschmidt, *Kälte* **9** (1976) 357.
268. C. L. Newton, *Chem. Process. Eng. (London)* **48** (1967) no. 12, 51 – 58.
269. G. G. Haselden: “Refrigeration and Liquefaction Cycles,” in G. G. Haselden (ed.): *Cryogenic Fundamentals*, Academic Press, London 1971, pp. 17 – 89.
270. W. Peschka: *Flüssiger Wasserstoff als Energieträger*, Springer Verlag, Wien – New York 1984, pp. 33 – 37.
271. H. Gutowski, M. Wanner, *Haus Tech. Vortragsveröff.* **448** (1982) 60 – 64.
272. C. R. Baker, R. L. Shanner: “A Study of the Efficiency of Hydrogen Liquefaction,” *Int. J. Hydrogen Energy* **3**, (1978) 321 – 334.
273. J. A. Barclay, ASME Paper 81-HT-82, 20th Joint ASME/AiChE National Heat Transfer Conf., Milwaukee, Wisc., Aug 1981.
274. W. A. Steyert, *J. Appl. Phys.* **49** (1978) 1227 – 1230. Links
275. J. G. Bednorz, K. A. Müller, *Z. Phys. B: Condens. Matter Quanta* **64** (1986) 189.
276. R. O. Voth, *Cryogenics* **25** (1985) 511 – 517. Links
277. C. Sindt, *Cryogenics* **10** (1970) 372. Links
278. W. Tunn, *Erdöl Kohle Erdgas Petrochem.* **26** (1973) 498.
279. O. Bernauer, C. Halene, *J. Less Common Met.* **131** (1987) 213.
280. E. Turnwald, *LINDE Ber. Techn. Wiss.* **36** (1975) 15.
281. W. Peschka, *Int. J Hydrogen Energy* **7** (1982) 661 – 670.
282. D. T. Wade, L. L. Ansell, W. R. Epperly, *Chem. Tech. (Heidelberg)* **12** (1982) 242.
283. K. Baumgärtner, R. Ewald, *Chem. Ing. Tech.* **59** (1987) 296.
284. G. Kaske, H. G. Fasold, P. Schmidt in D. Behrens (ed.): *Transport- und Verteilungsleitungen in Wasserstofftechnologie*, DECHEMA, Frankfurt/Main 1986, p. 462.
285. W. Diehl, *Gas Aktuell* **16** (1978) 20.
286. G. G. Leth: *Transmission of Gaseous Hydrogen*, Messer Griesheim, Company Brochure. F. J. Edeskuty, K. O. Williamson “Liquid Hydrogen Storage and Transmission,” in E. Cox, K. D. Williams Jr. (eds.): *Hydrogen, Its Technology and Implications*, vol. **2**, CRC Press, Boca Raton 1979.
287. G. Beghi *et al.*: *Hydrogen, Oxygen and Natural Gas by Pipelines*, Comparative Transport Costs. Rep. EUR-5103 e, Euratom-C.C.R., Ispra (Italy) 1973.
288. G. D. Brewer (ed): *LH2 Airport Requirements Study*, NASA CR 2700, Oct. 1976.
289. C. Carpetis in C. J. Winter, J. Nitsch (eds.): *Wasserstoff als Energieträger*, “Speicherung, Transport und Verteilung von Wasserstoff,” Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1986, p. 245.
290. C. Carpetis, *Int. J. Hydrogen Energy* **7** (1982) 191.
291. L. Paterson, *Int. J. Hydrogen Energy* **8** (1983) 53.
292. J. A. Paivanas, O. P. Roberts, D. I. J. Wang, *Adv. Cryog. Eng.* **10** (1965) 197.
293. W. Peschka (ed.): *Flüssiger Wasserstoff als Energieträger*, Springer, Wien – New York 1984.
294. J. Hord, W. R. Parrish in: K. E. Cox, K. D. Williamson (eds.): *Hydrogen—Its Technology and Implications*, “Economics of Hydrogen,” vol 5, CRC Press, Boca Raton 1979.
295. N. Hallet: *Study, Cost and System Analysis of Liquid Hydrogen Production*, NASA—Rep. CR 73–226, 1968.
296. H. R. Gray *et al.*: *Potential Structural Materials Problems in a Hydrogen Energy System*, NASA TM-X71 752, June 1975.
297. R. A. Oriani, J. P. Hirth, M. Smialowski (eds.): *Hydrogen Degradation of Ferrous Alloys*, Noyes Publications, Park Ridge, USA 1985.
298. W. Thomson, I. M. Bernstein (eds.): *Effect of Hydrogen on Behaviour of Metals*, Metal Soc. AIME, New York 1976.

299. C. G. Interrante, G. M. Pressouvre (eds.): *Current Solutions to Hydrogen Problems in Steel*, ASM Metals Park, Ohio 1982.
300. J. Eastman *et al.* in M. Bernstein, A. W. Thomson (eds.): *Hydrogen Effects in Metals*, AIME, New York 1981, p. 397.
301. C. S. Carter: *ARPA Handbook on Stress Corrosion and Corrosion Fatigue*, "Stress Corrosion Cracking and Corrosion Fatigue of Medium-Strength and High-Strength Steels," 1977.
302. G. A. Nelson: *Hydrogen Plant Steels*, Proc. API 29 M (III) (1949) 163.
303. W. A. Bonner, *Oil Gas J.* **77** (1977, July 25) 172.
304. F. H. Vitovec: *The Groth Rate of Fissures During Hydrogen Attack of Steels*, Proc. API 44 III 179 (1964).
305. W. Gayk: "Werkstoffauswahl bei Heißgasbeanspruchung," *DECHEMA Monogr.* **87** (1980) 73.
306. B. A. Hands (ed.): *Cryogenic Engineering*, Academic Press, London 1986.
307. D. Ervans, J. T. Morgan in [306] p. 271.
308. R. V. Homsy, C. A. Glaton, *Review of Literature on Catalytic Recombination of Hydrogen Oxygen*, ORNTL-MT 55, May 1968.
309. M. P. Sherman: *The Behaviour of Hydrogen During Accidents in Light Water Reactors*, Sandia Laboratories, NUREG/CR-1561, August 1980.
310. H. Karwat: *Hydrogen Mitigation*, Report at CEC Brussel, Belgium 1982.
311. L. B. Thompson, J. J. Haugh, R. G. Zalosh, *Nucl. Saf.* **25** (1984) 350.
312. R. Heck, *Atomkernenerg. Kerntech.* **53** (1988) no. 1, 56.
313. NASA Lewis Research Center, *Hydrogen Safety Manual*, NASA Tech. Memo TM-X 52 454 (1968).
314. J. Hord, *Int. J. Hydrogen Energy* **3** (1978) 157.
315. K. M. Kalyanam, D. R. Hay: *Safety Guide for Hydrogen*, National Research Council Hydrogen Safety Committee, Ottawa 1987.
316. R. L. Custer, R. G. Bright: *Fire Detection*, State of the Art, NASA CR 134 642, NBS TN 839 June 1974.
317. H. Eichert, M. Fischer, *Int. J. Hydrogen Energy* **11** (1986) 117.
318. W. Balthasar, J. W. M. van Rijnsoever: *Hydrogen is Safe*, KTI Publication, Zoetermeer, NL 1981.
319. K. Baumgärtner, R. Ewald, *Chem. Ing. Tech.* **59** (1987) 296.
320. J. Hord, *Int. J. Hydrogen Energy* **5** (1980) 579.
321. Occupational Safety and Health Act of 1970, PL 91–596, in United States Statutes at Large, *91st Congress*, Second Session 84, Part 2, US Government Printing Office, Washington, DC 1971, p. 1590.
322. C. Gentry, R. Schott, *Anal Chem.* **42** (1970) 7.
323. M. Riederer, W. Sawodny, *J. Chromatogr.* **179** (1979) 337. Links
324. K. Fujita, T. Kwan, *Int. Chem. Eng.* **8** (1968) 552.
325. C. Laycock, B. Foxall, *Iron and Steel Int.* (1978, Aug.) 223.
326. J. Plessers, R. Maes, E. Vangelooven, *Stahl Eisen* **108** (1988) 85.
327. Bruker Analytische Messtechnik, Company Brochure, *Bruker Minispec Application Notes*, Karlsruhe 1987.
328. J. Schulze, H. Gaenssten, *Chem. Ind.* **36** (1984) no. 3, 135 – 140; no. 4, 202 – 208. M. A. S. Guth, *Int. J. Hydrogen Energy* **10**, (1985) no. 1, 1 – 10.
329. H. Kalenda, G. Ruckelshauss, (Bundesministerium für Forschung und Technologie) – Forschungsbericht T 81–012 (1981).
330. G. Kaske: *Wasserstofftechnologie*, DECHEMA Studie, Frankfurt/Main 1986, p. 488.
331. H. G. Eickhoff, D. Martinsen, M. Walbeck, *Int. J. Hydrogen Energy* **9** (1984) 232.
332. S. Kasper, *Int. J. Hydrogen Energy* **8** (1983) 183.
333. *Chem. Bus. Suppl. of Chem. Market. Reporter* **11** (1985) 13.
334. M. A. S. Guth, *Int. J. Hydrogen Energy* **10** (1985) 1.
335. Y. Osumi, *CEER Chem. Econ. Eng. Rev.* (1985) June no. 5.
336. E. Wicke, *Chem. Ing. Tech.* **54** (1982) 41.
337. M. Fischer, G. Kreysa, G. Sandstede, *Int. J Hydrogen Energy* **12** (1987) 39.
338. W. Van Deelen: *Proc. 3rd World Energy Conference*, vol. **4**, Tokyo 1980, p. 2323.
339. J. G. Santangelo, G. T. Chen: "Hydrogen Recovery with Metal Hydrides," *ACS Symposium on Advances in Hydrogen Manufacture*, Los Angeles 1982.
340. T. Grundt, K. Christiansen, *Int. J. Hydrogen Energy* **7** (1982) 247.
341. L. Green Jr. *Int. J. Hydrogen Energy* **7** (1981) 335.
342. H. L. Blomelburg, *Plant Oper. Prog.* **1** (1982) 175.
343. H. F. Mason, R. S. Tolberg, *AIChE Symp. Ser.* **148** (1975) no. 71, 220.

344. H. J. Satrina (ed.): *Hydroprocessing Catalyst for Heavy Oil and Coal*, Noyes Data Corp., Park Ridge, NJ, 1982.
345. International Refining Catalyst Compilation 1987, *Oil Gas J.* **85** (1987, Oct. 5) 41.
346. 1986 Refining Process Handbook, *Hydrocarbon Process.* **65** (1986) no. 9, 83.
347. R. A. Meyers (ed.): *Handbook of Petroleum Refining Processes*, McGraw Hill, New York 1986.
348. M. Kuno, *Hydrocarbon Process.* **60** (1981) 124.
349. B. Schütze, H. Hofmann, *Erdöl Kohle Erdgas Petrochem.* **36** (1983) 457.
350. W. Krönig, Kohlehydrierung in J. Falbe (ed.): *Chemierohstoffe aus Kohle*, Thieme Verlag, Stuttgart 1977.
351. E. E. Donath, History of Catalysis in Coal Liquefaction, in J. R. Anderson, M. Bondart (eds.): *Catalysis*, vol. **3**, Springer Verlag, Berlin – Heidelberg – New York 1982, p. 1.
352. P. Nowacki (ed.): Coal Liquefaction Processes, Noyes Data Corp., New Jersey 1979, pp. 76 – 159.
353. S. Ueda, Y. Miki, M. Naka, M. Wada, Status of Synfuel Development in Japan, *19th IECEC*, San Francisco, CA, 1984.
354. K. H. van Heek, W. Wanzl, *Erdöl Erdgas* **105** (1989) 30.
355. R. Fladerer, L. Schrader, *Chem. Ing. Tech.* **54** (1982) 884.
356. J. Falbe (ed.): *Chemierohstoffe aus Kohle*, Thieme Verlag, Stuttgart 1977.
357. J. Fahey, “Industrial Chemicals via C_1-Processes,” *ACS Symp. Ser.* **328** (1986) .
358. J. Falbe (ed.): *New Syntheses with Carbon Monoxide*, Springer Verlag, Berlin – Heidelberg – New York 1980.
359. F. Asinger (ed.): *Methanol, Chemie- und Energierohstoff*, Springer Verlag, Berlin – Heidelberg – New York – London – Paris – Tokyo 1986.
360. C. D. Frohning et al. in [357], Chap. 8, pp. 219 – 299.
361. Ch. Masters, *Adv. Organomet. Chem.* **17** (1979) 61 – 103. Links
362. M. A. Vanice, *Catal. Rev. Sci. Eng.* **14** (1976) 153.
363. H. C. Anderson, J. L. Wiley, A. Newall, *Bibliography of the Fischer – Tropsch – Syntheses*, Bull. 544, Bureau of Mines, United States Goverment Printing Office, Washington 1954.
364. M. E. Dry, *CHEMTECH* **12** (1982) 744.
365. V. U. S. Rad, R. J. Germly, *Hydrocarbon Process.* **59** (1980) no. 11, 139.
366. L. König, J. Gaube, *Chem. Ing. Tech.* **55** (1983) 14.
367. C. D. Frohning, Polymethylene Synthesis in J. Falbe (ed.): *New Syntheses with Carbon Monoxide*, Springer Verlag, Berlin – Heidelberg – New York 1980.
368. F. W. Koeller, H. Roberts, B. Britz, *Hydrocarbon Process.* **53** (1974) no. 4, 69.
369. B. Cornils: “Hydroformylation, Oxo Synthesis, Roelen Reaction” in J. Falbe (ed.): *New Syntheses with Carbon Monoxide*, Springer Verlag, Berlin – Heidelberg – New York 1980.
370. H. Bahrmann, B. Cornils, *Chem. Ztg.* **104** (1980) 39.
371. R. L. Augustine, “Organic Functional Group Hydrogenation,” *Catal. Rev. Sci. Eng.* **13** (1976) 285.
372. M. Freifelder (ed.): *Practical Catalytic Hydrogenation*, Wiley Interscience, New York 1971.
373. M. Freifelder (ed.): *Catalytic Hydrogenation in Organic Synthesis, Procedures and Commentary*, Wiley Intercience, New York 1978.
374. R. J. Peterson (ed.): *Hydrogenation Catalysts*, Noyes Data Corporation, Park Ridge, NJ 1977.
375. R. Albrecht, A. Strätz, G. Vollheim, *Chem. Ing. Tech.* **52** (1980) 582.
376. R. Steffen, H. B. Lüngen, *Stahl Eisen* **108** (1988) 339.
377. E. Mazank, A. Brachany, C. Bryk, *Stahl Eisen* **102** (1982) 3.
378. F. Habashi, R. Dugdale, N. Nagmori, *Metall (Berlin)* **28** (1974) 1051.
379. Rockwell Int. Corp., DE-OS 30 428 878, 1981 (J. Friedmann *et al.*).
380. W. Pötke, R. Jeschdar, G. Kehse, *Stahl Eisen* **108** (1988) 581.
381. U. J. Gährs, P. Rutten, G. Schnorr, *Gas Aktuell*, Messer Griesheim brochure, Frankfurt 1987.
382. U. Gieger *et al.*, *Cryogenics* **11** (1976) 451.
383. P. Hoffmann: *The Forever Fuel*, Westview Press, Boulder, CO, 1981.
384. G. M. Woodwell, *Sci. Am.* **238** (1978) no. 1, 34 – 44. Links
385. C. D. Keeling *et al.*: “Atmospheric Carbon Dioxide Variations. Hawaii,” *Tellus* **28** (1976) no. 6, 538 – 551.
386. J. O. M. Bockris: *Energy: The Solar Hydrogen Alternative*, Wiley & Sons, New York 1980.
387. W. Bach *et al.* (eds.): *Renewable Energy Prospects*, Pergamon Press, Oxford 1980.
388. H. E. Landsberg, *Sol. Energy* **5** (1961) 95 – 98.
389. C. Voigt in C. J. Winter, J. Nitsch (eds.): *Wasserstoff als Energieträger*, Springer Verlag, Berlin 1986, p. 301.

390. H. Tributsch in H. Scheer (ed.): *Die gespeicherte Sonne*, Piper, München 1987, pp. 119 – 155.
391. J. Cotillon, *Houille Blanche* **37** (1982) 381 – 390.
392. C. Voigt in [389], p. 305.
393. M. Sappa, *Proc. 3rd World Hydrogen Energy Conf.*, vol. **3**, Tokyo 1980, pp. 1373 – 1406.
394. R. L. Le Roy: *The Metallurgical Society of CIM*, annu. vols. 1 – 10, 1978.
395. F. Conrad, *Energie* **40** (1988) no. 6, 46 – 53.
396. H. Wendt, *Chem. Ing. Tech.* **56** (1984) 265 – 272.
397. W. Schnurnberger, W. Seeger, H. Steeb in [389], pp. 205 – 243.
398. W. Doenitz, R. Schmidberger, *Int. J. Hydrogen Energy* **7** (1982) 321 – 330.
399. C. Carpetis, W. Peschka, *Int. J. Hydrogen Energy* **5** (1980) 539.
400. J. S. Noh, R. K. Agarwal, J. A. Schwarz, *Int. J. Hydrogen Energy* **12** (1987) 693.
401. G. Alefeld, J. Völkl (eds.): *Hydrogen in Metals*, vol. I and vol. II. Topics and Applied Physics, vols. 28, 29, Springer Verlag, Berlin – Heidelberg – New York 1978.
402. G. G. Libowitz, G. D. Sandrock (eds.): *Proceedings and Applications of Metal Hydrides*, vols. 1 and 2, Elsevier, Amsterdam 1980.
403. H. Wenzl, *Int. Met. Rev.* **27** (1982) 140.
404. G. D. Sandrock, S. Suda, *** Applications in L. Schlapbach (ed.): *Hydrogen in Intermetallic Compounds*, Springer Verlag, Berlin – Heidelberg – New York 1986.
405. F. A. Kuijpers, H. H. VanMal, *J. Less-Common Met.* **23** (1971) 395.
406. P. D. Goodell, G. D. Sandrock, E. L. Huston, *J. Less-Common Met.* **73** (1980) 135.
407. M. J. Song, M. Pezat, B. Darrier, P. Hagenmüller, *J. Solid State Chem.* **56** (1985) 191. Links
408. O. Bernauer, C. Halene, *J. Less-Common Met.* **131** (1987) 213.
409. HWT Gesellschaft für Hydrid- und Wasserstofftechnik mbH, Mülheim/Ruhr, company brochure 1986.
410. H. Buchner, *Int J. Hydrogen Energy* **3** (1981) 259.
411. F. B. Hill, Y. W. Wong, Y. N. I. Chan, *AIChe J.* **28** (1982) 1. Links
412. J. G. Santangelo, G. T. Chen, *Hydrogen Recovery with Metal Hydrides*, ACS Symposium on Advances in Hydrogen Manufacture, Las Vegas 1982.
413. E. L. Houston, *Development of a Commercial Metal Hydride Process for Hydrogen Recovery*, Final Report vols. 1 and 2, MPD Technology Corp., Wyckoff, New Jersey, 1984.
414. L. Schlapbach, M. Groll, *Metal Hydrides for Energy Technologies*, IEA Report Task VII, Storage, Conversion and Safety, March 1986.
415. Ergenics Inc., USA; company brochure.
416. M. Taube *et al.*, *Int. J. Hydrogen Energy* **10** (1985) 595.
417. G. Giacomazzi, CEC Joint Research Centre, Report EUR-11 330 – EN, Ispra (Italy) 1986.
418. K. Hoyermann, H. Janaer, H. G. Wagner, *Ber. Bunsenges. Phys. Chem.* **79** (1975) 858.
419. J. Pangborn, M. C. Scott, *Domestic Uses of Hydrogen in* K. E. Cox, K. D. Williamson, Jr. (eds.): *Hydrogen: Its Technology and Implications*, CRC-Press, Boca Raton, Florida 1979, p. 151.
420. *Einsatzmöglichkeiten neuer Energiesysteme*, Programmstudie des BMFT, part III: Wasserstoff, Bonn 1975.
421. M. Haruta, H. Sano, *Int. J. Hydrogen Energy* **7** (1982) 729, 737, 801.
422. J. L. Sloop, *Liquid Hydrogen as a Propulsion Fuel*, The NASA History Series, BASA-SP-4404, TL 785 S 58, 1978.
423. H. Dederra, E. Kirner, MBB: Vulcain HM60-Thrust Chamber Critical Development Items, 15th ISTS-Congress, Tokyo, May 1986.
424. D. Brewer, *Astronaut. Aeronaut.* **12** (1974) 40.
425. P. Hofmann *The forever Fuel. Hydrogen as Aircraft Fuel*, p. 139, Westview Press, Boulder, CO, 1982.
426. International Symposium (DGLR mit DFVLR) *Hydrogen in Air Transportation*, Sept. 11th to 14th, 1979, Stuttgart.
427. G. D. Brewer, *Int. J. Hydrogen Energy* **7** (1982) 21.
428. E. Reshoto, *J. Aircr.* **16** (1979) 584.
429. G. D. Brewer (ed.): NASA Contractor Report CR 2700, Washington, D.C., 1976.
430. H. P. Alder, *Int. J. Hydrogen Energy* **12** (1987) 571.
431. W. Peschka, *Flüssiger Wasserstoff als Energieträger*, Springer Verlag, Wien – New York 1984, pp. 129 – 184.
432. H. Buchner, *Int. J. Hydrogen Energy* **9** (1984) 501.
433. P. Hofmann: *The Forever Fuel: The Story of Hydrogen*, Westview Press, Boulder, CO, 1982, pp. 105 – 138.
434. BMW AG, *Hydrogen Vehicle, company brochure*, Munich 1986.
435. BMFT Studie, *Neuen Kraftstoffen auf der Spur, Bonn* 1974.

436. L. A. Sinclair, J. S. Wallace, *Int. J. Hydrogen Energy* **12** (1984) 123.
437. H. S. Homan, P. C. T. De Boer, W. J. McLean, *Int. J. Hydrogen Energy* **8** (1983) 131.
438. K. Feucht, R. Povel, W. Gelse, "Wasserstoff-Energietechnik," *VDI Ber.* **602** (1987) 185.
439. G. Withalm, W. Gelse, *The Mercedes Benz Hydrogen Engine for Application in a Vehicle Fleet*, VI World Hydrogen Energy Conference, Vienna 1986.
440. W. Strobl, W. Peschka, "Wasserstoff-Energietechnik", *VDI-Ber.* **602** (1987) 203.
441. H. J. Sternfed, H. Wojkowsky, W. Schnurrenberger, *Brennst. Wärme Kraft* **34** (1982) 351, 401.
442. H. J. Sternfeld, K. Wolfmüller, *VAB Kraftwerkstech.* **66** (1986) 793.
443. H. Liebhafsky, E. J. Cairns: *Fuel Cells and Fuel Cell Batteries*, Wiley, New York 1968.
444. K. K. Ushiba, *CHEMTECH* **14** (1984) 300.
445. H. Wendt, W. Jenseit, *Chem. Ing. Tech.* **60** (1988) 180.
446. K. Kordesch, J. C. T. Olivera, *Int. J. Hydrogen Energy* **13** (1988) 411.
447. W. Dönitz, *State of the Art and Future Developments of Solid Oxide Electrolyte Fuel Cells*, Proceedings of the UNESCO Workshop on Fuel Cells, June 1985.
448. A. P. Fickett, *Int. J. Hydrogen Energy* **8** (1983) 617.
449. S. Knable in E. A. Gillis, (ed.): Fuel Cell Seminar 1986, Program and Abstracts, Tucson Ariz., 1986, pp. 27.
450. N. Itoh, Science & Technology in Japan **5** (1986) no. 18, 16.
451. N. Getoff: *Wasserstoff als Energieträger*, Springer Verlag, Wien – New York 1977.
452. D. Behrens (ed.): *Wasserstofftechnologie. Perspektiven für Forschung und Entwicklung*, DECHEMA, Frankfurt 1986.
453. K. E. Cox, K. D. Williamson (eds.): *Hydrogen, its Technology and Implications*, vols. I – V, CRS Press, 1979.
454. J. Nitsch, C. Voigt in [389], pp. 288 – 310
455. R. Aureille, J. Pottier, *Proc. 5th World Hydrogen Energy Conf.*, vol. **1**, Toronto 1984, pp. 279 – 288.
456. A. G. Barnstaple, A. J. Petrella, Report no. 81 105 and 82 452, Ontario Hydro, Toronto 1981, 1982.
457. G. C. Crawford, A. K. Stuart, *Int. J. Hydrogen Energy* **9** (1984) no. 7, 619 – 625.
458. R. H. Shannon, R. D. Richardson, *Int. J. Hydrogen Energy* **8** (1983) no. 10, 783 – 792.
459. A. J. Petrella, D. Chisholm, A. G. Barnstaple, *Hydrogen study*, vol. 4, Report no. 82 645, Ontario Hydro, Toronto 1982.
460. J. Nitsch, C. Voigt in [389], p. 358.
461. T. Donakowski, C. Blazek, M. Novil, E. Daniels, *Int. J. Hydrogen Energy* **9** (1984) no. 7, 613 – 618.
462. *Hysolar—Solar Hydrogen Production and Utilization*, Projektvereinbarung Riad/Bonn, March 1986.
463. R. Meggle, *DECHEMA Monogr.* **112** (1988) 325 – 534.
464. J. Gretz, J. P. Baselt, O. Ullmann, H. Wendt: "The 100 MW Euro-Quebec Hydro-Hydrogen Pilot Project," *VDI-Rep.* **725** (1989) 81 – 91.
465. C. Marchetti, *Int. J. Hydrogen Energy* **10** (1985) no. 4, 215 – 219.
466. *Solare Wasserstoffenergiewirtschaft*, BMFT-Ausschuß, Bonn, Feb. 1988.
467. International Energy Agency, *Program of Work 1986 – 1989*, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt, Köln 1986.
468. *New Uses for H_2 in Canadian Energy Systems*, International Energy Agency, Task VII reports, Nov. 1985.
469. *The Sunshine Project*, Agenca of Ind. Sci. Tech. of MITI, Tokyo 1987. 1986 Annual Summary of Hydrogen Energy R + D, Japan Industrial Technology Association, Tokyo 1987.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007 John Wiley & Sons, Inc. All Rights Reserved.

Hydrogen Peroxide

Gustaaf Goor, Degussa AG, Hanau, Germany (Chap. 1, 2, 3, 4, 5, 6 and 7)
Jürgen Glenneberg, Degussa AG, Hanau, Germany (Chap. 1, 2, 3, 4, 5, 6 and 7)
Sylvia Jacobi, Degussa AG, Hanau, Germany (Chap. 8)

1. Introduction and Historical Aspects . . . 1
2. Physical Properties . . . 3
3. Chemical Properties . . . 5
4. Production . . . 6
4.1. Anthraquinone Process (AO Process) . . . 6
4.1.1. Principles . . . 6
4.1.2. Process Description . . . 9
4.1.2.1. Hydrogenation . . . 10
4.1.2.2. Oxidation . . . 14
4.1.2.3. Extraction and Drying . . . 15
4.1.2.4. Working Solution Purification and Regeneration . . . 15
4.1.2.5. Purification of Crude Hydrogen Peroxide . . . 16
4.1.2.6. Concentration of Hydrogen Peroxide . . . 17
4.2. 2-Propanol Process (Shell Process) 17
4.3. Electrochemical Processes . . . 18
4.3.1. Degussa–Weissenstein Process . . . 19
4.3.2. Münchner Process . . . 20
4.4. Other Processes . . . 20
4.4.1. Production from Peroxy Compounds 20
4.4.2. Production from Hydrogen and Oxygen . . . 20
4.4.3. Production from Carbon Monoxide, Oxygen, and Water . . . 22
4.4.4. Production by Autoxidation of Organic Compounds . . . 22
4.4.5. Production by Cathodic Reduction of Oxygen . . . 22
5. Storage and Transportation . . . 23
6. Safety . . . 24
7. Uses . . . 25
8. Toxicology and Occupational Health . . . 26
9. Environmental Effects . . . 28
9.1. Atmospheric Fate . . . 28
9.2. Aquatic Fate . . . 29
9.3. Biodegradation . . . 30
9.4. Environmental Releases . . . 30
9.5. Environmental Effects . . . 30
10. References . . . 31

This review deals with chemical and physical properties of hydrogen peroxide and gives detailed descriptions about the various production processes. Today, almost all hydrogen peroxide is produced by organic autoxidation processes, primarily the anthraquinone process. Other aspects in this review are handling, safety and usage of hydrogen peroxide. Hydrogen peroxide occurs naturally. Because of its chemical reactivity, hydrogen peroxide also has toxic properties. The formation and degradation of atmospheric hydrogen peroxide is photochemically mediated. Degradation of aquatic hydrogen peroxide is catalyzed by biological material or by transition metals. Toxicity data are collected for various organisms.

1. Introduction and Historical Aspects

The industrial production of hydrogen peroxide passed through three phases, starting with wet chemical processes, followed by electrochemical processes, and then by organic autoxidation processes. Almost all hydrogen peroxide is now produced by organic autoxidation (AO) processes, primarily the anthraquinone process.

Wet Chemical Processes. In 1818, L. J. Thenard [9] obtained hydrogen peroxide for the first time by treating barium peroxide [*1302-29-6*] with nitric acid. This process was improved by using hydrochloric acid to release hydrogen peroxide. The water-soluble barium chloride which

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_443.pub2

is formed simultaneously, was precipitated with sulfuric acid.

$$BaO_2 + 2\,HCl \rightarrow BaCl_2 + H_2O_2$$

$$BaCl_2 + H_2SO_4 \rightarrow BaSO_4 + 2\,HCl$$

$$BaO_2 + H_2SO_4 \rightarrow BaSO_4 + H_2O_2$$

With this reaction, THENARD established the foundation for the commercial manufacture of aqueous hydrogen peroxide solutions based on wet chemical processes, which began around 1880.

Industrial plants using the barium peroxide process were still operating until the middle of the 20th century. Around 1900, approximately 10 000 t/a of barium peroxide was processed by this technology, yielding ca. 2000 t/a of hydrogen peroxide. Sales opportunities for the coproduct barium sulfate ($BaSO_4$, blanc fixe) had a decisive effect on the profitability of the process.

The 3 % aqueous hydrogen peroxide solutions manufactured by the barium peroxide process had only a limited market because of their high production cost, low hydrogen peroxide content, and unsatisfactory stability due to impurities.

Electrochemical Processes. The introduction of electrochemical processes eliminated the disadvantages of the barium peroxide process.

In 1853, MEIDINGER discovered the formation of hydrogen peroxide during electrolysis of aqueous sulfuric acid [10]. In 1878, BERTHELOT showed that peroxodisulfuric acid [*13445-49-3*] is formed during this electrolysis and then hydrolyzed by water to give sulfuric acid and hydrogen peroxide via peroxomonosulfuric acid [*12188-01-1*] [11]:

$$2\,H_2SO_4 \rightarrow H_2S_2O_8 + H_2$$

$$H_2S_2O_8 + H_2O \rightarrow H_2SO_5 + H_2SO_4$$

$$H_2SO_5 + H_2O \rightarrow H_2SO_4 + H_2O_2$$

$$2\,H_2O \rightarrow H_2O_2 + H_2$$

In 1905, following the work of ELBS and SCHOENHERR [12], G. TEICHNER developed an electrolytic process at the Konsortium für Elektrochemische Industrie. The first hydrogen peroxide plant based on this new technology went on stream in 1908 at the Österreichische Chemische Werke in Weissenstein. The *Weissenstein process* was followed in 1910 by the *Münchner process* developed by A. PIETZSCH and G. ADOLPH at the Elektrochemische Werke, Munich, and then in 1924 by the *Riedel–Loewenstein process* developed by L. LOEWENSTEIN and first used by Riedel-de-Haen. In these processes, an ammonium sulfate solution was electrolyzed instead of free sulfuric acid, and the resulting ammonium peroxodisulfate (Riedel–Loewenstein) or potassium peroxodisulfate derived therefrom (Pietzsch–Adolph) was hydrolyzed. After introduction of these processes, production (calculated as 100 % hydrogen peroxide) increased steadily and, in 1950, reached approximately $(30–35) \times 10^3$ t/a [13].

Organic Autoxidation Processes. The decisive breakthrough in industrial production of hydrogen peroxide, which enabled the construction of modern large-scale plants, came with the development of organic autoxidation processes, especially the anthraquinone process.

In 1901, MANCHOT discovered that hydroquinones (or hydrazobenzenes) react quantitatively with oxygen to form quinones (or azobenzenes) and hydrogen peroxide [14]. In 1932, WALTON and FILSON in the United States went back to this work and proposed a cyclic process for the manufacture of hydrogen peroxide based on azobenzene–hydrazobenzene [15].

PFLEIDERER of I. G. Farbenindustrie AG, Ludwigshafen (BASF), subsequently developed a process for the alkaline autoxidation of hydrazobenzene in which sodium peroxide is obtained instead of free hydrogen peroxide [16]. This process was used on an industrial scale by Kymmene AB in Kuusankoski (Finland) under BASF licence (see Section 4.4.4.

The azobenzene process had two major technical drawbacks: hydrogenation of azobenzene was performed with sodium amalgam, and oxidation of hydrazobenzene proceeded satisfactorily only in alkaline solution. PFLEIDERER and RIEDEL overcame these problems by using alkylated anthraquinones instead of azobenzene. Between 1935 and 1945, BASF developed the Riedel–Pfleiderer process (generally referred to as the anthraquinone or AO process) in a pilot plant with a monthly output of 30 t. At the

end of World War II, this work was stopped in Germany because of Allied Control Commission Regulations, but it formed the basis for all present production processes. In 1953, Du Pont of Memphis, Tennessee, commissioned the first hydrogen peroxide production plant based on the AO process. Several other companies began to produce hydrogen peroxide by this process, and production capacity increased greatly. In 2002, the global production capacity excluding China was estimated to be 2.8×10^6 t/a (calculated as 100 % H_2O_2) [17].

Another organic autoxidation process, the 2-propanol process, used by Shell in Norco (United States) between 1957 and 1980, made only a limited contribution to the increase of hydrogen peroxide production with an annual production capacity of about 15 000 t.

Various patents on direct combination of hydrogen and oxygen have been published but up to now, the direct synthesis of hydrogen peroxide has not been implemented on a large industrial scale.

2. Physical Properties

Hydrogen peroxide [*7722-84-1*], M_r 34.016, is a clear, colorless liquid which is miscible with water in all proportions. Hydrogen peroxide and highly concentrated aqueous solutions (> 65 wt %) are soluble in a variety of organic solvents such as carboxylic esters.

Hydrogen peroxide and water do not form an azeotropic mixture and, in theory, can be separated completely by distillation. In practice, 100 wt % hydrogen peroxide is obtained by fractional crystallization of highly concentrated (ca. 90 wt %) aqueous solutions. Pure 100 wt % hydrogen peroxide is stable at room temperature. It is of scientific interest only and is not produced on an industrial scale.

Data pertaining to molecular structure (obtained by neutron diffraction on solid H_2O_2) are as follows [18]:

Bond length, O—O	145.3 ± 0.7 pm
Bond length, O—H	99.8 ± 0.5 pm
Bond angle, O–O–H	102.7 ± 0.3°
Azimuthal angle	90.2 ± 0.6°

Table 1. Physical properties of hydrogen peroxide and water

Property	Value	
	H_2O_2	H_2O
mp, °C	−0.43	0
bp (101.3 kPa), °C	150.2	100
Heat of fusion, J/g	368	334
Heat of vaporization, J g^{-1} K^{-1}		
At 25 °C	1519	2443
At *bp*	1387	2258
Specific heat, J g^{-1} K^{-1}		
Liquid (25 °C)	2.629	4.182
Gas (25 °C)	1.269	1.865
Relative density, g/cm^3		
0 °C	1.4700	0.9998
20 °C	1.4500	0.9980
25 °C	1.4425	0.9971
Viscosity, mPa·s		
0 °C	1.819	1.792
20 °C	1.249	1.002
Critical temperature, °C	457	374.2
Critical pressure, MPa	20.99	21.44
Refractive index n_D^{20}	1.4084	1.3330

A discussion of the molecular structure can be found in [19].

Table 2. Physical properties of aqueous hydrogen peroxide solutions

Property	H_2O_2 concentration, wt %			
	35	50	70	90
Relative density				
0 °C	1.1441	1.2110	1.3071	1.4136
20 °C	1.1312	1.1953	1.2886	1.3920
25 °C	1.1282	1.1914	1.2839	1.3867
Viscosity, mPa · s				
0 °C	1.82	1.87	1.93	1.88
20 °C	1.11	1.17	1.23	1.26
Refractive index				
n_D^{20}	1.3563	1.3672	1.3827	1.3995
mp, °C	−33	−52.2	−40.3	−11.9
bp (101.3 kPa), °C	107.9	113.8	125.5	141.3
H_2O_2 partial pressure (30 °C), kPa	0.05	0.11	0.17	0.29

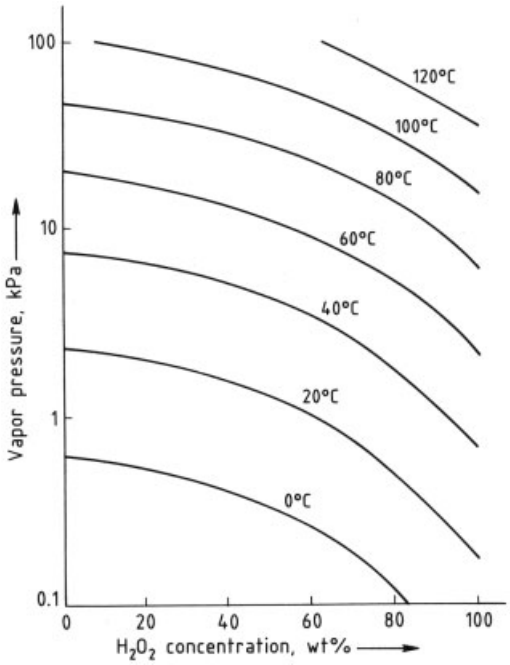

Figure 1. Vapor pressure of aqueous hydrogen peroxide solutions

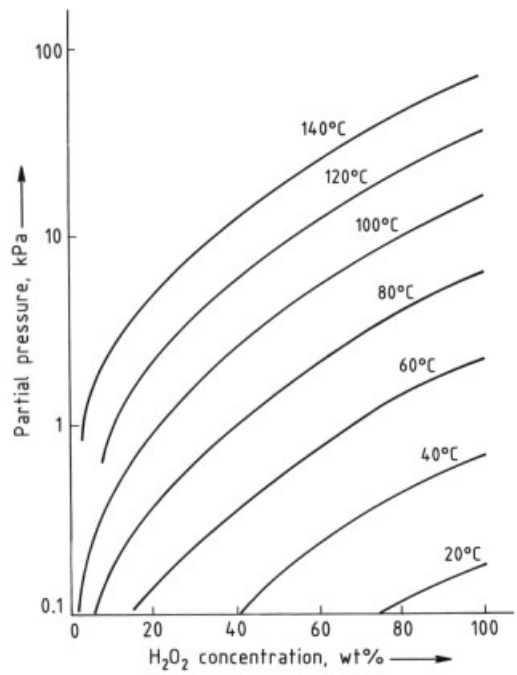

Figure 2. Partial pressure of hydrogen peroxide over aqueous hydrogen peroxide solutions

Physical constants of hydrogen peroxide and water are compared in Table 1. For a more detailed comparison, see [20]. Physical constants of some aqueous hydrogen peroxide solutions in commercially available concentrations are listed in Table 2.

The vapor pressure and partial pressure of aqueous hydrogen peroxide solutions are shown as a function of temperature in Figures 1 and 2, respectively. Figure 3 shows the vapor–liquid equilibrium curve for aqueous hydrogen peroxide solutions [21].

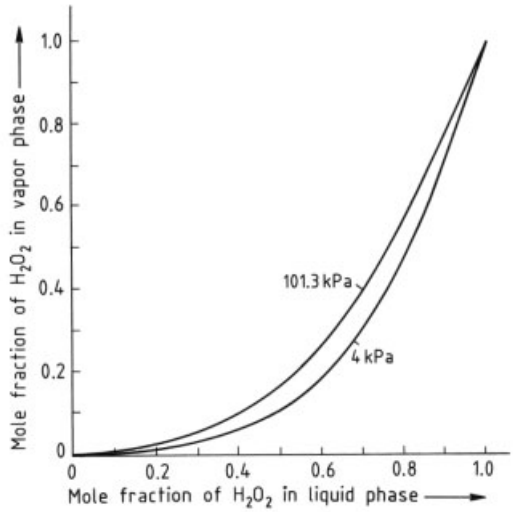

Figure 3. Vapor–liquid equilibrium curve for water–hydrogen peroxide

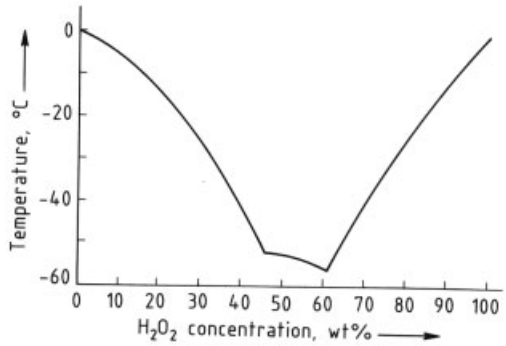

Figure 4. Freezing point curve for water–hydrogen peroxide

The solid–liquid phase diagram (Fig. 4) shows eutectic points for the mixtures

ice–$H_2O_2 \cdot 2\ H_2O$ at 45.2 wt % hydrogen peroxide and for solid H_2O_2–$H_2O_2 \cdot 2\ H_2O$ at 61.2 wt % hydrogen peroxide with a congruent melting point for the compound $H_2O_2 \cdot 2\ H_2O$ between them.

Heat of formation and of decomposition of hydrogen peroxide are:

$H_2(g) + O_2(g) \Rightarrow H_2O_2(g)$	−136.2 kJ/mol
$H_2(g) + O_2(g) \Rightarrow H_2O_2(l)$	−187.9 kJ/mol
$H_2O_2(g) \Rightarrow H_2O\ (g) + 1/2\ O_2(g)$	−105.8 kJ/mol
$H_2O_2(l) \Rightarrow H_2O\ (l) + 1/2\ O_2(g)$	−98.3 kJ/mol

3. Chemical Properties

Dissociation. Hydrogen peroxide is weakly acidic in aqueous solution, with a dissociation constant of 1.78×10^{-12} (pK 11.75) at 20 °C. As a weak acid, hydrogen peroxide forms salts with various metals.

Oxidation and Reduction. Hydrogen peroxide can behave both as an oxidizing and as a reducing agent (Tables 3 and 4). Systems with a redox potential $E_0 < -1.80$ V at pH 0 cannot be oxidized by hydrogen peroxide; systems with a redox potential $E_0 > -0.66$ V at this pH cannot be reduced by hydrogen peroxide [6, 22].

Table 3. Hydrogen peroxide as an oxidizing agent [20]

Redox reaction	Standard potential E_0, V *
pH 0	
$HOOH + 2\ H^+ + 2\ e^- \Rightarrow 2\ HOH$	+1.80
$HSO_3^- + HOH \rightarrow SO_4^{2-} + 3\ H^+ + 2\ e^-$	−0.17
$NO_2^- + HOH \rightarrow NO_3^- + 2\ H^+ + 2\ e^-$	−0.94
$2\ Cl^- \Rightarrow Cl_2 + 2\ e^-$	−1.36
$2\ Br^- \Rightarrow Br_2 + 2\ e^-$	−1.07
$2\ I^- \Rightarrow I_2 + 2\ e^-$	−0.54
pH 14	
$HOOH + 2\ e^- \Rightarrow 2\ OH^-$	+0.87
$Mn(OH)_2 + 2\ OH^- \Rightarrow MnO(OH)_2 + HOH + 2\ e^-$	+0.05

* Standard potential of redox reactions measured against a hydrogen electrode (25 °C, 100 kPa).

The reactions of hydrogen peroxide with potassium permanganate [23] or cerium(IV)sulfate are used for quantitative determination (volumetric titration) of hydrogen peroxide content.

Table 4. Hydrogen peroxide as a reducing agent [20]

Redox reaction	Standard potential E_0, V *
pH 0	
$HOOH \Rightarrow 2\ H^+ + O_2 + 2\ e^-$	− 0.66
$5\ e^- + MnO_4^- + 8\ H^+ \rightarrow Mn^{2+} + 4\ HOH$	+ 1.51
$1\ e^- + Ce^{4+} \Rightarrow Ce^{3+}$	+ 1.61
pH 14	
$HOOH + 2\ OH^- \Rightarrow 2\ HOH + O_2 + 2\ e^-$	+ 0.08
$1\ e^- + ClO_2 \rightarrow ClO_2^-$	+ 1.16
$2\ e^- + ClO^- + HOH \Rightarrow Cl^- + 2\ OH^-$	+ 0.89

* Standard potential of redox reactions measured against a hydrogen electrode (25 °C, 100 kPa).

Decomposition of hydrogen peroxide occurs with disproportionation

$$H_2O_2 \rightarrow H_2O + 1/2\ O_2$$

and is extremely important in handling hydrogen peroxide during storage and in the laboratory. This reaction is highly exothermic (see Chapter 2) and takes place in the presence of small amounts of catalyst even in aqueous solution. In the absence of catalyst, it occurs only in the gas phase at high temperature.

Decomposition can be catalyzed both homogeneously by dissolved ions (especially of the metals iron, copper, manganese, and chromium) and heterogeneously by suspended oxides and hydroxides (e.g., those of manganese, iron, copper, palladium, or mercury) and by metals such as platinum, osmium, and silver.

Substitution. The hydrogen atoms of hydrogen peroxide can be substituted by alkyl and acyl groups, leading to the formation of

H−O−O−R	alkyl hydroperoxides
R−O−O−R	dialkyl peroxides
H−O−O−Ac	percarboxylic acids
Ac−O−O−Ac	diacyl peroxides

Hydrogen peroxide forms peroxohydrates with a number of compounds. Addition compounds with sodium carbonate (sodium carbonate peroxohydrate [*15630-89-4*]) and with urea (urea peroxohydrate [*124-43-6*]) are industrially important. For a detailed description of peroxo compounds see → Peroxo Compounds, Inorganic; → Peroxy Compounds, Organic.

4. Production

4.1. Anthraquinone Process (AO Process)

4.1.1. Principles

In the AO process, 2-alkyl-9,10-anthraquinones react with hydrogen in the presence of a catalyst to form the corresponding hydroquinones.

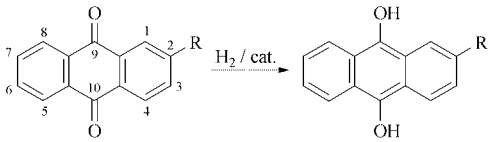

R = alkyl

After the catalyst is removed (otherwise, the hydrogen peroxide would decompose), the hydroquinones are oxidized to quinones with oxygen (usually air) with simultaneous quantitative formation of hydrogen peroxide:

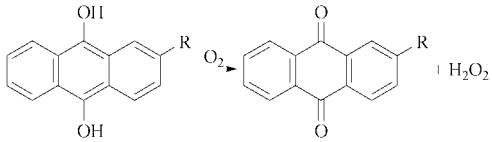

Hydrogen peroxide is extracted with water, and the quinones are returned to the hydrogenator to complete the loop.

The AO process, therefore, leads to the net formation of hydrogen peroxide from gaseous hydrogen and oxygen.

Solvents. Anthraquinones must be dissolved in a suitable solvent for hydrogenation, oxidation, and extraction (the so-called working solution). Although a number of individual solvents have been proposed (e.g., aromatic alcohols or their esters [24]), solvent mixtures are almost always used because the quinones and hydroquinones have different solubilities. Quinones dissolve readily in nonpolar, aromatic solvents (quinone solvents). Hydroquinones dissolve well in polar solvents, especially alcohols and esters (hydroquinone solvents). Proposed solvents or solvent mixtures are

Quinone solvents
- Benzene, xylene [25, 28]
- tert-Butylbenzene [26]
- Trimethylbenzene [28]
- Tetramethylbenzene (isodurene)[29]
- Mixtures of polyalkylated benzenes, e.g., C9-C11 aromatic solvents [30, 31]
- Methylnaphthalene [32]

Hydroquinone solvents
- Alkyl phosphates, e.g., tris-(2-ethylhexyl)-phosphate [33]
- Nonyl alcohols, e.g., diisobutylcarbinol [32]
- Alkylcyclohexanol esters, e.g., 2-methyl-cyclohexylacetate [34]
- *N*,*N*-Dialkyl carbonamides, e.g., *N*,*N*-dibutyl propionamide [35]
- *N*-Alkyl *N*-aryl carbonamides e.g. *N*-ethyl-*N*-phenyl benzamid [36]
- *N*,*N*-Dialkyl carbamates, e.g., 2-ethylhexyl-*N*-butylcarbamate [37]
- Tetraalkyl ureas, e.g., tetra-*n*-butylurea [42]
- Cycloalkyl ureas, e.g., dihexyl propyleneurea [38]
- Phenylalkyl ureas, e.g., *N*,*N*-dibutyl-*N*′-methyl-*N*′-phenylurea [39]
- *N*-Alkyl-2-pyrrolidones, e.g., octyl pyrrolidone [40]
- *N*-Alkyl caprolactams, e.g., *n*-octyl caprolactam [41]

Solvent mixtures [43, 44]:
- Polyalkylated benzenes and alkyl phosphates [30]
- Polyalkylated benzenes and tetraalkyl ureas [42]
- Trimethylbenzenes and alkylcyclohexanol esters [28]
- Methylnaphthalene and nonyl alcohols [32]

The following criteria must be fulfilled when choosing solvents and preparing solvent mixtures:

1) Good solubility of both quinone and hydroquinone
2) Good stability in the hydrogenator and oxidizer
3) Low solubility in water and aqueous hydrogen peroxide solutions
4) Sufficiently lower density than water to ensure separation of the two phases during extraction
5) Low volatility, i.e., high boiling point and flash point
6) High distribution coefficient for hydrogen peroxide in the solvent–water system
7) Low toxicity

Quinones. Criteria similar to those used for selecting solvents also apply to the quinone or quinone mixture which is used as the working compound:

1) Good solubility of the quinone form
2) Good solubility of the hydroquinone form
3) Good resistance to oxidation
4) Good availability

The choice of quinone depends mainly on the properties, especially the solubility, of the products formed in the AO process. Examples of working compounds are 2-ethylanthraquinone [*84-51-5*] [45], 2-tert-butylanthraquinone [*84-47-9*] [46], 2-neopentylanthraquinone [*123257-96-6*] [47], 2-isohexenylanthraquinone [*71308-16-2*] [48], diethylanthraquinones [49], mixtures of 2-amylanthraquinones [54], and mixtures of different alkylanthraquinones [50–53].

The formation of degradation products and their ability to be regenerated to active quinones also play a role in the decision. In addition to hydroquinone formation, a number of secondary reactions occur during the hydrogenation step; of these, hydrogenation of the anthraquinone ring system is particularly important. One ring of the anthraquinone molecule, preferably the one that is not substituted by an alkyl group, is hydrogenated. In the case of R=ethyl, this is accomplished with selectivities higher than 90 %.

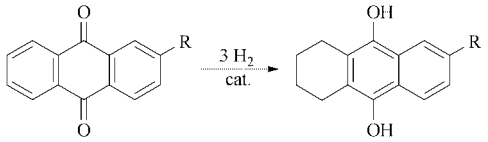

2-Alkyl-5,6,7,8-tetrahydro-9,10-dihydroxyanthracene is formed, which is also oxidized by oxygen to regenerate the 2-alkyl-5,6,7,8-tetrahydroanthraquinone (known as "tetra"), with simultaneous quantitative formation of hydrogen peroxide.

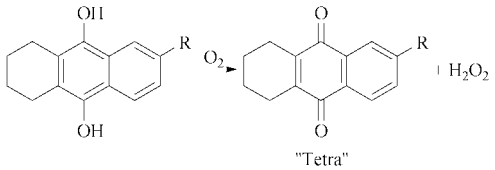

Although "tetra" is more readily hydrogenated than a 2-alkylanthraquinone, the resulting "tetra" hydroquinone is much more difficult to oxidize than the readily oxidizable anthrahydroquinone. The kinetics and basic physicochemical parameters of tetrahydroquinone oxidation have been studied [55–57]. Small amounts of amines accelerate hydroquinone oxidation [58].

The formation of "tetra" in the synthesis loop depends on process conditions and has led to two methods for carrying out this process: the "anthra" system and the "all-tetra" system.

Anthra System. The slower oxidation rate of the "tetra" hydroquinone causes difficulties in the oxidizer. Actions have therefore been proposed which suppress the formation of "tetra" or dehydrogenate it to anthraquinone.

When the "tetra" content of the working solution is kept low (i.e., if 2-alkylanthrahydroquinone is formed almost exclusively during hydrogenation), the process is referred to as the "anthra" system.

"Tetra" formation is suppressed by using selective catalysts [59], special solvents [60], special working compounds [61], and mild hydrogenation conditions [62] (e.g., the use of olefins has been recommend for the steady dehydrogenation of "tetra" [63]).

"Tetra" can be dehydrogenated in the presence of activated aluminum oxide:

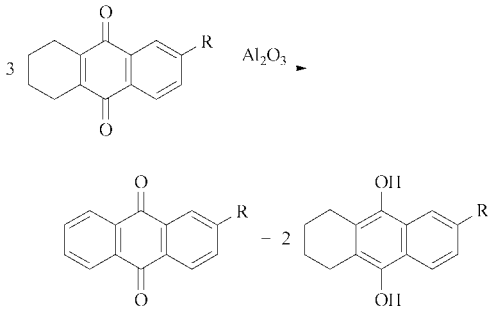

When the "anthra" system is used, tautomeric 2-alkyl-10-hydroxy-9-anthrone (oxanthrone) is also formed during hydrogenation [13, 64]:

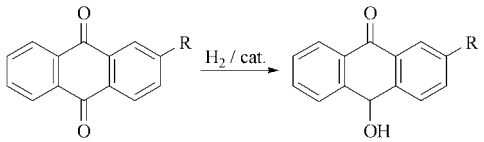

Oxanthrone

Oxanthrone is not oxidized by oxygen to form hydrogen peroxide and therefore leads to loss of active quinone. Further hydrogenation of oxanthrone leads to anthrone and tetrahydroanthrones [65]. Although oxanthrone and anthrone can be regenerated to active quinones, subsequent oxidative dimerization of anthrone may lead to dianthrones which represent a loss of quinone [13].

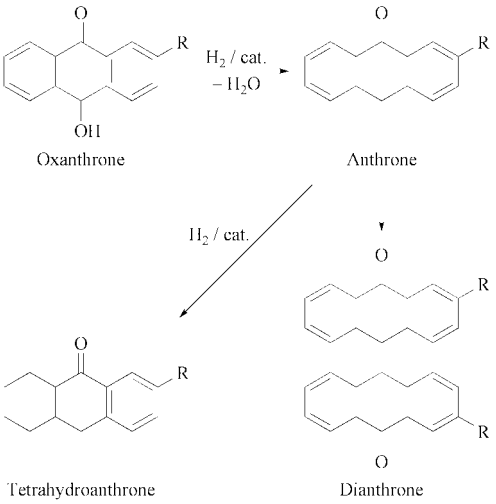

The buildup of anthrones and oxanthrones can be minimized by adding an aromatic tertiary amine to the working solution [66] or by pretreatment of the hydrogenation catalyst with halides [67].

"All-Tetra" System. If "tetra" formation is not suppressed during hydrogenation or "tetra" is not dehydrogenated, an equilibrium is reached in which the hydroquinone charged to the oxidizer consists exclusively of 2-alkyl-5,6,7,8-tetrahydroanthrahydroquinone. Such a system is called an "all-tetra" system.

If anthraquinone is hydrogenated in addition to "tetra" during the hydrogenation step, hydrogen transfer occurs [68]:

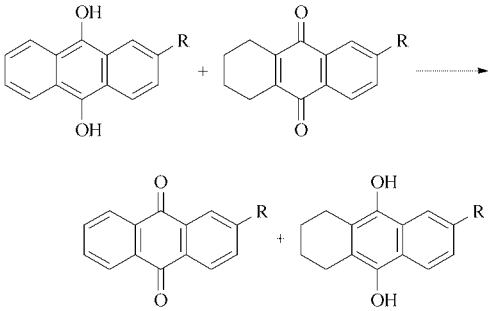

This means that the hydroquinone leaving the hydrogenation step is always the "tetra" hydroquinone. Dianthrone formation does not occur in the "all-tetra" system, but oxidation of "tetra" hydroquinone leads to a byproduct, the so-called epoxide:

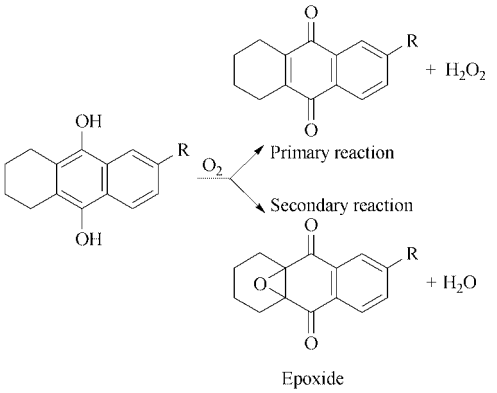

The epoxide does not participate in the formation of hydrogen peroxide and leads to a loss of active quinone. The formation of epoxides seems to be minimized when using high oxygen content [69], or mixing hydrogenated working solution with a portion of oxidized working solution containing hydrogen peroxide and then feeding to the oxidizer [70]. Different measures have been suggested for regenerating "tetra" from the epoxide [71, 73] (see Chapter 4.1.2.4)

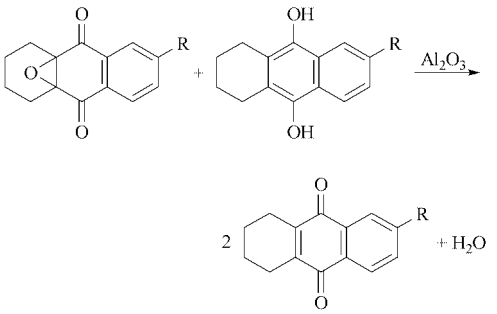

If the "tetra" content of the "all-tetra" system is high, another byproduct can be formed, namely, 2-alkyl-1,2,3,4,5,6,7,8-octahydroanthraquinone [13]:

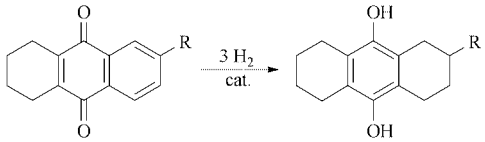

Although this "octa" hydroquinone is oxidized by oxygen to quinone with the formation of hydrogen peroxide, the reaction is too slow to be important in the formation of hydrogen peroxide. "Octa" thus represents a decomposition product from which quinone cannot be regenerated.

Octa formation depends on hydrogenation conditions and catalyst type. Other degradation products may be formed besides octa [65, 74, 75].

4.1.2. Process Description

The cyclic Riedel–Pfleiderer or BASF process (Chap. 1) forms the technological basis for all modern AO processes. Developments include improvement of the individual process steps, use of stable working solutions, and use of selective hydrogenation catalysts. The basic principles of the process are illustrated in Figure 5.

Hydrogenation. From the storage tank or hydrogenation feed tank (a), the working solution enters the hydrogenator (b) where it is hydrogenated in the presence of a suspended, supported, or fixed-bed catalyst.

If a suspended catalyst (e.g., palladium black or Raney nickel) or a supported catalyst (e.g., palladium) is used, the hydrogenation step includes a main filtration stage which retains the catalyst and allows it to be returned to the hydrogenator. The heat of reaction released during hydrogenation can be removed (1) before hydrogenation (by cooling the oxidized working solu-

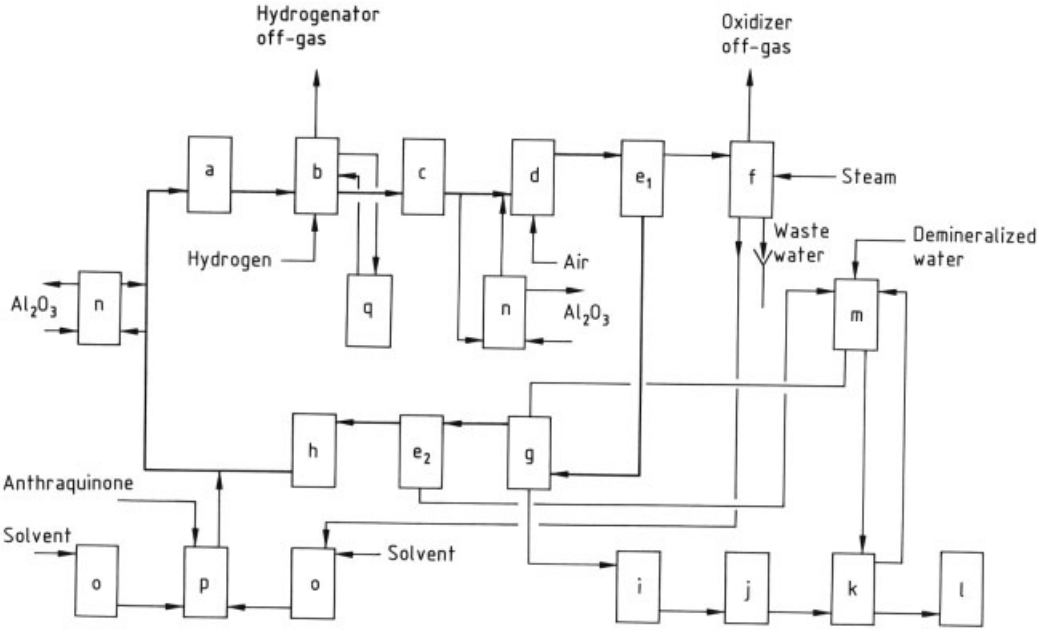

Figure 5. Production of hydrogen peroxide by the anthraquinone (AO) process
a) Storage tank for working solution or hydrogenator feed tank; b) Hydrogenator; c) Safety filtration; d) Oxidizer; e) Separator; f) Activated carbon adsorber; g) Extraction; h) Drying; i) Prepurification; j) Crude product storage tank; k) Hydrogen peroxide concentration; l) Hydrogen peroxide storage tank; m) Demineralized water feed tank; n) Regeneration and purification; o) Solvent storage tank; p) Working solution make up tank; q) Catalyst regeneration

tion), (2) during hydrogenation (by cooling the reactor), or (3) after hydrogenation (by cooling the hydrogenated working solution).

Oxidation. Before the hydrogenated working solution that contains hydroquinone can be fed to the oxidation step, it must pass through a safety filtration stage (c). This is particularly important because the hydrogenation catalysts used in the AO process (palladium and Raney nickel) also catalyze the decomposition of hydrogen peroxide. Even a small amount of these catalysts in the oxidation and extraction steps would lead to considerable loss of hydrogen peroxide and serious disturbances.

During the oxidation step (d), the hydrogenated working solution is gassed with air. Dissolved hydroquinones are oxidized to quinones, and hydrogen peroxide is formed.

After the working solution has been separated (e_1), air from the oxidation step passes over activated carbon adsorbers (f), and adsorbed solvent is recovered from them. Several adsorbers are usually loaded and regenerated alternately, for example, with steam.

Extraction and Recovery of the Working Solution. The oxidized working solution is then treated with water to extract hydrogen peroxide (g).

The working solution leaving the extraction unit must be adjusted to a specific water content before being returned to the hydrogenation step. Free water taken up by the working solution during extraction is separated (e_2) and the water content is adjusted to the desired level in the drier (h).

To purify the working solution and regenerate the quinone decomposition products into active quinones, some or all of the working solution is passed through a regeneration step (n).

Hydrogen Peroxide Concentration. Crude aqueous hydrogen peroxide from the extraction stage (H_2O_2 concentration 15–40 wt %) is fed into the crude product storage tank (j) via a prepurification unit (i).

From the crude product storage tank, aqueous hydrogen peroxide goes to the concentration unit (k) where it is distilled. Here, hydrogen peroxide is freed from most of its impurities and concentrated to the commercial concentration of 50–70 wt %; it is then collected in a storage tank (l).

Water vapor produced during distillation is condensed and fed into the water storage tank (m).

Auxiliary Processes. A number of additional processes are required to maintain the AO operation. For example, to maintain hydrogenation activity, part of the catalyst is removed, regenerated in the catalyst regeneration area (q), and returned to the hydrogenator. To compensate for quinone and solvent losses, working solution is periodically made up with anthraquinone and solvent (o, p).

4.1.2.1. Hydrogenation

The hydrogenation step is the most important step of modern AO processes. Quinone decomposition products that cannot be regenerated into active quinone are formed during this step (cf. Section 4.1.1). New hydrogenation catalysts and hydrogenation reactors have been developed that deviate totally from the BASF principle. Here, design of the hydrogenator depends largely on the type of catalyst used.

Four typical reactors for the three usual catalyst systems (suspended, supported, and fixed-bed catalysts) are discussed in this section.

BASF Hydrogenation Step. The hydrogenation step in the BASF plant (Fig. 6) uses a Raney nickel catalyst at a slight excess pressure of approximately 0.2 MPa and at 30–36 °C. Because Raney nickel is sensitive to oxygen, the working solution from the extraction or drying and purification steps cannot be fed directly into the hydrogenator. This working solution still contains residual hydrogen peroxide and must be passed over a decomposition catalyst (e.g., supported Ni–Ag), together with a fraction of the hydrogenated working solution (which also contains hydroquinone), to remove hydrogen peroxide completely:

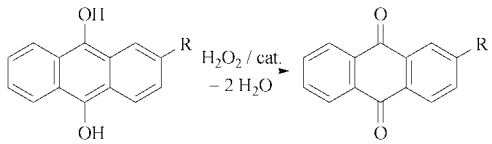

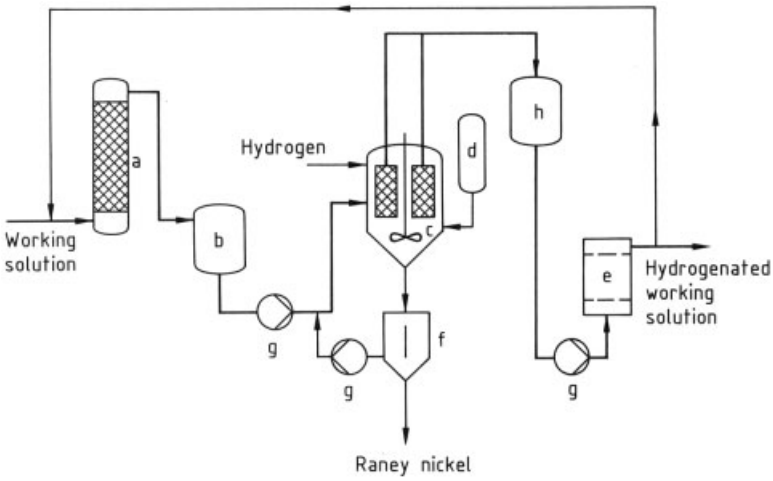

Figure 6. BASF hydrogenation step
a) Precontact column; b) Hydrogenator feed tank; c) Stirred tank reactor with internal filters; d) Catalyst feed tank; e) Safety filtration; f) Catalyst separator; g) Pump; h) Oxidizer feed tank

The solutions are passed through the precontact column (Fig. 6, a) and collected in the hydrogenator feed tank (b). The working solution is then pumped into the stirred vessel reactor (c) and is gassed with hydrogen in the presence of Raney nickel. Periodic addition of small amounts of hydrogenation catalyst from the catalyst feed tank (d) allows a constant rate of hydrogen conversion in the hydrogenator. Hydrogenated working solution is collected in the oxidizer feed tank (h) through the internal filters in the stirred vessel, thus exploiting the excess pressure in the reactor. The solution is then led into the oxidation step via the safety filter (e). A sidestream of hydrogenated working solution is withdrawn and recycled to the precontact column (a).

When the concentration of Raney nickel in the hydrogenation reactor reaches a certain limit, the content of the reactor is drained into the catalyst separator (f). Raney nickel settles to the bottom, and catalyst-free supernatant is pumped back to the hydrogenator.

A significant disadvantage of Raney nickel as catalyst is its limited selectivity, i.e., the ratio of hydroquinone formation to "tetra" formation. BASF largely eliminated this by pretreating the catalyst with ammonium formate [76].

Alternatives were subsequently suggested for pretreating Raney nickel (e.g., nitriles [77], amines [78], and aldehyde solutions [79]). In recent years, more selective nickel catalysts, based on Ni-B or Ni-Cr-B alloys were developed [80].

The pyrophoric properties of Raney nickel also require more stringent safety procedures when handling the material. Raney nickel is still used today in some AO plants, but palladium catalysts are preferred because of their higher selectivity and simpler handling.

Degussa Hydrogenation Step. Degussa proposed the use of palladium black as the hydrogenation catalyst. This exploits the advantages offered by a suspended catalyst and avoids the disadvantages of Raney nickel. Equipment that allows good conversion of hydrogen with very finely distributed palladium black is shown in Figure 7 [81].

The most important feature of the loop reactor (c) is the connection in series of pipes with different diameters. The working solution flows downward in the large pipes at a rate of 0.7–1.5 m/s and flows upward in the narrower pipes at 1.5–3 m/s. Efficiency can be improved by installation of internal plate-stacks with a channel width of 10-50 mm [82] or external premixing of gas and liquid [83].

Degussa proposed a carbon filter [84]. Ceramic [85] and cross-flow [86] filters can also be used. A decline in filter performance can be overcome by periodic back flush with hydrogenated working solution through the filter into the hydrogenator.

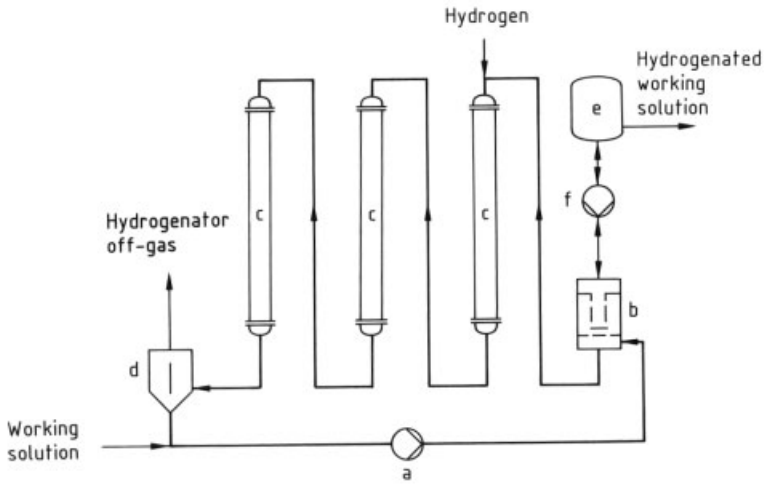

Figure 7. Degussa hydrogenation step
a) Reactor circulating pump; b) Filter; c) Loop reactor; d) Separator; e) Oxidizer feed tank; f) Back flush pump

Advantages of this hydrogenation system are

1) Almost complete conversion of hydrogen
2) Nonpyrophoric catalyst
3) Easy exchange of palladium black
4) Easy regeneration of the catalyst

Laporte Hydrogenation Step. Laporte Chemicals and other companies proposed the use of supported palladium catalysts [87]. Supports based on alumina [88], silica [89], and sodium aluminum silicate [90] are used. These catalysts have the advantage that their particle diameter in the range of 0.02–0.2 mm makes their filtration and recirculation to the reactor simpler than those of palladium black.

Laporte proposed the apparatus shown in Figure 8 for industrial hydrogenation [91]. The reactor contains a series of tubes whose lower ends lie just above the bottom of the reactor and whose top ends are just below the liquid surface. Hydrogen is fed into the bottom of each tube, and very small gas bubbles are formed by distributors. Upward flow occurs in the tube due to the density difference between the solutions in the tube and in the reactor. The catalyst suspension is drawn into the pipe by the continuous flow of working solution. To obtain a sufficiently high airlift effect in the tube, hydrogen must be circulated continuously. Continuously stirred tank reactors (CSTR) [92] and bubble-column reactors are also in use.

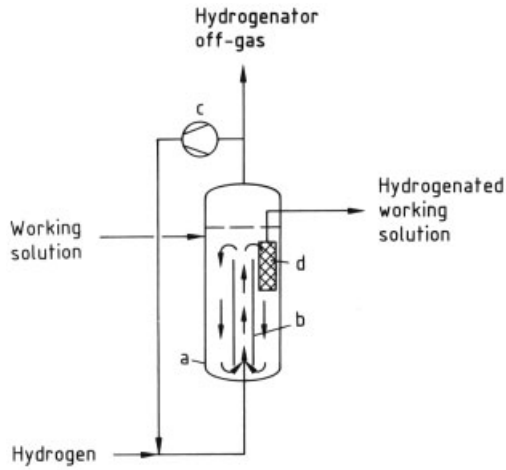

Figure 8. Laporte hydrogenation step
a) Hydrogenator; b) Reactor tubes; c) H_2 compressor; d) Internal filter

Fixed-Bed Hydrogenation Step. Fixed-bed hydrogenation represents a simple solution for the hydrogenation step; it involves a palladium catalyst and avoids the problem of filtration and recirculation of catalyst into the reactor. The first industrial fixed-bed hydrogenation unit for the AO process was commissioned by FMC (Fig. 9) [94].

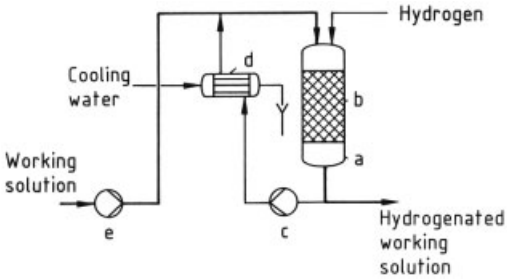

Figure 9. FMC hydrogenation step
a) Fixed-bed reactor; b) Catalyst; c) Reactor circulating pump; d) Heat exchanger; e) Hydrogenator feed pump

Most fixed-bed catalysts have a diameter of 0.2–5 mm and a palladium load of 0.1–0.5 %. The working solution is pumped to the top of the reactor. A side stream of the hydrogenated working solution is also fed into the fresh working solution after the heat of reaction has been removed in a heat exchanger (d). The catalyst must fulfill a number of requirements such as

1) High abrasion resistance to allow simplification of the filtration step
2) A long working life because replacing a fixed-bed catalyst is more complicated than replacing a suspended catalyst
3) Good productivity
4) Easy regeneration of the catalyst

External [98] or internal [99] static mixers may be used to ensure good liquid/gas dispersion. To enhance performance, operation in the foam regime with a liquid/gas ratio not higher than 1.5/10, temperatures of 45–75 °C, and a pressure of 0.18–0.5 MPa [93] is proposed.

A fixed-bed reactor with honeycomb monoliths is described by EKA [95]. The monoliths contain parallel channels with a diameter of 1–3 mm. Applied to the walls is a thin layer of porous silica gel, coated with palladium.

Desactivation plays an important role when operating fixed-bed catalysts. Impact of factors like type of support [96] or water [97] on side reactions was studied. Several measures, e.g., treatment with steam and hydrogen peroxide [100] or acid wash [101] are described for catalyst regeneration.

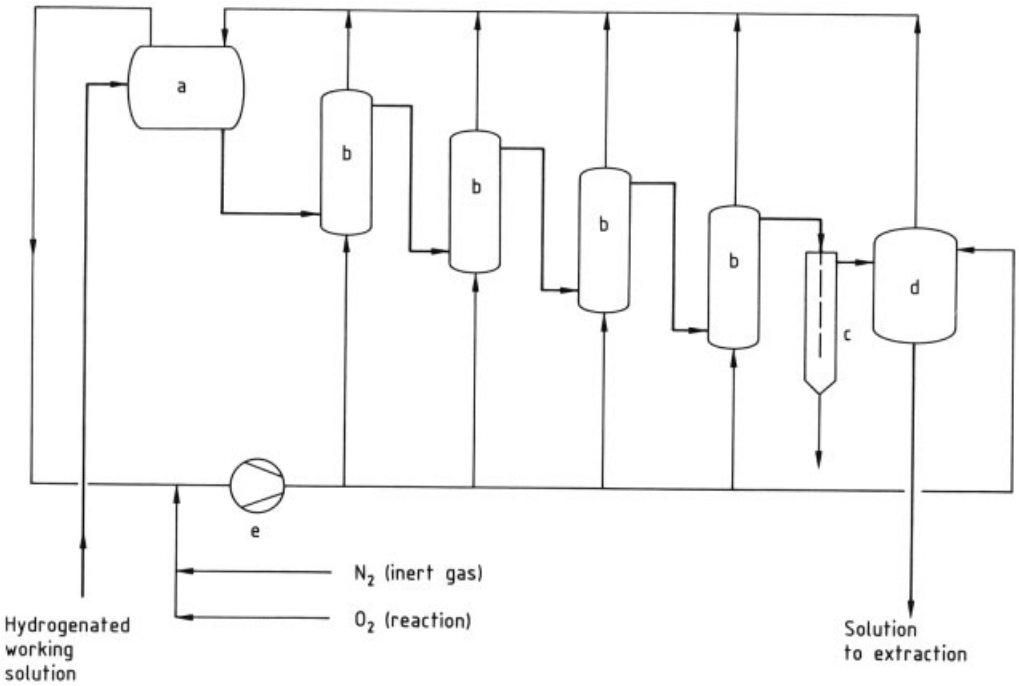

Figure 10. BASF oxidation step
a) Separator; b) Oxidation columns; c) Siphon; d) Extractor feed tank; e) Compressor

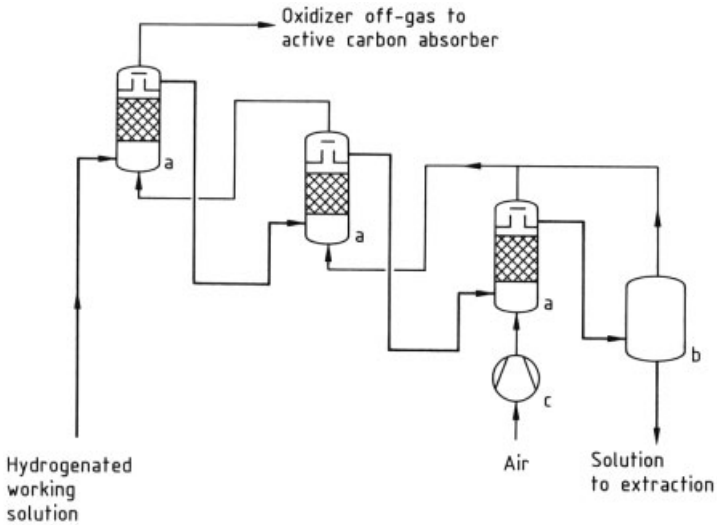

Figure 11. Degussa oxidation step
a) Concurrent-flow oxidation reactors; b) Separator; c) Air compressor

In recent years, especially AO plants built in China were equipped with fixed-bed hydrogenation.

4.1.2.2. Oxidation

In industrial plants, the catalyst-free hydrogenated working solution is usually oxidized with slightly pressurized air (up to 0.5 MPa). The oxidizer off-gas is then directed to a series of activated carbon adsorbers to purify it and recover the solvents escaping with it.

For economic reasons, the following objectives are sought in designing industrial oxidation reactors:

1) Efficient utilization of atmospheric oxygen to reduce the volume of the off-gas and the size of activated carbon adsorbers
2) Low compressor pressure to decrease energy costs
3) Small reactor volume or reduced holdup of working solution to lower investment costs for equipment and working solution

BASF Oxidation Step (Fig. 10). A working solution containing benzene as the quinone solvent has been used at BASF. For safety reasons and to minimize the volume of off-gas, the BASF oxidation step uses a nitrogen–oxygen mixture. The mixture is cycled in a closed circuit, and its composition is kept constant by adding pure oxygen to compensate for the amount consumed in the formation of hydrogen peroxide.

Hydrogenated working solution enters a separator (a) and then flows through four oxidation columns (b) arranged in series as a cascade. Oxidized working solution flows into the extractor feed tank (d) via a siphon (c). The nitrogen–oxygen mixture is compressed and fed into each of the four reactors.

Off-gas from the reactors is collected in a pipe and separated from entrained droplets of working solution. After compensation for consumed oxygen, the gas mixture is returned to the compressor (e).

Degussa Oxidation Step (Fig. 11) [102]. The hydrogenated working solution flows cocurrently upward with the oxidizing air through the first reactor. The working solution and air are separated in the upper part of the reactor. Off-gas from the reactor is fed to the activated carbon adsorption unit, and working solution flows into the second and third reactors, which are installed in series. The air needed for oxidation is pressurized with a compressor (c) and fed to the third reactor. The working solution and air flow

countercurrently, but both streams are divided into cocurrent flow segments.

Laporte Oxidation Step (Fig. 12). In Laporte's Warrington plant, oxidation was carried out in a packed cocurrent flow column [103, 104]. The entire volume of the oxidation reactor (a) was used for air gassing. The air and the working solution leaving the top of the column were fed together into a separator (b). The air then reached the alternately operated two-stage activated carbon adsorbers (c), and the working solution passed to the extraction stage. To enhance gas-liquid mass transfer, premixing of working solution and air by a venturi tube is proposed [105]. The use of a tubular reactor with static mixing elements is described by Kemira in [106].

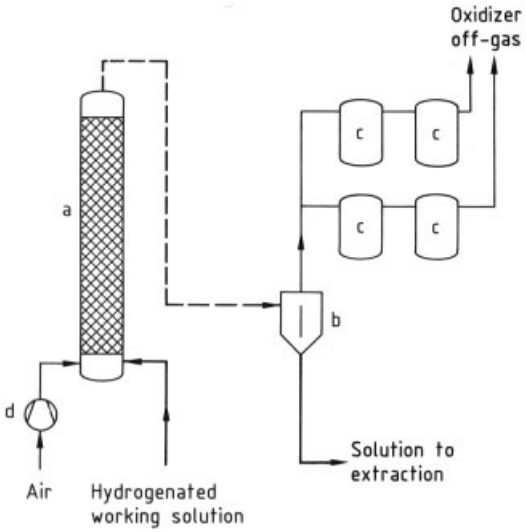

Figure 12. Laporte oxidation step
a) Oxidation reactor; b) Separator; c) Activated carbon adsorber; d) Air compressor

Allied Chemical Oxidation Step. To shorten the residence time of the working solution in the oxidation step, Allied Chemical suggested a counterflow reactor in which residence times are less than 2.5 min at a partial oxygen pressure of 70–90 kPa [107]. Space–time yield can be improved by using perforated trays with small hole size (cross-sectional area of holes below 3 mm^2) [108].

4.1.2.3. Extraction and Drying

In the BASF plant, a sieve-tray extraction column extracts hydrogen peroxide from the working solution. A number of other extractors, such as packed columns, pulsed packed columns, and Podbielniak extractors, have been proposed.

Besides extraction column design, the extraction efficiency is strongly influenced by the distribution coefficient, which is depending on working solution composition (e.g., type of hydroquinone solvent).

It has been found, that by injection of a small amount of air into the extraction column (gas-agitated sieve plate column) extraction efficiency is increased by 30–40 % [109]. Processes which integrate oxidation and extraction in a single apparatus are also known [112, 113].

Working solution leaving the extraction unit contains dispersed water droplets and is led into coagulators and separators to separate free water. The working solution is then dried to a specific water content.

Because the solubility of water in the working solution depends on temperature, its moisture content can be adjusted by carrying out the extraction at low temperature, separating the dispersed water, and then increasing the temperature of the working solution by about 30 °C before it reaches the hydrogenation step [110]. The working solution may be alternatively dried by using the off-gas from the oxidation step [111].

In extraction and drying according to BASF (Fig. 13), working solution leaving the head of the extraction column (a) is initially freed from entrained water in a water separator (b). The working solution then passes through aqueous potassium carbonate solution for drying (c).

4.1.2.4. Working Solution Purification and Regeneration

During the constant circulation of working solution, degradation products are formed not only from the working compound, but also from the solvents. These degradation products must be removed from the working solution to prevent (1) deterioration of the crude hydrogen peroxide (color, smell, dissolved organic compounds), (2)

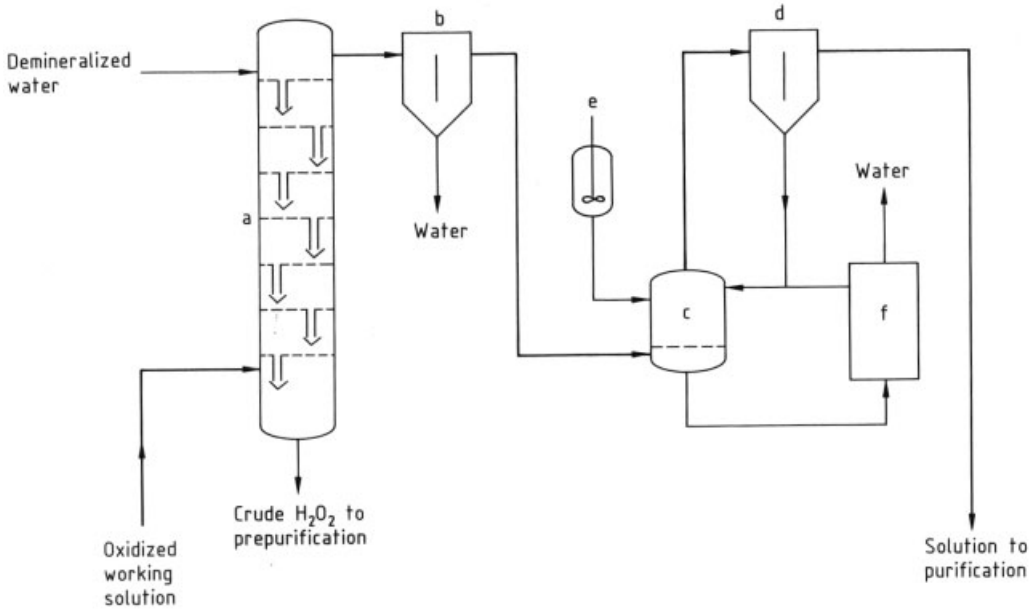

Figure 13. BASF extraction and drying step
a) Sieve–tray extraction column; b) Water separator; c) Drier; d) Potassium carbonate solution separator; e) Potassium carbonate solution make-up and feed tank; f) Potassium carbonate concentration

an increase in density and viscosity of the working solution, and (3) a decrease in interfacial tension of the working solution, which promotes the formation of an emulsion during extraction. In addition, degradation products decrease the activity and lifetime of the hydrogenation catalysts. Regeneration and purification can be performed with hydrogenated working solution or with working solution after extraction.

Numerous methods have been suggested for purifying the working solution and regenerating active quinone from the quinone degradation products. Examples include the following:

1) Treatment of the hydrogenated working solution with alkaline substances [71] or aqueous alkali hydroxide [72]
2) Treatment of the working solution with sodium aluminum silicates at 50–200 °C [114]
3) Treatment of the working solution after hydrogenation or extraction with active aluminum oxide at 20–150 °C [115–118]
4) Treatment with alkali hydroxide, calcium hydroxide, ammonia, or amines in the presence of oxygen or hydrogen peroxide [119]
5) Treatment of the working solution with aqueous sodium hydroxide or potassium hydroxide solution (8–17 mol/L) in the presence of oxidants [120]
6) Treatment with solvent and liquid carbon dioxide [121]
7) Treatment with a noncyclic hydrocarbon [122]

4.1.2.5. Purification of Crude Hydrogen Peroxide

The aqueous hydrogen peroxide leaving the extraction is impure. Separators are usually installed after extraction (Section 4.1.2.3) [123]. Measures for reducing the amount of dissolved organic compounds include treatment of the crude product with polyethylene [124], ion exchangers [125, 126], or hydrocarbons [127, 128]. Water-soluble organic compounds can also be oxidized by heating and then extracted with suitable solvents (e.g., the quinone solvent) [129]. The effect of purification on carbon content of crude hydrogen peroxide is illustrated in Table 5.

Table 5. Processes for purifying crude hydrogen peroxide

Process	Carbon content of H_2O_2, mg/L		Reference
	Before purification	After purification	
Extraction with xylene	580	450	[130]
Extraction with methylcyclohexanol acetate	580	390	[130]
Extraction with xylene and methylcyclohexanol acetate	580	370	[130]
Brief treatment with 0.5 wt % activated carbon	134	11	[131]
Treatment with adsorber resin	320	140	[132]
	1000	100	[133]
Reverse-osmosis membrane	130	18	[134]
Supercritical carbon dioxide	145	84	[135]

Purified crude product is then usually fed to a distillation unit where it is purified further and concentrated to the usual commercial concentration of 50–70 wt % hydrogen peroxide. This product can be used directly in a number of applications. Reverse osmosis membranes can also be used to purify crude, distilled or concentrated (up to 70 %) hydrogen peroxide [134, 137]. A disadvantage of this procedure is that only part of the feedstream leaves the unit as purified permeate, while the impurities are accumulated in the retentate.

For use in the electronics industry, hydrogen peroxide with a very low content (ppt level) of metal impurities is needed. Many patents deal with further purification of crude or prepurified hydrogen peroxide with a combination of cation- and anion- exchange resins [136–140]. Some methods make use of of reverse osmosis in combination with other methods [134, 141, 142], adsorption of metal impurities on activated carbon [143], stannic oxide [144], or zirconium phosphate [145].

4.1.2.6. Concentration of Hydrogen Peroxide

Water and hydrogen peroxide do not form an azeotrope; their boiling point difference at atmospheric pressure is 50.2 °C (Table 1). Distillation of water from aqueous hydrogen peroxide can therefore be used to concentrate the solution or to completely separate hydrogen peroxide from water. A number of technical safety requirements must be observed. Distillation of dilute solutions has been used industrially to produce commercial 50–70 wt % hydrogen peroxide solution; the technology is described in [3, 5, 146–152]. The use of structered packings [150] in distillation columns enhances perfomance. Concentrations higher than 90 % are preferably produced by crystallization [153, 154].

4.2. 2-Propanol Process (Shell Process)

In 1945 Harris discovered that primary and secondary alcohols react with oxygen to form hydrogen peroxide and an aldehyde or ketone, respectively [155]:

$$RCH_2OH + O_2 \rightarrow RCO + H_2O_2$$

$$R_2CHOH + O_2 \rightarrow R_2C{=}O + H_2O_2$$

Oxidation of the alcohol can be carried out in the liquid or the gas phase. Because aldehydes formed in the reaction with primary alcohols are oxidized easily, only the oxidation of secondary alcohols, especially 2-propanol [*67-63-0*], has industrial importance.

A process for the simultaneous production of hydrogen peroxide and acetone [*67-64-1*] by atmospheric oxidation of 2-propanol was first suggested by Rust [156, 157].

Reaction of 2-propanol with oxygen in the liquid phase does not require a special catalyst because it is catalyzed by the hydrogen peroxide product. However, a small amount (0.5–1 wt %) of hydrogen peroxide is added to 2-propanol to shorten the induction period. The reaction proceeds as follows:

$$HOOH \rightarrow 2\,HO^{\bullet}$$

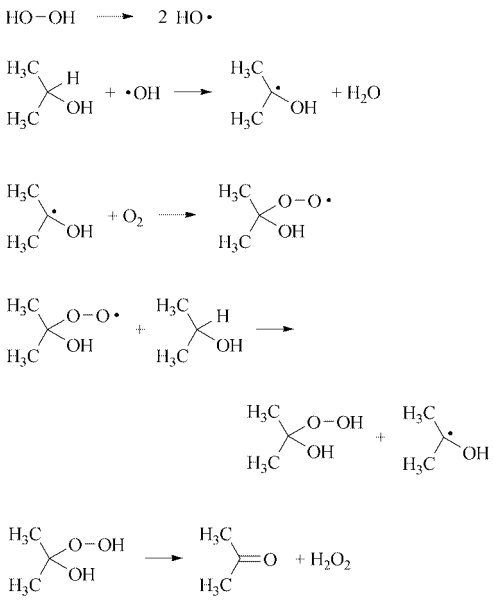

Secondary reactions also take place but are not discussed here. To reduce the formation of byproducts, particularly acetic acid, which affects the quality of hydrogen peroxide, only part of the 2-propanol is oxidized, and oxidation is carried out in several consecutive steps at decreasing temperature [158].

The 2-propanol process is used only in two plants in the former Soviet Union, which have been operating since 1968 and 1972. The Shell Oil Company decommissioned its plant in Norco, Louisiana in 1980 [159, 160]; that plant had been in operation since 1957. The technology of the Shell 2-propanol process is described in [161]. The hydrogen peroxide yield of the Russian plants based on 2-propanol or atmospheric oxygen is 87–98 %; the acetone yield is 92–94 % [162, 163].

An alternative process, which makes use of the liquid phase oxidation of methylbenzyl alcohol, coproduct in the production of propylene oxide by epoxidation of propylene with ethylbenzene hydroperoxide, was developed by Arco but has not been commercialized [164–166]. Similar to the AO process the operations include hydrogenation, oxidation, extraction, and purification steps. Oxidation of methylbenzyl alcohol with air at elevated temperature yields hydrogen peroxide and acetophenone. Methylbenzyl alcohol conversion in the reactor is 30 % with H_2O_2 selectivity about 80 %. The hydrogen peroxide content in the organic phase at the exit of the oxidizer is approximately 5 wt %. After addition of ethylbenzene hydrogen peroxide is extracted with water. The organic phase overflowing at the top of the extractor contains methylbenzyl alcohol, ethylbenzene and acetophenone. Whereas ethylbenzene is distilled off and recycled to the extractor acetophenone is hydrogenated to methylbenzyl alcohol. Advantages of the ARCO process compared with the AO process are: cheap working solution, methylbenzyl alcohol is used both as solvent and as working compound and high hydrogen peroxide concentration in the working solution after oxidation.

4.3. Electrochemical Processes

Electrochemical production of hydrogen peroxide is of minor industrial importance compared to the AO process.

Because electrochemical processes played an important part in the development of hydrogen peroxide and their technology reached a high standard, essential features of the three known cyclic processes are described below. These processes are described in detail in [3, 5].

The basis of electrochemical processes is anodic oxidative coupling of sulfate ions to peroxodisulfate ions. Sulfuric acid is oxidized in the Degussa–Weissenstein process, whereas ammonium hydrogensulfate solutions are used in the Münchner process.

$$2\,H_2SO_4 \rightarrow H_2S_2O_8 + H_2$$

$$2\,(NH_4)\,HSO_4 \rightarrow (NH_4)_2S_2O_8 + H_2$$

Hydrogen peroxide is formed by hydrolysis of the peroxodisulfate ion via the peroxomonosulfate anion:

$$S_2O_8^{2-} + H_2O \rightarrow HSO_4^- + SO_3(OOH)^-$$

$$SO_3(OOH)^- + H_2O \rightarrow HSO_4^- + HOOH$$

In the Degussa–Weissenstein and Riedel–Loewenstein processes, peroxodisulfuric acid or ammonium peroxodisulfate derived from

Table 6. Yields and energy consumption of electrochemical processes

Yield or consumption*	Degussa–Weissen-stein process	Münchner process	Riedel–Loewenstein process
Current yield, %	73	83	83
Hydrolysis and distilla-tion yield, %	93–96	87	82–85
Total yield, %	68–70	72	68–70.5
Electrical energy, kW · h	1280	1760	1700
Steam, t	1.85	2.5	4.5–4.8
Cooling water, m^3	200	240	250

* Parameters are expressed per 100 kg of hydrogen peroxide (calculated as 100 wt %) in the form of a distilled 35 wt % solution.

the electrolysis is hydrolyzed directly. In the Münchner process, ammonium peroxodisulfate is converted to potassium peroxodisulfate, which is then hydrolyzed. The Münchner process, therefore, consists of several coupled cyclic processes.

All three processes have a very high energy consumption and an unsatisfactory yield of ca. 70 % (Table 6).

Huron developed a modified Loewenstein process, which makes use of platinum–titanium anodes and achieves a yield of 85 % [167]

4.3.1. Degussa–Weissenstein Process

In the Degussa–Weissenstein process (Fig. 14), sulfuric acid is adjusted to a concentration of 550–570 g/L (a) before flowing through the catholyte chamber (b) where residual peroxide is destroyed and heavy-metal ions are precipitated. After addition of chemicals to raise the electrochemical potential (e.g., ammonium rhodanide or hydrochloric acid), sulfuric acid passes to the anolyte chamber where peroxodisulfuric acid is formed. Hydrolysis of peroxodisulfuric acid and

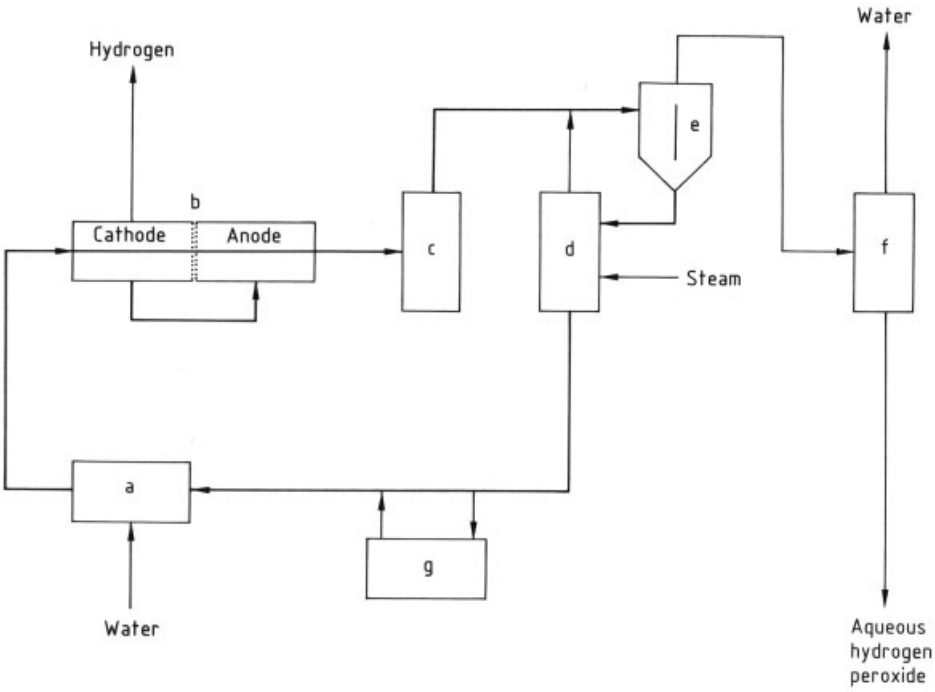

Figure 14. Degussa–Weissenstein process
a) Electrolyzer feed tank; b) Electrolyzer; c) Evaporation and hydrolysis; d) Hydrogen peroxide expulsion and hydrolysis; e) Acid separator; f) Water–hydrogen peroxide fractionating column; g) Acid purification

distillation of hydrogen peroxide are carried out in two steps (c, d). Hydrolysis is 80–90 % completed in (c), and distillation is achieved in approximately equal parts in two steps (c, d). In the expulsion column (d), hydrogen peroxide is expelled from the concentrated sulfuric acid solution by countercurrent steam. Entrained acid is separated in an acid separator (e), and hydrogen peroxide–water vapors are separated in a fractionating column (f). An aqueous solution with a maximum content of 45 wt % hydrogen peroxide is obtained. To purify the circulating solution, part of the stream is distilled constantly (g).

4.3.2. Münchner Process

In the Münchner process (Fig. 15), acid ammonium sulfate is electrolyzed in electrolyzers without diaphragms.

The circulating solution contains 300 g of $(NH_4)_2SO_4$, 40 g of K_2SO_4, 100 g of H_2SO_4, and 57.6 g of $S_2O_8^{2-}$ per liter before entering the electrolyzer. Its peroxodisulfate concentration is increased to 115.2 g/L by the electrolytic process. After electrolysis, the circulating solution is cooled and concentrated in a vacuum cooler (b) by evaporation of water and then fed into a crystallizer (c). Half of the peroxodisulfate is precipitated as potassium peroxodisulfate by addition of potassium hydrogen sulfate and separated in a centrifuge (d). The supernatant mother liquor diluted with washing water is returned to the electrolyzer. Solid potassium peroxodisulfate is hydrolyzed (e) and hydrogen peroxide is steam distilled.

The remaining potassium hydrogensulfate–sulfuric acid residue is separated in a centrifuge (f). Sulfuric acid is recycled to the hydrolysis and distillation stage (e) via a purification stage (g). Potassium hydrogensulfate is returned to the crystallizer (c) for precipitation of peroxodisulfate.

To purify the circulating solution, a side stream is mixed with ammonia in the purification step (h) to precipitate iron. The hydrogen peroxide–water vapors formed during hydrolysis and distillation are led through an acid separator (i) and separated in the rectification step (j). Because hydrolysis is carried out batchwise, the initial vapor mixture is rich in hydrogen peroxide, giving a distillate of 60 wt % which then drops to 20 wt %.

4.4. Other Processes

A number of reactions and processes for the production of hydrogen peroxide are described in the literature. Some of them are no longer economical (e.g., the barium peroxide, azobenzene.

Only a few examples can be described here; detailed information may be found in the literature [1, 3, 5].

4.4.1. Production from Peroxy Compounds

Peroxy compounds can be decomposed to form hydrogen peroxide. Examples of such reactions, which include the historically important Thenard reaction of barium peroxide with nitric or sulfuric acid, are listed in Table 7.

Table 7. Production of hydrogen peroxide from peroxy compounds

Peroxide	Reaction	Reference
BaO_2	reaction of barium peroxide with H_2SO_4 in the presence of a small amount of HCl allows quantitative conversion; ca 6 % aqueous H_2O_2 solutions are obtained	[90]
BaO_2	as above, but in the presence of HCl and H_3PO_4	[91]
Na_2O_2	sodium peroxide solutions are passed through acidic ion exchangers to exchange Na^+ for H^+	[92]
$(CH_3)_3CO–OH$	*tert*-butyl hydroperoxide is hydrolized to *tert*-butanol and H_2O_2 in the presence of H_2SO_4 in a fractionating column	[93]
$(CH_3)_3CO–OH$	H_2O_2 and isobutene are formed from *tert*-butyl hydroperoxide in an aqueous medium in the presence of an acidic ion-exchange resin	[94]
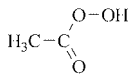	peracetic acid is hydrolyzed to CH_3COOH and H_2O_2 in the presence of H_2SO_4; to shift the equilibrium to H_2O_2, CH_3COOH is reacted with an alcohol and the ester is distilled	[95]
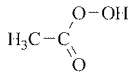	peracetic acid is hydrolyzed with acidic ion exchangers, and the equilibrium mixture is separated by distillation	[96]

4.4.2. Production from Hydrogen and Oxygen

Hydrogen peroxide can be formed by direct reaction of oxygen and hydrogen. However, the

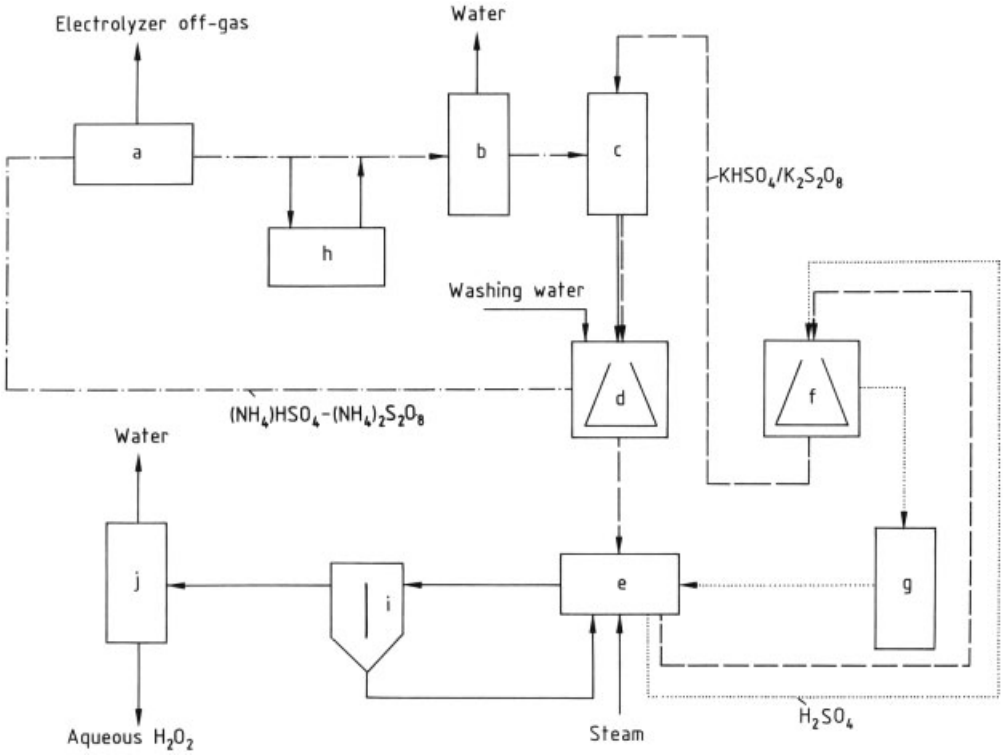

Figure 15. Münchner process
a) Electrolyzer; b) Vacuum cooler; c) Conversion (crystallization); d) Centrifuge; e) Hydrolysis–distillation; f) Centrifuge; g) Sulfuric acid purification; h) Purification of cycle solution; i) Acid separator; j) Fractionating column

most important hydrogen peroxide manufacturing process, the AO process, also allows production of hydrogen peroxide from the elements.

In 1935 [175] Pietzsch reported that silent electrical discharge into a gas containing 95 % hydrogen and 5 % oxygen in the presence of water vapor at 150–160 °C and 100 kPa, and subsequent condensation result in a 10 % aqueous solution of hydrogen peroxide. The power requirements are high (40 kWh per kilogram H_2O_2). Silent electrical discharge was proposed for producing ultrapure hydrogen peroxide for the semiconductor industry [176].

In 1985, DuPont developed and patented a new liquid-phase hydrogen peroxide process that directly combines hydrogen and oxygen and that allows a 50 % reduction in investment costs [177]. The DuPont patent [178] described the production of up to 25 % hydrogen peroxide solutions with a palladium on carbon catalyst in acidic aqueous solution containing halide ions as promoter at 10–25 °C and up to 17 MPa using an explosive oxygen/hydrogen gas mixture. The process was intensively investigated during the last two decades, a demonstration plant with a capacity of several kilotons of hydrogen peroxide in methanol is scheduled to be started up in the middle of 2006 [179]. For this direct combination of hydrogen and oxygen various catalytic systems, mostly with precious metals, were proposed. The process was especially investigated during the last decade, but has reached only pilot plant status. The ideal direct combination process should have:

1) High hydrogen peroxide production rate and low decomposition
2) A catalyst with long lifetime and minimal precious metal losses
3) Use of a non explosive gas mixture
4) Synthesis of hydrogen peroxide at low pressure and ambient temperature
5) Noncorrosive reaction medium
6) High hydrogen selectivity and hydrogen peroxide concentrations of at least 5 %

Heterogenous catalysts such as palladium or platinum or mixtures thereof on silica, alumina, charcoal or hydrophobic supports [180] or fixed-bed catalysts [181] are preferred. Fixing the catalyst on a halogenated resin eliminates dissolution of catalyst and corrosion of the equipment but results in low selectivity and product concentration [182, 183] Extremely high hydrogen selectivities (better than 90 %) using nonexplosive feedstreams can be reached with palladium on carbon black nanocatalysts [184].

To overcome the concerns over process safety while using explosive gas mixtures, special reactor designs are proposed: e.g., separately feeding hydrogen and oxygen into a completely liquid-filled tubular reactor [185], injecting dispersed minute bubbles of hydrogen and oxygen into a rapidly flowing liquid medium [186], or separately feeding very small bubbles of hydrogen and oxygen in the lower part of a mechanically agitated reactor in such a way that the molar ratio of hydrogen to oxygen falls within the explosive range but in the gas phase the ratio of the reactants lies outside this range [187]. Hydrogen peroxide concentrations as high as 20 % at hydrogen selectivities of 80–83 % can be reached. Use of methanol instead of water as reaction medium enables to lift the catalyst productivity from 207 to 879 gram H_2O_2 per gram catalyst per hour [188]. A vapour phase technique using a fluidized bed catalyst consisting of a thin layer (50 nm) of palladium on polytetrafluoroethylene is described [189, 190].

4.4.3. Production from Carbon Monoxide, Oxygen, and Water

Hydrogen peroxide can be formed from carbon monoxide, oxygen and water.

$$CO+O_2+H_2O \rightarrow H_2O_2+CO_2$$

The reaction was first described by Traube in 1884. Later palladium [191] or palladium phosphine and arsine complexes [192, 193] were used as catalysts, but the latter lacked oxidation stability. A palladium bathocuproine complex in a two-phase acidic solution should overcome these difficulties [194], but reaction pressure (5–7 MPa) and excess oxygen (ca. 10–20-fold) were high. The advantage of the process is that oxidation and reduction takes place in a single reactor. Improved catalyst stability is reported when using naphthoquinone as mediator [195]. The process is not used commercially.

4.4.4. Production by Autoxidation of Organic Compounds

Autoxidation processes for the production of hydrogen peroxide include the AO process and the 2-propanol process discussed in Sections Anthraquinone Process (AO Process) and 2-Propanol Process (Shell Process). A variation of the AO process is the use of H_2S instead of H_2 for the reduction of quinones [196, 197]. The azobenzene process has occasionally been used on a commercial scale. Analogous to the AO process, azobenzene is reduced to hydrazobenzene which is then oxidized to azobenzene. The reaction proceeds at alkaline pH and results in formation of $Na_2O_2 \cdot 8\ H_2O$ [198–200].

Other autoxidation processes have been suggested, such as the oxidation of hydrocarbons and alcohols, but they have no commercial significance. They are usually carried out in the gas phase at high temperature and give a multicomponent reaction mixture due to subsequent secondary reactions. Hydrogen peroxide can be also generated by the oxidation of hydrazine or hydroxylamine [201].

4.4.5. Production by Cathodic Reduction of Oxygen

The formation of hydrogen peroxide (max. 1 % solution) by cathodic reduction of oxygen was reported by Traube in 1882 [202]. The process was improved by Fischer and Priess who introduced pressurized electrolysis (ca. 10 MPa), which gave 1.3–2.7 % hydrogen peroxide solutions with current efficiencies of 83–90 % [203]. The cathodic reduction of oxygen can be described by the Equation (1).

$$O_2+H_2O+2\,e^- \rightarrow HOO^- + HO^- \quad (1)$$

Hydrogen peroxide is formed according to Equation (2).

$$HOO^- + H_2O \rightarrow H_2O_2 + HO^- \quad (2)$$

Side reactions are

$$HOO^- + H_2O + 2\,e^- \rightarrow 3\,HO^- \quad (3)$$

and

$$H_2O_2 - 2\,e^- \rightarrow O_2 + 2\,H^+ \quad (4)$$

Reaction (3) occurs at high current densities. To increase the space–time yield at low current densities, electrodes with high specific surface areas are used. Catalytic decomposition of hydrogen peroxide is avoided by using nonmetallic construction materials and limiting pH and temperature. The anodic decompostion (Eq. 4) is limited by using diaphragms or by controlling the anolyte flow. Oxygen evolved at the anode can be reused for cathodic reduction.

The first attempts to bring this process to technological maturity were made by Berl, followed by Kastening [204], and more recently by Huron and Dow (H-D Tech) [205–207] and others [208, 209]. An overview is given in [210, 211].

The Huron–Dow process seems to be the only one used on an industrial scale. It is intended for small on-site units, especially for the pulp industry. A unit with ca. 3000 t/a capacity has been operated since 1992 at Georgia Pacific in Muskogee/USA. The process uses a trickle-bed cathode and yields an alkaline solution with ca. 4 % hydrogen peroxide. The H_2O_2/alkali ratio is 1/1.7. The process is operated with an excess of $>$ 90 vol % oxygen near atmospheric pressure (ca. 0.11 MPa). A graphite/carbon black/teflon composite cathode with a lifetime of over one year is used, and the current yield is 80–95 %. The solution must have a minimum alkali concentration of 4 %; otherwise current yields are low.

For lowering the basicity of the hydrogen peroxide, combinations with dialysis cells are proposed [212]. Newer developments make use of fuel cell designs [213, 214]. While oxygen is reduced at the cathode, hydrogen at a pressure of 4.8–6.2 MPa is oxidized at the anode. No external power supply is needed, and alkaline-free hydrogen peroxide solutions with concentrations of up to 15 % can be produced.

5. Storage and Transportation [4]

The decomposition of hydrogen peroxide caused by catalytic impurities and the associated release of heat have been described in Chapters Physical Properties and Chemical Properties. Great care must be taken in production, storage, and transportation to prevent these impurities from entering the hydrogen peroxide solution and to ensure that hydrogen peroxide is put only into perfectly clean containers. Because commercial hydrogen peroxide always contains small quantities of catalytic impurities, the stabilization of peroxide solutions is extremely important. The stabilizing effects of inorganic and organic compounds are described in detail in [3].

Sodium pyrophosphate [*7722-88-5*] and sodium stannate [*12058-66-1*] are the preferred stabilizers and are added separately or together [215, 216]. Phosphonic- or aminophosphonic acids may also be used [217, 226].

Aluminum (99.5 %), aluminum–magnesium alloys, or stainless steels are good construction materials [218]. Because of their corrosion resistance, polyethylene containers and storage tanks are preferred for hydrogen peroxide concentrations up to 50 wt %. Before metallic tanks and containers can be used, their surfaces must be passivated. Iron particles may become attached to the surface during the rolling of aluminum, and they must be removed. Aluminum is therefore treated with dilute sodium hydroxide and then passivated with dilute nitric acid. Because hydrogen forms during the caustic treatment of aluminum with sodium hydroxide, safety precautions must be taken to avoid a detonating gas atmosphere. It is extremely important to ensure that no hydrogen peroxide is trapped, e.g. between closed valves. If decomposition occurs, ex-

Table 8. Transport regulations for hydrogen peroxide

	H_2O_2 concentration, wt %		
	8–20	**20–60**	**$>$ 60**
RID/ADR	5.1, No. 1 c	5.1, No. 1 b	5.1, No. 1a
IMDG Code	5.1	5.1	5.1
UN No.	2984	2014	2015
	packing	packing	packing
	group III	group II	group I

tremely high pressures result which lead to very serious explosions.

Hydrogen peroxide is admitted for air transport only up to a concentration of 40 wt % (IATA regulations).

No special regulations apply to the transport of solutions of up to 8 wt % hydrogen peroxide. Table 8 lists international regulations for rail (RID), road (ADR), and sea (IMDG code) transport of more highly concentrated solutions.

6. Safety

Hydrogen peroxide has a high energy content:

$$H_2O_2\,(l) \rightarrow H_2O\,(l) + 1/2\,O_2\,(g)\ \Delta H = -98.30\,\mathrm{kJ/mol}$$

Furthermore, its high oxygen content (47 wt %) is available for combustion reactions and the oxidation of organic compounds. This means that special safety conditions apply in handling hydrogen peroxide [219, 232]. Safety studies can be divided into three groups:

1) Examination of the pure vapor phase,
2) Examination of the pure liquid (aqueous) phase, and
3) Examination of mixtures or solutions of organic compounds with or in aqueous hydrogen peroxide.

Results are summarized below, but the literature must be consulted for details. In specific cases, especially when hydrogen peroxide is used with organic compounds, the intended work area must be checked for hazards if it has not already been certified as safe by a safety examination.

Pure Hydrogen Peroxide — Vapor Phase [5, 220–223]. Explosive vapor mixtures are formed at atmospheric pressure when the hydrogen peroxide concentration in the vapor phase exceeds 26 mol %. Figure 16 shows the explosive range as a function of temperature and hydrogen peroxide concentration.

For the industrial concentration of hydrogen peroxide solutions, it is important to note that at reduced pressure the critical hydrogen peroxide concentration of the vapor phase may exceed 26 mol % (Fig. 17).

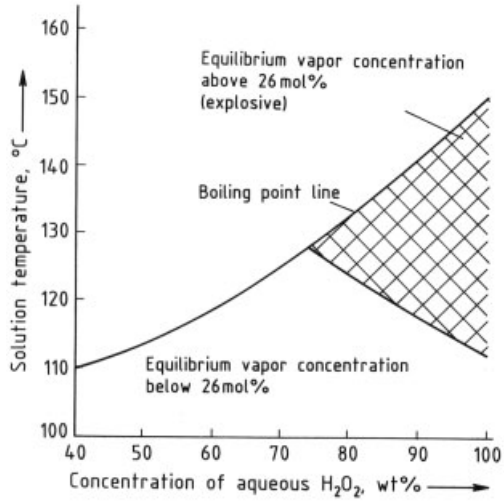

Figure 16. Dependence of the explosive range of the vapor phase (> 26 mol % H_2O_2) on the temperature and the hydrogen peroxide concentration of liquid phase (at atmospheric pressure)

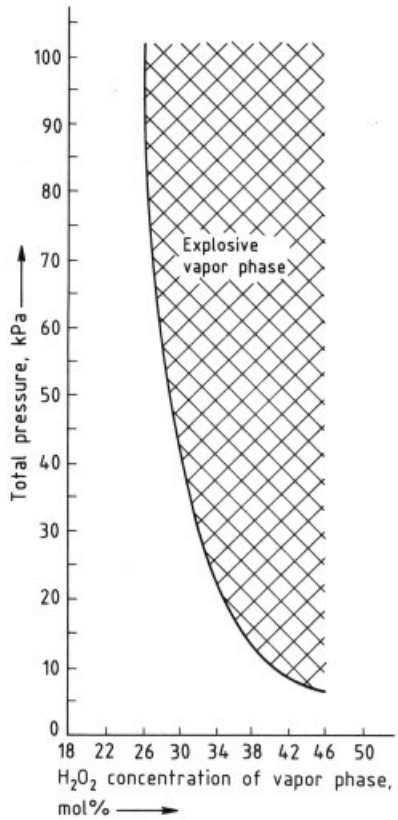

Figure 17. Dependence of the explosive range of vapor phase on pressure

Pure Hydrogen Peroxide — Liquid (Aqueous) Phase [224–226]**.** The explosive risk of pure, highly concentrated hydrogen peroxide solutions has been examined by various methods, but the sometimes contradictory findings are difficult to interpret consistently.

Results are influenced primarily by test conditions and not by hydrogen peroxide concentration. For example, 90.7 wt % hydrogen peroxide in a tube with an internal diameter of 2.67 cm (1.05 inches) could not be exploded by the detonation shock of a tetryl booster, whereas explosion occurred in a tube with an internal diameter of 4.09 cm (1.61 inches). Similarly, 86 % hydrogen peroxide in a tube of 4.09 cm internal diameter was exploded by raising the temperature above 50 °C [224, 225]. The decisive factors for explosion of pure aqueous hydrogen peroxide seem to be the degree of occlusion, the insulation, and the energy of detonation.

Mixtures or Solutions of Aqueous Hydrogen Peroxide with Organic Compounds [219, 227–234]**.** The hazards associated with mixtures of hydrogen peroxide and organic chemicals are important for its industrial production and, especially, for its uses. As an example, Figure 18 shows results obtained with the hydrogen peroxide–acetone–water system [227]. Many other organic compounds give similar results. The size of the explosive region depends on the organic compound and the test conditions.

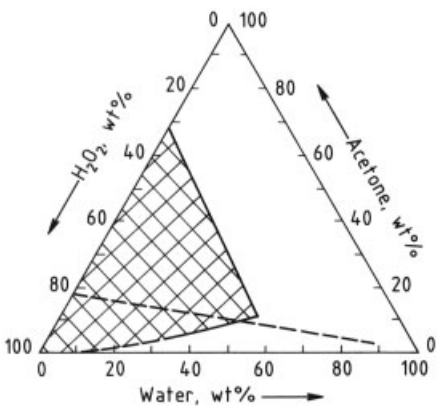

Figure 18. Explosive range (hatched area) of hydrogen peroxide–acetone–water mixtures
Dashed line, indicates the stoichiometry
$8\,H_2O_2 + CH_3COCH_3 \rightarrow 3\,CO_2 + 11\,H_2O$

For practical use, safety tests of the intended working area are very important. When systems such as hydrogen peroxide–formic acid–water are used, other products may be formed (in this case performic acid), that are more dangerous than the original mixture [228].

7. Uses

Development of the AO process permitted large-scale production of hydrogen peroxide, which is now used widely in almost all industrial areas. Its main use is in bleaching (→ Bleaching, Chap. 2.2.3). Uses in the chemical industry and in environmental protection are increasing because its great advantage is that the degradation product is water.

Hydrogen peroxide is used in the textile industry for bleaching cotton, linen, bast fibers, wool, silk, polyester fiber, and polyurethane fiber. In the pulp and paper industry, it is used to bleach sulfate and sulfite cellulose, wood pulp, and wastepaper, and to brighten wood veneers and wooden structures.

The chemical industry employs hydrogen peroxide for the production of peroxy compounds such as sodium perborate, sodium percarbonate, metallic peroxides, or percarboxylic acids. Hydrogen peroxide is very important in organic chemistry [229] for epoxidation and hydroxylation (manufacture of plasticizers and stabilizers for the plastics industry), oxidation (manufacture of amine oxides as surfactants for detergents), oxohalogenation, and initiation of polymerization. BASF and Dow Chemical have developed a cost-effective process for production of propylene oxide using hydrogen peroxide; the first commercial plant with a capacity of 300 000 t/a of propylene oxide using this new technology is scheduled to go on stream early 2008 [234].

Sulfuric acid solutions of hydrogen peroxide are used for the pickling and chemical polishing of copper, brass, and other copper alloys, as well as for etching and cleaning printed circuit boards. Highly purified hydrogen peroxide is used in wet chemical processes during the manufacture of silicon semiconductor devices to clean silicon wafers and to remove photoresist layers. It is also used for in situ leaching in underground uranium mining.

Hydrogen peroxide is increasingly used in environmental protection to detoxify effluents containing formaldehyde, phenols, or cyanide (e.g., wastewater from mines and tempering works, galvanizer concentrate, photochemical effluents), and to deodorize sulfur-containing effluents. Smoke and exhaust gases containing sulfur dioxide and even dioxins can be completely detoxified with hydrogen peroxide.

Hydrogen peroxide is a highly efficient disinfectant (especially for packaging materials); it is also used as a bleaching agent in hair preparations and as a propellant [226] in space technology.

8. Toxicology and Occupational Health

Hydrogen peroxide plays a major role in physiological oxidation processes in living organisms. Hydrogen peroxide is produced during normal aerobic cell metabolism, which involves a number of enzymes, in particular superoxide dismutase. Humans may exhale up to 1 mg/m3 of hydrogen peroxide resulting from endogenous formation [236].

Reactive oxygen species, including hydrogen peroxide, may also lead to oxidative damage to cells. Therefore a number of defense systems can be found in biological systems which lead to detoxification of reactive oxygen species. The most important detoxification reactions for hydrogen peroxide are its conversion to oxygen and water by enzymes such as catalase, peroxidase, and selenium-dependent glutathione peroxidase. The activity of these enzymes varies between tissues and between different animal species or strains [236]. Additionally, a variety of nonenzymatic antioxidants, such as Vitamin E and A, ubiquinols, ascorbic acid, glutathione, complete an efficient intra- and extracellular network of defenses [237].

There is limited information concerning the absorption of exogenously applied hydrogen peroxide. The permeability of biological membranes to hydrogen peroxide is comparable to that of water, but less than that of oxygen [236].

Although significant amounts of topically applied hydrogen peroxide can penetrate the epidermis or mucous membranes, the substance undergoes rapid spontaneous or enzyme-catalyzed decomposition to water and oxygen in the underlying tissue. This may lead to the formation of small gas bubbles and reversible blanching of the exposed tissue. Formation of larger volumes of gaseous oxygen can lead to detachment of cell layers and the rupture of tissues and organs. Locally formed oxygen is removed by the blood. However, the increase in oxygen content in the blood results in a hyperbaric response. There is no evidence that hydrogen peroxide itself is absorbed in amounts sufficient to produce systemic effects [236]. In in vitro studies with rat blood plasma additions of up to 1 mM of hydrogen peroxide were degraded to nondectable concentrations within less than 10 min [238]. This experimental result supports the assumption that hydrogen peroxide is not systemically available.

The major adverse effects of hydrogen peroxide are related to its irritating properties.

The acute oral toxicity of hydrogen peroxide in animals depends on the strength of its solution. In the rat LD_{50} values ranged from $>$ 5000 mg/kg for a 10 % solution to 1200 mg/kg for 35 % solution. In humans, toxic effects observed after accidental ingestion of concentrated solutions are generally related to the irritant or corrosive action and the rapid generation of large volumes of oxygen in the gastrointestinal tract. Single cases of deaths due to gas embolism following accidental swallowing of hydrogen peroxide solutions have been reported [239], however, in other case reports complete recovery was observed within 2-3 weeks [236].

Complete recovery has occurred within two to three weeks even in near fatal cases [236].

Effects of exposure to hydrogen peroxide by inhalation are limited to irritation of the respiratory tract and the eyes. Exposure to hydrogen peroxide vapor (338–427 mg/m^3) induced signs of irritation in rodents. In humans exposed for 4 h the irritation threshold for the respiratory tract was 10 mg/m^3 and for the skin 20 mg/m^3. At these concentrations, eye and throat irritation were reported [236].

Slight nasal irritation was reported in factory workers for short-term exposure to ca. 3.5 mg/m^3 [236]. Gradual bleaching of hair was reported for exposure levels of 0.7–1.4 mg/m^3, but it was uncertain whether bleaching was due to vapor exposure or to transfer of hydrogen peroxide from the hands (gloves) to the hair [236].

The dermal toxicity of hydrogen peroxide solutions of 35–75 % in rabbits was low (LD_{50}, dermal, rabbit: > 2000 mg/kg). LD_{50} values of a 90 % solution ranged from ca. 690 mg/kg in rabbits to > 2000 mg/kg in rats, pigs, and cats [236].

The irritating properties of hydrogen peroxide are dependent on the concentration of the solution. Hydrogen peroxide solutions of 10 % or less were nonirritating, and solutions of 35 % slightly irritating to rabbit skin. Higher concentrations (≥ 50 %) are corrosive to rabbit skin. Hydrogen peroxide (50 %) was corrosive to rabbit skin after 1 h of exposure, whereas 70 % hydrogen peroxide showed corrosive effects after only 3 min of exposure [236]. Dermal exposure of humans to dilute solutions has been reported to result in transient symptoms without permanent effects, such as paresthesia, blanching, and blistering [236].

With regard to eye irritation solutions of 5 % or less are not irritant to rabbit eyes; a concentration of 6 % caused slight reversible eye irritation in rabbits. Concentrations of 10 % or higher caused irreversible corneal damage. In humans, 1–3 % solutions have been used for eye treatment without causing injury. However, solutions containing more than 200 mg/L have been reported to cause hyperemia, and concentrations of greater 100 mg/L transient pain and stinging [236]. Hydrogen peroxide (6 %) was not a skin sensitizer in guinea pigs [236]. Human experience from dermatological hospital departments confirms that hydrogen peroxide is no skin sensitizer [239].

After repeated inhalation exposure (6 h/d, 2 to 3 d per week, 3 weeks), mortality in mice was observed at 80 mg/m^3 hydrogen peroxide. After prolonged exposure, in rats and dogs (5 or 6 h/d, 5 d per week for 16 or 26 weeks, respectively), respiratory irritation and transient skin thickening were observed at 10 mg/m^3 hydrogen peroxide [236]. In a 28-day inhalation study in rats the NOAEL for respiratory tract irritation was 2.9 mg/m^3. Clinical signs of irritation were observed from 14.6 mg/m^3 [240].

In humans, pulmonary function monitoring showed no evidence of adverse effects from occupational exposure for several years at levels ranging from not detectable to 0.79 mg/m^3 (8-h TWA) [236]. Further surveys of workers occupationally exposed to hydrogen peroxide in different production plants for up to 40 years with recent exposures below 1.4 mg/m^3 (8-h TWA), but occasional peak exposures up to 5 mg/m^3 confirmed the absence of effects on lung function or any other clinical effects apart from single observations of skin or throat irritation, skin or hair whitening after acute incidents of high short-term exposures [239]. In a population exposed to levels of 2 - 3 mg/m^3 (8-h TWA) and frequent peak exposure levels up to 11 mg/m^3 symptoms including eye and respiratory irritation, headaches and temporary loss of olfaction as well as skin and hair blanching and three cases of recurring bronchitis–sinusitis were reported. All persons recovered completely after the exposure was reduced to below the occupational exposure limit of 1.4 mg/m^3 and the repeated peak exposures were avoided. The authors assume that the recurrent peak exposure levels of up to 11 mg/m^3 may have contributed to the observed symptoms [241].

More detailed studies of the effects of hydrogen peroxide after prolonged exposure were performed by administration in drinking water. After prolonged exposure (146 d), decreased bodyweight gain was observed in rats at concentrations of 0.25 % and above [242] (other studies observed first effects at higher levels [236]).

In a 90-d drinking water study in catalase-deficient mice receiving 100–3000 ppm of hydrogen peroxide, irritation of the mucous membranes of the gastrointestinal tract was observed. The no observed adverse effect level (NOAEL) was 100 ppm, corresponding to 26 and 37 mg kg^{-1} d^{-1} for male and female mice, respectively. After cessation of treatment these effects were fully reversible. No systemic effects were observed [243]. In studies on the same mouse strain with a 6–7 month exposure period, signs of gastrointestinal irritation were observed at approximately the same dose level as in the 90-d study. Therefore it can be assumed that the effect is not cumulative and the NOAEL does not decrease with increasing exposure time. A NOAEL of 56.2 mg kg^{-1} d^{-1} was reported in a 12-week gavage study in rats [236].

Results of mutagenicity studies indicate that hydrogen peroxide induces gene mutations in vitro without metabolic activation in bacteria, yeast, and mammalian cells. Hydrogen peroxide also induced chromosome aberrations and micronuclei in mammalian cell cultures without

a metabolic activation system. However, addition of exogenous metabolic activators or catalase lowered or abolished the genotoxic response [236]. In a mechanistic study, it has be shown that hydrogen peroxide induced controlled cell death (apoptosis) rather than genotoxic lesions in cell cultures of human lymphocytes and did not induce 8-hydoxyguanosin formation [244].

In *in vivo* tests, no chromosomal aberrations or micronuclei were observed in the bone marrow of rats and mice after repeated oral or single intraperitoneal administration [236]. An in vivo/ex vivo study for unscheduled DNA synthesis (UDS) after intravenous administration of the maximum tolerated dose to rats did not show an increase in UDS compared to the solvent control group [245]. In a study of local genotoxicity to mouse skin, 70 % hydrogen peroxide did not cause any mutations, hydroxylation of DNA bases, or gave an indication of cellular proliferation, thus suggesting that the substance has no local genotoxic potential in vivo [239]. In conclusion, the available studies indicate that hydrogen peroxide has no genotoxic or mutagenic potential under *in vivo* conditions.

The carcinogenic effects of hydrogen peroxide were studied in rats and mice. Hydrogen peroxide induced an increase of duodenal tumors in mice chronically exposed to 0.1–0.4 % of hydrogen peroxide in drinking water. The tumor incidence correlated with a specific inflammatory response in this tissue and was more pronounced in those mice strains which had a low catalase activity [236]. The absence of preneoplastic changes and the reversibility of the effects in a 90-d drinking water study in mice of the same strain [243] (with 0.01–0.3 % hydrogen peroxide in drinking water), suggest that the tumors observed in the two-year study seem to be a secondary response to a prolonged inflammatory process in this specific strain of mice. Rats exposed to near lethal concentrations in their drinking water developed forestomach papilloma but no tumors in the glandular stomach and duodenum [236]. Chronic hydrogen peroxide exposure at concentrations which induce significant inflammation of the exposed tissue has been shown to give a weak tumor-promoting effect. Repeated topical application of 15 % hydrogen peroxide to Sencar mice did not induce skin tumors; this demonstrates the absence of initiating properties [236].

On the basis of these studies ACGIH has grouped hydrogen peroxide into category A3 for carcinogenicity: "The agent is carcinogenic in experimental animals at relatively high doses, by routes of administration, at sites of histological types, or by mechanisms that are not considered relevant to worker exposure. Available epidemiological studies do not confirm an increased risk of cancer in exposed humans. Available evidence suggests that the agent is not likely to cause cancer in humans except under uncommon or unlikely routes of exposure" [246]. Similar conclusions were drawn in the EU risk assessment on hydrogen peroxide [239].

Occupational exposure limits have been established for hydrogen peroxide in different countries, for example by ACGIH [246], the Dutch Social Economic Council [247], the UK HSE [251], Belgium [248], the French Ministry of Labour [249], and the German AGS (Ausschuss für Gefahrstoffe) [250]: 8-h TWA: 1.4 mg/m^3 or 1.5 mg/m^3. A 15-min STEL of 2.8 mg/m^3 was established by HSE (max. 1 h in a 24-h period) [251], and a 15-min STEL of 1.4 mg/m^3 by AGS [250].

9. Environmental Effects

Hydrogen peroxide occurs naturally. Its formation, use, and degradation are strongly related to the oxygen-containing atmosphere of our planet and adaptation of nature to oxygen. The concentration of hydrogen peroxide in environmental compartments results from a dynamic equilibrium between its natural production and degradation in photochemical, chemical, and biological processes.

Because of its chemical reactivity, hydrogen peroxide also has toxic properties that are sometimes used as a natural defense mechanism. The toxicity of hydrogen peroxide to organisms and cells is dependent on the balance of intoxication and detoxification. However, detoxification reactions can in turn lead to toxic effects due to the release of oxygen gas [252].

9.1. Atmospheric Fate

The atmospheric fate of hydrogen peroxide is governed by generation and degradation reactions. The formation of atmospheric hydrogen

peroxide is photochemically mediated. Consequently, natural concentrations are latitude-dependent. High solar irradiation leads to high hydrogen peroxide concentrations. Hydrogen peroxide in atmosphere is produced through radical reactions. Light, oxygen, hydrocarbons, and free radicals in the atmosphere mediate hydrogen peroxide formation.

In an unpolluted atmosphere, the formation of H_2O_2 is ultimately due to the photolysis of ozone (O_3) at wavelengths < 310 nm to give oxygen atoms. The latter are converted to free radical species, including $^{\bullet}OH$, $RO^{\bullet}$, $HO_2^{\bullet}$, and $RO_2^{\bullet}$, which initiate the breakdown of organic species and give rise to $HO_2^{\bullet}$ radicals which are precursors of H_2O_2 [253]. The superoxide radical $O^{2-\bullet}$ is the direct precursor of H_2O_2:

$$2\,O^{2-\bullet}+2\,H^{+}\rightarrow H_2O_2+O_2$$

The rate of H_2O_2 formation from $O^{2-\bullet}$ is of the second order and is also temperature- and humidity-dependent. Levels of H_2O_2 were raised by 70 % when air temperature increased from 10 to 30 °C, other factors remaining constant [254].

Water vapor is involved in another process of H_2O_2 formation, namely, the production of hydrated hydroperoxyl radicals, which rapidly react with $HO_2^{\bullet}$ radicals to generate H_2O_2. About twice as much H_2O_2 is formed at 100 % relative humidity than at 50 % humidity at ambient temperature (25 °C) [255].

Degradation in the atmosphere is due to direct and indirect photolysis reactions and to chemical reaction with organic substances. Degradation by indirect photolysis is expected to be the main degradation mechanism in air. Indirect photolysis is due to sensitization by secondary reactions with OH and O_2 radicals and organic substances. The most important degradation reaction is reaction with the hydroxyl radical:

$$H_2O_2+{}^{\bullet}OH\rightarrow H_2O+HO_2^{\bullet}$$

Direct hydrogen peroxide photolysis is mediated by radiation at wavelengths of 280–380 nm:

$$2\,H_2O+h\nu\rightarrow H_2O+{}^{\bullet}OH+HO_2^{\bullet}$$

$$^{\bullet}OH+HO_2^{\bullet}\rightarrow H_2O+O_2$$

The rate for the direct photolysis of hydrogen peroxide for moderate light intensities is directly proportional to the hydrogen peroxide concentration and the square root of the light intensity [256]. Kleinman estimated the direct photolysis half-life to be 2.14 days [257].

The formation of organic hydroperoxides (R−OOH) appears to be another pathway for the decomposition of H_2O_2 in the atmosphere. They are formed by hydroxy radical induced oxidation of hydrocarbons or direct oxidation of aldehydes by hydrogen peroxide:

$$R-CHO+H_2O_2\rightarrow R-OOH$$

However, the measured levels of R−OOH have been reported to be only 1 to 10 % of the ambient air concentrations of hydrogen peroxide [258].

The overall tropospheric half-life of hydrogen peroxide was reported to be 10 to 20 h [257]. Atmospheric levels of hydrogen peroxide ranged from 0.02 ppb to 4 ppb in unpolluted air and up to 180 ppb in smog situations. Concentrations in rainwater ranged from 0 to approximately 2000 µg/L, cloud water contained up to 8000 µg/L [252].

9.2. Aquatic Fate

Hydrogen peroxide levels in natural waters range from 0.14 to 58 µg/L in seawater and from 0.34 to 109 µg/L in freshwater [252]. The concentrations are dependent on the daytime and sunlight intensity and are decreasing with increasing depth. In surface water a photochemical process involving dissolved, light-absorbing organic matter (e.g., humic acids) and molecular oxygen produces hydrogen peroxide. The wavelengths needed are between 290 and 385 nm [256, 259, 260].

$$\text{Organic matter}+h\nu\rightarrow \text{Organic radical}^{\bullet}$$

$$\text{Organic radical}^{\bullet}+O_2\rightarrow \text{Organic radical}^{+\bullet}+O_2^{-}$$

$$2\,O_2^{-\bullet}+2\,H^{+}\rightarrow H_2O_2+O_2$$

In seawater, the Weiss mechanism also contributes to the formation of hydrogen peroxide [261].

$$Fe^{2+}+O_2\rightarrow Fe^{3+}+O_2^{-\bullet}$$

$$Fe^{2+}+O_2^{-\bullet}+2\,H^{+}\rightarrow Fe^{3+}+H_2O_2$$

Hydrogen peroxide concentrations in surface water can also originate from precipitation of rain [259].

9.3. Biodegradation

The main source of degradation of hydrogen peroxide in natural waters is the degradation by biological material, e.g., plankton. The half-life in water decreases with increasing size of the microbial population in water [252].

Consequently, the half-life in natural waters was reported to range from a few hours to several days [262]. Degradation is also catalyzed by transition metals. This may in particular play a role for the more rapid degradation in seawater. Direct photolysis does not play a major role in aquatic decomposition.

In municipal wastewater, hydrogen peroxide degrades rapidly because of the content of microorganisms and organic material. Half-lives were reported to be below 15 min [263]. Half-life was reported to be below 2 min in an activated sludge respiration test according to OECD guideline 209 [252].

9.4. Environmental Releases

Environmental releases of hydrogen peroxide, in particular to surface water, can occur during production and use. Where there is no wastewater treatment, higher concentrations may be released that may lead to toxicity to aquatic organisms. A review of the environmental releases from different production and use scenarios is given in the EU risk assessment of hydrogen peroxide [239].

9.5. Environmental Effects

Hydrogen peroxide is moderately toxic to aquatic organisms. For hydrogen peroxide, acute ecotoxicity data are available for three aquatic tropic levels (fish, daphnia, and algae). The toxicity of hydrogen peroxide to salt water species was of the same order of magnitude as that to freshwater species. The 96-h LC50 values for freshwater fish range from 16.4 to 37.4 mg/L H_2O_2 [252]. The 48-h EC50 value for *Daphnia pulex* was 2.4 mg/L and the 24-h EC50 for *Daphnia magna* was 2.3 mg/L [239].

For algae, the initial algae concentration influences the outcome of the study. The most appropriate studies are those with the OECD standard of 10 000 cells/mL at the moment of addition of hydrogen peroxide. Under these conditions, the 72-h EC50 (biomass) in the green algae *Chlorella vulgaris* was 2.5 mg/L, based on growth rate the 72-h EC50 was 4.3 mg/L [265]. Other algae species were less sensitive, for example the 96-h IC50 for *Scenedesmus quadricauda* was 27.5 mg/L [266].

For the bacterium *Pseudomonas putida*, a 16-h EC_{10} of 11 mg/l was determined [252]. At concentrations from 30 mg/mL (no observed effect concentration in an activated sludge respiration inhibition test), some toxicity to degrading microorganisms has been observed, the EC_{50} was 466 mg/L in this study [264]. However, at concentrations up to 200 mg/L, H_2O_2 does not affect the performance of biological treatment works (activated sludge process). It has even been shown to stimulate the biodegradation in wastewater treatment works [252].

Chronic toxicity studies on standard fish and daphnia species are not available, but studies on different life stages of fish [267] and prolonged (56-d) exposure of zebra mussels [268] do not indicate effects of higher sensitivity than in acute studies. The lowest long-term aquatic no-effect concentration (NOEC) was 0.1 mg/L in *Chlorella vulgaris* [265]. In a field study, the influence of hydrogen exposure on phytoplankton and bacterioplankton in a small humic lake in Canada has been examined. At H_2O_2 concentrations above 3.4 mg/L, both phytoplankton and bacterioplankton were inhibited. Bacteria were inhibited at concentrations between 0.12 and 3.4 µg/L. Photosynthetic activity of phytoplankton was increased at concentrations between 3.4 and 34 µg/L. Interestingly, the toxicity to phytoplankton was dependent on the time of the day. During the period of lowest background H_2O_2 concentrations at sunrise, no effect was observed even when concentrations as high as 170 µg/L were added. At sunset, a concentration of 1.7 µg/L resulted in a decrease of primary production by algae suggesting that the detoxification capacity of the cells might have been exceeded because of the higher natural concentrations [269].

The effect of hydrogen peroxide to plants has been tested on growth, seed germination, and early seed growth of rice, corn, soybean, pigweed, tomatoes, and barnyard grass. Concentrations up to 218 mg/L in irrigation water have no

significant effect on any of the above parameters [252].

When assessing the risk of hydrogen peroxide to environmental organisms, the background concentrations in the environment and the capacity to adapt to increased hydrogen peroxide concentrations have to be considered. The medium used in laboratory tests is quite different from the natural conditions. Depending on the environmental conditions, many materials (e.g., dissolved transition metals, solid metals, solid metal oxides, hydroxides, activated carbon, enzymes) may contribute to rapidly reduced hydrogen peroxide levels. It is thus assumed that under laboratory conditions the test organisms are more exposed and thus more sensitive than under natural conditions to the prolonged exposure of H_2O_2. In field conditions, the more sensitive species may also be protected from a direct effect of hydrogen peroxide by the presence of the total biomass and the suspended solids.

However, it is obvious that accidental releases, in particular of highly concentrated hydrogen peroxide, can immediately be lethal to aquatic organisms. With regard to long-term effects of lower levels, it can be expected that adaptation of ecosystems may well be possible.

10. References

General Reference

1. *Gmelin*, System no. 3, Sauerstoff.
2. *Kirk-Othmer*, 5th ed., 14, Wiley, 2005.
3. W. Machu: *Das Wasserstoffperoxid und die Perverbindungen*, 2nd ed., Springer, Wien 1951.
4. Wasserstoffperoxid, *Berufsgenossenschaft der chemischen Industrie*, Merkblatt M 009, BGI 792, Jedermann-Verlag, Heidelberg 2000.
5. W. C. Schumb, C. N. Satterfield, R. L. Wentworth: *Hydrogen Peroxide*, Reinhold Publ. Co., New York 1955.
6. *Ullmann*, 5th ed., A13, 443–465.
7. W. M. Weigert: *Wasserstoffperoxid und seine Derivate*, Hüthig, Heidelberg 1978.
8. *Winnacker-Küchler*, 5th ed., **3**, 682–760, Wiley-VCH, Weinheim, Germany 2005.

Specific Reference

9. L. J. Thenard, *Ann. Chim. Phys.* **8** (1818) 306.
10. H. Meidinger, *Ann. Chem. Pharm.* **88** (1853) 57.
11. H. Berthelot, C. R. Hebd. *Seances Acad. Sci.* **86** (1878) 71.
12. K. Elbs, O. Schönherr, Z. Elektrochem. **1** (1895) 417, 468.
13. W. M. Weigert, H. Delle, G. Käbisch, *Chem. Ztg.* **99** (1975) 101.
14. W. Manchot, *Liebigs Ann. Chem.* **314** (1901) 177;
    316 (1901) 318, 331.
15. J. H. Walton, G. W. Filson, *J. Am. Chem. Soc.* **54** (1932) 3228.
16. O. v. Schickh, *Chem. Ing. Tech.* **32** (1960) 462.
17. J. Lacson, S. Schlag and G. Toki: "CEH Product Review Hydrogen Peroxide" in *Chemical Economics Handbook*, SRI International, Menlo Park 2002, 741. 5000A.
18. W. R. Busing, H. A. Levy, *J. Chem. Phys.* **42** (1965) 3054.
19. P. A. Giguere, *J. Chem. Ed.* **60** (1983) 399.
20. C. L. Yaws, H. S. N. Setty, *Chem. Eng.* (1974) 67.
21. G. Scatchard, G. M. Karanagh, L. B. Ticknor, *J. Am. Chem. Soc.* **74** (1952) 3715.
22. W. M. Latimar: *The Oxidation States of the Elements and their Potentials in Aqueous Solutions*, 2nd ed., Prentice Hall, New York 1952.
23. Hydrogen Peroxide for Industrial Use – Determination of Content, International Organization for Standardization, ISO/DIS 7157.
24. Degussa, DE 1 039 500, 1956 (G. Käbisch, E. Richter, A. Gross).
25. I.G. Farbenindustrie AG, GB 508 081, 1939; US 2 215 883, 1939.
26. Degussa, DE-AS 1 945 750, 1969 (G. Käbisch, R. Trübe).
27. EDOGAWA, DE 1 112 051, 1957 (K. Hiratsuka, S. Matsumura, M. Matsuda, K. Takagi, T. Abe).
28. Laporte Chemicals Ltd., GB 747190, 1953 (C. W. Lefeuvre).
29. Akzo Nobel, WO 2001098204, 2000 (M. Nystroem, C. Jaernvik).
30. Degussa, DE 1 261 838, 1963 (G. Käbisch).
31. RO 108956, 1991 (M. Mihalcescu, C. Andrei, C. Dragoi).
32. E. I. du Pont de Nemours & Co., DE 888 840, 1950 (C. R. Harris, J. W. Sprauer).
33. BECCO, US 2 537 655, 1950 (L. H. Dawsey, C. K. Muehlhauser, R. R. Umhoefer).
34. Laporte Chemicals Ltd., DE 933 088, 1953 (C. W. Lefeuvre).
35. E. I. du Pont de Nemours & Co., US 4046868, 1973 (L. G. Vaughan).
36. Oxysnthese, EP 287 421, 1988 (C. Pralus).

37. Kemira, US 5 114 701, 1990 (R. Aksela, E. Suokas).
38. EKA, EP 284 580, 1988 (U. M. Andersson, E. A. Bengtsson).
39. Degussa, DE 3 027 253, 1980 (G. Goor, W. Kunkel).
40. FMC Corp., US 4 394 369, 1983 (R. Wayne).
41. EKA, EP 286 610, 1988 (U. M. Andersson, E. A. Bengtsson).
42. Degussa, DE 2 018 686, 1970 (G. Giesselmann, G. Schreyer, W. Weigert).
43. J. Fang, *Huagong Keji* **6** (1998) 18–23.
44. I. G. Farbenindustrie AG, US 2 215 883, 1937 (H.-J. Riedl, G. Pfleiderer).
45. I.G. Farbenindustrie AG, US 2 158 525, 1936 (H.-J. Riedl, G. Pfleiderer).
46. E. I. du Pont de Nemours & Co., US 2 689 169, 1949 (W. S. Hinegardner).
47. BASF, EP 348464, 1990 (M. Eggersdorfer, J. Henkelmann, W. Grosch).
48. Kawasaki Kasei, JP 58180452, 1983 (S. Numata).
49. Montecatini, IT 601 435, 1960 (C. Ferri, A. Niglierina, B. Ponzini).
BASF, DE 43 39 649, 1993 (H. Hiller).
50. Degussa, DE 1 063 581, 1956 (L. R. Darbee, D. F. Kreuz).
51. Kvaerner, EP 958 238, 1997 (E. Bengtsson, A. Casen, A. Unum).
52. Degussa, US 6 153 169, 1998 (J. Glenneberg, G. Goor, E. Staab, H. Angert).
53. Akzo Nobel, EP 1101733, 1999 (M. Nystroem, C. Jaernvik, H. Thor, S. Saari).
54. EDOGAWA, DE-AS 1 085 861, 1957 (K. Hiratsuka et al.).
55. E. Santacesaria, R. Ferro, S. Ricci, S. Carra, *Ind. Eng. Chem. Res.* **26** (1987) 155;
E. Santacesaria, M. Di Serio, A. Russo, U. Leone, R. Velotti, *Chem. Eng. Sci.* **54** (1999) 2799–2806.
56. P. Oinas, G. Wild, N. Midoux, H. Haario, *Chem. Eng. Proc.* **34** (1995) 503.
57. A. Golja, M. Krajnc, S. Flajs, M. Grcar: Zbornik Referatov s Posvetovanja Slovenski Kemijski Dnevi, Maribor, Slovenia, Sept. 20-21, 2001, Issue Part 2, 723–729.
58. L. Kurc, V. Vopravil, L. Cerveny, *Res. Chem. Intermed.* **27** (2001) 249–258.
Degussa, DE 10114 982, 2001 (T. Haas, J. Glenneberg, R. Wagner, R. Vanheertum).
59. Laporte Chemicals Ltd., GB 741 444, 1953 (W. R. Holmes, C. W. Lefeuvre, R. Lait, W. S. Wood).
60. Solvay & Cie., DE 1 041 928, 1955 (M. Gonze, E. Leblon).
61. Degussa, DE 1 051 257, 1957 (G. Käbisch, E. Richter).
62. E. I. du Pont de Nemours & Co., US 2 673 140, 1951 (J. W. Sprauer).
63. EDOGAWA, FR 1 319 025, 1962.
64. EDOGAWA, DE-OS 2 141 234, 1971 (H. Shin et al.).
65. J. Petr, L. Kurc, Z. Belohlav, L. Cerveny, *Chem. Eng. Proc.* **43** (2004) 887–894.
66. E. I. du Pont de Nemours & Co., US 4 668 499, 1986 (J. D. Rushmere).
67. Solvay Interox, EP 546 616, 1992 (V. Deremince, C. Vogels).
68. T. Berglin, N. H. Schöön, *Ind. Eng. Chem. Process Des. Rev.* **22** (1983) 150–153.
69. EKA, DE 2 419 534, 1973 (B. G. Franzen).
70. Degussa, EP 1274651, 2000 (T. Haas, J. Glenneberg, R. Wagner, R. Vanheertum).
71. Degussa, DE 1 273 499, 1964 (G. Schreyer, O. Weiberg, G. Giesselmann).
72. Degussa, WO 2005 014471, 2004 (G. Goor, E. Staab, J. Glenneberg).
73. F. Sandelin, P. Oinas, T. Salmi, H. Haario, *Stud. Surf. Sci. Catal.* **88** (1994) 635.
74. A. Drelinkiewicz, *J. Mol. Catal. A* **101** (1995) 61;
Bull. Pol. Acad. Sci. Chem. **43** (1995) 51.;
A. Drelinkiewicz, R. Laitinen, R. Kangas, J. Pursiainen, *Appl. Catal. A: General* **284** (2005), 59–67.
75. Solvay Interox, WO 98 15 350, 1997 (V. Mathieu, P. Pennetreau, N. Vanlautem).
76. BASF, DE 801 840, 1951 (G. Pfleiderer, H.-J. Riedl, W. Deuschel).
77. FMC Corp., US 2 720 532, 1952 (R. R. Umhoefer).
78. FMC Corp., US 2 730 533, 1952 (R. R. Umhoefer).
79. Kali-Chemie, DE-AS 1 137 719, 1960 (V. Hauschild, M. Stahl).
80. J. Fang et al., *J. Catal.* **229** (2005) 97–204.
81. Degussa, DE 1 542 089, 1965 (G. Käbisch, H. Herzog).
Degussa, US 4 428 923, 1982 (W. Kunkel, J. Kemnade, D. Schneider).
82. Degussa, WO 2000 015550, 1999 (H. Eickhoff, R. Schütte).
83. Degussa, EP 1 334 062, 2001 (B. Maurer, P. Korl).
84. Degussa, GB 1 096 260, 1965.
85. Degussa, EP 687 649, 1995 (H. Birkenbeil, U. Brand, G. Goor, W. Kunkel).
86. Degussa, EP-A 571 762, 1993 (H. Eickhoff, R. Goedecke, H. Huber, K. Mangartz, R. Schirrmacher).

87. Laporte Chemicals Ltd., GB 718 306, 1953 (C. A. Morgan).
88. Du Pont, US 5 772 977, 1995 (C.L. Jenkins, F. B. Kirby, T. A. Koch).
89. Mitsubishi, DE 19713376, 1997 (K. Ogasawara, T. Kato, N. Okuda, T. Konishi, K. Kato).
90. Laporte Chemicals Ltd., GB 746 385, 1953 (C. W. Lefeuvre, C. A. Morgan); GB 922 022, 1959 (R. Lait, D.R. Lloyd-Owen).
91. Laporte Chemicals Ltd., GB 718 307, 1952 (J. A. Williams, C. W. Lefeuvre).
92. Du Pont, US 4 374 820, 1981 (T. E. Guenter).
93. Solvay, WO 99 040024, 1998 (D. Vandenberg, J. P. Ganhy, N. Vanlautem).
94. FMC Corp., US 3 009 782, 1958 (D. H. Porter).
95. EKA, US 5 063 043, 1990 (E. A. Bengtsson). EKA, US 4 552 748, 1983 (T. Berglin, W. Herrmann).
96. A. Drelinkiewicz, A. Pukkinen, R. Kangas, R. Laitinen, *Catal. Lett.* **94** (2004) 157–170.
97. A. Drelinkiewicz, R. Laitinen, R. Kangas, J. Pursiainen, *Appl. Catal. A: General* **284** (2005) 59–67.
98. FMC Corp., US 4 428 922, 1982 (Q. G. Hopkins).
99. Kemira, US 5 071 634, 1990 (T. Maunula, E.-L. Mustonen, I. Turunen, P. Virta); Kemira, WO 93 02 960, 1992 (E. Maekiniemi, T. Maunula, A. Pukkinen, I. Turunen).
100. Q. Wang, L. Wang, Y. Wang, F. Hei, Z. Li, Z. Mi, *React. Kinet. Catal. Lett.* **81** (2004) 297–304.
101. EKA, EP 778 085, 1996 (M. Nystroem, M. Silverstroem).
102. Degussa, DE 2 003 268, 1970 (M. Liebert, H. Delle, G. Käbisch).
103. *Chem. Process Eng.* (London) **40** (1959) no. 1, 5.
104. *Br. Chem. Eng.* **4** (1959) 88.
105. Österr. Chemische Werke, WO 86 06710, 1986 (J. Kemnade, B. Maurer).
106. Kemira, DE-OS 4 029 784, 1990 (I. Turunen, E. L. Mustonen).
107. Allied Chemical Corp., US 2 902 347, 1954 (J. N. Cosby, R. L. Harris, E. H. Rissmiller).
108. Degussa, WO 2000 017098, 1999 (H. Eickhoff, R. Schütte).
109. S. Lue, Z. Mi, Y. Wang, L. Wang, *Chem. Eng. Technol.* **27** (2004) 903–908.
110. Pittsburgh Plate Glass Co., DE 977 130, 1965 (J. R. Cox, T. E. Brown).
111. EKA, EP 078 780, 1982 (B. G. Franzen, W. Herrmann).
112. Kemira, EP-A 726 227, 1996 (I. Turunen).
113. Research Disclosure 40735, March 1998 (Derwent No. 98-205225).
114. Pittsburgh Plate Glass Co., US 3 055 838, 1957 (J. W. Moore).
115. E. I. du Pont de Nemours & Co., US 2 739 875, 1951 (J. W. Sprauer, T. V. Williams).
116. Foret, US 4 566 998, 1981 (V. M. Ochoa, R. Pueyo).
117. Mitsubishi Gas, DE-OS 19 715 034, 1997 (K. Ogasawara, T. Kato, K. Minato, K. Kato, M. Sumitani).
118. FMC, US 6 180 804, 1998 (E. F. Dougherty).
119. Laporte Chemicals Ltd., DE-AS 1 138 743, 1959 (B. K. Howe, S. Andrus, G. N. Davis).
120. Laporte Chemicals Ltd., GB 991 413, 1961 (B. K. Howe).
121. FMC Corp., US 4 824 609, 1988 (D. S. Sethi).
122. FMC Corp., US 4 668 436, 1985 (D. S. Sethi).
123. Du Pont Australia, WO 95 04 702, 1994 (P. J. Collins, P. B. Humphries, G. J. J. Prendergast).
124. Columbia-Southern Chemical Corp., GB 795 433, 1955 (J. R. Cox, M. Korach).
125. Pittsburgh Plate Glass Co., GB 924 625, 1958 (M. Q. Freeland, P. R. Johnston).
126. Mitsubishi Gas, US 4 999 179, 1989 (Y. Sugihara, S. Shimokawa).
127. E. I. du Pont de Nemours & Co., DE-AS 1 036 225, 1956 (J. W. Sprauer).
128. Laporte Chemicals Ltd., GB 841 323, 1956 (H. B. Hopkins).
129. FMC Corp., US 3 043 666, 1960 (G. W. Siwinski).
130. Laporte Chemicals Ltd., DE-AS 1 467 091, 1964 (S. D. Williams).
131. E. I. du Pont de Nemours & Co., US 3 016 290, 1956 (A. A. Elston, W. S. Hinegardner).
132. Oxysynthèse, FR 1 539 843, 1967 (P. Thirion).
133. Foret, ES 547 571, 1985 (R. P. Gracia, V. O. Bendicho, J. M. Lopez Martinez).
134. E. I. du Pont de Nemours & Co., US 4 879 043, 1988 (J. H. Boughton, R. A. Butz, H. C. T. Cheng, J. R. Dennis, B. T. Hannon, J. H. Weigel).
135. Kemira, EP-A 703 189, 1995 (I. Turunen).
136. Mitsubishi Gas, EP 774 442, 1996 (M. Hattori, Y. Minamikawa, S. Murakami).
137. Ausimont, EP 930 269, 1999 (U. P. Bianchi, U. Leone, M. Lucci).
138. Mitsubishi Gas, US 5 534 238, 1995 (S. Kajiwara, H. Serizawa).
139. Solvay Interox, DE-OS 4 222 109, 1992 (H. Honig).

140. Solvay Interox, WO 92 06 918, 1991 (J. R. Crampton, F. R. F. Hardy, M. H. Millar, G. W. Morris).
141. Sumitomo, US 5 906 738, 1994 (Y. Matsuda, A. Morisaki, Y. Sawaguri).
142. Solvay, EP 1520839, 2003 (R. Owen, J. Bosse, M. Sell).
143. E. I. du Pont de Nemours & Co., US 4 985 228, 1990 (K. Kirksey).
144. Solvay Interox, WO 90 11 967, 1990 (N. D. Feasey, G. W. Morris).
145. FMC Corp., US 5 266 298, 1992 (T. Frianeza-Kullberg, D. Gibilisco, J. L. Manganaro, J. R. Reed).
146. Laporte Chemicals Ltd., US 3 073 755, 1963 (R. H. Banfield, A. M. Hildon).
147. Columbia-Southern Chemical Corp., US 2 960 448, 1960 (W. S. Straub, C. J. Sindlinger).
148. Degussa, US 3 060 105, 1962 (J. Müller); Degussa, US 3 714 342, 1970 (G. Käbisch).
149. FMC Corp., US 3 152 052, 1964 (T. M. Jenney, D. H. Porter, R. H. Rice).
150. Sulzer, EP 419 406, 1990 (A. Meili). Sulzer, EP 835 680, 1996 (A. Meili).
151. Kemira, EP-A 727 388, 1996 (P. Johnsson, T. Mattila, K. Saari).
152. Solvay Interox, EP 529 722, 1992 (L. Signorini, K. Glinos). Solvay, EP 1 090663, 1999 (S. Bloomfield, G. E. Williams).
153. FMC, US 2 724640, 1947 (G.G. Crewson, J. R. Ryan).
154. Degussa, WO 2002 044083, 2001, (M. Creutz, S. Gross, S. Nordhoff, R. Wagner).
155. E. I. du Pont de Nemours & Co., US 2 479 111, 1949 (C. R. Harris).
156. Shell Development Co., US 2 871 104, 1951 (F. F. Rust).
157. N. V. De Bataafsche Petroleum Maatschappij, GB 708 339, 1954.
158. N. V. De Bataafsche Petroleum Maatschappij, BE 1 002 295, 1953 (J. R. Skinner, S. E. Steinle).
159. *Chem. Eng. News* **35** (1957) no. **48**, 26.
160. *Chem. Week* **125** (1979) no. **18**, 13.
161. *Ullmann*, 4th ed., **17**, 694.
162. E. J. Pneva, *Chimitscheskoja Promyschlenost* **9** (1976) 19.
163. Khimprom, SU 1 685 865, 1991 (V. V. Chikurov, I. A. Militsin, L. K. Zharikov).
164. ARCO, US 4 897 252, 1989 (R. N. Cochran, L. M. Candela).
165. ARCO, US 4 975 266, 1989 (R. S. Albal, R. N. Cochran, L. M. Candela).
166. ARCO, US 5 039 508, 1990 (R. N. Cochran, L. M. Candela).
167. D. F. Dong, J. R. Jackson, *Int. Forum Electrolysis Chem. Ind.*, 9th Electrosynthesis, Lancaster, NY 1995, p. 215–228.
168. P. Askenasy, R. Rose, *Z. Anorg. Allg. Chem.* **189** (1930) 1, 10.
169. J. Lawrence, R. B. Peacock, C. W. Lefeuvre, *Bios Final Report* **1432**, (1946), no22.
170. D. J. Campbell, E. F. Lefson, US 2 497 812, 1950.
171. Atlantic Richfield Co., US 3 891 748, 1973.
172. Atochem, Oxysynthese, EP 237 402, 1987 (J.-P. Schirmann, M. Pralus).
173. Wacker-Chemie, US 312 421, 1964 (W. Lohringer, J. Sixt).
174. Celanese Corp., US 3 341 297, 1967 (A. F. MacLean, A. L. Stautzenberger).
175. Elektrochem. Werke München, US 2 015 040, 1935 (A. Pietzsch).
176. Mitsubishi Denki K. K., DE-OS 4 106 980, 1991 (S. Endoh, K. Namba, S. Yagi, K. Maeda).
177. *Chemical Week* **1987**, December 9, 20.
178. E. I. du Pont de Nemours & Co., US 4 681 751, 1985 (L. W. Gosser).
179. *Chem. Eng. News*, 17 Oct 2005, **83** (42), p 13.
180. Atomic Energy of Canada Ltd., EP 366 419, 1989 (K. T. Chuang).
181. J. R. Kosak, *Chem. Ind.* (Dekker) **62** (1995) 115–124.
182. Mitsubishi Gas Chemical Company Inc., EP 492 064 (Y. Hiramatsu, Y. Ishiuchi, H. Nagashima).
183. Mitsubishi Gas Chemical Company Inc., EP 498 166 (Y. Hiramatsu, Y. Ishiuchi, H. Nagashima).
184. Hydrocarbon Technologies Inc., US 6 168 775 (B. Zhou, L-K. Lee).
185. E. I. du Pont de Nemours & Co., US 5 194 242 (M. A. Paoli).
186. Advanced Peroxide Technology Inc., WO 98/31629 (H. Huckins).
187. Atofina, WO 01/05498 (M. Devic).
188. Hydrocarbon Technologies Inc., US 2004/0018143 A1 (B. Zhou, M. Rueter).
189. Eka Chemicals AB, WO 99/32398 (M. Nyström, J. Wanngard, W. Herrmann).
190. Eka Chemicals AB, WO 99/32399 (M. Nyström, J. Wanngard, W. Herrmann)
191. Halcon, US 4 462 978, 1982 (W. F. Brill).
192. BOC, US 4 711 772, 1986 (S. E. Jacobson).
193. Enichem, EP-A 788 998, 1997 (R. Bortolo, R. D'Aloisio, D. Bianchi).

194. Enichem, EP-A 808 796, 1997 (D. Bianchi, R. Bortolo, R. D'Aloisio, M. Ricci, S. Soattini). D. Bianchi, R. Bortolo, R. D'Aloisio, M. Ricci, *Angew. Chem.* **111** (1999) 734; *J. Mol. Catal. A: Chem* **150** (1999) 87.
195. C. Querci, R. D'Aloisio, R. Bortolo, M. Ricci, D. Bianchi, *J. Mol. Catal. A: Chem.* **176** (2001) 95–100.
196. Solvay Interox, US 5 545 391, 1992 (P. A. Clapp, D. F. Evans, A. F. E. McKenna).
197. Canadian Industries, BE 660 719, 1965 (D. Hudson, D. R. Lusby, F. Van Zeggeren).
198. Kymmene AB, DE 1 054 970, 1954 (R. M. R. Örnhjelm).
199. Kymmene AB, DE 1 036 829, 1955 (R. M. R. Örnhjelm).
200. Kymmene AB, DE 1 063 131, 1956 (R. M. Rafaelson).
201. T. S. Sheriff, *J. Chem. Soc. Dalton Trans.* 1992, 1051.
202. M. Traube, *Ber. Dtsch. Chem. Ges.* **15** (1882) 2434.
203. F. Fischer, O. Priess, *Ber. Dtsch. Chem. Ges.* **46** (1913) 698.
204. B. Kastening, W. Faul, *Chem. Ing. Tech.* **49** (1977) no. 11, 911.
205. J. A. McIntyre, R. F. Philips in T. R. Beck et al. (eds.): *Electrochemical Process and Plant Design*, Pennington, NJ, 1983, pp. 79–97.
206. Dow Chemical, EP 86 896, 1982 (J. A. McIntyre, R. F. Philips).
207. H-D Tech, EP 216 428, 1986 (A. L. Clifford, D. F. Dong).
208. Metallgesellschaft, WO 94 24 336, 1994 (E. Hillrichs, M. Kienberger, U. Sander).
209. Occidental, US 4 384 931, 1981 (R. J. Jasinski, C. G. Kuehn).
210. P. C. Foller, R. T. Bombard, *J. Appl. Elektrochem.* **25** (1995) 613–627.
211. C. Oloman: *Electrochemical Processing for the Pulp and Paper Industry*, The Electrochemical Consultancy Ltd., Romsey, Hants, England, 1996, pp. 143–152.
212. H-D-Tech, US 5 106 464, 1990 (D. J. Rogers, A. L. Clifford).
213. S. P. Webb, J. A. McIntyre: "Generation of Hydrogen Peroxide in a Shorted Fuel Cell", *Int. Forum Electrolysis Chem. Ind.*, 10th Electrosynthesis, Lancaster, NY 1996, p. 213–225.
(1996), p. 213–225.
214. Dow Chemical, WO 95 30 474, 1995 (J. A. McIntyre, S. P. Webb);
Dow Chemical, WO 97 13 006, 1996 (J. A. McIntyre, S. P. Webb).
215. Laporte Chemicals Ltd., GB 893 069, 1959 (P. V. Rutledge, D. S. Bell).
216. Columbia-Southern Chemical Corp., US 2 904 517, 1959 (P. H. Baker).
217. S. Croft, B. C. Gilbert, J. R. L. Smith, J. K. Stell, W. R. Sanderson, *J. Chem. Soc. Perkin Trans.* **2** 1992, 153.
218. E. Rabald, D. Behrens (ed.), *DECHEMA-Werkstoff-Tabelle Wasserstoffperoxid*, DECHEMA, Frankfurt 1971; 43.
Ergänzungslieferung, DECHEMA, Frankfurt, 2000;
R. Bloom, L.E. Weeks, Ch. W. Raleigh, *Corrosion* **16** (1960) 164t–170t.
219. J. Mackenzie, *Chem. Eng.* **1990**, 84–90.
220. C. N. Satterfield, G. M. Kavanagh, H. Resnick, *Ind. Eng. Chem.* **43** (1951) no. 11, 2507.
221. J. M. Monger, H. J. Baumgartner, G. C. Hood, C. E. Sanborn, *Journ. Chem. Eng. Data* **9** (1964) no. 1, 119.
222. E. S. Shanley, P. F. Strong, L. R. Darbee, N. D. Lee, *Ind. Eng. Chem. Fundam.* **9** (1970) no. 4, 574.
223. C. N. Satterfield, P. J. Cocotti, A. H. R. Feldbrugge, *Ind. Eng. Chem.* **47** (1955) 1040.
224. *Bur. Mines Inf. Circ.* **8387** (1968).
225. E. Haeuseler, *Explosionsstoffe* **6/7** (1953) 64.
226. a) N. Nimmerfroh, D. Pauls, S. McMahon, Propulse HTP Manufaxture: Quality, Transportation and Handling, 36 th AIAA / ASME / SAE / ASEE Joint Propulsion Conference, Huntsville / AL, 2000; b) N. Nimmerfroh, E. Walzer, C. Brossmer, European Space Agency, Special Publication SP-484 (2001) 65–73.
227. E. S. Shanley, F. P. Greenspan, *Ind. Eng. Chem.* **39** (1947) no. 12, 1536.
228. J. M. Monger, H. Selbo, D. C. Lehwalder, *J. Chem. Eng. Data* **6** (1961) no. 1, 23.
229. C. W. Jones: *Applications of Hydrogen Peroxide and Derivatives, RSC Clean Technology Monographs*, The Royal Society of Chemistry, Cambridge 1999.
230. *Bur. Mines Rep. Invest.* **5877** (1961).
231. O. Klais, *Thermochimica Acta* **225** (1993) 213–222.
232. P. Urban, L. Bretherick: *Handbook of Reactive Chemical Hazards*, 6th ed., Butterworth-Heinemann, Oxford 1999.
233. Shell Chemical Company, *Concentrated Hydrogen Peroxide*, Summary of Research Data on Safety Limitations, New York 1961.

234. A. Schreck, A. Knorr, K. D Wehrstadt, P. A. Wandrey, T. Gmeinwieser, J. Steinbach, *Journal of Hazardous Materials* **A108** (2004) 1–7.
235. *PetroChem. News* **43** (2005) no. 44, 1.
236. European Center of Ecotoxicology and Toxicology of Chemicals (ECETOC), Special Report No. 10, Hydrogen Peroxide OEL Criteria Document, ECETOC, Brussels 1996.
237. S. A. Kelly, C. M. Havrilla, T. C. Brady, K. H. Abramo, E. D. Levin, *Env. Health Perspect.* **106** (1998) 375–384.
238. CEFIC Peroxygens hydrogen peroxide risk assessment task force, Preliminary Experiments for the Determination of Hydrogen Peroxide in Blood, unpublished report, (1999).
239. European Union, Institute for Health and Consumer Protection, European Chemicals Bureau, Existing Substances 2nd priority list 38 (2003) Risk Assessment Report Hydrogen peroxide, European Commission Joint Research Centre, EUR 20844 EN.
240. CEFIC Peroxygen Sector Group, Hydrogen Peroxide: 28-day Inhalation Study in Rats unpublished report (2002).
241. V. Riihimäki, A. Toppila, P. Piirilä, E. Kuosoma, P. Pfäffli, P. Tuomela, *J. Occup. Health* **44** (2002) 433–438.
242. A. Romanovski, J. R. Murray, M. J. Huston, *Pharm. Acta Helv.* **35** (1960) 354–357.
243. CEFIC Peroxygen Sector Group: Hydrogen Peroxide 13-Week Drinking Water Study with 6-Week Recovery Period in C57BL/6NCrlBR Mice, unpublished report, (1997).
244. M. Fenech, J. Crott, J. Turner, S. Brown, *Mutagenesis*, **14** (6) (1999), 605–612.
245. CEFIC peroxygen sector group, Coming Hazelton Report No. 514/24, 1997 (unpublished).
246. Threshold Limit Values for Chemical Substances and Biological Exposure Indices, ACGIH, Cincinnati 2005.
247. Sociaal Economische Raad (NL), MAC Waarden, Waterstoffperoxide 90 % (2005).
248. Koninklijk besluit van 11 maart 2002 betreffende de bescherming van de gezondheid en de veiligheid van de werknemers tegen de risico's van chemische agentia op het werk (B.S. 14.3.2002, ed. 2; erratum B.S. 26.6.2002, ed. 2; erratum B.S. 4.12.2002) BIJLAGE I Grenswaarden voor beroepsmatige blootstelling.
249. Réferences françaises, INRS, ND2098, Valeurs limites d'exposition professionnelle aux agents chimiques en France (2005).
250. TRGS 900, Technische Regeln für Gefahrstoffe, Grenzwerte in der Luft am Arbeitsplatz, Bekanntmachung des BMA Ausschuß für Gefahrstoffe, BArbBl 7/8 (2004).
251. Health and Safety Executive, EH40/ 2005 Occupational Exposure Limits (2005).
252. European Center of Ecotoxicology and Toxicology of chemicals (ECETOC), Joint Assessments of Commodity Chemicals No. 22, Hydrogen Peroxide, ECETOC, Brussels 1993.
253. R. P. Wayne: The photochemistry of ozone. In: O. Hutzinger (ed.): The handbook of environmental chemistry 2E, reactions and processes, Springer, Berlin, 10 (1988).
254. H. Sakugawa, I. R. Kaplan, W. Tsai, and Y. Cohen, *Envir Sci. Technol.* **24** (1990) 1452–1462.
255. J. G. Calvert and W. R. Stockwell, *Envir. Sci. Technol.* **17** (1983) 28A–443A.
256. W. C. Schumb, C. N. Satterfield, R. L. Wentworth, (eds): *Hydrogen Peroxide*. Reinhold Publishing, New York, 1955, 64–68, 181, 458–464, 535–547.
257. L. Kleinman, *J. Geophys. Res.* **91** (10) (1986) 889–904.
258. E. Hellpointner, S. Gaeb, *Nature* **337** (1989) 631–634.
259. W. J. Cooper, R. G. Zika, *Science* **220** (1983) 711–712.
260. R. G. Zika, J. W. Moffett, R. G. Pestane, W. J. Cooper, E. S. Saltzman, *Geochim. Cosmochim. Acta* **49** (1985) 1173–1184.
261. J. M. Moffett, and R. G. Zika, *Environ. Sci. Technol.* **21** (1987) 804–810.
262. W. J. Cooper and D. R. S. Lean, *Environ. Sci. Technol.* **23** (11) (1989) 1425–1428.
263. W. Guhl, A. G. Berends, *Surfactants* **2** (2000) 98–102.
264. A. H. C. Groeneveld, W. A. de Groot: Activated sludge, respiration test with hydrogen peroxide, Unpublished report (1999), Solvay pharmaceuticals, report A.SOL.S.003.
265. Degussa AG, US-IT No. 91-0114-DKO, 1991 (unpublished).
266. C. A. Clark: Ph. D. Thesis, University of Bath, 1991.
267. M. Fitzpatrick, C. B. Schreck, R. L. Chitwood, *The Progressive Fish Culturist*, **57** (1995) 153–155.
268. P. Klerks, P. C. Fraleigh, *J. AWWA* 1991, 92–100.
269. M. A. Xenopoulos, D. F. Bird, *Photochem. Photobiol.* (1997) 471–478.

Hydrogen Sulfide

Francois Pouliquen, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 1)
Claude Blanc, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 2)
Emmanuel Arretz, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chaps. 3 and 8)
Ives Labat, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 3)
Jacques Tournier-Lasserve, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 4)
Alain Ladousse, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 5)
Jean Nougayrede, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chaps. 5 and 8)
Gérard Savin, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chaps. 6 and 8)
Raoul Ivaldi, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 7)
Monique Nicolas, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 9)
Jean Fialaire, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 10)
René Millischer, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 10)
Charles Azema, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 11)
Lucien Espagno, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 11)
Henri Hemmer, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 11)
Jacques Perrot, Elf Aquitaine, Tour Elf, Paris La Defense, France(Chap. 11)

1. Introduction 2
2. Physical Properties 2
3. Chemical Properties 4
3.1. Acid Properties 4
3.2. Reducing Properties 4
3.3. Reactions with Organic Compounds 4
3.3.1. Hydrogen Sulfide as a Nucleophilic Reagent 4
3.3.2. Hydrogen Sulfide as a Radical Reagent 6
3.3.3. Reaction of Hydrogen Sulfide in the Form of Hydrosulfide 6
4. Production 7
4.1. Production by Chemical Reaction . 7
4.2. Recovery from Gas 7
5. Environmental Protection 9
6. Analysis 11
7. Storage, Handling, and Transportation 12
7.1. Storage and Handling 12
7.2. Transportation 12
8. Uses 12
9. Economic Aspects 13
10. Toxicology 14
11. Occupational Health and Safety . . 15
12. References 18

1. Introduction

Hydrogen sulfide was first described as a natural compound by alchemists in the Middle Ages. The successful development of the sour gas field at Lacq (France) by the Société Nationale des Pétroles d'Aquitaine in 1957 modified the economy of hydrogen sulfide and allowed worldwide exploitation of natural gas with a high hydrogen sulfide content. Before this, hydrogen sulfide was regarded as a toxic, corrosive component of natural gas that had to be eliminated. However, it has now become an important raw material for producing sulfur and thioorganic compounds.

Some of the data and indications given are the result of Elf Aquitaine's experience, either from the Lacq field or from various gas and oil desulfurization plants that the company has built around the world.

Natural Occurrence The quantity of hydrogen sulfide naturally released into the atmo-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_467

sphere is thought to vary annually from 100×10^6 to 324×10^6 t [1–3], with half of this coming from volcanoes [4,5], flooded ground [6–9], or hydrogeological sources [10,11] and the other half from oceans [12,13].

Hydrogen sulfide has been identified in the atmosphere of stars by spectroscopic methods. Calculations have been made of the thermal equilibrium of chemical species found in the atmosphere of several planets [14].

Hydrogen sulfide is present as an occluded gas in minerals such as topaz, tourmaline, water chrysolite, obsidian, and granite. The hydrogen sulfide found in ore-bearing deposits may originate from bacteria: its reducing properties account for the presence of uranium in the form of nonoxidized mineral deposits.

Salt mines also contain hydrogen sulfide, sometimes mixed with other gases. Volcanoes release large quantities of hydrogen sulfide in the fumaroles, and in gases and vapors produced from solfataras.

Flooded ground (swamps, rice fields) also contains hydrogen sulfide which is generally formed in the soil by bacterial reduction of sulfates. However, its occurrence may also depend on the type of soil, the presence of organic material, and the degree of aeration.

Water is another natural medium that contains hydrogen sulfide. The dissolved or gaseous hydrogen sulfide found in lakes, saltwater ponds, and marine sediments originates from sulfate-reducing bacteria. Mineral waters from thermal springs are studied either in vitro, to examine sulfate reduction by bacteria, or in situ, where the hydrodynamic effects (water velocity in soil, mineralization, and temperature) are coupled with the biological production of hydrogen sulfide. Wastewater of industrial [15] or other origin [16,17] often contains hydrogen sulfide.

Finally, hydrogen sulfide is found in hydrocarbon reserves: natural gas, associated gas, and oil. An extensive bibliography exists on this subject. Further information on particular reservoirs can be found in the following references: Federal Republic of Germany [18,19]; Lacq (France) [20,21]; Astrakhan (former Soviet Union) [22–32]; Orenbourg (former Soviet Union) [33–41]; United States [42, ; gives the composition of 349 reservoirs]; and People's Republic of China [43] (reservoir containing the highest known percentage, i.e., 90 %, of hydrogen sulfide). Problems encountered in the development of very sour oil or gas fields are also discussed in the literature: avoidance of formation of SO_4^{2-} in edgewater [44]; tools for studying hydrogen sulfide wells [45,46]; neutralization of hydrogen sulfide during drilling [47–52]; and reserves located at great depth [53] or under high pressure [19].

2. Physical Properties

Hydrogen sulfide [*7783-06-4*], H_2S, is also known as monosulfane, M_r 34.08. The gas is colorless, heavier than air, and has a characteristic rotten-egg odor.

Physical properties are listed below:

bp (101.3 kPa)	−60.28 °C
Triple point (23.2 kPa)	−85.49 °C
Critical temperature	100.25 °C
Critical pressure	8.96 MPa
Critical density	347.62 kg/m^3
Density of solid [54] at	
−195 °C	1.217 kg/m^3
−170 °C	1.166 kg/m^3
Density of liquid [55] at	
bp	949.0 kg/m^3
−30 °C	893.8 kg/m^3
9.7 °C	813.8 kg/m^3
90.4 °C	532.6 kg/m^3
Density of vapor [55] at	
−80 °C	0.71 kg/m^3
bp	1.99 kg/m^3
−30 °C	6.82 kg/m^3
56.9 °C	71.65 kg/m^3
Vapor pressure of [55] at	
−30 °C	0.381 MPa
0 °C	1.032 MPa
30 °C	2.275 MPa
60 °C	4.335 MPa
Viscosity of vapor (101.3 kPa) [56] at	
0 °C	0.01179 mPa · s
15 °C	0.01241 mPa · s
50 °C	0.01398 mPa · s
100 °C	0.01608 mPa · s
Viscosity of liquid (101.3 kPa) [57] at	
−11.5 °C	0.00169 mPa · s
15 °C	0.00133 mPa · s
30 °C	0.00116 mPa · s
50 °C	0.00094 mPa · s
Surface tension at	
−60.2 °C [58]	25.43 kN/m
25 °C [59]	11.3 kN/m
40 °C [59]	8.7 kN/m
Refractive index (n_D^0) of gas at 0.1 MPa and 0 °C [60]	1.000644
Dipole moment [61]	3.269×10^{-30} C · m (0.98 debye)
Latent heat of fusion	2.378 kJ/mol

Flammability limits in air (20 °C and 101.3 kPa) [63]	
Lower	4 vol %
Upper	44 vol %
Autoignition temperature	
at 101.3 kPa	290 °C
Thermal conductivity of gas (101.3 kPa) [64] at	
0 °C	1.340 kJ m^{-1} s^{-1} K^{-1}
80 °C	1.784 kJ m^{-1} s^{-1} K^{-1}
120 °C	2.010 kJ m^{-1} s^{-1} K^{-1}
200 °C	2.433 kJ m^{-1} s^{-1} K^{-1}

Compressibility factors of hydrogen sulfide gas and liquid are given in Table 1. The latent heat of vaporization and specific heat of the liquid are listed in Table 2. Table 3 gives specific heats of hydrogen sulfide gas. The solubilities of hydrogen sulfide in water and some organic solvents are listed in Table 4.

Table 1. Compressibility factor z of gaseous and liquid hydrogen sulfide [55]

Temperature, K	Pressure, MPa	z (gas)	z (liquid)
190.00	0.0271	0.99321	0.00059
230.00	0.2247	0.96164	0.00436
270.00	0.94015	0.8888	0.01697
310.00	2.6642	0.77786	0.04705
350.00	5.9719	0.60792	0.11345
370.00	8.4622	0.42049	0.19604

Table 2. Heat of vaporization (Q_{vap}) and specific heats (c_p, c_v) of liquid hydrogen sulfide [55]

Temperature, K	Q_{vap}, kJ/mol	c_p, J mol^{-1} K^{-1}	c_v, J mol^{-1} K^{-1}
190	19.52	67.53	43.89
230	17.96	69.64	42.78
270	15.89	71.67	39.37
300	13.95	75.20	35.45
330	11.50	86.83	31.04
360	7.37	163.40	30.80

Table 3. Specific heats of gaseous hydrogen sulfide [55]

Pressure, MPa	Temperature, K	c_p, J mol^{-1} K^{-1}	c_v, J mol^{-1} K^{-1}
0.05	200	34.09	25.38
	300	34.39	25.98
	400	35.70	27.34
0.1013	220	34.54	25.59
	300	34.55	26.02
	400	35.76	27.35
1	300	39.13	27.41
	400	37.13	27.65
	500	38.02	29.08
5	400	47.82	30.20
	500	41.62	29.86
	600	41.38	31.15

Table 4. Solubility of hydrogen sulfide in water and organic solvents

Solvent	Temperature, °C	Partial pressure of H_2S, MPa	Mole fraction of H_2S in liquid phase
Water [64]	30	1.718	0.0238
	50	1.708	0.0170
Methanol [65]	0	0.0533	0.0286
	−25.6	0.0533	0.0655
Propylene carbonate [66]	40	2.378	0.7246
	100	4.96	0.5459
N-Methyl-2-pyrrolidone [67]	23.5	0.0415	0.0580
	35	0.1013	0.0971
n-Decane [68]	15	0.1013	0.0541
	25	0.1013	0.0465

3. Chemical Properties

3.1. Acid Properties

Hydrogen sulfide in aqueous solution is a weak acid with two dissociation constants:

$$H_2S \rightleftharpoons HS^- + H^+ \qquad K_1 = 10^{-7}\ (18\,^\circ\mathrm{C})$$
$$HS^- \rightleftharpoons S^{2-} + H^+ \qquad K_2 = 10^{-13} - 10^{-14}$$

Saturation of the first acid function with sodium hydroxide solution is an exothermic reaction:

$$NaOH + H_2S \longrightarrow NaSH + H_2O \qquad \Delta H = -36.7\ \mathrm{kJ/mol}$$

Hydrogen sulfide can precipitate salts of heavy metals (Hg, Pb, Ag, Bi, Cu, Cd, As, Sb, Sn, Co, Ni, Fe, Zn, Mn, Al) as sulfides. It dissolves and reacts with amines and can be released by heating. Numerous processes for removing hydrogen sulfide from natural gas (i.e., gas sweetening processes) are based on its absorption by amines (see Chap. 5).

3.2. Reducing Properties

The standard redox potentials of hydrogen sulfide at 25 °C follow [69]:

$$S + 2H^+ + 2e^- \rightleftharpoons H_2S \qquad E^0 = +0.14\ \mathrm{V}\ (\mathrm{pH} = 0)$$
$$S + 2e^- \rightleftharpoons S^{2-} \qquad E^0 = -0.48\ \mathrm{V}\ (\mathrm{pH} = 14)$$

At pH 0, hydrogen sulfide is barely dissociated; at a pH of ca. 12, partial dismutation modification of the oxidation level of sulfur can occur and the sulfide is oxidized to form oxygen–sulfur compounds.

Hydrogen sulfide reacts with numerous oxidizing agents to form sulfur compounds with varying degrees of oxidation. Depending on operating conditions and the oxidizing species concerned, sulfur oxides (SO_2, SO_3), salts ($S_2O_3^{2-}$, SO_3^{2-}, SO_4^{2-}), polythionic compounds, sulfur, and sulfanes are obtained.

1) Aqueous solutions of hydrogen sulfide can be oxidized in air.
2) In air, solutions of ammonium sulfide form orange polysulfides and even oxygen – sulfur compounds.
3) In an acid medium, iodine quantitatively oxidizes hydrogen sulfide to sulfur. This reaction is used to analyze hydrogen sulfide.
4) Strong oxidizing agents (e.g., MnO_4^-, ClO^-, Cl_2, and Br_2) react with hydrogen sulfide to give sulfate ions.
5) Hydrogen sulfide can react with sulfur to form sulfanes. The sulfanes can be converted back to hydrogen sulfide by base catalysis [70].
6) Hydrogen sulfide or alkaline hydrosulfides can reduce polysulfides to thiols [71]:

$$R-S-S-R + 2\,NaSH \longrightarrow 2\,RSH + Na_2S_2$$

This reaction is used for the preparation of thiokol polymers.

7) Hydrogen sulfide can reduce sulfoxides to sulfides [72]:

$$R-SO-R + H_2S \longrightarrow R-S-R + H_2O + S$$

8) The reaction of hydrogen sulfide and sulfur dioxide (Claus reaction) is used for the production of sulfur (→ Sulfur):

$$2\,H_2S + SO_2 \longrightarrow 3/8\,S_8 + 2\,H_2O$$

3.3. Reactions with Organic Compounds

Hydrogen sulfide reacts with numerous organic compounds directly, as an intermediate (e.g., sodium, potassium, or ammonium hydrosulfide), or as an alkaline sulfide.

Hydrogen sulfide itself can behave as either a nucleophilic reagent or a radical reagent, whereas hydrosulfides or sulfides behave solely as nucleophilic reagents.

3.3.1. Hydrogen Sulfide as a Nucleophilic Reagent

Hydrogen sulfide reacts as a nucleophilic reagent in the presence of acidic, basic, or metallic oxide catalysts.

Compounds such as alcohols or olefins react with hydrogen sulfide to form thiols (RSH). Thiols (with the exception of tertiary thiols) react like hydrogen sulfide to give the corresponding organic sulfides (RSR).

Substitution Reactions. Because of its widespread industrial applications, the best known substitution reaction is *sulfhydrolysis* (*thiolation*) *of alcohols*. This reaction occurs catalytically in the gas phase at ca. 300 °C. Thiols were first obtained by reacting hydrogen sulfide with alcohols on thorin or on pumice impregnated with thorin [73]. Subsequently, more selective, more active catalysts were developed based on activated alumina containing promoters such as soda or potash [74], metallic oxides, alkaline salts of transition-metal oxo acids [75], and heteropolyacids or their alkaline salts [76].

Primary alcohols are converted very efficiently into the corresponding thiols:

$$H_2S + ROH \longrightarrow RSH + H_2O$$

However, with secondary and especially tertiary alcohols, dehydration reactions are more prevalent.

Under catalytic reaction conditions similar to those used for alcohols, hydrogen sulfide reacts with aliphatic ethers to produce thiols. Cyclic ethers are converted to the corresponding sulfides by catalysis in the gas phase. For example, reaction of hydrogen sulfide with tetrahydrofuran yields tetrahydrothiophene [*110-01-0*] [77]:

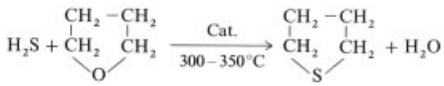

A similar reaction with furan gives thiophene [*110-02-1*] [78]:

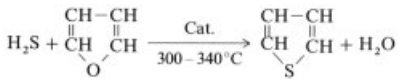

Hydrogen sulfide reacts with acetic anhydride in the presence of basic resins to form thioacetic acid [*507-09-5*] [79]:

$$H_2S + CH_3COOCOCH_3 \longrightarrow CH_3COSH + CH_3COOH$$

Addition Reactions. Catalytic addition of hydrogen sulfide to *olefins* follows Markovnikov's rule (i.e., the sulfhydryl group adds preferentially to the more highly substituted carbon of the double bond).

Linear olefins react with hydrogen sulfide in the presence of acid catalysts (e.g., acid zeolites, cation-exchange resins, alumina – silicas) to yield secondary thiols [80]. With branched olefins and acidic catalysts such as aluminum chloride, alumina – silica [81], zeolites [82], or cation-exchange resins [83], tertiary thiols are formed. Ethylene and cyclohexene are converted into thiols on metal oxide catalysts [84].

Hydrogen sulfide adds to the double bond of *acrylic derivatives* (nitrile, acids, esters) in the presence of basic catalysts such as amines [85], anion-exchange resins [86], or magnesium oxide [86].

Nucleophilic addition of the thiol function occurs at the carbon in the β-position with respect to the substituent group:

$$H_2S + CH_2{=}CHX \longrightarrow HSCH_2CH_2X$$
$$X = -CN,\ -COOH,\ -COOR$$

Base-catalyzed cleavage of *epoxides* with hydrogen sulfide gives mercaptoalcohols (hydroxythiols):

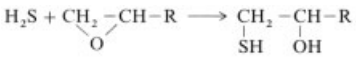

Basic catalysts may be mineral bases, amines, quaternary ammonium salts [87], anion-exchange resins [88], zeolites [89], or guanidines [90].

With ethylene oxide, the main product is mercaptoethanol [*60-24-2*]. Thiodiglycol or bis(2-hydroxyethyl) sulfide [*111-48-8*] is obtained as byproduct. The reaction can be catalyzed by thiodiglycol without a base [91].

The addition of hydrogen sulfide to *nitriles* is catalyzed by bases such as amines and gives thioamides:

An industrially important example of this reaction is the addition of hydrogen sulfide to cyanamide to form thiourea [*62-56-6*]:

With *aldehydes and ketones,* hydrogen sulfide yields dithiols or cyclic products, depending on conditions [92]:

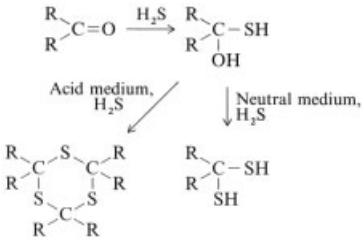

Thus, with formaldehyde, methanedithiol ($HSCH_2SH$) [*74-93-1*] is obtained [93]. Acetone reacts with sulfur and hydrogen sulfide in an ammoniacal medium to form a cyclic polysulfide [*4475-72-3*] [94]:

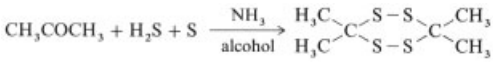

3.3.2. Hydrogen Sulfide as a Radical Reagent

Hydrogen sulfide can react as a free radical:

$$HS{-}H \longrightarrow HS\cdot + H\cdot$$

Generation of these two species by thermal dissociation is difficult. Use of chemical initiators or exposure to UV light is much more efficient.

Radical addition of hydrogen sulfide to olefins yields thiols. In practice, the reaction can be initiated as follows:

1) *Chemical initiators* can be used at temperatures corresponding to their thermal decomposition. Azo derivatives, in particular azobisisobutyronitrile, have been used [95] with

or without catalyst promoters such as organic phosphines and phosphites [96].

2) *Ultraviolet radiation* can be employed. Reaction proceeds either by direct photolysis ($\lambda = 253.7$ nm), or via intermediate radical initiators ($\lambda = 300-400$ nm). The initiators are either chemical initiators that are decomposed at a lower temperature by photons (e.g., azo derivatives [97], phosphites [98], acetophenone derivatives [99]) or aromatic carbonyl derivatives such as benzophenone [100] or thioxanthone [101], preferably associated with organic phosphites [100, 101].

Addition proceeds by a free-radical chain mechanism and is of the anti-Markovnikov type, i.e., the HS · radical adds to the least substituted carbon of the double bond.

With a terminal olefin, the corresponding primary thiol is selectively formed. However, the thiol may decompose into a thiyl radical that can also add to the olefin to give the corresponding organic sulfide. The percentage of sulfide formed can be minimized by using an excess of hydrogen sulfide.

Photochemical addition of hydrogen sulfide is also possible with olefins containing functional groups (e.g., allyl chloride, acrylic acid, and acrylates).

3.3.3. Reaction of Hydrogen Sulfide in the Form of Hydrosulfide

For certain ionic reactions, hydrogen sulfide must be supplied in the form of alkaline hydrosulfides, or ammonium hydrosulfide in an aqueous or aqueous – alcoholic solution.

Nucleophilic Substitution. Nucleophilic substitution of halogen compounds is facilitated if the halogen is activated by its electronic environment. Thus, nucleophilic substitution cannot be used to obtain thiol compounds from chlorinated hydrocarbons. However, if the chlorine is activated by an α, β, or γ heteroatom (O, Si, S), or by an electron-attracting group (C=O, C=N), substitution proceeds easily, often with total conversion of chlorine and excellent selectivity [102]:

$$2\,NaSH + Cl-(CH_2)_n-COOH \longrightarrow HS(CH_2)_n-COONa + H_2S + NaCl$$
$$\downarrow +HCl$$
$$HS(CH_2)_n-COOH + NaCl$$

This method is used to synthesize thioglycolic acid [*68-11-1*] and mercaptopropionic acid [*107-96-0*].

Other mercaptans are obtained by similar techniques [103, 104]:

$$NaSH + ClCH_2CH_2OR \longrightarrow HSCH_2CH_2OR + NaCl$$
$$NaSH + Cl(CH_2)_3Si(OCH_3)_3 \longrightarrow HS(CH_2)_3Si(OCH_3)_3 + NaCl$$
$$NaSH + C_6H_5CH_2Cl \longrightarrow C_6H_5CH_2SH + NaCl$$

Acid chlorides react particularly well with alkaline hydrosulfides [105]. For example, thiobenzoic acid [*98-91-9*] is synthesized by reaction of benzoyl chloride with sodium hydrosulfide:

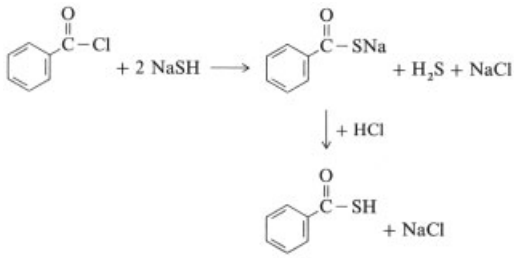

Some active esters may be subjected to sulfhydrolysis with hydrosulfides to form the corresponding thiols [106, 107]:

$$HSCH_2CH_2OCOCH_3 \xrightarrow[NH_3]{H_2S} HSCH_2CH_2SH + CH_3COOH$$

$$NH_2CH_2CH_2OSO_3Na \xrightarrow{NaSH} HSCH_2CH_2NH_2$$

Nucleophilic Addition. In an aqueous – alcoholic medium and in the presence of carbon disulfide, sodium hydrosulfide can react selectively with activated olefins to give thiols. Methyl mercaptopropionate [*2935-90-2*] has been prepared by this technique [108]:

$$NaSH + CH_2=CHCOOCH_3 \xrightarrow[\text{alcohol}]{CS_2} HSCH_2CH_2COOCH_3$$

4. Production

Hydrogen sulfide can be produced by (1) chemical reaction of sulfur with hydrogen or a hydrocarbon, or (2) hydrogen reduction or acid decomposition of a sulfide. It can also be recovered from natural gas, associated gases, and refinery gases.

4.1. Production by Chemical Reaction

Reaction of Sulfur with Hydrogen. Hydrogen sulfide is formed when hydrogen is fed into sulfur at 350 °C with or without a catalyst (e.g., bauxite, aluminum silicate, cobalt molybdate) [109, 110]. High pressure also aids the reaction.

$$H_2 + 1/2\,S_2 \rightleftharpoons H_2S$$

In the laboratory, hydrogen is fed into molten sulfur. Reaction occurs in a tube filled with pumice [111–113]. In the industrial production of hydrogen sulfide (Fig. 1), the reaction occurs at 450 °C and 0.7 MPa. Hydrogen sulfide is separated from sulfur vapor by cooling and solidifying the sulfur on the walls of a heat exchanger. The exchanger is alternately cooled and heated to remelt and eliminate the sulfur deposit. Operating data for a typical plant follow [114]:

Production capacity:	1 – 60 t/d
Consumption (per metric tonne of hydrogen sulfide produced):	
Hydrogen (purity 98 %)	67 kg
Sulfur (purity 99 %)	1066 kg
Electrical power	266 kW · h
Typical analysis of product	95 vol % H_2S, 3 vol % H_2, 2 vol % impurities

Reaction of Sulfur with Hydrocarbons. Sulfur reacts with methane to form a mixture of hydrogen sulfide and carbon disulfide. Carbon disulfide can be hydrolyzed, giving additional hydrogen sulfide [110], [115, 116]. With higher alkanes ($>C_{10}$) having a low vapor pressure, reaction takes place at 200 – 300 °C [115, 117]. Examples of laboratory preparation are given in [118, 119].

Reduction of Sulfides. Hydrogen sulfide is formed by reduction of sulfides (e.g., FeS_2, CuS) with hydrogen at > 500 °C.

Reaction of Sulfides with Dilute Acid. The most convenient method of preparing hydrogen

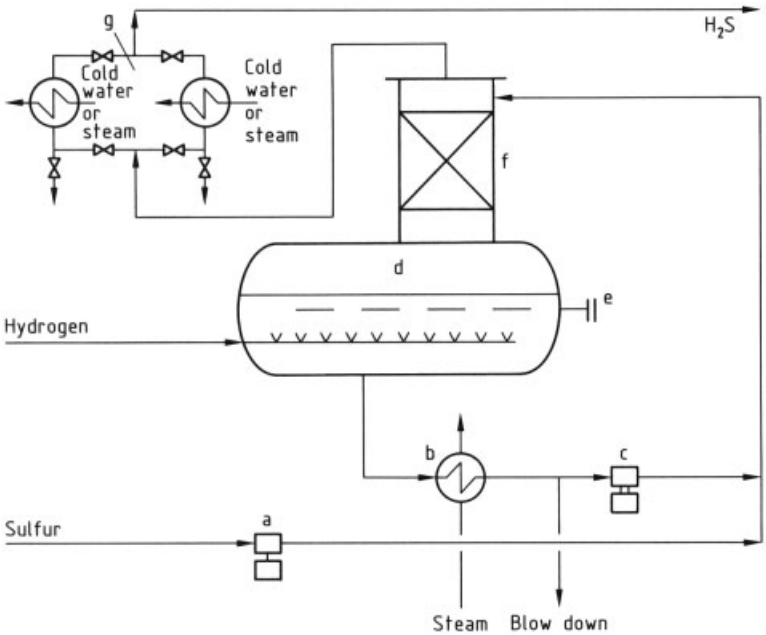

Figure 1. Industrial production of hydrogen sulfide from hydrogen and sulfur
a) Sulfur makeup pump; b) Sulfur cooler; c) Sulfur recycle pump; d) Reactor; e) Electric heater; f) Quench tower; g) Gas cooler

sulfide in the laboratory is to react sodium sulfide with dilute hydrochloric [120, 121] or sulfuric [122] acid. Other sulfides (CaS, ZnS, FeS, Sb_2S_3) can also be used. The purest hydrogen sulfide is obtained from Sb_2S_3.

4.2. Recovery from Gas (→ Gas Production, Chap. 5.4.2.; → Gas Production, Chap. 5.4.3.; → Natural Gas)

Natural gas or gas associated with crude oil is sour, i.e., it contains varying quantities of hydrogen sulfide, from traces to 70 – 80 vol %. This hydrogen sulfide is nearly always associated with varying quantities of carbon dioxide. Refinery gases produced in hydrodesulfurization or cracking units have a very high ratio of hydrogen sulfide to carbon dioxide. When the hydrogen sulfide level of the gas reaches a certain limit, partial or total sweetening (i.e., removal of hydrogen sulfide) is necessary to comply with transportation, distribution, and antipollution requirements. Sweetening gives an acid gas mixture of hydrogen sulfide and carbon dioxide that is generally used for sulfur recovery by the Claus process (→ Sulfur) but also serves as a source of hydrogen sulfide. Hydrogen sulfide is obtained from the gas

1) directly if the hydrogen sulfide – carbon dioxide ratio is high enough to give an acid gas with the desired hydrogen sulfide content (80 – 90 vol %) or if a selective sweetening process is used;
2) after additional treatment, consisting of selective reprocessing (desired purity, 90 – 95 %) or compression and subsequent distillation (minimum purity 99.5 %).

Absorption. Hydrogen sulfide can be recovered from gas containing sufficiently high quantities of it by (selectively) sweetening the gas with respect to carbon dioxide by chemical or physical absorption [123–126].

Absorption by a chemical solvent (amines or potassium carbonate) is a function of pressure and temperature. Absorption by a physical solvent is primarily a function of pressure.

Industrial absorption processes for gas sweetening are summarized in Table 5.

Table 5. Industrial absorption processes for hydrogen sulfide recovery from gases

Solvent	Trade name	Lincensing company
Chemical absorption with aqueous solutions		
N-Methyldiethanolamine (5 N solution)	Ucarsol-MDEA	Union Carbide
Diethanolamine (4 N solution)	SNPA-DEA	Elf Aquitaine
Diisopropanolamine (2 N solution)	ADIP	Shell
N-Methyldiethanolamine (4 N solution)	MDEA	Elf Aquitaine
Diglycolamine (6 N solution)	Econamine	Fluor
Hindered amine solution	Flexsorb	Exxon
Chemical absorption with alkaline salt solutions		
K_2CO_3, catalyst (amine borates)	Catacarb	Eickmeyer
K_2CO_3, catalyst (diethanolamine)	Benfield	Union Carbide
K_2CO_3, catalyst (arsenic salts)	Giammarco-Vetrocoke	Giammarco
Methylaminopropionate	Alkacid-M	BASF
Dimethylaminoacetate	Alkacid-DIK	BASF
Physical absorption		
Anhydrous propylene carbonate	Fluor	Fluor
Dimethyl ether of poly(ethylene glycol)	Selexol	Norton
Cold methanol	Rectisol	Lurgi
N-Methylpyrrolidone	Purisol	Lurgi
Aqueous solution of diisopropanolamine (D) or methyldiethanol-amine (M), sulfolane	Sulfinol D (or M)	Shell
Mixture of monoethanolamine or diethanolamine and methanol	Amisol	Lurgi
Methyl isopropyl ethers of poly(ethylene glycol)	Sepasolv	BASF

Figures 2 and 3 show typical process schemes for absorption with a chemical and a physical solvent, respectively. The sour gas is washed in a countercurrent absorber. The rich solvent is regenerated by flashing and stripping. Flashing (physical solvent) and reboiling (chemical solvent) are the main regeneration factors.

The production process used depends on the characteristics of the gas to be treated and the desired hydrogen sulfide specifications.

If maximum selectivity with regard to hydrogen sulfide is required, a process based on absorption with a selective amine is chosen (a tertiary alkanolamine such as *N*-methyldiethanolamine, MDEA) [127].

Hydrogen sulfide reacts readily and directly with tertiary amines. Carbon dioxide, however, can only react with tertiary amines indirectly in the form of carbonic acid to produce amine carbonates. Carbonic acid first has to be formed by

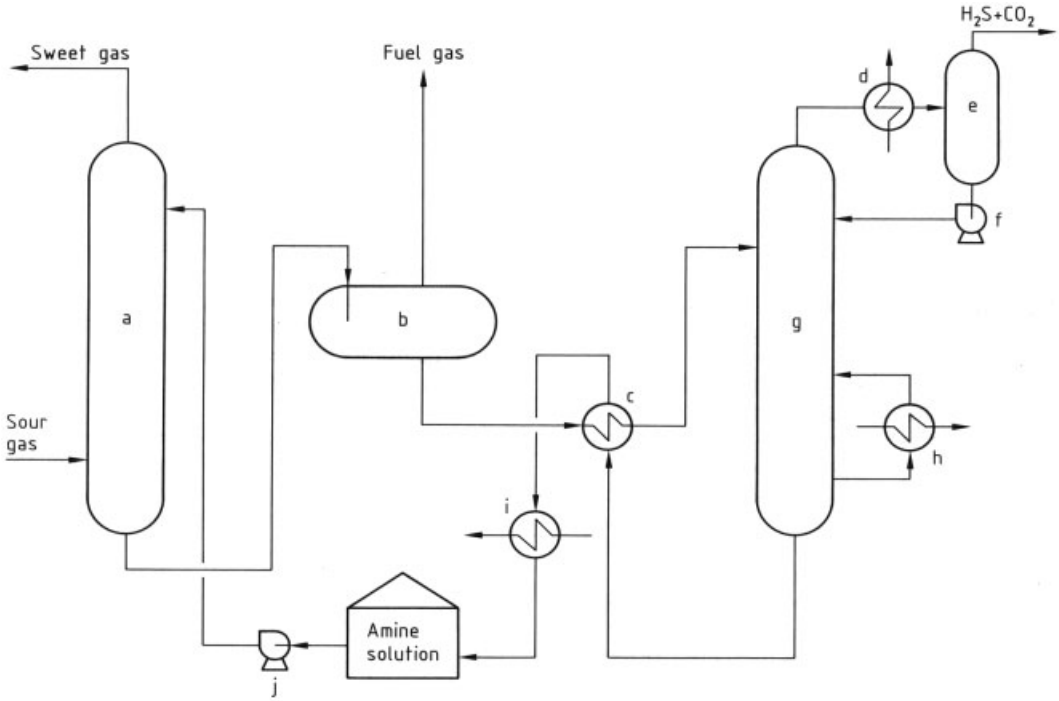

Figure 2. Recovery of hydrogen sulfide with a chemical solvent
a) Absorber; b) Flash drum; c) Amine – amine exchanger; d) Condenser; e) Reflux drum; f) Reflux pump; g) Regenerator; h) Reboiler; i) Cooler; j) Lean amine high-pressure pump

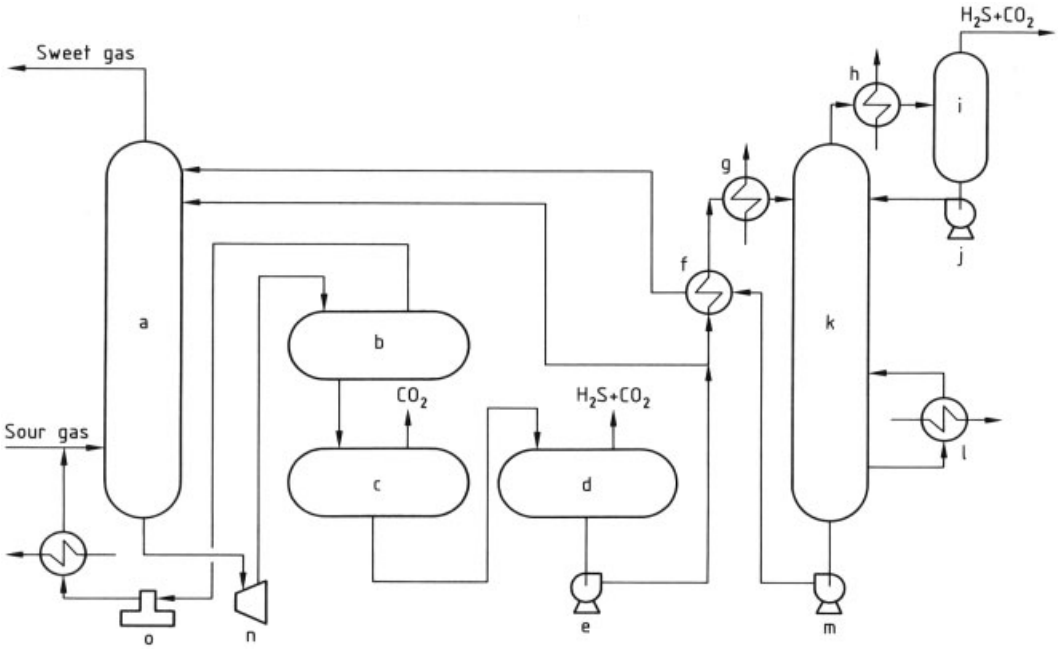

Figure 3. Recovery of hydrogen sulfide with a physical solvent
a) Absorber; b) – d) First, second, and third flash drums; e) Semilean amine high-pressure pump; f) Amine – amine exchanger; g) Reheater; h) Condenser; i) Reflux drum; j) Reflux pump; k) Hydrogen sulfide stripper; l) Reboiler; m) Lean amine high-pressure pump; n) Power recovery; o) Recycle compressor

reaction of carbon dioxide with water and this step is much slower than direct reaction. An important principle of selective absorption is therefore the selection of a gas – liquid contact time that is long enough to eliminate hydrogen sulfide, but short enough to minimize the coabsorption of carbon dioxide.

If the partial pressure of hydrogen sulfide in the sour gas is sufficiently high (≥ 0.3 MPa) and significant quantities of hydrocarbons (a few percent) can be tolerated, a physical solvent process may also be chosen.

Distillation. The industrial distillation of an acid gas in which the purity of hydrogen sulfide is insufficient was developed by Elf Aquitaine at its Lacq plant [128]. Hydrogen sulfide is obtained either as a liquid or as a gas with a purity > 99.5 %. The acid gas is first treated with activated carbon to eliminate heavy hydrocarbons and dried with a solid adsorbent (alumina, silica beads, or molecular sieves). It is then compressed to 2.5 – 3.5 MPa and distilled in a tray column or packed column. Liquid hydrogen sulfide is obtained at the bottom of the column and easily vaporized for more convenient distribution.

If the acid gas has a very low hydrogen sulfide content ($H_2S/CO_2 < 0.2$), separation of hydrogen sulfide and carbon dioxide can be improved by introducing a C_3– C_6 alkane above the column feed [129]. The hydrogen sulfide obtained is slightly less pure.

5. Environmental Protection

Hydrogen sulfide is not only produced industrially, it also originates from natural sources (see also Chap. 1):

1) diffusion into underground water from oil and natural gas wells,
2) formation in lakes by aerobic bacterial fermentation,
3) production during biological wastewater treatment, and
4) formation as a degradation product in the agricultural food industry, slaughterhouses, and the paper industry.

In the environment, the toxicity of hydrogen sulfide is due to the hydrogen sulfide molecule itself rather than to hydrosulfide or sulfur ions. For example, bacterial metabolism is inhibited when the hydrogen sulfide content of the surroundings reaches 70 – 200 mg/L [130]. Because hydrogen sulfide does not accumulate in living organisms, it cannot contaminate the food chain [131]. Natural elimination of hydrogen sulfide occurs mainly by diffusion into the atmosphere, but it is also subjected to biological or chemical degradation to form sulfur oxides and sulfur. In water, hydrogen sulfide is converted into elemental sulfur or insoluble heavy-metal sulfides.

Accidental leaks of hydrogen sulfide gas can be absorbed directly with aqueous solutions of sodium carbonate, sodium hydroxide, sodium hypochlorite, or iron(III) chloride. When liquid hydrogen sulfide is spilled on the ground, adsorbents such as activated carbon should be used to prevent the spread of liquid and reduce the emission of vapor [132]. Water polluted with hydrogen sulfide can be stripped and neutralized with lime, sodium carbonate, or sodium bicarbonate. Sulfur ions are then eliminated by precipitation.

Industrial effluents containing hydrogen sulfide must be treated or incinerated if their hydrogen sulfide content exceeds 0.001 vol %. The final procedure depends on the sulfur content and local discharge regulations. For example, for residual gases in sulfur plants, treatment can consist of

1) hydrogenation of residual sulfur compounds, followed by washing with amines that selectively absorb hydrogen sulfide and recycling of the effluent from the amine regenerator to the inlet of the sulfur plant: Scot (Shell), Elfapur (Elf Aquitaine) processes;
2) incorporation of an additional treatment stage based on the catalytic Claus reaction at low temperature (135 –150 °C). The sulfur obtained remains trapped in the catalyst and must be desorbed during a subsequent regeneration cycle: Sulfreen process (Elf Aquitaine, Lurgi).

When the sulfur concentration is not too high, washing the residual gas with chemically regenerable substances (as used in the LO-CAT process) or with nonregenerable substances (e.g., sodium hypochlorite or sodium hydroxide) may be more economical. Alternatively, the gases may be treated by conventional adsorption (activated carbon, metal oxides) or by adsorption

combined with biological regeneration [133]. The latter involves immobilizing bacteria on an adsorbent to incite autopurification.

6. Analysis

There is no universal method of analysis for the determination of hydrogen sulfide. Each method depends not only on the medium in which the quantity of hydrogen sulfide is to be determined (gaseous mixtures, liquid effluents) and on the concentration of hydrogen sulfide in the medium, but also on the nature of the desired results.

The following examples are by no means exhaustive:

1) Determination in a sour gas by iodometry [134], gas-liquid chromatography [135], and argentimetry [136].
2) Determination in a commercial purified gas by iodometry [137], polarography [138].
3) Determination in a fuel gas with lead acetate [139].
4) Determination in the atmosphere [140].

7. Storage, Handling, and Transportation

7.1. Storage and Handling

The most important problem in storing hydrogen sulfide is the material used for the storage vessel.

Anhydrous hydrogen sulfide does not react with steel at ambient temperature. However, in an aqueous medium, it attacks iron to form iron sulfide and hydrogen. The hydrogen can diffuse into the metal where it is trapped in cavities, causing blistering and embrittlement. The degree of corrosion depends on the grade of steel and is very limited with stainless steel. Ferritic steels are much more sensitive to hydrogen embrittlement than austenitic steels. Embrittlement can be important with highly resistant steel. High-strength steel exposed to hydrogen sulfide is likely to crack under stress. Stress-relieving treatment after welding reduces residual stresses. Thus, the steel used for hydrogen sulfide storage vessels and pipes must be chosen carefully.

Because of the problems associated with transportation of hydrogen sulfide and its toxicity, storage capacities are restricted; the product must be used close to its production site whenever possible. Nevertheless, hydrogen sulfide is transported by pipeline over several kilometers in Europe and the United States. Liquid hydrogen sulfide is also transported in specially designed containers. Particular attention should be given to emptying devices (valves) and their protection during transportation.

Storage tanks for liquid hydrogen sulfide should be equipped with a safety device system as shown in Figure 4. The blowdown tank is a reserve storage capacity that enables rapid emptying of liquid hydrogen sulfide in case of accident. It can be operated by an "emergency stop"

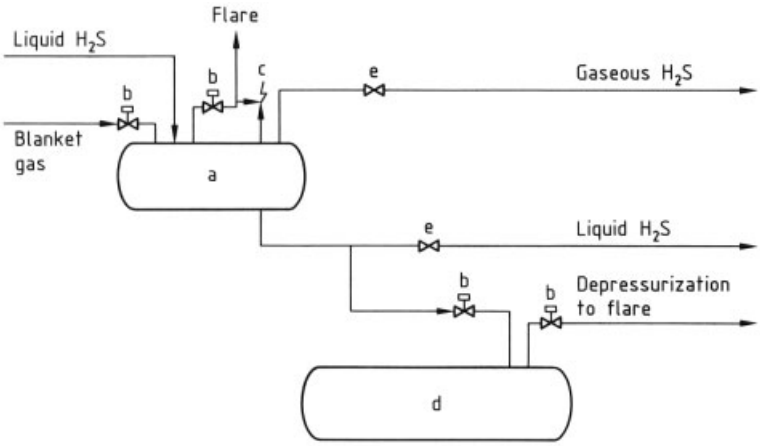

Figure 4. Storage tank for liquid hydrogen sulfide equipped with safety devices
a) Storage tank; b) Safety valve; c) Relief valve; d) Blowdown tank; e) Block valve

switch that automatically actuates the safety valves. The gas phase in the storage tank is connected directly to a flare by means of a relief valve that can be bypassed via a safety valve.

7.2. Transportation

Transportation of liquid hydrogen sulfide by air (IATA) is strictly prohibited. Transportation in bulk containers by sea is also prohibited (IMDG). Small containers of < 1 t or small bottles are generally used for road and rail transportation. The transport of hydrogen sulfide in Europe by rail has been abandoned; it is used instead at the site of production. In the United States, liquid hydrogen sulfide is transported in specially designed tankers. Hazard classifications for hydrogen sulfide are as follows:

IMDG Code, class 2.1, label 2.1 + 6
RID/ADR, class 2, label 3 + 6.1
United States, CFR 49 : 172.02, flammable
UN no. 1053

The sulfur content of crude oil can result in the production of hydrogen sulfide, representing a risk for field personnel and for the transporter. In sea transport, maintenance of storage tanks under a blanket of inert gas does not eliminate the risk of explosion. The risk of sulfur formation in the tankers should also be kept in mind. Wearing gas masks in the pump room of tankers is compulsory when the hydrogen sulfide content exceeds 10 ppm. The same measures should be taken for personnel connecting up the loading arms.

8. Uses

Production of Sulfur. Natural gas containing hydrogen sulfide is currently one of the most important sources of elemental sulfur. For commercial use, sour natural gas must first be sweetened to remove the acid compounds, mainly hydrogen sulfide and carbon dioxide. The resulting acid gas is used for the production of sulfur by the Claus process [141, 142]:

$$2\,H_2S + SO_2 \longrightarrow 3/n\,S_n + 2\,H_2O$$

For further details, see → Sulfur. Sulfur is marketed in liquid or solid form. In the liquid state, itmust first be degassed to remove all traces of hydrogen sulfide. The sales specification of sulfur in Europe requires a hydrogen sulfide content of < 10 ppm. This value is obtained by degassing liquid sulfur into sulfur pits with ammonia before dispatch.

Production of Metal Sulfides. The reaction of hydrogen sulfide with soda or lime solutions results in formation of the corresponding sulfides. Sodium sulfide is generally marketed in a solid hydrated form. Acid sulfides or hydrosulfides [e.g., NaSH or $Ca(SH)_2$] are marketed either as crystals or as concentrated aqueous solutions.

Sodium sulfide is used as a lignin demethylation agent in the Kraft process for manufacturing paper pulp. Mixtures of sodium and calcium sulfides are powerful reducing agents of disulfide bonds. They attack the cystine molecules of keratin and are used in the leather industry for chemical depilation of animal hides (→ Leather).

Flotation. Preliminary treatment with hydrogen sulfide is sometimes required for the purification of ore by selective flotation. Certain mineral species in the ore are sulfurized, which considerably increases the selectivity of flotation collectors and facilitates separation of valuable metallic species. This process is used in the purification of nickel and manganese.

Catalyst Activation and Poisoning. Certain catalysts reach their maximum activity only when their constituent metals are in the form of sulfides (e.g., the hydrotreating catalysts used in oil refining). Preliminary treatment with hydrogen sulfide is sometimes necessary.

However, hydrogen sulfide also poisons some hydrogenation and reforming catalysts with a platinum or palladium base. This can be used to advantage to reduce initial catalyst activity and limit coking (i.e., formation of carbonaceous deposits). Gasoline reforming catalysts, for example, are always selectively poisoned with a small quantity of hydrogen sulfide prior to use.

Treatment of Metallic Surfaces. Superficial sulfurization of metals with hydrogen sulfide allows modification of their physical or chemical properties. Hydrogen sulfide is used

for passivation of the walls of reactors operating at high temperature in petrochemical operations such as steam cracking and hydrodealkylation. This treatment prevents desirable secondary reactions.

Another use of hydrogen sulfide is the formation of a layer of sulfide on the surface of steel wires or plates that are to be coated with paint or plastic. This treatment protects the coating or improves the action of adhesive primers.

Analytical Chemistry. Hydrogen sulfide has long been used to selectively precipitate group 1, 2, and 3 metal cations. The metals are qualitatively identified by the color of the sulfides formed.

Production of Thioorganic Compounds. The most common industrial application of hydrogen sulfide is the production of organic compounds with a thiol function (i.e., RSH, where R is a hydrocarbon group) or a thiol group and another function (e.g., alcohol, acid).

Table 6 shows the most important thioorganic compounds (in terms of tonnage) produced from hydrogen sulfide in the United States and Europe. Other important thioorganic compounds derived from hydrogen sulfide are tetrahydrothiophene and mixtures of thiols or sulfides (often containing *tert*-butanethiol or diethylsulfide). These are used as strong-smelling additives for gas odorization.

Long-chain thiols also represent a range of useful products, especially *n*-dodecanethiol and *n*-octanethiol. These are used as intermediates in the synthesis of plant protection agents and as polymerization agents.

9. Economic Aspects

Total world sulfur production was 56.6×10^6 t in 1986, compared with 51.7×10^6 t in 1977. Sulfur recovered from the treatment of sour gas (i.e., from hydrogen sulfide) accounts for 25 % of this (ca. 14×10^6 t in 1986, 11.8×10^6 t in 1977).

The primary producing countries are listed below, with their 1986 production figures in 10^6 metric tonnes per year:

Western Canada	5.2
United States	2.2
former Soviet Union	1.8
Saudi Arabia	1.2
Federal Republic of Germany	1.0
France	0.9

Recovery of sulfur from hydrogen sulfide will continue to expand worldwide. The former Soviet Union is expected to become the main producer of sulfur from sour gas due to development of the Astrakhan gas field (hydrogen sulfide content: 26 %). Astrakhan I started producing early in 1987 and Astrakhan II is due to be put on stream early in 1989.

Canada should continue to be a major producer because of the development of the Caroline gas field in Alberta. The Middle East is also expected to play a major role, with substantial tonnage from Iran and Qatar. Production in

Table 6. Thioorganic compounds produced from hydrogen sulfide

Product	Formula	Raw materials	Annual world production, t	Producers	Uses
Methanethiol (methyl mercaptan) [*74-93-1*]	CH_3SH	methanol + H_2S	70 000	Pennwalt Rhone Poulenc Elf Aquitaine	base product for manufacture of methionine, dimethyl disulfide, and plant protection products
Thioglycolic acid (mercapto acetic acid) [*68-11-1*]	$HSCH_2COOH$	chloroacetic acid + NaSH	15 000	Elf Aquitaine Bock Merck	manufacture of esters, poly(vinyl chloride) stabilizers, and cosmetics
tert-Dodecanethiol [*25103-58-6*]	$C_{12}H_{25}SH$ (mercaptan mixture)	tetrapropylene or triisobutylene + H_2S	13 000	Pennwalt Phillips Elf Aquitaine Bayer	chain-transfer agent in radical polymerization; manufacture of polysulfides
Ethanethiol (ethyl mercaptan) [*75-08-1*]	CH_3CH_2SH	ethylene + H_2S	12 000	Phillips Pennwalt Elf Aquitaine	manufacture of plant protection products
Mercaptoethanol [*60-24-2*]	$HSCH_2CH_2OH$	ethylene oxide + H_2S	8000	Phillips Morton Thiokol Alcolac BASF Elf Aquitaine	soluble chain-transfer agent for polymerization in aqueous media; manufacture of plant protection products; poly(vinyl chloride) stabilizers

the United States and the Federal Republic of Germany is likely to stabilize at $(2.3-2.5)\times10^6$ and 1.0×10^6 t, respectively. In France, production will decline gradually due to depletion of the Lacq gas field (hydrogen sulfide content: 15 %). Sulfur recovered from hydrogen sulfide has been the major source of accumulated stocks, which are used to swing production to balance the market when needed. At the end of 1987, the stock of sulfur from sour gas, in tonnes, was

Western Canada	6.7×10^6
France	2.2×10^6
Saudi Arabia	1.5×10^6

10. Toxicology

Hydrogen sulfide is a toxic gas. Human poisoning is related to its acute toxicity. Chronic toxicity has not been demonstrated in people exposed to concentrations lower than those that induce acute or subacute poisoning [143, 144].

Animal Data. Much toxicological data has been published on animals exposed to massive dosages of hydrogen sulfide. These data confirm observations made on humans, but information on chronic or long-term effects remains limited. The acute toxicity of hydrogen sulfide is comparable in animals and humans.

Chronic toxicity studies have been carried out on rabbits, rats, and mice that inhaled daily doses of hydrogen sulfide at concentrations in air up to 80 ppm over three to five consecutive months [143–145]. Despite numerous detailed examinations, including neurological tests, no specific chronic effects were found. Mutagenicity testing has shown a lack of genotoxic effects on bacteria in vitro [146] and on rats in vivo [147].

Metabolism. Lack of a chronic toxic effect of hydrogen sulfide in animals and most probably in humans may be explained by the fact that it does not build up in the body. Hydrogen sulfide is metabolized rapidly, mainly into less toxic, oxidized metabolites that are quickly eliminated via the kidneys.

Hydrogen sulfide is readily absorbed by the pulmonary tract, which at the same time represents an excretory pathway for untransformed hydrogen sulfide. Hydrogen sulfide that is not eliminated by the respiratory tract binds to circulating proteins in the blood and is distributed rapidly to all tissues, including the nervous system. It is subsequently metabolized and detoxified as shown in Figure 5 [144].

Hydrogen sulfide is detoxified by oxidation to sulfate (the primary reaction) or by methylation (not yet confirmed). The toxicity of hydrogen sulfide is due to its reaction with protein constituents of essential enzymes or metalloproteins, especially the iron-containing metalloproteins of the respiratory cell chain.

During chronic exposure, the amount of hydrogen sulfide entering the organism is so small that it is neutralized by the above detoxification reactions. However, during massive acute exposure, the detoxification capacity is exceeded; the reaction of hydrogen sulfide with proteins can thus lead to poisoning.

Human Risk and Pathology. The acute toxic risk of hydrogen sulfide and the pathology observed in human poisoning are summarized in Tables 7 and 8 (see next pages) [148, 149].

11. Occupational Health and Safety

The main safety precautions listed here result from thirty years' experience at Lacq [150] and from the literature. Hydrogen sulfide has several properties that make it a highly dangerous gas:

1) high toxicity;
2) flammability: it ignites spontaneously at 260 °C (compared to 537 °C for methane) and forms an explosive mixture with air when the hydrogen sulfide content is 4 – 44 vol % (compared to 5 – 15 % for methane) [151];
3) density: being heavier than air, it accumulates in low-lying areas of installations;
4) ability to embrittle steel (Section 7.1); and
5) insidiousness: persons intoxicated with hydrogen sulfide no longer discern its characteristic smell of rotten eggs before the dangerous threshold concentration (50 – 100 ppm) is reached.

General Precautions. Areas in which hydrogen sulfide is produced, utilized, and stored must be isolated and delimited.

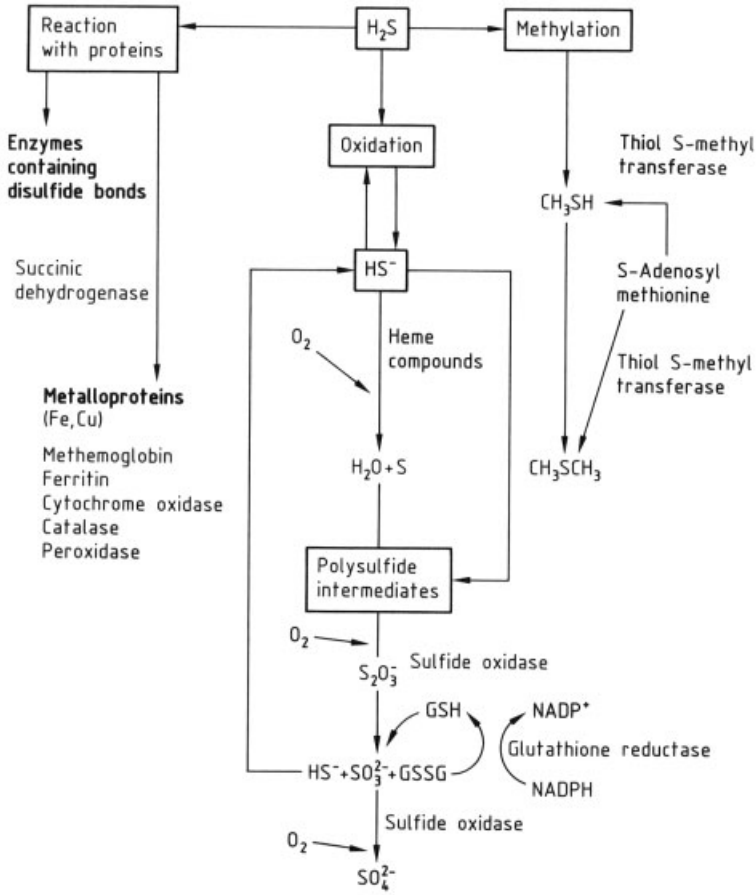

Figure 5. Metabolism of hydrogen sulfide in humans and animals
Abbreviations: GSH, reduced glutathione; GSSG, glutathione; $NADP^+$, oxidized nicotinamide – adenine dinucleotide phosphate; NADPH, reduced nicotinamide – adenine dinucleotide phosphate

Table 7. Acute toxic effects of hydrogen sulfide in humans

Concentration	Duration of exposure				
of H_2S, ppm	15 min	15 min – 1 h	1 – 4 h	4 – 8 h	Comments
10				eye irritation	maximum tolerable concentration for prolonged exposure
50 – 100	loss of olfactory perception	eye irritation	eye and bronchial irritation	danger in case of continuous exposure	working conditions necessitate protective measures
150 – 250	loss of olfactory perception	eye and bronchial irritation	serious respiratory distress and asthenia		working conditions necessitate protective measures
300 – 400	loss of olfactory perception, eye and bronchial irritation, asthenia	severe respiratory distress, acute asthenia	pulmonary edema and risk of death		risk of death if no appropriate measures taken
500 – 1000	loss of consciousness, respiratory distress	risk of pulmonary edema and death			risk of death if no appropriate measures taken
>1000	immediate loss of consciousness and respiratory distress				risk of death if no appropriate measures taken

Table 8. Human pathology in hydrogen sulfide poisoning

	Degree of poisoning		
	Weak	Intermediate	Severe
Early signs	equilibrium problems	warning signs of short duration	immediate coma
Immediate effects	loss of consciousness for a few seconds; rapid regain of consciousness	sudden loss of consciousness lasting a few minutes, followed by headache, respiratory distress, gastric pain, diarrhea (following day)	prolonged deep coma with cyanosis, respiratory distress, occasionally cardiac distress; regain of consciousness is difficult and agitated
Long-term effects	recovery	recovery or bronchial fragility; asthenia	bronchopulmonary sequelae
Emergency treatment	evacuation from danger area; administration of oxygen	evacuation from danger area; administration of oxygen; respiratory assistance if needed; intravenous vitamin C; medical observation for several hours	evacuation from danger area; administration of oxygen; respiratory assistance if needed; intravenous vitamin C; cardiotonics and neurosedatives if needed; hospitalization

Only authorized personnel should have access to these areas, i.e., personnel having undergone medical tests to detect ocular, respiratory, or cardiovascular disease, and signs of alcoholism. All authorized personnel should undergo special training to deal with the inherent risks of working in a hydrogen sulfide environment. They should be equipped with gas masks and undergo regular training in their use.

The usual precautions should be taken in all areas where an explosive atmosphere is likely to be found (sources of flame, electric equipment, ect.). "No smoking" signs should be displayed and strictly observed.

Only piping, apparatus, and containers made from special steel or alloys should be used; the grade should be suitable for the conditions of use. The use of copper and its alloys should be prohibited.

Storage and Transportation (see also Chap. 7). The storage area should be well ventilated and, if possible, equipped with a water cooling system in case of fire.

Hydrogen sulfide containers should be stored separately from incompatible products (oxidizing agents, inflammable products, powerful acids, etc.) and away from direct sunlight. A special parking area should be provided for vehicles transporting hydrogen sulfide.

Each container should carry a label indicating that the contents are highly flammable and toxic. The label should also list the special risks and include cautionary advice.

- Modified EEC directive 67/548
- EEC No. 016-001-00-4
- *Danger symbols:* highly flammable and toxic
- *Nature of special risks attaching to dangerous substances:*
- R.13 Extremely flammable liquefied gas
- R.26 Very toxic by inhalation
- *Safety advice concercing dangerous chemical substances:*
- S7/9 Keep container tightly closed in a well-ventilated place
- S.25 Avoid contact with eyes
- S.45 In case of accident or clinical symptoms medical advice should be sought immediately.

Installation and Handling. Apparatus should be properly closed and tightly sealed. Waste gas from discharge and pressure valves should be sent to a flare to avoid emission of hydrogen sulfide into the atmosphere. This flare can burn off hydrogen sulfide in dangerous situations, eliminate air from the discharge piping before discharge, and purge hydrogen sulfide after discharge.

Emissions produced under normal conditions should be recovered at source, and low-lying areas should be ventilated thoroughly.

Personnel working on apparatus containing (or having contained) hydrogen sulfide should be equipped with insulated gas masks and wear protective clothing, goggles, and long gloves. No one should be allowed to work alone in a zone where hydrogen sulfide is handled.

Frequent checks of the atmosphere should be made. Hydrogen sulfide detection devices should be installed (see below); smell should not be relied on for detection.

Programmed preventive maintenance should be carried out in all hydrogen sulfide installations, particularly on sealing devices. A fire per-

mit should be obtained to use tools producing flame, heat, or sparks.

Pyrophoric sulfides should be kept constantly humid. They should be extracted and incinerated as soon as possible. After maintenance work, all traces of humidity should be eliminated.

Emergency Measures. A plan should be drawn up for evacuation of personnel if the concentration of hydrogen sulfide exceeds the norm. Gas masks must be put on as soon as an alarm is given. Appropriate measures to be taken in the case of leakage, fire, or accident should be displayed prominently in the installation.

In the event of leakage without fire, the leak should be isolated and sprayed with water. Uncontrollable gas should be sent to a flare pit, and all sources of heat should be removed or shut down.

If the leaking material has caught fire, the fire should be put out only when it presents an especially serious threat to the environment or when the leak can be rapidly controlled.

In the event of accidental gassing, lifesaving intervention should be undertaken only after the rescuer has donned an insulating gas mask. The victim should be removed rapidly from the danger zone and should receive necessary first aid and medical treatment as soon as possible.

Detection of Hydrogen Sulfide in Air. Permanent detection devices are necessary both in and around the hydrogen sulfide storage area. They should set off both audible and visual alarms as soon as the concentration of hydrogen sulfide in air exceeds 10 ppm. Several models are available. An *electrochemical sensor* functions on the gas diffusion principle: hydrogen sulfide in air reacts chemically with an electrolyte (Sulfytron 4010, Draeger, FRG). In a *semiconductor* detector, air containing hydrogen sulfide diffuses into the specially adapted semiconductor and modifies its electrical resistance (General Monitors, United States).

Portable detectors can be used for random measurement of hydrogen sulfide concentration. An electrically or manually operated pump sucks air through a reactive tube divided into zones that become colored on contact with the polluting gas (Draeger model 21/31 with specific tube for hydrogen sulfide).

A simple but effective method of detecting a leak is to use lead acetate paper. The paper becomes black in the presence of hydrogen sulfide but gives no indication of concentration.

Exposure Limit Values [152]. The average permitted concentration of hydrogen sulfide in air varies from country to country:

TLV (TWA) 5 – 10 ppm (vol)
TLV (STEL) 10 – 15 ppm

12. References

1. E. Eriksson, *Tellus* **12** (1960) 63. H. Roffmann, M. C. Bender: "H_2S: new sources and potential impacts," *68th Annual Meeting of the Air Pollution Control Association* , June 1975.
2. B. Bonsang, *La Recherche* **13** (1982) no. 137, 1132 – 1142.
3. D. Möller, *Atmos. Environ.* **18** (1984) no. 1, 29 – 39.
4. T. P. Kudryautseva, Deposited document – VINITI 160, 7 – 15; *Chem. Abstr.* **94** (1981) no. 22, 177 614 v.
5. A. I. Kravtsov, G. I. Voitov, *Izv. Vyssh. Uchebn. Zaved., Geol. Razved.* **19** (1976) no. 4, 18 – 25; *Chem. Abstr.* **85** (1976) no. 24, 179 877 w.
6. H. G. Spratt, Jr., M. D. Morgan, R. E. Good, *Appl. Environ. Microbiol.* **53** (1987) no. 7, 1406 – 1411; *Chem. Abstr.* **107** (1987) no. 11, 95 855 x.
7. J. P. Hollis, Jr., *Acta Phytopathol. Acad. Sci. Hung.* **20** (1985) no. 3 – 4, 321 – 326; *Chem. Abstr.* **105** (1985) no. 21, 190 028 x.
8. I. Benda, *Arch. Mikrobiol.* **27** (1957) 337 – 374; *Chem. Abstr.* **53** (1959) 22 208 g.
9. R. Vàmos, M. Ando, *Acta Univ. Acta Biol., Szeged.* **5** (1959) 61 – 69; *Chem. Abstr.* **54** (1960) 17 770 f.
10. V. G. Popov, *Gazov. Prom-st Ser.: Geol. Razved. Gazov. Gazokondens. Mestorozhd.* 1981, no. 10, 10 – 18; *Chem. Abstr.* **96** (1982) no. 20, 165 091 a.
11. E. E. London, *Geol. Nefti Gaza* 1975, no. 8, 24 – 30; *Chem. Abstr.* **84** (1976) no. 6, 33 375 p.
12. C. A. Brook, *Geol. Surv. Prof. Pap. U.S.* 1141 (1981) 193 – 203.
13. R. J. Ferek, R. B. Chatfield, M. O. Andrae, *Nature (London)* **320** (1986) 514 – 516.
14. E. R. Lippincott, R. V. Eck, M. O. Dayhoff, C. Sagan, *Astrophys. J.* **147** (1967) 753 – 764; *Chem. Abstr.* **67** (1967) 58 795 n.

15. K. Pasternak, *Acta Hydrobiol.* **12** (1970) no. 4, 377 – 390; *Chem. Abstr.* **75** (1971) no. 6, 40 097 s.
16. N. F. Voznaya, *Lb. Tr. Leningr. Inzh. Stroit. Inst.* **25** (1957) 123 – 133; *Chem. Abstr.* **53** (1959) 22 637 i.
17. W. J. Mueller, *GWF Gas Wasserfach* **102** (1961) 986 – 989; *Chem. Abstr.* **56** (1962) 1303 e.
18. F. W. Brinkmann, *SPE/DDE Low Permeability Symposium,* Denver, CO, May 27 – 29, 1981.
19. H. Luebben, *Erdoel Erdgas Z.* **98** (1982) July, 225 – 231.
20. K. Le Tran, J. Connan, B. Van der Weide, *Adv. Org. Geochem. Proc. Int. Meet., 6th* 1973, 761 – 789; *Chem. Abstr.* **85** (1976) no. 14, 96 664 y.
21. J. L. Rumeau, C. Sourisse, *Adv. Org. Geochem. Proc. Int. Meet. 7th* 1975, 525 – 538; *Chem. Abstr.* **89** (1978) no. 20, 165 749 b.
22. L. I. Rovnin, N. V. Mizinov, N. I. Voronin, *Geol. Nefti Gaza* 1977, no. 10, 41 – 43; *Chem. Abstr.* **88** (1978) no. 2, 9336 p.
23. N. V. Mezinov, A. S. Zinger, V. G. Grushevoi, N. I. Voronin, *Geol. Nefti Gaza* 1979, no. 6, 1 – 5; *Chem. Abstr.* **91** (1979) no. 18, 142 832 a.
24. R. G. Pankina, V. L. Mekhtieva, S. P. Maksimov, *Geol. Nefti Gaza* 1983, no. 4, 45 – 51; *Chem. Abstr.* **99** (1983) no. 4, 25 095 s.
25. R. S. Aliev, *Geol. Stroenie i Neftegezonosn. Nizhn. Povolzh'ya Saratov* 1984, 86 – 93; *Chem. Abstr.* **102** (1985) no. 12, 97 906 k.
26. I. A. Lagunova, A. Ya. Vvedenskaya, *Geol. Nefti Gaza* 1986, no. 12, 20 – 23; *Chem. Abstr.* **106** (1987) no. 10, 69 776 g.
27. S. Kh. Aigistova et al., *Gazov. Prom-st* 1983, no. 8, 30 – 31; *Chem. Abstr.* **99** (1983) no. 20, 160 948 n.
28. N. B. Valitov, *Sov. Geol.* 1986, no. 3, 93 – 100; *Chem. Abstr.* **104** (1986) no. 24, 209 594 g.
29. N. B. Valitov, *Geol Nefti Gaza* 1986, no. 5, 23 – 28; *Chem. Abstr.* **105** (1986) no. 6, 45 921 d.
30. N. I. Voronin, E. I. Ben'ko, V. F. Laktynshina, L. A. Anisimov, *Geol. Nefti Gaza* 1986, no. 7, 58 – 60, *Chem. Abstr.* **105** (1986) no. 18, 155 735 d.
31. V. F. Perepelichenko, S. V. Eremeeva, A. F. Il'in, *Gazov. Prom-st* 1986, no. 12, 30 – 31; *Chem. Abstr.* **106** (1987) no. 10, 69 839 e.
32. V. I. Lapshin, Yu. Yu. Kruglov, A. P. Zheltov, *Geol. Nefti Gaza* 1987, no. 2, 41 – 43; *Chem. Abstr.* **107** (1987) no. 4, 25 703 x.
33. V. I. Danilov, V. P. Pronchnk, A. Ya. Litovchenko, *Tr. Gos. Inst. Proekt. Issled. Rab. Neft. Prom-sti* 1973, no. 19, 154 – 157; *Chem. Abstr.* **83** (1975) no. 26, 208 160 y.
34. Kh. I. Ismailova, *Razrab. Gazov. Mestorozhd. Transp. Gaza* 1974, no. 3, 40 – 46; *Chem. Abstr.* **84** (1976) no. 10, 62 087 b.
35. Yu. A. Zhurov, *Gazov. Prom-st* 1976, no. 9, 36 – 39; *Chem. Abstr.* **86** (1977) no. 20, 142 511 s.
36. P. M. Lomako, *Gazov. Prom-st* 1979, no. 11, 33 – 36; *Chem. Abstr.* **92** (1980) no. 18, 149 624 p.
37. V. N. Kortsenshtein, *Geol. Nefti Gaza* 1980, no. 4, 56 – 59; *Chem. Abstr.* **93** (1980) no. 6, 49 924 p.
38. A. I. Polivanova, G. S. Akopova, O. F. Tuzlukova, *Geol. Priro. Gaza Polezn. Kompon. Prir. Gazov Usloviya Form. Resur.* (1979) M 87 – 95; *Chem. Abstr.* **93** (1980) no. 18, 170 614 r.
39. V. M. Kir'yashkim, V. S. Goncharov, I. A. Borzenkov, A. K. Aleshina, *Gazov. Prom-st, Ser. Geol. Razved. Gazov. Gazokondens. Mestorozhd.* 1980, no. 6, 28 – 34; *Chem. Abstr.* **93** (1980) no. 16, 152 640 s.
40. G. P. Vdovykin, *Gazov. Prom-st* 1980, no. 10, 29 – 31; *Chem. Abstr.* **94** (1981) no. 14, 105 998 a.
41. A. G. Burmistrov, B. V. Surkov, *Probl. Dobychi Gaza* (*Primera Razrab. Orenb. Gazokondens. Mestorozhd.*) **1979,** 159 – 163;O. F. Khudyakov (ed.), Vses. Nauchno-Issled. Inst. Prir. Gazov. (Moscow) USSR; *Chem. Abstr.* **95** (1981) no. 18, 153 270 z.
42. F. R. Hertweck, Jr., D. D. Fox, *Inf. Circul. – U.S. Bur. Mines,* IC 8993, 1984, p. 130; *Chem. Abstr.* **102** (1985) no. 10, 81 281 s.
43. J. B. Hyne, *World Oil* **194/195** (1982) 75 – 78.
44. A. V. Kormukhina, A. I. Mishukovskaya, N. V. Verts, *Korroz. Zashch. Neftegazov. Prom-sti* 1976, no. 3, 5 – 7; *Chem. Abstr.* **85** (1976) no. 14, 96 679 g.
45. J. G. Connell, R. M. Shapiro, *Proc. Annu. Southwest. Pet. Short Course* **23** (76) 1 – 8; *Chem. Abstr.* **86** (1977) no. 14, 92 779 n.
46. T. L. Pugh, *Elastomerics* **116** (1984) no. 6, 35 – 39; *Chem. Abstr.* **101** (1984) no. 16, 133 459 g.
47. K. A. Shishin et al., *Neft. Khoz.* 1977, no. 12, 51 – 52; *Chem. Abstr.* **88** (1978) no. 16, 107 600 q.
48. R. M. Khasaev, L. V. Konysheva, L. S. Leon'teva, R. M. Abdullaev, SU 825 579; *Chem. Abstr.* **95** (1981) no. 8, 64 940 a.

49. Yu. N. Sokolov, B. V. Yakubenko, T. Ya. Deputatova, *Geol. Str. Neftegazonosn. Nizh. Povolzh'ya, Saratov* 1981, no. 4, 106 – 109; *Chem. Abstr.* **97** (1982) no. 26, 219 164 t.
50. A. P. Krezub, V. V. Kainash, *Burenie (Moskva)* 1981, no. 12, 7 – 9; *Chem. Abstr.* **98** (1983) no. 16, 128 805 g.
51. B. A. Andreson, G. P. Bochkarev, E. I. Ogarkova, R. Kh. Khazipov, SU 1127893 A1, 1984; *Chem. Abstr.* **102** (1985) no. 14, 116 396 n.
52. J. Raczkowski, A. Dudek, A. Hejnar, *Nafta (Katowice Pol.)* **42** (1986) no. 4, 106 – 110; *Chem. Abstr.* **106** (1987) no. 6, 35 660 z.
53. *Oil Gas J.* **82** (1984) 76.
54. Ullman, 4th ed., **21**, 169.
55. R. Goodwin: "Hydrogen sulfide provisional thermophysical properties from 188 to 700 K at pressures to 75 MPa," *NBS Report no. NBSIR 83 – 1694* (1983) 175.
56. "L'Air liquide," *Encyclopédie des gaz,* Elsevier, Amsterdam 1976, p. 935.
57. J. W. Hennel, K. Kikrynicki, *Acta Phys. Pol.* **18** (1959) 523 – 526.
58. *Gmelin,* **4 a/b,** p. 108.
59. P. Pascal, "Nouveau traité de chimie minérale," vol. XIII, Masson Ed., Paris 1956.
60. R. C. Weast (ed.): *Handbook of Chemistry and Physics,* 53rd ed., CRC Press, Cleveland, 1972/73, p. 222.
61. C. Huiszoom, A. Dynanus, *Physica* **31** (1965) 1049 – 1052.
62. A. Barua, A. Manna, P. Mukhopadhyay, *J. Chem. Phys.* **49** (1968) 2422 – 2425.
63. L'Air liquide: *Encyclopédie des gaz,* Elsevier, Amsterdam 1976, p. 936.
64. M. Burgess, R. Germann. *AIChE J* **15** (1969) 272 –275.
65. J. D. Zel'venski, A. V. Strunina, *Gazov. Promyšlennost* **6** (1960) no. 1, 42 – 47.
66. E. Isaacs, F. Otto, A. Mather, *Can. J. Chem. Eng.* **55** (1977) 751 – 752.
67. A. Kohl, F. Riesenfeld: *Gas Purification,* 3rd ed., Gulf Pub. Co., Houston 1976, p. 783.
68. M. King, H. Al Najjar, *Chem. Eng. Sci.* **32** (1977) 1241 – 1246.
69. G. Charlot: *Les réactions chimiques en solutions aqueuses,* Masson, Paris 1983.
70. Société Nationale des Pétroles d'Aquitaine, FR 1435 788, 1964 (J. Solinhac). Shell, FR 2 134 677, 1972 (R. S. Anderson). Exxon, EP 45 636, 1981 (L. T. Howard). Elf Aquitaine, FR 2 601 350, 1986 (R. Voirin, A. Pepy).
71. J. C. Jorczak, E. M. Fettes, *Ind. Eng. Chem.* **43** (1951) 324.
72. D. Martin, H. G. Hauthal: *Dimethylsulfoxide,* Van Nostrand Reinhold Ltd., Wokingham, England 1971.
73. P. Sabatier, A. Mailhes, *C. R. Acad. Sci.* **150** (1910) 1217.
74. H. O. Folkins, E. L. Miller, *Ind. Eng. Chem. Prod. Res. Dev.* **1** (1962) 271.
75. Pure Oil, US 2 820 062, 1958 (H. O. Folkins, E. L. Miller).
76. Pennsalt Chemicals Corp., US 3 035 097, 1962 (T. E. Deger, B. Buchholz, R. H. Goshorn).
77. Y. K. Yurev, M. N. Prokina, *J. Gen. Chem. USSR* (*Engl. Transl.*) **7** (1937) 1868. Pennsalt, GB 913 584, 1960 (T. E. Deger, R. H. Goshorn, B. Buchholz).
78. Pennsalt, US 3 197 483, 1961 (T. E. Deger, R. H. Goshorn, B. Buchholz).
79. Société Nationale des Pétroles d'Aquitaine, FR 1 417 154, 1964 (P. Vannel). Shell, EP 122 654, 1980 (F. H. Elliott).
80. Société Nationale Elf Aquitaine (Production), Demande brevet français 88 01879, 1988 (E. Arretz).
81. Pennsalt, US 2 951 875, 1960 (B. Loev, R. H. Goshorn).
82. Pennsalt, US 4 102 931, 1978 (B. Buchholz).
83. Société Nationale Elf Aquitaine (Production), EP 101 356, 1982 (E. Arretz, A. Mirassou, C. Landoussy, P. Augé).
84. Société Nationale Elf Aquitaine (Production), FR 2 330 674, 1975 (E. Arretz, A. Pfister). Phillips, US 3 963 785, 1975 (D. H. Kubicek). Seitetsu, JP 79 132 551, 1978 (M. Kawamura, M. Hatta, M. Nakamura, N. Koune, Y. Iwasaki).
85. R. Dahlbom, *Acta. Chem. Scand.* **5** (1951) 690.
86. Toyo Chemical Industry, JP 81 147 763, 1980. Phillips, EP 208 323, 1985 (J. S. Roberts).
87. Henkel, BE 731 879, 1968. Henkel, BE 731 881, 1968.
88. Société Nationale des Pétroles d'Aquitaine, FR 1 359 678, 1963 (B. Pflugfelder, P. Vannel). Toyo Kasei, JP 81 110 667, 1980.
89. Pennwalt, US 4 281 202, 1980 (B. Buchholz, C. B. Welsh, H. C. Henry).
90. Société Nationale Elf Aquitaine (Production), EP 274 934, 1988 (E. Arretz, P. Augé, A. Mirassou, C. Landoussy).
91. GB 585 655, 1947 (F. N. Woodward). Röhm and Haas, DE 1 221 217, 1965 (H. Walter, W. Gruber). Société Nationale Elf Aquitaine (Production), FR 2 483 917, 1980 (E. Arretz).

Société Nationale Elf Aquitaine (Production), US 4507505, 1985 (E. Arretz).
92. T. L. Cairns et al., *J. Am Chem. Soc.* **74** (1952) 3982. R. Mayer, *Angew. Chem. Int. Ed. Engl.* **2** (1963) 277.
93. Société Nationale des Pétroles d'Aquitaine, FR 1 578 891, 1968 (Y. Labat).
94. B. Magnusson, *Acta Chem. Scand.* **13** (1959) 1031.
95. Du Pont, US 2 551 813, 1949 (P. S. Pinkney).
96. Stauffer, US 3 689 568, 1969 (R. J. Eletto).
97. Phillips, US 3 223 738, 1965 (D. L. Crain).
98. Phillips, US 3 050 452, 1960 (R. P. Louthan).
99. Pennwalt, US 4 140 604, 1978 (D. A. Dimmig).
100. Société Nationale Elf Aquitaine (Production), EP 5 400, 1979 (J. Ollivier, G. Souloumiac, J. Suberlucq).
101. Société Nationale Elf Aquitaine (Production), EP 60754, 1981 (E. Arretz, C. Landoussy, A. Mirassou, J. Ollivier).
102. Akzo, US 3 927 085, 1974 (A. G. Zenkel, A. Toth, H. Nagerlein, G. Meyer).
103. Dow Corning Corp., US 4 082 790, 1975 (J. L. Speier). Stauffer, US 4 740 623, 1988 (J. B. Heather). Société Nationale Elf Aquitaine (Production) Demande brevet françois 8 804 693, 1988 (Y. Labat).
104. J. E. Bittel, J. L. Speier, *J. Org. Chem.* **43** (1978) 1687.
105. P. Noble, D. S. Tarbell, *Org. Synth.* **4** (1963) 924.
106. Société Nationale des Pétroles d'Aquitaine, US 3 781 366, 1973 (E. Arretz, A. Donche, A. Pfister).
107. Wakunaga, Yakuhin Kabushiki Kaisha, US 4 212 826, 1979 (H. Yamaguchi).
108. Sankyo Organic Chemicals Cy, US 3 840 586, 1974 (T. Chiba et al.). Degussa, US 4 490 307, 1984 (H. Kuenk).
109. S. Dushman, *J. Am. Chem. Soc.* **43** (1921) 427.
110. C. M. Thacker, E. Miller, *Ind. Eng. Chem.* **36** (1944) 182.
111. K. T. Zimmermann, *Angew. Chem.* **74** (1962) 151.
112. K. Pohl, G. Rebentish, *Chem. Tech. (Leipzig)* **19** (1967) 503.
113. L. T. Filimonova et al., *Neftepererab. Neftekhim. (Moscow)* **1** (1976) 42; *Chem. Abstr.* **85** (1976) no. 96, 155.
114. Girdler, Company publication, "Compact Package H2S generators" Licensee: Chemical and Industrial Corp., Girdler, Zurich, Switzerland.
115. R. F. Bacon, E. S. Boe, *Ind. Eng. Chem.* **37** (1945) 469.
116. R. C. Forney, J. M. Smith, *Ind. Eng. Chem.* **43** (1951) 1841.
117. W. A. Bryce, *J. Chem. Soc.* **4** (1949) 3379.
118. F. Seel, *Angew. Chem.* **68** (1956) 789.
119. F. Gomez Vigide et al., *Inf. Quim. Anal. (Madrid)* **13** (1959) 61.
120. M. A. Korshunov et al., *Zavod. Lab.* **34** (1968) 285; *Chem. Abstr.* **69** (1968) 7944.
121. E. N. Konovalov et al., *Gasov. Prom.* **6** (1972) 47; *Chem. Abstr.* **77** (1972) 77 435.
122. R. M. Levit, *Zavod. Lab.* **27** (1961) 352; *Chem. Abstr.* **55** (1961) 25 387.
123. A. L. Kohl, F. C. Riesenfeld: *Gas Purification,* 4th ed., Gulf Publishing Co., Houston 1985.
124. J. I. Lee, F. D. Otto, A. E. Mather, *J. Chem. Eng. Data* **17** (1972) 465.
125. R. N. Maddox: *Gas and Liquid Sweetening,* 3rd ed., Campbell Petroleum Series, Norman (OK) 1982.
126. *Proceedings of Gas Conditioning Conferences,* University of Oklahoma – Norman (OK) (since 1981).
127. C. Blanc, J. Elgue, *Hydrocarbon Process.* **60** (1981) no. 8, 111 – 116.
128. Elf Aquitaine proprietary documents, Licensee: Elf Aquitaine.
129. J. M. Ryan, A. S. Holmes, US 4 293 322, 1981; *Chem. Abstr.* **96** (1982) P 37667.
130. *Dangerous Properties of Industrial Materials,* N-I-SAX-1979. Van Nostrand, New York 1979.
131. *Oil and Hazardous Materials.* Technical Assistance Data System U.S., Environmental Protection Agency, Oil and special materials Control Division, Washington 1980.
132. Sulfure d'hydrogène – Collection Enviroguide. Environnement Canada-Ottawa 1984.
133. Société SAPS – Anticorrosion – Brevet Français no. 2 591 122.
134. Determination of H2S in Sour Gas by Iodometry, SNEA (P) Internal Method no. 52.
135. Determination of H2S in Sour Gas by Gas-liquid Chromatography, SNEA (P) Internal Method no. 80.
136. Determination of H2S in Sour Gas by Argentimetry, SNEA (P) Internal Method no. 53.
137. Determination of H_2S in Natural Gas by a Cadmium-Sulfate-Iodometric Method, ASTM D 2385.
138. Determination of H2S in a Purified Commercial Natural Gas by a Polarographic Method, SNEA (P) Internal Method no. 134.

139. Determination of H_2S in a Fuel Gas by a Lead Acetate Reaction Rate Method, ASTM D 4085.
140. Determination of H_2S in the Atmosphere, ASTM D 4323.
141. P. Grancher, *Hydrocarbon Process.* July 1978, 155 – 160; Hydrocarbon Process. Sept. 1978, 257 – 262.
142. B. Gamson, R. Elkins, *Chem. Eng. Prog.* **49** (1953) 203 – 215.
143. *Criteria for a Recommended Standard Occupational Exposure to Hydrogen Sulfide,* U.S. Department of Health Education, and Welfare Public Health Service – Center for disease control NIOSH – May 1977.
144. R. O. Beauchamps et al.: "A critical review of the literature on Hydrogen Sulfide toxicity," *CRC Crit. Rev. Toxicol.* **13** (1984) no. 1, 25 – 97.
145. Toxigenics, Inc.: "90 day vapor inhalation toxicity study of hydrogen sulfide in Fischer 344 rats, CIIT docket no. 22063, in Sprague Dawley rats, CIIT docket no. 32063, in B6C3F1 mice, CIIT docket no. 42063. Chemical Industry Institute of Toxicology, Research Triangle Park, N.C., U.S. vol. 1, 2, 1983.
146. T. J. Hughes et al., "Validation of chemical and biological techniques for evaluation of vapors in ambient air/mutagenicity testing of twelve (12) vapor-phase compounds," Report EPA-600/1-84-005; order no. PB-84-164219, p. 204. *Gov. Rep. Announce. Index U.S.* **84** (1984) no. 10, 65.
147. M. M. Derzhko, I. R. Barilyak, "Effect of Sulfur and Hydrogen Sulfide on the Chromosomal Apparatus of Somatic Cells," *Dopov. Akad. Nank Ukr. RSR, Ser. B: Geol. Khim. Biol. Nanki* **1** (1986) 70 – 72.
148. D. Demaret, J. Fialaire, "Les intoxications par hydrogène sulfuré dans une raffinerie de gaz naturel," *J. Eur. Toxicol.* **7** (1974) no. 1, 32 – 36.
149. D. Demaret, J. Fialaire, "Intoxications par hydrogène sulfuré; urgence médicale dans l'entreprise," *Arch. Mal. Prof. Méd. Trav. Sécur. Soc.* **39** (1978) 761 – 767.
150. C. Azema, "Règles generales de sécurité dans l'emploi de l'hydrogène sulfuré," Soc. Nat. Elf-Aquitaine (Prod.), in press.
151. "Inflammabilité et explosivité des gaz, liquides et solides usuees," Institut National de Recherche et de la Securité, Note no. 740-63-71 2nd quarter 1971.
152. "Valeur limite pour les concentrations des substances dangereuses dans l'air des lieux de travail," Institut National de Recherche et de la Securité, Note ND 1653-129-87-4th quarter 1987.

Hydrometallurgy → Metallurgy

Hydrogenation and Dehydrogenation

Heterogeneous Catalysis and Solid Catalysts is a separate keyword.

PAUL N. RYLANDER, (Engelhard Corp.), Iselin, New Jersey 08830, United States

1. Hydrogenation 1
1.1. Hydrogen Availability and Mass Transport 2
1.2. Catalysts 2
1.3. Techniques 3
1.3.1. Slurry Reactors 3
1.3.2. Trickle-Bed Reactors 3
1.3.3. Bubble Column and Ebulliated Bed Reactors 3
1.4. Hydrogenation of Individual Compounds 3
2. Dehydrogenation 9
2.1. Styrene 9
2.2. Dehydrogenation of *n*-Alkanes . . . 10
2.3. Catalytic Reforming of Hydrocarbons 11
2.4. Liquid-Phase Dehydrogenation . . . 11
3. References 12

1. Hydrogenation

Catalytic hydrogenation refers to the addition of hydrogen to an organic molecule in the presence of a catalyst. If the molecule undergoes cleavage, the reaction is referred to as hydrogenolysis. These reactions are used to produce a variety of both bulk and fine organic chemicals. In addition, hydrogenation is often used in various purification processes (e.g., selective reduction of trace amounts of acetylene in an ethylene stream or traces of butadiene in a butene stream, and removal of unwanted oxygen, hydrogen, or carbon oxides from a variety of systems).

A large number of hydrogenation catalysts are now available, and the scope of hydrogenation has become very large. Most functional groups can be reduced readily, often under mild conditions, and frequently with high chemo-, regio-, and stereoselectivity. Monographs detailing the scope and use of hydrogenation have been written by AUGUSTINE [1], FREIFELDER [2], RYLANDER [3–5], and ZYMALKOWSKI [6].

Hydrogenation is an exothermic reaction and the equilibrium usually lies far toward the hydrogenated product under most operating temperatures:

$$A + n\,H_2 \rightleftharpoons B \qquad \Delta H < 0$$

As the temperature increases, the equilibrium shifts toward the left, and the reverse reaction, dehydrogenation, can itself be a useful synthetic process (see Chap. 2). Industrial hydrogenation requires a means of limiting the temperature increase arising from the exothermic reaction; a higher temperature usually has a favorable effect on reaction rate, but may affect selectivity and catalyst life adversely.

Hydrogenation is influenced by a number of factors, including catalyst, solvent, substrate purity, and operating conditions. Temperature, pressure, agitation, and catalyst loading can each influence both the rate and the selectivity of hydrogenation. In general, as temperature, pressure, agitation, and catalyst loading increase, the rate of hydrogenation increases until a limit is reached. The influence of operating variables on selectivity is less straightforward but can often be linked to the effect of these and other variables on hydrogen availability at the catalyst surface.

1.1. Hydrogen Availability and Mass Transport

Hydrogenation is carried out in both the liquid and the gas phases, but liquid-phase reactions are much more common. In liquid-phase systems, transport of hydrogen to the catalyst surface is frequently the rate-limiting step. To react with the substrate, hydrogen must move from the gas phase into the liquid and then to the solid catalyst and into its porous structure. This net movement of hydrogen to the catalyst is the result of concentration gradients, which develop when hydrogen is consumed by the catalytic reaction.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_487

The concentration of hydrogen at the catalyst surface can vary widely, depending on the ratio of the rate of hydrogen consumption to the rate of hydrogen supply [7].

This rate ratio can have an important influence on selectivity. It is, for instance, an important factor in controlling selective removal of multiple unsaturation and the extent of cis–trans isomerization during hydrogenation of natural oils to make edible products.

1.2. Catalysts

Hydrogenation catalysts are of two types, heterogeneous and homogeneous. *Heterogeneous catalysts* are solids that form a distinct phase in the gas or liquid environment. *Homogeneous* catalysts dissolve in the liquid medium, forming only a single phase; a well-known catalyst of this type is the Wilkinson catalyst, $(Ph_3P)_3RhCl$.

Homogeneous Catalysts. Homogeneous hydrogenation catalysts are used infrequently in industry. They are difficult to separate from the reaction media and are not employed if a suitable heterogeneous catalyst is available. Nonetheless, they are useful in a variety of hydrogenation processes that are not possible with heterogeneous catalysts [8], [9]. Expanded use of homogeneously catalyzed hydrogenation awaits only industrial demand for its products.

A noteworthy industrial use of a homogeneous catalyst is the asymmetric hydrogenation of an achiral olefinic precusor of L-Dopa (3,4-dihydroxyphenylalanine, used in the treatment of Parkinson's disease; see → Parkinsonism Treatment) over a complex rhodium catalyst carrying chiral ligands. Optical yields are nearly quantitative.

Heterogeneous Catalysts. Heterogeneous catalysts can be divided into two types: those used in *fixed-bed processing,* in which the catalyst is stationary and the reactants pass over it, and those used in *fluidized-bed* or *slurry processing.* The former have a relatively large particle size, 0.79 – 6.35 mm, in various forms. Slurry or fluidized-bed catalysts are fine powders that can be suspended readily in a liquid or gas.

Many *metals* and *metal oxides* have general hydrogenation activity, but marked differences exist among them. Nickel, copper, cobalt, chromium, zinc, iron, and the platinum group (platinum, palladium, rhodium, and ruthenium) are among the elements frequently found in commercial hydrogenation catalysts. Combinations of these or other elements are sometimes used to confer additional activity, selectivity, stability, or life.

Palladium, platinum, rhodium, and ruthenium make exceptionally active hydrogenation catalysts and are used frequently when equipment or stability limitations permit operation only under mild conditions. They can effectively reduce most functional groups. Nickel is used to reduce a variety of functional groups, usually under more vigorous conditions than the platinum group. Copper chromite is used mainly to reduce esters to alcohols. Cobalt or iron, in the presence of ammonia, is an effective catalyst for the reduction of nitriles to primary amines.

Catalyst Selectivity. In series reactions of the type $A \rightarrow B \rightarrow C$, both mechanistic and thermodynamic selectivity are involved in the formation of B. These selectivities can vary widely with the catalyst used. *Thermodynamic selectivity,* which is related to the relative free energies of adsorption of A and B, refers to the competition of A and B for catalyst sites. The tendency for B to occupy a catalyst site relative to A increases as B accumulates and as A is depleted. High thermodynamic selectivity demands that A be adsorbed much more strongly than B and that the rate of displacement of B by A be high relative to the rate of hydrogenation of B to C. Thermodynamic selectivity is affected by depletion of A in the vicinity of the catalyst because of intra- and interparticle diffusion resistance. *Mechanistic selectivity* refers to the extent to which C is formed from A without the formation of desorbed B; it is related to the relative rates of $A \rightarrow B$ and $A \rightarrow C$ in the equation $B \leftarrow A \rightarrow C$. Either thermodynamic or mechanistic selectivity may be more important, but both must be high if B is to be obtained in good yield.

1.3. Techniques

All hydrogenation reactors serve to bring hydrogen, catalyst, and substrate together in the absence of oxygen. Liquid-phase substrates may be hydrogenated neat or in the presence of a suitable *solvent*; use of a solvent is one of the best available means for markedly altering selectivity. Solvents also help to moderate the heat of reaction, aid in catalyst handling and recovery, and permit the reduction of solid substrates. A convenient solvent may be the product itself or the solvent used in a previous or subsequent step. Militating against the use of solvent is diminished space – time yield and the cost of solvent and of its removal.

1.3.1. Slurry Reactors

Hydrogenation is exothermic, and some means of removing the reaction heat must be provided to avoid excessive temperature. In liquid-phase slurry reactions, cooling is usually accomplished by coils within the reactor. The reaction mass can also be cooled by circulating the mixture over external cooling elements. The latter save reactor space and permit the reactor to be cleaned more easily.

Slurry reactors must be provided with some sort of agitation to aid in the rapid transport of hydrogen from the gas phase to the catalyst suspended in the liquid medium. Agitation also helps in the removal of heat, and agitators can be designed to favor either hydrogen transport or heat removal, as required. Externally cooled systems can be engineered to return the reaction slurry to the reactor by a venturi tube through which hydrogen is drawn into the passing stream. Such a system gives efficient mixing, and much of the hydrogenation is thought to take place within the nozzle itself. Alternatively, the slurry can be returned to the reactor through a shower-type arrangement in which substrate and catalyst particles are sprayed into hydrogen in the head space.

1.3.2. Trickle-Bed Reactors

Continuous hydrogenation is carried out commonly in trickle-bed reactors (columns packed with catalyst over which substrate and hydrogen pass concurrently downward). This type of reactor is used extensively in the petroleum industry for various reactions such as hydrocracking and hydrodesulfurization. The residence time distribution is close to that of plug flow, which permits high conversion in a single vessel. Operation is continuous, handling is minimal, and catalyst separation from product is trivial. Heat removal, however, is a major problem unless small-diameter tubes are used. At times, reactors are placed in series, with hydrogen injected between beds for temperature control. Beds are usually of uneven height, becoming longer as the reactant is depleted [10].

1.3.3. Bubble Column and Ebulliated Bed Reactors

Bubble columns resemble trickle-bed reactors except for the direction of flow. In bubble column reactors, flow is upward; liquid fills the void space, and the catalyst is completely wetted. Ebulliated bed reactors are similar except that the catalyst is fluidized by high gas and liquid flow rates. This type of reactor is used in H-oil and H-coal processes (see also → Coal Liquefaction, Chap. 3.9.5.).

1.4. Hydrogenation of Individual Compounds

Alkynes. Alkynes are adsorbed on a catalyst surface much more strongly than the corresponding alkenes, thus permitting selective hydrogenation of traces of alkynes in massive amounts of alkenes. Hydrogenation of trace alkynes, usually over a low percentage of palladium on an alumina support with restricted hydrogen supply, is used widely to purify various alkene-containing streams.

The strong adsorption of alkynes also permits selective reduction to the cis alkenes in high yield, a reaction used in the production of a variety of fine chemicals, such as vitamins A and K, prostaglandins, and pheromones. The Lindlar catalyst (Pd – Pb on $CaCO_3$), with or without additional modifiers, is often used for this selective reduction. Ideally, the reaction comes to a

virtual halt after absorption of 1 mol of hydrogen, but terminating the reaction after consumption of the stoichiometric amount of hydrogen may be necessary. The largest volume hydrogenations are of alkynols and alkynediols. Butane-1,4-diol [*110-63-4*], derived from acetylene, has long been made by hydrogenation of butynediol over a base-metal catalyst such as nickel – copper – manganese on silica (→ Butanediols, Butenediol, and Butynediol, Chap. 1.3.). Nickel, the major component, is an effective catalyst for hydrogenation of both acetylenic and olefinic bonds. At the butenediol stage, some aldehyde is formed by rearrangement; this is removed effectively by copper. A small amount of manganese stabilizes the structure. The active binary catalyst palladium – ruthenium on carbon behaves in similar fashion.

Alkenes. Hydrogenation of olefins has extensive industrial application. Dripolene, or pyrolysis gasoline, is stabilized against oxidation by selective reduction of dienes in a complex mixture of olefins and aromatics. The preferred catalyst consists of pellets of palladium on alumina. If the aim of reduction is to produce gasoline, olefins are preserved to boost the octane rating, but if the product is to be used as a source of pure aromatics, olefins are removed in a second stage over a base-metal catalyst, such as nickel or cobalt – molybdenum. Two steps are used for complete olefin hydrogenation, because the more vigorous conditions required for complete reduction cause severe fouling of the catalyst by polymerization of the diene.

The largest single application of olefin hydrogenation is the partial hydrogenation of natural oils to produce margarine, shortening, toppings, salad oil, and other edible products (→ Margarines and Shortenings). A huge market also exists for products derived by more complete hydrogenation; the problem of selectivity described below is not present.

Edible products are derived by partial hydrogenation of various oils (e.g., cottonseed, soybean, sunflower, and rapeseed oil) all of which are glycerides of long-chain saturated and cis-unsaturated fatty acids (→ Fats and Fatty Oils, Chap. 4.3.). Partial hydrogenation of these products improves oxidative stability and alters plastic properties suitably. Oxidative instability is caused primarily by the presence of various homoconjugated dienes and trienes, and an aim of hydrogenation is preferential removal of multiple unsaturation. This task is facilitated by the tendency of hydrogenation catalysts to isomerize homoconjugated into conjugated olefins; the latter, relative to monoenes, are strongly adsorbed on the catalyst. Catalysts differ widely in their ability to promote isomerization, and a close parallel exists between this ability and the selective removal of polyunsaturation. As a consequence of double-bond migration, isomerization of the naturally occurring all-cis olefins into trans olefins may or may not occur, depending on the conformation of the chain at the time of migration. Many catalysts have been studied for these reactions, but industrial practice overwhelmingly favors some form of nickel. Hydrogenation of edible oils is carried out almost entirely in batchwise slurry operation at 140 – 190 °C and 0.2 – 0.3 MPa.

Notwithstanding the huge volume of product, processing is largely carried out in a slurry batch mode. Continuous fixed-bed operation is made difficult by intermittent production, variable feedstocks, and the problem of adjusting conditions to compensate for declining catalyst activity.

Aromatic and Heteroaromatic Rings. Compounds with aromatic or heterocyclic rings are hydrogenated readily and usually without difficulty if other reducible groups are absent. Cyclohexane [*110-82-7*] is obtained mainly by hydrogenation of sulfur-free benzene over supported fixed-bed nickel or platinum catalysts (170 – 230 °C, 2.5 MPa, → Cyclohexane, Chap. 4.1.). The reaction is reversed at higher temperature, which sets an upper limit to a permissible operating temperature of ca. 230 °C. With catalysts of controlled acidity, the product can be obtained virtually free of benzene and methylcyclopentane. Tetralin [*119-64-2*] is produced by hydrogenation of naphthalene over nickel sulfide or nickel – molybdenum at 400 °C and 2 – 6 MPa; a higher pressure gives decalin.

Hexahydrobenzoic acid [*98-89-5*], an intermediate for caprolactam (nylon 6 monomer), is made by hydrogenation of benzoic acid over palladium on carbon powder by using the product as solvent to help prevent excessive temperature (→ Caprolactam, Chap. 4.3.). When only base-metal catalysts were available, this was consid-

ered a very difficult reduction; base-metal catalysts are attacked by the carboxylic acid function.

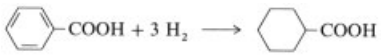

An important monomer 1,4-bis(hydroxymethyl)cyclohexane is made by two-stage hydrogenation of dimethyl terephthalate. The first stage (160 – 180 °C, 30 MPa), carried out in a fixed-bed reactor over supported palladium, produces dimethyl hexahydroterephthalate. Part of this product is recycled to moderate the temperature increase resulting from the highly exothermic hydrogenation (197.6 kJ/mol). The second stage uses a copper chromite catalyst under similar conditions.

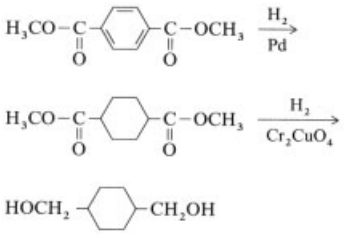

Perhydrogenated rosins are made by hydrogenation of rosin. Rosins contain dehydroabietic acid and similar compounds, which are difficult to reduce. The preferred catalyst is palladium, because it causes the least amount of unwanted decarbonylation and decarboxylation, reactions that lead to catalyst inhibition through competing adsorption with carbon monoxide for catalyst sites [11].

Cyclohexanone [*108-94-1*] is produced by hydrogenation of phenol either in the vapor phase or in a liquid slurry over supported palladium catalysts (→ Cyclohexanol and Cyclohexanone, Chap. 3.1.). In the latter, purified phenol and catalyst slurry enter the first of several reactors and are exposed to synthesis gas (75 % N_2, 25 % H_2). A controlled pressure drop between the reactors ensures forward flow without the need for pumps. The product with the lowest boiling point, cyclohexanone, is removed as vapor in the gas flow. In the last reactor, conversion of phenol is about 95 % complete. Slurry emerging from the last reactor is centrifuged, and recovered catalyst is returned for recycling. Reactor temperature is held between 155 and 170 °C, pressure at 0.5 – 1.4 MPa [12].

Carbonyl Compounds. Aldehydes and ketones behave similarly on hydrogenation, and the same order of catalyst activity usually applies to both functional groups. The difference between aromatic and aliphatic carbonyls is marked, however, and the preferred catalysts differ.

Aliphatic carbonyl compounds are reduced to the corresponding alcohols with little danger of overhydrogenation. Alcohols are in fact an excellent solvent for hydrogenation, even under vigorous conditions.

Hydrogenation of mesityl oxide [*141-79-7*] is a good example. Formed by aldol condensation of acetone, mesityl oxide can be quantitatively reduced to methyl isobutyl ketone (MIBK) over palladium; if desired, MIBK can be reduced further to 4-methylpentan-2-ol over base metals such as nickel or copper.

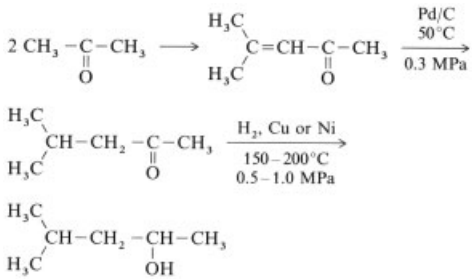

The first two steps can be combined by employing a dual-functional catalyst, i.e., one capable of catalyzing both condensation and hydrogenation; examples include palladium on zirconium phosphate or on a zeolite.

2-Ethylhexanol [*104-76-7*] an important chemical widely used in plasticizers, is also derived from aldol condensation, followed by hydrogenation of the resulting 2-ethyl-2-hexenal.

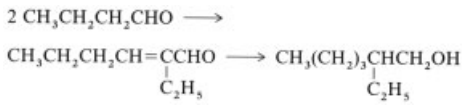

Reduction is usually carried out in two stages over supported nickel, with the carbon – carbon double bond reduced first, followed by the aldehyde. One-stage reduction is also performed, but

the two-stage process makes temperature control easier and yields better product quality (see also → 2-Ethylhexanol, Chap. 3.).

Sorbitol [*50-70-4*] is produced on a large scale by hydrogenation of D-glucose over a nickel or ruthenium catalyst. Sorbitol is used in a variety of foods, pharmaceuticals, and cosmetics, as well as in the production of polyurethanes, resins, and vitamin C.

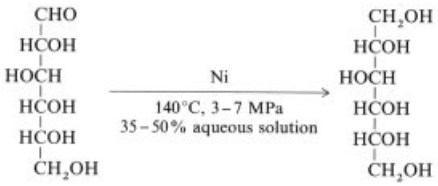

This hydrogenation is not without complications. The temperature must be high enough to give acceptable rates, but not so high as to produce various byproducts resulting mainly from cyclization. The process must be operated within a narrow pH range. Nickel catalysts are damaged if the solution becomes too acidic. If the solution becomes too alkaline, gluconic acid is produced by the Cannizzaro reaction, and mannitol is formed through epimerization at the α-carbon [13].

Aromatic carbonyl compounds are reduced easily, and either reduction to the carbinol or loss of the hydroxyl group can be achieved at will.

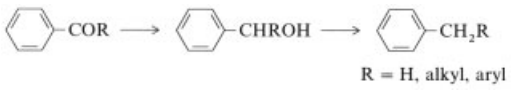

Both reactions provide a rich source of products and intermediates in the synthesis of fine chemicals. Palladium is often the catalyst of choice for either hydrogenation or hydrogenolysis. The same catalyst can be used in both steps; in neutral media, both the mechanistic and the thermodynamic selectivities toward the alcohol are very high. The reactions are usually conducted in the presence of small amounts of strong acid when the hydrogenolysis product is desired.

Nitro Compounds. Most nitro compounds are reduced in slurry reactors, although nitrobenzene itself is converted to aniline [*62-53-3*] in vapor-phase fixed-bed reactors as well as in fluidized beds. Catalysts for vapor-phase hydrogenation vary with the producer and include sulfided nickel chromates, nickel – aluminum, silver – manganese oxide, and copper – silica. The reaction temperature is high (300 – 475 °C), and the pressure is 0.5 – 1.5 MPa [14]. Aniline is also produced in continuous slurry reactors by using palladium on carbon black promoted by platinum and iron, palladium on activated carbon or nickel on silica [14] (see also → Aniline, Chap. 3.2.).

The course of hydrogenation of nitroaromatic compounds is complex, and a variety of products can be obtained. Partial hydrogenation yields phenylhydroxylamines. Reaction is best under mild conditions in neutral media with platinum on carbon, inhibited by addition of dimethyl sulfoxide to prevent overhydrogenation [15].

Phenylhydroxylamines are key intermediates in several important processes. In aqueous acid solution, phenylhydroxylamine rearranges to yield 4-aminophenol. If the hydrogenation of nitrobenzene is carried out in sulfuric acid, the intermediate phenylhydroxylamine can be made to rearrange as it forms. This process has become the preferred route to 4-aminophenol [*123-30-8*] (see also → Aminophenols). The hydroxylamine is susceptible to further hydrogenation, a reaction whose rate must be made minimal relative to the rate of rearrangement.

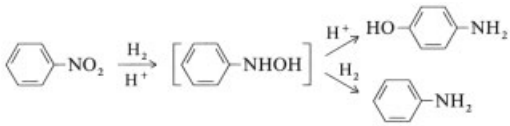

The preferred catalyst for this process is platinum on carbon. Carbon is inert to the hydrolytic environment, and platinum is resistant to attack by acid, as well as being the most effective catalyst for converting nitrobenzene to the hydroxylamine.

Reduction of nitrobenzene in aqueous sodium hydroxide solution over platinum on carbon as catalyst is used to produce hydrazobenzenes. Sodium hydroxide promotes coupling between the intermediates, nitrosobenzene and phenylhydroxylamine. The coupling reaction with subsequent benzidine rearrangement is a useful source of diaminobiphenyls [16].

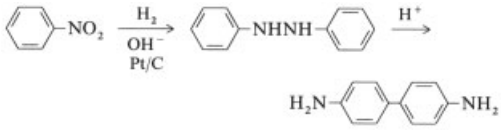

Diaminotoluene is produced by hydrogenation of dinitrotoluene in the liquid phase by using supported palladium or nickel catalyst. Depending on the producer, conditions vary widely: 50 – 130 °C and 0.2 – 20 MPa. The product is converted mostly to toluene diisocyanate for use in polyurethanes.

Hydrogenation of Nitrate Ion. The hydrogenation of nitrate ion in phosphoric acid or of nitrogen monoxide in sulfuric acid provides large-scale routes to hydroxylamine [*7803-49-8*] an intermediate for caprolactam (see also → Hydroxylamine, Chap. 4.1.). The latter process produces 0.9 t of ammonium sulfate (a sometimes unwanted byproduct) per ton of caprolactam; the former, the hydroxylamine – phosphate – oxime (HPO) process, produces none, and employs novel chemistry and engineering (→ Caprolactam). Reduction is carried out in a slurry with an aqueous solution of 20 % phosphoric acid and 20 % ammonium nitrate over palladium on activated carbon, promoted by germanium dioxide, an unusual promoter. The reduced solution is filtered through a bank of filter sticks mounted within the hydrogenator. The filtrate then passes in countercurrent fashion into a cyclohexanone –toluene mixture to form cyclohexanone oxime in toluene; the depleted aqueous solution is returned to the hydrogenator and enriched with more nitrate for further reaction [17].

Nitriles. Catalytic hydrogenation of nitriles yields a number of products, including primary, secondary, and tertiary amines, aldehydes, and alcohols. All of these products arise from intermediate imines.

$$RCN \xrightarrow{H_2} RHC{=}NH \xrightarrow{H_2} RCH_2NH_2$$

The primary amine formed in this sequence can give an addition product with the imine, which is converted to a secondary amine, either by elimination of ammonia and subsequent hydrogenation, or by direct hydrogenolysis.

$$2\,RCN \xrightarrow{H_2} \underset{\underset{NHCH_2R}{|}}{RCHNH_2} \xrightarrow{H_2} RCH_2NHCH_2R + NH_3$$

Tertiary amines are formed similarly from the addition of a secondary amine to an imine. Aldehydes are obtained in good yield by hydrogenation of a nitrile in aqueous media.

$$RCN \xrightarrow{H_2} RHC{=}NH \xrightarrow[H_2O]{H^+} RCHO + NH_4^+$$

Reduction of a nitrile in the presence of an added amine permits the formation of unsymmetrical amines.

$$R^1CN + R^2NH_2 \xrightarrow{H_2} R^1CH_2NHR^2 + NH_3$$

The major product obtained in the hydrogenation of nitriles depends strongly on the catalyst. For instance, hydrogenation of low molecular mass aliphatic nitriles over rhodium gives secondary amines in high yield, whereas excellent yields of tertiary amines are obtained by reduction over platinum or palladium.

Primary amines are formed from nitriles most easily by hydrogenation carried out in the presence of ammonia. The latter prevents secondary amine formation by competitively reacting with the intermediate imine.

$$RCN + NH_3 + H_2 \longrightarrow \underset{\underset{NH_2}{|}}{RCHNH_2} \xrightarrow{H_2} RCH_2NH_2 + NH_3$$

Cobalt, nickel, and iron are useful catalysts for the hydrogenation of nitriles; reduction over these catalysts is usually carried out under vigorous conditions. Platinum-group catalysts operate under mild conditions and are used in the production of fine chemicals.

Adiponitrile. A much studied and used reduction of nitriles is the synthesis of hexamethylenediamine [*124-09-4*] from adiponitrile (see also → Hexamethylenediamine, Chap. 4.). Hexamethylenediamine is a monomer of nylon 66, as well as an intermediate in the production of diisocyanates used for foams and resins. Reduction of dinitriles presents special problems: secondary amine formation can, in principle,

continue indefinitely, resulting in higher molecular mass compounds that cause catalyst fouling and a loss in yield. Additionally, intramolecular coupling can occur to produce hexamethyleneimine. Both of these reactions are suppressed by ammonia. Hydrogenation of adiponitrile also produces small amounts of 1,2-diaminocyclohexane, which is unusual in that carbon–carbon rather than carbon–nitrogen bond formation takes place. Cobalt, nickel, or iron catalysts are used at high temperature and pressure (100–180 °C, 30–60 MPa). Smaller quantities of the higher homologues of adiponitrile are produced in similar fashion.

Reductive Alkylation and Amination. Reductive alkylation is the term applied to the process of attaching alkyl groups derived from an aldehyde or ketone to ammonia or to primary and secondary amines. The term is also applied to those compounds that are not amines or carbonyls but are converted to them in the course of reduction. Reductive alkylation is a two-step process in which the amine and the carbonyl first react to give an intermediate that is converted to the final product either by elimination of water and subsequent hydrogenation of the resulting imine, or by direct hydrogenolysis of the addition product.

Nickel, rhodium, palladium, platinum, and Raney cobalt have all been used for industrial alkylation. Palladium and platinum are equally effective when reduction is carried out with small molecules (e.g., formaldehyde or acetone), but as the molecule increases in size, platinum becomes the better catalyst. Sulfided nickel, platinum, or rhodium catalysts are very effective for reductive alkylation. These catalysts can be used with lower quality feedstocks because they are more resistant to poisoning than nonsulfided catalysts; furthermore, they have a diminished tendency to reduce the carbonyl to an alcohol and are useful in reductive alkylation when dehydrohalogenation must be avoided, as in the conversion of halonitroaromatics to N-alkylated haloanilines [18].

Reductive alkylation is used widely for the synthesis of alkylated anilines. Compounds of this type are important antioxidants (→ Antioxidants).

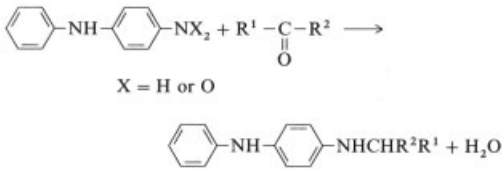

Either the aniline or the nitro compound makes a suitable reactant; both are used industrially.

Carboxylic Acids and Esters. Carboxylic acids and their esters are reduced only with difficulty; both require vigorous conditions, unless activated. Rhenium and ruthenium have been used to reduce carboxylic acids at 20–100 MPa and up to 300 °C. Decarboxylation occurs as a side reaction, the extent increasing with increasing temperature. Carboxylic esters are reduced more easily than free acids; the latter tend to attack both the equipment and the catalyst. Copper chromite catalysts are often used for the hydrogenolysis of esters.

An industrial route to the important monomer butane-1,4-diol [*110-63-4*] involves esterification of maleic anhydride with ethanol to obtain diethyl maleate, and hydrogenolysis of the diester over copper chromite; butyrolactone and tetrahydrofuran are byproducts. The relative amounts of these products can be controlled by reaction conditions. The equilibrium ratio of butyrolactone and butane-1,4-diol depends markedly on pressure: at 0.1 MPa and 270 °C, butyrolactone is highly favored; at 30 MPa, the diol is favored in the ratio 94 : 6 (see → Butyrolactone).

Fatty alcohols are produced in large volume by hydrogenolysis of fatty acids, or preferably their esters, under vigorous conditions (275–300 °C, 20–26 MPa). Reduction of acids requires the use of corrosion-resistant equipment; ester reduction is carried out in carbon steel equipment. Natural triglycerides are suitable substrates, but methyl esters are preferable. The latter allows easy separation of the product alcohols and avoids problems connected with high-temperature decomposition of glycerol. Catalysts are usually a type of copper chromite. These reduce double bonds as well as ester groups so that saturated alcohols are produced regardless of starting material. Unsaturated fatty alcohols are produced by hydrogenation of unsaturated esters over cadmium-based

catalysts at a temperature lower than that normally used in ester reduction (→ Fatty Alcohols, Chap. 2.3.3.2., → Fatty Alcohols, Chap. 3.).

2. Dehydrogenation

Dehydrogenation is an important industrial reaction, although the numbers of applications are few compared to hydrogenation. The lesser use of dehydrogenation stems from several aspects of its chemistry.

Dehydrogenation is an endothermic reaction, and the high temperature necessary to obtain a useful conversion to the dehydrogenated product restricts the reaction to those substrates and products with suitable thermal stability. In general, industrial dehydrogenation falls into one of the following categories: (1) the acceptable product is a complex mixture, as in the mixture of olefins obtained from alkane dehydrogenation, used for detergent alkylate; (2) an unsaturated system is extended, as in the conversion of benzene to biphenyl or ethylbenzene to styrene; (3) aromatization occurs, a reaction that provides a driving force as well as a convenient stopping place; (4) a point of attack exists, as in the conversion of methanol or ethanol to formaldehyde or acetaldehyde, respectively; (5) the choice of product is limited, as in the conversion of butane or butene to butadiene [9].

2.1. Styrene

Styrene [*100-42-5*], an important monomer, is obtained largely by dehydrogenation of ethylbenzene in the presence of superheated steam (→ Styrene).

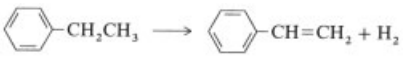

The catalyst for this reaction usually contains some major active ingredient, a stabilizer, and an alkali-metal promoter. Common commercial catalysts contain iron oxide promoted by potassium and a structural stabilizer. Potassium is a very effective promoter and increases the reaction rate by an order of magnitude. Potassium is volatile under reaction conditions and moves toward the reactor exit; redistribution of potassium occurs within the catalyst particles as well. Styrene formation is endothermic, and as a consequence, the center of a catalyst pellet in use becomes cooler than the edges; this temperature gradient causes inward migration of potassium, resulting in catalyst deactivation near both the center and the edges due to excess and deficient potassium, respectively.

Superheated steam plays an important role in styrene production and serves several purposes: it is introduced directly into the reactor in large molar excess to provide part of the necessary heat to drive the endothermic reaction; it serves as an oxidant to maintain the iron oxide catalyst in the appropriate oxidation state and removes carbonaceous deposits from the catalyst through the water – gas shift reaction; it reduces the partial pressure of hydrocarbons, thereby increasing conversion to styrene. Optimum conversion is determined by the balance between the advantage of increased conversion and the cost of increased steam. Over the last two decades, improved processing has permitted a continual decrease in the molar ratio of steam to hydrocarbon. The ratio is between 6 and 12, with 65 – 70 % conversion and 90 – 95 % selectivity, at 600 – 650 °C and 51 – 101 kPa. Some reactors operate in an isothermal mode, heat being supplied by a heat-transfer medium on the outside of the reactor tubes, which offers the advantage of reduced steam requirements; however, this advantage is offset by higher capital costs. Most processes operate in the adiabatic mode. In practice, conversion is limited by the equilibrium between styrene, ethylbenzene, and hydrogen. As styrene approaches equilibrium concentration, its rate of formation decreases but the rate of byproduct formation does not; the yield, therefore, falls with increased conversion [19], [20]. The limitation on conversion can be cirumvented if the reaction is run as an oxidative dehydrogenation, (i.e., with oxygen present to combine with liberated hydrogen and remove it from the equilibrium). Despite much effort, no large-scale plant operates in this mode.

The technology of ethylbenzene dehydrogenation has been extended to produce both α-methylstyrene and divinylbenzene by dehydrogenation of cumene and diethylbenzene, respectively.

2.2. Dehydrogenation of *n*-Alkanes

Olefins are hydrogenated easily and quantitatively to alkanes under ambient conditions, but the reverse reaction is not achieved as readily. In hydrogenation, the olefinic function provides a point of attack and the stability of the alkane product provides automatic termination. Alkanes, on the other hand, have no unique point of attack, and the alkene resulting from dehydrogenation is more easily dehydrogenated than the starting material itself. Moreover, at ambient conditions, the equilibrium between alkane and alkene lies completely in favor of the alkane [21]. Alkenes begin to appear in traces only at 300 °C; at least 400 °C is required for more substantial conversion (ca. 10 %). The high temperature necessary for alkene formation also favors side reactions such as polymerization and cracking.

Dehydrogenation of alkanes is greatly helped by a suitable catalyst where it is thermodynamically possible at a temperature below that of thermal decomposition. Propane, butane, and alkanes in the detergent alkylate range (ca. C_{10} – C_{14}) have all been catalytically dehydrogenated on a large scale. Little advantage is gained by catalysis in the dehydrogenation of C_2 hydrocarbons, because the equilibrium is unfavorable at a temperature below that required for thermal decomposition.

Cracking (cleavage of the carbon – carbon bond) is energetically favored over cleavage of the carbon – hydrogen bond, and catalysts such as nickel, which are excellent for hydrogenation of alkenes, lead to extensive formation of methane when applied to dehydrogenation.

The reaction expressing the general alkane – alkene equilibrium is

$$C_nH_{2n+2} \rightleftharpoons C_nH_{2n} + H_2$$

Alkenes are favored by lower pressure, as expressed by the equation

$$C^2/\left(1-C^2\right) = K/p$$

where C is the fraction dehydrogenated, K is the equilibrium constant, and p the pressure in atmospheres. This equation holds for the conversion of alkanes to monoenes and of monoenes to dienes. Conversion of alkanes to dienes (e.g., butane to butadiene)

$$C_nH_{2n+2} \rightleftharpoons C_nH_{2n-2} + 2\,H_2$$

is related to pressure by the equation

$$4C^3/(1-C)\,(1+2C)^2 = K/p^2$$

Conversion is increased by operation at lower pressure or by use of a diluent.

Butenes and butadiene are produced by dehydrogenation of butane, although the process finds increasing competition from oxidative dehydrogenation and from thermal cracking (→ Butadiene, Chap. 4.; → Butenes, Chap. 4.). Operating conditions vary: the temperature is 550 – 650 °C and the pressure 30 – 300 kPa, contact times are a few seconds. Catalysts coke rapidly and are oxidatively regenerated every 10 – 100 min. By mixing the catalyst with inert material, much of the heat liberated during oxidation can be captured and used to drive the endothermic dehydrogenation. Most catalysts in this service are based on chromia – alumina. These catalysts are strongly inhibited by water vapor; the butane feed must, therefore, be dried carefully.

Other catalysts, such as nickel – calcium phosphate stabilized with chromia, iron oxide promoted by potassium, and chromium oxides, require the presence of steam to help keep the catalyst in the proper oxidation state. Excessive coking occurs if these catalysts become reduced. Steam also acts as a source of heat and shifts the conversion toward the dehydrogenated product by lowering the partial pressure of the latter.

2.3. Catalytic Reforming of Hydrocarbons

The objective of catalytic reforming is the conversion of saturated hydrocarbons to aromatics as selectively as possible. This process is the chief source of high-octane fuel and of aromatic hydrocarbons. Catalytic reforming involves a complex series of reactions in which dehydrogenation figures prominently. Catalysts are dual-functional, having both a metallic element to catalyze dehydrogenation – hydrogenation and an acidic function to catalyze hydrocarbon rearrangements. Catalysts typically contain a small amount of highly dispersed platinum (<1 %)

supported on alumina (150 – 300 m^2/g), as well as a second metal such as rhenium. Polymetallic catalysts permit operation under more severe conditions; they also possess improved stability and activity.

Typical operating conditions are 430 – 530 °C and 1.0 – 3.5 MPa. A unit consists of several, frequently four, fixed-bed reactors in series. Heated naphtha enters the first reactor, where the major reaction is endothermic dehydrogenation of cycloalkanes accompanied by a decrease in temperature. The effluent is reheated and the process continued for all reactors. The liquid product contains 60 – 70 wt % aromatics; the gaseous product contains light hydrocarbons admixed with 60 – 90 mol % hydrogen. The gaseous portion is recycled to the first reactor in a ratio of 5 – 10 mol of recycle gas per mole of naphtha to maintain a high partial pressure of hydrogen in the system. The presence of hydrogen diminishes the yield of aromatics, but hydrogen is necessary to maintain catalyst activity by retarding formation of hydrocarbon residues that poison the catalyst. An aim of processing is operation at as low a partial pressure of hydrogen as possible, commensurate with good catalyst life.

Dehydrogenation of cyclohexanes is the most facile of all reactions that occur in reforming. On the other hand, only traces of olefins can be formed by dehydrogenation of alkanes under reforming conditions, because of an unfavorable equilibrium. Nonetheless, dehydrogenation is of considerable importance because olefins appear to be intermediates in the overall process [22].

2.4. Liquid-Phase Dehydrogenation

The constraints imposed on dehydrogenation by equilibrium considerations can be loosened if hydrogen is removed from the products as they are formed. Dehydrogenation in the liquid phase in an open vessel fulfills this goal. Hydrogen escapes from the system, whereas products and reactants remain in the vessel or are returned after condensation. The lower temperature limit is then set not by the equilibrium but by a convenient reaction rate.

In practice, the substrate, neat or dissolved in an inert solvent of convenient boiling point, is refluxed in the presence of catalyst. Hydrogen escapes while other volatile products are condensed and returned, with or without removal of the product. If the product is removed, the operation can be made continuous.

All the considerations involving the influence of mass transport on the rate of reaction (see Section 1.1) apply in reverse to dehydrogenation. Good stirring helps to remove liberated hydrogen from the vicinity of the catalyst, and reflux facilitates hydrogen removal from the liquid. If the substrate or product is thermally unstable at normal reflux temperature, the temperature can be lowered by reducing the pressure. Liquid-phase dehydrogenation is used in the production of a number of fine chemicals of limited thermal stability.

Dehydrogenation may be facilitated by the use of a hydrogen acceptor to remove hydrogen from the system as it is formed [23].

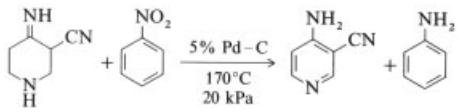

The hydrogen transfer reaction above may be written generally as follows:

$$AH_2 + B \longrightarrow A + BH_2$$

These hydrogenation – dehydrogenation reactions are useful for carrying out hydrogenations in places where hydrogen is unavailable or its use is not permitted; only simple equipment is required [24]. Hydrogen transfer reactions often give better yields of hydrogenated products than are obtained by direct hydrogenation.

3. References

1. R. L. Augustine: *Catalytic Hydrogenation,* Marcel Dekker, New York 1985.
2. M. Freifelder: *Practical Catalytic Hydrogenation,* Wiley-Interscience, New York 1971.
3. P. N. Rylander: *Catalytic Hydrogenation over Platinum Metals,* Academic Press, New York 1967.
4. P. N. Rylander: *Catalytic Hydrogenation in Organic Syntheses,* Academic Press, New York 1979.

5. P. N. Rylander: *Hydrogenation Methods,* Academic Press, London 1985.
6. F. Zymalkowski: *Katalytische Hydrierungen,* Enke Verlag, Stuttgart 1965.
7. G. W. Roberts in P. N. Rylander, H. Greenfield (eds.): *Catalysis in Organic Syntheses,* Academic Press, New York 1976, p. 1.
8. P. A. Chaloner: *Handbook of Coordination Catalysis in Organic Chemistry,* Butterworths, London 1986.
9. P. N. Rylander, *Organic Syntheses with Noble Metal Catalysts,* Academic Press, New York 1973.
10. C. N. Satterfield, *Am. Inst. Chem. Eng.* **21** (1975) no. 2, 209.
11. J. B. Montgomery et al., *Ind. Eng. Chem.* **50** (1958) 313.
12. J. F. VanPeppen et al. in R. L. Augustine (ed.) *Catalysis of Organic Reactions,* Marcel Dekker, New York 1985.
13. L. W. Wright, *Chemtech.* 1974, 42.
14. A. M. Stratz in J. R. Kosak (ed.): *Catalysis of Organic Reactions,* Marcel Dekker, New York 1984.
15. P. N. Rylander et al., *Ann. N.Y. Acad. Sci.* **172** (1970) 266.
16. C. C. Coe, J. W. Brockington in P. N. Rylander, H. Greenfield (eds.): *Catalysis of Organic Reactions,* Marcel Dekker, New York 1987.
17. G. M. Cor, M. VandeMoesdijk in J. R. Kosak (ed.): *Catalysis of Organic Reactions,* Marcel Dekker, New York 1984.
18. F. S. Dovell, H. Greenfield, *J. Am. Chem. Soc.* **87** (1965) 2767.
19. C. L. Thomas: *Catalytic Processes and Proven Catalysts,* Academic Press, New York 1970.
20. E. H. Lee, *Catal. Rev.-Sci. Eng.* **8** (1973) 284.
21. K. K. Kearby in P. H. Emmett (ed.): *Catalysis,* vol. 3, Reinhold Publishing Corp., New York 1955.
22. J. H. Sinfelt in J. R. Anderson, M. Boudart (eds.): *Catalysis Science and Technology,* vol. 1, Springer-Verlag, New York 1981.
23. J. F. Marschik, P. N. Rylander, US 3 517 021, 1970.
24. G. Brieger, T. J. Nestrick, *Chem. Rev.* **74** (1974) 567.

Hydrogen Bromide → Bromine Compounds
Hydrogen Chloride → Hydrochloric Acid
Hydrogen Cyanide → Cyano Compounds, Inorganic
Hydrogen Economy → Hydrogen
Hydrogen Fluoride → Fluorine Compounds, Inorganic
Hydrogen Iodide → Iodine and Iodine Compounds

Hydroquinone

PHILLIP M. HUDNALL, Tennessee Eastman Company, Kingsport, Tennessee 37662, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 3
4.1. **Hydroperoxidation of *p*-Diisopropylbenzene** 3
4.2. **Hydroxylation of Phenol** 4
4.3. **Oxidation of Aniline** 4
4.4. **Other Processes** 4
5. **Environmental Protection** 5
6. **Quality Specifications and Analysis** . 5
7. **Uses** 5
8. **Toxicology and Occupational Health** . 6
9. **References** 7

1. Introduction

Hydroquinone (**1**) [*123-31-9*] (1,4-benzenediol, 1,4-dihydroxybenzene, *p*-dihydroxybenzene, hydroquinol, quinol), $C_6H_6O_2$, M_r 110.11, first described in 1844 by WOEHLER, was obtained by the addition of hydrogen to 1,4-benzoquinone.

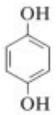

1

Hydroquinone occurs naturally as hydroquinone β-D-glucopyranoside (arbutin) in the leaves of several plants, including bearberry, cranberry, cowberry, and some varieties of pear [1], usually accompanied by its methyl ether, methylarbutin. Arbutin is easily hydrolyzed to hydroquinone and glucose in hot, dilute aqueous acid; hydrolysis of methylarbutin produces hydroquinone monomethyl ether and glucose.

2. Physical Properties

Hydroquinone is a colorless crystalline solid when pure. Three crystalline modifications exist: the stable α-form, *mp* 173.8 – 174.8 °C, is obtained as hexagonal needles by crystallization from water; the labile γ-form, *mp* 169 °C, is obtained as monoclinic prisms by sublimation; the labile β-modification is obtained as needles or prisms by crystallization from methanol or isopropyl alcohol.

Commercial grades of hydroquinone are typically white to off-white crystalline materials; solutions of hydroquinone are discolored by oxidation on contact with air. Physical constants and solubility data are listed in Table 1. Various spectra of hydroquinone are readily available: IR (prism) [2, a], IR (grating) [2, b], ^{13}C NMR [2, c], proton NMR [2, d], UV [2, e], and FTIR [3].

Table 1. Physical properties of hydroquinone

mp, °C	172
bp (101.3 kPa), °C	285
d_4^{15}	1.332
Flash point (closed cup), °C	165
Autoignition temperature (closed cup), °C	516
Flammability limits, vol %	1.6 – 15.3
Critical temperature, °C	549
Critical pressure, MPa	7.45
Critical volume, m^3/mol	3.0×10^{-4}
Liquid molar volume, m^3/mol	1.226×10^{-4}
Heat of formation, kJ/mol	−261.71
Gibbs free energy of formation, kJ/mol	−176.13
Absolute entropy, J mol^{-1} K^{-1}	344.17
Heat of fusion at melting point, kJ/mol	27.1
Standard net heat of combustion, kJ/mol	-2.74×10^3
Dipole moment, *D*	1.40
Van der Waals volume, m^3/mol	5.998×10^{-5}
Van der Waals area, m^2/mol	8.46×10^5
Dissociation constant (30 °C),	
K_1	1.22×10^{-10}
K_2	9.28×10^{-13}
Vapor pressure (132.4 °C), kPa	0.133
Solubility parameter, $(J/m^3)^{0.5}$	2.3753×10^4
Solubility ∗, g per 100 g solvent (30 °C)	
Ethanol	46.4
Acetone	28.4
Water	8.3
Benzene	0.06
Tetrachloromethane	0.01

∗ Reference [4].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_499

3. Chemical Properties

Hydroquinone is easily converted by most oxidizing agents to *p*-benzoquinone [*106-51-4*] (→ Benzoquinone); even neutral aqueous solutions of hydroquinone darken on exposure to air. The rate of oxidation of hydroquinone by air is accelerated in alkaline solution. The oxidation product can add water to give 1,2,4-benzenetriol; subsequent oxidation may result in the formation of humic acids.

The redox potential E_0 of hydroquinone is 699 mV [5]; the half-wave potential $E_{1/2}$ is 560 mV at pH 0 [5] and 234 mV at pH 5 – 6 [6]. Oxidation of hydroquinone involves loss of an electron and elimination of a proton to give the relatively stable semibenzoquinone radical (**2**). The ability of hydroquinone to act as an antioxidant is a consequence of the formation and stability of this radical. A second one-electron transfer and proton elimination results in the formation of *p*-benzoquinone (**3**).

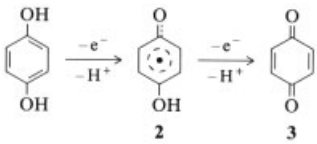

Hydroquinone and *p*-benzoquinone form the equimolar charge-transfer complex quinhydrone. The quinhydrone complex is typically dark greenish-black, *mp* 171 °C.

Hydroquinone is a reducing agent, whose reduction potential is suitable for reducing silver halide. Grains of silver halide that have been exposed to light are reduced much faster than unexposed or underexposed grains. This property is the basis for the use of hydroquinone as a developing agent in photography (→ Photography). The selectivity is a result of changes in the exposed silver halide crystals, which reduce the energy barrier for reaction with hydroquinone to form elemental silver and *p*-benzoquinone. Sodium sulfite is added to the developing solution to prevent formation of quinhydrone by converting the *p*-benzoquinone to sodium 2,5-dihydroxybenzenesulfonate (**4**), a weaker developing agent than hydroquinone.

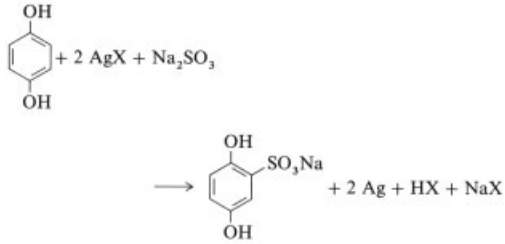

4

The reactivity of hydroquinone is generally similar to that of phenol (→ Phenol). One or both hydroxyl groups may be converted to an ether or ester. Commercially important oxygen derivatives include hydroquinone monomethyl ether (**5**) [*150-76-5*], hydroquinone dimethyl ether (**6**) [*150-78-7*], and hydroquinone bis(2-hydroxyethyl) ether (**7**) [*104-38-1*].

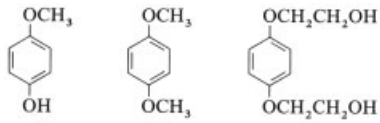

5 **6** **7**

Hydroquinone esters undergo the Fries rearrangement to give acyl-substituted hydroquinones.

Hydroquinone and its ethers may be C-alkylated under Friedel – Crafts conditions to produce a variety of mono- and disubstituted products. Among the commercially important alkylated derivatives are 2- *tert*-butylhydroquinone (**8**) [*1948-33-0*], 2,5-bis[2-(2-methylbutyl)]hydroquinone (**9**), and 2-*tert*-butyl-4-methoxyphenol (BHA) (**10**) [*121-00-6*].

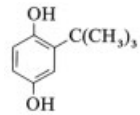

8

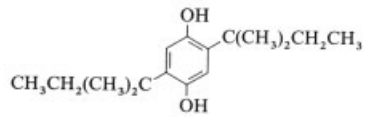

9

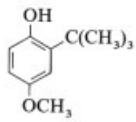

10

Hydroquinone reacts with alkylamines or arylamines to form substituted arylamines. Commercially important products include *p*-*N*-methylaminophenol (**11**) [*150-75-4*], usually marketed as the sulfate salt [*55-55-0*], and *N,N′*-diphenyl-*p*-phenylenediamine (**12**) [*74-31-7*].

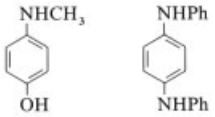

11 **12**

Hydroquinone may be chlorinated with chlorine or sulfuryl chloride to provide derivatives that range from mono- to tetrachlorinated. Kolbe–Schmitt carboxylation of hydroquinone with carbon dioxide leads to 2,5-dihydroxybenzoic acid (gentisic acid). Quinizarin [*81-64-1*] (1,4-dihydroxyanthraquinone) can be prepared by the condensation of hydroquinone with phthalic anhydride (see also → Anthraquinone Dyes and Intermediates, Chap. 2.6.1.1.). Catalytic hydrogenation of hydroquinone gives 1,4-cyclohexanediol (quinitol). Sulfonation of hydroquinone results in the formation of hydroquinone mono- or disulfonic acid, usually isolated as the corresponding potassium salt.

4. Production

Table 2. Production of hydroquinone

Process	Capacity, t/a	Major producers/location
Oxidation of *p*-diisopropylbenzene	25 000	Eastman Chemical Products/United States Goodyear/United States Mitsui/Japan
Hydroxylation of phenol	14 000	Rhône-Poulenc/France Enichem/Italy Ube/Japan
Oxidation of aniline	4 000	People's Republic of China
Total world capacity	43 000	

Three processes are used for the industrial production of hydroquinone: hydroperoxidation of *p*-diisopropylbenzene [*100-18-5*], hydroxylation of phenol [*108-95-2*], and oxidation of aniline [*62-52-3*]. The processes, production capacities, and major producers are listed in Table 2. Although other processes are known, they are not industrially important.

4.1. Hydroperoxidation of *p*-Diisopropylbenzene

p-Diisopropylbenzene (*p*-DIPB) is produced by Friedel–Crafts alkylation of benzene with propene (→ Acylation and Alkylation). Purified *p*-DIPB (**13**) is subsequently converted to the dihydroperoxide (DHP) by air oxidation under slightly alkaline conditions at 80 – 90 °C. The DHP (**14**) is separated from the reaction mixture

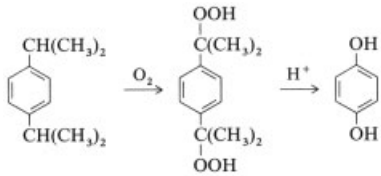

13 **14** **1**

either by extraction or by crystallization, and is then cleaved to hydroquinone (**1**) and acetone by an acid-catalyzed Hock rearrangement; the DHP solution is treated with sulfuric acid catalyst (0.2 – 1.0 %) at 60 – 80 °C. The hydroquinone is crystallized and isolated. The overall yield of (**1**) (based on *p*-DIPB) is ca. 80 %.

4.2. Hydroxylation of Phenol

The catalyzed hydroxylation of phenol at ca. 80 °C with 70 % aqueous hydrogen peroxide produces a mixture of hydroquinone and catechol (**15**). The catalyst may be a strong mineral acid, iron(II), or a cobalt(II) salt. Depending on catalyst selection, the ratio of catechol to hydroquinone can be varied from 3 : 1 to 0.1 : 1; in practice, the ratio is typically 1.5 : 1, i.e., catechol is the major product. The reaction proceeds by an ionic mechanism in which hydrogen peroxide is polarized by the strong acid catalyst and phenol is subsequently hydroxylated the resulting isomers are separated by a series of extractions and solvent-stripping operations.

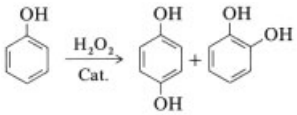

15

The ratio of the products hydroquinone and catechol may be influenced by the presence of superacids or shape-selective zeolites. Thus, use of a vanadium–modified Nafion perfluorosulfonate polymer for the oxidation of phenol gave a hydroquinone–catechol ratio of 12.5 : 1 [7]. The use of a shape–selective zeolite in the hydroxylation of phenol has yielded hydroquinone with 99 % selectivity [8].

4.3. Oxidation of Aniline

Oxidation of aniline [*62-53-3*] is the oldest process used for the production of hydroquinone. Aniline is oxidized with manganese dioxide (15 – 20 % excess) in aqueous sulfuric acid at 0 – 5 °C to produce *p*-benzoquinone (**3**). This intermediate is removed from the reaction mixture by steam stripping and collected. A byproduct, manganese sulfate, may be retrieved from the depleted reaction mixture and sold for agricultural applications.

Hydroquinone is obtained from the intermediate *p*-benzoquinone by reduction with iron at 55 – 65 °C or by catalytic hydrogenation. The product (usually technical grade) is crystallized, isolated from the typically aqueous stream by centrifugation, and dried in a vacuum dryer. The overall yield of hydroquinone from aniline is ca. 85 %.

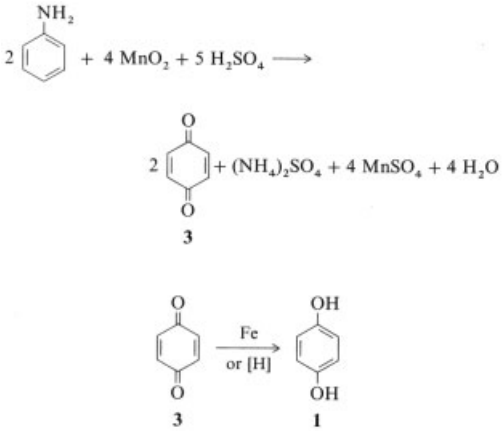

Aniline oxidation is a batchwise process and, therefore, relatively labor intensive. The use of ground manganese ore and finely divided iron leads to extremely abrasive processing conditions; a great deal of maintenance is usually required. Disposal of the inorganic coproducts (amounting to ca. 85 % of the total weight of products) is of environmental concern.

4.4. Other Processes

Carboxylate esters of aromatic diols can be prepared by Baeyer–Villiger oxidation of a 4-hydroxy-substituted aromatic ketone with 97 % conversion and 97 % efficiency [9]. Microbiological oxidation of either benzene or phenol produces hydroquinone with very high selectivity [10]. Copper-catalyzed air oxidation of phenol provides *p*-benzoquinone in greater than 90 % selectivity [11]. Benzene has been oxidized to hydroquinone in the presence of copper(I) chloride [12] or titanium [13]. Oxidation of benzene in aqueous solution with ozone gives *p*-benzoquinone and hydroquinone [14].

Other preparative methods include the reaction of *p*-isopropenylphenol with 30 % aqueous hydrogen peroxide under acidic conditions [15]; *p*-isopropenylphenol may be obtained from bisphenol A [*80-05-7*] by alkaline cracking. Hydroquinone can also be produced by the carbonylation of acetylene [16], [17]; catalytic hydrogenation of nitrobenzene in acidic solution [18]; acid hydrolysis of nitrobenzene [19] or *p*-nitrosophenol [20]; and electrochemical oxidation of benzene or phenol in dilute sulfuric acid, followed by reduction [21], [22].

5. Environmental Protection

An estimate of environmental effects can be obtained from the data available [23–29]. Hydroquinone has a high BOD and can cause oxygen depletion in aqueous systems. It has high potential for affecting some aquatic organisms and moderate potential for affecting microorganisms used in secondary waste treatment. Hydroquinone has a moderate potential to affect the germination or early growth of some plants.

It is readily biodegraded and has a low tendency to concentrate in biological systems. Direct, instantaneous discharge of hydroquinone to a large body of water at a final concentration of 0.005 mg/L or less, followed by secondary waste treatment, is not expected to cause adverse environmental effects.

6. Quality Specifications and Analysis

Five grades of hydroquinone are available: photographic, technical, U.S.P., inhibitor, and polyester. The most important grades (in terms of volume) are photographic and technical. The quality requirements for photographic-grade hydroquinone are defined in American National Standards Institute (ANSI) Specification PH 4.126-1972. Manufacturers specifications for photographic and technical grades are given in Table 3. Specifications for U.S.P.-grade hydroquinone, used in pharmaceutical applications, are described in U.S.P.-XXI [30]. Inhibitor and polyester grades are used in special applications, and the specifications for each grade are provided by the producer. Hydroquinone may be analyzed by a variety of techniques, including spectroscopic, chromatographic, titrimetric, and electrochemical methods.

Table 3. Specifications of hydroquinone

Property	Specifications	
	Photographic [32]	Technical [31]
Appearance	free-flowing white crystals	white to offwhite crystals
mp, °C (min.)	171	169
Assay, wt % (min.)	99.4	99.0
Ash content, wt % (max.)		0.04
pH (5 % aqueous solution)	4.1 – 4.7	
Iron, ppm (max.)	10	
Water, % (max.)	0.6	1.0
Resorcinol, ppm (max.)	50	
Solubility in water		complete
Molar absorptivity ∗ (max).	0.07 (700)	
mL g^{-1} cm^{-1} (λ_{max}, nm)	0.12 (520)	
	0.19 (455)	1.0 (455)

∗ In 0.1 M aqueous hydrochloric acid.

7. Uses

Hydroquinone and its derivatives are used in photographic applications, the rubber industry, monomer inhibitors, dyes and pigments, antioxidants, agricultural chemicals, and other diverse and special applications. Demand for hydroquinone and its derivatives is shown in Table 4, according to market segment. Consumption (in tonnes) of hydroquinone and derivatives by geographical area is as follows:

United States	13 100
Europe	9 600
Japan	3 800
Other	4 800
Total	31 300

The largest demand for hydroquinone is as a photographic developer, principally for black-and-white film, lithography, photochemical machining, microfilm, and X-ray film (→ Imaging Technology, → Photography). Many derivatives of hydroquinone are used in photographic applications, e.g., the sulfate salt of *p*-*N*-methylaminophenol (**11**) and potassium 2,5-dihydroxybenzenesulfonate.

Table 4. Worldwide demand for hydroquinone and its derivatives by market segment (1987)

Market segment	Demand	
	t	% of total
Photography	11 600	37
Rubber industry	8 100	26
Monomer inhibitors	3 100	10
Dyes and pigments	2 800	9
Antioxidants	650	2
Agricultural chemicals	650	2
Others	4 400	14
Total	31 300	100

The second largest consumer of hydroquinone is the rubber industry, which requires hydroquinone for the production of antioxidants and antiozonants (→ Antioxidants). Hydroquinone derivatives used in this area include *N,N′*-diaryl-*p*-phenylenediamines (e.g., *N,N′*-diphenyl-*p*-phenylenediamine), dialkylated hydroquinones, *N*-alkyl-*p*-aminophenols, dialkyl-*p*-phenylenediamines, and aralkyl-*p*-phenylenediamines.

Hydroquinone, hydroquinone monomethyl ether, and *p*-benzoquinone are used extensively in the vinyl monomer industry to inhibit free-radical polymerization during both processing and storage. Hydroquinone, *p*-benzoquinone, 2-methylhydroquinone (toluhydroquinone), 2-*tert*-butylhydroquinone, and 2,5-di-*tert*-butylhydroquinone are used as stabilizers for unsaturated polyester resins. Food-grade antioxidants include 2-*tert*-butylhydroquinone (**8**) and 2-*tert*-butyl-4-methoxyphenol (**10**) (→ Antioxidants, Chap. 5.1.).

Hydroquinone dimethyl ether (**6**) is used as a starting material for a family of dyes and pigments based on the 2-amino- and 2-amino-5-chloro derivatives. Quinizarin is an intermediate for textile dyes (→ Anthraquinone Dyes and Intermediates). The fungicide Chloroneb is produced from (**6**) as raw material. The herbicide ethofumesate is produced from *p*-benzoquinone. Fluazifop-butyl [33] is representative of a new family of herbicides based on the *o*-alkylation of hydroquinone with 2-halopropionic acid derivatives [34–37].

Hydroquinone bis(2-hydroxyethyl) ether (**7**), the product of the reaction of hydroquinone with ethylene oxide, is used as a chain extender in thermosetting urethane polymers [38].

Hydroquinone (U.S.P. grade) and several derivatives are used in topical formulations as skin bleaching and depigmenting agents.

Several new applications for hydroquinone and its derivatives are emerging. The oxygen-scavenging properties of hydroquinone are being exploited for use in boiler water treatment [39]. Hydroquinone and certain C-alkylated or Carylated derivatives are useful monomers for the preparation of a variety of polymers, including liquid crystal polyesters for high-performance plastics, composites, and fibers. In addition to high tensile and impact strengths, these materials exhibit good weatherability, solvent resistance, flame retardance, transparency to microwave radiation, and retention of strength at elevated temperature [40], [41].

8. Toxicology and Occupational Health

Acute toxicity data are shown in Table 5. In a two-year study of rats receiving a diet containing up to 1 % hydroquinone, no effects were observed on final body weight, hematology, or pathology; at 5 % in the diet, a 46 % weight loss together with toxic effects on the bone marrow and liver was found within nine weeks [43]. When applied to the skin of mice, hydroquinone was not carcinogenic and did not promote the carcinogenic activity of benzo[*a*]pyrene [44]. In a skin-painting study in rats, croton oil failed to promote any carcinogenicity from hydroquinone [45]. Hydroquinone produced a negative response in the *Salmonella* bacterial mutagenesis assay [46]. Positive results were observed in the micronucleus test [47] and in an in vitro test for sister chromosome exchanges in human lymphocytes [48].

Table 5. Toxicity data for hydroquinone [23]

Test	Species	Result
Acute oral LD_{50}	rat	400 mg/kg
Acute oral LD_{50}	mouse	100 – 200 mg/kg
Dermal LD_{50}	guinea pig	> 1000 mg/kg
Skin irritation	guinea pig	slight
Eye irritation	rabbit	moderate erythema, clearing by day 14
Sensitization (guinea pig)		
Freund's adjuvant		3/10 slight
Modified Buhler		7/15 low, 2/10 moderate
Maximization [42]		70 % sensitized, strong

Maternal toxicity and fetotoxicity were observed in pregnant rats given 300 mg/kg of hydroquinone daily, but no compound-related teratogenic effects occurred; no adverse effects were observed at doses up to 100 mg/kg per day [23]. Rats given a single dose of 200 mg/kg of (^{14}C)hydroquinone excreted 91.9 % in the urine within 2 – 4 d, 3.8 % in the feces, and 0.39 % in expired air, with 1.25 % remaining in the body; radiolabeled compounds identified in the urine included hydroquinone monoglucuronide (56 – 66 %), hydroquinone monosulfate (25 – 42 %), and hydroquinone (1.1 – 8.6 %) [49]. The percutaneous absorption rate of (^{14}C)hydroquinone in beagles was ca. 1.1 $\mu g\,cm^{-2}\,h^{-1}$, indicating that hydroquinone is poorly absorbed through the skin [23].

The OSHA Permitted Exposure Limit (PEL) and ACGIH Threshold Limit Value (TLV) for hydroquinone is 2 mg/m^3 TWA. Effects of hu-

man exposure are well documented (see below). Direct contact of hydroquinone dust and quinone vapor with the eye may cause irritation. Chronic exposure at high levels has caused brownish discoloration of the cornea and conjunctiva, as well as distortion of the cornea, leading in some cases to decreased visual acuity and even blindness. Exposure to levels below the TLV of 2 mg/m^3 has not produced these effects [50].

A low incidence of allergic contact dermatitis has been reported from handling large amounts of hydroquinone, where exposure to dust occurred, and from repeated contact with alkaline photographic developer solutions [23]. Skin sensitization was noted in 2 members of a group of 48 middle-aged males who used a depigmentation ointment containing 5 % hydroquinone; 38 subjects treated with a 2 % ointment and 26 subjects treated with a 3 % ointment showed no sensitization [51], [52].

In reported poisonings, the lethal dosage for adults is 5 – 12 g (70 – 170 mg/kg). Lower doses cause no adverse effects: two men ingested 500 mg/d for five months and 17 men and women ingested 300 mg/d for three to five months without abnormal symptoms or any abnormalities in blood and urine [43].

9. References

1. *Beilstein,* VI$_3$, 4374.
2. *Sadtler Standard Spectra,* Sadtler Research Laboratories, Division of Bio-Rad Laboratories, a. Infrared Prism Spectrum nos. 153, 13206; b. Infrared Grating Spectrum no. 47; c. Carbon-13 NMR Spectrum no. 38; d. Proton NMR Spectrum no. 10350; e. Ultraviolet Spectrum no. 60.
3. C. J. Pouchert (ed.), *The Aldrich Library of FTIR Spectra,* Aldrich Chemical Company, Milwaukee, Wisconsin 1984, vol. 1, no. 1107D.
4. A. Seidell: *Solubilities of Organic Compounds,* 3rd ed., **vol. 2,** Van Nostrand Publishing, New York 1941, p. 396.
5. H. Musso, H. Doepp, *Chem. Ber.* **100** (1967) 3267.
6. P. J. Elving, A. F. Krivis, *Anal. Chem.* **30** (1958) 1645.
7. FMC, EP-A 132 783, 1985 (R. A. Bull).
8. Mobil Oil, US 4 578 521, 1986 (C. D. Chang, S. D. Hellring).
9. Celanese, EP-A 178 929, 1986; US-A 661 552, 1984 (H. R. Gerberich).
10. Agency of Industrial Sciences & Technology, JP-Kokai 8 678 394, 1986 (A. Yoshikawa).
11. J. E. Lyons, C-Y. Hsu, Biological and Inorganic Copper Chemistry, in K. D. Karlin, J. Zubieta (eds.): *Proc. Conf. Copper Coord. Chem. (July 23 – 27, 1984),* **vol. 2,** Adenine Press, Guilderland, New York 1986, pp. 57 – 76.
12. J. Van Gent, A. A. Wismeijer, A. W. P. G. Peters-Rit, H. Van Bekkum, *Tetrahedron Lett.* **27** (1986) 1059 – 1062.
13. Mitsui Petrochemicals Industries, JP-Kokai 859 889, 1985.
14. C. H. Kuo, H. S. Soong, *Chem. Eng. J. (Lausanne)* **28** (1984) 163 – 171.
15. Upjohn, DE 2 214 971, 1972, and GB 1 344 602, 1974 (P. S. Charleton).
16. Ajinomoto, US 3 420 895, 1969 (H. Wakamatsu).
17. Lonza, GB 1 215 568, 1970 (P. Pino).
18. Koppers, US 3 953 509, 1974 (N. P. Greco).
19. Mitsui Toatsu, JP-Kokai 77 102 235, 1976 (F. Matsuda).
20. Upjohn, US 3 676 503, 1972 (W. v. E. Doering, W. J. Farrisey, Jr.).
21. M. Ya. Fioshin, *Elektrokhimiya* **13** (1977) 381.
22. M. Fremery, H. Hoever, G. Schwarzlose, *Chem. Ing. Technol.* **46** (1974) 635.
23. Unpublished results, Health & Environment Laboratories, Eastman Kodak, Rochester, New York.
24. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., Van Nostrand Reinhold, New York 1983, pp. 746 – 748.
25. *Environmental Effects of Photoprocessing Chemicals,* **vols. 1 and 2,** National Association of Photographic Manufacturers, and Hydroscience, Harrison, New York 1974.
26. G. M. De Graeve, D. L. Geiger, J. S. Meyer, H. L. Bergman, *Arch. Environ. Contam. Toxicol.* **9** (1980) 557.
27. J. E. McKee, H. W. Wolf (eds.): *Water Quality Criteria,* State of California, 1963, Publication no. 3-A.
28. I. Juhnke, D. Luedemann, *Z. Wasser Abwasser Forsch.* **11** (1978) 161.
29. A. J. Leo, C. Hansch (eds.): *Chemical Parameter Data Base,* Medicinal Chemistry Project, Pomona College, Seaver Chemistry Laboratory, Claremont, Calif. 1985.
30. The United States Pharmacopeia XXI, The United States Pharmacopeial Convention, Rockville, Maryland, 1985.

31. *Hydroquinone, Technical Grade, Sales Specification* no. 3503-9, Tennessee Eastman Company, Kingsport, Tenn.
32. *Hydroquinone, Photographic Grade, Sales Specification* no. 4305-1, Tennessee Eastman Company, Kingsport, Tenn.
33. Ishihara Sangyo Kaisha, DE 2 812 571, 1979 (R. Nishiyama et al.).
34. Dow Chemical Co., DD 217 694, 1985 (H. Johnson, L. H. Troxell).
35. Dow Chemical Co., EP-A 97 460, 1984 (H. Johnson, L. H. Toxell).
36. Bayer AG, DE 3 219 789, 1983 (H. Foerster et al.).
37. Zoecon Corp., US 4 529 438, 1985 (S. Lee).
38. Farbenfabriken Bayer Aktiengesellschaft and Mobay Chemical Company, US 3 016 364, 1962 (E. Muller).
39. S. Romaine, I. J. Cotton, *Proceedings American Power Conference* **48** (1986) 1066.
40. D. G. Baird, *Fiber Producer* (October1984) 66.
41. A. S. Wood, *Modern Plastics* (April1985) 78.
42. B. F. J. Goodwin, R. W. R. Crevel, A. W. Johnson, *Contact Dermatitis* **7** (1981) 248.
43. A. J. Carlson, N. R. Brewer, *Proc. Soc. Exp. Biol. Med.* **84** (1953) 684.
44. B. L. van Duuren, B. M. Goldschmidt, *J. Nat. Cancer Inst.* **56** (1976) 1237.
45. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* **vol. 15,** International Agency for Research on Cancer (IARC), Lyon, France 1977, pp. 155 – 175.
46. I. Florin, L. Rutberg, M. Curvall, C. R. Enzell, *Toxicology* **15** (1980) 219.
47. E. Gocke, M.-T. King, K. Eckhardt, D. Wild, *Mutat. Res.* **90** (1981) 91.
48. K. Morimoto, S. Wolff, A. Koizumi, *Mutat. Res.* **119** (1983) 355.
49. G. D. Divincenzo, M. L. Hamilton, R. C. Reynolds, D. A. Ziegler, *Toxicology* **33** (1984) 9.
50. F. L. Oglesby, R. L. Raleigh: *Eye Injury Associated with Exposure to Hydroquinone and Quinone,* 2nd ed., Laboratory of Industrial Medicine, Tennessee Eastman Co., Kingsport, Tenn. 1973, p. 2.
51. *NIOSH Criteria for a Recommended Standard for Hydroquinone,* National Institute for Occupational Safety & Health (NIOSH), U.S. Government Printing Office, Wahington, D.C., 1978.
52. M. C. Spencer, *J. Am. Med. Assoc.* **194** (1965) 962.

Hydroxyanthraquinones → Anthraquinone Dyes and Intermediates
Hydroxybenzaldehydes → Benzaldehyde

Hydroxycarboxylic Acids, Aliphatic

Butyrolactone, Citric Acid, Lactic Acid, and Tartaric Acid are separate keywords.

KARLHEINZ MILTENBERGER, Hoechst Aktiengesellschaft, Gersthofen, Federal Republic of Germany

1. Introduction 1
2. General Characteristics 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
3. Preparation 4
4. Analysis 6
5. Specific Aliphatic Hydroxycarboxylic Acids of Commercial Significance 7
5.1. Glycolic Acid (Hydroxyacetic Acid, 2-Hydroxyethanoic Acid) 7
5.1.1. Preparation 7
5.1.2. Uses 7
5.1.3. Derivatives 8
5.2. Hydroxypropionic Acids and β-Propiolactone 8
5.3. Hydroxybutyric Acids 9
5.4. (*R,S*)-Malic Acid [(*R,S*)-Hydroxysuccinic Acid, (*R,S*)-Hydroxybutanedioic Acid] . . 10
6. Toxicology 10
7. References 11

1. Introduction

Hydroxycarboxylic acids are widely distributed in nature, playing an important role as metabolic intermediates in both plants and animals. Examples can be found in the tricarboxylic acid cycle leading to the degradation (β-oxidation) of fatty acids, as well as in the fermentation of sugars. Hydroxycarboxylic acids are also excreted in the urine, especially in metabolic disorders. Thus, diabetics are incapable of further metabolizing the β-hydroxybutyric acid and acetoacetic acid derived from butyric acid, so these materials are simply eliminated.

One of the first reports of a naturally occurring hydroxycarboxylic acid occurred in 1780 when SCHEELE described the isolation of (*R*)-(−)-lactic acid from soured milk. The substance is formed during the fermentation of lactose by lactobacilli. SCHEELE also discovered *S*-(−)-malic acid (in 1785), *S*-(+)-tartaric acid (in 1769), and citric acid (in 1784) isolating each as its calcium salt. Another member of the group, glycolic acid, is a constituent of unripe fruit, especially grapes. Related natural substances include tartronic, hydroxystearic, mandelic (see → Hydroxycarboxylic Acids, Aromatic), and numerous other fruit- or sugar-derived acids.

Definition In contrast to the purely aromatic hydroxycarboxylic acids (phenolic carboxylic acids), which contain phenolic hydroxyl groups on the aromatic nucleus and thus display enhanced acidity, aliphatic hydroxycarboxylic acids are characterized by purely aliphatic alcohol hydroxyl groups.

Aliphatic hydroxycarboxylic acids may have one or several hydroxyl groups, together with one or more carboxyl groups. In the simplest case, the monohydroxymonocarboxylic acids, the compounds can be conveniently subdivided into three categories:

1) 2- or α-hydroxycarboxylic acids,
2) 3- or β-hydroxycarboxylic acids, and
3) hydroxycarboxylic acids with the hydroxyl group located in the 4- (i.e., γ-) position, or even more remote from the carboxyl group.

Table 1 lists several commercially important hydroxycarboxylic acids and related compounds, grouped in accordance with the above scheme.

2. General Characteristics

The overall characteristics of a hydroxycarboxylic acid are a function of the number of hydroxyl and carboxylic acid groups present as well as their relative placement.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_507

Table 1. Commercially important hydroxycarboxylic acids and derivatives

Type	Trivial name	CAS registry no.	Formula	Molecular mass, M_r
Monocarboxylic acids				
One −OH group in the α-position	glycolic acid	[*79-14-1*]	$HOCH_2-COOH$	76.03
	n-butyl glycolate	[*7397-62-8*]	$HOCH_2-COO-C_4H_9$	132.16
	(*R,S*)-mandelic acid	[*611-72-3*]	$C_6H_5-CHOH-COOH$	152.15
	(*R,S*)-lactic acid	[*598-82-3*]	$CH_3-CHOH-COOH$	90.08
	(*R,S*)-α-hydroxybutyric acid	[*600-15-7*]	$CH_3CH_2-CHOH-COOH$	104.11
One −OH group in the β-position	hydracrylic acid	[*503-66-2*]	$HOCH_2-CH_2-COOH$	90.08
	β-propiolactone	[*57-57-8*]	$CH_2-C=O$; CH_2-O	72.06
	(*R,S*)-tropic acid	[*552-63-6*]	$HOCH_2-CH(C_6H_5)-COOH$	166.18
	(*R,S*)-β-hydroxybutyric acid	[*625-71-8*]	$CH_3-CHOH-CH_2-COOH$	104.11
One −OH group in the γ-position (or more remote)	γ-hydroxybutyric acid	[*591-81-1*]	$HOCH_2-CH_2-CH_2-COOH$	104.11
	γ-butyrolactone	[*96-48-0*]	CH_2-CH_2-O ; $CH_2-C=O$	86.09
	(*R,S*)-ricinoleic acid	[*141-22-0*]	$CH-CH_2-CHOH-C_6H_{13}$; $CH-(CH_2)_7-COOH$	298.45
Two −OH groups	(*R,S*)-glyceric acid	[*600-19-1*]	$HOCH_2-CHOH-COOH$	
Dicarboxylic acids	tartronic acid	[*80-69-3*]	$HOOC-CHOH-COOH$	120.06
One −OH group	(*R,S*)-malic acid	[*617-48-1*]	$HOOC-CH_2-CHOH-COOH$	134.09
Two −OH groups	(*R,S*)-tartaric acid	[*133-37-9*]	$HOOC-CH(OH)-CH(OH)-COOH$	150.09
	meso-tartaric acid	[*147-73-9*]	$HOOC-CH(OH)-CH(OH)-COOH$	150.09
Tricarboxylic acids	citric acid	[*77-92-9*]	$CH_2(COOH)-C(OH)(COOH)-CH_2(COOH)$	192.13

2.1. Physical Properties

Most naturally occurring hydroxycarboxylic acids are chiral and thus exhibit optical activity. In pure form, they are usually crystallizable. Hydroxycarboxylic acids have exceptionally high boiling points relative to corresponding unsubstituted acids with the same chain length. This is a consequence of association among the hydroxyl groups. For this reason, many of these compounds cannot be distilled without decomposition, even under vacuum.

Table 2 summarizes the physical properties of the commercially important hydroxycarboxylic acids and their derivatives listed in Table 1.

2.2. Chemical Properties

Hydroxycarboxylic acids readily form the expected salts at the carboxyl group. The acidity of α- and β-hydroxycarboxylic acids is enhanced relative to the corresponding unsubstituted acids because of the proximity of the polar hydroxyl group. Compounds in this category may be esterified at either the carboxyl or the hydroxyl function by reaction with other alcohols or acids, respectively. The actual course of such reactions depends once again on the placement of the hydroxyl group relative to the carboxylic acid group.

Both α- and β-hydroxycarboxylic acids readily form esters with other alcohol groups but the latter also have a tendency to eliminate water under the reaction conditions. The γ- and

Table 2. Physical properties of the hydroxycarboxylic acids listed in Table .

Hydroxycarboxylic acid	*mp*, °C	*bp*, °C	Density, d_4^{20}	Refractive index, n_D^{20}	$pK_{A_1}^{25}$, (25 °C)
Glycolic acid	78 – 80	100 (decomp.)	1.49 (at 25 °C)		3.81
n-Butyl glycolate		178 – 186 or 82 – 97 (at 2.7 kPa)	1.015 – 1.023	1.423 – 1.4246	
(*R,S*)-Mandelic acid	118 – 121		1.300		3.37
(*R,S*)-Lactic acid	23 – 33	119 – 123 (at 1.6 – 2 kPa)	1.206 (25 °C)	1.4392	3.86
(*R,S*)-α-Hydroxybutyric acid	43 – 44	138 (at 1.9 kPa) or 260 (decomp.)	1.125		
Hydracrylic acid	syrup	decomp.	1.0474	1.4489	4.51
β-Propiolactone	−33.4	51 (at 1.4 kPa) or 162 (decomp.)	1.1460	1.4135	
(*R,S*)-Tropic acid	118	160 (decomp.)			
(*R,S*)-β-Hydroxybutyric acid	48 – 50 or syrup	94 – 96 (at 0.01 kPa) or 130 (at 1.5 – 1.9 kPa)		1.4424	4.39
γ-Hydroxybutyric acid	−17	decomp.			4.71
γ-Butyrolactone	−43.5	206	1.13	1.4352	
(*R,S*)-Ricinoleic acid	5	230 – 235 (at 1.2 kPa)	0.94 (at 15 °C)	1.46393 (at 45 °C)	
(*R,S*)-Glyceric acid	syrup				
Tartronic acid	156 – 158 (decomp.)				2.31
(*R,S*)-Malic acid	131		1.601		3.4
(*R,S*)-Tartaric acid	205				3.03
meso-Tartaric acid	159 – 160				3.11
Citric acid	153		1.665		3.13

δ-hydroxycarboxylic acids undergo intramolecular esterification to form lactones, this reaction being faster than intermolecular esterification. Compounds in which the hydroxyl group is more remote from the acid moiety are again subject to normal esterification.

α-Hydroxycarboxylic acids dimerize with elimination of water to produce six-membered cyclic diesters called 1,4-dioxane-2,5-diones according to IUPAC. These are commonly referred to as lactides, because the phenomenon was first encountered with lactic acid.

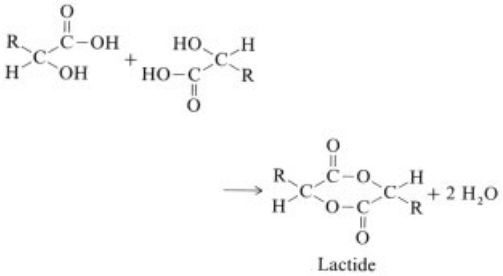

The process may be accompanied by intermolecular esterification, which produces linear polyesters.

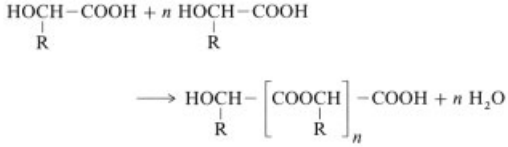

Note that the latter type of esterification, known as *estolide formation,* is not restricted to α-hydroxycarboxylic acids. Another important characteristic of both α- and β-hydroxycarboxylic acids is that the proximity of the two functional groups weakens the intervening C−C bond(s). Thus, treatment of such compounds with sulfuric acid results in the elimination of formic acid, which decomposes in the presence of concentrated sulfuric acid to carbon monoxide and water.

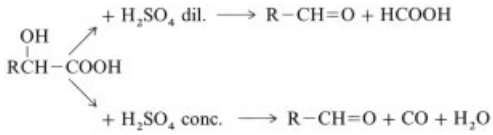

The rest of the chain is converted into aldehydes (in the case of α-hydroxycarboxylic acids) or ketones (in the case of β-hydroxycarboxylic acids) with one fewer carbon atom.

This elimination reaction proceeds so smoothly that it is used for quantitative determination of certain hydroxycarboxylic acids. Thus, lactic acid is converted easily to acetaldehyde, and β-hydroxybutyric acid to acetone. Oxidative cleavage is also possible. For example, treatment of α-hydroxycarboxylic acids with hydrogen peroxide in the presence of iron(II) ions leads to easy elimination of carbon dioxide to produce again aldehydes containing one fewer carbon atom in the chain.

β-Hydroxycarboxylic acids readily undergo intramolecular elimination of water to produce α,β-unsaturated carboxylic acids; e.g.,

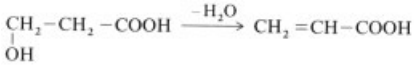

Thus, β-hydroxypropionic acid (hydracrylic acid) can be dehydrated to acrylic acid, and β-hydroxybutyric acid to crotonic acid.

The γ- and δ-*hydroxycarboxylic acids* undergo dehydration even at room temperature, yielding γ- and δ-lactones. The corresponding free acid may be regenerated in aqueous alkaline medium, but often it is stable only in the form of its salts. Lactones are internal esters of hydroxycarboxylic acids.

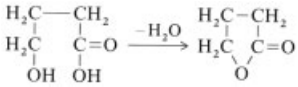

Treatment of a lactone with ammonia produces the corresponding *lactam.*

The related derivatives of α-hydroxycarboxylic acids, the three-membered α-lactones, are unstable and are considered only as reaction intermediates. Four-membered lactones (β-lactones from β-hydroxycarboxylic acids) are known, but special methods are required to prepare them [1], [2]. The exceptional reactivity of these compounds is a consequence of their highly strained rings. For example, they react with ammonia to form amides of the corresponding hydroxycarboxylic acids, which are subject to further conversion to amino acids.

Because of the properties described above, the hydroxyl group of a hydroxycarboxylic acid can be converted into various derivatives. Thus, treatment of lactic acid with hydrogen bromide results in α-bromopropionic acid. Hydrogen iodide can be used to reduce a hydroxycarboxylic acid to the corresponding unsubstituted acid, which is particularly easy with acids from the sugar family.

Saccharic acid lactones are converted easily to polyhydroxyaldehydes (aldoses).

Phosphorus pentachloride leads to replacement of the hydroxyl group in a hydroxycarboxylic acid by chlorine, with the carboxyl group simultaneously converted into an acid chloride moiety. Treatment with *thionyl chloride* results in an unstable sulfurous ester of the acid chloride, which will react, with an alcohol to give a hydroxyester or with ammonia to yield a hydroxyamide. If the hydroxyl group is protected by acetylation, thionyl chloride treatment leads only to the corresponding acid chloride.

3. Preparation

In addition to their isolation from natural sources, hydroxycarboxylic acids can also be obtained by a number of synthetic routes. One general approach involves hydrolysis of a halocarboxylic acid:

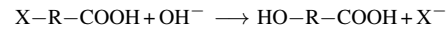

This route is in most cases quite straightforward for the preparation of α-hydroxycarboxylic acids (e.g., glycolic acid), but β- and γ-halocarboxylic acids lead instead to the corresponding lactones.

Certain hydroxycarboxylic acids may also be prepared by acid- or base-catalyzed hydration of an unsaturated carboxylic acid.

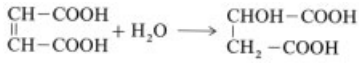

For example, either maleic or fumaric acid can serve as a source of racemic tartaric acid, and acrylic acid can be converted in this way to β-hydroxypropionic acid (hydracrylic acid).

α-Hydroxycarboxylic acids may be prepared by the action of zinc and an alkyl iodide on an oxalate ester. Cyanohydrin synthesis is another synthetic route to α-hydroxycarboxylic acids, one example being the synthesis of (*R,S*)-lactic acid from acetaldehyde via lactonitrile:

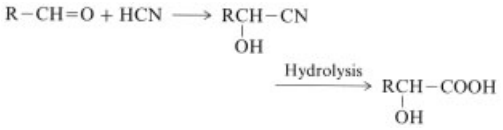

Another interesting preparative method involves diazotization as a means of deaminating an amino acid:

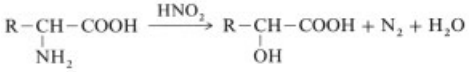

α-Olefins react with liquid dinitrogen tetroxide at 15 – 20 °C to produce nitrate esters of α-hydroxycarboxylic acids, compounds whose electrochemical reactivity is described in [3].

β-Hydroxycarboxylic acids can be prepared from epoxides by treatment with hydrogencyanide, followed by hydrolysis of the intermediate nitriles:

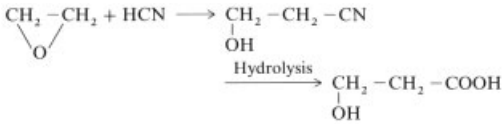

An alternative is the Kolbe nitrile synthesis involving addition of hypochlorous acid to an olefin. This results initially in a chlorohydrin, which can be converted with cyanide ion to a hydroxynitrile, which is then hydrolyzed:

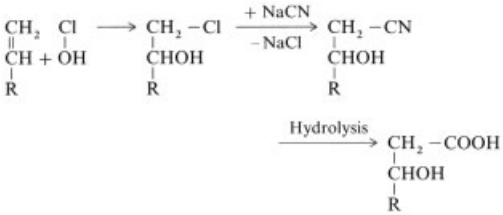

β-Hydroxycarboxylic acids can also be prepared from α-halocarboxylic acid esters by Reformatsky reaction with an aldehyde or ketone in the presence of activated zinc [4]. Another approach utilizes β-lactones, which are readily accessible through ketene treatment of aldehydes or ketones [5]. Thus, acetaldehyde reacts with ketene in the presence of zinc salts to give β-butyrolactone. Finally, esters of β-ketoacids are subject to catalytic hydrogenation, a method that is also applicable to the synthesis of γ- and δ-hydroxycarboxylic acids from the corresponding γ- and δ-ketoesters.

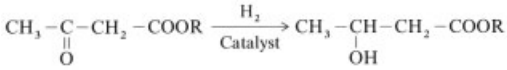

For example, catalytic hydrogenation of acetoacetic esters give the esters of (*R,S*)-β-hydroxybutyric acid.

Both γ- and *δ-hydroxycarboxylic acids* can be prepared by alkaline hydrolysis of the corresponding halo acids. In this case, the products are normally isolated in the form of alkali-metal salts.

Dihydroxycarboxylic acids may be made by reaction of unsaturated fatty acids (e.g., oleic acid) with hydrogen peroxide or peracid, in which epoxides are formed as intermediates. *Polyhydroxy acids* are also commonly prepared from unsaturated carboxylic acids, usually by permanganate oxidation.

Many of the methods described above lend themselves equally well to the synthesis of *polybasic hydroxycarboxylic acids.* For instance, *tartronic acid* has been prepared both by hydrolysis of monobromomalonic acid or by sodium amalgam reduction of mesoxalic acid (ketomalonic acid):

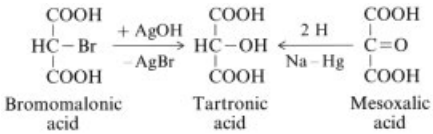

Another approach starts with glycerol, which upon oxidation with potassium permanganate yields tartronic acid, whereas treatment with nitric acid yields glyceric acid.

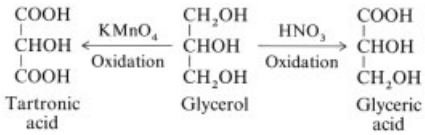

As in this case, careful selection of oxidizing agent often enables various hydroxycarboxylic acids to be obtained by starting with the same precursor, a point well illustrated by the oxidation of sugars (aldoses and ketoses). Mild oxidizing agents such as aqueous bromine or dilute nitric acid attack only the aldehyde functions of

aldoses (→ Carbohydrates), producing glyconic acids (e.g., gluconic acid from glucose). Such glyconic acids (also called "onic acids") are stable only in alkaline solution, where they exist as alkali-metal salts. The free acids rapidly cyclize to γ- or δ-lactones, with the γ-lactones preferred.

More powerful oxidizing agents such as concentrated nitric acid result in polyhydroxy dibasic acids (saccharic acids), which immediately cyclize to lactones. Thus, glucose leads to glucaric acid; mannose to mannaric acid; and galactose to mucic acid (galactaric acid). Even more vigorous oxidation causes cleavage of the carbon chain, yielding lower hydroxydicarboxylic acids.

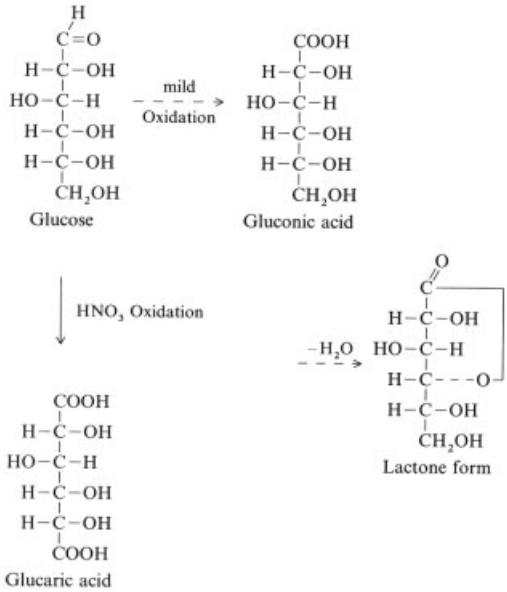

Numerous branched and unbranched hydroxycarboxylic acids may be prepared by cleavage or oxidation (with alkaline permanganate, chromic acid, dilute nitric acid, or concentrated nitric acid) of a variety of starting materials [6–9]. Enzymatic production by fermentation is becoming increasingly important.

4. Analysis

The literature describes an array of color reactions for *qualitative determination* of hydroxycarboxylic acids. These include the yellow color obtained in the presence of iron(II) chloride, characteristic complexation phenomena with copper salts, and the distinctive color produced with ammonium vanadate. Nevertheless, none of these reactions can be regarded as truly specific.

The most satisfactory approach to *quantitative analysis* involves acidimetric titration of carboxyl groups. Because of the nearly universal tendency to lactone formation in equilibrium with the free acid, free and total acids must be determined separately. Thus, lactones are treated in the same way as other esters: Hydrolysis with excess NaOH is followed by back titration with acid. Other recommended methods for quantitative analysis rely on acetylation of hydroxyl groups or esterification of carboxyl groups, with subsequent determination of the resulting water.

Instrumental methods of analysis are playing an increasingly important role, including IR, NMR and MS spectroscopy and, especially, chromatography (e.g., thin-layer chromatography and HPLC). Nevertheless, such methods prove reliable only if they are carefully adapted to suit the specific problem. A typical procedure involves methylation or silylation, followed by identification and quantification with GC – MS [7].

5. Specific Aliphatic Hydroxycarboxylic Acids of Commercial Significance

5.1. Glycolic Acid (Hydroxyacetic Acid, 2-Hydroxyethanoic Acid) [10]

The chemical structure, molecular mass, and CAS registry number of glycolic acid are given in Table 1. Its physical properties are listed in Table 2 and in the following material:

mp:	
α-Modification	78 – 80 °C
β-Modification, metastable	63 °C
Heat of combustion	−697.23 kJ/mol
Heat of solution in water (infinite dilution)	−11.71 kJ/mol
Dissociation constant (at 25 °C)	1.54×10^{-4}
pH (1 M solution)	2.4

Solid glycolic acid forms colorless, monoclinic, prismatic crystals. The acid is freely soluble in water, methanol, ethanol, acetone, and

ethyl acetate. Glycolic acid is only slightly steam volatile and cannot be destilled under vacuum; attempts result in self-esterification with loss of water, leading to di- and polyglycolides. However, newer results show that glycolic acid can be vaporized without decomposition. In the vapor phase it can be dehydrogenated catalytically to glyoxylic acid [11].

5.1.1. Preparation

Glycolic acid is usually produced by hydrolysis of molten monochloroacetic acid with 50 % aqueous sodium hydroxide at 90 – 130 °C. The resulting glycolic acid solution has a concentration of ca. 60 % and contains 12 – 14 % sodium chloride. The salt may be removed by evaporative concentration, followed by extraction of the acid with acetone [12]. Attempts have also been made to conduct the hydrolysis with acid catalysts at 150 – 200 °C with water or steam under pressure [13]. In this case, the byproduct is hydrogen chloride, rather than sodium chloride, which can be removed by distillation. The principal disadvantage of the method is the need for relatively large volumes of water.

Glycolic acid is produced commercially in the United States (Du Pont) by treating formaldehyde or trioxymethylene with carbon monoxide and water in the presence of acid catalysts at >30 MPa [14]. Another process, previously utilized by Degussa, involves electrolytic reduction of oxalic acid [15]. The compound can also be prepared in 90 % yield by hydrolysis of the corresponding nitrile, obtained by reaction of formaldehyde with hydrocyanic acid [16].

5.1.2. Uses

Glycolic acid is available commercially as either a 57 % (Hoechst) or a 70 % (Du Pont) aqueous solution. Total annual consumption worldwide is ca. 2000 – 3000 t of solution.

Glycolic acid is used in textile dyeing, printing, and creaseproofing. The fact that it can form a chelate with calcium(II) ions makes it well-suited to hide deliming in the leather industry, as well as to inclusion in alum and chrome mordants and in fur-processing operations. The compound has little tendency to cause corrosion, and this characteristic, coupled with its bactericidal properties, makes it suitable for incorporation into acidic cleansing agents. It is especially well-adapted to cleansing operations involving milk containers, milk-processing equipment, and drinking fountains, as well as rust and scale removal in heat exchangers and pipelines. Glycolic acid inhibits the growth of iron-oxidizing bacteria. The use of glycolic acid eliminates the need for simultaneous addition of chelating agents and bactericides. The effectiveness of glycolic acid as a complexing agent also contributes to its use in copper polishes, as an etching agent for lithographic plates, and in the preparation of electropolishing and galvanizing baths.

Safety Considerations. According to section 1.4 of the hazardous materials regulation, Gef Stoff V (Federal Republic of Germany), and to the EEC guidelines, glycolic acid solutions are "corrosive."

Analysis. Glycolic acid may be detected qualitatively by the violet color formed with 2,7-dihydroxynaphthalene [17]. The preferred method of quantitative analysis (in the absence of other acidic or hydrolyzable substances) is acidimetric titration. Because of the tendency of lactide formation free and total acid must be determined separately.

5.1.3. Derivatives

The glycolic acid esters *methyl glycolate* [*96-35-5*], *bp* 147 – 149 °C, and *ethyl glycolate* [*623-50-7*], *bp* 158 – 159 °C, both serve as starting materials for the laboratory preparation of pure glycolic acid [18] and were formerly employed as solvents for resins and for nitro- or acetyl cellulose. Apart from these two materials, only carboxymethyl cellulose [*9004-32-4*], (→ Cellulose Ethers, Chap. 6. and *n*-butyl glycolate have commercial significance.

***n*-Butyl Glycolate** (see Table 1).

Physical Properties. *n*-Butyl glycolate is a colorless liquid, which is miscible with most organic solvents. Its physical properties are summarized in Table 2. The solubility in water is limited to 8 wt % (20 °C), although the compound itself may contain up to 25 wt % water.

Production. *n*-Butyl glycolate is produced by treatment of sodium chloroacetate with *n*-butyl alcohol at 125 – 160 °C, followed by vacuum distillation [19].

Applications. *n*-Butyl glycolate is used primarily as a varnish additive, valued because of its low volatility. Trade names include Polysolvan-O (Hoechst) and GB-Ester (Wacker). It confers smooth spreading properties and high gloss on nitrocellulose varnish. For acetyl cellulose, it is effective as a blush inhibitor under conditions of high humidity. Due to its good blending properties, *n*-butyl glycolate is also used as an additive in alkyd resins and oil-based paints.

Safety Considerations. *n*-Butyl glycolate is not regarded as hazardous.

5.2. Hydroxypropionic Acids and β-Propiolactone

Of the two isomeric hydroxypropionic acids, only α-hydroxypropionic acid (→ Lactic Acid) is optically active.

β-Hydroxypropionic Acid (Hydracrylic Acid, 3-Hydroxypropanoic Acid) [20] (see Table 1).

Physical Properties. The most important physical properties of β-hydroxypropionic acid are listed in Table 2); its dissociation constant at 25 °C is 3.11×10^{-5}. β-Hydroxypropionic acid forms a highly acidic syrup, which loses water upon heating.

Preparation. β-Hydroxypropionic acid can be prepared by alkaline hydration of acrylic acid or by treatment of ethylene chlorohydrin with sodium cyanide, followed by hydrolysis of the resulting β-hydroxypropionitrile. It is also readily obtained by hydrolysis of the commercially important and easily accessible compound β-propiolactone.

β-Propiolactone (Oxetan-2-One) [21] (see Table 1).

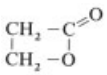

Physical Properties [1]. The physical properties of β-propiolactone are summarized in Table 2 and in the following material:

Standard heat of formation	−330.0 kJ/mol
Standard heat of combustion	−1422.0 kJ/mol

β-Propiolactone is a colorless, highly reactive liquid, that is soluble in water, alcohol, acetone, and chloroform (solubility in water at 25 °C, 37 vol %).

Chemical Properties. β-Propiolactone reacts with alcohols, acid chlorides, ammonia, and water to yield β-substituted propionic acid derivatives. The most important characteristic of the substance is its ability to polymerize. This highly exothermic process occurs simply by warming, although it is also catalyzed by both acid and base.

Preparation. β-Propiolactone is synthesized by passing equimolar amounts of ketene and formaldehyde into either acetone or β-propiolactone itself. The reaction is carried out at low temperature (<20 °C) with a yield of ca. 90 % [5]. Both aluminum chloride and zinc chloride have been employed as catalysts, and the use of methyl borate has also been suggested.

Applications. As late as 1974, β-propiolactone was used in the United States in the preparation of acrylic acid and acrylate esters [22]. Today, its principal significance is as a reactive intermediate in organic syntheses; a small amount is treated with ammonia to provide β-alanine. β-Propiolactone was also used as a disinfectant. It appeared to be an attractive replacement for formaldehyde due to its 25-fold greater disinfecting power, but it has since been abandoned because of its carcinogenic properties.

Safety Considerations. In the EEC hazardous materials guideline and in the Gef Stoff V (Federal Republic of Germany) β-propiolactone is classed as a strong carcinogen and is considered extremely dangerous at concentrations $\geq$1 wt %.

5.3. Hydroxybutyric Acids

Of the three structural isomers α-, β-, and γ-hydroxybutyric acid, only the α- and β-acids are optically active. Racemic mixtures of each can be separated into optical antipodes by way of the corresponding strychnine or brucine salts.

All three acids are soluble in water, ethanol, diethyl ether, and numerous organic solvents, and decompose readily on dry distillation. β-Hydroxybutyric acid easily loses water to give crotonic acid, and γ-hydroxybutyric acid is transformed even at low temperature to γ-butyrolactone.

(*R,S*-)-α-Hydroxybutyric Acid [(*R,S*-)-2-Hydroxybutanoic Acid] [23], (see Table 1). The physical properties of (*R,S*)-α-hydroxybutyric acid are listed in Table 2.

Preparation. (*R,S*)-α-Hydroxybutyric acid results as main product from treating butyraldehyde with alkaline sodium hypochlorite solution or from heating α-bromobutyric acid with formamide at 150 °C.

(*R,S*-)-β-Hydroxybutyric Acid [(*R,S*-)-3-Hydroxybutanoic Acid] [24], (see Table 1). *Physical Properties.* The *mp* of racemic β-hydroxybutyric acid is 48 – 50 °C; it is commonly described in the literature as a syrupy liquid (see Table 2). Pure (*R*)-(−)-β-hydroxybutyric acid [*625-72-9*] is a solid with *mp* 49 – 50 °C.

Preparation. (*R,S*)-β-Hydroxybutyric acid results from hydrolysis of β-butyrolactone, which in turn can be produced by the addition of ketene to acetaldehyde in the presence of boron trifluoride or zinc salts [5]. The compound is also prepared by reduction of acetoacetic acid with sodium amalgam in alkaline solution.

γ-Hydroxybutyric Acid (4-Hydroxybutanoic Acid) [25] (see Table 1).

Physical Properties. The most important physical properties of γ-hydroxybutyric acid are listed in Table 2. The dissociation constant at 25 °C is 1.93×10^{-5}. At ambient temperature, the acid is a liquid and consists in part of lactone.

Preparation. γ-Hydroxybutyric acid is obtained from γ-butyrolactone by alkaline hydrolysis and isolated as the alkali-metal salt. γ-Butyrolactone is produced by dehydrocyclization of 1,4-butanediol.

Applications. Hydroxybutyric acids were utilized in the manufacture of plasticizers, but their current applications are quite limited. The related lactones are some times used as starting materials or intermediates in organic synthesis.

Optically active *β-hydroxybutyric acid* is an important pharmacological agent [26]. Biodegradable *poly*(−(*R*)-(−)-*β-hydroxybutyric acids* (named PHB) have been used in the context of controlled long-term parenteral application of drugs through implantation. These materials offer a number of advantages relative to other implantation polymers [e.g., poly(methacrylic acid) derivatives, poly(glycolic acid), poly(lactic acid)] [27]. Poly(β-hydroxybutyric acid) has also been considered as a potential textile material.

γ-Hydroxybutyric acid has been used as an anesthetic and pain killer. It has a stabilizing effect on blood pressure and induces a state of unconsciousness, characteristics that have encouraged its application especially in geriatrics [28].

5.4. (*R,S*)-Malic Acid [(*R,S*)-Hydroxysuccinic Acid, (*R,S*)-Hydroxybutanedioic Acid] [29], [30](see Table 1.)

Physical Properties (see Table 2). Malic acid is a colorless, odorless, crystalline substance. It is highly soluble in water (100 g of water dissolves 126.3 g of malic acid at 20 °C), methanol, ethanol, acetone, ether, and other polar solvents. Because it is a dibasic acid, it has two dissociation constants: at 20 °C, $K_{A_1} = 3.9\times10^{-4}$ and $K_{A_2} = 1.4\times10^{-5}$.

Preparation. (*R,S*)-Malic acid is prepared commercially in the United States and Canada by hydration of maleic anhydride (→ Maleic and Fumaric Acids) [31]. The sole manufacturer in the United States is Alberta Gas, with an annual capacity of ca. 5000 t. In this process maleic acid is heated at ca. 180 °C (under a pressure of ca. 1 MPa), malic acid is yielded as the main product. Byproducts are maleic and fumaric acids. The latter can be separated by filtration and returned to the process stream because of its low water solubility.

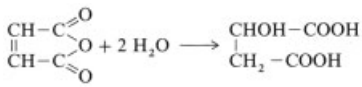

The filtrate is then concentrated; this causes separation of the malic acid, which is purified by multiple washings, evaporation, and recrystallization until the contents of fumaric and maleic

acids are reduced to 7.5 and <500 ppm, respectively. Additional purification is required to prepare pharmacological-grade material [32], [33].

Applications. (*R,S*)-Malic acid closely resembles both citric and tartaric acids in its physical and chemical properties. However, the compound has a more neutral flavor, so it is often preferred when the flavor of citric acid is considered objectionable. Examples include the impregnation of packing material for foodstuffs such as cheese or the acidification required during preparation of baked goods. It is widely used in the food industry as an acidulant and, to a lesser degree, as an acidity regulator. Relative to citric or tartaric acid, the flavor imparted by malic acid is not only more neutral but also of longer duration. Thus, citric acid imparts a strongly acidic taste to the tongue almost immediately, but the effect is brief. With malic acid, the sensation is less intense, but it lasts longer and is more consistent. Soft-drink manufacturers have discovered that this property helps mask the aftertaste associated with certain nonnutritive sweeteners and leads to a more balanced flavor profile. Moreover, synergistic effects are observed between these sweeteners and malic acid, which permit reductions of up to 20 % in the amount of sweetener required and 10 % in malic acid. Consequently, significant economic benefits may be realized. The economic picture is also influenced favorably by the fact that malic acid is available as an anhydrous powder.

Malic acid is used primarily as an ingredient in hard candys and other sweets, jams, jellies, and various canned fruits and vegetables. Most countries authorize its use as an additive in foodstuffs.

(*R*)-(+) *-Malic acid* [*636-61-3*] *and* (*S*)-(−)-*malic acid* [*97-67-6*] are isolated from natural sources by resolution of (*R,S*)-malic acid with the aid of an optically active base. (*S*)-(−)-Malic acid is also available through microbiological fermentation of fumaric acid [33]. The levorotatory *S*-enantiomer is widely distributed in fruit and displays the following characteristics that differ from those of the racemate: *mp*, 100 °C; d_4^{20}, 1.595; α_D^{18}, −2.3° (7 wt % in H_2O). The compound is soluble to the extent of 36.4 g in 100 g of water at 20 °C.

6. Toxicology

As previously noted, most important aliphatic hydroxycarboxylic acids occur as natural products. Because they serve as intermediates in plant and animal metabolic pathways, they do not exhibit many exceptionally toxic properties. Thus, only a few compounds listed in Table 1 are included in the *Register of Toxic Effects of Chemical Substances* prepared by NIOSH [35]. Certain hydroxycarboxylic acids (e.g., lactic, malic, tartaric, and citric) are even accepted as food additives or preservatives in most countries.

The following toxicity data for *citric acid* are taken from the NIOSH list [35, p. 945]:

LD_{50} (oral, rat)	11 700 mg/kg
LD_{50} (oral, mouse)	5 040 mg/kg
LD_{50} (oral, rabbit)	7 000 mg/kg

Glycolic acid was also permitted as a food preservative but later came to be considered suspect [36]. Today, it is described as moderately toxic, with the following data [35, p. 386],

LD_{50} (oral, rat)	950 mg/kg
LD_{50} (oral, guinea pig)	920 mg/kg

A derivative of glycolic acid *n-butyl glycolate* is described as displaying narcotic properties. It is also an irritant to both skin and mucous membranes. Nevertheless, it is markedly less toxic than glycolic acid, having an LD_{50} (oral, rat) of 4595 mg/kg.

The lowest reported lethal dose for *β-hydroxypropionic acid (hydracrylic acid)* derives from intravenous injection into rats: LD_{50} (i.v., rat) = 50 mg/kg [35, p. 447].

A derivative of β-hydroxypropionic acid *β-propiolactone* is of considerable interest because of its viricidal and bactericidal properties, and was used medically in sterilization and disinfection. Nevertheless, it has since been found to be a mutagen [37]. Animal studies have provided incontrovertible evidence that the compound displays carcinogenic properties irrespective of whether it is delivered orally, topically on the skin, or by intravenous, intraperitoneal, or subcutaneous injection [38], [35, p. 1326]. β-Propiolactone is a powerful irritant to the skin and eyes. The following toxicity data have been reported:

LC_{50} (inhalation, rat)	25 ppm, 6 h
LD_{50} (i.v., mouse)	345 mg/kg

Because of its cytotoxic properties, β-propiolactone has been added to Category A 2 of the official list of carcinogenic materials [39].

By contrast, *γ-butyrolactone* is completely free of these cytotoxic characteristics. Despite numerous animal studies, no evidence has been obtained to suggest carcinogenicity. Nonetheless, the compound is moderately toxic, having an LD_{50} (oral, rat) of 1800 mg/kg [35, p. 946]. The narcotic effects observed when γ-butyrolactone is administered orally or intraperitoneally have been ascribed to γ-hydroxybutyric acid, from which it is derived.

From animal studies, the sodium salt of γ-hydroxybutyric acid is known to possess narcotic properties [40]. A dose of ca. 40 mg/kg leads to nearly natural sleep in humans, with essentially no side effects.

7. References

1. H. E. Zaugg: "β-Lactones," *Org. React. (N.Y.)* **8** (1954) 305 ff.
2. A. Griesbeck, D. Seebach, *Helv. Chim. Acta* **70** (1987) 1320, 1326.
3. Jaroslawl. Technological Institute, DE-OS 2 326 297, 1974 (M. S. Rusakowa, et al.).
4. R. L. Shriner: "The Reformatsky Reaction," *Org. React. (N.Y.)* **1** (1942) 1 ff.
5. Goodrich Co., US 2 356 459, 1941 (F. E. Kung); US 2 424 589, 2 424 590, 1944 (T. R. Steadman).
6. Escambia Chem. Corp., US 2 811 545, 1956 (T. R. Steadman).
7. K. Niemela, E. Sjostrom, *Carbohydr. Res.* **144** (1985) no. 1, 87 – 92.
8. K, Niemela, E. Sjostrom, *Holzforschung* **40** (1986) no. 1, 9 – 14.
9. K. Garves in J. F. Kennedy, (ed.): *Cellulose Its Derivatives,* Ellis Horwood, Chichester 1985, pp. 487 – 494.
10. *Beilstein,* **3,** 228; **3 (1),** 88; **3 (2),** 167; **3 (3),** 367; **3 (4),** 571.
11. Hoechst, DE-OS 2 904 754, 1980 (H. Baltes, E. I. Leupold, F. Wunder).
12. Hoechst, DE-OS 2 812 682, 1979 (E. I. Leupold, B. Wojtech, H. J. Arpe).
13. Hoechst, DE-OS 2 810 975, 1979 (H. Klug, H. Woppert, H. Korbanka).
14. DuPont, US 2 152 852, 1939 (D. J. Loder); US 2 153 064, 1939 (A. T. Larson); US 2 443 482, 1948 (M. T. Shattuck).
15. Degussa, DE 194 038, 1903.
16. I. G. Farbenind., DE 459 602, 1925 (O. Schmidt).
17. E. J. E. Eegriwe, *Z. Anal. Chem.* **89** (1932) 124.
18. DuPont, US 2 331 094, 1940 (D. J. Loder).
19. I. G. Farbenind., DE 522 786, 1929 (W. Bülow).
20. *Beilstein,* **3,** 295; **3 (1),** 122; **3 (2),** 212; **3 (3),** 522; **3 (4),** 689.
21. *Beilstein,* **17 (1),** 130; **17/5 (3/4),** 4157; **17/9 (5),** 3.
22. *Ullmann,* 4th ed., **19,** p. 452.
23. *Beilstein,* **3,** 302; **3 (1),** 114; **3 (2),** 216; **3 (3),** 561; **3 (4),** 755.
24. *Beilstein,* **3,** 308; **3 (1),** 116; **3 (2),** 220; **3 (3),** 571; **3 (4),** 761.
25. *Beilstein,* **3,** 311; **3 (2),** 222; **3 (3),** 581; **3 (4),** 774.
26. D. Seebach, M. Züger, *Helv. Chim. Acta* **65** (1982) 495 – 503.
27. W. Korsatko, B. Wabnegg, G. Braunegg, H. M. Tillian et al., *Pharm. Ind.* **45** (1983) no. 5, 525 – 527; no. 10, 1004 – 1007; *Pharm. Ind.* **46** (1984) no. 9, 952 – 954.
28. W. Bushart, P. Rittmeyer: *Anästhesie mit γ-Hydroxybuttersäure,* Springer-Verlag, Berlin 1973.
29. *Beilstein,* **3,** 435; **3 (1),** 154; **3 (2),** 289; **3 (3),** 918; **3 (4),** 1124.
30. H. Rudy: *Fruchtsäuren, Wissenschaft und Technik,* Hüthig-Verlag, Heidelberg 1967.
31. T. Yoshida, O. Hibino, S. Morita, *Nippon Kagaku Kaishi (1921 – 47)* **53** (1950), 399 – 401; *Nippon Chem. Ind. JP* 4 360, 1950.
32. Chemie Linz, FR 2 193 810, 1973.
33. Allied Chem. Corp., DE-AS 1 518 522, 1965 (Winstrom, O. Leon, Ingleman, Miltin Russel).
34. Hüls, DE-OS 3 434 880 A1, 1986; Supplement to DE-OS 3 310 849.
35. Registry of Toxic Effects of Chemical Substances, 1983 – 84. Cumulative Supplement to the 1981 – 82 Edition NIOSH. U.S. Department of Health and Human Services Public Health Service, Centers for Disease Control. National Institute for Occupational Safety and Health.
36. W. F. Riker, G. Gold, *J. Am. Pharm. Assoc. Sci. Ed.* **31** (1942) 306.

37. H. H. Smith, A. M. Srb, "Induction of mutations with β-propiolacton," *Science* **114** (1951) 490.
38. D. J. Brusick, *Mutat. Res.* **39** (1977) no. 3 – 4, 241 –255.
39. D. Henschler, Vors. d. Arbeitsst.-Komm. d. Dtsch. Forschungsgemeinsch.: *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* **vol. IV,** 14. Lieferung, VCH-Verlagsgesellschaft Weinheim, 1988.
40. H. Laborit et al., *JAMA J. Am. med. Assoc.* **175** (1961) 520.

Hydroxycarboxylic Acids, Aromatic

Salicylic Acid is a separate keyword; Naphtholcarboxylic Acids see → Naphthalene Derivatives

EDWIN RITZER, Bayer AG, Leverkusen, Federal Republic of Germany

RUDOLF SUNDERMANN, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. General Properties 2
3. Production 2
4. Individual Aromatic Hydroxycarboxylic Acids 3
4.1. Monohydroxycarboxylic Acids . . . 3
4.1.1. 3-Hydroxybenzoic Acid 3
4.1.2. 4-Hydroxybenzoic Acid 4
4.1.3. Cresotic Acids 5
4.2. Dihydroxybenzoic Acids 6
4.3. Trihydroxybenzoic Acids 6
4.4. Mandelic Acid 7
5. Toxicology 8
6. References 8

1. Introduction

Aromatic hydroxycarboxylic acids (i.e., hydroxycarboxylic acids with phenolic hydroxyl groups) occur widely in nature in the form of their esters, ethers, intermolecular esters (depsides), or cyclic ether – esters (depsidones). Structural formulas of the most important depsides and of depsidone are given below:

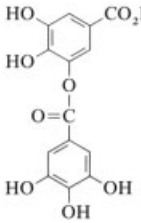

m-Digallic acid
[*536-08-3*]

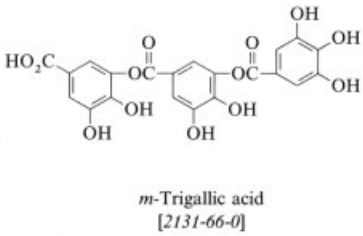

m-Trigallic acid
[*2131-66-0*]

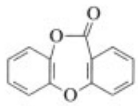

Depsidone
[*3580-77-6*]

Hydrolyzable tanning agents such as tannin (ester of gallic acid, tannic acid) [*5424-20*] are esters of aromatic hydroxycarboxylic acids with polyhydric alcohols and sugars. Aromatic hydroxycarboxylic acids are used mainly for pharmaceutical and cosmetic purposes: as disinfectants, preservatives, emulsifiers, analgesics, and antirheumatic and antipyretic drugs. Their use in the production of dyes and plastics is increasingly important, especially in color developing, and in liquid crystal polymers (→Liquid Crystals).

2. General Properties

Monohydroxybenzoic acids are dibasic acids. Thus, one mole of alkali forms the carboxylate and a second mole converts the phenolic hydroxyl group into a phenolate. The acidity of each group depends on its relative position (see Table 1). The carboxyl and hydroxyl group acidities of 3- and 4-hydroxybenzoic acids are comparable to those of benzoic acid and phenol. However, the corresponding acidities of 2-hydroxybenzoic acid (salicylic acid [*69-72-7*]) are different. The higher acidity of the carboxylate group in this case is the result of intramolecular chelate formation.

Table 1. Dependence of dissociation constants for aromatic hydroxycarboxylic acids on relative position of hydroxyl and carboxyl groups

		K_1(−COOH), $\times 10^{-5}$	K_2(−OH), $\times 10^{-10}$
Benzoic acid	[*65-85-0*]	6.3	
Phenol	[*108-95-2*]		9.9
2-Hydroxybenzoic acid	[*69-72-7*]	105	0.0004
3-Hydroxybenzoic acid	[*99-06-9*]	8.3	1.17
4-Hydroxybenzoic acid	[*99-96-7*]	2.9	4.8

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_519

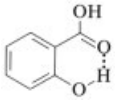

This accounts for its high steam volatility and ease of sublimation. 2-Hydroxy- and 4-hydroxybenzoic acids decarboxylate above their melting point; 2,4- and 2,6-dihydroxybenzoic acids decarboxylate even when heated in an aqueous solution. This process is catalyzed by strong acid. The carboxyl group is esterified with alcohols under acid catalysis without detectable etherization of the phenolic hydroxyl group. Ether – esters can be obtained easily from the dialkali salts by using alkyl halides or dialkyl sulfates. Alkaline hydrolysis of ether – esters produces alkoxycarboxylic acids. Electrophilic substitution (e.g., nitration, sulfonation, chlorination) leads to attack of the aromatic nucleus at the 2- or 4-position, with the 4-position preferred.

Analysis. Methods of analyzing carboxylic acids and phenols are also applicable to hydroxycarboxylic acids. In practice, alkalimetric titration (in special cases, potentiometric titration) is used. With potentiometric determination of the equivalence point, the hydroxyl group can also be determined. Trace analyses are carried out by colorimetric methods with iron(III) chloride [4]. Thin-layer chromatography is a suitable method for checking purity. Here, individual acids are made visible by means of alkaline (Na_2CO_3) coupling with diazotized 4-nitroaniline, by spraying with iron(III) chloride, or by fluorescence in UV light. For quantitative analysis, HPLC methods are suitable. Alternatively, the components can be derivatized and capillary gas chromatographic methods used to separate them [5].

3. Production

Kolbe – Schmitt or Marassé Carboxylation. Yields according to the Kolbe reaction of solid anhydrous alkali phenolates in a stream of carbon dioxide were improved by SCHMITT who carried out the reaction under pressure [6]. The Marassé method involves the conversion of phenol with potassium carbonate and does not work with less expensive sodium carbonate [7]. The Kolbe – Schmitt synthesis is a gas-solid reaction; thus the phenolate must be present as a fine powder rather than a melt to achieve better contact. For the same reason, inert high boilers, such as higher alkanes, are used for thorough mixing [8]. The alkali phenolates are converted into alkali-metal salts of the corresponding hydroxycarboxylic acids, generally by stirring, in pressurized vessels at 120 – 180 °C and a carbon dioxide pressure of 0.5 MPa. Conditions up to 260 °C and 120 MPa are required in the presence of electron-withdrawing groups or in the case of sterically hindered phenols. 3-Nitrophenol [*555-84-7*] is carboxylated with a yield of only 19 % [9]. Carboxylation occurs easily when the aromatic compounds are activated by electron-donating substituents, as in the case of aminophenols or di- and trihydroxyaromatic compounds. Reaction then occurs even in an aqueous alkali-metal carbonate solution.

The yield of dicarboxylic acid increases with increasing temperature and pressure above 1.0 MPa. At 10.0 MPa and 220 °C the equilibrium is shifted mainly toward the product [10]. The type of alkali metal is important for the course of the reaction. Sodium and lithium phenolates are carboxylated at medium temperature (150 – 180 °C) to form mainly 2-hydroxycarboxylic acids. Potassium phenolates give 2 – 4 mixtures at medium temperature, but above 220 °C the 4-hydroxycarboxylic acid is favored.

Exchange of a Sulfonic Acid, a Halide, or a Diazonium Group for the Hydroxyl Group. The most common method for producing 3-hydroxybenzoic acid [*99-06-9*] involves the alkali melt of *3-sulfobenzoic acid* [*121-53-9*] [11]. In some cases, if the carboxyl group is activated by electron-donating substituents, decarboxylation takes place during melting. Nucleophilic substitution of *halogen* by the hydroxyl group is particularly effective if the halogen atom is activated by electron-withdrawing groups in the *ortho* or *para* position [12]. The conversion of *aminobenzoic acids* into hydroxycarboxylic acids is possible via the diazonium salt [13].

Hydroxycarboxylic Acids via Oxidation. Oxidation of *alkylphenols* is another method of producing aromatic hydroxycarboxylic acids. In general the phenolic hydroxyl group must be

protected (e.g., by esterification with sulfuric acid). The alkyl moiety of the sulfuric acid ester is then oxidized easily by using potassium permanganate or chromic acid [14]. Aromatic hydroxycarboxylic acids can also be produced by oxidative alkali melt of appropriately substituted phenols. Alkyl groups ortho to the phenolic hydroxyl group are oxidized most easily [15]. The oxidative alkali melt of *aromatic hydroxyaldehydes* gives the corresponding hydroxycarboxylic acids in good yield [16]. In the presence of silver catalysts, oxidation is possible even in aqueous alkali solution [17]. *Hydroxyacetophenones* can be converted into hydroxycarboxylic acids by using iodine in pyridine (haloform reaction), followed by alkaline hydrolysis of the intermediate pyridinium iodide [18].

Direct introduction of a *hydroxyl* group into the aromatic nucleus by oxidation is possible in the case of ammonium benzoates by using hydrogen peroxide [19]. Thus, the first stage of the Dow process (production of phenols from benzoic acids) involves hydroxylation by air of the metal benzoate ortho to the carboxyl group [20].

Introduction of a *carboxyl group* is achieved by reaction of phenols with tetrachloromethane in the presence of alkali and powdered copper or copper salts [21]. Alkyl ethers of hydroxycarboxylic acids can be obtained directly by carboxylation of alkoxyphenyl metal halides (e.g., magnesium or lithium compounds) with carbon dioxide. Amides and thioamides of alkoxycarboxylic acids are formed by conversion of phenol ethers with carbamoyl chlorides, isocyanates, or isothiocyanates in the presence of aluminum chloride [22]. Rearrangement of arylurethanes in the presence of Lewis acids (e.g., aluminum chloride) may give good yields of the amides of the corresponding hydroxycarboxylic acids [23].

4. Individual Aromatic Hydroxycarboxylic Acids

4.1. Monohydroxycarboxylic Acids

4.1.1. 3-Hydroxybenzoic Acid

Physical Properties. 3-Hydroxybenzoic acid [*99-06-9*], $C_7H_6O_3$, M_r 138.12, crystallizes as white needles (from water) or rhombic prisms (from alcohol). Important physical properties follow: *mp* 203 °C; d_{25} 1.473; dissociation constants K_1 8.71×10^{-5}, K_2 1.18×10^{-10} (at 19 °C); solubility (in 100 g of solution): water 6.11 g (69 °C), 98 % ethanol 39.6 g (65 °C), and *n*-butanol 20.7 g (36.5 °C).

Chemical Properties. 3-Hydroxybenzoic acid does not undergo decarboxylation at elevated temperature and is stable up to 300 °C. This is in contrast to 2- and 4-hydroxybenzoic acids. Electrophilic substitution (e.g., nitration, halogenation, sulfonation) occurs *ortho* or *para* to the hydroxyl group. Thus, nitration with 62 % aqueous nitric acid gives 2-nitro-3-hydroxybenzoic acid as the main product, along with 4-nitro- and 6-nitro-3-hydroxybenzoic acids [24].

Production. 3-Hydroxybenzoic acid can be obtained by alkali melt of the sodium salt of 3-sulfobenzoic acid at 210 – 220 °C. The crude acid precipitates on acidification. It can be purified by recrystallization from water with the addition of activated carbon [11]. 3-Hydroxybenzoic acid can also be obtained from 3-nitrobenzoic acid esters. This occurs via catalytic hydrogenation followed by diazotization of the aminobenzoic acid ester formed. The diazonium salt is then heated [25] with aqueous sulfuric acid. Another production method is oxidation of 3-hydroxybenzaldehyde [*100-83-4*] by bubbling air through a hot aqueous sodium hydroxide suspension [26].

Uses. 3-Hydroxybenzoic acid is used as an intermediate in the manufacture of pharmaceuticals and pesticides. The 3-hydroxybenzoic acid ester of tropine is used in ophthalmology to dilate the pupils. The sodium salt of 3-hydroxybenzoic acid is a cholepoietic agent. Esters and metal salts have germicidal and preservative properties, and are used in food preservation. Ether – esters are suitable as plasticizers for vinyl and cellulose resins.

4.1.2. 4-Hydroxybenzoic Acid

Physical Properties. 4-Hydroxybenzoic acid [*99-96-7*] $C_7H_6O_3$, M_r 138.12, is a white granular crystalline powder consisting of mono-

Table 2. Properties of 4-hydroxybenzoic acid esters

	Methyl [*99-76-3*]	Ethyl [*120-47-8*]	*n*-Propyl [*94-13-3*]	*n*-Butyl [*94-26-8*]	Benzyl [*94-18-8*]
Physical properties					
M_r	152.12	166.17	180.2	194.22	228.24
mp, °C	131	118	97	70	111
Solubility at 25 °C in 100 g of					
Water	0.25	0.17	0.05	0.02	0.006
Water (80 °C)	2.0	0.86	0.30	0.15	0.09
Methanol	59	115	124	220	79
Ethanol	52	70	95	210	72
Acetone	64	84	105	240	102
Benzene	0.7	1.65	3.0	40	2.6
Diethyl ether	23	43	50	150	42
Tetrachloromethane	0.1	0.9	0.8	1.0	0.08
Bacteriostatic properties					
Phenol = 1.0	3.8	8.0	17.0	32.0	109.0

clinic prisms. The physical properties are summarized here: *mp* 216.3 °C; d_{20} 1.497; dissociation constants K_1 3.3×10^{-5}, K_2 4.8×10^{-10} (19 °C); solubility (in 100 g of solution): water 0.49 g (20 °C), 33.5 g (100 °C); 99 % ethanol 38.75 g (67 °C); and *n*-butanol 19.5 g (32.5 °C). 4-Hydroxybenzoic acid is sparingly soluble in chloroform.

Table 2 lists properties of the most commonly used esters.

Chemical Properties. Above 200 °C, 4-hydroxybenzoic acid undergoes decarboxylation to produce phenol and carbon dioxide. Electrophilic substitution (e.g., nitration, sulfonation, or halogenation) occurs *ortho* to the hydroxyl group. With iron(III) chloride, no color reaction occurs but a yellow precipitate is formed.

Production. 4-Hydroxybenzoic acid is produced by carboxylation of potassium phenolate above 220 °C and with a carbon dioxide pressure of 0.45 MPa. The yield of 4-hydroxybenzoic acid (partly present as the dipotassium salt) is 80 % of the theoretical amount in which the conversion of phenolate is 60 %. 4-Hydroxybenzoic acid can also be obtained by isomerization of dipotassium salicylate [27] at temperatures up to 240 °C and under carbon dioxide pressure (slightly above atmospheric pressure) [28], or by oxidation of the alkali-metal salt of *p*-cresol [*106-44-5*] on metal oxides at 260 – 270 °C [29]. Reaction of phenol with tetrachloromethane [*56-23-5*] in the presence of alkali produces salicylic acid [*69-72-7*] as the main product and smaller amounts of 4-hydroxybenzoic acid [*99-96-7*] [21].

Uses. *4-Hydroxybenzoic acid* is an intermediate in the manufacture of dyes, pharmaceuticals, pesticides [30], and plastics. It is used as an emulsifier and also as a corrosion-protection agent [31]. Its esters and corresponding sodium salts are antimicrobial substances with a high no-effect level (1200 mg per kilogram of body weight) [32]. They are, therefore, used under weakly acidic to weakly basic conditions as food preservatives (e.g., in fruit juice), cosmetics, and pharmaceuticals, as well as in technical products (e.g., in lubricants, anti-freezing agents) [33]. Benzyl esters are relatively soluble in preparations containing sugar. Table 2 lists observed bacteriostatic effects [34] compared to phenol.

4-Hydroxybenzoic acid is also used in polymers as a component in the manufacture of polyester and as a constituent of *liquid crystal polymers* [35] (→ Liquid Crystals).

Esters. The food preservative " PHB-Ester" is a mixture of the ethyl and *n*-propyl esters of 4-hydroxybenzoic acid. According to the preservatives regulations in the different countries [*EEC guideline* 64/52 date: 5. 11. 63, 24th amendment 85/585 date: 20. 12. 85; *United States:* Food and Drug Administration, HHS, 21 CFR Ch.I, 4. 1. 86, § 172.145, § 184.1490, § 184.1670; *Federal Republic of Germany:* Verordnung über die Zulassung von Zusatzstoffen zu Lebensmitteln, date: 22. 12. 81 (BGBl.

I, p. 1625/33) 1. amendment date: 20. 12. 84 (BGBl. I, p. 1652)], PHB-Ester in the amount of 0.1 – 10 g/kg is allowed in the preservation of foodstuffs. In Europe, PHB-Esters are identified as E 214 – E 219, in the United States as heptylparaben, methylparaben and propylparaben. A mixture of 60 % ethyl and 40 % propyl ester is particularly effective. Use of such preservatives is, however, restricted by law in selected foodstuffs (e.g., wine) [36]. Some higher esters are used for termite control [37].

The PHB-Esters that are solids as pure components give eutectic mixtures which are liquid at room temperature and are used as oil-in-water emulsions (→ Emulsions) [38]. See Table 2.

4-Hydroxybenzoic acid esters, particularly PHB-benzyl ester, are also employed as color developers in pressure-sensitive [39] and thermosensitive [40] writing systems. Here, purification and decoloration of the PHB-Ester are achieved through distillation with the addition of sodium hydrogen sulfite [41].

4.1.3. Cresotic Acids

Of the ten isomeric cresotic acids, $C_8H_8O_3$, M_r 152.15, *o*- and *p*-cresotic acids are of most interest.

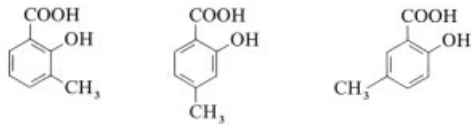

[83-40-9] [50-85-1] [89-56-5]

o-Cresotic acid [*83-40-9*] (2-hydroxy-3-toluic acid, 2-hydroxy-3-methylbenzoic acid), *mp* 165 °C, dissociation constant K_1 6.76×10^{-4}, is produced by carboxylation of the sodium salt of *o*-cresol [*95-48-7*] at 150 – 180 °C. The acid is sparingly soluble in water and has properties similar to salicylic acid. It is used as an intermediate in the preparation of pharmaceuticals and dyes. Its methyl ester is used as an aid in textile dyeing.

m-Cresotic Acid [*50-85-1*] (2-hydroxy-4-toluic acid, 2-hydroxy-4-methylbenzoic acid, 4-methylsalicylic acid), *mp* 177 °C, dissociation constant K_1 6.84×10^{-4}, is produced by carboxylation of the sodium salt of *m*-cresol [*108-39-4*]. It is used as an additive in galvanic baths [42].

p-Cresotic Acid [*89-56-5*] (2-hydroxy-5-methylbenzoic acid, *p*-homosalicylic acid), *mp* 153 °C, is produced by carboxylation of the sodium salt of *p*-cresol [*106-44-5*] with carbon dioxide at 150 – 180 °C. The acid is used as an intermediate in dye production.

4.2. Dihydroxybenzoic Acids

3,4-Dihydroxy- and 2,3-dihydroxybenzoic acids are both obtained by heating catechol [*120-80-9*] in aqueous ammonium carbonate solution at 140 °C.

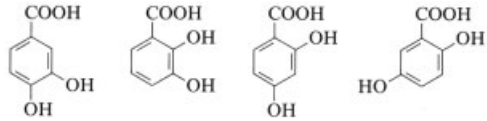

99-50-3] [303-38-8] [89-86-1] [490-79-9]

3,4-Dihydroxybenzoic acid [*99-50-3*] (protocatechuic acid), $C_7H_6O_4$, M_r 154.1, *mp* 203 °C, undergoes decomposition to catechol [*120-80-9*] and carbon dioxide. The acid crystallizes as a monohydrate from aqueous solution.

The most important derivative of this acid is the monomethyl ether, *4-hydroxy-3-methoxybenzoic acid* [*121-34-6*] (vanillic acid), $C_8H_8O_4$, M_r 168.14, *mp* 211 °C (from water). It is obtained by careful alkali melt of vanillin [*121-33-5*] or oxidation of vanillin with silver oxide [43]. Vanillic acid and its esters are constituents of preservatives and disinfectants.

2,3-Dihydroxybenzoic acid [*303-38-8*], $C_7H_6O_4$, M_r 154.1, *mp* 206 °C, undergoes decomposition to catechol and carbon dioxide. The acid crystallizes from water as a dihydrate. Among other uses, it serves as an iron-complexing agent in drugs [44].

2,4-Dihydroxybenzoic acid [*89-86-1*] (*p*-hydroxysalicylic acid, β-resorcylic acid), $C_7H_6O_4$, M_r 154.1, *mp* 226 °C (decomp.), is sparingly soluble in water. The dissociation constant K_1 is 6.05×10^{-4} (30 °C). The acid is produced by conversion of resorcinol [*108-46-3*] with aqueous potassium carbonate solution at 100 °C and a carbon dioxide pressure of 0.45 MPa. In this case, 2,6-dihydroxybenzoic

acid [*303-07-1*] (γ-resorcylic acid) is formed as byproduct and can be separated by recrystallization [45]. Pure 2,4-dihydroxybenzoic acid can be produced by oxidation of 2,4-dihydroxyacetophenone [*89-84-9*] with iodine in pyridine [18].

2,4-Dihydroxybenzoic acid is used as an intermediate in the manufacture of dyes, pharmaceuticals and of additives in photography, and cosmetics. Dyes made from resorcylic acids differ from those made from salicylic acid in their deeper and more intense colors.

The acid is used to complex titanium(IV) fluoride [46], in the passivation of aluminum [47], and as an additive in drilling fluids [48].

3,5-Dihydroxybenzoic acid [*99-10-5*] (α-resorcylic acid), $C_7H_6O_4$, M_r 154.1, *mp* 237 °C (decomp.), is obtained by alkali melt of 3,5-disulfobenzoic acid [49].

2,5-Dihydroxybenzoic acid [*490-79-9*] (gentisic acid, 2,5-DHBA), $C_7H_6O_4$, M_r 154.1, *mp* 205 °C, is soluble in water and alcohols but insoluble in benzene and chloroform. The acid is stable in boiling water and gives a deep blue color with iron(III) chloride. Here, the acid decomposes to form *p*-benzoquinone [*106-51-4*] and carbon dioxide.

2,5-Dihydroxybenzoic acid is obtained by carboxylation of hydroquinone [*123-31-9*] with aqueous potassium hydrogen carbonate solution at 130 °C. It can also be made by oxidation of salicylic acid with potassium peroxodisulfate [*7727-21-7*] in the presence of iron sulfate [50]. 2,5-Dihydroxybenzoic acid is not obtained by reaction of carbon dioxide with the dialkali salt of hydroquinone: instead hydroquinone-2,5-dicarboxylic acid is formed.

2-5-Dihydroxybenzoic acid is used as an antirheumatic drug and, more recently, as a color developer for thermal paper [51].

4.3. Trihydroxybenzoic Acids

3,4,5-Trihydroxybenzoic Acid [*149-91-7*].

Physical Properties. 3,4,5-Trihydroxybenzoic acid (gallic acid), $C_7H_6O_5$, M_r 170.12, is a white to pale yellow crystalline powder, which crystallizes from water in the form of silky needles as the monohydrate (*mp* 258–263 °C) and undergoes decomposition to pyrogallol [*87-66-1*] and carbon dioxde: d_{25} 1.694, dissociation constants K_1 4.63×10^{-3} (at 30 °C) and K_2 1.41×10^{-9}. 3,4,5-Trihydroxybenzoic acid is soluble in warm water, ethanol, diethyl ether, and acetone but insoluble in benzene and chloroform.

Chemical Properties. Solutions of gallic acid, particularly of the alkali-metal salts, absorb oxygen and turn brown in air like pyrogallol. The acid is a strong reducing agent that can reduce gold or silver salts to the metal. With iron(III) salts, an intensive blue complex is formed that is used for ink dyes. Gallnut ink, which consists of gallic acid and iron(II) sulfate, produces the blue iron(III)–gallic acid complex on paper in air.

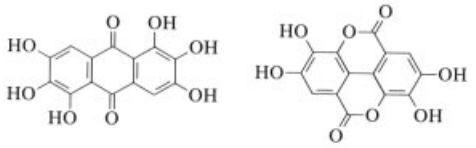

[*82-12-2*] [*476-66-4*]

When gallic acid is heated with concentrated sulfuric acid, hexahydroxyanthraquinone [*82-12-2*] (rufigallic acid) is produced by condensation. Reaction with *p*-nitrosodimethylaniline hydrogen chloride produces oxazine derivatives (→ Azine Dyes, Chap. 3.2., last paragraph). Oxidation with arsenic acid, permanganate, persulfate, or iodine yields ellagic acid [*476-66-4*], as does reaction of methyl gallate with iron(III) chloride. The carboxyl group undergoes azeotropic esterification with alcohols, or, by the Fischer method, with alcoholic hydrochloric acid. The ether–ester of gallic acid is obtained by conversion of dialkyl sulfates or alkyl halides in the presence of alkali, or by using diazomethane [*334-88-3*]. The hydroxyl group in the 4-position is the most reactive. 3,4,5-Trimethoxybenzoic acid [*118-41-2*] undergoes partial hydrolysis in strong acid to form 4-hydroxy-3,5-dimethoxybenzoic acid [*530-57-4*] (syringic acid).

Production. Gallic acid, a component of many tanning agents is an endogenous product in plants. The acid occurs free or bound to tannin [*1401-55-4*] (e.g., in divi-divi, oak bark, gallnuts, pomegranate roots, sumac, and tea). The acid is produced from tannin-rich aqueous gallnut extracts by acidic or alkaline hydrolysis. It

is also obtained by using the enzyme tannase [*9025-71-2*] or molds (*Penicillium glaucum, Aspergillus niger*) to cleave tannin by fermentation. Both the metabolism of gallic acid and its impact on plant growth enzymes have been studied [52].

Uses. Gallic acid is used for the manufacture of iron gallnut ink (see above) and as an intermediate in the production of dyes (e.g., anthragallol [*620-64-2*], gallocyanine [*1562-85-2*], galloflavin [*568-80-9*], and rufigallic acid [*82-12-2*]). Gallate esters, particularly methyl gallate [*99-24-1*] (gallicin), *mp* 157 °C, and propyl gallate [*121-79-9*], *mp* 150 °C, are used as antioxidants and food preservatives (e.g., for fats) [53]. Gallic acid is also employed as a reducing agent for pharmaceutical purposes (Dermatol, Airol, bismuth salt of gallic acid) [54] and for the production of photographic developers and pyrogallol.

4.4. Mandelic Acid

Racemic (*R,S*)-Mandelic Acid [*611-72-3*].

Physical Properties. (*R,S*)- Mandelic acid (α-hydroxyphenylacetic acid, phenylglycolic acid), $C_6H_5-CH(OH)-COOH$, $C_8H_8O_3$, M_r 152.15, *mp* 118 – 121 °C, is a colorless crystalline powder. It is readily soluble in ethanol and diethyl ether, less soluble in chloroform, and insoluble in petroleum ether. Its water solubility is 15 g per 100 mL of water. In alkaline solution, mandelic acid dissolves to form salts.

Chemical Properties. Mandelic acid turns brown when exposed to light. It can be esterified easily with alcohols in the presence of hydrogen chloride.

Production. (*R,S*)-mandelic acid is produced by hydrolysis of mandelic acid nitrile [*532-28-5*] (*mp* 10 °C) with hydrochloric acid. Mandelic acid nitrile is obtained by the conversion of benzaldehyde [*100-52-7*] and hydrogen cyanide in the nascent state (NaCN + HCl) under refrigeration. Another method of producing mandelic acid is the hydrolysis of α,α-dichloroacetophenone with alkali.

Uses. Mandelic acid esters have analgesic, antirheumatic, and spasmolytic effects. For examples, *cis*-3,5,5-trimethylcyclohexyl mandelate [*456-59-7*] (cyclandelate, Spasmocyclon, Clan-dilon), $C_{17}H_{24}O_3$, M_r 276.38, *mp* 54 – 58 °C, is used as a spasmolytic and to stimulate blood circulation. Mandelic acid tropine ester [*51-56-9*] (homatropine – HBr), $C_{16}H_{22}BrNO_3$, M_r 356.26, *mp* 212 – 215 °C, is readily soluble in water and ethanol, and is used in ophthalmology to dilate the pupils. Hexamethylenetetramine mandelate [*587-23-5*] (mandelamine, Mandropine, diuramine, hexydaline) $C_{14}H_{20}N_4O_3$, M_r 292.34, *mp* 127 – 130 °C, forms colorless crystals that are readily soluble in diethyl ether. This salt is used to treat urinary tract infections.

Levorotatory (*R*)(−)-Mandelic Acid [*611-71-2*]. (*R*)-Mandelic acid forms colorless crystals, *mp* 132 – 135 °C, specific rotation $[\alpha]_D^{20}$ ($c = 5$) −154 to 157 °C. It is produced by resolution of racemic (*R,S*)-mandelic acid with cinchonine [*118-10-5*] in aqueous solution, whereby the salt of (*S*)-(+)-mandelic acid crystallizes. If resolution is performed with *S*-(+)-aminodiol [*3306-06-7*], (*R*)-(−)-mandelic acid crystallizes, while (*S*)-(+)-mandelic acid [*17199-29-0*] remains in solution. (*R*)-(−)- and (*S*)-(+)-mandelic acid can again be subjected to racemic resolution after isomerization.

5. Toxicology

Toxicological and physiological properties of the different aromatic hydroxycarboxylic acids are summarized in the following material.

3-Hydroxybenzoic Acid. LD_{50} 3700 mg/kg (rat, i.p.), TDLo 400 mg/kg (rat, s.c., 11 d pregnant); classification: drug, teratogen [55, ; date: 1/84].

4-Hydroxybenzoic Acid. LD_{50} 2200 mg/kg (mouse, oral), LD_{50} 210 mg/kg (mouse, i.p.), LD_{50} 1050 mg/kg (mouse, s.c.); classification: drug, mutagen [55, ; date: 8/83].

o-Cresotic Acid. LD_{50} 1000 mg/kg (mouse, oral). LD_{50} 345 mg/kg (mouse, i.v.); classification: drug [55, ; date: 6/85].

m-Cresotic Acid. LD_{50} 1800 mg/kg (mouse, oral); classification: drug [55, ; date: 6/85].

p-Cresotic Acid. LD_{50} 1000 mg/kg (mouse, oral); classification: drug [55, ; date: 6/85].

4-Hydroxy-3-methoxybenzoic Acid. Classification: mutagen [55, ; date: 8/82].

2,4-Dihydroxybenzoic Acid. LDLo 642 mg/kg (rat, s.c., 11 d pregnant); classification: teratogen [55, ; date: 3/84].

3,5-Dihydroxybenzoic Acid. LD_{50} 2000 mg/kg (mouse, i.v.) [55, ; date: 6/84].

2,5-Dihydroxybenzoic Acid. 2,5 Dihydroxybenzoic acid is an intermediate in the metabolism of salicyclic acid [56]: LD_{50} 800 mg/kg (rat, oral) LD_{50} 3000 mg/kg (rat, i.p.), LD_{50} 374 mg/kg (mouse, i.v.), LDLo 380 mg/kg (rat, s.c., 9 d pregnant), LDLo 642 mg/kg (rat, s.c., 11 d pregant); classification: teratogen, natural product [55, ; date: 5/84].

3,4,5-Trihydroxybenzoic Acid. LD_{50} 320 mg/kg (mouse, i.v.), LD_{50} 5000 mg/kg (rabbit, oral), LDLo 800 mg/kg (mouse, i.p.), LDLo 5 mg/kg (mouse, s.c., 1 d pregnant); classification: agricultural chemical, mutagen, teratogen [55, ; date: 3/85].

6. References

General References

1. *Beilstein* **10,** 43, 134, 149, 192, 214, 375, 464, **10 (1),** 20, 63, 68, 83, 95, 173, 232, **10 (2),** 25, 79, 88, 112, 248, 331, **10 (3),** 87, 422, 1363, 2057, **10 (4),** 125, 315, 345, 1414, 1971.
2. E. Rodd: *Chemistry of Carbon-Compounds,* vol. 3, Elsevier, Amsterdam-London-New York 1956, pp. 756 – 815.
3. A. S. Lindey, H. Jeskey, *Chem. Rev.* **57** (1957) 583 – 620. *Houben Weyl* **8,**372 – 377.

Specific References

4. J. E. Heestermann, *Chem. Weekbl.* **32** (1935) 463.
5. C. G. Barroso, R. C. Torrijos, J. A. Perez-Bustamante, *Chromatographia* **17** (1983) no. 5, 249 – 252. R. Shinohara, *Suishitsu Odaku Kenkyu* **7** (1984) no. 9, 549 – 552; *Chem. Abstr.* **102** (1985) no. 26, 225 717 p.
6. H. Kolbe, *Justus Liebigs Ann. Chem.* **113** (1860) 125 –127. H. Kolbe, E. Lautermann, *Justus Liebigs Ann. Chem.* **115** (1860) 157 – 206. H. Kolbe, *J. Prakt. Chem.* **10** (1874) no. 2, 89 – 111. O. Hartmann, *J. Prakt. Chem.* **16** (1877) no. 2, 35 – 59. H. Kolbe, DE 426, 1877. F. W. v. Heyden, DE 29 939, 1884; 31 240, 1884(R. Schmitt). F. W. v. Heyden, DE 33 635, 1885 (R. Schmitt). F. W. v. Heyden, DE 38 742, 1886 (R. Seifert). S. Marassé, DE 73 279, 1893.
7. D. Cameron, H. Jeskey, O. Baine, *J. Org. Chem.* **15** (1950) 233 – 236. O. Baine, *J. Org. Chem.* **19** (1954) 510 – 514.
8. Rhône-Poulenc, DE-OS 27 54 239, 1976 (J. C. Choulet, J. Nouvel).
9. F. Wessely, *Monatsh. Chem.* **81** (1950) 1071 – 1091.
10. S. E. Hunt, J. I. Jones, A. S. Lindsey, *Chem. Ind. (London)* 1955, 417 – 418.
11. M. F. Clarke, L. N. Owen, *J. Chem. Soc.* 1950, 2108 – 2115.
12. K. W. Rosenmund, H. Harms, *Ber. Dtsch. Chem. Ges.* **53** (1920) 2226 – 2240.
13. F. F. Blicke, F. D. Smith, J. L. Powers, *J. Am. Chem. Soc.* **54** (1932) 1460 – 1471.
14. B. Heymann, W. König, *Ber. Dtsch. Chem. Ges.* **19** (1886) 704 – 706.
15. O. Jakobson, *Ber. Dtsch. Chem. Ges.* **11** (1878) 376, 570.
16. G. Lock, *Ber. Dtsch. Chem. Ges.* **62** (1929) 1177 – 1188.
17. I. A. Pearl, *J. Am. Chem. Soc.* **67** (1945) 1628 – 1629. I. A. Pearl, *J. Am. Chem. Soc.* **68** (1946) 429 – 432. I. A. Pearl, *J. Org. Chem.* **12** (1947) 79 – 84.
18. L. C. King, M. McWhirter, D. M. Barton, *J. Am. Chem. Soc.* **67** (1945) 2089 – 2095.
19. H. D. Dakin, M. D. Herter, *J. Biol. Chem.* **3** (1907) 419 – 434.
20. G. Navazio, A. Scipioni, *Chim. Ind. (Milan)* **50** (1968) 1086 – 1090.
21. J. Zelter, M. Landau, DE 258 887, 1912.
22. L. Gattermann, *Justus Liebigs Ann. Chem.* **244** (1888) 29 – 76.
23. G. J. Gershzon, *J. Gen. Chem. USSR (Engl. Transl.)* **13** (1943) 82.
24. H. Schäfer, *Memorial des Poudres* **27** (1937) 153, 156, 157.
25. H. E. Ungnade, A. S. Henick, *J. Am. Chem. Soc.* **64** (1942) 1737 – 1738.
26. I. G. Farben, DE 506 438, 1926 (F. Hans); *Friedländer* **17,** 472.
27. H. Kupferberg, *J. Prakt. Chem.* **16** (1877) no. 2, 424 – 447.
28. C. A. Buehler, W. E. Cate, *Org. Synth. Coll.* **2** (1943) 341 – 343. Dow Chemical, US 1 937 477, 1932 (L. E. Milk, W. W. Allen). S. E. Hunt, *J. Chem. Soc.* **1958,** 3152 – 3160. A. J. Rostron, A. M. Spivey, *J. Chem. Soc.* **1964,** 3092 – 3096.
29. P. Friedländer, O. Löw-Beer, DT 170 230, 1905.
30. G. W. K. Cavill, J. M. Vincent, *J. Chem. Soc. Ind.* **66** (1947) 175 – 182.
31. BASF Wyandotte, US 4 234 440, 1979 (St. T. Hirozawa, E. F. O'Brien, J. C. Wilson).
32. G. M. Cramer et al., *Food, Cosmet. Toxicol.* **16** (1978) 255.

33. A. J. Maga, K. Lorenz, *Cereala Sci. Today* **18** (1973) no. 10, 326 – 328, 350; *Chem. Abstr.* **84** (1976) no. 23, 163 164 e. H. P. Fiedler: *Lexikon der Hilfsstoffe für Pharmazie, Kosmetik und angrenzende Gebiete,* 2nd ed., Editio Cantor Aulendorf 1981, pp. 480 ff.
34. F. J. Bandelin, *Drug Cosmet. Ind.* **64** (1949) 430.
35. F. Hoffmann-La Roche, DE-OS 2 447 098, 1974 (A. Boller, H. Scherrer). Hitachi Ltd., EP 110 299, 1983 (S. Hattori, K. Iwasaki, T. Kitamura, A. Muko et al.). Chisso Corp., DE-AS 2 751 403, 1977 (T. Inukai, H. Sato, K. Yokohama).
36. C. S. Ough, *Chem. Eng. News* **65** (1987) 19 – 28.
37. Yoshitomi Pharmaceutical Industries, JP 62 76 651, 1987 (T. Inoue, M. Horie, T. Sato).
38. Mallinckrodt Inc., US 4 309 564, 1980 (D. F. Loncrini, J. J. Taylor).
39. Mitsubishi Paper Mills Ltd., JP 58-138 689, JP 58-151 292, JP 58-8 686, EP-AP 86638, 1983 (H. Tsukahara, T. Torii).
40. Ueno Pharma Co., JP 58-158 289, 1983 (R. Keno, H. Tsuchiya, S. Itou, T. Tsuchida). Honshupaper Co., JP 58-158 290, 1983 (M. Jsoda, S. Nishimatsu, K. Kuniyama). Fuji Photo Film Co., JP 58-87 089, 1983 (M. Watanabe). Konzaki Paper Mfg Co. Ltd., JP-P 210 057-82, 1982 (H. A. T. Suzuki, R. O. Hayashi, N. I. O. Arai, Y. I. N. Oeda).
41. Mitsui Toatsu Kagaku KK, JP 49-19 275, 1974 (R. Fujihara, K. Nitta).
42. Siemens, DE-OS 3 049 417, 1980 (H. Januschkowetz, H. Laub).
43. I. A. Pearl, *J. Am. Chem. Soc.* **68** (1946) 2180 – 2181. I. A. Pearl, *J. Am. Chem. Soc.* **68** (1946) 1100 – 1101.
44. A. Cerami, J. Graziano, R. W. Grady, C. M. Peterson, *Proc. Symp.: Dev. Iron Chelaters Clin. Use 1975,* 261 – 273; Chem. Abstr. 87 (1977) no. 7, 47 789 a.
45. K. Brunner, *Justus Liebigs Ann. Chem.* **351** (1907) 313 – 331.
46. Beecham Inc., US 4 601 898, US 4 601 899, 1985 (R. E. Stier, W. H. Dunn, J. D. Vidra).
47. Kaiser Aluminium & Chemical Corp., US 3 714 000, 1971 (G. A. Dorsey, Jr.).
48. Texaco Inc., US 3 537 992, 1967 (J. H. Kolaian).
49. L. Barth, K. Senhofer, *Justus Liebigs Ann. Chem.* **159** (1871) 216 – 230.
50. Chem. Fabr. AG vorm. E. Schering, DE 81 297, 1894. F. Mauthner, *J. Prakt. Chem.* **156** (1940) 150 – 153.
51. Fuji Photo Film Co., GB 1 360 581, 1971 (E. Katsusuke, I. Sadao).
52. E. A. Haddock, *Phytochemistry* **21** (1982) no. 5, 1049 – 1062. E. Hasalam, *Fortschr. Chem. Org. Naturst.* **41** (1982) 1 – 46. S. Kobayashi, M. Ariga, T. Ozawa, H. Imagawa, *Agric. Biol. Chem.* **48** (1984) no. 2, 389 – 395.
53. J. W. Thompson, E. R. Sherwin, *J. Am. Oil Chem. Soc.* **43** (1966) no. 12, 683 – 686.
54. P. Pfeiffer, E. Schmitz, *Pharmazie* **5** (1950) 517 – 520.
55. Cis. Inc., 7215 York Road, Baltimore, MD 21212, CIS-Data bank (RTECS and OHMTADS).
56. D. R. Boreham, B. K. Martin, *Br. J. Pharmacol.* **37** (1969) no. 1, 294 – 300; *Chem. Abstr.* **71** (1969) 121 980 u.

Hydroxyhydroquinone → Phenol Derivatives

Hydroxylamine

JOSEF RITZ, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

HUGO FUCHS, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

HOWARD G. PERRYMAN, BASF Corporation, Freeport, Texas, United States

1. Introduction 1
2. Physical Properties 1
2.1. Hydroxylammonium Sulfate 1
2.2. Hydroxylammonium Chloride 2
3. Chemical Properties 2
4. Production 3
4.1. Catalytic Hydrogenation of Nitric Oxide 3
4.2. Catalytic Hydrogenation of Nitrates . 3
4.3. The Raschig Process 4
4.4. Acid Cleavage of Nitroalkanes 4
4.5. Production of Crystalline Salts 4
5. Safety and Environmental Protection 4
6. Quality Specifications 5
7. Analysis 5
8. Storage and Transportation 5
9. Uses 6
10. Toxicology and Occupational Health . 6
11. References 7

1. Introduction

Hydroxylamine was first described in 1865 by LOSSEN [7], and the free base was isolated in 1891 by LOBRY DE BRUYN [8]. The first industrial process for production of hydroxylammonium sulfate was patented by RASCHIG [9] in Germany (1908) and in the United States (1911).

Salts of hydroxylamine and their solutions are of great industrial importance as intermediates, particularly in the production of caprolactam (→ Caprolactam). World capacity of hydroxylamine is estimated to be 800 000 t/a.

2. Physical Properties

Hydroxylamine [*7803-49-8*], NH_2OH, M_r 33.03, forms transparent, colorless, odorless crystals. The compound is readily soluble in water, methanol, and ethanol. Some physical properties of hydroxylamine are as follows:

mp	32.05 °C
bp (29 kPa)	56 °C
Vapor pressure	
at 0 °C	0.36 kPa
at 32 °C	7.1 kPa
Enthalpy of formation	−114 kJ/mol

2.1. Hydroxylammonium Sulfate

Crystalline hydroxylammonium sulfate [*10039-54-0*], $(NH_3OH)_2SO_4$ or $(NH_2OH)_2 \cdot H_2SO_4$, M_r164.14, is readily soluble in water but only slightly soluble in organic solvents. The density of the solid is 1.883 g/cm^3; the bulk density is ca. 1.10 g/cm^3. Hydroxylammonium sulfate decomposes above 120 °C. Aqueous solutions of hydroxylammonium sulfate are clear and colorless. The pH of a 1 wt % aqueous solution (20 °C) is 3.6. Some properties of hydroxylamine sulfate and its solutions are given in Table 1 [10].

Table 1. Properties of hydroxylammonium sulfate and its solutions

	Temperature, °C				
	20	40	60	80	100
Solubility (water), wt %	37.0	45.0	52.5	60.0	68.0
Density (aqueous solution), g/cm^3					
25 wt %	1.184	1.175	1.163		
Saturated	1.246	1.294	1.341	1.410	1.468
Dynamic viscosity (25 wt % aqueous solution), mPa · s	1.90	1.22	0.86	0.65	

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_527

The specific heat of aqueous hydroxylammonium sulfate solutions in the range 20 – 80 °C is as follows:

Concentration, wt %	25	30	40
Specific heat, kJ kg^{-1} K^{-1}	3.426	3.338	3.087

2.2. Hydroxylammonium Chloride

Hydroxylammonium chloride [*5470-11-1*], $(NH_3OH)Cl$ or $NH_2OH \cdot HCl$, M_r 69.50, is a hygroscopic, crystalline salt that is readily soluble in water and in methanol. The addition of 0.5 wt % of pyrogenic silica hinders caking; the silica gel is insoluble and inert. Hydroxylammonium chloride has a density of 1.676 g/cm^3 and a bulk density of ca. 0.780 g/cm^3. It decomposes above 120 °C. The pH of a 1 wt % aqueous solution (20 °C) is 3.2. Some properties of hydroxylamine chloride and its solutions are given in Table 2 [10].

Table 2. Properties of hydroxylammonium chloride and its solutions

	Temperature, °C			
	20	40	60	80
Solubility, wt %				
Water	46.6	54.5	62.6	70.8
Methanol	14	17	21	
Density (saturated solution), g/cm^3				
Water	1.215	1.245	1.278	1.314
Methanol	0.867	0.870	0.878	

3. Chemical Properties

Hydroxylammonium salts decompose on heating; runaway exothermic decomposition can take place as a result of local heating. Decomposition is promoted by heavy-metal impurities, particularly, copper, copper-containing alloys, and copper salts. The decomposition products of the sulfate are sulfur dioxide, dinitrogen monoxide (nitrous oxide), water, and ammonium sulfate; the chloride decomposes to hydrogen chloride, nitrogen, water, and ammonium chloride.

Hydroxylammonium salts react with alkali to give the easily decomposed hydroxylamine. With nitrites, decomposition to dinitrogen monoxide occurs.

Aqueous solutions of hydroxylammonium salts are acidic. Long periods at 80 °C lead to slow decomposition. Solutions of hydroxylamine and its salts are radical traps, whereby ammonia and nitrogen are formed.

Oxidation leads, depending on the oxidant and reaction conditions, to nitrogen compounds in various oxidation states. Strong reducing agents such as zinc or iron powder give ammonia. Many metal ions form complexes with hydroxylamine.

With oleum or chlorosulfonic acid, hydroxylammonium salts yield hydroxylamine-*O*-sulfonic acid. The latter is a suitable reagent for amination, reductive deamination, hydroxymethylation, and formation of heterocycles [11].

Hydroxylamine can react at both the nitrogen and the oxygen atoms; for preparation of O-substituted hydroxylamines the nitrogen atom must be protected to avoid N-alkylation [12].

Hydroxylammonium compounds react with aldehydes and ketones to form oximes:

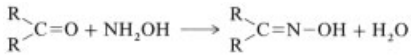

β-Diketones and β-dialdehydes form hydroxylamine isoxazoles.

With carboxylic acids and their derivatives, the products are the corresponding hydroxamic acids:

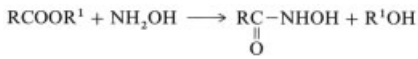

Reaction with isocyanates and nitriles gives *N*-hydroxyureas and amidoximes, respectively:

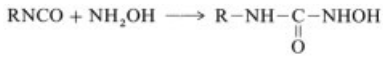

Hydroxylamine also reacts with olefinic double bonds:

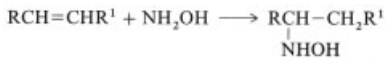

4. Production

Industrial production of hydroxylamine is carried out by reduction of the higher oxidation states of nitrogen. For example, nitric oxide or nitric acid can be hydrogenated catalytically to hydroxylamine. For the reduction of nitrous acid to hydroxylamine, sulfurous acid is used. In all cases the product is an aqueous solution of a salt $NH_2OH \cdot HX$ (where HX is an acid, e.g., sulfuric or phosphoric acid) or $(NH_3OH)X$, rather than free hydroxylamine. Acid hydrolysis of primary nitroalkanes, with simultaneous formation of the corresponding carboxylic acid, plays only a minor role in the production of hydroxylamine (see Section 4.4).

4.1. Catalytic Hydrogenation of Nitric Oxide

To produce nitric oxide of high purity, ammonia and oxygen are reacted together over a platinum–rhodium catalyst at a temperature above 900 °C, the reaction gas being diluted with enough water or water vapor to bring the mixture below the lower explosion limit.

The byproduct, nitrogen dioxide, and excess oxygen are subsequently hydrogenated over a silver catalyst to nitric oxide and water, respectively.

After condensation and simultaneous recovery of water vapor, nitric oxide is purified in a washer and converted to hydroxylamine by reaction with hydrogen below 50 °C over a suspension of partially poisoned platinum catalyst in sulfuric acid. Small amounts of ammonium sulfate and dinitrogen monoxide are produced as byproducts.

$$2\,NO + 3\,H_2 + H_2SO_4 \xrightarrow{\text{Cat.}} (NH_3OH)_2SO_4$$

A continuous multistep process, in which the exhaust gases are extensively recirculated, has proved itself industrially [13], [14]. Residual gas, composed of hydrogen, nitrogen, nitric oxide, and dinitrogen monoxide, can be burned for recovery of steam. Explosive gas mixtures can be avoided in the synthesis of hydroxylamine by the choice of suitable gas composition.

A modified process for the catalytic hydrogenation of nitrogen monoxide, in which both cyclohexanone oxime and hydroxylamine are produced (acidic oximation), is discussed in → Caprolactam.

4.2. Catalytic Hydrogenation of Nitrates

Hydroxylamine is produced in the hydroxylamine – phosphate – oxime (HPO) process by hydrogenation of ammonium nitrate solution in phosphoric acid in combination with synthesis of cyclohexanone oxime for production of caprolactam [15]. This process is described in → Caprolactam.

4.3. The Raschig Process

The Raschig process [9], sometimes modified [16], remains an important process for the production of hydroxylamine.

Water, ammonia, and carbon dioxide react together in an absorption column to give a solution of ammonium carbonate, which forms an alkaline solution of ammonium nitrite with nitrogen oxides at low temperature:

$$(NH_4)_2CO_3 + NO + NO_2 \longrightarrow 2\,NH_4NO_2 + CO_2$$

In a further step, the ammonium nitrite is converted to ammonium hydroxylamine disulfonate with sulfur dioxide:

$$2\,SO_2 + NH_4NO_2 + NH_3 + H_2O \longrightarrow HO{-}N(SO_3NH_4)_2$$

In this process, excess ammonium hydroxylamine disulfonate is recycled. The ammonium hydroxylamine disulfonate solution is drawn off and the salt is hydrolyzed and neutralized to give hydroxylammonium sulfate and ammonium sulfate as products.

$$HO{-}N(SO_3NH_4)_2 \xrightarrow{H_2O} HO{-}NHSO_3NH_4 + (NH_4)HSO_4$$

$$(NH_4)HSO_4 \xrightarrow{NH_3} (NH_4)_2SO_4$$

$$2\,HO{-}NHSO_3NH_4 \xrightarrow{H_2O} (NH_3OH)_2SO_4 + (NH_4)_2SO_4$$

4.4. Acid Cleavage of Nitroalkanes

Mixtures of nitropropane and nitromethane can be cleaved to hydroxylammonium sulfate and

the corresponding carboxylic acid by sulfuric acid at elevated temperature [17]:

$$2\ RCH_2NO_2 \xrightarrow[H_2SO_4]{H_2O} (NH_3OH)_2SO_4 + 2\ RCOOH$$

This process has been used only to a limited extent because of its cost and the limited availability of raw materials.

4.5. Production of Crystalline Salts

Catalytic hydrogenation of nitric oxide leads to solutions of hydroxylammonium salts, which, after removal of the hydrogenation catalyst and neutralization with ammonia if required, can be allowed to crystallize [18]. The Raschig process leads initially to an oxime; acid hydrolysis, removal of the ketone, and subsequent crystallization give the solid salt.

With the intregrated hydroxylamine – cyclohexanone oxime process for production of caprolactam, the solid salts are also, in principle, accessible by acid hydrolysis of the oxime.

5. Safety and Environmental Protection

Safety. Hydroxylammonium sulfate and chloride are not combustible. Above 120 °C, however, they decompose exothermically with evolution of gas. Neither salt is sensitive to shock or friction.

For hydroxylammonium sulfate, differential thermal analysis (DTA) shows that an exothermic reaction begins at 138 °C and reaches a maximum at 177 °C. For hydroxylammonium chloride, the decomposition reaches its maximum at 144 °C.

The heat accumulation test with hydroxylammonium chloride shows an exothermic reaction which begins at 120 °C and reaches a maximum temperature of 280 °C.

Decomposition of hydroxylammonium salts can be countered with water. Although dry hydroxylammonium salts are not caustic or corrosive, aqueous solutions are acidic, and even the damp salts cause considerable corrosion of many metals.

Environmental Protection. Concentrations of hydroxylammonium compounds above 5 mg/L can cause inhibition of activated sludge, thereby impairing sewage-plant performance. The toxicity of hydroxylammonium compounds can be neutralized by oxidation to nitrates with microorganisms. Hydroxylamine is a natural intermediate in the metabolism of nitrification and denitrification bacteria.

In aqueous solution, hydroxylamine can be decomposed with nitrite salts or hydrogen peroxide.

6. Quality Specifications

Hydroxylamine producers are, generally, also hydroxylamine consumers. Quality requirements and quality control are therefore determined by the standards of the specific company.

The product is free of nitro and nitroso compounds, and contains no amines other than hydroxylamine. The pH is ca. 1.2 (20 °C).

Table 3. Typical specificationsfor hydroxylammonium salts

	Hydroxylammonium salt	
	Sulfate	Chloride
Hydroxylammonium salt (min), wt %	99.0	99.0
Free acid (HCl) (max.), wt %	[a]	0.2
Ammonium salts (max.), wt %	1.0 [b]	0.3 [c]
Chloride (max.), mg/kg	1	[a]
Sulfate (max.), wt %	[a]	0.1
Combustion residue (max.), mg/kg	10	20
Heavy metals, [d] mg/kg	1	1
Iron (max.), mg/kg	3	3

[a] Not applicable.
[b] Sulfate.
[c] Chloride.
[d] Calculated as Pb, excluding Fe.

In the limited market for hydroxylammonium salt solutions, quality specifications are based on chemical aspects (e.g., hydroxylamine content), foreign matter such as ammonium salts, or other impurities. A typical quality specification (BASF) for hydroxylammonium sulfate solution is:

Hydroxylamine, NH_2OH (min.)	113 g/L
Hydroxylammonium sulfate, $(NH_3OH)_2SO_4$, (min.)	280 g/L
Free sulfuric acid (max.)	20 g/L
Ammonium sulfate (max.)	25 g/L
Iron (max.)	10 mg/L
Density, d^{20}, (min.)	1.170 g/cm^3
Solidification point	−7 °C

Typical specifications (BASF) for crystalline hydroxylammonium sulfate and chloride are given in Table 3. Specifications and analytical methods vary with producer.

7. Analysis

Hydroxylamine and ammonia can be determined titrimetrically. Reaction of hydroxylamine salts with cyclohexanone leads quantitatively to cyclohexanone oxime and a mineral acid. The amount of mineral acid formed is a measure of the hydroxylamine content of the solution; it can be determined by potentiometric titration at pH 5.5. Reaction of ammonium salts with formaldehyde gives quantitative formation of hexamethylenetetramine and mineral acid. The quantity of mineral acid is a measure of the ammonium content and can be determined by potentiometric titration at pH 8.5.

Determination of possible traces of heavy metals can be carried out by the usual methods.

8. Storage and Transportation

The transportation of hydroxylammonium salts is subject to the following regulations:

1)
Hydroxylammonium sulfate:

GGVE/GGVS:	Class 8, no. 27 c
RID/ADR:	Class 8, no. 27 c
GGVSee/IMDG Code:	Class 8
IATA-DGR:	2865-III
UN no.:	2865

2)
Hydroxylammonium chloride:

GGVE/GGVS:	Class 8, no. 27 c
RID/ADR:	Class 8, no. 27 c
GGVSee/IMDG Code:	Class 8
IATA-DGR:	1759-III
UN no.:	1759

Hydroxylammonium sulfate and chloride are classified as X_n, harmful to health.

Hydroxylammonium sulfate solutions can be stored at up to 40 °C in stainless steel or polyethylene containers. Crystalline hydroxylammonium sulfate and chloride may be stored in a well-aired place below 45 °C; protection from moisture is recommended. With longer storage periods, caking may occur, although the quality of the material is not affected. To avoid vigorous reactions, the contact of hydroxylammonium salts with powders of heavy metals (e.g., copper, zinc, or iron), heavy metal salts (including nitrates and nitrites), oxidants, and alkali must be prevented.

Storage in the vicinity of heat sources such as steam pipes, heating, naked flames, or welding operations, the direct action of ignition sources, and heating of closed containers must be avoided because of the danger of violent decomposition.

Hydroxylamine is used mainly as the sulfate salt, supplied as both an aqueous solution and crystalline material. Aqueous solutions of hydroxylammonium sulfate are shipped in vehicles with tanks or tank containers of stainless steel or polyethylene. Crystalline hydroxylammonium sulfate is supplied packed in polyethylene sacks.

9. Uses

More than 95 % of hydroxylamine production is used to produce cyclohexanone oxime or caprolactam. Depending on patents and available technology, various producers use either hydrogenation or the Raschig process. The differences lie mainly in the amount of ammonium sulfate produced in oximation (→ Caprolactam, Chap. 4.1.).

Hydroxylammonium salts are also used in many branches of the chemical industry as a stable form of hydroxylamine because of the ease with which they undergo reactions. Some applications include the following:

1) *Chemical industry:* a reactant for the production of oximes as intermediates; an oxidizing or reducing agent, depending on pH
2) *Paints and coatings:* production of special oximes for use in anticreaming agents
3) *Pharmaceuticals:* production of antibiotics, steroids, tranquilizers, spasmolytics, and antitubercular drugs
4) *Photography:* stabilizer for developers; additive for color emulsions
5) *Rubber industry:* regulator in the polymerization of synthetic rubber; antioxidant for natural rubber; accelerator for the vulcanization of synthetic rubber
6) *Soap:* auxiliary for refining fats for soap production
7) *Plastics:* regulator and inhibitor in various polymerizations
8) *Metallurgy:* additive for surface treatment of steel; auxiliary for metal extraction from electrolysis waste liquor and for reclaiming spoil tips; flotation agent for copper ores
9) *Nuclear industry:* auxiliary for separation of uranium and plutonium
10) *Textiles:* auxiliary for specific dyeing processes; fixative for textile dyes; starting material for textile dyes based on isatin and isoxazole
11) *Crop protection:* starting material for production of insecticides, herbicides, and plant growth regulators
12) *Food:* intermediate in the production of sweeteners

10. Toxicology and Occupational Health

Hydroxylammonium compounds are irritant and considered harmful to health [19]. The LD_{50} (oral, rat) is ca. 600 – 640 mg/kg. The toxicity to fish, determined as LC_{50} (golden orfe, 48 h), is 1 – 10 mg/L for hydroxylammonium sulfate [5], [10], [19].

The primary effect of hydroxylamine on the skin (rabbit, BASF method) is slight to moderate irritation. Sensitization is possible after skin contact. The primary effect on mucous membranes (rabbit eye, BASF method) is moderate to strong irritation. Hydroxylamine is absorbed through the skin and acts to form methemoglobin.

Examination of the teratogenic effect (in rats) shows no harmful effects on progeny. According to results obtained to date, hydroxylamine and its salts are not carcinogenic, although they are mutagenic toward lower organisms.

Contact with hydroxylammonium salts should be avoided because of the irritant effect and the effect of the dust on skin, eyes, and throat. In the event of contact with the skin or eyes, the affected parts should be rinsed copiously with water. Hydroxylammonium salts must not be swallowed; if necessary, a physician should be consulted. Hydroxylammonium salts are to be labeled as dangerous materials.

11. References

General References

1. *Winnacker-Küchler,* **2,** pp. 182 – 187.
2. J. Keck in: *Method. Chim.,* **vol. 6,** Georg Thieme, Stuttgart/Academic Press, New York 1974, pp. 53 –73.
3. G. R. Tauszik et al., *Appl. Catal.* **17** (1985) 1 – 21.
4. Sibner Hegner Ltd., Brochure *Hydroxylamines – An Introduction*, Beckenham, 1986.
5. Allied Chemicals, Brochure *Oximes – Hydroxylamine Sulfate,* 1984.
6. A. O. Ilvespää et al., *Chimia* **18** (1964) no. 1, 1 – 16.

Specific References

7. W. Lossen, *J. Prakt. Chem.* **96** (1865) 462.
8. C. A. Lobry de Bruyn, *Rec. Trav. Chim. Pays-Bas* **10** (1891) 100.
9. F. Raschig, DE 216 747, 1908; US 1 010 177, 1911.
10. BASF Broschure *Hydroxylamine*, 1986.
11. *Aldrichimica Acta* **13** (1980) no. 1, 1 – 11.
12. *Kirk-Othmer,* 2nd ed., **2,** pp. 493 – 508.
13. K. Jockers, *Nitrogen* **50** (1967) 27 – 30.
14. BASF, DE 1 177 118, 1963 (K. Jockers, H. Meier, E. Eberhardt, L. Taglinger); DE 1 542 628, 1966 (G. Büchler, H. Engelbach, H. Theobald).
15. J. Damme et al., *Chem. Ing. N.Y.* **79** (1972), July 10, 54 – 55. A. H. de Roij et al., *CHEMTECH* 1977, 309 – 315.
16. J. W. Mellor: *Comprehensive Treatise on Inorganic and Theoretical Chemistry,* **vol. 8,**

2nd Suppl., Nitrogen, Longmans Green, London 1967, pp. 115 – 157. Process Survey: Caprolactam, *Eur. Chem. News* (1969), May 2.
17. Commercial Solvents, US 2 755 171, 1955 (L. R. Jones); US 3 066 011, 1959 (R. J. Agne).
18. BASF, DE 1 088 037, 1957 (K. Jockers, H. Meier); DE 1 113 448, 1960 (K. Jockers); DE 1 124 024, 1960 (K. Jockers, H. Meier); DE 2 111 118, 1971 (K. Kartte, K. Jockers, H. Meier, L. Taglinger); Bayer, DE 1 567 511, 1965 (H. Zirngibel, H. Heine, R. Gerken); DE 1 567 513, 1965 (H. Zirngibel, A. Dornemann, R. Gerken, H. Heine, H. Meyer); DE 1 567 515, 1966 (H. Heine, R. Gerken); DE 1 567 532, 1966 (H. Zirngibel, R. Gerken, H. Heine); DE 1 567 533, 1966 (H. Zirngibel, H. Heine, R. Gerken); DE 1 806 537, 1968 (H. Heine). Inventa, CH 466 243, 1963 (H. Füeg, K. Matsuya, K. Haberthür); CH 444 128, (1964) (H. Füeg, K. Matsuya, K. Haberthür); CH 453 316, 1964 (G. Gehring, H. Schneller, H. Füeg, K. Haberthür, K. Matsuya); CH 451 097, 1965 (M. Capaul, K. Matsuya, M. Endinger); CH 507 023, 1968 (K. Germann, A. Gehring, H. Schneller).
19. BASF, DIN Sicherheitsdatenblatt Hydroxylammoniumsulfat, 1987.

Hypnotics

Hartmund Wollweber, Bayer AG, Wuppertal, Federal Republic of Germany

1. **Introduction** 1
1.1. **Amino Acids and Hormones Influencing Sleep** 2
1.2. **Influence of Hypnotics on Sleep** 3
1.3. **Requirements of Hypnotics** 4
1.4. **Applications of Hypnotics** 4
1.5. **Consumption of Hypnotics** 4
1.6. **Classification and Evaluation of Hypnotics** 4
2. **Sedatives, Antihistamines, and Tranquilizers** 5
3. **Alcohols and Aldehydes** 6
4. **Urethanes, Acyclic Urea Derivatives, and Amides** 8
5. **Cyclic Carboxamides, Carboximides, Cyclic Urea Derivatives, and Related Compounds** 9
5.1. **Barbituric Acids** 10
5.2. **Piperidinediones** 15
5.3. **Quinazolinones** 17
5.4. **Benzodiazepines** 18
6. **Other Hypnotics** 27
7. **References** 28

1. Introduction

Hypnotics (Greek $\upsilon\pi\nu o\zeta$, sleep) are drugs that promote the onset of sleep and induce a state similar to natural sleep by depressing the central nervous system [1–28]. Hypnotics are not a sharply defined group of drugs because their sleep-inducing effect greatly depends on the dose. Small doses of a hypnotic have a sedative action; increasing doses cause, in succession, a hypnotic effect, a narcotic effect, and finally severe respiratory depression and death.

Sleep that is properly induced by drugs appears to resemble physiological sleep closely. Both states share the following symptoms:

– reduction of heart rate
– reduction of metabolism
– decrease in concentration of calcium ions in blood
– decrease in muscle tone and blood pressure
– contraction of bronchi and
– contraction of pupils

In contrast to narcotized persons, sleeping persons can be awakened.

Physiology of Sleep. Marked differences between drug-induced sleep and physiological sleep can, however, be detected by using polygraphic methods, such as measurement of pH changes and eye movements (electrooculogram), electromyogram, and electroencephalogram (EEG). On the basis of data obtained from these methods, two types of sleep of importance for the restoration of the body have been described:

1) *Orthodox, synchronized sleep* or nonrapid eye movement (NREM, nonREM) sleep can be distinguished by δ waves corresponding to periods of reduced activity on the EEG (δ sleep). Since these waves are slow, this phase is also referred to as slow wave sleep (SWS).
2) *Paradox sleep* or rapid eye movement (REM) sleep has high EEG activity.

The two types of sleep form a cycle of 110–140 min; four to five cycles are observed in adults each night (Fig. 1). Sleep is extremely important for the renewal of strength. The main phases of natural sleep are an initial deep, dreamless sleep, during which the body temperature falls, and a light, dream-filled, restless sleep before waking. The depth of orthodox sleep has four levels, as seen on the EEG (Fig. 1A), which can partly be correlated with certain body functions [10].

The areas of the brain that regulate body temperature and the sleep–wake cycle are close to each other and are functionally related. The modulations of sleep are correlated with the circadian rhythm of body temperature. The onset of sleep at night coincides with a lowering of body temperature, characteristic of NREM sleep. People who display a temperature difference of less than 1 °C are short sleepers and those

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_533

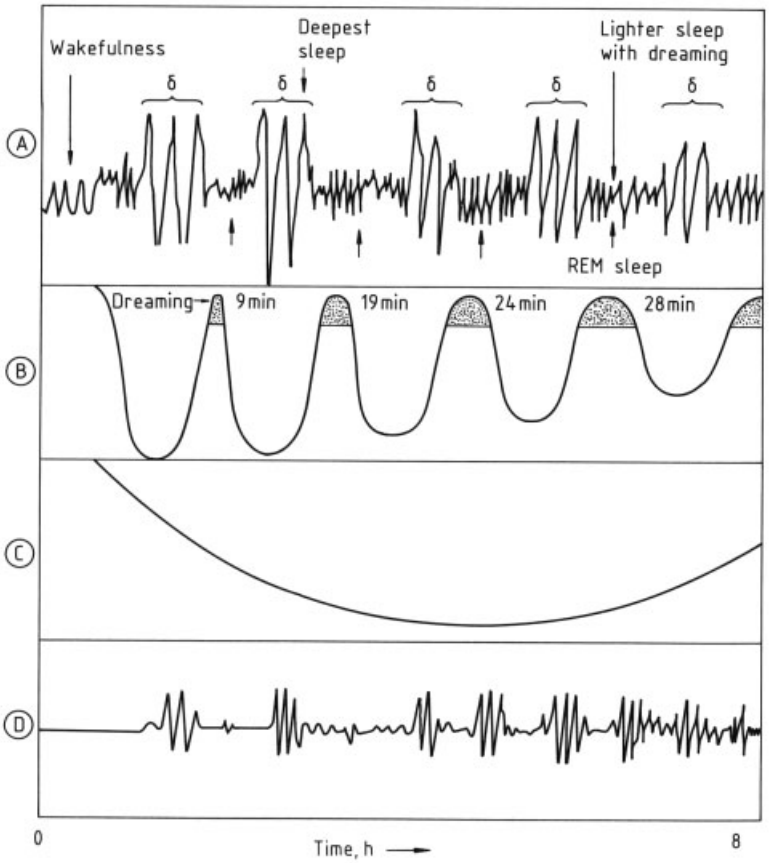

Figure 1. Different phases of normal adult sleep (ca. 8 h)
A) Electroencephalogram; B) Depth of sleep; C) Body temperature; D) Body movement

who display larger differences (>1 °C) are long sleepers. Body tissues are regenerated during orthodox sleep, and cerebral functions during the REM phase.

The REM phases account for ca. 20 – 25 % of sleep and appear to be periods of relatively high cerebral activity in which dreams occur. They are detected by measuring the spontaneous movements of the closed eye. In the absence of these dream phases sleep is not restful. Dreams are of great importance for learning and memory; they assimilate the experiences of the preceding days. Generally, dreams can be remembered slightly or not at all. With increasing age, the duration of REM sleep steadily decreases to about 20 – 30 % of its original duration. The REM phase is especially activated when a person is confronted with an emotionally arousing, intellectual task calling for creative effort, i.e., when alterations in thinking or behavior are necessary. Confrontation with problems also intensifies REM sleep, e.g., an astronaut experiences greatly increased REM sleep in his first night in space. Continuous suppression of REM sleep leads to psychological disturbances and psychosomatic symptoms.

1.1. Amino Acids and Hormones Influencing Sleep

Three *neurohormones* are involved in controlling sleep: serotonin, noradrenaline, and acetylcholine. Serotonin regulates orthodox sleep, whereas acetylcholine and noradrenaline influence the REM phase. The rates of cell division are the highest during sleep, and large amounts of *growth hormone* (GH), which stimulates pro-

tein synthesis, are secreted during the third and fourth levels of orthodox sleep. The EEG waves at this stage are also large and slow.

Melatonin is rhythmically secreted by the pineal gland. It stimulates specific receptors in the body's biological clock, the suprachiasmatic nuclei of the hypothalamus and may thus regulate the human circadian rhythm by increasing serotonin levels. The human biological clock can be "adjusted" with exogenous melatonin. Continous exposure of humans to light leads to sleep disorders by inhibiting melatonin synthesis [16], [29], [30].

A sleep inducing nonapeptide, *delta-sleep-inducing peptide* [*62568-57-4*] (DSIP), has been isolated from rabbit serum. This peptide, which also occurs in humans, has been synthesized and has been prepared by genetic engineering techniques. DSIP coordinates the rhythmic circadian sleep – wake cycles by promoting sleep functions (without having a sedative effect), and enhances mental performance when awake. At the same time, DSIP invokes a higher tolerance to stress. The action of this nonapeptide is attributed to modulation of the adrenergic receptors [31].

L-*Tryptophan,* a common amino acid found in the human body is a precursor of serotonin. In daily doses of 4 – 7 g it causes sedation during the day and a shortening of latency of sleep at night. It is registered in Switzerland as a "biological sleep-inducing drug." However, studies have questioned the sleep-modulating effect of this amino acid [32].

1.2. Influence of Hypnotics on Sleep

Most hypnotic drugs shorten both phases of sleep and suppress the paradoxical (REM) phase. Thus, the sleep curve is levelled out, and the EEG resembles that of a narcotized person. Hence, drug-induced sleep does not have the restorative quality of natural sleep and is no substitute for the treatment of insomnia. A research report of the Max Planck Institut für Psychiatrie (München, 1980) states that conventional sleep-inducing drugs do not produce restful sleep; since drug-induced sleep is not identical to natural sleep, taking sleeping pills is not advised. Pharmacologists also recommend the use of hypnotic drugs only for a few days, if at all, because of their many adverse effects. Alcohol should not be taken with sedatives or hypnotics because the reactions of the patient are then adversely affected.

Prolonged use of sedatives and hypnotics reduces their effect on the cells of the central nervous system (see also Section 5.1). Most hypnotics, particularly highly active preparations, can create psychological and physical dependence, resulting in withdrawal symptoms when treatment is terminated. In addition, many hypnotic drugs that suppress and shorten the REM phase produce a *rebound effect*, i.e., a sudden withdrawal after extended periods of use causing a prolongation of the REM phase. As in alcohol withdrawal, this effect can even lead to delirium tremens. In this way, the body tries to make up for lost REM sleep.

1.3. Requirements of Hypnotics

There is still no completely satisfactory hypnotic drug. The following requirements can be defined for an ideal hypnotic:

- reliable effect
- wide therapeutic range
- limited duration of action, no longer than 8 h
- physiological sleep
- no disturbance of deep sleep or REM sleep
- low toxicity
- no enzyme induction
- no respiratory depression
- no additive effect with alcohol
- no amnesia, confusion, or ataxia after arousal in the night
- no excessive daytime sedation
- no loss of potency on repeated use
- no risk of dependency, e.g., insomnia after termination of treatment
- no after effects (e.g., hangover)

1.4. Applications of Hypnotics

Sleep-inducing drugs are used by the following groups:

1) healthy people with demanding jobs or irregular working hours, who often take sleeping pills after stimulants such as coffee;

2) sick people with symptoms such as fever, sore throat, or cold;
3) old people with insomnia due to cerebral sclerosis usually have an inverted day – night rhythm; they are awake and mobile at night, and tired and peaceful in the morning.
4) sick people with uncomfortable and sleepless nights due to pain, who have difficulty in breathing, heart trouble, etc.; and
5) bedridden people with serious illnesses, who are deprived of physical exercise.

People belonging to groups 1 and 2 can be treated with mild relaxants by reducing the susceptibility of the reticular formation (site of the center of wakefulness in the brain) to external stimuli. These substances include plant components with sedative properties, such as valerian and hop, and tranquilizers or sedatives which, even in high doses, do not have narcotic properties. People belonging to groups 4 and 5 may have to be treated for extended periods with hypnotics and even with antipsychotic and antidepressant drugs.

1.5. Consumption of Hypnotics

The importance of hypnotic drugs is reflected in the fact that 21 % of the population of France (February 1975) and 32 % of the inhabitants of Los Angeles (1973) suffered from insomnia. In France, 2.5×10^6 tablets, capsules, and dragées are taken every evening to promote sleep [33–38]. The dramatic increase in the consumption of hypnotics with age can only be partly justified because older people require less sleep at night. Instead, they have a more pronounced endogenous sleep profile with increased cerebral readiness to sleep at 9.00, 13.00, and 17.00 h.

1.6. Classification and Evaluation of Hypnotics

Hypnotics may be classified as follows [34]:

1) benzodiazepine hypnotics;
2) nonbenzodiazepine hypnotics, e.g., chloral hydrate, barbiturates, piperidinediones, and methaqualone;
3) other drugs with sedative effects, (e.g., tranquilizers, antidepressants, and antihistamines); and
4) natural preparations, e.g., valerian, ethanol, and DSIP.

There is no sharp distinction between sedatives and hypnotic drugs; differences in activity are quantitative rather than qualitative. Both inhibit the activation system in the reticular formation. Drugs belonging to groups 1 and 2 are true hypnotics. Drugs belonging to group 3 are only occasionally used as hypnotics.

Der Bewertende Arzneimittelindex (The Valuating Drug Index) assesses "benzodiazepine sleep-inducing drugs" positively, especially because of their wide therapeutic range [37]. Other recommended drugs are chloral and its derivatives (see Sections page 7page 7), diphenhydramine, and doxylamine (see Sectionpage 5). Derivatives of barbituric acid, piperidinediones, and quinazolinones are not recommended. In comparison with the benzodiazepines, these cause more undesirable effects (respiratory depression, negative influence on REM sleep, enzyme induction, and problems of dependence). Their therapeutic range is narrow. Compounds containing bromine are also to be avoided because they may cause bromism (see Chap. 2). Valerian preparations (e.g., Valepotriate) and their components are not favored because of their in vitro cytotoxic and alkylating action. The therapeutic importance of these drugs cannot be assessed at present [37].

The lists of trade names cited for individual hypnotics are not complete.

2. Sedatives, Antihistamines, and Tranquilizers

Sedatives and Tranquilizers. There is no significant distinction between sedatives and hypnotics; the same drugs are often used for both purposes. Sedatives (→ Sedatives) and tranquilizers (→ Psychopharmacological Agents) are used to sedate and relieve anxiety during the day and, in addition, to induce sleep at night. Therapeutic doses of tranquilizers have a calming, relaxing effect and alleviate psychological symptoms such as fear, anxiety, restlessness, and tension. Ideally, they should not produce sedative

or hypnotic effects. The first therapeutically used hypnotics were plant extracts with either sedative (valerian root and hop) or analgesic action (opium alkaloids). In spite of the sedative effect of opium alkaloids [39], they should not be used for obvious reasons (addiction, euphoria, poor dose adjustment, and legality). Of all the plant extracts available, only one substance, valmane, is exactly defined from a chemical standpoint.

Valepotriate is a mixture of valtratum [*18296-44-1*] (**1**), acevaltratum [*25161-41-5*] (**2**), and didrovaltratum [*18296-45-2*] (**3**):

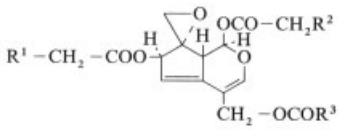

1: $R^1 = R^2 = CH(CH_3)_2$, $R^3 = CH_3$
2: $R^1 = CH(CH_3)_2$, $R^2 = C(CH_3)_2-O-CONH_2$, $R^3 = CH_3$

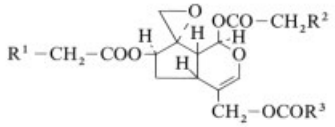

3: $R^1 = H$, $R^2 = CH(CH_3)_2$, $R^3 = CH_2-CH(CH_3)_2$

These three compounds, which occur in nature in the root of *Valeriana officinalis*, have a sedative, mildly psychotropic action. They act as mild hypnotic agents because they restore physiological equilibrium in vegetative disorders. The activity of valmane has been questioned [38].
Trade Names. Valmane (Kali-Chemie, FRG), Nevripan (Medopharm, FRG).

Simple *bromine-containing sedatives* (sodium, potassium, or calcium bromide, or the hydrobromides of magnesium glutamate or magnesium aspartate) should not be used under any circumstances. If given for long periods, they cause a mental condition known as bromism, which is characterized by symptoms such as loss of concentration and memory, insomnia, rash, and psychoses.

Antihistamines. The following antihistamines have sedative properties and are sometimes used as hypnotics:

Meclozine or Meclizine [*569-65-3*], 1-[4-chlorobenzhydryl]-4-(3-tolyl)piperazine dihydrochloride (→Antiallergic Agents, Chap. 2.1.4.)
Trade Names: Calmonal (Heiden, FRG), Bonine (Pfizer, USA).

Diphenhydramine hydrochloride [*147-24-0*], 2-(benzhydryloxy)-*N,N*-dimethylethylamine (→Antiallergic Agents)
Trade Names: Benadryl (Parke-Davis, USA), Halbmond (Much, FRG).

Doxylamine succinate [*562-10-7*], 2-dimethylaminoethoxyphenylmethyl-2-picoline
Trade Names. Decapryn, Meteprine (Merrell, USA), Unisom (Pfizer, USA).

Psychopharmacological Agents. Of the large number of tricyclic antipsychotropic drugs available, two are used as hypnotics:

Perimetazine [*13093-88-4*], perimethazine, 1-[3-(2-methoxyphenothiazin-10-yl)-2-methylpropyl]-4-piperidinol, $C_{22}H_{28}N_2O_2S$, M_r 384.54 (→Phenothiazines and Derivatives).
Trade Name. Leptryl (Roger Bellon, USA).

Perlapine [*1977-11-3*], 6-(4-methyl-1-piperazinyl)-11*H*-dibenz[*b,e*] azepine, $C_{19}H_{21}N_2$, M_r 291.40, *mp* 136 – 138 °C.

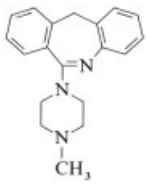

This compound, in the form of yellow prisms, is prepared by chlorinating 11*H*-dibenz[*b,e*] azepin-6-one with phosphoryl chloride to give the imide chloride, which is subsequently treated with *N*-methylpiperazine [40]. The compound thus obtained has hypnotic properties [41]; LD_{50} (oral) 415 mg/kg (mouse).
Trade Names. Hypnodine (Wander, FRG; Takeda, Japan), Pipnodine (Takeda, Japan).

3. Alcohols and Aldehydes

The discovery that after producing an initial excitation phase, ethanol has a general sedative action, led to the development of tertiary and halogenated alcohols as hypnotics. These substances are rarely used today. Chloral hydrate has long-lasting sedative and hypnotic effects because it is metabolized in the body to 2,2,2-trichloroethanol, which also has a sedative action.

Methylpentynol [*77-75-8*], 3-methyl-1-pentyn-3-ol, $C_6H_{10}O$, M_r98.14, *bp* 121–122 °C, is a highly mobile liquid with a pungent smell and burning taste. It solidifies at −30.6 °C and is soluble in almost all organic solvents. It is produced by reacting 2-butanone with sodium acetylide.

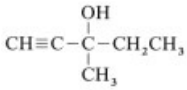

As a result of its short duration of action (0.5–1 g act for 1–2 h), this preparation is especially useful in promoting sleep in cases of mild insomnia.
Trade Names. Atemorin (Scherer, USA), Dalgol, Dorison, Dormison (Schering Corp., USA), Pentadorm (Ruef, Austria), Somnesin (British Drug Houses, UK), Util (Leo, Denmark).

Ethchlorvynol [*113-18-8*], 3-(*β*-chlorovinyl)-3-hydroxy-pent-1-ine, C_7H_9OCl, M_r 144.61, *bp* 173–174 °C, is a liquid with a pungent, aromatic smell. It darkens slowly when exposed to air and light. The compound is immiscible with water but miscible with most organic solvents.

It is prepared by reacting 1-chloropent-1-en-3-one with acetylene [42]. This preparation is a useful short-term hypnotic, but abuse can lead to habituation and addiction.
Trade Names. Roeridorm (Roerig, FRG), Arvynol (Pfizer, USA), Placidyl, Serenesil (Abbott, USA).

Paraldehyde [*123-63-7*], paracetaldehyde, 2,4,6-trimethyl-1,3,5-trioxane, $C_6H_{12}O_3$, M_r 132.16, *bp* 124 °C, is a colorless liquid with an aromatic smell and unpleasant taste. One part of paraldehyde is soluble in eight parts of water at 25 °C. For other properties and synthesis, see →Acetaldehyde, Chap. 8.1. This substance has an unpleasant taste and imparts a bad odor to the breath; addiction is possible. Paraldehyde is only used in clinics and is no longer prescribed.
Trade Names. Paraldehyde Thilo (Thilo, FRG; Fellows, USA), Paral (O. Neal, Jones & Feldmann, USA), Dannetène (Bottu, France).

Chlorethate [*5634-37-7*], bis(2,2,2-trichloroethyl)carbonate, $C_5H_4Cl_6O_3$, M_r324.82 is prepared by the reaction of phosgene with 2,2,2-trichloroethanol in the presence of a base [43], [44]. The insolubility of this substance prevents gastric irritation, which is produced by the hypnotically active parent compound, 2,2,2-trichloroethanol [44].
Trade Name. Clorets (Smith, Kline & French, USA).

Chloral hydrate [*302-17-0*], 2,2,2-trichloro-1,1-ethanediol, $C_2H_3Cl_3O_2$, M_r165.42, *mp* 57 °C, *bp* 98 °C (decomp. into chloral and water), is a crystalline substance with a pungent odor and taste. It is readily soluble in water, ethanol, diethyl ether, and chloroform. For production see →Chloroacetaldehyde, Chap. 3.3.

Doses of 0.5–0.75 g rapidly produce deep sleep, which lasts 4–8 h, without morning hangover. Standard doses of 0.5 g do not produce REM sleep, but doses of more than 0.8 g generally have a negative influence. Adverse effects are irritation of the mucous membranes and hepato- and nephrotoxicity. Prolonged administration can cause habituation, and physical and psychological addiction.
Trade Names. Ansopal (Ferrosan, Denmark), Aquachloral (Webcon, USA), Chloradorm (Knoll Lab., Australia), Chloraldurat (Pohl-Boskamp, FRG), Cohidrate (Coastal, USA), Escre (SS Pharmaceutical, Japan), H. S. Need (Hanlon, USA), Lanchloral (Lancet, Australia), Nervifene (Interdelta, Switzerland), Novochlorhydrate (Novopharm, Canada), Oradrate (Coast, USA), Somnos (Merck, Sharp & Dohme, USA), Suppojuvent Sedante (Juventus, Spain).

Chloral Derivatives.
Chloral betaine [*2218-68-0*], $C_7H_{14}Cl_3NO_4$, M_r 282.57, *mp* 122.5–124.5 °C, is prepared from betaine hydrate and chloral hydrate [45]. This adduct has the same hypnotic properties and contraindications as chloral hydrate, but it does not have the unpleasant physical properties.

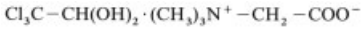

Trade Names. Beta-Chlor (Mead Johnson, USA) Somilan (British Drug Houses, UK), Somnalchlor (Schiaparelli, Italy).

Triclofos [*306-52-5*], 2,2,2-trichloroethyl dihydrogen phosphate, $C_2H_4Cl_3O_4P$, M_r229.39.

The sodium salt, *triclofos sodium* [*7246-20-0*], $C_2H_3Cl_3NaO_4P$, is soluble in water. This preparation has the same physiological properties as chloral hydrate, but it does not have the unpleasant physical characteristics. The compound is metabolized in the body mainly to the hypnotically active 2,2,2-trichloroethanol. A dose of 1.5 g is equivalent to 1 g of chloral hydrate.

Trade Names. Triclos (Lakeside, USA), Triclosyl (Glaxo, UK), Tricloran (C.T.S. Israel).

Petrichloral [*78-12-6*], pentaerythritchloral, $C_{13}H_{16}Cl_{12}O_8$, M_r725.76, *mp* 52 – 54 °C [46].

Chloralodol [*3563-58-4*], chlorhexadol, 2-methyl-(2′,2′,2′-trichloro-1′-hydroxy-4-ethoxy)-pentan-2-ol, $C_8H_{15}Cl_3O_3$, M_r 265.58, *mp* 102 – 124 °C [47].

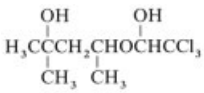

Trade Names. Lora (Wallace, USA), Mechloral, Mecloral (Dumex, Denmark).

Chloralose [*15879-93-3*], α-*O*-(2,2,2-trichloroethylidene)-1,2-α-D-glucofuranose-(*R*), $C_8H_{11}Cl_3O_6$, M_r309.54, *mp* 187 °C, $[\alpha]_D^{22}$ + 19 (5 % solution in 98 % ethanol) [48].

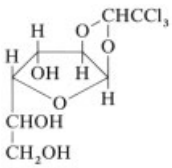

Trade Names. Chloralosan (Kuhlmann, France), Somio (Sidel, France).

Carbocloral [*541-79-7*], *N*-(2,2,2-trichloro-1-hydroxyethyl)- ethyl carbamate, $C_5H_8Cl_3NO_3$, M_r 236.49, *mp* 103 °C (decomp.) [49].

Trade Name. Prodorm (Parke-Davis, USA).

4. Urethanes, Acyclic Urea Derivatives, and Amides

These compounds are mild hypnotic agents. They are generally used to promote the onset of sleep or as daytime sedatives, often in combination with weak analgesics. They are well-absorbed and, depending on the structure, have few side effects. Dependence on these compounds occurs less frequently than on barbiturates and benzodiazepines.

Urethane [*51-79-6*], ethyl carbamate, a mild hypnotic agent, is no longer prescribed because its cytostatic properties can cause agranulocytosis on prolonged administration.

Ethinamate is the most important carbamate. Some other derivatives of carbamic acid are used as central muscle relaxants (→Skeletal Muscle Relaxants) and as psychotropic drugs (→Psychopharmacological Agents).

Ureides are still widely prescribed as mild sedatives and hypnotic agents. They are used alone and in combinations. However, bromine-containing derivatives should be given with great care because of the danger of bromism.

Most *amides,* especially those of simple branched fatty acids, have only a mild sedative action, but some are frequently used to promote sleep.

Ethinamate [*126-52-3*], 1-ethynylcyclohexylcarbamate, $C_9H_{13}NO_2$, M_r167.21, *mp* 96 – 98 °C, solubility in water 0.25 %, in ethanol 35 %, and

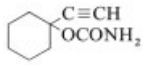

in hexane 2 %, is produced by treating 1-ethynylcyclohexan-1-ol with (1) carbamic acid chloride in the presence of sodium amide [50] or (2) with phosgene and ammonia [51].

Trade Names. Valamin (Schering, FRG), Valmid, Valmidate (Lilly, USA).

Hexapropymate [*358-52-1*], 1-(2-propynyl)cyclohexanol carbamate, $C_{10}H_{15}NO_2$, M_r 181.23, *mp* 99 °C, is synthesized as described for ethinamate [52].

Trade Names. Merinax (Labaz, France), Mekos (Helsingborg, Sweden).

Carfimate [*3567-38-2*], 1-phenyl-2-propynyl carbamate [53], $C_{10}H_9NO_2$, M_r 175.18, *mp* 86 – 87 °C, is a hypnosedative.

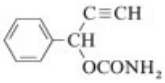

Trade Names. Equilium (Delagrange, France), Nirvetil, Nirvotin (Erba, Italy).

Ectylurea [*95-04-5*], *cis*-(2-ethylcrotonoyl) urea, $C_7H_{12}N_2O_2$, M_r 156.18, *mp* 191 – 193 °C (also *mp* 158 °C), is prepared by treating (2-bromo-2-ethylbutyryl)urea with sodium hydroxide or by reacting carbromal with silver oxide [54]. It is a sedative with a mild hypnotic action.

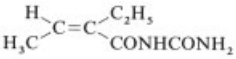

Trade Names. Cronil (Farmigea, Italy), Distesol (Locatelli, Italy), Ectyl (ACO, Sweden), Levanil (Upjohn, USA), Neuroprocin (Minerva-Chemie, Netherlands).

Carbromal [*77-65-6*], 2-bromo-2,2-diethylacetylurea, $C_7H_{13}BrN_2O_2$, M_r237.11, *mp* 116 – 119 °C; 1 g of this compound dissolves in 3000 mL of water, 18 mL of ethanol, 3 mL of chloroform, and 14 mL of ether. It is soluble in alkaline solution and in concentrated sulfuric, hydrochloric, and nitric acids. The reaction of α-bromo-α-ethylbutyric acid bromide with urea at 50 °C produces carbromal [55]. The effects are attributed to the unchanged molecule, even though all the bromine in the organism is broken down to bromide. The same applies to acecarbromal and bromisoval.

Trade Names. Adalin (Bayer, FRG), Bromadyl (Vicario, France), Diacid (Rieswerke, Austria), Fydalex (Boots, UK), Hyperysin (Hommel, Switzerland).

Acecarbromal [*77-66-7*], acetylcarbromal, *N*-acetyl-*N′*-(2-bromo-2-diethylacetyl)urea, $C_9H_{15}BrN_2O_3$, M_r279.14, *mp* 109 °C, is slightly soluble in ethanol and ethyl acetate. It is prepared by the acetylation of carbromal with acetic anhydride in the presence of zinc chloride [56]. Abuse can cause habituation and addiction.
Trade Names. Abasin (Bayer, FRG; Winthrop, USA), Adityl, Carbased (Mallard, USA).

Bromoisoval [*496-67-3*], α-bromoisovalerylurea, $C_6H_{11}BrN_2O_2$, M_r 223.08, *mp* 147 – 149 °C, is soluble in hot water, diethyl ether, ethanol, and alkali. The compound is prepared by the acylation of urea with α-bromoisovalerylbromide [57].
Trade Name. Bromuvan (Treibacher Chem. Werke, Austria), Bromovaleryl, Bromyl (ACO, Sweden), Dagrabromyl (Dagra, Netherlands), Dormigene (Pharmacobel, Belgium), Isobromyl (Clin-Comar-Byla, France), Isoval (Mission, USA), Melosan (Allied Lab., USA).

Apronalide [*528-92-7*], 2-allyl-2-isopropylacetylurea, 1-(2-isopropyl-4-pentenoyl)-urea, $C_9H_{19}N_2O_2$, M_r184.23, *mp* 194 °C [58]. This compound is no longer produced in the Federal Republic of Germany because of its severe adverse effects (purpura hemorrhagica).
Trade Names. Dormid (Rohto, Finland), Neurokin (Leerbeck & Holm, Denmark), Sedormid (Roche, Switzerland).

Pheneturide [*90-49-3*], ethylphenacemide, 2-phenylbutyrylurea, $C_{11}H_{14}N_2O_2$, M_r206.24, *mp* 149 – 150 °C, is prepared by acylating urea with the α-phenylbutyric chloride [59]. It is not only used as a hypnotic agent, but also in the treatment of epilepsy.

Trade Names. Benuride (Bengue, UK; Vinas, Spain), Deturid (Polfa, Poland; Sapos, Switzerland).

Capuride [*5579-13-5*], 2-ethyl-3-methylvalerylurea, $C_9H_{18}N_2O_2$, M_r186.25, *mp* 172 °C [60].
Trade Name. Pacinox (McNeil, USA).

Valdettamil, Novonal [*512-48-1*], 2,2-diethyl-4-pentenamide, $C_9H_{17}NO$, M_r155.23, *mp* 75 – 76 °C, is soluble in 120 parts of water and freely soluble in alcohol and ether [61].
Trade Names. Insomnia (ICN, USA), and as combinations: Arantil (Hoechst, FRG), Dentigoa (Scheurich, Switzerland).

Butoctamidesemisuccinate [*32838-28-1*], *N*-(2-ethylhexyl)-3-hydroxybutyramide hydrogen succinate, $C_{16}H_{29}NO_5$, M_r 315.41, *mp* 36.5 °C, is prepared by the aminolysis of ethyl 3-hydroxybutyrate with 2-ethylhexylamine and the subsequent formation of the semiester of succinic acid [62].

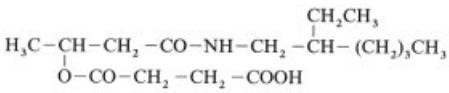

Butoctamide is related to (1-methyl)heptyl γ-bromoacetate, which is found in human cerebrospinal liquid. It increases the brain 5-hydroxytryptamine level and has an antitumor effect. It is given in doses of 200 mg (approximately as effective as 400 mg of bromoacetylurea [63].

5. Cyclic Carboxamides, Carboximides, Cyclic Urea Derivatives, and Related Compounds

Most of the synthetic sleep-inducing drugs used today have a carboxylic acid amide group (**4** in Fig. 2) incorporated into a ring structure. The ring nitrogen atom is linked to many different structural elements (Fig. 2).

The following groups of compounds belong to this category:

N-acyl derivatives: cyclic imides, e.g., (**5**) and 2,6-piperidinediones (**10**)
diacylated cyclic ureas: barbiturates (**8**) and thiobarbiturates (**9**)
N-(β-carbonyl)ethyl derivatives: 2,4-piperidinediones (**6**)
N-iminoacyl derivatives: quinazolinones (**7**)
phenyl N-iminoacyl derivatives: benzodiazepines (**11**)

All groups, except **5**, **7**, and **11**, have a doubly substituted carbon atom adjacent to the carbonyl group of the amide structure. These substituents greatly influence the action and the distribution of the drug in the body.

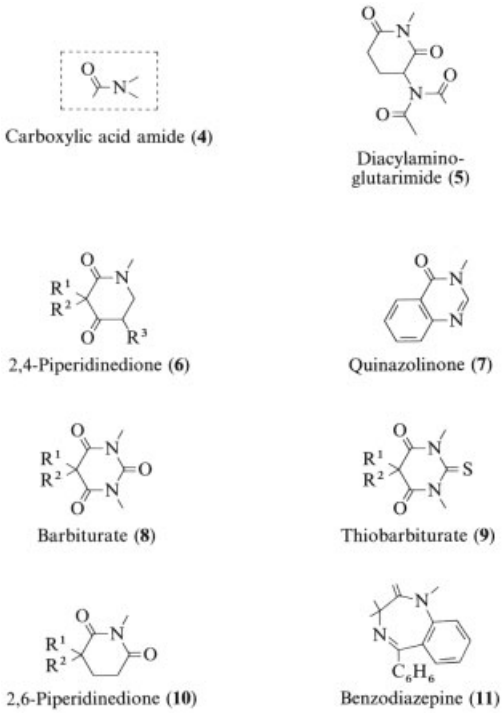

Figure 2. Cyclic carboxamide, carboximide, and urea structures used in hypnotics.

5.1. Barbituric Acids [64–66]

The first representative of this group of substances, barbituric acid, was discovered in 1845 [67]. A synthesis of the first barbiturate with hypnotic properties, 5,5-diethylbarbituric acid (barbital), was completed in 1882; its effects were described in 1903. Since then about 60 different barbiturates have been introduced, of which 20 – 25 are still used in the industrial

Table 1. Duration of action dose, biological half-life, and metabolism of barbiturates

Compound	Type *	Mean hypnotic dose, g	Duration of hypnotic effect, h	Onset of action, min	Plasma protein binding, % **	Biological half-life, h	Proportion metabolized, %	pK value
Barbital	L	0.25 – 0.5	10 – 24	30 – 60	2	72 – 120	< 5	7.88
Phenobarbital	L	0.1 – 0.2	4 – 12	30 – 60		48 – 96	80	7.29
Butobarbital	L	0.05 – 0.1	6 – 12	30 – 60			95	8.12
Probarbital	L – M	0.12 – 0.25	4 – 12	30 – 60				
Amobarbital	M – L	0.1 – 0.3	2 – 8	15 – 30		15	>95	
Cyclobarbital	M	0.1 – 0.2	2 – 8	15 – 30			>99 (in 3 d 42 %)	7.80
Aprobarbital	M	0.065 – 0.13	2 – 8	15 – 30			80	8.15
Allobarbital	M	0.1 – 0.3	2 – 8	15 – 30			70	7.92
Vinylbital	M	0.15	6 – 7	15 – 30			35 in 4 d	
Heptabarb	S – M	0.1 – 0.4	2 – 4	30			98	7.78
Secbutabarbital	S – M	0.1	2 – 4	30				
Pentobarbital	S – M	0.1 – 0.2	2 – 4	30	40	15	99 (in 1 d 25 %)	7.88
Butallyonal	S – M	0.2	2 – 4	30			>90	8.00
Talbutal	S	0.12 – 0.15	2 – 4	30				
Cyclopentobarbital	S	0.12 – 0.25	2 – 4	30				8.10
Propallyonal	S	0.1 – 0.3	2 – 4	30			>99	8.05
Secobarbital	VS – S	0.1 – 0.2	1 – 3	15	45		90	7.80
Hexobarbital	VS – S	0.25 – 0.4	1 – 4	15	2	3.5	95	8.30

* Duration of action: L = long; M = medium; S = short; VS = very short.
** Percentage of the barbiturate that is bound to the protein.

countries. Important properties are listed in Table 1. As hypnotics, barbituric acids all produce almost identical pharmacodynamic effects. However, they show wide differences in pharmacokinetics and metabolism. Differences in onset and duration of activity are related to lipid and water solubility, and rate of metabolism. Lipophilic barbiturates, e.g., the narcotically active, lipid-soluble thiobarbiturates, and the *N*-monoalkylbarbiturates such as hexobarbital rapidly disappear in adipose tissue. As a result of their redistribution in lipophilic compartments of the body, these compounds have a very short duration of action and are rapidly broken down (*short-acting narcotics*, →Anesthetics, General, Chap. 3.1.).

From a therapeutic standpoint, barbiturates used as hypnotic agents are classified according to their duration of action (see also Table 1):

1) *long-acting compounds* (barbital, phenobarbital),
2) *medium-acting compounds* (amobarbital, cyclobarbital),
3) *short-acting compounds* (hexobarbital), and
4) *compounds whose effects begin after 15 – 20 min* (secobarbital).

The duration of action is primarily determined by the rate of enzymatic degradation in the liver, which depends on the chemical structure of the barbiturate. Continual use causes *enzyme induction* in the liver and thus a gradual increase in the rate of barbiturate degradation. Increasing doses are required with prolonged use (*barbiturate habituation*). In some patients barbiturates not only have a sedative and hypnotic action, but also produce excitement and euphoria. This sense of well-being may easily lead to *barbiturate addiction*, a physical dependence with withdrawal symptoms. Barbiturate addicts take the hypnotic because of its euphoric properties and try to compensate for the central depression by using psychoanaleptic drugs.

Cases of accidental and suicidal barbiturate poisoning are fairly common. They are characterized by unconsciousness (which can precede a delirious stage), central respiratory depression accompanied by oxygen deficiency symptoms, and in severe poisoning circulatory collapse with death in 12 h to 4 d. The prognosis for barbiturate poisoning is favorable if treatment is immediate. Essential measures include artificial respiration, irrigation of the stomach, circulatory

support, forced osmotic diuresis using mannitol, exchange transfusion, and hemodialysis.

Barbiturates produce a condition superficially resembling natural sleep. However, the REM phase, which is very important for rest and recuperation, is shortened. Continued use of barbiturates leads to a very low REM value, and sudden withdrawal increases the REM phase (*rebound effect*). The adverse effects of barbiturates are well-documented after more than 50 years of experience. This is an important advantage over newly developed hypnotics [68].

The pK values of the hypnotically active barbituric acids, range between 7.78 and 8.30. Disubstituted barbituric acids are present, to ca. 50 %, in the undissociated form under physiological conditions (pH 7.4). Only this form of the drug can pass through biological membranes. The degree of dissociation is strongly influenced by slight alterations in pH.

Production. The 5,5-disubstituted barbituric acids are still synthesized by the classical condensation of substituted malonic esters (**12**), malonic amides, or cyanoacetic esters (**13**) with urea (**14**), thiourea (**15**), guanidine (**16**), or dicyanodiamide (**17**), followed by the hydrolysis of the resulting imino- or cyanobarbiturate [69] (Fig. 3).

Under acidic or neutral conditions, substituted malonic acids or malonic chlorides can also be used for the condensation with urea or thiourea. It is preferable for the substituents R^1 and R^2 to be present in the malonic ester or cyanoacetic ester molecule; disubstitution of barbituric acids at the 5-position is only possible with highly reactive halides, such as allyl halides.

Barbital [*57-44-3*], barbitone, 5,5-diethylbarbituric acid, $C_8H_{12}N_2O_3$, M_r 184.19, *mp* 188 °C, (sodium salt [*144-02-5*], $C_8H_{11}N_2NaO_3$, M_r 206.18), has a bitter taste. One part of the acid dissolves in 170 parts of water at 25 °C or 17 parts of water at 100 °C. It is freely soluble in acetone, ammonia, and lye. Barbital is prepared by the condensation of diethyl malonate with urea in the presence of sodium ethoxide [70]. Both the acid and its sodium salt are used as long-acting hypnotics; abuse can cause habituation or addiction.

Trade Names. Dormileno (Faes, Spain), Hypnox (Püschel, Austria), Veronal (Bayer, FRG; Merck, FRG), Veroletten (EAS-Labor, Austria).

Sodium salt: Barbinetten (Hormosani, Austria).

Phenobarbital [*50-06-6*], phenemalum, phenobarbitone, 5-ethyl-5-phenylbarbituric acid, $C_{12}H_{12}N_2O_3$, M_r 232.23, *mp* 174 – 178 °C. The acid crystallizes from water in the form of flakes. It is odorless, with a slightly bitter taste; an aqueous solution reddens litmus paper. One part dissolves in 1100 parts of water at 20 °C, 40 parts of boiling water, 10 parts of ethanol, 18 parts of diethyl ether, and 60 parts of chloroform. It is synthesized by the reaction of urea with ethyl phenyl malonate in sodium ethoxide solution [71].

The sodium salt of phenobarbital [*57-30-7*], $C_{12}H_{11}N_2NaO_3$, M_r 254.23, is a colorless, crystalline, hygroscopic, freely water-soluble powder. It is slightly soluble in ethanol, and insoluble in diethyl ether and chloroform. The aqueous solution is stable for only a short time.

Phenobarbital is a long-acting hypnotic with anticonvulsant properties. Long-term use in higher than therapeutic doses can cause physical addiction. An overdose resulting in blood levels of 8 – 12 mg/100 mL can cause death.

Trade Names. Agrypnal (Eggochemia, Austria), Aphenylbarbit (Streuli, Switzerland), Calminal (Wolfs, Belgium), Epsylone (Powell, Canada), Eskabarb Span (Smith, Kline & French, USA), Fenemal (DAK, Denmark), Gardenal (Specia, Canada), Gardenale (Farmitalia Carlo Erba, Italy), Hypnolone (Hartz, Canada), Pepinal, Lepinaletten (Arzneimittelwerk Dresden, GDR), Luminal (Bayer, FRG; E. Merck, FRG; Winthrop, USA), Mediphen (Medic, Canada), Phen Bar (Saunders, Canada), Phenobarbyl (Synochem, FRG), Seda-Tablinen (Beiersdorf, FRG), Solfoton (Poythress, USA), Teolaxin (Paul Maney, Canada); Phenobarbital diethylamine salt: Gratusminal (Farmasimes, Spain); Phenobarbital calcium salt: Fenileal (Turon, Spain), Lumcalcio (Abello, Spain).

Butobarbital [*77-28-1*], butobarbitone, 5-butyl-5-ethylbarbituric acid, $C_{10}H_{16}N_2O_3$, M_r 212.24, *mp* 124 – 127 °C, forms slightly bitter crystals; 1 g dissolves in 5 mL ethanol; preparation is described in [72]. It is a long-

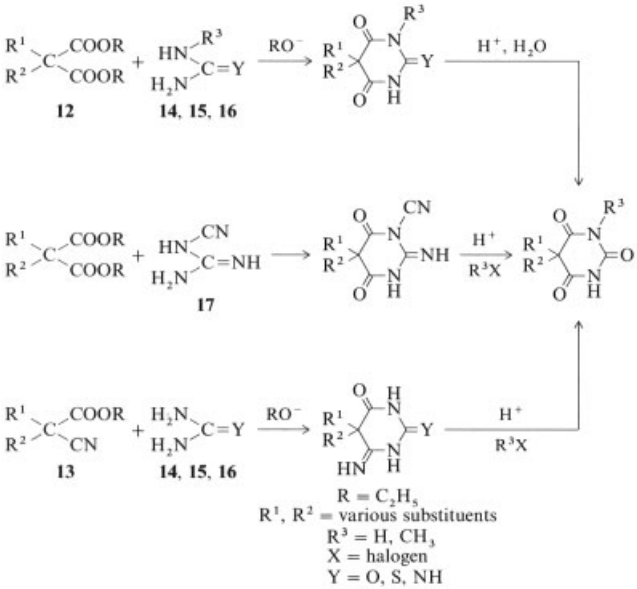

Figure 3. Synthesis of barbiturates.

acting hypnotic; abuse can cause habituation or addiction.
Trade Names. Butenil, Butynoct (Interpharm, Austria), Etoval, Longanoct, Sonabarb (Protea, Australia), Soneryl (Specia, Italy; May & Baker, USA).

Probarbital [*76-76-6*], 5-ethyl-5-isopropylbarbituric acid, $C_9H_{14}N_2O_3$, M_r 198.22, *mp* 197 – 198 °C, is slightly soluble in cold water, but more soluble in hot water and ethanol. Sodium salt [*143-82-8*], $C_9H_{13}N_2NaO_3$, M_r 220.20, is a hygroscopic powder that is soluble in water. The calcium trihydrate [*545-74-4*], $C_{18}H_{26}CaN_4O_6 \cdot 3\,H_2O$, M_r 488.55, forms slightly bitter crystals; 1 g dissolves in 40 g water; preparation is described in [73]. Probarbital is a long- to medium-acting hypnotic; abuse can cause habituation or addiction.
Trade Names. Ipral Sodium, Ipral Calcium (Squibb, USA).

Amobarbital [*57-43-2*], 5-ethyl-5-isopentylbarbituric acid, $C_{11}H_{18}N_2O_3$, M_r 226.28, *mp* 156 – 158 °C, forms slightly bitter crystals; 1 g dissolves in 1300 mL water and in 5 mL ethanol. The sodium salt, $C_{11}H_{17}N_2NaO_3$ [*64-43-7*], M_r 248.25, is hygroscopic and readily soluble in water; preparation is described in [74], and metabolism in [75].

Amobarbitol is a widely used medium- to long-acting hypnotic. It resembles secobarbital and pentobarbital in action with a somewhat longer duration. A blood level of 3 – 6 mg/100 mL of blood is fatal; abuse can cause habituation or addiction.
Trade Names. Amsal (Adams, Australia), Amycal (AFI, Sweden), Amydorm, Amytal (Lilly, USA), Isobec (Pharbec, Canada), Isonal (ICN, Canada), Neur-Amyl (Fawns & McAllen, Australia), Placidel (Miquel, Spain). Stadadorm (Stada, FRG).

Amobarbital sodium: Altinal, Amsal (Adams, Australia), Amsebarb (Paul Maney, Canada), Amylbarb Sodium (Protea, Australia), Amylobeta (Bramble, Australia), Amytal Sodium (Lilly, USA), Barbamyl (Teva, Israel), Dorminal (Boots, UK), Neur-Amyl (Fawns & McAllen, Australia), Novamobarb (Novopharm, Canada).

Cyclobarbital [*52-31-3*], cyclobarbitone, 5-(1-cyclohexen-1-yl)-5-ethylbarbituric acid, $C_{12}H_{16}N_2O_3$, M_r 236,26, *mp* 171 – 174 °C, forms bitter crystals and is moderately soluble in water. The calcium salt [*143-76-0*], $C_{24}H_{30}CaN_4O_6$, M_r 510.59, is prepared by heat-

ing ethyl(1-cyclohexen-1-yl) cyanoacetate with guanidine sulfate in sodium ethoxide, followed by hydrolysis with dilute sulfuric acid [76]. One part dissolves in 70 parts of water. This compound is a medium-acting hypnotic; abuse can cause habituation or addiction.

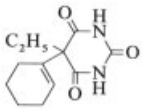

Trade Names. Fabadorm (Bayer, FRG), Phanodorm (Bayer, FRG; Merck, FRG; Winthrop, USA), Phanotal (Kwizda, Austria).

Aprobarbital [*77-02-1*], 5-allyl-5-isopropylbarbituric acid, $C_{10}H_{14}N_2O_3$, M_r210.23, *mp* 140 – 141.5 °C, forms slightly bitter crystals, and is insoluble in water, but soluble in ethanol. The sodium salt [*125-88-2*], $C_{10}H_{13}N_2NaO_3$, M_r232.21, a hygroscopic, bitter powder, is soluble in water. Preparation is described in [77]. This compound is a medium-acting hypnotic; abuse can cause habituation or addiction.
Trade Names. Alurate (Hoffmann-La Roche, Switzerland), Numal (United Chemicals, USA).

Allobarbital [*52-43-7*], allobarbitone, 5,5-diallylbarbituric acid, $C_{10}H_{12}N_2O_3$, M_r 208.21, *mp* 171 – 173 °C, is slightly bitter. One part dissolves in 300 parts of cold water and in 50 parts of boiling water. It is prepared by heating diallyl malonate with urea in sodium ethoxide. It can also be synthesized by the reaction of barbituric acid with allyl bromide in the presence of sodium acetate or copper sulfate [78], [79]. It is a mediumacting hypnotic; abuse can lead to habituation or addiction.
Trade Names. Diadol (Durst, USA), Dial (Ciba-Geigy, Switzerland); it is contained in Spasmo-Cibalgine (Ciba, Switzerland).

Vinylbital [*2430-49-1*], vinylbarbital, 5-(1-methylbutyl)-5-vinylbarbituric acid,$C_{11}H_{16}N_2O_3$, M_r 224.27, *mp* 90 – 91.5 °C, is prepared by the reaction of guanidine with the appropriate malonic ester and saponification. It can also be synthesized by the addition of acetylene to 5-(1-methylbutyl)barbituric acid in the presence of zinc stearate [80].

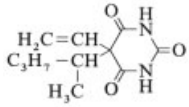

Vinylbital is a medium-acting hypnotic; abuse can cause addiction or habituation.
Trade Names. Bykonox (Byk Belga, Belgium), Optanox, Suppoptanox (Valpan, France), Speda (Byk-Gulden, FRG).

Heptabarb [*509-86-4*], heptabarbital, heptabarbitone, 5-ethyl-5-(1-cycloheptenyl)barbituric acid, $C_{13}H_{18}N_2O_3$, M_r 250.29, *mp* 174 °C; one part dissolves in 13 parts of methanol, 20 parts of ethanol, 13 parts of acetone, 40 parts of diethyl ether, and 5000 parts of water. It is synthesized by the condensation of ethyl(cyclohepten-1-yl)-cyanoacetic ester with urea [81].

This compound is a short- to medium-acting hypnotic; abuse can cause habituation or addiction.
Trade Names. Medomin, Medapan (Ciba-Geigy, Switzerland).

Secbutabarbital [*125-40-6*], secbutobarbital, butabarbital, secbutobarbitone, 5-ethyl-5-(2-butyl)barbituric acid, $C_{10}H_{16}N_2O_3$, M_r 212.24, *mp* 165 – 168 °C; the sodium salt [*143-81-7*], $C_{10}H_{15}N_2NaO_3$, M_r234.23, is a bitter powder; 1 g dissolves in 2 mL of water, and the pH of a 1 % aqueous solution is 9.0 – 10.2. Preparation is described in [82]. Secbutabarbital is a medium- to long-acting hypnotic. It is widely used in the United States as a sedative. Blood levels of 3.5 – 6 mg/100 mL are fatal; abuse can cause habituation or addiction.

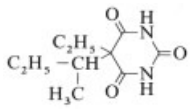

Trade Names. BBS (Reid-Provident, USA), Butabarbital Sodium (Parke-Davis, USA), Butabarpal (Philadelphia-Labs., USA), Butamide (Zenith, USA), Buticaps (Carter-Wallace, USA), Butisol (McNeil, USA), Day-Barb (Anca, Canada), Merisyl (Meriot, Canada), Neo-Barb (Neo, Canada), Sarisol (Halsey Drug, USA).

Pentobarbital [*76-74-4*], pentobarbitone, 5-ethyl-5-(1-methylbutyl)-barbituric acid, $C_{11}H_{18}N_2O_3$, M_r 226.28, *mp* 130 °C; the sodium salt [*57-33-0*], $C_{11}H_{17}N_2NaO_3$, M_r 248.26, is unstable in aqueous solution. It is prepared by boiling ethyl-1-methylbutylcyanoacetic ester with guanidine in sodium ethoxide solution and saponifying the product with dilute sulfuric acid [83]. This compound is a short- to medium-acting hypnotic and is more often used as a hypnotic than as a sedative. A dose of more than 3 g can cause death; blood levels of 1 – 2.5 mg/100 mL are fatal; abuse can lead to habituation or addiction.
Trade Names. Neodorm (Minden, FRG).

Pentobarbital sodium: Butylone (Hartz, Canada), Dorminal (Alfasan, Netherlands), Embutal (Abbott, USA), Hypnol (Stickley, Canada), Isoamytal, Isobarb (U.S. Standard, USA), Iturate, Lethobarb (Loveridge, UK), Napental (Beecham, UK), Narcoren (Veterinaria, Switzerland), Nembutal (Abbott, USA), Nova-Rectal (Nova, Canada), Novopentobarb (Novopharm, Canada), Pembule (Novocol, USA), Penbon (Adams, Australia), Pental (Van-Pelt & Brown, USA; Saunders, Canada), Pentone (Faulding, Australia), Pentosol (Chromalloy, USA), Prodormol (Teva, Israel), Sedanox (Troncin, France), Sombutol (Farmos Group, Finland), Somnopentyl (Pitman-Moore, USA), Vetanarcol (Veterinaria, Switzerland).

Pentobarbital calcium: Insom Rapido (Andreu, Spain), Repocal (Desitin, FRG).

Butallyonal [*1142-70-7*], 5-(2-bromoallyl)-5-*sec*-butylbarbituric acid, $C_{11}H_{15}BrN_2O_3$, M_r 303.16, *mp* 130 – 133 °C, slightly bitter crystals, is almost insoluble in water, but soluble in ethanol. The sodium salt [*3486-86-0*], $C_{11}H_{14}BrN_2NaO_3$, M_r 325.15, forms a bitter, crystalline powder and is soluble in water; the pH of a 10 % aqueous solution is ca. 9.5. Preparation is described in [84]. This compound is a short- to medium-acting hypnotic; abuse can cause habituation or addiction.

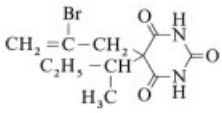

Trade Names. Pernocton (Riedel De Haen, FRG) Pernoston, Sonbutal (Tutag Pharm, USA).

Vinbarbital [*125-42-8*], 5-ethyl-5-(1-methyl-1-butenyl)barbituric acid, $C_{11}H_{16}N_2O_3$, M_r 224.3, *mp* 161 – 163 °C, forms slightly bitter crystals. Preparation is described in [85]; abuse can cause addiction.

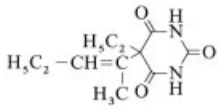

Trade Names. Delvinal (Merck, Sharp & Dohme, USA), Diminal (Astra, Sweden).

Talbutal [*115-44-6*], 5-allyl-5-*sec*-butylbarbituric acid, $C_{11}H_{16}N_2O_3$, M_r 224.25, *mp* 108 – 110 °C, forms slightly bitter crystals. Preparation is described in [86]. This compound is a short-acting hypnotic; abuse can cause habituation or addiction.
Trade Names. Lotusate (Winthrop, USA), Profundol (Promonta, FRG).

Cyclopentobarbital [*76-68-6*], 5-allyl-5-(2-cyclopenten-1-yl(barbituric acid, $C_{12}H_{14}N_2O_3$, M_r 234.25, *mp* 139 – 140 °C, forms slightly bitter crystals and is moderately soluble in hot water. This acid is synthesized by the condensation of urea with the appropriately disubstituted malonic ester [87]. It is a short-acting hypnotic; abuse can lead to habituation and addiction.

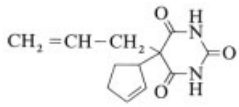

Trade Names. Cyclopal (Siegfried, Switzerland), Cyclopental (Pharmacia, Sweden).

Propallyonal [*545-93-7*], 5-(2-bromoallyl)-5-isopropylbarbituric acid, $C_{10}H_{13}BrN_2O_3$, M_r 289.13, *mp* 177 – 179 °C, forms slightly bitter crystals and is moderately soluble in water. Preparation is described in [88]. This drug is a short-acting hypnotic; abuse can cause habituation or addiction.

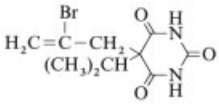

Trade Name. Noctal (Union Chimique, Belgium).

Secobarbital, 5-allyl-5-(2-pentyl)barbituric acid, $C_{12}H_{18}N_2O_3$, M_r 238.28, *mp* 100 °C, is slightly bitter. The sodium salt [*309-43-3*], $C_{12}H_{17}N_2NaO_3$, M_r260.27, is a hygroscopic, bitter powder; preparation is described in [89]. It resembles pentobarbital but has a shorter duration of action and is especially suitable for injection. Blood levels of 1 – 2.5 mg/100 mL are fatal; abuse can cause habituation or addiction.

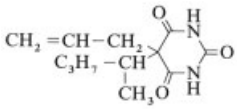

Trade Names. Seconal (Lilly, USA).

Secobarbital sodium (secobarbital solubile): Dormona (Wiedenmann, Switzerland), Immenoctal (Houde-ISH, France), Novosecobarb (Novopharm, Canada), Proquinal (Protea, Australia), S.C.B.Tal (Novo, Canada), Sebar (Vangard, USA), Secaps (Saunders, Canada), Secocaps (M.T.C. Canada), Secogen (Paul Maney, Canada), Seconal, Seotal (Lilly, USA), Quinbar (Adams, Australia).

Hexobarbital [*56-29-1*], 1,5-dimethyl-5-(cyclohexen-1-yl)barbituric acid, $C_{12}H_{16}N_2O_3$, M_r236.26, *mp* 146 °C. For preparation and properties, see →Anesthetics, General. It is a short-acting hypnotic, which rapidly induces sleep. Hexobarbital is frequently used as a short-acting narcotic in the form of its sodium salt [*50-09-9*], Evipan sodium, $C_{12}H_{15}N_2NaO_3$, M_r258.25. The short duration of action and the short biological half-life are related to the metabolism by cleavage of the *N*-methyl group and oxidation of the cyclohexenyl group at the α position. Evipan sodium is used as a narcotic.

Trade Names. Citopan (Nyco, Norway), Cyclopan (Interpharm, Austria), Sombucaps, Sombulex (Riker, USA), Tobinal (Siegfried, Switzerland), Toleran (Kwizda, Austria).

5.2. Piperidinediones

As already mentioned, barbituric acids are cyclic urea derivatives. Another large class of hypnotics is derived from the lactams of 5-aminoglutaric acid; they are divided into two groups: 2,4-piperidinediones and 2,6-piperidinediones (glutaric imides). They all have a disubstituted carbon atom located α to the carbonyl group of the lactam residue.

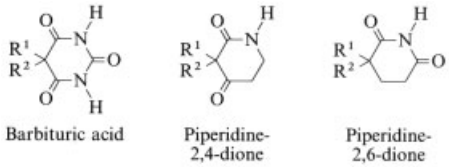

The hypnotic action of the piperidinediones qualitatively resembles that of the barbiturates, but they are less active and are used in larger doses. Adverse effects and toxic reactions are less frequent than with barbiturates. The piperidinediones are rapidly broken down in the body; the symptoms of poisoning resemble those caused by barbiturates.

Taglutimide (see page 16) and *thalidomide* [*50-35-1*] have a monosubstituted carbon atom at position 3. The notorious thalidomide (3-phthalimido-2,6-dioxopiperidine, Contergan) was removed from the market because it caused severe adverse effects (prolonged neuritis and malformation of the fetus).

Pyrithyldione [*77-04-3*], 3,3-diethyl-2,4-dioxo-1,2,3,4-tetrahydropyridine, $C_9H_{13}NO_2$, M_r 167.20 is structurally related to the piperidinediones and exists in three crystalline modifications: (1) *mp* 92 – 93 °C, (2) *mp* 97 – 98 °C, (3) *mp* 81 – 86 °C.

Trade Names. Persedon (Hoffmann-La Roche, Switzerland), Benedorm (Arzneimittelwerke, GDR).

This drug is no longer used because of its profound adverse effects (agranulocytosis). It has been replaced by the homologous methyprylon which does not cause blood dyscrasia.

Methyprylon [*125-64-4*], 3,3,diethyl-5-methyl-2,4-piperidinedione, $C_{10}H_{17}NO_2$, M_r 183.26, *mp* 74 – 77 °C, is soluble in water, ethanol, and chloroform. It is prepared by formylat-

ing 3,3-diethylpiperidine-2,4-dione (dihyprylone) and reducing the resulting 3,3-diethyl-5-hydroxymethylpiperidine-2,4-dione with hydrogen and Raney nickel [90]. The drug acts rapidly but longer than pyrithyldione; it is less toxic than phenobarbital. Prolonged use causes habituation and addiction.

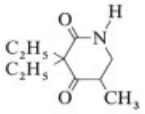

Trade Names. Noludar, Nolurate (Hoffmann-La Roche, Switzerland).

Glutethimide [*77-21-4*], 3-ethyl-3-phenylpiperidine-2,6-dione, $C_{13}H_{15}NO_2$, M_r217.26, *mp* 84 °C, is soluble in ethyl acetate, acetone, diethyl ether, chloroform, ethanol, and methanol, and almost insoluble in water. It forms a hydrate and a water-soluble hydrochloride. Glutethimide is prepared by the addition of 2-phenylbutyronitrile to acrylic ester to give 2-ethyl-2-phenylpentane-1,5-dicarboxylic acid mononitrile, followed by cyclization with sulfuric acid and acetic acid [91]. This compound is effective in inducing sleep that lasts through the night and is also suitable as a daytime sedative. Prolonged administration causes habituation and addiction.

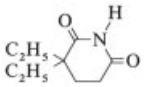

Trade Names. Doriden (Ciba-Geigy, Switzerland; Takeda, Japan), Elrodorm (Dtsch. Hydrierwerke, FRG), Glimide (Polfa, Poland), Gludorm (Knoll, FRG), Regenox (Richter, Hungary).

Alonimid [*2897-83-8*], 1,2,3,4-tetrahydro-4-oxonaphthalene-1-spiro-3′-piperidine-2′,6′-dione, $C_{14}H_{13}NO_3$, M_r243.25, *mp* 197–199 °C,

is prepared by the Michael addition of ethyl acrylate to phenylacetonitrile, followed by cyclization with polyphosphoric acid [92].

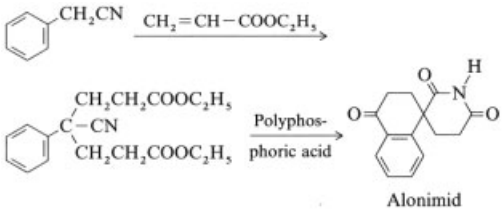

Trade Name. Alonimide (Merrell, USA).

Taglutimide [*14166-26-8*], biglumide, 2-(bicyclo [2,2,1] heptane-2-*endo*-3-*endo*-dicarboximido)-glutarimide, $C_{14}H_{16}N_2O_4$, M_r276.29, *mp* 237 °C, is unstable in alkaline solution, and is slightly soluble in water and in organic solvents. It is prepared by the reaction of *endo*-bicyclo [2,2,2] -heptane-1,2-dicarboxylic acid anhydride with L(+)-glutamic acid, followed by cyclization with urea in dimethyl sulfoxide at 180 °C [93].

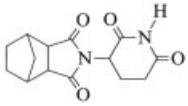

Unlike thalidomide, taglutimide does not produce any teratogenic effects [94]. It is a useful hypnotic because it causes little change in the electroencephalogram of normal sleep [95].
Trade Name. Synval (Kwizda, Austria).

5.3. Quinazolinones

Quinazolinones are cyclization products of anthranilamide. Some have analgesic and antipyretic properties. Only those derivatives in

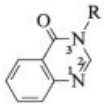

which the nitrogen atom at position 3 has an 2-substituted phenyl group are hypnotically active [96]. The effect is very fast in onset and lasts 3–4 h. These compounds are similar in hypnotic action to the barbiturates but are reported not to suppress REM sleep. They inhibit the biosynthesis of serotonin and catecholamines, which influence the physiology of sleep [97]. The quinazolinones are a group of hypnotics

with codeinelike, cough-suppressant, analgesic properties; they are useful in patients who do not tolerate or respond to barbiturates. Prolonged use may cause habituation and psychological dependence. No specific antagonists are available for quinazolinone poisoning.

Methaqualone [*72-44-6*], 2-methyl-3-*o*-tolyl-4-(3*H*)quinazolinone, $C_{16}H_{14}N_2O$, M_r250.30, *mp* 120 °C (also 114 – 116 °C), is soluble in ethanol, diethyl ether, and chloroform, but almost insoluble in water. The hydrochloride [*340-56-7*], $C_{16}H_{15}ClN_2O$, M_r286.85, *mp* 255 – 265 °C, is prepared by the condensation of *N*-acetylanthranilic acid with *o*-toluidine in the presence of phosphoryl chloride [98], [76]. Methaquolone is present in numerous combinations.

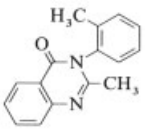

Prolonged administration of methaqualone has caused paresthesia of the extremities; daily doses of 0.3 g should not be taken for more than four weeks (in exceptional cases, eight weeks).
Trade Names. Bon-Sonnil (Diselen, Spain), Dormogen (Leciva, Czechoslovakia), Dormutil (Chem.-Pharm. Werk Oranienburg, GDR), Mequin (Lemmon, USA), Mozambin (Gerot, Austria), Pro Dorm (Schürholz, FRG), Somnotropon (Tropon, FRG), Torinal (Medicamentos y Productos Quimic, Spain), Tuazolona (Nessa, Spain).

Methaqualone hydrochloride: Cateudyl (Covor, Belgium), Dormir (Langley, Australia), Hyptor (Toraude, Canada), Melsed, Melsedin (Boots, UK), Mequelon (Merck-Frosst, Canada), Methasedil (Cooper, Switzerland), Nobadorm (Streuli, Switzerland), Normorest (Merck, FRG), Noxybel (Probel, Belgium), Optimil (Wallace, USA), Optinoxan (Robisch, FRG), Parest (Parke-Davis, USA), Parmilene (Chiesi, Italy), Pexaqualone (Therapex, Canada), Revonal (Merck, FRG), Riporest (Farmitalia, Italy), Sedalone (Pharbec, Canada), Somberol (Boots, UK), Somnafac (Smith, Miller & Patch, USA), Somnium (Fargal, Italy), Sovelin, Soverin (Weifa, Norway), Sovinal (ND & K, Denmark), Toraflon (Toraude, Canada), Tualone (ICN, Canada), Tuazol (Strasenburgh, USA).

Etaqualone [*7432-25-9*], 3-(2-ethylphenyl)-2-methyl-4(3*H*)-quinazolinone, $C_{17}H_{16}N_2O$, M_r 264.31, *mp* 81 °C. Hydrochloride [*97979-65-2*], $C_{17}H_{17}ClN_2O$, *mp* 247 °C. Etaqualone is prepared by the condensation of *N*-acetylanthranilic acid with 2-ethylaniline in the presence of phosphoryl chloride [99].

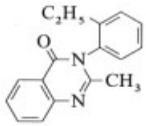

Trade Name. Aolan (Beiersdorf, FRG).

Nitromethaqualone [*340-52-3*], 3-(2′-methoxy-4′-nitrophenyl)-2-methyl-4(3*H*-) quinazolinone, $C_{16}H_{13}N_3O_4$, M_r311.29, *mp* 135 – 137 °C, and the hydrochloride [*3946-23-4*], *mp* 246 – 248 °C, are prepared as the preceding compounds [100]. For action, see [101].

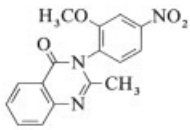

Trade Name. Parnox (Santos, Spain).

5.4. Benzodiazepines

Many benzodiazepine derivatives used as psychotropic agents (→Psychopharmacological Agents) promote sleep by acting as mental relaxants and also by having a hypnotic action of their own [102], [103]. The diverse action of this group of drugs clearly illustrates the difficulty of distinguishing hypnotics from sedatives and tranquilizers if the sleep-inducing effect is the only feature of interest. As with the barbiturates, the sedative and tranquilizing effects of benzodiazepines induce "better sleep." After an initially steep increase in the use of benzodiazepines as hypnotics, a drop in consumption is slowly becoming apparent in Europe and the United States [104], [105]. Even though benzodiazepines are

Table 2. Duration of action, biological half-life, and metabolism of benzodiazepines

Compound	Onset of action, min	Half-life, h	Main metabolic route *	Main active metabolite and half-life, h	Adult hypnotic dose, mg
Nitrazepam	16 – 30	20 – 50	nitroreduction, acetylation	none	5 – 10
Flunitrazepam	9 – 20	10 – 25	demethylation, nitroreduction	dealkylflunitrazepam, 20 – 30	1 – 2
Flurazepam	15 – 45	2.3	oxidation	desalkylfluorazepam, 50 – 100	15 – 30
Oxazepam	45 – 90	5 – 15	conjugation	none	15 – 30
Lorazepam	15 – 45	10 – 15 (20)	conjugation	none	2 – 4
Lormetazepam	15 – 45	9 – 15	conjugation	none	0.5 – 2
Temazepam	45 – 60	9.5 – 12 (17)	conjugation	none	15 – 30
Doxefazepam	15 – 40	7 – 8	dealkylation, conjugation	none	10 – 20
Cinolazepam	15 – 45	3.8	conjugation	none	15 – 30
Quazepam	60	39	oxidation, demethylation, conjugation		
				2-oxoquazepam, 39; demethyl-2-oxoquazepam, 73	7.5 – 15
Haloxazolam	45 – 60	2 – 4	cleavage of oxazoline residue	7-bromobenzodiazepin-2-one (**18**, R = Br, R^1 = H, R^2 = F)	5 – 10
Triazolam	18 – 30	1.5 – 3	oxidation	none	0.25 – 0.5
Estazolam	15 – 40	8 – 31	oxidation	none	2 – 4
Midazolam	16 – 30	1 – 3	oxidation, hydroxylation, conjugation	none	10 – 30
Loprazolam	15 – 30	10 – 16	oxidation	"loprazolam-*N*-oxide", 5 – 8	1 – 2
Brotizolam	10 – 20	4.5 – 9	oxidation	"hydroxymethyl brotizolam", 4 – 8; 3-hydroxy-brotizolam	0.25 – 1

* Plasma protein binding rates are high (usually 90 – 95 %). In most cases, conjugation indicates glucuronization, and oxidation denotes the introduction of a hydroxyl group at the 3-position or in the methyl group of a heterocycle condensed with the benzodiazepine moiety.

not ideal sleep-inducing agents, new products are constantly being marketed. Properties of the most important benzodiazepines are listed in Table 2.

Mode of Action. The benzodiazepines have the same hypnotic action as the barbiturates, with differences in selectivity, metabolism, and adverse effects. Large doses do not possess the narcotic properties exhibited by most barbiturates. Benzodiazepines do not solely have a sleep-inducing action; they exert a combination of effects: sedation, relief of anxiety (production of a better state of mind), muscle relaxation, and anticonvulsant action. These effects shorten both the intermittent waking times and deep sleep (reduction of δ sleep). In addition, a hangover always occurs on the next day. The effects of barbiturates and benzodiazepines are compared in Table 3. From a pharmacological standpoint, the hypnotically active benzodiazepines have a depressant action on the electrical discharges in the areas controlled by the limbic system and block the function of the reticular formation in the brain stem. They are similar to the barbiturates in this respect; however, quantitative differences could be of importance. Studies point to an interaction between benzodiazepines and their receptors. Thus, they activate the postsynaptic receptors for γ-aminobutyric acid (GABA, an inhibitory neurotransmitter) and increase the frequency of the opening of the chloride ion canals [106].

In other words, the benzodiazepines act in the central nervous system by interfering with GABA-mediated neurotransmission [107], [108]. Traces of benzodiazepine derivatives (e.g., diazepam and lorazepam) have been found in wheat and potato and are identical to the synthetic substances. Benzodiazepines also occur in untreated rats, as well as in human serum and brain. However, the concentration is very low (several nanograms per gram of potato) and a direct pharmacological effect on the central nervous system has not been demonstrated [108], [109].

Metabolism. All benzodiazepines are metabolized in the body [110], but in some cases the retention times are very long. In addition, the pharmacologically active metabolites of the benzodiazepines have long residence times in

the body, and the rate of metabolic degradation decreases considerably with age. A plasma half-life of 1.5 – 40 h, and in extreme cases 47 – 100 h, is observed [111]. The main metabolic transformations of benzodiazepines are (Fig. 4):

1) Hydroxylation at position 3 with formation of more active products, which are, however, usually rapidly eliminated after glucuronization
2) Demethylation at N-1 with formation of longer acting demethylbenzodiazepines
3) Ring opening at C-3, which again frequently produces active metabolites that can be recyclized in the organism
4) Heterocyclic moiety of benzodiazepines containing fused heterocyclic structures, (e.g., triazolam) rapidly broken down to form inactive metabolites

Therapy. The expectation that therapeutic problems could be solved with benzodiazepines of lower half-life has proved unfounded. Although the faster onset of effect, swift relief of anxiety, and rapid elimination of these drugs prevent a hangover, rebound insomnia occurs as a severe adverse effect to produce delirium and exogenous psychosis, (a type of premature withdrawal symptom) during the first night of use. A temporary shortening of REM sleep follows with subsequent paradoxical alertness, fear, agitation, and anterograde amnesia during the second half of the night [104]. In addition, abuse of the rapid onset of activity benzodiazepines (e.g., lorazepam and flunitrazepam) has been reported, especially among young people.

From a therapeutic standpoint, the benzodiazepine hypnotics are classified according to their duration of action:

1) long-acting compounds, e.g., flurazepam, quazepam, nitrazepam, and flunitrazepam;
2) medium-acting compounds, e.g., temazepam, lorazepam, loprazolam, oxazepam, and brotizolam; and
3) short-acting compounds, e.g., triazolam, midazolam, and cinolazepam.

Long-acting benzodiazepines should not be used to treat insomnia because of accumulation, occurrence of undesired effects before sleep induction, daytime drowsiness, lowered activity, ataxia, disorientation, and nightmares. Regular administration of most benzodiazepines increases the dose necessary to evoke the same response.

Indications for the use of benzodiazepine hypnotics are given in Table 4.

Table 3. Comparison of the effects and properties of barbiturates and benzodiazepines

Effect or property	Barbiturates	Benzodiazepines
Narcotic	yes	no
Sedative	yes	yes
Muscle relaxation	almost none	yes
Anticonvulsant	yes	yes
Shortening of REM sleep and increase in REM phase on withdrawal (rebound effect)	yes	yes
Retrograde amnesia	yes	almost none
Accumulation	yes	yes
Hangover	yes (clearly sedated waking phase)	yes (also positive influence on nervousness, apprehension, tension, listlessness in the waking phase)
Tolerance	yes	yes
Abuse	yes	yes
Enzyme induction	yes	no
Therapeutic range	narrow	wide
Toxicity	relatively high	relatively low

Synthesis. The hypnotically active benzodiazepines can be classified as benzo-1,4-diazepinones (**18**), or benzodiazepines containing fused heterocyclic moieties, e.g., triazolobenzodiazepines (**21a**–**21c**). The synthesis of these compounds is shown in Figures 5 and 6, respectively [102].

Nitrazepam [*146-22-5*], 1,3-dihydro-7-nitro-5-phenyl-2*H*-1,4-benzodiazepin-2-one, (**18a**), $C_{15}H_{11}N_3O_3$, M_r 281.26, *mp* 224 – 226 °C, is prepared by the cyclization of 2-(aminoacetamino)-5-nitrobenzophenone or by the nitration of benzodiazepinone with an unsubstituted phenyl substituent [112], [113].

Indications: problems in falling asleep, frequent awakening during the night, and early awakening in the morning. The duration of in-

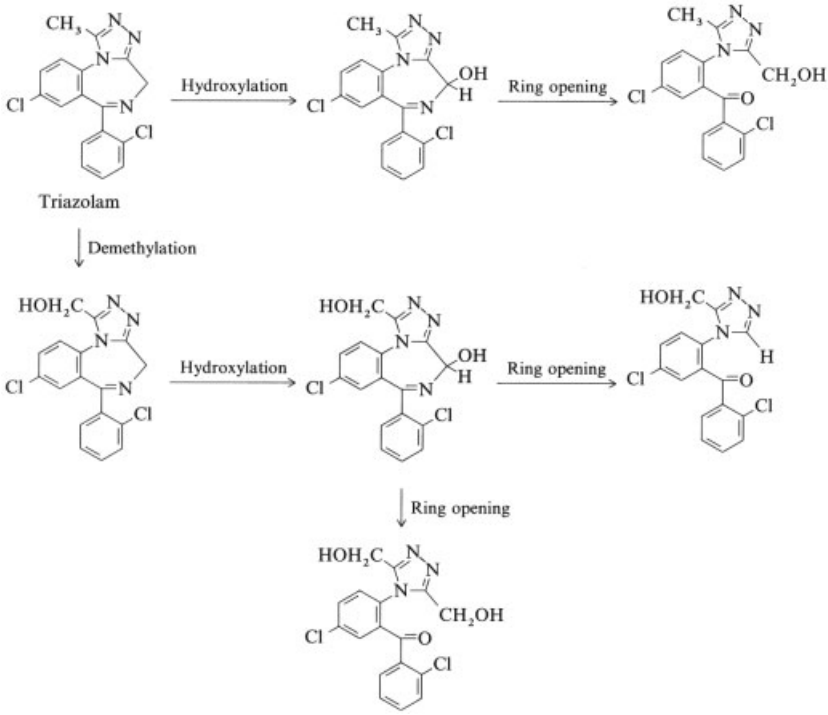

Figure 4. Metabolism of triazolam.

duced sleep is 7 – 8 h; a strong hangover occurs next day [114].

Trade Names. Alodorm (Alphapharm, Australia), Apodorm (Apoth. Lab., Norway), Arem (Lennon, Republic of South Africa), Atempol (Norgine, UK), Benzalin (Shionogi, Japan), Calsmin (Upjohn, USA), Dormo-Puren (Klinge, FRG), Dumolid (Dumex, Denmark), Eatan N (Desitin, FRG), Hipsal (Salvat, Spain), Imeson (Desitin, FRG), Insomin (Orion, Finland), Ipersed (Sidus, Italy), Lagazepam (Lagap, Switzerland), Megadon, Mogadan, Mogadon (Roche, Switzerland), Mitidin (Savorma, Italy), Nelbon (Sankyo, Japan), Neuchlonic (Taiyo, Japan), Nitepam (USV, USA), Nitrados (Berk, UK), Nitrempax (Lafi, Brazil), Novanox (Pfleger, FRG), Numbon (Ikapharm, Israel), Ormodon (Ormed, South Africa), Pacisyn, Paxisyn (Synthetic, Finland), Pelson (Infale, Spain), Radedorm (Berlin-Chemie, GDR), Relact (Lemonier, Argentina), Remnos (DDSA, UK), Somitran (Farmos Group, Finland), Somnased (Duncan Flockhart, UK), Somnite (Norgine, UK), Somnolin (Dima, Italy), Unisomnia (Unigreg, UK).

Flunitrazepam [*1622-62-4*], 5-(2-fluorophenyl)- 1, 3- dihydro-1-methyl-7-nitro-2*H*-1,4-benzodiazepin-2-one, (**18b**), $C_{16}H_{12}FN_3O_3$, M_r 313.30, *mp* 166 – 167 °C (also 170 – 172 °C). The synthesis of this compound is analogous to that of nitrazepam [113]. Pharmacologically, flunitrazepam is one of the most active benzodiazepinones; the hypnotic effects begin with doses as low as 1 – 2 mg [115]. It is used in anesthesiology in solution form to induce sleep.

Trade Names. Flunipam (A. L., Norway), Hypnodorm (Teva, Israel; Alphapharm, Australia), Hipnosedon, Narcozep, Rohpinol, Rohypnol, Roipnol (Hoffmann-La Roche, Switzerland).

Flurazepam [*17617-23-1*], 7-chloro-1-(2-diethylaminoethyl)-5-(2-fluorophenyl)-1,3-dihydro-2*H*-1,4-benzodiazepin-2-one, (**18c**),$C_{21}H_{23}ClFN_3O$, M_r 387.89, is oily. The dihydrochloride [*1172-18-5*], $C_{21}H_{25}Cl_3FN_3O$, M_r 460.83, *mp* 215.5 – 217.5 °C (decomp.), forms light yellow crystals. Flurazepam is synthesized by the alkylation of **18** (R = Cl,

Table 4. Indications for the use of benzodiazepine hypnotics *

Compound	Indications	Remarks
Lormetazepam (Flunitrazepam) Temazepam	insomnia of shift workers and in persons changing time zone or under temporary strain outpatient dose adjustment	no abuse (alcohol, coffee) should not be given at night before driving dose should be low and given daily for no longer than 3–4 weeks dose should be gradually reduced
Triazolam	for preferred inpatients dose adjustment	risk of rebound insomnia van der Kroef syndrome
Midazolam	short sleep during diagnostic operations (gastroscopy)	intravenous administration by anesthesiologist
Flunitrazepam Nitrazepam Oxazepam	hyposomnia accompanying complex psychological problems, daytime effect also desired (sedation, anxiety relief) suitable for inpatient treatment when used in combination with other psychotropic drugs	increased accumulation, paradoxical insomnia, and delirium in old people maximum duration of administration: 3 months one dose required every second or third night, gradual cessation as desired by the patient

* General contraindications: pregnancy, myasthenia gravis.

$R^1 = H$, $R^2 = F$) with diethylaminoethyl chloride [116]. Another method uses 2-(diethylaminoacetylamino)-5-chloro-2′-fluorobenzophenone as the starting material [117]. Flurazepam can also be synthesized by the oxidative ring extension of 2-aminomethyl-5-chloro-1-(2-diethylaminoethyl)-3-(2′-fluorophenyl)indole with chromium trioxide [118].

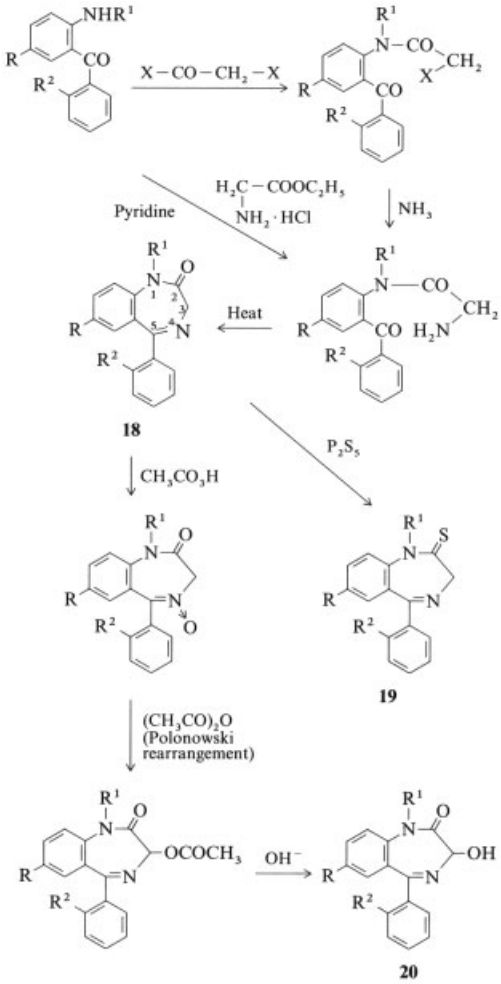

Figure 5. Synthesis of benzo-1,4-diazepinones (**18**), and their 2-thiono-(**19**) and 3-hydroxy derivatives (**20**)
Nitrazepam (**18a**): $R = NO_2$; $R^1 = H$; $R^2 = H$
Flunitrazepam (**18b**): $R = NO_2$; $R^1 = CH_3$; $R^2 = F$
Flurazepam (**18c**): $R = Cl$; $R^1 = CH_2-CH_2N(C_2H_5)_2$; $R^2 = F$
Nimetazepam (**18d**): $R = NO_2$; $R^1 = CH_3$; $R^2 = H$
Quazepam (**19**): $R = Cl$; $R^1 = CH_2-CF_3$; $R^2 = F$
Oxazepam (**20a**): $R = Cl$; $R^1 = H$; $R^2 = H$
Lorazepam (**20b**): $R = Cl$; $R^1 = H$; $R^2 = Cl$
Lormetazepam (**20c**): $R = Cl$; $R^1 = CH_3$; $R^2 = Cl$
Temazepam (**20d**): $R = Cl$; $R^1 = CH_3$; $R^2 = H$
Doxefazepam (**20e**): $R = Cl$; $R^1 = CH_2-CH_2OH$; $R^2 = F$
Cinolazepam (**20f**): $R = Cl$; $R^1 = CH_2-CH_2-CN$; $R^2 = F$

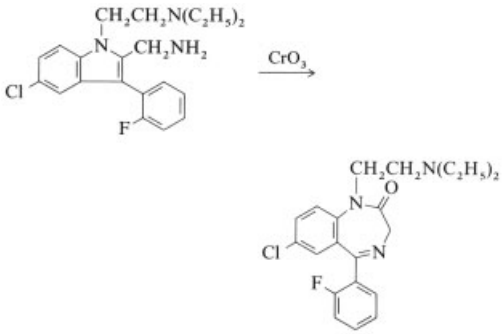

The full hypnotic response of flurazepam is evoked even after prolonged use. However, it has an extremely long plasma half-life (47 – 100 h), with long-lasting metabolites. It promotes the onset of sleep, reduces the frequency of interruptions at night, and increases the total duration of sleep. The effects of the drug begin within 20 – 30 min, and sleep usually lasts for 7 – 8 h. Flurazepam can reduce REM sleep, but does not produce a REM rebound and associated sleep disturbance on termination of treatment. This drug strongly suppresses "level 4" of orthodox sleep (see Chap. 1). It is given in doses of 30 – 60 mg [119].

Trade Names. Noctosom (Ikapharm Kfar-Sava, Israel), Staurodorm (Dolorgiet, FRG).

Flurazepam hydrochloride: Benzozil (Kyowa, Japan), Dalmadorm, Dalmane, Dalmate, Dalmene, Dormador, Dormodor (Hoffmann-La Roche, Switzerland), Felison (Sigurta, Italy), Insumin (Kyorin, Japan), Natam (Unifa, Argentina), Novoflupam (Novopharm, Canada), Remdue (Biomedica Foscama, Italy), Somnol (Horner, Canada), Som-Pam (ICN, Canada), Valdorm (Valeas, Italy).

Nimetazepam [*2011-67-8*], 1,3-dihydro-1-methyl-7-nitro-5-phenyl-2*H*-1,4-benzodiazepin-2-one (1-methylnitrazepam), (**18d**), $C_{16}H_{13}N_3O_3$, M_r295.30, *mp* 156.5 – 157.5 °C, forms light yellow crystals and is prepared analogously to nitrazepam [113], [120]. It has a long half-life and resembles nitrazepam in activity, but also has anticonvulsant and muscle relaxant properties [121].

Trade Name. Erimin (Sumitomo, Japan).

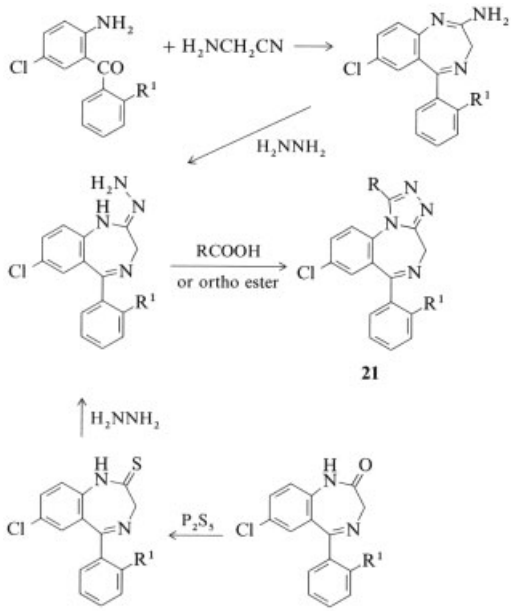

Figure 6. Synthesis of triazolobenzodiazepines (**21**)
Estazolam (**21a**): R = H; R^1 = H
Triazolam (**21b**): R = CH_3; R^1 = Cl
Midazolam (**21c**): R = CH_3; R^1 = F

Oxazepam [*604-75-1*], 7-chloro-1,3-dihydro-3-hydroxy-5-phenyl-2*H*-1,4-benzodiazepin-2-one,(**20a**), $C_{15}H_{11}ClN_2O_2$, M_r286.72, *mp* 205 –206 °C, is prepared by the oxidation of **18** (R = Cl, R^1 = H, R^2 = H) with peracetic acid, followed by the reactions described in Figure 5. Oxazepam is primarily a sedative, but also has sleep-inducing properties [122].

Trade Names. Adumbran (Thomae, FRG; Boehringer Ingelheim, FRG), Alepam (Alphapharm, Australia), Alopam (A. L., Norway), Anxiolit (Gerot, Austria; Medichemie, Switzerland), Aslapax (Asla, Spain), Benzotran (Protea, Australia), Durazepam (Durachemie, FRG), Enidrel (Syncro, Argentina), Expidet (Wyeth, UK), Isodin (Tosi, Italy), Murelax (Ayerst, Australia), Noctazepam (Brenner, Austria), Oxepam (Wyeth, UK), Oxpam (ICN, Canada), Praxiten (Wyeth, UK), Propax (Cipan, Portugal), Psicopax (Bama-Geve, Spain), Sedokin (Geymonat Sud, Italy), Senepax, Serax, Serenid (Wyeth, UK), Serepax (Wyeth, UK; Ferrosan, Denmark), Seresta (Wyeth, Netherlands), Sigacalm (Siegfried, Switzerland), Sobril (KabiVitrum,

Sweden), Uskan (Desitin, FRG), Vaben (Rafa, Israel), Zapex (Riva, Canada).

Lorazepam [*846-49-1*], 7-chloro-5-(2-chlorophenyl)-1,3-dihydro-3-hydroxy-2*H*-1,4-benzodiazepin-2-one, (**20b**), $C_{15}H_{10}Cl_2N_2O_2$, M_r 321.17, *mp* 166–168 °C, is prepared from **18** (R = Cl, R^1 = H, R^2 = Cl), analogously to oxazepam. Lorazepam is primarily an anti-anxiety drug; it also has sleep-inducing properties and is used in anesthesiology [123].
Trade Names. Ativan (Wyeth, USA), Bonton (Unipharma, Israel), Control (Sigurta, Italy), Donix (Llorens, Spain), Emotion (Alpes, Argentina), Emotival (Armstrong, Argentina), Larpose (Cipla, India), Laubeel (Desitin, FRG), Loram (Lek, Yugoslavia), Lorans (Schiaparelli, Italy), Lorax (Wyeth, UK), Lorivan (Disco, Israel), Merlit (Ebewe, Austria), Nervistop L (Ingram, Argentina), Orfidal (Orfi, Spain), Pro Dorm (Schürholz, FRG), Securit (Perrel, Italy), Sedatival (Raffo, Argentina), Tavor (Wyeth, FRG), Temesta (Ferrosan, Denmark; Wyeth, Netherlands), Tolid (Dologiet, FRG), Trapax (Wyeth, Argentina), Wypax (Yamanouchi, Japan).

Trimethylacetate: Drupal (Novag, Spain), Piralone (Ferrer, Spain), Placinor (Robert, Spain).

Lormetazepam [*848-75-9*], 7-chloro-5-(2-chlorophenyl)-1,3-dihydro-3-hydroxy-1-methyl-2*H*-1,4-benzodiazepin-2-one (*N*-methyllorazepam), (**20c**), $C_{16}H_{12}Cl_2N_2O_2$, M_r335.19, *mp* 192–194 °C, is prepared analogously to oxazepam [124]. The hypnotic effect of 1–2 mg is similar to that of nitrazepam [125].
Trade Names. Lembrol (Vinas, Spain), Loramet (Wyeth, Belgium, Netherlands), Minias (Farmades, Italy), Noctamid, Pronoctan (Schering, FRG), Nocton (Eurolab, Chile).

Temazepam [*846-50-4*], 7-chloro-1,3-dihydro-3-hydroxy-1-methyl-5-phenyl-2*H*-1,4-benzodiazepin-2-one (3-hydroxydiazepam, *N*-methyloxazepam, oxydiazepam), (**20d**), $C_{16}H_{13}ClN_2O_2$, M_r 300.74, *mp* 119–121 °C, prepared analogously to oxazepam [126]. It is used as a hypnotic drug and should be taken 1–2 h before going to sleep [127].
Trade Names. Euhypnos (Farmitalia Carlo Erba, Italy; Sigma, Australia), Levanxene (Farmitalia Carlo Erba, Italy; Aesca, Austria), Mabertin (Sidus, Argentina), Normison (Wyeth, UK), Planum (Farmitalia Carlo Erba, Italy), Remestan (Wyeth, FRG), Restoril (Sandoz, Switzerland; Anca, Canada), Somaz (Quantum, USA), Tenso (Castejon, Spain), Veroqual (Lek, Yugoslavia).

Doxefazepam [*40762-15-0*], 7-chloro-5-(2-fluorophenyl)-1,3-dihydro-3-hydroxy-1-(2-hydroxyethyl)-2*H*-1,4-benzodiazepin-2-one, (**20e**), $C_{17}H_{14}ClFN_2O_3$, M_r348.57, *mp* 138–140 °C, is prepared by the alkylation of 7-chloro-5-(2-fluorophenyl)-1,3-dihydro-2*H*-1,4-benzodiazepin-2-one *N*-oxide with 2-bromoethanol or 2-bromoethyl acetate, followed by Polonowski rearrangement and saponification [128], [129]. Doxefazepam is a hypnotic drug with two to four times the activity and half the toxicity of flurazepam; it causes significantly less hangover [130].
Trade Name. Doxans (Schiaparelli, Italy).

Cinolazepam [*75696-02-5*], 7-chloro-5-(2-fluorophenyl)-2,3-dihydro-3-hydroxy-2-oxo-1*H*-1,4-benzodiazepine-1-propionitrile, (**20f**),$C_{18}H_{13}ClFN_3O_2$, M_r356.84, *mp* 190–193 °C, is prepared by the cyanoethylation of compound **20** (R = Cl, R^1 = H, R^2 = F) in the presence of trimethylammonium hydroxide [131]. This drug resembles doxefazepam in action [132].

Quazepam [*36735-22-5*], 7-chloro-5-(2-fluorophenyl)-1-(trifluoro-2,2,2-ethyl)-1,3-dihydro-2*H*-1,4-benzodiazepine-2-thione, (**19**), $C_{17}H_{11}ClF_4N_2S$, M_r386.6, *mp* 138–139 °C, is synthesized by the reaction of compound **18** (R = Cl, $R^1 = CH_2CF_3$, R^2 = F) with phosphorus pentasulfide [133]. Quazepam is a sleep-inducing drug; it has a long duration of action and causes a hangover. It also produces long-acting metabolites in the body [134].
Trade Names. Dormalin (Schering Corp., USA), Prosedar, Selepam, Cetrane (Schering Plough, USA).

Haloxazolam [*59128-97-1*], 10-bromo-11*b*-(2-fluorophenyl)-2,3,7-11*b*-tetrahydrooxazolo-[3,2-*d*][1,4]benzodiazepin-6(5*H*)-one,$C_{17}H_{14}BrFN_2O_2$, M_r377.22, *mp* 185 °C, is synthesized by the reaction of bromo- (or tosyloxy-) acetamidobenzophenone

with ethanolamine, followed by cyclization in the presence of acetic acid [135].

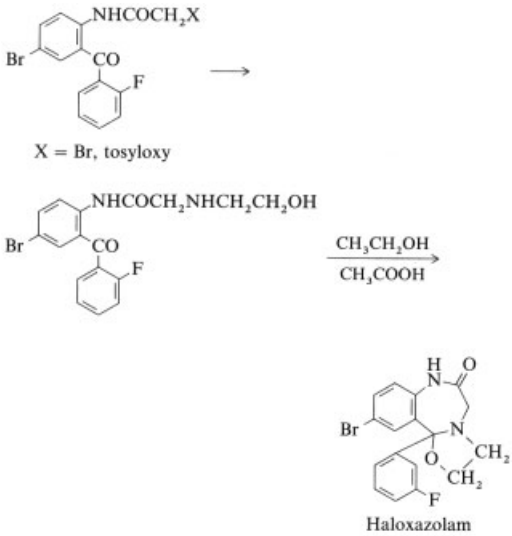

Haloxazolam, oxazolam, and cloxazolam belong to a group of derivatives with an oxazoline ring at the 4,5-position of the benzodiazepine ring, which causes a change in some biological parameters. Haloxazolam is primarily used as a hypnotic drug in the treatment of insomnia [136].

Trade Name. Somelin (Sankyo, Japan).

Estazolam [*29975-16-4*], 8-chloro-6-phenyl-4*H*-[1,2,4]triazolo-[4,3-*a*][1,4]benzodiazepine, (**21a**), $C_{16}H_{11}ClN_4$, M_r 294.75, *mp* 228 – 229 °C, is prepared as shown in Figure 6 [137], [138]. Indications are problems in falling asleep, insomnia, and sleep induction in anesthesiology [139]. The metabolism of the triazolobenzodiazepines differs from that of the benzodiazepinones in that the triazole ring remains intact in most of the metabolites. One of the first transformations in the metabolism of benzodiazepines is the cleavage of the substituent at position 1.

Trade Names. Domnamid (Lundbeck, Denmark), Esilgan (Amer. Cyanamid, USA; Takeda, Japan), Eurodin (Takeda, Japan), Nuetalon (Casenne, France), Tasedan (Takeda, Mexico), Somnatrol (Abbott, Peru), Noctal (Abbott, Brazil).

Triazolam [*28911-01-5*], 8-chloro-6-(2-chlorophenyl)-1-methyl-4*H*-[1,2,4]triazolo[4,3-*a*][1,4]-benzodiazepine, (**21b**), $C_{17}H_{12}Cl_2N_4$, M_r 343.22, *mp* 223 – 225 °C, is prepared as shown in Figure 6 [138], [140]. Triazolam is an extremely strong, but short-acting hypnotic drug, the hypnotic dose being 0.5 mg (equivalent to 30 mg of flurazepam) [141].

Doses of 0.125 – 1 mg are effective in inducing sleep [142]. The maximum effect is produced 1 – 2 h after administration and lasts up to 8 h; it is totally eliminated after 24 h. As with nitrazepam and flunitrazepam, withdrawal after prolonged use causes rebound insomnia [143]. Studies question the occurrence of the rebound effect when standard doses (0.125 – 0.25 mg) are taken [144]. For tolerance studies, see [145] and for a review [146].

Trade Names. Halcion (Upjohn, USA), Novidorm (Syntial, Argentina), Novodorm (Rubio, Spain), Nuctane (Bago, Argentina).

Midazolam [*59467-70-8*], 8-chloro-6-(2-fluorophenyl)-1-methyl-4*H*-imidazo[1,5-*a*][1,4]benzodiazepine,(**21c**), $C_{18}H_{13}ClFN_3$, M_r 325.77, *mp* 158 – 160 °C (maleate, *mp* 114 – 117 °C), is synthesized as shown in Figure 6 [147]. The presence of the imidazole ring ensures a stable solution, a short duration of action, and the formation of water-soluble injectable salts. Below pH 4, the diazepine ring is opened between positions 4 and 5 to give a water-soluble salt of the primary amine. Ring closure takes place above pH 4 with a half-life of 10 min (**22** → **21c**). Indications are premedication before operations and anesthesia [148].

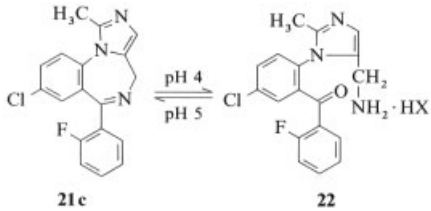

Trade Names. Dormicum, Dormonil, Hypnovel, Sorenor (Hoffmann-La Roche, Switzerland), Flormidal (Galenika, Yugoslavia).

Brotizolam [*54801-81-7*], 2-bromo-4-(2-chlorophenyl)-9-methyl-6*H*-thieno[3,2-*f*][1,2,4]triazolo[4,3-*a*][1,4]diazepine, (**24**), $C_{15}H_{10}BrClN_4S$, M_r 393.7, *mp* 212 – 214 °C, is synthesized by the condensation of 2-amino-3-(2-chlorobenzoyl)-thiophene (**23**), with aminoacetic ester, bromination, and fusion of the triazole ring [149].

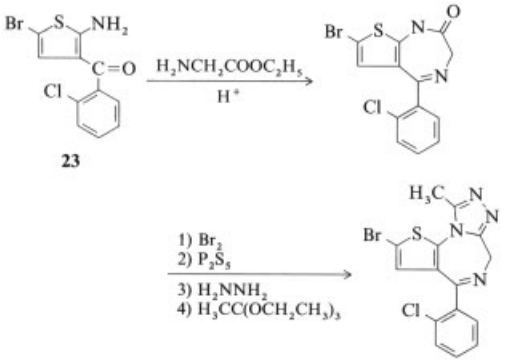

Brotizolam is very effective in prolonging sleep and is an especially useful hypnotic for light to moderate insomnia [150].

Trade Names. Ladormin, Lendorm, Lendormin (Boehringer, FRG).

Loprazolam [*61197-93-1*], 6-(2-chlorophenyl)-2,4-dihydro-2-[(4-methyl-1-piperazinyl)methylene]- 8- nitro- 1*H*- imidazo[1,2-*a*][1,4]benzodiazepin-1-one, (**25**), $C_{23}H_{21}ClN_6O_3$, M_r 464.91, *mp* 214 – 215 °C, (methanesulfonate, M_r 560.71, *mp* 205 – 210 °C), is synthesized by the condensation of the appropriate benzodiazepinethione with glycine, followed by cyclization. Condensation with dimethylformamide acetal and subsequent aminolysis with *N*-methylpiperazine gives loprazolam (Fig. 7) [151].

This drug is an effective, medium-acting hypnotic and is especially suitable for the short-term treatment of insomnia. It is also prescribed to promote the onset of sleep and prevent frequent awakening during the night [152].

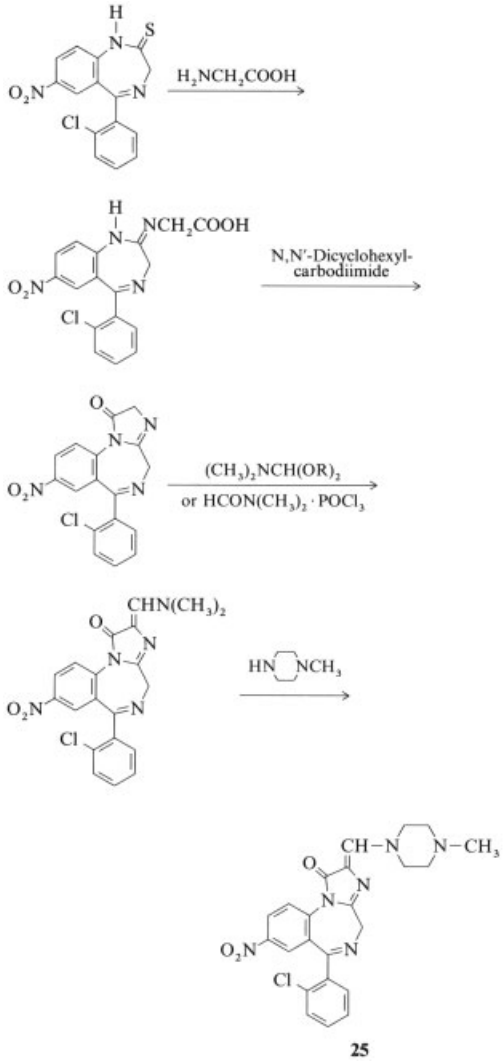

Figure 7. Synthesis of loprazolam (**25**)

Trade Names. Dormonoct (Roussel Lab., UK), Havlane (Diamant, France).

Rilmafazone [*99593-25-6*], 5-[(2-aminoacetamido)methyl]- 1- [4- chloro-2(2-chlorobenzoyl)-phenyl]- *N, N*- dimethyl- 1*H*-1,2,4-triazole-3-carboxamide, (**26**), $C_{21}H_{20}Cl_2N_6O_3 \cdot HCl \cdot 2\,H_2O$, M_r 547.82, *mp* 107 °C, is prepared from 2′,5-dichloro-2-aminobenzophenone as described in [153] (Fig. 8). Rilmafazone belongs to a group of open-ring benzodiazepinone analogs called peptide aminobenzophenones. It appears to be a precursor of the triazolobenzodiazepinones. It has both sedative

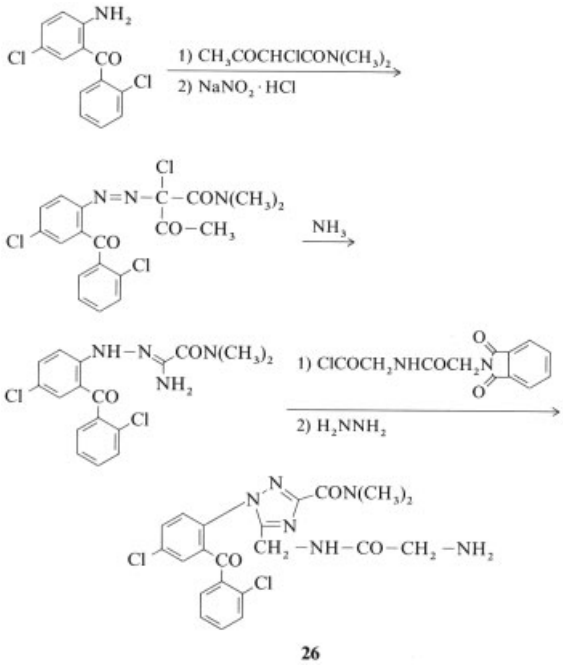

Figure 8. Synthesis of rilmafazone (**26**).

and antianxiety effects. Doses of 1 – 2 mg promote the onset of sleep. Indication is neurotic insomnia. This drug was introduced in 1988 (Shionogi, Japan) [154].

6. Other Hypnotics

Other substances interact with the benzodiazepine receptors in vitro and in vivo. They resemble benzodiazepines in general activity, with similar sedative, anti-anxiety, and hypnotic properties, but differences in adverse effects exist. The practical value of these drugs, which include zopiclone, suriclone (anxiety-relieving), zolpidem, and alpidem, will become apparent on extended use.

Zopiclone [*43200-80-2*], 6(5-chloro-2-pyridyl)-5-(4-methylpiperazin-1-yl)carbonyloxy-7- oxo- 6,7- dihydro- 5*H* - pyrrolo[3,4-*b*]pyrazine, (**29**), $C_{17}H_{17}ClN_6O_3$, M_r388.82, *mp* 178 °C, is synthesized by the condensation of 2,3-piperazinedicarboxylic acid anhydride with 2-amino-5-chloropyridine to give compound **27**, which is partially reduced to compound **28** and subsequently esterified with 4-methyl-1-piperazinyl chloride [155].

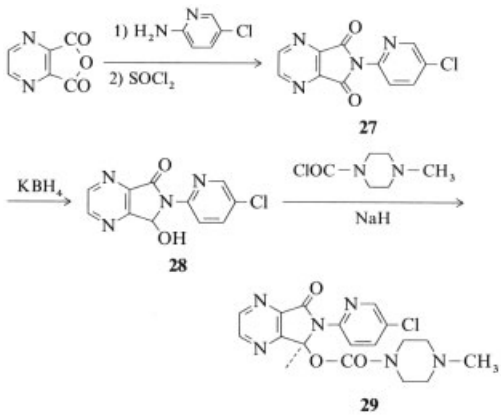

A dose of 7.5 mg is required for the treatment of insomnia, which is equivalent to the effect produced by 5 – 10 mg of nitrazepam. The adverse effects are similar to those produced by short-acting benzodiazepines, but zopiclone causes less hangover [156]. Zopiclone is, for the most part, converted to inactive metabolites. The elimination halftime for zopiclone and its active

N-oxido metabolite (11 % formation) is 3.5 – 6 h [157].
Trade Names. Imovane, Imovance, Theraplix (Rhône-Poulenc, France), Amoban, Amovane (Rhône-Poulenc, France).

Zolpidem [*82626-48-0*], 2-(4-methylphenyl)-6-*N,N*- trimethylimidazo- [1,2-*a*]pyridine-3-acetamide-L-(+)-tartrate (2 : 1) [*99294-93-6*], (**30a**), $(C_{19}H_{21}N_3O)_2 \cdot C_4H_6O_6$, M_r 764.88, is prepared as described in [158]. This short-acting hypnotic is rapid in onset; it reacts specifically with the benzodiazepine type 1 receptor [159]. It is registered as Stilnox (Synthelabo, France) and was put on the market in 1988 (licenced by Searle).

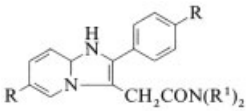

30

30a: R = CH_3, R^1 = CH_3 = Zolpidem
30b: R = Cl, R^1 = C_3H_7 = Alpidem

Sulfones. The first synthetic hypnotics to be introduced were sulfones (sulfonal and trional). These sedatives and hypnotics are no longer used because of their high toxicity and severe adverse effects.

Antiepileptic Agents. Chlormethiazole [*53-45-9*], 5-(2-chloroethyl)-4-methylthiazole, is an antiepileptic drug occasionally used as a hypnotic.
Trade Name. Distraneurin (Pharma Stern, FRG).

Fenadiazole [*1008-65-7*], 2-(2-hydroxyphenyl)-1,3,4-oxadiazole, $C_8H_6N_2O_2$, M_r 140.14, *mp* 111 – 112 °C, is prepared by the condensation of salicylhydrazide with ethyl orthoformate [161]. The LD_{50} is 940 mg/kg (mouse, i.p.).

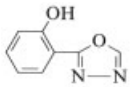

Trade Name. Viodor (Violani-Farmavigor, Italy).

7. References

General References

1. AMA Drug Evaluations: *Drugs Used for Anxiety and Sleep Disorders,* 6th ed., Chicago 1986, pp. 81 – 110.
2. J. A. Vida in W. O. Foye (ed.): "Central Nervous System Depressants: Sedative-Hypnotics," *Principles of Medicinal Chemistry,* Lea u. Febinger, Philadelphia 1976, pp. 155 – 183.
3. J. A. Vida in A. Burger (ed.): "Sedative-Hypnotics," *Medicinal Chemistry,* 4th ed., Part III, New York, Chichester, Brisbane, Toronto, 1980, pp. 787 – 828.
4. K. W. Wheeler: "Non-barbiturate Hypnotics," *Med. Chem.* (Academic) 6 (1963) 1 – 245.
5. W. J. Doran: "Barbituric Acid Hypnotics," *Med. Chem.* (Academic) 4 (1959) 1 – 334.
6. J. Hoffmann, B. Mencke, in G. Ehrhardt, H. Ruschig (eds.): "Hypnotika," *Arzneimittel,* 2nd ed., **vol. 1,** Verlag Chemie, Weinheim 1966 pp. 37 – 66.
7. S. Ebel: "Hypnotika, Sedativa," in *Synthetische Arzneimittel,* Verlag Chemie, Weinheim 1979, pp. 109 – 159.
8. E. Mutschler: *Arzneimittel-Nebenwirkungen, Hypnotika,* Wissenschaftl. Verlags GmbH, Stuttgart 1972, pp. 68 – 77.
9. S. C. Harvey in L. S. Goodman, A. Gilman, A. G. Gilman, G. B. Koelle (eds.): "Hypnotics and Sedatives," *The Pharmacological Basis of Therapeutics,* 7th ed.,Macmillan Publ., New York, Collier Macmillan, Toronto, Ballière Tindall, London 1985, pp. 339 – 371.
10. J. Oswald in E. Jucker (ed.): "Drug Research and Human Sleep," *Arzneimittelforschung,* **vol. 22,** Birkhäuser, Basel 1978, pp. 355 – 372.
11. M. Monnier et al., "Biology of Sleep. An Interdisciplinary Survey," *Experientia* **36** (1980) 1 – 142.
12. J. P. Olié, S. Lepastier, G. Dana, J. M. Birée, "Medicaments hypnotiques," *Med. Actuelle* **5** (1978) 53 – 61; *Chem. Abstr.* **88** (1978) 145 812.
13. M. D. Allen, D. J. Greenblatt, *Meyler's Side Eff. Drugs* **4** (1980) 29 – 36; **3** (1979) 29 – 38; **2** (1978) 36 –44; **1** (1977) 21 – 29; VIII 1972 – 1975, 64 – 83, Excerpta Medica, Amsterdam – Oxford.
14. H. P. Büch: "Medikamentöse Behandlung von Schlafstörungen und Erregungszuständen, Pharmakologie der Sedativa/Hypnotika," *Dtsch. Apoth. Ztg.* **121** (1981) 1951 – 1959.

15. A. A. Borbély: "Schlafmittel und Schlaf: übersicht und therapeutische Richtlinien," *Ther. Umschau* **43** (1986) 509 – 516.
16. E. Soos: "Schlaf und Schlafmittel," *OAZ Österr. Apoth. Ztg.* **28** (1974) 93 – 101.
17. A. N. Nicholson: "Hypnotics Their Place in Therapeutics," *Drugs* **31** (1986) 164 – 171.
18. E. A. Swinyard: *Sedatives und Hypnotics in Remington Pharmaceutical Sciences,* 7th ed., Mack Publ., Easton, Pa., 1985, pp. 1059 – 1074.
19. D. Schneider-Helmert: Schlafmittel: "Symptomatisch oder kurativ?" *Schweiz. ärzteztg.* **67** (1986) 1477 – 1483.
20. W. P. Koella: "The Organization and Regulation of Sleep," *Experientia* **40** (1984) 309 – 338.
21. W. Keup: "Mißbrauch und Abhängigkeit von Sedativa und Hypnotika," *Internist* **27** (1986) 746 – 756.
22. K. Laemmel: "Der Schlaf-Seismograph von Körper und Seele," *Schweiz. Rundsch. Med. Prax.* **74** (1974) 1305 – 1310.
23. A. Borbély: *Das Geheimnis des Schlafs,* Deutsche Verlagsanstalt GmbH, Stuttgart 1984.
24. T. Zilker, M. Clarmann: "Schlafmittelvergiftungen," *MMW Münch. Med. Wochenschr.* **126** (1984) 297 – 300.
25. V. B. Meyer-Rochow: "Vom Winterschlaf lernen," *Selecta* 1984, 1116 – 1120.
26. *Drug, Facts and Comparisons,* Facts and Comparisons Div., St. Louis 1982, 976 – 1016 (a division of J. B. Lippincott Co., Philadelphia-Toronto).
27. Martindale, *The Extra Pharmacopoeia, Hypnotic and Sedatives,* 28th ed., Pharmaceutical Press, London 1982, pp. 792 – 817.
28. E. A. Swinyard in *Remington's Pharmaceutical Sciences,* 17th ed., Mack Publ. Easton, Pa., 1985, 1059 –1074.

Specific References

29. A. A. Borbély, *J. Neural. Trans. Suppl.* **21** (1986) 243 – 254.
30. S. H. Reppert et al., *Science (Washington, D.C.)* **242** (1988) 78 – 81.
31. M. Monnier et al., *Experientia* **33** (1977) 548. A. E. Kornmüller, H. B. Lux, K. Wink, M. Klee, *Naturwissenschaften* **48** (1961) 503 – 505. D. Schneider-Helmert: "Die körpereigene Schlafsubstanz DSIP," *Schweiz. Rundsch. Med.* **74** (1985) 1311 – 1315. M. V. Graf, A. J. Kastin: "Delta-Sleep-Inducing Peptide (DSIP): An Update," *Peptides (N.Y.)* **7** (1986) 1165 – 1187; *Neurosci. Biobehav. Rev.* **8** (1984) 83 – 93.
32. A. A. Borbély, G. Youmbi-Balderer: "Ist Tryptophan ein biologisches Schlafmittel," *Schweiz. Rundsch. Med. Prax.* **75** (1986) 785 – 790.
33. J. P. Olié, S. Lepastier, S. Dana, J. M. Birée, *Med. Actuelle* **5** (1978) 53.
34. A. A. Borbély, *Ther. Umsch.* **43** (1986) 509.
35. M. B. Scharf, E. O. Bixler, A. Kales, C. R. Soldatos, *Pharmacology* **19** (1979) 173 – 181. A. Kales et al., *Pharmacology* **26** (1983) 138 – 149.
36. K. Laemmel, *Schweiz. Rundsch. Med. Prax.* **74** (1985) 1305 – 1310.
37. E. Greiser: Bewertender Arzneimittelindex: Hypnotika, vol. 2, Med. Pharm. Verlag, Wiesbaden 1983, pp. 1 – 25.
38. J. Krieglstein, D. Grusla, *Dtsch. Apoth. Ztg.* **128** (1988) 2041.
39. E. F. Domino, *Annu. Rev. Pharmacol.* **2** (1962) 215.
40. F. Hunzicker, F. Künzle, J. Schmitz, *Helv. Chim. Acta* **49** (1966) 1 433. Wander, GB 1 006 156, 1965; US 3 389 139, 1968.
41. G. Stille, A. Sayers, H. Lauener, E. Eichenberger, *Psychopharmakologia* **28** (1973) 325; *Chem. Abstr.* **78** (1973) 131 982 c. Y. Take et al., *Takeda Kenkyushoho* **29** (1970) 416; *Chem. Abstr.* **74** (1971) 21 815 n. H. Yokotani et al., *Takeda Kenkyushoho* **29** (1970) 441; *Chem. Abstr.* **74** (1971) 21 816 p.
42. W. M. McLamre, S. Y. P. An, A. Barley, *J. Org. Chem.* **20** (1955) 109. Pfizer, US 2 746 900, 1953.
43. Smith, Kline and French, DE-AS 1 183 333, 1963.
44. H. C. Caldwell, H. J. Adams, D. E. Rivard, J. V. Swintovsky, *J. Pharm. Sci.* **56** (1967) 920.
45. British Drug Houses, GB 874 246, 1959.
46. Am. Home Prod., US 2 784 237, 1957.
47. Danske Medicinal, US 2 931 838, 1960.
48. A. Pictet, F. H. Reichel, *Helv. Chim. Acta* **6** (1923) 621. W. Freudenberg, A. M. Vajda, *J. Am. Chem. Soc.* **59** (1937) 1955.
49. C. Bischoff, *Ber. Dtsch. Chem. Ges.* **7** (1874) 628.
50. Rheinpreussen, DE-AS 1 021 843, 1951.
51. Schering, US 2 816 910, 1953, Dtsch. Prior. 1952.
52. P. Läuger, M. Prost, R. Charlier, *Helv. Chim. Acta* **42** (1959) 2379. Labaz, US 2 931 828, 1960.

53. Carlo Erba, GB 736 340, 1955.
54. G. Newbery, *J. Chem. Soc.* **127** (1925) 295. Miles, US 2 854 379, 1958; US 2 931 832, 1960.
55. Bayer, DE 225 710, 1909. *Friedländer* **10,** 1 160.
56. Bayer, DE 286 760, 1913. *Friedländer* **12,** 704. Bayer, DE 327 129, 1917. *Friedländer* **13,** 809.
57. Knoll, DE 185 962, 1906. *Friedländer* **8,** 1217; US 914 518, 1909.
58. *Merck Index* **10** (1983) 773. E. Schröder, C. Rufer, R. Schmiechen: *Pharmazeutische Chemie,* Thieme Verlag, Stuttgart – New York 1982, pp. 251 – 252.
59. Bayer, DE 249 241, 1912. *Friedländer* **10,** 1 164. Labs. Aub., CH 374 644, 1964.
60. Millmaster Onyx, US 3 282 998, 1966. E. H. Volwiler, D. L. Tabern, *J. Am. Chem. Soc.* **58** (1936) 1 352.
61. M. Bockmühl, O. Schaumann, *Dtsch. Med. Wochenschr.* **54** (1928) 270.
62. Lion Dentifrice (Tokio), DE-OS 1 768 445, 1967; US 3 639 457, 1967; DE-OS 2 753 792, 1977.
63. N. Okudeira, S. Torii, S. Endo, *Psychopharmacology* **70** (1980) 117. K. Hillier, *Drugs Future* **6** (1981) 536 – 537.
64. W. J. Doran in F. F. Blicke, R. H. Cox (eds.): "Barbituric Acid Hypnotics," *Med. Chem. Ser. Monogr.* **4** (1959).
65. R. Y. Levina, F. K. Velichko: "Advances in the Chemistry of Barbituric Acids," *Russ. Chem. Rev. (Engl. Transl.)* **29** (1960) 437 – 459.
66. *Kirk-Othmer,* 2nd ed., **3,** 60 – 77.
67. A. Schlieper, *Justus Liebigs Ann. Chem.* **55** (1845) 251; **56** (1845) 1.
68. R. Gross, *Internist* **17** (1976) 229.
69. R. J. Levina, F. K. Velichko: "Advances in the Chemistry of Barbituric Acids," *Russ. Chem. Rev.* (Engl. Transl.) 29 (1960) 437.
70. E. Fischer, A. Dilthey, *Justus Liebigs Ann. Chem.* **335** (1904) 338. Merck, DE 146 496, 1903.*Friedländer* **7,** 651.
71. Bayer, DE 247 952, 1911, *Friedländer* **11,** 926.
72. Poulenc Frères, US 1 609 520, 1926.
73. Lambert Thorpe Co., US 1 255 951, 1917; US 1 576 014,1926.
74. E. Leyraud, GB 191 008, 1922. Eli Lilly, US 1 514 573, 1926; US 1 856 792, 1932.
75. H. H. Frey, M. P. Magnussen, *Arzneim. Forsch.* **16** (1966) 612.
76. Bayer, DE 442 655, 1924; DE 231 150, 1924.
77. Hoffmann-La Roche, US 1 444 802, 1923.
78. Ciba, DE 268 158, 1911.
79. Hoffmann-La Roche, DE 526 854, 1930.
80. M. Seefelder, *Justus Liebigs Ann. Chem.* **652** (1962) 107. BASF, Ludwigshafen, *Festschr. Carl Wurster 60. Geburtstag* 1960, 71; *Chem. Abstr.* **56** (1962) 9 927 i; BASF FP 869 Méd. 1961.
81. Geigy, FR 870 714, 1941; DE 756 489, 1941.
82. Eli Lilly, US 1 856 792, 1932.
83. Bayer, DE 293 163, 1915. Geigy, GB 650 354, 1948.
84. J. D. Riedel AG, US 1 739 662, 1929.
85. A. C. Cope, E. M. Hancock, *J. Am. Chem. Soc.* **61** (1939) 776. Sharp & Dohme, US 2 187 703, 1940.
86. E. H. Volwiler, *J. Am. Chem. Soc.* **47** (1925) 2236.
87. Compagnie de Bethune, DE 589 947, 1930. *Friedländer* **19,** 1 198.
88. J. D. Riedel AG, US 1 622 129, 1927; DE 481 733, 1923.
89. Eli Lilly, US 1 954 429, 1934.
90. O. Schneider, H. Frick, A. H. Lutz, *Experientia* **10** (1954) 135. Festschrift F. Reinhardt AG, Basel 1936, p. 195. Hoffmann-La Roche, US 2 151 047, 1938; US 2 680 116, 1954.
91. Ciba-Geigy, US 2 673 205, 1954, Schweiz.-Prior. 1951; DE 950 193, 1952. L. Hoffmann, E. Tagmann, *Helv. Chim. Acta* **35** (1952) 1 541.
92. Richardson-Merrell Inc. BE 753 846, 1971; DE-AS 2 035 636, 1969.
93. H. Koch, J. Kotlan, *Monatsh. Chem* **97** (1966) 1 648;Kwizda, DE-AS 1 695 346, 1966.
94. L. Stockinger, H. Koch, *Arzneim. Forsch.* **19** (1969) 167.
95. G. Argyropoulos, G. Lanner, E. Pasquali, *Wien. Med. Wochenschr.* **126** (1976) 75; *Med. Actual.* **16** (1980) 388.
96. K. H. Boltze et al., *Arzneim. Forsch.* **13** (1963) 688.
97. B. Dubnick, C. A. Towne in A. Cerletti, F. J. Bove (eds.): "The Present Status of Psychotropic Drugs," *Proc. of the VI. Internat. Congr. of the Coll. Internat. Neuro-Psychopharmacologicum,* Tarragona Apr. 1968 pp. 24 – 27; Excerpta Med. 1969, 77 – 83.
98. I. K. Kacker, S. H. Zaheer, *J. Indian Chem. Soc.* **28** (1951) 344; *Chem. Abstr.* **46** (1952) 11 210. General Aniline u. Film Corp., US 3 213 094, 1963. Lab. Toraude, GB 843 073, 1960.

99. Beiersdorf AG, BE 615 282, 1962; *Chem. Abstr.* **58** (1963) 13 971 ef. S. S. Parmar, K. Kishok, P. K. Seth, R. C. Arona, *J. Med. Chem.* **12** (1969) 138.
100. J. Klosa, *J. Prakt. Chem.* **14** (1961) 84; **20** (1963) 283.
101. E. Szirmai, *Ther. Umsch.* **20** (1963) 542.
102. L. H. Sternbach: "The Benzodiazepine Story," *Progr. Drug Res.* **22** (1978) 229 – 266. W. Schallek, W. D. Horst, W. Schlosser: "Mechanism of Action of Benzodiazepines," *Adv. Pharmacol. Chemother.* **16** (1979) 45 – 87. L. O. Randall, W. Schallek, L. H. Sternbach, R. Y. Ning in M. Gordon (ed.): "Chemistry and Pharmacology of the 1,4-Benzodiazepines," *Psychopharmacological Agents, A Series of Monographs in Medicinal Chemistry,* **vol. 3,** Academic Press, New York – San Francisco – London 1974, p. 175. V. C. Williams, G. C. Varnado, P. U. Nwangwu: "Perspectives on the Clinical Pharmacology of Benzodiazepines," *Med. Actual.* **21** (1985) 75 – 96. T. A. Hamor, J. L. Martin in G. P. Ellis, G. B. West (eds.): "The Benzodiazepines," *Progress in Medicinal Chemistry* **20** (1985) 158 – 223. A. Vela-Bueno, A. Kales: "Benzodiazepine Hypnotics in the Multidimensional Treatment of Insomnia," *Med. Actual.* **22** (1986) 271 – 281. C. Pina, A. Catedra: "Newer Benzodiazepines, Med. Actual. 11 (1975) 176; "Chemical Synthesis of Benzodiazepines (nur Literatur)," Med. Actual. 11 (1975) 211 – 212. H. Schütz: *Benzodiazepines: A Handbook,* Springer Verlag, Berlin 1982. C. Bellantuono, V. Reggi, G. Tognoni, S. Garattini: "Benzodiazepines – Clinical Pharmacology and Therapeutic Use," *Drugs* **19** (1980) 195 – 219. S. Garattini: *The Benzdiazepines,* Raven Press, New York 1973.
103. M. Lader (ed.): "New Perspectives in Benzodiazepine Therapy," *Arzneim. Forsch.* **30** (1980) no. 5 a, 851 – 913. U. Klotz: "Clinical Pharmacology of Benzodiazepines," in Progress in Clinical Biochemistry and Medicine, vol. 1, Springer, Berlin – Heidelberg – New York – Tokyo 1984, pp. 117 – 166.
104. H. Schrum: "Benzodiazepine als Schlafmittel," *Schweiz. Rundsch. Med. Prax.* **74** (1985) 1301 – 1304.
105. R. Weber, *Dtsch. Apoth. Ztg.* **128** (1988) 2643 – 2652.
106. L. L. Iversen, *Nature (London)* **275** (1978) 447. R. E. Study, J. L. Barker, *JAMA J. Am. Med. Assoc.* **247** (1982) 2147 – 2151. U. Klotz: *Progress in Clinical Biochemistry and Medicin,* Springer, Berlin – Heidelberg – New York – Tokyo 1984, pp. 127 – 129. T. A. Hamor, J. L. Martin in G. P. Ellis, G. B. West (eds.): *Progress in Medicinal Chemistry* **20** (1985) p. 183.
107. W. E. Müller, *MMW Münch. Med. Wochenschr.* **124** (1982) 879 – 884; *Dtsch. Apoth. Ztg.* **126** (1986) 1036 – 1044.
108. W. E. Müller, *Med. Actual.* **24** (1988) 649 – 663.
109. E. Müller, *Dtsch. Apoth. Ztg.* **128** (1988) 672 – 674. *Roche Magazin* 1988, no. 31, 3 – 13 (Hoffmann-La Roche).
110. S. Ebel, H. Schütz: "Biotransformation von Benzodiazepin-Derivaten," *Pharm. Unserer Zeit* **8** (1979) 87. U. Klotz, L. Kangas, J. Kanto: "Clinical Pharmacocineties of Benzodiazepines," Prog. Pharmacol. 3 (1980) 1 – 72. W. E. Müller: "über die Bedeutung der Plasmaproteinbindung für Verteilung und Wirkung von Tranquillantien vom Benzodiazepintyp," *Klin. Wochenschr.* **55** (1977) 105 – 110.
111. R. G. Borland, A. N. Nicholson, *Br. J. Clin. Pharmacol.* **2** (1975) 9. S. C. Harvey in Goodman Gilman, *The Pharmacological Basis of Therapeutics,* 7th ed., Macmillan Publ., New York, Toronto, London 1985, p. 352; see also [102] p. 82.
112. Hoffmann-La Roche, DE-AS 1 136 709, 1959; DE-AS 1 145 626, 1959; US 3 109 843, 1961; US 3 116 203, 1962; US 3 123 529, 1964. Delmar Chem., DE-AS 1 811 785, 1968.
113. L. H. Sternbach et al., *J. Med. Chem.* **6** (1963) 261. Hoffmann-La Roche, US 3 116 203, 1963; US 3 123 529, 1964; US 3 203 900, 1964; GB 1 039 372, 1966.
114. M. W. Johns, Drugs 9 (1975) 448. J. Rieder, *Arzneim. Forsch.* **23** (1973) 212. Martindale: *The Extrapharmacopoea,* 28th ed., Pharmaceutical Press, London 1982, pp. 807 –809. *Med. Actual.* **1** (1964) 132 – 138; **9** (1973) 237.
115. J. M. Monti, H. Altier, *Psychopharmacologia* **32** (1973) 343. W. Schallek, J. Kovacs, A. Kuehn, J. Thomas, *Arch. Int. Pharmacodyn. Ther.* **206** (1973) 161. S. A. Kaplan et al., *J. Pharm. Sci.* **63** (1974) 527. R. M. Joye, A. J. Hance, K. F. Killam, *Neuropharmacology* **10** (1971) 483. J. Freuchen, J. Østergaard, B. O. Mikkelsen, *Curr. Ther. Res.* **23** (1978) 90. R. T. Owen, *Med. Actual* **19** (1983) 163 – 167;

11 (1975) 176 – 178. M. A. K. Mattila, H. M. Larni, *Drugs* **20** (1980) 353 – 374.
116. L. H. Sternbach et al., *J. Med. Chem.* **8** (1965) 815. Hoffmann-La Roche, BE 629 005, 1963; US Prior. 1962; *Chem. Abstr.* **60** (1963) 1 589 f; NL 6 401 335, US Prior. 1963; *Chem. Abstr.* **62** (1965) 5288.
117. Hoffmann-La Roche, US 3 567 710, 1971.
118. S. Inaba, K. Ishizumi, H. Yamamoto, *Chem. Pharm. Bull.* **19** (1971) 263.
119. L. O. Randall et al., *Arch. Int. Pharmacodyn. Ther.* **178** (1969) 216. H. Jick, *Curr. Therap. Res* **9** (1967) 355. J. d'Armagnac et al., *Therapie* **26** (1971) 439. A. Schwartz, F. M. Vane, E. Postma, *J. Med. Chem.* **11** (1968) 770. Med. Actual. **6** (1970) 217. USP XXI, 447 – 449, Rockville, Md. 1985. AMA: *Drug Evaluations,* 6th ed., Am. Med. Assoc., Chicago 1986, pp. 99 – 100.
120. S. Inaba, K. Ishizumi, K. Mori, H. Yamamoto, *Chem. Pharm. Bull.* **19** (1971) 722. Hoffmann-La Roche, US 3 144 439, 1964; US 3 144 890, 1964; US 3 243 427, 1966. Dumex, ZA 6 706 791, 1968; *Chem. Abstr.* **70** (1969) 57 916 c. Sumitomo, DE-OS 1 816 046, 1970.
121. S. Sakai, S. Kitajawa, H. Yamamoto, *Arzneim. Forsch.* **22** (1972) 534. H. Yamamoto et al., *Sumitomo Kagaku, Tokushugo* 1977, 13 – 23; *Chem. Abstr.* **90** (1977) 26 w.
122. Martindale: *The Extra Pharmacopoeia,* 21st ed., The Pharmaceutical Press, London 1982, p. 1549. USP XXI, 762, Rockville, Md. (1985).
123. M. J. Bluckman et al., *Arzneim. Forsch.* **21** (1971) 1049 – 1102. Martindale: *The Extra Pharmacopoeia,* 21st ed., The Pharmaceutical Press, London 1982, p. 1543.
124. S. C. Bell et al., *J. Med. Chem.* **11** (1968) 457. A. Nudelman, *J. Pharm. Sci.* **63** (1974) 1886. Amer. Home Prod., BE 621 819, 1963; *Chem. Abstr.* **60** (1964) 2993 b; DE-OS 1 445 412, 1962; DE-OS 1 645 904, 1962; DE-OS 1 795 553, 1962; DE-OS 1 795 556, 1962.
125. A. Doenicke et al., *Anaesthesist* **28** (1979) 578. H. Ott et al., *Anaesthesist* **28** (1979) 29. R. Girkin et al., *Xenobiotica* **10** (1980) 401. M. Hümpel et al., *Eur. J. Drug. Metab. Pharmacokinet.* **4** (1979) 237. J. Oswald et al., *Br. Med. J.* 1982, I, 860. R. T. Owen, *Drugs Future* **5** (1980) 495 – 496. Pharmcast 87-10, 6-II-136 – 6-II-138, Technomic Inf. Service, Tokyo.
126. S. C. Bell, S. J. Childress, *J. Org. Chem.* **27** (1962) 1691. Am. Home Prod., US 3 197 467, 1965. Hoffmann-La Roche, US 3 340 253, 1967; US 3 374 225, 1968.
127. L. O. Randall, C. L. Scheckel, W. Pool, *Arch. Int. Pharmacodyn. Ther.* **185** (1970) 135. P. Sarteschi et al., *Arzneim. Forsch.* **22** (1972) 93. H. J. Schwandt, W. Sadee, K.-H. Beyer, *Xenobiotica* **4** (1974) 733. S. H. Curry, A. N. Nicholson, R. Whelpton, *Br. J. Pharmacol.* **57** (1976) 427. AMA: *Drug Evaluations,* 6th ed., Am. Med. Assoc. Chicago 1986, pp. 102 – 103. *Drug, Facts and Comparisons,* Facts and Comparisons Div., St. Louis 1982, pp. 991 – 993 (a division of J. B. Lipincott Co., Philadelphia –Toronto).
128. G. F. Tamagnone, R. De Maria, F. De Marchi, *Arzneim. Forsch.* **25** (1975) 720. Schiapparelli, DE-OS 2 338 058, 1972. Japk 7480 086, GB 1 431 282, 1972. Kleemann-Engel: *Pharmazeutische Wirkstoffe,* Suppl. 1982 – 1987, Thieme-Verlag, Stuttgart – New York 1987, pp. 1151 – 1152.
129. Hoffmann-La Roche, ZA 6 802 239, 1967.
130. G. F. Tamagnone, M. V. Torrielli, F. De Mardi, *J* . Pharm. Pharmacol. 26 (1974) 566. V. Monaco, A. Bontante, *Curr. Ther. Res.* **37** (1985) 626. L. De Angelis, *Drugs Future* **3** (1978) 145 – 148. S. Mardene, C. Biechie, G. M. Nano, *Ther. Drug Monit.* **3** (1981) 351 – 356. M. Babbini, M. V. Torielli, E. Strumia et al., *Arzneim. Forsch.* **25** (1975) 1294.
131. Gerot-Pharmazeutika, DE-OS 2 950 235, 1980; AT 379 391, 1984; *Chem. Abstr.* **104** (1986) 168 499 j.
132. H. Koch, *Drugs Future* **7** (1982) 99.
133. M. Steinmann, J. G. Topliss, R. Alekel et al., *J. Med. Chem.* **16** (1973) 1354. Schering, US 3 920 818, 1974. Scherico Ltd., DE-OS 2 138 733, 1971.
134. H. Winsauer, D. E. O'Hair, *Curr. Ther. Res.* **35** (1984) 228. A. Barnett, L. C. Jorio, E. Ongini, *Arzneim. Forsch.* **32** (1982) 1452. J. Castaner, *Drugs Future* **3** (1978) 139; **8** (1983) 165; **9** (1984) 152; **10** (1985) 171. S. J. Ankier, K. A. Goa, *Drugs* **35** (1988) 42 – 62. AMA: *Drug Evaluations,* 6th ed., Am. Med. Assoc., Chicago 1986, p. 102. *Annu. Drug Data Report* **8** (1986) 119; **2** (1979 – 1980) 192.
135. T. Miyadere et al., *J. Med. Chem.* **14** (1971) 520. Sankyo, DE-OS 1 812 252, 1969; J-Kokai 1967; *Chem. Abstr.* **71** (1969) 124516 h.
136. T. Kamioka, J. Nakayama, T. Hara, H. Takaji, *Arzneim. Forsch.* **28** (1978) 838. Y. Sakai, *Jpn. J. Pharmacol.* **33** (1983) 1017 – 1025. M. Yoshimoto et al., *Chem. Pharm. Bull.* **25**

(1977) 1378. *Med. Actual.* **17** (1981) 283 – 286. P. J. Thorpe, *Drugs Future* **6** (1981) 414 – 415.

137. J. B. Hester, D. J. Duchamp, G. C. Chichester, *Tetrahedron Lett.* 1971, 1609. J. B. Hester, A. D. Rudzick, B. V. Kamdar, *J. Med. Chem.* **14** (1971) 1078. K. Meduro, H. Tawada, Y. Kuwada, *Chem. Pharm. Bull.* **21** (1973) 1619, 2382. K. Kamiya, Y. Wada, M. Nishikawa, *Chem. Pharma. Bull.* **21** (1973) 1520. K. Megruo, Y. Kuwada, *Tetrahedron Lett.* 1970, 4039. D. L. Loffen, R. J. Fryer, D. A. Katonak, F. Wong, *J. Org. Chem.* **40** (1975) 894.
138. Upjohn, DE-OS 2 012 190, 1969; US 3 701 782, 1969; DE-OS 2 302 525, 1973; US 4 018 788, 1977; US 3 987 052, 1969. Toyama, DE-OS 1 955 349, 1971. Takeda, DE-OS 2 114 441, 1971; DE-OS 1 965 894, 1969; US 4 116 956, 1976.
139. S. Tanayama, Y. Shirakawa, Y. Kanai, Z. Suzuoki, *Xenobiotika* **4** (1974) 33. Y. Kanai, *Xenobiotika* **4** (1974) 441. F. S. Eberts, *Pharmacologist* **16** (1974) 196; *Drug Metab. Dispos.* **5** (1977) 547. H. Isozaki, M. Tanaka, K. Inanaga, *Curr. Ther. Res.* **20** (1976) 493. T. Momose, S. Shii, T. Kuge, *Curr. Ther. Res.* **19** (1976) 277. W. Ehrenstein, M. Müller-Limmroth, M. Opfermann, F. Weber, *Arzneim. Forsch.* **32** (1982) 280. J. K. Walsh et al., *Curr. Ther. Res.* **36** (1984) 866. *Med. Actual.* **12** (1976) 355 – 357. *Annual Drug Data Report* **2** (1979/1980) 85 – 86.
140. J. B. Hester, A. D. Rudzick, B. V. Kamdar, *J. Med. Chem.* **14** (1971) 1078.
141. A. Kales, M. B. Scharf, J. D. Kales, *Science (Washington, D.C.)* **201** (1978) 1039.
142. T. Roth, M. Kramer, T. Lutz, *Drugs Exp. Clin. Res.* **1** (1977) 279. D. A. C. Hegelbach-Feller, J. M. Tschopp, S. Christeller, J. Fabre, *Arzneim. Forsch.* **38** (1988) 387.
143. M. Leibowitz, A. Sunshine, *J. Clin. Pharmacol.* **18** (1978) 302. K. K. Okawa, G. S. Allen, *J. Int. Med. Res.* **6** (1978) 343. B. B. Knap, E. L. Boyd, G. Linsenmeyer, O. J. Linet, *Curr. Ther. Res.* **23** (1978) 230.
144. A. N. Sing, *Curr. Ther. Res. Clin. Exp.* **27** (1980) 627. A. N. Nicholson, *Br. J. Clin. Pharmacol.* **9** (1980) 223.
145. L. Lasagnon, *Lancet* **1** (1980), 815.
146. E. A. Swinyard in *Remingtons Pharmaceutical Sciences,* 17th ed., Mack Publ., Easton, Pa. 1985, p. 1064. AMA: *Drug Evaluation,* 6th ed., Am. Med. Assoc., Chicago 1986, p. 103. J. Castaner, *Drugs Future* **1** (1976) 393 – 396; **2** (1977) 558.
147. A. Walser et al., *J. Org. Chem.* **43** (1978) 936. Hoffmann-La Roche, DE-OS 2 540 522, 1976.
148. M. Gemper et al., *Arzneim. Forsch.* **31** (1981) 2177 – 2288. D. A. C. Hegelbach-Feller, J. M. Tschopp, S. Christeller, J. Fabre, *Arzneim. Forsch.* **38** (1988) 387. J. Kanto, *Med. Actual,* **19** (1983) 376 – 381.
149. K. H. Weber, W. Sirrenberg, O. Spohn, H. Daniel, *Arzneim. Forsch.* **36** (1986) 518. Boehringer, DE-OS 2 410 030, 1974 = US 4 094 984, 1975; DE-OS 2 503 235, 1975; DE-OS 2 533 924, 1975. Hoffmann-La Roche, DE-OS 2 221 623, 1971. K. Gewald, *Chem. Ber.* **98** (1965) 3571.
150. N. Ferrara et al., *Curr. Ther. Res.* **37** (1985) 295. J. Grünberger et al., *Curr. Ther. Res.* **24** (1978) 427. B. Saletu, J. Grünberger, J. Volavka, P. Berner, *Arzneim. Forsch.* **29** (1979) 700. M. S. Langley, S. P. Clissold, *Drugs* **35** (1988) 104 – 122. Pharmcast 87-10, Technomic Information Service, Tokyo 1987, pp. 6-II-151 – 6-II-152.
151. R. Agar et al., *J. Med. Chem.* **20** (1977) 1035. Russel-UCLAF, DE-OS 2 605 652, 1976; US 4 044 142, 1977.
152. T. G. Johns, D. C. Piper, G. W. L. James, *Arch. Int. Pharmacodyn. Ther.* **240** (1979) 53. J. Hindmarch, C. A. Clyde, *Drugs Exp. Clin. Res.* **2** (1980) 61. R. T. Owen, *Med. Actual.* **20** (1984) 59 – 63. Loprazolam Clinical Workshop, The Medicine Publishing Foundation, Oxford, 1983.
153. K. Hirai et al., *Heterocycl. Chem.* **19** (1982) 1363; *J. Med. Chem.* **23** (1980) 764; **24** (1981) 20. Shionogi, DE-OS 2 725 164, 1977; US 4 159 374, 1977. Hoffmann-La Roche, DE-OS 2 537 069, 1975.
154. K. Hirose, A. Matsushita, M. Eigyo, *Arzneim. Forsch.* **31** (1981) 63. *Drugs Future* **9** (1984) 522 – 524; **11** (1986) 626 – 628. Pharmcast 87-1, Technomic Information Service, Tokyo 1987, pp. 5-III-139 – 5-III-141.
155. Rhône-Poulenc, DE-OS 2 300 491, 1973; US 3 862 149, 1975. C. Jeanmart, C. Cotrel, *C. R. Ser. C* **287** (1978) 377.
156. R. Duriez, *Therapie* **34** (1979) 317. E. Wickström, K. E. Giercksky, *Eur. J. Clin. Pharmacol.* **17** (1980) 93; *Clin. Therap.* **3** (1978) 59. . J. C. Blanchard, A. Boireau, C. Garrett, L. Julou, *Life Sci.* **25** (1979) 2417. P. H. Wu, J. W. Phillis, A. S. Bender, *Life Sci.* **28** (1981) 1023. J. Castaner, *Drugs Future* **4** (1979) 59 – 60; **5** (1980) 56; **6** (1981) 62 – 63; **7** (1982) 69 – 70; **8** (1983) 79; **9** (1984) 81; **10**

(1985) 88 – 89; **11** (1986) 76 – 78; **12** (1987) 95; **13** (1988) 88 – 89. Pharmacast 87-7, Technomic Inf. Service, Tokyo 1987, 1989, 6-I-87 – 6-I-89. *Int. Pharmacopsychiatry* **17** (1982) (Suppl. 2), 1 – 252. K. A. Goa, R. C. Heel, *Drugs* **32** (1986) 48 – 65.

157. G. Houghton et al., *Int. J. Clin. Pharmacol. Biopharm.* **23** (1985) 97. R. Duriez et al., *Therapie* **34** (1979) 317. J. Gaillot et al., *Pharmacology* **27** (1983) (Suppl. 2) 76 – 91.
158. Synthelabor, EP 50 563, 1981. P. George, H. Depoortere, *J. Pharm-Sci. (in Press).*
159. S. Arbilla et al., *Naunyn-Schmiedebergs Arch. Pharmacol.* **330** (1985) 248. M. Dimsdale et al., *Drugs Future* **12** (1987) 777 – 780.
160. M. Dimsdale et al., *Drugs Future* **13** (1988) 106 – 109.
161. J. Maillard, M. Vincent, M. Bernard, *Bull. Soc. Chim. Fr.* **1961,** 529; **1962,** 1580. Logeais-Labs., GB 902 388, 1962;Fr. M. 379, 1962; *Chem. Abstr.* **57** (1962) 15251 g.

Hypochlorous Acid → Chlorine
Hypochlorous Acid → Chlorine Oxides and Chlorine Oxygen Acids
Hypoglycemic Drugs → Antidiabetic Drugs
Hypophosphorus Acid → Phosphorus Compounds, Inorganic

Ice Cream and Frozen Desserts

PHILIP G. KEENEY, Pennsylvania State University, University Park, Pennsylvania 16803, United States

1. **Introduction** 1
2. **Classification of Ice Cream Products** . 1
3. **Legal Aspects** 4
4. **Economic Aspects** 4
5. **Physical Structure** 4
6. **Mix Composition** 5
6.1. **Fats** 5
6.2. **Nonfat Milk Solids** 5
6.3. **Sweetener Solids** 5
6.4. **Stabilizers and Emulsifiers** 6
7. **Manufacture of Mix** 6
8. **Freezing Mix to Produce Ice Cream** . 7
8.1. **Freezing as Mix Agitates** 7
8.2. **Quiescent Hardening** 8
9. **References** 8

1. Introduction

The term ice cream is generally used to cover several types of desserts that are consumed in a frozen or partially frozen state. Ice cream is a frozen foam of an oil-in-water emulsion, which contains more than 55 wt % water and ice. Carbohydrate solutes contribute sweetness, but also control freezing and melting characteristics. The properties of the foam matrix are largely determined by lipid and protein; small amounts of emulsifier and hydrocolloid function as manufacturing aids and help to stabilize physical properties during storage and distribution. Flavor ingredients, mostly of botanical origin, (e.g., fruit, fruit products, cocoa, coffee, ginger, vanilla), allow great variations in aroma, taste, and texture impressions.

2. Classification of Ice Cream Products

The manufacturer has several variables to consider in the design and production of products to meet market needs. Common desirable traits are sweet and cool tastes that satisfy the expected refreshing qualities. Superimposed on these basic requirements are the influences of the characteristic flavor ingredients and the effects of fat, protein, other mix components, air, and ice crystals on creaminess, smoothness, and other physical qualities perceived upon consumption.

The "Codex Standard on Edible Ices and Ice Mixes" gives a comprehensive survey of types of ice cream and a detailed account of food additives (colors, emulsifers, stabilizers, thickening agents, flavors, etc.) including maximum levels and microbiological requirements [1]. The Codex Standards are merely recommendations for UN countries but the importance of the Edible Ice Standards has not declined since the EEC commission abandoned its plan to introduce Edible Ice Guidelines a few years ago.

Most countries have regulatory definitions consistent with local customs and needs. Almost always, one product class reflects the dominance of milk ingredients in determining consumption characteristics. In the United States these are *ice cream* and *ice milk*, which are similar except that the latter contains less fat. In the United Kingdom, ice cream made with milk fat is termed *dairy ice cream*. In Europe an increasing proportion of ice cream is sold as *premium ice cream*, which has a higher fat content than normal ice cream and is not as highly whipped.

In some countries (e.g. Japan, Sweden, England), vegetable fats may be used in place of milk fat. In fact ca. 80 % of ice cream consumed in these countries contains vegetable fat. *Mellorine* is a product in which all or a portion of the milk fat is replaced by another fat of plant or animal origin.

A second class of products is dominated by the soluble carbohydrate fraction and contains little or no fat and protein. These are the *sherbets, sorbets*, and *ices*. A third category comprises products with a dietary attraction. Examples are artificially sweetened products for people who desire low-calorie foods or who must control dietary sugar intake. In these products

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_563

Table 1. International compositional requirements in ice cream and related products [a]

Country	Component	Cream ice (≥5% milk fat)	Cream ice with fruit	Milk ice (<5% milk fat)	Frozen dessert with egg yolk	Frozen dessert with nondairy fat	Fruit sherbet	Water ice
Australia [b]		Ice cream		Dairy ice		Frozen confection		Ice confection
	milk fat	100 g/kg		50 g/kg		80 g/kg		prepared from water, foods, or water and foods
	food solids	168 g/L				150 g/L		
Belgium and The Netherlands		Roomijs, crème glacée		Melkijs, glace au lait	Eiroom-ijs, Ijs, glace glace aux oeufs			Waterijs, glace à l'eau
	milk fat	9	7	2.5	7			
	msnf [c]			5				
	msnf + milk fat	16	14		14			
	total solids	30	30	25	30	25		15
	egg yolk				7			
	fruit, chocolate sauce		15					
	citrus fruit		10					
Canada		Ice cream	[d]	Ice milk			Sherbet	
	milk fat	10	8	3–5			with fruit or fruit juice, not more than 5% milk solids incl. milk fat	
	solids	36	36	33				
	solids of which milk fat must be	1.8 lb/gal 0.50 lb	1.8 lb/gal 0.40 lb	1.6 lb/gal 0.14 lb				
Federal Republic of Germany [f]		Eiskrem	Fruchteis-krem	Milch-speiseeis	Kremeis (Eier-kremeis)		Fruchteis	Kunstspeise-eis
	milk fat	10	8					artificial flavors and colors are allowed
	milk			70				
	egg yolk				100 g/L milk or 270 g/L egg			
	fruit						20	
	citrus fruit						10	
France		Glace à la crème, crème glacée, ice-cream [e]			Glace aux oeufs		Sorbet	Glace à ..., glace au sirop [e] ...
	milk fat	7	5		2			
	sucrose	14	14		16			18
	total solids	31	29		20			28
	fruits		15				35	
	acid fruits		10				15	
	egg yolk				7			
Japan		Ice cream [g]		Ice milk [g]				
	milk fat	8		3				
	milk solids (incl. fat)	15		10				

Table 1. Continued.

Sweden		Gräddglass		Mjölkglass	Glass	Sherbet	Saftis
	milk fat	8		3			
	milk protein	2.5		2.5	1 (msnf)		
	total solids	30		28	28		
	fat				3	1–3	
	protein				2.5		
	fruit					15	15
	citrus fruit					10	10
United Kingdom		Dairy ice cream	Dairy ice cream with fruit	Milk ice	Ice cream	Ice cream with fruit	Parev ice
	milk fat	5	5	2.5			no milk fat or other milk derivative
	msnf	7.5	7.5	7	7.5	7.5	
	fat		or		5	5	
	milk fat		7.5			or	
	msnf		2			2	
	milk solids incl. (milk) fat		12.5			12.5	

[a] This table was provided by Dr. Fritz Timm (retired from Langnese, Hamburg, Federal Republic of Germany). Unless otherwise stated figures refer to minimum contents expressed in weight percent. For United States requirements, see text.
[b] Requirements according to proposed draft regulations and model food standards regulation.
[c] msnf = milk solids not fat.
[d] Ice cream containing cocoa, chocolate syrup, fruit, nuts, or confections.
[e] Different requirements apply for added ingredients (chocolate, praliné, coffee, vanilla, pistachio, malt, caramel).
[f] A further German ice cream type is "Rahmeis (Sahneeis)" with ≥60% cream. Another frozen dessert is "Einfacheiskrem" with ≥3% milk fat.
[g] Terms are not official. Another Japanese frozen dessert, "Lacto ice" contains ≥ 3% milk solids.

sucrose is replaced by sugar substitutes such as fructose or sorbitol to ensure that the ice cream retains desirable freezing and melting characteristics.

Ice cream is sold in a variety of forms, shapes, and flavors. Forms of supply include:

1) single portions (sticks, cups, tubs, cones, sandwiches, puppet ice),
2) family-sized portion (bricks, bombes, multipacks), and
3) catering packs.

A further differentiation among products is based on whether freezing occurs quiescently, and/or under agitation. Most stick confections are subjected to *quiescent freezing* by pouring liquid mix into molds immersed in a brine solution at −35 °C. The finished product with inserted stick has the geometry of the mold, which allows for a great variety of shapes and sizes. *Freezing under agitation* refers to products in which 35–55 % of the water content is frozen as the temperature is lowered to between −4 and −6 °C in a swept-surface heat-exchange cylinder, which also distributes air as overrun. (Overrun refers to volume expansion caused by air incorporation, it being 80 % when 1 L of liquid mix yields 1.8 L of frozen product.) Product extruded from the freezer can be dispensed into cones or cups to be consumed in soft form which cannot be stored or transported. Alternatively, it can be packaged and placed in a colder environment (at below −18 °C) to allow quiescent freezing of most of the remaining water to yield ice cream in its typically hardened, packaged form. The latter process is a combination of freezing quiescently and under agitation. Hardened ice cream (including varieties that are easily scoopable at −20 °C) can be stored and transported.

3. Legal Aspects

International compositional requirements for ice cream and related products are summarized in Table 1.

In the United States ice cream, ice milk, sherbet, fruit ice, and mellorine are defined under the standards of the federal Food and Drug Administration [2]. Guidelines for other products are specified by individual states. Key provisions for

ice cream are a minimum of 10 % milk fat and 20 % total milk solids, and a product weight of at least 4.5 lb per gallon (0.54 kg/L). The weight requirement prevents more than a doubling of mix volume through air incorporation during manufacture. *Ice milk* is low-fat ice cream (with a fat content of 2 – 7 wt %). In *mellorine* some or part of the milk fat is replaced by another plant or animal fat. *Sherbet* must contain 1 – 2 % milk fat with not more than 5 % total milk solids and a minimum weight of 6 lb per gallon (0.71 kg/L). Most sherbets have a tart taste due to the addition of an edible acid, usually citric acid. *Fruit ice* requirements are similar to those of sherbet, except that milk solids are not allowed.

4. Economic Aspects

In the United States, 65 % of the 1.4×10^9 gallon (5.3×10^6 m^3) annual production of frozen desserts is ice cream, 22 % is ice milk, 7 % sherbet, 1 % mellorine, and 5 % other products [3]. Soft-serve products account for 21 % of the total production, most being ice milk. At 22 L per capita, production is highest in the United States, followed closely by Finland at 21 L, New Zealand and Australia at 19 L, then Canada and Sweden with 14 – 19 L.

5. Physical Structure

Ice cream being a foam of an oil-in-water emulsion is a composite of liquid, solid, and gaseous phases (Fig. 1). The numbers and sizes of air cells and ice crystals are influenced by several manufacturing and storage variables. A representative diameter for an ice crystal is 30 µm with air cells averaging 60 µm. The proportion of the mass which is present as a continuous liquid phase depends on water – ice equilibria as determined by temperature.

Changes in the nonfrozen phase during freezing are considerably more complex than a mere concentration of soluble constituents as water crystallizes. The beating and scraping action which facilitates heat transfer and air incorporation in the freezing cylinder can profoundly alter colloidal and emulsion properties. An electron micrograph of ice cream (Fig. 2) shows that the fat particles are located at the surface of the air cell [5]. Evidence exists for accumulation of protein and fat near the interface between air and the liquid, unfrozen continuous phase (also called serum) [5]. Emulsions are thermodynamically unstable [6] and changes occurring are the result of the shearing of the mix during freezing and whipping in the swept-surface cylinder. Particle size distribution for fat tends towards larger globules and formation of agglomerates is evident. Depending on the amount of change, the physical characteristics of the frozen emulsion matrix are profoundly influenced.

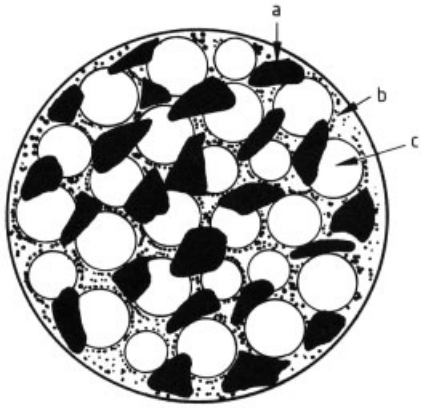

Figure 1. Illustration of ice cream structure
a) Ice crystal; b) Fat globules; c) Air cell

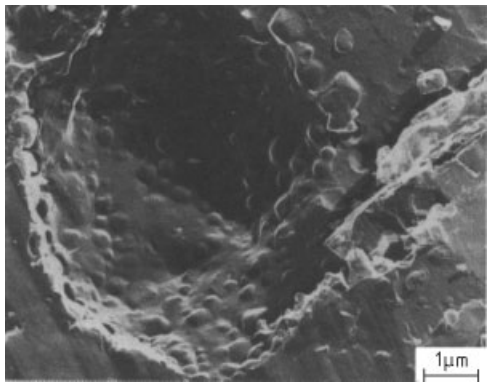

Figure 2. Wall of air cell in ice cream revealed by freeze-etching and transmission electron microscopy [4].

6. Mix Composition

Table 2 shows typical mix compositions for two types of ice cream, an ice milk, and sherbet [7].

Mix for fruit ice is similar to that for sherbet, except for the absence of milk solids. Although each mix constituent is important for determining the qualities perceived upon consumption of ice cream, the amount and type of fat is especially critical. Thus, the fat level of the mix is decided first and then the proportions of the other constituents are selected to complement it.

Table 2. Representative compositions for ice cream, ice milk, and sherbet (wt %)

Ingredient	Regular ice cream	Super premium ice cream	Ice milk	Sherbet
Milk fat	10.0	16.5	4.0	1.5
Nonfat milk solids	11.5	9.0	13.0	3.0
Sucrose	10.2	15.5	9.0	23.0
Corn sweetener solids	6.8		9.0	7.0
Stabilizer/ emulsifier	0.3	0.2	0.4	0.3
Total solids	38.8	41.2	35.4	34.8

6.1. Fats

Fat is primarily responsible for the creaminess and mellowing qualities of ice cream products. It also depresses the intensity of cold sensations experienced upon consuming a partially frozen product. An important role for fat in establishing the physical matrix has been referred to in Chapter 5.

Fresh cream (milk and milk products) provides most of the fat for ice cream in the United States. It is readily available and, being a liquid, has handling advantages. Cream is recognized best from the standpoint of flavor. For economic reasons, butter is more commonly used in Europe; where allowed, a vegetable fat can also be used, either alone or in combination with butter or cream. Preferred are partially hydrogenated fats of the lauric type from coconut and palm kernel.

6.2. Nonfat Milk Solids

The nonfat solids of milk consist of 35 % protein, 55 % lactose, and 10 % mineral ash. The protein fraction is particularly important for ice cream. Protein coats fat particles and stabilizes the emulsion in homogenized mix, it also allows controlled colloid and emulsion changes during freezing. Protein is essential for the encapsulation of air as overrun, and its hydrophilic properties enhance desirable texture characteristics.

Because the nonfat solids content of milk is only 8.6 wt %, a concentrate must be employed to reach the levels specified in Table 2. These solids are best provided by whole or skim milk concentrated to 25 – 35 wt % solids, or by skim milk powder (→Milk and Dairy Products). In the United States, 25 % of the nonfat milk solids in ice cream can be supplied by whey, a cheese industry byproduct (→Cheese, Processed Cheese, and Whey, Chap. 7.2.). Since whey solids contain only 13 % protein compared to 35 % in skim milk solids, ice cream composition is altered somewhat when whey is used as a partial substitute for skim milk. However, if whey is combined with commercially available casein, the composition of skim milk is duplicated quite closely.

6.3. Sweetener Solids

Ice cream is a sweet-tasting food, but the preferred sweetness intensity varies among markets around the world. In the United States, the sweetness equivalency of a 15 % sucrose solution is preferred. Sweetness in ice cream is provided by soluble carbohydrates which themselves vary in sweetness. Lactose has a fifth and 36 DE corn sweetener solids half the sweetness intensity of sucrose, but solids containing fructose may be even sweeter than sucrose.

Of importance equal to sweetness control is the dominant role of the soluble carbohydrates in determining the colligative freezing properties of the mix. The water – ice equilibrium at any temperature mostly reflects the combined concentrations of these solutes.

Depending on the combination of carbohydrate ingredients, total mix solids may be increased or decreased to influence texture and body characteristics. In this regard, the higher saccharides of low-conversion corn syrups have a favorable impact [8].

Sucrose from sugar cane and sugar beet has been the sweetener of choice for ice cream. Primarily for economic reasons, syrups of hydrolyzed corn starch now partially replace su-

crose in most frozen dessert products. Corn syrups of the low conversion 36 DE type compensate to a degree for the gradual reduction over many years in the milk solids content of economy and regular ice cream brands. Since their introduction in the 1970s, fructose syrups obtained by hydrolysis of starch syrup to glucose and enzyme isomerization have gained a strong foothold. Conventional corn syrups are not as sweet as sucrose, but fructose syrups are equally sweet or sweeter. Although they are the most economical source of sweetness, fructose syrups increase the molar concentration of solutes, thus requiring lower freezing and storage temperatures to maintain proper hardness in packaged products. For a detailed description of syrups, see →Glucose and Glucose-Containing Syrups.

6.4. Stabilizers and Emulsifiers

In most countries the addition of stabilizers and emulsifiers to ice cream is legally restricted. Most ice cream products contain 0.1 – 0.2 wt % *hydrocolloid stabilizer* which functions as a water control agent and increases the viscosity of the unfrozen aqueous phase. By mechanisms which are poorly understood, the hydrocolloids delay growth of large ice crystals and prevent deterioration of quality during storage and distribution. *Emulsifiers,* used at the same or slightly higher levels than hydrocolloids, serve as a manufacturing aid by helping to control emulsion changes and lead to good extrusion properties for ice cream from the freezing cylinder. For a description of emulsifiers, see →Foods, 3. Food Additives, Chap. 3.6.

A large number of hydrocolloids function in ice cream systems; guar gum, locust bean gum, cellulose gum, and carrageenan are used most frequently. Most are used in combination as stabilizer blends, selection being based on the rheological properties desired by a particular ice cream manufacturer. Generally, carrageenan is a component of stabilizer blends, e.g., with guar or locust bean gum at a ratio of 1 : 10 [9] or at a concentration of 0.018 wt % [10]. It functions as a manufacturing aid to prevent separation of mix into layers during quiescent storage prior to freezing into ice cream.

Monoglycerides in mixture with diglycerides, and polysorbate-80 have a long history of use as emulsifiers for ice cream. The polysorbates are especially effective in hastening controlled emulsion destabilization leading to desirable extrusion properties for ice cream [11].

7. Manufacture of Mix

Preparation of a mix for freezing into ice cream is a multistep process involving the assembly of the mix ingredients, pasteurization of the resulting fluid mix, homogenization to stabilize the emulsion, and finally the cooling and storage of the finished mix to await freezing. Equipment can be sized to meet production needs. Mix output varies from 2000 L/h for a small ice cream factory to 20 000 L/h for a high-volume operation. Hygiene plays an important role in the manufacture and treatment of ice cream. The hygienic preparation of ice cream requires not only pasteurization, but also regular cleaning and disinfection of equipment, proper packing and storage, i.e., Good Manufacturing Practice. Pasteurization is not compulsory in all countries, e.g., in the Federal Republic of Germany it is only necessary for the types "Eiskrem" (the main type of industrially produced edible ice) and "Einfacheiskrem". The microbiological state of ice cream is subject to very stringent regulations in most countries, e.g., total microbial count, coliforms, and pathogenic microorganisms (especially *Salmonella*).

Pasteurization. The primary goal of pasteurization is the guaranteed destruction of pathogenic microorganisms. In the process, almost all nonpathogens are simultaneously destroyed, but the mix is not sterile and must be stored at 4 °C to prevent multiplication of surviving microorganisms. Pasteurization regimes are defined and strictly controlled by federal and state regulatory agencies.

In *batch pasteurization*, all ingredients in the desired proportions are fed into a jacketed vat, heated for 10 – 30 min under agitation to 68.3 °C, and held at this minimum pasteurization temperature for at least 30 min. The vat contents must then be pumped through the homogenizer before a new batch can be started.

Continuous high temperature – short time (HTST) pasteurization is more energy efficient than batch pasteurization and better suited to

high-volume plants. Mix ingredients are assembled in a vat similar to the batch process. However, the mix is pasteurized by pumping it from the vat as a thin film through a plate heat exchanger. Thin film heating is almost instantaneous and continuous HTST pasteurization is the preferred approach to pasteurization in the ice cream industry. Temperature – time conditions for HTST pasteurization are 78 – 80 °C for 20 – 40 s (US minimum requirements are 79.4 °C for 25 s).

The times and temperatures for batch and HTST pasteurization are acceptable minimum combinations. Often, a higher temperature or longer hold is employed to provide added protection relative to microbial kill. Furthermore, a significant number of manufacturers believe that an extra measure of heat energy, especially at elevated temperature, improves ice cream body and texture.

Homogenization. Following batch pasteurization or as an integral part of the HTST process, mix passes through a homogenizer. Homogenizers are high-pressure pumps which force mix through a restricted valve under defined conditions of turbulence and shear. During homogenization, fat globules are broken up to less than 2 µm in diameter; the resulting small globules become coated with milk protein, emulsifier, and other surface-active constituents. The temperature of the mix should be at least 50 °C to ensure that the fat is completely melted.

Storage of Mix. The prepared mix is held at 4 °C or less in a storage tank until it can be frozen into ice cream. In the United States, freezing is usually carried out the same day the mix is made, with only a 2 – 3 h hold in the storage tank to allow complete crystallization of fat globules. The former practice of holding the cold mix for at least 6 h (usually overnight) before freezing commenced has been been superseded by the use of hydrocolloid stabilizers.

8. Freezing Mix to Produce Ice Cream

Mix water begins to freeze at −2.5 °C, the exact temperature being a function of the solution concentration. Dissolved components in the aqueous phase that control this colligative property are the lactose and soluble salts of milk, as well as the carbohydrates supplied by sweetening and flavoring ingredients. Measured and calculated values for the freezing point depression are in close agreement, at least in regard to the initial freezing temperature [12]. As water freezes, the solutes become more concentrated, thus requiring even lower temperatures for the continuing change of state of water to ice.

Consider a mix containing 15 wt % sucrose, 7 wt % lactose and milk salts (i.e., 22 wt % soluble solids), and 38 wt % total solids (62 wt % water). The ratio of soluble solids to unfrozen water is 35 : 100 at the start of freezing. When soft ice cream at −5.5 °C leaves the swept-surface cylinder of the freezer to be packaged, the ratio will be 70 : 100 since 50 % of the mix water will have crystallized as ice. At a typical storage temperature near −20 °C the ratio will be 275 : 100. Obviously, as temperature decreases and water freezes, ice cream stiffens progressively to a hardened mass.

8.1. Freezing as Mix Agitates

Ice cream freezers are of two types, batch and continuous. In *batch freezers*, the mix is frozen and whipped in single batches of 10 – 20 L depending on machine size. The refrigerant (either ammonia or Freon) in the jacketed rim of the cylinder, absorbs heat from the mix causing it to freeze as a thin film. The film is continuously scraped off the cylinder wall by revolving scraper blades. Depending on conditions, 4 – 10 min is required to reach −4.4 °C, which is a typical discharge temperature for batch-frozen ice cream. The proportion of mix water frozen (ca. 40 %) at this temperature results in a viscosity or thickness for the whipping mass that is ideal for incorporation and retention of air as overrun. If the mix is either too highly frozen or not frozen enough, air cannot be retained within the matrix.

With *continuous freezers*, mix is pumped through the freezing cylinder and exits semifrozen in a continuous stream. Because the cylinder barrel is pressurized, product stiffness is not nearly as important for air incorporation as in batch freezing. Consequently, ice cream flow-

ing from a continuous freezer can be at −6 °C or even lower, which means more extensive nucleation during the fast freezing stage of ice cream manufacture. This produces smaller ice crystals and a smoother texture. Because continuous freezers use a pumping system, they can be adapted for molding ice cream by extrusion, for incorporating variegating syrups and a great variety of particulate flavor ingredients, and the automatic filling of packages. Several sizes of continuous freezers allow mix pumping rates between 200 and 2000 L/h over an almost infinite range of overruns. Because of their many advantages, continuous freezers are preferred over batch machines, except in the smallest ice cream factories.

8.2. Quiescent Hardening

The final stage of ice cream manufacture involves further freezing to convert packages of soft ice cream to a hardened form. Hardening takes place in either a hardening room or tunnel. Ideally in a *hardening room*, currents of air at −30 °C sweep across all sides of a carton or package. For this to occur, the packages must be kept separated on shelves until core temperatures are well below −20 °C, this may require several hours depending on package size. Only then can the packages be bundled and placed on pallets for storage until shipment from the factory. In *tunnel hardening*, packages move through the tunnel on a conveyor in −50 °C high velocity air. Hardening is much faster than in a hardening room, and upon exiting the tunnel packages are ready to be bundled and placed on pallets for storage. The main advantages of tunnel hardening are efficiencies of time, labor, and space utilization.

9. References

1. Codex Standard for Edible Ices and Ice Mixes 137, 1981; *IDF Bulletin* **172** (1984) Appendix II.
2. Code of Federal Regulations 21, Food and Drugs Parts 110 – 199, U.S. Government Printing Office, Washington, D.C., 1987, pp. 228 – 236.
3. *The Latest Scoop,* International Ice Cream Association, Washington, D.C., 1987.
4. K. C. Berger in S. Friberg (ed.): *Food Emulsions,* **vol. 5,** Dekker, New York 1976, pp. 141 – 213.
5. P. G. Keeney, J. A. Maga, *J. Dairy Sci.* **48** (1965) 1591 – 1596.
6. S. Friberg in S. Friberg (ed.): *Food Emulsions,* **vol. 5,** Dekker, New York 1976, pp. 1 – 37.
7. P. G. Keeney: *Commercial Ice Cream and Other Frozen Desserts,* Extension Circular 553, The Pennsylvania State University, University Park, Pa., 1972, pp. 1 – 50.
8. P. G. Keeney, D. V. Josephson, *Ice Cream Trade J.* **57** (1961) no. 5, 28 – 34.
9. K. A. Hyde, J. Rothwell: *Ice Cream,* Churchill Livingstone, Edinburgh 1973.
10. P. G. Keeney, M. Kroger in B. H. Webb, A. H. Johnson, J. A. Alford (eds): *Fundamentals of Dairy Chemistry,* AVI Publishing 6; Washington 1974.
11. P. G. Keeney, *Food Technol. (Chicago)* **36** (1982) no. 11, 65 – 70.
12. W. Arbuckle: *Ice Cream,* 4th ed., AVI Publishing Co., Westport, CN, 1986, p. 46.

Illites → Clays

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Imaging Technology

Standard Article

Detlef Winkelmann[1], Manfred Lutz[2], Damodar M. Pai[3], Andrew R. Melnyk[4], Richard Hann[5], Walter Crooks[6], Keith S. Pennington[7], Francis C. Lee[8], C. Wayne Jaeger[9], Don R. Titterington[10], Walter Lutz[11], Arno Bräuninger[12], Luc De Brabandere[13], Frans Claes[14], Rene De Keyzer[15], Wilhelmus Janssens[16], Johan Verelst[17], Werner Frass[18], Thomas Telser[19], Horst Hoffmann[20], Bernd Bronstert[21], Karl-August Springstein[22], Rod Potts[23], Hartmut Steppan[24], Donald C. Mammato[25], Thomas Stoudt[26], Michael C. P. Watts[27], David Allen[28]

[1]Hoechst Aktiengesellschaft, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany
[2]AEG Elektrofotografie GmbH, Warstein, Federal Republic of Germany
[3]Xerox Corporation, Joseph C. Wilson Center for Technology, Rochester, New York 14644, United States
[4]Xerox Corporation, Joseph C. Wilson Center for Technology, Rochester, New York 14644, United States
[5]ICI Imagedata, Brantham Industrial Estate, Manningtree, United Kingdom
[6]IBM, San Jose, California 95193, United States
[7]IBM, San Jose, California 95193, United States
[8]IBM, San Jose, California 95193, United States
[9]Xerox, Wilsonville, Oregon 97070, United States
[10]Xerox, Wilsonville, Oregon 97070, United States
[11]Kalle ReproMedia, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany
[12]Hoechst Aktiengesellschaft, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany
[13]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[14]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[15]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[16]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[17]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[18]Hoechst Aktiengesellschaft, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany
[19]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[20]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[21]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[22]Hamburg, Federal Republic of Germany
[23]Agfa-Gevaert N.V., Mortsel (Antwerpen), Belgium
[24]Hoechst Aktiengesellschaft, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany
[25]Hoechst Celanese Corporation, Somerville, New Jersey 08876, United States
[26]Hoechst Celanese Corporation, Somerville, New Jersey 08876, United States
[27]Hoechst Celanese Corporation, Somerville, New Jersey 08876, United States
[28]College of Manufacturing, Cranfield Institute of Technology, Cranfield, United Kingdom

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a13_571.pub2
Article Online Posting Date: March 15, 2003

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction

2. Copying and Nonimpact Printing Processes
2.1. Office Copying and Printing
2.1.1. Electrophotography
2.1.1.1. Photoreceptor
2.1.1.2. Corotron Charging
2.1.1.3. Light Exposure
2.1.1.4. Image Development
2.1.1.5. Image Transfer and Fusing
2.1.1.6. Photoreceptor Cleaning and Erase
2.1.1.7. Other Electrophotographic Imaging Systems
2.1.2. Thermographic Printing
2.1.2.1. Thermal Print Transducers
2.1.2.2. Resistive Ribbon Technology
2.1.2.3. Imaging Materials and Mechanisms
2.1.2.4. Comparison of Thermal Technologies
2.1.3. Ink-Jet Printing
2.1.3.1. Continuous Ink Jet
2.1.3.2. Impulse (Drop-on-Demand) Ink Jet
2.1.3.3. Ink-Jet Nozzle Orifice
2.1.3.4. Ink-Jet Inks
2.1.3.5. Colorants Used in Ink Jet Inks
2.1.3.6. Color Printing with Ink Jets
2.2. Technical Copying
2.2.1. Diazotyping
2.2.1.1. Photolysis
2.2.1.2. Coupling Reaction
2.2.1.3. Production of Diazotyping Material
2.2.1.4. Processing
2.2.2. Other Photochemical Systems
2.2.3. Silver Process
2.2.3.1. Black and White Copying Materials Based on Silver Diffusion Transfer Technology
2.2.3.2. Modern Lithographic Printing Applications
2.2.3.3. Color-Copying Materials Based on Dye Diffusion Chemistry
2.3. Microfilms and Microfiches
2.3.1. Diazotyping
2.3.2. Vesicular Film
3. Imaging in Graphic Arts
3.1. Graphic Arts Photography
3.1.1. Classes of Photomaterials Used in Prepress Production
3.1.2. Flow Scheme of Graphic Prepress Process
3.2. Basics of the Use of Light-Sensitive Nonsilver Materials in the Graphic Arts
3.2.1. Application
3.2.2. Chemical Fundamentals
3.2.2.1. Photosolubilising Systems
3.2.2.2. Photoinsolubilizing (Photocuring) Systems
3.2.2.3. Photophysical Principles of Image Generation
3.3. Color Proofing
3.3.1. Basics
3.3.2. Applications
3.3.3. Overlay Systems
3.3.4. Single-Sheet Systems
3.3.4.1. Precolored Systems
3.3.4.2. Systems With Colors Generated During Processing
3.3.5. Digital Proofing
3.4. Platemaking
3.4.1. Lithography (Planographic Printing, Offset)
3.4.1.1. Substrates
3.4.1.2. Coating and Further Processing
3.4.1.3. Processing of Presensitized Plates

3.4.2. Letterpress and Flexography
3.4.2.1. Structure of Photopolymeric Letterpress and Flexographic Plates
3.4.2.2. Production and Requirements of Photopolymer Letterpress and Flexographic Plates
3.4.2.3. Chemistry of Photopolymer Relief Printing Plates
3.4.2.4. Future Developments
3.4.3. Gravure Printing
3.4.3.1. Conventional Method of Gravure Platemaking
3.4.3.2. Electromechanical Gravure Platemaking
3.4.3.3. Electron-Beam and Laser Gravure
3.4.4. Screen Printing
3.4.4.1. The Screen
3.4.4.2. The Stencil
4. Imaging for Electronics
4.1. Photoresists
4.1.1. Industrial Applications of Photoresists
4.1.2. Function and Chemistry of Photoresists
4.1.2.1. Positive Photoresists
4.1.2.2. Negative Photoresists
4.1.3. Testing Methods
4.1.4. Economic Aspects and Suppliers of Photoresists
4.1.5. Occupational Health and Environmental Protection
4.2. Printed Circuits (Printed Circuit Boards)
4.2.1. Methods of Producing Printed Circuit Boards
4.2.2. Masking Techniques
4.2.2.1. Exposure Masks
4.2.2.2. Screen Printing
4.2.2.3. Photoprinting
4.2.2.4. Other Techniques
4.2.3. Economic Aspects
4.2.4. Occupational Health and Environmental Protection
4.3. Microelectronic Devices
4.3.1. Integrated Circuit Manufacturing
4.3.2. Photoprocess
4.3.3. Techniques for Submicron Lithography
5. Photochemical Machining
5.1. Introduction
5.2. Artwork Generation and Phototool Production
5.3. Materials
5.4. Photoresist Systems
5.5. Etching Technology
5.6. Process Capability
5.7. Economic Aspects
5.8. Products

[Top of Page]

1. Introduction

Detlef Winkelmann and Manfred Lutz

The roots of modern imaging go back to the invention of photography in the middle of the 19th century. In those early days photography was used primarily in portraiture, illustration, and documentary imaging. In the first half of the 20th century new applications of imaging technology emerged, which altogether had the main purpose to store information for later use or for distribution. Examples include microfilm recording, prepress printing processes, reprography, or office copying.

The emergence of solid state electronics in the middle of the 20th century had an impact on imaging in two ways. The first one was that now images could be recorded, transmitted, or distributed electronically. Fax machines and wireless transmitting of photographs became the first applications of this new technology.

Secondly, imaging techniques were used to produce for example integrated circuits, and micromechanical or microelectronic components. In consequence imaging techniques created the basis for today's imaging and communication technology which may probably end up in the course of the 21st century in the cross-media concept. This will convert information such as text, images, photographs, in combination with animation or interactive components into a “document” that can be produced as

either a printed report, digital hardware in form of a CD-ROM or as an information platform in the internet.

In spite of this vision still a wide variety of imaging techniques exists and will stay for a while in application in analog, hybrid, or digital systems. But mere analog imaging systems will not see any more growth in the market.

In principle, two main fields for imaging exist today:

1. The production of permanent, visually perceptible two-dimensional information,
2. The production of goods or devices as printed circuits boards, microelectronic, and micromechanical devices as well as photochemical machining.

For both fields sophisticated materials and processes have been developed in the past for analog image taking, processing, and output. With the new digital systems image taking and processing will be handled by computer systems and only the processing software and the output hardware remain unique for the different fields of imaging. This means for example that output on paper will still require another technique than the writing process for a mask to be used for the manufacturing of microelectronic devices.

Copying became very popular and widespread in the office environment with the invention of the electrophotographic process which easily could duplicate a document for further distribution. Now with the availability of highly integrated circuits, documents are digitized by scanners and a hard copy is prepared by a printer using one of the nonimpact output technologies as electrophotography, ink jet, thermography, and electro- or magnetostatics. Such a system could work either as a digital copier, a fax machine, or a computer printer.

Furthermore the availability of high resolution scanners and printers both in black and white and in color enables the replacement of analog processes in other areas than office communication. Examples are imaging in graphic arts, wide format copying and reprography, color proofing, plate making, and offset printing.

It is only a matter of time that analog diazotyping — still used in wide format copying and reprography — will be replaced completely by digital copiers or plotters using electrophotograhic, electrostatic, or ink jet processes, since drawings and layouts are nowadays almost all designed on CAD systems and are digitized anyway.

Print shops are still using to a very large extent lithographic materials, processes, and machinery. But printing industry talks about computer to plate, computer to press, and computer to paper.

All three of them include digital processing of information during the prepress steps but require special output technologies.

Computer to plate needs a photopolymer sensitive to the light of a raster image systems, mostly laser exposure, to change the physicochemical behavior necessary for the printing process.

Computer to press means there is no printing plate. The print master is directly generated on a sleeve or a cylinder in the printing press by converting light or temperature sensitive materials into durable print masters and washing away the unexposed areas. When the print job is finished the sleeve or the cylinder are cleaned and can be prepared for a new job. Besides the savings in materials and set up time, there are no plates to be stored. For a reprint the digital file has to be reloaded only.

Finally, *computer to paper* is the fastest and most convenient way to print small editions. Such systems employ one of the nonimpact printing techniques as electrophotography, ink jet, electrostatics, or magnetostatics, depending one the market segment they target.

In the field of imaging methods for the manufacture of *integrated circuits*, and *micromechanical* or *microelectronic components* again digital image processing is used to design the masks for the following process steps.

In the course of time efficient photoresists, exposure systems, and process units have been developed for the various applications.

To achieve smaller size of e.g., electronic devices and therefore a higher integration is the driving force for the *submicron lithography*. Two lithographic systems have to be distinguished: One generates the mask — today mostly an electron beam system — and the other one structures the wafer. Current 256 MBit DRAMs (dynamic random-access memory) with 250 nm structures are exposed with a KrF laser (λ = 248 nm). Future 4 and 16 GBit DRAMs require shorter wavelength lasers or other technologies as UV, X-Ray, electron beam, or ion lithography. These are expected to be capable to create structures smaller than 100 nm corresponding to 64 GBit-DRAMS and even more within the next decade.

If this target can be reached it would have again an impact on imaging technologies since more memory and faster processors will be available for image processing.

[Top of Page]

2. Copying and Nonimpact Printing Processes

Manfred Lutz, Damodar M. Pai, Andrew R. Melnyk, Richard Hann, Walter Crooks, Keith S. Pennington, Francis C. Lee, C. Wayne Jaeger, Don R. Titterington, Walter Lutz, Arno Bräuninger, Luc De Brabandere, Frans Claes, Rene De Keyzer, Wilhelmus Janssens, Rod Potts

2.1. Office Copying and Printing

Since the invention of xerography by CHESTER CARLSON in 1938, a variety of imaging systems have been developed into

commercial products [1], [2] for the purpose of duplicating analog information — like in office copiers, or plain paper fax machines — or for the purpose of distribution and remote printing of originally analog, or digital information.

As systems components *analog copiers* are provided with optics to image the original onto the process unit in order to produce the copy (Fig. 1). *Printers*, however, need a character generator and raster or scanning optics and sufficient memory for the intermediate data storage (Fig. 2). For *digital copiers* and *plain paper fax machines* a scanner unit has to be included into the system to digitize the initial analog original (Fig. 3).

Figure 1. Components of modern electrophotographic systems — Copier

Figure 2. Components of modern electrophotographic systems — Printer

Figure 3. Components of modern electrophotographic systems — Fax machine or digital copier

Together they all need a process unit to produce the copy or printout finally. Electrophotography is the technology used in the process unit in virtually all copiers commercially available today and is also the most prevalent printer technology.

2.1.1. *Electrophotography*

In electrophotography, a charge pattern replicating the light image is formed on a photoconducting film.

Charged, pigmented, thermoplastic particles, called toner, are selectively attracted to the charge pattern, thereby making it visible. The toner image is then transferred to the paper, and fixed by softening and fusing the toner to the paper. This imaging process is called the Carlson or xerographic process.

Other imaging principles in electrophotography include Electrofax, thermoplastic and overcoated xerography, ionography, and magnetography [1-7]. The major portion of the following section is devoted to xerography; the other forms of electrophotographic imaging are only briefly described.

Since electronic imaging is developing very fast it is recommended to consult the web sites of IS&T (The Society for Imaging Science and Technology) and SPIE (International Society for Optical Engineering) for updates on technical trends [4].

The six steps of xerographic imaging are (Fig. 4):

1. *Charging* of a high-resistivity photoreceptor in the dark.
2. Imagewise *exposure* of the charged photoreceptor; resulting in a charge pattern, the latent image, on the surface.
3. *Developing* the latent image by applying charged toner particles that come into contact with the charge patterns.
4. *Transferring* the toner image onto the paper or other media.
5. *Fusing* the toner to the media.
6. *Cleaning* and conditioning the photoreceptor for the next imaging cycle.

2.1.1.1. *Photoreceptor*

Description. The photoreceptor is the device or transducer (Fig. 4) that converts the light pattern into an electrostatic pattern. At minimum, it consists of a *photoconductive layer* on a *conductive substrate* (Fig. 5). The essential steps of photoreceptor operation—charging, charge generation, and charge transport—are also indicated in Figure 5. The first step is the production of a uniform electric field across the photoconductor layer by corona charging. In the second step, exposure, light photons incident on the photoreceptor in the pattern of the image are absorbed and generate conducting charge (electron-hole pairs). Under the influence of the electric field produced by the surface charge, these drift across the photoconductor, neutralize the surface charge, and produce the electrostatic image. The photogeneration and charge transport steps may be performed in separate material layers (Fig. 5 B) that are optimized for their respective functions. In addition to the charge generation and transport layers, other layers for adhesion or charge blocking (to minimize charge injection into the photoconductor) may also be present.

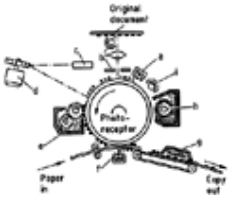

Figure 4. Schematic of a xerographic reproduction machine, indicating the important process steps

a) Corona charging; b) Light exposure; c) Laser; d) Rotating polygon; e) Development; f) Transfer; g) Fixing; h) Cleaning; i) Erase

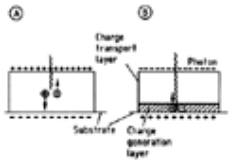

Figure 5. Photoreceptor structures

A) Single-layer photoreceptor; B) Dual-layer photoreceptor

In the single-layer photoreceptor, photogeneration can take place at the top surface or in the bulk. In two-layer photoreceptors, charge photogeneration takes place in the thin photogeneration layer which is located either at the bottom or at the top.

The earliest photoreceptor (used in the Xerox model D) was a *flat plate* similar to a photographic plate, which required manual handling. *Drums* allowed automatic machines of the type shown in Figure 4. *Flexible photoreceptors* were fabricated initially as sheets of film (Electrofax ZnO) [5], later as scrolls, and more recently as belts. Belts enable full frame flash exposure that increases the throughput speed of duplicators.

The essential characteristic of the photoreceptor is the photo-discharge curve, which relates light exposure to the voltage on the photoreceptor (Fig. 6) [6]. The dependence of photogeneration efficiency on electric field generally determines the shape of the photo-discharge curve [7], [8], which relates to the gray (density) scale reproduction characteristics of the xerographic system [9].

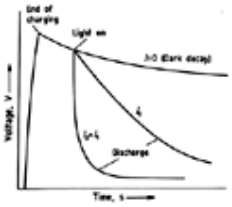

Figure 6. Photoinduced discharge characteristics of typical photoconductors showing the potentials at the end of charging step, dark decay and discharge curves

Photoreceptor Requirements. To be useful in xerography, a photoreceptor should meet the following criteria [10]:

1. The photoreceptor must hold the corona charge image in the dark. This requires trapping the surface charge and, in some cases, a blocking contact between the conductive substrate and the photoconductor. Additionally, the rate of charge loss in the dark (dark discharge) must be minimized.
2. In a *copier application,* the photoreceptor should discharge efficiently at all wavelengths of the visible spectrum. In *printer applications*, the wavelength of the laser or exposure system is fundamental, for example, 800 nm for gallium arsenide lasers. The figure of merit is the efficiency of photogeneration, which is generally less than unity and decreases with decreasing electric field.
3. The photogenerated charge carriers must traverse the photoreceptor in times that are short compared to the image development time. This sets a lower limit to the charge carrier mobility (velocity per unit electric field).
4. The charged, partially discharged, and residual voltages must remain stable with repetitive cycling during a multiple copy run. A condition termed *cycle-up* results from a buildup of residual voltage caused by accumulation of trapped charge through repetitive use. A condition termed *cycle-down* results from increased dark discharge with repetitive use. Charge trapping is caused by certain impurities, which necessitates stringent purity requirements ($<10^{13}$ impurity atoms per cubic centimeter).
5. Photoreceptor materials must be capable of being fabricated into large-area, defect-free films. Materials must be sufficiently stable to perform in a corona environment containing ozone, nitrogen oxides, and other effluents, as well as to withstand wear by the development and cleaning processes.

Brief Survey of Photoreceptors.
Inorganic Photoreceptors. Amorphous selenium (a-Se) photoreceptors, consisting of vacuum-deposited films of amorphous selenium 20 – 100 µm thick, on aluminum drums, were the most widely used photoreceptors in xerography during the 1970s and 1980s. Amorphous selenium meets most of the photoreceptor requirements, except sensitivity throughout the visible spectrum; it is highly sensitive to blue light only (Fig. 7). Addition of tellurium to a-Se increases the red sensitivity [11], and multilayer devices consisting of a thick a-Se charge transport layer overcoated with a thin, red-sensitive, selenium-tellurium layer are used commonly as photoreceptors [12-15]. Another weakness of a-Se is that it crystallizes readily at elevated temperature, and crystallized Se does not accept charge. Arsenic is added to a-Se to retard the rate of crystallization and increase red sensitivity [16]. Photoreceptors containing 40 atom % arsenic and 60 atom % selenium have high sensitivity in the entire visible spectrum and are employed in high-speed printers with helium – neon lasers or LED image bars (Fig. 7).

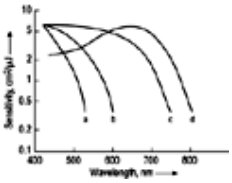

Figure 7. Relative spectral photosensitivities of inorganic photoconductors

a) a-Selenium; b) a-Selenium with 10 wt % Te; c) a-Arsenic triselenide; d) a-Silicon

A more recent addition to the family of inorganic photoreceptors is *hydrogenated amorphous silicon* (a-Si: H) [17-20] which is sensitive in the entire visible spectrum and has sufficient 780 nm sensitivity to be used with gallium arsenide lasers (Fig. 7). It is vacuum-coated by chemical vapor decomposition of silane (SiH_4). Like crystalline silicon, a-Si: H can be made positively or negatively conducting by doping with boron or phosphorus. Other inorganic photoreceptor materials applied in xerography are solvent-coated layers of *zinc oxide* [5], [21] and *cadmium sulfide* [22], [23] in an organic polymer binder.

Organic Photoreceptors. When CARLSON invented xerography, he employed sulfur and anthracene as photoreceptors, and the initial commercialization of his idea relied on inorganic photoreceptors [1]. The current trend is toward the use of organic photoreceptors because of their material variety, economy, and flexibility. High-speed copying machines use belts coated with organic photoreceptors, whereas personal copiers employ aluminum drums dip-coated with multilayered organic photoreceptors. The first organic photoreceptor consisted of a single-layer charge-transfer complex of an electron-donor polymer (polyvinylcarbazole) with an electron-acceptor (2,4,7-trinitrofluorenone) [24-26]. This charge-transfer complex absorbs visible light and enables charge transport of both electrons and holes. The photogeneration efficiency of this photoreceptor is relatively low and highly field dependent [27].

To enhance photosensitivity of organic photoreceptors dual-layer structures are employed, as with selenium – tellurium photoreceptors [28], [29]. Hole transporting polyvinylcarbazole is replaced by other electron-donor molecules in solid solutions of mechanically durable polymers such as the polycarbonates of bisphenol A or Z. Typical electron-donor molecules include triphenylamines and diamines [30-32], pyrazolines [33], hydrazones [34], [35], oxadiazoles [36], [37], and stilbenes [38]. Transport layers generally consist of a 40 – 50 wt % solution of the electron-donor compound, 10 – 25 µm thick, and have hole mobilities of 10^{-6} – 10^{-5} $cm^2\ V^{-1}\ s^{-1}$. Because these organic charge transport layers are transparent, the photogeneration layer is overcoated with the transport layer and the photoreceptor is charged negatively. The photogeneration layers consist of pigments such as phthalocyanines [39], [40], bisazocompounds [33], [41], thiapyrylium salts [42], [43], perylenes [36], [44], and many other pigments. The relative spectral sensitivity of several of these is shown in Figures 8 and 9. A very comprehensive survey on photoconductivity and carrier transport phenomena of both inorganic and organic materials can be found in [36], [46]. More recent studies and papers are summarized in [47] and [48].

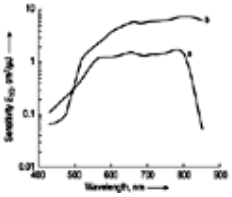

Figure 8. Relative spectral photosensitivities of organic photoconductors for printers

a) Metal-free phthalocyanine; b) Titanyl phthalocyanine

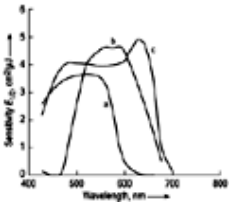

Figure 9. Relative spectral photosensitivities of organic photoconductors for copiers

a) Dibromoanthanthrone; b) and c) Bisazo pigments

2.1.1.2. *Corotron Charging*

The first step in the operation of a photoreceptor is to charge it to a (uniform) voltage by depositing charge on the free surface of the photoreceptor with a device called a corotron (Fig. 10 A). The simplest and most common corotron design consists of a thin wire (50 – 150-µm diameter) stretched between two insulating supports and electrically connected to a high-voltage supply [1], [2]. The wire is surrounded on three sides by a conducting shield spaced ca. 1 cm from the wire, as shown in Figure 10 A. A voltage of 3 – 8 kV on the wire produces a high electric field that ionizes the surrounding air, and ions are deposited on the photoreceptor surface, thereby charging it.

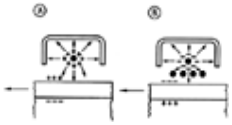

Figure 10. Schematic of a simple corotron (A) and a scorotron (B)

With a *positively biased wire*, electrons and negative ions in the surrounding air are drawn to the wire, causing secondary ionization consisting of electrons and positive and negative ions in the vicinity of the wire. Positive ions, primarily hydrated

protons, $(H_2O)_nH^+$ [49], are repelled from the wire and driven by the voltage difference to the shield and photoreceptor. With a *negatively biased wire*, electrons are emitted from the wire and ionize the air. The electrons attach to molecules so that anions are formed, predominantly carbonate ions, which are deposited on the photoreceptor [49].

The charging current flowing to the photoreceptor surface depends on the electric field and, hence, on photoreceptor voltage. The charging current is inversely proportional to the voltage. At some high voltage, all current flow to the photoreceptor is cut off. This current-voltage characteristic minimizes the voltage variation on the photoreceptor surface. It is exploited in a device called a *scorotron* (from screened corotron; Fig. 10 B), in which a wire screen or metal mesh is placed between the corotron and the photoreceptor. The cutoff voltage becomes the voltage applied to the screen, which provides a convenient way of controlling the voltage on the photoreceptor.

Primary charge rollers (PCR) (Fig. 11) are used in many desk top printers and copiers to charge the organic photoreceptor. The mechanism of the charging process is generally understood as ionization of air in the gap between the PCR and the photoreceptor. The advantages are that contact charging realizes extremely low ozone emission, compactness, and low charging voltage (below 1 kV) as compared with conventional charging devices, such as corotrons and scorotrons [50], [51]. There are two types of contact charge rollers. One consists of a steel sleeve with a conducting or semiconducting rubber coating leaving a small gap between the roller and the photoreceptor [52]. Charging is controlled by the electrical discharge in the vicinity of the gap. Stable operation and uniformity of surface potential across the photoreceptor is achieved by superimposing an a.c. or pulsed voltage over the d.c. voltage.

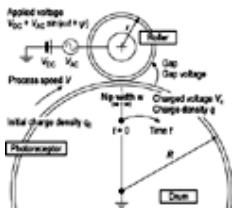

Figure 11. Contact roller charging subsystem

The other one is a softer, more brush type roller, which has a closer contact to the photoreceptor. Not only the electrical discharge, but also a charge injection at the contact area, controls the charging. The amount of charge injection depends on the effective size of the contact area and increases with the humidity in the ambient air.

2.1.1.3. *Light Exposure*

Light Exposure in Copiers. The model D, a manually operated copier, was the first commercial copier introduced by Xerox and employed standard camera projection optics to image and photodischarge the photoreceptor. This type of light exposure optics is still used in flash exposure systems of high-speed duplicators. The full image is projected on a flat area of a belt photoreceptor with a xenon flash lamp to freeze the image on the moving belt.

Copiers with *drum photoreceptors* require scanning optics that project a moving narrow strip of the original image through a slit on, and synchronously with, the moving photoreceptor surface. Many techniques have been used to accomplish this. One is to project the original image through a lens on a rotating mirror that scans the image across a slit synchronously with photoreceptor motion. Another involves scanning the original image with moving lamps and mirrors, thereby projecting the image through folded optics that maintain the focus of the original on the photoreceptor through a slit. In small, personal copiers, another approach is to move the original with rollers or a translating platen while the optics remain stationary. One advantage of this system is the use of very low-cost fiber optics in a bar, called selfoc lenses, in place of lenses and mirrors. Early copiers, with a-Se photoreceptors, had fluorescent tube lamps in scanning optics and xenon lamps in full frame exposures. Current copiers with scanning optics tend to have high-temperature incandescent tube lamps in place of fluorescent lamps.

Light Exposure in Printers. Printers discharge the photoreceptor digitally with light spots that are on or off. Digital imaging is performed in several ways. The majority of xerographic printers use a polygon laser flying spotlight exposure. The laser light beam is projected onto a rotating polygon mirror which causes the beam to be swept repetitively across the photoreceptor. By timing the sweeps to repeat after the photoreceptor has advanced the width of the beam spot, the surface of the photoreceptor is raster scanned much like the electron beam in a television tube. The beam is turned on or off by an optoacoustic modulator or by directly modulating a gallium arsenide diode laser. Other approaches used are image bar arrays of individually addressable light-emitting diodes, or liquid crystal and magneto-optic light valves [53]. The two most commonly used lasers, helium – neon and gallium – aluminum – arsenic, emit at 630 and 780 – 820 nm, respectively, whereas the light-emitting diodes used in image bars emit at 660 – 720 nm. Hence, photoreceptors for the xerographic printers require photosensitivities in the red to near-infrared spectrum.

2.1.1.4. *Image Development*

Electric Field of the Charge Image. After the charge and exposure steps, the image consists of a charge pattern across the photoconductor. This produces an electric field pattern with field lines flowing from the surface charge to the countercharge on the conductive substrate [54], [55]. Unless an electrode is closer than the thickness of the photoreceptor, the electric field is very weak above a large charge area. With a distant electrode, field lines extend into the space outside the photoconductor only at the edges of lines or solid patterns (Fig. 12) [56]. Because development occurs by attraction of charged toner particles along the field lines, the toner is attracted only on the fringe of wide (solid) image areas. To develop solid image areas, a development electrode must be introduced, as illustrated in Figure 12. Generally, the electrode is biased to a voltage higher than the white image background potential to prevent toner development in the white areas [57].

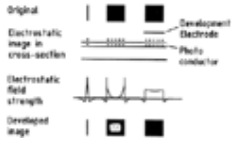

Figure 12. Effect of a development electrode on electric fields and toner development

The original image, electrostatic image in cross section, electrostatic field strength, and developed image are illustrated

Development Materials. The charge image is rendered visible by the development step. A typical developer consists of a black or colored powder called the toner which contains thermoplastic particles, typically 5 – 25 µm in size, with 5 – 10 % pigment to give the desired color [1], [2], [58]. The function of the thermoplastic is to fuse and fix the image to the paper by heat, pressure, or a combination of the two. Commonly employed thermoplastics are random copolymers of styrene with methacrylates or acrylates. Black toner contains a pigment dispersion of carbon black, about 1 µm in size; color toners contain mixtures of dyes and pigments of the required color.

The widely used, two-component development employs a mixture of a few weight percent of toner with larger size (ca. 100 – 700 µm) carrier beads. Carrier beads serve two functions; they provide (1) a method for mechanically transporting the fine toner powders and (2) a means of charging the toner. Because the fine toner powders are difficult to control and their leakage contaminates xerographic hardware, they are triboelectrically attached to the much larger and heavier carrier beads. With agitation, charge is exchanged between the toner and the carrier bead, and the toner is charged by triboelectrification. Typically, as many as 10^3 toner particles may be attached to a carrier bead.

In *charged area development* which is always employed in photocopying, the toner is charged to a polarity opposite that of the photoreceptor, and the high-potential areas are developed black whereas partially discharged areas produce gray densities. Printers can also employ *discharge area development*, in which the polarity of the toner is the same as that of the photoreceptor and only low-voltage or discharged areas are developed.

Magnetic Brush Development. Magnetic brush development is the most widely used toner development process. The development device consists of a series of magnets inside a nonmagnetic shell [1], [2], [56]. As the photoreceptor moves into the development zone, toner-covered carrier beads are transported by the shell (Fig. 13) counterrotating to the photoreceptor. The magnetic carrier beads tend to form brushlike filaments along the magnetic lines of force, as shown in Figure 14, hence the name magnetic brush development. The toner is stripped from the carrier beads by the electric field of the charge pattern, and toner-depleted carrier beads are returned to the mixing region where they are covered with toner particles. The carrier beads are coated with low surface energy materials, such as fluorinated polymers to optimize toner charging, minimize permanent adhesion of the toner to the carrier, and control the conductivity of the toner – carrier mixture. The effective developer – electrode spacing, which controls the field in large solid area development as discussed above, can be reduced from the roll-to-photoreceptor distance by the conductivity of carrier beads. By varying the degree of polymer coverage, the toner – carrier mixture can be made insulating (called insulating magnetic brush) or conducting [59]. The former enhances fringe-field development which is useful for fine lines, while the latter enhances solid area development.

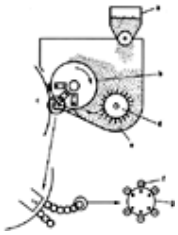

Figure 13. Schematic of a magnetic brush development system. The inserts schematically illustrate a developer brush fiber and a developer bead with toner attached.

a) Toner dispenser; b) Rotating shell; c) Photoreceptor; d) Paddle mixer; e) Developer; f) Toner; g) Carrier

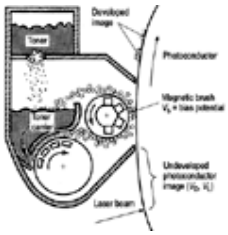

Figure 14. Magnetic brush developer unit. Toner is mixed and charged with carrier beads. Magnet arrays are stationary while the cylinder around them rotates. The cylinder is biased to assist in developing large exposed areas.

Print quality, as determined by line resolution, gray density range and uniformity, and white background density (or cleanliness), is controlled by several factors [6]. These include electrostatics (the shape of the photodischarge curve and the image and cleaning voltage difference or contrast), the size and charge distribution of toner particles, the toner concentration, and the developer flow rate. The copy darker – lighter buttons on office copiers adjust the electrostatics, and some modern copiers maintain copy quality by sensing and adjusting the electrostatics. Toner must be added constantly to the developer mixture to maintain a constant toner concentration. Too much leads to white background printout (dirt) and dark copies, and too little leads to light copies. Most copiers have sensors that measure toner concentration or image density and control the addition of toner. Although carrier beads are not used up, they must be replaced over regular intervals because constant use causes surface damage.

Other Development Systems.
Cascade development is an older two-component development technology not in use today [56], [60], [61]. The mixture of

carrier beads and toner is tumbled (cascaded) over the photoreceptor surface by gravity. Because the carriers used for this system are large, heavy beads of glass or steel, their size limits the developer electrode spacing to distances >500 µm, making this system unsuitable for solid area densities.

Mixing the thermoplastic polymer with magnetic iron oxide makes the toner itself magnetic, eliminating the need for carrier bead transport in magnetic brush development systems [62]. In this *single-component development*, the toner is charged either by induction or by triboelectric charging against a donor roll. Single-component development enables smaller and lighter developer transport hardware, used in personal copiers, because the carrier beads and mixing hardware take up most of the volume and require heavy magnets. Because magnetic iron oxide is brownish red, it limits the ability to produce bright color toners.

Aerosol development, also called *powder cloud development,* is one of the earliest forms of single-component development [63], [64]. Toner particles are charged by blowing them through fine-bore tubes constructed from metal or ceramic. A shortcoming of this process is that the toner charge distribution is broad and not easily controlled. This process is especially suited to develop X-ray images in xeroradiography or mammography where continuous gray scale, high-resolution images are required to reproduce slight differences in contrast potentials.

Electrophoretic or liquid development relies on charge exchange between micrometer-size toner particles and a liquid medium (in which they are suspended) to charge the toner [1], [2], [65], [66]. The liquid may be an alkane such as kerosene mineral oil, or isopar (a mixture of branched $C_8 – C_{14}$ isoparaffins) to which charge control agents are added. Liquid development produces high-resolution images because the particle size is small and can approach the image very closely [67]. A disadvantage is the delivery of an objectionable residual solvent onto the copy.

2.1.1.5. *Image Transfer and Fusing*

The developed image is transferred by placing a paper on the photoreceptor and applying a corona charge of the same polarity as the photoreceptor charge on the back side of the paper [1]. The reversed electric field breaks the adhesion between the toner and photoreceptor surface, transferring the toner particles to the paper. A biased roll can also be used for transfer.

To fix the transferred toner image to the paper, four types of fusers are employed commercially: hot roll, cold pressure roll, radiant, and flash fusers. Because rheological requirements for toners, such as melting point and melt viscosity, differ for the four types of fusers, toner design must take into account the fuser type.

Hot roll fusing is the most common type of fusing. Here the paper with the toner image passes between two rollers, one of which is heated, for example by an internal quartz lamp. To penetrate the paper fibers, the toner image must be heated to a temperature substantially higher than the glass transition temperature of the toner polymer. Since the paper spends only 5 – 10 ms in the nip between the rollers, it is desirable that the glass transition temperature of the polymer be low to minimize the requirements on the fuser heating power.

To prevent melted toner from sticking, the rollers are either permanently coated with a low surface energy material such as Teflon or repeatedly coated with a release agent such as silicone oil [68]. Alternately, the release agent, usually a polyolefin wax such as polyethylene or polypropylene, can be incorporated into the toner, eliminating the need for external application. Image gloss, an important consideration for copy appearance, may be controlled by the fuser coating. A noncompliant coating such as Teflon produces a high-gloss image, whereas a compliant coating such as Viton produces a low-gloss image.

In *cold roll fusing,* toned paper passes between two polished steel rolls at room temperature, and the image fuses by pressure alone. The main advantage of cold roll fusing is low power consumption, with no need for standby power. In *radiant fusing*, the toned paper is passed under a heated coil or quartz lamp. The residence time required for radiant fusing is 200 – 500 ms longer than for roll fusing. *Flash fusing* is similar to radiant fusing except that the toner is fused by a high-intensity light flash of approximately 5 ms duration. Finally, the toner can also be fused to the paper by exposing it to solvent vapors.

2.1.1.6. *Photoreceptor Cleaning and Erase*

Before resuming the cycle, the photoreceptor must be cleaned of residual charge and toner. Residual toner is removed by a brush, wiper blade, or magnetic roll (similar to that used in development). Prior to cleaning, the toner may be charge-neutralized with an a.c. corotron. The cleaning systems can have an auger or vacuum system that collects the removed toner, and in some cases circulates it back into the developer system. Residual charge is usually removed by exposure to a high-intensity lamp.

2.1.1.7. *Other Electrophotographic Imaging Systems*

Electrofax. Electrofax is the trade name of a xerographic system that uses a photosensitive paper on which the final image appears, eliminating the transfer and cleaning steps [5], [20]. The photoconductor layer consists of a layer of silicone resin binder, 5 – 15 µm thick, containing 50 wt % zinc oxide particles which give it a white color. Because zinc oxide absorbs only UV light, it is dye-sensitized with, e.g., rose bengal, fluorescein, or methylene blue, and the appearance of the paper is made as close as possible to plain paper by selecting an appropriate mixture of dyes to produce a light gray color.

To produce *transparencies*, direct formation of the final image on the photoreceptor has also been employed with transparent organic photoconductors. By using liquid development and omitting the transfer step, very high image resolution can be achieved. Another variant involves peeling the thin photoconductor layer, with the fused image, from the substrate and laminating it to the paper [69].

Thermoplastic and Overcoated Xerography.
Thermoplastic xerography produces images by surface deformation of a thermoplastic. The photoreceptor can consist of an

organic thermoplastic photoconductor or a photoconductive layer coated with a thin layer of thermoplastic material [70], [71]. In the latter, the charge pattern, formed by conventional means (Fig. 15 A), remains across both layers in the dark regions and across the thermoplastic layer in light-exposed areas (Fig. 15 B). By charging with an a.c. corotron to zero potential, the charge is removed in the unexposed region but remains in the exposed region (Fig. 15 C). Upon heating, the electric field in the exposed region causes the thermoplastic to deform into a wrinkling pattern (Fig. 15 D), and differential light scattering renders the image visible.

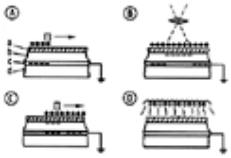

Figure 15. Process steps of overcoated thermoplastic xerography

A) Thermoplastic layer; B) Transparent photoconducting layer; C) Conductive layer; D) Supporting layer

Overcoated Xerography. A variant of the xerographic process is to form the image on a photoconductor overcoated with an insulating layer as described above, develop it with toner, and transfer it to paper by conventional means [2]. Several variations of this charge image formation have been proposed [23], [72], and one, the Canon NP process, has been used extensively in copiers throughout the 1970s and early 1980s. This process utilizes a photosentive element with a conductive base, a photoconductive layer, e.g., *cadmium sulfide* in an organic polymer binder, and an insulating top layer. The steps involved in forming an electrostatic image are shown in Figure 16. The difference to the Carlson process is that an additional a.c. corona discharge and the imagewise exposure take place simultaneously. After a final overall exposure the electrostatic image is formed on the insulator surface and can be processed by conventional xerographic methods.

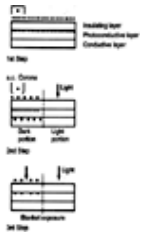

Figure 16. Electrostatic image formation by the Canon NP process

Another related process, called *TESI (transfer of electrostatic image)*, forms or transfers the electrostatic image on an insulating layer that is separate from the photoconductor [2].

Ionography is a xerographic process in which the light exposure step is eliminated and the charged image pattern is created directly on an insulating dielectric by ion deposition. An array of addressable charging pins or ion apertures is used to electronically create the charge pattern (Fig. 17), which is then developed by liquid or dry toner.

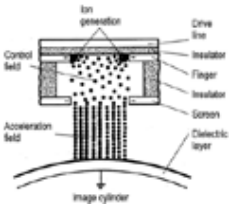

Figure 17. Schematic cross section of an ion generation site in the print cartridge. The three electrodes, the drive or RF line, the finger electrode, and the screen electrode, are separated by insulating layers.

TonerJet is a direct printing process where charged toner particles are deposited directly onto a plain paper surface to form a visible image pattern. An array of electrodes creates dot-sized electrostatic fields to draw charged toner particles through openings in a mesh, and deposit them in an image pattern on a plain paper surface.

A scheme of a TonerJet print zone is shown in Figure 18. Monocomponent magnetic toner is transported on a developer sleeve that carries the toner into the print zone while charging each particle by contact charge exchange with the developer sleeve material.

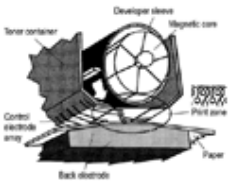

Figure 18. Simplified TonerJet print zone; inset: Etched circuit aperture array

A uniform electric field is created between a high potential on the back electrode below the paper and a low potential on the developer sleeve. That uniform field is modified by control potentials on individual ring electrodes in a two-dimensional array placed in the print zone. The dot-sized electrostatic fields draw the toner from the developer sleeve through apertures in the ring electrodes, and deposit them on the paper surface where they are fused by pressure or heat [74].

Magnetography is the magnetic analog of electrophotography. In this printing process, a magnetic latent image is created on a film of ferromagnetic material, using an array of magnetic inductive heads. The latent image is then developed with magnetic toner. A vertical magnetization, perpendicular to the ferromagnetic film, can be created by a pin-shaped magnetic pole aligned perpendicular to the film (Fig. 19). The return path for the flux is through a magnetic sublayer under the film and a large counter pole. The force attracting the toner containing soft magnetic material is proportional to the product of the magnetic field and its gradient and is by an order of magnitude smaller than in the forces in an electrophotographic developer unit [75], [76].

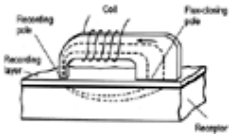

Figure 19. Energizing the recording head's coil creates a magnetic flux in the circuit composed of the head and drum cores. Because the field under the flux-closing pole is below the coercivity of the recording layer, a dot is recorded only under the higher-intensity recording pole

2.1.2. *Thermographic Printing*

In *thermal transfer printing*, heat generated from an electrically resistive source (the thermal transducer) works on a thermally sensitive material (the dye or ink) to create a permanent image. The image must be created before heat is lost to a thermal reservoir (the paper). The thermal transfer printing process is shown schematically in Figure 20.

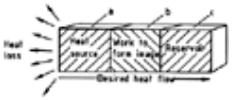

Figure 20. Thermal transfer printing process

a) Thermal transducer; b) Thermally reactive dye or ink; c) Paper

Basically, two thermal transducer technologies exist. One uses a transducer of electrically resistive elements contained in multilayer *rigid head structures*. In the more recently developed technology, the electrically resistive elements are part of a *ribbon structure* [85]. The ribbon structure normally contains a thermoplastic image-forming material on one side. Heat generated in the head or the ribbon transducer either melts a thermoplastic ink and transfers it to a receiving substrate. In direct thermal transfer printing, heat generated in the head causes thermally sensitive materials on the paper to react, forming a colored image. The technology using thermally sensitive paper is known as *direct thermal transfer printing*. Print heads can consist of a small array of heating elements that print a matrix character (e.g., 5 horizontal × 7 vertical (Fig. 21 A)); or a simple vertical linear array (Fig. 21 B). The head may also consist of a page width of elements printing up to a line of dots at a time (Fig. 21 C).

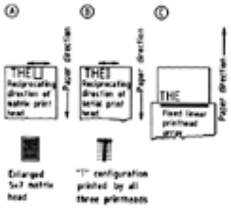

Figure 21. Thermal print head configurations

A) Matrix; B) Serial; C) Line

THE = thermal heat element

2.1.2.1. *Thermal Print Transducers*

Rigid thermal heads are generally subdivided into three types: (1) the transistor – resistor silicon mesa (i.e., plateau), (2) thin-film resistive heads, and (3) thick-film resistive heads [86].

Silicon Mesa Technology. The basic structure of the silicon mesa (patented by Texas Instrument) is shown in Figure 22 [87].

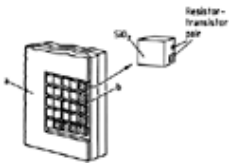

Figure 22. Silicon mesa technology

a) Semiconductor chip; b) Mesas (operating temperature 175 – 200 °C)

The printing elements have a driver transistor – resistor pair deposited on the base of a silicon chip. The silicon chip is part of a heater element array of air-isolated mesas. Each chip is epoxy-bonded to a ceramic substrate, which in turn is epoxy-bonded to a metal heat sink. A small base current through the transistor controls a larger current in the resistor. Joule heating of the resistor heats the silicon mesa. Mesa elements, when in contact with thermally sensitive paper, produce an image corresponding to the elements addressed. Silicon mesas are coated with silicon dioxide, which provides thermal isolation and prevents wear of the silicon. Operating temperature of the mesa elements is ca. 200 °C.

Thin-Film Technology. Thin-film thermal heads consist of a thin-film resistive heating element sandwiched between a protective layer and a heat-insulating layer. The heat-insulating layer is bonded to a heat sink. A typical structure is shown in Figure 23.

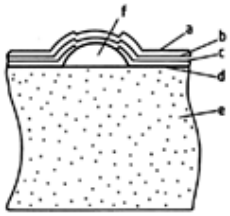

Figure 23. Structure of a thin-film thermal head [115], operating temperature ≈ 400 °C

a) Ta_2O_5-protective layer (5 µm); b) SiO_2-protective layer (2 µm); c) Aluminum lead (2 µm); d) Ta_2N resistor; e) Aluminum substrate; f) Glass glaze (40 µm)

The actual material used in a thermal head varies with the manufacturer. *Protective layers* are chosen from materials such as SiO_2, Ta_2O_5 [88], [89], or SiC [89]. For reliability, the wear life of a protective layer must exceed the cycle life of the resistor and other head parts.

The *resistor elements* are chosen from materials such as Ta_2N [88], [94], Ta – SiC [90], Zr – N [91], Ta – SiO_2, Cr – Si – O, Si–Ta [93], Ni – Cr, and Ta – Al. *Thermal insulating layers* can be glass or Al_2O_3 [93], and *conductors* are usually NiCr – Au [88] or Ni – Au [86].

All materials must be selected judiciously; similar coefficients of expansion in the various layers help prevent crack formation at high operating temperature. When employed for direct thermal printing, protective layers must be resistant to abrasion by the paper. The resistive layer must have a low temperature coefficient of resistance and good electrical stability under operating conditions.

The *Meander heat element* is an alternate thin-film heater design [95] (Fig. 24).

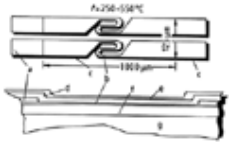

Figure 24. Plane and sectional view of the meander heat element

a) NiCr – Au lead; b) Exothermic elements; c) Ta_2N lead; d) Protective layer (Ta_2O_5); e) Protective layer (SiO_2); f) Heat-insulating layer (glaze layer); g) Alumina board

Meander elements are thicker than conventional rectangular thin-film elements and are considered more resistant to oxidative degradation. They provide more uniform heating of the protective layers, resulting in more uniform image density.

Thin-film elements can provide up to 16 dots per millimeter for high-resolution printing. A print speed of 20 lines per second can be achieved on thermal paper with a linear thermal head.

Head temperature varies from 250 to 550 °C depending on the duration of heating. Both parameters are determined by the total energy required to create the printed image.

Thick-Film Technology. In thick-film resistive technology [86], a resistor paste (borosilicate glass, lanthanum glass – ruthenium oxide) (DuPont 1211/1221, 1411/1421, 9145, Mathey-Bishop RRIOOR/RRIKOO [92]) is silk-screened onto a ceramic substrate (Tektronix alumina, Tektronix fosterite [96]) to produce the individual print elements shown in Figure 25. Another way to make the resistor is by laser machining a continuous bar of printed resistor paste [96]. Either way, the elements are approximately 25 µm (1 mil) thick. The element conductors (Owen Illinois 99+ Au [96]) must be thin enough to permit high resistor relief. However, their resistance must be low enough to accommodate the return path current, particularly when the route is through a common current return path.

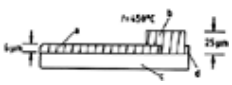

Figure 25. Print head for thick-film technology

a) Conductor; b) Resistor; c) Ceramic substrate

Current passing through thick-film resistors causes Joule heating while the head is in contact with the thermally sensitive paper. This heat creates a permanent image. Resistance drift of the resistor elements represents greater threat of failure than mechanical wear.

The required print head temperature is determined by the sensitivity of the thermal paper and the duration of heating. A peak temperature of 450 °C can be reached.

2.1.2.2. *Resistive Ribbon Technology*

In resistive ribbon technology, the heating function is transferred from the head to a ribbon structure. A 40-electrode flexible head selectively injects electrical current into a resistive ribbon [97]. The resistive ribbon is a 16-µm composite film of polycarbonate imbedded with electrically conductive carbon black. It has a sheet resistance in the range of 500 – 900 Ω per square [98]. On the side opposite the electrodes, the composite film is overcoated with a thin layer of aluminum (100 – 200 nm) applied by vacuum deposition. The electrically conductive aluminum assures that current flows anisotropically rather than randomly through the carbon-filled polycarbonate layer. During the printing process, current flows from the activated electrodes in the 40-stylus head through the conductive polycarbonate layer to the aluminum layer. Tunneling conduction has been described as the mechanism responsible for electrical conduction in the carbon black doped composite

material [99]. Localized heat is generated by (1) contact resistance between the electrode and the carbon-filled polymer, (2) bulk resistance of the carbon-filled polymer, and (3) interface resistance between the carbon-filled polycarbonate and the aluminum layer. The natural formation of a 4 nm oxide layer on the aluminum increases this interface resistance [100] and contributes to the overall Joule heating that occurs during a current pulse. A large return electrode minimizes Joule heating as the current flows to ground. The resistive ribbon configuration is shown in Figure 26. As can be seen from the figure, the 100-nm aluminum layer is overcoated with a thermoplastic transfer layer. Heat generated in the resistive substrate melts the thermoplastic layer, transferring it to the paper.

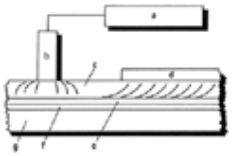

Figure 26. Resistive ribbon printing technology

a) Pulse generator; b) Print electrode; c) Carbon-loaded polycarbonate, conductive substrate, 16 µm; d) Grounded plate; e) Aluminum layer, 100 nm; f) Transfer coating; g) Paper

The operating temperature in organic ribbon structures must be similar to that attained in rigid head structures but must not affect the integrity of the resistive substrate. A modeling of ribbon temperatures with the 5796-PBH Advanced Statistical Analysis Program (ASTAP) is shown in Figure 27 [101], [107].

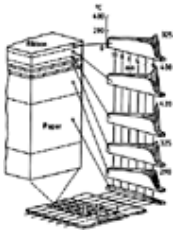

Figure 27. Simulated temperature distribution in resistive ribbon technology

The predicted temperatures within the structure after a 12.5 ms, 25-mA current pulse are highest directly beneath the electrode, and a maximum peak temperature (420 °C) occurs at the interface between the carbon-filled polycarbonate and the 100 nm aluminum layer. Ink temperature is predicted to be 325 °C.

Figure 27 predicts that the composite temperature should decrease to 100 °C in 10 ms; this is considerably lower than the glass transition temperature of the polycarbonate (T_g = 145 °C). The short cycle time permits the organic structure to retain its integrity during printing. Additionally, relatively slow lateral thermal spread allows the resistive ribbon to produce high-resolution printing.

2.1.2.3. *Imaging Materials and Mechanisms*

Thermal Paper. Thermal paper consists of a 5 to 10 µm-thick coating on a smooth base paper [105]. The coating contains a leuco dye and a phenol developer. The developer can be modified to improve the thermal stability of the coating [106]. Both components are shown on the page.

The color-forming reactants are dispersed separately from the binder (e.g., styrene – maleic anhydride copolymer) and other coating constituents, and then combined in the correct ratio prior to application. After application to the smooth base, the paper is gently dried below the activation temperature for the reaction between leuco dye and phenol.

When heated by a thermal head, both compounds react to form a color.

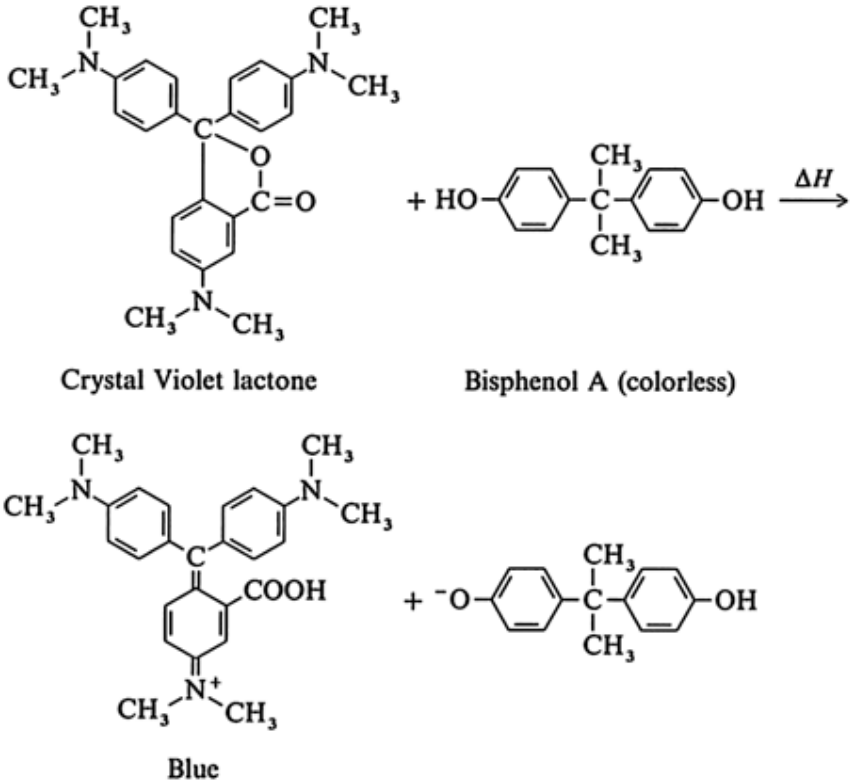

The higher the temperature (up to the point where all reactants are consumed), the higher is the optical density. Optical density as a function of temperature in a static 5 s test is shown in Figure 28 [105].

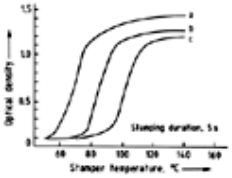

Figure 28. Optical density vs. temperature for three papers of different sensitivities

a) High; b) Medium; c) Low

Heating duration with a thermal head depends on print velocity and is on the order of 2 – 8 ms. Thin- and thick-film head temperatures greatly exceed the reaction temperature of the chemicals and compensate for the short print residence times. Temperatures as high as 450 °C can be reached with both thin-film and thick-film heads.

Print images are also developed on thermal paper if the paper is placed adjacent to the aluminum-coated side of a resistive ribbon transducer (without thermoplastic ink). Pulsing the head in the normal manner then results in an image.

Inks for Thermal Ribbons. In thermal ink transfer printing, imaging is accomplished by selectively pulsing a thermal head in contact with a polyester (Mylar, DuPont) or similarly stable base. The base contains, on the side opposite that contacted by the head, an ink layer which is in contact with paper (Fig. 29) during printing [109].

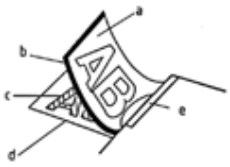

Figure 29. Thermal ink transfer printing

a) Transparent film; b) Ink layer; c) Transferred ink; d) Plain paper; e) Head

Ink transfer occurs as a result of the ink melting, wetting the paper, separating from the substrate, and solidifying on the paper. The melting point of the ink, heat capacity, ink thickness, melt viscosity, ink open time (i.e., the time the ink remains in a molten stage), paper roughness, and head pressure determine the amount of energy required for printing. Because in the thermal ink transfer process, heat generated in the head must pass through Mylar or a similar thermal insulator which adds to heat loss, inks with the lowest permissible melting point are chosen. The composition of 60 °C melting point ink is shown below, in weight percent [109]:

Carnauba wax	20
Ester wax	40
Pigment	20

Oil	10
Additives	10

Wax-based inks tend to smear and, for this reason, low-melting thermoplastic polymer-based inks are currently being developed to replace them.

Inks for Resistive Ribbons. Ink formulations for resistive ribbon printing can be relatively simple if they are for printers only. Ink composition is much more complex in typewriter applications where lift-off correction is desired.

For *printer applications*, polyamide resins with a carbon loading of ca. 8 wt % are very satisfactory [110]. Polyamide inks based on Versamid 940 or 950 have the correct melting-temperature viscosity and open time for high-quality images. Poly (vinyl acetate) (Aldrich Chemical), an amorphous polymer with a glass transition temperature of 30 °C and a melting point of 65 – 70 °C, can also be used as the thermoplastic binder for resistive ribbon inks. Addition of ca. 10 wt % Regal 330 carbon black and a trace of Methyl Violet produces an ink with good transfer characteristics and good optical density.

These inks are modified for use with the *resistive ribbon-saving feature*. Here, ribbon is fed at a rate slower than the movement of the head with respect to the paper. For example, if the rate of printing is 80 cps (characters per second) (8 inches/s, 10 pitch), then the ribbon is fed at 4 inches/s for a 2 : 1 ribbon saving. This feature improves ribbon mileage and provides a cost-per-character saving at increased print velocities. Inks used with the ribbon-saving feature must have good flow properties. They must print two or more characters from an amount of ink normally required for one character, with minimum loss of optical density of the print; thus, formulation adjustments are necessary. A low melting point (70 °C), low melt viscosity resin is blended with a medium melting point (90 – 100 °C), medium melt viscosity resin. These blends provide engraved character print quality at paper: ribbon velocity ratios of 2 : 1.

Ink formulations for *typewriter ribbons* must allow for lift-off correction; i.e., the image on the typewritten page must be removable. Resistive ribbon printing occurs by activating the electrodes at an energy level sufficient to transfer ink from the aluminum side of the polycarbonate substrate to the paper. The electrodes in the 40-electrode linear array are activated according to the desired print pattern, and electrical current passing through the ribbon to the aluminum ground causes Joule heating. The heat melts the ink, which, at the print pressure, transfers to the paper (Fig. 30).

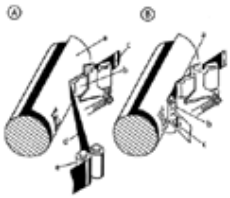

Figure 30. Basic components of resistive ribbon technology

A) Printing process: a) Paper; b) Print head; c) Ribbon supply; d) Ribbon take-up; e) Pinch rollers

B) Correction process: a) Electrodes; b) Correction roller; c) Correction block

In the correction mode, all 40 electrodes are turned on at a reduced energy level for the entire block matrix of the character to be corrected [111]. Energy is reduced for character correction by modulating the pulse width of the current to the electrodes (typically a 50 % duty cycle). The head pressure (2 N) is the same for both correcting and printing. The correction roller force is 2.5 N [111]. The block technique heats the ink to a temperature at which it permanently bonds to the character to be corrected, but not to the paper. The unwanted character is removed from the paper when the ribbon is lifted away. The block technique can correct any character with a single 40-electrode pattern.

An ink formulation that allows both printing and correction to be performed is relatively complex. When the ribbon is manufactured, the aluminum layer is first coated with a low melting point release layer [112]. This intermediate release layer is overcoated with a water-based pigmented ink layer consisting of ca. 11 % carbon black, and a mixture of ethylene – vinyl acetate copolymer and a poly(ethyl acrylate) – polyacrylonitrile latex.

The two-layer structure produces high-quality printing [250 printing elements per inch (100 µm) vertical ×360 printing elements per inch (75 µm) horizontal] [111] when sufficient thermal energy is used to give complete ink transfer. It also permits lift-off correction when the thermal energy is reduced appropriately (see Fig. 30).

2.1.2.4. *Comparison of Thermal Technologies*

A comparison of the various thermal print technologies based on the literature is given in Table 1, which is intended to serve as a guide. Data were obtained (or calculated) from a number of sources.

Table 1. Comparison of various thermal print technologies

Technology	Application	Resolution, dots/mm	Energy to print, J/cm^2	Peak temperature, °C
Silicon mesa	direct thermal paper			200 [87]
Thin film	direct thermal	8 [113]	4 [113]	400 [104]

	paper			
	thermal ink	3.6 [109]	6 [109]	
Thin film meander	direct	8 – 16 [102]	4 [102]	550 [109]
Thick film	direct	1 – 3 [103]	8 [96]	450 [114]
Resistive ribbon	thermal ink	9.4 (vertical) [111] 14.2 (horizontal)	3	420 [101]

Table 1 shows that all technologies require approximately the same order of energy for printing. They probably also have approximately the same level of thermal efficiency. The efficiency of *resistive ribbon thermal transfer printing* is ca. 6 %; that is, ca. 94 % of the heat generated in the ribbon is lost to the head, ribbon, ink, paper, and environment. In *direct thermal transfer printing*, 85 % of the thermal energy is dissipated in the head and 15 % goes to the paper [114]. What fraction of the 15 % is used to develop the image and what fraction is lost to the paper is not known, but thermal efficiency could be very similar to that of resistive ribbon printing [108].

Although thermal transfer printing is relatively inefficient, the energy required to print is still considerably less than that required for impact printers (the wire matrix, for example, is ca. 100 J/cm^2). On the other hand, the energy required for bubble or thermal ink jet (see Section Ink-Jet Printing) is on the order of 0.35 J/cm^2.

Selection of Appropriate Thermal Transfer Printing Technology. The type of thermal technology used depends on the application. Calculators, which require neither permanence nor high resolution, use *direct thermal transfer printing* with thick- or thin-film heads [115]. Direct technology is also used in typewriters and computer printers, with thin-film heads being employed when good-quality printing is required. If high print speed is required, a full-page linear array of thermal elements is employed in the printer. This type of head lends itself well to graphics.

If high-quality printing with good archival properties is required, *thermal ribbon printing* is employed; a high-resolution thin-film head or the resistive ribbon printing process is used. Both technologies have APA (all points addressable) and graphic capabilities, but with resistive ribbon printing a greater variety of papers can be used because of the head flexibility. It also has the potential for higher print speeds than thermal head printing because no duty cycle is required during operation. However, because the heating function is in the ribbon and not in the head, resistive ribbon printing supply costs are higher than thermal head – thermal transfer supply costs. The dual print and correction function offsets this higher cost somewhat for typewriters. The process of reducing the rate of relative movement of the ribbon past the paper lowers ribbon costs for printers considerably.

2.1.3. *Ink-Jet Printing*

In ink-jet technology, tiny drops of ink fluid are projected directly onto a surface for printing without physical contact between the printing device and the surface. The placement of each drop on the printing substrate is controlled electronically; and ink jet printing has become one of the preferred methods for printing data that has been generated or manipulated by a computer. Printing is accomplished by moving the print head across the substrate or vice versa. This direct printing process has been incorporated into many different low-cost and relatively simple printing mechanisms [116]. In addition, the noncontact feature of the technology allows for a variety of printing applications in which high printing speed, quietness, and substrate insensitivity are required. High-quality printing can be achieved by manipulating the drop size (with print resolution up to at least 40 dots per millimeter), by using inks containing multiple levels of dyes, or both [117]. Color printing is achieved by simply placing the proper colorant in the ink vehicles.

Ink-jet technology can be traced back more than 200 years. In 1754, L'ABBÉ NOLLE, Master of Physics for the French Dauphin, carried out research on the effects of static electricity on the flow of liquid drops from capillary tubes. In the late 1800s, Lord RAYLEIGH discovered that a fluid jet emerging from a nozzle breaks up into droplets and that the size of the droplets can be controlled by applying a uniform disturbance [118]. The first practical ink-jet apparatus was disclosed by THOMSON (Lord KELVIN) in his patent filed in 1867 [119]. Since then, patent activity has increased worldwide for various types of ink-jet marking devices and has accelerated in the last 10 – 15 years to the point that hundreds of ink jet related patents are issued worldwide each year [120-129]. A very large number of companies and many universities and research institutes have been involved in the development of ink-jet printing, and many will continue to invest heavily for the forseeable future [130-134].

The principle of drop generation varies among different ink-jet technologies. Electrostatic, magnetic, piezoelectric (acoustic pressure), electrothermal (Joule heating), mechanical microvalve, and spark discharge are among those that have been demonstrated. Regardless of the method employed ink jets can be classified into two basic categories: continuous ink jet and impulse (drop-on-demand) ink jet.

Continuous ink jet is characterized by pressurized ink which is emitted through a nozzle to generate drops of ink directed to the marking substrate in a continuous stream. A method of steering the drops to the desired printing location and disposing of unwanted ink drops is part of the print mechanism. This technology has been widely used for industrial marking, coding and labeling. It has also been adapted for very high quality color printing and computer-based high speed commercial printing. Continuous ink jet systems were the first to mature into well accepted products. Although this technology is capable of very high performance in both print speed and print quality, it has been somewhat limited because of its complexity.

Impulse ink jet differs from continuous ink jet in that the ink supply is maintained at or near atmospheric pressure. An ink drop is ejected from a nozzle only on demand when a controlled excitation is applied to the drop-generating transducer. A huge amount of effort has gone into the devlopment of drop on demand systems in the 1980s and 1990s, particularly in the

integration of the technology into small, relatively cheap printing devices that integrate hundreds of nozzles.

Applications of ink-jet printing appear endless. The printing process is termed nonimpact because no high mechanical impact pressure is involved in transferring ink to the receiving surface as in letterpress or wire matrix printing. Hence, ink-jet printing can be carried out on fragile objects such as eggs. Additionally, because it is nonimpact and no platen is required to support the receiving substrate, the process can be used to print on cans, bottles, textiles, and many other objects. Ink-jet printing is particularly suited to three- or four-color subtractive printing by having a set of nozzles for each of the cyan, magenta, and yellow (also black for four-color printing) primary colors. The process is also well-suited to more conventional printing such as word processing, facsimile printing, high-speed computer printout, black and white or color copiers (if attached to an optical scanner), typewriters, bar code, label printers, and graphics images. However, the technology has found its widest acceptance in the area of color computer-based printing, for which it is particularly suited.

Many products have been developed to satisfy a variety of these applications. *Continuous ink-jet technology* is used, for instance, in labeling, coding, and marking systems available from Videojet, Domino Printing Systems, Imaje, and others. Iris Graphics manufactures a color graphic drum printer specifically intended for graphic arts proofing. Scitex Digital Printing Systems is developing a relatively high speed, full color, computer driven commercial printing system using a version of this approach.

Impulse ink-jet technology is used in color printers developed by Tektronix, Hewlett-Packard, Canon, Lexmark, and many others [135].

2.1.3.1. *Continuous Ink Jet*

Three types of continuous ink jets are available: the Hertz type, the Sweet type [122] (or the raster-scan type), and the binary type. They differ from each other by the way the drops break off and the way they are steered.

Hertz-Type Continuous Ink Jet [123]. The Hertz continuous ink-jet device consists of a nozzle connected to a high-pressure ink supply, a charge electrode, and an aperture plate (see Fig. 31). During operation, an unperturbed stream of ink, emanating from the nozzle orifice, is broken into drops of *irregular size* because of Rayleigh instability. Turning the drop stream on and off to the print medium is accomplished by turning the charge electrode on and off. Only uncharged ink drops pass straight through the small aperture and reach the printing substrate. When ink drops are not wanted on the substrate during printing, a sufficiently high-voltage pulse is applied to the charge electrode located at the drop break-off point, and ink drops are inductively charged beyond their stability limit. At this point, the Coulombic force within each drop overcomes the surface tension, causing the drop to explode into a fine mist. The receiving substrate is shielded from this unwanted mist by the aperture plate. Unwanted ink can be discarded or recirculated to the ink reservoir through proper filtration because only 2 – 5 % of the expelled ink arrives on the print substrate.

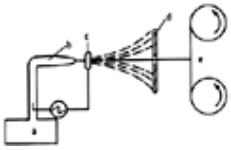

Figure 31. The Hertz continuous ink jet

a) Pressurized ink supply; b) Nozzle; c) Charge electrode; d) Aperture plate spray catcher; e) Print substrate

In this printing method, more than one drop is used to print a single spot on the print medium. Controlling the number of tiny drops per printed spot allows multiple gray level printing with very high-quality results. The nozzle size is typically ca. 10 µm. Pump pressure can be up to ca. 4 MPa to achieve a jet velocity of 25 m/s. Charge electrode voltage on the order of several hundred volts has been used. Drop frequencies up to 1 MHz have been reported. One shortcoming of this technology is that the aperture plate cannot effectively block 100 % of the unwanted mist. Some of this fine mist ends up on the printing medium as background and impairs print quality. This approach is the basis for the very high quality proofing printers made by Iris Graphics.[136].

Raster-Scan or Sweet-Type Continuous Ink Jet [137]. The Sweet-type continuous ink-jet device is analogous to an electron gun, except that the electron beam is replaced by a drop stream of conductive ink. The Rayleigh method of creating *uniform-size drops* is employed. As illustrated in Figure 32, a piezoelectric crystal is used to perturb the pressurized ink stream expelled from the nozzle orifice by the action of the pump. Under this regular excitation by the piezoelectric crystal, the jet stream breaks into uniform drops at a short distance from the nozzle orifice. A charge electrode is placed at the point of drop break-off, which selectively charges the ink drops according to the printing signal. The charged drop is steered away from the normal trajectory as it passes through the pair of deflection electrodes which provide a constant, high electric field. The larger the electrical charge on the drop, the more it is deflected. Hence, by controlling the sequence and level of drop charging, information can be printed by raster scanning of the charged ink drops. Uncharged drops (unwanted drops) are collected in the gutter and either discarded or recirculated.

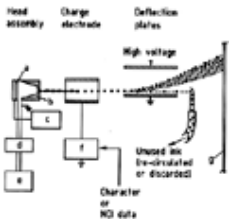

Figure 32. The raster-scan continuous ink jet

a) Piezo crystal; b) Nozzle; c) Crystal driver; d) Pump; e) Ink supply; f) Charging control; g) Paper

This technique projects one drop to one print spot on the receiving substrate. Various nozzle sizes have been used for

different applications. Typical drop size is about twice the nozzle diameter. The pump pressure is lower than that in the Hertz device, typically < 34 kPa. Drop-generating frequency can be up to several hundred kHz; drop velocity is ca. 20 m/s. Drop charging can be achieved with electrical pulses <1 00 V. The drops are deflected with the aid of an electric field of ca. 10 – 20 kV/cm. Most of the commercial marking and labeling devices use this technology

Binary Continuous Ink Jet [138] is similar to the Sweet-type except that uncharged drops are used for printing (Fig. 33). Charged drops are deflected into the gutter. The advantage of this approach, compared to the raster scan, is that charging of drops need not be very precise because the relatively large gutter can capture deflected drops with moderate deflection error. However, because printing is done with one nozzle per print spot, a multiple nozzle print head is normally employed. This technology is being used in a page-wide array of jets by Scitex Digital Printing for their digital printing press.

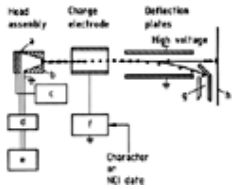

Figure 33. The binary continuous ink jet

a) Piezo crystal; b) Nozzle; c) Crystal driver; d) Pump; e) Ink supply; f) Charging control; g) Gutter; h) Paper

Technology Issues in Continuous Ink Jet. Although utlization of a continuous ink jet appears to be a simple direct printing process, the operational tolerance of each component of the printing device requires rather stringent control. This creates complicated engineering problems. For example, breaking up the jet stream is not always neat and straightforward. Smaller drops, called satellite drops, traveling at a different velocity can occur between two consecutive main drops. These satellite drops are charged differently and, therefore, can be deflected onto the printing substrate at undesirable locations. Print quality is degraded by such background spots. To some extent, satellite drops can be controlled by proper selection of ink properties, excitation methods, nozzle geometry, and jet velocity.

The break-off point of the ink stream, a short distance from the nozzle exit, depends strongly on ink properties that are a function of temperature. Proper charging of the drop can be achieved only when the break-off point resides within the charge electrode which is at a fixed distance from the nozzle. Hence, ink properties must be controlled precisely to ensure proper ink-jet operation.

When consecutive drops are charged, electrostatic repulsion exists among these drops, which leads to deviation from the normal flight path (Fig. 34). One solution is to maintain a large spacing between drops by either increasing drop velocity or reducing drop frequency. However, drops with too high a velocity cause splashing on the print substrate, and the throughput decreases when jets are operated at too low a frequency. Another approach is to charge only every other drop (the alternative uncharged drops are called guard drops).

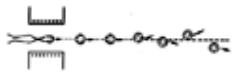

Figure 34. Charge repulsion of ink drops

Aerodynamic factors also affect drop motion. In a continuous stream of traveling ink drops, the trailing drop is pulled along by the influence of the wake of the leading drop. As the drops move toward the receiving medium, the very first drop experiences a higher drag than the trailing drops. Therefore, the first drop slows down and tends to merge with the following drops, causing misplacement of printing information.

When ink is recirculated in the continuous ink-jet device, the loss of ink solvent through evaporation causes a shift of ink properties. Furthermore, stopping and starting of the ink jet can also lead to crusting of ink in the nozzle orifice and splashing of ink on the charge and deflection electrodes. Contamination of the charge electrode and deflection plates not only produces corrosion on hardware but also affects the uniformity of the electric field, resulting in inconsistent drop projection. In multiple-nozzle ink-jet arrays, problems are compounded by compliance with channel uniformity requirements. An excellent description of the physics of continuous ink-jet technology can be found in [144-151].

Although, all of these problems can be managed through careful engineering, the increase in cost and complexity has undoubtedly minimized some of the attractive attributes of continuous ink-jet technology.

2.1.3.2. *Impulse (Drop-on-Demand) Ink Jet*

The three typical impulse ink-jet devices are the piezoelectric, the thermally driven “bubble-jet”, and the mechanical microvalve. This latter method generates huge drops and has been widely used for box labeling where large, crude characters are appropriate and for mural or advertising printing where the image produced is large and is viewed from a large distance.

Piezoelectric Impulse Ink Jet [125-127], [152]. Piezoelectric materials such as barium titanate and lead zirconate titanate (⟶ Ceramics, Electronic), once properly polarized, change their physical dimensions when exposed to an electric field through their surface electrodes. This is essentially the principle used by all piezoelectric impulse ink-jet drop generators. When an electric pulse is applied to the transducer, it induces motion on the piezoelectric element which generates a pressure wave inside the ink (Fig. 35). A portion of the pressure wave travels to the nozzle exit where it interacts with the ink meniscus, causing a drop of ink to be ejected from the nozzle. Ink drops are expelled only when the transducer is excited. Refill of the nozzle is accomplished by capillary action and the effect of inertia.

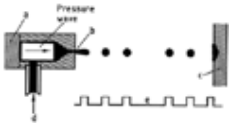

Figure 35. Operational principle of impulse ink jet

a) Transducer; b) Nozzle; c) Paper; d) Ink supply at atmospheric pressure; e) Data pulse train

One of the driving forces in the practice of this technology has been the continued miniaturization and integration of the drive elements and fluid cavities of the devices to pack more and more nozzles into less and less space. An analogy to the advances in integrated circuit technology is a good one and in fact many of the strategies employed and materials used for this integration come from that industry. One of the advantages of smaller devices is that they are inherently amenable to high drop generation frquencies. Drop frequencies of well over 20 kHz has been demonstrated and are continuing to climb. Energy requirements are extremely low, on the order of only a few microjoules per drop. Discussions of the physics of piezoelectric impulse ink-jet technology can be found in [153-158].

Until the 1990s, integration of piezoelectric ink jet devices into large arrays that were cost competitive with bubble jet technology had lagged significantly behind. However, efforts at Epson, Xaar, Tektronix and Spectra, among others, are rapidly closing that gap and are eliminating this as an inherent disadvantage of this method. One of the inherent advantages of piezo-driven devices is that they can accommodate a wide variety of inks, including phase change or hot melt ink technology.

Thermal-Bubble Jet [128]. Thermal-bubble jets differ from the piezoelectric type by their use of drop propulsion. A tiny resistive heater produced by thin-film vacuum deposition is used. Figure 36 A shows a typical thin-film heater structure on a silicon substrate (g). The heat delay layer (f), underneath the resistive film, improves thermal efficiency. The passivation layer (b, d) on top of the resistive film protects the heater from mechanically enhanced corrosion. Two typical designs have been used in the industry, the top-shooter (Fig. 36 A) and the end-shooter (Fig. 36 B).

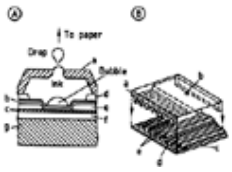

Figure 36. Structure and design of thermal-bubble ink jets

A) Top-shooter: a) Nozzle plate; b) Inorganic overcoats; c) Resistor; d) Polyimide overcoat; e) Conductor; f) Thermal barrier; g) Substrate

B) End-shooter: a) Glass body; b) Ink supply inlet; c) Ceramic substrate; d) Resistive element; e) Conductive lead

The principle of bubble jet operation is shown in Figure 37. When an electric pulse is applied to the heating element, Joule heating for a few microseconds causes the ink adjacent to the heater to be superheated and eventually exceed its nucleation threshold. A vapor bubble forms over the heater and expands rapidly because of the energy supplied from the superheated ink layer. The growing bubble acts like a piston pump to expel a drop of ink from the nozzle. As usable energy in the superheated layer is depleted, the vapor bubble starts to collapse and ultimately disappears. The nozzle is then refilled by capillary action and an inertial effect.

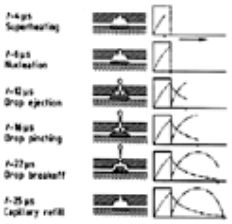

Figure 37. Operational principle of thermal-bubble ink jet

—— Voltage; –·–·– Temperature; –··–··– Bubble volume

As mentioned above, a significant advantage of this technology is the easy manufacturability of the device. Because batch fabrication techniques are used to produce the bubble jet head, hundreds or thousands of nozzles can be produced with precision and at low cost. A good review is given in [159].

An inherent limitation of bubble jet technology is the need for the ink to generate a vapor and to be able to survive the very rapid and very large temperature transients that it sees. This has been overcome quite well in many commercial systems that use aqueous ink, but does limit the selection of ink materials and types.

Mechanical Microvalve Printing. This technology is based on electrically actuated microvalves which cut off a drop of ink delivered by a pressurized supply. The orifice sizes for this technology are very large and the repetition rates very low. However, it is very well suited for the applications in which it is used and amenable to a very wide variety of inks.

Other Drop on Demand Technologies. Several other ways of generating drops have been demonstrated in the laboratory or have been adapted into pilot products that did not achieve market acceptance. These methods include jetting devices driven by spark discharge, electrostatic forces, magnetic forces, and acoustic lenses. They are described in the references listed and will not be mentioned further.

Technology Issues in Impulse Ink Jet. The advantages of impulse ink-jet technology, its low energy requirements and ease of integration into arrays, can also be pitfalls. Small amounts of air bubbles, foreign particles, and contamination at the nozzle exit can have devastating effects on the performance of the print head. Variations in ink properties, whether caused

by a change in environmental conditions (e.g., temperature and humidity), evaporative loss, or time-dependent chemical reaction in the ink, can render the print head nonfunctional. Channel to channel variability due to the above effects and manufacturing variability is also a huge issue for practitioners. The key to a successful impulse ink jet relies on the development of a unique ink, a simple automated maintenance procedure, and careful control of the manufacturing process for both inks and print heads. The requirements of ink-jet inks are discussed in Section Ink-Jet Inks.

As briefly mentiond above, material selection is particularly important in thermal-bubble jets because operating conditions are quite harsh. Repetitive high thermal stresses are imposed on the thin-film structure. Cavitational impact during bubble collapse erodes the top layers of the heater. Aggressive chemical corrosion also occurs at the elevated operating temperature. Discussions of the materials and failure mechanisms in thermal-bubble ink-jet devices appear in [161-163].

2.1.3.3. *Ink-Jet Nozzle Orifice*

In addition to the motion transducers, the nozzle orifice plays a crucial role in the operation of an ink-jet print head because jet and drop formation depend greatly on its quality and uniformity. These requirements are even more stringent for ink-jet arrays. Accurate positioning of the nozzles, uniform stream directionality, and identical drop break-off must be maintained across the array. The surface on the front side of the nozzle plate should not be wetted by the ink, yet wettable surfaces are needed for the inside of the nozzle cavity. These requirements place a significant burden on the manufacturing of the nozzle plate.

Drawn-glass nozzles as well as mechanically punched or laser-drilled metal nozzle plates were used formerly. The current techniques for nozzle plate fabrication are batch-manufacturing methods. Photolithographic etching on silicon, electroforming (see Chap. Photochemical Machining) with nickel, or laser ablation using a photomask and excimer laser technology are three of the most popular methods. Many nozzle plates and thousands of nozzles can be manufactured at one time by these processes. Details of these methods have been described [159], [164].

2.1.3.4. *Ink-Jet Inks*

The requirements for ink-jet inks are different and in some ways much more stringent than those for inks for traditional impact printing methods (see Section Thermographic Printing). The patent literature on ink jet ink technology has grown exponentially over the last 20 years and a wide variety of novel ink chemistries and applications have been proposed. For the practitioner of this technology the field is still wide open and the increasing demands placed on ink jet printing systems require increasing sophistication and complexity in the ink systems.

Physical Properties. The physical properties of ink-jet inks should be fully compatible with the ink-jet technology used because ink is an integral part of the printing mechanism during operation. Inks used in *continuous ink jet* are generally electrically conductive (resistance < 1000 Ω/ cm) and low in viscosity (< 4 mm^2/s). *Impulse ink-jet inks* need not be conductive but can require higher viscosity (< 20 mm^2/s) for better damping of pressure wave reverberation after drop ejection. *Thermal-bubble jets* require, in addition, ink components that are thermally stable above 300 °C. In many applications, purified versions of dyes and colorants may be required to avoid ionic corrosion of print head hardware and first-drop delay because of undesirable absorption phenomena or chemical reactions at the nozzle meniscus or elsewhere in the system.

Printing Property Requirements. No long-term problems such as crusting, corrosion, precipitation, or first-drop delay should occur in the ink-jet print head in which ink resides for the life of the printer. Ideally, printing on a variety of print substrates should produce the same high-quality (sharp, dense, and precise) results with rapid dry time, waterfastness, and lightfastness. Finally, ink-jet inks should be nontoxic, have low flammability, and offer a good and stable shelf life. The ink should show good chemical, smear, and rub resistance and high optical density.

Three types of ink-jet inks are available, (1) water-base inks, (2) oil – nonaqueous-base inks, and (3) the hot-melt system [165].

Water-Based Inks. The dilemma with aqueous ink jets is the desire to never have the inks dry out while in the print head, but have the inks instantly dry when they are printed on paper or a transparency. Water-based inks usually contain > 70 % water. Small amounts of humectants such as glycols are added to reduce the rate of evaporation and prevent precipitation of dyes when evaporation occurs at the orifice. Biocides are added to prevent the growth of microorganisms. For *continuous ink jets*, some salt must be included in the formulation to obtain the necessary electrical conductivity. However, because high conductivity generally leads to electrochemical corrosion, basic corrosion inhibitors may be needed. For *impulse ink jets*, corrosion is a smaller problem because conductivity is not required. Thickeners such as water-soluble polymers may be used to increase the viscosity of inks formulated for piezoelectric impulse jets. These polymers can also enhance print quality on the substrate by limiting drop spread by various mechanisms. However, these thickeners can cause maintenance problems because restarting the jet after it has been out of use can be difficult if the wrong materials are chosen. To shorten dry time, surfactants or penetrants may be added. However, print quality degradation must be avoided because feathering usually occurs on the print substrates with these additives. Hundreds of patent applications representing a huge amount of work have been devoted to overcoming these problems and achieving high print quality on so-called “plain papers”. The quality of prints obtained from aqueous ink jet printers on typical office papers has radically improved in the past decade. However, for color printing applications the best image quality is still achieved on specially coated papers designed as part of the printing system that uses aqueous inks.

Oil – Nonaqueous-Base Inks. This type of ink can contain more than 50 % nonvolatile vehicles such as glycols and fatty acids or their esters. They have low toxicity, minimal corrosive effects, and low volatility. First-drop delay and long-term maintenance are normally not problems. The viscosity of these inks is higher than that of water-based inks, which makes them most suitable for application in *piezoelectric impulse ink jets*. However, because drying of these inks on paper (or other print substrates) relies on absorption into the paper fibers (or the substrate matrix), less than optimal print quality usually occurs on printing paper because the ink bleeds along certain directions (e.g., long fibers). However, this property has recently been greatly minimized by the use of ink systems which contain pigments that precipitate from the ink upon printing

and do not tend to follow the solvent as it wicks out into the paper. Oil – nonaqueous-base inks are not used in continuous ink-jet applications because of the difficulty of dissolving high concentrations of salts in the ink to make it conductive. However, many solvent based inks are used in continuous jets for industrial marking applications. These systems are particularly good for dissolving high polymers that give the inks good durability on substrates such as glass and metals. Because the solvents are usually polar organic solvents such as methyl ethyl ketone, methanol, ethanol, and the like, they can dissolve certain salts in high enough quantities to provide the appropriate conductivity. Another interesting development in this area is the adaptation of UV curable ink systems to ink jet marking and coding technology. Conceptually, this eliminates any volatile solvents from the process, and in fact, these ink systems are now available in commercial marking applications.

Hot-Melt Inks [165] are solid at room temperature. They have also been referred to as phase-change or solid inks. Print heads using these inks are kept at elevated temperature so that the ink contained in the print head system is in the liquid state. Molten ink is ejected in discrete drops and flies through the air as a liquid. The very low mass of the ink drops compared to the substrate enables extremely rapid ink solidification upon reaching the receiving surface so that very fast dry times and little feathering are obtained. Print quality is usually outstanding, and print substrate sensitivity is minimal. The initial implementation of this technology was in devices that jetted the ink directly on to the final substrate. However, the most widely accepted commercial use of the technology is in color printers that use the system in an offset printing fashion. In this implementation, the ink is initially jetted onto an intermediate drum, on which it solidifies but still remains soft and malleable at the drum's elevated temperature. After the complete image has been assembled on the drum, it is transferred in one step to the final substrate by means of a pressure nip. The transfer process is very analogous to the final transfer step in offset lithographic printing or to toner transfer from the imaging drum or belt to the substrate in photocopiers or laser printers. Unlike those technologies, offset phase-change ink jet printing requires complete, clean transfer of the image, which is achieved by the use of a sacrificial oil layer on the drum. This process provides for high dot spread while maintaining tight control of the final dot size and shape. It also provides (in combination with the ink) for high image durability and a minimization of the final ink film thickness, which were two large limitations of the technology as originally implemented. The process also is amenable to high speed printing and a simple, realizable paper path.

The ink vehicles are typically waxes with melting points of ca. 80 – 95 °C. The selection of colorants for hot melt inks is very stringent because their solubility in the waxy carrier is usually low. The waxes, as well as the colorants, must be nontoxic and environmentally safe. The colorants must have the right shade of cyan, magenta, yellow, or black, and be soluble in the waxes that are usually not particularly good solvents for colorants. The colorants and inks must be thermally stable for extended periods while the inks are molten when the printer is held in the ready-to-print state. The resulting prints are expected to have excellent fastness to light and good resistance to abrasion and scratching.

Advantages of hot melt inks over aqueous inks are their lack of drying at the ink jet nozzle. Hot melt ink's rapid solidifying of the ink drop gives precise control of dot gain and excellent image quality on almost any media. Prints made with hot melt inks are not marred by water. The loading of the solid ink sticks into the printer is a simple, clean operation.

Printer start-up is slower from a complete power off because time is required to liquefy a block of solid ink. A power on – off cycle causes volume changes during an ink phase change in the print head, usually requiring a maintenance purge to remove any ingested air bubbles. These problems are mitigated by leaving the power on continuously. The printer goes into a standby mode when not being used for several hours. In the standby mode the temperature is dropped to just above the freezing point of the ink. In this manner, the printer can be brought back to the ready-to-print state very quickly. Hot melt ink jet printers have found commercial success particularly in local area networks of approximately a dozen personal computers, where the printer is used often and the speed of the printer is more than sufficient to keep up with the demand.

2.1.3.5. *Colorants Used in Ink Jet Inks [166]*

Colorants for ink jet inks originally came from the textile trade inventory and, more commonly, from a limited list of food dyes. Because the ink jet orifice is about the width of a human hair, ink developers quickly discovered the dye purity requirements for ink jets are much more stringent than for coloring fabric or even food. Both textile and food dyes require significant purification before inks made from them will perform reliably in ink jet printers. After the first generation of ink jet inks, ink formulators discovered that purification of commercially available dyes alone was insufficient. Even after purification, the dyes lacked some desired properties, such as wetfastness or lightfastness. Color chemists are developing a second generation of colorants specifically designed for ink jet applications.

Blacks. C.I. Food Black 2 was one of the first black dyes used in *aqueous ink-jet inks* [167]. Printing with ink made with Food Black 2 gives an image easily marred by water. To overcome the lack of wetfastness, ICI chemists eliminated or replaced some of the hydrophilic sulfonic acid groups with carboxyl groups in Food Black 2 [168], [169]. The modified dye has much improved wetfastness while retaining the desirable characteristics of the original dye.

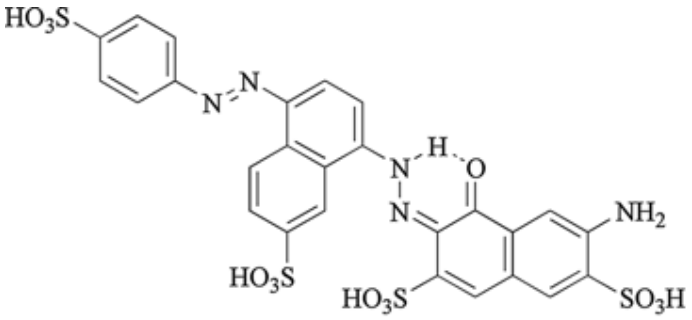

Food Black 2

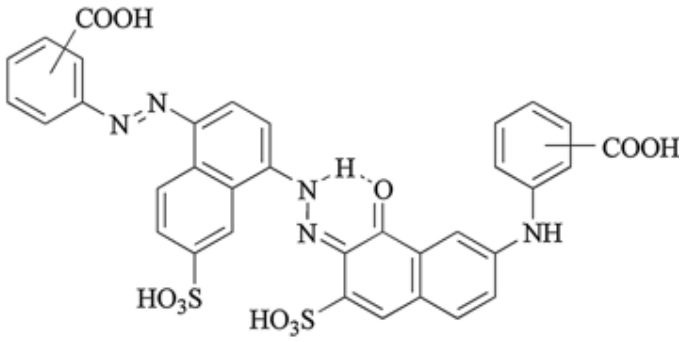

Black Dye modified for better wetfastness

There is significant patent activity in the field of pigmented aqueous inks and major improvements have been made in pigment dispersions [170]. Dye-based inks tend to have brighter colors and pigment-based inks generally have duller colors but better lightfastness. The main reasons why color pigments have not yet come into general use include the difficulty in producing good dispersions and getting reliable performance in ink-jet print heads. To achieve better lightfastness than that provided by dye-based inks, Du Pont has pushed the development of aqueous pigment dispersions for ink-jet applications [171]. The only commercial application of aqueous ink jet inks using colored pigments has been in wide format ink-jet printers. Encad utilizes disposable print heads printing at 300 dots per inch. Mutoh uses piezoelectric driven “permanent” print heads with much larger than normal apertures for the printing of outdoor signs about 100 dots per inch. In the printing of signs, lightfastness is more important than a large color gamut. In the wide format printer application, the ink-jet print heads are used almost continuously. It is yet to be determined in typical office and home use, whether ink-jet printers will be used often enough to prevent clogged nozzles due to the settling and agglomeration of the color pigments. Improved dispersion formulations appear to have mitigated some of the reliability concerns of pigmented inks, but not yet solved the lack of bright primary colors. Within Hewlett Packard's ink jet color set, only the dye-based black ink has been changed to the Du Pont carbon black aqueous dispersion. Hewlett Packard's cyan, magenta, and yellow inks are still dye-based [172].

Black dyes consisting of a 2 : 1 dye – metal complex, such as C.I. Solvent Blacks 27, 28, 29, 35, 45, etc., afford excellent thermal stability, solubility, and lightfastness in *hot-melt ink applications*.

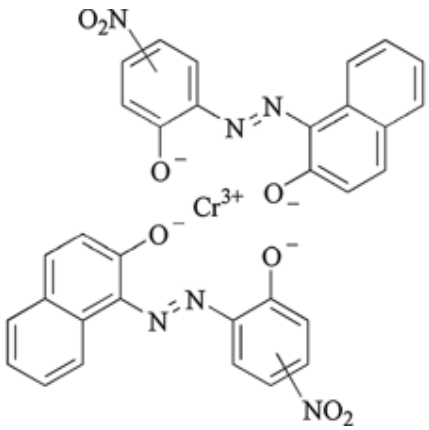

Solvent Black
2:1 Azo dye to
metal complex

Cyans. The greenish blue color required for a good cyan in an aqueous ink was originally accomplished with C.I. Acid Blue 9 (also known as Food Blue 2):

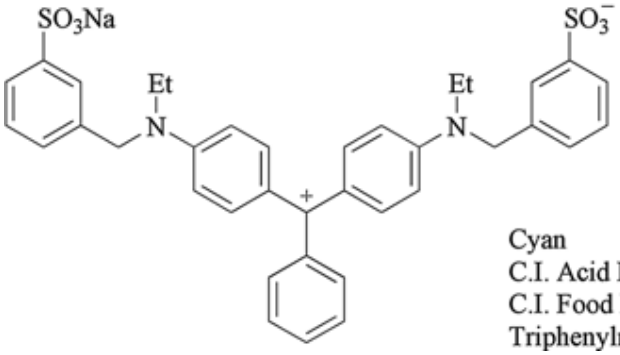

Cyan
C.I. Acid Blue 9
C.I. Food Blue 2
Triphenylmethane

Unfortunately, this triphenylmethane dye has relatively poor lightfastness. Most inks for current ink-jet applications use a cyan dye based on the copper phthalocyanine chromogen. The type and number of substituents on the chromogen determines the ink system for which the dye is suitable. Aqueous inks have used C.I. Direct Blue 86 (see below, W = X = SO_3Na, and Y = Z = H) and C.I. Direct Blue 199 (W = X = SO_2NH_2, Y = SO_3Na, and Z = H) [173].

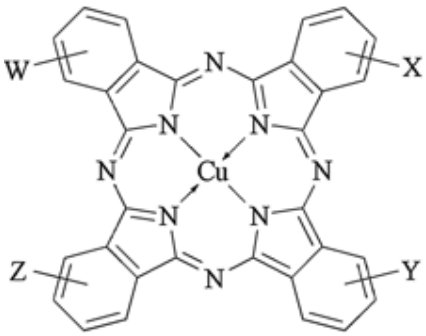

Cyan
Copper phthalocyanine

Oil and hot melt inks require more oleophilic substituents. such as those found in C.I. Solvent Blues 25, 44, 48, 67, 70, etc. These cyan dyes normally have two to four sulfonamide substituents. In C.I. Solvent Blue 25, for instance, W = X = Y = Z = SO_2NHR where R is a six carbon aliphatic tail. The other phthalocyanine cyan dyes differ only by the number of sulfonamide groups present and the aliphatic chain length [174]. In the manufacture of these cyan dyes the sulfonic acid ammonium salt is formed as byproduct instead of the desired sulfonamide. Increasing amounts of the polar ammonium salt decrease the dye solubility in nonpolar oils and waxes. In pigment applications the material most commonly used is C.I. Pigment Blue 15:3 where W = X = Y = Z = H.

Magentas. Bright magentas are commonly produced with the *xanthene chromogen*. Xanthene dyes are tinctorially strong (generating a lot of color per unit weight of dye), thermally stable, and produce clean, bright, almost ideal magenta shades that are difficult to reproduce at the same cost with other chromogens. However, xanthene dyes are notoriously poor in lightfastness.

C.I. Acid Red 52 (A = B = SO_3^-) produces a strong, bright magenta shade in aqueous inks.

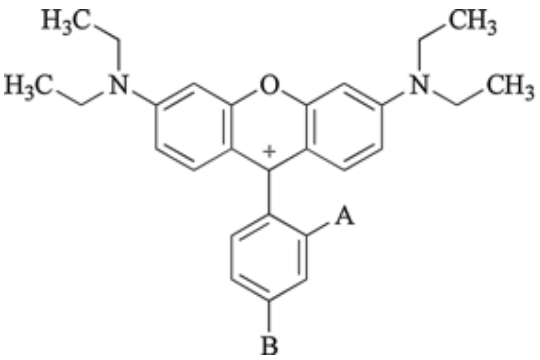

Magenta
Xanthene

In some applications, the less bright, but more lightfast C.I. Reactive Red 180 has been used in combination with the poor lightfast Acid Red 52 [175].

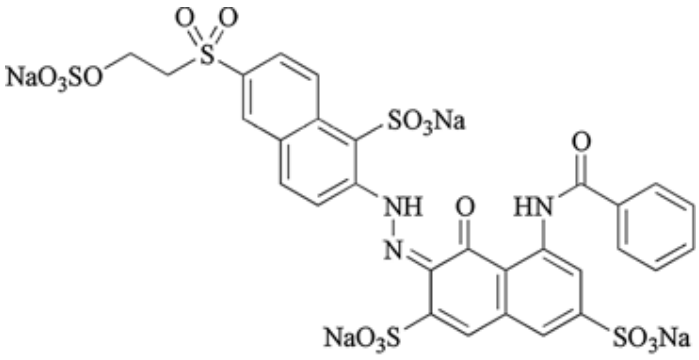

Magenta
C.I. Reactive Red 180

When the water-soluble, acid form of C.I. Acid Red 52 reacts with a tertiary alkyl primary amine, it produces $^{+}NH_3CR_3$ (where R_3 represents a total of 12 to 14 carbons), which is soluble in oil and wax-based inks. The dye – amine complex is quite stable to oxidation [176]. A magenta dye successfully used in Tektronix' hot melt inks is also based on the xanthene chromogen. It is C.I. Solvent Red 49 (A = COO^{-} and B = H), which requires an acid medium to develop color.

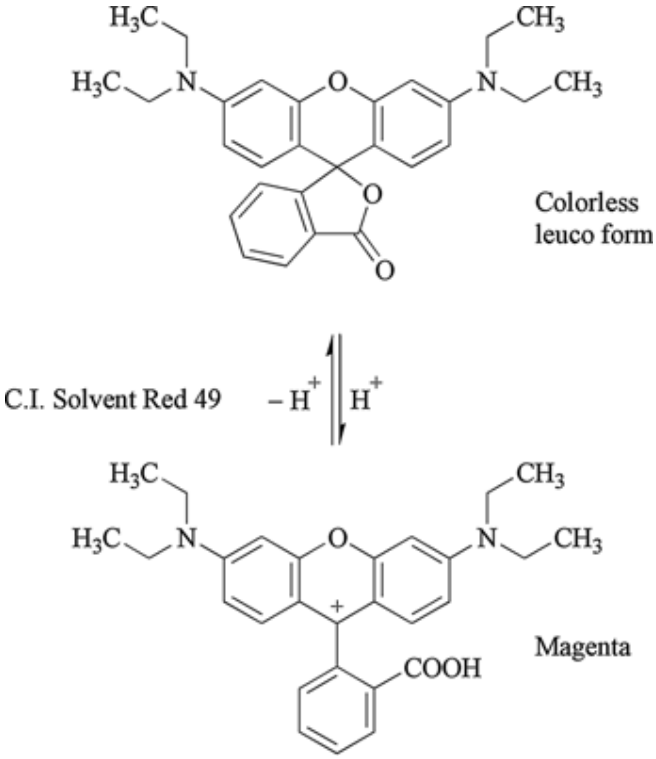

If hot melt ink prints using Solvent Red 49 are laminated and the adhesive comes in direct contact with the ink, some laminate adhesives will cause a significant loss in color by enabling formation of the colorless lactone (leuco form). It is best to use the laminates recommended by the printer manufacturer.

The anthraquinone chromogen has been used to produce acceptable lightfast magentas. C.I. Disperse Red 60 (see below, X = H, Y = OC_6H_6, and Z = OH) and C.I. Solvent Red 172 (X = $C_6Br_2H_2CH_3$, Y = H, and Z = OH) are soluble in hot melt inks:

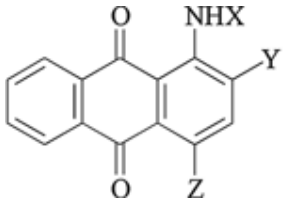

Magenta
Anthraquinone

There are several drawbacks with anthraquinones. They are not nearly as tinctorially strong as the xanthenes and therefore require more dye to produce an equivalent color level. The amount of anthraquinones that can be added to an ink is limited by their tendency to crystallize and migrate particularly under the influence of fingerprint oils. Crystallization and migration of the dye causes an objectionable shift in the original color of the print. In pigment applications the material most commonly used is C.I. Pigment Red 122.

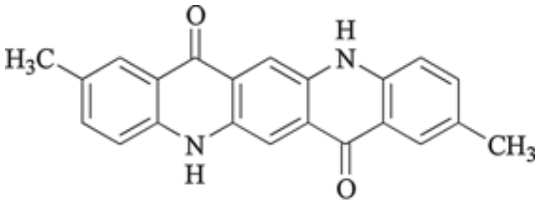

C.I. Pigment Red 122
Quinacridone

Yellows are easily obtained from many azo derivatives and azo dyes can be used in aqueous, oil, or hot melt inks. C.I. Acid Yellow 23 and C.I. Direct Yellow 86 have each been used as a process shade yellow in aqueous inks.

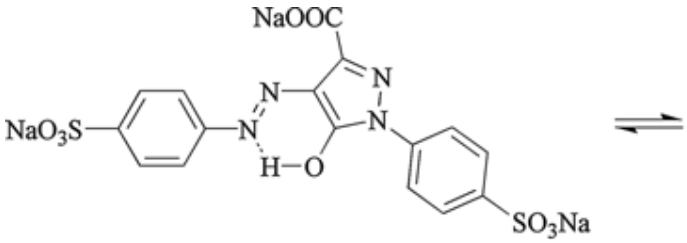

C.I. Acid Yellow 23

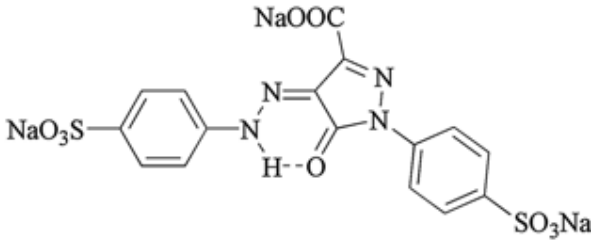

Hydrazone tautomer of
mono azo pyrazolone dye

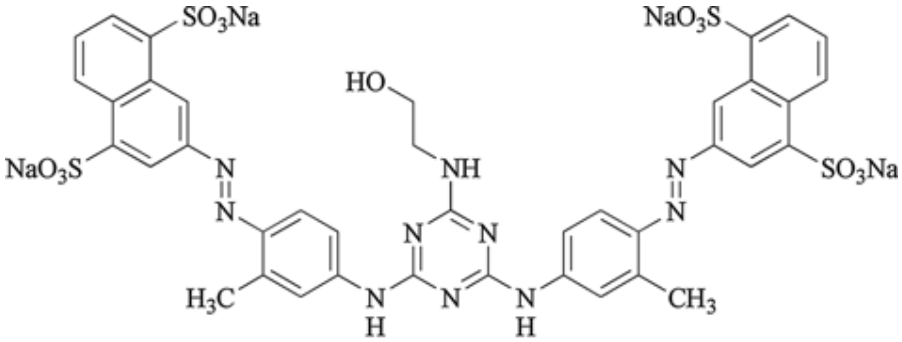

C.I. Direct Yellow 86

In general, yellows from azo derivatives are not as thermally stable as anthraquinones. Yellows that can form the hydrazone tautomer such as C.I. Acid Yellow 23, C.I. Disperse Yellow 119 (see below, where A = H, B = NO_2, R = Et), and C.I. Solvent Yellow 162 (see below, where A = $SO_2NHCH_2CH(Et)(CH_2)_3CH_3$, B = H and R = *n*-butyl), appear to be more thermally stable than azo dyes that can not readily form the hydrazone tautomer.

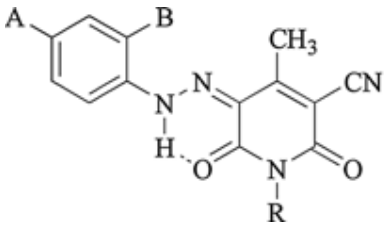

Hydrazone tautomer of
mono azo pyridone dye

Solvent Yellow 146, as Orasol Yellow 4GN from Ciba, is structurally similar to C.I. Solvent Yellow 162 and has been successfully used in hot-melt inks [177].

In pigment applications the material most commonly used for a yellow shade is C.I. Pigment Yellow 17 (see below, where X = OCH_3) or the redder yellow, Pigment Yellow 12 (X = H).

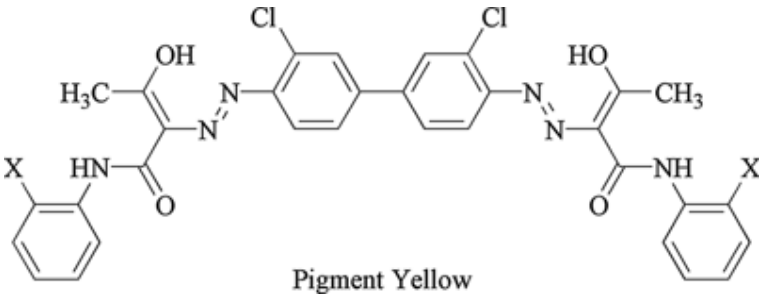

2.1.3.6. *Color Printing with Ink Jets*

Ink-jet printers are excellent for color printing because the amount of colorant required to create the image can be metered precisely with each drop of ink. The potential applications of ink-jet color printers are many. They range from limited color business graphics to full color prints from conventional silver halide color negatives or still video cameras (solid-state cameras which take still pictures like conventional cameras) [178].

In color ink-jet printing, the substractive system (see ⟶ Photography) is used to present a variety of colors to the eye [179]. Typically, amounts of red, green, and blue are subtracted from a white reflecting (paper) substrate. This is accomplished by applying proportions of cyan, magenta, and yellow dyes that absorb red, green, and blue, respectively. By overlaying saturated yellow, magenta, and cyan primary inks in appropriate combinations, saturated patches of red, green, blue, and black colors can be obtained.

Creation of more complex colors requires the addition of primary colors with varying degrees of *saturation* (white component). Total saturation denotes the absence of a white component; pink is a desaturated red. In ink-jet printing, saturation adjustment is affected by placing the transparent ink on only a portion of a small white picture element. This element should be small enough so that the eye cannot resolve the spots of primary ink themselves but integrates over the element and reacts to the composite spectrum.

In color printing, hue relates to the dominant absorption wavelength and determines whether the color appears blue, green, or orange. *Lightness* refers to the relative amount of reflected light. Adding black to red decreases lightness [180-182]. These principles are illustrated in Figure 38.

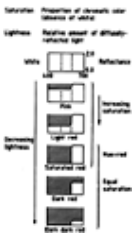

Figure 38. Elements determining chromatic expression

Ink and Paper for High-Quality Color Ink-Jet Printing. The complex list of requirements for ink-jet inks described in Section Ink-Jet Inks applies to inks for color printing but with additional requirements for the dye or colorants. To reproduce a large gamut of colors, dyes must be chosen in which the peak wavelength corresponds to cyan, magenta, and yellow.

In addition to providing a wide color range, the inks must be colorfast and pH stable, and they should not aggregate at the concentrations used [179], [183]. To obtain high-quality images from water-base inks applied from an ink-jet nozzle, the properties of the surface on which printing occurs must allow controlled spreading and penetration of the ink. In this case, ink drops of specific dimensions and color can be printed. A photographic-grade base paper, coated with barytes, to which a thin film of poly(vinyl alcohol) has been applied, is an excellent substrate for ink-jet printing if ultrahigh-quality color images are desired.

Creating Colors by Using Ink-Jet Matrix Printing. Consideration is given here to *full color ink-jet printing*. The straight forward seven-color printing used for business graphics and charts is only a small subset of full color printing. Color print data can be obtained from *stored data* or from a *color scanner*, as shown schematically in Figure 39; the latter describes ink-jet matrix printing. Additionally, the matrix printing method is used for continuous ink-jet printing as well as impulse printing [178], [179].

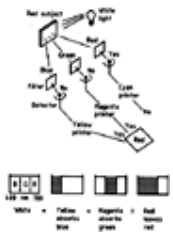

Figure 39. Substractive color reproduction

To demonstrate how a variety of composite colors can be achieved, a small white matrix composed of four elements (shown in Fig. 40) should be considered. The matrix is small enough so that the eye does not resolve the individual elements but

integrates over the whole matrix. By covering one element of the matrix with yellow ink, one-fourth of the reflected blue intensity is absorbed. The visual effect is a yellow-hued, very light, desaturated color (Fig. 40 B). If yellow coverage is increased, the hue remains the same, but saturation increases and lightness decreases (Fig. 40 C). A black component can be included by adding an equal amount of cyan and magenta as demonstrated in Figure 41 A, and a hue shift can be attained by adding unequal amounts of the primary inks as shown in Figure 41 B.

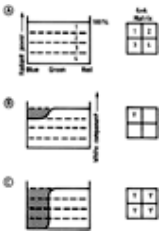

Figure 40. Varying saturation

A) White paper: hue neutral, lightness maximum

B) Light yellow: hue yellow, lightness high, saturation low

C) Saturated yellow: hue unchanged, lightness reduced, saturation maximum

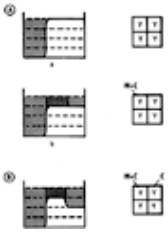

Figure 41. A) Varying lightness: a) Saturated yellow; b) Hue yellow saturation unchanged, lightness decreased

B) Varying hue: hue shifted to green, saturation full, lightness decreased

One means of determining the amount of ink to be printed in ink-jet printing is to compare the analog level of the scanner to a preset threshold value. If the scanner falls below the threshold at any print element, a drop of ink is placed in that element. This basic print or no-print access is limited with respect to quality. To copy continuous tone prints, such as photographs, which contain varying levels of color, a means of reproducing these levels such as the *threshold matrix technique* must be employed.

Threshold Matrix Technique. Associated with the print matrix is a multilevel threshold matrix, with threshold levels selected so that the color scale is reproduced correctly. An example is shown in Figure 42.

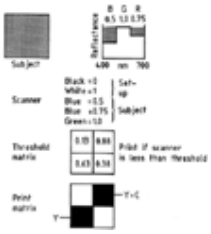

Figure 42. Threshold printing

Here, 50 % of the original blue is reflected. Thus, to satisfy reproduction, 50 % of the blue must be absorbed from the four-element print matrix of the reproduction. This is accomplished by printing two elements with blue-absorbing yellow ink. The decision to print is made by comparing the scanner and threshold values, and printing occurs on elements where the scanner is below the threshold setting. The same logic is applied to green – magenta printing and red – cyan printing to obtain the same red, green, and blue reflectance in the copy as in the original. In this example, a four-element gray scale matrix is described.

Techniques Required for High-Quality Color Reproduction. By using the multilevel threshold matrix technique, good-quality color reproduction is obtained. The limitations of this method relate to drop size, drop placement errors, and nonideal dyes used in inks.

Drop size has been adjusted experimentally in black and white printing to produce a resolution of 57 pels/mm (i.e., printing elements = no. of drops per millimeter). This resolution is excellent; however, problems of reliability of the printing device are severe. To obtain a resolution of 57 pels/mm, nozzle diameters of ca. 12 µm are required. With such nozzles, clogging occurs frequently, even with sophisticated ink filtration procedures.

Drop placement errors, which are inevitable when the multilevel threshold matrix technique is employed, can be alleviated by using error carry or error diffusion techniques. The multiple error correction algorithm (MECCA) is a method available for this purpose [184]. In the MECCA method, drop printing is controlled by a single-level, simple threshold, but the error incurred at each scan – print element is carried to the next element. To be more specific, the analog level of the scanner indicates some level of gray (color lightness), and the print result after thresholding is either a color or white, depending on whether or not a drop of ink is printed. The error then is the difference between the desired gray (color lightness) and the resultant color or white. This error is determined and added to the scan value of the succeeding element. Further print enhancement is obtained by weighing surrounding elements.

This method not only reproduces gray and color levels, but also reduces some of the objectionable texture and patterns of the threshold matrix technique. Edge definition is sharper and moiré patterns are decreased significantly. However,

considerably more data processing is required.

The problem of reproducing color with inks containing real dyes is that the yellow, magenta, and cyan inks have secondary absorptions in unwanted regions of the visible spectrum. If ideal inks could be used, the amount of ink required to reproduce a color could be calculated in a straightforward way based on the principal absorption bands of the primary inks. For example, if the green scanner measures some amount of green absorbed in the original, then by knowing the green absorbance of magenta ink, the amount of magenta ink required to obtain a similar green absorption can be determined. However, cyan and yellow inks also have some secondary absorption bands in the green region, and if some of these inks must be printed with magenta, an excess amount of absorbed green occurs in the reproduction. Other secondary additions compound the color-matching error.

If secondary absorption can be calculated quantitatively, it can be compensated in a number of ways. One simple way is to hold back a preset amount of ink of low equivalent density while maintaining full density reproduction of the saturated primary colors. A more accurate method of reproducing color is to include secondary absorption in color-matching equations. Hence, when a certain density of red is required, the secondary red absorption characteristics of the yellow and magenta inks are taken into consideration in a quantitative way. Algorithms have been derived to do this for red, blue, and green densities [185]. Application of these algorithms to the scanned data permits extremely accurate color reproduction. An ink-jet color-copying robot is shown in Figure 43.

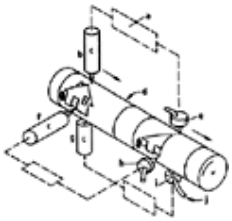

Figure 43. Ink-jet color-copying robot

a) Electronic interface; b) Red separation scanner; c) Photomultiplier tube (PMT); d) Original; e) Cyan ink-jet printer; f) Green scanner; g) Blue scanner; h) Magenta printer; i) Yellow printer; j) Ink supply

Undercolor Removal. The black component of dark colors can be produced by applying all three primary inks to a print element. With undercolor removal, the three primary inks required to create black are replaced by a single amount of black ink. This requires the addition of a black print head to the cyan, magenta, and yellow print heads.

Such a system has several advantages. One is that no print element has more than two layers of ink. This saves ink and reduces printing problems caused by large amounts of wet ink on the printing surface. Another advantage is that more freedom is allowed in formulating the three primary color inks because the need to produce a good three-color black is eliminated. One disadvantage is that drop placement must be accurate because the black drop must not fall on the same element as the primary colors, which would negate their effect and degrade color quality.

2.2. Technical Copying

Technical copying refers to the copying of technical documents: technical drawings, construction designs, sketches, drafts, patterns, graphics, parts lists, and so on. Unlike office copying, the documents produced in technical settings are often characterized by their large size.

Indeed, special photochemical methods have always been used to copy large-size originals, and the blueprinting technique remained in use until the 1920s. Blueprinting is based on the reduction of iron(III) salts and the formation of Berlin blue. It was finally replaced by diazotyping. Other photochemical procedures for special applications, such as the Dylux process, have come into use more recently.

Documents of smaller size can, of course, be copied by using electrophotographic office copiers, which are available even for sizes as large as DIN A 0. However, diazotyping is still a method used for copying technical papers because of its simplicity, versatility, and cost effectiveness.

The change has taken place in the preparation of technical drawings [computer-assisted design (CAD) technique] is also of considerable importance. These documents are no longer filed as drawings on paper or film, but are stored in digital form. When required, the file can be used to produce drawings with the aid of a plotter.

2.2.1. *Diazotyping*

The diazotyping process is based on two fundamental properties of certain aromatic diazonium salts:

1. They can react with aromatic coupling components or those with enolizable hydrogen to yield dyes (coupling reaction). This reaction shows a strong pH dependence.
2. On exposure to light, they lose nitrogen and, consequently, their coupling ability (photolysis) [186-188].

2.2.1.1. *Photolysis*

Photolysis proceeds as follows:

$$\mathrm{Ar{-}N_2^+X^-} \xrightarrow{h\nu} \mathrm{Ar^+ + X^- + N_2}$$
$$\mathrm{Ar^+ + H_2O} \longrightarrow \mathrm{ArOH + H^+}$$

Heterolytic cleavage of the C–N bond occurs from the excited singlet state and is the rate-limiting step. Subsequent reaction

of the carbonium ion with the nucleophilic partner proceeds very rapidly.

In a nonpolar environment (e.g., in a polymer matrix), homolytic photolysis is also considered:

$$Ar-N_2^+X^- \xrightarrow{h\nu} Ar^\cdot + N_2 + X^\cdot$$
$$Ar^\cdot + X^\cdot \longrightarrow ArX$$

The quantum yield from the photolysis of commercially useful diazonium salts is, without exception, < 1. Until now, sensitization has been achieved only in exceptional cases (e.g., in solution or in a special polymer matrix) [189]. Hence, suitable diazonium salts must absorb in the emission range of strong light sources (e.g., mercury lamps) and, in addition, should be sufficiently stable and produce light-decomposition products that are colorless.

Examples of commonly used diazonium salts are:

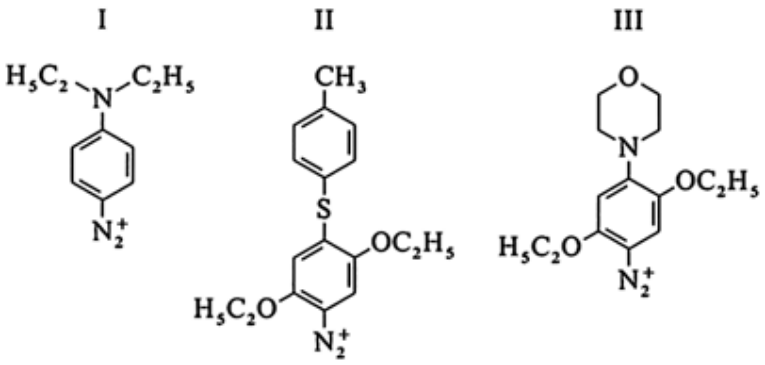

These diazonium salts with electron-donor substituents in the para position absorb in the “actinic range”, i.e., in the near ultraviolet (380 – 420 nm) (Fig. 44). The light sources used for diazotyping are mercury high-pressure lamps, as well as special fluorescent lamps (superactinic or ultraactinic). Because an irradiation energy of 50 – 100 mJ/cm^2 is required, diazotyping materials are used exclusively for contact copying. They can be processed in daylight and do not require a darkroom.

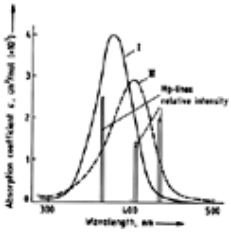

Figure 44. Light absorption of diazonium salts I and III and line spectrum of the mercury lamp

2.2.1.2. *Coupling Reaction*

A barely visible yellow diazo picture is discernible on the diazotype material after it has been exposed to light under a transparent original. This image can be destroyed by further exposure and has much too little contrast for practical use. In the second step, development, the coupling reaction converts the diazo image to a high-contrast, light-stable color picture. Hence, diazotyping is a positive process: i.e., dark parts of the original are reproduced as such.

The coupling reaction proceeds as follows [190]:

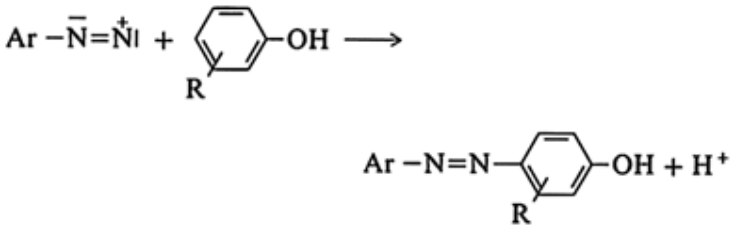

A base is required to convert the phenol or naphthol to the corresponding phenolate or naphtholate. This base must be added during development. Conversely, acids are added to the diazotype materials to prevent early coupling and stabilize the diazonium salts. The coupling component is more important than the diazonium salt for the color of the resulting dyes. For example, blue shades are obtained with naphthols; compounds with enolizable hydrogen or monohydric phenols give yellow coupling products; resorcinol and its derivatives produce brown to red shades; and very dark dyes are obtained from phloroglucinol. Black printing diazotype material represents a special case because it is produced by mixing various couplers, usually blue, yellow, and brown. Examples of commonly used couplers are:

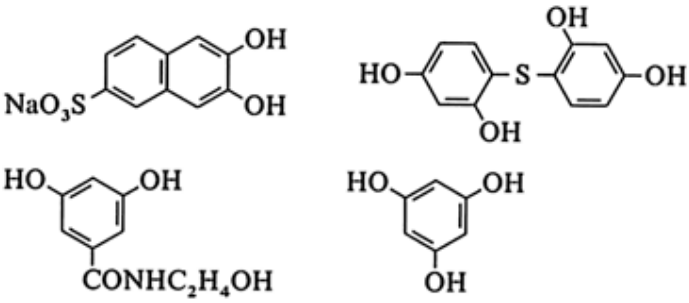

2.2.1.3. *Production of Diazotyping Material*

Opaque or transparent paper, transparentized paper, lacquered paper, and polyester films are employed as base materials.

Special wood-free, sized papers are used. They are usually precoated with an aqueous dispersion of fine filling material (white pigments; ⟶ Pigments, Inorganic) and binders before the light-sensitive coating is applied. The precoat homogenizes the surface of the paper, avoids penetration of light sensitive solution into the paper and improves contrast. Transparent papers are frequently prelacquered (e.g., with cellulose esters) to improve transparency and mechanical strength. The most extensively used synthetic films are polyesters because they are dimensionally and mechanically stable. They are usually chemically pretreated to achieve the adhesion of the lacquer layer. The functions of the layers are to accept the diazo coating and to allow writing with both pencil and ink.

After pretreatment of the base material on one or both sides, it is coated with the solution of diazotype chemicals. Pretreatment and coating are carried out in one working process in large coating machines that have several coating and drying stations. The running speed varies from a few meters per minute to >100 m/min, depending on material. *Aqueous coating solutions* can be used on opaque paper of 40 – 210 g/m^2. However, solutions in *organic solvents* are necessary for lacquered paper and for film. The film thickness usually varies from 36 to 125 µm.

2.2.1.4. *Processing*

Diazo printing machines with different degrees of efficiency are available for copying. High-performance printers are largely automated and capable of releasing folded copies. They achieve processing speeds >20 m/min [191]. Two different development techniques are available:

1. In *one-component materials*, the light-sensitive layer contains only the diazonium salt and stabilizers. The coupler is added during development as a buffered aqueous solution. Two different qualities of paper are used. One quality needs to be dried in a high-performance developping unit immediately after development (semiwet method). For the other quality 3 g/m^2 of developer is sufficient for application on one side so that drying can be dispensed with altogether [192].
2. In *two-component materials*, the light-sensitive layer contains not only the diazonium salt and stabilizers, but also the couplers. Development is carried out by passing the material through an ammonia atmosphere in which coupling occurs (dry method). The ammonia technique is more versatile and simpler in the choice of materials: black, blue, red, and brown printing types can be employed.

Two-component material can also be developed by applying a small amount (approximately 2 g/m^2) of an alkaline organic solution [193]. However, the copying quality of ammonia development is not achieved in this process.

Annual consumption of diazo printing material in the Western world is estimated at 1.1×10^9 m^2.

Some important producers of diazotype material are:

Azon Corporation, Johnson City, New York.
OCE Imaging, Charleston, Illinois.
Dietzgen Corporation, North Brunswick, New Jersey.
Oce van der Grinten NV, Venlo, The Netherlands.
Ozalid Company, Loughton, United Kingdom.
Regma SA, Arques-la-Bataille, France.
Ricoh Company, Tokyo, Japan.

2.2.2. ***Other Photochemical Systems***

The *Dylux method*, [194], [195], developed by DuPont, has attained some importance, especially in the graphics industry as proof material. This method requires only light and no chemicals for processing. In *normal processing*, selective exposure to UV light (λ_1) produces the corresponding dyed picture (negative method). Upon subsequent overall exposure to visible light (λ_2), the UV-sensitive substances are deactivated and the copy is fixed. Conversely, *positive copies* are obtained if the selective exposure to visible light is followed by overall exposure to UV light.

Dye formation occurs as a result of the photochemical oxidation of a leuco dye (**1**) by an oxidizing agent:

$$\underset{(1)}{\text{Leuco dye}} + \underset{(2)}{\text{Bisimidazole}} \xrightarrow[\lambda_1\,(250-370\ \text{nm})]{h\nu} \text{Dye} + 2\ \text{Imidazole}$$

(1) e.g., 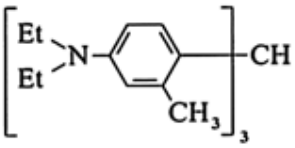 as tosylate whereas

(2) may be

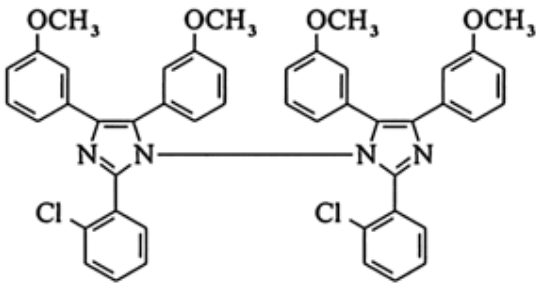

The energy of irradiation required to attain a color density of 1.0 is 15 – 50 mJ/cm^2.

A second photoredox system consisting of a quinone (**3**) and a hydrogen donor (**4**) is required for inactivation. The quinone is reduced photochemically to hydroquinone, which is a radical trap and stops further dye formation.

$$\underset{(3)}{\text{Quinone}} + \underset{(4)}{\text{HR}} \xrightarrow[\lambda_2\,(400-500\ \text{nm})]{h\nu} \text{Hydroquinone} + \text{R}$$

$$\text{Hydroquinone} + \text{Bisimidazole} \xrightarrow{\text{Dark reaction}} \text{Quinone} + 2\ \text{Imidazole}$$

(**3**) may be pyrene quinone and phenanthraquinone; (**4**) can be a polyether, a nitrilotriacetic acid derivative, or a triethanolamine ester.

Inactivation requires 0.5 – 2 J/cm^2 of light energy. Dylux material is available on paper as well as on film base. It is suitable for the reproduction of continuous tone, line, and screen pictures.

Another interesting technique has been developed by the Mead Corporation [196]. This method is based on the *joint microencapsulation* of a monomer, an initiator for photopolymerization, and a dye precursor; sheets are coated with these microcapsules. During exposure, the monomer polymerizes in the exposed areas and the microcapsules harden.

If the exposed sheet and a receptor sheet are subject to pressure, the capsules that were not hardened burst and their contents produce a dye on those areas of the receptor sheet that were unexposed. Hence, a positive copy of the original is obtained.

This system can form the basis for a *color-copying technique* (cycolor) if three different types of microcapsules are employed, each containing a dye precursor for one of the three primary colors in combination with an appropriate selectively absorbing initiator. No chemicals are required for processing. However, this system has not yet been adopted commercially.

2.2.3. *Silver Process*

The silver halide process has been used for about 150 years to reproduce images and documents of all kinds and is described extensively elsewhere (see ⟶ Photography).

The silver halide process is capable of producing negative and positive copies with excellent archival permanence and remarkable resolving power. Further useful properties are the wide range of speeds (i.e., sensitivity) and contrast values obtainable.

Silver halide materials are not used for office copying because of the availability of electrostatic photocopiers using plain paper (see Section Electrophotography). However, the silver process remains for special and professional copying applications. For example, pre-press copy operations such as page make-up demand high quality halftone reproduction and high-density images. Until the advent of electronic page make-up, diffusion transfer materials were extensively employed in pre-press work.

Before the introduction of electrophotography, the silver diffusion transfer process was the most important copying technique (1950–1960). The *black and white silver diffusion transfer system* (*Copyrapid*) was introduced in the 1950's. The process was independently invented by ROTT at Gevaert and WEYDE at Agfa [197] in 1938 (⟶ Photography – History). In principle, two

coated papers are employed which in one processing step provides a negative and a positive image within ca. 1 min (see Section Black and White Copying Materials Based on Silver Diffusion Transfer Technology) by diffusion of the unexposed silver in complexed form from the negative photo-sensitive paper to the positive paper.

Compared to classical negative processing, in which printing the positive is followed by a second exposure and processing, this meant a great improvement in speed and convenience.

The silver halide diffusion transfer process has recently gained widespread use as the light sensitive media for Computer-to plate (CtP) lithographic printing plates (see Section Color-Copying Materials Based on Dye Diffusion Chemistry).

Related to black and white copying systems is the *color diffusion transfer process*. This was first designed by Polaroid for instant imaging and then developed for professional purposes by Agfa-Gevaert (Copycolor) [198]. Dyes generated during color development of the negative migrate to a positive receiving layer (see Section Color-Copying Materials Based on Dye Diffusion Chemistry).

2.2.3.1. *Black and White Copying Materials Based on Silver Diffusion Transfer Technology*

Diffusion Transfer. The diffusion transfer process consists of a sequence of physical and chemical processes coupled by diffusion. The process starts with an *exposure step*. In a camera, a light pattern is exposed, which produces a latent image in the emulsion layer, as shown in Figure 45.

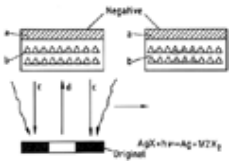

Figure 45. Diffusion transfer mechanism

a) Support; b) Negative; c) Emulsion layer; d) Light absorption; e) Light reflection

⌓ = AgX grain; ⌓ with dot = AgX grain with latent image

Then the exposed negative (b) and a receiving material (a) are passed through the developing bath while being pressed together (c), as shown in Figure 46 [201].

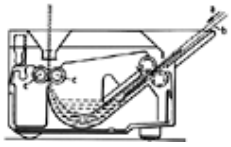

Figure 46. Developing bath

The exposed negative and the receiving material are passed through the developing bath while being pressed together

a) Receiving material; b) Exposed material; c) Squeegee rollers

Before both materials reach the squeegee rollers (c), the emulsion is fully developed according to the scheme in Figure 47.

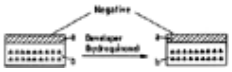

Figure 47. Schematic of diffusion transfer development. Before the exposed negative and the receiving material reach the squeegee rollers the emulsion is fully developed

a) Support; b) Emulsion layer ⌓ = AgX grain; ⌓ with dot = AgX grain with latent image; ▲ = Fully developed AgX grain

Silver ions in the silver halide grains which bear a latent image are reduced by reaction with the developer (hydroquinone).

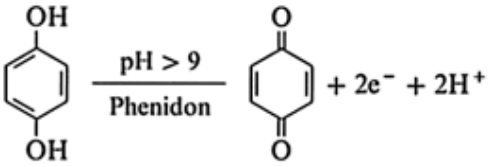

$$AgCl + e^- \xrightarrow{h\nu} Ag^\circ + Cl^-$$

Parallel to this reaction *silver halide* is dissolved by an aqueous solution of the complexing agent (sodium thiosulfate) [7772-98-7], with the formation of water-soluble silver (I) complexes.

$$AgCl + 2\,S_2O_3^{2-} \longrightarrow Ag(S_2O_3)_2^{3-} + Cl^-$$

From the squeegee rollers, these complexes are transferred by diffusion to the receiving layer where a positive image forms on the nuclei by physical development (Fig. 48). The nuclei in the receiving layer are (heavy) metal sulfide clusters, which act as a nucleating catalyst for silver deposition. During this physical development, the silver complexes deposit on the nuclei where silver is reduced. The electrons are supplied by a superadditive pair of developers (usually hydroquinone [123-31-9], and methyl-Phenidone) (⟶ Photography).

Figure 48. Schematic of a positive image formed on the nuclei by physical development

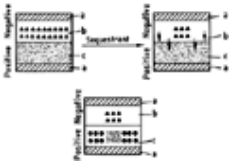

a) Support; b) Emulsion layer; c) Negative; d) Positive; e) Nuclei-containing layer

⌂ = AgX grain; ▲ = Fully developed AgX grain; •••• = Ag in the nuclei-containing layer; Dotted area = Nuclei

After the transfer phase, the positive is separated from the negative, rinsed (optional), and dried [202].

System Elements. A diffusion transfer system consists of a *light-sensitive material* (negative), a *receiving material* (positive), and a *processing solution* [197].

Most of the negatives are composed of two, three, or more layers. The top layer shields the emulsion layer from the environment and is called the antistress layer. Transport properties and contact rheology in this polymer layer are controlled by the addition of wetting agents, adhesion control agents, and matting agents.

The emulsion layer below the antistress layer consists of AgCl, AgClBr, or AgClBrI microcrystals between 0.1 and 1.0 µm in size. This emulsion is chemically and spectrally sensitized, and can be negative- as well as positive-working (direct-positive emulsion). After transfer, the former yields a positive legible image; the latter, a negative one.

Often the emulsion layer is undercoated with a third layer which is used as an antihalation layer and contains antihalation dyes. In some materials, this layer also contains developer substances so that the system becomes processible by activation with an alkaline solution (⟶ Photography).

Figure 49 shows a transmission electron micrograph (TEM) of the layer composition of a diffusion transfer negative. This layer assembly is coated on polyethylene-coated paper or a subbed polyester film. The film version can be used for obtaining a reverse-reading image by exposure through the back [203].

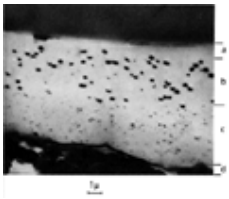

Figure 49. TEM micrograph of the layer composition of a negative

a) Antistress layer; b) Emulsion layer; c) Anti-halation layer; d) Polyethylene – paper

Receiving Material (Positive). Receiving materials are also made of several layers and are applied to polyethylene-coated paper or polyester. The *protective layer* (antistress layer) functions as in the light-sensitive material and, at the same time, avoids the formation of superficial metal deposits (bronzing).

Under this layer, the *nuclei-containing* layer is located. The nuclei catalyze the physical development. Mixtures of metal sulfides are generally used as nuclei, but single sulfides or colloidal metal sols are also possible (Ag, Au, PdS, Ag_2S, etc.). The radius of these colloidal particles ranges from 3 to 7 nm. The nuclei are dispersed in a polymer phase. In addition to gelatin, cellulose derivatives, poly(vinyl alcohol) [9002-89-5], polyacrylamide, and colloidal silica [7631-86-9] are used. Swelling of these polymer layers controls the diffusion of reactants through the gel and is decisive for transfer rate and sensitometry. A hardening system regulates the final swelling.

A third layer is sometimes located under the nuclei-containing layer [204]; dyes, toning agents, stabilizers, developing agent systems, and silver ligands can be incorporated into it. In Figure 50, the composition of a two-layer positive after silver transfer is illustrated by means of a TEM micrograph.

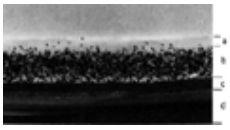

Figure 50. TEM micrograph of positive layer composition after silver transfer

a) Antistress layer; b) Nuclei layer; c) Adhesion layer; d) Polyethylene terephthalate

In the nuclei-containing layer, the image is formed on the nuclei and results in a finely divided silver deposit, whose particles are a maximum of 20 nm in size. This is considerably smaller than that of the chemically developed AgX grains of the emulsion layer (300 – 600 nm).

Processing Solution. Processing of diffusion transfer systems occurs in developers as well as in activators, depending on whether or not developing substances have been incorporated into the negative material.

Processing occurs at pH values >10. Therefore, inorganic bases (NaOH, Na_3PO_4, Na_2CO_3, etc.), organic bases (secondary and tertiary alkanolamines), and mixtures of both are used.

The diffusion transfer step is made possible by the formation of a soluble and reducible silver complex. Thiosulfates and

thiocyanates are used as inorganic silver complexing agents. They are incorporated in the form of their ammonium, potassium, and sodium salts. For special purposes, organic silver ligand systems are also used (hydantoin and pyrimidine derivatives).

These chemicals are protected from air oxidation by the addition of sodium sulfite. To control the structure and transfer speed on the silver deposition process, inorganic potassium bromide and organic (mercaptotetrazoles, benzimidazoles, etc.) inhibitors that inhibit the physical development process are used. Viscosity and surface tension of the bath are regulated with polymeric thickening agents (cellulose derivatives) and low molecular mass surfactants.

In addition, a superadditive system of developing substances consisting of a hydroquinone derivative and an alkyl- or dialkylpyrazolidinone is added to the developing liquid [205].

2.2.3.2. *Modern Lithographic Printing Applications*

The *Copyproof* black-and-white repro products from Agfa remain the most important diffusion transfer copy materials. The *Copyproof* system consists of negatives and a range of positive receiving films and papers.

However, this application of diffusion transfer technology is under pressure from electronic and digital reproduction methods. Of growing importance today is the use of the silver diffusion transfer process to make lithographic printing plates. Silver, when deposited by a diffusion transfer process, is in the form of densely packed clusters of metallic silver, unlike the filamentary morphology that is formed with usual photographic development. In this densely packed form, silver is capable of accepting lithographic ink. Thus, if an image of silver is formed on a lithographic surface by the diffusion transfer process and ink and water are applied, the silver accepts the ink whilst the water lies in the non-image background. On press, the ink on the silver image is transferred via an impression cylinder to the paper.

In so-called **analogue litho plate-making**, a photographic film is generated which carries the image to be printed. This image is placed as a mask in contact with a printing plate and flood exposed to UV light. In such a process, the printing plate is a low light sensitive resin-based material. Silver halides on the other hand are highly sensitive to a wide range of visible wavelengths of light. Consequently, the diffusion transfer process has been incorporated into lithoplates that can be imaged directly without the need for the film intermediate. For many years, plates based on the diffusion transfer process have been used for direct exposure in a camera to a paste-up copy or original. In the Agfa *Copyrapid Plus* system (Fig. 51 A) a paper donor sheet (i.e., the light sensitive negative) is exposed to the original or paste-up on a standard process camera and then processed (see Fig. 46) in contact with a grained and anodised aluminium receiving sheet (i.e., the positive). After further treatment to improve the oleophilicity of the silver image, the plate is loaded onto the press.

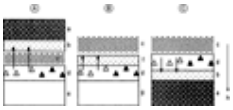

Figure 51. Simplified schematics of lithographic printing plates based on the diffusion transfer process.

a is the two-sheet *Copyrapid* transfer system, where a) is aluminium, b) is a grained and anodised alumina film containing the nucleation particles, c) is the diffusion transfer developer, d) is the exposed silver halide emulsion coating, e) is the paper support.

b is the mono-sheet *Setprint*/*Supermaster* or *Digiplate*/*Silver Master* system, where f) is the lithographic layer containing nucleating particles, and g) is the paper or PET support.

c is the mono-sheet *Silverlith*, *Lithostar*, or *Digiplate* aluminium system.

All systems are positive-working lithographic plates. h) Developer diffusion into emulsion and nucleation layers.

Copyrapid Plus is a *two-sheet* system. A printing plate, which combines the donor and receiver functions, is classed as a *mono-sheet system* (Figs. 51 B, 52). Mitsubishi Paper Mills and Agfa manufacture flexible plates in large volumes in which the substrate can be either PET or resin-coated paper (*Silver Digiplate* and *Silver Master* from Mitsubishi Paper Mills, and *Setprint Plus* and *Supermaster* from Agfa respectively for imagesetter and camera exposures). Mono-sheet products with a transparent PET base can be exposed on a standard process camera through the base from a right-reading original. For paper based materials, a camera fitted with a reversing mirror is required to give a right-reading image on exposure.

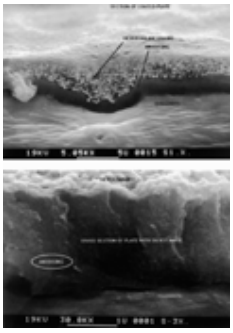

Figure 52. Cross-section of a *Silverlith* plate before and after processing.

Both the two-sheet and mono-sheet plate systems described above are better suited to small offset, single or 2-color work. With the advent of digital page make-up and the development of stable lasers, computer-to-plate (CtP) became possible in the late 1980's. CtP lithoplates capable of commercial and newspaper 4-color work became the major theme of the 1990's pre-press market. *Silverlith* from the lithoplate manufacturer Howson-Algraphy was the first CtP fully grained and anodised aluminium printing plate (Fig. 51 C, 52). It used the diffusion transfer process and combined the mono-sheet idea with a true lithographic substrate. Thus, the grained and anodised substrate is coated with a colloidal silver nucleating agent followed by a chlorobromide/gelatin photo-emulsion. After exposure, the plate is developed in a diffusion transfer developer where the

exposed areas are rapidly converted into filamentary silver that stays in the gelatin medium. The unexposed areas dissolve at a slower rate to form soluble silver complexes that diffuse to the nuclei lying on the anodic film where they are reduced to metallic silver. Because of the rough grain structure of the substrate, the silver becomes mechanically attached. When the plate is washed with water after development, a positive image of silver remains adhered to the anodised substrate. Only about 0.8 gm/m^2 of silver is deposited in the image areas at a thickness of around 10 to 50 nm but despite this print runs of many 100,000's of impressions are possible.

Continued developments in this field at DuPont-Howson and at Agfa-Gevaert have led to the present range of *Lithostar Plus (LAP)* aluminum plates from Agfa. These printing plates can now be exposed on platesetters using 488 nm argon-ion lasers (*LAP-B*), 532 nm frequency-doubled YAG lasers (*LAP-O*), 633 nm HeNe and 670 nm diode lasers (*LAP-R*) and, in 2000/01, 405 nm solid state lasers (*LAP-V*). Mitsubishi Paper Mills have also entered the market with a similar family of aluminium diffusion transfer plates (*Silver Digiplate SDP-* α*A* and *SDP-* α*R* for argon-ion and HeNe or red diode lasers respectively).

No other CtP consumable has the breadth of application as these plates. The diffusion transfer process finds itself, once more, at the center of a revolution in applied imaging technology.

2.2.3.3. *Color-Copying Materials Based on Dye Diffusion Chemistry*

Although the *Ektaflex* and *Agfachrome* Speed systems described below are no longer available they are included for their technical interest.

In the first half of the 1980s, color-copying materials based on dye diffusion chemistry came on the market; until then, these had been used only in the manufacture of instant color materials for amateur cameras (Kodak PR 10, Polaroid SX 70, and Polacolor).

Kodak Ektaflex System. Kodak elaborated the Ektaflex system with the PR 10 chemistry (⟶ Photography). After some years, this was withdrawn from the market because of patent litigation with Polaroid. In this system, a two-sheet material was used, consisting of a light-sensitive sheet and a receptor sheet on which the color image formed by diffusion transfer. Positive-working and negative-working versions were available.

In the *negative-working* version, common silver halide emulsions were used. The *positive* character of the second material was obtained with special direct-positive emulsions such as those in the PR 10 instant material.

For both types, the same receptor sheet was used. A special processing apparatus was developed that moistened only the light-sensitive sheet for 20 s. Then the film was laminated onto the receiving paper by a pair of pressure rollers. The sandwich could be separated after a contact time of 8 – 15 min, depending on the temperature and type of paper. Rinsing was not necessary because the light-sensitive sheet contained a neutralization layer and a timing layer, which lowered the pH to neutral (a timing layer is a binder layer which delays the penetration of OH^- ions to the neutralization layer). The copy was dry after a few minutes. An overhead film for making transparencies was also available.

Agfa – Gevaert Color Diffusion System. In Agfa-Gevaert diffusion materials, a completely new chemical concept is used. This is a system of dye release, which enables the replacement of direct positive emulsions with grains containing fog centers by negative emulsions that are simpler to prepare. In this way, direct-positive color images are obtained.

The principle is deduced from a reaction described in the chemical literature: 4-hydroxybenzylphenylsulfones are cleaved in alkaline medium to form a sulfinate and a quinone methide, although other sulfones are highly alkali-stable [207].

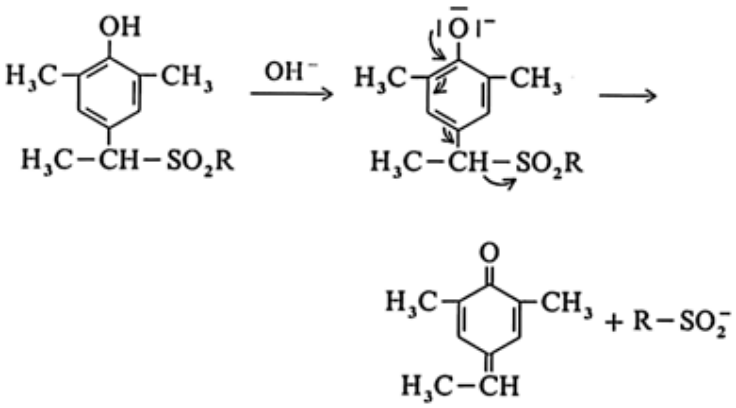

The reaction can be redox-controlled by introducing an additional hydroxyl group in the phenol nucleus. In this way, *o*- and *p*-hydroquinone derivatives are obtained, which convert to quinones on oxidation. These quinones become more alkali-stable with increasing number of substituents attached to the ring. Azo dye sulfinic acids are used as sulfinate components.

The following molecule has its bulky substituent on the hydroquinone residue:

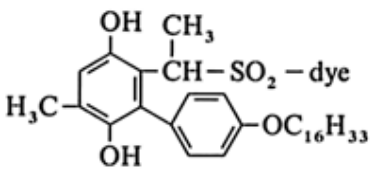

This hydroquinone derivative can basically be oxidized by exposed silver halide in the presence of an auxiliary developing substance (electron transfer agent, e.g., Phenidone) to a stable *p*-quinone that no longer releases its dye moiety [208].

For practical purposes, however, dye releasing appears to occur too quickly being only partly controllable by exposed silver halide. Therefore, the chemical system is adapted further as follows [209].

Starting materials are two compounds, the alkali-stable quinone and an electron donor. At *unexposed areas*, the electron donor is capable of reducing the quinone to the corresponding hydroquinone, which releases the dye – sulphinic acid moiety. At *exposed image areas*, the electron donor is oxidized by exposed silver halide via the electron transfer agent so that the stable quinone continues to exist there. The following compound is used as the electron donor:

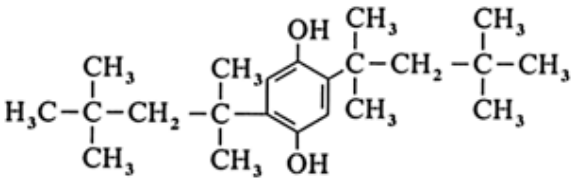

Stable nondiffusing quinone derivatives that supply the dye – sulfinic acid moiety in the material include the following:

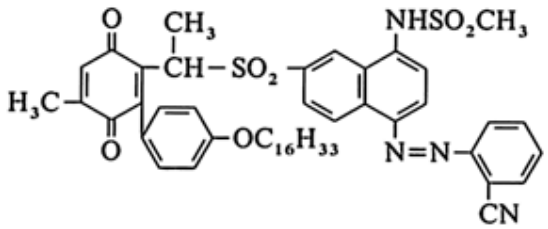

magenta

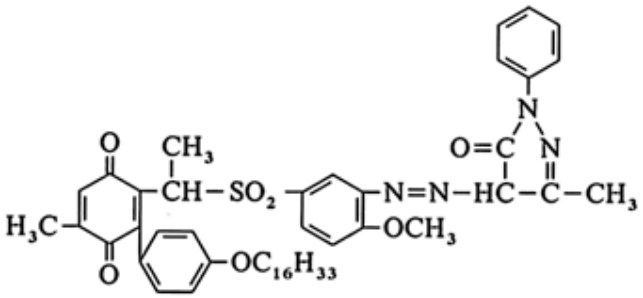

yellow

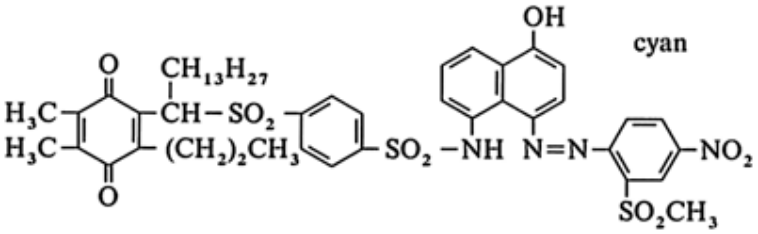

cyan

At the exposed image areas, two competing reactions take place:

1. oxidation of the electron donor by exposed silver halide via the electron transfer agent, and
2. oxidation of the electron donor by dye – quinone after which the dye – sulfinic acid moiety is released. This reaction is undesirable and must be prevented. For this reason, the most rapidly developing silver halide emulsions (mainly AgCl) are used.

The released, diffusing dyes are very pure, spectral-grade azo dyes without developer residue, so that very brilliant color images can be obtained by this mechanism.

Without further sophistication the absorption of the incorporated dyes would interfere with the exposure of the silver halide grains, resulting in a loss of sensitivity. Because of the pH-dependence of the absorption spectra this can be prevented by keeping the pH of the material sufficiently low. Only during and after the alkali activation are the absorption spectra shifted to a longer wavelength and the correct color is obtained. The dye-releasing compounds can be incorporated directly into the silver halide emulsion layers without any loss of sensitivity. Thus, the material can be structured in a considerably simpler way.

On the basis of this chemical principle, Agfa – Gevaert put two photographic materials on the market in the early 1980s: a two-sheet version (Copycolor) and a mono-sheet version (Agfachrome Speed).

The *Copycolor material* consists of a light-sensitive and a receiving sheet to which the dyes diffuse during processing and on which they are immobilized by the mordanting agent. The composition of both sheets, as well as the operation, are represented schematically in Figure 53.

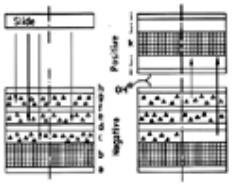

Figure 53. Schematic of a light sensitive (A) and a receiving sheet (B)

a) Backing layer; b) Polyester support; c) Red-sensitive layer with cyan-releasing compound; d) Intermediate layer with sodium light-absorbing filter; e) Green-sensitive layer with magenta-releasing compound; f) Intermediate layer with yellow-filter; g) Blue-sensitive layer with yellow-releasing compound; h) Antistress layer; i) Antistress layer (receiving sheet); j) Mordanting layer; k) Support

The receiving sheets are coated symmetrically, which allows diffusion images to be generated on both sides. Different versions exist: mat paper, glossy paper, and transparent film for the production of transparencies for overhead projection.

The light-sensitive sheet has a special absorption filter for sodium vapor light in the intermediate part of the red-sensitive layer, which allows work in the darkroom under sodium vapor lamps.

Gradation of the material is rather steep (g = ca. 2.6) and can be regulated as desired by use of the contrast-reducing screens.

Thus, a universal copying material is available for copying various originals, such as normal color photographs, color prints, colored line originals and graphs, aquarelles, transparencies, and graphic selections. It can also be used for making transparencies for overhead projection.

Light sources may be color enlargers, repro cameras, contact exposure units, etc.

The system does not need any special processors as in the case of Kodak Ektaflex; processing can take place in diffusion transfer processors according to the Eisbein principle (see Fig. 46). The processing liquid is a highly alkaline activator.

On the basis of the same chemical principle, Agfa-Gevaert also developed a mono-sheet material: Agfachrome Speed [210]. The light-sensitive silver halide layers are covered with a white titanium oxide layer coated with the mordanting layer.

Exposure is through the support layer. The structure of the material is represented schematically in Figure 54.

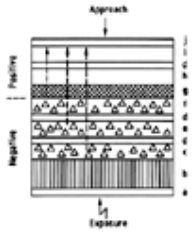

Figure 54. Schematic of a monosheet material (Agfachrome speed)

a) Backing layer; b) Transparent support; c) Blue-sensitive layer with yellow-releasing compound; d) Intermediate layer; e) Green-sensitive layer with magenta-releasing compound; f) Red-sensitive layer with cyan-releasing compound; g) Carbon black layer; h) TiO_2 layer; i) Mordanting layer; j) Antistress layer

2.3. Microfilms and Microfiches

2.3.1. *Diazotyping*

Although silver halide film is preferred for microfilm archives, working copies (“active microfilm”) are often made with diazo film, especially when several copies are required [211].

The diazotyping process is simple and economical, and the diazo material is especially suitable because it is grain-free and, consequently, gives very high resolution.

In the 1980s, polyester film has been used almost exclusively as the base material. The film is chemically pretreated to increase adhesiveness and then coated with a laquer that usually contains the chemicals for diazotyping. In general, the binders employed are cellulose esters, and the diazo compounds used are, in principle, the same as those described in Section Diazotyping. The diazonium tetrafluoroborates and diazonium sulfosalicylates are preferred to zinc chloride double salts because of their greater solubility in organic solvents and their higher thermal stability.

Diazo duplicating films, which give blue-lined copies, are employed only as reading film; black-lined duplicates can be used to make further copies (generations). Diazo films are suited primarily for the duplication of writing because of their steep gradation. Other types of film with specially adjusted, flatter gradation are employed for continuous tone pictures. Diazo duplicating films are available in the same forms as photographic films (i.e., as roll film, flat film (microfiche), and aperture cards). Processing of these films is done exclusively with the dry (ammonia) method (see Section Processing). Duplicating machines with different classes of efficiency are available for roll films, microfiche, and aperture cards. High-performance duplicators produce up to 2000 duplicates of size DIN A 6 per hour.

2.3.2. *Vesicular Film*

Like diazotyping, the vesicular method is also based on the photolysis of diazonium salts by UV light. In practice, the same diazonium salts are employed in this procedure. They are applied in a thin layer of special thermoplastic material to a transparent base material which is usually a polyester film.

During development, the picture is produced not with the aid of the diazo compound present on the areas covered by the picture, but with the aid of nitrogen formed by photolysis on the nonpicture areas. Heating for a short time (≤1 s) to 100 –

150 °C softens the thermoplastic material; the nitrogen, which is at first dissolved, expands and microscopically small bubbles (0.5 – 5 µm) are formed. Light is scattered at the interfaces between bubble and polymer, and these areas appear dark when examined in transmitted light. This is normally a negative process and thus represents a welcome addition to the positive diazotype duplicating films.

During fixation, the diazonium salt still present is destroyed by overall exposure to light. Without subsequent heating, the nitrogen produced escapes slowly from the layer with no bubble formation.

The light sensitivity of this system is somewhat higher than diazotyping because the diazo compound need not be entirely decomposed for picture formation, and development exerts a certain intensification effect. Both methods are used only for contact exposure. Gradation with the vesicular film process is even steeper than with diazotyping.

Although this method was developed a long time ago [212], it was not used commercially until the 1950s [213]. The duplication of computer output microfilm (COM) represents the most important application of this technique. At present, all important producers of diazo duplicating film also supply vesicular film. Duplicating machines can often be used for the processing of either diazo duplicating film or vesicular film.

Worldwide consumption of duplicating film (diazo and vesicular film) was estimated at 60×10^6 m^2 for 1996. Some producers (and trade names) of diazo and vesicular film are:

Rexam, Runcorn, Cheshire, United Kingdom.
Messerli AG, Glattbrugg/Zürich, Switzerland.
Rexam, South Hadley, Massachusetts.
Somar Manufacturing Company, Tokyo, Japan (Somic-D).
SKC, Sunnyvale, California.

[Top of Page]

3. Imaging in Graphic Arts

Johan Verelst, Werner Frass, Thomas Telser, Horst Hoffmann, Bernd Bronstert, Karl-August Springstein, Rod Potts

3.1. Graphic Arts Photography

3.1.1. *Classes of Photomaterials Used in Prepress Production*

Graphic photomaterials are characterized by their *gradation*, maximum *density, energy requirement* (i.e., the inverse of their photographic speed), *spectral sensitivity* (see ⟶ Photography), and recommended processing.

Phototypesetting Paper [217]. Most phototypesetting papers consist of a waterproof, resin-coated paper base covered with a photographic emulsion having a gradation of 2.5 – 5, a maximum density of > 1.5, and requiring an exposure of 0.1 – 10 mJ/m^2, depending on the type of light source in the phototypesetter. This also dictates the spectral sensitization: cathode-ray tube (CRT) recorders necessitate orthochromatic materials; helium – neon lasers require peak sensitivity at 633 nm; and the newest diode – laser devices need spectral sensitization to peak at 780 – 820 nm. In most phototypesetters, the exposure times of a single image element range from 100 to 500 ns; for these very short exposures, the photoemulsion should be optimized.

Phototypesetting Film [217]. Phototypesetting films are manufactured on a transparent film base consisting of cellulose triacetate or poly(ethylene terephthalate) (PETP). The emulsion properties are similar to those of phototypesetting paper. However, the silver halide content is higher, allowing for a maximum density > 4 and a gradient in the range of 3 – 6. Phototypesetting materials are usually processed in fast Phenidone – hydroquinone developers (rapid access developers).

Scanner Recording Films [218]. Recording films for color-separation scanners and for output recorders of computer-based color image assembly systems are always manufactured on a stable PETP base. The emulsion coatings are quite similar to those of phototypesetting films; some have even higher gradient and higher maxi-mum density. The spectral sensitivity can have a peak at 488 nm for argon ion lasers or at 633 nm for helium – neon lasers; both are commonly used light sources in these applications. Scanner recording films are usually processed in *rapid access developers*; when extreme edge acuteness of the recorded halftone image is required, *lith developers* are used.

Lith and Line Films [219]. *Lith films* produce very high gradients (> 10) and very high maximum density (> 5) when processed in *lith developers*. Lith developers contain only hydroquinone as the developing substance and almost no sulfite. The most common lith developers are orthochromatically sensitized and have their speed adapted for use in reprographic cameras. Lith films are typically used for conversion of continuous tone pictures into halftones.

Line films are quite similar in density and spectral sensitivity, but they are usually processed in rapid access developers. The gradients obtained are typically in the range 5 – 10. Their major application is the reproduction of line art.

Silver Halide Diffusion Transfer Materials [220]. Making a direct positive reproduction in one step is possible with silver halide diffusion transfer materials (e.g., Copyproof of Agfa; see Section Silver Process). A silver halide-containing donor sheet is exposed imagewise. By means of a developer containing a silver complexant, unexposed silver is transferred to a receptor sheet consisting of paper or film. This process yields particularly high gradients in the transferred image, which allows for excellent halftone images and very clean line art reproductions.

Contact Film [221]. Contact films are low-speed, nonorthochromatic line films. The most popular can be handled in bright room light without fogging; the exposure is done with intense UV light, and processing relies on Rapid Access Developers.

3.1.2. *Flow Scheme of Graphic Prepress Process.*

In general, a publisher's job file contains four kinds of information: (1) text copy; (2) color pictures (most often continuous tone color transparencies); (3) miscellaneous graphic elements and monochrome pictures; and (4) layout, which describes how all the information is to be arranged on the pages.

In most printing processes (with the exception of gravure printing), ink transfer onto the paper is a bilevel mechanism [222], [223] (i.e., an image element is either printed at full ink strength or with no ink at all). To incorporate continuous tone pictures on the printed page, they must first be converted into a screened, bilevel halftone format [224].

Color-Separation Images. In color printing, pictures are reproduced by making a color-separation image for each primary printing color (i.e., yellow, magenta, cyan, and the supporting black). The desired color print is obtained by printing the four color-separation images on top of each other, by using the corresponding printing inks [214], [225].

Photo-Optical Color Separation Process. Color separations can be made by using a reprographic camera or an enlarger, and suitable photographic films. The process of making color separations typically involves successive exposure of panchromatic continuous tone film to the orginal picture through blue, green, red, and amber filters. The resulting continuous tone negatives contain the yellow (Y), magenta (M), cyan (C), and black (K = key) information of the original, respectively. The separation negatives are then copied onto orthochromatic lith film through a contact screen, to produce four halftone separation positives that can be used to produce the printing plates.

Color Separation with Electronic Separation Scanners. Since the early 1970s the photo-optical process of making color separations has largely been replaced by electronic separation scanners (see below).

Prepress Work Flow for Producing Color Pages (Fig. 55). *Layout data* pass through the central branch and dictate how all informational elements should be positioned on the page (in many cases, the trade shop which makes the color separations must verify and clean up the layout for inaccuracies).

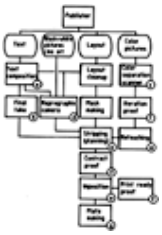

Figure 55. Typical prepress workflow with stand-alone color separation scanner

Photomaterials: a) Typesetting paper; b) Typesetting film; c) Scanner film; d) Lith and line; e) Contact film; f) Color proofing; g) Printing plates

In the left branch, *text* is composed according to the layout and set on phototypesetting paper for proofreading and correction [226]. Correction proofs are often produced on plain paper by electronic page printers [227], [228].

After approval, the typesetter produces a final take, again in phototypesetting paper. Then a film negative is shot in a camera with lith or line film. This negative is mounted onto the final page by the stripping department. For high-quality publications, the final take is often made on phototypesetting film. (In many cases, text composition is done in a separate typesetting trade shop.)

Bringing line art and monochrome continuous tone pictures to the final size, and converting the latter to halftones, are most frequently done with a reprographic camera by using film or paper having very high contrast (i.e., line or lith materials as well as silver halide diffusion transfer materials) [229].

On the right of the flow chart, color pictures are analyzed by the color-separation scanner. In real time, image data are converted into Y,M,C,K data, which take into account the spectral characteristics of the printing inks [214], [230]. Moreover, a scaling transformation is executed according to the dimensional instructions of the layout. Four color-separation halftone films are obtained, one for each of the subtractive primary printing colors (yellow, magenta, and cyan) and for the supporting black. During the conversion of continuous tone data into raster halftones, the nonlinear transfer characteristics of the printing are compensated. Color reproduction can be corrected by retouching, often under guidance of a color proof [231]. Both wet (*dot etching*) and dry retouching techniques are currently used in most shops [232], [233], with the latter being preferred. After approval, color separations of each picture go to the stripping department where they are assembled into pages.

The *stripping department* is the nodal point at which all page elements come together in their final form. After manual assembly, a clean copy should be made on contact film. Then, a color proof can be made of the page for the client's approval.

A number of individual page films are usually mounted together to accommodate the size of the printing plates. The imposition of the pages assures the proper page order after folding and binding.

Computer-Based Color Image Assembly Systems. Manual stripping and dot etching are the most labor-intensive nodes in the work flow. Investments in image-processing technology are made in these areas to reduce production costs. The functionality of modern computer-based color image assembly systems (Fig. 56) goes far beyond the automation of stripping images onto a page. All pictures for a page are scanned, digitized, and stored on hard disks. Image manipulation and page assembly are performed at an interactive color work station [234], [235]. The color display allows the operator to decide whether interacting with the system will produce the desired effect on printing. (This soft proof at the work station can also be transmitted to a distant client for approval; remote proofing and the hard copy proof by means of digital proof recorders are still in their pioneering stage.)

Figure 56. Typical prepress workflow with color image assembly system

Photomaterials: a) Typesetting paper; b) Typesetting film; c) Scanner film; d) Lith and line; e) Contact film; f) Color proofing; g) Printing plates

In most color system trade shops, the assembled images are recorded on scan film. Next, the text parts are combined with the images by making copies on contact film. Color proofs of these finished pages are usually submitted to the client for approval. Whenever the client requires any color changes, either a new run must be made through the system, or the work already done can be saved in part by retouching. In many cases, the latter proves to be more economical. This explains the interest of most shops in computer-assisted retouching systems.

3.2. Basics of the Use of Light-Sensitive Nonsilver Materials in the Graphic Arts

To reproduce information stored in the photographic original in a printing process, this information must be transferred to a printing plate, which is generally coated with a light-sensitive material. This process is known as platemaking (in a narrower sense “plate” also means the unexposed, light-sensitive plate. In German it is differentiated by using the term “Druckformherstellung” for platemaking).

However, prior to platemaking the appearance of the final printed result must be determined. This is especially important in four-color printing because (1) the human eye is extremely sensitive to color deviations; (2) many color originals are intended to be reproduced in a slightly different shade of color and it is of interest to determine in advance whether the color variation is correct; and (3) the printed result has economic consequences (e.g., in the case of a mail-order catalogue).

Plates can be prepared to obtain a preprint on a small printing press so that the printed result can be viewed in advance. Preprints of this type are called *press proofs*. Although this process is still found in several regions around the world, it is less common these days. It is very useful when several copies of the four-color prints are needed, the process is very time consuming and it has now largely been superseded by digital proofing techniques.

An alternative way to press proofs to obtain an advance impression of the printed result is to use photochemical methods similar to those employed in platemaking to prepare four-color copies of the photographic original. This technique is sometimes *called color-proofprinting, color proofing*, or *off-press proofing* and can be generated either by analogue or digital means. Because the application and chemical fundamentals of both color-proof materials and light-sensitive printing plates are to a great extent identical, these principles are described in a separate section.

3.2.1. *Application*

The recording materials carry on a substrate (e.g., metal, paper, or plastic film) a more or less thin layer (0.1 – 7000 µm) of light-sensitive coating. These light-sensitive coatings undergo photochemical or photophysical change [236] on exposure to light from a suitable source. Usually, the light reproduces an image of the photographic original with its transparent and covered picture elements on the light-sensitive recording material. The light sources used are xenon lamps and metal-halide doped mercury lamps that emit in the spectral range 350 – 450 nm. Modulated lasers, (e.g., of 488 nm wavelength) are used for line-by-line scanning exposure. A more recent development is that of thermally sensitive plates which are also exposed by modulated lasers typically at 830 nm.

The processing of light-sensitive recording materials must occur in the total absence of daylight. Yellow light (light without the blue component) is generally used for lighting the workplace. Thermal plates on the other hand may be safely handled under daylight conditions.

After imagewise exposure of the recording material, the unexposed areas must be protected against the influence of light, which would result in the destruction of recorded information. This process is generally known as development. It can be achieved by chemical reaction of the exposed or unexposed areas of the coating, by dissolving the exposed or unexposed areas from the carrier (*subtractive* development), or by selective deposition of a developing substance on the exposed or unexposed areas of the coating (*additive* development). The photomechanical method is presented in Figure 57 as an example of subtractive development.

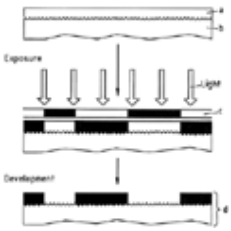

Figure 57. Photochemical platemaking

a) Light-sensitive layer, negative working; b) Carrier material; c) Film original; d) Printing plate

The process is defined as either negative or positive working, depending on whether the transparent or the opaque areas of the photographic original appear as the image in print. These terms cannot be related to particular chemical processes but depend equally on the composition of the lightsensitive recording material and the read-out process (i.e., type of

development and intended use).

As far as chemistry is concerned, two basic types of systems exist: coatings that become soluble after exposure to light (*photosolubilising coatings*) and those that become insoluble after exposure (*photo cross-linking, photopolymerising* or *photoinsolubilising coatings*). A photo cross-linking formulation can be used to produce a negative lithographic plate or a positive screen-printing stencil because the mechanism of ink transfer in lithography is different from that in screen printing. The third system applied in information recording in graphic arts is a photophysical one: *photoconductors* can be used as active agents that are several orders of magnitude more sensitive than photochemical materials.

3.2.2. *Chemical Fundamentals*

3.2.2.1. *Photosolubilising Systems*

Derivatives of 1,2-Naphthoquinone Diazide (2-Diazo-1-oxo-1,2-dihydronaphthalenes). In most cases in which photosolubilising systems are used, a derivative of 1,2-naphthoquinone diazide is the active substance. Photolysis of this class of substances was elucidated by SÜs in 1944 [237]. The mechanism corresponds to the Wolff rearrangement of a 1,2-diazoketone. The first step is the light-induced elimination of nitrogen from **1** with the intermediate formation of a carbene (**2**), which in turn rearranges to give a ketene (**3**). Layers of this type usually contain traces of water, which are sufficient to convert this ketene to the indenecarboxylic acid (**4**). Substance **4** probably rearranges to give the isomeric acid (**5**).

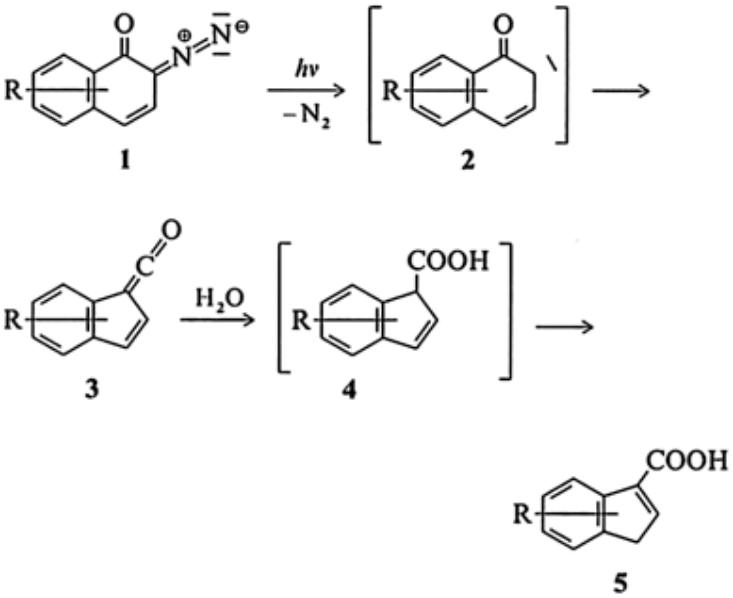

The exposed layer is brought into contact with an aqueous alkaline salt solution (typically Na_2SiO_3) for development. In this process, the indenecarboxylic acid formed in the *exposed areas* dissolves; in the *unexposed regions*, the naphthoquinone diazide remains unchanged on the carrier.

For practical reasons, the 4- or 5-sulfonic or carboxylic acid esters of naphthoquinone diazides with phenol derivatives are usually employed instead of the parent substances. These esters are synthesised by reaction of the sulphonyl chlorides or carboxylic acid chlorides of naphthoquinone diazide with the corresponding mono- or polyfunctional phenols (e.g., 4-cumylphenol [238] or the condensation product of pyrogallol and acetone [239].

Naphthoquinone diazide derivatives are not the only substances present in light-sensitive layers because, despite their relatively high molecular mass, they are not good film formers. Instead, they are usually employed in combination with a polymeric binder having good film-forming properties. Novolaks have proved especially useful. The solubility of the light-sensitive layer is greatly reduced by interactions between the naphthoquinone diazide and the novolak [240]. These interactions cease after exposure of the naphthoquinone diazide. In this way, the difference in solubility between exposed and unexposed areas is greater than it would be if the naphthoquinone diazide were used alone.

Other Systems. Two related systems are described briefly here because they have attained a certain practical status.

Photolysis of o-Nitrophenylcarbinol Esters. The carboxylic acid esters of 2-nitrobenzylalcohol, for example, undergo a light-induced rearrangement and cleavage to 2-nitrosobenzaldehyde and free carboxylic acid when irradiated with shortwave UV light of 300 – 400 nm. On treatment of the exposed recording material with aqueous alkali, the free carboxylic acid and the products resulting from 2-nitrosobenzaldehyde dissolve from the exposed areas [241], [242].

Photoinitiated Hydrolysis of Acetals and Orthoesters. A whole series of substances is known which eliminates acids upon photolysis. Examples are diazonium salts of halogen complex anions [243], certain naphthoquinone diazide derivatives [244], and trichloromethyl-s-triazines [245] or -oxadiazoles [246] with chromophoric substituents. If these photolytic acid donors are combined with easily hydrolysable compounds (e.g., acetals or orthoesters), coatings are obtained that are highly sensitive to light because of the catalytic action of the acid formed [247-251].

3.2.2.2. *Photoinsolubilizing (Photocuring) Systems*

Many more photochemical systems are known that become insoluble after exposure to light. The first photochemical nonsilver materials used to record information belonged to this type of system (asphalt, gelatin-dichromate).

Several different mechanisms may be utilised in the formulation of such systems [252]. However, only a few are of commercial relevance today:

1. photoreduction of chromium(VI) compounds,
2. photolysis of aromatic diazonium salts,
3. photolysis of *p*-diazoiminoquinones,
4. photolysis of aromatic azides,
5. photodimerization of activated double bonds,
6. photoaddition of mercapto groups to double bonds, and
7. photoinitiated polymerisation.

Photoreduction of Chromium(VI) Compounds. is of historical interest only [253]. By the middle of the nineteenth century, alkaline or ammoniacal mixtures of water-soluble organic colloids, such as gelatin, isinglass, starch, gum arabic, albumin, casein [and later, poly(vinyl alcohol)], and water-soluble dichromates were known to be sensitive to light. The colloids lose their water solubility in the process and retain only a more or less pronounced swelling capacity; i.e., they are cured. The use of this process is seldom if ever encountered today because of the environmental issues associated with the use of chromium (VI) compounds.

Photolysis of Aromatic Diazonium Salts. It was found as early as 1931 that high molecular mass polyfunctional aromatic diazonium salts can be used advantageously in the formulation of light-sensitive printing plates and other photomechanically processible materials [254]. Due in part to the ease of development in aqueous based developers, it remains today the most commonly utilized process in commercial negative working plate coatings.

Diazonium salts of this type can be produced by the condensation of *p*-diazodiphenylamine derivatives with lower aldehydes in an acidic medium and precipitated as sulfonates. Activated diethers and bismethylol compounds can also be used instead of aldehydes [255].

The photochemical reactions occurring during exposure have not been elucidated fully. However, the higher molecular mass diazonium compounds can be assumed to form similar reaction products on exposure as lower molecular mass compounds, which can be investigated more easily. Nitrogen is known to be eliminated from diazonium compounds on exposure to UV light. The *aryl cation* produced can react with nucleophilic substances in the layer to form substitution products such as phenols or phenol ethers. These products have a much lower solubility in polar solvents. However, it is unlikely that this process occurs exclusively. Diazonium compounds in the binder matrix can also be photolyzed to give *radicals*, as in thermal decomposition. The resulting radicals can cause cross-linkages by hydrogen abstraction and similar reactions, which also lower the solubility of the matrix. Both mechanisms result in a product that is less soluble than the starting material. Unexposed areas are removed from the carrier during development, and the exposed portions are retained.

Photolysis of p-Diazoiminoquinones. *p*-Diazoiminoquinones (**6**), with R_1, R_2, R_3 = alkyl, undergo ultraviolet-induced nitrogen elimination to form carbenes, which can react further to

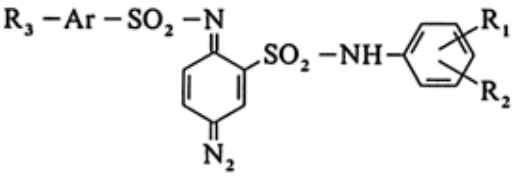

6

produce higher molecular mass substances [256]. These substances are less soluble in organic solvents and dilute alkali or acid than the parent diazo compounds. Synthesis of p-diazoiminoquinones can be achieved, for example, by starting with 1-chloro-4-nitrobenzene-2-sulfonyl chloride [257].

Photolysis of Aromatic Azides. Low molecular mass aromatic azides (e.g., 4,4′-diazidostilbene-2,2′-disulfonic acid) can also be used in the photochemical curing of organic colloids. On exposure to UV light, azides eliminate nitrogen to form highly reactive nitrenes, which can react further by hydrogen abstraction, insertion in C–H bonds, reaction with olefinic double bonds in the matrix, or dimerization. Cyclorubber (⟶) may be used as the binding agent.

Polyfunctional azides are the active substances of choice because they lead to a higher cross-linking density and can also be sensitised [258]. During development, which is often carried out with organic solvents, unreacted azides and the binding agent are dissolved from the carrier. Lithographic plates made with such materials are characterised by long run length under very harsh press conditions. However the increasing trend towards aqueous developers in recent years means that the use of such plates has diminished rapidly.

Photodimerisation of Activated Double Bonds. α, β-Unsaturated carbonyl compounds, such as chalcones (phenyl styryl ketone derivatives) and cinnamic acid derivatives are accessible to photochemical dimerisation. Eastman Kodak made a very detailed study of the photochemical behaviour of cinnamic acid esters in the 1950s [259]. The dimerisation reaction apparently starts from *trans*-cinnamic acid and proceeds via the excited triplet state of the cinnamate molecule and radical intermediates to give derivatives of α-truxillic acid (diphenyleyclobutanedicarboxylic acid). Radical side reactions complicate the course of the overall reaction. This reaction mechanism offers numerous possibilities for sensitisation with triplet sensitizers. Thus, the light-sensitive region of cinnamates, which lies below 350 nm, can be extended to > 400 nm [260].

Polymeric cinnamates, preferentially poly(vinyl cinnamate), are used widely in industry to increase the difference in solubility between exposed and unexposed portions of the recording material. Poly(vinyl cinnamate) can be produced from poly(vinyl

alcohol) and cinnamoyl chloride. Another possibility is the use of *polyesters* from *p*-phenylenebisacrylic acid and dihydric alcohols. A single cross-linking step between two polymer chains results in an approximate doubling of the molar mass and hence, a dramatic decrease in solubility. Development is usually carried out with organic solvents and again their use today is diminishing as the demand for solvent-free developer increases.

"Thiol – Ene" Addition. Photochemical addition of mercapto groups to terminal double bonds of allyl or vinyl ethers also leads to an increase in molar mass and a decrease in solubility. This effect is especially pronounced when polyfunctional mercaptans react with polyfunctional olefins because a high cross-linking density is generated in the process. Initiators play an important part in starting the reaction. They absorb light and lead to the formation of mercapto radicals, which then add to the double bond. In general, light-sensitive materials of this type are viscous liquids which are applied to the desired carrier by the user [261].

Radical Photoinitiated Polymerization [262]. Radical photoinitiated polymerisation was first suggested by GATES [263] in 1945 and then investigated in detail by Du Pont. The reaction principle is based on the photochemical generation of radicals that initiate polymerisation of a reactive monomer [264]. Only the initial reaction is dependent on light; the chain reaction represents an intensification mechanism, and thus the quantum yield of the total reaction $\Phi > 1$. Hence, a system of this type should be very sensitive to light. However, unlike the photoreaction of cinnamates in which a single reaction step causes a doubling in molar mass, each step in the chain reaction increases the molar mass only slightly. Consequently, many chain propagation steps are required before the difference in solubility between image and nonimage areas is sufficient for development.

Nonvolatile low molecular mass polymerisable compounds, as well as high molecular mass compounds, can be used as monomers. Polyfunctional compounds (e.g., esters of acrylic or methacrylic acid, such as pentaerythritol triacrylate or trimethylolpropane triacrylate) are preferred for yielding extensive three-dimensional cross-linked networks [265]. Various substances and combinations have been proposed as initiator systems, e.g., aromatic carbonyl compounds, heterocycles, azo or diazo compounds, mercaptans or disulphides, and photoredox systems [266].

A light-sensitive coating that consists only of monomer and initiator is soft, sticky and flowable. Addition of a sufficient amount of suitable, polymeric binder yields a nonslip, stable coating.

These properties can be exploited directly for imaging. A coating of this type can lose its surface stickiness after exposure. The sticky image can then be made visible by application of a colored powder [266].

Unhindered propagation of the photoinitiated chain reaction in the exposed areas can be maintained only if atmospheric oxygen is largely excluded from the reaction site. This may be achieved by using a covering layer of poly(vinyl alcohol) or a covering film (e.g., of polyester) [266].

3.2.2.3. *Photophysical Principles of Image Generation*

Electrophotography [267], [268] (see Section Electrophotography). The basic principle of electrophotography is the photoconductor effect. Photoconductors are compounds that are insulators in the dark (conductivity 10^{-13} S/m), but whose conductivity increases by a factor of ca. 10^8 after exposure to light. Lithographic plate systems based on this principle were very popular during the 1980's and early '90's, but their use has diminished significantly of late.

For imaging, the photoconductor layer is charged electrostatically in the dark under a high-voltage corona. Thereafter, it behaves like the dielectric material of a charged condenser. During exposure, pairs of charge carriers are generated in the photoconducting layer. These pairs are separated under the influence of the electric field and migrate to the two poles of the condenser, and the voltage in the exposed areas collapses. A latent charge image remains in the unexposed areas, which can be developed in different ways. Resin particles with an electrically opposite charge can be applied to the latent charge image in powder form (dry toner) or as a dispersion (liquid toner). These particles are deposited only in the charged image areas. The toner image can, in principle, be processed further in different ways (Fig. 58).

Figure 58. Further processing of a toner image generated by electrophotography

a) Charged plate; b) Exposed plate; c) Developed plate; d) Molten toner; e) Hydrophilized plate; f) Uncoated plate

The *light sensitivity* of electrophotographic recording materials is 10^3 – 10^4 times higher than that of photochemically acting materials described above [236]. As with silver halide layers, the spectral sensitivity to light of electrophotographic materials can be extended to cover a wide range by the use of sensitiser dyes. Thus, the sensitivity of the recording materials can be adjusted very specifically to various light sources (argon-ion laser 488 nm, helium – neon laser 632 nm, and tungsten lamps).

Different types of photoconductors are known: Zinc oxide is the most common *inorganic photoconductor*. It is applied as a pigment dispersion in a binder matrix. In the past, layers on paper were used in office copying; today, they are applied in small-size offset printing [269-271]. Sputtered cadmium sulphide layers are also used, e.g., as electrophotographic film on polyester film and as light-sensitive material on specially prepared steel plates for the production of color-proof prints [272].

Furthermore, a large number of *organic photoconductors* are available, e.g., nitrogen heterocycles such as poly-N-vinylcarbazole and derivatives of oxadiazole [273], oxazole, thiazole, thiadiazole, or imidazole with suitable electron donor and electron acceptor substituents.

Like zinc oxide, *organic pigments* can also be used in binder-containing dispersions for the production of photoconductive

coatings. Copper phthalocyanine is especially suited to this application. The different crystal modifications of copper phthalocyanine vary somewhat in their electrophotographic effectiveness [274].

The properties of photoconductors, i.e., the formation of pairs of charge carriers in light and the migration of these charges in an electric field, are not necessarily combined in the same chemical substance. Cases are known in which two substances or coatings are combined to achieve the photoconductor effect. The charges are generated in one of the substances, whereas the other is responsible for charge transport (e.g., perylene pigments and low molecular mass photoconductors) [275].

Photolysis of Silver Halides. The principle of silver halide photographic film is also applied in different forms to record information in the graphic arts [276].

On the one hand, the corresponding films can be used in reprophotography; on the other hand, silver halide coatings can be applied to layers of lower light sensitivity. In this case, the higher sensitivity of the silver halide coating is exploited for image production, so an image-carrying mask is obtained. Subsequent overall exposure transfers the information in the mask to the layer below, which better fulfills the requirements of printing.

Furthermore, the principle of silver salt diffusion (see Section Silver Process) can be used for the production of proofs and printing plates (see ⟶ Photography).

The high sensitivity of the silver salt diffusion transfer system means that printing plates can be produced directly from hard copy (via a camera) or, more usually nowadays, from digital copy (via a laser platesetter) without the requirement for a film intermediate. Lithographic printing plates based on the silver diffusion transfer process are described in Section Modern Lithographic Printing Applications. They hold a significant share in both Direct-to-Plate and Computer-to-Plate markets. There are two types of assembly, two-sheet systems where the negative light-sensitive donor is separate from the positive receiver sheet, and single-sheet systems where the donor and receiver layers are incorporated onto the same substrate carrier. Products based on single-sheet assemblies are dominant. The substrate can be reinforced paper, polyethylene terephthalate or lithographic grained and anodised aluminium.

3.3. Color Proofing [277]

3.3.1. *Basics*

Four-color prints are produced by the subsequent printing of the three subtractive primary colors (cyan, yellow, and magenta). All color nuances, including black, can be reproduced in this way. However, black is usually added as the fourth color to increase contrast or, as an independent color, to reproduce deep shadows [278].

It is necessary, at different points in the production process of a four-color print, to check the image quality prior to printing or give the customer a reliable impression of the final printed result. *Press proofs* are still used today when the highest fidelity of reproduction is demanded or when a medium-to-large number of proofs is required that must be shown to people who are involved in decision making at different places.

In many other cases, the same purpose can be achieved by making a single four-color copy from the screened color separation films, the *color proof*. Thus, the technician and the customer can be given an advance visual impression of the colors before production of the final printing plates and before printing. In principle, this can be achieved on a color monitor as well (*soft proof*). In this case, screened subtractive primary colors must be reproduced by using additive primaries. The latter are emitted by luminescent substances placed at the monitor in a raster arrangement. This poses a problem.

In contrast to soft proofs, a color copy on paper or other materials is known as a *hard proof*. The appropriate materials are exposed in contact with the color separation films. A negative color-proofing material must be used for negative films and a positive color-proofing material for positive films.

The color pigments or dyes applied in the production of color proofs must correspond as closely as possible to the pigments in the printing inks to be used later in printing. Because standard colors vary in definition in different countries, different color-proofing systems must be employed with regard to color reproduction. Even the light source used to judge color proofs is standardized.

3.3.2. *Applications*

Color-proofing materials are used [277] in various applications: (1) for quality control checks in the production of color separation films and as indicators for necessary corrections, (2) as internal quality proofs during production, (3) to check pictures from different sources for comparable quality of reproduction, (4) as layout proofs, (5) as color proofs for customer approval, or (6) as guideline for the printer at the printing press.

In commercial printing color proofs may be more economical than pressprinting if only a very limited number of large-size four-color images is required as drafts for advertising agencies, and in the production of animated films.

The quality of reproduction required for various applications differs. For example, a layout proof need not be as exact as the proof shown to a customer for approval. However, for simulation of the highest quality, the characteristics of the printing process to be used must be taken into account in producing the color-proof separations. This means that other colors and, therefore, other color-proofing results are required for offset printing, for flexography, or for gravure proofing with offset means, but even then the characteristics of a special offset press must be simulated in the color proof. Thus, quality levels differ even among color-proofing processes.

On the one hand, the (older) *overlay* systems result in four subtractive color images on transparent plastic film which when placed one upon another in register give a four-color picture. On the other hand, in the (more recent) *single-sheet* systems the four-color picture appears on an opaque white background (ideally, on the final printing paper).

3.3.3. *Overlay Systems*

Overlay materials generally consist of a transparent (film) substrate which is coated in the factory with one or more dyed and/or light-sensitive layers. The light sensitivity is caused by photosolubilizing (in the case of positive transparent originals) or photocuring (in the case of negative transparent originals) compounds in the coating. After respective exposure through the four corresponding screened color separation films and after development, four imagewise colored films are produced which, when accurately registered, give a colored impression of the expected result.

The *advantage* of this procedure is that it also allows examination of a combination of single films. So the printed result can be compared with the set point at all stages in production. The *disadvantage* of overlay systems is that four separate films with their imperfect transparency of the film substrate cause a marked graying of the picture and background.

Trade Names. Overlay systems available on the market differ in structure and processing procedures. The overlay proof Color Key, produced by 3 M, is available only for negative copies [279]. It has a light-sensitive and a colored coating on the carrier film. The more soluble (in this case, the unexposed) areas of the coating are removed during development in the MR 424 processor.

Hoechst – Celanese offers an overlay system called Colorlink which is available in both a negative and a positive version [280], [281]. In these products, the light-sensitive diazo layer contains the dye as well. Here, too, a developing liquid and a processing device (CF 262) are available. The developing liquid is an aqueous solution.

DuPont has developed a product that does not require a liquid developer. This product, called Cromacheck, is available only for negative copies and has a sandwich structure: (1) opaque film, (2) light-sensitive colored photopolymer layer, and (3) clear film. After exposure, the exposed areas adhere more strongly to the clear than to the opaque film. Conditions are exactly reversed for the unexposed portions. Image areas are then separated from nonpicture areas by physical separation of the two films (peeling apart). The color image is present on the clear film. An advantage of proofing with Cromacheck is its quickness [282].

3.3.4. *Single-Sheet Systems*

In comparison with overlay systems, the advantage of single-sheet products is that the proof not only is made of one piece like a press print, but can also be handled like a print.

A distinction must be made between systems that are *produced* as colored layers and those in which the color or pigment is applied during *processing*. In the former, the color density (and the color hue) is predetermined by the production process. In the latter, the color density can be influenced within certain limits during processing.

3.3.4.1. *Precolored Systems*

In the simplest precolored systems, the *user* applies to a substrate a standard pigmented or colored, light-sensitive dispersion or solution that hardens on exposure to light. The material is then dried, exposed, and developed (usually with tap water). The procedure is repeated with the four process colors to give a four-color proof.

This system has many advantages including low cost of materials and miscibility of colored liquids plus the fact that the color density can be varied to some extent by varying coating conditions.

Trade names are Watercote, Kwik-Proof, and dr Color from the Direct Reproduction Corp. These systems differ mainly in the method of coating. Two other products of this kind, Aqua Color Proof from Graph Mark and Quadracolor from Castcraft Industries, also use porous carriers, such as paper, after the pores have been impregnated with a special covering layer to prevent penetration of the light-sensitive formulations into the paper.

Another approach involves *presensitized products*. These offer not only a guaranteed constant coating quality but also greater ease of processing. Here, the color density of the final product cannot be manipulated.

Trade names are 3 M Transfer Key, 3 M Matchprint II [283], Hoechst – Celanese Colorlink Pressmatch [284], [285], and Fuji Photo Film Color Art. Apart from Transfer Key, which is offered only as a negative system, all other types are offered in both negative and positive versions.

The light-sensitive principles used are the photolysis of quinone diazides, photohardening of diazonium salts, or photoinitiated polymerization.

Another single-sheet system that is processed similarly, Agfa – Gevaert Agfaproof, was introduced in 1986. It is based on the silver halide technology and is used for the production of four-color proofs with negative transparent films.

Despite the somewhat important differences in application, the processing principle of these systems can be simplified as follows: The light-sensitive layers, which are often embedded between two foils for transport and storage, are first laminated onto a receiver material. This is achieved via a special adhesive layer that is activated by heat or pressure, the latent adhesiveness of the light-sensitive layer, or wet transfer (Agfaproof). The layer is subsequently exposed in contact with the corresponding color separation film and then developed. This process is repeated four times to give a four-color proof.

Only in the case of Color Art from Fuji Photo Film is the light-sensitive layer exposed and developed before being transferred to a receiver sheet. The advantage of this system is that all four process colors can be exposed simultaneously side by side. However, accurate transfer of the single color pictures causes certain difficulties.

In some of the products mentioned above, the four-layer structure can be transferred from the receiving sheet to printing paper. This gives a particularly good simulation of the four-color print.

3.3.4.2. *Systems With Colors Generated During Processing*

Ektaflex PCT from Eastman Kodak has a silver salt diffusion color film structure in which the colors are generated during the development process (see Section Color-Copying Materials Based on Dye Diffusion Chemistry; ⟶ Photography).

Photochemical Imaging. As described in Section Photoinsolubilizing (Photocuring) Systems, not only can radical photoinitiated polymerization be used to prepare materials that are developed by dissolving the nonpolymerized areas off the carrier, but can also be used to formulate the starting material so that it has a sticky surface which imagewise loses its stickiness during exposure. In an *additive development process*, the layer is then treated with a colored powder that adheres permanently only to the sticky areas of the layer [266]. This principle has been used by DuPont in the concept of its positive acting Cromalin system [286].

The toner can be applied manually with pads. However, more uniform results are obtained when the toner is applied in a toning device (e.g., ATM III). The application of the color powder can be totally avoided by coating the toner on a polyolefin film which is brought in contact with the sticky image. In this way, the toner layer is transferred imagewise to the exposed proof material without soiling the machine with toner powder [287].

A negative version of the Cromalin process has been developed as well. Here, removal of the cover film from the exposed material leaves the *exposed areas sticky*. This material is processed similarly to positive material.

Additive development is also used in another negative proofing process called Spectra, which was developed by KEUFFEL and ESSER [288] and is marketed by Polaroid. The light-sensitive material consists of a polyester film coated with a sticky polymer layer, a light-sensitive naphthoquinone diazide layer, and an overcoat. On imagewise exposure and subsequent development, the sticky layer is uncovered in the exposed areas. The quinone diazide layer is completely solubilized in a second exposure step, which is integrated in the processor. After additive development with color powder, the four color pictures are developed again and finally successively transferred in register to a receiving material (e.g., the printing stock).

Electrophotographic Imaging. The advantage of electrophotographic imaging lies in the extreme light sensitivity of electrophotographic materials. An added advantage in proofing is that color density and screen dot diameter can be controlled independently.

The disadvantage of this system is that imaging and reproduction quality depend to a great extent on the developer, usually a dispersion of fine, dyed resin particles in an aliphatic hydrocarbon (isopar). The key issue is to maintain the stability of this dispersion over a long period under varying climatic conditions.

The first process of this type, the Remak method, was developed by Research Laboratories of Australia. However, the method was not adopted commercially because of the extreme sensitivity of the employed zinc oxide materials to atmospheric humidity.

Kimoto of Japan has put a similar proofing system, Kimofax, on the market. This method uses an organic photoconductor and four colored dry toners. However, it can be employed only to proof simple line work (e.g., in map printing).

In the late 1970s, a high-resolution electrophotographic recording material was made available that used photoconducting layers of cadmium sulfide [243] which were sputtered on steel at high frequency (KC-crystalplate). This material, together with the high-resolution color toners of the Remak type, is ideally suited to the production of an electrophotographic proofing system. Considerable development in the area of electrophotographic proofing occurred as a result of collaboration between Coulter Systems Corporation of the United States and Stork of the Netherlands. The final version of the imaging cycle, conducted in a largely automated device, consists of the following steps: (1) charging (ca. 35 V) of the KC plate, (2) contact exposure, (3) toner application, (4) transfer of the toner image to an intermediate, and then (5) transfer to the final receiving carrier (e.g., the printing paper). The image quality obtained is comparable to that of photochemically produced proofs.

In 1986 Eastman Kodak introduced another electrophotographic proofing system called Signature. Instead of working with a double toner transfer, this system uses an organic photoconductor applied to a transparent, conducting carrier which is exposed from the back through a color separation film. All four subtractive color pictures are collected on the photoconductor and transferred to coated paper in a laminator outside the imaging device.

The equipment requirements of all electrophotographic proofing systems are relatively high. Because the light sensitivity of the materials used is very high, they must be processed in a closed device to avoid working in a darkroom.

3.3.5. ***Digital Proofing***

Color separation is carried out at present in color separation scanners. Color data are stored in an electronic memory and can be recalled and inspected on a high-resolution color screen (*soft proof*). However, an additional proof on paper or other substrate (*hard proof*) is often needed. A satisfying technical solution to this problem is not yet marketed. However, a number of possibilities exist, and various announcements have been made that new systems will be introduced in the near future. In principle, the approaches being considered are silver halide materials, photochemical systems, electrophotography (see Section Electrophotography), and other digitally controlled imaging methods (thermotransfer, ink jet, see Sections Thermographic Printing and Ink-Jet Printing).

The use of silver halide (i.e., color film materials) as continuous tone materials for the simulation of screened printed images is possible only to a limited extent. Polaroid is developing a silver – halide based system, which can be directly exposed in a scanner. The 3 M Company has proposed the use of photochemical systems for the production of digital color proofs (Dataproof). Ultraviolet lasers are employed for imaging.

Digital Electrophotographic Color Proofing. In 1986 DuPont announced a new digital electrophotographic color proofing method DDP (direct digital proofing). Digital data from the scanner are stored in an optical memory and recalled into the DDP

system. The charge images are developed with liquid toner and transferred to the printing paper.

Eastman Kodak is working on the adaptation of the Signature method to digital data input. Stork is developing a *scanner* that can be employed for the digital exposure of the inorganic photoconductor in use.

Digital Ink-Jet Color Proofing. The ink-jet method can be used for the multicolored digital imaging of plain paper. This method has the advantages of low cost and high speed. On the other hand, color matching, uniformity of larger color areas, and reproduction of conventional screened structures have not yet been perfected, although improving rapidly. This method can be employed for position and layout proofing. It is manufactured by Iris Graphics.

Digital Thermotransfer Color Proofing. In the *thermotransfer* (see Section Thermographic Printing) method, the color substance is transferred thermally point by point from a coated paper to a receiver sheet. The digitally controlled transfer is repeated four times to give a four-color image. Transfer is achieved by sublimation of the dye or by contact transfer of a thermoplastic substance under the influence of a thermo head. The result resembles a continuous-tone picture and has a dot density of 120 dots per centimeter. The imaging device is manufactured by Dainippon Printing Co., Tokyo.

3.4. Platemaking

Printing is defined, according to the DIN standard [289], as “the reproduction in any quantity of a presentation in text and/or picture form by transfer of printing ink or coloring substances to the material to be printed using a printing form”. Each of the four printing processes uses a different kind of plate [214].

In *lithography*, image and nonimage areas are essentially in the same level, but they differ in their physicochemical behavior. The image areas are oleophilic; the nonimage areas, hydrophilic.

In *letterpress printing*, the image areas are raised above the surface level of the plate, and the nonprinting areas are lowered to different extents.

In *gravure printing*, the image areas consist of engraved or etched wells from which ink is transferred. The surface of the gravure cylinder acts as nonprinting area.

In *screen printing*, nonprinting areas (i.e., ink-impermeable areas) are protected by a stencil. Areas not covered by the stencil represent image areas.

In the Federal Republic of Germany these processes were used in 1986 for printing as follows: 57 % lithography, ca. 26 % letterpress, and ca. 17 % gravure printing. Screen printing is used mainly for textiles and in the production of printed circuits (see Section Printed Circuits (Printed Circuit Boards)) [290].

3.4.1. *Lithography (Planographic Printing, Offset)*

Lithography was invented by SENEFELDER in Munich in 1797. Because he used limestone from Solnhofen as the carrier, this method is called “writing on stone” or lithography. The American W. RUBELS made lithography an industrially useful printing process, by transferring the printing ink from the plate to an intermediate carrier, the rubber blanket, and then from the blanket to printing paper (offset).

Because of its technological characteristics, offset printing is widely employed worldwide in the production of high quality one- and multicolor prints for medium runs in advertising material, calendars, art prints, books, brochures, newspapers, and packaging. Runs ranging from 10000 to 250000 are typical, but under certain press conditions, much higher runs are achievable.

An important characteristic of offset printing is that platemaking is simple, quick, and economical. On a suitable hydrophilic carrier (e.g., a plastic film coated with silicates in a binder) oleophilic image areas can be made by hand or with a typewriter for simple black and white line work.

For more demanding work plates with a light-sensitive coating are used. *Presensitized (PS) plates* of this type are produced by various manufacturers and are available commercially. World sales in 2001 were ca. 360×10^6 m^2.

The photomechanical method, which uses subtractive development, is described in Section Application. Once the information for printing exits, either in digital (computer) or analogue (film) form, it can be transferred to the plate by exposure in a matter of minutes. The exposed plate becomes a printing plate on development. A lithographic plate can be prepared in ca. 5–8 min.

Wipe-on plates are still being used in some countries, mainly the United States and East-bloc countries. Worldwide consumption in 2001 amounts to ca. 40×10^6 m^2. These plates are not precoated.

Printers buy the uncoated carrier, and then coat it with a light-sensitive solution shortly before imaging and then dry it. Wipe-on solutions are supplied only for negative plates. The resulting coated plates usually have a storage life of a few days before they are no longer developable. Once exposed, these plates are developed with an emulsion lacquer. The lacquer dissolves the unexposed light-sensitive coating (<1 µm thick) from the substrate, and deposits resins and pigments on exposed areas to improve the stability of these image areas during the printing process.

All light-sensitive lithographic plates, whether wipe-on or presensitized, have a simple structure: (1) substrate with sufficiently high mechanical strength and a hydrophilic or potentially hydrophilic surface, and (2) a light-sensitive layer.

3.4.1.1. *Substrates*

Demands made on substrates in lithography arise from the stresses to which the plate is exposed during both platemaking

and printing [291]. During development, the exposed plate usually comes into contact with acidic and alkaline salt solutions, and occasionally with organic solvents. The surface of the plate is often brushed or sponged. The finished plate is punched, bent and finally mounted on the metal cylinder of the press. During each printing cycle, the surface of the plate comes in contact with an acidic (pH ≈ 5) dampening solution and then with solvent-containing ink, which is transferred from the plate to paper via the rubber blanket.

For these reasons, the substrates must be resistant to kink marks, creases, and scratching during use. The material must be totally flat and finally should have a high dimensional stability, as well as high abrasion resistance in the press.

Great demands are also made on the chemical resistance of the surface to acidic and alkaline agents. Not only should the surface be stable to corrosion during use, it should also have a well defined microroughness and a high hydrophilicity for spreading the dampening agent. The hydrophilicity should not be lost on repeated and prolonged contact with printing ink. Microroughness also improves adhesion of the printing layer.

The presslife expected of an offset printing plate varies between some ten and several hundred thousand runs depending on the substrate and the printing layers.

Special water-resistant paper with a hydrophilic coating can be employed as substrate for runs of a few hundred or thousand.

Composite materials made of paper and aluminium foil or plastic film (usually polyester film with a hydrophilic coating) are suitable for runs up to ca. 10 000. These materials have better wet strength and dimensional stability.

Highly pure (99.5 %) or low-alloy (ca. 98.5 %) *aluminum* is generally used as the substrate for lithographic plates. The surface quality and material properties of this litho-type aluminium, used in the production of offset plates, are stipulated by extremely narrow specifications. At first, the raw litho-type aluminium was processed in sheets only; however, the processing of coils is the method of choice today. After this treatment, the final product is generally an aluminium web that is finished on one side and often coated in line. However, there also exists a market for plates that are grained and coated on both sides (Mainly in the USA), this market is declining rapidly. These plates are used for shorter runs (up to 20000–30000) and average printing quality.

Surface Graining. Mechanical graining is one of the oldest methods used for surface finishing. If the aluminium surface is treated with rotating wire brushes, it acquires a silky luster. The surface structure then shows a depth of roughness of R_Z = 3 – 6 µm, with a marked *anisotropic* roughness in the direction of brushing. When abrasives such as suspensions of quartz, corundum, or pumice powder in water are used, the carrier surface acquires an *isotropic* roughness, with R_Z values of 2 – 10 µm. The abrasive suspension is either pressed onto the surface of the aluminium sheet, with the help of porcelain balls, or applied with rotating nylon brushes, or sprayed on under high pressure [292].

Electrochemical graining has been used increasingly since the mid-1960s. The surface quality of the aluminum plays an important part in this method. The raw aluminum surface is first degreased and then uniformly corroded by using an a.c. treatment in dilute nitric acid, hydrochloric acid, or mixtures of these with other acids. This process of graining is a kind of pitting corrosion which attacks the entire surface of the aluminum leaving a very uniform rough surface [293], [294].

The depth of roughness can be controlled within wide limits (2.5 – 9 µm) by varying the electrical parameters. Both processing speed and the uniformity of the resulting substrates depend on temperature, flow rate of the electrolyte, geometry of the tank, and other parameters. Combination processes have also been used in graining. In these processes, mechanical graining is followed by electrochemical treatment [295].

Graining with wire brushes produces the lowest increase in surface area of the plate, and electrochemical graining the highest. This influences not only the quality of reproduction and the optical resolution but also the printing properties (water spreading and length of run). The finer the graining of an aluminum surface, the lower is the mechanical resistance. With wire-brushed aluminum, a second hardening step is not required. However, an anodic oxidation step is essential for electrochemically grained aluminum.

Anodic Oxidation. Anodic oxidation involves conversion of the uppermost layer of the aluminium workpiece to aluminium oxide, without affecting the surface topography too much [296]. Two procedures are primarily used for this electrochemical reaction in which aluminum is connected as the anode: anodic oxidation in sulphuric acid [297] or in a phosphoric acid electrolyte [298]. The former leads to smaller pores and thicker oxide layers than the latter. Mixtures of the above-mentioned acids and multistep processes are described in the patent literature [299].

In general, not more than *5 g* of oxide is produced per square meter; depending on the roughness of the surface, this corresponds to a layer thickness of ca. 1 µm. The oxide formed during anodising is X-ray amorphous and an extremely strong adsorbent. The properties of the substrate are greatly improved during this step, i.e., the adhesion to the light-sensitive coating; resistance of the surface to chemicals; its hydrophilicity, hardness and abrasion resistance, and thus the potential of running length. However, the extent to which the aforementioned aspects can be effective depends on process parameters and on the process itself.

Posttreatment. Posttreatment of the oxide with certain hydrophilising solutions has proved advantageous in the production of PS-plates. The main purpose of this treatment is to improve the hydrophilicity of the aluminum oxide surface. However, the adhesive properties, developability, and shelf life of the plates are influenced at the same time. Solutions of the following substances have been proposed for this purpose: alkali silicates [300], phosphoric acids [301], hexafluorometalates of group 4 elements (titanium, zirconium, hafnium) [302], heteropolyacids, oxo acids of pentavalent phosphorus [303], and hydrophilic colloids [304]. All of these compounds coat the surface of the aluminium oxide to generate a very polar covering layer, which improves the hydrophilicity of the surface.

3.4.1.2. *Coating and Further Processing [305]*

Preliminary treatment of the carrier is followed by in-line coating of the plate in a fully continuous process. A solution of the components of the light-sensitive coating in a more or less volatile solvent mixture is applied to the carrier in a defined wet film thickness. The coatings are generally applied with a fishtail die, with a roller coater, or by spraying. The thickness of dry coatings is 0.3 – 5 µm.

The line speed in modern plants is typically in the range 40–80 m/min. The length of individual steps is adjusted to this process speed. After drying, the web is automatically inspected for coating defects. It is cut apart, and defective plates are removed. Cutting of the plates requires special attention. On the one hand, the plates must be cut to size with an accuracy of some tenth of a millimeter. Plate measurements must also be able to be changed quickly and accurately, which is not an easy task with these high process speeds. On the other hand, the plates must be cut at right angles, and finally, no burrs should be produced during cutting because these would cause problems for the customer. Presensitized lithographic plates are available in many sizes up to ca. 3 m^2 (1500 × 2000 mm). The finished faultless plates are stacked by using a suitable interleaving paper and packed in lightproof wrapping paper (if necessary, after shrink-film packaging).

The shelf life required for these presensitised lithographic plates depends on the extent of their distribution. Also, the more diverse the climatic conditions under which the plates are to be stored and used, the more diverse are the requirements.

3.4.1.3. *Processing of Presensitized Plates*

The light-sensitive layer applied in the coating step is a mixture of numerous components, which is formulated to meet the requirements of the techniques used in plate manufacturing and to fulfil the demands of the customers in using the plates. For various applications, different types of plate are available.

In principle, similar processing procedures are applicable to all presensitised plates. In the first step, light exposure, certain selected areas of the plate are photochemically or photophysically changed, leaving the remaining portions unexposed. How this step is undertaken depends very much on whether the information to be printed is available in digital (computer) or analogue (film or paper) format.

In analogue format, the method of transfer depends very much on the format of the film or paper original. This step is usually conducted in a copying frame if an original-sized transparent film is used as the original. The copying frame is evacuated in the process to make sure that the film original is in close contact with the light-sensitive coating on the plate during exposure. If an enlargement or a reduction in size of the original is required, plate exposure is carried out through the film-original in a projection camera. Exposures of especially sensitive presensitised plates can also be made in a projection camera with opaque originals and reflected light. However, with the advent of Computer-to-Plate technology and digital workflows, the use of this particular system has diminished significantly in recent years.

Where the information is in digital format, plates are exposed with a platesetter. Here, all areas of the plate are not exposed at the same time (parallel exposure), but one area is exposed after another (sequential exposure). A focused beam of light, generally a laser beam, is used to scan the light-sensitive material line by line while being modulated imagewise as shown in Figure 59. A complete picture is created by the close lining up and partial overlapping of many modulated lines of light.

Figure 59. Image composition at line-by-line exposure

In the case of photosensitive plates, the actinic spectral region is limited at the lower end by the absorption of UV light by glass and the physiological dangers of shortwave UV radiation, and at the upper end by the sensitivity of the light-sensitive material to yellow light (room lighting).

In the next step, development, the exposed plate is stabilised in a specific way to make sure that the image information is not erased by further action of light. This can be achieved by a chemical reaction, by washout development (subtractive), or by the deposition of a covering substance (additive; see Section Application). Suitable developers are offered and recommended by plate producers. These preparations are designed specifically for the development of a particular plate and can usually be employed for that plate only. The plate and the developer form a “processing system.” Development is usually carried out in a processor in which the developer is applied and allowed to act on the plate in a very specific manner. The system dependence between plate and developer is in general stronger than that between plate, chemical and processor.

The third step that generally follows development directly is gumming. It serves the purpose of protecting the hydrophilic nonprinting areas from contamination by fatty substances in the air. An aqueous solution of film-forming hydrophilic colloid (dextrin, gum arabic) is applied to the plate and dried.

Despite the fact that plates of various makes have these processing steps in common, a number of different ways of converting a presensitised plate to a finished printing plate still exist. These methods differ in the starting material, on the type of original copy available (e.g., positive film, negative film, microfilm, paper, or a data carrier); in the time available for platemaking (e.g., in newspaper printing or packaging); and in the use of the plate (e.g., to print an art calendar, a business form, or a metal sheet). The chemical systems applied in several process variations for the preparation of conventional (analogue) plates will be described.

Negative Transparent Film. Platemaking for newspaper printing is a typical example of the use of a negative film in full plate size as the original. Most of the newspapers produced in Europe by the offset principle are printed with the help of presensitized plates. In the United States however wipe-on plates are still being used widely although this market is also declining as newspapers move to CtP.

To transfer information on the negative film to the plate so that the positive image is obtained, the coating on the plate corresponding to the transparent areas of the film original must be converted to the printing areas. This is achieved with a photoinsolubilising light-sensitive layer. Plates of this type are called negative plates. Negative plates are used in more than 66 % of the printing done in the United States. However, this is declining rapidly as the USA is moving more quickly to CtP solutions than most markets.

Presensitized negative plates suitable for newspaper printing must be able to be exposed and developed quickly because many printing plates have to be made just before the start of the presses. Also, these plates must have good wear characteristics in the press because newsprint has a rough, dusty surface and the web offset presses employed in newspaper printing have very high paper speed (ca. 10 m/s). The plates should produce up to 150 000 good prints. They must be anodized.

Various light-sensitive systems are used by individual producers, which may be broadly categorised as either photoinsolubilizing, photopolymerising, or photocrosslinking systems:

1. Oligomeric diazonium salts are widely employed. In some cases, the coating also contains a special binder, e.g., an acetal resin. The ease of development, as well as the printing properties of the plate, depend on the modification of the resin.
2. Photoinitiated polymerization is also used in the production of negative plates. These layers can be built up with binders containing carboxyl groups and then developed with aqueous alkaline salt solutions. Plates based on this principle are extremely robust and abrasion-resistant.
3. The reaction products of azide-substituted aromatic carboxylic acid derivatives with polymers having hydroxyl groups can also be used in negative plate coatings.
4. The same applies to plates with unsaturated polyesters that can undergo light-induced cross-linking. These polyesters are made from phenylenebis(acrylic acid) and 1,4-bis(hydroxymethylcyclohexane).

All these coating formulations contain dyes or pigments, in addition to the actual active substance, and change color on exposure. Thus, the user can distinguish between exposed and unexposed plates. The plates also show a good color contrast between image and nonimage areas after development. The increasing trend on environmental grounds towards aqueous-based developers in recent years mean that plates of types 1) and 2) tend to dominate in today's market.

Aqueous diazonium salt solutions are used for *wipe-on plates*. These plates are not dyed and have very thin coatings. Because their shelf life is limited, they are always prepared shortly before use. Wipe-on plates are developed with an emulsion lacquer that has a double function: On the one hand it dissolves the unexposed areas. On the other hand, a resin is deposited from the lacquer on the exposed images to strengthen the thin exposed layer.

Presensitised negative plates are produced by Agfa-Gevaert, Fuji Photo Film Co., KPG, Lastra, and others. *Wipe-on plates* are produced by American Litho, Anocoil, Citiplate, and Western Lithotech.

Positive Transparent Film. In vast parts of Europe and Asia, the use of positive transparent film as the original copy is the method of choice for platemaking—to print advertising, pamphlets, brochures, and art prints, in most cases in four-color print. If the transparent areas of the film original are to become the nonprinting areas after plate development, a presensitised offset plate must have a photosolubilising coating. Plates of this type are called positive plates. Positive plates are usually used for art prints (e.g., exhibition catalogues). For a four-color print (see Section Basics), four plates are produced by exposing positive plates through the positive color separation films.

Whereas the photocuring layer of negative plates capable of sufficiently long runs requires a definite minimum amount of light to cure the image areas, the number of prints made with a positive plate is independent of exposure time. The light merely serves to solubilise the nonprinting areas of the coating. For this reason, image reproduction can be controlled by varying the exposure time of positive plates without the risk of insufficient run length. This is the starting point for the standardizing system introduced to the European market, which provides the prerequisites for a highly uniform quality of reproduction. A further advantage of the positive copy is that the original, especially for pictorial depictions, can be checked more easily for mistakes, and the color separations (i.e., the film originals for color printing plates) can be mounted more easily in register on film than with negative originals.

Positive plates use grained, anodized (1 – 4 g of oxide per square meter) aluminum as a substrate. The light-sensitive formulations contain naphthoquinone diazides, binders, and additives such as dyes, indicators, and pigments. These additives bring about good handling properties, i.e., they improve color contrast after exposure, colour contrast after development, and developer resistance.

After exposure, positive plates are developed with aqueous alkaline solutions (pH 12.5 – 13.5) of alkali phosphates or alkali silicates. Plates made this way can be subjected to heat treatment (6 min at 230 °C) to make them capable of longer press runs. The diazonaphthalinone derivatives decompose at this temperature. They subsequently undergo a complicated cross-linking reaction which improves the mechanical properties of the layer to such an extent that these baked plates are capable of press runs two to three times longer than those obtained with untreated plates [306]. In practice, the plate is covered with a protective layer of heat-resistant, hydrophilic, film-forming material before being baked to make sure that nonimage areas are not soiled with oleophilic substances [307].

Diazonaphthalinone on exposure forms derivatives of indenecarboxylic acid, which can easily undergo decarboxylation under

certain conditions. This reaction helps to utilize these plates not only in photosolubilisation of quinone diazide layers, but also in exposure of the same plate through a negative film, followed by the decarboxylation of the indenecarboxylic acid in the exposed areas (2 min at 100 – 140 °C) [308] to give image areas insoluble in the developer. A subsequent overall or imagewise exposure solubilises areas of the plate that were previously unexposed. In this way, both positive and negative originals can be economically used to expose the same plate, thereby saving the expense of additional film.

Producers of *presensitised positive plates* are Agfa-Gevaert, Fuji Photo Film, KPG, Lastra, and others.

The *waterless plate* represents another modern variation of platemaking with positive transparent originals for special application. Whereas in conventional offset printing, the nonimage areas are wetted with water (the image areas are hydrophobic) to prevent the ink from wetting these areas, the nonprinting areas of a waterless plate consist of a silicone rubber, which repels ink even without water [309]. These plates are also available in a negative version [310]. They are produced by Toray.

Projection Platemaking using Negative Film or Microfilm Originals. Processing microfilm was very popular during the 1980s as it involved the saving of film costs. This method was used advantageously in printing books. Today this is very much a *niche* market mainly for poster printing. Annual sales are < 1,000,000 m^2. The equipment used includes a source of UV light and an optical system for enlargement and projection of the microfilm onto the plate.

Highly light-sensitive special products must be employed as presensitised printing plates. The coating of these plates is based on various photocrosslinking mechanisms: (1) photolysis of aromatic diazonium salts, (2) photodimerisation of cinnamic acid derivatives, (3) photoinitiated polymerization, and (4) silver halide photolysis. The light sensitivity of the materials increases in this order. Although the energy requirement of type 1 can be reduced to approximately 10 mJ/cm^2 at best, values of 0.1 – 1 mJ/cm^2 have been reached with types 2 and 3. In comparison, silver halide materials require still less energy (<1 $\mu J/cm^2$) for imaging. Of the individual highly light-sensitive plates, types 2 and, especially, 3 are capable of the longest press runs. Producers of highly light-sensitive negative plates are Agfa-Gevaert (types 1 and 3), Fuji Photo Film Co. (type 4), and KPG (types 1 and 2).

Positive Paper Originals. Positive paper originals can be used as input in two ways: parallel exposure in a camera or sequential exposure in a scanner [311].

Camera Imaging. Camera imaging is increasingly important in small-size offset printing for short runs and in newspaper printing with different material–equipment combinations.

1. In the Opti-Copy camera or in the Sixt RHI, the image of a positive paper original is projected, by means of an optical system, onto a silver halide-sensitized plate which uses the *silver salt diffusion principle* for imaging. In this case, a transparent polyester film is employed as substrate. Because these cameras do not contain deflecting mirrors, exposure is made through the back of the plate to make sure that a right reading printing image (required in offset printing) is obtained. Prints in quantities up to 50 000 can be produced. Producers are 3M, Agfa-Gevaert, and Mitsubishi Paper Mills. A xenon discharge tube (with filter) is used as the light source in both cases.
2. Paper printing plates coated with zinc *oxide* (*inorganic photoconductor*) have found wide application for shorter runs in small-size offset printing. This is an economical method for printing several hundred A4 or A3 copies. Special cameras that take a roll of zinc oxide-coated material and are equipped with high-voltage corona, table for the original, exposure platen, illuminating device, optics, and development device are produced by different firms (Eskofot, Itek, Iwatsu Electric Co., Ricoh, and Fuji Photo Film).
   The photoconducting zinc oxide pigment is embedded in a resin and sensitised with sensitizer dyes. An electrically conductive paper serves as the substrate. Processing is carried out as described in Section Photophysical Principles of Image Generation. A suitable salt solution (e.g., K, [Fe (CN),] or, more recently, other compounds such as phytic acid) is finally used to convert the zinc oxide surface that is not covered with toner to the corresponding zinc salts, which function as hydrophilic nonimage areas in offset printing. The energy requirement for imaging a zinc oxide plate is 1–10 $\mu J/cm^2$.
   Producers are AM international, Oji Paper, Ricoh, Scott Graphics, and Tomoegawa.
3. Another variation of electrophotography is used extensively by newspaper printers for platemaking [312]. In this process, *organic photoconductors* are employed as the light-sensitive substance. They are embedded together with sensitiser dyes in an alkali-soluble binder matrix. Imaging occurs as described in Section Photophysical Principles of Image Generation. Finally, the areas of lithographic carrier not covered by toner are uncovered with the help of an alkaline decoating solution. Only the toned areas still contain organic ink receptive substances. This plate can then be used in a web offset press in the same way as a negative or positive plate. An automatic camera of the type employed for processing such electrophotographic plates is shown in Figure 60.
   The advantage of this method of platemaking, in addition to lower material cost, is a reduction of labor and time required for the last (and latest) pages of the newspaper.

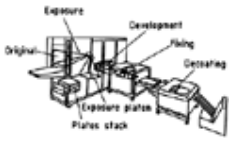

Figure 60. Automatic camera for processing electrophotographic plates

Producers of presensitised electrophotographic plates of this type are Agfa-Gevaert and KPG.
A variation of electrophotographic platemaking is a process that utilizes an effect called persistent conductivity. The material is exposed without first electrically charging the surface. The electrons released on exposure to light lead to a certain conductivity of the material which persists even after exposure [313].

Scanner Imaging. The sequential imaging of radiation-sensitive plates may be utilized for the facsimile transmission of imaging signals and the transfer of data from computer by direct modulation of the light source.

A variety of light sources are employed for sequential (scanner) imaging: water-cooled argon-ion lasers in UV mode (351 and 364 nm), helium – cadmium lasers (442 nm), water-cooled argon-ion lasers in visible mode (488 – 514 nm), air-cooled argon-ion lasers (488 – 514 nm), helium-neon lasers (633 nm), LED (light-emitting diodes), laser diodes, normal incandescent lamps, and YAG lasers (yttrium-aluminum-garnet) (1060 nm).

A correspondingly large number of radiation-sensitive materials are used for the production of lithographic plates. The light source must be adjusted to the recording material in such a way that the following requirements are fulfilled:

1. The radiation employed must correspond to an absorption and sensitivity band of the recording material.
2. The energy density of the radiation must be high enough to induce a change of the recording layer within a time span that is justifiable for practical use and which is sufficient for printing purposes. In other words, a weaker light source would require a more sensitive recording material.

The closer the arrangement of the exposure lines, the better is the quality of reproduction achieved in sequential, line-by-line exposure. However, the sequential exposure of a presensitised plate becomes more time-consuming for larger size plates. A simple estimation shows that with an area of 600 μm^2 per image element (which corresponds to a screen dot of 2 % area coverage in the 60 metric screen) and with a plate size of just 400 × 600 mm^2, only 0.15 µs is available for the exposure of a single element if exposure of the entire plate is to be accomplished in 1 min.

Lasers beams that satisfy the first condition for a given recording material, after passing through the optical system and reaching the plate with a light energy of 0.4 W still require a radiation-sensitive coating, capable of being imaged with a light energy of not more than 10 mJ/cm^2 for imaging. Highly light-sensitive coatings must be used [311]. Coatings based on both diazonium salts or on photodimerising polymers can be formulated to give a light sensitivity of ca. 10 mJ/cm^2. In particular, further development of photoinitiated polymerisation has led to the production of highly efficient plates that require only 50–150 $\mu J/cm^2$ for imaging. The use of such plates and systems is particularly popular in newspapers where fast make-ready and durability of run are important factors. Producers: Agfa-Gevaert and Mitsubishi Chemical Corporation.

The flatbed scanner shown in Figure 61 illustrates the principle of image production.

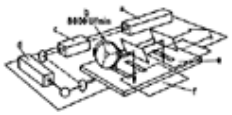

Figure 61. Flatbed scanner for reading (——) and writing (----) an original. Fiber optics of the reading beam, photomultiplier, and signal input to modulator are not shown

a) Argon-ion laser; b) Pyramidal mirror (up to 8000 rpm); c) Modulator; d) Helium – neon laser; e) Original; f) Plate

Although conventional UV-sensitive plates require several times this density of energy and are not suitable for imaging in the manner described above, a system has been recently developed by BasysPrint that enables conventional UV-sensitive plates to be integrated into the digital workflow. The use of such a UV platesetter offers significant cost advantages over other fully integrated digital systems.

Electrophotographic materials can be sensitized for different wavelengths of the *visible spectrum,* including the region around 500 nm. The energy requirement for imaging is ca. 40 $\mu J/cm^2$. Hence, although principally developed for use with projection cameras, plates of this type can be imaged with small, air-cooled argon-ion lasers that have a power of only a few milliwatts of light energy (on the plate). Flatbed scanners were used during the 1980s for the direct recording of facsimile transferred data on electrophotographic plates [311] (producers: Agfa-Gevaert, KPG).

The above describes early attempts at direct lithographic plate making. From the early 1990s computer memory and laser optical design were able to store and output digital image information with increasing efficiency and cost effectiveness. Computer-to-Plate (CtP) products then became the objective of lithographic plate manufacturers. Early entrants in this race were electrophotographic plates described above but adapted for laser exposure [311] (producers Agfa-Gevaert, KPG). However because of image quality problems and processing complexity they were soon replaced with plates using the high light-sensitive silver diffusion transfer process (see Section Modern Lithographic Printing Applications) or plates based on photoinitiated polymerisation.

The first platesetters all used visible laser; air cooled Argon-ion (488 nm), Helium-Neon (633 nm), semi-conductor diode (670 nm), and frequency-doubled YAG (532 nm) were important development milestones. In the mid-1990s infra-red laser diodes (830 nm) started a trend for thermally imaged printing plates. Development of thermal platesetters by CREO Inc. and the launch of thermally sensitive printing plates by Kodak (later Kodak Polychrome Graphics) started a counter-revolution to the marketing of visible light CtP products. Based on Novolac resins (phenol or cresol formaldehyde polymers) traditionally used in positive analogue plates, the thermal plates function by having a cationic dye (typically tri-phenyl methane structures) that insolubilises the Novolac resin but which, on thermal (laser) irradiation, causes a structural disruption so making the resin soluble in alkali developer. Thermal CtP plates are characterised by having sharp reproduction and can be handled in normal room lighting conditions.

Further important developments can be expected in the field of thermally exposed CtP whilst visible-light sensitive CtP are moving into a mature phase of development. Thermal exposure also opens up the possibility of designing plates that only require exposure, that is they can be produced without chemical processing (producers: Agfa-Gevaert, KPG, Fuji Photo Film, Lastra).

Imaging of Plates from Digital Data. Steadily increasing data storage capacities permit the fast processing of ever growing quantities of information and text. Not only the composition of text columns, but also page makeup of type and electronically screened four-colour illustrations, can be carried out by computer and checked on the screen. The completely integrated digital system from text preparation to platemaking is described in Section Graphic Arts Photography).

Because of the digital structure of the type and image information in the computer, printing plates can be exposed directly in output scanners without any intermediate imagewise recording of information on a carrier material. These systems are largely independent of an original copy and, consequently, of silver film as well. The steps involved from the information to the plate are presented schematically in Figure 62.

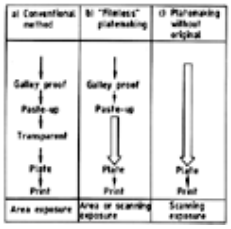

Figure 62. Comparison of different pathways from original to printing plate

3.4.2. *Letterpress and Flexography*

Letterpress is the oldest method of printing. In it, printing areas of the printing plate are raised above nonprinting areas. Ink is applied to the raised areas with the help of ink rollers and then transferred to the printing material. In the past, letterpress printing plates were made of metal (lead, zinc, or magnesium). The appearance of photo setting (i.e., setting information directly on film) necessitated the development of a direct path from film to plate. This led to the emergence of photopolymer relief plates in the 1960s; today, letterpress almost exclusively uses plates of light-sensitive plastics. The polymeric binders used mainly in letterpress printing plates are polyamide or poly(vinyl alcohol).

Flexography is a special form of relief printing (called aniline or rubber printing in the past). In flexography the printing plate is a soft and flexible material. Before 1970 rubber printing forms were used. Production of a rubber printing form involved the following steps. First a hard letterpress printing form is produced, which is used to mold a duroplastic flong. Then the rubbery raw material is filled into the flong and vulcanized to yield the rubber printing form. Today most flexographic applications use light-sensitive photopolymer printing plates, which contain flexible thermoplastic elastomers (e.g., styrene – isoprene – styrene block copolymers) as polymeric binders. These flexographic printing plates can be processed more easily than rubber and show better print results.

Letterpress printing (platen and rotary printing procedures) ranges from job printing (e.g., visiting cards, programs, and business papers) to printing of adhesive labels, forms, tubes, cups, newspapers, and paperback books. The market share of letterpress newspaper printing is decreasing due to the dominance of lithography, whereas the market share or letterpress label printing is increasing. Flexography is especially suited for printing paper, cellophane, aluminum, and plastic films, and is also used to print uneven materials (e.g., corrugated board). Nearly all products printed by flexography are packaging materials.

3.4.2.1. *Structure of Photopolymeric Letterpress and Flexographic Plates*

A solid photopolymer printing plate consists of a light-sensitive photopolymer layer, which forms the relief layer after processing (see Fig. 63). Depending on the application the thickness of this layer varies from 700 to 7000 µm (flexographic printing plates) or 30 to 2000 µm (letterpress printing plates). The very sensitive photopolymer layer is covered by a protective film to prevent soiling. The plates are provided with an additional thin layer (release layer) between the protective film and the photopolymer layer. This release layer is usually made of poly(vinyl alcohol), polyamide, or cellulose derivatives [315], [316]. The function of this layer is to ensure a tack-free surface and to provide a good contact to the film negative during exposure. The adhesive layer mechanically anchors the photopolymer layer to the base material. Steel, aluminum, or polyester are used as base materials. High demands have to be fulfilled by the adhesive layer, because the relief is subjected to considerable mechanical and chemical stress during printing. Solvent-resistant two-component lacquers based on the cross-linking reaction between multifunctional polyols and polyisocyanates are often employed for the adhesive layer [317].

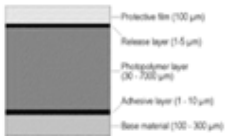

Figure 63. Structure of photopolymer printing plates

3.4.2.2. *Production and Requirements of Photopolymer Letterpress and Flexographic Plates*

Processing of letterpress and flexographic printing plates involves the following steps:

Exposure;
Washout;
Drying;
Posttreatment (if necessary).

Exposure (see Fig. 64). First the protective film is peeled off, then a film negative is placed on the plate. During exposure

with UV-A light (wavelength range 350 – 400 nm, UV tubes or mercury lamps) the polymer layer polymerizes at the areas corresponding to the transparent areas of the film negative. The exposed areas of the photopolymer layer cross-link and become insoluble. The nonexposed areas of the photopolymer layer remain soluble. The exposure is performed under vacuum to ensure a good contact between plate and film negative.

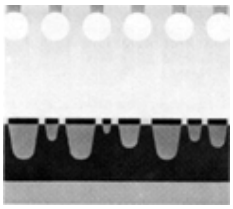

Figure 64. Exposure

Washout (see Fig. 65). After exposure, the soluble parts of the plate are washed out in special devices (spraying or rubbing system). Letterpress plates are washed with alcohol – water mixtures or with pure water. Most flexographic printing plates have to be washed with organic solvents (see Table 2). Before 1990 a mixture of tetrachloroethylene with butanol was used as washout solution. Today less harmful solvents, such as paraffinic and naphthenic hydrocarbons, aromatic hydrocarbons, or alkylesters have been substituted for tetrachloroethylene in most countries [318]. Strong efforts are made to develop flexographic printing plates, which can be washed out with water or aqueous solution (see Section Gravure Printing).

Table 2. Trade names and manufacturers of relief printing plates

	Trade name	Company	Washout
Flexographic printing plates	AFP	Ashai	organic solvent
	nyloflex	BASF	organic solvent
	Cyrel	DuPont	organic solvent
	Flex-Light	Polyfibron	organic solvent
	Flexceed	Supratech	water
	Okha-Flex	Tokyo Okha	organic solvent
	Cosmolight	Toyobo	aqueous solution
Letterpress printing plates	nyloprint	BASF	water or water – alcohol
	Napp plate	Napp Systems	water
	Miraclon	Tokyo Okha	water or water – alcohol
	Rigilon	Tokyo Okha	water
	Torelief	Toray	water
	Printight	Toyobo	water

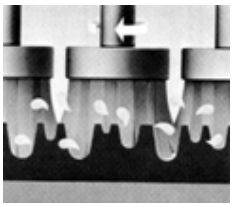

Figure 65. Washout

Drying and Posttreatment (see Fig. 66). Once washed, the plates are dried and subsequently reexposed for complete polymerization. In this way, they attain their final mechanical properties for the printing process. Flexographic plates are sticky after production and must be subjected to an additional posttreatment process. By exposure of the plates with UV-C light (wavelength 250 nm) the surface of the relief is hardened and detackified [319].

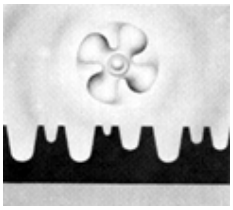

Figure 66. Drying

Both letterpress and flexography have a wide spectrum of applications. The demands made on the plate in different applications are so divers that they cannot be met by a single plate. These requirements depend on the ink, the material to be printed, and the type of press. In flexography printing inks of low viscosity are used. Jobs on paper, napkin, or corrugated

board are performed with water-based inks. Plastics and metal foils are printed with alcohol-based inks or UV inks. The uneveness of the substrate determines the thickness of the flexographic printing plates. Corrugated board is printed with plates with a relief height of about 3.0 mm. Plastics is printed with thin plates with a relief height of about 0.7 mm. The trend is to use plates as thin as possible to reduce thickness tolerances, to increase print quality, and to reduce cost [320]. In letterpress applications UV inks or oil-based inks are used. Due to the higher viscosity of the printing inks the relief heights of letterpress plates are lower (0.3 – 0.7 mm).

3.4.2.3. *Chemistry of Photopolymer Relief Printing Plates*

The principle of all photopolymer relief printing plates is based on the UV-induced photopolymerization of olefinic unsaturated compounds [321], [322]. Other processes used for reprographic products (e.g., photolysis of *o*-quinone diazides or dimerization of cinnamic acid; see Section Lithography (Planographic Printing, Offset)) are of little importance for letterpress because of the low utilizable photochemical quantum efficiency. Relief printing plates are up to 1000-fold thicker than lithographic printing plates. Hence, in keeping with the latest findings, only a chain reaction such as radical polymerization can reach the required low exposure times of 10 – 1000 s.

The development of commercially available photopolymer relief printing plates, based on polymeric binder, vinyl monomer, and photoinitiator was accelerated considerably by the work of PLAMBECK et al. in the mid 1950s [323]. This was based on the work of GATES on photopolymerization of methyl methacrylate through a film negative, which paved the way for photomechanically produced relief printing plates [324].

In the absence of a sensitizer, olefinic compounds absorb light only in the shortwave UV range below 300 nm. However, this radiation cannot be used for cross-linking because light sources with sufficient energy output at this wavelength are not available and common film negative materials are not transparent enough below 300 nm. To initiate photopolymerization in relief printing plates with long-wave UV-A light (wavelength range 350 – 400 nm), the photopolymer formulation must contain substances that are capable to absorb UV-A light and generate polymerization-initiating radicals. Usually compounds are added to the photopolymer formulation that either undergo a homolytic bond cleavage to yield a polymerization-initiating radical pair or form suitable initiating radicals by abstraction of a hydrogen radical from another compound [325].

Examples of hydrogen abstraction initiators are derivatives of anthraquinone, thioxanthone, benzophenone, and acetophenone. Some alpha splitters are derivatives of benzoin, benzil, benzoylphosphines, and benzoylphosphine oxides [326].

Examples of cross-linkable vinyl compounds (polymeric, oligomeric, or monomeric) are:

1. mono- or polyfunctional acrylates and methacrylates [327], [328];
2. mono- or polyfunctional acrylamide or methacrylamide derivatives [329];
3. allyl ethers (e.g., polyurethane oligomers with allylic functionality that cross-link easily with polyfunctional mercaptans [330]);
4. mono- or polyfunctional esters of maleic or fumaric acid [331].

Apart from the initiator and the polymerizable vinyl compound, many relief printing plate formulations contain a polymeric binder, which controls the physical chemical properties, e.g., elasticity, hardness, resistance to solvents. The polymeric binder is, as a rule, the main component of the photopolymer formulation and is the matrix that takes up the generally liquid monomers and makes them manageable. Often the polymer takes part in the UV-induced cross-linking reaction and, together with the monomers, forms a three-dimensional cross-linked network. Some binders in commercially available relief printing plates are:

1. soluble polyamides [329];
2. partly saponified poly(vinyl acetate) [e.g., poly(vinyl alcohol)] [328];
3. polyacrylates, polyesters, or polyurethanes;
4. thermoplastic elastomers, such as styrene – isoprene – styrene rubbers or styrene – butadiene – styrene block copolymers, ethylene – propylene – diene rubbers, or nitrile rubbers [327], [332].

As a rule, polar polymers, such as polyamide or poly(vinyl alcohol), are used in letterpress plates. Because of the high polarity of the polymers the plates can be washed out in polar solvents such as water or alcohol and can be printed with nonpolar printing inks such as UV or oil-based inks.

In contrast, nonpolar polymers such as thermoplastic elastomers are used in flexographic printing plates. Due to the nonpolar nature of these polymers, flexographic printing plates have to be washed out in nonpolar organic solvents such as hydrocarbons or chlorinated hydrocarbons and can be printed with polar printing inks based on water or alcohol.

Because of ecological reasons strong efforts are made to develop flexographic printing plates which, in the unexposed state, can be washed with water and, in the exposed state, do not swell in water or alcohol-based printing inks. Today nearly all manufacturers of flexographic printing plates have developmental plates which can be washed in water or aqueous solution. But only some of these are commercially available (see Table 2). The photopolymer layer of these plates often contains a mixture of polymers of different polarity [333]. Another possibility is to incorporate monomers with different polarity into one polymer [334-336].

Further auxiliary agents that may be added to photopolymer formulations are:

1. Polymerization inhibitors;
2. Plasticizers;
3. Dyes and pigments which improve the relief form and the resolution of the printing relief [337].

3.4.2.4. *Future Developments*

Photopolymeric letterpress and flexographic plates have connected relief printing to modern computer technology (phototypesetting and reproduction technology). The print quality of letterpress and flexography improved substantially by the use of photopolymer plates and reprographic techniques which adjust the film negative to an optimum print result.

The next generation of photopolymer plates will not even need a film. Computer-to-plate (CTP) techniques which are state of the art in lithography will be introduced in letterpress and flexography. To overcome the difficulties arising from the plate thickness of relief printing plates, multilayer plates are being developed. Two filmless exposable flexographic printing plates were launched by BASF and DuPont in 1996. Both plates have a thin black layer of high optical density on top of the photopolymer layer. This black layer is imagewise laser-ablated by exposure with a yttrium – aluminum – garnet (YAG) laser (1064 nm). By this process the film negative is created in situ on the plate. Then the plate is processed like a conventional flexographic printing plate by UV exposure, washout, drying, and posttreatment [338], [339]. The processing of conventional flexographic printing plates is compared with the CTP-process in Figure 67.

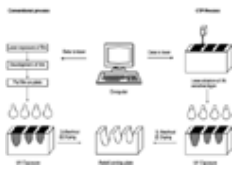

Figure 67. Conventional processing of photopolymer plates and CTP process

Further developments tend strongly toward improved environmental protection and occupational health. Skin-sensitizing monomers will be replaced in photopolymer plates. Further progress is expected with respect to water-washable flexographic printing plates. In flexographic printing water-based inks and UV inks will be substituted for alcohol-based inks. Photopolymer plates with a better resistance against UV inks will be developed. The print quality of letterpress and flexography will increase further.

3.4.3. *Gravure Printing*

Rotogravure is the printing process for highest printing quality. Printing can be performed on nearly any kind of printing material (also on inexpensive paper). Two types of rotogravure exist: *rotary printing* and *sheet-fed printing*.

Rotary printing is the most important process for mass production of publications, mainly magazines and catalogues. The printing forms for gravure printing are cylinders. These cylinders have a steel jacket which, in turn, has either a copper or a plastic coating depending on the process used to make the gravure cells. In the majority of practical applications, a base copper coating is electrolytically applied to the steel jacket. A plastic coating is applied to the steel jacket only when lasers are used in gravure platemaking (see Section Electron-Beam and Laser Gravure). The layer thickness of the base copper is approximately 0.5 – 3 mm, depending on whether the gravure cells are placed directly in the copper base or in a detachable copper skin (=Ballard Shell) on the base copper (etched or engraved). If a Ballard Shell is used, the thickness of the base copper is only 0.5 – 1 mm. After application of the base copper, the cylinder must be turned on a lathe very precisely to the desired diameter. Then the cylinder surface must be polished.

A copper-coated cylinder of this type serves as the carrier for the Ballard Shell used for printing. The thickness of a Ballard Shell ranges from 0.1 to 0.15 mm. This shell should not combine with the copper base but should remain easily detachable. For this reason, a separating layer must be applied between the base copper and the shell. A solution of silver nitrate (or silver cyanide) and calcium cyanide is generally used for this purpose. The advantage of this process is that the cylinder always has a uniform diameter due to the application of the copper.

The next step involves placing the printing elements — the gravure cells — in the gravure cylinder (raw cylinder) prepared as described above. Until the early 1960s, only *chemical methods* were used for this purpose. These methods have been improved and refined continuously, especially by using electronic controlling and standardizing devices. This type of gravure platemaking is known as the conventional process to distinguish it from other processes.

With increased industrialization of many process steps in the printing industry, especially in gravure printing, *electromechanical* gravure is becoming better established in gravure platemaking. This process represents application of the technology developed by Dr.-Ing. Rudolf Hell (Kiel) for letterpress platemaking. The appliances required for this purpose were called *Klischographs*. In letterpress printing, they were used to remove nonprinting areas from the plate material. In gravure platemaking, exactly the opposite occurs: the printing elements or gravure cells are engraved into the plate material. The appliances required for gravure platemaking are called *Helio-Klischographs*, analogous to the trade name Klischograph, and are used worldwide for publication rotogravure (magazines and catalogues).

3.4.3.1. *Conventional Method of Gravure Platemaking*

Because of the present status of gravure platemaking and the numerous variations that exist, only the fundamentals of the conventional process are discussed here. The problems connected with this method are easily discernible and have determined the trend toward methods that are more easily controllable and reproducible. Specific details of the conventional method are described in the literature [340-342].

For the production of printing cylinders, the information to be printed (text or pictures) must be available as complete pages, in the form of a page makeup or stored data. In the conventional process, the text is present as a phototypesetting film (positive, black images on a light background), and the picture material is available as a continuous-tone diapositive. For multicolor gravure, positive color separation diapositives are prepared with the printing colors yellow, magenta, cyan, and black, corresponding to the page layout. These originals cannot be transferred directly to the raw cylinder. An intermediate carrier, carbon tissue, is required. This paper is available in rolls and is coated with gelatin, which gives an etchable washout

relief. Shortly before use, the gelatin coating is made sensitive to light with a chromium salt solution. Use of carbon tissue requires considerable care. Workrooms must be air-conditioned and dust free. The sheets of paper can be cut to the required size only after they have become acclimatized. Prepared sheets are usually dried in an oven and then stored in a deep freeze until used.

In gravure printing the range of tone values is realized by varing the volume of the printing elements (cells) on the printing form (cylinder). As in other traditional printing processes (e.g., letterpress printing, lithography) the printing ink must be transferred to, e.g., paper in the form of very small portions (80 – 200 lines per inch). These portions are a result of the so-called screening process. Each original (image), may it be text or picture, must be screened. This process takes place in the pre-press department for making the films that are to be copied on, e.g., pigment paper (carbon tissue) [342], [343]. This results in the formation of area elements of uniform size. Finally, these areas must be etched to a depth corresponding in size to the required tone value. These are the *gravure cells*, and the entire net is called the *gravure screen* [342]. It serves two purposes simultaneously. On the one hand, the quantity of ink in each individual cell is defined by the volume of that cell.

On the other hand, the doctor blade, which wipes the ink off the surface of the cylinder, before printing, is provided with a surface of uniform height to ride on.

To produce the cells and partitions, a screen is copied on the carbon tissue. The *variable-depth process* normally uses a cross screen with a 45° rotation with respect to the side lines. It consists of a transparent net and squares (Fig. 68). The ratio of partition width to length of a square (i.e., screen – partition ratio) is usually 1 : 3. The screen ruling (i.e., the number of lines per centimeter) is normally 70 for printing publications. Larger screens are usually used to printing packaging and textiles; finer screens are used for stamps.

Figure 68. Gravure screen for production of cells and partitions

Carbon tissue is sensitive to light ca. 300 – 450 nm in wavelength. However, the maximum sensitivity is at 360 – 380 nm. Consequently, only light sources emitting predominantly shortwave radiation are suitable for gravure. To obtain an especially sharp halftone copy of the screen partitions, a point source of light projected through a lens is employed. The areas on the carbon tissue that are exposed to light (the areas under the transparent net) lose their water solubility. The chromium – gelatin layer is thus hardened to varying degrees depending on the intensity of light.

After the screen has been copied on the carbon paper, the picture (i.e., the fully stripped page) is copied in a second step. Depending on the tone value (degree of density) of the different image areas, light penetrates more or less into the chromium – gelatin layer. Thus, the curing of the gelatin layer is directly proportional to the density of the original to be copied. The gelatin remains water-soluble in those areas where no light passes through. The screen copied in the previous step is not affected by picture copying. It is maintained in lines on all the printing areas, both in the text and in the smooth dark areas as well.

The finished sheets of carbon tissue are then aligned on the copper cylinder with a special transfer machine. The highly polished copper cylinder must be degreased carefully. The copied sheets of carbon tissue are briefly wetted and then pressed onto the cylinder. This must be a highly precise step because carbon tissue is not dimensionally stable.

Despite standardization and very careful operating methods, inaccuracies in positioning cannot always be avoided. As a result, other materials have been developed as substitutes for carbon tissue. These materials use a stabilizing foil or film instead of a paper base but undergo similar processing steps.

The copies transferred onto the cylinder are developed with warm water (40 °C) in a *developing machine*. After a few minutes, the paper backing is wet enough to be removed carefully from the rotating cylinder. Development continues (ca. 10 – 15 min) until the soluble gelatin is completely washed out, leaving the exposed (cured) gelatin relief clearly visible. The completely blank areas give the deepest wells after etching, and a layer of hardened gelatin of varying thickness remains on the other wells. The cylinder is then cooled to room temperature and dried with alcohol. These steps are carried out in fully automatic and programmable developing machines.

The *etching process* is the most difficult step in gravure platemaking. All areas that are not to be etched are covered with an acid-resistant asphalt lacquer before etching. The etching process is now largely automated in specially constructed etching machines. Solutions of iron(III) chlorid, commonly called perchloride of iron, of varying densities are used for etching. The heavier solutions are able to soften and penetrate the thin gelatin layers (equivalent to the shaded areas in the film positive). Lighter solutions are used progressively as it becomes necessary to penetrate the thicker gelatin layers (equivalent to the middle tone to highlight range) [340-342]. Etching is done in several stages, with four to six solutions at different Baumé density levels (Bé). The most concentrated solution is applied first, and less dense solutions are used stepwise. A variation of the etching process uses one etching bath only with an etching solution of a single Baumé level. The numerous problems associated with manual and mechanical etching are described [340], [341].

3.4.3.2. *Electromechanical Gravure Platemaking*

Modern electronics, together with precision mechanics, guarantees the highest standards of operational safety and reproducibility. Gravure cylinders with accurately programmed printing characteristics can be produced with the Helio-klischographs, which were introduced to the public in 1962. This process is generally called “cylinder gravure.”

An important advantage of cylinder gravure is that it involves only one operation which starts with the original (stripped page)

and leads directly to the ready-for-press cylinder. Furthermore, it eliminates all the difficult-to-control problems that can occur during chemical processes, during the use of intermediate carriers (e.g., carbon paper), and during conventional etching. In addition, all problems associated with the halftone copy are eliminated because the halftone process is carried out electronically.

Cylinder gravure can be considered a procedure for the transfer of information. The originals to be engraved are mounted in a ribbon-by-ribbon manner as black and white originals, right side up, on the *image cylinder*. Each line of the original is scanned with a scanning system. This is the same process used for the transfer of press photos or for electronic color correction in color scanners. The optical signals scanned from the picture cylinder are converted to electrical signals, which are fed into a computer. The desired gravure gradation is adjusted in this data-processing unit, which can be switched to scan either positive or negative originals. Electrical signals obtained this way are then conveyed to the engraving heads, which engrave the *printing cylinder*. Based on the principle of transmission of information, the unit with the image cylinder can be located at any distance from the engraving unit (long-range transmission). Electrical signals from an image cylinder can also be used to control several engraving units at the same time. In general, however, the scanning unit and the engraving unit have a common machine bed (Fig. 69).

Figure 69. Helio-Klischograph type K 201 (Dr.-Ing. R. Hell, Kiel). The scanning cylinder is in the middle of the figure; the gravure cylinder to be engraved on the right.

Electrical signals transmitted to the engraving unit operate an engraving system for each string. These engraving systems are moved with great precision synchronously with the scanning systems. Each engraving system has a sharply cut diamond stylus that moves up and down ca. 4000 times each second at right angles to the surface of the cylinder. Every time the stylus penetrates the cylinder, a pyramid-shaped cell is engraved into the surface. The depth of penetration of the stylus and, thus, the volume of the cell correspond to a particular tone value on the scanning side. In the surface of the cylinder, the cells have a rhombic shape and are arranged in the gravure tracks so that the cells of neighboring tracks are staggered [340], [341].

An engraving efficiency of 0.27 m^2/h is attained with a cylinder of 1000-mm circumference and a screen of 70 lines per centimeter. It is customary to use as many scanning or engraving heads as ribbons are provided for printing. For this reason, the engraving of a cylinder requires as much time as the engraving of a string (i.e., one hour) independent of the number of strings.

Because cylinder gravure basically processes electrical signals, this method can be used for purposes that are difficult or impossible to achieve with conventional gravure (e.g., repetition in the circumferential direction, repetition in the direction of the cylinder axis, and postgravure [340], [344]).

3.4.3.3. *Electron-Beam and Laser Gravure*

Since the beginning of the 1960s, repeated attempts have been made to develop a method that uses an electron beam for cylinder gravure. This idea was initiated by the European Rotogravure Association (ERA). Development of an electron-beam gravure (EBG) machine was undertaken by the company Dr.-Ing. Rudolf Hell (Kiel) [345] and was continued after the company changed its name to Linotype-Hell following a takeover by Linotype. However, the project was discontinued in March 1993 after detailed cost analyses showed that the likely running costs of an EBG machine would have made it uneconomic to operate.

In contrast to EBG (electron-beam gravure), the application of lasers does not require the creation of a vacuum. Research on the use of a laser engraving machine was begun in the mid 1970s by Crosfield Electronics (London). This work finally resulted in 1982 – 1983 in an entire system for the production of gravure plates. The Laser Gravure System 700 was presented to the public in 1983 in a form that had been tested at the same time by Odham-Sun Printers Ltd. (Watford, England) under practical gravure printing conditions (Fig. 70) [346]. However, this development also ultimately proved not to be economically viable and was itself abandoned.

Figure 70. Engraving machine of the system Lasergravure 700 (Crosfield Electronics, United Kingdom)

The laser gravure process differs considerably from all other gravure platemaking processes. Because laser beams are strongly reflected from metallic surfaces. The problems which this causes were solved after many years of development work by the company MDC Max Dätwyler AG in Bleienbach, Switzerland, whose decisive breakthrough came with the development of a special alloy which could be applied to the gravure cylinder. The “Laserstar” laser gravure system, based on this idea, gives an engraving frequency of 35 kHz with up to 240 lines per centimeter. Printing results so far obtained with these cylinders confirm the precision and high state of technical development of the “Laserstar” system. The first practical printing press of this type is expected to be operating in 1997.

Development work is aimed at increasing the engraving frequency to 70 – 140 kHz and at producing a second generation of alloys for the process.

Doubling of the engraving frequency incurs a cost increase of ca. 25 %.

3.4.4. *Screen Printing [347], [348]*

Screen printing was the successor of stencil printing, a method used in China and Japan in very early times. This printing process was reportedly invented by MIYASAKI (1664 – 1736) in Japan. Not only is screen printing suited to flat materials such as stickers, road signs, large advertisements, pictorial depictions on automobiles, and decals, but also items such as glass, tile, T-shirts, plastic bottles, and printed circuits can be printed on by this process.

The stencil in screen printing consists of a screen mounted on a frame and the parts of the stencil that are fixed to the screen. Printing occurs under the screen by forcing the ink through the screen openings with a squeegee onto the printing material. Not only the frame, but also the screen material, fabric thickness, mesh size, yarn diameter, type of weave, and structure of the stencil can vary from job to job. Thus, screens are not prefabricated or presensitized; rather, the individual materials required for stencil making are available commercially and the final stencil must be made at the printer's from these materials.

3.4.4.1. *The Screen*

Materials used for screens are textile fabrics such as silk, polyamide, and polyester fibers, but metals such as steel, nickel, and phosphor bronze may also be employed. The screens are pretreated before coating to improve adhesion of the stencil and printing quality. The stencil is removed from the screen after use so that the screen can be reused for other jobs.

3.4.4.2. *The Stencil*

Like other printing processes, the printing form in screen printing can be produced manually or photomechanically.

Manual Methods. For instance, a stencil can be made by sticking pieces of film or firm paper on the nonimage areas of the screen. In another method, a suitably coated film or paper is used to produce a stencil. The image is engraved into the coating, the coating is removed from the image areas, and the remaining layer is transferred to the screen with heat or solvent.

Still another method employs a lacquer, which is applied to the nonimage areas on the screen, or a covering agent, which is applied to the entire surface over a resistant image that is subsequently washed out with a suitable solvent to produce the printing screen.

Photomechanical Platemaking. A light-sensitive coating is employed in both photomechanical methods: the direct and the indirect method. In the first case, the coating is applied to the screen before exposure and development; in the other, it is exposed on a carrier, developed, and then transferred to the screen.

In the *direct method*, the screen mounted on the frame is coated with a light-sensitive solution called an emulsion. The carefully dried coating can then be exposed to UV light through a positive original in a copying frame. After exposure the soluble areas are removed by a water rinse. Various formulations are being used as the emulsion:

1. Colloid – dichromate solutions (see Section Photoinsolubilizing (Photocuring) Systems) are highly sensitive to light, can be developed easily with water, are very resistant to solvent-containing ink, and allow easy regeneration of the screen. Substances such as gelatin are used as colloids. The emulsion is made at the print shop from its individual components because the light-sensitive mixture is stable for only a limited time at room temperature.
2. Colloid – diazonium salt solutions (see Section Photoinsolubilizing (Photocuring) Systems) have been available for some time as substitutes for chromium(VI)-containing emulsions. Poly(vinyl alcohol) or poly(vinyl alcohol) – poly(vinyl acetate) is used as the colloid. The processing procedure is comparable to that of colloid – dichromate solutions. Colloid – diazonium emulsions are sometimes said to have a lower sensitivity to light than the corresponding dichromate solutions. The reason for this may be that the spectral sensitivity of dichromate coatings exceeds 500 nm, whereas that of aromatic diazonium compounds is limited to the range 360 – 420 nm.
3. In more recent times, photopolymer formulations have also been offered for screen printing [349]. These fomulations may be epoxy resins with unsaturated side groups or acetals of 4-(4′-ethenylpyridinium)benzaldehyde (4-stilbazolium aldehyde) with poly(vinyl alcohol), which are susceptible to cross-linking dimerization. The water solubility of the emulsion is guaranteed by the very polar structure of the pyridinium units. In some preparations, these polymers are used in combination with acrylate monomers.

In the *indirect method*, a binder film, which may be either sensitive or insensitive to light, is used on a carrier as the starting material for stencil making. In the print shop the insensitive film is wetted with a solution of a light-sensitive agent (sensitized) and dried before exposure. From this point on, their processing is basically comparable: exposure, development, transfer of the stencil to the screen, and removal of the film carrier. The advantages of the indirect method are accurate reproduction, even with mesh-crossing contours, as well as constant layer thickness which, in turn, is advantageous for print quality. The precoated materials used consist of poly(vinyl alcohol) [or poly(vinyl alcohol) with water-soluble diazonium salts] and are dyed.

The *combination stencil* is a photo stencil which is produced commercially as a light-sensitive film on a carrier. This film is first transferred to the fabric with water or emulsion. After the emulsion has been dried, the carrier is removed and the light-sensitive stencil is exposed. Development of the exposed stencil is carried out with a jet of tap water, as in the direct method.

Some advantages of the combination stencil are accurate positioning of the image because of direct exposure and sharp reproduction favored by the smooth surface of the screen. The light-sensitive formulations that can be used for these coatings are colloids which are sensitized with diazonium salts or with styrylpyridinium salts.

Producers of light-sensitive products for making stencils are Hoechst, Kissel und Wolf, Ulano, Autotype, Sati, Sericol,

NazDar, Aicello, and Murakami.

[Top of Page]

4. Imaging for Electronics

Hartmut Steppan, Donald C. Mammato, Thomas Stoudt, Michael C. P. Watts

This chapter occupies a special position because “imaging” in the classical sense is discussed here only in exceptional cases. Here, imaging refers to the selective coating of substrate surfaces to protect coated areas from the effects of etching reagents, metallization baths, etc. Because of the controlling function of selective coating, this technique allows the overall, undifferentiated action of working media (chemical tools) on the entire workpiece, which is a prerequisite for mass production.

Photoresists play a very important part in the production of selective coatings (Section Photoresists). Other important aspects of imaging for electronics are the production of *printed circuits* (see Section Printed Circuits (Printed Circuit Boards)) and the production of *integrated circuits* (see Section Microelectronic Devices).

4.1. Photoresists

The term *resist* is derived from *resistant*. It refers to a material that is used to coat areas of a substrate surface to protect them from the reagent in question during chemical or physical treatment (e.g., acidic or alkaline etching, etching with plasma, galvanic metallization, diffusion of foreign substances, to dope the substrate etc.).

Generally, two methods are used for the production of selective resist coatings:

1. selective application only to those areas of the substrate surface that are to be protected: e.g., application by screen printing (see Section Screen Printing) [350], magnetography [351], or other printing processes; and
2. covering the entire substrate surface with a layer of resist, which is subsequently removed from the areas to be processed.

The latter includes the photoresist technique [352-354], which plays an important part in the production of certain printing plates (multi-metal plates, zinc and magnesium stereotype plates), but is of paramount importance in electronics for the production of printed circuit boards (PCBs) and integrated circuits (ICs).

Photoresists are defined as radiation-sensitive or, in a narrower sense, light-sensitive materials whose solubility is altered when exposed to radiation. The selective resist coating is produced by exposing the photoresist layer through a mask. In this way, only certain areas of the resist layer are exposed to radiation. Subsequent development with a suitable liquid, the developer, results in removal of the more soluble coating areas, leaving these areas of the substrate uncovered.

In the field of integrated circuits, however, tendency to avoid wet processes is growing, not only in etching and stripping, but also in developing [355-360].

Classification of Photoresists. A number of different photoresists are available on the market today. The characteristic features are *mode of operation* (negative, positive), *physical state* (liquid, solid), *differentiating radiation* (e.g., short- and long-wave UV, X-rays, or electron beams), *development* (with organic solvents, aqueous solutions, etc.), *application* (e.g, resists for the production of solder masks), *number of components* (e.g., one-, two-, or multicomponent systems), and *chemical composition* [352], [353], [361-364].

If the solubility of a photoresist *increases* during exposure to radiation, it is termed a *positive* resist. If the solubility *decreases*, it is a *negative* resist (see Fig. 71). Whether a photoresist is positive or negative depends not only on its composition, but also on the method of processing.

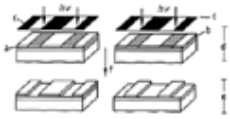

Figure 71. A) Application of a resist mask using a photoresist;

B) Development by dissolving the more soluble areas of the layer a) Positive photoresist; b) Negative photoresist; c) Exposure mask; d) Substrate covered with photoresist; e) Substrate with selective resist coating

Liquid resists are available in two versions:

1. solutions of photoresist components in solvents in which the wet film dries on evaporation of the solvent, to give a uniformly smooth layer; and
2. solvent-free liquid systems (without exception negative systems) with reactive diluents, e.g., liquid acrylates or methacrylates, which polymerize or cross-link during radiation and thus become a component of the solid resist coating [365].

Dry film resists are solid photoresists (thickness ca. 15 – 100 µm) coated and dried on temporary carriers (usually polyester) [366]. In general, the layers are also covered with a protective foil (polyolefin). After removal of the protective foil, these photoresist materials adhere to (largely flat) substrates when they are heated under pressure.

For many years, a spectral light sensitivity for photoresists in the range 350 – 450 nm was customary and adequate. Today, photoresists sensitive to more energetic radiation (e.g., UV light of 249 or 313 nm, as well as electron beam, ion beam and X-rays) [361-364], [367] are becoming increasingly important for microlithography because they allow higher resolution. Despite the different nature of these rays, the term “photoresist” is retained. On the other hand, photoresists sensitive to lasers of

488 nm are now fairly important in the field of printed circuit boards.

In general, the photochemistry of photoresists sensitive to electromagnetic radiation of wavelengths >300 nm is quite comprehensible. Selective reactions cause predictable changes in the solubility of the resist material. However, more energetic radiation excites not only conjugated bond systems but also isolated double bonds, non-bonding electron pairs, and finally, σ-bonds. Ionizing radiation and electron beams, ion beams or X-rays produce very reactive intermediates, which in turn take part in successive nonselective reactions. Fragmentation and depolymerization often occur simultaneously with grafting and cross-linking processes, and can be influenced by adjusting the radiation dose.

4.1.1. *Industrial Applications of Photoresists*

A series of rules, not entirely without exceptions, applies to the use of photoresists; these are described briefly below. In general, thin photoresist layers (Langmuir – Blodgett layers represent the extreme case [368]) resolve better than thick layers; positive-acting layers are better than negative layers; and layers that can be developed in aqueous alkaline solutions (less swelling) resolve better than those developable with organic solvents.

A list of important properties of industrial photoresists follows:

Solubility
Purity (particle-free)
Sensitivity to differentiating radiation
Resolving power and accuracy of the structure transfer
Processing properties (development, stripping, etc.)
Processing latitude
Adhesion to substrate
Resistance to etching
Resistance to electroplating
Thermal stability of the resist mask
Product constancy from batch to batch
Shelf life
Toxicological safety

Photoresists are applied in the production or implementation of certain products and processes. A list of these, along with some photoresist characteristics and their importance in individual cases, appears in Table 3. No single photoresist can fulfill all these requirements which, in fact, must be extended for special techniques. A large number of photoresists are necessary, each one functioning optimally in a particular area only.

Table 3. Application and importance of some properties of photoresists

Property	Product or process *				
	Printed circuit boards [352], [353], [369]	**Photochemical machining [370]**	**Electroforming [371], [372]**	**Integrated circuits [373][374] [375]**	**Micromechanics [376]**
Resolving power	(+)	+	(+)	+	+
Resistance to etching	+	+	–	+	+
Resistance to electroplating	+	–	+	(+)	+
Heat resistance of resist mask	+ **	–	–	+	–

* symbols: + important; (+) less important; – unimportant.
** only for solder masks.

4.1.2. *Function and Chemistry of Photoresists*

The photoresist technique dates back to the early nineteenth century [352], [377]. Until now, an exceptionally large number of chemical systems and compositions have been tested for their suitability as photoresists for many different applications [352], [353], [361], [362], [378-381]. Many of these systems have been used in practice, but only a small selection is presented here.

4.1.2.1. *Positive Photoresists*

Diazonaphthalinone Systems. All positive photoresists used on a large scale in industry for the production of integrated circuits are *liquid* systems. These systems are based on the light-induced Wolff rearrangement [382-384] of diazonaphthalinone derivatives (2-diazo-1-naphthalinones) (**1**) which give 3-indenecarboxylic acid in the last step.

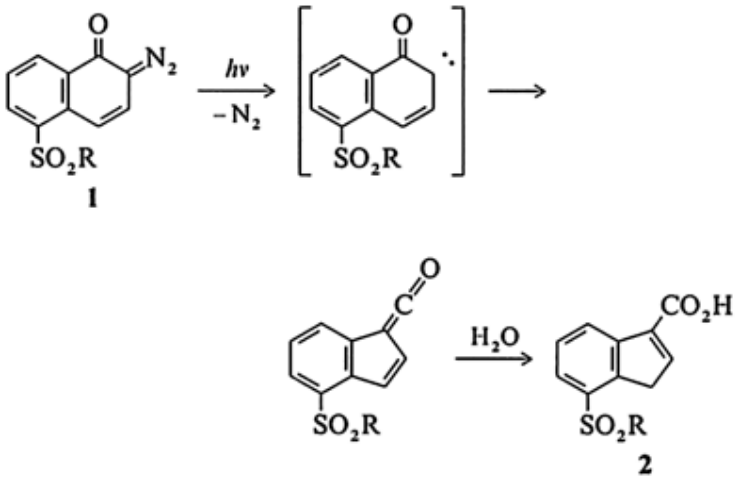

Apart from compounds of type **1**, derivatives of diazonaphthalinone-4-sulfonic acid are also used for positive photoresists (R represents primarily aryloxy groups, but occasionally also substituted amine residues).

The differentiation principle is also clearly demonstrated: namely, the conversion of oleophilic compound **1**, which is insoluble in aqueous systems, to alkali-soluble compound **2**. However, the situation is more complicated in the case of industrial resists because they contain a novolac binder (**3**) (a non-self-curing phenolic resin) in addition to the light-sensitive component (**1**) [353].

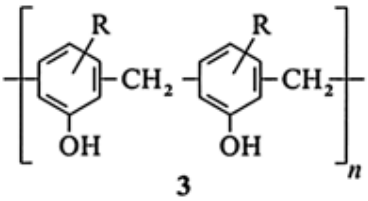

The fact that these resist masks do not swell during development in buffered aqueous alkali-metal hydroxide or tetramethylammonium hydroxide solutions contributes to their high resolution potential. They are used in the long- and middle-wavelength UV range. Their sensitivity to more energetic radiation is generally too low.

Systems That Are Fragmented or Depolymerized by Radiation. High molecular mass poly(methyl methacrylate) (**4**; see Fig. 72) is fragmented by energy-rich radiation (shortwavelength UV light of 230 – 280 nm; electron beams, ion beams, and X-rays) to give fragments that are more readily soluble (**5**) [385]. Mixtures of alcohols and ketones are used as developer. This high-resolution positive resist is used primarily as an *electron beam resist* and has low radiation sensitivity.

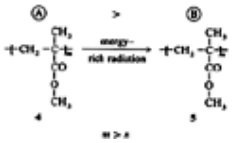

Figure 72. Fragmentation of high-molecular mass poly(methyl methacrylate) by energy-rich radiation

A) High-molecular, slightly soluble; B) Low-molecular, readily soluble

A more sensitive electron beam resist is polybutene sulfone (**6**). This compound can be partially depolymerized to its volatile comonomers by electron beams. Here, too, mixtures of alcohols and ketones are used as developer.

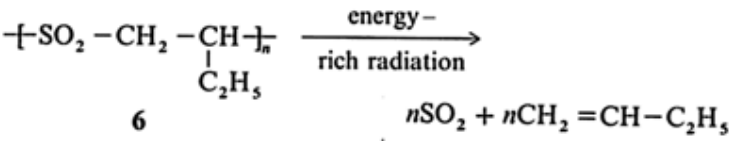

Both basic types of resists have been modified in many ways [381], [386].

Photocatalytic Systems. As a rule, these systems contain only small amounts of a compound that produces a catalyst (usually an acid) on exposure to radiation. This, in turn, catalyzes the conversion of other components of the layer that are not affected by radiation, causing an increase in the solubility of the layer.

Examples of such systems are

1. the radiation-induced, acid-catalyzed hydrolysis of orthocarboxylic acid esters (**7**) in combination with novolacs, to give alcohols and carboxylic acid esters [353] (Fig. 73); and
2. the radiation-induced, acid-catalyzed fragmentation of poly(*p-tert*-butyloxycarbonyloxystyrene) (**8**), which is insoluble in aqueous alkali, to give alkali-soluble poly(vinyl phenol) (**9**), as well as isobutene and carbon dioxide (see Fig. 74) , [388].

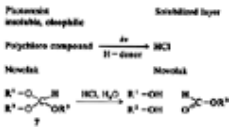

Figure 73. Photocatalytic system, photolytic acid donor novolac orthoester

R^1, R^2, R^3 = alkyl groups

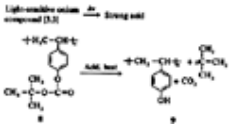

Figure 74. Photocatalytic system, photolytic acid donor poly(*p-tert*-butyloxycarbonyloxystyrene) [352]

Photoresists for all common types of radiation can be formulated by suitable selection of compounds that yield acids on exposure to this type radiation. Thick layers are possible.

4.1.2.2. *Negative Photoresists*

Negative-working photoresists quantitatively represent the largest part of all photoresists used commercially.

Azide Systems. The negative system most often used in the production of integrated circuits consists of partially cyclized polyisoprene (**10**) and an organic azide (e.g., a bisazide such as **11**; see Fig. 75) as the light-sensitive component. On exposure to radiation, nitrogen is eliminated from the azide groups to yield highly reactive nitrenes, which in turn react with the cyclorubber and reduce its solubility [353]. The resist mask swells considerably during development with aromatic hydrocarbons.

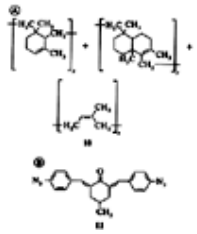

Figure 75. Azide-based photoresist

A) Partially cyclized polyisoprene (important structural elements)

B) Organic azide sensitive to radiation

Negative photoresists based on poly(vinyl phenol) – organic azide, which can be developed with aqueous alkaline solutions, react more favorably in this respect [389]. Azide-based photoresist systems are applied preferentially in the short- to long-wave UV range.

Photodimerization Systems [390]. The oldest commercially employed negative photoresists based on (sensitized) polymeric cinnamic acid esters [391], e.g., poly(vinyl cinnamate) [392], belong to this group. On exposure to light, the (sensitized) photodimerization of neighboring cinnamate groups to cyclobutane structures results in the reduction of solubility [393].

Another system used is photodimerization of chalcone (phenyl styryl ketone) groups (e.g., in epoxy resins based on 4,4′-dihydroxychalcones). Once exposed and developed, systems of this type permit the production of a resist mask that can be thermally cured (solder masks) [394].

The spectral sensitivity to light of the system (a part of which is sensitized) lies in the short- to long-wave UV range. Most of these systems must be developed with organic solvents.

Photoinitiated Systems. Analogous to photocatalytic systems, photoinitiated systems contain only small amounts of a photoinitiator which is converted to an active species on exposure to radiation. This species, in turn, initiates a chain reaction that results in a reduction of solubility (application in thick layers is possible).

Photoinitiated Polymerization. A number of review articles describing the principles and versatile commercial aspects of these systems are available [354], [395-397]. Only radical polymerization has attained real importance. Negative dry film resists, which are generally formulated as described below, represent the most important application.

Main components

1. Photoinitiator system that yields radicals on irradiation [397]
2. Radically polymerizable monomers (e.g., esters and amides of acrylic and methacrylic acids)
3. Polymeric binders (e.g., polymethacrylates)

Modifying components (among others)

1. Thermal polymerization inhibitors
2. Adhesives
3. Plasticizers
4. Dyes

As far as quantity is concerned, components 2 and 3 account for the largest part of the photoresist composition. Depending

on the polymeric binder, either an organic solvent or aqueous alkali is required for development. In fact, the latter is becoming increasingly important. Examples of layer formulations are given in [398-400]. Photopolymerizable polyamide – carboxylic acids are used in the production of high-temperature-resistant masks for the manufacture of integrated circuits [401].

Thiol – Ene Systems. Thiols add to alkenes in photoinitiated radical chain reactions to give thioethers [402]. This principle is exploited in solvent-free liquid systems [365].

Other Systems. Copolymers of the glycidyl ester of methacrylic acid and ethyl acrylate are of some importance as negative electron beam resists.

4.1.3. *Testing Methods*

Apart from controlling purity and identity by standard chemical and physical techniques, the testing of photoresists includes a series of function tests, e.g., determination of light sensitivity, resist adhesion, coating properties, and particle content. The methods used are naturally more numerous and extensive for photoresists intended for microapplication [403] than for those used to produce relatively coarser structures [404].

4.1.4. *Economic Aspects and Suppliers of Photoresists*

Liquid photoresists are applied mainly in the field of microlithography. However, liquid systems are again gaining importance in the printed circuit sector because of economic factors and the higher resolution attainable. An assessment of market development is given below [405].

	World market, ($×10^6)	
	1984	1989
Liquid resists (microapplications)		
Negative	73	195
Positive	73	265
Dry film resists		
Negative	325	875

Some suppliers of liquid and dry films resists are listed in Tables 4 and 5, respectively.

The following companies have a significant share of the market: Hoechst, Shipley, and Tokyo Ohka in the liquid resist (positive) sector; DuPont and Dynachem in the dry film resist sector.

Table 4. Suppliers of dry film resists

Company	Trade name	Mode of operation	
		Positive	Negative
Du Pont	Riston		x
Morton Thiocol (Dynachem)	Laminar		x
Hercules	Aquamer		x
Hitachi Chemical	Photec		x
Hoechst	Ozatec	x	x

Table 5. Suppliers of liquid resists for microapplications

Company	Trade Name	Mode of operation	
		Positive	Negative
Hoechst	AZ	x	
Olin Hunt Specialty Prod.	Waycoat	x	x
Shipley	Microposit	x	

Tokyo Ohka Kogyo	OFPR, TSMR, etc.	x	x
Union Carbide Corp. (KTI)		x	x

4.1.5. *Occupational Health and Environmental Protection*

Safety regulations and instructions for use available from manufacturers must be followed carefully when handling photoresists. Some comments on the *physiological effects* and safe handling of a series of components of UV-hardenable systems are given in [406] and [407].

Careful attention must be paid to avoid possible hazards from light sources and UV radiation [408].

Injunctions provided by law on the limitation of solvent release into the environment and the prevention of water pollution must be met when handling photoresists (coating with liquid resists; wash-out development of imagewise exposed resist layers with organic solvents or aqueous solutions). Advanced techniques are available for the removal of solvents from outlet air (e.g., adsorption by active charcoal, followed by recovery) and the separation of dissolved photoresists from aqueous developer or stripper [409-412].

4.2. Printed Circuits (Printed Circuit Boards)

The term *printed circuits* implies that an electrical circuit is produced by *printing*. In a strict sense, only the thick-film circuit printed with the aid of the screen printing technique satisfies this description. In this case, conductors, resistors, dielectrics, etc., are produced by imprinting: e.g., conductivity pastes, resistance pastes, and dielectric pastes on ceramic substrates (usually aluminum oxide) and subsequent burning-in [413].

Today, the term printed circuit includes all variations of the printed circuit board. The printed circuit board can be made from a rigid or a flexible insulant carrier. Wiring paths are arranged on one or both sides of the board, with conducting connections between the two levels if necessary.

To achieve the high wiring path density demanded by today's sophisticated applications and to prevent exceedingly fine conducting tracks, these can also be distributed among several parallel levels connected to each other through the insulating material (*multilayer technique*) [414]. The printed circuit board is the base for the assembly of functional components; it contains all the conductive connections required among components. In case of high circuit density, wiring paths are usually covered with a solder mask, which leaves only the solder pads exposed. The printed circuit board frequently carries information for the assembly as well (lettering print).

4.2.1. *Methods of Producing Printed Circuit Boards*

The many different and versatile techniques available for the production of printed circuit boards are described in detail in [369], [415-420].

The actual production of wiring paths and plated-through holes can be achieved by using real printing processes (screen printing, magnetography, and others) or photoresist techniques, also called photoprinting. In each case, the substrate (e.g., an insulating plate covered with a copper sheet) is first coated with a *temporary* selective resist coating. Wiring paths are then produced by etching (substractive step), electroplating, or chemical metal deposition (additive step), possibly in combination with etching. After this step, the resist coating is removed (stripped).

A special technique (fully additive process) for making wiring paths is available, which does not require a selective resist coating [421], [422].

Solder masks can also be applied via both screen printing and photoprinting. These masks usually remain as the *permanent* coatings part of the printed circuit board.

4.2.2. *Masking Techniques*

4.2.2.1. *Exposure Masks [423], [424]*

Production of a printed circuit board begins with a layout, which takes into account the requirements of the wiring diagram. In fact, complicated printed circuit boards require the use of CAD (computer-assisted design) techniques to draw up the layout. The layout is then transferred to a silver film of the required size (polyester or glass carrier) by using a computer-controlled plotter (photoplotter or laser plotter). The resulting film is called an exposure mask (photomask). An exposure mask of this type is required for each conductive level and for each solder mask. Exposure masks are used both in the production of screen printing forms (see Section Screen Printing) and in the exposure of photoresist layers in photoprinting.

4.2.2.2. *Screen Printing*

The screen printing process is the oldest technique for the application of resist coatings stable to etching, electroplating, and soldering, as well as lettering in the production of printed circuit boards. The large number of special screen inks can be simplified as follows:

	Drying or hardening
One-component systems	physical

	oxidative
	by heating
	by UV light
Two-component systems	by chemical reactions (reactive hardening)

Two-component systems (e.g., mixtures of epoxide resin and hardening agent) and UV-hardening systems, in particular, fulfill the requirements for a modern system because they yield resist masks of higher quality (resistance to chemicals, physical properties).

Monofil (if necessary, yellow colored) polyester and steel screens are used preferentially in the production of screen printing forms (see also Section Screen Printing) for circuit printing. Factors responsible for achieving the highest possible resolution are screen selection (number of filaments per centimeter and filament thickness), sharpness of edges of the stencil as well as its size, resistance to deformation, dimensional stability, properties of the printing ink, and finally, execution of the printing process [425].

As a rule, screen printing is used to make printed circuit boards on which the pattern width of the printing is at least 0.15 mm. However, it is applied mainly in the production of much coarser structures. For more detailed information, see [426-429].

4.2.2.3. *Photoprinting*

Photoprinting gives considerably better resolution than screen printing and is now used widely in the production of printed circuit boards. This masking technique makes use of photoresists (see Section Photoresists) to produce coatings that are resistant to etching, plating baths, and soldering.

The entire surface of the substrate is coated with photoresist (dipping, curtain coating, roller coating, electrostatic spraying of liquid resists, and lamination of dry film resists) and, in the case of solvent-containing photoresists, dried. Exposure is achieved through an exposure mask, generating differences in the solubility of the coating. Proximity exposure is used in the case of solvent-free liquid systems [365]. Also, laser exposure without the use of an exposure mask is being discussed at present [422], [430].

Once exposed, the areas that have become soluble (positive process) or that have remained soluble (negative process) are washed out by development, thereby selectively uncovering the substrate surface to be processed.

Photoprinting of printed circuit boards requires the application of negative photoresists, predominantly dry film resists, almost exclusively [366]. In certain applications (e.g., the production of solder masks), liquid resists are being used increasingly for technical or economic reasons [431].

4.2.2.4. *Other Techniques*

Magnetographic [351] and electrophotographic [432] techniques are also of some interest. Special techniques are required to produce wiring paths on three-dimensional printed circuit boards [433].

4.2.3. *Economic Aspects*

Worldwide total sale of printed circuit boards in 1984 was estimated at \$ 10×10^9. Materials used in the production of printed circuit boards accounted for \$ 9.2×10^9, with an annual growth rate of ca. 12 % [405].

4.2.4. *Occupational Health and Environmental Protection*

Questions connected with the use of photoresists are dealt with in Section Occupational Health and Environmental Protection. Safe handling of strongly acidic or alkaline solutions and of heavy-metal compounds is the subject of a number of special publications and of laws specific to each country. For further details, see [410] and [434].

4.3. Microelectronic Devices

Imaging technology for modern microelectronic applications must meet some unique performance requirements, such as

1. resolving features only two times larger than the exposure wavelength;
2. “unbelievably perfect” patterns with defect densities of fewer than one defect >0.2 µm per 10 cm^2 (defects are defined as flaws ≥0.2 µm);
3. the resist must mask a number of aggressive wet and dry etch environments;
4. imaging of features on substrates with widely varying reflectivity and steps equal in height to the final line width.

The characteristics follow directly from the manufacturing requirements of silicon-based integrated circuits and the available exposure technology. *Integrated circuits* are the dominant microelectronic commercial devices produced by imaging.

4.3.1. *Integrated Circuit Manufacturing [435-437]*

An integrated circuit consists of a threedimensional structure of electrically conductive regions in the surface of a silicon wafer (Fig. 76). The regions vary in physical properties as evidenced by the resistivity changing from 10^{-6} Ω/cm^2 (metal) through 10^{-1} – 10^4 Ω/cm^2 (doped silicon, ⟶ Semiconductors) to infinity (silicon dioxide). The conductive regions are assembled into transistors, capacitors, etc., and linked together to process either analog or digital signals. Digital signal processing leads to computer applications with either metal – oxide – silicon (MOS) field effect transistors or bipolar

transistors (⟶ Semiconductors); MOS is the simplest technology and is used in high-density, cheap memories and microprocessors. The bipolar technology is more complex but gives faster switching speeds and is used in high-speed applications.

Figure 76. Perspective schematic of a 1 µm transistor [435]

A typical integrated circuit is ca. 15 mm on one side and is manufactured on a silicon wafer 100 – 200 mm in diameter, so that >100 circuits can exist on an individual wafer. After manufacture, the circuits are tested, and the wafer is diced to remove the good die. Yields can vary from 1 – 2 % for an experimental product to >90 % for a proven one. The yield determines the commercial viability of a production circuit.

The processes used to form the three-dimensional structures in silicon can be divided into vertical and horizontal operations. *Vertical operations* generate layers in the silicon by

1. growth (e.g., oxidation of silicon to silicon dioxide);
2. deposition (e.g., chemical vapor deposition of silane and oxygen to form silicon dioxide) or processes that transfer material from one solid phase to another via a gas (e.g., evaporation or sputtering); and
3. doping, either by diffusion of gas at high temperature or by scanning an accelerated beam of ions across the wafer (ion implantation).

Horizontal structures are produced by transferring a pattern from a mask into the photoresist, developing the photoresist, and then transferring the pattern into the underlying device by etching (Fig. 77). The mask usually consists of a pattern etched in chrome on a glass substrate; the manufacturing process is discussed in the following section. Transfer from the photoresist into the substrate proceeds via wet or dry etching. A *wet etch* reacts with the surface to give water-soluble products. In *dry etching*, either a plasma is used which reacts with the surface to yield volatile products, or the surface is sputtered away in a solid – gas – solid transfer. *Reactive ion etching* is a combination of plasma and sputter etch designed to give anisotropic (depth etches faster than width) etching. In ion implantation (vertical operation), a resist or etched silicon dioxide is often used as an implantation mask so that a vertical layer is formed and the pattern is transferred simultaneously. A typical device requires 10 – 15 layers and, hence, a similiar number of mask levels and etching operations.

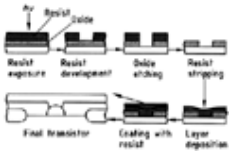

Figure 77. Process sequence for a “mask level”. The sequence is repeated about ten times to produce the final device.

Defect Density Requirements. The scale of integrated circuits is most difficult to comprehend; 1 Mbit random access memory is taken as a practical reference scale. The minimum line width of 1 µm is imaged by using light with a wavelength of 436 nm. Transistor performance must be controlled to ± 10 %, so line width must be controlled to the same tolerance (i.e., ± 1/4 of the exposure wavelength). These lines must be controlled across a 150-mm wafer, with a wafer width – line width ratio of $>10^5$. The *critical defect size* is 0.2 µm, which follows from line width control requirements. The *defect density* can be derived from the circuit size (≅1 cm^2) and the number of mask levels (≃10). If a defect exists on any level, the device will fail; therefore, defect density should be <1 (defect) per 10 cm^2.

4.3.2. *Photoprocess [438-441]*

The high degree of control and low defect density described in the previous section place some unique constraints on the photoprocess. All processing must be done under clean-room conditions, on automated equipment that minimizes particles generated by handling. The schematic in Figure 77 can be used to follow the discussion.

Substrate. One of the most significant differences between microelectronic lithography and most other lithographic applications arises because the optical properties of the substrate change from mask level to mask level, and across individual wafers. The optical properties are a function of the properties of the top layer of the wafer at each mask level. In most cases, this top layer is uniform and unpatterned, but the refractive index varies from mask level to mask level, thus causing reflectivity changes. If the top layer is a dielectric, it will be transparent and the patterned underlying layers will cause local reflectivity variations across the wafer.

Steps also occur in the substrate that are caused by the underlying patterned layers which can be of the same height as the line width. These steps cause variations in resist thickness which, in turn, lead to variations in line width. If the steps are covered by highly reflective metal or polysilicon, scattered light from the steps also causes line width variation. Finally, grain in materials is also a source of light scattering.

Surface Preparation. Surface preparation is essential to ensure adhesion of many photoresists. Freshly prepared surfaces without adsorbed water give the best adhesion. Adhesion of substrates is often improved by removing the surface water by dehydration baking. Aqueous liquids with a high affinity for hydrophilic surfaces (metals or silicon dioxide) exhibit the worst adhesion. *Adhesion promoters* are often used, especially with optical positive resists that employ aqueous developers. The most common adhesion promoter is hexamethyldisilazane, which makes surfaces hydrophobic before the resist is applied.

Selection of Photoresists. The selection of a photoresist is the result of cost – performance analysis. Two classes of material are in common use: (1) negative cross-linking systems based on cyclized butadiene, and (2) positive materials based on novolac resins, or naphthoquinone diazides which improve the dissolution rate in aqueous alkali on exposure.

The performance criteria and relative performance of the materials are summerized in Table 6. The dominant factor is resolution in terms of the minimum resolvable feature and the profile quality of the resolved lines. Cross-linking negative resists swell in the developer and limit the resolution to 2 – 3 µm. Positive resists developed in aqueous alkali do not appear to swell at all, and standing waves <0.1 µm in size can be resolved easily. Swelling also produces rounded profiles in negative resists in contrast to almost vertical walls in positive ones.

Table 6. Comparison of positive and negative resists

	Negative	Positive
Resolution	<2 – 3 µm	<0.1 µm
Profile	round	vertical
Process control	good	good
Adhesion	excellent	good (adhesion promoter)
Dry etch resistance	good	excellent
Sensitivity, mJ/cm^2	20	200 *

* Modern equipment can expose at these doses.

Positive resists are three to five times more expensive than negative ones. If features of <4 µm must be resolved, positive resists should be chosen.

Coating. The objective of coating is to produce layers with a uniformity >20.0 nm to ensure uniform optical properties (no “thin-film interference”): 2 – 3 mL of resist is placed in the middle of the wafer, and the wafer is spun at 3000 – 4000 rpm, causing rapid solvent evaporation in the presence of high shear forces on the photoresist. The puddle that is initially $\simeq$ 0.5 cm thick is spun down to a thin film of 1×10^{-4} cm with the desired uniformity. The formation of convection cells, which produce surface irregularities (striations), can be prevented by modifying the surface tension of the resist [442].

Baking. The objective of baking is to remove the solvent without degrading the photosensitizer. The three commercial options are

1. *convection ovens* in which the wafers are placed in a hot circulating airstream;
2. *hot plates* in which the wafers are held in contact with a large thermal mass (a large mass, thermally heated); and
3. *infrared ovens* that use radiant heating of the wafer.

Hot plate baking gives better thermal coupling, can be automated easily, and is the current standard technique. In many resists, a significant amount of solvent remains in the layer at a temperature that can degrade the photosensitizer [443]. Because the solvent affects the dissolution rate, control of the baking operation is critical.

Exposure. The objective of exposure is to transfer the image from the mask to the wafer and convert the photosensitizer locally to form a “latent image.” All commercial exposure equipment uses high-pressure mercury lamps, with the 436-, 408-, and 365-nm lines overlapping the sensitizer absorption curve (see Section Diazotyping).

A variety of commercial exposure equipment is available. From the point of view of imaging alone, the differences can be summarized as (1) image transfer process, (2) image field shape, (3) mask size, and (4) illumination spectrum.

Two image transfer processes exist: (1) *contact – proximity process* in which the mask is placed in contact with, or close to, the wafer; and (2) projection in which the mask is imaged through a lens. In the contact – proximity process, image quality is determined by near field diffraction from the mask, which results in excellent images in contact printing. The major problems in contact printing are the defects caused by forcing mask and wafer together. Projection imaging is controlled by far field diffraction effects associated with the lens.

Projection technologies are further classified by the shape of the image field. The whole wafer cannot be imaged directly, and different optical configurations produce differently shaped image fields that must be moved to cover the whole wafer. *Scanning* cameras have optics that produce a narrow slit field as wide as a wafer, which is scanned to cover the whole wafer. *Step and repeat* cameras produce a rectangular field that is stepped across the wafer.

The mask size relative to the final image size is also a function of optics. In contact printers and projection machines with reflective optics, mask and image are of the same size “1 x.” By contrast, machines with transmissive optics use reduction so that the mask is five to ten times larger than the final pattern, which facilitates mask making.

The illumination spectrum is also a function of optics. Machines with *reflective* optics use multiple mercury lines, whereas machines with *transmissive* optics use a single wavelength (usually 436 nm) because a chromatic correction must be made.

Exposure systems are selected according to resolution of the lens, throughput, alignment, and focus accuracy. Step and

repeat cameras dominate high-resolution applications.

Postexposure Operations. Process operations after exposure are designed to modify the latent image in some way. The best known *physical* modification is postexposure baking, which induces diffusion of the sensitizer in positive resists to eliminate standing waves. Standing waves are generated by exposure machines that use a single wavelength. *Chemical* modifications include image reversal of positive resists by baking in an ammonia atmosphere or simply by baking one particular resist (AZ 5200) [444].

Developing. The objective of developing is to remove the exposed photoresist uniformly. The two commercial methods are (1) batch developing in which a boat of 20 wafers is immersed in a tank, and (2) spray developing in which the developer is sprayed on each wafer as it rotates. Spray development can be automated easily and is the dominant technology in modern facilities.

Hard Baking. The objective of hard baking is to remove the solvent and cross-link the resist to improve thermal stability and adhesion. The commercial implementation options are the same as in any other baking operations.

Stripping. After all processing steps have been completed, including etching and ion implantation, the photoresist must be removed without residues. The two options are wet or dry stripping. *Wet stripping agents* can be subdivided further into solutions that *oxidize* the resist (sulfuric acid – hydrogen peroxide mixtures) and solutions that *dissolve* the resist (polar solvent – water mixtures). *Oxidizing* systems remove all water-soluble oxides and produce the cleanest surface. Of course, they cannot be used if the wafer surface is oxidizable (e.g., a metal). Solutions that *dissolve* the resist remove all soluble material, so strongly cross-linked resist may be left behind.

Dry stripping agents are either oxygen-plasma or ozone based, and they remove all compounds that form volatile oxides. Unfortunately, metal contaminants in the resist, such as sodium and iron, form nonvolatile oxides so they are always left after dry stripping. Some form of combined wet – dry stripping is used for the most severe cases.

Mask Making. Many types of mask are available, including silver halide and bright or antireflective chrome masks, generated by electron-beam or optical equipment. For large-scale integrated circuits, mask making is nearly always done by electron-beam lithography on antireflective chrome. The mask making process is a very simple integrated process using photoresists on a chrome layer on a glass plate. The chrome is wet-etched to generate the mask. Antireflecting chrome is chrome that has been surface oxidized. The mask will be 1, 5, or 10 times the size of the final pattern, depending on the optics of the production lithography machine. The various lithography machines are discussed in more detail in the following sections.

4.3.3. *Techniques for Submicron Lithography [440], [445]*

Because of the increased integration of devices, techniques must be developed to integrate smaller features.

These can be obtained by

1. resolving smaller features with existing lenses,
2. improving the resolution of lenses by using either shorter wavelengths or lenses with a larger numerical aperture (NA), or
3. higher resolution imaging technology such as electron beam or X-ray.

Resolving Smaller Features with Existing Lenses. To resolve smaller features, line width control must be maintained although the quality of the image delivered by the lens becomes poorer. The image quality, as measured by the modulation transfer function (MTF) for sets or equal lines and spaces, decreases rapidly with decreasing feature size (Fig. 78 B).

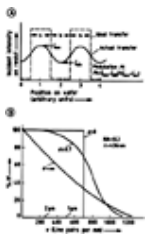

Figure 78. Schematic showing the definition of modulation transfer function (MTF) (A) and the variation of MTF with resolution (B). A partial coherence σ of 0.7 is the value for most imaging systems

Experience has shown that *single-layer* lithography can be produced down to an MTF of 0.8 [445]. A 436-nm lens with an NA of 0.35 will print 1-µm lines and spaces with an MTF of 0.8. *Multilayer* techniques have been used to make devices at an MTF of 0.6 [446]. Below an MTF of 0.6, the image becomes so blurred that light from neighboring features overlaps and produces line widths that vary depending on the number and size of these neighbors (“optical proximity effect”) [446]. These effects cause very complex line width variations that are probably uncorrectable; hence, an MTF of 0.6 represents the practical resolution limit of a lens [445].

Practical lithography at modulation transfer functions between 0.8 and 0.6 must use processes with exceptional control because the image is so poor. These process variations that determine the line width can be grouped as follows:

1. Variations of incident image (e.g., dose, focus)
2. Variations of material dissolution rate (e.g., reflectivity, steps)

The classical solution to the first two problems is to use thin, very transparent resists. Unfortunately, thin resists give poor line

width control across steps, and transparent resists are very sensitive to reflections.

Many advanced resist processes have been described that attempt to reduce the effects of the three different process variations on line width. Examples are image reversed positive resists [444], contrast enhancement materials [447], and multilayer resist systems [446] including the “built on mask” (BOM, a process for producing a chemical mask in situ) system [448]. In general, all these systems give excellent wall profiles and, hence, good line width control over steps. Most of them improve sensitivity relative to image variations. Substrate reflection remains the most difficult problem to eliminate. Any processes in which light from the initial or any pattern transfer exposure hits the underlying substrate can cause problems. Only processes in which etching is used for pattern transfer are completely free of reflectivity problems. All these systems attempt to gain process control at the expense of additional process complexity. This complexity results in higher costs and additional defects.

Improved Lenses. The resolution of a lens is proportional to the ratio of wavelength to numerical aperture; i.e., as the wavelength decreases, diffraction is reduced, and as the NA (“size”) of the lens increases, more of the diffracted light is captured and resolved back onto the wafer (Fig. 79 A). Reducing the wavelength becomes progressively more difficult because of decreasing glass transmission, and it becomes extremely difficult at ≤310 nm where quartz is the only conventional lens material available. Consequently, chromatic abberations at short wavelengths are very difficult to correct, and very narrow band excimer laser light sources are being investigated.

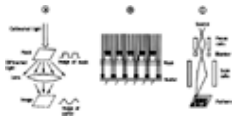

Figure 79. Advanced resist techniques

A) Optical lithographic system; B) X-ray system; C) Electron-beam system

Conventional positive resist based on diazo compounds works well down to 310 nm, although special diazo sensitizers give the best results [444]. At ≤310 nm, the aromatic polymers that provide high dry etching resistance are very absorbing, so alternative resins (e.g., copolymers of styrene) are needed. Very intense laser light sources can also result in different (potentially beneficial) photochemistry.

Resolution may also be increased by increasing the NA, with the disadvantage of a significantly reduced depth of field. The depth of field is inversely proportional to the square of NA and increases linearly with wavelength. The sensitivity of the process to focus changes can be reduced by using a thin imaging layer and by planarizing the substrate as much as possible. A multilayer system may be essential for use with lenses of very high NA (>0.4).

X-Ray Lithography. X-ray lithography involves proximity printing by using very short-wavelength photons to reduce diffraction effects (Fig. 79 B). The mask is the crucial point in X-ray lithography because the mask base must transmit soft X-rays. The base is typically a 5-µm membrane of silicon carbide or boron nitride, which provides mechanical strength and stability, with a gold absorber pattern. The mask must be of the same size as the final pattern (“IX“). The technological problems associated with masks of low defect density represent the greatest unknowns in X-ray lithography.

Three X-ray sources are possible: generation by (1) electron impact, (2) plasma, and (3) synchrotron rings. Only *synchrotron rings* have sufficient intensity for submicron step and repeat lithography. The synchrotron ring forces high-energy electrons around corners to generate X-rays. The principal problem with the synchrotron ring is its cost, which is estimated to be $ 5 – 10×10^6.

Even with a synchrotron source, photoresist must have a quantum efficiency one to two orders of magnitude greater than conventional materials. Catalytic resist systems in which a photogenerated species catalyzes degradation of 50 – 100 dissolution inhibitor molecules appear to be the most satisfactory.

Electron-Beam Lithography. Electron-beam lithography differs from other technologies in that it is a primary pattern generation technique. A focused electron beam is scanned across a wafer and switched on and off at each location to delineate the pattern. Throughput is a function of the number of locations that must be visited and the length of time that the beam must reside at each location to expose the resist. The residence time is a function of beam intensity and resist sensitivity. To meet a throughput of 30 wafers an hour requires sophisticated high-speed electronics and very fast resists. The major drawback for its use in submicron lithography is that the number of locations varies with the square of the minimum feature sizes, which makes throughput very sensitive to minimum feature size.

The resist requirements are similar to X-ray, and similar catalytic systems are being evaluated. The exposure of patterns by electrons backscattered from the substrate (proximity effect) causes line width variation.

[Top of Page]

5. Photochemical Machining

David Allen

5.1. Introduction

Using techniques similar to those employed for the production of printed circuit boards (see Section Printed Circuits (Printed Circuit Boards)) and integrated circuits (see Section Microelectronic Devices), the photochemical machining (PCM) industry plays a valuable role worldwide in the production of precision parts and decorative items. Examples are color television shadow masks; integrated circuit lead frames; magnetic recording head laminations; heat ladders, plates, and sinks; attenuators, light choppers, and encoder disks; grills, grids, sieves, mesh, and screens; washers, shims, and gaskets; jewelry; and decorative signs, plaques, and nameplates (Fig. 80).

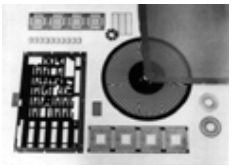

Figure 80. A selection of parts made by PCM

In the multistage PCM process shown in Figure 81, the use of photoresists (see Section Photoresists) ensures higher resolution than that obtained by traditional metal-cutting techniques. Details are given in [449].

Figure 81. The modern PCM process

Terminology. Photochemical machining is also known as *photoetching, photochemical milling, photomilling, photofabrication,* and *photochemical etching*. In the United States, *chemical blanking* and *photoforming* (a term used in Europe for a very different technique, see Section Economic Aspects) are also synonyms for this technique.

5.2. Artwork Generation and Phototool Production

Various methods exist for the production of phototools. Some employ the manufacture of enlarged artwork (ink drawings; scraperboard art; adhesive colored tapes or lettering on card; or, most common today, computer-aided artwork generated from scribed “cut and peel” film) which is subsequently photoreduced to the correct size.

Other methods produce a phototool of the correct size by using a photoplotter or a laser pattern generator to expose the photographic emulsion selectively according to computer-aided design data.

Once a master image has been made, it may be processed photographically by “step and repeat” and contact printing to form a multiple-image, registered, double-sided phototool.

Ultimately, the future may see the demise of phototooling. At present, photoresists can be imaged directly from CAD data by using UV lasers. To achieve economical, rapid, high-quality imaging, precision control of a high-resolution, high-powered, low-cost laser is essential.

5.3. Materials

Virtually all materials can be etched, although some are etched more easily than others. Etching is basically rapid, controlled corrosion. Thus, corrosion-resistant materials are difficult to etch and require extremely corrosive etchants. Some metals (tantalum, gold, titanium) [450] and alloys (Elgiloy and Duratherm 600, containing 43 wt % and 41 wt % cobalt, resp.) [451] fall into this category, together with ceramics [452], glass, and plastics. However, many materials used commonly in manufacturing can be etched readily by using aqueous metal salt solutions or dilute acid or alkali (see Section Etching Technology).

5.4. Photoresist Systems

Photoresists are described in Section Photoresists.

Due to the nature of PCM, which involves the production of parts from sheet metal and foil with substantial thickness, photoresist resolution is not usually a problem.

Adequate resolution can be obtained by using a relatively thin photoresist coating und UV light as an exposure source. This coating can be applied to the metal as a liquid or solid film. Photoresists are commonly applied by

1. dipcoating (withdrawal of the metal sheet through a meniscus of liquid);
2. flowing (used, for example, in the mass production of television shadow masks from continuous coils of mild steel); and
3. laminating dry films of photoresist onto metal sheets by a combination of heat and pressure.

In addition to the photoresists mentioned in Section Photoresists, other negative-working photoresists used in PCM include formulations based on fish gelatin and casein sensitized with ammonium dichromate.

The photoresist is exposed with UV light through a phototool and forms a stencil after development.

The developer may be an organic solvent or an aqueous solution. For environmental reasons, the use of organic solvent developers is decreasing and that of aqueous developers is increasing, especially for dry film photoresists.

After development and rinsing, liquid resists are baked to increase their resistance to the etchant. Because etching usually produces appreciable undercut (see Section Etching Technology), most of the resists used in PCM are negative-working as their stencil edges are less prone to fracturing (cf. positive-working resists).

After etching, the resist stencil is usually stripped from the substrate. Again, environmental considerations favor the use of aqueous strippers such as potassium hydroxide solution for dry film photoresists.

5.5. Etching Technology

Material not covered by the photoresist stencil is attacked chemically by the etching solution to form soluble products. For etching to be economical, the etchants must be relatively innocuous, fast-acting on a wide variety of materials, and easily disposed of when spent.

For these reasons, etching is usually performed with either iron(III) chloride solutions (suitable for copper or nickel and their alloys, ferrous materials including stainless steel, and aluminum) or copper(II) chloride solutions which etch copper and its alloys.

As material dissolves into the etchant, the concentration of impurities increases and the activity of etchant decreases. To maintain a constant etchant activity and to lower costs, the technique of *etchant regeneration* is now being used widely. In the ideal case of etching iron with iron(III) chloride and regenerating the etchant, excess iron(III) chloride is produced

$$\begin{array}{c} Fe + 2\,FeCl_3 \xrightarrow{\text{Etching}} 3\,FeCl_2 \xrightarrow{\text{Regeneration}} 3\,FeCl_3 \\ \uparrow \qquad\qquad \text{Dilution} \qquad\qquad \downarrow \end{array}$$

The situation is more complicated when etching of alloys is considered. Regeneration methods involve (1) *chemical treatment* with gases such as chlorine and ozone, or oxidizing agents such as sodium chlorate; and (2) *electrolytic treatment* based on anodic oxidation of iron(II) to iron(III) ions.

Similarly, economic advantages accrue if copper(II) chloride solution, used as an etchant for copper, is regenerated:

$$\begin{array}{c} Cu + CuCl_2 \xrightarrow{\text{Etching}} 2\,CuCl \xrightarrow{\text{Regeneration}} 2\,CuCl_2 \\ \uparrow \qquad\qquad \text{Dilution} \qquad\qquad \downarrow \end{array}$$

The economics of each process depends on plant and chemical availability and energy costs, but cost reduction factors of ten have been quoted in the United States [453]. High carbon content in alloys results in a lower cost reduction factor [454].

When a metal is being etched downward, a sideways etching under the stencil also results, leading to *undercut* (see Fig. 82). This is responsible for some of the limitations of PCM (see Section Process Capability), such as the dependence of minimum hole size on material thickness.

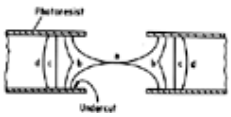

Figure 82. Development of etched edge profiles and undercut profile (a) ⟶ (b) ⟶ (c) ⟶ (d) as etch time increases

5.6. Process Capability

The main advantages of PCM are as follows:

1. The method is a low-temperature technique that does not affect the physical or chemical properties of the starting material.
2. The saying “If you can draw it, PCM can make it” is basically true, provided the limitations discussed below are considered. The PCM process is ideal for creating complex metal parts involving arrays of apertures and shapes that are difficult to obtain by its traditional rival, stamping (see Section Economic Aspects).
3. Although usually associated with the production of planar components, PCM may be used to etch the surfaces of three-dimensional components based on cylindrical, conical, or spherical geometry [455].
4. Lead times are short so that the field test of a new product can start very quickly. If modifications are required, they are inexpensive and quick to make.
5. No burrs or distortions are produced during manufacture.
6. Components may be produced separately or tabbed into a sheet for ease of handling.

In general, the main limitations of PCM, in material of thickness T, are

1. minimum hole diameter $\Phi_{min} = 1.1\,T$,
2. upper thickness limitation of 1.5 mm (although 6 mm is possible for low-resolution work),
3. some deviation from a classical straight profile at the edges of parts, usually controlled to <0.25 T.

5.7. Economic Aspects

Every manufacturing method has its advantages and disadvantages, and should be selected accordingly. If the part to be

produced is at the limit of the process capability of the method, the yield of acceptable products will be poor, which is reflected in higher production costs.

Other processes, which may be competitive with PCM, are, therefore,

1. photoforming, (i.e., electroforming through a photoresist stencil);
2. wire electro-discharge machining, i.e., spark erosion using a thin wire as the cathode, a dielectric fluid such as water, and sheet metal as the anode and workpiece;
3. laser beam machining, i.e., laser cutting with an intense, focused heat source, as obtained, e.g., with a carbon dioxide laser;
4. stamping or fine blanking and piercing.

In deciding which method to use for part production, the technical aspects of the methods should be checked to ensure the selection of appropriate techniques for the material used and its thickness as well as part specifications in terms of edge profile, flatness, and dimensions. A study of model parts capable of being manufactured by all five methods determined that the factors influencing the most economical method of production were part complexity, batch size, and part thickness [456]. Generally, large-volume production favors *stamping*; very low volumes favor *wire electro discharge machining*; high complexity favors *photochemical machining*; thicker components favor *laser beam machining*; thinner components favor *photochemical machining*, and extremely thin (<0.025 mm thick), high-resolution components in metals capable of being electroplated favor *photoforming*.

5.8. Products

The range of products is described in the introduction, but the number of products made worldwide by PCM is difficult to estimate. However, 1988 output figures from Japan include 14×10^9 integrated circuit lead frames, 110×10^6 vacuum fluorescent display meshes, and 36×10^6 color television shadow masks (including 10×10^6 high-resolution masks) made by this technique [457].

[Top of Page]

References

1. J. H. Dessauer, H. E. Clark: *Xerography and Related Processes*, Focal Press, London 1965.
2. R. M. Schaffert: *Electrophotography*, 2nd ed., Focal Press, London 1975.
3. W. E. Haas, in J. Sturge, V. Walworth, A. Shepp (eds.): *Imaging Processes and Materials*, Neblette's Eighth Edition, Van Nostrand Reinhold, New York 1989, pp. 375 – 405.
4. IS&T - The Society for Imaging Science and Technology http://www.imaging.org SPIE - International Society for Optical Engineering http://www.spie.org
5. C. J. Young, H. G. Grieg, *RCA Rev.* **15** (1954) 441 – 484.
6. M. E. Scharfe, F. W. Schmidlin in L. Morton (ed.): *Advances in Electronics and Electron Physics*, Academic Press, New York 1975, pp. 83 – 146.
7. D. M. Pai, R. C. Enck, *Phys. Rev. B* **11** (1975) 5163 – 5174.
8. I. Chen. J. Mort, *J. Appl. Phys.* **43** (1972) 1165 – 1170.
9. M. E. Scharfe: *Electrophotography, Principles and Optimization*, Research Studies Press, Letchworth, England 1984.
10. D. M. Pai, A. Melnyk, *Proc. SPIE Int. Soc. Opt. Eng.* **81** (1986) 83 – 95.
11. J. H. Neyhart, *Photogr. Sci. Eng.* **10** (1966) 126 – 132.
12. L. Cheung, G. M. T. Foley, P. Fournia, B. E. Springett, *Photogr. Sci. Eng.* **26** (1982) 245 – 249.
13. K. Kiyota, A. Teshima, M. Tanaka, *Photogr. Sci. Eng.* **24** (1980) 289 – 292.
14. M. Lutz, *SPIE Proceedings* **759** (1987) 35 – 39.
15. H. Pinsler, H. Walsdorfer, M. Lutz, *J. Imaging. Technol.* **16** (1990) 55 – 59.
16. E. J. Felty in O. Helwich (ed.): *Proceedings of the Second International Congress on Reprography*, Cologne, Germany 1967, pp. 40 – 46.
17. W. E. Spear, P. G. LeComber, *Solid State Commun.* **17** (1976) 1193 – 1195.
18. I. Shimizu, *J. Non Cryst. Solids* **35, 36** (1983) 773 – 778.
19. N. Yamamoto, *Jpn. J. Appl. Phys.* **20** (1981) 305 – 310.
20. D. M. Pai in J. Adler, H. Fritzche, S. R. Ovshinsky (eds.): *Physics of Disordered Materials*, Plenum Press, New York 1985, pp. 579 – 592.
21. J. A. Amick, *RCA Rev.* **20** (1959) 770 – 784.
22. P. Mark, *Photogr. Sci. Eng.* **18** (1974) 254 – 261.
23. K. Yoshida, K. Kinoshita, T. Kawamura, *Appl. Opt. Suppl.* **3** (1969) 170 – 175.
24. M. Lardon, E. Lell-Doller, J. W. Weigl, *Mol. Cryst.* **2** (1967) 241 – 246.
25. W. D. Gill, *J. Appl. Phys.* **43** (1972) 5033 – 5040.
26. G. Weiser, *J. Appl. Phys.* **43** (1972) 5028 – 5033.
27. P. J. Melz, *J. Chem. Phys.* **57** (1972) 1694 – 1699.
28. P. J. Regensburger, *Photochem. Photobiol.* **8** (1968) 429 – 440. Links
29. D. M. Pai, J. Yanus, *Photogr. Sci. Eng.* **27** (1983) 14 – 19.
30. G. Pfister, *Phys. Rev.* **B 16** (1977) 3676 – 3687.

31. P. M. Borsenberger, P. M. Mey, A. Chowdry, *J. Appl. Phys.* **49** (1978) 273 – 279. Links
32. M. Stolka, J. F. Yanus, D. M. Pai, *J. Phys. Chem.* **88** (1984) 4707 – 4714.
33. P. J. Melz *et al.*, *Photogr. Sci. Eng.* **21** (1977) 73 – 78.
34. A. Kakuta, T. Mori, H. Morishita, *IEEE Trans. Ind. Appl.* **IA-17** (1981) 4 – 10.
35. L. B. Schein, A. Rosenberg, S. L. Rice, *J. Appl. Phys.* **60** (1986) 4287 – 4292. Links
36. E. G. Schlosser, *J. Appl. Photogr. Eng.* **4** (1978) 118 – 122.
37. E. S. Baltazzi, *J. Appl. Photogr. Eng.* **6** (1980) 147 – 152.
38. D. Winkelmann, *J. Appl. Photogr. Eng.* **4** (1978) 187 – 194.
39. R. O. Loutfy, A. M. Hor, A. Rucklidge, *J. Imaging Sci.* **31** (1987) 31 – 37.
40. N. C. Khe *et al.*, *Photogr. Sci. Eng.* **28** (1984) 196 – 203.
41. K. Arishima, H. Hiratsuka, A. Tate, T. Okada, *Appl. Phys. Lett.* **40** (1982) 279 – 281. Links
42. W. J. Dullmage *et al.*, *J. Appl. Phys.* **49** (1978) 5543 – 5554.
43. P. Borsenberger, A. Chowdry, D. C. Hoesterey, W. Mey, *J. Appl. Phys.* **4** (1978) 5555 – 5564.
44. D. Winkelmann, K. Reuter in J. Gaynor (ed.): *Advances in Non-Impact Printing Technologies for Computer and Office Applications*, Van Nostrand, New York 1982, pp. 446 – 462.
45. D. Adam, H.-J. Humpert, S. Dreihöfer, H. Pinsler, M. Lutz, *Proc. 13. Int. Conf. on Digital Printing Technologies,* IS&T, Seattle 1997, 245 – 247
46. J. Mort, D. M. Pai: *Photoconductivity and Related Phenomena*, Elsevier, New York 1976.
47. A. S. Diamond (ed.): *Handbook of Imaging Materials*, Marcel Dekker, New York 1991.
48. P. M. Borsenberger, D. S. Weiss: *Organic Photoreceptors for Xerography*, Marcel Dekker, New York 1998.
49. M. M. Shahin, *Appl. Opt. Suppl.* **3** (1969) 106 – 110.
50. Canon, US Patent 4 851 960, 1989 (S. Nakamura, H. Hirabayashi, J. Araya, N. Koitabashi).
51. S. Matsumoto *et al.*, *J. Imaging Sci. Technol.* **39** (1995) 132.
52. H. Kawamoto, H. Satoh, *J. Imaging Sci. Technol.* **38** (1994) 383.
53. M. Lutz, *J. Imaging Sci. Technol.* **11** (1985) 256.
54. R. M. Schaffert, *Photogr. Sci. Eng.* **6** (1962) 197 – 215.
55. H. E. J. Neugebauer, *J. Appl. Opt.* **3** (1964) 385 – 393.
56. T. L. Thourson, *IEEE Trans. Electron Devices*, **ED 19** (1972) 495 – 511.
57. L. B. Schein: *Electrophotography and Development Physics*, Springer Verlag, Berlin 1987.
58. J. W. Weigl in M. Hair, M. D. Croucher (eds.): *Colloids and Surfaces in Reprographic Technology*, American Chemical Society, Washington, DC 1982, pp. 139 – 182.
59. L. B. Schein, *Photogr. Sci. Eng.* **19** (1975) 255 – 265.
60. W. A. Sullivan, T. L. Thourson, *Photogr. Sci. Eng.* **11** (1967) 115 – 120.
61. J. T. Bickmore, K. W. Gunther, J. F. Knapp, W. A. Sullivan, *Photogr. Sci. Eng.* **14** (1970) 42 – 51.
62. T. Takahasi, N. Hosomo, J. Kanbe, T. Toymono, *Photogr. Sci. Eng.* **26** (1982) 254 – 261.
63. R. B. Lewis, H. M. Stark in W. F. Berg, K. Hauffe (eds.): *Current Problems in Electrophotography*, "Chapter X", de Gruyter, Berlin 1972, pp. 322 – 335.
64. J. T. Bickmore, M. Levy, J. Hall, *Photogr. Sci. Eng.* **4** (1960) 37 – 43.
65. K. A. Metcalfe, *J. Sci. Instrum.* **32** (1955) 74 – 75.
66. J. Dahlquist, I. Brodie, *J. Appl. Phys.* **40** (1969) 3020 – 3027. Links
67. C. E. Case, T. M. Lawson, *14. Int. Conf. on Digital Printing Technologies,* IS&T, Toronto, 1998, pp. 226 – 230.
68. D. A. Seanor, *Photogr. Sci. Eng.* **22** (1978) 240 – 243.
69. T. Nylund, C. Cowan, J. Spence, L. Steele, 3rd *Int. Cong. on Advances in Non-Impact Printing Technologies,* SPSE, San Francisco, USA, 1986, pp. 74 – 80.
70. R. W. Gundlach, C. J. Klaus, *Photogr. Sci. Eng.* **7** (1963) 14 – 19.
71. P. Cressman, *J. Appl. Phys.* **34** (1963) 2327 – 2330.
72. M. J. Mitsui, *IEEE Trans. ED* **19** (1972) 397.
73. J. R. Rumsey, D. Bennewitz, *J. Imaging Technol.* **12** (1986) 144.
74. J. Johnson, *10. Int. Congr. on Advances in Non-Impact Printing Technologies,* IS&T, New Orleans, 1994, 311 – 313.
75. H. A. Hermanson, R. E. Drews, D. G. Parker, F. Tomak, S. Swackhamer, in S. W. Ing, M. D. Tabak, W. E. Haas (eds.): *Electrophotography, Fourth International Conference*, SPSE, Springfield, VA 1983, pp. 541 – 570.
76. J. J. Eltgen, in S. W. Ing, M. D. Tabak, W. E. Haas (eds.): *Electrophotography, Fourth International Conference*, SPSE, Springfield, VA 1983, p. 519.
77. V. Tulagin, *J. Opt. Soc. Am.* **59** (1969) 328 – 331.
78. P. Warter, V. Tulagin, L. Carreira, 3rd *Int. Cong. on Advances in Non-Impact Printing Technologies,* SPSE, San Francisco, USA, 1986, pp. 228–332.
79. P. Cressman, G. C. Hartman, *J. Chem. Phys.* **61** (1974) 2740 – 2746. Links
80. G. Hartmann, F. Schmidlin, *J. Appl. Phys.* **46** (1975) 226 – 278.
81. W. L. Goffe, *Photogr. Sci. Eng.* **15** (1971) 304 – 308.
82. A. L. Pundsak, *Photogr. Sci. Eng.* **18** (1974) 642 – 651.
83. P. Selenyi, *J. Appl. Phys.* **9** (1938) 637 – 641.

84. G. L. Pressman, *2nd International Conf. on Electrophotography*, SPSE, 1974, pp. 38 – 45.
85. K. S. Pennington, W. Crooks, “Resistive Ribbon Thermal Transfer Printing (R 2 T 2): A New Printing Technology”, *Technical Abstracts Second International Congress on Non-Impact Printing*, Society for Photo-Optical Scientists and Engineers, Springfield, Virginia 1984, p. 236.
86. W. Boles, *Thermal Printhead Technology and Design*, Hewlett-Packard Company, Desktop Computer Division, Fort Collins, Colorado, Computer Design October, 1979.
87. Texas Instruments, Inc., US 3 982 093, 1976 (W. S. Henrion).
88. S. Shibata, K Murasugi, K. Kaminishi: “New Type Thermal Printing Head Using Thin Film”, *IEEE Trans. Parts Hybrids Packag.* **12** (1976) no. 13, 223 – 230.
89. K. Nunomura, N. Koyama, T. Matsubara: “Ta-SiC Thin Film Resistors of Highly Reliable Thermal Printing Heads”, *IEEE Trans. Components Hybrids Manuf. Technol.* **CHMT-6** (1983) no. 4, 260 – 263.
90. R. Shaw, *J. appl. Photogr. Eng.* **2** (1976) 17 – 27
91. J. W. Weigl, *Angew. Chem. Int. Ed. Engl.* **16** (1977) 374 – 392
92. L. B. Schein: *Electrophotography and Development Physics*, Springer Verlag, Berlin 1987.
93. T. Nozawa, T. Suzuki, “Zirconium Nitride Thin Film Resistors for Thermal Print Heads”, *Int. J. Hybrid Microelectron.*, **7** (1984) no. 3, 63 – 67.
94. W. Butler, “Technological Aspects of the Development of a Thin Film Thermal Printhead”, *Int. J. Hybrid Microelectron.*, **5** (1982) no. 1, 1981 European Hybrid Microelectronics Conference, Avignon, France.
95. S. Shibata, M. Ito, “Evaluation of Characteristics of Meander Heat Elements for a Thermal Printing Head”, *Electronics and Communications in Japan*, 66-C, (1983) no. 4, 321 – 327.
96. P. A. McLaughlin, E. W. Stapleton, J. E. Tumbaugh: “Development of a Thick Film Thermal Print Head”, *Tektronix, Inc., Beaverton, Oregon, International Microelectron Symposium*, San Francisco, California October 22 – 24, 1973, Sess. Paper 1 A-4. Available from International Society for Hybrid Microelectron, Inc. Park Ridge, Illinois.
97. W. Crooks, C. Hafer, W. Weiche, US 4 456 915, 1984.
98. G. F. Brooks, W. Crooks, W. Weiche, US 4 103 066, 1978.
99. E. K. Sichel, P. Scheng, J. I. Gittleman, *Phys. Rev. Lett.* **40** (1978) 1197; *Phys. Rev. B Condens. Matter* **18** (1978) 5712; *Phys. Rev. B. Condens. Matter* **24** (1981) 6131.
100. K. K. Shih, D. B. Dove: “Electrical Properties of Resistive Ribbon”, *IBM J. Res. Dev.* **29** (1985) no. 5, 519 – 526.
101. T. G. Twardeck, “Characterization of a Resistive Ribbon Thermal Transfer Printing Process”, *IBM J. Res. Dev.* **29** (1985) no. 5, 470 – 477.
102. S. Shibata, K. Hihei: “Development of 16 dots/mm Thermal Printing Head”, *Int. J. Hybrid Microelectron.* **6** (1983) no. 1, International Microelectronics Symp. Philadelpia, PA, October 31 to November 2, 1983, pp. 531 – 536.
103. P. Barnwell: “Thick Film Thermal Print Heads”, *Electrocomponent Sci. Technol.* **6** (1980) no. 3 – 4, 135 – 139.
104. Y. Tokunaga, M. Yoshida, M. Zazawa: “Life Estimation of Thermal Print Heads Consisting of Ta^2N Thin-Film Resistors”, *IEEE Trans. Components, Hybrids Manuf. Technol.* CHMT **4** (1981) no. 1, 148 – 153.
105. K. Mashio: “Current Developments of Thermal Papers”, *TAPPI Print Reprog. Test Conf. Proc., Rochester, New York* TAPPI, Atlanta, Georgia November 11 – 14,1979, pp. 133 – 139.
106. Graphic Controls Corp., US 4 500 354, 1982 (B. J. Katsen *et al.*).
107. IBM branch offices, IBM Program Product NO. 5796-PBH, Advanced Statistical Analysis Program (ASTAP) Reference Manual, No. SH 20–1180–0.
108. K. J. Fowler, C. H. Hafer: “Thermal Paper Printer”, *IBM Technical Disclosure Bulletin*, December, 1983, pp. 3781 – 3782.
109. Y. Tokunaga, K. Sugiyama: “Thermal Ink-Transfer Imaging”, *IEEE Trans Electron Devices* **ED-27** (1980) no. 1, 218 – 222.
110. W. J. Weiche, US 4 308 318, 1981.
111. S. Applegate *et al.*: “Implementation of the Resistive Ribbon Technology in a Printer and Correcting Typewriter”, *IBM J. Res. Dev.* **29** (1985) no. 5, 459 – 469.
112. H. Findlay, K. Jones, US 4 453 839, 1984.
113. T. Saito *et al.*: “A Large Scale Hybrid Thermal Print Head”, *Electro component Sc. Technol.* **10** (1983) 261 – 267.
114. J. C. Hurt, W. B. Baril: “Thick Film Thermal Printer Temperature Studies”, *Int. Microelectron Conf., Proc. of the Tech. Prog., Anaheim, CA, February 11 – 13, 1975, New York, June 17 – 19, 1975*, Industrial and Scientific Conference Management, Inc., Chicago, Illinois 1975, pp. 129 – 134.
115. L. F. Valenti: “Thermal Printhead Technology”, *Int. IGC Conference for Continuing Education Program*, Amsterdam, Holland April 1985.
116. H. Taub: “Thermal Ink Jet Printing”, *IGC Conference*, Key Biscayne, Florida October 27 – 29, 1985.
117. T. Yamada *et al.*: “High-Resolution Full Color Printer by Microdot Ink-Jet Printing Method”, *SID Digest*, **14** (1983) 102.
118. Lord Rayleigh: “Capillarity”, *The Theory of Sound*, 2nd ed. vol. **2**, "CLXX", Dover Publications Inc., New York 1945.
119. W. Thomson, GB 156 897, 1867.
120. C. W. Hansell, US 1 941 001, 1929.
121. R. Elmquist, US 2 566 443, 1950.
122. R. G. Sweet, US 3 596 275, 1964.
123. C. H. Hertz, S. I. Simonsson, US 3 416 153, 1968.
124. C. W. Hansell, US 2 512 743, 1946.
125. E. L. Kyser, S. B. Sears, US 3 946 398, 1970.

126. S. J. Zoltan, US 3 683 212, 1970.
127. N. G. E. Stemme, US 3 747 120, 1971.
128. T. Hara *et al.*, US 4 296 421, 1979.
129. M. Bovio *et al.*, US 4 238 807, 1978.
130. C. H. Hertz, G. Rydgren: "Ink Jet Marking by Ultrasonic Jet Control", *The Second International Congress on Advances in Non-Impact Printing Technologies*, SPSE Arlington, USA Nov. 4 – 8, 1984, p. 99.
131. N. Kiyohiro *et al.*: "Measurement of Droplet Velocity Using High Speed Motion Pictures", *The Second International Congress on Advances in Non-Impact Printing Technologies*, SPSE, Arlington, USA Nov. 4 – 8, 1984, p. 20.
132. N. Bugdayci *et al.*: "Axisymmetric Motion of Radially Polarized Piezoelectric Cylinders Used in Ink Jet Printing", *IBM J. Res. Dev.* **27** (1983) 171 – 180.
133. E. P. Hofer: "Spot-Size Modulation in Drop-on-Demand Ink Jet Technology", *SID Digest, New Orleans, USA* **16** (1985) 321.
134. W. Runge: "One-Dimensional Simulation of a Phase-Change Cycle in Thermal Ink Jet Printers", *SID Digest, New Orleans, USA* **18** (1987) 189.
135. M. Ikeda: "The Bubble Jet Printing Technology for BJ-80 Printer", *IGC 18th Annual Conference on Ink Jet Printing*, Amsterdam, March 27 – 29, 1985.
136. C. H. Hertz, B. Samuelsson: "High Quality Ink Jet Printing of Color Images", *The Third International Congress on Advances in Non-Impact Printing Technologies*, SPSE, San Francisco, California, USA August 24 – 28, 1986.
137. F. Schmidlin, *IEEE Trans. Electron Devices* **19** (1972) 448 – 457.
138. R. N. Mills, Display, *1987 International Symposium*, New Orleans, Louisiana, USA May 11 – 15, 1987.
139. G. S. Minto: "Ink Jet Printing", *IGC Lecture*, Carmel, California, USA Dec. 10 – 12, 1978.
140. B. Kazan, US 3 287 734, 1966.
141. P. F. Kurz, L. A. Nash, US 3 715 219, 1969.
142. B. Hermanrud: "The Compound Jet and Its Application", *The Second International Congress on Advances in Non-Impact Printing Technologies*, SPSE, Arlington, Virginia, USA Nov. 4 – 8, 1984, p. 35.
143. F. J. Kamphoefner: "Ink Jet Printing", *IEEE Trans. Electron Devices* **ED-19** (1972) no. 4, 584.
144. W. L. Buehner *et al.*: "Application of Ink Jet Technology to a Word Processing Output Printer", *IBM J. Res Dev.* **21** (1977) no. 1, 2 – 9.
145. S. A. Curry *et al.*: "Scale Model of an Ink Jet", *IBM J. Res Dev.* **21** (1977) no. 1, 10 – 20.
146. W. T. Pimbley *et al.*: "Satellite Drop Formation in a Liquid Jet", *IBM J. Res Dev.* **21** (1977) no. 1, 21 – 30.
147. T. G. Twardeck *et al.*: "Effect of Parameter Variations on Drop Placement in an Electrostatic Ink Jet Printer", *IBM J. Res. Dev.* **21** (1977) no. 1, 31 – 36.
148. G. L. Fillmore *et al.*: "Drop Charging and Deflection in an Electrostatic Ink Jet Printer", *IBM J. Res. Dev.* **21** (1977) no. 1, 37 – 47.
149. H. C. Lee *et al.*: "Boundary Layer around a Liquid Jet", *IBM J. Res. Dev.* **21** (1977) no. 1, 48 – 51.
150. J. M. Carmichael *et al.*: "Controlling Print Height in an Ink Jet Printer", *IBM J. Res. Dev.* **21** (1977) no. 1, 52 – 55.
151. M. Levanoni *et al.*: "Study of Fluid Flow through Scaled-Up Ink Jet Nozzles", *IBM J. Res. Dev.* **21** (1977) no. 1, 56 – 68.
152. S. D. Hawkins, US 4 459 601, 1984.
153. F. C. Lee *et al.*: "The Application of Drop-on-Demand Ink Jet Technology to Color Printing", *IBM J. Res. Dev.* **28** (1984) no. 3, 307.
154. D. B. Bogy *et al.*: "Experimental and Theoretical Study of Wave Propagation Phenomena in Drop-on-Demand Ink Jet Devices", *IBM J. Res. Dev.* **28** (1984) no. 3, 314.
155. J. E. Fromm: "Numerical Calculation of the Fluid Dynamics of Drop-on-Demand Jets", *IBM J. Res. Dev.* **28** (1984) no. 3, 322.
156. F. C. Lee: "Pressure-Wave Generation in Impulse Ink-Jet Head Using Elongated Fluid-Filled Piezoelectric Transducer", *SID Digest*, **18** (1987) 186.
157. H. Gerhauser *et al.*: "The Effect of Pulse Shape on the Drop Volume and the Frequency Response of Drop-on-Demand Ink Jet Transducer", *SID Digest*, **14** (1983) 110.
158. F. C. Lee: "Drop Break-Off Dynamics of Drop-on-Demand Ink Jet", *The Second International Congress on Advances in Non-Impact Printing Technologies*, SPSE, Springfield, USA Nov. 4 – 8, 1984, 9.
159. Hewlett-Packard Journal 36, May, 1985.
160. E. Manini *et al.*: "The Liquid Spark Ink Jet Technology", *The Third International Congress on Advances in Non-Impact Printing Technologies*, August, 1986, p. 145.
161. G. Olive *et al.*: "Materials and Processing Studies for Thermal Ink-Jet Devices", *SID Digest* **17** (1986) 105.
162. L. Chang, G. Olive: "Factors Influencing the Lifetime of Thermal Ink-Jet Heaters", *SID Digest* **18** (1987) 192.
163. A. R. Forouhi, J. Eldridge: "Prediction of Thermal Ink-Jet Lifetime from Chemical Dissolution Rate of Passivation Layers", *SID Digest* **18** (1987) 196.
164. E. Bassous *et al.*: "Ink Jet Printing Nozzle Arrays Etched in Silicon", *Appl. Phys. Lett.* **31** (1977) 135. Links
165. A. R. Lin: "Advances in Ink for Ink Jet Technology", *Annual Guide to Ribbons & Toners*, 1985, p. 51.
166. P. Gregory: "High-Technology Applications of Organic Colorants, Topics in Applied Chemistry", Plenum Press, New York 1991.
167. Canon, US 4 585 484, 1986 (M. Haruta, T. Nakousai, K. Iwata, K. Nakousai).
168. J. S. Campbell, P. Gregory, P. Mistry, Proceedings of the IS&T's Seventh International Congress on Advances in Non-Impact Technologies. Springfield, Virginia, IS&T 1991, vol. 2, pp. 142 – 145.

169. US 5 203 912, 1993 (D. Greenwood, N. Hughes, R. W. Kenyon, S. L. Hindagolla).
170. Lexmark, US 5 719 204, 1998 (B. L. Beach, A. M. Pearson, J. X. Sun, R. B. Watkins).
171. Du Pont, US 5 712 338, 1998 (D. H. Donovan, G. A. Lane, C. C. Tang).
172. Hewlett-Packard, US 5 745 140, 1998 (J. L. Stoffel *et al.*).
173. Hewlett-Packard, US 5 534 051, 1996 (H. P. Lauw).
174. Seiko Epson JP HEISEI4–164975, 1992; *Chem. Abstr.* **117** (1992) no. 18, 173524m.
175. Hewlett-Packard, US 5 536 306, 1996 (L. E. Johnson *et al.*).
176. Tektronix, US 5 084 099, 1992 (C. W. Jaeger, C. F. Sheley).
177. Tektronix, US 5 372 852, 1994 (D. R. Titterington, L. V. Bui, L. M. Hirschy, H. R. Frame).
178. J. Maed: “Color Printer Performance Is Improved, But Demand Had Yet to Blossom”, Sanyo Electric Company, OEP, May, 1985.
179. W. Crooks, E. W. Luttman, A. B. Jaffe: “High Quality Color Printing with Continuous Ink-Jet”, *SPSE 1st International Conference on Advances in Non-Impact Printing Technologies*, Venice, Italy June, 1981.
180. C. Rainwater: *Light and Color*: Golden Press, Western Publishing Company 1971.
181. F. W. Billmeyer, Max Saltzman: *Principles of Color Technology*, Wiley, New York 1966.
182. D. B. Judd, G. Wyszecki: *Color in Business, Science, and Industry*, Wiley, New York 1975.
183. E. Coates: “Aggregation of Dyos in Aqueous Solutions”, *ISDC*, 1969, 355 – 368.
184. IBM Report, P. Stucki, IBM Zurich Research Laboratory: “Mecca—A Multiple Error Correction Compalation Algorithm for Bi-Level Image Hardcopy Reproduction”, Switzerland.
185. IBM internal report, E. W. Luttman, ISD Advanced Technology: “Color Reproduction with Ink Jet”, San Jose, California.
186. J. Kosar: *Light Sensitive Systems*, Wiley Interscience, New York – London – Sidney 1965.
187. M. S. Dinaburg: *Photosensitive Diazo Compounds*, The Focal Press, London – New York 1964.
188. J. Munder: “Plenarvorträge”, *4th Intern. Kongress für Reprografie und Information*, Hannover 1975, p. 120; AWV Fachbericht 349, Frankfurt/Main 1975. Ausschuß für wirtschaftliche Verwaltung in Wirtschaft und öffentlicher Hand.
189. J. Epperlein *et al.*, *J. Signalaufzeichnungsmaterialien* **11** (1983) no. 6, 403 – 424.
190. H. Zollinger: *Azo and Diazo Chemistry, Aliphatic and Aromatic Compounds*, Chap. 10, 11, Interscience, New York 1961.
191. O. Frank: *Die Lichtpaustechnik*, Dorotheen-Verlag, Stuttgart 1959.
192. Océ-van der Grinten, DE 23 25 579, 1973.
193. Adressograph-Multigraph Corp., DE 15 22 449, 1966.
194. R. Dessauer, *Chimia* **29** (1975) no. 11, 467 – 469.
195. E. Brinkman, G. Delzenne, A. Poot, J. Willems: *Unconventional Imaging Processes*, The Focal Press, London – New York 1978, pp. 68 – 73.
196. The Mead Corporation, US 4 399 209, 1983 (F. W. Sanders, G. F. Hillenbrandt); US 4 440 846, 1984 (J. S. Arney, R. F. Wright) R. F. Wright).
197. E. A. Rott: *Photographic Silver Halide Diffusion Transfer Processes*, Focal Press London – New York 1972.
198. E. H. Land, H. G. Rogers, V. K. Walworth in J. M. Sturge (ed.): *Neblette's Handbook of Photography and Reprography*, 7th ed., Van Nostrand, New York – London 1977.
199. L. W. Tregilius *et al.*, US 3 179 517, 1965; R. Tarkinson, *Photogramm. Eng.* **31** (1965) no. 1, 126 – 130.
200. W. R. Hawken, *Copying Methods Manual*, LTP Publications no. 11, American Library Association, Chicago 1966.
201. W. Eisbein, US 2 657 618, 1953.
202. G. P. Levenson in T. H. James (ed.): *Theory of the Photographic Process*, “Diffusion transfer & mono-baths”, Macmillan Publ. Co., New York 1977.
203. H. Vermeulen, A. Poot, J. W. Van den Houte, US 435 7407, 1982.
204. R. M. De Keyzer EP-A 87 2017 009, 1989.
205. Y. Tsubai, K. Okazaki, S. Horii, A. Yoshida, GB 2 159 968 A, 1985.
206. D. H. Schultze: “Graphic arts applications of silver diffusion transfer systems”, *J. Appl. Photogr. Eng.* **5** (1979) no. 3, 163.
207. J. Velek *et al.*, *Collect. Czech. Chem. Commun.* **41** (1976) 1419. Links
208. Agfa-Gevaert, DE-OS 2 654 213, 1977 (E. Meier, W. Laessig).
209. Agfa-Gevaert US 4 232 107, 1980 (W. Janssen).
210. M. Peters, *Lecture given on SPSE congress in Boston*, MA, USA, on May 20 – 24, 1984. *J. Imag. Technol.* **11** (1985) 101.
211. G. A. Plunka (ed.): *Micrographic Film Technology*, The National Micrographics Association, Silver Spring, Maryland, USA 1983.
212. Kalle, DE 559 795, 1932 (W. Krieger).
213. H. H. Mc Gregor, *J. Microgr.* **9** (1975) 13 – 20.

General References

214. J. Yule: *Principles of Color Reproduction*, J. Wiley & Sons, New York 1967.
215. M. H. Bruno: *Pocket Pal—A Graphic Arts Production Handbook*, International Paper Company, New York 1983.
216. H.-J. Teichmann, *et al.*: *Neue technische Kommunikation in der Druckindustrie*, Bundesverband Druck E.V., Wiesbaden 1982.

Specific References

217. H. Philippaerts, R. Vanesch, E. Hofman, J. Verelst, *Dtsch. Drucker* **19** (1983) no. 28, 38 – 42.
218. E. Dutoit, *Druckspiegel* **42** (1987) no. 8, 801 – 802.
219. H. Zwicky, *Polygraph* **38** (1985) no. 18, 1348, 1350, 1352 – 1354.
220. A. Rott, E. Weyde: *Photographic Silver Halide Diffusion Processes*, Focal Press, London 1972.
221. J. Verelst, *Chimia* **34** (1980) no. 4, 190 – 199. Links
222. J. Heller: *Printmaking Today*, Holt, Rinehart & Winston, New York 1972.
223. W. Stupp, *MFM Mod. Fototech.* **27** (1979) no. 6, 290 – 292.
224. E. Born: *Handbuch der Rasterfotografie*, Verlag Ambipress, Basel 1983.
225. in [215] pp. 79 – 80.
226. P. Müller, *Druckspiegel* **40** (1985) no. 7, 682, 684 –685.
227. E. Fritz, *Polygraph* **38** (1985) no. 22, 1740, 1742, 1747, 1750, 1752.
228. D. Spencer, *Seybold Report on Publishing Systems* **16** (1987) no. 11, 3 – 20.
229. E. Seubert, *Druckspiegel* **39** (1984) no. 2, 152 – 158.
230. J. Van der Schans *et al.*: *Working with Color*, Elsevier, Amsterdam 1984.
231. M. H. Bruno: *Principles of Color Proofing*, Gamma Communications, Salem 1986.
232. H. Zwicky, *Fachh. Chemigr. Lithogr. Tiefdruck* (1983) no. 4, 183 – 186.
233. N. N., *Druckspiegel* **39** (1984) no. 5, 486 – 489.
234. J. Birkenshaw: “Trends in Text and Image Processing”, *J. Photogr. Sci.* **32** (1984) 219 – 220.
235. P. Dyson, J. Seybold, W. Solimeno, D. Stivison, *Seybold Report on Publishing Systems* **14** (1985) no. 19, 3 – 26.
236. W. Fraß, *Chem. Unsere Zeit* **17** (1983) 10 – 20, 33 – 40.
237. O. Süs, *Justus Liebigs Ann. Chem.* **556** (1944) 65 – 84.
238. Kalle AG, DE 1 543 721, 1966 (O. Süs, F. Uhlig, H. Schäfer).
239. Polychrome Corp., DE 1 522 478, 1965; US 3 635 709, 1969 (K. Kobayashi).
240. M. J. Bowden, L. F. Thompson, S. R. Fahrenholtz, E. M. Doerries, *J. Electrochem. Soc.* **128** (1981) 1304 – 1313. Links
241. BASF, DE 2 150 691, 1973 (H. Barzynski, M.-J. Jun, D. Sänger, O. Volkert).
242. H. Barzynski, D. Sänger, *Angew. Makromol. Chem.* **93** (1981) 131 – 141. Links
243. American Can Co., DE-OS 2 329 407, 1972 (Sh. I. Schlesinger, V. Cochran).
244. Kalle AG, DE 2 331 377, 1973 (P. Stahlhofen, R. Beutel).
245. Hoechst AG, DE 2 718 259, 1977 (G. Buhr).
246. Fuji Photo Film Co. Ltd., DE 2 851 471, 1977 (M. Iwashaki); DE 2 851 472, 1977 (M. Iwashaki, S. Sato, N. Yasuo).
247. Hoechst AG, DE 2 610 842, 1975 (G. Buhr, H. Ruckert, W. Fraß).
248. Hoechst AG, DE 2 718 254, 1977 (J. Sander, G. Buhr, H. Ruckert).
249. Hoechst AG, EP 0 006 626, 1978 (J. Sander, G. Buhr, H. Ruckert).
250. Hoechst AG, EP 0 006 627, 1978 (J. Sander, D. Skaletz, G. Buhr, G. Lohaus).
251. Minnesota Mining and Manuf. Co., US 3 779 778, 1972 (G. H. Smith, J. A. Bonham).
252. J. Kosar, *Light-Sensitive Systems, Chemistry and Application of Nonsilver Halide, Photographic Processes*, J. Wiley & Sons, New York – London – Sydney 1965.
253. B. Duncalf, A. S. Dunn, *J. Appl. Polym. Sci.* **8** (1964) 1763 – 1776. Links
254. Kalle AG, DE 581 697, 1931 (R. Zahn).
255. Kalle AG, DE 2 024 242, 1969 (H. Steppan).
256. O. Süs, K. Möller, H. Heiss, *Liebigs Ann. Chem.* **598** (1956) 123 – 138.
257. O. Süs, J. Munder, H. Steppan, *Angew. Chem.* **74** (1962) 985 – 988.
258. Howson-Algraphy Ltd., DE 2 251 828, 1971 (A. P. Gates).
259. Eastman Kodak Co., GB 695 262, 1953; US 2 725 372, 1955 (L. M. Minsk).
260. Eastman Kodak Co., DE-AS 2 717 778, 1976 (J. A. Vanallen, M. P. Cunningham, D. P. Specht, S. Y. Farid).
261. W. R. Grace u. Co., US 3 661 744, 1972; US 3 697 395, 1972; US 3 697 402, 1972; US 3 700 574, 1972 (C. L. Kehr, W. R. Wszolek).
262. H. Steppan: “Photopolymerisierbare und photovernetzbare Systeme”, *Plenarvorträge auf dem 4. Internationalen Kongreß für Reprographie und Information 1975,* 13. – 17. 04. 1975, Hannover, 147 – 162.
263. Imperial Chemical Industries Ltd., GB 566 795, 1945 (W. E. F. Gates).
264. D. W. Woodward, V. C. Chambers, A. B. Cohen, *Photographic Science and Engineering* **7** (1963) 360 – 368.
265. DuPont, US 3 261 686, 1966 (J. R. Celeste, S. Bauer).
266. P. Walker, V. J. Webers, G. A. Thommes, *J. Photogr. Sci.* **18** (1970) 150 – 158.
267. R. M. Schaffert: *Electrophotography*, Focal Press. London, 5th ed. 1980.
268. J. W. Weigl, *Angew. Chem.* **89** (1977) 386 – 406.
269. H. J. Gerritsen, W. Ruppel, A. Rose, *Helv. Phys. Acta* **30** (1957) 504 – 512.
270. K. Hauffe, *J. Photogr. Sci.* **10** (1962) 321 – 330.
271. R. B. Comizzoli, D. A. Ross, *Photogr. Sci. Eng.* **13** (1969) 265 – 270.
272. M. R. Kuehnle, *J. Appl. Photogr. Eng.* **4** (1978) 155 – 159.

273. Kalle & Co. AG, DE 1 058 836, 1956 (W. Neugebauer, M. Tomanek, H. Behmenburg).
274. Xerox Corp., DE 1 497 205, 1969 (J. F. Byrne).
275. W. Wiedemann: “Organische Photoleiter – Ein überblick”, *Chem. Ztg.* **106** (1982) 3 – 14; *Chem. Ztg.* **106** (1982) 15 – 29.
276. *Winnacker-Küchler*, 4th ed., **7, Organische Technologie III**, 537 – 562.
277. M. H. Bruno: *Principles of Color Proofing*, Gama Communications, Salem N. H. 1986.
278. G. Werner: *Standardisierter Mehrfarben-Offsetdruck*, Polygraph Verlag, Frankfurt/M. 1982.
279. Minnesota Mining and Manuf. Co., US 3 136 637, 1958 (G. W. Larson); DE 1 912 801, 1968 (P. C. van Beusekom).
280. Azoplate Corp., DE 2 032 947, 1969 (H. G. Fernekess, T. N. Gillich).
281. Kalle AG, DE 1 291 197, 1963 (F. Endermann, F. Uhlig).
282. DuPont, US 4 489 153, 1983 (R. W. Ashcraft, J. G. Moehlmann).
283. Minnesota Mining and Manufacturing Co., EP-A 0 115 899, 1983 (P. M. Koelsch, R. I. Krech).
284. American Hoechst Corp., US 4 650 738, 1984 (St. J. W. Platzer, G. I. Koletar).
285. American Hoechst Corp., US 4 659 642, 1984 (St. J. W. Platzer, G. I. Koletar, R. L. Shadrach).
286. DuPont, US 3 649 268, 1969 (V. F. H. Chu, A. B. Cohen).
287. DuPont (Deutschland) GmbH, DE 3 625 014, 1986 (H. Fröhlich).
288. Keuffel and Esser, US 4 376 158, 1977 (D. S. Speckler).
289. DIN 16 544, Begriffe der Reproduktionstechnik im graphischen Gewerbe, Jan. 1966.
290. Jahresbericht 1986 des Bundesverbandes Druck e.V., Wiesbaden, p. 12.
291. *Winnacker-Küchler*, 4th ed., **7, Org. Technologie III**, 524 – 526.
292. Polychrome Corp., US 3 891 516, 1967 (S. L. Chu).
293. Hoechst AG, US 4 666 576, 1985; US 4 661 219, 1985 (E. Pliefke).
294. Vickers Ltd., DE 2 816 307, 1977 (M. Ould, G. N. Stevens).
295. Davidson Manufacturing Corp., US 2 344 510, 1944 (W. T. Hagelin). Polychrome Corp., GB 1 582 620, 1977 (S. L. Chu, E. Golda, J.-Ch. Huang, A. L. Wilkes).
296. A. W. Brace, P. G. Sheasby: *The Technology of Anodizing Aluminium*, Technicopy Limited, Stonehouse, Glos. GL 10 2 NA, GB 1968.
297. Hoechst AG, DE 2 811 396, 1978 (G. Usbeck).
298. W. L. Howson Ltd., CA 875 681, 1968 (L. Watkinson, B. Moore).
299. Hoechst AG, DE 2 836 803, 1978 (G. Usbeck).
300. Minnesota Mining and Manufacturing Co., US 2 714 066, 1954 (C. L. Jewett, J. M. Case).
301. Kalle AG, DE 1 621 478, 1967 (G. Berghäuser, F. Uhlig).
302. Polychrome Corp., US 3 440 050, 1965.
303. Hoechst, DE 3 206 470, 1982 (D. Mohr).
304. Minnesota Mining and Manuf. Co., DE 1 091 433, 1954/1955 (C. L. Jewett, J. M. Case, J. F. Dowdall).
305. in [291] pp. 527 – 532.
306. Kalle AG, DE 1 447 963, 1965 (O. Velten, F. Uhlig).
307. Vickers Ltd., GB 1 513 368, 1974 (H. Wielinga).
308. Fuji Photo Film Co. Ltd., DE-OS 2 461 912, 1974 (M. Watanabe); Hoechst, EP 0 131 238, 1983 (P. Stahlhofen, G. Mack); EP 0 133 216, 1983 (P. Stahlhofen).
309. Toray Industries Inc., US 347 303, 1978 (M. Asano, Y. Mori, Y. Takayama).
310. Toray Industries Inc., US 4 342 820, 1979 (T. Kinashi, T. Fujita, N. Kawabe).
311. H. W. Vollmann, *Angew. Chem.* **92** (1980) 95 –106.
312. Kalle AG, DE 1 117 391, 1959 (F. Uhlig).
313. Minnesota Mining and Manuf. Co., US 3 563 734, 1971 (B. L. Shely).
314. James River Graphics Inc., US 4 245 003, 1979 (R. L. Oransky, D. G. Savage).
315. BASF, DE 3 226 779, 1982 (J. Lynch, R. Vyvial, M. Zuerger, K. Borho).
316. BASF, DE 3 908 763, 1990 (T. Telser, W. Hümmer, K. Kurtz).
317. BASF, DE 2 202 357, 1990 (C. Srna *et al.*); DE 1 597 515, 1967 (H. Henkler, H. Hartmann, K. Gulbins, H. Wilhelm).
318. BASF, EP 0 470 071, 1992 (T. Telser *et al.*). DuPont, DE 3 828 551, 1988 (M. Schober *et al.*).
319. Asahi, EP 0 017 927, 1984 (S. Nakamura *et al.*).
320. D. Finna, V. Jansen, R. Michels, *Flexo + Tiefdruck* **2** (1995).
321. J. Kosar: *Light Sensitive Systems*, J. Wiley & Sons, New York 1965.
322. H. Barzynski, K. Penzien, O. Volkert, *Chem. Ztg.* **96** (1972) no. 10, 545 – 51.
323. DuPont, US 2 760 863, 1952 (L. Plambeck).
324. W. E. F. Gates, CA 441 538 (1947).
325. H. J. Timpe, H. Baumann: *Photopolymere*, Dt. Verlag für Grundstoffindustrie, Leipzig 1988.
326. BASF, DE-OS 2 232 497 1972 (M.-J. Jun, M. Fischer, O. Volkert, E. Hickmann); DE-OS 2 232 498, 1972 (E. Hickmann); DE-OS 2 909 992, 1979 (P. Lechtken *et al.*).
327. DuPont, DE 2 215 090, 1981 (G. Chen).

328. Nippon Paint, US 3 801 328 1974 (Y. Takimoto *et al.*).
329. BASF, DE-OS 1 522 463, 1966 (G. Faulhaber *et al.*).
330. W. R. Grace, US 3 700 574, 1970 (C. L. Kehr, S. Spring, W. R. Wszolek);
331. Asahi, DE 2 815 678, 1980 (S. Nakamura *et al.*).
332. BASF, EP 0 027 612, 1982 (G. Heinz, P. Richter, M. Jun); DE 3 803 457, 1988 (K. Kurtz, H. Koch, T. Telser, M. Zürger).
333. Toyo Boseki, EP 0 273 393, 1988 (A. Tomita *et al.*).
334. BASF, DE 3 540 950, 1987 (H. Koch, W. Ziegler).
335. Nippon Paint, EP 0 261 910, 1994 (K. Ishikawa, K. Konishi, H. Kusuda).
336. DuPont, EP 0 356 953, 1990 (M. Fryd, E. Leberzammer).
337. BASF, DE-OS 2 722 421, 1977 (B. Bronstert, W. Kuesters, M. Jun, M.-J. Jun, G. Wallbillich).
338. DuPont, US 5 262 275, 1993 (R. N. Fan).
339. BASF, DE-OS 19 53 68 05.3, 1995 (T. Loerzer *et al.*).
340. R. Golpon: *Lehrbuch der Druckindustrie*, “Tiefdruck”, vol. 5, Polygraph Verlag, Frankfurt/M. 1974.
341. O. Leidung (ed.): *Technical Guide for the Gravure Industry*, Gravure Technical Association, Inc., New York, N.Y. 1975.
342. E. D. Stiebner: *Bruckmann's Handbuch der Drucktechnik*, 4. ed.. Bruckmann, München 1986.
343. F. Bauer: *Lexikon der Reproduktionstechnik*, Verlag Beruf + Schule, Itzehoe 1986.
344. K.-A. Springstein: *Elektronische Bildverarbeitung von A – Z*, Verlag Beruf + Schule, Itzehoe 1982.
345. W. Boppel: Die Elektronenstrahlgravur von Hell zur Herstellung von Tiefdruckzylindern, *Dtsch. Drucker* **16** (1986) 44 – 49.
346. K.-A. Springstein: “Erfolgreicher Praxiseinsatz des ersten ‘Lasergravure-700'-Systems in England.” *Dtsch. Drucker* **25** (1984) 48 – 60.
347. J. V. Duppen: *Handbuch für den Siebdruck*, Verlag Der Siebdruck, D-2400 Lübeck 1986.
348. O. Zeman, *IS + L, Fachzeitschrift für industriellen Siebdruck + Leiterplattentechnik* **5** (1984) 24 – 31.
349. Agency of Industrial Science and Technology, MITI, DE 2 932 376, 1978 (Ichimura, Konihiro).
350. see [347-349]
351. DuPont, EP-A 4383, 1979 (G. R. Nacci, D. G. Pye).
352. W. S. De Forest: *Photoresist, Materials and Processes*, McGraw-Hill, New York 1975.
353. H. Steppan, G. Buhr, H. Vollmann, *Angew. Chem.* **94** (1982) 471 – 485.
354. C. G. Roffey: *Photopolymerization of Surface Coatings*, Chap. 6, John Wiley & Sons, Chichester, Reprint 1985, pp. 263 – 304.
355. S. J. Moss, *Polym. Degrad. Stab.* **17** (1987) 205 – 222.
356. G. N. Taylor, T. M. Wolf, L. Stillwagon, *Solid State Technol.* (1984, Feb.) 145 – 155.
357. N. G. Einspruch, G. B. Larrabee (eds.): “Materials and Process Characterization”, *VLSI Electronics Microstructure Science*, vol. **6**, "Chap. 4", Academic Press, New York 1983, pp. 217 – 253.
358. M. Tsuda, *J. Imaging Technol.* **11** (1985) 158 – 163.
359. F. Coopmans, B. Roland, *Solid State Technol.* (1987, June) 93 – 99.
360. R. Srinivasan, *Polym. Degrad. Stab.* **17** (1987) 193 – 203.
361. L. F. Thompson, C. G. Willson, J. M. J. Fréchet (eds.): “Materials for Microlithography-Radiation-Sensitive Polymers”, *ACS Symp. Ser.* **266** (1984) 39 – 117.
362. L. F. Thompson, C. G. Willson, M. J. Bowden (eds.): “Introduction to Microlithography—Theory, Materials, and Processing”, *ACS Symp. Ser.* **219** (1983) 87 – 159.
363. J. Bargon (ed.): *Methods and Materials in Microelectronic Technology*, Plenum Press, New York 1984, pp. 181 – 241.
364. E. R. Howells (ed.): *Technology of Chemicals and Materials for Electronics*, Chap. 8, Ellis Horwood Ltd., Chichester 1984, pp. 128 – 148.
365. Grace, GB 2 091 493, 1982 (F. J. Rendulic, R. K. Trasavage, P. A. Boduch).
366. R. H. Wopschall, *Solid State Technol.* (1986, June) 153 – 157.
367. J. E. Jensen, *Solid State Technol.* (1984, June) 145 – 150.
368. S. T. Kowel *et al.*, *Thin Solid Films* **152** (1987) 377 – 403.
369. M. Hummel: *Einführung in die Leiterplattentechnologie*, E. G. Leuze Verlag, Saulgau/Württ. 1985.
370. G. W. W. Stevens, *J. Appl. Photogr. Eng.* **8** (1982) p. 90 – 100.
371. “Galvanoformung, Berichtsband über das 8. Ulmer Gespräch der Deutschen Gesellschaft für Galvano- und Oberflächentechnik”, 24./25. April 1986 in Neu-Ulm, E. G. Leuze Verlag, Saulgau/Württ., 1986.
372. W. Becht, W. Ehrfeld, A. Maner, D. Schmidt, *Galvanotechnik* **77** (1986) 2695 – 2705.
373. D. J. Elliott: *Integrated Circuit Fabrication Technology*, McGraw-Hill, New York 1982.
374. L. K. White, *RCA Rev.* **47** (1986) 345 – 379.
375. W. R. Brunsvold, D. M. Crockatt, G. J. Hefferon, C. F. Lyons, *Opt. Eng.* **26** (1987) 330–336. Links
376. J. B. Angell, S. C. Terry, P. W. Barth, *Sci. Am.* **248** (1983) no. 4, 36 – 47. Links
377. M. Hepher, *J. Photogr. Sci.* **12** (1964) 181 – 190.
378. R. Phillips: *Sources and Applications of Ultraviolet Radiation*, Academic Press, London 1983, pp. 33 – 37, 56 – 111.
379. W. Schnabel, H. Sotobayashi, *Prog. Polym. Sci.* **9** (1983) 297 – 365. Links

380. T. Davidson (ed.): “Polymers in Electronics”, *ACS Symp. Ser.* **242** (1984) Beiträge 1 – 21.
381. Landolt-Börnstein: *Zahlenwerte und Funktionen aus Naturwissenschaften und Technik,* Neue Serie, Gruppe III: “Kristall- und Festkörperphysik”, vol. 17, Teilband c, Springer-Verlag, Berlin – Heidelberg pp. 250 – 280.
382. H. Meier, K.-P. Zeller, *Angew. Chem.* **87** (1975) 52 – 63.
383. O. Süs: *Justus Liebigs. Ann. Chem.* **556** (1944) 65 – 84.
384. J. Pacansky, J. R. Lyerla, *IBM J. Res. Dev.* **23** (1979) 42 – 55. Links
385. in [362] pp. 122 ff.
386. A. S. Gozdz, *Solid State Technol.* (1987, June) 75 – 79.
387. IBM, US 4 491 628, 1982 (H. Ito, C. G. Willson, J. M. J. Fréchet).
388. S. P. Pappas, *J. Imaging Technol.* **11** (1985) 146 – 157.
389. M. Hashimoto, T. Iwayanagi, H. Shiraishi, S. Nonogaki, *Polym. Eng. Sci.* **26** (1986) 1090 – 1095. Links
390. W. L. Dilling, *Chem. Rev.* **83** (1983) 1–47. Links
391. Kodak, US 1 973 493, 1932 (C. J. Malm, C. R. Fordyce).
392. Kodak, GB 695 197, 1949 (L. M. Minsk, W. P. Vandeusen).
393. A. Reiser, P. L. Egerton, *Photogr. Sci. Eng.* **23** (1979) 144 – 150.
394. Ciba-Geigy, DE 2 342 407, 1973 (S. A. C. Zahir, H. Rembold, E. Losert).
395. D. W. Woodward, V. C. Chambers, A. B. Cohen, *Photogr. Sci. Eng.* **7** (1963) 360; J. Kosar: *Light-Sensitive Systems: Chemistry and Application of Nonsilver Halide Photographic Processes*, Wiley, New York 1965 – 1970, "Chap. 3, 4 and 5"; P. Walker, V. J. Webers, G. A. Thommes, *J. Photogr. Sci.* **18** (1970) 150; H. Barzynski, K. Penzien, O. Volkert, *Chem.-Ztg.* **96** (1972) 545; H. Steppan, *4. Ing. Kongr. Reprogr. Inf.* Hannover 1975, p. 147; A. Ledwith, *Pure Appl. Chem.* **49** (1977) 431; Links G. A. Delzenne, *Adv. Photochem.* **11** (1979) 1; *Makromol. Chem. Suppl.* **2** (1979) 169; J. L. R. Williams, R. C. Daly, S. Y. Farid, *Euchem. 80, Conf. Photopolymer. Photocrosslink. Org. Coat.*, Stockholm 1980; H. Böttcher, *J. Signalaufzeichnungsmaterialien* **8** (1980) 405.
396. in [378] pp. 56 – 111.
397. H. J. Hageman, *Prog. Org. Coat.* **13** (1985) 123 – 150.
398. DuPont, BE 685 011, 1966 (J. R. Celeste).
399. Dynachem Corp., US 4 610 951, 1985 (M. A. Lipson, G. A. Derrico, S. Y. Tark, T. Yamazaki).
400. Fuji Photo Film Co., DE-OS 3 601 997, 1986 (K. Maemoto *et al.*).
401. R. Rubner, H. Ahne, E. Kühn, G. Kolodziej, *Photogr. Sci. Eng.* **23** (1979) 303; H. J. Boos, G. Kolodziej, R. Rubner, *Galvanotechnik* **68** (1977) 984; Siemens, DAS 2 462 105, 1979 (R. Rubner, W. Kleeberg, E. Kühn).
402. G. A. Delzenne, *Adv. Photochem.* **11** (1979) 1; C. L. Kehr, W. R. Wszolek, *Am. Chem. Soc. Div. Org. Coat. Plast. Chem. Pap.* **33** (1973) 295. C. R. Morgan, A. D. Ketley, *ibid.* **33** (1973) 281.
403. in [373] Chapter 13, pp. 311 – 337.
404. R. H. Wopschall: *Proceedings of the Technical Program National Electronic Packaging and Production Conference*, Anaheim, California March 1 – 3, 1983, pp. 11 – 15.
405. C. Tiby, *Chem. Ind. (Düsseldorf)* **109** (1986) 1044 – 1050.
406. in [354] Chap. 7, pp. 305 – 338.
407. S. G. Wentink, S. D. Koch (eds.): *UV Curing in Screen Printing for Printed Circuits and the Graphic Arts*, Chap. 9, Technology Marketing Corp. Norwalk, Conn. USA 1981, pp. 221 – 242.
408. in [378] pp. 406 – 425.
409. J. C. Knaff, W. Liebsch, *Galvanotechnik* **75** (1984) 115 – 120.
410. J. Eason, R. DeBisschop, *Galvanotechnik* **75** (1984) 936 – 940.
411. U. v. Mylius, J. Leudolph, *Galvanotechnik* **77** (1986) 402 – 406.
412. H. Dengler, W. Röcker, *Galvanotechnik* **77** (1986) 2237 – 2240.
413. K. G. Faas, J. Swozil (eds.): *Verdrahtungen und Verbindungen in der Nachrichtentechnik*, Akade-mische Verlagsgesellschaft Frankfurt am Main 1974, pp. 196 – 250.
414. J. A. Scarlett (ed.): *The Multilayer Printed Circuit Handbook*, Electrical Publications Ltd. 1985; *cf. Galvanotechnik* **77** (1976) 2196.
415. in [413] pp. 66 – 195.
416. J. A. Scarlett: *Printed Circuit Boards for Microelectronics*, 2nd ed., Chap. 2, Electrochemical Publications Ltd., Ayr (Scotland) 1980, pp. 19 – 80.
417. H. Müller: *Konstruktive Gestaltung und Fertigung in der Elektronik,* vol. 1, Chap. 3.2, “Bildübertragung”, pp. 46 – 67, Chap. 3.3, “Verfahren zur Metallstrukturierung”, pp. 68 – 96, Vieweg & Sohn, Braunschweig/Wiesbaden 1981.
418. J. A. Scarlett: *An Introduction to Printed Circuit Technology*, Electrochemical Publications Ltd., Ayr (Scotland) 1984.
419. G. Herrmann *et al.*: *Handbuch der Leiterplattentechnik*, “Laminate – Manufacturing – Assembly – Test, “2nd ed., Chap. 1 – 23, E. G. Leuze Verlag, Saulgau (Württ.) 1982.
420. R. H. Clark: *Handbook of Printed Circuit Manufacturing*, Van Nostrand, New York 1985.
421. in [419] Chap. 15, pp. 274 – 283.
422. A. Gemmler, J. L. Jostan, *Galvanotechnik* **77** (1986) 51 – 60.
423. P. Lund: *Präzisionsvorlagen für Leiterplatten*, Hanser Verlag, München – Wien 1979.
424. in [417] pp. 20 – 46.
425. S. A. Partridge, *Circuit World* **13** (1987) no. 2, 4 – 13.
426. in [419] Chap. 6, pp. 90 – 124, Chap. 7, pp. 125 – 151.

427. in [407] Chap. 6, pp. 133 – 156.
428. W. Peters, *Galvanotechnik* **74** (1983) 1395 – 99.
429. C. W. Ruff, *Metalloberfläche* **41** (1987) 135 – 139.
430. P. Britton, *Circuit World* **13** (1986) no. 1, 4 – 7.
431. J. Tuck, *Circuits Manufacturing* (1987) 24 – 33.
432. A. Gemmler, J. L. Jostan, *Galvanotechnik* **77** (1986) 1182 – 88.
433. R. Germany, *Circuit World* **13** (1987) no. 3, 68 – 71.
434. in [419] Chap. 29, pp. 535 – 54.
435. C. Mead, L. Conway: *Introduction to VLSI Systems*, Addison-Wesley Publishing Reading 1980.
436. S. Wolf, R. Tumbler: *Silicon Processing for the VLSI Era, Process Technology*, Lattice Press, Sunset Beach 1987.
437. R. Muller, T. Kamins: *Device Electronics for Integrated Circuits*, J. Wiley, New York 1977.
438. W. S. DeForest: *Photoresist Materials and Processes*, McGraw Hill, New York 1975.
439. D. J. Elliott: *Integrated Circuit Fabrication Technology*, McGraw Hill, New York 1982.
440. L. F. Thompson, C. Willson, M. Bowden: "Introduction to Microlithography. Theory, Materials and Processing" *ACS Symp. Ser.* **219** (1983) 16.
441. L. White, D. Meyerhofer, *RCA Rev.* **44** (1983) 110.
442. B. Daniels, C. Szmanda, M. Templeton, P. Trefonas, *Proc. SPIE Int. Soc. Opt. Eng.* **631** (1986) 192.
443. J. Shaw. M. Frisch, F. Dill, *IBM J. Res. Dev.* **26** (1976) 220.
444. V. Marriott, C. Garza, M. Spak, *Proc. SPIE Int. Soc. Opt. Eng.* **771** (1987) 221.
445. A. Broers, *Solid State Technol.* **28** (1985) no. 6, 119. Links
446. K. Bartlett *et al.*, *Proc. SPIE Int. Soc. Opt. Eng.* **394** (1983) 49.
447. P. West, B. Griffing, *Proc. SPIE Int. Soc. Opt. Eng.* **394** (1983) 33.
448. F. A. Vollenbroeck *et al.*, *Microcircuit Eng.*, **245** (1985).
449. D. M. Allen: *The Principles and Practice of Photochemical Machining and Photoetching*, Adam Hilger, Bristol 1986.
450. D. M. Allen, P. J. Gillbanks, *Proc. Tech. Program Nat. Electron. Packag. Prod. Conf.* 1986, 487 – 494.
451. D. M. Allen, *Proc. Tech. Program Nat. Electron. Packag. Prod. Conf.* 1987, 583 – 589.
452. E. Makino, T. Sato, Y. Yamada, *Precis. Eng.* **9** (1987) 153 – 157.
453. P. M. Wible, *Photo Chem. Mach. Inst. J.* **6** (1981) 7 – 11.
454. R. Markle, *Photo Chem. Mach. Inst. J.* **23** (1986) 10 – 12.
455. D. M. Allen, J. R. C. Brightman, P. J. Gillbanks, P. M. Smith, *Annals of the CIRP* **36/1** (1987) 91 –94.
456. D. M. Allen, P. J. Gillbanks, A. J. Palmer, *Proc. Int. Sym. Electric. Mach.*, ISEM-9, (1989) 246 –249.
457. Photofabrication Guide Book published by the Japan Photofabrication Association, Sapporo 1989.

[Top of Page]

Related Articles

⟶ Photography is a separate keyword.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Imidazole and Derivatives

Klaus Ebel, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2–6)
Hermann Koehler, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2–6)
Armin O. Gamer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 7.1)
Rudolf Jäckh, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 7.2)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Production 2
4.1. The Radziszewski Reaction 2
4.2. Dehydrogenation of Δ^2-Imidazolines 2
4.3. Functionalized Imidazoles by Modification of Simple Imidazoles 3
5. Quality Specifications and Analysis . 4
6. Uses 4
6.1. Intermediates for Herbicides and Pharmaceuticals 4
6.2. Other Uses 4
7. Toxicology and Occupational Health . 4
7.1. Imidazoles without Nitro Substituents 4
7.2. Nitroimidazoles 7
8. References 8

1. Introduction

Imidazole [*288-32-4*], $C_3H_4N_2$, M_r 68.10, was first synthesized in 1858 by Debus from ammonia and glyoxal; it was originally named glyoxalin. The name imidazole was introduced by Hantzsch. Industrial production of imidazole began in the 1950s; a wide range of derivatives is now available in industrial quantities.

2. Physical Properties

Some important physical properties of imidazole are listed in Table 1. Molecular masses, boiling points, melting points, and CAS registry numbers of the most important imidazole derivatives are listed in Table 2. Substitution at nitrogen generally causes a reduction in the melting and boiling points, because the N−H group is no longer available for intermolecular hydrogen bonding. Distillation of nitroimidazoles is accompanied by the risk of spontaneous decomposition.

Table 1. Physical properties of imidazole

Flash point	154 °C
d_4^{110}	1.0257
Bulk density	0.55 – 0.63 g/cm^3
Viscosity (100 °C)	2.696 mPa · s
Solubility in water∗	241
Enthalpy of fusion	12.96 kJ/mol
Enthalpy of vaporization (220 – 265 °C)	57.12 kJ/mol

∗ In grams per 100 g of solvent at 20 °C.

3. Chemical Properties

Imidazole is a moderately strong base (pK_b = 7.0 [1]), and a weak acid (pK_a = 14.9 [1]). Imidazoles substituted with electron-withdrawing groups are stronger acids than imidazole itself; e.g., 4(5)-nitroimidazole has a pK_a of 9.3 [1].

Imidazole is stable at 400 °C, possesses considerable aromatic character, and undergoes the usual electrophilic aromatic substitution reactions. Nitration and sulfonation require, however, far more drastic conditions than the corresponding reactions with benzene. Other substitution reactions of imidazole include halogenation, hydroxymethylation, coupling with aromatic diazonium salts, and carboxylation. Catalytic hydrogenation has not been described [1], [2].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a13_661

Table 2. Physical properties of imidazoles

	CAS registry no.	M_r	*mp*, °C	*bp*, °C
Imidazole	[*288-32-4*]	68.1	90	256
1-Methylimidazole	[*616-47-7*]	82.1	−1	198.5
2-Methylimidazole	[*693-98-1*]	82.1	145	270
4(5)-Methylimidazole	[*822-36-6*]	82.1	46	271
2-Ethylimidazole	[*1072-62-4*]	96.1	77	268
2-Isopropylimidazole	[*36947-68-9*]	110.1	134	278
2-Ethyl-4(5)-methylimidazole	[*931-36-2*]	110.1	45	260
2-Phenylimidazole	[*670-96-2*]	144.2	145	336
1-Vinylimidazole	[*1072-63-5*]	94.1		80∗
4(5)-Methyl-5(4)-hydroxymethylimidazole	[*29636-87-1*]	112.1	138	
4(5)-Nitroimidazole	[*3034-38-6*]	113.1	305∗∗	
2-Methyl-4(5)-nitroimidazole	[*696-23-1*]	127.1	244∗∗	
2-Isopropyl-4(5)-nitroimidazole	[*13373-32-5*]	155.2	184	
1-Methyl-5-nitroimidazole	[*3034-42-2*]	127.1	59	
1,2-Dimethyl-5-nitroimidazole	[*551-92-8*]	141.1	139	
1-Methyl-2-isopropyl-5-nitroimidazole	[*14885-29-1*]	169.2	56	
1-(2-Hydroxyethyl)-2-methyl-5-nitroimidazole	[*443-48-1*]	171.2	155	

∗ At 1.3 kPa.
∗∗ Decomposes.

4. Production

Of the many known methods for producing imidazoles, only the following are of industrial importance.

4.1. The Radziszewski Reaction

In the generally applicable Radziszewski reaction, a 1,2-dicarbonyl compound is condensed with an aldehyde and ammonia ($R^1 = H$) in a molar ratio of 1 : 1 : 2, respectively. Replacement of a molar equivalent of ammonia with a primary amine (R^1 = alkyl or aryl) leads to the corresponding 1-substituted imidazoles.

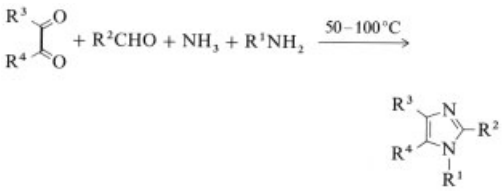

The reaction is usually carried out in water or a water – alcohol mixture at 50 – 100 °C. Work-up may involve the usual processes (e.g., distillation, extraction, and crystallization). Distillation leads to imidazole with a purity 99 %. The yield is generally 60 – 85 % [3]. This process allows the most important simple imidazoles to be prepared from the starting materials listed in Table 3.

4.2. Dehydrogenation of Δ^2-Imidazolines

Δ^2- Imidazolines can be obtained by several routes from 1,2-diamino compounds and carboxylic acid derivatives:

1) by reaction of a diamine with a carboxylic acid over an acidic heterogeneous catalyst (e.g., alumina – phosphoric acid) in the gas phase [4];
2) by reaction of a diamine with a carboxylic acid nitrile in the presence of sulfur [5] or copper salts [6] in the liquid phase; or
3) by preparation of diformyl derivatives from a diamine and formic acid esters, followed by conversion to imidazoline in the gas phase at 200 – 350 °C over a heterogeneous catalyst (e.g., zinc oxide – alumina) [7].

The Δ^2-imidazoline is then dehydrogenated in the gas phase over a precious metal at ca. 300 °C [8] or on alumina – zinc oxide at 400 – 500 °C [9]. In some cases this synthesis gives even better yields than the Radziszewski reaction (e.g., for 2-aryl-substituted imidazoles).

Table 3. Compounds used to prepare simple imidazoles by the Radziszewski reaction

Imidazole	Dicarbonyl compound	Aldehyde	Amine
Imidazole	glyoxal	formaldehyde	ammonia
1-Methylimidazole	glyoxal	formaldehyde	ammonia, methylamine
2-Methylimidazole	glyoxal	acetaldehyde	ammonia
2-Ethylimidazole	glyoxal	propionaldehyde	ammonia
2-Isopropylimidazole	glyoxal	isobutyraldehyde	ammonia
4(5)-Methylimidazole	methylglyoxal	formaldehyde	ammonia
2-Ethyl-4(5)-methylimidazole	methylglyoxal	propionaldehyde	ammonia

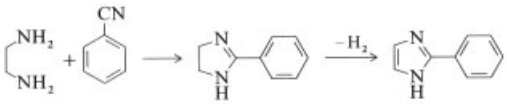

4.3. Functionalized Imidazoles by Modification of Simple Imidazoles

Nitroimidazoles. Reaction of imidazole with 60 – 80 % nitric acid in the presence of concentrated sulfuric acid at 120 – 150 °C leads to 4(5)-nitroimidazoles [10], [11]; nitration at C-2 has not been observed. 4(5)-Nitroimidazole, 2-methyl-4(5)-nitroimidazole, and 2-isopropyl-4(5)-nitroimidazole have been prepared by this method.

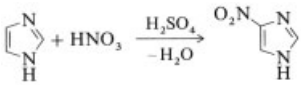

1-Substituted Imidazoles. Imidazole and alkylimidazoles can be alkylated with alcohols in the gas phase at 300 °C over acidic catalysts such as alumina – phosphoric acid or silica – phosphoric acid [12], [13].

In the liquid phase, imidazoles may be alkylated with dialkyl sulfates or alkyl halides in polar solvents such as alcohols and in the presence of a base, e.g., alkali-metal hydroxides or alkoxides [14].

Of particular importance for the preparation of 1-alkyl-5-nitroimidazoles is the treatment of 4(5)-nitroimidazoles with dialkyl sulfates in formic acid as solvent [15]. This method has been used for 1-methyl-, 1,2-dimethyl-, and 1-methyl-2-isopropyl-5-nitroimidazoles.

1-Vinylimidazole is produced by base-catalyzed addition of acetylene to imidazole in the liquid phase in an autoclave [16].

1-(2-Hydroxyethyl)-2-methyl-5-nitroimidazole (metronidazole) is obtained by addition of ethylene oxide to 2-methyl-4(5)-nitroimidazole in formic acid at 25 – 40 °C [17].

Michael addition of imidazole to acrylonitrile followed by catalytic hydrogenation of the intermediate 1-(2-cyanoethyl)imidazole [*23996-53-4*] leads to 1-(3-aminopropyl)imidazole [*5036-48-6*] [18].

Hydroxymethylimidazoles. 4(5)-Methyl-5(4)-hydroxymethylimidazole can be obtained by treatment of 4(5)-methylimidazole with aqueous formaldehyde under acidic [19] or basic [20] conditions.

Quaternary Salts. Treatment of 1-alkylimidazoles with benzyl chloride at 100 °C in solvent-free melt results in formation of 3-benzyl-1-alkylimidazolium chloride [21].

5. Quality Specifications and Analysis

Simple imidazoles are generally available in 99 % purity, which is particularly important for their use in pharmaceutical production. Purity is usually determined by reverse-phase HPLC; volatile derivatives may be analyzed by capillary column gas chromatography. A typical quality specification for imidazole is shown below:

Imidazole (min.)	99.5 %
2-Methylimidazole (max.)	0.2 %
Water (max.)	0.2 %
Other compounds (max.)	0.1 %

6. Uses

6.1. Intermediates for Herbicides and Pharmaceuticals

The imidazole ring is a constituent of several important natural products, including purine, histamine, histidine, and nucleic acids. Therefore, the bulk of imidazole produced is used in the preparation of biologically active compounds.

1-Substituted imidazoles such as *N*-propyl-*N*-[2-(2,4,6-trichlorophenoxy)ethyl]imidazole-1-carboxamide (prochloraz) [*67747-09-5*], 1-[2-(2,4-dichlorophenyl)-2-(2-propenyloxy)ethyl]-1*H*-imidazole (imazalil) [*35554-44-0*], and 1-(4-chlorophenoxy)-1-(1*H*-imidazolyl-1-yl)-3,3-dimethylbutanone (climbazole) [*38083-17-9*] are used as fungicides in crop protection [22]. Clotrimazole [*23593-75-1*] and bifonazole [*60628-96-8*] are used as antimycotics (→ Antimycotics).

4(5)-Methylimidazole and 4(5)-methyl-5(4)-hydroxymethylimidazole are used as intermediates for the H_2-blocker cimetidine [*51481-61-9*] (→ Antiulcer Drugs).

Nitroimidazoles are important chemotherapeutic agents [23]: e.g., 1,2-dimethyl-5-nitroimidazole (dimetridazole), 1-methyl-2-isopropyl-5-nitroimidazole (Ipronidazole), and 1-methyl-2-carbamoyloxymethyl-5-nitroimidazole (ronidazole) [*7681-76-7*] are used as veterinary products. Human chemotherapeutics based on 4(5)-nitroimidazole or 2-methyl-4(5)-nitroimidazole include metronidazole [*443-48-1*], tinidazole [*19387-91-8*], nimorazole [*6506-37-2*], ornidazole [*16773-42-5*], secnidazole [*3366-95-8*], and carnidazole [*42116-76-7*] (→ Chemotherapeutics).

Quaternary imidazolium derivatives possess algicidal and bactericidal properties [21].

6.2. Other Uses

1-Vinylimidazole is employed as a copolymer in the production of cationic polymers for various uses. Alkylimidazoles are used as hardeners for epoxy resins and for polyurethanes [24]. Less important uses for alkyl- and arylimidazoles include photography and dyes.

7. Toxicology and Occupational Health

7.1. Imidazoles without Nitro Substituents

The prominent feature in imidazole toxicity is the corrosive nature of the compounds. This property is eliminated by 2-phenyl-substitution. In terms of the labelling directives of the EEC the acute systemic toxicity ranges from toxic (T) to harmful to health (X_n).

Acute poisoning results in excitement of the central nervous system and convulsions; therefore, imidazole has been tested as an antidepressant [25]. Interaction of imidazole with the dopaminergic and noradrenergic systems in the brain was postulated [25]. Repeated application of some of the compounds revealed no specific toxic effects. The liver appears to be affected after treatment with high doses. None of the available in vitro short-term tests revealed a mutagenic potential for nonnitroimidazole compounds [26–29].

Imidazole. Acute toxicity data (see Table 4) show only slight toxic effects on animals exposed via oral or parenteral routes. The main symptoms of intoxication at high doses are seizures, tremors, loss of balance, opisthotonus (spasm of muscles in the back), and restlessness. Single oral exposure of pregnant rats to a dose of 240 mg/kg of imidazoles on the 12th or 13th day of gestation produced no harmful effect in the progeny [30]. Topically, imidazole is corrosive to the skin and eye, resulting in necrosis and irreversible damage to these organs (Table 4).

Repeated administration to rats for 28 days by gavage (through a stomach tube) revealed no observable effect at a level up to 62.5 mg kg^{-1} d^{-1}. Higher doses up to 500 mg/kg led to swelling of the liver and other unspecific toxic effects without mortality. The alteration of the liver may be due to the enhancement of metabolic activity in this organ [26].

1-Alkyl-Substituted Imidazoles. These compounds become more noxious with increasing length of the side chain: 1-*n*-butylimidazole is fairly toxic, regardless of exposure route. Further increase in chain length reduces the toxicity to that of the parent compound (see Table 4).

Table 4. Toxicity of imidazoles containing saturated substituents

Substance	Acute toxicity, LD_{50},	mg/kg	Effect on skin and eye	References
Imidazole	Rat, oral	970	Corrosive	[25], [26], [31], [35]
	Mouse, oral	880		
	Mouse, oral (male)	1880		
	Rat, s.c.	626		
	Mouse, s.c.	560; 817		
	Rat, i.p.	620		
	Mouse, i.p.	520		
	Mouse, i.v.	475		
1-Methylimidazole	Rat, oral	1139	Corrosive	[26], [31]
	Mouse, oral (male)	1400		
	Rabbit, dermal	400 < 640		
	Mouse, i.p.	445		
1-Ethylimidazole	Rat, oral	900 ∗	Corrosive	[26]
	Mouse, i.p.	300 ∗		
1-Propylimidazole	Rat, oral	400 ∗	Corrosive	[26]
	Rabbit, dermal	200 ∗		
	Mouse, i.p.	90 ∗		
1-*n*-Butylimidazole	Rat, oral	80	Strongly irritant	[26]
	Rabbit, dermal	<200		
	Rat, dermal	50 <200		
	Rat, inhalation 4 h (aerosol)	0.11 mg/L		
	Mouse, i.p.	40		
1-Decylimidazole	Rat, oral	470	Corrosive	[26]
	Mouse, i.p.	150		
1-Dodecylimidazole	Rat, oral	1100∗	Corrosive	[26]
	Rabbit, dermal	400		
	Mouse, i.p.	125∗		
2-Methylimidazole	Rat, oral	1300	Strongly irritant to corrosive	[26], [31]
	Mouse, oral (male)	1400		
	Rat, dermal	ca. 2000		
	Rabbit, dermal	200		
	Mouse, i.p.	350		
2-Ethylimidazole	Rat, oral	1400	Corrosive	[26]
	Rabbit, dermal	200		
	Mouse, i.p.	400		
2-Isopropylimidazole	Rat, oral	>1000 <2000	Irritant	[26]
	Rabbit, dermal	400		
	Mouse, i.p.	50 <200		
4-Methylimidazole	Rat, oral	350	Strongly irritant to corrosive	[26], [31], [36]
	Mouse, oral (male)	370		
	Rabbit, dermal	440		
	Mouse, i.p. (male)	165		
1,2-Dimethylimidazole	Rat, oral	1300	Strongly irritant to corrosive	[26]
	Rabbit, dermal	200		
	Rat, inhalation 4 h (aerosol of 50 % solution)	3 mg/L		
	Mouse, i.p.	300		
2-Ethyl-4(5)-methylimidazole	Rat, oral	731; 1000	Strongly irritant to corrosive	[26]
	Rabbit, dermal	400		
	Mouse, i.p.	200		
1-Vinylimidazole	Rat, oral	1100	Strongly irritant to corrosive	[26]
	Mouse, s.c.	650		

∗ LD_{50}, µL/kg.

2-Methylimidazole behaves like 1-methylimidazole and imidazole (Table 4). Its irritant properties seem to be slightly less than those of the other two compounds but are nevertheless pronounced.

A 28-day gavage study in rats revealed at the lowest dose of 100 mg kg^{-1} d^{-1} only a slight decrease in blood protein content. Higher doses (up to 800 mg/kg) led to unspecific clinical symptoms as well as changes in clinicochemical parameters and organ weights without mortality [26].

4-Methylimidazole is the most toxic of the methyl-substituted imidazoles and is thought to be the toxic agent in ammoniated molasses and hay, which may be responsible for postnatal losses in ruminants fed with these products [31], [32]. It is fairly toxic via the oral and dermal routes. 4-Methylimidazole is the most potent convulsive within this group of compounds. In addition to the data in Table 4, further studies in rabbits [31] and calves [32] demonstrate that single oral doses of 120 – 150 or 200 mg/kg (rabbit) and 400 mg/kg (calf) lead to convulsion and death. No clinical, clinicochemical, or pathological effects were observed in the study with calves after four repeated doses of 25 or 50 mg/kg within 96 h. At a level of 100 mg/kg, mild behavioral changes appeared.

Intraperitoneal dosage of 43 mg/kg to rats for up to 10 weeks led to enlargement of the liver in some animals, but with no effect on the blood or clinicochemical parameters [33].

1-Vinylimidazole. Resorption of 1-vinylimidazole through the skin in fatal amounts is possible in a relatively short exposure time. The symptoms of intoxication differ with dose and species but excitation and convulsion are a common feature at high doses.

After repeated application (rat, rabbit, cat, oral) for up to 7 weeks, unspecific clinical symptoms (weight loss, changes in blood count) and pathological findings (sporadic incidence of liver damage) were reported. The level of no observable effect was between 10 and 100 µL/kg [26].

No toxic effects were observed (mouse, rat, rabbit, cat) after repeated inhalation (five 6-h exposures) of an atmosphere saturated with vinylimidazole vapor at room temperature [26], [34].

N-Methyl-2-vinylimidazole and *N*-butyl-2-vinylimidazole showed a sensitizing effect on the skin (guinea pig) in addition to the corrosive action [34].

Phenylimidazoles. 2-Phenyl-substituted imidazoles display reduced irritancy but no change in acute systemic toxicity compared with imidazole. The acute oral LD_{50} for 2-phenylimidazole [26], 1-methyl-2-phenylimidazole [26], and 1-vinyl-2-phenylimidazole [26] is ca. 1000 mg/kg; the LD_{50} (mouse, i.p.) is ca. 200 mg/kg.

The symptoms of intoxication indicate depression of the central nervous system by 2-phenylimidazole and a convulsive action for the other two compounds.

Inhalation Hazard Tests. With the exception of imidazole, 2-isopropylimidazole, and 1-vinylimidazole, all of the compounds listed in Table 4 have also been subjected to seven- or eight-hour practice-oriented inhalation hazard tests in which rats were exposed to an atmosphere highly enriched in or saturated with the vapor or the volatile components of the test substance [26]. Of the compounds tested, only 1-*n*-butylimidazole resulted in mortality; three out of six rats died after exposure for 8 h [26].

7.2. Nitroimidazoles

Nitroimidazoles are a well-established group of antiprotozoal and bactericidal agents. Due to their aromatic nitro group, they may undergo metabolization to give electrophilic intermediates under certain conditions [37], [38] and thus entail bacteriotoxic, but also mutagenic and potentially carcinogenic properties. Metabolization of the nitro group may be achieved under anaerobic conditions by nitroreductases of intestinal bacteria or of the liver [39]. Bone marrow cells also have the potential to reduce the nitro group even under aerobic conditions [40].

Acute Toxicity. The acute toxicity data for nitroimidazoles indicate a moderate to considerable acute toxicity depending on the substituents. In some cases the pH may also play a

role. The materials so far investigated are mostly eye irritants, but not skin irritants.

The acute toxicities (LD_{50}, mg/kg) for some nitroimidazoles are as follows:

4(5)-Nitroimidazole [26]	1660 (rat, oral)
1,2-Dimethyl-5-nitroimidazole hydrochloride (dimetridazole hydrochloride) [26]	3000 (rat, oral) 200 – 2000 (cat, oral)
1-Methyl-5-nitroimidazole [26]	ca. 200 (rat, oral)
1-(2-Hydroxymethyl)-2-methyl-5-nitroimidazole (metronidazole) [41]	3800 (mouse, oral)
α-Methoxymethyl-2-nitroimidazole-1-ethanol (misonidazole) [42]	2130 (rat, oral)
2-Methyl-4(5)-nitroimidazole [26]	1540 (rat, oral)

Four-hour dust inhalation tests on rats with 4(5)-nitroimidazole and dimetridazole gave LC_{50} values of 6.56 and 5.3 mg/L, respectively. Dimetridazole and 1-methyl-5-nitroimidazole have an irritant action on the eyes of rabbits, while 4(5)-nitroimidazole is non-irritating [26]. None of the three compounds exhibits irritancy towards the skin of rabbits [26].

Mutagenicity. Several nitroimidazoles have been subjected to bacterial point mutation assays (such as the Ames test), and, in most cases, were found to exert mutagenicity [37], [43–51]. Among these were the following:

1-Methyl-5-nitroimidazole
2-Methyl-4(5)-nitroimidazole
1,2-Dimethyl-5-nitroimidazole (dimetridazole)
1-(2-Hydroxymethyl)-2-methyl-5-nitroimidazole (metronidazole)
2-Isopropyl-4-nitroimidazole
1-Methyl-2-isopropyl-5-nitroimidazole (ipronidazole)
1-(3-Chloro-2-hydroxypropyl)-2-methyl-5-nitroimidazole (ornidazole)
(1-Methyl-5-nitroimidazole-2-yl)methyl carbamate (ronidazole)
α-Methoxymethyl-2-nitroimidazole-1-ethanol (misonidazole)

Nitroimidazoles also induce streptomycin-resistant mutants (forward mutations) in *Klebsiella pneumoniae* [44], [45].

Conflicting results were earlier reported with 4(5)-nitroimidazole, 2-methyl-4(5)-nitroimidazole, and 2-methyl-4(5)-nitroimidazole-1-yl-acetate [43], [46], [47]. Negative results may be due to the inadequate metabolic activation achieved in the test systems. Materials bearing a carboxylate or sulfate group may in fact be non-mutagenic [43], [45], but the data are insufficient for a firm conclusion.

Urine and other body fluids of patients treated with metronidazole in doses of 750 mg/d showed mutagenic activity in the Ames test [52], [53]. Patients who received 200 – 1200 mg/d for several months were found to have two- to four-fold increases in chromosomal abnormalities in peripheral leukocytes [54].

Efforts have been undertaken to lower the mutagenicity of nitroimidazoles by derivatisation and substitution [47], [48], [55]. Note, however, that mutagenicity in vitro is not quantitatively extrapolable to in vivo situations. Furthermore, a lower mutagenicity does not necessarily indicate a decrease in the carcinogenic potential as well.

Carcinogenicity. The carcinogenic potential has been investigated for some nitroimidazoles used or foreseen as drugs. However, the study designs were not fully in accordance with present standards. The carcinogenic responses obtained so far were moderate or low. Human experience with nitroimidazole drugs has not revealed cancer cases due to therapeutic usage. Metronidazole is listed as a carcinogen by IARC [56]. Mice receiving 0.06, 0.15, 0.3, or 0.5 % (ca. 90, 225, 450, or 750 mg/kg) of metronidazole in the diet throughout their life-span showed a dose-dependent increase in the incidence of lung tumors and lymphomas. The lowest dose did not significantly increase the incidence of tumors [57].

No carcinogenic effects were found in rats in a feeding experiment with 0.135 % (ca. 200 mg $kg^{-1} d^{-1}$) of metronidazole in the diet for 66 weeks and a post-observation period of 10 weeks [58]. This administration period is, however, too short to fully rule out carcinogenic effects.

More recent investigations with metronidazole in the diet of mice (75, 150, or 600 mg $kg^{-1} d^{-1}$ for 92 weeks) and rats (0.06, 0.3 or 0.6 %, equivalent to 30, 150 or 300 mg $kg^{-1} d^{-1}$ for 150 weeks) showed dose-related increases in tumor incidence (liver and mammary tumors, Leydig cell and pituitary tumors) [59], [60].

After gavage administration of metronidazole to mice during pregnancy and lactation (2 mg per animal per day for 5 d per week) an increased tumor incidence was observed in the offspring [61].

Metronidazole (0.15 or 0.3 %) in the diet of hamsters did not increase the incidence of tumors; however, published details of the study are insufficient for a reliable assessment of the carcinogenic effect [62].

A group of patients who had been treated with metronidazole during the period 1960 – 1969 did not show increased tumor rates in a follow-up study published in 1988 [63].

Dimetridazole was found to have tumorigenic activity in rats receiving 0.2 % in the diet for 46 weeks. Of 35 Sprague – Dawley rats, 25 developed a benign, but possibly premalignant mammary tumor (fibroadenoma) compared to 4 from 35 rats in the control group [58]. The administration period was, however, too short to fully determine the tumorigenic potential.

Ornidazole (25, 100, or 400 mg kg^{-1} d^{-1}) was reported to cause no tumors in a two-year feeding experiment in rats [64].

The overall conclusion from all experiments is that the carcinogenic potential of nitroimidazoles has to be taken into account, but the amount and quality of the available data do not allow firm conclusions.

Neurotoxic effects. There are several case reports on peripheral neuropathy due to Metronidazole treatment. The symptoms appear to be reversible upon ceasing treatment [65].

8. References

General References

1. M. R. Grimmett in K. T. Potts: *Comprehensive Heterocyclic Chemistry,* **vol. 5,** Pergamon Press, 1984, pp. 345 – 456.
2. K. Hofmann: *The Chemistry of Heterocyclic Compounds, Imidazole and Derivatives,* Part I, Intersci. Publ. Inc., New York 1953.

Specific References

3. BASF, DE-OS 2 360 175, 1979 (A. Frank, H. Karn, H. Spänig).
4. BASF, EP 0 012 371, 1982 (T. Dockner, U. Kempe, H. Krug, P. Magnussen, W. Prätorius, H. Szymanski).
5. BASF, DE 2 512 513, 1976 (A. Frank, T. Dockner).
6. BASF, EP 0 111 073, 1987 (A. Frank, T. Dockner).
7. BASF, EP 0 036 521, 1983 (T. Dockner, U. Kempe, H. Krug, P. Magnussen, W. Prätorius).
8. Houdry Process Corp., US 2 847 417, 1958 (W. E. Erner).
9. BASF, EP 0 000 208, 1979 (T. Dockner, A. Frank).
10. C. Corsar et al., *Arzneim. Forsch.* **16** (1966) 23.
11. BASF, DE 2 645 172, 1982 (M. Wetzler, T. Dockner).
12. BASF, DE 1 670 293, 1976 (H. Spänig, A. Steimmig, J. Sand).
13. BASF, DE-AS 2 233 908, 1977 (T. Dockner, H. Krug).
14. M. Häring, *Helv. Chim. Acta* **42** (1959) 1845 – 1846.
15. A. Grimison, J. H. Ridd, B. V. Smith, *J. Chem. Soc.* 1960, 1357 – 1362.
16. W. Reppe, *Justus Liebigs Ann. Chem.* **601** (1956) 81 – 138.
17. Rhône-Poulenc, DE-OS 1 470 200, 1970 (C. Podesva, K. Vagi).
18. Y. Masashige, M. Masaichiro, *Chem. Pharm. Bull.* **24** (1976) 1480 – 1484.
19. BASF, EP 0 015 516, 1982 (U. Kempe, T. Dockner, F. Frank, H. Karn).
20. Delalande S.A., DE-OS 2 934 925, 1978 (Y. A. Hubert-Brierre).
21. E. R. Shepard, H. A. Shonle, *J. Am. Chem. Soc.* **69** (1947) 2269 – 2270.
22. W. Perkow: *Wirksubstanzen der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* 2nd ed., Verlag Paul Parey, Berlin – Hamburg 1985.
23. M. D. Nair, K. Nagarajan, *Prog. Drug Res.* **27** (1983) 163 – 252.
24. Shell Oil Co., US 3 538 039, 1970 (W. L. Lantz, J. P. Manasia).
25. F. Ferrari, *Arch. Int. Pharmacodyn. Ther.* **227** (1985) 303 – 312.
26. BASF Aktiengesellschaft, unpublished results.
27. L. E. Voogd et al., *Mutat. Res.* **66** (1979) 207 – 221.
28. A. Momii et al., *Iyakuhin Kenkyu* **10** (1979) 351 – 357.
29. E. Grzybowska et al., *Acta Microbiol. Pol.* **34** (1985) 111 – 120.
30. J. A. Ruddick et al., *Teratology* **13** (1976) 263 – 266.

31. K. Nishie et al., *Toxicol. Appl. Pharmacol.* **14** (1969) 301 – 307.
32. T. E. Fairbrother et al., *Vet. Hum. Toxicol.* **29** (1987) 312 – 315.
33. M. H. Idaka, *Okayama Igakkai Zasshi* **88** (1976) 673 – 680.
34. *Ullmann,* 4th ed., **23,** 617.
35. Niosh: Registry of Toxic Effects of Chemicals.
36. M. Hidaka, *Okayama Igakkai Zasshi* **88** (1976) 653 – 657.
37. D. G. Lindmark et al., *Antimicrob. Agents Chemother.* **10** (1976) 476 – 482.
38. D. I. Edwards, M. Dye, H. Carne, *J. Gen. Microbiol.* **76** (1973) 135.
39. J. L. Blumer et al., *Cancer Res.* **40** (1980) 4599 – 4605.
40. M. Isildar et al., *Toxicol. Appl. Pharmacol.* **94** (1988) 305 – 310.
41. B. Cavalleri et al., *J. Med. Chem.* **20** (1977) 1522.
42. National Technical Information Service USA, PB 81-121212.
43. C. E. Voogd et al., *Mutat. Res.* **86** (1981) 243 – 277.
44. C. E. Voogd et al., *Mutat. Res.* **26** (1974) 483 – 490.
45. C. E. Voogd et al., *Mutat. Res.* **66** (1979) 207 – 221.
46. H. S. Rosenkranz et al., *Mutat. Res.* **38** (1976) 203 – 206.
47. J. B. Chin et al., *Mutat. Res.* **58** (1978) 1 – 10.
48. J. S. Walsh et al., *J. Med. Chem.* **30** (1987) 150 – 156.
49. M. Suzangar et al., *Biochem. Pharmacol.* **36** (1987) no. 21, 3743 – 3749.
50. P. G. Wislocki et al., *Chem. Biol. Interactions* **49** (1984) 27 – 38.
51. H. S. Rosenkranz et al., *Biochem. Biophys. Res. Commun.* **66** (1975) 520 – 525.
52. M. S. Legator et al., *Science* **188** (1975) 1118 – 1119.
53. E. Mohtashamipur et al., *Mutagenesis* **1** (1986) 371 – 374.
54. F. Mitelman et al., *Lancet,* 1976, 802.
55. G. Cantelli-Forti et al., *Teratog. Carcinog. Mutag.* **3** (1983) 51 – 63.
56. IARC Monograph **13**, 113 – 121.
57. M. Rustia, P. Shubik, *J. Nat. Cancer Inst.* **48** (1972) 721 – 729.
58. S. M. Cohen et al., *J. Nat. Cancer Inst.* **51** (1973) 403 – 417.
59. J. H. Rust, An assessment of metronidazole tumorigenicity studies in the mouse and the rat, in S. M. Finegold (ed.): Metronidazole, Proc. Intern. Metronidazole Conf. Montreal, Excerpta Medica, Amsterdam (1977) 138 – 144.
60. M. Rustia, P. Shubik, *J. Nat. Cancer Inst.* **63** (1979) 863 – 868.
61. M. Chacko et al., *Cancer Res. Clin. Oncol.* **112** (1986) 135 – 140.
62. F. C. Roe, Proc. Intern. Metronidazole Conf. Montreal, Excerpta Medica, Amsterdam (1977) 132 – 137.
63. C. M. Beard et al., *Mayo Clin. Proc.* **63** (1988) 147 – 153.
64. R. Richle et al., *Arzneim. Forsch.* **28** (1978) 612 – 625.
65. W. G. Bradley et al., *Br. Med. J.* **2** (1977) 610 – 611.

Immobilized Biocatalysts

Jiří E. Přenosil, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland

Ömer M. Kut, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland

Irving J. Dunn, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland

Elmar Heinzle, Eidgenössische Technische Hochschule Zürich, Zürich, Switzerland

1. Introduction . . . 3
2. Immobilization Techniques . . . 5
2.1. Introduction . . . 5
2.2. Methods of Enzyme Immobilization . . . 5
2.2.1. Carriers for Enzyme Immobilization 5
2.2.2. Methods for Insoluble Enzymes . . 6
2.2.2.1. Carrier Binding . . . 6
2.2.2.2. Cross-Linking . . . 8
2.2.2.3. Enzyme Copolymerization . . . 11
2.2.2.4. Entrapment . . . 11
2.2.3. Methods for Soluble Enzymes . . . 12
2.3. Methods of Cell Immobilization . 12
2.3.1. Introduction . . . 12
2.3.2. Cell Supports . . . 13
2.3.3. Immobilization Techniques . . . 14
2.3.3.1. Cell Immobilization without a Support . . . 14
2.3.3.2. Binding of Cells to a Carrier . . . 14
2.3.3.3. Immobilization of Cells by Entrapment . . . 15
2.4. Methods of Organelle Immobilization . . . 17
2.5. Co-immobilization of Biocatalysts 17
3. Activity and Kinetics of Immobilized Biocatalysts . . . 18
3.1. Effects of Immobilization on Enzyme Activity . . . 18
3.2. Kinetics of Immobilized Biocatalysts . . . 20
3.3. Assay of Immobilized Biocatalysts 21
4. Mass Transfer in Immobilized Biocatalyst Systems . . . 21
4.1. External Mass Transfer . . . 21
4.2. Prediction and Correlation of External Mass-Transfer Coefficients 22
4.2.1. Catalyst Beds . . . 22
4.2.2. Particles in an Agitated Tank . . . 23
4.3. Internal Mass Transfer . . . 24
5. Design of Immobilized Biocatalyst Reactors . . . 28
5.1. Batch Stirred-Tank Reactor . . . 28
5.2. Plug-Flow Tubular Reactor . . . 30
5.3. Continuous Stirred-Tank Reactor 30
5.4. Packed- or Fixed-Bed Reactors . 30
5.5. Fluidized-Bed Reactors . . . 31
5.6. Membrane Reactor Systems . . . 32
5.7. Laboratory Reactors for Kinetic Measurements . . . 32
5.7.1. Batch and Continuous Tank Reactors 32
5.7.2. Experimental Differential Recycle Reactor . . . 33
6. Immobilization Media . . . 34
6.1. Characteristics of Solid Supports 34
6.2. Characteristics of Semipermeable Separators . . . 34
7. Application of Immobilized Biocatalysts . . . 34
7.1. Analysis . . . 34
7.2. Immobilized Enzymes in Therapeutic Medicine . . . 36
7.3. Ethanol Production Using Immobilized Cells . . . 37
7.4. Industrial Applications of Immobilized Biocatalysts . . . 37
7.4.1. Production of High-Fructose Corn Syrup . . . 38
7.4.2. Amino Acid Production . . . 39
7.4.3. Environmental Applications . . . 41
8. Economic Aspects . . . 41
9. Safety, Environmental, and Legal Aspects . . . 43
9.1. Immobilized Enzymes . . . 43
9.2. Immobilized Cells . . . 44
10. References . . . 46

Symbols:

A area, m^2
C concentration, g/m^3
d diameter, m
D diffusivity, m^2/s
F Faraday constant; volumetric flow rate, m^3/s
j diffusion flux, $g\,m^{-2}\,s^{-1}$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_001

j_D	factor in mass-transfer correlation
k	Boltzmann constant; mass-transfer constant; rate constant
K_M	Michaelis – Menten constant
L	total length, depth, or thickness, m
m	mass
n	exponent or reaction order
N	dimensionless number
P	power, $g\,m^{-2}\,s^{-3}$
P_{D-R}	parameter for diffusion – reaction process
r	reaction rate, $g\,m^{-3}\,s^{-1}$ or $g\,m^{-2}\,s^{-1}$
R	gas constant
S	substrate concentration, g/m^3 or mol/m^3
S'	dimensionless substrate S/S_0
t	time, s
t'	dimensionless time $t/(L^2/D)$
T	temperature, K
v	flow velocity, m/s; reaction velocity, $g\,m^{-3}s^{-1}$
V	volume, m^3 or m^3/mol
z	valence
Z	length, m
Z'	dimensionless length Z/L
ε	void fraction; volume fraction
η	effectiveness factor
μ	viscosity, $g\,m^{-1}\,s^{-1}$
μ/ϱ	kinematic viscosity, m^2/s
ϱ	density, g/m^3
τ	residence time, s
Φ	Thiele modulus
ψ	electrostatic potential, V

Indices

A	location in bulk, substance A
app	apparent
avg	average value
B	location on surface, substance B
bulk	bulk solution
c	catalyst
i	component i
I	impeller
L	liquid
m	mass
max	maximum
M	matrix
n	nth position or number
p	particle
Pe	Péclet number
r	reactor
R	recycle stream
Re	Reynolds number
s	surface; system
s	surface
S	solid; substrate
Sc	Schmidt number
Sh	Sherwood number
SL	liquid diffusion film at the solid – liquid interface
SLt	liquid diffusion film at the solid – liquid interface using terminal velocity
t	terminal velocity
T	total system
0	bulk; initial value; inlet; zero order
1	first order; outlet
2	second order

1. Introduction

Catalytic activity is an important property of many biological materials termed biocatalysts, which include whole cells, parts of cells, and enzymes. Biocatalysts catalyze chemical reactions that usually occur in living organisms.

The catalytic power of enzymes has been exploited by humans for thousands of years. The final product is often formed by a number of reaction steps catalyzed by a system of enzymes acting in ordered fashion. The enzymatic system is usually localized within a living cell (microorganism, animal, or plant cell). The catalytic activity of the whole cell or parts of it can be used in the same manner as isolated enzymes.

The use of single purified enzymes may be desirable for reaction modeling or process management, but in many cases this approach is associated with practical difficulties (primarily enzyme stability and purification cost). Therefore, the use of whole cells or parts of cells containing an appropriate enzyme system may be advantageous, especially when complex reactions are involved. Furthermore, the problems of residual biomass and wastewater disposal involved in large-scale enzyme production, as well as the costs of enzyme isolation and purification, may become prohibitive. The lower specific activity of whole cells and a potentially lower selectivity are often balanced by lower cost and simplicity of use.

Many processes utilized by humans mimic nature, and enzyme immobilization is no exception. All native intracellular enzymes seem to be bound to the cell membrane or localized in the gel-like surroundings of a cellular organelle [1].

Cellular enzyme systems vary greatly from two or more sequentially acting independent enzymes to multienzyme complexes consisting of tightly aggregated enzymatic or regulatory subunits which lose their activity upon separation [2], [3].

Under natural conditions, microorganisms tend to grow in aggregates attached to solid surfaces or to form microbial films (*biofilms*) on a variety of materials. Studies on microbial populations in rivers show that very few microorganisms exist free in suspension; instead, they are associated with solid surfaces (e.g., silt particles) [4].

Biologically active materials are immobilized mainly to allow repeated or continuous use of their activity in a controlled environment (e.g., a bioreactor). Anchorage of an enzyme molecule to a carrier can also stabilize its tertiary structure and prolong its useful life.

Enzymes can be retained in a reactor by covalent, ionic, or adsorptive attachment to surfaces, cross-linking, entrapment in gels and polymers, inclusion in micelles, or retention by membrane systems. The methods used for retaining whole cells in a reactor are immobilization in gels and polymers, entrapment in porous solids or beds of particulates, formation of biofilms and flocs, and the utilization of membrane systems. In continuous processes, attention must be paid to the problem of enzyme loss, especially when the cell wall has been made permeable to facilitate substrate and product transport and the carrier matrix does not retain the enzyme.

Most of the techniques and principles applied to immobilized enzymes are applicable to immobilized whole cells. In fact, from the standpoint of engineering, little reason exists for distinguishing between these two types of biocatalyst. The application of biofilms or microorganisms that are naturally immobilized by their ability to form flocs has no counterpart in enzyme technology, but the same principles of diffusion and reaction are also involved here.

Similarly, a dissolved enzyme differs little from a suspended organism in a membrane reactor. Therefore, this article deals with biocatalysts consisting of either enzymes or whole cells, which occur in suspended form or are immobilized in solids, gels, or biofilms.

Historical Aspects. The use of fixed biocatalysts is closely connected with the history of human civilization. Vinegar manufacturing was one of the first industrial processes to use immobilized living cells. Alcoholic solutions were trickled through vats packed with wood shavings. Microbial films developed on the shavings and converted the ethanol to acetic acid ("Quick" process, invented in 1823). Attempts to immobilize enzymes date back as far as 1916, when Nelson and Griffin showed that immobilized invertase was active [5].

An overview of the historical development of immobilized biocatalysts is given in Table 1.

Table 1. Historical development of immobilized biocatalysts (partly based on [6])

Year	Biocatalyst, process, or event	Comments
1823	vinegar manufacture from alcoholic solutions by using biofilm on wood chips ("Quick" process)	
1916	invertase adsorbed on charcoal [5]	
1948	urease insolubilization [7]	
1949	physiologically active protein immobilization [8]	
1951	albumin bound to diazonium derivative of 4-aminobenzyl cellulose [9]	
1953	immobilization of carboxypeptidase, diastase, pepsin, and ribonuclease [10]	
1969	continuous optical resolution of D,L-amino acids by immobilized aminoacylase [11]	first industrial process employing an immobilized biocatalyst
1971	First Enzyme Engineering Conference at Henniker, New Hampshire [12]	definition of immobilized enzymes
1972	isomerization of glucose to fructose by immobilized glucose isomerase [13]	
1973	production of L-aspartic acid by immobilized *Escherichia coli* (aspartase) [14]	first industrial use of immobilized microbial cells
1975	adsorption of rat liver mitochondria on alkylsilanized glass beads [15]	first immobilization of subcellular particles
1979	adsorption of algae (*Anabaena cylindrica*) on glass beads for the evolution of hydrogen [16]	

Table 1. Continued

Year	Biocatalyst, process, or event	Comments
1979	immobilization of plant cells (*Morinda, Catharanthus, Digitalis*) by the calcium alginate method [17]	first immobilization of plant cells
1979	slice of porcine kidney immobilized in the membrane of an ammonia electrode for amino acid detection [18]	first immobilization of animal cells
1987	*Corynebacterium* spp. immobilized in polyacrylamide gel for acrylamide production [19]	first reported industrial production of a commodity chemical by immobilized cells

2. Immobilization Techniques

2.1. Introduction

Since the late 1960s a variety of techniques have been developed for immobilizing biocatalysts. Complete coverage of the vast number of papers, books, and patents on this topic is almost impossible. The reader is referred to comprehensive reviews and monographs [20–38].

The definition of biocatalysts used in this chapter is that drawn up by the Working Party on Applied Biocatalysts within the European Federation of Biotechnology: "Immobilized biocatalysts are enzymes, cells, or organelles (or combinations of them) which are in a state that permits their reuse" [39].

Biocatalysts dissolved in aqueous buffer solutions are usually referred to as *soluble or native* enzymes, cells, cell parts, or organelles. The terms *immobilized, fixed, or insolubilized* enzymes, cells, etc., denote biocatalysts that are bound to a support. The material to which enzymes or cells are bound is called the *carrier, support*, or *matrix.* The agent enabling enzyme – support binding is referred to as a *cross-linking agent, bifunctional agent,* or *carrier activator*. For clarity, techniques for immobilization of enzymes and cells are discussed separately (Sections 2.2 and 2.3, respectively).

2.2. Methods of Enzyme Immobilization

(see also → Enzymes, Chap. 3.3.)

The definition of immobilized biocatalysts given in Section 2.1 includes enzymes that are

1) modified into a water-insoluble form,
2) retained by an ultrafiltration membrane inside a reactor, or
3) bound to another macromolecule to restrict their mobility.

Several schemes have been suggested for classifying immobilization techniques [24], [30], [40], [41]. The classification used here is based on that of Kennedy and Cabral [41] and attempts to combine the nature of the interaction responsible for immobilization with that of the support (Fig. 1).

A perfect, universal immobilization method does not exist. Each end use requires evaluation of the individual steps according to criteria such as the purpose of immobilization, activity, stability, simplicity, and economic feasibility [42], [43].

2.2.1. Carriers for Enzyme Immobilization

The major components of an immobilized enzyme system are the enzyme, the support, and the mode of interaction of the enzyme with the support. The support may be a membrane, a water-insoluble solid, or a polymer matrix. The support (carrier) should have the following properties [44], [45]:

1) Large surface area and high permeability
2) Sufficient functional groups for enzyme attachment under nondenaturing conditions
3) Hydrophilic character
4) Water insolubility
5) Chemical and thermal stability
6) Mechanical strength
7) High rigidity and suitable particle form
8) Resistance to microbial attack
9) Regenerability
10) Toxicological safety
11) Low or justifiable price

Carriers can be classified according to their morphology (porous, nonporous, or gel type) or their chemical composition. The most widely used supports are listed in Table 2. The characteristics of these materials are presented in [44], [45]. Sources of commercial carriers are listed in [46].

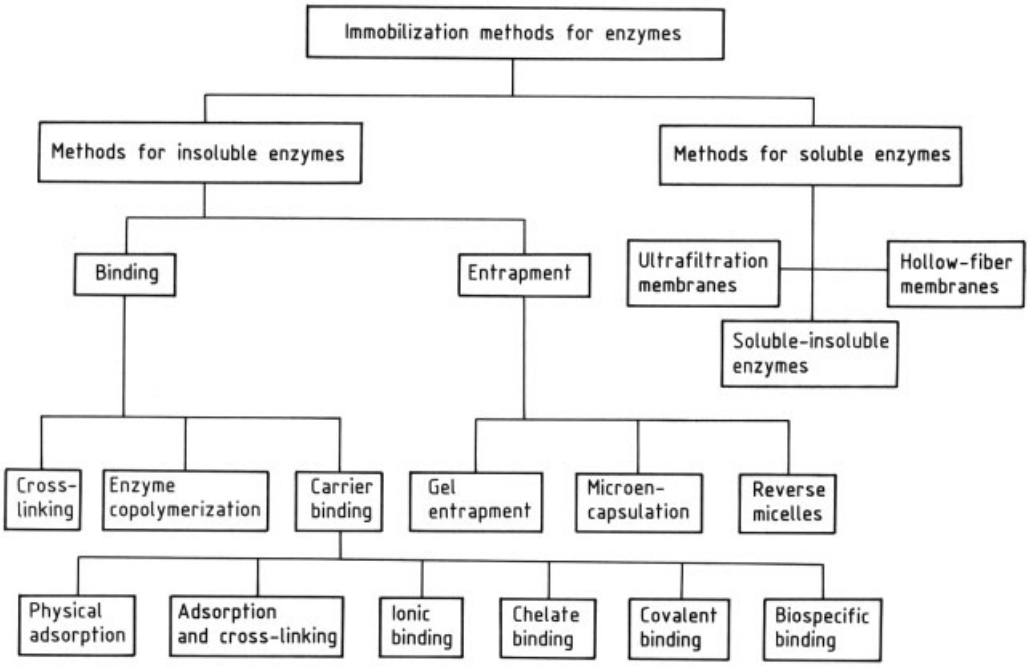

Figure 1. Classification of enzyme immobilization methods

Table 2. Chemical classification of matrixes used for enzyme immobilization [45]

Organic matrixes	Inorganic matrixes
Natural polymers	*Minerals*
Polysaccharides	Attapulgite clays
Cellulose	Bentonite
Starch	Kieselghur
Dextran	Pumice stone
Agar and agarose	
Alginate	
Carrageenan	
Chitin and chitosan	
Proteins	
Collagen	
Gelatin	
Albumin	
Silk	
Carbon materials (activated carbon)	
Synthetic polymers	*Synthetic materials*
Polystyrene	Nonporous glass
Polyacrylates and polymethacrylates	Controlled pore glass
Polyacrylamide	Controlled pore metal oxides
Hydroxyalkyl methacrylates	Metals
Glycidyl methacrylates	
Maleic anhydride polymers	
Vinyl and allyl polymers	
Polyamides	

2.2.2. Methods for Insoluble Enzymes

2.2.2.1. Carrier Binding

Enzymes are large protein molecules with chemically reactive groups, ionic groups, and hydrophilic domains that can all participate in the immobilization process through physical adsorption, ionic binding, or covalent linkage.

Physical Adsorption. In physical adsorption, the enzyme adheres to the surface of a support which has not been specially functionalized for covalent binding [47]. Attachment is achieved by physical interactions such as van der Waals forces, hydrogen bonding, or hydrophilic – hydrophobic effects. In the absence of chemical reactions, practically no conformational changes occur in the enzyme and associated loss of activity is minimal. However, binding forces between the enzyme and carrier are generally weak, and leakage of enzyme by desorption may occur in flow systems. Changes of operational parameters such as pH, ionic strength, or temperature can increase leakage dramatically. Commonly used absorbents are

alumina, activated carbon, clay, diatomaceous earth, glass, and hydroxyapatite. The problems of enzyme immobilization by physical adsorption are discussed in [47], [48], and examples are given in [49], [50]. In some cases such problems can be eliminated by cross-linking the adsorbed enzyme using multifunctional reagents (Section 2.2.2.2).

Ionic Binding. Ionic binding is an established, simple way of immobilizing enzymes. The binding forces are ion – ion interactions and are stronger than those in simple physical adsorption. Again, however, binding stability is affected by changes of pH or ionic strength. Both commercial anion and cation exchangers can be used as carriers. Examples are given in [24], [49].

Chelate Binding. The chelating properties of a transition metal such as titanium or zirconium are employed to couple enzymes to an organic material (e.g., cellulose, sawdust, alginic acid) or an inorganic support such as celite or glass [26], [41], [51]. This method was initally used by Novais in 1971 [52] and was developed further by Kennedy and co-workers [53]. The chemistry of metal – enzyme coupling, properties of the immobilized system, and detailed working procedures are given in [54].

Biospecific Binding. Biospecific interactions between enzymes and other molecular species (e.g., lectins or antibodies) are also used for binding. Lectins such as concanavalin A [*11028-71-0*] are glycoproteins, mainly of plant origin, that bind strongly to specific carbohydrate residues. For example, a carbohydrate-containing enzyme (e.g., L-ascorbate oxidase) can be bound specifically to commercial Concanavalin A-Sepharose [55]. The same method has been used to immobilize invertase for continuous conversion of sucrose [56].

A novel method for preparing highly active immobilized enzymes is to bind them with specific monoclonal antibodies attached to carriers [57–59] (Fig. 2). The enzyme is bound at a specific, well-defined site without affecting its catalytic activity. Transglutaminase (glutaminylpeptide γ-glutamyltransferase [*80146-85-6*] E.C. 2.3.2.13) has been immobilized in this way on cross-linked agarose gel beads (AffiGel 10) to produce a very specific, highly active biocatalyst starting from a crude enzyme preparation [57]. In the immobilization of carboxypeptidase A [*11075-17-5*] (E.C. 3.4.17.1), the highest enzyme activity was shown with a mouse monoclonal antibody – enzyme complex on Eupergit C (Röhm Pharma) or on a Protein A-Sepharose CL 4B (Pharmacia) [58].

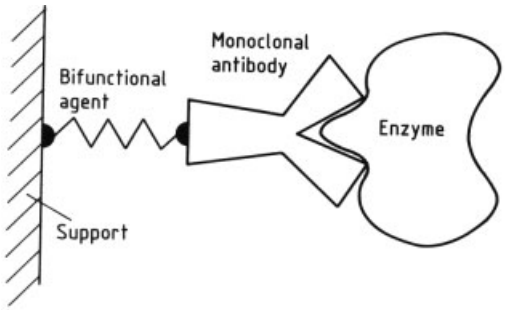

Figure 2. Schematic representation of biospecific immobilization

Covalent Binding. A water-insoluble carrier can be covalently bound to the enzyme via the reactive side groups of amino acid residues (e.g., amino, hydroxyl, thiol, or phenolic groups) that are not associated with the active site or the substrate binding site (see Table 3). This technique includes the greatest number and variety of published immobilization methods [43]. Under normal conditions, covalent binding is stable and no enzyme leakage occurs if substrate concentration or ionic strength changes during the reaction. Coupling reactions are often complicated and are carried out under relatively severe conditions in comparison with physical adsorption methods. Covalent binding can alter the conformation and active site of the enzyme, resulting in significant loss of activity or selectivity. The advantage of this approach is the wide range of possible binding methods and carriers, which allows great flexibility in designing biocatalysts with specific chemical and physical properties [55]. For example, if a hydrophilic environment is required, a coupling agent containing a number of hydroxyl groups may be utilized. An aromatic coupling agent can improve the compatibility of a substrate – enzyme reaction when the substrate is also aromatic [48]. Glycoenzymes can be coupled selectively via their carbohydrate moieties [60], [61].

Table 3. Reactive residues of proteins [45]

Group	Description
$-NH_2$	terminal amino group of L-lysine and N-terminal amino group
$-SH$	thiol group of L-cysteine
$-COOH$	carboxyl group of L-aspartate and L-glutamate and C-terminal carboxyl group
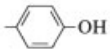	phenolic group of L-tyrosine
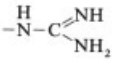	guanidino group of L-arginine
	imidazole moiety of L-histidine
$-S-S-$	disulfide group of L-cystine
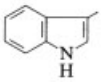	indole moiety of L-tryptophan
CH_3-S-	thioether group of L-methionine
$-CH_2OH$	hydroxyl group of L-serine and L-threonine

A survey of coupling methods and sources of commercial matrix materials can be found in [46].

Support materials in their original form do not usually possess reactive groups for direct coupling with enzymes. Their existing surface hydroxyl, amino, amido, or carboxyl groups must be chemically activated. The resulting reactive groups on the carrier surface and the coupling reactions are given in Table 4. The chemical reactions employed for coupling can be classified according to the chemical characteristics of the carrier [43], [48] or the coupling reaction [24], [41].

Diazo Coupling is based on coupling of the enzyme (e.g., the imidazole side chain of histidine or the phenolic group of tyrosine) to the aryldiazonium groups of the carrier. The aromatic amino groups on the surface of the carrier can be diazotized easily with nitrite in an acidic medium. Typical carrier structures with aromatic amines are given in [50]. Examples of applications of the method are given in [24], [49].

Peptide Bond Formation. The nucleophilic (amino, hydroxyl, thiol) groups of the enzyme can attack the activated functional groups to form a covalent (amide, peptide, or other) bond. Two different surface activation processes can be used [24], [44]:

1) Supports containing carboxyl groups are converted to reactive derivatives (e.g., acyl azide, acid chloride, acid anhydride, isocyanate, isothiocyanate, cyclic carbonate, or cyclic imidocarbonate), which can react directly with free amino functions of the proteins. The well-known activation of polysaccharides with cyanogen bromide [*506-68-3*] is an example of this method [62]. The vicinal hydroxyl groups of the polysaccharide matrix (e.g., dextran, cellulose, or agarose) react with cyanogen bromide at pH 10 – 11.5 to give the reactive imidocarbonate derivative. At pH 9 – 10 the free amino groups react to give essentially substituted isoureas [63].
2) Condensing reagents such as carbodiimides or Woodward's reagent K (*N*-ethyl-5-phenylisoxazolium-3′-sulfonate) [*4156-16-5*] can form amide linkages between the free amino and carboxyl groups of the carrier surface and the enzyme.

Examples of enzyme immobilization by amide binding are presented in [24], [49].

Alkylation or Arylation. Free amino, phenolic, or thiol groups of an enzyme are alkylated or arylated with an activated functional group of the support such as halide, epoxy, or vinylsulfonyl groups. Supports with hydroxyl groups (e.g., polysaccharides or minerals) can be activated by treatment with cyanuric chloride derivatives. Another widely used support contains reactive epoxy groups. Details of modifications of the method and practical examples are given in [24], [41], [49], [50].

Miscellaneous Methods. Several other covalent binding methods have been described for the immobilization of enzymes. These include Schiff base formation, Ugi reaction [64], thiol – disulfide interchange, amidination, mercury – enzyme interaction, and γ-radiation induced coupling. These methods have limited applicability and are reviewed in [20], [24], [41], [45], [49], [50].

2.2.2.2. Cross-Linking

The enzyme can be immobilized by cross-linking it to other enzyme molecules or to

Table 4. Carrier and enzyme functional groups used for the immobilization of enzymes by covalent binding [45]

Reactive group of carrier	Reacting group of enzyme	Coupling reaction
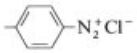 Diazonium salt	$-NH_2$ $-SH$ 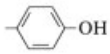	diazo coupling
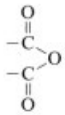 Acid anhydride	$-NH_2$	peptide bond formation
$-CH_2CON_3$ Acylazide	$-NH_2$ $-SH$ 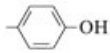	peptide bond formation
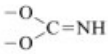 Imidocarbonate	$-NH_2$	peptide bond formation
$-R-NCS$ Isothiocyanate	$-NH_2$	peptide bond formation
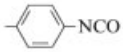 Isocyanate	$-NH_2$	peptide bond formation
$-CH_2COCl$ Acyl chloride	$-NH_2$	peptide bond formation
Cyclic carbonate	$-NH_2$	peptide bond formation
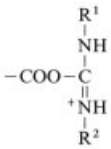 *O*-Acylisourea	$-NH_2$	peptide bond formation
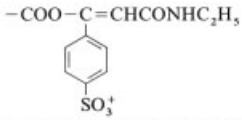 Woodward's reagent K derivative	$-NH_2$	peptide bond formation
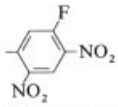 5-Fluoro-2,4-dinitroanilide	$-NH_2$	arylation

Table 4. Continued

Reactive group of carrier	Reacting group of enzyme	Coupling reaction
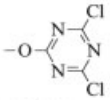 Triazinyl	$-NH_2$	arylation
$-O-CH-CH_2$ (X)	$-NH_2$	alkylation
X = NH, O, S	$-OH$ $-SH$	alkylation
$-O-CH_2-CH_2-SO_2-CH=CH_2$ Vinylsulfonyl	$-NH_2$ $-SH$ $-OH$	alkylation
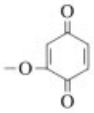 Vinylketo	$-NH_2$ $-SH$ $-OH$	Schiff's base formation
$-CHO$ Aldehyde	$-NH_2$	
$-\overset{+}{N}H=C(R^1)(R^2)$ Imine	$-COOH$ $-NH_2$	Ugi reaction
$-\overset{NH}{\overset{\|\|}{C}}-OC_2H_5$ Imidoester	$-NH_2$	amidination
$-CN$ Cyanide	$-NH_2$	amidination
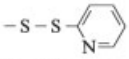 Disulfide residue	$-SH$	thiol–disulfide interchange
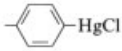 Mercury derivative	$-SH$	mercury–enzyme interaction
Matrix radical	enzyme radical	γ-radiation induced coupling
$-NH_2$ Amine	$-NH_2$ $-COOH$	peptide bond formation (in presence of condensing reagents)
$-CONHNH_2$ Acylhydrazide	$-NH_2$ $-COOH$	peptide bond formation (in presence of condensing reagents)

an inert protein such as albumin (co-cross-linking) and precipitating the resulting aggregate [24], [25], [49]. This method can also be used in combination with a carrier such as a membrane, where the physically adsorbed enzymes are cross-linked on the membrane surface [65]. Bi- or multifunctional reagents such as glutaraldehyde [*111-30-8*], toluene diisocyanate [*1321-38-6*], or bisdiazobenzidine derivatives can be used as cross-linking agents. The primary disadvantages of this method are the relatively severe reaction conditions and difficulties in controlling the reaction. Hence, activities of such biocatalysts are generally low [24], [25], [40], [41], [66].

2.2.2.3. Enzyme Copolymerization

The enzyme can be immobilized by copolymerizing it with the polymer matrix. Although the enzyme is then a covalently bound part of the matrix, this method can be regarded as a modification of gel entrapment (see Section 2.2.2.4). In the original method, proteins were vinylated with acylating or alkylating monomers and copolymerized with other monomers. The advantages of this method over matrix entrapment are the higher activity and higher stability due to the lack of enzyme bleeding [67], [68].

In enzyme graft copolymerization, enzyme copolymer is grafted onto a matrix such as a polysaccharide carrier [69], [70]. High mechanical strength can be attained, depending on the carrier chosen. High intra particle diffusion resistances can be circumvented if the enzyme copolymer is grafted to the surface of a carrier with low porosity [55]. For example, horseradish peroxidase (E.C. 1.11.1.7) was immobilized under strictly controlled conditions to give a biocatalyst with kinetic constants close to those of the native enzyme but with higher thermal stability [71].

2.2.2.4. Entrapment

Entrapment may be regarded as the physical confinement of an enzyme in a semipermeable matrix. The matrix lattice must be tight enough for the enzyme to be retained but must allow permeation of substrate and product(s). The enzyme is generally not chemically bound to the matrix as in enzyme copolymerization. Entrapment methods can be subdivided into gel entrapment and microencapsulation and reverse micelles.

Gel Entrapment. In this procedure, free enzyme is entrapped within the interstitial spaces of a cross-linked, water-insoluble polymeric gel. Gel-forming materials such as polysaccharides, proteins, or synthetic polymers can be employed. Among the natural polysaccharides, alginate, agar, and κ-carrageenan are widely used in preparing immobilized biocatalysts.

Polyacrylamide gel was the first material employed for entrapping enzymes [e.g., trypsin (E.C. 3.4.21.4), papain (E.C. 3.4.22.2), β-amylase (E.C. 3.2.1.2), and D-fructose 1,6-diphosphate aldolase (E.C. 4.1.2.13)] [72]. The polymerization of acrylamide can be initiated with potassium persulfate and by X- or γ-radiation [73]. Another method uses prepolymers that can be cross-linked by UV radiation; photo-cross-linkable resin prepolymers with varying hydrophilic – hydrophobic ratios, chain lengths, and anionic or cationic functional groups are available [74]. Such biocatalysts can also be used for bioconversion of highly lipophilic or slightly water-soluble substrates in reaction media consisting of water and a water-miscible organic cosolvent or water-containing organic solvents. Hydrophilic gel matrixes and organic solvents of low polarity are generally preferred for bioconversion of lipophilic substrates [75]. For further details of this method, the literature should be consulted [20], [24], [41], [43], [45], [55].

A special modification of this method involves enzyme entrapment during the spinning of synthetic fibers such as cellulose triacetate [76], [77]. High surface areas can be achieved by using very fine fibers, which also show chemical resistance to weak acid, alkali, and some organic solvents. Fiber-entrapped enzymes are used in the pharmaceutical industry (penicillin acylase E.C. 3.5.1.11 is used in the synthesis of antibiotics with a β-lactam ring) and food industry (β-D-galactosidase, E.C. 3.2.1.23, is used for the hydrolysis of lactose). The risk of microbial contamination is one disadvantage of this method; a reactor densely packed with fibrous material cannot be sterilized easily.

Microencapsulation. In microencapsulation, enzymes are immobilized by enclosing them in membranes that are permeable to the substrate and the product. The resulting microcapsules generally have a diameter of 1–100 μm. CHANG, who pioneered this approach, describes the procedures in detail [78]; see also → Microencapsulation. An emulsion is usually prepared from an organic phase and an aqueous enzyme-containing phase in the presence of a surfactant. The membrane-forming polymer, dissolved in an organic solvent, is then added. On standing, polymer precipitates on the microdroplet interface to form stable microcapsules [55], [79].

Unmodified coenzymes or multienzyme systems can also be immobilized by microencapsulation with lipid–polyamide membranes [80]. Microencapsulation gives the highest enzyme concentrations per unit volume of immobilized preparation [48]. However, as in all semipermeable membrane systems, the substrate diffusion rate limits the applications of microencapsulated biocatalysts.

Reverse Micelles. Amphiphilic surfactant molecules can form reverse micelles in hydrocarbon solvents. The micelles contain water (water pools) and can be used as hosts for biopolymers. The solubilized biopolymers (enzymes or even whole cells) retain their biological activity and are protected against the organic solvent by the surfactant envelope. Such systems may be applied for biocatalytic conversion of water-insoluble substrates or other reactions requiring the presence of an organic phase. The micelles must be kept in an appropriate reactor which allows physical separation of enzyme and reagents; use of membrane reactors seems to offer substantial advantages [81]. A review of the use of enzymes and nucleic acids in hydrocarbon micellar solutions is presented in [82].

2.2.3. Methods for Soluble Enzymes

In the immobilization methods described above, modification of the native enzyme or its microenvironment often decreases enzyme activity or selectivity. To keep the enzymes in their native state, an enzyme solution can be separated from the substrate and product by a *semipermeable membrane*. The membrane pores allow substrate and product diffusion and physically confine the larger enzyme molecules. Flat sheet ultrafiltration or microfiltration membranes or hollow-fiber membranes can be used (Section 5.6). Cofactors, which are normally small molecules that could diffuse through the membrane, can be retained in the reaction zone by coupling them to larger molecules. For example, nicotinamide–adenine dinucleotide (NAD) is completely retained by a hollow fiber cell when it is attached to a soluble poly(ethylene glycol) chain [83]. The enzyme solution is also protected by the membrane from microbial attack. Process selectivity can be controlled by proper design of membrane porosity and hydrophilicity. The main problems associated with membrane reactors are membrane fouling caused by the adsorption of diffusion species and mass-transfer resistances [45]. A review of membrane immobilization is given in [84]; see also → Membranes and Membrane Separations.

Limitation of diffusion through membranes can be circumvented by using *soluble–insoluble immobilized enzymes.* Such preparations are very useful for conversion of water-insoluble substrates such as cellulose. Cellulase from *Aspergillus niger* has been immobilized on poly(L-glutamic acid), which is soluble in neutral and alkaline solution but can be precipitated by lowering the pH without loss of enzyme activity [85]. Preparation of soluble–insoluble immobilized proteases such as papain (E.C. 3.4.22.2) or chymotrypsin (E.C. 3.4.21.1), their properties, and possible applications are discussed in [86].

2.3. Methods of Cell Immobilization

2.3.1. Introduction

Although self-immobilization of cells on surfaces commonly occurs in nature, the intentional attachment of cells to a support is a recent development. Immobilized whole-cell systems (microbial, yeast, plant, or mammalian cells) are much more versatile than the corresponding immobilized enzymes. Such biocatalysts can consist of intact or disintegrated dead cells that contain active enzymes. However, viable resting or

growing cells can also be immobilized; this technique is used especially with eukaryotic cells where the whole metabolic machinery is often required for their specific application.

The state of the art in immobilized cell technology is described in detail in [21], [22], [25], [40], [43], [66], [87–102]. As in enzyme immobilization, no "ideal" general method exists. This chapter merely gives an overview of the advantages and disadvantages of immobilized whole-cell systems and the most generally applied immobilization techniques.

The primary advantage of immobilized cell cultures as opposed to suspension cultures is the reduced cost of bioprocessing as a result of higher cell densities, easy separation, and repeated or continuous use of biocatalyst. Immobilization may also increase the shear resistance of shear-sensitive eukaryotic cells. The enzymes retain their native geometry and microenvironment, and are protected against changes in the bulk system (temperature, pH, ionic strength, mechanical stress). Support materials generally stabilize biocatalyst activity and storage stability, but this phenomenon is not fully understood.

Correct choice of the hydrophobicity – hydrophilicity balance of the support can alter the microenvironment by increasing the concentration of limiting substrate or by decreasing the concentration of inhibiting product in the vicinity of the biocatalyst [92], [103]. Poor diffusion and permeability into the carrier matrix and through the cell membrane limit the applicability of such biocatalyst systems to low molecular mass substrates or products. This transport barrier can protect the cells against microbial or viral infection if the pore width of the matrix is sufficiently small [92].

The reactions catalyzed by immobilized whole-cell biocatalysts can be classified as follows:

1) Reactions involving single enzymes (bioconversions)
2) Reactions involving multienzyme systems with or without cofactors
3) Reactions involving a complete metabolic pathway yielding primary or secondary metabolites

The use of immobilized microbial cells is advantageous in the following areas [96]:

1) when the desired enzymes are intracellular and the extracted, purified enzymes become unstable after immobilization;
2) when the microorganism does not contain interfering enzymes or when such enzymes can be inactivated without loss of desired catalytic activity; and
3) when substrates and products do not have a high molecular mass and can diffuse through the cell membrane.

The use of whole cells instead of isolated enzymes obviates costly, tedious purification procedures. In multienzyme reactions, the cell structure provides the proper geometric enzyme arrangement which is extremely difficult to achieve by enzyme co-immobilization. For reactions requiring cofactor regeneration or for the production of primary or secondary metabolites, the immobilized cells must be viable. If cell viability is not a prerequisite, diffusion problems can be solved by chemical or electrical permeabilization of the cell membrane.

The major advantage of such biocatalytic systems is that the operational stability can often be markedly enhanced by reactivating the enzymes in the immobilized cells [104]. Such reactivation can be achieved either by brief incubation with growth medium or by continuous controlled growth of the immobilized cells [105].

2.3.2. Cell Supports

A variety of inorganic and organic matrices are available for cell immobilization; they interact with the cell surface by physical or chemical bonds [106], [107]. A comprehensive list of commercial support materials is given in [46]. A number of factors must be considered in the selection of a suitable carrier for large-scale use [98], [100], [103]:

1) The material should be available in sufficient quantities and at low price.
2) The material should have a large surface area accessible to cells and reactants.
3) The material must be mechanically, chemically, and thermally stable under process and storage conditions.
4) The matrix should contain a sufficient number of functional groups to bind the cells.

5) The material should not reduce cell activity or initiate cell lysis.
6) The material should be easy to handle in the immobilization procedure.
7) The material should be capable of recycling or safe disposal.
8) In the case of viable growing cells, the matrix should have a sufficiently large void volume or be elastic enough to accommodate new cells.

2.3.3. Immobilization Techniques

Most of the common cell immobilization techniques are extensions or modifications of those discussed in Section 2.2.2 for enzymes. These methods can be grouped according to the way in which cells bind to the matrix (Fig. 3). An alternative way of classifying these methods is to use the route by which the biocatalyst is prepared [103]:

1) Carrier-free immobilization
2) Immobilization of a given biomass onto a preformed carrier surface
3) Immobilization of a given biomass during the course of carrier formation (e.g., by polymerization)
4) Immobilization by controlled growth of an inoculum or by germination of immobilized spores

2.3.3.1. Cell Immobilization without a Support

The simplest way to immobilize biocatalysts is to use their intrinsic tendency to aggregate or flocculate at high cell densities. Yeast, mold, and plant cells aggregate [90], [95], [99], [104] and are relatively stable to shear fields in fluidized-bed reactors [99], [108]. Flocculation can also be induced by polyelectrolytes such as chitosan [93], [99], [104].

Cell aggregation can be induced by low molecular mass bi- or multifunctional reagents such as glutaraldehyde, diazotized diamines, or toluene diisocyanate, but the toxicity and the high reactivity of these substances limit their applicability in immobilizing viable cells. Examples of immobilization by intercellular aggregation are given in [93], [100], [104].

2.3.3.2. Binding of Cells to a Carrier

Physical Adsorption. Immobilization of microorganisms or cells by adhesion or adsorption to a preformed matrix is based on their inherent affinity for solid surfaces during growth. Such immobilization can generally be achieved by keeping the inert carrier and the actively growing cells in contact for a defined period. Additional chemicals are usually unnecessary, and fixation is carried out under growth conditions; viable cell preparations can, therefore, be obtained. Examples of cells immobilized by adhesion or adsorption to various support materials are listed in [100], [104].

Although the method is very simple, mild, and versatile, its use is limited by adhesion capacity and adhesive strength, which usually depend on environmental conditions (pH, ionic strength) [93]. Use of a high flow rate as a method of selection in a fluidized bed yields good results. Such surface-attached cell systems (*biofilms*) have been used extensively in wastewater treatment [95], [109], [110].

A promising application of this method is the use of mammalian cells bound to preformed surfaces to produce therapeutic biochemicals. The cells are immobilized on microcarriers (small-diameter beads, 100 – 200 μm) manufactured from different synthetic polymers (e.g., polystyrene, gelatin, dextran, polyacrylamide, or glass) that offer a large specific surface area for cell growth [111–113]. The primary weakness of this method is the exposure of surface-grown cells to mechanical stress resulting from bead – bead collisions and impeller impact. The number of cell divisions is also restricted by the limited outer surface area. Both problems have been partially solved by the development of macroporous microcarriers prepared from gelatin, dextran, polyacrylamide, or agarose [113], [114]. Cells grow on the much larger interior surface of the porous structure and are also protected against mechanical stress.

A monolithic ceramic matrix has been developed for large-scale cell cultures. It consists of a ceramic cylinder with uniform longitudinal channels of square cross section. The matrix has a surface area of 32 cm^2 per cubic centimeter of volume. A matrix with a very smooth surface

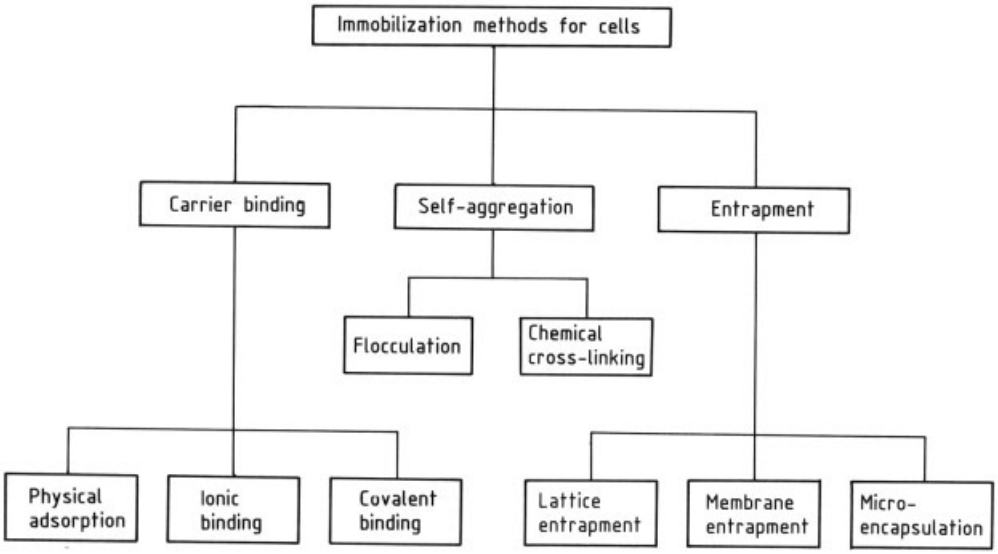

Figure 3. Classification of cell immobilization methods

is used for the growth of anchorage-dependent cells. Nonadherent cell types (e.g., hybridoma cells) are cultured in a matrix with a very rough surface and are physically entrapped in the pores of the channel walls. Unlike microcarrier cell cultures, no shear problems occur in the ceramic matrix [113], [115].

For adherent cell growth, the scalability is almost linear and depends on the available surface area. For immobilized hybridoma cultures, scaleup is based on the volume of the porous matrix [116].

Ionic Binding. Ionic binding is a special case of physical adsorption where charged microbial cells can electrostatically interact with the ions on a carrier surface to form stable complexes. Synthetic ion-exchange resins, modified cellulose derivatives, or inorganic materials can be used as carriers. Cell adsorption is mainly affected by factors such as pH, ionic strength, surface charge, cell age, or composition of the carrier surface. Examples for the application of this method are given in [93], [98].

Covalent Binding. Cells can also bind to the functional groups of the carrier surface by covalent bonds. This technique is frequently used for enzyme immobilization but has only limited use in whole-cell immobilization because the toxicity of the coupling agents often results in loss of cell viability or enzyme activity [66], [104].

The main advantages of this method are reduced or eliminated cell washout and improved mechanical strength of the biocatalyst without increasing diffusion barriers. The problems associated with various organic and inorganic supports, along with coupling techniques, are discussed in [104]; applications are reviewed in [100]. The preparation, properties, and experimental applications of eukaryotic cells covalently bound to solid supports are described in [117].

2.3.3.3. Immobilization of Cells by Entrapment

Cells may be immobilized by entrapment with a polymer lattice, a membrane, or microcapsules.

Lattice Entrapment. The widely used method of lattice entrapment involves retention of cells within the network of a polymer matrix. The lattice is tight enough to prevent release of cells, while allowing diffusion of substrates and products. The cell does not bind to the matrix, and the polymer network is generally formed in the presence of the cells to be entrapped. Hydrophilic porous matrices are formed under mild enough conditions to allow cell entrapment

with minimum loss of viability and activity. The prepared biocatalyst forms pellets which should have good mechanical and chemical stability, as well as high cell loading capacity. The high viable cell density results in an enhanced demand for nutrients and oxygen, which must overcome the diffusion barriers of the gel structure.

In this section, only the most extensively used matrices (e.g., polyacrylamide gel, alginate gel, κ-carrageenan, and photo-cross-linkable resins) are discussed. More detailed information on other systems can be found in [21], [90].

Polyacrylamide Gel. Polyacrylamide gel was used to immobilize whole cells in the first successful application of lattice entrapment used for the biotechnological production of biochemicals [118]. The same technique was also employed in the production of L-aspartic acid [119]. The method is based on free-radical polymerization in an aqueous acrylamide monomer solution containing the cells at low temperature. The degree of cross-linking determines the porosity and mechanical properties of the pellets. The major disadvantage of the method is the toxicity of the acrylamide monomer, the cross-linking agent (e.g., *N,N′*-methylenebisacrylamide [*110-26-9*]), and the polymerization initiator (e.g., *N,N,N′,N′*-tetramethylethylenediamine [*110-18-9*]) which can decrease cell viability and enzyme activity [96], [120]. Hydroxyethyl methacrylate [*868-77-9*] can be used as a less toxic alternative to acrylamide [43], [121].

With acrylamide monomers, optimal results are achieved in isotonic buffered solution at low temperature (4 – 12 °C) by keeping exposure time to a minimum (3 – 4 min). Examples of successful applications of this method are given in [96], [98].

Alginate [*9005-38-3*] is extracted from seaweed and is a linear copolymer of β-D-mannuronic acid and α-L-guluronic acid linked by 1,4-glycosidic bonds. It forms a gel in the presence of multivalent ions, usually calcium or aluminum. The controlled entrapment of cells is simple and generally nontoxic. Various cell types can be immobilized with negligible loss of viability. However, the matrix can be solubilized in the presence of Ca^{2+}-chelating agents such as phosphate, citrate, or ethylenediaminetetraacetic acid. The alginate matrix is mechanically weak so that the growing cells (especially plant cells) can be released from, or even disintegrate, the beads. Another drawback of the alginate method for viable animal cells is the difficulty of producing sufficiently small beads to overcome oxygen limitation in their interior [113].

Strengths and weaknesses of the calcium alginate method are discussed in [122]; examples of applications are given in [96], [98], [100].

κ-Carrageenan [*11114-20-8*] is a readily available, nontoxic, high molecular mass, sulfonated polysaccharide extracted from seaweed. It consists of β-D-galactose 4-sulfate and 3,6-anhydro-D-galactose units. Only the sodium salt of κ-carrageenan is soluble in cold water. The gelation characteristics of κ-carrageenan are discussed in [123]. Induction of gelation with potassium ions can be performed under very mild conditions without the use of chemicals that decrease enzyme activity. Another advantage of this technique is that the immobilized cell biocatalysts can be tailor-made into pellets, sheets, beads, etc. [96].

Two entrapment procedures are available. In the one-step procedure, a large number of cells are entrapped in the gel. The cell suspension in physiological saline is mixed with soluble κ-carrageenan gel under sterile conditions, and the gel beads are incubated in a nutrient medium [98], [122]. Another very elegant method involves the use of photo-cross-linkable resins [74], [124], [125]. Active groups (e.g., vinyl groups) are coupled to polyethylene or poly(propylene glycol) oligomers of controlled chain length to prepare prepolymers. The prepolymers are then mixed with the cell culture, and cross–linking is initiated with UV radiation. The advantages of this method are:

1) entrapment is simple and proceeds under very mild conditions;
2) prepolymers do not contain toxic monomers;
3) the network structure of the gels can be adapted as required; and
4) optimal physicochemical gel properties can be achieved by selecting suitable prepolymers [125].

Membrane Entrapment. In membrane entrapment, the cells are not held in a porous matrix (as in lattice entrapment), but are retained by a semipermeable membrane barrier. The membrane allows the diffusion of soluble material

to and from the immobilized cells while retaining and protecting enclosed organisms. Hollow-fiber bioreactors, in which the organism is confined on one side of the porous fiber and the soluble substrate and products on the other, have found the widest application. Higher cell densities can be achieved than with lattice-entrapped biocatalysts [126], but growth must be controlled to prevent excessive buildup of biomass and rupture of the membrane [99]. The supply of oxygen and nutrients to the cells and the removal of carbon dioxide from the cells occur by diffusion and may be insufficient. Model equations for mass transfer in hollow-fiber reactors are discussed in [127], [128].

The state of the art of membrane reactors is described in [84]. Developments in hollow-fiber reactors are discussed in [129], and for flat-bed hollow-fiber reactors in [130]. A flat membrane reactor with different compartments for cells, medium, and product for large-scale cultivation of mammalian cells is described in [131].

Microencapsulation. In microencapsulation, cells are placed inside an alginate bead that is coated with a semipermeable membrane. The membrane is usually composed of calcium alginate and poly-L-lysine. After being coated the alginate gel can be solubilized with sodium citrate, leaving the cells entrapped in hollow spheres where they can be cultured. Different capsule diameters (from 20 nm to 2 mm) and controlled membrane porosity can be achieved [127]. In larger microcapsules, however, radial concentration gradients of nutrients or oxygen can result in a necrotic core [127].

This well-known enzyme immobilization technique (Section 2.2.2.3) was first adapted for the whole-cell immobilization of *Micrococcus denitrificans* [132] and of viable mammalian cells [133]. Applications of this type of immobilized biocatalyst [134] include the production of monoclonal antiboides from hybridoma cells [135].

2.4. Methods of Organelle Immobilization

Enzymes in living cells may be associated in supramolecular structures, which are assembled into organelles such as mitochondria, chloroplasts, peroxisomes (microbodies), vacuoles, and nuclei. Such organelles have their own distinct metabolic functions and, therefore, provide excellent immobilized biocatalysts that can carry out multistep, multifunctional reactions. Immobilization generally increases organelle longevity.

Examples of the uses of immobilized organelles are adenosine 5′-triphosphate (ATP) regeneration, metabolite oxidation, and photoproduction of hydrogen [136]. Most research has been devoted to the isolation and immobilization of chloroplasts as potential solar energy photoconverters [136], [137]. Older work on organelle immobilization is reviewed in [137], [138]. Entrapment techniques using photo-cross-linkable prepolymers are described in [125].

Immobilization of rat liver mitochondria on a solid support consisting of alkysilanized glass beads was first reported in 1975 [140]. Rat liver mitochondria have also been entrapped in polyacrylamide gel [141]; yeast mitochondria have been entrapped in a polyurethane matrix [142].

2.5. Co-immobilization of Biocatalysts

Co-immobilization refers to the simultaneous immobilization of specific combinations of cells, organelles, or enzymes that show complementary biocatalytic activity. The synergistic advantages of co-immobilization are, however, limited by the biological compatibility of the components [143].

Co-immobilization can be achieved by coentrapment of preimmobilized enzymes and whole cells in a single matrix or by direct coupling of enzymes to the microorganisms [25], [143]. The former method was used to immobilize galactosidase and *Saccharomyces cerevisiae* for the conversion of cellobiose to ethanol [144]. Yeast cells or dead mycelia have been employed for direct coupling of enzymes to cell walls in food-processing applications [25], [143]. Another interesting application is the bioproduction of hydrogen gas with hydrogenase and chloroplasts [145]. Co-immobilization results in a fivefold increase in hydrogen production compared with the system employing two separately immobilized catalytic species. Recent results on integral

hydrogen-producing biocatalysts are presented in [139].

Further information on the state of the art of co-immobilization, its limitations and prospects can be found in [25], [90], [143], [146], [147].

3. Activity and Kinetics of Immobilized Biocatalysts

3.1. Effects of Immobilization on Enzyme Activity

Immobilization can be considered a random binding process which may, therefore, negatively influence the active site of an enzyme to varying degrees, even in the same immobilization procedure. A lower net specific activity of the immobilized enzyme is observed. The degree of *activity retention* depends on the immobilization method used. Major reasons for the loss of enzymatic activity in immobilization are listed below and illustrated in Figure 4:

1) Enzyme may be denatured by reactants or products involved in the immobilization procedure (entrapment matrix formation, covalent binding reaction, or cross-linking reaction).
2) Immobilization conditions may lead to denaturation or deactivation.
3) Enzyme denaturation may be caused by a reactive group in the active site participating in the binding reaction.
4) Binding forces may hold the enzyme molecule in an inactive configuration (conformational changes).
5) Orientation of the bound enzyme molecule relative to the support may prevent substrate access to the active site or product release (steric hindrance).

Measurement of activity retention after immobilization is difficult because of mass-transfer effects. It seems to decrease with increasing amount of enzyme bound to the matrix and increasing immobilization time. It also varies greatly with different supports. The *coupling yield* represents the binding capacity of a carrier and is defined as

$$\text{Coupling yield} = \frac{(\text{Protein bound}) \times (\text{Volume of immobilization solution})}{(\text{Total protein used in immobilization}) \times (\text{Mass, volume, or surface area of carrier})}$$

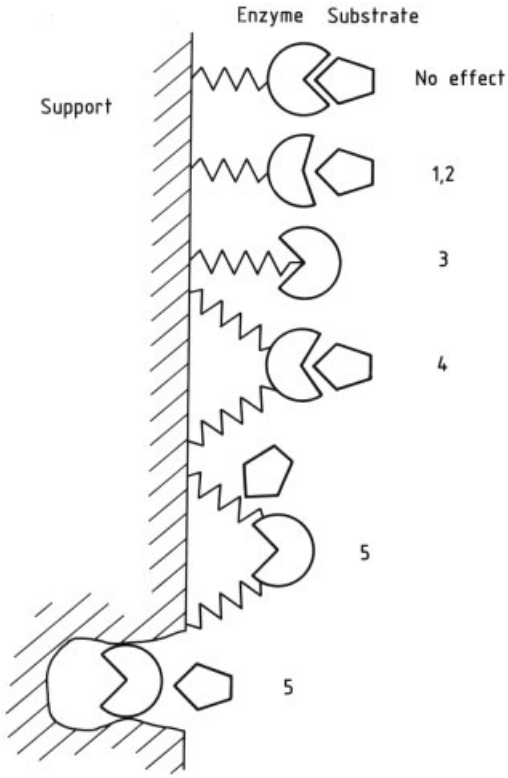

Figure 4. Effects of immobilization on enzyme activity (see text for further explanation)

Coupling yield gives no information about the amount of active protein immobilized. In practice, however, the amount of activity retained after immobilization must be known. In the presence of mass-transfer effects, the amount of immobilized active protein can only be determined accurately by *active site titration*; this method is based on stoichiometric and reversible binding of a specific titrant. Unfortunately, it can be applied to very few enzymes (e.g., trypsin [148]). Other methods aiming at the separation of mass transfer and kinetics are usually less precise and require substantial experimental and computational effort [149] (see also Chap. 4).

Stabilization. Immobilized enzymes and native enzymes are largely deactivated by the

same effects. However, immobilization often increases enzyme stability, probably due to stabilization of the tertiary protein structure.

The elimination of autolytic denaturation of hydrolases is another beneficial effect of immobilization, especially by covalent binding methods [150]. Some physical methods (e.g., the use of membranes or microencapsulation) in which the enzyme remains in the soluble form, do not have a protective effect.

An effective application of immobilized biocatalysts is their use in continuous reactors. This requires high biocatalyst stability, which can be obtained by using appropriate immobilization methods and modifications of the biocatalyst microenvironment. Various additives such as antioxidants, salts, polyols, sugars, and other polymers may be especially useful for this purpose. Biocatalyst stabilization by immobilization or solute addition is discussed in [151]. The effect of cross-linking is treated in [152].

Electrostatic Effects. An apparent shift in the pH optimum of a biocatalyst immobilized on an ionic support may be observed due to microenvironmental effects caused by the charge of the matrix. The local pH of a negatively charged matrix (in which the enzyme is located) is lower than that of the bulk solution, which causes the pH optimum observed in the bulk solution to increase (Fig. 5, curve c). A positively charged matrix has a higher local pH than the bulk solution, thus shifting the apparent pH optimum to a lower value (Fig. 5, curve a).

A thermodynamic expression has been derived for the shift of the pH activity profile for an enzyme immobilized to a charged support [153]:

$$\Delta \mathrm{pH} = \mathrm{pH}^{\mathrm{local}} - \mathrm{pH}^{\mathrm{bulk}} = 0.43\,\frac{zF\psi}{kT}$$

where z is the valence of the substrate, F is the Faraday constant, ψ is the electrostatic potential, k is the Boltzmann constant, and T is the absolute temperature.

Analogously, for the change in the Michaelis constant K_{M} (for definition of K_{M}, see Section 3.2):

$$\Delta \mathrm{p}K_{\mathrm{M}} = \mathrm{p}K_{\mathrm{M}}^{\mathrm{app}} - \mathrm{p}K_{\mathrm{M}} = \log\frac{K_{\mathrm{M}}}{K_{\mathrm{M}}^{\mathrm{app}}} = 0.43\,\frac{zF\psi}{kT}$$

where the superscript app denotes apparent.

Similarly, transfer of a charged substrate through the external liquid film to an enzyme located on the surface of the support may influence enzyme kinetics. Several expressions for $K_{\mathrm{M}}^{\mathrm{app}}$ are given in [154].

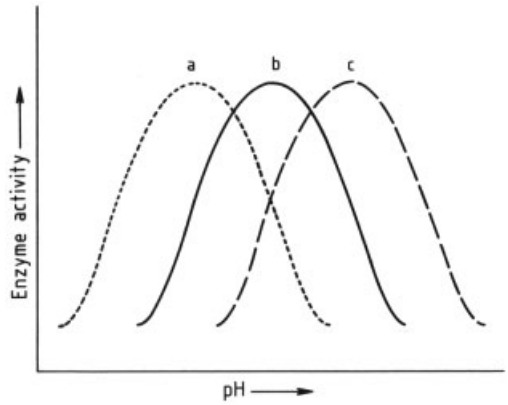

Figure 5. Electrostatic effects on pH optimum of enzyme activity
a) Positively charged matrix; b) Native enzyme; c) Negatively charged matrix

Polarity. The polarity of a support is one of the parameters determining its affinity for the enzyme substrate. The attractive or repulsive forces alter the substrate concentration in the microenvironment of the support (i.e., partition of substrate), thus influencing the enzymatic reaction rate.

The reaction rate of a polar substrate can be increased by using polar supports for immobilization; hydrophobic supports increase the reaction rates of apolar substrates.

Nonuniform Distribution of Enzyme in the Matrix. Biocatalysts with nonuniform activity are of two main types:

1) Composite catalysts consisting of uniformly active and inactive layers
2) Biocatalysts with a continuous activity profile. The distribution can be convex (obtained from incomplete adsorption in a support) or concave (obtained by incomplete back washing of a support of uniform initial profile).

Numerical simulation shows that a biocatalyst with activity concentrated in the shell (convex distribution) is always more effective

for enzymatic reactions of positive order such as Michaelis – Menten kinetics or competitive product inhibition. A catalyst whose activity increases toward its core (concave distribution) can give greater utilization efficiency for substrate-inhibited reactions [155].

3.2. Kinetics of Immobilized Biocatalysts

The rate of reaction catalyzed by a soluble enzyme can usually be described by a Michaelis – Menten type of kinetic equation (→ Enzymes, Chap. 2.). The simplest mechanism for such a reaction assumes the reversible formation of an enzyme – substrate (ES) complex, which decomposes irreversibly into product (P) and free enzyme (E):

$$S + E \rightleftharpoons ES \longrightarrow E + P$$

The well-known Michaelis – Menten equation is

$$r_S = \frac{v_{max} C_S}{K_M + C_S}$$

where r_s is the reaction rate of the substrate, v_{max} is the maximum reaction rate, K_M is the Michaelis constant, and C_S is the substrate concentration. The reaction rate is a function of the substrate and enzyme concentrations, pH, temperature, solvent composition, and ionic strength.

The corresponding reaction mechanism for an immobilized enzyme is more complicated because (1) the enzyme activity may be changed by immobilization and (2) the substrate must be transported to the support and the product must be transported from the reaction site back into the bulk solution:

$$\begin{array}{ccccc} S_{bulk} & & & & P_{bulk} \\ \uparrow\downarrow & & & & \uparrow\downarrow \\ S_{surface} & + & E_{immob.} & \rightleftharpoons & ES_{immob.} \longrightarrow P_{surface} \end{array}$$

The reaction rate is governed by additional parameters resulting from immobilization, i.e., conformational changes, chemical changes, partitioning effects (electrostatic, hydrophobic – hydrophilic interactions), and diffusional effects (i.e., external and internal mass transfer; see Chap. 4).

Mass-transfer resistance must be considered in kinetic measurements or when modeling processes with immobilized biocatalysts. The familiar Michaelis – Menten equation can be adapted by using the effectiveness factor η:

$$r_S = \eta \frac{v_{max} C_{SA}}{K_M + C_{SA}}$$

or in its linearized form

$$\frac{1}{r_S} = \frac{K_M}{\eta v_{max}} \frac{1}{C_{SA}} + \frac{1}{\eta v_{max}}$$

where C_{SA} is the substrate concentration at the surface. Since both the substrate concentration at the surface C_{SA} and the effectiveness factor η can vary with the bulk substrate concentration C_S, the double reciprocal Lineweaver – Burk diagram and other linearized forms of the Michaelis – Menten equation are no longer linear. In Figure 6 it can be seen that the influence of mass transfer decreases at high substrate concentration and disappears totally when this concentration becomes infinite. The effectiveness factor η changes with the bulk substrate concentration, reaching its lowest value $\eta(0)$ when $C_{SA} = 0$.

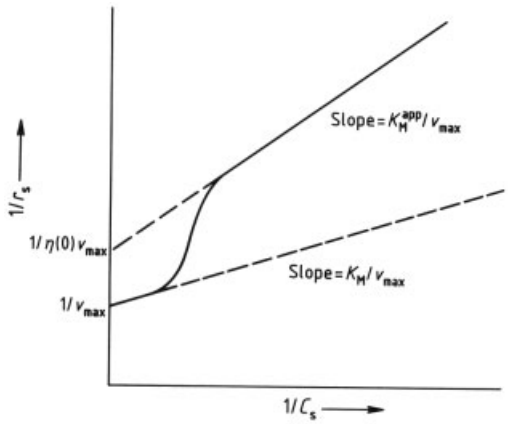

Figure 6. Influence of mass-transfer resistance on the kinetics of immobilized enzymes in a Lineweaver – Burk plot
At low substrate concentrations, the values of the apparent kinetic constants are determined. At high substrate concentrations, the effectiveness factor η approaches unity and, at least in theory, intrinsic kinetic parameters can be determined.

3.3. Assay of Immobilized Biocatalysts

Immobilized biocatalysts are usually particulate or filamentous; therefore, conventional assay methods for soluble enzymes can rarely be applied without modification. Activity can be measured in a stationary catalyst bed, a homogeneous biocatalyst suspension, or a fluidized-bed reactor.

The *stationary catalyst bed* (column reactor) has the disadvantages of axial substrate and product concentration gradients and the danger of channeling. These can be overcome by using a very low catalyst bed height or a high flow rate.

Homogeneous biocatalyst suspension (stirred-tank reactor) avoids concentration gradients. However, to obtain a homogeneous suspension, intensive mixing is required and this can lead to mechanical destruction of the biocatalyst.

The *fluidized-bed reactor* may be considered a compromise between the two preceding methods. To achieve sufficiently dense fluidized beds, high-density particles and columns with internal baffles may be needed. Advantages are a low pressure drop across the bed, minimum particle attrition, no channeling, and no bed clogging.

Analytical methods are discussed in [149], [156]. They may be discrete or continuous; in the latter case, conventional methods cannot be applied without some modification. Most important is to work under well-defined, reproducible conditions.

Summarizing, immobilization is one of the major considerations in the design of bioreactors. Beside the advantages of continuous or repetitive processes, high stability, fast reaction, and easy separation from the reaction medium, immobilized biocatalysts also have disadvantages such as mass-transfer resistance, necessity of immobilization procedures, and additional costs of immobilization reagents. A comprehensive review of these engineering aspects with respect to reactor design is given in [157].

4. Mass Transfer in Immobilized Biocatalyst Systems

The retention of immobilized enzymes and cells usually requires the presence of solid carrier phase or flocculant cell mass. Thus, mass transfer through a liquid diffusion film from the bulk liquid to the surface of the biocatalyst may have a significant effect on biocatalyst activity; it is referred to as *external mass transfer.* Because the reaction site may be located within a gel, porous solid, biofilm, or biofloc, transfer of substrate from the exterior surface to a point within the carrier is usually necessary. This is called *internal mass transfer* or intra particle transport. Reviews on transfer effects with immobilized biocatalysts are found in [158–160].

4.1. External Mass Transfer

To reach the reaction site, the substrate is transported from the bulk liquid by convection to the stagnant diffusion film (point A in Fig. 7) and by diffusion through the film (from A to B) to the reaction site, which may be on the surface of the carrier (point B) or within the carrier (from B to C).

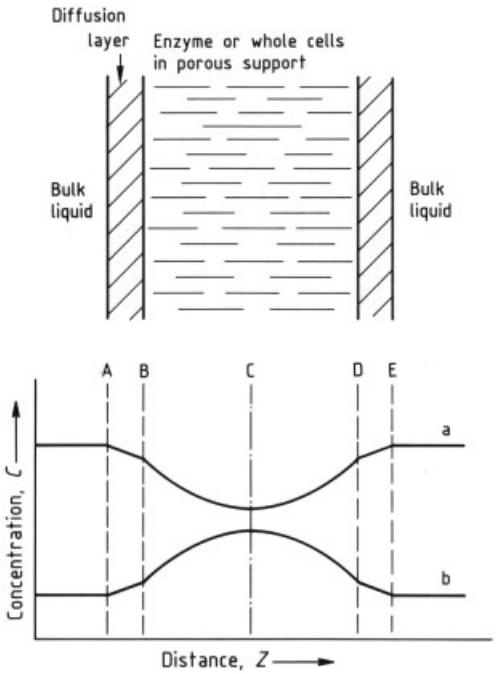

Figure 7. Concentration profiles around and in a porous matrix symmetrically exposed to bulk liquid
a) Substrate; b) Product

Consider the reaction of a substrate which is transferred from the bulk liquid to an enzyme immobilized on a nonporous surface (point B in Fig. 7). For a first-order reaction (or Michaelis–Menten kinetics with $S < K_M$) at steady state (i.e., diffusion equal to surface reaction rate), the apparent reaction rate per unit surface area (r_{app}) is

$$r_{app} = k_{SL}\,(S_A - S_B) = k_s S_B \tag{1}$$

where S_A is the substrate concentration at point A, S_B the substrate concentration at point B, k_{SL} the mass-transfer coefficient for the liquid film at the solid–liquid interface, and k_s the surface reaction rate constant. Solving for S_B and rearranging give

$$r_{app} = \frac{k_s}{(k_s/k_{SL}) + 1} \cdot S_A \tag{2}$$

Two extremes can be identified. For $k_s/k_{SL} \gg 1$, S_B approaches zero (Eq. 1), and the reaction is completely mass transfer controlled with an apparent rate of $k_{SL}\,S_A$. For $k_s/k_{SL} \ll 1$, S_B approaches S_A, and the reaction is kinetically controlled with an apparent rate of $k_s S_A$. In the intermediate situation given by Equation (2), the apparent reaction rate is influenced by both the rate constant (k_s) and the mass-transfer coefficient (k_{SL}).

For a *zero-order reaction,*

$$k_{SL}\,(S_A - S_B) = k_S \tag{3}$$

where k_s is a zero-order rate constant. The substrate concentration at the reaction surface (S_B) is thus

$$S_B = S_A - (k_s/k_{SL}) \tag{4}$$

The ratio of the rate constant to the mass-transfer coefficient determines S_B, but does not influence the overall rate.

For Michaelis–Menten kinetics between zero and first order, Equation (1) becomes

$$r_{app} = k_{SL}\,(S_A - S_B) = v_{max} S_B / (K_M + S_B) \tag{5}$$

Solving Equation (5) for S_B shows that all coefficients influence the overall reaction rate and also that external transfer changes the overall reaction in such a way that it does not follow the Michaelis–Menten form with respect to bulk concentration. Thus, the apparent kinetics of an immobilized enzyme differ from the intrinsic kinetics of the native enzyme. This can be seen most easily from Equation (5) when S_B is low. In this case, the surface rate reduces to $v_{max}\,S_B/K_M$ and the apparent rate becomes

$$r_{app} = \frac{(v_{max}/K_M)\,S_A}{(v_{max}/K_M k_{SL}) + 1} \tag{6}$$

i.e., the apparent rate depends on the transfer coefficient k_{SL}.

Generally, true kinetics can be determined only by operating in the kinetic regime without diffusion. This can be achieved by reducing the ratio of v_{max}/k_{SL} sufficiently as seen from Equation (6). Thus, transfer-promoting parameters (flow rate in packed and fluidized beds, stirring speed in tanks) should be increased, and those that promote reaction rate should be decreased (enzyme loading, temperature).

4.2. Prediction and Correlation of External Mass-Transfer Coefficients

4.2.1. Catalyst Beds

The approximate value of k_{SL} for a bed of packed or fluidized particles can be predicted by using data from the literature, which are usually presented as dimensionless correlations having the form

$$\frac{k_{SL} d_p}{D} \sim \left(\frac{d_p v_s \varrho}{\mu}\right)^{n_1} \left(\frac{\mu}{\varrho D}\right)^{n_2} \tag{7}$$

where ϱ is the liquid density, μ the viscosity, v_s the superficial linear velocity, and n_1 and n_2 are empirical constants.

$$N_{Sh} \sim (N_{Re})^{n_1}\,(N_{Sc})^{n_2} \tag{8}$$

These are referred to as Sherwood–Reynolds–Schmidt number (N_{Sh}, N_{Re}, N_{Sc}) correlations. Experimentally, k_{SL} has been found to vary with $D^{2/3}$, which on rearrangement of Equations (7) and (8) gives the j_D factor or Chilton–Colburn correlation:

$$j_D = \frac{k_{SL}}{v_s}\left(\frac{\mu}{\varrho D}\right)^{2/3} = f\,(N_{Re}) \tag{9}$$

Correlations of the above form can be found in [161–163]. For the region of low relative velocity, theory and experiment show that when the Péclet number N_{Pe} (= *d/D*) is greater than 500 [165], the following holds:

$$\frac{k_{\mathrm{SL}} d_{\mathrm{p}}}{D} = 0.99\,(N_{\mathrm{Re}})^{1/3}\,(N_{\mathrm{Sc}})^{1/3} = 0.99\,(N_{\mathrm{Pe}})^{1/3} \tag{10}$$

For high flow rates ($N_{\mathrm{Re}} > 1$), theory and experiment give

$$N_{\mathrm{Sh}} = 0.5\,(N_{\mathrm{Re}})^{1/2}\,(N_{\mathrm{Sc}})^{1/3} \tag{11}$$

Fluidized beds are more complex because the void fraction ε also varies with velocity. This effect has been taken into account by replacing v_{s} by $v_{\mathrm{s}}/(1-\varepsilon)$ [164].

Dimensionless groups are thus useful in correlating experimental data and predicting mass-transfer coefficients. However, during experimentation, measurement of the dependency of individual operating parameters and fluid properties on the mass-transfer coefficient or on the overall reaction rate is of interest. Their independent influence on k_{SL} can be deduced from dimensionless correlations given in the literature. From Equation (7),

$$k_{\mathrm{SL}} = f\left(d_{\mathrm{p}}^{n_1-1}, v^{n_1}, \mu^{n_2-n_1}, D^{1-n_2}, \varrho^{n_1-n_2}\right) \tag{12}$$

In Table 5, the values of these exponents for stagnant flow ($N_{\mathrm{Sh}} = 2$), low velocities (Eq. 10), and high velocities (Eq. 11) are summarized.

4.2.2. Particles in an Agitated Tank

The prediction and correlation of stirred-tank mass-transfer data are complex because of the difficulty in defining a fluid velocity and the fact that the particle – fluid relative velocity depends on particle size and density. These effects have been characterized by the terminal velocity of a falling particle as calculated by Stoke's law, with

$$N_{\mathrm{Sh}} = 2.0 + 0.6\,(N_{\mathrm{Re}})^{1/2}\,(N_{\mathrm{Sc}})^{1/3}$$

being used to calculate a terminal velocity mass-transfer coefficient k_{SLt} [164]. Much of the data can be approximated by

$$k_{\mathrm{SL}} = 2k_{\mathrm{SLt}}$$

The power input of the stirrer per unit volume of fluid (*P/V*) has been used to replace the velocity in the term N_{Re}, which results in the following nondimensionless correlation [165]:

$$\frac{k_{\mathrm{SL}} d_{\mathrm{p}}}{D} \sim \left(\frac{(P/V)\,\mu}{\varrho}\right)^{1/4} \left(\frac{\mu}{\varrho D}\right)^{2/3}$$

The individual dependencies are

$$k_{\mathrm{SL}} \sim (P/V)^{1/4},\ d_{\mathrm{p}}^{0},\ \mu^{-5/12},\ \varrho^{5/12},\ D^{2/3}$$

Other data give

$$k_{\mathrm{SL}} \sim (P/V)^{0.2},\ d_{\mathrm{p}}^{0},\ \mu^{-0.4},\ D^{0.7}$$

Another dimensionless correlation of the form [161]

$$\frac{k_{\mathrm{SL}} d_{\mathrm{p}}}{D} \sim \left(\frac{(P/V)\,d_{\mathrm{p}}^{4}\varrho^{2}}{\mu^{3}}\right)^{1/3} \left(\frac{\mu}{\varrho D}\right)^{1/3}$$

has been used, which gives the individual dependencies

$$k_{\mathrm{SL}} \sim (P/V)^{1/3},\ d_{\mathrm{p}}^{1/3},\ \mu^{-2/3},\ \varrho^{1/3},\ D^{2/3}$$

Because *P/V* cannot be measured easily, the Reynolds number of the impeller (N_{ReI}) is also used, which for a given geometry in the fully turbulent regime gives

$$\frac{k_{\mathrm{SL}} d_{\mathrm{p}}}{D} \sim \left(\frac{N d_{\mathrm{I}}^{2}\varrho}{\mu}\right)^{3/4}$$

or

$$N_{\mathrm{Sh}} \sim (N_{\mathrm{ReI}})^{3/4}$$

This correlation suggests that k_{SL} is proportional to $N^{3/4}$, as is found for glass – enzyme particles, whereas small gel – enzyme particles, which have a low density difference, exhibit a much lower sensitivity to stirrer speed, $k_{\mathrm{SL}} \sim N^{1/5}$ [165].

Table 5. Values of exponents for parameter dependency of external mass-transfer coefficients in packed columns

Variable	Exponent		
	Stagnant flow: $v=0$, $N_{Sh}=2$, $k_{SL}=2D/d_p$	Low flow: $N_{Re}<1$, $k_{SL}=(2-5)\times10^{-3}$ cm/s	High flow: $N_{Re}>1$, $k_{SL}=(10-40)\times10^{-3}$ cm/s
Particle diameter d_p	−1	−2/3	−1/2
Flow velocity v	0	+1/3	+1/2
ϱ_L	0	0	+1/6
μ	0	0	−1/6
Diffusivity D	+1	+2/3	+2/3

4.3. Internal Mass Transfer [161], [166], [167]

The uptake of substrates and products within solid matrices (internal mass transfer) occurs by diffusional transport, whose driving force is a concentration gradient. The reaction rate within the matrix normally depends on the substrate and product concentrations. A mathematical model is extremely useful for understanding and predicting the competing effects of diffusion and reaction. The fundamentals of modeling methods are discussed below.

Consider a solid matrix, such as a membrane, and a substrate A with an initial concentration $C_A = C_{A0}$ on the left of the matrix and zero on the right ($C_A = 0$). This concentration difference causes a flux of substrate A from left to right, and the flux rate j_A at any point in the matrix is given by Fick's law

$$j_A = -D_{A-M}\frac{dC_A}{dZ}$$

where D_{A-M} is the diffusivity of A in the matrix material. If no reaction occurs in the matrix, a steady-state concentration gradient with a constant slope is established (solid line, a, in Fig. 8). Because the gradient is constant, the amount of A entering any region from the left is equal to the amount of A leaving at the right and no accumulation occurs within the matrix.

If cells or enzymes are present within the matrix which convert substrate A to product, the substrate profile is curved (curve b in Fig. 8). This occurs because the substrate is consumed in the reaction and the flux of substrate from the left exceeds that leaving at the right. A mass balance can be written over a segment of the matrix.

Under steady-state conditions, the accumulation is zero, and in the absence of reaction the diffusion rates in and out must be equal. If reaction occurs, the difference between the diffusion rates is equal to the reaction rate. Balances of this form can be written for any component; a complete description usually requires balances for every component participating in the reaction.

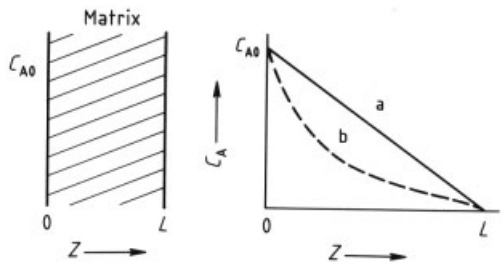

Figure 8. Flux of substrate A (initial bulk concentration C_{A0}) across a solid matrix
a) No reaction; b) Reaction

The reaction rate at each position in the immobilized biocatalyst matrix is generally a function of local substrate and product concentrations and thus varies from position to position. These local rates cannot normally be measured; only the net rate can be determined by monitoring concentration changes in the bulk liquid.

Consider a situation in which the substrate diffuses into the matrix from one side and does not penetrate completely because it is consumed by reaction with a biocatalyst (Fig. 9). Under steady-state conditions, the diffusion rate of substrate into the matrix equals the net reaction rate. Thus, the flux J_A multiplied by the matrix surface area A can be equated to the average volumetric reaction rate r_{avg} multiplied by the volume of the matrix V_M

$$(j_{A,Z=0})\,A = r_{avg}V_M$$

In both batch reactors and well-mixed continuous reactors, the diffusion flux at the interface

and hence the substrate profile within the biocatalyst are determined by the substrate concentration in the reactor C_{A1}. In the batch reactor, concentrations and rates vary with time.

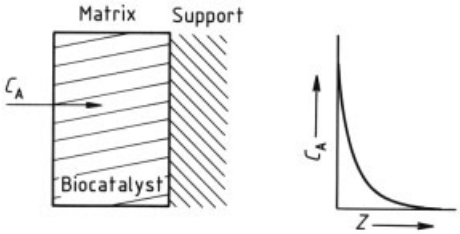

Figure 9. Unilateral flux of substrate A and concentration gradient in a matrix containing a biocatalyst

The inner region of a thin matrix is more accessible to substrate than that of a thick matrix with its longer diffusion path. Thus, the effectiveness of the biocatalyst is generally greatest when the matrix thickness (or particle size) is smallest. The effective overall reaction rate per unit volume increases with decreasing matrix thickness. The overall influence of diffusion is described by the effectiveness factor η:

$$\eta = \frac{\text{Overall reaction rate}}{\text{Reaction rate achieved at bulk liquid conditions}}$$

Consider the matrix again but this time with substrate diffusing in from both sides (Fig. 10). The concentration profiles are the result of competition between reaction and diffusion. The ratio of maximum diffusion rate to maximum reaction rate gives a dimensionless parameter P_{D-R}

$$P_{D-R} = \frac{jA}{r_{max}AL} = \frac{D\,(C_{A0}/L)\,A}{r\,(C_{A0})\,AL}$$

where D is the same as D_{A-M} and L is the matrix thickness.

For a first-order reaction, $r = kC_{A0}$;
therefore,

$$P_{D-R} = \frac{D}{kL^2}$$

For a zero-order reaction, $r = k$ and

$$P_{D-R} = \frac{DC_{A0}}{kL^2}$$

Therefore, L plays an important role. The higher P_{D-R}, the higher is the diffusion reaction ratio and the smaller the substrate gradient.

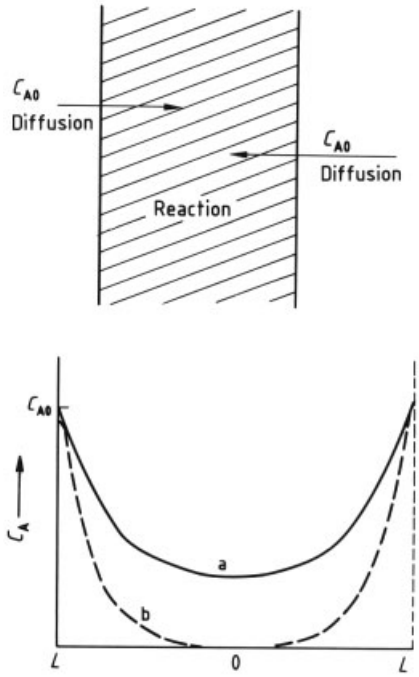

Figure 10. Diffusion of substrate A from both sides of a matrix
a) Complete penetration; b) Incomplete penetration

Derivation of a Finite Difference Model for Diffusion – Reaction Phenomena. Consider the case shown in Figure 11A in which substrate concentration varies from S_0 in the bulk liquid to a zero gradient at position L (a wall or the center of a symmetrical particle). The matrix is divided into a suitably large number of sections such that the change in concentration between sections is not too large. Diffusion fluxes between the sections are portrayed in Figure 11B.

A component mass balance is written for each segment and for each component:

$$\begin{bmatrix}\text{Accumulation}\\ \text{rate}\end{bmatrix} = \begin{bmatrix}\text{Diffusion}\\ \text{rate in}\end{bmatrix} - \begin{bmatrix}\text{Diffusion}\\ \text{rate out}\end{bmatrix} + \begin{bmatrix}\text{Production rate}\\ \text{by reaction}\end{bmatrix}$$

$$A\Delta Z\frac{dS_n}{dt} = j_{n-1}A - j_nA + r_SA\Delta Z$$

where r_S is the production rate of S by reaction per unit volume of matrix.

By using Fick's law in the form

$$j_n = \frac{(S_{n-1} - S_n)}{\Delta Z} D_S$$

the differential difference equation form is obtained:

$$\frac{dS_n}{dt} = D_S \frac{(S_{n-1} - 2S_n + S_{n+1})}{\Delta Z^2} + r_{S_n}$$

Thus, N equations are obtained, one for each segment of the catalyst matrix. This is the finite difference form, which is especially suitable for simulation programming. The equivalent partial differential equation is

$$\frac{\partial S}{\partial t} = D_S \frac{\partial^2 S}{\partial Z^2} + r_S$$

The boundary conditions must be described separately. For the above case, $dS/dZ = 0$ at $Z = L$, and $S = S_0$ at $Z = 0$. The equations for the first and last elements must be written accordingly.

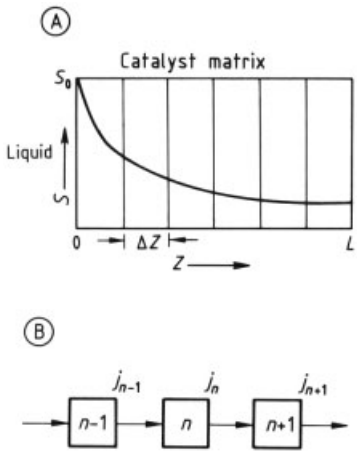

Figure 11. Diffusion of a substrate from a bulk liquid in a catalyst matrix
A) Substrate concentration; B) Diffusion fluxes between sections shown in A

Dimensionless Model. The governing dimensionless model is obtained by defining the variables to range between 0 and 1. Thus, $S' = S/S_0$, $Z' = Z/L$, and for time, $t' > = t/\left(L^2/D\right)$.

Substituting into the differential equation for the first-order case $r_S = -kS$ gives

$$S_0 \frac{dS'}{(L^2/D_S)\, dt'} = D_S S_0 \frac{d^2 S'}{L^2 dZ'^2} - k_1 S_0 S'$$

or

$$\frac{\partial S'}{\partial t'} = \frac{\partial^2 S'}{\partial Z'^2} - k_1 \left(L^2/D_S\right) S'$$

Thus, the solution depends only on $k_1 L^2/D_S$, a dimensionless reaction – diffusion parameter for a first-order reaction (P_{D-R1}), whose square is known in the literature as the Thiele modulus.

For a zero-order reaction, the equation becomes

$$\frac{\partial S'}{\partial t'} = \frac{\partial^2 S'}{\partial Z'^2} - k_0 \left(L^2/D_S S_0\right)$$

where $k_0 L^2/D_S S_0$ is the governing parameter for a zero-order reaction (P_{D-R0}).

Complete and Incomplete Substrate Penetration. Complete penetration of the substrate results in a zero concentration gradient either because of wall conditions at a solid support or because of the symmetry of diffusion at the center. Incomplete penetration results in the substrate profile shown in Figure 12. The shape of the substrate profile is obtained by integrating the diffusion – reaction equations.

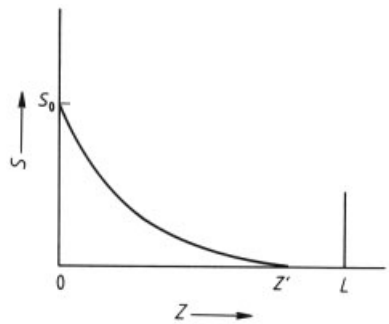

Figure 12. Substrate profile after incomplete penetration of the matrix

Integration for a single substrate with incomplete penetration due to a zero-order reaction gives

$$r_{app} = -\frac{A_M}{V_M} \sqrt{2 D_S k S_0}$$

where A_M and V_M are the area and volume of the matrix, respectively. Thus, when diffusional effects lead to incomplete penetration, the zero-order reaction has an apparent order of 1/2, a value midway between the first-order dependency of diffusion and the zero order of the reaction.

Effectiveness Factor. Analytical solutions of the general diffusion – reaction equation for simple reaction orders are given in [161], [166]. In Figure 13 the values of η for zero, first, and second reaction order are plotted against the square root of the *Thiele modulus* φ where

$$\varphi = \sqrt{L^2 k S_0^{n-1}/D}$$

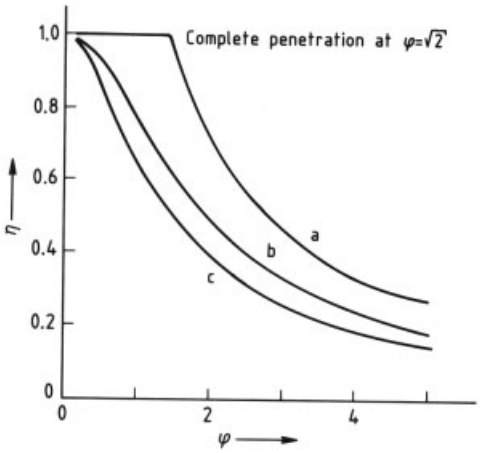

Figure 13. Dependence of effectiveness factor η on Thiele modulus for reactions in a flat film with diffusion from one side (adapted from [161])
a) Zero order; b) First order; c) Second order

As seen from the foregoing, the diffusion – reaction phenomena for single reactions with simple kinetics are governed by a single parameter. For a reaction of order n, the parameter is $(L^2 k S_0{}^{n-1}/D)$ and is the ratio of the reaction rate to the diffusion rate. The magnitude of the parameter can be used to predict the influence of diffusion on overall kinetics. A value much less than 1 means that diffusion has no influence and the concentration profile is flat. A value much above 1 indicates that diffusion is important and that the concentration gradients in the matrix are large. For a zero-order reaction where the parameter is L^2k/DS_0, the overall rate would be influenced only if the substrate were exhausted within the porous support. A first-order reaction (parameter L^2k/D) is influenced by the shape of the profile. This difference is reflected in the effectiveness factor plot of Figure 13.

For zero-, first-, and second-order, and Michaelis – Menten kinetics, η is always less than 1, but for substrate inhibition, the decreased concentration inside the matrix can cause η to be greater than 1. For zero-order kinetics, the limited penetration of substrate causes η to become less than unity.

Experimental Strategies for Overcoming Internal Diffusion. In practice, η must be increased to 1 if diffusion-free enzyme or cell kinetics are to be measured. This can be achieved by

1) decreasing the particle size d_p,
2) decreasing the enzyme or cell loading k,
3) increasing the diffusivity by immobilizing on a support with a more open pore structure (impossible in gels or biofilms), and
4) decreasing the temperature to decrease k.

Experimental strategy is given in Table 6, which summarizes the dependency of the observed kinetics on the experimental variables for a reaction influenced by internal diffusion.

Table 6. Effect of internal mass transfer on the apparent kinetics of a reaction influenced by internal diffusion

Experimental variable change	Observation
Decrease in particle diameter d_p	r_app increases; $K_\mathrm{M,app}$ decreases; apparent enzyme concentration increases
Increase in initial substrate concentration S_0	in low S_0 range, r_app increases with S_0; in high S_0 range, r_app increases with $S_0^{1/2}$ and reaches a maximum, $K_\mathrm{M,app} > K_\mathrm{M}$
Increase in flow rate or mixing rate	no effect on r_app above the point where external mass transfer is eliminated
Decrease in enzyme loading	$K_\mathrm{M,app}$ decreases; value of d_p at which r_app is maximum increases
Decrease in temperature	apparent enzyme concentration increases

Modeling Complex Diffusion – Reaction Systems. The diffusion – reaction phenomena for complex, multireaction systems within a biocatalyst matrix can be adequately described by the quasi-homogeneous diffusion – reaction model presented previously. A batch or continuous biofilm nitrification reactor can be simulated by combining the above model for the solid phase with the appropriate one for the batch or continuous liquid phases as shown in Figure 14.

Thus mass balances can be written for each component in the bulk liquid and coupled to the solid matrix by a mass transfer flux, j_s

$$j_\mathrm{s} = -D_\mathrm{S-M}\,(S_1 - S_0)\,/\Delta Z$$

where D_{S-M} is the diffusivity of substrate in the matrix and S_0 is the concentration in the liquid.

The final model for numerical solution then consists of N equations for each component in the solid matrix plus equations for each component in the liquid phase. Solution can easily be made using a simulation language on a digital computer [166].

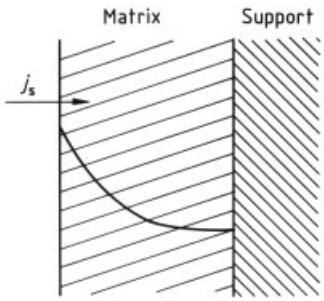

Figure 14. Coupling of bulk liquid by the diffusion flux to the matrix

5. Design of Immobilized Biocatalyst Reactors

In immobilized biocatalyst reactors, the biocatalysts may be immobilized in or on a carrier, immobilized by linkage among one another to form larger particles, or confined within membrane barriers [168–170].

Various reactor configurations are shown in Figure 15. Their advantages and disadvantages are summarized in Table 7. For purposes of modeling, the tank and tubular models, together with the external and internal transfer concepts, and the intrinsic kinetics, provide adequate flexibility to describe all reactor types. The choice of a particular reactor configuration depends on many factors including

1) batch or continuous operation required;
2) immobilized biocatalyst characteristics (particle size, catalyst density, reaction rates, stability, etc.);
3) mass-transfer requirements (nutrients, oxygen etc.);
4) downstream processing requirements (product stability, undesirable byproducts);
5) shear imposed on the biocatalyst (plant and animal cell systems);
6) energy requirements;
7) control of environmental conditions (temperature, pH, substrates, inhibitory and toxic products);
8) substrate availability and waste treatment; and
9) aseptic operation.

Most of the reactors can be run in a batch, fed-batch, or continuous mode (fed-batch denotes semicontinuous operation with unequal feed and effluent streams). For batch operation, packed and fluidized-bed systems, and membrane systems, must be run in a recycle mode.

To describe reactor systems quantitatively, mass balances for total mass and for individual components, as well as energy balances, must be set up. The general form of a component balance for component i is

$$\begin{bmatrix}\text{Rate of accumulation}\\ \text{of component } i \text{ within}\\ \text{boundary region}\end{bmatrix} = \begin{bmatrix}\text{Inflow}\\ \text{rate of com-}\\ \text{ponent } i\end{bmatrix} - \begin{bmatrix}\text{Outflow}\\ \text{rate of com-}\\ \text{ponent } i\end{bmatrix} + \begin{bmatrix}\text{Production rate of}\\ \text{component i within}\\ \text{boundary region}\end{bmatrix}$$

For a well-mixed tank, this gives

$$\frac{\mathrm{d}\,(VC_i)_1}{\mathrm{d}t} = (FC_i)_0 - (FC_i)_1 + (r_i V)_1 \tag{13}$$

where V is the volume in cubic meters, C_i is the concentration of component i in moles per cubic meter, F is the flow rate in cubic meters per second, and r_i is the reaction rate in moles per cubic meter per second. Subscript 0 characterizes the stream entering the reactor; subscript 1, the stream leaving the reactor and in the reaction because the liquid in the reactor is assumed to be well mixed. Equation (13) may be used for a batch, fed-batch, or continuous stirred-tank reactor.

5.1. Batch Stirred-Tank Reactor

For a constant-volume batch stirred-tank reactor (Fig. 15A), the general balance given in Equation (13) becomes

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = +r_i \tag{14}$$

The parameter r_i is defined as the rate per unit volume; for substrates $r_i < 0$ and for products $r_i > 0$. In catalyst systems, the reaction rate

Table 7. Advantages and disadvantages of batch and continuous reactors

Mode of operation	Reactor type	Advantages	Disadvantages
Batch	stirred tank	simple equipment for small-scale production; aseptic operation easier	time-consuming downtime for loading and cleaning; reaction conditions vary with time; difficult to control
	tubular reactors (e.g., packed bed)	recycle mode for kinetic studies	not used for production in this mode
Continuous		constant reaction conditions; better product quality	requires flow-regulating equipment (pumps, valves)
	stirred tank	good external mass transfer; easy control of environmental conditions; easy recharge or withdrawal of biocatalyst	usually low conversion; shear problems
	tubular reactors	high conversion	external mass transfer may be insufficient; environmental control difficult
	packed bed	simple construction; high biocatalyst density	high pressure drop; channeling; mechanical strength of biocatalyst required
	fluidized bed	good external mass transfer; easy recharge or removal of catalyst	high flow rates; low biocatalyst density
	membrane reactor	separation of catalyst from substrates and products	mass-transfer limitations; membrane fouling; pressure drop often high

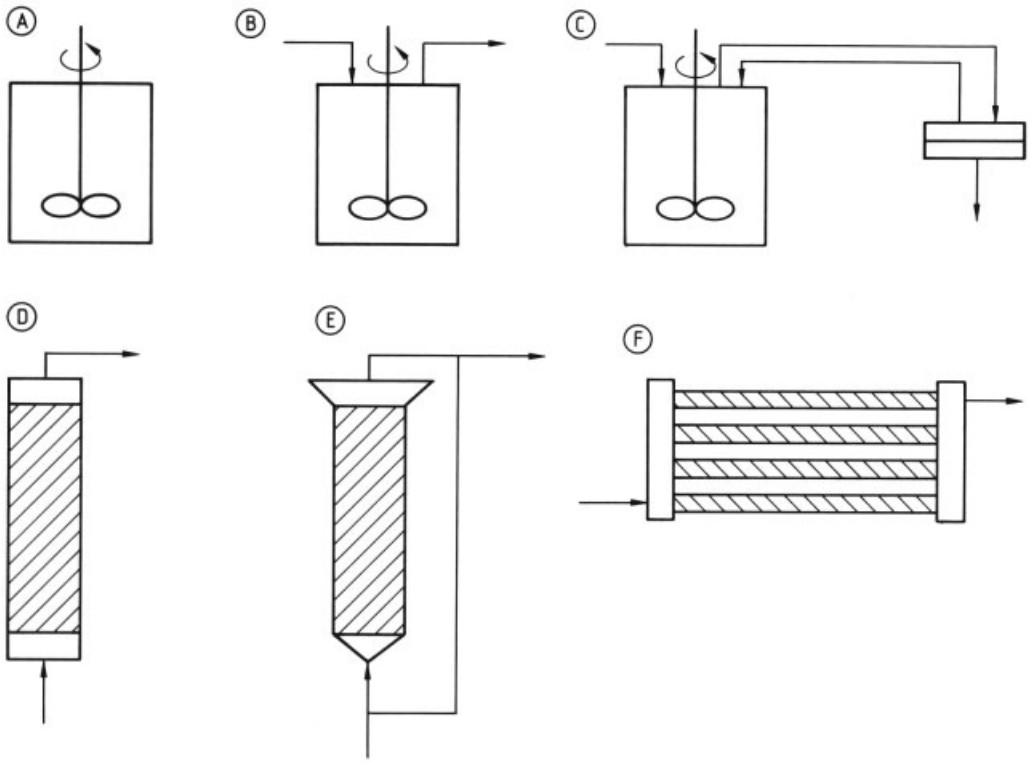

Figure 15. Immobilized biocatalyst reactor types
A) Batch stirred-tank reactor; B) Continuous stirred-tank reactor; C) Continuous stirred-tank membrane reactor with recycle; D) Packed-bed column reactor; E) Fluidized-bed reactor with recycle; F) Tubular membrane reactor
The hatched area indicates the immobilized biocatalyst.

($r_{\mathrm{m},i}$, mol kg^{-1} s^{-1}) is often related to the catalyst mass m_{c}:

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = +\frac{r_{\mathrm{m},i} m_{\mathrm{c}}}{V} \tag{15}$$

The residence time τ necessary for a certain conversion can be obtained by integration of Equation (14)

$$\tau = \int_{Ci0}^{Ci} \frac{\mathrm{d}C_i}{r_i} \tag{16}$$

A plot of $1/r_i$ versus concentration gives the reactor residence time necessary for the required conversion. Three typical cases (Michaelis–Menten kinetics, substrate inhibition, and autocatalytic reaction) are shown in Figure 16. The hatched area represents the necessary residence time.

A number of applications including the production of high-fructose corn syrup, lactase-free milk, and 6-aminopenicillanic acid are listed in [171].

5.2. Plug-Flow Tubular Reactor

Under steady-state conditions, an idealized plug-flow tubular reactor with no axial mixing but complete radial mixing shows the same fluid effluent residence time behavior as a batch tank reactor. Therefore, Equation (14) and Figure 16 are also valid for a steady-state tubular reactor.

Immobilized biocatalysts are usually employed in packed-bed, fluidized-bed, or membrane reactors. In some analytical applications, the biocatalyst is immobilized on the wall of the tube. These reactor types are discussed in Section 7.1.

5.3. Continuous Stirred-Tank Reactor

In an ideal continuous stirred-tank reactor (CSTR, Fig. 15B), each entering fluid element is mixed instantaneously with the entire reactor volume. The reactor operates at reactor outlet conditions. For steady-state conditions and constant volume, Equation (14) gives

$$\tau = -\frac{C_{i0} - C_{i1}}{r_{i1}} \tag{17}$$

The necessary residence time τ can be determined as shown in Figure 17 (hatched area).

Comparison of these results with those of Figure 16 shows that a CSTR is favorable for substrate inhibition (B) and autocatalytic growth kinetics (C) at low and medium conversion (large values of C_i). A tank reactor can be run in a fed-batch mode to keep the concentration of inhibitory substrate at a desirable level and, therefore, obtain a performance similar to that of a CSTR. Batch and plug-flow reactors are favorable with reaction kinetics having a reaction order $n > 0$ and with product inhibition kinetics. In the case of zero-order kinetics, both reactor types show the same performance.

Mass transfer in CSTRs may be very high and control is simple. They are, therefore, used when high mass transfer is required (e.g., aeration) and control of environmental conditions (e.g., pH) is critical.

5.4. Packed- or Fixed-Bed Reactors

Table 8. Literature on fixed-bed biocatalyst applications

Application	Biocatalyst	Reference
Production of L-amino acids	adsorbed aminoacylase	[178]
Aspartate production	aspartase in immobilized cells	[178]
Diverse applications, laboratory and large-scale	enzymes and cells immobilized in gels	[172]
Anaerobic wastewater treatment	biofilm on porous ceramic	[179]
Drinking water denitrification	whole cells in alginate	[180]
Ethanol production	cells on wood chips	[181]
High-fructose corn syrup, packed-bed enzyme reactor and moving-bed adsorption	enzyme immobilized and product adsorbed on zeolite	[192]
Inulin hydrolysis	cells immobilized by adsorption on wood shavings	[183]
Commercial ethanol production	yeast cells in alginate gel	[174]
Continuous production of sucrose	enzyme on corn grits	[184]
Acid whey hydrolysis	enzyme on ion-exchange resin	[185]
Ethanol production, modeling	immobilized yeast	[186]
Hydrolysis of starch, modeling	immobilized enzyme	[187]
Mammalian cell culture	various cell types	[188], [189]

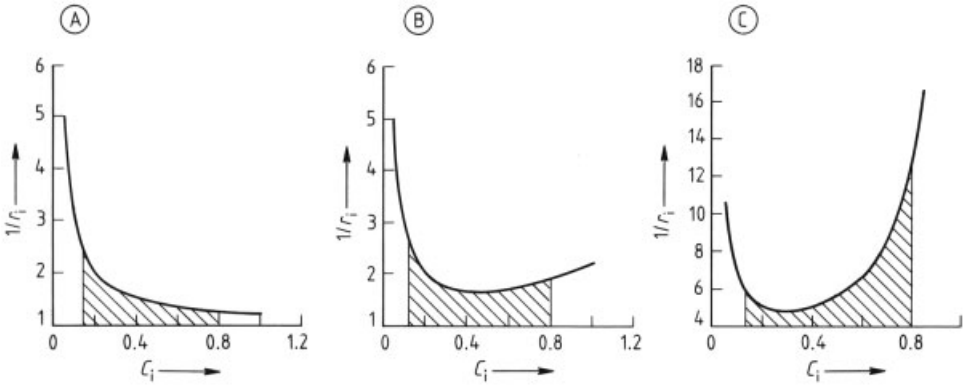

Figure 16. Graphical determination of residence time (hatched area) for given conversion and different kinetic characteristics in a batch reactor
A) Michaelis – Menten kinetics; B) Substrate inhibition kinetics; C) Autocatalytic reaction (e.g., growth)

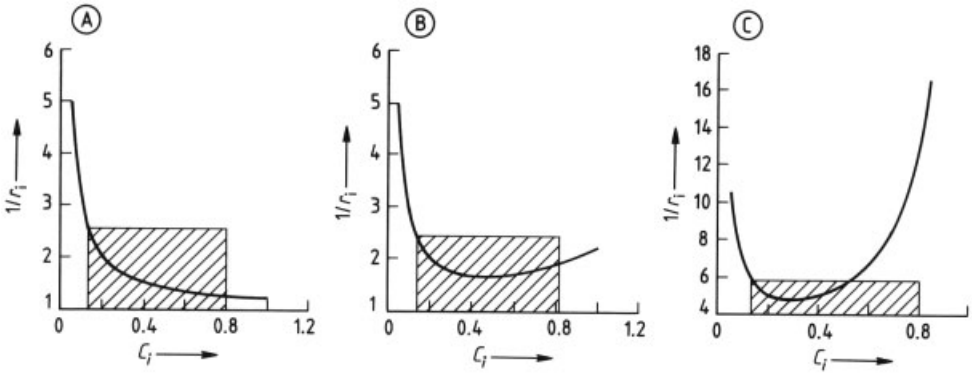

Figure 17. Graphical determination of residence time (hatched area) for given conversion and different kinetic characteristics in a continuous stirred-tank reactor
A) Michaelis – Menten kinetics; B) Substrate inhibition kinetics; C) Autocatalytic reaction (e.g., growth)

Most continuous laboratory and large-scale production units are of the packed- or fixed-bed type (Fig. 15D) [172–174]; they are usually tubular plug-flow reactors. In addition to axial gradients, internal and external diffusion processes may cause gradients at or within the biocatalyst. In large-scale reactors, the mechanical strength of biocatalyst particles, pressure drop, and control are important. Deviations from ideal plug flow may be observed, especially if channeling occurs in the bed. Applications of fixed-bed reactors are listed in Table 8.

5.5. Fluidized-Bed Reactors

Fluidized-bed reactors (Fig. 15E) have found increased application because they provide high external mass transfer to small reactive particles consisting of immobilized cells, enzymes, or biofilms. This is important because growing organisms would otherwise block fixed- or packed-bed columns. Fluidized-bed reactors are of the tubular type and models are, therefore, similar to those for an ideal plug-flow reactor. Figure 18 shows a large-scale anaerobic fluidized-bed reactor for wastewater treatment [175]. A minimum flow rate is required because fluidization occurs at a minimum superficial liquid velocity which depends on the density difference and the particle size. To ensure a minimum flow rate, a certain amount of recycle is required. The recycle ratio decreases with increasing reactor size. Small laboratory-scale fluidized-bed biocatalyst reactors are, therefore, often run as

differential recycle reactors with low conversion per passage.

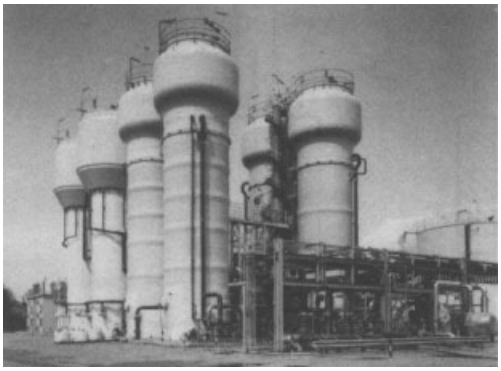

Figure 18. Large-scale anaerobic fluidized-bed reactor for wastewater treatment
Courtesy of Gist-brocades, Delft, Netherlands.

Literature on the application of fluidized-bed biocatalyst reactors is listed in Table 9.

Table 9. Literature on fluidized-bed biocatalyst applications

Application	Biocatalyst	Reference
Review on the application of fluidization	various types	[190]
Biological application	biofilms on various supports	[191]
Ethanol production	cells in alginate	[192]
Selection of particles	flocculant biomass on support particles	[193]
2-Propanol – butanol production	*Clostridium* spp. on calcium alginate	[194]
Anaerobic wastewater treatment	biofilms	[175], [195], [196]
Design of fluidized-bed reactors	various types	[197]
Wastewater treatment	biofilms	[198]
Animal cell culture	special particles	[199]
Plant cell culture	spontaneously growing pellets	[210]
Three-phase fluidized beds	various types	[201–203]

5.6. Membrane Reactor Systems

(→ Membranes and Membrane Separations)

Membrane systems provide a means of retaining enzymes and cells within a biological reactor. Progress in membrane technology has produced membrane reactor systems that can operate in a variety of modes [176]. Ultrafiltration or microfiltration membranes may be used in a recycle stirred-tank reactor (Fig. 15C) to retain enzymes, cells, or cofactors. A tubular membrane reactor is shown in Figure 15F.

Membrane reactors may be classified according to several criteria:

1) Membrane type (ultrafiltration, microfiltration, dialysis, liquid membrane, pervaporation)
2) Biocatalyst type (enzyme, whole cells)
3) Number of liquid phases
4) Number of different membranes
5) Permeation (product or substrate)
6) Location of biocatalyst (in liquid phase or membrane)
7) Membrane geometry (flat, tubular, spiral, etc.)
8) Driving force for membrane transport (pressure or concentration gradient)
9) Flow of individual liquid phases (batch or continuous)
10) Sterility

Applications of membrane biocatalyst are listed in Table 10.

Table 10. Literature on membrane bioreactor applications

Application	Biocatalyst	Reference
Mammalian cell culture	hollow fiber reactor	[204–209]
Large-scale animal cell culture	various	[195]
Hybridoma culture	membrane gas supply and perfusion in animal cell culture	[210]
Microbial cultures	hollow fiber reactor	[211], [212]
Membranes and membrane processes	various	[213]
Scaleup, engineering aspects	various	[214], [215]

5.7. Laboratory Reactors for Kinetic Measurements

5.7.1. Batch and Continuous Tank Reactors

Both batch and continuous tank reactors lend themselves to the determination of kinetic parameters on a laboratory scale. According to Equation (14), the reaction rate in a batch reactor is calculated from the slope of the concentration – time curve. The CSTR, on the other hand,

requires measuring the steady-state difference between outlet and inlet concentrations and the mean residence time (Eq. 17). The CSTR usually yields better results, except when reaction rates are very low.

In a batch experiment, substrate and product concentrations vary with time and may be difficult to monitor and interpret. Suitable design of experiments (e.g., variation of initial concentrations, supply of an excess of one component) is, therefore, required to establish reaction rate kinetics and determine rate constants.

At steady state the CSTR can be used to measure rates. New steady states can be established by changing the flow rate, inlet concentration,etc.

Evaluation of experimental data may be performed in serveral ways:

1) Differential equations can be integrated algebraically and rearranged to give a linear function for kinetic parameters. In this way, model applicability and model parameters may be estimated in one plot (e.g., Lineweaver – Burk plot for Michaelis – Menten kinetics [169].
2) If accumulation terms are equal to reaction rates, a plot of terms versus concentration may result in useful concentration – rate relationships and, after rearrangement, yield reaction rate parameter values.
3) A third method uses numerical integration and linear or nonlinear parameter estimation methods to check model equations and to estimate model parameter values. A user-friendly, numerically powerful tool is available (SIMUSOLV [177]). Modern program packages including excellent simulation methods may help greatly in optimal design of kinetic experiments.

A short summary of experimental strategies to determine rate parameters for complex biochemical kinetics is given in [179].

5.7.2. Experimental Differential Recycle Reactor

Reactions occurring in and on particulate solids (e.g., in beads of immobilized enzymes or whole cells) are generally sensitive to external mass transfer (Section 4.1). Study of the influence of flow rate on the overall reaction rate is, therefore, of interest; a convenient reactor configuration is shown in Figure 19. The system has a reactor that holds the biocatalyst particles, either as a packed or fluidized bed. Variable flow rates through the reactor are achieved by recycle. The reactor is usually operated at low residence time (high flow, small volume) so that the conversion per pass is small. The axial gradients are then sufficiently low for concentrations in the reactor to be approximated by the average between inlet and outlet. Thus, steady-state balances take the form

$$0 = F_{\mathrm{R}}\,(C_{i0} - C_{i1}) \pm r_{i,\mathrm{avg}} V_{\mathrm{r}}$$

where F_{R} is the flow rate of the recycle stream, $r_{i,\,\mathrm{avg}}$ is the reaction rate at the mean concentration $C_{i,\,\mathrm{avg}}$, and V_{r} is the volume of the reactor. The difference $C_{i0} - C_{i1}$ must be large enough for accurate measurement but small enough for average concentration to be a valid approximation.

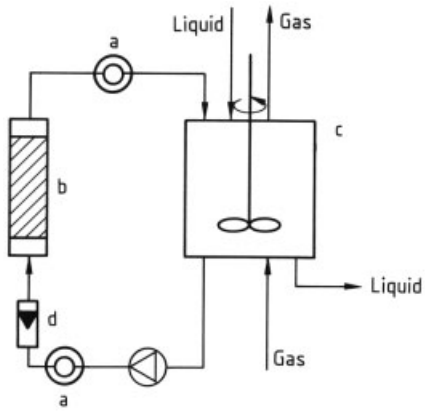

Figure 19. Differential recycle loop reactor for kinetic studies
a) Sampling sensor; b) Biocatalyst bed; c) Sensor and control (pH, temperature, etc.); d) Flow meter

The recycle system shown in Figure 18 can be operated batchwise, with no flow entering or leaving the recycle loop and a small amount of conversion occurring per pass. The balances become

$$V_{\mathrm{T}} \frac{\mathrm{d}C_i}{\mathrm{d}t} = \pm\, r_i V_{\mathrm{r}}$$

where V_{T} is the volume of the total system. Corresponding balances for a continuously fed recycle system can be set up [169].

6. Immobilization Media

Biocatalysts must be immobilized in a different phase or physically separated from the phase that leaves the reaction vessel and contains the products.

6.1. Characteristics of Solid Supports

A solid-state carrier or support is generally used for biocatalyst immobilization in a different phase. Because the catalytic activity of a biocatalyst reactor is proportional to the active enzyme (or whole cell) concentration, carriers with high surface area are commonly employed. Porous solids provide areas ranging from one to several hundred square meters per gram. However, a high-area support is characterized by small-diameter pores which cause diffusional limitations. Very small pores prohibit the entrance of large protein molecules, rendering at least part of the support area inactive. This phenomenon can be used for selective immobilization of enzymes of different molecular mass [216]. The support should not always be assumed to be entirely inert because it may modify enzymatic properties (e.g., conformation) and may itself exhibit partition effects (adsorption, pH shift, hydrophobicity, etc.). Adsorption of inhibitors from the reaction mixture may also be detrimental to biocatalyst activity.

Many other carrier properties, although equally important for the function of an enzyme reactor, are beyond the scope of this chapter. For nonporous particles, these include size, size distribution, shape, abrasion in stirred vessels or fluidized beds, apparent viscosity when suspended in liquid, density, voidage, swelling characteristics, and osmotic characteristics. Important properties of porous media are porosity, permeability, pore size distribution, accessibility of inner surface for the binding of enzymes and cells, wetting characteristics, and tortuosity. Most of these properties and their influence on enzyme reactor performance are treated in [217], [218]. Physical characterization of entrapment networks for immobilization of whole cells is discussed in [219].

6.2. Characteristics of Semipermeable Separators

Membranes with suitable pore size can be used as physical barriers to retain a biocatalyst in a reactor and allow free passage of substrate and product. The biocatalyst can be used in either a soluble or an immobilized form. Possible configurations of biocatalyst and membrane with respect to flow regime are given in Figure 20. A large number of combinations and operational modes are conceivable based on

1) membrane type (ultrafiltration, dialysis, liquid membrane),
2) biocatalyst type,
3) number of liquid phases,
4) number of different membranes,
5) permeation of product or substrate,
6) location of biocatalyst (in liquid phase, in membrane),
7) membrane geometry (flat, tubular, spiral, etc.), and
8) driving force for membrane transport.

7. Application of Immobilized Biocatalysts

7.1. Analysis

Most of the applications of immobilized biocatalysts are analytical. The specific molecular recognition and molecular transformation of biocatalysts are combined with a suitable transducer to allow a given property (e.g., optical) to be monitored. Specific analysis in complex mixtures and detection of reaction products are thus possible. Bioanalytical devices with intimate contact of biological material and transducer are called *biosensors* (→ Chemical and Biochemical Sensors).

A number of analytical systems have been designed:

1) *Flow systems:* biocatalyst column with subsequent detection and segmented flow or flow injection systems
2) *Dipsticks and multilayer films:* biocatalyst within a thin layer with colorimetric or photometric detection

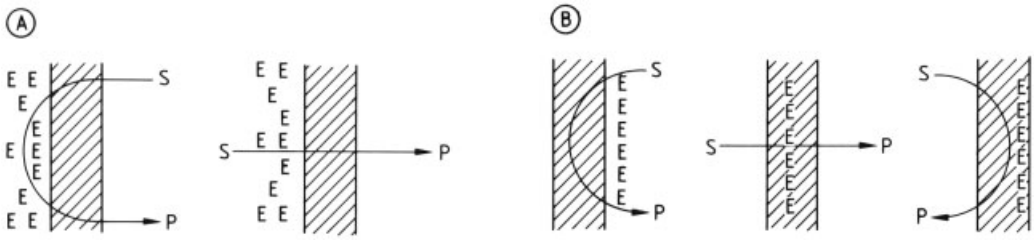

Figure 20. Location of a soluble or suspended biocatalyst (A) and an immobilized biocatalyst (B) and possible transport fluxes in a membrane reactor
Abbreviations: E = enzyme (biocatalyst); P = product; S = substrate

3) *Electrodes:* amperometric or potentiometric detection
4) *Optodes:* optical detection by using light absorption, fluorometry, or luminometry
5) *Thermistor probes:* calorimetric measurement of heat of reaction
6) *Transistors:* similar to electrodes, but semiconductor devices

An extensive literature exists on analysis via immobilized biocatalysts (see Table 11).

Table 11. Reviews on analytical application of immobilized biocatalysts

Title	Comments	Reference
Biosensors: a survey of what is done and used commercially	commercial sensors in medicine, defense, environmental analysis	[220]
Immobilized enzymes in analysis	optical thin-film enzyme assays and immunoassays	[221]
Electrochemical sensors in the analysis and control of bioprocesses	developments in on-line monitoring by biocatalyst electrodes	[222]
Whole-cell biosensors	whole-cell electrodes	[223]
Immunosensors: antibody-based biosensors	immunosystems and transducers	[224]
Analytical uses of immobilized enzymes	monograph	[225]
Biosensors		[226]
Applications of immobilized biocatalysts in chemical analysis	various systems, diffusion effects	[227]
Immobilized whole-cell biocatalysts	monitoring of various chemicals	[228]
Biosensors	international workshop	[229]
Immobilized enzymes and cells	section on analytical applications	[230]
Chemical sensors	book on chemical and biological sensors	[231]

Enzyme Sensors for Clinical Analysis. Electrodes are available for the measurement of a number of important parameters [e.g., hydrogen ions (pH), oxygen, hydrogen peroxide, and ammonia]. Because such compounds and ions are often involved in enzymatic reactions as reactants or products, these electrodes can be employed in conjunction with a suitable enzyme to monitor substrate concentration. Important reviews on enzymes and microbial sensors are given in [232–237].

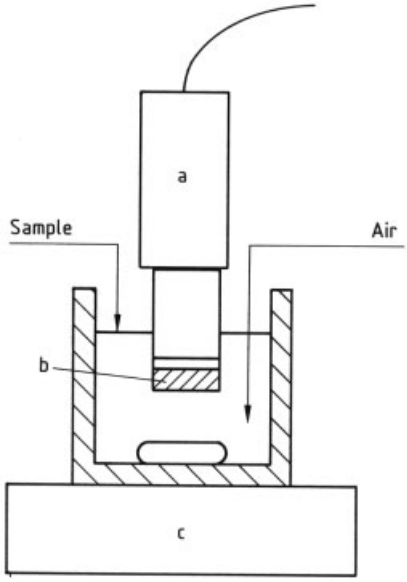

Figure 21. Enzyme electrode used in clinical analysis
a) Electrode; b) Enzyme membrane; c) Stirrer

A typical setup for this procedure is shown in Figure 21. It can be used, for example, to measure glucose concentration. A membrane-covered oxygen electrode is employed together with a membrane containing immobilized glucose oxidase [*9001-37-0*] (E.C. 1.1.1.49). Glucose and dissolved oxygen react to form gluconic acid and hydrogen peroxide. The oxygen

electrode detects oxygen consumption which is proportional to reaction rate and thus to glucose concentration. At low glucose levels, the resulting electrode signal is linearly related to this concentration. Alternatively, a hydrogen peroxide electrode can be used.

Enzyme electrodes have been constructed for at least 40 different substances, but very few have been commercialized (Table 12). The most important factor is enzyme stability, which is critical for the lifetime of the enzyme electrode, and depends largely on the immobilization procedure. Stabilities of three to four months have often been attained.

Table 12. Examples of enzyme electrodes

Analysis	Enzyme system	Electrode
Alcohols	alcohol oxidase	H_2O_2
Amino acids	L-amino acid oxidase	H_2O_2
Cholesterol	cholesterol oxidase	O_2, H_2O_2
Creatinine	creatinase	NH_4^+
Glucose	glucose oxidase on poly(vinyl chloride) membrane	O_2, H_2O_2
Sucrose	invertase, mutarotase, and glucose oxidase on collagen	O_2
Urea	urease on Teflon	NH_4^+

Microorganisms can also be immobilized on the outside of the electrode. Thus, for example, phenol can be detected by using an oxygen electrode together with phenol-oxidizing microorganisms. Stabilities of 5 – 30 d have been achieved for over 20 compounds by using immobilized organisms [237].

For certain analyses, the immobilized enzyme is better employed in a reactor. The sample is passed through the reactor, and the concentration of the desired substance in the reactor effluent is then measured downstream by an appropriate electrode sensor. This arrangement allows adjustment of the enzyme quantity.

7.2. Immobilized Enzymes in Therapeutic Medicine

The ability of enzymes to carry out specific biological reactions is being exploited as a therapeutic method in medicine (→ Enzymes, Chap. 5.4.). A number of applications have been tested on animals, but very few on humans. The enzymes require immobilization and contact with body fluids and organs, either inside or outside the body (intra- or extracorporeal). Review articles in this rapidly developing field are given in [238–240].

Intracorporeal methods require immobilization in or on microparticles or blood cell fragments, or with macromolecules [238]. Distinction is made between enzymes circulated with the blood and those used for local treatment of specific organs. Soluble polymers and microparticles can be used for immobilization in the blood stream. Slow release by diffusion or biodegradation from particles is employed for local use.

Enzymes introduced into the body may have the advantage of targeting specific sites but increase the danger of toxic or immune response. The carrier and the immobilized enzyme must exhibit a minimum of allergic and immunogenic effects. Fortunately, immobilization often reduces immunogenicity. A partial list of intracorporeally employed enzymes, their therapeutic use, and the type of carrier employed is given in Table 13. Microencapsulation methods are important for a variety of enzyme drug applications and have the following advantages [238], [240]:

1) ultrathin (20 nm) synthetic membrane capsules;
2) small particle size (100 μm);
3) high transfer rates;
4) biological particles (normal cells, cell ghosts, synthetic liposomes);
5) wide range of immobilization applications (single enzymes, enzyme mixtures, whole cells, absorbents);
6) intra- and extracorporeal use;
7) possibility of biodegradable materials (proteins, lipids);
8) wide range of administration methods (oral, intravenous, local injection); and
9) minimal immune reactions.

Extracorporeal application requires access to the blood stream and anticoagulant treatment [239]. The equipment is usually a small immobilized enzyme reactor connected to the patient (much like a hemodialysis unit) and, therefore, is not suitable for continuous treatment. To obtain the necessary activity, a large surface area of particles (packed and fluidized beds) or membranes (hollow-fiber cartridges) is necessary, which increases the danger of blood clotting. Extracor-

Table 13. Methods of immobilizing intracorporeal enzymes

Carrier	Enzyme	Therapy	Reference
Artificial cells	asparaginase	cancer	[238–240]
or microcapsules	urease	model system	[238], [240]
Erythrocyte ghosts	urokinase	thrombosis	[238]
	asparaginase	cancer	[238–240]
Liposomes	asparaginase	cancer	[238–240]
Soluble polymers	arginase	hyperargininemia	[238], [239]
Poly(ethylene glycol)	uricase	gout	[239]
Albumin, dextran	glutaminase	cancer	[238], [239]

poreal applications, including the reactor type, are given in Table 14.

Table 14. Applications of extracorporeal enzyme therapy [239]

Reactor device	Enzyme	Therapy
Fluidized bed and packed beds of sepharose beads	heparinase	anticoagulant removal
	bilirubin oxidase	liver failure
	urea cycle enzymes	liver failure
Hollow-fiber cartridges	arginase	hyperargininemia
	asparaginase	cancer
Packed bed of agarose beads	UDP – glucuronyl transferase	liver failure

7.3. Ethanol Production Using Immobilized Cells

Catalyzed by the oil crisis, research activity has been particularly intense in the field of ethanol production for fuel. Review articles on ethanol production using fermentation of glucose with immobilized cells are given in [241–245]; see also → Ethanol.

Because raw materials constitute two-thirds of the production costs, yield is particularly important. Another major cost is separation of the alcohol, usually by distillation, from water. Because ethanol is toxic to microorganisms, its final concentration is usually well below 100 g/L.

In conventional suspended-culture batch fermentation, ethanol production rates are 1.8 – 2.5 $g\,L^{-1}\,h^{-1}$; rates in continuous tanks are 6 – 8 $g\,L^{-1}\,h^{-1}$. Suspended cell systems can be improved by cell separation and recycle systems (production rate 30 – 40 $g\,L^{-1}\,h^{-1}$), but this entails considerable capital and operational expense [243]. Immobilized yeast (*Saccharomyces* spp.) and a bacterium (*Hymomonas mobilis*) have been investigated.

An alternative way to obtain high cell concentrations and thus high conversion rates is to immobilize the cells by adhesion or adsorption on porous surfaces and entrapment in gels. Adhesion has been achieved with organic materials (wood chips, cotton fibers, sawdust, anion-exchange resins, and gelatin-coated glass beads) and with porous inorganics (ceramics, glass fibers, silica gel, and vermiculite). Production rates in the range of 20 – 60 $g\,L^{-1}\,h^{-1}$ or higher have been reported. Entrapment in polymer gel matrices is the most frequently investigated method. Natural polymers such as alginate and carrageenan give the best results. Stable operations over 100 d have been reported with packed and fluidized beds, 1 – 3-h residence times, and production rates of 15 – 50 $g\,L^{-1}\,h^{-1}$.

Flocculating cells can also be retained in a vertical column with an upper sedimentation section and slow upflow. Operations over 30 d have given production rates of 14 – 100 $g\,L^{-1}\,h^{-1}$.

Closest to industrial application is the pilot plant of the Kyowa Hakko Kogyo Co. in Japan. With a total reactor volume of 4 m^3 in five fluidized-bed columns, 8.5 % ethanol is produced at a rate of 20 $g\,L^{-1}\,h^{-1}$. Immobilization is with alginate.

7.4. Industrial Applications of Immobilized Biocatalysts

The more important industrial applications of immobilized biocatalysts are summarized in Table 15.

Table 15. Industrial applications of immobilized cells [245]

Product	Cell system	Year introduced on a commercial scale
Acetic acid	*Acetobacter* on wood shavings in a trickle-filter reactor	1880
L-Aspartic acid	*Escherichia coli* in various gels	1973 – 1976
High-fructose corn syrup	various systems	1966 – 1974
L-Malic acid	*Brevibacterium* in carrageenan gel	1974 – 1977
Penicillin	gel-entrapped *Escherichia coli*	1969 – 1978
Raffinose hydrolysis	pellets of *Mortierella* fungus	1968 – 1974
Prednisolone (steroid transformation)	*Curvularia* and *Arthrobacter* coentrapped in polyacrylamide gel	1978

7.4.1. Production of High-Fructose Corn Syrup

High-fructose (42 %) corn syrup (HFCS) from cornstarch is by far the most important product of immobilized biocatalyst technology; the estimate annual worldwide production for 1982 was 6×10^6 t on a dry basis (5.5×10^6 t in the United States by 11 companies and 0.4×10^6 t in Japan by 16 companies). Glucose from cornstarch is converted to fructose with glucose isomerase [*9055-00-9*] (E.C. 5.3.1.18) in the form of a variety of immobilized whole cells. Immobilization methods include gel entrapment with cross-linking, binding on ion-exchange resin, flocculation with electrolytes, extrusion of cells and cross-linking, and heat treatment (→ Enzymes, Chap. 4.2.1., → Enzymes, Chap. 4.3.; → Glucose and Glucose-Containing Syrups, Chap. 6.2.). Developed in Japan in 1966, the rights to the process were acquired by a U.S. company which by 1978 had improved the glucose isomerase enzyme production of *Streptomyces* by mutation by a factor of 200 [245].

The process starts with a relatively inexpensive raw material, corn. Cornstarch is converted to glucose syrup by soluble amyloglucosidase from *Aspergillus niger* in a batch process. The glucose-containing syrup is then purified and treated by a continuous process using immobilized glucose isomerase. The resulting fructose composition is typically 42 %, which is close to the theoretical maximum for the equilibrium (57 %). This syrup has essentially the same sweetness as sucrose sugar.

Mutant *Streptomyces murinus* cells, that are disrupted mechanically, cross-linked with glutaraldehyde, flocculated, filtered, extruded to pellets (0.3 – 1 mm), and dried [246]. Activity decreases to 10 % in 127 d.

One of the largest enzyme suppliers for this process, the Novo Company in Denmark, gives the following information (Table 16) [246]:

Table 16. Operational details of the Novo Sweetzyme reactors [246]

Parameter	Comments
Biocatalyst	mutant *Streptomyces murinus* cells, disrupted mechanically, cross-linked with glutaraldehyde, flocculated, filtered, extruded to pellets (0.3 – 1 mm), and dried
Reactor	Capacity 100 t/d (dry material) fixed bed with downflow (diameter 0.95 m, height 3.1 m, total volume 17.4 m^3)
	two lines of four reactors in series (with recycle) or parallel
	5240 kg biocatalyst in all reactors, 30 kg replaced per day, 16 day replacement
	flow rate of syrup (dry basis) 100 t/d
Feed composition	40 % solids (95 % glucose)
Product composition	42 % fructose, 53 % glucose, 5 % oligosaccharides, 70 – 75 % total solids after evaporation
Control	cycle flow adjustment necessary due to decreasing catalyst activity to keep fructose product concentration constant. Flow varies linearly from 10 to 8 m^3/h over 400 h
Catalyst lifetime	activity decreases to 10 % in 127 days, catalyst replaced alternately in the eight reactors every 16 days
Conditions	reactor outlet pH 7 – 7.5, feed pH 7.5 – 8.2, Mg : Ca ratio 15, 60 °C, oxygen-free
Yield	0.44 t of fructose per tonne of glucose
Productivity	3.3 t HFCS (dry) per kilogram of catalyst
	1.4 t fructose (dry) per kilogram of catalyst
Daily production rates	100 t HFCS (dry)
	43 t fructrose
	16.4 t fructose per kilogram of catalyst
	5.7 t/m^3 HFCS (dry)

The syrup is purified by filtration, concentrated from 30 % to 40 % by evaporation, treated with activated carbon to remove soluble impurities (peptides and amino acids), and by ion exchange to remove calcium ions. Oxygen must finally be removed before the syrup enters the fixed-bed isomerization reactors. Purification is critical for obtaining high product quality and maximum enzyme stability.

The feed stream is brought to pH 7.5 – 8 with sodium carbonate. Magnesium, which activates and stabilizes the enzyme, is continuously added in the form of magnesium sulfate so that the Mg : Ca ratio is above 15. Before entering the reactors the syrup is passed through a heat exchanger and brought to 60 °C.

The isomerization reactors are columns filled with immobilized enzyme (" Sweetzyme") consisting of microbial cells which have been cross-linked and extruded to small pellets (see Table 16). Since the enzyme has a limited lifetime, the reactor columns must be refilled alternately with fresh biocatalyst when the activity reaches 10 % of the maximum. The continual fall in activity requires that the flow rate be reduced to maintain the desired conversion to fructose. After enzyme replacement the flow can be increased. Temperature variation cannot be used to compensate for decreased enzyme activity because higher temperatures increase activity but lead to poorer stability and product quality. Lower temperatures give longer catalyst life but increase the risk of microbial infection.

Post-isomerization often involves contacting with activated carbon and ion-exchange resin to remove color-forming material. The syrup is then evaporated to 70 – 75 % dry substance and stored at 30 °C.

Separation of the fructose (usually by ion exchange) allows unreacted glucose to be recycled to the reactors and results in syrups containing 55 % fructose (enriched fructose corn syrup, EFCS) and 90 % fructose (very enriched fructose corn syrups, VEFCS). The new higher fructose syrups were widely accepted by the U.S. soft drink industry when introduced in 1976; more than half of the sugar used in soft drinks is HFCS.

7.4.2. Amino Acid Production

Chemically synthesized D, L-amino acids require optical resolution to obtain the biologically active L-form. For many years the Tanabe Seiyaku Co. processed acyl-D,L-amino acid mixtures by stereospecific hydrolysis with soluble L-amino acid acylases according to the reaction

$$\text{Acyl-D,L-amino acid} \xrightarrow{\text{Aminoacylase}} \text{L-Amino acid} + \text{Acyl-D-amino acid}$$

The D-amino acid was racemized by heating to the D,L-form and recycled. Removal of the remaining enzyme required complicated purification (→ Enzymes, Chap. 4.9.1.).

In 1969 the company introduced an immobilized enzyme process based on mold aminoacylase attached by ionic binding to DEAE-Sephadex in a packed-bed column reactor (Fig. 22). The L-amino acid can be separated easily by selective crystallization from the concentrated product stream. The D-amino acid can then be recycled after racemization. The production cost of the immobilized enzyme process is 40 % lower than with the soluble enzyme process. A number of L-amino acids are reported to have been synthesized this way: phenylalanine, methionine, tryptophan, and valine.

Reviews of the production of amino acids by immobilized enzymes and whole cells are given in [245], [247–250]. Industrial processes are listed in Table 17. Other processes, not yet industrialized, are listed in Table 18.

l-Aspartic Acid. In 1973 a Tanabe Seiyaku Co. process went onstream to produce L-aspartic acid [*56-84-8*] from fumaric acid and ammonia by using *Escherichia coli* cells immobilized in polyacrylamide gel; this was the first large-scale industrial process using immobilized microorganisms. The catalyst had a half-life of four months, and the activity was sufficient to produce 1800 kg of L-aspartic acid per day in a 1000-L reactor. In 1979, immobilization in carrageenan and treatment with glutaraldehyde increased the biocatalyst half-life to 680 d and the production to 3400 kg/d. In 1982, a strain of *E. coli* with sevenfold higher aspartase activity, specially treated to eliminate fumarase activity, was introduced into the process.

l-Alanine. L-Alanine [*56-41-7*] is important in pharmaceuticals and as a taste additive in foods. It can be produced by decarboxylation of L-aspartic acid with L-aspartate 4-decarboxylase [*9024-57-1*] (E.C. 4.1.1.12).

This process was operated batchwise with soluble enzyme by the Tanabe Seiyaku Co. In 1982 the company industrialized a continu-

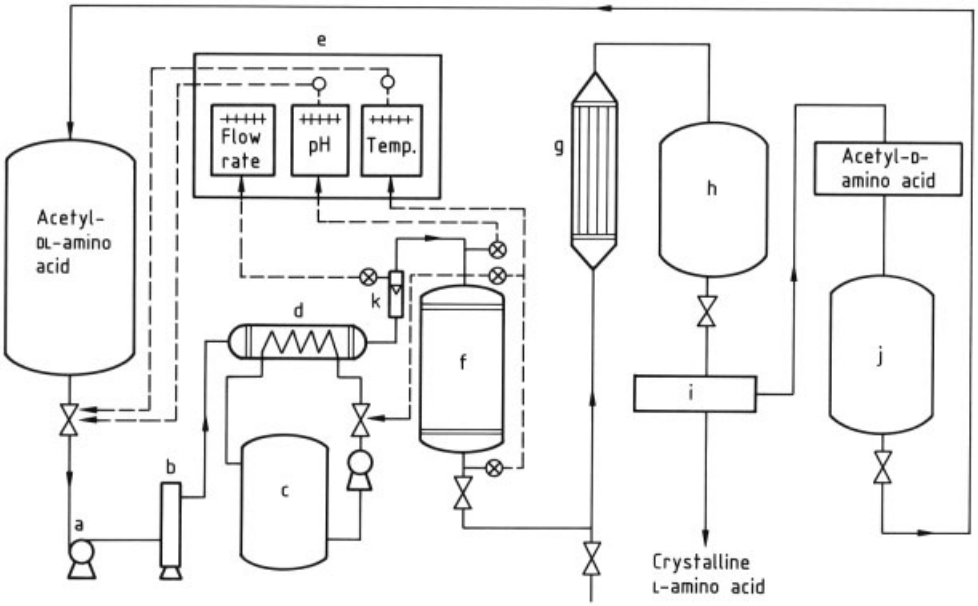

Figure 22. Flow diagram for the continuous production of L-amino acids by immobilized aminoacylase (Tanabe Seiyaku, Japan)
a) Pump; b) Filter; c) Hot-water tank; d) Heat exchanger; e) Control panel; f) Enzyme column; g) Continuous evaporator; h) Crystallizer; i) Separator; j) Tank for racemization; k) Flow meter

Table 17. Industrial processes for amino acids using immobilized biocatalysts

Product	Substrates	Catalyst	Reactor
L-Alanine L-Methionine L-Phenylalanine L-Tryptophan L-Valine	D,L-amino acid racemic mixtures	aminoacylase on DEAESephadex	packed bed (Chibata)
L-Aspartic acid	fumaric acid and ammonia	*Escherichia coli* in agar and carrageenan	packed column
L-Alanine	L-aspartic acid	*E. coli* in carrageenan and *Pseudomonas dacunhae* in carrageenan	two packed columns in series

Table 18. Amino acid syntheses with biocatalysts not yet used commercially

Product	Substrates	Biocatalyst	Carrier
L-Glutamic acid	glucose	*Brevibacterium flavum* *Corynebacterium*	collagen polyacrylamide
L-Lysine	D,L-α-amino-ε-caprolactam	caprolactam hydrolase and caprolactam racemase	Sephadex
L-Tryptophan	indole, pyruvate, and ammonia	trytophanase	Sepharose
L-Tryptophan	indole and serine	*Escherichia coli*	polyacrylamide
L-Tyrosine	phenol, pyruvate, and ammonia	β-tyrosinase	Sepharose

ous process based on carrageenan-immobilized *Pseudomonas dacunhae* in a pressurized packed-bed column. The pressure keeps the carbon dioxide in solution, which results in a smaller pH change and less axial mixing. The L-aspartic acid substrate is produced by immobilized *E. coli* cells as described previously, making this method the first use of two different immobilized cells in series (Fig. 23).

7.4.3. Environmental Applications

Mixed microbial cultures can be adapted by continuous techniques under anaerobic and aerobic conditions to attach and adhere to almost any surface and to degrade a variety of wastes. The organisms develop into a thin film termed a biofilm, which is typically 0.1 – 0.5 mm thick.

If a large surface area is available in the reactor for biofilm growth, then a high concentration of organisms can be retained within a small reactor volume to give a high activity per unit volume. Retention of the biomass as a biofilm is particularly important for slow-growing mixed bacterial cultures, such as those involved in anaerobic treatment of food industry wastes, aerobic nitrification for ammonia removal, and removal of toxic organic wastes, such as phenol. Immobilization allows the process to be operated at high flow rates, which exceed the rates that would otherwise wash suspended organisms out of the reactor. Retention of the active biomass in large quantities allows the reactor to operate at rates up to 50 times higher than those for nonimmobilized systems.

Biofilm reactors can be designed in a variety of ways, depending on whether the process is anaerobic or aerobic. Column reactors lend themselves to both processes. Packed beds of solid particles (e.g., bricks, stones, or plastic media) retain organisms by entrapment and film growth, and are known as aerobic trickling filters or anaerobic filters. Expanded or fluidized beds use small particles (e.g., sand) which are lifted by the upflowing liquid. The organisms must form a tight biofilm, otherwise they would be swept out of the reactor. A variety of other materials (plastic media, honeycomb-type packing, wound steel wire particles, activated charcoal) have been used for biofilm development. Important requirements are (1) maximal surface area, (2) absence of clogging and thus low-pressure drop flow properties for packed beds, and (3) sufficiently high-density particles for fluidized beds.

Plastic foam and other light porous material can be added to conventional activated sludge reactors to entrap biofilms for biomass retention. Another type of reactor, the upflow anaerobic sludge blanket reactor, uses the tendency of anaerobic cultures to form granules by natural flocculation and to thus be retained in a column reactor with slow upward flow.

Most of the above reactors are classified as high-rate processes and have been tested on various scales. They have found limited large-scale application in the biological treatment of industrial waste. The field is still in the developmental stage. Review articles provide more detailed information [251–255].

8. Economic Aspects

The world biotechnology market has a very high growth rate: 25 % of the world pharmaceutical market is already covered by biotechnologically manufactured products. In the chemical industry, this fraction is approaching 10 %. According to market growth projections, U.S. biotechnology industry sales increased from about $\$ 400\times10^6$ in 1987 to $\$ 2.5\times10^9$ by 1990, and more than $\$ 25\times10^9$ by 2000 [256]. The world biotechnology market in the year 2000 was predicted to be $\$ 65\times10^9$ [257].

An algorithm has been developed for preliminary analysis of the economic viability of an immobilized biocatalyst production process for high-value, intermediate-value, and commodity products [258]. A detailed list of market values, in dollars per kilograms, and U.S. market sizes, in kilograms per year, of potential biotechnological products is given in that article. Manufacturing costs of immobilized cell systems are discussed along with the effects of feedstock costs and biocatalyst characteristics (e.g., selectivity, longevity, inherent productivity, and yield) on process economics [258].

Immobilized Enzymes. The total market for industrial enzymes was about $\$ 400\times10^6$ in 1987 [259] and is growing by about 15 % each year. Enzyme immobilization has not initiated

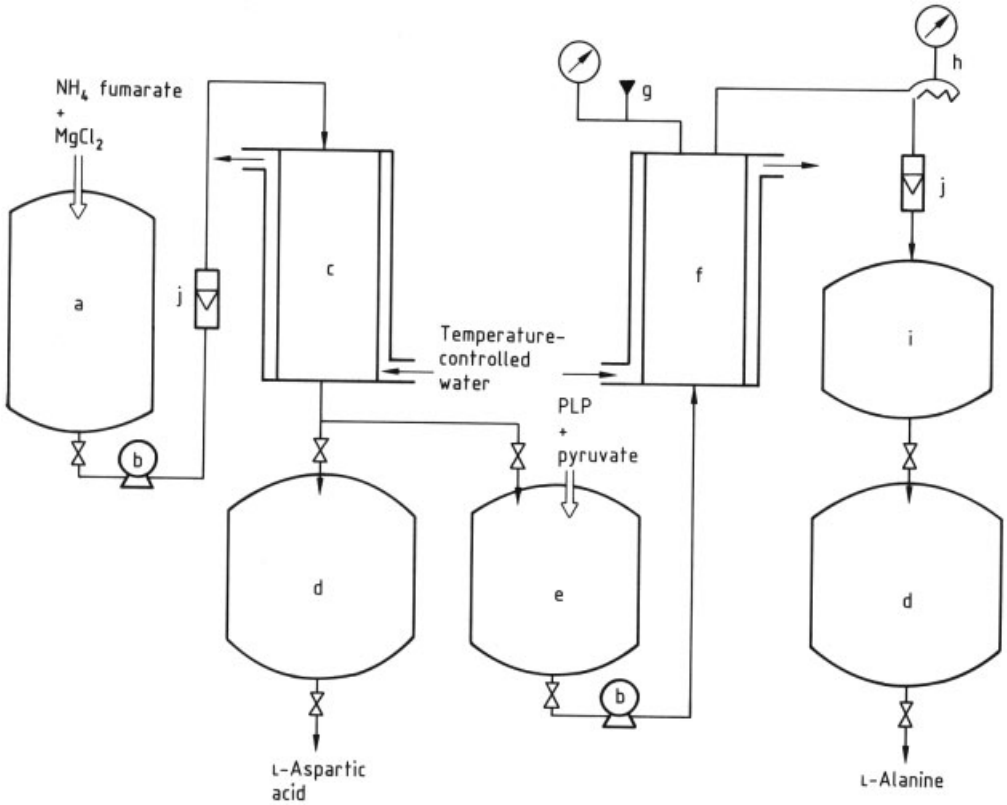

Figure 23. Flow diagram for continuous production of L-aspartic acid and L-alanine by immobilized cells (Tanabe Seiyaku, Japan)
a) Substrate solution tank; b) Pump; c) Immobilized *Escherichia coli* column (1000 L); d) Creptallinger; e) pH control tank; f) Immobilized *Pseudomonas dacunhae* closed column (2000 L); g) Safety valve; h) Pressure control valve; i) Evaporator; j) Flow meter

the expected revolution in the enzyme industry [260]. Soluble use-and-discard enzymes have, by far, the most dominant market share. Only one immobilized enzyme product, immobilized glucose isomerase, is used in amounts > 50 t a year; 1500 – 1750 t are used worldwide annually [261]. The three other major immobilized enzymes are aminoacylase (about 5 t/a), lactase (< 5 t/a), and penicillin G acylase (3 – 4 t/a) [261]. Other immobilized enzymes are also available but are used in even smaller amounts: glucoamylase, hydantoinase, invertase, nitrilase, RNAse, and penicillin V acylase [260].

Many immobilized enzyme biocatalysts are developed and manufactured by the company that uses them. Most economic data for such processes are confidential; published data are extremely sparse. A theoretical economic analysis of the advantages and disadvantages of a continuous immobilized enzyme process, compared to a batchwise soluble enzyme system, is presented in [262].

Immobilized microorganisms are preferred biocatalysts for processes with multiple enzyme reactions especially where cofactor regeneration is essential. However, ascertaining which processes have been industrialized is difficult; in many cases, whether immobilized cells or enzymes are used is unclear [263].

Recombinant *Escherichia coli* can be used for the production of relatively simple *peptide hormones* with pharmaceutical applications such as human insulin, human growth factor, or human interferons (→ Genetic Engineering, Chap. 5.1., → Genetic Engineering, Chap. 5.2.).

An immobilized biocatalyst is reportedly used for the conversion [264]. A *hepatitis B surface antigen vaccine* prepared in recombinant yeast was approved for marketing, and many other products from genetically engineered microorganisms and yeast are in the clinical test stage (interferons, lymphokines, hormones, enzymes, and vaccines) [265].

Immobilized Mammalian Cells. Many complex proteins of pharmacological interest have precise folding and glycosylation requirements that cannot be fulfilled in simple microorganisms. By using hybridoma or recombinant DNA techniques, mammalian cells can often be persuaded to produce and secrete useful quantities of these compounds [265].

Tissue plasminogen activator (TPA) is one of the latest products of this technology. It dissolves blood clots that cause heart attacks and can be produced in recombinant *E. coli* or yeast or in glycosylated form in mammalian cells. The first three months' sales in the United States in 1987 were almost $\$60\times10^6$. Worldwide sales are projected to reach about $\$500\times10^6$ in the next few years [266].

Monoclonal antibodies represent another major group of genetically engineered mammalian cell products. More than 100 monoclonal antibodies are offered, with annual sales of nearly $\$500\times10^6$ [266].

Generally, the economics of microbial systems are more favorable than those of mammalian cell culture systems for recombinant proteins. When the use of recombinant microbial systems is impossible or impractical, cell culture systems will gain a market position if the products are high-value biologicals ($>\$100$/g) [265].

Whether immobilized cell processes will play an important role in classical fermentation is still an open question. Large-scale cell culture technologies are discussed in [267], [268]. Immobilized cell cultures using hollow fibers, ceramic matrices, or microcarriers and microencapsulated cell systems for industrial processes are being developed. β-Interferon was produced from human diploid fibroblasts grown on microcarriers at a scale >1000 L. Anchorage-independent hybridoma cells can be immobilized in perfusion systems [269]. Calculation of reactor productivity and process economics for large-scale cell cultures is described in [268].

Immobilized Plant Cells. Plant cell cultures can also be used for the production of metabolites such as pharmaceuticals, chemicals, flavors, and fragrances. The first product obtained from mass plant cell cultures was shikonin [*517-89-5*], a red pigment composed of eight naphthoquinone molecules. Shikonin is produced by a two-stage fermentation process and is a high-value chemical ($\$4000$/kg) with a limited annual market capacity of ca. 15 kg [270].

Immobilized plant cell systems will be used mainly for products of cells in the stationary growth phase. The release of intracellularly stored products by intermittent permeabilization of immobilized cells can be a great economic advantage, allowing reutilization of the biomass [271]. The continuous immobilized plant cell process in combination with strain selection and improved product leakage allows production of plant-derived chemicals in the range of $\$20$–25/kg [272]. However, in the present industrial state of technology for plant cell cultures, a relatively small number of products have both high value per weight and sufficient market size [273].

9. Safety, Environmental, and Legal Aspects

9.1. Immobilized Enzymes

Although enzymes are found throughout nature, they cannot always be regarded as harmless. Enzymes do not show systemic toxicity in animal tests and have not been reported to be mutagenic, carcinogenic, or teratogenic [274]. However, an enzyme can cause allergic reactions in humans; proteolytic enzymes such as papain can be deleterious to skin or mucous membranes. Futhermore, commercial enzyme preparations contain foreign materials (e.g., additives, stabilizers). In the safety evaluation of commercial enzyme preparations three major areas should be considered:

1) catalytic activity of the pure enzyme,

2) allergenic reactions produced by the enzyme itself or by other proteins in the preparation, and
3) presence of toxic metabolites such as mycotoxins or antibiotics.

The preparation and use of industrial enzymes are subject to a variety of national laws and regulations. Generally, regulations concerning biocatalysts are determined primarly by the end use of the enzyme, such as food-processing, agricultural, or medical applications.

Safety aspects of enzyme preparations in food processing were considered by the Joint FAO/WHO Expert Committee on Food Additives (JECFA) in several sessions. The committee laid down criteria for evaluating enzymes according to the source material [275]. Enzymes obtained from animal, plant, or microbial sources commonly used as food, or used in the preparation of food are accepted as foods if satisfactory chemical and microbiological specifications can be established. No toxicological studies are required for such enzymes. For enzymes obtained from nonpathogenic microorganisms commonly found as contaminants in food, JECFA recommended short-term toxicity experiments. More extensive toxicological studies, as well as chemical and microbiological specifications, are required for enzymes obtained from less well-known microorganisms [276].

In the United States and Canada, enzymes from microbial sources can be marketed only if they are produced according to the guidelines for "good manufacturing practices" (GMP) [277] and the U.S. Food Chemicals Codex (FCC) [278]. In many other countries, regulations are based on these guidelines or on the JECFA recommendations.

In the United States, enzymes used in food, animal feed, medical devices, and diagnostic systems, as well as enzyme-based drugs and bioengineered food products, are regulated by the Food and Drug Administration (FDA). Generally, any product that is considered food can be marketed without undergoing premarket clearance if it is generally recognized as safe (GRAS) [279]. Some enzymes are assigned to this category. Enzymes that are not GRAS are subject to approval as food additives and must undergo comprehensive safety testing before being submitted to the FDA. Lists of enzymes permitted in food manufacture in the United States, Canada, and the United Kingdom can be found in [276].

The carrier and immobilization techniques used for immobilized enzymes must also be studied in safety evaluations [275]. The U.S. regulations require that the fixing agents consist of GRAS substances or that the immobilized enzyme preparation be washed thoroughly to remove fixing-agent residues. The only two fixing agents approved for the preparation of immobilized glucose isomerase in high-fructose corn syrup production are diethylaminoethyl cellulose (DEAE cellulose) and glutaraldehyde [280]. Examples of pending or unsuccessful petitions for GRAS affirmation are given in [280], [281]. Cross-linking agents are highly reactive and often toxic. Many of them lack sufficient toxicity data and are not approved for use in food. On the other hand, the use of nontoxic natural matrix materials such as κ-carrageenan, alginate, or monomer-free prepolymers for enzyme entrapment seems to be a feasible way of producing safe bioctalyst preparations for food processing.

The safety program required to obtain affirmation of GRAS status for immobilized Lipozyme through scientific procedures is presented in [281]. Lipozyme is a lipase from *Mucor miehei* that is immobilized on a macroporous anion-exchange resin; it is used for modification of edible fats and oils. The information required to obtain official approval of immobilized enzymes should be presented by enzyme producers to the authorities as discussed in [282].

As long as the materials used for immobilizing enzymes are toxicologically acceptable or do not to leak measurably into the substrate or reactor effluent streams, the use of immobilized enzymes represents a very safe technique [274].

9.2. Immobilized Cells

A major area of risk assessment in biotechnology is associated with the use of living cells, especially if they are genetically engineered. Extensive public discussion has focused on the possible danger to humans and the environment from the intended or accidental release of recombinant microorganisms (→ Genetic Engineering, Chap. 6.).

The safety of mammalian cell products used in medicine is discussed in [283], [284]. Biochemical purity and the hypothetical risk of transmitting biologically active agents to product recipients should be considered. The risk of contamination by adventitious agents such as fungi, bacteria, and viruses, and the putative risk of tumorigenicity can be eliminated by effective process design, characterization studies, and routine biochemical and microbiological testing [284].

Substantial experimental evidence indicates that even if mammalian tumor cells are introduced accidentally into a human, tumor formation does not occur because the foreign cells are rejected and destroyed. Only large inocula ($> 10^6$ cells) can initiate tumor formation in immunosuppressed rodents [283]. Scientific methods are available to assure that recombinant products derived from continuous mammalian cell lines are safe for use as human therapeutics [284].

Because the highly selected mammalian or plant cell lines used in fermentation cannot survive in the environment without optimized culture media, accidental release of such cells does not present a serious problem. However, unintended release can become a central problem with genetically engineered microorganisms. In industrial processes employing microorganisms, nonpathogens are preferred. If the use of pathogens is unavoidable, they should be physically and biologically contained to prevent or limit their release. Safety limits depend on the risk class of microorganism involved; a risk classification list has been drawn up by the Safety in Biotechnology Working Party of the European Federation of Biotechnology (EFB) [285]. Containment is achieved by employing good laboratory practice, appropriate equipment, and special design. Immobilization should contribute positively to this.

For recombinant DNA work, large-scale industrial processes are arbitrarily considered to involve culture volumes $> 10\,\mathrm{L}$ and are subject to special regulations. Guidelines have been established by the U.S. National Institutes of Health (NIH) [286], [287] and by the Organization for Economic Cooperation and Development (OECD) [284], [288], [289].

These guidelines conclude that the hazards associated with recombinant DNA microorganisms can be assessed and managed in a way similar to that used for other organisms. For EFB Class 1 microorganisms (nonpathogens) free of adventitious agents and having limited capacity for survival without adverse consequences in the environment, the use of good industrial largescale practice (GILSP) is allowed. For Classes 2 and 3 with higher risks, more stringent rules must be obeyed [287], [289].

With immobilized living cell biocatalysts, safety problems exist only in the case of leakage or escape. To prevent such escape, any waste gas should be sterilized by filtering and heat treatment if necessary. Contaminated effluents must be inactivated by chemical or physical means, and accidental spills must be collected. Biotechnological operations can be performed with the same level of safety as microbiological laboratory work if good microbiological techniques (GMT) are applied [289].

The history of the development of government regulation of recombinant DNA research and manufacturing processes in the United States is reviewed in [290], [291]. Efforts to coordinate scientific review procedures, approval policies, and rulings have resulted in a "coordinated framework" [292] that defines the operation of a flexible case-by-case risk assessment [293].

The fundamental question in work with recombinant DNA is whether genetically engineered organisms pose environmental risks that require entirely new, specific regulations. Overall consensus has been reached that the nature of the organism to be released is what matters, not the method by which it was produced [294], [295].

In Europe, an EEC commission is planning to announce a series of biotechnology directives that define the levels of physical and biological containment and accident control, as well as provide the basis for authorizing the planned release of genetically engineered organisms into the environment [295]. The European Committee on Regulatory Aspects of Biotechnology (ECRAB) has suggested a stepwise framework for risk analysis in biotechnological processes from initiation to commercialization [296].

10. References

1. E. S. Kempner, J. H. Miller, *Exp. Cell Res.* **51** (1968) 141.
2. K. Mosbach, B. Matthiasson in K. Mosbach (ed.): *Methods of Enzymology,* **vol. 44,** Academic Press, New York 1976, p. 453.
3. "Multistep Enzyme Systems and Coenzymes," K. Mosbach (ed.): *Methods of Enzymology,* Sect. 1, **vol. 136,** Academic Press, New York 1987.
4. H. Heukelekian, N. C. Dondero: *Principles and Applications in Aquatic Microbiology,* Wiley-Interscience, New York 1964.
5. J. M. Nelson, E. G. Griffin: "Adsorption of Invertase," *J. Am. Chem. Soc.* **38** (1916) 1109 – 1115.
6. I. Chibata, T. Tossa in "Immobilized Microbial Cells," I. Chibata, L. B. Wingard, Jr. (eds): *Applied Biochemistry and Bioengineering,* **vol. 4,** Academic Press, New York 1983, pp. 1 – 9.
7. J. B. Sumner, *Science* **108** (1948) 410.
8. F. Micheel, J. Ewers, *Macromol. Chem.* **3** (1949) 200.
9. D. H. Campbell, E. Luescher, L. S. Lermann, *Proc. Natl. Acad. Sci. USA* **37** (1951) 575.
10. N. Grubhofer, L. Schleith, *Naturwissenschaften* **40** (1953) 508.
11. I. Chibata et al., *Ferment. Technol. Today Proc. Int. Ferment. Symp., 4th,* Kyoto 1972, p. 383.
12. L. B. Wingard Jr. (ed): *Enzyme Engineering,* vol. 1, Wiley-Interscience, New York 1972.
13. K. L. Smiley, G. V. Strandberg, *Adv. Appl. Microbiol.* **15** (1972) 13 – 38.
14. I. Chibata, T. Tosa, T. Sato, *Appl. Microbiol.* **27** (1974) 878.
15. B. Arkles, W. S. Brinigar, *J. Biol. Chem.* **250** (1975) 8856.
16. G. R. Lambert, A. Daday, G. D. Smith *FEBS Lett.* **101** (1979) 125.
17. P. Brodelius, B. Deus, K. Mosbach, M. H. Zenk, *FEBS Lett.* **103** (1979) 93.
18. G. A. Rechnitz, M. A. Arnold, M. E. Meyerhoff, *Nature (London)* **278** (1979) 466.
19. S. Nishikawa, Y. Ashina, I. Watanabe *Kagaku Kogaku* **51** (1987) 423.
20. K. Mosbach (ed.): *Immobilized Enzymes,* **vol. 44**, Academic Press, New York 1976.
21. K. Mosbach (ed): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135**, Academic Press, Orlando 1987.
22. K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part C, Methods in Enzymology,* **vol. 136**, Academic Press, Orlando 1987.
23. L. B. Wingard Jr., E. Katchalski-Katzir, L. Goldstein (eds.): *Immobilized Enzyme Principles, Applied Biochemistry and Bioengineering,* vol. 1, Academic Press, New York 1976.
24. I. Chibata: *Immobilized Enzymes, Research and Development,* Kodansha, Tokyo 1978.
25. W. Hartmeier: *Immobilisierte Biokatalysatoren,* Springer Verlag, Berlin 1986.
26. R. A. Messing (ed.): *Immobilized Enzymes for Industrial Reactors,* Academic Press, New York 1975.
27. M. D. Trevan: *Immobilized Enzymes: Introduction and Applications in Biotechnology,* J. Wiley, New York 1980.
28. A. Wiseman (ed.): *Handbook of Enzyme Biotechnology,* 2nd ed., Ellis Horwood, Chichester UK 1985.
29. J. Woodward (ed.): *Immobilized Cells and Enzymes – A Practical Approach,* IRL Press, Oxford UK 1985.
30. O. R. Zaborsky: *Immobilized Enzymes,* CRC Press, Cleveland, Ohio 1973.
31. L. B. Wingard Jr. (ed.): *Enzyme Engineering,* vol. 1, Wiley-Interscience, New York 1972.
32. E. K. Pye, L. B. Wingard Jr. (eds.): *Enzyme Engineering,* **vol. 2**, Plenum Press, New York 1974.
33. E. K. Pye, H. H. Weetall (eds.): *Enzyme Engineering,* **vol. 3**, Plenum Press, New York 1978.
34. G. B. Broun, G. Manecke, L. B. Wingard Jr. (eds.): *Enzyme Engineering,* **vol. 4**, Plenum Press, New York 1978.
35. H. H. Weetall, G. P. Royer (eds.): *Enzyme Engineering,* **vol. 5**, Plenum Press, New York 1980.
36. I. Chibata, S. Fukui, L. B. Wingard Jr. (eds.): *Enzyme Engineering,* **vol. 6**, Plenum Press, New York 1982.
37. A. I. Laskin, G. T. Tsao, L. B. Wingard Jr. (eds.): *Enzyme Engineering, vol. 7, Ann. N.Y. Acad. Sci. No. 434,* New York 1984.
38. A. I. Laskin, K. Mosbach, D. Thomas, L. B. Wingard Jr. (eds.): *Enzyme Engineering 8, Ann. N.Y. Acad. Sci. No. 501*, New York 1987.
39. P. B. Poulsen, *Enzyme Microb. Technol.* **5** (1983) 304 – 307.
40. A. M. Klibanov, *Science (Washington, D.C.)* **219** (1983) 722 – 727.
41. J. F. Kennedy, J. M. S. Cabral in W. H. Scouten (ed.): *Solid Phase Biochemistry,* J. Wiley & Sons, New York 1983, pp. 253 – 391.

42. Y. Y. Linko. P. Linko in J. Tramper, H. C. van der Plas, P. Linko (eds.): *Biocatalysis in Organic Synthesis,* Elsevier, Amsterdam 1985, pp. 159 – 178.
43. A. Rosevear, *J. Chem. Technol. Biotechnol. Biotechnol.* **34 B** (1984) 127 – 150.
44. J. Tramper in W. H. Scouten (ed.): *Solid Phase Biochemistry,* J. Wiley & Sons, New York 1983, pp. 393 – 477.
45. J. F. Kennedy, J. M. S. Cabral in H.-J. Rehm, G. Reed (eds.): *Biotechnology, A Comprehensive Treatise in 8 Volumes,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, pp. 347 – 404.
46. W. H. Scouten in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 30 – 65.
47. R. A. Messing (ed.): *Immobilized Enzymes for Industrial Reactors,* Academic Press, New York 1975.
48. R. A. Messing in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 2,** Pergamon Press, Oxford 1985, pp. 191 – 201.
49. J. F. Kennedy, C. A. White in A. Wiseman (ed.): *Handbook of Enzyme Biotechnology,* 2nd ed., Ellis Horwood, Chichester UK 1985, pp. 147 – 207, 1985 a.
50. J. F. Kennedy, C. A. White in A. Wiseman (ed.): *Handbook of Enzyme Biotechnology,* 2nd ed., Ellis Horwood, Chichester UK 1985, pp. 380 – 420, 1985 b.
51. A. N. Emery, J. S. Hough, J. M. Novais, T. P. Lyons, *Chem. Eng. (London)* **259** (1972) 71 – 76.
52. J. M. Novais: Ph. D. Thesis, University of Birmingham 1971.
53. J. F. Kennedy, *Chem. Soc. Rev.* **8** (1979) 221 – 257.
54. J. F. Kennedy, J. M. S. Cabral in J. Woodward (ed.): *Immobilized Cells and Enzymes – A Practical Approach,* IRL Press, Oxford UK 1985, pp. 19 –37.
55. F. R. Bernath, K. Venkatasubramanian in A. L. Demain, N. A. Solomon (eds.): *Manual of Industrial Microbiology and Biotechnology,* American Society for Microbiology, Washington D.C. 1986, pp. 230 – 247.
56. J. Iqbal, M. Saleemuddin, *Enzyme Microb. Technol.* **4** (1982) 175 – 178.
57. K. Ikura et al., *J. Appl. Biochem.* **6** (1984) 222 – 231.
58. B. Solomon, R. Koppel, E. Katchalski-Katzir, *Biotechnol. Bioeng.* **28** (1986) 1213 – 1221.
59. B. Solomon, Z. Hollander, R. Koppel, E. Katchalski-Katzir in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 160 – 170.
60. M. Marek, O. Valentova, J. Kas, *Biotechnol. Bioeng.* **26** (1984) 1223 – 1226.
61. G. P. Royer in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 141 – 146.
62. R. Axen, J. Porath, S. Ernbaeck, *Nature (London)* **214** (1967) 1302 – 1304.
63. J. Kohn, M. Wilchek in W. H. Scouten (ed.): *Solid Phase Biochemistry,* J. Wiley & Sons, New York 1983, pp. 599 – 630.
64. I. Ugi, *Angew. Chem. Int. Ed. Engl.* **1** (1962) 8 – 21.
65. J. E. Prenosil, T. Hediger, *Biotechnol. Bioeng.* **31** (1988) 913 – 921.
66. I. Chibata, T. Tosa, T. Sato, *J. Mol. Catal.* **37** (1986) 1 – 24.
67. D. Jaworek in E. K. Pye, L. B. Wingard Jr. (eds.): *Enzyme Engineering,* **vol. 2,** Plenum Press, New York 1974, pp. 105 – 114.
68. D. Jaworek, H. Botsch, J. Maier in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 44,** Academic Press, New York 1976, pp. 195 – 201.
69. L. D'Angiuro et al., *Biotechnol. Bioeng.* **22** (1980) 2251 – 2272.
70. P. Cremonesi, L. D'Angiuro, *Biotechnol. Bioeng.* **25** (1983) 735 – 744.
71. L. D'Angiuro, C. de Lalla, P. Cremonesi, *Biotechnol. Bioeng.* **27** (1985) 1548 – 1553.
72. P. Bernfeld, J. Wan, *Science (Washington, D.C.)* **142** (1963) 678 – 679.
73. K. Kawashima in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 146 – 153.
74. S. Fukui, A. Tanaka, *Adv. Biochem. Eng. Biotechnol.* **29** (1984) 1 – 33.
75. S. Fukui, A. Tanaka, T. Iida in C. Laane, J. Tramper, M. D. Lilly (eds.): *Biocatalysis in Organic Media,* Elsevier, Amsterdam 1987, pp. 21 – 41.
76. D. Dinelli, *Process Biochem.* **7** (1972) 9 – 12.
77. D. Dinelli et al. in E. K. Pye, H. H. Weetall (eds.): *Enzyme Engineering,* **vol. 3,** Plenum Press, New York 1978, pp. 477 – 481.
78. T. M. S. Chang in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 44,** Academic Press, New York 1976, pp. 201 – 218.
79. H. E. Klein, D. W. Sundstrom, D. Shim in J. Woodward (ed.): *Immobilized Cells and Enzymes – A Practical Approach,* IRL Press, Oxford UK 1985, pp. 49 – 54.
80. Y. T. Tu, T. M. S. Chang, *Enzyme Microb. Technol.* **4** (1982) 327 – 331.

81. P. L. Luisi et al. in [37] , p. 549.
82. P. L. Luisi, L. J. Magid, *CRC Crit. Rev. Biochem.* **20** (1986) no. 4, 409 – 474.
83. R. Wichmann, C. Wandrey, A. F. Buchmann, M. R. Kula, *Biotechnol. Bioeng.* **23** (1981) 2789 – 2802.
84. V. G. Gekas, *Enzyme Microb. Technol.* **8** (1986) 450 – 460.
85. T. Takeuchi, K. Makino, *Biotechnol. Bioeng.* **29** (1987) 160 – 164.
86. M. Fujimura, T. Mori, T. Tosa, *Biotechnol. Bioeng.* **29** (1987) 747 – 752.
87. I. Chibata, L. B. Wingard Jr. (eds.): *Immobilized Microbial Cells, Applied Biochemistry and Bioengineering,* vol. 4, Academic Press, New York 1983.
88. B. Mattiasson (ed.): *Immobilized Cells and Organelles,* **vol. 1, 2,** CRC Press, Boca Raton, Florida 1983.
89. C. Webb, G. M. Black, B. Atkinson (eds.): *Process Engineering Aspects of Immobilized Cell Systems,* The Institution of Chemical Engineers, Rugby UK 1986.
90. J. Tampion, M. D. Tampion: *Immobilized Cells: Principles and Applications,* Cambridge University Press, Cambridge 1987.
91. M. Moo-Young (ed.): *Bioreactor Immobilized Enzymes and Cells, Fundamentals and Applications,* Elsevier Applied Science, London 1988.
92. S. Birnbaum, P.-O. Larsson, K. Mosbach in W. H. Scouten (ed.): *Solid Phase Biochemistry,* J. Wiley & Sons, New York 1983, pp. 679 – 762.
93. J. Klein, K. D. Vorlop in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 2,** Pergamon Press, Oxford 1985, pp. 203 – 224.
94. J. Tramper, *Trends Biotechnol.* **3** (1985) 45 – 50.
95. S. F. Karel, S. B. Libicki, C. R. Robertson, *Chem. Eng. Sci.* **40** (1985) 1321 – 1354.
96. I. Chibata, T. Tosa, T. Sato in A. L. Demain, N. A. Solomon (eds.): *Manual of Industrial Microbiology and Biotechnology,* American Society for Microbiology, Washington D.C. 1986, pp. 217 – 229.
97. P. K. Robinson, A. L. Mak, M. D. Trevan, *Process Biochem.* **21** (1986) 122 – 127.
98. P. Brodelius, E. J. Vandamme in H.-J. Rehm, G. Reed (eds.): *Biotechnology, A Comprehensive Treatise in 8 Volumes,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, pp. 405 – 464.
99. C. D. Scott, *Enzyme Microb. Technol.* **9** (1987) 66 – 73.
100. C. Akin in G. E. Russell (ed.): *Biotechnology and Genetic Engineering Reviews,* **vol. 5,** Intercept, Wimborne UK 1987, pp. 319 – 367.
101. P. Brodelius in M. Moo-Young (ed.): *Bioreactor Immobilized Enzymes and Cells, Fundamentals and Applications,* Elsevier Applied Science, London 1988, pp. 167 – 196.
102. J. Klein in M. Moo-Young (ed.): *Bioreactor Immobilized Enzymes and Cells, Fundamentals and Applications,* Elsevier Applied Science, London 1988, pp. 1 – 8.
103. J. Klein, F. Wagner in I. Chibata, L. B. Wingard Jr. (eds.): *Immobilized Microbial Cells,* Academic Press, New York 1983, pp. 11 – 51.
104. J. F. Kennedy, J. M. S. Cabral in I. Chibata, L. B. Wingard Jr. (eds.): *Immobilized Microbial Cells,* Academic Press, New York 1983, pp. 189 – 280.
105. B. Mattiasson in B. Mattiasson (ed.): *Immobilized Cells and Organelles,* **vol. 2,** CRC Press, Boca Raton, Florida 1983, pp. 23 – 40.
106. F. B. Kolot, *Process Biochem.* **16** (1981) no. 5, 2 – 9.
107. F. B. Kolot, *Process Biochem.* **16** (1981) no. 6, 30 – 33.
108. M. E. Hegglin: Ph. D. Thesis no. 8628, Swiss Federal Institute of Technology, Zürich 1988.
109. H. Tanaka, S. Uzman, I. J. Dunn, *Biotechnol. Bioeng.* **23** (1981) 1683 – 1702.
110. I. J. Dunn in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 300 – 307.
111. A. L. van Wezel, *Nature (London)* **216** (1967) 64 – 65.
112. J. M. Clark, M. D. Hirtenstein, *J. Interferon Res.* **1** (1981) 391 – 400.
113. K. Nilsson, *Trends Biotechnol.* **5** (1987) 73 – 78.
114. K. Nilsson, F. Buzaky, K. Mosbach, *Bio/Technology* **4** (1986) 989 – 990.
115. B. K. Lydersen et al. in J. Feder, W. R. Tolbert (eds.): *Large-Scale Mammalian Cell Culture,* Academic Press, Orlando 1985, pp. 39 – 58.
116. G. G. Pugh, G. J. Berg, C. H. J. Sear in G. W. Moody, P. B. Baker (eds.): *Bioreactors and Biotransformations,* Elsevier Applied Science, London 1987, pp. 121 – 131.
117. V. Jirku, J. Turková in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 341 – 357.
118. K. Mosbach, R. Mosbach, *Acta Chem. Scand.* **20** (1966) 2807 – 2810.

119. I. Chibata, T. Tosa, T. Sato, *Appl. Microbiol.* **27** (1974) 878 – 885.
120. G. K. Skryabin, K. A. Koshcheenko in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 198 – 222.
121. M. Kumakura, M. Yoshida, I. Kaetsu, *Biotechnol. Bioeng.* **21** (1979) 679 – 688.
122. C. Bucke in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 175 – 189.
123. I. Chibata, T. Tosa, T. Sato, I. Takata in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 189 – 198.
124. T. Omata, A. Tanaka, T. Yamane, S. Fukui, *Eur. J. Appl. Microbiol. Biotechnol.* **6** (1979) 207 – 215.
125. S. Fukui, K. Sonomoto, A. Tanaka in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 230 – 252.
126. M. L. Shuler in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 372 – 387.
127. C. Heath, G. Belfort, *Adv. Biochem. Eng. Biotechnol.* **34** (1987) 1 – 31.
128. K. Ku et al, *Biotechnol. Bioeng.* **23** (1981) 79 – 95.
129. J. Van Brunt, *Bio/Technology* **4** (1986) 45 – 50.
130. J. Feder in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 393 – 399.
131. G. Klement, W. Schreier, H. W. D. Katinger in M. Moo-Young (ed.): *Bioreactor Immobilized Enzymes and Cells, Fundamentals and Applications,* Elsevier Applied Science, London 1988, pp. 53 – 58.
132. R. R. Mohan, N. N. Li, *Biotechnol. Bioeng.* **17** (1975) 1137 – 1156.
133. F. Lim, A. Sun, *Science (Washington, D.C.)* **210** (1980) 908 – 910.
134. A. Jarvis, T. Gridina, *Bio Techniques* **1** (1983) 22 –27.
135. R. G. Rupp in J. Feder, W. K. Tolbert (eds.): *Large-Scale Mammalian Cell Culture,* Academic Press, Orlando 1985, pp. 19 – 36.
136. J. N. Barbotin et al. in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 454 – 472.
137. A. Tanaka, S. Fukui in B. Mattiasson (ed.): *Immobilized Cells and Organelles,* **vol. 1,** CRC Press, Boca Raton, Florida 1983, pp. 101 – 119.
138. H. Ochiai, A. Tanaka, S. Fukui in I. Chibata, L. B. Wingard Jr. (eds.): *Immobilized Microbial Cells,* Academic Press, New York 1983, pp. 153 – 187.
139. P. E. Gispy, K. K. Rao, D. O. Hall in K. Mosbach (ed.): *Immobilized Enzymes and Cells, Part B, Methods in Enzymology,* **vol. 135,** Academic Press, Orlando 1987, pp. 440 – 454.
140. B. Arkles, W. S. Brinigar, *J. Biol. Chem.* **250** (1975) 8856 – 8862.
141. S. S. Godbole, R. Kaul, S. F. D'Souza, G. B. Nadkarni, *Biotechnol. Bioeng.* **25** (1983) 217 – 224.
142. A. Tanaka, N. Hagi, G. Gellf, S. Fukui, *Agric. Biol. Chem.* **44** (1980) 2399 – 2405.
143. W. Hartmeier, *Trends Biotechnol.* **3** (1985) 149 – 153.
144. B. Hahn-Hägerdal, *Biotechnol. Bioeng.* **26** (1984) 771 – 774.
145. M. F. Cocquempot et al., *Biotechnol Lett.* **4** (1982) 313 – 318.
146. B. Hahn-Hägerdal in B. Mattiasson (ed.): *Immobilized Cells and Organelles,* **vol. 2,** CRC Press, Boca Raton, Florida 1983, pp. 79 – 94.
147. M. Charles, J. Phillips in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 2,** Pergamon Press, Oxford 1985, pp. 225 – 229.
148. K. Buchholz (ed.): "Characterization of Immobilized Biocatalysts," *DECHEMA Monogr.* **84** (1979) 198 – 200.
149. J. E. Prenosil, I. J. Dunn, E. Heinzle in H.-J. Rehm, G. Reed (eds.): *Biotechnology,* vol. 7 a, Chap. 10, VCH Verlagsgesellschaft, Weinheim 1987.
150. A. Bar-Eli, E. Katschalski, *Nature (London)* **188** (1960) 856.
151. P. Monsan, D. Combes in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 137,** Academic Press, New York 1988, pp. 584.
152. K. Martinek, V. P. Torchilin in K. Mosbach (ed.): *Methods in Enzymology,* **vol. 137,** Academic Press, New York 1988, pp. 615.
153. L. Goldstein, Y. Levin, E. Katchalski, *Biochemistry* **3** (1964) 1913 – 1919.
154. D. I. C. Wang et al., *Fermentation and Enzyme Technology,* J. Wiley & Sons, New York 1979.

155. Hor-Da Juang, Hung-Shan Weng, *Biotech. Bioeng.* **26** (1984) 623 – 626.
156. K. Mosbach (ed.): *Methods in Enzymology,* vol. 44, Acad. Press, New York 1976.
157. S. Furusaki, *J. Chem. Eng. Japan* **21** (1988) , 219 –230.
158. K. Buchholz: "Reaction engineering parameters for immobilized biocatalysts," *Adv. Biochem. Eng.* **24** (1982) 39.
159. S. F. Karel, S. B. Libicki, C. R. Robertson, *Chem. Eng. Sci.* **40** (1985) 1321 – 1354.
160. J. M. Radovich, *Enzyme Microb. Technol.* **7** (1985) 2 – 10.
161. C. N. Satterfield, T. K. Sherwood: *The Role of Diffusion in Catalysis,* Addision Wesley, Reading, Mass. 1963.
162. C. N. Satterfield, *Mass Transfer in Heterogeneous Catalysis,* MIT Press, Cambridge, Mass. 1970.
163. P. H. Calderbank in V. W. Uhl, J. B. Gray (eds.): *Mixing,* vol. 2, Academic Press, London 1967.
164. T. K. Sherwood, R. L. Pigford, C. Wilke: *Mass Transfer,* McGraw-Hill, New York 1975.
165. K. Buchholz: "Mass transfer influence on effectiveness," *DECHEMA-Monogr.* **84** (1979) 212.
166. S. F. Karel, S. B. Libicki, C. R. Robertson, *Chem. Eng. Sci.* **40** (1985) 1321 – 1354.
167. H. Tanaka, I. J. Dunn: "Kinetics of Biofilm Nitrification," *Biotechnol. Bioeng.* **24** (1982) 669.
168. D. I. C. Wang et al.: *Fermentation and Enzyme Technology,* J. Wiley & Sons, New York 1979.
169. J. E. Prenosil, I. J. Dunn, E. Heinzle in H.-J. Rehm, R. Reed (eds.): *Biotechnology,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, p. 489.
170. R. E. Spier, J. B. Griffiths: *Animal Cell Biotechnology,* vol. 3, Academic Press, London 1988.
171. C. L. Cooney, *Science (Washington, D.C.)* **219** (1983) 728.
172. I. Chibata in P. Linko, J. Larinkari (eds.): *Food Process Engineering,* **vol. 2,** Applied Science Publ., London 1979, p. 1.
173. K. Venkatasubramanian in P. Linko, J. Larinkari (eds.): *Food Process Engineering,* **vol. 2,** Applied Science Publ., London 1979, p. 162.
174. H. Samejima et al., *Ann. N.Y. Acad. Sci.* **434** (1984) 394.
175. J. J. Heijnen et al., *GWF Gas Wasserfach: Wasser/Abwasser* **126** (1985) 81.
176. J. Haggin, *Chem. Eng. News* **66** (1988) no. 23, 7.
177. E. C. Steiner, G. E. Blau, G. L. Agin: *Simusolv, Modeling and Simulation Software,* Dow Chemical Company, Midland, Michigan 1986.
178. I. Chibata, T. Tosa, T. Sato in J. F. Kennedy (ed.): *Biotechnology,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, p. 653.
179. R. A. Messing, *Biotechnol. Bioeng.* **24** (1982) 1115.
180. I. Nilsson, S. Ohlson, *Eur. J. Appl. Microbiol. Biotechnol.* **14** (1982) 86.
181. M. A. Gencer, R. Mutharasan, *Biotechnol. Bioeng.* **25** (1983) 2243.
182. K. Hashimoto, S. Adachi, H. Noujima, Y. Ueda, *Biotechnol. Bioeng.* **25** (1983) 2371.
183. J. P. Guiraud, A. M. Bajon, P. Chautard, P. Galzy, *Enzyme Microb. Technol.* **5** (1983) 185.
184. P. Monsan, D. Combes, *Biotechnol. Bioeng.* **26** (1984) 347.
185. J. E. Prenosil, E. Stuker, T. Hediger, J. R. Bourne, *Ann. N.Y. Acad. Sci.* **434** (1984) 136.
186. F. Godia, C. Casas, C. Sola, *Biotechnol. Bioeng.* **30** (1987) 836.
187. J. M. S. Cabral, J. M. Novais, J. P. Cardoso, J. F. Kennedy, *J. Chem. Tech. Biotechnol.* **36** (1986) 247.
188. J. B. Griffiths in R. E. Spier, J. B. Griffiths (eds.): *Animal Cell Biotechnology,* **vol. 3,** Academic Press, London 1988, p. 179.
189. O.-W. Merten, *Trends Biotechnol.* **5** (1987) 230.
190. C. G. J. Baker, A. Margaritis, M. A. Bergougnou in M. Moo-Young (ed.): *Advances in Biotechnology,* **vol. 1,** Pergamon Press, Toronto 1981, p. 635.
191. P. F. Cooper, B. Atkinson: *Biological Fluidized Bed Treatment of Water and Wastewater,* Ellis Horwood, Chichester UK 1981.
192. M. Nagashima et al., *Biotechnol. Bioeng.* **26** (1984) 992.
193. G. F. Andrews, *Biotechnol. Prog.* **2** (1986) 16.
194. G. H. Schoutens et al., *J. Chem. Technol. Biotechnol.* **36** (1986) 335.
195. S. J. Chen, C. T. Li, W. K. Shieh, *J. Chem. Technol. Biotechnol.* **35 B** (1985) 101.
196. M. Denac, I. J. Dunn, *Biotechnol. Bioeng.* **32** (1988) 159.
197. G. F. Andrews, J. Przezdziecki, *Biotechnol. Bioeng.* **28** (1986) 802.
198. W. K. Shieh, J. D. Keenan, *Adv. Biochem. Eng. Biotechnol.* **33** (1986) 131.
199. P. W. Runstadler Jr., S. R. Cernek in R. E. Spier, J. B. Griffiths (eds.): *Animal Cell Biotechnology,* **vol. 3,** Academic Press, London 1988, p. 305.

200. M. Hegglin, J. E. Prenosil in O. M. Neijssel, R. R. van der Meer, K. Ch. A. M. Luyben (eds.): *Proceedings 4th European Congress on Biotechnology, 1987,* **vol. 2,** Elsevier, Amsterdam 1987, p. 400.
201. S. B. Karkare, R. C. Dean, Jr., K. Venkatasubramanian, *BioTechnology* **3** (1985) 247.
202. L.-S. Fan, K. Fujie, T.-R. Long, W.-T. Tang, *Biotechnol. Bioeng.* **30** (1987) 498.
203. N. Bernet et al., *J. Ind. Microbiol.* **2** (1987) 129.
204. E. Adema, A. J. Sinskey, *Biotechnol. Prog.* **3** (1987) 74.
205. J. A. Schonberg, G. Belfort, *Biotechnol. Prog.* **3** (1987) 80.
206. S. L. Gallagher, J. P. Tharakan, P. C. Chau, *Biotechnol. Tech.* **1** (1987) 91.
207. J. P. Tharakan, P. C. Chau, *Biotechnol. Bioeng.* **28** (1986) 329.
208. O. T. Schönherr, P. J. T. A. van Gelder in R. E. Spier, J. B. Griffiths (eds.): *Animal Cell Biotechnology,* **vol. 3,** Academic Press, London 1988, p. 337.
209. M. A. Tyo, B. J. Bulbulian, B. Z. Menken, T. J. Murphy in R. E. Spier, J. B. Griffiths (eds.): *Animal Cell Biotechnology,* **vol. 3,** Academic Press, London 1988, p. 357.
210. J. Lehmann, J. Vorlop, H. Büntemeyer in R. E. Spier, J. B. Griffiths (eds.): *Animal Cell Biotechnology,* **vol. 3,** Academic Press, London 1988, p. 222.
211. J. S. Hwang, H. N. Chang, *Biotechnol. Lett.* **9** (1987) 237.
212. T. B. V. Roy, H. W. Blanch, C. R. Wilke, *Trends Biotechnol.* **1** (1983) 135.
213. H. Strathmann, *Trends Biotechnol.* **3** (1985) 112.
214. R. S. Tutunjian, *BioTechnology* **3** (1985) 615.
215. H. N. Chang, *Biotechnol. Adv.* **5** (1987) 129.
216. R. A. Messing, *Biotechnol. Bioeng.* **16** (1974) 897 – 908.
217. K. Buchholz: "Mass transfer influence on effectiveness," *DECHEMA-Monogr.* **84** (1979) 212.
218. K. Buchholz, J. Klein in K. Mosbach (ed.): *Methods in Enzymology,* vol. 135, Acad. Press, Orlando 1987.
219. J. Klein, F. Wagner: "Immobilized whole cells," *DECHEMA-Monogr.* **84** (1979) 265 – 311.
220. V. M. Oven, *Ann. N.Y. Acad. Sci.* **501** (1987) 233.
221. L. J. Kricka, G. H. G. Thorpe, *Trends Biotechnol.* **4** (1986) 253.
222. J. V. Twork, A. M. Yacynych, *Biotechnol. Progress* **2** (1986) 67.
223. D. M. Rawson, *Intern. Ind. Biotechnol.* **8** (1988) no. 2, 18.
224. J. N. North, *Trends Biotechnol.* **3** (1985) 180.
225. G. G. Guilbault, *Analytical Uses of Immobilized Enzymes,* Marcel Dekker, New York 1984.
226. C. R. Lowe, *Trends Biotechnol.* **2** (1984) 59.
227. L. D. Bowers, *Anal. Chem.* **58** (1986) 513 A.
228. K. P. Flint in *Biotechnology of Waste Treatment and Exploitation,* Ellis Horwood, Chichester 1987, p. 32.
229. R. D. Schmid, *Biosensors: International Workshop 1987,* VCH Verlagsgesellschaft, Weinheim 1987.
230. K. Mosbach, "Immobilized Enzymes and Cells," in: *Methods in Enzymology,* Part D, vol. 137, Academic Press, San Diego 1988.
231. T. E. Edmonds, *Chemical Sensors,* Blackie, Glasgow 1988.
232. S. W. May, S. R. Padgette, *Bio/Technol.* **10** (1983) 677 – 686.
233. G. G. Guilbault, *Enzyme Eng.* **6** (1982) 395 – 404.
234. S. Suzuki, I. Karube, *Appl. Biochem. Bioeng.* **3** (1981) 145 – 174.
235. I. Karube, S. Suzuki in A. I. Laskin (ed.): *Enzymes and Immobilized Cells in Biotechnology,* Benjamin/Cummings, Menlo Park 1985, pp. 209 – 227.
236. G. G. Guilbault, G. de Olivera Neto in A. I. Laskin (ed.): *Enzymes and Immobilized Cells in Biotechnology,* Benjamin/Cummings, Menlo Park California 1985, pp. 227 – 261.
237. I. Karube in J. F. Kennedy (ed.): *Biotechnology,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, pp. 685 – 708.
238. V. P. Torchilin in *Adv. Drug Delivery Rev.* **1** (1987) 41 – 86.
239. M. D. Klein, R. Langer, *Trends Biotechnol.* **4** (1986) 179 – 186.
240. T. M. S. Chang in A. I. Laskin (ed.): *Enzymes and Immobilized Cells in Biotechnology,* Benjamin/Cummings, Menlo Park, California 1985, pp. 263 –288.
241. A. Margaritis, F. J. A. Merchant, *CRC Crit. Rev. Biotechnol.* **1** (1984) 339 – 393.
242. F. Godia, C. Casas, C. Sola, *Proc. Biochem.* **22** (1987) 43 – 48.
243. M. J. Nunez, J. M. Lema, *Enzyme, Microb. Technol.* **9** (1987) 642 – 651.
244. C. Akin, *Biotechnol. and Genet. Eng. Rev.* **5** (1987) 320 – 367.

245. P. Linko, Y.-Y. Linko, *CRC Crit Rev. Biotechnol.* **1** (1984) 289 – 337.
246. Novo Technical Report, 175d-GB, Novo Allé, Bagsvaerd, Denmark 1985.
247. I. Chibata, T. Tosa, T. Sato in A. I. Laskin (ed.): *Enzymes and Immobilized Cells in Biotechnology,* Benjamin/Cummings, Menlo Park, California 1985, pp. 37 – 70.
248. I. Chibata, T. Tosa, T. Sato in J. F. Kennedy (ed.): *Biotechnology,* **vol. 7 a,** VCH Verlagsgesellschaft, Weinheim 1987, pp. 653 – 684.
249. I. Chibata, T. Tosa, T. Sato, *J. Mol. Cat.* **37** (1986) 1 – 24.
250. Y.-Y. Linko, P. Linko in J. Tramper, H. C. van der Plas, P. Linko (eds.): *Biocatalysts in Organic Synthesis,* Elsevier, Amsterdam 1985, pp. 159 – 178.
251. M. S. Switzenbaum, *Enzyme Microb. Technol.* **5** (1983) 242 – 250.
252. P. F. Cooper, B. Atkinson (eds.): *Fluidized Bed Treatment of Water and Wastewater,* Ellis Horwood, Chichester 1981.
253. C. Webb, B. M. Black, B. Atkinson (eds.): *Process Engineering Aspects of Immobilized Cell Systems,* Inst. Chem. Engineers, Rugby 1986.
254. C. Webb in C. F. Forster, D. A. J. Wase (eds.): *Environmental Biotechnology,* Ellis Horwood, Chichester 1987, pp. 347 – 376.
255. D. Barnes, P. A. Fitzgerald in C. F. Forster, D. A. J. Wase (eds.): *Environmental Biotechnology,* Ellis Horwood, Chichester 1987, pp. 57 – 113.
256. R. E. Shamel, J. J. Chow, *Chem. Eng. Prog.* **83** (1987) no. 10, 41 – 44.
257. C. G. Rosén, *CHEMTECH* **17** (1987) no. 10, 612 – 618.
258. W. E. Goldstein, *Biotechnol. Bioeng. Symp.* **17** (1986) 763 – 776.
259. P. L. Layman, *Chem. Eng. News* **64** (1986) no. 37, Sept. 15, 11 – 14.
260. P. B. Poulsen in [38] , pp. 413 – 419.
261. P. B. Poulsen: *Third European Congress on Biotechnology, München 1984, Proceedings,* **vol. 4,** VCH Verlagsgesellschaft, Weinheim 1985, pp. 339 – 344.
262. M. M. Domach, T. J. Gricus, *Biotechnol. Bioeng.* **27** (1985) 1458 – 1465.
263. P. Linko: *Third European Congress on Biotechnology, München 1984, Proceedings,* **vol. 4,** VCH Verlagsgesellschaft, Weinheim 1985, pp. 357 – 363.
264. K. Kieslich: *Third European Congress on Biotechnology, München 1984, Proceedings,* **vol. 4,** VCH Verlagsgesellschaft, Weinheim 1985, pp. 39 – 72.
265. A. S. Lubiniecki, *Dev. Ind. Microbiol.* **28** (1987) 133 – 138.
266. L. H. Smithson, *CHEMTECH* **18** (1988) no. 3, 168 – 173.
267. W. R. Arathoon, J. R. Birch, *Science (Washington, D.C.)* **232** (1986) 1390 – 1395.
268. R. W. Swartz in M. Phillips, S. P. Shoemaker, R. D. Middlekauff, R. M. Ottenbrite (eds.): "The Impact of Chemistry on Biotechnology," *ACS Symp. Ser.* **362** (1988) 102 – 120.
269. H. Katinger in O. M. Neijssel, R. R. van der Meer, K. Ch. A. M. Luyben (eds.): *Proceedings 4th European Congress on Biotechnology,* **vol. 4,** Elsevier, Amsterdam 1987, pp. 21 – 42.
270. V. Petiard, P. Steck in J. M. Cardoso et al. (eds.): *Perspect. in Biotechnol. NATO ASI Ser. A* **128** (1987) 139 – 153.
271. P. Brodelius in A. I. Laskin (ed.): *Enzymes and Immobilized Cells in Biotechnology,* Benjamin/ Cummings, Menlo Park, California 1985, pp.109 – 147.
272. O. Sahai, M. Knuth, *Biotechnol. Prog.* **1** (1985) 1 – 9.
273. V. Petiard et al. in O. M. Neijssel, R. R. van der Meer, K. Ch. A. M. Luyben (eds.): *Proceedings 4th European Congress on Biotechnology, 1987,* **vol. 4,** Elsevier, Amsterdam 1987, pp. 157 – 170.
274. P. F. Noordervliet, D. A. Toet in H.-J. Rehm, G. Reed (eds.) *Biotechnology, A Comprehensive Treatise in 8 Volumes*, **vol. 7 a**, VCH Verlagsgesellschaft, Weinheim 1987, pp. 711 – 741.
275. "Evaluation of Certain Food Additives," 21st Report of the Joint FAO/WHO Expert Committee on Food Additives (JECFA), *W.H.O. Tech. Rep. Ser.* **617** (1978).
276. W. H. B. Denner in T. Godfrey, J. R. Reichelt (eds.): *Industrial Enzymology,* The Nature Press, New York 1983, pp. 111 – 137.
277. *Code of Federal Regulations,* 21 CFR Sec. 110.1, Office of Federal Register, National Archives and Records Service, Washington D.C. 1981.
278. *Food Chemicals Codex,* 3rd ed., National Academy of Science/National Research Council, Food and Nutrition Board, Committee on Codex Specifications, National Academy Press, Washington D.C. 1981.
279. J. S. Kahan, J. N. Gibbs, *Appl. Biochem. Biotechnol.* **11** (1985) 489 – 506.
280. H. H. Weetall, W. H. Pitcher Jr., *Science (Washington D.C.)* **232** (1986) 1396 – 1403.

281. J. R. Fordham, N. H. Block, *Dev. Ind. Microbiol.* **28** (1987) 25 – 31.
282. J. R. Reichelt in T. Godfrey, J. R. Reichelt (eds.): *Industrial Enzymology,* The Nature Press, New York 1983, pp. 138 – 156.
283. J. C. Petricciani, *Swiss Biotech.* **5(2a)** (1987) 19 – 20.
284. A. S. Lubiniecki, *Dev. Ind. Microbiol.* **28** (1987) 133 – 138.
285. "Safe Biotechnology – General Considerations," report prepared by the Safety in Biotechnology Working Party of the European Federation of Biotechnology, *Third European Congress on Biotechnology,* **vol. 3**, VCH Verlagsgesellschaft, Weinheim 1984, pp. 647 – 662.
286. "Guidelines for Research Involving Recombinant DNA Molecules," National Institute of Health (NIH) revised version, US Fed. Reg. **49** (Nov. 23, 1984) 46 266.
287. C. Sekhar, *Chem. Eng. (N.Y.)* **63** (1985) no. 17, 57 – 59.
288. *Recombinant DNA Safety Considerations,* OECD, Paris 1986.
289. S. P. Vranch in O. M. Neijssel, R. R. van der Meer, K. Ch. A. M. Luyben (eds.): *Proceedings 4th European Congress on Biotechnology,* **vol. 4,** Elsevier, Amsterdam 1987, pp. 355 – 363.
290. R. L. Greenstein, *Appl. Biochem. Biotechnol.* **11** (1985) 489 – 506.
291. T. O. McGarity, K. O. Bayer in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 4,** Pergamon Press, Oxford 1985, pp. 587 – 608.
292. "Coordinated Framework for the Regulation of Biotechnology," U.S. Fed. Reg. **51** (June 26, 1986) 23 302 – 23 393.
293. D. T. Kingsbury, *Trends Biotechnol.* **6** (1988) 39 – 42.
294. *Introduction of Genetically Engineered Organisms into the Environment: Key Issues,* US National Academy of Sciences, Washington D.C. 1987.
295. B. P. Ager, *Trends Biotechnol.* **6** (1988) 42 – 44.
296. European Committee on Regulatory Aspects of Biotechnology (ECRAB), *Swiss Biotech.* **4(5)** (1986) 15 – 20.

Immunodiagnostic Reagents → Diagnostic Reagents

Immunotherapy and Vaccines

Stanley J. Cryz, Jr., Swiss Serum and Vaccine Institute Berne, Switzerland (Chaps. 1 and2)

Marta Granstrom, Departments of Clinical Microbiology and of Vaccine Production, National Bacteriological Laboratory, Karolinska Hospital, Stockholm, Sweden (Chap. 3)

Bruno Gottstein, Institut für Parasitologie, Universität Zürich, Zürich, Switzerland (Section 4.1)

Luc Perrin, Division d'Hematologie, Hôpital Cantonal Universitaire, Genève, Switzerland (Section 4.2)

Alan Cross, Department of Bacterial Diseases, Walter Reed Army Institute of Research, Washington D.C. 20307-5100, United States (Chap. 5)

James Larrick, Genelabs Incorporated, Redwood City, California, United States (Chap. 6)

1. Introduction . . . 3
1.1. Historical Aspects . . . 3
1.2. Principles and Definitions . . . 4
1.2.1. Antigens . . . 4
1.2.2. Antibodies . . . 5
1.2.3. Immune Response . . . 7
1.2.4. Active Immunization . . . 8
1.2.5. Passive Immunization . . . 9
1.2.6. Genetic Engineering . . . 9
2. Bacterial Vaccines . . . 10
2.1. Diphtheria Vaccine . . . 10
2.2. Tetanus Vaccine . . . 11
2.3. Pertussis Vaccine . . . 12
2.4. Typhoid Fever Vaccine . . . 14
2.5. *Streptococcus pneumoniae* Vaccine 15
2.6. Shigella Vaccines . . . 17
2.7. Cholera Vaccine . . . 17
2.8. Vaccines Against Nosocomial Pathogenes . . . 19
2.9. Meningococcal Meningitis Vaccine . . . 20
2.10. Tuberculosis Vaccine . . . 21
2.11. *Escherichia coli* Vaccines . . . 23
2.12. *Neisseria gonorrhoeae* Vaccine . . 24
2.13. *Hemophilus influenzae* Type b Vaccines . . . 25
3. Viral Vaccines . . . 26
3.1. Measles Vaccine . . . 26
3.2. Mumps Vaccine . . . 27
3.3. Rubella Vaccine . . . 28
3.4. Combined Measles – Mumps – Rubella Vaccine . . . 29
3.5. Polio Vaccine . . . 30
3.6. Hepatitis B Vaccine . . . 31
3.7. Rabies Vaccine . . . 33
3.8. Influenza Vaccine . . . 34
3.9. Varicella Vaccine . . . 35
3.10. Yellow Fever Vaccine . . . 36
3.11. Tick-Borne Encephalitis Vaccine 37
3.12. Japanese Encephalitis Vaccine . . 38
3.13. Smallpox Vaccine . . . 39
3.14. Rift Valley Fever Vaccine . . . 39
4. Vaccines against Parasites . . . 39
4.1. Vaccines against Helminths . . . 39
4.1.1. Vaccines against Schistosoma . . . 40
4.1.2. Vaccines against Nematodes . . . 41
4.1.2.1. Gastrointestinal Nematodes . . . 41
4.1.2.2. Tissue-Invading Nematodes (Filariidae) . . . 42
4.1.3. Vaccines against Cestodes . . . 43
4.2. Malaria Vaccine . . . 44
4.2.1. Strategy for Malaria Vaccine Development . . . 45
4.2.2. Sporozoite Vaccines . . . 45
4.2.3. Asexual Blood Stage Vaccine . . . 47
4.2.3.1. Merozoite Surface Antigens . . . 48
4.2.3.2. Rhoptry Antigens . . . 48
4.2.3.3. Antigens Associated with the Membrane of Infected Erythrocytes . . . 49
4.2.3.4. Other Proteins and Synthetic Peptides . . . 49
4.2.4. Sexual Stages–Transmission Blocking Immunity . . . 50
5. Immunotherapy . . . 50
5.1. Gamma Globulin Preparations . . 51
5.1.1. Standard Immune Serum Globulin . 51
5.1.2. Immunoglobulin for Intravenous Use . . . 52
5.1.3. Hyperimmune Globulins and Antitoxins . . . 53
5.1.4. Production Requirements . . . 53
5.2. Prophylaxis with Immune Serum Globulin . . . 54
5.3. Prophylaxis with Hyperimmune Globulins . . . 55

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_049

5.4. Therapy with Immune Serum Globulin 56
5.5. Prophylaxis and Therapy with Intravenous Immunoglobulin (IVIG) 56
5.5.1. Viral Infection 57
5.5.2. Bacterial Infection 57
5.5.3. Noninfectious Diseases 59
5.5.3.1. Therapeutic Effect of IVIG 59
5.5.3.2. Mechanism of Action 59
5.6. Prophylaxis and Therapy with Plasma and Other Blood Products 60
5.7. Adverse Effects of Gamma Globulin Preparations 61
5.8. Future Prospects 63
6. Immunotherapeutic Uses of Monoclonal Antibodies 63
6.1. Introduction 63
6.2. Bacterial Targets 64
6.3. Viral and Chlamydial Targets . . 65
6.4. Parasite Targets 66
7. References 66

Abbreviations used in this article:

AIDS acquired immune deficiency syndrome
BCG acillus Calmette-Guerin
CMV Cytomegalovirus
CPS capsular polysaccharide
CS circumsporozoite
Da dalton
DNA deoxyribonucleic acid
DPT diphtheria–pertussis–tetanus
DPT-Pol. diphtheria–pertussis–tetanus–polio
DT diphtheria–tetanus
FHA filamentous hemagglutinin
HBIG human anti-HBV immune globulin
HBsAg hepatitis B surface antigen
HBV hepatitis B virus
HIV human immunodeficiency virus
HRIG human rabies immune globulin
humab human monoclonal antibodies
Ig immunoglobulin
IPV inactivated polio vaccine
ISG immune serum globulin
ITP idiopathic thrombocytopenic purpura
IU international units
IVIG intravenous immunoglobulin
Lf limit of flocculation
LPS lipopolysaccharide
MMR measles–mumps–rubella
NANP asparagine–alanine–asparagine–proline
NVDP asparagine–valine–aspartic acid–proline
OPV oral polio vaccine
PFU plaque forming units
PT pertussis toxin
RESA ring-infected erythrocyte surface antigen
RNA ribonucleic acid
TCID tissue culture infectious dose

1. Introduction [1–9]

1.1. Historical Aspects

Immunization is the most efficient, cost-effective means of preventing infectious diseases. The concept of preventing disease by vaccination is old: in China and India, the practice of "variolation," whereby small quantities of material from disease pustules were used to immunize people against smallpox, was practiced before 1000 B.C. The first "rational" approach to vaccination was taken by JENNER in 1798, who used naturally attenuated cowpox to immunize against smallpox. About 100 years later, PASTEUR introduced vaccines against anthrax and rabies based upon attenuated virulent organisms. The discovery by VON BEHRING in 1890, that serum antibodies could neutralize diphtheria toxin opened the door for a new avenue of vaccine development and passive therapy, whereby preformed antibodies were transferred to at-risk patients. By the beginning of the 20th century, serum obtained from immunized animals was used to treat a variety of diseases including diphtheria and tetanus. The use of human serum followed shortly in 1907.

Since the turn of the century, a wide range of vaccines and antisera have been introduced to manage infectious and noninfectious diseases (Table 1). The development of such agents has required expertise from a variety of disciplines including microbiology, biochemistry, immunology, and molecular biology.

Table 1. Vaccines and immunoglobulins

Type of vaccine or immunoglobulin	Disease
Bacterial vaccines	cholera
	diphtheria
	Hemophilus influenzae b
	meningococcal meningitis
	pertussis
	Streptococcus pneumoniae
	tetanus
	tuberculosis
	typhoid fever
Viral vaccines	hepatitis B
	influenza
	Japanese encephalitis
	measles
	mumps
	polio
	rabies
	Rift Valley fever
	smallpox
	tick-borne encephalitis
	varicella
	yellow fever
Immunoglobulins against bacterial diseases	diphtheria
	Hemophilus influenzae, meningococcal meningitis, *Streptococcus pneumoniae* (polyvalent preparation)
	pertussis
	tetanus
Immunoglobulins against viral diseases	cytomegalovirus
	hepatitis A
	hepatitis B
	human immunodeficiency virus (HIV) (normal intravenous immunoglobulin preparation administered to HIV-positive infants)
	measles
	mumps
	rabies
	rubella
	vaccinia
	varicella
Immunoglobulins against noninfectious diseases	hypogammaglobulinemia
	rhesus factor
	idiotypic thrombocytopenia purpura

1.2. Principles and Definitions

1.2.1. Antigens

An antigen is a molecule that can elicit an immune response (either humoral, i.e., antibody-mediated, or cellular, i.e., cell-mediated) or an immune reaction, such as an allergic reaction. An antigen that evokes an immune response is commonly referred to as an *immunogen*. Only foreign or "non-self" molecules are immunogenic. Usually the larger and more complex a molecule is, the more immunogenic it will be. For example, the gram-negative bacterial cell envelope shown in Figure 1 contains many different somatic (cell-associated) antigens, such as lipopolysaccharide (LPS), outer-membrane proteins, and phospholipids. Numerous factors determine the immunogenicity of a purified molecule. Size is of critical importance: proteins with a molecular mass of $\leq 20\,000$ are poorly immunogenic. Similarly, simple polysaccharides composed of a limited number of repeating monosaccharides are not immunogenic unless their molecular mass exceeds 500 000. Small molecules can be rendered immunogenic by covalently coupling them to larger molecules, forming conjugates.

A single antigen may contain many *epitopes,* which are specific areas of the molecule with a three-dimensional configuration that induces an immune response. Complex molecules, such as large proteins composed of many different amino acids, contain more epitopes than a comparatively simple polysaccharide composed of two or three monosaccharide repeats. The immune response to a given antigen can vary greatly among species and individuals within a species due to immune regulation genes.

1.2.2. Antibodies

Antibodies are proteins found primarily in the serum which are produced by B cells in response to contact with a foreign antigen. There are several different classes of antibodies with characteristic functions (see Chap. 5, Table 3). A schematic of an *immunoglobulin G* (IgG) molecule is shown in Figure 2. Immunoglobulins are composed of light and heavy chains held together by disulfide bonds. Each chain has a variable and a constant region. The tertiary structure of the variable region accounts for the specificity of antibody binding. An antibody produced from a given clone of B cells (Section 1.2.3) recognizes and binds to a given epitope or closely related epitopes. Antibodies which recognize more than one epitope are termed *cross-reactive*. The strength with which an antibody binds to an antigen is termed *affinity* and is determined by

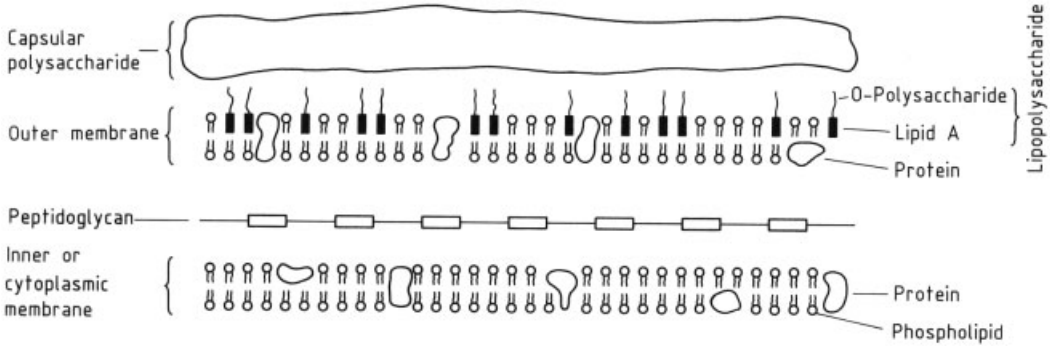

Figure 1. Schematic representation of the gram-negative bacterial cell envelope

the "fit" between the immunoglobulin binding site and the epitope.

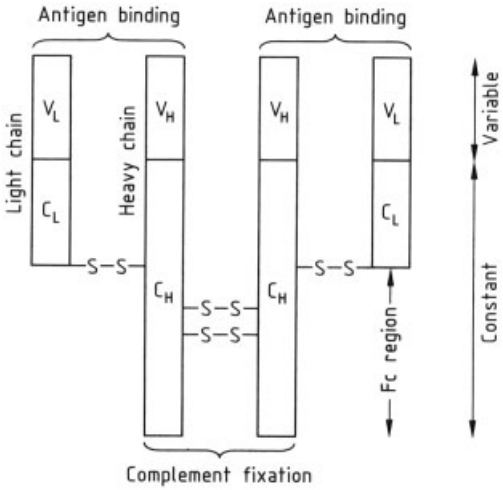

Figure 2. Schematic representation of an immunoglobulin G (IgG) molecule

Antibodies produced by a single clone of B cells are termed *monoclonal antibodies* and recognize only a single epitope (see Chap. 6). *Polyclonal antibodies* are produced by several B cell clones which recognize the same antigen but bind to different epitopes.

Immunoglobulins are divided into five classes termed IgG, IgM, IgA, IgD, and IgE based upon physical and structural differences; they are all glycoproteins. The vast majority of immunoglobulins circulate in the plasma fraction, but certain cells of the immune system can express immunoglobulin on their surface.

Immunoglobulin G (IgG) has a molecular mass of 150 000 and has two antigen binding sites; it represents approximately 80 % of all immunoglobulins in normal serum (8 – 16 mg/mL). Four IgG subclasses (IgG 1 – IgG 4) are found which account for ca. 70, 19, 8, and 3 % of total IgG, respectively; they differ with respect to their antigenic properties in the constant region of the heavy chains. The functional attributes of these subclasses are detailed elsewhere (see Section 5.1). Immunoglobulin G readily crosses the placenta and provides protection against a variety of infectious diseases in the neonate. For example, immunization of the mother against tetanus shortly before delivery ensures a protective level of antibody for the infant. Immunoglobulin G also diffuses into the extravascular tissue more readily than other immunoglobulin classes. The IgG antibody is thought to be responsible for neutralizing the majority of bacterial toxins (e.g., tetanus and diphtheria toxin) formed during an infection. Upon binding with invading bacteria, IgG activates the complement system (a group of interacting serum proteins) which attracts phagocytic cells. The binding of complement components to the Fc region of IgG permits the uptake and killing of the bacteria by phagocytes.

Immunoglobulin M (IgM) is a 900 000 dalton pentamer whose five IgG-like units are held together by intramolecular disulfide bonds and stabilized by a polypeptide "anchor" termed the "J-chain." IgM comprises 5 – 10 % of normal circulating immunoglobulins (1 – 2 mg/mL). Due to its large size, IgM is confined to the intravascular space. It binds complement but, unlike IgG, does not bind directly to phagocytic cells. Immunoglobulin M is usually the first antibody class formed in response to foreign anti-

gen exposure. Due to its multivalency, IgM is extremely efficient at agglutinating bacteria. This phenomenon is important in the control of bacteremia and the neutralization of LPS released by gram-negative bacteria.

Immunoglobulin A (IgA) occurs as 160 000 dalton monomers or 320 000 dalton dimers. Monomers are found primarily in the intravascular space and comprise approximately 10 – 15 % (1.4 – 4 mg/mL) of total serum immunoglobulins. There are two subclasses, IgA 1 and IgA 2. Dimeric IgA, termed *secretory IgA*, is formed by noncovalent interaction with the "secretory piece," a glycoprotein of about 60 000 daltons. Secretory IgA is the predominant immunoglobulin found in mucous secretions, such as tears, colostrum, pulmonary, intestinal, and urinogenital fluids; IgA can bind complement and react with phagocytic cells.

Although parenteral immunization stimulates a vigorous serum IgA response, secretory IgA is considered to be of greater importance due to its "first line" defensive role in body secretions. Secretory IgA plays a critical role in providing protection against respiratory, intestinal, and urinogenital tract bacterial pathogens.

Immunoglobulin D (IgD) is a monomer of ca. 180 000 daltons; concentrations vary widely in humans ranging from 0 to ca. 0.4 mg/mL. Compared to other immunoglobulin classes, IgD is very susceptible to proteolysis and possesses a short half-life (about 3 days). Since IgD does not fix complement or bind to phagocytic cells, it is not considered to be a "protective" immunoglobulin as are IgA, IgG, and IgM. However, IgD is abundant on the surface of B lymphocytes and may play a central role in the activation process which leads to the development of antibody-secreting plasma cells.

Immunoglobulin E (IgE) is a monomer of 200 000 daltons; its average serum concentration (ca. 250 ng/mL) is the lowest of any immunoglobulin. Although IgE does not fix complement or react directly with phagocytic cells, it has a very high affinity for most cells and basophils. Immunoglobulin E appears to mediate both a protective and a detrimental immune response. At the mucoid surface, pathogens binding to IgE stimulate an acute inflammatory response by triggering the release of potent mediators from mast cells and basophils. A protective role for IgE is indicated in several chronic parasitic infections, most notably schistosomiasis (see Chap. 4) where high levels ($> 100\,\mu g/mL$) of serum IgE have been noted. Immunoglobulin E induces an inflammatory response and "recruits" effector cells to the area. In contrast, IgE-mediated degranulation of effector cells leads to the release of vasoreactive molecules. Individuals suffering from certain allergies can also have elevated serum IgE levels.

1.2.3. Immune Response

The chain of events leading to an immune response is extremely complicated and not yet completely understood. A simple model is shown in Figure 3. The human is capable of forming an immune response to thousands of foreign antigens. The humoral (antibody-mediated) response depends on the *B cells* which are lymphocytes derived from bone marrow stem cells. Upon maturation, B cells form antibody-secreting *plasma cells.*

Each clone of B cells has an immunoglobulin molecule with a recognition site for a specific antigen on its surface. Binding of the appropriate antigen to the B cell causes proliferation of the clone whereby the progeny cells secrete antibody whose specificity is identical to that of the cell-surface immunoglobulin. A foreign antigen can be taken up by macrophages, processed, and presented upon the cell surface, or may remain in a soluble state. All antigens can be termed T-dependent or T-independent depending on whether they require or do not require the interaction of T cells for antibody synthesis (T cells are lymphocytes derived from the thymus). In the case of a *T-independent antigen* (usually polymers, such as bacterial capsular polysaccharides), the antigen can cross-link the B cell surface antibody molecules of the B cell; this initiates proliferation and antibody synthesis. A *T-dependent antigen* requires, in addition to binding to B-cell surface immunoglobulin, the release of T cell factors which act upon the B cell to initiate proliferation.

A human exposed to an antigen in this way is considered *primed*. This is a critical factor in immunization. After initial exposure to a T-dependent antigen by immunization, the induced serum antibody concentration is low and returns to near-basal levels comparatively quickly

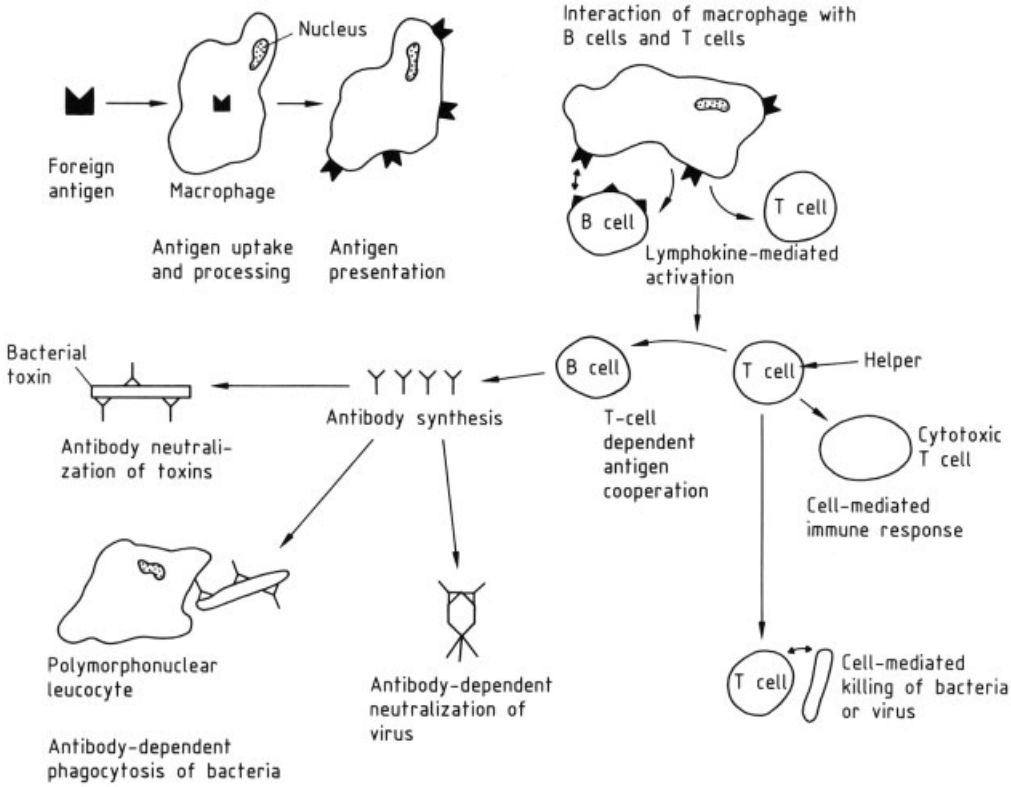

Figure 3. Immunological response cascade to a foreign antigen or vaccine, leading to antibody response or cell-mediated response

(the average half-life of IgG is ca. 22 days, whereas IgM is ca. 5 days). Upon revaccination (boosting) or natural exposure to the pathogen or toxin, a vigorous antibody response occurs in which high antibody levels are synthesized over a longer period of time. This is termed an *anamnestic response.* Therefore, high levels of serum antibodies do not have to be present for the individual to be protected by prior vaccination. A T-independent antigen does not evoke an anamnestic response.

The induction of a *cell-mediated immune (CMI) response* is rather more complicated and less well defined. The CMI response plays a critical role in immunity to pathogens able to live and proliferate within host cells. The first step is "activation" of T cells. Like B cells, T cells also recognize specific antigens. Recognition of an antigen on the surface of a macrophage by T cells results in the release of *interleukin 1* (IL-1), a lymphokine which activates them. One subpopulation of T cells synthesizes interleukin 2 (IL-2), also known as T-cell growth factor, which causes a second subpopulation of activated T cells expressing the IL-2 receptor to proliferate and become cytotoxic T cells. Cytotoxic T cells recognize the specific antigen they are activated against when expressed on the surface of infected cells; they thus kill these infected cells by lysis. A third subpopulation of T cells evolve to be "primed" memory cells which undergo rapid proliferation upon reexposure to the same antigen.

Immunization with a given vaccine may induce a CMI, a humoral antibody response, or both depending on the infecting pathogen. Both types of immunity are usually desirable and may act synergistically. For example, vaccine-induced immunity to many viral diseases, such as rabies, measles, mumps, and rubella, is confirmed by measuring serum antibody levels. However, infected cells can only be destroyed

by cytotoxic T cells. Circulating antibody probably prevents the spread of the virus, while cytotoxic T cells eliminate already infected cells. Tuberculosis vaccines stimulate a good antibody response but are less proctective due to the suppression of the critical CMI response.

1.2.4. Active Immunization

Active immunization entails the administration of one or more antigens to a host in the form of a vaccine in an attempt to elicit a protective immune response. In this way an individual can be rendered immune to a variety of diseases. Most vaccines are administered to infants or young children. The trend is to combine several *monovalent vaccines* (vaccines containing a single antigen or vaccine strain) to form *multivalent vaccines* capable of simultaneously inducing immunity to several diseases. Infants are routinely immunized against diphtheria, tetanus, pertussis, and, in some countries, polio by administering the vaccines in a single injection. Similarly, young children are immunized simultaneously with a multivalent measles, mumps, and rubella vaccine. Administration of multivalent vaccines reduces the number of visits to healthcare centers, a critical point in developing countries.

Historically, *parenteral vaccines* have usually been used (i.e., injected with a syringe and needle). This approach is extremely successful for systemic diseases such as diphtheria or measles. However, in localized diseases (e.g., intestinal infections such as cholera), parenteral vaccines are of limited use since the immune system needs to be stimulated at the site of infection (e.g., local immune system). Therefore, effort is now being directed at controlling intestinal infectious diseases with orally administered vaccines. *Oral vaccines* can consist of

1) killed intact bacteria,
2) toxoids (i.e., toxins that are still immunogenic but are rendered biologically inactive by treatment with a chemical, heat, or mutation)
3) subunit vaccines, in which only the nontoxic portion of a molecule is used, or
4) live-attenuated vaccines, in which a viral or bacterial strain is rendered nonpathogenic (e.g., by passaging the virus in cell culture or deletion of bacterial genes), but is still able to multiply to a limited degree, thereby eliciting a protective immune response in the absence of disease symptoms.

One major problem associated with highly purified antigens or subunit vaccines is reduced immunogenicity. Unfortunately, as protective antigens are purified from either bacteria, viruses, or parasites in order to free them from nontoxic substances such as LPS, they lose their ability to evoke an immune response. Their immunogenicity can, however, be improved by using an *adjuvant.* Adjuvants function either to present the antigen to the immune system for a prolonged period of time, or to nonspecifically stimulate the immune system by releasing immune modulators such as lymphokines. To date, only aluminum gels have been licensed as adjuvants for human use. Tetanus or diphtheria toxoid is absorbed onto the gel which allows the toxoid to persist longer at the injection site. Biologically active peptides that can nonspecifically stimulate the immune system via lymphokine release or downregulate the suppression of the immune system, are being evaluated.

1.2.5. Passive Immunization

Passive immunization entails the transfer of preformed immunoglobulins to a host. Passive immunization is usually employed after known or suspected exposure to a given pathogen (for example, after being bitten by a rabid animal). It is used in cases were disease progress is rapid if the pathogen or toxin is not neutralized. The time needed for the host to mount a protective immune response following active immunization (7 – 14 days) would be too long. In some instances, globulin may be passively administered as a prophylactic measure, as in the case of travellers entering an area where hepatitis A is endemic.

Immunoglobulin preparations are assayed to confirm that they contain a high titer of neutralizing antibodies against the disease in question. Such *hyperimmune globulins* must contain at least five-fold higher levels than normal globulin and are obtained by screening plasma units (human) for a given antibody. Alternatively, vol-

unteers may be vaccinated to enrich their plasma for a given antibody.

Immunoglobulin is administered either intramuscularly or intravenously. Only a limited volume of globulin can be given intramuscularly. Intravenous administration allows for comparatively large quantities (400 mg of immunoglobulin per kilogram body weight) of immunoglobulin to be administered quickly.

Several problems are associated with the preparation of immunoglobulin for passive therapy. Firstly, the identification of plasma donors who are "hyperimmune" to a given antigen requires expensive and tedious screening programs. Secondly, the administration of immunoglobulins entails the possibility of transmitting viral diseases, such as hepatitis B and non-A, non-B hepatitis. Additional screening procedures are therefore required.

An alternative to antibody obtained from donors is the use of human monoclonal antibodies (humabs, see Chap. 6). Humabs can be synthesized by hybridoma cell lines that are produced by fusing a human B cell secreting a desired antibody to a nonsecreting cell line. Hybridomas can be grown in fermentors containing up to 1000 L of serum-free medium. Antibody yields can reach 100 mg per liter of medium. The use of humabs can circumvent the need to obtain plasma from many donors and the risk of viral disease transmission, all at a lower cost. Humabs have been produced against a variety of infectious agents (e.g., cytomegalovirus, diphtheria and tetanus toxins) and will be tested in the near future.

1.2.6. Genetic Engineering

The application of recombinant DNA technology to the field of vaccine development has led to remarkable progress since the late 1970s. Critical protective antigenic determinants can be defined, cloned, produced on a large scale by using an appropriate expression system (vector and host), and purified. The first vaccine for human use to be produced by genetic engineering is that against hepatitis B; the surface antigen was cloned, then expressed in yeast, and purified to homogeneity. These vaccines have virtually replaced the first-generation, plasma-derived antigen vaccines due to increased safety and lower costs.

Numerous licensed, live-attenuated vaccines (measles, polio, mumps, rubella, varicella, and typhoid fever) have been obtained by rather empirical techniques, such as passage on tissue culture or chemical mutagenesis. Surprisingly, the precise genetic alterations responsible for avirulence are unknown. Several live-attenuated bacterial vaccine strains have been developed by using genetic engineering. The most advanced, in regards to clinical testing, are for cholera and typhoid fever. These strains have been attenuated by inactivating a gene or genes essential for virulence. A related approach has been to clone a given protective antigen and express it in a suitable, nonvirulent "carrier" strain. For example, surface glycoprotein from HIV-1 and rabies virus have been introduced into vaccinia virus (smallpox vaccine). In addition, antigens from several enteric pathogens such as *Shigella sonnei, S. dysenteriae,* and *Vibrio cholerae* have been introduced into the licensed live-oral typhoid vaccine strain, Ty21a. Such recombinant strains may prove useful as "bivalent" vaccines, conferring protection against two enteric pathogens.

2. Bacterial Vaccines

The chemotherapy of bacterial disease is treated elsewhere (→ Antibiotics; → Chemotherapeutics, Chap. 2.).

2.1. Diphtheria Vaccine

Etiological Agent and Pathogenesis. The causative agent of diphtheria is *Corynebacterium diphtheriae,* first isolated by Loeffler in 1884. The disease is spread by inhalation of infected droplets and is characterized by the formation of a pseudo-membrane in the nasopharynx caused by localized bacterial replication. Disease symptoms are caused exclusively by the production of a lethal toxin which spreads from the site of infection via the bloodstream. Death is due to the inhibition of protein synthesis by the toxin in vital organs [10].

History of Immunization. Studies by ROUX and YERSIN in 1888 demonstrated that diphtheria is a "toxicosis." Two years later, BEHRING and KITASATO established that the disease could be prevented in animals by immunization with a crude diphtheria toxoid. Attempts to extend these findings to humans used a "toxoid" prepared by combining diphtheria toxin and antitoxin. The first large-scale immunization program with such a product was performed on New York school children in 1922. Soon after, crude toxoids prepared by formaldehyde treatment of *C. diphtheriae* filtrates replaced toxin-antitoxin toxoids. Routine mass vaccination against diphtheria was initiated in Western Europe and North America shortly after World War II.

Production and Properties of Diphtheria Vaccine. Diphtheria toxoid is prepared by detoxification of diphtheria toxin with formaldehyde. Derivatives of the hypertoxinogenic Park Williams 8 strain are used by most manufacturers. Large quantities of toxin are produced in fermentors (ca. 200 – 5000 L) and yields approach 500 mg/L. Although manufacturing processes vary, most employ the following steps. Formaldehyde [*50-00-0*] (0.4 – 0.6 %) is added to cell-free culture supernatants which are then stored at 35 – 37 °C for 3 – 5 weeks. Detoxification is accomplished first by reaction of formaldehyde with the ε-amino groups of lysine and then by formation of irreversible methylene bridges between aromatic amino acids. Upon confirmation of detoxification by animal testing, the toxoid is purified by diafiltration, ammonium sulfate or ethanol fractionation, or anion-exchange chromatography.

Immunization Recommendations. Diphtheria toxoid is rarely used as a monovalent vaccine, but is usually combined with tetanus toxoid (DT, see Section 2.2), pertussis vaccine (DPT, see Section 2.3), or inactivated polio vaccine (DPT-Pol, see Section 3.5) absorbed onto an aluminum salt adjuvant. Toxoid content is usually expressed in limits of flocculation, Lf (1 Lf $\approx$ 2 µg of toxoid). For primary immunization of infants, ca. 25 Lf of toxoid is administered intramuscularly starting at 6 – 12 weeks of age. Three doses are usually administered at 4 – 8 week intervals with a fourth dose 4 – 12 months later. To maintain life-long immunity, booster doses, together with tetanus toxoid, are recommended every 7 – 10 years.

Adverse Reactions. The absolute reactogenicity of diphtheria toxoid is difficult to determine since it is usually administered with tetanus toxoid or pertussis vaccines. Overall, the vaccine is considered to be safe; it primarily evokes mild, transient, local reactions.

Vaccine Efficacy. No large-scale field trials have been performed to determine vaccine efficacy. This is because it would be unethical to withhold the vaccine from control subjects in view of the dramatic decrease in disease following immunization in the 1920s. Furthermore, overwhelming data indicate that proper vaccination can provide absolute protection. For example, mass vaccination with diphtheria toxoid in Romania in 1958 showed that within seven years morbidity and mortality due to diphtheria declined by more then 99 %. In Western Europe and North America, where vaccination is universal, diphtheria has been virtually eradicated [11]. Rare cases occur almost exclusively in individuals lacking a proper history of vaccination.

Future Prospects. Given the combination of vaccine safety, efficacy, and cost, there has been little impetus to develop and test a "second generation" diphtheria toxoid. Any such vaccine would have to completely eliminate the possibility of toxic reversion and be more economical to produce. In another approach, synthetic peptides that express key epitopes of diphtheria toxin are conjugated to carrier proteins; these conjugates can elicit neutralizing antibodies [12]. The drawback with this approach is the cost of purifying the carrier protein.

The most promising approach to toxoid development is the use of nontoxic mutant proteins [13], obtained by recombinant DNA technology. Such proteins can be constructed by deletion of specific regions responsible for toxicity [14]. In addition, these "toxoids" can be synthesized in yields equal to those obtained by production strains of *C. diphtheriae.* The efficacy of such proteins is being evaluated.

2.2. Tetanus Vaccine

Etiological Agent and Pathogenesis. The causative agent of tetanus is the spore-forming bacillus, *Clostridium tetani*. The disease, first described by HIPPOCRATES, is spread via contamination of wounds or abrasions with soil containing *C. tetani* spores. Bacteria multiplying at the wound site release a potent neurotoxin which enters the bloodstream and acts upon the central nervous system evoking the "spastic paralysis" characteristic of tetanus [15]. The majority of tetanus cases occur in neonates due to the nonsterile severing of the umbilical cord.

History of Immunization. The identification of tetanus as a toxicosis by FABER in 1890 led to the development of crude tetanus toxoids as early as 1893. Initial attempts to vaccinate against tetanus were carried out during World War I. By 1926, parenteral immunization of humans with a safe, effective formalin toxoid was achieved. Until the mid-1960s, only such crude toxoids were available.

Production and Properties of Tetanus Vaccine. The presently used tetanus toxoid is a partially purified preparation. The hypertoxinogenic Harvard strain of *C. tetani* is most widely used for production of tetanus toxin. The bacteria are grown in fermentors until the cells autolyse, thereby releasing the toxin into the medium. Intact cells are removed by filtration. Formaldehyde is added to the filtrate to a final concentration of 0.4 – 0.6 %, the pH adjusted to 7.4 – 7.6, and the mixture held at 35 – 37 °C for about four weeks. After confirmation of detoxification by animal testing, tetanus toxoid is purified by diafiltration and ammonium sulfate fractionation.

Immunization Recommendations. Tetanus toxoid is most frequently adsorbed to an aluminum salt adjuvant and administered together with diphtheria toxoid (DT, Section 2.1), pertussis vaccine (DPT, Section 2.3), or inactivated polio vaccine (DPT-Pol, Section 3.5). For primary immunization of infants, each dose of vaccine contains 20 – 30 Lf's of tetanus toxoid as DPT or DPT-Pol. Immunization starts at 6 – 12 weeks of age and consists of three doses given at 4 – 8 week intervals with a fourth dose 4 – 12 months later. To maintain life-long immunity (≥ 0.01 IU/mL of serum), booster doses, often combined with diphtheria toxoid, are recommended every 7 – 10 years.

Tetanus vaccine is often administered as a routine prophylactic measure following puncture wound trauma. This practice can lead to hyperimmunization and attendant reactions. Alternatively, an individual may be "tolerized" by repeated vaccination and, therefore, unable to mount a protective immune response. Vaccination of pregnant women is recommended with the final dose given about three weeks before expected delivery. The immune response of humans to tetanus vaccine varies widely [16]. The immunization schedule used appears to influence the magnitude of the immune response [17].

Adverse Reactions. Serious reactions following vaccination with tetanus toxoid are infrequent: swelling, pain, and/or redness at the injection site are the most common. However, immediate and delayed hypersensitivity, an Arthus-like reaction, and in rare instances (< 1 case in 2×10^6), neurological sequelae have been noted [15]. The majority of severe reactions are associated with high levels of pre-existing antitetanus antibody. Therefore, attention should be paid to the interval between booster doses to avoid these occurrences.

Vaccine Efficacy. Data supporting the efficacy of tetanus toxoid vaccine comes primarily from retrospective analysis of attack rates in immune versus nonimmune populations. The first such study was the evaluation of soldiers during World War II, where tetanus occurred primarily in nonimmune individuals. Immunization of pregnant women completely prevented neonatal tetanus [18]. Recently, cases of tetanus observed in the United States occurred exclusively in nonimmune individuals. Numerous studies have also shown that abbreviated one- or two-dose immunization regimens are able to induce long-lasting immunity [15]. This is of particular importance in developing countries where a multidose immunization regimen may not be feasible due to an inadequate health care system or economic constraints.

Future Prospects. As with diphtheria toxoid, tetanus vaccine is an extremely economic, effective, safe vaccine. Efforts directed to a "new generation" of tetanus vaccine are aimed at providing a toxoid more amenable to mass vaccination. The main thrust has been to produce a safer, more immunogenic toxoid capable of inducing long-lasting immunity after one or two administrations without reactogenic adjuvants. The substitution of glutaraldehyde for formaldehyde yields a safe, immunogenic vaccine not requiring adjuvants [19]. Production of toxoids by recombinant DNA techniques similar to those described for diphtheria toxin is also feasible [14].

2.3. Pertussis Vaccine

Etiological Agent and Pathogenesis. *Bordetella pertussis*, the cause of pertussis or "whooping cough," is a gram-negative rod that was first isolated by Bordet and Gengou in 1906. Humans are the only known host for *B. pertussis*. The disease is spread by infectious droplets with the vast majority of cases occurring in young children. *B. pertussis* shows a marked trophism for the ciliated epithelial cells of the upper respiratory tract. Initial symptoms are similar to those of the common cold, but worsen within 10 – 20 days. The paroxysmal stage, lasting an average of 15 – 20 days, is characterized by severe bouts of coughing. The convalescent stage can last for many months when secondary infections can present a major problem. *B. pertussis* remains localized within the initial disease stages which implies that the severe symptoms are toxin-induced [20]. This is supported by the fact that *B. pertussis* can synthesize a large number of toxic extracellular factors including pertussis toxin (also referred to as lymphocytosis promoting factor), adenylate cyclase, dermonecrotic toxin, and tracheal cytotoxin [21].

History of Immunization. Attempts at immunization against pertussis were initiated in the 1920s using killed whole-cell vaccines [22]. These studies demonstrated that vaccination could not only prevent a substantial proportion of disease, but that symptoms in vaccinated individuals who became infected were milder than in unvaccinated controls. Routine large-scale immunization against pertussis began shortly after World War II using standardized whole-cell vaccines of known potency. Even though the safety of these vaccines has been under attack, they are still used in most areas of the world except Japan, where acellular vaccines have been used since 1980 (see below).

Production and Properties of Pertussis Vaccine. At present, there is no standardized method, culture medium, or bacterial strain for producing pertussis vaccine. A given manufacturer's production procedure is designed to yield a vaccine which will meet the minimal requirements of the appropriate regulatory agency. *B. pertussis* is usually grown in fermentors on a synthetic or semisynthetic medium. The cells are harvested by centrifugation and resuspended to a given opacity using a reference standard. Inactivation is accomplished by heating and/or adding formaldehyde or thimerosal. In view of the diversity of production techniques employed, toxic components which escape inactivation can vary considerably from manufacturer to manufacturer.

Immunization Recommendations. Pertussis vaccine is used almost exclusively in combination with diphtheria and tetanus vaccines (see Sections 2.1 and 2.2). The vaccine is adsorbed onto an aluminum salt adjuvant. Immunization usually commences at 6 – 12 weeks of age and consists of three doses of vaccine given intramuscularly at 4 – 8 week intervals with a fourth dose given 4 – 12 months later. Vaccination with the acellular vaccine used in Japan consists of 2 – 3 doses of vaccine given at 4 – 8 week intervals starting at two years of age.

Adverse Reactions. The majority of children receiving pertussis vaccine experience an adverse reaction. Approximately 40 – 50 % of children have a local reaction and 30 – 40 % a systemic reaction (anorexia, vomiting, fretfulness, fever, or persistent crying) [23]. Of greater concern is the infrequent occurrence of convulsions and hypotonia. As a result of concern regarding vaccine safety, a large-scale study was initiated in England to determine the incidence of serious reactions in children aged 2 – 36 months [24]. The risk of vaccine-

induced neurological illness was estimated to be 1/110 000 vaccinations and that for encephalopathy, 1/310 000 vaccinations. It is extremely difficult to establish a causal relationship in individual cases in such a study; incidence rates were evaluated on a temporal basis. Furthermore, neurological symptoms believed to be vaccine-related constitute only a small proportion of similar cases seen in this age group. Review of these data by health authorities has led to the conclusion that the benefit of vaccination outweighs the attendant risk in view of the fact that clinical pertussis can often result in neurological sequela.

Vaccine Efficacy. The first indication of vaccine efficacy came from a trial conducted in 1929 during an epidemic on the Faroe Islands; vaccination afforded 73 % protection against clinical disease [22]. Extensive information has been obtained in Japan, England, and Sweden where disease incidence correlates inversely with the overall immune status of the general population.

Routine immunization against pertussis in Japan was initiated in 1947 – 1949 with an acceptance rate of roughly 90 %. The number of pertussis cases declined from 152 072 in 1947 to less than 400 by 1971. In 1975, controversy concerning vaccine safety briefly halted vaccine usage. When it was resumed, the acceptance rate declined to ca. 25 – 30 %. Pertussis returned to epidemic proportions by 1979 with more than 13 000 cases reported nationally. At the urging of federal health authorities, vaccine acceptance rates increased to 65 – 70 % by 1982 with a concomitant decline in cases.

Similarly, concern about vaccine safety in England resulted in a dramatic decline in vaccine acceptance from ca. 80 % in 1973 to ca. 30 % in 1978. As in Japan, the disease became epidemic with more than 100 000 cases reported between 1977 and 1980. In Sweden, routine vaccination against pertussis was initiated in the early 1950s. One decade later, > 90 % of children were considered to be immune. Due to changes in the manufacturing technique, the pertussis vaccine used in the mid-1970s was not potent. Shortly after, a marked increase of pertussis cases was seen. When an effective vaccine was reintroduced, disease incidence was dramatically reduced.

Case-contact studies have shown the whole-cell vaccine to be 63 – 95 % effective at preventing overt disease. Efficacy of acellular vaccines is discussed below.

Future Prospects. Efforts to develop a "second generation" pertussis vaccine have centered on using purified detoxified antigen preparations containing a minimum amount of lipopolysaccharide. Several such acellular vaccines have been developed and clinically evaluated. They are composed of detoxified pertussis toxin (PT) alone or in combination with filamentous hemagglutinin (FHA), two key protective antigens [25]. The first generation acellular vaccines were produced in Japan using supernatant from static cultures as a source of antigen. Cell-free supernatants were subjected to ammonium sulfate precipitation followed by sucrose density gradient ultracentrifugation to simultaneously enrich PT and FHA and eliminate lipopolysaccharide; PT was inactivated by formaldehyde treatment. Preliminary testing in Japan showed the vaccine to be far better tolerated than whole-cell vaccine; case-contact studies revealed that it was ca. 90 % effective.

Second generation acellular vaccines of a greater purity have been produced on a large scale using fermentor-grown cultures. A monovalent formalin PT toxoid and a bivalent PT toxoid – FHA vaccine have been evaluated in a placebo-controlled trial in 5 – 11 month old children in Sweden [26]. Children received 2 doses of vaccine 8 – 12 weeks apart. The small number of children vaccinated (ca. 3000) does not allow evaluation of vaccine safety regarding the rare neurological reactions. However, the vaccine evoked far fewer local reactions in comparison to the whole-cell vaccine. After 15 months of observation, the PT toxoid vaccine was 54 % effective whereas the PT toxoid-FHA vaccine was 69 % effective against all forms of clinical pertussis. Both vaccines were equally effective (ca. 80 %) against severe disease.

2.4. Typhoid Fever Vaccine

Etiological Agent and Pathogenesis. *Salmonella typhi,* the causative agent of typhoid fever, is a gram-negative bacillus first isolated

by GAFFKY in 1884. Infection is due to ingestion of the organisms in contaminated food or water. The bacteria penetrate the epithelium of the small intestine and are ingested by reticuloendothelial cells. Unlike most bacteria, the typhoid bacillus can survive and multiply within phagocytic cells and then spreads to the spleen, liver, lymph nodes, and gallbladder. Symptoms appear 1 – 2 weeks after exposure when the bacteria enter the bloodstream. Despite appropriate treatment, 1 – 2 % of those infected will become chronic asymptomatic carriers serving as an infectious reservoir. Typhoid primarily occurs in developing countries or in areas of poor sanitation. In endemic areas, the disease is primarily contracted by school-aged children, but travellers of all ages entering an endemic area are at risk.

History of Immunization. Attempts at immunizing against typhoid fever began in 1896. The first mass vaccination was conducted by the British Army during World War I using a killed whole-cell vaccine. The attack rate among vaccinated soldiers was far less than for nonimmunized soldiers. Consequently, the use of whole-cell vaccines became common practice in the armed forces.

In the 1960s and 1970s, attempts were made to use killed oral immunization. Killed bacteria were administered to volunteers at 1×10^{11} per dose; twelve doses resulted in a modest, but significant protection rate of 30 % [27]. Field trials in India showed that such killed preparations were ineffective at preventing disease.

An alternative approach to killed oral vaccines against typhoid was to use attenuated live-oral vaccine strains. The first candidate was a streptomycin-dependent mutant developed in 1967 [28]. When orally administered to volunteers in doses of up to 10^{11} bacteria, this strain did not evoke adverse reactions. Good levels of protection (66 – 78 %) were achieved with freshly harvested cultures. However, lyophilized preparations afforded little or no protection [28], which diminished interest in this strain.

The second live-oral vaccine candidate was a gal ε mutant termed Ty21a which lacked glucose 1,4-epimerase activity and therefore could not synthesize complete lipopolysaccharide [29]. Freshly harvested cultures of this strain were safe and effective in volunteer challenge studies [30]. A lyophilized Ty21a preparation afforded >90 % protection for three years in a field trial in Egypt [31].

Production and Properties of Typhoid Vaccines. Killed whole-cell vaccines are usually produced from the Ty2 strain of *S. typhi.* Fermentor-grown cultures are inactivated either by heating in combination with phenol or formaldehyde or by acetone drying. Although the latter method yields a more efficacious vaccine (see below), only the vaccine produced by the former method is readily available. Each 0.5 mL dose of vaccine contains $(1-3)\times10^{9}$ killed organisms. Vaccine potency is estimated by a mouse-protection test.

The attenuated live-oral Ty21a vaccine strain is produced from fermentor-grown cultures. The harvested cells are suspended in a cryoprotective medium consisting of sugar and amino acids and lyophilized. The lyophilizate is placed in gelatin capsules which are then coated with an acid-resistant layer. Each capsule contains $(1-5)\times10^{9}$ viable bacteria. Vaccine potency is based upon the number of viable organisms per capsule [32].

Immunization Recommendations. People living in or travelling to an endemic area are vaccinated with two doses of whole-cell vaccine administered subcutaneously or intramuscularly 14 days apart. A single booster dose is advised for individuals travelling to an endemic area if 2 – 3 years have elapsed since primary immunization.

Three doses of attenuated live-oral vaccine are ingested, each dose on an alternate day. An additional complete immunization course is advised when entering an endemic area if more than 1 – 2 years have elapsed since primary immunization. The vaccine is for use in children six years of age or older and in all ages of non-immunocompromised adults.

Adverse Reactions. Reactions following immunization with the parenteral whole-cell vaccine are frequent and consist of local pain, swelling, and redness, often accompanied by fever, chills, and malaise. Transient debilitating reactions occur in ca. 10 % of vaccine recipients. Due to the high rate of adverse reactions, parenteral typhoid vaccine is not widely used.

Reactions following ingestion of the live oral Ty21a vaccine are rare (< 1 per 100 000 doses). Reactions are usually mild, transient, and resolve of their own accord. Gastrointestinal disturbances and rash are most frequently reported.

Vaccine Efficacy. To accurately assess the efficacy of parenteral typhoid vaccine, several field trials were conducted in the 1960s [33]. In most instances, two doses of either heat–phenol or acetone-dried vaccine were administered. Protection rates ranged from 51 – 77 % and 79 –93 %, respectively, for up to seven years. A further trial was conducted in Tonga in 1966 – 1973 where school-aged children received one or two doses of acetone-dried vaccine. Surprisingly, one dose gave no protection whereas two doses afforded only 40 % protection.

Efficacy of the live oral Ty21a vaccine administered in enteric-coated capsules has been evaluated in field trials in school-aged children in Chile [34]. Three doses of vaccine provided about 70 % protection over a three-year period.

Future Prospects. Two additional typhoid vaccines are undergoing clinical evaluation. The first is a double auxotrophic mutant of *S. typhi* requiring 4-aminobenzoate and adenine [35]. Although ingestion of up to 10^{10} organisms only evoked a poor humoral immune response to *S. typhi* cellular antigens, most vaccines manifested a specific cell-mediated immune response to *S. typhi* lipopolysaccharide.

The second candidate vaccine is a capsular polysaccharide termed V_i. Purified V_i antigen is safe upon parenteral immunization. A single dose of vaccine provided ca. 70 % protection for ca. one year in Nepal [36].

2.5. *Streptococcus pneumoniae* Vaccine

Etiological Agent and Pathogenesis. *Streptococcus pneumoniae,* often referred to as "pneumococcus," is a gram-positive diplococcus first isolated by Sternberg and Pasteur in 1881. The pneumococcus usually causes disease subsequent to a viral infection or in a compromised individual, i.e., is an opportunistic pathogen. Pneumonia is the most common disease syndrome caused by *S. pneumoniae*; bacteremia and meningitis are often sequelae to pneumococcal pneumonia. *Streptococcus pneumoniae* is also a leading cause of otitis media in young children. The pneumococcus often colonizes the nasopharynxes of healthy individuals which then serve as infectious foci. The organisms can then spread to another person by infected droplets or translocate to the inner ear or lungs following a viral infection.

History of Immunization. Early attempts at immunization against pneumococcal pneumonia employed heat-killed, whole-cell vaccines (reviewed in [37]). These trials and related studies led to the conclusion that immunity is mediated by serospecific antibody produced against the polysaccharide capsule surrounding the pneumococcus. A large-scale trial performed in the mid-1930s with a bivalent capsular polysaccharide (CPS) vaccine showed a modest decrease in disease. Conclusive proof of vaccine efficacy came from trials conducted with polyvalent vaccines; protection was serospecific. Further studies showed that vaccines comprising up to six distinct capsular antigens could be effectively used [38]. The first pneumococcal polysaccharide vaccines were hexavalent and were licensed in the early 1950s. However, owing to the belief that antibiotics could effectively control pneumococcal disease, vaccine acceptance was low, leading to a halt in production.

Production and Properties of *Streptococcus pneumoniae* Vaccine. This vaccine is composed of 23 types of purified CPS that have been selected on the basis of seroepidemiological surveillance studies of bacteremic isolates [39].

Pneumococcal CPSs are purified by a variety of techniques, depending upon serotype, which include precipitation with organic solvents, treatment with detergent, digestion with DNAse, RNAse, and protease, and ultracentrifugation. The vaccine contains only traces of protein and nucleic acids (≤ 2 wt %) and is analyzed for its CPS constituents, serological purity, and pyrogenicity. Vaccine potency is based upon the determination of the molecular mass of the CPS, because a minimum antigen size is required for immunogenicity in humans. One human dose consists of 25 µg of each antigen in 0.5 mL administered intramuscularly or subcutaneously.

Immunization Recommendations. Pneumococcal vaccine is recommended for immunocompetent individuals with underlying clinical conditions that increase their risk of acquiring pneumococcal bacteremia or pneumonia. These conditions include alcoholism, asplenia, sickle cell anemia, reduced pulmonary function, congestive heart failure, diabetes, reduced renal function, and cirrhosis [40]. Routine vaccination of the elderly is also warranted due to the increased incidence of disease in this group. Patients with underlying conditions (leukemia, Hodgkin's disease, myeloma, and treatment with immunosuppressive drugs) that render them susceptible to pneumococcal disease may be considered candidates for vaccination even though a proportion may not respond to immunization [40]. Revaccination is not advised due to severe reactions.

Adverse Reactions. Pneumococcal vaccine is very safe [40]. Mild local reactions occur in ca. 30 – 40 % of vaccine recipients, fever and severe local reactions in ca. 1 %. Anaphylactic reactions are extremely rare with a frequency of $5/1 \times 10^6$ doses administered. Severe local and systemic reactions are more likely to occur upon revaccination.

Vaccine Efficacy. Considerable controversy exists concerning the degree of protection provided by pneumococcal vaccine. The vaccine was highly effective in reducing the incidence of pneumococcal pneumonia among South African gold miners [41]. However, this population differs substantially in age and in general health from those patients for whom the vaccine is now recommended. Several studies have attempted to determine vaccine efficacy among debilitated and/or elderly patients. The distribution of *S. pneumoniae* serotypes among vaccinated and nonvaccinated individuals with invasive pneumococcal disease has been compared [42]; efficacy was estimated to be 61 – 65 %. In contrast, a vaccine trial on high-risk patients, aged > 55, did not demonstrate any significant benefit [43], probably due to the inability of these patients to mount or maintain a protective antibody response. Similar findings were reported in a study on patients suffering from chronic pulmonary disease [44]. In spite of these divergent findings, health authorities still strongly recommend immunization based upon risk–benefit analysis.

Future Prospects. The current pneumococcal vaccine is poorly immunogenic in children under two years of age and in certain debilitated patient populations [40]. Attempts have been made to improve immunogenicity by covalently coupling pneumococcal CPS to carrier proteins [45]. Such conjugates evoked a more vigorous response in mice and rhesus monkeys when compared to native CPS [46]. However, a type 6A-tetanus toxoid conjugate administered to young adult volunteers was only slightly more immunogenic than native 6A-CPS, and a booster dose did not significantly increase anti CPS antibody levels [47]. However, these volunteers had been previously immunized with tetanus toxoid and possessed low, but detectable, levels of antibody to 6A-CPS.

2.6. Shigella Vaccines

Etiological Agent and Pathogenesis. Shigellosis or bacillary dysentery is caused by *Shigellae,* most notably *S. sonnei, S. flexneri* 2 a and 3, and *S. dysenteriae* 1. Disease results from ingestion of as few as 10 bacteria, making it the most infective bacterial enteric pathogen. In developing countries, where the attack rate can exceed $100\,000/1 \times 10^6$ population, shigellosis is primarily a disease of young children aged six months to six years.

Once ingested, the organisms penetrate the epithelium of the colon and rapidly multiply. This leads to localized inflammation and ulceration, possibly due to synthesis of a potent cytotoxin [48]. Symptoms include fever, bloody diarrhea, cramps, tenesmus, and shock. The mortality rate for untreated disease can exceed 10 %.

History of Immunization. Immunization of humans with killed whole-cell vaccines administered parenterally does not provide significant protection against shigellosis [49]. Subsequent attempts at immunization have therefore centered on use of attenuated live-oral vaccine strains. Vaccines based on streptomycin-dependent derivatives of *S. flexneri* and *S. sonnei* are well tolerated and provide significant immunity for 6 – 12 months [50], [51]. However, these

vaccines were not pursued due to the necessity of administering four doses for primary immunization, yearly boosting to maintain efficacy, and genetic instability.

Attempts have been made to produce live-oral shigella vaccines by conjugal transfer of genes coding for *S. flexneri* surface antigens into *E. coli*. Multiple doses evoked no significant adverse reactions when fed to volunteers, but failed to protect against subsequent challenge [52].

The identification of critical *Shigellae*-protective antigens has allowed their expression in the attenuated live-oral typhoid vaccine strain, *S. typhi* Ty21a [53]. One strain, *S. typhi* Ty21a-5076-1C, carries a plasmid coding for the form I (O-polysaccharide) antigen of *S. sonnei.* This strain is safe and has afforded significant protection in volunteer studies [54].

Future Prospects. At present, no vaccines are available for use against shigellosis. As noted above, *S. typhi* 5076-1C that expresses *S. sonnei* form I antigen has shown promise in volunteer studies. However, efficacy varies from lot to lot and probably depends on the degree of flagellation of the bacteria. The construction of a *S. typhi* Ty21a strain expressing *S. flexneri* 2a type and group antigens has been described [55], but this strain has not yet been clinically evaluated.

Genes coding for the production of the O-antigen of *S. dysenteriae* 1 have been cloned and inserted in a plasmid [56]. The *S. dysenteriae* O-antigen can be expressed in *E. coli* K12 and *S. typhi* Ty21a by introduction of the recombinant plasmid. Volunteer studies to determine their safety are scheduled.

2.7. Cholera Vaccine

Etiological Agent and Pathogenesis. The causative agent of cholera is *Vibrio cholerae*, a gram-negative motile bacillus which is transmitted via the fecal–oral route. Upon passage through the stomach acid barrier, *V. cholerae* colonizes the ileum where it rapidly multiplies. The release of a potent enterotoxin (cholera toxin) causes massive watery diarrhea [57]. In most cases, oral or intravenous fluid replacement is sufficient treatment.

History of Immunization. Parenteral whole-cell cholera vaccines were employed by Fenan as early as 1885. Starting in the 1960s, several field trials were performed using inactivated whole-cell vaccines or cholera antigens (see "Vaccine Efficacy").

Numerous attempts have been made to orally vaccinate against cholera. Multiple oral doses of heat-killed *V. cholerae* evoked both a local and humoral antibody response [58]. Ten daily oral doses of 1.6×10^{10} killed vibrios provided protection against a homologous challenge [59]. However, parenteral vaccine gave slightly better protection. Although promising, the massive quantities of organisms used and the need for multiple doses made the oral approach expensive and cumbersome.

Several environmental isolates of *V. cholerae* with reduced virulence have been evaluated as live-oral vaccines. These strains failed to provide significant protection due to their inability to colonize the small intestine and evoke a protective immune response [60]. A hypotoxinogenic mutant of *V. cholerae* isolated by chemical mutagenesis afforded significant protection against diarrhea. However, due to its propensity to revert, it was not studied further [61]. Chemical mutagenesis has provided a strain of *V. cholerae* 3083 (Texas Star-SR), that produces only the B subunit of cholera toxin responsible for binding to cells and lacks the enzymatically active A subunit. Although vaccination with this strain afforded good levels of protection against cholera, about 25 % of vaccinees experienced mild to moderate diarrhea [62].

Production and Properties of Cholera Vaccine. The licensed cholera vaccine is composed of inactivated whole cells of *V. cholerae* and is administered parenterally. Fermentor-grown cultures are inactivated by a combination of heat and phenol or formaldehyde. Usually, two cultures of Ogawa-classical and Inaba-El Tor serotypes-biotypes are produced and then mixed. Cell concentration is adjusted to $(5-8)\times10^{9}$ cells/dose (0.5 mL). Vaccine potency is estimated by an intraperitoneal mouse challenge test. Two doses of vaccine given at two-week intervals are recommended. A booster dose is recommended whenever entering an endemic area.

Adverse Reactions. Approximately 20–30 % of vaccinees have a mild to moderate local reaction consisting of pain, redness, and/or swelling. Systemic reactions such as fever, headache, or malaise occur in ca. 5 % of vaccinees.

Vaccine Efficacy. Various parenteral vaccines, including inactivated whole-cell, purified lipopolysaccharide, cell-free Inaba antigen, and a whole-cell vaccine in adjuvant have been evaluated in field trials [58]. The results can be summarized as follows:

1) two doses of vaccine are superior to one;
2) protection afforded by vaccination is greater in adults than children;
3) vaccination is only moderately effective in nonendemic areas; and
4) protection lasts for only 3–6 months.

Therefore, the currently available cholera vaccine is not an effective tool to control endemic cholera, it is best employed by travellers entering an endemic area for a short time.

Future Prospects. Due to advances in understanding the pathogenesis and molecular genetics of *V. cholerae*, one or more safe and effective cholera vaccines will probably be introduced in the near future.

Several attenuated live-oral vaccine strains have been developed by deleting the cholera toxin gene or its enzymatically active A subunit [63], [64]. A derivative of *V. cholerae* 16961, in which the cholera toxin gene was deleted afforded 89 % protection against clinical disease although 50 % of the volunteers had mild to moderate diarrhea. A second vaccine strain was produced by deletion of the A subunit of cholera toxin from strain 395. Again, good protection was afforded by a single dose, but mild diarrhea was observed in 60 % of the subjects. The discovery, that strains of *V. cholerae* can produce a Shiga-like cytotoxin suggested that at least part of the diarrhea could be due to this toxin [65]. Therefore, Kaper and coworkers (personal communication) deleted the A subunit of cholera toxin from strain 569B which is naturally cytotoxin negative. This strain, called CVD-103, is far less reactinogenic than previously evaluated strains, eliciting mild diarrhea in ~ 10 % of subjects. Vaccination with CVD-103 afforded excellent protection (87 %) against a homologous challenge. Furthermore, significant protection (67–78 %) was seen even when the challenge was of a different serotype or biotype.

Orally administered inactivated vaccines consisting of 1×10^{11} killed vibrios and 1 mg of the B subunit of cholera toxin or 1×10^{11} killed vibrios alone per dose have been evaluated [66]. After six months of surveillance, three doses of the combined vaccine afforded 85 % protection, whereas the cells alone were 58 % effective. After one year, protection had declined to 62 % for the combined vaccine and 53 % for the cells alone.

2.8. Vaccines Against Nosocomial Pathogenes

The routine use of broad-spectrum antibiotics, more frequent invasive surgery, increasing use of immunosuppressive agents in the management of cancer, and the ability to prolong the lives of critically ill patients have led to a dramatic increase of hospital-acquired (nosocomial) bacterial infections [67]. In the United States alone, nosocomial infections are a contributing factor in ca. 70 000 deaths per year; their treatment costs $> 1\times10^{9}$ \$ [68], [69]. The mortality rate for nosocomial bacteremia and pneumonia has remained unacceptably high (≥ 25 %) despite the introduction of numerous antibacterial agents [68]. Therefore, much effort has been devoted to developing immunological agents for controlling these infections [70].

Etiological agents and pathogenesis. *Pseudomonas aeruginosa*, *Escherichia coli*, *Klebsiella* spp., and *Staphylococcus aureus* account for the majority of serious nosocomial infections, such as bacteremia and pneumonia [71]. Infection with these agents usually arises due to the translocation of bacteria from the skin, intestine, or nasopharyngeal cavity following disruption of the host's defense systems by trauma, invasive surgical procedures, treatment with antibiotics, or immunosuppressive agents. Foci of infection are usually formed from which the bacteria enter the bloodstream. Death is most likely caused by shock.

Future Prospects. No vaccines are licensed against the four pathogens listed above. Attempts to control bacterial nosocomial infections by immunological means are complicated because of the diverse patient populations who are at high risk and the substantial number of bacterial pathogens of varying serotypes involved. Furthermore, vaccination of at-risk patients may not be feasible because of the time needed to mount a protective immune response and because a substantial portion of these patients are unable to mount a significant antibody response. A more effective approach is to passively immunize patients by using a hyperimmune globulin for intravenous use (see Chap. 5): vaccines developed against nosocomial pathogens could then be used to vaccinate healthy donors whose plasma would be processed into such a globulin [70].

Two different approaches are being taken to develop appropriate hyperimmune products. The first is based upon developing vaccines against the relevant serospecific bacterial antigens. The second is to evoke an antibody response to a common epitope expressed by the lipopolysaccharide of gram-negative bacteria.

Serospecific Vaccine against *Pseudomonas aeruginosa.* Human immunity to *P. aeruginosa* depends on the presence of humoral antibody directed against serospecific lipopolysaccharide (LPS) determinants and toxin A [72]. Several lipopolysaccharide-containing vaccines have been developed and clinically evaluated [73]. These vaccines have not gained wide acceptance due to their toxicity, poor immunogenicity, and poorly characterized antigenic content. Two serotypes of *P. aeruginosa* high molecular mass polysaccharides have been purified and found to be safe and immunogenic in humans [74]. In an attempt to produce both an antitoxin A and an antiLPS antibody response, toxin A – O-polysaccharide conjugate vaccines were synthesized. Such conjugates are nontoxic, nonpyrogenic, safe when administered to humans, and evoke an immune response to both vaccine moieties [75]. A polyvalent vaccine based upon O-polysaccharide serotypes is undergoing clinical evaluation; this vaccine would "cover" >90 % of *P. aeruginosa* bacteremic isolates.

Serospecific Vaccine against *Klebsiella spp.* Antibody to *Klebsiella* capsular polysaccharide (CPS) is highly protective against experimental infections [76]. Several experimental vaccines composed of purified *Klebsiella* CPS were found to be safe and immunogenic in human volunteers [70]. Based upon the seroepidemiology of *Klebsiella* blood isolates, a 24-valent CPS vaccine has been developed which is safe and immunogenic in humans. It also elicits production of antibody to 11 "cross-reactive" CPS serotypes not included in the vaccine and therefore, offers potential "coverage" against ca. 75 % of all *Klebsiella* bacteremic isolates.

Serospecific Vaccine against *Escherichia coli.* Antibodies to both O (LPS) and K (capsular) antigens can provide protection against experimental *E. coli* infections [77], [78]. However, ca. 40 % of *E. coli* bacteremic isolates do not have a polysaccharide capsule [79]. In addition, the two most frequently encountered K serotypes among blood isolates, K 1 and K 5, are nonimmunogenic in humans. Approximately 90 % of *E. coli* bacteremic isolates can be typed according to their O-antigen and most of them can be grouped within 11 serotypes, making a polyvalent formulation feasible [79]. The toxicity of native LPS precludes its use as a vaccine. However, nontoxic, serologically reactive O-polysaccharide can be isolated from *E. coli* LPS and O-polysaccharide – protein conjugates are being constructed in a manner similar to that described for *P. aeruginosa* (see above).

Serospecific *Staphylococcus aureus* Vaccine. *Staphylococcus aureus* produces numerous somatic antigens and extracellular toxins which have been implicated as virulence factors. These include teichoic acid, lipoteichoic acid, exopolysaccharide, capsular polysaccharide, α-toxin, and β-toxin. Patients with deep-seated *S. aureus* disease usually mount an antibody response to teichoic acid, α-toxin, and β-toxin [80]. However, the relative protective capacities of antibodies against these antigens are unknown.

A serotyping scheme has been described for *S. aureus* based upon cell-surface capsular polysaccharides [81]. Approximately 80 – 90 % of *S. aureus* clinical isolates belong to one of these serotypes [82]. Serospecific CPS is pro-

duced during experimental *S. aureus* infections and can be isolated from blood [83]. The protective capacity of antibody to CPS is being evaluated.

Crossreactive Anticore Glycolipid Vaccine. Antibody produced against the "core" glycolipid of *E. coli* strain J 5 crossreacts with the LPS of virtually all gram-negative bacteria [84]. Polyclonal or monoclonal antiJ 5 antibody (see also Section 6.1) can afford significant protection against several gram-negative pathogens [85]. Protective antibody is directed against the lipid A moiety of LPS and neutralizes the toxic activities of LPS. When used prophylactically or therapeutically, human plasma enriched with antiJ 5 antibody provides significant protection against death due to "gram-negative shock" [86], [87].

2.9. Meningococcal Meningitis Vaccine

Etiological Agent and Pathogenesis. The causative agent of meningococcal disease is *Neisseria meningitidis*, a gram-negative diplococcus. Humans are the only known reservoir for *N. meningitidis*. The meningococcus colonizes the nasopharynx; infrequently, it enters the bloodstream and infects the meninges. Meningococcal disease is most often seen in children less than 18 months of age and in settings such as military recruit camps where individuals from different geographical areas come into constant close contact [88]. The disease is endemic worldwide with an annual incidence of about 3/100 000 population. Numerous epidemics occurred in the 1970s and 1980s with attack rates exceeding 500/100 000. Eleven meningococcal capsular serotypes are known, with the majority of disease (ca. 95 %) caused by groups A, B, C, W 135, and Y [88], [89].

History of Immunization. The first attempts at immunizing against meningococcal disease used vaccines made of killed bacteria and were unsuccessful [90]. Attention then centered on using purified capsular antigens. In 1945 Kabat demonstrated that up to 1 mg of purified antigen could be administered to humans with no untoward reactions. However, these studies were not continued because the prophylactic treatment of meningococcal meningitis with antibiotics appeared to be highly effective. However, the appearance of resistant strains in the early 1960s revived interest in vaccine development. In the late 1960s, highly purified group A and group C capsular polysaccharides with a high molecular mass were shown to be safe and immunogenic in humans [91]. Human immune sera contained elevated levels of antibody capable of lysing meningococci in the presence of complement.

Production and Properties of Meningococcal Vaccines. Available vaccines contain serogroup A, C, W 135, and Y capsular antigens. The most widely used formulations are the tetravalent A, C, W 135, and Y, and the bivalent A and C vaccines. Capsular antigens are purified from fermentor-grown cultures by coprecipitation with detergents followed by ethanol fractionation, extraction with organic solvents, and ultracentrifugation. Preservation of the high molecular mass of the capsular antigens is of critical importance since this governs the immune response. Each vaccine dose contains 50 µg of capsular antigen per serotype in lyophilized form. Potency of the vaccine is based upon various physicochemical properties (size and purity). No animal test is required.

Immunization Recommendations. Meningococcal vaccine can be administered to any immunocompetent individual over 18 months of age. A single dose is given either subcutaneously or intramuscularly. The group A antigen is immunogenic in children $\geq$6 months of age, but two doses must be administered within a 2 – 3 month interval. Due to the low incidence of the disease in the general population, vaccine acceptance has been poor, it is usually used to immunize military recruits or travellers.

Adverse Reactions. The meningococcal vaccines are very safe–about 250×10^6 doses of vaccine have been administered worldwide with no vaccine-associated fatalities reported. Mild, transient local reactions occur in about 25 % of vaccinees.

Vaccine Efficacy. Conclusive proof that group C meningococcal vaccine prevents disease was presented in 1970 [92]. Subsequently, the group A vaccine was found to be effective

in controlling endemic and epidemic disease in adults and children [93], [94]. Group C vaccine elicited a protective antibody response only in children $\geq$18 months of age, whereas the two doses of group A vaccine were immunogenic in children $\geq$6 months of age [95]. Efficacy for the W 135 and Y capsular antigens has not been clinically demonstrated, but is assumed because they engender appropriate levels of relevant functional antibodies in humans.

Future Prospects. Current meningococcal vaccines have two major limitations. First, they do not provide protection against group B organisms which may account for up to 50 % of endemic cases. Secondly, the group C vaccine is nonimmunogenic in children $\leq$18 months of age. The major difficulty in producing a group B meningococcal vaccine is that the group B capsular antigen, a homopolymer of α (2$\rightarrow$8)-linked scialic acid, is poorly immunogenic in humans, probably because similar structures are found in human gangliosides and fetal proteins. Even when coupled to carrier proteins [96] or mixed with group B outer membrane proteins [97], the group B antigen is a poor immunogen. An alternative approach is to use serotype-specific outer membrane proteins, 18 such serotypes have been identified. Outer membrane protein vaccines containing group B or C capsular antigens were found to be safe and immunogenic in humans [98], [99]. A large-scale field trial to determine the efficacy of these vaccines has been conducted.

In an attempt to construct vaccines capable of engendering a protective immune response in young children, capsular antigens have been coupled to carrier proteins to form conjugate vaccines [96]. Group A and C antigens coupled to tetanus toxoid are more immunogenic in animals than native capsular antigens. Studies are in progress to determine the safety and immunogenicity of group A and C conjugates in humans.

2.10. Tuberculosis Vaccine

Etiological Agents and Pathogenesis. Human pulmonary tuberculosis is caused by the bacilli, *Mycobacterium tuberculosis* and *M. bovis*, the latter causing disease primarily in children. *M. tuberculosis* is spread by infected droplets, whereas *M. bovis* is disseminated by ingestion of contaminated milk. The vast majority of clinical disease occurs in middle-aged individuals in developed countries and in children under 10 years of age in underdeveloped countries. The incidence of tuberculosis ranges from ca. 10/100 000 population in Europe and North America to 500/100 000 in parts of Africa. Although 95 – 98 % of people exposed to *M. tuberculosis* become infected, i.e., become tuberculin-positive, they do not manifest signs of clinical disease and are immune [100]. Only a small percentage of infections progress to a disease state. Once the inhaled bacilli reach the lower respiratory tract, they are ingested by alveolar macrophages. Rapid intracellular growth with inflammatory cell recruitment leads to granuloma formation. In the majority of cases, this is as far as the disease progresses. Active disease presents clinical symptoms ranging from self-limiting pulmonary involvement to acute disseminated disease which rapidly leads to death.

History of Immunization. Efforts to protect humans against tuberculosis have almost exclusively employed *Bacillus Calmette – Guérin* (BCG) vaccine (see below). Oral BCG vaccine was first evaluated in the 1920s in newborns. Subsequent studies showed BCG vaccine to be safe when given orally or parenterally. In the early 1930s, large-scale trials with BCG vaccine were started. Subsequent trials involving American Indians and Eskimos showed that intradermally administered BCG conferred protection against disease in infants. Additional proof of vaccine efficacy was obtained in a small-scale trial with Canadian Indians [101]. Even in the face of a high disease transmittance rate and a mortality rate of 800/100 000, the vaccine was more than 80 % effective over a 9-year observation period.

Production and Properties of Tuberculosis Vaccine. The original BCG vaccine strain was developed at the Pasteur Institute over 50 years ago by in vitro cultivation of *M. bovis* for 10 years. Numerous substrains are now used to produce commercial products. The bacteria are grown on Sauton's liquid media either in bottles with a high surface area or in submerged culture. They are then collected, suspended in a sta-

bilizer solution, homogenized, and lyophilized. The bulk material is standardized by opacity, dry mass, and number of viable bacteria per unit mass. The vaccine is usually filled into glass ampules which are sealed under vacuum.

Controls of BCG vaccine are stringent. First, the culture is examined for purity, viability, and identity. Innocuity is tested by injection of vaccine into guinea pigs. The number of live organisms are determined by viability counting. General safety tests to confirm absence of unexpected toxicity are also performed. Resistance of the bacteria to elevated temperature is also documented. Vaccine potency is based upon the total viable bacteria per dose. However, this is not totally satisfactory because (1) different BCG strains may vary in their ability to evoke a protective cell-mediated immune response, and (2) the ratio of viable to nonviable bacteria can influence the magnitude of both the humoral and cell-mediated immune response to vaccination [102].

Immunization Recommendations. A single intradermal dose containing $(4-8)\times 10^5$ viable organisms in 0.1 mL is normally administered. The preferred use of BCG vaccine depends on the disease incidence in the area in question. In developed areas where disease is infrequent, vaccination should be limited to individuals at high risk to exposure, i.e., family members of tuberculosis patients, health care workers expected to come into contact with tuberculosis patients or infected clinical specimens, or travellers entering an area of high endemicity for a prolonged period of time. In developing areas with a high rate of disease incidence, routine immunization of infants is a cost-effective method of controlling tuberculosis. In addition, all tuberculin-negative children upon entry into school and tuberculin-negative adults should be vaccinated.

Adverse Reactions. Correctly administered BCG vaccine is considered to be safe; one vaccination in 100 000 will result in a noticeable reaction [103]. A significant proportion of these reactions are ulcers caused by inadvertent subcutaneous administration of the vaccine. Of greater concern are the reports of mycobacterial disease following immunization [104]. Great care must be taken to insure that a potential vaccinee is not immunocompromised as the vaccine organisms can then rapidly multiply and lead to death [105].

Vaccine Efficacy. In numerous field trials the efficacy of BCG vaccine ranged from 0–80 % [100]. Although three trials failed to document significant vaccine-induced protection, considerable support still exists for large-scale vaccination in developing countries [106].

Future Prospects. At present, no "second-generation" tuberculosis vaccines are being clinically evaluated. Although the tuberculin skin test offers a practical method to measure "immunity" to disease, the molecular mechanism responsible for inducing a tuberculin-positive state is not known.

Several *M. tuberculosis* cell surface antigens have been identified which can induce a cell-mediated response in experimental animals [107]. The cloning of such antigens and their high-level expression in bacteria or yeast would permit economical large-scale purification. Alternatively, such cloned genes could be inserted into a suitable vector such as vaccinia virus [108]. Since most tuberculosis vaccine are used in underdeveloped nations, any new vaccine must not only be safe and effective, but also economical to produce and control.

2.11. *Escherichia coli* Vaccines

Etiological Agent and Pathogenesis. *Escherichia coli*, a gram-negative motile rod, is a major cause of diarrheal disease, urinary tract infections, neonatal meningitis, and nosocomial infections (see Section 2.8). *Escherichia coli* produces a variety of virulence factors including a polysaccharide capsule antigen (K-antigen), lipopolysaccharide antigen (O-antigen), fimbriae, enterotoxins, and cytotoxins. The majority of *E. coli* strains causing a given disease fall within a limited number of K or O serotypes.

Escherichia coli is the major cause of diarrhea worldwide. Three "categories" are recognized, i.e., enterotoxigenic, enteroinvasive, and enteropathogenic *E. coli* strains.

1) *Enterotoxigenic E. coli* invariably causes watery diarrhea which may be accompanied

by cramps, fever, or vomiting. The disease is acquired by ingesting contaminated food and affects infants, young children, and travellers to underdeveloped countries. The strains produce either a heat-stable enterotoxin and/or a heat-labile enterotoxin; they also possess fimbriae which mediate attachment to the ileal epithelium. Three distinct fimbriae (CFA/I, CFA/II, and E 8775) have been identified on enterotoxigenic *E. coli* strains isolated from human disease.

2) *Enteroinvasive E. coli* causes a dysentery-like disease characterized by fever, diarrhea, cramps, and stools containing blood and mucus. The disease is acquired by ingestion of contaminated food. The bacteria readily penetrate the ileal and colonic epithelia where they rapidly multiply resulting in tissue destruction. They can then infect adjacent cells or penetrate the underlying lamina propria.
3) *Enteropathogenic E. coli* affects primarily infants and young children after ingestion of contaminated food causing watery diarrhea often accompanied by fever and/or vomiting. These *E. coli* strains rarely produce a heat-labile or heat-stable enterotoxin and are unable to invade eukaryotic cells. They colonize the ileal epithelium with destruction of the surrounding brush border in the absence of invasion. Enteropathogenic *E. coli* strains can synthesize a cytotoxin similar to that produced by *Shigella dysenteriae* 1. This so-called Shiga-like toxin is thought to elicit the clinical symptoms.

Approximately 50 % of neonatal meningitis is due to *E. coli* transmitted from mother to infant. The bacteria apparently penetrate the intestinal epithelium of the neonate, enter the bloodstream, and are then translocated to the meninges. The majority of strains (>80 %) causing meningitis express the K 1 capsular antigen. However, the O-antigen also appears to play a role in virulence. *Escherichia coli* meningitis has a high attendant fatality rate.

Escherichia coli is the causative agent of ca. 90 % of nonobstructive urinary tract infections which are much more common among women than men. The clinical syndromes can include bacteriuria, cystitis, and pyelonephritis. Renal scarring associated with pyelonephritis occurs most often in children ≤5 years of age. The infecting organisms are thought to be acquired from the flora colonizing the vagina and periurethral areas. P and type I fimbriae are believed to play an important role by mediating tissue colonization [109].

History of Immunization. Initial attempts to vaccinate against *E. coli* diarrhea used crude cellular extracts derived from O types 111, 55, and 86 [110]. In one study, hospitalized infants up to one year of age received multiple doses of oral vaccine. Overall efficacy was 41 %. Efforts to develop vaccines against *E. coli* diarrhea center around the use of purified antigens or live-oral attenuated strains to elicit an antitoxic or antiadhesion response. Calves or neonatal piglets reared by mothers immunized with purified K 88, K 99, or 987 P fimbriae, procholeragnoid (heat-aggregated cholera toxin with reduced toxicity) or heat-labile enterotoxin are protected against *E. coli* bacillosis [111–113].

Graded doses (45 – 1800 µg) of purified type I fimbriae have been parenterally administered to human volunteers [114]. The vaccine was well tolerated and elicited both a humoral IgG and an intestinal secretory IgA antibody response. A moderate level of serospecific protein was obtained in volunteer challenge studies.

Oral administration of 1 mg of purified CFA/I evoked a significant secretory IgA response in about 50 % of vaccinated humans [115]. Responders were protected against challenge with an St^+, LT^+, CFA/I^+ *E. coli* strain.

Future Prospects. No vaccines are available to prevent diarrhea urinary tract infections, or meningitis caused by *E. coli*. At present, both attenuated live-oral and inactivated vaccines are being evaluated as to their ability to prevent *E. coli* diarrhea. The production of a killed, inactivated, oral, whole-cell vaccine from a CFA/I^+ strain has been described [115]. The majority of vaccinees responded with a good secretory IgA antibody response to CFA/I. The diarrhea attack rate for volunteers challenged with a toxinogenic CFA/I^+ *E. coli* strain was 89 % for those in the placebo group and 20 % for those in the vaccine group.

A spontaneous mutant of *E. coli* has been isolated which has lost the genes for production of the heat-labile and heat-stable enterotoxins but still expresses CFA/II. This strain has been

proposed as a likely carrier for cloned genes expressing relevant protective antigens such as colonization factors or "enterotoxoids" [116]. Several purified antigens show promise as oral vaccine candidates including conjugates of the two enterotoxins [117], the B subunit of the heat-labile toxin [118], and toxoid termed "procoligenoid" produced by heating the heat-labile enterotoxin [119].

Since the majority of *E. coli* strains associated with neonatal meningitis possess the K 1 capsule antigen, this would appear to be an excellent vaccine candidate. However, the K 1 antigen (which is identical to the group B meningococcal capsule) is not immunogenic in humans (see Section 2.9).

The most promising vaccines to prevent serious *E. coli* urinary tract infections (pyelonephritis) contain type I fimbrial antigens. Two fimbrial amino acid sequences have been shown to prevent colonization by a homologous *E. coli* strain in a murine pyelonephritis model [120]. Antibody to one sequence also afforded protection against an *E. coli* strain expressing a distinct fimbrial serotype.

2.12. *Neisseria gonorrhoeae* Vaccine

Etiological Agent and Pathogenesis. *Neisseria gonorrhoeae* was identified as the causative agent of gonorrhea by Bumm in 1885. It is a gram-negative diplococcus with humans serving as its only known natural host. Gonorrhea is a sexually transmitted disease of epidemic proportions worldwide. More than 500 000 cases are reported annually in the United States. The infection is usually restricted to the entry site, i.e., the genitalia, pharynx, and/or rectum. More than 90 % of infected males display symptoms characterized by pain upon urination and an urethral exudate [121]. However, up to 50 % of infected women remain asymptomatic and can serve as carriers. In approximately 10 % of women with genital gonorrhea, the bacteria invade the bloodstream and result in pelvic inflammatory disease. Disseminated gonococcal infections are far less frequent in male patients.

When cultured on solid medium, four colonial variants (T 1 – T 4) can be discerned [122]; however, only the piliated T 1 and T 2 variants can cause disease [123]. The first step in the infectious process is pili-mediated attachment of the gonococcus to mucus-secreting epithelial cells. Shortly thereafter, the bacteria are internalized and enter the submucosa [121]. Treatment of gonorrhea has been complicated by the occurrence of multiple antibiotic-resistant clinical isolates.

History of Immunization. The first attempt to prevent gonorrhea by immunization used a killed whole-cell vaccine administered parenterally [124], but did not provide protection in Eskimos. Subsequent human vaccine trials have used purified intact pili: antipili antibody blocks attachment of the gonococcus to epithelial cells, and, an antipili secretory IgA antibody response is mounted during natural infection [125]. Vaccination of human volunteers with purified *N. gonorrhoeae* pili elicited a humoral and a local vaginal antibody response. These antibodies blocked the attachment of *N. gonorrhoeae* to epithelial cells in vitro [125]. Two 2 mg doses of a monovalent pili vaccine resulted in significant protection against challenge with the homologous strain [126]. However, no protection was observed when the challenge strain possessed a pilus serotype modestly cross-reactive with the vaccine [121]. A subsequent field trial with this monovalent vaccine failed to show any protection [125].

Future Prospects. No vaccine has been licensed for the prevention of gonorrhea. The above human volunteer studies showed that protection against gonorrhea mediated by antipili antibody is serospecific [125]. Two different approaches are being taken to circumvent the problem of antigenic variation among pili:

1) Use of a polyvalent vaccine containing those types of pili which predominate in a given area [126].
2) Use of conserved antigenic domains from the antigenic pili molecule. A cyanogen bromide cleavage fragment can induce production of an antibody that can recognize various pili types [127]. Peptides corresponding to this conserved region block attachment of the gonococcus to epithelial cells in vitro [128]. If such peptides can induce a broadly protective immune response in

humans, synthetic peptides containing the proper sequence could be coupled with carrier proteins to produce a conjugate vaccine.

A possible alternative to pili as vaccine antigens are the outer membrane proteins of *N. gonorrhoeae*, the most promising being the PI protein. The PI protein has a sufficiently restricted antigenic variation to make construction of a polyvalent vaccine feasible. Furthermore, antiPI antibody is bactericidal and may block entry of the gonococcus into epithelial cells [129]. A parenterally administered, outer membrane preparation enriched with PI protein elicits an antibody response in humans [130].

2.13. *Hemophilus influenzae* Type b Vaccines

Etiological Agent and Pathogenesis. *Hemophilus influenzae* is a gram-negative bacillus first isolated by Pfeiffer in 1892. The vast majority (>90 %) of human disease is caused by type b capsular serotype organisms in children between the ages of three months to two years [131]. *Hemophilus influenzae* type b can cause several serious disease syndromes predominated by meningitis which is endemic worldwide and is one of the three leading causes of bacterial meningitis in developed countries. Infection is by inhalation of contaminated aerosols. After localized multiplication, the organisms enter the bloodstream and translocate to the meninges. Even though effective antibiotic treatment against *H. influenzae* type b has been available for many years, meningitis due to this organism still has a high attendant morbidity and mortality rate.

History of Immunization. Intensive efforts to understand human immunity to *H. influenzae* type b disease have largely replaced clinical trials with crude, poorly characterized vaccines. The capsular polysaccharide is a critical virulence factor because it confers resistance to lysis by serum [132]. Antibody-dependent, complement-mediated bacteriolysis is the basis for immunity against *H. influenzae* type b [132]. Antibodies directed against the type b capsular polysaccharide are bactericidal in the presence of complement and are believed to be the predominant protective antibody [131].

Production and Properties of *H. influenzae* Type b Vaccines. Two vaccines are licensed for use against *H. influenzae* type b: a purified capsular polysaccharide and a conjugate vaccine composed of capsular polysaccharide covalently coupled to diphtheria toxoid. The capsular antigen is purified by coprecipitation of the capsule with detergent, digestion with nucleases and protease, extraction in phenol, and ultracentrifugation. In the conjugate vaccine, the capsule is covalently linked to diphtheria toxoid by a bifunctional spacer molecule. The potency of *H. influenzae* type b vaccines is not determined in animals but by analysis of physicochemical characteristics which confirm that the capsular antigen is in an immunogenic form.

Immunization Recommendations. The capsular polysaccharide vaccine was first licensed for use in the United States in 1985. A single 25 µg dose, administerd parenterally is recommended for children 2 – 6 years of age; no booster dose is needed.

For the conjugate vaccine licensed in 1987, a single dose is recommended for children 18 months to 6 years of age; no booster dose is needed.

Vaccine Efficacy. The purified capsular polysaccharide vaccine was evaluated in a trial on Finnish children [133]. Vaccine efficacy was approximately 90 % for children who were vaccinated at 2 – 5 years of age or older. Vaccination provided no protection in children 3 – 18 months of age due to poor immunogenicity [132]. However, retrospective studies performed in the United States have found vaccine efficacy to range from 55 – 88 % [133]. In contrast, a capsular polysaccharide – diphtheria toxoid conjugate vaccine evoked protective levels of antibody in a majority of 18-month old children [135].

Future Prospects. A *H. influenzae* type b vaccine is clearly needed which is immunogenic in children <18 months of age, where approximately 75 % of serious disease occurs. Several candidate vaccines including the diphtheria toxoid conjugate mentioned above and a *Neisseria meningitides* outer membrane protein – capsular

polysaccharide conjugate are being evaluated in clinical trials [135], [136]. A large-scale Finnish trial is being performed to evaluate the diphtheria toxoid conjugate vaccine. Vaccine efficacy was 83 % in children vaccinated at 3, 4, 6, and 14 months of age [137]. In contrast, this vaccine afforded insignificant protection when used to immunize Alaskan Eskimo children at 2, 3, and 4 months of age.

3. Viral Vaccines

3.1. Measles Vaccine

Etiological Agent and Pathogenesis. Measles (rubeola) is caused by a paramyxovirus first isolated by Enders and Peebles in 1954. This highly contagious disease is transmitted in an aerosol of virus-infected droplets from the respiratory tract. During the prevaccination era, children in industrialized countries generally had a self-limiting disease characterized by conjunctivitis, bronchitis, fever, and a rash; complications included pneumonia (mostly bacterial) and encephalitis. A late complication of measles in early childhood is the development of a progressive, fatal neurological disease, termed subacute sclerosing panencephalitis. Death due to complications following measles infection remains one of the leading causes of child mortality in developing countries.

History of Immunization. Two approaches to vaccine development were adopted, one was based on killed (inactivated) strains and the other on live, attenuated strains [138]. The formaldehyde-inactivated, alum-precipitated vaccine was licensed in the United States in 1963 and withdrawn in 1967 because its protective effect was insufficient. In addition, the vaccinees exposed to the wild virus developed atypical measles with fever and rash but also pneumonitis.

The attenuated live vaccine with the Edmonston B strain was also licensed in 1963 in the United States. A large-scale field trial in the United Kingdom showed high seroconversion rates, a long antibody response, and a 84 – 94 % decrease in attack rate among the 36 000 immunized children. Further attenuation of the vaccine strain in chick embryo cells yielded the Schwartz and the Moraten strains. Both strains are included in current measles (and combined measles – mumps – rubella) vaccines. Attenuation of the original Edmonston strain in human diploid cells yielded the Edmonston-Zagreb strain, also in current use [139]. With the introduction of general immunization against measles, the disease has almost disappeared in many countries (see below).

Production and Properties of Measles Vaccines. The attenuated measles vaccine strain is usually grown in primary culture using chick embryo cells or human diploid cells. The WHO test requirements for the seed virus and cell substrates are described in[140, pp. 52 – 73],[141, p. 179]. Tests for monitoring measles vaccine grown in chick embryo cells include tests for nonadsorbing viruses and avian leukosis viruses. For vaccines grown in human diploid cells, chromosomal monitoring requirements have been formulated. Potency is checked by titration of the virus in tissue culture, (minimum requirement, 1000 $TCID_{50}$ per human dose). To test for stability, a sample of the final freeze-dried vaccine is incubated at 37 °C for seven days: the sample must then contain at least 1000 $TCID_{50}$ in each human dose. The vaccine is lyophilized; reconstituted vaccine should be used immediately or stored at 0 – 10 °C for not more than 8 h.

Immunization Recommendations. The vaccine is given at the age of 15 – 24 months [142] or even younger in developing countries [139]. In the United States and many European countries, a combined measles – mumps – rubella (MMR) is usually given (see Section 3.4). As for all live vaccines, immunization of immuncompromised individuals is not generally recommended. Immunization of pregnant women should be avoided although no increased risks have been documented for measles vaccine.

Adverse Reactions. Current attenuated measles strains give a few mild reactions, mostly consisting of fever and rash 7 – 10 days after immunization. The neurologic reaction of the Guillian-Barré syndrome has been reported but is extremely rare.

Vaccine Efficacy. The rapid decrease of measles morbidity in the United States after introduction of general immunization is a good example of vaccine efficacy [142], [143]. Over a period of 20 years the rate of notified cases fell by more than 99 %. Subacute sclerosing panencephalitis also decreased. The overall protective efficacy of the current strains is 90 % but is lower in children of 12 months of age due to the presence of maternal antibody. Life-long immunity after one dose of vaccine has not been proven. Increased disease incidence in the United States has raised the question of the possible need for a two-dose schedule [142].

Future Prospects. Given the efficacy and acceptability of the currently used attenuated measles strains, there has been little incentive for vaccine development. However, current vaccines are not very stable at high temperature, which is a concern in many developing countries. Furthermore, a large proportion of morbidity and mortality in children in developing countries occurs before the recommended age for immunization. Research is centered on solving these two problems. More immunogenic strains and/or higher doses of attenuated vaccines are being investigated [139]. Another approach would be to renew work on the development of a killed (inactivated) vaccine. Inactivated vaccines have so far failed, this is probably the result of a lack of neutralizing antibodies to the viral fusion surface protein that is destroyed in the inactivation process [144]. Subunit vaccines based on the hemagglutinin and the fusion proteins of measles virus have successfully been tested in animals and may provide a vaccine especially suited for developing countries [145].

3.2. Mumps Vaccine

Etiological Agent and Pathogenesis. Mumps is caused by a member of the paramyxovirus group and is also known as (epidemic) parotitis due to its predominant symptom–infection of the salivary (parotid) gland. Transmission occurs via infected aerosol droplets. The disease is generally characterized by moderate fever and swelling of the salivary glands. Subclinical infection occurs in ca. one third of infected individuals. Most cases occur in children 5 – 10 years of age. The most frequent complication is meningoencephalitis, giving clinical symptoms in >10 % of patients. Orchitis, a less common but feared complication in postpubertal males, results in impairment of fertility in 10 – 20 % of cases but absolute sterility is rare.

History of Immunization. Isolation of the causative agent for mumps by Habel in 1945 from chick embryo was followed by attenuation studies for vaccine development. Development of an inactivated (killed) vaccine, was also initiated; a killed vaccine was licensed in the US in 1950 – 1978 but had low long-term protective efficacy [146]. An inactivated mumps vaccine was also produced in Finland for immunizing military recruits; it decreased disease incidence by 94 % [147]. Immunization with an attenuated mumps vaccine (Jeryl Lynn strain cultured in chick embryonic cells), licensed in 1967 in the United States resulted in a 97 % decline of disease incidence in the general population by 1981 [146], [148], [149]. Chick embryonic cells are used to propagate another highly attenuated strain, Urabe Am 9 developed in Japan [150]. The Rubini strain, developed in Switzerland, is used in a vaccine in which the virus is grown in human diploid cells [151]. Vaccines based on the Leningrad-3 strain have been used in the former Soviet Union and other countries since 1974.

Production and Characterization of Mumps Vaccines. The WHO requirements for the above-mentioned live mumps vaccine have been formulated [152, pp. 139 – 164]. Controls, thermostability tests, and requirements are as for measles vaccines, i.e., the minimum potency should be retained after inoculation for one week at 37 °C (see Section 3.1). No minimum requirement of potency has been established by the WHO but a commonly used minimum dosage is 5000 $TCID_{50}$ per human dose. The vaccine is lyophilized, reconstituted vaccine should be used without delay.

Immunization Recommendations. Mumps vaccine is recommended to be given to all susceptible individuals over the age of 12 months [153]. It is usually given in combination with measles and rubella vaccines in industrialized countries (see Section 3.4).

Vaccine Efficacy. Clinical efficacy is 75–90 % for the live vaccine containing the Jeryl Lynn strain [149]. Similar results have been shown for the other strains or inferred from serologic comparisons. The long-term protective efficacy of the inactivated mumps vaccine used in Finland has not been established but antibody responses indicate that it may give a shorter-term immunity than a live vaccine due to a lack of antibody response to the fusion protein [154] as in the case of the inactivated measles vaccine (see Section 3.1).

Adverse Reactions. Mumps vaccine is one of the least reactogenic attenuated vaccines. Side effects are rare and mild. No cases of atypical mumps were found after immunization of Finnish recruits with killed mumps vaccine [154].

Future Prospects. The present vaccine is highly satisfactory and little effort has been devoted to further development. Subunit vaccines based on the hemagglutinin neuraminidase and the fusion surface proteins of mumps virus are protective in animal models [155].

3.3. Rubella Vaccine

Etiological Agent and Pathogenesis. Rubella, also known as German measles, is caused by a member of the togavirus group. The virus was isolated in 1962 by two independent American groups. In 1941 Gregg reported that this mild disease could cause severe congenital cataracts in children born to mothers who were infected during the first trimester of pregnancy. Transmission of the virus occurs by the respiratory route. The clinical picture is a discrete rash and low-grade fever, 25–50 % of cases remain subclinical. Arthritis is a common complication; encephalitis is less common than in measles or varicellae (1/6000 cases). The congenital rubella syndrome is usually characterized by hearing impairment, ocular lesions, cardiac malformation, microcephaly, and mental retardation. The severity of defects is related to fetal age at the time of maternal infection, with the most severe damage seen after infection during the first month of pregnancy.

History of Immunization. Live attenuated rubella vaccines were licensed in several countries in 1969–1970. One of the first vaccines contained the Cendehill strain, grown in rabbit kidney cells [156]. Other vaccines were based on the HPV-77 strain, grown in either duck embryo cultures or in dog kidney cultures. The latter vaccine produced arthritis symptoms; 50–60 % of adult female vaccinees as compared to 10–20 % for the other two vaccines. A vaccine based on the RA 27/3 strain, isolated and propagated in human diploid cells, was licensed in the mid-1970s in Europe and in 1979 in the USA [157]. This vaccine gave a better immune response and had less side effects than the other vaccines; it has replaced other rubella strains in the current vaccines.

Production and Properties of Rubella Vaccines. The attenuated RA 27/3 strain is grown in human diploid cells. The WHO requirements for rubella vaccine include control of normal karology of the human diploid cells[158, pp. 54–84], [159, pp. 313–316]. The minimal potency requirement, determined by titration in tissue culture, is 1000 $TCID_{50}$ per human dose. The vaccine is lyophilized; reconstituted vaccine should be used immediately or stored at 2–8 °C for not more than 8 h.

Immunization Recommendations. Vaccination strategies range from protection of the individual as in the U.K. to indirect protection as in the US [160], [161]. Protection is achieved by general immunization of adolescent girls, usually combined with post partum vaccination of seronegative women. Indirect protection of adult women by immunization of young children is often combined with vaccination of seronegative women.

A combination of these strategies is used in some European countries, e.g., Sweden [162]. This program includes routine screening of all pregnant women, post partum vaccination of seronegatives, and two-dose immunization of children with combined measles–mumps–rubella vaccine. Although the risks for the fetus seem small, the vaccine should not be given to pregnant women and contraceptive measures are recommended for three months after immunization.

Adverse Reactions. The RA 27/3 strain has few and mild side effects. The joint manifestations, mainly arthralgia, are more common in adults. Intrauterine infections during pregnancy have been documented for all three strains (Cendehill, HPV-77, and RA 27/3) [163]. No abnormalities related to congenital rubella infections have been documented for any of the strains.

Vaccine Efficacy. The efficacy of the current strain is ca. 90 %. The incidence of congenital rubella virus infection decreased in countries with a routine immunization program against the disease. The duration of immunity has not yet been determined. The risks to the fetus upon a maternal reinfection are also unknown and may be very small.

Future Prospects. The current vaccine is considered safe and efficient and no efforts are devoted to further development.

3.4. Combined Measles – Mumps – Rubella Vaccine

History of Immunization. The first combined measles – mumps – rubella (MMR) vaccine was licensed in the US in 1971. It contains the Moraten measles strain (Section 3.1), the Jeryl Lynn mumps strain (Section 3.2), both grown in chick embryo cultures, and the RA 27/3 rubella strain (Section 3.3), grown in human diploid cells. Three other combined vaccines were licensed later. The triple vaccine has replaced the monovalent vaccines used in general immunization programs for children in the United States and in many European countries.

Production and Properties of Combined Vaccines. The minimum potencies for the attenuated strains of measles, mumps, and rubella virus, are the same as in the monovalent vaccines, i.e., 1000 $TCID_{50}$ for measles, >5000 $TCID_{50}$ for mumps, and >1000 $TCID_{50}$ for rubella. The vaccine is lyophilized and should be stored at 2 – 8 °C.

Immunization Recommendations. In many countries, one dose of vaccine is administered to children at the age of 15 – 24 months. Two doses are given in some European countries [162]. In Sweden, a first dose is administered at 18 months and a second at 12 years of age, replacing the monovalent programs for measles and rubella immunization respectively. In the United States, the combination of MMR with oral polio vaccine and diphteria – tetanus – pertussis (DTP) at 15 months has been recommended [164], [165].

Adverse Reactions. The side effects are the same as for the monovalent vaccines, the measles component being the most reactogenic. The rate of adverse reactions is <0.5 – 4 % [166].

Vaccine Efficacy and Future Prospects. The efficacy of the combined preparation is the same as for the monovalent vaccines with an overall efficacy of 90 – 95 % [143]. The life-long protective efficacy of a single injection has not been proven. Further combinations with an attenuated varicella component were investigated [167].

3.5. Polio Vaccine

Etiological Agent and Pathogenesis. Polio (infantile paralysis) is caused by a picornavirus of the genus enteroviruses. A poliovirus strain was first isolated in cell culture by Enders, Weller, and Robbins in 1949. In 1951, polio virus isolates were officially grouped into three serotypes, type 1 (Brunhilde), type 2 (Lansing), and type 3 (Leon). Transmission is mainly via the oral – fecal route and the virus reaches the central nervous system by way of the blood stream. The vast majority of infections are subclinical. The nonparalytic disease has a mild or minor form with fever and general malaise and a more severe or major form with additional symptoms of meningitis/meningoencephalitis. Paralytic polio, with its most severe bulbar form, is estimated to represent 5 – 10 % of clinical cases. The mortality rate is 5 – 10 % of clinical cases, i.e., 1 – 2 % of all infections. Severe paralysis is seen in 10 – 20 % of clinical cases, and mild or moderate paralysis in 30 %.

With improved sanitation and standards of living, a shift towards infection at a higher age was observed in industrialized countries in the prevaccination era. During the first half of

this century, peak incidence was noted in the age group 5 – 14 years. A large proportion of cases occurred in young adults. In the developing countries, polio has maintained its character of infantile paralysis with the majority of children being infected during the first few years of life. Cases of paralytic polio occur only in the youngest age groups.

History of Immunization. The first large-scale field trial of an inactivated (killed) polio vaccine was launched in the US in the early 1950s, only a few years after successful propagation of the virus in tissue culture. Inactivated polio vaccine (IPV) was licensed for general use in the United States in 1955 [168], [169]. In 1955, cases of atypical paralytic polio were reported, most of them were associated with two lots of vaccine from CUTTER [170]. Although live poliovirus was recovered from vaccine supplied by other manufacturers, this failure of inactivation is known as the Cutter incident. The total toll was 269 cases, of which 192 were paralytic (with ten deaths). Clinical trials with a live attenuated oral polio vaccine (OPV) were started in 1958 and the vaccine was licensed in the United States in 1962 [171], [172]. Immunization with OPV alone has been used from the late 1950s – early 1960s in Europe. One of the largest immunization campaigns was launched in the former Soviet Union in 1960 when OPV was given to ca. 77×10^6 people. A few European countries have used only IPV [173–175]. In Sweden, clinical trials with IPV were started in 1955 and general immunization was introduced in 1957 [173]. All three strategies were effective in reducing the incidence of paralytic polio. With IPV alone, an 80 % reduction rate was achieved in the US with less than 50 % of the population immunized [168], [169], [171]. The decrease of polio continued at the same rate after introduction of OPV. Introduction of the massive IPV campaign in Sweden resulted in a 93 % decrease of paralytic disease after six years of immunization and elimination of the disease by 1964 [173].

Production and Properties of Polio Vaccines. Oral polio vaccines are manufactured with the attenuated Sabin strains. Inactivated polio vaccines are mostly produced with the type 1 Mahoney strain, type 2 MEF-1 strain, and type 3 Saukett strain. Primary and secondary monkey (Cynomolgus) kidney cultures or human diploid cells are most commonly used for culture. The WHO requirements for both OPV and IPV produced in primary monkey kidney cells include tests

1) for the absence of cytopathogenic viruses with the exception of some foamy viruses[141, pp. 40 – 84] , [159, pp. 157 – 173], [176],[182, pp. 108 – 110]
2) the absence of simian B virus in rabbits, and
3) the absence of SV 40 in sensitive cells.

(usually primary green monkey *Cercopitecus* kidney cultures). Minimal potency requirements for OPV were formulated by the WHO in 1987[152, pp. 165 – 166]. A single human dose of trivalent oral vaccine should contain approximately $10^{5.5} - 10^{6.5}$ infectious units of type 1, $10^{4.5} - 10^{5.5}$ of type 2, and $10^{5.0} - 10^{6.0}$ of type 3. For IPV, recommendations of 40 : 8 : 32 D-antigen units per human dose for types 1, 2, and 3 respectively were issued in 1981 [141]. The potency tests for OPV (and IPV prior to inactivation) are performed by titration in tissue culture. In vivo potency tests are also required for IPV but neither the animal species nor the number of injections is specified. Inactivation of polio vaccine by formaldehyde is controlled by titration of polio-sensitive cells (usually *Cynomolgus* or *Cercopitecus*) in tissue culture. Oral polio vaccines are supplemented with a stabilizer, usually magnesium chloride or sorbitol. Both OPV and IPV contain antibiotics, usually neomycin; OPV is best stored at −20 °C but can be stored at 2 – 4 °C for a variable length of time; IPV is stored at 2 – 8 °C.

Immunization Recommendations. Recommendations vary but at least three doses of OPV are given. In the US, five doses are recommended, in Sweden four, and in Finland six. The WHO recommends four doses for infants in developing countries.

Adverse Reactions. Serious adverse reactions of paralytic polio have only been reported for OPV with a predominance for type 3 [169], [177] (incidence = $1/1 \times 10^6$ vaccine recipients, including both nonimmune and immune individuals). No cases of IPV-induced paralytic polio have been reported after 30 years of use (with

the exception of the Cutter incident). IPV is considered to be the least reactogenic of all the vaccines used for childhood immunization.

Vaccine Efficacy. Both OPV and IPV have proven highly efficient in eliminating polio from industrialized countries with general immunization programs. The protective efficacy is close to 100 % after at least three doses of vaccine with adequate immunogenicity. Serologic data reported for OPV from developing countries have been less than encouraging. Both vaccines have eliminated the circulation of wild type strains in industrialized countries with high immunization rates. Outbreaks among groups refusing immunization have occurred [168]. In 1984, nine cases of paralytic polio caused by type 3 occurred in Finland [177]. The low immunogenicity of the IPV used in Finland has been known since the late 1960s [178]. The new IPV developed in the Netherlands with its high antigen content is given in European countries in at least three doses [179]. Two doses or even one dose of the new IPV preparations have been claimed to be highly protective. A two dose-schedule has been used in Africa and protective efficiency was ca. 89 % [180]. Inactivated polio vaccine is more heat-stable and can be combined with diphteria – tetanus – pertussis vaccines.

Future Prospects. Further development of OPV is aimed at reducing production costs by use of microcarrier cultures (cell cultures on beads etc.) in fermentors [179]. Use of continuous cell lines is being investigated. An IPV produced in a continuous cell line is currently licensed in France [181]. The WHO requirements for production of vaccines in continuous cell lines are formulated in [182, pp. 93 – 107].

3.6. Hepatitis B Vaccine

Etiological Agent and Pathogenesis. Hepatitis B is caused by the hepatitis B virus (HBV), a member of the hepadna virus group. The presence of a new antigen called the Australia (Au) antigen or hepatitis B surface antigen (HBsAg) in the blood of patients with leukemia, Down's syndrome, and hepatitis was shown by Blumberg in 1964 – 67. Other antigens (HBeAg and HBcAg) were described later and correlated with infectivity; they are found in the 42-nm Dane particle, i.e., the infectious virion.

The HBV is a pathogen only for humans, but certain monkeys, in particular chimpanzees, are also susceptible to infection. Transmission in humans occurs mainly by inoculation with infected blood. The virus is excreted in body fluids, causing infection through saliva and sexual contact. A chronic carrier state is established in 5 – 10 % of adult cases; chronic active hepatitis develop in 25 – 30 % of the carriers, often leading to cirrhosis of the liver: 90 % of infants infected at birth become carriers. Hepatitis B occurs throughout the world; the incidence varies from <0.5 % in Western Europe and the US to 5 – 15 % in Southeast Asia and Southern Africa [183], [184].

Hepatitis B virus plays an important role in hepatocellular carcinoma although the exact mechanism has not yet been determined [183].

History of Immunization. Heat-inactivated HBsAg-positive serum was used by Krugman in 1971 to immunize children. A 70 % protection rate was found upon challenge with active virus. The high protective efficacy of hepatitis B vaccine derived from purified, inactivated HBsAg from positive plasma was demonstrated in 1980 [185]. The vaccine, developed in the US, was licensed for general use in 1981. Studies in 1980 – 1983 also showed that the HBV vaccine in combination with human antiHBV immune globulin (HBIG) prevented development of the carrier state in infants born to carrier mothers. Hepatitis B vaccine was the first vaccine to be routinely produced by DNA recombinant technology in yeast (→ Genetic Engineering, Chap. 5.2.). Serologic studies indicate that the protective efficacy of the recombinant vaccines is the same as that of the plasma-derived vaccine. Genetically engineered vaccines were licensed in the US in 1986 and subsequently in several European countries.

Production and Properties of Hepatitis B Vaccines. The WHO requirements for plasma-derived HBV specify guidelines for selection of donors of HBsAg-positive plasma[186, pp. 70 – 101]. The plasma pool must be subjected to extensive tests in animals, fertile eggs and cell cultures for extraneous viruses and *Mycobacterium tuberculosis*. The purified HBsAg (>95 % pure) is inactivated by treatment with pepsin followed

by urea (8 mol/L), and formaldehyde or by heat with or without formaldehyde. After controls for purity and antigen content, an alum adjuvant is added; the vaccine must be stored at $5 \pm 3\,^{\circ}C$. No minimal requirement of antigen content has been formulated but a commonly used vaccine contains 20 µg HBsAg per human dose.

The WHO requirements for HBV made by recombinant DNA technology state that such vaccines may contain the S gene products or the S/pre-S combination [152, pp. 106 – 138]; licensed recombinant vaccines contain only the S gene product. A full description of the host cell (*Saccharomyces cerevisiae*) and the expression vector is required.

The HBsAg is commonly purified by precipitation, ultrafiltration, and chromatography. Tests for HBsAg and residual cell or plasmid DNA are required before addition of adjuvant. No minimal antigen requirements have been formulated but a potency assay in mice has been outlined. Commonly used recombinant yeast vaccines contain 10 or 20 µg HBsAg per human dose.

Immunization Recommendations. Most Western European countries and the US have issued recommendations for vaccination of risk groups [187–189]. In general, pre-exposure prophylaxis is recommended for medical and other staff in frequent contact with high-risk groups or with blood from such groups. Immunization of patient groups frequently receiving blood or blood products, as well as patients in certain institutions is also recommended.

For pre-exposure prophylaxis, three 20 µg doses of plasma-derived vaccine is given at day 0, 30 and 6 months intramuscularly in the deltoid region. Children <11 years of age should receive 5 µg at each injection. Postexposure prophylaxis for infants to HBsAg, HBeAg-positive mothers is 0.5 mL intramuscular injection of human antiHBV immune globulin (HBIG, see also Section 5.3). Vaccine (10 µg) should be given at birth and then at one and six months of age. For postexposure prophylaxis of adults, 1 mL HBIG should be given immediately together with three doses of vaccine administrated as for preexposure prophylaxis.

Adverse Reactions. Side effects are mild and mainly local at the site of injection.

Vaccine Efficacy. The overall vaccine efficacy is 90 % with no response upon immunization in 5 – 10 % of the vaccines. In individuals with seroconversion to HBsAg, protective efficacy is almost 100 %. Studies on the duration of protection beyond 5 – 10 years are not yet available. The protective efficacy of HBIG and vaccine administered to newborns born to carrier mothers has been estimated to be more than 90 % [183].

Future Prospects. Current HBV vaccines are safe and have a high efficacy but also high production costs. Research is aimed at the development of cheaper, more immunogenic vaccines. The inclusion of the pre-S (P 31) region of the viral genome in addition to the present S (P 25) region is being investigated in yeast-derived recombinant vaccines [190], [191]. Other antigens such as HBcAg are also being studied. Polypeptide vaccines, based on the major determinants of HBsAg, are undergoing clinical trials. Synthetic peptides may provide the ultimate solution to the problem of immunogenic, readily available HBV vaccines.

3.7. Rabies Vaccine

Etiological Agent and Pathogenesis. Rabies is a lethal disease caused by a neutropic rhabdovirus affecting humans and warm-blooded animals. The most common route of transmission to humans is by infected saliva through the bite from a rabid animal. The virus ascends along peripheral nerves to the central nervous system. The incubation period varies from ten days to several months. Symptoms include a nonspecific prodromal stage followed by an acute neurological phase ending in coma and death. The disease is endemic in animals in most parts of the world with some exceptions such as the United Kingdom and parts of Scandinavia [192]. In the urban form, stray dogs and cats act as vectors, whereas the sylvatic form involves wild animals. Urban rabies is well-controlled in most countries, but progressive spread of the sylvatic form by foxes is a major problem in Europe. In the US, other animal species are involved including bats. In Europe and Africa, bats are the vector for an antigenic variant of the virus [193].

History of Immunization. In 1885 PASTEUR administered the first ever vaccination to a young boy who had been bitten by a rabid dog [194]. Serial injections containing a virus strain that had been propagated and attenuated in the spinal cord of rabbits ("fixed" virus) were given in an increasingly virulent form. The treatment was rapidly adopted as standard postexposure prophylaxis. The attenuated vaccine was later replaced by an inactivated (killed) vaccine also produced in neural tissue [194]. However, in some patients the myelin content of the vaccine caused sensitization resulting in neurological disease. An inactivated vaccine produced in duck embryos decreased the risk of neural tissue sensitization. The most widely used rabies vaccine in Europe and in the United States was developed in the 1960s and is produced in human diploid cells. A highly purified, concentrated duck embryo vaccine is also available. In Europe, other vaccines are produced in primary animal cells and continuous cell lines; non human primate diploid cells are used in the US.

Production and Properties of Rabies Vaccines. Attenuated strains of rabies virus are grown in tissue culture or embryonated duck eggs (see above). The virus suspension is usually concentrated and in some cases purified. Only inactivated vaccines are allowed for human use. Inactivation is mostly achieved by treatment with β-propiolactone, but phenol, formaldehyde, or ultraviolet irradiation are also used. The WHO has established requirements for controls of the cells used for virus propagation [152, pp. 167 – 194],[159, pp. 54 – 95]. The potency of the vaccine is determined in a mouse challenge assay (minimal requirement 2.5 IU per human dose). The vaccine is then usually lyophilized although adjuvanted preparations are also in use.

Immunization Recommendations. The recommendations vary by vaccine and country. As pre-exposure prophylaxis, three 1 mL doses are usually recommended on days 0, 7, and 28 or on days 0, 28, and 90 (or 365) as intramuscular injections in the deltoid or suprascapular region [192], [195]. As postexposure prophylaxis, recommended measures include local wound care, administration of vaccine, and administration of human rabies immune globulin (HRIG, 20 IU per kilogram of body weight; see also Section 5.3). Purified equine rabies immune globulin (40 IU/kg) is used in many developing countries with few side effects [196]. Postexposure vaccine prophylaxis is given in five 1 mL doses on days 0, 3, 7, 14, and 28 by intramuscular injection in previously unimmunized individuals. In individuals who have received pre-exposure prophylaxis, two doses are given on days 0 and 3. Children under four years of age receive 0.5 mL injections.

Vaccine Efficacy. The protective efficacy of the human diploid cell vaccine given (with HRIG) as postexposure prophylaxis was first documented in field trials conducted in Iran in 1974 – 75; the survival rate was 100 %. Other studies have confirmed the protective effect of pre- and postexposure immunization with this vaccine. The general opinion that postexposure vaccine prophylaxis in combination with immune globulin conveys 100 % protection has been challenged by two cases of vaccine failure [197]. In other vaccines, a high protective efficacy has been proven in field trials or inferred from comparative serologic studies.

Adverse Reactions. The most common side effects are local redness and induration at the injection site and fever. Desquiting episodes of a serum sickness-like reaction upon repeated immunization have been reported [198] and are possibly due to a sensitizing complex formation between β-propiolactone and human serum albumin [199] or to the high content of bovine serum residues in the vaccine [200].

Future Prospects. Work is centered on the development of large-scale vaccine production techniques to decrease the high costs and thereby increase availability in poorer countries. One such approach is the use of microcarrier culture systems in fermentors [179]. A vaccine produced in this manner in a continuous cell line has been evaluated. A subunit vaccine based on the surface glycoprotein of the virus is a further possibility. The peptide segment of the glycoprotein that induces the production of neutralizing antibodies has been identified and synthesized. A synthetic peptide vaccine may therefore be the most attractive future alternative [201].

3.8. Influenza Vaccine

Etiological Agent and Pathogenesis. Influenza is caused by two antigenically distinct members of the orthomyxovirus group–influenza viruses A and B. In 1934 ANDREWES managed to transfer the influenza A virus from human material to ferrets and later to mice. In 1940, FRANCIS and MAGILL independently isolated influenza B by transmission to ferrets. The virus was subsequently propagated in embryonated hen's eggs and tissue culture. The disease is transmitted by droplet infection from the respiratory tract; it is characterized by high fever, muscle pain, and a dry cough. Pneumonia is the main cause of mortality in elderly people and persons with chronic underlying diseases. Influenza occurs worldwide with regular epidemics during the winter months. The recurrent epidemics are caused by small changes (antigenic drift) in the main pathogenic determinants of the virus, i.e., hemagglutinin and neuraminidase. Large pandemics occur at 10 – 20 year intervals and are due to substantial changes (antigenic shift) in the pathogenic determinants. Antigenic drift is seen in both influenza A and B whereas antigenic shift has only been noted in influenza A [202].

History of Immunization. The first influenza vaccines produced in hens' eggs were tested in humans in the early 1940s. These inactivated *whole virus vaccines* had a low content of viral antigens and a high content of contaminating egg protein. Consequently, the protective effect was low and the rate of adverse reactions high. Whole virus vaccines purified by ultracentrifugation showed clearly improved immunogenicity and a decreased rate of side reactions. Zonal centrifugation further improved the vaccine for adults but still caused adverse reactions in children. Disruption of influenza vaccine with detergent was used to develop *split vaccines* in the mid 1960s. *Subunit vaccine* introduced in the mid 1970s contain concentrated, purified hemagglutinin and neuraminidase [202].

The development of live, attenuated virus strains was pursued in the former Soviet Union by serial passages in hens' eggs [202]. However, doubt was cast on their stability. More stable attenuated vaccines were obtained by isolation of temperature-sensitive and cold-adopted mutants [203]. These strains were successfully tested in humans in the late 1970s but are not in general use. Attenuated vaccines, based on avian – human recombinant strains, are being evaluated in clinical trials.

Production and Properties of Influenza Vaccines. The most commonly used influenza vaccines are of the whole or split virus type. The strains used are specified in annual WHO recommendations. The virus is grown in embryonated hens' eggs (usually pathogen-free) and the allantoic fluid is harvested. According to the WHO requirements, the inactivation method used (usually addition of formaldehyde or β-propiolactone) should inactivate avian leukosis viruses and mycoplasma[204, pp. 148 – 170]. The virus is concentrated and purified by high-speed centrifugation either before or after inactivation. Effective inactivation is controlled by inoculation of embryonated hens' eggs. Hemagglutinin content is usually checked by single radial immunodiffusion against a WHO standard. Whole virus and split virus vaccines usually contain 10 – 15 µg hemagglutinin per human dose. The inactivated vaccines are usually supplemented with a preservative and stored at 2 – 8 °C.

The WHO requirements for attenuated (live) vaccines stipulate that pathogen-free eggs must be used; the absence of other pathogens must be controlled in tissue cultures and animals[204, pp. 171 – 194]. Vaccine potency is determined by titration in embryonated eggs. No minimal requirements of infective dose have been formulated.

Immunization Recommendations. Most countries have recommendations for yearly immunization of high-risk groups. Some countries and the WHO recommend vaccination of all individuals over 65 years. In the United States, immunization of children with chronic pulmonary or cardiac disorders or with other chronic diseases residing in institutional care is recommended; 0.25 and 0.5 mL of split virus vaccine are given in the age groups 6 – 35 months and 3 – 12 years, respectively [205]. For individuals older than 12 years, 0.5 mL of vaccine is recommended.

Adverse Reactions. The incidence of systemic (febrile) and local reactions is <10 % and <20 % for the whole virus vaccines and split vaccines, respectively. An increased rate of the neurological reaction of Guillain – Barre's syndrome was reported in the United States after large-scale immunization against swine influenza in 1976.

Protective Efficacy. The protective efficacy of influenza vaccines is controversial. Discrepancies are probably due to different vaccines, number of injections given, the age groups under study, and occurrence of antigenic drift during the study period. The newer purified whole virus and split virus vaccines are considered to be 70 – 90 % effective in healthy adults. The duration of protection is largely dependent on the degree of antigenic drift. In the case of antigenic shift, little or no protection can be expected. The immune response and protective efficacy in children receiving chemotherapy and in debilitated elderly persons are lower [206]. The protective efficacy in preventing death has been estimated at 74 % [207].

Future Prospects. Research on influenza vaccines is directed toward the synthesis of the hemagglutinin and the neuraminidase antigens and to improvement of their immunogenicity [201]. Another line of development is the attenuated live vaccine approach [203].

3.9. Varicella Vaccine

Etiological Agent and Pathogenesis. The varicella-zoster virus, a member of the herpes virus group, causes two distinct clinical manifestations [208]–varicella and herpes zoster. *Varicella,* also known as chickenpox, is the primary infection. *Herpes zoster* is caused by the reactivation of the latent varicella virus in ganglion tissue. Varicella is usually a mild infection characterized by a vesicular rash. Transmission occurs by droplet infection from the respiratory tract and by direct contact with vesicle fluid. Complications are encephalitis (1/1000 – 1/3000 cases) and pneumonia. The disease occurs worldwide, with 90 – 95 % of cases occurring before the age of 15 years. Suspected cases of congenital varicella have been described mainly after varicella during the first trimester of pregnancy. Neonatal varicella with a 30 % mortality rate occurs with maternal varicella within one week prior to term without prophylaxis. Varicella in the compromised host is a severe disease with a 10 – 20 % mortality rate. Herpes zoster, a localized, often painful vesicular rash is most common in the elderly; it occurs in 10 % of the population.

History of Immunization. The varicella-zoster virus was first isolated and propagated in tissue culture by Weller and Stoddard in 1952. Attenuated, live varicella vaccine was developed in Japan in 1970 using the Oka strain [209]. After serial passages, the attenuated strain was adapted to human diploid cell cultures. The immunogenicity and protective efficacy of the vaccine in healthy children has been demonstrated [209], [210]. Efficacy of the vaccine in children with malignant disease, in particular leukemia, has also been shown [211]. The Oka strain is used for vaccine production in Japan, Europe, and the United States.

Production and Properties of Varicella Vaccines. The attenuated, live Oka strain is propagated in human diploid cell cultures. The WHO requirements include control for the absence of adventitious agents and the usual conditions for culture in human diploid cells [186, pp. 102 – 133]. No minimal potency requirements have yet been formulated. Varying doses have been used in clinical trials. The most common formulation is ≥2000 plaque-forming units per human dose. The vaccine is lyophilized; reconstituted vaccine should be used without delay.

Immunization Recommendations. Varicella vaccine is not yet recommended for general immunization in Europe or the United States. It is given to children with leukemia during remission or when chemotherapy is withheld for one week prior to and after vaccination [211]. A second dose is often given to children who remain seronegative after the first injection.

Adverse Reactions. In healthy children and adults, local swelling and pain occurred in 1 – 5 % of healthy children and 20 % of adults. Systemic reactions with fever and a rash are reported in 5 – 10 % of both healthy children and adults.

Fever and rash occurred in 40 % of vaccine recipients with chemotherapy suspended for two weeks.

Vaccine Efficacy. Protective efficacy in healthy children is 95 – 100 % with persistence of immunity over a 5 – 10 year period. In adults, protective efficacy is 60 – 80 % and possibly of shorter duration. In children with malignancies, protective efficacy is 60 – 92 %. Spread from vaccine-induced vesicular rash has been documented in household contacts.

Future Prospects. The currently investigated, attenuated Oka strain vaccine is effective but long-term immunity and zoster incidence remain to be established. The problems of latency and of vaccine-induced rash could be overcome by the development of a subunit vaccine produced either by a recombinant DNA technique or by using synthetic peptides.

3.10. Yellow Fever Vaccine

Etiological Agent and Pathogenesis. Yellow fever is a hemorrhagic fever caused by a member of the toga virus group [212], [213]. The virus was transmitted to rhesus monkeys by Mathis and coworkers in 1927. The virus strain was then propagated by serial passages of intracerebral inoculations in white mice. Yellow fever is transmitted by mosquitoes of the genus *Haemagogus* in South America and of the genus *Aedes* in Africa. The animal reservoir is mainly monkeys. Subclinical infections are common. The incubation period in clinical cases is 3 – 6 days; symptoms range from transient fever and headache to high fever with meningoencephalitis, followed by jaundice and hemorrhagic manifestations. In the malignant form all these symptoms are present and death occurs within one week. Mortality rates are 40 – 50 % in the severe forms of the disease.

History of Immunization. Two types of attenuated vaccine were developed in the early 1930s cultured in neural tissue (Dakar vaccine) and in chick embryo [214]. Initially, both types were given together with human immune serum. In the late 1930s, the vaccine cultured in neutral tissue was used for mass immunization in Senegal and the chick embryo vaccine was tested in Brazil. The vaccines were administered by scarification using normal human serum as stabilizer. The chick embryo cultured vaccine (first the 17 E vaccine and later the 17 D vaccine) has been most widely used. Numerous problems and accidents were associated with both vaccines. The human serum used as stabilizer caused hepatitis affecting many vaccinees in the armed forces during World War II. Systemic reactions were common. Severe postvaccination encephalitis with a high mortality rate occurred mainly with the neural tissue cultured vaccine. This reaction (mostly in children $<$ one year of age) was also reported with some lots of the 17 D vaccine with increased neurotropism; furthermore, loss of protective efficacy was noted in tropical climates due to low thermostability. The 17 D vaccine is the result of developments aimed at careful definition of the properties of the seed virus and at improved thermostability [215].

Production and Properties of Yellow Fever Vaccines. Only certain institutes are approved by the WHO for production of yellow fever vaccine[159, pp. 34 – 53]. The seed lot virus, usually a substrain of 17 D-204, has to be shown to be free from neurotropism by testing in monkeys[159, pp. 34 – 53],[182, pp. 113 – 141],[216] . Most producers use seed lots that are free of leukosis virus for production but their use is not mandatory. Virus-infected embryos are harvested, homogenized, and the supernatant is used as vaccine. Several tests for adventitious agents are performed. Virus titrations are performed in a mouse assay with a minimal potency requirement of 1000 LD_{50} per human dose. The vaccines are lyophilized in the presence of stabilizer. Most current vaccines retain the minimal requirement for two weeks at 22 °C. A vaccine stable for two weeks at 37 °C is requested by WHO for use in tropical areas.

Immunization Recommendations. Vaccination of visitors to endemic areas in equatorial Africa and northern parts of South America is recommended. Immunization is mandatory in several countries for visitors from endemic areas. One 0.5 mL injection is given to both adults and children for both primary and booster immu-

nization. Booster injections are recommended every ten years. The vaccine is not recommended for children under one year of age or for pregnant women.

Adverse Reactions. Yellow fever vaccines are safe and induce only minor local reactions. Transient headache can occur.

Protective Efficacy. Current vaccines are estimated to be 90 – 95 % protective. Immunization has decreased or eliminated the disease in many endemic areas.

3.11. Tick-Borne Encephalitis Vaccine

Etiological Agent and Pathogenesis. The tick-borne encephalitis virus, a member of the toga virus group, was first isolated in the former Soviet Union in 1937 [212]. Two antigenically distinct forms of the virus cause the disease in Europe and in the eastern former Soviet Union. The main vector for the European form is *Ixodes ricinus* and for the Eastern form *I. perulcatus*. Many wild and domestic animals can be infected; the main animal reservoirs are small mammals such as field mice. Transmission is usually by tick bite but infection can occur by drinking untreated cow milk. The incubation period is 7 – 14 days before onset of fever and malaise. After a 1 – 2 week recovery period, a second stage with fever and neurological symptoms can follow: meningitis (40 % of cases), meningoencephalitis (40 %), and severe meningoencephalomyelitis (20 %). Mortality rates are 1 – 2 % in the European form and 20 – 25 % in the Eastern form; neurological sequelae are seen in 15 – 40 % of cases. The disease is subclinical or abortive with only the first stage in 75 % of infections. Tick-borne encephalitis is endemic in Central Europe, in the Balkan countries, and in Finland and Sweden.

History of Immunization. The first inactivated (killed) vaccine was developed and used in humans in the former Soviet Union in 1939 followed by an attenuated, live vaccine in the 1960s [217]. An inactivated vaccine, developed in Europe with a virus strain isolated from a tick in Austria, was subjected to clinical trials in 1973 [218]. The vaccine used in Western Europe is a purified, concentrated version of this.

Production and Properties of Tick-Borne Encephalitis Vaccine. The inactivated vaccine is produced by propagation of the virus in hens' eggs, followed by purification by continuous flow zonal ultracentrifugation, and inactivation with formaldehyde. No WHO requirements have been formulated. The vaccine contains not less than 25 protective doses per human dose assayed in a mouse protection test. Human albumin is used as stabilizer and aluminum hydroxide as adjuvant.

Immunization Recommendations. In Austria and Bavaria, immunization of children older than one year is recommended. Other endemic countries recommend immunization of forest workers and other high-risk groups. Primary immunization consists of two 0.5 mL intramuscular injections at 1 – 3 month interval, followed by a third 0.5 mL dose 9 – 12 months after the second. Booster injections are recommended every three years.

Adverse Reactions. Local and systemic reactions are rare and mild. Low-grade fever is occasionally seen, mainly in children after the first injection.

Protective Efficacy. The vaccine is at least 95 % protective against all European virus strains in children and young adults. Seroconversion rates of about 90 % are reported for persons over 65 years of age.

3.12. Japanese Encephalitis Vaccine

Etiological Agent and Pathogenesis. The Japanese encephalitis virus, belonging to the toga virus group, was first isolated in 1935 in Japan [212]. The disease is transmitted by the mosquito *Culex tritaeniorhynchus*; the main animal reservoirs are pigs and birds. The incubation period is 5–15 days. The disease is subclinical in at least 95 % of infections. In clinical cases symptoms vary from mild febrile disease with headache to severe encephalitis. Paralytic forms more commonly affect the upper extremities. In endemic areas the disease affects mainly

younger children but also elderly people with mortality rates of 50 %. Neurological sequelae have been reported in 30 – 40 % of survivors of the severe clinical forms.

History of Immunization. A formaldehyde-killed vaccine, consisting of a 5 % suspension of infected mouse brain tissue, was introduced for human use in Japan in 1954 [219]. The vaccine (Nakayama strain) has been purified by protamine sulfate precipitation since the late 1950s and by absorption with charcoal or kaolin since the early 1960s. A protective efficacy of 80 – 90 % for the vaccine was shown. The highly purified, mouse brain cultured vaccine produced in Japan is also used for immunizing travellers to endemic areas. In China, a vaccine produced in primary hamster kidney cells has been extensively used since the late 1950s.

Production and Properties of Japanese Encephalitis Vaccines. Several Japanese manufacturers produce the purified vaccine from culture of mouse neural tissue by similar methods. No WHO requirements have been formulated. A vaccine used for immunizing travellers to endemic areas is prepared by infecting mouse brain with the Nakayama strain. The brain homogenate is purified by protamine sulfate treatment and then inactivated with formaldehyde. Further purification involves ultracentrifugation on a sucrose density gradient. The vaccine is lyophilized; the reconstituted vaccine must be used immediately.

Immunization Recommendations. Primary immunization consists of subcutaneous injection of two 1 mL doses at a 1 – 2 week interval. A third 1 mL dose is recommended one month later as is a regular booster injection every 1 – 3 years. Extensive immunization in endemic areas has been considered by the WHO. Children less than three years of age should receive 0.5 mL doses. Immunization is generally recommended for health care workers and other people with extended stay in endemic areas.

Adverse Reactions. Only a few mild, local and systemic reactions have been reported.

Protective Efficacy. Immunization has drastically reduced disease incidence in Japan. The protective efficacy is estimated to be 90 – 95 % from serologic studies.

3.13. Smallpox Vaccine

Smallpox (variola) was caused by a member of the pox virus group, which also includes the vaccinia virus used for immunization. The disease was one of the most devastating infections in human history. Eradication of smallpox is the success story of immunization. In 1967 the WHO launched a massive eradication program– smallpox was still reported from 42 countries. In May 1980, WHO officially declared the world free from smallpox. No proven cases of variola have occurred in the past decade.

Two types of vaccine were manufactured, calf lymph vaccine and egg vaccine [220]. Both liquid and lyophilized forms were used. Requirements for manufacturing, control, and potency were formulated by the WHO. Immunization by multiple puncture with a bifurcated needle was most commonly used. Severe adverse reactions included postvaccination encephalitis and disseminated vaccinia.

General immunization against smallpox was withdrawn in most European countries in the mid 1970s. Requirement for smallpox vaccination was abandoned for international travel in 1982.

Recommendations for civilian immunization in the United States include only laboratory workers handling variola virus or other closely related orthopox viruses. Military personnel in the United States and the former Soviet Union are routinely vaccinated against smallpox [221].

3.14. Rift Valley Fever Vaccine

Rift Valley fever is an arthropod-borne disease known only in Africa [212], [213]. It is caused by a member of the Bunya virus group. The major vectors are mosquitoes (*Culex theileri* and *Aedes cabbalus*). The main natural hosts are sheep, cattle, and goats. The incubation period is 2 – 6 days. Rift Valley fever is a febrile disease with headache and abdominal pain lasting less than one week. Hemorrhagic fever with liver

necrosis and encephalitis are the severe manifestations causing mortality and sequelae. Outbreaks occurred in Africa during the 1970s, with the largest outbreak in Egypt in 1977 – 1978.

An inactivated (killed) vaccine (NDBR-103) was produced by the US army in 1967 [222]. The Entebbe strain of the virus was grown in primary monkey kidney cells, inactivated with formaldehyde, and lyophilized. The adverse reactions are few and mild. One case of Guillain – Barré occurred in a Swedish military vaccinee [223]. Seroconversion rates after subcutaneous injection of three 1 mL doses given at 1 – 2 week intervals were over 95 %. A newer vaccine (GSD-200) is based on a cloned version of the original seed virus (Entebbe strain) and grown in diploid rhesus monkey cells. The WHO has formulated requirements for inactivated Rift Valley fever vaccines produced in primary monkey kidney cells and in human or non-human primate diploid cells[141, pp. 104 – 143]. No minimal potency requirements have been formulated.

4. Vaccines against Parasites

4.1. Vaccines against Helminths

Helminths represent one of the major causes of infectious diseases affecting humans and domestic animals. This results not only in a deleterious effect to the health of the hosts, but also in great economic losses. Although major advances have been made in the chemotherapy and epidemiology of diseases caused by helminths (→Anthelmintics) immunotherapy has produced only minor breakthroughs in the field of veterinary parasitology.

As a result of a long evolutionary development and a close parasite – host relationship, helminths have evolved strategies for circumventing complete elimination by the host immune response. Nevertheless, the immune response usually exerts deleterious effect upon their growth and proliferation. Thus, the use of a vaccine resulting in partial or complete protection might be one of the most cost-effective means of controlling helminth diseases.

Only a few reliable vaccines are available on a commercial scale for immunoprophylaxis of helminthoses in livestock. A vaccine against lungworms in cattle is a commercial success. Another vaccine against hookworms in dogs, although immunologically efficient, has failed commercially.

Important in the development of vaccines is the identification, isolation, and testing of putative protective antigens. Successful vaccination against helminth infections requires the priming of those responses which may subsequently be triggered during natural infection. Since antigen presentation plays a central role in the acquired immune response, the development of accessory cells and the activation of T and B lymphocytes has to be taken into account [224]. Developments have focussed on vaccines produced by recombinant DNA techniques. In vitro cultivation methods have been used for producing helminth antigens for immunoprophylaxis and combined with recombinant DNA technology [225].

The helminthoses described in this chapter were chosen on the basis of the parasites' importance to human health or because of their interesting biology.

4.1.1. Vaccines against Schistosoma

Etiologic Agents, Pathogenesis, and Epidemiology. Human schistosomiasis occurs primarily in tropical countries where it is one of the most threatening diseases. The infection affects about $(200 – 300) \times 10^6$ people throughout the world; more than 600×10^6 people live in *Schistosoma*-endemic areas. The importance of schistosomiasis has also increased in industrialized countries with intensive tourism and influx of high numbers of refugees from endemic areas [226].

The main causative agents of schistosomiasis are helminths of the genus *Schistosoma* which use aquatic or amphibious snails as intermediate hosts:

1) *Schistosoma haematobium* uses aquatic snails of the genus *Bulinus* as intermediate hosts and occurs mainly in Africa and some middle-eastern countries.
2) *S. mansoni* uses aquatic snails of the genus *Biomphalaria* as intermediate hosts; it occurs in Africa, parts of Arabia, northern and eastern parts of South America, and some Carribean islands.

3) *S. japonicum* uses amphibious snails of the genus *Oncomelania* as intermediate hosts; it occurs in Japan, the Philippines, and parts of China, Thailand, and Indonesia.

The life cycle of all three *Schistosoma* species is similar. The fully developed *miracidium* hatches from the egg in water and infects the intermediate host (i.e., the snail) where it multiplies asexually to produce numerous *cercariae*. The cercariae are released into the water and infect humans by penetration through the skin. They develop into immature worms (schistosomula) which migrate to the lungs and liver. The adult, sexually mature worms mate and migrate to their final destination which varies according to the species: *S. haematobium* migrates to the veins of the vesical plexus, *S. mansoni* and *S. japonicum* to the mesenteric veins. The adult paired worms produce eggs (300 – 3000 per pair per day) which pass through the vessels and tissues into the lumen of the gut and bladder. The eggs escape from the host in the feces and urine and the cycle is repeated.

Eggs from *S. haematobium* are mainly found in the bladder and urogenital tract causing hematuria and fibrosis of the bladder. In severe cases malignancies may develop. Eggs from *S. mansoni* and *S. japonicum* are trapped in the liver and bowels causing hepatomegaly. Subacute disease is probably due to the passage of worms through lungs leading to cough, pulmonary infiltrates, and fever [227]. Mainly in *S. japonicum*, acute schistosomiasis occurs 5 – 7 weeks after heavy primary infection. The illness if often associated with diarrhea, fever, hepatosplenomegaly, resulting in liver fibrosis and ascites.

Immunity and Vaccine Design. An age-dependent resistance to reinfection after chemotherapy was demonstrated with *S. haematobium* and *S. mansoni* [228], [229]. This is associated with an increased lymphocyte proliferation in response to egg, cercarial, and adult worm antigen [230], [231]. Lack of reinfection is also related to the eosinophil count [228], [232]; human eosinophils mediate antibody-dependent damage to the schistosomula of *S. mansoni* [233]. The main immune mechanisms involved in schistosomiasis are summarized in Table 2 and reviewed in [234].

Table 2. Main cells and antibodies active against *Schistosoma* in experimental animal models or in vitro

Developmental stage of parasite	Cells and antibodies involved	Mechanisms
Egg	effector T lymphocytes	initiation of granuloma formation (lymphokines)
Schistosomula	neutrophils, IgG	tegumental damage and killing of young larvae under certain conditions
Schistosomula	eosinophils, IgG	killing of young larvae in the presence of complement
Schistosomula	eosinophils, IgE	killing of young larvae
Schistosomula	macrophages, IgE	IgE-dependent cytotoxicity

The immune response may result in resistance to reinfection but not to simultaneous expulsion of an established parasite population from primary exposure [235]. In this way, the parasites evade the immune response. An alternative escape mechanism is that the worms may be covered by bound host IgG [236]. The immunoglobulin is partially cleaved by a parasite protease to produce peptides which may inhibit macrophage activation and thus depress macrophage-mediated, IgE-dependent destruction (cytotoxicity) of the schistosomula.

In the development of schistosomiasis vaccines, the findings regarding age-dependent host resistance will require application and efficacy at a very early age, before the child is exposed to natural infections [237]. The search for antigens that mediate protective immunity has concentrated on the exposed outer surface of the young schistosomula [238]. About 90 % of exposed epitopes consist of carbohydrates that crossreact with *Schistosoma* egg antigen. Antibodies to surface polypeptide antigen do not generally crossreact with egg antigen but are present on the surface membrane of adult worms. Monoclonal antibodies to some of these molecules seem to confer partial resistance when passively administered to animals and could thus be potentially protective antigens [239–241]. A range of candidate vaccine antigens of *S. mansoni* have been identified, several have been cloned and expressed in *Escherichia coli*, yeast, or vaccinia virus [238]. This important achievement

will facilitate the production of large amounts of polypeptides for vaccination trials.

One of the most promising candidates is a schistosomula surface polypeptide with a molecular mass of 28 000. The gene coding for this polypeptide has been cloned [242] and expressed in *E. coli*. Immunization with this recombinant antigen induced a high level of serum cytotoxicity towards schistosomulas in the rat, hamster, and monkey. Significant protection against a natural challenge infection with live cercariae was obtained in rats and hamsters [242]. These results can be viewed with optimism for the development of a schistosomiasis vaccine.

4.1.2. Vaccines against Nematodes

4.1.2.1. Gastrointestinal Nematodes

The most prevalent and pathogenic gastrointestinal nematode parasites of humans belong to the genera *Ascaris, Strongyloides, Trichinella, Trichuris, Ancylostoma,* and *Necator.* Although many species parasitize deeper tissues of the body, the majority are intestinal. The intestine has been maintained as a site for adult stage development, whereas larval stages often invade other host tissues. The intestine consists of a series of distinct parasite habitats (gut sections, lumen, mucosa, etc.) each having its own characteristics. Large worms such as *Ascaris* must live within the lumen, smaller species such as hookworms are associated with the mucosa. For a long time research on immunity against intestinal worms was given low priority due to the lack of knowledge about local immune responses in the gut. Studies have shown that intestinal worms are indeed subject to protective immune responses, although these responses differ somewhat from classical immunity in the body because the worms live in the gut lumen. Special features are (1) macromolecular antigen uptake across the intact mucosal epithelium or by specialized epithelial cells overlying Peyer's patches and (2) complexation of the antigen by dimeric IgA secreted from the mucosa or intestinal IgG Fab fragments. Cells from the underlying lamina propria participate in cytotoxicity and hypersensitivity reactions and thus affect mucosal structure and function. Furthermore, a large variety of nonlymphoid effector cells occur within the intestine, including natural killer cells, macrophages, neutrophils, eosinophils and basophils; their numbers increase during parasite infections.

No vaccine against gastrointestinal nematodes is presently available for human application. However, promising, successful trials in veterinary parasitology have initiated interesting work in human gastrointestinal nematode infections and will probably result in vaccine supply in the near future.

Hookworm Disease. Various species of the family Ancylostomatidae are responsible for hookworm disease in humans, the most important being *Ancylostoma duodenale* and *Necator americanus*. Over a fifth of the world's population is afflicted by this disease, mainly in tropical and subtropical regions. Adult hookworms have a length of about 0.7 – 1.8 cm and a hook-like anterior end with a distinct mouth area. Females release eggs in the small intestine which are excreted in the feces of the host. In humid surroundings first-stage larvae hatch from the eggs and develop into infective larvae, which penetrate through the skin into new hosts. After migration through lymph or blood vessels, the larvae finally develop into mature adult worms in the small intestine, completing the parasite's life cycle. Hookworms suck blood from microlesions in the host's small intestine, thereby causing chronic gastrointestinal blood loss, anemia, and hypoalbuminemia. With a large worm burden death may occur. Patients usually suffer from extreme weakness, pallor, secondary respiratory tract infections, skin irritations, heart palpitations, and gastrointestinal distress.

Immunity and Vaccine Design. Nematode parasites present special problems for the host's protective immune response, because they possess a tough, protective, external cuticle. The cuticle is both antigenic and immunogenic, but it is doubtful whether responses directed against its surface play a major role in immunity against intestinal species. Protective responses are more likely initiated by antigens released through the orifices of the parasite [243]. One such antigen with a potential vaccine function is a secreted proteolytic enzyme [244]. Hookworms attached to the mucosa secrete the enzyme from glands in their mouths. The enzyme degrades the host

proteins and inhibits blood coagulation which permits the hookworms to feed for an indefinite period of time. Immunizing the host against this enzyme would result in inhibition of any enzyme secreted by worms after a subsequent challenge infection. The worms would then be unable to feed. The gene coding for the enzyme in question has been cloned. Future investigations will have to demonstrate the applicability of this kind of vaccine.

A vaccine has been developed for the control of hookworm infections in dogs [245]. It was based upon irradiation-attenuated infective larvae and was administered parenterally. Although the vaccine was highly effective in preventing hookworm disease, it was withdrawn because it did not completely prevent infection, and effective anthelminthic chemotherapy was available. This example suggests that protective immunity may also occur in human hookworm infections.

4.1.2.2. Tissue-Invading Nematodes (Filariidae)

Many nematode species which live as adults in the intestine, (e.g., *Ascaris*, hookworms, and *Trichinella*) undergo development in parenteral tissues. Other species are wholly confined to these tissues and have no contact with the intestine (tissue-invading nematodes). This closed habitation site requires special conditions in order to obtain biological contact with the outside world, especially for reproduction. In the major group of tissue-invading nematodes, the Filariidae, the worms overcome this problem by using bloodfeeding arthropods as intermediate hosts. The female worms release microfilariae larvae which circulate in the blood or accumulate in the skin of the host. The arthropod feeds on the blood and takes up the microfilariae which develop into infective larvae. At a following blood meal, infective larvae are reinoculated into new human hosts.

Filariasis. The human disease filariasis comprises an extremely heterogeneous group of diseases. The main filiarial parasite species in humans are *Wuchereria bancrofti* and *Brugia malayi*, which are both transmitted by mosquitoes. The disease is widespread in tropical and subtropical regions affecting ca. 90×10^6 persons. Its early symptoms are fever, lymphangitis, and lymphadenitis. A following chronic stage is frequently characterized by more serious clinical manifestation including elephantiasis, hydrocele, and pulmonary eosinophilia. Adult worms are found in the lymphatic system, microfilariae may be found in blood.

Filiarial infections caused by *Onchocerca volvulus* occur in Africa and South and Central America, affects millions of patients, many of whom become blind. Adult worms are generally located under the skin, forming typical nodules; less often they penetrate deeply into the tissues. Pathology in onchocerciasis is due entirely to the microfilarial stage. Microfilariae are present in the skin and may penetrate into the eye, thus leading to severe eye lesions and blindness.

Loa loa is prevalent in the forest areas of West Africa. Adult worms penetrate into the tissues, provoking transient edema. Severe clinical symptoms are rare, loiasis being generally regarded as a benign infection [246]. *Dipetalonema perstans, D. streptocerca,* and *Mansonella ozzardi* infect humans (Africa, South America) but usually asymptomatically and rarely cause significant diseases [247].

Many of the changes associated with filarial infection are immunopathological in origin and hypersensitivity reactions are important in their development.

Immunity and Vaccine Design. Although filarial parasites provoke a strong immune response in the human host, the chronicity of these infections implies the absence of a protective response or the evasion of such responses by the worms. There is also no direct evidence that filarial infections in nature confer resistance to reinfection with the same parasite species.

No vaccines against filarial parasites are available. The development of new strategies for immunological control depends on a thorough understanding of immunological host – parasite relationship. Many studies on protective immunity in animals have concentrated upon responses directed against larval stages. A vaccine against microfilariae would inhibit transmission of the disease, a vaccine against infective larvae would provide protective immunity against primary infection of hosts. Inoculation with infective larvae attenuated by irradiation [248], [249], only confers partial protection to a subsequent challenge infection. Filarial vaccines based on

irradiated larvae cannot be used in humans without first determining whether these attenuated larvae induce pathological changes [250].

The target antigens of antimicrofilarial immune responses are probably located on the surface of the microfilariae; they are currently being characterized [251]. Antibodies to surface antigens mediate adherence of host cells to the microfilariae; this can result in worm killing [243]. Using *Dipetalonema vitae* as a model, IgE was found to be the primary immunoglobulin involved in cell adherence to the worm cuticle. The first cell type to adhere is the eosinophil. Subsequent, adherence of macrophages is followed by release of lysosomal enzymes which degrade the cuticle. In humans, a comparatively long time is needed to establish such immunity. This may be due to pronounced immunosuppression or because the microfilariae cover their surface with host components and thus render their antigenic surface epitopes less accessible to the host's immune system. The search for filarial antigens that can safely and successfully be used in humans is still continuing; it is still not known how restricted series of antigens can be used to protect natural hosts against first or persistent infections with a complex, adaptable, genetically diverse parasite. Recombinant DNA technology and hybridoma technology may provide potential candidates for successful filarial vaccines.

4.1.3. Vaccines against Cestodes

Etiologic Agents, Pathogenesis, and Epidemiology. This section deals with the hydatidosis and cysticercosis disease complexes, which are caused by the larval stages of tapeworms belonging to the family Taeniidae. The most important causative agents of *hydatidosis* (echinococcosis) are *Echinococcus granulosus* and *E. multilocularis,* whose life cycles involve a definitive and an intermediate mammalian host. The definitive hosts are carnivores (mainly dogs for *E. granulosus* and foxes for *E. multilocularis*) in whose intestines the adult stage worms occur. Intermediate hosts are herbivorous and omnivorous species in which the larvae (metacestodes) develop. Humans and other intermediate hosts become infected by ingesting eggs passed in the feces of definitive hosts. The diseases caused by the metacestodes are referred to as cystic echinococcosis for *E. granulosus* and alveolar echinococcosis for *E. multilocularis. Echinococcus granulosus* is prevalent throughout the world and is a public health and economic problem in many areas. *Echinococcus multilocularis* only exists in the northern hemisphere and is relatively frequently seen in the former Soviet Union (Siberia), central Europe, northern China, Japan, and Alaska. The fully developed metacestode of *E. granulosus* is a typically unilocular, fluid-filled cyst, which is located in the liver, lungs, and other organs. *Echinococcus multilocularis* metacestode conforms a vesiculated parasitic mass in the liver of the host, it proliferates by continuous exogenous budding with possible metastasis formation in other organs.

The causative agent of *cysticercosis* is the tapeworm *Taenia solium*. Humans are the obligatory definitive hosts, pigs act as intermediate hosts. Metacestode stage infection can also occur in humans and may result in infection of the central nervous system by parasite larvae (neurocysticercosis). Morbidity includes intracranial hypertension, basal arachnoiditis, focal neurological deficits, and dementia. Hyperendemic areas are found in Latin America, Africa, and Asia; areas of lower endemicity occur in southern and eastern Europe.

Immunity and Vaccine Design. Host protective immunity is a striking feature of repeated infection with cestodes in mammalian intermediate hosts [225]. It plays a major role in regulating natural transmission of these parasites, and substantial research efforts have been undertaken towards development of vaccines against cestodes of veterinary importance [252]. For human cestode infections, it is debatable whether there is sufficient importance to warrant the research required to develop a vaccine. In areas of high egg contamination (e.g., the Turkana district (Kenya) for *E. granulosus*, St Lawrence Island for *E. multilocularis*, or Mexico for *T. solium*), vaccination should be given to very young persons, as patients usually become infected at very young age. An ideal human vaccine requires complete, long-lasting protection. Experiments in veterinary parasitology, however, demonstrated exactly opposite results. Vac-

cination of animals against *Taenia* species resulted only in marked reduction of cyst numbers which persisted only for a maximum of one year. In addition, strong adjuvants had to be employed which are not tolerated by humans. Other criteria may be important; a review of protective immune mechanisms is given in [253].

Little attention has been paid to the vaccination of definitive hosts (dogs and foxes for *Echinococcus*, and humans for *T. solium*). Experiments [254], [255] demonstrated an immune response after infection as well as highly significant suppression of egg production by *E. granulosus* after immunization of dogs with secretory antigens derived from adult tapeworms. This approach seems to be the most likely control measure for reducing infection risk in humans.

4.2. Malaria Vaccine

Malaria remains a major health problem in many tropical and subtropical countries and affects hundreds of million of people each year. A major effort was made to control malaria from 1950 – 1970 with insecticides (→ Insect Control) and antimalarial drugs (→ Chemotherapeutics, Chap. 3.3.). The initial remarkable results have been difficult to maintain. The advent of drug-resistant parasite strains and of insecticide-resistant mosquito vectors are major obstacles in the effort to control malaria. Since the early 1970s new approaches have been explored such as vector control through biological agents and control of malaria infection through vaccines.

The development of malaria vaccine has received considerable impetus: first, because immunization of animals with whole parasites can induce a degree of protection equal or superior to that induced following natural infection [256–258]; second, because in vitro culture systems have been developed for the blood and hepatic stages of the malarial parasite *Plasmodium falciparum* [259], [260]; and third, because monoclonal antibody and recombinant DNA techniques have been used to identify and produce the parasite polypeptides possibly involved in the development of protective immunity.

4.2.1. Strategy for Malaria Vaccine Development

More than 100 species of malarial plasmodia are known, but only four infect humans: *Plasmodium falciparum*, which is responsible for the majority of human deaths, *P. vivax, P. malariae,* and *P. ovale*.

The life cycle of the plasmodia is complex (Fig. 4). The female anopheline mosquitoes inoculate *sporozoites* into the blood of the vertebrate host. Within minutes the sporozoites invade the liver parenchymal cells (hepatocytes) where they divide asexually and develop into *merozoites* which rupture the hepatocytes and reenter the blood. In the subsequent *erythrocytic cycle,* the merozoites invade the red blood cells and mature into *schizonts* within 48 – 72 h depending on the species. The mature schizonts release merozoites which invade new erythrocytes. The erythrocytic cycle is responsible for the clinical manifestations of malaria. Some merozoites differentiate into sexual stages called *gametocytes* which are ingested by the mosquito. Fertilization of the gametes occurs solely in the midgut of mosquito. The resulting zygotes develop into *ookinetes* and then into *oocysts*. Sporozoites are released from mature oocysts and migrate to the mosquito salivary glands. The cycle is then repeated.

This complex life cycle involves continuous morphologic, enzymatic, and antigenic changes which are linked to the parasite's environmental adaptation to the host. The invasive stages of the parasite (sporozoites, merozoites, gametes) have unique, stage-specific surface determinants. Furthermore, immunologic crossreactivity exists between plasmodia species and between the developmental stages of a given species. However, immunization experiments in animal models have demonstrated that the antigenic determinants involved in protective responses are species- and stage-specific. Three types of malaria vaccine can therefore be devised, based on:

1) sporozoites,
2) asexual stages (merozoites, schizonts), and
3) sexual stages (gametes).

In addition, stagespecific parasitic antigens are expressed on liver cells containing merozoites [261] and may also be candidates for vac-

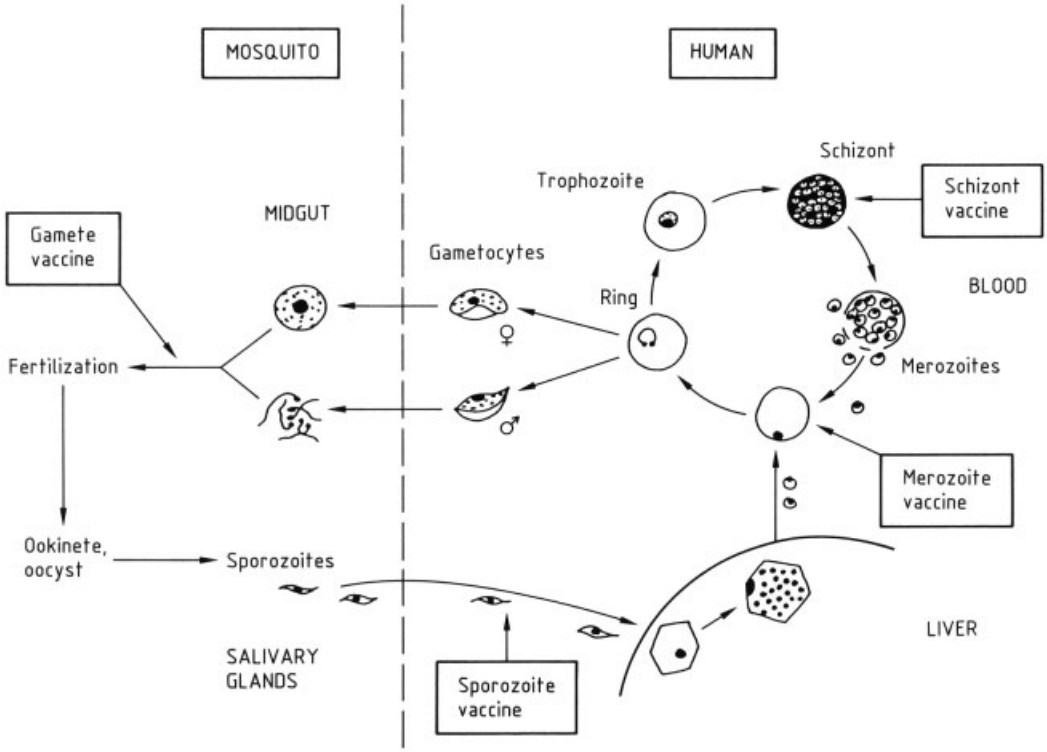

Figure 4. *Plasmodium falciparum* life cycle showing targets for vaccine development

cine development. A favored approach is the development of a multivalent vaccine containing components of several malaria stages. The selection of defined parasite components versus whole parasites (sporozoites, merozoites, gametes) for vaccine development is indicated by the following reasons:

1) Large-scale production and purification of whole parasites is not feasible.
2) Parasites cannot be obtained free of host components (e.g., mosquito salivary glands, erythrocyte membranes) which may induce adverse autoimmune reactions.
3) Most of the parasite components are irrelevant to the induction of protective responses and their inclusion may impair truly protective responses or induce immunopathologic lesions in the host.

The strategy for the development of malaria vaccine is as follows:

1) Identification and selection of malaria antigens that are the target of protective immune responses.
2) Functional, biochemical, and immunological characterization of these antigens, including identification of B and T cell epitopes.
3) Cloning of the genes coding for protective antigens, determination of DNA and amino acid sequences–what is the level of antigenic diversity?
4) Production of candidate protective antigens or epitopes by genetic engineering or chemical synthesis.
5) Evaluation of candidate protective antigens in terms of production of antigens, safety tests, adjuvants, carriers, etc.
6) Immunization trials in monkeys and human volunteers.

4.2.2. Sporozoite Vaccines

Following invasion of the hepatocytes, the sporozoites develop into thousands of mero-

zoites, each of which may invade an erythrocyte. Obviously a vaccine which neutralizes sporozoites before their entry into liver cells or within liver cells would optimally prevent malaria infection. Vaccination with attenuated, irradiated sporozoites in rodents, monkeys, and humans induced complete protection against malaria [262], [263]. The induced immunity is species- and stage-specific but not strain-specific. It is at least partially mediated by antibodies as is shown by the protection against sporozoite challenge afforded by the passive transfer of antisporozoite monoclonal antibody [268] and by neutralization of sporozoite infectivity following incubation with the serum of protected animals. Deposition of specific antibodies on the sporozoite surface results in the formation of a tail-like precipitate; this is called the circumsporozoite (CS) reaction. Indirect evidence suggests that antisporozoite antibodies may play a role in human malaria [264], [265].

The control of sporozoite-induced infection also involves T cell-dependent mechanisms. In animal models, effector T cells can mediate antisporozoite immunity via antibody-independent mechanisms; for example, immunization with irradiated sporozoites can protect B cell-deficient mice against subsequent challenge infection with viable sporozoites [266]. Activation of malaria-specific T cells by malaria antigens induces the secretion of lymphokines which may act directly on the malaria parasite or indirectly by activating host effector systems [267].

The CS protein has been identified in several plasmodia species. Passive transfer of monoclonal antibodies directed against the CS protein protects mice from sporozoite challenge infection [268]. Indirect evidence suggests that CS protein is involved in the binding and penetration of sporozoites into liver cells. Similarly, fragments of monoclonal antibodies against CS prevent the attachment of sporozoites to hepatocytes in vitro [269].

The gene coding for the CS protein has been cloned in several plasmodial species [270], [271]. The protein contains a central block of tandemly repeated amino acids which vary in number and sequence among the different malaria species [272], [273]. For example, in *P. falciparum* the central area consists of 37 copies of the tetrapeptide asparagine – alanine – asparagine – proline (NANP) and four copies of the tetrapeptide asparagine – valine – aspartic acid – proline (NVDP). The regions flanking the repeats are more highly conserved between species than the repeats. Within a species, limited variations also occur outside the repeats [274].

The repeats cover the surface of mature sporozoites; for *P. falciparum* ca. 10^8 molecules of NANP are expressed on the membrane of mature sporozoites. The NANP repeats are the target of protective monoclonal antibodies passively transferred in vivo. The antibody response against sporozoites in humans is also mainly directed against the NANP repeats [275], [276].

In view of these findings, and because the NANP repeats are present on all the isolates of *P. falciparum* tested, two malaria vaccines based on NANP repeats have been prepared and tested on human volunteers. The first, produced by DNA recombinant technology, consisted of 32 repeats of NANP and NVDP fused to a 32 amino acid tail. The second was composed of three NANP repeats (NANP 3) conjugated to tetanus toxoid. Both formulations use aluminum hydroxide as adjuvant [277], [278]. The vaccines were safe and did not induce adverse reactions. Volunteers with high antisporozoite antibody titers were challenged with sporozoites and some were protected or presented a delay in appearance of parasitemia. Protection was shown in individuals with the highest antibody titers.

An antisporozoite vaccine must induce high antibody titers for a prolonged period and ideally a boosting effect should occur following exposure to sporozoites. Higher antibody responses can be obtained by changing the formulation and concentration of the immunogens and by using other adjuvants or live-attenuated vectors (vaccinia virus, salmonella) that carry and express the gene coding for the CS protein. Optimal antibody formation is dependent on the collaboration of T helper cells and B cells. In the two sporozoite vaccines tested, T cell help was provided by foreign protein (tetanus toxoid) but was unable to boost the antiNANP antibody response following exposure to sporozoites. More efficient sporozoite vaccines should contain T cell epitopes derived from sporozoites and ideally from the CS protein; in this context, proper help is provided by sporozoite-specific, primed T cells and can lead to optimal antiNANP anti-

body production by B cells. In a mouse model the response to some of the T cell epitopes of the CS protein is restricted by antigens of the major histocompatibility complex (MHC class II) [279–282]. In human populations, only three immunodominant epitopes are located in polymorphic regions of the CS protein outside the repetitive area. Since T cells have exquisitely specific reactivity, it follows that the polymorphism of T cell determinants may be responsible for a lack of proper help following exposure to sporozoites with T cell areas on CS that are different from those present on sporozoites responsible for previous infections in the same individual.

Therefore, it seems that more efficient vaccines should be based on either native malaria polypeptide(s) or cocktails of synthetic polypeptides containing multiple B and T cell epitopes. These epitopes should be selected in relation to constant and variant parasite components and in relation to epitope binding and recognition by components of the major histocompatibility complex of the human host.

4.2.3. Asexual Blood Stage Vaccine

The multiplication of asexual blood stages (merozoites and schizonts) is responsible for the morbidity and mortality associated with malaria. The level of parasitemia usually correlates with the severity of malaria infection. Immunity to malaria is mostly acquired but natural immunity also plays a role. Several single-gene disorders affecting erythrocytes, (e.g., sickle cell anemia, the thalassemias, and glucose phosphate deficiency) reduce the severity of malaria infection. Another genetic characteristic, the lack of the Duffy blood group antigens, is associated with complete resistance to *P. vivax* infection [283]; the Duffy blood group antigen or a closely associated antigen may be the receptor for *P. vivax* merozoites at the surface of erythrocytes.

The development of acquired resistance to malaria depends on the frequency and duration of the exposure to the parasite [284]. In endemic areas, babies born to immune mothers are resistant to malaria during the first three months of life due to the presence of maternal antibodies transferred during gestation. Later they suffer from severe, recurrent attacks; most deaths due to malaria occur in young children. From adolescence to adulthood there is a decrease in the severity and frequency of malaria attacks but sterile immunity is probably never achieved. In this context two types of vaccine based on asexual blood stages can be envisaged; (1) a vaccine which is more efficient than nature and leads to sterile immunity (i.e., infection no longer detectable) or (2) a vaccine capable of attenuating the parasite load by transforming the immune system of a non immune individual into that of an adult living in an endemic area.

The immune response to blood stages is complex and is directed against several antigens. Both antibody-mediated responses and cell-mediated, antibody-independent responses control asexual blood-stage infection. In humans, passive transfer of immunoglobulins purified from the sera of immune adults abort malaria infection in nonimmune infected children [285]. The antibodies may react with the surface of the merozoites and provoke their lysis upon addition of complement, or enhance their phagocytosis by mononuclear cells, or simply inhibit the binding of merozoites to erythrocytes. Other targets for antibodies are antigens on the surface of erythrocytes containing schizonts; binding of antibodies to schizonts may also lead to their destruction by phagocytosis [286] or induce the endothelial release of schizonts which may be later destroyed in the spleen [287]. Immunity to asexual blood stages also operates through a variety of antibody-independent mechanisms: T cell-dependent release of lymphokines, induction of oxidizing radicals leading to intracellular death of the malaria parasites, and activation of mononuclear cells in the spleen.

Immunization with merozoites and/or schizonts in a variety of plasmodia – host systems resulted in partial to almost complete protection [288]. Subsequent investigations were aimed at the characterization of components capable of inducing immunity. Several hundreds of asexual blood stage components can raise an immune response but only very few of the evoked responses are helpful to the host. Characteristics of some of the candidate antigens for asexual bloodstage vaccines are discussed in Sections 4.2.3.1, 4.2.3.2, 4.2.3.3, 4.2.3.4.

4.2.3.1. Merozoite Surface Antigens

A protein with a molecular mass of 190–200 kDa has been identified at the surface of *P. falciparum* schizonts and merozoites [289], [290]. During maturation of schizonts this polypeptide is processed into several components, one of them (M_r83 000) being the main surface component of the merozoites [290]. An important feature of the 190–200 kDa protein is its genetic polymorphism [291–294]. The gene coding for the protein can be divided into blocks ranging in homology among different *P. falciparum* isolates from 10–87 % at the amino acid level. A relatively short region of variable tripeptide repeats is found close to the N terminus. The blocks encoding for the N and C terminal sequences are highly conserved.

Immunization with the 190–200 kDa protein derived from *P. falciparum* in monkeys [295–297] and with an analogous protein from *P. yoelii* [298] can induce at least partial protection. Immunization with synthetic polypeptides corresponding to defined parts of the molecule (for example, the N terminus and amino acids 277–287) also confer partial protection in monkeys [299], [300].

An antigen with a molecular mass of 51 kDa is also expressed at the surface of *P. falciparum* merozoites and is the target of inhibitory monoclonal antibodies. It contains variant and constant epitopes for various *P. falciparum* isolates.

There is considerable antigenic diversity among *P. vivax* isolates as regards the components exposed at the surface of merozoites.

4.2.3.2. Rhoptry Antigens

Rhoptries are apical organelles of the merozoites which release their contents onto the erythrocyte membrane during invasion. A monoclonal antibody directed against a rhoptry protein of a rodent malaria, *P. yoelii*, reduced the virulence of the infection and a monoclonal antibody directed against 82 and 41 kDa components of *P. falciparum* inhibited the growth of *P. falciparum* in vitro [301–303]. The 82 kDa component is processed into 82 and 65 kDa components [304].

The 82 kDa polypeptide is membrane-bound through a glycosyl–phosphatide–inositol anchor; hydrolysis of its anchor activates the proteolytic activity of the 76 kDa polypeptide and may play a role in the invasion of erythrocytes by merozoites [305]. The 41 kDa polypeptide displays aldolase activity [306]. Interestingly, both the 76 kDa and the 41 kDa components can induce at least partial protection against *P. falciparum* infection in monkeys [307], [308]. The gene coding for the 41 kDa protein has been cloned and presents two interesting characteristics in terms of vaccine development; absence of variable amino acid repeats and almost complete conservation of the amino acid sequences among isolates from *P. falciparum* [306].

Another rhoptry antigen of *P. falciparum* with a molecular mass of 225 kDa has been identified in the peduncle of the rhoptries. It is synthesized as a 240 kDa polypeptide which is processed into a 225 kDa protein during schizogony and is quantitatively recovered in the culture supernatant following merozoite invasion [309].

A third set of rhoptry-associated proteins is the 105–130–140 kDa complex composed of three coprecipitating but unrelated proteins [310]. The 225 kDa proteins and the 105–130–140 kDa complex have not been evaluated in immunization trials.

4.2.3.3. Antigens Associated with the Membrane of Infected Erythrocytes

The ring-infected erythrocyte surface antigen (RESA) is a *P. falciparum* antigen with a molecular mass of 155 kDa. It is synthesized in trophozoites, accumulates in the merozoite, and following invasion becomes associated with the membrane of erythrocytes containing ring forms of the parasite [311], [312] but is not directly accessible on the external erythrocyte surface. AntiRESA antibodies inhibit the multiplication of asexual blood stages in vitro and may interfere with the invasion process [313]. The gene coding for RESA has been cloned and sequenced [314]. It contains two blocks of repetitive amino acid sequences which are the immunodominant regions of the molecules in terms of antibody response. Antibodies directed against the RESA repeats crossreact with at least six other asexual blood stage components. Aotus monkeys have been immunized with fusion proteins corresponding to various areas of the RESA and

with synthetic polypeptides corresponding to the repetitive sequences [315]. Partial protection was observed in some groups of animals; work is in progress to optimize the efficacy of immunization based on RESA-derived molecules.

Erythrocytes containing mature asexual blood stages of *P. falciparum* attach to endothelial cells lining the venules of deep tissues. This cytoadherence of mature parasites prevents their passage through the spleen and thus their exposure to localized destructive mechanisms. Electron-dense protuberances (knobs) on the plasma membranes of infected erythrocytes are implicated in cytoadherence [316]. The genes coding for two knob components (knob-associated histidine-rich protein M_r85 – 105 kDa) and mature parasite-infected erythrocyte surface antigen (M_r240 – 300 kDa) have been cloned [317], [318]. The two proteins differ antigenically among isolates and contain repeated amino acid sequences. Cytoadherence can be inhibited by antisera in a strain-specific manner [319]. However, an antigenically invariant epitope has also been identified on the surface of infected erythrocyte isolates and may be an important antigen for vaccine development [320].

4.2.3.4. Other Proteins and Synthetic Peptides

A number of other antigens are also candidates for vaccine development. *P. falciparum* requires exogenous iron in the form of ferrotransferrin. A malaria transferrin receptor at the surface of infected erythrocytes transports bound ferrotransferrin to the parasite and may be used as a target for the vaccine [321]. Glycophorins exposed at the erythrocyte surface may act as ligands for *P. falciparum* merozoites and *P. falciparum* proteins have been identified which either bind to glycophorins or to human erythrocytes [322], [323]. A prominent antigen of *P. falciparum* with an apparent molecular mass of 126 – 140 kDa is associated with merozoite release. The gene coding for this protein has been cloned and contains at least two stretches of amino acid repeats, one being composed of polyserine repeats [324]. Monkeys immunized with this protein are protected from a lethal challenge infection [295].

Several immunization trials have been conducted in monkeys using synthetic peptides derived from asexual blood stages of *P. falciparum* coupled to carrier proteins [298], [299], [315]. A partial protective response was observed with peptides corresponding to various areas of the 190 – 200 kDa protein, to RESA, and to fragments of parasite components identified by their molecular mass of 55 and 35 kDa [299]. Synthetic hybrid polymer – proteins containing several peptides corresponding to epitopes of 195 – 200 kDa, RESA, 55 kDa, 35 kDa, and CS protein have been used for immunization of human volunteers [324]. The vaccine was well tolerated and no adverse effects were observed. All the immunized and control volunteers had patent parasitemia but the majority of the immunized volunteers were able to control their parasitemia in the absence of drug therapy. The immune response was low in terms of specific antibody production, and cell mediated responses as measured by proliferation assays was undetectable.

4.2.4. Sexual Stages–Transmission Blocking Immunity

The transmission of malaria from the vertebrate host to the mosquito vector is effected by sexual parasite stages–the gametocytes–which develop from merozoites. Within the vertebrate host the gametocytes are surrounded by the erythrocyte membrane; following ingestion by the mosquito vector, the gametes become extracellular. The female gametes are fertilized by the male gametes in the midgut of the mosquito to produce zygotes which develop into ookinetes. The ookinetes penetrate the midgut wall where they remain to form oocysts in which the sporozoites develop.

In the vertebrate host, the sexual stages do not produce illness, and their intracellular localization prevents direct attack by host effector mechanisms. Within the midgut of the mosquito the extracellular gametes are exposed to antigamete and/or antizygote antibodies from the vertebrate host that are ingested with the mosquito's blood meal. The antibodies partially or completely prevent the development of sexual stages and subsequent production of sporozoites. Transmission of the parasite is therefore blocked. This phe-

nomenon is termed *transmission blocking immunity*.

The development of vaccines based on sexual blood stages could have an important impact on the epidemiology of malaria in endemic areas by reducing the level of malaria transmission. Ideally, transmission blocking vaccines have to be used in combination with vaccines based on sporozoite and/or asexual blood stages (see Sections 4.2.2 and 4.2.3).

Transmission blocking immunity has been induced by immunization with extracellular gametes in several animals [325–328]. The induced antigamete response is long lasting and in some cases is boosted by malaria infection, probably due to the gametocyte antigens in the circulation of the vertebrate host [328], [329]. There is also evidence that in *P. vivax* malaria in humans the antigamete response is boosted during natural infection [330]. Addition of sera of previously infected individuals to gametes can prevent fertilization and development of oocysts in mosquitos.

Specific targets for antigamete immunity have been identified using monoclonal antibodies in species including the human parasites *P. falciparum* and *P. vivax* [331], [332]. The antibodies act against the gametes by preventing fertilization and against the zygotes and ookinetes by preventing further development. In *P. falciparum* the target antigens for inhibition of fertilization are polypeptides with a molecular mass of 230 kDa and 45 – 48 kDa [331], [332]. Some of the epitopes on the 45 – 48 kDa antigen have been defined and are the targets of inhibitory monoclonal antibodies [333]. New antigens are expressed at the surface of the zygote and ookinete and one of them (M_r= 25 kDa) is a probable target of inhibitory monoclonal antibodies [331].

5. Immunotherapy

Since the late nineteenth century considerable progress has been made in our concepts of passive immunotherapy and in the development of preparations safe for human use; however, the proper role for such therapy in clinical medicine still needs to be defined.

By 1900 immune serum from various animal species had been used to treat pneumonia, tetanus, diphtheria, and rabies. Human serum was first used in 1907 by Cenci for the modification of measles and later for mumps and pertussis [334]. Placental extracts prepared by ammonium sulfate precipitation [335], [336] were also employed and may be considered as the first immunoglobulins prepared for human therapy [336]. Placental material is still used as a source of immunoglobulin.

The serious hypersensitivity reactions associated with animal serum proteins and the risk of viral hepatitis with convalescent human serum limited the use of serum therapy to life-threatening infections. In the 1920s attempts were made to separate the immune substances from animal serum by alcohol or acetone treatment. One such preparation, Huntoon's antibody solution, was administered intravenously to over 400 patients without the occurrence of anaphylaxis or serum sickness; however, pyrexia, cyanosis, and dyspnea did occur and were implicated in the deaths of three patients [337]. The introduction of antibiotics in the 1930s decreased the demand for serum therapy [338]. During this period, however, the experimental basis for combination therapy with antimicrobials and hyperimmune animal serum was established [339], [340].

A wide variety of biological products is now available for immunotherapy, the most important being purified gamma globulin for intramuscular injection (standard immune serum globulin, ISG) and globulin for intravenous injection (standard intravenous immunoglobulin, IVIG). Hyperimmune globulins with a high antibody titer against specific pathogens are also used. Additional preparations include antitoxins, plasma, and other blood products.

5.1. Gamma Globulin Preparations

5.1.1. Standard Immune Serum Globulin

Historical Aspects. In 1936 Arne Tiselius separated serum proteins into four major fractions by electrophoresis; subsequently he and Kabat found that immunoglobulin occurred predominantly in the gamma electrophoretic fraction [341]. Cohn and colleagues devised a procedure for recovering immunoglobulins (gamma globulins) from serum on a large scale.

The serum proteins were fractionated by precipitation with ethanol under carefully controlled conditions of pH, temperature, protein concentration, and ionic strength [336], [342]. This process enriched and stabilized the gamma globulin from plasma at a relatively uniform antibody content while also denaturing most viruses [334]. Such gamma globulin prepared from large pools (>500 donors) of donor plasma were used in the treatment of infectious diseases during World War II [343].

Shortly after World War II gamma globulin was shown to contain antibody titers adequate for the prevention or attenuation of measles [344], infectious hepatitis [345], and polio [346]. Since the report of hypogammaglobulinemia in 1952, and the demonstration that gamma globulin administered on a monthly basis decreased the incidence of infection [347], antibody replacement of this deficiency has been routine.

Properties of Gamma Globulins. Gamma globulins occur at a serum concentration of 600 – 1200 mg/100 mL in adults, they represent approximately 11 – 14 % of total serum proteins [341] and 80 % of serum antibody [348]. Immunoglobulin G has a half-life in the normal circulation of approximately 25 days (35 – 40 days in patients with agammaglobulinemia) and is synthesized by adults at a daily rate of 35 mg/kg body weight [348]. Its rate of synthesis is regulated by serum IgG levels. Immunoglobulin M and IgA have lower serum concentrations and shorter half-lifes than IgG. The molecular mass of IgG has been estimated to be 145 000 [341] with 2.5 wt % being carbohydrate that is associated with the heavy chain (see also Fig. 5).

The IgG isotype can be divided into four subclasses (Table 3). The most abundant is IgG 1 (60 – 70 % of total IgG) which binds to the Fc receptors of neutrophils and mononuclear cells and to the first component of complement [349]. The IgG 2 subclass (20 – 30 % of total IgG) activates the classical complement pathway poorly but can activate complement by the alternate pathway. It is more resistant to proteolysis than the other subtypes [348]. The antibody response to pneumococcal and hemophilus polysaccharides may be related to the preimmune levels of this subclass [350]. The subclasses IgG 3 and IgG 4 bind to the Fc receptors of phagocytes (mononuclear cells and neutrophils) and basophils, respectively [349]. The IgG 3 has a short half-life (nine days); IgG 4 is unable to bind complement.

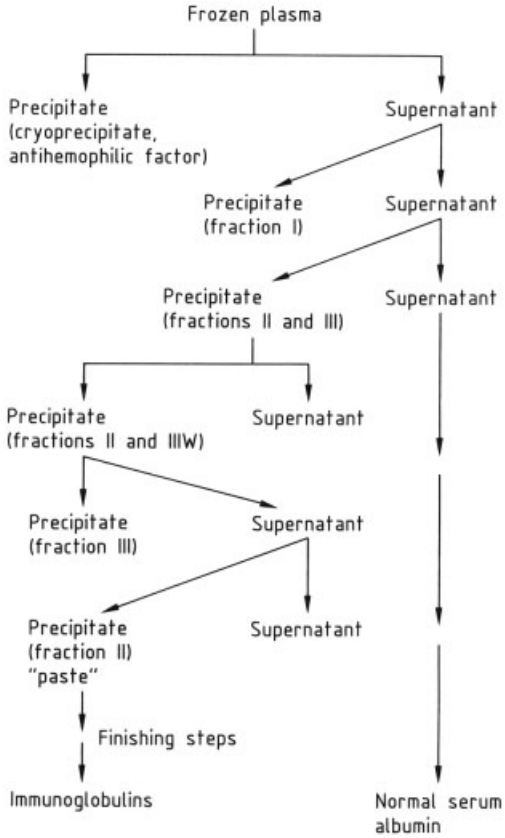

Figure 5. Preparation of immunoglobulin by Cohn – Oncley fractionation ("6/9 method")
Cohn fractionation of plasma by cold – alcohol procedures yields a fraction II precipitate or "paste". This is followed by applying the Oncley procedure to Cohn fraction II material to give immunoglobulin. The procedure is based on solubility differences that depend on ethanol concentration, ionic strength, pH, temperature, and protein concentration. Precipitates are usually collected by centrifugation.

Preparation of Immune Serum Globulin. Standard immune serum globulin (ISG, gamma globulin) is prepared from the plasma of pools of >1000 donors (see Fig. 5); → Blood, Chap. 2.2.5.). This minimizes differences between individual antibody levels to specific antigens and ensures a broad range of antibody specificities.

Table 3. Immunoglobulin G subclasses *

Property	IgG 1	IgG 2	IgG 3	IgG 4
Percentage of total IgG (adult)	60 – 70	20 – 30	5 – 10	<5
Molecular mass (daltons)	146 000	146 000	170 000	146 000
No. of interchain disulfide bonds	2	4	11	2
No. of amino acids in hinge region	15	12	62	12
Half-life (days)	21 – 23	20 – 23	7 – 8	21 – 23
Fc receptor binding				
mononuclear cells **	++	+	++	±
neutrophils	++	±	++	+

* Adapted from [239].
** ++ strong binding; + detectable binding; ± binding detected in some, but not all studies.

Standard immune serum globulin is usually prepared as a 16.5 % injectable solution (165 mg per mL) and represents a ca. 20 – 25-fold concentration of plasma IgG [336], [343]. This provides an effective dose of antibody in a relatively small volume [343]. Because of the viscosity of the preparation, it can only be given intramuscularly or subcutaneously [343], usually through a large gauge needle (16 – 18 ga). It has been estimated that 1 g of ISG contains 4×10^{18} antibody molecules with $>10^7$ specificities [334]. The product, although highly stable at 4 °C, can still undergo proteolysis, presumably due to plasmin contamination [336]. The ISG may also contain blood group substances, IgA, and IgG dimers. The presence of IgA can result in anaphylactic reactions in patients who lack IgA [336].

5.1.2. Immunoglobulin for Intravenous Use

Intravenously administered gamma globulin is needed to allow greater patient comfort, increase acceptability, and provide larger quantities of antibody. This is particularly important for patients who have a small muscle mass (children) or insufficient skin surface (burns), who are at risk from uncontrollable bleeds (Wiskott-Aldrich or other bleeding diatheses), or who need either large repeated doses (immunodeficient) or rapid onset of peak levels (intoxicated patients). Furthermore, intravenous administration avoids the local degradation of ISG antibody that may occur at the injection site [351].

In 1948 Cohn administered 25 – 50 mL of ISG intravenously and saw no adverse effects [336], [352]. Janeway and Cohn repeated the experiment with a new preparation and induced severe reactions after only 2 mL. This reaction was attributed to contamination of the preparation by staphylococcal enterotoxin [352]. The intravenous administration of ISG was repeated by Janeway in 1970 [336], and induced severe cardiovascular (tachycardia, arrhythmias, hypotension, severe chest pain), tachypneic, and pyrogenic (fever, chills, malaise) reactions [343]. These adverse effects were attributed to the presence of immunoglobulin aggregates that activated complement [353], [354].

To avoid the above-mentioned severe reactions, methods were introduced to rid the preparations of the high molecular mass immunoglobulin aggregates. Although they could be removed by centrifugation, reaggregation usually occurred and this method was not practical on a large scale [355]. Consequently, attention was turned to chemical modification.

Various methods of chemical modification have been tried. Degradation with the enzymes pepsin and plasmin resulted in the formation of antibody fragments with immunologic activity [336], [356]. In addition, since the proteolytic enzyme was not removed, it may have continued to be active in the absence of lyophilization; porcine pepsin also induced antibody formation [354]. Reaggregation was also prevented by acidification with hydrochloric acid [357] and reduction – oxidation with dithiothreitol, iodoacetamide, sulfite, or tetrathionite. Alkylation and acylation of immunoglobulin has also been accomplished with β-propiolactone.

The IVIG preparations that have been reduced and alkylated or otherwise modified may have impaired complement binding and altered subclass distribution [358], impaired opsonic activity in vitro [359], shortened serum half-life, and decreased protective efficacy [360]. Con-

sequently, a new generation of native or intact IVIG preparations has been produced without modification by methods which include adjustment to pH 4, use of poly(ethylene glycol), ethanol precipitation, ultra- or diafiltration, and ion-exchange chromatography [353]. Four unmodified products are licensed in the United States: one prepared by adjustment to pH 4 in the presence of traces of pepsin, one prepared by ion-exchange chromatography and ultrafiltration, and a third by diafiltration, ultrafiltration, and adjustment of pH to 4 – 4.5. A fourth product (Venoglobulin-1 from Alpha Therapeutic Corporation) licensed in the United States but produced in Japan is prepared by poly(ethylene glycol) fractionation and ion-exchange adsorption.

Commercially prepared IVIG is usually stabilized with a mono- or disaccharide, such as 10 % maltose [338] or glucose. This considerably decreases the incidence of side effects that accompany the infusion of IVIG [336], [357], [361], perhaps by minimizing precipitation or aggregation. Maltose can cause a mild diuretic effect [362].

5.1.3. Hyperimmune Globulins and Antitoxins

Hyperimmune globulins are high-titered preparations of ISG against a specific antibody. They are prepared from the plasma of patients who have recently recovered from the disease or who have a high titer as a result of a previous vaccination or natural infection. Antitoxins for use in intoxication with botulinus or diphtheria toxin are prepared from snake venom or in horses [363].

5.1.4. Production Requirements

The WHO requirements stipulate that current immunoglobulin products

1) are prepared from pools of >1000 donors;
2) are free of kinins, plasmin, and prekallikrein activity;
3) have a low IgA content;
4) are as free as possible from immunoglobulin aggregates;
5) contain at least 90 % intact IgG without fragments;
6) are as unmodified as possible so that opsonic, complement, and other biologic activities are maintained;
7) contain all IgG subclasses;
8) have high levels of antibody to at least two bacterial species or toxins and two viruses (to be ascertained by neutralization tests), and
9) contain at least 0.1 International Units (IU) of antibody to hepatitis B and have a 1 : 1000 titer to hepatitis A [364].

In addition, manufacturers should specify the diluent and any chemical modifications used. Finally a product can only be classified as a hyperimmune immunoglobulin if the antibody level is five times that of standard ISG preparations.

There is no consensus on which laboratory tests should be used to predict the safety of IgG (contact activation, anticomplementary activity) [336]. Safety requirements should, however, consider the IgG half-life and virus transmission (particularly of human immunodeficiency virus HIV, but also of hepatitis B and non-A, non-B hepatitis viruses) [353].

5.2. Prophylaxis with Immune Serum Globulin

Passive immunization with ISG has been recommended for the short-term prevention of disease when vaccines for active immunization are unavailable or when active immunization was not given before disease exposure; it should be given before expected contact or early in the disease incubation. In these situations active immunization is always preferable because the antibodies subsequently formed by the vaccine provide longterm protection. Immune serum globulin is also used for antibody replacement in patients who lack adequate serum levels of immunoglobulin (i.e., hypogammaglobulinemia) [335].

The efficacy of ISG for the short-term prophylaxis of specific infections was established shortly after its widespread availability in the 1940s. These include:

Hepatitis. Stokes and Neefe showed that ISG could prevent or modify the course of hepatitis A [335], [336], [345], but Krugman found

that ISG modified, but did not prevent this disease [365]. The ISG dose for treating hepatitis A is 0.02 mL/kg body weight [336]. A hyperimmune antibody preparation is also available commercially. In the case of hepatitis B, ISG is not recommended. Instead, hyperimmune globulin is given (see Section 5.3). Although ISG can decrease the incidence of post-transfusion hepatitis [366], its use for non-A, non-B hepatitis is considered optional [364].

Measles. The ability of convalescent serum to modify the course of measles was demonstrated in 1907 [335]. The ISG has also been shown to prevent or attenuate the disease [343]. The introduction of active immunization with a measles vaccine in 1963 (see Section 3.1) significantly decreased the incidence of this disease in the United States. Standard ISG is now recommended for infants under one year of age or for immunodeficient patients within six days of acute exposure to a case of measles [364].

Polio. If given early, ISG can modify the paralytic complications of polio [335], [346]. An unexposed individual should receive 0.15 mL/kg body weight.

Prevention of Infection in Patients with Hypogammaglobulinemia. An intramuscular ISG dose of 100 mg/kg given every 3 – 4 weeks is recommended for patients with hypogammaglobulinemia. This maintains a serum level of circulating IgG above 200 mg/mL [336] and confers protection against a wide variety of infections [342]. This minimum recommended dose is based on a study by the British Medical Research Council conducted between 1956 – 66 on 176 patients [367]. Although serum IgG levels at this dose rarely rise to the normal range [368], total replacement of IgG does not appear necessary for preventing infection [369]. Although a monthly dose of 200 mg/kg was found superior to 100 mg/kg, most patients did not tolerate more than the 100 mg/kg [361]. Two other studies also showed that higher doses of ISG decreased the incidence of acute infections [334]. Individualization of doses and their frequency has also been suggested [368].

Rubella. Standard ISG is unreliable in the modification of rubella [335], [337], [365]. High titers of antibody are needed [365]. Nevertheless, the use of ISG is recommended for women exposed to the disease during early pregnancy [364].

Clinical Studies. On the basis of data available in 1980, a committee of the WHO observed that it was "inappropriate" to use standard immunoglobulin for the prevention of infection in premature infants, during the physiologic hypogammaglobulinemia of infancy, or for malnutrition. Its use is contraindicated in patients with selective IgA deficiency [364]. Previous attempts in the 1960s showed that ISG did not prevent infection in a variety of clinical situations [370], [371] including multiple myeloma [372]. Although ISG was unable to prevent upper or lower respiratory tract infections in children or institutionalized, elderly adults, a significant decrease was seen in the incidence of mumps in children and fevers of unknown origin in adults [373]. Its use in the prevention of infection in burned patients has yielded conflicting data [335]. In a study on Peruvian children with burns over at least 10 % of their body surface area, there was a 41 % incidence of septicemia and 15 % mortality in control patients compared to a 21 % incidence of septicemia and 6 % mortality among those who received either plasma or ISG [374]. However, in another study in the United States administration of ISG did not beneficially affect either the rate of septicemia or mortality in burned patients [375]. More recently, patients admitted to a burn unit in India were randomized into groups that received active immunization against *Pseudomonas aeruginosa*, passive immunization with a cold ethanol precipitate of plasma from normal immunized volunteers, both immunologic treatments, or neither. In children, but not adults, there was a significant reduction in mortality following a daily dose of just over 20 mg protein for three days. There was a decrease in the incidence of not only *P. aeruginosa*, but also other gram-negative bacilli among those treated immunologically [376].

The ability of ISG to prevent infections has been most extensively studied in high-risk, premature infants. This population has a hypogammaglobulinemia that exposes them to a high incidence of bacterial infection [377]. The more premature the infant the lower the IgG, since most of

the transplacental transfer of IgG occurs during the last six weeks of gestation [370]. Children do not attain adult levels of immunoglobulin until two years of age. In two studies using 0.5 – 3 mL/kg, no prophylactic effect was observed [371], [378], although there was a suggestion that the high dose regimen may have had a beneficial effect.

5.3. Prophylaxis with Hyperimmune Globulins

Several hyperimmune globulins are commercially available or undergoing development. Hyperimmune IVIG preparations are discussed in Section 5.5.

Diphtheria [342]. Diphtheria antitoxin is obtained from the blood of horses immunized against diphtheria toxin. Individuals should be tested for sensitivity to horse serum before being given the product. For prophylaxis, 1000 – 5000 IU of antitoxin is administered to Schick-positive individuals exposed to diphtheria. Higher doses (20 000 – 80 000 IU) are given for treatment.

Hepatitis. A hyperimmune globulin is available for hepatitis A. In the case of hepatitis B, a hyperimmune globulin is administered after mucosal or percutaneous exposure (including sexual contact) to an antigen-positive individual [336]. This is also recommended for newborns of antigen-positive mothers [365].

Mumps. Immune serum globulin has no efficacy in the modification of this disease, but hyperimmune globulin does [343].

Pertussis. Hyperimmune globulin modifies the disease. For example, 2.5 mL of hyperimmune antipertussis gamma globulin with a followup dose at 5 – 7 days led to a 75 % reduction in disease among exposed, nonimmune individuals [343]. However, use of hyperimmune globulin has been superseded by antibiotics [379].

Rabies. Hyperimmune globulin is recommended in addition to active immunization following exposure to a possible or proven case of rabies [364] (see Section 3.7). Combined active and passive immunization against rabies have been shown to be superior to active immunization alone [336].

Rho(D) immune globulin. This hyperimmune globulin is recommended for rhesus-negative mothers who deliver rhesus-positive infants [364].

Tetanus. The efficacy of tetanus immune globulin in either the prophylaxis or treatment of tetanus has not been clearly shown in clinical study. Current recommendations are for 250 IU injected intramuscularly for individuals whose tetanus immunization is not known and whose wound is of a sufficiently serious nature. For treatment of clinical tetanus, doses of 3000 – 6000 units are recommended.

Vaccinia. Hyperimmune globulin is used for prophylaxis against smallpox and for treatment of the dermal complications of vaccination [343].

Varicella. The efficacy of standard ISG in the prophylaxis of this disease is not well-established [335], [365]. Use of the hyperimmune product, however, is indicated in individuals who have never had chicken pox, who are exposed to acute cases and belong to a high risk group (newborns or immunocompromized patients), and women who are pregnant [364].

Hyperimmune Globulins under Development. Globulins with high titers to cytomegalovirus, and *Pseudomonas aeruginosa* have been tested. A product (bacterial polysaccharide immune globulin) has been tested that is prepared from the sera of donors immunized with licensed vaccines against *Hemophilus influenzae* (type b), pneumococci, and meningococci.

Certain subpopulations of children, such as native American Indians and Eskimos, are at particularly high risk of acquiring serious infection with encapsulated bacteria. Standard ISG preparations have failed to show a beneficial prophylactic effect for many types of infection. Immune serum globulin made hyperimmune to polysaccharide antigens by immunizing volunteers with licensed vaccines against pneumococci, *H. influenzae*, and meningococci has been

used to immunize Apache infants. There was a significant reduction in the incidence of systemic disease caused by *H. influenzae* and pneumococci during the first six followup months as well as a significant decrease in the incidence of bacteremia [380].

5.4. Therapy with Immune Serum Globulin

Data suggest ISG is active in bacterial infections in animal models [379], [381] and may be synergistic with antibiotics [339], [340], [382]. However in 1968 Schless and Harrell [379] and others [336] observed that there was little evidence to support its therapeutic efficacy in established infection in humans; they suggested the need for a controlled clinical trial of ISG in the treatment of systemic infection in patients that were not deficient in antibody [379]. Indeed, since the most functionally-active antibody to gram-negative bacteria is IgM and not the IgG found in ISG, little benefit was to be expected [342]. The ISG is active against a wide variety of human pathogens in animal models of infection [379]. When used with antimicrobial agents, ISG has possible benefits in experimental infection with a wide variety of organisms [382], [383] and in clinical infections in humans [339], [340], [384], [385]. Large daily doses of ISG administered intravenously for 10 days to patients with leukemia and fever was well tolerated [386]. However, no benefit was found for patients who received the ISG in addition to antibiotics (compared to antibiotics alone). The administration of ISG to shorten the course of infection in children under two years of age was ineffective [387].

5.5. Prophylaxis and Therapy with Intravenous Immunoglobulin (IVIG)

In studies on the prophylactic or therapeutic efficacy of ISG, the possibility that higher doses of standard immunoglobulin might improve efficacy was a recurring theme. With the availability of IVIG, larger volumes of immunoglobulin could be administered directly into the bloodstream (see Section 5.1.2). In addition, with the ability to screen large numbers of samples for antibody levels or to immunize volunteers with an increasing number of vaccines, an increasing number of publications have examined the efficacy of passive immunotherapy, particularly with hyperimmune IVIG preparations, in the treatment and prophylaxis of infectious diseases.

Intravenous immunoglobulin G is effective in the prevention of infection in patients with hypogammaglobulinemia [358]. It has also been shown to be effective for treating chronic infection in such patients who developed sinopulmonary infection despite ISG maintenance therapy [351].

5.5.1. Viral Infection

Since cytomegalovirus (CMV) is a frequent cause of infection in patients undergoing organ transplantation and effective antiviral therapy is lacking, there has been considerable interest in the use of IVIG that is hyperimmune in CMV antibody for both the prophylaxis and treatment of CMV infections.

Prophylaxis. Hyperimmune CMV-IVIG (total dose 550 mg/kg) was shown to decrease the attack rate of symptomatic CMV infection among CMV-negative patients who received kidneys from CMV antibody-positive donors from 60 – 21 % [388]. This treatment also decreased the incidence of fungal and parasitic infection. Hyperimmune CMV-IVIG (3×200 mg/kg doses) also prevented mortality and interstitial pneumonia from CMV for 120 days in leukemic patients who underwent bone marrow transplantation [389]; however in another study it did not prevent acquisition of infection [390]. In another trial, nonimmune IVIG with a high antiCMV titer did not decrease the incidence of CMV seroconversion in bone marrow transplantation patients [391]; however, the incidence of symptoms and interstitial pneumonia decreased. In contrast, CMV-IVIG given to CMV-negative patients undergoing bone marrow transplantation had no effect on either the prevention of new disease or the amelioration of established disease [392]. Therapy of bacterial infections following infection with human immunodeficiency virus are discussed in Section 5.5.2.

Therapy. Immunoglobulin has been used sporadically in the treatment of viral disease. High-titered CMV-IVIG from screened donors did not show efficacy in the treatment of bone marrow transplant patients with documented CMV infection [393]. IVIG has been shown to alter the course of echovirus encephalitis infection in three patients with hypogammaglobulinemia. Although no benefit occurred from giving the IVIG intravenously in two of these patients, intraventricular administration resulted in clinical cures [394], [395]. The use of IVIG was unable to alter the lethal course of polymyositis secondary to echovirus in another patient with hypogammaglobulinemia [396]. IVIG has also been used experimentally in the successful treatment of herpes infection in mice [397].

5.5.2. Bacterial Infection

Numerous studies have demonstrated the efficacy of hyperimmune IVIG in the prophylaxis and, if used early after infection, the treatment of bacterial infection. In 1943 ALEXANDER showed that the combination of a sulfa drug and animal hyperimmune sera was more effective in reducing mortality than either agent alone [339], [340]. A number of studies have demonstrated the efficacy of IVIG in both the prevention and treatment of experimental infection with *H. influenzae* in neonatal rat models [398]; with *E. coli* [399] and *Klebsiella* [400] in mouse models; and with *P. aeruginosa* in neutropenic and burned rodent models [401], [402]. Intravenous immunoglobulin hyperimmune to group B streptococcal surface antigens can prevent and treat serious bacteremia in experimental infection in monkeys [403].

Prophylaxis. Earlier data with ISG in both experimental bacterial infection in animals and in clinical infection in humans indicated that in some situations (e.g., patients with hypogammaglobulinemia), exogenous standard gamma globulin could prevent the acquisition of serious bacterial infection. Similar studies with IVIG have established the efficacy of these preparations [358].

Since earlier investigators believed that larger doses of ISG might be effective in the prevention of bacterial infection in high-risk neonates, it is not surprising that similar studies have now been reported with IVIG. In one nursery with a high rate of infection, a single dose of IVIG (120 mg/kg within 2 h of birth) decreased acquisition of infection and mortality in preterm, lowbirthweight neonates [404]. A second dose at eight days, conferred no further advantage. Prophylaxis with IVIG (0.5 mg/kg/week for four weeks) was significantly better in preventing infection and death in neonates, but only in the subpopulation that weighed less than 1500 g and had a gestational age of less than 34 weeks [370]. In another study, antibiotics were given either alone or with IVIG to women 27 – 36 weeks pregnant who had chorioamnionitis. Only high doses (24 g/d for five days) of IVIG given after the 32nd week of pregnancy prevented infection in the delivered babies [405]. The investigators concluded that little transplacental transfer of IgG occurred before the 32nd week of gestation.

Data on the use of standard IVIG for the prevention of bacterial infections in patients without hypogammaglobulinemia is limited. High levels of specific antibody are needed; these may not be found in standard, nonimmune IVIG, despite the possibility of delivering larger amounts of IVIG than was the case with ISG. Hyperimmune products have been shown to prevent specific infections in experimental models (see Sections 5.4, 5.5.2, and 5.6). The administration of nonimmune IVIG (1000 mg/kg) before bone marrow transplantation and weekly for 17 weeks thereafter had no effect on the acquisition of either bacterial or fungal infection [406]. Clinical trials are being performed to test the efficacy of IVIG in the prophylaxis of bacterial infection in adults with chronic lymphocytic leukemia, a condition which may be complicated by hypogammaglobulinemia [407].

Therapy. Patients with human immunodeficiency virus (HIV) infection (i.e., acquired immune deficiency syndrome, AIDS) have a dysfunction in their humoral immune system which includes both a decreased antibody response to bacterial antigens and an altered distribution of IgG subclasses [408]. Unlike adults, who tend to acquire opportunistic infections (infections in which immunodeficient individuals are infected by organisms that are normally withstood by immunocompetent individuals), young children and particularly infants infected with

HIV often resemble patients with primary humoral immunodeficiency and tend to suffer from bacterial infections. Consequently, prophylaxis with monthly doses of IVIG has been used in the management of these patients. In one pilot study, a decreased incidence in episodes of fever and bacteremia was noted. This was accompanied by clinical improvement, prolongation of life, and improvements in other immunologic parameters [408]. In a 37 month old child with AIDS, monthly doses of IVIG produced increases in IgG 2 antibody and antibody to 12 pneumococcal serotypes, and prevented subsequent episodes of bacteremia [409]. On the basis of these preliminary data, IVIG prophylaxis has been advocated in the treatment of childhood AIDS [408]. Periodic administration of IVIG decreased lactic dehydrogenase activity (a proposed indicator of pulmonary interstitial inflammation) in adults and children with HIV infection [410].

Attempts to demonstrate a significant effect in the therapy of bacterial infections with IVIG have been much less successful than with its use in prophylaxis. This may be attributable to the shorter half-life of IVIG in the blood during infection [368], [381]. The addition of IVIG to standard regimens of antibiotics decreased the mortality from documented bacteremia, particularly among preterm infants. The number of subjects studied was too small for statistical analysis, however [411].

In adults, little data is available to support the use of IVIG in the treatment of bacterial infections. This may be due to the need for high levels of antibody specific for the invading organism (rather than simply high levels of nonspecific antibody) as well as the need for prompt initiation of therapy. In most experimental studies with hyperimmune IVIG, little benefit can be shown if the exogenous immunoglobulin is given more than 8 h after infection (see [381], [383]). In clinical medicine, however, identification of the time of onset of infection is often difficult.

5.5.3. Noninfectious Diseases

5.5.3.1. Therapeutic Effect of IVIG

In 1981 it was reported that a patient who received IVIG for hypogammaglobulinemia also recovered from bleeding secondary to a coincidental platelet deficiency (*thrombocytopenia*) following the infusion [412]. Following this fortuitous observation, experiments showed that a high dose of IVIG could indeed reverse thrombocytopenia in the absence of hypogammaglobulinemia; furthermore, a regimen of 400 mg/kg/d for five days was effective treatment for idiopathic thrombocytopenic purpura (ITP) in children [413]. The use of IVIG and oral steroids was also compared. Among those who responded rapidly to treatment (62 %), IVIG was as efficacious as steroids; however the slower responders responded better to the IVIG treatment. Administration of IVIG was accompanied by a doubling of serum IgG; IgM levels rose under both treatment regimens. The IVIG may possibly reduce clearance of antibody-coated platelets via an Fc-mediated mechanism (see Section 5.5.3.2) [414]. The dosage regimen used in the above study (400 mg/kg/d for five days) has been used in many other studies on the noninfectious uses of IVIG. However, in another study 800 – 1000 mg/kg was given as a single infusion to children with ITP with similar results and no reported untoward effects [415]. Experience with IVIG for ITP has not been uniformly successful however [416].

IVIG has been used in the treatment of *immunologically-mediated blood disorders* in both pediatric and adult populations. These include autoimmune hemolytic anemia [417], [418], autoimmune neutropenia [419], post-transfusion purpura [420], ITP in adults [421], [422], chronic ITP in adults and children [423], [424], thrombocytopenia secondary to alloimmunization in leukemic patients receiving platelet transfusions who became refractory to subsequent platelet transfusions [425], thrombocytopenia secondary to transplacental passage of antiplatelet antibodies [426], ITP of pregnancy [427], antibody-mediated red cell aplasia [428], during pregnancy in women with severe rhesus immunization [429], and in conjunction with cyclophosphamide used to treat antibody to factor VIII in hemophilia [430].

In addition to these hematologic disorders, two large trials conducted in patients with *Kawasaki's disease* have compared the use of aspirin alone to that of aspirin and IVIG (400 mg/kg/d for 4 and 3 days) [431], [432].

The IVIG reduced the fever and the incidence of coronary artery disease [431], [432]. It suppressed the T and B cell activation characteristic of patients with this disease and decreased the levels of spontaneous immunoglobulin synthesis in vitro [433].

Placentally derived immunoglobulin has been administered to patients with severe *rheumatoid arthritis* because the symptoms of some patients improved during pregnancy [434]. None of five patients given IVIG derived from control plasma improved whereas 3/5 improved under the placentally derived globulin. An antihuman leukocyte antigen – DR surface antigen (anti HLA-DR) antibody in the placental preparation was presumed to be a possible mechanism for this improvement.

High dose immunoglobulin treatments have also been used in patients with *Felty's syndrome* [435], *myasthenia gravis* [436], *epilepsy* [437], and *multiple sclerosis* [438]. In the latter study one-third of patients worsened with IVIG therapy.

5.5.3.2. Mechanism of Action

Many immunologic mechanisms have been invoked to explain the beneficial effects of IVIG in noninfectious illnesses. IVIG is a potent immune modulator; the role of the Fc portion of the immunoglobulin molecule in this respect is still a subject of active investigation.

Antibody-coated material is removed by phagocytes in a process known as *Fc-receptor-mediated clearance.* The Fc portion of the antibody coating the bacterium cell, protein etc. binds to the Fc receptor of the phagocyte. The material is then taken up by the phagocyte and digested. Initially, the Fc receptors in the reticuloendothelial system were thought to become saturated by the high dose of immunoglobulins in IVIG resulting in a decreased clearance of IgG-coated particles due to competitive inhibition [417], [439]. This is consistent with studies in which radiolabelled, autologous erythrocytes were cleared more slowly after IVIG infusion than before [414], [423]; this treatment altered Fc receptor affinity, not receptor number [440]. Other suggested mechanisms include decreased synthesis of autoantibodies [441], the clearance of persistent, occult viral infections [396], [441], the protection of platelets or megakaryocytes against antiplatelet antibodies [441], antiidiotypic suppression of antibody synthesis [436], [442], blocking of the Fc receptor by antibodies, and the production of antilymphocyte antibody [443]. However, patients with ITP have responded to IVIG without demonstrable alteration in Fc-receptor mediated clearance [441]. In addition, preincubation of erythrocytes in IVIG failed to inhibit the phagocytosis of sensitized erythrocytes by cultured macrophages [444]. Thus, IVIG may lead to improvement in such patients by multiple mechanisms.

IVIG is a potent modulator of antibody production both in vitro and in vivo. It inhibits lectin-driven B cell differentiation in vitro [445], [446] and immunoglobulin production by peripheral blood mononuclear cells stimulated with pokeweed mitogen [446]. For these effects the Fc portion of immunoglobulin must be present: the Fc portion alone was 100 fold more effective than the intact IgG. The observation that IgM antibody rises after IVIG infusion has raised the possibility that it may stimulate some immunoglobulin-producing cell populations [423]. Interestingly, the monthly administration of ISG to premature infants during the first year of life resulted in a significantly lower level of immunoglobulin compared to the control group [371], [447].

In adults with ITP, the infusion of IVIG (400 mg/kg) led to a decrease in T 4 (helper) lymphocytes and elevation in T 8 (suppressor) lymphocytes (and decrease in T 4/T 8 ratio) [421]. Similarly, the administration of IVIG to patients with hypogammaglobulinemia led to increased suppressor T cell activity, decreased total T cells, a significant decrease in the T 4/T 8 ratio, and a decrease in lectin-induced immunoglobulin syntheses in vitro [448].

The ability of IVIG to decrease antibody production may be desirable in patients producing autoantibody, but dangerous in those with infection [445]. In patients with acute otitis, repeated monthly [379] infusions of IVIG resulted in higher IgG levels if the initial antipneumococcal antibodies were low. However, in the presence of high initial levels of antibody, specific antipneumococcal antibody levels decreased [449]. The suggestion has also been made that high levels of nonspecific antibody could lead to a reduction in the survival of specific antibody [450],

[451]. Evidence suggests that IVIG may even exacerbate infections. A patient with neutropenia died due to acceleration of the yeast infection shortly after infusion of IVIG; the IVIG may have blocked the normal Fc-mediated clearance mechanism [452].

Immunoglobulin G can bind both native C 3 and C 3 b during complement activation by soluble immune complexes and by bacteria [453], [454]. Finally since immunoglobulin has unique antigenic determinants, antibodies to the immunoglobulin may develop [455], [456]. These antibodies may play a role in the anti-idiotype network.

In summary, exogenous immunoglobulin has a diverse, potent effect on a wide variety of immune regulatory mechanisms. These interactions may often work to the patient's benefit, but may also result in previously unsuspected adverse effects.

5.6. Prophylaxis and Therapy with Plasma and Other Blood Products

Although the regular, large-scale, clinical use of plasma and other blood products is considered impractical (both for production and safety reasons), passive immunotherapy with these agents may provide some efficacy in the prophylaxis and treatment of infections in humans.

Plasma has been used to treat infections, particularly those caused by *Pseudomonas* [374], [457]. Plasma was superior to ISG in the prevention of infection among patients who suffered >30 % burns [457]. Similar observations were made during experimental infection [458]. One possible explanation for the higher efficacy of plasma is that it contains IgG, IgA, and IgM (ISG contains only IgG). This may improve the distribution of antibody at different anatomic sites. Since the isotypes have functional differences, the antibacterial activity is also increased (natural antibody against gram-negative bacilli is thought to be predominantly of the IgM isotype). However, plasma infusions carry the risk of transmitting hepatitis (there is no manufacturing process for plasma that inactivates the virus as is the case with ISG).

Further efforts in the use of passive immunotherapy have also focused on septic shock. The rapid infusion of large volumes of lyophilized human plasma from blood with >40 μg/mL of antibody to a mixture of lipopolysaccaride serotypes resulted in nearly 7-fold decrease in mortality in South African women treated in an obstetrical–gynecological ward for septic shock [459]. Equine plasma similarly screened for antilipopolysaccaride antibodies has also been used routinely in veterinary practice in South Africa [460].

Opsonins. The defective neonatal antibody (opsonic) response that mediates the uptake and killing of bacteria by phagocytes can be partially corrected with gamma globulin [461]. The ability passively administered opsonins to prevent infection was first suggested in one study in which fresh whole blood was administered to infants to prevent group B streptococcal sepsis. All nine infants transfused with blood having antibody to group B streptococci lived versus 3/6 transfused with blood having undetectable antibody titers. Protection was correlated with opsonic antibody titers in the infants' blood and with having >40 % of their blood volume replaced [462].

Postimmune Serum. Donors were immunized with a J 5 mutant of *E. coli* to elicit antibodies to widely shared core epitopes in the lipopolysaccharides of gram-negative bacteria. The passive administration of postimmune serum from these individuals to patients in bacterial endotoxic shock decreased the incidence of death as well as the need for pressor therapy in patients with profound shock. However, since this protection did not correlate with the presence of antibody to the core epitopes, it is not clear whether the protection was conferred by the antibody [463]. In a follow-up study, the prophylactic administration of postimmune plasma appeared to significantly prevent shock and death from gram-negative infections in patients admitted to a surgical intensive care unit. Such treatment had no effect on infection rate, however [464]. In both studies the protective moiety in the plasma and serum was presumed to be antibody to the endotoxin. In any event, for this to be a practical therapy, the protective portion of the postimmune product must be mass produced in the form of a safe, standardized preparation (e.g., made into an IVIG from

a large pool of donors immunized with the J 5 vaccine or into a monoclonal antibody preparation). Attempts to show a protective effect from an IVIG prepared from immunized donors have been unsuccessful to date [465]. Nevertheless, monoclonal antibody preparations directed toward a J 5 epitope were evaluated in human volunteers (see also Section 6.1).

5.7. Adverse Effects of Gamma Globulin Preparations

Immune serum globulin G is one of the safest biological products available with an overall incidence of reactions of 3 – 12 % [334]. Rare systemic (anaphylactic) reactions may occur during or within minutes of administration of ISG (ca. 1/500 – 1/1000 injections). Late-occurring systemic reactions (within hours or days) include arthralgias, pyrexia, and diarrhea and are not uncommon in immunodeficient patients [364]. Local reactions occur with intramuscular administration of ISG and are related primarily to the large volume administered. Local reactions to ISG can be prevented by using small volumes or by pretreatment with analgesics. Systemic reactions may be modified by the use of aspirin, hydrocortisone, or antihistamines.

Systemic reactions with IVIG are more common than with ISG and often depend on the rate of infusion and the type of preparation. Patients receiving IVIG for hypogammaglobulinemia have a reported incidence of 2.5 % reactions which are usually mild [358]; the incidence of these reactions decreases after the first two months of therapy. The cause of reactions to IVIG are not clear but may depend on aggregate formation, IgA contamination, and activated enzymes that initiate the release of inflammatory mediators [364].

The experimental work discussed above and clinical experience give reason for caution in the use of immunoglobulin preparations. Clinical experience primarily involves concern about viral transmission, anaphylactic reactions among patients with IgA deficiency, and isolated case reports of unusual occurrences. The indications for the use of gamma globulin should be well founded [363].

Viral Transmission. Immune serum globulin G has had an outstanding safety record with regard to lack of transmission of virus to recipients [466]. For example, ISG is free of serum hepatitis, even if the virus occurs in the original plasma. Ten children who received plasma for prophylaxis against measles in England became jaundiced and three died. In contrast, when ISG prepared from the same plasma was given to 56 children, only one child became jaundiced [343]. In the United States, 4 out of 15 volunteers who were inoculated subcutaneously with hepatitis-positive plasma developed hepatitis. Five further volunteers received large doses of ISG and none developed hepatitis [343]. The ability of the Cohn – Oncley fractionation (see Fig. 5) to inactivate virus may in part be due to the high concentration of antivirus antibody [353]. Non-A, non-B hepatitis has been reported in pastes produced in this procedure: in one, the lyophilization step was omitted, and in two others an ion-exchange step was added to lower IgA but it lowered IgG 4 levels as well [353].

In contrast, non-A, non-B (NANB) hepatitis has been reported following IVIG infusion [467–469]. In one such episode 16 out of 77 patients who received IVIG for immunodeficiency acquired hepatitis with death occurring in 5 cases. Interestingly, no hepatitis was observed following administration of ISG prepared from the same serum pool [467]. Since neither of the IVIG products was subjected to a recognized virucidal finishing step, such as treatment with acid, β-propiolactone, or ultraviolet radiation, the virus may still have been viable. Thus ethanol fractionation by itself might not fully inactivate putative virus [470], though it is virucidal for enveloped viruses [471]. Supplementary virucidal procedure could make IVIG safe from NANB hepatitis transmission [471]. Unlike plasma, IVIG has not been known to transmit hepatitis B [472].

There have been no confirmed cases of transmission of HIV virus in IVIG but antibody to HIV has been isolated from two patients with primary hypogammaglobulinemia [471]. When HIV was added to plasma and then processed into IVIG, $>10^4$ PFU/mL (PFU denotes plaque-forming units) of HIV were inactivated during alcohol fractionation and poly-(ethylene glycol) fractionation to IVIG [473]. Antibody to HIV was detected by an enzyme-linked im-

munoabsorbent assay (ELISA) in all 10 lots of a reduced, alkylated IVIG and in 4/8 lots of a pH 4/pepsin IVIG; 8/10 and 3/8 lots were also positive by Western blot analysis. The HIV antibody-positive lots, which were negative for HIV by culture and reverse transcriptase activity, were infused into patients. Comparison of pre- and post-infusion sera demonstrated that antibody to HIV was detectable for up to one month before converting to a negative antibody status [472].

One woman who had received hyperimmune Rho (D) immunoglobulin subsequently was found to be infected with HIV but belonged to a group with an increased risk of acquiring infection with HIV [474].

Anaphylaxis. Anaphylaxis has been documented following the administration of IVIG to some patients with hypogammaglobulinemia. Indeed, patients with antibody deficiency have an increased incidence of adverse reactions to IVIG [342]. Low levels of IgA are associated with an increased risk of such anaphylactic reactions [475]; IVIG products differ markedly in their IgA content [476]. In two patients with common variable immunodeficiency, IgE antibody to IgA in the infused IVIG was felt to account for these reactions [475]. In addition, IgE has been detected in commercial preparations of IVIG [477].

Other clinical problems have been described which involve the use of IVIG. IVIG prophylaxis precipitated cryoglobulinemic nephropathy in a patient with hypogammaglobulinemia secondary to B cell neoplasm [478].

The passive transfer of antibodies against specific blood types in IVIG may also be associated with hemolytic anemia or cause problems in the crossmatching of blood prior to surgery [479]. Immunologically normal patients may receive antibody to immunoglobulin that is of a different genotype.

5.8. Future Prospects

Large amounts of immunoglobulin can now be safely delivered directly into the bloodstream by using IVIG. This need was recognized by Cohn in the 1940s. Since immunoglobulin is now known to be a potent modulator of the immune system, IVIG has been used as a form of therapy for many noninfectious diseases. However, the impact of such treatment on the ability to respond to infectious diseases often complicates these conditions and still has to be assessed. Studies to date do not clearly show that the indications for the use of IVIG in infectious diseases should be expanded beyond those already shown for ISG. In some conditions involving functional hypogammaglobulinemia (e.g., HIV infection in young children, chronic lymphocytic leukemia), prophylaxis with standard IVIG may ultimately demonstrate efficacy. In certain subpopulations that are at risk of infection with specific organisms over a finite period of time, immunoglobulin preparations enriched in antibody to those organisms (i.e., hyperimmune preparations) may prove useful. The most promising use appears to be in the prophylaxis of specific infections; this usually involves hyperimmune serum and, as shown by studies in neonates, the substantial replacement of blood volume.

The use of IVIG to treat established infections with pharmacological (not replacement) doses of immunoglobulin, (with the possible exception of antitoxin treatment of septic shock), still has no proven role. This may be related to the need for prompt intervention and a large amount of a specific antibody, requiring rapid identification of the microorganisms involved. Preparations that are known to be immunotherapeutically effective are less efficacious when given 8 h or more after the onset of infection. Substantial evidence suggests that passive immunotherapy may provide an enhanced therapeutic benefit; however, this still remains to be proved in a prospective, controlled therapeutic trial.

Finally, more recent studies that show some benefit from passive immunotherapy have not been performed with IVIG preparations, e.g., [376], [463], [464], [459]. Thus despite the substantial progress made in the development of IVIG preparations for human use, their firm recommendation for widespread use in infectious disease requires further well-designed studies.

6. Immunotherapeutic Uses of Monoclonal Antibodies

When a vertebrate is vaccinated with a foreign antigen, it produces a mixture of antibodies that can bind to the antigen. Each antibody binds to a different epitope on the immunogen and is made by an individual clone of B lymphocytes. Industrial quantities of a single (i.e., monoclonal) antibody recognizing a specific epitope can be made by fusing a B lymphocyte producing the antibody of interest with an immortal tumor cell line to give hybridomas. For a detailed description of monoclonal antibodies, see → Monoclonal Antibodies.

Infectious diseases continue to cause much morbidity and mortality despite the discovery of a wide variety of novel antimicrobial agents and improved methods for immunizing at-risk populations. Therapy with human or animal serum is an effective adjunct to conventional drug therapy and, prior to the development of antimicrobials, was the only effective therapy for many infectious diseases (see Chap. 5). The efficacy of serum treatment for various infectious diseases suggests that this approach could be a useful adjunct to drug therapy if safety, efficacy, and production problems could be overcome. Human monoclonal antibodies (humabs) may solve many of the problems associated with serum therapy. Therapeutic humabs have been developed against many target organisms and are being tested in clinical trials.

6.1. Introduction

Despite the fact that the first reports describing the generation of antigen-specific murine monoclonal antibodies were published in the mid 1970s, routine production of humabs has been more difficult. Methods to generate and produce these molecules have improved and serious efforts are underway to develop them as therapeutic products for which passively administered human antisera against specific target antigens are in use in a number of clinical settings:

1) Red cell antigens: Rh (hemolytic disease of the newborn)
2) White cell antigens: antilymphocyte/thymocyte globulin
3) Viral antigens: hepatitis A and B, rabies, CMV, HSV, and varicella zoster
4) Bacterial antigens: tetanus, endotoxins, and pneumococcus
5) Elimination of circulating drugs (overdoses)
6) Antisnake venom
7) Fertility control (e.g., anti-β-human chorionic gonadotropin)

Humabs that recognize many of these antigens have been produced, and will probably augment or replace pooled antisera in the near future. Concerns, which include the possible contamination of pooled globulins by infectious agents (e.g., human immunodeficiency virus HIV, various hepatitis viruses, and cytomegalovirus) and the diminished availability of serum donors, as well as the relative ease of reproducible humab manufacture, will accelerate this trend.

Humabs will be preferred for the diagnosis and treatment of a variety of diseases because they minimize the problems encountered when a foreign animal monoclonal antibody is administered (e.g., anaphylaxis, clinical manifestations of immune complex formation, and reduction of efficacy by anti-antibodies). In well over half of the patients treated to date with murine monoclonal antibodies, the human antimouse immunoglobulin response has limited their usefulness [480]. Only a fraction of the antimouse immune response is directed to the variable region (idiotype) of the rodent immunoglobulins. This suggests that humabs will be more effective than murine monoclonal antibodies. Furthermore, humabs are more likely to have species-specific carbohydrates which may be important in a number of effector functions, such as Fc receptor-mediated, antibody-dependent, cellular cytotoxicity, complement activation, and phagocytosis [481]. Serum half-life and effector functions of immunoglobulin subclasses are very important for designing the optimal anti-infectious agent therapeutic. There are a number of reasons why passively administered humabs will be preferred for therapeutic use. A good summary and technical background are to be found in [482], [483].

6.2. Bacterial Targets

Various bacterial targets are recognized by humabs:

Tetanus toxoid [484–488]
Diphtheria toxoid [489]
Gram-negative endotoxins [491–494]
Pseudomonas aeruginosa lipopolysaccharide [496] and exotoxin A [497]
Hemophilus influenzae [498], [499]
Mycobacterium leprae [500]
Neisseria meningitides [501]
Pneumococcus [502], [503]

Tetanus-neutralizing humabs have been frequently generated because of the ease of obtaining immune B cells from vaccinated persons [484–488].

Because immunization against tetanus in the United States is universal and most individuals have relatively high titers, it is unlikely that any of these monoclonals will be scaled up and used in this country. The same is true of antidiphtheria monoclonal antibodies [489]. In other countries, vaccination against tetanus and diphtheria is not widespread and administration of immune human or animal sera is still practiced. Concern about HIV-contaminated serum has stimulated efforts to generate humabs against tetanus for use in India, South America, and southern Europe.

Gram-negative bacterial infections account for 1 – 2 % of hospital admissions and up to 100 000 deaths each year in the USA. Despite therapies including antibiotics and various support measures, mortality rates remain as high as 50 – 70 %. Much interest has been generated by the idea that lipopolysaccharide, the lethal component of gram-negative bacteria, might have antigenic determinants that are shared by many bacterial species. The outer carbohydrate domain of lipopolysaccharide varies highly from one species to the next whereas the inner carbohydrate core and lipid A regions appear to be much more conserved. In a clinical trial antisera were obtained from firemen vaccinated with a rough mutant *E. coli* strain called J 5 that lacks the outer carbohydrate. This antiserum gave significant protection to patients with gram-negative sepsis [490]. Several laboratories have reported the generation of humabs that are highly cross-reactive with bacterial lipopolysaccharide [491–494]. They are all of the IgM class and protect mice against infection when given at high doses. Two of these and a murine monoclonal of the same specificity are in clinical trials for the treatment of gram-negative sepsis. At present it is too early to determine whether they are successful, however several points should be made regarding this approach:

Firstly, it is unclear whether antibody was the active factor in the antisera in the original study [490]. Secondly, no documentation on the binding of the antisera to the bacteria causing a specific case of sepsis was made, i.e., no correlation could be made between response and nonresponse. Thirdly, immune serum is known to contain many other factors that might have benefited the treated patients. Fourthly, although cross-reactive antisera and monoclonal antibodies appear to bind to a wide number of bacteria and blotted lipopolysaccharide, questions still exist regarding the accessibility of the core and lipid A region for antibody binding. Fifthly, several trials with slightly different design have given inconsistent results [495]. Thus for a number of reasons the so-called core concept is questionable.

Type-specific horse antibodies recognizing typhoid and pneumococcus immunotypes were used from the turn of the century up to the beginning of the antibiotic era in the 1940s. The generation of humabs recognizing type-specific determinants on gram-negative bacteria is a reasonable alternative to the above-mentioned highly cross-reactive monoclonals. Although the diversity of possible immunotypes of gram-negative bacteria is large, a particular subset is probably responsible for most of the invasive bacteremias. Several groups are trying to generate a number of type-specific humabs and administer a cocktail of antibodies. This approach has already been tried for *Pseudomonas aeruginosa* [496]. In model systems, murine monoclonal antibodies recognizing a single immunotype were much more effective than those recognizing a core determinant that was shared between Pseudomonas species. If the cocktail approach is to prove effective, it may be necessary to add humabs that neutralize various virulence factors. In the case of *Pseudomonas* infections, humabs have also been generated to exotoxin A [497].

Humabs have been reported that recognize *Hemophilus influenza* type B capsular polysaccharide [498]. The successful introduction of an *H. influenza* type B capsular polysaccharide vaccine [499] (see Section 2.13) has made the clinical use of these humabs less attractive than originally anticipated. However, a safe humab for this life-threatening disease may still be developed for clinical use in conjunction with the vaccines.

Anti-mycobacterium leprae humabs have been reported [500]. In view of the central role of cell-mediated immunity in resistance and recovery from this disease, these reagents are unlikely to find clinical application. Nevertheless, these antibodies and those made by patients suffering from other infectious diseases can be used to probe the human humoral immune response, to clone antigens recognized by the humabs, and to investigate antibody-mediated autoimmunity initiated by microorganisms.

6.3. Viral and Chlamydial Targets

Antiviral humabs are especially attractive therapeutic targets. Immunoprophylaxis of cytomegalovirus (CMV) infections in immunosuppressed patients (e.g., transplantation) is of major clinical interest. Several groups have generated CMV-neutralizing humabs [504], [505] and clinical trials were scheduled to begin in 1989.

Humabs that neutralize hepatitis B virus [506–508] will replace the antisera currently used after acute exposure to this agent. These humabs may also be useful for "active – passive" immunization of at-risk infants. A very large population of infants, particularly in less developed countries, become chronically infected in the perinatal period. Evidence suggests [509] that administration of immune sera will prevent lifelong infection and could have an impact on the prevalence of the most common human cancer in the world, hepatitis-B-positive hepatoma [509]. The widespread use of anti-hepatitis A antisera suggests that the humabs generated to this target may also find clinical application.

A minor yet important target for humabs is the treatment of varicella zoster infections [510]. These humabs may have to be developed as an orphan drug given the small numbers of patients. (The development of drugs for rare diseases is supported by the US government; these pharmaceuticals are called orphan drugs.)

The acquired immune deficiency syndrome (AIDS) is caused by the human immunodeficiency viruses (HIV) and is currently the major infectious viral disease problem. Many groups are generating humabs that recognize one or more HIV strains. Many of these humabs recognize gene product gp 160 that is found in the viral envelope [511], [512]. At present, there is no evidence that a humoral immune response can prevent or change the pattern of HIV infection in people or animals. The viruses undergo rapid mutation (five times the rate of influenza). Furthermore, they can spread by cell to cell contact and reside permanently out of reach of humoral immunity inside the mononuclear phagocytes scattered throughout the body. Trials are underway to test the efficacy of a high-titer anti-HIV antiserum in the prevention of neonatal AIDS. The dismal failure of all attempted vaccines and passive serum to delay the spread of the virus in subhuman primate models is very discouraging. Antiviral humabs attached to toxins or engineered to bring cytotoxic cells into contact with infected cells might improve the potency of the antibody approach.

Several humabs have been generated against chlamydia [513] and other viruses:

Epstein Barr Virus [514]
Herpes simplex [515–517]
Human T cell leukemia virus I [518]
Measles (SSPE) virus [519], [520]
Rabies [521]
Rubella [522]
X 31 influenza virus [523]

However, clinical interest in the therapeutic purposes and use of these antibodies is limited.

6.4. Parasite Targets

The major therapeutic target of interest among parasitic diseases is malaria (see Section 4.2). Much less is known about the humoral immune response of humans suffering from other parasitic diseases (Section 4.1). Passively administered humabs may have therapeutic potential in acute *Falciparum malaria*. Immunoglobulin

M and IgG humabs have been generated [524], [525]. Monoclonal antibodies are also useful research tools in the development of malaria vaccines (Chap. 4).

7. References

1. I. Roitt: *Essential Immunology,* 5th ed., Blackwell Scientific Publications, Oxford, United Kingdom 1984.
2. B. D. Davis, R. Dubecco, H. N. Eisen, H. S. Ginsberg: *Microbiology: Including Immunology and Molecular Genetics,* 3rd ed., Harper and Row, Hagerstown, MD, 1980.
3. R. H. Schwartz, A. Yano, W. E. Paul: "Interaction between antigen-presenting cells and primed T-lymphocytes," *Immunol. Rev.* **40** (1978) 153.
4. W. Stroler, L. A. Hanson, K. W. Sell (eds.): *Recent Advances in Mucosal Immunity,* Raven Press, New York 1982.
5. T. Yamamura, T. Tada: *Progress in Immunology V,* Academic Press, Tokyo 1984.
6. J. B. Robbins, J. C. Hill, J. C. Sadoff (eds.): *Bacterial Vaccines, Seminars in Infectious Disease,* vol. IV, Thieme-Stratton, New York 1982.
7. U. E. Nydegger (ed.): *Immunotherapy, A Guide to Immunoglobulin Prophylaxis and Therapy,* Academic Press, London 1981.
8. R. E. Black, M. M. Levine, M. L. Clements, G. Losonsky et al.: "Prevention of shigellosis by a Salmonella typhi – Shigella sonnei bivalent vaccine," *J. Infect. Dis.* **155** (1987) 1260 – 1265.
9. M. M. Levine, J. B. Kaper, D. Herrington, J. Ketley et al.: "Safety, immunogenicity, and efficacy of recombinant live oral cholera vaccines, CVD 103 and CVD 103-HgR," *Lancet* **2** (1988) 467 – 470.
10. R. J. Collier: "Diphtheria toxin: mode of action and structure," *Bacteriol. Rev.* **39** (1975) 54 – 85.
11. A. M. Pappenheimer, Jr. in R. Germanier (ed.): "Diphtheria," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 1 – 36.
12. F. Audibert et al.: "Active antitoxic immunization by a diphtheria toxin synthetic oligopeptide," *Nature (London)* **289** (1981) 593 – 594.
13. T. Uchida, D. M. Gill, A. M. Pappenheimer, Jr.: "Mutation in the structural gene for diphtheria toxin carried by temperate phage β," *Nature (London)* **283** (1971) 8 – 11.
14. L. Greenfield, H. L. Dovey, F. C. Lawyer, D. H. Gelfand: "High level expression of diphtheria toxin peptides in Escherichia coli," *Bio/Technology* **4** (1986) 1006 – 1011.
15. B. Bizzinin in R. Germanier (ed.): "Tetanus," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 37 – 68.
16. M. C. Hardegree et al.: "Titration of tetanus toxoids in international units; relationship to antitoxin responses of rhesus monkeys," *6th Proc. Int. Conf. Tetanus 1981,* 1982, pp. 409 – 423.
17. M. Huet: "La standardisation des vaccins tetaniques," *6th Proc. Int. Conf. Tetanus 1981,* 1982 pp. 425 – 433.
18. K. W. Newell, A. Dueñas-Lehmann, D. R. Leblanc, N. Garces-Osorio: "The use of toxoid for the prevention of tetanus neonatorum; final report of a double-blind controlled field trial," *Bull. W.H.O.* **35** (1966) 863 – 871.
19. N. Guerin, C. Fillastre: "Vaccin D.T.C. Polan," *6th Proc. Int. Conf. Tetanus 1981,* 1982, pp. 477 – 479.
20. M. Pittman: "Pertussis toxin; the cause of the harmful effects and prolonged immunity of whooping cough," *Rev. Infect. Dis.* **1** (1979) 401 – 412.
21. C. R. Manclark, J. L. Cowell in R. Germanier (ed.): "Pertussis," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 69 – 106.
22. T. Madsen: "Vaccination against whooping cough," *JAMA J. Am. Med. Assoc.* **101** (1933) 187 – 188.
23. C. L. Cody et al.: "Nature and rates of adverse reactions associated with DTP and DT immunizations in infants and children," *Pediatrics* **68** (1981) 650 – 660.
24. R. Aldersdale et al.: *Whooping Cough. Reports from the Committee on Safety of Medicines and the Joint Committee on Vaccination and Immunization,* **vol. 4,** H. M. Stationary Office, London, England 1981, pp. 79 – 169.
25. M. Oda, J. L. Cowell, D. G. Burstyn, R. Manclark: "Protective activities of the filamentous hemagglutinin and the lymphocytosis-promoting factor of Bordetella pertussis in mice," *J. Infect. Dis.* **150** (1984) 823 – 833.
26. Ad Hoc Group for the study of Pertussis Vaccines: "Placebo controlled trial of two acellular pertussis vaccines in Sweden–protective efficacy and adresse events." *Lancet* **1** (1988) 955 – 960.

27. C. S. Chuttani et al.: "Controlled field trial of a high-dose oral killed typhoid vaccine in India," *Bull. W.H.O.* **55** (1977) 643 – 644.
28. M. Reitman: "Infectivity and antigenicity of streptomycin-dependent Salmonella typhosa," *J. Infect. Dis.* **117** (1967) 101 – 107.
29. R. Germanier, E. Fürer: "Isolation and characterization of gal ε mutant Ty 21 a of Salmonella typhi; a candidate strain for a live, oral typhoid vaccine," *J. Infect. Dis.* **131** (1975) 553 – 558.
30. R. H. Gilman, R. B. Hornick, W. E. Woodward, H. L. Dupont et al.: "Evaluation of a UDP-glucose-4-epimeraseless mutant of Salmonella typhi as a live oral vaccine," *J. Infect. Dis.* **136** (1977) 717 – 723.
31. M. H. Wahdan, C. Sérié, Y. Cerisier, S. Sallam et al.: "A controlled field trial of live Salmonella typhi Ty 21 a oral vaccine against typhoid; three year results," *J. Infect. Dis.* **145** (1982) 292 – 295.
32. World Health Organization Committee on Biological Standardization: Requirement for typhoid vaccine (live-attenuated, Ty21 a, oral). World Health Organization Technical Report, series 700; Geneva, Switzerland: 34th report, 1984, pp. 48 – 68.
33. R. Germanier in R. Germanier (ed.): "Typhoid fever," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 137 – 165.
34. M. M. Levine, C. Ferreccio, R. E. Black, R. Germanier, Chilean Typhoid Committee: "Large-scale field trial of Ty21 a live oral typhoid vaccine in enteric-coated capsule formulation," *Lancet* **ii** (1987) 1049 – 1052.
35. M. M. Levine, D. Herrington, J. R. Murphy, J. G. Morris et al.: Safety, infectivity, immunogenicity, and in vivo stability of two auxotrophic mutant strains of *Samonella typhi,* 541 Ty and 543 Ty, as live oral vaccine in humans, *J. Clin. Invest.* **79** (1987) 888 – 902.
36. I. L. Acharya, C. V. Lowe, R. Thapa, V. L. Gurubacharya et al.: "Prevention of typhoid fever in Nepal with the Vi capsular polysaccharide of *Salmonella typhi," N. Engl. J. Med.* **317** (1987) 1101 –1104.
37. R. Austrian: "Of gold and pneumococci; a history of pneumococcal vaccines in South Africa," *Trans. Am. Clin. Climatol. Assoc.* **89** (1977) 141 – 161.
38. C. M. MacLeod, M. R. Krauss: "Stepwise intratype transformation of pneumococcus from R to S by way of a variant intermediate in capsular polysaccharide production," *J. Exp. Med.* **86** (1947) 439 – 453.
39. J. B. Robbins, R. Austrian, C.-J. Lee, S. C. Rastogi et al.: "Considerations for formulating the second-generation pneumococcal capsular polysaccharide vaccine with emphasis on the cross-reactive types," *J. Infect. Dis.* **148** (1983) 1136 – 1159.
40. Health and Public Policy Committee, American College of Physicians: "Pneumococcal vaccine," *Ann. Intern. Med.* **104** (1986) 118 – 120.
41. P. Smit, D. Oberholzer, S. Hayden-Smith, H. J. Koornhof et al.: "Protective efficacy of pneumonococcal polysaccharide vaccines," *JAMA J. Am. Med. Assoc.* **238** (1977) 2613 – 2616.
42. G. Bolan, C. V. Broome, R. R. Facklam, B. D. Plikaytis et al.: "Pneumococcal vaccine efficacy in selected populations in the United States," *Ann. Intern. Med.* **104** (1986) 1 – 6.
43. M. S. Simberkoff, A. P. Cross, M. Al-Ibrahim, A. L. Baltch et al.: "Efficacy of pneumococcal vaccine in high-risk patients," *N. Engl. J. Med.* **315** (1986) 1318 – 1327.
44. J. A. Leech, A. Gervais, F. L. Ruben: "Efficacy of pneumococcal vaccine in severe chronic obstructive pulmonary disease," *Can. Med. Assoc. J.* **136** (1987) 361 – 365.
45. C. Chu, R. Schneerson, J. B. Robbins, S. C. Rastogi: "Further studies on the immunogenicity of Haemophilus influenzae type b and pneumococcal type 6 A polysaccharide-protein conjugates," *Infect. Immun.* **40** (1983) 245 – 256.
46. R. Schneerson, J. R. Robbins, C. Chu, A. Sutton et al.: "Serum antibody response of juvenile and infant rhesus monkeys injected with Haemophilus influenzae type b and pneumococcus type 6 A capsular polysaccharide-protein conjugates," *Infect. Immun.* **45** (1984) 582 – 591.
47. R. Schneerson, J. B. Robbins, J. C. Parke, Jr., C. Bell et al.: "Quantitative and qualitative analyses of serum antibodies elicited in adults by Haemophilus influenzae type b and pneumococcus type 6 A capsular polysaccharide-tetanus toxoid conjugates," *Infect. Immun.* **52** (1986) 519 – 528.
48. A. D. O'Brien, M. R. Thompson, P. Gemski, B. P. Doctor et al.: "Biological properties of Shigella flexneri 2 A toxin and its serological relationship to Shigella dysenteriae 1 toxin," *Infect. Immun.* **15** (1977) 796 – 798.
49. A. R. Higgins, T. M. Floyd, M. A. Kader: "Studies in shigellosis. III. A controlled

evaluation of a monovalent Shigella vaccine in a highly endemic environment," *Am. J. Trop. Med. Hyg.* **4** (1955) 281 – 288.
50. D. M. Mel, E. J. Gangarosa, M. L. Radovanović, B. L. Arsić et al.: "Studies on vaccination against bacillary dysentery. 6. Protection of children by oral immunization with streptomycin-dependent Shigella strains," *Bull. W.H.O.* **45** (1971) 457 – 464.
51. M. M. Levine, P. A. Rice, E. J. Gangarosa, G. K. Morris et al.: "An outbreak of Sonnei shigellosis in a population receiving oral attenuated shigella vaccines," *Am. J. Epidemiol.* **99** (1973) 30 – 36.
52. S. B. Formal, M. M. Levine in R. Germanier (ed.): "Shigellosis," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 167 – 186.
53. S. B. Formal, L. S. Baron, D. J. Kopecko, O. Washington et al.: "Construction of a potential bivalent vaccine strain: introduction of Shigella sonnei Form I antigen genes into the gal ε Salmonella typhi Ty 21 a typhoid vaccine strain," *Infect. Immun.* **34** (1981) 746 – 750.
54. R. E. Black, M. M. Levine, M. L. Clements, G. Losonsky et al.: "Prevention of shigellosis by a Salmonella typhi – Shigella sonnei bivalent vaccine," *J. Infect. Dis.* **155** (1987) 1260 – 1265.
55. L. S. Baron, D. J. Kopecko, S. B. Formal, R. Seid et al.: "Introduction of Shigella flexneri 2 a type and group antigen genes into oral typhoid vaccine strain Salmonella typhi Ty 21 a," *Infect. Immun.* **55** (1987) 2797 – 2801.
56. S. Sturm, K. Timmis: "Cloning of the rfb gene region of Shigella dysenteriae 1 and construction of an rfb-rfp gene cassette for the development of lipopolysaccharide-based live anti-dysentery vaccines," *Microb. Pathog.* **1** (1986) 289 – 297.
57. R. A. Finkelstein: "Immunology of cholera," *Immunol.* **69** (1975) 137 – 195.
58. R. A. Finkelstein: "Cholera," *CRC Crit. Rev. Microbiol.* **2** (1973) 553 – 623.
59. R. A. Cash, S. I. Music, J. P. Libonati, J. P. Craig et al.: "Response of man to infection with Vibrio cholerae. II. Protection from illness afforded by previous disease and vaccine," *J. Infect. Dis.* **130** (1974) 325 – 333.
60. R. A. Cash, S. I. Music, J. P. Libonati, A. R. Schwartz et al.: "Live oral cholera vaccine; evaluation of the clinical effectiveness of two strains in humans," *Infect. Immun.* **10** (1974) 762 – 764.
61. W. E. Woodward, R. H. Gilman, R. B. Hornick, J. P. Libonati et al.: "Efficacy of a live oral cholera vaccine in human volunteers," *Dev. Biol. Stand.* **33** (1976) 108 – 112.
62. M. M. Levine, R. E. Black, M. L. Clements, C. Lanata et al.: "Evaluation in humans of attenuated *Vibrio cholerae* El Tor Ogawa strain Texas Star-SR as a live oral vaccine, *Infect. Immun.* **43** (1984) 515 – 522.
63. J. B. Kaper, M. M. Levine: "Cloned cholera enterotoxin genes in study and prevention of cholera," *Lancet* **ii** (1981) 1162 – 1163.
64. J. J. Mekalanos, D. J. Suartz, G. D. N. Pearson, N. Harford et al.: "Cholera toxin genes: nucleotide sequence, deletion analysis, and vaccine development," *Nature* (London) 306 (1983) 551 – 557.
65. A. D. O'Brien, M. E. Chen, R. K. Holmes et al.: Environmental and human isolates of *Vibrio cholerae* and *Vibrio parahaemolyticus* produce a *Shigella dysenteriae* 1 (Shiga)-like cytotoxin," *Lancet* **i** (1984) 77 – 78.
66. J. D. Clements, J. R. Harris, M. R. Khan, B. A. Kay et al.: "Field trial of oral cholera vaccine in Bangladesh," *Lancet* **ii** (1986) 124 – 127.
67. J. E. McGowan, M. W. Barnes, M. Finland: "Bacteremia at Boston City Hospital; occurrence and mortality during 12 selected years (1935 – 1972), with special reference to hospital-acquired cases," *J. Infect. Dis.* **132** (1975) 316 – 335.
68. C. S. Bryan, K. L. Reynolds, E. R. Brenner: "Analysis of 1186 episodes of gram-negative bacteremia in non-university hospitals; the effects of antimicrobial therapy," *Rev. Infect. Dis.* **5** (1983) 629 – 638.
69. R. W. Haley: *Preliminary cost-benefit analysis of hospital infection control programs (The SENIC Project),* Proc. Internat. Workshop, Baiersbromm, Germany (1977) pp. 93 – 95.
70. S. J. Cryz, Jr.: "Prospects for the prevention and control of gram-negative nosocomial infections," *Vaccine* **5** (1987) 261 – 265.
71. W. R. Jarvis, J. W. White, V. P. Munn, J. L. Mosser et al.: "Nosocomial infection surveillance," 1983. Centers for Disease Control Surveillance Summaries **33**, no. 299, pp. 955 – 2155 (1985).
72. M. S. Pollack, L. S. Young: "Protective activity of antibodies to exotoxin A and lipopolysaccharide at the onset of Pseudomonas aeruginosa septicemia in man," *J. Clin. Invest.* **63** (1979) 276 – 286.

73. S. J. Cryz, Jr. in R. Germanier (ed.): "Pseudomonas aeruginosa infections," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 317 – 351.
74. G. B. Pier, S. E. Bennett: "Structural analysis and immunogenicity of Pseudomonas aeruginosa immunotype 2 high molecular weight polysaccharide," *J. Clin. Invest.* **77** (1986) 491 – 495.
75. S. J. Cryz, Jr., E. Fürer, A. S. Cross, A. Wegmann et al.: "Safety and immunogenicity of a Pseudomonas aeruginosa O-polysaccharide-toxin A conjugate vaccine in humans," *J. Clin. Invest.* **80** (1987) 51 – 56.
76. S. J. Cryz, Jr., A. S. Cross, E. Fürer, N. Chariatte, et al.: "Activity of intravenous immune globulin against *Klebsiella," J. Lab. Clin. Med.* **108** (1986) 182 – 189.
77. W. D. Welch, W. J. Martin, P. Stevens, L. S. Young: "Relative opsonic and protective activities of antibodies against Kl, O and lipid A antigens of *Escherichia coli," Scand. J. Infect. Dis.* **11** (1979) 291 – 298.
78. B. Kaijser, S. Ahlstedt: "Protective capacity of antibodies to Escherichia coli O and K antigens," *Infect. Immun.* **17** (1977) 286 – 289.
79. A. S. Cross, P. Gemski, J. C. Sadoff, F. Ørskov et al.: "The importance of the Kl capsule in invasive infections caused by *Escherichia coli," J. Infect. Dis.* **149** (1984) 184 – 193.
80. M. Granström, I. Julander, R. Möllby: "Serological diagnosis of deep Staphylococcus aureus infections by enzyme-linked immunosorbent assay (ELISA) for staphylococcal hemolysin and teichoic acid," *Scand. J. Infect. Dis. Suppl.* **41** (1983) 132 – 139.
81. W. W. Karakawa, J. M. Fournier, W. F. Vann, R. Arbeit et al.: "Method for the serological typing of the capsular polysaccharides of *Staphylococcus aureus," J. Clin. Microbiol.* **22** (1985) 445 – 447.
82. H. K. Hochkeppel, D. G. Braun, W. Visher, A. Imm et al.: "Serotyping and electron microscopy studies of Staphylococcus aureus clinical isolates with monoclonal antibodies to capsular polysaccharide types 5 and 8," *J. Clin. Microbiol.* **25** (1987) 526 – 530.
83. R. D. Arbeit, J. M. Nelles: "Capsular polysaccharide antigenemia in rats with experimental endocarditis due to *Staphylococcus aureus," J. Infect. Dis.* **155** (1987) 242 – 246.
84. L. M. Mutharia, G. Crockford, W. C. Bogard, Jr., R. E. W. Hancock: "Monoclonal antibodies specific for Escherichia coli J 5 lipolysaccharide: cross-reaction with other gram-negative bacterial species," *Infect. Immun.* **45** (1984) 631 – 636.
85. E. J. Ziegler, J. A. McCutchan, H. Douglas, A. I. Braude: "Treatment of E. coli and Klebsiella bacteremia in agranulocytopenic animals with antiserum to a UDP-GAL epimerase-deficient mutant," *J. Immunol.* **111** (1973) 433 – 438.
86. E. J. Ziegler, J. A. McCutchan, J. Fierer, M. P. Glauser et al.: "Treatment of gramnegative bacteremia and shock with human antiserum to a mutant of *Escherichia coli," N. Engl. J. Med.* **307** (1982) 1225 – 1230.
87. J. D. Baumgartner, M. P. Glauser, J. A. McCutchan, E. J. Ziegler et al.: "Prevention of gramnegative shock and death in surgical patients by antibody to endotoxin core glycolipid," *Lancet* **ii** (1985) 59 – 63.
88. J. D. Band, M. E. Chamberland, T. Platt, R. E. Weaver et al.: "Trends in meningococcal disease in the United States, 1975 – 1980," *J. Infect. Dis.* **148** (1983) 754 – 758.
89. E. C. Gotschlich in R. Germanier (ed.): "Meningococcal meningitis," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 237 – 255.
90. L. Lapeyssonnie: "La meningite cerebro-spinale en Afrique," *Bull. W.H.O.* **28** (1963) 1 – 114.
91. E. C. Gotschlich, I. Goldschneider, M. S. Artenstein: "Human immunity to the meningococcus. IV. Immunogenicity of group A and group C meningococcal polysaccharides in human volunteers; *J. Exp. Med.* **129** (1969) 1367 – 1384.
92. M. S. Artenstein, R. Gold, J. G. Zimmerly, F. A. Wyle et al.: "Prevention of meningococcal disease by group C polysaccharide vaccine," *N. Engl. J. Med.* **282** (1970) 417 – 420.
93. P. H. Mäkelä et al.: "Effect of group-A meningococcal vaccine in army recruits in Finland," Lancet **ii** (1975) 883 – 886.
94. M. H. Wahdan, S. A. Sallam, M. N. Hassan, A. A. Gawad et al.: "A second controlled field trial of a serogroup A meningococcal vaccine in Alexandria," *Bull. W.H.O.* **55** (1977) 645 – 651.
95. H. Peltola, P. H. Mäkelä, H. M. Käyhty, H. Jousimies et al.: "Clinical efficacy of meningococcus group A capsular polysaccharide vaccine in children three

months to five years of age," *N. Engl. J. Med.* **297** (1977) 686 – 691.
96. H. J. Jennings, C. Lugowski: "Immunochemistry of groups A, B and C meningococcal polysaccharide-tetanus toxoid conjugates," *J. Immunol.* **127** (1981) 1011 – 1018.
97. C. E. Frasch: "Prospects for the prevention of meningococcal disease: special reference to group B," *Vaccine* **5** (1987) 3 – 4.
98. W. D. Zollinger, R. E. Mandrell, J. M. Griffis in J. B. Robbins, J. C. Hill, J. C. Sadoff (eds.): "Enhancement of immunological activity by monovalent complexing of meningococcal group B polysaccharide and outer membrane protein," *Seminars in Infectious Disease,* vol. 4, Thieme-Stratton, New York 1982, pp. 254 – 262.
99. C. E. Frasch, G. Coetzee, D. M. Zahradrik, L. Y. Wang in G. K. Schoolnik (ed.): "New developments in meningococcal vaccines," *Pathogenic Neisseria,* American Society for Microbiology, Washington, D.C. 1985, pp. 633 – 640.
100. F. M. Collins in R. Germanier (ed.): "Tuberculosis," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 373 – 418.
101. R. G. Ferguson, A. B. Simes: "BCG vaccination of Indian infants in Saskatchewan," *Tubercle* **30** (1949) 5 – 11.
102. S. V. Boyden: "The effect of previous injections of tuberculoprotein on the development of tuberculin sensitivity following B.C.G. vaccination in guinea pigs," *Br. J. Exp. Pathol.* **38** (1957) 611 – 617.
103. A. Lotte, O. Wasz-Höckert, F. Lert, N. Dumitrescu et al.: "Balance of risks and balance of cost regarding tuberculosis and antituberculosis vaccination," *Dev. Biol. Stand.* **43** (1979) 111 – 119.
104. R. Torriani, A. Zimmermann, A. Morell: "Die BCG-Sepsis als letale Komplikation der BCG-Impfung," *Schweiz. Med. Wochenschr.* **109** (1979) 708 –713.
105. W. Hennessen, H. Freudenstein, H. Engelhardt: "Observations on a BCG vaccine causing adverse reactions in newborns," *J. Biol. Stand.* **5** (1977) 139 – 146.
106. D. B. Travers: "BCG vaccination," *Lancet* **i** (1981) 1001 – 1002.
107. M. C. Lu, M. H. Lien, R. E. Becker, H. C. Heine et al.: "Genes for immunodominant protein antigens are highly homologous in Mycobacterium tuberculosis, Mycobacterium africanum, and the vaccine strain Mycobacterium bovis BCG," *Infect. Immun.* **55** (1987) 2378 – 2382.
108. B. Moss, C. Flexner: "Vaccinia virus expression vectors," *Ann. Rev. Immunol.* **5** (1987) 305 – 324.
109. H. Leffler, C. Svanborg-Eden: "Glycolipid receptors for uropathogenic E. coli attaching to human urinary tract epithelial cells and agglutinating human erythrocytes," *FEMS Microbiol. Lett.* **8** (1980) 127 – 134.
110. K. Rauss, I. Kétyi, E. Matusovits, L. Szendrei et al.: "Specific oral prevention of infantile enteritis. III. Experiments with corpuscular vaccine," *Acta. Microbiol. Acad. Sci. Hung.* **19** (1972) 19 – 28.
111. E. Fürer et al.: "Protection against colibacillosis in neonatal piglets by immunization of dams with procholeragenoid," *Infect. Immun.* **35** (1982) 887 – 894.
112. L. Doberescu, C. Huygelen: "Protection of piglets against neonatal *E. coli* enteritis by immunization of the sow with a vaccine containing heat-labile enterotoxin (LT), *Zentralbl. Veterinärmed. Reihe A* **23** (1976) 79 – 88.
113. R. L. Morgan, R. E. Isaacson, H. W. Micon, C. C. Brinton et al.: "Immunization of suckling pigs against enterotoxigenic Escherichia coli-induced diarrheal disease by vaccinating dams with purified 987 or K 99 pili; protection correlates with pilus homology of vaccine and challenge," *Infect. Immun.* **22** (1978) 771 – 777.
114. M. M. Levine, R. E. Black, C. C. Brinton, Jr., M. L. Clements et al.: "Reactogenicity, immunogenicity and efficacy studies of *Escherichia coli* type 1 somatic pili parenteral vaccine in man, *Scand. J. Infect. Dis. Suppl.* **33** (1982) 83 – 95.
115. D. G. Evans, D. J. Evans, Jr., A. R. Opekun, D. Y. Graham in A. Tagliabue, R. Rappuoli, S. E. Piazzi (eds.): "Oral whole cell vaccine protective against enterotoxigenic Escherichia coli diarrhea," *Bacterial Vaccines and Local Immunity,* Edita Da Sclavo Sp.A., Siena, Italy 1986, pp. 155 – 156.
116. M. M. Levine in R. Germanier (ed.): *Escherichia coli* infections," Bacterial Vaccines, Academic Press, Orlando, FL 1984 pp. 187 – 235.
117. F. A. Klipstein, R. F. Engert, R. A. Houghton: "Protection in rabbits immunized with a vaccine of Escherichia coli heat-stable toxin cross-linked to the heat-labile toxin B subunit," *Infect. Immun.* **40** (1983) 888 – 893.

118. J. Holmgren: "Actions of cholera toxin and the prevention and treatment of cholera," *Nature (London)* **292** (1981) 413 – 417.
119. R. A. Finkelstein, C. V. Sciortino, L. C. Rieke, M. F. Burks et al.: "Preparation of procholigenoids from Escherichia coli heat-labile enterotoxins," *Infect. Immun.* **45** (1984) 518 – 521.
120. M. A. Schmidt, P. O'Hanley, G. K. Schoolinik in A. Tagliabue, R. Rappuoli, S. E. Piazzi (eds.): Synthetic peptides protect against infection in a model of experimental pyelonephritis in mice," *Bacterial Vaccines and Local Immunity,* Edita Da Sclavo Sp.A., Siena, Italy 1986, pp. 389 – 398.
121. E. C. Gotschlich in R. Germanier (ed.): ''Gonorrhea," *Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 353 – 371.
122. D. S. Kellogg, Jr., I. R. Cohen, L. C. Norins, A. L. Schroeter et al.: "Neisseria gonorrhoeae. II. Colonial variation and pathogenicity during 35 months in vitro," *J. Bacteriol.* **96** (1968) 596 – 605.
123. J. Swanson, S. J. Kraus, E. C. Gotschlich: "Studies on gonococcus infection. I. Pili and zones of adhesion; their relation to gonococcal growth patterns," *J. Exp. Med.* **134** (1971) 886 – 906.
124. L. Greenberg, B. B. Diena, F. A. Ashton, R. Wallace et al.: "Gonococcal vaccine studies in Inuvik," *Can. J. Public Health* **65** (1974) 29 – 33.
125. E. C. Tramont, J. W. Boslego: "Pilus vaccines," *Vaccine* **3** (1985) 3 – 10.
126. C. C. Brinton, S. W. Wood, A. Brown, A. M. Labik et al. in : J. B. Robbins, J. Hill, J. C. Sadoff (eds.): "The development of a Neisseria pilus vaccine for gonorrhea and meningococcal meningitis," *Seminars in Infectious Disease,* vol. 4, Bacterial Vaccines, Thieme-Stratton, Inc., New York 1982, pp. 140 – 159.
127. G. K. Schoolinik, J. Y. Tai, E. C. Gotschlich: "A pilus peptide vaccine for prevention of gonorrhea," *Prog. Allergy* **33** (1983) 314 – 331.
128. J. B. Rothbard, R. Fernandez, L. Wang, N. N. H. Teng et al.: "Antibodies to peptides corresponding to conserved sequence of gonococcal pilins block bacterial adhesion," *Proc. Nat. Acad. Sci. U.S.A* (2) (1985) 915 – 919.
129. M. S. Blake in G. G. Jackson, H. Thomas (ed.): "Functions of the outer membrane proteins of *Neisseria gonorrhoeae," The Pathogenesis of Bacterial Infections*, Springer-Verlag, Berlin 1985, pp. 51 – 66.
130. T. M. Buchanan, M. S. Siegel, K. C. S. Chen, W. A. Pearce in J. B. Robbins, J. Hill, J. C. Sadoff (ed.): "Development of a vaccine to prevent gonorrhea," *Seminars in Infectious Disease,* vol. 4, Bacterial Vaccines, Thieme-Stratton, Inc., New York 1982, pp. 160 – 164.
131. J. B. Robbins, R. Schneerson, M. Pittman in R. Germanier (ed.): *Haemophilus influenzae type b infections," Bacterial Vaccines,* Academic Press, Orlando, FL 1984, pp. 289 – 316.
132. A. Sutton, R. Schneerson, S. Kendall-Morris, J. B. Robbins: "Differential complement resistance mediates virulence of Haemophilus influenzae type b," *Infect. Immun.* **35** (1982) 95 – 104.
133. H. Peltola, H. Käyhty, A. Sivonen, P. H. Mäkelä: "Haemophilus influenzae type b capsular polysaccharide vaccine in children; a double-blind field study of 100 000 vaccines 3 months to 5 years of age in Finland," *Pediatrics* **60** (1977) 730 – 737.
134. E. A. Mortimer, "Efficacy of Haemophilus b Polysaccharide Vaccine: an Enigma," *J. Am. Med. Assoc.* **260** (1988) 1454 – 1455.
135. H. Käyhty, J. Eskola, H. Peltola, M. G. Stout et al.: "Immunogenicity in infants of a vaccine composed of Haemophilus influenzae type b capsular polysaccharide mixed with DPT or conjugated to diphtheria toxoid," *J. Infect. Dis.* **155** (1987) 100 – 106.
136. A. A. Lenoir, P. D. Granoff, D. M. Granoff: "Immunogenicity of Haemophilus influenzae type b polysaccharide-Neisseria meningitidis outer membrane protein conjugate vaccine in 2- to 6-month-old infants," *Pediatrics* **80** (1987) 283 – 287.
137. J. Eskola, H. Peltola, A. K. Takala, H. Käyhty et al.: "Efficacy of Haemophilus influenzae type b polysaccharide-diphtheria toxoid conjugate vaccine in infancy," *N. Engl. J. Med.* **317** (1987) 717 –722.
138. C. D. Mitchell, H. H. Balfour, Jr.: "Measles control: So near and yet so far," *Prog. Med. Virol.* **31** (1985) 1 – 42.
139. L. E. Markowitz, R. H. Bernier: "Immunization of young infants with Edmonston-Zagreb measles vaccine," *Pediatr. Infect. Dis. J.* **6** (1987) 809 – 812.
140. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **329** (1966) 52 – 73.

141. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **673** (1982).
142. Centers for Disease Control: "Measles – United States, 1986," *Morb. Mort. Wkly. Rep.* **36** (1987) 301 – 305.
143. K. J. Bart, W. A. Orenstein, A. R. Hinman: "The virtual elimination of rubella and mumps from the United States and the use of combined measles, mumps and rubella vaccines (MMR) to eliminate measles," *Dev. Biol. Stand.* **65** (1986) 45 – 52.
144. E. Norrby, G. Enders-Ruckle, V. ter Meulen: "Differences in the appearance of antibodies to structural components of measles virus after immunization with inactivated and live virus," *J. Infect. Dis.* **132** (1975) 262 – 269.
145. T. M. Varsanyi, B. Morein, A. Löve, E. Norrby: "Protection against lethal measles virus infection in mice by immune-stimulating complexes containing the hemagglutinin or fusion protein," *J. Virol.* **61**(1987) 3896 – 3901.
146. M. R. Hilleman: "Mumps vaccination," in: *Modern Trends in Medical Virology,* vol. 2, Butterworths, London (1970) pp. 241 – 261.
147. K. Penttinen, K. Cantell, P. Somer, A. Poikolainen: "Mumps vaccination in the Finnish defense forces," *Am. J. Epidemiol.* **88** (1968) 234 – 244.
148. M. R. Hilleman, R. E. Weibel, E. B. Buynak, J. Stokes, Jr. et al.: "Live, attenuated mumps-virus vaccine; 4. Protective efficacy as measured in a field evaluation," *N. Engl. J. Med.* **276** (1967) 252 – 258.
149. Centers for Disease Control: "Efficacy of mumps vaccine – Ohio," *Morb. Mort. Wkly. Rep.* **32** (1983) 391–398.
150. F. E. Andre, J. Peetermans: "Effect of simultaneous administration of live measles vaccine on the "take rate" of live mumps vaccine," *Dev. Biol. Stand.* **65** (1986) 101 – 107.
151. R. Glück, J. M. Hoskins, A. Wegmann, M. Just et al.: "Rubini, a new live attenuated mumps vaccine virus strain for human diploid cells," *Dev. Biol. Stand.* **65** (1986) 29 – 35.
152. WHO Expert Committee on biological standardization: *WHO Tech. Rep. Ser.* **760** (1987).
153. Recommendation of the Immunization Practices Advisory Committee (ACIP): "Mumps vaccine," *Morb. Mort. Wkly. Rep.* **31** (1982) 617 – 625.
154. K. Penttinen, E.-P. Helle, E. Norrby: "Differences in antibody response induced by formaldehyde inactivated and live mumps vaccines," *Dev. Biol. Stand.* **43** (1979) 265 – 268.
155. A. Löve, R. Rydbeck, G. Utter, C. örvell et al.: ''Monoclonal antibodies against fusion protein are protective in necrotizing mumps meningoencephalitis," *J. Virol.* **58** (1986) 220 – 222.
156. F. T. Perkins: "Licensed vaccines," *Rev. Infect. Dis.* **7** (1985) Suppl. 1, 73 – 76.
157. S. A. Plotkin, F. Buser: "History of RA 27/3 rubella vaccine," *Rev. Infect. Dis.* **7** (1985) Suppl. 1, 77 – 78.
158. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **610** (1977) 54 – 84.
159. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **658** (1981).
160. J. A. Dudgeon: "Selective immunization; Protection of the individual," *Rev. Infect. Dis.* **7** (1985) Suppl. 1, 185 – 190.
161. Recommendation of the Immunization Practices Advisory Committee (ACIP): "Rubella prevention," *Morb. Mort. Wkly. Rep.* **33** (1984) 301 – 318.
162. B. Christenson, M. Böttiger, L. Heller: "Mass vaccination program aimed at eradicating measles, mumps, and rubella in Sweden; first experience," *Br. Med. J.* **287** (1983) 389 – 391.
163. Centers for Disease Control: "Rubella vaccination during pregnancy – United States, 1971 – 1986," *Morb. Mort. Wkly. Rep.* **36** (1987) 457 – 461.
164. Recommendation of the Immunization Practices Advisory Committee (ACIP): "New recommended schedule for active immunization of normal infants and children," *Morb. Mort. Wkly. Rep* **35** (1986) 577 – 579.
165. A. Deforest, S. S. Long, H. W. Lischner, J. A. C. Girone et al.: "Simultaneous administration of measles-mumps-rubella vaccine with booster doses of diphtheria-tetanus-pertussis and poliovirus vaccines," *Pediatrics* **81** (1988) 237 – 246.
166. H. Peltola, O. P. Heinonen: "Frequency of true adverse reactions to measles-mumps-rubella vaccine," *Lancet* **i** (1986) 939 – 942.
167. A. M. Arbeter, L. Baker, S. E. Starr, S. A. Plotkin: "The combination measles, mumps, rubella and varicella vaccine in healthy children," *Dev. Biol. Stand.* **65** (1986) 89 – 93.

168. L. B. Schonberger, J. Kaplan, R. Kim-Farley, M. Moore et al.: "Control of paralytic poliomyelitis in the United States," *Rev. Infect. Dis.* **6** (1984) Suppl. 2, 424 – 426.
169. J. Salk: "Commentary: Poliomyelitis vaccination – choosing a wise policy," *Pediatr. Infect. Dis. J.* **6** (1987) 889 – 893.
170. G. S. Wilson: *The hazards of immunization,* University of London, The Athlone Press 1967.
171. A. B. Sabin: "Oral poliovirus vaccine: History of its development and use and current challenge to eliminate poliomyelitis from the world," *J. Infect. Dis.* **151** (1985) 420 – 436.
172. A. B. Sabin: "Commentary: Is there a need for a change in poliomyelitis immunization policy?" *Pediatr. Infect. Dis. J.* **6** (1987) 887 – 889.
173. M. Böttiger: "Experiences of vaccination with inactivated poliovirus vaccine in Sweden," *Dev. Biol. Stand.* **47** (1981) 227 – 232.
174. K. Lapinleimu, M. Stenvik: "Experiences with polio vaccination and herd immunity in Finland," *Dev. Biol. Stand.* **47** (1981) 241 – 246.
175. H. Bijkerk: "Poliomyelitis in the Netherlands," *Dev. Biol. Stand.* **47** (1981) 233 – 240.
176. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **687** (1983) 107 – 174.
177. A. M. Mcbean, J. F. Modlin: "Rationale for the sequential use of inactivated poliovirus vaccine and live attenuated poliovirus vaccine for routine poliomyelitis immunization in the United States," *Pediatr. Infect. Dis. J.* **6** (1987) 881 – 887.
178. K. Lapinleimu: "Sero-immune pattern of poliomyelitis type I and III in a population vaccinated with inactivated polio vaccine," in: *Proceedings of the 11th Symposium of the European Association Against Poliomyelitis and Allied Diseases,* Rome 1966, Brussels; European Association Against Poliomyelitis, 1967, pp. 119 – 25.
179. A. L. van Wezel, G. van Steenis, Ch. A. Hannik, H. Cohen: "New approach to the production of concentrated and purified inactivated polio and rabies tissue culture vaccines," *Dev. Biol. Stand.* **41** (1978) 159 – 168.
180. S. E. Robertson, H. P. Traverso, J. A. Drucker, E. Z. Rovira et al.: "Clinical efficacy of a new, enhanced-potency, inactivated poliovirus vaccine," *Lancet* **i** (1988) 897 – 899.
181. B. J. Montagnon, B. Fanget, A. J. Nicolas: "The large-scale cultivation of vero cells in micro-carrier culture for virus vaccine production. Preliminary results for killed poliovirus vaccine," *Dev. Biol. Stand.* **47** (1981) 55 – 64.
182. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **745** (1987).
183. Prevention of liver cancer: *WHO Tech. Rep. Ser.* **691** (1983).
184. A. A. McLean: "Development of vaccines against Hepatitis A and Hepatitis B," *Rev. Infect. Dis.* **8** (1986) 591 – 598.
185. W. Szmuness, C. E. Stevens, E. J. Harley, E. A. Zang et al.: "Hepatitis B vaccine; Demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States," *N. Engl. J. Med.* **303** (1980) 833 – 841.
186. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **725** (1985) 102 – 133.
187. Recommendation of the Immunization Practices Advisory Committee (ACIP): "Recommendations for protection against viral hepatitis," *Morb. Mort. Wkly. Rep.* **34** (1985) 313 – 335.
188. Recommendation of the Immunization Practices Advisory Committee (ACIP): "Update on Hepatitis B prevention," *Morb. Mort. Wkly. Rep.* **36** (1987) 353 – 366.
189. G. Eder, J. L. McDonel, F. Dorner: "Hepatitis B vaccine," in: *Progress in Liver Diseases,* vol. 8, Grune and Stratton, New York 1986, pp. 367 – 394.
190. A. Zuckerman: "Tomorrow's hepatitis B vaccines," *Vaccine* **5** (1987) 165 – 167.
191. D. N. Standring, W. J. Rutter: "The molecular analysis of hepatitis B virus," in: *Progress in Liver Diseases,* vol. 8, Grune and Stratton, New York 1986, pp. 311 – 333.
192. WHO Expert Committee on Rabies: *WHO Tech. Rep. Ser.* **709** (1984).
193. Centers for Disease Control: "Bat-rabies – Europe," *Morb. Mort. Wkly. Rep.* **35** (1986) 430 – 432.
194. P. Sureau: "Rabies vaccine production in animal cell cultures," *Adv. Biochem. Eng. Biotechnol.* **34** (1987) 111 – 128.
195. B. D. Perry: "Rabies," *Vet. Clin. North. Am. Small Anim. Pract.* **17** (1987) 73 – 89.
196. H. Wilde, P. Chomchey, S. Prakongsri, P. Punyaratabandhu: "Safety of equine rabies immune globulin," *Lancet* **ii** (1987) 1275.

197. Centers for Disease Control: "Human rabies despite treatment with rabies immune globulin and human diploid cell rabies vaccine – Thailand," *Morb. Mort. Wkly. Rep.* **36** (1987) 759 – 765.
198. Centers for Disease Control: "Systemic allergic reactions following immunization with human diploid cell rabies vaccine," *Morb. Mort. Wkly. Rep.* **33** (1984) 185 – 187.
199. M. C. Swanson, E. Rosanoff, M. Gurwith, M. Deitch et al.: "IgE and IgG antibodies to β-propiolactone and human serum albumin associated with urticarial reactions to rabies vaccine," *J. Infect. Dis.* **155** (1987) 909 – 913.
200. M. Granström, M. Eriksson, G. Edevåg: "A sandwich ELISA for bovine serum in viral vaccines," *J. Biol. Stand.* **15** (1987) 193 – 197.
201. B. Morein, K. Simons: "Subunit vaccines against enveloped viruses; virosomes, micelles and other protein complexes," *Vaccine* **3** (1985) 83 – 93.
202. C. H. Stuart-Harris, G. C. Schild: "Influenza," *The Viruses and Disease,* E. Arnoldt Ltd, London 1976.
203. P. F. Wright, D. T. Karzon: "Live Attenuated Influenza Vaccines," *Prog. Med. Virol.* **34** (1987) 70 – 88.
204. WHO Expert Committee on Biological Standardization: *WHO Techn. Rep. Ser.* **638** (1979).
205. Recommendations of the Immunization Practices Advisory Committee (ACIP): "Prevention and Control of Influenza," *Morb. Mort. Wkly. Rep.* **36** (1987) 373 – 387.
206. P. A. Gross, A. L. Gould, A. E. Brown: "Effect of cancer chemotherapy on the immune response to influenza virus vaccine: Review of published studies," *Rev. Infect. Dis.* **7** (1985) 613 – 618.
207. M. A. Strassburg, S. Greenland, F. J. Sorvillo, L. E. Lieb et al.: "Influenza in the elderly: report of an outbreak and a review of vaccine effectiveness reports," *Vaccine* **4** (1986) 38 – 44.
208. J. M. Ostrove, G. Inchauspe: "The biology of varicella-zoster virus" in: S. E. Straus "Varicella-zoster virus infections: biology, natural history, treatment, and prevention," *Ann. Intern. Med.* **108** (1988) 221 –237.
209. M. Takahashi: "Clinical overview of varicella vaccine; Development and early studies," *Pediatrics* **78** (1986) Suppl., 736 – 741.
210. R. E. Weibel, B. J. Neff, B. J. Kuter, H. A. Guess et al.: "Live attenuated varicella virus vaccine. Efficacy trial in healthy children," *N. Engl. J. Med.* **310** (1984) 1409 – 1415.
211. A. A. Gershon, S. P. Steinberg, L. Gelb, and The National Institute of Allergy and Infectious Diseases Varicella Vaccine Collaborative Study Group: "Live attenuated varicella vaccine use in immunocompromised children and adults," *Pediatrics* **78** (1986) Suppl., 757 – 762.
212. Arthropod-borne and rodent-borne viral diseases: *WHO Tech. Rep. Ser.* **719** (1985).
213. Viral haemorrhagic fevers: *WHO Tech. Rep. Ser.* **721** (1985).
214. K. C. Smithburn, C. Durieux, R. Koerber, H. A. Penna et al.: *Yellow Fever Vaccination,* World Health Organization, Geneva 1956.
215. J. C. Roche, A. Jouan, B. Brisou, R. Rodhain et al.: "Comparative clinical study of a new 17 D thermostable yellow fever vaccine," *Vaccine* **4** (1986) Suppl., 163 – 165.
216. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **594** (1976) 23 – 49.
217. A. A. Smorodintsev, A. V. Dubov, V. I. Ilyenko, V. G. Platonov: "A new approach to development of live vaccine against tick-borne encephalitis," *J. Hyg.* **67** (1969) 13 – 20.
218. C. Kuntz, F. X. Heinz, H. Hofmann: "Immunogenicity and reactogenicity of a highly purified vaccine against tick-borne encephalitis," *J. Med. Virol.* **6** (1980) 103 – 109.
219. A. Oya in H. Fukumi (ed.): "Japanese encephalitis vaccine," *The vaccination. Theory and practice,* International Medical Foundation of Japan (1975) pp. 69 – 82.
220. WHO Expert Committee on Biological Standardization: *WHO Tech. Rep. Ser.* **323** (1966) 56 – 71.
221. Recommendation of the Immunization Practices Advisory Committee (ACIP): *Smallpox Vaccine Morb. Mort. Wkly. Rep.* **34** (1985) 341 – 342.
222. R. Randall, L. N. Binn, V. R. Harrison: "Immunization against Rift Valley fever virus," *J. Immunol.* **93** (1964) 293 – 299.
223. B. Niklasson: "Rift Valley fever virus vaccine trial; Study of side effects in humans," *Scand. J. Infect. Dis.* **14** (1982) 105 – 109.
224. P. M. Kaye, *Parasitol. Today* **3** (1987) 293 – 299.
225. M. D. Rickard, M. J. Howell in E. E. R. Taylor, J. R. Baker (ed.): *In Vitro Methods for Parasite Cultivation,* Academic Press, London 1987, pp. 407 – 451.

226. G. Piekarski: *Medizinische Parasitologie,* Springer Verlag, Berlin 1987.
227. V. Houba in V. Houba (ed.): *Immunological Investigation of Tropical Parasitic Diseases,* Churchill Livingstone, Edinburgh 1980, pp. 130 – 147.
228. P. Hagan, P. J. Moore, A. B. Adjukiewicz, B. M. Greenwood et al.: *Parasite Immunol.* **7** (1985) 617 – 624.
229. A. E. Butterworth, M. Capron, J. S. Cordingley, P. R. Dalton, *Trans. R. Soc. Trop. Med. Hyg.* **79** (1985) 393 – 408.
230. G. Gazzinelli, J. R. Lambertucci, N. Katz, R. S. Rocha, *J. Immunol.* **135** (1985) 2121 – 2127.
231. C. W. Todd, R. W. Goodgame, D. Colley, *J. Immunol.* **122** (1979) 1440 – 1446.
232. R. F. Sturrock, R. Kimani, B. I. Cottrell, A. E. Butterworth, *Trans. R. Soc. Trop. Med. Hyg.* **77** (1983) 363 – 371.
233. A. E. Butterworth, R. F. Sturrock, V. Houba, A. A. F. Mahmoud, Nature (*London*) **256** (1975) 727 – 729.
234. A. E. Butterworth, *Adv. Parasitol.* **23** (1984) 143 – 235.
235. S. R. Smithers, R. J. Terry, *Adv. Parasitol.* **14** (1976) 399 – 422.
236. A. Capron, *Fortschr. Zool.* **27** (1982) 259 – 264.
237. A. E. Butterworth, P. Hagan, *Parasit. Today* **3** (1987) 11 – 16.
238. A. J. G. Simson, D. Cioli, *Parasit. Today* **3** (1987) 26 – 28.
239. C. Dissous, I. M. Crzych, A. Capron, *J. Immunol.* **129** (1982) 2232 – 2234.
240. Q. D. Bickle, B. I. Andrews, M. G. Taylor, *Parasit Immunol.* **8** (1986) 95 – 107.
241. C. Kelly, A. I. G. Simpson, E. Fox, S. M. Phillips et al.: *Parasit Immunol.* **8** (1986) 193 – 198.
242. J. M. Balloul, P. Sondermeyer, D. Dreyer, M. Capron, *Nature (London)* **326** (1987) 149 – 153.
243. D. Wakelin: *Immunity to Parasites,* Edward Arnold Ltd., London 1984.
244. S. Friedmann, *Gen Eng. News* (1985) 24 – 25.
245. V. K. Vinayak, N. K. Gupta, A. K. Chopra, G. L. Sharma et al., *Parasitology* **82** (1981) 375 – 382.
246. P. Ambroise-Thomas in V. Houba (ed.): *Immunological Investigation of Tropical Parasitic Diseases,* Churchill Livingstone, Edinburgh 1980, pp. 84 – 103.
247. B. M. Ogilvie, M. J. Worms in S. Cohen, E. H. Sadun (eds.): *Immunology of Parasitic Infections,* Blackwell Scientific Publ., Oxford 1976, 397 – 402.
248. M. M. Wong, H. I. Fredericks, C. P. Ramachandran, *Bull. W.H.O.* **40** (1969) 493 – 501.
249. C. P. Ramachandran, *Southeast Asian J. Trop. Med. Public Health* **1** (1970) 78 – 92.
250. A. Haque, A. Capron in T. W. Pearson (ed.): *Parasite Antigens,* Marcel Dekker Inc., New York 1986, pp. 317 – 402.
251. R. M. Maizels, F. Partono, Sri Oemijati, D. A. Denham et al., *Parasitology* **87** (1983) 249 – 263.
252. M. D. Rickard in C. Arme, P. W. Pappas (eds.): *Biology of the Eucestoda,* Academic Press, London 1983, pp. 539 – 579.
253. M. D. Rickard, F. F. Williams, *Adv. Parasitol.* **21** (1982) 229 – 296.
254. R. P. Herd, R. I. Chappel, D. Biddell, *Int. J. Parasitol.* **5** (1975) 395 – 399.
255. D. J. Jenkins, M. D. Rickard, *Aust. Vet. J.* **63** (1986) 40 – 42.
256. R. S. Nussenzweig, J. Vanderberg, H. Most, C. Orton, *Nature (London)* **216** (1967) 160 – 162.
257. W. A. Siddiqui, *Science* **197** (1977) 388 – 389.
258. G. H. Mitchell, W. H. G. Richards, G. A. Butcher, S. Cohen, *Lancet (i)* (1977) 1335 – 1358.
259. W. Trager, J. B. Jensen, *Science* **193** (1976) 673 – 675.
260. D. Mazier, R. L. Beaudoin, S. Mellouk, P. Druilhe et al., *Science* **227** (1985) 440 – 442.
261. C. Guerin-Marchand, P. Druilhe, B. Galey, A. Londono et al., *Nature (London)* **329** (1987) 164 – 167.
262. A. H. Cochrane in J. P. Kreier (ed.): *Malaria, Immunology and Immunization* vol. 3, Chap. 4, Academic-Press, New-York, 1980.
263. D. F. Clyde, *Am. J. Trop. Med. Hyg.* **24** (1975) 397 – 401.
264. E. H Nardin, R. S. Nussenzweig, I. A. McGregor, J. H. Bryan, *Science* **206** (1979) 597 – 599.
265. G. Del Giudice, H. D. Engers, C. Tougne, S. S. Biro et al., *Am. J. Trop. Med. Hyg.* **36** (1987) 203 – 212.
266. D. H. Chen, R. E. Tigelaar, F. I. Weinbaum, *J. Immunol.* **118** (1977) 1322 – 1327.
267. A. Ferreira, L. Schofield, V. Enea, H. Schellekins et al., *Science* **232** (1986) 881 – 884.
268. P. Potocnjak, N. Yoshida, R. S. Nussenzweig, V. Nussenzweig, *J. Exp. Med.* **151** (1980) 1504 – 1513.
269. D. Mazier, S. Mellouk, R. L. Beaudoin, B. Texier et al., *Science* **231** (1986) 156 – 159.

270. J. B. Dame, J. L. Williams, T. F. McCutchan, J. L. Weber et al., *Science* **225** (1984) 593 – 599.
271. V. Enea, J. Ellis, F. Zavala, D. E. Arnot et al., *Science* **225** (1984) 628 – 630.
272. D. J. Kemp, R. L. Coppel, R. F. Anders, *Ann. Rev. Microbiol.* **41** (1987) 181 – 208.
273. V. F. De la Cruz, A. A. Lal, T. F. McCutchan, *J. Biol. Chem.* **262** (1987) 11935 – 11941.
274. M. J. Lockyer, R. T. Schwarz, *Mol. Biochem. Parasitol.* **22** (1987) 101 – 107.
275. G. Del Giudice, A. S. Verdini, M. Pinori, J. P. Verhave et al., *J. Clin. Microbiol.* **25** (1987) 91 – 96.
276. S. L. Hoffman, R. Wistar, Jr., W. R. Ballou, M. R. Hollingdale et al., *New-Engl. J. Med.* **315** (1986) 601 – 603.
277. D. A. Herrington, D. F. Clyde, G. Losonsky, M. Cortesia et al., *Nature (London)* **328** (1987) 257 – 259.
278. W. R. Ballou, S. L. Hoffman, J. A. Sherwood, M. R. Hollingdale et al., *Lancet* **I** (1987) 1277 – 1279.
279. G. Del Giudice, J. A. Cooper, J. Merino, A. S. Verdini et al., *J. Immunol.* **137** (1986) 2952 – 2955.
280. M. F. Good, J. A. Berzofsky, W. L. Maloy, Y. Hayashi et al., *J. Exp. Med.* **164** (1986) 655 –660.
281. A. R. Togna, G. Del Giudice, A. S. Verdini, F. Bonetti et al., *J. Immunol.* **137** (1986) 2956 – 2960.
282. M. F. Good, D. Pombo, I. A. Quakyi, E. M. Riley et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 1199 – 1203.
283. L. H. Miller, S. J. Mason, D. F. Clyde, M. H. McGinis, *New-Eng. J. Med.* **295** (1976) 302 – 304.
284. M. J. Miller, *Trans. Roy. Soc. Trop. Med. Hyg.* **52** (1958) 152 – 158.
285. S. Cohen, I. A. McGregor, S. C. Carrington, *Nature (London)* **192** (1961) 733 – 735.
286. A. Celada, A. Cruchau, L. H. Perrin, *Clin. Exp. Immunol.* **47** (1982) 635 – 641.
287. P. H. David, *Proc. Natl. Acad. Sci. USA* **80** (1983) 5075 – 5081.
288. L. H. Miller, R. J. Howard, R. Carter, M. F. Good et al., *Science* **234** (1986) 1349 – 1356.
289. L. H. Perrin, E. Ramirez, L. Er-Hsiang, P.-H. Lambert; *Clin. Exp. Immunol.* **41** (1980) 91 – 106.
290. A. A. Holder, R. R. Freeman, *J. Exp. Med.* **156** (1982) 1528 – 1538.
291. J. S. McBride, D. Walliker, G. Morgan, *Science* **217** (1982) 254 – 256.
292. A. A. Holder, M. J. Lockyer, K. G. Odink, J. S. Sandhu et al., *Nature (London)* **317** (1985) 270 – 273.
293. K. G. Oding, M. J. Lockyer, S. C. Nicholls, Y. Hillman et al., *FEBS Lett.* **173** (1984) 108 – 112.
294. M. Mackay, M. Goman, N. Bone, J. E. Hyde et al., *EMBO J.* **4** (1985) 3823 – 3829.
295. L. H. Perrin, B. Merkli, M. Loche, C. Chizzolini et al., *J. Exp. Med.* **160** (1984) 441 – 447.
296. R. Hall, J. E. Hyde, M. Goman, D. L. Simmons et al., *Nature (London)* **311** (1984) 379 – 382.
297. W. A. Siddiqui, L. Q. Tam, K. J. Kramer, G. S. N Hui et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 3014 – 3018.
298. A. A. Holder, R. R. Freeman, *Nature (London)* **294** (1981) 361 – 363.
299. M. E. Patarroyo, P. Romero, M. L. Torres, P. Clavijo et al., *Nature (London)* **328** (1985) 629 – 631.
300. A. Cheung, J. Leban, A. R. Shaw, B. Merkli et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 8323 – 8332.
301. L. H. Perrin, R. Dayal, *Immunol. Rev.* **61** (1982) 245 – 268.
302. R. R. Freeman, A. J. Trejosiewicz, G. A. Cross, *Nature (London)* **284** (1980) 366 – 369.
303. L. H. Perrin, E. Ramirez, P.-H. Lambert, P. A. Miescher, *Nature (London)* **289** (1981) 301 – 303.
304. C. Braun-Breton, M. Jendoubi, E. Brunet, L. Perrin et al., *Mol. Bioch. Parasitol.* **20** (1986) 33 – 38.
305. C. Braun-Breton, T. L. Rosenberg, L. Pereira Da Silva, *Nature (London)* **332** (1988) 457 – 459.
306. U. Certa, P. Ghersa, H. Dobeli, H. Matile et al., *Science* **240** (1988) 1036 – 1038.
307. P. Dubois, J. P. Dedet, T. Fandeur, C. Roussilhon et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 229 – 335.
308. L. H. Perrin, B. Merkli, M. S. Gabra, J. Stocker et al., *J. Clin. Invest.* **75** (1985) 1718 – 1725.
309. J. F. Dubremetz, P. Delplace, B. Fortier, G. Tronchin et al., *Mol. Biochem. Parasitol.* **27** (1988) 135 – 142.
310. R. L. Coppel, J. G. Culvenor, A. E. Bianco, P. E. Crewther et al., *Mol. Biochem. Parasitol.* **25** (1987) 73 – 81.
311. H. Perlamnn, K. Berzins, M. Wahlgren, J. Carlsson et al., *J. Exp. Med.* **159** (1984) 1686 – 1693.

312. G. V. Brown, I. G. Culvenor, P. E. Crewther, A. E. Bianco et al., *J. Exp. Med.* **162** (1985) 774 – 779.
313. B. Wahlin, M. Wahlgren, H. Perlamnn, H. Berzins et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 7912 – 7916.
314. J. M. Favarolo, R. L. Coppel, L. M. Corcoran, S. J. Foote et al., *Nucl. Acid. Res.* **14** (1986) 8265 –8277.
315. W. E. Collins, R. F. Anders, M. Pappaioanou, G. H. Campbell et al., *Nature (London)* **323** (1986) 259 – 262.
316. S. A. Luse, L. H. Miller, *Am. J. Trop. Med. Hyg.* **20** (1971) 655 – 670.
317. T. Triglia, H. D. Stahl, P. E. Crewther, D. Scanlon et al., *EMBO J.* **6** (1987) 1413 – 1419.
318. M. Koenen, A. Scherf, O. Mercereau, G. Langsley et al., *Nature (London)* **311** (1984) 382 – 385.
319. I. J. Udeinya, L. H. Miller, I. A. McGregor, J. B. Jensen, *Nature (London)* **231** (1983) 429 – 431.
320. K. Marsh, R. J. Howard, *Science* **231** (1986) 150 – 153.
321. S. Pollack, J. Fleming, *Brit. J. Haematol.* **58** (1984) 289 – 294.
322. M. E. Perkins, *J. Exp. Med.* **160** (1984) 788 – 797.
323. D. Camus, T. J. Hadley, *Science* **230** (1985) 553 – 557.
324. P. Delplace, B. Fortier, G. Tronchin, J. F. Dubremetz et al., *Mol. Biochem. Parasitol.* **23** (1987) 193 – 202.
325. M. E. Patarroyo, R. Amador, P. Lavijo, A. Moreno et al., *Nature (London)* **332** (1988) 158 – 161.
326. R. W. Gwadz, *Science* **193** (1976) 1150 – 1153.
327. K. N. Mendis, G. A. T. Targett, *Nature (London)* **277** (1979) 289 – 391.
328. R. W. Gwadz, L. C. Koontz, *Infect. Immun.* **44** (1984) 137 – 145
329. N. Kumar, R. Carter, *Mol. Biochem. Parasitol.* **13** (1984) 333 – 341.
330. K. Mendis, P. Udayama, R. Carter, P. H. David, *J. Cell Biochem. Suppl.* **10 A** (1986) 149 – 156.
331. A. N. Vermuelen, T. Ponnudurai, P. J. A. Beckers, J. P. Verhave et al., *J. Exp. Med.* **162** (1985) 1460 – 1476.
332. J. Rener, P. M. Graves, R. Carter, J. L. Williams et al., *J. Exp. Med.* **158** (1983) 976 – 980.
333. P. M. Graves, R. Carter, T. R. Burkot, J. Rener et al., *Infect. Immun.* **48** (1985) 611 – 619.
334. E. R. Stiehm, E. Ashida, K. S. Kim, D. J. Winston et al., *Ann. Intern. Med.* **107** (1987) 367 – 82.
335. E. R. Stiehm, *Pediatrics* **63** (1979) 301 – 19.
336. J. A. Finlayson, in C. S. F. Easmon, J. Jeljaszewics (eds.): *Medical Microbiology,* **vol. 1,** Academic Press, London 1982, pp. 129 – 82.
337. R. Heffron: *Pneumonia, With Special Reference to Pneumococcus Lobar Pneumonia,* The Commonwealth Fund, New York 1939.
338. R. H. Rousell, M. S. Collins, M. B. Dobkin, R. E. Louie et al., *Am. J. Med. Suppl.* **76** (1984) 40 – 145.
339. H. E. Alexander, *Am. J. Dis. Child.* **66** (1943) 172 – 187.
340. H. E. Alexander, *Am. J. Dis. Child.* **66** (1943) 160 – 171.
341. E. Merler, F. S. Rosen, *N. Engl. J. Med.* **275** (1966) 480 – 86; 536 – 542.
342. C. A. Janeway, F. S. Rosen, *N. Engl. J. med.* **275** (1966) 826 – 831.
343. P. A. M. Gross, D. Gitlin, C. A. Janeway, *N. Engl. J. Med.* **260** (1959) 170 – 178.
344. J. Stokes, E. P. Maris, S. S. Gellis, *J. Clin. Invest.* **23** (1944) 531 – 540.
345. J. Stokes, J. R. Neefe, *JAMA J. Am. Med. Assoc.* **127** (1945) 531 – 540.
346. W. M. Hammon, L. L. Coriell, J. Stokes et al., *JAMA J. Am. Med. Assoc.* **150** (1950) 739 – 49; 750 – 756.
347. C. A. Janeway, L. Apt, D. Gitlin, *Trans. Assoc. Am. Physician* **66** (1953) 200 – 202.
348. P. A. M. Gross, D. Gitlin, C. A. Janeway, *N. Engl. J. Med.* **260** (1959) 121 – 125.
349. R. G. Hamilton, *Clin. Chem. (Winston Salem N.C.)* **33** (1987) 1707 – 1725.
350. G. R. Siber, P. H. Schur, A. C. Weitzman, G. Schiffman, *N. Engl. J. Med.* **303** (1980) 178 – 182.
351. C. M. Roifman, H. M. Lederman, S. Lavi, L. D. Stein et al., *Am. J. Med.* **79** (1985) 171 – 174.
352. S. Barandun, H. Isliker, *Vox Sang.* **51** (1986) 157 – 160.
353. A. Hassig, *Vox Sang.* **51** (1986) 10 – 17.
354. D. D. Schroeder, M. L. Dumas, *Am. J. Med. Suppl.* **76** (1984) 33 – 39.
355. C. A. Janeway, E. Merler, F. S. Rosen, S. Salmon et al., *N. Engl. J. Med.* **278** (1968) 919 – 923.
356. J. Passwell, F. S. Rosen, E. Merler in B. Alving, J. Finlayson (eds.): *Immunoglobulins; characteristics and uses of intravenous preparations,* U.S.Dept. H.H.S., Washington 1979, pp. 139 – 142.

357. H. D. Ochs, S. H. Fischer, R. J. Wedgwood, *J. Clin. Immunol Suppl.* **2** (1982) 22 S – 29 S.
358. C. Cunningham-Rundles, F. P. Siegal, E. M. Smithwick, A. Lion-Boule et al., *Ann. Intern. Med.* **101** (1984) 435 – 439.
359. M. S. Collins, J. H. Dorsey, *J. Infect. Dis.* **151** (1985) 1171 – 1173.
360. J. R. Schreiber, V. A. Barrus, G. R. Siber, *Infect. Immun.* **47** (1985) 142 – 148.
361. H. D. Ochs et al., *Lancet* **2** (1980) 1158 – 1159.
362. *Medical Letter* **24** (1982) 81 – 82.
363. R. H. Johnson, R. J. Ellis, *Ann. Intern. Med.* **81** (1974) 61 – 67.
364. C. Cunningham-Rundles, L. A. Hanson, W. H. Hitzig, W. Knapp et al., *Bull. W.H.O.* **60** (1982) 43 – 47.
365. S. Krugman, *N. Engl. J. Med.* **269** (1963) 195 – 201.
366. E. B. Grosman, S. G. Stewart, J. Stokes, *JAMA J. Am. Med. Assoc.* **129** (1945) 991 – 994.
367. Medical Research Council Workingparty, *Lancet* **1** (1969) 163 – 168.
368. C. L. S. Leen, P L. Yap, D. B. L. McClelland, *Vox Sang.* **51** (1986) 278 – 286.
369. *Lancet* **1** (1983) 105 – 106.
370. G. Chirico, G. Rondini, A. Plebani, A. Chiara et al., *J. Pediatr.* **110** (1987) 437 – 442.
371. J. Amer, E. Ott, F. A. Ibbott, D. O'Brien et al., *Pediatrics* **32** (1963) 4 – 9.
372. A. A. Hertler, S. C. Ross, *J. Clin. Lab. Immunol.* **21** (1986) 177 – 181.
373. S. Baron, E. V. Barnet, R. S. Goldsmith, S. Silbergeld et al., *Am. J. Hyg.* **79** (1964) 186 – 195.
374. N. A. Kefalides, J. A. Arana, A. Bazan, M. Bocanegra et al., *N. Engl. J. Med.* **267** (1962) 317 – 324.
375. H. H. Stone, C. D. Graber, J. D. Martin, L. Kolb, *Surgery St. Louis* **58** (1965) 810 – 814.
376. R. J. Jones, E. A. Roe, J. L. Gupta, *Lancet* **2** (1980) 1263 – 1265.
377. R. L. Wasserman, *Pediatric Infectious Diseases* **5** (1986) 620 – 621.
378. J. A. Steen, *Arch. Pediatr.* **77** (1960) 291 – 304.
379. A. P. Schless, G. S. Harell, *Amer. J. Med.* **44** (1968) 325 – 329.
380. M. Santosham, R. Reid, D. M. Ambrusino, M. C. Wolff et al., *Engl. J. Med.* **317** (1987) 923 – 9.
381. S. M. Rosenthal, R. C. Millican, J. Rust, *Proc. Soc. Exp. Biol. Med.* **94** (1957) 214 – 217.
382. M. W. Fisher, *Antibiot. Chemother. (Washington D.C.)* **7** (1957) 315 – 321.
383. M. W. Fisher, M. C. Manning, *Antibiot. Chemother. (Washington D.C.)* **8** (1958) 29 – 31.
384. B. A. Waisbren, *Antibiot. Chemother. (Washington D.C.)* **7** (1957) 322 – 333.
385. B. A. Waisbren, D. Lepley, *Arch. Intern. Med.* **109** (1962) 712 – 716.
386. G. P. Bodey, B. A. Nies, N. R. Mohberg, E. J. Freireich, *JAMA J. Am. Med. Assoc.* **190** (1964) 1099 – 1102.
387. K. C. Finkel, J. C. Haworth, *Pediatrics* **25** (1960) 798 – 806.
388. D. R. Snydman, B. G. Werner, B. Heinzer-Lacey, V. P. Berardi et al., *N. Engl. J. Med.* **317** (1987) 1049 – 1054.
389. R. M. Condie, R. J. O'Reilly, *Am. J. Med.* **76** (Suppl. 30 March 1984) 134 – 141.
390. A. Hagenbeek, H. G. J. Brummelhuis, A. Donkers, A. M. Dumas et al., *J. Infect. Dis.* **155** (1987) 897 – 902.
391. D. J. Winston, W. G. Ho, C.-H. Lin, M. D. Budinger et al., *Am. J. Med.* **76** (Suppl. 30 March 1984) 128 – 133.
392. R. A. Bowden, M. Sayers, N. Flournoy, B. Newton et al., *N. Engl. J. Med.* **314** (1986) 1006 – 1010.
393. E. C. Reed et al., *J. Infect. Dis.* **156** (1987) 641 –644.
394. K. Erlendsson, T. Swartz, J. M. Dwyer, *N. Engl. J. Med.* **312** (1985) 351 – 353.
395. P. J. Mease, H. D. Ochs, R. J. Wedgood, *N. Engl. J. Med.* **304** (1981) 1278 – 1281.
396. J. M. Crennan, R. E. Van Scoy, C. H. McKenna, T. F. Smith, *Am. J. Med.* **81** (1986) 35 – 42.
397. K. S. Erlich, J. Mills, *Rev. Infect. Dis.* **8** (1986) S 439 – 445.
398. D. Ambrosino, J. R. Schreiber, R. S. Daum, G. R. Siber, *Infect. Immun.* **39** (1983) 709 – 714.
399. T. E. Harper, R. D. Christensen, G. Rothstein, *Pediatr. Res.* **22** (1987) 455 – 460.
400. S. J. Cryz, A. S. Cross, E. Furer, N. Chariatte et al., *J. Lab. Clin. Med.* **108** (1986) 182 – 189.
401. I. A. Holder, J. G. Naglich, *Am. J. Med.* **76** (1984) 161 – 167.
402. J. E. Pennington, G. J. Small, *J. Infect. Dis.* **155** (1987) 973 – 978.
403. V. G. Hemming, W. T. London, G. W. Fischer, B. L. Curfman et al., *J. Infect. Dis.* **156** (1987) 655 – 658.
404. K. N. Haque, M. H. Zaidi, S. K. Haque, H. Bahakim et al., *Pediatric Infectious Diseases* **5** (1986) 622 – 625.
405. D. Sidiropoulos, U. Herrmann, A. Morell, G. von Muralt et al., *J. Pediatr. St. Louis* **109** (1986) 505 – 508.

406. W. G. Ho, D. J. Winston, K. Bartoni, R. E. Champlin et al., Abstr. 456, 1983 Interscience Confer. Infect. Dis. Chemother., Las Vegas, Nevada.
407. C. Bunch, H. Chapel, K. R. Rai, R. P. Gale, *Blood* **70** (1987) 224 a (Abstr); *N. Engl. J. Med.* **319** (1988) 902 – 907.
408. H. D. Ochs, *Pediatr. Infect. Dis.* **6** (1987) 509 – 511.
409. C. C. Wood, J. G. McNamara, D. F. Schwarz, W. W. Merrill et al., *Pediatr. Infect. Dis.* **6** (1987) 564 – 566.
410. B. A. Silverman, A. Rubinstein, *Am. J. Med.* **78** (1985) 728 – 736.
411. D. Sidiropoulos, U. Boehme, G. von Muralt, A. Morell et al., *Pediatr. Infect. Dis.* **5** (1986) 193 –194.
412. S. Barandun, P. Imbach, A. Morrell, H. P. Wagner in U. E. Nydegger (ed.): *Immunohemotherapy. A guide to immunoglobulin prophylaxis and therapy,* Academic Press, New York 1981, pp. 275 – 282.
413. P. Imbach, H. P. Wagner, W. Berchtold, G. Gaedicke et al., *Lancet* **2** (1985) 464 – 468.
414. J. Fehr, V. Hoffmann, U. Kappeler, *N. Engl. J. Med.* **306** (1982) 1254 – 1258.
415. S. Rosthoj, G. K. Steffensen, T. K. Guld, *Acta Paediatr. Scand.* **76** (1987) 631 – 635.
416. R. Bohm, C. Hofstaetter, R. C. Briel, *Blut* **48** (1984) 469 – 470.
417. F. E. Leickly, R. H. Buckley, *Am. J. Med.* **82** (1987) 159 – 162.
418. G. W. Richmond, I. Ray, A. Korenblitt, *J. Pediatr. (St. Louis)* **110** (1987) 917 – 919.
419. M. W. Hilgartner, J. Bussel, *Am. J. Med.* **83** (1987) 35 – 39.
420. T. Becker, S. Panzer, D. Maas, V. Kiefel et al., *Br. J. Haematol.* **61** (1985) 149 – 155.
421. F. Dammacco, G. Iodice, N. Campobasso, *Br. J. Haematol.* **62** (1986) 125 – 135.
422. A. C. Newland, J. G. Treleaven, R. M. Minchinton, A. H. Waters, *Lancet* **1** (1983) 84 – 87.
423. J. B. Bussel, R. P. Kimberly, R. D. Inman, I. Schulmann et al., *Blood* **62** (1983) 480 – 486.
424. J. R. Duran-Suarez, A. Martin, M. C. Botella, S. de la Torre et al., *Haematologica* **68** (1983) 564 – 566.
425. C. A. Schiffer et al., *Blood* **64** (1984) 937 – 940.
426. G. Chirico, M. Duse, A. G. Ugazio, G. Rondini, *J. Pediatr. (St. Louis)* **103** (1983) 654 – 655.
427. E. C. Besa, M. W. McNab, A. J. Solan, M. J. Lapes et al., *Am. J. Hematol.* **18** (1985) 373 – 379.
428. W. A. McGuire, H. H. Yang, E. Bruno, J. Brandt et al., *N. Engl. J. Med.* **317** (1987) 1004 – 1008.
429. C. De la Camara, R. Arrieta, A. Gonzalez, E. Iglesias et al., *N. Engl. J. Med.* **318** (1988) 519 – 520.
430. I. M. Nilsson, E. Berntorp, O. Zettervall, *N. Engl. J. Med.* **318** (1988) 947 – 950.
431. J. W. Newburger, M. Takahashe, J. C. Burns, A. S. Beiser et al., *N. Engl. J. Med.* **315** (1986) 341 – 347.
432. M. Nagashima, M. Matsushima, H. Matsuoka, A. Ogawa et al., *J. Pediatr. (St. Louis)* **110** (1987) 710 – 712.
433. D. Y. M. Leung, J. C. Burns, J. W. Newburger, R. S. Geha, *J. Clin. Invest.* **79** (1987) 468 – 472.
434. B. Combe, B. Cosso, J. Clot, M. Bonneau et al., *Am. J. Med.* **78** (1985) 920 – 928.
435. F. C. Breedveld, A. Brand, W. G. van Aken, *J. Rheumatol.* **12** (1985) 700 – 702.
436. E. L. Arsura, A. S. Bick, N. G. Brunner, T. Namba et al., *Arch. Intern. Med.* **146** (1986) 1365 – 1368.
437. K. Kawada, P. I. Terasaki, *Exp. Hematol (N.Y.)* **15** (1987) 133 – 136.
438. E. Schuller, A. Govaerts, *Eur. Neurol.* **22** (1983) 205 – 212.
439. A. Salama, C. Mueller-Eckhardt, V. Kiefel, *Lancet* **2** (1983) 193 – 195.
440. R. P. Kimberly, J. E. Salmon, J. B. Bussel, M. K. Crow et al., *J. Immunol.* **132** (1984) 745 – 750.
441. U. Budde et al., *Scand. J. Haematol.* **37** (1986) 125 – 129.
442. A. Etzioni, S. Pollack, A. Benderly, *N. Engl. J. Med.* **318** (1988) 994.
443. G. P. Sandilands, H. I. Atrah, G. Templeton, J. E. Cocker et al., *J. Clin. Lab. Immunol.* **23** (1987) 109 – 115.
444. T. W. Jungi, S. Barandun, *Vox Sang.* **49** (1985) 9 – 19.
445. W. Stohl, *Clin. Exp. Immunol.* **62** (1985) 200 – 207.
446. F. Hashimoto, Y. Sakiyama, S. Matsumoto, *Clin. Exp. Immunol.* **65** (1986) 409 – 415.
447. H. H. Hodes, *Pediatrics.* **32** (1963) 1 – 3.
448. W. B. White, C. R. Desbonnet, M. Ballow, *Am. J. Med.* **83** (1987) 431 – 434.
449. K. Prellner, P. Christensen, O. Kalm, K. Offenbartl, *Acta Pathol. Microbiol. Scand. Sect.* **94 C** (1986) 207 – 211.

450. K. K. Christensen, P. Christensen, *Pediatric Infectious Diseases* **5** (1986) S 189 – 192.
451. J. G. Kelton, C. J. Carter, C. Rodger, G. Bebenek et al., *Blood* **63** (1984) 1434 – 1438.
452. A. S. Cross, B. M. Alving, J. C. Sadoff, P. Baldwin et al., *Lancet* **1** (1984) 912.
453. M. Berger, P. Rosencranz, C. Y. Brown, *Clin. Immunol. Immunopathol.* **34** (1985) 227 – 236.
454. J. Kulics, E. Rajnavolgyi, G. Fust, J. Gergely, *Mol. Immunol.* **20** (1983) 805 – 810.
455. E. F. Ellis, C. S. Henney, *J. Allergy* **43** (1969) 45 – 54.
456. R. C. Williams, O. J. Mellbye, G. Kronvall, *Infect. Immun.* **6** (1972) 316 – 323.
457. D. S. Feingold, F. Oski, *Arch. Intern. Med.* **116** (1965) 226 – 228.
458. R. C. Millican, J. D. Rust, *J. Infect. Dis.* **107** (1960) 389 – 394.
459. E. Lachman, S. B. Pitsoe, S. L. Gaffin, *Lancet* **1** (1984) 981 – 983.
460. M. A. Thomson, *J. S. Afr. Vet. Assoc.* **54** (1983) 279 – 281.
461. M. L. Forman, R. Stiehm, *N. Engl. J. Med.* **281** (1969) 926 – 931.
462. A. S. Shigeoka, R. T. Hall, H. R. Hill, *Lancet* **1** (1978) 636 – 638.
463. E. J. Ziegler, J. A. McCutchan, J. Fierer, M. P. Glauser et al., *N. Engl. J. Med.* **307** (1982) 1225 – 1230.
464. J.-D. Baumgartner, M. P. Glauser, J. A. McCutchan, E. J. Ziegler et al., *Lancet* **2** (1985) 59 –63.
465. T. Calandra, J. Schellekens, J. Verhoef, M. P. Glauser, *Program and Abstracts for the 4th International Symposium on Infections in the Immunocompromised Host,* Ronneby Brunn, Sweden, 1986. Abstract No. 128.
466. A. H. Levy, *J. Chronic Dis.* **15** (1962) 589 – 598.
467. J. Bjorkander, C. Cunningham-Rundles, P. Lundin, R. Olsson et al., *Am. J. Med.* **84** (1988) 107 – 111.
468. R. S. Lane, *Lancet* **2** (1983) 974 – 975.
469. H. D. Ochs, S. H. Fischer, F. S. Virant, M. L. Lee et al., *Lancet* **1** (1985) 404 – 405.
470. C. L. S. Leen, P. L. Yap, G. Neill, D. B. L. McClelland et al., *Vox Sang.* **50** (1986) 26 – 32.
471. B. Cuthbertson, R. J. Perry, P. R. Foster, K. G. Reid, *J. Infect.* **15** (1987) 125 – 133.
472. C. C. Wood, A. E. Williams, J. G. McNamara, J. A. Annunziata et al., *Ann. Intern. Med.* **105** (1986) 536 – 538.
473. Y. Hamamoto, S. Harada, N. Yamamoto, Y. Uemura et al., *Vox Sang.* **53** (1987) 65 – 69.
474. *Morb. Mort. Wkly. Rep.* **36** (1987) 728 – 729.
475. A. W. Burks, H. A. Sampson, R. H. Buckley, *N. Engl. J. Med.* **314** (1986) 560 – 564.
476. R. Apfelzweig, D. Piskiewicz, J. A. Hooper, *J. Clin. Immunol.* **7** (1987) 46 – 50.
477. R. Paganelli, I. Quinti, G. P. D'Offizi, C. Papetti et al., *Vox Sang.* **51** (1986) 87 – 91.
478. J. C. Barton, G. A. Herrera, J. H. Galla, L. F. Bertoli et al., *Am. J. Med.* **82** (1987) 624 – 629.
479. A. G. Brox, D. Cournoyer, M. Sternbach, G. Spurll, *Am. J. Med.* **82** (1987) 633 – 635.
480. R. Schroff, K. Foon, S. Beatty, R. Oldham et al.: "Human anti-murine immunoglobulin responses in patients receiving monoclonal antibody therapy," *Can. Res.* **45** (1985) 879 – 885.
481. M. Nose, H. Wigzell: "Biological significance of carbohydrate chains on monoclonal antibodies," *Proc. Natl. Acad. Sci. U.S.A.* **80** (1983) 6632 – 6636.
482. E. Englemann, S. Foung, J. Larrick, A Raubitschek (eds.): *Human Hybridomas and Monoclonal Antibodies,* Plenum Press, New York 1985.
483. A. Strelkelkaus (ed.): *Human Hybridomas; Diagnostic and Therapeutic Applications,* Marcel Dekker, New York 1986.
484. D. Kozbor, J. Roder, T. Chang, Z. Stplewski et al.: "Human antitetanus toxoid monoclonal antibody secreted by EBV-transformed human B cells fused with murine myeloma," *Hybridoma* **1** (1982) 323 – 328.
485. J. Larrick, K. Truitt, A. Raubitschek, G. Senyk et al., "Characterization of human hybridomas secreting antibody in tetanus toxoid," *Proc. Natl. Acad. Sci. U.S.A.* **80** (1983) 6376 – 6380.
486. N. Chiorazzi, R. Wasserman, H. Kunkel: "Use of Epstein-Barr virus-transformed B cell lines for the generation of immunoglobulin-producing human B cell hybridomas," *J. Exp. Med.* **156** (1982) 930 – 935.
487. F. Gigliotti, R. Insel: "Protective human hybridoma antibody to tetanus toxin," *J. Clin. Invest.* **70** (1982) 1306 – 1309.
488. L. Olsson, T. Mazauric, J. Vincent-Falquet, J. Armand: "A human monoclonal antibody specific for tetanus toxoid," *Dev. Biol. Stand.* **57**(1984) 87 – 91.
489. S. Tsuchiya, S. Yokoyama, O. Yoshie, Y. Ono: "Production fo diphtheria antitoxin antibody in Epstein-Barr virus induced lymphoblastoid cell lines." *J. Immunol.* **124** (1980) 1970 – 1976.

490. E. Ziegler, J. McCutchan, J. Fierer et al.: "Treatment of Gram-negative bacteremia and shock with human anti-serum to a mutant Escherichia coli," *N. Engl. J. Med.* **307** (1982) 1225 – 1230.
491. W. Bogard, E. Hornberger, P. Kung in E. Englemann, S. Foung, J. Larrick, A. Raubitscheck (eds.): "Production and characterization of human monoclonal antibodies against Gram-negative bacteria," *Human Hybridomas and Monoclonal Antibodies,* Plenum Press, New York 1985, pp. 95 – 112.
492. N. Teng, H. Kaplan, J. Hebert et al.: "Protection against Gram-negative bacteremia and endotoxemia with human monoclonal IgM antibodies," *Proc. Natl. Acad. Sci. U.S.A.* **82** (1985) 179 – 184.
493. J. Larrick, M. Jahnsen, G. Senyk, S. Weiss et al. in H. Friedman (ed.): Production of human monoclonal antibodies recognizing cross-reactive determinants on lipopolysaccharides," *The Immunology and Immunopharmacology of Endotoxins,* 1986, pp. 75 – 81.
494. M. Pollack, A. Raubitschek, J. Larrick: "Cross-reactive human monoclonal antibodies that recognize conserved epitopes in the core-lipid A region of endotoxin," *J. Clin. Invest.* **79** (1987) 1421 – 1430.
495. J. Baumgartner, M. Glauser, J. McCutchan, E. Zeigler et al.: "Prevention of gramnegative shock and death in surgical patients by antibody to endotoxin core glycolipid." *Lancet* **2** (1985) 59 – 63.
496. J. Larrick, S. Hart, D. Lippman, M. Glembourtt et al. in A. Strelkelkaus (ed.): "Generation and characterization of human monoclonal anti-Pseudomonas aeruginosa antibodies," *Human Hybridomas; Diagnostic and Therapeutic Applications,* Marcel Dekker, New York 1986, pp. 65 – 80.
497. J. Larrick, B. Dyer, G. Senyk et al. in E. Engleman, S. Foung, J. Larrick, A. Raubitschek (eds.): "In vitro expression of human B cells for the production of human monoclonal antibodies," *Human Hybridomas and Monoclonal Antibodies,* Plenum Press, New York 1985, pp. 149 – 165.
498. K. Hunter, Jr., G. Fischer, V. Hemming, S. Wilson et al.: "Antibacterial activity of a human monoclonal antibody to Haemophilus influenzae type B capsular polysaccharide," *Lancet* **2** (1982) 789 – 799.
499. H. Peltola, H. Kayhty, M. Virtanen, P. Makela: "Prevention of Hemophilus influenza type b bacteremic infections with the capsular polysaccharide vaccine," *N. Engl. J. Med.* **310** (1984) 1561 –1565.
500. T. Atlaw, D. Kozbor, J. Roder: "Human monoclonal antibodies against Mycobacterium Leprae," *Infect. Immun.* **49** (1985) 104 – 110.
501. B. Brodeur, L. Lagace, Y. Larose, M. Martin et al. in B. Schook (ed.): "Mouse-human myeloma partner for the production of heterohybridomas," *Monoclonal Antibodies* , Marcel Dekker, 1986, p. 51.
502. M. Steinitz, S. Tamir, A. Goldfarb: "Human anti-pneumococci antibody produced by an Epstein-Barr virus (EBV)-immortalized cell line," *J. Immunol.* **132** (1984) 877 – 882.
503. J. Schwaber, M. Posner, S. Schlossman, H. Lazarus: "Human-human hybrids secreting pneumococcal antibodies," *Hum. Immunol.* **9** (1984) 137 – 142.
504. D. Emanuel, J. Gold, J. Colacino, C. Lopez et al.: "A human monoclonal antibody to cytomegalovirus (CMV)," *J. Immunol.* **133** (1984) 2202 – 2205.
505. C. Amadei, S. Michelson, J. Frot, M. Fruchart et al.: "Human anticytomegalovirus (CMV) immunoglobulins secreted by EBV-transformed B-lymphocytes cell lines," *Dev. Biol. Stand.* **57** (1984) 283 – 286.
506. K. Burnett, J. Leung, J. Marinis in E. Engleman, S. Foung, J. Larrick, A. Raubitschek (eds.): "Human monoclonal antibodies to defined antigens," *Human Hybridomas and Monoclonal Antibodies,* Plenum Press, New York 1985, pp. 113 – 133.
507. Y. Ichimori, K. Sasano, H. Itoh, S. Hitosumachi et al.: "Establishment of hybridomas secreting human monoclonal antibodies against tetanus toxin and hepatitis B virus surface antigen," *Biochem. Biophys. Res. Commun.* **129** (1985) 26 – 38.
508. E. Stricker, R. Tiebout, P. Lelie, W. Zeijlemaker: "A human monoclonal IgG antihepatitis B surface antibody; production, properties and applications," *Scand. J. Immunol.* **22** (1985) 337 – 345.
509. R. Beasley et al.: "Hepatitis B immune globulin (HBIG) efficacy in the interruption of perinatal transmission of hepatitis B virus carrier state," *Lancet* **2** (1981) 388 – 393.
510. S. Foung et al.: "Human monoclonal antibodies neutralizing Varicella Zoster virus," *J. Infect. Dis.* **52** (1985) 280 – 285.

511. L. Evans, J. Homsy, W. Morrow, I. Gaston et al.: "Human monoclonal antibody directed against gag gene products of the human immunodeficiency virus," *J. Immunol.* **140** (1988) 941 – 943.
512. B. Banapour et al.: "Characterization and epitope mapping of a human monoclonal antibody reactive with the envelope glycoprotein of human immunodeficiency virus," *J. Immunol.* **138** (1987) 4027 – 33.
513. A. Rosen, K. Persson, G. Klein: "Human monoclonal antibodies to a genus-specific chlamydial antigen, produced by EBV-transformed B cells," *J. Immunol.* **130** (1983) 2899 – 2902.
514. S. Koizumi, S. Fujiwara, H. Kikuta et al.: "Production of human monoclonal antibodies against Epstein-Barr virus specific antigens by the virus immortalized lymphoblastoid cell lines," *Virology* **150** (1986) 161 – 170.
515. J. Seigneurin et al.: "Herpes simplex virus glycoprotein D; human monoclonal antibody produced by bone marrow cell line," *Science* **221** (1983) 173 – 175.
516. Y. Masuho et al.: "Generation of hybridomas producing human monoclonal antibodies against herpes simplex virus after in vitro stimulation," *Biochem. Biophys. Res. Commun.* **135** (1986) 495 – 505.
517. L. Evans, C. Maragos, J. May: "Human lymphoblastoid cell lines established from peripheral blood lymphocytes secreting immunoglobulins directed against herpes simplex virus," *Immunol. Lett.* **8** (1984) 39 – 50.
518. S. Matsushita, M. Robert-Gurhoff, J. Trepel, J. Cossman et al.: "Human monoclonal antibodies directed against an envelope glycoprotein of human T-cell leukemia virus type 1," *Proc. Natl. Acad. Sci. U.S.A.* **83** (1986) 2672 – 2677.
519. C. Croce, A. Linnenbach, W. Hall, Z. Steplewski et al.: "Production of human hybridomas secreting antibodies to measles virus," *Nature (London)* **288** (1980) 488 – 489.
520. R. Ritts, Jr., A. Ruiz-Arguelles, K. Weyl et al.: "Establishment and characterization of a human non-secreting plasmacytoid cell line and its hybridization with human B cells," *Int. J. Cancer* **31** (1983) 133 – 151.
521. J. Hilfenhaus, E. Kanzy, R. Kohler, W. Willems: "Generation of human anti-rubella monoclonal antibodies from human hybridomas constructed with antigen-specific Epstein-Barr virus transformed cell lines," *Behring Inst. Mitt.* **80** (1986) 31 – 40.
522. F. van Meel, P. Steenbakkers, J. Oomen: "Human and chimpanzee monoclonal antibodies," *J. Immunol. Methods* **80** (1985) 267 – 280.
523. D. Crawford et al.: "Production of human monoclonal antibody to X 31 influenze virus nucleoprotein," *J. Gen. Virol.* **64** (1983) 697 – 700.
524. R. Schmidt-Ullrich, J. Brown, R. Whittle, P. Lin: "Human-human hybridomas secreting monoclonal antibodies to the M.W. 195 000 Plasmodium falciparum blood stage antigen," *J. Exp. Med.* **163** (1986) 179 – 189.
525. R. Udomsangpetch et al.: "Human monoclonal antibodies to Pf-155, a major antigen of malaria parasite Plasmodium falciparum," *Science* **231** (1986) 55 – 59.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Inclusion Compounds

Standard Article

Jerry L. Atwood[1]

[1]Department of Chemistry, University of Alabama, Tuscaloosa, Alabama, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a14_119
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Classification
2.1. Inorganic Hosts
2.2. Organic Hosts
2.3. Biochemical Inclusion
3. Uses
3.1. Zeolites
3.2. Cyclodextrins

[Top of Page]

1. Introduction

Inclusion compounds are those formed between two substances that are intimately linked, but not with covalent bonds. One substance is regarded as the host and the other as the guest. The term clathrate was originally used to describe the situation in which the guest is completely encapsulated by the host. The first clathrate structures were elucidated in the 1940s and current understanding dates from that time. Other aspects of the chemistry of inclusion compounds include molecular recognition, supramolecular chemistry, and the chemistry of non-covalent bonds.

Although the field has developed rapidly since the fundamental ideas were developed in the 1940s, it is possible to find much earlier reports of various types of inclusion compounds. For example, the chlorine clathrate hydrate was described by FARADAY in 1923, graphite intercalates date from 1841, and cyclodextrins were reported in 1891. The early results were not always well understood, because inclusion compounds are often non stoichiometric.

[Top of Page]

2. Classification

2.1. Inorganic Hosts

Werner Compounds. Werner compounds are named after the coordination chemist who made major contributions to the field around 1900. The host molecules have the formula MX_2A_4, where M is a divalent transition metal cation, X is an anionic ligand, and A is a neutral ligand, usually a substituted pyridine. A typical example is Ni $(NCS)_2$(4-methylpyridine)$_4$. The structure shows octahedral coordination about the metal with a trans arrangement of the NCS^- ligands. The pyridine groups are arranged in a propeller fashion with an angle of twist of about 50° from a coplanar arrangement. In the solid state the molecules pack in layers with the guest molecules included between the layers in a space bounded by the NCS^- ligands. The cavities occupied by the guests are connected by two-dimensional windows in a fashion similar to that found in zeolites.

Werner compounds were used to separate xylene isomers in the late 1950s, but the problems associated with handling crystalline solids prevented commercialization of this process [1].

Hofmann-Type Compounds. A typical Hofmann compound has the formula Ni $(NH_3)_2$· Ni $(CN)_4$· 2 C_6H_6. In the structure of this compound there are two types of nickel atom, one that exhibits square planar four-coordination and one that has an octahedral arrangement of nitrogen atoms around the nickel atom. The host possesses a layered structure with the ammonia

ligands protruding into the layers. The benzene guest molecule is completely enclosed, as is shown in Figure 1.

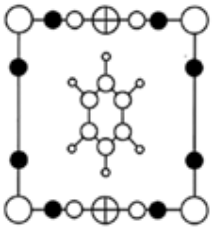

Figure 1. Hofmann's benzene clathrate [1]

⊕ = Ni(D_{4n}); ○ = Ni(O_n); ○ = C; • = N; ○ = H

Because of the limitations of space, only molecules of very specific shape and size are included. There are many ways to modify Hofmann compounds to allow the inclusion of different guests.

Gas Clathrate Hydrates. These substances are based on the structure of ice. Small molecules, usually gases, are included in voids in the hydrogen-bonded framework of water molecules. The clathrate hydrates have a geometrical similarity to zeolites in that both are based on three-dimensional four-connected networks. The guests range in size from argon to dioxane. Gas clathrate hydrates have been used for the separation of gases.

This type of host lattice is also formed by quaternary ammonium salt hydrates and alkylamine hydrates. Many of the gas clathrate hydrates possess structures which are thermodynamically more stable than that of ice. Several have melting points in the range 0 – 15 °C and one is reported to melt at 31.5 °C. Applications in heat storage have been sought but not realized.

With few exceptions, gas clathrate hydrates crystallize in one of two structures, I and II. The smaller guests belong to structure I whereas the larger prefer structure II. The ideal unit cell content for structure I is 6 X · 2 Y · 46 H_2O and for structure II, 8 X · 16 Y · 136 H_2O, where X refers to guests in 14-hedra and higher, and Y to those in 12-hedra. The polyhedra are illustrated in Figure 2. Both X and Y type spaces can be filled by a small molecule, such as methane, which forms the clathrate 8 CH_4· 46 H_2O. This clathrate hydrate, which has a melting point well above that of ice, is believed to be responsible for blocking natural gas pipelines at temperatures above 0 °C in some cases.

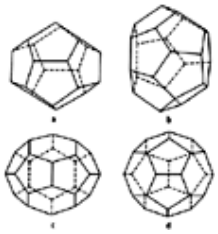

Figure 2. Illustration of the void spaces in clathrate hydrates

a) 12-hedra; b) 14-hedra; c) 15-hedra; d) 16-hedra

Zeolites (⟶ Zeolites) are porous tectosilicates with formulas such as $Ca_4(Al_8Si_4O_{12})$ · 24 H_2O or $(Na_2,Ca,Mg)_{29}(Al_{58}Si_{134}O_{384})$ · 240 H_2O. There are about sixty naturally occurring framework structures and a large number have also been synthesized. Each structure has a unique system of channels and cavities such as that shown in Figure 3, which places the total emphasis on the void system, rather than the covalent network of the host substance. The cations are usually exchangeable and share cavities along with the guests.

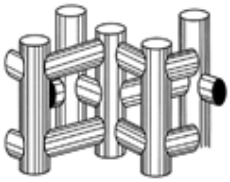

Figure 3. Channel structure of zeolite ZSM-5

The cavities are rigid and adsorption is controlled by the size of the channels and by the guests that inhabit them.

The channel system may be classified as one-, two-, or three-dimensional. A one-dimensional system arises when diffusion between channels is not possible because of size restrictions. The most important windows between channels are composed of 8-, 10-, and 12-rings. A typical window opening for an 8-ring is 420 pm. A representation of the window concept is presented in Figure 4, which emphasizes the 10-ring opening of the zeolite ZSM-5. This opening allows the entrance of normal alkanes and simple aromatic hydrocarbons. The opening is so narrow, however, that *p*-xylene (of critical dimension 630 pm) may be differentiated from *o*-xylene (critical dimension 690 pm), while molecules such as pentamethylbenzene (critical dimension 780 pm) are totally excluded. This aspect will be emphasized in Section Uses.

Figure 4. Representation of the openings between cavities in zeolites

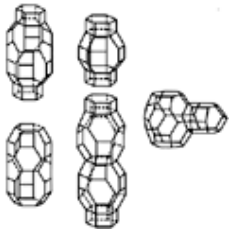

Applications in the area of electronic devices such as semiconductors will probably be realized, but have not reached commercialization as yet [2].

Intercalates. Intercalation compounds are formed by the insertion of guest species into a host lattice. Graphite is the most common substance which possesses this property (⟶ Carbon – Graphite Compounds). The process of insertion is one in which the host matrix remains essentially unchanged with regard to structure and composition, even though the layers may often be forced far apart.

There are three types of hosts from a structural standpoint, one-, two-, and three-dimensional. Graphite represents the two-dimensional type.

Hosts with special conductivity properties are finding applications in energy storage, and new analytical techniques based on intercalation are being developed.

Other Inorganic Hosts. Of the many new types of inorganic hosts, two have potential applications which are under investigation. Cyclophosphazenes are paddle-wheel shaped molecules which can entrap long-chain guest molecules. The specific orientation of molecules in the channels has led to interesting solid-state polymerization results.

The liquid inclusion complexes referred to as liquid clathrates are based on salts which can entrap aromatic hydrocarbons. Applications in the separation of aromatic from aliphatic hydrocarbons, and for the separation of the para and meta isomers of aromatic compounds are under investigation.

2.2. Organic Hosts

Crown Ethers. Crown ethers are cyclic polyethers which were first recognized as a new class of macrocyclic compounds in 1967 (⟶). Crown ethers have been typically used for the complexation of cations, but it is also possible to complex neutral molecules and anions with many members of the series.

The crown ethers are essentially two-dimensional complexing agents. Three-dimensional, oligocyclic analogues of crown ethers, the cryptands, are also known [1].

The crown ethers represent an important class of synthetic macrocycles. Applications are being sought in pharmaceuticals, metal ion removal and recovery, and in chemical sensors [2].

Cyclodextrins. Cyclodextrins are cyclic oligosaccharides which are naturally formed by the enzymatic degradation of starch. The three types most commonly obtained are alpha- (6 glucose units), beta- (7 glucose units), and gamma-cyclodextrin (8 glucose units). A schematic representation of these structures is shown in Figure 5. Since the subunit of the structures (glucose) is rigid, the cyclodextrin molecule possesses a cavity, even in the absence of a guest. The approximate dimensions of the cavities, given in Figure 5, are comparable to the molecular dimensions of many simple organic molecules. The fit of guest in host is illustrated in Figure 6 for 4-iodoaniline and alpha-cyclodextrin.

Figure 5. Representations of the structures of alpha-, beta-, and gamma-cyclodextrins (values are given in picometers)

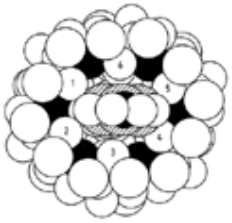

Figure 6. Structure of the inclusion complex of alpha-cyclodextrin with 4-iodoaniline using space-filling models

The cyclodextrins have hydrophobic interiors and hydrophilic exteriors, which leads to good water solubility. The exterior is easily modified chemically.

Other Organic Hosts. Urea was found to form inclusion complexes in 1940. Here the host is the lattice of urea (or thiourea) and the guest is typically a long-chain hydrocarbon. Since urea is such an inexpensive host, attempts have been made to exploit its inclusion properties. In particular, its use in the separation of benzene and cyclohexane from *n*-heptane has been

investigated, but not commercialized.

Some phenolic compounds, such as hydroquinone, phenol, and Dianin's compound have interesting inclusion properties.

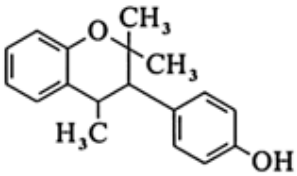

Dianin's compound

Considerable attention is now directed towards the calixarenes [3]. These molecules have cavities of molecular size for many simple organic molecules and they can be readily and economically synthesized. A schematic view of the calixarene structure and the crystal structure of an inclusion complex are shown in Figure 7A and B.

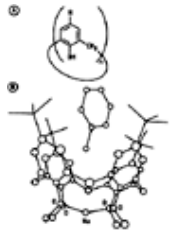

Figure 7. A) Schematic view of the structure of a calix(n)-arene (for *n* = 4, 6, or 8, the molecules are readily accessible)

B) The methyl ether of calix(4) arene with included toluene and complexed sodium ion

Other promising organic hosts under active investigation include deoxycholic acid, perhydrotriphenylene, cyclotriveratrylene, and tri-o-thymotide.

2.3. Biochemical Inclusion

Since the processes of biochemical molecular recognition and catalysis are so clearly tied to the principles of inclusion chemistry, much endeavor is aimed at duplication or improvement of certain aspects of biochemistry. In particular, enzyme mimics are being sought. These should catalyze reactions specifically and under mild conditions. The most studied enzyme mimics are the cyclodextrins. An example of a reaction of interest is the chlorination of anisole.

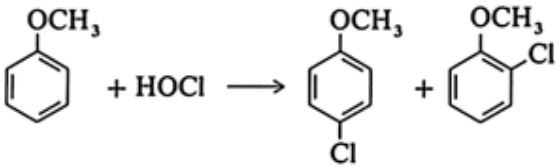

In this reaction the ratio of ortho- to para-isomer is 40 : 60. In the presence of alpha-cyclodextrin the ratio changes to 4 : 96. The blocking of the ortho position is illustrated in Figure 8. Cyclodextrin also accelerates the reaction by providing a new reaction pathway.

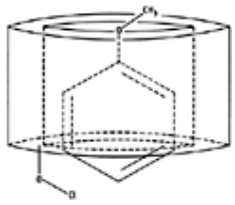

Figure 8. Schematic view of the structure of the complex of anisole and cyclodextrin

Crown ethers and cryptands have been studied for their action in cation transport. In biochemical systems it is important to solubilize cations in media of low dielectric constant. Possible applications of this research in the pharmaceutical industry may be anticipated.

[Top of Page]

3. Uses

3.1. Zeolites

Worldwide sales of naturally occurring zeolites exceeded 250 000 t in 1979, and the market is growing yearly. The major applications can be divided into the four categories addressed below.

Adsorptive Separation of Hydrocarbons. By far the largest industrial application of inclusion compounds to date involves the use of zeolites in adsorptive separations (⟶ Adsorption – Oxidic Hydrophilic Adsorbents). The zeolites contain rigid channels and cavities which are only accessible to molecules of appropriate size and shape. The first industrial application came in 1962 with the Molex process for the separation of normal alkanes from branched-chain alkanes [4]. The process uses 5 Å (500 pm) molecular sieves, and only straight-chain molecules can enter the zeolite channels. Branched-chain

alkanes are completely excluded.

The ideas which came to fruition with the Molex process have given rise to the Sorbex family of processes. In these large-scale separations, a simulated moving-bed technique is used. Sorbex processes are now used in 60 plants worldwide (Table 1) and produce a total of 4.5×10^9 t/a of purified products.

Table 1. Licensed sorbex plants [4]

Process	Separation	Number of plants
Parex	*p*-xylene/C_8 hydrocarbons	34
Olex	alkenes/alkanes	4
Molex	n-alkanes/branched and cyclic hydrocarbons	16
Sarex	fructose/dextrose and polysaccharides	4
Cresex	*p*- or *m*-cresol/cresol isomers	1
Cymex	*p*-cymene or *m*-cymene/cymene isomers	1
		60

Several features of these adsorptive processes are of note. First, in most instances one component is included, but not to the total exclusion of the others. The xylene isomers can all gain access to the channels, but some are adsorbed more strongly than others. This might be as a result of their differing basicities. The base strength of aromatic hydrocarbons relative to hydrogen fluoride [5] (relative basicity) at 0.1 molal are as follows:

toluene	0.63
p-xylene	1.0
o-xylene	1.1
m-xylene	26
mesitylene	13 000

In zeolites, the acidity of the cavities or channels, and thus the strength of adsorption, can be controlled by varying the Si : Al ratio.

It is also worthwhile to compare adsorptive separations using zeolites to those based on liquid – liquid extraction. Figure 9 shows a typical plot of selectivity versus chain length for an alkane – alkene separation using liquid – liquid extraction. The solubilities of alkenes and alkanes are similar and separation of their mixtures is ineffective. In Figure 10 an equivalent plot is shown for an adsorptive separation with zeolites. Separation of the alkanes and alkenes is effected.

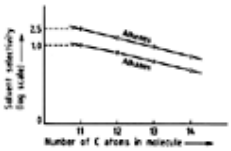

Figure 9. Liquid – liquid extraction selectivity of polar solvents in alkane – alkene systems [4]

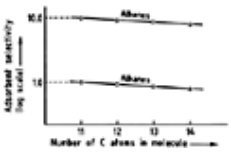

Figure 10. Adsorption selectivity of Olex adsorbent in alkane – alkene systems [4]

Purifications of Gases and Liquids. Zeolites (molecular sieves) have long been used to remove water from organic solvents. The molecular sieve is selected with a pore size that excludes the organic molecule. Regeneration is accomplished by heating the molecular sieve under vacuum to remove the guest water molecules. This same technology can be used for the purification of solvents from a range of contaminants.

Zeolites have been used to remove water from natural gas and air. They also find application in the sweetening of LPG (liquified petroleum gas) where they both dehydrate and remove sulfur-containing compounds. Zeolites are also used to

remove traces of water which may cause problems in sealed systems, such as closed cycle refrigerants and the space between dual-pane windows.

A recent application of molecular sieves is in exhaust systems for automobiles. Cooling of the system, particularly the catalytic converter, results in condensation of moisture which ultimately leads to rusting. Appropriate placement of a packet of zeolitic material in the system affords adsorption of the water, and when the engine is working the heat of the exhaust gases removes the water molecules regenerating the zeolite.

Catalytic Cracking of Hydrocarbons. The use of zeolites in catalysis represents a major share of the total zeolite market. The largest industrial process of this type is the catalytic cracking of hydrocarbons. The catalysts contain up to 40 % zeolite, often lanthanide-ion exchanged, on a clay matrix. Such catalysts improve the yield and octane number of gasoline and produce less coke than other methods. Zeolites also find extensive use in hydrocracking and in alkane isomerization. Their application in shape-selective catalysis can be expected to grow [3].

Ion Exchange. As has been noted in previous sections, it is possible to exchange the cations in zeolites for others. In general, more highly charged cations replace those having a lower charge. The major application of this is in the production of detergents. The primary function of phosphates in detergents is the complexation of Mg^{2+} and Ca^{2+} ions (water softening). Since the trend is now away from the use of phosphates, zeolites such as Type A, $(Na_2)(Al_2SiO_7) \cdot 4.5\ H_2O$, are being used as replacements. However, Type A zeolite cannot remove highly hydrated Mg^{2+} ions since its pore size is too small. Moreover, zeolites have low solubility so that their use is expected to decline as the demand for liquid detergents increases. A secondary application is in point-of-use water filters. Not only is water softening accomplished, but also heavy metal ions may be removed.

3.2. Cyclodextrins

Pharmaceuticals. Extensive studies have shown that beta-cyclodextrin is nontoxic. The first studies on humans were performed in the early 1970s. Since cyclodextrins are oligosaccharides, it is believed that their structures decompose readily under physiological conditions.

The numerous patents for pharmaceutical applications of cyclodextrins attest to the level of interest in this area [6]. Cyclodextrins are being used as drug complexation agents and as drug additives. Complexation of oral drugs provides many potential benefits: (1) liquid drugs can be transferred into crystalline materials which can then be manufactured as tablets, (2) compounds which react with one another may be formulated together, (3) smell and taste problems may be overcome, and (4) cyclodextrin – drug complexes are not hygroscopic. More generally, both oral and liquid drugs show enhanced chemical and physical stability with respect to oxidation by air, sensitivity to light, rate of disproportionation or polymerization, and losses due to evaporation. Furthermore, the solubility of drugs may be greatly enhanced. A general scheme for drug complexation and absorption is given in Figure 11.

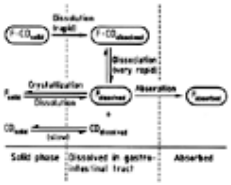

Figure 11. Representation of the dissolution – dissociation – absorption process for a drug – cyclodextrin complex, F.CD [6]

Food Additives. Complexes of cyclodextrins with food flavor molecules show the same enhanced physical and chemical properties as do those of drugs. In addition, significant improvements in the formulation of the additives make it easier to protect against contamination.

Beta-cyclodextrin has a sweet taste, comparable to that of sucrose. An interesting application is in the complexation of aspartame. The complex is marketed in Japan under the name PalSweet and is said to have no bitter aftertaste.

Although it is well established that the cavity of beta-cyclodextrin can complex molecules such as aspartame, vanillin, and menthol, the complexation of molecules such as naringin, a bitter component of grapefruit juice, is doubtful.

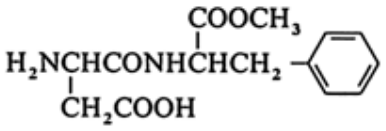

Aspartame

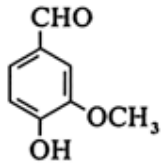

Vanillin

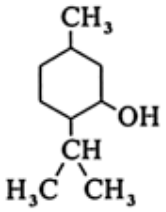

Menthol

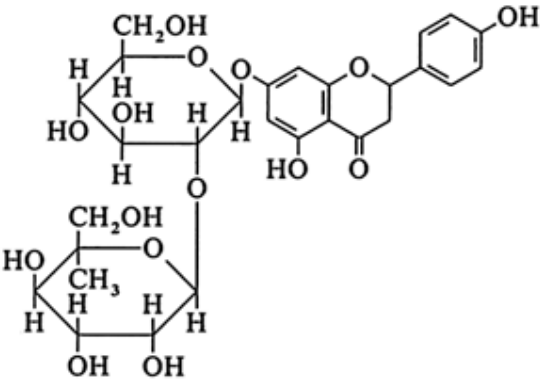

Naringin

Pesticides. The advantages of the use of cyclodextrin complexes in the pesticide industry are like those for drugs and food additives, but with two additional features. First, the pesticide is enclosed in a hydrophilic cover (the cyclodextrin). This means that absorption into the hydrophilic environment in the intestine of a pest is increased, whereas contact absorption through the hydrophobic surface of the insect exoskeleton is decreased. This in turn means that these complexes can be used to selectively attack herbivorous pests, while other insects remain largely unaffected by contact absorption. The second feature of cyclodextrin – pesticide inclusion compounds concerns the aspect of timed-release activity. A complex may be applied to dry soil and only be released when it rains.

[Top of Page]

References

1. J. L. Atwood, J. E. D. Davies, D. D. MacNicol (eds.): *Inclusion Compounds*, vol**s. 1 – 3**, Academic Press, Orlando Florida 1983, 1984.
2. J. L. Atwood (ed.): *Perspectives on Inclusion Phenomena and Molecular Recognition*, Plenum, New York 1989.
3. J. L. Atwood, J. E. D. Davies, D. D. MacNicol (eds.): *Inclusion Compounds,* vols. 4 – 5, Oxford University Press, Oxford 1990.
4. D. B. Broughton, *Sep. Sci. Technol.* **19** (1984, 1985) 723.
5. H. A. Zinnen, S. H. Hobbs, S. A. Gembicki, *Sep. Sci. Technol.* **19** (1984 – 85) 737.
6. J. Szejtli: *Cyclodextrin Technology,* Kluwer, Dordrecht 1988.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Indicator Reagents

Standard Article

Erwin Ross[1], Jutta Köthe[2], Renate Naumann[3], Wolfgang Fischer[4], Wilhelm-Dietrich Mayer[5], Gerhard Wieland[6], Ernest J. Newman[7], Charles M. Wilson[8]

[1]Merck KGaA, Darmstadt, Federal Republic of Germany
[2]Merck KGaA, Darmstadt, Federal Republic of Germany
[3]Merck KGaA, Darmstadt, Federal Republic of Germany
[4]Merck KGaA, Darmstadt, Federal Republic of Germany
[5]Merck KGaA, Darmstadt, Federal Republic of Germany
[6]Merck KGaA, Darmstadt, Federal Republic of Germany
[7]Merck Ltd. (formerly British Drug Houses Ltd.), Poole, United Kingdom
[8]EM Science, Cincinnati, Ohio, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a14_127
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. pH Indicators
2.1. Theory
2.2. Errors
2.3. Dye Classes and Individual pH Indicators
2.4. pH Indicator Papers
2.5. pH Indicators for Nonaqueous Solutions
3. Adsorption Indicators
4. Fluorescent Indicators
5. Chemiluminescent Indicators
6. Redox Indicators
6.1. Standard Potential and rH Value
6.2. Classes of Redox Indicators
7. Metal Indicators
7.1. Volumetry (Complexometry)
7.2. Colorimetry and Photometry
7.3. Structure
7.4. Ligand Classes
7.4.1. Ligands Forming Five-Membered Chelate Rings
7.4.2. Ligands Forming Six-Membered Chelate Rings
8. Analytical Test Kits

[Top of Page]

1. Introduction

Indicator reagents are substances that enable the course of a chemical reaction to be followed or the state of a chemical system to be characterized. Their names are derived from the areas in which they are used, e.g., pH indicators, metal indicators, or redox indicators.

Indicator reagents are often used to detect equivalence points during titration. They are also utilized for qualitative, semiquantitative, or quantitative analysis, particularly of inorganic ions. In this case, indicator reagents can be employed in the form of individual reagents, as components of a test kit, or as test strips or indicator papers.

Indicators that are added directly to a solution are called *internal indicators.* The concentration of an internal indicator must be very low compared to that of the substance being examined. Indicators that are not used within the observed system are called *external indicators.* These include indicator papers and test strips.

The pH (acid – base) indicators are the oldest known indicator reagents. In 1660, BOYLE recognized that certain plant extracts could be used as indicators because they changed color on treatment with acid or base. The indicator reagent in longest use is the dye known as *litmus*, which is obtained from the orseille lichen (*Roccella tinctoria*). At an early stage, litmus was coated onto cellulose to form litmus paper, which is still used widely today.

Indicator reactions that incorporate enzyme-catalyzed stages are not discussed in this article (⟶ ; ⟶ Enzymes – Amino Acid Analysis and Protein Sequence Analysis, ⟶ Enzymes – Site-Specific RNases in RNA Sequence Analysis, ⟶ Enzymes – Enzymes in Diagnosis, ⟶ Enzymes – Enzymes for Food Analysis, [6], [30]).

[Top of Page]

2. pH Indicators

The term “indicator” is often employed without further explanation to refer to pH (acid – base) indicators.

2.1. Theory

The pH indicators are substances (usually organic dyes) that change color within defined pH ranges. According to OSTWALD, the change in spectral properties as a function of pH is due to dissociation of the indicator molecule on accepting or donating protons:

$$\mathrm{In_A} \rightleftharpoons \mathrm{H^+} + \mathrm{In_B} \tag{1}$$

where In_A is the indicator acid and In_B the indicator base. The absorption spectra of In_A and In_B differ. If the laws of mass balance are applied to Equation (1) and the activity coefficients are neglected, then

$$K_A = \frac{[\mathrm{H^+}]\,[\mathrm{In_B}]}{[\mathrm{In_A}]} \tag{2}$$

$$-\log[\mathrm{H^+}] = -\log K_A + \log\frac{[\mathrm{In_B}]}{[\mathrm{In_A}]} \tag{3}$$

$$\mathrm{pH} = \mathrm{p}K_A + \log\frac{[\mathrm{In_B}]}{[\mathrm{In_A}]} \tag{4}$$

where K_A is the dissociation constant of the indicator and pK_A its negative logarithm to the base 10. According to Equation (4), the ratio $[In_B]/[In_A]$ is a function of pH. Consequently, the color of an indicator solution depends upon the pH.

If both forms of the indicator exist in equal concentrations, Equation (4) simplifies to

$$\mathrm{pH} = \mathrm{p}K_A \tag{5}$$

Indicators that are able to donate or accept several protons have several corresponding pK_A values, and the above equations must to be extended appropriately. Sometimes, for example with phenolphthalein, two pK_A values can be used to derive the following equation:

$$\mathrm{pH}_{1/2} = \frac{\mathrm{p}K_{A1} + \mathrm{p}K_{A2}}{2} \tag{6}$$

The dissociation equilibrium is not always sufficient to describe indicator properties. If, for example, a reaction takes place prior to dissociation (as is the case with nitrophenols), a new equilibrium constant K_i must be defined:

$$K_i = K_A \cdot K_U \tag{7}$$

where K_U represents the equilibrium constant of the reaction product.

The pH value corresponding to the negative logarithm to the base 10 of K_i is defined as the *transition point*:

$$\mathrm{pH} = \mathrm{p}K_i \tag{8}$$

The pK_A or pK_i value is the pH at which the color change occurs. The color change of a given indicator generally extends over a pH range. This range is called the *transition region* or *transition interval* and is defined by $pK_i \pm 1$. In this region, colors of various forms of the indicator coexist. If one form has an absorbance maximum at the wavelength λ_1 and the other at λ_2, then the plot shown in Figure 1 is obtained when the pH is varied. Curve (a) represents the spectrum of the acidic form of the indicator; curve (b), that of the alkaline form; and curve (c), a mixture of the two. All of the curves intersect at the *isosbestic point* (P).

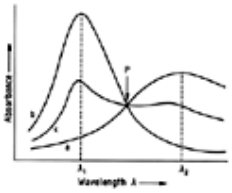

Figure 1. Absorption spectrum of a pH indicator as a function of pH (for explanation, see text)

a) Acidic form of indicator; b) Alkaline form of indicator; c) Mixture of acidic and alkaline forms

2.2. Errors

Errors occurring during the use of indicators include the following.

Salt Errors [13]. Neutral salts affect the activity coefficients of both indicator structures to varying extents. At salt concentrations > 0.2 mol/L, changes in pK_i values are generally encountered.

Protein Errors. The presence of charged macromolecules can disrupt the equilibrium because they bind to indicator molecules of the opposite charge. With some indicators (e.g., Bromophenol Blue), the resulting color shifts can be so pronounced that they are used for the analytical determination of protein.

Indicator Errors. Because the indicator can accept or donate protons, it may itself cause a pH change (e.g., in weakly buffered solutions). During titration, the consumption of reagents by the indicator must also be considered.

Alcohol Errors. The addition of alcohols and other solvents to aqueous solutions can alter the dissociation constants of indicators and, consequently, change the transition point.

2.3. Dye Classes and Individual pH Indicators [14]

The most commonly used pH indicators include azo dyes, nitrophenols, phthaleins, and sulfonphthaleins. Some other acid – base indicators belong to the classes aniline – sulfonphthaleins, triphenylmethane dyes, or benzaurins.

Variation of the substituents of the basic chromophoric indicator structures yields pH indicators with transition points that cover the complete pH range from 0 to 14. The indicators are generally used as 0.1 % solutions in water, ethanol, or water – ethanol mixtures. Typical volumes range between 0.05 and 0.5 mL.

The most commonly used pH indicators are listed in Table 1 in order of increasing pH of their transition points. A graphic overview of the pH ranges and color changes is given in Figure 2.

Table 1. Commonly used pH indicators

Transition range, pH	Indicator	CAS registry no.	Color change	Composition of indicator solution, %
0.0 – 2.0	Malachite Green (C.I. 42 000)	[2437-29-8]	yellow – green blue	0.1 (in water)
0.0 – 2.6	Brilliant Green (C.I. 42 040)	[633-03-4]	yellow – green	0.1 (in water)
0.1 – 2.3	Methyl Green (C.I. 42 590)	[82-94-0]	yellow – blue	0.1 (in water)
0.2 – 1.0	Picric acid (C.I. 10 305)	[88-89-1]	colorless – yellow	0.1 (in 70 % ethanol)
0.5 – 2.5	Cresol Red	[1733-12-6]	red – yellow	0.1 (in 20 % ethanol)
0.8 – 2.6	Crystal Violet (C.I. 42 555)	[548-62-9]	yellow – blue violet	0.1 (in 70 % ethanol)
1.2 – 2.3	Metanil Yellow (C.I. 13 065)	[587-98-4]	red – yellow	0.1 (in 20 % ethanol)
1.2 – 2.8	*m*-Cresol Purple	[2303-01-7]	red – yellow	0.04 (in 20 % ethanol)
1.2 – 2.8	Thymol Blue	[76-61-9]	red – yellow	0.1 (in 96 % ethanol)
1.2 – 2.8	*p*-Xylenol Blue	[125-31-5]	red – yellow	0.1 (in 50 % ethanol)
1.2 – 2.8	Thymol Blue sodium salt	[62625-21-2]	red – yellow	0.05 (in water)
1.4 – 2.2	Quinaldine Red	[117-92-0]	colorless – pink	0.1 (in 60 % ethanol)
1.4 – 2.6	Tropaeolin OO	[3012-37-1]	red – yellow	0.1 (in water)
2.0 – 4.4	2,6-Dinitrophenol	[573-56-8]	colorless – yellow	0.1 (in 70 % ethanol)
2.2 – 3.2	Phloxine B (C.I. 45 410)	[18472-87-2]	colorless – purple	0.1 (in water)

2.8 – 4.0	2,4-Dinitrophenol	[51-28-5]	colorless – yellow	0.1 (in 70 % ethanol)
2.9 – 4.0	4-Dimethylaminoazobenzene (Methyl Yellow, C.I. 11 020)	[60-11-7]	red – orange yellow	0.1 – 0.5 (in 90 % ethanol)
3.0 – 4.5	Bromochlorophenol Blue	[2553-71-1]	yellow – blue violet	0.1 (in 20 % ethanol)
3.0 – 4.6	Bromophenol Blue	[115-39-9]	yellow – blue violet	0.1 (in 20 % ethanol)
3.0 – 4.6	Bromophenol Blue sodium salt	[34725-61-6]	green yellow – blue violet	0.05 (in water)
3.0 – 5.2	Congo Red	[573-58-0]	blue – orange red	0.2 (in water)
3.1 – 4.4	Methyl Orange (C.I. 13 025)	[547-58-0]	red – yellow orange	0.04 (in 20 % ethanol)
3.5 – 5.8	2,5-Dinitrophenol	[329-71-5]	colorless – yellow	0.1 (in 70 % ethanol)
3.7 – 5.0	1-Naphthyl Red	[131-22-6]	red – yellow	0.1 (in 70 % ethanol)
3.8 – 5.4	Bromocresol Green	[76-60-8]	yellow – blue	0.1 (in 20 % ethanol)
3.8 – 5.4	Bromocresol Green sodium salt	[62625-32-5]	yellow – blue	0.05 (in water)
4.3 – 6.3	Alizarin (Red) S (C.I. 58 005)	[130-22-3]	yellow – pink	0.1 (in water)
4.4 – 6.2	Methyl Red (C.I. 13 020)	[493-52-7]	red – yellow – orange	0.1 (in 96 % ethanol)
4.5 – 6.2	Methyl Red sodium salt	[845-10-3]	red – yellow – orange	0.05 (in water)
4.7 – 6.3	Bromophenol Red	[2800-80-8]	yellow – purple	0.1 (in 20 % ethanol)
4.8 – 6.4	Chlorophenol Red	[4430-20-0]	yellow – purple	0.1 (in 20 % ethanol)
5.0 – 7.2	Hematoxylin (C.I. 75 290)	[517-28-2]	yellow – violet	0.05 (in 96 % ethanol)
5.0 – 8.0	Litmus (C.I. 1242)	[573-56-8]	red – blue	4 (in water)
5.2 – 6.8	Bromocresol Purple	[115-40-2]	yellow – purple	0.1 (in 20 % ethanol)
5.4 – 7.5	4-Nitrophenol	[100-02-7]	colorless – yellow	0.2 (in 96 % ethanol)
5.7 – 7.5	Bromoxylenol Blue	[40070-59-5]	yellow – blue	0.1 (in 96 % ethanol)
5.8 – 7.2	Alizarin (C.I. 58 000)	[72-48-0]	pale yellow – violet red	0.5 (in 96 % ethanol)
5.8 – 7.6	Bromothymol Blue	[76-59-5]	yellow – blue	0.1 (in 20 % ethanol)
5.8 – 7.6	Bromothymol Blue sodium salt	[34722-90-2]	yellow – blue	0.05 (in water)
6.0 – 7.0	Nitrazine Yellow (C.I. 14 890)	[5423-07-4]	brownish yellow – blue violet	0.05 (in 70 % ethanol)
6.4 – 8.2	Phenol Red	[143-74-8]	yellow – red violet	0.1 (in 20 % ethanol)
6.5 – 8.0	Phenol Red sodium salt	[34487-61-1]	yellow – red violet	0.05 (in water)
6.5 – 8.5	Cresol Red	[1733-12-6]	orange – purple	0.1 (in 20 % ethanol)
6.6 – 8.6	3-Nitrophenol	[554-84-7]	colorless – yellow	0.3 (in 96 % ethanol)
6.8 – 8.0	Neutral Red	[553-24-2]	blue red – orange yellow	0.1 (in 70 % ethanol)
7.1 – 8.3	1-Naphtholphthalein	[1301-55-9]	pink brown – blue	0.1 in (96 % ethanol)
7.4 – 9.0	*m*-Cresol Purple	[2303-01-7]	yellow – purple	0.04 (in 20 % ethanol)
7.8 – 9.5	Thymol Blue sodium salt	[62625-21-1]	green – blue	0.05 (in water)
8.0 – 9.6	Thymol Blue	[76-61-9]	yellow – blue	0.1 (in 96 % ethanol)
	p-Xylenol Blue	[125-31-5]	yellow – blue	0.1 (in 50 % ethanol)

8.0 – 9.6				
8.2 – 9.5	*o*-Cresolphthalein	[596-27-0]	colorless – red violet	0.02 (in 50 % ethanol)
8.2 – 9.8	Phenolphthalein	[81-90-3]	colorless – pink	0.1 (in 96 % ethanol)
9.3 – 10.5	Thymolphthalein	[125-20-2]	colorless – blue	0.1 (in 50 % ethanol)
9.4 – 12.0	Alizarin (Red) S (C.I. 58 005)	[130-22-3]	brown orange – violet	0.1 (in water)
10.2 – 11.8	Alizarin (C.I. 58 000)	[72-48-0]	brownish red – violet	0.5 (in 96 % ethanol)
10.2 – 12.1	Alizarin Yellow GG (C.I. 14 025)	[584-42-9]	pale yellow – brownish yellow	0.1 (in water)
11.5 – 13.0	Alkali Blue (C.I. 42 765)	[568-02-5]	blue – pink violet	0.1 (in 90 % ethanol)
11.6 – 13.0	Epsilon Blue	[73904-21-9]	orange – violet	0.1 (in water)
11.7 – 13.2	Malachite Green (C.I. 42 000)	[569-64-2]	blue – colorless	0.1 (in water)
11.7 – 14.0	Indigo Carmine (C.I. 73 015)	[860-22-0]	blue – yellow	0.25 (in 50 % ethanol)
12.0 – 14.0	Acid Fuchsin (C.I. 42 685)	[3244-88-0]	red – colorless	0.1 (in water)

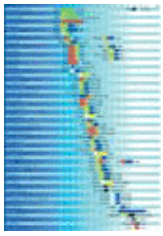

Figure 2. Color changes of commonly used pH indicators (courtesy of Merck)The pH ranges and color shades shown are approximations.

Azo Dyes. By using *4-dimethylaminoazobenzene* as an example, the chromophoric –N=N– group can be seen to display antiauxochrome (electron acceptor) properties. In combination with auxochromic (electron-donating) groups and the π-electron system of the benzene ring, mesomers can be formed.

The yellow – red color transition observed on addition of acid is caused by stabilization of the mesomer system on transformation to the dye ion.

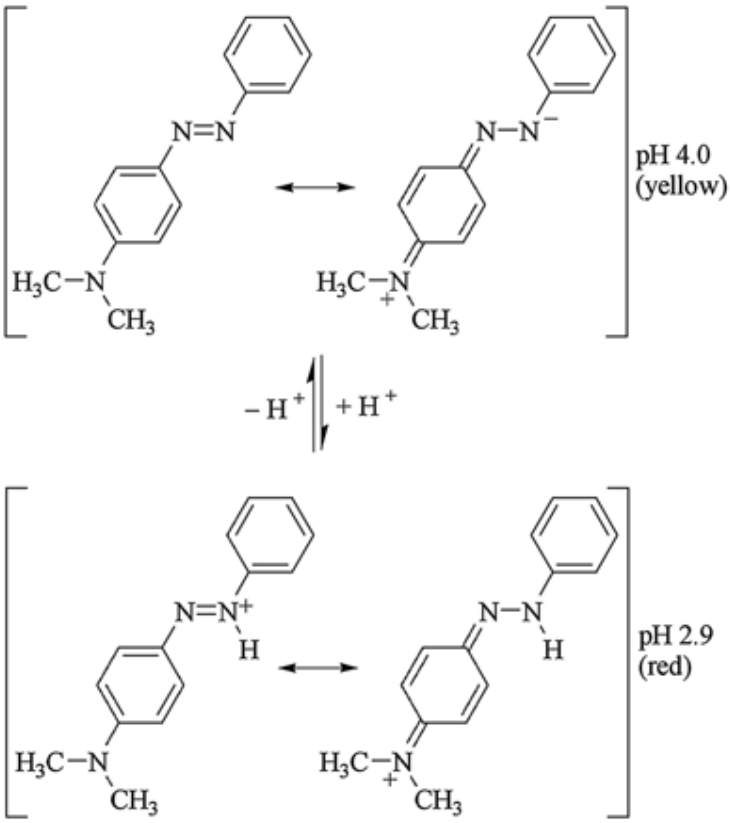

Water-soluble azo indicators can be obtained by introducing the sodium salts of carboxylic or sulfonic acids as substituents.

Azo dyes with specified transition ranges and solubility properties can be produced by changing the functionality and position of these groups on the aromatic ring, as well as by replacing the *N*-methyl group with other *N*-alkyl residues. An example is:

4-Dimethylaminoazobenzene (Methyl Yellow, Butter Yellow) is soluble in ethanol and employed along with Methylene Blue for the titration of carbonates (see Mixed Indicators.).

Methyl Orange (sodium-4-dimethylaminoazobenzene 4′-sulfonate) is soluble in water and is one of the most frequently employed acid – base indicators. It is generally used for titration of strong acids and weak bases. The drawback of this indicator is its weak color transition (red ⟶ orange ⟶ orange yellow), which makes detection of the transition point difficult.

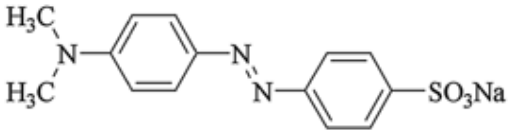

4-Dimethylaminoazobenzene-4′-sulfonamide has a direct transition from yellow to red.

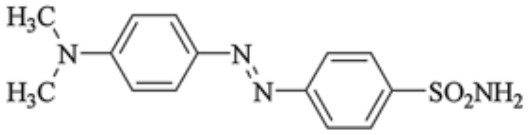

Methyl Red (4-dimethylaminoazobenzene-2'-carboxylic acid) is also frequently used for titrating strong acids and weak bases. Its color transition is more easily detected than that of Methyl Orange.

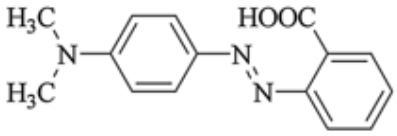

Nitrophenols. All of the nitrophenols used as pH indicators show a transition from colorless to yellow, which occurs when the colorless indicator acid is converted into the yellow aci form:

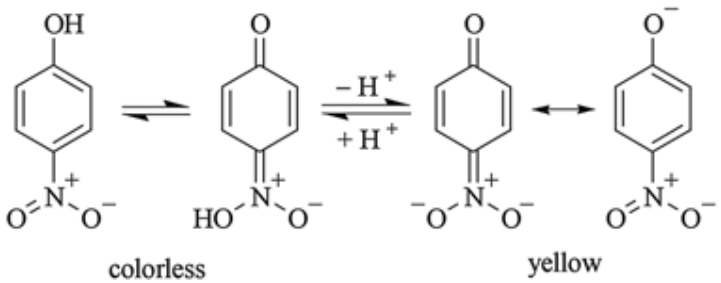

The bathochromic effect here is also attributed to improved resonance stabilization of the resulting dye ion.

The transition range of nitrophenols is governed by the number and position of nitro groups. Mononitrophenols show a transition in the pH range 5.0 – 8.6; dinitrophenols between pH 2.0 and 5.8; and picric acid between pH 0.2 and 1.0 (see Table 1).

Phthaleins. Perhaps the best known phthalein indicator is phenolphthalein. The molecular structure of phenolphthalein and its pH dependence are representative of all phthaleins:

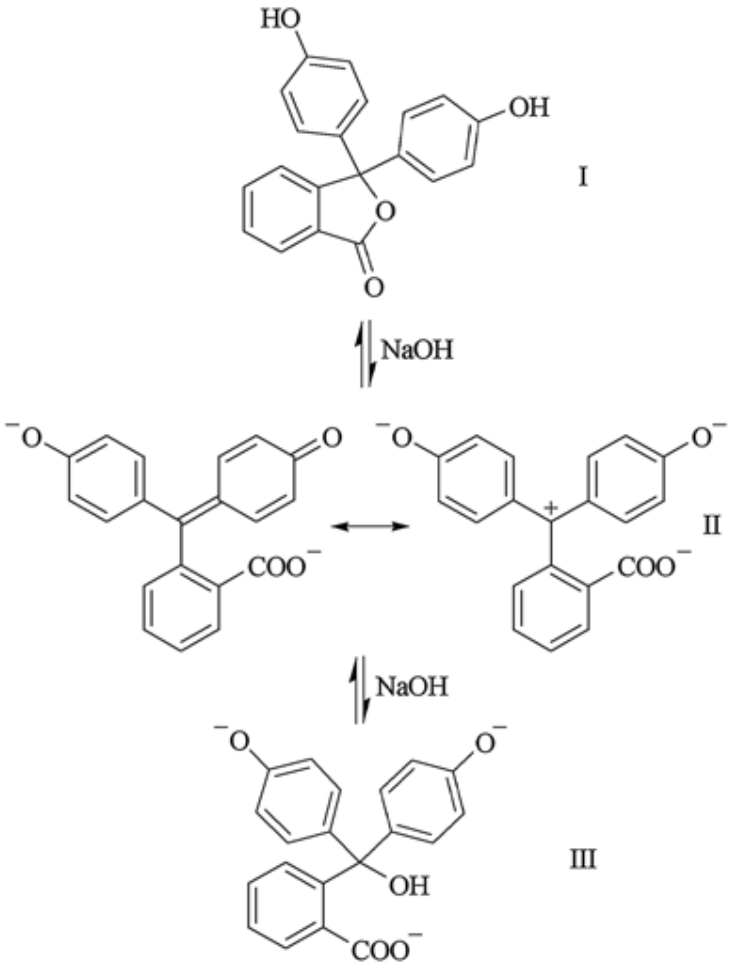

I Phenolphthalein (colorless lactone, acidic)
II Resonance-stabilized anion (red, alkaline)
III Carbinol base (colorless, strongly alkaline)

In acidic media, phthalein derivatives are colorless because the central tetrahedral carbon atom is unable to participate in any resonance structure. In alkaline media, a planar, resonance-stabilized anion is formed that corresponds to a triphenylmethane dye. At extremely high pH, nucleophilic substitution of a hydroxyl group on the central carbon atom occurs; this eliminates the possibility of resonance stabilization and leads to formation of the colorless triphenylmethanol base.

The transition range of phthaleins can also be influenced by the number and properties (electronic and steric) of one or both phenyl residues. Phenolphthalein and thymolphthalein can be used to titrate weak acids.

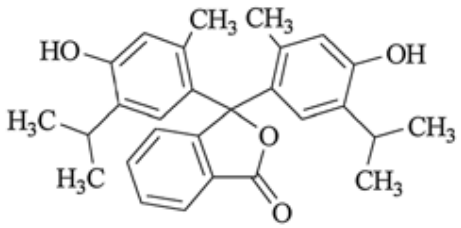

Thymolphthalein

Sulfonphthaleins. Sulfonphthaleins are closely related to phthaleins: the carboxylic acid function is merely replaced by a sulfonic acid group. The various pH-dependent structures are illustrated here with *m*-Cresol Purple, which has two transition ranges:

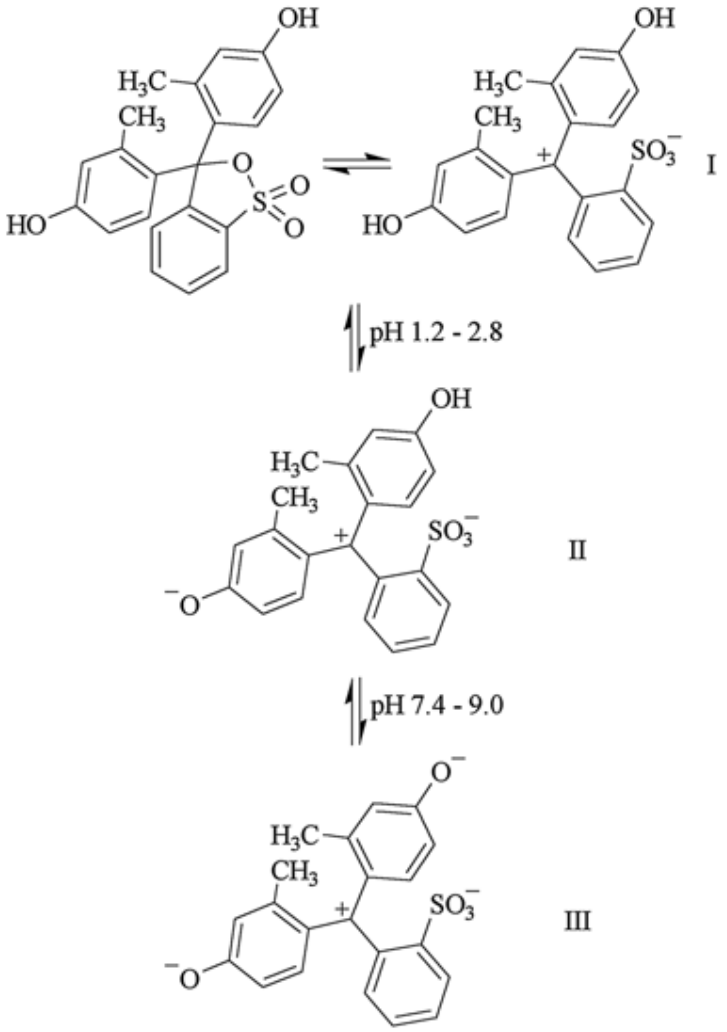

I Red (strongly acidic)
II Yellow (two mesomeric forms, acidic)
III Purple (three mesomeric forms, alkaline)

The alkaline form of sulfonphthaleins is more stable than that of phthaleins; i.e., formation of a triphenylmethanol base and, hence, loss of color under strongly alkaline conditions does not take place. The sulfonphthaleins have a sharp color transition in both acidic and alkaline media. Table 2 lists other commonly used sulfonphthalein pH indicators.

Table 2. Important pH indicators from the sulfonphthalein dye series

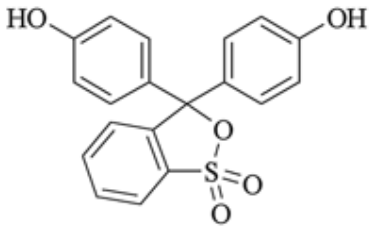

Phenol Red

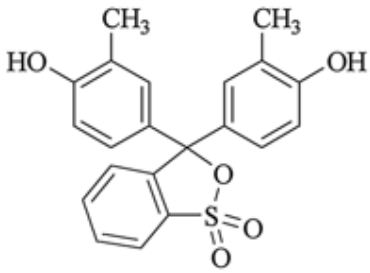

Cresol Red

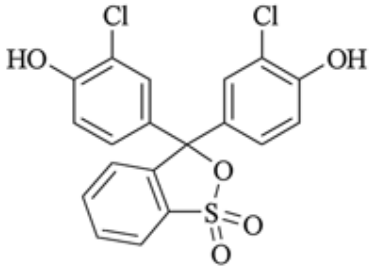

Chlorophenol Red

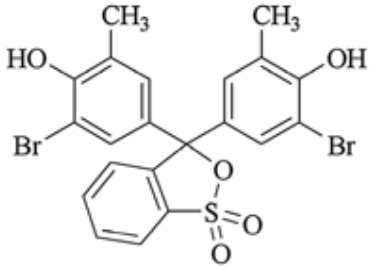

Bromocresol Purple

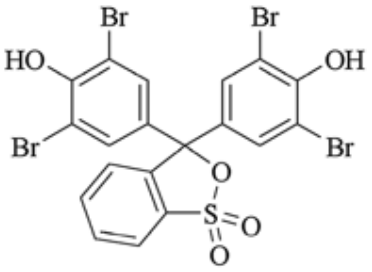

Bromophenol Blue

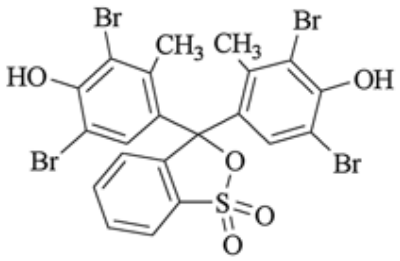

Bromocresol Green

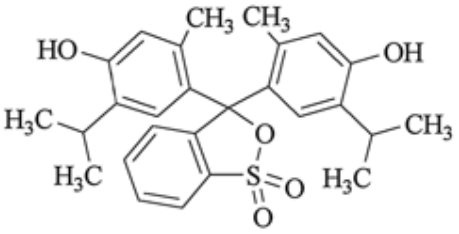

Thymol Blue

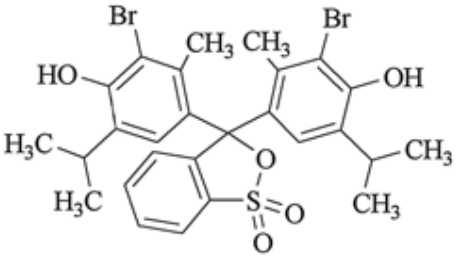

Bromothymol Blue

Other pH Indicators. If the hydroxyl groups in the 4′- and 4′′-positions of Phenol Red (see Table 2) are replaced by amino groups, *anilinesulfonphthalein* [4538-11-8] is obtained:

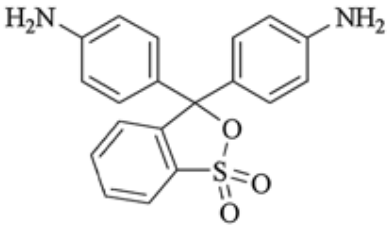

Anilinesulfonphthalein shows two transition regions: yellow – violet at pH 1.3 – 1.9 and violet – yellow at pH 11.7 – 12.5. Other anilinesulfonphthaleins are derived from this compound.

Benzaurin compounds are very similar to phthaleins in their basic structure and their tendency to be stabilized via intermediate structures. Because these indicators lack hydrophilic groups, they are not water soluble and hence are used mostly in nonaqueous systems (see Section pH Indicators for Nonaqueous Solutions).

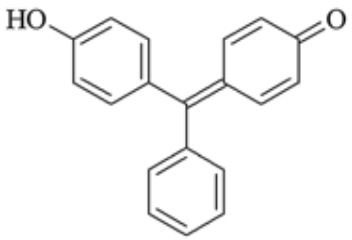

Benzaurin [569-60-8]

Triphenylmethane dyes are structurally related to benzaurins. Malachite Green and Crystal Violet, the most important of these dyes, are obtained by the introduction of dimethylamino groups in the 4′- and 4′′-positions of the benzaurin structure.

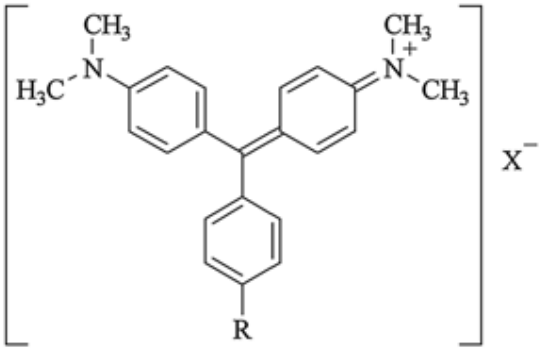

R = H: Malachite Green
R = $N(CH_3)_2$: Crystal Violet

Neutral Red is an *azine dye* (⟶ Azine Dyes – Diaminophenazine Dyes).

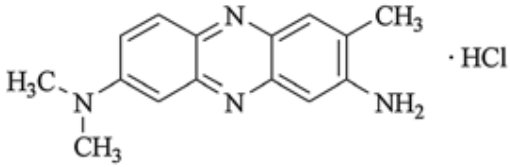

Neutral Red

Triphenylmethane and azine dyes are also used as redox indicators (see Chap. Redox Indicators).

Mixed Indicators. An important criterion for the selection of an indicator is that the transition interval be clear, sharp, and visually detectable. Ideally, the acidic and alkaline forms should possess complementary colors, e.g., red and green. Unfortunately, no indicator fulfills this requirement. Furthermore, a mixture of colors exists in the transition range during which both indicator structures occur simultaneously.

This situation can be improved in two ways. First, a pH-neutral dye can be added to the system to provide a *screened indicator* (Table 3). Second, several indicators with similar transition ranges can be combined to form a *mixed indicator* (Table 4).

Table 3. Screened pH indicators

Components of screened indicator mixture	Parts by volume	Transition point, pH	Color change: Acid	Color change: Alkali
Dimethylaminoazobenzene (0.1 % in ethanol)	1	3.3	blue violet	green
Methylene Blue (0.1 % in ethanol)	1			
Methyl Orange (0.1 % in water)	1	4.1	violet	green
Indigo Carmine (0.25 % in water)	1			
Methyl Orange (0.2 % in water)	1	4.5 – 4.8	red	green fluorescence
Fluorescein (0.2 % in ethanol)	1			
Methyl Red (0.2 % in ethanol)	1	5.4	red violet	green
Methylene Blue (0.1 % in ethanol)	1			
Neutral Red (0.1 % in ethanol)	1	7.0	violet blue	green
Methylene Blue (0.1 % in ethanol)	1			
Phenol Red (0.1 % in ethanol)	1	7.3	green	violet
Methylene Blue (0.02 % in water)	1			
Phenolphthalein (0.1 % in ethanol)	1	8.9	green	violet

Methyl Green (0.1 % in ethanol)	2			
Phenolphthalein (0.1 % in ethanol)	1	10.0	blue	red
Nile Blue (0.2 % in ethanol)	2			
Nile Blue (0.2 % in water)	2	10.8	green	red brown
Alizarin Yellow (0.1 % in ethanol)	1			

Table 4. Mixed pH indicators

Components of mixed indicator	Parts by volume	Transition point, pH	Color change Acid	Color change Alkali
Bromocresol Green (0.1 % in ethanol)	1	4.3	orange	green
Methyl Orange (0.02 % in water)	1			
Bromocresol Green (0.1 % in ethanol)	3	5.1	wine red	green
Methyl Red (0.2 % in ethanol)	1			
Bromocresol Green-Na (0.1 % in water)	1	6.1	yellow green	blue violet
Chlorphenol Red-Na (0.1 % in water)	1			
Bromocresol Purple-Na (0.1 % in water)	1	6.7	yellow	violet blue
Bromothymol Blue-Na (0.1 % in water)	1			
Bromothymol Blue-Na (0.1 % in water)	1	7.5	yellow	violet
Phenol Red-Na (0.1 % in water)	1			
Cresol Red-Na (0.1 % in water)	1	8.3	yellow	violet
Thymol Blue-Na (0.1 % in water)	3			
Thymol Blue (0.1 % in 50 % ethanol)	1	9.0	yellow	violet
Phenolphthalein (0.1 % in 50 % ethanol)	3			
Phenolphthalein (0.1 % in 50 % ethanol)	2	9.6	slightly pink	violet
1-Naphtholphthalein (0.1 % in 50 % ethanol)	1			
Phenolphthalein (0.1 % in ethanol)	1	9.9	colorless	violet
Thymolphthalein (0.1 % in ethanol)	1			
Thymolphthalein (0.1 % in ethanol)	2	10.2	yellow	violet
Alizarin Yellow (0.1 % in ethanol)	1			

For example, detection of the red – yellow color transition of Methyl Red, Methyl Orange, or 4-dimethylaminoazobenzene is facilitated by adding Methylene Blue, a pH-independent blue dye. Under acidic conditions, a violet color is formed; on addition of alkali, a green mixed color is produced. If these azo dye indicators are mixed instead with sulfonphthaleins with similar transition ranges (e.g., addition of Bromocresol Green to Methyl Red), easily recognizable, sharp color changes result (yellow – red to blue – green).

Universal Indicators. If several indicators of differing transition regions are combined, mixed indicators are obtained that can be used over an extremely wide pH range. These “universal indicators” are typically employed in the form of test papers or test strips for rapid pH determinations (see Section pH Indicator Papers). Defined values of pH are indicated within a large pH range (usually in about five stages) by a variety of colors. The pH can be quickly determined to within 0.5 pH unit by comparing the color of the indicator in solution with a reference color scale. Table 5 gives examples of indicators with extended application ranges.

Table 5. Some pH indicators with extended measuring ranges

Name	Measuring range, pH	Color change	Number of colors on evaluation scale	Accuracy
Indicator fluid, pH 0 – 5, Merck	0 – 5	orange red, ocher, lime green, light green, dark green, brown green, blue	11	± 0.25 pH units
Universal indicator fluid, pH 4 – 10, Merck	4 – 10	pink, yellow, yellow green, gray green, gray, bluish gray, gray violet, violet	13	± 0.25 pH units
Indicator fluid, pH 9 – 13, Merck	9 – 13	yellow green, lime green, gray green, gray violet, violet	5	± 0.5 pH units

2.4. pH Indicator Papers

"Bleeding" Indicator Papers. Absorbent, uncoated, unfilled paper is impregnated with solutions of acid – base indicators (see Section Dye Classes and Individual pH Indicators) and dried in an inert atmosphere under constant conditions. One or more indicator dyes can be applied, depending on the intended end use.

Three types of pH indicator papers are produced, which differ with regard to measuring range and accuracy. *Simple pH indicator papers* (e.g., congo, litmus, or phenolphthalein papers) can be used to determine only whether a solution is acidic, neutral, or alkaline. *Universal indicator papers* permit measurements over the total pH range 0 – 14 with an accuracy of ca. 1 pH unit. *Special indicator papers* cover ranges of 2 – 5 pH units; under ideal conditions, the pH can be determined with an accuracy of 0.5 or even 0.2 – 0.3 units. In both universal and special indicator papers, the color obtained during contact with the test solution is compared with a reference color scale.

The pH indicator papers produced by impregnation with dye solutions are prone to "bleeding"; i.e., a portion of the indicator dye dissolves in the test solution during immersion of the paper or strip. This occurs particularly in an alkaline medium (because the solubility of the dye reaches a maximum) and in weakly buffered solutions (because indicator papers must remain in the solution longer). Dissolution of the indicator dye leads to areas of varying dye concentration and color on the test strip, which makes analysis difficult and inaccurate. In addition, the test solution becomes contaminated. To avoid these drawbacks, manufacturing techniques and the composition of the test paper have been modified. However, progress has been limited, and fundamental difficulties associated with the impregnation technique have not been eliminated.

"Nonbleeding" Indicator Papers. Nonbleeding papers can be prepared by using direct [15] or reactive [16] dyes. In the latter case, pure linters (raw material for paper) are suspended in water, and an aqueous solution of a reactive dye is then added. An example of such a dye is

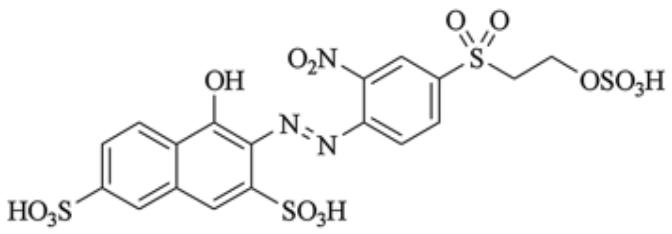

Yellow – violet, pH 5.5 – 8.0

which reacts via the sulfonic acid group of the side chain. The mixture is finally made alkaline, and the dye reacts with the linters. After completion of the reaction, the fiber pulp is centrifuged, washed free of electrolytes, and processed on a paper machine to form pH paper. The pH papers produced in this way are mostly bonded onto a plastic base material and used as pH test strips.

Those pH papers with covalently bonded indicator dyes have the advantage of being able to remain indefinitely in the test liquid, without the indicator dyes mixing with each other or dissolving. This improves measuring accuracy significantly, particularly in alkaline solution, weakly buffered media, and hot solutions (up to ca. 70 °C). Furthermore, pH determination can be carried out on turbid solutions or suspensions, because the strip can be rinsed briefly with distilled water without changing its color. Another advantage is that covalently bonded indicator dyes are not prone to protein errors (see Section Errors); hence, pH measurements can also be performed on biological specimens.

The introduction of mobile small reflectometers (cf. Chap. Analytical Test Kits) which allow the quantitative reading of test strips also led to pH strips which can be numerically evaluated. It is to be mentioned that these products (commercial name Reflectoquant) enable to perform pH measurements even in very complex and difficult matrices like cooling lubricants.

Recently, a new type of indicators was coupled to paper by the above-mentioned direct dye method. The resulting indicator strips allow quick visual pH determination of concentrated acids [17] and bases [18], respectively. These products

(commercial names Aci-Test and Alka-Test) allow to substitute acid – base titration in process analysis [19].

Manufacturers of pH papers and test strips are:

Hydrion (United States)
Kloz (FRG)
Kyoritsu Chemical (Japan)
Macherey-Nagel (FRG)
Merck (FRG)
Riedel-de Haën (FRG)
Toyo Roshi (Japan)
Whatman (United Kingdom)

2.5. pH Indicators for Nonaqueous Solutions

The most widely used indicator dyes for nonaqueous solutions are listed in Table 6. Transition regions are not specified because of inadequate correlation between the pH range in nonaqueous solutions and that in the aqueous phase. The indicators are employed almost exclusively for titrations. Before routine use of a pH indicator in nonaqueous solutions, the coincidence of the transition and equivalence points under the chosen conditions should be checked by potentiometric titration. Detailed information is available in [20-22].

Table 6. Commonly used pH indicators for titrations in nonaqueous solvents

Substance to be titrated	Indicator	CAS registry no.	Color change
Weak bases			
In glacial acetic acid with perchloric acid	Crystal Violet	[548-62-9]	violet – blue green – green – yellow
	Methyl Violet	[8004-87-3]	violet – blue – blue green – yellow
	Malachite Green	[2437-29-8]	brown green – green – yellow
	Nile Blue A	[3625-57-8]	blue – colorless
In acetic anhydride with perchloric acid	Crystal Violet	[548-62-9]	blue – green yellow – yellow
	α-naphtholbenzein	[6948-88-5]	yellow – green
Weak acids			
In dimethylformamide			
With tetraalkylammoniumhydroxide	Thymol Blue	[76-61-9]	red – yellow – blue
With *n*-butylamine	Azo Violet	[74-39-5]	red – blue
With pyridine	2-nitroaniline	[88-74-4]	yellow – red

[Top of Page]

3. Adsorption Indicators

Adsorption indicators provide information about a specific chemical condition by means of color or fluorescence reactions that take place on the surface of a precipitate. In titrimetry, they are employed to indicate the end points of precipitation reactions. The distinct change in color of the precipitate is generally easier to perceive than corresponding changes in the color of the solution. This is because the first ions that bind to the precipitate are those present in excess in the solution to be titrated. At the neutralization point, this excess is replaced by an excess of ions of opposite charge, and the charge on the surface of the precipitate is reversed. This reversal is made visible by the simultaneous binding of anionic or cationic dyes to the surface.

The best known titration – precipitation reactions include the argentometric determination of halogens and pseudohalogens with such anionic dyes as fluorescein, eosin, phloxine, and Rose Bengal. A comparable but reversed titration reaction employs the cationic dye rhodamine 6G for the determination of silver.

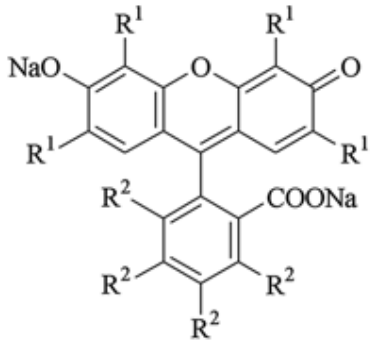

Fluorescein: $R^1 = H, R^2 = H$

Eosin: $R^1 = Br, R^2 = H$

Phloxine: $R^1 = Br, R^2 = Cl$

Rose Bengal: $R^1 = I, R^2 = Cl$

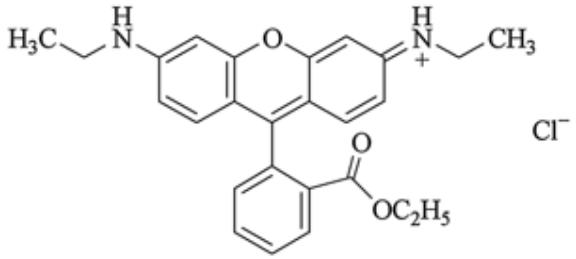

Rhodamin 6G

The uses and color transitions of the adsorption indicators are given in Table 7. Further indicators, along with more detailed information, can be found in [5, vol. 2, pp. 267 – 315], [23-26].

Table 7. Adsorption indicators

Name	CAS registry no.	Use	Color change (solution ⟶ precipitate)
Fluorescein (C.I. 45 350)	[2321-07-5]	titration of Cl^-, Br^-, I^-, and SCN^- with Ag^+	pale green ⟶ orange
Eosin (C.I. 45 380)	[17372-87-1]	titration of Br^- and I^- with Ag^+	orange ⟶ red violet
Phloxine (C.I. 45 405)	[6441-77-6]	titration of Br^- and I^- with Ag^+	orange ⟶ red violet
Rose Bengal (C.I. 45 440)	[632-68-8]	titration of I^- with Ag^+	red ⟶ blue red
Rhodamine 6G (C.I. 45 160)	[989-38-8]	titration of Ag^+ with Br^-	orange ⟶ red violet

[Top of Page]

4. Fluorescent Indicators

Fluorescent indicators are organic compounds whose changes in fluorescence allow determination of chemical parameters. Fluorescence may be influenced by factors such as pH, the partial pressure of oxygen, or the presence of metal or halide ions. The intensity of fluorescence can be increased by addition of surfactants [27]. Examples of fluorescent indicators are listed in Table 8. Applications of fluorescent indicators include the *titration* of solutions in which conventional indicators yield an indistinct color change at the end point (e.g., pH titration of beverages, synthetic resins, oils, or detergents). Titration is normally carried out in diffuse (ideally red) light or in a darkened room, and the solution is irradiated with UV light (wavelength 300 – 400 nm).

Table 8. Fluorescent indicators

Name	CAS registry no.	Structural formula	Comments
Fluorescent pH indicators			

Erythrosin B (C.I. 45 430)	[16423-68-0]	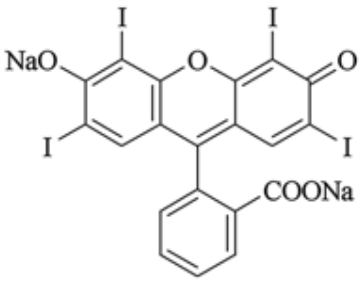	transition range: pH 2.5 (no fluorescence) to pH 4.0 (pale green fluorescence)
2-Naphthionic acid	[81-16-3]	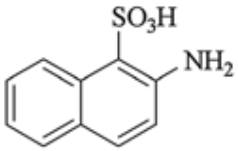	transition range: pH 12.0 (blue fluorescence) to pH 13.0 (violet fluorescence)
Umbelliferone	[93-35-6]	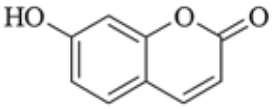	transition range: pH 6.5 (no fluorescence) to pH 8.5 (blue fluorescence)
Fluorescent metal indicators			
Morin (C.I. 75 660)	[480-16-0]	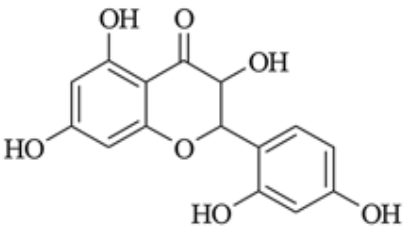	for titration of gallium (pH 3.8); transition: green fluorescence to no fluorescence for titration of indium (pH 5.0); transition: green fluorescence to no fluorescence
3-Hydroxy-2-naphthoic acid	[92-70-6]	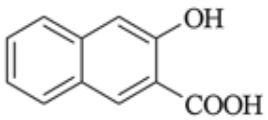	for titration of aluminum (pH 3.0); transition: blue fluorescence to green fluorescence

Fluorescent indicators are used in the construction of fluorescent *optical sensors* (optodes or optrodes). The indicator is fixed to a substrate such as glass, an ion-exchange membrane, or cellulose. Alternatively, an indicator solution may be separated from the test solution by a membrane that is permeable to the substance to be analyzed. In both cases, transmission of light occurs via optical fibers [28]. Such sensors have been used to determine oxygen content [29] or pH [29], [30]. Combination with optical fibers, however, has also increased interest in the use of fluorescent indicators for the titration of turbid or colored solutions. This applies to titration of pH [31], metals [32], and precipitates [33]. New fluorescent pH indicators with a wide transition range suitable for rapid pH determination are described in [34].

Special fluorescent indicators have been employed for the *intracellular determination* of pH [35], [36] and calcium [37], [38]. Indicators are introduced into the cell in the form of membrane-permeable hydrophobic esters. Inside the cell, the esters are enzymatically cleaved to form the hydrophilic fluorescent species. The use of fluorescent metal indicators for fluorometric *elemental analysis* is described in [39].

[Top of Page]

5. Chemiluminescent Indicators

Chemiluminescent indicators are organic substances whose changes in luminescence enable quantitative analysis of various compounds. The energy for chemiluminescence is produced by an oxidation – reduction process, particularly by the formation of transannular peroxides. Very low-intensity light is emitted without the evolution of thermal energy. This light is usually detected by using equipment operated with the exclusion of external light. However, titration end points determined with the assistance of chemiluminescent indicators can also be detected visually. Both increases and decreases in the intensity of luminescence have been used for analytical purposes [40-42].

Chemiluminescent indicators have the advantage that determination can be carried out in highly colored samples. A disadvantage is that metal contaminants can lead to nonluminescent reactions and, therefore, interfere with the determination. Chemiluminescent indicators can also be used as redox indicators and, because the redox potential depends on pH, pH determinations can be made.

An example of this application is the determination of the acid number in highly colored fats and oils by using *lucigenin* [2315-97-1], which is oxidized in alkaline media:

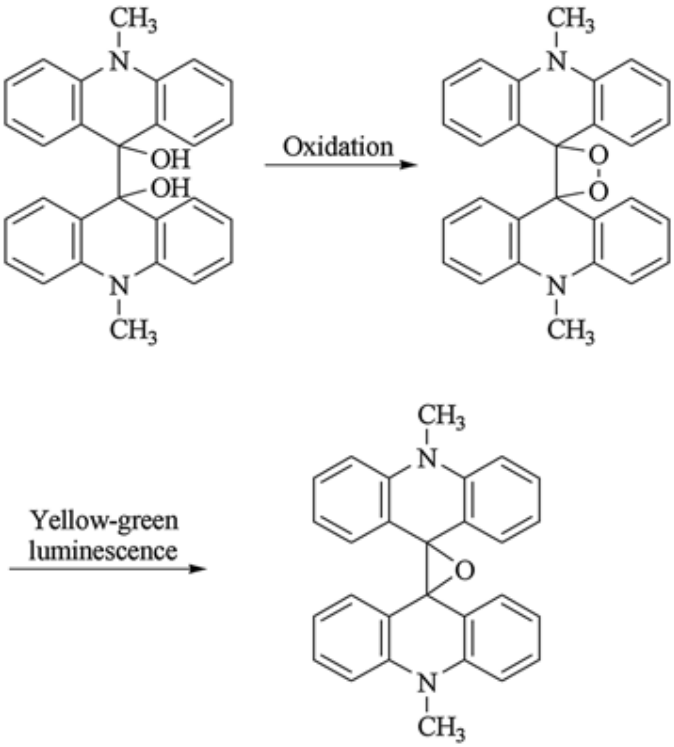

Other chemiluminescent indicators include *luminol* [521-31-3] and *lophine* [484-47-9], both of which are similarly oxidized in alkaline media; their peroxides have blue and pale yellow luminescence, respectively.

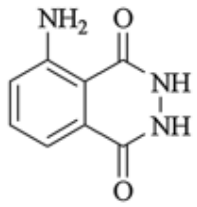

Luminol

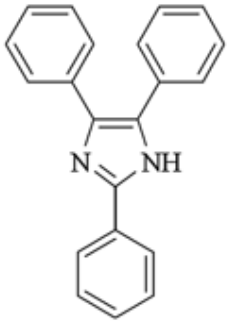

Lophine

Other indicators, along with a review of the applications of chemiluminescence in analytical chemistry and biochemistry, are discussed in [43-45] (see also ⟶ Luminescent Materials).

[Top of Page]

6. Redox Indicators

Redox indicators are organic compounds whose oxidized and reduced forms display different colors. The two levels of oxidation form a redox system that can be described by its standard potential or rH value.

6.1. Standard Potential and rH Value

The standard potential U^0 of a redox indicator can be derived from the Nernst equation:

$$U = U^0 + \frac{RT}{zF} \ln \frac{c_{ox}}{c_{red}}$$

where U is the potential of the redox system; R is the gas constant; T, the absolute temperature; z, the number of electrons involved in the reaction; F, the Faraday constant; and c_{ox} and c_{red} are the molar concentrations of the oxidized and the reduced species, respectively.

The standard potential U^0 results when the activities of the oxidized and reduced species are both assumed to be 1. For the electrochemical determination of U^0, a platinum electrode is immersed in the test solution and the potential is measured relative to a standard hydrogen electrode.

Generally, the redox potential depends strongly on pH because hydrogen ions are directly involved in the redox process and, therefore, must also be considered in the Nernst equation. The redox system of phenolindophenolate for example

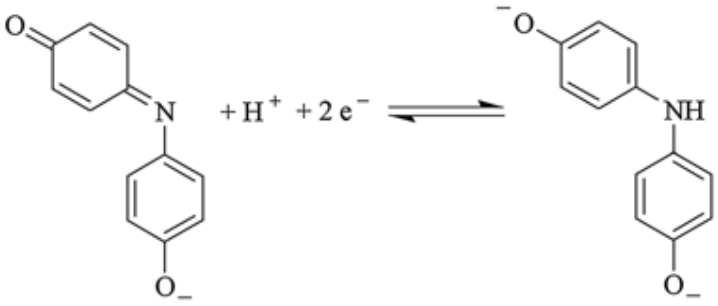

can thus be described by

$$U = U^0 + \frac{RT}{zF} \ln \frac{c_{ox} - c_{H^+}}{c_{red^{2-}}}$$

Another way of describing the oxidizing and reducing properties of a redox system is to use the rH value, which was introduced as an analogue of pH. A platinum electrode immersed in a solution of a redox system is regarded as a hydrogen electrode. The rH value is defined as the negative logarithm of the hydrogen pressure of this electrode. Measurement is once again carried out relative to a standard hydrogen electrode. The redox potential is given by

$$U = 0.058 \log a_{H^+} - 0.029 \log p$$

where a_H^+ is the activity of the hydrogen ion and p the pressure of hydrogen. By incorporating pH and rH, the following relationship is obtained:

$$U = -0.58\text{pH} + 0.029\text{rH}$$

From this equation, redox potential and rH value can be interconverted, provided the pH is known. Increasing rH values indicate increasing oxidation.

To compare redox indicators, their respective redox potentials must be measured under standard conditions, i.e., in neutral solution and with the redox system present in a semioxidized state.

The color of redox indicators does not change abruptly at a defined potential but rather over a certain range. The midpoint of this range should be regarded as the rH value. Frequently, however, the rH transition range is also specified.

Practical Aspects. Redox indicators are usually added to the test solution. A minimal amount of indicator should be used because the indicator also represents a redox system, which could influence the system under investigation.

The redox potential of the indicator must be greater than that of the titrant for an oxidizing titrant (i.e., when an oxidizing solution is being added) and lower than that of the titrant for a reducing titrant. Furthermore, redox reactions often proceed slowly and, therefore, require some time to reach equilibrium.

6.2. Classes of Redox Indicators

Table 9 lists a selection of the most commonly employed redox indicators. In this section, the most important classes of redox indicators are discussed. Redox indicators are dealt with comprehensively in [1, pp. 469 – 678]. Detailed examples of the use of redox indicators are described in [11], [12], [46], [47].

Table 9. Redox indicators

Indicator	CAS registry no.	Standard potential U^0, V	Approximate rH	Color (reduced – oxidized)
Neutral Red (C.I. 50 040)	[553-24-2]	– 0.32 [a]	2 – 4.5	colorless – red
Safranine T or O (C.I. 50 240)	[477-73-6]	– 0.29 [a]	4 – 7.5	colorless – red
Indigo Carmine (C.I. 73 015)	[860-22-0]	– 0.11 [a]	8.5 – 10.5	yellow tinge – blue
Methylene Blue (C.I. 52 015)	[61-73-4]	+ 0.01 [a]	13.5 – 15.5	colorless – blue
Thionin (C.I. 52 000)	[581-64-6]	+ 0.06 [a]	15 – 17	colorless – violet
Thymolindophenol	[2667-28-9]	+ 0.18 [a]	17.5 – 20	colorless – blue (red at pH 9)

2,6-Dichlorophenol indophenol	[620-45-1]	+ 0.23 [a]	20 – 22.5	colorless – blue
Gallocyanine (C.I. 51 030)	[1562-85-2]	+ 0.02 [b]		colorless – violet blue
Safranine T or O (C.I. 50 240)	[477-73-6]	+ 0.24 [b]		colorless – blue violet
Indigo Carmine (C.I. 73 015)	[860-22-0]	+ 0.29 [b]		colorless – blue
Nile Blue (C.I. 51 180)	[2381-85-3]	+ 0.41 [b]		colorless – blue red
Methylene Blue (C.I. 52 015)	[61-73-4]	+ 0.53 [b]		colorless – blue
Thionin (C.I. 52 000)	[581-64-6]	+ 0.56 [b]		colorless – violet
2,6-Dichlorophenol indophenol	[620-45-1]	+ 0.67 [b]		colorless – red
Variamine Blue (C.I. 37 255)	[3566-44-7]	+ 0.71 [c]		colorless – blue
Diphenylamine	[122-39-4]	+ *0.76* [b]		colorless – violet
Diphenylamine-4-sulfonic acid, barium salt	[6211-24-1]	+ 0.85 [d]		colorless – red violet
Tris(2,2-dipyridyl)iron(II) sulfate		+ 1.03 [d]		red – blue
		+ 1.02 [e]		
N-Phenylanthranilic acid	[91-40-7]	+ 0.89 [d]		colorless – purple red
Ferroin	[14708-99-7]	+ 1.06 [d]		orange red – blue
		+ 1.14 [e]		
Nitroferroin	[4199-88-6]	+ 1.25 [d]		pink – colorless
		+ 1.26 [e]		

[a] pH 7.
[b] In strongly acidic solution.
[c] Standard potential is strongly pH dependent.
[d] In 1 mol/L H_2SO_4 .
[e] In 0.5 mol/L H_2SO_4 .

Indophenols. The reaction equation describing the redox process with phenolindophenol [500-85-6] has already been given in Section Standard Potential and rH Value. Thymolindophenol and 2,6-dichlorophenolindophenol are two redox indicators with a similar basic structure. Applications include the titration of copper(I) ions with iron(III) sulfate [48] as well as the determination of free halogen, hypohalogenites, and halogenates with ascorbic acid [49].

Azines, Oxazines, Thiazines. Azines used as redox indicators include safranine T and Neutral Red. Applications include the titration of iron(III) salts with titanium(II) chloride [50] or chromium(II) salts [51]. Oxazine redox indicators include Nile Blue and gallocyanine. The best known thiazines are Methylene Blue and thionin. Applications include the titration of iron(III) salts with vanadium(II) compounds [52]. Many of these dyes are described in detail elsewhere (⟶ Azine Dyes).

Indigo Derivatives (⟶). An example of an indigo dye is the frequently employed Indigo Carmine. The redox process can be expressed as follows:

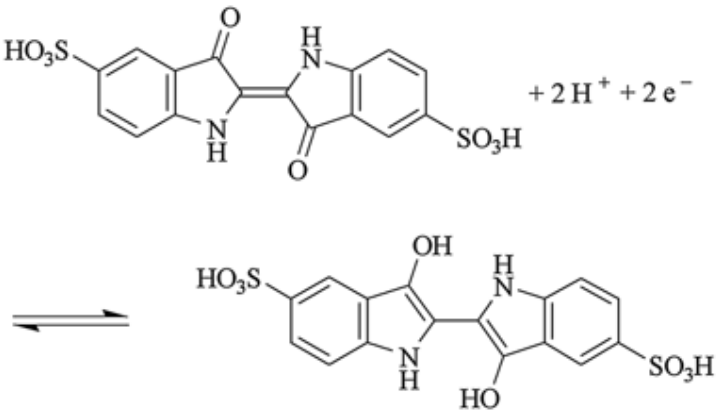

Applications include the determination of rhenium with tin(II) solutions [53].

Diphenyl and Benzidine Derivatives. Diphenylamine-4-sulfonic acid and Variamine Blue are derivatives of diphenylamine. The redox process for Variamine Blue can be described as follows:

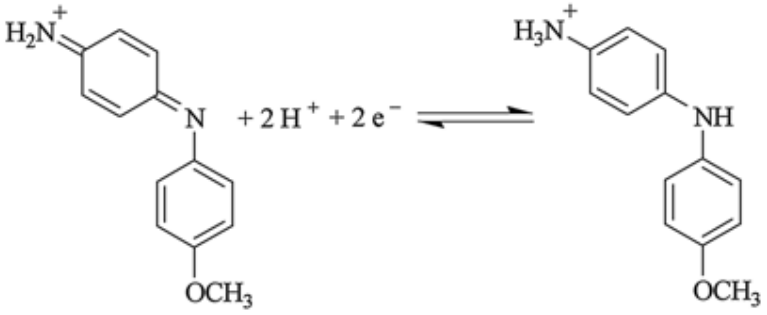

A multitude of applications have been discussed in [46].

The benzidine derivatives formerly employed as redox indicators include *o*-tolidine, *o*-dianisidine, and 3,3-dimethylnaphthidine. However, because benzidine and most of its derivatives are carcinogenic, they are not included in Table 9.

1,10-Phenanthroline Complexes. The best known 1,10-phenanthroline complex is ferroin, an iron(II) complex with three 1,10-phenanthroline molecules (see Section Ligands Forming Five-Membered Chelate Rings). Utilization of these compounds as redox indicators with particularly brilliant color transitions is discussed comprehensively in [46], [54]. An example is the titration of chromium(VI) with iron(II) in determination of the chemical oxygen demand (COD) according to DIN 38 409, Parts 41 and 43, ISO 6060–1986, or Norme Française NF T 90–101.

[Top of Page]

7. Metal Indicators

Metal indicators are organic compounds that form specifically colored metal ion complexes in aqueous media. These dye reactions can be employed for detection of the equivalence point in volumetry (complexometry), as well as for concentration determinations in colorimetry or photometry. The demands placed on metal indicators differ slightly depending upon the area of application.

The most important metal indicators are listed in Table 10. Detailed information is available on their applications in volumetry (complexometry) [55-59]; colorimetry and photometry [60-65]; and masking [66], [67].

Table 10. Metal indicators

Indicator	CAS registry no.	For complexometric determination of	For photometric determination of
Alizarin Complexone	[3952-78-1]	Ba, Ca, Cd, Co, Hg, In, Pb, Sr, Zn	fluoride
Alizarin S	[130-22-3]	Al, lanthanides, Mo, Sc, Th, Y, Zr	Al, lanthanides, Sc, Th, Y, Zr
Arsenazo III	[62337-00-2]	lanthanides, Th, Y	lanthanides, Sc, Th, U, Y, Zr
Aurintricarboxylic acid (Aluminon, C.I. 43 810)	[4431-00-9]	Al, Ca, Fe, Mg	Al, Be, Fe, Ga, lanthanides, Pd, Y
2,2′-Bipyridine	[366-18-7]		Ag, Cd, Cu, Fe, Mo
Bromopyrogallol Red	[16574-43-9]	Bi, Cd, Co, lanthanides, Mg, Mn, Ni, Pb	Ag, Bi, Co, Cu, lanthanides, Pb, Sb, Sn, Th, Ti, U, W, Y, Zr
Calcon (Eriochrome Blue Black R)	[2538-85-4]	Ca, Cd, Mg, Mn, Zn	
Calconcarboxylic acid	[3737-95-9]	Ca	
Chrome Azurol S (C.I. 43 825)	[1667-99-8]	Al, Ca, Cu, Fe, lanthanides, Mg, Ni, Th, U, V, Y, Zr	Al, Be, Bi, Cr, Cu, Fe, Ga, In, lanthanides, Mg, Mn, Ni, Pb, Pd, Sc, Sn, Th, Y, Zr
Chromotropic acid, disodium salt	[5808-22-0]		Ti
Cuprizone	[370-81-0]		Cu
5-(4-Dimethylaminobenzylidene) rhodanine	[536-17-4]		Ag, Au, Cu, Hg, Pd

Dimethylglyoxime	[95-45-4]		Co, Fe, Ni, Pd
1,5-Diphenylcarbazide	[140-22-7]	Hg, Ni, Pb, V	Cr, Cu, Hg
Dithizone	[60-10-6]	Ag, Au, Bi, Cd, Co, Cu, Fe, Hg, In, Mn, Ni, Pb, Sn, Zn	Ag, Au, Bi, Cd, Co, Cu, Hg, Ni, Pb, Pd, Zn
Eriochrome Black T	[1787-61-7]	Ba, Ca, Cd, Hg, In, lanthanides, Mg, Mn, Pb, Zn, Zr	
Eriochrome Blue SE	[1058-92-0]	Ca, Cd, Mg, Mn, Ni, Pb, Zn	
Eriochrome Blue Black B	[3564-14-5]	Ca, Cd, Mg, Mn, U, Zn, Zr	
Eriochrome Cyanine R (C.I. 43 810)	[3564-18-9]	Al, Ca, Cu, Fe, Mg, Th, Zr	Al, Be, Cr, Ga, Mg, Th, Tl, Zr
Fluorescein Complexone (Calcein)	[1461-15-0]	Ba, Ca, Cr, Cu, Mg, Sr	Ca
Glyoxalbis(2-hydroxyanil)	[1149-16-2]		Ca, Sc, U
Hematoxylin	[517-28-2]	Al, Bi, Cu, Th, Zr	Sn
8-Hydroxyquinoline	[148-24-3]		Al, Bi, Ca, Cd, Cr, Cu, Fe, Ga, Ge, In, La, lanthanides, Mg, Nb, Sc, Th, Ti, U, V, Y, Zn, Zr
2-Mercaptobenzothiazole	[149-30-4]		Au, Cd, Cu, Pb, Pd
Methylthymol Blue	[1945-77-3]	Ba, Bi, Ca, Cd, Cu, Fe, Hg, In, lanthanides, Mg, Mn, Pb, Sc, Sn, Th, Tl, Zn, Zr	Al, Be, Bi, Ca, Co, Cr, Cu, Fe, Ga, In, lanthanides, Mg, Mn, Ni, Pb, Pd, Sc, Th, U, V, Y, Zn, Zr
Murexide	[3051-09-0]	Ag, Ca, Co, Cu, Mn, Ni, Sc, Th, Zn	Ca, Cu, Sc
1-Nitroso-2-naphthol	[131-91-9]		Co, Fe, Mo, Pd
2-Nitroso-1-naphthol	[132-53-6]		Co, Pd
Nitroso-R-salt	[525-05-3]	Cu, Ni	Co, Fe, Pd
1,10-Phenanthroline	[66-71-7]		Ag, Cu, Fe, Pd, V, Zn
Phenylfluorone	[975-17-7]		Ge, Mo, Nb, Sb, Sn, Ta, U
Phthalein Purple		Ba, Ca, Cd, Mg, Sr	Ba, Ca, Mg, Mo
1-(2-Pyridylazo)-2-naphthol (PAN)	[85-85-8]	Al, Bi, Cd, Co, Cu, Fe, Ga, Hg, In, Mn, Ni, Pb, Th, Tl, U, V, Zn	Ag, Bi, Cd, Co, Fe, Ga, lanthanides, Pd, Sc, U, V, Y, Zn, Zr
4-(2-Pyridylazo)resorcinol (PAR)	[1141-59-9]	Al, Bi, Cd, Cu, Fe, Ga, Hg, In, lanthanides, Mn, Ni, Pb, Sr, Th, Tl, Zn	Ag, Bi, Cd, Co, Cr, Cu, Fe, Ga, Hg, La, Mn, Nb, Ni, Pb, Pd, Sc, U, V, Y, Zn, Zr
Pyrocatechol Violet	[115-41-3]	Bi, Cd, Co, Cu, Fe, Ga, In, Mg, Mn, Ni, Pb, Th, Zn	Al, Bi, Cr, Cu, Fe, Ga, In, Mg, Mn, Ni, Pb, Sc, Sn, Th, Y, Zn, Zr
Pyrogallol Red	[32638-88-3]	Bi, Co, Ni, Pb	Ag, Mo
Sulfonazo III	[1738-02-9]	Ba,	Ba
5-Sulfosalicylic acid	[5965-83-3]	Fe, Tl	Fe, Ti
4-(2-Thiazolylazo)resorcinol (TAR)	[2246-46-0]	Co, Cu, lanthanides, Ni, Tl	as for PAR
Thorin	[3688-92-4]	Bi, Sc, Th, U, Y	lanthanides
Thymolphthalexon	[1913-93-5]	Ag, Ba, Ca, Cr, Mn, Sr	Ba, Ca, lanthanides, Mg, Mo, Sr
Tiron	[149-45-1]	Fe	Fe, lanthanides, Mo, Ti, U, W
Toluene-3,4-dithiol	[496-74-2]		Ge, Mo, Sn, W
Xylenol Orange	[1611-35-4]	Bi, Ca, Cd, Co, Cu, Fe, Hg, In, lanthanides, Mg, Mn, Pb, Sc, Th, Tl, U, V, Y, Zn, Zr	Ag, Bi, Cd, Cr, Fe, Ga, Hg, In, La, lanthanides, Mn, Mo, Nb, Ni, Pb, Sc, Sn, Th, Ti, Tl, U, V, Y, Zn, Zr
Zincon	[135-52-4]	Cd, Hg, Pb, Zn	Cu, Ga, Ni, Zn

7.1. Volumetry (Complexometry) [55-59]

A small amount of indicator is added to the solution to be titrated. The color of the metal-free indicator must differ from that of

the metal –indicator complex. Another complexing agent, usually an aminopolycarboxylic acid such as ethylenediaminetetraacetic acid (Titriplex) is used as titrating agent, which binds the metal ions as a colorless complex. The equivalence point is indicated by the color change between the metal – indicator complex and the free indicator.

Metal indicators for complexometry must satisfy the following requirements:

1. *Complex Stability.* The metal – indicator complex must be less stable than the metal – aminopolycarboxylic acid complex.
2. *Detection.* The color difference between the metal – indicator complex and the free indicator must be large and clearly visible.
3. *Selectivity.* Ideally, the indicator should be specific for only one element. Although this condition is rarely met, selectivity can often be increased by correct choice of reaction conditions.
4. *Shelf Life and Solubility.* Both indicator and metal – indicator complex should be water soluble. The indicator should have a long shelf life in aqueous or water-miscible mixtures.

7.2. Colorimetry and Photometry [60-65]

In visual colorimetry (color comparison) and photometry (measurement of light absorption), the formation of colored complexes is used to determine metal ion concentration. The intensity of the resulting color is a function of the metal ion concentration. The actual concentration is determined by comparing the color with standards or by measurement of the absorbed light.

Metal indicators for colorimetry or photometry must satisfy the following requirements:

1. *Complex Stability.* Metal – indicator complexes should be stable with respect to time. Contrary to the situation with complexometry, thermodynamic stability is less important. By selecting the conditions (indicator excess), even weak complexes can be used to determine concentration.
2. *Sensitivity.* Sensitivity is governed primarily by the magnitude of the molar extinction coefficient. Metal – indicator complexes typically used for colorimetric and photometric determinations have extinction coefficients (ϵ) in the region of 5×10^3 to 1×10^5 L mol^{-1} cm^{-1}.
3. *Selectivity.* Maximum selectivity is desired to minimize the number of processing steps required to remove interfering contaminant ions. Selective or specific reagents are, however, still very rare. Selectivity can sometimes be enhanced by modifying the indicator molecule; however, selectivity is usually improved by optimizing the reaction conditions and the use of additives (e.g., masking agents) [66], [67].

7.3. Structure

The metal indicator and metal ions form complexes consisting of an electron donor and an electron acceptor. If the donor molecule (ligand) contains two or more atoms with donor properties, the complex is known as a *chelate.* The number of ligand donor atoms bound to the central ion determines the ligand functionality; for example, ligands with a single donor atom are referred to as monodentate; those with two donor atoms, as bidentate.

Electron-donating atoms of the “functional group” of the ligand molecule include oxygen, nitrogen, and sulfur. The fundamental properties of a complexing agent are determined by these functional groups, whereas other molecular substituents influence sensitivity, selectivity, and color contrast. Incorporation of sterically hindered groups in an indicator molecule may increase its selectivity. Increasing the size of the resonance system enhances sensitivity in all cases. The latter is particularly advantageous if the change in resonance shifts the absorption maximum of the complex to longer wavelengths because the human eye is more optically receptive to these wavelengths (important for visual techniques).

Most organic complexing agents can be regarded as bidentate and form five- or six-membered chelate rings. According to the type of ligand, chelating reagents can be classified into ligands with O,O-, O,N-, N,N-, and N,S-donor atoms.

7.4. Ligand Classes

7.4.1. *Ligands Forming Five-Membered Chelate Rings*

O,O Donor Atoms. The O–C=C–O group is often present in chelating agents. The resulting complexes have the following structure:

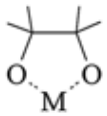

Important examples of this group include 1,2-dihydroxybenzenes and 1,2-dihydroxynaphthalenes, e.g., Tiron, Pyrocatechol Violet, Pyrogallol Red, and Bromopyrogallol Red, as well as the fluorones.

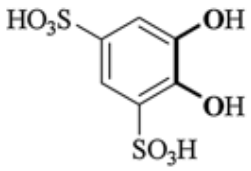

Tiron

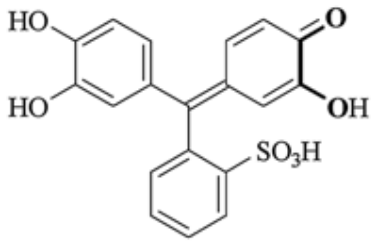

Pyrocatechol Violet

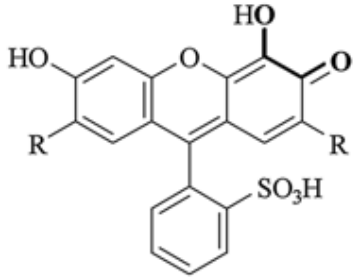

R = H: Pyrogallol Red
R = Br: Bromopyrogallol Red

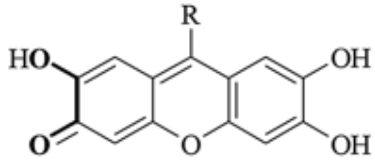

R = CH_3, C_6H_5: Fluorones

A range of triphenylmethane dyes derived from Pyrogallol Red and Pyrocatechol Violet are frequently employed as indicators in complexometry.

O,N Donor Atoms. The functional group in one of the oldest and best known complexing agents, 8-hydroxyquinoline, consists of O–C=C–N. Because this functional group has only slight selectivity, the compound is a group reagent and forms yellow complexes of varying intensity with more than 40 ions.

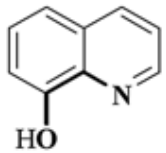

8-Hydroxyquinoline

The O–C–N=N group present in diphenylcarbazone [538-62-5] also forms five-membered chelates, with a structure analogous to the 8-hydroxyquinoline complex. Chelates can be derived from the tautomeric hydroxy form:

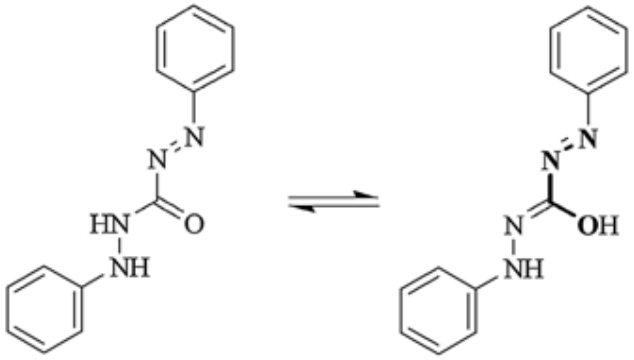

Diphenylcarbazone

Diphenylcarbazone and the related diphenylcarbazide are used in the photometric determination of chromium and mercury.

N,N Donor Atoms. Many of the more important metal indicators contain N,N donor atoms. The most common groups are N=C–C=N and N=C–N=N. The former is present in α-dioximes, which are highly selective reagents for nickel and palladium. The best known compound is the classic nickel reagent diacetyldioxime (dimethylglyoxime).

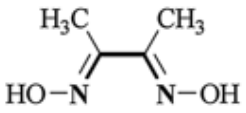

Diacetyldioxime

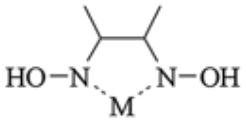

Chelate

If the functional complexing group N=C–C=N is incorporated in a heterocyclic system, then a class of indicators is obtained that is specific for iron and copper, e.g., 2,2′-dipyridyl (**1**) and 1,10-phenanthroline (**2**):

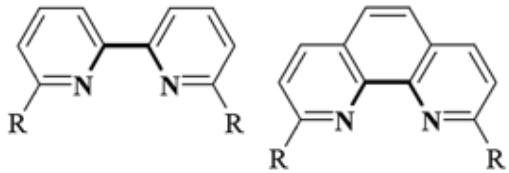

1 Copper indicators: R = CH_3, C_6H_5

2 Iron indicators: R = H

The indicators form intensely colored 1 : 2- and 1 : 3-chelates with iron(II) and copper(I) compounds. In these chelates, however, the central ion can easily be oxidized to iron(III) or copper(II), which give weakly colored complexes. This reaction allows the chelates to be used as redox indicators (see Section Classes of Redox Indicators). Due to steric reasons, only the copper compounds can form the tetrahedral 1 : 2-chelate; this explains their highly specific nature.

The N=C–N=N group occurs in the pyridylazoresorcinols (PAR) and pyridylazonaphthols (PAN).

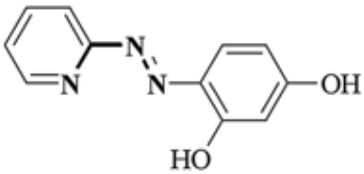

4-(2-pyridylazo)resorcinol (PAR)

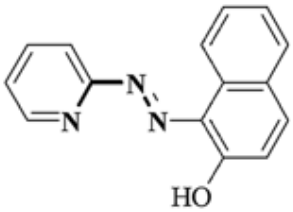

1-(2-pyridylazo)-2naphthol (PAN)

Chelation is thought to occur by formation of an additional five-membered ring with the N,O group:

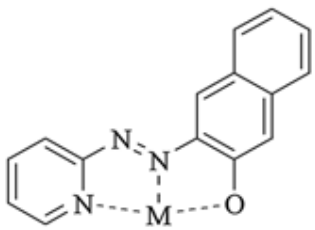

This does not increase selectivity, but significantly improves the stability of the complex.

The pyridylazoresorcinols and pyridylazonaphthols are excellent indicator reagents for heavy-metal ions. They are frequently employed in complexometry and photometry.

N,S Donor Atoms. The N=C–C–S group is present in the thio analogue of 8-hydroxyquinoline, i.e., 8-mercaptoquinoline [491-33-8]. Due to the different reactivity of sulfur, this indicator is more selective than 8-hydroxyquinoline; it reacts only with transition metals of pronounced sulfur affinity, as well as molybdenum, tungsten, and vanadium.

The N=N–C=S group in the sulfur analogue of diphenylcarbazone, diphenylthiocarbazone (dithizone), is of greater importance. Dithizone forms complexes with a number of transition metals and thus is a group reagent:

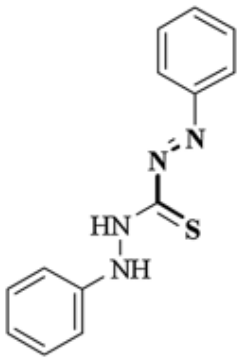

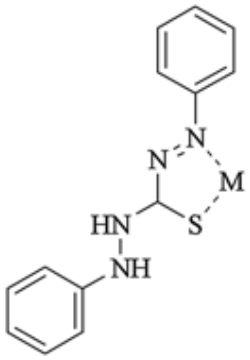

Nevertheless, dithizone can be employed for very selective and also extremely sensitive determinations by maintaining a constant pH, employing masking reagents, and replacement (substitution) reactions [65].

7.4.2. *Ligands Forming Six-Membered Chelate Rings*

Six-membered chelate rings occur with a frequency similar to five-membered rings in important analytical complexes. However, only O and N are donor atoms. The most important reagents are anthraquinones, 4,5-dihydroxynaphthalenes, 2-hydroxybenzoic acids, 2-nitrosonaphthols, and methylaminodiacetic acid derivatives.

Anthraquinones. Anthraquinone indicators include the commonly used indicator, Alizarin S (1,2-dihydroxyanthraquinone-3-sulfonic acid, sodium salt):

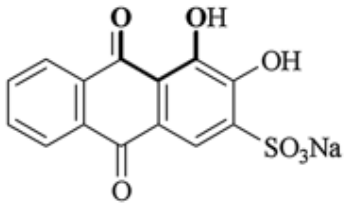

Chelate formation is limited mainly to cations from the second, third, and fourth groups of the periodic table.

4,5-Dihydroxynaphthalenes. Chromotropic acid [148-25-4] (4,5-dihydroxy-2,7-naphthalenedisulfonic acid) represents a typical member of this group:

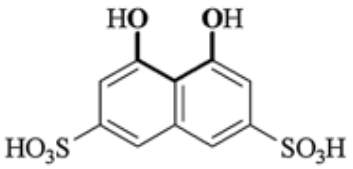

A multitude of compounds containing additional azo functions can be derived from the basic structure. An example is phenylazochromotropic acid [4197-07-3], which is able to function in two ways. Normally, complex formation takes place via the O,O ligand atoms, whereas the azo group forms a second ring as a result of hydrogen bonding:

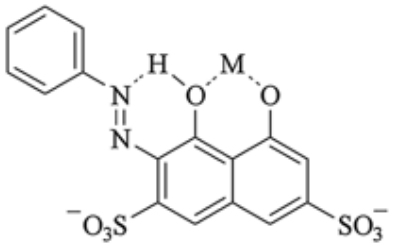

Cations with a pronounced affinity toward nitrogen, however, can also form complexes via the N,O group.

Other important indicators from this group include arsenazo I [520-10-5], arsenazo II [3861-75-4], arsenazo III [1668-00-4], sulfonazo III [1738-02-9], chlorophosphonazo III [1914-99-4], and thorin [3688-92-4].

2-Hydroxybenzoic Acids. The simplest 2-hydroxybenzoic acids are salicylic [69-72-7] and 5-sulfosalicylic [5965-83-3] acids, but they are of limited analytical importance.

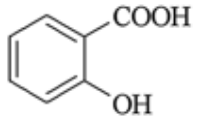

Salicylic acid

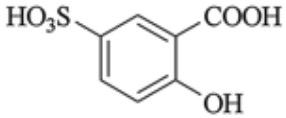

5-Sulfosalicylic acid

Frequently used complexometric indicators and photometric reagents include Chrome Azurol S, aurintricarboxylic acid, and Eriochrome Cyanine R:

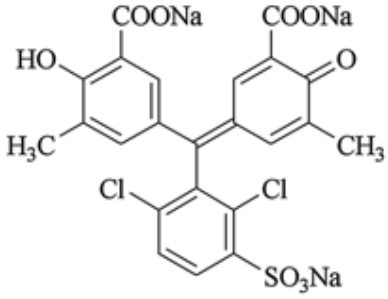

Chrome Azurol S

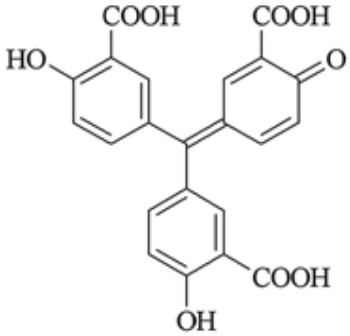

Aurintricarboxylic acid

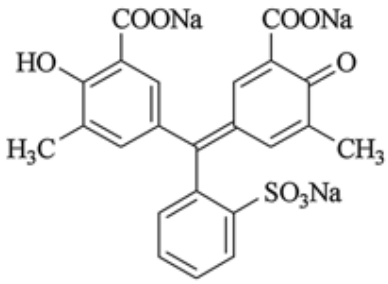

Eriochrome Cyanine R

2-Nitrosonaphthols. 2-Nitrosonaphthol indicators are employed preferentially in the analysis of cobalt; however, they can also be used as group reagents. Typical examples include 1-nitroso-2-naphthol (**3**), 2-nitroso-1-naphthol (**4**), and the water-soluble nitroso-R-salt (**5**).

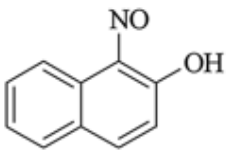

3

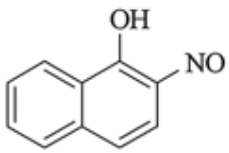

4

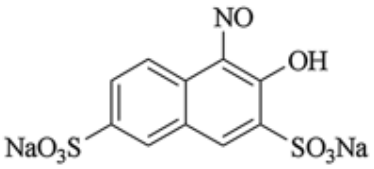

5

Methylaminodiacetic Acid Derivatives. Compounds containing a methylaminoacetic acid group and a triphenylmethane or anthraquinone group can also form six-membered ring chelate complexes. Dyes of this type are preferred as indicators in complexometry, but they are also suitable photometric reagents provided selectivity is not critical. The most important examples include:

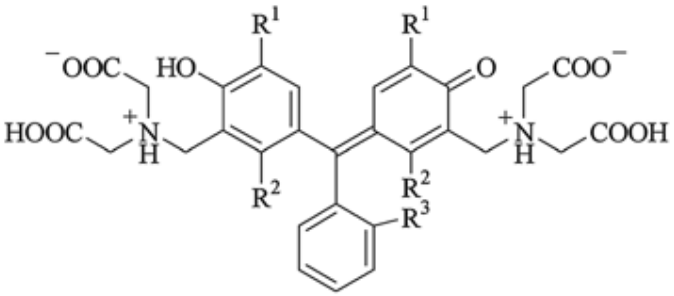

	R^1	R^2	R^3
Xylenol Orange	CH_3	H	SO_3H
Methylthymol Blue	$CH(CH_3)_2$	CH_3	SO_3H
Thymolphthalexone	$CH(CH_3)_2$	H	COOH
Phthalein Purple	CH_3	H	COOH

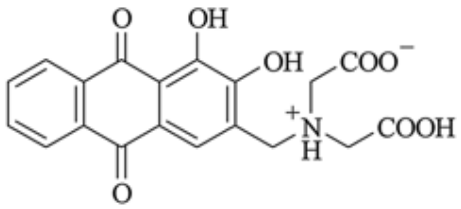

Alizarin Complexone

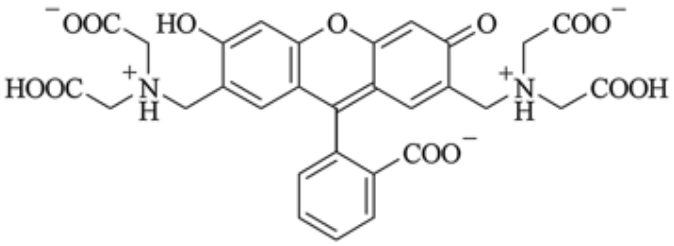

Fluorescein Complexone

[Top of Page]

8. Analytical Test Kits

Test kits are combinations of indicator reagents, miscellaneous reagents, and other auxiliaries that allow simple, usually rapid analyses. They can be reasonably well classified according to their applications:

1. Determination of chemical parameters (especially ions) in nonbiological media
2. Determination of chemical and biological parameters in biological media such as body fluids (i.e., diagnostic test kits, ⟶)
3. Determination of chemical parameters in gases (e.g., analytical tubes in work safety applications)

In this article, only the first group of analytical test kits is discussed. The pH papers and strips discussed in Section pH Indicator Papers represent the first examples of such products.

Properties and Limitations. Most analytical test kits are designed for use without additional laboratory equipment, which

means that

1. reagents and auxiliary chemicals are provided ready for use;
2. correct dosage is possible with very simple equipment generally included in the kit;
3. complex operations such as weighing and thermostating are generally not necessary;
4. test kits usually contain a buffer system which, to a certain extent, can counteract possible pH deviations and prevent associated interference;
5. the interferences of contaminating ions and other compounds are minimized by optimizing reaction conditions and reagent composition (e.g., incorporation of masking agents);
6. manufacturers may also provide information pertaining to the tolerance levels of contaminant ion concentrations;
7. auxiliary materials necessary for evaluation of results are included in the kits; and
8. batch-to-batch reproducibility of kit properties is guaranteed by the manufacturer.

Test kits are, of course, also prone to certain limitations:

1. Like classical test procedures, test kits can only detect free or easily released ions. To detect complexed or bound ions, a preliminary digestion stage is required.
2. Samples with extreme pH values must be adjusted prior to the determination.
3. When extreme conditions with regard to contaminant ions are involved, the sample must be appropriately treated.

Classification. Ideally, a test kit should combine the following properties:

1. high detection sensitivity,
2. maximum measuring range,
3. highest possible differentiation between divisions on the indicator scale,
4. absolute selectivity, and
5. convenient use.

Because a combination of all these requirements is not possible in practice, a compromise must be made. One or two of the above requirements are satisfied, and the others are met to an extent acceptable for the problem in question. Hence, a number of quite different types of test kits have been developed:

1. Test papers providing a yes – no result
   Manufacturers. Macherey-Nagel, Merck (FRG); Kyoritsu Chemical, Sibata, Toyo Roshi (Japan).
2. Test papers or strips giving semiquantitative results (generally, visual differentiation between four to eight concentrations)
   Manufacturers. As for (1) but also ETS and Serim (both USA).
3. Test strips providing quantitative results that require instrumental evaluation [68-72]
   Manufacturers. Merck (FRG).
4. Test kits for semiquantitative determinations (with up to ten, visually discernible concentrations)
   Manufacturers. Macherey-Nagel, Merck (FRG); Kyoritsu Chemical, Sibata, Yoshitomi (Japan); Lovibond (UK); CHEMetrics, Hach, Hellige, LaMotte (USA).
5. Test kits providing quantitative results that require instrumental evaluation (For example, manufacturers offer photometers that have been optimized for specific test kits and, hence, simplify the test applications considerably. However, the kits can also be used with conventional photometers.)
   Manufacturers. Hoelzle & Chelius, Lange, Macherey-Nagel, Merck, Riedel-de Haen (FRG); Kyoritsu (Japan); Lovibond, WPA (UK); CHEMetrics, Hach, LaMotte (USA).

Figure 3 shows nitrite test kits (Merck) that illustrate the possible options for sensitivity, measurement range, and measurement intervals.

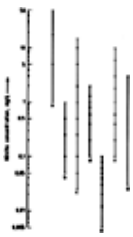

Figure 3. Measuring range and measuring intervals of nitrite test kits (Merck)

a) Test strip; b) – e) Test kits with visual evaluation in reflected light; f) Test kit with visual evaluation in transmitted light; g) Test kit with instrumental evaluation

Horizontal bars indicate divisions of the color scale used to read the concentration.

Parameters that can be determined by using currently available tests are listed below:

Acid binding capacity
Acidity
Acids, organic
Alkalinity
Aluminum
Amines
Ammonium
Amphetamines
Antimony

Adsorbable organic halogens (AOX)
Arsenic
Ascorbic acid
Bacteria, coliform
Barbiturates
Bismuth
Biological oxygen demand (BOD)
Boron
Bromide
Bromine
Cadmium
Calcium
Cannabinoids
Carbohydrates
Carbon dioxide
Carbon tetrachloride
Carbonate hardness
Chalk requirement (soil)
Chloride
Chlorine
Chlorine dioxide
Chromium
Cobalt
Cocaine
Chemical oxygen demand (COD)
Complexing agents
Copper
Cyanide
Cyanuric acid
Detergents
Diethylhydroxylamine
Ethanol
Ethylenediaminetetraacetic acid
Formaldehyde
Glycols
Gold
Heroin
Hydrazine
Hydrogen peroxide
Hypochlorite
Iodine
Iron
Lead
Lysergic acid diethylamide (LSD)
Magnesium
Manganese
Molybdenum
Morphine
Nickel
Nicotine
Nitrate
Nitrite
Nitrobenzene
Phenols
Oil
Opium
Oxidizing agents
Oxygen
Ozone
Penicillin
Pentachlorophenol
Peracetic acid
Permanganate value
Peroxidase
Peroxides
pH
Phenols
Phosgene
Phosphatases
Phosphates
Potassium
Pyridine
Quaternary ammonium compounds
Silicates
Silver

Sodium
Styrene
Sugar
Sulfate
Sulfide
Sulfite
Tannin – lignin
Tartaric acid
Thiocyanate
Thioglycolic acid
Thiophene
Tin
Titanium
Total organic carbon (TOC)
Triazole
Trichloroacetic acid
Urea
Vanadium
Water
Water hardness (total, carbonate, residual)
Water, inherent color
Zinc
Zirconium

Applications. The ranges of applications for test kits differ widely; the following are some important examples.

1. *Sample screening:* determination on site to establish whether laboratory analysis is necessary; determination of approximate concentrations in the laboratory
2. *Process control:* sewage plants, fermentation equipment, chemical production plants, drinking water treatment plants, and agriculture
3. *Rapid analysis of wastewater:* determination of substances that cannot withstand transportation, screening, control of threshold values
4. *Rapid analysis of surface water:* emergency analysis following water pollution, determination of special parameters (see above), testing the suitability of water in the construction industry, monitoring water quality in fish farming

Test kits involve a great deal of extremely innovative chemistry; however, the basic principles of analytical chemistry (e.g., correct sampling, validating) must not be neglected when using them.

[Top of Page]

References

General References

1. E. Bishop (ed.): *Indicators*, “International Series of Monographs in Analytical Chemistry no. 51,” Pergamon Press, Oxford 1972.
2. G. Jander, K. F. Jahr: *Maßanalyse, Theorie und Praxis der Titration mit chemischen und physikalischen Indikatoren*, 15th ed., De Gruyter, Berlin 1989.
3. G. Kraft, J. Fischer: *Indikation von Titrationen*, De Gruyter, Berlin 1972.
4. I. M. Kolthoff, E. B. Sandell, E. J. Meehan, S. Bruckenstein: *Quantitative Chemical Analysis*, 4th ed., Collier-Macmillan, London 1969, pp. 691 – 696.
5. V. Schmidt, W. D. Mayer in R. Bock, W. Fresenius,H. Günzler, W. Huber, G. Tölg (eds.): *Analytiker-Taschenbuch*, vol. 2,pp. 267 – 315 (1981); vol. 3, pp. 37 – 86 (1983) Springer Verlag, Berlin-Heidelberg-New York.
6. H. U. Bergmeyer (ed.): *Methods of Enzymatic Analysis*, 3rd ed., vol. **1**, VCH Verlagsgesellschaft, Weinheim, Germany 1983, pp. 197 – 232.
7. O. Sonntag: *Trockenchemie, Analytik mit trägergebundenen Reagenzien*, Thieme Verlag, Stuttgart-New York 1988.
8. F. W. Küster, A. Thiel, A. Ruland (eds.): *Rechentafeln für die chemische Analytik*, 104th ed., De Gruyter, Berlin 1993.
9. O.-A. Neumüller (ed.): *Römpps Chemie-Lexikon*, 10th ed., vol. **3**, Georg Thieme Verlag, Stuttgart 1997, pp. 1906 – 1908.
10. E. B. Sandell, T. S. West, *Pure Appl. Chem.* **19** (1969) 427 – 436.
11. I. M. Kolthoff, V. A. Stenger: *Volumetric Analysis*, Interscience, New York vol. 1, 1942; vol. 2, 1947.
12. I. M. Kolthoff, R. Belcher: *Volumetric Analysis*, vol. 3, Interscience, New York 1957.

Specific References

13. G. D. Christian: “The Effect of Salts on Titration,” *CRC Crit. Rev. Anal. Chem.* **5** (1975) 119 – 163.
14. G. Ackermann, L. Sommer, W. I. Stephen, *Pure Appl. Chem.* **57** (1985) 845 – 848.
15. Behringwerke, DE 2 436 257, 1974 (E. Pfeil).
16. Merck, DE 1 698 247, 1968 (K. Neisius, W. Bäumer).
17. Merck, DE 4 139 302, 1991 (W. Fischer, R. Klink).
18. Merck, DE 4 333 696, 1993 (W. Fischer, S. Baum, T. Hartig, M. Schleehahn).

19. Merck, company information.
20. I. Gynenes: *Titrationen in nichtwäßrigen Medien*, 3rd ed., Enke Verlag, Stuttgart 1970.
21. K. Stammbach: *Titrationen in nichtwäßrigen Lösungsmitteln*, Schweiz. Laborantenzeitung, Separatdruck 1970.
22. J. S. Fritz: *Acid-Base Titrations in Nonaqueous Media*, Allyn and Bacon, Boston 1973.
23. E. Singh, *Fresenius Z. Anal. Chem.* **282** (1977) 299 – 300.
24. S. K. Garg, D. C. Agarwal, G. K. Chaturverdi, *J. Indian Chem. Soc.* **58** (1981) 662 – 663. Links
25. J. F. Coetzee in I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry*, 2nd ed., Part I, vol. 3, Wiley-Interscience, New York-ChichesterBrisbane-Toronto-Singapore 1983, pp. 362 – 378.
26. R. Kumar, D. C. Agarwal, G. K. Chaturverdi, *Indian J. Chem.* **24 A** (1985) 631 – 632.
27. W. L. Hinze, H. N. Singh, Y. Baba, N. G. Harvey, *Trends Anal. Chem.* **3** (1984) 193 – 199.
28. O. S. Wolfbeis, *Trends Anal. Chem.* **4** (1985) 184 – 188.
29. N. Opitz, *Chem. Ind. (Düsseldorf)* **36** (1984) 742 – 744.
30. O. S. Wolfbeis, E. Fürlinger, H. Kroneis, H. Marsonner, *Fresenius Z. Anal. Chem.* **314** (1983) 119 – 124.
31. O. S. Wolfbeis, *Fresenius Z. Anal. Chem.* **320** (1985) 271 – 273.
32. O. S. Wolfbeis, B. P. H. Schaffar, *Talanta* **33** (1986) 867 – 870. Links
33. O. S. Wolfbeis, P. Hochmuth, *Mikrochim. Acta* 1984, 129 – 148. Links
34. O. S. Wolfbeis, H. Marhold, *Fresenius Z. Anal. Chem.* **327** (1987) 347 – 350.
35. T. J. Rink, R. Y. Tsien, T. Pozzan, *J. Cell. Biol.* **95** (1982) 189; Links *Chem. Abstr.* **97** (1982) 179 455 e.
36. J. A. Thomas, *Soc. Gen. Physiol. Ser.* **40** (1986) 311; *Chem. Abstr.* **106** (1987) 152 380 u.
37. G. Grynkiewicz, M. Poenie, R. Y. Tsien, *J. Biol. Chem.* **260** (1985) 3440. Links
38. G. H. R. Rao, *Biochem. Biophys. Res. Commun.* **132** (1985) 652. Links
39. G. Schwedt: *Fluorimetrische Analyse*, Verlag Chemie, Weinheim, Germany 1981.
40. O. V. Zui, A. V. Terletskaya, A. T. Pilipenko, *Zh. Anal. Khim.* **40** (1985) 120 – 124; *Chem. Abstr.* **102** (1985) 124 774 z.
41. I. E. Kalinichenko, T. M. Tkachuk, A. T. Pilipenko, *Zh. Anal. Khim.* **40** (1985) 1581 – 1585; *Chem. Abstr.* **104** (1986) 61 172 v.
42. L. L. Klopf, T. A. Niemann, *Anal. Chem.* **57** (1985) 46 – 51. Links
43. W. R. Seitz: "Chemiluminescence and Bioluminescence Analysis: Fundamentals and Biomedical Applications," *CRC Crit. Rev. Anal. Chem.* **13** (1981) 1 – 58.
44. L. J. Kricka, G. H. G. Thorpe, *Analyst (London)* **108** (1983) 1274 – 1296.
45. J. Schölmerich, R. Anderson, A. Kapp, M. Ernst, W. G. Woods (eds.): *Bioluminescence and Chemiluminescence—New Perspectives*, Wiley-Interscience, New York – Chichester – Brisbane – Toronto – Singapore 1987.
46. A. Hulanicki, S. G łab, *Pure Appl. Chem.* **50** (1978) 463 – 498.
47. J. Barek, E. Berka: "Redox Titrations in Nonaqueous Media," *CRC Crit. Rev. Anal. Chem.* **15** (1984) 163 – 222.
48. G. J. Misra, J. P. Tandon, *Fresenius Z. Anal. Chem.* **214** (1965) 94.
49. L. Erdey, G. Svekla, *Fresenius Z. Anal. Chem.* **167** (1959) 164.
50. P. V. K. Rao, G. G. Rao, *Anal. Chim. Acta* **29** (1962) 410.
51. J. P. Tandon, R. C. Mehrotra, *Fresenius Z. Anal. Chem.* **187** (1962) 40.
52. K. L. Chawla, J. P. Tandon, *Talanta* **12** (1965) 665. Links
53. G. Henze, R. Geyer, *Fresenius Z. Anal. Chem.* **200** (1964) 434.
54. A. A. Schilt: *Analytical Applications of 1,10-Phenanthroline and Related Compounds*, Pergamon Press, Oxford 1969.
55. E. Merck, company in formation, *Komplexometrische Bestimmungsmethoden mit Titriplex*, Darmstadt 1987.
56. A. Hulanicki, S. G łab, G. Ackermann, *Pure Appl. Chem.* **55** (1983) 1137 – 1230. Links
57. H. Freiser in I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry, Part I, vol. 3*, Wiley-Interscience, New York-Chichester-Brisbane-Toronto-Singapore 1983, pp. 395 – 505.
58. G. Schwarzenbach, H. Flaschka: *Die komplexometrische Titration*, 5th ed., Enke Verlag, Stuttgart 1965.
59. R. P řibil: *Applied Complexometry*, Pergamon Press, Oxford 1982.
60. O. G. Koch, G. A. Koch-Dedic: *Handbuch der Spurenanalyse*, 2nd ed., Springer Verlag, Berlin-Heidelberg-New York 1974.
61. J. Fries, H. Getrost: *Organische Reagenzien für die Spurenanalyse*, Merck, Darmstadt 1975.
62. S. B. Savrin: "Organic Reagents in Photometric Analysis," *CRC Crit. Rev. Anal. Chem.* **8** (1979) 55 – 109.
63. F. Umland: *Theorie und praktische Anwendung von Komplexbildnern in der Analyse*, Akadem. Verlagsgesellschaft, Frankfurt 1971.
64. Z. Marczenko: "Spectrophotometric Detection of Trace Elements," *CRC Crit. Rev. Anal. Chem.* **15** (1984) 163 – 222.
65. H. M. N. H. Irving: "The Analytical Applications of Dithizone," *CRC Crit. Rev. Anal. Chem.* **8** (1980) 321 – 366.
66. D. D. Perrin: *Masking and Demasking of Chemical Reactions*, Wiley-Interscience, New York 1970.
67. D. D. Perrin: "The Selection of Masking Agents for Use in Analytical Chemistry," *CRC Crit. Rev. Anal. Chem.* **5** (1975) 85 – 115.
68. G. Schwedt, *UmweltMagazin* **22** (1994) 134 – 135.
69. L. C. Waters, R. A. Jenkins, R. W. Counts, *DOE Methods*, Method Number MS 100.
70. L. C. Waters, R. W. Counts, A. Palausky, R. A. Jenkins, *J. Hazard. Mater.* **43** (1995) 1 – 12. Links
71. H. Rehbein, T. Schmidt, *Fischwirtschaft* **43** (1996) 37 – 39.

72. R. Gromes, D. Tanzer, *UmweltMagazin* **25** (1997) 84 – 85.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Indigo and Indigo Colorants

Standard Article

Elmar Steingruber[1]
[1]Frankenthal, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a14_149.pub2
Article Online Posting Date: August 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Physical Properties
2. Chemical Properties
3. Occurrence and Raw Materials
4. Production
5. Environmental Protection
6. Quality Specifications
7. Chemical Analysis
8. Storage and Transportation
9. Uses
10. Commercial Forms
11. Economic Aspects
12. Toxicology and Occupational Health
13. Indigo Derivatives

[Top of Page]

Indigo (1) is a natural dye obtained from the plants *Indigofera tinctoria* and *Isatis tintoria*, which have been used to color textiles in India, China, and Egypt for 4000 years. Since its discovery, the cultivation, extraction, and use of indigo have spread throughout the world, and trade in indigo became a significant economic factor, reaching a volume of about 8000 t in 1900.

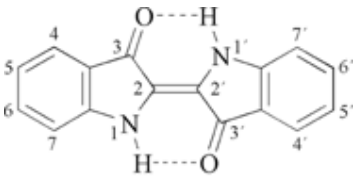

Indigo (**1**)

With the development and technical realization of chemical syntheses at the end of the nineteenth century, indigo became an important product of the growing chemical industry. After the proof of its structure by ADOLF VON BAEYER in 1883, experiments to prepare indigo synthetically culminated in 1887 in the commercialization and large-scale production of synthetic indigo by the Badische Anilin- und Soda-Fabrik (now BASF Aktiengesellschaft) in Ludwigshafen.

Today, more than 100 years later, synthetic indigo is still an important product of the dye manufacturing industry and has almost completely replaced the natural indigo from plants. Synthetic indigo is the largest single textile dye in the world, supplied by only a small number of producers to over 500 dye houses, whose main product is blue denim (i.e., cotton twill weave where the warp is dyed with indigo). Indigo is used for the manufacture of approximately 10^9 pairs of blue jeans annually. The apparently unending life cycle of indigo is, in view of its moderate fixation on the fiber, a phenomenon that is difficult to understand.

Worldwide acceptance of easily washed out, faded, indigo blue garments, not only as work clothes but also as fashionable sport and leisure articles, opened up new markets for indigo and ensured increasing demand for dyers and outfitters.

Apart from indigo itself, only few derivatives and related compounds, primarily halogenated indigo and thioindigo, have reached a position of steady economic importance (see Chap. Indigo Derivatives).

The considerable list of publications concerning indigo began nearly 170 years ago [1] and is indicative of the unusual cultural, as well as chemical, position occupied by this compound up until recent times [2-10].

[Top of Page]

1. Physical Properties

Indigo [482-89-3], indigotin, ($\Delta^{2,2'}$-biindoline)-3,3′-dione, 2-(1,3-dihydro-3-oxo-2, *H*-indol-2-ylidene)-1,2-dihydro-3, *H*-indol-3-one, $C_{16}H_{10}N_2O_2$, M_r 262.27, C.I. Vat Blue 1, C.I. 73000 exists at ambient temperature and pressure as dark blue–violet needles or leaves with a reddish bronze metallic appearance. It sublimes above 170 °C (*mp* 390 – 392 °C). Indigo is practically insoluble in water, aqueous acid, aqueous base, and nonpolar solvents, but slightly soluble in high-boiling polar solvents, (e.g., aniline, nitrobenzene, phenol, phthalic anhydride, dimethyl formamide, and dimethyl sulfoxide); it crystallizes in pure form from hot solutions. The color of indigo is unusually intense, compared to conjugated molecules of similar size [11], and is influenced strongly by solvents. Despite the possibility of cis – trans isomerism about the central C ═C double bond, only the trans isomer, stabilized by intra- and intermolecular hydrogen bonding, is observed [12, 13]. This fact also explains the extremely low solubility, high melting point, and vibrational and electronic absorption spectra of indigo. Bond orders and charge densities of the indigo chromophore have been investigated and calculated [12, 14, 15] as has the X-ray crystal structure [16, 17].

[Top of Page]

2. Chemical Properties

Indigo is very resistant to light and elevated temperature, even in the presence of air. Substitution by electrophiles or nucleophiles is difficult; only sulfonation in concentrated sulfuric acid, to give the di- and tetrasulfonic acid, and halogenation in nitrobenzene, to introduce up to six halogen atoms, are successful (See Chap. Indigo Derivatives). With strong mineral acids, indigo forms salts that decompose in water.

Reducing agents (e.g., sodium dithionite, hydroxyacetone, zinc, hydrogen or biological material in fermentation, recently also electric current [18, 19]) convert indigo to the leuco form indigo white [6537-68-4](**2**), which dissolves in sodium hydroxide solution to form the disodium salt [894-86-0] (**3**). This is the effective dyeing agent (vat indigo) which penetrates into the fiber and, after reoxidation adheres rather superficially to the surface of cotton fibers; in contrast, bonding to the polypeptide fibers of wool or silk is more saltlike.

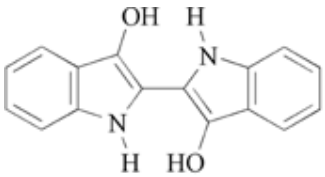

Leucoindigo,
Indigo White (**2**)

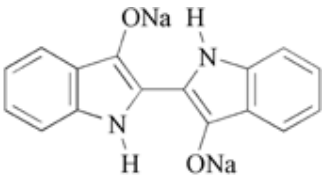

Leucoindigo Disodium Salt,
Indigo Vat (**3**)

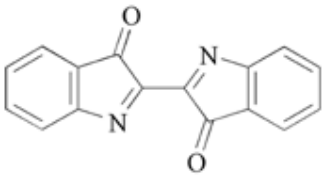

Dehydroindigo (**4**)

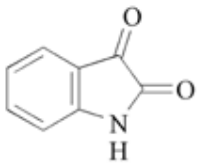

Isatin (**5**)

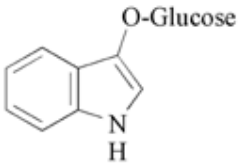

Indican (**6**)

Leuco indigo can be di- and tetraacetylated as well as methylated at the nitrogen and oxygen atoms.

The action of oxidizing agents leads to dehydroindigo (**4**); oxidation with permanganate or chromate cleaves the molecule at the central double bond, forming isatin [91-56-5] (**5**).

Concentrated alkali at elevated temperature leads to decomposition of indigo with formation of aniline, *N*-methylaniline, and anthranilic acid.

In biological systems, indigo is extremely stable. For example, in a biological wastewater treatment plant, indigo is not degraded, but simply adsorbed to the activated sludge, probably as a result of its low solubility; the elimination is 90–100 % [37]. This fact accounts for the low acute and chronic toxicity of indigo (see Chap. Toxicology and Occupational Health).

[Top of Page]

3. Occurrence and Raw Materials

Natural indigo occurs worldwide in numerous types of *Indigofera*, particularly in *Indigofera tinctoria* (in Asia and America), and in Cruciferae such as *Isatis tinctoria* (dyer's woad, in Europe). The concentration of indican [487-60-5] (6), a glucoside of indoxyl found in the stems and leaves, amounts to only 0.2 – 0.8 %. After indoxyl is released with enzymes, it oxidizes rapidly to indigo.

Before the introduction of synthetic indigo in 1897 by BASF, the cultivation of indigo plants, especially in India, on an area of more than 7000 km^2 was a huge business with a reported output of 19 000 t of indigo in 1897. The natural indigo production collapsed to 1000 t per year in 1914.

For the industrial synthesis of indigo, only two starting materials are now important. The first is *N*-phenylglycine [103-01-5] (7) produced from aniline and chloroacetic acid, or by reaction of aniline with formaldehyde and sodium cyanide or hydrogen cyanide followed by saponification of the nitrile (8). In an alternate reaction, the addition product **9** of aniline, formaldehyde, and sodium bisulfite is treated with sodium cyanide to give the *N*-phenylglycine nitrile **8**. Another route to the nitrile **8** is the reaction of aniline with hydroxyacetonitrile (**10**) [20-22].

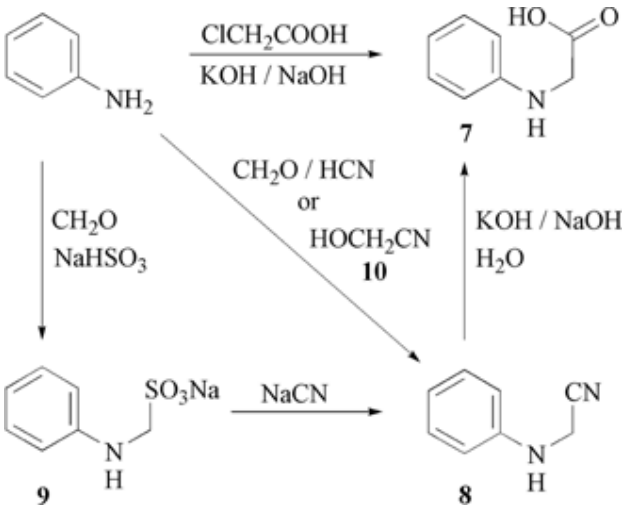

All four synthetic routes were used industrially [23]: the aniline-formaldehyde-hydrogen cyanide route by DyStar (which took over the textile dye business from BASF, Germany), the aniline-formaldehyde-sodium cyanide variation by Buffalo Color Corp. (USA; the plant has been closed in the meantime), the bisulfite version by Bann Química (Brazil), and the chloroacetic acid route by Chinese manufacturers. ICI and Mitsui have dicontinued their indigo manufacture.

The second raw material still used for the industrial synthesis of indigo is anthranilic acid [118-92-3] (**11**), which is either condensed with chloroacetic acid and alkali to give the salt of *N*-phenylglycine-*o*-carboxylic acid, or converted to the nitrile (**12**) with formaldehyde and hydrogen cyanide; subsequent saponification yields (**13**). The *N*-phenylglycine-*o*-carboxylic acid process was applied by BASF in 1897 to produce the first synthetic indigo on an industrial scale and employed in Ludwigshafen until 1924. The ring closure used in this process to form indoxyl is easier because no sodium amide is required, but the overall cost is considerably higher than the process based on aniline. The anthranilic acid route to synthesize indigo is therefore no longer of any commercial importance.

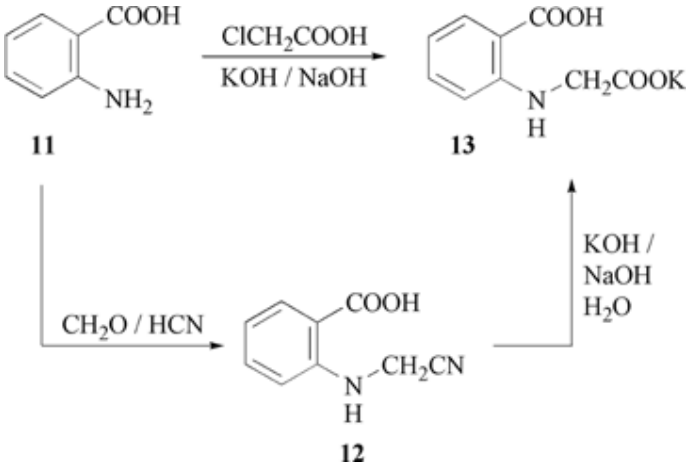

Other intermediates used in the pioneering phase of synthetic indigo were *N*-(2-hydroxyethyl)aniline and *N*-ethyl-*N*-phenylglycine; the use of these starting materials ceased in 1914 and 1904, respectively. The numerous other synthetic routes to indigo have no commercial importance. The Baeyer – Drewsen process of 1882 is no longer in use; in this method,

2-nitrobenzaldehyde was used to synthesize 4,4′-dichloroindigo, which is inaccessible by direct chlorination.

During the first 20 years of synthetic indigo manufacture, the decision as to which intermediate to use depended on the availability of the petrochemical raw materials benzene and naphthalene. These two chemicals, then accessible only from coal tar, were converted to aniline via nitrobenzene and anthranilic acid via phthalic anhydride, respectively. Today, the decision is influenced predominantly by the process technology available and the production costs; in this respect, the aniline process requires a higher expenditure for process equipment, but consumes cheaper raw materials.

Since about 1980, attempts to manufacture indigo by biosynthetic methods from glucose have been successful, using recombinant microorganisms [24-26]. The disadvantage of handling dilute solutions in huge bioreactors, the difficulties of separating indigo from biomass, and the treatment of large volumes of organic byproducts prevented a commercialization of this process.

[Top of Page]

4. Production

The extraction and isolation of *natural indigo* are described in detail in [27], [28]. The indigo plants are cut and fermented in water for 15 h. The glycoside indican (**6**) is enzymatically cleaved to glucose and indoxyl. Oxidation of indoxyl with air yields indigo. The most important producer countries were India (particularly Bengal), Java, and Guatemala. Purification, based on reduction and reoxidation, or fractional precipitation from sulfuric acid, was usually necessary because of the low purity of the natural indigo of only 15–70 %.

The rapid success of *synthetic indigo* at the beginning of the 20th century had a devastating effect on the plantations of natural indigo; by 1925 the production in India had dropped to only a few percent.

For the industrial synthesis of indigo the most important and also the most economic production process currently in use is ring closure of *N*-phenylglycine with sodium amide [7782-92-5] in anhydrous sodium hydroxide and potassium hydroxide solvent. The original process used by KARL HEUMANN in 1890 did not employ sodium amide and gave an indigo yield of only ca. 10 % [3].

After the observation by J. PFLEGER in 1901 that the addition of sodium amide resulted in a high yield of indigo, this process was introduced industrially by Farbwerke Hoechst. The process was also used by BASF in Ludwigshafen in 1905 with cheaper calcium oxide as desiccant. When the price of sodium dropped sharply, BASF changed to sodium amide in 1926 [29].

The Heumann – Pfleger process provides indigo of adequate purity in yields up to 90 %. Investigation of reaction conditions and raw material ratios is described in [2]. Ring closure of alkali salts of *N*-phenylglycine is carried out with about 2 mol of sodium amide at ca. 200 °C and releases ammonia. The indoxyl formed in the anhydrous melt exists as the stable dialkali-metal salt (**14**).

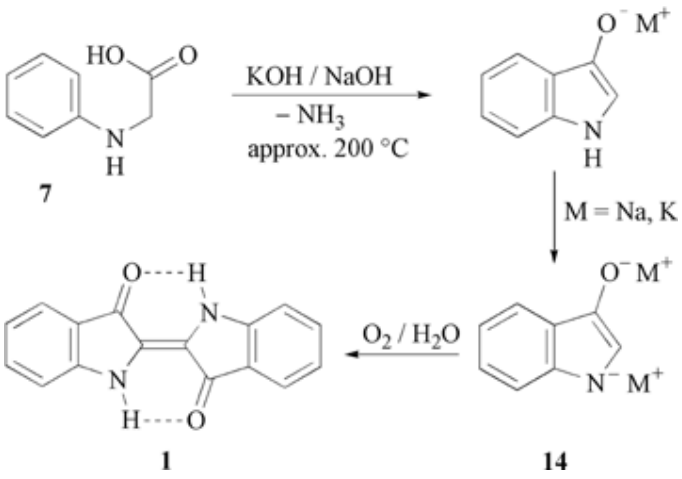

After ring closure, the melt is dissolved in water; the unstable indoxyl monoalkali salt is converted to indigo by oxidation with air. Some decomposition occurs, leading to aniline, *N*-methylaniline, and anthranilic acid as byproducts. Further details concerning the indoxyl melt are described extensively in [3, 23, 27-34].

The crystalline indigo formed by oxidation is separated from the aqueous alkali by filtration and washed with water. The usual commercial forms are indigo powder, prepared by drying the aqueous filter cake, and reduced indigo in the form of indigo vat 40 % solution and 60 % grains; some manufacturers offer a paste or liquid containing 20 % indigo. The filtrate of the oxidation mixture can be concentrated and the alkali used again for the indoxyl melt. The alkali solution concentrated to 50% is filtered to remove precipitated anthranilic acid, sodium carbonate, and other impurities. Recovery of the alkali makes the process more cost attractive because only a small amount of alkali is lost from the production cycle.

The production of indigo based on the more expensive anthranilic acid via *N*-phenylglycine-*o*-carboxylic acid (**13**) or its salts is simpler because no anhydrous alkali or dehydrating agent like sodium amide is required. Nevertheless, the reaction mixture is more viscous and therefore requires considerably more alkali to carry out the fusion [27, 28, 35].

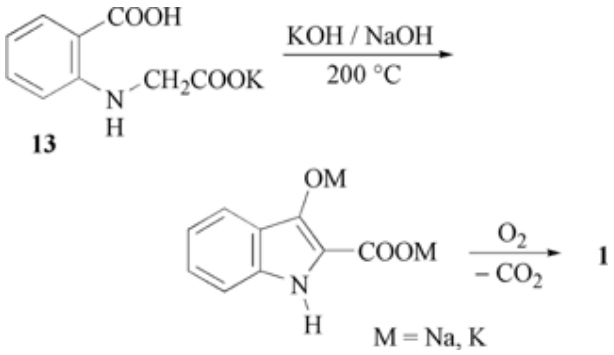

In this process, ring closure to the indoxyl carboxylic acid also occurs at ca. 200 °C. The yield is 70 – 90 %, depending on reaction conditions and the quality of the dicarboxylic acid (**13**). The fusion mass is dissolved in water, decarboxylation and oxidation by air yields indigo. Isolation of indigo is similar to that in the *N*-phenylglycine process; the product quality is also comparable.

The indoxyl fusion of *N*-phenylglycine (**7**) under anhydrous conditions can be carried out in carbon steel vessels, which suffer practically no corrosion. Stainless steel is preferred for the considerably more aggressive water-containing melt from *N*-phenylglycine-*o*-carboxylic acid (**13**). Both indoxyl oxidation and isolation of indigo use normal steel, which is sufficiently resistant to dilute caustic alkali.

In addition to the two processes described above, another synthetic route starts from *N*-(2-hydroxyethyl)aniline. Although an easily accessible raw material is involved and the process was used by BASF from 1907 to 1914, it is no longer of industrial importance. A considerably higher temperature is required, at which hydrogen is released spontaneously; the yield and purity of the indigo are lower. Further process improvements by Farbwerke Hoechst [23] could not revive this route.

Indigo is also marketed as the leuco indigo disodium salt **3**. Prepared originally by reduction of indigo with zinc or iron, it is now obtained by catalytic hydrogenation of indigo. Until 1980 indigo white (**2**) was marketed as a 50 % paste and used as a syrupy product, predominantly in the Far East; it is no longer in use. Indigo vat 60 % is a granular product obtained by drying a mixture of the sodium salt **3** and molasses, adjusted to an indigo content of 60 %. Since 2000, also a 40 % solution of indigo vat is manufactured and sold; the prereduced indigo gained interest in the dye houses by saving costs and reducing the salt content in the effluent.

[Top of Page]

5. Environmental Protection

Because of their chemical properties and toxicity, the chemicals used in the production of *N*-phenylglycine and indigo (particularly aniline, formaldehyde, hydrogen cyanide, potassium or sodium hydroxide, sodium, ammonia, and sodium amide), demand high standards for safe handling and processing. These are achieved by extensive use of closed systems.

Indigo possesses very low mammalian toxicity [36], [37] (see Chap. Toxicology and Occupational Health) and does not have an adverse effect on adapted activated sludge systems (e.g., in rivers or biological effluent treatment plants). The rate of biological degradation is quite low because of its low solubility, but indigo is adsorbed on activated sludge and thereby eliminated from wastewater.

Byproducts from indigo synthesis may be present in wastewater resulting from filtrates, wash water, and flue-gas scrubbers; the wastewater must then be purified. The byproducts aniline and anthranilic acid are easily degraded in biological water treatment plants, but *N*-methylaniline is degraded more slowly. The ammonia released in the sodium amide process and in the saponification of the nitriles **8** or **12** is preferably recovered but can also be removed from waste streams by nitrification. Any alkali reaching the wastewater must be neutralized with acid.

The extraction of natural indigo also involves ecological problems, because 98 % of the dry plant material do not consist of indigo and have to be degraded in the fermentation process or deposited in the indigo plantations.

[Top of Page]

6. Quality Specifications

Indigo is available in the following forms: technically pure powder, microgranules, or lumps considered to contain 100 % indigo; liquid dispersions in water with an indigo content of 20–30 %; indigo vat 60 % grains and indigo vat 40 % solution.

The powder and microgranules still contain traces of byproducts from the synthesis, together with small amounts of additives that influence the dispersibility. The following parameters serve as quality criteria: indigo content, dusting and wetting properties, reduction rate, properties of the vat solution, and dyeing characteristics.

With the liquid dispersions, which usually contain 20 – 30 % indigo and are free-flowing liquids, quality criteria include indigo content, alkali content (0 – 6 %), viscosity, specific gravity, storage stability, sedimentation behavior, and those parameters used for the powder (i.e., reduction behavior, the quality of the reduced solution, and dyeing properties).

Indigo vat 60 % grains and indigo vat 40 % solution are tested for indigo, alkali, and water content, and for the amount of

reoxidized material.

For natural indigo, the most important quality criteria are its purity, solubility after reduction, and dyeing behavior.

[Top of Page]

7. Chemical Analysis

Numerous methods have been suggested for the determination of indigo content, some of which have been developed in the era of natural indigo [27], [32]. The preferred method consists of sulfonation of an indigo sample with concentrated sulfuric acid, followed by titration of the diluted sulfonic acid with sodium dithionite solution [38], [39]. Indigo sulfonic acid can also be determined spectrophotometrically [40]; highly purified indigo, obtained by extraction with nitrobenzene or phenol, or by recrystallization, is used as a standard. Other quality criteria are discussed in Chapter Quality Specifications. The effective indigo content of indigo vat 40 % and 60 % is obtained after air oxidation of the vat solution and filtration, and of liquid paste indigo products, after extraction with dilute mineral acids and filtration.

[Top of Page]

8. Storage and Transportation

No specific limitations have been placed on indigo because of the low toxicity. The danger of dust explosions must be considered in storing indigo powder (dust explosion class 2). With a particle size of $(20 – 60)\times10^{-6}$ m, ignition may occur at concentrations exceeding 50 g/m^3. Indigo powder, indigo grains, indigo vat, and neutral liquid products are similarly regarded as nonhazardous for transportation [37]. Alkaline liquid forms with 1 – 5 % alkali are regarded as corrosive for transportation because of the action of alkali on aluminum [41].

[Top of Page]

9. Uses

Only small quantities of indigo are used for dyeing woven fabrics of cotton, wool, and silk (e.g., wax batik for fashionable hand-crafted objects). The greater part of the indigo produced worldwide is used in continuous dyeing of cotton yarn. The dyed cotton warps are woven with undyed weft in various twill weaves to give the typical blue denim cloth. The predominantly heavy, sailcloth-like material is used by about 400 producers to manufacture mainly blue jeans. The jeans “phenomenon” was examined in detail in 1981 from historical, economic, fashion, artistic, and business perspectives [42, 43]. Thanks to the popularity of cotton clothes dyed with indigo (of which ca. 90 % are blue jeans and only ca. 10 % other fashion articles including printed textiles), the consumption of indigo has attained a continuous growth rate worldwide since about 1975, which is above average of the textile industry.

In dye houses specializing in indigo dyeing, continuously operated slasher or warp dyeing machines are used in which the cotton yarn is immersed successively in three to six dye baths of vatted indigo. Indigo is reduced to the soluble leuco form with alkali and hydrosulfite, or pre-reduced indigo vat is used, or electric current is used in specialized equipment to reduce the indigo [18, 19]. After each immersion into the indigo vat, the yarn is squeezed and the adsorbed indigo oxidized by air [44]. Worldwide, more than 100 dye houses run about 200 dyeing machines, with about 70 % being in the Far East, particularly the Hong Kong area, followed by the United States and Europe. In 1980, a “Loop Dye 1 for 6” dyeing machine was introduced, in which cotton yarn passes through the same vat up to six times [45].

[Top of Page]

10. Commercial Forms

Commercial products based on indigo include the following:

DyStar Indigo gran.
Indigo Bann 20 (paste)
Indigo Bann 30 (paste)
DyStar Indigo Vat 60 % Grains
DyStar Indigo Vat 40 % Solution
DyStar Indigo Black 5006 (mixture of indigo with a yellow dye)

[Top of Page]

11. Economic Aspects

From 1960 to 1980, the growing demand for blue jeans resulted in a steady increase in the indigo production capacity worldwide. By 1988, the world capacity had reached 14 000 t per year of synthetic indigo and was still distributed over more than ten countries, 12 000 t were supplied by the five major manufactureres BASF (Germany), Buffalo Color Corp. (USA), ICI (UK and Brazil), Mitsui Toatsu (Japan) and Taisung (Taiwan). Smaller quantities were produced in China, the Soviet Union, Mexico, Korea, and India.

Fierce price competition and globalization of world trade started to cause major changes, beginning about 1990. Local indigo production was started in Brazil by Bann Química, and low price indigo from China began to compete on all markets. In 1996, the Zeneca indigo plant in the UK (formerly ICI), was sold to BASF, which then reached a total capacity of 7000 t of indigo

per year. In 2000, BASF included its textile dye business in the joint venture DyStar, which consisted of the textile dye activities of Bayer and former Hoechst. In the meantime, the UK plant was closed, Mitsui terminated their indigo production in Japan, and in 2004 also Buffalo Color Corp. in the USA ended their operations. The world indigo consumption had reached about 30 000 t in 2002.

The remaining indigo suppliers in the world are DyStar with a 7000 t/a capacity, Bann Química in Brazil with about 4000 t/a, and manufacturers in China who constructed huge new capacities, an estimated 40 000 t/a or more. The indigo production coming from China into the world market was expected to amount to 14 000 – 18 000 t in 2003.

The consumption of indigo for dyeing cotton yarn amounts to only 1.5 – 3 % of the weight of the cotton. The weight of blue denim varies from ca. 150 to 450 g/m^2. With 1.5 m^2 of material per pair of trousers, only 3 – 12 g of indigo are used for each pair of blue jeans. The world consumption of 30 000 t/a of indigo in 2003 allows the production of 2–4 × 10^9 pairs of blue jeans with an enormous sales value, of which indigo itself represents only a small fraction.

As a result of the enormous indigo production capacities in China and the export pressure onto the world market, indigo market prices have suffered a steep drop from $ 15 – 20 per kilogram in 1988 to $ 7 – 8 per kilogram in 2003 in Europe; Chinese indigo is even offered at less than $ 4 per kilogram.

The estimated distribution of the indigo consumption in 2002 is shown in Table 1. Since 1988, the demand has increased particularly in Turkey, India, and China. The Far East continues to be the largest growing indigo market.

Table 1. Indigo consumption in 2002 (t)

Europe	3300
USA, Canada	3300
Central America	2700
South America	2700
Turkey, Near East, Middle East	2000
Africa	480
India, Pakistan	1950
China	16 000
Rest of Asia	6300
Total	38 780

[Top of Page]

12. Toxicology and Occupational Health

Despite the long use of natural and synthetic indigo, only few toxicological data were known until a few decades ago. Since 1962, however, the toxicological properties of indigo have been increasingly examined and reported, in connection with the application for registration of purified indigo as D & C Blue No. 6 for use in pharmaceuticals and cosmetics [36].

The low acute oral toxicity (LD_{50}> 5000 mg/kg) has been confirmed in mammals [36] fish toxicity LC_{50} is > 1000 mg/L (96 h, *Leuciscus idus*) [37]. Experiments to determine the acute dermal and inhalation toxicity demonstrated no toxic effects. Subchronic and chronic feeding experiments on rats and dogs with purified indigo at up to 3 % of feed weight showed, after a certain adaptation time, no serious negative effects other than coloring of the organs. Similarly, no negative effects were found in rats over three generations in studies of reproduction or in teratogenic investigations. Repeated application of purified indigo to the skin yielded no negative effects. With alkaline forms of commercial indigo, the corrosive action of the alkali can be observed. No indication of sensitization in humans was found after repeated skin application of neutral indigo.

The Ames test for mutagenicity towards bacteria was applied to synthetic indigo in 1979 and 1982 and revealed a minimal positive mutagenic effect, which was later found to be due to impurities [46-48].

[Top of Page]

13. Indigo Derivatives

In comparison with the use of indigo as a special product for dyeing blue jeans, other indigoid dyes (e.g., thioindigo and substituted indigo derivatives) have considerably less economic importance.

Thioindigo [522-75-8], C.I. Vat Red 41, C.I. 73 300, $C_{16}H_8O_2S_2$, M_r 296.37, available from thiophenol-*o*-carboxylic acid and chloroacetic acid via 2-hydroxythionaphthene (**15**), is not used for cotton dyeing because of its poor color fastness and limited affinity for cotton, but is a brilliant red disperse dye for polyester (C.I. Disperse Red 364), marketed by DyStar.

4,7,4′,7′-Tetrachlorothioindigo, obtained from the dichloro compound **16** by cyclization or by direct chlorination of thioindigo, is an industrially important pigment (see ⟶ Pigments, Organic – Thioindigo Pigments).

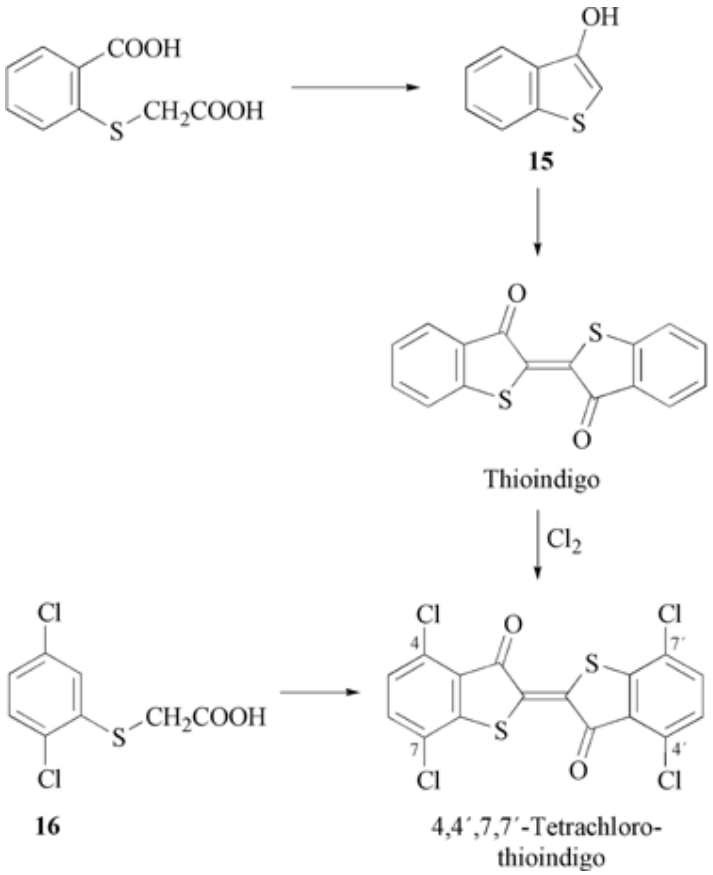

Among the halogen derivatives, *Tyrian purple* was obtained from historic times to the Middle Ages from various sea snails. Despite its dull red-violet shade, this dye was much sought after. In 1909 the compound was identified as 6,6′-dibromoindigo [19201-53-7], but it cannot be prepared by direct bromination of the parent indigo.

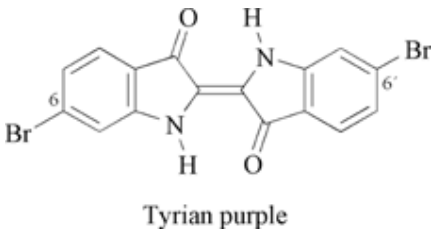

The halogenated indigo dyes still in use today are specialty products; their sales volume amounts to much less than 1 % of the unsubstituted indigo products.

5,7,5′,7′-Tetrabromoindigo [2475-31-2] , C.I. Vat Blue 5, C.I. 73 065, is synthesized by bromination of indigo in acetic acid. On account of its interesting neutral blue color the product is still sold as DyStar Brilliant Indigo 4 B (for dyeing) and DyStar Indigo 4B sfx. (for printing). The commercial forms consist of a mixture of tri- and tetrabromoindigo with a low chlorine content.

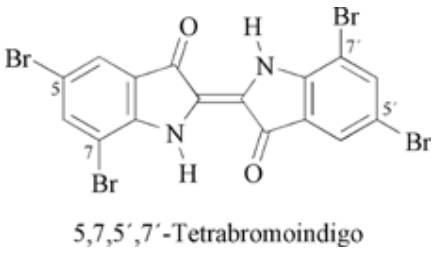

5,5′-Dibromo-4,4′-dichloroindigo, C.I. Vat Blue 2, C.I. 73045, obtained by bromination of 4,4′-dichloroindigo and sold as Brilliant Indigo 4 G, had been a desirable dye for printing because of its interesting greenish blue shade. It is no longer commercially available.

Among the sulfonation products of indigo the disodium salt of 5,5′-indigodisulfonic acid $C_{16}H_8N_2Na_20_8S_2$ is known as indigotine I or indigo carmine [860-22-0], C.I. Acid Blue 74, C.I. 73015, M_r 466.36.

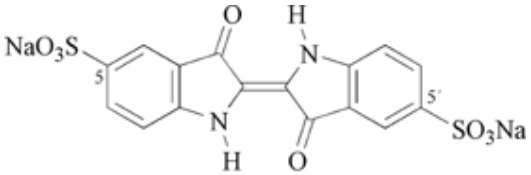

5,5′-Indigodisulfonic acid disodium salt,
Indigocarmin, Indigotin I, C.I. Food Blue 1

Registered also as C.I. Food Blue 1 and in the USA as FD&C Blue 2, it has found an application as colorant for food, cosmetics, drugs, and animals feeds. It is marketed by BASF as Sicovit Indigotin 85 E 132.

[Top of Page]

References

1. Dumas, *Annalen der Pharmacie* **21** (1837) 72 – 77. Links
2. F. Henesey, *J. Soc. Dyers Colour.* **54** (1938) 105 – 115. Links
3. K. Holzach, *Angew. Chem.* **A/60** (1948) 200 – 204. Links
4. H. Cassebaum, *Melliand Textilber* **48** (1967) 207. Links
5. M. Seefelder: *Indigo: Kultur, Wissenschaft und Technik*, ecomed, Landsberg, 1994.
6. H. Schmidt, *Chemie in unserer Zeit* **31** (1997), 121 – 128. Links
7. A. Dutly: Geschichte des Indigos, ETH Zürich, 2002, http://www.dutly.ch/indigohtml/indigo 1.html.
8. P. Miederer in: K. Hunger (ed.), *Industrial Dyes: Chemistry, Properties, Applications*. Wiley-VCH, Weinheim, Germany, 2003, p. 40 – 43, 204 – 214.
9. H. Zollinger: *Color Chemistry*, 3rd. rev. ed. VHCH Verlag Helvetica Chimica Acta, Zürich, and Wiley-VCH, Weinheim, Germany, 2003, p. 259 – 268.
10. M. Seefelder: *Indigo in Kultur, Wissenschaft und Technik*, BASF Aktiengesellschaft, Ludwigshafen am Rhein 1982.
11. W. Lüttke, M. Klessinger, *Chem. Ber.* **97** (1964) 2342 – 2357. Links
12. M. Klessinger, W. Lüttke, *Tetrahedron* **19** (1963) Suppl. 2, no. 315 – 335. Links
13. M. Klessinger, W. Lüttke, *Chem. Ber.* **99** (1966) 2136 – 2145. Links
14. E. Wille, W. Lüttke, *Angew. Chem.* **83** (1971) 853; Links *Angew. Chem. Int. Ed. Engl.* **10** (1971) 803. Links
15. E. Wille, W. Lüttke, *Chem Ber.* **106** (1973) 3240. Links
16. H. v. Eller, *Bull. Soc. Chim. Fr.* 1955, 1426 – 1444. Links
17. M. Klessinger, *Dyes and Pigments* **3**, (1982) 235. Links
18. W. Schrott T. Bechtold, *Textilveredlung* 2003 (9/10), 19 – 22. Links
19. T. Bechtold E. Burtscher O. Bobleter, *Recent Res. Dev. in Electrochem.* 1998, 245 – 264. Links
20. E. Knoevenagel, *Ber. Dtsch. Chem. Ges.* **37** (1904) 4080. Links H. Bucherer, *Ber. Dtsch. Chem. Ges.* **39** (1906) 2800. Links
21. H. Bucherer, *Ber. Dtsch. Chem. Ges.* **39** (1906) 2800. Links
22. BASF, GB 1 119 256, 1966 (E. Tolksdorf).
23. J. G. Kern, H. Stenerson, *Chem. Eng. News* **24** (1946) 3164 – 3181. Links
24. D. Murdock, B. D. Ensley, C. Serdar, M. Thalen, *Biotechnology* 1993 **11**, 381. Links
25. J. W. Frost, J. Lievense, *New J. Chem.* 1994 **18**, 341. Links
26. H. Bialy, *Nature Biotechnol.* 1997 **15**, 110. Links
27. *Ullmann*, 1st ed. **6**, 505 – 508.
28. *Ullmann*, 2nd ed. Urban & Schwarzenberg, Berlin, Wien 1930, **6**, 244 – 246.
29. *Ullmann*, 3rd ed., Urban & Schwarzenberg, München, Berlin **7**, 1956, 113 – 114. **8** (1957) 748 – 763.
30. *Kirk-Othmer*, 2nd ed., **7**, (1965) 621 – 625; **11**, (1966) 562 – 580, Interscience Publishers, New York.
31. *Kirk-Othmer*, 3nd ed., **8**, 363 – 368;
32. K. Venkataraman: *Chemistry of Synthetic Dyes*, vol. **2**, Academic Press, New York 1952, pp. 1003 – 1022.
33. *Ullmann*, 4th ed, **13**, 177 – 181.
34. *Ullmann's*, 5th ed., **A 14**, 149 – 156.
35. M. Phillipps, *J. Ind. Eng. Chem.* **13** (1921) 759 – 762. Links
36. K. H. Ferber, *J. Environ. Pathol. Toxicol. Oncol.* **7** (1987) 73 – 83. Links
37. *Sicherheitsdatenblatt DyStar Indigo gran*, DyStar Textilfarben GmbH & Co., Leverkusen, 01.08.2002.
38. BASF Indigo rein, BASF Aktiengesellschaft Ludwigshafen am Rhein 1907, p. 14.
39. Badische Anilin- und Soda-Fabrik *Z. Farben Ind.* **8** (1909) 121 – 127, 137 – 141. Links
40. Lunge-Berl: *Chemisch-technische Untersuchungsmethoden*, 7th ed., Springer Verlag, Berlin 1924, pp. 1033 – 1059.
41. Sicherheitsdatenblatt DyStar Indigo Vat 40% Lsg., DyStar Textilfarben GmbH & Co., Leverkusen, 03.09.2002.
42. B. Wehdemeier-Pusch, A. Pusch, *Das Phänomen "Jeans"*, Universität Bielefeld 1981.
43. *Blaue Wunder für individuelle Jeansmode*, BASF-Gruppe, 09.07.2003, http://www.pressi.com/de/Mitteilung/69301.html.

44. Process Data Sheet: *Continuous Dyeing of Cotton Warps with Indigo*, BASF Aktiengesellschaft, Ludwigshafen am Rhein 1976 (German 1983).
45. B. Kramrisch, *Am. Dyest. Rep.* (Nov. 1980) 34 – 38.
46. B. N. Ames, J. McCann, E. Yamasaki, *Mutat. Res.* **31** (1975) 347 – 364. Links
47. J. M. Muzzall, W. L. Cook, *Mutat. Res.* **67** (1979) 1 – 8. Links
48. W. F. Jongen, G. M. Alink, *Fd. Chem. Toxic.* **20** (1982) 917 – 920. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Indium and Indium Compounds

NOËL FELIX, Metallurgie Hoboken-Overpelt, Hoboken, Belgium

1. **Introduction** 1
2. **Properties** 1
3. **Resources and Raw Materials** 2
4. **Production** 2
4.1. **Production from Zinc Circuits** 3
4.2. **Production from Lead Circuits** 5
4.3. **Production from Tin Circuits** 6
4.4. **Other Production Developments** . . . 6
4.5. **Refining** 7
5. **Quality Specifications** 7
6. **Analysis** 7
7. **Indium Compounds** 7
8. **Uses** 8
9. **Economic Aspects** 10
10. **Toxicology and Occupational Health** . 11
11. **References** 11

1. Introduction

Indium [*7440-74-6*], In, atomic number 49, A_r 114.82, is a metallic element of group 13 of the periodic table.

Indium has two natural isotopes: ^{115}In (95.72 %) and ^{113}In (4.28 %). The abundance of indium in the earth crust is comparable to that of silver, i.e., 0.1 ppm [1].

Indium was discovered in 1863 by F. REICH and T. H. RICHTER during spectrometric analysis of sphalerite ores at the Freiberg School of Mines in Germany [2]. Indium is named after the indigo blue spectral lines that led to its identification.

2. Properties [3–8]

Physical Properties. Indium is a crystalline, silvery white metal, which is very soft (softer than lead), ductile, and malleable. Indium retains its highly plastic properties at cryogenic temperatures. Indium can be deformed almost indefinitely under compression, is easily cold-welded, and, like tin, emits a high-pitched "cry" on bending.

Indium generally increases the strength, corrosion resistance, and hardness of an alloy system to which it is added. Even small concentrations of indium can have a considerable influence.

Molten indium wets clean glass. Indium has a low melting point (429.75 K) but a high boiling point (2353.15 K). Indium becomes superconducting at 3.37 K.

Some physical properties of indium are as follows [3], [5], [7]:

Electronic configuration	[Kr] $4d^{10}\ 5s^2\ 5p^1$
A_r	114.82
Thermal neutron cross section at 2200 m/s	
absorption	$(194 \pm 2) \times 10^{-28}$ m^2/atom
scattering	$(2.2 \pm 0.5) \times 10^{-28}$ m^2/atom
Crystal structure	tetragonal A 6 $a_o = 0.458$ nm $c_o = 0.494$ nm
Number of atoms per unit cell	4
Atomic radius (coordination number 12)	0.162 nm
Ionic radius (In^{3+}, coordination number 6)	0.081 nm
Atomic volume	15.73×10^{-6} m^3/mol
Density at 293.15 K	7.310 g/cm^3
mp	429.75 K
bp	2353.15 K
Enthalpy of fusion	3.26 kJ/mol
Enthalpy of vaporization	231.2 kJ/mol
Specific heat at 298.15 K	26.70 kJ mol^{-1} K^{-1}
Entropy at 298.15 K	57.7 kJ mol^{-1} K^{-1}
Entropy of fusion	7.58 kJ mol^{-1} K^{-1}
Entropy of vaporization	98.56 kJ mol^{-1} K^{-1}
Coefficient of linear expansion (273.15 – 373.15 K)	3×10^{-5} K^{-1}
Vapor pressure p, kPa, at temperature T, K, between *mp* and *bp*	$\log p = 9.835 - \frac{12860}{T} - 0.7 \log T$
Thermal conductivity (273.15 – 373.15 K)	71.1 W m^{-1} K^{-1}
Surface tension at temperature T, N/m	$0.602 - 10^{-4}\,T$
Electrical resistivity	
at 3.38 K	superconducting
at 273.15 K	8.4×10^{-8} Ωm
temperature coefficient (273.15 – 373.15 K)	4.9×10^{-3} K^{-1}

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_157

Standard electrode potentials	
$In^{3+} + 3e \rightleftharpoons In$	−0.338 V
$In^{3+} + 2e \rightleftharpoons In^{+}$	−0.40 V
$In^{2+} + e \rightleftharpoons In^{+}$	−0.40 V
$In^{3+} + e \rightleftharpoons In^{2+}$	−0.49 V
Brinell hardness	0.9
Tensile strength	2.645 MPa
Modulus of elasticity	10.8 GPa

Chemical Properties. Metallic indium is not oxidized by air or oxygen at room temperature. It reacts directly with arsenic, antimony, the halogens, oxygen, phosphorus, sulfur, selenium, and tellurium when heated.

Indium dissolves only slowly in cold dilute mineral acids, but more readily in hot dilute or concentrated acids. Alkalis do not attack the massive metal.

Indium forms alloys with most other metals. Extensive solid solutions are formed with lead, thallium, and mercury.

3. Resources and Raw Materials [9]

Indium does not occur in the native state. It is widely spread in nature, generally in very low concentrations. The content of indium in the earth crust is estimated to be 0.1 ppm. Indium is found as a trace element in many minerals (see Table 1).

Table 1. Indium contents of some minerals [9] ∗

Mineral	Composition	Indium content, ppm
Sphalerite	ZnS	0.5 – 10 000
Galena	PbS	0.5 – 100
Chalcopyrite	$CuFeS_2$	0 – 1500
Enargite	Cu_3AsS_4	0 – 100
Bornite	Cu_5FeS_4	1 – 1000
Tetrahedrite	$(Cu,Fe)_{12}Sb_4S_{13}$	0.1 – 160
Covellite	CuS	0 – 500
Chalcocite	Cu_2S	0 – 100
Pyrite	FeS_2	0 – 50
Stannite	Cu_2FeSnS_4	0 – 1500
Cassiterite	SnO_2	0.5 – 13 500
Wolframite	$(Fe,Mn)WO_4$	0 – 16
Arsenopyrite	FeAsS	0.3 – 20

∗ Reproduced from Roskill: The Economics of Indium 4th ed., 1987, by permission of the publisher, Roskill Information Services Ltd.

Sphalerite is the most important indium-containing mineral followed by lead and copper sulfides. The amount of indium in sphalerite may vary widely even within a deposit. The indium concentration is typically 10 – 20 ppm, but may be as high as 1 % (10 000 ppm). A sphalerite deposit with a very high indium content (0.2 – 0.3 %) is the Huari Huari zinc deposit in Bolivia. Tin deposits with high indium contents (up to 0.21 wt %) are found in Cornwall (United Kingdom) and New Brunswick (Canada).

4. Production

During smelting processes for the recovery of base metals, indium concentrates in byproducts such as residues, flue dusts, slags, and metallic intermediates. The processes for the recovery and production of indium are often complex and sophisticated; they are characterized by a low direct extraction efficiency. Therefore, the recycling of the indium contained in intermediate products is important.

4.1. Production from Zinc Circuits

A process for the recovery of indium from secondary zinc oxide is described in [10], [11]. The process is illustrated in Figure 1. In the first step, most of the zinc is removed by leaching with dilute sulfuric acid. The residue is leached with dilute hydrochloric acid to dissolve the indium. Tin is removed from the indium solution by neutralization to pH 1. Further neutralization then precipitates indium. The indium residue is leached with sodium hydroxide to give crude indium hydroxide as an intermediate product. Dissolution of this indium hydroxide in dilute hydrochloric acid gives an indium solution which is purified by cementation of copper and arsenic with iron, followed by cementation of tin and lead with indium. Indium is recovered from the purified solution by cementation with aluminum.

In the process developed by the Anaconda Copper Mining Company (Fig. 2), indium is recovered from zinc calcine (roasted zinc blende) or zinc oxide fume [12]. Zinc calcine or zinc oxide fume is first leached in sulfuric acid at moderate pH, dissolving the bulk of the zinc and leaving indium in the residue. In the second step, indium is dissolved from the residue by leaching with sulfuric acid (20 – 25 g/L H_2SO_4).

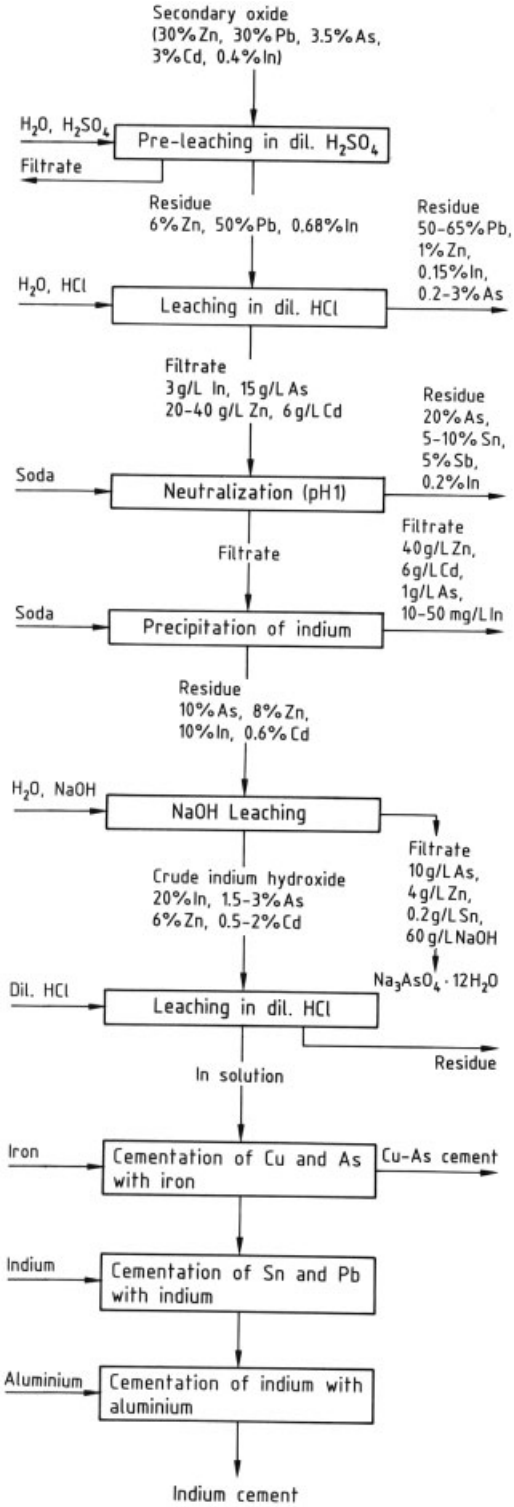

Figure 1. Recovery of indium from secondary zinc oxide [10], [11]

Indium is precipitated from the resulting solution by adding zinc oxide and sodium hydrogen sulfite. The indium precipitate is purified by leaching with concentrated sodium hydroxide, washing with water and with dilute sulfuric acid to remove zinc. The purified residue is dissolved in sulfuric acid, heavy metals are precipitated with hydrogen sulfide, and indium is recovered as a sponge by adding zinc. The indium sponge may be melted with sodium hydroxide and cast into bars. Alternatively, the sponge may be purified by dissolving it in hydrochloric acid and adding barium chloride to remove sulfate, and hydrogen sulfide to remove heavy metals. After filtration, indium is recovered from the solution by electrolysis.

The extraction of indium from sphalerite of the Rammelsberg and Bad Grund deposits in the Federal Republic of Germany is described in [13]. On refining the retort zinc by distillation, indium remains with the lead. After oxidation of the lead residue, indium is recovered from the litharge, which contains 1 – 4 % indium. The litharge is leached with sulfuric acid, which dissolves 88 – 97 % of the indium. Indium is recovered by addition of zinc slabs to give a crude indium sponge containing ca. 95 % indium. The crude indium is redissolved in sulfuric acid. The resulting solution is purified first by cementation with indium strips to remove tin and then with slabs of a zinc alloy, containing 1 % indium and 0.3 % cadmium, to remove thallium and cadmium. Indium with a purity of greater than 99.99 % is recovered from the purified solution by precipitation with aluminum strips followed by melting under sodium hydroxide and sodium cyanide.

Chlorination of the lead retort residue under a molten salt cover for the extraction of indium has been patented by Asarco [14]. The chloride slag is dissolved in water and indium is recovered by cementation on zinc or aluminum.

At the zinc refinery of Mitsubishi in Akita [15], indium is recovered from the leach residue (400 ppm In) of the zinc concentrate (120 ppm In). The residue is treated by a sulfation roasting process. Leaching of the sulfated product in sulfuric acid dissolves about 60 % of the indium to give a leach solution containing 50 – 70 mg/L of indium. This solution is concentrated to an in-

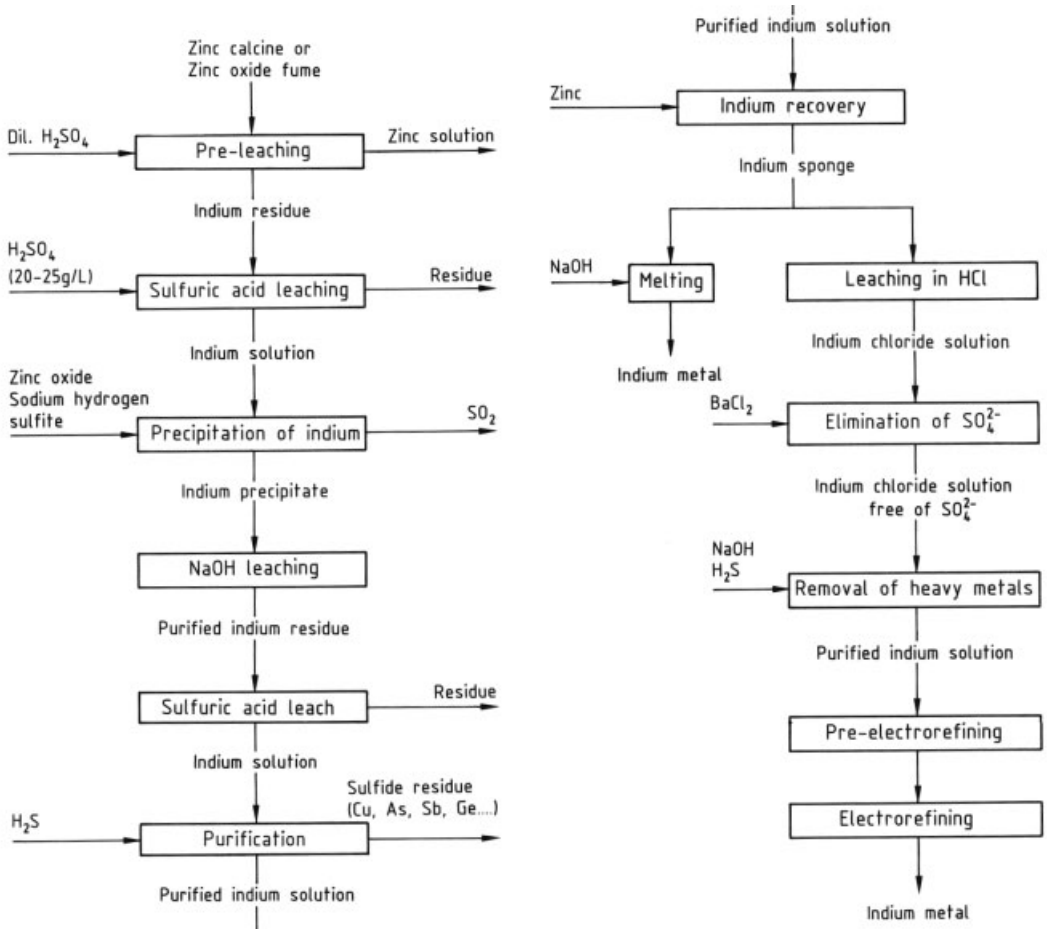

Figure 2. Indium recovery by Anaconda Copper Mining Company [12]

dium content of 2 g/L by double neutralization with calcium carbonate. The enriched solution is then subjected to solvent extraction followed by back extraction with hydrochloric acid. The aqueous solution is concentrated to 20 – 30 g/L indium. Indium is recovered by immersing aluminum plates in the solution. Indium sponge adhering to the aluminum plates is removed for melting and casting into anodes. Indium metal of 99.99 % purity is then obtained by electrorefining.

4.2. Production from Lead Circuits

At Cominco's integrated lead – zinc smelter at Trail in Canada, tin and indium accumulate in the fumes of the continuous drossing furnace [16]. The flow sheet for the process is illustrated in Figure 3. A mixture of fumes, coke, and antimony-containing slag is melted in a direct-fired rotary furnace. Arsenic is removed by adding scrap iron to form iron speiss. The residual tin – indium – antimony – lead alloy is cast into anodes and electrolyzed in a fluorosilicate solution, giving a solder cathode (Sn – Pb alloy) and indium – antimony anode slimes. The slimes are roasted with sulfuric acid. Indium is dissolved from the sulfated product by leaching with water. Copper is removed from the indium solution by cementation on indium sheets. Indium is then recovered from the solution as a sponge by addition of aluminum or zinc. The indium sponge is melted and electrorefined to give indium of 99.99 % purity.

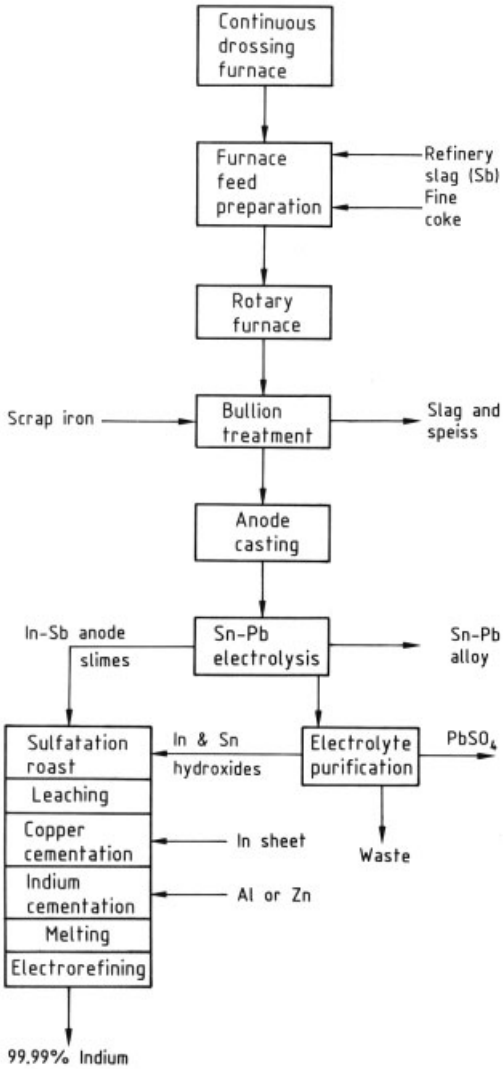

Figure 3. Indium recovery by Cominco [16]
Previously published in the Proceedings of the Symposium on Quality Control in Non-Ferrous Pyrometallurgical Processes from the 24th Annual Conference of Metallurgists, August, 1985. Reprinted with permission of the Canadian Institute of Mining and Metallurgy.

Indium is recovered from lead bullion by Centromin of Peru at the La Oroya smelter, and by Metallurgie Hoboken-Overpelt (MHO) at the Hoboken smelter. At MHO, indium concentrates in a complex residue during refining of lead by the Harris process (→Lead) [17]. At the La Oroya refinery, lead bullion is first drossed to remove copper (→Lead), then tin and indium are removed at a higher temperature [18], [19]. The tin – indium dross is reduced to metal using coke, a mixture of lead and zinc chlorides is added, and indium is recovered as indium chloride in the slag.

The chloride slag of indium, lead, zinc, and tin is leached by wet grinding. The solution is purified by cementation of lead and tin on indium sheet. Indium is recovered as a sponge by adding zinc.

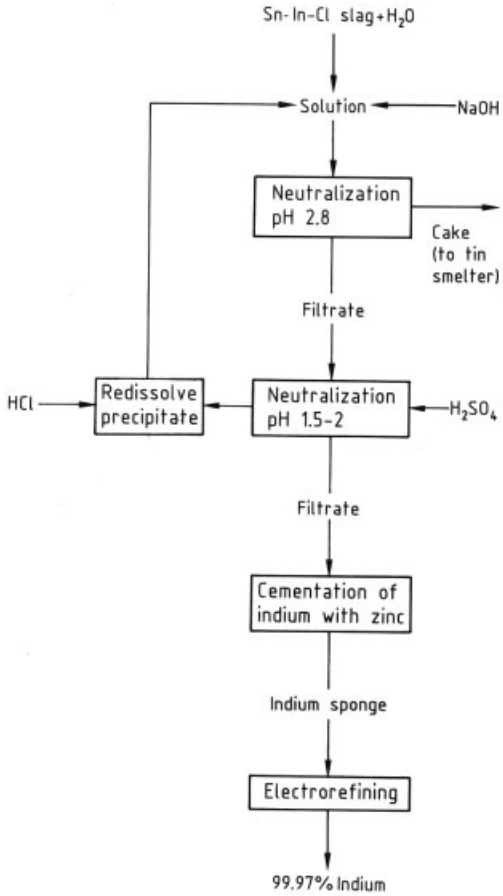

Figure 4. Indium recovery by Capper Pass [21]
Reproduced by permission of the Institution of Mining and Metallurgy, from P. Halsall: Indium–Extraction from Lead, Zinc and Tin circuits (Fig. 9). Trans. Instn. Min. Metall. (*Sect. C: Mineral Process. Extr. Metall.*) **97**, 1988, C 98.

4.3. Production from Tin Circuits

At Capper Pass, chlorination of electrorefined tin gives a tin chloride slag which contains 2.7 % indium [20]. This slag is treated as shown in Figure 4 [21]. Most of the tin is precipitated in the

first neutralization stage. The precipitate from the second neutralization stage, which contains some indium, is recycled for indium recovery. Indium is recovered from the filtrate as a sponge by cementation with zinc. The indium sponge of 95 % purity is electrorefined to give a 99.5 % indium cathode, which is recast as the anode for the second electrorefining stage, in which indium of 99.97 % purity is produced.

4.4. Other Production Developments

Solvent extraction systems have been investigated to extract indium from various solutions.

The use of a mixture of hydroxyquinolines (Kelex 100 and LIX 26) to extract indium from acidic or alkaline aqueous solutions has been reported [22].

Di-(2-ethylhexyl) phosphoric acid (D2EHPA) in kerosene, and tributyl phosphate (TBP) in kerosene are used to extract indium from the solution obtained by leaching lead-containing residues with sulfuric acid [23]. The concentration of arsenic, cadmium, zinc, copper, antimony, and iron in the extract is greatly reduced. For good separation, iron should be present in the ferrous state.

Solutions of D2EHPA and TBP in kerosene are also reported to extract indium from sulfuric acid solutions that contain tin. The addition of soluble fluorides to the acid solution prior to solvent extraction inhibits the extraction of tin [24].

A process for the separation of indium from gallium and germanium in sulfuric acid solution using solvent extraction is described in [25].

Supported liquid membranes are used to extract indium from a copper-dross leach solution [26], [27]. The feed solution contains indium (0.5 g/L), copper (60 g/L), and sulfuric acid (30 g/L). The extractant is a 40 % solution of D2EHPA in Escaid 100 (a mixture of 80 % aliphatic and 20 % aromatic long-chain hydrocarbons). A microporous polypropylenic support is used.

The recovery of indium and gallium from spent semiconductor material has also been reported [28].

4.5. Refining

Sponge indium with a purity up to 99.5 % requires upgrading for most uses. For the semiconductor industry, a purity of at least 99.9999 % is required. Electrorefining is a commonly used technique.

In electrorefining, cotton-bagged indium anodes and pure indium cathodes are used. Metallic impurities collect as anode slime in the cotton bag, which thus prevents their entering the bulk electrolyte and contaminating the indium deposit. The electrolyte is generally a chloride solution, but the use of solutions of sulfate [29], cyanide [30], fluoroborate [31], and sulfamate [32] has also been reported.

A typical composition of the electrolyte is 20 – 75 g/L In, 80 – 100 g/L NaCl, 1 g/L glue, pH 2.2 – 2.5. The cathode current density is 1 – 2 A/dm^2. Apart from the dendritic nature of the indium deposit, the difficulty with electrorefining lies in the fact that impurities which generally accompany indium have standard potentials close to that of indium (see Table 2).

Table 2. Standard potentials, in volts

Cd^{2+}/Cd	−0.402
In^{3+}/In	−0.338
Tl^{+}/Tl	−0.336
Sn^{2+}/Sn	−0.136
Pb^{2+}/Pb	−0.126

For purities up to 99.999 %, multistage electrorefining is required [33], or electrorefining must be supplemented by vacuum distillation or zone refining or both of them.

The difference between the melting point and boiling point of indium is greater than for most other elements. This property allows the elimination of impurities such as cadmium, sulfur, selenium, tellurium, zinc, lead, and thallium by vacuum distillation [34].

Elimination of cadmium, copper, zinc, gold, silver, and nickel by zone refining is described in [35]. Intensive cooling is necessary because of the low melting point of indium.

A method for refining indium to semiconductor grade, in which a chelating agent is used to absorb indium selectively from solution, has been reported [36].

5. Quality Specifications

Qualities ranging from 99.97 % to 99.99999 % indium are commercially available. Impurities may be metallic (Ag, Cu, Pb, Sb, Sn, Cd, Tl, Zn, Fe, Al, Ga) and nonmetallic (S, Cl, O, Se, N). Their concentrations depend on the origin of the indium-containing feed material and on the production and refining process.

Commercial grades are 99.97 % (3N7), 99.99 % (4N), 99.999 % (5N), 99.9999 % (6N), and 99.99999 % (7N).

6. Analysis

Spectroscopic analysis enables the detection of indium down to 10 ppm due to its characteristic lines in the indigo blue region at wavelengths of 4511.36, 4101.76, 3256.09, and 3039.39 nm. The quantitative analysis of indium in ores, compounds, and alloys is described in [37]. Spark source mass spectrometry (SSMS) and graphite furnace atomic absorption spectrometry (GFAAS) are used in the analysis of high purity indium (≥99.9999 %). About 40 elements can be determined with detection limits of 10–30 ppb by spark source mass spectrometry [38].

7. Indium Compounds

The most common valence of indium is three. Monovalent and bivalent compounds of indium with oxygen, sulfur, and halogens are also known. The trivalent indium compounds are the most stable. The lower valence compounds tend to disproportionate to give the corresponding trivalent compounds and indium metal.

The properties of some indium compounds are given in Table 3 [7], [39].

Indium(III) oxide, In_2O_3, is formed by thermal decomposition of indium hydroxide, carbonate, nitrate, or sulfate in air, or by burning indium in air. It is a pale yellow, amorphous solid when formed at low temperature, but forms red-brown crystals when it is prepared at high temperature. The amorphous form is soluble in acids, but not in alkalies. Indium trioxide is reduced to indium by hydrogen, ammonia, sodium, aluminum, and carbon. *Indium(II) oxide* [*12136-26-4*], InO, and *indium(I) oxide* [*12030-22-7*], In_2O, are formed by reduction of indium-(III) oxide under carefully controlled conditions.

Indium hydroxide [*56108-30-6*], $In(OH)_3$, is precipitated from solutions of indium salts by adjusting the pH in the range 4–10.

Indium trichloride, $InCl_3$, is obtained by burning indium in excess chlorine. It is a white, crystalline substance which is easily sublimed, and is very soluble in water. *Indium dichloride,* $InCl_2$, is obtained by the reaction of indium with hydrogen chloride at 200 °C. *Indium monochloride,* InCl, is formed by passing indium dichloride vapor over heated indium. The lower chlorides, InCl and $InCl_2$, disproportionate in water to give indium trichloride and indium metal.

Indium orthoborate [*13709-93-8*], $InBO_3$, is prepared by melting a mixture of boric oxide and indium hydroxide.

Indium forms alkyl and aryl compounds such as trimethyl indium [*3385-78-2*], $In(CH_3)_3$; triethyl indium [*923-34-2*], $In(C_2H_5)_3$; and triphenyl indium [*3958-47-2*], $In(C_6H_5)_3$. They are made by treating indium with the corresponding mercury alkyls or mercury aryls, or by reaction of indium trichloride with a Grignard reagent.

The semiconducting compounds *indium antimonide*, InSb, *indium arsenide,* InAs, and *indium phosphide*, InP, are prepared by direct combination of the high-purity elements at high temperature. Indium phosphide is also obtained by thermal decomposition of a mixture of a trialkyl indium compound and phosphine (PH_3).

8. Uses

Table 4 summarizes the western-world consumption of indium by end use [9], [40], [41].

Low-melting alloys represent a major use of indium. Indium is added to *solder alloys*, mostly composed of tin and lead, to increase thermal fatigue resistance, and to improve malleability at low temperature and corrosion resistance. Indium solders are used in *electronics* for fixing semiconductor chips to a base, for assembling semiconductor devices and hybrid integrated circuits, and to seal glass to metal in vacuum tubes. These solders are mainly used in the form of a paste, whose main components are Pb, Ag, Cd, In, Sn, and an organic binder.

Table 3. Properties of some indium compounds [7], [39]

Compound	M_r	Density, g/cm^3	*mp*, K	*bp*, K	Enthalpy of formation, kJ/mol	CAS Registry Number
In_2O_3	277.63	7.180	2183		−925	[*1312-43-2*]
InCl	150.27	4.190	498	881	−186	[*13465-10-6*]
$InCl_2$	185.73	3.620	509	796	−362.4	[*13465-11-7*]
$InCl_3$	221.18	3.450	856	771 ∗	−535.0	[*10025-82-8*]
InAs	189.47		1215		− 57.7	[*1303-11-3*]
InSb	236.57		798		− 31.1	[*1312-41-0*]
InP	145.79		1328		− 75.2	[*22398-80-7*]

∗ Sublimes

Fusible indium alloys are usually based on bismuth in combination with lead, cadmium, tin, and indium.

Fusible indium alloys are used to bend thin-walled tubes without wrinkling the wall or changing the original cross section. The alloy is cast in the tube, rendering it effectively solid. After bending, the alloy can be melted out using hot water.

Table 4. Estimated western-world consumption of indium in 1985 [9]

Use	Consumption, t
Low-melting alloys	12
Bearings	9
Dental alloys	7
Nuclear reactor control rods	3.3
Low-pressure sodium lamps	4 – 5
Electrical contacts	8
Alkaline dry batteries	6.8 ∗
Phosphors	6 ∗∗
Semiconductors	
Lasers, photodetectors, integrated circuits (Ga – As In – P, In – P)	5
CCDs in infra-red video cameras (image orthicons, In – Sb)	3.5
Liquid crystal displays	7

∗ 1987 data [40].
∗∗ 1986 data for Japan [41].

These alloys are also used in fire-control systems. Restraining links that hold alarm, water valve, and door operating mechanisms in place are soldered with a low-melting alloy. A rise in temperature sufficient to melt the alloy results in the system coming into operation.

Fusible alloys are used for making disposable patterns and dies for the founding of ferrous and non-ferrous metals.

They are also used as temperature indicators in situations where other methods of temperature measurement are impracticable. Small rods of alloys of known melting point are inserted into equipment such as aircraft header tanks, test bearings, and experimental rigs. The melting of the alloy thus indicates the temperature attained in particular parts of the test system.

Addition of indium to lead – tin *bearings* for heavy-duty and high-speed applications provides particularly high resistance to fatigue and seizure. These bearings are used in piston-type aircraft engines, high-performance automobile engines (formula 1) and in turbo-diesel truck engines.

The addition of indium to gold *dental alloys* improves their mechanical characteristics and increases resistance to discoloring. Small amounts of indium are used to improve the machinability of gold alloys for *jewelry*.

Indium is used in *nuclear reactor control-rod* alloys (80 % Ag, 15 % In, and 5 % Cd). The consumption of indium in this field is estimated at 3.3 t/a for the replacement of control rods in existing nuclear power plants [9].

In *low pressure sodium lamps* (SOX lamps), indium is applied as a thin layer of indium – tin oxide on the inside of the outer glass envelope. The indium – tin oxide film increases the operating temperature and the efficiency of the lamp. Low pressure sodium lamps, characterized by their orange color, are mainly used in the United Kingdom, Belgium, and the Netherlands. The production of SOX lamps accounts for 4 – 5 t/a indium [9].

Indium is now used as a corrosion inhibitor in *alkaline dry batteries* as a substitute for mercury. This allows the mercury content of the amalga-

mated zinc powder used in their production to be reduced from 7 % to 1 – 3 % [42]. A target of very low mercury content, or even complete removal, is being pursued. The low-mercury zinc powders have corrosion and degassing characteristics which are similar to those of the 7 % mercury powders.

Indium orthoborate is used as a *phosphor* in *cathode-ray tubes*, where it is applied as a thin layer. Indium phosphors emit light for a longer time than zinc sulfide phosphors but are more expensive. They are used particularly in computer monitors where image stability is of great importance. In Japan, consumption of indium for this application was 6 t in 1986 and 8 t in 1987.

Intermetallic compounds of indium find application as *semiconductors*. The development of the second generation of optical fiber communication systems in the early 1980s with substantially lower levels of attenuation at wavelengths near the infrared region of 1.3 µm and 1.55 µm has led to the development of new semiconductor lasers and photodetectors. *Laser diodes* based on indium – gallium arsenide – phosphide emit in the 1.27 – 1.6 µm region. Indium – gallium arsenide *photodetectors* have high responsivity at 1.3 and 1.55 µm.

The development of integrated circuits and opto-electronic integrated circuits based on indium is still in the research phase.

High-purity indium is used for laser diodes, photodetectors, and integrated circuits.

Charge coupled devices (CCDs) operating in the infrared region are based on indium antimonide. Infrared detectors and infrared video cameras, such as orthicons, use indium antimonide.

Thin films of indium – tin oxide are able to transfer electricity to light emitting or light reflecting elements. They find application in *liquid crystal displays* and to a lesser extent in electroluminescent displays and touch panels.

Minor uses of indium include thin plastic films coated with indium – tin oxide that are being developed for use in touch panels (e.g., for computer screens), liquid crystal devices, and electroluminescent lamps. Indium oxide is used as a coating for *glass*. Window glass coated with an indium – tin oxide film to reflect infrared radiation is being investigated. Indium oxide films doped with tin oxide are used to prevent icing and fogging on aircraft windshields. Their application as windshield coatings in automobiles is in development, particularly in Japan. Indium antimonide is used in *Hall generators* to measure magnetic fields. *Solar cells* based on copper – indium diselenide are being developed [43]. Small amounts of indium compounds are used as additives for *lubricating oils*, and to produce *yellow glass*.

9. Economic Aspects

The world production of indium in 1986 has been estimated at 80 t. Table 5 lists the estimated indium production by country and company for 1986. The main producers are located in Belgium, Japan, France, and Italy. Important indium refiners are located in the United States (Indium Corporation of America, 25 t in 1986 [9]) and the United Kingdom (Mining and Chemical Products, 10 t in 1986 [9]). The U.S: Bureau of Mines estimates world production in 1992 140 metric tons. [45]

Table 5. Estimated indium production in 1986 [9]

Country	Company	Production, t
Belgium	Metallurgie Hoboken-Overpelt	15
France	Peñarroya	19
Italy	Pertusola	19
United Kingdom	Capper Pass	3
	Johnson Matthey	2
Netherlands	Billiton	1
German Democratic Republic	VEB	1
Federal Republic of Germany	Preussag	1
Japan	Nippon Mining	16
China		6
Former Soviet Union		6
Peru	Centromin-Peru	4
Canada	Cominco	6
World production		80

Reproduced from Roskill: The Economics of Indium 4th ed., 1987, by permission of the publisher, Roskill Information Services Ltd.

Figure 5 shows the evolution of the producer price for standard grade indium (99.97 %) from the Indium Corporation of America. For indium of 99.99 % purity, a premium of 2 – 5 $/kg is charged. Very pure indium (7N or 99.99999 %) costs ca. 1500 $/kg [9].

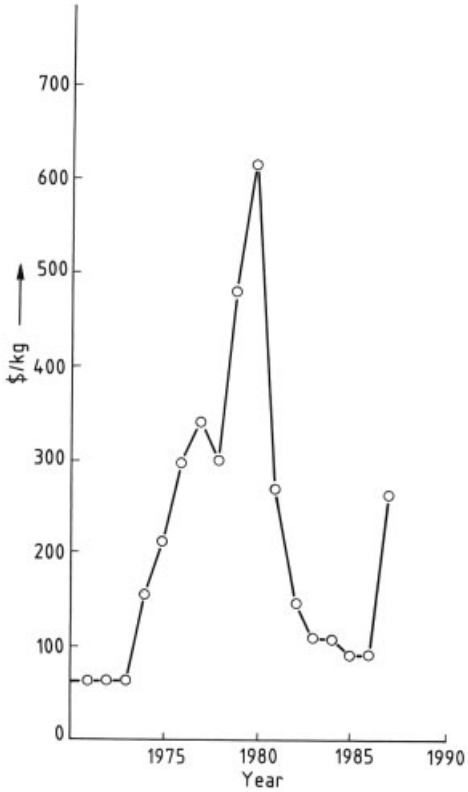

Figure 5. Evolution of the producer price for standard grade indium (99.97 %) (Indium Corporation of America)

10. Toxicology and Occupational Health

There is no evidence of any health hazard from industrial use of indium. No systemic effects in humans exposed to indium have been reported [44]. The TLV for indium is 0.1 mg/m^3.

11. References

1. H. E. Suess, H. C. Urey, *Rev. Mod. Phys.* **28** (1956) 56.
2. F. Reich, T. H. Richter, *J. Prakt. Chem.* **89** (1863) 441.
3. *Kirk Othmer,* 3rd ed., **13,** p. 207.
4. J. R. Mills, R. A. King, C. E. T. White: *Rare Metals Handbook,* 2nd ed., Chap. 13, Reinhold Publishing Corporation, London 1961, p. 220.
5. *Ullmann,* 4th ed., **13,** 197.
6. *Handbook of Chemistry and Physics,* 62nd ed., The Chemical Rubber Co., Cleveland, Ohio, 1981/82.
7. J. C. Bailar et al.: *Comprehensive Inorganic Chemistry,* 1st ed., Pergamon Press, Oxford 1973, pp. 983 – 1117.
8. J. W. Mellor, *A Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol. 5. Chap. XXXV, Longmans, Green and Co., London 1946.
9. Roskill: *The Economics of Indium 1987,* 4th ed., Roskill Information Services Ltd., London 1987.
10. E. Theurich, *Freiberg. Forschungsh.* **B 90** (1963) 93.
11. L. Müller, *Freiberg. Forschungsh.* **B 90** (1963) 105.
12. Anaconda Copper Mining Comp., US 2 384 610, 1940 (H. M. Doran et al.).
13. R. Kleinert, "Indium from the Rammelsberg Ores," *Min. Mag.* **83** (1950) 146.
14. Asarco, US 2 433 770, 1947 (Y. C. Lebedeff).
15. S. Takayanagi, K. Yajima, paper presented at the 22nd Annual Conference of Metallurgists of the Metallurgical Society of CIM, Edmonton, August 1983, paper no. 19.4.
16. E. F. Milner, W. A. Van Beek, "Rotary furnace reduction of lead drossing fume for recovery of tin and indium," paper presented at the 24th Annual Conference of Metallurgists of the Metallurgical Society of CIM, Vancouver, August 1985.
17. J. De Keyser, W. Jespers, "The Harris Refinery of Metallurgie Hoboken-Overpelt," paper presented at the 110th AIME meeting, Chicago, Feb. 1981, AIME-TMS paper no. A-81-10.
18. J. Coyle, *Trans. Electrochem. Soc.* **85** (1944) 223.
19. T. Quarm, *Trans. Inst. Min. Metall.* **60** (1950) 77.
20. Capper Pass Limited, GB 2 109 008 A, 1982 (P. Halsall).
21. P. Halsall, *Trans. Inst. Min. Metall. Sect. C* **97** (1988) 93.
22. Preussag Aktiengesellschaft Metall, US 4 666 686, 1987 (W. Krajewski, K. Hanusch).
23. Hazen Research, Inc. USA, CA 1 188 105, 1985 (E. Reynolds Jones, A. R. Williams).
24. Cominco Ltd., CA 1 218 237, 1987 (R. Perri et al.).
25. Tian Run-cang Zou Jia-yun, Zhou Ling-Zhi in M. J. Jones, P. Gill (eds.): *Mineral Processing and Extractive Metallurgy,* London IMM, 1984, pp. 615 – 624.

26. Samim Societa Azionarva Minero Metallurgica S. P. A., BE 902 890, 1986 (R. Guerziero et al.).
27. R. Guerziero, L. Meregalli, X. Zhang, *ISEC Int. Solvent Extr. Conf.,* Dechema, Frankfurt 1986, pp. 585 – 589.
28. Nippon Mining, JP-KK 88 016 338, 1988.
29. H. B. Linford, *Trans. Electrochem. Soc.* **79** (1941) 443.
30. Oneida Community Ltd., US 1 965 251, 1934 (M. S. Murray, D. Gray).
31. General Motors, US 2 409 983, 1946 (W. M. Martz).
32. Indium Corporation of America, US 2 458 839, 1949 (J. R. Dyer, T. J. Rowan).
33. G. S. Rao, G. H. Phatak, K. Gangadharan, *Indian J. Technol.* **12** (1974) 256.
34. A. Arkady et al., US 4 287 030, 1981.
35. U. Wiese, *Erzmetall* **34** (1981) no. 4, 190.
36. Sumitomo Chem. Ind., JP-KK 204 830, 1986.
37. C. E. T. White, *Metal. Bull.* **47** (1977).
38. Y. D. Liu, J. Verlinden, F. A. Lams, E. Adriaenssens, *Bull. Soc. Chim. Belg.* **95** (1986) 309.
39. I. Barin, O. Knacke, O. Kubaschewski: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin-Heidelberg-New York, 1977.
40. D. B. Winter, *Met. Bull.,* August 1988.
41. *Jpn. Ind. Rare Met. Ann. Rep.,* April 1988.
42. M. Meeus, Y. Strauven, L. Groothaert, "Power Sources 11," *Proc. Int. Power Sources Symp.,* 15th, (1987) 281 – 299.
43. C. E. T. White: *Advanced Materials & Processes Inc., Metal Progress,* December 1986, 69 – 72.
44. L. Friberg, G. F. Nordberg, V. Vouk: *Handbook on the Toxicology of Metals,* Elsevier/North-Holland Biomedical Press, Amsterdam 1979, pp. 429 – 436.
45. *Kirk-Othmer,* **14** 1992, p. 157.

Indole

GERD COLLIN, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany (Chaps. 2–5)

HARTMUT HÖKE, Weinheim, Federal Republic of Germany (Chap. 6)

1. Introduction 1
2. Properties 1
3. Production 1
4. Derivatives 1
5. Uses 2
6. Toxicology 3
7. References 4

1. Introduction

Indole [*120-72-9*], 1-benzo[*b*]pyrrole, C_8H_7N, was discovered by A. v. BAEYER and C. A. KNOP in 1866 as the basic structure of the natural dye, indigo, from which it was obtained. In 1910, R. WEISSGERBER found indole in coal tar.

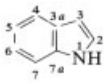

2. Properties

Physical Properties. Indole, M_r 117.15, *mp* 52.5 °C, *bp* 254.7 °C (101.3 kPa), *d* 1.22 g/cm^3 (18 °C), forms colorless, shiny flakes with a slight jasmine aroma. It is readily soluble in ethanol, diethyl ether, and benzene, soluble in hot water, slightly soluble in cold water, and volatile in steam. The heat of combustion is 3.650 kJ/kg (25 °C), the enthalpy of vaporization 597.5 kJ/kg (10.3 – 27.4 °C), and the dipole moment 2.11 D (benzene).

Chemical Properties. As a secondary amine, indole has a hydrogen atom which can be substituted by alkali metals. Oxidation leads to indigo, mild hydrogenation to 2,3-dihydroindole (indoline). Diindole, triindole, and resinification products are obtained upon treatment with acids.

3. Production

High-temperature coal tar contains on average just under 0.2 % indole. In coal tar distillation, the indole is concentrated in a biphenyl – indole fraction which boils between 245 and 255 °C. Following extraction of phenols and bases, the indole is isolated from this fraction and separated from the other major component biphenyl, whose boiling point is only 0.3 °C higher. This is achieved by melting with potassium hydroxide to give the potassium salt of indole, by azeotropic distillation with diethylene glycol, or by extraction with selective solvents (e.g., glycols, aqueous dimethyl sulfoxide, or monoethanolamine). The crude indole can then be purified by crystallization from aliphatic hydrocarbon solvents.

In addition to its recovery from coal tar, indole is also synthesized in technical quantities. Methods used include Madelung synthesis of formyltoluidine (from *o*-toluidine and formic acid), followed by cyclization [8–10], dehydrogenating cyclization of 2-ethylaniline [11–20], cyclocondensation of aniline and ethylene glycol in the liquid or gas phase [21–23], and cyclization of 2-(2-nitrophenyl)ethanol [24].

4. Derivatives

3-Methylindole [*83-34-1*], skatole, C_9H_9N, M_r 131.18, *mp* 95 °C, *bp* 266.4 °C (101.3 kPa), forms colorless flakes which are soluble in ethanol and slightly soluble in water. It can be recovered from coal tar or synthesized by Fischer reaction from the phenylhydrazone of propionaldehyde or, together with indole, by catalytic dehydration of 2-(2-aminophenyl)-1,3-propanediol [25].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_167

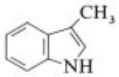

3-(Dimethylaminomethyl)indole [*87-52-5*], gramine, $C_{11}H_{14}N_2$, M_r 174.25, *mp* 138–139 °C, forms colorless needles. The compound is soluble in ethanol, diethyl ether, and chloroform and insoluble in water. It can be synthesized from indole by the Mannich reaction with dimethyl-amine and formaldehyde. Gramine is used to synthesize D,L-tryptophan, the plant growth regulators indole-3-acetic acid and indole-3-butanoic acid as well as the intermediate tryptamine.

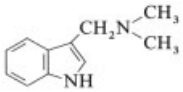

Indole-3-aldehyde [*487-89-8*], C_9H_7NO, M_r 145.16, *mp* 194 °C, forms colorless crystals. It is readily soluble in ethanol and slightly soluble in cold water. It can be produced by Vilsmeier reaction of indole with formylmethylaniline or with dimethylformamide and phosphorus oxychloride. Indole-3-aldehyde can be used to make D,L-tryptophan and cyanine dyes.

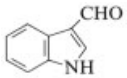

2,3-Dihydroindole [*495-15-1*], indoline C_8H_9N, M_r 119.16, *bp* 229–230 °C (101.3 kPa), is a colorless liquid, which is volatile in steam and soluble in diethyl ether, acetone, and benzene, but only slightly soluble in water. Indoline is obtained by hydrogenation of indole or by catalytic cyclodehydration of 2-(2-aminophenyl)ethanol. A range of pharmaceuticals, as well as fungicides and bactericides, can be produced from indoline.

3-(2-Aminoethyl)indole [*61-54-1*], tryptamine, $C_{10}H_{12}N_2$, M_r 160.22, *mp* 118 °C, forms colorless needles. It is readily soluble in ethanol and acetone, and slightly soluble in water, diethyl ether, chloroform, and benzene. It can be produced by reacting indole with ethylene imine, by catalytic hydrogenation of indole-3-acetonitrile, or by reacting phenylhydrazine with 4-aminobutyraldehyde diethyl acetal. Tryptamine is used to synthesize the vasodilator and antihypertensive, Vincamine.

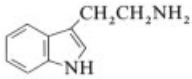

3-(2-Hydroxyethyl)indole [*526-55-6*], tryptophol, $C_{10}H_{11}NO$, M_r 161.20, *mp* 59 °C, *bp* 174 °C, (2 kPa) forms colorless prisms or flakes that are readily soluble in methanol, ethanol, diethyl ether, chloroform, and acetone, soluble in benzene, and slightly soluble in water. It is produced by reacting methyl-1*H*-indole-3-acetate with lithium aluminum hydride, or through reduction of methyl-1*H*-indole-3-glyoxylate with sodium borohydride. The antihypertensive Indoramin is produced from tryptophol.

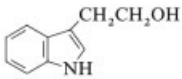

5. Uses

In line with its natural occurrence as a component of jasmine oil and orange-blossom oil, indole has been used for many years to fix fragrances and is therefore found in many perfumes [26].

A further important application is the production of the essential amino acid tryptophan (→ Amino Acids, Chap. 3.2.17.). Tryptophan can be synthesized from indole by chemical reactions as a racemate (via the indole derivatives gramine and indole-3-aldehyde). It is also produced by biotechnology in the enantiomerically pure L-form (e.g., from indole and L-serine).

Indole is also used as a feedstock in the synthesis of plant growth regulators, such as indole-3-acetic acid [*6250-86-8*] (heteroauxin) and

indole-3-butanoic acid [*3127-44-4*] (→ Plant Growth Regulators). Fungicidal and bactericidal plant protectives are synthesized from indoline [27].

Derivatives of indole-3-acetic acid find application as mild analgesics, e.g., indomethacin, (→ Anti-inflammatory – Antirheumatic Drugs). Indole and its derivatives are basic building blocks for many other pharmaceuticals. Thus, for example, indoline is used to synthesize analgesics [28] and antihypertensives [29], and tryptamine [*61-54-1*] [3-(2-aminoethyl)-indole] for the synthesis of the vasodilator and antihypertensive, Vincamine [30]. Tryptophol [*526-55-6*] [3-(2-hydroxyethyl)indole] is used to prepare the antihypertensive Indoramin [31] (→ Antihypertensives, Chap. 4.2.2.2.). (Pyridylalkyl)indoles can be produced from indole and can be used as antidepressants, antihistamines, and antihypertensives [32–34]. 3-Piperidinylindoles can be obtained via the Grignard reaction of 4-chloroindole with 3-piperidinone, and can be used in the prevention of anoxia [35]. Indole-2-carboxylic acid derivatives are useful, e.g., as antihypertensives [36–38]. 5-Chloroindole [*17422-32-1*] can be synthesized from indoline via the intermediates 1-acylindoline, 1-acyl-5-chloroindoline, and 5-chloroindoline [39]. It can be used as an intermediate in the production of 5-chlorotryptamine (for the production of tranquilizers and blood pressure lowering agents), as well as in the production of antidepressives, antiemetics, and drugs for the treatment of Parkinson's disease [40–42]. 2-Methylindole [*95-20-5*] (as the heterocyclic coupling component) is a starting material for commercially important cationic diazo dyes. Cyanine dyes are synthesized from 2-methylindole, 1-methyl-2-phenylindole, 2-methylindole-3-aldehyde, and indole-3-aldehyde (→ Methine Dyes and Pigments). In addition, indolylmethane dyes are produced from 1-methyl-2-phenylindole (→ Triarylmethane and Diarylmethane Dyes). Reacting indole with 3,6-bis-(2-carboxybenzoyl)carbazole affords carbazole phthalein dyes, which are used as optical filter agents in photographic emulsions [43].

6. Toxicology

Apart from contact during production and manufacture of indole, human exposure results from smoking and intestinal degradation of nutritional L-tryptophan. Indole is readily absorbed from the gastrointestinal tract and rapidly converted by the liver, e.g., to indican and oxindole, which are excreted mainly into the urine [44], [45]. The acute oral LD_{50} is 1000 – 1100 mg/kg in rats [46], [47] and about 500 mg/kg in mice [48]. Autopsy after acute exposure revealed hemorrhages and hyperemia of inner organs and tissues. Upon application to the skin and eyes of rabbits, slight temporary irritation was observed [49], [50]. Unlike 3-methylindole (skatole), indole does not induce pulmonary injury in cattle after repeated oral application [51]. At high chronic dosage (20 – 200 mg kg^{-1} d^{-1} orally and subcutaneously in mice, rats, rabbits, and dogs for more than three months), indole resulted in a shift in the blood cell differential picture (anemia, leukocytosis, or leukopenia); with persistently high intoxication, the development of leukemia cannot be ruled out [52–54]. In cattle, hemolysis and hemoglobinuric nephrosis were observed after several oral doses of indole (100 or 200 mg/kg) [51]. Although there is no other experimental evidence for its cancerogenic potential [55], indole may influence the formation of tumors caused by other agents. Thus, very high oral doses (800 mg kg^{-1} d^{-1}) have been reported to accelerate the development of tar-induced skin tumors in mice [53] and 2-acetylaminofluorene-induced bladder carcinomas in hamsters and rats [56]. On the other hand, it simultaneously seemed to suppress 2-acetylaminofluorene-induced hepato-carcinogenesis [56]. A genotoxic effect of indole was confirmed in the Ames test in combination with other mutagenic agents [57]. However, indole itself was neither mutagenic in the Ames test [57], [58], nor in a cell transformation test [59], although in a DNA repair assay with *Bacillus subtilis,* indole caused reparable DNA damage [59].

7. References

General References

1. *Beilstein,* **20** 304, **20(1)** 121, **20(2)** 196, **20(3)** 3176 – 3180.
2. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung,* Verlag Glückauf, Essen 1951, pp. 146 – 148, 151 – 152.
3. Rütgerswerke, *Erzeugnisse aus Steinkohlenteer,* 1958, Frankfurt, Federal Republic of Germany.
4. H.-G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Berlin 1968, pp. 60 – 61, 180 – 181.
5. H.-G. Franck, J. W. Stadelhofer: *Industrial Aromatic Chemistry,* Springer Verlag, Berlin 1988, pp. 417 – 418.
6. R. J. Sundberg: *The Chemistry of Indoles,* Academic Press, New York 1970.
7. W. J. Houlihan: *Indoles,* John Wiley & Sons, New York 1972.

Specific References

8. Mitsui Toatsu, JP 74 20 587, 1970 (R. Fujiwars, Y. Yamada, N. Kuroda).
9. Kyowa Fermentation Industry, JP 73 76 864, 1972 (T. Yamauchi, S. Yada, S. Kudo).
10. Ube Industries, JP 78 95 965, 1977 (M. Nakai, T. Komata, M. Oda).
11. Rütgerswerke, DE 2 224 556, 1972; DE 2 401 017, 1974 (G. Grigoleit, R. Oberkobusch, G. Collin).
12. Houdry, US 2 891 965, 1956 (S. E. Voltz, J. H. Krause, W. E. Erner).
13. Snam Progetti, DE 2 148 961, 1971 (M. M. Mauri, P. Moggi).
14. Teijin, JP 73 76 862, 1972 (T. Fukada, K. Tanaka).
15. E. I. du Pont de Nemours, US 2 409 676, 1944 (W. F. Grasham, W. M. Brunner).
16. Socony Mobil Oil, NL 6 505 911, 1965.
17. Lonza, DE 2 441 439, 1974 (C. O'Murchu).
18. Société Nationale Elf Aquitaine, DE 2 328 284, 1973 (M. Petinaux, J. Metzger, J.-P. Aune, H. Knoche).
19. Sumitomo Chemical, JP 7 831 660, 1976 (M. Tamura).
20. Mitsui Toatsu, JP 76 113 868, 1975 (K. Yamamoto, K. Nitta, S. Ichikawa).
21. Mitsui Toatsu, EP 86 239, 1982 (T. Honda, K. Terada).
22. Ube Industries, JP 83 128 371, 1982 (N.N.).
23. Keishitsu Ruibun Shinyoto Kaihatsu Gijutsu Kenkyu Kumiai, JP 86 44 862, 1984 (M. Imanari, T. Seto).
24. Bofors, DE 2 052 678, 1970 (J. M. Bakke, H. E. Heikmann).
25. Rütgerswerke, DE 2 328 330, 1973 (G. Grigoleit, R. Oberkobusch, G. Collin).
26. G. T. Walker, *Seifen öle Fette Wachse* **100** (1974) 375 – 377.
27. Pfizer, GB 1 394 373, 1975; GB 1 394 374, 1975 (R. J. Bass, R. C. Koch, H. C. Richards, J. E. Thorpe).
28. Hoechst-Roussel Pharmaceuticals, US 4 448 784 784, 1982 (E. J. Glamkowski, J. M. Fortunato).
29. Science Union, FR 2 003 311, 1968 (L. Beregi, P. Hugon, M. Laubie).
30. Roussel-UCLAF, DE 2 115 718, 1970 (J. Warnant, A. Farcilli, E. Toromnoff).
31. Wyeth, GB 1 399 608, 1971 (J. L. Archibald).
32. Pharmindustrie, EP 65 907, 1982 (F. Audiau, G. R. Le Fur).
33. Boehringer Ingelheim, EP 58 975, 1982 (K. Freter, V. Fuchs, J. T. Oliver).
34. Wyeth, EP 2 886, 1979 (J. L. Archibald, T. J. Ward).
35. Roussel-UCLAF, EP 21 924, 1981 (J. Guillaume, L. Nedelec, C. Dumont, R. Fournex).
36. Hoechst, DE 3 151 690, 1981 (H. Urbach, R. Henning, V. Teetz, R. Geiger, R. Becker).
37. McNeilab, US 4 529 724, 1983 (C. Y. Ho).
38. American Cyanamid, US 4 619 941, 1984 (W. B. Wright, J. B. Press).
39. Rütgerswerke, DE 3 035 403, 1980 (M. Maurer, W. Orth, W. Fickert).
40. MARPHA, DE 2 618 152, 1977 (A. Champseix, C. Gueremy, G. Le Fur).
41. Roussel-UCLAF, DE 2 719 294, 1977 (F. Clemence, D. Humbert).
42. Roussel-UCLAF, DE 2 738 646, 1978 (C. Dumont, J. Guillaume, L. Nedelec).
43. Polaroid, US 3 932 455, 1976 (R. C. Bilofsky, R. D. Cramer).
44. A. H. Beckett, D. M. Morton, *Biochem. Pharmacol.* **15** (1966) 937 – 946.
45. L. J. King, D. V. Parke, R. T. Williams, *Biochem. J.* **98** (1966) 266 – 277.
46. National Institute for Occupational Safety and Health (NIOSH): *Registry of Toxic Effects of Chemical Substances,* US Government Printing Office, Washington D.C. 1987, p. 12992.
47. Huntingdon Research Centre, Rütgerswerke Order No. 338 a, 1979, Münster, Federal Republic of Germany.

48. G. J. Martin, E. H. Rennebaum, M. R. Thompson, *Ann. Intern. Med.* **18** (1943) 57 – 71.
49. Huntingdon Research Centre, Rütgerswerke Order No. 338 b, 1979, Münster, Federal Republic of Germany.
50. Huntingdon Research Centre, Rütgerswerke Order No. 338 c, 1979, Münster, Federal Republic of Germany.
51. A. C. Hammond, J. R. Carlson, *Vet. Rec.* **107** (1980) 344 – 346.
52. W. T. Taylor, *Med. Ann. D. C.* **8** (1939) 362 – 363.
53. K. Kaiser, *Z. Krebsforsch.* **59** (1953) 448 – 495.
54. H. Ehrhart, W. Stich, *Klin. Wochenschr.* **35** (1957) 504 – 511.
55. M. J. Shear, J. Leiter, *J. Natl. Cancer. Inst. (US)* **2** (1941) 241 – 258.
56. M. Matsumoto, M. L. Hopp. R. Oyasu, *Invest. Urol.* **14** (1976) 206 – 209.
57. T. Matsumoto, D. Yoshida, S. Mizusaki, *Mutat. Res.* **56** (1977) 85 – 88.
58. M. Curvall, J. Florin, T. Jansson, *Toxicology* **23** (1982) 1 – 10.
59. T. Osawa, M. Namiki, K. Suzuki, T. Mitsuoka, *Mutat. Res.* **122** (1983) 299 – 304.

Industrial Hygiene → Occupational Health and Safety

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Information Storage Materials

Standard Article

Roman E. Wochele[1], Barry K. Middleton[2], Henk van Houten[3], Johan P. W. B. Duchateau[3], Hans J. G. Kloosterboer[3], Jan A. Th. Verhoeven[4], René van Vlimmeren[5], Peter E. J. Legierse[6], Dirk J. Gravesteijn[3], C. David Wright[7], Herman J. Borg[3], Heinrich Heitmann[8], Jacques Heemskerk[9]

[1]Gramm Oberflächentechnik, Ditzingen, Germany
[2]Manchester University, Manchester, United Kingdom
[3]Philips Research Laboratories, Eindhoven, The Netherlands
[4]Toolex International N.V., Veldhoven, The Netherlands
[5]Toolex International N.V., Veldhoven, The Netherlands
[6]Optical Disc Mastering (ODM), Eindhoven, The Netherlands
[7]Manchester School of Engineering, University of Manchester, United Kingdom
[8]Philips Research Laboratories, Hamburg, Federal Republic of Germany
[9]Philips, Eindhoven, The Netherlands

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a14_171.pub2
Article Online Posting Date: August 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Magnetic Recording
1.1. Introduction
1.1.1. History
1.1.2. Magnetism and Magnetic Materials: Definitions
1.2. Principles
1.2.1. Digital Record and Replay
1.2.2. Media Requirements for Digital Recording
1.2.3. a.c. Bias Audio Recording
1.2.4. Video Recording
1.2.5. Transports
1.2.6. Electronic Systems
1.3. Magnetic Recording Media
1.3.1. γ-Fe_2O_3
1.3.2. Chromium Dioxide
1.3.3. Cobalt-Modified γ-Fe_2O_3
1.3.4. Metal Particles
1.3.5. Metal-Evaporated Tape (MET)
1.3.6. Metal Film Disk Media
1.3.7. Perpendicular Recording
1.4. Recording Heads
1.5. Economic Aspects
2. Optical Recording
2.1. General Introduction
2.1.1. Optical Storage Media
2.1.2. The Optical Channel
2.1.2.1. Focusing and Tracking (Servo Control)
2.1.2.2. Data Channel
2.1.3. Compact Disc Family

2.1.3.1. Read-Only, Replicated Discs
2.1.3.2. Recordable and Rewritable Discs
2.1.4. Digital Versatile Disc (DVD)
2.1.5. HD-DVD and Blue Laser Discs
2.1.6. Magneto-Optical Storage
2.1.7. Future Developments
2.1.7.1. Evolutionary Progress in Optical Disc Storage
2.1.7.2. Three-Dimensional Optical Storage
2.2. Optical Disc Mastering and Stamper Production
2.2.1. General
2.2.2. Requirements
2.2.3. Mastering Methods
2.2.4. Electroforming of Stampers
2.3. Read-Only Optical Recording Materials
2.3.1. Production Processes
2.3.2. Material Specifications
2.3.3. Molding Processes
2.3.4. Photoreplication Process
2.3.5. Metallization
2.3.6. Protective Coating
2.3.7. Nonstandard Processes
2.4. Write-Once Optical Recording Materials
2.4.1. Material Specifications
2.4.2. Deposition Techniques
2.4.3. Tellurium Alloys
2.4.4. Organic Dyes
2.4.5. Metal Particle Colloids
2.4.6. Metal-Polymer Bilayers
2.4.7. Alloying Bilayers
2.4.8. Agglomeration
2.4.9. Surface Texture
2.5. Reversible Optical Recording Materials
2.5.1. Magneto-Optic Materials
2.5.2. Phase-Change Recording Media
2.5.3. Liquid Crystals
2.5.4. Photochromics
2.5.5. Photochemical Hole Burning
2.6. Economic Aspects of Optical Recording
2.6.1. Consumer Electronics (CE)
2.6.2. Data Applications
3. Acknowledgment

[Top of Page]

1. Magnetic Recording

Barry K. Middleton

1.1. Introduction

Magnetic recording technology is widespread in its adoption and its impact on our lives. Be it in portable audio recorders, hi-fi audio systems, video recorders, or in floppy or rigid disk systems in computers, the products of the technology are widely used and appreciated today. Novel data storage systems are being prepared and queuing up for possible launch into the market place. All these are propelled by mighty industrial concerns in various parts of the world primarily in America and Japan, but also in Europe.

The majority of audio and video recording has traditionally taken place using analogue techniques although in recent times products operating in the digital domain have been introduced. The future will see an accelerated shift to digital techniques and this means that there is a continued convergence of most recording technologies into the data storage arena where all information from whatever source can be recorded and manipulated using digital methods.

1.1.1. *History*

Magnetic recording [1] was first demonstrated by the Danish inventor VALDEMAR POULSEN in 1898. He developed a machine for recording voice messages onto steel wire and was able to demonstrate recording and replay in a recognizable form. This

achievement quickly received acclaim. Despite the low signal levels and high levels of distortion in the recordings, developments of this equipment were made and a telephone message recording machine called a Telegraphone was eventually put on the market. However the Telegraphone was not widely adopted, partly through its lack of clarity, and the full exploitation of the technology had to await the later developments of electronic technology and the means to reduce distortions.

A major step forward came in 1935 with the development in Germany by BASF and AEG of a machine called the Magnetophone. This machine, and its later variants, brought into play two new developments which remain today as the core of many magnetic recording systems. The first was the introduction of magnetic particles as the recording medium and the second was the use of a.c. bias to linearize the characteristics of the recording medium. With these developments record and replay with high clarity was achieved and it led eventually, with further progress, to the widespread adoption of a.c. biased recording systems. The most popular current manifestation of the technology is in the compact cassette audio system introduced by Philips in 1963.

Digital recording emerged during the late 1940s and 1950s in response to the need for back-up storage for computer devices. By these times magnetic tape recording for audio was in widespread use and the technology was adapted for the recording of digital waveforms without the use of a.c. bias. In the early machines bit packing densities were very low but this increased rapidly with successive generations of equipment. Record and replay took place at data rates which were slow compared with the processing rates of computers and access times to information became a critical issue. This led to the development of faster access mechanical systems such as the magnetic drum and later the very successful magnetic disk. The latter is now the mainstay of storage for computers and disk systems exist in a variety of forms and formats employing rigid and flexible disk media.

Video recording first took place in “fixed head” systems similar to those used in audio and data recording. However, the wide bandwidths required by video signals meant that very high tape speeds had to be used and this required very long tapes for program length recordings. To alleviate this problem scanning head technology was developed in which tracks were recorded very close together in directions across the tape length to give very high area packing densities of information. First came the transverse scan technology introduced by Ampex in 1956 and later the helical scan technologies were introduced for domestic use and are now most widely manifested in VHS recorders [2]. Helical scan technology has now been adapted for use in data storage systems especially for computer back-up applications.

Recording technology continues to move forward and new developments invariably break fresh ground in terms of the amount of data stored per unit area of recording surface. Here comes the pressure for new and advanced recording media.

1.1.2. *Magnetism and Magnetic Materials: Definitions (* ⟶ Magnetic Materials*)*

For a more detailed description of the terms used in the specification of magnetic materials, particularly those used in recording media and in magnetic heads, see, e.g., [3].

The magnetic properties of materials are often characterized by the *magnetic induction* or *flux density* within them. This quantity is represented by the symbol *B* and measured in units of tesla (T). In free space *B* is given by

$$B = \mu H0 \tag{1}$$

where $\mu_0 = 4\pi \cdot 10^{-7}$ H/m is the permeability of free space and *H* is the magnetic field strength measured in A/m. The magnetic field is usually generated by a coil carrying a current and the magnitude of the field depends on that current and the geometry of the coil hence leading to a field strength which is expressed in A/m.

When the induction is measured inside a magnetic material it may be represented by

$$B = \mu(H + M)0 \tag{2}$$

where *H* is the applied field strength and *M* is an additional contribution, measured in A/m, arising from the magnetic material and which is usually referred to as its magnetization. Sometimes Equation (2) is written in a modified form

$$B = \mu H + J0 \tag{3}$$

where *J* (sometimes the symbol *I* is used) is also referred to as the magnetization but it is now in units of Tesla. Both representations of magnetization appear in the literature but in this article both *M* and *H* will be quoted in the same units of A/m. This is the approach adopted in most publications although there is still a widespread use in many parts of industry and academia of cgs and some imperial units.

In ferromagnetic and ferrimagnetic materials the relationship between the magnetization produced in a material and the applied magnetic field is usually nonlinear. A typical relationship is shown in Figure 1. Starting with a demagnetized sample, that is one with zero net magnetization, the application of a magnetic field increasing from zero takes the magnetization along the curve a. First of all the magnetization rises very slowly, but as the field increases the rate of rise increases. Eventually the magnetization approaches a value known as the saturation magnetization M_s. This curve is known as the initial magnetization curve. On removal of the field the magnetization drops back to the value M_r, known as the remanent magnetization. On increasing the field in a negative direction the magnetization eventually starts to reduce and is brought to zero at a field labeled $-H_c$ which is the negative value of the coercivity. Further increase of this field in the negative direction eventually takes the magnetization through to its negative saturation value of $-M_s$. Bringing the field back to zero and then increasing it again in a positive direction traces out a curve with the positive going arrows, which is a mirror image of that produced by negative going fields, and brings the material back to the positively saturated state. The curve just traversed is known as the major hysteresis loop and the quantities coercivity H_c, remanent magnetization M_r, saturation magnetization *M*

$_s$, and hysteresis and loop squareness M_r/M_s are all properties very important in characterizing materials and especially those used in magnetic recording.

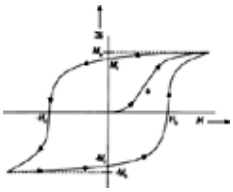

Figure 1. Initial magnetization curve (a) and major hysteresis loop of a typical magnetic material.

Sometimes magnetic materials are quoted as being magnetically hard or magnetically soft. *Magnetically hard materials* (⟶ Magnetic Materials – Parameters of Hard Magnetic Materials) are those in which very large magnetic fields have to be applied to change their states of magnetization. These types of materials have high coercivities and are permanent magnet materials. Magnetic recording media must be made of hard magnetic materials so they retain their recorded magnetizations permanently until they are purposely overwritten with new information or are erased. *Soft magnetic materials* (⟶ Magnetic Materials – Parameters of Soft Magnetic Materials) are those with extremely low coercivities. In this case the application of only small magnetic fields is sufficient to change their magnetizations. Such materials are used in magnetic recording heads where small currents flowing in the coils of the heads are sufficient to change their states of magnetization and therefore change the fields produced by them.

1.2. Principles

1.2.1. *Digital Record and Replay*

Figure 2 shows, in diagrammatic form, a simple record and replay system. It consists of a magnetic recording medium moving from left to right firstly under a record head and then under a replay head. The medium is made of permanent magnet material so that once a pattern is recorded into it it remains permanently magnetized, while the record and replay heads are made of soft magnetic material in which the magnetizations are easily alterable. The heads are of the ring type (see Section Recording Heads), so named because of their open shape, and they are designed to produce a high magnetic field in the recording medium sufficient to alter its magnetization. The replay head senses the magnetic fields arising from the magnetized medium as it passes underneath.

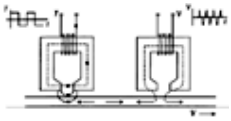

Figure 2. The digital record and replay processes. See text for description.

Recording takes place as follows. A current *I* flowing in the record head in the direction shown produces a field in the coil of the head which magnetizes the core of the head as shown. This produces a very high magnetic field in the gap of the head while below the head gap the field leaks out and penetrates the recording medium. With a good design of the pole region of the head this leakage or fringe field is sufficiently high to magnetize the recording medium. As shown in Figure 2 the field produced under the head is directed to the left and while this field remains in place the medium that is transported through it is magnetized to the left. A reversal of the record current to the opposite sense (direction) produces a magnetic field under the record head which is directed to the right. This then leads to the magnetization of the medium also to the right. Hence an alternating polarity of record current leads to an alternating polarity of magnetization along the tape.

While it is possible to change the sense of the record current extremely quickly with respect to time, the magnetization needs some distance along the tape over which its sense can be changed. This imposes limitations on the number of reversals or changes of magnetization that can be recorded per unit length of medium. These changes of magnetization, or transitions, are used in recording to carry the bits of information and so the number of these transitions per unit length is often referred to as the bit packing density. Obviously an advanced and good recording tape will have transitions which are extremely narrow so that a large number of them can be recorded per unit length of medium. Much of modern recording technology centers on developing media which can sustain narrow transitions and heads which are able to record them.

The replay process takes place as follows. The replay head is sensitive to the magnetization in the recording medium as it passes under its gap. The head produces in its magnetic circuit a flux which is related to the flux in the recording medium. This flux passes through the coil of the head and its rate of change gives the output voltage. The latter is therefore related to the rate of change of flux in the medium below the head gap and so only changes of magnetization in the medium produce an output. The input current and the corresponding output voltage are shown in Figure 2. Figure 3 top, again shows the output voltage as a function of time for transitions widely spaced on the medium. The narrower the recorded transitions the higher and narrower are the replayed pulses. The converse is true in that the wider the transitions the smaller and wider are the pulses.

Figure 3. Output waveforms at different bit packing densities.

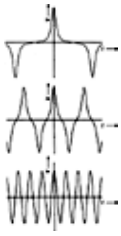

Obviously it is desirable to record as much information as possible per unit length of recording medium. When this happens the replay pulses are moved closer together and eventually they overlap. This is illustrated in Figure 3 where in subsequent diagrams the packing density of information increases. It can easily be seen that as pulses overlap they detract from one another and the output amplitude is reduced. Thus, the price of recording at high densities is a loss of signal amplitude, and, as with all electronic systems, ultimately performance is limited by the signal-to-noise ratio and so the reducing signal level will determine the performance of the recording system.

Figure 4 shows a typical curve of output voltage as a function of packing density which in this case is for an advanced thin-film disk recording medium. The recording densities involved are high and the transition widths narrow. A high performing recording medium will have narrow transitions recorded in it and output voltage will fall off slowly with increasing packing density. A lower performing recording medium will have wider transitions recorded in it and its output will fall off quickly as packing density increases. More detailed discussions of this are in the literature (see, e.g., [4]) but more will also be said about this in later sections.

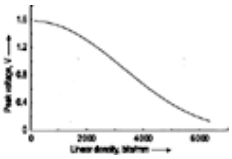

Figure 4. Output voltage amplitude (after amplification) as a function of packing density for a thin film disk recording medium of thickness 30 nm, coercivity 175 kA/m, head gap 0.25 µm, head to medium spacing 0.06 µm, and estimated transition width parameter of 0.095 µm.

1.2.2. *Media Requirements for Digital Recording*

There are two aspects of recording with which the recording media must be adequately designed to cope [4]. The first is their ability to have recorded within them narrow transitions to facilitate small bit sizes, and thus high bit packing densities, and the second is to retain those transitions once they have been recorded.

A transition is created in a recording medium when the current in the record head and therefore the field under it is reversed (see Fig. 2). This field is very large immediately under the center of the gap and falls away with distance to the right because this side of the head produces the recorded transition. The new magnetization under the gap likewise gives way to the previously recorded magnetization to the right of the head gap. The head field is therefore decaying rapidly with distance within the transition and the magnetization gradient is given by

$$\frac{\mathrm{d}M}{\mathrm{d}X} = \frac{\mathrm{d}M}{\mathrm{d}H}\frac{\mathrm{d}H}{\mathrm{d}x} \tag{4}$$

where (dH/dx) is the head field gradient and (dM/dH) is the variation of magnetization with applied field. The center of a recorded transition is at $M = 0$ and according to the hysteresis loop of Figure 1 this is where $H = H_c$ and the slope of the hysteresis loop is highest. For narrowest transitions high hysteresis loop squarenesses are required of the media and high field gradients are required of the record heads.

However the treatment involving only head fields is a simplification. Changes of magnetization in a magnetic medium produce stray (or demagnetizing) fields which depend on the magnitude of the magnetization recorded in the media and also on the thicknesses of the media. These add to the head fields and together they shape the recording process as indicated by Equation (4) (see, e.g., [4-6]). Taking these into account and assuming the magnetization distribution along the medium through a transition is of the form [7]

$$M = -\frac{2M_r}{\pi}\arctan(x/a) \tag{5}$$

where a is a parameter proportional to the transition width, as shown in Figure 5. It can be shown that in a thin recording medium of thickness D and with a rectangular hysteresis loop [4]

$$a = \left[\frac{M_r D d}{\pi H_c}\right]^{1/2} \tag{6}$$

where d is the record head to medium spacing. However, when the recording medium is unable to retain such narrow recorded transitions as predicted by Equation (6) because the self demagnetizing fields in the transitions are high, there is a minimum value for transition width determined by the fact that the demagnetizing fields should not exceed the coercivity of the medium [8]. Then [9]

$$a = \frac{M_r D}{2\pi H_c} \tag{7}$$

Thus, the write process in a thin film medium will produce a transition of width given by Equation (6) provided the width is larger than the minimum which self demagnetizing fields allow according to Equation (7). Thus the resultant transition width takes the larger of the values given by Equations (6) or (7). Either way the controlling factor is the ratio ($M_r D/H_c$) and a small value is required for narrow transitions and high density recording.

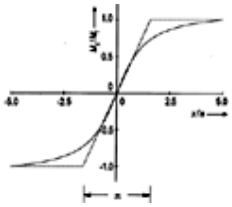

Figure 5. Assumed magnetization distribution as a function of distance through a transition.

The tangent at the origin shows the overall transition width to be πa.

Media suited for narrow transitions and high density (short wavelength) recording therefore need to be thin and have a high coercivity. These requirements are put alongside that of high hysteresis loop squareness and together they have persisted as the guiding requirements for advanced new media.

1.2.3. *a.c. Bias Audio Recording*

In audio recording the magnetization level which is induced in the tape must be linearly related to the recording signal. This is achieved by using a.c. bias recording [4], [10]. The principle of a.c. bias recording is illustrated in Figure 6 which shows graphs of magnetization remaining in a tape after the application and removal of magnetic fields H (remanence curves). Take the curve labeled with a 0 where the field has been applied to a previously demagnetized specimen of tape and the relationship between the magnetization and field is clearly not linear. Therefore a signal field applied by the recording head would not induce in the medium a magnetization directly proportional to it. However if an a.c. bias field H_{ac} is added to the applied field H and taken away in step with the applied field, the relationship between magnetization and field is altered according to the amount of a.c. bias added. Figure 6 shows curves for a number of values of a.c. bias field and it is clear that increasing the a.c. bias removes the non-linearities of the magnetization curve eventually producing, at low d.c. fields, the required linear relationship. The use of an a.c. bias field in conjunction with the applied field leads to a set of curves which are known as anhysteretic remanence curves. Where the field and bias reduce in unison, as in the recording process, they are known as modified remanence curves, whereas the “standard” anhysteretic remanence curves are obtained by reducing the bias field to zero before the d.c. field is removed [11], [12].

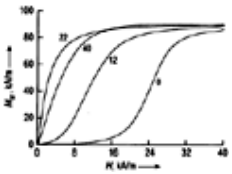

Figure 6. Modified remanence curves for γ-Fe_2O_3 recording tapes. The numbers alongside each curve indicate the amount of a.c. bias.

In the recording situation the applied field represents the signal which is at low frequency and to this is added a much higher frequency a.c. bias field. For example for audio signals of up to 20 kHz bandwidth the a.c. bias added must have a frequency at the very least of 65 kHz. The sum of the bias and the signal is passed through the record head and this then creates a magnetization in the medium which is linearly related to the record signal. The role of the bias field is purely to linearize the relationship between magnetization and signal field and is itself not recorded on the recording medium.

The requirements for the recording process are therefore for high linearity of the relationship between applied field and magnetization, but for retention of the recorded pattern at short wavelengths the tape must be thin to keep demagnetizing fields low and the coercivity must be high to resist demagnetization. Thus the requirements for good short wave length audio recording mirror those for high bit density digital recording.

The output voltage generated in the replay head is, as before, related to the magnetization gradient recorded in the medium. For a record signal which takes the form of a sine wave the differential is a cosine whose amplitude increases linearly with frequency. The output from an ideal recorder therefore has an amplitude which rises linearly with frequency but has suffered a 90 ° phase shift from the record signal. This linear curve is something of an idealization because at high frequencies losses in the replay process come into play [12-14]. These are caused by the finite spacing between the head and the medium, the finite thickness of the medium, and the finite width of the gap. They conspire to reduce the output at high signal frequencies as shown in Figure 7. Therefore the overall characteristic of a recording system is of a bandpass nature with no output at d.c., very little signal at low frequencies, and also little output at high frequencies due to the replay losses.

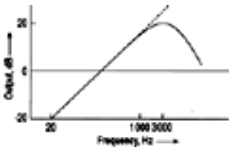

Figure 7. Output as a function of frequency for an audio recording tape.

1.2.4. *Video Recording*

Video recording takes place as an analogue process but does not use a.c. bias. It achieves its linearity by the use of

frequency modulation in which the video signal is applied to a high frequency carrier wave so that the instantaneous frequency of the modulated waveform is linear with the instantaneous magnitude of the video signal [2], [13]. It is therefore the variations in frequency, and thus the spacings between the crossovers in the waveform, which carry the video information. In the recording machine the frequency modulated waveform is applied directly to the record head. Despite the distortions in unbiased recording the signal, represented by the spacings between the crossovers, is faithfully recorded and so after replay and demodulation the video information can be recovered [2], [13].

The overall characteristics of the recorder in terms of transfer characteristics are much the same as described in Section a.c. Bias Audio Recording, and shown in Figure 7, except that the bandwidths are now much wider and reach out to several MHz. The requirements of the recording medium are likewise similar, i.e., thinness and high coercivity.

1.2.5. *Transports*

Audio recording systems use heads which are fixed in position and tapes which move past and in contact with them. With a tape speed v of 4.76 cm/s a length of tape of approximately 172 m would be needed for 1 h playing time. Such lengths of tape are accommodated in a spool of small diameter which can be carried in a cassette. The same tape with a recording bandwidth of 15 kHz would at the highest frequency have recorded a wavelength on the tape which can be shown (using $v = f\ \lambda$) to be 3.2 µm.

In a video recording system [2] where, say, a 5 MHz bandwidth is required and where it is assumed that recording can take place down to 1 µm wavelength a head to medium speed of 5 m/s would be required. With a fixed head recording system this would then mean that 18 000 m of tape would be needed for 1 h playing time and this is clearly impractical. A different type of transport system is necessary and the first approach to a practical alternative was delivered in 1956 by the introduction of the transverse scan system developed by Ampex Corporation in the USA [2], [13]. In this system a tape moves at very low speed past and in contact with the circumference of a spinning wheel on which are mounted four heads equally spaced (see Fig. 8). Each of those heads in turn passes transversely across the tape at high speed, sufficient to record at the highest frequency and to create tracks of recorded information across the tape. With the tape moving at a low speed it moves only a short distance before the next head starts its traverse across the tape. In that way tracks are recorded extremely close together and the amount of information per unit area of tape is very high. By packing the information this way a low usage of tape results, because of the low tape speed, but a high bandwidth operation, because of the high head to medium speed.

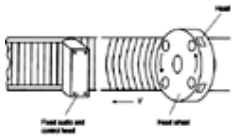

Figure 8. Simplified picture of a transverse scan tape transport.

Transverse scan technology has today been superseded by helical scan technology as depicted in Figure 9. This is employed commonly in a wide range of video recorders and in the VHS domestic system. Here a rotating drum, which replaces the wheel of the transverse scan system, has two heads on its circumference and spins at high speed. The axis of the rotation of the drum is no longer parallel to the tape direction but has been rotated through an angle such that the recorded track is now diagonally across the tape. The tracks are now several centimeters long and a single head scan along its track is sufficient to record or reproduce one frame of a television picture. Again there is a high head to medium speed and a high track density to enable a high area density of information, a low tape speed and low tape consumption. These helical scan systems work extremely efficiently and are now in widespread usage for video, digital audio, and digital data storage systems.

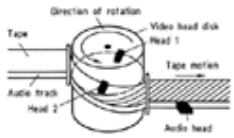

Figure 9. Simplified picture of a helical scan tape transport.

For computer memory applications floppy disk and rigid disk systems are the norm. A transport consists of a disk which is rotated at high speed on a suitable spindle and a mechanism for moving heads above appropriate tracks. In rigid disk systems a head cantilever is used which is pivoted, like a gramophone pickup arm, and sweeps in an arc across the disk surface to the appropriate track. In floppy disk systems the head is positioned on a cantilever which is moved radially across the disk to the appropriate track.

1.2.6. *Electronic Systems*

As shown in Section a.c. Bias Audio Recording a sine wave record current of constant amplitude gives rise to an output current which varies very strongly as a function of frequency and which also has a phase shift of 90 °. Thus, the magnetic recording channel introduced distortions which, though they are linear, give an output which is not recognizable immediately as a replica of the input. To correct for this distortion, equalizers are employed which take the output from the replay head and filter it to a form which is a replica of the input. In an audio recorder, and usually in a video recorder, the equalizer has characteristics which are the exact inverse of the recording characteristics and, since the decrease of output at high frequencies depends on the tape and recording characteristics, they compensate for the limitations introduced by them. In addition to this amplitude correction also a phase correction of 90 ° is needed. Practical equalizers are associated with circuits having particular time constants which are specified in the literature [15]. These types of equalizers are referred to as *flat equalizers* since they make the overall gain of the system flat, i.e., independent of frequency.

In digital recorders flat equalizers have also been widely used but they are just one of the equalization strategies which are

available [13]. In digital recording the requirement is not necessarily that the output signal from the equalizer is a replica of the input, but that from the output from the equalizer makes it possible to detect whether a digit 0 or digit 1 had been recorded. Thus the role of the equalizer is to render visible the recorded pattern in such a form that the data originally recorded can be easily recovered. A number of equalizer strategies are available for use in digital recording systems. These include peak detection, Nyquist filtering, and class IV partial response equalization. In all cases the equalizer compensates in some way for the distortions introduced to the signal by the recording properties of the recording medium.

1.3. Magnetic Recording Media

A typical recording tape consists of a dispersion of magnetic fine particles in a nonmagnetic binder which is spread onto a plastic tape backing, as shown very diagrammatically in Figure 10. The fine particles are usually less than 0.5 µm in length which is an essential requirement for creating the high coercivity that is needed for magnetic recording media. The dispersion of the particles is necessary to avoid their close proximities and the resultant effect of interparticle interactions and to achieve a uniform coating. Figure 11 shows dispersions of three different types of particle and demonstrates how variable the sizes, shapes, and orientations of particles can be. Despite these variations, tapes with high uniformity and with the required magnetic hysteresis properties and recording performances can be obtained.

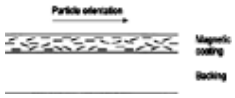

Figure 10. Simplified diagram of a magnetic recording tape.

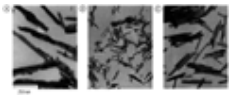

Figure 11. Transmission electron micrographs of magnetic pigment particles

A) Cobalt-modified γ-Fe_2O_3; B) Passivated metallic iron; C) CrO_2

Fine particles are used because their dimensions are sufficiently small not to have room for domain walls and so they contain a single domain. Thus, domain wall motion, which takes place at low applied fields, is avoided and so are low coercivities. But, if the particles have either shape anisotropy or a high uniaxial crystal anisotropy [16] (Fig. 12), the magnetizations are in stable orientations when they are oriented along either senses of the easy axes. Consequently only the application of very high magnetic fields in the opposite directions to those of the magnetizations will cause the magnetizations to change their polarities from one sense to the other [16-18]. A magnetization which exists in only one of two stable states gives rise to a rectangular hysteresis loop of high coercivity and when particles are packed in a binder with their easy axes aligned the resultant tape has high coercivity and high hysteresis loop squareness. A detailed description of the properties of magnetic tapes is given in [16].

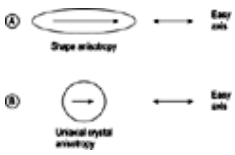

Figure 12. Magnetic fine particles with A) shape anisotropy, B) uniaxial crystal anisotropy

In the production of magnetic tape the particles are mixed into a slurry and then milled for an extended period to separate the particles and break down any agglomerates. The slurry is provided with various agents and chemicals for wetting and lubrication, and with solvents. The slurry is then applied in thin layers to a plastic backing using coating machines with either gravure roles, doctor blades, extrusion slips, or reversed role systems. While the coating is still wet it is passed through a magnet assembly which aligns the easy axes of the particles along the coating direction, and it then passes through an oven to drive off the solvents and solidify the tape. After removal of the solvents the coating on the tape contains close to 50 vol % magnetic particles (Fig. 10). The tape may also have a coating on its reverse side which contains carbon to reduce static electricity and/or friction, and sometimes carbon may be included in the magnetic coating itself to reduce friction between tape and the record head.

The final coated tape often has a rough surface with particles protruding from it. To improve the surface finish the tapes are passed between rollers to flatten their surfaces, a process called calendering. When a coating has been produced for use in floppy disks their surfaces may well be further improved by polishing using a ceramic polisher.

1.3.1. γ-Fe_2O_3

γ-Ferric oxide recording tape is easily recognized by its red brown appearance and is widely used as a standard grade audio tape. In addition, it has also been used in digital tapes for computer and instrumentation applications, and also in floppy and rigid disks. As a recording material it dates back to the late 1930s and for most, if not all, of that period it has been the mainstay of tape recording. Nowadays it is no longer considered an advanced medium. It is giving way to other materials in the latest applications which demand higher density recording performances.

γ-Ferric oxide particles can be produced by a number of methods. Usually they involve chemical precipitation processes to produce, first of all, needle shaped particles of α-FeOOH which are then dehydrated to α-Fe_2O_3, reduced to Fe_3O_4 and finally reoxidized to γ-Fe_2O_3. The process steps require careful control to ensure the appropriate morphology of the particles, and that they have the required needle shaped uniaxial anisotropy for high coercivity. This is particularly important since γ-Fe_2O_3 has cubic anisotropy which would result in coercivities which are too low. Particle sizes are typically less than 0.5 µm.

Properties of γ-Fe_2O_3 are shown in Table 1 where typical coercivities are seen to be in the range of approximately 25 – 30 kA/m.

Table 1. Magnetic properties of tapes and particles

Material	Coercivity H_c, kA/m	Tape coating thickness, µm	Remanent magnetization coating, kA/m	Remanent magnetization bulk material, kA/m	Particle lengths, µm
γ-Fe_2O_3	25 – 35	5 – 8	80 – 130	375	0.25 – 1
CrO_2	40 – 60	3 – 5	100 – 140	480	0.2 – 0.4
Co- γ-Fe_2O_3	50 – 75	3 – 6	100 – 140		0.3 – 0.5
Fe, FeCo	100 – 200	0.2 – 3	200 – 300	1700 – 1900	0.2 – 0.5
MET *	80 – 160	0.1 – 0.15	400 – 500		
CoP	60 – 120	0.5 – 0.1	300 – 800		
Co, CoNi, CoNiCr	60 – 200	0.02 – 0.05	300 – 800		
CoCr (perpendicular)	60 – 200	–1	300 – 800		

* MET = Metal-evaporated tape.

1.3.2. *Chromium Dioxide*

Chromium dioxide particles were introduced as recording media in the late 1960s. The fabrication processes produce particles which are almost perfect crystalline whiskers of submicron dimensions and high axial ratios. They have higher coercivities than γ-Fe_2O_3 particles and because of their smoothness can easily be oriented within the tapes to produce media with high coercivities and high hysteresis loop squarenesses. Table 1 gives some properties.

CrO_2 is produced by reacting CrO_3 and Cr_2O_3 in an aqueous medium and at high temperature and pressure. The CrO_2 particles are formed in parallel bundles and their shapes and sizes are influenced by the presence of additives during the preparation process. Some passivation of the particles is needed because of their chemical reactivity. Typical particles are less than 0.5 µm long, have axial ratios of 10 or more, and have coercivities in the range 40 – 60 kA/m. Figure 11 C is a micrograph of typical CrO_2 particles.

Since tapes produced using chromium dioxide particles have higher coercivities than tapes using γ-ferric oxide they are capable of superior recording performances. They are used in higher grade audio tapes, video tapes (VHS) and some computer tape applications. They are more expensive than γ-ferric oxide particles and to justify their premium price their higher performances are a requisite.

1.3.3. *Cobalt-Modified γ-Fe_2O_3*

The coercivity of γ-Fe_2O_3 is too low for modern high density recording needs and increases in its coercivity have been obtained by modification of γ-Fe_2O_3 with cobalt. In *cobalt-substituted iron oxide* some of the iron atoms are replaced by cobalt atoms on appropriate atomic sites. This is achieved by introducing cobalt in a suitable form during the preparation process for γ-Fe_2O_3 and allowing the various heating processes to diffuse the cobalt into the γ-Fe_2O_3. The presence of cobalt results in higher crystalline anisotropy and ultimately a higher coercivity. However, the coercivities of these particles depend on temperature and on time. These dependencies are undesirable. In *cobalt-adsorbed γ-Fe_2O_3* layers of cobalt are produced on the surfaces of the γ-Fe_2O_3 and without heat treatment the cobalt remains on the surfaces as a coating. This increases the anisotropies of the particles to produce increased coercivities. The cobalt-adsorbed particles are found to have a higher thermal stability than the cobalt-substituted ones and have been most widely used. Typical properties for tapes prepared using these particles are shown in Table 1. A micrograph of some particles is shown in Figure 11 A. These tapes have found widespread application in video recording systems (VHS) and also in audio tapes, digital tapes, and floppy disks.

1.3.4. *Metal Particles*

Particles of metallic iron are formed by the reduction of γ-Fe_2O_3 with hydrogen. This yields particles of pure iron which are extremely reactive and prone to spontaneous ignition and which need some passivation before they can be used. The passivation is usually achieved through a surface oxidation to produce very thin protective layers of oxides of iron. Inclusion of the particles into tape occurs by a similar process to that used with γ-Fe_2O_3 although some special problems exist with the dispersion of the metallic particles because of their tendency to cluster together due to high magnetostatic interactions. Nevertheless very high coercivity tapes have been produced with good uniformity and very high recording performances.

Many different particle tapes have been produced which have different proportions of nickel or cobalt included in addition to the iron. Only a few percent of nickel but more than 25 % of cobalt may be included. Some properties are shown in Table 1.

Besides benefiting from very high coercivities metal particle tapes also benefit from reductions in their thicknesses. The most

spectacular improvement has come via the dual layer coating process in which a thin layer of magnetic particles is coated on the backing at the same time as an underlayer of TiO_2 particles [19]. Layers of less than 0.5 µm thickness have been achieved and this in combination with the high coercivity results in very high density recording performances which are comparable with those achieved by metal film media.

Metal particle tapes are intended for use in the latest high performance digital recording systems. Digital audio, digital video, and digital data recording are all relevant application areas where high density performance is a requisite.

1.3.5. *Metal-Evaporated Tape (MET)*

Thin metal film recording media have been produced by evaporation of cobalt or cobalt nickel alloy (typically $Co_{80}Ni_{20}$) in an oxygen atmosphere and subsequent deposition onto plastic substrates. Deposition is usually at an angle of incidence other than normal to the planes of the substrates (see Fig. 13). The films produced by this method are granular and the grains have been shown to be columnar and inclined to the normal to the film planes [20]. The grain boundaries, which contain metallic oxides, appear to create discontinuities between the grains and disrupt exchange coupling but also introduce significant shape anisotropy in the grains. The crystalline easy axes are usually along the directions of the grains and these, along with the shape anisotropies, cause the magnetizations to tip out of the tape planes and towards the axes of the grains.

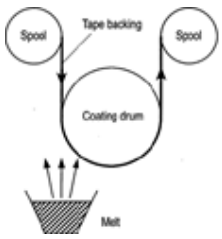

Figure 13. Diagrammatic view of the vacuum evaporation process for producing metal evaporated tape.

The tipping of the magnetization out of the media planes results in recording properties which are slightly different from typical longitudinal recording media [21]. The recording properties also depend on which way the media are moved passed the recording heads. For example the granular structure, illustrated in a rather idealized way in Figure 14, moving from left to right through a head field would give rise to slightly different recording properties if the tape were reversed and moved through the same head field. Improved performances are observed with the orientation shown in the right-hand diagram. However, the recording properties experienced as a result of the movement of the tapes in the two ways are only slightly different and they remain essentially longitudinal in character [21]. Some properties of the tapes are given in Table 1.

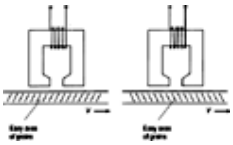

Figure 14. The passage of a metal evaporated tape past a record head

The two pictures show the tape being moved in such a way as to produce different recording geometries.

The different recording characteristics for different directions of tape motion have been circumvented by the production of two-layer and three-layer tapes as illustrated diagrammatically in Figure 15. These have been produced by sequential depositions of the metals onto substrates moving in different directions for each subsequent layer. The increased production difficulties are obviously undesirable and the three-layer tape is no longer produced as the two-layer tape gives rise to similar recording properties for both directions of tape motion.

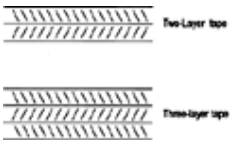

Figure 15. Representation of two-layer and three-layer metal evaporated tapes.

Metal films have suffered from a number of mechanical problems especially crazing of the coated films and low wear resistance [22]. Wear resistance has been vastly improved by growth induced surface features (asperities/nodules) and by the use of either diamond-like or graphite coatings on their wearing surfaces.

1.3.6. *Metal Film Disk Media*

Early disk recording machines employed γ-Fe_2O_3 media but these have now been exclusively superseded by thin-film media. The latter are used partly because of their thinness but also because of the very high coercivities that are available.

The first thin films to come into use were produced by the electroless chemical deposition of cobalt phosphorus onto nickel phosphorus-coated aluminum alloy substrates [20], [23]. The small percentage of phosphorus, up to about 5 %, is almost insoluble in the depositing cobalt and so is pushed into the grain boundaries. This nonmagnetic layer between the grains prevents exchange coupling across the boundaries and isolates them from one another (see Fig. 16). The result is that the films behave like planar arrays of closely packed single domain particles and this leads to high coercivities. Film thicknesses were of around 50 nm and coercivities were up to 100 kA/m and more.

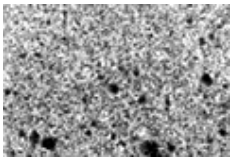

Figure 16. Grain structure of a CoP film of thickness 90 nm and coercivity 60 kA/m. The grains are around 40 nm in diameter.

The chemical deposition process has now given way to film deposition by sputtering. Cobalt films and cobalt – nickel films have been prepared this way by deposition onto aluminum alloy or glass substrates which have been previously coated with a chromium underlayer. The chromium of the underlayer permeates the magnetic coating and provides nonmagnetic separating layers between the grains [24]. This again ensures discontinuity of the magnetic structure causing the grains to be effectively separate particles and this results in high coercivity. This process has been supported by the inclusion of small quantities of metals such as platinum or tantalum [25]. Film thicknesses are now down to less than 30 nm and coercivities are up to around 200 kA/m. These properties are sufficient to allow extremely high densities of recording where linear densities of over 6000 bits/mm and area densities of over 15 Mbits/mm^2 have been demonstrated.

1.3.7. *Perpendicular Recording*

In perpendicular recording the recording medium is magnetized alternately in positive and negative directions normal to the medium plane [26] as shown in Figure 17, contrary to the situation in longitudinal recording (see Fig. 2). To achieve this orientation of magnetization a recording medium is needed which has an easy axis pointing in a direction normal to its plane and there need to be heads which produce large magnetic field components in the same direction. Such an arrangement has been achieved in practice with the use of high anisotropy media and it has been shown that perpendicular recording can sustain very narrow recorded transitions and extremely high linear packing densities [26]. Despite this success it has proved difficult to bring the technology into widespread and common usage.

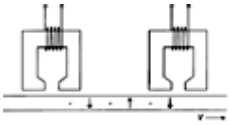

Figure 17. Perpendicular magnetic recording on a medium with an easy axis directed normal to its plane.

The materials used for perpendicular recording are invariably cobalt – chromium alloys with cobalt : chromium ratios around 80 : 20 [27]. The recording layers are prepared by sputtering or evaporation, have thicknesses in the region 0.1 to 1 µm and coercivities in the perpendicular direction of up to 200 kA/m.

The chromium is included with the cobalt to produce a lower magnetization so as to avoid demagnetizing effects arising from the perpendicular orientation of the magnetization and thus make it easier for the magnetization to point out of the film media planes. A typical crystal structure for a cobalt – chromium medium is shown in Figure 18 where it can be seen that in comparison to longitudinal recording the grains have grown to become columnar in nature and have boundaries which are relatively easily discernible. These boundaries are chromium rich and help to produce magnetic discontinuity between the grains. It has been regularly proposed that these films behave as arrays of interacting fine particles, with magnetic anisotropies due to shape and crystalline *c* axes being parallel to the long axes of the grains.

Figure 18. Transmission electron micrograph of a cobalt chromium thin film with its easy axis normal to the film plane.

Two-layer recording media have also been produced for perpendicular recording in which the cobalt – chromium layer has been deposited onto a previously deposited soft magnetic nickel – iron coating, whose function is twofold. Firstly the coating reduces the demagnetizing effects within the medium during the recording process, and secondly influences the nature of fields produced by the record heads and thus also the record process (Fig. 19) [26].

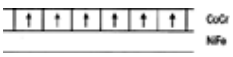

Figure 19. Diagram of a two-layer perpendicular recording medium.

A high coercivity cobalt chromium thin film with its easy axis directed normal to its plane sits on a soft NiFe thin film.

Perpendicular recording has required the development of some new head structures, which provide large fields in the directions perpendicular to the recording media planes. Nevertheless, ring heads have been used and these have performed sufficiently well to facilitate high density recording. However, they are not the optimum and the range of new heads has also been produced [26]. Figure 20 shows a single pole head situated above a recording medium. This consists of a single sheet of material whose thickness corresponds to the gap in the ring head. The head is driven by a coil which is often mounted on an auxiliary pole as shown in Figure 20 A and which is situated below the recording medium. When a current is passed into this coil the auxiliary pole is magnetized and it produces a low magnetic field which magnetizes the main pole. The latter produces a high localized field near its tip which penetrates into the recording medium and magnetizes it to achieve recording. This type of structure has been shown to work but it is clumsy in that it requires poles either side of the medium and the spacing between the two causes many difficulties. These difficulties have been alleviated by using a two-layer

medium shown in Figure 20 B, where the single pole above the medium carries its own coil and when driven this head induces an image of itself in the soft magnetic underlayer below the recording medium. Therefore there is effectively a pole above and below the medium and together they form a suitable record head structure. With these pole heads extremely high density recording has been achieved and they represent the highest densities so far to appear in the literature.

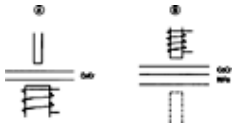

Figure 20. Pole heads for perpendicular recording

A) A single pole and an auxiliary pole; B) A single pole and a recording medium with underlayer.

Particulate media have also been prepared for perpendicular recording although, as a generalization, high quality media employing the same particles as used for longitudinal recording have not been successfully produced. However, barium ferrite particles have been used [26] which tend to form hexagonal platelets with their magnetic easy axes normal to their planes (see Fig. 21). When coated onto tapes the platelets usually lie with their planes parallel to the tape planes and so have their easy axis in the perpendicular direction. Thus media have been produced with high anisotropy and this has given rise to stable perpendicular magnetizations and high coercivities in that direction.

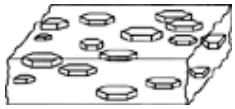

Figure 21. Recording tape containing barium ferrite particles.

Perpendicular media have been used for high density recording and have also been included in a commercial perpendicular recording floppy disk drive. However, exploitation of this medium and its corresponding technology has been rather limited to date mostly due to the continued success of the competitive longitudinal recording. In spite of that perpendicular recording remains as having potential for future application in very high density disk storage systems.

1.4. Recording Heads

Ring Heads. The most widely used type of recording head is the so-called ring head [28]. Ring heads have been the mainstay of magnetic recording throughout its existence and are expected to remain so for some time ahead although other and more varied types may appear. A ring head is illustrated in a simplistic way in Figure 2, which gives a good idea of the shape of the heads. They are often formed by bringing together two C cores (i.e., circular cores in the shape of the letter C) on each of which a coil is wound and which are joined together. Gaps are formed at both the front and back of the head, the one at the rear is simply a result of the fabrication process of the head and is not required, while the one at the front is specifically designed along with pole shaping to produce a large field just below the gap. Variations on this design abound and another common shape is made by bringing together a C and an I core. Here one half of the head is essentially straight (I core) while the other is circular. This latter head was often used when two heads were to be placed side by side, and back to back, to effect record and replay with one package.

Metal heads, typically of molybdenum permalloy (Ni – Fe – Mo alloy), Alfenol (Al – Fe) or Sendust (a Fe – Si – Al alloy), have been widely used in audio and data recording and are fabricated from stacks of thin laminations. The higher the bandwidth required the thinner the laminations. Although metal heads have high relative permeabilities (4000 – 11 000) and high saturation magnetizations, (up to 800 kA/m), they are very much susceptible to eddy currents. For very high frequency operation ferrite heads, typically Ni – Zn or Mn – Zn ferrites, are used as these are of very low electrical conductivity and very much less susceptible to eddy currents. They suffer, however, from low magnetization (up to 400 kA/m) and, often but not in all cases, low relative permeability (up to 10 000) and so they may have difficulty in recording on very high coercivity media. Some alleviation from this problem can be obtained by using very high saturation materials to make the pole tip region of the heads.

Metal in gap heads [29] are also used for a number of applications and in these various metals are included in the gap region of the head. Their function is either to divert the flux from passing through the gap between the poles of the head and out into the recording medium below, or to provide high magnetization in the pole tips to avoid saturation during recording. Such heads are widely used in video recording systems.

Thin-film heads are now gaining wider acceptance. They are widely used on disk machines, in some digital data tape systems, and also in some video and audio applications. A thin-film head is essentially a ring head but it is made using thin-film technology and lithographic processes to produce a very small magnetic circuit of a ring type. Such a head made of 80 : 20 NiFe is shown in Figure 22. Overall their efficiencies are rather similar to those of the monolithic heads but they have the advantage that they are more amenable to mass production.

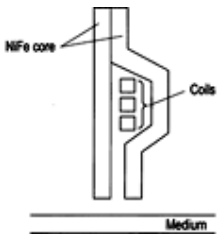

Figure 22. Cross-section of a thin-film recording head.

Thin-Film Flux-Sensitive Heads. For replay thin-film flux-sensitive heads have been produced and in diagrammatic form are shown in Figure 23. The one in Figure 23 A consists of an element of soft magnetic material which is held in close proximity to the tape and therefore it experiences magnetic fields emanating from the tape as it moves by. In magnetoresistive elements any change of magnetization causes a change of resistance and so as the recording medium moves by and its magnetic fields alter the magnetization in the element there is a change in resistance. This can be detected using electronic circuitry. There are many different designs of flux-sensitive heads which use a number of different magnetoresistive effects in single or multi-layer NiFe films. Such heads are very sensitive to magnetic fields and produce high output voltages. This is good because of sensitivity reasons but it has a special benefit in that being particularly sensitive it allows thinner recording media to be used and thinner recording media allow narrower transitions and higher recording densities. Hence for special applications such as high density recording on rigid disks the use of magnetoresistive heads is a major step forward.

The head shown in Figure 23 B is a shielded magnetoresistive head. Here the shields “absorb” the flux emanating from the tape and shield the sensitive element except when the recorded transition is directly below. In this way the head only sees the transition when it is directly below the element and the influence of nearby transitions is reduced. Hence the resolution of the read back process is improved.

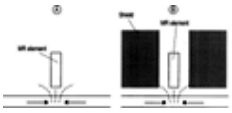

Figure 23. Idealized picture of (A) unshielded and (B) shielded magnetoresistive (MR) replay heads.

1.5. Economic Aspects

Magnetic recording is a technology which pervades all aspects of our lives. Be it audio, video or data recording its products are in every household and in every company. Each of the categories of business is at the $ 10^9 per annum level and drives huge industrial concerns to compete for the corresponding lucrative businesses. Both America and Japan have excelled in these technologies although their research and development, and production facilities are dispersed and spread into many parts of the world particularly in the Far East. In this way many other countries have benefited from the distribution of the production facilities to parts of the world where appropriate labor at highly economic prices is available.

The total values of media delivered worldwide during 1997 are given below. Figures are in $ × 10^9.

Rigid disks	4.4
VHS video tapes	3.5
Audio cassettes	2.0
Data tapes	1.4
Floppy disks	1.4
Other audio visual tapes	0.9
Other disks	0.3

The total media business amounted to around $ 14.4 × 10^9 but when the media are incorporated into products and systems the turnover of this enlarged industry is at least ten times or more of the figure quoted.

The audio market has been dominated by demands for cassette, particularly the compact cassette introduced by Philips, and open reel tapes for a.c. bias recording systems. The cassettes are universal, and remain one of the most popular vehicles for audio entertainment. The digital variants are DAT (Digital Audio Tape) and DCC (Digital Compact Cassette), which both delivered very high quality audio, comparable in quality to compact discs, but which have not achieved such wide acceptance in the domestic market. Both have, however, found their places in the professional world where their qualities are fully recognized.

Analogue video recording systems are dominated in the domestic market by the helical scan VHS system which sustains its leading position despite the fact that technology can now provide better performances on smaller cassettes. 8 mm wide tape systems are available and are used in portable video camera systems. Cassette systems are also in use for professional video applications. For the future on the domestic market Data VHS and digital video cassettes are likely to appear soon. Many new systems, some based on fixed head technology, are likely future products and will have the ability to interface into computer systems to increasingly enable multimedia applications. At the professional level digital video recorders and high definition television (HDTV) systems have been developed and will find increasing application. Overall there is a massive demand for recording media for video applications and this situation looks likely to continue for the foreseeable future although there is expected to be a rising demand and competition from various compact and other optical disc systems.

Data recording for instrumentation and computer applications may involve tape, cassette, cartridge, floppy disk, and rigid disk. All have there place in the system of storage hierarchy and have corresponding importance. Disk technology is central to computer storage and it is especially in the area of rigid disk technology where the pioneering advances in recording performance take place leading to extremely rapid improvements in terms of information storage densities and data rates. 3½ inch and 5¼ technologies predominate although 2½ inch and 1.8 inch disks have achieved positions in the market. Smaller disk sizes have also been introduced for special applications.

[Top of Page]

2. Optical Recording

Roman E. Wochele, Henk van Houten, Johan P. W. B. Duchateau, Hans J. G. Kloosterboer, Jan A. Th. Verhoeven, René van Vlimmeren, Peter E. J. Legierse, Dirk J. Gravesteijn, C. David Wright, Herman J. Borg, Heinrich Heitmann, Jacques Heemskerk

2.1. General Introduction

2.1.1. *Optical Storage Media*

Optical storage started on a small scale with the analogue video disc, commercially launched in 1978. Since then, it has become a mainstream technology due to the success of the *Compact Disc digital audio system* (CD-A) in the 1980s, the *Compact Disc Read-Only Memory* (CD-ROM) in the 1990s and the *Digital Versatile Disc* (DVD) in the late 1990s. In all cases, the digital information is encoded on a 12 cm diameter disc in the form of a relief structure, consisting of pits with varying lengths. This relief structure can be replicated at low cost and in high volume by injection molding of polycarbonate using a stamper carrying the desired pattern (see Section Optical Disc Mastering and Stamper Production).

The information layer in a Compact Disc is coated with a metallic mirror and covered by a protective lacquer. The layer is addressed by a converging laser beam, incident on the 1.2 mm thick transparent polycarbonate substrate, and focused to a diffraction-limited light spot [30], [31] (see Fig. 24). Reading through the substrate has the key advantage that the detrimental effects of dust and small scratches on the read-out signal are minimized, since these are out of focus. As a result, optical discs present a rugged, highly reliable, removable storage medium.

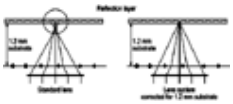

Figure 24. Substrate incident read-out of the information layer. The objective lens, corrected for the 1.2 mm polycarbonate substrate, guarantees an aberration-free focused spot. Dust particles on the substrate surface are out of focus and therefore they hardly deteriorate the read-out signal.

Cheap replicability, removability, and a long data life up to 100 years and more are the key technical reasons for the success of the Compact Disc as the medium of choice for the distribution of prerecorded digital information. It should be realized that removability also implies interchangeability of discs and players made by different manufacturers. This explains why worldwide standardization is so important in optical disc storage.

In addition to read-only optical storage systems, *write-once/read many* (WORM)-type compact disc systems (CD-R, DVD+/–R) as well as rewritable compact disc systems (CD-RW, DVD+/–RW, DVD–RAM) have been standardized [32], [33], [34].

Again, compatibility with their read-only counterparts is an important condition for widespread market acceptance. Such compatibility has been achieved in the case of write-once discs by employing laser-induced pit formation in organic dye films as a recording mechanism. An alternative solution is the laser-induced amorphous-to-crystalline phase change in thin films made of inorganic alloys (see Section Write-Once Optical Recording Materials). Phase-change recording with appropriately optimized materials is used for rewritable optical discs. Yet another possibility is magneto-optical recording, based on magnetic domain reversal (see Section Reversible Optical Recording Materials), which is used, for example, in the 64 mm diameter MiniDisc. However, magneto-optical discs are not compatible with CD/DVD, because they cannot be read by the nonpolarizing optics of a standard CD/DVD player. The bit pattern in the various types of information layers is illustrated in Figure 25.

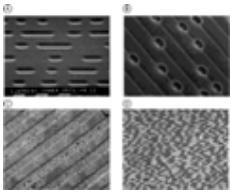

Figure 25. Microphotographs of the information layer for various types of discs

A) Replicated CD; B) Ablative write-once system, based on a tellurium alloy; C) CD-RW (phase-change); D) MiniDisc (MO)

A compact disc offers 650 up to 900 Megabytes (MB) of storage capacity for ROM, WORM, and rewritable media. The DVD–ROM provides 4.7 Gigabytes (GB) for a single information layer, 8.5 GB for a dual layer disc, and 17 GB for a double sided/dual layer disc. Recordable and rewritable versions of the DVD offering 4.7 GB in a single-sided/single-layer disc (DVD+/–R, DVD+/–RW, DVD–RAM) and 8.5 GB for a single-sided/double-layer disc (DVD+R9) are close to standardization. The new high-density versions of optical storage media (HD-DVD) using a blue-violet laser for writing/reading provide 27 GB (BD = Blu-ray Disc) [35], [36] and 30 GB (UDO = Ultra Density Optical) [37].

Future generations of optical storage should address the capacity domain over 100 GB. The roadmap for UDO shows for the third generation a capacity of 120 GB. Innovation will be required on many fronts: the optical head, short-wavelength semiconductor lasers, improved signal processing, modulation codes and error correction, single and multi-layer media, and shorter wavelength mastering equipment. Some of the evolutionary and revolutionary options are briefly mentioned in Future Developments. In Figure 26 a comparison of the bit pattern density of CD, DVD, and BD is shown.

Figure 26. Comparison of the bit pattern density of optical storage media

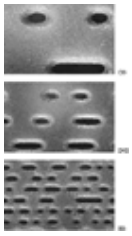

2.1.2. *The Optical Channel*

When it is reflected by the information layer, the diffraction-limited focused laser beam is modulated in amplitude, phase, or polarization [30]. The relief structure in the replicated Compact Disc predominantly modulates the phase of the reflected light beam. Phase modulation is also the main effect in write-once ablative dye systems, but phase-change recording results in a strong amplitude effect. Magneto-optical storage exploits the polarization modulation on reflection of perpendicularly magnetized domains caused by the magneto-optic Kerr effect. Discussions of magneto-optical data storage can be found in [38], [39]. In the following, some of the generic physical characteristics of the optical channel, including tracking and focusing, as well as the data path are described. For convenience, the Compact Disc system will be taken as a point of reference.

2.1.2.1. *Focusing and Tracking (Servo Control)*

The basic optical read-out principle is shown in Figure 27. A laser beam is focused through the substrate by an objective lens onto the information layer on the spinning disc. Reflected light returns through the lens and is directed to a photodiode detector. The diameter (full width at half maximum) of a diffraction-limited focused spot is given by $\lambda/2NA$, with *NA* the numerical aperture of the objective lens. In Compact Disc players, *NA* is usually chosen in the range 0.40 – 0.52 (DVD: 0.6; UDO/BD: 0.7/0.85), giving a spot diameter of about 1 µm. The high brightness of a laser is required to achieve an adequate signal-to-noise ratio [30].

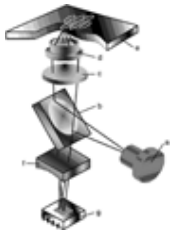

Figure 27. The light path in an optical recording system consists of a

a) Semiconductor laser; b) Half-silvered mirror acting as a beam splitter; c) Collimator lens; d) Objective lens for focusing the beam. The lens is mounted on an actuator for the adjustment of the disc-lens separation; e) Optical disc; f) A double wedge structure for the Foucault focus measurement; g) Detector for the beam reflected from the disc. The detector is a multiple sensing device, and information about focusing and positioning of the beam with respect to the track on the disc is also obtained during the read out.

The Compact Disc system uses a laser diode of group III-V elements (e.g., gallium arsenide) emitting at a wavelength in the near-IR (780 nm). The CD-R system employs the same wavelength. The DVD standard is based on a red laser (650 nm), the BD/UDO standard on a blue-violet laser (405 nm), using gallium nitride.

The diameter of the objective lens is chosen so that a free working distance (the separation between the lens and the nearest surface of the disc) of more than 1 mm is achieved. The diameter of the converging beam as it enters the substrate is typically 0.7 mm, resulting in reduced sensitivity to dust and scratches on the disc. As the substrate is part of the optical path, its thickness, variations in thickness and angular orientation, optical homogeneity, birefringence, and surface flatness must be included in the list of specifications of the optical system (see Sections Optical Disc Mastering and Stamper Production and Read-Only Optical Recording Materials).

In order to be able to scan the information layer on the disc, two functions must be provided by the optical system of the recorder: focusing and tracking.

Focusing. The depth of focus Δz of the laser beam is given in a good approximation by the equation:

$$\Delta z \simeq \pm \lambda/2(NA)^2$$

This is of the order of a few micrometers for CD. It is completely impractical to specify the flatness and alignment of the information surface to this degree. The lens – disc separation is therefore continuously adjusted during read-out. The focus adjustment is made using servo control of an electromechanical actuator for the lens movement. A dynamic measurement of the amount and direction of de-focus is required as input to the servo system. Several optical strategies are used, a brief description of the two most popular methods will illustrate the general principles.

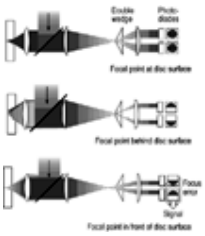

Figure 28. Schematic drawing of the Foucault double wedge principle for generation of the focus error signal. The incident laser beam (arrow) is deflected by a beam splitter, and is focused onto the disc. The reflected light is transmitted by the beam splitter, and reaches a second focus close to the apex of a double wedge. With four detectors the focus error signal is extracted by coupling the detectors according to the scheme (top + bottom detector) minus (sum of two middle detectors). The focus error signal changes from a negative value (bottom) to a positive value (middle). At the focal point (upper) the error signal is exactly zero.

Foucault Method. In the Foucault method, depicted in Figure 28, the apex of a prism (or double wedge) is placed on the optical axis at the ideal focal point of the returning beam. Four photodetectors, arranged in a line perpendicular to both the optical axis and the apex, are placed behind the focal point. The difference signal from the photodetectors yields an S-shaped focus error signal over a range of ± 4 µm with respect to the ideal focal position. The acquisition range, which is of importance during the initial focusing action, is much larger (up to 25 µm). This signal can be used to position the objective lens.

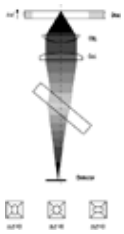

Figure 29. Astigmatic focusing method. In this picture a plan-parallel plate is used for generating perpendicularly-oriented astigmatic lines on either side of the best focus position.

Astigmatic Focusing Method. In the astigmatic focusing method (Fig. 29) a cylindrical lens is placed in the returning beam so that perpendicularly oriented astigmatic lines are generated on either side of the best focus position. The relative intensity of these lines can be measured using an appropriately oriented quadrant photodetector. The output is again used to give the desired S-shaped focus error signal.

Tracking. In addition to focusing, a dynamic measurement is required in the optical system to detect and follow the information track on the disc. In the Compact Disc, the track pitch is 1.6 µm, and the laser spot should be centered on the track to within 0.1 µm. Several methods are used, as illustrated in Figure 30.

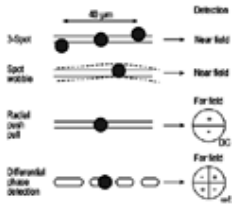

Figure 30. Different radial tracking methods

A) Three-spot tracking: the two satellite spots, generated by a grating, have a radial offset of 1/4 of the track pitch and an offset in the track direction of typically 20 µm. The d.c. content of the satellite spots is measured on separate detectors and will vary as a function of the radial position of the scanning spot.

B) Radial wobble method: an extra movement is given to the scanning spot relative to the tracks. From the intensity fluctuations on the detector it is possible to extract a radial error signal.

C) Radial push-pull method: the light flux through the pupil of the objective is separated in two parts and from the difference signal a tracking error signal is derived. Variations in light intensity over the pupil are due to the track structure, acting as a diffraction grating.

D) Differential phase detection: this method uses a quadrant detector over the cross-section of the returning beam and extracts a radial error signal by comparing the phases of the high frequency signals on the four quadrants.

Three-Spot Tracking Method. The three-spot tracking method (Fig. 30, first mechanism) employs a grating structure in the light path to generate two satellite spots from the main beam. The grating is oriented so that the radial offsets of the satellite spots are 1/4 of the track pitch, with a considerable (ca. 10 to 20 µm) tangential offset. If the main spot is on the track, the satellite spots ride partly on the land between the tracks, both preceding and following the central read-out spot. By measuring the d.c. content of the satellites on separate detectors and by establishing their difference signal, a tracking error signal can be generated.

Radial Wobble Method. In the radial wobble method (Fig. 30, second mechanism), the scanning spot is given an oscillatory movement in the radial direction. A radial error signal is deducted from the phase and intensity of the modulated reflected signal at the oscillation frequency.

Radial Push-Pull Method. The radial push-pull tracking method (Fig. 30, third mechanism) exploits the fact that the track structure of the disc acts as a diffraction grating, with for CD a period of 1.6 µm. Optical interference occurs at the detector between the zeroth and two first-order beams diffracted from the track structure. Using a detector split along the plane of symmetry of the far-field radial diffraction pattern, a tracking error signal is derived from the difference signal. The same split detector can be used for the read-out signal by summing the two contributions. This is the *central aperture signal*.

For prerecorded discs, the information itself is sufficient to form the required grating structure. In this case, the phase depth of the relief structure is chosen to be near to, but not equal to, the value $\lambda/4n$, The exact value $\lambda/4n$ leads to a maximum in the central aperture read-out signal, but to a minimum push-pull signal. For recordable discs, no information is initially present. In this case, it is customary to manufacture a continuous pregroove (a groove structure replicated to the disc) for tracking purposes. The depth of the pregroove is usually chosen to be close to $\lambda/8n$. The exact value leads to a maximum in the push-pull tracking signal, but the central aperture read-out signal is significantly reduced. A variant of the push-pull tracking method is the three-spot push-pull method, frequently used in CD drives.

Sampled Servo Method. The sampled servo tracking method (not shown in Fig. 30), used in some magneto-optical drives, requires special patterns of optical effects, present on the disc at predetermined intervals. The read-out of the pattern is used to adjust tracking and time-base errors; the servo system is then disabled until the next special pattern is encountered.

Differential Phase Detection. Finally, differential phase detection (DPD) (Fig. 30, fourth mechanism) is a method which derives a tracking error signal from the timing difference of signals detected by different parts of a four-quadrant detector. Such a phase difference occurs when the diffraction limited light spot impinges with a radial offset on a pit or written mark on a disc.

Lens Adjustment. The focus and tracking error signals serve to adjust the objective lens, using actuators. The objective lens is commonly mounted on an electromechanical actuator driven by the optical focus error signal. The maximum vertical displacement of the lens for *focusing* is of the order of 1 mm. For *tracking*, the entire optical pick-up unit (OPU) can be mounted in a linear or rotating carriage assembly. This low-bandwidth coarse displacement actuator is commonly supplemented with a high bandwidth, fine displacement actuator which drives the objective in the radial direction, to compensate for the inevitable track eccentricity.

In an alternative approach, most of the components of the optical pick-up unit (such as laser, detectors, beam splitters, and collimating lens) are mounted rigidly into the optical recorder. The collimated laser beam is sent to a unit that normally includes an actuated deflecting mirror for fine radial tracking and the objective lens, both mounted on a movable carriage [30], [31].

These servo systems are designed to relax the tolerances that must be placed on the geometrical parameters of the optical disc. Nevertheless, the performance of the servo systems is limited in terms of the magnitude and bandwidth of the corrections that can be accomplished (see, e.g., [30]).

2.1.2.2. *Data Channel*

For analogue *video disc systems*, the standard composite video signals are directly frequency modulated, and recorded on the disc (at carrier frequencies of 7.6 or 8.6 MHz). The signal path of a digital optical storage system is analogous to that of a digital communications system, as depicted in Figure 31 for a rewritable optical disc system.

Figure 31. Block diagram of the read–write channel of an optical disc recorder.

In the recording process (write channel), an analogue input signal is digitized (A/D converter), and an *error correction code* (ECC) is added. For the CD digital *audio system*, the audio signal is sampled at a rate of 44.1 kHz, which according to Nyquist's sampling criterion is sufficient to reproduce audio signals up to 20 kHz [30]. The number of bits is 32 per sample (16 per stereo channel), resulting in a bit stream of 1.41 Mbit/s. The audio bits are grouped into frames containing 6 samples each. For each 3 original audio bits, 1 bit is added for error correction, leading to a bit stream of 1.83 Mbit/s. Even though error correction is used, the raw bit error rate must be low (below 10^{-4}). The final data stream is interleaved or dispersed over several code words on the disc in order to circumvent the effects of optical disc defects such as pinholes [cross-interleaved Reed-Solomon code (CIRC)]. One more byte is added to each audio frame for playback and display functions.

To adapt the binary data pattern to the modulation transfer characteristics of the optical channel, a *modulation code* is applied, before the data are stored on the disc. A run-length limited sequence of bits is chosen, with a minimum run length of 3 and a maximum of 11 channel bits. This choice has been made to achieve an optimal match to the spatial frequency characteristics of the optical channel, and to realize an optimal data density. The chosen modulation scheme is known as *Eight-to-Fourteen Modulation* (EFM), because each byte of audio data (including CIRC and sub code data) is transposed through a look-up table into 14 channel bits [41]. Three further merging bits are added. The channel bit rate is 4.3 Mbit/s.

Both the encoding and the error correction are treated in detail elsewhere [41]. Recordable and rewritable discs require additional signal processing to generate the appropriate laser pulses needed for the chosen write strategy, as discussed in Section Compact Disc Family. The general aspects of encoding and signal processing of magneto-optical recording systems are similar, although the details differ [38], [39]. In the case of a read-only, replicated disc, the write process is replaced by the disc mastering process.

The read channel is similar, but is followed in reverse order (see Fig. 31). In the read-out process, extra steps are needed to determine the proper timing information (clock recovery), and to compensate for the particular frequency response of the channel (equalization). The read-out of such a disc is discussed in some detail in what follows. Modulation of the reflected light intensity from the optical disc is caused by the interaction of a diffraction-limited spot with optical effects on the disc having characteristic dimensions of the order of the spot size. In such cases the modulation can be modeled by considering the effects of light diffraction by supposedly periodic structures (amplitude or phase gratings) on the disc [30], [42]. A rigorous treatment of the coupling requires a complete vector diffraction theory solution of Maxwell's equations; the complexity of this approach has so far prohibited its widespread application. Fortunately, scalar diffraction theory can give a reasonable prediction of the relationship between optical amplitude or phase structures on the disc and the read-out signal characteristics. The optical response to the spatially modulated phase structures on the disc is described by the *modulation transfer function* (MTF). For central aperture detection of the optical signal, the MTF decreases monotonically from its maximum value at low spatial frequencies on the disc, to zero at a *spatial cut-off frequency* given by $2NA/\lambda$. The modulation approaches zero at the cut-off frequency because the length of a pit plus the space between pits then equals the diameter of the diffraction limited laser spot. The cut-off frequency determines the maximum information density that can be stored on the disc.

In good approximation, the amplitude of the detected signal is proportional to the modulation depth of the effects on the disc, modified by the frequency dependent MTF. Translated into the signal domain, the maximum signal frequency that can be extracted from an optical disc is $2vNA/\lambda$, with v the linear velocity of the track. The frequency dependence of the MTF results in timing jitter due to *inter-symbol interference* (ISI). An additional source of jitter is cross talk from adjacent tracks, but this is not yet significant in the compact disc. Electronic filtering may be used to reduce ISI by boosting the high frequency response, and thereby compensating for the MTF decay (equalization [43]). The practical frequency range of the recorded signal is typically limited to about 75 % of the cut-off frequency. The noise sources in the optical detection system also include electronic effects such as the dark current noise of the photodetector and the current amplifiers, the ultimate limit

being photon shot noise. From the media the important noise sources are the microroughness of the information-carrying surface and optical inhomogeneities and anisotropy of the substrate (e.g., birefringence). All these sources produce amplitude noise in the detected signal [44]. In addition, irregularities in the definition of the pits or recording marks on the disc are a cause of noise. Especially in write-once and rewritable media timing jitter associated with the microstructure of the information layer and the recorded mark edge definition are critical factors. The geometry of the optical marking effects in these media is difficult to predict, as it is only indirectly related to the intensity profile of the laser beam, because of the nonlinear and nonuniform thermal processes involved.

The EFM modulation code, jointly with the frequency dependence of the modulation transfer function of the optical channel, and the non-linearities in the photodiode transfer characteristic give rise to the eye pattern of a CD bit detector, as illustrated in Figure 32 bottom. (The term eye pattern derives from the similarity of the lozenge-shaped pattern formed by adjacent zero-crossings of the oscilloscope traces). The EFM choice yields a read-out signal with a small low frequency content of the information on the disc, so that the focus and tracking servo systems are not influenced by the low-frequency components of the data stream.

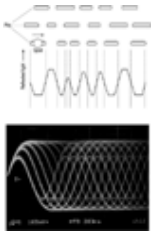

Figure 32. The information is present on the disc in the form of pits in a reflective surface. When the laser light hits a pit, the intensity of the reflected light is much lower than of the light reflected back from the reflective surface between the pits. These differences in intensity are translated into electrical signals. The eye pattern is the result of triggering the digital signal to the rising edges of the signal. The clearly opened eye pattern is the result of a virtually perfect combination of optical disc and read-out light path.

In the ensuing sections on optical recording materials the concept of *carrier-to-noise ratio* (CNR) is used as a figure of merit. This CNR is the signal-to-noise ratio of a single frequency written on the disc (the carrier). The signal and noise are measured in narrow band widths centered at and adjacent to that frequency, respectively. Although precise predictions on the performance of optical recording media cannot be made on the basis of CNR, it gives a quick impression about the potential usefulness of the optical recording material.

2.1.3. *Compact Disc Family*

2.1.3.1. *Read-Only, Replicated Discs*

The Compact Disc (CD) digital audio system appeared in a coordinated international introduction in 1983. Since then, derivatives of the CD system have emerged that use extensions of the original CD standard prepared jointly by Philips and Sony in 1979. The major derivatives are CD-ROM and CD-Recordable. Their technical specifications are described in the so-called “Red Book”, “Yellow Book” and “Orange Book”, respectively [45], [46]. Table 2 shows an overview of the CD books.

Table 2. Overview of the CD books, standards, and specifications

Book	Description, specifications	Dates (of latest version), comments
Red Book (CD Audio)	defines the physical properties of the compact disc and the digital audio encoding: • audio specification for 16-bit PCM. • disc specification, including physical parameters. • optical stylus and parameters. • deviations and block error rate. • modulation system and error correction. • control and display system and CD graphics allows for up to 74 min of digital sound. Sample rate of 44.1 kHz, or 44 100 samples per second. Transfer rate of 150 kilobytes per second. Also known as "single-speed" or 1X. can contain up to 99 tracks.	CD Audio (Aug 1995, Red Book); subcode/control and display system (Nov 1991, extension to Red Book for CD Graphics); CD Text (Sep 1996, extension to Red Book) the Red Book is also available as the international standard ISO/IEC 60908
Yellow Book (CD-ROM & CD-ROM XA)	defines the CD-ROM specification plus an extension for CD-ROM XA: • disc specification, optical stylus parameters, modulation and error correction and control and display system (from Red Book) • digital data structure (the sector structure and the ECC and EDC for a CD-ROM disc) • disc format including Q channel and sector structure using Mode 2 sectors (CD-ROM XA)	CD-ROM (Nov 1988, original Yellow Book); CD-ROM XA (May 1991, CD-ROM XA system description) The Yellow Book is also available as the international standard, ISO/IEC 10149 Yellow Book, Mode 1, ISO 9660 Level 1 is the most widely utilized standard since it will allow virtually all PCs to access its data. Mode 2 will permit the

	• data retrieval structure based on ISO 9660 including file interleaving which is not available for Mode 1 data (CD-ROM XA) • audio and graphics encoding The Yellow Book standard defines two data types. Mode 1 describes CD-ROM data with Logical Error Correction Code (LECC), which provides room for 2048 bytes of user data and is the mode used to store data that is unforgiving of error, like computer programs or databases. Mode 2 describes sector data stored without LECC, which provides increased room for 2336 bytes of user data. This mode is typically used for data that is more tolerant of error, like audio, video, or graphics.	"combining" of Mode1 and Mode2 data on a single disc. Mode 2 is primarily the entertainment portion of the Yellow Book standard. CD-ROM/XA, Bridge discs (including Photo CD, Karaoke CD, and Video CD), andGreen Book, or CD-i, fall under the Mode2 standard of Yellow Book. Mode2 discs are intended to be played on dedicated consumer electronics platforms.
White Book (Video CD)	defines the Video CD specification: • disc format including use of tracks, Video CD information area, segment play item area, audio/video tracks and CD-DA tracks • data retrieval structure, compatible with ISO 9660 • MPEG audio/video track encoding • segment play item encoding for video sequences, video stills and CD-DA tracks • play sequence descriptors for pre-programmed sequences • user data fields for scan data and closed captions • Examples of play sequences and playback control	Karaoke CD (1993, Original VCD for Karaoke); Video CD (Apr 1995, Current Video CD specification); Internet (Apr 1997, extension for linking to websites); Super Video CD (Nov 1998, Higher quality video using MPEG-2) the White Book specification for Video CD was announced by JVC, Philips, SONY and Matsushita in July 1993. A special CD-ROM/XA subset designed to hold MPEG-1 video. White Book defines Mode 2, Form 2 data standard which creates a disc that can contain up to 74 min of full-motion video. Limited with a slow playback rate and only 680 MB of storage space, the DVD standard is designed to replace the troubled Video CD. MPEG stands for Motion Picture Experts Group.
Blue Book (enhanced CD)	defines the (also known as CD Extra) specification for multisession pressed disc: • disc specification and data format including the two sessions (audio and data) • directory structure (to ISO 9660) including the directories for CD Extra information, pictures and data • format of the CD Plus information files, picture file formats and other codes and file formats • MPEG still picture data format	Enhanced Music CD (Dec 1995) a subset of the Orange Book specification, the Blue Book standard was designed expressly for stamped multisession discs limited to two sessions, one music and one data. Also known as CD Plus, Blue Book is a defined, licensed standard supported by Philips, Sony, Microsoft, and Apple. Discs play on CD-Audio players with no possibility of producing static and on computers with newer CD-ROM drives.
Photo CD Book	defines the specification for Photo CD: • general disc format including example of program area layout, index table, volume descriptor, data area, subcode Q-channel skew, CD-DA clips and microcontroller readable sectors • data retrieval structures including directory structure • image data coding including a description of image coding and image packs • ADPCM files for simultaneous playback of audio and images by interleaving • playback program system including playlist files	specified by Kodak and Philips based on the CD-i Bridge specification Last updated: Dec 1994
Orange Book (CD-R and CD-RW)	defines CD-Recordable discs with multisession capability: • disc specification for unrecorded and recorded discs	Part I: CD-MO (Nov 1990) Magneto Optical disks Part II: CD-R (Dec 1998) CD-WO -

	• pre-groove modulation • data organization including linking • multisession and hybrid discs • recommendations for measurement of reflectivity, power control etc.	Write Once Part III: CD-RW (May 1998) Rewritable
Purple Book (DDCD)	defines the specification of double-density compact disk (DDCD): • narrowing the track pitch to 1.1 µm from 1.6 µm • Shortening the minimum pit length to 0.623 µm from 0.833 µm • new type of error correction (CIRC7) • adaptation of the ISO 9660 file format • scanning velocity of 0.9 m/s	the informal name for specification produced by Philips and Sony in 2000 to describe their double-density compact disc (DDCD) format, which increases the storage capacity of the disc to 1.3 GB through means such as increasing its number of tracks and pits. Purple Book-defined products were released in recordable and rewritable - rather than read-only - formats, and require special CD-RW drives to handle them.
CD-i Bridge	defines specification for CD-i Bridge • disc format defining CD-i Bridge discs as conforming to the CD-ROM XA specification • data retrieval structure as per ISO 9660. A CD-i application program is mandatory and stored in the CDI directory • audio data coding which includes ADPCM and MPEG • video data coding for compatibility with CD-i and CD-ROM XA • multisession disc structure including sector addressing and volume space • CD-i related data since all CD-i players must be able to read CD-i Bridge data.	CD-I Bridge is a Philips/Sony specification, for discs intended to play on CD-i players and other platforms such as the PC. Confusion reigned temporarily when Philips put MPEG-1 video on Green Book discs, which could only be played on CD-i machines. However, all new MPEG-1 film titles conform to the White Book standard for Video CDs, which may be read on any White Book-compatible machine including CD-i and suitable CD-ROM drives. The video and sound are compressed together using the MPEG 1 standard, and recorded onto a CD Bridge disc.
Green Book (CD-i)	defines the CD-i disc format: • Volume Compact Disc Interactive Full Functional Specification • Microware OS/9 Technical Manual	CD-i is a real interactive system delivering various qualities of video, various qualities of audio together with full user interactivity in real time. CD-i was created by Sony and Philips mainly to provide entertainment by interfacing with televisions and stereo systems (May 1994). Compact Disc Interactive is the only specification which not only defines the disc and the data, but also the entire hardware and operating platform.
ISO 9660	ISO 9660 (International Standards Organization) is the file and directory naming standard that refined High Sierra to be readable on any platform, regardless of the content format. ISO-9660 specifies each file name consist of three components: name, extension and version. A name or extension may consists of zero or more characters of the set [A through Z], [0 through 9] and the _ (underscore). The version number ranges from 1 to 32767. There are three levels of this standard defined by ISO-Files are listed in alphabetical order. A directory name consists of one or more characters of the set. The Volume Name of the disc can contain up to 11 of the set. Directories are sorted alphabetically. ISO 9660 Level 1 - Known as the "DOS" Character set, the number of characters is restricted to 8 and the number of characters of the extension is restricted to 3. Also referred to as 8.3 (eight-dot-three). ISO 9660 Level 3 - The total length of the name and extension is restricted to 30 characters (excluding the point). ISO 9660 (Joliet) - The total length of the name and extension is restricted to 64 characters. ISO 9660 Rock Ridge Interchange Protocol - Permits UNIX and variants to use the naming standards of the UNIX platform and performs like a native volume. Hybrid - Permits "Partitioning of the disc" to allow the disc to perform natively in two or more platforms, operating systems, or environments.	

Further members of the CD family, such as CD-i (interactive CD), Photo-CD, Video-CD, and the rewritable CD-RW, have also been standardized. It is generally recognized that a key factor behind the success of the Compact Disc system is the high level of interchangeability between the different members of the family. As discussed in [46], this interchangeability is a result of standardization at various levels. The standardizations related to the media hardware include test conditions of the optical system, characteristics of the disc, the read, write and erase characteristics (i.e., the interaction of the laser with the information layer), and the data and servo formats. Software related standardization deals with file labels and formats. To allow true interchangeability, the CD-related standards are well protected internationally. Aside from the system and media components, the CD (audio as well as ROM) has had a very significant effect on the recording (software/content) industry.

This industry, therefore, plays an important role in setting future standards for optical disc systems.

The *CD-ROM*, containing up to 650 MB of user storage capacity, has been designed for the large scale distribution of computer software or large-scale data bases. After a slow start (it was first introduced on the market as early as 1985), it has now become a standard attachment for a PC. Since computer applications typically have a random bit stream, the CD-ROM cannot rely on some of the interpolation techniques used in CD-Digital Audio to handle detection errors, whereas also the addressing information needs to be bit-accurate to individual data frames. For this reason the CD-ROM requires some additional error detection and correction encoding compared to the basic CD-Digital Audio disc [43]. The last years have witnessed a rapid succession of CD-ROM drive generations, with ever increasing data rates (from the original CD-Audio data rate up to 52 ×). The drive spins at 52 × at more than 10 000 rpm. To reach these speeds several writing/reading strategies such as CLV, CAV, Z-CLV and P-CAV have been developed (Fig. 33.)

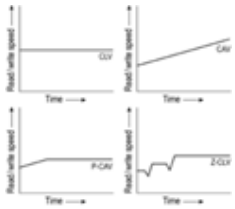

Figure 33. Overview of CD writing/reading strategies

Constant Linear Velocity (CLV). The disc is read/written at a constant speed. The data transfer rate is kept at a fixed level by changing the rotation speed.

Constant Angular Velocity (CAV) The disc is read/written at a constantly increasing speed. The drive maintains a consistent rotation speed which will result in a variable transfer rate.

Partial-Constant Angular Velocity (P-CAV): The disc is being read/written at an increasing speed until a certain point (speed). After this point the speed will not increase anymore and remain at this level. P-CAV is a combination of CLV and CAV. The transfer rate increases until the drive reaches its maximum speed (CAV). From that point the drive will slow down and the transfer rate will be constant (CLV). Because P-CAV drives reach their maximum speed much sooner than CAV drives, the average transfer rate should be higher.

Zoned-Constant Linear Velocity (Z-CLV). The disc is divided into several CLV zones. After each zone the write speed is increased (note that the Z-CLV mode is not used when reading a disc). Most high-speed recorders are using this technology.

Regarding the actual trend CDs will most likely continue to be popular for music recording and playback and for smaller quantities of computer data storage. Newer technologies like the digital versatile disc (DVD) store much more data in the same space and have higher transfer rates.

2.1.3.2. *Recordable and Rewritable Discs*

Various recordable versions of the compact disc have been proposed. An early magneto-optic version of the CD (CD-MO, described in the orange book, part 1) was not a success, because of incompatibility with the CD. A 12 inch write-once/read many (WORM) disc was introduced as early as 1984, based on organic dyes in the information layer. It was intended as a replacement for magnetic (Winchester) discs in computers. Again success was rather limited. A 12 cm disc with 650 MB capacity, compatible with CD, was defined based on the WORM format (orange book, part 2). Known as a *recordable compact disc* (CD-R) it is now rapidly penetrating the market. It can be used in a multi-session mode, by allowing consecutive recording sessions for a single disc (at the expense of capacity loss due to the need to define lead-in and lead-out sections for each session). The CD-R can be used for applications such as photo-archiving, back-up, or the copying of read-only discs for personal use [45].

A *rewritable CD* based on phase-change recording technology has been announced in 1995, and shown at industry briefings in fall 1996 under the name of CD-RW. In phase-change recording, laser pulse sequences are used to write a mark (see Fig. 34)

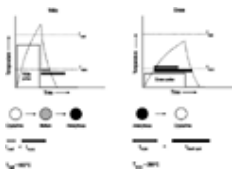

Figure 34. Principles of phase-change recording. Amorphous marks are written by using a short, high-power laser pulse to melt the recording material and subsequently quench it to below the crystallization temperature, so that nucleation and growth of the crystalline phase is avoided. Information is erased by heating the amorphous state to a temperature between the crystallization temperature and the melting point long enough to regain the crystalline state.

In between the pulses, the laser power is turned to a low level, so that after each laser pulse the molten material is quenched to below the crystallization temperature, resulting in an amorphous dot. A series of overlapping amorphous dots forms a recording mark. The reflection coefficient in the amorphous state is low, so that a recording mark corresponds to a pit in a CD. Direct overwrite can be accomplished by switching the laser power to the erase level in the regions between two marks (see Fig. 35). For an optimized erasing sequence the material is heated to a lower temperature (below the melting point), for a time long enough to regain the high-reflectance polycrystalline state (similar to the high reflectance of the CD “lands”, formed by the regions in between the pits replicated in the polycarbonate substrate).

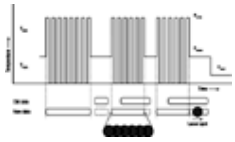

Figure 35. Direct overwrite strategy for phase-change recording. Each recording mark is formed by a train of short laser pulses which result in overlapping amorphous dots. Direct overwrite can easily be accomplished by switching the laser to the erase level between the writing of successive information marks.

The drives for recordable and rewritable CDs are similar to their read-only counter parts, although a higher-power 780 nm semiconductor laser is required, and some additional electronics to generate the write strategies.

Almost every CD-R can hold between 660 and 690 MB data but usually the total space of a CD-R is physically set to 650 – 659 MB. The extra space is normally used for the mandatory 90-s lead-out area. By shortening this lead-out area, extra data can be written on the CD-R. This method of writing a CD-R is called oversizing or overburning. The reason that the total space of a CD-R is set to maximum 659 MB is because most CD-writers are not able to write more than this amount of data to a CD-R (because of mechanical or firmware shortcomings). Thus bigger CD-Rs are useless for these CD-writers. A CD-R is in fact recordable up to the end of the physical pre-groove. To oversize CD-R's the Disc-At-Once (DAO) mode of the CD-writer is utilized. In the DAO mode, instead of TAO (Track-At-Once), the writer must provide its own lead-out (which now can be moved to the very end of the CD-R).

In 1998 the 80 min/700 MB CD-R's came on the market. When they were released they were quite expensive ($ 25) but within a year the price dropped considerably ($ 2). The secret behind the CD-R 700 (80 min), 800 (90 min) and 900 (100 min) is that every CD-R has a spiral (pregroove), which starts on the inside and ends on the outer part of the CD, where the data is written. The laser head will always follow this spiral from beginning to end to read (or write) data. The distance between the start- and end-position of the spiral is defined in the CD Red-Book and is the same for all 63, 74, 80, 90, and 100 min CD-R media. So the spiral has to be wound tighter to fit 80 min (or more) data in the same space where normally only 74 min data resides. 80, 90 and 100 min CD-R's can be written by almost any CD-writer. However, a few CD-writer manufacturers (such as HP and Sony) do not support “big CD-R’s” as these do not comply to the CD Red-Book standard.

2.1.4. *Digital Versatile Disc (DVD)*

The success of the compact disc family is based on the availability of replicated read-only discs, which are a removable medium ideally suited to the low-cost distribution and preservation of digital information. Recordable (and rewritable) versions became successful only after compatibility with read-only discs had been achieved. Removability and compatibility require standardization, and this hampers incremental increases in the storage capacity, in contrast to non-removable media such as hard discs. At first, such a capacity increase was not needed, because a 650 MB CD-ROM offered far more capacity than could be handled by early PCs. However, due to the rapid progress in processing power typical file sizes have increased dramatically (from 1 MB in 1982 to 500 MB in 1995 [45]). In addition, the digitalization of audio, graphical, and video information has led to the emergence of multimedia PCs. Thus, a high density successor to the CD was needed.

After considerable debate about the respective merits of various proposed solutions [45], a new standard has been laid down in the blue book for the digital versatile disc, or DVD-ROM, in fall 1996. A DVD-ROM [47] can store 4.7 GB data on a single information layer — a 7-fold capacity increase over a CD-ROM. The required increase in areal data density was made possible by a number of changes in physical parameters, listed in Table 3. The basic idea is to reduce the diffraction-limited spot size used for read-out, given by $\lambda/2NA$, through the choice of a shorter wavelength diode laser (650 nm instead of 780 nm), and a larger numerical aperture of the objective lens (NA = 0.6 instead of 0.45). This enables a reduction of the track pitch and of the channel bit length. The physical bit density was thereby increased by a factor of 4.7 (see Table 3 and Fig. 37). This comes at the expense of a severe reduction in operating margins, such as disc tilt and focus error. These margins were kept at an acceptable level by reducing the substrate thickness from 1.2 mm to 0.6 mm. Because of the large installed base of CD-Audio and CD-ROM media, the same physical format was required (12 cm diameter, 1.2 mm thickness). A DVD therefore consists of two 0.6 mm substrates bonded back-to-back. A double-sided version of the DVD was thus made possible, allowing a total capacity of 9.4 GB. In addition, a double-layer version of the DVD has been standardized, with a capacity of 8.5 GB per side (see Fig. 36). Fig. 38 shows an overview of the DVD read only formats.

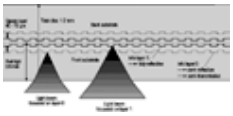

Figure 36. DVD double layer structure. Information layer 0 is covered with a semi-reflective layer and is separated by a spacing layer from information layer 1. The latter is covered with a full reflective layer. Read-out of both information layers is assured by switching focus.

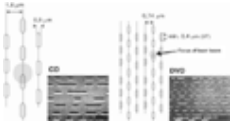

Figure 37. Comparison of CD and DVD

Figure 38. Overview of the DVD read only formats, a) DVD-5 (single-sided, single-layer, 4.7 GB), b) DVD-9 (single-sided, double-layer, 8.5 GB), c) DVD-10 (double-sided, single-layer, 9.4 GB), d) DVD-18 (double-sided, double-layer, 17.0 GB)

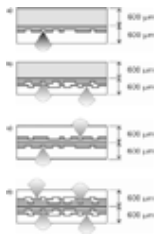

In addition to the changes in physical bit density, changes were also made in the data format and the channel code (from EFM to EFM$^+$, an eight-to-sixteen code). The DVD is more sensitive to disc defects due the increased density, and thus requires a more powerful error correction scheme. The cross-interleaved Reed–Solomon code (CIRC) of the compact disc was replaced by the Reed–Solomon Product Code (RSPC), which is so effective that it takes up a smaller fraction of the capacity (13 % instead of 30 % [46]).

To enable multimedia applications in both consumer electronics appliances and as a computer peripheral, an open file format was adopted for the DVD [Universal Disc Format (UDF) an extension of the CD-ROM ISO 9660 file format]. The DVD standard accommodates high-quality digital compressed video according to the MPEG-2 variable bit rate standard. When used for video applications, a DVD can also accommodate 8 channels of audio and up to 32 channels of sub-titles. The format is backwards compatible with the older MPEG-1 standard used in the Video-CD.

Table 3. From CD to DVD

Parameter	CD	DVD
Wavelength λ, nm	780	650
Numerical aperture NA	0.45	0.60
Track pitch, µm	1.6	0.74
Channel bit length, nm	277	133
Shortest pit/land length, nm	831	399
Modulation code*	EFM	EFM+
Physical bit density, Mbit/cm^2	106	508
Reference velocity *CLV*, m/s	1.2	4.0
Substrate thickness *D*, mm	1.2	0.6
Spot size $\lambda/2NA$, µm	0.9	0.55
Capacity, GB	0.65	4.7
Depth of focus ($\propto \lambda/2NA^2$), mm	± 2	± 0.9
Tilt margin ($\propto \lambda/NA^3D$), °	± 0.6	± 0.3
Thickness margin ($\propto \lambda/NA^4$), mm	± 0.1	± 0.02

* For EFM one has 17 channels bits (14 modulation bits and 3 merging bits) for 8 data bits. Each channel bit corresponds to $\frac{1}{3}$ of the minimum mark length. The physical bit density equals 1/(track pitch × channel bit length × 17/8). For EFM$^+$ the factor 17/8 is replaced by 16/8.

The coexistence of DVD and CD calls for the development of players that can accommodate both types of disc. This is nontrivial, because of the differences in physical parameters listed in Table 3. For example, the difference in substrate thickness requires optical pick-up units that can read out information layers through substrates of 1.2 and 0.6 mm thickness, while the objective cannot correct the different amounts of spherical aberration simultaneously. Acceptable and creative solutions have nevertheless been found [48].

Similar to the rainbow books for the Compact Disc, there are also specifications for the DVD (see Table 4).

Table 4. Overview of the DVD books, standards, and specifications

		Part 1	Part 2	Part 3
Book name		**Physical**[a]	**File system**[b]	**Application**[c]
A	DVD–ROM	read-only	ISO 9660 + UDF	undefined
B	DVD–Video	read-only	UDF	MPEG-2 video
C	DVD–Audio	read-only	UDF	high quality audio
D	DVD–R	write once	UDF	not defined

	DVD–RAM and DVD–RW rewritable	UDF	
E	RW		not defined

[a]Part 1 defines the physical format of a DVD–ROM disc. Various categories of information are defined:

- Physical - Disc size, curvature, reflectivity, etc.
- Sectors - Data format of the sector headers and ECC blocks.
- Layout - Lead-in/out areas, data areas. Differences between physical and logical sector numbering.
- Basic data-format and location of sectors that describe the disc type and layout (e.g. layer count, physical, and logical sector address mapping, etc.)

[b]Part 2 defines the file system DVD uses a UDF-Bridge file system. That is, both an ISO-9660 and a UDF file system structure are present on the disc. This is to allow backward compatibility for file access.

[c]Part 3 defines the DVD–Video The DVD–Video specification defines the format of all files stored in the VIDEO_TS directory on a DVD–Video disc (IFO, BUP, VOB). Unlike the physical and file system specifications, this information is not officially available.

In contrast to recordable and rewritable CDs, two main competing standards and three recordable DVD formats exist on the DVD writing side, i. e., DVD–RAM, DVD–R/RW and DVD+R/RW. DVD–R/RW (this is a combination of DVD–R and DVD–RW-note the minus in the name), mainly developed and promoted by Pioneer, has taken an early lead. The other competing standard, DVD+RW (note the "plus" in place of the dash), mainly developed and promoted by Philips, making a strong bid. All in all there are six recordable versions of DVD: DVD–R for general, DVD–R for authoring, DVD–RAM, DVD–RW, DVD+RW, and DVD+R. DVD–R and DVD+R can record data once, like CD-R, whereas DVD–RAM, DVD–RW, and DVD+RW can be rewritten thousands of times, like CD-RW. DVD–R was first available in fall 1997. DVD–RAM followed in summer 1998. DVD–RW came out in Japan in December 1999, but was not available in the USA until spring 2001. DVD+RW became available in fall 2001. DVD+R was released in mid-2002.The situation is similar to that of the competing VHS and Beta video standards decades ago. But unlike the video tapes of old, the two main competing DVD standards are designed to create discs that may be read by most consumer DVD movie players. Drives of both standards can read commercial DVD movies and work as CD readers and burners. Current DVD–recorders are multiple-format writers, that support both formats, DVD+R/RW and DVD–R/RW.

The third DVD rewritable standard, DVD–RAM, runs in the PC-world a distant third. A DVD–RAM is initially more a removable storage device for computers than a video-recording format, although it has become widely used in DVD video recorders, because of the flexibility the system provides in editing a recording. This technology is incompatible with commercial DVD players — data stored on a DVD–RAM disc cannot be read by DVD movie players and many other DVD drives. DVD–RAM discs are housed in cartridges and hold between 2.6 and 9.4 GB.

The double-layer format for recordable DVDs, announced in 2003 by Pioneer (DVD–R9) and Philips (DVD+R9) right after each other, has once again shown that the format fight is not over. Double-layer rewritable discs are still under development. The double-layer DVD+R system uses a thin organic dye film as data storage layer [34]. The recording principle is the same as used in single-layer recordable DVDs and is based on irreversibly modifying the dye's physical and chemical structure, induced by heating with a focused laser beam. The recorded marks have optical properties different from their unmodified surroundings, giving the same read-out signals as read-only discs. The digital information is contained in the length of the recorded marks and the unmodified spaces between them. Adding an additional storage layer, that can be accessed from the same side of the disc, provides challenging issues in organic dye-based optical recording. The transmission of the upper recording layer is > 50% to allow for read-out and recording of the lower recording layer. In the upper layer of DVD+R9 a thin silver-alloy layer is used as reflector material. The reflection of the upper layer should be at least 18% to ensure compatibility with the double-layer DVD–ROM standard. The lower recording layer has a high reflection (> 50%) and high power sensitivity, because the upper recording layer absorbs and reflects part of the incoming light. Effectively, at least 18% reflection of the lower layer is required. The transparent spacer layer is about 55 µm thick and serves to physically separate the two recording layers. The laser beam can easily be focused on either of the two layers by adjusting the position of the objective lens. So the double-layer recordable DVD will probably be readable in all DVD players, the point is, that the existing DVD recorders have to be updated or exchanged. The future will tell if this will be as a big success as the current single-layer DVD–Recordable. In Table 5 an overview of the currently available recordable DVDs (April 2004) and their specifications is given.

Table 5. Overview of the specifications for recordable DVDs

Media	DVD–RAM	DVD–RW	DVD+RW	DVD–R	DVD+R	DVD+R9
Published by DVD–Forum	yes	yes	no	yes	no	no
Hardware maker	Panasonic / Toshiba / Hitachi / Samsung	Pioneer / Sharp / Sony	Philips / Ricoh / Sony	Pioneer / Panasonic	Philips / Ricoh / Sony	Philips / MKM
Promotion organization	Recordable DVD Council	RWPPI	DVD +RW Alliance	RWPPI / Recordable DVD Council	DVD +RW Alliance	DVD +RW Alliance
Book version	Ver 2.1	Ver 1.1	Ver 1.1	Ver 2.0 / for general	Ver 1.0	n/a
Capacity/ sided			4.7 GB			8.5 GB (double

						layer)
Wavelength, nm (NA)	650 (0.6)	650 (0.6)	650 (0.6)	650 (0.6)	650 (0.6)	650 (0.6)
Channel bit length, µm	0.14~0.146	0.133	0.133	0.133	0.133	0.133
Track pitch , µm	0.615	0.74	0.74	0.74	0.74	0.74
Depth, nm	~70	~30	~30	~180	~180	n/a
Media type		Phase change			Dye	
Active layer material	GeSbTe	AgInSbTe	AgInSbTe	cyanine / azo dyes		
Reflectivity, %	15~25	18~30%	18~30%	45~85%	45~85%	> 18 %
Recording speed, kB/s	2700	1350-2700	1350-10800	1350-10800	1350-21600	3240-n/a
Tracking method	land/groove	groove	groove	groove	groove	groove
Addressing	CAPA	LPP	wobble	LPP	Wobble	Wobble
Format	ZCLV	CLV	CLV/CAV	CLV	CLV/CAV	CLV/CAV
Tracking	P-P	DPD	DPD	DPD	DPD	DPD
Recording	random	random / sequential	random / sequential	sequential	sequential	sequential
Finalization	without	with	with	with	with	with
Defect management	yes	no	no	no	no	no
Copy protection	CPRM	CPRM	CPRM	-	-	n/a
Cartridge	selectable	no	no	no	no	no
Cyclability, times	100 000	1000	1000	once	once	once

2.1.5. *HD-DVD and Blue Laser Discs*

Blue laser systems are able to store more data than DVDs because of the shorter wavelength of blue light. This means that the laser, which is used to record data on the disc, produces a smaller spot on the recording layer, which in turn means that the space needed for each bit of data is smaller. HD-DVD (HD stands for both high-density and high-definition) was under development before DVD came out and finally emerged in 2003. Some high-definition versions of HD-DVD use the original DVD physical format but depend on new video encoding technology such as H.264 to fit high-definition video in the space, that used to hold only standard-definition video. High-density formats use blue or violet lasers to read smaller pits, increasing data capacity to around 15 to 30 GB per layer. High-density formats use high-definition MPEG-2 video and may also use advanced encoding formats. Early in 2004 are six systems are under development, announced or released (see Table 6).

Table 6. Summary of specifications for HD-DVD (under development, announced, or released)

Format	Backers	Data depth	Laser	Video	Capacity (single layer/dual layer)	Data rate
HD-DVD (AOD)	DVD Forum	0.6 mm	blue (405 nm)	HD MPEG-2, H.264, VC9,MPEG-4	15 GB / 30 GB (ROM)20 GB / 40 GB (recordable)	36 Mbit/s
Blu-ray	Blu-Ray Disc Founders	0.1 mm	blue (405 nm)	HD MPEG-2	27 GB / 50GB	36 Mbit/s
UDO	Sony/ Plasmon	0.1 mm	blue (405 nm)	n/a	30 GB / 60 GB	64 Mbit/s
EVD	eWorld (Govt. of China)	0.6 mm	red (650 nm)	HD MPEG-2 (later AVC)	n/a / 8.5 GB (ROM)	22 Mbit/s
FVD 1	ITRI (Taiwan)	0.6 mm	blue (405 nm)	AVC	17GB / n/a	25.05 Mbit/s
FVD 2	ITRI (Taiwan)	0.1 mm	blue (405 nm)	AVC	17GB / n/a	31.59 Mbit/s

HD-DVD (AOD). The DVD Forum's next-generation DVD, once called Advanced Optical Disc (AOD), currently being called HD-DVD, is a modification of the existing DVD physical format to enable storage of about 15 GB per layer using a blue-ultraviolet readout laser. The same 0.6-mm data depth as in DVD is used. AOD is designed to improve data capacity while theoretically being able to use existing replication equipment. It is primarily supported by Toshiba and NEC.

Blu-Ray Disc (BD). The first HD-DVD format announced was the Blu-ray disc. This format was developed outside of the DVD Forum, that supports the successful DVD–Video format. Blu-ray is a new high-density physical format that will hold 23 to 27 GB per layer using a blue-ultraviolet laser and a 0.1-mm data depth. Because of the 0.1-mm cover layer it will require significant changes to production equipment. Blu-ray is initially intended for home recording, professional recording, and data recording. Mass-market distribution of prerecorded movies will come later, after the read-only format, called BD-ROM, is developed and the details of video, audio, interactivity, and copy protection are hammered out. Blue-ray supporters are LG, Panasonic, Philips, Pioneer, Hitachi, Mitsubishi, Samsung, Sharp, Sony, and Thomson. Sony released the first BD recorder in Japan in April 2003. In Figure 39 the development in data density from the CD to the Blu-ray disc is shown.

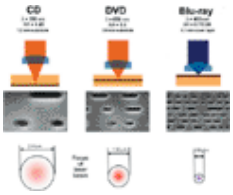

Figure 39. Comparison of CD, DVD and Blu-ray technology

The most important technical details of the Blu-ray disc are up to 27 GB per layer using 0.1-mm recording depth (to reduce aberration from disc tilt), a 405-nm blue-violet semiconductor laser with 0.85 *NA* lens design to provide 0.32 µm track pitch (half that of DVD) and a pit length as small as 0.138 µm. Variations include 23.3 GB capacity with 0.160-µm minimum pit length (used by Sony's Professional Disc system) and 25 GB capacity with 0.149-µm minimum pit length. The physical discs use phase-change groove recording on a 12-cm diameter, 1.2-mm thick disc, similar to DVD–RW and DVD+RW. The data transfer rate is 36 Mbit/s. Recording capacity on a single layer is about 2 h of HD video (at 28 Mbit/s) or about 10 h of standard-definition video (at 4.5 Mbit/s). The cartridge size is 129 ×131 × 7 mm. There are plans to produce dual-layer recordable discs, holding about 50 GB per side, but such discs will take a few additional years to appear.

UDO. UDO (Ultra Density Optical) is a 5.25-inch format with a recording density of 7.4 Gb/in^2, mainly developed by Plasmon and Sony. The target market for UDO is professional optical data storage and secure long-term archival storage for valuable business information. The system is technically similar to the Blu-ray disc format, that was developed by a consortium of nine companies led by Sony, although both systems are incompatible. The main difference between the two formats is in the data transfer rate. Blu-ray, which is aimed at consumer recording of high-definition television, can record data onto the discs at a rate of up to 36 Mbit/s to match the data rate of digital television. However, UDO can record data in the first version at 64 Mbit/s or double the rate of Blu-ray. The first version of the UDO format is a single-sided, single-layer optical disc with a capacity of 30 GB. The roadmap extends to a 60 GB capacity version by 2006 at 96 Mbit/s and a 120 GB capacity version by 2008 at 144 Mbit/s. UDO uses a patented phase-change recording process, that permanently alters the molecular structure of true write-once media, ensuring data integrity at the most fundamental level. UDO is also available with rewritable media for archive environments, where data needs to be deleted or media capacity reused. Unlike true write-once media, rewritable media allow the phase change recording process to be reversed. Enhanced error correction algorithms (Reed-Solomon) and read-ahead defect management complement UDO's phase change recording process to guarantee high data integrity.

EVD. A government-backed consortium of companies in China, called eWorld, has developed a domestic version of DVD called EVD (Enhanced Versatile Disc). EVD uses red laser discs but with tighter tolerances than DVD to hold more data. Video is encoded in HD MPEG-2, although a future version will use the China-developed AVC compression format. EVD players first appeared in China at the beginning of 2004.

FVD 1 and FVD 2. The Advanced Optical Storage Research Alliance (AOSRA), formed by Taiwan's Industrial Technology Research Institute (ITRI) has its own variations of blue-laser formats. FVD 1 uses a 0.6-mm data depth similar to AOD, and FVD 2 uses a 0.1-mm data depth similar to Blu-ray.

2.1.6. *Magneto-Optical Storage*

Magneto-optic recording is intrinsically a rewritable optical storage technology [38], [39]. The medium is a disc consisting of a transparent substrate, covered with a thin magnetic film with perpendicular magnetic anisotropy (so that the magnetization vector is oriented perpendicular to the plane of the film). A typical material is the rare earth – transition metal alloy TbFeCo. Initially, the disc is magnetized uniformly. Writing occurs by a thermomagnetic process (see Fig. 40). A focused laser spot is used to heat the thin permanent magnetic layer locally above its Curie point (typically 200 °C). A small coil is positioned at the other side of the disc, opposite to the laser spot. It creates a weak magnetic field (a few hundred Gauss) that causes reversal of the magnetization vector in the heated region of the film. The size of the written domain can be less than the spot diameter. The written domain is quite stable at room temperature, because of the large coercive field of the materials used in magneto-optic recording. In comparison with hard discs or floppy discs, MO discs are much less sensitive to stray magnetic fields, so that they are an attractive removable storage medium.

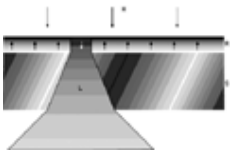

Figure 40. Principle of magneto-optical recording. A focused laser beam *L* heats the magnetic layer *R* to above the Curie point, so that local magnetization reversal because of the applied external field *H* can occur. The magneto-optical effect rotates the plane of polarization of the linearly polarized read-out spot. The sense of the rotation inverts when the direction of the magnetization is reversed.

Read-out of an MO disc can be done with the same laser, used at a lower power. The read-out mechanism is based on the magneto-optic Kerr effect: the direction of polarization of the laser light is rotated on reflection off the magnetized domain. This is detected by means of polarizing optics.

A disadvantage of magneto-optic recording in its simplest form (i.e., writing with a constant external magnetic field, and modulated laser power) is that it does not allow for direct overwrite (erasure requires an external field of opposite sign). This limitation may be overcome by keeping the laser power constant, and switching the direction of the external magnetic field (*magnetic field modulation*, or MFM). This technique requires a very small electromagnet, flying at a very short distance from the disc surface. An alternative way to achieve a direct overwrite capability is to make use of exchange-coupled thin film stacks, in combination with laser-power modulation [50], [51].

To circumvent the recording density limit imposed by the size of the focused laser spot, the *magnetic amplification technique* has been proposed. A second magnetic layer with a lower Curie point, for example GdFeCo, is sputtered on top of the recording layer. When heated above about 100 °C by means of the focused laser spot, the material covered by the spot copies — through exchange coupling — the magnetization of the underlying information layer. In the presence of an external magnetic field of the proper orientation, the size of the magnetized area is comparable to that of the laser spot, which can be considerably larger than the size of the bit mark written in the recording layer. In reverse external field, no magnetic field amplification occurs, and the read-out signal vanishes [50], [51].

MO recording drives have not been very successful as a PC peripheral, or on the consumer electronics market. This may be due in part to the fact that no standard file format exists for magneto-optic discs (comparable to the ISO 9660 format for CD-ROM). The exchange-ability of MO discs is therefore limited. MO recording has thus mainly addressed professional and niche markets.

An exception is the recordable version of the 64 mm diameter MiniDisc system, which is aimed at the consumer electronics market. It is gaining acceptance, especially in Japan. Its general specifications are very similar to those of the CD: it is based on a laser with 780 nm wavelength, an objective with a numerical aperture of 0.45, the modulation code is EFM, and error correction is done by an advanced form of the cross-interleaved Reed-Solomon code. The MiniDisc can be used to record 74 min of compressed digital audio. For portable applications the shock-sensitivity is improved by using a semiconductor buffer memory.

2.1.7. *Future Developments*

2.1.7.1. *Evolutionary Progress in Optical Disc Storage*

The introduction of the DVD and the new blue laser discs is following the increase in file size, which itself follows the increase in capacity of hard disk drives. The storage capacity of magnetic hard disks is currently increasing at a rate of about 50–60% per year. The areal density of current high-end hard disks is more than 60 Gbits/in^2. As shown in Figure 41, the physical bit density of state-of-the art magnetic hard disks exceeds that of an optical medium. There is currently much debate on the fundamental limits of both optical and magnetic storage. Magnetic recording areal density has increased approximately 2.5×10^6 fold since 1957, but researchers are now debating the ultimate limits imposed by superparamagnetic effects and media limitations such as grain size. On the optical side, new advances in recording films such as “superresolution” and more sophisticated encoding/decoding schemes are extending fundamental limits imposed by the available laser technologies. It is expected that this rapid progress will continue unabated in the next decade. A similar rate of progress is anticipated for semiconductor memories (DRAM). Also the processing power of microprocessors is rapidly increasing. Typical file sizes will also keep growing, allowing new applications in domains such as multi-media, or 3D interactive graphics. It is, therefore, to be expected that in the future a higher-capacity successor of current optical media will be required.

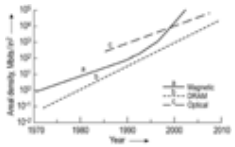

Figure 41. Areal density development of physical density of optical and magnetic storage media and DRAM

A possible path to an even higher capacity replicated optical disc, as already used in the blue laser discs, is offered by the combination of shorter wavelengths (GaN lasers will enable 410 nm or even 360 nm recording), thinner substrates, and higher numerical apertures. A high *NA* may be achieved by a dual objective [52]. The extra lens could either be mounted on a second actuator (requiring additional error signals for servo control), or on a slider, as in a hard disk drive. In its ultimate form, the extra lens is a solid immersion lens (SIL) in near-contact with the optical disc. An *NA* > 1 is then possible [53]. Points of concern are contamination and damage of the disc surface. This is a problem shared with other forms of near-field (or scanned probe) storage [54].

In addition, the possibility exist to use *optical superresolution*. This technique uses an active layer or additional coatings which change their optical properties under intense laser illumination. The net effect is to spatially filter the central part of the laser spot. In one form (called rear-aperture detection) the coating replaces the aluminum mirror in a compact disc. If a chalcogenide material such as $Ge_2Sb_2Te_5$ is used, it melts under strong laser illumination, and becomes highly reflective. In a second form (called front aperture detection) a complementary effect is used, and only the central part of the focused laser spot can reach the information layer. In both cases it is possible to read out premastered marks with a size below the diffraction limit [55].

Finally, as in magnetic storage, more sophisticated detection techniques, such as partial response maximum likelihood (or Viterbi) detection, in combination with new formats (such as land/groove recording) may be considered.

In view of the nature of optical disc storage as a standardized, removable medium, starting in the market with mass-distributed replicated prerecorded media, it should be expected that the capacity increase will proceed by big steps, as in the transition from CD to DVD, rather than incrementally as in the hard disk drive industry.

2.1.7.2. *Three-Dimensional Optical Storage*

It is quite conceivable that even higher storage capacities, 100 GB and beyond, will be required for the consumer market in the future. One can then resort to juke boxes of optical discs, or to magnetic or optical tape recording [56]. Alternatively, forms of three-dimensional optical storage are under development. The third dimension can be spatial (holography, multi-layered bit recording [57], two-photon addressing of thick media), spectral (spectral hole burning [58], or a combination of both (holography combined with spectral coding [59]). Early in 2004 InPhase Technologies presented the first generation of a 200 GB holographic drive, that uses the blue laser technology (405 nm). The system will arrive at the market in 2006. In holographic data storage, light from a coherent laser source is split into two beams, signal (data-carrying) and reference beams. Digital data to be stored are encoded onto the signal beam via a spatial light modulator. The data or strings of bits are first arranged into pages or large arrays. The 0's and 1's of the data pages are translated into pixels of the spatial light modulator that either block or transmit light. The light of the signal beam traverses through the modulator and is therefore encoded with the “checkerboard” pattern of the data page. This encoded beam then interferes with the reference beam through the volume of a photosensitive recording medium, storing the digital data pages. The interference pattern induces modulations in the refractive index of the recording material yielding diffractive volume gratings. The reference beam is used during readout to diffract off the recorded gratings, reconstructing the stored array of bits. The reconstructed array is projected onto a pixelated detector that reads the data in parallel. This parallel readout of data provides holography with its data transfer rates up to 800 of Mbit/s. The readout of data depends sensitively on the characteristics of the reference beam. By varying the reference beam, for example by changing its angle of incidence or wavelength, many different data pages can be recorded in the same volume of material and read out by applying a reference beam identical to that used during writing. This process of multiplexing data exploits the enormous storage capacity of holography.

2.2. Optical Disc Mastering and Stamper Production

2.2.1. *General*

Commercial production of optical media requires large scale and reliable replication of the optical media substrates with the relevant physical structures. These submicron structures have to provide the well specified optoelectronic signals. Based on technical and commercial criteria, molding or embossing technologies are presently the method of choice. Both these technologies require a “mother disc” which is called “stamper”. This is generally a nickel-based disc which can be mounted in a given substrate replication system.

Mastering and stamper production refer to a family of processes for making this “mother disc”. The first step in the creation of the relevant mother disc is the mastering process. Master manufacturing is followed by making a one-to-one replica from that master by a well-established electroforming procedure which yields the stamper [30], [60-62]. A typical production chain of optical discs is shown in Figure 42 [63]. The information to be recorded is brought into a standardized format by “data formatting” (including error correction methods).

Figure 42. Flow diagram for the production of optical media.

The formatted data is written by a laser beam recorder (LBR) on the prepared master disc, which is subsequently developed and a thin metal layer is applied before a nickel copy can be made from the master by electroforming. This nickel stamper is then used for mass replication of optical media substrates. The substrates are covered with a reflective metal layer, a protective coating and a printed label in the disc finishing processes.

The media requirements for playability (staying in track and being synchronized, mark formation during recording of user data) and for manufacturing strongly determine the shape and the tolerances of the physical structures. Therefore first a review on these requirements will be presented.

2.2.2. *Requirements*

Playability Requirements. For *writable media* (write-once, rewritable media), the physical structures have to provide reliable (virtually error-free) tracking and allocation or timing. As depicted in Figure 43 tracking is normally done by a pre-embossed groove (often called pregroove), by a “pseudo pregroove” consisting of a continuous sequence of pits, or by a structure of regularly positioned pits (sampled servo tracking). The tracking technologies based on these concepts are well proven to provide reliable tracking of the laser beam during writing, erasing and/or reading (see Section General Introduction).

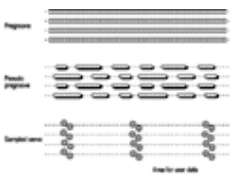

Figure 43. Examples of tracking provisions. Continuous groove, pseudo groove, and sampled servo.

Addressing a package of data on the disc can be done by allocation procedures using (1) sectors with pre-embossed sector headings; (2) by using discrete time flags along the tracks; (3) by timing procedures using embedded time marking along the pregrooves either by depth and/or width modulation; or (4) by a radial wobbling around the central track line (Fig. 44).

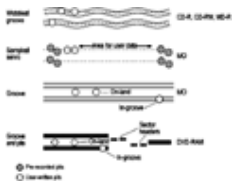

Figure 44. Typical provisions for addressing data on optical media.

For MO both sampled servo and continuous groove are shown.

The *CD-R concept* embodies a pregroove for tracking and a wobbling feature of the track for timing and/or addressing. The *ISO-MO standards* use a combination of a pregroove with pits on the land between the grooves for sectorization to provide data allocation. For *DVD-RAM* a combination of on-land and in-groove recording is chosen with sector headers positioned half way between the track center lines.

Read-only media have to provide the data too. A common principle is to use the sequence of pits, which represents the encoded user data (audio, video, or data and their error detection and correction information) also for tracking and timing. The sequence of pits acts as a pseudo-track, and the data is grouped in specific frames in such a way that a number of these groups can be used to provide addressing (and timing) information.

Table 7 gives an overview of specific functionalities for the physical structures and their implementations. The dimensions of the physical structures depend strongly on the vacuum wavelength of the read-out laser λ and the refractive index *n* of the substrate. In going to shorter wavelengths for read-out, smaller physical structures have to be made by mastering.

Table 7. Physical structures of various optical media and their functionalities

Optical medium	Tracking provisions	Timing location	Data recording	Track pitch	Typical pit sizes*
Laser Disc	pseudo groove	time flags	analog pits	1.7 µm	depth: 0.13 µm width: 0.5 µm
Compact Disc	pseudo groove	frame no.	digital pits	1.6 µm	depth: 0.13 – 0.15 µm width: 0.5 µm
CD-Recordable	grooves	radial wobble	in groove	1.6 µm	
CD-Rewritable	grooves	radial wobble	in groove	1.6 µm	
Mini Disc (recordable)	grooves	radial wobble	in groove	1.6 µm	
ISO-MO	grooves	pre-recorded sector headings	on land	1.6 µm – 0.9 µm	
DVD-ROM	pseudo groove	frame no.	digital pits	0.74 µm	depth: 0.11 – 0.12 µm width: 0.25 µm
DVD-Recordable DVD-RAM	grooves			0.74 µm	

* Only for pre-recorded media. Writable media show a large variation in sensitive layers and therefore in depths and widths of the physical structures.

In addition, the recording performance of writable media depends on the groove shape too. Rewritable systems struggle in general with signal-to-noise problems. The presence of a groove decreases the reflected signal and thereby the signal-to-noise figure. In order to avoid this, the signal-to-noise ratio can be improved by either recording the data pits on the flat area between the grooves (“on-land” recording) or by making the grooves rather wide and flat (“wide groove” recording). This is illustrated in Figure 45.

Figure 45. Examples of groove recording

a) In-groove (small groove); b) On-land; c) Wide-groove

The marks represent the user written data.

The “on-land” recording format is standardized for the ISO-MO media. Figure 46 shows an atomic force microscopy (AFM) picture of the pre-recorded header of “on-land” recording.

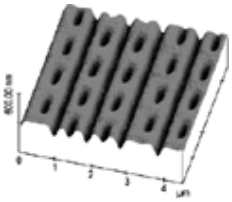

Figure 46. AFM picture of MO format showing grooves and sector headers.

Specific microscopic structures have to secure a good recording performance. Hence, optoelectronic tracking and synchronization specification as well as recording performance determine the optimal physical structure.

Each deviation from the intended structure will be reproduced in the final product. In addition, statistical phenomena as well as reproduction imperfections contribute cumulatively with the subsequent production steps. It is the final disc which has to provide virtually error-free tracking, clocking and data recovery. This puts high requirements on the mastering technology, both for its resolution as well as for its quality level and secure large number replication. A first parameter to characterize the quality of the formed pits during mastering is jitter; being the statistical deviation from the pit/land and land/pit transitions with respect to the clock.

Disc Manufacturing Requirements. The playability imposes, in general, specific requirements on the microscopic structure of the physical structures. For example, data pits give optimal optoelectronic signals if their microstructure shows a flat bottom, steep edges, and a depth of one quarter of the effective wavelength of the play back laser. For ease of replication, however, a stamper should have no sharp cornered pits. The slopes involved would be as shallow as possible. This makes the replication or embossing process easier. Control of the parameters which determine the physical microstructures enables an optimal compromise between the requirements involved.

2.2.3. *Mastering Methods*

Several technologies have been explored to produce a master disc. Mechanical pressing of pits into a relatively soft and ductile metal layer by actuation of a piezo driven diamond-stylus impression tool was one of them [64]. Technologies based on ion and electron beams have been investigated too.

On the basis of commercial criteria and on technical requirements for resolution, reliability, and microstructuring of the physical structures, optical laser mastering is presently the preferred industrial choice.

Optomechanical System for Laser Beam Cutting. The production of a spiral or concentric track is achieved by rotation of the master disc and horizontal shifting of the laser spot. The light beam's intensity is modulated in the optical modulator in order to generate the required physical structures (see Fig. 47).

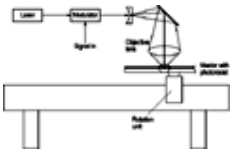

Figure 47. Schematic of a laser beam recorder writing on a master disc.

Laser beam recorder systems for advanced formats, such as ISO-MO, use dual beam systems to generate simultaneously both a groove and pit structures. Given the submicron precision of the location and the size tolerances of these structures, a variety of technologies are used in order to secure stability in the optical, mechanical, and electrical subassemblies of the optomechanical system for mastering.

In the laser beam mastering process, the limits of the recording process are determined by the size of the focused spot of the laser. The most ideal spot formation can be obtained by focusing an optical beam free of aberrations (i.e., from an aberration-free laser source) and homogeneously distributed over the opening of the diaphragm. Figure 48 A presents the intensity distribution of the aberration-free and diffraction-limited spot. Even this ideal spot consists in reality of a central spot with full width at half maximum (FWHM) diameter *d*:

$$d = 0.51\frac{\lambda}{NA}$$

The numerical aperture *NA* is given by the effective opening of the lens system (see Fig. 48 B). Around this central spot are, even for the most optimal aberration-free situation, annuli of decreasing brightness. Any deviation from the ideal situation will lead to larger spot diameters as well as to more intensities in the annuli. In practice, the resolution depends not merely on the basic parameter λ/NA, but also on the quality of the laser, on the optical system — including homogeneity of the distribution of the laser light over the limiting aperture, on lens quality, on alignment, and on system stability.

Figure 48. Intensity distribution of an ideal laser spot

A) Intensity distribution; B) Effective opening of the lens system and numerical aperture

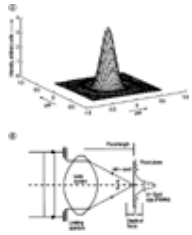

In the conventional optical-disc mastering process, many disc manufacturers have been using laser beam recorders (LBR), which utilize ultraviolet (UV) or deep-ultraviolet (deep-UV) lasers as light source. In April 2004 Pioneer presented a high-precision electron beam recorder (EBR), that employs an electron beam as a recording beam to sharply narrow the beam diameter. This can realize even finer pattern processing in the mastering process than LBRs. The high-precision EBR also achieves high levels of record-positioning accuracy, thanks to the high-precision recording-position control technology. Research and development for the EBR started with the production of laser discs in 1993. With EBR master discs for high-density optical discs including Blu-ray discs can be manufactured, as well as discrete track media and patterned media - higher-density hard disks regarded as highly promising future technologies.

The diameter of the spot size is the first parameter which controls the resolution of the final master recording (see Section d-2.1.4-d-2.1.7).

Further optical requirements and characteristics of the mastering process have been described in detail [30], [65], [67], [44], [66], [68-71]. An atomic force microscopic (AFM) picture of a DVD master is shown in Figure 49.

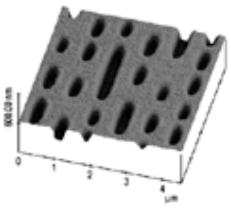

Figure 49. AFM picture of a DVD master.

Photoresist Master and Stamper Making. The used master disc is a circular, flat, polished glass substrate that has been spin-coated with a layer (130 – 150 nm thick) of a conventional positive working photoresist. Before spincoating, the substrate is usually treated with an adhesion promoting layer. The information is written into the resist layer in the air-incident mode using a laser beam master recorder. After writing is complete, the exposed areas of the photoresist layer are selectively removed by developing the disc in an aqueous alkali solution. In the optically exposed areas, a photochemical reaction occurs such that the rate of dissolution in the developer is increased by several orders of magnitude. A description of the novolac – *o*-naphthoquinone diazide photoresist system used for mastering is given in [60]. This type of photoresist is also widely used in making integrated circuits (⟶ Imaging Technology – Printed Circuits (Printed Circuit Boards)).

Development results in selective removal of the photoresist material from the exposed areas. The master disc is then rinsed with water, dried, and coated with a metal layer of silver or nickel by means of wet chemical or vacuum technologies. The master disc can now be read in reflection, so that its optical quality can be assessed. Next this master disc is used for stamper production (see Section Electroforming of Stampers).

Direct Stamper Recording. In 1997 a modified photoresist technology was introduced for direct stamper making. This process uses a negatively working photoresist in which the laser irradiated areas remain after the development process. The negative photoresist is spin-coated on a metal substrate with the thickness of a conventional nickel stamper. After recording and development the photoresist stamper is treated to harden the soft photoresist structures to withstand the high temperatures in the molding. The process enables shorter cycle times to produce a stamper by skipping electroforming.

Ablative Dye Mastering. Instead of using conventional, chemically developable, photoresist materials, the mastering process can also be carried out by direct laser writing in a thin layer of a degradable polymer such as nitrocellulose or acetocellulose. The polymer is again deposited by spin-coating from a solution that also contains a dye. The laser light is absorbed by the dye, converted into heat, and the polymer decomposes locally into volatile products.

This approach does not require chemical development. The ablative writing mode can, however, lead to debris and a pit rim that increase the noise levels of the master and the replica. An additional feature and requirement for this approach is the possibility of direct-read-after-write, which offers immediate optical quality assessment by read-out during writing and thereby write-power control.

Photolithography with Dry Ion Etching. A modified photoresist mastering technique uses the exposed photoresist layer to serve as a dry ion etch mask. Indium oxide, silicon oxide, or metals such as chromium are sputtered or otherwise deposited on a glass substrate to the correct thickness (near $\lambda/4n$ at read-out). The photoresist layer is spin-coated onto this sputtered layer, exposed, and then chemically developed and dissolved. Reactive ion or reactive ion beam dry etching is used to selectively etch only the oxide or metal layer. The remaining unexposed photoresist layer is then removed by dissolution, leaving the desired relief structure of oxide or metal. This technique requires three more steps than normal photoresist mastering, so that the latter is generally the preferred mastering technique.

Direct Metal Mastering. Instead of laser writing in a photoresist layer, direct writing in a very thin (20 – 30 nm) metal coating on a glass substrate is also possible. The heat generated by the laser beam evaporates the very thin metal layer. This procedure also offers the advantages of no chemical development and immediate optical quality assessment by direct read-out during writing. This approach has the disadvantage of creating debris and possibly a rim on the pit or hole during laser writing leading to an increased noise in the master. Thus, these masters cannot be used to provide electroformed stampers for molding and embossing based replication without complex steps to overcome this problem.

2.2.4. *Electroforming of Stampers*

Replicas and stampers are produced from the master by electroforming, also called electroplating, plating, or galvanics (artref Electrochemical and Chemical Deposition a09_127). The photoactive layer of the glass master is first covered with a thin (50 – 100 nm) metallization layer of nickel or nickel-vanadium. The end product leaving mastering is most widely referred to as metallized glass master. From the metallized glass master one nickel metal copy (a 300 µm stamper father) is generated by electrolysis. Cleaning, coating, polishing, and punching finish the stamper to fit the mold in the molding process.

Basically electroforming has remained the same process since the old vinyl discs were produced, with certain refinements, but it is still a highly manual process. The example of a typical stamper specification below gives a rough idea as to which extent the refinements took place.

Figure 50. Stamper specification

1. Thickness T = 300 µm (+0/–10 µm)
2. Inner diameter D_i = 34 mm (+0.010/–0 mm)
3. Outer diameter DO = 138 mm (+0.5/–0 mm)
4. Absence of burrs
5. Absence of pinholes and nodules
6. Back roughness R_Z / R_Y = 0.5 – 2.5 µm
7. Low material stress /deformation

The specifications in Fig 50, together with information on how they are mainly influenced by electroforming parameters are discussed in detail in the following sections.

Basic Model of Electroforming. The basic functionality of the electroforming process (shown in Fig. 51) can be summarized as the transportation of nickel ions from the anode basket to the object at the cathode, on which they are deposited. The medium for this transportation is the aqueous solution of nickel sulfamate. Controlling the electrical field and the current over time allows to control thickness, respectively thickness variation, of the galvanoform, that is to be copied in the family process or reinforced in the father process, respectively.

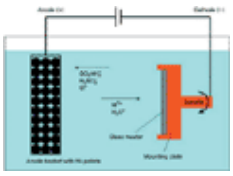

Figure 51. Basic electrolyte / electroforming model

Following the principle, that all ions in the solution are free motion-independent charge carriers, it is easy to picture the force of the electrical field driving the ions. If the solution is well balanced ▪in which respect? ▪, then practically the only conversions taking place are the two desired main reactions at the electrodes:

Anode reaction	Cathode reaction
$Ni \longrightarrow Ni^{2+} + 2e^-$	$Ni^{2+} + 2e^- \longrightarrow Ni$

At the cathode the positively charged nickel ions are discharged and the Ni atoms are built into the crystal lattice of the already deposited nickel, building up the metal layer that forms the stamper. Simultaneously the negatively charged ions ($SO_3NH_2^-$, $H_2BO_3^-$, Cl^-) of the solution assist in the process of dissolving the solid nickel pellets. Thus the nickel metal of the anode is brought to the state of freely moving ions, replenishing the nickel ions consumed at the cathode. The amount of nickel precipitated on the cathode is governed by Faraday's law, that can be summarized in the electrochemical equivalent of nickel: 1.095 g of nickel is precipitated per ampere hour (Ah) of current. For practical application a rule of thumb can be used: For a 240 mm diameter glass master 1 Ah ≈ 3 µm of average thickness. A typical absolute working amperage for a 240 mm glass master may be 80 – 120 A d.c. at 15 – 20 V.

Electrolyte Recipe and Process Parameters. Typically most electrolyte recipes in practical CD electroforming can be found in the range of parameters listed in Table 8.

Table 8. Typical process parameters

Nickel sulfamate c[$Ni(SO_3NH_2)_2$] 300–450 g/L	
Nickel chloride c [$NiCl_2 6H_2O$]	0 – 20 g/L
Boric acid c[H_3BO_3]	30 – 70 g/L
pH -Value	3.5 – 4.5
Temperature	45 – 65 °C
Anode material	Sulfur-activated pellets (Ni-S-pellets)
Additives:	generally as little as possible
Wetting agents	
Brightening agents	
Anode activators	
Current densities	15 – 30 A/dm^2

As the complexity of the recipe already indicates, the desired reactions are not the only reactions to take place, especially at the desired deposition rates. One reason for this is that water also influences the process. Additionally all negatively charged ions move towards the anode and all positively charged ions move to the cathode. This leads to undesired codepositions and secondary reactions. By using a well-balanced nickel sulfamate electrolyte, these side effects can be reduced to a level, that fulfills the specifications of the stamper. This requires some consideration with regard to the recipe as well as to the design of the unit used for electroforming.

Recipe Considerations. Concentration of nickel sulfamate (Ni^{2+}): Generally a higher concentration of nickel leads to a higher availability of charge carriers that increase the conductivity of the solution, leading to a reduction of the anode cathode voltage at a given working amperage. At concentrations tending toward the top limit of the range specified above, a saturation peak is reached where no further gain in conductivity is possible. Similar to the effect on conductivity, the internal material tensions (material stress) pass through a minimum at a certain, relatively high, concentration.

Concentration of nickel chloride (Cl^-): The contribution of nickel chloride to the nickel concentration can generally be neglected. The active component is the chloride content. The function of Cl^- is to ensure good and constant anode corrosion and reduce the risk of anode passivation at high current densities. One effect of anode passivation is depletion of (nickel) ions in the electrolyte. Anode passivation must also be avoided, in order not to generate low pH values which would support the hydrolysis of sulfamate ions to sulfate and ammonia. Both sulfate and ammonia lead to elevated levels of material stress, even at low concentrations (≤ 1 g/L). It has become common knowledge that by using sulfur-activated nickel pellets the use of nickel chloride can be avoided, provided certain precautions are taken. This is of considerable interest, because an increased concentration of chloride ions in the electrolyte inevitably leads to proportionally increased levels of material stress.

Concentration of boric acid: Boric acid mainly serves as a pH buffer in the electrolyte solution. A minimum concentration is required to ensure high-speed deposition of nickel at the cathode. Without boric acid during the process a film with an extremely high value pH could be formed on the nickel surface, binding nickel as its hydroxide. This film would lead to increased anode cathode voltages and in the worst case to blackened (burnt) nickel depositions.

pH Value: The pH value of the electrolyte solution has to be kept between 3,5 – 4.5. A pH value in this range has only little effect on the quality of the metalwork. However, it is good manufacturing practice to keep the pH in rather tighter tolerances than specified above. Extremely low pH values of 2.0 – 2.5 have to be avoided because they support the hydrolysis of the sulfamate ion. As the H_3O^+-ions are mainly responsible for codeposition of hydrogen at the cathode, a well balanced pH value is essential to avoid increased pitting or pinholes.

Temperature: Usually a somewhat higher temperature is desirable, as it leads to an increased mobility of the ions in the solution. This increases the conductivity of the solution, reducing the anode - cathode voltage at a given working amperage. Unfortunately, extremely high temperatures will also support the hydrolysis (decomposition) of the sulfamate ions. The process of electroforming generates a considerable amount of heat. The actual temperature of the electrolyte between anode and cathode will generally be higher than in the main tank and makes an appropriate cooling system imperative in the system. Most modern electroforming systems use a titanium tube solenoid for this purpose, externally supplied with cooling water. Tank temperature in nickel sulfamate in 60°C (± 2°C). Temperatures higher than 70°C lead to bad results.

Anode material: In CD/DVD electroforming so-called Ni-S -pellets are most widely used. Basically these pellets consist of pure nickel metal with a small addition (≈ 0.2%) of sulfur. The sulfur content reduces the electrochemical potential of nickel, supporting a stable anode corrosion, even at very low chloride concentrations or with no chloride at all in the electrolyte. As a side effect of this sulfur addition a small amount of anode sludge (small black granules) is left over when the pellets are consumed. This sludge is one of the major reasons for the necessity of good filtration to avoid nodules.

Additives: In CD/DVD electroforming the general rule applies that as few additives as possible should be used. Nearly all additives are organic chemicals influencing the nickel electroforming process with respect to pitting (wetting agents), back roughness (brightening agents) and anode activity. Unfortunately, most of the additives have unavoidable side effects although a range of modern agents exist, that have only minor negative effects. But like all organic chemicals used as additives, their utilization leaves organic degradation products, that contaminate the solution. The contamination increases the need for active carbon treatment. Generally organic contamination may be a major source for process problems such as increased material stress.

Current densities: In electroforming the key figure for the rate of nickel deposition is the current density applied on the surface of the cathode area. A working amperage of 100 A on a 240 mm diameter glass master for example results in a theoretical current density of approximately 22.1 A/dm^2. Apart from the limitations imposed on the process from the material side (see above), the current density cannot be increased at will in order to reduce cycle times. Material stress is increased proportionally to increasing current densities. As a result a compromise has to be found between acceptable material stress and shortening of cycle times in electroforming. Unfortunately, no standardized method has yet been implemented to measure material stress quantitatively at current densities and deposition thicknesses typical for CD electroforming. The fact that it is often impossible to distinguish mechanical deformation and material stress in a stamper is the main reason for the absence of such a measuring standard. Typical current programs feature a starting ramp (especially necessary for a glass master's thin initial metal layer), leveling off to the working current. Modern control systems allow easy programming and control of the ramping as well as direct ampere-hour setting to program the average thickness.

Schematic Layout of an Electroforming System. In Figure 52. the schematic layout of an electroforming system is depicted. The reproducibility of metalwork with tightly controlled thickness variations requires a stable geometry of the electrical field between anode and cathode. As mentioned before, the average thickness of an electroformed piece of metalwork can be controlled by achieving well controlled ampere-hour values. Thickness variations over the surface of the working piece require a suitably designed geometry of the electrical field between anode and cathode.

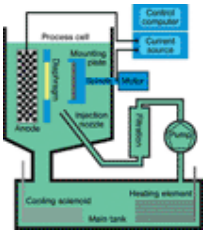

Figure 52. Schematic electroforming system (Source: technotrans AG)

Rotating spindles and control of the electrical field: Achieving a perfectly homogeneous electrical field should result in the required thickness tolerances. But as the simple model of a plate condenser (Fig. 53) easily visualizes, in practice this is virtually impossible to achieve.

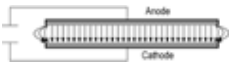

Figure 53. Plate condenser model of a homogeneous electrical field

Not only is the field inhomogeneous at the edges of the conductive areas, but also the necessity of having anode and cathode both perfectly parallel and flat can hardly be achieved in reality. The need of being able to install and remove anode basket, mounting plate, and for example the glass master makes certain tolerances unavoidable. The consequence for the electroforming process is that a relatively strong electrical field will lead to relatively higher deposition rates of nickel. In most systems used in CD/DVD electroforming a rotating spindle at the cathode helps to generate a concentrically homogeneous deposition of nickel. The evenness of nickel deposition can then be relatively easily controlled by adequate shielding. A positive side effect of rotation is, that codeposited hydrogen will be washed away mechanically, reducing pitting. A total thickness variation of ≤± 3 µm for a stamper father will be guaranteed by manufacturers of good quality electroforming systems.

Flow control of electrolyte: For thermal reasons the electrolyte in a process cell must be exchanged at relatively high rates, to ensure stable working temperatures of the electrolyte. Injecting conditioned electrolyte between anode and cathode is typically the solution chosen in most electroforming systems. The design of the flow needs considerable care. Unwanted thickness variations may be the result if the flow rate or its course are not designed and controlled adequately. A good design typically allows to monitor and adjust flow rates as needed. Additionally a good flow design supports mechanical removal of codeposited hydrogen (reducing the need for wetting agents as used for antipitting) and does not affect thickness variations when the flow rate is changed.

Anode basket and diaphragm: Anode baskets are typically made from titanium, ensuring a flat anode surface for the electrical field, while facilitating good and stable anode corrosion. Titanium is used, as its oxide layer is chemically inert and will not allow corrosion of metallic titanium for anode - cathode voltages up to 20 – 22 V. To avoid the codeposition of anode sludge on the cathode (nodules) the anode has to be clearly separated from the cathode. Preferably a diaphragm should be used for this purpose combined with a well designed electrolyte flow. This minimizes the risk of sludge reaching the cathode. Historically so-called anode bags made from fabric have been used to contain the anode sludge around the anode basket.

Filtration: High-quality pump filtration systems, made from similar materials as the process cells, are required to avoid carrying over particles from the tank into the process cells. These particles could be nuclei that trigger the formation of nodules. Preferably a pre- and a fine filtering stage, based on filter cartridge technology, should be included in the system. This also facilitates active carbon treatment with cartridges, whenever an organic contamination is be suspected.

A schematic representation of the process is depicted in Figure 54

Figure 54. Schematic representation of the production of replicas and stampers for optical discs

A) Build-up of master

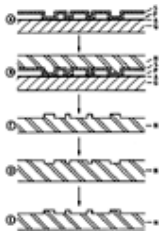

Glass disc (G); Adhesive layer (A); Photoresist (P); Silver (S)

B) Master plated with nickel (N)

C) Nickel negative or father stamper after removal of photoresist, adhesive layer and silver

D) Nickel positive or mother stamper

E) Nickel negative or son production stamper

In the process of separating the father stamper from the master, the relatively soft photoresist master is destroyed. When a silver layer is used as a metal coating, the silver has to be removed too. The nickel father stamper is replicated several times by electroplating to obtain positive *mother* nickel stampers (for small series of discs the father stamper can be used directly for disc replication). From the mother stampers, negative *son* nickel production stampers are replicated by electroplating. In each of the nickel-to-nickel replication steps, a separation layer is required. Rinsing with a dilute aqueous solution of potassium dichromate or an anodic treatment to form an oxide layer are typically used. Many negative production son nickel stampers can be made by this family method from a single photoresist master disc.

The electroplating process and nickel solution used are quite similar to those used for gramophone record production. The tolerance demands on the thickness of the nickel metal layer, the flatness of the metal backing, and the cleanliness of the nickel solution are much more stringent for stampers used to produce optical discs.

Polymer Stampers. Polymer production stampers can also be made by photoreplication from a mother nickel stamper or from an original photoresist master disc. A liquid, photopolymerizable coating is squeezed between a flexible plastic substrate and the stamper or master, followed by photopolymerization of the coating. This process is similar to a disc replication production process (2p or photopolymerization). For use with the typical elevated temperatures occurring in production molding processes, the photopolymerized stamper is coated with a metal such as chromium or aluminum.

2.3. Read-Only Optical Recording Materials

2.3.1. *Production Processes*

Numerous processes have been investigated for the replication of optical discs. At present *photopolymerization* (for digital optical recording, pregroove structures), *injection molding* (for video discs such as Laser Vision (LV), Compact Disc audio, and DOR pregroove structures), and *injection – compression molding* (for CD audio) are among the most important. These processes are shown schematically in Figure 55 and will be discussed in detail in Sections Molding Processes and Photoreplication Process.

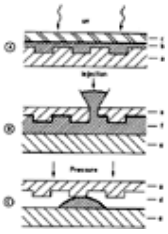

Figure 55. Production of optical discs

A) Photopolymerization

B) Injection molding

C) Injection-compression molding

a) Mold with stamper; b) Photopolymerizable coating; c) Transparent plastic substrate; d) Injection – compression molding material; e) Mirror block; f) Injection molding material [72]

The choice of a particular production process and disc material depends mainly on economic factors, i.e., the choice will depend on the state-of-the-art of the processes considered and on the availability of suitable materials.

The conventional production process for gramophone records is compression molding of a vinyl chloride – vinyl acetate copolymer. This process is unsuitable for making optical discs, because of the high birefringence of the discs, the local noncircularity of the tracks introduced on cooling, and problems with stamper contamination. The absence of birefringence is of particular importance because anisotropy of the refractive index of the substrate will change the plane (linearly) polarized laser beam used for read-out into an elliptically polarized one, which may cause severe disturbance of the read-out process.

Historically, photopolymerization (Fig. 55 A) was the first process used with LV discs because neither injection molding nor compression molding could be carried out with sufficient accuracy in the 1970s for making 300 mm diameter discs. In the late 1970s and early 1980s, injection molding of poly(methyl methacrylate) (PMMA) became feasible. PMMA injection molding has replaced photopolymerization, notably in mass production, for reasons of economy and simplicity (Fig. 55 B).

In the meantime, the 120 mm diameter CD was developed. CDs can be produced by injection molding and by injection – compression molding of polycarbonate (PC), because of the development of special grades of PC that allow the birefringence problem to be overcome (Fig. 55 C).

Compared to PMMA, PC has definite advantages with respect to moisture uptake and therefore dimensional stability. This is of special importance with present Compact Discs because they are single-sided. Uptake of water through the nonmetallized side of a PMMA disc leads to severe warpage.

With the double-sided video disc, water uptake is a smaller problem because of the symmetrical, sandwich-like structure of the disc. These discs can also be made from a blend of PMMA and poly(vinyl chloride) (PVC) [73]. DOR discs can have a glass or polymer substrate. For glass substrates, the preformatting (i.e., the provision of a pregroove and optionally clocking

and address information) is performed by means of the photopolymerization process (see Section Photoreplication Process). For thermoplastic substrates, injection molding is used to make the substrate and the preformatting information in one step. So-called mini-DOR substrates with 130 mm diameter are routinely made by injection molding of PC. For thermosetting materials, similar single-step processes have been described [74]. A possible alternative is the use of photosetting materials such as acrylates [75], [76].

2.3.2. *Material Specifications*

The specifications for discs and players are matched. The disc must therefore satisfy a number of requirements relating to optics, geometry, and stability. This will be illustrated for video (LV) and audio discs (CD).

Optical Properties. At least 70 % optical reflection of the substrate-incident laser beam is required for proper read-out of the stored information. Absorption and scattering losses that occur as the laser beam passes through the substrate twice depend strongly on the type and purity of the material; the same applies to the reflection at the metal – polymer interface. Only a limited number of combinations of substrate and mirror or reflective coating (and replication coating in the case of photopolymerization) materials satisfy the requirements for high reflectance. The substrate used with photopolymerization must also be transparent in the near UV (320 – 380 nm). Turbidity, microvoids, and dust particles must be absent, and the substrate must be optically homogeneous.

Another optical requirement is connected with the use of linearly polarized laser light in the optical system of the player for separating the incident from the reflected beam. Any change in polarization makes this separation less effective and causes a drop in detection efficiency and an undesired feedback of modulated light to the laser. Birefringence invariably changes the plane polarization of the incident laser beam into elliptical polarization.

For the video disc the birefringence must be < 20 nm for a single pass of a parallel beam of light. This imposes a serious restriction on the choice of substrate material and the method of production. For CD discs, the requirements are less severe: the birefringence must not exceed 50 nm for a single pass of a parallel beam of light.

Geometrical Properties. Specifications for discs and players are such as to ensure that any disc can be played on any player.

The thickness of the video disc and the CD have been set to 1.25 ± 0.10 mm and 1.20 ± 0.10 mm, respectively. Much thinner substrates are difficult to handle and to produce and are usually not stiff enough. Moreover, at a thickness of ca. 1 mm, scratches or dust on the outer surface of the disc are so far outside the depth of focus of the objective lens of the player that only tolerable signal degradation results.

The tolerances for flatness and parallelism are more severe than those for thickness. Any departure from planarity of the inner surface of the disc causes vertical movement of the information track. The high rotational speed of the video disc (25 or 30 revolutions per second for PAL or NTSC formats respectively) and the restricted acceleration, velocity, and amplitude of the focus control system impose restrictions on the tolerable local thickness variations. Tolerances depend on the radial position as well as on the wavelength of the disturbance. If a sinusoidal disturbance is assumed along the track of a video disc, it can be shown that, depending on the radial position on the disc, disturbances with a wavelength of 8 – 22 mm may cause vertical amplitude deviations no greater than 2 µm and still be tolerated. With increasing wavelength of the disturbance, the vertical tolerance also increases up to ca. 1000 µm. The tolerances with respect to radial displacements are smaller. Radial amplitude deviations of 0.1 µm can be tolerated for wavelengths of 4 mm at the inner radius of the active area of the disc and for wavelengths of 10 mm at the outer radius. If the wavelength exceeds 100 mm at the inner radius or 280 mm at the outer radius, radial amplitude deviations as large as 80 µm can be tolerated. The influence of radial and tangential variations and tolerances have been discussed in detail [77].

Stability. The optical and geometrical characteristics should remain unaltered under climatic conditions likely to be encountered during transportation, storage, and use. Desirable properties are high scratch resistance, toughness, and fracture strength. Dimensional stability under wide variations in temperature and relative humidity, a certain inertness with respect to cleaning agents, and the absence of metal mirror corrosion are also prerequisites.

Service life is tested in cyclic humidity tests in which discs are subjected to periodic alterations of temperature (usually from 20 – 45 °C) and relative humidity (usually from 50 – 96 %) [77].

Materials. Few substrate materials satisfy all the requirements mentioned above. In Figure 56, five polymers and an inorganic glass are compared qualitatively for those material properties that are most relevant to optical disc production and performance [78]. The larger the radius in the polar diagrams, the more favorable the respective property of the substrate material.

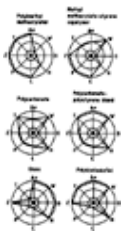

Figure 56. Qualitative comparison of substrate materials for optical discs. The larger the radius, the more favorable the indicated property of the substrate material. The properties shown are:

Δn, birefringence; W, water uptake; O, permeability of oxygen and water; H, heat distortion temperature; C, cost; P, ease of processing; E, tensile modulus; T, toughness [78]

For PC in particular, but also for other polymers, considerable work has been carried out to develop a suitable grade for optical disc manufacturing. Reduction of birefringence has required adaptation of the material and the processing used. The

properties of various substrate materials have been reviewed [78-81]. At present, PC and PMMA are the most widely used polymers. Properties of discs made from these materials are listed in Table 9 [79], [80].

Table 9. Material requirements and properties for video discs (PMMA) and Compact Discs (PC) [79]

Property	Video disc	CD, DVD
Optical		
Light transmission, %		
630 nm	≥90	≥90
300 – 380 nm	≥80 [a]	
Birefringence (optical path difference single pass [b]), nm	≤20	≤50
Optical defects >200 µm	none	none
Mechanical		
Thickness (nominal), mm	1.25 ± 0.1	1.20 ± 0.1
Thickness tolerance, µm	± 20 [c]	± 50 [d]
Distortion, degrees	± 0.4	± 0.6
Surface roughness, nm	≤10	≤15
Rockwell hardness	105	86
Impact strength, kJ/m^2	12	no failure
Thermal		
Vicat softening temperature, °C	96 – 107	145
Melt index, g/10 min	10 – 22 (230 °C)	30 (260 °C)
Physical-chemical		
Water absorption (100 % R.H.), (wt %)	2.1	0.4
Water diffusion coefficient (23 °C), mm^2/s	5×10^{-7}	48×10^{-7}

[a] For photopolymerization.
[b] Parallel light beam.
[c] Disc diameter 305 mm.
[d] Disc diameter 120 mm.

2.3.3. *Molding Processes*

For injection molding of optical discs, hot molten polymer is injected at high pressure into a thin cold mold cavity at a high flow rate through the gate located at the center of the cavity (see Fig. 55 B). A clean room environment is required. The molding process and the molding material have been optimized to attain the required birefringence for CDs of <50 nm (single pass for a parallel beam of light).

The birefringence of a disc is caused by two effects [82], frozen-in thermal stress and frozen-in orientation. Thermal stress occurs because of inhomogeneous cooling combined with a change of mechanical properties upon solidification. The amount of thermal stress is determined mainly by the difference between the mold temperature and the glass transition temperature T_g of the polymer.

The orientation of the macromolecules, which is caused by the shear and elongational stresses in the melt during mold filling, can be frozen-in by rapid cooling and vitrification. The amount of frozen-in orientation depends mainly on the injection temperature of the melt.

A special problem with PC is the high stress optical coefficient (SOC), defined as the ratio of birefringence to stress. The SOC of PC is more than ten times larger than that of PMMA. This causes the birefringence of PC substrates to be much more sensitive to internal stress than the birefringence of PMMA substrates. Much effort is therefore spent in structural modification of PC in order to reduce its intrinsic optical anisotropy [83], [84] and in blending materials with low stress optical coefficients of opposite sign [85].

Careful analysis of the birefringence distribution in CDs made of PC has revealed that both the radial and the tangential refractive indices vary in a complicated fashion throughout the thickness of the disc. The birefringence of a normally incident beam is equal to the difference of these two components. Therefore, the birefringence also varies strongly through the thickness of the disc (see Fig. 57) [86]. At the outer surface ($z = 0.6$), the sample is almost uniaxial; the birefringence nearly vanishes. Just below the surface, the birefringence becomes strongly negative, then increases to a positive value and next decreases to almost zero near the center plane of the disc. The observed overall birefringence is, therefore, determined by the compensation of two large contributions.

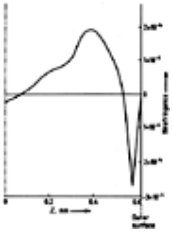

Figure 57. Birefringence of a CD in the direction normal to the disc (*Z*), determined at a distance of 35 mm from the center axis [86].

This explains why the observed birefringence is so sensitive to small variations in processing conditions. The art of processing lies in finding the right balance between the injection or filling stage and the compression stage of the injection molding process [86].

Important properties of the disc such as birefringence, residual stress (which causes birefringence), variation in thickness, and replication of the pit pattern from the stamper are critically dependent on the flow behavior of the molten polymer. The non-Newtonian dependence of the melt viscosity on the shear rate is therefore an important parameter. At the shear rates commonly used in injection molding, the dependence of the viscosity of PC on shear-rate is much smaller than that, for example, of PMMA. To obtain good moldability with PC, a relatively high temperature is required, but polymer degradation sets an upper limit to the molding temperature [87].

Special low molecular mass grades of PC have been developed to extend the processing window to lower temperatures and pressures while maintaining good flow properties [88-90]. The granulate polymer should contain neither gel nor dust particles.

In addition to processing conditions and material properties, the molding equipment has also been optimized with respect to product properties and cycle time [91], [92]. Without the increased precision, the injection molding of compact discs using PC would not have been possible. The injection molding process has been described in detail [92], [93].

Injection – compression molding (see Fig. 55 C) is distinguished from injection molding in that a heated mold is filled by injection of even hotter fluid polymer before complete closure of the mold occurs. Excess polymer is expelled from the mold (flash). This is removed after cooling. Unlike injection molding, polymer dosage is not critical.

The cycle time for CD injection molding is ca. 10 s. For injection – compression molding, the cycle time is somewhat longer, because the mold itself has to be heated up and then cooled before opening and removal of the disc. Since a heated mold is used, however, considerably less polymer orientation is frozen-in. Process conditions and mold tolerance are, therefore, less critical for obtaining acceptable birefringence than with injection molding. This is also reflected in the higher yield of the injection – compression molding process. Metallization and protective coating of the discs are described in Sections Metallization and Protective Coating.

2.3.4. *Photoreplication Process*

The photoreplication of LV video discs is shown in Figure 58 [77]. In the first step (Fig. 58 A), a liquid photosensitive coating C is deposited on the metallic stamper or mold Mo. The slightly curved substrate S (PMMA, 1.25 mm thick) is pressed onto the coated mold (Fig. 58 B). The latter forms a thin layer C (20 – 30 µm) between the substrate and the mold. The coating is then irradiated with an array of fluorescent lamps (λ = 350 nm) for 6 s. The substrate and the coating are then released from the mold (Fig. 58 C), inverted, and a metallic reflective layer M is deposited on top of the organic coating. The reflective mirror layer is covered in turn with an organic protective coating P (Fig. 58 D). Finally, the P-sides of two such discs are glued together to yield a double-sided video disc. The principles of photopolymerization and the formation of densely cross-linked, glassy networks have been described in detail [94].

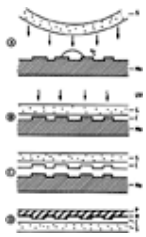

Figure 58. The 2 p photopolymerization process [77]

C = Replication coating; Mo = mold; S = substrate; M = Mirror; P = protective coating

The photopolymerizable coating used for copying the information pattern must have a high curing rate, a low viscosity, and it must wet the substrate surface. When cured, it must have high dimensional stability, durability, and lack of odor, easy release from the mold, and good adhesion to the substrate and to the reflective layer.

Coatings that meet all these requirements are not easy to find, mainly because some of the requirements are conflicting, e.g., easy release from the metallic mold but good adhesion to the metallic reflective coating.

Acrylates are preferred monomers because of their high rate of polymerization. The most efficient photoinitiator is 2,2-dimethoxy-2-phenyl acetophenone (DMPA). The absorption spectrum of DMPA strongly overlaps the emission spectrum of readily available fluorescent lamps [95].

The monomer coating's components were chosen with a view of desirable ultimate properties such as behavior upon metallization and adhesion to the metallic reflective coating.

Correlating the monomer components' chemical functionality versus their content of inactive, saturated hydrocarbon groups along with hardness, adhesive and release properties [95], [96] has made it possible to delineate a region on the overlapping maps where potentially useful monomers and mixture compositions can be found. One single monomer coating and a few mixtures have been selected for further investigation in this way (see Table 10).

Table 10. Video disc coating properties [95], [96]

Coating composition, wt % *			Viscosity 23 °C, mPa·s	Curing time, s 23 °C	Curing time, s 80 °C
1	TPGDA 57 %	NVP 29 %	7.8	1.5	1.0
	TMPTA 10 %	DMPA 4 %			
2	TPGDA 61 %	NVP 17.5 %	12.3	1.7	1.0
	TMPTA 17.5 %	DMPA 4 %			
HDDA	HDDA 96 %	DMPA 4 %	6.7	3.0	1.4

* DMPA = 2,2-dimethoxy-2-phenylacetophenone [24650-42-8]; HDDA = 1,6-hexanediol diacrylate [13048-33-4]; BDDA = 1,4-butanediol diacrylate; EHA = 2-ethylhexyl acrylate [103-11-7]; TPGDA = tripropyleneglycol diacrylate [42978-66-5]; TMPTA = trimethylolpropane triacrylate [3290-92-4]; NVP = *N*-vinyl-2-pyrrolidone [88-12-0].

An important parameter of the coating is its *hardness*. The influence of chemical composition on hardness and dimensional stability is shown in Figure 59 [96]. Measurement of coating hardness and its influence on signal-to-noise ratio have been described [97]. Hardness is related to the cross-link density. Cross-link density is related to the extent of double bond conversion. Unexpectedly, it turned out that at room temperature the highest cross-link densities can be obtained with the diacrylates [95].

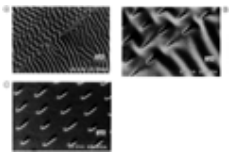

Figure 59. Scanning electron micrographs of the LaserVision video disc pit pattern after metal deposition on a coating of a mixture of EHA and BDDA. The composition of coatings in pictures (A) and (B) is 80 % EHA – 20 % BDDA. After metallization, the soft coating appears as a wave pattern, due to shear deformation caused by stresses in the aluminum layer. Picture (C) has the composition 20 % EHA – 80 % BDDA. There is no wave pattern because the cured coating is more densely cross-linked [95], [96]

After comparison of various parameters, 1,6-hexanediol diacrylate (HDDA) was selected as the single monomer coating. It has been further studied together with the mixtures given in Table 10. Parameters studied included curing time under process conditions, extractability of low molecular mass material, and the energy required for separation from the mold [95], [96] (see Table 11).

Table 11. Separation of video discs from a mold (constant pulling rate of 8.33 mm/s)

Coating *	Total separation, energy, J	Substrate deformation energy, J	Coating separation energy, J	Separation time, s	Maximum pulling force, N
1	1.27	0.39	0.88	4.2	51
2	0.75	0.27	0.48	3.6	48
HDDA	0.30	0.08	0.22	2.4	31

* see Table 10.

During release of the substrate with the cured information layer from the stamper or mold, the disc can be warped in a

complicated way. Part of the total separation energy is required for the actual separation (release of adhesion), the remaining part for the (in)elastic deformation of the substrate with the information-containing layer. Mold release is easiest for the single-monomer system, although satisfactory results were obtained for all three mixtures (Table 11) [95], [96]. Video discs of excellent quality have been made with both the single-monomer coating and with a number of mixtures.

2.3.5. *Metallization*

The metallic reflective coating should have a reflectance of at least 80 – 85 % (ca. 8 – 10 % is lost by scattering and by reflection at the air – polymer interface) because at least 70 % of the laser light must be reflected by the disc. For this reason, together with material cost, stability, and reactivity, the choice of material is limited to aluminum, silver, copper, gold, or selected alloys of these metals [98]. The reflectances of the pure metals in the near-IR, where the discs are read, are 85 %, 99 %, 95 %, and 97 % respectively. The thickness of the reflective layers is about one hundred nanometers.

Aluminum can be deposited easily by thermal or electron-beam vacuum evaporation and, in contrast to silver, it adheres well to the polymerized coatings, as well as to PC and PMMA. A disadvantage of this deposition technique is that it must be carried out in batches, which makes it difficult to integrate into the otherwise continuous production process. Each batch requires a relatively long time before the desired vacuum deposition pressure $<10^{-3}$ Pa is reached, notably when PMMA is used as a substrate. PMMA may contain 1 – 2 wt % water under ambient conditions [78], and this must be largely removed before vacuum deposition may be started. If the deposition is started at an insufficiently low pressure, a mirror coating with reduced reflectance is formed, owing to reaction of the growing layer with water vapor. The use of PMMA requires the design of special evaporation equipment with large liquid nitrogen-cooled surfaces. The problems are much less severe with PC, which absorbs much less water under ambient conditions.

For PMMA and PC, alternative techniques such as electroless silver deposition and magnetron sputtering have been developed [77]. *Electroless silver deposition* is quite similar to the process used for many years in the production of glass mirrors. The cured coating is simultaneously sprayed with an ammoniacal silver salt solution and a reducing agent. Adhesion to the organic surface is poor, as with an evaporated silver coating. A special pretreatment is therefore necessary [99].

Magnetron sputtering uses a cold, solid target that is bombarded with ions, usually argon(I) ions. At an argon pressure of ca. 0.1 Pa and a sufficiently high d.c. voltage a discharge is formed between the target, acting as cathode, and an anode. With a magnetic field, maintained by permanent magnets behind the cathode, a concentrated discharge plasma is produced immediately above the target surface. Argon(I) ions with an energy of 300 – 500 eV can be extracted from the plasma; these ions bombard the target surface and sputter its material.

The disc to be coated is situated directly opposite the target and outside the plasma region. The production rate is determined mainly by the sputtering time, not by the time taken for loading and unloading, so that d.c. magnetron sputtering of aluminum is also widely used in CD production. Sputtering also allows more freedom in the choice of the reflective mirror material because binary and even ternary alloys can be deposited in this way. Some of these alloys have desirable properties for optical disc applications, including high stability and low internal stress [98]. The layout of a magnetron sputter is shown in Figure 60.

Figure 60. Schematic of a magnetron sputter

2.3.6. *Protective Coating*

The metallized side of the video disc or CD is spin-coated with a solution of a protective organic coating such as nitrocellulose or with a UV-curable material. For Compact Discs, a label is applied by tampon printing to this side of the disc. The discs are inspected, tested, and packed for shipment.

For video discs, the cured protective coating is sprayed with an acrylic-based contact adhesive. After being dried, two discs are pressed together, care being taken that no inclusion of air occurs: this might introduce local variations in thickness, which could produce unacceptably large accelerations that are difficult for the focus-servo system of the video disc player to handle during playback. Moreover, deformation will also induce local stress, which may initiate degradation of the reflective metal coating. If necessary, the video disc is mechanically balanced by removing some material from the perimeter of the sandwiched double-sided disc. An adhesive and label are applied to each side and the discs are inspected, tested, and packed for shipment.

2.3.7. *Nonstandard Processes*

Photopolymerization of Complete Discs. Instead of using a thin photopolymerizable coating, it is also possible to cast the complete disc by photopolymerization of liquid monomers without using a substrate. In this case, provisions must be made to cope with shrinkage during polymerization. In one design, the liquid monomer is confined between a stamper or mold and a glass plate and irradiated through the glass. The glass plate is spring-loaded so that it can move towards the mold during polymerization and attendant shrinkage [75]. Another invention used a construction that allows for replenishment of liquid monomer during the polymerization and shrinkage [76]. The use of thermosetting polymers has also been proposed [74].

Embossing or Printing. A heated mold is pressed against the surface of a preformed polymer sheet, held in a press, for long enough to transfer the encoded information without deforming the sheet. Nonplasticized calendered PVC sheet and extruded PC film have been used [100-103]. A continuous web manufacturing process is possible. In addition to high throughput, this allows the use of a relatively thin and soft sheet or foil, which is then metallized and laminated to a stiff thick substrate such as PC. A further variation is the use of metallized polymer films [104-106], even directly in an injection-molding machine [107].

Photographic Reproduction. The master disc is converted into a photographic exposure or imaging mask by deposition of chromium on the developed areas, followed by removal by dissolution of the photoresist. A contact copy can then be made using a polymer substrate disc coated with a high-resolution photographic emulsion. Systems based on physical development (PD) are particularly suitable [108]. The disc with the contact image can be read in transmission or, alternatively, in reflection after coating with a reflective metal layer [109]. In another method, a reflective image is made directly during the PD process [110]. Although good discs have been obtained in this last way, the whole process of contact-copying suffers from sensitivity to dust and contamination of the imaging mask. It is also very complicated to create the imaging mask and the process requires the use of extremely smooth and flat polymer substrates. A noncontact holographic replication of CDs has been shown on photosensitive layers [111].

Photopolymer Reproduction.A photopolymer process, which also requires that the master disc be converted into an imaging mask, with the contact copy steps and difficulties as in photographic reproduction, is also possible. The contact copy is made using a polymer substrate disc coated with a high resolution photopolymer. The optical read-out signal is due to a spatial variation in the refractive index.

Photopolymers are polymers in which optical density and refractive index change (typically $\Delta n < 10^{-4}$) under the influence of light, mainly as a result of photopolymerization of dissolved monomers [112], [113], [114].

The most suitable materials are polyacrylates, polyacrylamides, polystryrenes, polycarbonates and their copolymers, mixed with various monomers, photoinitiators and photosensitizers [80]. PMMA is the preferred matrix polymer. Pattern-wise irradiation causes a local depletion of unreacted monomer, which is replenished by diffusion from the unexposed areas. Next, a flood exposure is applied in order to fix the refractive index pattern induced by the diffusion of the monomer [112], [113]. Especially large local changes in refractive index ($\Delta n > 10^{-2}$) have been obtained in styrene-doped PMMA, exposed to UV radiation (λ > 280 nm) [115], [116]. Green light (λ = 514 nm) can also be used for optical recording in a layer where titanocene chloride is the photoinitiator [116]. In addition to refractive index changes, photomechanical materials exist in which relief structures can be attained with a spatial resolution <1 µm. This is possible by the local photopolymerization of styrene, dissolved in PMMA, followed by removal of the unreacted monomer [115-117]. As mentioned earlier, PMMA is not an ideal substrate for single-sided optical discs, primarily due to its high water uptake [78].

2.4. Write-Once Optical Recording Materials

Write-once recording mechanisms can be divided into those that involve macroscopic material transport and those that do not. These mechanisms are illustrated in Figure 61. All these mechanisms are of the heat-mode category, that is, being induced by the heat generated from the optical absorption spectra of the recording materials.

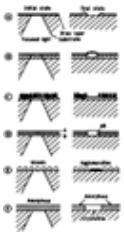

Figure 61. Write-once optical recording mechanisms

A) Hole burning (ablation); B) Bubble formation; C) Texture change; D) Alloying of bilayers; E) Island agglomeration; F) Phase change ■is this figure still relevant? ■

2.4.1. *Material Specifications*

Write-once optical discs are generally used for the archival storage of information. A prime prerequisite, therefore, is chemical and physical stability of the recording media and the written information. It is not sufficient that the substrate and the films themselves are stable because interfacial phenomena often have a very detrimental effect on media stability. Furthermore flat, homogeneous discs are required, without pinholes, cracks, or blisters. The disc must also have sufficient reflectivity at the wavelength of the laser used, in order to focus and track the pregroove structure and to read out the stored information.

For optical recording the laser must produce permanent optically-detectable changes in the sensitive layer of the discs. These changes should be formed with the pulse powers and pulse durations available from a semiconductor diode laser. Typically, a pulse duration of 20 – 100 ns and a power up to 10 mW are used. A temperature of the order of 1000 – 1500 °C can thereby be attained [118-123].

A large difference in reflectivity is required between marked (written) and unmarked (unwritten) areas in order to obtain a large read-out signal. A high signal-to-noise ratio is desired which implies both a large reflection change and a low noise level. Contributions to the noise include irregularity in the size and position of the bits (write noise) and the microstructure of the film (disc noise). The bit error rate or probability is the ratio of bits erroneously detected to the number of written bits. For a practical storage system, the uncorrected or raw bit error rate is more important than the noise. Usually, a value of $<10^{-4}$ is desired [124].

Read-out is usually performed with the same laser as used for writing, but at a much lower power level. The heating caused by the reading laser should not induce changes in the film or data, implying that mark formation must have a threshold temperature.

To attain high sensitivity, very thin, strongly absorbing films are primarily used as active layers. The complex refractive index of a material is given by the equation:

$$\bar{n} = n - ik$$

where *n* is the refractive index and *k* is the absorption index [125]. Then, the absorption length, α, is related to *k* by:

$$\alpha = \lambda / 4\pi k$$

The generation of heat will usually be most effective for a layer with a thickness less than α. With thicker layers, deeper portions receive less light and are hardly heated.

In the optical design of a disc, interference effects must be taken into account to calculate optimum layer thicknesses [125], [126]. A compromise often has to be found between a high reflectivity and a low absorption. The dependence of reflectivity and transmission on layer thickness for an organic dye is shown in Figure 62.

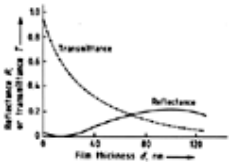

Figure 62. Dependence of reflectance (substrate-incident) and transmittance of a near-IR dye (λ = 800 nm) on layer thickness [119].

The reflectivity, absorption, or contrast can be optically tuned by using multilayer configurations; this is more complex, and therefore expensive [126]. This concept is illustrated in Figure 63. A disadvantage of such constructions is that they are relatively difficult to produce because of the often very narrow tolerances on the thicknesses of the layers. Extra interfacial problems can also arise. In some applications, however, (see Section Magneto-Optic Materials), such structures must be used to enhance the read-out signals to the necessary minimum level.

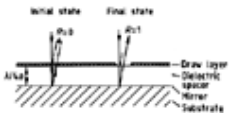

Figure 63. Principle of the anti-reflective trilayer structure [126].

2.4.2. *Deposition Techniques*

The thin films used in optical recording are generally metals, semiconductors, dielectrics, and organic materials. Inorganic films are deposited by vacuum techniques such as thermal, flash- and electron-beam evaporation [127], and d.c. or r.f. sputtering [128]. The evaporation techniques are all performed under high vacuum. The differences are mainly the methods used to vaporize the materials.

Alloys are difficult to deposit by thermal evaporation because of the different vapor pressures of the components; sputtering is the preferred technique. A further advantage of sputtering over thermal evaporation is the ability to deposit many materials with a high accuracy in composition. Moreover, this technique is well suited for mass-production.

Organic layers are usually applied by means of spin solvent coating [129]. A solution of the organic material is deposited on a slowly rotating disc, which is then rotated at a high speed. The solvent evaporates and the film is thereby formed.

2.4.3. *Tellurium Alloys*

The conceptually simplest optical recording mechanism is heating the film with the focused laser beam and melting the film locally, followed by ablation of the material. The advantages of this mechanism are that the contrast is large and the process is self-regulating. Ablation is the most widely used write-once optical recording mechanism. The most important write-once optical recording materials are alloys consisting mainly of tellurium.

Tellurium has high absorption and reflection at 800 nm, a low melting point [124], and an acceptable chemical stability, at least when it is alloyed with other elements such as selenium. Materials used include Te – Se – Sb – S [121], [124], [130], [131] and Pb – Se – Te alloys [132]. Some indications of the selection procedure for write-once materials will be given by treating the tellurium alloys in detail.

If laser heating is to be efficient, the layer must be as thin as possible [124]. At a layer thickness <20 nm, however, the films become inhomogeneous, prone to oxidation, and the reflectivity becomes very low [127]. Pure tellurium films are easily oxidized, resulting in a rapid decrease in absorption and reflection. The oxidative stability of tellurium, as extrapolated from a thermal cycling test in a high R.H. environment, the so-called Z/AD test, can be increased by at least four orders of magnitude by adding selenium (see Fig. 64) [130]. In this test, the films are cycled between 25 and 65 °C at 93 % relative humidity [133]. Adding more than 15 atom % selenium results in decreasing absorption at wavelengths in the near-IR with hardly any increase in corrosion resistance.

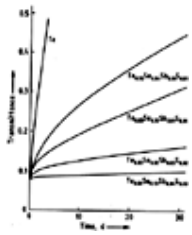

Figure 64. Transmittance (750 nm) of Te – Se alloys as a function of time in the Z/AD test [130].

The films are usually amorphous when deposited and crystallize rapidly at ambient or slightly elevated temperature [130]. The accompanying contraction is appreciable, resulting in stress and sometimes even cracks in the film. Adding antimony has the advantage of inhibiting the crystallization (which is still rapid at ca. 90 °C) and simultaneously reducing the volume contraction. The addition of 3 atom % antimony appears to be sufficient to prevent the above problems.

The advantage of antimony over other elements such as indium, thallium, germanium, tin, lead, arsenic, or bismuth, which also reduce stress, is that antimony inhibits corrosion, gives more well-defined pits, and does not form volatile products at 90 °C [130]. If more than 5 atom % antimony is added, then the crystallization temperature becomes excessively high. With 3 – 5 atom % antimony and preferably 1 – 2 atom % sulfur, the activation energy for the amorphous-to-polycrystalline transition is such that a reasonably rapid transition occurs at 90 °C (typically within 15 min). The optimum composition appears to be $Te_xSe_ySb_qS_z$, with $0.81 < x < 0.85$, $0.01 < y < 0.13$, $0.03 < q < 0.04$, and $0.01 < z < 0.02$.

Pit formation is rather complicated [121], [122], [134], [135]. It depends on alloy composition, film thickness, and also on the substrate surface [134], [135]. The pits usually consist of a hole surrounded by a rim containing most of the removed material with some pit material ejected as debris. A prerequisite (though not a sufficient condition) for pit formation is that the layer is melted locally by the laser pulse [122], [124]. For long laser pulses (1 μs), the temperature required for pit formation appears to be only slightly above the melting point. For short laser pulses, a temperature close to the boiling point is required [122]. The influence of the substrate is also extremely important. Apart from the thermal properties of the substrate, the interfacial energy between the substrate and the film is important. Lowering the interfacial energy by using fluorinated polymers in the substrate lowers the threshold energy, but the stability of the film also decreases appreciably [134], [135].

Optimization leads to discs with high carrier-to-noise ratio (CNR) values, for pulse energies of the order of 0.6 nJ. The stability (as determined from the Z/AD test) of hermetically sealed, air-sandwich glass disc constructions is at least 20 years. The air-sandwich prevents moisture from entering the 2 p lacquer, which could cause leaching of sodium from the glass into the tellurium alloy active layer. Such leaching of sodium would result in an undesirable chemical degradation [130].

2.4.4. *Organic Dyes*

Organic dye films used as optical recording media in the ablative writing mode have advantages over tellurium alloy films. Organic dye films combine high absorption coefficients with a low decomposition or melting temperature and low thermal conductivities [119], [120], [136-142]. Moreover, films of organic dyes are usually applied by means of solvent spin-coating [129], which is simpler than the vacuum techniques used for inorganic films. These potential advantages were recognized early. Problems with organic dyes are their limited thermal and ambient light (daylight) stability, and the relatively narrow absorption bands compared with inorganic materials.

In early work, before the advent of reliable, economical, high-power semiconductor lasers operating in the near-IR region, dyes absorbing in the visible region were chosen [136], [137]. Visible dyes are used in high data density applications. Usually, bi- or trilayer structures were used to obtain sufficient reflection [137], [138] (see Fig. 63). The dye (up to 30 wt %) is dissolved in a polymer matrix. At higher concentrations, phase separation and dye crystallization occur. Polymeric dyes can also be used. For both types of dye system, stable, sensitive media can be designed [139].

The threshold energy for laser marking appears to be determined only by the thermal efficiency of the organic dye layer. The viscosity does not appear to be important [142]. The CNR, however, depends on the polymer. The higher the viscosity, the shallower the pits and hence the lower the contrast and CNR [140], [142]. Furthermore, diffusion of the dye into the substrate sometimes plays a role in the bit formation process [136].

The most important pit formation mechanism is probably melting of the film, followed by partial decomposition or evaporation. The gaseous components exert a recoil pressure on the molten material and push it into the rim [138], [143], with some pit material ejected as debris. The resulting pit shape is somewhat different from that observed with tellurium alloys (see Fig. 25, Section General Introduction).

For semiconductor laser wavelengths, dyes absorbing in the near-IR are needed. They usually consist of conjugated polymethine chains [143], [144]. These dyes are often susceptible to oxidation. Certain tetraarylpentamethines have sufficient thermal and ambient light stability. The choice of anion is also important, both for stability and for solubility. The solubility in the spin-coating solvent should be at least 1 – 2 wt % to obtain a sufficient layer thickness. The solubility in water should be very low to ensure long-term integrity of the dye layer. For example, chlorides dissolve readily in alcohols and water, but they are often unstable.

Another type of dye that can be used is the squarylium class [119], [120]. The absorption maximum is usually near 750 nm. Substitution of the end groups by thiopyrilium units shifts the absorption maximum to 820 nm. With bulky substituents such as *tert*-butyl groups, the solubility can be increased to 10 – 20 wt % in *n*-propanol.

For semiconductor laser write-once recording, pure dyes are coated in a single layer on a pregrooved substrate. In this way, maximum dye loading is obtained. The reflectivity of these films, substrate-incident, is 15 – 20 %, sufficient for tracking, focusing, and reading. The near-IR dye films appear to be up to 50 % more sensitive than tellurium alloys [119], [145]. The organic dye discs can be made using a hermetically sealed air-sandwich construction or with an in-contact protective overcoat.

Their high CNR values are, however, much more important. The well-defined, reliable, and reproducible pit shape and position are also important. These properties make organic dye media well suited for high density data and video recording [119], [138]. Written data can be read at least 10^5 times without deterioration at normal reading power [119]. Furthermore, long term stability of up to 50 years is claimed [145]. A metal-overcoated organic dye system which combines a high initial reflectivity (> 70 %) with a high contrast such that a written disc could be read by a standard CD player [146].

For DVD–R and DVD+R dyes absorbing in the red (650 nm) are used. Because of the high demands for dyes in high-speed

DVD–media every manufacturer uses his own special dye. In Table 12. an overview of some actually used dyes is given.

Table 12. Overview of some actually used dyes for CD-R and DVD+/–R

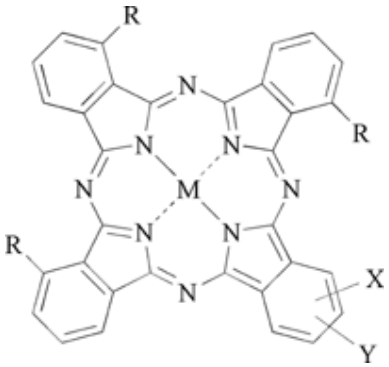	Phthalocyanine (780 nm, CD-R)
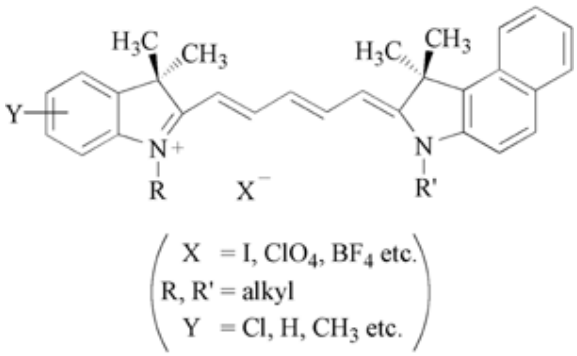	Cyanine (780 nm, CD-R)
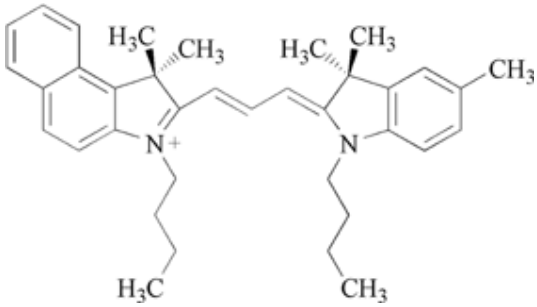	Cyanine (650 nm, DVD+/–R)
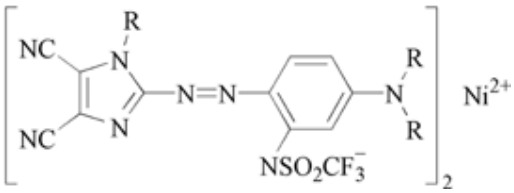	Azo metal chelate (650 nm, DVD+/–R)
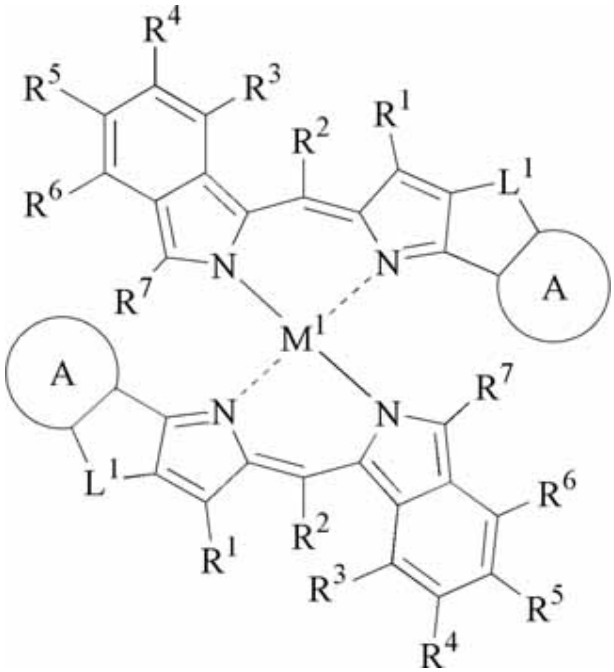	Dipyrromethene metal chelate (650 nm, DVD+/–R)

2.4.5. *Metal Particle Colloids*

Small metal particles dispersed in a gelatin matrix can also give a highly absorbing film with low thermal conductivity [147], [148]. The gelatin is loaded with filamentary silver grains (to absorb the laser light strongly) and with circular silver grains (to enhance the reflectivity). A reflectivity of 50 % can be attained with this design [148]. The writing mechanism is based on the absorption of the laser light by the silver particles, which heats and melts the gelatin film, resulting in a surface deformation that can be read optically. The maximum CNR of the media, however, is not very high.

The colloidal film, on the other hand, is very sensitive; maximum CNR is obtained for a very low energy density (0.7 nJ/μm^2). This film can be overcoated with an in-contact protective lacquer. An interesting application is their use as an optically recordable flexible card [148].

2.4.6. *Metal-Polymer Bilayers*

In metal deformation optical recording, a thin absorbing metal layer is plastically deformed by the gaseous products formed by heating an underlying polymer layer [149]. The laser heats the metal film, and, by thermal conduction, the polymer bottom layer also heats up and starts to decompose. When the gas pressure exceeds the elastic limit of the metal film a bubble is formed that remains intact after the bilayer film is cooled. A similar marking mechanism has been found in films which are composed of tellurium crystals dispersed in an organic polymer matrix [150], [151].

Material flow is not desired, only decomposition of the polymer. High molecular mass polymers are therefore used to obtain a high softening temperature. A practical molecular mass is in excess of 80 000. Moreover, it is possible to spincoat flat, high molecular mass polymer films. The polymers used are poly(methyl methacrylate) and polystyrene.

The metal films are required to withstand the large deformation without breaking at the speed of bubble growth (up to 10 m/s). This demands that the metal has high ductility, break resistance, and resilience. The metal layer should be continuous, adhere to the polymer film, and be resistant to corrosion. Typical metals used are gold or platinum-based alloys with <25 % additional elements.

Reproducible, well-defined bubbles can be written with reasonable laser power leading to a relatively high CNR and low bit error rate. Furthermore, the long term stability is claimed to be high [149]. One problem of this recording mechanism is that it is not self-limiting. Only a rather narrow range of laser power gives high CNR values. If the energy is increased too much, the metal film breaks and irregular pits result. These metal – polymer bilayer films are used in hermetically-sealed air-sandwich disc designs.

2.4.7. *Alloying Bilayers*

Another writing mechanism involving no macroscopic material transport uses the alloying or reaction of thin films, resulting in a layer with different optical properties [157], [158]. In the proposed system, a complicated quadrilayer construction is needed. An amorphous Sb_2Se_3 layer, followed by an amorphous Bi_2Te_3 layer, is deposited on the substrate. These two layers react by laser heating to form a four element alloy. The third layer consists again of a second amorphous Sb_2Se_3 layer. This layer is used as a thermal barrier and to tune the optical properties of the quadrilayer to give a high signal. The fourth thin film is an aluminum reflective layer. The reflectivity, substrate-incident, is chosen to be as low as possible considering the tracking and focusing needs (10 – 14 %).

The recording sensitivity appears to be high: 6 mW at 5 m/s and the CNR is reasonably high. The long term stability is expected to be excellent. Diffusion between the two thin film materials is negligible at ambient temperature.

2.4.8. *Agglomeration*

A third example of nonablative optical recording is agglomeration of very thin films of noble metals such as gold [118], [159], [160], silver [161] or tellurium alloys such as Te – Cu [162], [163]. The noble metal films are typically 3 nm thick, whereas the films of Te – Cu are typically 10 nm thick. The noble metal films, as deposited, are not uniform. They form very small islands with radii of 5 – 10 nm. The Te – Cu films are uniform.

Laser heating close to or above the melting point induces coalescence of the noble metal islands into larger islands up to four times their original size. In the Te – Cu uniform films, laser heating induces coalescence thereby producing a discontinuous film with very small islands within the marks. These coalesence or agglomeration mechanisms induce a change in the optical properties, which can be detected. To optimize the read-out contrast, an absorption trilayer configuration is used in the media design.

The advantage of the films is that they are approximately twice as sensitive as ablative tellurium or tellurium alloys on identical substrates. They are useful for high data rate applications. The media can be overcoated with in-contact protective organic lacquers.

2.4.9. *Surface Texture*

When a textured surface or interface is locally heated by a laser above the melting point, the molten area will resolidify under the influence of surface tension as a flat domain. The textured surface absorbs light strongly because its average periodicity is much smaller than the wavelength of the incident laser light. The smooth domain area reflects strongly when the material is a good reflector (see Fig. 61 C). The optical contrast can be high, and the disc noise may be often high.

The most investigated material for producing surface textured media is germanium [164]. The texture periodicity is ca. 40 nm. High reflection contrast is obtained with ca. 3 nJ of input energy.

Another material is a textured polymer layer (moth-eye) coated with a metal reflector [165]. In the textured polymer layer, the mark is formed not by melting, but by plastic deformation (blistering) of the polymer at high temperature. Moderately high

CNRs and good bit error rates are obtained.

2.5. Reversible Optical Recording Materials

In reversible optical recording materials, the laser beam used for recording must produce reversible, optically detectable changes in the sensitive layer of the disc. These can be changes (1) in the direction of polarization, (2) in the intensity, (3) in the phase of the reflected laser light, or (4) a combination of these. At present, the technologically most advanced erasable optical storage technologies are magneto-optic (MO) and phase-change (PC) recording.

Read-out in *magneto-optic recording* is based on the magneto-optic Kerr effect, which causes a rotation of the direction of polarization of the laser light when it is reflected at the perpendicularly magnetized recording layer. Magneto-optic materials have been investigated since the late 1950s; generally, amorphous rare-earth transition-metal alloys (RE-TM) such as Tb – Fe – Co and Gd – Fe – Co are used.

Favorable aspects of MO recording are its high recording density, which can be increased beyond the diffraction limit by using magnetic super-resolution techniques [166-168], and its high data rate, which is not intrinsically limited by the recording mechanism. Additionally, the overwrite repeatability (i.e., the number of overwrite cycles) of MO media is almost unlimited. Disadvantages of MO recording are its rather complex media (complex thin-film stacks, tight optical tolerances), and the requirement of polarization optics and (in some cases) a magnetic head in the drive. The compatibility with read-only drives and discs is therefore rather limited.

Phase-change recording is based on the change in the intensity reflection coefficient when the recording material undergoes a crystalline-to-amorphous phase transition. Phase-change materials for optical recording have been studied since 1970. Typically, chalcogenide alloys such as Ge – Sb – Te or Ag – In – Sb – Te are used.

The signal amplitudes in phase-change recording are much higher than in MO, due to the large reflectivity difference between the amorphous and crystalline state. The detection of nonpolarized, intensity-modulated light allows for compatibility with read-only discs and simplifies the optics in a phase-change optical head, leading to a low cost drive. Direct overwrite is easily realized by modulation of laser power. Disadvantages of phase-change recording are its inherently poor write power sensitivity, due to the high melting temperature of the phase-change material, and the limited overwrite repeatability, due to degradation of the media. The data rate is closely related to the crystallization rate of the phase-change material, which is limited by atomic mobility.

The aforementioned recording mechanisms are induced by the heat generated from the absorption of optical energy by the recording material. A second category is formed by *photon mode recording mechanisms*, which involve electronic transitions in the optical absorption spectra. Examples of such recording technologies are photochromic and photochemical (spectral) hole burning. Photon mode materials for optical recording have been investigated since the late 1970s. However, the progress in this field has been relatively slow and is still far from applications for the mass consumer market.

2.5.1. *Magneto-Optic Materials*

History. Magneto-optic (MO) recording has been commercially available longer than any other reversible optical storage format. The first generation of the ISO 5¼ inch (130 mm) MO drive was introduced in 1988 and offered 650 MB capacity on a double-sided disc (i.e., 325 MB per side). The 3½ inch (90 mm) MO drive arrived on the market in 1991 with 128 MB capacity. Since then considerable progress has been made in improvement of storage capacity, data transfer rates, and cost : performance ratio. Current (1998) 130 mm drives, the so-called 4X MO, have a capacity of 2.6 GB (1.3 GB per side), while the 90 mm format offers 640 MB per disc [169]. The other major MO storage product family is the Sony MiniDisc system, introduced in 1992 and designed primarily for digital audio applications [170]. In 1993 a version of this for data storage, MD DATA, entered the market with a capacity of 140 MB on a 2½ inch (64 mm) disc. MD DATA2, with a capacity of 650 MB, identical to that of a standard CD-ROM, has now been announced [171]. Looking to the immediate future, a number of industrial companies have formed the Advanced Storage Magneto-Optical (ASMO) consortium with the aim of producing a MO system with 6 GB capacity on a 120 mm single-sided disc, alongside compatibility with the popular DVD and CD formats [172]. One fundamental problem with such high storage densities is that the size of the bit that is recorded in the MO disc is actually too small (typically less than 0.2 µm in length) to be read back using conventional optical techniques. To overcome this limitation the ASMO format proposes to read the data using a form of magnetically induced super-resolution (MSR) [172], [173].

Paralleling these increases in storage capacities, MO recording has shown similar improvements in data transfer rates (i.e., how quickly data is transferred between the host system and the disc drive). However, unlike for conventional magnetic recording, it is not straightforward to implement a direct overwrite (DOW) scheme in MO recording. Rather, old data is first erased in one pass of the optical head before new data is written on a second pass. This obviously slows down data transfer, and much effort has been put into finding reliable MO – DOW methods. The scheme adopted by Sony for their MiniDisc system is essentially a form of thermally assisted magnetic recording known as Magnetic Field Modulation (MFM) [170], [171]. Another approach, realized in the laboratory over a decade ago but only recently becoming a commercially reality, is Light Intensity Modulation Direct Overwrite (LIMDOW) [174], [175]. Ways of combining these direct overwrite techniques with super-resolution read-out are being developed [176].

All of the above developments place special materials requirements on the storage medium. The materials that have proved consistently capable of meeting these requirements are based on alloys of rare earth (RE) and transition metal (TM) elements, RE – TM films for short. All MO systems currently available commercially, and those under commercial development, use RE – TM films as the storage medium. This does not mean that other materials are not suitable for MO storage applications. Indeed, much research has been carried out into alternative MO media, including platinum – cobalt multilayers and alloys for use with short-wavelength lasers [177], [178], garnet films for enhanced stability [179], [180], manganese-based materials for high read-out signal levels [181], and, perhaps most exotic of all, europium and uranium chalcogenides with enormous Kerr rotations [182]. One reason for the continued dominance of RE – TM alloy films is that their magnetic, thermal, and optical properties can be varied in a controlled fashion by varying the alloy composition, the

deposition conditions, or both. Thus film properties can be tailor-made to suit a particular recording technique, be it conventional MO recording or more advanced formats such as MFM, LIMDOW, or MSR [183].

Recording Principles. A more accurate name for MO recording is thermomagnetic-magneto-optic (TMO) recording. This reflects the fact that the writing of data to the disc is effected by thermal and magnetic effects, whereas the read-out of data from the disc relies on magneto-optic effects. The principle is as follows. A small region of a uniaxial magnetic film with the easy axis perpendicular to the film plane is heated to a critical temperature, invariably the Curie temperature, at which the coercivity falls to a very low value. The film then cools to ambient temperature in an external magnetic field. Upon cooling, the heated region acquires the magnetization direction of the external field. The state of magnetization of the neighboring unheated regions is not affected by this reversal. Provided that the coercivity at ambient temperature is sufficiently large, the written magnetic domains, or bits, are stable even when the applied field is removed. This process is reversible and repeatable; hence, MO recording is both erasable and rewritable [184]. The source of heat is a high-power, solid-state, laser focused to a diffraction-limited spot of about 1 µm diameter and modulated by the information signal. The magnetic field is supplied by a small electromagnet or rotatable permanent magnet.

Reading of the stored information is performed using the polar magneto-optic Kerr effect. The laser is again focused on the MO film, this time at low power. The state of polarization of the beam reflected from the disc is dependent on the magnetization direction, which in turn depends on whether a binary '1' or '0' has been written. A bipolar signal representing the two states of magnetization is extracted by analyzing the polarization direction of the reflected beam using polarization sensitive optics in the read-out head [185]. A typical MO drive is shown in Figure 65.

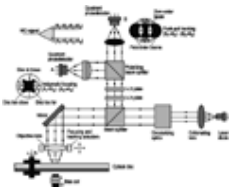

Figure 65. A typical magneto-optical disc drive with astigmatic focus and push – pull tracking.

Magneto-Optic Materials Requirements. To be a candidate for MO recording, a material must satisfy certain magnetic, thermal, and optical requirements. For conventional MO recording the fundamental requirements include: (1) perpendicular magnetic anisotropy, (2) a large coercivity at ambient temperature (to provide stability against stray magnetic fields and allow for high recording densities), (3) a temperature variation of coercivity that allows writing and erasing at the moderate temperatures achieved by heating with short duration pulses from solid state lasers, and (4) sufficiently large magneto-optic effects (Kerr rotation) to generate a usable read-out signal. Additionally the recording medium should be environmentally and chemically stable, homogeneous, low-noise and be suited to commercial production. It has not proved possible to meet all these requirements using a single layer recording medium, and so the usual approach is to combine active, protection, reflector, and enhancement films in a quadrilayer disc structure, as shown in Figure 66. Recording and read-out is achieved by focusing through the substrate. This provides immunity to dirt and scratches on the disc surface and helps to ensure that the disc is removable, one of the main assets of optical recording.

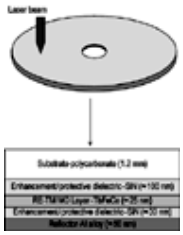

Figure 66. A typical quadrilayer disc structure based on RE – TM alloy as the active MO layer.

RE – TM Materials. The first reported studies of MO recording took place in 1957. This pre-dates the invention of the laser and a thermal pen was used to write magnetic domains in a Mn – Bi film, with reading achieved by white light in a polarizing microscope. Mn – Bi had suitable magnetic properties and large magneto-optic effects. However, Mn – Bi is thermally unstable as it exists in several crystalline phases and its Curie temperature is close to its melting point. It also has a polycrystalline morphology with relatively large grain size, resulting in high media noise levels that in turn results in a low signal-to-noise ratio. In the beginning of the 1990s Mn – Bi films have been revisited, in particular by studying the effects of additives — notably aluminum to form Mn – Bi – Al — for the purpose of reducing grain size, eliminating temperature-induced phase transitions and enhancing MO effects [181], [186].

The major breakthrough in the development of MO recording materials came in 1973 when IBM researchers discovered a new class of MO materials, the amorphous rare earth-transition metal (RE – TM) alloys [187]. These overcame many of the shortcomings of early materials, and are nowadays the only media used in commercial products. A host of binary, ternary, and quaternary RE – TM alloy compositions have been investigated. In cases where the RE constituents are from the heavy RE metals (Gd, Tb, Dy, Ho, Er, Tm, Yb, and Lu), the magnetic moments of the TM and RE subnetworks tend to couple antiferromagnetically, resulting in ferrimagnetic behavior. Where light RE elements (La, Ce, Pr, Nd, Pm, Sm, and Eu) are involved, the subnetwork coupling is generally ferromagnetic. For commercial applications the TM elements of choice are Co and Fe, whilst the RE elements are Tb and Gd, sometimes also Dy. In these cases, ferrimagnetic behavior is observed and the magnetic and magneto-optic properties vary as a function of temperature as shown schematically in Figure 67.

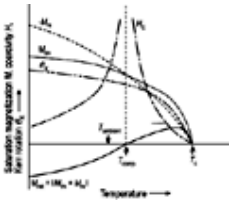

Figure 67. Variation of the magnetic and magneto-optical properties of antiferromagnetically coupled amorphous RE – TM materials as a function of temperature.

The RE and TM subnetwork moments, M_{RE} and M_{TM}, are equal in magnitude at the compensation temperature T_{COMP} where, since M_{RE} and M_{TM} are coupled antiferromagnetically and $M_{NET} = |M_{TM} + M_{RE}|$, the net magnetization is zero. Above T_{COMP} the net magnetization initially increases and then goes to zero again at the Curie temperature, T_C, where the material becomes paramagnetic [188].

Straightforward binary RE – TM alloys are not very suitable for MO recording. For instance, the Curie temperature of Tb – Fe is too low, whereas Gd – Fe has a very high T_C in addition to an in-plane anisotropy. Consequently, practical MO RE – TM materials are invariably ternary or quaternary alloys. Adding Tb to Gd – Fe, for example, increases the perpendicular magnetic anisotropy and the coercivity, whilst at the same time decreasing the Curie temperature. Adding Co to Tb – Fe increases both the Curie temperature and the Kerr rotation [189], [190]. The ability to vary magnetic properties over a wide range by changing the RE – TM alloy composition provides enormous flexibility in material design, and is one of the main reasons for the continuing predominance of RE – TM media. The compensation temperature is determined primarily by the ratio of RE to TM in the film, with changes of a few percent shifting T_{COMP} by more than 100 °C [183], [191]. Curie temperature depends mainly on the Co : Fe ratio; for example T_C varies from 135 °C for Tb – Fe to over 400 °C for Tb – Co [192]. Coercivity (H_C) and anisotropy (K_U) are strongly affected by the percentage RE content and the deposition conditions. A materials design strategy can be summarized as [183]:

1. Determine T_C by selecting Co : Fe ratio;
2. Determine T_{COMP} by selecting RE : TM ratio;
3. Determine K_U and H_C by deposition conditions and selecting Tb : Gd ratio (if appropriate).

Typical useful compositions in conventional MO recording applications are $Gd_{20}Tb_5Fe_{75}$ and $Tb_{27}Fe_{65}Co_8$ [193]. Additional elements, such as Al, Cr, Ta or Ti, may be included to improve the oxidation, corrosion and pitting resistance of the RE – TM film [194], [195]. However, the primary environmental protection for the RE – TM film is provided by the two dielectric layers between which it is sandwiched (see Fig. 66). These are usually aluminum or silicon nitride (AlN or SiN) and without them the lifetime of the RE – TM film can be measured in hours or even minutes [196].

The magneto-optic effects in RE – TM films are primarily brought about by the TM subnetwork, and only indirectly dependent upon the RE content. Incoming photons excite the 3*d* valence electrons of the transition elements altering their orbital angular momenta. Through the spin-orbit interaction, photons that are re-emitted (reflected) when the excited electrons relax are polarized. For perpendicularly magnetized material and linearly polarized incoming light, this process manifests itself as the polar Kerr effect. In general, the Kerr rotation decreases with increasing RE content. Also, as increasingly heavier RE elements in the series Gd, Tb, Dy, Ho, are used, a monotonic decrease in Kerr rotation is observed [183]. The intrinsic Kerr rotation angle for all RE – TM compositions of practical interest is small, having absolute values at visible wavelengths of less than 0.5° [197]. Furthermore, the rotation decreases with decreasing wavelength. Since future optical recording systems are expected to move to shorter, blue, wavelengths where the size of the focused laser spot is smaller and thus the storage densities are larger, this decrease of Kerr rotation is of some concern and has led to much research into alternative short-wavelength materials.

Platinum – Cobalt Multilayer Films. Of the materials considered as replacements for the RE – TM media, the most promising are those of platinum-cobalt multi-layered thin films made of a stack of alternated Co and Pt layers, as shown in Figure 68. These can be produced with perpendicular anisotropy and exhibit a Kerr rotation that increases considerably as the wavelength is reduced. Although extremely thin, Pt – Co multilayer films are remarkably stable on exposure to air and provide an inherently stable and simple disc structure. The films exhibit ferromagnetic behavior and are nanocrystalline.

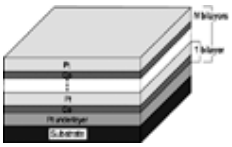

Figure 68. Structure of a Pt – Co multilayer, $N \times (x\ Co + y\ Pt) + z\ Pt$ where N is typically 15, x is ≈ 0.4 nm, y is ≈ 1 nm and z is ≈ 1 to 5 nm

The development of Pt – Co followed studies in the mid-1980s of the use of Pd – Co multilayers for perpendicular magnetic recording applications [177]. Perpendicular anisotropy occurs for cobalt layer thicknesses < 0.8 nm, where the interface anisotropy dominates over volume anisotropy [198], [199]. Optimum magnetic properties are found for Co layer thicknesses in the range 0.3 to 0.6 nm and Pt layer thicknesses in the range 0.8 to 2 nm. The number of bilayers are typically in the range 10 to 15, giving an overall film thickness of around 10 to 40 nm [200], [201]. Use of Pt – Co multilayers as MO recording media was demonstrated for the first time in 1989 [202], [203], [204]. Curie temperatures are typically around 350 °C to 400 °C, somewhat higher than for RE – TM films and necessitating higher laser record powers. T_C can be controlled to some

extent by variation of the Pt and Co layer thicknesses, by the addition to the Co layer of a few percent of Os or Re or by using Co*X* – Pt type multilayers, where X includes a large number of TM, RE, or semiconductor elements [205]. The coercivity of the Pt – Co multilayers depends very strongly on a number of extrinsic factors such as deposition conditions, substrate type and treatment, and underlayer type and thickness. Films prepared by vacuum evaporation have room temperature coercivities comparable to RE – TM films. However, films prepared by standard argon sputtering techniques, the preferred commercial approach, have relatively low coercivities. H_C can be increased by sputtering in krypton and xenon instead of argon, and by using Pt or other underlayers [206], [207]. Comparative studies of the recording properties of Pt – Co multilayers versus RE – TM films at a number of different wavelengths show that at 820 nm the RE – TM film provided the best SNR, whereas at 458 nm the Pt – Co system had the highest SNR [208]. In spite of such promising results, Pt – Co multilayer films show no sign of being adopted commercially. This in part is due to the established order, but also to the versatility provided by RE – TM materials for the design of advanced media formats.

Garnets and Oxide-Based Materials. Garnets are oxides of RE metals and iron, expressed as $R_3Fe_5O_{12}$ or $3\ R_2O_3 \cdot 5\ Fe_2O_3$. They have outstanding structural and chemical stability and very favorable MO properties in the visible and soft-UV region [209]. Their characteristics are deposition dependent and can also be tailored by chemical substitution. The perpendicular anisotropy of sputtered garnets is primarily stress-induced and can be increased by substitution of ions with high magnetostrictive effects such as dysprosium. Coercivity depends on composition and grain size, and thus on preparation, with typical values 40 kA/m to 240 kA/m [209], [210]. Garnets of compositions $(Bi,Dy)_3(Fe,Al\ or\ Ga)_5O_{12}$ [209] $Ce_1Dy_2Ga_{0.4}Fe_{4.6}O_{12}$ [211], and $Bi_3Fe_5O_{12}$ [212], have good potential as candidates for MO recording. However, a major drawback of all garnets is their highly polycrystalline nature. This, as in the case of the first MO media, Mn – Bi, increases media noise to an extent that SNR levels are too low to be of practical use [213]. Research efforts are directed towards reducing the grain size [214], [215]. One possible future application of garnets is in so-called volumetric recording, where information is stored in more than one layer on each side of a disc [216]. The semitransparency of garnet films is an attractive feature in such applications. Other iron oxide-based materials, such as Co-ferrite $CoFe_2O_4$ and hexagonal Ba-ferrites, have a high MO figure of merit and an unsurpassed chemical stability and have been proposed for MO recording. However, low SNR values are again obtained due to grain-induced noise.

Advanced Media. Advances in MO media development have enabled the technology to overcome two substantial drawbacks of conventional MO recording: firstly its inability to perform direct overwrite; secondly the limit on storage density imposed by optical diffraction which states that it is not possible to “see” features smaller than the wavelength used for observation. Two types of DOW exist, Magnetic Field Modulation (MFM) and Light Intensity Modulation (LIMDOW). MFM media are virtually identical to conventional RE – TM MO media. LIMDOW media are substantially more complicated, although they are also made from amorphous RE – TM alloys.

LIMDOW was first demonstrated in 1987 and relies on exchange-coupling between various RE – TM layers of different compositions [174]. In its most basic form the LIMDOW medium has two RE – TM layers; a memory layer and a reference layer. The memory layer has a high coercivity H_{CM} and a low Curie temperature T_{CM}, whereas the reference layer has a lower coercivity H_{CR} and a higher Curie temperature T_{CR}. The recorded information is always read from the memory layer. Two magnets, an initializing magnet and a bias magnet, are needed for this scheme to operate. The initializing magnet generates a magnetic field oriented in the direction corresponding to the magnetization state of an erased disc, i.e., logic '0'. The field is large enough to saturate the reference layer at room temperature, but small enough so as not to affect the memory layer. The bias field points to the opposite direction, i.e., logic '1', and is too small to alter the magnetization in either layer at ambient temperature. The laser is operated at two power levels, P_{LOW} and P_{HIGH}, during writing. At P_{LOW} the temperature of the film exceeds the Curie temperature of the memory layer, but not that of the reference layer whose magnetization remains unaffected by the bias field. On cooling, a copy temperature, T_{COPY}, below T_{CM} is reached at which the magnetization of the reference layer is copied to the memory layer by exchange coupling [217]. Thus, a '0' has been written. To record a '1', the laser is pulsed at high power, P_{HIGH}, to heat the entire film above T_{CR}. This time on cooling the reference layer's magnetization assumes the orientation of the bias field. When the T_{COPY} is reached the memory layer also acquires the magnetization direction of the bias field via the exchange coupling with the reference layer. Finally the reference layer is reinitialized as it passes above the initializing magnet (see Fig. 69). The need for two permanent magnets with one of them, the initializing magnet, generating a high field complicates disc drive design. A solution to this problem is the quadrilayer exchange-coupled structure that has two additional layers, an initializing layer and a switching layer, that remove the requirement for an initializing magnet [217]. Commercial systems employing such materials have been launched on the market by several manufacturers.

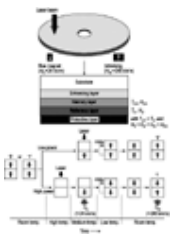

Figure 69. The LIMDOW direct overwrite mechanism in exchange-coupled double-layer films showing the disc structure (top) and the overwrite mechanism (bottom).

MSR technology improves the read-out resolution by incorporating a read-out layer that copies information from the memory layer only within a limited temperature range. Thus, data from the memory layer is read though what is effectively a very small, substantially submicron, aperture that is placed directly above the memory layer [173]. The optical resolution is thus determined by the size of the aperture, rather than the wavelength used for read-out.

Several MSR schemes are possible, the commonest being Rear-Aperture-Detection (RAD), Front-Aperture-Detection (FAD) and Central-Aperture-Detection (CAD); the name arising from the location of the effective aperture in the focused laser spot that reads the data and induces the necessary temperature rise for aperture formation [169]. MSR techniques require at least two active MO layers, one for the memory layer and one for read-out and aperture formation. In reality extra layers are needed to control the interlayer coupling and switching. A typical CAD – MSR system is shown in Figure 70 [172], [218]. Here, a read-out layer of RE – TM material, typically Gd – Fe – Co, has an in-plane magnetization at room temperature. As the temperature of the film is increased slightly by the read-out beam, a small aperture forms in which the magnetization, due to magneto-static coupling with the memory layer, becomes perpendicular. The nonmagnetic intermediate layer, usually a dielectric, moderates the level of magneto-static coupling. The memory layer is typically Tb – Fe – Co, as in conventional MO films. Using such a scheme the bit size in both circumferential and radial directions can be substantially smaller than when using a conventional recording approach. Over 6 Gbytes of storage on a single-sided 120 mm disc has been reported [172], [218].

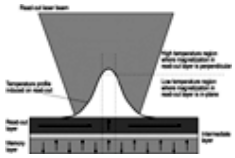

Figure 70. Central aperture detection system in which the readout layer is typically Gd – Fe – Co (≈ 30 nm thick) and the memory layer is Tb – Fe – Co (≈ 40 nm thick).

Magnetically Amplified Magneto-Optic System. Yet another MO media format that is the topic of much research is the Magnetically Amplified Magneto-Optic System (MAMOS). Here the opposite approach to that of MSR is adopted to achieve high density read-out in so much as instead of reading out data through a small effective aperture, small bits are copied from a memory layer to a read-out layer and then expanded in size before being read in a conventional way [219]. Again the materials of choice for this scheme are RE – TM films due to the design flexibility that they offer.

2.5.2. *Phase-Change Recording Media*

Principles. Phase-change recording is based on the formation of (sub)micron-sized amorphous marks in a crystalline thin film by using a focused laser beam. The amorphous and crystalline phase have different refractive indices, leading to amplitude and/or phase modulation of the reflected laser beam during readout. Since the transition between the crystalline and amorphous phase is reversible, information can be erased and/or overwritten (see Fig. 71).

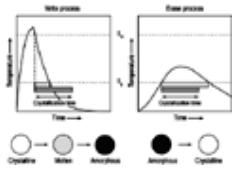

Figure 71. Principle of phase-change rewritable optical recording. The write process involves the application of a short high-power laser pulse to locally melt the phase-change layer. After turning the laser power to a lower level, the molten material is quenched and an amorphous dot is formed. Recrystallization is prevented because the time interval between the melting temperature T_m and the glass transition temperature T_g (the gray bar in the figure) is much shorter than the crystallization time (the black bar).

The erase process involves the application of an intermediate d.c. erase power level. The phase-change layer is heated to a temperature between the glass transition temperature T_g and the melting temperature T_m sufficiently long to recrystallize the material.

An amorphous mark is written by using a short, high-power laser pulse that melts the recording layer locally. Immediately after the pulse the laser is switched to a low level, so that the molten material is rapidly quenched; if the cooling rate is sufficiently large, recrystallization can be avoided and an amorphous dot is formed. A recorded mark is erased by operating the laser at an intermediate d.c. erase level, thus heating the recording layer to temperatures between the glass transition temperature T_g and the melting point T_m. In this temperature range the atoms have sufficient mobility to enable recrystallization of the material. For direct overwrite applications, the crystallization rate of the phase-change material should be sufficiently fast to erase the recorded marks completely within the dwell time of the laser spot. The crystallization rate of the material is thus of key importance to the data transfer rates that can be achieved.

Material Specifications. In order to understand the properties that are relevant for rewritable phase-change materials, it is necessary to examine the crystallization process in some detail [220], [221]. Crystallization of the amorphous phase involves nucleation and growth processes, similar to the crystallization of a liquid during undercooling. In fact, both cases can be considered equivalent, because during heating the amorphous phase usually undergoes a glass transition (at the glass transition temperature T_g) to become a supercooled liquid prior to crystallization.

Crystal Growth. The driving force for crystallization is the free energy difference ΔG between the supercooled liquid (L) and crystalline state (C), given by the equation:

$$\Delta G = \Delta H_{\mathrm{LC}} - T\Delta S_{\mathrm{LC}}$$

where ΔH_{LC} and ΔS_{LC} are the enthalpy and entropy difference between the two states, respectively. This relation can, to first order, be approximated by:

$$\Delta G = \Delta H_{\mathrm{f}} \frac{(T_{\mathrm{m}} - T)}{T_{\mathrm{m}}}$$

where ΔH_{f} is the latent heat of fusion and T_{m} the melting temperature. The equation shows that the driving force for crystallization increases with the undercooling $T_{\mathrm{m}} - T$. In contrast, the mobility of the atoms in the supercooled liquid decreases with increasing undercooling. The combination of both effects results in an initial increase of the crystal growth

rate when the material is cooled below T_m. At a certain undercooling the growth rate maximizes, and it decreases again when the material is cooled further towards the glass transition temperature T_g. Below T_g, the mobility and consequently the growth rate are negligible (Fig. 72).

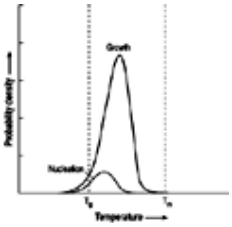

Figure 72. Probability density of the nucleation and growth processes as a function of temperature T. Nucleation takes place only at considerable undercooling $T_m - T$, with T_m the melting temperature. The growth rate maximizes at higher temperatures. Below the glass transition temperature T_g the mobility of the atoms is too small to enable crystallization.

Nucleation. In order to start crystallization, nuclei have first to be formed. The energy required to create the solid-liquid interface, given by the solid-liquid interfacial energy σ, forms a barrier towards nucleation. The work of formation ΔG_n for a cluster of n atoms is given by:

$$\Delta G_n = n\Delta G' + n^{\frac{2}{3}} A\sigma$$

where $\Delta G'$ is the free energy difference between an atom in the solid and an atom in the liquid, and A is a geometry factor. Since $\Delta G'$ is negative and σ positive, a critical nucleus size n^* exists at which the heat of formation is maximal. Below this size the nuclei will on average shrink, above this size the nuclei tend to grow further to crystallites. At small undercooling, where the critical nucleus size n^* is large, the nucleation rate is negligible. The onset of nucleation occurs only at substantial undercooling. At temperatures just above T_g, the nucleation rate is small due to the low atomic mobility (Fig. 72).

The density of the nuclei determines the number and size of crystallites that are being formed. This is an important parameter in optical recording, because the grain size of the crystalline phase determines the ultimate bit density that can be obtained. However, the nature of the actual nucleation site is often unclear: Nuclei may form spontaneously within the material itself (homogeneous nucleation), or at impurities or interfaces where the interfacial energy σ is decreased (heterogeneous nucleation).

On the basis of crystallization theory, the following criteria for obtaining fast crystallization rates can be discerned:

1. The free energy difference between the supercooled liquid and the crystalline phase should be large. In an alloy system, the free energy level of the liquid state varies slowly with composition, whereas the free energy of the crystalline state drops steeply at stoichiometric compositions. Thus, the driving force for crystallization and consequently the crystallization rate will be maximal for stoichiometric compounds.
2. The mobility of the atoms in the amorphous and supercooled liquid phase should be high. The mobility is affected by the viscosity of the liquid phase, which depends on the binding energy of the atoms. When the binding forces between atoms are strong, the mobility is decreased and crystallization is slow. In general, if compound materials do not have strong covalent bonds the crystallization rate is relatively large.
3. The atomic diffusion distances between the amorphous and crystalline state should be small. These distances are closely related to material composition and crystal structure. A nonstoichiometric compound generally crystallizes into different phases. Phase separation requires atomic diffusion over longer distances and is time-consuming. Therefore, crystallization of a stoichiometric compound into a single phase is generally faster. Crystal structure plays also a role. The atomic distribution in the amorphous state is most probably isotropic as in the liquid state. Therefore, it is likely that materials with a simple isotropic crystal structure crystallize more rapidly than materials with a more complex anisotropic crystal structure.

A suitable phase-change optical recording material should have the following properties:

1. The optical constants should allow sufficient contrast between the amorphous and crystalline state at the recording wavelength. This suggests the use of semiconductors instead of metals or insulators. Semiconductors generally exhibit a larger difference between optical constants of the amorphous and crystalline phase, due to the fact that the absorption edge shifts in the visible range during the phase transition.
2. The sensitivity of the material should be appropriate: The melting point must be low enough to melt the material with the available laser power. On the other hand, the amorphous phase must be sufficiently stable against self-crystallization at ambient and storage temperatures up to 50 °C. In practice, this means that the melting point T_m should be between 500 and 1000 °C, while the glass transition temperature (typically 50 – 70 % of T_m) should be well above 100 °C.
3. The amorphous-to-crystalline (a – c) and crystalline-to-amorphous (c – a) transitions should both be rapid. For high-speed direct overwrite applications, the transitions should proceed within the dwell time of the laser spot, typically 100 ns or less. This criterion sets an upper limit on the complete erasure time, i.e., the time required to erase (crystallize) an amorphous mark completely. On the other hand, the complete erasure time should not be too small to prevent recrystallization of the molten material during the write (melt – quench) process.
4. The amorphous-to-crystalline (a – c) and crystalline-to-amorphous (c – a) transitions should be possible many times. Depending on the application of the disc, a repeatability of up to 10^5 write – erase cycles is required. It appears beneficial to use single phase materials, in order to prevent progressive phase separation during the many write and erase cycles.

Material Compositions. Since the first observation of rapid reversible laser-induced amorphous – crystalline transitions by OVSHINSKY in 1970 [152], [257], a variety of materials for reversible phase-change recording has been investigated. Mainly, alloys based upon elements from the chalcogen family (sulfur, selenium, and especially tellurium) have been used, because these materials can be molten by low-power laser irradiation, quenched into an amorphous state easily, and have appropriate optical differences between the two structural states.

In the early years of phase-change materials research (1970 – mid-1980s), the media consisted of a phase-change layer sandwiched between dielectric layers on a plastic substrate. The relatively low cooling rate of these stacks led to the search for phase-change materials which could be easily vitrified. Eutectic (low melting point) compositions of Te-based alloys were found to be promising candidates: These alloys have a relatively low mobility in the liquid phase, resulting in an easy amorphization process. The low mobility is explained by the presence of long Te-chains, even in the liquid state, which are extensively cross-linked by the alloying elements that are used, such as Ge, As, or Sn [258], [260]. The negligible mobility at room temperature results in long data retention times. However, the inherent disadvantage of the small mobility is the large crystallization times of these materials, typically on the order of several hundreds of nanoseconds or microseconds [152], [261] which excludes their use in high data rate direct overwrite applications.

During the 1980s, it was gradually appreciated that virtually every material can be vitrified if the cooling rate is sufficiently large. By using thin-film stacks as described below (with a so-called “rapid cooling structure”), quench rates on the order of 10^{10} K/s can easily be obtained. Therefore, the rate of the crystallization process is nowadays a more important issue than the ease of vitrification of the material. In order to shorten the crystallization times, stoichiometric compositions have been studied from the mid-1980s. In 1986, the feasibility of direct overwriting with one laser beam was demonstrated for In – Se based films [262], [263]. Around the same time, it was shown that a stoichiometric GeTe alloy can be crystallized within 30 ns [258]. However, practical application of GeTe appeared to be limited by the large decrease in crystallization rate for compositions that are only slightly off-stoichiometric, due to the decrease of the nucleation rate [264] and/or the phase separation of excess Ge or Te [259].

A major breakthrough in phase-change materials research was the discovery of Ge – Sb – Te ternary alloys as recording material [265-267]. These alloys crystallize within 100 ns for a relatively wide range of compositions near the GeTe – Sb_2Te_3 pseudo-binary line (Fig. 73). Especially the stoichiometric compounds on this line ($Ge_2Sb_2Te_5$, $GeSb_2Te_4$, and $GeSb_4Te_7$), with crystallization times below 50 ns, have interesting recording characteristics: Phase segregation does not take place in these compounds during repeated melting and recrystallization, resulting in a direct overwrite repeatability of more than 10^5 cycles.

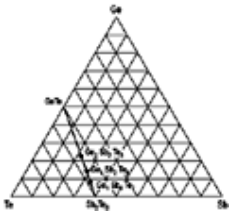

Figure 73. The Ge – Sb – Te ternary alloy system. The indicated stoichiometric compounds on the pseudo-binary GeTe-Sb_2Te_3 tie line have crystallization times on the order of 30 – 50 ns, and are the nowadays most frequently used materials for high-density phase-change recording.

In their thermodynamically stable configuration, the structure of the stoichiometric GeTe – Sb_2Te_3 compounds is built up of planar homo-elemental layers which are fcc-stacked to form hexagonal lattices [266], [268], [269]. The periodicity of these lattices varies from 9 ($Ge_2Sb_2Te_5$) up to 21 layers ($GeSb_2Te_4$), which makes these structures rather complex. However, under laser irradiation a metastable NaCl-type structure is formed, which is built up of Te-planes and planes with lattice sites randomly occupied by Ge and Sb [260]. This isotropic cubic structure resembles the isotropic atomic distribution of the amorphous state, so that only small atomic movement and bond angle adjustment is required for crystallization, which explains the fast crystallization rate of these materials. Transformation from the metastable cubic phase to the hexagonal equilibrium phase is kinetically inhibited under short laser-pulse irradiation.

Regarding their short crystallization time, pseudo-binary GeTe – Sb_2Te_3 alloys are particularly suitable recording materials for media that are used at high linear velocities (5 m/s or higher), such as the high-density rewritable DVD media. These compositions appear less practical for use at lower linear velocities, because then the recording layer (partially) recrystallizes during the melt – quench process, due to the decreased cooling rate of the disc. Therefore, the crystallization rate of the recording materials for low velocity rewritable media (such as CD-Rewritable) should be decreased intentionally. A possible way to decrease the crystallization rate of the Ge – Sb – Te alloy is the choice of compositions remote from the pseudo-binary GeTe – Sb_2Te_3 line, e.g., by adding Sb to a composition on the GeTe – Sb_2Te_3 tie line [270], or by adding an additional element, such as Co [271].

In CD-RW media, introduced in 1996, the phase-change layer generally consists of a quaternary Ag – In – Sb – Te alloy [272-274]. The crystallization time of these materials is typically on the order of 100 – 200 ns, making them suitable for the range of linear velocities that is currently covered in CD-RW (CD-1X: 1.2 – 1.4 m/s, CD-2X: 2.4 – 2.8 m/s). An advantage of these materials relative to (doped) Ge – Sb – Te alloys is their high sensitivity. Unfortunately, the overwrite repeatability is limited to 10^4 cycles, possibly due to phase segregation in this multiphase material [275].

Structure of Phase-Change Recording Media. A rewritable phase-change optical disc consists of a thin-film stack which is sputter-deposited onto a pregrooved plastic substrate (Fig. 74). The stack comprises at least four layers, a dielectric underlayer, the phase-change recording layer, a dielectric upper layer, and a reflective mirror layer. The disc is covered with a protective lacquer layer, which is applied onto the quadrilayer stack by spincoating. As in the case of prerecorded (read-only) media, the information in the recording layer is written and read through the substrate, making the system very robust

against dust and scratches. The pregrooves in the substrate form a grating structure which is used for tracking with the radial push – pull method (see Section The Optical Channel).

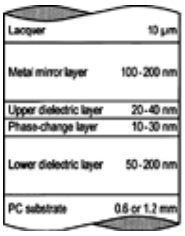

Figure 74. Typical four-layer stack design of current phase-change rewritable discs. The phase-change layer is sandwiched between protective dielectric layers, which are also used to optimize the optical and thermal properties of the disc. The metal layer has a mirror and heat-sink function.

In current commercially available recording formats, such as CD-RW, information is only recorded in the grooves (*groove recording*, see Fig. 75). Typically, groove depths of approximately $\lambda/8n$ (with λ the wavelength of the laser and n the refractive index of the substrate) are used to optimize the magnitude of the push – pull tracking signal. The distance between the grooves (track pitch) is an important parameter in optimizing the storage capacity of a rewritable disc. In practice, the minimum track pitch in groove recording is limited by optical cross-talk between adjacent data tracks and by the stability of push – pull tracking.

A further reduction of the track pitch is possible by applying *land/groove recording* (Fig. 75), in which information is recorded in both land and groove tracks. Although adjacent data tracks are closer together, optical cross-talk between them can now be canceled to a large extent by optimizing the depth (approximately at $\lambda/6n$) and shape of the grooves. The track pitch that can be achieved in land/groove recording is typically 75 % of that in groove recording. Radial tracking is not really a critical factor in land/groove recording, because the groove pitch (twice the track pitch) is much larger than its critical value approached in groove recording. The limiting factors in reducing the land/groove recording track pitch are optical cross-talk and thermal cross-erase, which is the erasure (recrystallization) of amorphous marks in the neighboring tracks during the write process [276]. A possible solution against thermal cross-erase is the use of steep and/or deep grooves, so that the land and groove tracks become thermally isolated [277].

Figure 75. Principles of groove recording (left panel) and land/groove recording (right panel).

Thin-Film Stacks Design. Materials choice and design of the thin-film stack is a rather complex matter, because the recording performance of the disc depends critically on the physical and chemical properties of the various layers in the stack.

The sensitivity of the disc is closely related to the optical and thermal characteristics of the stack: A large part of the incident optical energy should be converted into heat within the recording layer, while at the same time the cooling rate of the molten material in the recording layer must be sufficiently large (on the order of 1 – 10 K/ns) to prevent recrystallization during quenching. The cooling rate of the stack is mainly determined by the thermal resistance of the upper dielectric, and the heat capacity and thermal conductivity of the metal mirror heat sink. In practice, these requirements limit the thickness range of the upper dielectric layer. The requirement that the incident optical energy has to be used to a large extent for heating the phase-change layer sets a lower limit to the layer thickness of the upper dielectric. If this layer would be too thin, the heat would almost instantly flow into the metal mirror, which has a much larger thermal conductivity than the other materials. On the other hand, the requirement of a sufficiently large cooling rate sets an upper limit to the thickness of the poorly heat conducting upper dielectric layer. The lower dielectric layer should be taken sufficiently thick, so that the plastic substrate does not heat too much during recording.

The optical contrast between the amorphous and crystalline phase is optimized by tuning the thickness of the phase-change and the lower dielectric layer. As for prerecorded media, the background (crystalline phase) is designed to have a high reflectivity, while the recorded marks (amorphous phase) have a low reflectivity. For sensitivity reasons, the crystalline reflectivity is typically between 20 – 25 %, and the amorphous reflectivity below 5 %. In groove recording, the phase difference between the light reflected from the crystalline or amorphous phase can be used to further optimize the signal modulation. In land/groove recording, cancellation of optical cross-talk between adjacent data tracks requires a very small phase difference between light reflected from the amorphous and crystalline phase.

The choice of dielectric layer materials is particularly important for the overall recording performance. The chemical nature of the dielectrics influences the crystallization kinetics of the phase-change material, because the initial stage of crystallization (nucleation) is sensitive to the energy of the phase-change/dielectric interface (see above). The dielectric layers should act as diffusion barriers to prevent the phase-change material against oxidation, contamination, or vaporization, at storage temperatures but also during the write process, when the temperature in the phase-change layer increases up to 1000 °C. Stability of the dielectrics themselves during repeated heating is a key issue. Mechanical properties play also a role: The thermal expansion of the various layers in the stack while locally heating the phase-change layer induces large stresses. This results in void formation and mass transport in the phase-change layer, the formation of microcracks in the dielectric layers, and/or delamination of the phase-change layer and dielectric layers. These phenomena may seriously limit the number of direct overwrite cycles (DOW repeatability). The choice of suitable dielectric materials can to some extent reduce these degradation effects.

In current phase-change recording media, the most frequently used dielectric is a mixture of 80 % ZnS and 20 % SiO_2. This mixture appears to give good power sensitivity and overwrite repeatability. A disadvantage of this material is its low sputter deposition rate and consequently its cost. Therefore, alternative dielectrics (oxides, nitrides, carbides) are currently under investigation. Table 13. shows the layers of a DVD–RAM phase-change stack.

Table 13. Phase-change stack of a 4.7-GB DVD–RAM

Layer	Name	Thickness	Role in stack
Polycarbonate	substrate	1.2 mm	grooved substrate with address information
$ZnS\text{-}SiO_2$	1st dielectric layer	140 nm	thermal isolation of the substrate, also tunes/enhances reflectivity
GeCrN	interface	5 nm	changes interface energy and nucleation rate; reduced S diffusion
$Ge_6Sb_2Te_9$	active layer	9 nm	Phase-change layer
GeCrN	interface	5 nm	changes interface energy and nucleation rate; reduced S diffusion
$ZnS\text{-}SiO_2$	2nd dielectric layer	25 nm	thermal control of cooling rate
GeCr	absorption control	55 nm	optical compensation (normalized absorption between amorphousand crystallization marks); also isolates ZnS and Ag
Ag-alloy	heat sink	120 nm	heat sink
UV lacquer	corrosion protection	6 µm	

Recording Strategy. In early optical storage systems, information was recorded by using the pulse position modulation (PPM) method. A PPM recording strategy uses a single high-power laser pulse to write an amorphous dot. The central position of this amorphous dot corresponds to “1” in the binary code. An inherent advantage of the PPM method is that small distortions of the recording mark do not introduce bit errors during read-out.

In current high-density optical storage systems the so-called pulse-width modulation (PWM) or mark-edge recording technique is applied. In this method, the lengths of the amorphous marks and of the crystalline spaces between these marks are varied in a stepwise fashion. Now, the edges of the amorphous marks correspond to “1” in the binary code, leading to an increased recording density with respect to PPM recording.

The PWM direct overwrite strategy of the CD-Rewritable system is given in Figure 76. Marks are written by using a train of short, high-power laser pulses. Between these pulses the laser power is switched to a low level, so that after each pulse the molten material is quenched into the amorphous state. The written mark thus consists of a series of overlapping amorphous dots, its length being defined by the number of pulses in the sequence. The advantage of using a train of short pulses rather than a single long pulse is the reduction of heat accumulation and recrystallization during writing, resulting in a constant mark width irrespective of mark length. Direct overwrite is accomplished by switching the laser power to an intermediate d.c. erase level between successive write pulse trains.

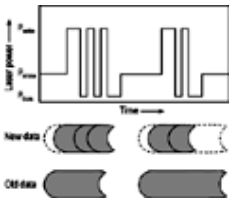

Figure 76. The (direct over)write strategy of the CD-Rewritable recording system. Amorphous marks are written by using a sequence of high power laser pulses (P_{write}). The width of the first write pulse is equal to the channel clock period T, subsequent pulses have a width of $T/2$. Between the write pulses the laser is switched to a low power level (P_{bias}) to reduce the accumulation of heat and to quench the molten material into the amorphous state. Between the write pulse trains, an intermediate d.c. erase power level (P_{erase}) is used to recrystallize previously written amorphous marks.

Read-out of recorded information involves the detection of the edges of amorphous marks. One of the most critical sources of noise is the irregularity in the definition of recorded marks, which leads to timing jitter [fluctuation of the (detected) mark edge position relative to its nominal position]. Reduction of jitter is a rather complex issue, because the geometry of written marks is only indirectly related to the intensity profile of the laser beam, due to the nonlinear and nonuniform thermal processes involved. One of the sources of jitter is the accumulation of heat in the recording layer during the (over)write process, resulting in thermal interference between an amorphous mark and its preceding and succeeding crystalline spaces. Write strategies have been developed which correct for these thermal interference effects, e.g., by shifting the first and last write pulse in the sequence [278] or by using a variable preheating pulse before the write pulse train and a variable cooling gap after the last write pulse.

Interference between old and new data is the main reason for the significant increase of “overwrite jitter” relative to “write-once” jitter. The differences in optical absorption and thermal parameters of the amorphous and crystalline phase lead to mark shape variations in overwriting existing data. A typical four-layer stack has a low reflective amorphous phase and a high reflective crystalline phase, resulting in a higher absorption of the amorphous phase (A_a) than that of the crystalline phase (A_c). The reflectivity (and absorption) of the recording stack is the result of optical interference between light reflected at the various interfaces in the stack. Absorption of light may take place in several layers, depending on the optical constants and thickness of the employed materials. The amount of light absorbed in the phase-change layer depends on the actual (amorphous or crystalline) state of the material at the position of the laser spot. “Overwrite” jitter can be reduced by designing the thin-film stacks so that A_c is slightly higher than A_a, to compensate for the latent heat of melting of the crystalline phase. Stacks with $A_c > A_a$ can be achieved in various ways, e.g., by incorporating an absorbing Si-layer between the upper

dielectric and the aluminum mirror [279] or by using a semi-transparent Si-mirror instead of aluminum [280]. Another possibility is a stack with a high reflective amorphous state and a low reflective crystalline state, so that $A_c > A_a$. This can be realized by inserting a semi-transparent layer (e.g., a thin Au layer) between the substrate and the lower dielectric [281].

2.5.3. *Liquid Crystals*

A liquid crystal consists of rod-shaped molecules (⟶ Liquid Crystals). Order-to-disorder transitions occur at low temperature (250 °C). Certain types of highly ordered liquid crystals (smectic) have been developed as erasable optical storage media [223-225]. The writing mechanism is the formation of disordered light-scattering centers in micron-sized areas by heating an adjacent metal electrode with the laser beam. The erase mechanism for the disordered light-scattering centers is reorientation by laser heating in the presence of an external electric field generated using adjacent a.c. electrodes. The liquid crystal layers are ca. 8 µm thick.

The light scattering is due to a spatial variation in the refractive index. The write laser pulse times can be very short (< 100 ns), but the formation (freezing-in) of the disordered scattering center takes much longer (ca. 5 µs) compared to the transition times for inorganic phase change alloys (see Section Phase-Change Recording Media). The reorienting erasure also has a long relaxation time (ca. 5 µs). For the nonscattering ordered texture to have few defects (low noise), the smectic liquid crystal layers must be oriented parallel, and the liquid crystal molecules oriented perpendicular to the enclosing substrates.

This critical alignment can be achieved using well-known methods from liquid crystal display technology. The liquid crystal layers must be totally encapsulated because of their low viscosity (< 20 mPa · s). In spite of several difficulties, moderate CNRs are obtained and liquid crystal media can store data at densities and data rates compatible with optical media requirements. The kinetics of this mechanism have created difficulties with DRAW or DRAE (direct-read-after-write or erase) capabilities and with direct overwriting of information.

All of the above difficulties (except for the kinetic aspect) have been addressed using liquid crystalline polymers [226-230]. In these polymers, the ordered molecular groups are side groups covalently linked to a main polymeric backbone. The ordered structures generated in such polymers are frozen-in below the T_g of the backbone polymer. Such ordered polymers are readily deposited on substrates in micron thickness using spin coating.

These polymers also contain dye side groups covalently linked to the backbone polymer. The dye groups allow localized laser heating, resulting in high sensitivity and contrast. Formation and reorientation of the light-scattering centers (write – erase) takes a relatively long time (>10 µs). This time is longer than for nonpolymeric liquid crystals because of backbone rotation and steric hindrances. The use of an electric field during reordering helps to speed up erasure of the disordered scattering center. Organic order – disorder transitions can be extended to the monomolecular regime [231]. The deposition technique is the classical Langmuir – Blodgett method (⟶ Thin Films).

2.5.4. *Photochromics*

Photochromism involves a reversible electronic level transformation induced in the writing direction by absorption of short wavelength light and in the erasing direction by absorption of longer wavelength light (photon mode recording mechanism) [232], [80], [233-240], [241]. The two chemical species in these bistable switches are related, but have distinct optical properties such as absorption or refractive index. Many photochromic reactions are completed within a few nanoseconds so that no cooling time effects occur as in the heat mode optical recording mechanisms considered so far throughout this entire article [232]. A wide range of organic photochromics with absorption maxima in the visible region is available. The spectral sensitivity of photochromics has not yet been extended to the technically important near-IR region.

The reverse transition (erasure) is accelerated by light or heat. This spontaneous reversal or erasure to the more stable initial species (bleaching) or to a different species (fatigue) is the major technical obstacle to the use of organic photochromics for optical recording. Considerable progress has been made in designing organic photochromics with improved bleaching and fatigue properties but stability sufficient for optical recording has still to be fully demonstrated. Organic photochromics can be extended to the monomolecular level using Langmuir – Blodgett deposition techniques [239], [240]. Other photochemically triggered mechanisms such as induced circular dichroism and molecular changes (e.g., hydrophilic to lipophilic, *trans – cis* isomerization) are also being investigated [232].

Photochromics are often very finely dispersed in polymeric binders. Reversible photochromism in polymers is rare. A soluble organometallic polysilane derivative that exhibits thermal photochromism in the near-UV has been synthesized [242-244]. Photostimulation and luminescence of electron traps in phosphors is another type of photochromism [245].

Bistable optical organic switches can be made from other organometallics. Cu and Ag TCNQ (7,7,8,8-tetracyano-1,4-quinodimethane) undergoes an electronic transformation (by electron transfer) to elemental copper or silver and TCNQ [246], [247]. This redox reaction occurs in a few nanoseconds and produces high contrast marks. The reaction is driven in one direction by absorption of light. The reaction is claimed to be driven in the reverse direction by laser heating. Reproducing the erasure results is difficult. A composition with excess TCNQ may expedite the erasure kinetics.

2.5.5. *Photochemical Hole Burning*

Cryogenic (<20 K), frequency-selective, photochemical hole-burning (PHB) offers optical storage densities greater than the diffraction limited ca. 100 Mbits/cm^2 [248-256]. Various photoreactive organic molecules and inorganic color centers can be dispersed as the carrier species in amorphous or crystalline host materials. These carrier species have inhomogeneously broadened absorption lines, up to 10^3 times broader than the homogeneous absorption band of the species alone, due to the random local environments encountered in the host matrix.

PHB writing consists of burning a hole in the broad absorption band by depleting the subset of species whose frequency-

shifted absorption matches the wavelength of a frequency-scanned laser. PHB read-out uses special frequency modulation spectroscopy techniques. Erasure is performed by heating the material above 20 K. New materials are being sought that exhibit this effect at higher temperature, e.g., near the boiling point of liquid nitrogen or even at room temperature.

Photochromics, PHB, and other photochemically- and photophysically-triggered optical recording mechanisms (photon mode materials) have been reviewed [232], [61], [241].

2.6. Economic Aspects of Optical Recording

Today's worldwide optical disc device market has gone through drastic changes since 2000 until now (2004). In 2000 the CD-R and CD-RW markets were expanding, DVD–ROM was starting up and now these products have already become commodities. The Audio-CD has reached its peak and still has the biggest part of the market, followed by CD-ROM, but since 2001 the CD market has slightly decreases, due to the growth of DVD. Extensive research has occurred in the field of optical storage and many new products have appeared on the market. The market of DVD Recordables and Rewritables (DVD+/–R/RW and DVD–RAM), the different formats were released from 1997 to 2002, keeps on expanding without unified standard. Furthermore the launching of blue laser products such as Blu-ray and UDO has begun. On the other hand, another Blue Laser standard, which was introduced as AOD, is now entering the crucial stage of standardization. Today, both cutting-edge type and commodity-type optical disc devices are making waves in the market. Various categories of devices, such as CD-R/RW, Combo, DVD–ROM, are sold in particular markets for specific applications, and have already established strong positions in the PC market. Therefore, the entire optical disk market has been uplifted and is constantly growing.

2.6.1. *Consumer Electronics (CE)*

The optical storage market for consumer applications can be divided into three types of applications: audio, video and games.

Audio. The audio market has been the largest part of the optical storage market over the last years and has reached its peak. Now the customers are starting to buy DVD's preferably. The decline affected virtually all major markets, with Western Europe showing particularly sharp falls compared to recent years. Sales in Germany went down by 19% in 2003 and by more than 30% in value since 1999. Denmark, France, Sweden, Belgium, Greece, Ireland, Portugal and Switzerland also experienced double digit declines. Since 2000 the industry has suffered global losses of 20% year after year. In 1983, the first year they were available, 800 000 compact discs were sold in the United States. By 1990 that number reached nearly 10^9 discs sold worldwide. It is estimated that more than 12×10^9 prerecorded discs were sold worldwide in 2003 (14×10^9 in 2002). The number of sold CD-R and CD-RW — also used for audio recording — is still increasing (8×10^9 in 2003) and the peak is expected in 2005 at 9×10^9 discs. Apart from the 120 mm CD discs, also the market for recordable MD discs (a 64 mm disc in a cartridge of $72 \times 68 \times 5$ mm^3) has already reached its peak in 2002 at 220×10^6 units and is now declining.

Video. Until the DVD was released the video market has been served for a long time by the laser disc; a 300-mm analogue video format and the VCR (video cassette recorder). The laser disc, the 12-inch optical disc format that had been available since 1978, is obsolete since 2001. Pioneer Entertainment, the long-time champion of laser disc, abandoned laser disc production in the U.S. in June 1999 but continued to release small runs in Japan until 2001. The laser disc still fills niches in education, training, and video installations, but it is fading even there. Existing laser disc players and discs will be around for a while, though essentially no new discs are being produced. There is a total of over 35 000 titles worldwide that could be played on over 7×10^6 laser disc players. It took DVD several years to reach this level, and there are still rare titles available on laser disc but not on DVD. Looking at the video tape: DVD player sales exceeded VCR sales in 2001 and sales of DVD recorders are expected to reach similar levels of VCR decks in 2006. DVD recorders are expected to slowly take over for DVD players, in stand-alone products as well as in combination products such as DVD+VCR and TV+DVD. In 2004, the first DVD recorder combination products are entering the market. There are several DVD recorder + VCR and “Home Theater in a Box”+ DVD recorder systems scheduled to be launched in 2004. In 2003 alone, DVD recorder unit shipments increased by more than 200% and for 2007 the shipment of more than 50×10^6 DVD recorders is expected. DVD recorders will accelerate the death of VCRs once the price difference is small enough. DVDs have many advantages over tapes, such as no rewinding, quick access to any part of a recording, and fundamentally lower technology cost for hardware and disc production. Some projections show DVD recorder sales passing VCR sales in 2005. By 2010 VHS may be as dead as vinyl records were in 2000. DVD–Video is a huge success, in fact the most successful format ever introduced into the consumer electronics industry. In 1997, the first year DVD–Video players became available in the USA, 349 000 players were sold, together with 5.5×10^6 discs. DVD player unit shipments in 2003 were 98×10^6 units and will top 100×10^6 in 2004. Worldwide production of DVD recordings exceeded 2.5×10^9 discs in 2003, nominally half the number of audio CDs sold. Sales of DVD recorders are expected to reach 13.6×10^6 units in 2004 and 22.6×10^6 in 2005. Since these recorders can also play back both DVD and CD recordings, there may be little need in the future for playback-only units of any kind in the home. Automotive DVD systems and portable DVD players are expected to grow by over 30% annually for the next five years. In addition, while households in Europe, North America, and Japan will convert to DVD recorders, markets like China will continue to consume DVD players. The total market will then begin its decline in 2005, though the portable and automotive segments will still continue to grow.

Games. A very fast growing, no longer niche market for Consumer Electronics is the market of CD- and DVD–based computer games and interactive information and entertainment programs. A distinction must be made between PC games on the one hand and console games on the other. The most popular consoles (sold units until 2003) are Sony PlayStation 2 (51.2×10^6), Nintendo GameCube (9.55×10^6) and Microsoft Xbox (9.4×10^6). According to one forecast, sales of games will overtake Audio-CDs in Europe in 2005. The global video game market will expand from $\$21.2 \times 10^9$ in 2002 to $\$35.8 \times 10^9$ in 2007, growing at an 11% compound annual rate.

2.6.2. *Data Applications*

The distinction between “Consumer Electronics” and “Data Applications” is very useful for analysis; at the same time it may

be very misleading. For instance: virtually any CD/DVD–ROM drive is able to play Audio-CDs or movies on DVDs. The distinction becomes even more blurred when interactive applications are considered. The catchword "multimedia" points directly to the observation that the PC world and the CE world, which have expanded along two unrelated tracks for a long time, are now beginning to overlap. The currently available PCs for home use are called Mediacenter-PC and this not only because of the multimedia extension of the Microsoft Windows operating system released in 2004.

CD-ROM. The first CD-ROM drive was already introduced in 1984, but the market was completely stagnant until advances in CD technology reduced the drive cost sufficiently and until sufficient titles became available. At the time of writing (2004), the CD-ROM market is as big as the CD-Audio market. The market for PC CD-ROM drives is declining because state-of-the-art PCs are loaded with combi-drives, a combination of DVD–ROM/CD-ROM, DVD–ROM/CD-recorder or DVD–recorder/CD-recorder. The recorders play as a matter of course the prerecorded version of the discs.

CD-R, CD-RW. In 2003 about 10×10^9 discs were sold and the market is still increasing, but the peak is expected for 2005.

5.25″MO. The last version of this professional optical storage technology was introduced in 2000 with rewritable 9.1 GB double-sided discs. There are many bases installed for backup purposes but it is expected that these will be replaced by UDO solutions.

UDO is the next generation 5.25" professional optical storage technology introduced in 2003. It is a convergent technology that delivers the performance of 5.25" MO, the longevity of 12-inch WORM, and the cost-effectiveness of DVD. UDO utilizes blue-violet lasers and phase-change media recording technology from DVD consumer products to provide a quantum leap in data storage densities. First generation UDO products have 30 GB capacity, future generations will increase capacity to 60 GB and 120 GB and will provide full backward-read compatibility. WORM and rewritable media are/will be available. Forecasts expect 30 000 units sold in 2004 increasing to 60 000 units in 2006.

DVD–R(9), DVD+R(9), DVD–RW, DVD+RW, DVD–RAM. DVD–R and DVD+R are WORM media, DVD–RW and DVD+RW rewritable media. The first recordable version of DVD was introduced in 1998. The data capacity of 4.7 GB is equivalent to DVD–ROM. The first double-layer versions (8.5 GB) became available in 2004. The third DVD rewritable standard, DVD–RAM, runs a distant third, but has the best defect management and reliability. This technology is incompatible with most commercial DVD players. DVD–RAM discs are housed in cartridges and hold between 2.6 GB and 9.4 GB. The market for recordable and rewritable DVDs is dramatically increasing. In 2003 about 1×10^9 discs were sold worldwide, for 2004 sales of nearly 2×10^9 units are expected.

Blu-ray Disc. The Blu-ray Disc, presented in 2003, enables the recording, rewriting and play-back of up to 27 GB of data on a single-sided single-layer 12 cm CD/DVD size disc using a 405 nm blue-violet laser. Because the Blu-ray Disc utilizes global standard "MPEG-2 Transport Stream" compression technology, highly compatible with digital broadcasting for video recording, a wide range of content can be recorded. In addition, the adoption of a unique ID written on a Blu-ray Disc realizes high quality copyright protection functions. The initial market for Blu-ray Discs is expected to be for data storage. The forecast predicts that approximately 167 400 Blu-ray devices will be sold worldwide for PC data storage in 2005, rising to 584 000 in 2006. In contrast, only 84 500 will be sold as set-top box devices in 2005, rising to 185 000 in 2006.

[Top of Page]

References

1. M. Camras: *Magnetic Recording Handbook*, Van Nostrand, New York 1988.
2. H. Sugaya, K. Yokoyama: “Video Recording,” in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording*, vol. **III, Video, Audio, and Instrumentation Recording**, "Chap. 2", McGraw-Hill, New York 1988.
3. D. Jiles: *Introduction to Magnetism and Magnetic Materials*, Chapman and Hall, London 1991.
4. B. K. Middleton: “Recording and Reproducing Process,” in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording Technology*, 2nd ed., Chap. 2, McGraw-Hill, New York 1996.
5. M. L. Williams, R. L. Comstock, *AIP Conf Proc* **5** (1972) 738.
6. V. A. J. Maller, B. K. Middleton, *Radio Electron Eng.* **44** (1974) 281.
7. J. J. Miyata, R. R. Hartel, *IRE Trans. Elect. Comp.* **EC-8** (1959) 159.
8. D. W. Chapman, *Proc. IEEE* **51** (1963) 394.
9. P. I. Bonyhard, A. V. Davies, B. K. Middleton, *IEEE Trans. Magn.* **MAG-2** (1966) 1.
10. E. D. Daniel, J. Watkinson: “Audio Recording,” in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording*, vol. **III**, Video, Audio, and Instrumentation Recording, Chap. 3, McGraw-Hill, New York 1988.
11. E. D. Daniel, I. Levine, *J. Acoust. Soc. Am.* **32** (1960) 1.
12. C. D. Mee: *The Physics of Magnetic Recording*, North-Holland, Amsterdam 1964.
13. J. C. Mallinson: *The Foundations of Magnetic Recording*, 2nd ed., Academic Press, New York 1993.
14. H. Neal Bertram: *Theory of Magnetic Recording*, 2nd ed., Cambridge University Press, Cambridge 1994.
15. J. Earl: *Cassette Tape Recorders*, Fountain Press, Watford 1977.
16. E. Koster: “Particulate Media,” in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording*, 2nd ed., McGraw-Hill, New York 1995.
17. E. C. Stoner, E. P. Wohlfarth, *Trans. Roy. Soc.* **A240** (1948) 599.
18. I. S. Jacobs, C. P. Bean, *Phys. Rev.* **100** (1955) 1060.
19. H. Inaba *et al.*, *IEEE Trans. Magn.* **29** (1993) 3602.
20. J. C. Lodder: “Magnetic Properties and Preparation of Thin-Film Magnetic Recording Media,” in K. H. J. Bushow, G. J.

Long, F. Grandjean (eds.): *High Density Digital Recording*, Chap. 3, NATO ASI Series, Dordrecht 1992.

21. S. R. Cumpson, B. K. Middleton, J. J. Miles, *J. Magn. Mat.* **154** (1996) 382.
22. B. Bhushan: "Tribology of Magnetic Head-Medium Interface," in K. H. J. Bushow, G. J. Long, F. Grandjean (eds.): *High Density Digital Recording*, Chap. 10, NATO ASI Series, Dordrecht 1992.
23. R. D. Fisher, W. H. Chilton, *J. Electrochem. Soc.* **109** (1962) 485. Links
24. Y. Maeda, K. Takei, *IEEE Trans. Magn.* **27** (1991) 4721.
25. G. Tsang, M. Chen, T. Yogi, *Proc. IEEE* **81** (1993) 1344.
26. S. Iwasaki, J. Hokkyo (eds.): *Perpendicular Magnetic Recording*, Ohmsha, Tokyo 1991.
27. J. C. Lodder: "Preparation, Microstructure and Magnetic Properties of Co – Cr Thin Films," in K. H. J. Bushow, G. J. Long, F. Grandjean (eds.): *High-Density Digital Recording*, NATO ASI Series, Dordrecht 1992.
28. R. E. Jones, C. D. Mee, C. Tsang: "Recording Heads," in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording Technology*, 2nd ed., Chap. 6, McGraw-Hill, New York 1996.
29. J. J. M. Ruigrok: *Short Wavelength Magnetic Recording*, Elsevier, Amsterdam 1990.
30. G. Bouwhuis *et al.*: *Principles of Optical Disc Systems*, Adam Hilger Ltd., Bristol 1985.
31. A. Marchant: *Optical Recording* Addison-Wesley Pub. Co., Reading MA 1990.
32. http://www.dvdforum.com (international association of hardware manufacturers, software firms and other users of Digital Versatile Discs).
33. http://www.iecee.org/ (official website of IECEE: Worldwide System for Conformity Testing and Certification of Electrical Equipment).
34. http://www.research.philips.com/ (Global information center of Koninklijke Philips Electronics N.V., NL).
35. http://www.blu-raydisc-official.org (official website of Blu-ray Disc founders).
36. http://www.blu-raydisc.info (Blu-ray Disc license website).
37. http://www.udo.com (official website of UDO).
38. P. Asthana, B. I. Finkelstein, A. A. Fennema, *IBM J. Res. Dev.* **40** (1996) 543; T. W. McDaniel, P. C. Arnett, *IBM J. Res. Dev.* **40** (1996) 311.
39. M. Mansuripur: *The Physical Principles of Magneto-optical Recording*, Cambridge University Press, Cambridge 1995.
40. I. Akasaki *et al.*, *Electron. Lett.* **32** (1996) 1105; Links S. Taniguchi *et al.*, *Electron. Lett.* **32** (1996) 552. Links
41. K. Schouwhamer-Immink: *Coding Techniques for Digital Recorders*, Prentice Hall, Englewood Cliffs, NJ 1991; J. P. J. Heemskerk, K. A. Schouwhamer-Immink, *Philips Tech. Rev.* **40** (1982) 157.
42. H. H. Hopkins, *J. Opt. Soc. Am.* **69** (1979) 4.
43. E. F. Stikvoort, J. A. C. van Rens, *IEEE J. On Selected Areas in Communications* **10** (1992) 191.
44. J. P. J. Heemskerk, *Appl. Opt.* **17** (1978) 2007.
45. J. A. Th. Verhoeven, Q. A. G. van Vlimmeren, B. P. Th. M. van Dijk: *High density CD, an industrial approach*, ODME, Veldhoven 1995.
46. B. van Dijk, J. Nijboer: "The CD family," in R. K. Jörgen (ed.): *Digital Consumer Electronics Handbook*, McGraw-Hill, New York in print.
47. J. Taylor: *DVD Demystified*, 2nd. Ed., McGraw-Hill, New York ▪year? ▪
48. J. Braat: *On the influence of substrate thickness in optical disc read-out*, in print.
49. J. P. W. B. Duchateau, B. A. J. Jacobs, Proc. Magneto-Optical Recording Int. Symp. 1996, *J. Magn. Soc. Jpn.* **20**, Suppl. no. S1 (1996) 193.
50. T. W. McDaniel, P. C. Arnett, "Optical Data Storage Media," *IBM J. Res. Develop* **40** (1996) 311.
51. M. Kaneko, *IEEE Trans. Magn.* **28** (1992) 2494.
52. S. M. Mansfield, W. R. Studenmund, G. S. Kino, K. Osato, *Opt. Lett.* **18** (1993) 305.
53. H. J. Mamin *et al. IBM. J. Res. Develop.* **39** (1995) 681.
54. D. W. Pohl, *IBM J. Res. Develop.* **39** (1995) 701.
55. K. Yashuda *et al. Jpn. J. Appl. Phys.* **32, pt. 1** (1993) 5210; Y. Kasami *et al. Jpn. J. Appl. Phys.* **3** (1996) 423.
56. P L M. Put, G. E. van Rosmalen, J. A. H. Kahlman, *Jpn. J. Appl. Phys.* **36** (1997) 572.
57. T. Tanaka, S. Kawata, *J. Opt. Soc. Am.* **A 13** (1996) 935.
58. W. E. Moerner, *Jpn. J. Appl. Phys.* **28** (1989) S-3, 221.
59. I. Lorgere *et al.*, *J. Opt. Soc. Am.* **B 13** (1996) 2229.
60. P. L. M. Put, P. A. Kraakman, R. D. Morton, J. Pasman, *Proc. SPIE-Int. Soc. Opt. Eng.* **1663** (1992) 336.
61. J. Isailovic: *Videodisc and Optical Memory Systems*, Prentice Hall Press, New York 1985.
62. K. C. Pohlmanh: *The Compact Disc: A Handbook of Theory and Use*, A-R Editions, Inc., Madison, Wisconsin 1989.
63. J. A. Th Verhoeven *et al.*: *Corporate CD-ROM publishing*, ODME, Veldhoven 1994.
64. H. Redlich, G. Joschko, *J. Audio Eng. Soc.* **35** (1987) 130.
65. J. S. Winslow, *IEEE Transactions on Consumer Electronics* **11** (1976) 318.
66. P. E. J. Legierse, J. H. T. Pasman, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **29** (1988) no. 2, 223.
67. J. G. Dil, C. A. Westdorp, *Appl. Opt.* **18** (1979) no. 18, 3198.
68. J. H. T. Pasman, *J. Audio Eng. Soc.* **41** (1993) no. 1/2, 19.
69. P. L. M. Put, R. D. Morton, J. J. Rush, H. P. Urbach, "The resolution limit of optical disc mastering," *Jpn. J. Apl. Phys.* **36** (1997) 539.

70. Y. J. Huh, J. S. Kim, T. Y. Nam, S. C. Kim, *Jpn. J. Apl. Phys.* **36** (1997) 403.
71. Y. Mitsuhashi, *Appl. Opt.* **25** (1986) 4013.
72. P. E. J. Legierse, *Trans. Inst. Met. Finish.* **65** (1985) 13.
73. N. V. Philips, US 4 668 549, 1987 (G. J. M. Lippits, P. E. J. Legierse, J. P. Dujardin).
74. S. Ohsawa *et al.*, *Appl. Opt.* **25** (1986) 4027.
75. N. V. Philips, DE 30 19 189, 1979 (E. Broeksema and A. Smeets).
76. Hitachi, EP 0 211 330 A 2, 1987 (R. Suko, H. Miwa, T. Tajima).
77. H. C. Haverkorn van Rijsewijk, P. E. Legierse, G. E. Thomas, *Philips Tech. Rev.* **40** (1982) 287.
78. J. Hennig, *Jpn. J. Appl. Phys.* **26 Suppl. 26–4** (1987) 9.
79. J. Hennig, *Kunststoffe* **75** (1985) 425.
80. G. Kaempf, H. Loewer, M. W. Witman, *Polym. Eng. Sci.* **27** (1987) 1421. Links
81. R. G. Gossink, *Angew. Makromol. Chem.* **145/146** (1986) 365.
82. J. Javornicki: *Photoplasticity* Chap. 5, Elsevier, Amsterdam 1974.
83. G. H. Werumeus-Buning, R. Wimberger-Friedl, H. Janeschitz-Kriegl, T. M. Ford in L. A. Kleintjens, P. J. Lemstra (eds.): *Integration of Fundamental Polymer Science and Technology II*, Elsevier Applied Science Publishers, London 1988, p. 405.
84. G. H. Werumeus – Buning, R. M. Gijsen, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **29** (1988) no. 2, 211.
85. H. Saito, T. Inoue, *J. Polym. Sci. Polym. Phys. Ed.* **25** (1987) 1629.
86. R. Wimberger-Friedl, H. Janeschitz-Kriegl in L. A. Kleintjens, P. J. Lemstra (eds.): *Integration of Fundamental Polymer Science and Technology II*, Elsevier Applied Science Publishers, London 1988, p. 410.
87. S. Tokuhara, Y. Onisawa, S. Iyoshi, H. Toba, *Proceedings of the Optical Society of America Meeting on Optical Data Storage*, paper FC-A 5–1, Monterey (California) 1984.
88. W. Siebourg, *Kunststoffe* **76** (1986) 917.
89. J. M. Newcome, *Proceedings of the Society of Plastics Engineers Annual Technical Conference (ANTEC '85)*, (1985) p. 773.
90. R. Riess, H. Loewer, *Plast. Eng*. (1985) no. 7, 27.
91. Meiki Seisakusho, DE 20 29 353, 1980 (M. Kunogi).
92. S. Anders, B. Hardt, *Kunststoffe* **77** (1987) 21.
93. M. Eichlseder, N. Eusemann, *Kunststoffe* **77** (1987) 150.
94. J. G. Kloosterboer, *Adv. Polym. Sci.* **84** (1988) 1. Links
95. J. G. Kloosterboer, G. J. M. Lippits, *J. Radiat. Curing* **11** (1984) no. 2, 10.
96. J. G. Kloosterboer, G. J. M. Lippits, H. C. Meinders, *Philips Tech. Rev.* **40** (1982) 298.
97. A. J. M. van den Broek, P. E. J. Legierse, *Photogr. Sci. Eng.* **28** (1984) 3.
98. A. G. Dirks, J. J. van den Broek, *Thin Solid Films* **96** (1982) 257. Links
99. J. J. Ponjée, P. N. T. van Velzen, *Philips Tech. Rev.* **44** (1988) 81.
100. P. Palermo, A. Korpel, G. Dickinson, W. Watson, *Opt. Laser Technol.* **8** (1977) 169.
101. *Plast. Technol.* **15** (1987) 13.
102. Hallmark, US 4 790 893, 1988 (J. B. Watkins).
103. P. de Lancle, *Mix* (1987) no. 12, 466. V. W. Schild, *Funkschau* (1987) no. 6, 71.
104. Dai Nippon Printing, JP 01 115 621 A, 1989 (N. Kazahaya), JP 01 152 044 A, 1989 (M. Tabita).
105. 3 M Company, US 4 363 844, 1982; US 4 519 065, 1985 (T. W Lewis, R. J. Anderson, D. J. Kerfeld).
106. Matsushita, JP-Kokai 63–273231, 1988 (T. Higashi).
107. Kurz L. GmbH, PT 087 372 A, 1989 (W. Reinhart).
108. L. K. H. van Beek, *Photogr. Sci. Eng.* **20** (1976) 88.
109. N. V. Philips, CA 1 119 451, 1982 (E. J. Spiertz, A. J. G. op het Veld).
110. N. V. Philips, US 4 125 401, 1978 (E. J. Spiertz, C. F. W. Flinsenberg, L. K. H. van Beek).
111. Thorn EMI EP 0 271 300, 1988 (A. Lang and P. Bloch).
112. W. S. Colburn, K. A. Haines, *Appl. Opt.* **10** (1971) 1636.
113. R. T. Ingwall, H. L. Fielding *Opt. Eng.* **24** (1985) 809.
114. W. J. Tomlinson, E. A. Chandross in J. N. Pitts *et al.* (eds.): *Adv. Photochem.* **12** (1982) 201.
115. H. Franke *et al.*, *Colloid Polym. Sci.* **262** (1984) 213. Links
116. J. Marotz *et al.*, *Physikertagung* **48** (1984) München; *Physik Verhandl.* **391** (1984) 1.
117. T. Luckemeyer, H. Franke, *Polym. Prepr.* **29** (Am. Chem. Soc. Div. Polym. Chem.) (1988) no. 2, 206.
118. V. B. Jipson, H. N. Lynt, J. F. Graczyk, *Appl. Phys. Lett.* **43** (1983) 27. Links
119. D. J. Gravesteijn, J. van der Veen, *Philips Tech. Rev.* **41** (1983/1984) 325.
120. V. B. Jipson, C. R. Jones, *J. Vac. Sci. Technol.* **18** (1981) 105.
121. M. Chen, V. Marello, *J. Vac. Sci. Technol.* **18** (1981) 75.
122. P. Kivits, B. Jacobs, P. Zalm, *J. Vac. Sci. Technol.* **18** (1981) 68. P. Kivits, R. de Bont, B. Jacobs, P. Zalm, *Thin Solid Films* **87** (1982) 21.
123. A. H. M. Holtslag, *J. Appl. Phys.* **66** (1989) no. 4, 1530.

124. D. Y. Lou, G. M. Blom, and G. C. Kenney, *J. Vac. Sci. Technol.* **18** (1981) 781.
125. O. Heavens: *Optical Properties of Thin Solid Films*, Academic Press, New York 1955.
126. A. E. Bell, R. A. Bartolini, *Appl. Phys. Lett.* **34** (1979) 275. Links A. E. Bell, F. W. Spong, *IEEE J. Quantum Electron.* **QE 14** (1978) 487.
127. R. Glang in R. Glang and E. Massel (eds.): *Handbook of Thin Film Technology*, McGraw Hill, New York 1970.
128. J. J. Scheer, J. Visser, *Philips Tech. Rev.* **39** (1980) 246.
129. D. Meyerhofer, *J. Appl. Phys.* **49** (1978) 3993.
130. L. Vriens, B. A. J. Jacobs, *Philips Tech. Rev.* **41** (1983/1984) 313.
131. J. A. Markvoort *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 134.
132. A. Saito, T. Maeda, H. Sugiyama, Y. Tsunoda, Topical Meeting on Optical Data Storage, *Technical Digest Series* **10**, 1987, p. 64.
133. International Electrotechnical Commission Publication 68–2-38, Test Z/AD, Geneva (1974).
134. D. J. Broer, L. Vriens, *Appl. Phys.* **32** (1983) 107.
135. C. R. Davis, W. Y. Lee, M. Chen, H. Ito, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 260.
136. V. Novotny, L. Alexandru, *J. Appl. Phys.* **50** (1979) 1215. Links V. Novotny, L. Alexandru, *J. Polym. Sci.* **24** (1979) 1321.
137. K. Y. Law *et al.*, *Appl. Phys. Lett.* **36** (1980) 884. Links
138. D. G. Howe, J. J. Wrobel, *J. Vac. Sci. Technol.* **18** (1981) 92. J. J. Wrobel, A. B. Marchant, D. G. Howe, *Appl. Phys. Lett.* **40** (1982) 928. Links
139. J. W. Wheeler, M. C. Gupta, C. E. Kurtz, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 39.
140. K. Y. Law, G. E. Johnson, *J. Appl. Phys.* **54** (1983) 4799. Links
141. G. E. Johnsonm, K. Y. Law, *Soc. Photo Opt.Instrum. Eng.* **420** (1983) 336.
142. M. F. Molaire, Topical Meeting on Optical Data Storage, *Technical Digest Series* **10**, Optical Society of America, Washington 1987, p. 90.
143. D. A. Hill, D. S. Soong, *J. Appl. Phys.* **61** (1987) 2132. Links
144. J. Griffiths, *Color and Constitution of Organic Molecules*, Academic Press, London 1976.
145. M. Itoh, S. Esho, K. Nakagawa, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 332.
146. E. Hamada *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng* 1078 (1989) 80.
147. J. M. Drexler, *J. Vac. Sci. Technol.* **18** (1981) 87.
148. T. L. Lichtenstein, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 96.
149. J. Cornet, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 86.
150. M. Mashita, N. Yasuda, *Proc. Soc. Photo Opt. Instrum. Eng.* **329** (1981) 190.
151. N. Yasuda, M. Mori, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 319.
152. J. Feinleib, J. de Neufville, S. C. Moss, S. R. Ovshinsky, *Appl. Phys. Lett.* **18** (1971) 254. Links
153. T. Ohta *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **695** (1986) 2.
154. D. J. Gravesteijn *et al.*, *Appl. Opt.* **26** (1987) no. 22, 4722.
155. Y. D. Tyan *et al.*, Topical Meeting on Optical Data Storage, *Technical Digest Series* **10**, 1987 , p. 44.
156. R. Barton, C. R. Davis, K. Rubin, G. Lim, Appl. Phys. Lett. **48** (1987) 1255.
157. Y. Nakane, N. Sato, H. Makino, S. Miyaoka, *Proc. Soc. Photo Opt. Instrum. Eng.* **529** (1985) 76.
158. S. Y. Suh, *Proc. Soc. Photo Opt. Instrum. Eng.* **695** (1986) 16.
159. V. B. Jipson, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 344.
160. J. H. Apfel, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 104.
161. A. Werner, H. Hibst, *Appl. Opt.* **28** (1989) no. 7, 1422.
162. P. F. Carcia *et al.*, *J. Appl. Phys.* **64** (1988) no. 4, 1671. Links
163. Du Pont, US 4 857 373, 1989 (P. F. Carcia, F. D. Kalk).
164. S. Y. Suh, *Proc. Soc. Photo Opt. Instrum. Eng.* **382** (1983) 196. S. Y. Suh, H. G. Craighead, R. E. Howard, L. M. Schiavone, *Proc. Soc. Photo Opt. Instrum. Eng.* **382** (1983) 199.
165. K. Gardner, R. J. Longman, P. R. Helfet, Topical Meeting on Optical Data Storage, *Technical Digest Series* **9**, Optical Society of America, Washington 1985, papers WDD 4 and WDD 5.
166. K. Aratani *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **1499** (1991) 209.
167. H. Awano *et al.* , *J. Appl. Phys.* **69** (1996) 4257.
168. T. Shiratori, E. Fujii, Y. Miyaoka, Y. Hozumi, *J. Magn. Soc. Jpn.* **22, Supp. S2** ,(1998) 47., Proc. MORIS/ISOM (1997), to be published in Jpn. J. Appl. Phys.
169. T. Suzuki, *MRS Bulletin* **21** (1996) 42. Links
170. T. Yashida, *Proc. IEEE* **82** (1994) 1492.
171. M. Shinoda *et al.*, *J. Magn. Soc. Jpn.* **22** S2 (1997) 173.
172. Y. Tanaka *et al.*, *Jpn. J. Appl. Phys.* **37** (1998) 2150.
173. K. Aratani *et al.*, *Proc. SPIE* **1499** (1991) 209.
174. J. Saito, M. Sato, H. Matsumoto, H. Akasaka, *Jpn. J. Appl. Phys.* **26 Suppl. 26–4** (1987) 155.
175. J. Saito *et al.*, *J. Magn. Soc. Jpn.* **22** S2 (1997) 133.
176. T. Hosokawa, A. Okamuro, M. Hashimoto, K. Miyata, *Proc. SPIE* **3109** (1997) 74.

177. P. F. Carcia, A. D. Meinhaldt, A. Suna, *Appl. Phys. Lett.* **47** (1985) 178. Links
178. W. B. Zeper, F. J. M. Greidanus, P. F. Carcia, C. R. Fincher, *J. Appl. Phys.* **65** (1989) 4971. Links
179. J. P. Krumme *et al.*, *J. Appl. Phys.* **66** (1989) 4393. Links
180. B. Bechevet *et al.*, *J. Appl. Phys.* **69** (1991) 4767. Links
181. G. Q. Di, S. Iwata, S. Uchiyama, *J. Magn. Mag. Mater.* **131** (1994) 242.
182. J. Schoenes: "Magneto-Optical Properties of Metals, Alloys and Compounds," in R. W. Cahn, P. Haasen, E. J. Kramer (eds.): *Materials Science and Technology*, vol **3A** VCH Publishers, Weinheim, Germany 1992, p. 147.
183. M. H. Kryder: "Magnetic Information Storage," in R. Gerber, C. D. Wright and G. Asti (eds.): *Applied Magnetism*, Kluwer Academic Publ., Doredrecht 1994.
184. M. H. Kryder, *J. Magn. Mat.* **83** (1990) 1.
185. C. D. Wright: "The Magneto-Optical Readout Process," in T. W. McDaneil, R. H. Victora (eds.): *Handbook of Magneto-Optical Data Recording*, Noyes Publications, Westwood 1997, p. 476.
186. Y. Wang *et al.*, *J. Magn. Soc. Jpn.* **17** (1993) 294.
187. P. Chaudhari, J. J. Cuomo, R. J. Gambino, *Appl. Phys. Lett.* **22** (1973) 337. Links
188. M. H. Kryder, *J. Appl. Phys.* **57** (1985) 3913. Links
189. M. Hartmann *et al.*, *J. Magn. Soc. Jpn.* **15** S1 (1991) 165.
190. P. Hansen: "Magnetic Amorphous Alloys," in K. H. J. Buschow (ed.): *Handbook of Magnetic Materials*, vol. **6** North-Holland, Amsterdam, 1991.
191. D. Raasch, *IEEE Trans. Magn.* **29** (1993) 34.
192. S. Takayama *et al.*, *J. Appl. Phys.* **61** (1987) 2364.
193. D. Mergel, P. Hansen, S. Klahn, *SPIE Proc.* **1274** (1990) 270.
194. T. Niihara *et al.*, *IEEE Trans. Magn.* **24** (1988) 2437.
195. T. K. Hatwar, *J. Appl. Phys.* **70** (1991) 6335. Links
196. F. E. Luborsky, *IEEE Trans. Magn.* **20** (1984) 1305.
197. P. Hansen, H. Heitmann, *IEEE Trans. Magn.* **25** (1989) 4390.
198. P. F. Carcia, *J. Vac. Sci. Technol.* **A5** (1987) 1975.
199. P. F. Carcia, *J. Appl. Phys.* **63** (1988) 5066. Links
200. F. J. A. den Broeder, H. C. Donkersloot, H. J. G. Draaisma, W. J. M. de Jonge, *J. Appl. Phys.* **61** (1987) 4317. Links
201. H. J. G. Draaisma, W. J. M. de Jonge, F. J. A. den Broeder, *J. Magn. Magn. Mat.* **66** (1987) 351.
202. W. B. Zeper, F. J. A. M. Greidanus, P. F. Carcia, *IEEE Trans. Magn.* **25** (1989) 3764.
203. Y. Ochiai, S. Hashimoto, K. Aso, *IEEE Trans. Magn.* **25** (1989) 3755.
204. F. J. A. M. Greidanus *et al.*, *Jpn. J. Appl. Phys.* **28** S28–3 (1989) 37.
205. H. W. van Kesteren, W. B. Zeper, *J. Magn. Mag. Mater.* **120** (1993) 271.
206. P. G. Pitcher, J. Miller, D. P. A. Pearson, P. D. Gurney, *J. Magn. Soc. Jpn.* **17** (1993) 91.
207. P. F. Carcia, M. Reilly, *Appl. Phys. Lett.* **58** (1991) 191. Links
208. W. B. Zeper *et al.*, *IEEE Trans. Magn.* **28** (1992) 2503.
209. M. Gomi, M. Abe, *Mat. Res. Soc. Symp. Proc.* **150** (1989) 121.
210. K. Shono et al. *MRS Conf., San Diego*, paper F 3–2 (1989).
211. B. Bechevet *et al.*, *J. Appl. Phys.* **69** (1991) 4767. Links
212. T. Okuda, T. Katayama, K. Satoh, H. Yamamato, *J. Appl. Phys.* **69** (1991) 4580. Links
213. M. H. Kryder, *Thin Solid Films* **216** (1992) 174. Links
214. A. Itoh, K. Nakagawa, *Jpn. J. Appl. Phys.* **31** (1992) L790.
215. T. Suzuki, *J. Appl. Phys.* **69** (1991) 4756. Links
216. K. A. Rubin *et al.*, *SPIE Proc.* **2338** (1994) 247.
217. K. Tsutsumi, T. Fukami, *J. Magn. Magn. Mat.* **118** (1993) 213.
218. A. Takahahi *et al.*, *J. Magn. Soc. Jpn.* **22** (1997) S2, 67.
219. H. Awano *et al.*, *SPIE Proc.* **3109** (1997) 83.
220. D. Turnbull, *Contemp. Phys.* **10** (1969) 473. Links
221. D. R. Uhlmann, *J. Non-Cryst. Solids* **7** (1972) 337. Links
222. T. Minewura, H. Andoh, Y. Maeda, *J. Appl. Phys.* **63** (1988) no. 9, 4632. Hoechst, US 4 478 782, 1984 (A. J. East, J. E. Kuder).
223. Hewlett Packard Company, US 4 405 993, 1983 (H. Birecki, R. A. Burmeister, F. J. Kahn).
224. H. Birecki, S. Naberhuis, F. J. Kahn, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 194.
225. J. Pereda, *Proc. Soc. Photo Opt. Instrum. Eng.* **369** (1982) 284.
226. Röhm GmbH, EP-A 231–857 A-858 A; DE 3603–267 A, 1987 (H. Eich, J. Wendorff, B. Reck).
227. H. W. Schmidt: Ph. D. Thesis, University of Mainz 1985.
228. W. D. McIntyre, D. S. Soane, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **29** (1988) no. 2, 197.
229. I. Cabrera *et al.*, *Polym* . Prepr. (Am. Chem. Soc. Div. Polym. Chem.) 29 (1988) no. 2, 200.
230. H. E. Ringsdorf, *Makromol. Chem.* **183** (1982) 1245. V. P Shibaev *et al.*, *Polym. Commun.* **24** (1983) no. 12, 364.
231. Canon KK, JP-Kokai 62149–482 A, JP - 483 A-484 A, JP - 487 A, JP 62151–384 A, JP -385 A, 1987 (H. Matsuda *et*

al.).
232. S, Tazuke, *Jpn. J. Appl. Phys.* **26, Suppl. 26–4** (1987) 3.
233. Plessey, Ltd., GB 1 464 603, 1977 (H. G. Heller);Traqson Ltd., WO 88/0 95 49, 1988 (S. C. Wallace).
234. Japan Agency of Science and Technology, JP-Kokai 62209–186 A, 1987 (Y. Suzuki, K. Ichimura, M. Sakuragi).
235. W. Tomlinson, *Appl. Opt.* **23** (1984) no. 22, 3990.
236. L. M. Ralston, *Proc. Soc. Photo Opt. Instrum. Eng.* **420** (1983) 186.
237. G. H. Brown (ed.): *Techniques in Chemistry III: Photochromism*, Wiley-Interscience, New York 1971.
238. M. Irie, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **29** (1988) no. 2, 215.
239. C. Ishimoto *et al.*, *Appl* . Phys. Lett. 49 (1986) no. 25, 1677.
240. Sony, EP 0 218 263, 1987 (C. Ishimoto, J. Seto, H. Tomimuro).
241. J. M. Pearson, *CRC Crit. Rev. Solid State Mater. Sci.* **13** (1986) no. 1, 1.
242. R. West, *J. Organomet. Chem.* **300** (1986) 327. Links
243. R. Miller, *J. Am. Chem. Soc.* **107** (1985) 2172. Links
244. R. Miller, *Macromolecules* **19** (1986) 611.
245. Quantex Company, US-A 870 877, 1986 (J. A. Lindmayer). *Laser Focus World* **25** (1989) no. 11, 119
246. Johns Hopkins University, US 4 371 883, 1983 (D. O. Cowan, T. O. Poehler, R. S. Potember).
247. R. C. Hoffman, R. S. Potember, *Appl. Opt.* **28** (1989) no. 7, 1417.
248. IBM Corporation, US 4 101 976, 1978 (G. Castro, D. Haarer, R. M. McFarlane, H. P. Trommsdorff).
249. W. E. Moerner, T. P. Carter, C. Brauchle, *Appl. Phys. Lett.* **50** (1987) no. 8, 430. Links
250. W. E. Moerner, R. M. McFarlane, W. Lenth, Topical Meeting on Optical Data Storage, *Technical Digest Series* **10**, Optical Society of America, Washington 1987, p. 151.
251. M. Yoshimura *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **1078** (1989) 326.
252. W. Lenth *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **695** (1986) 216.
253. T. Tani *et al.*, *Jpn. J. Appl. Phys.* **26, Suppl. 26–4**, (1987) 77.
254. D. Haarer, *Jpn. J. Appl. Phys.* **26, Suppl. 26–4**, (1987) 227.
255. U. P. Wild *et al.*, *Jpn. J. Appl. Phys.* **26, Suppl. 26–4**, (1987) 233.
256. U. Bogner, K. Beck, P. Schatz, M. Maler, Topical Meeting on Optical Data Storage, *Technical Digest Series* **9**, Optical Society of America, Washington (1985) paper WBB 2.
257. J. Feinleib, S. R. Ovshinsky, *J. Non-Cryst. Solids* **2** (1970) 564.
258. M. Chen, K. Rubin, R. W. Barton, *Appl. Phys. Lett.* **49** (1986) 502. Links
259. S. R. Ovshinsky, *J. Non-Cryst. Solids* **141** (1992) 200. Links
260. N. Yamada, *Mater. Res. Soc. Bulletin*, (1996) Sept., 48.
261. A. W. Smith, *Appl. Opt.* **13** (1974) 795.
262. M. Terao *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **695** (1986) 106.
263. T. Nishida *et al.*, *Appl. Phys. Lett.* **50** (1987) 667. Links
264. J. H. Coombs, A P. J. M. Jongenelis, W. van Es-Spiekman, B. A. J. Jacobs, *J. Appl. Phys.* **78** (1995) 4918. Links
265. N. Yamada *et al.*, *Jpn. J. Appl. Phys.* **26 Suppl. 26–4** ,(1987) 61.
266. N. Yamada, E. Ohno, K. Nishiuchi, N. Akahira, *J. Appl. Phys.* **69** (1991) 2849. Links
267. J. Gonzales-Hernandez *et al.*, *Appl. Phys. Commun.* **11** (1992) 557.
268. K. A. Agaev, A. G. Talybov, *Kristallografiya* **11** (1966) 454.
269. I. I. Petrov, R. M. Imamov, Z. G. Pinsker, *Kristallografiya* **13** (1968) 417.
270. T. Ohta *et al.*, *Jpn. J. Appl. Phys.* **28 Suppl. 28–3**, (1989) 123.
271. M. Terao *et al. Proc. Soc. Photo Opt. Instrum. Eng.* **1078** (1989) 2.
272. H. Iwasaki *et al.*, *Jpn. J. Appl. Phys.* **31** (1992) 461.
273. H. Iwasaki *et al.*, *Jpn. J. Appl. Phys.* **32** (1993) 5241.
274. Y. Kageyama, H. Iwasaki, M. Harigaya, Y. Ide, *Jpn. J. Appl. Phys.* **35** (1996) 500.
275. T. Handa, J. Tominaga, S. Haratani, S. Takayama, *Jpn. J. Appl. Phys.* **32** (1993) 5226.
276. H. J. Borg, J. P. W. B. Duchateau, *Proc. Soc. Photo Opt. Instrum. Eng.* **3109** (1997) 20.
277. S. Morita, M. Nishiyama, T. Ueda, *Jpn. J. Appl. Phys.* **36** (1997) 444.
278. T. Ishida *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **2338** (1994) 121.
279. M. Tero *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **3109** (1997) 60.
280. M. Okada *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **2514** (1995) 329.
281. H. Kobori *et al.*, *Proc. Soc. Photo Opt. Instrum. Eng.* **2514** (1995) 127.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Inorganic Polymers

Charles E. Carraher, Jr., College of Science, Florida Atlantic University, Boca Raton, Florida 33431, United States Florida Center for Environmental Studies, Palm Beach Gardens, Florida 33410, United States

Charles U. Pittman, Jr., Department of Chemistry, Mississippi State University, Mississippi State, Mississippi 39762, United States

1. Introduction 1
2. Polysilanes, Polygermanes, and Polystannanes 2
3. Poly(Carborane – Siloxanes), Polycarbosilanes and Polysilazanes 4
4. Polyphosphazenes 6
5. Poly(Boron Nitride) 7
6. Poly(Sulfur Nitride) 8
7. Polysulfur 9
8. Organometallic Polymers 10
9. Sol – Gel Inorganic Polymers 11
10. Inorganic Fibers and Whiskers 14
11. Biomedical and Antimicrobial Polymers 15
12. Boron Polymers 17
13. Aluminum Polymers 19
14. Polymers with –Si–O–M–O– Backbones 20
15. Poly(Carbon Disulfide), Poly(Carbon Diselenide), and Polythiocyanogen . . . 21
16. Phosphorus-Containing Polymers . . . 23
17. References 24

1. Introduction

Inorganic polymers are macromolecules linked by covalent bonds, whereby there is an absence or near-absence of organic units within the backbone. Inorganic polymers include many of the most important "natural" materials as well as many synthetic materials. Natural inorganic polymers include many of the rocks about us, diamond, graphite, sulfur, boric oxide, silica, polyphosphates, quartz, and glass. Sulfur, selenium, and tellurium form macromolecules. Chalcogenide glasses are three-dimensional polymers, as are the asbestoses and a number of silicates. Most of these are industrially important, and some exhibit properties that are utilized in high-technology applications. Thus, arsenic sulfide, a chalcogenide glass, is used as an infrared-transparent window, and modified zeolites as selective chelating agents and catalysts.

Many scientific societies, including the National Research Council [1], have emphasized that the major hindrance to progress in most technological areas is the lack of suitable materials. The search for new materials in the 1970s and 1980s generally focused on rearranging, mixing, blending, and copolymerizing well-known organic polymers. Although great success has been achieved, many new applications require materials whose performance greatly exceeds that of currently available materials. Such performance requirements can be achieved only by materials containing other elements in addition to those found in the usual carbon-based polymers. For example, although the first polysilane, poly(dimethylsilylene), was prepared in 1924 by Kipping [2], the initial products were infusible and insoluble. In fact, these properties were so unfavorable that polysilanes were not developed until more than five decades later. Today, polysilanes offer combinations of chemical, mechanical, and electrical properties not common to any other commercially available class of materials.

Major applications of inorganic polymers include their use as bulk and specialty building materials, biologically inert and biologically active materials, high-strength materials, catalysts, and speciality coatings. The breadth of the field in relation to more traditional polymer chemistry is exemplified by comparing the number and variety of elements typically found in commercial organic polymers (C, H, N, Cl, F, O) with the number and variety of readily available inorganic materials. The relative abundance of elements in the earth's crust also attests to the importance of noncarbon-based materials. In fact, the most abundant natural building materials are

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_241

based on silicon – oxygen polymers, including the majority of rocks, soil, and sand.

The alternative of utilizing heteroatom backbones or noncarbon-based polymers is attractive because of the great variety of available reactants. Thus, a wide range of compounds can be exploited to obtain materials with specific properties. Furthermore, these materials are of interest because the resulting bonds may have bond energies greater than that of the C–C single bond in alkanes (ca. 343 kJ/mol). Approximate bond energies for other elements, in kilojoules per mole, are B–N 389, Si–N 439, Si–O 795, P–N 615, and Be–O 448. Bonding in these materials is a combination of ionic (nondirectional) and covalent (directional) components; the ionic character increases as the electronegativity difference between adjacent atoms increases. Furthermore, the tendency to form multiple bonds is high for many of the elements involved. The thermal stability of many inorganic and organometallic materials has been described [15].

Inorganic and organometallic polymers represent a rapidly growing field of chemical research and already have many applications [16–26]. Any division between inorganic and organometallic polymers is necessarily somewhat arbitrary. The classification used here is convenient but far from perfect. The reader is therefore referred to several other articles which review the scope of organometallic polymers and discuss many other systems which could also be regarded as inorganic polymers [20], [22–27].

2. Polysilanes, Polygermanes, and Polystannanes

Carbon has an unique ability to form strong single, double, and triple bonds to itself. Because of the relatively high bond energies (C–C single bonds are of the order of 340 kJ/mol) compounds containing carbon in their backbone do not readily cleave. By comparison, fellow members of group 14 have only recently been synthesized because of the low bond energies (generally on the order of 85 – 125 kJ/mol). Polysilanes may have been prepared by Kipping in 1924 [2]. Only recently have soluble, well-characterized polysilanes been synthesized [3–5]. (Topics related to polysilanes have been recently reviewed [6–14].)

Production. Polysilanes have been synthesized by four methods:

1) Reductive coupling
2) Dehydrogenative coupling
3) "Masked" dienes
4) Anionic ring opening

Attempts to prepare polysilanes in the late 1940s and 1950s gave insoluble, intractable materials; the major route employed was the reaction of dichlorosilanes with sodium (Fig. 1). For example, polydimethylsilane [*28883-63-8*] was investigated at Union Carbide but found to be insoluble [28]. The key developments leading to soluble polysilanes were the introduction of large substituent groups to decrease order, the use of two different substituents, and the copolymerization of different monomers (Fig. 1) [29], [30]. This resulted in high molecular mass (> 400 000), tractable, soluble polymers, which could be purified easily, cast into films, and molded.

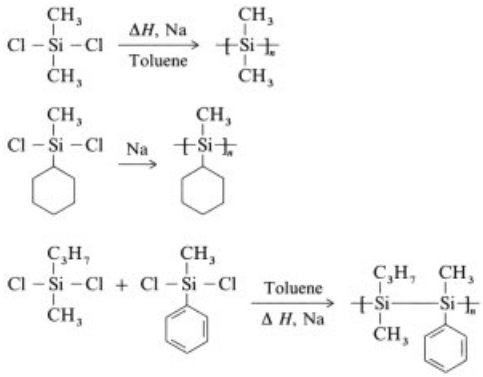

Figure 1. Synthesis of polysilanes

Improved methods for preparing high molecular mass polysilanes are being sought, because the reaction of dichlorosilanes with sodium frequently gives low yields of high molecular mass polymers and because binodal molecular mass distributions are obtained [28], [29],[31–35]. Carrying out the coupling reaction at low temperature by using ultrasonication gives mononodal polymers with low polydispersity [35].

Alternating copolymers are formed from the coupling of dichlorodisilanes with dilithium salts of 1,2-diethynyldisilanes [36] or dilithio

polythiophenes [36], [37]. These products exhibit $\sigma - \pi$ conjugation. Some of these materials also undergo solid-state transitions to liquid-crystalline mesophases [36].

Water-soluble products are formed by coupling silanes with oligomeric poly(ethylene glycol) side chains [38].

Dehydrogenative coupling of dialkyl and monoalkyl silanes in the presence of transition metal catalysts such as titanocene and zirconocene dialkyls has given polysilylenes [36–48]. Functionalized polysilanes have also been produced. When the side chains contain amine functional groups, the products can be made water-soluble by formation of an amine salt network [49–51]. Dendrimeric [52–54] products can also be formed by dehydrogenative coupling. These products exhibit σ conjugation in many directions and show good conductivity. They have been used as negative-type photoresists because exposure to the air or light gives insoluble products.

Polygermanes and polystannanes have also been prepared by similar routes [55–58]. Many of these materials show better σ conductivity than the corresponding polysilanes and have been used as light-emitting diodes [59].

Properties and Uses. Much of the interest in the polysilanes, polygermanes, and polystannanes involves their σ delocalization and their $\sigma - \pi$ delocalization when coupled with arenes or acetylenes. This is not unexpected since silicon exists as a covalent network similar to diamond. In exhibiting electrical conductivity, germanium and tin show more typical "metallic" bonding. Some polystannanes have been referred to as "molecular metals" [60]. Conductivity is increased by doping [61], illumination [62], and application of an electric field.

Because of a number of interesting electronic and physical properties exhibited by polysilanes, a number of potential uses have been suggested and/or shown to exist, including precursors of β-SiC fibers [63–71], impregnation of ceramics [71], [72], polymerization initiators [73], [74], photoconductors for electrophotography [75–78], nonlinear optical materials [79–87], mid-UV photolithography bilayer materials [88–91], contrast enhancement layers in photolithography [92], deep UV-sensitive photoresists [90], [93–100], and self-developing by excimer laser or deep UV exposure [99–101]. The unusual absorption spectra of polysilanes have indicated potential uses in a number of conducting areas. The copolymer of polysilastyrene, when doped with arsenic pentafluoride, becomes a semiconductor. Conducting films have been reported [102].

Polysilanes display UV absorption due to silicon – silicon bonds of the backbone. For example, poly(cyclohexylmethylsilane) exhibits an absorption band at 326 nm [30]. This absorption is not found in saturated carbon chains. Polysilanes with phenyl side chains show strong absorption near 330 nm, resulting from interaction between the phenyl groups and the silicon backbone, which acts as a $\sigma \rightarrow \sigma*$ or $\sigma \rightarrow \pi*$ chromophore.

Polysilanes have a high sensitivity to degradation by UV light. After light is absorbed, chain cleavage occurs with measured quantum yields of 0.20 – 0.97. This has led IBM to examine polysilanes as resists for bilayer UV lithography [103–105]. They serve as excellent reactive-ion etching barriers for bilevel resist applications because a protective layer of silicon dioxide is formed during exposure of the polymers to oxygen plasmas [104], [105].

The wavelength at which maximum absorption occurs ($\lambda_{\max}$) depends on the polysilane chain length. During irradiation nonlinear bleaching (decrease in UV absorption) occurs, because the polymer undergoes chain cleavage to give shorter fragments of lower absorptivity. As the irradiation continues, the polysilanes become increasingly transparent. This unique bleaching latency is useful in contrast-enhanced lithography [104], [105].

The $\lambda_{\max}$ for UV absorption also depends on the conformation of the polysilane backbone [107], [108]. Polysilanes having a planar zigzag backbone conformation absorb at much longer wavelengths than those in which the backbone is disordered or helical. The solid materials are strongly thermochromic [107]. For example, the $\lambda_{\max}$ of poly(di-*n*-hexylsilane) in solution or immediately after baking at 100 °C is 317 nm, but after 3 h at 21 °C films exhibit $\lambda_{\max} = 371$ nm.

Polysilanes are excellent photocatalysts for making high molecular mass vinyl polymers. On irradiation, chain cleavage occurs to give free radicals [106], which initiate vinyl polymerization and curing. The 3M Company has

commercialized this process. Polysilane initiators are relatively insensitive to termination of polymerization by oxygen.

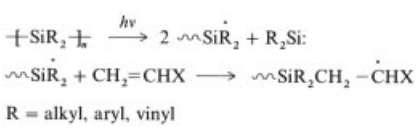

Pioneering work by YAJIMA, [109] showed that poly(dimethylsilane) can be converted into β-SiC fibers with very high tensile strength ($350\,kg/mm^2$) by a series of pyrolysis steps. Thermolysis of poly(dimethylsilane) at 400–450 °C gives a poly(carbosilane) which is fractionated and then melt-spun into fibers. The resulting fibers are oxidized at the surface in air at 350 °C (cross-linking occurs) to provide rigidity, then pyrolyzed at 1200 °C under nitrogen to give crystalline β-SiC (see also → Fibers, 5. Synthetic Inorganic, Chap. 3.2.2.).

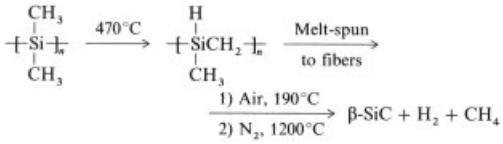

Objects have been manufactured from silicon carbide by Shin Nisso Kako using poly(silastyrene):

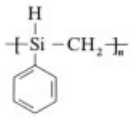

These materials contain 20–30 vol% of microvoids, which lower the ultimate strength slightly but also inhibit crack growth, so that they can be machined to exact dimensions.

Delocalization in the Si–Si σ-bond framework makes polysilanes potential electrical conductors. Although neutral polysilanes are insulators, they become semiconductors when doped with AsF_6 or SbF_5 [31]. Conductivities up to $0.5\,\Omega^{-1}\,cm^{-1}$ have been measured. Polysilanes are excellent photoconductors operating through hole migration with high hole mobilities ($10^{14}\,cm^2\,V^{-1}\,s^{-1}$ at ambient temperature) [110]. This has possible applications in electrophotography and communications technology. The discovery that poly(methylphenylsilane) has nonlinear optical properties suggests eventual applications in laser technology [111].

A variety of linear, two- and three-dimensional polysilanes are under investigation for use as specialty caulks, adhesives, oils, sealants, viscosity regulators [31], [112], [113], biologically active materials [113], [114], as well as in electrophotography [115] and reprography [116]. Although these applications are largely in the research stage, the unique all-silicon backbone provides an alternative to carbon and siloxane polymers.

3. Poly(Carborane–Siloxanes), Polycarbosilanes and Polysilazanes

Silicon forms bonds with itself to give polysilanes, but it can also bond to other elements. Here several of these families are considered. (For polysiloxanes, see → Silicones).

Poly(Carborane–Siloxanes). The poly(carborane–siloxanes) have a linear structure (**1**) in which R^1 and R^2 can be alkyl, fluoroalkyl, or aryl groups. The main chain contains carborane polyhedra. Although the polyhedral carborane $C_2B_{10}H_{10}$ is most commonly employed in the preparation of poly(carborane–siloxanes), the *closo*-carboranes $C_2B_5H_7$ and $C_2B_{10}H_{12}$ have also been used as precursors [117], [118]. Major interest in this class of polymers results from (1) the need for enhanced flame resistance and (2) their high thermal and oxidative stability.

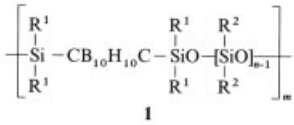

1

Production. The hydrolysis–condensation of carborane–siloxane monomers containing terminal Si–Cl bonds is one synthetic route to the poly(carborane–siloxanes) [119]. Cohydrolysis–condensation leads to other compositions of **1** [119]:

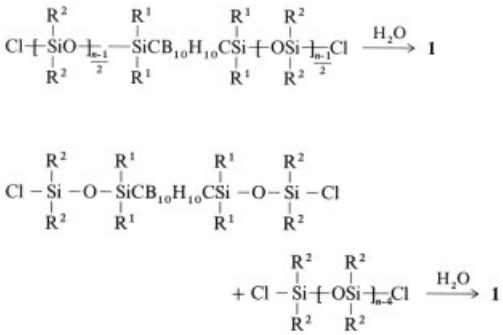

Another route involves reacting a dihalosilane with a carborane-containing dihydroxydisilane [120]. Synthetic routes to the required monomers have been reported [121], [122].

Properties and Uses. For polymers of series **1**, where $R' = R = CH_3$ and $n = 1-5$, the glass transition temperature (T_g) decreases with increasing siloxane content from −42 to −88 °C as n increases from 1 to 5 [123]. Replacing R^2 by phenyl groups gives a completely amorphous polymer, but T_g increases to −12 °C. When R^2 is a mixture of methyl and phenyl groups, T_g is lower. Trifluoropropyl-modified polymers are amorphous and exhibit T_g values that decrease with increasing n.

The thermal stability of poly(carborane–siloxanes) is excellent. In an inert atmosphere, rapid weight loss occurs only upon heating above 400 °C [123]. Phenyl-substituted polymers are more stable than their methyl analogues [124]. The amorphous polymers are elastomers and are formulated with fillers and vulcanized by using standard silicon technology. The high-temperature capabilities have resulted in the application of poly(carborane–siloxanes) as stationary phases in gas chromatography and the fabrication of these polymers into O-rings, gaskets, and wire coatings for operating temperatures >300 °C [117], [120].

Polycarbosilanes have generally been made as precursors for silicon carbide. Recently, polysilaethylene and related functionalized products have been prepared [125–127]. Polysilaethylene has a low T_g (about − 140 °C) and melts at about room temperature (ca. 25 °C). Ease of substitution of the chloro derivative by alkoxide, amine, and other functional groups may lead to a number of interesting products with derivative chemistry like that of polyphosphazenes. Fluorinated polysilaethylenes have also been made [128], [129]. It is suggested that these materials might exhibit piezo and pyroelectric properties similar to those of poly(vinyl fluoride).

Polysilazenes and Poly(*N*-methylsilazenes). Polysilazenes. Most research focuses on the use of polysilazanes as precursors to ceramics [130–133]. These products have not proved to be very useful since the ceramic yield is low (about 50 % at 1000 °C) and the free carbon content is high (about 30 % at 1000 °C).

They are mainly synthesized by aminolysis and ammonolysis of dichlorosilanes [134], [135], deamination or redistribution reactions of aminosiliazanes [136], [137], and more recently by the anionic and cationic ring-opening polymerization of cyclosilazanes. The latter products are soluble in typical organic liquids. The Si–N bond energy is greater than the Si–O bond energy. The temperature of initial weight loss for the products formed from the ring-opening reaction is about 50 °C higher than that of polydimethylsiloxane. Future applications will focus on areas similar to those of polysiloxanes themselves.

4. Polyphosphazenes

Polyphosphazenes are a well developed versatile class of inorganic polymers with over 700 different polymers formed in the past three decades [138–142].

Polyphosphazenes are polymers containing alternating nitrogen and phosphorus atoms in the backbone and alternating single and double bonds [143]. Phosphorus is pentavalent and has two pendant substituents (X and Y) which can be alkyl, aryl, alkoxy, aryloxy, arylamino, alkylamino, halogen, or pseudohalogen. Variation of X and Y leads to a wide variety of materials with highly diverse properties and applications [143–147].

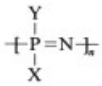

2

Production. Polyphosphazenes can be prepared in several ways [143–147]. Reaction of ammonium chloride with phosphorus pentachloride yields a mixture of cyclic and linear oligomeric phosphazenes with a degree of polymerization ($\overline{DP}$) of up to 20:

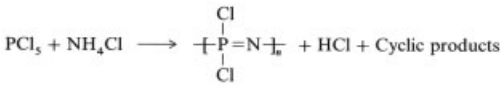

The ring opening polymerization of hexachlorocyclotriphosphazene [*940-71-6*], [*16422-79-0*] (**3**) leads to a rubbery material, but when a carefully purified starting material is polymerized under controlled conditions, a degree of polymerization of up to 15 000 can be obtained and the polymer can be purified readily.

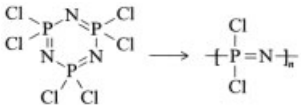

3

Poly(dichlorophosphazene) [*26085-02-9*] degrades slowly in a moist environment. However, the chlorine atoms are readily (and completely) displaced by many different nucleophiles. This reaction leads to a variety of highly stable polymers [143], [146–150].

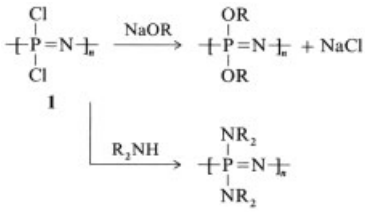

The condensation polymerization of appropriate *N*-silylphosphoramines has resulted in the successful preparation of the first polyphosphazene to have only alkyl substituents on phosphorus [151]:

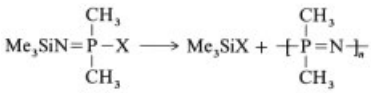

4

Recently, polyphosphazenes have been prepared by room-temperature cationic (living) polymerization of phosphoranimines [152–154]. Certain initiators allow the formation of multiple initiating sites and the formation of star-branched polyphosphazenes with a narrow polydispersity. Block structures with polymers such as poly(ethylene oxide) have been made that contain reactive amine end groups [155]. Copolymers with polyurethanes have also been produced [156].

A number of polythionylphosphazenes have been produced with a variety of aryloxy, alkoxy, and amino appendages [157].

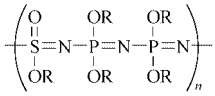

Polyphosphazenes have also been coupled with organosilicon compounds [158].

Properties. The poly(organophosphazenes), unlike their precursors the poly(dichlorophosphazenes), are resistant to hydrolysis and can be vulcanized to form elastomers. Copolymers with trifluoroethoxy and heptafluorobutoxy (1 : 1) groups are elastomeric with $T_g = -77\,°C$ [148], whereas T_g values for poly(arylaminophosphazenes) are much higher (see Table 1). Poly(aryloxyphosphazenes) tend to be intermediate between these two classes.

Table 1. Glass transition temperatures (T_g) and crystalline melting points (T_m) for some polyphosphazenes, $-[P(Y)_2=N-]_n$

Substituent (Y)	T_g, °C	T_m, °C
Cl	−63	30
OCH_3	−76	
OCH_2CF_3	−66	242
OC_6H_5	6	390
$N(H)C_6H_5$	105	

Polyphosphazenes exhibit some of the lowest known rotational barriers for skeletal backbone bonds in polymers (as low as 420 J/mol of repeating units), which is consistent with their low T_g values. Two first-order transitions are usually found, with a temperature interval of ca. 150 – 200 °C [143], [146], [147]. The nature of the lower temperature transition resembles the change into a mesomorphic state similar to that observed in nematic liquid crystals. In practice,

polyphosphazenes are usually soft just above this lower transition temperature, which allows compression molding of films to be carried out. Table 1 lists representative data.

The thermal stability of some polyphosphazenes is outstanding. Several classes exhibit onset of weight loss above 300 °C, as shown by thermogravimetric analysis (TGA), but these studies overestimate the stability. Poly(diphenoxyphosphazene) [*28212-48-8*], for example, undergoes rapid depolymerization above 150 °C. Fluoroalkoxyphosphazene vulcanizates have excellent resistance to solvents (e.g., lubricants and fuels) and display low volume swells after long immersion in lubricants [159]. Polyphosphazenes also display good flame-retardant properties, with oxygen index values in the range 20 – 70, and they give off only moderate amounts of smoke [160], [161].

Uses. The favorable properties of polyphosphazenes, as well as the high elastomer elongations, which are greater than those of polysiloxanes from −60 to 200 °C, have resulted in the use of various polyphosphazenes as gaskets, specialty damping materials, and petroleum piping for arctic applications. They have also been used as specialty membrane materials [150]. Polyphosphazene – salt complexes, which have the highest ionic conductivities at ambient temperature of all polymer – salt electrolytes, are reported to be semiconducting materials [150]. Other uses include catalytic templates [150], biomedical materials [160], and their most prominent use as specialty elastomers in O-rings, gaskets, and fuel hoses [143], [161–165].

The first commercial polyphosphazene elastomers were synthesized and subsequently developed by the Firestone Tire and Rubber Co. [166] and represent the first new class of semi-inorganic elastomers to be developed commercially since silicone rubber. Ethyl Corporation holds the rights to many commercial uses and produces both commercial polymers and finished products from this interesting class of materials [167].

Thin films of amino-substituted polythionyl phosphazenes exhibit an unusual permeability to molecular oxygen [168]. When a phosphorescent dye that is quenched by oxygen is dispersed in the polymer film, the product can be used as a visually interpretable oxygen sensor. One potential use is to follow differences in pressure across a surface area such as a wing of a plane.

5. Poly(Boron Nitride)

Boron nitride [*10043-11-5*] occurs in three crystalline forms – α, β, and γ (→ Boron Carbide, Boron Nitride, and Metal Borides, Chap. 2.). The α-form is a layered, hexagonal polymer much like graphite. A crystalline, tetrahedral network polymer similar to diamond also exists [169–172]. Although boron nitride is isoelectronic with the diamond and graphite forms of carbon, it exhibits different properties. For example, layered boron nitride is white and an electrical insulator. Thus its electrons appear to be highly localized in contrast to those of graphite. Like graphite, layered boron nitride is soft and easily machined. Articles are fabricated by sintering or machining. These articles can be used in air up to 800 °C or in an inert atmosphere up to 1600 °C. Boron nitride has a Young's modulus : density ratio of 20 (versus 4 for iron) and a tensile strength : density ratio of 8 (versus 1.7 for glass and 0.3 for iron). The tetrahedral form of boron nitride is almost as hard as diamond and is therefore employed as a substitute for diamond in both cutting and grinding and in jewelry. Industrial applications include its use as an abrasive, insulator, and refractory material.

The hexagonal modification of boron nitride has been prepared from combinations of cheap boron and nitrogen compounds, such as $B(OH)_3$, $(NH_2)_2CO$, and N_2, or KBH_4 and NH_4Cl [162–167, 169–175]. More recently, chemical vapor deposition (CVD) methods have used mixtures of BCl_3 and NH_3, BF_3 and NH_3, or B_2H_6 and NH_3 to produce hexagonal boron nitride [175, 178–180].

Polymeric precursors to boron nitride are beginning to attract attention. The reaction of difunctionalized borazines with bis(trimethylsilyl)-amines gives soluble oils that form films which can be pyrolyzed at 1200 °C to yield a boron carbide – boron nitride ceramic, but removal of all of the carbon has not yet been achieved [173]. The use of trifunctional borazines (shown below) leads to hexagonal boron

nitride in good yields, with good crystallinity via intermediates such as **5**:

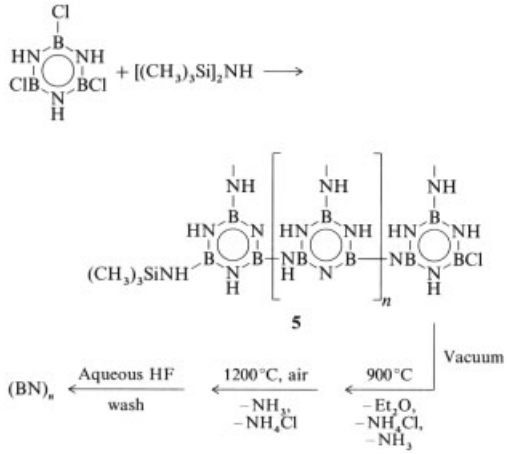

The formation of boron nitride nanocomposites has been reviewed [176].

6. Poly(Sulfur Nitride)

The topic of poly(sulfur nitride) has been reviewed [177].

Poly(sulfur nitride) [*56422-03-8*] was first prepared in 1910 by BURT, who passed vapors of cyclic tetrasulfur tetranitride [*28950-34-7*], S_4N_4, over heated silver gauze or silica wool [181]. However, it was not studied extensively until the 1970s, when purer products were obtained [182]. Poly(sulfur nitride) films have a lustrous golden color and are very striking in appearance. Polymeric sulfur nitride is a crystalline, fibrous material that is malleable under mild pressure. Electron micrographs reveal that the crystals are composed of fiber layers which are stacked parallel to one another along the crystal axis; X-ray studies show that the polymer consists of almost planar chains of alternating sulfur and nitrogen atoms. All the S–N bond lengths are similar and correspond to a sulfur – nitrogen bond order intermediate between a single and a double bond. The polymer decomposes on heating at ca. 140 °C (see below).

The polymer exhibits highly anisotropic electrical conductivity. Room-temperature conductivities as high as $3700\,\Omega^{-1}\,cm^{-1}$ along the crystal axis have been measured [184], [185]. Conductivity increases markedly upon cooling to 4.2 K, and the polymer becomes superconducting at 0.26 K [186]. Poly(sulfur nitride) was the first example of a polymeric superconductor. The conductivity increases on doping with bromine, and $(SNBr_{0.4})_n$ at room temperature has a conductivity of $3.8\times10^4\,\Omega^{-1}\,cm^{-1}$ parallel to the fibers, but only $8\,\Omega^{-1}\,cm^{-1}$ perpendicular to the fiber axis. This strong directionality is related to efficient conductivity along the molecular chain axis, whereas a lower degree of order in chain packing along the perpendicular axis results in lower conductivity.

The large-scale production of poly(sulfur nitride) does not appear to have attracted much attention because of its rather low thermal stability and because S_4N_4 is explosive.

Further, polymerization is slow (6 – 8 weeks). Efforts have aimed at developing alternate synthetic routes.

Poly(sulfur nitride) has been prepared from solution by addition of trimethylsilyl azide to $(NSCl)_3$ and other similar routes [187–189]. In these processes, a powdered product is obtained rather than the larger crystalline grains. Microcrystalline layers of poly(sulfur nitride) have been prepared on platinum electrodes by electrolysis of $[S_5N_5]^+$ [190], [191]. Additional synthetic routes have also been found [177]. Even so, the classical route appears to be the method of choice for the synthesis of pure poly(sulfur nitride).

The fine structure of poly(sulfur nitride) continues to be elucidated [177]. Currently, it is believed the polymer chains are planar and lie along the (crystallographic) *b* axis that develops along the *a* axis of S_2N_2. There are two chains in each unit cell with two SN units per chain. Disordered chains are present between the "ordered" chains. Secondary interaction between the chains is significant.

Polymer crystals are stable in air and water for several days. Stress and strain properties are like those of an anisotropic metal. The Young's modulus parallel to the axis of the chain is on the order of $10^{10}\,N\,m^{-2}$.

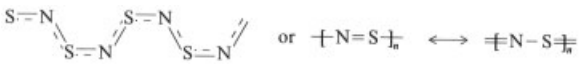

Poly(sulfur nitride) becomes superconductive at 0.26 K [192]. This property has been extensively studied [177]. The onset of superconductivity T_c increases with increasing pressure up to about 9000 bar, whereupon it decreases, but at higher pressures T_c again increases to about 3 K. It is possible that at higher pressures small amounts are converted to a –SNNS– polymer [193]. Poly(sulfur nitride) exhibits the Meissner – Ochsenfeld effect (→ Superconductors, Chap. 2.2.) below about 20 K. Below 4.2 K negative transverse magnetoresistance occurs in the presence of weak magnetic fields (< 3 T), though a normal positive magnetoresistance is found above 77 K or in stronger fields [177].

Thin films have been produced with properties that at times are similar to and under certain conditions appears to differ from those of the crystals themselves. The reason(s) for these differences are not currently fully understood. The ability to make large crystals has been reported [194], [195].

Recent interest has focused on electrical applications because of its unusual magnetic and electrical properties. Suggested uses include light-emitting diodes, solar cells, and transistors [196], [197]. A poly(sulfur nitride) – Ga – As solar cell with an efficiency of 6.2 % has been described [198]. Blue light-emitting diodes have been made on ZnS with either gold or poly(sulfur nitride) barrier electrodes [199]. The polymer gives barriers about 0.75 eV higher with a hundredfold increase in quantum efficiency in comparison to the gold-based electrode. A power output of 440 W h/kg at 200 mA was found for a lithium cell with a poly(sulfur nitride) electrode [192]. Patents for the use of poly(sulfur nitride) have been issued for a variety of uses including the cathode in batteries [200], [201], coatings on an image recording sheet [202], etching mask [203–205], explosive initiator [206], electroconductive resins [207], multilayered wiring patterns [208], sublimable polishing grains for semiconductor devices [209], and in light-emitting diodes [210].

Extensive work has been done on poly(sulfur nitride) electrodes [177]. A ruthenium(II)-modified poly(sulfur nitride) electrode generated hydrogen continuously in aqueous sulfuric acid solution at − 0.1 V under visible light [211]. Molybdenum(VI) electrodes using film and crystalline poly(sulfur nitride) can convert ethyne to ethene at rate 10^6 times that achieved by chemical systems [212]. Paste electrodes of poly(sulfur nitride) exhibit metallic electrochemical behavior in nonaqueous liquids [213].

7. Polysulfur

Because of its removal from industrial waste to meet environmental standards, sulfur has been stockpiled and is available in large quantities at a modest price. Although the stable form at room temperature is cyclooctasulfur, S_8, linear polysulfur is formed on heating. Unfortunately, the thermodynamically stable form of sulfur is the S_8 monomer and the polymer undergoes depolymerization after some time. Various methods have been used to retard depolymerization to the S_8 monomer. Removal of minute amounts of the monomer through judicious extraction with carbon disulfide retards depolymerization to some extent. Probably the most effective method involves the addition of olefins, such as limonene, cyclopentadiene, and myrcene [214]. Some of these result in polysulfur that is stable for more than five years. Such stabilized polysulfur has been incorporated into concrete and asphalt mixes to strengthen them. Concrete blocks, posts, highway pavement, and parking tire restrainers that contain stabilized polysulfur are produced.

8. Organometallic Polymers

A variety of organometallic polymers have been synthesized for a number of applications. Examples are shown in Figure 2. The metals may be an integral part of the polymer chain (**6**) or side chain (**7**); they can be connected through carbon (**8**) or other atoms (**9**); they may be bonded by coordination (**9**), covalent (**6–8**), or π – metallocene-like bonding (**10**); or they can be part of a linear (**6**, **8–10**) or cross-linked (**7**) product. A classical coordination polymer is described by **11**, and stacked metal-chain complexes are depicted by **12**, where the circles represent organic macrocycles such as phthalocyanine.

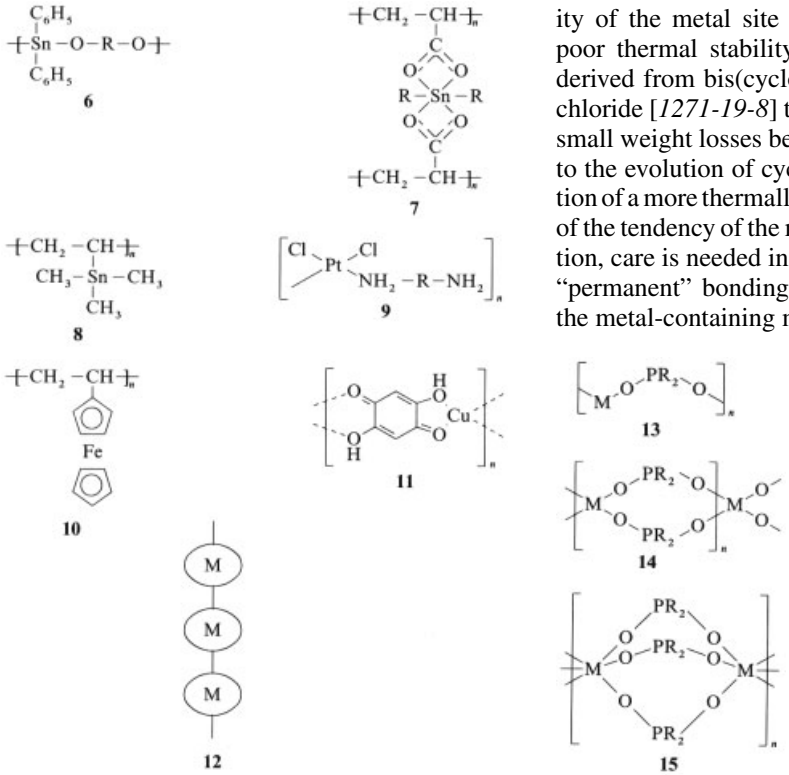

Figure 2. Some organometallic polymers

Polymeric metal phosphinates may have single-, double-, or triple-bridged structures, such as **13**, **14**, and **15**, respectively. Such structures are known for aluminum, beryllium, cobalt, chromium, nickel, titanium, and tin [229]. They form films, with thermal stability up to 450 °C claimed in some instances. The chromium(III) polyphosphinates have been used as thickening agents for silicone grease to improve its high-pressure physical properties.

The metal moiety of an organometallic polymer imparts particular properties. First, the metal atom is typically a site for reaction with Lewis bases and redox reactants. These sites have been exploited for catalysis and may be responsible for internal rearrangement and degradation. Second, the metal generally contributes to polymer stiffness through added volume occupation and π-bond formation. Thus, flexibility must be derived from the organic moiety. Third, the metal-containing portion typically contributes to the dipolar, ionic character of the products. The high chemical reactivity of the metal site results in polymers with poor thermal stability. For example, products derived from bis(cyclopentadienyl)titanium dichloride [*1271-19-8*] typically begin to undergo small weight losses between 60 and 150 °C due to the evolution of cyclopentadiene and formation of a more thermally stable material. Because of the tendency of the metal site to undergo reaction, care is needed in selecting solvents so that "permanent" bonding does not occur between the metal-containing moiety and the solvent.

The variety of structures and properties offered by organometallic polymers is enormous and cannot be summarized in a representative fashion in this short chapter. The reader is referred to reviews on organometallic polymers [23–27], [215–228]. Important uses of organometallic polymers are as follows:

1) Catalysis [226], [214], [231]
2) Biological uses, general [225], [232], [233–238], [340, 376–378, 385]
3) Additives: coatings, textiles, plastics, and paper [225], [230], [231], [239–242]
4) Electrical conductors, semiconductors, and piezoelectric devices [243–246], [247–251]
5) Electrocatalysis, photoelectrocatalysis, photovoltaic cells, specialized electrodes, sensor devices, electrochromism, electroluminescence, and multicolor displays [225], [226], [246], [252–258]
6) Analytical reagents [259–263]
7) Photoprotective agents [230], [231], [239–242]

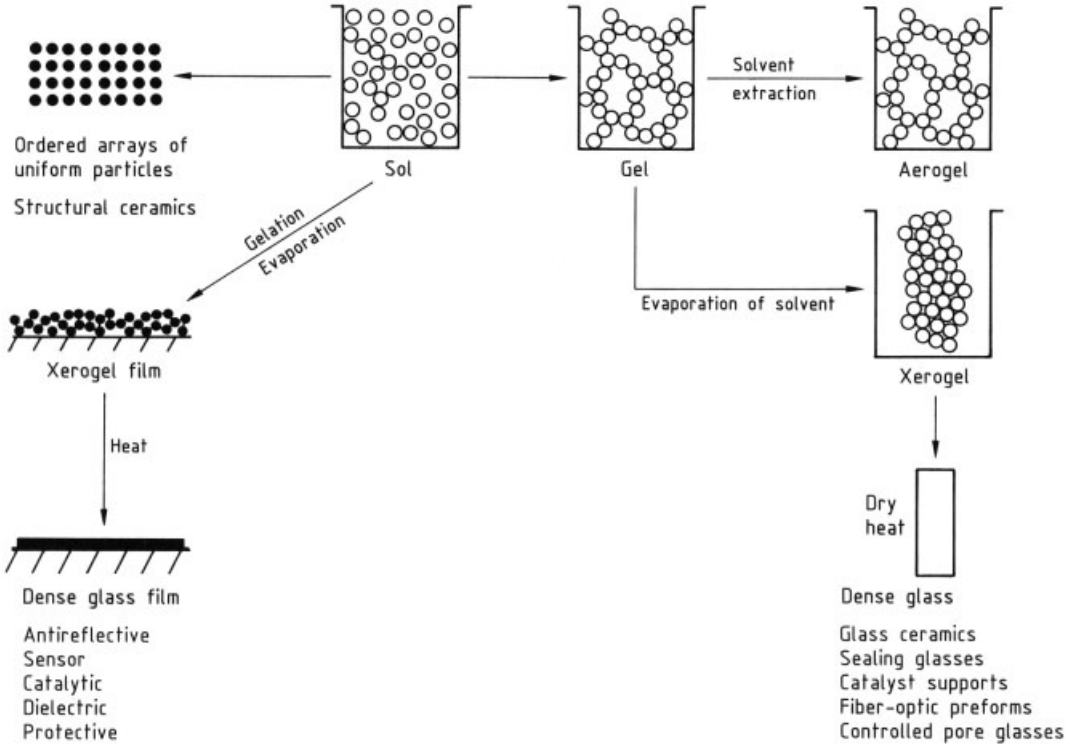

Figure 3. Schematic of the sol – gel process

8) Thickening agents [229]
9) Energy-transfer agents [265–267], [247], [248]
10) Permanent coloring agents [230], [239–242]
11) Composites [268], [269]
12) Metal deposition [270–274]
13) Solar energy conversion [275], [276]
14) Ceramic precursors [277–282]
15) Superconductivity [283]
16) Functionalized electrodes (electrocatalysis, photoelectrocatalysis, photovoltaic cells) [284–287]
17) Ferromagnetic materials [288]
18) Photoconductive materials [262], [263], [287]
19) Poly(vinyl alcohol) heat stabilizers [225]
20) Electrochemical sensors [222]
21) Nonlinear optics [222]
22) One-dimensional conductivity [222]
23) Nanostructures and nanotechnology [226].

9. Sol – Gel Inorganic Polymers

In the 1980s, several symposia were devoted to sol – gel inorganic polymer preparation, indicating the explosive growth in this field [289–291]. This field has continued explosive growth under additional "titles" such as organic/inorganic composite materials. Reviews are given in [292–294].

In sol – gel processes, ceramic polymer precursors are formed in solution at ambient temperature; shaped by casting, film formation, or fiber drawing; and then consolidated to furnish dense glasses or polycrystalline ceramics [295]. The most common sol – gel processes employ alkoxides of elements such as silicon, boron, titanium, and aluminum. In alcohol – water solution, the alkoxide groups are removed stepwise by hydrolysis under acidic or basic catalysis and replaced by hydroxyl groups, which then form −M−O−M− linkages. Thus, branched polymeric chains grow and interconnect, as illustrated below for a silicate sol. Gelation eventually occurs as the growing polymers link to-

gether to form a network that spans the entire solution volume. At this point (the gel point), both the viscosity and the elastic modulus increase rapidly.

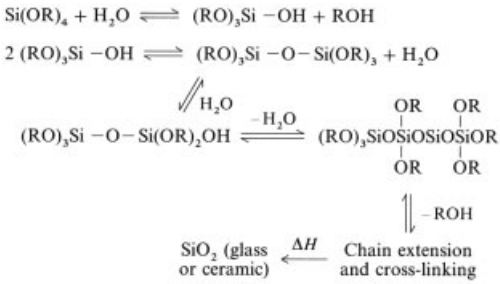

The gel can be viewed as a viscoelastic material composed of interpenetrating liquid and solid phases. The solid network retards the escape of the liquid and prevents structural collapse. Gelation is advantageous in processing because the gel freezes in shapes. Thus, the sudden increase in viscosity locks in place the shapes formed by casting, drawing of fibers, or film formation. Some sols and gels may be oriented or modified by drawing or shearing. The gel can then be dried by evaporation to form a *xerogel* or by supercritical fluid extraction to give an *aerogel*. Consolidation to dense glasses or ceramics is finally carried out by thermal treatment and sintering. Figure 3 gives some generalized sol–gel processes for ceramic or glass preparation.

Both aerogels and xerogels have high surface area ($> 500\,m^2/g$) and small pore diameter ($< 20\,nm$). They have been used as ultrafiltration media, antireflective coatings, and catalyst supports. Final densification is accomplished by viscous sintering in which the viscosity exceeds $10^{12}\,Pa \cdot s$.

The rate of silicate sol and gel formation is exceptionally sensitive to pH and to the water–alcohol molar ratio in the reaction medium, as is the solubility of the amorphous silica that is formed. Silica networks are based on $[SiO_4]^{4-}$ tetrahedra modified by $[O_3Si{-}O^-M^+]$ units. Addition of B_2O_3, Al_2O_3, TiO_2, ZrO_2, and related network-forming units can result in the formation of mixed glasses and ceramics. The sol–gel approach uses the corresponding metal alkoxides, as illustrated below, for the formation of borosilicate glasses:

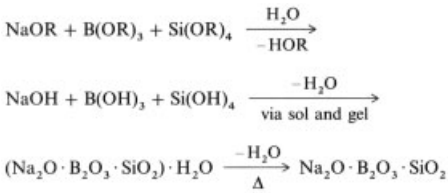

The chemistry of the sol–gel process can thus be tailored to design new polymeric glasses or ceramics. For example, two or more metal alkoxides can be mixed in varying ratios to form mixed sol–gel polymers. Mixing $Zr(OEt)_4$ with $Si(OEt)_4$ could lead to a gel with structure **19**. Similarly, a zirconia–alumina–silicate gel polymer of structure **20** could be obtained from $Zr(OR)_4$, $Al(OR)_3$, and $Si(OR)_4$.

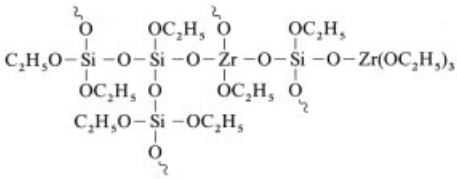

19

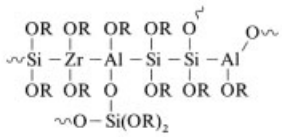

20

The development of organically modified silicates (ceramers) has resulted in a variety of new materials [296]. Thus, vinyl groups, epoxides, and methacrylate functions have been polymerized to give interpenetrating networks with a wide variety of structures and properties [297–299]:

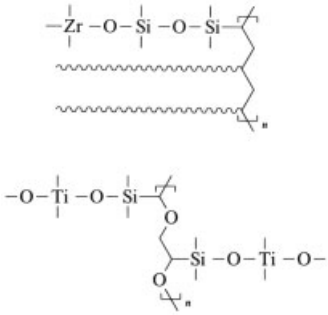

Applications of ceramers include adhesives for glass surfaces [297], protective coatings for medieval stained glass [300], and scratch-resistant coatings for plastic eyeglass lenses.

Sol – gel preparations of tetraethoxysilane – water – alcohol – HCl can be spun into fibers once the appropriate viscosity has been reached. These fibers are only slightly weaker than silica-glass fibers; ZrO_2 – SiO_2 fibers have also been made in this way. Sol – gel processing of Ti(*i*-$C_3H_7O)_4$ – H_2O – ethanol – HCl gives thin film coatings of controlled thickness which exhibit interference colors that vary with film thickness and they are n-type semiconductors [301].

Hybrid materials have been made by incorporating end-capped poly(tetramethylene oxide) (PTMO) blocks of molecular mass 650 – 2000 into tetraethoxysilane sol – gel glasses [302].

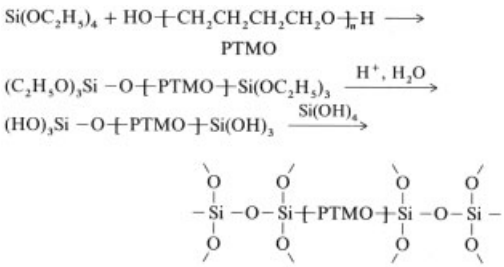

These materials exhibit high extensibility. They contain interdispersed organic polymer and inorganic polymer regions. Although the PTMO blocks are well dispersed, some degree of local phase separation occurs. Titanium isopropoxide has been used to incorporate TiO_2 into these block systems [302].

Thus, although sol – gel chemistry has been extensively investigated since the early 1960s [304], [305], only now is it rapidly becoming important in material synthesis.

Uses. Sol – gels offer properties between those of organic polymers and inorganic materials. They are particularly useful in areas where they can replace materials to give better properties, better processability, and/or lower cost. Thin-film applications are promising, and the production of numerous nonlinear optical materials has been reported [292]. Thin-film protective coatings are being considered in a number of areas. Hot-melt adhesives for glass containers have been developed [293]. Hard contact lens materials have also been developed [293]. Another area of practical application is the reinforcement of plastics and elastomers [293]. Recent production of sol – gels with nanoscale pores allows their potential use as catalysts, porous supports, and as selective absorbents [293].

10. Inorganic Fibers and Whiskers

Inorganic fibers have been important materials for over 50 years. Asbestos, glass, and carbon fibers are particularly well known. Many of the materials described herein are discussed in more detail under → Fibers, 5. Synthetic Inorganic.

Many inorganic fibers exhibit very high tensile strength and thermal stability (> 1000 °C) but suffer from high density and, high cost. This chapter discusses the preparation and properties of some inorganic fibers which are of interest in the production of high-performance materials, particularly advanced resin-matrix, ceramic-matrix, and metallic-matrix composites. The patent literature on inorganic fibers and composites is extensive and growing rapidly [302].

Asbestos, once widely used as insulation, now has more limited applicability because of its health hazard. Small fibers lodge in the lungs and can cause lung cancer and other lung damage. Asbestos-containing shingles, paper, welding rods, floor tiles, and brake linings are, however, still manufactured. These products avoid fibers of length 5 – 20 µm, which are the most dangerous. *Glass fibers* are used widely as insulation and in fiber-reinforced plastics. *Carbon fibers* are of increasing importance in high-performance composites. Several metallocene-type condensation polymers (e.g., **21**) exhibit fiber formation directly upon polymerization or upon application of stress [306].

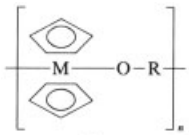

21

Inorganic fibers composed of alumina, silica, titania, zirconia, aluminum silicate, boron, boron carbide, boron nitride, silicon boride, silicon carbide, silicon nitride and others are known.

Polymeric *alumina, titania,* and *zirconia fibers* are produced industrially [307–310]. Zirconia textiles are produced by Union Carbide under the tradename Zircar.

Alumina fibers with a high Al_2O_3 content have high temperature stabilities and high moduli. They have been made by dispersing aluminum salts on rayon fiber, heating to burn off the organic material, and sintering. Alternatively, slurries of hydrated alumina can be extruded, dried, and heated to give polycrystalline filaments [311], [312]. Fibers may also be extruded and drawn from a viscous organic solution, followed by heating to high temperature [313]. Polyaluminoxanes are spun into fibers and sintered to give Al_2O_3 fibers, or mixed with silicon compounds, spun, and calcined at 1000 – 1200 °C to give aluminosilicate fibers.

High-purity *silicon dioxide fibers* can be used up to 1100 °C and are of interest for advanced composites [314]. They are made by melting a glass composed of 75 % SiO_2 and 25 % Na_2O, and then spinning and stretching the fiber simultaneously at 1100 °C. The fibers are then acid-leached to remove Na_2O and dried above 300 °C [315]. In another process, silicon alkoxide polymers are spun by a sol – gel process to make high-purity SiO_2 fibers. The fibers are heated to 1000 °C to give a 99.999 % pure quartz fiber which is stable at 1000 °C [316].

A variety of *metal oxide fibers* are now available. Du Pont has developed an alumina – zirconia fiber with a tensile strength > 2100 MPa and a modulus of 380 GPa for use in high-performance composites [317]. Even after exposure to 1400 °C for 100 h, its tensile strength was 1400 MPa and its modulus remained unchanged. Stabilized zirconia and alumina – silica – boria fibers have also been produced [315]. The patent literature reports fibers of such materials as alumina – chromia, thoria (thorium oxide), titania, titania – silica, thoria – silicate, zirconia – alumina, zirconium silicate, and zirconia – yttria, as well as whiskers of these and many other metal oxides [318].

Boron – tungsten fibers have densities of 2.4 – 2.6 g/cm^3 and elastic moduli as high as 390 GPa. However, they lose both tensile strength and modulus at high temperature. When coated with silicon carbide and boron carbide, they retain about half their initial strength in air up to 800 – 1000 °C. *Boron fibers* are much less dense and are, therefore, useful in the construction of lightweight composites [318] found in tennis racket frames and aircraft. However, their cost is high because of production complexities. Processes involve chemical vapor deposition of boron onto carbon fiber and other cores, followed by a variety of treatments [328]. *Boron carbide fibers* are made from mixtures of methane and boron trichloride by using CVD to create a boron carbide layer on tungsten core filaments [302]. *Boron nitride fibers* are used with boron nitride matrices to give composites that are employed as electrical insulators which are also good thermal conductors. Boron nitride fiber mats are useful as electrtic cell separators in lithium sulfide batteries [302]. The fibers are made from boric oxide fibers that are nitrided by ammonia at ca. 800 °C followed by stretching at 2000 °C. They have a tensile strength of 2100 MPa and a modulus of 345 GPa.

Silicon carbide can withstand 1800 °C under oxidizing conditions. Two methods of production are employed. In the first, CVD is used to deposit silicon carbide on tungsten filaments. In the second method, polysilanes are melt-spun to give a precursor fiber, which on heat treatment yields fibers 10 – 30 µm in diameter (see Chap. 2), [109], [319], [320]. Silicon carbide fibers are more resistant to oxidation at high temperature than carbon fibers and have greater compressive strength and electrical resistance. *Silicon carbide – carbon* fibers, can be spun from mixtures of pitch and a poly(dimethylsilane), followed by vacuum heating at 1400 °C [321]. In a similar fashion, melt-spinning of a polycarbosilazane followed by high-temperature treatment gives *silicon carbide – nitride* fibers with a tensile strength of 1.2 MPa and a modulus of 187 MPa, which resist oxidation up to 1200 °C [332].

Silicon nitride fibers are used in advanced composites for aircraft, radomes, and electrical components. They are prepared either by reduction of silica in the presence of nitrogen in an electric furnace [315] or by nitriding polycarbosilane precursor fibers with ammonia [323].

Sol – gel *polytitanocarbosilanes* have been prepared and used to melt-spin a precursor fiber which, on pyrolysis in nitrogen at 800 – 1500 °C, gives Si – C – Ti – O fibers. At 1300 °C, a tensile strength of 3.0 GPa and a modulus of 220 GPa were achieved [324]. Polycarbosilanes can be

treated with titanium tetrabutoxide to give a polymer from which fibers may be spun. After heat treatment, Si – C – Ti – C fibers with a Young's modulus of 75 – 150 GPa and tensile strength of 0.1 – 11 GPa were obtained [335].

More recently, a number of inorganic materials have also been produced as whiskers (see also → Whiskers). In general, the tensile strengths of whiskers are greater than those of fibers. For instance the approximate tensile strength of graphite in the bulk form is 1000 MPa, for fibers it is 2800 MPa, and for whiskers it is 15 000 MPa. Carbon whiskers are produced by depositing a layer of pyrocarbon from the vapor phase on a catalytically grown carbon filament [325].

Inorganic fibers and whiskers exhibit some of the highest tensile strengths recorded. They are useful when light weight and high strength are needed. Uses include turbine blades, heat-resistant re-entry vessels, golf club shafts, automotive radio aerials, fishing poles, dental fillings, and parts of aircraft, including stealth aircraft [326]. Many of these inorganic fibers and whiskers are used as reinforcing agents in composites [326].

11. Biomedical and Antimicrobial Polymers

Silicones. Silicones (polysiloxanes) are polymers having a silicon – oxygen backbone. Among the numerous materials having biomedical applications, silicone polymers play an important and versatile role. Medical devices made from polysiloxanes include artificial wrist, toe, and finger joints; hip implants; oviductal plugs [350], [351]; brain membranes; mammary implants; tracheotomy vents; and artificial hearts. For a detailed description of polysiloxanes, see → Silicones [328–331]. This chapter deals with their biomedical applications [332]. Polysiloxanes are widely applied in medicine for numerous reasons including their generally extremely low toxicity, high gas permeability, good mechanical properties, and the wide range of fabrication processes that can be used (e.g., coating, encapsulation, casting, molding, sealing, and extrusion). Furthermore, they are generally inert to the body and to the immune system. The biomedical applications of polysiloxanes are reviewed in [232]. Monomeric and oligomeric siloxanes are generally not sold as medical-grade materials and may contain toxic trace contaminants that are not important in industrial applications but may be critical in biomedical applications. Proper procedures must be adopted to ensure quality control and absence of contaminants that may be toxic and may adversely affect long-term, repeated use of the end product.

Polysiloxanes are extensively employed in biomedical devices as specialty surface treatments [332–334]. For example, octadecylpolysiloxane (Siliclad, Glassclad 18) is applied to hydrophilic surfaces to reduce blood – surface interactions and protein adsorption [334], [335]. Other polysilane-derived surface treatments are sold under tradenames such as Glassclad HP, D.C. 5700, Glassclad 6C, Dri-Film, and Glassclad IM [232, 340]. Silicone membranes play critical roles in oxygenation and dialysis. Polysiloxanes are about ten times more permeable to oxygen than low-density polyethylene and a hundred times more permeable than nylon or butyl rubber [336]. Therefore, poly(dimethylsiloxane) [*1646-73-7*] soft lenses and poly(dimethylsiloxane-*co*-methacrylate) hard lenses constitute a major portion of the contact lens market. Silicone-based interpenetrating polymer networks give clean, resilient gels with cohesive properties, which are used in breast implants [337–339, 232]. Bilayer membranes employing a top layer of cured silicone polymer are used as artificial skin [348], [349].

Low-consistency, room-temperature vulcanizing silicones cure to give low modulus mechanical plugs when injected into the fallopian tubes [350], [351]. This blocks the ovum from reaching the uterus and serves to prevent conception. A retrieval ring at the uteral end makes the procedure reversible. Silicone polymers are used to make dental impression materials [342–345]. Vinyl addition cured systems are used in biomedical encapsulation and mold making [346] and in maxillofacial prostheses. High-consistency elastomers are used in compression molding of shunts, flexing joints, and catheter tubing [347]. Silicone – urethanes (e.g., cardiothane 51) are used in blood pumps, intraaortic balloons, and artificial hearts because of their good fatigue strength, flexibility, tough-

ness, and retarded interaction with plasma proteins [348].

Silicone – polycarbonate block copolymers can be injection-molded, extruded, and cast into films. Their synthesis is described in [349–351]. They have high tear strength as well as high oxygen and water permeability. Applications include blood oxygenation, dialysis, and microelectrode materials [352–354].

Metal-Containing Polymers. A few metalcontaining polymers are used as therapeutic agents in the treatment of cancer [355–359], juvenile diabetes [360–362], Cooley's anemia [363], [364], and in the control of bacteria [359], [365–367] and viruses [359–362]. The polymer derived from the reaction of tetrachloroplatinate and methotrexate can inhibit the virus that causes juvenile diabetes-like symptoms [360], [361].

Tin-containing compounds are inhibitory to certain bacteria. A variety of methyl-, ethyl-, and propyl-substituted organotin polymers effectively inhibit *Pseudomonas aeruginosa* strains which are responsible for about one-half of the deaths of burn patients. Some suggested uses are as additives to paints, insulation, bandages, and clothing [369]. Degradation resistance has been obtained by in situ polymerization of tin-containing esters in wood [366], [367]. Impregnated poles and planks show total stability when exposed to seaside environment or seawater for more than five years [374], [375]; untreated wood is destroyed within a year. Similar results have been obtained for tin-containing cross-linked coatings in the control of ship hull fouling. Nevertheless, plants and sea creatures survive within centimeters of test objects, suggesting that the use of tin-containing polymers might be environmentally acceptable. Tin-containing polymers have been examined in the United States and Japan as antifouling coatings [370–375] and mildew-resistant paints [376]. Commercial products are available; however, environmental concerns in harbor areas remain.

12. Boron Polymers

The topic of boron polymers has been reviewed [377].

A variety of boron-containing polymers have been made by heating trialkyl phosphates with boron trichloride at 300 °C [378], [379]. At 900 °C this reaction produces boron phosphate glass. The boron polymers form networks prior to glass formation.

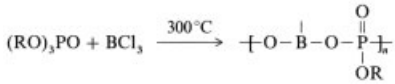

Reaction of benzene boronic acid with diphenylsilane gives transparent, film-forming, polymeric borosiloxanes which can be drawn into fibers [380]. Similar materials resulting from boronic esters and chlorosilanes form electrically insulating films when baked on copper plates [381]. A variety of polymers containing Si–O–B units have been reported [382]:

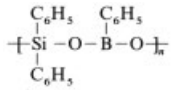

Boric acid forms condensation polymers with polyhydroxy compounds. With ethylene glycol and phenol, a tacky polymer is formed that is hydrolyzed readily by water [383]. Many such compositions have been studied as adhesives, binders, coatings, and resins.

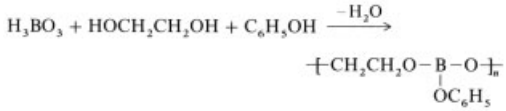

Polymers with all-boron backbones [384] or alternating boron – phosphorous backbones [385] have been reported. Pyrolysis of dimethylphosphineborane in the presence of a base gives linear high molecular mass polymers [386]. In many cases, ring systems form instead of linear polymers, with cyclic trimers and tetramers predominating. Use of triethylamine as a chain end blocking group favors the formation of linear polymers.

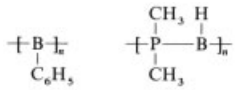

The action of water on diboron tetrachloride

gives subboric acid, $B_2(OH)_4$, which when heated at 220 °C loses water to yield a white boron monoxide polymer [387]. This remains unchanged up to 500 °C, but at 650 °C is transformed into brown boron monoxide which has lower solubility in water.

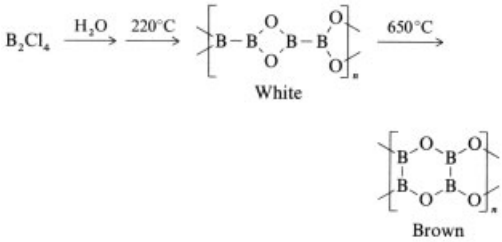

Low molecular mass polymers result from the reaction of 1,1′-dilithioferrocene with phenylboron dichloride, followed by treatment with water [388]:

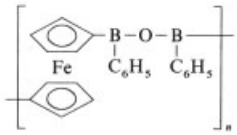

A large number of polyborazines have been prepared, and many are reviewed in [389]. The condensation of bis(alkylthio)borazines with diphenylsilicondiol gives borazine – siloxane polymers [390]. Cross-linked materials result from the corresponding reactions with trialkylthioborazines:

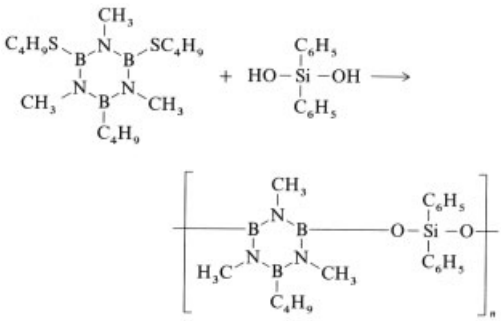

Reaction of β-chloroborazines with sodium sulfide, sodium disulfide (Na_2S_2), or sodium tetrasulfide (Na_2S_4) leads to polyborazines with sulfur linkages between the rings and molecular masses up to 53 000 [391]. Other polyborazines with oxygen [392], nitrogen [393], and phosphite bridges [394] between rings have also been prepared.

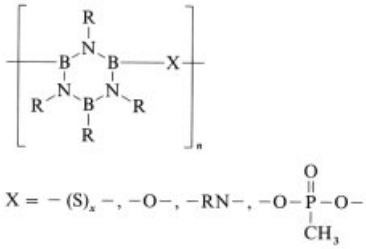

Carborane polymers with a variety of structures have been made from the $C_2B_{10}H_{10}$ icosahedral cage systems, in which the carbon atoms are located *ortho*, *meta*, or *para* to one another. The best known examples are the Dexsil polymers of the Olin Corporation [395], [396]. This series has siloxane groups attached to the carbon atoms of *ortho*-carborane (1,2-$C_2B_{10}H_{12}$) and varying numbers of dimethylsiloxane groups between adjacent cages:

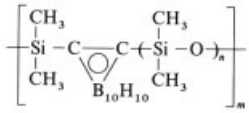

$n = 2$, Dexsil 200
$n = 3$, Dexsil 300
$n = 4$, Dexsil 500

Many other groups have been used to connect carborane cages. Treatment of dilithiocarborane with phosphorus trichloride and sodium azide leads to the very unusual structure [397]:

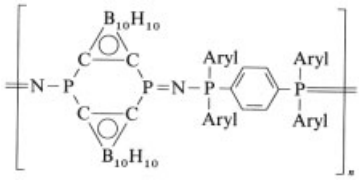

Dilithiocarboranes are reacted with dialkyl metal dihalides to introduce metal functions between the cages. For example, reaction of *ortho*- or *para*-dilithiocarborane with dimethyldichlorostannane introduces dimethyltin bridges between the cages. Dimethylgermanium bridges are introduced in a similar manner [398], [399].

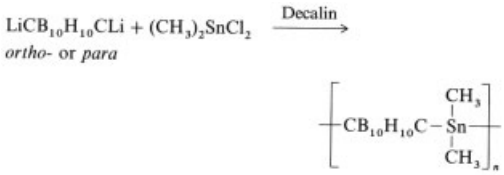

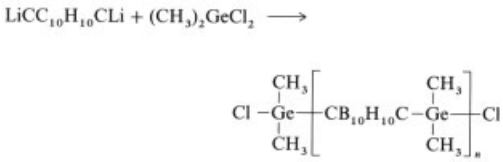

The dimethyltin-bridged *para*-carborane polymer softens only above 420 °C and decomposes at 425 °C, whereas its *ortho* analogue melts at 250 – 255 °C. Random mixtures of *meta*-carborane (1,7-$C_2B_{10}H_{12}$) and *para*-carborane (1,12-$C_2B_{10}H_{12}$) have been used to prepare dialkyltin-bridged polymers.

A polymer bridged by $-Ge(CH_3)_2NH-$ units was prepared from dilithiocarborane by using excess dimethyldichlorogermane and then ammonia:

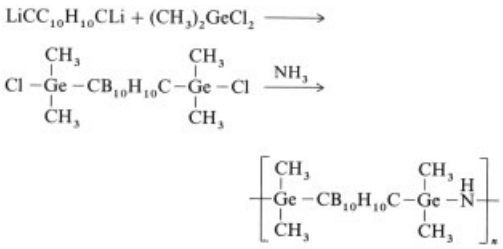

Borazine bridges have also been introduced [400]:

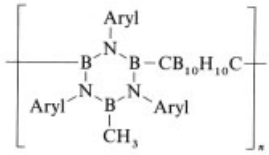

Refluxing the bis(sulfinylchloride) of *meta*-carborane in water gives a poly(carboranethiosulfinate), whereas treatment with dilithiocarborane gives the corresponding poly(carboranesulfide) (*mp* 219 – 222 °C) [401]:

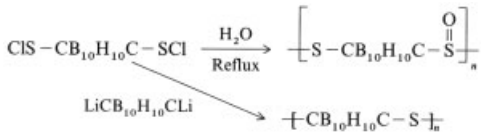

13. Aluminum Polymers

Polymers containing only aluminum atoms in the backbone are unknown. Aluminum – oxygen and aluminum – nitrogen backbones are known, but they are hydrolyzed readily by aqueous acid and alkali. Polymers with an aluminum – oxygen backbone are usually obtained by reaction of aluminum alkoxides with water, diols, organic acids, or amides. They occur as gums to brittle solids; applications include use as cross-linking agents, drying agents, gelling agents, fuel additives, water repellents, paints, varnishes, linoleum, and inks [402–406]. An idealized linear structure is

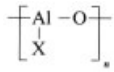

where X can be RO– [407], RCOO– [408], R_3SiO- [407], [409], or R(RO)P(O)O– [410]. These polymers are highly cross-linked, and the aluminum is actually four-coordinate; cyclic structures often predominate.

To increase the hydrolytic stability of Al–O bonds and permit the synthesis of linear polymers, four-coordinate aluminum ethyl acetoacetate complexes have been made as chelate rings along the chain [411]. The use of diols gives hydrolytically unstable polymers [412].

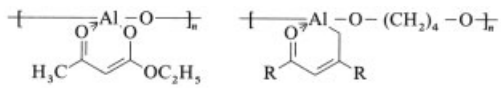

A variety of coordination polymers involving aluminum chelates have been made, but these are considered true organometallic polymers and are not discussed here. ANDRIANOV pioneered the use of R_3SiO- blocking groups to prevent formation of three-dimensional networks. Although oligomers such as

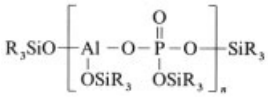

have been prepared [409], useful higher polymers remain elusive.

Poly(aluminosiloxanes) are polymers containing an Si–O–Al–O backbone. A typical example results from the reaction of the sodium salt of a poly(dimethylsiloxane) with aluminum chloride.

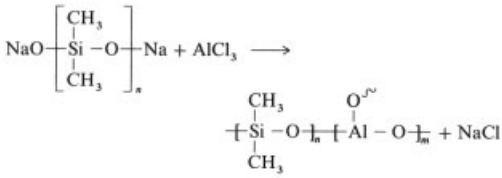

Polymers with silicon – aluminum ratios of 0.8 – 23 have been made [413], [414]. Polymers with low silicon – aluminum ratios are brittle and insoluble, indicating a three-dimensional structure; those with silicon – aluminum ratios of 7 – 23 are soluble and stable in both polar and nonpolar solvents even at 150 °C.

The poly(aluminophenylsiloxane) having a silicon – aluminum ratio of 4 is completely infusible but very soluble in organic solvents; it can be plasticized. A ladder structure has been assigned:

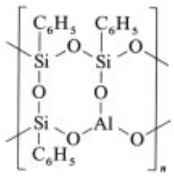

Regular poly(aluminosiloxane) structures are prepared by polycondensation of dialkyldiacetoxysilanes with aluminum alkoxides.

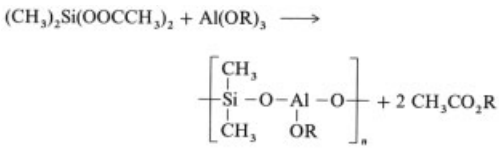

14. Polymers with –Si–O–M–O– Backbones

The early driving force in preparing poly-(metallosiloxanes) was to enhance thermal stability. The group 13 elements boron and aluminum have been incorporated into siloxane chains (see Chaps. 12 and 13). The metal can also be from group 14 (Ge, Sn, Pb), 15 (P, As, Sb), or 16 (S), or a transition metal (Ti, V, Cr, Fe, Ni, Zr, Hf). The thermal and hydrolytic stabilities of these polymers are usually inferior to those of the parent poly(organosiloxanes), but a variety of property modifications have been achieved. The hydrolytic stability (to HCl) of poly(titanoxane –phenylsiloxane) was shown to be greater than that of its aluminum analogue poly(aluminoxane – phenylsiloxane), but less than that of poly-(diphenylsiloxane). Thus, the order of stability with respect to cleavage by hydrochloric acid at 90 °C is
– Si – O – Si – O – > – Si – O – Ti – O – > – Si – O – Al – O –.

Random –Si–O–Ge–O– and –Si–O–Sn–O– polymers can be made by the cohydrolysis of dialkyldichlorosilanes with dialkyldihalogermanes or dialkyldihalostannanes [415]. Low molecular mass components are treated further with sulfuric acid to give rubber-like materials.

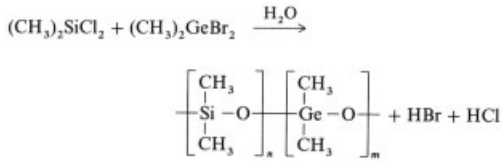

The random incorporation of tin is achieved by cohydrolysis of dihalostannanes with dihalosilanes [416] or by reaction of dialkyltin oxides with a silanediol [417].

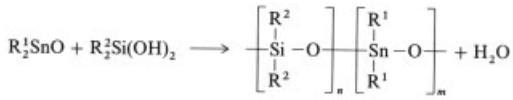

The most stable polymers in this series have high silicon: tin ratios. Regular alternating structures

are achieved by reaction of dialkyldiacetoxystannanes with dialkyldialkoxysilanes [418]:

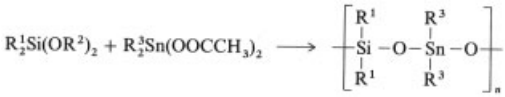

Poly(titanosiloxanes) can be prepared by cohydrolysis of dichlorosilanes with titanium alkoxides to give random incorporation of $-Si(R)_2O-$ and $-Ti(R)_2O-$ units. Alternatively, diphenylsilanediol can react with titanium alkoxides [419]. The resulting polymers are hard coatings, which are opaque to UV light and suitable for use as refrigeration enamels [419].

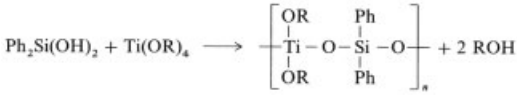

Partial hydrolysis of alkylalkoxysilanes with titanium alkoxides, acylates, or alkylamides gives copolymers that can be used to help catalyze the drying of printing inks [420], [421]. Similar silicon – zirconium ($-Si-O-Zr-O-$) preparations have also been reported [422]. A different type of titanium unit is represented by the titanocene fragment. Condensation of dimethylsilanediol [423] or its disodium salt [424] with titanocene dichloride results in the incorporation of titanocene units into the polymers.

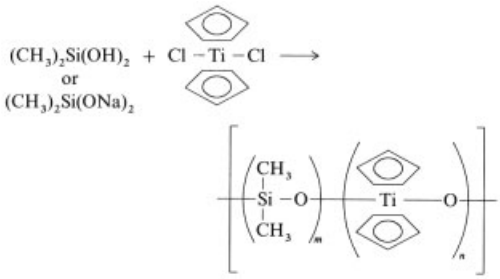

A related polymer containing $-O-Si-S-Ti-S-Si-$ units in the backbone is made by polycondensation of dimercaptotitanocene [425]:

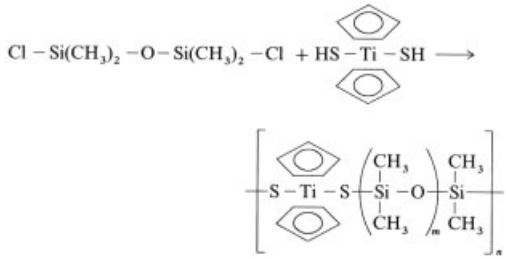

Reactive silicones and titanium alkoxides give polymers with $-Si-O-Ti-O-$ backbones when heated in petroleum solvents. These are said to be suitable for making waterproof leather and textiles [426], [427].

Polymers with titanium – oxygen backbones, related to siloxanes, are prepared by heating trialkoxytitanium acylates and dialkoxytitanium diacylates at 75 °C. These materials have low molecular mass, and whether they are linear polymers or cyclic oligomers is as yet unresolved. Titanium alkoxides polymerize on treatment with acetic acid, but problems with cross-linking and three-dimensional network formation complicate this method.

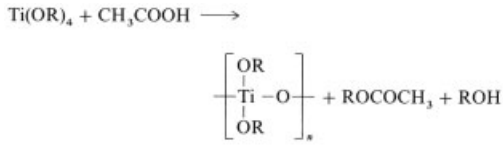

Incorporation of a group 15 element is illustrated by the preparation of $-Si-O-As-O-$ chain polymers. Condensation of dimethyldichlorosilane with $CH_3AsO(OH)_2$ gives a transparent, rubbery polymer that is used as an impregnation rubber additive [428]. The arsenic content inhibits the growth of molds.

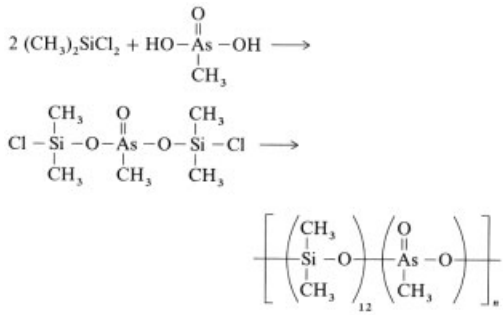

15. Poly(Carbon Disulfide), Poly(Carbon Diselenide), and Polythiocyanogen

The delocalized chalcogen-based polymers are of special interest due to the observation of superconductivity in $[SN]_n$ [429] and $[Se]_n$ [430]. BRIDGMAN first showed that carbon disulfide can be polymerized under high pressure (> 4.5 GPa) at 175 °C to give a black semiconducting solid [431]. The $(CS_2)_n$ prepared under high pressure was later assigned a head-to-tail linear structure [432]. The head-to-head structure is 16.3 kJ/mol lower in energy and, by analogy to $(CSe_2)_n$, might be the more likely structure [433].

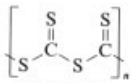

Head-to-tail

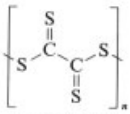

Head-to-head

Reaction products vary with experimental conditions. CHAN and JONSHER isolated a low-conductivity material ($< 10^{-13}$ S/cm) and a semiconducting material (10^{-3} S/cm) which consisted of free sulfur in a (CS_{2-x}) matrix [434]. They also proposed a cross-linked structure.

Anionic polymerization of carbon disulfide, initiated by sodium dispersed in dipolar aprotic solvents, results in polymers with less than two sulfur atoms per carbon [433], [435]. Poly(carbon disulfide) has also been prepared by photolysis [436], irradiation [437], and plasma deposition [438] to give structures described as $-[C_3S_2-]_n$, $-[CS(S)_m-]_n$, or mixtures [439].

Unlike carbon disulfide, carbon diselenide can be polymerized both at high pressure and under ambient conditions. The products obtained range from insulators to metals and superconductors. Carbon diselenide polymerizes at a rate of about 1 wt % per month at ambient temperature and pressure [440], [441]. When dissolved in solvents such as methylene chloride and dioxane, it polymerizes readily under high pressure (ca. 0.5 GPa) at 100 °C [442]. This polymer is an amorphous insulator. When the polymer is heated to 130 – 160 °C, a black semiconducting (10^{-3} S/cm) material results. After high-pressure polymerization and annealing at 130 °C, a crystalline product is obtained with a room-temperature conductivity of 50 S/cm; it has been assigned a two-dimensional sheet structure [443–445]. The polymer is thought to have a head-to-head structure [440], [441], which is significantly more stable than the head-to-tail structure [433]:

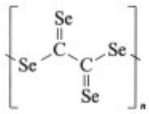

It is typically contaminated with trigonal selenium domains, depending on annealing and polymerization conditions. The observed crystallinity is due to both small and large trigonal selenium crystallites within the polymer matrix, and this selenium is the reason for the high conductivity [433]. As selenium crystallites form, the polymer becomes highly disordered, with the formation of −C−C−, −C=Se−, and −C−Se−Se−C− units.

Carbon selenosulfide (CSeS) does not polymerize in solution under conditions similar to those used to polymerize carbon diselenide [433]. Polymerization of carbon diselenide in the presence of excess carbon disulfide gives only the selenium-containing product. The slow polymerization of carbon diselenide at atmospheric pressure and 25 °C gives a black product with an electrical conductivity $< 10^{-6}$ S/cm, consistent with a band gap of ca. 2 eV [433].

Polythiocyanogen. Polythiocyanogen or parathiocyanogen is formed by the spontaneous polymerization of thiocyanogen $(SCN)_2$ [446]. While a number of structures are possible, a linear structure with the SCN atoms connected with a bond order of 1.5 is currently preferred [447]. The polymer is a semiconductor that can be doped with halogens [447].

16. Phosphorus-Containing Polymers

Phosphorus-containing polymers are among the three most important basic blocks of life. They serve as the building blocks of the critical nucleic acids, DNA, RNA, phosphosaccharides, phosphoproteins, and related biomaterials. Here, we will focus only on synthetic phosphorus-containing polymers. Reviews are given in [450–455]. The reviews cover the synthesis and uses of organic and inorganic phosphorus polymers having a wide variety of compositions and structures. Polymer chains containing phosphorus can be made by a variety of approaches, including those summarized below:

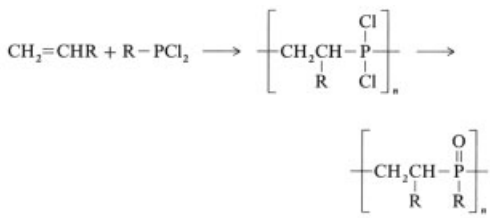

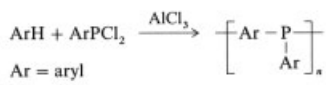

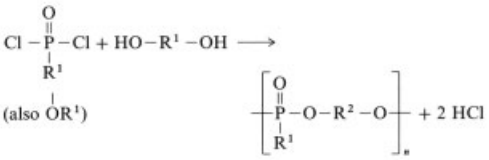

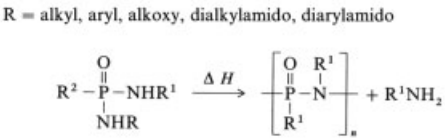

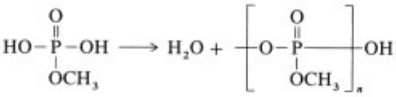

Polymers with a glasslike structure are formed by cocondensation of metal alkoxides or metal acetates with phosphorus oxychloride or phosphoric acid:

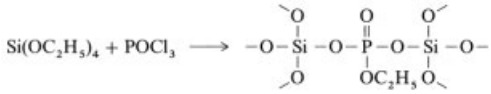

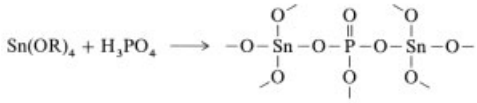

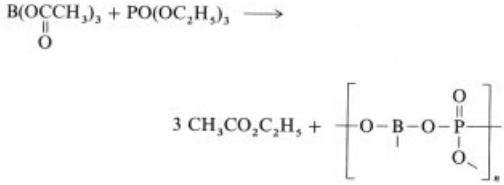

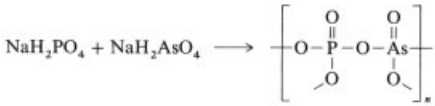

Such reactions can be used to prepare artificial glass, paint and lubricant additives, plasticizers, and adhesives.

An interesting class of linear phosphorous – boron backbone polymers has been reported [449]. When the dimethylphosphine – borane complex is pyrolyzed, a linear polymer with $\overline{DP} = 80$ is obtained which is soluble in hot benzene:

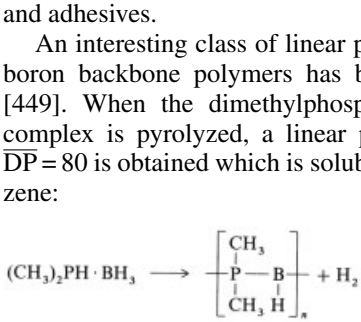

A similar process yields borophane polymers from tris(trialkylsilyl)phosphines [449]:

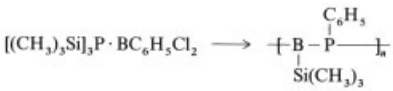

An interesting application of electrophilic substitution involves the incorporation of phosphorus bridges between ferrocene units [450], [456]. However, this route only gives low molecular mass materials.

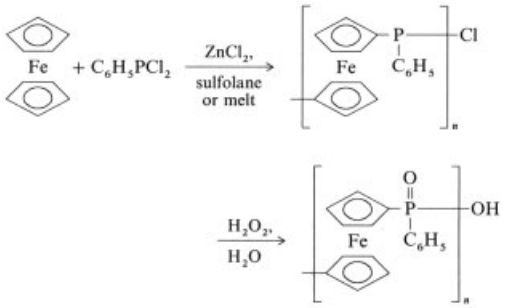

Polyphosphazenes are discussed in Chapter 4. However, a few carboranylphosphazene polymers have been reported [457]. Ring-opening polymerization of the carboranyl-substituted phosphazene ring proceeds without complications to yield a high molecular mass polymer:

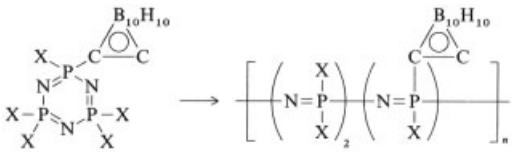

17. References

1. NRC: *Polymer Science and Engineering,* National Academy Press, Washington, D.C. 1981.
2. F. S. Kipping, *J. Chem. Soc.* **125** (1924) 229.
3. R. West et al., *J. Amer. Chem. Soc.* **103** (1981) 7352.
4. J. Wesson, T. Williams, *J. Polymer Sci., Polym. Chem. Ed.* **17** (1979) 2833.
5. R. Trujillo, *J. Organometallic Chem.* **198** (1980) C27.
6. S. Sawan, S. Ekhorutomwen, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6722.
7. S. Hayase, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6734.
8. K. Matyjaszewski, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6741.
9. A. Soum et al., in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6747.
10. C. Tien, A. Savoca, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6758.
11. E. Hengge, *J. Inorg. Organomet. Polym.* **3** (1993) no. 4, 287.
12. S. Thames, K. Panjnani, *J. Inorg. Organometal. Polym.* **6** (1996) no.2, 69.
13. A. Eckhardt, W. Schnabel, *J. Inorg. Organometal. Polym.* **6** (1996) no.2, 95.
14. K. Matyjaszewski, *J. Inorg. Organometal. Polym.* **1** (1991) no. 4, 463.
15. C. Carraher, *J. Macromol. Sci. Chem.* **A17** (1982) no. 8, 1293.
16. F. G. A. Stone, W. A. G. Graham: *Inorganic Polymers,* Academic Press, New York 1962.
17. K. A. Andrianov: *Metalorganic Polymers,* Interscience, New York 1962.
18. S. N. Borisov, M. G. Voronkov, E. Lukevits: *Organosilicon Heteropolymers and Heterocompounds,* Plenum Press, New York 1970.
19. H. R. Allcock: *Phosphorus-Nitrogen Compounds,* Academic Press, New York 1972.
20. H. R. Allcock: *Heteroatom Ring Systems and Polymers,* Academic Press, New York 1967.
21. M. G. Voronkov, V. P. Mileshkevich, Yu. A. Yuzhelevskii: *The Siloxane Bond,* Consultants Bureau, New York 1978.
22. A. L. Rheingold: *Homoatomic Rings, Chains and Macromolecules of Main Group Elements,* Elsevier Scientific Publ. Co., New York 1977.
23. C. E. Carraher, J. E. Sheats, C. U. Pittman, Jr., (eds.): *Organometallic Polymers,* Academic Press, New York 1978.
24. C. E. Carraher, J. E. Sheats, C. U. Pittman, Jr., (eds.): *Advances in Organometallic and Inorganic Polymers,* Marcel Dekker, New York 1982.
25. M. Zeldin, K. J. Wynne, H. R. Allcock (eds.): "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).
26. C. U. Pittman, Jr., C. E. Carraher, Jr., J. R. Reynolds in H. Mank, N. Bikales, C. Overberger and G. Menges (eds.): *Encyclopedia of Polymer Science and Engineering,* "Organometallic Polymers," vol. 10, 2nd ed., J. Wiley, 1987, p. 541.
27. M. D. Rausch et al., "New Vinyl Organometallic Monomers: Synthesis and Polymerization Behavior" in B. M. Culbertson, C. U. Pittmann, Jr. (eds.): *New Monomers and Polymers,* Plenum Press, New York 1984; F. R. Hartley: *Supported Metal Complexes,* Reidel Publishing Co., Boston 1985; J. E. Sheats, C. E. Carraher, Jr., C. U. Pittman, Jr.:

Metal-Containing Polymeric Systems, Plenum Press, New York 1985; C. U. Pittman, Jr., "Vinyl Polymerization of Organic Monomers Containing Transition Metals" in E. J. Becker, M. Tsutsui (eds.): *Organometallic Reactions,* vol. 6, Wiley-Interscience, New York 1977, p. 1; C. U. Pittman, Jr. "Polymer Supported Catalysts" in G. Wilkinson and F. G. A. Stone (eds.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Chap. 55, Pergamon Press, Oxford 1982.
28. J. P. Wesson, T. C. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **19** (1981) 65.
29. X. H. Zhang, R. West, *J. Polym. Sci. Polym. Chem. Ed.* **22** (1984) 159, 225.
30. R. West, L. D. David, P. I. Djurovich, H. Yu, R. Sinclair, *Am. Ceram. Soc. Bull.* **62** (1983) no. 8, 899.
31. R. West, *J. Organomet. Chem.* **300** (1986) 327.
32. R. West et al., *J. Am. Chem. Soc.* **103** (1981) 7352.
33. L. A. Harrah, J. M. Zeigler, *J. Polym. Sci. Polym. Lett. Ed.* **23** (1985) 209.
34. J. M. Zeigler, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 424.
35. K. Matyjaszewski, Y. L. Chen, H. K. Kim: "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).
36. R. West, S. Hayase, T. Iwahara, *J. Inorg. Organometal. Polym.* **1** (1991) 545.
37. J. Wilderman, J. Herrema, G. Hadziioannou, E. Schomaker, *J. Inorg. Organometal. Polym.* **1** (1991) 567.
38. C. Van Wakreem, T. Cleiji, J. Zwikker, L. Jenneskens, *Macromolecules* **28** (1995) 8696.
39. J. Herrema et al., *Macromolecules* **28** (1995) 8102.
40. C. Aitken, J. Harrod, E. Samuel, *Organomet. Chem.* **279** (1985) C11.
41. C. Aitken, J. Harrod, E. Samuel, *J. Amer. Chem. Soc.* **108** (1986) 4059.
42. J. Harrod, *ACS Symposium Ser.* **360** (1988) 89.
43. H. Woo, T. Tilley, *J. Amer. Chem. Soc.* **111** (1989) 3757.
44. T. Tilley, *Accounts Chemical Research* **26** (1993) 22.
45. J. Banovetz, K. Stein, R. Waymouth, *Organometallics* **10** (1991) 3430.
46. J. Corey, X. Zhu, T. Bedard, L. Lange, *Organometallics* **10** (1991) 924.
47. H. Li, F. Gauvin, J. Harrod, *Organometallics* **12** (1993) 575.
48. C. Forsyth, S. Nolan, T. Marks, *Organometallics* **10** (1991) 2543.
49. P. Bianconi, T. Weidman, *J. Amer. Chem. Soc.* **110** (1988) 2342.
50. W. Szymanski, G. Visscher, P. Bianconi, *Ultrasonics* **28** (1990) 310.
51. W. Szymanski, G. Visscher, P. Bianconi, *Macromolecules* **26** (1993) 869.
52. A. Sekignchi, M. Naugo, C. Kabuto, H. Sakurai, *J. Amer. Chem. Soc.* **117** (1995) 4195.
53. H. Suzuki, Y. Kimata, S. Satoh, A. Kuriyama, *Chem. Lett.* (1995) 293.
54. J. Lambert, J. Pfing, C. Stern, *Angew. Chem. Int. Ed. Engl.* **34** (1995) 98.
55. T. Imori, T. Tilley, *J. Chem. Soc. Commun.* (1993) 1607.
56. N. Devylder, M. Hill, K. Molloy, C. Price, *Chem. Commun.* (1996) 711.
57. V. Lu, T. Tilley, *Macromolecules* **29** (1996) 5763.
58. J. Reichi, C. Popoff, L. Gallagher, E. Remsen, D. Berry, *J. Amer. Chem. Soc.* **118** (1996) 9430.
59. M. Fujino, T. Hisaki, N. Matsumoto, *Macromolecules* **28** (1995) 5017.
60. S. Adams, J. Drager, *Angew. Chem. Int. Ed. Engl.* **26** (1987) 1255.
61. J. Mark, H. Allcock, R. West: *Inorganic Polymers,* Prentice Hall, New York 1992.
62. H. Suzuki, H. Meyer, J. Simmerer, J. Yang, D. Naarer, *Adv. Mater.* **5** (1993) 743.
63. R. West, L. David, P. Djurovick, H. Yu, *Am. Ceram. Soc. Bull.* **62** (1983) 899.
64. M. Kumada, K. Tamao, *Adv. Organomet. Chem.* **6** (1968) 19.
65. R. Barney, G. Chandra: *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vol. 13,** John Wiley & Sons, New York 1988, p. 312.
66. C. Schilling, *Br. Polymer J.* **18** (1986) 355.
67. J. Pilot et al., in A. Bassindale, P. Gaspar (eds.): *Frontiers in Organosilicon Chemistry,* Royal Society of Chemistry, Cambridge 1991.
68. W. Schmidt et al., *Chem. Mater.* **3** (1991) 257.
69. Y. Shieh, S. Sawan, *Polymer Preprints* **33** (1992) 1042.
70. D. Seyferth et al., *Polymer Reprints* **34** (1992) no. 1, 223.
71. D. Seyferth, P. Czubarow, *Chem. Mater.* **6** (1994) no. 1, 10.
72. S. Yajima, Y. Hasegawa, L. Hayasi, M. Iimura, *J. Mater. Sci.* **13** (1978) 2569.
73. R. West, A. Wolff, D. Peterson, *J. Radiation Curing* **35** (1986) 40.
74. A. Wolf, R. West, *Applied Organomet. Chem.* **1** (1987) 7.

75. R. Miller, G. Wallraff, N. Clecak, R. Sooriyakumaran, *PMSE* **60** (1989) 40.
76. A. Eckhardt, N. Yars, T. Wollny, W. Schnabel, S. Nespurek, *Ber. Bunsen-Ges. Phys. Chem.* **98** (1994) 853.
77. G. VanderLaan et al., *Macromolecules* **27** (1994) no. 7, 853.
78. I. Kminek et al., *Collect. Czech. Chem. Commun.* **58** (1993) no. 10, 2337.
79. F. Kajar, J. Messier, C. Rosilio, *J. Appl. Phys.* **60** (1986) 3040.
80. C. Callender, L. Robitaille, M. Lecleric, *Opt. Eng. (Bellingham, Wash.)* **32** (1993) 2246.
81. S. Grigoras, T. Banton et al., *Annu. Tech. Conf. Soc. Plast. Eng.* **50** (1992) no. 2, 2265.
82. S. Grigoras et al., *Polymer Preprints* **33** (1992) 655.
83. K. Matyjaszewski, J. Chrusciel, H. Kim, J. Maxka, *Polymer Preprints* **34** (1993) 67.
84. R. Corriu, W. Douglas, Z. Yang, Y. Karakus, G. Cross, D. Bloor, *J. Organomet. Chem.* **69** (1993) 455.
85. S. Grigoras et al., *Synth. Met.* **49** (1992) 293.
86. Y. Iwasa, T. Hasegawa, T. Koda, Y. Tokura, H. Tachibana, Y. Kawabata, *Synth. Met.* **50** (1992) 415.
87. F. Kajzar, J. Messier, C. Rossilio, *J. Appl. Phys.* **60** (1986) 3040.
88. D. Hofer, R. Miller, G. William, *APIE Advances in Resist Technology* **16** (1984).
89. R. Miller et al., *Polym. Eng. Sci.* **29** (1989) 882.
90. T. Weidman, A. Joshi, *Proc. SPIE-Int. Soc, Opt. Eng.; Advances in Resist Technology and Processing* **1925** (1993) 145.
91. G. Wallraff et al., *Polym. Mater. Sci. Eng.* **66** (1992) 105.
92. D. Hofer, R. Miller, C. Wilson, A. Neurath, *SPIE-Advances in Resist Technology* **469** (1984) 108.
93. C. Rosilio, A. Rosilio, B. Serre, *Microelectronic Eng.* **6** (1987) 399.
94. P. Trefonas, R. West, R. Miller, D. Hofer, *J. Polym. Sci., Polymer Lett. Ed.* **21** (1983) 823.
95. A. Gozdz, *Polym. Adv. Technol.* **5** (1994) 70.
96. A. Joslie, T. Weidman, A. Johnson, J. Miner, D. Ibbotson, *Proc. SPIE-Int. Soc. Opt. Eng., Adv. Resist Technology and Processing* **1925** (1993) 709.
97. M. Sakata, T. Ito, Y. Yamashita, *J. Photopolym. Sci. Technol.* **3** (1990).
98. R. Miller et al., "Polymer Microlithography," *ACS Symp. Ser.* (1989).
99. J. Su, T. Hsu, S. Sawan, J. Lavine, *Polymer Preprints* **33** (1992) 1046.
100. C. G. Wilson, in L. Thompson, G. Wilson, M. Bowden (eds.): *Introduction to Microlithography,* American Chemical Society, Washington DC 1983.
101. J. Zeigler, L. Harrah, A. Johnson, *SPIE-Advances in Resist Technology and Processing* **539** (1985) 166.
102. *Chem. Abstr.* **107** (1987) 218592u.
103. I.B.M. Corp., US 4 464 460, 1984 (H. Hiraoka).
104. D. C. Hofer, R. D. Miller, C. G. Willson, *J. Soc. Photo Opt. Instrum. Eng.* **469** (1984) 16.
105. D. C. Hofer, R. D. Miller, C. G. Willson, A. R. Neureuther, *J. Soc. Photo Opt. Instrum. Eng.* **469** (1984) 108.
106. R. West, A. R. Wolff, D. J. Peterson, *J. Radiat. Curing* **13** (1986) 35.
107. R. D. Miller et al.: "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).
108. J. Michl et al.: "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).
109. Y. Hasegawa, M. Iimura, S. Yajima, *J. Mater. Sci.* **15** (1980) 1209.
110. R. G. Kepler, J. M. Zeigler, L. A. Harrah, S. R. Kurtz, *Phys. Rev. B* **35** (1987) 2818.
111. F. Kazar, J. Messier, C. Rosilio, *J. Appl. Phys.* **60** (1986) 3040.
112. R. West "Organopolysilanes" in G. Wilkinson, F. G. A. Stone, E. Abel (eds.): *Comprehensive Organometallic Chemistry,* vol. 9, Pergamon Press, Oxford, U.K., 1983.
113. R. Seymour, C. Carraher: *Polymer Chemistry,* Marcel Dekker, New York 1987.
114. C. Gebelein, C. Carraher: *Polymeric Materials in Medication,* Plenum Press, New York 1985.
115. R. West, J. Maxka, R. Sinclair, P. Cotts, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 387.
116. J. F. Harrod, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 403.
117. R. W. Williams, *Pure Appl. Chem.* **29** (1972) 569.
118. E. N. Peters, *J. Macromol. Sci. Rev. Macromol. Chem.* **C17** (1979) 173.
119. K. O. Knollmueller, R. N. Scott, H. Kawasnik, J. R. Sieckhaus, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 1071.
120. E. N. Peters et al., *Rubber Chem. Technol.* **48** (1979) 14.
121. S. Papetti, B. B. Schaeffer, A. P. Gray, T. L. Heying, *J. Polym. Sci. Polym. Chem. Ed.* **4** (1966) 1623.
122. E. N. Peters, D. D. Stewart, *J. Polym. Sci. Polym. Lett. Ed.* **17** (1979) 405.

123. M. B. Roller, J. K. Gillham, *Polym. Eng. Sci.* **8** (1974) 567.
124. E. N. Peters et al., *J. Polym. Sci. Polym. Phys. Ed.* **15** (1977) 723.
125. L. Interrante et al., *J. Amer. Chem. Soc.* **116** (1994) 12025.
126. H. Wu. L. Interrante, *Macromolecules* **25** (1992) 1840.
127. I. Rushkin, L. Interrante, *Macromolecules* **28** (1995) 5160.
128. M. Lienhard et al., *J. Amer. Chem. Soc.* **116** (1994) 12020.
129. M. Lienhard, C. Weigand, T. Apple, B. Farmer, L. Interrante, *Polymer Preprints* **39** (1998) no. 2, 808.
130. S. Yamjima, Y. Hasegawa et al., *J. Mater. Sci.* **12** (1978) 2569.
131. K. Wynne, R. Rice, *Ann. Rev. Mater. Sci.* **14** (1984) 297.
132. R. Laine, Y. Blum et al., *ACS Symp. Ser.* **360** (1988) 142.
133. J. Mark, H. Allcock, R. West, *Inorganic Polymers,* Prentice Hall, Englewood Cliffs, NJ 1992, Chap. 6.
134. D. Seyferth, G. Wiseman, C. Prud'homme, *J. Am. Ceram. Soc.* **66** (1983) C13.
135. Rhône-Poulence, EP 197 863 B1, 1988 (J. Lebrun, H. Porte).
136. Bayer, US 3 853 567, 1974 (W. Verbeck).
137. Dow Corning, US 4 340 619, 1982 (J. Gaul, Jr.).
138. H. Allcock, *Adv. Mater.* **6** (1994) 106.
139. H. Allcock, S. Kuharcik, *J. Inorg. Organometal. Polym.* **6** (1996) no. 1, 1.
140. P. Bortolus, M. Gleria, *J. Inorg. Organometal. Polym.* **4** (1994) no. 2, 95 and **4** (1994) no. 3, 205.
141. *J. Inorg. Organometal. polym.* **6** (1996) no. 3, and **6** (1996) no. 4.
142. G. Bosscher, A. Jekel, J. van der Grampel, *J. Inorg. Organometal. Polym.* **7** (1997) no. 1, 19.
143. H. R. Allcock: *Phosphorus-Nitrogen Compounds,* Academic Press, New York 1972.
144. L. Hagnauer, C. Carraher, J. Sheats, C. U. Pittman, Jr. (eds.): *Advances in Organometallic and Inorganic Polymer Science,* Chap. 17, Marcel Dekker, New York 1982.
145. R. Neilson, P. Wisian-Neilson in C. E. Carraher, Jr., J. E. Sheats, C. U. Pittman, Jr. (eds.): *Advances in Organometallic and Inorganic Polymer Sciences,* Chap. 19, Marcel Dekker, New York 1982, p. 425.
146. H. R. Allcock, *CHEMTECH* **5** (1975) 552.
147. H. R. Allcock, *Angew. Chem. Int. Ed. Engl.* **16** (1977) 147.
148. S. H. Rose, *J. Polym. Sci. Part. B* **6** (1968) 837.
149. C. Carraher, *J. Macromol. Sci. Chem.* **A17** (1982) no. 8, 1293.
150. H. R. Allcock, *Polymer Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 437.
151. R. H. Neilson et al., in M. Zeldin, K. J. Wynne and H. R. Allcock (eds.): "Inorganic and Organometallic Polymers," Chap. 22, *ACS Symp. Ser.* **360** (1988).
152. H. Allcock, J. Nelson, S. Reeves, C. Honeyman, I. Manners, *Macromolecules* **29** (1996) 7740.
153. J. Nelson, H. Allcock, *Macromolecules* **30** (1997) 1854.
154. H. Allcock, S. Reeves, J. Nelson, C. Crane, *Macromolecules* **30** (1997) 22B.
155. J. Nelson, H. Allcock, *Polymer Reprints* **39** (1998) no. 2, 631.
156. I. Dez, R. de Jaeger, *J. Inorg. Organometal. Polym.* **6** (1996) no. 2, 111.
157. D. Gates, I. Manners, *J. Chem. Soc., Dalton Trans.* (1997) 2525.
158. H. Allcock, S. Kuharcik, *J. Inorg. Organometal. Polym.* **5** (1995) no. 4, 307.
159. J. C. Vicic, K. A. Reynard, *J. Appl. Polym. Sci.* **21** (1977) 3185.
160. R. Singler, G. L. Hagnauer in C. Carraher, J. Sheats, C. U. Pittman, Jr. (eds.): *Organometallic Polymers,* Chap. 26, Academic Press, New York 1978.
161. R. E. Singler, N. S. Schneider, G. L. Hagnauer, *Polym. Eng. Sci.* **15** (1975) 321.
162. H. R. Penton, *Pol. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 437.
163. C. Kim, H. R. Allcock, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 446.
164. R. Singler, R. Willingham, R. Lenz, A. Furukana, H. Finkelmann, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 448.
165. G. S. Kyker, T. A. Antkowiak, *Rubber Chem. Technol.* **47** (1974) 32.
166. D. F. Lohr, J. A. Beckman, *Rubber and Plastics News* **16** (1982).
167. H. R. Penton in M. Zeldin, K. J. Wynne and H. R. Allcock (eds.): "Inorganic and Organometallic Polymers," Chap. 21, *ACS Symp. Ser.* **360** (1988).
168. Z. Pang et al., *Adv. Mater.* **8** (1996) 768.
169. W. Balmain, *Phil. Mag.* **21** (1842) 10.

170. R. Wentorf, *J. Chem. Phys.* **26** (1957) 956.
171. T. O'Connor, *J. Amer. Chem. Soc.* **84** (1962) 1753.
172. F. Bundy, R. Wentorf, *J. Chem. Phys.* **38** (1963) 1144.
173. *Gmelin,* Boron Compounds, 2nd Suppl. 1 (1983) 304.
174. N. J. Archer, *Spec. Publ. Chem. Soc.* No. 30 (1977) 167.
175. C. S. Kalyoncu, *Ceram. Eng. Sci. Proc.* (1985) 1356.
176. K. Gonslaves, X. Chen, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 5,** CRC Press, Boca Raton, FL 1996, p. 3259.
177. A. J. Banister, I. B. Gorrell, *Adv. Mater.* **10** (1998) no. 17, 1415.
178. T. Takahashi, H. Itoh, A. Takeuchi, *J. Cryst. Growth* **47** (1979) 245.
179. M. Sano, M. Aoki, *Thin Solid Films* **83** (1981) 247.
180. K. L. Chopra et al., *Thin Solid Films* **126** (1985) 307.
181. F. P. Burt, *J. Chem. Soc.* 1910, 1171.
182. C. Mikulski et al., *J. Am. Chem. Soc.* **97** (1975) 6358.
183. M. Akhtar et al., in C. Carraher, J. Sheats, C. U. Pittman, Jr. (eds.): *Organometallic Polymers,* Chap. 30, Academic Press, New York 1978.
184. R. H. Baughman, P. A. Apgar, R. R. Chance, A. G. MacDiarmid, A. F. Garito, *J. Chem. Phys.* **66** (1977) 401.
185. A. G. MacDiarmid in R. B. King (ed.): *Inorganic Compounds with Unusual Properties,* Chap. 6, American Chemical Society, Washington D.C. 1976.
186. R. L. Green et al., *Lect. Notes Phys.* **65** (1977) 603.
187. J. Passmore, M. Rao, *J. Chem. Soc., Chem. Commun,* **1980** 1268.
188. F. Kennett, G. MacLean, J. Passmore, M. Rao, *J. Chem. Soc., Dalton Trans* **1982** 851.
189. A. Banister, Z. Hauptam, J. Passmore, C. Wong, P. White, *J. Chem. Soc., Dalton Trans* **1986** 2371.
190. A. Banister, Z. Haupton, A. Kendrick, *J. Chem. Soc., Chem. Commun.* **1983** 1016.
191. A. Banister, Z. Hauptman, A. Kendrick, R. Small, *J. Chem. Soc., Dalton Trans* **1987** 915. H. Fritz, R. Bruchhaus, *Z. Naturforsch. B.* **38** (1983) 1375.
192. R. L. Greene, G. B. Street, L. J. Suter, *Phys. Rev. Lett.* **34** (1975) 577.
193. W. H. Jones, R. Bardo, *J. Phy. Chem.* **97** (1993) 4974.
194. I. Kakada, *J. Cryst. Growth* **55** (1981) 447.
195. H. Kahlert, B. Kundu, *Mater. Res. Bull.* **11** (1976) 967.
196. R. Scranton, *J. Appl. Phys.* **48** (1977) 3838.
197. R. Scranton, J. Best, J. McCaldin, *J. Vac. Sci. Technol.* **14** (1977) 930.
198. M. Cohen, J. Harris, *Appl. Phys. Lett.* **33** (1978) 812.
199. A. Thomas, J. Woods, Z. Hauptman, *J. Phys. D: Appl. Phys.* **16** (1983) 1123.
200. Asahi Chemical Industry Co. Ltd., JP 55 137 671, 1980; 58 93 164, 1983; 58 93 177, 1983; 59 49 155, 1984.
201. Electricité de France Service National, EP 97 078 B1, 1986 (J. Badoz et al.).
202. Sony Corp., JP 55 108 944 A, 1980 (T. Yoshioka, M. Ogawa, K. Hayashi).
203. Sony Corp., EP 517 165 A1, 1992; EP 517 165 B1, 1997; JP 07 230 984, 1955 (S. Kadomura).
204. Sony Corp., JP 07 263 420, 1995 (T. Nagayama).
205. Sony Corp., JP 07 263 426, 1995 (N. Akiba).
206. The United States of America as represented by the Secretary of the Army, US 4 206 705, 1980 (Z. Iqbal, H. Fair, D. Downs).
207. Nippon Denso Co. Ltd., JP 59 129 261, 1984 (H. Murametsu).
208. Sony Corp., JP 07 130 735, 1995 (T. Nagayama).
209. Sony Corp., JP 09 148 285, 1997 (J. Sato).
210. Toshiba Corp., JP 05 136 458, 1993 (T. Kamakura).
211. H. B. Mark, A. Voulgaropoulos, C. Meyer, *J. Chem. Soc., Chem. Commun.* (1981) 1021.
212. J. Rubinson, T. Behymer, H. Mark, *J. Amer. Chem. Soc.* **104** (1982) 1224.
213. R. Nowak, C. Joyal, D. Weber, *J. Electroanal. Chem.* **143** (1983) 413.
214. B. R. Currell, A. J. Williams, A. J. Mooney and B. J. Nash, *Adv. Chem. Ser.* **1** (1974) 1.
215. J. Sheats, C. Carraher, C. Pittman, M. Zeldin, B. Currell: *Inorganic and Metal-Containing Polymeric Materials,* Plenum, New York 1990.
216. C. Pittman, C. Carraher, M. Zeldin, B. Culbertson, J. Sheats: *Metal-Containing Polymeric Materials,* Plenum, New York 1996.
217. M. Zeldin, K. Wynne, H. Allcock: *Inorganic and Organometallic Polymers,* ACS, Washington, DC 1998.
218. Macromolecule-Metal Complexes III, Seclected Papers from seminar of July 1989; *J. Macromol. Sci. Chem.* **A27** (1990) nos. 9 – 11, 1109.

219. I. Manners, *Ann. Rep. Prog. Chem., Sect. A. Inorg. Chem.* **88** (1991) 77; **89** (1992) 93; **90** (1993) 103; **91** (1994) 131; **92** (1995) 127; **93** (1996) 129.
220. J. Mark, H. Allcock, R. West, *Inorganic Polymers,* Prentice Hall, New York 1992.
221. J. Sheats, C. Carraher, C. Pittman, M. Zeldin, in C. Craver, C. Carraher (eds.): *Applied Polymer Science,* Elsevier, New York 1999.
222. S. Takahashi, K. Sonogashira, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4804.
223. A. D. Hunter, A. Guo, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4813.
224. T. Endo, I. Tomita, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4822.
225. R. Wei, L. Ya, W. Jinguo, X. Qifeng, in, J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4826.
226. W. L. Driessen (ed.): *Macromolecule-Metal Complexes,* Huthig & Wepf, Oxford, CT 1998.
227. F. Ciardelli, E. Tsuchida, D. Wohrle (eds.): *Macromolecule-Metal Complexes,* Springer-Verlag, Berlin 1996.
228. A. Pomogailo, V. Savost'yanov: *Synthesis and Polymerization of Metal-Containing Monomers,* CRC Press, Boca Raton, New York 1994.
229. B. P. Block, *Inorg. Macromol. Rev.* **1** (1970) no. 2, 115.
230. C. Carraher, R. Schwarz, J. Schroeder, M. Schwarz, *J. Macromol. Sci. Chem.* **A15** (1981) no. 5, 773.
231. E. Neuse, H. Rosenberg: *Metallocene Polymers,* Marcel Dekker, New York 1970.
232. Dow Corning, GB 1 582 081, 1980.
233. C. Carraher, F. Li, D. Siegman-Louda, C. Butler, J. Ross, *PMSE* **77** (1997) 499.
234. C. Carraher, V. Saurino, C. Butler, D. Sterling, *PMSE* **72** (1995) 192.
235. C. Carraher, D. Sterling, C. Butler, T. Ridgway, *Biotechnical Polymers,* Chap. 13, Technomic, Lancaster, PA 1995.
236. C. Carraher, M. Nagata, H. Stewart, S. Miao, S. Carraher, A. Gaonkar, C. Highland, F. Li, *PMSE* **79** (1998) 52.
237. C. Carraher, A. Gaonkar, H. Stewart, S. Miao, D. Mitchell, C. Barosy, M. Colbert, R. Duffield, *Polymer Reprints* **38** (1997) no. 2, 572.
238. C. Carraher, H. Stewart, W. Soldani, J. Dela Torre, B. Pandya, L. Reckleben, *Metal-Containing Polymeric Materials,* Plenum, NY 1996.
239. C. Carraher et al., *Polym. Mater.* **55** (1986) 469.
240. C. Carraher, V. Foster, R. Linville, D. Stevison, *Polym. Mater.* **56** (1987) 401.
241. C. Carraher, R. Schwarz, J. Schroeder, M. Schwarz, H. M. Molloy, *Org. Coat. Plast. Chem.* **43** (1981) 798.
242. C. Carraher, R. Schwarz, J. Schroeder, M. Schwarz in C. Carraher, J. Preston (eds.): *Interfacial Synthesis,* vol. 3, Chap. 6, Marcel Dekker, New York 1982.
243. C. Carraher et al., *Org. Coat. Plast. Chem.* **44** (1982) 753.
244. J. Wrobleski, D. B. Brown, *Inorg. Chem.* **18** (1979) 498, 2738.
245. H. Matsuda, H. Nakaniski, M. Kato, *J. Polym. Sci. Polym. Lett. Ed.* **22** (1984) 107.
246. J. E. Sheats, C. Carraher, C. Pittman (eds.): *Metal-Containing Polymeric Systems,* Plenum Press, New York 1985.
247. X. Zhang, R. West, *J. Polym. Sci. Polym. Chem. Ed.* **22** (1984) 159.
248. R. Liepins, M. Timmons, N. Morosoff, J. Surles in J. Sheats, C. Carraher, C. Pittman (eds.): *Metal-Containing Polymeric Systems,* Plenum Publishing, New York 1985, pp. 225 – 236.
249. H. Yasuda et al., in J. Sheats, C. Carraher, C. U. Pittman (eds.): *Metal-Containing Polymeric Systems,* Plenum Publishing, New York 1985, pp. 275 – 290.
250. C. Carraher et al., *Organic Coatings Plastics Chemistry* **44** (1981) 753.
251. E. Savinova et al., *J. Mol. Catal.* **32** (1985) 149.
252. M. Kaneko, A. Yamada, *Adv. Polym. Sci.* **55** (1983).
253. M. S. Wrighton (ed.): *Interfacial Photoprocesses: Energy Conversion and Synthesis,* American Chemical Society, Washington 1980.
254. U. T. Müller-Westerhoff, A. I. Nazzal, US 4 379 740, 1983.
255. F. C. Anson et al., *J. Am. Chem. Soc.* **106** (1984) 59.
256. E. R. Savinova et al., *J. Mol. Catal.* **32** (1985) 149, 159.
257. in [23] pp. 137 – 147.
258. B. J. Spalding, *Chemical Week* **135** (1986) no. Oct. 8, 29.
259. C. Carraher, S. Tsuji, J. DiNunzio, W. Feld, *Polym. Mater.* **55** (1986) 875.

260. C. Carraher, J. Schroeder, *Polym. Prepr.* **16** (1975) 659, *J. Polym. Sci. Polym. Lett. Ed.* **13** (1975) 215.
261. L. Donaruma, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **22** (1981) 1.
262. B. Jiang et al., *Coord. Chem. Rev.* **171** (1998) 365.
263. W. E. Jones, L. Hermans, B. Jiang, in V. Ramamurthy, K. Schanze (eds.): *Molecular and Supramolecular Photochemistry,* **vol. 2,** Marcel Dekker, New York 1999.
264. C. Pittman, *J. Paint Technol.* **39** (1967) 585.
265. N. E. Wolff, R. J. Pressley, *Appl. Phys. Lett.* **2** (1963) 152.
266. in [17] p. 425.
267. J. E. Sheats et al., in B. M. Culbertson, C. U. Pittman, Jr. (eds.): *New and Unusual Monomers and Polymers,* Plenum Press, New York 1983, pp. 83 – 98; *Polym. Mater. Sci. Eng.* **49** (1983) no. 2, 363.
268. N. Bilow, H. Rosenberg, *J. Polym. Sci. Part A-1* **7** (1969) 2689.
269. N. Bilow, H. Rosenberg, US 3 640 961, 1972.
270. C. U. Pittman, Jr., R. F. Felis, *J. Organomet. Chem.* **72** (1974) 389.
271. C. U. Pittman, R. F. Felis, *J. Organomet. Chem.* **72** (1974) 399.
272. R. Tannenbaum, E. P. Goldberg, C. L. Flenniken in J. Sheats, C. Carraher, C. Pittman (eds.): *Metal-Containing Polymeric Systems,* Plenum Publishing, New York 1985, pp. 303 – 340.
273. C. Carraher, X. Xu, *PMSE* **75** (1996).
274. C. Carraher, X. Xu: *Modification of Polymers,* Plenum, New York 1997.
275. C. Carraher, V. Forster, R. Linville, D. Steverson, R. Ventatachalam, *Adhesives, Sealants, and Coatings for Space and Harsh Environments,* Chap. 19, Plenum, NT 1988.
276. C. Carraher, A. Taylor-Murphy, *PMSE* **76** (1997) 409.
277. S. Yajima, H. Hayashi, M. Omori, *Chem. Letters.* (1975) 931 and 1209.
278. R. West, *Am. Ceram. Soc. Bull.* **62** (1983) 825.
279. D. Seyferth, G. H. Wiseman, *J. Amer. Ceram. Soc.* **67** (1984); Massachusetts Institute of Technology, US 4 482 669, 1984 (D. Seyferth, G. H. Wiseman).
280. R. Laine, Y. Blum, A. Chow, R. Hamlin, K. Schwartz, D. Rowecliff, *Polymer Reprints* **28** (1987) no. 1, 393.
281. M. Arai, S. Sakurada, T. Isoda, T. Tomizawa, *Polymer Reprints* **28** (1987) no. 1, 407.
282. C. Narula, R. Paine, R. Schaeffer, *Polymer Reprints* **28** (1987) no. 1, 454.
283. C. Carraher, H. Zhuang, F. Medina, D. Baird, R. Pennisi, B. Landreth, F. Nounou, *PMSE* **62** (1990) 633.
284. M. Kaneko, A. Yamada, *Adv. Polym. Sci.* (1983) 55.
285. US 4 379 740, 1983 (U. Muller-Westerhoff, A. Nazzal).
286. F. Anson, *J. Amer. Chem. Soc.* **106** (1984) 59.
287. B. J. Spaulding, *Chemical Week, Oct 8 (1986)29 and C. M. Elliott, Abstracts, 193 ACS Meeting, Denver CO, April, 1987, INOR 232.*
288. M. MacLachlan, P. Aroca, N. Coombs, I. Manners, G. Ozin, *Adv. Mater.* **10** (1998) 144.
289. *Proceedings of the IIIrd Internat. Conference on Ultrastructure Processing of Glasses, Ceramics and Composites,* Feb. 24 – 27, 1987.
290. C. J. Brinker, D. Clark, D. R. Ulrich (eds.): *Better Ceramics Through Chemistry II,* Materials Res. Soc., Pittsburgh, Pa., 1986.
291. L. L. Hench, D. R. Ulrich (eds.): *Science of Ceramic Chemical Processing,* Wiley-Interscience, New York 1986.
292. J. O'Reilly, B. Coltrain, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4772.
293. G. Wilkes, J. Wen, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4782.
294. Y. Chujo, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4793.
295. C. J. Brinker et al., "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988); D. R. Ulrich, *CHEMTECH* **18** (1988) 242.
296. H. K. Schmidt, "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).
297. H. Schmidt, H. Scholze, G. Tunker, *J. Non Cryst. Solids* **80** (1986) 557.
298. H. Schmidt, H. Scholze, *Springer Proceedings in Physics,* **vol. 6,** Aerogels, Springer Publishing Co., 1986, p. 49.
299. G. Philipp, H. Schmidt, *J. Non Cryst. Solids* **63** (1984) 261.
300. G. Tunker, H. Patzelt, H. Schmidt, H. Scholze, *Glastechn. Ber.* **59** (1986) 272.
301. S. Sakka, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) no. 1, 430.
302. P. Brooke, H. Schurmans, J. Verhoest: *Inorganic Fibers and Composite Materials: A Survey of Recent Developments,* Pergamon Press, Oxford, UK, 1984.
303. H. Huang, R. H. Glaser, G. L. Wilkes, "Inorganic and Organometallic Polymers," *ACS Symp. Ser.* **360** (1988).

304. H. Dislich, *Angew. Chem. Int. Ed. Engl.* **106** (1970) 363.
305. B. E. Yoldas, *J. Mater. Sci.* **12** (1977) 203; **14** (1979) 1843.
306. C. Carraher, *CHEMTECH* 1972, 741.
307. V. C. Farmer, A. R. Fraser, J. M. Tait, *J. Chem. Soc. Chem. Commun.* 1977, 462.
308. K. Wada et al., *Clay Miner.* **8** (1970) 487.
309. G. Alberti, U. Costantino, *J. Chromatogr.* **102** (1974) 5.
310. N. H. Ray: *Inorganic Polymers,* Chap. 7, Academic Press, New York 1978.
311. J. D. Birchall, J. A. Bradbury, J. Dinwoodie in W. Watt, B. V. Perov (eds.): *Handbook of Composites,* North Holland, Amsterdam 1985, p. 115.
312. J. D. Birchall: "Inorganic Fibers" in M. B. Bever (ed.): *Encyclopedia of Material Sciences and Engineering,* Pergamon Press, Oxford, UK 1986, p. 2333.
313. Sumitomo Chemical Co., US 4 101 615, 1973 (S. Horikiri, K. Tsuji, Y. Abe); 4 152 149, 1977 (S. Horikiri, K. Tsuji, Y. Abe).
314. J. Tretzel et al., "Production, Processing and Application of Enka Silica Fibers," *57th Annual TRI Conference,* Charlotte, N.C., April 21 – 22, 1987.
315. W. C. Miller in M. Grayson (ed.): *Encyclopedia of Textiles, Fibers and Nonwoven Fabrics,* J. Wiley and Sons, New York 1984, p. 443.
316. *Chem. Week* **13** (1987) no. April 22.
317. J. C. Romine, *Ceram. Eng. Sci. Proc.* **8** (1987) 755.
318. J. O. Carlsson: "Boron Fibers" in M. B. Bever (ed.): *Encyclopedia of Material Science and Engineering,* Pergamon Press, Oxford, UK, 1986, p. 402.
319. J. O. Carlsson: "Silicon Carbide Fibers" in M. B. Beuer (ed.): *Encyclopedia of Material Science and Engineering,* Pergamon Press, Oxford, UK, 1986, p. 4406.
320. S. Yajima, Y. Hasegawa, J. Hayashi, M. Imura, *J. Mater. Sci.* **13** (1978) no. 12, 2569.
321. Y. Hasegawa, K. Okamura, *Yogyo Kyokaishi* **95** (1987) 99, 103.
322. W. Verbeck, DE-OS 2 218 960, 1973.
323. K. Okamura, M. Sato, Y. Hasegawa, *Ceram. Int.* **13** (1987) 55.
324. T. Yamamura et al., "Proc. World Congress on High Tech Ceramics, 6th Int. Meeting on Modern Ceramics Technologies," *Mater. Sci. Monogr.* **38** (1987).
325. N. Tsubokawa, T. Yoshihara, H. Ueno, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 2,** CRC Press, Boca Raton, FL 1996, p. 973.
326. C. Carraher: *Polymer Chemistry: An Introduction,* 4th ed., Dekker, New York 1996, p. 122.
327. K. Okamura, *Composites* **18** (1987) 107.
328. W. Noll: *Chemistry and Technology of Silicones,* Academic Press, New York 1968.
329. W. Lynch: *Handbook of Silicon Rubber Fabrication,* Van Nostrand, New York 1978.
330. M. Ranney: *Silicones,* Noyes Data Corp., Park Ridge, N.J., 1977.
331. B. Arkles, W. Peterson, R. Anderson: *Organosilicon Compounds: Register and Review,* Petrarch Systems, Bristol, Pa., 1982.
332. B. Arkles, *CHEMTECH* **13** (1983) 542.
333. A. Isquith, L. Abbot, P. Walters, *Appl. Microbiol.* **24** (1972) 859.
334. B. Arkles, W. Brinigar, A. Miller in D. Leyden, W. Collin (eds.): *Silylated Surfaces,* Gordon Breach, New York 1980.
335. E. Sako, O. Kawamura, US 3 746 196, 1973.
336. W. Robb, *Ann. N.Y. Acad. Sci.* **146** (1968) 119.
337. M. Nelson, US 3 020 260, 1962.
338. E. Jeram, US 4 072 635, 1978.
339. T. Cronin, US 3 293 663, 1965.
340. I. V. Yannas, J. F. Burke, *J. Biomed. Mater. Res.* **14** (1980) 65, 107.
341. N. Dagalakis et al., *J. Biomed. Mater. Res.* **14** (1980) 511.
342. R. A. Erb, US 4 245 623, 1981.
343. R. A. Erb, T. P. Reed, *J. Reprod. Med.* **23** (1979) 65.
344. A. Zaffaroni, US 3 905 360, 1975.
345. P. Hittmaier, W. Hechtel, H. Eckhart, US 4 035 453, 1977.
346. C. N. Raptis, R. Yu, J. G. Knapp, *J. Prosthet. Dent.* **4414** (1980) 447.
347. A. B. Swanson, US 3 875 594, 1975.
348. T. Honda et al., *J. Thorac. Cardiovasc. Surg.* **69** (1975) 92.
349. D. Laurin, US 3 994 988, 1976.
350. D. G. LeGrand, J. Magila, *Eng. Sci.* **10** (1970) 349.
351. J. Riffle, R. Freelin, A. Banthia, J. McGrath, *J. Macromol. Sci. Chem.* **A15** (1981) 967.
352. E. Pierce, N. Dibelus, *Trans. Am. Soc. Artif. Intern. Organs* **14** (1968) 220.
353. J. Montalvo, US 3 869 354, 1975.
354. J. Brown, G. Slusarczuk, O. LeBlanc, US 3 743 588, 1971.
355. H. Allcock in C. Carraher, J. E. Sheats, C. U. Pittman, Jr. (eds.): *Organometallic Polymers,* Academic Press, New York 1978, p. 283.

356. H. Allcock, R. Allen, J. O'Brien, *J. Chem. Soc. Chem. Commun.* 1976, 717.
357. C. Carraher, W. J. Scott, J. Schroeder, D. J. Giron, *J. Macromol. Sci. Chem.* **A15** (1981) no. 4, 625.
358. C. Carraher et al., in C. Gebelein, C. Carraher (eds.): *Polymeric Materials in Medication,* Chap. 15, Plenum Publishing, New York 1985.
359. C. Carraher in C. Gebelein, C. Carraher (eds.): *Bioactive Polymeric Systems,* Chap. 22, Plenum Publishing, New York 1985.
360. C. Carraher, I. Lopez, D. Giron, *Polym. Mater.* **53** (1985) 644.
361. C. Carraher, I. Lopez, D. Giron in C. Gebelein (ed.): *Advances in Biomedical Polymers,* Plenum Publishing, New York 1987, p. 311.
362. C. Carraher, N. Bigley, M. Trombley, D. Giron, *Polym. Mater.* **57** (1987) 177.
363. A. Winston, in C. Gebelein, C. Carraher (eds.): *Bioactive Polymeric Systems,* Chap. 21, Plenum Publishing, New York 1985.
364. A. Winston, *Polym. News* **10** (1984) 6.
365. C. Carraher et al., in C. Carraher, C. Gebelein (eds.): *Biological Activities of Polymers,* Chap. 2, American Chemical Society, Washington, D.C., 1982.
366. D. Anderson, J. Mendoza, B. Garg, R. V. Subramanian in C. Carraher, C. Gebelein (eds.): *Biological Activities of Polymers,* Chap. 3, American Chemical Society, Washington, D.C., 1982.
367. R. V. Subramanian, B. Garg, J. Corredor in C. Carraher, J. Sheats, C. U. Pittman, Jr. (eds.): *Organometallic Polymers,* Chap. 19, Academic Press, New York 1978.
368. C. Carraher et al., *Polym. Mater.* **57** (1987) 186.
369. C. Carraher, D. Giron, J. Schroeder, C. McNeely, US 4 312 981, 1982.
370. J. A. Montemarano, E. J. Dyckman, *J. Paint Technol.* **47** (1975) 59.
371. A. T. Phillip, *Prog. Org. Coat.* **2** (1973/74) 159.
372. J. Leebrick, US 3 167 473, 1965.
373. D. Atherton, J. Verborgt, M. A. Winkeler, *J. Coat. Technol.* **51** (1979) 88.
374. S. Matsuda, H. Kudara, US 4 174 339, 1979.
375. Shipcare International, 10 (1978) no. 11, 23.
376. R. W. Drisko, T. B. O'Neill, L. K. Schwab, Technical Note no. N-1480, Naval Facilities Engineering Command, Point Hueneme, CA, May 1977.
377. Y. Chujo, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 6,** CRC Press, Boca Raton, FL 1996, p. 4795.
378. W. Gerrard, *J. Chem. Soc.* 1960, 3170.
379. R. Bedell, M. J. Frazer, W. Gerrard, *J. Chem. Soc.* 1960, 4037.
380. R. W. Upson, US 2 511 310, 1950.
381. K. Hizawa, E. Nojimoto, JP 4791, 1953.
382. J. M. Thompson: "Other Boron-Containing Polymers" in M. F. Lappert, G. J. Leigh (eds.): *Developments in Inorganic Polymer Chemistry,* Elsevier, Amsterdam 1962, p. 57.
383. R. M. Washburn, E. Levens, C. F. Albright, F. A. Billig: "Metal-Organic Compounds," *Adv. Chem. Ser.* **23** (1959) 139, 153.
384. W. Kuchen, R. Brinkmann, *Angew. Chem.* **72** (1960) 564.
385. A. B. Burg, *Angew. Chem.* **72** (1960) 183.
386. R. I. Wagner, F. F. Caserio, *J. Inorg. Nucl. Chem.* **11** (1959) 259.
387. T. Wartik, E. F. Apple, *J. Am. Chem. Soc.* **77** (1955) 6400.
388. H. Rosenberg, F. L. Hedberg, Tech. Rept. AFML-TR-69-68, AD 863, 817 (1969).
389. H. Steinberg, R. J. Brotherton: *Organoboron Chemistry,* **vol. 2,** J. Wiley and Sons, New York 1966, pp. 329 – 340, 363 – 364, 418.
390. B. M. Mikhailov, A. F. Galkin, *Bull. Acad. Sci. USSR, Div. Chem. Sci. (Engl. Transl.)* **12** (1963) 575.
391. R. H. Toeniskoetter, R. Didchenko, CA 685 580, 1964 to Union Carbide.
392. R. H. Toeniskoetter, K. A. Killip, *J. Am. Chem. Soc.* **86** (1964) 690.
393. V. Gutmann, A. Metter, R. Schlegel, Monatsh. Chem. **94** (1963) 1071.
394. G. Chainani et al., *J. Appl. Chem.* **15** (1965) 372.
395. K. O. Knollmuller, R. N. Scott, H. Kwasnik, J. F. Sieckhaus, *J. Polym. Sci. Polym. Chem. Ed.* A1, 9 (1971) 1071.
396. S. Papetti et al., *J. Polym. Sci. Polym. Chem. Ed.* **4** (1966) 1623.
397. R. P. Alexander, H. Schroeder, *Angew. Chem.* **76** (1964) 278.
398. H. A. Schroeder et al., *Inorg. Chem.* **8** (1969) no. 11, 2444.
399. H. A. Schroeder, *Inorg. Macromol. Rev.* **1** (1970) 45.
400. I. B. Atkinson, D. B. Clapp, W. F. Flavell, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **13** (1972) 770.
401. N. Semenuk, S. Papetti, H. Schroeder, *Inorg. Chem.* **8** (1969) no. 11, 2441.
402. J. Rinse, *Ind. Eng. Chem.* **56** (1964) no. 5, 42.
403. DuPont, US 2 744 071, 1956 (C. N. Theobald).
404. K. C. Pande, R. C. Mehrotra, *Z. Anorg. Allg. Chem.* **286** (1956) 291.

405. Peter Lunt and Co., GB 573 083, 1945 (E. J. Lush).
406. Hardman and Holden Ltd., GB 772 144, 1957 (J. Rinse).
407. A. L. McCloskey et al., WADC Tech. Rept. 59-761 (1959) pp. 117 – 200.
408. Koninklijke Industrieele Maatschappij voorheen Noury, GB 783 679, 1957 (N. V. Van der Lande).
409. K. A. Andrianov: *Metalorganic Polymers,* Chap. 5, Interscience Publishers, New York 1965.
410. K. A. Andrianov, V. M. Novikov, *Vysokomol. Soedin.* **1** (1959) 1390.
411. V. Kugler, *J. Polym. Sci.* **29** (1958) 637.
412. T. R. Patterson et al., *J. Am. Chem. Soc.* **81** (1959) 4213.
413. J. I. Jones: "Polymetallosiloxanes" in M. L. Lappert, G. J. Leigh (eds.): *Developments in Inorganic Polymer Chemistry,* Part I, Part II, Chap. 7, 8, Elsevier, New York 1962.
414. D. C. Bradley: "Polymeric Metal Alkoxides, Organometalloxanes and Organometalloxanosiloxanes" in F. G. A. Stone, W. A. G. Graham (eds.): *Inorganic Polymers,* Chap. 7, Academic Press, New York 1962.
415. I. K. Stavitskii, *Vysokomol. Soedin.* **1** (1959) 1502.
416. S. N. Borisov, N. G. Sviridova, *Vysokomol. Soedin.* **3** (1961) 50.
417. R. D. Crain, P. E. Koenig, WADC Tech. Rept. 69-427 (1960) p. 15.
418. F. A. Henglein, R. Lang, L. Schmack, *Makromol. Chem.* **22** (1957) 103.
419. Du Pont, US 2 512 058, (H. C. Gulledge).
420. T. Boyd, GB 728 751, 1955.
421. T. Boyd, US 2 716 656, 1955.
422. A. Hancock, R. Sidlow, *J. Oil Colour Chem. Assoc.* **35** (1952) 28.
423. W. H. Post, W. T. Schwartz, Jr., AD 255 545, U.S. Dept. Commerce Office Tech. Serv., Washington D.C., 1961.
424. K. A. Andrianov, S. V. Pichkhadze, *Vysokomol. Soedin.* **3** (1961) 577.
425. C. Ungurenasu, I. Haidec, *Rev. Roumaine Chim,* **13** (1968) 957.
426. C. C. Currie, US 2 672 455, 1954.
427. Monsanto Inc., GB 723 989, 1955.
428. R. M. Kary, US 2 646 440, 1953.
429. W. P. Gill et al., *Phys. Rev. Lett.* **38** (1977) 1305.
430. F. B. Bundy, K. T. Dunn, *J. Chem. Phys.* **71** (1979) 1550.
431. P. W. Bridgman, *Proc. Am. Acad. Arts Sci.* **74** (1941) 399.
432. E. Whalley, *Can. J. Chem.* **38** (1960) 2105.
433. Y. Okamoto, Z. Iqbal, R. H. Baughman, *J. Macromol. Sci. Chem.* **A25** (1988) no. 5 – 7, 799.
434. W. S. Chan, A. K. Jonsher, *Phys. Status Solidi* **32** (1969) 749.
435. J. Tsukamoto, A. Takahashi, *Jpn. J. Appl. Phys.* **25** (1986) L338.
436. M. Berthelot, *Ann. Chim. Phys.* **11** (1936) 15.
437. Y. Asano, *Jpn. J. Appl. Phys.* **22** (1983) 1618.
438. L. A. Wall, D. W. Brown, *J. Polym. Sci. Part C* **4** (1964) 1151.
439. Z. Iqbal et al., *J. Chem. Phys.* **85** (1986) 4019.
440. G. Grimm, H. Metzger, *Ber. Dtsch. Chem. Ges.* **69** (1936) 1356.
441. A. J. Brown, E. Whalley, *Inorg. Chem.* **7** (1968) 1254.
442. Y. Okamoto, P. S. Wojciechowski, *J. Chem. Soc. Chem. Commun.* 1982, 386.
443. H. Kobayashi et al., *Chem. Lett.* 1983, 1407.
444. H. Kobayashi, A. Kobayashi, Y. Sasaki, *Mol. Cryst. Liq. Cryst.* **118** (1985) 427.
445. A. Kobayashi et al., *Bull. Chem. Soc. Jpn.* **59** (1986) 3821.
446. H. Emeleus, J. Anderson: *Modern Aspects of Inorganic Chemistry,* 3rd ed., Routledge & Kegan, London 1960.
447. F. Cataldo, *J. Inorg. Organomet. Polym.* **7** (1997) no. 1, 35.
448. Y. L. Gefter: *Organophosphorus Monomers and Polymers,* Pergamon Press, London 1962.
449. H. Noth, W. Schragle, *Z. Naturforsch. B* **166** (1961) 473.
450. C. U. Pittman, Jr., *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 371.
451. C. Shimasaki, H. Kitano, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 7,** CRC Press, Boca Raton, FL 1996.
452. *Phosphorus, Sulfur, Silicon, Relat. Elem.,* ongoing journal.
453. D. Corbridge: "Studies in Inorganic Chemistry 10: Phosphorus, An Outline of Its Chemistry," *Biochemistry and Technology,* 4th ed., Elsevier, Amsterdam 1990.
454. G. Borissov, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 7,** CRC Press, Boca Raton, FL 1996, p. 5092.
455. D. Liaw, D. Wang, in J. Salamone (ed.): *Polymeric Materials Encyclopedia,* **vol. 9,** CRC Press, Boca Raton, FL 1996, p. 6569.
456. E. W. Neuse, G. J. Chris, *J. Macromol. Sci.* A-1 (1967) 371.
457. H. R. Allcock: "Rings, Clusters and Polymers of the Main Group Elements," *ACS Symp. Ser.* **232** (1983).

Insect Control

Acaricides is a separate keyword

ROBERT L. METCALF, Department of Entomology, University of Illinois, Urbana, Illinois 61801, United States

1. Insects as Enemies of Man 3
1.1. Development of Insecticides 4
1.2. Insect Resistance to Insecticides . 4
1.2.1. Cross-Resistance and Multiple Resistance 5
1.2.2. Persistence of Genes for Insecticide Resistance 6
1.3. Resurgence and Secondary Pests 7
1.4. Insecticides and Environmental Quality 7
1.5. Integrated Pest Management . . . 8
2. Inorganic Insecticides 9
2.1. Arsenicals 9
2.2. Fluorides 10
3. Natural Product Insecticides . . . 10
3.1. Nicotinoids 11
3.2. Rotenoids 11
3.3. Veratrine Alkaloids 12
3.4. Ryanodine 12
3.5. Pyrethroid Insecticides 12
3.5.1. Synthetic Pyrethroids 13
3.5.1.1. Chrysanthemate Esters 13
3.5.1.2. Chrysanthemate Analogues 15
3.5.1.3. Phenylacetate Esters 16
3.5.2. Mode of Action 17
3.5.3. Environmental Toxicology 17
3.5.4. Pyrethroid Synergists 17
4. Dinitrophenols 18
5. Organochlorine Insecticides 19
5.1. DDT and Derivatives 19
5.1.1. Mode of Action 20
5.1.2. Environmental Toxicology 20
5.2. Hexachlorocyclohexanes 21
5.3. Cyclodienes 21
5.4. Chlorinated Camphenes 25
6. Organophosphorus Insecticides . 25
6.1. Phosphoric Acid Anhydrides . . . 26
6.2. Vinyl Phosphates 27
6.3. Aliphatic Phosphorothioate Esters 28
6.4. Phosphorothioate Esters of Phenols 31
6.5. Phosphonothioate Esters of Phenols 33
6.6. Phosphorothioate Esters of Heterocyclic Enols 33
6.7. Phosphorothioate Esters of S-Methyl Heterocycles 34
6.8. Miscellaneous Organophosphorus Esters 35
6.9. Mode of Action and Toxicity . . . 36
6.9.1. Inhibition of Acetylcholinesterase . 36
6.9.2. Delayed Neuropathy 38
6.9.3. Selective Toxicity 38
6.10. Environmental Toxicology 39
7. Carbamate Insecticides 39
7.1. Dimethylcarbamates of Heterocyclic Enols 40
7.2. Methylcarbamates of Phenols . . 40
7.3. Methylcarbamates of Oximes . . . 43
7.4. Mode of Action 43
7.5. Environmental Toxicology 44
8. Formamidines 44
9. Insect Growth Regulators 44
9.1. Juvenoids 44
9.2. Chitin Synthesis Inhibitors 45
10. Other Types of Insecticides 45
11. Fumigants 46
12. Microbial Insecticides 47
13. Genetic Control 48
13.1. Sterile Male Release 48
13.2. Chemosterilants 49
13.3. Host Plant Resistance 49
14. Repellents 50
14.1. Repellents for Plant-Feeding Insects 50
14.2. Repellents for Blood-Feeding Insects 50
14.3. Repellents for Wood-Feeding Insects 52
14.4. Repellents for Fabric-Eating Pests 52
15. Attractants 53
15.1. Sex Pheromone Attractants 53
15.2. Aggregation Pheromones 55
15.3. Kairomone Attractants 55
16. Insecticide Formulation 57
17. Insecticide Residues in Foods . . . 57
17.1. Residue Persistence 57
17.2. Environmental Chemistry of Insecticide Residues 58
17.2.1. Inactivation 58
17.2.2. Activation 62

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_263

17.3. Toxicology of Insecticide Residues 62
18. References 63

1. Insects as Enemies of Man [1]

Insects attack humans and domestic animals; act as vectors for human, animal, and plant diseases; destroy structures (man-made buildings, etc.); and compete for supplies of food and fiber. The ca. 10^6 insect species that have been described represent about 70 % of all animal species. At least 600 insect species are important pests in the United States; the world total exceeds 10 000 species [1].

Worldwide crop losses caused by insect pests have been estimated as wheat 5.0 %, rice 26.7 %, maize 12.4 %, potatoes 6.5 %, sugar beet and cane 16.5 %, vegetables 8.7 %, fruit 5.8 %, oil crops 11.5 %, and fiber and rubber crops 14.2 % [2]; see also → Crop Protection. Total production loss is estimated at 13.8 % or $\$ 29.7 \times 10^9$. In the United States, total losses from insect attacks calculated in 1988 dollars are estimated at $\$ 14 \times 10^9$ [1].

Losses resulting from the attacks of insect vectors of human and animal disease are enormous. Some 60 species of *Anopheles* mosquitoes act as vectors of the plasmodia that cause human malaria; worldwide, they are responsible for as many as 200 million clinical cases and a million deaths annually. At least 250 million people suffer from lymphatic filariasis caused by the nematode parasites *Wuchereria bancrofti* and *Brugia malayi*. This disease is transmitted largely by the common house mosquito *Culex fatigans* that breeds in polluted water. Trachoma, a viral disease spread by the housefly *Musca domestica* is the most common cause of human blindness, infecting an estimated 40 million victims. American trypanosomiasis (Chagas' disease) is caused by *Trypanosoma cruzi* and transmitted to as many as 40 million humans by about 95 species of blood-sucking Triatominae bugs, with a variety of animal reservoirs. The dozen or so species of tsetse flies (*Glossina* spp.) are vectors of trypanosomiasis in Central Africa. Blackflies of the *Simulium damnosum* complex are vectors of the filarian *Onchocerca volvulus* that infects about a million people in the Volta River Basin of Africa. The parasite may attack the eye; about 70 000 of its victims are considered blind. Other important human diseases transmitted by insects and related arthropod vectors include plague (causative organism *Yersinia pestis*) transmitted by the asiatic rat flea *Xenopsylla cheopis* and other fleas, endemic typhus (*Rickettsia prowazeki*) spread by the human body louse *Pediculus humanus humanus*, and scrub typhus (*R. orientalis*) spread by *Trombicula* mites [3].

Arthropods are particularly adept at transmitting viral disease (arborviruses); insects, mites, and ticks are vectors for at least 450 viruses. The most feared are yellow fever, dengue, and dengue hemorrhagic fevers (transmitted largely by the domestic mosquito *Aedes aegypti*), as well as the human encephalitides such as St. Louis, western and eastern equine, and Venezuelan encephalitis for which the mosquitoes *Culex tarsalis* and *C. pipiens* are important vectors [4].

1.1. Development of Insecticides [5–13]

Humans have been striving to ameliorate the enormous depredations caused by insects since the dawn of history. Chinese writings of the Shang Dynasty (1523 – 1027 B.C.) describe organized efforts to use fire to destroy plagues of the migratory locust *Locusta migratoria manilensis*. However, the causal agents of malaria, *Plasmodium* spp., and the role of *Anopheles* mosquitoes in their transmission have been recognized only since 1871.

Plant-derived insecticides (e.g., rotenone, veratridines, pyrethrins, and nicotine) have been used for insect control since antiquity. Major developments in the use of chemicals for plant protection began with Paris green or copper acetoarsenite in 1865, used as a stomach poison to control the Colorado potato beetle *Leptinotarsa decemlineata*. Lead arsenate, introduced in 1892, and calcium arsenate, introduced in 1907, became the most important insecticides during the early twentieth century; production reached a maximum of 41.1×10^6 kg of lead arsenate in

1944 and 38.1×10^6 kg of calcium arsenate in 1942 [11]. Much of this was applied to cotton to control the boll weevil *Anthonomus grandis* and the pink bollworm *Pectinophora gossypiella*. Arsenicals had the pronounced disadvantages of high toxicity to humans and domestic animals, considerable phytotoxicity, and extreme environmental persistence [6], [11]. Cryolite (sodium fluoroaluminate) was introduced as a stomach poison insecticide in 1928 because of widespread concern about toxic arsenical residues on fruit and vegetables. These stomach poison insecticides were effective only against chewing insects and had little contact action.

Although potassium dinitro-*o*-cresylate, the first synthetic organic insecticide, was introduced in Germany in 1892 as a dormant spray for tree fruit insects, synthetic organic insecticides had little success until the patenting of 1,1,1-trichloro-2,2-bis(4-chlorophenyl)ethane in 1939 [14]. This insecticide was enormously successful in vector control of malaria, typhus, and other human diseases during World War II, and widespread publicity attending its civilian debut in 1945 led to the synthesis and evaluation of hundreds of thousands of synthetic organic insecticides. With its efficient contact action, DDT became the mainstay of vector control programs and rapidly replaced the arsenicals in crop protection. It became the standard remedy for control of the codling moth *Laspeyresia pomonella*, which attacked deciduous fruit, and for the pink bollworm *P. gossypiella*, which attacked cotton. For the first time, the use of DDT offered prospects for chemical control of the two great defoliators of North American forests, the gypsy moth *Porthetria dispar* and the spruce budworm *Choristoneura fumiferana*. The name DDT became a household word for controlling insect pests in flower and vegetable gardens and for mothproofing woolens. The success of DDT residual spray operations, which eliminated malaria in the Latina province of Italy and on Sardinia within four years, inaugurated the malaria eradication campaign of the World Health Organization in 1955. The plan was to eliminate malaria by spraying, at a level of 1 g/m^2, all human dwellings where malaria transmission occurred [15]. The enormous success of DDT in insect control stimulated the research and development of other synthetic insecticides. Large-scale use of other organochlorine compounds followed (e.g., lindane; toxaphene; and the cyclodienes chlordane, heptachlor, aldrin, dieldrin, and endrin).

1.2. Insect Resistance to Insecticides [16–18]

The ability of insects to rapidly develop resistance to the lethal action of insecticides has become a major factor in the control of insect pests. Insecticide resistance depends on random mutation that establishes an R-allele (a gene coding for resistance) in the natural population of the insect species. The mutant individual possesses biochemical, physiological, or even behavioral characteristics that confer some degree of immunity to the insecticide. Widespread application of the insecticide propagates the R-allele through preferential survival, and it is dispersed throughout the population. As the R-allele becomes sufficiently common, insecticide effectiveness is reduced. If the R-allele is partially dominant, completely resistant RR homozygotes are selected rapidly. Thus insecticide resistance has been described as accelerated microevolution. These R-alleles are thought to be present in typical insect populations at frequencies of $10^{-4} - 10^{-2}$, with RR homozygotes present at $10^{-8} - 10^{-4}$. With the resistance of *Anopheles gambiae* mosquito to dieldrin used for malaria eradication, R-alleles were present in African populations at frequencies of 0.04 – 6.0 %, and a single spray treatment produced a population of 86 % RS and RR genotypes, with virtual immunity [16].

The first description of insecticide resistance in 1914 was that of the San Jose scale *Quadraspidiotus perniciosus* to lime sulfur spray. In 1916, resistance to hydrogen cyanide fumigant was reported in the California red scale *Aonidiella aurantii* and the black scale *Saissetia oleae* which attacked citrus. The number of scientifically validated examples of insecticide resistance increased gradually, and by 1946 resistance was known in eleven species including the codling moth *Laspeyresia pomonella* and the peach twig borer *Anarsia lineatella* to lead arsenate, the citrus thrips *Scirtothrips citri* to potassium antimonyl tartrate (tartar emetic), the walnut husk fly *Rhagoletis completa* to cryolite, and the cattle ticks *Boophilus microplus* and *B. de-*

Table 1. Rates of development of insect resistance to insecticides

Insecticide	Number of resistant species								
	1938	1948	1954	1970	1975	1980	1984	1989	Average doubling time, years
DDT, methoxychlor	0	3	13	98	203	229	233	263	6.3
Lindane, cyclodienes	0	1	5	140	225	269	276	291	5.0
Organophosphorus compounds	0	0	3	54	147	200	212	260	4.0
Carbamates	0	0	0	3	36	51	64	85	2.5
Pyrethroids	0	0	0	3	6	22	32	48	1 – 2
Total	7	14	25	224	364	428	447	505	8.3

coloratus to sodium arsenite dip. However, little scientific attention was given to insecticide resistance until the introduction of DDT in 1945, with resistant strains of the housefly *Musca domestica* appearing almost simultaneously in Sweden and Denmark in 1946, of the mosquitoes *Culex tarsalis* in Italy and *Aedes sollicitans* in Florida in 1947, of the bedbug *Cimex lectularius* in Hawaii in 1947, and of the human body louse *Pediculus humanus humanus* in Korea and Japan in 1951 [16].

Burgeoning use of insecticides after World War II resulted in an exponential increase in the number of scientifically documented examples of resistance, which reached 224 species in 1970, 364 in 1975, 447 in 1984, and 505 in 1988 (Table 1). Most of the early examples of resistance were found in insect vectors of human disease because of the widespread use of DDT, lindane, and dieldrin in vector control programs. By 1970, resistance was documented in 119 pests of crops, forests, and stored products, compared to 166 pests of human and animal health; in 1980 the respective figures were 283 agricultural pests compared to 198 pests of humans and animals [18].

As each new class of insecticides was introduced and used widely, the rate of development of resistant species followed an almost identical pattern of exponential growth until high-level resistance or legal restrictions slowed further usage (Table 1). Doubling time for the development of resistant species has decreased steadily with the introduction of each new class of insecticide (i.e., DDT and methoxychlor, lindane and cyclodienes, organophosphates, carbamates, and pyrethroids). This accelerated development of multiple resistance is a result of the persistence of R genes in the species genome and of their interaction through a variety of resistance mechanisms that affect both the detoxification of and the target-site insensitivity to various insecticides.

1.2.1. Cross-Resistance and Multiple Resistance

The exponential growth in numbers of insect species resistant to insecticides does not adequately describe the impact of resistance on insect control problems. Two general types of resistance are common: cross-resistance and multiple resistance.

Cross-resistance enables resistant species to survive exposure to chemically related insecticides, such as DDT and methoxychlor, lindane and heptachlor, parathion and diazinon, carbaryl and carbofuran, or permethrin and fenvalerate. Cross-resistance results from a common detoxification mechanism or a change in susceptibility to a common biochemical lesion.

Multiple resistance enables resistant species to survive exposure to several classes of insecticide with differing detoxification systems and biochemical modes of action. Multiple resistance results from interactions among several types of evolutionarily developed survival mechanisms so that entire categories of insecticides become useless. Examples include the altered acetylcholinesterase target site that produces multiple resistance to organophosphorus and carbamate insecticides, which is found in the housefly, the green rice leafhopper *Nephotettix cincticeps*, the cattle tick *Boophilus microplus*, and the red spider mite *Tetranychus urticae*. Another type of multiple resistance is the

"knockdown" mechanism of nerve axon insensitivity that promotes resistance to both DDT and the pyrethroids, and has been demonstrated in the housefly; a variety of mosquitoes; bedbugs of *Cimex* spp.; lepidopteran pests including *Spodoptera, Plutella,* and *Heliothis*; and the cattle tick *B. microplus*. Multiple resistance is now present in at least 44 families of insects in 10 orders [19]. An increasing number of major insect pests are resistant to all the principal classes of insecticides (Table 1). These include the housefly; the malaria vectors *Anopheles albimanus, A. sacharovi,* and *A. stephensi*; the German cockroach *Blattella germanica*; the Colorado potato beetle *Leptinotarsa decemlineata*; the green peach aphid *Myzus persicae*; the pear psylla *Psylla pyricola*; the diamondback moth *Plutella xylostella*; the cotton bollworms *Heliothis* spp. and *Spodoptera* spp.; the red flour beetle *Tribolium castaneum*; and the granary weevil *Sitophilus granarius*.

In summary, cross-resistance limits the choice of available insecticides, whereas multiple resistance reflects the past history of insecticide use and precludes a return to those used previously. Both types of insecticide resistance severely deplete insecticide resources.

1.2.2. Persistence of Genes for Insecticide Resistance

Gene frequency of a specific resistance allele may decrease upon removal of insecticide pressure; however, a changed background of residual inheritance persists and causes resistant strains to regain high-level resistance as soon as the insecticide is reapplied. Thus genes for DDT and cyclodiene resistance in Danish houseflies have persisted for more than 20 years, and the insecticides became ineffective within two months after reapplication. The citrus thrips *Scirtothrips citri* has retained its resistance to tartar emetic bait for more than 45 years and to DDT for more than 35 years. No signs of regression have appeared in the multiple resistance patterns of the Colorado potato beetle and the cotton bollworm *Heliothis virescens* over as much as 30 years.

The combined effects of cross- and multiple resistances and the persistence of resistant genes have had catastrophic effects on many insect control programs. At least 51 of the 60 important *Anopheles* mosquito vectors of malaria are resistant to one or more of the few residual insecticides that can be used safely in human habitations. This resistance has so increased the cost and decreased the effectiveness of malaria control that the World Health Organization abandoned its widely heralded program for global eradication of malaria in 1976 [20]. Important insect pests of man, such as the housefly and the German cockroach, can scarcely be controlled by insecticides. Practical chemical control of agricultural pests such as the Colorado potato beetle, the diamondback, and the *Heliothis* cotton bollworms is in jeopardy. Resistance has severely affected the economy of insect control in agriculture and public health by necessitating replacement of the relatively inexpensive organochlorine insecticides with the increasingly expensive organophosphates, carbamates, and pyrethroids. Adverse effects on human health followed the withdrawal of relatively safe organochlorines, such as DDT and lindane, and their replacement by highly toxic organophosphates, such as parathion and methyl parathion. Insecticide resistance is a major factor in efforts to replace programs for the chemical control of insects by more ecologically oriented programs for insect pest management.

1.3. Resurgence and Secondary Pests

[20], [21]

The overuse and injudicious use of broad-spectrum insecticides has upset the biological balance between insect pests and their natural enemies so that dramatic population increases (resurgence) follow. This resurgence is usually the result of repeated applications of insecticides that select resistant races of the pest and coincides with depletion or virtual elimination of their important natural enemies. The associated development of secondary pests results from resistance and resurgence of nontarget insect herbivores which then become so numerous that they attain pest status. Destructive resurgence has become a major factor in the development of ecologically oriented programs for integrated pest management.

Proliferation of the use of the broad-spectrum insecticides DDT, parathion, carbaryl, and related compounds in 1946 was responsible for the

enormous outbreak of redspider mites (Tetranychidae) in deciduous and citrus fruit orchards and on cotton and other row crops. These mites were rarely abundant enough before 1946 to require pesticide application. The sudden appearance in 1947 of the new words miticide and acaricide for special pesticides to control these mite pests provided dramatic proof of the impact of broad-spectrum insecticides on the balance between phytophagous mites and their predators. The imbalances created were all the more dramatic because of the very short generation time of the mites and the rapidity with which they developed pesticide resistance.

Today, the use of broad-spectrum pesticides has almost inevitably been followed by pest resistance, pest resurgence, and the outbreak of secondary pests. These occur most dramatically on crops that are treated heavily with pesticides, such as cotton, rice, and tree fruit. Invariably, ecological imbalance has been related to the suppression of key natural enemies. Cotton culture worldwide has been almost destroyed by the resurgence of *Heliothis* spp. and *Spodoptera* spp. bollworms that became devastating secondary pests when their predators *Chrysopa, Orius, Geocoris,* and *Nabis* spp. were destroyed. Attempts to control cotton bollworms with more frequent applications of insecticides led to the destruction of populations of the *Encarsia* and *Eretmocerus* parasites of the whitefly *Bemisia tabaci*. This secondary pest, which transmits a variety of plant diseases, has become a critical factor in profitable cotton production in several parts of the world.

The green rice leafhopper *Nephotettix cincticeps* and related leafhoppers are vectors of rice diseases and have emerged as important secondary pests of rice as a result of the very heavy use of benzene hexachloride (BHC), parathion, and other broad-spectrum insecticides to control the rice stem borers *Chilo suppressalis* and *Tryporyza incertulas*. Insecticide application destroyed predatory spiders that suppressed leafhopper populations.

1.4. Insecticides and Environmental Quality [22–27]

Insecticides have been highly controversial because of their widespread and persistent effects on the environment. This concern results from their intrinsic toxicity to animals and humans, as well as their purposeful application to fields, orchards, forests, ponds and streams, and human habitations. Insecticide application is generally inefficent: much less than 0.1 % of the amount applied is required to kill pests; the remainder (> 99.9 %) is distributed throughout the ecosystem.

With rapidly increasing use, environmental contamination from insecticides has become massive. In 1976, ca. 73.5×10^6 kg of active ingredients of insecticides was applied to 30.335×10^6 ha of United States cropland; average contamination was about 2.4 kg/ha [28]. For cotton, the most heavily treated crop, 29.5×10^6 kg of active ingredient was applied to 2.835×10^6 ha, an average of 10.3 kg/ha. Early insecticides such as the *arsenicals* and *nicotine* were highly toxic, and lead arsenate persisted and accumulated in orchard soil to levels that killed young trees.

The *organochlorine insecticides* had the greatest impact on environmental quality because of their relatively long persistence, their very high lipid – water partition coefficients, and their relative resistance to biological degradation. Estimated average soil half-lifes in years are for DDT 2.8, dieldrin 2.5, endrin 2.2, lindane 1.2, chlordane 1.0, and heptachlor 0.8; in some soil types, persistence is much longer. These insecticides have very low water solubility (in the ppb-ppm range) and have octanol – water partition values of $10^3 - 10^6$.

They accumulate in living organisms (bioconcentration), and trace levels in water or food increase by a factor of $10^3 - 10^6$ in fish or animals at the top level of the food chain. During their maximum use in 1950 – 1970, organochlorine insecticides became ubiquitous micropollutants of the environment (see Chap. 17). Their perturbing effects on wildlife and humans have been major factors in their stringent regulation or prohibition in the United State, Canada, and both Eastern and Western Europe: e.g., U.S. government withdrawal of registration for DDT (1973), aldrin and dieldrin (1974), heptachlor and chlordane (1976), chlordecone (1980), endrin (1981), mirex (1982), and toxaphene (1983). The average levels of organochlorine insecticide residues in water, soil, and human tissue have begun to decline, and threatened

populations of birds of prey have been restored. However, worldwide use of these compounds between 1946 and 1976 was so extensive (DDT $> 2.3 \times 10^6$ t; toxaphene 4.5×10^5 t; cyclodienes including aldrin, dieldrin, heptachlor, chlordane, and endrin $> 2.7 \times 10^5$ t) that traces will remain in the environment for many years.

Environmental pollution resulting from overuse of organochlorine insecticides focused attention on the importance of biodegradability. *Organophosphate insecticides* react in the presence of light, air, and moisture, and are readily degraded in living organisms by oxidation and hydrolysis. Therefore, the use of these insecticides has not led to serious problems of bioaccumulation and food chain transfer. Their soil persistence is low, with half-lifes ranging from 1 – 2 weeks for malathion, to 3 – 6 months for parathion, diazinon, azinphosmethyl, and chlorpyrifos.

However, organophosphates are highly toxic to almost all animals including humans because they irreversibly inhibit the enzyme acetylcholinesterase. Acetylcholinesterase is required to break down the neurotransmitter acetylcholine after it has been released from the nerve ending. Unlike most organochlorine insecticides, many organophosphate insecticides are almost as toxic upon dermal exposure as upon ingestion. Furthermore, chronic exposure can progressively deplete vital acetylcholinesterase levels to a point where slight additional exposure can result in severe illness or death from cholinergic crises. Human intoxication resulting from the use of the organophosphates is intensified by insecticide resistance, which necessitates increased dosage and more frequent application. The incidence of human intoxication after use of organophosphorus insecticides has become acute; the World Health Organization has estimated that 500 000 hospitalizations and as many as 20 000 deaths occur annually worldwide [29].

Introduction of *carbamate insecticides* provided some amelioration of both environmental persistence and acute toxic hazards. Carbamates are generally biodegradable and do not bioaccumulate. Soil persistence is low, with half-lifes of 1 – 2 weeks for carbaryl and 1 – 4 months for carbofuran and aldicarb. Many carbamates are highly toxic to humans and animals, but acetylcholinesterase inhibition is readily reversible so that severe intoxication is unlikely.

Pyrethroids have the advantage of extraordinary effectiveness against insect pests, so that the required dosages are only about one-tenth to one-twentieth those of organochlorines, organophosphates, and carbamates. However, the newer pyrethroids are more persistent and lipophilic; they are extremely toxic to beneficial insects and fish. Furthermore, unlike organophosphates and carbamates, which have atropine as an antidote, no effective antidote exists for pyrethroid poisoning.

1.5. Integrated Pest Management [30–36]

The discovery of DDT and subsequent development of hundreds of organochlorine, organophosphate, carbamate, and pyrethroid insecticides produced a euphoria about chemical pest control that led to the pest eradication philosophy prevalent during the 1950s and 1960s. Widespread application of cheap, effective insecticides was supposed to provide a rapid solution to every insect pest control problem. The World Health Organization began the global eradication of malaria through annual residual (long-lasting) house spraying directed at *Anopheles* mosquito vectors. The United States undertook the insecticide eradication of pests such as the gypsy moth *Porthetria dispar*, the Japanese beetle *Popillia japonica*, the yellow fever mosquito *Aedes aegypti*, and the red imported fire ant *Solenopsis invicta*. As each eradication program faltered because of insecticide resistance, destruction of wildlife, and widespread environmental pollution, the initial reaction was to develop another insecticide of different chemistry.

The publication of Rachel Carson's *Silent Spring* in 1962 [37] inaugurated public challenge to the notion that all insect control problems could be solved solely with insecticides. Surveys demonstrated the continuing, unabated damage resulting from insect pest attack despite a greater than 10-fold increase in insecticide use and a 20-fold proliferation in chemicals available for insect control (Table 2). Unbridled use of broad-spectrum insecticides over the past 40 years had obviously not solved the manifold problems of insect pest control. The benefits to cost ratio of pest control activities to the farmer and to public health have steadily decreased and

Table 2. Estimates of U.S. crop losses from insect attacks over 60 years ∗

Crop	Loss of total crop, %			
	1900 – 1904	1910 – 1935	1942 – 1951	1951 – 1960
Alfalfa	10	5	9	15
Apple	20	10	14	13
Cabbage	10	20	8	17
Corn	8	12	4	12
Cotton	10	15	15	19
Potato	10	22	16	4

∗ USDA data [20].

the nontarget effects of pesticides on environmental quality have continued to multiply.

Pest control must be directed away from exclusive reliance on toxic chemicals and toward the optimization of pest control tactics in an economically and ecologically sound manner, i.e., *integrated pest management* (IPM). A fundamental premise of IPM is to relegate pesticides to emergency use and spray only when necessary.

The primary goals of IPM are [33]

1) to determine how the life system of the pest must be modified to reduce its numbers to tolerable levels (i.e., below the economic injury level);
2) to apply biological knowledge and current technology to achieve the desired modifications (i.e., applied ecology); and
3) to devise procedures for pest control compatible with economic and environmental control aspects (i.e., economic and social acceptance).

Successful IPM programs integrate a variety of suppressive measures including both density-dependent factors (natural enemy predators, parasitoids, and pathogens) and density-independent factors (host plant resistance and cultural practices that deter pest attack). The basics of IPM practice are

1) to establish sound economic injury levels in which the cost of treatment balances the value of pest damage;
2) to monitor regularly to determine pest population levels; and
3) to implement control practices promptly when the economic injury level is approached.

Insecticides are important components of IPM programs; indeed, IPM is a dynamic strategy for the judicious and selective use of insecticides. *Insecticide management* is concerned with the safe, efficient, and economical handling of insecticides during manufacture, utilization, and disposal [38–40]. It involves selection of the proper insecticide for the IPM program; selection of the mode, timing, and dosage of application; and selection with regard to resistance and resurgence. The nature of insecticide residues in food and in the environment, and their impact on humans, must be considered.

2. Inorganic Insecticides

2.1. Arsenicals [6], [11]

The first synthetic insecticides were arsenical salts used as stomach poison insecticides to control chewing insects such as the Colorado potato beetle *Leptinotarsa decemlineata* and the cotton boll weevil *Anthonomus grandis.*

Paris green, copper acetoarsenite [*12002-03-8*], $[Cu(CH_3COO)]_2 \cdot 3\,Cu(AsO_2)_2$, is a bright-green pigment, used originally as a wall and shutter paint. Its insecticidal properties were discovered around 1865. Paris green contains 44.3 % arsenic, with 2 – 3 % of it water-soluble; the LD_{50} (rat, oral) is 100 mg/kg.

Lead arsenate was developed in 1892 to combat the gypsy moth *Porthetria dispar.* It is a safer arsenical for plant use because of its lower content of water-soluble arsenic. Acid lead arsenate [*7784-40-9*], $PbHAsO_4$, contains 21.6 % arsenic with 0.25 % of it water-soluble. The LD_{50} (rat, oral) is 1050 mg/kg.

Basic lead arsenate [*62648-18-4*], $Pb_4(PbOH)(AsO_4)_3$, with 14 % arsenic, was introduced as a safer form to use on delicate foliage in foggy or humid regions where acid lead arsenate is hydrolyzed to liberate arsenic acid.

Calcium arsenate [*7778-44-1*], $Ca_3(AsO_4)_2$, was introduced in 1920 to control the boll weevil on cotton. It contains 37.6 % arsenic and has an LD_{50} (rat, oral) of 298 mg/kg. Basic calcium arsenate [*74966-85-1*], $[Ca_3(AsO_4)_2]_3 \cdot Ca(OH)_2$, contains 25 % arsenic with ca. 0.4 – 0.5 % of it water-soluble. It is a safer form that does not hydrolyze readily.

Arsenic trioxide [*1327-53-3*], As_2O_3, with 75.7 % arsenic has an LD_{50} (rat, oral) of 14.1 mg/kg. Sodium arsenite [*7784-46-5*], $NaAsO_2$, with 55.7 % arsenic has an LD_{50} (rat, oral) of 41 mg/kg. These two compounds have been used as poisons in bran bait to kill migrating grasshoppers and in water dips to control cattle ticks.

Mode of Action. The As^{3+} ion of arsenicals specifically inhibits the sulfhydryl-containing enzymes involved in cell respiration, e.g., pyruvate oxidase (E.C. 1.2.3.3.).

2.2. Fluorides [6]

Cryolite [*15096-52-3*], sodium fluoroaluminate, Na_3AlF_6, 54.2 % fluorine, water-soluble fluorine 0.035 – 0.061 %, was developed as a safer stomach poison insecticide than the arsenicals, especially because of concern about persistent lead arsenate residues on edible produce such as apples. The LD_{50} (rat, oral) of cryolite has been given as 200 – 13 500 mg/kg.

Sodium fluorosilicate [*16893-85-9*], Na_2SiF_6, 60.6 % fluorine, water-soluble fluorine 0.65 %, LD_{50} (rat, oral) 125 mg/kg, is used for mothproofing.

Barium fluorosilicate [*17125-80-3*], $BaSiF_6$, 40.8 % fluorine, water-soluble fluorine 0.026 %, LD_{50} (rat, oral) 175 mg/kg, has been used as a stomach poison for plant-feeding insects.

Sodium fluoride [*7681-49-4*], NaF, 45.2 % fluorine, water-soluble fluorine 4.3 %, LD_{50} (rat, oral) 180 mg/kg, is used as a stomach poison for cockroaches, for chewing lice, and in poison baits.

Mode of Action. Inorganic fluorides complex magnesium ions as magnesium fluorophosphate. They inhibit enzymes such as enolase (E.C. 4.2.1.1) that require Mg^{2+} as a prosthetic group. Thus, fluoride poisoning prevents phosphate transfer in oxidative metabolism.

3. Natural Product Insecticides

In this and the following chapters two values are often cited for the LD_{50}. The first value refers to male animals and the second to female animals.

3.1. Nicotinoids [8], [9]

Nicotine [*54-11-5*], L-1-methyl-2-(3′-pyridyl)-pyrrolidine, *bp* 247 °C, d_4^{20} 1.009, one of the earliest insecticides, was recommended in 1793 as a tea for the destruction of aphids. Nicotine is found in the leaves of *Nicotiana tabacum* and *N. rustica* in amounts of 2 – 14 %, as well as in *Duboisia hopwoodii* and *Aesclepias syriaca*. It occurs as the principal alkaloid along with small amounts of 12 other alkaloids, including the water-miscible insecticidal compounds nornicotine and anabasine. Nicotine is readily volatile (vapor pressure 5.5 Pa at 25 °C) and dibasic ($K_{b1} = 1 \times 10^{-6}$, $K_{b2} = 1 \times 10^{-11}$), easily forming salts with acids and many metals. Its LD_{50} (rat, oral) is 30 – 55 mg/kg.

Nicotine sulfate [*65-30-5*], $(C_{10}H_{14}N_2)_2 \cdot H_2SO_4$, has been used as an aphicide on fruit, vegetables, and ornamentals because it is less volatile than nicotine and safer to handle.

Nornicotine [*494-97-3*], 2-(3′-pyridyl)pyrrolidine, *bp* 270 °C, d_4^{20} 1.07, occurs naturally as both the D and the L isomer, the former in *D. hopwoodii* and the latter in *Nicotiana*.

Anabasine [*494-52-0*], L-2-(3′-pyridyl)piperidine, *bp* 281 °C, d_4^{20} 1.047, is the principal alkaloid of *Anabasis aphylla*, present at 1 – 2 % in the shoots, and is also found to about 1 % in *N. glauca*.

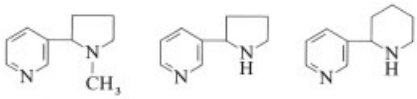

Imidacloprid [*105827–78–9*], 1-[(6-chloro-3-pyridinyl)-methyl]-*N*-nitro-2-imidazolidinime, vapor pressure 0.2 μPa at 20 °C, solubility in water 0.51 g/L, is a synthetic nicotinoid with contact and systhemic activity to aphids, leafhoppers, and white flies. Its LD_{50}values are 424 – 495 (rat, oral) and > 5000 (rat, dermal).

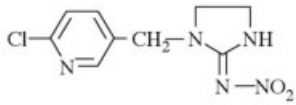

Mode of Action. The nicotine alkaloids are agonists of acetylcholine and affect the postsynaptic receptors of the insect nervous system, facilitating synaptic transmission at low concentration and blocking transmission at higher concentration. The nicotine molecule has dimensions similar to acetylcholine. Ionization of the *N*-methylpyrrole group affects penetration through the ionic barrier of the nerve sheath to the site of action; the quaternary form has an increased affinity for the negatively charged acetylcholine receptor.

3.2. Rotenoids [8], [41]

Rotenoids are found in at least 68 plant species, including *Tephrosia, Derris, Lonchocarpus, Milletia*, and *Mundulea*. The important economic species are, *D. elliptica* and *D. malaccensis* from Malaysia and Indonesia, and cube or timbo (*L. utilis* and *L. urucu*) from South America. Insecticidal compounds are found in the roots and seeds of these plants.

Rotenone [*83-79-4*], *mp* 165 °C (a dimorphic form, *mp* 185 °C) by far the most important rotenoid, is five to ten times more effective than elliptone [*478-10-4*], *mp* 159 °C (furan ring at E); sumatrol [*82-10-0*], 15-hydroxyrotenone, *mp* 188 °C; malaccol [*478-07-9*], 15-hydroxyelliptone, *mp* 244 °C; L-α-toxicarol [*82-09-7*], 15-hydroxyrotenone with a 2,2-dimethyl-2*H*-pyranyl ring in place of ring E, *mp* 101 °C; and deguelin [*522-17-8*], with a hydrogen on C-15 in place of the hydroxyl group of toxicarol, *mp* 165 – 171 °C. The content of these components in commercial sources is variable: roots of *D. elliptica* average 5 – 9 % rotenone and up to 31 % ether extractives (organic constituents extracted by diethyl ether), whereas *L. utilis* averages 8 – 11 % rotenone and 25 % extractives. When exposed to light and air, rotenone undergoes hydroxylation at C-7, followed by dehydration to form dehydrorotenone (C=C between rings B and C). These reactions inactivate rotenone after 5 – 10 d of exposure to normal sunlight. Rotenone is soluble in water only to ca. 0.2 mg/L but is readily soluble in petroleum solvents. The LD_{50} (rat, oral) is ca. 130 mg/kg.

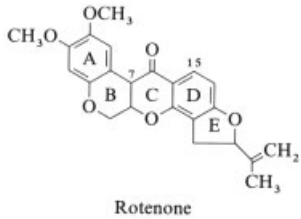

Rotenone-containing insecticides have been used as dusts for ground roots, dispersible powders, and emulsive extracts for application to edible products shortly before harvest, as well as for the control of animal ectoparasites and cattle grubs.

Mode of Action. Rotenone is a specific inhibitor of the respiratory enzyme NADH dehydrogenase (E.C. 1.6.99.3) and interferes with the electron transport necessary for oxidation of NADH (reduced nicotinamide adenine dinucleotide) to NAD by cytochrome b. Poisoning, therefore, inhibits the mitochondrial NAD-dependent oxidation of the citric acid cycle intermediates.

3.3. Veratrine Alkaloids [41]

Sabadilla, obtained from the ground seeds of *Schoenocaulon officinale* (Liliaceae) from South and Central America, has been used in native louse powders for centuries. More recently, it has been used in sugar bait sprays for thrips control. The toxic principles of sabadilla are the veratrine alkaloids which are present in the seeds to 2 – 4 % and include ca. 13 % cevadine and 10 % veratridine, as well as lesser amounts of cevadilline and sabadine.

Sabadilla preparations are activated by heating to 75 – 80 °C for 4 h, moistening with 10 % sodium carbonate solution, or milling with hydrated lime. Veratrine alkaloids are destroyed rapidly by exposure to light; sabadilla, which is both a contact and a stomach poison insecticide, is safe for use on food crops within a few days of harvest. The alkaloids most responsible for its insecticidal action are *cevadine* [*62-59-9*], $C_{32}H_{49}O_9N$, *mp* 213 °C, the angelic acid ester of cevine; and *veratridine* [*71-62-5*], $C_{36}H_{51}O_{11}N$, *mp* 180 °C, the veratric acid ester of cevine at C-3. These alkaloids are highly poisonous; the intraperitoneal LD_{50} of cevadine in mouse is 3.5 mg/kg, and it may cause severe irritation of mucous membranes and the respiratory tract. Ground sabadilla seeds have an LD_{50} (rat, oral) of 5000 mg/kg.

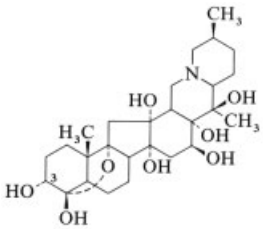

Cevine

3.4. Ryanodine [41]

Ryania, the ground roots and stems of *Ryania speciosa* (Flacourtiaceae) native to South America, contains 0.16 – 0.25 % of insecticidal components of which ryanodine is the most important. *Ryanodine* [*15662-33-6*], $C_{25}H_{33}O_9N$, *mp* 219 – 220 °C is both a contact and a stomach poison insecticide; it is more stable to the action of light and air than rotenone or pyrethrum. It is used as a residual insecticide for codling moth control, particularly in home plantings. Ryania is applied either as a wettable powder or as a methanolic extract. The LD_{50} (rat, oral) is 750 mg/kg. Ryania uncouples the adenosine 5′-triphosphate – adenosine 5′-diphosphate (ATP – ADP) actomyosin cycle in the contraction of striated muscle.

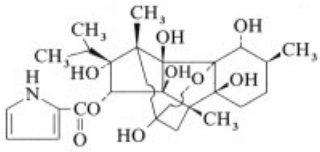

Ryanodine

3.5. Pyrethroid Insecticides [41–43]

The insecticidal properties of pyrethrum from the ground flowers of *Chrysanthemum cinerariaefolium* and *C. coccineum* (Compositae) were recognized in the Transcaucasus region of Asia around 1800. Manufacture of flea and louse powders began about 1828, and when the nature and source of the material was published in 1851, it became used worldwide. Pyrethrum flowers were the most important export of Dalmatia until World War I, when Japan became the leading producer. Pyrethrum production began in Kenya about 1932, and it was the principal supplier by 1940. World production of dried pyrethrum flowers was about 23×10^3 t in 1975. Kenya flowers contain ca. 1.3 % of the toxic constituents, pyrethrins and cinerins; Japanese flowers, ca. 1 %; and Dalmatian flowers, ca. 0.7 %. Pyrethrum concentrates are prepared by extracting the flowers with petroleum ether, ethylene dichloride, methanol, acetone, or acetic acid. The extracts are purified from waxes by reextraction with nitromethane or methanol, and absorption on activated charcoal, to produce a 25 % pyrethrum concentrate containing about 10 % pyrethrin I, 9 % pyrethrin II, 2 % cinerin I, 3 % cinerin II, 1 % jasmolin I, and 1.1 % jasmolin II (Fig. 1).

Pyrethrum is an especially valuable insecticide because of its rapid paralysis (“knockdown”) of flying insects and because its toxic principles are inactivated on exposure to light. The main use of pyrethrum insecticide is in pressurized aerosol “bombs” that contain

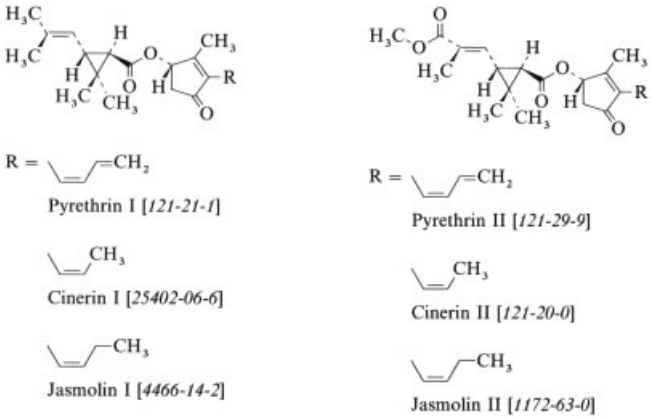

Figure 1. Structures of pyrethroids from *Chrysanthemum cinerariaefolium*

0.04 – 0.25 % active ingredient, together with five to ten times this amount of piperonyl butoxide or other synergists to retard detoxification in insects. Deodorized oil or water-based sprays are also used. The principal areas of use for pyrethrum products are

1) household insecticides for the control of mosquitoes, flies, cockroaches, bedbugs, ants, silverfish, and fabric pests;
2) livestock or cattle sprays to protect animals from annoyance by biting and nonbiting flies;
3) protective sprays for grain in storage; and
4) limited use on fruit and vegetables to control insect pests prior to harvest when toxic residues are very objectionable.

Toxic Constituents. Complete identification of the active principles of pyrethrum required approximately 50 years, dating from the work of STAUDINGER and RUSZICKA in 1924. There are six esters of the two cyclopropanecarboxylic acids, chrysanthemic acid and pyrethric acid (Fig. 1), and three cyclic keto alcohols, pyrethrolone, cinerolone, and jasmolone. The esters are optically active with absolute configurations of 1R, 3R, and 4S; the double bond in the acid side chain is *cis* and in the alcohols, *trans*.

3.5.1. Synthetic Pyrethroids [44–46]

Elucidation of the chemical structure of the naturally occurring pyrethroids stimulated the search for effective synthetic derivatives of these costly natural products. Since the early 1940s, structural optimization of pyrethroid molecules has led to a series of broad-spectrum insecticides with greatly enhanced activity and, in many cases, extended residual action. Pyrethroids have become one of the most important classes of insecticides.

3.5.1.1. Chrysanthemate Esters

Allethrin [*584-79-2*], 2-methyl-4-oxo-3-(2-propenyl)-2-cyclopenten-1-yl 2,2-dimethyl-3-(2-methyl-1-propenyl)cyclopropanecarboxylate, the first successful synthetic pyrethroid, was developed in 1949 and is the allyl homologue of cinerin I. The product is a brownish liquid, $d\,^{20}_{4}$ 1.005 – 1.015, containing 75 – 95 % of a mixture of eight optical and geometric isomers [70 % (±)-*trans* and 30 % (±)-*cis* acids] esterified with (±)-methylcyclopentenolone alcohol. Allethrin has an LD_{50} (rat, oral) of 920 mg/kg; it is as effective as natural pyrethrum against flies and mosquitoes but has a narrow spectrum of activity. Bioallethrin is the (+)-*trans*-chrysanthemate ester of the same alcohol with enhanced "knockdown" activity.

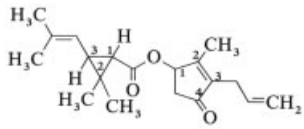

Allethrin

Benzyl Chrysanthemates. Esterification of (±)-*cis,trans*-chrysanthemic acid with substituted benzyl alcohols produced *barthrin* [*70-43-9*], the 6-chloro-1,3-benzodioxol-5-yl ester, and *dimethrin* [*70-38-2*], the 2,4-dimethylbenzyl ester. These compounds have very low mammalian toxicity (LD_{50}, rat, oral > 20 g/kg) but a limited spectrum of insecticidal activity. Dimethrin has been used as a safe mosquito larvicide for drinking water in cisterns.

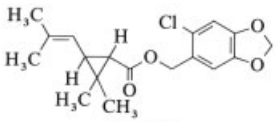
Barthrin

Resmethrin [*10453-86-8*], 5-benzyl-3-furanyl (±)-*cis,trans*-chrysanthemate, *mp* 43 – 48 °C, *bp* 169 – 172 °C (1.3 Pa), is a white waxy solid. The compound contains 20 – 30 % (±)-*cis* and 70 – 80 % (±)-trans isomers and has an LD_{50} (rat, oral) of 2 000 mg/kg. The benzylfuranyl moiety is readily decomposed by light, and resmethrin is an outstanding household insecticide, 20 – 50 times more effective than pyrethrum and of very short persistence.

Bioresmethrin [*28434-01-7*] is the corresponding (±)-trans-(1R,3R) ester, which is about twice as toxic to insects; it has an LD_{50} (rat, oral) > 8 000 mg/kg.

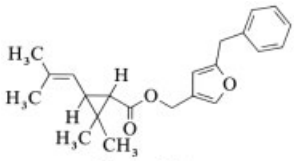
Resmethrin

Tetramethrin [*7696-12-0*], (1,3,4,5,6,7-hexahydro-1,3-dioxo-2*H*-isoindol-2-yl)methyl (±)-*cis,trans*-chrysanthemate, *mp* 65 – 80 °C, vapor pressure 6.2 μPa at 20 °C, has an LD_{50} (rat, oral) of 1 000 mg/kg and produces rapid "knockdown". It is used in veterinary hygiene.

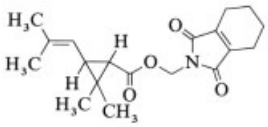
Tetramethrin

Phenothrin [*26002-80-2*], (3-phenoxyphenyl)methyl (±)-*cis,trans*-chrysanthemate, a colorless liquid, d_4^{25} 1.058, is soluble in water to 0.2 mg/L. The LD_{50} (rat, oral) is 5000 mg/kg.

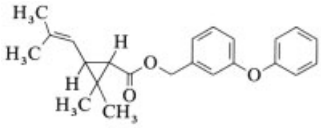
Phenothrin

3.5.1.2. Chrysanthemate Analogues

More recent development in pyrethroid chemistry has involved substantial changes in the structure of the acid moiety and esterification with a variety of optimized alcohol moieties.

Permethrin [*52645-53-1*], (3-phenoxyphenyl)methyl 3-(2,2-dichloroethenyl)-2,2-dimethyl cyclopropanecarboxylate is a mixture of 70 % (±)-*trans* and 30 % (±)-*cis* configurations, *mp* 35 °C, water solubility 0.2 mg/L, vapor pressure 45 μPa at 25 °C. The LD_{50} (rat, oral) is 1500, 2000 mg/kg. Substitution of chlorine atoms for the dimethyvinyl group greatly reduces photolability and increases persistence; thus permethrin is a widely used broad-spectrum insecticide for cotton and other field crops.

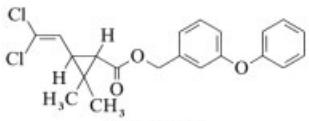
Permethrin

Cypermethrin [*52315-07-8*], α-cyano(3-phenoxyphenyl)methyl 3-(2,2-dichloroethenyl)-2,2-dimethylcyclopropanecarboxylate, *mp* 60 – 80 °C, water solubility 0.01 mg/L, is a mixture of 70 % (±)-*trans* and 30 % (±)-*cis* configurations; its LD_{50} (rat, oral)

is 310 mg/kg. Cypermethrin is widely used on agricultural crops and in cockroach control.

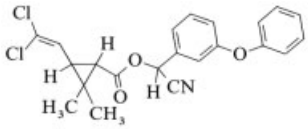
Cypermethrin

Cyfluthrin [*68359–37–5*], (*R,S*)-α-cyano-(4-fluoro-3-phenoxyphenyl)-methyl-(1*R,S*)-*cis*,*trans*-3-(2,2)-dichloroethenyl)-2,2–dimethylcyclopropanecarboxylate, *mp* 60 °C, solubility in water 2 µg/L. Cyfluthrin is a mixture of eight isomers; its LD_{50}values (rat) are 500, 800 (oral), and 600 (dermal). Cyfluthrin is a wide-spectrum pyrethroid, especially effective against greenhouse pests.

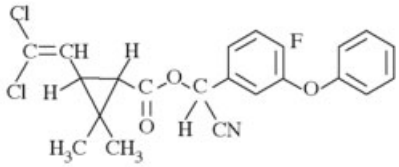
Cyfluthrin

Deltamethrin [*52918-63-5*], (*S*)-α-cyano-3-phenoxybenzyl(+)-*cis*-(1*R*,3*R*)-2,2-dimethyl-3-(2,2-dibromovinyl)cyclopropanecarboxylate, *mp* 98 – 101 °C, vapor pressure 2 μPa at 25 °C, the most active and persistent of all pyrethroid insecticides, is effective against field crop insects at doses of 100 g/ha. However, it is highly toxic to mammals; its LD_{50} (rat, oral) is 25, 60 mg/kg.

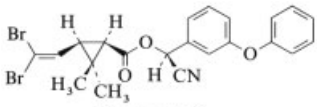
Deltamethrin

Tralomethrin [*66841-25-6*], (*S*)-α-cyano-(3-phenoxyphenyl)-methyl-(1*R*,3*S*)-2,2-diemthyl-3-(*RS*)-(1,2,2,2-tetrabromoethyl)cyclopropanecarboxylate, *d* 1.70, vapor pressure 17 mPa at 25 °C, the tetrabromo analogue of deltamethrin, is a mixture of two isomers. Solubility in water 70 mg/L. Its LD_{50} values are 1250, 1070 mg/kg (rat, oral) and > 2000 mg/kg (rabbit, dermal). Tralomethrin is used as a household, stored grain, foliar, and wood protectant insecticide.

Cyhalothrin [*68085-85-8*], (*RS*)-α-cyano-(3-phenoxyphenyl)methyl (±)-*cis*-(1*R,S*)-3-(2-chloro-3,3,3-trifluoro-1-propenyl)-2,2-dimethylcyclopropanecarboxylate, *mp* 49 °C, vapor pressure 0.2 μPa at 25 °C soluble in water (5 µg/L); its LD_{50} (rat, oral) is 79, 56 mg/kg. Cyhalothrin is a persistent pyrethroid used on cotton and other field crops, which has acaricidal activity.

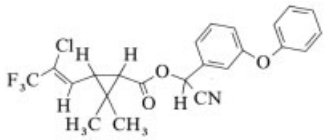
Cyhalothrin

Kadethrin [*58769-20-3*], 5-(benzyl-3-furanylmethyl) 3-[(dihydro-2-oxo-3(2*H*)-thienylidene)-methyl]-*cis*-(1*R*,3*S*)-2,2-dimethylcyclopropanecarboxylate, is a pyrethroid of very rapid "knockdown". The LD_{50} (rat, oral) is 140, 1300 mg/kg.

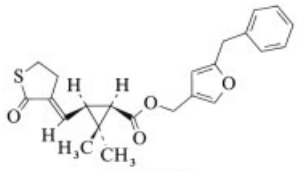
Kadethrin

Bioethanomethrin [*22431-62-5*], 5-(benzyl-3-furanylmethyl)3-(cyclopentyl-idenemethyl)-(+)-*trans*-(1*R*,3*S*)-2,2-dimethyl-cyclopropanecarboxylate. The LD_{50} (rat, oral) is 100 mg/kg.

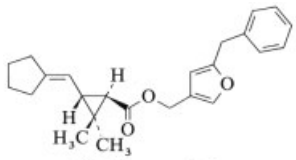
Bioethanomethrin

Fenpropathrin [*39515-41-8*], α-cyano(3-phenoxyphenyl)methyl 2,2,3,3-tetramethylcyclopropanecarboxylate, a white crystalline solid, has an LD_{50} (rat, oral) of 25 mg/kg. It is a broad-spectrum soil insecticide and acaricide.

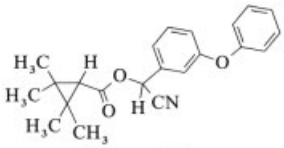
Fenpropathrin

Tefluthrin [*79538-32-2*], 2,3,5,6-tetrafluoro-4-methylphenylmethyl (2)-(1*RS*,3*RS*)-3-(2-chloro-3,3,3-trifluoro-1-propenyl)-2,2-dimethylcyclopropanecarboxylate, is a white solid, *mp* 44 – 60 °C, vapor pressure 1.0 μPa at 25 °C, water solubility 2 µg/L. The LD_{50} (rat, oral) is 23, 35 mg/kg. It is a persistent soil insecticide.

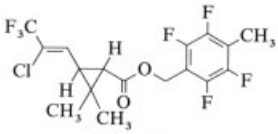
Tefluthrin

3.5.1.3. Phenylacetate Esters

The phenylacetate esters of pyrethroids are structurally optimized in both the acid and the alcohol moieties.

Fenvalerate [*51630-58-1*], α-cyano(3-phenoxyphenyl)methyl ($\pm$)-(2*RS*)-4-chloro-α-(1-methylethyl)benzeneacetate, a clear liquid, d_4^{23} 1.17, vapor pressure 1.4 μPa at 25 °C, is a widely used broad-spectrum insecticide. The LD_{50} (rat, oral) is 450 mg/kg.

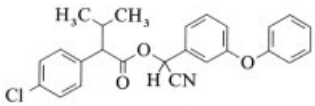
Fenvalerate

Esfenvalerate [*66230-04-4*] is the (+)-2-(*S*) isomer of fenvalerate, *mp* 59 °C.

Flucythrinate [*70124-77-5*], α-cyano-(3-phenoxyphenyl)methyl($\pm$)-(2*S*)-4-(difluoromethoxy)-α-(1-methylethyl)benzeneacetate, is a broad-spectrum insecticide. The LD_{50} (rat, oral) is 53, 87 mg/kg.

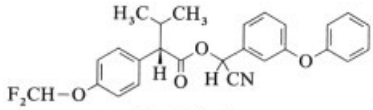
Flucythrinate

Fluvalinate [*69409-94-5*], α-cyano(3-phenoxyphenyl)methyl *N*-[2-chloro-4-(trifluoromethyl)phenyl]-DL-valine, vapor pressure 13 μPa at 25 °C, solubility in water 2 µg/L, is a yellowish liquid, and a general-purpose insecticide.

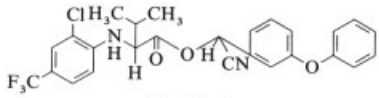
Fluvalinate

3.5.2. Mode of Action [47]

Pyrethroids, particularly the optimized synthetic derivatives, have very high lipid – water partition coefficients and penetrate rapidly into the insect nervous system. Their action is characterized by excitation, uncoordination, and paralysis, producing the characteristic "knockdown effect".

The natural pyrethrins, resmethrin, permethrin, and DDT, make the Na^+ channel close very slowly after each action potential, producing a prolonged afterpotential. Pyrethroids with the α-cyano group, deltamethrin and cypermethrin, delay closing of the Na^+ channel so that spontaneous generation of the action potential is suppressed. These actions are a function of the Ca^{2+}-dependent protein kinases associated with the Na^+ channels that phosphorylate neural protein.

Extensive structure – activity studies of the pyrethroids have defined the following essential features for high insecticidal activity:

1) *Acid Moiety.* The 2,2-methyl groups of cyclopropanecarboxylic acid or its stereochemical equivalent (the methyl groups of the α-(1-methylethyl)benzeneacetic acid) are es-

sential. The C-3 substituent in cyclopropane can be isobutenyl, 2,2-dihalovinyl, or trifluoromethylchlorovinyl. The configuration at C-1 and C-3 of the cyclopropane must be either 1R, 3S or 1R, 3R.

2) *Alcohol Moiety.* The alcohol moiety should contain a cyclopentenolone ring or stereochemically similar furan or benzyl groups and an unsaturated side chain or aromatic group. The α-cyano group apparently acts as a dimensional spacer to promote maximum complementarity with the axonal receptor.

3.5.3. Environmental Toxicology

The synthetic pyrethroids are insecticidally active at dosages of 20 – 100 g/ha; i.e., they are about ten times more effective than other insecticides. Thus, their use lessens the environmental burden on nontarget organisms considerably. However, pyrethroids are extremely toxic to fish, with LC_{50} values in the 0.001-ppm range. Persistence varies from 1 – 2 d for the natural pyrethrins and resmethrin to 1 – 3 months for deltamethrin and tefluthrin.

Acute toxicity to humans and higher animals ranges from very safe compounds such as resmethrin and permethrin to highly toxic compounds such as deltamethrin, flucythrinate, and tefluthrin. Many of the more active compounds are highly irritating to the skin and nasal membranes. A major drawback of the highly toxic compounds is lack of an effective antidote for acute poisoning.

3.5.4. Pyrethroid Synergists [5], [48]

Synergists are nontoxic compounds used in conjunction with insecticides to produce considerably more than additive toxicity. Strong synergistic interactions have been found with carbamates and with rotenone or ryanodine, but the use of synergists has practical importance only for pyrethroid insecticides. Natural pyrethroids are rapidly detoxified in the insect by mixed-function oxidases; thus, insects paralyzed with pyrethroids often recover. These enzymes act through cytochrome P450 to produce hydroxylating (- OH) groups that attack the C=C double bonds of pyrethrins, the cinerin side chains, and methyl groups of the isobutenyl moiety. Synergists of methylenedioxyphenyl(1,3-benzodioxole) react with the iron atom of cytochrome P450 and prevent detoxification. Such synergists are usually included with natural pyrethrins at ratios of (10 – 20) : 1 in fly sprays and aerosols. They may increase the activity of pyrethrins as much as 30-fold. Extensive structural optimization of synthetic pyrethroids has eliminated most of the detoxification sites. Synergists are generally ineffective with the persistent pyrethroids.

Piperonyl butoxide [*51-03-6*], 3-{[2-(2-butoxyethoxy)ethoxy]methyl}-6-propyl-1,3-benzodioxole, *bp* 180 °C (0.13 kPa), d_4^{25} 1.04 – 1.07, LD_{50} (rat, oral) 7 500 – 6 150 mg/kg, is the synergist most commonly used with natural pyrethrins in pressurized aerosol sprays.

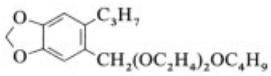

Piperonyl butoxide

Other synergists with similar properties include the following:

Sulfoxide [*120-62-7*], 5-(2-octylsulfinyl)propyl-1,3-benzodioxole, has an LD_{50} (rat, oral) of 2 000 mg/kg.

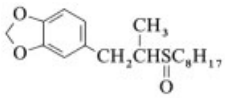

Sulfoxide

Sesamex [*51-14-9*], 5-{1-[2-(2-ethoxyethoxy)-ethoxy]-ethoxy}-1,3-benzodioxole, has an LD_{50} (rat, oral) of 2 000 mg/kg.

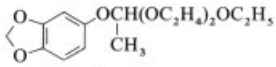

Sesamex

MGK-264 [*113-48-4*], *N*-(2-ethylhexyl)-[2.2.1]-5-heptene-2,3-dicarboximide, *bp* 158 °C (0.26 kPa), d_4^{18} 1.05, LD_{50} (rat, oral) 2800 mg/kg, is an effective pyrethroid and carbamate synergist of totally different chemical structure.

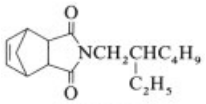

MGK-264

4. Dinitrophenols [8]

The first synthetic organic insecticide was the potassium salt of 4,6-dinitro-2-methylphenol patented as a dormant spray for orchard pests in 1892. Dinitrophenols have a wide range of biocidal activity and have been used as insecticides, acaricides, herbicides, and fungicides.

Dinitro-*o*-cresol (DNOC) [*534-52-1*], 4,6-dinitro-2-methylphenol, is a yellow crystalline compound, *mp* 87 °C, vapor pressure 6.8 mPa at 25 °C, water solubility 0.013 wt %. It is a pseudoacid ($K_a 2.5 \times 10^{-6}$) and readily forms water-soluble ammonium, potassium or sodium salts. The LD_{50} values (rat) are 30 mg/kg (oral) and 200 mg/kg (dermal).

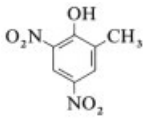

Dinocap [*39300-45-3*], 2-(1-methylheptyl)-4,6-dinitrophenyl 2-butenoate, *bp* 138 – 140 °C (6.7 Pa), LD_{50} (rat) 980, 1 190 mg/kg (oral) and > 4 700 mg/kg (dermal), is an acaricide and fungicide.

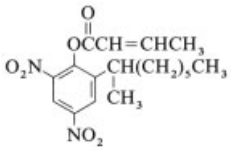

Dinoseb [*88-85-7*], 4,6-dinitro-2-(1-methylpropyl)phenol, *mp* 42 °C, has LD_{50} values (rat) of 136, 225 mg/kg (oral) and > 1 000 mg/kg (dermal). Triethanolamine salts have been used as a dormant spray for deciduous fruit pests.

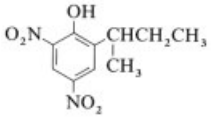

Binapacryl [*485-31-4*], the 3-methyl-2-butenoate ester of dinoseb, *mp* 70 °C, LD_{50} values (rat) 63, 58 mg/kg (oral) and 810, 720 mg/kg (dermal), has been used as an acaricide for foliage.

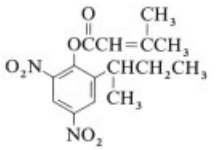

Mode of Action. The dinitrophenol insecticides uncouple oxidative phosphorylation and are highly toxic to plants, insects, higher animals, and humans. Poisoning increases (three- to tenfold) cellular respiration. Death results from metabolic exhaustion because of an inability to utilize the energy provided by respiration and glycolysis. This general biocidal activity has rendered them obsolete as insecticides.

5. Organochlorine Insecticides [49], [50]

5.1. DDT and Derivatives [5], [9], [51]

Although DDT was synthesized by O. Zeidler in 1874, its insecticidal properties were discovered by P. Müller in 1939 [51]. This discovery marked the beginning of the era of synthetic organic insecticides. Because DDT appeared at the onset of World War II, it rapidly became the standard remedy for control of typhus and malaria in both military and civilian use. At its peak use in 1962, about 85×10^6 kg of DDT was produced

in the United States. The compound has an appropriate combination of properties: (1) broad spectrum of insecticidal activity, (2) long residual persistence, (3) low acute toxicity to mammals, and (4) low cost. The enormous publicity it received after World War II heightened its use as a panacea for all insect pest control problems, and DDT quickly became the standard remedy for control of the codling moth on deciduous fruit, the pink bollworm on cotton, the Colorado potato beetle, and the European corn borer. During the 1960s, DDT was registered for use on 334 agricultural commodities in the United States. It was highly effective against forest defoliators such as the gypsy moth and the spruce budworm and as a home mothproofing agent. The safety and long residual action of DDT stimulated the WHO program for global eradication of malaria, and it still has major place in malaria control.

The compound known as DDT [*50-29-3*], 1,1′-(2,2,2-trichloroethylidene)bis(4-chlorobenzene), is a white crystalline solid, *mp* 109 °C, vapor pressure at 20 °C 0.025 mPa. Technical DDT is a white amorphous powder containing 65 – 80 % of the active 4,4′-DDT; 14 – 21 % of the nearly inactive 2,4′-DDT [*789-02-6*], *mp* 73 – 74 °C; up to 4 % of the insecticidal DDD [*72-54-8*], 1,1′-(2,2-dichloroethylidene)bis(4-chlorobenzene), *mp* 110 °C; and traces of 2,2′-DDT and bis(4-chlorophenyl)sulfone. Technical DDT, *mp* 80 –94 °C, is almost insoluble in water (1.2 µg/L, readily soluble in xylene and tetralin (600 g/L), and moderately soluble in mineral oil and kerosene (50 – 80 g/L). In alkaline solution, DDT is dehydrochlorinated readily to the noninsecticidal 1,1-dichloroethenylidenebis(4-chlorobenzene) [*72-55-9*] (DDE), *mp* 85 °C. This reaction also occurs in biological systems and accounts for the storage and accumulation of DDE in lipid tissue of humans and higher animals. The rat LD_{50} values of DDT are 113, 118 mg/kg (oral) and 2510 mg/kg (dermal).

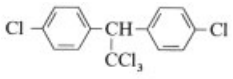

DDT

Methoxychlor [*72-43-5*], 1,1′-(2,2,2-trichloroethylidene)bis(4-methoxybenzene), *mp* 89 °C, is less easily dehydrochlorinated than DDT in alkaline solution or in biological systems. However, the methoxy groups are readily dealkylated in vivo by microsomal oxidases to produce phenols that are eliminated easily. Thus, methoxychlor does not bioaccumulate like DDT and is favored for general environmental use. Methoxychlor is one of the safest of all insecticides with an LD_{50} (rat, oral) of > 6 000 mg/kg. Unfortunately, insects resistant to DDT are generally cross-resistant to methoxychlor.

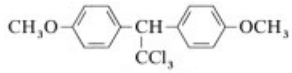

Methoxychlor

The compound known as DDD [*72-54-8*] (TDE), 1,1′-(2,2-dichloroethylidene)bis(4-chlorobenzene), occurs as an impurity in technical DDT and has been marketed as an insecticide. Technical DDD contains mainly the 4,4′-isomer, *mp* 110 °C, and 7 – 8 % of the 2,4′-isomer, *mp* 76 °C. The properties of DDD are almost indistinguishable from DDT and the product is even safer with LD_{50} values (rat) of 3 400 mg/kg (oral) and > 10 000 mg/kg (dermal). Like DDT, DDD is also bioaccumulated and its registration was also canceled in the United States in 1974.

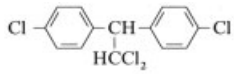

TDE

Perthane [*72-56-0*], 1,1′-(2,2-dichloroethylidene)bis(4-ethylbenzene), *mp* 60 – 61 °C, is the 4,4′-diethyl analogue of DDT. It is very biodegradable and has low mammalian toxicity, LD_{50} (rat, oral) 8 170 mg/kg. Perthane has been used indoors, where very low acute and chronic toxicities are desirable.

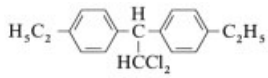

Perthane

The compound DFDT [*475-26-3*], 1,1′-(2,2,2-trichloroethylidene)bis(4-fluorobenzene), 1,1′-(2,2,2-trichloroethylidene)bis-

(4-fluorobenzene), *mp* 44 °C, has chemical and insecticidal properties similar to DDT. It was used as a sanitary insecticide by the German army in World War II but is too phytotoxic for agricultural use. The LD_{50} (rat, oral) is 900 mg/kg.

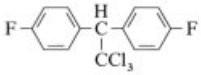

5.1.1. Mode of Action

Both DDT and its analogues specifically affect the peripheral sensory organs of insects to produce violent trains of afferent impulses that cause hyperactivity and convulsions. Subsequent paralysis and death are the result of metabolic exhaustion or the production of an endogenous neurotoxin. The insecticidal action of DDT is facilitated by its highly lipophilic character, which promotes absorption through the lipoprotein matrix of the chitinous insect cuticle. The DDT molecule is absorbed into the sodium gate mechanism of the insect nerve axon, prolonging the negative afterpotential and promoting multiple, repetitive nerve impulse firing. The critical site of action may be the inhibition of Ca^{2+} ATPase.

5.1.2. Environmental Toxicology

Because DDT is the most durable and persistent of all the commonly used contact insecticides, the very properties that made it such an effective insecticide–its low vapor pressure, stability to photooxidation, high fat solubility (ca. 100 g/kg), and low water solubility (1.2 µg/L) – resulted in its becoming the prototype environmental micropollutant. As a result of its enormous use worldwide and the environmental stability of its primary degradation product DDE, DDT – DDE residues are ubiquitous in the global environment. By 1969, the average U.S. inhabitant had 2.3 – 4.0 mg/kg of DDT and 4.3 –8.0 mg/kg of DDE stored in body fat; levels in other countries were even higher. This widespread persistence and bioaccumulation were key factors in the withdrawal of DDT registrations by the EPA, as of January 1974. The DDT residues in humans declined slowly to about one-half the above values by 1978. In vivo DDT is slowly dechlorinated reductively to DDD and dechlorinated further to DDA or 4,4′-dichlorodiphenylacetic acid, the predominant excretory metabolite. In anaerobic microorganisms, DDT may form 4,4′-dichlorodiphenylacetonitrile (DDCN). However, most DDT that enters the environment is sequestered as DDE. In humans, DDT is secreted in milk as DDT, DDD, and DDE, with traces of their 2,4′-isomers.

Whereas DDT is highly toxic to fish (LC_{50} for trout and bluegill 0.002 – 0.008 mg/kg), it is only moderately toxic to birds (oral LD_{50} for mallard 1 300 mg/kg and for pheasant $>$ 2 240 mg/kg). However, because of its high lipid – water partition value, DDT bioaccumulates as much as 3×10^6-fold from the waters of Lake Michigan 6 ng/L in the bodies of lake trout *Salvelinus namaycush* ($>$ 20 mg/kg). In raptorial birds such as eagle, osprey, and brown pelican, bioaccumulation of DDE from fish has decreased eggshell thickness by 10 – 15 %, resulting in widespread reproductive failure. Because of the many environmental problems resulting from massive use of DDT, its registration for agricultural use was canceled in the United States and Canada; Japan and many European countries have enacted similar prohibitions.

5.2. Hexachlorocyclohexanes [49]

Lindane [*58-89-9*], γ-1,2,3,4,5,6-hexachlorocyclohexane, is the active component of hexachlorocyclohexane (benzene hexachloride) insecticides. Benzene hexachloride was first produced by FARADAY in 1825 by chlorination of benzene in sunlight. Its insecticidal properties were discovered independently by DUPIRE and RAUCOURT in France and by SLADE et al. in England around 1941. Crude benzene hexachloride is a grayish to brownish amorphous solid, with a characteristic musty odor, which begins to melt at 65 °C. It consists of 10 – 18 % of the active γ-isomer, *mp* 112 °C, configuration *aaaeee* (*a* is axial, *e* is equatorial), together with at least four other nearly inactive stereoisomers: α-isomer, *mp* 157 °C (*aaeeee* or *aeeeea*), 55 – 70 %; β-isomer, *mp* 309 °C (*eeeeee*), 5 – 14 %; δ-isomer, *mp* 138 °C (*aeeeee*), 6 – 8 %; ε-isomer,

mp 219 °C (*aeeaee*), 3 – 4 %; and a trace of η-isomer, *mp* 90 °C (*aeaaee*). A heptachlorocyclohexane and an octachlorocyclohexane are present as impurities and contribute to the unpleasant odor of benzene hexachloride.

Lindane is the only component with appreciable insecticidal activity; it is prepared by treating crude BHC with methanol or acetic acid in which the α- and β-isomers are nearly insoluble, followed by chromatographic adsorption or fractional crystallization.

Lindane, vapor pressure 0.12 Pa at 20 °C, is a white crystalline product that is practically odorless, soluble in water to 7.3 mg/L, and soluble in petroleum solvents. It is very stable to sunlight and oxidation, and can be burned without appreciable decomposition; however, it is decomposed by alkali to form mainly 1,2,4-trichlorobenzene. The rat LD_{50} is 88, 91 mg/kg (oral) and 900, 1 000 mg/kg (dermal).

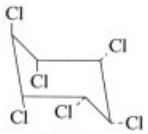

Lindane

Lindane is used predominantly as a seed dressing, as a soil insecticide, and for the control of locusts and grasshoppers. It is also used to control ectoparasites of humans and domestic animals and as a residual spray to control the *Anopheles* vectors of malaria. Because of its relatively high volatility, it is useful for controling wood-boring insects of timber, fruit trees, and ornamental plants.

The mode of action of lindane is not well understood, but like DDT and the cyclodienes it causes distortion and consequent excitation of ionic transmission in the arthropod nerve. The mechanism is believed to be competitive blockage of γ-aminobutyric acid (GABA) synaptic transmission.

Crude benzene hexachloride causes severe off-flavors in root crops grown in treated soil, but the problem does not occur after the use of purified lindane. Lindane is highly toxic to fish, with LD_{50} values to trout of 0.027 ppm and bluegill 0.068 ppm. It has low oral toxicity in birds, an LD_{50} of $> 2\,000$ mg/kg for mallards. The average soil half-life is about two years.

5.3. Cyclodienes [49]

The enormous commercial success of DDT stimulated further investigation of organochlorine insecticides. Julius Hyman developed chlordane, the first cyclodiene insecticide, in 1944. The cyclodienes are polychlorinated cyclic hydrocarbons with endomethylene-bridged structures. They are prepared by the Diels – Alder reaction, with hexachlorocyclopentadiene as the primary dienophile to form adducts with a variety of dienes.

Chlordane [*12789-03-6*] is made by chlorinating hexachlorodicyclopentadiene (chlordene) to form two isomers of 1,2,4,5,6,7,8,8-octachloro-2,3,3a,4,7,7a-hexahydro-4,7-methano-1*H*-indene, α-*trans, mp* 106.5 °C, and β-*cis*, *mp* 104.5 °C. The β-isomer has significantly greater insecticidal activity. Technical chlordane is an amber liquid of d_4^{25} 1.61 which contains about 60 % of these isomers, 10 – 20 % heptachlor, and lesser amounts of chlordene and a nonachloro derivative. A total of 14 components have been identified by gas – liquid chromatography.

Technical chlordane, vapor pressure 1.3 mPa at 25 °C, is soluble in petroleum solvents but has very low water solubility (0.009 ppm). It has rat LD_{50} values of 335, 430 mg/kg (oral) and 840, 690 mg/kg (dermal). Chlordane has had very wide use as a soil insecticide for termite control and as a household insecticide for the control of ants and cockroaches, together with limited agricultural use as a soil insecticide.

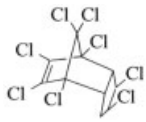

β-Chlordane

Heptachlor [*76-44-8*], 1,4,5,6,7,8,8-heptachloro-3a,4,7,7a-tetrahydro-4,7-methanoindene, *mp* 95 °C is three to five times as active insecticidally as chlordane. The pure crystalline material, vapor pressure 0.039 Pa at 20 °C, is soluble in petroleum solvents but very insoluble in water to 0.056 ppm. The rat LD_{50} values

are 100, 162 mg/kg (oral) and 195, 250 mg/kg (dermal). Heptachlor is oxidized readily to heptachlor epoxide [*1024-57-3*], *mp* 159 – 160 °C, LD_{50} (rat, oral) 47 mg/kg, which is an important and highly persistent environmental pollutant.

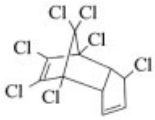

Heptachlor

Endosulfan [*115-29-7*] is the adduct of hexachlorocyclopentadiene and 1,4-dihydroxy-2-butene reacted further with $SOCl_2$ to produce 6,7,8,9,10,10-hexachloro-1,5,5a,6,9,9a-hexahydro-6,9-methano-2,4,3-benzodioxathiepin-3-oxide. The technical product is a brownish solid, *mp* 70 – 100 °C, vapor pressure 1.3 mPa at 25 °C, soluble in petroleum solvents but having low solubility in water. The rat LD_{50} values are 43, 18 mg/kg (oral) and 130, 74 mg/kg (dermal). Technical endosulfan consists of about seven parts of the α-isomer, *mp* 108 – 109 °C (cis with regard to the sulfite groups) and three parts of the β-trans isomer, *mp* 206 – 208 °C. The α-isomer has somewhat greater insecticidal activity and is slowly converted to the more stable β-isomer at high temperature. Both isomers oxidize slowly in air and in biological systems to endosulfan sulfate [*1031-07-8*], *mp* 181 – 182 °C. In acid media, both isomers form endosulfan diol [*2157-19-9*], *mp* 203 – 205 °C.

Endosulfan is a broad-spectrum insecticide used to control pests of vegetables, fruit, field crops, and ornamentals. Unlike other cyclodiene insecticides, it is biodegradable by hydrolysis at the sulfite ester bonds.

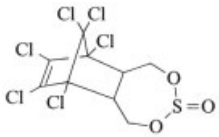

Endosulfan

Isobenzan [*297-78-9*], 1,3,4,5,6,7,8,8-octachloro-1,3,3a,4,7,7a-hexahydro-4,7-methanoisobenzofuran is produced by photochlorination of the Diels – Alder adduct of hexachlorocyclopentadiene and 2,5-dihydrofuran. It has a *mp* of 122 – 123 °C and a vapor pressure of 0.39 mPa at 20 °C; it is soluble in water to about 0.1 mg/L at 20 °C. Isobenzan had uses in Europe similar to heptachlor, but production was discontinued in 1966 because of its very high mammalian toxicity, LD_{50} (rat, oral) 7, 8.7 mg/kg.

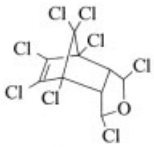

Isobenzan

Aldrin [*309-00-2*], 1,2,3,4,10,10-hexachloro-1,4,4a,5,8,8a-hexahydro-1,4-*endo,exo*-5,8-dimethanonaphthalene, is the Diels – Alder adduct of hexachlorocylopentadiene and bicyclo[2.2.1]-hepta-2,5-diene. It is a colorless solid, *mp* 104 °C, vapor pressure 5.2 mPa at 20 °C. This compound is very slightly soluble in water (to 0.027 mg/L) but soluble in petroleum hydrocarbons. Aldrin has rat LD_{50} values of 39, 60 mg/kg (oral) and 98 mg/kg (dermal). It has been widely used as a seed treatment and soil insecticide where it is gradually oxidized to its epoxide dieldrin.

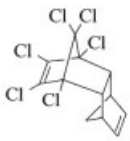

Aldrin

Dieldrin [*60-57-1*], 3,4,5,6,9,9-hexachloro-1a,2,2a,3,6,6a,7,7a-octahydro-2,7 : 3,6-dimethanonaphth[2,3-*b*]oxirene, is the 6,7-epoxide of aldrin, *mp* 176 °C, vapor pressure 0.023 mPa at 20 °C. It is produced by oxidation of aldrin with peracetic or perbenzoic acid, and is very slightly soluble in water (to 0.05 ppm) and soluble in petroleum solvents. Dieldrin has rat LD_{50} values of 46 mg/kg (oral) and 90, 60 mg/kg (dermal). It is a highly persistent insecticide and was used extensively as a residual treatment for control of the *Anopheles* mosquito vectors of malaria and the tsetse fly vectors of African

trypanosomiasis. However, resistance to it developed very rapidly. In agriculture, it had more limited use as a seed treatment and soil insecticide. Because of the environmental persistence of dieldrin and its propensity for bioaccumulation, the registrations of both dieldrin and aldrin, which is readily epoxidized to dieldrin, were canceled in the United States in 1974.

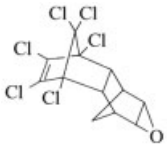

Dieldrin

Endrin [*72-20-8*], 3,4,5,6,9,9-hexachloro-1a,2,2a,3,6,6a,7,7a-octahydro-2,7 : 3,6-dimethanonaphth[2,3-*b*]oxirene, *mp* 245 °C (decomp.), vapor pressure 0.022 mPa at 25 °C, is the *endo-endo*-isomer of dieldrin and is produced by a Diels – Alder reaction of hexachloronorbornadiene (formed from hexachlorocyclopentadiene and acetylene or vinyl chloride) with cyclopentadiene, followed by epoxidation. Endrin is soluble in water to about 0.06 ppm and is somewhat less soluble in petroleum solvents than dieldrin. The rat LD_{50} values are 17.8, 7.5 mg/kg (oral) and 15 mg/kg (dermal). Endrin has been used as a cotton insecticide, but because of its high toxicity to fish and its propensity for bioaccumulation, its uses have been restricted substantially. It has also had very limited use as a rodenticide for field mouse control.

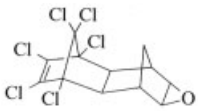

Endrin

Mirex [*2385-85-5*] is a dimer of hexachlorocyclopentadiene dodecachloropentacyclo-[5.3.0.0^{2,6}.0^{3,9}.0^{4,8}]decane, *mp* 485 °C, rat LD_{50} 306, 600 mg/kg (oral) and > 2 000 mg/kg (dermal). Mirex is extremely resistant to biodegradation. It was once considered the perfect stomach poison for control of the imported fire ant *Solenopsis invicta* but, even when used in bait at a dosage of a few grams per hectare, was found to bioaccumulate in birds and other animals. Its registration was canceled in the United States in 1976. Mirex is environmentally very stable; it is slowly oxidized to chlordecone and dechlorinated in sunlight to photomirex.

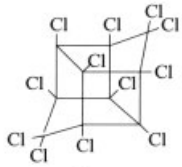

Mirex

Chlordecone [*143-50-0*], decachloro-5-oxo-pentacyclo[5.3.0.0^{2,6}.0^{3,9}.0^{4,8}]decane, *mp* 349 °C (decomp.), is the 2-keto analogue of mirex and is soluble in water to 4 g/L at 100 °C. The rat LD_{50} values are 95, 140 mg/kg (oral) and > 2 000 mg/kg (dermal). Like mirex, chlordecone is a stomach poison insecticide; it has been used in bait for cockroaches and ants and for the control of banana thrips. After widespread environmental contamination of the James River and Chesapeake Bay with chlordecone from manufacturing effluent, its registration was canceled in the United States in 1978.

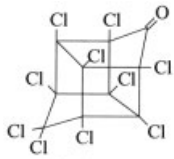

Chlordecone

Kelevan [*4234-79-1*], decachloro-5-(4-ethoxycarbonyl-2-oxo-butyl)-5-hydroxypentacyclo-[5.3.0.0.^{2,6}.0^{3,9}.0^{4,8}]decone, *mp* 91 °C, is a chlordecone analogue that is produced by reaction with ethyl levulinate. It has been used in Europe as a stomach poison for the control of the Colorado potato beetle.

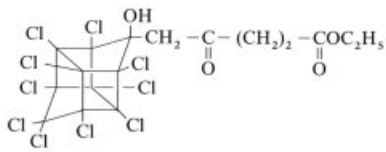

Kelevan

Bromodan [*1715-40-8*], 5-bromomethyl-1,2,3,4,7,7-hexachlorobicyclo[2.2.1]hept-2-ene, *mp* 75 – 79 °C, is the Diels – Alder adduct of hexachlorocyclopentadiene and allyl bromide. It has very low mammalian toxicity, LD_{50} (rat, oral) 1.2×10^4 mg/kg, and has been used in Europe to control ectoparasites of domestic animals.

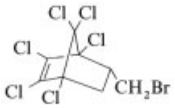

Bromodan

Mode of Action [52]. The cyclodienes appear to act by a common mechanism affecting the nerve axon to produce hypersensitivity, convulsions, prostration, and death. Spray crewmen with lengthy exposure to dieldrin have suffered from prolonged and repeated central nervous system epileptiform convulsions; similar nervous disturbances have occurred in workers exposed to chlordecone. The lesion appears to be blockage of GABA synaptic transmission by competitive inhibition of the neurotransmitter binding site. This mode of action is common to the cyclodienes, toxaphene, and lindane. The highest toxic action is shown by compounds containing a polychlor center and an adjacent, properly oriented electronegative site (e.g., a double bond adjacent to a Cl, O, S, or N atom).

Environmental Toxicology. Cyclodienes add oxygen readily; e.g., aldrin and heptachlor form the more stable epoxides dieldrin and heptachlor epoxide. The epoxides are formed in the tissues of animals that ingest or absorb the compounds, and they are concentrated and stored preferentially in fatty tissue. Epoxidation also occurs in plants and soil. Because dieldrin and heptachlor epoxide are more stable (e.g., the half-life on alfalfa leaves is 7 – 8 d, compared to < 1 d (for aldrin and heptachlor), in situ oxidation markedly affects the residual behavior of these cyclodienes. Average soil half-lifes for cyclodienes are estimated as aldrin 1 – 4 years, dieldrin 1 – 7 years, chlordane 2 – 4 years, and endrin 4 – 8 years. The high lipophilicity of dieldrin and heptachlor epoxide has resulted in severe environmental pollution. They bioaccumulate in fish from water up to 10^5-fold and are present in human milk and fatty tissue. Dieldrin residues in the body fat of U.S. inhabitants averaged 0.22 – 0.35 ppm in 1970 and declined to 0.12 – 0.14 ppm by 1978, after cancellation of the use of aldrin and dieldrin in 1974. Oxychlordane [*26880-48-8*], mirex, and chlordecone are also highly persistent and bioaccumulative.

The cyclodienes are extremely toxic to fish with the following LC_{50} values (in milligrams per kilogram) to trout and bluegill: endrin 0.001 – 0.002, endosulfan 0.001 – 0.003, aldrin 0.006 – 0.01, dieldrin 0.003 – 0.015, heptachlor 0.03 –0.026, and chlordane 0.022 – 0.095. The LD_{50} values to pheasant and mallard, respectively, are aldrin 16.8 and 520, dieldrin 79 and 381, and endrin 1.6 and 5.6 mg/kg. Lifetime animal feeding studies have shown that heptachlor, aldrin, and dieldrin are carcinogens. The data with endrin are equivocal.

5.4. Chlorinated Camphenes

Toxaphene, $C_{10}H_{10}Cl_8$, is a mixture of more than 180 chlorinated cyclic compounds produced by photochlorination of the bicyclic terpene camphene. The product is a yellowish, semicrystalline gum, *mp* 65 – 90 °C, d_4^{25} 1.64, vapor pressure 0.13 mPa at 25 °C. Toxaphene is soluble in water to about 0.4 mg/L and very soluble in petroleum solvents. It is unstable in alkali, upon prolonged exposure to sunlight, and above 155 °C, liberating hydrogen chloride and losing some of its insecticidal potency. The rat LD_{50} values are 80, 90 mg/kg (oral) and 1 075, 780 mg/kg (dermal). The most active ingredients identified from technical toxaphene are 2,2,5- *endo*-6-*exo*-8,9,10-heptachlorobornane [*51775-36-1*], with a mouse intraperitoneal LD_{50} of 6.6 mg/kg, and 2,2,5-*endo*-6-*exo*-8,9,9,10-octachlorobornane [*58002-18-9*], with a mouse intraperitoneal LD_{50} of 3.1 mg/kg. Each constitutes 2 – 6 % of the technical mixture. Toxaphene is a broad-spectrum persistent insecticide with major use on cotton and for the control of animal ectoparasites.

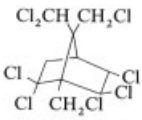

2,2,5-*Endo*-6-*exo*-8,9,9,10-octachlorobornane

Environmental Toxicology. Toxaphene is extremely toxic to fish, with LC_{50} values to trout and bluegill of 0.003 – 0.006 mg/kg. At a water concentration as low as 50 ng/L, fish exposed to toxaphene residues suffer broken-back syndrome, a crippling collagen deformity. From water, bioaccumulation occurs in fish at levels up to 10^5-fold. Toxaphene is also highly toxic to birds: the oral LD_{50} values for pheasant mallard are 40 and 71 mg/kg, respectively. Because of the complexity of the toxaphene mixture, its environmental persistence and fate are difficult to assess. Substantial toxaphene residues have been found throughout the Great Lakes, as a result of atmospheric distribution from heavy use on cotton in the southern states. The estimated soil half-life ranges from 2 months to 10 years. Toxaphene is carcinogenic in rats and mice, and its registration was canceled by the U.S. EPA in 1983.

6. Organophosphorus Insecticides

[53–57]

The organophosphorus (OP) insecticides are primarily triesters of phosphoric acid $P(O)(OH)_3$ and phosphorothioic acid $P(S)(OH)_3$.

Historical Aspects. Discovery of the biological activity of organophosphate esters began with the chance observation in 1932 that exposure to the vapor of diethyl phosphorofluoridate, $(C_2H_5O)_2P(O)F$, produced strong cholinergic effects in humans [58]. Development of this class of compounds as insecticides resulted from the work of Gerhard Schrader in 1937, who devised tetraethyl pyrophosphate (TEPP) as a substitute for nicotine. Schrader's research also led to the related, extremely toxic compounds, the "nerve gases" tabun and sarin; because of wartime secrecy, the development of insecticides did not become evident until 1946. Because TEPP hydrolyzes rapidly in water, it was ineffective in conjunction with Bordeaux mixture used as a fungicide for potatoes. To find a more stable compound to control the Colorado potato beetle *Leptinotarsa decomlineata*, Schrader turned to phosphorothionate esters and produced methyl and ethyl parathions, the most widely used of the organophosphate insecticides. Schrader's development of schradan, demeton (the first practical systemic insecticide), azinphosmethyl (the standard remedy for the codling moth *Laspeyresia pomonella*), trichlorfon (a stomach poison), coumaphos (for animal ectoparasites), and numerous other insecticidal molecules, followed from his elucidation of the principles of organophosphorus toxicity based on the formula:

where R^1 and R^2 are short-chain alkoxy, alkyl, or amino residues, X is a displaceable group with a high-energy phosphorus bond, such as fluorine or an acyl anhydride; and pentavalent phosphorus is bonded directly to oxygen or sulfur. At present ca. 100 commercially successful OP insecticides are marketed in estimated quantities of $(1.5–2.5)\times10^5$ t. More than 10^5 different organophosphorus esters appear to have been synthesized and evaluated as pesticides. The potential number of toxic structures that can be devised from Schrader's formula is in excess of 25×10^6 [59].

Organophosphorus insecticides are available with a variety of properties. Some, such as TEPP and mevinphos, have a very short residual action; others, such as diazinon, azinphos-methyl, and chlorpyrifos have a lengthy persistence. Some have a broad spectrum of action (e.g., parathion and malathion), whereas others have a very specific action (e.g., schradan and trichlorfon). The unique properties of plant absorption and translocation found in dimethoate and demeton have resulted in successful plant systemic insecticides. Persistent compounds used in soil and seed treatment (e.g., disulfoton, phorate, and terbufos) protect newly developed seedlings. Animal systemics such as cruformate can be fed to cattle to kill the cattle grubs (*Hypoderma* spp.) living in their bodies.

The organophosphorus insecticides are biodegradable because of their chemical reactivity, especially through the phosphate ester bonds, and their use does not present serious problems of bioaccumulation and food chain transfer. Their major use as replacements for organochlorine insecticides has had favorable consequences on environmental quality.

However, organophosphorus compounds act as specific inhibitors of the enzyme acetylcholinesterase, which is an essential component of nerve impulse transmission in both vertebrates and invertebrates. Therefore, they generally lack selectivity. Careless use and improper storage have caused hundreds of thousands of cases of human poisoning and thousands of deaths. By taking advantage of differences in the detoxification processes and the structures of the target enzyme in insects and mammals, a few insecticides including malathion, ronnel, fenitrothion, stirifos, and phoxim have been produced, which incorporate an effective degree of insecticidal action with adequate safety to the user and to domestic animals.

6.1. Phosphoric Acid Anhydrides

The aliphatic organophosphorus insecticides developed by SCHRADER hydrolyzed rapidly and had very high mammalian toxicity. They are now of largely historical interest.

Tetraethyl pyrophosphate [*107-49-3*] is a clear, colorless liquid, *bp* 104 – 110 °C (10.7 Pa), d_4^{25} 1.1845. It is miscible with water and hydrolyzes rapidly (half-life 6.8 h at 25 °C). The rat LD_{50} values are 1.0 mg/kg (oral) and 2.4 mg/kg (dermal).

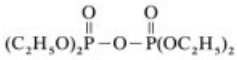

Tetraethyl pyrophosphate

Sulfotepp [*3689-24-5*], *O,O,O′,O′*-tetraethyl dithiopyrophosphate is a yellowish liquid, *bp* 110 – 113 °C (26.7 Pa), d_4^{25} 1.196. It is soluble in water to 20 mg/L and hydrolyzes only in alkaline solution. The LD_{50} (rat, oral) is 5 mg/kg. Sulfotepp is used in aerosols for greenhouse pest control.

Sulfotepp

Dimefox [*115-26-4*], *N,N,N′,N′*-tetramethylphosphorodiamidofluoridate, *bp* 67 °C (0.5 kPa), d_4^{25} 1.12, LD_{50} (rat, oral) 3.5 mg/kg, and Schradan [*152-16-9*], octamethyl pyrophosphoramide, *bp* 154 °C (0.5 kPa), d_4^{25} 1.13, LD_{50} (rat, oral) 10 mg/kg, were the first practical systemic insecticides for mites and aphids.

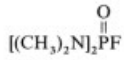

Dimefox

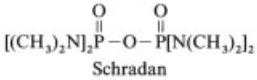

Schradan

6.2. Vinyl Phosphates

Dichlorvos [*62-73-7*], *O,O*-dimethyl *O*-(2,2-dichlorovinyl) phosphate, *bp* 140 °C (2.7 kPa), vapor pressure 1.6 Pa at 20 °C, d_4^{20} 1.314, is a colorless liquid that is soluble in water to about 10 g/L. The half-life in water is about 8 h. The LD_{50} values (rat, oral) are 80, 56 mg/kg. Dichlorvos is a rapid-acting insecticide with strong fumigant action, used in bait and aerosols for flies and mosquitoes. It has been incorporated in slow-release plastic strips and in pet collars against ectoparasites.

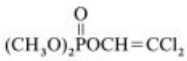

Dichlorvos

Naled [*300-76-5*], *O,O*-dimethyl *O*-(1,2-dibromo-2,2-dichloroethyl) phosphate, *mp* 27 °C, *bp* 110 °C (66.7 Pa), vapor pressure 0.26 Pa at 20 °C, has rat LD_{50} values of 250 mg/kg (oral) and 800 mg/kg (dermal). Naled is a rapid-acting, relatively nonpersistent insecticide.

Naled

Trichlorfon [*52-68-6*], *O,O*-dimethyl-1-hydroxy-2,2,2-trichloroethyl phosphonate, *mp* 83 –84 °C, d_4^{20} 1.73, vapor pressure 1 mPa at 20 °C, is soluble in water to 15.4 % at 20 °C. The rat LD_{50} values are 630, 560 mg/kg (oral) and

> 2 000 mg/kg (dermal). Trichlorfon is a stomach poison insecticide used on foliage for chewing insects and in dry sugar bait for fly control. Above pH 6, trichlorfon is converted rapidly to dichlorvos, which accounts for its toxicity.

Trichlorfon

Mevinphos [*7786-34-7*], *O,O*-dimethyl *O*-(2-methoxycarbonyl-1-methylvinyl) phosphate, *bp* 106 – 107 °C (0.13 kPa), d_4^{20} 1.25, vapor pressure 0.38 Pa at 21 °C, is a mixture of one part *cis* and two parts *trans* isomer, of which the former is about a ten times stronger insecticide. Mevinphos is miscible with water; the rat LD_{50} values are 6.1, 3.7 mg/kg (oral) and 4.7, 4.2 mg/kg (dermal). Mevinphos is a nonpersistent systemic insecticide suitable for treatment of edible produce close to harvest because the residue dissipates rapidly by volatilization and hydrolysis.

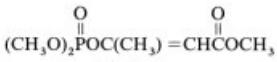

Mevinphos

Phosphamidon [*13171-21-6*], *O,O*-dimethyl *O*-(2-chloro-2-diethylcarbamoyl-1-methyl)vinyl phosphate, *bp* 160 °C (0.2 kPa), d_4^{25} 1.21, vapor pressure 3.2 mPa at 20 °C, is miscible with water. The rat LD_{50} values are 24 mg/kg (oral) and 143, 107 mg/kg (dermal). Phosphamidon is absorbed rapidly by plant surfaces and decomposes quickly in the plant to form a nonpersistent systemic insecticide.

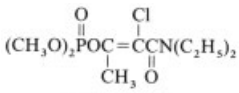

Phosphamidon

Dicrotrophos [*141-66-2*], *cis-O,O*-dimethyl *O*-(2-dimethylcarbamoyl)-1-methylvinyl phosphate, *bp* 115 – 120 °C (0.13 Pa), d_4^{23} 1.19, is water miscible. The rat LD_{50} values are 21, 16 mg/kg (oral) and 43, 42 mg/kg (dermal). Dicrotophos is a contact and systemic insecticide which has been injected into trees to control bark and foliage pests.

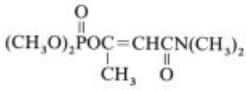

Dicrotophos

Monocrotophos [*6923-22-4*], *cis-O,O*-dimethyl-*O*-(1-methyl-2-methylcarbamoyl)vinyl phosphate, *mp* 54 – 55 °C, vapor pressure 0.9 mPa at 20 °C, is miscible with water. The rat LD_{50} values are 18, 20 mg/kg (oral) 126, 112 mg/kg (dermal). Monocrotophos, a metabolite of dicrotophos, is a systemic and contact insecticide.

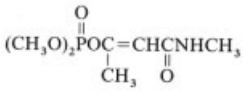

Monocrotophos

Crotoxyphos [*7700-17-6*], *O,O*-dimethyl *O*-[1-methyl-2-(1-phenylcarbethoxy)vinyl] phosphate, *bp* 135 °C (4 Pa), d_4^{25} 1.19, vapor pressure 1.8 mPa at 20 °C, is soluble in water to 1 g/L. The rat LD_{50} values are 110, 74 mg/kg (oral) and 375, 202 mg/kg (dermal). Crotoxyphos is used primarily to control animal ectoparasites.

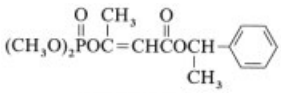

Crotoxyphos

Chlorfenvinphos [*470-90-6*], *O,O*-diethyl *O*-[2-chloro-1-(2,4-dichlorophenyl)vinyl] phosphate, *bp* 168 – 170 °C (66.7 Pa), d_4^{15} 1.36, vapor pressure 0.97 mPa at 25 °C, is soluble in water to 145 mg/L. The rat LD_{50} values are 15, 13 mg/kg (oral) and 31, 30 mg/kg (dermal). Chlorfenvinphos is a contact and soil insecticide for agricultural pests.

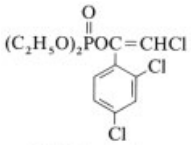

Chlorfenvinphos

Tetrachlorvinphos [*22248-79-9*], Stirifos, *O,O*-dimethyl *O*-[2-chloro-1-(2,4,5-trichlorophenyl)vinyl] phosphate, *mp* 97 – 98 °C, vapor pressure 5 μPa at 20 °C, is soluble in water to 11 mg/L; its rat LD_{50} values are 1100, 1125 mg/kg (oral) and > 5 000 mg/kg (dermal). Stirifos is used for protection against stored product pests, against vegetable and fruit insects, and in veterinary hygiene.

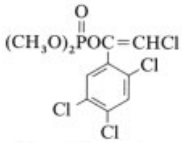

Tetrachlorvinphos

Propetamphos [*31218-83-4*], *O*-[2-(isopropoxycarbonyl)-1-methylvinyl] *O*-methyl *N*-ethyl phosphoramidothioate, *bp* 87 – 89 °C (0.7 Pa), d_4^{20} 1.129, is water-soluble to 110 mg/L. The LD_{50} values (rat, oral) are 75, 82 mg/kg. Propetamphos is used as a household insecticide for flies and cockroaches and to control insect pests of cotton.

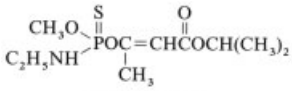

Propetamphos

6.3. Aliphatic Phosphorothioate Esters

Demeton O [*8065-48-3*], *O,O*-diethyl *O*-(2-ethylthio)ethyl phosphorothioate (thiono-isomer, demeton O), *bp* 123 °C (0.13 kPa), d_4^{21} 1.119 is water-soluble to 66 mg/L. The LD_{50} (rat, oral) is 30 mg/kg. Demeton S [*126-75-0*], *O,O*-diethyl *S*-(2-ethylthio)ethyl phosphorothioate (thiolo-isomer, demeton S), *bp* 128 °C (0.13 kPa), d_4^{21} $_4^{21}$ 1.132, is water-soluble to 0.2 %. The technical isomer mixture (2 : 1 ratio of Demeton S to Demeton O) has been used widely as a systemic insecticide for ornamentals and nursery stock. The LD_{50} (rat, oral) is 1.5 mg/kg. The rat LD_{50} values for the mixture are 6.2, 2.5 mg/kg (oral) and 14, 8.2 mg/kg (dermal).

Demeton O

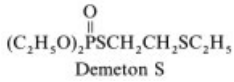

Demeton S

A series of related compounds are also marketed as systemic insecticides but are less persistent and less toxic. The P=O ester is largely responsible for systemic activity.

Demeton methyl [*8022-00-2*] is a mixture of the *O,O*-dimethyl P=S and *O,O*-dimethyl P=O esters analogous to demeton. The respective LD_{50} values (rat, oral) are 180 and 40, 60 mg/kg.

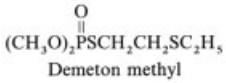

Demeton methyl

Oxydemeton-methyl [*301-12-2*], *O,O*-dimethyl *S*-(2-ethylsulfinyl)ethyl phosphorothiolate has rat LD_{50} values 47, 52 mg/kg (oral) and 173, 158 mg/kg (dermal).

Metasystox S. *O,O*-dimethyl *S*-(2-ethylsulfinyl-1-methyl)ethyl phosphorothioate, LD_{50} (rat, oral) 105 mg/kg, is also a commercial insecticide.

Oxydemeton methyl

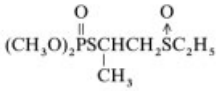

Metasystox S

The P(S)S analogues of the demeton type are much less water-soluble, more persistent systemics than P(O)S compounds. They have been

very widely used in granular soil and seed treatment for row crops and ornamentals.

Disulfoton [*298-04-4*], *O,O*-diethyl *S*-(2-ethylthio)ethyl phosphorodithioate is a pale-yellow liquid, *bp* 62 °C (1.3 Pa), d_4^{21} $_4^{20}$ 1.144. It is soluble in petroleum solvents and to 66 mg/L in water. The rat LD_{50} values are 6.8, 2.3 mg/kg (oral) and 25, 6 mg/kg (dermal).

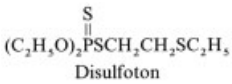
Disulfoton

Thiometon [*144-41-2*], *O,O*-dimethyl *S*-(2-ethylthio)ethyl phosphorodithioate, has an LD_{50} (rat, oral) of 85 mg/kg.

Thiometon

Phorate [*298-02-2*], *O,O*-diethyl *S*-(2-ethylthio)methyl phosphorodithioate, *bp* 118–120 °C (0.1 kPa), is a yellowish liquid, d_4^{21} $_4^{21}$ $_4^{25}$ 1.167, vapor pressure 0.11 Pa at 20 °C, soluble in petroleum solvents and in water to 85 mg/L. The rat LD_{50} values are 2.3, 1.1 mg/kg (oral) and 6.2, 2.5 mg/kg (dermal).

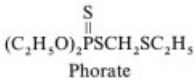
Phorate

Terbufos [*13071-79-9*], *O,O*-diethyl *S*-(1,1-dimethylethylthio)methyl phosphorodithioate is a yellow liquid, *bp* 69 °C (1.3 Pa), d_4^{24} 1.165, soluble in water to 10–15 mg/L. The LD_{50} values (rat, oral) are 4.5, 9.0 mg/kg. Terbufos is more persistent than phorate and is used as a granular soil treatment for the control of corn root worms.

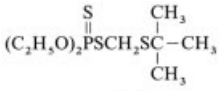
Terbufos

Dimethoate [*60-51-5*], *O,O*-dimethyl *S*-(*N*-methylcarbamoyl)methyl phosphorodithioate, *mp* 51–52 °C, d_4^{65} 1.277, is soluble in water to 70 g/L. The rat LD_{50} values are 215, 245 mg/kg (oral) and 610 mg/kg (dermal). Dimethoate is a widely used contact and systemic insecticide for vegetables, fruit, and ornamental plants.

Dimethoate

Dimethoate Analogues. *Omethoate* [*1113-02-6*], the P=O analogue of dimethoate, is an activated metabolite of dimethoate and a practical systemic insecticide that is especially effective against aphids and red spider mites. The LD_{50} (rat, oral) is 50 mg/kg.

Omethoate

Further dimethoate analogues include *formothion* [*2540-82-1*], *O,O*-dimethyl *S*-(*N*-formyl-*N*-methylcarbamoyl)methyl phosphorodithioate, *mp* 25 °C, LD_{50} (rat, oral) 275 mg/kg; *morphothion* [*144-41-2*], *O,O*-dimethyl *S*-(morpholinocarbonyl)methyl phosphorodithioate; and *vamidothion* [*2275-23-2*], *O,O*-dimethyl *S*-[2-(1-methylcarbamoyl)ethylthio]ethyl phosphorothioate, *mp* 46–48 °C, soluble in water to 40 g/L, LD_{50} (rat, oral) 64, 105 mg/kg.

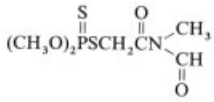
Formothion

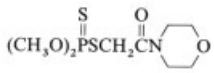
Morphothion

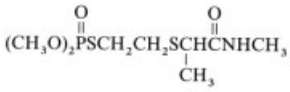
Vamidothion

Phenthoate [*2597-03-7*], *O,O*-dimethyl *S*-(α-carboethoxy)phenylmethyl phosphorodithioate, *mp* 17–18 °C, d_4^{20} 1.226 has LD_{50} values (rat, oral) of 300, 400 mg/kg. Phenthoate is used as a contact and systemic insecticide for tree fruit pests.

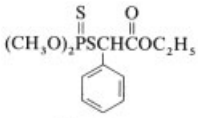

Phenthoate

Carbophenothion [*786-19-6*], *O,O,*-diethyl *S*-(4-chlorophenylthio)methyl phosphorodithioate, *bp* 82 °C (1.3 Pa), d_4^{20} 1.285, rat LD_{50} values 30, 10 mg/kg (oral) and 54, 27 mg/kg (dermal), and its *O,O*-dimethyl analogue **carbophenothion methyl** [*953-17-3*], rat LD_{50} values 98, 120 mg/kg (oral) and 215, 190 mg/kg (dermal), are persistent general-purpose contact insecticides used on fruit, vegetables, and forage.

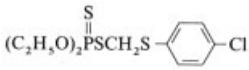

Carbophenothion

Malathion [*121-75-5*], *O,O*-dimethyl *S*-(1,2-dicarbethoxy)ethyl phosphorodithioate, *bp* 156–157 °C (93.3 Pa), is a yellow oil, d_4^{25} 1.23, vapor pressure 5.2 mPa at 30 °C, soluble in water to 145 mg/L. Malathion hydrolyzes easily above pH 7.0 and below pH 5.0. The rat LD_{50} values are 1375, 1000 mg/kg (oral) and $>$ 4 000 mg/kg (dermal). Malathion is a persistent general-purpose insecticide, and its relatively low toxicity makes it especially suitable in household, home garden, vegetable, and fruit insect control, as well as in control of insects for public health (e.g., flies, mosquitoes, and lice).

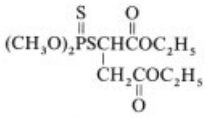

Malathion

Ethion [*563-12-2*], *O,O,O′,O′*-tetraethyl *S,S′*-methylene diphosphorodithioate, *bp* 165 °C (53.3 Pa), d_4^{20} 1.22, vapor pressure 0.2 mPa at 20 °C, rat LD_{50} values 65, 27 mg/kg (oral) and 245, 62 mg/kg (dermal), is an acaricide and insecticide used on fruit and vegetables.

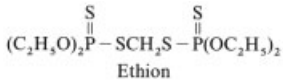

Ethion

Dioxathion [*78-34-2*], *O,O,O′,O′*-tetraethyl *S,S′*-(1,4-dioxane-2,3-diyl) diphosphorodithioate, d_4^{26} 1.26, rat LD_{50} 43, 23 mg/kg (oral) and 235, 63 mg/kg (dermal), is a general-purpose insecticide and acaricide for deciduous fruit.

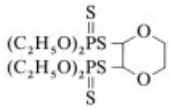

Dioxathion

6.4. Phosphorothioate Esters of Phenols

The phosphorothioate esters of phenols contain a high-energy phosphate bond–the pseudoanhydride bond of phosphoric acid and the phenol. The 4-nitrophenyl phosphate esters, the parathions, are the cheapest and most widely used OP insecticides, but they are also among the most toxic to nontarget organisms and humans. The hazard of human poisoning from the use of parathions has stimulated the development of safer analogues.

Parathion [*56-38-2*], *O,O*-diethyl *O*-4-nitrophenyl phosphorothioate is a dark-brown liquid with a garlic odor, *bp* 375 °C, d_4^{25} 1.265, and vapor pressure 4.9 mPa at 20 °C. Parathion is resistant to aqueous hydrolysis, the half-life is 120 d in water and 8 h in lime water. Above 130 °C, parathion isomerizes slowly to form *O,S*-diethyl *O*-4-nitrophenyl phosphorothiolate [*597-88-6*], which is a much less stable and less effective insecticide. Parathion is reduced readily to the nontoxic *O,O*-diethyl *O*-4-aminophenyl phosphorothioate and oxidized with difficulty to the very highly toxic paraoxon *O,O*-diethyl *O*-4-nitrophenyl phosphorothiolate [*311-45-5*]. Parathion is very soluble in aromatic petroleum solvents, very slightly soluble in kerosene and mineral oil, and soluble in water to 0.2 mg/L. The rat LD_{50} values are 13, 3.6 mg/kg (oral) and 21, 6.8 mg/kg (dermal).

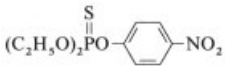

Parathion

Methyl parathion [*298-00-0*], *O,O*-dimethyl *O*-4-nitrophenyl phosphorothioate, *mp* 36 °C, d_4^{20} 1.358, has very similar properties to parathion but hydrolyzes and isomerizes more readily. The rat LD_{50} values are 14, 24 mg/kg (oral) and 67 mg/kg (dermal). Because of its lower acute toxicity, methyl parathion is used more widely than parathion.

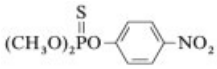

Methyl parathion

Fenitrothion [*122-14-5*], *O,O*-dimethyl *O*-(3-methyl-4-nitrophenyl) phosphorothioate, a brownish oil, *bp* 140 – 145 °C (13.3 Pa), d_4^{25} 1.322, is soluble in aromatic hydrocarbons and in water to 20 mg/L. The rat LD_{50} values are 740, 570 mg/kg (oral) and 300, 400 mg/kg (dermal). Fenitrothion is a general-purpose insecticide with about the same spectrum of activity as methyl parathion. It is also safe for residual house spraying in malaria control.

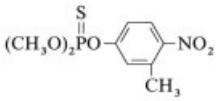

Fenitrothion

Dicapthon [*2463-84-5*], *O,O*-dimethyl *O*-(2-chloro-4-nitrophenyl) phosphorothioate, *mp* 51 °C, rat LD_{50} 400, 330 mg/kg (oral) and 790, 1250 mg/kg (dermal), and **chlorthion** [*500-28-7*], *O,O*-dimethyl *O*-(3-chloro-4-nitrophenyl) phosphorothioate, *mp* 21 °C, d_4^{20} 1.437, rat LD_{50} 880, 980 mg/kg (oral) and 4 500, 4 100 mg/kg (dermal), are two safer analogues of methyl parathion that have been used as household insecticides.

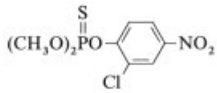

Dicapthon

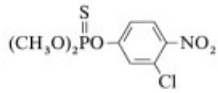

Chlorthion

Ronnel [*299-84-3*], fenchlorphos, *O,O*-dimethyl *O*-2,4,5-trichlorophenyl phosphorothioate, *mp* 35 – 37 °C, is soluble in petroleum hydrocarbons and in water to 80 mg/L. The LD_{50} values (rat, oral) are 1 250, 2 630 mg/kg. Ronnel is a safe, persistent insecticide that is used in veterinary hygiene especially to control cattle grubs. It is too phytotoxic for agricultural use.

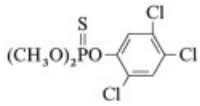

Ronnel

Bromophos [*2104-96-3*], *O,O*-dimethyl *O*-(2,5-dichloro-4-bromophenyl) phosphorothioate, *mp* 54 °C, vapor pressure 17 mPa at 20 °C, water solubility 40 mg/L, rat LD_{50} 1 600, 1 730 mg/kg (oral) and > 5 000 mg/kg (dermal), and **iodofenphos** [*18181-70-9*], *O,O*-dimethyl *O*-(2,5-dichloro-4-iodophenyl) phosphorothioate, *mp* 76 °C, water solubility 20 mg/L, LD_{50} (rat, oral) 2100 mg/kg, are very closely related to ronnel.

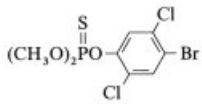

Bromophos

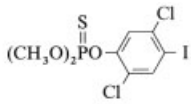

Iodofenphos

They are relatively safe used in veterinary hygiene and as household insecticides.

Fenthion [*55-38-9*], *O,O*-dimethyl*O*-(3-methyl-4-methylthiophenyl) phosphorothioate, is a colorless oil, *bp* 105 °C (1.3 Pa), d_4^{20} 1.250, vapor pressure 3.9 mPa at 20 °C, which is soluble in water to 55 mg/L. The rat LD_{50} values are 215, 245 mg/kg (oral) and 330 mg/kg (dermal). Fenthion is a persistent insecticide used in veterinary hygiene and as a mosquito larvicide; it is especially effective in polluted water. It has also been used for bird pest control.

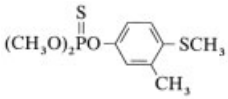

Fenthion

Fensulfothion [*115-90-2*], *O,O*-diethyl *O*-(4-methylsulfinylphenyl) phosphorothioate, *bp* 138 – 141 °C (1.3 Pa), rat LD_{50} 4.1, 1.8 mg/kg (oral) and 19, 4.1 mg/kg (dermal), is used as a soil insecticide and nematocide.

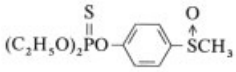

Fensulfothion

Temephos [*3383-96-8*], *O,O,O′,O′*-tetramethyl *O,O*-(thiodi-4,1-phenylene) phosphorodithioate, *mp* 30 °C, d_4^{20} 1.32, water solubility 0.025 mg/L, rat LD_{50} 8 600, 13 000 mg/kg (oral) and > 4 000 mg/kg (dermal), is used as a larvicide for blackflies and mosquitoes.

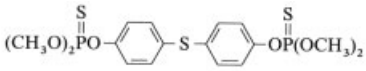

Temephos

Coumaphos [*56-72-4*], *O,O*-diethyl-*O*-3-chloro-4-methyl-2-oxo-2*H*-1-benzopyran-7-yl phosphorothioate, *mp* 91 °C, vapor pressure 0.013 mPa at 20 °C, rat LD_{50} 41, 16 mg/kg (oral) and 1000 mg/kg (dermal), is applied as a 0.25 – 0.5 % spray to control animal parasites and cattle grubs.

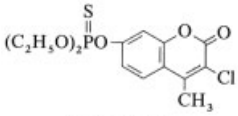

Coumaphos

Isofenphos [*25311-71-1*], *O*-ethyl *O*-[2-(isopropoxycarbonyl)phenyl] *N*-isopropyl phosphoroamidothioate, *bp* 120 °C (1.3 Pa), d_4^{20} 1.13, vapor pressure 0.52 mPa at 20 °C, water solubility 24 mg/L, LD_{50} (rat, oral) 28 mg/kg, is a soil insecticide.

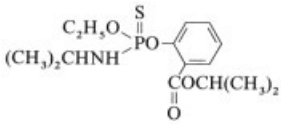

Isofenphos

Sulprofos [*35400-43-2*], *O*-ethyl *O*-[(4-methylthio)phenyl] *S*-propyl phosphorodithioate, *bp* 155 – 158 °C (13.3 Pa), d_4^{20} 1.20, LD_{50} (rat, oral) variously given as 22.7, 60, and 227 mg/kg, is used for control of lepidopterous larvae that attack cotton.

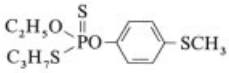

Sulprofos

6.5. Phosphonothioate Esters of Phenols

A number of relatively persistent *O*-alkyl *O*-aryl phosphorothioates have strong insecticidal properties. Some produce delayed neuropathy in hens and in humans, and have been withdrawn from use.

EPN [*2104-64-5*], *O*-ethyl *O*-(4-nitrophenyl) phenylphosphonothioate, *mp* 41 °C, d_4^{20} 1.268, rat LD_{50} 36, 7.7 mg/kg (oral) and 230, 25 mg/kg (dermal), is used as a cotton insecticide, but is a delayed neurotoxin.

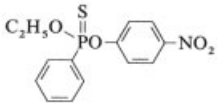

EPN

Cyanofenphos [*13067-93-1*], *O*-ethyl *O*-(4-cyanophenyl) phenylphosphonothioate, *mp* 83 °C, vapor pressure 1.7 mPa at 25 °C, water solubility 6 mg/L, has an LD_{50} (rat, oral) of 44 mg/kg but is a delayed neurotoxin.

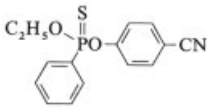

Cyanofenphos

Leptophos [*21609-90-5*], *O*-methyl *O*-(4-bromo-2,5-dichlorophenyl) phenylphosphonothioate, *mp* 71 – 72 °C, d_4^{25} 1.53, soluble in water to 0.03 mg/L, LD_{50} (rat, oral) 53, 45 mg/kg, is a persistent insecticide that is especially effective against lepidopterous larvae, but it is a delayed neurotoxin.

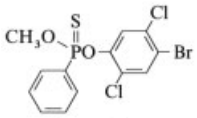

Leptophos

Fonofos [*944-22-9*], *O*-ethyl *S*-phenyl ethylphosphonodithioate, *bp* 130 °C (13.3 Pa), d_4^{25} 1.16, water solubility 13 mg/L, LD_{50} (rat, oral) 8, 17 mg/kg, is a persistent soil insecticide.

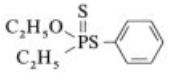

Fonofos

Trichlornate [*327-98-0*], *O*-ethyl *O*-2,4,5-trichlorophenyl ethylphosphonothioate, LD_{50} (rat, oral) 37.5 mg/kg, has been used as a soil insecticide but is a delayed neurotoxin.

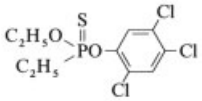

Trichlornate

6.6. Phosphorothioate Esters of Heterocyclic Enols

Diazinon [*333-41-5*], *O,O*-diethyl *O*-[(2-isopropyl-4-methyl)pyrimidin-6-yl] phosphorothioate, is one of the oldest broad-spectrum organophosphorus insecticides. Diazinon is a brownish oil, *bp* 83 – 84 °C (0.3 Pa), d_4^{20} 1.116, vapor pressure 18.2 mPa at 20 °C, soluble in water to 40 mg/L. The rat LD_{50} values are 250, 285 mg/kg (oral), and the rabbit LD_{50} (dermal) is > 1 000 mg/kg. Diazinon is used as a soil insecticide and to control agricultural pests.

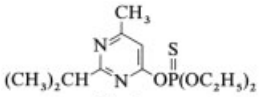

Diazinon

Chlorpyrifos [*2921-88-2*], *O,O*-diethyl*O*-(3,5,6-trichloro-2-pyridinyl) phosphorothioate, *mp* 42 – 43 °C, vapor pressure 2.5 mPa at 25 °C, water solubility 2 mg/L, rat LD_{50} 135, 165 mg/kg (oral) and 2000 mg/kg (dermal), is used widely as a soil insecticide, in veterinary hygiene, and as a household insecticide for cockroach and termite control.

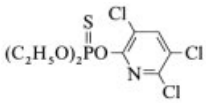

Chlorpyrifos

Chlorpyrifos-methyl [*5598-13-0*], the *O,O*-dimethyl analogue of chlorpyrifos, LD_{50} (rat, oral) 941, 2140 mg/kg, is much safer and is used to control blackfly and mosquito larvae.

Fospirate [*5598-52-7*], dimethyl 3,5,6-trichloro-2-pyridinyl phosphate, *mp* 87 °C, LD_{50} (rat, oral) 869 mg/kg, is the P = O analogue of chlorpyrifos and is used as a veterinary parasiticide.

Pirimiphos-ethyl [*23505-41-1*], **pirimithate,** *O,O*-diethyl*O*-(2-diethylamino-6-methyl-4-pyrimidinyl) phosphorothioate, *bp* 128 – 132 °C (5.3 Pa), *d* 1.165, LD_{50} (rat, oral) 140, 200 mg/kg, is used as an agricultural insecticide and acaricide.

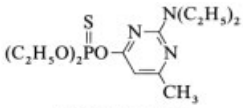

Pirimiphos-ethyl

Pirimiphos-methyl [*29232-93-7*], *O,O*-dimethyl *O*-(2-diethylamino-6-methyl-4-pyrimidinyl) phosphorothioate, d_4^{30} 1.157 rat LD_{50} 2 050 mg/kg (oral) and > 2 000 mg/kg (dermal), the *O,O*-dimethyl analogue of pirimiphos-ethyl is a much safer insecticide

used to control aphids, insects in stored products and for veterinary hygiene.

6.7. Phosphorothioate Esters of S-Methyl Heterocycles

Azinphos-methyl [*86-50-0*], *O,O*-dimethyl *S*-(3,4-dihydro-4-oxobenzo-[*d*]-(1,2,3)-triazin-3-yl methyl) phosphorodithioate, *mp* 73 – 74 °C, d_4^{20} 1.44, water solubility 30 mg/L, has rat LD_{50} values of 13, 11 mg/kg (oral) and 220 mg/kg (dermal). The compound is a persistent insecticide used extensively to control the codling moth and other pests of deciduous fruit, as well as those of vegetables, ornamentals, and cotton.

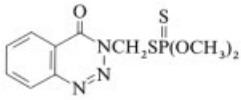

Azinphos-methyl

Azinphos-ethyl [*2642-71-9*], *mp* 53 °C, d_4^{20} 1.284, vapor pressure 0.029 mPa at 20 °C, LD_{50} (rat, oral) 17.5 mg/kg, is the diethyl analogue of azinphos-methyl and is also a commercial insecticide.

Phosmet [*732-11-6*], *O,O*-dimethyl *S*-(*N*-phthalimidomethyl) phosphorodithioate, *mp* 72 °C, vapor pressure 0.13 Pa at 50 °C, water solubility 24 mg/L, has rat LD_{50} values of 113, 160 mg/kg (oral) and > 1 500 mg/kg (dermal). It is used to control pests of deciduous fruit.

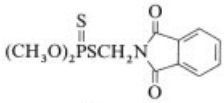

Phosmet

Dialifor [*10311-84-9*], *O,O*-diethyl *S*-(2-chloro-1-phthalimidoethyl) phosphorodithioate, *mp* 67 – 69 °C, rat oral LD_{50} 43, 71 mg/kg and rabbit dermal LD_{50} 145 mg/kg, is used to control pests of deciduous fruit.

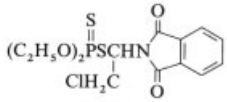

Dialifor

Phosalone [*2310-17-0*], *O,O*-diethyl *S*-(6-chloro-2,3-dihydro-2-oxobenzoxazol-3-yl)methyl phosphorodithioate, *mp* 48 °C, LD_{50} (rat, oral) 120, 175 mg/kg, is an insecticide and acaricide for tree fruit pests.

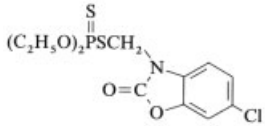

Phosalone

Methidathion [*950-37-8*], *O,O*-dimethyl *S*-[5-methoxy-2-oxo-1,3,4-thiadiazol-3(2*H*)-yl]methyl phosphorodithioate, *mp* 39 – 40 °C, LD_{50} (rat, oral) 25, 48 mg/kg, is used as an insecticide and acaricide for fruit and vegetables.

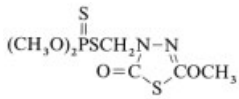

Methidathion

Azamethiphos [*35575-96-3*], *O,O*-dimethyl *S*-[(6-chloro-2-oxooxazolo[4,5-*b*]pyridin-3(2*H*)-3-yl)methyl] phosphorodithioate, *mp* 89 – 90 °C, LD_{50} (rat, oral) 1 100, 1 750 mg/kg, is used for household pest control.

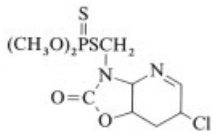

Azamethiphos

Menazon [*78-57-9*], *O,O*-dimethyl *S*-(4,6-diamino-1,3,5-triazin-2-yl)methyl phosphorodithioate, *mp* 160 – 162 °C, vapor pressure 0.13 mPa at 25 °C, water solubility 2.4 g/L, LD_{50} (rat, oral) 1 950 mg/kg, is a systemic aphicide.

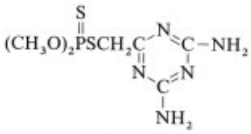

Menazon

Endothion [*2778-04-3*], *O,O*-dimethyl *S*-[(5-methoxy-oxo-4*H*-pyran-2-yl)methyl]phosphorothioate, *mp* 96 °C, LD_{50} (rat, oral) 30, 50 mg/kg, is a systemic insecticide.

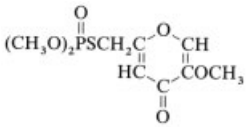

Endothion

6.8. Miscellaneous Organophosphorus Esters

Crufomate [*299-86-5*], *O*-methyl *O*-[2-chloro-4-(1,1-dimethylethyl)phenyl] methylphosphoramidate, *mp* 61 °C, LD_{50} (rat, oral) 635, 460 mg/kg, is an animal systemic insecticide. It is used primarily for cattle grub control either as a 0.75 % spray or when fed at 20 – 25 mg/kg body weight.

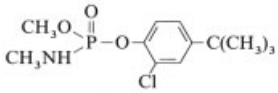

Crufomate

Phoxim [*14816-18-3*], *O,O*-diethyl*O*-(α-cyanobenzilideneamino) phosphorothioate, *bp* 102 °C (1,3 Pa), d_4^{20} 1.176, LD_{50} (rat, oral) 2 500 mg/kg, is especially useful for control of stored product pests.

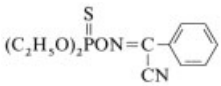

Phoxim

Methamidophos [*10265-92-6*], *O,S*-dimethyl phosphoramidothioate, *mp* 39 – 41 °C, $d_4^{44.5}$ 1.31, vapor pressure 39 mPa at 30 °C, LD_{50} (rat, oral) 19, 21 mg/kg, is a contact and systemic insecticide for insect pests of cotton, potato, and vegetable crops.

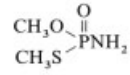

Methamidophos

Acephate [*30560-19-1*], *O,S*-dimethyl *N*-acetyl phosphoramidothioate is the *N*-acetyl derivative of methamidophos, *mp* 72 – 80 °C, water soluble to 650 g/L, LD_{50} (rat, oral) 866, 945 mg/kg. It has much lower acute toxicity than methamidophos, and is effective against a wide variety of agricultural and household insect pests.

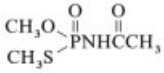

Acephate

Ethoprophos [*13194-48-4*], Prophos, Ethoprop, *O*-ethyl *S,S*-dipropyl phosphorodithioate, *bp* 86 – 91 °C (26.6 Pa), d_4^{20} 1.094, vapor pressure 45 mPa at 26 °C, soluble in water to 750 mg/L, rat LD_{50} 61 mg/kg (oral) and 26 mg/kg (dermal), is used to control insect pests and nematodes in vegetables and ornamentals.

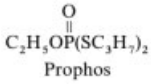

Prophos

6.9. Mode of Action and Toxicity [60–66]

6.9.1. Inhibition of Acetylcholinesterase

The biological activity of the organophosphorus insecticides results from the capacity of the central phosphorus atom to phosphorylate the active site of acetylcholinesterase (E.C. 3.1.1.7), which is an essential enzyme of invertebrates and vertebrates. The phosphorylated enzyme is irreversibly inhibited and, therefore, cannot perform its normal function of rapidly hydrolyzing and removing the neurotransmitter acetylcholine from the nerve synapse. Acetylcholine

consequently accumulates and disrupts the normal functioning of the nervous system, producing typical cholinergic symptoms (i.e., hyperactivity, tremors, convulsions, paralysis, and death). In higher animals and in humans, these cholinergic effects are translated into muscarinic effects (e.g., nausea, salivation, lachrymation, and myosis of the pupil of the eye); nicotinic effects (e.g., muscular fasiculations); and central effects (e.g., giddiness, tremulousness, convulsions, and coma).

The reaction between the esterase and the phosphorus inhibitor involves attack on the electrophilic phosphorus atom by a nucleophilic serine hydroxyl group with the neighboring imidazole ring of a histidine residue as the active site of acetylcholinesterase. The normal three-step reaction (A) that occurs at the synapse between acetylcholinesterase (EH) and the substrate acetylcholine is

(A)

$$CH_3\overset{O}{\overset{\|}{C}}OCH_2CH_2\overset{+}{N}(CH_3)_3 + EH \underset{(1)}{\overset{K_E}{\rightleftharpoons}}$$

Acetylcholine

$$CH_3\overset{O}{\overset{\|}{C}}OCH_2CH_2\overset{+}{N}(CH_3)_3 \cdots EH \xrightarrow[(2)]{k_1}$$

$$CH_3\overset{O}{\overset{\|}{C}}E + HOCH_2CH_2\overset{+}{N}(CH_3)_3 \xrightarrow[(3)\ H_2O]{k_2} CH_3\overset{O}{\overset{\|}{C}}OH + EH$$

In the normal process (A), step 3 occurs very rapidly and step 1 is the rate-determining step. However, in the inhibition process (B), step 3 is rate-determining because it occurs very slowly (from a matter of hours with *O,O*-dimethyl phosphates to days with *O*,*O*-diethyl phosphates).

(B)

$$(RO)_2\overset{O}{\overset{\|}{P}}X + EH \underset{(1)}{\overset{K_I}{\rightleftharpoons}} (RO)_2\overset{O}{\overset{\|}{P}}X \ldots EH \xrightarrow[(2)]{k_1}$$

$$(RO)_2\overset{O}{\overset{\|}{P}}E + HX \xrightarrow[(3)\ H_2O]{k_2} (RO)_2\overset{O}{\overset{\|}{P}}OH + EH$$

The organophosphorus insecticides may, therefore, be regarded as synthetic neurotransmitters with very slow turnover numbers. In vitro, the organophosphorus insecticide (e.g., tetraethyl pyrophosphate, mevinphos, or dichlorvos) engages in a first-order reaction with acetylcholinesterase. The inhibited phosphorylated enzyme is relatively stable and contains 1 mol of phosphorus per mole of enzyme. An equimolar quantity of alcoholic, phenolic, or acidic product HX is liberated as the leaving group.

Delay Factor. The reactivity of organophosphorus insecticides is determined by the electrophilicity of the phosphorus atom, the strength of the "high-energy" P—X bond, and steric effects of the substituents. The electrophilicity of the central phosphorus is determined by the relative electronegativities of the atoms bonded to it (e.g., P 2.1, O 3.5, S 2.5, N 3.0, and C 2.5). Therefore, the phosphorus in a phosphate ester (P=O) is much more electrophilic than the oxygen, and these compounds are much more reactive than the phosphorothioate esters (P=S). The latter are generally so stable as to be unreactive with acetylcholinesterase. Their insecticide activity and general toxicity result from in vivo oxidation by a microsomal oxidase in the insect gut and fat body tissue and in the mammalian liver. A typical example is the oxidation of parathion to paraoxon [*311-45-5*]:

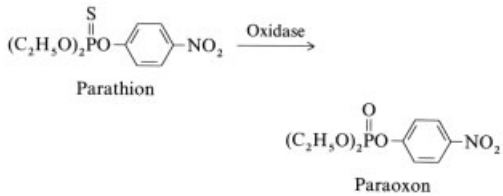

The time required for this in vivo oxidation or activation (usually several hours) is called the delay factor and is important in human toxicology because at least a portion of the insecticide can be detoxified and eliminated during this period. The delay factor also provides opportunity for therapy with atropine which protects the postsynaptic nerve membrane from the accumulating acetylcholine. Because of the delay factor, phosphorothionates (P=S) are almost always less toxic and less hazardous to the user than their equivalent phosphates (P=O).

The substituents on phosphorus strongly affect its electrophilic character by inductive and mesomeric effects. If the substituents are electron-attracting, the phosphorus atom becomes more reactive and the insecticide more effective. For the simple *O,O*-dimethyl and *O,O*-diethyl phenyl phosphates and phosphorothioates of the parathion – paraoxon type, a very

high degree of correlation exists between the quantitative measure of electron-withdrawing power of the substituent (Hammett's σ) and anticholinesterase activity, and toxicity [58], [59]. This explains the high insecticidal activity of molecules with 4-NO_2 (methyl parathion), 4-CN, 4-CH_3SO (fensulfothion), and 2,4,5-Cl (ronnel) phenyl leaving groups.

Activation. In vivo, microsomal oxidation is also important for activation or lethal synthesis in which electron-donating or inert groups (e.g., CH_3S) are converted to strongly electron-withdrawing groups (e.g., CH_3SO and CH_3SO_2). Thus, insecticides such as fenthion are activated in vivo by microsomal oxidation to produce strongly electrophilic phosphorus atoms. This type of reaction involves a built-in delay that is particularly effective in certain systemic insecticides, such as phorate and disulfoton, where both P=S and C_2H_5S oxidation occur in plant tissues:

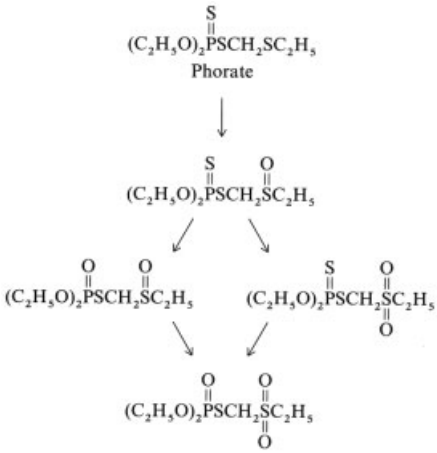

The originial insecticide is relatively lipophilic and stable. It is absorbed readily and translocated through plant tissue without destructive hydrolysis until microsomal oxidation occurs. The activated (oxidized) insecticide is highly toxic to sucking insects and relatively unstable so that it hydrolyzes to nontoxic products. Microsomal oxidation can also take place in the mammalian liver, which makes many of these systemic insecticides extremely toxic to humans.

6.9.2. Delayed Neuropathy [57], [65], [66]

Certain organophosphorus insecticides [especially phosphonates and phosphonothionates with P(O)C bonds] phosphorylate an enzyme neurotoxic esterase to produce a delayed neuropathy that is essentially irreversible and may occur from one to four weeks after acute or chronic exposure. Neurotoxic esterase is evidently required for growth and maintenance of the long axons of the central nervous system. After its inhibition, degenerative changes develop in the long axons and are manifest by tingling and burning sensations in the limb extremities that progress to weakness in the lower limbs and to ataxia. The final stage is irreversible paralysis which, in severe poisoning, may affect upper limbs as well. Recovery is very slow and seldom complete in adults, with flaccid paralysis changing to spasticity. The phosphorylated esterase is modified by "aging" so that the remaining ester bond of the organophosphorus inhibitor attached to the active site of the neurotoxic esterase is cleaved to form a negatively charged residue bound to the enzyme.

Delayed neuropathy was originally observed in humans after excessive exposure to tri-*o*-cresyl phosphate and has been produced by mipafox, EPN, leptophos, and methamidophos. Many other related compounds such as cyanofenphos, trichlormat, and isofenphos produce delayed neuropathy in hens.

6.9.3. Selective Toxicity

The organophosphorus insecticides are generally highly toxic to nontarget organisms, and their use can be hazardous to humans. Human poisoning can result from acute inhalation, dermal exposure, or oral ingestion. Many of the insecticides are nearly as toxic when applied to the skin as when ingested (cf. oral and dermal LD_{50} values). Fatalities have resulted from ingestion of as little as 2 mg of parathion in small children to 120 mg in adults. Some of the most acutely toxic organophosphorus insecticides (e.g., phorate, disulfoton, and terbufos) are marketed only as granular or pelletized products that are applied directly to the soil and thus reduce the toxic hazard.

Acetylcholinesterase inhibition in the blood and central nervous system of humans is cumulative and can increase to fatal levels upon repeated chronic exposure of agricultural workers. This is due to the irreversible inhibition of acetylcholinesterase with *O,O*-diethyl phosphates. Therefore this vital enzyme may have to be resynthesized. Hazards resulting from use of the more toxic organophosphorus insecticides (Class I poisons with $LD_{50} < 50$ mg/kg) can be reduced by wearing protective clothing including mask, goggles, and rubber gloves. Workers with frequent exposure should have regular determinations of blood acetylcholinesterase levels and should be removed from exposure, when these blood acetylcholinesterase levels show progressive decrease. In cases of severe exposure and incipient cholinergic symptoms, atropine should be administered promptly and the patient should be hospitalized.

Selectivity is implicit in the definition of an insecticide as " an agent for destroying insects ". Nevertheless, many highly toxic organophosphorus compounds are still used in massive quantities throughout the world. Their selectivity is apparent only through careful application and confinement to the area of treatment. Selectivity is important in providing safety not only for the user of pesticides and domestic animals, but also for wildlife and beneficial insects: honey bees and other pollinators, parasites, and predators.

The development of malathion in 1950 was an important milestone in the use of selective insecticides. Malathion is about one-half to one-twentieth as toxic as parathion to insects, but it is only about one one-hundredth as toxic to humans and higher animals. The use of thousands of metric tons of malathion over several decades has demonstrated substantial safety to users, coupled with effectiveness in insect pest control.

A high degree of selectivity has been demonstrated by incorporating chlorine or methyl groups in the 3-position of the aryl ring of parathion. Such groups interact stereochemically with acetylcholinesterase, increasing the affinity for the insect enzyme and decreasing it for the mammalian enzyme. Such compounds, including fenitrothion, fenthion, ronnel, bromophos, and iodofenphos, have been used widely and safely in household, animal health, and public health pest control. Safety is also enhanced by differences in the rates of microsomal oxidation of P=S to P=O (see Section 6.9.1) and in hydrolytic detoxification between insects and mammals.

The user of OP insecticides can enhance selectivity and safety by choosing only compounds with LD_{50} values > 100 mg/kg, by using *O,O*-dimethyl instead of *O,O*-diethyl insecticides, and by using P=S in preference to P=O insecticides. Great care should be exercised in handling, applying, and storing all organophosphorus insecticides.

6.10. Environmental Toxicology

Organophosphorus insecticides are reactive and are degraded readily by oxidation and hydrolysis in vivo. Therefore, the use of organophosphorus esters does not result in serious problems of bioaccumulation and food chain transfer. Soil persistence is low, with typical half-life values ranging from one or two weeks for malathion to three to six months for parathion, diazinon, and azinphos-methyl. However, organophosphorus insecticides are general biocides and are toxic to nearly all animals. Oral LD_{50} values, in milligrams per kilogram, are phorate, mallard 0.62, pheasant 7.1; parathion, mallard 2.0, pheasant 12; monocrotophos, mallard 4.8, pheasant 2.8; diazinon, mallard 3.5, pheasant 4.3; fenthion, mallard 5.9, pheasant 18; chlorpyrifos, mallard 75, pheasant 12; and azinphos-methyl, mallard 136, pheasant 75. The organophosphorus compounds are generally toxic to fish, and LD_{50} values, in parts per million, are phorate, bluegill 0.0055; parathion, bluegill 0.10, trout 0.047; diazinon, bluegill 0.052, trout 0.38; malathion, bluegill 0.11, trout 0.13; fenthion, bluegill 1.4, trout 0.93; and azinphos-methyl, bluegill 0.005, trout 0.004. Organophosphorus insecticides are highly toxic to bees, beneficial parasites, and predators.

7. Carbamate Insecticides [67–71]

The carbamates are synthetic analogues of the toxic alkaloid physostigmine [*57-47-6*] isolated from the calabar bean *Physostigma venenosum*

in 1864. Its structure was elucidated by STEDMAN and BARKER (1925), and its cholinergic nature was discovered by LOEWI and NAWRATIL in 1926. STEDMAN found that 3-dimethylaminophenyl methylcarbamate methiodide was a highly effective cholinergic drug in aqueous solution. The more stable 3-dimethylaminophenyl dimethylcarbamate methiodide was developed as the drug prostigmine. Because of their positively charged quaternary structure, prostigmine and its medicinally effective relatives cannot penetrate the lipid sheath of insect nerves and are noninsecticidal. GYSIN (1954) developed the *N,N*-dimethylcarbamate of heterocyclic enols as insecticides [72]. KOLBEZEN, METCALF, and FUKUTO (1954) produced the first *N*-methylcarbamates of phenols that have a broader spectrum of activity and are now the most widely used carbamate insecticides [73]. Subsequently, *N*-methylcarbamates of a variety of oximes were found to be extraordinarily toxic and have been used extensively as systemic insecticides.

Carbamates were the third major group of synthetic organic insecticides, developed after organochlorine and organophosphorus compounds. Because of their high degree of biodegradability, they are free from the persistence and bioaccumulation associated with organochlorines. Carbamates, although often highly toxic to humans and higher animals, are reversible inhibitors of acetylcholinesterase and much safer to use than the irreversible organophosphorus inhibitors. Additionally, carbamates provided a welcome alternative to the serious problem of insect pest resistance that had developed with organochlorines and organophosphates. Therefore, carbamates have become widely used for soil insects and in foliar application to field and vegetable crops. Approximately 11×10^6 kg (active ingredient) were applied to U.S. farm crops in 1976. Carbaryl production was estimated at 25×10^6 kg in 1971, and global consumption of carbamate insecticides has been estimated to be as high as 35×10^6 kg. Propoxur was developed as a residual insecticide for controlling adult mosquitoes in human habitations and has been used as a replacement for DDT.

7.1. Dimethylcarbamates of Heterocyclic Enols

Isolan [*119-38-0*], 1-isopropyl-3-methyl-5-pyrazoyl dimethylcarbamate, is a colorless liquid, *bp* 105 – 107 °C (40 Pa), LD_{50} (rat, oral), 54 mg/kg, used as a systemic aphicide.

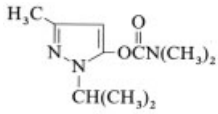

Isolan

Dimetilan [*644-64-4*], 2-dimethylcarbamoyl-3-methyl-5-pyrazolyl dimethylcarbamate, *mp* 68 – 71 °C, is a water-soluble carbamate used in sugar bait for housefly control. The rat LD_{50} values are 25, 64 mg/kg (oral) and 600 mg/kg (dermal).

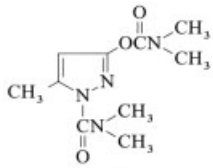

Dimetilan

Pirimicarb [*23103-98-2*], 2-(dimethylamino)-5,6-dimethyl-4-pyrimidinyl dimethylcarbamate, *mp* 90.5 °C, vapor pressure 4 mPa at 30 °C, is water-soluble to 0.25 %. The rat LD_{50} values are 147 mg/kg (oral) and > 500 mg/kg (dermal). Pirimicarb is a systemic aphicide used largely on grain crops.

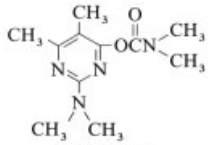

Pirimicarb

7.2. Methylcarbamates of Phenols

Aminocarb [*2032-59-9*], (4-dimethylamino-3-methyl)phenyl methylcarbamate, *mp* 93 °C, rat LD_{50} 40, 38 mg/kg (oral) and 280, 320 mg/kg (dermal), is particularly effective against lepidopterous larvae, snails, and slugs.

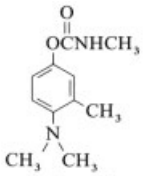

Aminocarb

Bendiocarb [*22781-23-3*], 2,2-dimethyl-1,3-benzodioxol-4-yl methylcarbamate, *mp* 129–130 °C, vapor pressure 0.7 mPa at 25 °C, is water-soluble to 0.04 g/L. The rat LD_{50} values are 179 mg/kg (oral) and 1000 mg/kg (dermal). Bendiocarb is used to control cockroaches, other household pests, and soil insects.

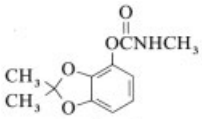

Bendiocarb

BPMC [*3766-81-2*], 2-(1-methylpropyl)-phenyl methylcarbamate, *mp* 32 °C, *bp* 112–113 °C (40 Pa), d_4^{20} 1.050, 48 mPa at 20 °C, LD_{50} (rat) 410 mg/kg, is used to control leafhoppers and plant hoppers on rice.

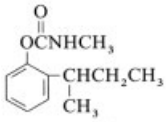

BPMC

Butacarb [*2655-19-8*], 3,5-di-*tert*-butylphenyl methylcarbamate, *mp* 97–99 °C, LD_{50} (rat, oral) > 4 000 mg/kg, has been used in veterinary entomology to control sheep blowfly larvae in wounds.

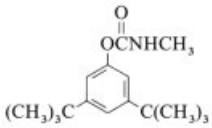

Butacarb

Carbaryl [*63-25-2*], 1-naphthyl methylcarbamate, *mp* 142 °C, vapor pressure < 5 mPa at 20 °C, is soluble in water to 0.12 g/L. The rat LD_{50} values are 540 mg/kg (oral) and > 4 000 mg/kg (dermal). Carbaryl is a broad-spectrum insecticide registered on more than 100 crop plants in the United States. It is detoxified and eliminated rapidly in animals, and is neither concentrated in fat nor secreted in the butterfat of milk. Therefore, carbaryl is favored for application to food crops.

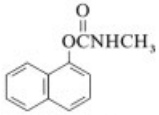

Carbaryl

Carbofuran [*1563-66-2*], 2,3-dihydro-2,2-dimethylbenzofuranyl methylcarbamate, *mp* 150–152 °C, vapor pressure 2.7 mPa at 33 °C, is soluble in water to 0.7 g/L. The rat LD_{50} values are 4.8 mg/kg (oral) and > 10 000 mg/kg (dermal). Carbofuran is a broad-spectrum insecticide used largely to control soil insects.

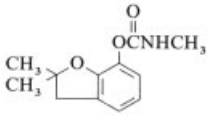

Carbofuran

Dioxacarb [*6988-21-2*], 2-(1,3-dioxolan-2-yl)phenyl methylcarbamate, *mp* 114–115 °C, vapor pressure 0.05 mPa at 20 °C, is soluble in water to ca. 6 g/L. The rat LD_{50} values are 125 mg/kg (oral) and 3 000 mg/kg (dermal). Dioxacarb is used to control cockroaches and a variety of household or stored product insects.

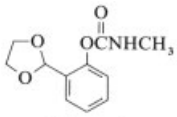

Dioxacarb

Ethiofencarb [*29973-13-5*], 2-(ethylthiomethyl)phenyl methylcarbamate, is a yellowish oil, d_4^{20} 1.147, vapor pressure 13 mPa at 30 °C, water solubility 1.8 g/L. The rat LD_{50} values are 411, 499 mg/kg (oral) and > 1 150 mg/kg

(dermal). Ethiofencarb is a systemic insecticide effective against aphids.

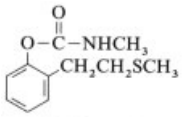

Ethiofencarb

Etrofol [*3942-54-9*], CPMC, 2-chlorophenyl methylcarbamate, *mp* 87 °C, LD_{50} (rat, oral) 648 mg/kg, LD_{50} (mouse) 150 mg/kg, is used to control leafhoppers and plant hoppers in rice.

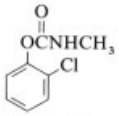

Etrofol

Formetanate [*22259-30-9*], *N,N*-dimethyl-*N′*-{3-[(methylamino)carbonyl]oxy}phenyl methanimidamide, 3-dimethylaminomethyleneiminophenyl methylcarbamate, vapor pressure 0.027 mPa at 25 °C, water solubility 6.3 g/L, rat LD_{50} 15 mg/kg (oral) and $>$ 5 60 house flies

0 mg/kg (dermal), is marketed as the water-miscible hydrochloride and used as an acaricide for deciduous fruit.

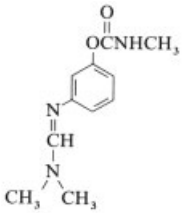

Formetanate

Isoprocarb [*2631-40-5*], 2-(1-methylethyl)-phenyl methylcarbamate, *mp* 93 °C, LD_{50} (rat, oral) 403, 485 mg/kg, is used to control pests of rice and cacao.

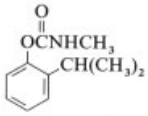

Isoprocarb

Metalkamate [*8065-36-9*], bufencarb, a mixture of three parts 3-(1-methylbutyl)phenyl carbamate and one part 3-(1-ethylpropyl)phenyl methylcarbamate, *mp* 26 – 36 °C, LD_{50} (rat, oral) 87, 170 mg/kg, has been used extensively as a soil insecticide.

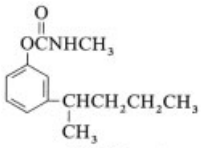

Metalkamate

Methiocarb [*2032-65-7*], (3,5-dimethyl-4-methylthio)phenyl methylcarbamate, *mp* 121 °C, 15 mPa at 25 °C, rat LD_{50} 70, 60 mg/kg (oral) and $>$ 2 000 mg/kg (dermal), is used to control insect pests of fruit and vegetables.

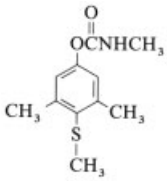

Methiocarb

Mexacarbate [*315-18-4*], (4-dimethylamino-3,5-dimethyl)phenyl methylcarbamate, *mp* 85 °C, is soluble in water to 0.1 g/L. Rat LD_{50} values are 37, 25 mg/kg (oral) and 1 500 mg/kg (dermal). Mexacarbate has been used principally to control lepidopterous forest pests.

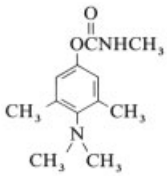

Mexacarbate

MPMC [*2425-10-7*], 3,4-dimethylphenyl methylcarbamate, *mp* 79 – 80 °C, vapor pressure 70 mPa at 25 °C, water solubility 0.58 g/L, LD_{50} (rat, oral) 380 mg/kg, has been used to control leafhoppers and plant hoppers that attack rice.

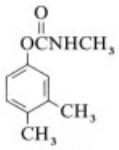
MPMC

MTMC [*1129-41-5*], 3-methylphenyl methylcarbamate, *mp* 76 – 77 °C, 145 mPa at 25 °C, LD_{50} (rat, oral) 268 mg/kg, has also been used to control leafhoppers and plant hoppers that attack rice.

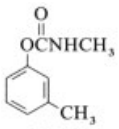
MTMC

Promecarb [*2631-37-0*], 3-methyl-5-(1-methylethyl)phenyl methylcarbamate, *mp* 87 – 88 °C, water solubility 92 mg/L, LD_{50} (rat, oral) 74 mg/kg, has been used to control orchard pests.

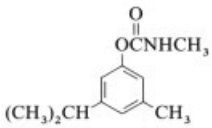
Promecarb

Propoxur [*114-26-1*], 2-(1-methylethoxy)-phenyl methylcarbamate, *mp* 91 °C, vapor pressure 0.84 mPa at 20 °C, water solubility 2 g/L, has rat LD_{50} values of 83 mg/kg (oral) and > 1 000 mg/kg (dermal). Propoxur is used to control household insects such as cockroaches, bedbugs, and wasps, and as a replacement for DDT in malaria control.

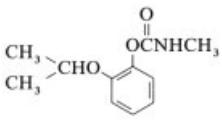
Propoxur

TBPMC [*780-11-0*], 3-*tert*-butylphenyl methylcarbamate, *mp* 140 – 142 °C, LD_{50} (mouse, oral) 470 mg/kg, has been used to control rice and fruit pests.

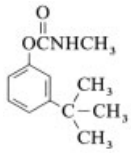
TBPMC

Trimethacarb [*2655-15-4*], 3,4,5-trimethylphenyl methylcarbamate, *mp* 117 – 119 °C, LD_{50} (rat, oral) 178 mg/kg, is used as a soil insecticide for corn.

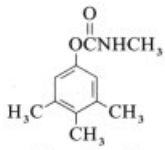
Trimethacarb

7.3. Methylcarbamates of Oximes [74]

Aldicarb [*116-06-3*], 2-methyl-2-(methylthio)-propanal *O*-[(methylamino)carbonyl]oxime, *mp* 99 – 100 °C, vapor pressure 13 mPa at 25 °C, water solubility 6 g/L, rat LD_{50} values 0.8, 0.65 mg/kg (oral) and 3, 2.5 mg/kg (dermal) is the most toxic commercial insecticide. Aldicarb is a broad-spectrum systemic insecticide used for seed and soil treatment, and as a nematocide.

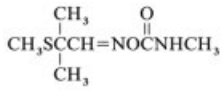
Aldicarb

Methomyl [*16752-77-5*], *S*-methyl-*N*-[(methylcarbamoyl)oxy]thioacetimidate, *mp* 78 – 79 °C, vapor pressure 9.1 mPa at 25 °C, water solubility 58 g/L, rat LD_{50} 17, 26 mg/kg (oral) and > 15 000 mg/kg (dermal), is a broad-spectrum insecticide.

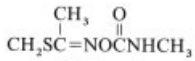
Methomyl

Oxamyl [*23135-22-0*], *S*-methyl 1-(dimethylcarbamoyl)-*N*-[(methylcarbamoyl)oxy]thioformimidate, *mp* 108 – 110 °C, vapor pressure 20 mPa at 25 °C, rat LD_{50} 5.4 mg/kg (oral) and 2 960 mg/kg (dermal), is a systemic insecticide and nematocide.

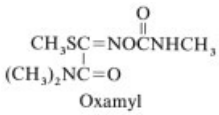

Oxamyl

Thiofanox [*39196-18-4*], 3,3-dimethyl-1-(methylthio)-2-butanone-*O*-[(methylamino)-carbonyl]oxime, vapor pressure 22.6 mPa at 25 °C, water solubility 5.2 g/L, LD_{50} (rat, oral) 8.5 mg/kg, is a systemic soil insecticide effective against red spider mites.

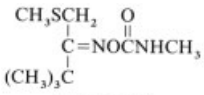

Thiophanox

7.4. Mode of Action [62], [69]

The insecticidal carbamates are cholinergic. Poisoned insects and animals exhibit violent convulsions and other neuromuscular disturbances. These insecticides carbamylate acetylcholinesterase and may also have a direct action on acetylcholine receptors because of their pronounced structural resemblance to acetylcholine. The mechanism of interaction with acetylcholinesterase is analogous to the normal three-step hydrolysis of acetylcholine (cf. Section 6.9.1). However, the third reaction step is much slower for the carbamylated enzyme than for the acetylated one. The importance of structural complementarity of the insecticidal carbamates to the active site of acetylcholinesterase is demonstrated by the pronounced difference in activities of D- and L-2-(*sec*-butylphenyl) methylcarbamate (the L isomer is five times more toxic) and of the 2-, 3-, and 4-substituted phenyl methylcarbamates where the 4-isomers are virtually inactive [68], [69].

Detoxification of carbamate insecticides occurs in vivo through microsomal hydroxylation, *N*-demethylation of carbamyl nitrogen, side chain oxidation, and ring hydroxylation. Methylenedioxyphenyl synergists (Section 3.5.4) prevent this largely by inhibiting the microsomal enzymes.

7.5. Environmental Toxicology

The *N*-methylcarbamates are generally biodegradable and have low to moderate soil persistence (one to four months). Aldicarb oxidizes readily to toxic CH_3SO and CH_3SO_2 derivatives which persist in groundwater. Certain carbamates are highly toxic to birds with the following oral LD_{50} values, in milligrams per kilogram: carbofuran, mallard 0.40, pheasant 4.2; mexacarbate, mallard 3.0, pheasant 4.5; and methomyl, mallard 16, pheasant 15; compared to carbaryl $> 2\,000$. Fish toxicity is generally low, but carbamates are highly toxic to bees. Carbamates with low mammalian LD_{50} values (e.g., isolan, aldicarb, methomyl, oxamyl, carbofuran, and mexacarbate) have caused severe poisoning in humans and domestic animals.

8. Formamidines [75–77]

The formamidine insecticides act as competitive agonists of the neurotransmitter octopamine in insects.

Chlordimeform [*6164-98-3*], *N′*-(4-chloro-2-methylphenyl)-*N,N*-dimethylmethanimidamide, *mp* 35 °C, *bp* 156 – 157 °C (53.3 Pa), d_4^{25} 1.105, vapor pressure 0.045 Pa at 20 °C, LD_{50} (rat, oral) 238 mg/kg, is used primarily against eggs and early growth stages of cotton bollworm and rice stem borer larvae (instars). It has both vapor and systemic action.

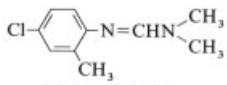

Chlordimeform

Amitraz [*33089-61-1*], *N′*-(2,4-dimethylphenyl)-*N*-{[(2,4-dimethylphenyl)imino]methyl}-*N*-methylmethanimidamide, *mp* 86 – 87 °C, water solubility 1 mg/L, LD_{50} (rat, oral)

400 mg/kg, is used principally to control red spider mites and pear psylla.

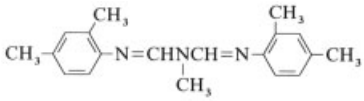

Amitraz

9. Insect Growth Regulators [78–85]

Most conventional insecticides are neurotoxins that produce biochemical lesions at specific target sites (e.g., acetylcholinesterase or the nerve axon). Therefore, many organochlorines, organophosphates, carbamates, and pyrethroids lack desirable insecticide selectivity. Insect growth regulators interfere with unique biochemical processes that are peculiar to insects: molting or ecdysis, and formation of the chitinous exoskeleton; they are truly selective insecticides (see also → Biological Control, Chap. 5.1.1.).

9.1. Juvenoids

Juvenoids are structurally optimized derivatives of the insect juvenile hormones (JH) or neotenin: JH I [*13804-51-8*], methyl 7-ethyl-9-(3-ethyl-3-methyloxiranyl)-3-methyl-2,6-nonadienoate, JH O [*55333-14-7*], (7-ethyl, 11-ethyl), JH II [*34218-616*], (7-methyl, 11-ethyl), and JH III [*22963-93-5*], (7-methyl, 11-methyl) have also been isolated from various insect species.

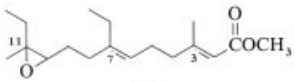

JH I

The JH hormones are insecticidal when applied exogenously to the egg and last instar larval stage, and interfere with the development of physiologically competent adults. However, they are of short persistence, and a variety of structurally optimized analogues have much longer residual action and improved specificity.

Methoprene [*40596-69-8*], 1-methylethyl (*E,E*)-11-methoxy-3,7,11-trimethyl-2,4-dodecadienoate, *bp* 100 °C (6.7 Pa), vapor pressure 3.1 Pa at 25 °C, soluble in water to 1.4 mg/L, has an LD_{50} (rat, oral) of > 34 000 mg/kg and is effective against mosquito and fly larvae.

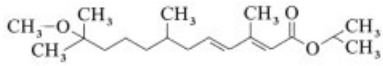

Methoprene

Hydroprene [*41096-46-2*], ethyl (*E,E*)-3,7,11-trimethyl-2,4-dodecadienoate, water solubility 0.54 mg/L, LD_{50} (rat, oral) > 34 000 mg/kg, is especially effective against aphids and in bait for cockroach control.

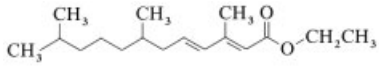

Hydroprene

Epophenonane [*57342-02-6*] is 2-ethyl-3-[3-ethyl-5-(4-ethylphenoxy)pentyl]-2-methyloxirane. The LD_{50} (rat oral) is 4 000 mg/kg.

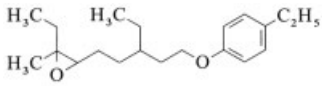

Epophenonane

Fenoxycarb [*79127-80-3*], ethyl [2-(4-phenoxyphenoxy)-ethyl] carbamate, *mp* 53 °C, vapor pressure 7.8 μPa at 20 °C, water solubility 6 mg/L, has a broad spectrum of activity. Its LD_{50} (rat, oral) is > 16 800 mg/kg.

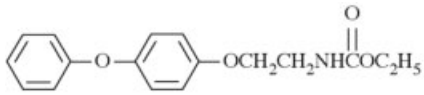

Fenoxycarb

9.2. Chitin Synthesis Inhibitors [85], [86]

Some insect growth regulators inhibit the formation of the chitinous insect exoskeleton and thus produce a critical biochemical lesion at the time of molting, pupation, or ecdysis.

Diflubenzuron [*35367-38-5*], *N*-{[(4-chlorophenyl)amino]carbonyl}-2,6-difluorobenzamide, *mp* 239 °C, is soluble in water to 0.3 mg/L and has an LD_{50} (rat, oral) of > 4 640 mg/kg. Diflubenzuron is the best known compound with the same general structure that inhibits the synthesis of chitin and is thus very specific for the control of arthropods. Diflubenzuron is an ovicide and a slow-acting broad-spectrum insecticide.

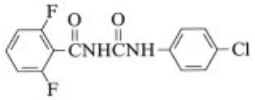

Diflubenzuron

10. Other Types of Insecticides

Hydramethylnon [*67485-29-4*], tetrahydro-5,5-dimethyl-2(1*H*)-pyridinone(3-[4-(trifluoromethyl)-phenyl]-1-{2-[4-(trifluoromethyl)phenyl]ethenyl}-2-propenylidene)hydrazone, *mp* 189 – 191 °C, is a slow-acting toxicant used successfully in bait for the fire ant, other ants, and cockroaches.

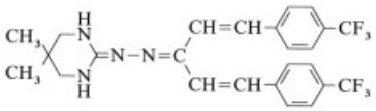

Amdro

Azadirachtin [*11141-17-6*], is the most active of the growth regulators (limonoids), occurring at 0.2 – 0.4 % in the seeds of the neem tree, *Azadirachta indica* (Meliaceae). Neem extracts are wide-spectrum insecticides and they act by blocking the action of the molting hormone ecdysone [87].

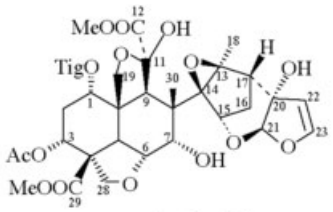

Azadirachtin

11. Fumigants [1]

Fumigants are distributed to the target site as gases and, therefore, must exist in air at a concentration that is lethal to the insect pest. This physical requirement limits the insecticides that can be employed usefully as fumigants. Compounds that boil around room temperature are the most generally useful and include hydrogen cyanide, methyl bromide, methyl formate, and ethylene oxide. For soil fumigation, the slower release of vapor from ethylene dibromide, dibromochloropropane, and dichlorodiethyl ether, which boil as high as 180 °C, has proved effective. Organic compounds with higher vapor pressures, such as 1,4-dichlorobenzene and naphthalene, sublime readily and have fumigant action in tight enclosures such as museum cases and cedar chests. Certain contact insecticides, such as azobenzene, lindane, and dichlorvos, may kill insects by vapor action especially when volatilized by atomization, heat, and combustion in pyrotechnic mixtures.

The major uses of fumigants are for insect-infested mills, warehouses, grain elevators, groceries, museums, and habitations to control structural pests; for stored products in bulk or package; for soil fumigation of grubs, wireworms, ants, nematodes, etc.; and for living plants in greenhouses and nursery stock.

The proper choice of fumigant for insect pest control is determined not only by the effectiveness of the fumigant gas, but also by safety to humans, animals, and plants; flammability, penetrability, and reactivity with furnishings; effect on seed germination; and cost. Many of the compounds listed below are described in detail elsewhere in this encyclopedia.

Acrylonitrile [*107-13-1*], CH_2=CHCN, *bp* 78 °C, liquid d_4^{20} 0.797, gas *d* 1.8, vapor pressure 11 kPa at 20 °C, TLV 20 ppm, is used to fumigate mills and commodities.

Carbon disulfide [*75-15-0*], CS_2, *bp* 46.3 °C, liquid d_4^{20} 1.263, gas *d* 2.6, vapor pressure 41.9 kPa at 20 °C, TLV-TWA 10 ppm, MAK 10 ppm, has been used as a household fumigant.

Carbon tetrachloride [*56-23-5*], CCl_4, *bp* 76 °C, liquid d_4^{20} 1.595, gas *d* 5.3, vapor pressure 12 kPa at 20 °C, TLV-TWA 5 ppm, MAK 10 ppm, is a grain fumigant and carrier for other fumigants.

Chloropicrin [*76-06-2*], Cl_3CNO_2, *bp* 112 °C, liquid d_4^{20} 1.651, gas *d* 5.7, vapor pressure 2.7 kPa at 20 °C, TLV-TWA 0.1 ppm, MAK 0.1 ppm, has been used as a grain and soil fumigant, but is an obnoxious lachrymator.

1,2-Dibromo-3-chloropropane [*96-12-8*], $CH_2BrCHBrCH_2Cl$, *bp* 199 °C, liquid d_4^{20} 2.08, vapor pressure 0.08 kPa at 20 °C, was used widely as a soil fumigant, but U.S. registration was withdrawn because of its carcinogenicity.

2,2′-Dichlorodiethyl ether [*111-44-4*], $ClCH_2CH_2OCH_2CH_2Cl$, *bp* 178 °C, liquid d_4^{20} 1.222, gas *d* 4.9, vapor pressure 0.10 kPa, at 20 °C, TLV 15 ppm, MAK 10 ppm, is a soil fumigant.

1,1-Dichloro-1-nitroethane [*594-72-9*], $CH_3CCl_2NO_2$, *bp* 111 °C, liquid d_4^{20} 1.224, vapor pressure 2.47 kPa at 20 °C, TLV 10 ppm, MAK 10 ppm, has been used to fumigate grain and stored products.

1,2-Dichloropropane [*78-87-5*], propylene dichloride, $CH_2ClCHClCH_3$, *bp* 95.4 °C, liquid d_4^{20} 1.159, vapor pressure 28 kPa at 20 °C, TLV-TWA 50 ppm, MAK 75 ppm, is a soil fumigant.

***trans*-1,3-Dichloropropene** [*542-75-6*], $ClCH{=}CHCH_2Cl$, *bp* 111 °C, liquid d_4^{20} 1.224, vapor pressure 2.47 kPa at 20 °C, TLV-TWA 1 ppm, is a soil fumigant.

Ethylene chlorobromide [*107-04-0*], $ClCH_2CH_2Br$, 1-bromo-2-chloroethane, *bp* 107 °C, liquid d_4^{20} 1.689, vapor pressure 5.3 kPa at 20 °C, has been used as a soil fumigant.

Ethylene dibromide [*106-93-4*], $BrCH_2CH_2Br$, 1,2-dibromoethane, *bp* 131.6 °C, liquid d_4^{20} 2.172, gas *d* 6.5, vapor pressure 1.0 kPa at 20 °C, has been used widely as a soil and commodity fumigant, but its U.S. registration was canceled because of carcinogenicity.

Ethyl formate [*109-94-4*], $HCOOC_2H_5$, *bp* 54 °C, liquid d_4^{20} 0.917, TLV-TWA 100 ppm, MAK 100 ppm, is used to fumigate packaged foods.

Ethylene oxide [*75-21-8*], $(CH_2)_2O$, *bp* 10.7 °C, liquid d_4^{7} 0.887, gas *d* 1.5, vapor pressure 146 kPa at 20 °C, TLV-TWA 1 ppm, is used to fumigate packaged foods.

Hydrogen cyanide [*74-90-8*], HCN, *bp* 26 °C, liquid d_4^{20} 0.688, gas *d* 0.9, vapor pressure 84 kPa at 20 °C, TLV-TWA 10 ppm, is a very hazardous general fumigant.

***β*-Methallyl chloride** [*563-47-3*], $CH_2{=}C(CH_3)CH_2Cl$, *bp* 72 °C, liquid d_4^{20} 0.925, is no longer used.

Methyl bromide [*74-83-9*], CH_3Br, bromomethane, *bp* 4.3 °C, liquid d_4^{0} 1.732, gas *d* 3.3, vapor pressure 189.2 kPa at 20 °C, TLV-TWA 5 ppm, MAK 5 ppm, is a widely used general fumigant.

Methyl formate [*107-31-3*], $HCOOCH_3$, *bp* 32 °C, liquid d_4^{20} 0.974, vapor pressure 83.2 kPa at 20 °C, TLV-TWA 100 ppm, is used to fumigate dried fruit.

Naphthalene [*91-20-3*], $C_{10}H_8$, *bp* 218 °C, gas *d* 4.4, vapor pressure 0.01 kPa at 20 °C, TLV-TWA 10 ppm, is used to fumigate fabric and museum specimens.

1,4-Dichlorobenzene [*106-46-7*], ClC_6H_4Cl, *bp* 173.4 °C, *mp* 53 °C, gas *d* 5.1, vapor pressure 0.1 kPa at 20 °C, TLV-TWA 75 ppm, is used to fumigate fabric and museum specimens.

Phosphine [*7803-51-2*], PH_3, *bp* − 87.4 °C, liquid d^{-90} 0.746, gas *d* 1.2, TLV-TWA 0.3 ppm, is used as a grain fumigant, generated from aluminum phosphide and ammonium carbamate in the presence of moisture.

Sulfuryl fluoride [*2699-79-8*], SO_2F_2, *bp* − 55.2 °C, liquid d_4^{25} 1.342, gas *d* 3.5, vapor pressure 1601.3 kPa at 20 °C, TLV-TWA 5 ppm, is used in tent fumigation for drywood termites and powder post beetles.

Trichloroacetonitrile [*545-06-2*], Cl_3CCN, *bp* 85 °C, liquid d_4^{25} 1.44, has been used as a grain fumigant.

Trichloroethylene [*79-01-6*], $ClCH{=}CCl_2$, *bp* 86.7 °C, liquid d_4^{15} 1.470, gas *d* 4.5, vapor pressure 9.7 kPa at 20 °C, TLV-TWA 50 ppm, is a grain fumigant.

12. Microbial Insecticides [88–92]

Insects are attacked by numerous pathogens; ca. 450 viruses, 80 bacteria, 460 fungi, 250 protozoa, and 20 rickettsial insect diseases have been identified. A number of these can be adapted for mechanical dissemination as microbial insecticides in the form of spores, microbial toxins, or viral suspensions (→ Biological Control, Chap. 2.3., → Biological Control, Chap. 5.1.2.,→ Biological Control, Chap. 5.1.3.,→ Biological Control, Chap. 5.2.1.,→ Biological Control, Chap. 5.2.2.). These microbial insecticides are being used increasingly widely for integrated pest management programs because of their selectivity and freedom from environmental hazards and toxic residues. Possible disadvantages are their instability in light and air, their dependence upon suitable climatic conditions, the precise timing of application for maximum effectiveness, and their slow lethal action.

Bacillus thuringiensis insecticide is the crystalline endotoxin that is made from the bacterium as a fermentation product (→ Biological Control, Chap. 2.3.). Commercial wettable powders are standardized against a reference strain and are especially effective against lepidopterous larvae such as the bagworm, cabbage looper, and gypsy moth. *Bacillus thuringiensis* subspecies *isralensis* is a related toxin that is especially effective against mosquito and blackfly larvae; *B. sphaericus* is a spore-forming bacterium that produces a specific toxin against mosquito larvae.

Bacillus popillae is a spore-forming bacterium that causes "milky disease" in the larvae of the Japanese beetle *Popillia japonica* and related species. The spores have been used for many years as a standardized talc dust, containing 10^8 spores per gram, for treatment of turf to kill Japanese beetle larvae.

Avermectins and Ivermectin. The avermectins are pentacyclic lactones isolated as fermentation products of *Streptomyces avermitilis.* Ivermectin is 22,23-dihydroavermectin. These complex products (→Biological Control, Chap. 5.1.3., →Biological Control) are broad-spectrum acaricides, insecticides, and parasiticides, and are effective in quantities of grams per hectare for control of ants, mites, and other pests.

Other microbial insecticides include the nuclear polyhedrosis viruses (NPV) of the cotton bollworms *Heliothis,* the armyworms *Prodenia* and *Spodoptera,* the cabbage looper *Trichoplusia*, and the sawfly *Neodiprion*. The spores of the microsporidan *Nosema locustae* are marketed for grasshopper control. The fungal spores of *Beauvaria bassiana* are standardized and are used in China to control the European corn borer and in the Soviet Union to control the Colorado potato beetle.

13. Genetic Control [93–98]

Genetic control refers to a variety of methods for suppressing insect pests through manipulation of the inheritance mechanisms of the population.

The mass release of sterilized males, *sterile male release,* has proved most successful, but other techniques have been evaluated including the use of chemosterilants and the mass introduction of deleterious mutations (e.g., conditional lethals, cytoplasmic incompatibility, chromosomal translocations, and meiotic drive; (→ Biological Control, Chap. 5.2.2.).

13.1. Sterile Male Release

The release of relatively large numbers of males, sterilized either by exposure to γ-radiation or by chemosterilants, into a population of virgin females has reduced the number of fertile female offspring, as shown in Table 3.

The most effective use of sterile male release was in the eradication of the screwworm fly *Cochliomyia hominivorax* from the southeastern United States in 1959 by periodic release

Table 3. Rate of decrease in wild pest population after sustained release of sterile males [a]

Generation	Wild female population		Sterile male population released	Ratio of sterile to fertile males	Number of reproducing females [b]
	Unsuppressed	Suppressed			
Parental	1×10^6	1×10^6	9×10^6	9 : 1	100 000
F_1	5×10^6	0.5×10^6	9×10^6	18 : 1	26 316
F_2	25×10^6	131 580	9×10^6	68 : 1	1 907
F_3	1.25×10^8	9 535	9×10^6	942 : 1	10
F_4	1.25×10^8 [c]	50	9×10^6	180 000 : 1	0

[a] After KNIPLING [93].
[b] Each female produces five reproductive daughters.
[c] Maximum sustainable population.

over 18 months of ca. 2×10^9 male flies, sterilized with γ-radiation from cobalt-60, over an area of 181 000 km^2. A similar program along the Texas – Mexico border has substantially controlled the screwworm over the past 25 years, through the annual release of as many as 10×10^9 sterile males. Dramatic population reduction of the Mediterranean fruit fly *Ceratitis capitata* was also achieved in California in 1981 when ca. 40×10^6 flies were released weekly. Sterile male release has been used with limited success to suppress isolated populations of the codling moth *Laspeyresia pomonella* in the Pacific Northwest and is proposed as an integral part of a cotton boll weevil eradication scheme in the southeastern United States.

Success in sterile male release programs depends on several factors:

1) method for rearing large numbers of insects,
2) adequate dispersion of released sterile males,
3) a sterilization procedure that does not decrease male vigor and competitiveness in mating,
4) a pest species that mates only once per lifetime or in which the sperm from sterile males compete with those of fertile males, and
5) a low pest population density or a means of reducing the population to a low level (e.g., by insecticide application).

Many species of insects do not conform to these criteria, and the rearing and release of sterile males can be a costly procedure.

13.2. Chemosterilants [99–101]

The availability of a variety of radiomimetic (i.e., mimic the effect of γ-radiation) and antimetabolite cancer suppressive drugs that can sterilize large segments of wild insect populations has led to many suggestions for their use in genetic control programs.

Aziridines have a highly reactive ethylenimine group. They alkylate deoxyribonucleic acid and cause sterility in both sexes of house flies, mosquitoes, fruit flies, and other insects. Compounds used experimentally include *TEPA* [*57-39-6*], triethylenephosphoramide, *mp* 41 °C, LD_{50} (rat, oral) 37 mg/kg, and its analogue *thio-TEPA* [*52-24-4*] (→ Cancer Chemotherapy, Chap. 3.5.), triethylenethiophosphoramide, *mp* 51.5 °C. Other radiomimetic compounds that have been investigated as chemosterilants include *apholate* [*52-46-0*], 2,2,4,4,6,6-hexakis(1-aziridinyl)-2,2,4,4,6,6-hexahydro-1,3,5,2,4,6-triazaphosphorine, *mp* 147.5 °C, LD_{50} (rat, oral) 98, 113 mg/kg; and *HEMPA* [*680-31-9*], hexamethylphosphoric triamide, *bp* 105 – 107 °C (1.5 kPa), d_4^{20} 1.03, LD_{50} (rat, oral) 2650, 3360 mg/kg. *Busulfan* [*55-98-1*], 1,4-butanediol dimethanesulfonate, $CH_3SO_2O(CH_2)_4OSO_2CH_3$, LD_{50} (rat, oral) 1.8 mg/kg, has been used in the mass production of sterile insects. Unfortunately, all of these radiomimetic compounds are mutagens and carcinogens, and produce degenerative changes in human germ plasm at dosages of one-hundredth or less of the LD_{50}. Their casual and indiscriminate use in insect control is precluded.

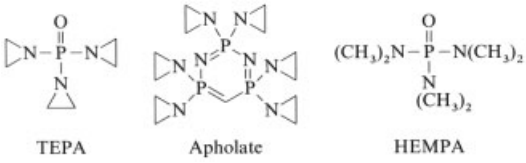

Antimetabolites suggested for use as insect sterilants include *methotrexate* [*59-05-2*], amethopterin, *N*-{4-[(2,4-diamino-6-pteridinyl)methyl]methylamino}benzoyl-L-glutamic acid, a folic acid antimetabolite; and *fluorouracil* [*51-21-8*], 5-fluoro-2,4-(1*H*,3*H*)-pyrimidinedione, a uracil antimetabolite (→ Cancer Chemotherapy, Chap. 2.2.).

13.3. Host Plant Resistance

Breeding plants to incorporate genetic factors that deter insect pest attack is a most important technology for integrated pest management (IPM). Genetic host plant resistance is produced by antixenosis, antibiosis, or by tolerance. *Antixenosis* (nonpreference) results from the removal of kairomone attractants (see Section 15.3) or alteration of morphological characters such as pubescence. *Antibiosis* results from the production of deleterious allomones that deter insect growth and development. *Tolerance* is produced by enhanced plant vigor that results in rapid repair of insect injury.

Genetic engineering provides opportunities to produce host plant resistance through transfer of genes for the production of antibiotic factors such as *Bacillus thuringiensis* (Bt) toxin into the genomes of crop plants such as tobacco, tomato, potato, corn, and soybean. The insecticidal properties of the toxin should protect the crop plant from damage by insect herbivores. However, widespread utilization of this technology may also provide increased opportunities for the development of Bt resistance.

14. Repellents [102], [103]

Repellents are substances that protect humans and domestic animals, plants, or products from insect attack by making food or living conditions unattractive or offensive to the insect. Repellents usually have low toxicity and are rarely, if ever, effective against a wide variety of insect pests. They can sometimes be employed to advantage in insect control where use of toxic insecticides is impossible. Examples of the use of insect repellents are

1) repellents against crawling insects (e.g., cresote lines and dinitro-*o*-cresol barriers for migrating chinch bugs, pentachlorophenol and trichlorobenzene soil barriers for subterranean termites, and sticky bands around tree trunks to protect against gypsy moth larvae); and
2) repellents against insect feeding (e.g., Bordeaux mixture to prevent the feeding of leafhoppers, flea beetles, and psyllids on plants; fly and mosquito repellents to lessen the attack of blood-sucking insects, mites, and ticks; treatment of logs with pentachlorophenol to prevent attack of bark beetles; and mothproofing treatment to protect woollen goods from attack by clothes moths and carpet beetles).

14.1. Repellents for Plant-Feeding Insects

Bordeaux mixture originated in France as a spray to control downy mildew disease, caused by the fungus *Peronospora viticola* of grapes in about 1882 (→ Fungicides, Agricultural). Although primarily a fungicide, Bordeaux mixture is repellent to leafhoppers, psyllids, and flea beetles when sprayed on leaves of grapes and potatoes. The mixture is formed by mixing 3.6 – 4.5 kg of hydrated lime and 1.9 – 2.7 kg of copper sulfate in 380 L of cold water to produce a colloidal precipitate of tetracupric sulfate. The colloid remains in suspension for several hours and, when sprayed on foliage, covers the leaves with a thin tenacious film that gradually produces soluble copper to give some residual toxic effect.

Two fungicides that have repellent effects on the Japanese beetle *Popillia japonica*, which feeds on foliage, are *thiram* [*137-26-8*], *bis*(dimethylthiocarbamoyl) disulfide, *mp* 155 – 156 °C, *d* 1.29, LD_{50} (rat, oral) 640 mg/kg, and *nabam* [*142-59-6*], ethylenebis(dithiocarbamic acid) disodium salt, LD_{50} (rat, oral) 395 mg/kg.

Thiram

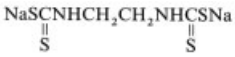

Nabam

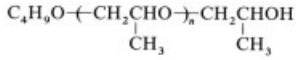

14.2. Repellents for Blood-Feeding Insects

The requirements for effective repellents against blood-feeding pests in order of importance are:

1) effective protection of the treated area for several hours on all types of subjects and under all climatic conditions;
2) complete freedom from toxicity and irritation when regularly applied to human or animal skin;
3) cosmetic acceptability including freedom from unpleasant odor, taste, or touch, and harmlessness to clothing;
4) protection against a wide variety of biting mosquitoes, flies, fleas, mites and ticks; and
5) low cost and ready availability.

Evaluation of tens of thousands of chemicals as pest repellents have shown that few meet all requirements satisfactorily. The routine use of repellents for arthropod vectors of human disease has been an important part of military hygiene to protect troops from such diseases as malaria and scrub typhus. The use of fly repellents on cattle and milk cows has been shown to decrease weight loss and to improve milk production in animals subject to severe annoyance and blood loss from constant infestation. The following repellents have found widespread practical application:

Benzyl benzoate [*120-51-4*], *bp* 323 °C, *mp* 21 °C, d_4^{25} 1.12, LD_{50} (rat, oral) 1900 mg/kg, is used as a clothing treatment for chigger mites *Trombicula* spp.

Butoxypoly(propylene glycol) [*9003-13-8*], 400 and 800 molecular mass fractions, *d* 0.973 – 0.990, LD_{50} (rat, oral) 11, 200 mg/kg, is a fly repellent for cattle.

***N*-Butylacetanilide** [*91-49-6*], *bp* 277 – 281 °C, *mp* 22 °C, *d* 0.99, LD_{50} (rat, oral) 2 830 mg/kg, is used as a clothing treatment to repel fleas and ticks.

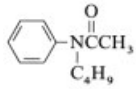

2,3,4,5-*bis*(Butyl-2-ene)tetrahydrofurfural [*126-15-8*], *bp* 307 °C, *d* 1.121, LD_{50} (rat, oral) 2 500 mg/kg, repels biting flies on cattle.

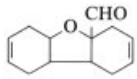

Butopyronoxyl [*532-34-3*], butyl-3,4-dihydro-2,2-dimethyl-4-oxo-2*H*-pyran-6-carboxylate, *bp* 256 – 270 °C, d_4^{25} 1.052 – 1.060, LD_{50} (rat, oral) 7 800 mg/kg, is a mosquito and fly repellent.

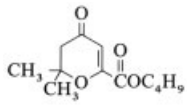

Dibutyl adipate [*105-99-7*], *bp* 183 °C (1.9 kPa), *d* 0.965, LD_{50} (rat, oral) 12 900 mg/kg, is a tick repellent.

Dibutyl phthalate [*84-74-2*], *bp* 340 °C, d_4^{20} 1.045, LD_{50} (rat, oral) 21 000 mg/kg, is used on clothing to repel chiggers.

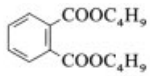

Di-*n*-butyl succinate [*141-03-7*], *bp* 274 °C, d_4^{20} 0.9768, LD_{50} (rat, oral) 8 000 mg/kg, is a fly repellent for cattle.

Deet [*134-62-3*], *N,N*-diethyl-3-methylbenzamide, *bp* 160 °C (2.5 kPa), d_4^{20} 0.996, LD_{50} (rat, oral) 2 000 mg/kg, is a general-purpose repellent that is used widely by outdoor recreationists.

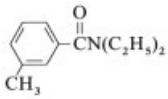

Dimethyl phthalate [*131-11-3*], *bp* 284 °C, d_{25}^{25} 1.189, LD_{50} (rat, oral) 8 200 mg/kg, is a general-purpose repellent for flies and mosquitoes.

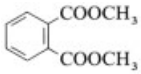

2-Butyl-2-ethyl-1,3-propanediol [*115-84-4*], *mp* 40 °C, d_4^{20} 0.931, LD_{50} (rat, oral) 5 040 mg/kg, is an effective, persistent mosquito repellent.

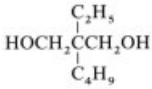

2-Ethyl-1,3-hexanediol [*94-96-2*], "612," *bp* 244 °C, d_4^{20} 0.942, LD_{50} (rat, oral) 2 400 mg/kg, is a widely used mosquito repellent.

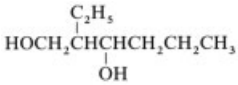

Di-*n*-propylpyridine 2,5-dicarboxylate [*136-45-8*], *bp* 186 °C (0.13 kPa), *d* 1.08, LD_{50} (rat, oral) 6 230 mg/kg, is a fly repellent used on cattle.

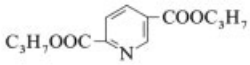

Application. Repellents are often applied to clothing, gloves, and head nets, where their action may remain effective for a week or more. Application can be made by rubbing, spraying, or dipping the clothing in an emulsion of the repellent. A repellent mixture of three parts each of benzyl benzoate, *N*-butylacetanilide, and 2-ethyl-2-butyl-1,3-propanediol with one part of nonionic emulsifier is the standard clothing impregnant (M – 1960) for the U.S. Armed Services.

14.3. Repellents for Wood-Feeding Insects

Wooden building materials have been protected against termites by treatment with *pentachlorophenol* [*87-86-5*].

Trichlorobenzene, bp 205 – 250 °C, d_{25}^{25} 1.460, has been used as a soil treatment to repel or poison termites around the foundation of buildings. The technical product is a mixture of the 1,2,3- [*87-61-6*] and 1,2,4- [*120-82-1*] isomers.

14.4. Repellents for Fabric-Eating Pests

Two major groups of insect pests can utilize the keratinaceous proteins of wool and animal hides as complete dietary sources and, thus, are major pests of woolen clothes, rugs, upholstered furniture, and museum specimens. The clothes moths (family Tineidae) include four species with worldwide distribution: the case-making clothes moth *Tinea pellionella*, the webbing clothes moth *Tineola bisselliella*, the carpet moth *Trichophaga tapetzella*, and the brown house moth *Hoffmannophila pseudospretella*. Carpet beetles (family Dermestidae) include the black carpet beetle *Attagenus piceus*, along with the carpet beetle *Anthrenus scrophulariae*, the varied carpet beetle *A. verbasci,* and the furniture carpet beetle *A. flavipes*. These fabric pests cause direct damage amounting to more than \$ 10^9 annually in the United States alone. A number of mothproofing agents are available that protect woolen fabrics from attack by a combination of

repellent and toxic action. Ideally, such mothproofing materials should

1) be fixed into the cloth during dyeing by chemical reaction with the wool protein,
2) be completely effective against fabric pests,
3) give protection for up to 15 years after repeated washing and dry cleaning, and
4) be nonvolatile, stable to light, odorless, colorless, and nontoxic to humans.

Sodium fluorosilicate [*16893-85-9*], alone or with aluminum, magnesium, or ammonium salts, is an effective mothproofing agent when employed in 0.5 – 0.7 % aqueous solution with 0.3 % potassium alum and 0.03 % oxalic acid. Such formulations may be applied by spraying or dipping and cannot be removed by dry cleaning.

Formerly, *DDT* at 0.25 – 0.75 % and *dieldrin* at 0.05 % of the cloth weight were applied by spraying in a volatile petroleum solvent or by dipping the cloth in an aqueous emulsion. This protected the cloth against fabric pests for months to years. However, because of environmental pollution problems with these persistent bioaccummulative organochlorines, they are no longer used. Several persistent *synthetic pyrethroids* including permethrin, fenvalerate, and fenpropanate are effective in situ mothproofing agents.

The most widely used commercial mothproofing agents are the (polychloro-2-chloromethylsulfamido) diphenyl ethers (*Eulan type*) and the chlorinated diphenyl ether urea sulfonic acid (*Mitin type*). They are applied to woolens during fabrication at about 1 – 3 % of the weight of the cloth, and fixed by chemical reaction with the wool protein.

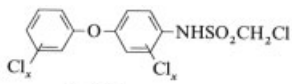

Eulans

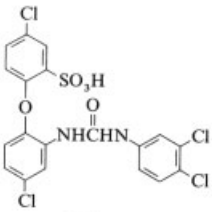

Mitin FF
[*3567-25-7*]

15. Attractants [104–117]

With a growing need for more specific, less environmentally polluting pest control, the use of specific insect attractants has become an important part of control technology. Chemical attractants can be used in conjunction with simple, inexpensive sticky traps for monitoring insect pest populations in relation to the economic threshold, for removal trapping of pests, for luring them to toxic bait, or for mating confusion. Two general types of attractants have been studied widely: (1) the sex pheromones used in intraspecies chemical communication leading to mating, and (2) plant kairomones used in interspecies chemical communication for host plant selection.

15.1. Sex Pheromone Attractants

Sex pheromone attractants have now been identified in several hundred species of insect pests[113]. For the most part, they are volatile esters, alcohols, and aldehydes that facilitate sexual communication and mating over distances up to several hundred meters and can be perceived in nanogram to microgram quantities. In Lepidoptera, sex pheromones are typically produced by the female in eversible glands at the tip of the abdomen. A number of pheromones used in commercial insect control are identified in Table 4. The naturally occurring sex pheromones are often complex mixtures, and maximum response to synthetic mixtures depends on specific blends. For example, the male European corn borer *Ostrinia nubilalis* responds to a mixture of the *Z*- and *E*-isomers of 11-tetradecenyl acetate. Populations of the corn borer in the midwestern United States respond optimally to a 97 : 3 isomer blend, but those in the eastern United States respond to a 3 : 97 isomer blend [113].

The male tobacco budworm *Heliothis virescens* produces a pheromone blend of (*Z*)-11-hexadecenal with traces of (*Z*)-9-hexadecenal, (*Z*)-7-hexadecenal, (*Z*)-9-tetradecenal [*53939-27-8*], (*Z*)-11-hexadecenol [*56683-54-6*], hexadecanal [*629-80-1*], and tetradecanal [*124-25-4*].

In the Coleoptera, sex pheromones are generally produced by the female, but in some species

Table 4. Sex pheromones (Lepidoptera) used in insect control

Compound	Structure	CAS registry number	Species
(*Z*)-11-Hexadecenal		[*53939-28-9*]	corn earworm, *Heliothis zea*
(*Z*)-9-Hexadecenal		[*56219-04-6*]	tobacco budworm, *H. virescens*
(*Z*)-7-Hexadecenal		[*56797-40-1*]	
(*Z,Z*)-11,13-Hexadecadienal		[*71317-73-2*]	navel orangeworm, *Amyelois transitella*
(*E,E*)-8,10-Dodecadienol		[*33956-49-9*]	codling moth, *Laspeyresia pomonella*
(*Z*)-7-Dodecenyl acetate	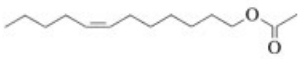	[*14959-86-5*]	cabbage looper, *Trichoplusia ni* soybean looper, *Pseudoplusia includens* alfalfa looper, *Autographa californica* black cutworm, *Agrotis ipsilon*
(*Z*)-8-Dodecenyl acetate	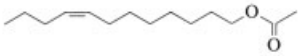	[*28079-04-1*]	oriental fruit moth, *Grapolitha molesta*
(*Z*)-9-Dodecenyl acetate		[*16974-11-1*]	grape berry moth, *Paralobesia viteana*
(*Z*)-9-Tetradecenyl acetate	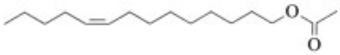	[*16725-53-4*]	beet armyworm, *Spodoptera exigua* southern armyworm, *S. eridanea* fall armyworm, *S. frugiperda*
(*E*)-11-Tetradecenyl acetate	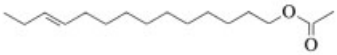	[*39298-60-7*]	European corn borer, *Ostrinia nubilalis*
(*Z,E*)-9,12-Tetradecadienyl acetate	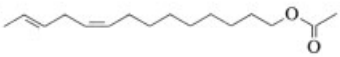	[*31654-77-0*]	almond moth, *Cadra cautella* fig moth, *C. figulilella* Mediterranean flour moth, *Anagasta kuehniella* tobacco moth, *Ephestia elutella*
(*Z,Z*)-7,11-Hexadecadienyl acetate	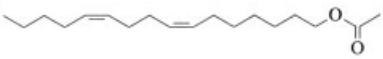	[*52207-99-5*]	pink bollworm, *Pectinophora gossypiella* Angoumois grain moth, *Sitotroga cerealella*
(*Z,Z*)-3,13-Octadecadienyl acetate	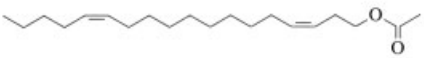	[*53120-27-7*]	peach tree borer, *Sanninoidea exitiosa*
(3 *E*,13 *Z*)-Octadecadienyl acetate	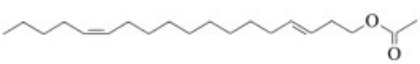	[*53120-26-6*]	lesser peach tree borer, *Synanthedon pictipes*
(*Z*)-7,8-Epoxy-2-methyloctadecane	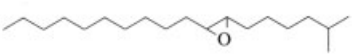	[*29804-22-6*]	gypsy moth, *Porthetria dispar*

(e.g., cotton boll weevil *Anthonomus grandis*), a four-component sex pheromone is produced by the male and liberated in the feces (frass). Each component is essential for effective aggregation, and the mixture is used commercially as grandlure [*11104-05-5*].

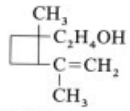

D-*cis*-2-isopropenyl-1-methylcyclobutaneethanol [*30820-22-5*]

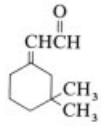

cis- and *trans*-3,3-Dimethyl-Δ′, α-cyclohexaneacetaldehyde [*24393-78-0*] (*cis*) [*24398-89-8*] (*trans*)

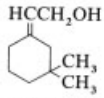

cis-3,3-Dimethyl-Δ′, β-cyclohexaneethanol [*26532-23-0*]

The sex pheromone of the female Japanese beetle *Popillia japonica* is (*R,Z*)-5-(1-decenyl)-dihydro-2-(3*H*)-furanone [*64726-91-6*]; that of the western corn rootworm *Diabrotica virgifera virgifera* and its close relative the northern corn rootworm *D. barberi* is (2*R*,8*R*)- 8-methyl-2-decyl propanoate [*104487-05-0*]. The female sex pheromone of the southern corn rootworm *D. undecimpunctata* is (*R*)- 10-methyl-2-tridecanone [*82621-53-2*].

5-(1-Decenyl)dihydrofuranone

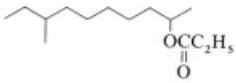

8-Methyl-2-decyl propanoate

10-Methyl-2-tridecanone

These beetle pheromones are used in trapping to monitor pest populations.

15.2. Aggregation Pheromones [108], [114]

Bark beetles (Coleoptera, Scolytidae) attack dying trees and bring about aggregation and mass attack by using a blend of pheromones produced by males and females, together with kairomones produced by weakened trees. Three types of aggregation pheromone blends have been used in mass trapping of bark beetle pests.

In *Dendroctonus brevicomis*, the western pine beetle, oxygenated terpenoid pheromones *exo*-brevicomin [*20290-99-7*], (1*R*)-*exo*-7-ethyl-5-methyl-6,8-dioxabicyclo[3.2.1]octane (female pheromone), and frontalin [*28401-39-0*], 1,5-di-methyl-(1*S*)-6,8-dioxabicyclo[3.2.1]octane (male pheromone), are produced in the hind gut of the beetles after contact with the tree kairomone terpenoid myrcene [*123-35-3*].

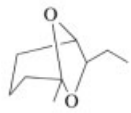

Exo-brevicomin

Frontalin

Myrcene

In *Ips typographus*, the Norway spruce beetle, boring males convert the tree terpenoid myrcene to the pheromone ipsdienol [*35628-00-3*], 2-methyl-6-methylene-(*S*)-2,7-octadien-4-ol. These beetles also produce the pheromones 2-methyl-3-buten-2-ol [*115-18-4*] and *cis*-verbenol [*473-67-6*], *cis*-4,6,6-trimethylbicyclo[3.1.1]hept-3-en-2-ol.

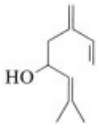

Ipsdienol

2-Methyl-3-buten-2-ol

cis-Verbenol

Scolytus multistriatus, the lesser European bark beetle, is a vector of the pathogen *Ceratocystis ulmi*, which causes Dutch elm disease. The beetle frass pheromones 4-methyl-3-heptanol [*14979-39-6*] (female) and α-multistriatin [*59014-03-8*], (1*S*)-*endo, endo*-5-ethyl-2,4-di-

methyl-6,8-dioxabicyclo[3.2.1]octane, (male) are produced after exposure to the tree terpenoid α-cubebene [*17699-14-8*], [3a*S*-(3aα,3bβ,4β,7α,*S*)]-3a,3b,4,5,6,7-hexahydro-3,7-dimethyl-4-methyl-ethyl-1*H*-cyclopenta[1,3]cyclopropa[1,2]benzene. The three-component aggregation pheromone is called multilure [*60018-97-5*].

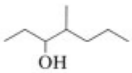

4-Methyl-3-heptanol

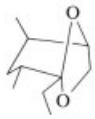

α-Multistriatin

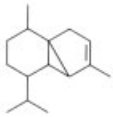
α-Cubebene

15.3. Kairomone Attractants

Kairomone attractants are volatile chemicals released by plants that serve as major cues for host plant selection by phytophagous insects. Increasing numbers of kairomones have been identified and found to be useful insect attractants[109](see also → Biological Control). Among the most successful are mixtures of eugenol [*97-53-0*] with geraniol [*106-24-1*], or phenylethyl propanoate [*122-70-3*], which are used widely for monitoring new infestations of the Japanese beetle *Popillia japonica* and for removal trapping in orchards.

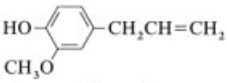

Eugenol

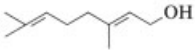

Geraniol

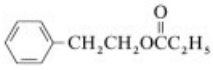

Phenethyl propanoate

Other kairomone lures for Coleoptera include caproic acid [*142-62-1*] for the green June beetle *Cotinus nitida*; propyl 1,4-benzodioxane-2-carboxylate [*24902-02-1*] for the European chafer *Amphimallon majalis*; and ethyl 3-isobutyl-2,2-dimethylcyclopropanoate [*33419-38-4*] for the coconut rhinoceros beetle *Oryctes rhinoceros*.

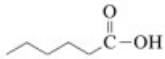

Caproic acid

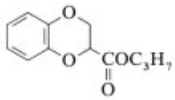

Propyl benzodioxane-carboxylate

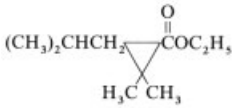

Ethyl 3-isobutyl-2,2-dimethylcyclopropanoate

Adult corn rootworms are attracted strongly to a variety of kairomones, including indole [*120-72-9*], estragole [*140-67-0*], and 4-methoxycinnamaldehyde [*71277-11-7*] for the western corn rootworm; eugenol and *trans*-cinnamyl alcohol [*104-54-1*] for the northern corn rootworm; and cinnamaldehyde [*104-55-2*] for the southern corn rootworm.

Indole

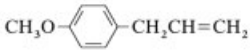

Estragole

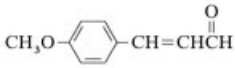

4-Methoxycinnamaldehyde

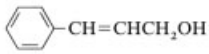

Cinnamyl alcohol

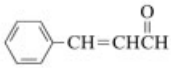

Cinnamaldehyde

Attractants have proved especially effective for the control of fruit flies (Diptera: Tephritidae). Methyl eugenol [*93-15-2*], eugenol methyl ether, is very strongly attractive for the oriental fruit fly *Dacus dorsalis* and many closely related species. Raspberry ketone [*5471-51-2*], 1-(4-hydroxyphenyl)-2-butanone, is a strong attractant for the melon fly *D. cucurbitae* and a large group of its close relatives. The 1-(4-acetyl)derivative, cue-lure [*3572-06-3*], is more volatile and is a synthetic kairomone especially effective for monitoring and for bait traps. Methyl eugenol and cue-lure have been used for male annihilation programs for several *Dacus* spp. by applying them to fiberboard blocks containing the volatile attractants at 15 g/ha and contact insecticides such as malathion or naled at 1 g/ha.

Similarly, the Mediterranean fruit fly *Ceratitis capitata* is attracted strongly to the kairomone trimedlure [*12002-53-8*], *tert*-butyl 2-methyl-4-chlorocyclohexanoate. Mixtures of apple volatiles including butyl 2-methylbutanoate, propyl hexanoate, butyl hexanoate, hexyl propanoate, and benzyl butanoate are used to trap the apple maggot *Rhagoletis pomonella*; dipropyl disulfide [*629-19-6*], $C_3H_7S{-}SC_3H_7$, is an attractant for the onion maggot fly *Hylemya antiqua*; and allyl isothiocyanate [*57-06-7*], 3-isothiocyanato-1-propene, for the cabbage maggot fly *H. brassicae*. Yellow jackets, *Vespula* spp. (Hymenoptera: Vespidae), are trapped with heptyl butyrate [*5870-93-9*], $C_3H_7COOC_7H_{15}$.

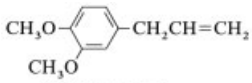
Methyl eugenol

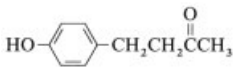
Raspberry ketone

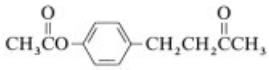
Cue-lure

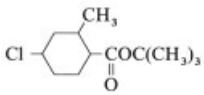
Trimedlure

16. Insecticide Formulation

Effective use of any chemical for insect control depends on its formulation into a product that guarantees safety to the applicator, to plants, and to domestic animals. Insecticides are commonly formulated as dust, water-dispersible powder, emulsive concentrates, solution, and bait. Preparation of such formulations involves the use of accessory agents such as dust carriers, solvents, emulsifiers, wetting and dispersing agents, stickers, deodorants or masking agents, and attractants or feeding stimulants.

Dusts are the simplest formulations and generally contain low concentrations (1 – 20 %) of the active ingredients, although ground botanical preparations such as pyrethrum, derris, or ryania may be used undiluted. Dust carriers include organic flours, sulfur, silicon oxides, lime, gypsum, talc, pyrophyllite, bentonites, attapulgite, and volcanic ash. They are selected on the basis of compatibility with the active ingredient (including pH, moisture content, and stability), particle size, abrasiveness, absorbability, density, wettability, and cost. The mixture of active and inactive ingredients is produced by operations including solvent impregnation, fusing, grinding, and milling. Particle diameter is usually 0.5 – 4.0 µm.

Granulars are pelleted mixtures containing 2 – 20 % active ingredients in a carrier of absorptive clay, bentonite, or diatomaceous earth, with particle diameter ranging from 250 – 420 µm (40 – 60 mesh) to 420 – 840 µm (10 – 40 mesh). Ground organic matter such as corn grits or corn cobs is also used. Granulars are prepared by impregnation of the carrier with a solution or slurry of the active ingredient, and are used chiefly as soil insecticides or mosquito larvicides.

Water-dispersible powders contain the active ingredient in high concentration (15 – 95 %), with a dust carrier such as attapulgite which wets and suspends well in water. From 1 to 2 % of a surfactant is usually added to improve suspension and wetting of the carrier. Wettable powders are used widely for spraying plants and animals because of their relative safety.

Emulsives are emulsions of the active ingredient in water-immiscible organic solvents, commonly at 15 – 50 %, with a small percentage of surfactant to promote emulsification, wetting, and spreading. Factors influencing the choice of solvent are solubility, safety to plants and animals, volatility, flammability, compatibility, odor, and cost. The most commonly used solvents are kerosene, xylene, and related petroleum fractions; methyl isobutyl ketone; and amyl acetate. Water emulsion sprays are used widely for plant protection and household insect control.

Slow-release formulations are especially useful for compounded pheromone and

Table 5. Average organochlorine insecticide residues in U.S. inhabitants (1964) ∗

Insecticide	Food, daily intake, mg	Human fat, mg/kg	Human blood, mg/L	Human milk, mg/L
DDT	0.037	2.3	0.0068	0.022
DDE	0.024	8.0	0.0114	0.041
Aldrin, dieldrin	0.008	0.29	0.0014	0.0073
Heptachlor, heptachlor epoxide	0.003	0.24	0.0008	0.0027
BHC	0.004	0.60	0.0031	0.0024

∗ Data from [121–123].

kairomone attractants, in which reduced volatility of the active ingredient is important and decreased toxicity is essential. Formulations are made with polymeric carriers such as starch, cross-linked polyamides, carboxylated polyacrylates, or polyisobutylenes. Incorporation of the active ingredient in poly(vinyl chloride) (e.g., dichlorvos), or in hollow microfibers is used to release the sex pheromone attractants of the gypsy moth *Porthetria dispar* or the pink bollworm *Pectinophora gossypiella* for male confusion.

Bait. The use of bait-containing attractants and feeding stimulants together with a minimum amount of insecticide is a growing technology for integrated pest management. Examples include dried ground apple (pomace) and carbaryl for the black cutworm *Agrotis ipsilon*, soybean oil and amdro for the imported fire ant *Solenopsis invicta*, cue-lure and naled for the melon fly *Dacus cucurbitae*, and methyl eugenol and naled for the oriental fruit fly *D. dorsalis*. The use of bait greatly reduces the amount of toxic chemical applied and improves the selectivity of the insect pest management procedure.

17. Insecticide Residues in Foods

[118], [119]

17.1. Residue Persistence

Approximately 70 % of all insecticides are used in agriculture where they are typically applied directly to raw agricultural commodities for plant and animal protection. Despite weathering of crop residues during maturation and further attenuation during food processing and preparation, measurable amounts of insecticides (i.e., pesticide residues) are detectable in food. A pesticide residue is defined by the World Health Organization as "any substance . . . resulting from the use of a pesticide [that] includes any specified derivatives such as degradation products, metabolites, and reaction products which are of toxicological significance" [120].

Residues of persistent pesticides in the human diet are absorbed readily, stored in human adipose tissue, and secreted in body fluids such as milk [121–123]. The effects of 20 years of contamination of the U.S. diet by persistent organochlorine insecticide residues are summarized in Table 5. Withdrawal of these insecticides in the 1970s was followed by a decline in the average residue in human adipose tissue, as shown by comparison of surveys made in 1970 and 1978: DDT(T) 7.5 vs. 3.6 mg/kg, dieldrin 0.44 vs. 0.19 mg/kg, and β-BHC 0.44 vs. 0.19 mg/kg[124].

Residue Persistence Curve. The magnitude and duration of pesticide residues on raw agricultural commodities depend on the dosage applied, the chemical nature of the insecticide, and the physical nature of the treated plant surface. Standard technology involves determination of the content of insecticide residue, in milligrams per kilogram, in produce at intervals after application. Because of the curvilinear nature of the results (Fig. 2A), estimation of the rate of residue loss is difficult. For rapidly degradable insecticides (most organophosphates and carbamates), residue persistence follows first-order kinetics. When plotted on a semilogarithmic scale (Fig. 2B), straight lines are obtained whose statistical reliability can be determined

Table 6. FAO/WHO acceptable daily intakes (ADI) and maximum residue limits (MRL) together with EPA residue tolerances for various insecticides

Insecticide	Trade name	ADI, mg kg^{-1} d^{-1}	MRL, mg/kg	Tolerance, mg/kg
Organochlorines				
Aldrin/dieldrin		0.0001		0 – 0.1 [a]
Chlordane		0.001		0.3 [a]
DDT		0.005		7 [a]
Endosulfan	Thiodan	0.0075		0.1 – 2
Endrin		0.0002		
Heptachlor		0.0005		0 – 0.1 [a]
Lindane		0.0125		1 – 7
Methoxychlor	Marlate	0.1		1.25 – 100
Chlorobenzilate		0.02		0.2 – 5
Dicofol	Kelthane	0.025		0.1 – 30
Toxaphene				0.3 – 7 [a]
Chlorbenside		0.01		
Tetradifon	Tedion			1 – 100
Ovex	Chlorfenson, Ovotran	0.01		3 – 5
Organophosphates				
Acephate	Orthene	0.0005 [b]	0.1 – 10	0.5 – 15
Azinphos methyl	Guthion	0.0025	0.2 – 4	0.2 – 10
Bromophos	Nexion	0.04	0.5 – 5	
Carbophenothion	Trithion	0.0005	0.2 – 2	0.1 – 10
Chlorfenvinphos	Supona, Birlane	0.002	0.05 – 1	
Chlorpyrifos	Dursban, Lorsban	0.01	0.05 – 2	0.05 – 1.5
Coumaphos	Co-ral, Asuntol	0.0005 [c]		0.1 – 10
Cruformate	Ruelene	0.1		1
Demeton	Systox	0.005 [c]	0.1 – 1	0.3 – 12
Dialifor	Torak	0.003 [c]	1 – 40	3
Diazinon	Basudin, Spectracide	0.003	0.5 – 2	0.1 – 60
Dichlorvos	Vapona, DDVP	0.004	0.1 – 2	0.02 – 1
Dicrotophos	Bidrin			0.05
Dimethoate	Cygon, Rogor	0.002 [b]	0.05 – 2	0.002 – 5
Dioxathion	Delnav	0.0015	0.1 – 5	0.14 – 4.9
Disulfoton	Thiodemeton	0.002	0.1 – 5	0.1 – 12
Ethion		0.0005 [b]	0.1 – 2	0.1 – 14
Ethoprop	Mocap		0.02	0.02
Fenaminphos		0.0003 [b]	0.05 – 0.5	
Fenchlorphos	Ronnel, Korlan, Nankor	0.01	0.01	0 – 10
Fenitrothion	Sumithion, Folithion	0.003 [b]	0.05 – 2	
Fensulfothion	Dasanit	0.0003	0.02 – 0.1	0.02 – 5
Fenthion	Baytex	0.001	0.05 – 2	0.01 – 18
Fonofos	Dyfonate			0.1 – 0.5
Formothion		0.02	0.2	0.2
Isofenphos	Amaze	0.0005 [b]	0.02 – 1	
Malathion	Cythion	0.02	0.5 – 8	8 – 135
Methamidophos	Monitor, Tamaron	0.0006	1 – 2	0.1 – 10
Methidathion	Supracide	0.005	0.2 – 5	0.2 – 6
Mevinphos	Phosdrin	0.0015	0.05 – 1	0.24 – 4
Monocrotophos	Azodrin	0.0006	0.05 – 1	0.05 – 5
Naled	Dibrom			0.5 – 3
Omethoate	Folimat	0.0003	0.05 – 2	0.5 – 3
Parathion, ethyl	Alkron, Niron	0.005	0.5 – 1	0.1 – 5
Parathion, methyl	Metacide	0.02	0.05 – 0.2	0.1 – 5
Phorate	Thimet	0.0002	0.1 – 1	0.05 – 1.5
Phosalone	Zolone	0.006	5 – 10	0.05 – 85
Phenthoate	Cidial	0.003	1	

Table 6. (Continued)

Insecticide	Trade name	ADI, mg kg^{-1} d^{-1}	MRL, mg/kg	Tolerance, mg/kg
Phosphamidon	Dimecron	0.0005	0.05 – 0.5	0.01 – 1
Phosmet	Imidan	0.02	0.05 – 15	0.1 – 4
Pirimiphos methyl	Actellic	0.01	0.05 – 5	
Thiometon	Morphothion, Ekatin M	0.003		
Trichlorfon	Dylox, Dipterex	0.01	0.05 – 2	0.1 – 240
Temephos	Abate			0.1
Stirofos	Rabon, Gardona			0.1 – 110
Phoxim	Volaton	0.001	0.05 – 2	
Leptophos	Phosvel	0.001 [b]		
Triazophos	Hostathion	0.0002	0.05 – 2	
Carbamates				
Aldicarb	Temik	0.005	0.1 – 0.5	0.002 – 1
Bendiocarb	Ficam	0.002 [b]		
Carbaryl	Sevin	0.01	0.1 – 2.5	0.5 – 100
Carbofuran	Furadan	0.01		0.02 – 20
Ethiofencarb	Crotenon	0.1		
Methomyl	Lannate			0.1 – 10
Oxamyl	Vydate	0.03		
Propoxur	Baygon	0.02		0.05 – 5
Pirimicarb	Pirimin	0.02		
Pyrethroids				
Allethrin	Pynamin			2 – 4
Deltamethrin	Decis	0.01	0.02 – 0.2	
Permethrin	Pounce, Ambush	0.05	0.1 – 1	
Pyrethrins		0.04	1 – 2	1 – 3
Cypermethrin	Ripcord	0.05	0.01 – 0.2	
Fenvalerate	Pydrin, Sumicidin	0.02	0.1 – 10	
Miscellaneous				
Chlordimeform	Galecron, Fundal	0.01		0.5 – 25
Binapacryl	Morocide	0.0025		
Methoprene	Altosid			0.01 – 1
Piperonyl butoxide				0.1 – 20

[a] Cancelled. [b] Temporary. [c] Withdrawn. [a] Cancelled. [b] Temporary. [c] Withdrawn.

by linear regression analysis. With highly persistent insecticides such as organochlorines and certain pyrethroids, residue persistence may follow second-order kinetics due to initial photochemical or hydrolytic effects and the formation of persistent metabolites, as well as to gradual compartmentalization of the residue in plant waxes or sequestration in oil glands.

Safe Interval. The linear, semilogarithmic residue persistence curve (Fig. 2) provides ready determination of the residue half-life ($R_{1/2}$), which is independent of application rate. From such residue persistence curves, the required interval after application for residue attenuation to toxicologically safe levels (maximum residue limits) can be determined easily and expressed as a safe interval before harvest on the pesticide label directions.

17.2. Environmental Chemistry of Insecticide Residues [125][126]

Exposure to air, light, and moisture, and biochemical interactions with enzymes in plant and animal products, cause insecticide residues to undergo inactivating or activating reactions, generally within periods of days or weeks.

17.2.1. Inactivation

Inactivation processes result in the formation of chemical residues that no longer exert specific target site interactions (e.g., inhibition of acetylcholinesterase by an organophosphate or carbamate). Inactivation processes typically attack the chemical bonds of degradophores (i.e., chemical groups sensitive to hydrolysis, oxidation, reduction, or conjugation) so that the insecticide

structure is altered to form smaller, more water-soluble moieties.

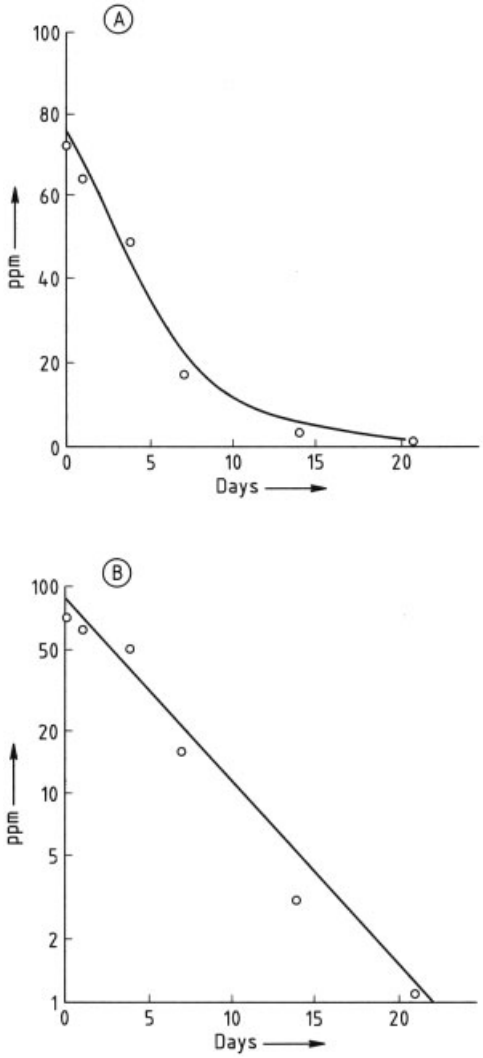

Figure 2. Residue persistence curves for carbaryl on corn, *Zea mays*
A) Degradation of residue vs. time; B) Semilogarithmic transformation [134]

Hydrolysis. Highly lipophilic, organochlorine insecticides with strong C–Cl bonds are not subject to rapid environmental hydrolysis, which accounts for their lack of biodegradability. On the other hand, the three C–O–P ester links in organophosphate insecticides are hydrolyzed readily, which accounts for their general biodegradability. Parathion, for example, is hydrolyzed principally at the reactive P–O–aryl bond to form diethylphosphate and 4-nitrophenolate. More profound hydrolysis may split P–O–alkyl bonds of organophosphates to form partial phosphate esters. The reduced toxic hazard associated with malathion is the result of hydrolysis at both carbethoxy and phosphorothiolate ester bonds.

Hydrolysis is also a major detoxification pathway for carbamates; for example, carbaryl is hydrolyzed to 1-naphthol and methylamine.

Oxidation occurs readily via light, atmospheric oxidants, and microsomal processes catalyzed by cytochrome P 450 enzymes in plant and animal tissues where $\cdot OH$ is the initiating radical. Common oxidation reactions include the following [125]:

1) Desulfuration of phosphorothionates (P=S) to phosphates (P=O)
2) Oxidation of thioethers (C–S–C) to sulfoxide (C–SO–C) and sulfone (C–SO_2C) (see Section 6.9.1)
3) Ring hydroxylation of aromatic rings (e.g., carbaryl to naphthalene diols, permethrin and fenvalerate to phenols)
4) Side chain oxidation of $CH_3 \rightarrow CH_2OH \rightarrow COOH$ (e.g., in pyrethrins, carbamates)
5) O-Dealkylation of aryl methoxy (CH_3O) to phenol (HO) (e.g., in methoxychlor)
6) N-Dealkylation of $NHCH_3$ to NH_2 (e.g., in carbamates)
7) Epoxidation of C=C characteristic of cyclodienes (e.g., aldrin to dieldrin, and heptachlor to heptachlor epoxide)

Reduction commonly occurs only under anaerobic conditions (e.g., reduction of the NO_2 group of parathion to NH_2 in the animal rumen).

Dehalogenation occurs by glutathione conjugation of organochlorines (e.g., DDT is converted to DDE; lindane, to trichlorothiophenol).

Conjugation by enzymatic action enhances water solubility and excretion in animals. Phenols are conjugated with glucosides, glucuronides, sulfates, and phosphates; carboxylic acids with glycine.

17.2.2. Activation

Activation is a special residue reaction in which the first transformation produces an insecticide derivative that is of equal or enhanced toxicity:

1) Epoxidation of cyclodienes (e.g., aldrin to dieldrin and heptachlor to heptachlor epoxide) occurs rapidly on plant surfaces, in the soil, or in animal tissues. The epoxides are more persistent and usually more toxic than their precursors.
2) Phosphorothioates are oxidized to phosphates which are over a thousand fold more reactive with the target site, acetylcholinesterase (see Section 6.9.1).
3) Sulfides are oxidized readily to sulfoxides, and the enhanced electron withdrawal activates the phosphorus atom of many alkylthiophosphates such as fenthion, demeton, disulfoton, and phorate (see Section 6.9.1).
4) *N*-Dealkylation of carboxamide organophosphates enhances their toxicity (e.g., dicrotophos to monocrotophos).

17.3. Toxicology of Insecticide Residues [128]

No Observed Adverse Effect Level (NOAEL). Determination of the NOAEL for a pesticide fed over the lifetime of at least two and preferably three species of laboratory animal (mouse, rat, dog) provides the basis for risk assessment in predictive toxicology. The NOAEL, in milligrams of pesticide ingested per kilogram of body weight per day, is determined by lifetime (two-year) feeding studies with a range of dosages. Evaluated factors include general health, food intake, weight gain, blood chemistry, enzyme activity, organ weight, and gross histopathology of organs and tissues. The NOAEL is the highest dosage administered that does not produce adverse (toxicological) effects.

Acceptable Daily Intake (ADI) [120], [127–129]. To determine the safe dosage for humans exposed to pesticide residues, the NOAEL obtained from animal feeding studies is divided by an additional safety factor to correct for both interspecies and intraspecies variations. The result is the acceptable daily intake (ADI). Three levels of safety factors are recognized:

1) a safety factor of 10 when valid chronic exposure data for humans are available,
2) a safety factor of 100 when no valid human exposure data exist, but adequate chronic exposure data are available from studies with laboratory animals, and
3) a safety factor of 1000 when no adequate chronic exposure data exist.

The ADI is an estimate of the daily intake of a pesticide, in milligrams per kilogram per day, which over an entire human lifetime appears to be without appreciable risk. Leadership in the establishment of ADI values is taken by the Joint FAO – WHO Meeting on Pesticide Residues, which periodically reviews and revises ADI values as pertinent scientific data become available. ADI values for many insecticides are presented in Table 6.

Maximum Residue Limits, Residue Tolerances [120], [129–134]. The basic guidelines for estimating the risk from human exposure to pesticide residues in food are called the maximum residue limits (MRL) or tolerances. They are established by various national and international agencies. The MRL value represents the smallest amount of pesticide that is practical and toxicologically acceptable [128]. The MRL values (in parts per million) for agricultural commodities are determined as the residues resulting after a pesticide has been used according to good agricultural practice, given the minimum quantities necessary to achieve adequate pest control. These definitions leave substantial latitude, and differing MRL values have been established for similar uses in various countries. However, standardization is highly desirable because of international trade in agricultural commodities, and the Joint FAO – WHO Codex Committee on Pesticide Residues has established international MRL values for more than 100 pesticides in $> 1\,600$ specific foods. These values are reviewed and revised at intervals as new information becomes available [131].

In the United States, registration and regulation of pesticide tolerance are the responsibility of the Environmental Protection Agency, under

the Federal Insecticide, Fungicide, and Rodenticide Act (1972) [132], [133]. This law requires "that a pesticide may be registered if it is effective, properly labeled, and when properly used will not cause unreasonable adverse effects on the environment taking into account the economic, social and environmental costs and benefits of the use of any pesticide." The EPA residue tolerance values for most insecticides are presented in Table 6.

Carcinogenicity. The Food Additive Amendment (1958) to the U.S. Food, Drug, and Cosmetic Act (1938) states "that no additive shall be deemed to be safe if it is deemed to cause cancer when ingested by man or animal." This concept has been used by EPA to cancel registration and tolerance for organochlorine insecticides such as DDT, aldrin, dieldrin, chlordane, heptachlor, and toxaphene.

18. References

1. R. L. Metcalf, R. A. Metcalf: *Destructive and Useful Insects,* 5th ed., McGraw-Hill, New York 1993.
2. H. H. Cramer: "Plant Protection and World Crop Production," *Pflanzenschutz Nachr.* **20** (1967).
3. J. W. Wright in R. L. Metcalf, J. J. McKelvey, Jr. (eds.): *The Future for Insecticides,* Wiley-Interscience, New York 1976, p. 17.
4. World Health Organization: "Arthropod-Borne and Rodent-Borne Viral Diseases," *W.H.O. Tech. Rep. Ser.* **719** (1985).
5. K. H. Büchel (ed.): *Chemistry of Pesticides,* Wiley-Interscience, New York 1983.
6. D. E. H Frear: *Chemistry of Pesticides,* 3rd ed., Van Nostrand Princeton, N.J., 1955.
7. N. N. Melnikov: *Chemistry of Pesticides (Eng. Transl.),* Springer, New York 1971.
8. R. L. Metcalf: *Organic Insecticides,* Interscience, New York 1955.
9. R. D. O'Brien: *Insecticides: Action and Metabolism,* Academic Press, New York 1967.
10. W. Perkow: *Wirksubstanzen der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* Parey, Berlin 1976.
11. H. H. Shepard: *The Chemistry and Action of Insecticides,* McGraw-Hill, New York 1951.
12. E. Y. Spencer: *Guide to the Chemicals Used in Crop Protection,* 7th ed., Research Branch Agriculture, Canada, 1980.
13. R. Wegler (ed.): *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 1,** 1970; vol 3, 1976; Springer-Verlag, Berlin-Heidelberg.
14. R. L. Metcalf, *J. Agr. Food Chem.* **21** (1973) 511.
15. A. W. A. Brown, J. Hayworth, A. R. Zahar, *J. Med. Entomol.* **13** (1976) 1.
16. A. W. A. Brown, R. Pal: *Insect Resistance in Arthropods,* WHO, Geneva 1971.
17. G. P. Georghiou, T. Saito: *Pest Resistance to Pesticides,* Plenum Publishing, New York 1984.
18. G. P. Georghiou, A. Lagunes-Tejeda: "The Occurrence of Resistance to Pesticides in Arthropods," *FAO Plant Protection Series,* Rome 1991.
19. National Academy of Sciences: *Pesticide Resistance: Strategies and Tactics for Management,* Washington, D.C., 1986.
20. R. L. Metcalf, *Annu. Rev. Entomol.* **25** (1980) 219.
21. R. L. Metcalf in M. Kogan (ed.): *Ecological Theory and Integrated Pest Management Practice,* Wiley-Interscience, New York 1986.
22. A. W. A. Brown: *Ecology of Pesticides,* Wiley-Interscience, New York 1978.
23. C. A. Edwards (ed.): *Environmental Pollution by Pesticides,* Plenum Publishing, New York 1973.
24. C. A. Edwards: *Persistent Pesticides in the Environment,* 2nd ed., CRC Press, Cleveland 1973.
25. M. A. Q. Khan (ed.): *Pesticides in Aquatic Environments,* Plenum Publishing, New York 1977.
26. F. L. McEwen, G. R. Stephenson: *The Use and Significance of Pesticides in the Environment,* Wiley-Interscience, New York 1979.
27. D. Pimentel: *Ecological Effects of Pesticides on Non-Target Species,* President's Office Science, Technology, Washington, D.C., 1971.
28. T. R. Eichers, P. A. Andrilenas, T. W. Anderson: "Farmers' Use of Pesticides in 1976," *USDA Agr. Econ. Rept. No. 418,* Washington, D.C., 1978.
29. J. F. Copplestone in D. L. Watson, A. W. A. Brown (eds.): *Pesticide Management and Insecticide Resistance,* Academic Press, New York 1977, p. 147.
30. M. L. Flint, R. van den Bosch: *Introduction to Integrated Pest Management,* Plenum Publishing, New York 1981.

31. C. B. Huffaker (ed.): *New Technology of Pest Control,* Wiley-Interscience, New York 1980.
32. M. Kogan (ed.): *Ecological Theory and Integrated Pest Management Practice,* Wiley-Interscience, New York 1986.
33. R. L. Metcalf, W. H. Luckmann (eds.): *Introduction to Insect Pest Management,* 3rd ed., Wiley-Interscience, New York 1994.
34. National Academy of Sciences: *Scientific Aspects of Pest Control,* Publ. 1402, Washington, D.C., 1966.
35. National Academy of Sciences: *Insect Pest Management and Control,* Publ. 1695, Washington, D.C., 1969.
36. National Academy of Sciences: *Pest Control Strategies for the Future,* Washington, D.C., 1972.
37. R. Carson: *Silent Spring,* Houghton-Mifflin, Boston 1962.
38. R. L. Metcalf, *Annu. Rev. Entomol.* **25** (1980) 219.
39. R. L. Metcalf in R. L. Metcalf, W. H. Luckmann (eds.): *Introduction to Insect Pest Management,* 3rd ed., Wiley-Interscience, New York 1994.
40. D. L. Watson, A. W. A. Brown: *Pesticide Management and Insecticide Resistance,* Academic Press, New York 1977.
41. M. Jacobsen, D. G. Crosby (eds.): *Naturally Occurring Insecticides,* Marcel Dekker, New York 1973.
42. J. E. Cassida (ed.): *Pyrethrum, the Natural Insecticide,* Academic Press, New York 1971.
43. C. B. Gnadinger: *Pyrethrum Flowers,* 2nd ed., McLaughlin, Gormley, King, Co., Minneapolis 1945.
44. M. Elliott in R. L. Metcalf, J. J. McKelvey, Jr. (eds.): *The Future for Insecticides,* Wiley-Interscience, New York 1976, p. 163.
45. M. Elliott, *Bull. W.H.O.* **44** (1971) 315.
46. R. A. Fuchs, R. Schroder in K. H. Buchel (ed.): *Chemistry of Pesticides,* Wiley-Interscience, New York 1983.
47. T. Narahashi in R. L. Metcalf, J. J. McKelvey, Jr. (eds.): *The Future for Insecticides,* Wiley-Interscience, New York 1976, p. 371.
48. C. F. Wilkinson in R. L. Metcalf, J. J. McKelvey, Jr. (eds.): *The Future for Insecticides,* Wiley-Interscience, New York 1976, p. 195.
49. G. T. Brooks: *Chlorinated Insecticides,* **vols. 1, 2,** CRC Press, Cleveland 1974.
50. R. L. Metcalf in R. L. Metcalf, J. J. McKelvey, Jr. (eds.): *The Future for Insecticides: Needs and Prospects,* Wiley-Interscience, New York 1976, p. 223.
51. P. Müller (ed.): *DDT Insektizid Dichlordiphenyltrichloroäthan und seine Bedeutung,* Birkhäuser, Basel 1958.
52. J. E. Cassida, L. J. Lawrence, *Environ. Health Perspect.* **61** (1985) 123.
53. G. Schrader: *Die Entwicklung neuer insektizider Phosphorsäure-Ester,* 3rd ed., Verlag Chemie, Weinheim 1963.
54. C. Fest, K.-J. Schmidt: *The Chemistry of Organophosphorus Pesticides,* 2nd ed., Springer-Verlag, New York 1980.
55. M. Eto: *Organophosphorus Pesticides: Organic and Biological Chemistry,* CRC Press, Cleveland 1974.
56. R. D. O'Brien: *Toxic Phosphorus Esters,* Academic Press, New York 1960.
57. WHO Environmental Health Criteria 63: *Organophosphorus Insecticides,* WHO, Geneva 1986.
58. W. Lange, G. von Kreuger, *Chem. Ber.* **65** (1932) 1998.
59. R. L. Metcalf, *Bull. Entomol. Soc. Am.* **5** (1959) 3.
60. T. R. Fukuto, *Bull. W.H.O.* **44** (1971) 31.
61. T. R. Fukuto in C. F. Wilkinson (ed.): *Pesticide Biochemistry and Physiology,* Plenum Publishing, New York 1978.
62. R. D. O'Brien in C. F. Wilkinson (ed.): *Pesticide Biochemistry and Physiology,* Plenum Publishing, New York 1976, p. 271.
63. T. Namba, *Bull. W.H.O.* **44** (1971) 289.
64. S. D. Murphy in C. D. Klassen, M. O. Amdur, J. Doull (eds.): *Toxicology,* 3rd ed., Macmillan Publ. Co., New York 1986, p. 519.
65. R. L. Metcalf, *Neurotoxicology* **3** (1982) 269.
66. M. K. Johnson: "The Delayed Neuropathy Caused by Some Organophosphorus Esters: Mechanism and Challenge," *CRC Crit. Rev. Toxicol.* **3** (1975) 289.
67. R. J. Kuhr, H. W. Dorrough: *Carbamate Insecticides, Chemistry, Biochemistry, and Toxicology,* CRC Press, Cleveland 1976.
68. R. L. Metcalf, T. R. Fukuto, *J. Agric. Food Chem.* **13** (1965) 220.
69. R. L. Metcalf, *Bull. W.H.O* **44** (1971) 43.
70. WHO Environmental Health Criteria 64: *Carbamate Pesticides: a General Introduction,* Geneva 1986.
71. R. L. Metcalf, T. R. Fukuto, *J. Agric. Food Chem.* **15** (1967) 1022.
72. H. Gysin, *Chimia* **8** (1954) 205.
73. M. J. Kolbezen, R. L. Metcalf, T. R. Fukuto, *J. Agric. Food Chem.* **2** (1954) 864.

74. T. A. Magee in J. R. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p. 71.
75. C. O. Knowles in J. R. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p. 243.
76. R. M. Hollingworth, A. E. Lund in J. R. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p. 190.
77. F. Matsumura, R. W. Beeman in J. R. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p. 231.
78. V. B. Wigglesworth, *J. Insect Physiol.* **15** (1969) 73.
79. L. I. Gilbert (ed.): *The Juvenile Hormones,* Plenum Publishing, New York 1976.
80. W. S. Bowers, *Bull. W.H.O.* **44** (1971) 381.
81. G. B. Stahl, *Annu. Rev. Entomol.* **20** (1975) 417.
82. J. P. Edwards, J. J. Menn in R. Wegler (ed.): *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 6,** Springer Verlag, Berlin 1981, p. 185.
83. N. Wakaboyashi, R. M. Waters in E. D. Morgan, B. Mandava (eds.): *Handbook of Natural Pesticides,* vol. 1, CRC Press, Boca Raton, Florida, 1987.
84. E. P. Marks, T. Leighton, F. Leighton in J. L. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p.283.
85. L. C. Post, B. J. deJong, W. R. Vincent, *Pestic. Biochem. Physiol.* **4** (1974) 473. C. A. Hendrick in J. L. Coats (ed.): *Insecticide Mode of Action,* Academic Press, New York 1982, p. 315.
86. A. Verloop, C. D. Ferrell in J. R. Plimmer (ed.): *Pesticide Chemistry for the 20th Century,* Am. Chem. Soc., Washington, D.C., 1977.
87. H. Schmutterer: *The Neem Tree,* VCH, Weinheim1995.
88. E. A. Steinhaus: *Insect Pathology: An Advanced Treatise,* **vol. 1, 2,** Academic Press, New York 1973.
89. E. Kurstak (ed.): *Microbial and Viral Pesticides,* Marcel Dekker, New York 1982.
90. C. M. Ignoffo in M. Jacobsen (ed.): *Insecticides of the Future,* Marcel Dekker, New York 1975.
91. R. A. Dyback in P. Doyle, T. Fujita (eds.): *Pesticide Chemistry: Human Welfare and the Environment, Synthesis and Structure Activity Relationships,* **vol. 1,** Pergamon Press, Oxford 1982, p. 83.
92. C. C. Beegle, L. C. Lewis, T. Yamamoto in C. M. Ignoffo (ed.): *Handbook of Naturally Occurring Pesticides: Microbial Insecticides,* CRC Press, Boca Raton, Florida, 1986.
93. E. F. Knipling: "The Basic Principles of Insect Population Suppression and Control," *USDA Agricultural Handbook,* no. 512, Washington, D.C., 1979.
94. M. A. Hoy, J. J. McKelvey, Jr.: *Genetics in Relation to Insect Management,* Rockefeller Foundation, New York 1979.
95. A. H. Baumhover, *JAMA J. Am. Med. Assoc.* **196** (1966) 240.
96. R. Pal, M. J. Whitten (eds.): *The Use of Genetics in Insect Control,* Elsevier, Amsterdam 1974.
97. J. W. Wright, R. Pal (eds.): *Genetics of Insect Vectors of Disease,* Elsevier, Amsterdam 1967.
98. R. Pal, L. E. LaChance, *Annu. Rev. Entomol.* **19** (1974) 269.
99. A. B. Borkovic: *Insect Chemosterilants,* vol. 7, Adv. Pest Control Res., Wiley-Interscience, New York 1966.
100. G. C. LaBrecque, C. N. Smith: *Principles of Chemosterilization,* Elsevier, Amsterdam 1968.
101. D. G. Campion, *Bull. Entomol. Res.* **61** (1972) 577.
102. S. A. Hall, N. Green, M. Beroza, *J. Agr. Food. Chem.* **5** (1957) 66.
103. C. N. Smith, "Repellents for Anopheline Mosquitoes," *Misc. Publ. Entomol. Soc. Am.* **7** (1970) 99.
104. W. J. Bell, R. T. Carde (eds.): *Chemical Ecology of Insects,* Sinauer Assoc., Sunderland, Mass., 1984.
105. M. Beroza: "Attractants and Repellents for Insect Control," *in Pest Control Strategies for the Future,* Nat. Acad. Sci., Washington, D. C., 1972.
106. W. H. Bossert, E. O. Wilson, *J. Theor. Biol.* **5** (1963) 443.
107. V. G. Dethier: *The Physiology of Insect Senses,* Wiley-Interscience, New York 1963.
108. E. R. Mitchell (ed.): *Management of Insect Pests with Semiochemicals,* Plenum Publishing, New York 1981.
109. D. A. Nordlund, R. L. Jones, W. J. Lewis (eds.): *Semiochemicals: their Role in Pest Control,* Wiley-Interscience, New York 1981.
110. H. H. Shorey, J. J. McKelvey, Jr. (eds.): *Chemical Control of Insect Behavior,* Wiley-Interscience, New York 1977.
111. M. C. Birch: *Pheromones,* Elsevier, Amsterdam 1974.

112. J. M. Brand, J. C. Young, R. M. Silverstein, "Insect pheromones: a critical review of recent advances in their chemistry, biology, and application," *Prog. Chem. Org. Natural Prod.* **37** (1979) 1.
113. Y. Tamaki in G. A. Kerkut, L. I. Gilbert (eds.): *Comprehensive Insect Physiology, Biochemistry, and Pharmacology,* "Sex pheromones," vol. 9, Pergamon Press, Oxford 1985, p. 145.
114. J. H. Borden in G. A. Kerkut, L. I. Gilbert (eds.): *Comprehensive Insect Physiology, Biochemistry, and Pharmacology,* "Aggregation pheromones," vol. 9, Pergamon Press, Oxford 1985, 257.
115. P. A. Hedin, F. G. Maxwell, J. H. Jenkins in F. A. Maxwell, F. A. Harris (eds.): *Proc. Summer Inst. of Biol. Control of Plants, Insects, and Diseases,* "Insect plant attractants, feeding stimulants, repellents, deterrents, and other related factors affecting insect behavior," University Press, Jackson, Missisippi, 1974, p. 494.
116. R. L. Metcalf, E.. R.. Metcalf: *Plant Kairomones in Insect Ecology and Control,* Chapman & Hall, New York1992.
117. S. Finch in T. H. Coaker (ed.): *Applied Biology* "Chemical attraction of plant feeding insects to plants," V., Academic Press, New York 1980, p. 67.
118. F. A. Gunther, J. Gunther (eds.): *Residue Revs.,* **vols. 1 – 86,** Springer Verlag, Hamburg 1962 – 1987.
119. H. Maier-Bode: *Pflanzenschutzmittel-Rückstände,* E. Ulmer, Stuttgart 1965.
120. N. A. Smart, *Residue Rev.* **98** (1985) 99.
121. W. F. Durham, *Ann. N.Y. Acad. Sci.* **160** (1969) 183.
122. R. E. Duggan, *Ann. N.Y. Acad. Sci.* **160** (1969) 173.
123. A. Curley, R. Kimbrough, *Arch. Environ. Health* **18** (1969) 156.
124. T. Nakatsugawa, M. A. Morelli in C. F. Wilkinson (ed.): *Insecticide Biochemistry and Physiology,* Plenum Publishing, New York 1976, p. 61.
125. C. M. Menzie: *Metabolism of Pesticides,* U.S. Dept. Interior, Spec. Sci. Rept. Wildlife no. 96, 1966 ; no. 184, 1974, no. 212, 1978.
126. H. Aizawa: *Metabolic Maps of Pesticides,* Academic Press, New York 1982.
127. G. Vettorazzi: "Toxicological Decisions and Recommendations Resulting from the Safety Assessment of Pesticide Residues in Food," *CRC Crit. Rev. Toxicol.* **4** (1975) 125.
128. WHO/FAO Tech. Rept. Ser. no. 370, 1963; 417, 1969; 458, 1970; 474, 1971; 502, 1972; 525, 1973; 545, 1974; 574, 1975.
129. U.S. Environmental Protection Agency: *Compendium of Registered Pesticides,* Washington, D.C.
130. L. P. Van Dyk, I. H. Wiese, J. E. C. Muller, *Residue Rev.* **82** (1984) 37.
131. G. Vettorazzi: *International Regulatory Aspects of Pesticide Chemicals,* **vols. 1, 2,** CRC Press, Cleveland, Ohio, 1979.
132. R. A. Merrill in C. D. Klassen, M. O. Amdur, J. Doull (eds.): *Toxicology,* 3rd ed., McMillan Publ. Co., New York 1986, p. 917.
133. National Academy of Sciences: *Regulating Pesticides,* Washington, D.C., 1980.
134. P. Dahl, L. Rattan: *J. Assoc. Off. Anal. Chem.* **54** (1971) 1095.

Insect Repellents → Crop Protection
Insect Repellents → Insect Control
Instant Photography → Photography

Insulation, Electric

ADOLF KAMM, Prüfinstitut für elektrisches Verhalten, Leverkusen, Federal Republic of Germany (Chaps. 1 and 2 in part, 3.2, 3.3, and 4)

KARL HEINZ SCHÜLLER, Fachhochschule Nürnberg, Fachbereich Werkstofftechnik, Nürnberg, Federal Republic of Germany (Chaps. 1 and 2 in part, Section 3.1)

1. Introduction 2
2. Properties and Testing Methods . 3
2.1. Resistance to Tracking 4
2.2. Volume Resistivity 5
2.3. Surface Resistivity 6
2.4. Relative Permittivity 7
2.5. Dielectric Dissipation Factor . . . 7
2.6. Loss Index 8
2.7. Breakdown Voltage 8
2.8. Further Electrical Tests 9
2.9. Mechanical and Thermal Properties 11
3. Insulating Materials 12
3.1. Inorganic Insulating Materials . . 12
3.1.1. Insulating Materials Based on Natural Minerals 13
3.1.1.1. Mica and Related Products 13
3.1.1.2. Asbestos and Related Products . . . 14
3.1.2. Synthetic Inorganic Insulating Materials 14
3.1.2.1. Porcelain 15
3.1.2.2. Steatite 17
3.1.2.3. Materials for Electrical Heat Engineering 19
3.1.2.4. Alumina Ceramics 20
3.1.2.5. Special Oxide Ceramics 21
3.1.2.6. Insulating Materials Containing Titanium Oxide and Titanates 22
3.1.2.7. Nonoxide Ceramics 25
3.1.2.8. Glass 25
3.2. Organic Insulating Materials . . . 26
3.2.1. Thermoplasts 27
3.2.2. Thermosets 37
3.2.3. Elastomers 40
3.2.4. Polymer Processing 42
3.3. Liquid Insulating Materials 44
4. Selection of an Insulating Material 45
5. References 48

1. Introduction

All nonmetallic organic and inorganic materials are semiconductors. They provide a vastly greater resistance towards a flowing electric current than metallic conductors. Their resistivity may be 27 powers of ten greater than that of a conductor.

The differing electrical conduction behavior of conductors and semiconductors is a result of their atomic structure. In metallic conductors, the electric current results from the movement of free electrons. In contrast, the extremely small, but nevertheless existant conductivity of the so-called nonconductors arises from electrons and ions which migrate under the influence of electric fields. Since dissociation induced by cosmic radiation always generates a significant level of freely moving electrons and ions, an absolute nonconductor is not possible.

Within the field of electrotechnics, nonconductors (insulators) are generally employed to insulate the current-carrying components of conducting circuits from their surroundings. Insulators should prevent contact between conductors and energy losses caused by stray tracking currents. The function of insulators therefore is to prevent the undesirable flow of current between two conductors subjected to a potential difference. Insulators must possess a maximum electrical resistance under all the conditions to which the system can be exposed, e.g., at elevated temperature (electrothermal applications), at high voltages (energy technology), or high frequency (telecommunications). Insulation can be improved by appropriate design of the insulator. The choice of a suitable insulating material, however, remains critical for successful insulation. If insulating materials are used to fill the free space between the electrodes of capacitors,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_321

then insulators act as dielectrics and have the additional function of storing electrical energy.

Not all materials of high resistivity (see Section 2.2) are suitable for use as electrical insulating materials. Additional chemical, mechanical, thermal, and, depending upon the application involved, further properties are required. Stability toward various electrical and nonelectrical influences [e.g., toward electrical discharges in cavities and at interfaces (see Section 2.8), toward heat produced by dielectric dissipation (see Section 2.5) and toward ignition and burning] is of major importance. The working life of an insulated electric circuit depends mainly upon the thermal aging behavior. Although a material may be technically suitable as an insulator, unfavorable processing properties or the manufacturing price can restrict its applications.

The selection of an insulating material lies in the hands of the designer, provided the properties necessary for the function in question are guaranteed. To ensure this, materials are tested according to national standards or IEC publications. The results of these tests permit comparison of the insulating materials under the specified conditions. Data for a specific material cannot generally be applied to an insulating component because the component and its insulation properties are influenced by design and manufacturing processes. The final decision regarding functionality must be made by testing the component itself.

Insulating materials can be subdivided into inorganic and organic materials. Inorganic insulators include natural products such as mica or asbestos. Synthetic insulators made of ceramics, glass, or more recently glass ceramics are of greater significance, however. Organic insulating materials are primarily synthetic polymers. All of these insulating materials have a secure position within the electrical industry, because of their special material properties or the efficient processes used to manufacture components from them. Developments have afforded new materials with improved properties. Developments within the electronics and electrical industries make the availability of new improved insulating materials essential

In their original form, natural materials are seldom able to meet the diverse requirements demanded of an electric insulating material; however, purification, preparation, or chemical modification, often increase their value. Polymers (thermoplasts, thermosets, and elastomers) are mainly synthetically manufactured products. Electric insulating materials are generally industrial products of complex chemical structure which have been adapted to suit their intended application by treatment, modification, and blending.

Compared with metallic conductors and magnetic substances, insulating materials represent the most comprehensive material group within the field of electrotechnics and possess the greatest number of technically interesting properties. Nonelectrical properties have been considered here only to an extent sufficient to permit a brief characterization.

Groups of organic insulating materials are listed in Chapter 3 according to their chemical composition. In most cases, organic products based on rubber or other polymers can be modified by processing. Hence an absolute value for a property of a given insulator group does not exist. For this reason, data are not presented in tabular form but as performance relief diagrams. The values shown in each block diagram merely indicate the range of the properties of main interest. Organic products which find application in the electrotechnics industry are generally employed with additives, which can often considerably influence electrical and dielectric properties. The most common additives include processing aids, stabilizers, pigments, fillers, and flame retardants.

2. Properties and Testing Methods

The electrical behavior of electrical insulating materials results from the physical processes of dissociation, polarization, and ionization.

Certain assumptions about the formation of free electrons and bound charge carriers under the influence of electrical fields are associated with these terms. They are responsible for electrical resistance, dielectric polarization, and electrical breakdown. In the broadest sense, these measurable parameters can be regarded as physical properties. Electrical insulating materials are materials with a resistivity $>10^6\,\Omega\cdot\mathrm{m}$ ($>10^8\,\Omega\cdot\mathrm{cm}$). The energy values which the electrons in a solid are able to assume can be

classified in an energy step scheme known as the "band model". Various energy levels exist:

1) energy levels which are occupied by electrons in the ground state of the atom;
2) energy levels which are accessible to electrons after introduction of energy (excited state); and
3) energy levels which cannot be occupied because of the Pauli exclusion principle.

The description of conducting processes in a solid material involves only the outermost electron-filled energy levels (the basic or valence band) and the next highest permissible energy range. In an insulating material, the valence band is completely filled with electrons, whereas the conduction band is empty. Following the introduction of energy, for example in the form of thermal energy, electrons are excited from the valence band and enter the conduction band. Here, they possess increased mobility and consequently lead to electrical conductance. In this way, an insulating material can attain a semiconductor character at elevated temperature.

In an insulating material, a prohibited energy zone occurs between the valence and conduction bands which is so large that an activation energy of >2 eV is required for an electron to make the transition. For this reason, practically no electrons occupy the conduction levels at room temperature.

Test procedures exist for reproducing or simulating technical conditions. The principles of these tests are often not evident, but can also be traced back to physical processes. The most important process in low-voltage technology is that used to determine resistance to tracking. Additionally, methods are used to determine promotion of corrosion, to observe electrostatic charging behavior, to investigate behavior towards partial discharges, and to assess the damage caused by arcing. Studies of thermally stimulated discharge currents provide further physical information [22].

2.1. Resistance to Tracking

Tracking is the progressive decomposition of the surface of a solid insulating material caused by local discharge and subsequent development of a conducting or partially conducting tracking path [23]. The international report of the International Electrotechnical Commission (IEC Publication 664/664 A) classifies insulation materials for low-voltage technology into four groups, according to their comparative tracking indices (CTI) [24]. The electric strength and the insulation resistance are disregarded, but they should be in a range appropriate for an insulating material.

Insulating material I:	CTI = 600
Insulating material II:	CTI = 400 – 575
Insulating material III a:	CTI = 175 – 375
Insulating material III b:	CTI = 100 – 150

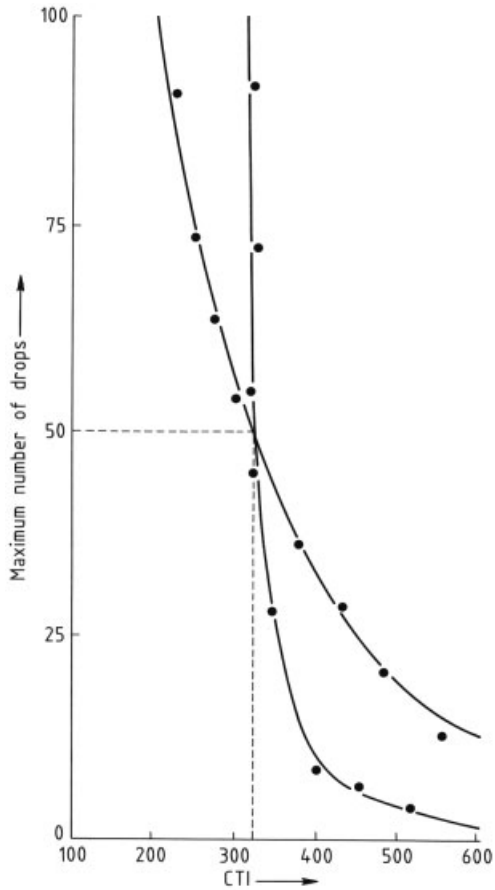

Figure 1. Diagram for determining the comparative tracking index (CTI) according to ASTM D 3638/UL 764 A
Curves are shown for two materials of differing behavior but with a CTI of 325.

The international standard IEC Publication 112-1979 stipulates conditions for measuring

resistance to tracking. The United States standard ANSI/ASTM 3638-1977 [25] forms the basis for tests according to the testing regulations ANSI/UL 746 A-1984 [26]; these differ from the currently valid international standard (IEC Publication 112-1979) in the type of evaluation procedure employed and in the lack of opportunity for using a more aggressive testing solution (B).

During testing, surface contamination is simulated by allowing an appropriate test fluid to drip onto the surface. The IEC Publication 112 specifies five separate determinations of a.c. voltage (50 – 60 Hz) at which 50 drops of test solution between the platinum electrodes do not lead to failure. In contrast, ASTM 3638-1977 stipulates one measurement at a number of test voltages and comparison with a standard curve. The point of intersection of this curve with the 50-drop line is determined on the plot of droplet number as a function of the testing voltage; the test voltage is read off the abscissa (voltage axis) and the CTI is thus determined (Fig. 1).

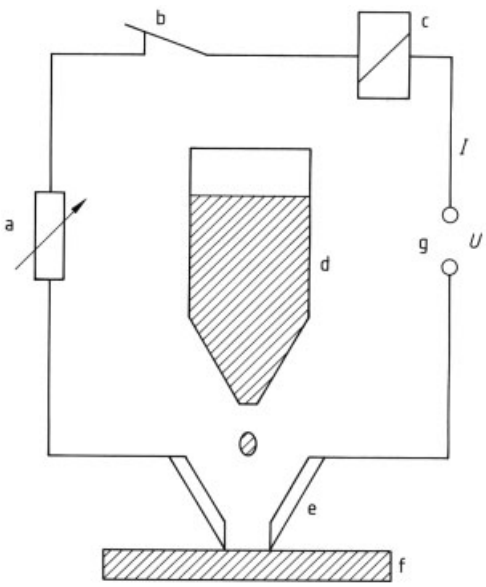

Figure 2. Schematic representation of measurement of low-voltage tracking
a) Current limiting resistor; b) Excess current switch is activated by relay (c) when $I \geq 0.5$ A for 2 s; c) Relay; d) Contamination fluid; e) Platinum electrode; f) Specimen; g) Voltage supply

Test solution A comprises 1 g of ammonium chloride dissolved in 1 L of deionized water. The concentration of ammonium chloride in testing solution B is identical, but a specified amount of a given wetting agent is also added. The testing solutions therefore differ in their electrolytic conductivity because the wetting agent introduces almost the same contribution to the total conductivity as the ammonium chloride salt. The testing solutions also differ in their surface tension.

According to the definition given in the standards, a track has developed when the tracking current on the tracking path is greater than or equal to 0.5 A for at least 2 s, or when the specimen begins to burn as a result of the heat generated. A circuit used for evaluating tracking is shown in Figure 2.

The test is carried out at voltage steps between 100 and 600 V. The maximum test voltage which can be divided by 25 without a remainder represents the comparative tracking index CTI, provided that the test specimen (which must be at least 3 mm thick) can also withstand a further five tests with 100 drops at a voltage 25 V lower than the maximum value. Should a sample not fulfill this requirement, the voltage must be reduced in steps of 25 V and the value must be determined at which 100 drops can be withstood without failure. The magnitude of this voltage is given in parentheses following the CTI.

To avoid the incorrect characterization of strongly eroding materials, the standard recommends that the maximum depth of erosion from the five evaluated results of the tracking path formation is determined and indicated along with the CTI. This also applies when no tracking path is generated, e.g., due to sample ignition or when, although tracking paths are formed at a test voltage of 600 V, a current of 0.5 A is not exceeded.

2.2. Volume Resistivity

Volume resistance is obtained by dividing the d.c. voltage between two electrodes placed on two opposite faces of a specimen by the steady-state current between these electrodes. The volume resistivity ϱ is defined as the volume resistance relative to a cube with unit volume [27].

When a d.c. voltage is placed across two electrodes which are in intimate contact with a specimen (Fig. 3), the current passing through the sample decreases asymptotically towards a

constant end value. The drop in current can be caused by dielectric polarization and by migration of highly mobile ions toward the electrodes. The insulation resistance is generally determined 1 min after the beginning of the electrification.

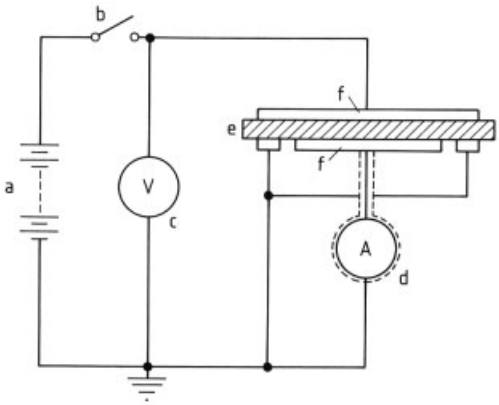

Figure 3. Circuit diagram for measuring volume resistivity
a) Voltage supply; b) Switch; c) Voltmeter; d) Shielded ammeter; e) Specimen; f) Electrode

The current flowing along the surface, and the current arising from the polarization phenomena on the electrodes during careful and exact measurements must be excluded from the measurement. This is achieved by using appropriate circuits and measurement procedures.

Electrodes used for these tests are metal-like conductors or metallic conducting materials of suitable shape, size, and design; they are placed in intimate contact with the specimen and form the transition regions between the electron current and the ion-mediated current flow in the insulator. With liquid insulators, the electrodes are provided by the completely wetted container in which the measurements are being carried out (fluid electrode). With solid insulators, the electrodes must be adhered to the material (Fig. 4). Even under pressure, placing metallic components of dimensions stipulated in the standards on or beneath the sample material is not adequate in order to determine publishable results. None of the insulating materials has a perfectly flat surface. Therefore, in extreme cases, disk-shaped electrodes only make defective contact at three elevations on the sample surface.

The international standard testing process is specificied in IEC Publication 93 [27] for solid materials and IEC Publication 247 [28] for liquid insulators. Recommendations for measuring the insulation resistance at elevated temperature are given in IEC Publication 345 [29].

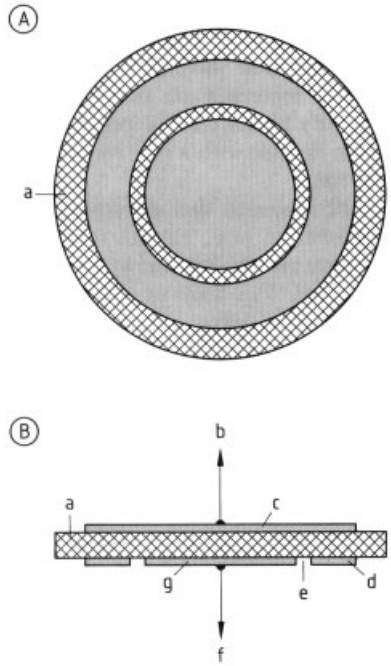

Figure 4. Top view (A) and side view (B) of volume resistivity measurement in a solid specimen using adhering electrodes
a) Sample; b) To voltage supply; c) Unprotected electrode; d) Guard electrode; e) Guard gap; f) Connection for ammeter; g) Protected electrode

2.3. Surface Resistivity

The electrical insulation behavior of an insulator is influenced considerably by the nature of its surface, as well as by adsorbed or contamination layers. Conduction across the surface produced by these effects is known collectively as surface conductance. To decrease this property, the length of the conduction pathway can be increased, for example by mounting ribs or shielding on the insulator. The removal of contamination layers is simplified by using a smooth surface, which is achieved in ceramic insulating components by glazing.

The contribution of adsorbed and contamination layers to surface conduction illustrates that this parameter is not a material property. However, the deposition and enrichment of adsorbed

layers is also influenced by the nature and structure of the surface of the material. For example, in inorganic insulating materials the formation of adsorbed water layers is related to the phase characteristics of the surface: the polished surfaces of ceramic insulator components often exhibit a greater surface conductance than the fired, glass-enriched surface. A significant increase in surface conductance can be observed with increasing air humidity.

Surface resistance is obtained by dividing the d.c. voltage between two electrodes in contact with the surface of a sample by the current flowing between these electrodes 1 min after electrification (see Section 2.2). The surface resistivity σ is the surface resistance expressed relative to a square area. The length of the sides of this square is, in theory, immaterial [27].

Generally, current mainly flows through the surface layer of the specimen and through the moisture and surface contamination associated with it. However, the measured surface resistance also includes a nondefinable contribution from the resistance of the bulk of the sample. Thus, surface resistance cannot be measured precisely but only approximately. The measured value is generally an indication of the state of the surface at the time of measurement. The relative permittivity (see Section 2.4) of the sample under investigation affects the form and extent of contamination [27]. The current which flows as a result of contamination is, however, also influenced by the surface structure and surface properties of the specimen. Hence the surface resistivity is not a material property in the normal sense, but it can be regarded as such if the effects of contamination are taken into consideration, e.g., by appropriate preconditioning.

The test procedure has been specified by the international standard IEC 93 [27].

2.4. Relative Permittivity

Polarization leads to the internal charging of the material in a direction opposing that of the existing field. This causes a reduction of the field strength within the capacitor and a corresponding increase in the capacity when the volume between the electrodes is completely and exclusively filled with an insulating material as opposed to being empty. The ratio between the capacity of a capacitor filled with a dielectric and that of a capacitor of identical dimensions but incorporating a vacuum between the electrodes is known as the relative permittivity ε_r of the dielectric. Under d.c. conditions, the relative permittivity is characteristic for most materials, but is temperature dependent. Under a.c. conditions, it moreover depends on frequency and in certain cases on field strength.

The relative permittivity is a dimensionless comparative value. The procedure for determining capacitance is given in the international standard IEC Publication 250 [30].

Insulating materials can be divided into three groups according to relative permittivity:

1) With most materials, the positive and negative charge centers coincide. Under the influence of an electric field, electron polarization, and in some cases ion polarization, occur. Such insulating materials possess low relative permittivities of 2 – 10.
2) In the second group, a dipole moment is already present because of the molecular or crystal structure. In addition to the electron or ion polarization, the external field introduces orientation polarization, assuming of course that the dipoles are sufficiently mobile. However, this is rarely the case with organic insulating materials, or even with liquid insulation oils of relatively high viscosity. With inorganic insulating materials, the dipoles in the crystal structure may reorientate, which leads to a considerable increase in capacity. Such insulating materials are characterized by a significantly higher relative permittivity (e.g., titanium dioxide with a relative permittivity of ca. 100).
3) In ferroelectric materials that undergo spontaneous polarization (e.g., barium titanate), complete regions are reorientated by an external electrical field. This leads to relative permittivity values of ≥ 1000.

The transportation of charge carriers in an electric field and the continuous reversal of polarization in the a.c. field both incur energy losses. They cause a certain degree of heating of the capacitor dielectric. The power requirement for this is taken from the a.c. electrical field.

2.5. Dielectric Dissipation Factor

The dielectric dissipation factor (dielectric loss) is the tangent of the loss angle δ by which the phase displacement between current and voltage in a capacitor deviates from $\pi/2$ when the dielectric of the capacitor consists solely of the insulating material under investigation (Fig. 5). At the same time, it is also the ratio of the loss or effective power N_e to the reactive power N_r of a capacitor completely filled with the insulating material and subjected to an a.c. voltage.

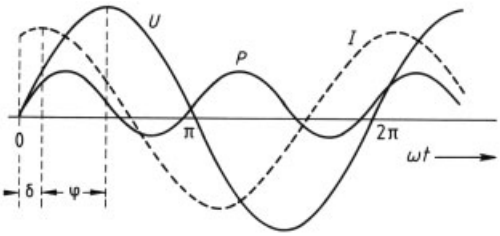

Figure 5. Sinusoidal alternating current: voltage (U), current strength (I) and power (P) as a function of time (t)
δ = loss angle; φ = phase angle; ω = frequency

The following equation describes the relationship between the dielectric dissipation factor ($\tan\delta$) and the power factor ($\cos\varphi$) of a capacitor, where $\varphi = 90° - \delta$ and the apparent power,

$$N_a = \sqrt{N_e^2 + N_r^2}$$

$$\cos\varphi = \frac{N_e}{N_a} = \frac{N_e}{\sqrt{N_e^2 + N_r^2}} = \frac{\tan\delta}{\sqrt{1 + \tan^2\delta}}$$

The dielectric dissipation factor alone only permits comparison between the specific dissipation of different materials when the relative permittivity is the same (see Section 2.6).

The procedures for measuring the dielectric dissipation factor are described in the international standard IEC Publication 250 [30].

The highest-quality dielectric materials possess dissipation factors ($\tan\delta$) of < 0.0002 within a temperature range of 20 to 100 °C and a frequency range of 50 Hz to 1 MHz. With some insulating materials the dielectric dissipation factor is greater at low frequencies than at high frequencies. The dissipation factor generally increases with increasing temperature. In high-power systems, the heat generated by the energy losses has detrimental effects.

With glasses and ceramic materials incorporating a glassy phase, the network-modifying alkali ions have a higher mobility than the doubly bonded alkaline-earth ions. In low-loss insulating materials, the fluxing agents (which are normally based on alkali compounds) are substituted by equivalent materials based on alkaline earths. Examples include "low-loss steatite" and "alkaline-earth porcelain". Improvements in dielectric properties are, however, associated with manufacturing problems; alkali compounds used as fluxing agents cannot usually be substituted.

2.6. Loss Index

The loss index $\varepsilon'_r{}'$ of an insulating material is defined as the product of the relative permittivity and the dissipation factor

$$\varepsilon''_r = \varepsilon_r \cdot \tan\delta$$

When comparing different materials, the dielectric loss is to be preferred to the dissipation factor because the dielectric losses of identical volumes under identical field strengths and at identical frequencies behave in the same way as the respective loss indices (see Section 3.2). The dielectric loss performance P_v, relative to a unit volume of, for example, 1 cm^3 can be calculated using the following equation:

$$P_v\ (\mathrm{per\,cm^{-3}}) = \omega \cdot \varepsilon_0 \cdot \varepsilon''_r \cdot E^2 \cdot 10^{-14}\ \left(\mathrm{W/cm^3}\right)$$

where ω = angular frequency = $2\pi f$ (frequency in Hz), ε_0 = electric constant (0.08854 pF/cm), $\varepsilon'_r{}' = \tan\delta \cdot \varepsilon_r$, E = field strength (V/cm).

2.7. Breakdown Voltage

The breakdown voltage is the voltage level at which the potential between the electrodes cannot be maintained and the specimen is destroyed. It is not a material-specific property and depends on numerous parameters:

sample thickness,
rate of voltage increase
type of test voltage (d.c. or a.c.),
duration of voltage exposure,
environmental temperature,
humidity,

mechanical state of the specimen,
homogeneity,
surrounding medium, and
electrode size and positioning.

Hence all of the specifications listed in the standards (e.g., international standard IEC Publication 243) must be met exactly if comparable results are to be obtained. Even if all secondary effects can be discounted from the results (e.g., parasite discharges at the edges of the electrodes), the short-term breakdown voltage does not in general possess the character of a material-specific property. It is more a criterion for the degree of defectiveness of the insulator because the weakest regions (e.g., inhomogeneities, voids, gas inclusions) are monitored during measurement.

Breakdown behavior of insulating materials varies according to material type and testing conditions; differences depend on various breakdown mechanisms (Fig. 6). In general, a purely *electrical breakdown* is observed during impulse tests (e.g., exposure to peak loadings), and the values of breakdown voltage are correspondingly high. During long-term field strength testing, *thermal breakdown* occurs as a result of local instability caused by the heat generated by increased conductivity losses and dielectric losses (see Section 2.6). Hence the values of breakdown voltage obtained are lower, possibly also because of the low thermal conductivity. The *discharge breakdown* is of particular importance, especially during the testing of compacted materials and liquids. It is caused by discharge and destructive processes located around inhomogeneities or gas inclusions.

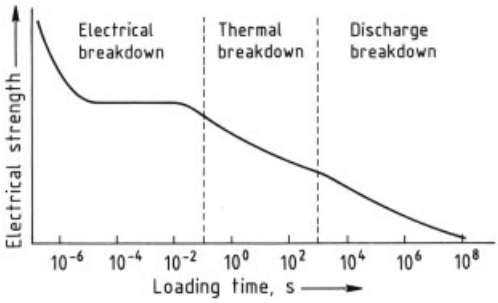

Figure 6. Electrical strength as a function of loading time

The electrical strength is defined as the ratio of the breakdown voltage to the distance between the two electrodes across which the voltage is applied. The term electrical strength is only valid when a uniform field exists between the electrodes.

Determination of the breakdown voltage is specified in the international standards IEC Publication 243 [31] for solids and IEC Publication 156 [32] for liquid insulating materials.

2.8. Further Electrical Tests

For special applications or investigations, supplementary test procedures have proved valuable; they have gained international recognition and have been standardized.

High-Voltage Tracking Test. This test procedure (Fig. 7) is utilized for materials that are used under severe ambient conditions in electrical engineering (e.g., high humidity, high voltages). The test enables assessment of the stability towards track formation and erosion. *Track formation* is defined as a process in which tracking and erosion are the result of an electrical discharge at or adjacent to the contaminated surface of an insulating material.

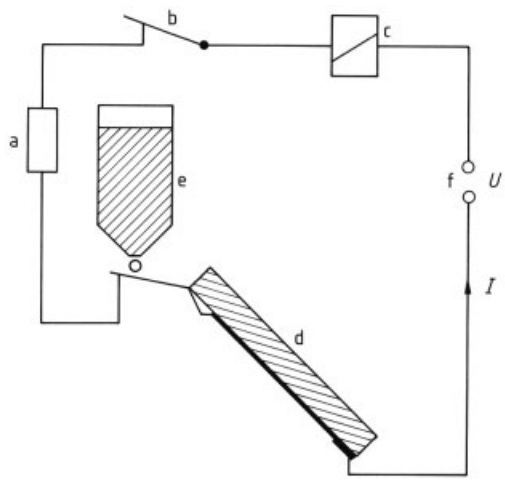

Figure 7. Schematic representation of measurement of high-voltage tracking
a) Current limiting resistor; b) Excess current switch activated by relay (c) when $I \geq 60$ mA for 2 s; c) Relay; d) Specimen (inclined plane); e) Contamination fluid; f) Voltage supply

The test must be carried out with plane specimens that are inclined at 45° and moistened with

a contamination liquid. The contamination liquid consists of distilled water containing ammonium chloride and a specified wetting agent. The specimen has the dimensions of 5×12 cm. High voltages of 2.5, 3.5, or 4.5 kV are applied to the two electrodes fastened on the smaller edges of the specimen. The contamination is allowed to flow at a specified rate over the lower surface of the specimen. The time (up to 6 h) is observed for failure caused by tracking or burning.

The tested materials are classified into four categories according to their resistance to tracking which depends on the test voltage and time. A correlation exists between this test and the low-voltage tracking test (see Section 2.1). The procedure is specified in the international standard IEC Publication 587 [33].

Arc Resistance. The test procedure according to ASTM-D 495 uses a high-voltage supply of 12.5 kV in order to generate arcing across the surface of the specimen [34]. Contrary to the high-voltage tracking procedure according to IEC Publication 587 (see above), the samples are tested in the absence of arbitrary contaminants. The electrodes are round tungsten rods. The angled tips of these rods are placed in contact with the upper or under surface of the specimen under investigation at a defined angle and at a distance of 6.35 mm apart (Fig. 8). The insulating material is classified according to the duration of the arc in seconds. At the end of each test minute, the power effect on the surface of the specimen is doubled.

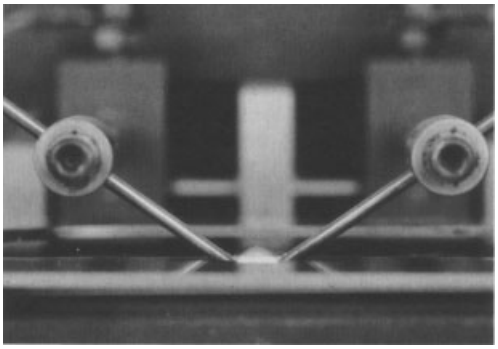

Figure 8. Arc resistance test (ASTM)
Arcing occurs between the tungsten electrodes across the sample surface.

Low Voltage – High Current Arc Test. This test is used to investigate the behavior of heat-resistant materials. These are generally inorganic substances or organic materials which are highly filled with inorganic substances. The test determines (1) the distance over which a carbon arc can be extended across the surface, (2) whether the arc path is conducting, and (3) whether the material underneath the surface changes significantly after cooling. The test procedure is specified in DIN VDE 0303 part 5 [35].

Assessment of Electrostatic Behavior. Electrostatic behavior can be assessed in a test method in which either disk-shaped samples (rigid materials) or endless strips or tapes (flexible materials), are charged using contacting substrates. The field strength generated by separation of the specimen and the substrate is measured as well as the change of the charge as a function of time. Since the tests are designed to allow the selection of nonchargeable or only slightly chargeable materials, they are only of advantage for materials with a resistance at the lower end of the insulation range ($10^9 - 10^{11}\ \Omega$). The test is related to the determination of surface resistance (see Section 2.3). Procedures are specified in DIN VDE 0303 parts 8 and 14 [36], [37].

Determination of Corrosion-Promoting Effects. The corrosion-promoting effects of insulating materials can be determined by three standardized procedures specified in IEC Publication 426 [38]. A visual procedure, a tensile strength method, and an insulation resistance method are available to select between materials which have corrosion-promoting effects on nonprecious metals. With the first two tests, effects are evaluated according to the damage following contact with copper or brass. For special applications, other metals can be employed. During the evaluation of the resistance change, a decrease in resistance between taper pin electrodes or bar electrodes (measurement procedure in IEC Publication 167 [39]) after storage under humid conditions is regarded as being a criterion for the presence of a corrosion-promoting effect.

Within the fields of telecommunications, process control, and data processing, insulating materials are often employed to insulate conductors placed under a d.c. voltage. Under unfavorable

conditions, the metallic components may corrode. This can be simulated by testing under the simultaneous affects of a d.c. supply, heat (e.g., 40 °C), and environmental humidity (e.g., 92 % R.H.). The test reveals which combinations of insulating material and metal are particularly susceptible to corrosion. Judgement is made by observing color changes, destruction of test metal strips, or decreases in insulation resistance. Testing procedures are specified in IEC Publication 426 [38].

Relative Resistance to Breakdown by Surface Discharges. This method is used to estimate the working life of materials by exposing them to voltages above the discharge voltage. As with the determination of the breakdown voltage (Section 2.7), the working life decreases rapidly with increasing field strength. The partial discharges at the surface damage the material in a way similar to discharges at defects within the material.

The resistance breakdown by discharges is tested. Electric strengths refer to one year and are only comparable if sample thicknesses are identical and all test parameters are kept constant. A minimum of nine samples are tested at a minimum of three voltage settings. Large deviations must be expected because defects are irregularly distributed and the test area is relatively small. Hence the median of each set of results is used for evaluation; extrapolation to one year is performed either graphically or using regression calculations.

The test procedure is described in IEC Publication 343 [40].

Detection of Transition Regions. The detection of regions of relaxation transition using thermally stimulated discharge currents is an extremely effective method for studying inelastic effects in polymers. The high resolution allows the complex behavior of the materials to be broken down into elementary processes (Fig. 9) [22].

The interpretation of the informative discharge curves requires further intensive investigations with physical or other methods. Thermal stimulation of discharge currents can be used to solve a variety of problems [22]; examples include effects of additives on electrical and mechanical properties; compatibility in copolymers; deformation; aging; irradiation; cryogenic processing; crystallinity; conducting mechanisms; activation energy; stereo complexes; thermal histories; cross-linking; moisture content; and relationships between mechanical relaxation, elongation, and calorimetry.

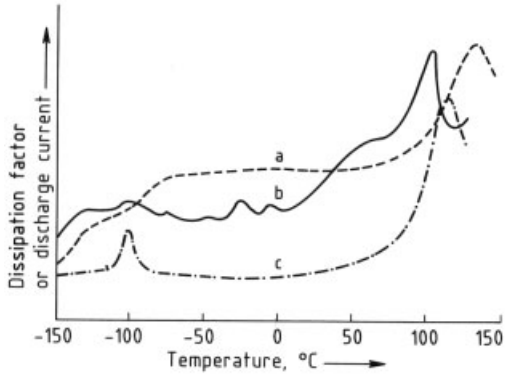

Figure 9. Relaxation regions of an acrylonitrile – butadiene – styrene (ABS) product
a) Dielectric damping (1 kHz); b) Thermally stimulated discharging (0.01 Hz); c) Mechanical damping (1 Hz)

2.9. Mechanical and Thermal Properties

Strength. The applicability of an insulator is primarily determined by its electrical properties. In many applications, however, mechanical properties are of equal importance.

The assessment of the mechanical strength of insulating materials is carried out using special specimens. However, the strength characteristics determined in this way cannot simply be applied to an insulation component because the strength decreases with increasing sample volume. This particularly applies to brittle, nonmetallic, inorganic materials. During the design of components and equipment, this must be taken into account and safety corrections must be made. These are known as a result of experience but they can also be approximately calculated by using techniques such as fracture mechanics and Weibull statistics.

Due to experimental difficulties however, acceptably reproducible results are only obtained in bending strength tests. Therefore standards have only been drawn up for this test. However, tensile tests are also routinely carried out when testing insulators.

The strength of ceramic insulators can be increased considerably by application of suitable glazes. The thermal expansion coefficient of the glaze must be slightly lower than that of the ceramic substrate. Following cooling after firing, the glaze is subjected to compressive stresses. When an external tensile force is applied to a glazed article, these compressive stresses in the surface must be counteracted before the tensile stresses can become effective. In glasses, similar effects are achieved by thermal hardening, i.e., by quenching from temperatures above the glass transition temperature. Chemical hardening (via ion exchange at elevated temperature) is used less widely.

With organic polymers, similar increases in strength can be gained by employing composite materials: glass or carbon fibers are incorporated into the polymer.

Material Space Filling. Apart from completely solid materials, porous substances are also used in electrical engineering. Porosity lowers the breakdown strength, therefore porous materials are not considered for high-voltage applications. In contrast to this, an open porous structure is favorable for the thermal shock resistance because porous materials have a lower Young's modulus. Porous materials are therefore often employed in thermal electrical engineering. An increase in the surface conductivity caused by the larger surface does not pose a threat because the water film is evaporated by the current flow.

Under high-vacuum conditions, porous materials are also utilized as electrical insulators to prevent the formation of continuous vapor layers.

Thermal Expansion. The thermal expansion of inorganic insulators is generally relatively low which is favorable for thermal shock resistance. A low thermal expansion also ensures good shape retention following temperature change. Inorganic insulators can withstand rapid temperature increases better than rapid cooling.

Organic substances are prone to material changes as a result of prolonged exposure to high temperature, which affects both the mechanical and electrical properties. Special organic compounds have recently been developed which are also suitable for application at elevated temperatures.

Thermal Conductivity. The thermal conductivity of the insulating materials is generally low. Hence, they can be simultaneously employed as electrical and thermal insulators. However, oxidic and nonoxidic materials are available which have a similar thermal conductivity to metallic substrates. Such insulators are normally employed when heat transfer or removal is necessary.

Specific Heat. The specific heat of the insulating material may be of significance in heat storage applications. Magnesium oxide with a specific heat of greater than $1000\,\mathrm{J\,kg^{-1}\,K^{-1}}$ is particularly suitable.

Combustion Behavior. The possibility of combustion can be a considerable obstacle regarding the use of organic insulators. Materials that do not burn at all when placed in contact with an electrical source of heat are given priority. Classification criteria include the extinguishing of the flame following removal of the heat source, or the speed at which the flames spread.

3. Insulating Materials

3.1. Inorganic Insulating Materials

Inorganic electrical insulating materials include several natural minerals (generally mica or asbestos) but also industrially manufactured products. The wide range of industrially manufactured materials includes pure glass, materials containing glassy and crystalline structural components, purely crystalline materials, and even single crystals. This broad spectrum of materials ensures a wide variety of properties. Depending upon the intended application, one or more of the following properties may be particularly pronounced: low or high relative permittivity; extremely low dielectric loss; high electrical strength at low and high frequencies; excellent insulation characteristics even at temperatures well above 1000 °C; low or high thermal conductivity; high mechanical strength.

The ferrites are also inorganic insulating materials. Ferrites are sintered products consisting

of iron(III) oxide and oxides of mono- or divalent metals. Their application within electrical engineering is primarily governed by their magnetic properties. Ferrites are generally regarded as magnetic substances (→ Magnetic Materials), although their insulation characteristics are of interest for certain applications.

3.1.1. Insulating Materials Based on Natural Minerals

3.1.1.1. Mica and Related Products
(→ Mica)

Micas are sheet silicates. Their properties are determined by mixtures of homo- or heteropolar bonds within the lamellar planes, which in turn are held together by van der Waals' forces. This type of bonding is responsible for several of mica's properties, e.g., the excellent cleaving properties parallel to the plane of the lamellae. Muscovite [*1318-94-1*], $KAl_2[AlSi_3O_{10}(OH)_2]$, and phlogopite [*12251-58-0*], $KMg_3[AlSi_3O_{10}(OH)_2]$, havefound widespread application in the electronics industry. Materials selected for insulation purposes must not contain contaminants which may adversely effect insulation properties. The mica products employed in electronics mainly originate from India.

Production and Processing. The mica is selected by hand from the mined rock and sorted according to size. Small fragments are ground and processed into powder. Larger specimens are cleaned and cleaved into sheets of thickness as low as 0.01 mm. Commercially available products are classified according to purity and surface area.

Cleaned, cleaved mica is generally processed by punching. The requirements of the punched component regarding tolerances and precise finishing are extremely high. During manufacture of capacitors, the metallizing preparation (e.g., a silver suspension) is applied directly onto the mica substrate and fired onto the surface at 600 °C following drying.

Mica products can be processed to vacuum-sealed components. Lead borate glasses are particularly suitable solders because of their low melting temperatures. Such composites can be used up to 300 °C.

Composite Insulating Materials. Large amounts of mica powder and scrap are processed to nonwoven substrates or sheets. In one process, scrap mica is heated to 800 °C; a portion of the water of crystallization escapes and causes partial cleavage of the lamellae. The mica is then immersed in sodium carbonate solution whilst still hot. Carbon dioxide gas is evolved during subsequent treatment with dilute acid, and causes complete separation of the mica lamellae. The very fine mica platelets can be processed into thin, endless, nonwoven substrates in a procedure analogous to that used in the manufacture of paper.

In another process, the mica is cleaved using a high-pressure water jet and again processed to a nonwoven fabric. The product is slightly less densely packed than that resulting from the procedure mentioned previously.

Such nonwoven fabrics, and small rejected, cleaved mica fragments are combined with natural or synthetic binders to give sheets, films, and tapes (micanite, VDE 0332, IEC 371). Suitable binders include asphalt, shellac, alkyd resins, or more recently, silicone and epoxy resins.

Micanite is a binder-containing sheet product that is manufactured using heated hydraulic presses. Commutator micanite should not contain more than 4 % binder. Sheets are ground to meet the required thickness within the tolerance limits, and then punched into commutator segments.

Heating micanite is designed for use as insulation in electric heaters; here, the binder immediately ignites when the device is first utilized. The remaining mica layer is responsible for insulation. With an appropriate binder, only short-term inconvenience is caused by the combustion products. The binder should not produce an electrically conducting residue after combustion. Shellac is a very satisfactory binder in amounts up to 3 %. Heating micanite is also manufactured from fine mica and contains 10 to 15 % silicone resin as an additive. The organic components ignite to form gaseous combustion products; the residual silica functions as a cement and confers the required strength to the insulator.

Rigid micanite, which cannot be shaped, is obtained on addition of increased amounts of binder. If subsequent shaping is desired, thermoplastic binder materials must be employed.

Mica foil consists of a thin substrate covered with one or two layers of cleaved mica or fine nonwoven mica material that are bonded together with a binder. Depending upon the type of substrate employed, the products are known as paper micanite, fabric micanite, nonwoven micanite, or film micanite (VDE 0332, IEC 371). *"Glimmerband"* (mica tape) denotes mica foil which has been slit into tapes.

Glass – mica composite materials can also be manufactured with low-melting lead borate glasses. With appropriate selection of the glass component, the semifinished article can be machined.

Properties. The properties of mica products are strongly influenced by the nature of the binder. Fine mica articles are more flexible than their cleaved mica counterparts. The application range is essentially governed by the type of binder used, but generally lies below 120 °C. With silicone resin binders, prolonged exposure to temperatures as high as 180 °C is possible. Heating micanite bound with silicone resins is able to withstand temperatures of 500 °C.

Uses. Mica wafers in their naturally occurring form are generally employed to fix electrode systems in valves and guarantee a high degree of stability, independent of the operating temperature. Further areas of application include discharge windows for high frequency energy, e.g., in transmitters operating in the microwave range, or as admission windows in counter tubes. Their reasonably high relative permittivity means that they can be used as insulators in capacitors in constant-frequency devices. The good thermal stability is utilized for insulating heating systems.

Micanites and mica foils are suitable for a variety of insulation applications, particularly in the construction of electrical equipment and plants.

3.1.1.2. Asbestos and Related Products
(→ Asbestos)

Asbestos [*1332-21-4*] is a fibrous form of magnesiumsilicate. It has an extremely high thermal stability as well as good electricaland thermal insulation capabilities even at elevated temperatures.

Asbestos materials cause severe lung damage (asbestosis), which has led to a ban on their use. They have been replaced by fiber materials produced from ceramics, particularly alumina ceramic fibers. Their properties are equal to or even better than those of asbestos. It has not yet, however, been clarified whether these alternative fibers can also cause cancer. No regulations have been laid down so far. The IEC standards and draft standards for the standardization of asbestos products have not yet been withdrawn.

3.1.2. Synthetic Inorganic Insulating Materials

Synthetic inorganic materials which are commonly used as electrical insulators are described in IEC Publication 672 [41], which has also been recently adopted as the German Standard DIN VDE 0335. Part 1 specifies terms and classification, Part 2 describes basic testing procedures, and Part 3 contains characteristic data for the individual material types.

Classification. The IEC 672 standard concerns insulating materials made from ceramics, glass, and glass ceramics, although the latter have not yet been standardized as a material type. The ceramic insulators are classified into eight groups:

C-100	materials based on alkali – aluminum silicate (porcelains)
C-200	materials based on magnesium silicate (steatites and forsterites)
C-300	materials based on titanium dioxide, titanates, stannates, or niobates (ceramics of high relative permittivity)
C-400	materials based on alkaline earth – aluminum silicates
C-500	porous materials based on aluminum silicate and magnesium silicate
C-600	materials based on aluminum silicate (mullite ceramics)
C-700	materials of high alumina content
C-800	special oxide ceramics

The glass insulators are classified into six groups (see next page).

G-100	alkali – lime – silica glasses
G-200	chemically resistant borosilicate glasses
G-300	electrically resistant borosilicate glasses
G-400	alumina – lime – silica glasses
G-500	lead oxide – alkali – silica glasses
G-600	barium oxide – alkali – silica glasses

According to IEC 672, *ceramic materials* are those substances which are shaped before firing; their main components are usually polycrystalline silicates, aluminosilicates, titanates, or oxides. *Glasses* are dense inorganic materials manufactured from a melt or by sintering, which solidify without crystallization. *Glass ceramics* are manufactured from glass, they contain a considerable crystalline content as a result of a special treatment. *Glazes* are coatings with a smooth surface which are generated in a melt process and which essentially consist of glass. They can contain coloring components or opacifiers.

The above classification of ceramic insulators is not systematic. Apparent inconsistencies are due to historical reasons. Because the preceding standard, DIN 40 685, had proven valuable for several decades and was internationally accepted, its classification system was adopted in the new IEC 672 standard. Some groups are partially defined on the basis of material aspects. However, structure is also decisive, in particular the presence of an open pore system. For example group C-500 contains only porous materials, whereas dense substances of the same chemical composition are classified in another group. Porous materials are not limited to group C-500 however. Certain groups (e.g., C-100 and C-200) contain both porous and dense varieties of a given composition.

Testing. IEC 672, Part 2 describes the testing procedures used for determining the mechanical, thermal, and electrical properties of the individual materials. Precise specification of test procedures, including sample preparation, is necessary because data are strongly influenced by the manufacturing process, shape, and dimensions of the sample, and also by the measuring procedure itself. Accordingly, data are strictly valid only for the stipulated sample and cannot be applied to samples that are of another shape or are manufactured by another process. However, modern fracture mechanics and Weibull statistics allow the properties of actual construction components to be estimated from the characteristic data obtained with standardized samples.

With glass, the thermal history of the material is important. Prestressed glass is influenced by the conditions experienced during thermal hardening. Here, it is even more difficult to deduce the properties of an insulation component from those of a standard specimen.

3.1.2.1. Porcelain

Porcelain was introduced as an electrical insulating material by W. von Siemens in the middle of the nineteenth century. Various types of porcelain have since become the material of choice for high-voltage insulators in overhead lines and transformer stations. Normal porcelain is less suitable for high-frequency and electrical heating applications because the desired properties cannot be obtained due to the presence of alkali from the fluxing agent. For special applications, alkaline-earth porcelains are available in which the alkali elements are eliminated. Only small components can be manufactured from these materials, however, because their firing interval (i.e., the temperature range between vitrification and distortion) is very small.

Material Types. The C-100 group according to IEC 672 Part 3, comprises five material types which differ with regard to their phase composition and required properties, particularly the mechanical data:

C-110	quartz porcelain
C-111	compressed porcelain (with a permitted residual porosity)
C-112	cristobalite porcelain
C-120	alumina porcelain, strong
C-130	alumina porcelain, ultrastrong

The compressed porcelains of type C-111 are only suitable for special low-voltage applications due to their permitted open porosity. Cristobalite porcelains exhibit better mechanical properties than the quartz porcelains. Alumina porcelains are now the preferred porcelains for high-voltage insulators. Their classification in the two types C-120 and C-130 takes into consideration recent developments with high-strength materials being classified as type C-130. Their manufacture is, however, difficult which explains why C-130 type materials are

only employed when extreme mechanical demands have to be met by the insulator. For normal requirements type C-120 materials, which present fewer manufacturing problems, suffice.

The use of the various porcelain types is based on differing philosophies regarding selection criteria. In western industrial nations, quartz porcelain is only used when the insulator is not exposed to mechanical loading. In general, alumina porcelains are applied even when they more than satisfy the requirements of the problem in hand. However, the universal utilization of alumina porcelains eliminates a range of problems that could arise if insulators were made in a single factory from different materials. Moreover, this improves the safety of overhead lines and transformer stations in which failure of an insulator could have a serious effect on energy supply. In contrast, in other countries (e.g., the Comecon countries), the minimization of the import of expensive synthetic alumina is the predominant factor. Here, quartz porcelain is used whenever acceptable.

Properties. Important properties of the porcelain types C-110, C-120, and C-130 are summarized in Table 1.

Raw Materials and Body Composition. The traditional hard porcelain is made from china clay, quartz, and feldspar. The plasticity of the china clays is more important than its ability to produce whiteness after firing. China clay can thus be partially substituted by plastic clay in order to improve the workability of the porcelain body.

Quartz porcelains have a high quartz content of up to 40 % which accounts for their high strength. During firing, certain types of quartz are readily converted to cristobalite, which gives rise to the formation of cristobalite porcelain types [42]. The substitution of quartz by alumina produces the alumina porcelains; type C-120 contains up to 30 wt% and type C-130 up to 40 wt% alumina.

Manufacture (→ Ceramics, General Survey). Molding and shaping are carried out using conventional processes employed in the ceramic industry. Low-voltage insulation components are manufactured by pressing techniques. High-voltage insulators are produced by extrusion of a pug with a vacuum extruder and subsequent cutting either after partial drying in a "leather-hard" state (Fig. 10), or after complete drying. The pug itself can be manufactured by isostatic pressing to give a dry pug, suitable for immediate machining. Care must be taken to ensure that any dust generated during machining is removed completely.

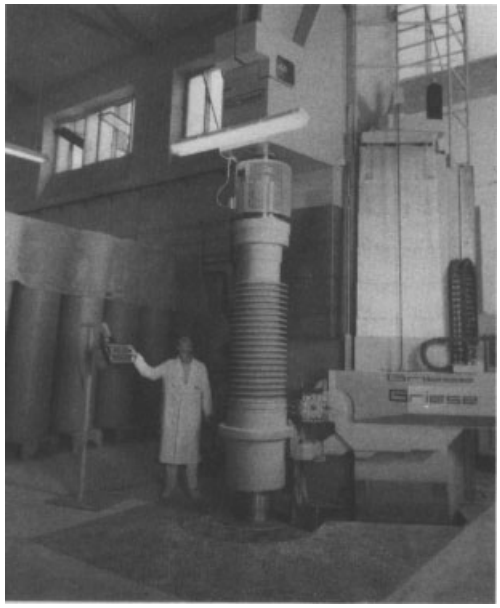

Figure 10. Machining of an insulator with photoelectric control

The shaped components are coated with a raw glaze, dried, and fired in a tunnel, chamber, or oven kiln.

Semiconducting glazes coated onto insulators may influence the shape of the electric field and contribute to the prevention of short circuiting due to condensation. However, they can cause problems during operation, and have therefore not yet made a general breakthrough, although they would permit the more compact construction of transmission lines.

Uses. Low-voltage insulation components made from compressed porcelain are employed in a variety of applications including coupler terminal strips, cartridge fuses, terminal blocks, and inserts for plugs and sockets. They may be used in combination with organic polymers.

A wide range of porcelain insulators are employed in high-voltage applications: support in-

Table 1. Important physical properties of porcelain types C-110, C-120, and C-130 classified according to IEC 672

Property	C-110	C-120	C-130
Open porosity, vol%	0	0	0
Bulk density, g/cm^3	>2.2	>2.3	>2.5
Bending strength, N/mm^2			
Unglazed	>50	>90	>140
Glazed	>60	>110	>160
Linear expansion coefficient (20–600°C), 10^{-6} K^{-1}	4–7	4–7	5–7
Electrical strength, kV/mm	>20	>20	>20
Withstand voltage (1 min), kV	>30	>30	>30
Relative permittivity at 20 °C and 50 Hz	6–7	6–7	6–7.5
Dielectric dissipation factor (tan δ) at 20 °C and 50 Hz	<0.025	<0.025	<0.030
Volume resistivity, $\Omega \cdot$ cm			
at 20 °C	$>10^{13}$	$>10^{13}$	$>10^{13}$
at 600 °C	$>10^{4}$	$>10^{4}$	$>10^{4}$

sulators, suspension insulators, and terminal insulators for overhead line construction; insulator housings able to withstand extreme mechanical loading; or as external shielding for voltage transformers of heights up to 5 m and diameters of 1 m. Lead-through insulators can also be produced with multiple bore holes running parallel to the longitudinal axis.

The alkaline earth porcelains are mainly employed as resistor bodies.

3.1.2.2. Steatite

The term steatite [*14807-96-6*] has two meanings, firstof all as a synonym for the fine-grained variety of talc, and secondly fora special ceramic material manufactured from this talc.

To avoid confusion, the term steatite is used here to denote the ceramic product. The term talc (and not steatite) is used for the ceramic raw material regardless of its crystal size. Normally, however, the term talc solely denotes the coarse-grained variety of the mineral.

Material Types. Group C-200 according to IEC 672 contains six types of material based on talc:

Type C-210	steatite for low-voltage applications (not dense)
Type C-220	steatite
Type C-221	low-loss steatite
Type C-230	steatite, porous
Type C-240	forsterite, porous
Type C-250	forsterite, dense

The most important materials in group C-200 are steatite (type C-220) and low-loss steatite (type C-221). Type C-250 is formed from a MgO-enriched body and contains forsterite [*15118-03-3*] (Mg_2SiO_4) as the main crystalline phase. The porous types (C-230 and C-240)allow the manufacture of high precision components from fired semifinished products using conventional machining techniques.

Properties. Several of the most important properties of the C-220 and C-221 types are listed in Table 2.

Type C-220 steatite is manufactured using feldspar as a fluxing agent. This yields an adequate firing range for the production of larger items. With low-loss steatite (type C-221), dielectric properties are considerably improved when barium carbonate is used as a fluxing agent. This material is thus also suitable as a dielectric for capacitors. For some applications its relatively high mechanical strength is also of significance. With the exception of high-voltage insulators, low-loss steatite is the most popular variety of steatite.

Compared with the porcelains, the steatites have outstanding electrical properties. Only the best alumina porcelains have mechanical properties comparable with those of steatite. Recent investigations have shown that the resistance of steatite to crack propagation is slightly greater than that of alumina porcelain [43].

Raw Materials. The preferred raw material for the manufacture of steatite is talc [*14807-96-6*], $Mg_3[Si_4O_{10}(OH)_2]$ in its fine crystalline form (→ Talc). The random orientation of the crystals in this talc allows the manufacture of ceramic products with low particle

Table 2. Important physical properties of steatite types C-220 and C-221, as well as forsterite type C-250 classified according to IEC 670

Property	C-220	C-221	C-250
Open porosity, vol%	0	0	0
Bulk density, g/cm^3	>2.6	>2.7	>2.8
Bending strength, unglazed, N/mm^2	>120	>140	>140
Linear expansion coefficient (20 – 600 °C), 10^{-6} K^{-1}	7 – 9	7 – 9	9 – 11
Electrical strength, kV/mm	>15	>20	>20
Withstand voltage (1 min), kV	>20	>30	>30
Dielectric dissipation factor (tan δ) at 20 °C			
50 Hz	<0.005	<0.0015	<0.0015
1 MHz	<0.003	<0.0012	<0.0005
Volume resistivity, $\Omega \cdot$ cm			
at 20 °C	$>10^{13}$	$>10^{13}$	$>10^{13}$
at 600 °C	$>10^{5}$	$>10^{7}$	$>10^{7}$
Relative permittivity at 20 °C and 50 Hz	6	6	7

orientation. One of the most important talc deposits is found near Göpfersgrün – Thiersheim in the Federal Republic of Germany. It represents the main basis for the worldwide manufacture of large, high-voltage steatite insulators. Talc for use in ceramics originates in Australia, the United States, Spain, and China.

Only talcs with low levels of contamination can be used as a raw material for the manufacture of steatite. The maximum upper limits which can be tolerated are 1.5 wt% Al_2O_3 and 2.0 wt% Fe_2O_3 [44]. Recently, calcia has been regarded as less critical and may be present in amounts up to 1 wt% [45]. The steatite body contains 80 to 90 % talc along with 5 to 10 wt% of a fluxing agent (in steatite, feldspar; in low-loss steatite, barium carbonate) and plastic clay to improve the processing properties.

Manufacture. Components from steatite are produced according to the general principles employed in the ceramics industry (→ Ceramics, General Survey). All molding and shaping techniques can be employed such as casting, shaping in the plastic state, or pressing, and when necessary subsequent machining in the leather-hard or dried state. The most important technique, particularly for small components, is pressing. The dry pressing process only gained widespread use after its development with talc bodies. Talc bodies have proved to be particularly suitable for this technique because the low hardness of talc (Mohs hardness = 1) guarantees a long service life.

The steatite components are usually fired at 1300 – 1400 °C in an oxidizing or reducing atmosphere. The firing interval is lower than that of porcelain, particularly with low-loss steatite. The lower firing interval for steatites is mainly responsible for the fact that steatite insulators can only be manufactured with maximum heights of ca. 1.2 m.

During firing, dry-pressed components shrink by only 10 to 12 %, and hence are suitable as insulator components with high dimensional tolerances.

Metal layers of silver, platinum, iron, or molybdenum can be used as electrodes on steatite or as a soldering base for joining steatite to metals of other ceramic components.

During firing, the talc is transformed to protoenstatite, $MgSiO_3$. The additives (clay and feldspar or barium carbonate) form a melt phase which vitrifies on cooling; the glass content of the fired body may be as high as 25 to 50 vol% [46]. An excess of silica leads to crystallization of cristobalite from the melt; this occurs in certain steatite bodies, but particularly in feldspar steatites. Improved thermal shock resistance is observed with steatites in which the silica has been completely taken up by the melt phase, for example in most low-loss steatite products containing silica-free fluxing agents.

Uses. Steatite finds widespread use as an insulating material in electrical engineering. In low-voltage applications, it is employed for the manufacture of insulation components in which a high degree of dimensional precision is required (e.g., for mechanically mounted components). Typical insulators made from steatite include sockets, inserts for plugs, levers and rollers

for rotary and rocker switches, holders, and cartridge fuses. Steatite has many advantages over porcelain when greater mechanical loading has to be withstood.

In high-voltage applications, steatite can now be replaced by alumina porcelain. Abrasion of machining tools is, however, considerably lower with steatite bodies. On the other hand, alumina porcelain bodies are easier to fire and are suitable for the manufacture of components much larger than those made of steatite. Steatite shows clear advantages in high-frequency applications, for example as bases for antenna masts and for terminal insulators in transmission masts.

Low-loss steatites are suitable for the manufacture of capacitors, e.g., the complex thick-walled capacitors used in high-voltage applications.

Talc Bodies Enriched in Magnesia. The crystalline phase *forsterite,* Mg_2SiO_4, is deliberately formed in fired bodies of the material types C-240 and C-250 by enriching the talc with magnesia (e.g., by addition of magnesite, $MgCO_3$). Several properties of type C-250 materials are summarized in Table 2.

The outstanding properties of forsterite ceramics include the particularly low dielectric losses and the large thermal expansion coefficient. The latter permits low-stress soldering to iron or nickel substrates which possess coefficients of thermal expansion of similar magnitude. This advantage of forsterite is, however, obtained at the expense of low thermal shock resistance.

Zircon-Containing Steatite. Zircon [*1490-68-2*] , $ZrSiO_4$, can be sinteredto produce a dense material only with difficulty. When used as an additivefor steatite, and, in some cases, for porcelain, the thermal expansion of the resulting material is reduced, and hence the thermal shock resistance is improved. The amount of zircon additive can reach 60 wt%. The dielectric properties of the zircon steatites are adversely affected because losses increase significantly above 1 MHz. The relative permittivity of zircon steatite is 8, this is sometimes undesirably high for certain insulators.

3.1.2.3. Materials for Electrical Heat Engineering

Two of the most important demands for the use of ceramic materials in electrical heat engineering are adequate stability at high temperature and favorable thermal shock resistance.

The thermal shock resistance is determined by the phase composition: compositions which approach that of a eutectic melt at relatively low temperatures and are therefore unsuitable. Thermal shock resistance increases with increasing mechanical strength and thermal conductivity, but also with decreasing modulus of elasticity and decreasing thermal expansion. It also depends upon external conditions, such as shape and dimensions of the article, as well as on the rate of heat exchange with the surroundings. In this respect, the thermal shock resistance is not a material constant.

Young's modulus is related to the structure of the material; it is lower with porous structures than with dense structures.

The most important factor controlling thermal shock resistance is thermal expansion. In materials with a low thermal expansion coefficient, the product is only exposed to low stresses, even if the thermal shock is high.

Hence the most suitable materials for electrical heat engineering are those which are porous and possess a low thermal expansion coefficient.

Material Types. Group C-400 defined in IEC 672 contains two types of material:

C-410	cordierite dense
C-420	celsian dense ($BaAl_2Si_2O_8$)

Type C-410 materials are used more widely.

Group C-500 contains the following porous materials:

C-510	materials based on grog (fireclay, chamotte)
C-511	materials based on grog containing magnesium oxide
C-512	materials based on grog containing magnesium oxide
C-520	materials of high cordierite content
C-530	materials of high alumina content

Types C-511 and C-512 contain cordierite [*12182-53-5*], $Mg_2Al_4Si_5O_{18}$, and mullite [*55964-99-3*], $Al_6Si_2O_{13}$. They differ in the number andsize of the pores, and hence also in the strength and thermal shock resistance. The composition of type C-520 materials approaches that of cordierite. The high alumina content of the type C-530 materials enables high application temperatures up to 1300 °C.

Properties. Table 3 contains typical properties of porous and dense materials used in electrical heat engineering.

Materials Containing Cordierite. Even as early as the 1930s, addition of talc to clay-based material had been shown to reduce their thermal expansion coefficient, and hence increase their thermal shock resistance. Talc and clay react during firing to give cordierite. Optimum results are obtained with approximately 60 % clay and grog (chamotte) and 30 % talc (type C-520 material according to IEC 672). Types C-511 and C-512 contain ca. 10 % talc.

A distinct disadvantage of cordierite materials is their narrow firing range. Consequently it is normal to refrain from vitrifying the articles, and concentrate on porous products that have a lower Young's modulus and therefore exhibit improved thermal shock resistance. If dense materials of type C-410 are required for special applications, then small amounts of a fluxing agent must be added; feldspar has proved particularly effective in this respect. Such materials are not only inferior as regards their thermal shock resistance, but also in their electrical insulation characteristics. This is because of the contribution of the alkali ions to their electrical conductivity at high temperatures. Furthermore, the porous materials allow higher application temperatures (ca. 1200 °C) than their dense counterparts. However, dense cordierite materials are required for some applications.

Lithium Aluminum Silicates. Lithium aluminum silicates possess an even lower thermal expansion coefficient than cordierite.

The materials have not been standardized. Raw materials are petalite, $LiAlSi_4O_{10}$, and to a lesser extent spodumen, $LiAlSi_2O_6$, or eucryptite, $LiAlSiO_4$. Widespread application of these materials has been somewhat hindered because of the scarcity of the raw materials, and because of the extremely narrow firing range. Therefore their excellent thermal shock resistance is only used in exceptional cases.

3.1.2.4. Alumina Ceramics (→ Ceramics, Electronic)

The inorganic insulating materials discussed in Section 3.1.1 are all produced from natural materials. Aluminas and the other materials discussed in this section are based on synthetic materials produced by the chemical industry.

Material Types. In IEC 672, alumina materials are classified in group C-700. They differ in their minimum alumina contents. In all cases, the alumina content is so high that the insulating material contains α-aluminum oxide as the predominant or even exclusive crystal phase.

Properties. The important physical properties of alumina ceramics are summarized in Table 4. Properties generally improve with increasing alumina content (higher number).

Alumina ceramics combine a range of remarkable properties. With the high percentage types (C-799), a strength of $\geq 500\,N/mm^2$ can be achieved, particularly if crystal growth can be suppressed during sintering. Mechanical strength is maintained up to temperatures of 1000 °C. Thermal conductivity reaches a level equivalent to that of metallic lead, especially in the high percent types. Alumina substrates are amongst the best electrical insulating materials under all application conditions.

Production. Raw materials include calcined aluminas which are obtained from natural bauxites by the Bayer process (→ Aluminum Oxide). The precipitation and calcination conditions govern the most important properties in ceramic applications, such as crystal size and sintering capability. Since the aluminas are not plastic, molding and shaping necessitates the addition of organic plasticizers which must be removed prior to firing.

Table 3. Important properties of materials classified in groups C-400 and C-500 of IEC 672

Property	Fine fire clay	Materials containing cordierite			Materials containing alumina	
	C-510	C-410	C-511	C-512	C-520	C-530
Open porosity, vol%	30	0.5	20	40	20	30
Bulk density, g/cm^3	>1.9	>2.1	>1.9	>1.8	>1.9	>2.1
Bending strength, N/mm^2	>25	>60	>25	>15	>30	>30
Linear expansion coefficient (20–600 °C), 10^{-6} K^{-1}	3–6	2–4	4–6	3–6	2–4	4–6
Thermal shock resistance, K	>150	>250	>200	>250	>300	>350
Electrical strength, kV/mm		>10				
Dielectric dissipation factor (tan δ) at 50 Hz		<0.025				
Volume resistivity, $\Omega \cdot$ cm						
at 200 °C	$>10^9$	$>10^8$	$>10^9$	$>10^9$	$>10^9$	$>10^{10}$
at 600 °C	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^6$

Table 4. Important properties of alumina materials classified in group C-700 according to IEC 672

Property	C-780	C-786	C-795	C-799
Open porosity, vol%	0	0	0	0
Bulk density, g/cm^3	>3.2	>3.4	>3.5	>3.7
Bending strength, N/mm^2	>200	>250	>280	>300
Coefficient of linear expansion (20–600 °C), 10^{-6} K^{-1}	6–8	6–8	6–8	7–8
Thermal conductivity, W m^{-1} k^{-1}	10–16	14–24	16–28	19–30
Electrical strength, kV/mm	>10	>15	>15	>17
Withstand voltage (1 min), kV	>15	>18	>18	>20
Dielectric dissipation factor at 20 °C				
50 Hz	<0.001	<0.0005	<0.0005	<0.0002
1 MHz	<0.0015	<0.001	<0.001	<0.001
Volume resistivity, $\Omega \cdot$ cm				
at 20 °C	$>10^{14}$	$>10^{14}$	$>10^{14}$	$>10^{14}$
at 600 °C	$>10^7$	$>10^7$	$>10^8$	$>10^8$

For solder connections, alumina components are metallized, preferably by the moly–manganese paste process (→ Ceramics, Ceramic–Metal Systems, Chap. 4.1.). Alternatives are active soldering with titanium hydride or use of solders containing titanium under conditions of high vacuum or in an oxygen-free atmosphere.

Uses. Alumina ceramics find many applications within the field of electronics and electrical engineering due to their outstanding insulation properties even at very high frequencies. Furthermore, their large thermal conductivity means that any heat generated as a result of the dielectric losses in electrical circuits and components can easily be removed. Examples of uses include housings for high-frequency electron tubes, rectifier housings, high-frequency coupling windows, solderable bypasses, resistor substrates, substrates for potentiometers, and substrates for microelectronics.

3.1.2.5. Special Oxide Ceramics

Group C-800 of IEC 672 comprises several special oxide ceramic insulating materials: beryllia (type C-810), magnesia (type C-820), and zirconia (type C-830).

Beryllia Ceramics. Beryllia ceramics (type 810 according to IEC 672) possess similar properties to alumina ceramics; the main differences are the lower density (<3.0 g/cm^3) and the large thermal conductivity which is approximately one order of magnitude higher than that of alumina. In substrates, potentiometer mounting plates, and housings for semiconductor components, the material effectively removes heat produced by dielectric losses. However, beryllia dust is highly toxic and attacks the lungs, necessitating comprehensive safety precautions during manufacturing (beryllia in finished products is of no danger provided that no dust is generated). This, along with the limited availability of

raw material deposits, means that beryllia components are extremely expensive. Their use is restricted to a small number of applications that involve particularly high demands, e.g., in the aerospace industry.

Magnesia Ceramics. Magnesia is standardized in IEC 672 as type C-820, a porous ceramic insulating material. Its main applications are found within the field of electrical heat engineering, primarily because of its high specific heat for use in heat storage elements, but also because of its favorable thermal conductivity and excellent insulation properties for use as an insulator in tubular heating elements. In the latter case, the ceramic is inserted into the cavity between the heating wire and the outer tubes and subsequently crushed, compacted, and solidified by rolling, hammering, and pressing.

Zirconia Ceramics. Zirconia ceramics are standardized as type C-830 in IEC 672. They can only be used as insulators at room temperature because the oxygen ions of zirconia become highly mobile at higher temperatures and the material then acts as a semiconductor. This principle forms the basis of the widespread application of zirconia as a sensor for determining the partial pressure of oxygen.

3.1.2.6. Insulating Materials Containing Titanium Oxide and Titanates

Whereas the materials discussed so far are generally employed purely as insulators, the use of components containing titanium dioxide and titanates is governed by their high relative permittivity.

Material Types. Group C-300 of IEC 672 is subdivided into seven types:

C-310	titanium dioxide
C-320	magnesium titanate
C-330	titanium dioxide incorporating other oxides as additives
C-331	as for type C-330
C-340	strontium, calcium, and bismuth titanates
C-350	barium titanate
C-351	barium titanate

The materials differ with respect to their relative permittivity, the associated temperature coefficient, and the dielectric dissipation factor. Types C-310, C-320, C-330, C-331, and C-340 have been classified according to international regulations as Type I capacitor materials; according to DIN 41 920, they are subdivided on the basis of the temperature coefficients of their relative permittivities. Material types C-350 and C-351 have a high relative permittivity and are classified as Type II substances.

Titanium Dioxide and Mixed Dielectrica Based on Rutile. Titanium dioxide [*13463-67-7*] crystallizes in the rutilestructure, which is always formed at temperatures above 1000 °C evenwhen the metastable anatase is used as a starting material.

The relative permittivity of rutile is strongly anisotropic. In polycrystalline materials with random crystal orientation, the relative permittivity has an average value of 117. The temperature coefficient of the relative permittivity is highly negative ($-0.001\,K^{-1}$).

Properties (see Table 5) can be modified by additives, such as zirconium oxide, lanthanum oxide, or magnesium titanate. These additives possess considerably lower relative permittivities and positive temperature coefficients. By following a logarithmic mixing relationship for relative permittivity, or a linear mixing relationship for the temperature coefficient, specific materials with tailored properties can be obtained. The temperature coefficient and the relative permittivity decrease with increasing additive concentrations. Dissipation factors are low ($\leq 10^{-3}$); with mixed dielectrics containing zirconium oxide, they are independent of frequency up to 10^8 Hz.

Dielectrics of this type are used to compensate for the increase in capacitance with temperature produced by other components. The primary criterion for the selection of the dielectrics is the magnitude of the negative temperature coefficient of the relative permittivity rather than the permittivity itself.

Materials Containing Titanates. The properties of titanate-based insulating materials are listed in Table 6.

Type C-320 IEC 672 comprises materials based on magnesium titanate. The outstanding

Table 5. Important physical properties of dielectrics based on titanium dioxide classified according to IEC 672

Type	C-310	C-330	C-331
Open porosity, vol%	0	0	0
Bulk density, g/cm^3	>3.5	>4	>4.5
Bending strength, N/mm	>70	>80	>80
Electrical strength, kV/mm	>8	>10	>10
Withstand voltage (1 min), kV	>15	>15	>15
Relative permittivity	40–100	25–50	30–70
Dielectric dissipation factor (tan δ)			
1 kHz	<0.0065	<0.020	<0.007
1 MHz	<0.002	<0.0008	<0.001
Resistivity at 20 °C, $\Omega \cdot$ cm	$>10^{12}$	$>10^{11}$	$>10^{11}$
Temperature coefficient of the relative permittivity, 10^{-6} K^{-1}	−280 to −900	+70 to −120	−120 to −700

Table 6. Important physical properties of titanate-based dielectrica classified according to IEC 672

Type	C-320	C-340	C-350	C-351
Open porosity, vol%	0	0	0	0
Bulk density in raw state, g/cm^3	>3.1	>3.0	>4.0	>4.0
Bending strength, N/mm^2	>70	>70	>50	>50
Electrical strength, kV/mm	>8	>6	>2	>2
Withstand voltage (1 min), kV	>15	>8	>2	>2
Relative permittivity	12–40	100–700	350–3000	>3000
Dielectric dissipation factor (tan δ) at 1 MHz	<0.0015	<0.005	<0.035	<0.035
Volume resistivity at 20 °C, $\Omega \cdot$ cm	$>10^{11}$	$>10^{11}$	$>10^{10}$	$>10^{10}$
Temperature coefficient of the relative permittivity, 10^{-6} K^{-1}	+130 to −150	−1200 to −6000	nonlinear	nonlinear

feature of these materials is the excellent temperature stability of the relative permittivity, the values of the latter lie between 12 and 40.

Type I capacitors with extremely high negative temperature coefficients for the relative permittivity (up to 0.006 K^{-1}) can be obtained with materials based on strontium and/or calcium titanate (type C-340 in IEC 672).

Barium titanate [*12047-27-7*], $BaTiO_3$, is the most important alkaline-earth titanate (→ Ceramics, Electronic). It also finds application as a piezoelectric material.

Because of their piezoelectric properties, considerable interest has also been centered on titanates that contain lead in the form of mixed crystals with zirconates, stannates, and niobates.

Like all of the other titanates listed above, barium titanate crystallizes to give a deformed perovskite structure. Slight distortions lead to a tetrahedal structure at room temperature, a rhombic structure below 0 °C, and trigonal symmetry below −90°C. The most significant property of the distorted structure without a center of symmetry is its extremely large relative permittivity, particularly in the direction perpendicular to the main crystal axis (Fig. 11). Maximum relative permittivity values are observed at each of the crystallographic transition temperatures, but are accompanied by a maximum in the dielectric loss (Fig. 12). This results in a strong, nonlinear temperature dependence of the relative permittivity. The dipoles in the deformed crystal interact mutually and are aligned in parallel in small areas known as domains. The direction of dipole orientation in neighboring domains is reversed so that the whole system assumes an energy minimum and does not have a net external charge. On applying an external electrical field, the domains become oriented in the direction of the field to an extent proportional to the electrical field strength. Because of the anisotropy of the dielectric properties, this reorientation of the domains is coupled with a change in the relative permittivity, i.e., the relative permittivity is also dependent upon field strength. The combination of phenomena discussed here is known as fer-

roelectricity (see also → Ceramics, Electronic, Chap. 3.).

Above a certain temperature known as the Curie temperature (in the case of barium titanate 120 °C), the structure becomes strictly cubic, and the ferroelectric properties are lost. Above the Curie temperature T_c, the changes in the relative permittivity with temperature are described by the Curie – Weiss law:

$$\varepsilon_T = \frac{C}{T - T_c}$$

where C is the Curie constant and ε_T is the relative permittivity at temperature T. Compounds exhibiting a perovskite structure tend to form mixed crystals. Hence Ba^{2+} can be replaced by Sr^{2+}, Pb^{2+}, and to a certain extent also by Ca^{2+}; Ti^{4+} can be substituted by Zr^{4+} or Sn^{4+}. Similarly, ion combinations can be incorporated providing the ionic radii are equivalent and a balance of charge is attained (complex perovskites). These substitutions produce changes in the relative permittivity and its temperature dependency, as well as in the dielectric loss and the Curie temperature. In this way, a wide range of insulating materials with differing properties can be produced.

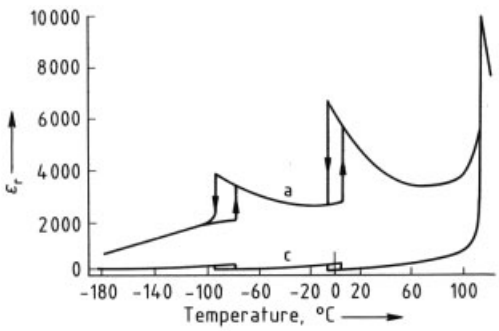

Figure 11. Relative permittivity of barium titanate along the a and c crystal axes as a function of temperature

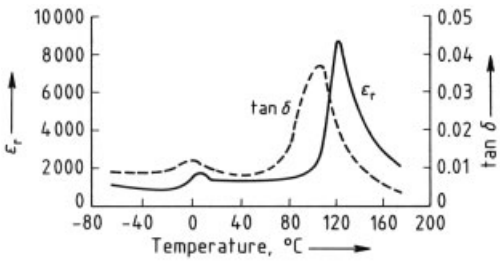

Figure 12. Temperature dependence of the relative permittivity (ε_r) and the dielectric loss ($\tan\delta$) of polycrystalline barium titanate

Certain oxide additives (e.g., a small excess of titanium dioxide) act as grain boundary phases which prevent crystal growth. In small crystals, ferroelectric domains are still formed, but their changes of position and size in an external field are considerably hindered. Materials of this kind show reduced hysteresis loss, and the curve of $\varepsilon_T = f(T)$ flattens out in the vicinity of the Curie temperature. Because of their low losses, such dielectrics are employed as contact-protective capacitors. Extremely small crystals (ca. 50 nm) cannot be reorientated in an external electric field and hence do not exhibit ferroelectric properties.

Other additives (e.g., bismuth oxide, bismuth titanate, zinc oxide, or zinc titanate) extend the firing interval, but also change the shape of the curve $\varepsilon_T = f(T)$. Bismuth titanate additives produce dielectric materials of low temperature dependency in their electrical properties.

If a dielectric is made from two perovskite materials that do not form or only partially form mixed crystals, then the change in the relative permittivity between the relevant Curie temperatures is low. Examples of such dielectrics include those containing calcium and barium titanate.

In titanate-based products, the raw materials are thoroughly mixed and calcined onto a frit, which is subsequently ground, shaped, and fired. In order to prevent chemical reduction of the titanium, which would cause the article to have semiconductor properties, capacitor materials must be fired in an oxidizing atmosphere. The electrodes and contact layers are formed by depositing metal coatings, generally silver.

Types C-350 and C-351 in IEC 672 have a relative permittivity below and above 3000, respectively.

The main use of titanate-containing dielectrics is as capacitors in high-frequency applications. Examples include the bridging of d.c. resistors and the manufacture of capacitive couplers. Furthermore, perovskite-based materials containing titanate have gained considerable significance as piezoelectric components.

When titanates are fired under reducing conditions, semiconductor materials are obtained which reoxidize at the interfacial regions during the firing of metallizing layers. This process is used to produce intergranular barrier layer capacitors which possess particularly high capac-

ities due to the low thickness of the dielectric layer (Type III dielectrics).

3.1.2.7. Nonoxide Ceramics

More recently, increased interest has been shown in nonoxide inorganic materials because of several unusual properties. They have a very high mechanical strength, even at elevated temperature. The most important materials are compounds of elements of groups 3, 4, and 5 of the periodic table. Silicon nitride (Si_3N_4), silicon carbide (SiC), and aluminum nitride (AlN) deserve special mention.

The essentially homopolar bonding character between the atoms of these compounds determines their physical properties to a large extent, but also causes problems during the manufacture of components when using traditional procedures employed in the ceramic industry. Production is thus more complicated and more expensive. Nonoxide ceramic materials are only suitable as insulating materials when their extraordinary properties provide a particularly elegant solution for a given problem. Silicon nitride [*12033-89-5*] andaluminum nitride [*24304-00-5*] are used as electrical insulators.

Silicon carbide is not suitable for electrical insulation purposes because its electrical conductivity increases strongly with temperature; it is even employed as a high-temperature heating element.

The thermal conductivity of aluminum nitride is considerably greater than that of aluminum oxide. The low density of 3.2 g/cm^3 makes it a suitable alternative to beryllium oxide when the effective removal of heat generated by dielectric losses is important. At increasing temperatures, the decrease in thermal conductivity of aluminum nitride is considerably lower than that of the oxide materials. A further advantage of aluminum nitride, especially as a substrate for high-power silicon components, is the fact that the thermal expansion coefficient is closer to that of silicon than any other potential substrate materials.

Silicon nitride exhibits similar insulation properties to aluminum oxide, but its thermal expansion coefficient is only half as large, and correspondingly shows greater thermal shock resistance. This characteristic is of considerable importance in glow plugs in diesel engines.

3.1.2.8. Glass (→ Glass)

Glass materials have been used as insulators in electrical engineering for many years. The systematic classification and establishment of minimum requirements occurred during their incorporation in the IEC standard 672.

Material Types. Glass materials used as electrical insulators can be classified into six groups:

G-100	alkali – lime – silica glasses
G-200	chemically resistant borosilicate glasses
G-300	electrically resistant borosilicate glasses
G-400	alumina – lime – silica glasses
G-500	lead oxide – alkali – silica glasses
G-600	barium oxide – alkali – silica glasses

Group G-100 consists of:

Type G-110 alkali – lime – silica glasses, annealed
Type G-120 alkali – lime – silica glasses, toughened

The mechanical strength is increased tremendously by prestressing (Type G-120).

Group G-300 consist of:

Type G-310 borosilicate glasses, low loss
Type G-320 borosilicate glasses, stable to high voltage

Properties. Several of the requirements specified in the IEC 672 are summarized in Table 7. A range of properties of glass are governed by its chemical composition. Characteristic data can be determined by multiplying the content of a glass component (expressed in percent) by a property-specific factor, and summing the calculated values. Properties that can determined in this way include the thermal expansion and the surface energy.

Production. Glass is produced by melting appropriate mixtures of raw materials and molding them in the molten state (→ Glass). On cool-

Table 7. Important physical properties of glass-insulating materials classified according to IEC 672. Figures in bold face indicate properties that are of importance for intended applications.

Property	G-110	G-120	G-200	G-310	G-320	G-400	G-500	G-600
Bending strength, N/mm^2	**>30**	**>150**	**>30**	**>30**	**>30**	**>40**	**>30**	**>30**
Linear expansion coefficient, 10^{-6} K^{-1}	**8.5–10**	**8.5–10**	**3–5**	**4.6–5.1**	**4.6–5.5**	**4–4.6**	**8–10**	**9–10**
Glass transition temperature, °C	500–560	500–560	520–560	**480–510**		**620–730**	**430–470**	430–500
Electrical strength, kV/mm	**>25**	**>25**	**>30**	**>30**	**>30**	**>30**		
Relative permittivity at 20 °C and 1 MHz	6.5–7.6	7.3–7.6	4.0–5.5	4.9–5.5	5–6	5.5–7.5	6–8	6.5–7.5
Dielectric dissipation factor at 20 °C								
50 Hz	<0.030	<0.060	<0.020	**<0.0035**	<0.030	<0.0025	<0.003	<0.004
1 MHz	<0.010	<0.060	<0.010	<0.002	<0.008	<0.003	<0.002	<0.0025
Volume resistivity, Ω · cm								
at 20 °C	$>10^{12}$	$>10^{12}$	$>10^{14}$	$>10^{14}$	$>10^{14}$	$>10^{14}$	$>10^{14}$	$>10^{14}$
at 200 °C	$>\mathbf{10^{7}}$	$>\mathbf{10^{7}}$	$>\mathbf{10^{9}}$	$>\mathbf{10^{10}}$	$>\mathbf{10^{8}}$	$>\mathbf{10^{12}}$	$>\mathbf{10^{10}}$	$>\mathbf{10^{10}}$

ing, the melt vitrifies to form the desired component. Residual stress can be removed by annealing. However, compressive stresses can also be deliberately generated in order to increase the mechanical strength of the component.

Insulator components are mainly shaped by pressing the viscous melt in steel molds, although recently, extrusion has been introduced as an additional shaping method. Glass can also be processed by blowing.

Uses. The applications for glass in electrical insulator are determined by properties that can be particularly well optimized in certain types of glass; these properties are indicated by bold type in Table 7.

Glass materials from group G-100 are frequently employed for overhead line insulators. Cap insulators are most common and are combined to form long chains. Long rod insulators are also used but are more difficult to manufacture.

Glass insulators have been increasingly employed with overhead lines. Their major advantage is that faults can easily be identified because they are transparent; slightly defective components can therefore be exchanged before a major failure occurs.

The high requirements placed on the mechanical properties of suspension insulators for overhead lines are met by toughened glass. Glass ceramics (which have not yet been included in the IEC standard) have also been recently employed because they have a high strength combined with a low thermal expansion coefficient and correspondingly an improved thermal shock resistance (→ Glass Ceramics).

Glass materials of type G-200 (chemically resistant borosilicate glass) are used in discharge lamps. The low coefficient of thermal expansion resulting from the use of boron oxide ensures that the glass has a high thermal shock resistance. Boron also improves chemical resistance.

Type G-310 glass materials exhibit favorable properties for lead through insulators in high-frequency applications. They are also suitable for glass components designed to have capacitor functions. Type G-320 materials are employed in valves for high-voltage applications, e.g., for X-ray tubes.

Type G-400 glass materials were developed to have optimum properties for high-performance halogen lamps; their properties have now been standardized.

Type G-500 and G-600 glasses are employed when metallic components, particularly those based on nickel–iron alloys, have to be sealed into electric lamps and valves. They have a low glass transition temperature and a large thermal expansion coefficient to ensure that only weak mechanical stresses are generated in the composite system. Because of the incorporation of heavy metal oxides, they are also used in televisions to absorb the generated X-rays. Glass materials containing lead can change on exposure to X-rays and electron beams, and thus become discolored or even produce tracking currents. In applications where these effects must be excluded, lead-free glasses of type G-600 must be employed.

3.2. Organic Insulating Materials

The following discussion of the most common organic materials employed as electrical insulators includes products that meet all or many of the imposed requirements. This does not mean that all of the products are employed in large amounts in the electronics industry. The compilation is by no means complete, and is intended merely to provide an overview. Description of the products has been kept brief. Details are available from manufacturers' information and specialist publications. Examples of trade names are given for each group of polymers but these lists are by no means exhaustive. Most of the polymers are described in detail elsewhere in the encyclopedia. Emphasis has been placed on graphic representations–the relief diagrams. As opposed to tables, these diagrams allow the rapid visual assessment of the suitability of a material for a given electrotechnical project. Different materials can also be compared. The block lengths demonstrate the quality variation of a particular material class. Each diagram shows:

1) the *comparative tracking index* (CTI) measured according to IEC 112 (see Section 2.1)
2) the *volume resistivity* (ϱ) measured according to IEC 93 at 23 °C and 50 % R.H. (see Section 2.2)
3) the *surface resistivity* (σ) measured according to IEC 93 at 23 °C and 50 % R.H. (see Section 2.3)
4) the *loss index* ε''_r measured according to IEC 250 at 1 kHz, 23 °C, and 50 % R.H. (see Section 2.6)
5) and the *relative permittivity* ε_r measured according to IEC 250 at 1 kHz, 23 °C, and 50 % R.H. (see Section 2.4)

The electric strength is not shown because although the test results obtained in accordance with the standard can be used for detecting changes or deviations from normal characteristics resulting from processing variables, aging conditions, or other manufacturing or environmental situations, they can seldom be used directly to determine the behavior of insulating materials in a real application. Measured values of the electric strength are affected by many factors (Section 2.7), but these factors are not specified when the results are published.

The significance and determination of the parameters presented in the graphics are described in Chapters 1, 2, and 4. Values can easily be read from the plots. The dielectric dissipation factor is obtained by dividing the loss index (column 4) by the corresponding value for relative permittivity (column 5).

3.2.1. Thermoplasts

Thermoplasts, also referred to as plastomers, are solid synthetic resins which can be repeatedly softened without chemical modification by heating in a relatively narrow temperature range. They can be shaped while soft and generally consist of linear molecules. Physical properties of the thermoplasts are compiled in Table 8.

Acrylonitrile – Butadiene – Styrene. Physical properties are given in Table 8 and electrical properties in Figure 13 (see also → Polystyrene and Styrene Copolymers).

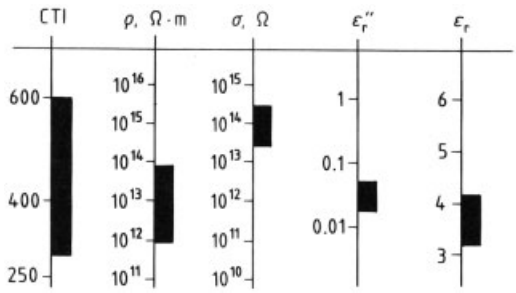

Figure 13. Relief diagram of electrical properties of acrylonitrile – butadiene – styrene

Special properties: high impact toughness, thermally stable, unlimited shaping characteristics, limited stability to UV radiation.

Uses : casings for office machinery, electrical industry, electronics, telecommunications, protective insulating components.

Trade Names: Cycolac (Borg Warner Chemicals, USA); Kralastic (USS Chemicals, USA); Novodur (Bayer, FRG); Terluran (BASF, FRG); Diapet (Mitsubishi Rayon Co., Japan); Ravikral (EniChem, Italy).

Acrylonitrile – Styrene – Acrylic Ester. Physical properties are given in Table 8 and electrical properties in Figure 14.

Table 8. Physical properties of thermoplasts

Polymer	CAS registry number	Density, g/cm^3	Tensile strength, N/mm^2	Ultimate strain, %	Long term thermal stability, °C
Acrylonitrile – butadiene – styrene copolymer (ABS)	[*9003-56-9*]	1.1	45	20	80
Acrylonitrile – styrene – acrylic ester (ASA)	[*9003-54-7*]	1.1	50	15	85
Cellulose acetate (CA)	[*9004-35-7*]	1.2	35	30	40
Cellulose acetate butyrate (CAB)	[*9004-36-8*]	1.3	35	30	40
Cellulose propionate (CP)	[*9004-64-2*]	1.2	20 – 50	4	40
Perfluoro(ethylene – propylene) (FEP)	[*25067-11-2*]	2.2	20	300	70
Polyamide (PA)		1.1	60	200	60 – 90
Polyamide 12-resin (PA)	[*24937-16-4*]	1	55	200	150
Polyamideimide (PAI)		1.4	220	6	220
Polybutene (PB)	[*9003-29-6*]	0.92	10	15	100
Poly(butylene terephthalate) (PBTP)	[*24968-12-5*]	1.3	60	5 – 200	120
Polycarbonate (PC)	[*25037-45-0*]	1.2	60	60 – 100	90
Polycarbonate alloy (PC – ABS)		1.2	60	60 – 100	90
Polycarbonate alloy (PC – PBTP)		1.3	60	60 – 100	90
Polyethylene (PE)	[*9002-88-4*]	0.92 – 0.96	9 – 23	650	60 – 80
Cross-linked polyethylene (PE-X)		1.3	9 (low density) 23 (high density)	350	130
Polyetherimide (PEI)	[*61128-24-3*]	1.3	105	60	170
Polyethersulfone (PES)	[*25667-42-9*]	1.4	85	20	200
Poly(ethylene terephthalate) (PETP)	[*25038-59-9*]	1.4	240	90	130
Polyimide (PI)	[*25036-53-7*]	1.4	90	10	250
Poly(methyl methacrylate) (PMMA)	[*9011-14-7*]	1.2	80	5	70
Poly(4-methyl-1-pentene) (PMP)	[*9016-80-2*]	0.8	25	10 – 50	150
Polyoxymethylene (POM)	[*9002-81-7*]	1.4 – 1.7	2 – 20	30	100
Polypropylene (PP)	[*9003-07-0*]	0.9	30	600	100
Poly(phenylene oxide) (PPO)	[*9041-80-9*]	1.1	60	30	175
Poly(phenylene sulfide) (PPS)	[*9016-75-5*]	1.7	200	2	240
Polystyrene (PS)	[*9003-53-6*]	1.1	60	3	60
Polysulfone (PSU)	[*25135-51-7*]	1.2	85	2 – 70	160
Polytetrafluoroethylene (PTFE)	[*9002-84-0*]	2.2	25	300	260
Poly(vinyl chloride) (PVC)	[*9002-86-2*]	1.4	55	30	60
Styrene – acrylonitrile copolymer (SAN)	[*9003-54-7*]	1.1	70	4	85
Styrene – butadiene copolymer (SB)	[*9003-55-8*]	1.1	40	35	60

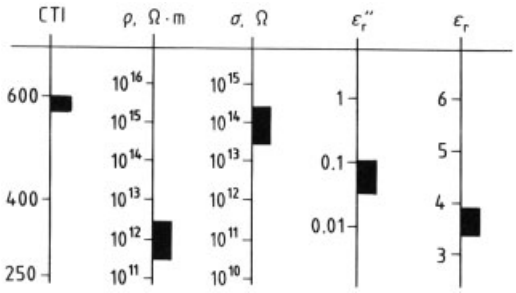

Figure 14. Relief diagram of electrical properties of acrylonitrile – styrene – acrylic ester

Special properties: stable to UV radiation, weather-resistant, thermally stable over long periods, unlimited shaping characteristics.

Uses: insulators, insulating material, ventilators, casings for electronic equipment, electrical industry, electronics, telecommunications, protective insulating components.

Trade Names: Richform (Richmond Div. of Dixico, USA); Luran S (BASF, FRG); Geloy (General Electric, USA).

Cellulose Acetate. Physical properties are given in Table 8 and electrical properties in Figure 15 (see also → Cellulose Esters).

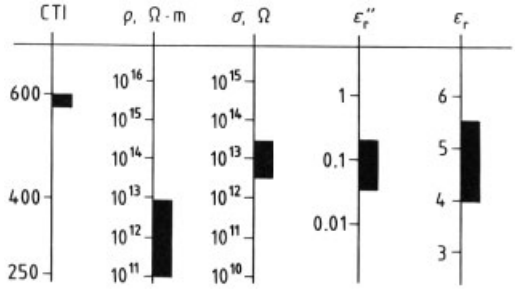

Figure 15. Relief diagram of electrical properties of cellulose acetate

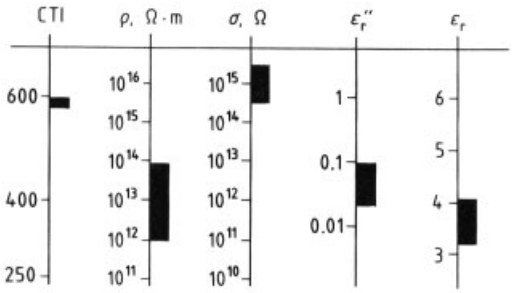

Figure 16. Relief diagram of electrical properties of cellulose acetate butyrate

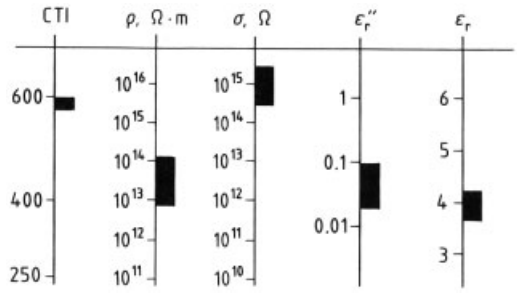

Figure 17. Relief diagram of electrical properties of cellulose propionate

Special properties: antistatic properties despite high electrical resistance, crystal clear, tough, hard, scratch-resistant, insensitive to stress cracking, readily dyeable with brilliant colors, not permanently weather-resistant, less suitable for precision components.

Uses: tool handles, electric insulation films, lights, casings.

Trade Names: Dexel (Courtaulds Plastic Group, UK); Ampol (American Polymers, USA); Tenite (Eastman Chemical International, USA).

Cellulose Acetate Butyrate. Physical properties are given in Table 8 and electrical properties in Figure 16 (see also → Cellulose Esters).

Special properties: antistatic despite high electrical resistance, crystal clear, tough, hard, scratch resistant, insensitive to stress-cracking, readily and brilliantly dyeable, not suitable for precision components.

Uses: tool handles, electric insulation films, lights, casings.

Trade Names: Ampol (American Polymers, USA); Tenite (Eastman Chemical International, USA).

Cellulose Propionate. Physical properties are given in Table 8 and electrical properties in Figure 17 (see also → Cellulose Esters).

Special properties: antistatic despite high electrical resistance, crystal clear, tough, scratch-resistant, insensitive to stress cracking, readily dyeable with brilliant colors, not suitable for precision components.

Uses: tool handles, electric insulation films, lights, casings.

Trade Names: Ampol (American Polymers, USA); Tenite (Eastman Chemical International, USA).

Perfluoro(Ethylene – Propylene). Physical properties are given in Table 8 and electrical properties in Figure 18 (see also → Fluoropolymers, Organic, Chap. 2.3.).

Special properties: nonflammable, weather-resistant, resistant to solvents, no water uptake, physiologically inert, stable to pressure.

Uses: flexible printed circuits, coaxial cables for computers, cable insulation, shrinkable tubes.

Trade Names: Teflon (Du Pont, USA); Neoflon (Daikin Kogyo Co., Japan).

Polyamide. Physical properties are given in Table 8 and electrical properties in Figure 19 (see also → Polyamides).

Special properties: normal minimum water content of 2.5 %, prone to stress cracking when freshly injected, impact resistant, viscoelastic, puncture resistant, dynamic-loading capability, good slip properties.

Uses: high-voltage insulation, multiple connecting plugs, relays, extinguishing chambers, wear-resistant insulation and sleeve materials for cable technology.

Trade Names: Ultramid (BASF, FRG); Durethan (Bayer, FRG); Technyl (Rhône-Poulenc Specialities Chimiques, France); Grilon (Ems-Chemie, Switzerland); Zytel (Du Pont, USA); Maranyl (ICI, UK).

Polyamide 12-Resin. Physical properties are given in Table 8 and electrical properties in Figure 20.

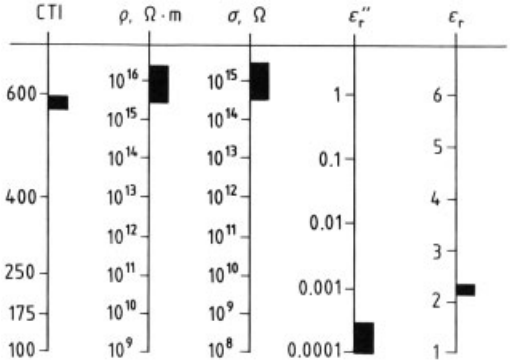

Figure 18. Relief diagram of electrical properties of perfluoro(ethylene – propylene)

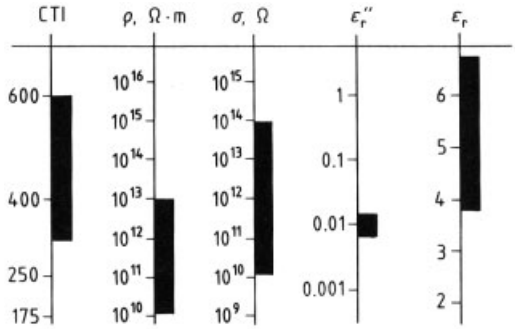

Figure 19. Relief diagram of electrical properties of polyamide

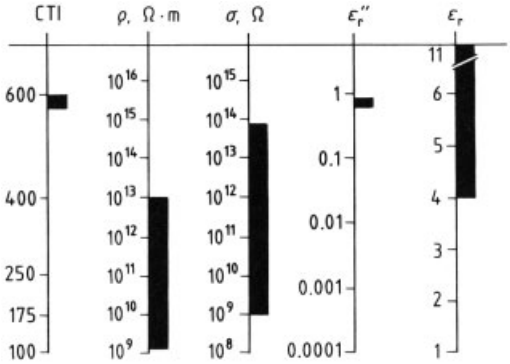

Figure 20. Relief diagram of electrical properties of polyamide 12-resin

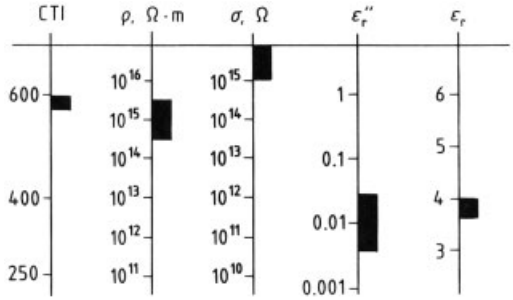

Figure 21. Relief diagram of electrical properties of polyamideimide

Special properties: resistant to organic solvents, fuels, oils, weak acids and alkalis.

Uses: casings, components for electrical devices.

Trade Names: Vestamid (Hüls, FRG); Rilsan (Atochem, France); Grilamid (Ems-Chemie, Switzerland).

Polyamideimide. Physical properties are given in Table 8 and electrical properties in Figure 21.

Special properties: resistant to elevated and cryogenic temperatures; nonmeltable; resistant to hydrocarbons, weak acids, and weak alkalis; resistant to highly energetic radiation sources; very low coefficient of friction.

Uses: special material for relays and multiple connection plugs used under extreme conditions, printed circuits, suitable for solderable connecting cable.

Polybutene. Physical properties are given in Table 8 and electrical properties in Figure 22 (see also → Polyolefins).

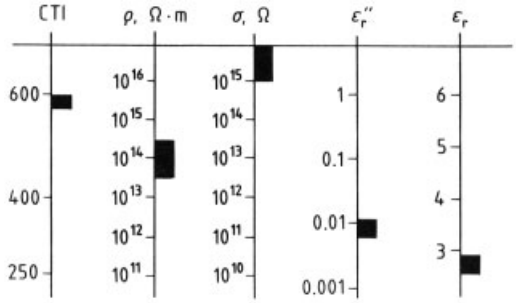

Figure 22. Relief diagram of electrical properties of polybutene

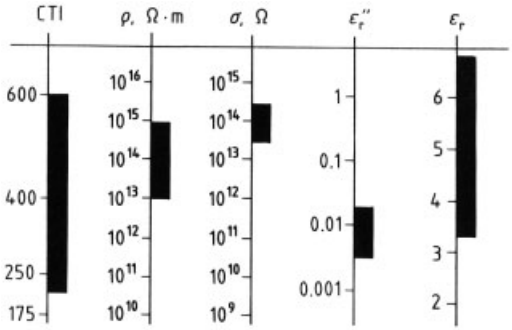

Figure 23. Relief diagram of electrical properties of poly(butylene terephthalate)

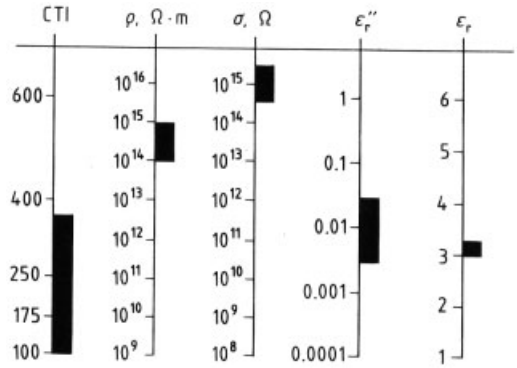

Figure 24. Relief diagram of electrical properties of polycarbonate

Special properties: high long-term stability, low tracking propensity at high temperature, high stress cracking stability, resistant to low temperature, easily weldable.

Uses: insulation jackets for wires, cable sleeves, pipes.

Trade Names: Duraflex (Shell International Chemical Co., UK); Krene (Union Carbide, USA).

Poly(Butylene Terephthalate) (PBT). Physical properties are given in Table 8 and electrical properties in Figure 23 (see also → Polyesters).

Special properties: impact toughness, rigid, maintains shape, scratch-resistant, good slip and abrasion behavior, electrical properties are essentially frequency- and temperature-independent.

Uses: electrical insulation in household equipment, relay capstans, connecting cable, components for switches, sparking plug cases.

Trade Names: Hostadur (Hoechst, FRG); Pocan (Bayer, FRG); Vestodur (Hüls, FRG); Ultradur (BASF, FRG); Celanex (Celanese Engineering Resins, USA); Novadur (Mitsubishi Chemical Industries, Japan).

Polycarbonate (PC). Physical properties are given in Table 8 and electrical properties in Figure 24 (see also → Polycarbonates).

Special properties: low combustion, rigid elastic, high impact toughness above 90 °C, long term stability at high temperature, limited suitability under abrasive conditions, electrical properties practically independent of humidity.

Uses: moldings, casings, appliances, relay elements, condenser housings, plug connections, high-voltage insulation, cable distribution panels, mounting boards, screens, compact discs, information storage media.

Trade Names: Makrolon (Bayer, FRG); Orgallan (Atochem, France); Merlon (Mobay Chemical Corp., USA); Lexan (General Electric, USA); Jupilon (Mitsubishi Gas Chem, Japan); Sinvet (EniChem, Italy).

Polycarbonate Alloy (PC – ABS). Physical properties are given in Table 8 and electrical properties in Figure 25.

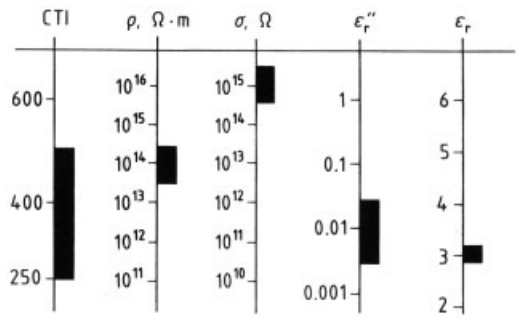

Figure 25. Relief diagram of electrical properties of polycarbonate alloy (PC – ABS)

Special properties: resistant to combustion, rigid elastic, tough above 90 °C, long term stability at high temperatures, limited suitability under abrasive conditions, electrical characteristics practically independent of humidity.

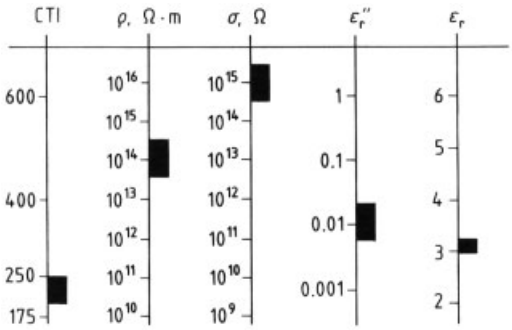

Figure 26. Relief diagram of electrical properties of polycarbonate alloy (PC – PBTP)

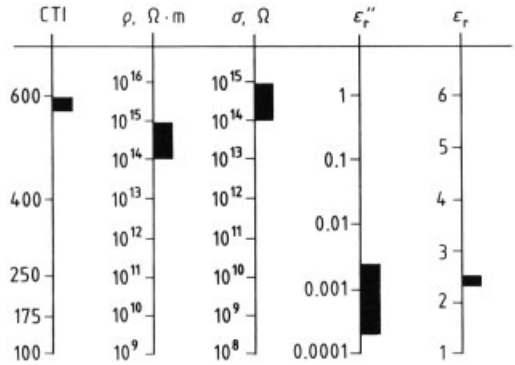

Figure 27. Relief diagram of electrical properties of polyethylene

Uses: moldings, casings, appliances, relay elements, condensor housings, plug connections, high-voltage insulation, cable distribution panels, mounting boards, screens.
Trade Names: Cycolac (Borg Warner Chemicals, USA); Bayblend (Bayer, FRG); Cycoloy (Borg Warner Chemicals, USA).

Polycarbonate Alloy (PC – PBTP). Physical properties are given in Table 8 and electrical properties in Figure 26.

Special properties: resistant to combustion, rigid elastic, tough above 90 °C, long term stability at high temperatures, limited suitability under abrasive conditions, electrical characteristics practically independent of humidity.

Uses: moldings, casings, consumer equipment, relay elements, condensor housings, plug connections, high voltage insulation, cable distribution panels, mounting boards, screens.
Trade Names: Makroblend (Bayer, FRG); Xenoy (General Electric, USA).

Polyethylene. Physical properties are given in Table 8 and electrical properties in Figure 27 (see also → Polyolefins).

Special properties: resistant to water, saline solutions, alkalis, acids, (exception: strong oxidizing agents), can be subjected to high mechanical stress, hard, resistant to low temperatures.

Uses: insulation and sleeve materials for cables, pipes, with 30 % carbon black deflects lines of force in conductors, transformers, etc.
Trade Names: Lupolen (BASF, FRG); Carlona (Shell International Chemical Co., UK); Alathone (Du Pont, USA); Novatec (Mitsubishi Chemical Industries, Japan); Dylan (Arco Chemical Co., USA).

Cross-linked Polyethylene. Physical properties are given in Table 8 and electrical properties in Figure 28.

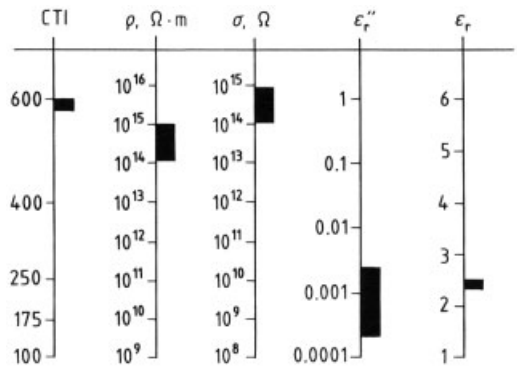

Figure 28. Relief diagram of electrical properties of cross-linked polyethylene

Special properties: stable at high temperature, high impact strength at low temperature, resistant to stress cracking, greater resistance to elevated temperatures than polyethylene.

Uses: insulation and sleeve materials for cables, shrinkable sleeves.
Trade Names: Lubonyl (N. Lundbergs Fabrics AB, Sweden); Alveolit (Sekisui Chemical Co., Japan); Alveolux (Alveo AG, Switzerland); Blanex (Reichhold Chemicals Inc., USA); Elapor (Elastogran, FRG).

Polyetherimide. Physical properties are given in Table 8 and electrical properties in Figure 29 (see also → Polymers, High-Temperature).

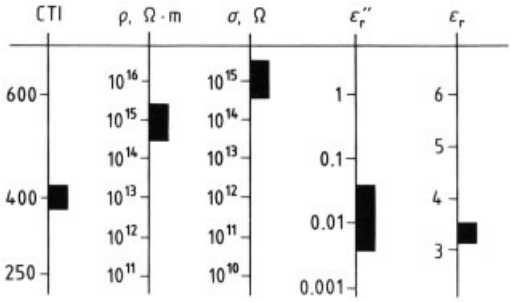

Figure 29. Relief diagram of electrical properties of polyetherimide

Special properties: transparent yellow; high heat resistance; very high resistance to combustion; resistant to alcohols; fuels; lubricants; detergents, acids and weak alkalis; resistant to UV light, high energy sources of radiation, hydrolysis, hot water, and steam.

Uses: casings, equipment, protective relays for high-voltage supplies, multiple connecting plugs, soldered connecting cables, components for integrated circuits.

Trade Names: Ultem (General Electric Co., USA).

Polyethersulfone. Physical properties are given in Table 8 and electrical properties in Figure 30 (see also → Polymers, High-Temperature).

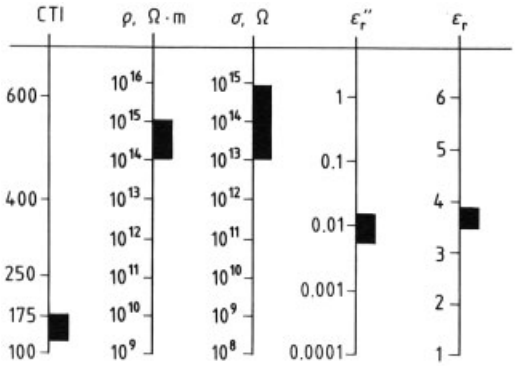

Figure 30. Relief diagram of electrical properties of polyethersulfone

Special properties: transparent yellow, high long-term stability, low tracking propensity until the glass transition temperature is approached, resistant to acids and alkalis, not susceptible to hydrolysis, resistant to oxidation, highly incombustible, dielectric values essentially temperature- and frequency-independent.

Uses: involving high demands in electrical engineering, insulation films.

Trade Names: Udel (Amaco Chemicals, USA); Ultrason (BASF, FRG); Victrex (ICI, UK).

Poly(Ethylene Terephthalate). Physical properties are given in Table 8 and electrical properties in Figure 31 (see also → Polyesters).

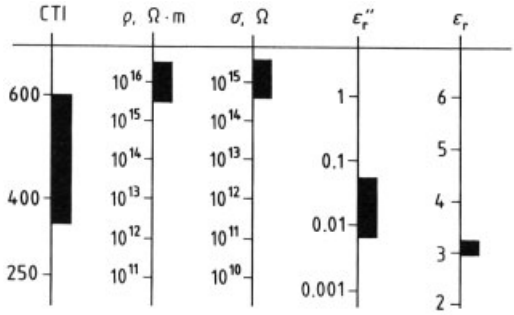

Figure 31. Relief diagram of electrical properties of poly-(ethylene terephthalate)

Special properties: tough, stiff, maintains shape at high temperature, good slip and abrasion behavior, electrical properties are only slightly temperature- and frequency-dependent.

Uses: insulation and sleeve materials for cables, sonic substrates, insulation materials, insulation films.

Trade Names: Hostaphan (Hoechst, FRG); Crastin (Ciba-Geigy, Switzerland); Melinar (ICI, UK); Rynite (Du Pont, USA).

Polyimide. Physical properties are given in Table 8 and electrical properties in Figure 32 (see also → Polyimides).

Special properties: high thermal stability, resistant to most dilute or weak acids, attacked by strong alkalis and ammonia, permanent humidity may cause cracking, pronounced resistance to high-energy radiation.

Uses: electrical and electronic components, films, shrinkage films.

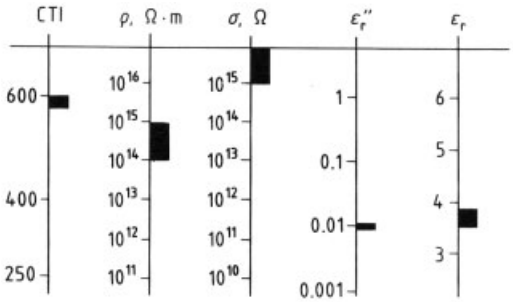

Figure 32. Relief diagram of electrical properties of polyimide

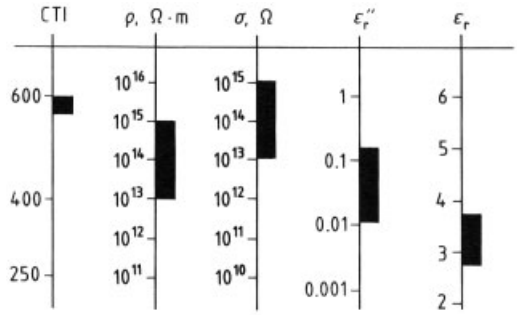

Figure 33. Relief diagram of electrical properties of poly(methyl methacrylate)

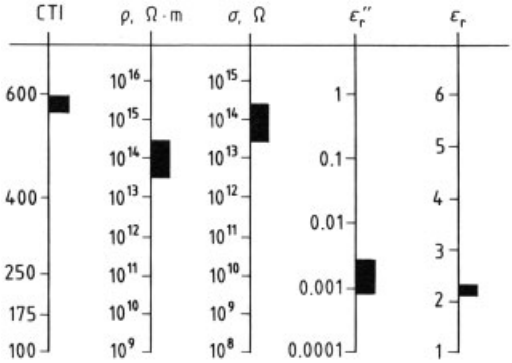

Figure 34. Relief diagram of electrical properties of poly(4-methyl-1-pentene)

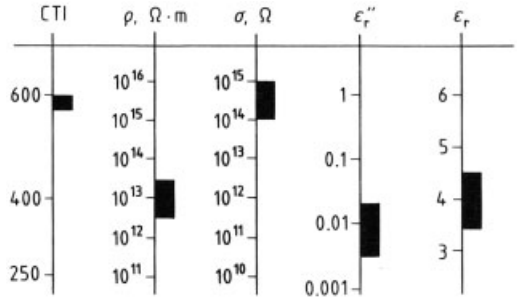

Figure 35. Relief diagram of electrical properties of polyoxymethylene

Trade Names: Kapton, Vespel (Du Pont, USA); Kinel (Rhône-Poulenc Specialities Chimiques, France).

Poly(Methyl Methacrylate). Physical properties are given in Table 8 and electrical properties in Figure 33 (see also → Polymethacrylates).

Special properties: hard, tough, highly transparent, weather-resistant, chemically inert to medium concentrations of acids, resistant to fuels, physiologically inert.

Uses: lamps, traffic lights, nautical aids, optical fibers, high-pressure mercury lamps, optical data storage media.

Trade Names: Lucite (Du Pont, USA); Plexiglas (Röhm, FRG); Altulite (Altulor Groupe CdF Chimie, France); Sumipex (Sumitomo Chemical Co., Japan).

Poly(4-methyl-1-pentene) (PMP). Physical properties are given in Table 8 and electrical properties in Figure 34 (see also → Polyolefins).

Special properties: light weight, crystal clear, rigid, not suitable for exterior applications, good thermal stability.

Uses: insulation for coaxial and electrical cables.

Trade Names: TPX (Mitsubishi Petrochemical Co., Japan).

Polyoxymethylene (POM). Physical properties are given in Table 8 and electrical properties in Figure 35 (see also → Polyoxymethylenes).

Special properties: high strength and stiffness, high dynamic loading capacity over a wide temperature range, high impact toughness even at −40 °C, high surface hardness, low coefficient of friction, good wear properties, low gas permeability, not resistant to UV light, physiologically

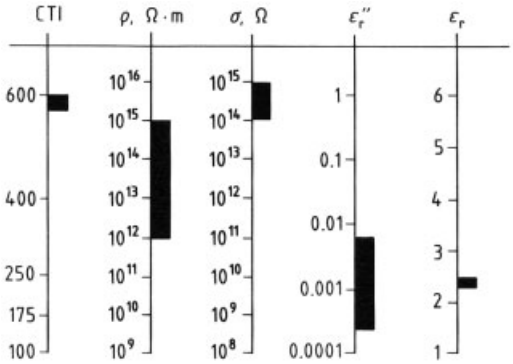

Figure 36. Relief diagram of electrical properties of polypropylene

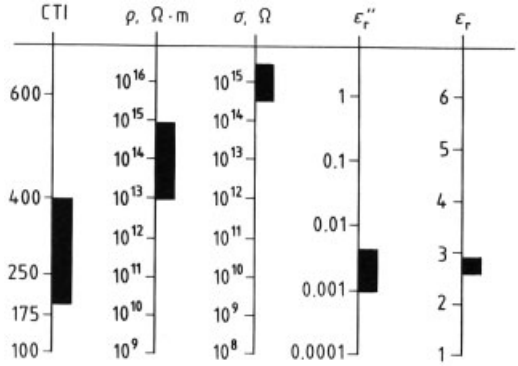

Figure 37. Relief diagram of electrical properties of poly(phenylene oxide)

inert, dielectric characteristics are practically independent of frequency and temperature.

Uses: precision components for measuring instruments in electronics and precision engineering, household appliances.

Trade Names: Delrin (Du Pont, USA); Ultraform (BASF, FRG); Tenac (Asahi Chemical Industries, Japan).

Polypropylene. Physical properties are given in Table 8 and electrical properties in Figure 36 (see also → Polyolefins).

Special properties: stiff, impact toughness, no stress cracking, resistant to hot water and chemicals.

Uses: accumulator boxes, casings for electrotechnical equipment, electrotechnical components.

Trade Names: Vestolen P (Hüls, FRG); Hostalen PP (Hoechst, FRG); Propathene (ICI, UK); Profax (Himont, USA); Noblen (Mitsubishi Petrochemical Co., Japan).

Poly(Phenylene Oxide). Physical properties are given in Table 8 and electrical properties in Figure 37 [see also → Poly(Phenylene Oxides)].

Special properties: high long-term stability; low tendency to tracking until the glass transition temperature is approached; resistant to acids, alkalis, hydrolysis, and oxidation; highly incombustible; transparent.

Uses: cable insulation for data processing equipment and equipment for telecommunications.

Trade Names: Luranyl (BASF, FRG); Noryl (General Electric, USA); Vestoran (Hüls, FRG).

Poly(Phenylene Sulfide). Physical properties are given in Table 8 and electrical properties in Figure 38.

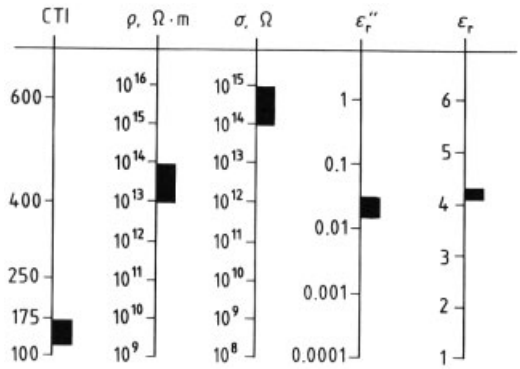

Figure 38. Relief diagram of electrical properties of poly(phenylene sulfide)

Special properties: transparent, requires special reinforcement, low glass transition temperature, high thermal loading capacity, dielectric properties essentially independent of temperature and frequency.

Uses: encapsulation of electronic components, insulating films.

Trade Names: Ryton (Phillips Petroleum Chemicals, Belgium).

Polystyrene. Physical properties are given in Table 8 and electrical properties in Figure 39 (see also → Polystyrene and Styrene Copolymers).

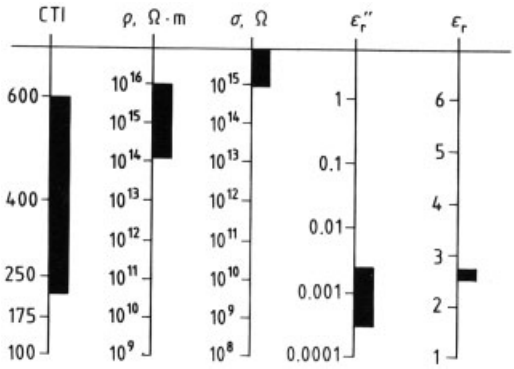

Figure 39. Relief diagram of electrical properties of polystyrene

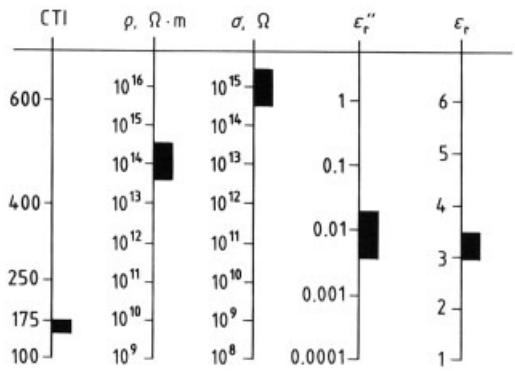

Figure 40. Relief diagram of electrical properties of polysulfone

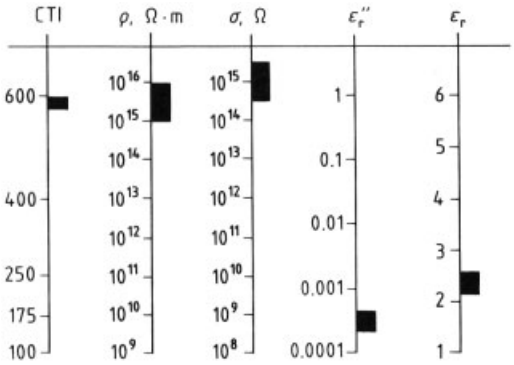

Figure 41. Relief diagram of electrical properties of polytetrafluoroethylene

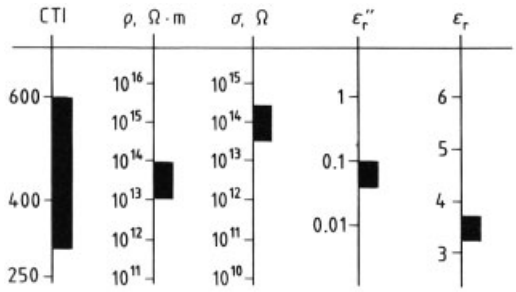

Figure 42. Relief diagram of electrical properties of poly(vinyl chloride)

Special properties: crystal clear; high gloss; brittle; resistant to humidity, salt solutions, alkali and nonoxidizing acids; mechanical properties only slightly time- and temperature-dependent; dielectric properties are independent of temperature and frequency; susceptible to damage by UV light.

Uses: encapsulation of electronic components, insulating films, dielectric for capacitors.
Trade Names: Styroflex (Norddeutsche Seekabelwerke AG, FRG); Vestyron (Hüls, FRG); Sternite (Chemical Products, UK); Dylene (Arco Chemical Co., USA); Esbrite (Sumitomo, Japan); Toporex (Mitsui Toatsu Chemicals, Japan).

Polysulfone. Physical properties are given in Table 8 and electrical properties in Figure 40 (see also → Polymers, High-Temperature).

Special properties: transparent yellow; high mechanical loading capacity; dielectric properties only slightly temperature- and frequency-dependent, resistant to acids, bases, oxidation and hydrolysis; highly incombustible.

Uses: insulation films, circuit boards.
Trade Name: Ultrason (BASF, FRG).

Polytetrafluoroethylene. Physical properties are given in Table 8 and electrical properties in Figure 41 (see also → Fluoropolymers, Organic, Chap. 2.2.).

Special properties: difficult to wet; antiadhesive; favorable coefficient of friction; noncombustible; weather-resistant; physiologically inert; extremely resistant to corrosion, solvents, acids, and bases; dielectric properties independent of temperature and frequency.

Uses: films, insulation, insulating and sleeve material in cables.
Trade Names: Hostaflon TF (Hoechst, FRG); Teflon (Du Pont, USA); Polyflon (Daikin Kogyo Co., Japan); Fluon (ICI, UK).

Poly(Vinyl Chloride). Physical properties are given in Table 8 and electrical properties in Figure 42 [see also → Poly(Vinyl Chloride)].

Special properties: highly incombustible, rigid, weldable, thermoformable, resistant to salt solutions, bases, most acids, and fuels.

Uses: insulation, casings, appliance components, insulating and sleeve material in cables.
Trade Names: Vestolit (Hüls, FRG); Trosiplast (Hüls-Troisdorf, FRG); Vinoflex (BASF, FRG); Duval (Alpha Chemical & Plastics Corp., USA); Corvic (ICI, UK); Nipolit (Chisso Corp., Japan).

Styrene – Acrylonitrile Copolymer. Physical properties are given in Table 8 and electrical properties in Figure 43 (see also → Polystyrene and Styrene Copolymers).

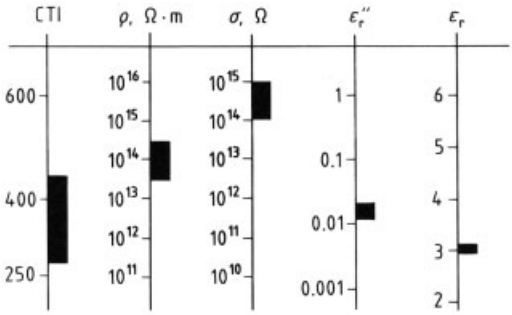

Figure 43. Relief diagram of electrical properties of styrene – acrylonitrile copolymer

Special properties: resistant to UV light, weather-resistant, good long-term stability at elevated temperature, good formability.

Uses: insulators, insulation material, ventilators, casings for electrical equipment, electrical, electronic, telecommunications, protectively insulated components.
Trade Names: Acrylafil (Fiberfil Inc., USA); Thermocomp (LNP Corp., USA); Luran (BASF, FRG); Lustran (Monsanto Co., USA); Litac (Mitsui Toatsu Chemicals Inc., Japan).

Styrene – Butadiene Copolymer. Physical properties are given in Table 8 and electrical properties in Figure 44 (see also → Polystyrene and Styrene Copolymers).

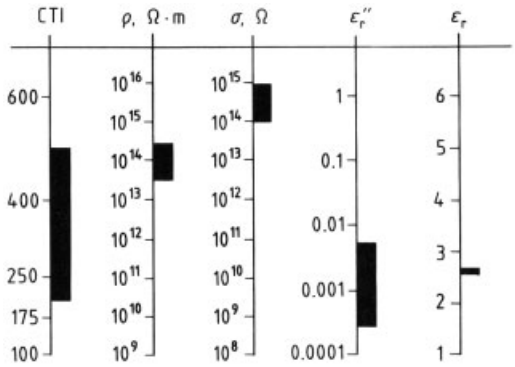

Figure 44. Relief diagram of electrical properties of styrene – butadiene copolymer

Special properties: impact toughness, long-term thermal stability, good formability, limited resistance to UV light.

Uses: casings for office equipment, electrical, electronic, telecommunications, protectively insulated components.
Trade Names: Superflex (Carl Gordon Ind., USA).

3.2.2. Thermosets

Thermosets are curable synthetic resins and polymers that undergo irreversible cross-linking and become insoluble on heating within a certain temperature range or on addition of catalysts.

The term also includes cured products which do not change significantly below their decomposition temperature. Thermosets can be modified with fillers. Physical properties are compiled in Table 9.

Epoxy Resins. Physical properties are given in Table 9 and electrical properties in Figure 45 (see also → Epoxy Resins).

Special properties: strong adhesion to metals, good flow properties, no material losses, thermally stable, light colors.

Uses: jacketing, contact substrates in electrical applications, insulation of converters, substrate material for motor armatures, collector rings in electrical motors and generators.
Trade Names: Beckopox (Hoechst, FRG); Araldit (Ciba-Geigy, Switzerland); Epodite (Showa High Polymer Co., Japan); Epocast (M

Table 9. Physical properties of thermosets

Polymer	Density, g/cm^3	Tensile strength, N/mm^2	Ultimate strain, %	Long-term thermal stability, °C
Epoxy Resins (EP)	1.2	60	4	150
Melamine – formaldehyde resins (MF)	1.5	30	1	65 – 100
Phenol – formaldehyde resins (PF)	1.5	25	1	150
Synthetic resin (chip board)	1.2	130	5	100
Polyurethane (PUR)	1.3	45	50	80
Silicone resin (SI)	1.6	90	2	180
Urea – formaldehyde resins (UF)	1.5	30	1	120
Unsaturated polyester (UP)	1.2	60	2	70

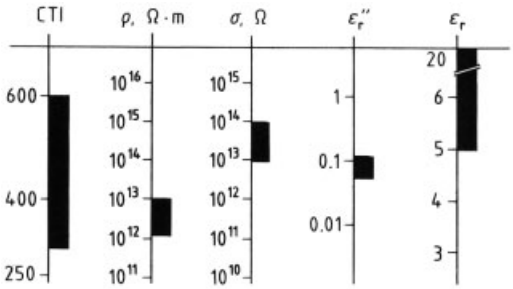

Figure 45. Relief diagram of electrical properties of epoxy resins

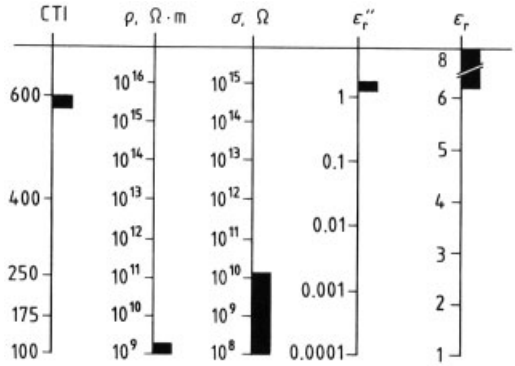

Figure 46. Relief diagram of electrical properties of melamine – formaldehyde resins

& T Chemicals Inc., USA); Epikote (Shell International Chemical Co., UK).

Melamine – Formaldehyde Resins. Physical properties are given in Table 9 and electrical properties in Figure 46 (see also → Amino Resins).

Special properties: flame resistant; high mechanical loading capacity; resistant to fuels, oils, solvents, alcohols; low stability to acids and bases; prone to stress cracking; can be dyed with light colors.

Uses: electrical installations, electrical components, fuse elements.

Trade Names: Ultrapas (Hüls-Troisdorf, FRG); Tufnol (Tufnol Ltd., UK); Textolite (General Electric, USA); Synthane (Synthane-Taylor Corp., USA); Nikamelamine (Nippon Carbide Ind., Japan).

Phenol – Formaldehyde Resins. Physical properties are given in Table 9 and electrical properties in Figure 47 (see also → Amino Resins).

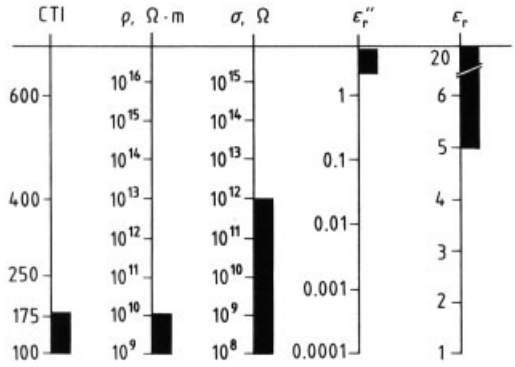

Figure 47. Relief diagram of electrical properties of phenol – formaldehyde resins

Special properties: insensitive to heat and humidity, thermally stable, dimensionally stable, only dark colors.

Uses: insulating material, casings for electrical and electronic applications.

Trade Names: Acell (BP Chemicals, UK); Alveophen (Synthetic Resins Ltd., UK); Bakelite (Bakelite, FRG); Compolet (Nobel In-

dustries Sweden, Sweden); Dynapor (Hüls-Troisdorf, FRG).

Synthetic Resin – Chip Board. Physical properties are given in Table 9 and electrical properties in Figure 48.

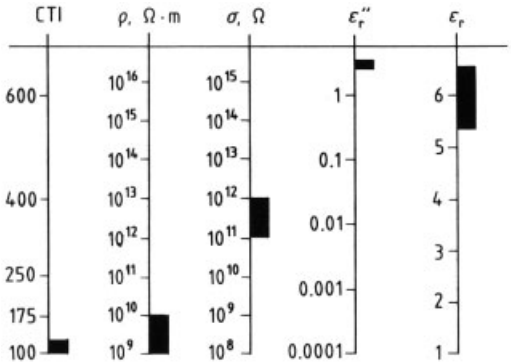

Figure 48. Relief diagram of electrical properties of synthetic resin – chip board

Special properties: properties are mostly those of the impregnating agent.

Uses: high-voltage insulation, slot liners.

Trade Names: Delignit (Blomberger Holzindustrie, FRG); Pag-Holz (Pag-Preßwerk, FRG).

Polyurethane. Physical properties are given in Table 9 and electrical properties in Figure 49 (see also → Polyurethanes).

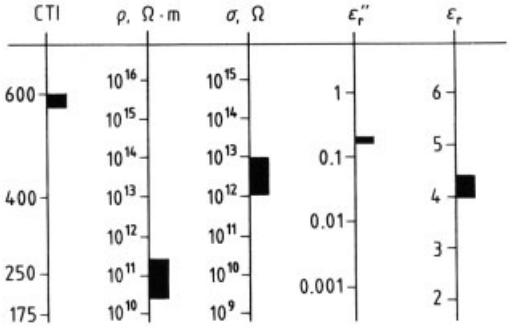

Figure 49. Relief diagram of electrical properties of polyurethane

Special properties: constant mechanical behavior over a wide temperature range; resistant to surface and sea water; resistant to biological contamination; limited resistance to hydrolysis by steam, hot water, acids and bases; physiologically inert when completely cross-linked; castable; curing takes place even at low temperature.

Uses: cable end caps, converters, transformers, ignition coils.

Trade Names: Elastan (Elastogran, FRG); Desmodur (Bayer, FRG); Suprasec (Polyurethanes Group ICI, UK).

Silicone Resin. Physical properties are given in Table 9 and electrical properties in Figure 50 (see also → Silicones).

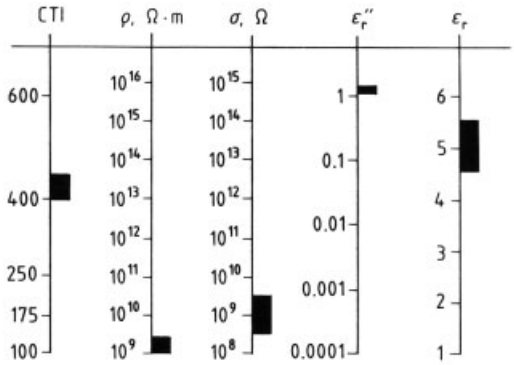

Figure 50. Relief diagram of electrical properties of silicone resin

Special properties: chemically inert, physiologically inert, water-repellent, physical and electrical properties remain almost unchanged from very low to high temperatures.

Uses: thermally stable impregnating resins.

Trade Names: Baysilon (Bayer, FRG); Wacker Silicone (Wacker Chemie, FRG); GE-Silicone (General Electric, USA); DC-Silicone (Dow Corning Corp., USA).

Urea – Formaldehyde Resins. Physical properties are given in Table 9 and electrical properties in Figure 51 (see also → Amino Resins).

Special properties: stable to light, resistant to solvents and oils, but not to acids and bases, light colors possible.

Uses: electrical installations.

Trade Names: Pollopas (Hüls-Troisdorf, FRG); Beetle, Scarab (BIP Chemicals, UK); Urochem (Chemiplastica S.p.A., Italy).

Unsaturated Polyester. Physical properties are given in Table 9 and electrical properties in Figure 52 (see also → Polyester Resins, Unsaturated).

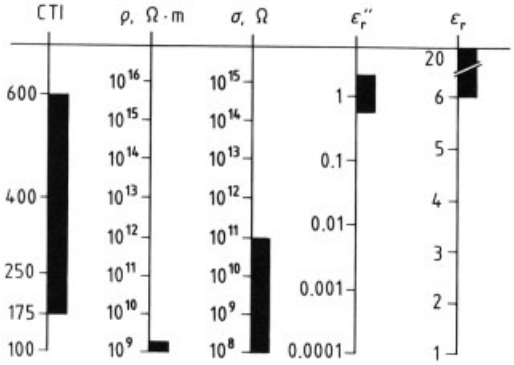

Figure 51. Relief diagram of electrical properties of urea–formaldehyde resins

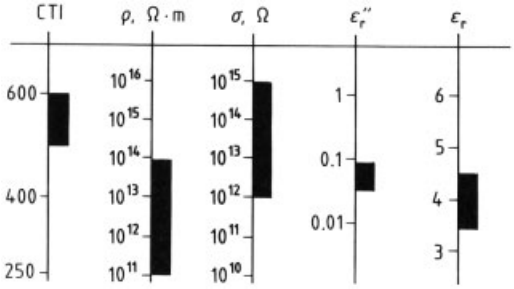

Figure 52. Relief diagram of electrical properties of unsaturated polyester

Special properties: meltable, can be cast when liquid, no material losses, dimensionally stable, not prone to stress cracking, stable to light, easy to process.

Uses: mounting electrical components, also for high-voltage applications.

Trade Names: Palatal (BASF, FRG); Vestopal (Hüls, FRG); Alpolit (Hoechst, FRG); Crystic (Scott Bader Co., UK); Glastic (Glastic Corp., USA); U-Pica (Japan-Upica, Japan).

3.2.3. Elastomers

Elastomers are natural or synthetic polymers (→ Rubber, 2. Natural; → Rubber 3., Synthetic) which exhibit rubberlike behavior between at least 20 °C and their respective decomposition temperature. They solidify at or below 0 °C. The rubber – elastic behavior is a result of the tangled arrangement of the molecular chains or of their degree of cross-linking. Physical properties of elastomers are compiled in Table 10.

Chloroprene Rubber. Physical properties are given in Table 10 and electrical properties in Figure 53.

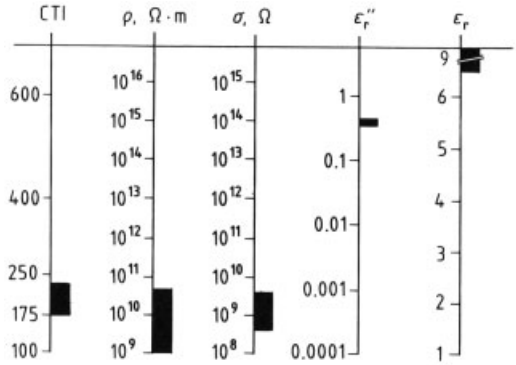

Figure 53. Relief diagram of electrical properties of chloroprene rubber

Special properties: highly incombustible, leatherlike toughness, weather-resistant, corrosion-resistant.

Uses: insulating and sleeve material for cables.

Trade Names: Baypren (Bayer, FRG); Butaclor (Rhône-Poulenc Specialities Chimiques, France); Neoprene (Du Pont, USA).

Ethylene – Propylene Terpolymer. Physical properties are given in Table 10 and electrical properties in Figure 54.

Special properties: rubber-elastic, abrasion resistant, stable towards aging and weathering (oxygen, ozone, water, UV radiation, high temperature), inert towards many chemicals.

Uses: insulation and sleeve material for cables, plugs, cable filling, with carbon black causes deflection of the lines of force in conductors, transformers, etc.

Table 10. Physical properties of elastomers

Polymer	Density, g/cm^3	Tensile strength, N/mm^2	Ultimate strain, %	Long-term thermal stability, °C
Chloroprene rubber (CR)	1.3 – 1.7	55	250	60 – 90
Ethylene – propylene – terpolymer (EPM – EPDM)	0.9	18	380	140
Ethylene – propylene rubber (EPR)	1.3	4	200	90
Ethylene – vinyl acetate rubber (EVA)	1.4	7	200	120
Styrene – butadiene rubber (SBR)	0.94	5 – 25	500	70
Silicone rubber (SI-R)	1.3	4 – 9	300 – 700	200

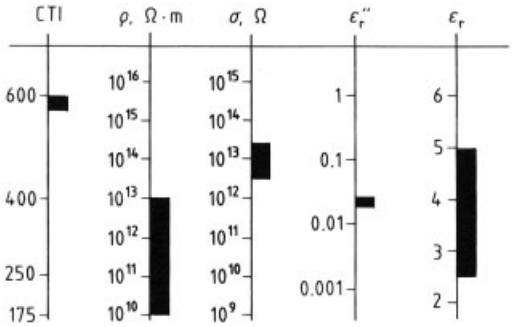

Figure 54. Relief diagram of electrical properties of ethylene – propylene terpolymer

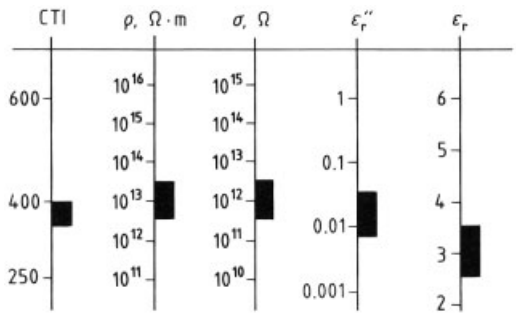

Figure 55. Relief diagram of electrical properties of ethylene – propylene rubber

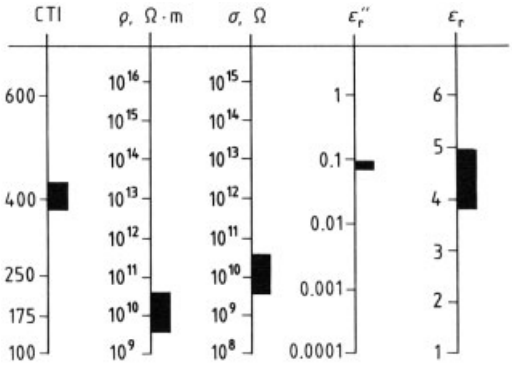

Figure 56. Relief diagram of electrical properties of ethylene – vinyl acetate rubber

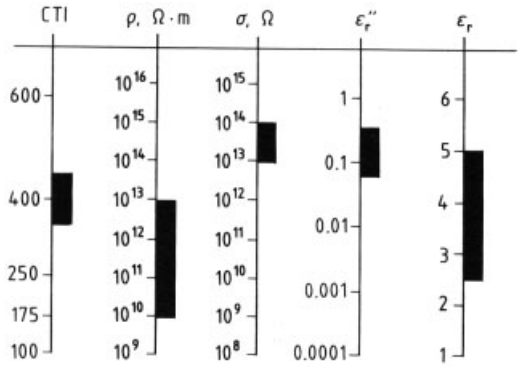

Figure 57. Relief diagram of electrical properties of styrene – butadiene rubber

Trade Names: Buna AP (Hüls, FRG); Vistalon (Exxon Chemical Co., USA).

Ethylene – Propylene Rubber. Physical properties are given in Table 10 and electrical properties in Figure 55.

Special properties: resistant to chemicals, ozone, aging, water, acids, and alkalis; weather resistant; unstable to mineral oils and organic solvents.

Uses: insulating and sleeve material for cables.

Trade Names: Dutral TP (Montepolmeri, Italy); Intolan (Dow Corning Corp., USA); Vistalon (Exxon Chemical Co., USA).

Ethylene – Vinyl Acetate Rubber. Physical properties are given in Table 10 and electrical properties in Figure 56. Special properties: ther-

mally stable, excellent adhesion, rubber-elastic, impact resistant.

Uses: insulating and sleeve material for cables.

Trade Name: Levapren (Bayer, FRG).

Styrene – Butadiene Rubber. Physical properties are given in Table 10 and electrical properties in Figure 57.

Special properties: resistant to organic solvents, weldable, limited stability fuels, oils, acids, and alkalis.

Uses: insulating and sleeve material for cables, capacitor components, accumulator boxes components for telecommunications, connectors, insulators.

Trade Name: Buna-Hüls (Hüls, FRG).

Silicone Rubber. Physical properties are given in Table 10 and electrical properties in Figure 58.

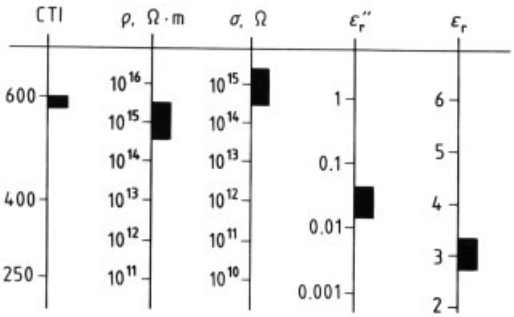

Figure 58. Relief diagram of electrical properties of silicone rubber

Special properties: difficult to wet, physiologically safe, electrical characteristics show little change over large regions.

Uses: insulating and sleeve material for cables, electric insulation, casting resins.

Trade Names: Silastic (Dow Corning Corp., USA); Silopren (Bayer, FRG); Wacker-Silicongummi (Wacker-Chemie, FRG).

3.2.4. Polymer Processing

Once polymeric materials have been synthesized and, if necessary, treated with fillers, pigments, plasticizers, processing aids, and agents to minimize aging and flammability, they are processed into powders or granulates. Subsequent forming operations result in products that frequently do not require further physical modification. The forming procedures most applicable to single items or small numbers of insulators include reshaping and bonding, for example by heat sealing. Thermoplastic materials can also be machined, including boring and cutting. A molded polymeric insulator, fashioned to the proper shape in a single operation without recourse to subsequent machining, may often replace conventional insulators consisting of multiple parts, each requiring its own series of machining operations. However, the equipment required to produce complicated polymeric objects are usually quite expensive, so economic factors tend to limit this approach to mass-produced items. For a more detailed description of the techniques described below see → Plastics, Processing.

Casting without the Use of Pressure. Certain polymeric materials can be formed without the use of pressure by methods such as the rotational casting of hollow objects in heated forms and centrifugal casting. Such materials include thermoplastic polymers such as poly(methyl methacrylate), polyamide 6, poly(vinyl chloride) mixed with liquid plasticizers, certain elastomers, and thermosetting polymers (casting resins, e.g., epoxy resins and polyurethanes).

Powder Techniques. The powder method entails fusing a low-melting polymer powder to a heated metal surface, for example by whirl sintering. This approach lends itself to materials such as polyolefins, cellulose acetobutyrate, and casting resins.

Low-Pressure Casting. Low-pressure casting is appropriate for polymers reinforced with glass or carbon fibers. The precursors are resin mats and prepregs made from unsaturated polyesters, epoxides, and other casting resins. These are converted to their final form either by the hand lay-up or the plane-type press process. A mixture of resin and fibers can also be sprayed at low pressure.

Casting under Thermoplastic Conditions. The thermoplastic casting method plays an important role in polymer processing. It requires a temperature in the range of 150 – 250 °C and a pressure of several hundred bar.

Calendering denotes the process of rolling a polymeric mass in the plastic state until it reaches a desired thickness, usually 0.06–0.6 mm.

Extruding. The most versatile of the processes used for forming thermoplastic materials is extrusion with a continuous screw extruder. Proper choice of dies permits the manufacture of films, pipes, tubes, sheets, slabs, and even items of more complex cross-sectional shape. Screw extruders are often coupled with calibration devices, smoothing rolls, and cooling facilities. With this approach it is possible to completely surround a metallic wire, conductor, or cable with a sprayed plastic insulator.

Pressing. Thermosetting laminates (e.g., laminated paper, glass–fabric laminates, cotton laminates, melamine–formaldehyde laminates, or phenolformaldehyde–fiberglass fabrics) are prepared with the aid of a multiplaten press in which the carrier sheet is continuously soaked with a resin solution. Suitably sized pieces of the material are then stacked in layers and warmed under pressure in a heated press, causing the resin to flow and subsequently solidify. Provided the material is cooled while still under pressure, such a press can also be used to produce thermoplastic blocks of any desired thickness with a high-quality surface finish.

Injection Molding. Injection molding machines operate on the batch principle. A heated injection cylinder operates in tandem with a completely closed forming die, the two units being separated spatially. Plastic material is introduced in portions into the die through a nozzle. Once the substance has hardened, the die is opened and the molded object ejected. The injection molding process is also applicable to thermosetting polymers.

Compression molding requires continuous heating of the forming die. Appropriate amounts of a preheated thermosetting material are first introduced into an open die. Once the die is closed, heat and pressure cause the polymer to flow and expand, entirely filling the available space. After curing is complete, the die is opened and the finished product is ejected while still hot.

Transfer Molding. In transfer molding the polymer is warmed to a plastic state in a separate compartment prior to being forced into a forming die.

Reshaping of a semifinished thermoplastic polymer into a thermoelastic state requires relatively little applied force. Cut sheets or blocks of material are brought to working temperature with the aid of infrared radiation and then drawn into monolithic forming dies where they acquire the desired shape. Finished parts are ejected after cooling.

Heat sealing complements the technique of molding. Small numbers of complex objects can be economically assembled from components by heat-sealing either with the aid of warm air or by techniques based on high-frequency radiation, heat impulses, or ultrasonics.

Soaking is especially suitable for coating electronic products with casting resins. However, it may also be utilized with insulating liquids and certain thermoplastic materials, where the latter are either preheated or applied as low-viscosity solutions. The purpose of soaking is to reduce the risk of short circuiting, for example in the windings of a current-carrying metallic conductor, by ensuring that all gaps are filled. Improved heat transfer is achieved simultaneously (the specific thermal conductivity of air is $0.026\,\mathrm{W\,m^{-1}\,K^{-1}}$, while that of synthetic resin filler is $0.6\,\mathrm{W\,m^{-1}\,K^{-1}}$). The most common soaking resins are solvent-free unsaturated polyester resins, which may be cured by activation with a setting agent. Resins generally have a higher filling power than impregnating varnishes.

Impregnation provides almost total protection against undesired environmental influences. It also confers increased mechanical stability and protection against the adverse effects of motion. Impregnation itself is normally carried out under vacuum. Most impregnating agents are liquid in the working state, but they can be transformed into a thermosetting state by curing. The method is also applicable to solvent-free trickling resins prepared from unsaturated polyesters. These materials break into droplets in the course of winding and revert to a gel-like state with

subsequent electrical conduction warming of the conductor. Rapid heating to the curing temperature induces such an impregnating agent finally to harden.

Varnishing. Metallic electrical conductors can also be electrically insulated with wire varnish. This method is often employed in transformers, motors, generators, relay coils, and similar devices. The applied polymer (usually a polyesterimide, polyurethane, polyamideimide, or polyhydantoin) is dissolved in an appropriate solvent, e.g., aromatic and aliphatic hydrocarbons, chlorinated hydrocarbons, aqueous alcohols. The conductor is passed several times through a varnish bath, and becomes uniformly coated. Solvent is evaporated in a radiation oven, permitting deposition and annealing of a multilayered coating graded according to hardness. Thermoplastic polycarbonate films can be obtained from solution in a fashion similar to varnish films, whereby the film is extracted from a drum after evaporation of solvent.

A more environmentally acceptable coating technique utilizes either a melt varnish of the polyesterimide type or a water-soluble wire varnish.

Embedding. Solvent-free, soaking, trickling, and casting resins are used for embedding components in electrotechnics. They protect costly components against environmental effects, provide mechanical stabilization, and prevent the formation of dangerous contacts. Reaction resins, monomers, or prepolymers are generally used that harden in the presence of suitable reactants. Most important are methyl methacrylate, unsaturated polyesters, epoxy resins, polyurethanes, modified phenolic resins, and furan resins. The cross-linkable hydrocarbons and methyl methacrylate casting resins have favorable dielectric values at high frequencies (Sections 2.6) and are therefore important for microwave technology (television, radar, dielectric warming).

3.3. Liquid Insulating Materials

In transformers, starter motors, and other electrical equipment that generate a considerable amount of heat, liquids are employed both as insulators and coolants. In cables, capacitors, and connections, insulation liquids are used as a functional component of the dielectric. In relays, they function as insulators, coolants, and extinguishers for arcing. Most liquid insulators consist of mineral oils or synthetic liquids.

Mineral oils are complex hydrocarbon mixtures that are obtained from petroleum oil by fractional distillation and refining. They combine important properties including high breakdown strength, good impregnation characteristics, and the ability to conduct large amounts of heat via convection. Mineral oils can also bind any gas produced during arcing with pronounced evaporation and decomposition. Depending upon the predominant hydrocarbon, the oils can be classified as *naphthene-* or *paraffin-based oils;* however, the transition between these groups is continuous–the *mixed petroleum oils.* The naphthene-based oils have a higher density, and a lower solidification point than the paraffin oils; their viscosity is also more temperature dependent. Since they are also more stable at low temperatures, they are preferred for devices used in the open air. They are also preferable for cable applications because paraffin oils are more prone to decomposition at high field strengths with formation of hydrogen and wax products. Decomposition can lead to partial discharges and gas inclusions, which may interact resulting in breakdown of the cable insulation (see section 2.7). Naphthene oils are also suitable for cable impregnation systems comprising oils and refined resins (e.g., collophonium) because the refined resins are incompatible with paraffin oils [47].

Insulation oils for transformers, converters, and relays are subject to standard regulations, e.g., IEC Publication 296, which only specifies minimum values for the breakdown voltage [48].

Since the electric strength of oils depends strongly on the measuring technique used (Section 2.7), breakdown behavior is characterized by the breakdown voltage determined under standardized conditions (specified electrode configuration, electrode separation, and test procedure, IEC Publication 156 [32]).

When insulation liquids simultaneously act as a coolant, they must have a low viscosity. They should be able to withstand cooling to $-30\,^{\circ}C$ without thickening. Furthermore, the oxidation and aging behavior of new oil and used oil should

be monitored during operation when the temperature is elevated and atmospheric oxygen has access to the equipment.

The viscosity of cable insulation oils varies within wide limits, ranging from low viscosity oils for oil cables to highly viscous impregnation systems for earthing and pressure cables. Furthermore, extremely viscous cable immersion systems can be prepared by addition of thickeners (natural resins, polymers).

The most important property of cable immersion systems is the dielectric loss index (see Section 2.6) which should be as low as possible because of the danger of thermal breakdown of the cable insulation at higher field strengths. The electrical strength of the impregnation system is less important because the high electrical strength of the cable insulation is a result of the combination of the multilayer insulation film (paper or polymeric) and impregnating agent; it depends upon the breakdown behavior of both components.

The viscosity and electrical properties of mineral oils used for high-tension capacitors (phase displacers) are similar to those of the cable insulation oils.

Synthetic Liquids. The drawbacks of mineral insulation oils (oxidizability, flammability, and aging with formation of detrimental decomposition products and deterioration of electrical properties) led to the development of synthetic insulation liquids with improved properties. A general advantage of synthetic insulation liquids is the fact that they can be reproducibly manufactured with identical structures and compositions [49]. They include chlorinated hydrocarbons, the most common being aromatic compounds such as chlorinated biphenols, chlorinated benzenes, and their mixtures. These noncombustible materials are known as *Ascarels.* Damage to health and the environment can result from careless handling and use of such toxic agents. Consequently, sale of Ascarel-containing products is prohibited in many countries. Closed systems (transformers, capacitors, etc.) have been exempted from these regulations.

During decomposition in electric arcs, Ascarels do not emit explosive or combustible gas mixtures provided that they contain equivalent molar amounts of chlorine and hydrogen; hydrogen chloride is then produced. The Ascarels are both chemically and thermally stable. They possess dipole character; their dielectric dissipation factor is therefore frequency- and temperature-dependent and they have a high relative permittivity ($\varepsilon_r \approx 4-6$ at 50 Hz).

Liquid fluorine compounds show excellent properties. They are noncombustible, prevent arcing, and are more thermally stable than Ascarels. They have a lower boiling point and are highly volatile. Because of their low viscosity, heat transfer properties are favorable. They have a high breakdown voltage, high volume resistivity (Section 2.2), and an essentially frequencyindependent dielectric dissipation factor.

Liquid silicones are polymeric organosilicon compounds with a highly variable degree of polymerization. Commercial silicone oils have differing viscosities, which are only slightly temperature-dependent. They combine excellent resistance to low temperatures (setting point $-5\,°C$ to $-80\,°C$) with outstanding high temperature stability ($240\,°C$) and are not prone to oxidation. They remain noncombustible up to operating temperatures of $300\,°C$. Silicone oils have a high breakdown strength, low dielectric dissipation factor ($\leq 1\times10^{-4}$), low relative permittivity (2.5 – 2.9), and an average volume resistivity ($10^{12}\,\Omega m$). The low temperature dependence of the dielectric dissipation factor (3×10^{-4}) is remarkable.

Other Liquid Insulators. Apart from the Ascarels and liquid silicones, further synthetic insulation fluids find application mostly in transformers, capacitors, and cables. The most important commercially available examples include compounds based on alkyl benzenes, polybutene, diarylalkanes, esters, alkyl diphenyl ethers, and high molecular mass paraffin hydrocarbons. These liquids match the excellent electrical properties of the Ascarels but not their flame retardation. They differ according to application regarding relative permittivity, dielectric dissipation factor, ignition temperature, viscosity, and the behavior toward low temperatures.

4. Selection of an Insulating Material

Within its technical committee 15 for "Insulating Materials", the International Electrotechnical Commission (IEC) has installed two subcommittees: subcommittee 15 A for "Short

Term Tests" and subcommittee 15 B for "Endurance Tests". Electrical testing methods for insulation materials are developed in the working groups of these subcommittees. The recommendations of national electrotechnical commissions serve as the basis for their development.

The intention and objective of the IEC is that the international standards drawn up by numerous experts from the member states should be adopted as the national standards. All published results should then be comparable.

The values of electrical properties given in this publication have been extracted from numerous sources; some have been supplemented and amended using the results of our own investigations. In the available literature, detailed information pertaining to testing conditions has often been omitted, consequently the data plots provide nothing more than rough guides. This can only be improved by testing identical samples of the huge range of insulating materials used in the electrotechnics industry in one and the same laboratory, under identical testing conditions that strictly comply with the international regulations. Although this undertaking would be highly desirable, it is almost impossible.

To solve a specific insulation problem, materials can be preselected from the relief diagrams. The manufacturers must then be consulted, or reference to technical data sheets can be made. Measurements or the results of an independent testing institute ensure that the material is able to meet the requirements.

What makes an outstanding material for electrical applications? The answer to this question depends upon the intended use, e.g., whether it involves low-voltage or high-tension supplies, d.c., or low-, mid-, or high-frequency sources. For applications with low-voltage supplies, the *tracking parameter* should be at least CTI 400, i.e., group II according to IEC Publication 664 [24], group I includes materials with CTI 600. Similarly, the *volume resistivity* should be greater than $10^{12}\ \Omega \cdot \mathrm{m}$, the surface resistivity at 23 °C and 50 % R.H. should be at least $10^{14}\ \Omega$.

At mains frequency, no particular demands are placed on the *dielectric dissipation factor* or the *relative permittivity* except for the application as a dielectric in capacitors. If, for example, the power loss per cubic centimeter should not exceed 0.1 W at 50 Hz and the field strength should be no greater than 1000 V/cm (see Section 2.6), then a loss index of $\varepsilon_r'' \approx 3600$ can be tolerated. An increase in the field strength to 10 000 V/cm causes a decrease in the threshold value for the loss index by a factor of 100. In the same way the loss index threshold value decreases linearly with an increase in frequency.

The magnitude of the *relative permittivity* is particularly important in determining the intended application of a material. To avoid dielectric warming, its value should be low. A lower relative permittivity is also associated with a higher breakdown voltage in air. (As a general rule, the maximum attainable breakdown voltage is achieved when the relative permittivity of the material is identical with that of the surrounding medium). For applications in capacitors, the relative permittivity should be as large as possible to minimize the geometric dimensions.

The *electrical strength* is primarily intended to allow comparison between materials. However, this is only the case when all parameters, including sample thickness, are absolutely identical. The significance of the electrical strength is thus disputable (Section 2.7). It cannot be used directly for design purposes. Test models are imperative, studies on the effect of loading time are often required.

A "single value", i.e., result at a discrete temperature and frequency does not characterize the behavior of a material under application conditions. The influences of parameters such as temperature, frequency, loading time, age, or humidity cannot be discerned from this single piece of information. A knowledge of the interdependence of the various parameters provides far more information than a single data value. In many cases, the development of technical polymers and their successful application in the electrical industry has only been possible because the complete characteristics of the material had been previously compiled and evaluated. The application range must be considered as a function of temperature, frequency, voltage, and environmental conditions. Other influences including manufacturing processes, mechanical loading, and the effects of fillers and additives are better reflected by the functions of the electrical parameters rather than by individual results at a single temperature or frequency. Automatic measurement processes have made the practical

determination of such relationships relatively inexpensive.

The following diagrams represent examples for characterizing a technical insulation material–a thermoplast with typical properties: high flame resistance, rigid-elastic, high impact toughness above 90 °C, mechanical durability at 90 °C. On the basis of the short-term electrical properties, the product quality lies within the following limits:

Tracking	CTI 250 – CTI 500
Volume resistivity	$(0.5-5)\times 10^{16}\ \Omega\cdot\mathrm{cm}$
Surface resistivity	$(0.5-5)\times 10^{15}\ \Omega$
Loss index	$(0.3-3)\times 10^{-2}$
Relative permittivity	2.9 – 3.2
Electrical strength	25 – 35 kV/mm

Figure 59 illustrates the dielectric dissipation factor (Section 2.5) and the relative permittivity (Section 2.4) as a function of temperature at frequencies of 50 Hz, 1 kHz, and 1 MHz. At temperatures below the long-term thermal stability threshold (90 °C) the curves remain relatively constant and are typical of a medium-quality material. In the vicinity of − 50 °C, a relaxation phenomenon occurs, which is indicated by the decrease in the dielectric dissipation factor at low frequencies. Above 90 °C, the dielectric dissipation factor increases sharply; the effects of conductivity losses are apparent.

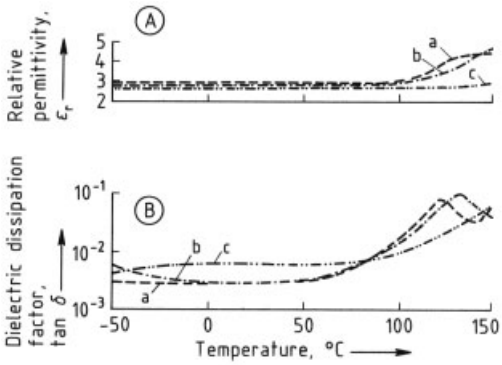

Figure 59. Relative permittivity (A) and dielectric dissipation factor (B) of a thermoplast as functions of temperature a) 50 Hz; b) 1 kHz; c) 1 MHz

Figure 60 illustrates both dielectric parameters as functions of frequency. The decrease in the relative permittivity with increasing frequency is low. The relative permittivity must decrease with increasing frequency because it assumes a value equal to the square of the refractive index at the optical relative permittivity. The dielectric dissipation factor reaches a broad maximum at 10 MHz.

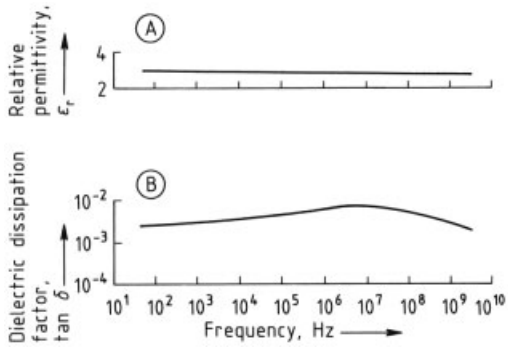

Figure 60. Relative permittivity (A) and dielectric dissipation factor (B) of a thermoplastic polymer blend as functions of frequency at 23 °C

Figure 61 shows the plot of volume resistivity (Section 2.2) as a function of temperature. The decrease in the volume resistivity by a factor of five powers of ten is considerable and illustrates that the number of mobile charge carriers strongly increases with increasing temperature.

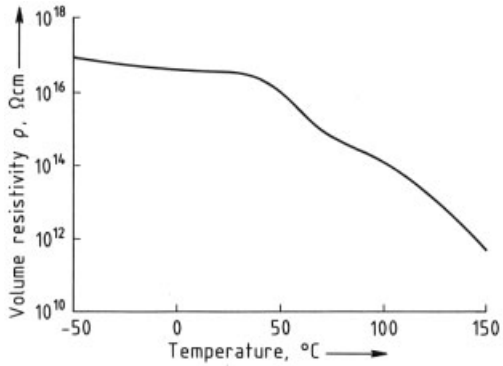

Figure 61. Volume resistivity of a thermoplastic polymer blend as a function of temperature

The surface of the sample is hygroscopic. This is shown in Figure 62 by the decrease in surface resistivity with increasing air humidity (Section 2.3).

Because of the large deviation in measured values, the variation in the measurement of surface resistivity as a function of temperature (see Section 2.3) has been included in Figure 63. A pronounced decrease in resistance surprisingly

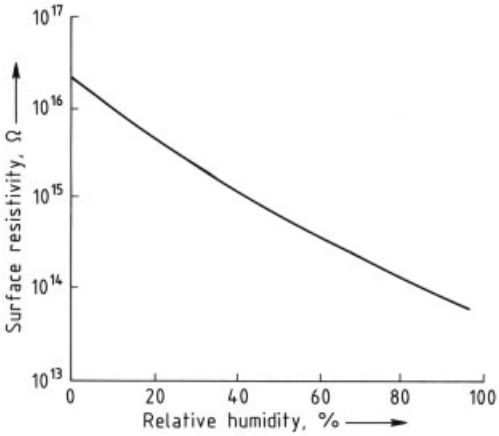

Figure 62. Surface resistivity of a thermoplastic polymer blend as a function of atmospheric humidity

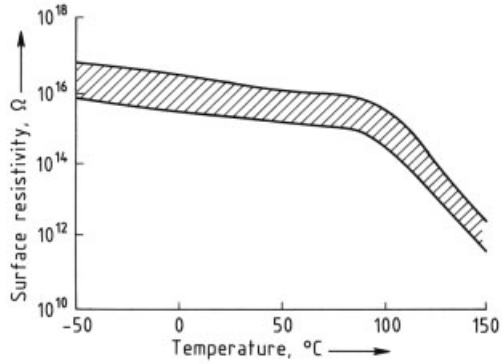

Figure 63. Surface resistivity of a thermoplastic polymer blend as a function of temperature

appears at approximately 90 °C but the reason for this is not known.

The electric strength is shown as a function of temperature in Figure 64. A pronounced decrease in the breakdown voltage occurs between 90 and 100 °C. This represents the limitations in thermal stability as determined using nonelectrical procedures (see also Section 2.7).

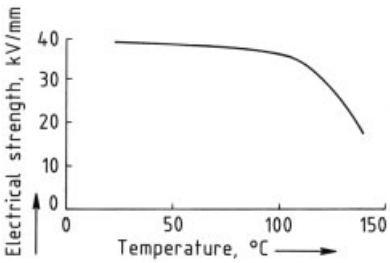

Figure 64. Electrical strength of a thermoplastic polymer blend as a function of temperature

On the basis of the above properties, the insulating material would be suitable for applications in the low- and middle-voltage range in a normal climate (e.g., 20 °C, 55 % R.H.), and at a working temperature up to 70 °C. Small distances between conductors subjected to potential differences are possible. The material is also qualified for integrated circuits and for high frequencies. It is highly suited for applications in audio, video, and telecommunication equipment.

5. References

General References

1. J. B. Birksetal.(eds.): *Progress in Dielectrics,* **vols. 1 – 7,** Heywood, London 1959 – 1967.
2. S. Flügge (ed.): *Handbuch der Physik,* Springer Verlag, Berlin Göttingen Heidelberg, vol. 17, Dielektrika, 1956; vols. 19, 20, Elektrische Leitungsphänomene, 1958, 1959.
3. S. Flügge,F. Trendelenburg:*Ergebnisse der exakten Naturwissenschaft,* vol. 25,1951; vol. 27, 1953;vol. 31, 1959, Springer Verlag, Berlin .
4. H. Fröhlich: *Theory of Dielectrics,* Oxford University Press, Oxford 1987.
5. P. M. Pflier, H. Jahn, G. Jentsch: *Elektrische Meßgeräte und Meßverfahren,* Springer Verlag, Berlin 1978.
6. K. Küpfmüller: *Einführung in die theoretische Elektrotechnik,* Springer Verlag Berlin, Göttingen, Heidelberg 1984.
7. A. von Hippel: *Dielectric Materials and Applications,* M.I.T. Press, Cambridge, USA, 1954.
8. A. von Hippel: *Molecular Science, Molecular Engineering,* Chapman and Hall, London 1959.
9. C. P. Smith: *Dielectric Behaviour and Structure,* McGraw-Hill, New York 1955.
10. K. Potthoff, W. Widmann: *Meßtechnik der hohen Wechselspannungen,* Vieweg & Sohn, Braunschweig – Wiesbaden 1965.
11. H. Bilke: *Handbuch der Keramik,* Z. Vlg. Schmid, Freiburg i. Brg.

12. *Engineering Property Data on Selected Ceramics:* **vol. I, Nitrides (1976); vol. II, Carbides (1979); vol. III, Single oxides (1981);** Metals and Ceramics Information Center, MCIC Report HB-07, Battelle Columbus Laboratories, Columbus, Ohio.
13. W. Espe: *Werkstoffkunde der Hochvakuumtechnik,* **vol. II: Silikatwerkstoffe,** VEB Dtsch. Vlg. d. Wissensch., Berlin 1960.
14. P. Guillery: *Werkstoffkunde für Elektroingenieure,* Vieweg, Braunschweig 1971.
15. T. Haase: *Keramik,* VEB Dtsch. Vlg. f. Grundstoffind., Leipzig 1970.
16. A. Hecht: *Elektrokeramik.* Springer, Berlin – Heidelberg – New York 1976.
17. R. Racho et al.: *Werkstoffe der Elektrotechnik,* VEB Vlg. Technik, Berlin 1968.
18. H.-W. Rotter: *Glimmer und Glimmererzeugnisse, Eigenschaften, Entwicklungen, Anwendungen,* Siemens AG, Berlin – München 1985.
19. E. Ryschkèwitch: *Oxide Ceramics,* Academic Press, New York – London 1960.
20. H. Stäger: *Werkstoffkunde der elektrotechnischen Isolierstoffe,* Bornträger, Berlin 1955.
21. H. Salmang, H. Scholze: *Die physikalischen und chemischen Grundlagen der Keramik,* Springer, Berlin – Heidelberg – New York 1968.

Specific References

22. A. Kamm, *Kunststoffe* **75** (1985) no. 12, 900 – 903.
23. IEC Publication 112 – 1979, Bureau Central de la Commission Electrotechnique International, Genève.
24. IEC Publication 664/664 A – 1981, Bureau Central de la Commission Electrotechnique International, Genève.
25. ANSI/ASTM 3638 – 1977, ASTM Philadelphia, Pa., USA.
26. ANSI/UL 764 A – 1984, UL Publications Stock, Northbrook, Ill., USA.
27. IEC Publication 93 – 1980, Bureau Central de la Commission Electrotechnique International, Genève.
28. IEC Publication 247 – 1978, Bureau Central de la Commission Electrotechnique International, Genève.
29. IEC Publication 345 – 1971, Bureau Central de la Commission Electrotechnique International, Genève.
30. IEC Publication 250 – 1969, Bureau Central de la Commission Electrotechnique International, Genève.
31. IEC Publication 243 – 1967, Bureau Central de la Commission Electrotechnique International, Genève.
32. IEC Publication 156 – 1963, Bureau Central de la Commission Electrotechnique International, Genève.
33. IEC Publication 587 – 1984, Bureau Central de la Commission Electrotechnique International, Genève.
34. ANSI/ASTM-D 495, ASTM Philadelphia, Pa., USA.
35. DIN VDE 0303, 1989.
36. VDE 0303, part 8, Beuth Verlag, Berlin 1975.
37. DIN VDE 0303, 1989.
38. IEC Publication 426 – 1973. Bureau Central de la Commission Electrotechnique International, Genève.
39. IEC Publication 167 – 1964, Bureau Central de la Commission Electrotechnique International, Genève.
40. IEC Publication 343 – 1970, Bureau Central de la Commission Electrotechnique International, Genève.
41. IEC Publication 672-2 – 1980, Part 1 Definition and Classification, Part 2 Methods of Test, Part 3 Individual Materials, Bureau Central de la Commission Electrotechnique International, Genève.
42. K. H. Schüller: "Hochfeste Porzellane auf Quarz- und Cristobalitbasis," *Ber. Dtsch. Keram. Ges.* **44** (1967) 212 – 223, 284 – 293, 387 – 391.
43. V. Frith, R. O. Heckroodt, K. H. Schüller: "Characterization of Ceramic High-Voltage Insulator Materials," *Ber. Dtsch. Keram. Ges.* **64** (1987) 216 – 219.
44. H. Kromer, D. Rose, K. H. Schüller: "Investigations on Commercial Talcs and Steatites, IV formation of Mineral Phase During Firing," *Sprechsaal* **121** (1988) no. 8, 1 – 8.
45. H. Schmidt, K. H. Schüller: "Untersuchungen an Sondersteatiten mit Kalk," *Silikattechnik* **39** (1988) 137 – 138.
46. W. Büssem, C. Schusterius, K. Stuckard: "über die Konstitution des Steatis," *Wiss. Veroeff. Siemens-Werken* **17** (1938) 59 – 89.
47. A. C. M. Wilson: *Insulated Liquids,* The Institute of Electric Engineers, London 1980.
48. IEC Publication 296 – 1969, Bureau Central de la Commission Electrotechnique International, Genève.

49. VDEW: *ölbuch,* Versorgungs- und Wirtschaftsgesellschaft der Elektrizitätswerke mbH, Frankfurt 1983.
50. H. Saechtling: *Kunststoff-Taschenbuch,* 23rd ed., C. Hanser Verlag, München 1986.
51. R. Vieweg et al., *Kunststoff Handbuch,* vols. I – XI, Hanser, München 1963.

Insulin → Peptide and Protein Hormones
Integrated Circuits → Imaging Technology
Integrated Circuits → Semiconductors

Interferons

TATTANAHALLI L. NAGABHUSHAN, Schering – Plough Research, Bloomfield, New Jersey 07003, United States

PAUL P. TROTTA, Schering – Plough Research, Bloomfield, New Jersey 07003, United States

1. Introduction 1
2. Classification and Origin 2
3. Biological Activity 3
3.1. Antiviral Activity 3
3.2. Anticellular Activity 4
3.3. Immunomodulatory Activity 4
4. Mechanism of Action 5
5. Physicochemical Properties 6
5.1. Primary Structure 6
5.2. Secondary and Tertiary Structure . . 7
5.3. Solution Properties 8
6. Purification 8
6.1. Natural Leukocyte IFN 8
6.2. Recombinant IFNα 9
6.3. Removal of Pyrogens 10
7. Quality Specifications 12
7.1. Purity 12
7.2. Pyrogenicity Testing 14
8. Clinical Studies 14
9. References 15

Abbreviations used in this article:

AIDS	acquired immune deficiency syndrome
cAMP	cyclic adenosine 5′-monophosphate
cGMP	cyclic guanosine 5′-monophosphate
CD	circular dichroism
CPE	cytopathic effect
DEAE	diethylaminoethyl
DNA	deoxyribonucleic acid
HPLC	high-performance liquid chromatography
IFN	interferon
IU	international antiviral units
MHC	major histocompatibility complex
RIEF	recycling isoelectric focusing
RNA	ribonucleic acid
SDS – PAGE	sodium dodecyl sulfate – polyacrylamide gel electrophoresis
Tris	tris(hydroxymethyl)aminomethane

1. Introduction

In 1957, A. ISAACS and J. LINDENMANN observed that mammalian cells incubated with a heat-inactivated influenza virus produced a substance that could confer on fresh cells a resistance to infection by live virus [1]. This substance, called *interferon*, was subsequently shown to consist of a system of distinct proteins, many of which are structurally related, that are products of multigene families as well as separate genes [2]. Interferons (IFNs) are unique antiviral agents because they act directly on the target cell to confer a state of resistance to viral infectivity at one or more phases of the viral replication cycle, but have no direct toxicity toward the virus itself. The biological effects of IFN are not limited to antiviral activity, however: highly purified preparations of genetically engineered IFN have potent antiproliferative and immunomodulatory effects both in vitro and in animal models [3], [4]. Many types of animal cells can produce IFN in response to a variety of external stimuli or "inducers," including certain types of double-stranded ribonucleic acid (RNA), antigens, and mitogens [5].

Purification of IFN from natural sources for structural and biological characterization, as well as for initiation of clinical trials, proved difficult because IFNs exhibit a high specific biological activity and are normally produced by cells in low absolute quantities. Thus, only relatively impure preparations containing up to 1 % IFN could be partially purified from normal human leukocytes for use in clinical trials in the mid 1970s. Not until the late 1970s could sufficient quantities of naturally occurring IFNs be purified from lymphoblastoid cells for initiation of structural studies [6], [7]. However, the achievement of WEISSMANN and coworkers in cloning the gene for a single subtype of human leukocyte IFN, IFNα-2, in *Escherichia coli* allowed the preparation of gram quantities through genetic engineering [8]. The availability of vir-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_365

tually unlimited quantities of IFN has permitted detailed biological and structural characterization, as well as extensive clinical trials worldwide [9], [10].

Genetically engineered IFNα-2 b (Intron A) has now been approved for marketing in over 30 countries, including the United States, for both cancer and viral indications such as hairy cell leukemia, acquired immune deficiency syndrome (AIDS)-related Kaposi's sarcoma, renal cell carcinoma, malignant melanoma, multiple myeloma, laryngeal papillomatosis, and condyloma acuminata (genital warts). Extensive clinical trials have also been performed on IFNγ [11–14] and IFNβ [15], [16]. This article reviews the properties, production, and specifications of recombinant human IFNα, for which the most data are available. However, for comparative purposes selected data are also provided for human IFNβ and IFNγ, as well as natural leukocyte IFN. For further information, see [2–4], [9], [17–44].

2. Classification and Origin

The first classification of IFN subtypes was based on the observation that exposure to pH 2 inactivated the antiviral activity of some IFN preparations. The IFN stable at pH 2 was designated class I, whereas IFN labile at pH 2 was designated class II. Thus, heterogeneity exists within the IFN system. Subsequently the type I IFNs produced by leukocytes and fibroblasts were shown to differ in their antigenicity and in selected chemical properties. As a result of this observation, type I IFNs were renamed according to the cells of origin, i.e., leukocyte and fibroblast IFN, respectively. These definitions, however, were not adequate because IFN identified as fibroblast with specific antisera was also shown to be produced by leukocytes; leukocyte IFN could likewise be produced by fibroblasts. Therefore, international agreement was reached that IFNs should be categorized as *alpha* and *beta*, depending principally on their antigenic properties. A third antigenic category, referred to as *gamma* (or *immune*) IFN, was also recognized; this IFN class is produced principally by T lymphocytes stimulated by specific antigens or mitogens and is categorized as type II (i.e., acid-labile).

In addition to viruses, a number of natural and synthetic agents are capable of inducing IFN production in a wide variety of cell types [5]. A general summary of cell types and IFN inducers for the three IFN classes is presented in Table 1. Interferon-inducing agents include bacteria, protozoa, mycoplasma, rickettsia, and bacterial products (e.g., bacterial cell wall extracts, lipopolysaccharides, and endotoxins). The effect of synthetic and natural nucleic acids and polyanions, including double-stranded RNA, synthetic polynucleotides, and synthetic polycarboxylates, has been of special interest. Low molecular mass compounds such as anthraquinones, propanediamines, acridines, pyrimidines and fluorenones can also induce IFN titers in the serum of certain animals. Specific structural requirements for these low molecular mass compounds to induce IFN have not been identified yet. Metabolic inhibitors (e.g., actinomycin D), mitogens, antigens, and tumor cells have also been identified as IFN inducers.

Table 1. Classes, sources, and inducers of IFN

Class	Cell source	Inducers
Alpha (leukocyte, type 1)	macrophages	viruses
	B cells	foreign cells and their products
	null cells ∗	low molecular mass inducers (e.g. tilorone)
Beta (fibroblast, type I)	epithelial cells	viruses
	fibroblasts	polynucleotides
	null cells ∗	foreign cells and their products
	lymphoblasts	metabolic inhibitors B-cell mitogens
Gamma (immune, type II)	T cells	T-cell mitogens
	null cells ∗	foreign antigens metabolic inhibitors

∗ Null cells are non-B, non-T lymphocytes.

3. Biological Activity

As previously noted, IFN was first discovered as an antiviral agent that could protect cells from viral infectivity. However, a variety of other biological effects can be elicited by both natural and recombinant IFNs. Utilizing relatively crude preparations, Paucker et al. [45] described the first nonantiviral activity of IFN, i.e., the ability to inhibit cell growth. Subsequently, IFN was also shown to exert potent immunomodulatory

effects as measured by its ability to enhance the activity of immune effector cells (e.g., natural killer cells) [46]. The range of biological activity demonstrated for IFN has increased considerably [47]. One of the important features of IFNs is their species specificity; all classes of human IFN are most effective on human cells and are generally much less potent on animal cells. However, because many of the early studies on nonantiviral activity were performed with impure preparations of naturally occurring IFNs, these biological effects were not known to be inherent properties of IFN. Subsequent use of highly purified recombinant IFN has demonstrated that most observed biological effects can be attributed to IFN [2], [3].

These properties of IFN provide strong support for the clinical application of this class of biological response modifiers as both antiviral and antitumor agents. Aspects of the three categories of biological activity of IFN are reviewed briefly below.

3.1. Antiviral Activity

Early studies indicated that the replication of a broad spectrum of both RNA- and DNA-containing viruses, either oncogenic or nononcogenic, could be inhibited by IFN [37]. Examples of such DNA viruses are herpes virus types 1 and 2 and cytomegalovirus. Examples of RNA viruses are rhinovirus and respiratory syncytial virus. In addition, IFNα can act synergistically with other antiviral agents [48]. However, the degree of antiviral activity depends not only on the nature of the virus, but also on the characteristics of the target cell, the type of IFN, and the ratio of infecting virus to cell number. In vivo studies in nonhuman primate models have also demonstrated the antiviral activity of IFNα and IFNβ against herpesvirus [49].

The specific antiviral activities of highly purified recombinant IFNα subtypes utilizing human foreskin fibroblast cells and encephalomyocarditis virus are summarized in Table 2 [50]. Four of these (α-1, -2, -4, and -7) are naturally occurring gene products, whereas the remaining two subtypes (δ-4 α-1 and the δ-4 α-2/α-1 hybrid) are products of recombinant DNA technology in which the first four amino acids of the amino terminus are deleted. Both of these molecules lack a cysteine residue at position 1 and, hence, are predicted to contain only one disulfide bond (see Chap. 5). The hybrid molecule consists of residues 5 – 62 of IFNα-2 and residues 64 – 166 of IFNα-1. Despite the high degree of purity of these preparations and their amino acid sequence homology (at least 70 %), the specific antiviral activities span approximately two orders of magnitude. Interferon α-2 has the highest antiviral activity (1.7×10^8 IU/mg).

Table 2. Specific antiviral activity of IFNα subtypes $*$

IFN subtype	Activity, IU/mg
α-1	7.1×10^6
α-2	1.7×10^8
α-4	1.0×10^8
α-7	2.0×10^7
δ-4 α-1	2.8×10^6
δ-4 α-2/α-1	1.0×10^8

$*$ Activity was determined by the CPE-inhibition assay employing encephalomyocarditis virus and human foreskin cells; data represent the mean of at least nine assays [50].

Interferon confers resistance to viral infectivity by affecting various stages of the viral replication cycle. It may inhibit early events (e.g., attachment, receptor-mediated endocytosis, uncoating, and transcription), the translation of messenger RNA, and virus maturation, including budding of the virion at the cell membrane. Other evidence also indicates that IFN lowers the infectivity of the progeny virions.

3.2. Anticellular Activity

Interferon can inhibit the growth of a variety of normal and malignant cells in vitro. For example, it can inhibit the growth of fresh tumor cells in the clonogenic assay [51]. This assay is an in vitro cloning technique in which cancer cells form colonies in a semisolid medium. The antigrowth activity of IFN is generally cytostatic (i.e., prevents cell division) rather than cytotoxic (i.e., kills cells directly). Sensitivity to growth inhibition by IFN is variable, even within cells of the same histologic type; this may be due to differences in IFN receptor levels or in the affinity of IFN for the receptor.

Combinations of IFN and chemotherapeutic agents may have a synergistic (enhancing) or additive interaction as antitumor agents [52], [53].

Synergy may be the result of a complex interaction of factors, including the nature of the tumor, the mechanism of action of the cytotoxic agent, and the administration schedule of IFN and the cytotoxic agent. Thus, synergistic interactions between IFN and cytotoxic drugs have been observed in the human tumor clonogenic assay utilizing recombinant IFNα and vinblastine, cisplatin, or doxorubicin [54], [55]. However, synergy is highly schedule-dependent. When cells were first treated with IFN

α followed by doxorubicin or vice versa, no synergy was observed. In contrast, dramatic synergy was observed when both agents were administered simultaneously. Other drugs tested in this assay (e.g., bleomycin, methotrexate, or vinca alkaloids) were not synergistic with IFNα. The mechanism of the observed synergy as well as of drug failure remains unknown.

Another important aspect of the anticellular activity of IFNα is its ability to modulate the state of cellular differentiation. Both inhibition and acceleration of differentiation have been observed [56]. Interferon also interacts synergistically with other agents that inhibit differentiation, e.g., phorbol ester tumor promoters [57]. The induction of differentiation in tumor cells mediated by IFN may account for at least a part of its antitumor activity because the degree of cellular differentiation appears to be inversely related to the rate of cell division.

3.3. Immunomodulatory Activity

A variety of immune functions can be affected by IFN [41]. Immune response may be mediated either by cells or by antibodies. Interferon can activate or inhibit both cellular and humoral (antibody) immune function: higher concentrations generally inhibit, whereas lower concentrations activate. Natural killer cell activity as well as macrophage tumoricidal and tumoristatic activities can be significantly enhanced by both IFNα and IFNγ. Other types of immune effector cells activated by IFN include cytotoxic T lymphocytes, cells involved in antibody-dependent cytotoxicity, and mast cells. Interferons of all three types induce expression of class I antigens of the major histocompatibility complex (MHC), and IFNγ induces expression of class II MHC antigens. This property is significant because class I MHC antigens play a role in the lysis of virus-infected cells and tumor cells by cytotoxic T lymphocytes. Class II MHC antigens are essential for the ability of macrophages to function as antigen-presenting cells. However, both classes of MHC antigens are fundamentally important for achieving maximal immune responsiveness. All types of IFN appear to inhibit antibody production, although under certain conditions of dosing and time of addition, an enhancement of antibody production has been observed.

The role of immune stimulation in clinical responses to IFN remains unknown. For example, measurements of natural killer cell activity after administration of recombinant IFNα have demonstrated both enhancement and depression of activity, as well as no response [58], [59]. However, stimulation by IFN of cytotoxic effector cell function and antibody production may still represent major mechanisms of its therapeutic activity.

4. Mechanism of Action

The first step in expression of the biological activity of IFN is its binding to specific cell surface receptors [60]. The number of receptors for IFNα on human cells is reported to be ca. 650 – 13 000 per cell. The ligand and receptor form a high-affinity complex with a dissociation constant of 10^{-10} – 10^{-11} mol/L. The IFN-receptor complex has an apparent molecular mass of 140 000 – 150 000 deduced from chemical cross-linking of radioactively labeled IFN to the cell surface. Thus, the apparent molecular mass of the receptor is 120 000 – 130 000 if one molecule of IFN is bound per receptor. The failure of certain cells to respond to IFN may be due to absence of the receptor or a low affinity of IFN for the receptor.

Ultrastructural studies indicate that bound IFNα enters cells by a clustering of the receptor – ligand complex into "coated pits." These pits, which are surrounded by the protein clathrin, function in the transport of a variety of molecules into the cell. The pits pinch off to form vesicles (so-called receptosomes), which can transport their contents to a variety of targets within the cell. The role of internalization of IFNα into the cell in the mechanism of action is

yet to be elucidated. Interferon alone, the IFN–receptor complex, or IFN degradation products may trigger a biological response; alternatively, binding of IFN to the cell surface may itself be sufficient.

Binding of IFNα to its receptor induces the synthesis of about a dozen different proteins ranging from 15 000 to 120 000 in molecular mass. Many of these have not been identified, and their role in the mechanism of IFN activity is unknown. Interferon also affects the activity of specific proteins or enzymes. For example, levels of ornithine decarboxylase and *S*-adenosylmethionine decarboxylase, enzymes involved in polyamine biosynthesis, are reduced by IFN treatment [61]. Because these enzymes are critical for cellular proliferation, their inhibition may account for the antigrowth activity of IFN. Interferon treatment can increase cyclic guanosine 5′-monophosphate (cGMP) and cyclic adenosine 5′-monophosphate (cAMP); both of these compounds have also been implicated in growth regulation [61]. However, the exact biochemical mechanism for the antiproliferative effects of IFN remains unknown.

As noted above, all three IFN classes can induce the expression of class I MHC antigens [62], which mediate the lysis of virus-infected cells and tumor cells by cytotoxic T lymphocytes. In addition, IFN can enhance the expression of the immunoglobulin G receptor on lymphocytes and macrophages; this, in turn, may enhance antibody-dependent cytotoxicity or macrophage-mediated phagocytosis [63].

An extensively characterized enzyme that is induced by IFN treatment and appears to play a critical role in interferon's antiviral and antiproliferative activities is 2′,5′-oligoadenylate synthetase (E. C. 2.7.7.-) [64]. This enzyme catalyzes the synthesis of a family of novel 2′,5′-oligoadenylate nucleotides that regulate the breakdown of RNA in cells. Reported levels of induction of the synthetase vary between 10- and 10 000-fold. The enzyme requires double-stranded RNA, which may be produced by the virus as a replicative intermediate, for expression of enzyme activity. Although the exact mechanism for this activation is not known, the double-stranded RNA may bind to the enzyme and convert it to its biologically active form. The 2′,5′-oligoadenylates activate a latent endoribonuclease, known as RNase L, which catalyzes the cleavage of both messenger and ribosomal RNA. Thus, the localized activation of endonuclease by double-stranded RNA present during viral replication may account for the selective cleavage of viral versus host RNA.

Interferon induces another system that also requires double-stranded RNA for activation, namely, a protein kinase that catalyzes the phosphorylation of protein initiation factor eIF-2, as well as the ribosomal protein P 1 [65]. In cell-free systems, the kinase is activated by much lower concentrations of double-stranded RNA than is 2′,5′-oligoadenylate synthetase; higher concentrations of double-stranded RNA inhibit the kinase but activate the synthetase. Thus, the level of double-stranded RNA may determine which biochemical pathway is important in the expression of antiviral activity.

5. Physicochemical Properties

Only very low (picogram) quantities of IFN can normally be produced from natural sources because the biological activity is expressed at very low protein levels. Thus, characterization of its properties was for many years slow and incomplete. Studies on IFNα prepared from a variety of natural sources, including virally induced natural leukocytes, lymphoblastoid cells, and chronic myelogenous leukemic cells, suggested that IFNα consists of a family of multiple species with apparent molecular masses determined by sodium dodecyl sulfate – polyacrylamide gel electrophoresis (SDS – PAGE), ranging from 16 000 to 27 000. Amino acid analysis of the purified proteins indicated that their compositions were similar but that IFNα consisted of a family of proteins with different amino acid sequences. However, further characterization of the physicochemical properties of IFNs was severely hindered by the fact that IFNs purified from natural sources were frequently contaminated with other proteins and IFN subtypes. In addition, the extremely low levels of substance produced resulted in large losses of protein because of adsorption and surface denaturation. These problems were solved by the large-scale production of genetically engineered IFNα-2 b (Intron A) in *E. coli* [3, pp. 1 – 12], [10].

5.1. Primary Structure

The amino acid sequence of IFNα-2 deduced from the DNA sequence of the cloned gene is depicted in Figure 1. This sequence has been confirmed by *N*-terminal sequencing through automated Edman degradation of the native protein and of fragments produced by either proteolytic or cyanogen bromide cleavage. The DNA sequences of 16 other IFN genes indicate an amino acid sequence homology of 75 % or more and an identity of amino acids at approximately 60 % of all positions for the majority of IFNα subtypes. However, some IFNα subtypes differ by only one amino acid and yet are produced by distinct, nonallelic genes, e.g., IFNα-A and IFNα-2. The IFNα subtypes generally consist of 165 or 166 amino acid residues preceded by a 23-amino acid signal sequence. These signal sequences are not found in the mature protein because they are removed by proteolysis during secretion of IFN by the cell.

All IFNα subtypes contain four cysteine residues. As shown in Figure 1, for IFNα-2 b these residues occur at positions 1, 29, 98, and 138. Chromatographic analysis of proteolytic fragments has shown that these amino acids form two disulfide bonds between residues 1 and 98, and residues 29 and 138. Pairing of the cysteine residues is probably the same in all IFNα subtypes. Studies with reducing agents indicate that the 29 – 138 disulfide bond is required for the full expression of biological activity, whereas the 1 – 98 disulfide bond may be disrupted without a significant loss in activity.

Based on examination of the biological activity of various recombinant hybrid IFNα molecules, IFNα appears to contain at least two distinct binding sites located at the amino and the carboxyl terminals, respectively. Each binding site is thought to interact with the IFN receptor, and the nature of the biological response is determined by the quality of fit between ligand and receptor [66]. However, other studies suggest that the final 13 carboxyl-terminal amino acids of IFNα are not required for biological activity.

Two forms of IFNβ have been cloned, and the amino acid sequence predicted from the complementary DNA sequence [67], [68]. The principal form of IFNβ produced by fibroblasts, IFNβ-1, contains 166 amino acids and exhibits approximately 30 % homology in amino acid sequence with IFNα-1. However, the region of highest homology between IFNβ-1 and IFNα-1 lies between residues 115 and 151, which suggests an important functional role for this region.

A single form of IFNγ has been identified [69]. The gene structure predicts that this protein contains 146 amino acids, significantly fewer than IFNα or IFNβ. Virtually no homology exists at the amino acid level with either IFNα or IFNβ. Another form of IFNγ that lacks three amino acids (cysteine – tyrosine – cysteine) at the amino terminus has also been cloned. The specific antiviral activity of this species appears to be higher than that of the 146-amino acid form.

5.2. Secondary and Tertiary Structure

The secondary structure of purified recombinant IFNα-2 b has been examined by circular dichroism (CD) in both the near and the far UV range (Fig. 2).

Intense, negative bands at 218 and 208 nm are characteristic of an alpha helix. Calculations based on the amplitude of these bands indicate an alpha helicity of ca. 50 %. Although the CD spectra do not indicate the presence of beta sheets, a small percentage (e.g., 0 – 10 %) may be present but undetectable in the spectrum. Near-ultraviolet CD spectra indicate two intense negative bands at 291 and 286 nm, which can be assigned to transitions of tryptophan. The intensity and location of these bands suggest that at least one of the two tryptophan residues is present in a hydrophobic environment. Weak, negative CD bands at 255, 262, and 268 nm can be assigned to phenylalanine. Titration of IFNα-2 b solution to pH 2 results in virtually complete loss of tryptophan CD bands as well as substantial depression in the CD bands representative of the alpha helix. However, the apparent loss in native conformation that occurs on acidification is completely reversible on retitration to pH 7.4. In parallel with these findings, the antiviral activity of IFNα-2 is also retained completely after exposure to pH 2 and neutralization.

Precise information on the tertiary structure of IFNα is not yet available. However, crystallization of recombinant IFNα-2 b (Fig. 3) has provided a basis for X-ray diffraction analysis

CYS	ASP	LEU	PRO	GLN	THR	HIS	SER	LEU	GLY	SER	ARG	ARG	THR	LEU
MET	LEU	LEU	ALA	GLN	MET	ARG	ARG	ILE	SER	LEU	PHE	SER	CYS	LEU
LYS	ASP	ARG	HIS	ASP	PHE	GLY	PHE	PRO	GLN	GLU	GLU	PHE	GLY	ASN
GLN	PHE	GLN	LYS	ALA	GLU	THR	ILE	PRO	VAL	LEU	HIS	GLU	MET	ILE
GLN	GLN	ILE	PHE	ASN	LEU	PHE	SER	THR	LYS	ASP	SER	SER	ALA	ALA
TRP	ASP	GLU	THR	LEU	LEU	ASP	LYS	PHE	TYR	THR	GLU	LEU	TYR	GLN
GLN	LEU	ASN	ASP	LEU	GLU	ALA	CYS	VAL	ILE	GLN	GLY	VAL	GLY	VAL
THR	GLU	THR	PRO	LEU	MET	LYS	GLU	ASP	SER	ILE	LEU	ALA	VAL	ARG
LYS	TYR	PHE	GLN	ARG	ILE	THR	LEU	TYR	LEU	LYS	GLU	LYS	LYS	TYR
SER	PRO	CYS	ALA	TRP	GLU	VAL	VAL	ARG	ALA	GLU	ILE	MET	ARG	SER
PHE	SER	LEU	SER	THR	ASN	LEU	GLN	GLU	SER	LEU	ARG	SER	LYS	GLU

Figure 1. Amino acid sequence of human IFNα-2 b (Intron A)

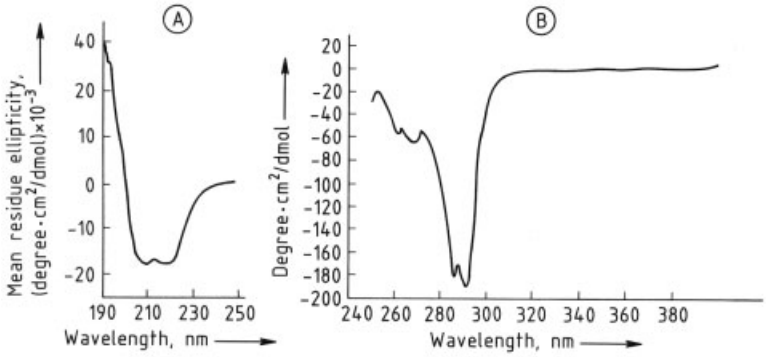

Figure 2. Circular dichroism spectra of human IFNα-2 b (Intron A)
A) Far ultraviolet; B) Near ultraviolet

of three-dimensional structure. Needle-shaped crystals, with dimensions of ca. 0.1×0.01 mm^2, are prepared by incubation of a purified IFNα-2 solution at pH 6.0, which is near the isoelectric point of the protein [10]. Crystal dimensions vary depending on the conditions of crystallization.

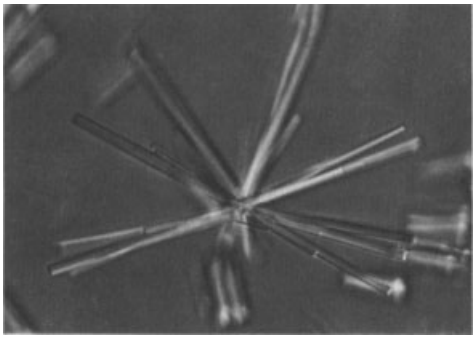

Figure 3. Human IFNα-2 b (Intron A) crystals

Crystallization of a form of human IFNγ (IFNγ-D$'$), in which the five amino acids at the carboxyl terminus are deleted, has also been reported [70].

Crystals suitable for high-resolution X-ray diffraction have been prepared by vapor-diffusion equilibration against a solution of 30 % saturated ammonium sulfate in 50 mmol/L sodium acetate, pH 5.9 [70]. Rhombohedral crystals ($0.45 \times 0.45 \times 0.40$ mm) are formed after 5-d incubation at room temperature (Fig. 4). The X-ray precession photographs indicate that the crystals are trigonal. The crystals are moderately stable to X radiation and diffracted to a resolution of 0.33 nm on a rotating anode generator. The three dominant faces of the rhombohedral crystal correspond to the (110), (101), and (011) planes of the rhombohedral unit cell. The measured density of the crystals is 1.224 g/cm^3 in a Ficoll – 45 % ammonium sulfate solution. The calculated mass of the protein per unit volume is 63 200, which indicates that four molecules of IFNγ-D$'$ are present in the asymmetric unit. Diffraction data from native crystals and from several heavy-atom derivatives have been collected by oscillation photography with synchrotron radiation [70]. The intensity data sets from native crystals (resolution to 0.285 nm) are being analyzed to determine the

three-dimensional structure at this degree of resolution.

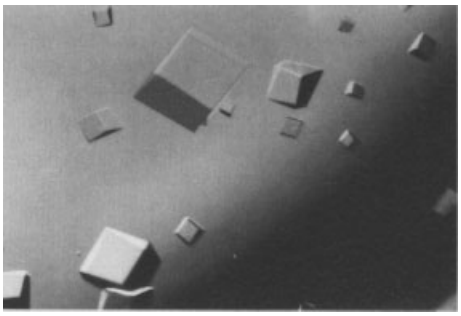

Figure 4. Human IFNγ crystals in a suspended droplet

5.3. Solution Properties

Based on their predicted amino acid compositions, the polypeptide molecular masses for human IFNα, β, γ are 19 269, 19 990, and 17 145, respectively. However, under certain conditions the apparent molecular mass of IFNα in solution appears to be considerably higher than the polypeptide molecular mass [10]. Sedimentation velocity experiments have established that IFNα-2 undergoes self-association to yield a molecular mass at least three times that of the monomer. Lowering the pH from 7 to 2 favors monomer formation. Under the low protein concentration in a biological assay, it is anticipated that IFNα is completely monomeric; thus, the monomer is probably the active species. Even under highly concentrated conditions (e.g., 10 mg/mL or higher), IFNα-2 b appears to be very soluble. At its isoelectric point (ca. pH 6.0), IFNα is relatively insoluble and can be crystallized readily. In contrast, IFNγ appears to form a stable dimer that does not dissociate on dilution and has a highly basic isoelectric point of ca. 9.5.

6. Purification

6.1. Natural Leukocyte IFN

Starting materials used for the purification of human IFN include leukocytes isolated from whole blood, neonatal fibroblasts, and lymphoblastoid and various leukemic cell lines. The principal source for purification of IFN prior to the use of genetically engineered microorganisms was human cell cultures treated with an appropriate inducer (e.g., a virus or a double-stranded polynucleotide). The principal problem with the isolation of IFN from such sources is the low level of expression of IFN due to its potent biological activity. Large starting volumes and multiple purification steps are thus required, and only small quantities of material can be purified. For example, in 1978, two laboratories in Helsinki together produced only 1.3 g of IFNα of <1 % purity from 90 000 bags (45 000 L) of donated blood. Lack of precision in the antiviral assay commonly employed for monitoring the degree of purification (see Section 7.1) has also been a problem in the development of purification protocols.

Partially purified human leukocyte IFN was first developed by CANTELL and his associates in Finland and has been available in limited quantities for clinical trials since 1973 [71]. In this procedure, fresh blood from normal donors is centrifuged, and the leukocyte fraction (i.e., the buffy coat) is removed. Pooled leukocytes are treated with 0.83 % ammonium chloride to lyse residual erythrocyte contaminants, suspended in a culture medium (e.g., Eagle's minimum essential medium), and then primed with high-titer crude IFN. Interferons production (primarily by monocytes) is then induced by addition of Sendai virus. After overnight incubation, the cells and debris are removed by centrifugation; the supernatant containing crude IFN is the starting material for purification.

Crude IFN is first precipitated with potassium thiocyanate at pH 3.4 and extracted with 95 % ethanol. Extraction with acidified ethanol inactivates residual Sendai virus and partially purifies the IFN because not all proteins redissolve. The pH is then increased stepwise to precipitate contaminating proteins, which are removed by centrifugation. The IFN is precipitated at pH 8, concentrated, and dialyzed against phosphate-buffered saline.

The specific antiviral activity of the IFN preparation is approximately 10^6 IU/mg of protein, which represents approximately 1 % purity. The net yield per unit of blood (i.e., 0.5 L) is ca. $(5-7)\times10^6$ antiviral units. The final product contains a number of different species of IFNα,

including α-1, α-2, and α-4. Other lymphokines (i.e., hormones released by leukocytes such as interleukin-1 and 2) probably also contaminate the preparation.

The partially purified leukocyte IFN preparation can be purified further to a specific activity of ca. 10^8 IU/mg of protein by chromatography on an immobilized monoclonal antibody referred to as NK-2. The IFN is eluted from the column by exposure to low pH. However, the relative composition of IFNα subtypes in this preparation differs from that in crude IFN prior to chromatography because the NK-2 antibody does not bind all IFNα species equally.

6.2. Recombinant IFNα (→Genetic Engineering, Chap. 5.1.)

Purification of recombinant IFN from a genetically engineered microorganism poses unique problems distinct from those encountered in purification from natural sources [72]. The following requirements must be satisfied:

1) A procedure for efficient extraction of IFN from the cell that does not denature the protein must be developed. High concentrations of a conformation-disrupting agent such as urea or guanidine should be avoided for extraction because renaturation of the extracted protein may be difficult or impossible. This consideration is especially important for human IFNα because two disulfide bonds must be reformed in the native structure and the use of a chaotrope may lead to incorrect disulfide pairing. Furthermore, even if the final product is biologically active and appears to be properly refolded, small quantities of incorrectly folded material may be highly antigenic and remain undetected by standard analytical methods.
2) The extraction and purification procedure must yield a protein that is uncontaminated with bacterial pyrogens (gram-negative bacterial lipopolysaccharides that increase body temperature when injected parenterally). Pyrogens are not readily removed from purified IFN preparations because effective methods for their removal from solution (e.g., adsorption on activated charcoal or destruction with ionizing radiation) may damage the biological activity or structure of the recombinant protein.
3) The purified IFN must be substantially free of nucleic acids and proteins derived from the host microorganism if it is to be employed in human clinical trials.
4) The product must be a homogeneous IFN species that is free of aggregates and proteolytic fragments. These IFN-related components are of special concern because they could render the product antigenic when injected into humans.
5) To be commercially viable, purification must be achieved with a good yield (e.g., 20 %) of biological activity (e.g., antiviral units).

Purification of recombinant IFNα from *E. coli* for the IFNδ-4 α-2/α-1 hybrid that was genetically engineered to contain amino acid residues 4 – 62 of IFNα-2 and 54 – 166 of IFNα-1 is summarized in Table 3. The vector for this species was a PBR-322-based plasmid under control of the *lac* promoter [73] (see also →Genetic Engineering, Chap. 4.1.1.). The purification of IFN is monitored by following changes in specific antiviral activity (IU) expressed per milligram of protein.

The *E. coli* extract is prepared by acidification of the bacteria to pH 2, followed by adjustment to neutral pH (see Section 6.1) [74]. This procedure exploits the acid stability of type I (α and β) IFNs (see chap. 2) and avoids high concentrations of denaturants. The first purification step removes nucleic acids and contaminating *E. coli* proteins by acidification to pH 4.5 followed by centrifugation and is accompanied by a 20-fold increase in specific antiviral activity. The stability of IFNα at low pH and in organic solvents permits subsequent precipitation by addition of trichloroacetic acid and extraction from the precipitate with ethanol. This step increases the specific activity by an additional 20-fold and also concentrates the IFN solution with virtually no loss in activity. The 118 % yield indicates complete recovery of antiviral activity within the experimental accuracy of the assay. The next step consists of affinity chromatography on a blue dextran affinity medium [Matrex Gel Blue A (Cibracron Blue F 3 GA) Sepharose]. This resin is used because both human and murine IFNs bind to blue dextran. Elution from this column is achieved with

Table 3. Purification of recombinant human IFNδ-4 α-2/α-1

Purification step	Volume, L	Antiviral activity			Purification ∗, fold
		10^7 IU/mg	Total, 10^9 IU	Yield, %	
Escherichia coli extract ∗∗	765.0	0.05	629	100	(1.0)
Acidification (pH 4.5)	765.0	1.12	466	67	2.4
Trichloroacetic acid – ethanol extract	21.0	2.1	740	118	42
Matrex Gel Blue A Sepharose	15.3	7.7	177	28	154
Concentration by ultrafiltration followed by	1.0	8.2	192	31	164
Selective precipitation	0.069	7.0	30	5	140
DEAE Sepharose chromatography	0.031	7.0	18	3	140

∗ Based on antiviral units.

∗∗ Extract was prepared by acidification of the *E. coli*, followed by adjustment to neutral pH for release of the IFN.

sodium chloride solution (2 mol/L). The final purification step can be performed in two ways: (1) repeated cycles of selective precipitation by dialyzing to low ionic strength and redissolving in 2 mol/L sodium chloride or (2) selective precipitation followed by ion-exchange chromatography on diethylaminoethyl (DEAE) Sepharose. The specific activity of the final product prepared by either procedure is 7×10^7 IU/mg, which represents a 140-fold purification.

An alternative approach for purifying recombinant IFNα employs immunosorbent chromatography with an immobilized monoclonal antibody [72]. Purification of IFNα-5 by this technique is summarized in Table 4.

The IFNα-5 is extracted by acidification of *E. coli* followed by neutralization (see Section 6.1). Extracted IFN is adsorbed on silica gel and eluted by lowering the pH to 2.0, which results in an almost tenfold increase in specific activity. Silica gel chromatography also removes the majority of contaminating nucleic acids in the unbound fraction. Matrex Gel Blue A chromatography results in a further 3.3-fold purification. Next, IFN is adsorbed onto an anion-exchange resin (i.e., DEAE Sepharose) and eluted with a pH gradient from 8.0 to 6.0, resulting in an additional tenfold increase in specific activity. The DEAE Sepharose eluate is purified further on an immobilized monoclonal antibody, YOK Sepharose. Elution is achieved with 0.1 mol/L sodium citrate, pH 2.0, to give a final specific activity of 1.9×10^8 IU/mg.

6.3. Removal of Pyrogens

The final IFN preparation should be low in pyrogenic substances if it is to be employed in clinical trials (see Section 6.2). Although the purification protocols described in Sections 6.1 and 6.2 generally result in a product that is very low in pyrogens (i.e., 5 – 10 ng of endotoxin per milligram of protein, determined as described in Section 7.2), preparations with high pyrogen levels are occasionally encountered. An effective procedure for lowering the endotoxin level of these preparations is recycling isolelectric focusing (RIEF). Isoelectric focusing is based on the principle that charged substances migrate in an electric field and a stable pH gradient to a pH at which their net change is zero. In the RIEF apparatus designed by Bier et al. [75], the protein solution is continuously recycled between a multichannel heat exchange reservoir and a multichannel focusing cell. Thus, heat is dissipated in the heat exchange reservoir and not in the focusing cell. As shown in Figure 5, when RIEF is applied to a solution of IFNα-2 with a high pyrogen content, the majority of endotoxins and the IFNα-2 migrate to different pH values [76]. Thus, whereas endotoxins reach maximum levels at pH 5.07 (channel 5), the antiviral activity of IFNα-2 peaks at pH 6.4 (channel 10).

In the removal of contaminating pyrogens from IFNα-2, a 40-fold purification is obtained from a single RIEF experiment, which results in a pyrogen level of 12.5 ng/mg of protein, compared to 500 ng/mg of protein prior to RIEF. No apparent loss in specific antiviral activity is observed, and the total recovery of IFN activity

Table 4. Purification of recombinant human IFNα-5

Purification step	Volume, L	Antiviral activity			Purification, fold
		10^7 IU/mg	Total, 10^9 IU	Yield, %	
Escherichia coli extract	200.0	0.019	42	(100)	(1.0)
Silica (Si 200)	1.8	0.18	32	76	9.5
Matrex Gel Blue A	0.8	0.6	25	60	31.6
DEAE Sepharose CL-6 B	0.2	5.8	10	24	305
YOK Sepharose	0.04	19.0	8	19	1000

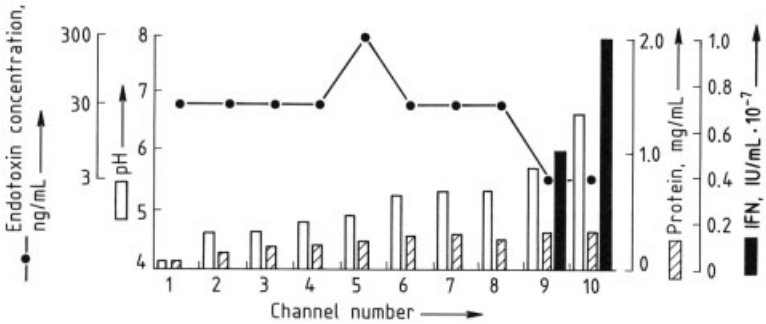

Figure 5. Recycling isoelectric focusing (RIEF) of human recombinant IFNα-2 b (Intron A) containing *Escherichia coli* pyrogens (endotoxins)
The apparatus was prefocused with 0.1 % ampholines (pH range 3 – 10) in urea (4 mol/L) for 60 min followed by addition of the sample. Protein concentration was determined from the optical density reading at 280 nm; pyrogen levels were assayed by the limulus amebocyte lysate assay, and antiviral activity was determined by CPE-inhibition assay.

is >90 %. The ampholines used to establish the pH gradient are finally removed from the preparation by crystallization of the protein, gel filtration chromatography, or ammonium sulfate precipitation.

7. Quality Specifications

7.1. Purity

Purified IFN must be shown to be biologically active, and its activity must remain constant from batch to batch. Because IFN was originally defined in terms of its antiviral activity, quantitation of antiviral activity per milligram of purified protein is routinely employed for characterization of purified product. The most commonly employed assay, which is readily adapted to screening large numbers of samples, is the *cytopathic effect- (CPE-) inhibition assay*. This assay is based on the ability of IFN to interfere with the cytopathic effects (e.g., cell lysis) caused by viruses. In a simple CPE assay, IFN is serially diluted (usually by a factor of two) in the cells of a microtiter plate, and target cells (e.g., human foreskin diploid fibroblasts) are then added to each well. Cells to which IFN has not been added are also included as controls. After an incubation period (e.g., 4 h), a challenge virus (e.g., encephalomyocarditis virus) is added, followed by a further period of incubation. In the presence of IFN, cells are protected from viral infection and can be detected by staining with a dye such as Crystal Violet. However, in the absence of IFN, the cells lyse and are not stained by the dye because they do not adhere to the microtiter plate. The end point is arbitrarily defined (e.g., 50 % protection against viral infectivity). An appropriate international reference standard provided by the World Health Organization is included in each microtiter plate and allows conversion to the international antiviral units (IU) established by international accord. Further details of the CPE-inhibition assay are described in [77].

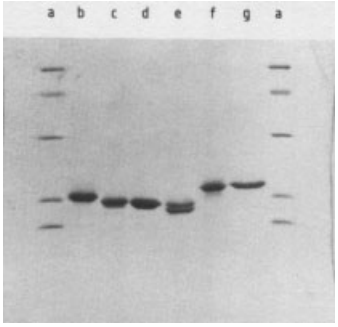

Figure 6. Polyacrylamide gel electrophoresis of purified recombinant IFNα subtypes
Before loading, 0.01 mg of each subtype was boiled for 2 min in 1 % 2-mercaptoethanol, 1 % sodium dodecyl sulfate, and 10 % glycerol dissolved in 0.06 mol/L Tris hydrochloride, pH 6.8, containing 0.001 % bromophenol blue. Discontinuous SDS – PAGE was run according to [79].
a) Standard proteins (lysozyme, β-lactoglobulin, carbonic anhydrase, ovalbumin, and bovine serum albumin); b) IFNα-1; c) IFNδ-4 α-1; d) IFNα-2; e) IFNδ-4 α-2/α-1; f) IFNα-4; g) IFNα-7

Sodium dodecyl sulfate – polyacrylamide gel electrophoresis is a powerful technique for demonstrating the purity of the final IFN preparation [78]. The quantitation of impurities and sensitivity of the method depend on the nature of the stain employed for protein detection. Quantities of protein as low as 0.05 – 0.1 µg can be detected by staining with Coomassie Blue R 250, whereas silver staining may result in detection of levels of impurities up to 100-fold lower. However, silver-stained gels must be interpreted with caution because staining intensity varies widely among proteins (up to a factor of ten). A typical gel electrophoresis of purified recombinant IFNα subtypes stained with Coomassie Blue R 250 is presented in Figure 6; SDS-PAGE reveals a purity of $\geq$95 % for each subtype, which was confirmed by silver staining. The two components observed for the δ-4 α-2/α-1 hybrid (Fig. 6, lane e) were both demonstrated to be IFN-related. Approximate quantitation of impurities can be achieved by electrophoresis at IFN loading levels ranging from 25 to 0.05 µg. The loading level at which an impurity disappears relative to the main component can be used to calculate the approximate percentage of impurity. The SDS-PAGE technique can also be employed for estimation of apparent molecular mass by comparison to the migration of standards with known molecular mass. Thus, for example, the average apparent molecular mass of IFNα-2 was observed to be 18 500 $\pm$ 600, the standard deviation representing gel-to-gel variation. The calculated molecular mass of IFNα-2 is 19 270, in good agreement with the experimental value.

Reversed-phase high-performance liquid chromatography (HPLC) is also employed to characterize the purity of the final product [80]. Reversed-phase HPLC of a clinical-grade sample of IFNα-2 b prior to formulation is presented in Figure 7. Results indicate a high degree of homogeneity, in agreement with those obtained by SDS-PAGE.

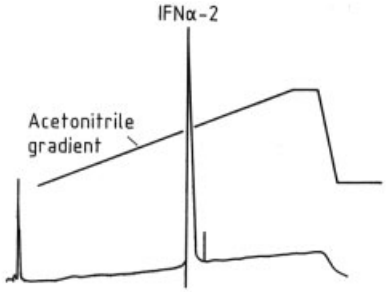

Figure 7. Reversed-phase HPLC of purified recombinant human IFNα-2 b (Intron A)
Chromatography was performed on a Water Associates μ-Bondapak C-18 column (30 cm $\times$ 3.9 mm). The eluting gradient extends from 31.4 to 62.7 % acetonitrile containing 0.01 mol/L trifluoroacetic acid at a flow rate of 1 mL/min for 30 min. Detection was accomplished by monitoring fluorescence after post-column derivatization with *o*-phthalaldehyde. The first peak is caused by the solvent.

Amino acid analysis can also be used to characterize the purified preparation. Table 5 contains the average amino acid analysis of 15 batches of purified IFNα-2 b. Interferon was first hydrolyzed with 6.7 mol/L hydrochloric acid for 18 h at 160 °C, followed by chromatographic resolution of the individual amino acids on a C-18 HPLC reversed-phase column utilizing paired ion chromatography [81]. The amino acids were detected by fluorescence following postcolumn derivatization with *o*-phthalaldehyde. Excellent agreement was observed between experimental and theoretical values.

Amino acid sequencing provides a more sensitive method than amino acid analysis for detecting contaminating proteins, as well as for demonstrating that the recombinant protein has the correct amino acid sequence predicted from

the nucleotide sequence of the gene. For example, *N*-terminal sequencing has been achieved for recombinant IFNα-2 by using automated Edman degradation in a gas-phase sequenator [82].

Table 5. Amino acid analysis of recombinant human IFNα-2 b

Amino acid	Theoretical content, mol %	Measured content, mol %
Aspartic acid, asparagine	7.74	9.06 ± 0.46
Serine	9.03	10.33 ± 0.48
Glycine	3.23	3.80 ± 0.91
Glutamic acid, glutamine	16.77	19.03 ± 0.79
Threonine	6.45	5.54 ± 0.58
Alanine	5.16	5.61 ± 0.58
Valine	4.52	2.62 ± 0.23
Methionine	3.23	2.82 ± 0.59
Tyrosine	3.23	2.81 ± 0.45
Isoleucine	5.16	3.52 ± 0.39
Leucine	13.55	13.96 ± 0.46
Phenylalanine	6.45	6.51 ± 0.20
Histidine	1.94	1.95 ± 0.15
Lysine	7.10	6.53 ± 0.58
Arginine	6.45	5.99 ± 0.70

The phenylthiohydantoin – amino acids were resolved by reversed-phase HPLC on a cyanopropyl silane column, with acetonitrile – methanol (4 : 1) as eluent. Sequencing was performed after reduction with dithiothreitol and alkylation with iodoacetamide to permit identification of cysteine residues. The first 57 amino acids agreed completely with the amino acid sequence predicted from the complementary DNA (Fig. 1). In addition, by utilizing peptide fragments produced by cyanogen bromide treatment (which cleaves the protein at methionyl residues), 90 % of the amino acid sequence was determined. Tryptic digestion of the C-terminal cyanogen bromide fragment followed by reversed-phase HPLC and amino acid analysis of the separated components was performed to identify glutamic acid as the C-terminal amino acid. The amino acid sequence was also surveyed by gas chromatography – mass spectrometry of fragments of IFNα-2 produced by subtilisin cleavage followed by trifluoroacetylation and permethylation. Selected sequences with IFNα-2 were identified and demonstrated to correspond to the gene structure.

7.2. Pyrogenicity Testing

Amebocytes circulating in the horseshoe crab (*Limulus*) hemolymph contain a coagulation system that is highly sensitive to gram-negative bacterial endotoxins [83]. The *Limulus* amebocyte lysates form a firm gel in the presence of low levels of endotoxin. The "limulus test" for detection of endotoxins, based on this gelation reaction, is extensively employed in the pharmaceutical industry. As shown in Figure 8, extremely low endotoxin levels are detectable in purified preparations of recombinant IFNα-2, in which the values generally are significantly lower than 1.0 ng/mg of protein.

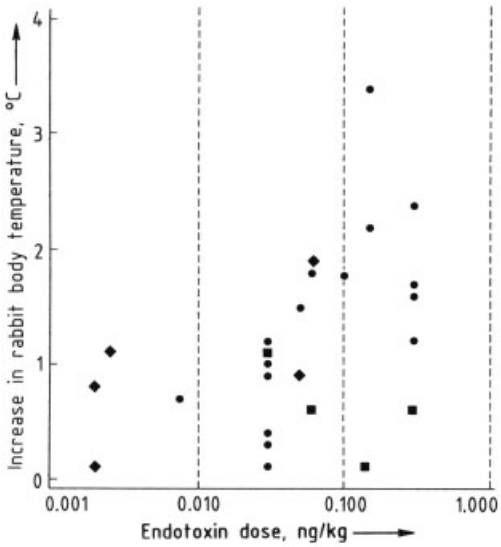

Figure 8. Pyrogenicity of recombinant human IFNα-2 b (Intron A)
IFN dose, IU/kg: ■ 0.75; ● 1.5; ♦ 2.5

An alternative, traditional method for investigating pyrogen content is to measure the *increase in body temperature* following intravenous injection into rabbits. However, as shown in Figure 8, no apparent correlation existed between temperature increase and either IFN dose or endotoxin level measured in the limulus test. The in vivo test is further complicated by the fact that IFN itself has inherent pyrogenic properties. Thus, the limulus test appears to be a convenient and easily performed in vitro test for estimation of endotoxin levels.

8. Clinical Studies

Although IFN was first discovered in 1957, not until 1981 were large enough quantities of purified recombinant IFNα available to permit the initiation of phase I/II human clinical trials. To date over 2000 patients with either advanced malignancies or viral infections have been treated with IFNα-2 b (Intron A). Despite the fact that cancer patients treated with Intron A generally displayed poor prognosis as well as resistance to standard chemotherapy, a number of patients demonstrated complete recovery of long-term duration.

Phase I trials with Intron A were generally performed on patients with advanced malignancies and on a small number of normal volunteers [9], [84]. These studies indicate that Intron A is safe in humans; its toxic effects are dose-related and completely reversible. No evidence for cumulative toxicity or drug intolerance was obtained. Three routes of administration (subcutaneous, intramuscular, and intravenous) were employed in a variety of schedules. For example, Intron A was administered in an escalating daily dose for 28 d to determine the maximum tolerable dosage. Dosages were also elevated on a weekly schedule. The total daily doses examined were generally in the range of $(1-100)\times 10^6$ IU. The principal toxic effect observed in 90 % of all patients consisted of flu-like symptoms (e.g., fatigue, fever, headache). However, these dose-related effects reversed rapidly when the dosage was lowered or treatment discontinued. Other hematologic toxicities, including mild depression of white blood cell and platelet counts, and elevation of hepatic enzymes were also observed in some patients but were all reversible upon discontinuance of IFN administration. Single Intron A doses of up to 200×10^6 IU per square meter of body surface area could be tolerated when administered intravenously. However, for 5-d intravenous scheduling, 50×10^6 IU/m^2 was the maximum dose tolerated.

The pharmacokinetics of Intron A were also evaluated in phase I trials. Intravenous infusion for 30 min produced peak serum levels within 15 – 30 min. Because of a relatively short half-life, serum levels returned to base line within 4 h. However, sustained levels could be achieved by continuous intravenous infusion. Subcutaneous and intramuscular dosing demonstrated distinct pharmacokinetics with slowly rising blood levels that peaked at 4 h and maintained sustained levels for 8 – 12 h. Based on in vitro studies, the peak serum levels obtained by any of these three routes of administration should be sufficient to activate immune effector cells or to exert a direct antiproliferative effect on tumor cells.

Phase II clinical trials of Intron A have been conducted on a variety of malignancies. Results are as follows:

Active:	hairy cell leukemia
	non-Hodgkin's lymphoma
	mycosis fungoides
	Kaposi's sarcoma
	multiple myeloma
	malignant melanoma
	renal cell cancer
	bladder cancer (superficial tumors)
	ovarian cancer
Inactive:	breast cancer
	colon cancer
	non-small cell lung cancer
	prostate cancer
	acute myelogenous leukemia
To be determined:	chronic myelogenous leukemia
	chronic lymphocytic leukemia
	preleukemia
	Hodgkin's disease
	osteogenic sarcoma
	astrocytoma

Efficacy has been observed most often in hematological malignancies, particularly those of pre-B-cell and B-cell origin. For example, in hairy-cell leukemia, Intron A administered subcutaneously at relatively low doses for continuous periods produced hematological responses in ca. 90 % of patients and objective responses in ca. 75 %. Complete responses of relatively long duration have also been observed in multiple myeloma and low-grade lymphomas.

Most adult solid tumors (e.g., breast, lung, and colon) have not responded well to Intron A therapy. However, clear activity has been observed in malignant melanoma, in which the therapeutic effects were apparently schedule-dependent because continuous dosing was more efficacious than intermittent dosing. Tumor burden is probably also important in determining the response to Intron A because malignant melanoma patients with large tumor burdens appeared to have a lower incidence of response. Objective responses have also been observed in superficial bladder cancer, renal cell cancer,

ovarian cancer, and AIDS-related Kaposi's sarcoma.

Intron A has also demonstrated activity in viral indications. For example, intralesional administration into genital warts (three times weekly for three weeks) can result in complete regression of the injected warts between four and twelve weeks. Therapeutic activity has also been observed in laryngeal papillomatosis, in chronic hepatitis B, and as a prophylactive treatment for the common cold.

In all these clinical trials, the incidence of human antibody formation to Intron A was very low [85]. An antibody response was elicited in $<3\,\%$ of more than 500 patients treated systemically and in $<1\,\%$ of more than 1000 patients treated intralesionally or intranasally. Significantly higher rates of neutralizing antibody production have been reported for recombinant IFNβ (ca. 65 % of treated patients) and recombinant IFNα-2 a (Roferon) (ca. 25 – 30 % of treated patients).

The presence of neutralizing antibodies to Roferon has been reported in 7 out of 12 responding patients and in 9 out of 29 nonresponding patients with metastatic renal cell carcinoma [86]. A comparison of the medium duration of remission between antibody-positive and antibody-negative patients in this study suggested that antibody formation contributed to early relapse. Similarly, 17 out of 23 patients in a phase I clinical trial with a form of recombinant IFNβ containing a genetically engineered serine residue at position 17 showed evidence for antibody formation [87]. These data strongly support that Intron A is significantly less antigenic in human subjects than other recombinant IFNs.

9. References

1. A. Isaacs, J. Lindenmann, *Proc. R. Soc. (London)* **Sec. B 147** (1957) 258 – 267.
2. P. P. Trotta, S. K. Narula in S. Baron, F. Dianzani, G. J. Stanton, W. R. Fleischmann, Jr. (eds.): *Interferon Genes,* Univ. of Texas Press, Austin 1987, pp. 137 – 147.
3. P. P. Trotta in P. L. Kisner, J. F. Smyth (eds.): *Interferon Alpha-2: Preclinical and Clinical Evaluation,* Martinus Nijhoff Medical Division, The Hague 1985, pp. 13 – 27.
4. P. P. Trotta, R. J. Spiegel in F. M. Muggia (ed.): *Concepts, Clinical Developments, and Therapeutic Advances in Cancer Chemotherapy,* Martinus Nijhoff Publishers, Boston 1987, pp. 4 – 159.
5. D. A. Stringfellow (ed.): *Interferon and Interferon Inducers-Clinical Applications,* Marcel Dekker, New York 1980.
6. K. C. Zoon, D. Miller, D. Zur Nedden, M. W. Hunkapiller, *Proc. Natl. Acad. Sci. U.S.A.* **76** (1979) 5601 – 5605.
7. K. C. Zoon in E. DeMaeyer, G. Galasso, H. Schellekens (eds.): *The Biology of the Interferon System 1981,* Elsevier/North-Holland Biomedical Press, Amsterdam 1981, pp. 47 – 55.
8. M. Streuli, S. Nagata, C. Weissmann, *Science* **209** (1980) 1343 – 1349.
9. R. J. Spiegel in [3]pp. 43 – 50.
10. T. L. Nagabhushan et al. in K. C. Zoon, P. D. Noguchi, T. Y. Liu (eds.): *Interferon: Research, Clinical Application and Regulatory Consideration,* Elsevier, New York 1984, pp. 79 – 88.
11. C. L. Bennett, N. J. Vogelzang, M. J. Ratain, S. D. Reich, *Cancer Treat. Rep.* **70** (1986) 1081 – 1084.
12. K. Foon et al., *Cancer Immunol. Immunother.* **20** (1985) 193 – 197.
13. E. S. Kleinerman et al., *Cancer Res.* **46** (1986) 5401 –5405.
14. R. Kurzrock et al., *Cancer Res.* **45** (1985) 2866 – 2872.
15. J. Rinehart, L. Malspeis, D. Young, J. Neidhart, *Cancer Res.* **46** (1986) 5364 – 5367.
16. E. C. Bordon et al., *Cancer Res.* **45** (1985) 5914 – 5920.
17. L. H. Kronenberg in P. N. Cheremisinoff, R. P. Quellette (eds.): *Biotechnology,* Technomic Publishing Co., Lancaster 1985, pp. 451 – 462.
18. S. Pestka, J. H. Langer, K. C. Zoon, C. E. Samuel in C. C. Richardson, P. D. Boyer, I. B. David, A. Meister (eds.): *Annual Reviews of Biochemistry,* **vol. 56,** Annual Reviews, Palo Alto 1987, pp. 727 – 777.
19. S. Pestka, *Arch. Biochem. Biophys.* **221** (1983) 1 –37.
20. S. Baron, F. Dianzani, G. J. Stanton (eds.): *The Interferon System: A Review to 1982,* Univ. of Texas Medical Branch Publishers, Galveston 1982, Part 1. Reference 20, Part 2.
21. S. Baron, F. Dianzani, G. J. Stanton (eds.): *The Interferon System: A Review to 1982,* Univ. of Texas Medical Branch Publishers, Galveston 1982, Part 2.

22. S. Baron, F. Dianzani, G. J. Stanton, W. R. Fleischmann, Jr. (eds.): *The Interferon System-A Current Review to 1987,* Univ. of Texas Press, Austin 1987.
23. P. E. Came, W. A. Carter: *Interferons and Their Applications,* Springer Verlag, New York 1984.
24. E. DeMaeyer, H. Schellekins (eds.): *The Biology of the Interferon System 1983,* Elsevier, New York 1983.
25. R. M. Friedman, T. M. Merigan, T. Sreevalson: *Interferons as Cell Growth Inhibitors and Antitumor Factors,* Alan R. Liss, New York 1986.
26. I. Gresser (ed.): *Interferon 1-1979,* Academic Press, London-New York 1979.
27. I. Gresser (ed.): *Interferon 2-1980,* Academic Press, London-New York 1980.
28. I. Gresser (ed.): *Interferon 3-1981,* Academic Press, London-New York 1981.
29. I. Gresser (ed.): *Interferon 4-1982,* Academic Press, London-New York 1982.
30. I. Gresser (ed.): *Interferon 5-1983,* Academic Press, London-New York 1983.
31. I. Gresser (ed.): *Interferon 6-1985,* Academic Press, London-New York 1985.
32. I. Gresser (ed.): *Interferon 7-1986,* Academic Press, London-New York 1986.
33. H. Kirchner, H. Schellekens (eds.): *The Biology of the Interferon System 1984,* Elsevier, New York 1985.
34. T. C. Merigan, R. M. Friedman (eds.): *Interferons,* Academic Press, New York 1982.
35. S. Pestka (ed.): *Methods in Enzymology,* **vol. 78,** Academic Press, New York 1981, Part A.
36. S. Pestka (ed.), *Methods in Enzymology,* **vol. 79,** Academic Press, New York 1981, Part B.
37. W. E. Stewart II: *The Interferon System,* 2nd ed., Springer-Verlag, New York 1981.
38. W. E. Stewart II, H. Schellekins (eds.): *The Biology of the Interferon System 1985,* Elsevier, Amsterdam-New York 1986.
39. K. C. Zoon, P. D. Noguchi, T.-Y. Liu (eds.): *Interferon: Research, Clinical Application and Regulatory Consideration,* Elsevier Science Publishers, New York 1984.
40. A. Billiau (ed.): *Interferon 1. General and Applied Aspects,* Elsevier, Amsterdam 1984.
41. J. Vilcek, E. DeMaeyer (eds.): *Interferon 2. Interferons and the Immune System,* Elsevier, Amsterdam 1984.
42. R. M. Friedman (ed.): *Interferon 3. Mechanisms of Production and Action,* Elsevier, Amsterdam 1984.
43. N. B. Finter, R. K. Oldham (eds.): *Interferon 4. In Vivo and Clinical Studies,* Elsevier, Amsterdam 1985.
44. J. Taylor-Papadimitriou (ed.): *Interferons. Their Impact in Biology and Medicine,* Oxford University Press, Oxford 1985.
45. K. Paucker, K. Cantell, W. Henle, *Virology* **17** (1962) 324 – 334.
46. P. Lindahl, P. Leary, I. Gresser, *Proc. Natl. Acad. Sci. U.S.A.* **69** (1972) 721 – 726.
47. J. Taylor-Papadimitriou in A. Billiau (ed.): *Interferon 1. General and Applied Aspect,* Elsevier, Amsterdam 1984, pp. 139 – 166.
48. D. A. Eppstein, Y. V. Marsh, *Biochem. Biophys. Res. Commun.* **120** (1984) 66 – 73.
49. E. Lvovsky et al., *J. Natl. Cancer Inst.* **66** (1981) 1013 – 1019.
50. N. B. Lydon et al., *Biochemistry* **24** (1985) 4131 – 4141.
51. S. E. Salmon et al., *J. Clin. Oncol.* **1** (1983) 217 – 225.
52. I. Gresser, C. Maury, M. Tovey, *Eur. J. Cancer* **14** (1978) 97 – 99.
53. M. Tozawa et al., *Cancer Treat. Rep.* **66** (1982) 1575 – 1577.
54. M. S. Aapro, D. S. Alberts, S. E. Salmon, *Cancer Chemother. Pharmacol.* **10** (1983) 161 – 166.
55. C. E. Welander, T. M. Morgan, H. D. Homesley, P. P. Trotta et al., *Int. J. Cancer* **35** (1985) 721 – 729.
56. J. L. Taylor, J. L. Sabran, S. E. Grossberg, *Eur. J. Immunol.* **10** (1980) 3.
57. P. B. Fisher, R. H. Mufson, I. B. Weinstein, *Biochem. Biophys, Res. Commun.* **100** (1981) 823 – 830.
58. E. Lotzova, C. A. Saavary, J. U. Gutterman, E. M. Hersh, *Cancer Res.* **42** (1982) 2480 – 2488.
59. A. E. Maluish et al., *J. Immunol.* **131** (1983) 503 – 507.
60. K. C. Zoon, H. Arnheiter, *Pharmacol. Ther.* **24** (1984) 259 – 278.
61. T. Sreevalson in R. M. Friedman (ed.): *Interferon-3. Mechanisms of Production and Action,* Elsevier, Amsterdam 1984, pp. 343 – 387.
62. M. E. Hokland, *Acta Pathol. Microbiol. Immunol. Scand. ,[Suppl.] 93* (1985) 4 – 35.
63. K. Stoh, M. Inoue, S. Kataoka, K. Kumagai, *J. Immunol.* **124** (1980) 2589 – 2595.
64. B. R. G. Williams, R. H. Silverman (eds.): *The 2-5 A System: Molecular and Clinical Aspects of the Interferon-Regulated Pathway,* Alan R. Liss, New York 1985.

65. B. Lebleu, G. C. Sen, S. Shaula, B. Carrer, *Proc. Natl. Acad. Sci. U.S.A.* **73** (1976) 3107 – 3111.
66. M. Streuli et al., *Proc. Natl. Acad. Sci. U.S.A.* **78** (1981) 2848 – 2852.
67. T. Taniguchi, S. Ohno, Y. Fuji-Kuriyama, M. Muramatsu, *Gene* **10** (1980) 11 – 15.
68. A. Zilberstein, R. Ruggieri, J. H. Horn, M. Revel, *EMBO J.* **5** (1986) 2529 – 2537.
69. P. W. Gray, D. V. Goeddel, *Nature (London)* **298** (1982) 859 – 863.
70. S. Vijay-Kumar et al., "Crystallization and Preliminary X-ray Investigation of a Recombinant Form of Human Gamma Interferon," *J. Biol. Chem.* **262** (1987) 4804 – 4805.
71. K. E. Mogensen, K. Cantell, *Pharmacol. Ther.* **1** (1977) 369 – 383.
72. P. P. Trotta, H. V. Le, B. Sharma, T. L. Nagabhushan in G. Pierce (ed.): *Developments in Industrial Microbiology,* **vol. 27,** Elsevier, Amsterdam 1987, pp. 53 – 64.
73. P. Leibowitz, M. Ryan, *Eur. Patent Publication No. 146* (1985) 903.
74. Schering Corp., US 4 315 852, 1982 (P. J. Leibowitz, M. J. Weinstein).
75. M. Bier et al.: "New Developments in Isoelectric Focusing," in E. Gross, J. Meienhofer (eds.): *Peptides: Structures and Biological Function*, Pierce Chem. Co., Rockford 1979, pp. 79 – 89.
76. T. L. Nagabhushan, B. Sharma, P. P. Trotta, *Electrophoresis* **7** (1986) 552 – 557.
77. S. Rubinstein, P. C. Fameletti, S. Pestka, *J. Virol.* **37** (1981) 755 – 758.
78. K. Weber, M. Osborn, *J. Biol. Chem.* **244** (1969). 4406 – 4412
79. U. K. Laemmli, *Nature (London)* **227** (1970) 680 – 685.
80. M. Rubinstein, S. Chen-Kiang, S. Stein, S. Odenfriend, *Anal. Biochem.* **111** (1981) 184 – 188.
81. M. K. Radjai, R. T. Hatch, *J. Chromatogr.* **196** (1980) 319 – 322.
82. P. Edman, *Acta Chem. Scand.* **4** (1956) 761 – 768. P. Edman, G. Begg, *Eur. J. Biochem.* **1** (1967) 80 – 91.
83. J. Levin, F. B. Bang, *Bull. Johns Hopkins Hosp.* **115** (1975) 265 – 274.
84. R. J. Spiegel in K. Hellman (ed.): *Cancer Treatment Reviews,* **vol. 12,** Academic Press, New York 1985, supplement B, pp. 5 – 16.
85. R. J. Spiegel et al., *Am. J. Med.* **80** (1986) 223 – 228.
86. R. R. Quaseda et al., *J. Clin. Oncol.* **3** (1985) 1522 –1528.
87. M. Hawkins et al., (eds.): *The Biology of the Interferon System 1984,* Elsevier, New York 1985, pp. 503 – 508.

Iodine and Iodine Compounds

PHYLLIS A. LYDAY, US Bureau of Mines, Washington, DC 20241, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Occurrence** 3
5. **Production** 4
6. **Quality Specifications, Transportation, and Storage** 5
7. **Uses** 5
8. **Economic Aspects** 7
9. **Inorganic Iodine Compounds** 8
10. **Organic Iodine Compounds** 10
11. **Toxicology and Occupational Health** . 10
12. **References** 12

1. Introduction

Iodine, I [*7553-56-2*], atomic mass 129.9044, atomic number 53, a nonmetallic element of the halogen family, appears in group 17 of the periodic table between bromine and astatine. The electronic configuration of the iodine atom is [Kr] $4d^{10}\,5s^2\,5p^5$. The relative atomic mass of the only stable isotope of iodine is 127. There are 22 artificial isotopes with masses between 117 and 139, and fourteen of these yield significant radiation, ^{122}I through ^{123}I and ^{126}I through ^{136}I (→Radionuclides). The half-lifes of the isotopes vary; the shortest, ^{136}I, has a half-life of 86 s and the longest, ^{129}I, has a half-life of 1.6×10^7 a. The isotope ^{129}I is a product of the fission of ^{235}U with slow neutrons. Because of their use in radioactive tracer work and radiotherapy, the most important radioisotopes are ^{131}I, which has a half-life of 8 d, and ^{125}I, which has a half-life of 60 d.

Iodine was first discovered by COURTOIS in 1811. While cleaning kelp pans with hot sulfuric acid, he noticed the evolution of a violet vapor that condensed as crystals on the cooler parts of the vessel. COURTOIS provided samples of this unknown material to other scientists, and in 1813 DAVY and GAY-LUSSAC identified it as a new element. GAY-LUSSAC named the substance "iode" after the Greek word for violet-colored. In 1814 GAY-LUSSAC published the results of his investigations in "Mémoire sur l'Iode".

2. Physical Properties

Iodine is the first member of the halogen family that is solid at ordinary temperature. Iodine vapor exhibits fluorescence at low pressure, emitting a yellowish-green light when illuminated by sunlight or the light from an electric arc. It also rotates the plane of polarized light when placed in a magnetic field between crossed Nicol prisms, and is thermoluminescent above 500 °C. Some physical properties of iodine are as follows:

Atomic radius	133.1 pm
Covalent radius	136 pm
Ionic radius	220 pm
Electronegativity (Pauling scale)	2.66
Electron affinity	301.9 kJ/mol
Ionization potential	10.451 eV
Solid iodine	
mp	113.6 °C
Density at 101.3 kPa	
d_4^{20}	4.93 g/cm^3
d_4^{60}	4.866 g/cm^3
Refractive index, n_D^{20}	3.34
Vapor pressure	
at 25 °C	0.031 kPa
at 113.6 °C	9.17 kPa
Electrical resistivity	
at 25 °C	5.85×10^6 Ωcm
at 110 °C	8.33×10^5 Ωcm
Thermal conductivity	0.421 W m^{-1}K^{-1}
Dielectric constant at 25 °C	10.3
Entropy at 298.2 K	116.81 JK^{-1}mol^{-1}
Heat of fusion at *mp*	62.17 J/g
Heat of sublimation at *mp*	238.4 J/g
Specific heat, J g^{-1}K^{-1}	0.21163 + 19.615
at 25 – 113.6 °C	$\times10^{-5}\,t$
Liquid iodine	
bp	184.35 °C
Critical temperature	546 °C
Critical pressure	11.6 MPa
Critical compressibility factor, Z_c	0.268
Density	
d^{120}	3.960 g/cm^3
d^{180}	3.736 g/cm^3

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_381

Electrical resistivity at 140 °C	1.1×10^5 Ωcm
Kinematic viscosity	
at 116 °C	0.5727 mm^2/s (=cSt)
at 184 °C	0.3785 mm^2/s (=cSt)
Dynamic viscosity	
at 116 °C	2.268 mPa · s
at 184 °C	1.414 mPa · s
Dielectric constant at 118 °C	11.08
Heat of vaporization at *bp*	164.46 J/g
Specific heat at 113.6 – 184 °C	0.3163 J $g^{-1}K^{-1}$
Gaseous iodine	
Vapor density	
at 101.3 MPa, 185 °C	6.75 g/L
Entropy at 298.2 K	260.63 $JK^{-1}mol^{-1}$
Specific heat 25 – 1200 °C	0.1464 J $g^{-1}K^{-1}$

The solubilities of iodine in water and organic solvents and the color of the solutions are given in Table 1.

Table 1. Solubility of iodine in various solvents at 25 °C

Solvent	Solubility, g/kg	Color
Benzene	164.0	red
Butan-2-ol	97	brown
Carbon disulfide	197.0	red
Carbon tetrachloride	19.2	violet
Chloroform	49.7	violet
Ethanol	271.7	brown
Ether	337.3	brown
Ethyl acetate	157.0	brown
Hexane	13.2	violet
Toluene	182.5	red
Water	0.34	brown

3. Chemical Properties

The principal oxidation states of iodine are −1, +1, +3, +5, and +7, but an oxide of formula IO_2 with a formal oxidation state of +4 also exists. Except for the +4 state, thermodynamically stable compounds are formed in all these oxidation states. Iodine is the least electronegative of the halogens and is usually less violent in its reactions than fluorine, chlorine, or bromine. It exhibits mild oxidizing properties in acid solution. Iodine is a stronger oxidizing agent than astatine, the fifth halogen. A basic chemical characteristic of the halogens is the formation of a salt by direct reaction with a metal. Iodine reacts slowly with those metals that form nonvolatile iodides, for example, magnesium, especially if they are not finely divided.

Iodine forms compounds with all the elements except the noble gases, sulfur, and selenium, but only reacts indirectly with carbon, nitrogen, oxygen, and some of the noble metals. Unlike chlorine, iodine does not form compounds with carbon monoxide, nitric oxide, or sulfur dioxide. Ferric and cupric salts, and compounds of vanadium, chromium, and manganese in their highest oxidation states are reduced in acid solution by iodide ion, with the liberation of free iodine.

Reaction with Organic Compounds. Iodine reacts with hydrocarbons but the equilibria are often unfavorable because hydrogen iodide is liberated, which may reduce the alkyl iodide to the hydrocarbon with formation of iodine. Iodination can sometimes be made to go to completion by preventing the formation of free hydrogen iodide. The following three methods are commonly used:

1) Addition of an oxidizing agent, such as nitric acid, iodic acid, or a peroxide
2) Neutralization of hydrogen iodide with a base
3) Combination with mercuric salts to form mercuric iodide or an iodomercurate(II) complex [7], [8]

Iodine also adds to carbon – carbon double bonds in compounds such as ethylene, but the reactions often do not go to completion due to their reversibility.

Iodine is used as a catalyst in many organic reactions (e.g., acylation, halogenation, dehydration, isomerization, and pyrolytic decomposition).

Detection of Iodine. Iodine as the free element can be detected by the characteristic blue color given with a starch solution, or by the violet color of its vapor and of solutions in carbon tetrachloride or carbon disulfide. Iodine is also detected by the precipitation of yellow silver iodide, black palladium iodide, or yellow or red mercuric iodide. Iodine can be quantitatively determined in acidic aqueous solutions by titrating with a standard solution of sodium thiosulfate using starch as the indicator. Iodides can also be assayed by the Volhard method in which an excess of a standard silver nitrate solution is added and then backtitrated with a standard

thiocyanate solution with ferric alum as the indicator. Quantities of up to 0.05 µg of iodine can be determined by colorimetric methods.

4. Occurrence

According to the United States Geological Survey, iodine is the forty-seventh most abundant element in the earth's crust, counting the rare earths as a single element. Small quantities of iodine are widely distributed in rocks, soils, underground brines, and seawater.

Minerals. Iodine is found in nature in the following mineral forms: bruggenite, $Ca(IO_3)_2 \cdot H_2O$; dietzeite, $Ca_2(IO_3)_2 \cdot CrO_4$; iodoargyrite, AgI; iodoembolite, Ag (Cl, Br, I); lautarite [*7789-80-2*], $Ca(IO_3)_2$; and marshite [*24401-69-2*], CuI. The most important iodine-containing mineral is lautarite (anhydrous calcium iodate) [9]. Iodate occurs in the minerals bruggenite, dietzeite, and lautarite, in caliche deposits (between 2 and 50 feet below ground level) in the Atacama Desert of Chile's Antofagasta and Tarapaca Provinces. The iodate is believed to have formed from oceanic emissions during the late Tertiary and Quaternary epochs. The atmospheric photochemical oxidation of iodide to iodate is thought to have occurred in the troposphere and at ground level.

Seawater contains approximately 0.05 ppm of iodine [11]. The amount of iodine is related to the salinity. For example, the Red Sea contains 0.06 ppm and the Baltic Sea, 0.01 ppm. Iodine is concentrated by marine organisms, (e.g., sponges and seaweed) and by fish, especially in the thyroid gland and in liver oil, which contain 510 – 69 200 ppm [11]. The brown algae, *Laminaria* and *Fucus*, which grow in temperate zones of the Northern Hemisphere, have the highest iodine content (0.0279 – 0.4535 wt % after drying). The name kelp was originally applied to the ash obtained by burning seaweed, but has been extended to include the seaweed itself, particularly the varieties yielding potash, iodine, and other commercially important products.

Brines. Iodine occurs frequently as potassium or sodium iodide in subsurface brines associated with oil and gas deposits.

In Oklahoma, the United States, iodine occurs in brines with concentrations of 150 – 1200 ppm in the Morrowan formation of the Pennsylvanian age [12]. In California, brines containing 30 – 70 ppm of iodine are found in the Monterey Formation of the Middle Miocene age and in the Repetto Formation of the Pliocene age. In Michigan, brines containing about 30 ppm of iodine occur in the Sylvania Formation of the Devonian age. In playa lakes in the western United States, iodine occurs at a concentration of 14 ppm [13]. In Japan, the major iodine production occurs in the southern Kanto gas field that underlies the Chiba, Tokyo, and Kanagawa Prefectures. The primary iodine reservoirs occur in the Kiwada, Otadai, and Umegase Formations of the Pliocene to middle Pleistocene. In the Miyazaki Prefecture, iodine and natural gas are produced from the late Pliocene Sadowara and Tonogori members of the Miyazaki group. Iodine concentrations in the Sadowara gas field range between 59 and 101 ppm. Iodine is contained in the Tomigusuku Formation of the Shimajiri group of late Miocene to early Pliocene. Iodine concentrations in the southern Okinawa gas field range between 40 and 110 ppm. In the Niigata Prefecture, iodine concentrations range between 7 and 68 ppm in brines in the Niigata and Hachimori gas fields. The Niigata basin is of the Pliocene to Pleistocene age [14]. In Indonesia, iodine occurs with trace amounts of bromine in brines associated with oil. The most important iodine-producing area is the Gujangon anticline of sandstone and diatomaceous marls of the Upper Pliocene, Kailibeng Formation. The brines were reported to contain about 50 ppm of iodine [15]. Information on iodine production in other parts of the world is incomplete, but the former Soviet Union has reported iodine concentrations of 67 ppm in brines associated with oil fields at Lake Bejuk-Schor, near Baku in the Azerbaidzhan area on the Caspian Sea. Brines in the Slavyansko – Troitsko area were reported to have iodine concentrations of 80 – 120 ppm in the Neftechinski oil field near the Black Sea. At the Nebit-Dag iodine – bromine plant in Turkmenistan, iodine is produced from brines that contain 250 g/L of a combination of boron, calcium, magnesium, sodium compounds, and iodine [16].

5. Production

There have been three phases in the history of iodine production:

1) Production from seaweed began in 1817 and continued until 1959.
2) Production as a byproduct of sodium nitrate production began in 1852.
3) Production from brines began in 1854.

Six processes have been used to produce iodine from brines; the three that are currently used are described here in detail. Other, obsolete processes include the nitrate – sulfite and silver iodide methods [16].

Production from Chile Salpeter. Chile is the only location in the world where an iodine-containing deposit is commercially mined and processed. Iodine is present in the deposits as calcium iodate (lautarite and dietzeite). Processing of the deposits began prior to the 1800s. The first iodine plant opened in 1852, but closed during the first year of production. Iodine production did not begin again until about 1868 and has continued ever since. Iodine is obtained as a byproduct of sodium nitrate production.

Crushed ore is leached to give a solution containing sodium nitrate and calcium iodate. After removal of sodium nitrate by precipitation, the iodate-rich mother liquor is treated with aqueous sodium bisulfite to reduce the iodate to iodide as follows [17]:

$$2\,NaIO_3 + 6\,NaHSO_3 \rightarrow 2\,NaI + 3\,Na_2SO_4 + 3\,H_2SO_4$$

Addition of an equivalent quantity of the mother liquor then liberates free iodine.

$$5\,NaI + NaIO_3 + 3\,H_2SO_4 \rightarrow 3\,Na_2SO_4 + 3\,H_2O + 3\,I_2$$

The iodine precipitate is collected by filtration and is then fed into a melting tank where it is indirectly heated with steam to 120 °C. The molten iodine flows to a settler tank for flaking. The flaker consists of two water-cooled cylindrical drums made of Hastelloy C, an alloy which is resistant to corrosion by iodine. The molten iodine flows into the trough formed by the drums and is removed by scraping. The iodine flakes are carried by conveyor belt to be packed in fiber drums lined with double-ply polyethylene.

Production from Brine. In the United States, iodine is produced in Oklahoma by two brine processes. The brines contain ca. 300 ppm of iodine. In the *blowing-out process*, brine is pumped from the production well into a surge tank. The iodide-rich brine is fed into a column, where it is acidified with hydrochloric acid and treated with chlorine to oxidize the iodide to iodine. The iodine is stripped from the brine by steam and sent to a second column, where it is reduced and absorbed in an aqueous solution of hydriodic acid and sulfuric acid. The iodine is reduced to iodide and absorbed into solution. The iodide solution is continuously fed to a storage tank where chlorine gas precipitates crystals of elemental iodine. The iodine is filtered from the solution, and the acid stream is returned to the absorption column. The iodine is purified by melting under concentrated sulfuric acid to remove water and organic contaminants. It is then dried, and packaged in polyethylene-lined cardboard cartons for shipping [12].

The second process is an *adsorption process* that is used with brines from oil field disposal plants. The brines are skimmed in holding tanks to remove suspended oil. Oil and suspended matter are removed by filtration through sand. Hydrochloric acid is added to the brine and chlorine gas to oxidize the iodide to iodine. The brine is then passed through the absorber, where iodine is absorbed, and the waste brine is neutralized and sent to a disposal well. The adsorbent is treated with sodium hydroxide solution that elutes the iodine as sodium iodide. The iodide-rich eluate is sent to a refinery that uses a process similar to the blowing-out process.

In Japan, iodine is produced from underground brines associated with gas fields by a blowing-out process and by two adsorption processes (the ion-exchange and the sloping fluidized-bed processes) [16]. The blowing-out process is similar to that used in the United States.

In the *ion-exchange process*, the resin bed is fixed, and filtration of the brine is necessary to remove any solids before adsorption. The acidified brine is treated with chlorine to liberate iodine, which is adsorbed on an ion-exchange resin (Amberlite) as polyiodides. The iodine is released from the resin in two stages by elution with sodium hydroxide and then with sodium

chloride. After elution, the resulting iodate–iodide solution is acidified with sulfuric acid to form iodine. The iodine is then recovered from the resulting slurry by filtration or centrifugation, and purified by sublimation or by melting under sulfuric acid.

In the *sloping fluidized-bed process*, sodium iodide is also adsorbed on an ion-exchange resin, but in this case the resin is mobile and filtration of the brine is not necessary. The resin is transferred to a recovery plant, and the sodium iodide is removed as hydrogen iodide by elution with sulfuric acid. Iodine is precipitated by addition of chlorine.

The Wengfu Mine in China has reported iodine concentrations of 0.004 %. Calcining at 1050 °C yielded an iodine-rich gas that could be further processed. Iodine has been recovered during process design testing [18].

6. Quality Specifications, Transportation, and Storage

Commercial crude iodine normally has a minimum purity of 99.5 %. Impurities are chiefly water, sulfuric acid, iron, and insoluble materials. The U.S. Pharmacopoeia (U.S.P.) XVII specifies an iodine content not less than 99.8 %. The Committee on Analytical Reagents of the American Chemical Society (ACS) allows a maximum of 0.005 % total bromine and chlorine, and 0.010 % nonvolatile matter.

Crude iodine is packed in fiber drums lined with double-ply polyethylene that contain 25–50 kg. There are no specific transportation, shipping, or safety requirements. As low temperatures reduce vapor pressure and therefore decrease the rate of loss, a cool, well-ventilated storage area is recommended.

Under normal conditions, the rate of loss from intact commercial containers is negligible. Investment and maintenance costs are high for a new iodine plant using underground brines because brine sources usually do not contain high concentrations of iodine and because 1 kg of chlorine is required for the production of 1.4–1.8 kg of iodine. Electrical costs for maintaining the pumps to bring the brine to the surface, for air blowing the iodine, and for reinjecting the brine are major components of maintenance costs. Capital costs include high-quality stainless steels because iodine is highly corrosive. In addition, the brine must be leased from private landowners over large acreages for many years.

7. Uses

Iodine was first exploited commercially in 1916 in the form of potassium iodide as a remedy for goiter. In 1916, tincture of iodine and iodoform, CHI_3, were used as disinfectants for cuts and abrasions, and for sanitation. In 1939, Louis Daguerre published details of his method for making photographs on plated silver by using iodine vapor to form a thin coating of light-sensitive silver iodide crystals. Clinical uses of iodine have evolved from simple tinctures to radiopaque media, iodophors, and batteries in subcutaneously implanted pacemakers [20]. Iodine is also used in a wide variety of specialty chemicals.

Establishing an accurate end-use pattern for iodine is difficult because iodine-containing intermediates are marketed before reaching their ultimate end uses.

Uses include dietary supplements, catalysts, inks and colorants, pharmaceuticals, radiopaque diagnostic media, photographic equipment, sanitary and industrial disinfectants, and stabilizers, as well as the production of motor fuels, smog inhibitors, lubricants, high-purity metals, and iodized salt. Iodine also has applications in cloud seeding; worldwide about 11 000 kg of silver iodide is used in amounts of 10–50 g per cloud seeding.

Catalysts. The major end use of iodine is in catalysts. After the catalytic Monsanto process for producing acetic acid was introduced in the 1960s, 90 % of new acetic acid capacity used the process (→Acetic Acid, Chap. 4.2.). The process involves methanol carbonylation with an iodide-promoted rhodium complex as catalyst [20]. In the final distillation, the acetic acid often contains organic iodides; these are removed by crystallization or distillation [21]. Iodide catalysts, such as titanium tetraiodide [*7720-83-4*] and aluminum iodide [*7784-23-8*], are also significant in the dehydrogenation of butane and butene to butadiene [22], and in the preparation of stereoregular polymers [23].

Stabilizers. Consumption of iodine as a stabilizer in the manufacture of nylon for tire cord and carpets, and for converting rosins, tall oil, and other wood products to more stable forms is the second major end use [24].

Butadiene reacts with iodine and copper cyanide to yield a copper(I) iodide complex of dehydroadiponitrile. Reaction with hydrocyanic acid gives a high yield of adiponitrile, which can be catalytically hydrogenated to give hexamethylenediamine, a precursor of nylon 66 [26]. In the reaction with HCN the starting materials, copper(I) cyanide and iodine, are regenerated.

Animal Feeds. The third area of major iodine consumption is in the form of additives for animal feeds. In 1986, the recommended amount of ethylene diamine dihydroiodide [*5700-49-2*] (EDDI) consumed in animal feeds was lowered from 50 to 10 mg per head per day, in accordance with the U.S. Food and Drug Administration Compliance Policy Guideline (7125.18). Cattle and sheep are fed iodine compounds for the prevention of soft tissue lumpy jaw. The use of iodized salt has significantly reduced the incident of goiter in livestock in certain iodine-deficient geographical regions. In addition, iodized protein in fowl and cattle feed increases yields of eggs and milk. About 25 % of the reported domestic consumption of iodine was in animal feeds, primarily as the compound EDDI, but also as potassium iodide, calcium iodate [*7789-80-2*], and calcium periodate [*7790-28-5*]. Iodine is one of 14 mineral commodities listed by the National Feed Ingredients Association as being used for domestic meat production.

Disinfectant. Iodine is an effective germicide for a wide range of microorganisms. The use of iodine in sanitary preparations, such as disinfectants, is the fourth major area of consumption. Iodine is often used in conjunction with iodine-complexing nonionic surfactants or polymers (iodophors) in disinfectants that are used in dairies, laboratories, and food processing plants, and for the sanitation of dishes in restaurants [26]. Polyvinyl pyrrolidinone – iodine complex [*25655-41-8*], PVP-iodine, is extensively used because of its germicidal, bactericidal, fungicidal, and general disinfecting properties [27] (→Disinfectants). The solution contains about 10 % available iodine and about 5 % inactive iodide [28]. Globaline tablets are used by the army to disinfect water supplies without boiling [29]. Iodine is used as a disinfectant in swimming pools and is an efficient microbiocide. It is more effective than chlorine against the standard fecal indicators. High iodine concentrations are easily maintained and are more stable than chlorine [30]. A study in Florida showed that concentrations up to 5 ppm of iodine in potable water are not deleterious to health and that 1 ppm is sufficient to safely disinfect a water supply with no loss of effectiveness at high pH values [31].

Pharmaceuticals. The fifth major area of iodine consumption is in pharmaceuticals, primarily as radiopaque media, which are injected intravenously or intracerebrally to produce X-ray photographs of the internal structure of the body (→Radiopaque Media). Clinical use includes cardioangiography, cerebral angiography, arteriography, urography, and phlebography. Media listed by the U.S. International Trade Commission include meglumine diatrizoate [*131-49-7*], sodium diatrizoate [*737-31-5*], iopanoic acid [*96-83-3*], iothalamic acid [*2276-90-6*], and sodium tyropanoate [*7246-21-1*]. These substances contain between 47 and 67 wt % iodine. A nonionic contrast medium, *N,N′*-bis-(2,3-dihydroxypropyl-5-[*N*-(2,3-dihydroxypropyl)acetamidol]-2,4,6-triiodoisophthalamide, which contains 46 wt % iodine, has also been marketed [32] . Potassium iodide is widely used as an expectorant in cough medicines. Hydriodic acid and potassium iodide are used in the synthesis of amphetamine, methamphetamine, and ethylamphetamine.

Miscellaneous Uses. One of the oldest industrial uses of iodine is in photography. Up to 7 % of the silver salt in negative emulsions is iodide. Ammonium iodide [*12027-06-4*] is used as a photographic developer. Iodine is also a constituent of some dyes including 4′,5′-diiodofluorescein [*38577-97-8*], rose bengal [*11121-48-5*], and erythrosine [*16423-68-0*].

Erythrosine, certified food color Red No. 3, contains 58 % iodine and is used in carbonated soft drinks, powdered drinks, gelatin desserts, icings, and pet food. 4′,5′-Diiodofluorescein [*38577-97-8*], C.I. Solvent Red 73, is used in dyeing and printing cotton, in printing

half-silk, and in dyeing jute and straw products. Sodium iodate can also be used in the dyeing process as an oxidizing agent. Fluoroiodocarbons, such as $CF_3(CF_2)_7I$ [*507-63-1*] and $CF_3(CF_2)_7CH_2CH_2I$ [*2043-53-0*], are intermediates in the manufacture of oil and water repellents, surfactants, and fire-extinguishing foams [33], [34].

The addition of iodine to aromatic hydrocarbons, such as *n*-butylbenzene [*104-51-8*], results in the formation of charge-transfer complexes that are effective lubricants for some metals, e.g., titanium or steels [35]. Iodine is also used to produce high-purity forms of titanium, silicon, hafnium, and zirconium in the van Arkel process, which involves the formation of volatile iodides [36].

8. Economic Aspects

From 1982 to 1986, world iodine production remained fairly static and amounted to ca. 12 500 t/a (Table 2). In 1986, Japan accounted for 57 % of world production, followed by Chile (23 %), the Soviet Union (15 %), and China (3.5 %); production figures for the United States were withheld.

Table 2. Production of crude iodine [a]

Country	Production, t/a			
	1994[a]	1996[b]	2000[b]	2001[b]
Chile	5 600	5 000	9 800	10 500
China		500	500	500
Indonesia		80	70	70
Japan	7 200	6 200	6 760	6 100
Former Soviet Union		760	570	570
Total	16 700	13 800	19 100	19 600

[a] [44]
[b] [45]

In Japan, iodine is produced entirely as a by-product of natural gas production. When gas and oil prices are low, the import of world supplies of petrochemicals becomes economically viable for Japan and therefore leads to a decrease in production of domestic natural gas and the associated iodine.

Chile has large reserves of iodine associated with sodium nitrate deposits, but until 1986 had not provided capital for expansion or process improvements. In 1987, Chile completed construction of a 150-t/a iodine plant to extract iodine from waste tailings from sodium nitrate production. These tailings still contain ca. 50 % of the original iodine content. This new plant will guarantee Chile a secure source of iodine from the accumulated waste tailings.

In the United States, iodine is produced by IoChem, North American Brine Resources, and Woodward Iodine.

The United States has been dependent upon imported iodine for roughly half of domestic consumption since ca. 1960. United States imports of iodine and potassium iodide according to the exporting country are listed in Table 3.

Table 3. United States imports of iodine and potassium iodide

Product and country	1985		1986	
	Quantity, t	Value, 10^3 $	Quantity, t	Value, 10^3 $
Crude iodine				
Belgium	10	174		
Canada	1	3		
Chile	749	8 105	627	7 622
Japan	1 496	18 479	746	9 576
Total *	2 254	26 761	1 373	17 199
Resublimed iodine				
Federal Republic of Germany		4		3
Japan	185	2 158	1 204	15 530
Mexico	1	13		
Sweden	1	17	1	24
Total *	186	2 193	1 205	15 530
Potassium iodide				
Belgium	1	17		3
Brazil			1	10
Canada			3	37
Federal Republic of Germany	3	30		3
India	5	64	23	279
Italy	1	3	28	5
Japan	19	226	8	111
Switzerland				1
United Kingdom	1	54	1	42
Total *	30	396	37	492
Grand total	2 470	29 350	2 615	33 253

* Data may not add to totals shown because of independent rounding.

U.S. companies began importing large amounts of potassium iodide in 1984, due to increased demand (see Table 3).

Changes in the official price of iodine have in the past been initiated by Japan followed by Chile and the United States. The price of crude iodine in the United States increased from $ 10.40/kg in 1980 to $ 17.40/kg in 1988.

Table 4 summarizes the consumption of crude iodine in the United States according to product; iodine consumption according to use is given in Table 5.

Table 4. Consumption of crude iodine in the United States.

Product	Consumption, t/a	
	1985	1986
Resublimed iodine	87	70
Potassium iodide	488	474
Sodium iodide	66	62
Other inorganic compounds	566	565
Ethylenediamine dihydroiodide	595	461
Other organic compounds	595	698
Total	2346	2329 ∗

∗ Data do not add to total shown because of independent rounding.

Table 5. Estimated consumption of iodine in the United States according to use

Use	Consumption, t/a				
	1981	1982	1983	1984	1985
Animal feed additives	816	726	635	544	499
Catalysts	727	635	816	862	816
Inks and colorants	227	136	181	181	181
Pharmaceuticals	544	317	317	317	363
Photography	181	181	272	272	272
Sanitary uses	635	454	544	544	544
Stabilizers	726	544	726	635	680
Other	181	181	181	181	181
Total	3991	3129	3673	3537	3537

Consumption of iodine in the United States has decreased for of a variety of reasons. Some companies have turned to importing intermediate products or sell products produced in other countries. For example, the majority of radiopaque media are now imported. Furthermore, use of the major animal feed additive, EDDI, has now been limited by the United States Department of Agriculture. In 1987, IoChem began production at a new plant in Oklahoma, in a joint venture with Schering (Federal Republic of Germany). A major portion of all iodine produced at the plant is exported to Schering.

9. Inorganic Iodine Compounds

Iodine forms binary compounds with hydrogen, metals, other halogen elements, and oxygen.

Iodides range from completely ionic to covalent structures; for example, potassium iodide is ionic and titanium tetraiodide is covalent. In general, the ionic iodides are the most soluble of the halides of a given element.

Commercially, iodides are the most important class of iodine compounds. In general, they are less fusible and volatile than chlorides or bromides. At higher temperature, even the more stable iodides show dissociation. Volatile covalent iodides, such as those of aluminum and titanium, form inflammable mixtures with air. Dry nitrogen triiodide, NI_3, formed by reacting iodine with excess ammonia, will explode at ordinary temperature under the slightest shock.

Potassium iodide [*7681-11-0*], KI, M_r 166.02, *mp* 686 °C, forms colorless cubic crystals, which are soluble in water, ethanol, methanol, and acetone. It is used in animal feeds, catalysts, photographic chemicals, and for sanitation. Potassium iodide is produced by reaction of potassium hydroxide with iodine. The product is purified by crystallization from water.

Sodium iodide [*7681-82-5*], NaI, M_r 149.92, *mp* 662 °C, forms colorless cubic crystals which are soluble in water, ethanol, acetone and glycerol. It is used in photography, for the production of organic chemicals, and as an expectorant in cough medicines. Sodium iodide is produced by the addition of sodium hydroxide or sodium carbonate to an acidic iodide solution. The anhydrous product is obtained from the solution by evaporation.

Ammonium iodide [*12027-06-4*], M_r 144.96, forms white tetragonal crystals which are soluble in water, alcohol, glycerol, and methanol. Upon heating, ammonium iodide sublimes with decomposition. It is used as an expectorant and in photographic chemicals. Ammonium iodide is prepared from ammonia, iodine, and hydrogen peroxide; from ammonia and hydrogen iodide; or from ammonium carbonate and hydrogen iodide. To prevent the crystals or solution from turning yellow due to formation of free iodine, up to 1 % of ammonium hypophosphite is added as a stabilizer.

Iodates have the general formula MIO_3, where M is a univalent metal or electropositive radical. Iodates can be prepared by dissolving

iodine in a caustic alkali solution and recovering the iodate by filtration. Iodine can be oxidized with concentrated nitric acid to iodic acid, followed by neutralization with an oxide or hydroxide; oxidation of an alkaline iodide solution with chlorine; or electrolytic oxidation of a neutral iodide solution containing chromate with a platinum anode and an iron cathode in a nondiaphragm cell. Iodates are precipitated when solutions of metal salts react with alkali-metal iodates.

Most iodates are white. The solid metallic iodates are stable and safe to handle in the pure state, but should be kept out of contact with organic substances and other combustible materials because such mixtures can be explosive. The salts of the alkali metals and magnesium are soluble in water.

Iodates and periodates can be determined by addition of excess alkali iodide in an acid solution followed by titration of the liberated iodine with thiosulfate. Periodates can be determined in the presence of iodates by adding potassium iodide to a solution buffered with borax and boric acid, and titrating the liberated iodine with sodium arsenite solution.

Hydrogen iodide [*10034-85-2*], HI, M_r 127.93, *mp* −50.9 °C, *bp* −35.4 °C, is a colorless gas which fumes in moist air. It is the least stable of the common hydrogen halides, and the best reducing agent. It cannot be prepared using strong acids because it is readily oxidized; therefore, it must be prepared from a weak acid and iodine, or by the hydrolysis of certain iodides.

Hydriodic acid is the colorless solution formed when hydrogen iodide gas dissolves in water, and at commercial strength, typically contains 47 % hydrogen iodide. Hydriodic acid is used in the preparation of iodides and is an important reagent in organic chemistry. Dilute hydriodic acid reacts with metals or their hydroxides, carbonates, and other salts to yield metal iodides.

Iodic acid [*7782-68-5*], HIO_3, M_r 175.93, forms colorless rhombohedral crystals, which decompose at 110 °C. It is prepared by oxidizing iodine with nitric acid. It is a strong oxidizing agent, oxidizing iodide to iodine, sulfite to sulfate, and hydrogen sulfide to sulfur. It reacts vigorously with dry carbon, phosphorus, and organic matter.

Polyiodides and polyhalides form when iodine or iodine halides combine with halide salts. Some examples are RbI_3 [*12298-69-0*], $[N(CH_3)_4I_9$ [*3345-37-7*], $[N(C_2H_5)_4IBr_2]$ [*20445-98-1*], and $KICl_4$ [*14323-44-5*]. Polyiodide and polyhalide salts usually form highly colored orthorhombic crystals; however, KIF_6 [*20916-97-6*] is white.

Iodine Halides. Seven iodine halides are known: IF, IF_3, IF_5 [*7783-66-6*], IF_7 [*16921-96-3*], I_2Cl_6 [*865-44-1*], ICl [*7790-99-0*], and IBr [*7789-33-5*]. All are readily formed by direct reaction of the two halogen elements.

10. Organic Iodine Compounds

As with inorganic iodides, organic iodides have higher specific gravities and indices of refraction than the corresponding chloro and bromo derivatives, and higher melting and boiling points. Compounds containing an iodine atom attached to a primary carbon atom are the most stable. Secondary compounds are less stable, and tertiary are the least stable of all. Aliphatic iodo compounds can be prepared by reacting alcohols with a mixture of iodine and red phosphorus. Other methods for preparing aliphatic iodo compounds include: the addition of iodine halides or iodine to olefins; the substitution reactions between an alkyl bromide or chloride and an alkali metal iodide; and the reaction of alcohols with triphenyl phosphite and methyl iodide.

Organic iodo compounds are highly reactive. The alkyl iodides are generally 50 – 100 times more reactive than the corresponding chlorides. Reactivity of unsaturated organic iodides is affected little if iodine is attached to one of the carbon atoms involved in the double bond, but reactivity is promoted if iodine is attached to the carbon atom in the α position to the double bond. Iodine in organic compounds can be determined after being converted into an inorganic form by heating the compound with one of the following: (1) fuming nitric acid in a sealed strong glass tube, the Carius method; (2) sodium peroxide

in a combustion container; or (3) a mixture of chromic and sulfuric acids, followed by addition of phosphoric acid and distillation into a solution of sodium arsenite [37].

Iodoform [*75-47-8*], triiodomethane, CHI_3, *mp* 119 °C, crystallizes from acetone as yellow hexagonal plates, which have a characteristic penetrating odor and are insoluble in water. The compound can be prepared by the reaction of iodine chloride with acetone in the presence of potassium hydroxide, or by reaction of chloroform with methyl iodide. Iodoform is used as a disinfectant.

Methylene iodide [*75-11-6*], CH_2I_2, M_r 267.85, *mp* 5 °C, *bp* 180 °C (101.3 kPa), is a colorless liquid when pure. It has a relative density of 3.325 at 20 °C and can be mixed with benzene (d_{20}^{20} 0.879) to produce liquids with different densities. These mixtures are used in the determination of the refractive indices of minerals and for the separation of minerals according to density.

Ethyl iodide [*75-03-6*], C_2H_5I, M_r 155.98, *bp* 72.3 °C, is a colorless liquid, which is completely miscible with ethanol and ether, but only partially miscible with benzene. Ethyl iodide is prepared by reaction of ethanol with iodine and red phosphorus.

Methyl iodide [*74-88-4*], CH_3I, M_r 141.95 *mp* −66.1 °C, *bp* 42.5 °C, is a colorless, transparent liquid, which turns brown upon exposure to light. Methyl iodide is prepared from methyl alcohol, iodine, and red phosphorus. Methyl iodide is used as a methylating agent.

Iodobenzene, [*591-50-4*], M_r 204.02, *mp* −30 °C, *bp* 188.7 °C, is a colorless liquid that rapidly turns yellow and has a characteristic odor. Iodobenzene is prepared by diazotization of aniline followed by treatment of the solution with an aqueous solution of potassium iodide, or by the action of nitric acid on a benzene and iodine mixture.

11. Toxicology and Occupational Health

Iodine is absorbed by the body and concentrated in the form of diiodotyrosine and triiodothyronine, which form thyroxine that is stored as thyroglobulin. Thyroxine is secreted by the thyroid, enters the circulation, and is carried to the peripheral tissues where it controls tissue metabolism primarily through regulation of enzyme activities (→ Hormones, Chap. 2.1.). Iodine is essential to higher animals and humans. A normal person requires about 75 mg of iodine per year, which is usually consumed as iodized salt that contains one part sodium or potassium iodide to 100 000 parts of sodium chloride. Iodine deficiency is a major cause of goiter. Studies on chickens link autoimmune thyroiditis directly with consumption of large amounts of iodine [38]. These results are also applicable to humans.

Iodine is safer to handle at ambient temperature than chlorine or bromine because it is a solid with a lower vapor pressure. When handling properly packaged containers, usual work clothes are sufficient. In the handling of solid, unpacked iodine, rubber gloves, a rubber apron, and safety goggles are recommended; more drastic situations may necessitate the use of a gas mask. The maximum safe concentration for short-term exposure of up to 1 h is 1.0 ppm. Exposure of the lungs and eyes can be irritating at concentrations of 0.1 ppm and should be avoided. Exposure in concentrations higher than 0.1 ppm for extended periods causes severe irritation to the eyes and the respiratory tract, and may lead to pulmonary edema [39]. The mean lethal oral dose for an adult is 2 – 4 g. Prolonged contact with the skin may be harmful.

Treatment includes irrigating the eyes with water or washing the contaminated body area with a 5 % solution of sodium thiosulfate, followed by saline catharsis. Iodine and concentrated solutions, including U.S.P. tincture, are classed as poisons because of their irritating effects to the gastrointestinal tract. The gastric treatment antidote is a 5 % solution of sodium thiosulfate followed by saline catharsis. Medical attention should be provided.

In the United States, the Occupational Safety and Health Administration (OSHA) has set the maximum concentration of iodine vapor permitted in the working atmosphere as 0.1 ppm [40].

The American Conference of Government and Industrial Hygienists (ACGIH) established 0.1 ppm as the TLV (TWA) for iodine. The MAK value is also 0.1 ppm. To detect iodine at this concentration, air monitoring is necessary [41].

Empty containers may be destroyed in an incinerator or decontaminated by washing with a dilute sodium thiosulfate solution. Bulk waste should be treated by controlled chemical operations to reclaim the iodine or convert it to harmless waste.

Iodine can be mixed safely with ammonia solutions in the presence of a reactive organic substance; however, contact of gaseous ammonia or its solutions with free iodine and its halogen derivatives in any form can form explosive nitrogen triiodide [*13444-85-4*], NI_3. If this dark brown to black solid forms, it should be discarded while wet or destroyed by adding a reducing agent, such as ethanol, or more of the compound to be iodinated.

In general, iodides are acutely toxic only when ingested in large amounts. Chronic ingestion or absorption through the skin may produce iodism, which is manifested by a rash, nasal drip, or headache. In severe cases, weakness, anemia, and general depression may occur. Care should be taken in handling organic iodine compounds.

In 1979 and 1986, nuclear accidents caused the release of radioactive iodine, ^{131}I, into the atmosphere. Ingestion of potassium iodide, KI, can reduce the radiation dose to the thyroid gland by up to 90 % if administered before or shortly after exposure. A dosage of 130 mg was recommended for adults. If administered 4 h after exposure, KI can block 50 % of the uptake of ^{131}I. The United States Federal Emergency Management Agency provided guidelines to state and local agencies in the United States regarding the distribution of KI for use as a thyroidal blocking agent by the general public in the vicinity of nuclear power plants [42]. Guidance on the use of radioprotective substances was delegated to the Food and Drug Administration (FDA). The FDA recommended that KI be used by people who are likely to receive more than 10 – 20 rad to the thyroid [43].

12. References

General References

1. G. L. Clark (ed.): *The Encyclopedia of Chemistry,* 2nd ed., Van Nostrand, New York 1966, pp. 556 – 559.
2. D. M. Considein et al. (ed.): *Van Nostrand's Scientific Encyclopedia,* 6th ed. Van Nostrand, New York 1983, pp. 1550 – 1652.
3. *McGraw-Hill Encyclopedia of Science & Technology,* 5th ed., **vol. 7,** McGraw-Hill, New York 1982, pp. 321 – 325.
4. *Kirk-Othmer,* 3rd ed., **13,** 649 – 676.
5. M. Windholz (ed.): *The Merck Index,* 9th ed., Merck & Co., Rahway, N.J., 1976.
6. C. A. Hampel: *Encyclopedia of Chemical Elements,* Reinhold Book Corp., New York 1968.

Specific References

7. L. Jurd, *Aust. J. Sci. Res., Ser. A:* **2** (1949) 111 – 1169.
8. R. G. Shepherd, C. E. Fellows, *J. Am. Chem. Soc.* **70** (1948) 157 – 160.
9. G. E. Ericksen, *Geol. Surv. Prof. Pap. (U.S.)* (1981) 1188.
10. F. G. Sawyer, M. F. Ohman, F. E. Lusk, *Ind. Eng. Chem. Fundam.* **41** (1949) no. 8, 1547 – 1552.
11. A. Gene Collins: *Geochemistry of Oilfield Waters,* Elsevier Scientific Publishing Co., New York 1975, p. 228.
12. H. M. Cotten in K. S. Johnson, J. A. Russel (eds.): "Iodine in Northwestern Oklahoma," *Thirteenth Annual Forum on the Geology of Industrial Minerals,* Norman, OK 1978 pp. 89 – 94.
13. P. A. Lyday: *Mineral Facts and Problems,* BuMines Bulletin 675, Washington 1985, pp. 377 – 384.
14. O. Fukuta in W. J. Schlitt ed.: Japanese Iodine–Geology and Geochemistry, *Salts & Brines 85,* New York 1985, pp. 151 – 168.
15. R. W. von-Bemmelen: *Geology of Indonesia,* **vol. 2,** The Hague, Netherlands 1949, pp. 103 – 111.
16. P. A. Lyday, *Indus. Min. Mag.* **222,** London 1986, pp. 65, 66, 69, 70, 75, 76.
17. R. D. Crozier, *Min. Mag.* **146** (1982) no. 4, 282 – 289.
18. *World Mining Equipment,* **10** (1986) no. 4, 40 – 44.
19. P. A. Lyday, *J. Chem. Ed.* **64** (1987) 152.

20. F. E. Paulik, *Catal. Rev.* **6** (1972) 49. R. P. Lowry, A. Aguilo, *Hydrocarbon Process.* **53** (1974) no. 11, 103.
21. *Kirk-Othmer,* 3rd ed., **1,** 135.
22. Shell Oil Co., US 3 106 590, 1963 (C. W. Bittner); US 3 119 881, 1964 (R. L. Hodgson).
23. Phillips Petroleum Co. GB 920 244, 1963; GB 931 440, 1963.
24. G and A Laboratories, Inc., US 2 299 577, 1942 (T. Hasselstrom, E. A. Brenhan).
25. Esso Research & Engineering, *J. Chem. Eng.* **78** (1971) 107.
26. R. R. Whitehead, *Water Pollut. Control (Don Mills Can.)* **119** (1981) no. 12, 10 – 12.
27. West Chem. Products, Inc., US 4 271 149, 1981 (M. W. Winicor, M. Oberland).
28. GAF Corp., US 4 128 633, 1978 (D. H. Larenz, E. P. Williams).
29. J. C. Morris et al., *Ind. Eng. Chem.* **45** (1953) 1013 – 1015.
30. M. S. Favero, C. H. Drake, *Public Health Rep.* **79** (1964) no. 3 251 – 257.
31. W. C. Thomas et al. *Trans. Clin. Climatol. Assoc.* **90** (1978) 153 – 162.
32. E. Lindgren, *Acta Radiol. Suppl.* **362** (1980) 134.
33. *Kirk-Othmer,* 2nd ed., **22,** 145 – 146.
34. *Kirk-Othmer,* 2nd ed., **Supplement Volume,** 964 – 973.
35. General Electric Co., US 3 228 880, 1966 (R. S. Owens, R. W. Roberts).
36. R. F. Rolsten: *Iodine Metals and Metal Iodide,* J. Wiley & Sons, New York 1961.
37. J. W. Thomas et al., *Anal. Chem.* **22** (1950) 726 – 747.
38. N. Bagchi et al., *Science (Washington D.C.)* **4723** (1985) 325 – 327.
39. N. I. Sax: *47th Dangerous Properties of Industrial Materials,* 4th ed., Van Nostrand Reinhold Co., New York 1975, p. 32.
40. CFR, 29th, 1910.1000, Table Z – 1, July 1, 1986.
41. *Encyclopaedia of International Occupational Health and Safety,* **vol. 1** , International Labor Office, Geneva 1971.
42. Federal Emergency Management Agency: "Federal Policy on Distribution of Potassium Iodide Around Nuclear Power Sites for Use as a Thyroidal Blocking Agent," *Fed. Regist.* **50** (1985, Jun. 20) no. 119, 25624 – 25625.
43. Federal Emergency Management Agency: "Federal Policy on Distribution of Potassium Iodide Around Nuclear Power Sites for Use as a Thyroidal Blocking Agent," *Fed. Regist.* **50** (1985, Jul. 24) no. 142, 30258 –30259.
44. Büchel, Moretto, Woditsch *Industrial Inorganic Chemistry* Wiley-VCH, Weinheim
45. http://minerals.usgs.gov/minerals/pubs/commodity/iodine

Ion Exchange Membranes → Membranes and Membrane Separations

Ion Exchangers

Francois de Dardel, Rohm and Haas, Paris, France

Thomas V. Arden, Cobham, United Kingdom

1. Introduction 2
2. Structures of Ion-Exchange Resins 4
2.1. Polymer Matrices 4
2.2. Functional Groups 4
2.2.1. Cation-Exchange Resins 4
2.2.2. Anion-Exchange Resins 5
2.2.3. Other Types of Ion-Exchange Resins 6
2.3. Adsorbent Resins and Inert Polymers 7
3. Properties 7
3.1. Degree of Cross-Linking and Porosity 7
3.2. Exchange Capacity 8
3.3. Stability and Service Life 10
3.4. Density 12
3.5. Particle Size 12
3.6. Moisture Content 13
4. Ion-Exchange Reactions 14
4.1. Cation Exchange 14
4.2. Anion Exchange 15
4.3. Cation and Anion Exchange in Water Treatment 15
5. Ion-Exchange Equilibria 15
5.1. Dissociation and pK Value 15
5.2. Mono – Monovalent Exchange . . 17
5.3. Mono – Divalent Exchange (Water Softening) 18
5.4. General Case 20
6. Exchange Kinetics 20
6.1. Principles 20
6.2. Kinetic Curves 22
6.3. Strongly Acidic or Strongly Basic Resins 23
6.3.1. Film Diffusion 23
6.3.2. Particle Diffusion 24
6.4. Weakly Acidic or Weakly Basic Resins 24
7. Practical Results of Ion-Exchange Equilibrium and Kinetics 24
7.1. Operating Capacity, Regeneration Efficiency, and Regenerant Usage 24
7.2. Permanent Leakage 25
7.3. Water Analysis 26
7.4. Calculations in the Design of Ion-Exchange Plants for Water Purification 28
7.5. Example of Calculation 30
7.5.1. Principle 30
7.5.2. Basic Data 30
7.5.3. Demineralization Unit 31
7.5.4. Polishing Unit 32
8. Industrial Use of Ion Exchange . 33
8.1. Description of the Ion-Exchange Cycle 33
8.2. Methods for Overcoming Equilibrium Problems 33
9. Ion-Exchange Resin Combinations 35
9.1. Pretreatment 35
9.2. Softening 37
9.3. Demineralization (Primary System) 37
9.4. Polishing 38
9.5. Choice of Resin 38
10. Plant Design 41
10.1. General Considerations 41
10.2. Fixed-Bed Ion-Exchange Units . 41
10.2.1. Column Diameter 41
10.2.2. Small-Scale Units 42
10.2.3. Industrial Co- and Counterflow Plants 42
10.2.4. Mixed Beds 47
10.2.5. Other Ion-Exchange Polishers . . . 50
10.3. Continuously Circulated Ion-Exchange Resins 51
10.4. External Valves and Pipework . . 51
10.5. Control Systems 52
11. Special Processes in Water Treatment 52
11.1. Removal of Organic Matter . . . 52
11.2. Treatment of Potable Water . . . 53
11.3. Treatment of Brackish Water . . 55
11.4. Processes Involving Sea Water . . 56
11.5. Treatment of Condensates 56
11.5.1. Conventional Resins 57
11.5.2. Powdered Resins 58
11.6. Water Treatment in the Nuclear Industry 58
11.7. Production of Ultrapure Water . 58

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_393

12. Special Applications of Ion Exchange 59
12.1. Processing Steps 59
12.1.1. Purification 61
12.1.2. Ion Substitution 64
12.1.3. Recovery and Concentration 65
12.1.4. Separation 65
12.1.5. Diffusion 67
12.1.6. Catalysis 67
12.1.7. Dehydration 68
12.1.8. Coalescence on Oleophilic Resins 68
12.1.9. Liquid Ion Exchangers 69
12.1.10. Ion-Exchange Membranes 69
12.2. Technical Considerations 70
13. References 72

Abbreviations and Acronyms:

BV bed volume
DG atmospheric degasser
DM dry matter
DVB divinylbenzene
EMA equivalent mineral acidity
eq equivalents
FMA free mineral acidity
FR flow rate
IN inert resin
MHC moisture-holding capacity
R resin
SAC strongly acidic cation exchanger
SBA strongly basic anion exchanger
SBC strongly basic capacity
TAlk total alkalinity
TDS total dissolved solids
TH total hardness
TOC total organic carbon
U.C. uniformity coefficient
WAC weakly acidic cation exchanger
WBA weakly basic anion exchanger

Symbols:

C_P weight capacity, eq/L
C_V volume capacity, eq/L
D diffusion coefficient, mol/cm^3
J ion flux, mol s^{-1} cm^{-2}
k mass-transfer coefficient, cm/s
K equilibrium constant
K_A^B selectivity coeffient between components A and B
N fouling factor
P weight of resin
Q flow rate, m^3/h
r radius of ion-exchange bead
S salinity, meq/L
t operating time, h
V volume of resin, m^3
X mole fraction
α_A^B separation factor between components A and B
γ activity coefficient
δ thickness of Nernst film

1. Introduction

Definition and Principles. In ion exchange, ions of a given charge (either cations or anions) in a solution are adsorbed on a solid material (*the ion exchanger*) and are replaced by equivalent quantities of other ions of the same charge released by the solid.

The ion exchanger may be a salt, acid, or base in solid form that is insoluble in water but hydrated. Exchange reactions take place in the water, retained by the ion exchanger; this is generally termed *swelling water* or *gel water*. The water content of the apparently dry material may constitute more than 50 % of its total mass.

Figure 1 shows the partial structure of a cation exchanger; each positive or negative ion is surrounded by water molecules.

Ion exchange forms the basis of a large number of chemical processes which can be divided into three main categories: substitution, separation, and removal of ions.

1) *Substitution.* A valuable ion (e.g., copper) can be recovered from solution and replaced by a worthless one. Similarly, a toxic ion (e.g., cyanide) can be removed from solution and replaced by a nontoxic ion.
2) *Separation.* A solution containing a number of different ions passes through a column containing beads of an ion-exchange resin. The ions are separated and emerge in order of their increasing affinity for the resin.
3) *Removal.* By using a combination of a cation resin (in the H^+ form) and an anion resin (in

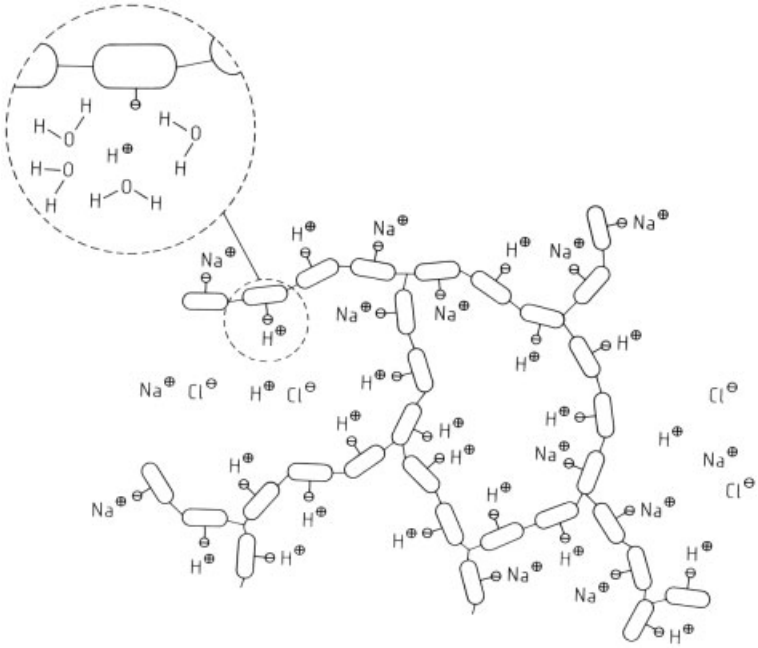

Figure 1. Structure of a cation exchanger that exchanges H^+ for Na^+ ions
Swelling water is represented in the inset.

the OH^- form), all ions are removed and replaced by water (H^+OH^-). The solution is thus demineralized.

Historical Aspects. The discovery of ion exchange dates from the middle of the nineteenth century when THOMSON [1] and WAY [2] noticed that ammonium sulfate was transformed into calcium sulfate after percolation through a tube filled with soil.

In 1905, GANS [3] softened water for the first time by passing it through a column of sodium aluminosilicate that could be regenerated with sodium chloride solution. In 1935, LIEBKNECHT [4] and SMIT [5] discovered that certain types of coal could be sulfonated to give a chemically and mechanically stable cation exchanger. In addition, ADAMS and HOLMES [6] produced the first synthetic cation and anion exchangers by polycondensation of phenol with formaldehyde and a polyamine, respectively. Demineralization then became possible. At present, aluminosilicates and phenol – formaldehyde resins are reserved for special applications and sulfonated coal has been replaced by sulfonated polystyrene.

Polystyrene Resins. The first polystyrene-based resin was invented by D'ALELIO in 1944 [7]. Two years later, MCBURNEY produced polystyrene anion-exchange resins by chloromethylation and amination of the matrix [8].

The anion exchangers known until then were weakly basic and took up only strong mineral acids. The new resins produced by the McBurney process were stronger bases and could adsorb weak acids such as carbon dioxide or silica, allowing complete demineralization of water with a purity previously obtainable only by multiple distillation in platinum. Even today, ion exchange is still the only process capable of producing the water quality needed for high-pressure boilers. Reverse osmosis and electrodialysis can demineralize solutions with 50 – 90 % efficiency. Only ion exchange can "polish" the predemineralized solution with a demineralization efficiency of 99 – 99.99 %

Macroporous Resins. Two of the problems encountered in the use of ion-exchange resins are the fouling of the resin by natural organic acids present in surface waters and the mechanical stress imposed by plants operating at high

flow rates. To cope with these, three manufacturers [9–11] invented resins with a high degree of cross-linking but containing artificial open pores in the form of channels with diameters up to 150 nm that can adsorb large molecules. Resins in which the polymer is artificially expanded by the addition of a nonpolymerizable compound that is soluble in the monomer are known as *macroporous* or *macroreticular resins* (see Section 3.1). Other naturally porous resins are known as *gel resins.*

Polyacrylic Anion Exchangers. Between 1970 and 1972, a new type of anion-exchange resin with a polyacrylic matrix appeared on the market. This possesses exceptional resistance to organic fouling and a very high mechanical stability due to the elasticity of the polymer.

Uniform Size Resins. In the 1980s and 1990s, several producers developed new manufacturing technologies aimed at producing resins with particles of almost identical size.

2. Structures of Ion-Exchange Resins

An ion exchanger consists of the polymer matrix and the functional groups that interact with the ions. This article deals only with organic ion exchangers; inorganic ion exchangers are of minor importance and are primarily layer silicates and zeolites (→ Silicates, → Zeolites).

2.1. Polymer Matrices

Polystyrene Matrix. (→ Polystyrene and Styrene Copolymers). The polymerization of styrene [*100-42-5*] (vinylbenzene) under the influence of a catalyst (usually an organic peroxide) yields linear polystyrene [*9003-53-6*]. Linear polystyrene is a clear moldable plastic which is soluble in certain solvents (e.g., styrene or toluene) and has a well-defined softening point. If a proportion of divinylbenzene is mixed with styrene, the resultant polymer becomes cross-linked and is then completely insoluble.

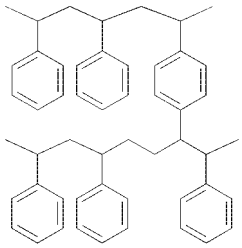

Cross-linked polystyrene

In the manufacture of ion-exchange resins, polymerization generally occurs in suspension. Monomer droplets are formed in water and, upon completion of the polymerization process, become hard spherical beads of the polymer.

Polyacrylic Matrix. Matrices for ion exchangers can also be obtained by polymerizing an acrylate, a methacrylate, or an acrylonitrile, any of which can be cross-linked with divinylbenzene [*105-06-6*], DVB [→ Polyacrylamides and Poly(Acrylic Acids); → Polyacrylates].

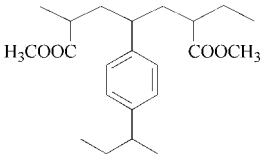

Cross-linked polymethacrylate

Other Types of matrix. Other types of matrix include

1) *Phenol–formaldehyde* resins (→ Phenolic Resins) which show interesting adsorption properties
2) *Polyalkylamine resins*, obtained from polyamines by condensation with epichlorohydrin, which gives an anion exchanger directly in a single step.

2.2. Functional Groups

2.2.1. Cation-Exchange Resins

Cation-exchange resins in current use can be separated into two classes according to their active groups:

1) Strongly acidic (sulfonic groups)
2) Weakly acidic (carboxylic groups)

Strongly Acidic Cation-Exchange Resins. Chemically inert polystyrene beads are treated with concentrated sulfuric or chlorosulfonic acid to give cross-linked polystyrene 3-sulfonic acid. This material is the most widely used cation-exchange resin and is strongly acidic .

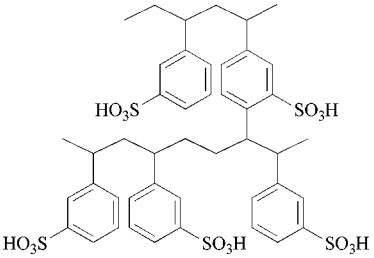

Cross-linked polystyrene 3-sulfonic acid

Examples: Amberlite IR 120, Dowex HCR, Lewatit S 100.

Weakly Acidic Carboxylic Cation-Exchange Resins. The weakly acidic resins are almost always obtained by hydrolysis of polymethylacrylate or polyacrylonitrile to give a poly(acrylic acid) matrix.

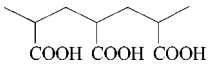

Examples: Amberlite IRC 86, Lewatit CNP.

2.2.2. Anion-Exchange Resins

Polystyrene Materials. Cross-linked polystyrene beads are treated with chloromethyl methyl ether under anhydrous conditions, with either aluminum chloride or tin(IV) chloride as catalyst. Chloromethylated polystyrene is obtained:

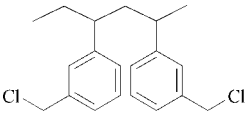

In a second stage, the chlorine in the chloromethylated group can be replaced by an amine or even by ammonia. Depending on the reaction selected, the anion exchanger obtained may be strongly to weakly basic. The degree of basicity can be "made to measure" because of the large number of amines available. The anion exchangers listed below are arranged in order of decreasing basicity:

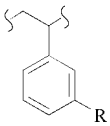

where R can be

$-CH_2N^+(CH_3)_3Cl^-$ (type 1 resin)	e.g., Amberlite IRA402
$-CH_2N^+(CH_3)_2CH_2CH_2OHCl^-$ (type 2 resin)	e.g., Amberlite IRA410
$-CH_2N(CH_3)_2$	e.g., Amberlite IRA96

Resins with quaternary ammonium groups are strongly basic. Those with benzyltrimethylammonium groups are known as type 1 and are the most strongly basic, whereas those with benzyldimethylethanolammonium groups are known as type 2 and are slightly less basic.

Type 1 resins are used when total removal of anions, even those of weak acids (including silica), is essential. Type 2 resins are also basic enough to remove all anions, but they release the anions more easily during regeneration with caustic soda; as a result, they have a high exchange capacity and a better regeneration efficiency (see Section 7.1). Unfortunately, they are chemically less stable and produce greater silica leakage than type 1 resins.

Resins whose active group is an amine are generally denoted as weakly basic, although their basicity may vary considerably. Tertiary amines are sometimes called medium-base or intermediate-base resins, whereas primary amines are very weakly basic and are rarely used.

The most widely used weakly basic resins contain tertiary amino groups and adsorb any strong acids present in the solution to be treated but do not affect neutral salts or weak acids.

Manufacturers do not always indicate the chemical structure of their exchangers in their literature. Care should therefore be taken not to assume that resins are chemically identical merely because they have similar general characteristics.

Secondary and Tertiary Cross-Linking. During chloromethylation, a side reaction may occur in which the chloromethyl group of a chloromethylated benzene ring reacts with an unconverted ring, to yield a methylene bridge. These bridges form additional cross-links in the polystyrene matrix:

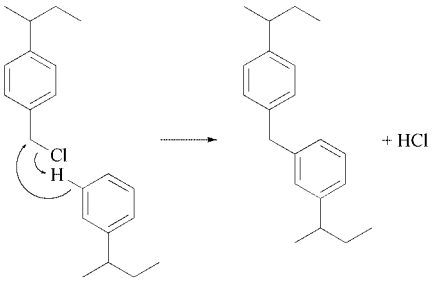

The amount of this secondary cross-linking can be adjusted by varying the conditions (quantity and type of catalyst, temperature) of the chloromethylation reaction. Most strongly basic and weakly basic polystyrene resins have some degree of secondary cross-linking.

Furthermore, during the amination of weakly basic resins, another type of cross-linking may be produced. This is called tertiary cross-linking and yields strongly basic quaternary groups in addition to the weakly basic tertiary groups.

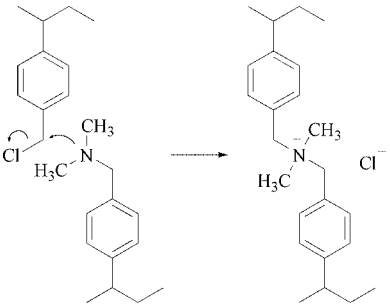

Polyacrylic Resins. Polyacrylic resins are manufactured in a manner analogous to that used for polystyrene resins. Beads are prepared from an acryclic ester copolymerized with divinylbenzene by using suspension polymerization and free-radical catalysis. The polyacrylate formed is then given active groups by reaction with a polyfunctional amine containing at least one primary amino group and one secondary or, more frequently, tertiary amino group. The primary amino group reacts with the polyester to form an amide, whereas the secondary or tertiary amino group forms the active group of the anion exchanger. This method always yields a weakly basic exchanger, which can be further treated with chloromethane or dimethyl sulfate to give a quaternary strongly basic resin:

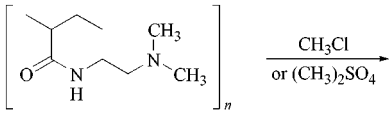

Tertiary amine
(weakly basic, e.g.,
Amberlite IRA 67)

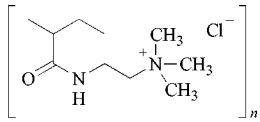

Quaternary amine
(strongly basic, e.g.,
Amberlite IRA 458)

In principle, a wide range of anion-exchange resins can be obtained by varying the type of ester chosen as the starting material and the polyamine used for activation. In practice, the range is limited by the availability and cost of raw materials.

2.2.3. Other Types of Ion-Exchange Resins

By using polymerization and activation methods analogous to those described above, a wide variety of functional groups can be grafted onto a given polymer. Some of these groups can be used for selective uptake of ions, principally metals (Table 1).

The thiol group forms very stable bonds with certain metals, particularly mercury. The iminodiacetic, aminophosphonic, and amidoxime groups form metal complexes whose stability depends mainly on the pH of the solution. Selective adsorption of certain metals can thus be achieved by varying the pH. These types of material are known as *chelating* or *complexing resins.*

The *N*-methylglucamino group is used to make resins specific for boric acid, which is taken up as a complex.

Table 1. Principal active groups of ion exchangers used for selective uptake of metals

Active group	Formula *	Example
Thiol	$-SH$	Imac TMR
Iminodiacetic acid	$-CH_2N(CH_2COOH)_2$	Lewatit TP 207
Aminophosphonic acid	$-CH_2NHCH_2CH_2PO_3H$	Duolite C 467
Amidoxime	$-C(=NOH)NH_2$	Duolite ES 346
N-Methylglucamine	$-CH_2N(CH_3)-(CHOH)_4CH_2OH$	Amberlite IRA 743

* The active groups are substituents (R) of polystyrene with the following formula:

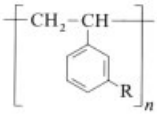

2.3. Adsorbent Resins and Inert Polymers

Strictly speaking, adsorbent resins are not ion exchangers but resemble them very closely. They have a high porosity and are used for the adsorption of nonionic or weakly ionized species as a complement to ion exchange. They may have cation- or anion-exchange groups or no ion-exchange groups at all. The latter are ionically inert. In order of decreasing polarity, adsorbent resins can be classified in the following manner:

1) *Ionized adsorbents* are strongly basic exchangers used in chloride form for color removal from sugar juices or as "organic scavengers" (see Section 11.1, e.g., Amberlite IRA958).
2) *Phenolic adsorbents* contain weakly basic amine and phenolic groups or phenolic groups, only. They are used to remove color bodies (colored impurities) from solutions of organic acids and food-processing streams (e.g., Duolite A561, XAD761).
3) *Inert adsorbents* are macroporous copolymers of styrene and divinylbenzene with a very high degree of cross-linking and a large surface-to-volume ratio. These resins are used to remove organic, weakly ionized, or nonionic substances, such as phenols, chlorinated solvents, antibiotics, and complexing agents, from aqueous or organic solutions (e.g., Amberlite XAD4, Diaion HP20).

Inert polymers without measurable porosity and without active groups can be used either to separate two resin layers or to keep a resin separate from a collector system.

3. Properties

3.1. Degree of Cross-Linking and Porosity

An increase in the degree of cross-linking (i.e., the weight percentage of DVB related to the total amount of monomer prior to polymerization) produces harder, less elastic resins. Resins with higher degrees of cross-linking show more resistance to oxidizing conditions that tend to decross-link the polymer. Above 10–12 % DVB, however, the structure becomes too hard and dense. Activation (i.e., chemical transformation of the inert copolymer into an ion-exchange resin) becomes difficult because access to the interior of the bead is hindered by the high density of the matrix. In addition, osmotic stress cannot be absorbed by the elasticity of the structure, which causes the bead to shatter. Finally, the rate of exchange increases in proportion to the mobility of the ions inside the exchanger bead: if the structure is too dense, ionic motion is slowed down, thus reducing the operating capacity of the resin.

For sulfonic resins, maximum operating capacity (Section 3.2) is obtained with approximately 8 % DVB.

Cross-Linking and Affinity. The greater the ionic mobility in the resin, the poorer is the differentiation between the adsorption of ionic species with the same charge. Consequently, the degree of cross-linking in the resin must be increased when greater differences in ionic affinity are required.

In water treatment, the sulfonated polystyrene resins usually have a DVB concentration of ca. 8 %. Resins with a slightly higher degree of cross-linking (10 – 12 %) are sometimes used to increase the retention of mineral ions when water of very high purity is being produced. Resins with slightly lower levels of cross-linking (5 – 7 %) may be chosen when easier desorption and, hence, better regeneration efficiency are required, especially in water softening.

Nonuniformity in the Matrix. Cross-linking reduces the retention of water in ion-exchange resins (Section 3.6). The volume occupied by this water is a measure of the resin's porosity. Cross-linking is not uniform because the DVB – DVB reaction is more rapid than that between DVB and styrene. Polymerization begins to occur around the catalyst molecules, and polymer growth is faster at sites rich in DVB than at those rich in styrene. Material with an average of 8 % DVB may contain local microscopic regions with more than 20 % DVB, whereas other regions may have less than 4 %.

Macroporous resins are made by mixing the monomers with a compound (e.g., heptane, saturated fatty acids, $C_4 - C_{10}$ alcohols or polyalcohols, or low molecular mass linear polystyrene) which expands the resin. The substance does not itself polymerize and, thus, although it acts as a solvent for the monomers, it causes the polymer to precipitate from the liquid.

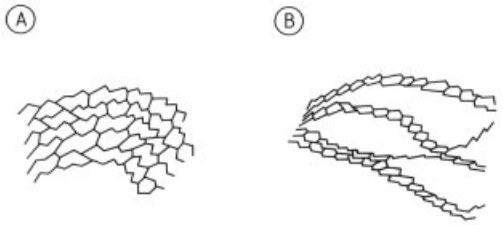

Figure 2. Arrangement of structural units in gel (A) and macroporous (B) resins

Channels are formed inside the beads, producing an artificially high porosity. Resins containing such channels are described as *macroporous*, whereas other resins with natural porosity are known as *gel resins* (Fig. 2).

Macroporous resins have a higher degree of cross-linking than gel resins to strengthen the matrix and compensate for voids left by the added solvent. The porosity and mechanical strength of the resin can be modified by varying the degree of cross-linking or the amount of solvent added. Therefore, various macroporous resins are available, with different moisture-holding capacities and internal structures.

The pore diameter is ca. 100 nm in a macroporous resin and ca. 1 nm in a gel resin. The macropores form a network of channels filled with free water, and large molecules can move freely in the resin into the center of a bead. Once inside the resin, ions generally have a much shorter distance to travel before they encounter an active group: ca. 100 nm in macroporous resins and up to 500 µm in gel resins. Exchange is thus faster in a macroporous resin.

Macroporous resins are highly resistant to physical stress and generally withstand osmotic shock very well. They are therefore used in systems where mechanical and osmotic stress would otherwise cause gel resins to deteriorate rapidly, such as those involving circulation of resin, fluidized beds, high flow rates, oxidizing conditions, concentrated solutions, and short cycles.

Finally, macroporous resins are used when reversible uptake of large molecules is necessary, without fouling the resin. The adsorbents described in Section 2.3 have a macroporosity that allows selective retention of various molecules.

3.2. Exchange Capacity

Total Capacity. The total exchange capacity of a resin, expressed in equivalents per unit weight (or per unit volume), represents the number of active sites available. In polystyrene exchangers, the maximum number of active sites corresponds to the "grafting" of one active group per benzene ring. The capacity is expressed in equivalents (eq) per kilogram of dry resin (the weight capacity C_p) or equivalents per liter of wet settled resin (the volume capacity C_V). To-

Table 2. Typical capacities of ion-exchange resins ∗

Type	Amberlite	C_p, eq/kg	C_v, eq/L
Strongly acidic gel, 8 % DVB	IR120	4.5 (Na)	2.05 (Na)
Weakly acidic gel	IRC86	11.0 (H)	4.2 (H)
Strongly basic, type 1	IRA402	3.9 (Cl)	1.3 (Cl)
Strongly basic, type 2	IRA410	3.6 (Cl)	1.3 (Cl)
Strongly basic acrylic	IRA458	4.2 (Cl)	1.3 (Cl)
Weakly basic styrene	IRA96	4.7 (free base)	1.25 (free base)
Weakly basic acrylic	IRA67	5.6 (free base)	1.6 (free base)

∗ Values depend to some extent on the analytical method used for capacity determination. Macroporous strongly acidic resins have a dry weight capacity of 4.3 – 4.5 eq/kg. Their volume capacity depends greatly on their porosity, which can be adjusted within relatively broad limits. Similarly, macroporous strong base resins have dry weight capacities close to their gel counterparts, but generally lower volume capacities.

tal capacity values for some of the most common resins are given in Table 2. Equivalent resins of other brands have similar capacity values (see Table 10).

Operating Capacity. The operating capacity is defined as the proportion of total capacity used during the exchange process. It can amount to a large or small proportion of the total capacity and depends on a number of process variables including

1) Concentration and type of ions to be absorbed
2) Rate of percolation
3) Temperature
4) Depth of resin bed
5) Type, concentration, and quantity of regenerant

In a packed column, reaction between the ions in solution and those in the resin occurs over a well-defined region of the resin bed known as the *reaction zone*. When the selectivity of a resin for a dissolved ion is high (see Chap. 5), a sharp exchange wavefront is formed which moves toward the column outlet, making maximum use of the resin. The depth of the reaction zone depends on factors such as flow rate (kinetics) and ionic concentration, but is independent of column length [12].

When selectivity is low, a diffuse wavefront develops. The depth of the reaction zone then depends on the selectivity coefficient (Section 5.2) and also on the bed depth. The longer the column, the deeper is the reaction zone and the greater is the operating capacity of the resin. Figure 3 illustrates this: the top of the column contains the completely exhausted resins (a), whereas the reaction zone (b) contains partially exhausted resin. The time at which the lowest point of this zone reaches the bottom of the column (i.e., when the adsorbed ions break through the bottom of the bed) is generally taken to be the time at which the service phase is complete. A proportion of the resin is still not exhausted at the time of breakthrough.

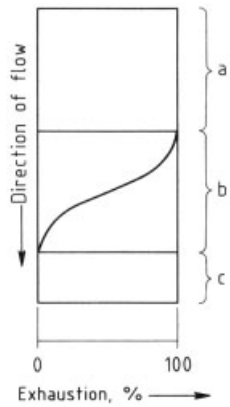

Figure 3. Reaction zone in a resin column during percolation
a) Exhausted resin; b) Reaction zone; c) Regenerated resin

In practice, strongly acidic and strongly basic resins are never 100 % regenerated at the beginning of a cycle. The operating capacity thus represents the difference between the available capacity at the beginning of a cycle and that remaining at the end point. The most important factor is the amount of regenerant used to convert the resin to the regenerated form required at the beginning of the service cycle (c). The calculation of capacity is described in Chapter 7.

3.3. Stability and Service Life

Because ion-exchange resins must give several years of service, their stability over long periods of time is of prime importance.

Chemical Stability of the Matrix. Industrially available resins have a degree of cross-linking high enough to make them insoluble. A new resin may release minute quantities of short-chain polymers or other soluble substances, but this effect is short-lived.

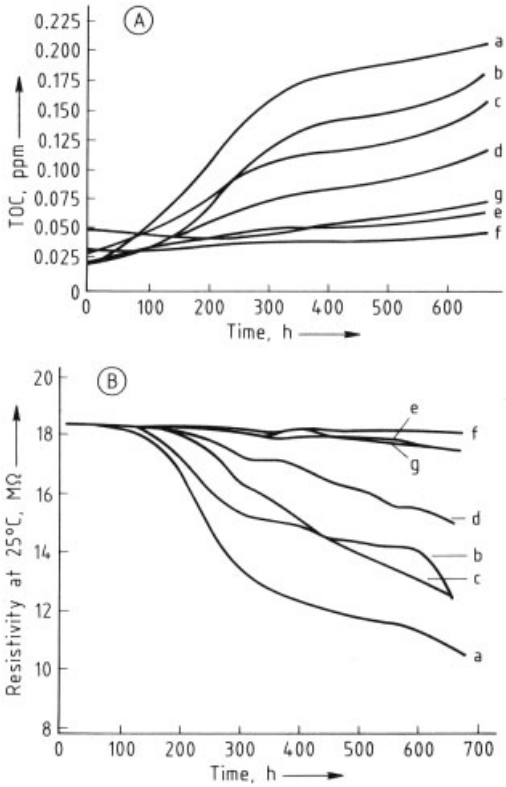

Figure 4. Suitability of resins for producing ultrapure water in mixed beds [14]
Resins were tested at ambient temperature during first chlorine exposure by using an influent with 0.30 ppm active chlorine dosed as NaOCl and 0.02 – 0.03 ppm TOC.
Resins: a) Gel cation – porous gel anion; b) Gel cation – standard gel anion; c) Macroporous cation – porous gel anion; d) Macroporous cation – standard gel anion; e) Macroporous cation – macroporous anion; f) Macroporous cation – developmental anion; g) Gel cation – developmental anion

Highly oxidizing conditions (presence of chlorine or chromic acid) can attack the matrix and destroy cross-linking. A sulfonated polystyrene cation-exchange resin with 8 % DVB cross-linking withstands 0.2 mg/kg of chlorine at ambient temperature for several years and is also completely stable at 120 °C in the absence of oxidants. However, 1 mg/kg of chlorine oxidizes the polymer at a rate dependent on temperature; this breaks down the cross-linking, releases sulfonated organic compounds and causes the resin to swell until it softens, resulting in excessive head loss (sloughage) [13]. When oxidizing agents are present, highly cross-linked resins with a greater resistance to oxidation, such as the macroporous resins, should be used.

Degradation products from a cation-exchange resin may foul anion resins [14], [15]. This is particularly critical in processes designed to produce ultrapure water (Fig. 4). Oxidants break the cross-links to produce soluble, short-chain oligomers that can be measured as the total organic carbon (TOC) in the treated water (Fig. 4 A). The oligomers bear ionized active groups and thus decrease the resistivity of the treated water (Fig. 4 B). Under normal conditions of water treatment, resins can operate continuously for many years (sometimes up to 20 years) without deterioration of their physical and chemical properties.

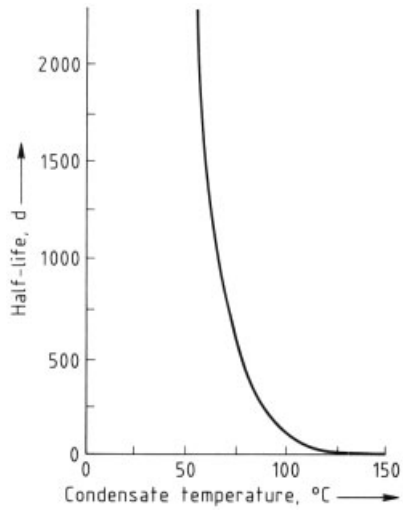

Figure 5. Half-life of Amberlite IRA 402 as a function of temperature
Half-life is the contact time of resin with hot solution required to reduce total capacity by 50 %.

Thermal Stability of Active Groups. [16] The sulfonic group of cation-exchange resins

is extremely stable. Anion-exchange resins, on the other hand, are temperature-sensitive. When heated, Hofmann degradation may transform quaternary ammonium groups (strongly basic) into tertiary amines (weakly basic) or even destroy the active group completely. Because this reaction occurs under alkaline conditions, anion exchangers are more stable in the form of a salt than as a base.

Strongly basic type 1 resins are the most stable; the Hofmann degradation reaction becomes significant only above 50 °C (Fig. 5). At ambient temperature, these resins can last for five to seven years or more.

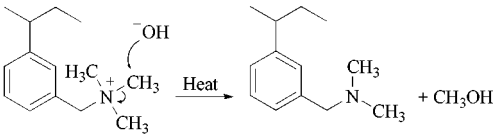

Type 1 resin

Type 2 resins are more liable to undergo Hofmann degradation because the hydroxyethyl group weakens the bond:

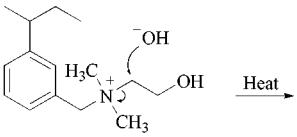

Type 2 resin

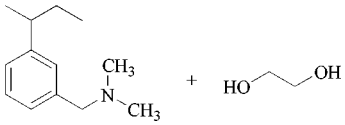

At ambient temperature, these resins lose ca. 50 % of their strongly basic groups in five years; the same effect occurs in about one year at 50 °C. (A resin in which 50 % of the strongly basic groups have been converted into weak groups, often has an unchanged operating capacity because the weakly basic groups are fully capable of taking up mineral acids).

Mechanical Stability. Polycondensation-type resins that are manufactured in bulk and broken up into irregular grains are comparatively fragile and used only in fixed beds (Section 10.2). Polystyrene and polyacrylic resins made by suspension polymerization are perfect spheres and suffer little damage when used in continuous moving-bed ion-exchange plants. However, mechanical strength can vary considerably from one product to another, and resin beads which are seen to have many internal cracks under the microscope are more likely to break under mechanical stress than crack-free products.

Gel-type anion resins are generally weaker than cation materials and are particularly poor at withstanding compression. However, new polymerization techniques produce a more uniformly structured polystyrene matrix. As a result, gel resins with high physical stability are now available. To prevent fragile resins from breaking, it is important to keep the bed clean by frequent backwashing because the water pressure on a layer of fine debris can be as much as 60 kPa.

Acrylic resins are more elastic than polystyrene materials and can normally withstand any mechanical stress encountered in practice. Macroporous resins are often the strongest of all and are used widely for the most severe stress conditions.

The less elastic resins, i.e., those with the highest degree of cross-linking (gel resins with > 8 % DVB and macroporous resins with >15 % DVB), have the disadvantage that, when they do break, they explode into minute fragments, whereas other resins break into two or three usable pieces.

Osmotic Stability. During ion exchange, the configuration around each active group in the resin changes: the adsorbed ion generally has a different size and, more important, a different hydration layer than the displaced ion. The resin bead may therefore swell or contract appreciably during the reaction. The stresses to which the resin is subjected during these volume changes are known as osmotic forces. They are very intense and can produce local pressures of several thousand kilopascals – much greater than purely mechanical stress.

Resins for industrial use must be able to withstand hundreds of cycles of exhaustion and regeneration. The nature and, hence, the strength of osmotic shock vary according to the ionic species in solution and their concentration.

Higher mechanical and osmotic strengths are obtained with resins whose matrix is sufficiently strong to withstand physical shock (attrition) but sufficiently flexible and porous to deform without breaking under the effect of osmotic shock. Some macroporous resins combine both of these qualities.

Resistance to Drying. Repeated drying and rewetting produce stresses analogous to those due to osmotic shock and can lead to fragmentation of most gel resins. Resins must therefore be kept permanently moist.

3.4. Density

Resin density is an important property because it determines the hydrodynamic behavior in counterflow systems. Resin density normally lies in the following ranges (figures in parentheses are the most common values for standard resins):

Strongly acidic cation exchangers	1.18 – 1.38 (1.28)
Weakly acidic cation exchangers	1.13 – 1.20 (1.18)
Strongly basic anion exchangers	1.07 – 1.12 (1.10)
Weakly basic anion exchangers	1.02 – 1.10 (1.05)

By choosing suitable particle sizes, several different types of resin can be used in the same column. If necessary, they can be kept separate by an upflow of water. This is used mainly in layered beds (see Section 9.3).

Resin density can be increased artificially, for example, by attaching chlorine or bromine atoms to the matrix. This yields high-density anion exchangers, which are useful when fluidized-bed operation is required. However, such "high-density" resins are not available commercially.

3.5. Particle Size

For industrial use, particle size is a compromise between the speed of the exchange reaction (which is greater with small beads) and high flow rates (which require coarse particles to minimize the head loss).

The size of the polymer droplets formed during polymerization, and hence the size of the resin beads, is determined by the polymerization technology, the suspension medium and, the monomer concentration.

Traditional polymerizations are carried out in batches in a stirred reactor. The beads produced in this way have a range of particle sizes rather than a uniform size (Table 3). The population of a resin sample (i.e., the number of beads classified according to bead diameter) has an approximately Gaussian distribution. Because the volume of a fraction is a logarithmic function of the number of beads, the particle-size distribution is often represented by a straight line on probability graph paper (Fig. 6).

Table 3. Particle-size distribution of ion-exchange resins ∗

Sieve aperture, mm	Noncumulative total (between sieves), %	Cumulative total (passing through), %
<0.315	0.2	0.2
0.315 – 0.40	1.3	1.5
0.40 – 0.50	8.5	10.0
0.50 – 0.63	30.8	40.8
0.63 – 0.80	33.3	74.1
0.80 – 1.00	20.4	94.5
1.00 – 1.25	4.9	99.4
>1.25	0.6	100.0

∗ Effective particle size 0.50 mm, mean diameter 0.68 mm, uniformity coefficient 0.73/0.50 = 1.46

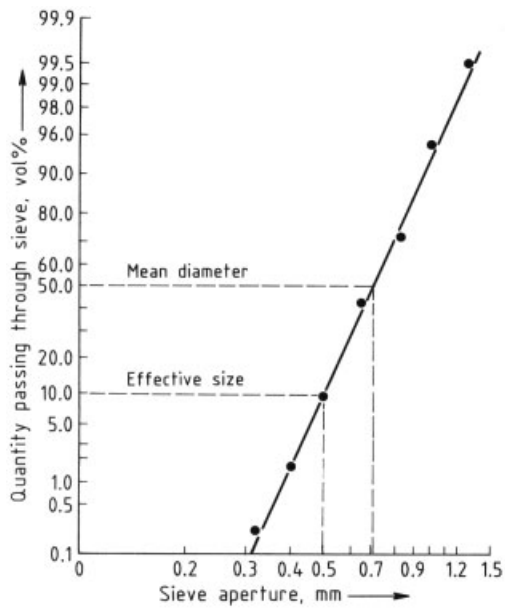

Figure 6. Typical particle-size distribution in a normal resin

For Gaussian distribution, the particle-size distribution is defined by

1) The *mean diameter* (corresponding to the mesh size of a sieve allowing 50 % of the beads to pass)
2) The *uniformity coefficient*, U.C. (given by the ratio between the aperture of a sieve allowing 60 % of the beads to pass and that of a sieve allowing 10 % to pass. The latter theoretical sieve is known as the *effective size.*

$$\text{U.C.} = x_{60\,\%}/x_{10\,\%}$$

The closer the uniformity coefficient is to unity, the narrower is the Gaussian curve and hence the smaller is the range of particle size. Resins produced with the traditional stirred-reactor process usually have a uniformity coefficient of 1.5 to 1.9.

Several producers now offer ion-exchange resins with a very uniform particle size distribution, which are produced with different polymerization technologies. In one technique, the monomers are injected into the suspension medium through a plate perforated with thousands of very small holes. Droplets of the monomer mixture with an almost identical volume are expelled. The uniformity coefficient of resins produced in this way is always less than 1.2, often in the range of 1.05 – 1.10. Examples of such resins are Amberjet (Rohm and Haas), Dowex Monosphere (Dow Chemical), and Lewatit Monoplus (Bayer).

3.6. Moisture Content

Ion-exchange resins carry both fixed and mobile ions which are always surrounded by water molecules located in the interior of the resin beads. The water retention capacity governs the kinetics, exchange capacity, and mechanical strength of ion-exchange resins.

The moisture content or *moisture-holding capacity* (MHC) is defined as

$$\text{MHC} = \left(P_{\text{Hydr}} - P_{\text{Dry}}\right)/P_{\text{Hydr}}$$

where P_{Hydr} is the weight of the hydrated resin sample and P_{Dry} the weight of the same sample after drying.

The MHC of an ion-exchange resin is an inverse function of the degree of cross-linking unless the porosity or degree of cross-linking in the polymer is artificially increased (as in macroporous resins). Figure 7 shows how the moisture content varies with the proportion of DVB for gel-type sulfonic polystyrene resins.

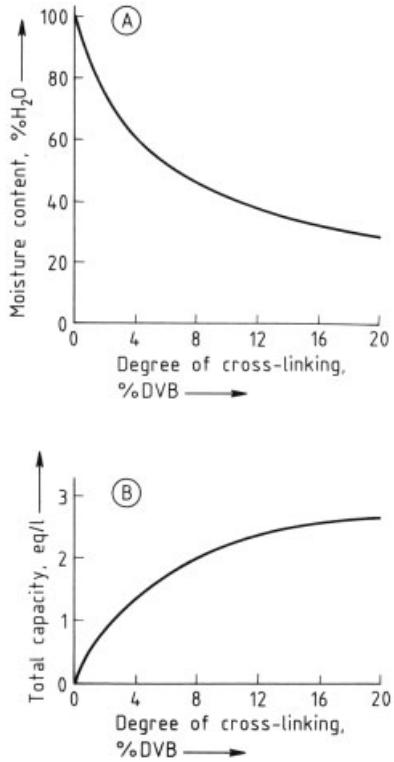

Figure 7. Variation of moisture content (A) and total capacity (B) with the degree of cross-linking in a sulfonated polystyrene resin in sodium form

Another useful quantity is the *dry matter* (DM) defined as

$$\text{DM} = P_{\text{Dry}}/V_{\text{Hydr}}$$

where P_{Dry} is the weight of the dry resin sample and V_{Hydr} the volume of the sample before drying. Although they are closely connected, no simple arithmetical relationship exists between dry matter and moisture content. In all cases, the ionic form of the resin at the time of measurement should be quoted.

4. Ion-Exchange Reactions

4.1. Cation Exchange

General cation exchange is used widely to remove undesirable ions from a solution without changing the total ionic concentration or pH. The resin can be used in many ionic forms, but the sodium form is usually preferred because the resin has a relatively low affinity for sodium, which facilitates the adsorption of other metals. Furthermore, sodium chloride is an inexpensive regenerant.

The following reaction is used to treat wine (R denotes the resin):

$$R^-Na^+ + K^+\,(HTartrate)^- \longrightarrow R^-K^+ + Na^+\,(HTartrate)^-$$

The reaction used in water softening is

$$2\,R^-Na^+ + Ca^{2+}(HCO_3^-)_2 \longrightarrow (R^-)_2Ca^{2+} + 2\,Na^+HCO_3^-$$

In each case, the resin is regenerated by reversal of the reaction with sodium chloride solution.

Hydrogen Exchange in Strongly Acidic Resins. The replacement of metal ions with hydrogen ions leads to a reduction of the total dissolved solids in solution and the production of free acid:

$$R^-H^+ + Na^+Cl^- \longrightarrow R^-Na^+ + H^+Cl^-$$

This reaction is used as the first stage in the demineralization of water and other solutions. It is sometimes called salt splitting. Regeneration is carried out with a mineral acid.

Hydrogen Exchange in Weakly Acidic Resins. Carboxylic resins are such weak acids that they ionize only slightly under acid conditions. However, they have such a high affinity for divalent metals that in the presence of these metals, they are forced to ionize and therefore remain active under slightly acidic conditions down to ca. pH 4.5.

$$2\,RCOOH + Mg(OH)_2 \longrightarrow (RCOO)_2Mg + 2\,H_2O$$

$$2\,RCOOH + Na_2CO_3 \longrightarrow 2\,RCOONa + H_2O + CO_2$$

$$RCOOH + NaCl \text{ No reaction}$$

$$2\,RCOOH + CaCl_2 \longleftarrow (RCOO)_2Ca + 2\,HCl$$

However, H^+ from carboxylic resins cannot remove significant amounts of metals from solutions of mineral acid salts because the acid produced quickly lowers the pH and prevents further ion exchange. In this case, ion exchange is controlled by the basicity of the anion in solution. A strongly nucleophilic anion "tears off" hydrogen from the carboxylic group, which takes up the associated cation in exchange, whereas a stable anion does not react significantly.

Because the resin has a very high affinity for divalent ions (the effect of chelation), but only moderate affinity for monovalent ions, it has a high capacity for removing calcium and magnesium from bicarbonate solution but takes up only a small amount of sodium. This occurs because sufficient carbonic acid is formed to suppress the exchange of monovalent ions:

$$RCOOH + NaHCO_3 \longleftarrow RCOONa + H_2O + CO_2$$

$$2RCOOH + Ca(HCO_3)_2 \longrightarrow (RCOO)_2Ca + 2H_2O + 2CO_2$$

Carboxylic exchangers (weakly acidic) are selective: they preferentially remove divalent or trivalent cations until competition arises from alkaline anions present in solution.

Selective Exchange in Weakly Acidic and Complexing Resins. Because carboxylic resins take up multivalent cations in preference to monovalent ones and are regenerated very easily, they can be used for selective removal of divalent and trivalent metals, even when high concentrations of alkaline cations are present. To achieve this, the exchanger is first converted to the monovalent sodium form:

$$RCOOH + NaOH \longrightarrow RCOONa + H_2O$$

which allows exchange of neutral salts:

$$2\,RCOONa + ZnSO_4 \longrightarrow (RCOO)_2Zn + Na_2SO_4$$

At the end of the cycle, the resin is regenerated with acid:

$$(RCOO)_2Zn + H_2SO_4 \longrightarrow 2\,RCOOH + ZnSO_4$$

and is reconverted to the monovalent sodium form by reaction with sodium hydroxide.

4.2. Anion Exchange

General Anion Exchange. The most widely used resin for general anion exchange is a strongly basic exchanger in chloride form:

$$R^+Cl^- + Na^+NO_3^- \longrightarrow R^+NO_3^- + Na^+Cl^-$$

This process is used to remove natural organic acids (e.g., humic acid, see Section 7.3) and nitrate from water and in hydrometallurgy to selectively adsorb metals that form anionic complexes:

$$n\,R^+Cl^- + (Na^+)_n\ \mathrm{Humate}^{n-} \longrightarrow (R^+)_n\ \mathrm{Humate}^{n-} + n\,Na^+Cl^-$$

$$R^+Cl^- + Na^+[Au(CN)_2]^- \longrightarrow R^+[Au(CN_2]^- + Na^+Cl^-$$

The resins are regenerated by reversal of the reactions with sodium chloride solution.

Acid Absorption in Strongly Basic Resins is the most widely used form of anion exchange. When it follows hydrogen exchange in strongly acidic resins (Section 4.1), it completes the demineralizing process:

$$R^+OH^- + H^+Cl^- \longrightarrow R^+Cl^- + H_2O$$

$$R^+OH^- + CO_2 \longrightarrow R^+HCO_3^-$$

Regeneration of the bicarbonate form of the resin requires two equivalents of OH^- ions per equivalent of HCO_3^- taken up, because half the OH^- neutralizes the bicarbonate and converts it to carbonate. The same applies to silica.

Acid Absorption in Weakly Basic Resins. Weakly basic resins, in which the active groups are usually amines, do not have a true hydroxide form. They ionize only under acidic conditions:

$$RN(CH_3)_2 + H^+Cl^- \longrightarrow RN^+H(CH_3)_2Cl^-$$

Under alkaline conditions, they exist as free bases and can therefore adsorb acids in the same way that free ammonia reacts with hydrochloric acid to form ammonium chloride. In practice, they are called the "chloride form". Weakly basic resins can only adsorb strong acids. Neutral salts are not "split":

$$RN(CH_3)_2 + NaCl \text{ No reaction}$$

This is because no H^+ ion exists to which the nucleophilic base can give its electrons and thus balance the anion. Similarly, weak acids do not have dissociated H^+ ions, so they are taken up only in very small amounts, if at all:

$$RN(CH_3)_2 + CO_2 + H_2O \rightleftharpoons RN^+H(CH_3)_2 + HCO_3^-$$

$$RN(CH_3)_2 + SiO_2 \text{ No reaction}$$

Weakly basic exchangers can be regenerated with ammonia or sodium carbonate:

$$RN(CH_3)_2 \cdot HCl + NH_3 \longrightarrow RN(CH_3)_2 + NH_4Cl$$

$$RN(CH_3)_2 \cdot HCl + Na_2CO_3 \longrightarrow RN(CH_3)_2 + NaCl + NaHCO_3$$

4.3. Cation and Anion Exchange in Water Treatment

Figure 8 summarizes the way in which the various forms of ion exchange described in Sections 4.1 and 4.2 are used in water treatment. The composition of raw water is described in detail in Section 7.3.

5. Ion-Exchange Equilibria

5.1. Dissociation and pK Value

Dissociation of the acid group in a cation-exchange resin is described by the equilibrium reaction

$$R \cdot H \rightleftharpoons R^- + H^+$$

where R^- is the co-ion fixed in the matrix structure. The acidity of the resin is defined by its degree of dissociation at equilibrium

$$K = [\underline{R}^-] \cdot [\underline{H}^+] \,/\, [\underline{RH}]$$

where K is the equilibrium constant; the quantities in square brackets represent concentrations, and underlines indicate the resin phase rather than the aqueous phase. The pK is defined as

$$pK = -\log K$$

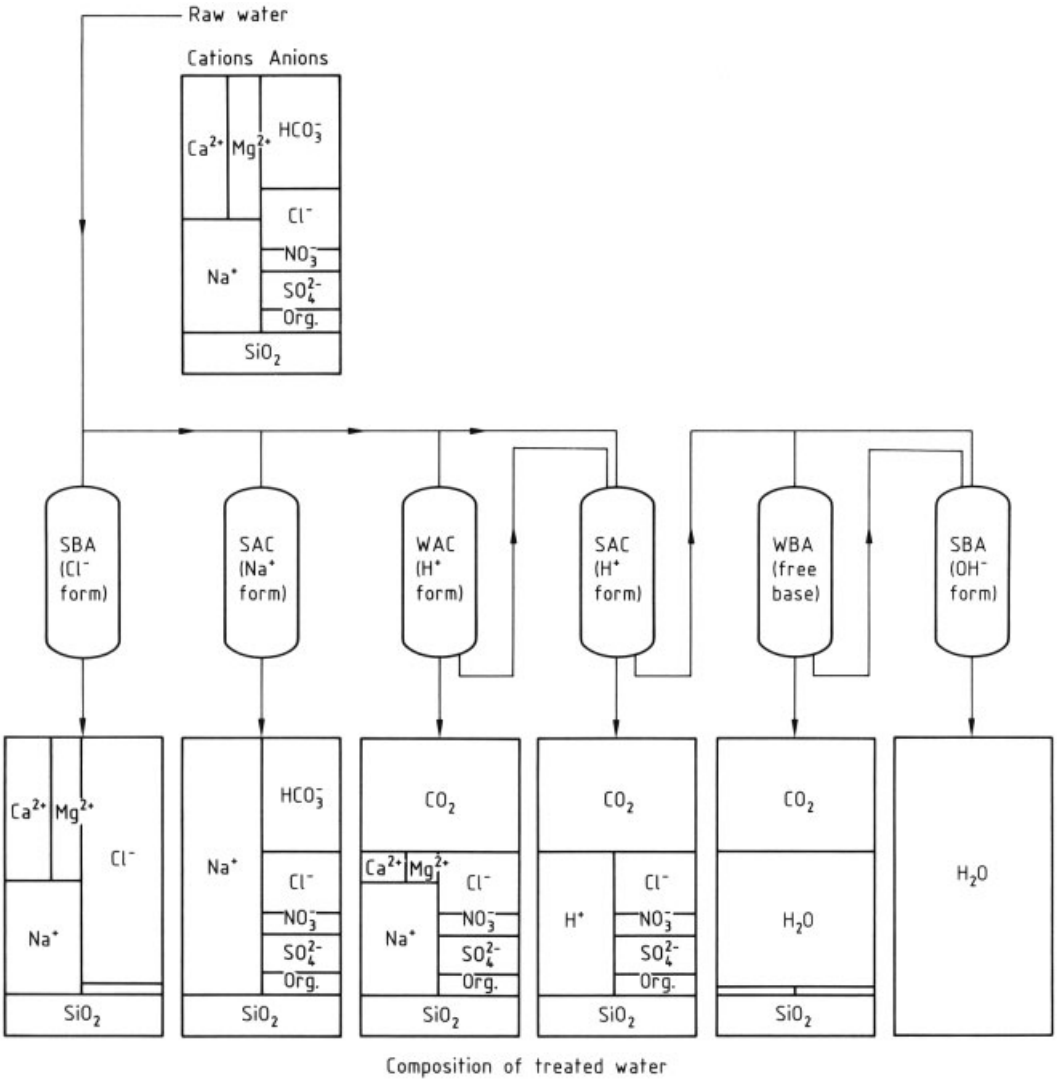

Figure 8. Summary of the kinds of ion exchange used in water treatment
SAC = strongly acidic cation exchangers; SBA = strongly basic anion exchangers; WAC = weakly acidic cation exchangers; WBA = weakly basic anion exchangers

Cation-exchange resins can be titrated (Fig. 9) with sodium hydroxide in the presence of sodium chloride. Uptake of Na^+ by the cation-exchange resin in H^+ form occurs:

$$R^-H^+ + Na^+ \longrightarrow R^-Na^+ + H^+ \quad (1)$$

The H^+ ions released by the resin combine immediately with the OH^- ions from the alkali

$$H^+ + OH^- \longrightarrow H_2O$$

and drive the reaction shown in Equation (1) to completion.

At the beginning of titration, the strongly acidic (sulfonic) resin Amberlite IR 120 exchanges sodium ions from sodium chloride, and released H^+ ions lower the pH of the aqueous phase (Fig. 9, curve a). The pH remains low as long as the resin releases H^+ ions to replace the H^+ ions picked up from the added sodium hydroxide solution. As soon as all H^+ ions have

been replaced in the resin, further addition of sodium hydroxide raises the pH sharply.

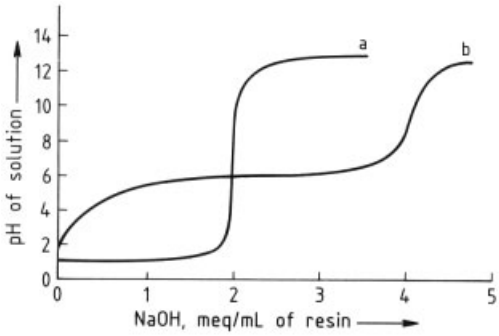

Figure 9. Titration curves of cation-exchange resins
a) Amberlite IR 120, sulfonic acid resin; b) Amberlite IRC 86 carboxylic acid resin

The weakly acidic (carboxylic) resin Amberlite IRC 86 behaves differently. Because its acidic groups are weakly dissociated, the exchange reaction (Eq. 1) remains incomplete, and the pH of the solution increases progressively even during the early stages of titration (Fig. 9, curve b). A sharp rise is noticed, however, when the resin is converted completely to the Na^+ form.

In both cases, the total capacity of the resin can be read from the titration curves and corresponds to the amount of sodium hydroxide that produces the sharp rise in pH. In Figure 9, Amberlite IR 120 has a total capacity of 2 meq/mL, whereas Amberlite IRC 86 has a capacity > 4 meq/mL.

Similar considerations apply to anion-exchange resins.

5.2. Mono – Monovalent Exchange

The law of mass action applied to the reaction

$$R^-H^+ + Na^+ \rightleftharpoons R^-Na^+ + H^+$$

gives

$$\frac{[\underline{Na}^+]\,\underline{\gamma}_{Na}}{[\underline{H}^+]\,\underline{\gamma}_{H}} = K\,\frac{[Na^+]\,\gamma_{Na}}{[H^+]\,\gamma_{H}} \qquad (2)$$

where $[Na^+]$ and $[H^+]$ are the equivalent concentrations in the liquid phase and γ_{Na} and γ_H are the corresponding activity coefficients. The concentrations and coefficients for the resin are underlined. Thus,

$$K \cdot \frac{\gamma_{Na}}{\gamma_H} \cdot \frac{\underline{\gamma}_H}{\underline{\gamma}_{Na}} = \frac{[\underline{Na}^+]}{[\underline{H}^+]} \cdot \frac{[H^+]}{[Na^+]}$$

The activity coefficients may be assumed constant, so that

$$\frac{[\underline{Na}]}{[\underline{H}]} = K_H^{Na} \cdot \frac{[Na]}{[H]}$$

(the + signs are omitted for simplicity). The parameter K_H^{Na} is known as the *selectivity coefficient* for the Na^+/H^+ exchange.

Selectivity coefficients differ for each pair of ions because the affinity of the resin for an ion is governed by the size of its hydrated form. The larger the ion, the more the resin bead must expand to accommodate it. Expansion is opposed by the restraining cross-links, so that large ions require a greater force to penetrate the resin than small ones.

Table 4. Relative selectivities of sulfonic resins for cations

Cation	Degree of cross-linking, % DVB			
	4	8	12	16
Monovalent				
H *	1.0	1.0	1.0	1.0
Li	0.90	0.85	0.81	0.74
Na	1.3	1.5	1.7	1.9
NH_4	1.6	1.95	2.3	2.5
K	1.75	2.5	3.05	3.35
Rb	1.9	2.6	3.1	3.4
Cs	2.0	2.7	3.2	3.45
Cu	3.2	5.3	9.5	14.5
Ag	6.0	7.6	12.0	17.0
Divalent				
Mn	2.2	2.35	2.5	2.7
Mg	2.4	2.5	2.6	2.8
Fe	2.4	2.55	2.7	2.9
Zn	2.6	2.7	2.8	3.0
Co	2.65	2.8	2.9	3.05
Cu	2.7	2.9	3.1	3.6
Cd	2.8	2.95	3.3	3.95
Ni	2.85	3.0	3.1	3.25
Ca	3.4	3.9	4.6	5.8
Sr	3.85	4.95	6.25	8.1
Hg	5.1	7.2	9.7	14.0
Pb	5.4	7.5	10.1	14.5
Ba	6.15	8.7	11.6	16.5

* Reference value

Table 4 lists the relative selectivities of sulfonic resins for mono- and divalent cations, and Table 5 gives the selectivities of strongly ba-

sic type 1 and type 2 resins for monovalent anions. Values increase with the degree of cross-linking and tend to unity as the cross-linking tends to zero. Data are only approximate and merely demonstrate the scale of selectivities.

Table 5. Relative selectivities of quaternary ammonium exchangers for monovalent anions

Anion	Resin	
	Type 1	Type 2
Hydroxide *	1.0	1.0
Benzenesulfonate	>500	75
Salicylate	450	65
Iodide	175	17
Phenolate	110	27
Bisulfate	85	15
Chlorate	74	12
Nitrate	65	8
Bromide	50	6
Cyanide	28	3
Bisulfite	27	3
Bromate	27	3
Nitrite	24	3
Chloride	22	2.3
Bicarbonate	6.0	1.2
Iodate	5.5	0.5
Formate	4.6	0.5
Acetate	3.2	0.5
Propionate	2.6	0.3
Fluoride	1.6	0.3

* Reference value

The degree of cross-linking in resins is not uniform, so that an ion entering a resin first seeks the most favorable regions. All exchange reactions become less favorable as they progress toward completion. The practical effects of this general phenomenon on the exchange rate, exchange capacity, and leakage from resins are considered in Chapter 7.

5.3. Mono – Divalent Exchange (Water Softening)

In the exchange reaction

$$2\,R^-Na^+ + Ca^{2+} \rightleftharpoons (R^-)_2Ca^{2+} + 2\,Na^+$$

the law of mass action gives

$$\frac{[\underline{Ca}]\,\underline{\gamma}_{Ca}}{[\underline{Na}]^2\,\underline{\gamma}^2_{Na}} = K\,\frac{[Ca]\,\gamma_{Ca}}{[Na]^2\,\gamma^2_{Na}}$$

If the activity coefficients are constant, then

$$\frac{[\underline{Ca}]}{[\underline{Na}]^2} = K^{Ca}_{Na}\,\frac{[Ca]}{[Na]^2}$$

The equivalent fraction X_i of an ion i in a solution with a total concentration of ions $[C]$ is defined by

$$X_i = [i] \,/\, [C]$$

Similarly, the equivalent fraction of ion i in the resin is

$$[\underline{X}_i] = [\underline{i}] \,/\, [\underline{C}]$$

The total concentration of ions in the resin $[\underline{C}]$ (in equivalents per liter) is the same as the total capacity (defined in Section 3.2).

In exchange involving only two ions (e.g., Na^+ and Ca^{2+}), the following equations apply to the solution:

$$[C] = [Na] + [Ca]$$

$$X_{Na} + X_{Ca} = 1$$

Similar equations hold for the resin. Equation (2) thus becomes

$$\frac{\underline{X}_{Ca}}{(1-\underline{X}_{Ca})^2} = K^{Ca}_{Na}\cdot\frac{[\underline{C}]}{[C]}\cdot\frac{X_{Ca}}{(1-X_{Ca})^2} \qquad (3)$$

where $[\underline{C}]$ is constant and K^{Ca}_{Na} is approximately constant at low concentrations. Equation (3) shows that as the solution becomes more dilute (decreasing C) with a constant fraction of calcium ($\underline{X}_{Ca}$ constant), X_{Ca} increases. The resin takes up more calcium as the total cation concentration of the solution decreases. This effect is illustrated in Figure 10, which shows the equilibrium curves for a given resin at various total concentrations of the solution. Each point on a curve corresponds to the equivalent fraction of calcium in the resin at equilibrium, i.e., the curve gives the proportion of active sites in the resin in Ca^{2+} form as a function of the proportion of calcium in solution. These curves are called *ion-exchange isotherms*.

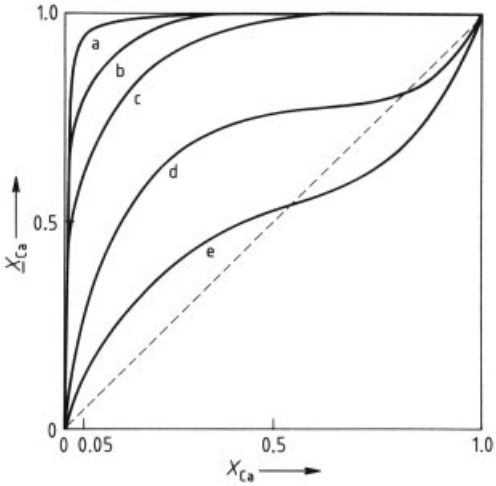

Figure 10. Mono – divalent equilibrium curves for Na^+ – Ca^{2+} solutions of different total concentration
Total concentration of Na^+ and Ca^{2+}, eq/L: a) 0.005; b) 0.01; c) 0.1; d) 1; e) 5

Water Softening. For water with a total salinity of 5 meq/L, i.e. $[C] = 0.005$ N, Figure 10 shows that water containing only 5 % Ca^{2+} ($X_{Ca} = 0.05$ or [Ca] = 0.25 meq/L) is in equilibrium with a resin loaded with 95 % calcium ($\underline{X}_{Ca} = 0.95$). In other words, the resin is capable of removing calcium from the water even if the concentration accounts for only 5 % of total cations ($X_{Ca} = 0.05$), as long as the resin contains more than 5 % sodium ($\underline{X}_{Ca} < 95$ %). At high concentrations, the affinity of the resin for calcium over sodium decreases until sodium becomes favored. This is the effect that makes regeneration of the resin possible.

Figure 11 shows the complete softening cycle and regeneration with 12 % brine (ca. 2 N NaCl) using curves representing successive equilibrium states of the system. At the starting point O, all of the resin is in the sodium form. The water, with a total sodium and calcium content of 5 meq/L (0.005 N), comes in contact with the resin, and establishment of an equilibrium leads to the uptake of calcium by the resin and its removal from the water. The first upward-pointing arrow on the curve indicates the displacement of the equilibrium. A new equilibrium is then established at P′ where the resin loaded with 20 % calcium is in equilibrium with completely softened water. The equilibrium then shifts to P, where resin loaded with 40 % calcium is still in equilibrium with completely softened water, and finally to Q where the resin loaded with 90 % calcium is in equilibrium with water which is about 98 % softened ($X_{Ca} = 0.02$) and therefore contains $0.02 \times 5 = 0.1$ meq/L of calcium. At this point, known as the end point, a new stage of the process is initiated to regenerate the resin (to continue as before would lead to the removal of less calcium). The system is displaced to point R on the other curve as the sodium chloride concentration changes from 0.005 to 2 N. At R, the resin is in equilibrium with a solution containing about 95 % calcium and therefore tends to release calcium ions so as to produce this concentration in the liquid. In this regeneration process, the resin is converted continuously into the sodium form by exchanging the calcium ions adsorbed during the first stage for sodium ions from the regenerating brine.

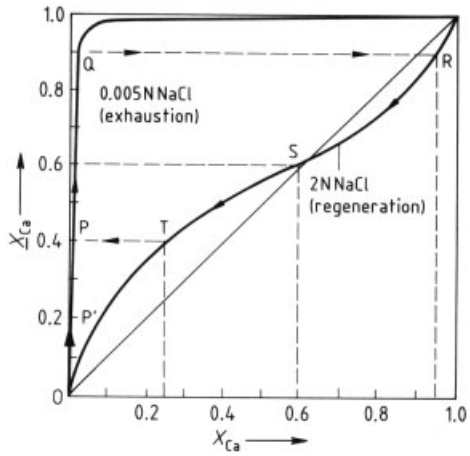

Figure 11. Successive equilibria in the water-softening cycle

As the system moves toward the bottom of the curve, the equilibrium becomes less favorable: at R, 95 % of the sodium chloride in the solution is used up in regeneration of the resin (because $X_{Ca} = 0.95$), whereas at point S the proportion is already less than 60 % and at point T, 25 %. At the latter point, the resin is 60 % regenerated ($\underline{X}_{Ca} = 0.4$) and the process stops because any attempt to reach point O would need increasing amounts of brine, only a small proportion of which would be used in actual regeneration. The resin is therefore rinsed at point T and another new stage is initiated by displacing the system to P and beginning a second cycle in which the

water is again almost completely softened. All subsequent cycles follow the path PQRT.

The regeneration efficiency (percentage of brine exchanged) can be evaluated at each point on the curve RSTO. Between R and T the mean regeneration efficiency is 60 %.

Regeneration was carried out between $\underline{X}_{\mathrm{Ca}} = 0.9$ and 0.4, so that half the total capacity of the resin was used. With a total capacity of 2 eq/L, the operating capacity would thus be 1 eq/L. A regeneration efficiency of 60 % is assumed, 1/0.6 = 1.67 eq of sodium chloride would be used, i.e., 97.5 g per liter of resin.

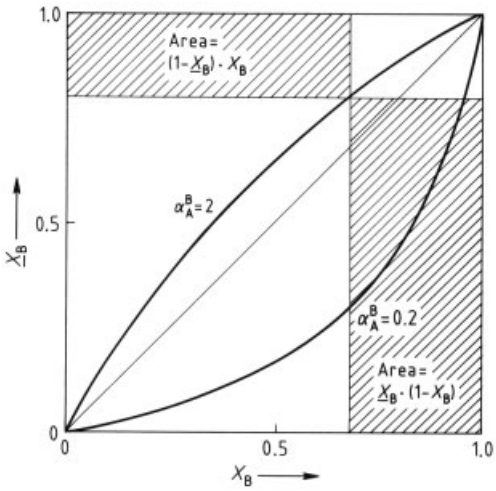

Figure 12. Ion-exchange isotherms for mono – monovalent ions

5.4. General Case

In the general exchange reaction between ions A and B of ionic valences a and b, respectively, the *selectivity coefficient* $K_{\mathrm{A}}^{\mathrm{B}}$ is defined as

$$K_{\mathrm{A}}^{\mathrm{B}} = \frac{[\underline{\mathrm{B}}]^b}{[\underline{\mathrm{A}}]^a} \cdot \frac{[\mathrm{A}]^a}{[\mathrm{B}]^b}$$

$K_{\mathrm{A}}^{\mathrm{B}}$ depends on experimental conditions such as concentration and temperature.

Another useful quantity is the *separation factor*, defined as

$$\alpha_{\mathrm{A}}^{\mathrm{B}} = \frac{[\underline{\mathrm{B}}]}{[\underline{\mathrm{A}}]} \cdot \frac{[\mathrm{A}]}{[\mathrm{B}]} = \frac{\underline{X}_{\mathrm{B}}}{1-\underline{X}_{\mathrm{B}}} \cdot \frac{1-X_{\mathrm{B}}}{X_{\mathrm{B}}}$$

For mono – monovalent exchange, the separation factor $\alpha_{\mathrm{A}}^{\mathrm{B}}$ and selectivity coefficient $K_{\mathrm{A}}^{\mathrm{B}}$ are identical and practically independent of the total salt content. For mono – divalent exchange, $\alpha_{\mathrm{A}}^{\mathrm{B}}$ is a function of the term $K \cdot [\underline{C}]/[\mathrm{C}]$. A separation factor $\alpha_{\mathrm{A}}^{\mathrm{B}}$ greater than unity means that the resin will take up ion B in preference to ion A.

Figure 12 shows ion-exchange isotherms for two monovalent ions A and B. The upper curve ($\alpha = 2$) represents a case in which the resin has a higher affinity for B than for A; the lower curve ($\alpha = 0.2$) represents the opposite case. For any point on a given isotherm, the ratio between the two hatched areas is equal to the separation factor α.

In many cases, simple selectivity calculations show whether proposed ion-exchange processes will function [17], [18].

6. Exchange Kinetics

6.1. Principles [19], [20]

Mass action equations apply only to systems in equilibrium. In industrial practice where a solution flows through the resin, equilibrium is not necessarily reached and the results are influenced by kinetic considerations.

In fully ionized systems, the rate-determining step of ion exchange is the diffusion of the mobile ions toward, from, and in the resin phase, rather than the chemical reaction between fixed ions of the resin and mobile counterions. If a cation-exchange resin with negative fixed ions (e.g., sulfonate ions) is used as an example, the cation concentration in the resin is much greater than that in solution. However, any cations diffusing out of the resin into the dilute solution create a net negative charge in the solid phase and a net positive charge in solution. The resulting potential difference is called the *Donnan potential*; it prevents anions from penetrating the (negatively charged) resin. This phenomenon is called *Donnan exclusion.* Therefore, the co-ion (i.e., the anion in the case of cation exchange) does not participate in the ion-exchange process.

Figure 13 illustrates the uptake of Na^+ ions by a bead of H^+-form resin. The bulk solution contains a large excess of available Na^+ ions, with an effectively constant concentration. A static layer of solution, known as the *Nernst film*, surrounds the bead. This film is defined such

that it is unaffected by convection (i.e., flow) around the bead; ion transport takes place by diffusion only. Strong convection (i.e., high flow rate) decreases film thickness. The ion concentration is practically constant outside the Nernst film, and a concentration gradient occurs within it (Fig. 14).

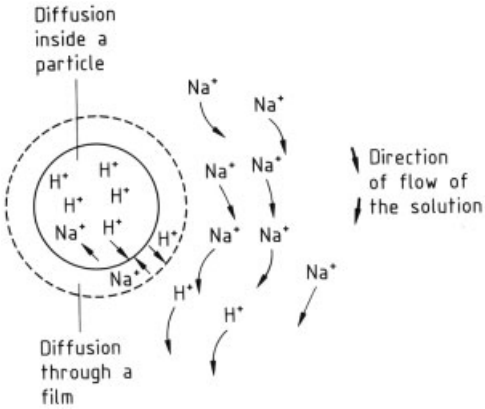

Figure 13. Diffusion through a film and inside a particle

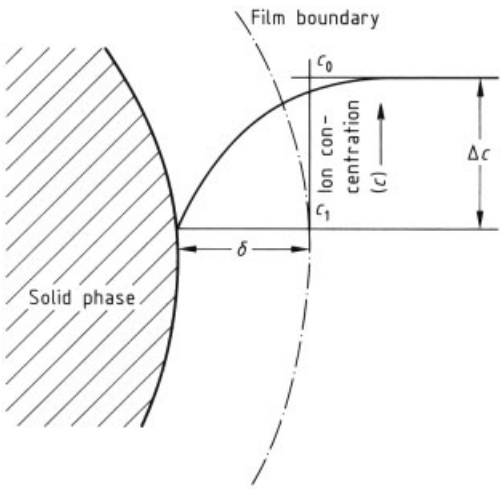

Figure 14. Nernst film

Diffusion through the film and in the solid phase occurs at different rates and two steps may be rate-determining:

1) Diffusion of ions within the resin (*particle diffusion*)
2) Diffusion in the Nernst film (*film diffusion*)

The slower step controls the overall ion-exchange rate. A criterion has been established by HELFFERICH [19] to determine which process is rate-determining:

$$H = (\underline{C}\underline{D}\delta / CDr) \cdot \left(5 + 2\alpha_{\mathrm{A}}^{\mathrm{B}}\right)$$

where

C = total ion concentration in solution
$\underline{C}$ = total ion concentration in the solid phase (total capacity)
D, $\underline{D}$ = diffusion coefficients
H = Helfferich constant
δ = thickness of Nernst film
r = radius of ion-exchange bead
α = separation factor (defined in Section 5.4)

When $H \gg 1$, film diffusion is rate-limiting; when $H \ll 1$, particle diffusion is rate-limiting.

The definition of H shows that film diffusion is favored by the following:

High resin capacity $\underline{C}$
Thick Nernst film δ (i.e., low flow rate)
Low concentration C in solution
Small resin beads
High selectivity (high α)

In general, forward and reverse exchange rates differ (Fig. 15). In this example of H^+/Na^+ exchange in a conventional strongly acidic cation-exchange resin, forward exchange (uptake of Na^+ ions) is faster than reverse exchange (regeneration). Because the H^+ ion generally is more mobile, exchange is faster when the faster ion is in the resin initially. The rates of sodium – calcium exchange are several times slower.

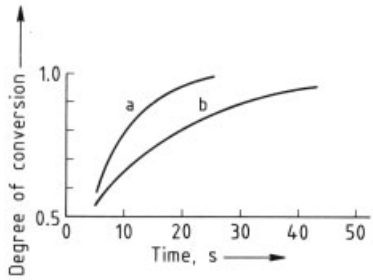

Figure 15. Forward and reverse exchange rates in a polystyrene sulfonic resin
a) Conversion from H^+ to Na^+ ; b) Conversion from Na^+ to H^+

For a given ion, mass-transfer flux through the film can be described by Fick's law:

$$J = -D \cdot \mathrm{grad} C$$

where

J = ion flux, mol s^{-1} cm^{-2}
grad C = ion concentration gradient, mol/cm^3
D = diffusion coefficient cm^2/s

When D is constant, the ion concentration independent of C, and the concentration gradient linear, the relationship becomes

$$J = k \cdot (C_1 - C_0)$$

where k is the *mass-transfer coefficient* (cm/s) given by D/δ. Measurement of the mass-transfer coefficient of new and used resins can provide useful information about their ability to operate under critical conditions [21].

Diffusion in the resin phase is always slower than that in solution, due to the obstruction created by the resin matrix. Highly cross-linked materials have smaller diffusion coefficients.

For a comprehensive treatment of diffusion and rate laws in ion exchange, the reader is referred to [19], [20]. The remainder of this chapter deals with practical consequences of the above kinetic principles.

6.2. Kinetic Curves

As with ionic equilibria, exchange kinetics are difficult to appreciate from a purely mathematical point of view. However, exchange rates for different resins can be compared under the same conditions.

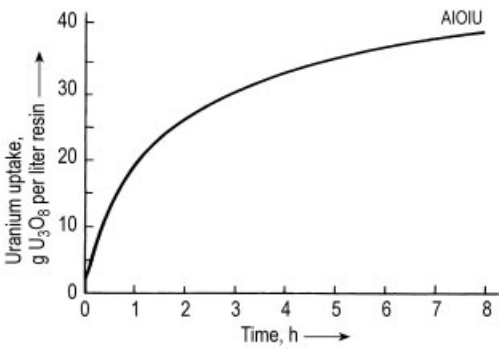

Figure 16. Kinetics of Duolite A 101U in uranium extraction

Rate of Ion Uptake (Exhaustion Rate). The exhaustion rate of a resin through which a given solution flows is measured under standard conditions. In the general case of a single ion species, the exhausted fraction X_i of the resin is plotted vertically against time (Fig. 16). The example illustrates the recovery of uranium from a sulfuric acid leaching solution in a stirred tank with strongly basic exchangers in the sulfate form. Duolite A 101U is a high-capacity polystyrene material.

Kinetic Leakage. The concentration of ions that is not taken up by the resin but passes into the treated solution is termed the ion leakage. Kinetic ion leakage can be measured as a function of specific flow rate, which is the ratio of the flow rate of the solution to be treated to the volume of resin. The specific flow rate is expressed as cubic meters of solution per hour per cubic meter of resin or as bed volume (BV) per hour.

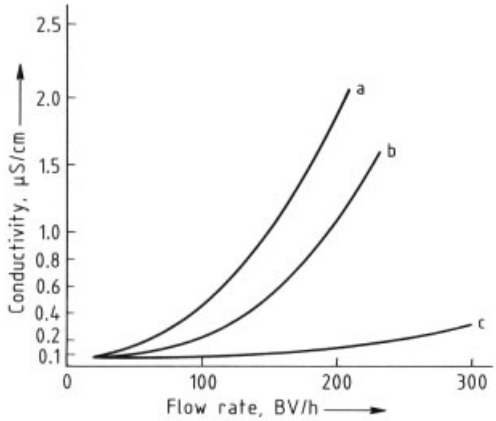

Figure 17. Kinetic leakage: conductivity vs. flow rate for two resin mixtures at two concentrations
a) Used cation resin, NaCl 16 mg/L; b) New cation resin, NaCl 16 mg/L; c) Used and new resins, NaCl 1.7 mg/L

In the example shown in Figure 17, new and used macroporous strongly acidic cation-exchange resins were tested for their ability to remove sodium from dilute sodium chloride solutions percolated through a mixed bed made of the cation resin and of a new, highly regenerated nuclear-grade anion resin. The conductivity of effluent water was measured. Both mixtures give identical low leakage for the more dilute sodium chloride solution (1.7 mg/L, curve c). However, the mixture containing used cation resin is more sensitive to the flow rate than that with new cation resin when the sodium chloride

concentration is increased to 16 mg/L (curves a and b); both mixtures then become strongly sensitive to flow rate.

This type of kinetic test is useful to assess resin performance, particularly when very high-purity water is required [22], [23].

6.3. Strongly Acidic or Strongly Basic Resins

6.3.1. Film Diffusion

At concentrations up to 10 meq/L and flow rates up to 120 m/h used in water treatment, diffusion rates through the resin mass are much greater than through the surrounding film. The film thus controls the rate of exchange, and the process exhibits film-controlled kinetics. When sodium chloride solution passes through a column of resin, originally in H^+ form, the concentration of the ion under consideration in the emergent liquor exhibits variations of the form shown in Figure 18.

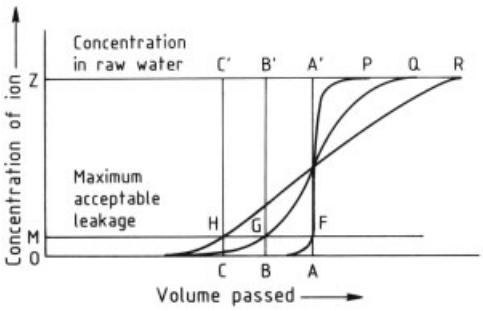

Figure 18. Film kinetics: exhaustion curves and volumes passed at maximum leakage
For explanation see text.

If the flow rate is slow enough, equilibrium is established as the solution reaches a new layer of the resin. The concentration in the emergent liquor is represented by the curve OFP. At F, the breakthrough point, the concentration reaches its maximum permitted value and flow stops. If it were continued up to P where the concentration in the emergent liquor equals that of the raw solution, the resin would be 100 % exhausted.

In industrial practice, flow is stopped when the concentration of the ion under consideration in the emergent liquor reaches a small fraction (e.g., 1 %) of the concentration in raw solution. At F, a volume A of the solution has flown through the column. The capacity used is given by the area OFA′Z, which differs very little from the area OFPZ representing the total capacity of the resin.

Figure 18 can also be taken to represent the progress of the exhaustion wavefront through a resin column. In the slow flow described above, the wavefront is only slightly diffuse, each successive layer of resin being almost completely exhausted before leakage occurs. As flow rate increases, equilibrium is no longer reached and the exhaustion curve OGQ is obtained. If the process is continued beyond the breakthrough point G as far as Q, the area OGQZ gives another measure of the total capacity of the resin. If, however, the process is stopped at G, the capacity used is only that given by OGB′Z, which is less than the total capacity.

If the flow rate is increased further, a curve such as OHR is obtained, which has a still smaller operating capacity at the new breakthrough point H.

Because the Nernst film thickness is an inverse function of flow rate, the film becomes thinner as the flow rate increases. Thus, doubling the flow rate does not mean that the leakage curve is spread out over twice the distance. In film-controlled kinetics, capacity depends very little on the rate of presentation of ions or kinetic load (i.e., the product of flow rate and concentration in the solution to be treated). In the complete absence of kinetic effects, the operating capacity is entirely independent of flow rate; therefore, the volume of the exhaustion stage is inversely proportional to the ionic concentration. As described in Section 5.3, the operating capacity of a strongly acidic resin depends mainly on the amount of regenerant. For a given quantity of regenerant, a corresponding capacity is available which is less than the total capacity.

If the exhaustion time is too short (i.e., if resin volume is small and flow rate per unit volume is high), the operating capacity is less than the available capacity. This reduction in operating capacity (which is mainly confined to strongly and weakly basic anion resins) can increase as the resin ages due to partial blocking of its pores with organic matter absorbed during several years of service. Fouling of this type reduces ion diffusion rates. In designing a

demineralizing plant, the anion-exchange column should have sufficient capacity to allow the plant to run for at least 8 h at the maximum loading rate.

6.3.2. Particle Diffusion

As the concentration of ions in solution increases, the mass-transfer rate through the film rises until it exceeds the diffusion rate through the resin beads. Diffusion through the resin then becomes the controlling factor, and the system is said to exhibit particle-controlled kinetics. This condition occurs mainly during regeneration of resins with solutions having concentrations between 1 and 3 N.

Breakthrough curves are similar to those in Figure 18, except that the length of the wavefront is a linear function of flow rate. Because of the high concentration gradient through the resin, the whole process is much faster than the exhaustion stage. Virtually complete equilibrium can be achieved in 15 min, but if a shorter regeneration time is used, operating capacity can be significantly reduced. Nevertheless, small automatic water softeners sometimes have brine injection times as short as 5 min.

6.4. Weakly Acidic or Weakly Basic Resins

Weakly acidic or weakly basic resins are not fully ionized in the regenerated form. As a result, the adsorption process involves not only diffusion through the resin but also the formation or destruction of covalent bonds, which are slower processes. The exchange rate of these resins is therefore controlled by slow particle diffusion and the resin cannot be exhausted as quickly as with strongly ionized materials. Weakly acidic and weakly basic resins are said to have slow kinetics: their capacity depends strictly on the rate of presentation of ions.

As shown in Figure 19, the carboxylic resin Amberlite IRC 86 requires a cycle time of over 24 h to utilize its full capacity (in this example, the ions to be removed from solution have a concentration of 5 meq/L). In practice, usually less than half the total capacity of weakly acidic resins is utilized. The kinetic effect is not as dramatic for weakly basic resins, which are usually operated at ca. 75 % of their total capacity.

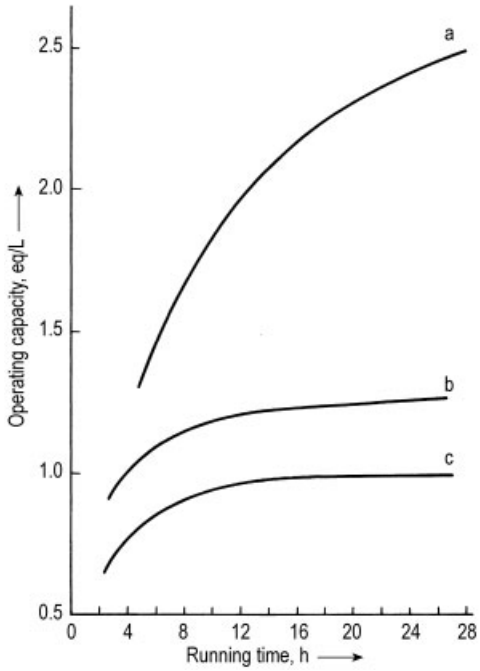

Figure 19. Effect of cycle time on usable capacity for a weakly acidic cation resin (a) and two weakly basic anion resins (b, c)
a) Amberlite IRC 86 (weakly acidic); b) Amberlite IRA 67 (weakly basic, acrylic); c) Amberlite IRA 96 (weakly basic, styrenic)

7. Practical Results of Ion-Exchange Equilibrium and Kinetics

As shown in Chapter 6, under normal conditions the exchange process is always incomplete in both the service and the regeneration stages, which means that the total capacity of the resin can never be fully used.

7.1. Operating Capacity, Regeneration Efficiency, and Regenerant Usage

Calculation of operating capacity must take into account the following:

Raw water analysis
Required quality of treated water (acceptable leakage)

Service flow rate
Temperature of water to be treated
Type and amount of regenerant
Regeneration flow rate
Regenerant temperature
Required duration of cycle

The concept of operating capacity has been introduced in Section 3.2 and developed in Section 5.3 for the special case of water softening ($Na^+ - Ca^{2+}$ equilibrium), in both instances starting from 100 % regenerated resin. In practice, this is seldom the situation because an enormous amount of regenerant would be required to remove all the ions from the resin that were taken up during the previous cycle. The variable that has the greatest effect on operating capacity is therefore the amount of regenerant used.

Regeneration efficiency is defined as the ratio of the operating capacity to the amount of regenerant used; both expressed in equivalents per liter. The reciprocal of this efficiency is known as the *regenerant usage* or *regenerant ratio* and is always >1. In practice, these quantities are often expressed as percentages.

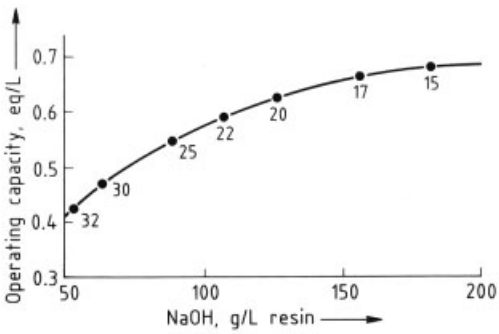

Figure 20. Operating capacity of a strongly basic type 1 resin as a function of regenerant dosage (caustic soda)
Numbers on the curves indicate percentage of regenerant efficiency.

With most strongly acidic or strongly basic resins more than 60 % of their total capacity is seldom used. As an example, the total capacity of the strongly basic type 1 resin shown in Figure 20 is ca. 1.2 eq/L. A basic operating capacity of 0.46 eq/L can be obtained by using 60 g of caustic soda per liter of resin (1.5 eq/L). In this case, the regeneration efficiency (0.46/1.5) barely exceeds 30 %. By doubling the amount of regenerant, a basic operating capacity of 0.615 eq/L can be achieved: this is only one-third greater than before, and regeneration efficiency falls to 21 %. Thus, such a resin is seldom operated at more than 50 % of its total capacity. However, these very strongly basic type 1 materials are less easily regenerated than type 2 resins and therefore yield a lower operating capacity for the same amount of regenerant than the latter. Weakly acidic and weakly basic resins, on the other hand, can be completely regenerated, but their operating capacity is limited by their poor kinetics (Section 6.4).

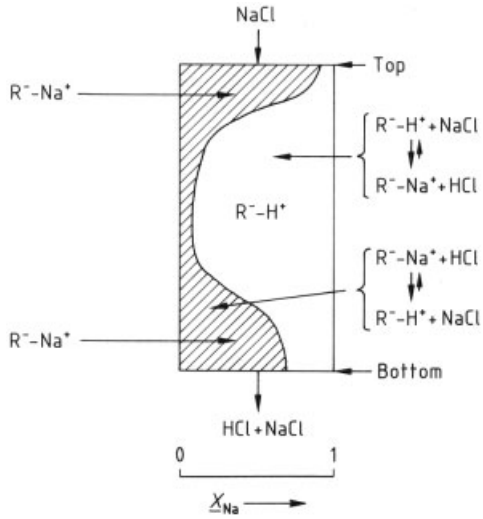

Figure 21. Ion leakage due to self-regeneration in a coflow regenerated column
For explanation see text.

7.2. Permanent Leakage

Strongly Acidic and Strongly Basic Resins. When strongly basic or strongly acidic resins are returned to service after regeneration, they are only partially regenerated. When coflow operation is used (i.e., when flow is in the same direction in the service and regeneration stages), the most poorly regenerated layer is usually the lower one through which the treated solution emerges. This affects the quality of the treated solution due to self-regeneration. In practice, the bed does not consist of two distinct (completely regenerated, completely exhausted) regions, but has a continuously varying level of regeneration along the column. Figure 21 shows the exchange reactions that occur in the column during the

service stage, in which salts from the solution (represented here by NaCl) are retained by the regenerated resin layers. Each Na^+ ion that is taken up ejects an H^+ ion from the resin. The H^+ ions move down the column and reach a region where the resin is only partially regenerated. Reverse exchange then takes place (self-regeneration), leading to leakage of Na^+ ions into the treated water.

Figure 22 shows the operating capacity (A) and sodium leakage (B) of Amberlite IR 120 (a standard, gel-type, strongly acidic cation-exchange resin) as a function of regenerant level.

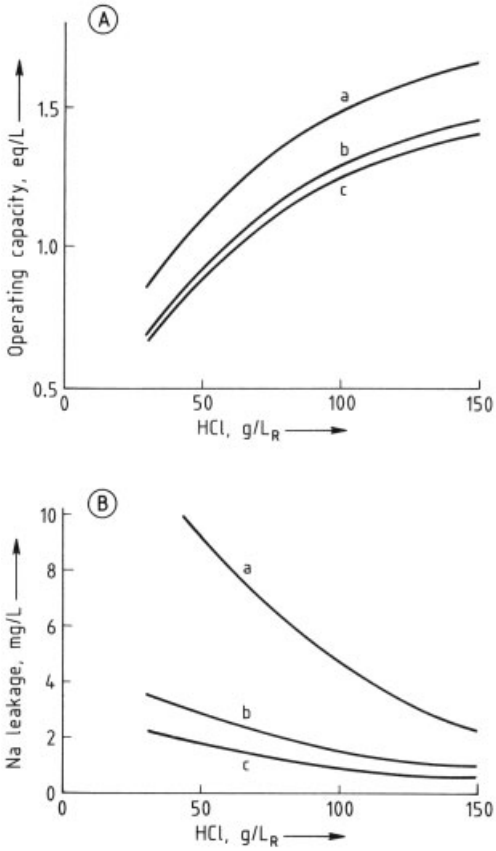

Figure 22. Capacity (A) and sodium leakage (B) of a standard strongly acidic cation resin (Amberlite IR 120) regenerated in coflow mode.
Sodium (NaCl) content with respect to total cation concentration: a) 100 %; b) 50 %; c) 25 %Flow rate, 10 BV/h; total salinity of raw water, 10 meq/L; acid dosage expressed as 100 % HCl per liter of resin; water temperature, 20 °C.

In raw water, the main cations are usually sodium, calcium, and magnesium. The selectivity of strongly acidic resins for these cations increases in the order $Na^+ < Mg^{2+} < Ca^{2+}$ (Section 5.2). Therefore, chromatographic separation occurs in the resin column, and the sodium ion is the first to emerge in the treated solution, both other cations being more strongly held by the resin. This explains the high leakage obtained for water having only sodium as a cation (Fig. 22 B, curve a).

Because the quality of demineralized water is generally expressed in terms of its electrical conductivity, sodium leakage from the cation-exchange column is the main contributor to the conductivity of treated water. A sodium leakage of 1 mg/L corresponds to 9 µS/cm conductivity (as NaOH). In coflow regenerated systems, the level of regenerant is chosen according to the desired conductivity. For high-sodium water, the only way to obtain low leakage is to use counterflow regeneration (see Section 8.2).

Weakly Acidic and Weakly Basic Resins. Weakly acidic and weakly basic resins are almost 100 % regenerated at the start of the service stage, but their exchange rate is relatively low so that some ions are not taken up and pass through the bed (i.e., kinetic leakage; see Section 6.2).

7.3. Water Analysis

Calculations for designing water treatment plants that use ion-exchange resins involve some simple concepts related to the composition of the raw water.

Units. Ion concentration is usually measured in equivalents per liter. The following units are also used to express ionic concentrations in water:

the French degree: 1° f = 0.2 meq/L
the German degree: 1° dH = 0.357 meq/L (which corresponds to 10 mg of CaO per liter)

The following units are used in Great Britain and other English-speaking countries:

1 g of $CaCO_3$ per liter = 0.02 eq/L
1 mg of $CaCO_3$ per liter = 0.02 meq/L

Although resin capacities are internationally expressed in equivalents per liter, in the United

States capacity values are widely given in kilograins (kgr) where 1 kgr as calcium carbonate per cubic foot corresponds to 0.0458 eq/L.

Water Composition. For ion-exchange calculations, the exact composition of the solution to be treated must be known. The composition of a typical water is given in Figure 23. The total concentration of all the anions and cations in solution is known as the *total dissolved solids* (TDS). Ions normally encountered in water treatment are the cations Na^+, Ca^{2+}, Mg^{2+}, and the anions OH^-, CO_3^{2-}, HCO_3^-, Cl^-, NO_3^-, SO_4^{2-}. Other ions may be present (K^+, NH_4^+, Mn^{2+}, Fe^{2+}), but their concentration in natural water is usually very low. For ionic equilibrium, the total cation and anion concentrations measured as equivalents per liter must be equal.

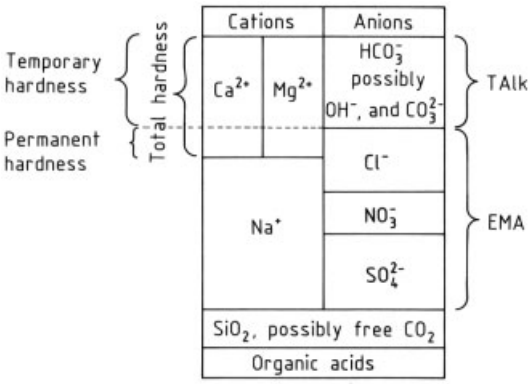

Figure 23. Composition of water

Both Ca^{2+} and Mg^{2+} ions (and, possibly, Fe^{2+} and Mn^{2+}) are classified as ions producing hardness, and their concentration gives the *total hardness* (TH) measured in milliequivalents per liter. In water softening, Ca^{2+} and Mg^{2+} are exchanged for Na^+. The Cl^-, NO_3^-, and SO_4^{2-} ions are grouped together, and their total concentration is called the *equivalent mineral acidity* (EMA), or *free mineral acidity* (FMA) after cation exchange.

The total concentration of OH^-, CO_3^{2-}, and HCO_3^- ions, measured in equivalents per liter gives the *total alkalinity* (TAlk). This is often referred to as m-Alk or the *m* value (alkalinity to methyl orange):

$$\text{TAlk} = m = \left[HCO_3^-\right] + \left[CO_3^{2-}\right] + \left[OH^-\right]$$

When TH exceeds TAlk, the difference (TH−TAlk) is called the *permanent hardness,* whereas TAlk is known as the *temporary (or bicarbonate) hardness*. When TH is lower than TAlk, only temporary hardness (no permanent hardness) exists. For water with alkaline pH values, the *caustic alkalinity*, usually described as p-Alk or the *p* value (alkalinity to phenolphthalein) is measured separately.

Carbonate Equilibrium.

The *p* value includes all anions with a p$K > 8.3$, i.e., in practice the OH^- ions and a value arithmetically equal to half the total carbonate present because CO_3^{2-} is first neutralized to HCO_3^-, which itself has a pK of only 6.7:

$$p = \left[OH^-\right] + \left[CO_3^{2-}\right]/2$$

In a demineralization system, where the first step is usually cation exchange, ions of *p* alkalinity must not be considered in the anion-exchange balance, because they are neutralized by the H^+ ions produced in the cation-exchange part of the process:

$$OH^- + H^+ \longrightarrow H_2O$$

$$CO_3^{2-} + 2H^+ \longrightarrow CO_2 + H_2O$$

As demonstrated in Section 4.2, carbon dioxide is taken up by the strongly basic resin as monovalent bicarbonate, but must be regenerated as a divalent species:

$$RHCO_3 + 2\,NaOH \longrightarrow ROH + Na_2CO_3 + H_2O$$

Because in the series OH^-, CO_3^{2-}, HCO_3^-, and CO_2, only two adjacent species can coexist in water, raw water does not contain free carbon dioxide when *p* is positive.

In practice, the equivalent concentration of carbon dioxide to be considered after cation exchange (and before the degasifier, if any) is

$$[CO_2] = m - p + [\text{free}\,CO_2]$$

The case in which $p = 0$ is the most common one, except when the raw water is lime-decarbonated. Besides, free carbon dioxide exists only with a pH close to or lower than neutral.

The concentrations of nonionized substances must also be known; these include free silicon

dioxide and carbon dioxide. The amount of organic substances present in the raw water should be considered on the anion side, because they are mainly acidic. Concentration is usually expressed in terms of the quantity of potassium permanganate needed to oxidize these substances (milligrams of $KMnO_4$ per liter of solution). The amount of organic matter determines the choice of anion-exchange resin (see Section 10.1). The permanganate value obtained in the oxidizing test depends on the test method. Here, oxidation is assumed to be performed with 0.0125 N potassium permanganate in the presence of sulfuric acid, the solution being boiled for 10 min. The composition of a typical water is given in Figure 23.

Fouling with Organic Material. The most common organic constitutents in natural waters are the high molecular mass carboxylic acids, humic and fulvic acid. These large molecules enter the ion-exchange resins and are trapped in the most highly cross-linked regions, their chains becoming tangled with the resin matrix. This eventually reduces the capacity of anion exchangers. Rinsing becomes more difficult, and problems arise with the quality of treated water because the carboxylic character of the organic matter means that, as the pH varies, caustic soda is initially taken up and then leached out. The problem is aggravated by synthetic ionic surfactants.

Although qualitative and quantitative assessment of organic materials is difficult, their concentration is usually expressed in milligrams of potassium permanganate required per liter of water to oxidize them under the given conditions. The *fouling factor N* is the quantity of organic matter (in milligrams per liter of $KMnO_4$) divided by the total anion concentration (in milliequivalents per liter).

Anion resins can be listed in order of their resistance to fouling, starting with the least resistant:

Polystyrene, gel type 1	2
Polystyrene, macroporous type 1	4
Polyamine, weakly basic	4
Polystyrene, gel type 2	6
Polystyrene, macroporous type 2	8
Polyacrylic, strongly basic	20
Polystyrene, weakly basic	20
Polyacrylic, weakly basic	40

These values merely indicate the fouling risk of each individual resin and give no information about the amount of organic material taken up. Anion resins, even those with high fouling resistance, do not remove all organic matter from the influent solution. Reversible uptake can range from 20 to 80 %; a reasonably safe figure is 50 %.

If N is too large (e.g., > 8 for a strongly basic polystyrene material), the risk of fouling can be reduced by increasing the volume of resin. The working capacity of the resin will be less than its potential operating capacity. Weakly basic resins can also protect strongly basic resins against fouling.

7.4. Calculations in the Design of Ion-Exchange Plants for Water Purification

General Method. The various types of combination process, the criteria used in choice of resin, and the recommended operational conditions are considered in Chapter 9. The operating capacity for water softening is greater when the amount of regenerant is large, the sodium ion (self-regenerant) concentration is low, and the flow rate is low.

Resin manufacturers provide standard charts and curves for each type of resin, enabling calculations to be made of the volume of each resin, amount of regenerant, operating capacity, and leakage. The starting point is normally provided by a curve that applies to standard conditions, and corrections are made for each condition that differs from the set standard. The method of calculation is too complicated to be described here; only the main factors are considered.

Weakly acidic carboxylic resins have a very high capacity for divalent ions (Ca^{2+} and Mg^{2+}): they remove temporary hardness (see Section 7.3) and are normally used only when the temporary hardness is high and the total hardness of the water is even greater (TH > TAlk). Their relatively slow exchange rate means that they are very sensitive to the specific flow rate.

Regeneration is carried out with an amount of acid calculated to be slightly in excess of the design capacity, with a regenerant ratio on the order of 105 – 110 %. These resins are therefore very efficient.

Strongly Acidic Sulfonic Resins. In hydrogen exchange, capacity depends on the type of acid used for regeneration: hydrochloric acid is the most efficient because sulfuric acid is not completely dissociated at the concentration normally used. In addition (this also applies to carboxylic resins), if the resin has taken up a lot of calcium, the sulfuric acid must be highly diluted to avoid precipitation of calcium sulfate during regeneration. The amount of regenerant is chosen according to the permitted ion leakage: in other words, it depends on the permissible level of electrical conductivity in the treated water (see Section 7.2). The greater the amount of regenerant, the lower is the electrical conductivity and the greater is the operating capacity (see Fig. 22). The operating capacity of the resin depends on the cation composition of the water, more in relation to proportions than to absolute values. Reference should be made to manufacturers' curves.

Weakly basic resins also have slow kinetics and are rate-sensitive. In addition, their capacity depends on the composition of the water, increasing as the concentration of strong acids decreases. Only small quantities of weak acids are removed. However, carbon dioxide improves the capacity of weakly basic styrene resins so that, if an atmospheric degasser is used it should be located between the weakly basic and strongly basic exchange steps. Weakly basic acrylic resins have a higher pK value, so that they effectively remove carbon dioxide from water, and the benefit of the degasifier is lost if it is located downstream.

The quantity of regenerant normally used is on the order of 130 % of the operating capacity of the resin.

Strongly basic resins remove strong and weak acids. Because the uptake of silica is poorer than that of other anions, it is the first to "leak". The regenerant level thus depends on the acceptable silica leakage. In addition, because silica has a tendency to polymerize on the resin, regenerating with hot sodium hydroxide is sometimes worthwhile. However, only type 1 materials withstand high temperature (see Section 3.3).

In practice, regenerant requirements are from 150 to ca. 500 % (sometimes > 1000 % when the strongly basic resin is preceded by a weakly basic one), so the efficiency of these resins is only moderate. However, a new type of strongly basic resins was developed by Duolite International, Bayer, and Rohm and Haas in the mid 1980s. These are partially quaternized weakly basic resins, which thus exhibit a strong base capacity of ca. 40 % (i.e., enough to efficiently remove silica from waters containing up to about 15 % SiO_2), whereas they offer a regenerability close to that of weakly basic resins. Examples of such resins are Amberlite IRA478 and Lewatit AP246.

Calculation of Resin Volume. From the calculated operating capacities, the volume of each type of resin can be determined as a function of the operating time chosen between two regenerations:

$$V = (Q \cdot t \cdot S) / C$$

where

V = resin volume, m^3
Q = flow rate, m^3/h
t = operating time, h
S = salinity to be adsorbed by the resin, eq/m^3 of water or meq/L
C = operating capacity of the resin, eq/m^3 of resin

Service Water. The operating cycle for water treatment includes backwashing, regeneration, and rinsing stages, all of which consume water. The salinity of this service water is thus added to the ionic load of the production cycle and reduces resin capacity. The volume of resin must therefore be increased in proportion to the amount of service water used. The additional volume depends on the salinity of the raw water, which varies between 5 % (low-TDS water) and 35 % (semibrackish water).

Specific Flow Rate and Cycle Duration. The specific flow rate should not be too low because a very even flow must always be obtained; irregularities in flow may cause "channeling" in the resin bed, thereby impairing complete exchange. Too high a rate produces an excessive head loss. The specific flow rate Q/V should be between 4 and 40 BV/h. If this is not the case, the volume V must be appropriately adjusted by shortening or lengthening the operating time t.

7.5. Example of Calculation

In this section, the calculation of a system for demineralizing water, based on Amberlite resins is described. One solution of the problem is fully developed here, further options are discussed in Section 9.3 and column design in Section 10.2.

7.5.1. Principle

A set of curves for each type of resin is loaded into a computer, equipped with a program designed to yield the optimum system, given the characteristics required for the treated water and the consumption of regenerants. The calculation can also be carried out by hand. Basic data include the following:

- An analysis of the raw water
- Flow rate to be treated
- Required quality of the treated water
- Combination system required
- Types of resin to be used
- Length of the production cycle
- Types of regenerant
- Temperature of the caustic soda regenerant

The calculation leads to a determination of the volume, operating capacity, and regeneration rate for each resin. The diameter of each filter is also calculated as a function of its volume. A set of limiting criteria controls the calculation. Examples of these are the following:

1) The specific flow rates for each column must lie between 4 and 40 BV/h.
2) The linear velocity of flow should not exceed a value leading to high head loss across the bed; this linear flow rate is the ratio between the flow rate (in cubic meters per hour) and the area of the column (in square meters). The recommended head loss limit is 100 to 150 kPa.
3) The depth of the resin beds must be between 0.7 and 2.5 m.
4) The amounts of organic matter determine the choice of anion-exchange resins (see Section 7.3).

The above values should serve as guidelines for industrial water treatment. For special cases (e.g., condensate polishing or specific ion-exchange systems), the process limits are considerably different.

The demineralization of 80 m^3/h of water with a salt content of 8 meq/L can be used as an example. Each of the items is treated in the set of basic data (Table 6) and in the results of the calculation [demineralizing (Table 7) and polishing (Table 8) units].

7.5.2. Basic Data

Basic data are summarized in Table 6.

Analysis of Water. The water analysis must be available in detail and be ionically balanced.

Required Quality. Quality is expressed in terms of the electrolytic conductivity of the treated solution and the residual concentration of silica, which is important in boiler feedwaters.

Combination Process for Demineralization. The type of ion-exchange line and regeneration system is chosen according to the composition of the raw water, the flow rate to be processed, and the required water quality (see Chap. 9). Each of the pairs strong acid – weak acid and strong base – weak base resins is regenerated in thoroughfare, i.e., the weakly functional resin is regenerated with the excess from the strongly functional resin.

Role of Each Resin.

1) Amberlite IRC86, a weakly acidic cation resin, removes the temporary hardness, i.e. the calcium and magnesium ions, up to the limit of TAlk (4 meq/L).
2) Amberlite IR120, a strongly acidic cation resin, removes the rest of the cations ($Ca^{2+} + Mg^{2+} + Na^{+}$), i.e., $8 - 4 = 4$ meq/L.
3) Amberlite IRA96, a weakly basic anion resin, removes the EMA ($Cl^{-} + NO_3^{-} + SO_4^{2-}$), i.e., 4 meq/L.
4) Amberlite IRA458, a strongly basic anion resin, removes the other anions after degassing together with residual carbon dioxide (0.3 meq/L) and the silica (0.2 meq/L).

In addition, the two anion-exchange resins remove a considerable proportion of the organic matter contained in the water. Since the initial amount is quite large, a strongly basic polyacrylic material is chosen to avoid fouling.

Table 6. Examples of design data for a plant calculation

1) Water analysis, meq/L					
Ca^{2+}	5.0	Cl^-	3.0	P-Alk	0.5
Mg^{2+}	1.0	NO_3^-	0.0	M-Alk	4.0
$Na^+ + K^+ + NH_4^+$	2.0	SO_4^{2-}	1.0		
Total cations	8.0	EMA	4.0		
Silica	12 mg/L as SiO_2 (= 0.20 meq/L)				
Organic matter	10 mg/L $KMnO_4$				
Residual CO_2 (after degassing)	0.30 meq/L				
Water temperature	15 °C				
2) Production					
Running time (production) = 16 h	Net flow rate		80 m^3/h^3		
	Gross flow rate		82.7 m^3/h		
	Net throughput		1280 m^3		
	Gross throughput		1324 m^3		
3) Required water quality					
	Before polishing		After polishing		
Conductivity µS/cm	1.0		0.1		
SiO_2, ppm	0.1		0.01		

Table 7. Calculated results for the demineralization unit

Parameter	Amberlite resin type			
	IRC86	IR120	IRA96	IRA458
Volume per line, L	2800	4150	3800	3850
Gross ionic load, eq	5371	5371	3736	2268
Overrun, % of ionic load	0		30	
Operating capacity, eq/l	1.92	1.29	0.98	0.59
Regeneration with	HCl	HCl	NaOH (25 °C)	NaOH (25 °C)
Regenerant ratio, %	110	210	125	306
Regenerant level, g/L		99		72
Concentration, wt %		5.0		4.0
Regenerant mode :	Amberpack	Amberpack	Amberpack	Amberpack
Suggested diameter, mm	1900	1900	1900	1900
Resin height, mm	988	1464	1340	1358
Linear velocity, m/h	29	29	29	29
Specific flow-rate BV/h	29.7	20.0	21.9	21.6
Pressure drop, kPa	38	53	49	40

7.5.3. Demineralization Unit

Duration of Production. The figure of 16 h was chosen for hydraulic reasons because the specific flow rate must not exceed 40 BV/h.

Total Ionic Load. This is the product of the gross production (percolated volume in cubic meters, taking into account the water required for regenerant dilution and rinse) and the salt content taken up by each exchange resin (in equivalents per cubic meter).

Operating Capacity. This is calculated for each resin from the curves supplied by the manufacturer.

Volume. The volume of each resin can be determined as a function of the operating time chosen between two regenerations from the formula:

$$V = (Q \cdot t \cdot S)\,/C$$

where $(Q \cdot t \cdot S)$ is the ionic load and C the operating capacity of the resin. The volume of resin is calculated by dividing the ionic load by the operating capacity.

Amount of Regenerant. This is determined from the manufacturer's curves as a function of the raw water composition and the required quality of the treated water. Both resin couples

(WAC – SAC and WBA – SBA) are regenerated in thoroughfare (see Section 9.3), i.e., the fresh regenerant first passes through the strong resin, then through the weak resin.

Regenerant Temperature. To improve elution of the silica, the caustic soda regenerant may be heated to 45 °C, but only if the strongly basic exchange resin is of type 1. For type 2 or polyacrylic materials, the temperature should not exceed 30 °C. Since Amberlite IRA 458 is a polyacrylic resin, regeneration at 25 °C is preferred.

Acid Concentration, Flow Rate, and Contact Time. Hydrochloric acid is generally used at a concentration of between 5 and 8 wt % and caustic soda at 4 wt %. Recommended contact time for all regenerants is between 15 and 60 min. If sulfuric acid is used, care must be taken that its strength is low enough to avoid precipitation of calcium sulfate on the cation exchange resins. Regeneration in this case is often carried out in several stages with sulfuric acid concentration increasing from 0.8 to 7 wt %.

Column Diameter. The diameter is chosen according to the volume of resin and flow rate so as to obtain a bed depth between 0.7 and 2.5 m of resin and a pressure drop of less than ca. 120 kPa across the bed.

Overrunning the Weakly Basic Resin. Calculating resin volumes on the basis of the ionic loads (see page 30) results in a large discrepancy between the volume of the weakly basic resin and that of the strongly basic resin. This is because the quantity of weak acids (carbon dioxide and silica) is small compared to that of strong acids. The volume of strongly basic resin would be less than 1000 L, and the corresponding specific flow rate (ca. 80 BV/h) would exceed the maximum recommended value of 40 BV/h. A better solution, always used in real projects, is to "overrun" the WBA resin. The volume of the WBA resin is made deliberately too small, so that part of the ionic load of the strong acids must be removed by the SBA resin, the amount of which, as a result, is made larger. In practice, the resin volumes are balanced in such a way that they are approximately equal. When the bicarbonate hardness in the feed water is very high (say 75 % or more of the dissolved solids), overrunning of the WAC resin is also possible.

Two Resins in One Column. As is discussed in in Chapter 10, modern ion-exchange systems use packed beds in which the WAC/SAC and the WBA/SBA resin couples are placed in the same column. This reduces considerably equipment costs.

Results of The Calculation. Using the overrun technique and packed bed design, and taking into account the higher capacity of the Amberjet resin and a specific rinse procedure that reduces the amount of waste, the final results of the calculation are summarized in Table 7.

Table 8. Calculated results for the polishing unit

Parameter	Amberjet 1200 H	Amberjet 4200 Cl
Volume per line, L	1000	1000
Production time, h	200	200
Max. throughput, m^3	16000	16000
Regeneration with	HCl	NaOH (25 °C)
Regenerant level, g/L	100	100
Concentration, wt %	5.0	4.0
Suggested diameter, mm	1500	1500
Resin height, mm	1130	1130
Linear velocity, m/h	45.2	45.2
Specific flow rate, BV/h	40	40
Pressure drop, kPa	71	71

7.5.4. Polishing Unit

Calculated results for the polishing unit are listed in Table 8. The mixed bed for polishing (here a Triobed unit) is calculated in such a way as to have a specific flow rate of about 40 BV/h. Regeneration is carried out with 80 – 120 g of acid per liter of cation resin and 80 – 100 g of caustic soda per liter of anion resin.

The operating capacity is usually limited by the silica content, because the anion-exchange resin in the mixed bed almost entirely removes the silica that has leaked from the primary system. In addition, the proportion of anion-exchange resin in the mixture is generally selected between 50 and 65 %. The diameter of a mixed bed unit is calculated to give a bed depth between 0.9 and 1.5 m and a linear flow rate between 30 and 60 m/h. Because the influent to the mixed bed varies during the service run of the

primary line, the run length of the mixed bed is only approximate, and most users decide to regenerate their polishing unit after a fixed running time.

8. Industrial Use of Ion Exchange

The practical use of ion exchangers, based on the general principles discussed in the previous chapters, is described here. Although many examples refer to water treatment, other fluids are treated in a very similar way.

8.1. Description of the Ion-Exchange Cycle

Figure 24 shows a vertical section through a coflow ion-exchange column. The solution to be treated is introduced into the column through a distributor. It passes through the resin and emerges through a collector or collection system. In a conventional plant, the resin fills only half the available space so that the bed can be decompacted by an upward flow of water. This expands the resin and removes suspended matter and fragments of resin accumulated during the previous cycle.

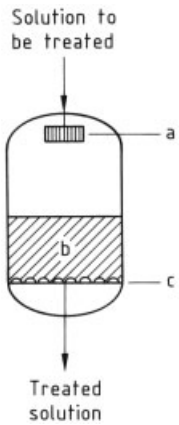

Figure 24. An industrial ion-exchange column
a) Distributor; b) Resin; c) Collector

The ion-exchange cycle is divided into four stages:

1) exhaustion (or service),
2) decompaction (or backwash),
3) regeneration, and
4) rinsing.

Exhaustion (or Service). The solution to be treated passes through the resin bed and exhausts it. As soon as the quantity of ions taken up reaches the operating capacity (i.e., the breakthrough point) and leakage reaches a predetermined limiting value, the service stage is stopped.

Backwash. After the service stage, the resin is decompacted for about 15 min by an upward current of water. This treatment also removes any particles deposited on the surface of the bed, together with any fragments of resin. The bed is then allowed to settle.

Regeneration. The regenerant solution is introduced, usually at a concentration of a few percent, and slowly percolates through the bed. This takes about 15 – 60 min.

Rinsing. The regenerant is then displaced by water at a low flow rate until the resin bed contains no more than traces of regenerant. This displacement or slow rinse stage is followed by a rapid rinse stage at a higher flow rate to remove the last traces of regenerant. The concentration of the residual regenerant is measured at the end of the operation and, as soon as this falls to an appropriate limiting value, the next cycle begins. When water is being demineralized, the electrolytic conductivity of the emergent liquor is measured.

8.2. Methods for Overcoming Equilibrium Problems

Counterflow Regeneration [24], [25]. In the traditional method, the resin is regenerated by flow in the same direction as that during the service stage (coflow). Figure 21 illustrates the exhaustion of the bed in this case. For economically acceptable quantities of regenerant, the bottom of the bed is only partially regenerated, which results in high leakage (see Section 7.2). This effect is of major importance in the demineralization of water. The leakage problem can be overcome by regenerating in the reverse direction (i.e., from bottom to top). The

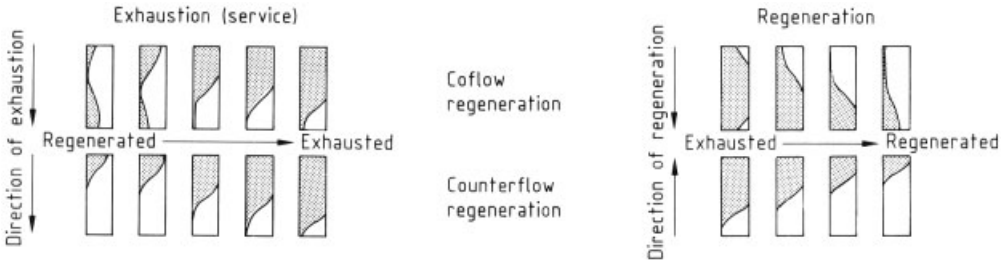

Figure 25. Effects of coflow and counterflow regeneration on the condition of resin beds during loading and regeneration Hatched areas indicate 100 % exhausted resin, and white areas, regenerated resin.

lowest resin layers reach equilibrium with fresh, uncontaminated acid and are therefore completely regenerated. This minimizes permanent leakage during the next cycle. Several counterflow regeneration systems are described in Section 10.2.2. Among these, some have an upflow exhaustion phase with downflow regeneration. The bed must be kept perfectly compacted when operating in upflow mode to maintain contact between the ions in solution and the resin beads.

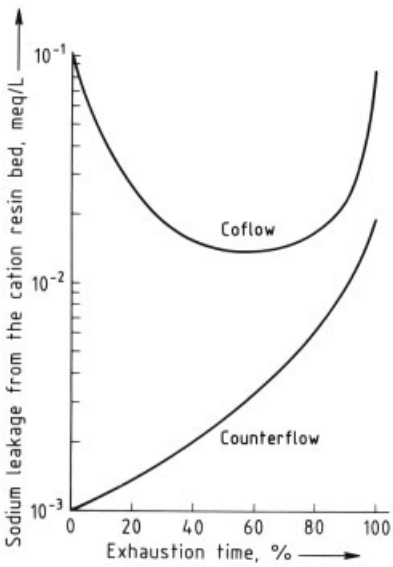

Figure 26. Effects of coflow and counterflow regeneration on ionic leakage

Reverse flow regeneration is often called "countercurrent," which implies two streams flowing in the opposite direction. The terms "counterflow" or "reverse flow" are used here, both meaning only that the direction of percolation is reversed between service and regeneration.

The difference between coflow and counterflow regeneration is illustrated in Figure 25, and their effects on treated water quality are shown in Figure 26. Two-bed demineralization with counterflow regeneration produces water with a conductivity $\leq 2.0\,\mu$S/cm, even from feedwater with a high sodium content, and the consumption of regenerant is much lower than with coflow regeneration. Figure 25 also shows that in counterflow regeneration, ions taken up by the resin are desorbed over the shortest path instead of passing through the whole bed; the lower layers are always regenerated well.

In order not to disturb the arrangement of resin layers, the bed is only backwashed every 20 – 50 cycles.

Rinse Recycling. Considerable economy can be achieved by recycling the rinse water to remove traces of salts and regenerant, which slowly diffuse out of the resin beads into the treated water and thus interfere with its quality during the early part of the service run. Each column is first rinsed until the conductivity of the rinse water decreases to a value equal to or less than that of the feedwater. The rinse water is recycled through the two columns in series in a closed circuit, so that each resin removes the last traces of regenerant from the other. The amount of rinse water used is reduced significantly and the quality of the treated water is improved.

Buffer Column. The final ion leakage at the output from the demineralization units consists of sodium hydroxide and silica (see Section 7.4). The pH is slightly alkaline, normally between 7 and 9. If a neutral effluent is required, sodium

hydroxide leakage can be taken up by a cation-exchange material, the most efficient being a weakly acidic resin. This is known as a buffer column. Since the reaction is a pure neutralization, it is very fast and efficient. Because the resin operates in the H – Na cycle, the selected carboxylic resin may swell considerably in operation, and it must have excellent osmotic stability.

Mixed Beds. When beads of a cation- and an anion-exchange resin (R_C and R_A, respectively) are mixed intimately, the normal equilibrium for each resin is displaced:

$$R_C^-H^+ + Na^+Cl^- + R_A^+OH^- \longrightarrow R_C^-Na^+ + R_A^+Cl^- + H_2O$$

This reaction produces water. The exchange process is no longer reversible and continues until completion. Mixed-bed resins in well-designed equipment reduce the concentration of dissolved salts to 0.01 mg/L, giving a conductivity of 0.055 µS/cm at 25 °C, equal to that of completely pure water. Regeneration is carried out after the two resins have been separated (see Section 10.2.4). The advantages of mixed beds are twofold: (1) they produce water of excellent quality in a single stage, and (2) final rinsing is very rapid because the regenerants neutralize each other. However, the operating capacities of the resins are quite low, and the end of the cycle (breakthrough) is very abrupt.

If the ionic concentration of the feedwater is low (i.e., for polishing), the limiting factor is the specific flow rate. Ordinary gel-type resins are used up to about 40 BV/h and macroporous resins up to 100 – 130 BV/h. The amounts of regenerant used are on the order of 100 g of acid per liter of cation-exchange resin and 100 g of sodium hydroxide per liter of anion-exchange resin. The ratios of resin volumes used ($R_AOH : R_CH$) vary between 2 : 1 and 1 : 1. If the feedwater contains neutral salts, the 2 : 1 ratio is chosen; if only traces of sodium hydroxide from the demineralization train are present, the 1 : 1 ratio is selected.

In mixed beds for general use where the feedwater has a high ionic concentration, each of the exchange resins is calculated as if it were in a separate bed. The capacities are multiplied by a reduction factor of about 0.8 to allow for imperfect regeneration caused by the resin mixing zone after backwash and separation, as well as by the mutual diluting effect of the resins.

Continuous Systems. In continuous systems (always used with counterflow regeneration), the exhausted resin layers are extracted continuously or, more frequently, semicontinuously at intervals of a few minutes. They are replaced by freshly regenerated resin introduced at the point where the treated liquid emerges from the column. The total volume of resin is reduced by rapid circulation. In large industrial demineralization plants, a constant supply of treated water is often required. With fixed-bed systems, this necessitates two alternately operating lines or a large treated-water tank to secure water supply during regeneration period. With continuous systems, a single unit can be used, which ensures that the quality of the treated liquid is virtually constant.

The plant must be completely automated. The main disadvantage of these systems lies in the complexity of their construction. Examples are described in Chapter 10.

9. Ion-Exchange Resin Combinations

Many combinations of ion-exchange units are employed in water treatment; only the most widely used are considered here and summarized in Figure 27. They include pretreatment, softening, demineralization (primary system), and polishing. Softening exchanges Ca^{2+} and Mg^{2+} for Na^+ ions, whereas demineralization exchanges all cations for H^+ and all anions for OH^- ions. Polishing removes the cations and anions that have "escaped" the first-stage demineralization.

9.1. Pretreatment

Before being allowed to percolate through an ion-exchange column, the solution to be treated must be thoroughly filtered to free it from suspended matter. Organic matter can also be partly removed by flocculation, by using ferric chloride or aluminum sulfate with the possible addition of polyelectrolytes.

When the raw water is highly alkaline, the classical process of lime softening may be used to remove alkaline salts as calcium carbonate.

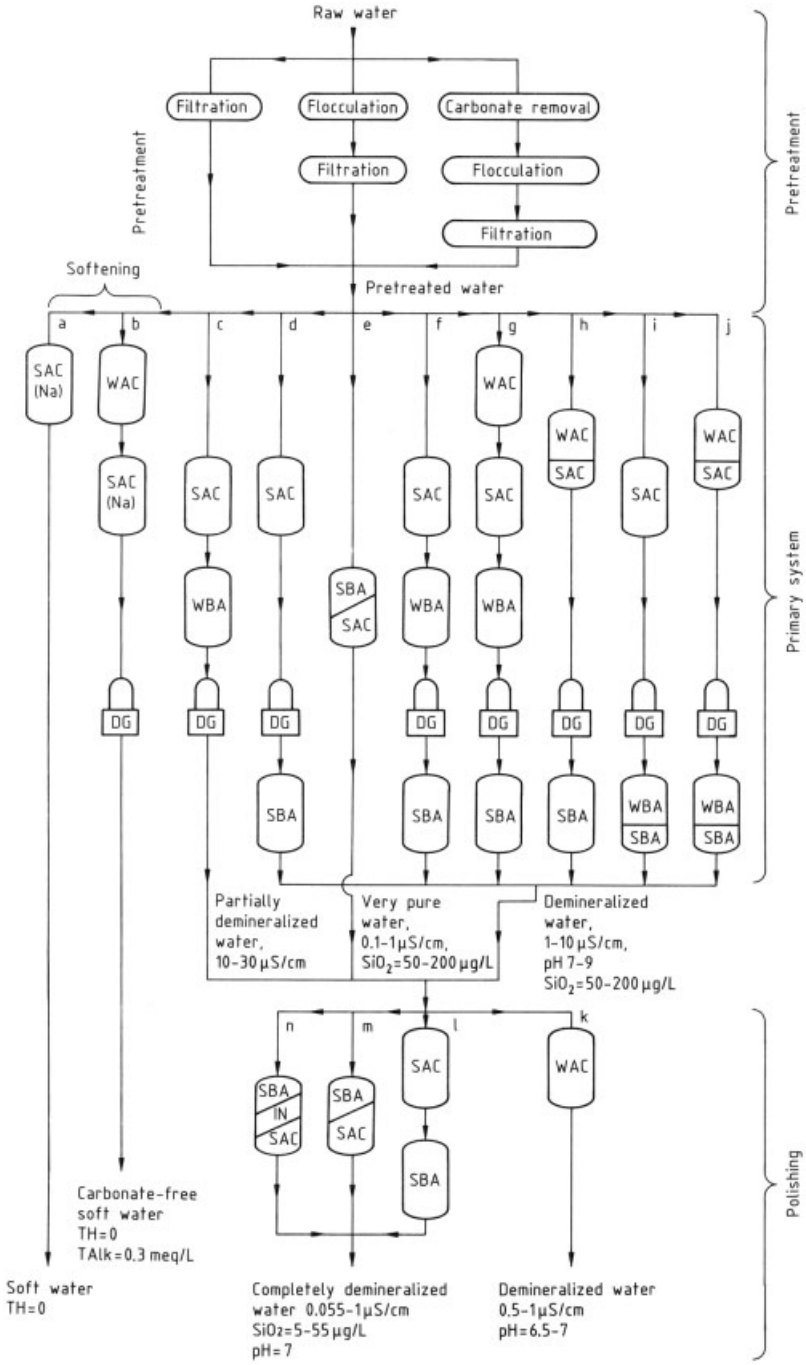

Figure 27. Combination ion-exchange systems for water treatment
DG = atmospheric degasser; IN = inert resin; SAC = strongly acidic cation exchanger; SBA = strongly basic anion exchanger; WAC = weakly acidic cation exchanger; WBA = weakly basic anion exchanger

This process, known as de-alkalization or carbonate removal, reduces the ionic load on subsequent ion-exchange units.

9.2. Softening

Pretreatment can be followed by softening (Figure 27, route a) with a single strongly acidic cation-exchange resin regenerated with sodium chloride.

Carbonate removal can also be combined with softening (Fig. 27, route b) to reduce the salinity of hard alkaline water and remove the hardness for use in low-pressure boilers and as cooling water. The pH may have to be increased from 4.5 to ca. 8 with sodium hydroxide.

9.3. Demineralization (Primary System)

The abbreviations used in this section are defined in Figure 27.

Degasser. As soon as the alkalinity of the raw water exceeds ca. 1 meq/L, an atmospheric degasser (DG) should be used in the system. When the solution flows through the cation-exchange resin, bicarbonate ions (and carbonate ions if any) are converted into carbon dioxide, which is removed in the degasser instead of passing to the strongly basic resin and being taken up there. A degasser is assumed to be present in all systems except those with mixed beds. The cation-free water at the degasser outlet contains 0.2–0.3 meq of free carbon dioxide per liter.

SAC – WBA – DG System (Fig. 27, route c). If the raw water is hard, the SAC – WBA – DG system produces water with < 5 mg of dissolved salts per liter, together with silica and ca. 15 mg of carbon dioxide per liter. If the raw water is soft, however, sodium leakage can be quite high unless the SAC is regenerated by a counterflow system. This type of arrangement is used to produce water for medium-pressure boilers and for chemical processes. The degasser may be placed after the weakly basic anion-exchange unit to obtain maximum capacity from the resin (see Section 7.4).

SAC – DG – SBA System (Fig. 27, route d). This simple system produces very pure water suitable for most applications. The water quality obtained depends on the regeneration system and the amount of regenerant. With coflow regeneration, a conductivity of 5 – 25 µS/cm and a residual silica content of 50 – 200 µg/L can be achieved, depending on the quality of raw water. With counterflow regeneration, the corresponding values can reach 1 µS/cm and 5 – 25 µg of silica per liter. Small plants are often built without a degasser.

General-Purpose Mixed Bed (SAC+SBA). A working mixed bed system (Fig. 27, route e) produces very pure water (0.1 – 1 µS/cm; 10 – 100 µg of SiO_2 per liter), but its regeneration efficiency is moderate because it is impossible to fit a degasser. The system is therefore restricted to small laboratory units or water with very low salinity.

SAC – WBA – DG – SBA System (Fig. 27, route f). This system produces the same quality of water as the SAC – DG – SBA system (route d). Its advantage lies in the higher regeneration efficiency due to the weakly basic resin. The anion exchangers are regenerated in series in the direction from SBA to WBA, so that an excess of clean caustic soda passes through the final resin, usually in sufficient amounts to regenerate the weak resin.

WAC – SAC – WBA – DG – SBA System (Fig. 27, route g). A weakly acidic cation resin can be used in all the above systems for appropriate types of raw water (i.e., high alkalinity and even greater total hardness; see Section 7.4).

The system considered here should be used for large plants dealing with highly saline water in order to use the high efficiency of the WAC and WBA resins to reduce the cost of regeneration. Acid regeneration is carried out in the direction from SAC to WAC, and sodium hydroxide flows from SBA to WBA as in route f.

Layered Beds. Weakly and strongly acidic resins can be superposed in the same column, because a carboxylic resin is less dense than a sulfonic resin. In addition, fine particle size is used for the WAC and coarse particle size for

the SAC so that the two layers are still well-separated after the backwash. The same applies to anion exchangers. Ion-exchange units containing superposed weakly and strongly active resins are called layered or stratified beds.

These dual resin systems (Fig. 27, routes h – j) have a high efficiency only if counterflow regeneration is used. They then produce water of excellent quality (1 µS/cm, 5 – 20 µg of SiO_2 per liter) with low regenerant consumption (regeneration efficiencies close to 100 %) – all in a plant with only two exchange columns. Calculation of the capacities and maximization of the performance of the system are complex operations, but they can be carried out rapidly on a suitably programmed computer.

A difficulty can occur with some weakly basic anion-exchange resins when the difference between their density and that of their strongly basic partner is too small (polyacrylic resins). A solution to this problem is to use a high-density resin as the strongly basic exchanger (see Section 3.4).

9.4. Polishing

To remove the last traces of salts, a polishing unit can be installed after the primary demineralization system.

A final *buffer column* containing a carboxylic, or sometimes a sulfonic, exchanger (Fig. 27, route k) can remove sodium leakage from an anion-exchange resin (see Section 8.2, buffer column).

Two-column polishing (Fig. 27, route l) was often used before counterflow regeneration became widespread. The sodium and silica leakages were taken up by a strongly acidic and a strongly basic exchanger in series. Regeneration was carried out from SBA2 to SBA1 and from SAC2 to SAC1, a primitive kind of counterflow system, often referred to as thoroughfare regeneration.

The mixed-bed polisher (Fig. 27, route m) is the most widely used polishing system, with its three-component variation (Fig. 27, route n; see also Section 10.2.4). With a mixed bed, a theoretical conductivity of pure water (0.055 µS/cm at 25 °C) can be achieved; ion and silica leakages are ca. 1 µg/L. Because the mixed-bed polishing system has to deal with only a low ionic load, it has a long service run, and is regenerated only every 5 – 10 cycles of the primary system. If the weak acid concentration in the raw water is low, the SAC – WBA – DG – MB system can be used and all the weak acids are then retained in the strongly basic resin of the mixed-bed polishing system.

9.5. Choice of Resin

General Selection Criteria. To ensure efficient operation of an ion-exchange installation, minimize running costs, and obtain a maximum resin lifetime, the resins must be selected carefully. Before individual resins are selected, the following questions should be answered:

1) Is the solution to be treated aggressive? Does it contain oxidants?
2) What is the operating temperature?
3) Is the flow velocity high or moderate?
4) Is the resin bed deep or shallow?
5) Is a high operating capacity required?
6) Is the expected leakage very low?
7) Are the cycles long or short?
8) Is the regenerant hot?
9) Is the regenerant very concentrated?
10) Are the resins used in a fixed, packed, or fluidized-bed unit or in a continuous system?
11) Are the resins used in stratified-bed or in mixed-bed units?
12) Does the regeneration process have a resin transfer step?
13) Is an effluent total organic carbon (TOC) strictly limited?
14) Is the application in the food-processing or nuclear industry?

The answers to these questions provide useful hints for resin selection:

1) In the presence of oxidants, a resin with a high degree of cross-linking should be used (see Section 3.3).
2) High operating temperature is critical for the choice of strongly basic resins: above 40 °C, only type 1 resins should be used (see Section 3.3).

3) For a high flow rate, physically stable resins and a relatively coarse particle size should be used.
4) With deep beds, coarse resins should be used to minimize the pressure drop.
5) If a high capacity is required, gel-type resins are preferred over macroporous types; type 2 or acrylic anion resins give a higher capacity than type 1. For condensate polishing applications, or for strongly basic resins placed downstream from a weakly basic resin, a high operating capacity is not of prime importance, but flow rate is a dominating factor.
6) For low leakage, resins with high cross-linking (i.e., high selectivity) associated with high regenerant dosage will give the lowest concentration of residuals.
7) Long cycles increase possible fouling risks and tend to compress resin beds and create large pressure drops; mechanically strong resins are thus advisable. Short cycles mean frequent osmotic shocks; in this case, macroporous resins are generally more stable.
8) For a hot regenerant, see item 2.
9) Concentrated regenerant also mean stronger osmotic shocks; see item 7.
10) Packed beds, fluidized beds, and continuous systems call for special particle sizes.
11) In stratified-bed or mixed-bed units, the particle size and density of the resins must be selected to obtain perfect resin separation after backwash.
12) Resin transfers result in mechanical stress: tough resins must be selected.
13) For low effluent TOC, special resin grades must be used that give rise to a minimum of leachables. On the other hand, if the influent solution contains organics, the anion resin must be capable of removing them reversibly so that it does not become fouled (see Section 7.3).
14) For food processing or nuclear waste treatment, special resin grades are prepared by major resin manufacturers to comply with applicable regulations or specifications.

Weakly Acidic Cation Exchangers. For water dealkalization or as a first unit in a series of demineralization steps, a high-capacity carboxylic resin should be selected. The osmotic stress for such a resin operating between the H^+ and the Ca^{2+} or Mg^{2+} forms is generally moderate, and most weakly acidic resins are mechanically very stable. In applications involving the Na^+ or NH_4^+ forms, however, osmotic stability is essential; the best choice is a resin with limited swelling and a macroporous structure.

Strongly Acidic Cation Exchangers. For softening or demineralization of water, gel-type cation-exchange resins are generally chosen. Sulfonic resins have a maximum capacity for a degree of cross-linking corresponding to 8 % DVB (Section 3.1). Gel resins with 6 % DVB can be used for water softening, and 10 % DVB resins for the production of highly pure water. Macroporous resins are restricted to special applications, continuous processes, and the treatment of condensates.

Weakly Basic Anion Exchangers. A wide variety of weakly basic anion exchangers are available with different matrices and porosities. The main types are as follows:

1) *Macroporous polystyrene resins* are the most widely used and the strongest, but the capacity of the standard grades is not very high. For high ionic loads and low flow rates, some manufacturers offer special grades with lower moisture and higher volume capacities.
2) Polyacrylic resins have a higher capacity and excellent resistance against organic fouling. They are, however, more expensive and sometimes difficult to rinse after regeneration.
3) *Phenol–formaldehyde resins* are generally restricted to special applications, but they have a strong adsorptive power for organic matter.

Figure 28 illustrates the comparative operating capacities of types 1 and 2. The capacity is plotted against an arbitrary factor $K*$ which takes into account the water salinity and the flow rate.

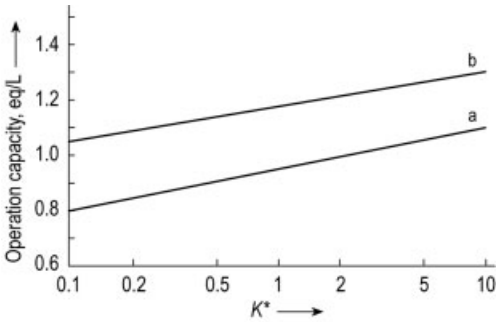

Figure 28. Operating capacities of weakly basic resin ($SO_4^{2-}:Cl^- = 1:1$)
a) Polystyrene resin; b) Polyacrylic resin $K^* = 10^4 \frac{[CO_2]^2}{(FMA)(FR)^2}$ where FMA is the free mineral acidity and FR is the flow rate in bed volumes per hour.

Strongly Basic Anion Exchangers. Like cation exchangers, strongly basic anion exchangers are gel-type resins which are widely used for water treatment because of their high capacity. Macroporous resins are restricted to continuous systems, condensate polishing and special applications.

A distinction is made between the following three types:

1) *Polystyrene resins type 1* are very strongly basic and therefore achieve extremely low silica leakage. They withstand high temperature quite well and can be regenerated with sodium hydroxide up to 50 °C. However, their regeneration efficiency is only moderate, so that their capacity is rather low and they are subject to irreversible fouling by organic matter.
2) *Polystyrene resins type 2* are slightly less basic and have a better regeneration efficiency and higher capacity (about 20 % higher than type 1). They become fouled less quickly but withstand high temperature less well (see Section 3.3). Their silica leakage is higher.
3) *Polyacrylic resins* combine some of the advantages of types 1 and 2 polystyrene resins: low silica leakage, good capacity, moderate regeneration efficiency, high physical stability, and good elution of organic matter adsorbed. However, they are 25 % more expensive than polystyrene resins and can withstand only moderately high temperature (≤ 35 °C). The partially quaternized acrylic resins described in Section 7.4 (strongly basic resins) have a particularly high capacity.

Figure 29 shows the basic operating capacity of the three types of resin for water with a 1 / 1 Cl^-/SO_4^{2-} ratio with coflow regeneration. Figure 30 indicates the regenerability of the resins.

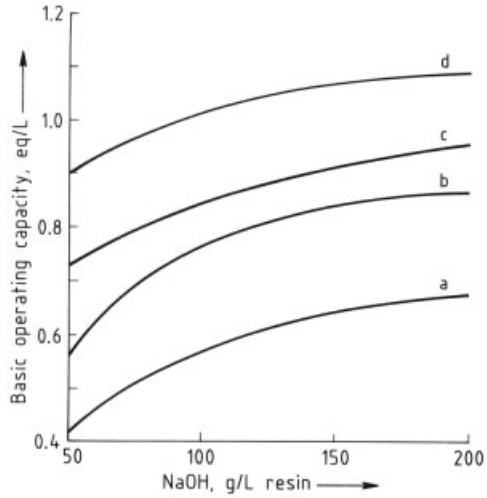

Figure 29. Basic operating capacity of strongly basic anion exchangers as a function of the amount of regenerant
a) Polystyrene resin type 1; b) Polyacrylic resin; c) Polystyrene resin type 2; d) Dual base acrylic resins

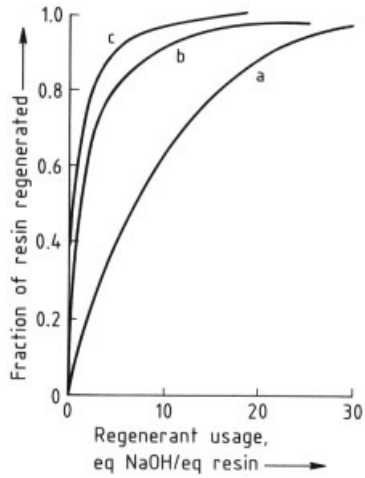

Figure 30. Regenerability of strongly basic anion exchangers
a) Polystyrene resin type 1; b) Polyacrylic resin; c) Polystyrene resin type 2

One or Two Anion Resins? Most water to be demineralized contains salts of both strong and weak acids, the former frequently in high proportion. The economic advantage of using weakly and strongly basic anion-exchange resins in series, due to considerable savings in caustic regenerant, is demonstrated in Section 9.3.

For small and medium-sized plants (up to ca. 40 m^3/h), capital cost considerations are generally predominant, so that a series of the type SAC – DG – SBA is often selected. Because of its relatively good regeneration efficiency, a type 2 polystyrene resin is a good candidate for the anion exchanger.

Dual base acrylic resins (see Section 7.4) offer an even higher operating capacity (0.8 to >1 eq/L) with low regenerant consumption if the influent water contains less than 5 % silica. The regenerant usage is 140 % of the capacity in reverse flow regeneration, and 150 – 190 % in coflow, depending on water composition.

Layered-bed systems are used for bigger plants. In very large installations, however (>100 m^3/h), hydraulic considerations often call for separate weak base and strongly basic resin columns. The strongly basic resin is generally a type 1 or an acrylic resin with a strong base functionality.

10. Plant Design

For a detailed discussion, see [66].

10.1. General Considerations

The size of the water treatment plant, number of columns, choice of resins, and type and amount of regenerant are largely determined by

1) The rate at which treated water is required
2) The specified quality of treated water
3) The capital cost of the plant
4) The running costs of the plant

Small, mass-produced ion-exchange plants generally use large amounts of regenerant to maximize water quality and operating capacity. Running costs are high, but the user benefits from low capital cost and operational simplicity.

Where very large volumes of water are required (e.g., plants for municipal softening and nitrate removal or in heavy industry), emphasis is placed on the reduction of running costs. A very high degree of removal is not required, and regenerant levels are kept as low as possible to raise the regeneration efficiency. Thus, large volumes of resin are required, which increases capital costs.

When treated water of ultrahigh quality is required in large quantities as feedwater for high-pressure boilers, high regenerant levels can be used, which result in high operating capacities, compact plants, and high running costs. However, reverse flow regenerated systems are now universally employed which have a lower regenerant dosage and hence improved regeneration efficiency. In water demineralization, the combined use of weak and strong electrolyte resins further improves the chemical efficiency of the system. A complete treatment cycle consists of four phases (Section 8.1):

1) Service, i.e., the phase during which the resin is exhausted
2) Decompaction of the bed by backwash
3) Regeneration
4) Rinsing

In units regenerated by reverse flow, decompaction is only carried out occasionally so as not to disturb the resin layers and lose the advantages associated with counterflow operation. Some systems provide for cleaning the upper layer each cycle by backwashing only the top 15 cm of the bed.

10.2. Fixed-Bed Ion-Exchange Units

10.2.1. Column Diameter

Column diameters vary from a few centimeters (in laboratory units) to 5 m. Industrial units usually have diameters of 0.8 – 3.6 m. The depth of the resin bed also varies from 10 cm to about 3.5 m. For standard industrial units, bed depths should generally not be less than 70 cm. The maximum depth depends on the head loss that can be tolerated. Most units have resin beds from 1 to 2 m in depth. The main problem lies in producing even flow across the bed, because the liquid to be treated and the regenerant must be in

uniform contact with the whole volume of resin to allow stable horizontal wavefronts to develop and move vertically through the beds.

10.2.2. Small-Scale Units

Figure 31 shows a typical column designed for domestic water softening with a diameter up to about 50 cm. The outer cylinder is constructed from polyester reinforced with fiber glass. The water and regenerant inlet and the backwash outlet are combined in a single large strainer formed by a plastic cylinder with slits ca. 0.25 mm wide through which resin beads cannot pass. The treated water outlet is formed by a pipe concentric with the input strainer and fitted with a polyamide screen at the bottom. This type of unit is fully automatic; a multiport valve is fitted directly on top of the cylinder so as to form a compact arrangement with the system of inlet and outlet pipes.

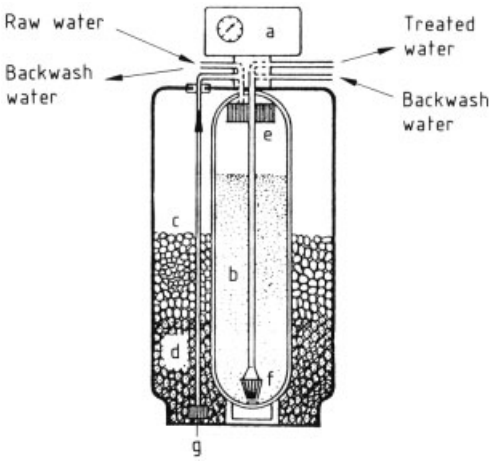

Figure 31. Small-scale water softener
a) Control unit; b) Resin; c) Solid salt; d) Saturated brine

Small counterflow units, such as those normally used for domestic softeners, can be extremely simple, consisting of a column similar to that shown in Figure 31 but completely filled with resin. Although backwashing is not possible, the units operate satisfactorily because any dirt is filtered out at the top of the resin bed and washed out during reverse-flow regeneration.

10.2.3. Industrial Co- and Counterflow Plants

Industrial demineralization units are typically rubber-lined and fabricated from steel. A rubber layer about 3 mm thick is needed to protect the steel against acid corrosion. Various distribution and recycling systems are used to ensure even flow across the resin. Four examples are illustrated in Figure 32.

In coflow regenerated units, a regenerant distributor should be located just above the resin bed, as shown in Figure 32 A. In some simple columns, the regenerant is introduced at the top of the column. Regeneration is less efficient, as the regenerant is diluted before it reaches the resin. Because the columns are constantly filled with water during the whole cycle, space above the resin bed must be sufficient to allow for expansion of the bed during the backwash phase. When introduced at the same level as the upper surface, dilution is negligible. Uniformity of flow is controlled by the collection system at the bottom of the column rather than by the inlet.

In Figure 32 A, the collection system is buried in successive layers of gravel, or pieces of crushed flint (h, i) or anthracite from 2 to 10 mm in size; the coarser layers are at the bottom and the finer at the top. The collection system consists of a main header pipe onto which a set of perforated lateral pipes is welded or screwed, the whole system being supported by a base plate. The space between this plate and the convex bottom of the column is filled with concrete (m) for strength and rigidity. This long-established system is still used for softeners or simple demineralizers and gives an extremely uniform flow.

Figure 32 B shows a unit with counterflow (upflow) regeneration. The treated water collector (j) has lateral pipes fitted with small plastic filtering nozzles or strainers (two types of strainer are illustrated in Fig. 33). Slits in the strainers allow water to pass but retain the resins. This system may cause problems when water of the highest purity is required. To obtain uniform collection, the lateral pipes must be spaced at intervals of less than 15 cm and form a network at the bottom of the column. This network can retain traces of regenerant that subsequently leak into the treated water.

In large industrial units, the particles filtered out of the resin bed must be removed with a

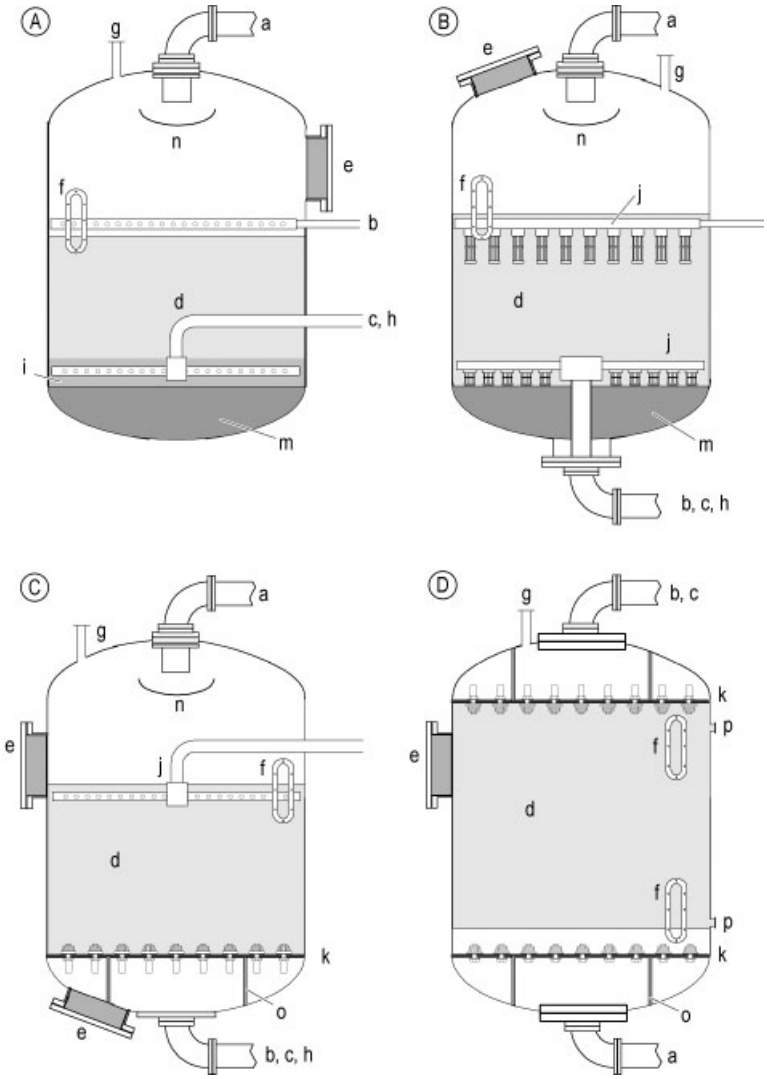

Figure 32. A) Co-flow-regenerated unit; B) Counterflow regenerated unit with "plunging" strainers; C) Counterflow-regenerated unit with nozzle plate; D) Floating bed unit with two nozzle plates
a) Inlet for raw water and outlet for backwash water; b) Regenerant inlet; c) Outlet for treated water; d) Resin; e) Manhole; f) Sight glass; g) Breather or vent; h) Inlet for backwash water; i) Support layer; j) Collector system with strainers; k) Nozzle plate; m) Concrete; n) Splash plate; o) Pillar support; p) Resin transfer port

backwash. Many systems are available to keep the bed compact during service and regeneration while allowing it to expand during an occasional backwash. A broad distinction can be made between designs with a large freeboard, in which a backwash can be carried out in the ion-exchange vessel itself, and packed-bed designs, in which the ion-exchange vessel is fully packed with resin, without a freeboard.

Buried collector systems, as shown in Figures 32 B and 32 C, are widely used in designs that have a freeboard. Figure 34 shows how they are operated. During the exhaustion phase (A), the solution to be treated flows from the top to the bottom of the unit; regeneration is carried out by introducing the regenerant at the bottom and extracting it through the collector (C); the resin bed is maintained in a compact state by a

current of water from the top of the column or by compressed air which leaves with the spent regenerant through the collector. This type of construction enables the upper resin layer to be backwashed from the collector without disturbing the lower layers (B).

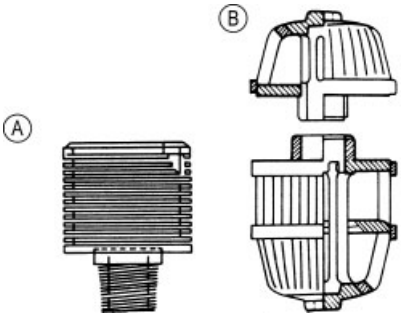

Figure 33. Examples of strainers used for industrial water-softening plants
A) Simple plate strainer (Johnson); B) Double strainer for two-chamber floating beds (Kleemeier, Schewe & Co, Herford, Germany)

In a variation of this process, the buried collector is replaced by a number of downpipes with strainers at their ends (see Figure 32 B); these long, candle-shaped tubes penetrate the resin. This system restricts the water flow less and gives a more even collection. In both systems, the layer of resin above the collector does not take part in the ion-exchange process because it is not regenerated. Air-holddown systems use an inert polymer instead of this inactive resin layer, to protect the ion-exchange resin and prevent its drying out during regeneration.

A very frequent and efficient collection method for water treatment is shown in Figures 32 C and D. It consists of a flat base plate, perforated at 15 cm intervals, and fitted with screwed-on strainers leading into a lower chamber. This method of construction is, however, relatively expensive. Since the base plate is not supported by concrete, it must be either very thick or supported by small metal pillars to withstand operating pressure. Moreover, the space underneath must be lined with vulcanized rubber and be large enough to accommodate a worker to carry out this operation. Another method is to fit the cylinder with a coupling flange, which adds to the cost.

Packed-Bed Column. A packed bed column is shown in Figure 32 D: the resin is kept between two nozzle plates (k), and there is almost no freeboard. The particular design here is a floating bed unit (Amberpack) with upflow loading.

For the treatment of valuable solutions (see Chap. 12), special units with a flat bottom and a precisely engineered collection system are frequently used to minimize the void volume and

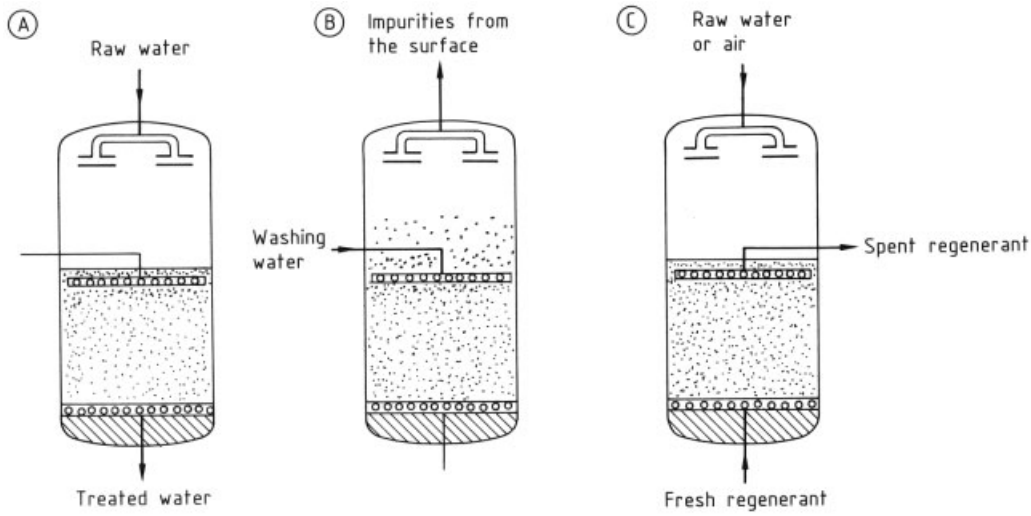

Figure 34. Buried collector system for counterflow regeneration
A) Service (exhaustion) phase; B) Surface backwash phase; C) Regeneration phase

the risk of regenerant "hide-out", i.e., a dead zone from which spent regenerant is difficult to wash out during rinsing steps. Hide-out is usually due to poor hydraulic distribution (channeling) or to an imperfect collecting system.

Stratified Bed Units. In upflow regenerated columns using the air- or water-holddown technology, a combination of weakly and strongly active resins can be used in theform of a layered or stratified bed (Figure 35). Selected resin grades with a suitable particle size and density are required to prevent mixing of the layers.

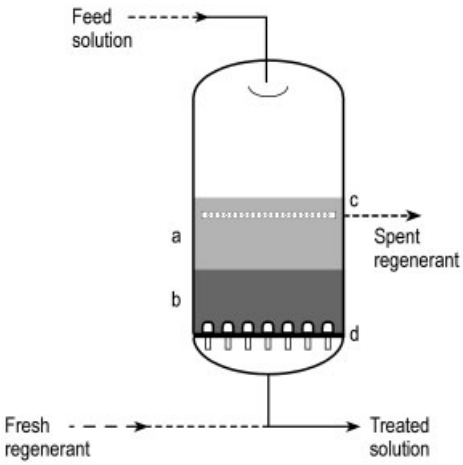

Figure 35. Stratified-bed unit
a) Weakly acidic or weakly basic resin; b) Strongly acidic or strongly basic resin; c) Regenerant collector; d) Nozzle plate

Split-Flow System [32]. As shown in Figure 36, the split-flow system has a regenerant collector (b) that is buried below the bed surface at about one-third of the resin bed depth. The regenerant distributor (a) is located just above the bed surface. The regenerant is introduced simultaneously through the distributor and at the bottom of the column. The top stream keeps the bed compacted and regenerates the most exhausted part of the resin, whereas the bottom stream ensures that the lowest layers are always completely regenerated. This system also allows the upper part of the bed to be backwashed without disturbing the lower layers.

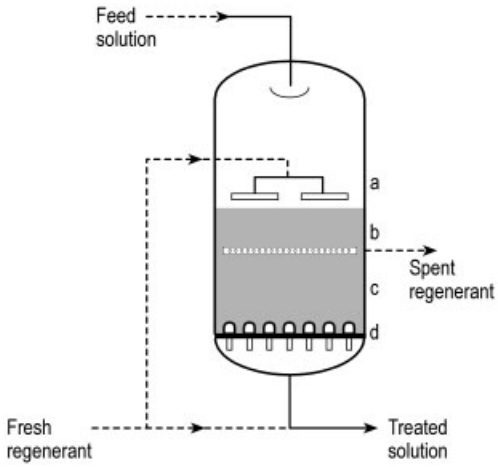

Figure 36. Split-flow system
a) Regenerant distributor; b) Regenerant collector; c) Resin; d) Nozzle plate

Econex System. In the Econex system of Davy Bamag (Fig. 37) [31], the upper part of the column is filled with inert material (b), which is used for compaction and can be removed to decompact the bed.

Floating-Bed Unit. The floating bed, invented and patented by Bayer in 1963 [27], is a reverse counterflow system (Fig. 38). During the service phase, the flow of water is from bottom to top, whereas the column is regenerated from top to bottom. The unit consists of a vessel fitted with an upper and a lower strainer rack. In practice, the vessel is filled almost completely with resin to prevent its becoming too fluid. A layer of light, inert, unbreakable resin granules may be used to prevent the fine particles or pieces of active resin from clogging the upper strainers. The unit is compact and has been quite successful, but has some limitations: it must operate at a minimum flow rate to prevent fluidization of the bed, and it is sensitive to the presence of suspended solids in the feed.

A variation of the floating bed has two chambers that are separated by a floor fitted with double strainers (d) and allows two resins to be placed in a single column (Fig. 39). The resins are regenerated in the direction from SAC to WAC or from SBA to WBA (see Section 9.3).

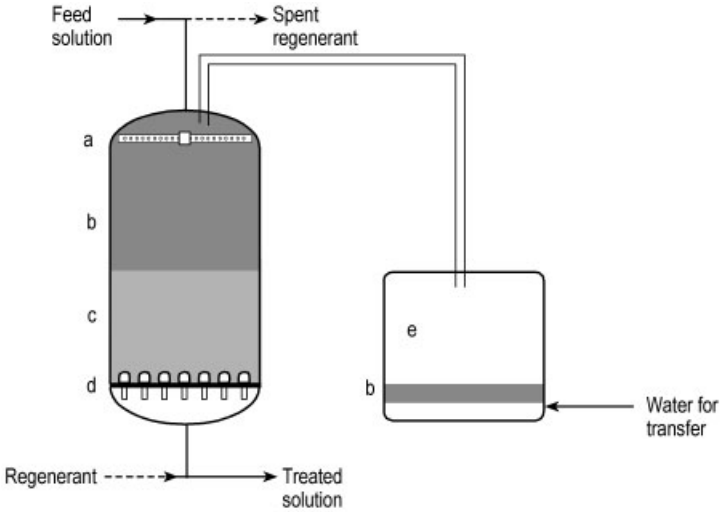

Figure 37. Econex system
a) Water distributor and regenerant collector; b) Coarse inert resin for bed compaction; c) Ion-exchange resin; d) Nozzle plate; e) Tank for inert resin during backwash

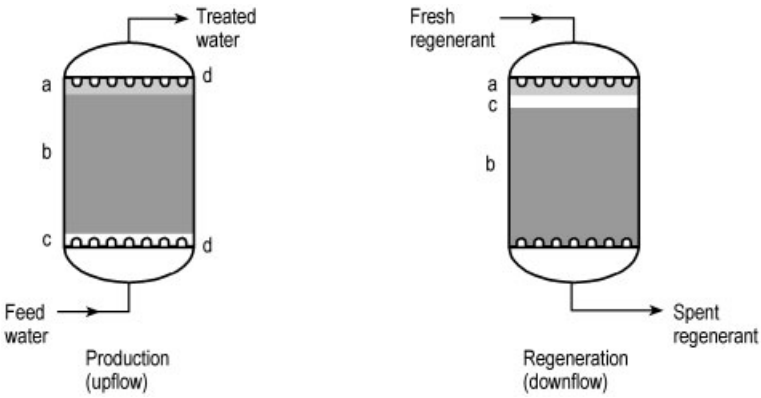

Figure 38. Floating-bed unit
a) Inert floating resin; b) Ion-exchange resin; c) Void space; d) Nozzle plate

Amberpack System. The Amberpack system of Rohm and Haas [28] is an improved floating-bed unit. It has one or two small ancillary backwash columns into which portions of the resin beds can be transferred for the removal of fines or suspended solids, as shown in Figure 40.

UFD System. In the UFD system of Degrémont (Fig. 41) [29], the resin fills the column completely (as in the floating bed), but percolation occurs in the conventional direction (downflow), and regeneration in the floating-bed mode (upflow).

UpCoRe System. This system, proposed by Esmil [30] and licensed to Dow Chemical under the name UpCoRe, is very similar to the UFD system. It is less sensitive to the presence of suspended solids than the floating-bed system. However, regeneration of the bed at a high flow rate makes it slightly less efficient.

Recoflo System. The Recoflo system is a shallow-bed process offered by Eco-Tec (Fig. 42) [33], using fine mesh resins (particle size 75 – 150 μm) between two distributing and collecting plates. The bed depth is usually 15 cm,

and the run length is typically only 5 – 30 min. Regeneration time is correspondingly short.

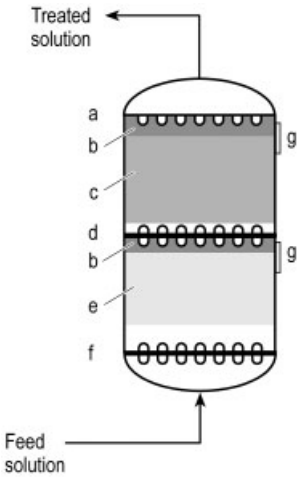

Figure 39. Double-compartment floating bed
a) Outlet nozzle plate; b) Inert resin for protection of nozzles; c) Strongly acidic or strongly basic resin; d) Separation plate with double nozzles; e) Weakly acidic or weakly basic resin; f) Inlet nozzle plate with double nozzles; g) Sight glass

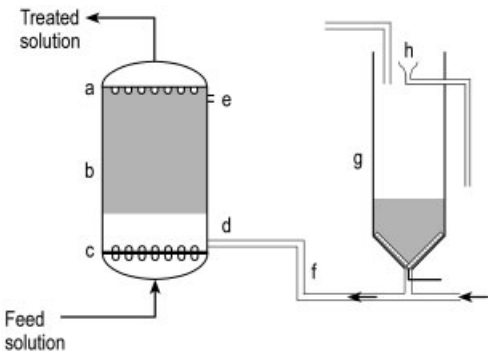

Figure 40. Amberpack system
a) Nozzle plate; b) Resin; c) Nozzle plate with double nozzles; d) Lower transfer port; e) Upper transfer port; f) Flexible transfer hose; g) Backwash column; h) Backwash collector

10.2.4. Mixed Beds

General Case. The mixed-bed principle was described in Section 8.2. Successful operation depends on the fact that the resins are separated by backwashing and settling before each component is regenerated. The densities of the resins normally used are 1.07 – 1.10 for the strongly basic resin and 1.20 – 1.30 for the strongly acidic resin. These are sufficiently different for the resin to be well-separated by backwashing, provided the cation exchanger does not contain too many fine particles or the anion exchanger too many large beads.

A typical mixed-bed design is shown in Figure 43. Untreated water enters at the top of the column through a strainer that also acts as the backwash outlet and prevents loss of resin. A sodium hydroxide distributor (c) is placed just above the bed top, and an intermediate central collector (d) is located at the interface between the two resins.

The operational cycle is illustrated in Figure 44 and consists of the following stages:

1) Service (Fig. 44 A)
2) Backwash for bed separation and subsequent settling (B)
3) Regeneration of the anion exchanger (C), followed by a rinse (the spent regenerant and rinse water leave via the central collector)
4) Regeneration of the cation exchanger with acid introduced through the bottom of the column and extracted through the central collector (D) followed by a rinse
5) Re-mixing of the resins with compressed air (E), followed by a final rinse

A small counterflow of water is advisable during the two regeneration stages to restrict the diffusion of acid into the anion exchanger and of sodium hydroxide into the cation exchanger. The two resin beds can also be regenerated simultaneously, with both spent regenerants leaving through the central collector. Although this system is very efficient, problems may arise due to precipitation of calcium carbonate or ferric hydroxide when the acid and the alkaline effluents meet. However, this situation is encountered mainly in general-purpose mixed beds and only rarely in polishing systems.

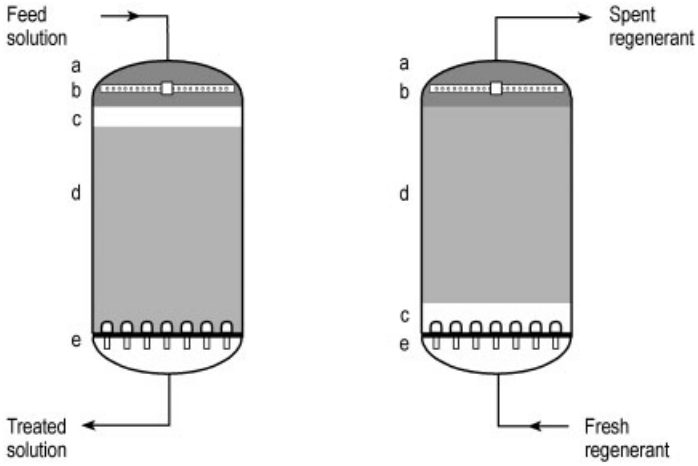

Figure 41. UFD system
a) Inert resin for protection of collector; b) Water distributor and regenerant collector; c) Void space; d) Resin; e) Nozzle plate with counterflow valves

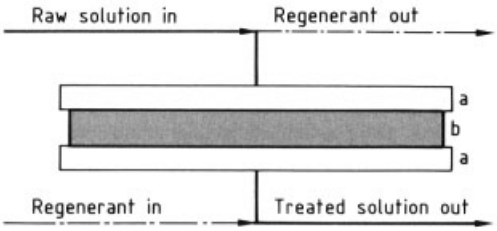

Figure 42. Recoflo system
a) Distributor – collector; b) Bed of fine resin, depth ca. 15 cm

Three-Component Mixed Beds. Three-component mixed beds, known under the trade names Triobed (Duolite) or Ambersep (Rohm and Haas), overcome the problem of mixing of resins at the interface [34]. A layer of inert polymer beads, with a density intermediate between those of the cation- and anion-exchange resins, is mixed with both. Backwashing causes the mixture to split into three distinct layers, giving complete separation between the two ion exchangers (Fig. 45). This prevents cross-contamination, improves regeneration efficiency, and considerably reduces ionic leakage [35], [36].

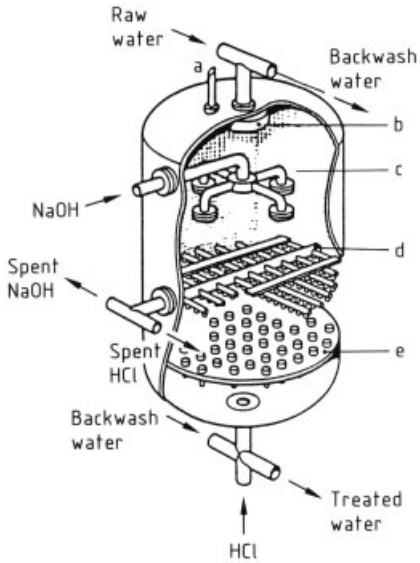

Figure 43. Mixed-bed ion-exchange system
a) Breather; b) Raw water distributor; c) Sodium hydroxide distributor; d) Intermediate collector; e) Strainer rack

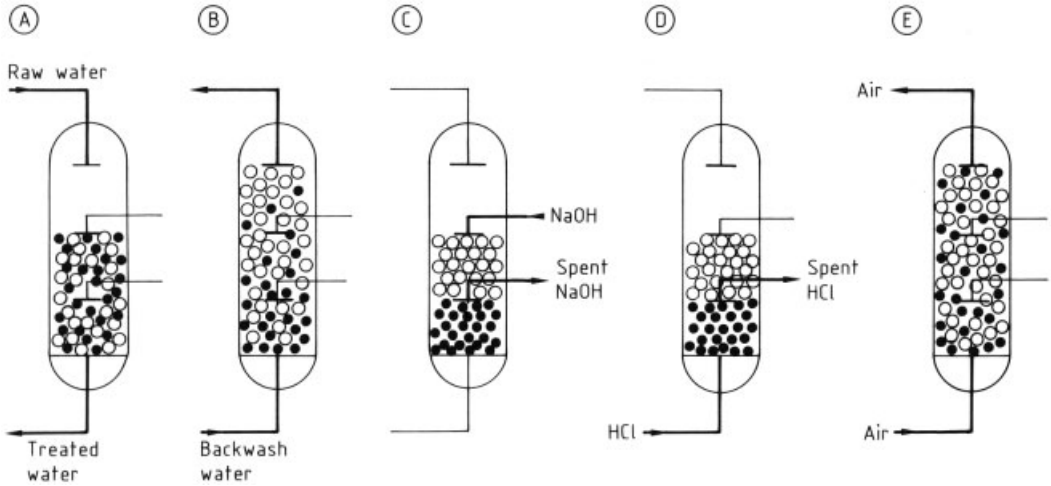

Figure 44. Operational phases of a mixed-bed cycle
A) Service; B) Separation; C) Regeneration of anion exchanger; D) Regeneration of cation exchanger; E) Remixing with air
○ Anion-exchange resin; ● Cation-exchange resin

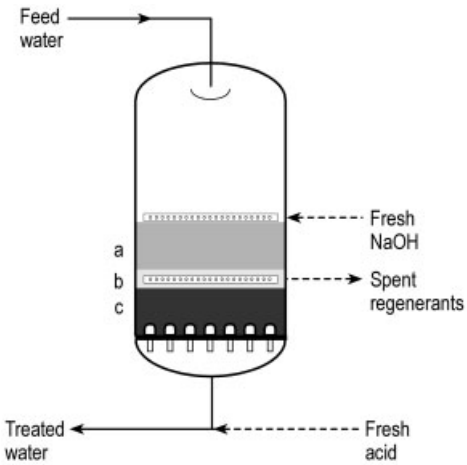

Figure 45. Triobed process
a) Anion-exchange resin; b) Inert resin; c) Cation-exchange resin

However, the inert beads, being less hydrophilic than ion exchange resins, can cause operational problems. The availabilty of resins with uniform particle size and sophisticated regeneration technologies have gradually reduced the popularity of the Triobed concept.

External Regeneration. In the treatment of condensates or in polishing that follows a demineralizing system, the water to be treated by the mixed bed has a very low salinity. As a result, only a small exchange capacity is required and the units are therefore made as small as possible. They operate at very high specific flow rates of up to 130 BV/h. Any internal obstruction in the column can cause high pressure; therefore, these units are constructed as simply as possible (one rack of strainers, bed depths of ca. 1 m to avoid excessive head loss). To prevent accidental inlet of regenerant chemicals into the condensate circuit, the resins are transferred to separate regeneration units that operate at much lower flow rates. A single regeneration station can serve a series of mixed-bed units, as shown in Figure 46.

It may consist of several separate columns, specially designed to produce maximum separation of the resins and the best possible regeneration efficiency:

1) A separation column, also used for regenerating the cation exchanger
2) A column to which the anion exchanger is transferred for regeneration
3) A mixing and waiting column

An example of a mixed-bed unit as used for condensate polishing is shown in Figure 47.

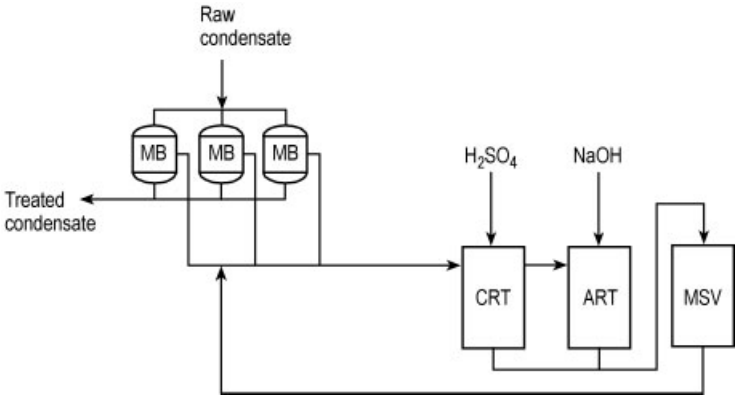

Figure 46. Mixed-bed external regeneration station
MB: Mixed-bed unit; CRT: Cation-regeneration tower; ART: Anion-regeneration tower; MSV: Mixing and storage vessel

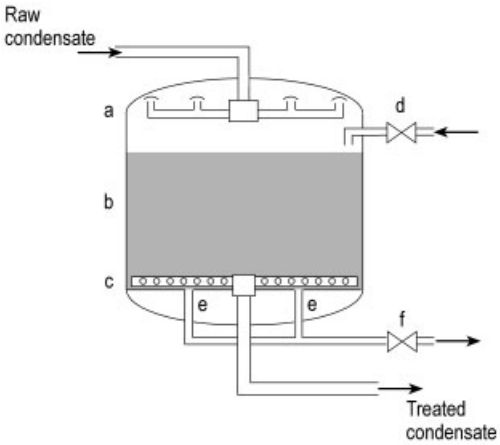

Figure 47. Condensate polishing unit
a) Condensate distributor; b) Mixed resin; c) Condensate collector; d) Clean resin return; e) Resin transfer ports; f) Dirty resin transfer

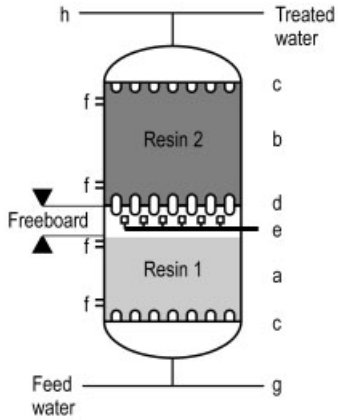

Figure 48. Amberpack Sandwich polisher
a) Strongly acidic resin; b) Strongly basic resin; c) Nozzle plate; d) Separation plate with double nozzles; e) Regenerant inlet 1 and regenerant outlet 2; f) Transfer port; g) Regenerant outlet 1; h) Regenerant inlet 2

10.2.5. Other Ion-Exchange Polishers

In the *Amberpack Sandwich polisher* of Rohm and Haas (Figure 48), a strongly acidic cation-exchange resin is combined with a strongly basic anion-exchange resin in two separate compartments of a column. The two compartments communicate via a plate fitted with double nozzles. A collecting device located just below the separation plate serves as a regenerant collector for the upper compartment and a regenerant distributor for the lower compartment.

In the *Tripol* process of Permutit-Boby [55], [68], three resins, (SAC1/SBA/SAC2) are placed in three shallow compartments of a single column. The second cation-exchange resin compartment, which is always very highly regenerated, acts as a final sodium trap.

The *Multistep* process of Bayer [56] is essentially the same as the Tripol process but operates in the opposite direction, in the floating-bed mode.

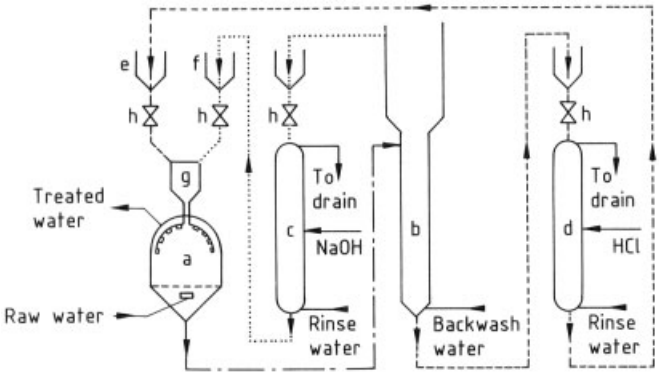

Figure 49. Mixed-bed demineralization (Asahi process)
a) Service column; b) Backwash and separation column; c) Anion exchanger regeneration column; d) Cation exchanger regeneration column; e) Regenerated cation-exchange resin; f) Regenerated anion-exchange resin; g) Mixing chamber; h) Valve
—— water; ——·—— mixed resin; – – – cation exchanger; · · · · anion exchanger

10.3. Continuously Circulated Ion-Exchange Resins [37]

As discussed in Section 8.2, automatic units in which the resin circulates continuously or intermittently should offer the following advantages:

- Smaller floor areas
- Smaller resin volume
- Continuous flow of regeneration liquids

In all resin-circulating units the resin flows counter to the solution to be treated. The exhaustion, regeneration, and rinsing stages occur simultaneously in different columns or in different sections of the same column.

The difficulties associated with circulation of the resin, its more rapid exhaustion than in fixed-bed units, and the complex construction of continuous systems have restricted their successful application. The counterflow processes described in Section 10.2.2 are much more widely used at present. However, mention should be made of some continuous processes that are still in operation:

1) The *Asahi process* [38], invented in Japan and finally developed in France by Degrémont (Fig. 49);
2) The *Kontimat process* [39] developed by Hager und Elsässer; and
3) The *Higgins process* of Chem-Seps [40].

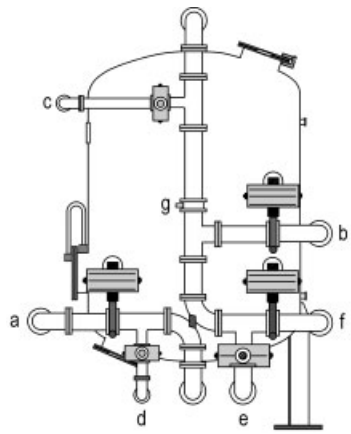

Figure 50. External pipework and valves
a) Raw water inlet; b) Treated water outlet; c) Regenerant inlet; d) Regenerant outlet; e) Recycle outlet; f) Fast rinse outlet; g) Sample point for quality analysis

10.4. External Valves and Pipework

A typical arrangement of external valves and pipework is shown in Figure 50. In the case of automatic softeners, separate valves can be replaced by a single multiport valve actuated by means of an electronic scheduler. For large systems, separate pneumatically or hydraulically controlled valves are used.

10.5. Control Systems

The quality of water produced by a plant must be monitored. As soon as it falls below a previously specified level, the service stage is stopped and the regeneration stage initiated. This section briefly describes the measuring instruments used in monitoring.

Timers. For domestic water softeners, daily consumption is assumed to be constant, and the softener is scaled to require regeneration at fixed intervals (e.g., once every three days). A timer initiates the process automatically, and regeneration frequency can be varied according to changing requirements. The method is simple and reliable, but wasteful, because regeneration is always initiated before the softener is exhausted.

Quantity Measurement. In communal buildings or laundries, where a cheaper and more flexible method is required, a contacting-head flow meter can be installed which automatically initiates regeneration as soon as a given volume of water has passed. This method is accurate as long as the hardness of the water supply does not vary.

Automatic Hardness Detector. A colorimetric hardness analyzer can be installed at the output of the softener to initiate regeneration when the allowed level of leakage is reached. This avoids waste of regenerant, but the analyzer is relatively expensive and the method is used chiefly in large industrial installations.

Conductivity Meters. Demineralizers are normally controlled by measuring electrical conductivity. Simple, reliable instruments can be installed to initiate regeneration when the conductivity of the treated water rises above a preset alarm value (e.g., 0.25 μS/cm for mixed-bed units or 10 μS/cm for two-bed units).

The breakthrough of silica, which has no detectable conductivity, can cause a problem. A common, reasonably reliable way of dealing with this is to place a conductivity probe above the bottom of the bed. The position of the probe is calculated in such a way that the wavefront of chlorides moving down the bed reaches the probe before the silica wavefront reaches the bottom. However, this method is not reliable enough to be used for high-pressure boiler water, in which the absence of silica is vital.

Silicometers are automatic colorimetric analyzers that operate like hardness detectors. They normally analyze water taken from a sampling point located above the bottom to ensure that no silica leaves the unit. They require very careful maintenance.

Differential Conductivity. For monitoring hydrogen-ion exchangers, a small separate cation-exchange column is set up and maintained in a highly regenerated condition by passing a large excess of acid through it at each regeneration of the main unit. During the service stage, a continuous sample of water from the main unit is passed through the test column and the difference in conductivity is monitored continuously. The difference is small and can stay constant or decrease during the run. A sudden increase marks the end of the cycle, indicating breakthrough from the main unit (see Section 6.3).

pH Measurement [67]. Although pH is easily measured, it is not a very useful parameter for plant control because unbuffered demineralized water tends to give variable values. Furthermore, the change in pH at the end of the cycle is less distinct and more subject to random changes than the variation in conductivity.

Sodium Electrodes. For cation exchangers, sodium-sensitive electrodes afford an extremely sensitive and accurate method of control.

11. Special Processes in Water Treatment

This chapter deals with special ion-exchange processes used in water treatment applications.

11.1. Removal of Organic Matter

In certain water with low salt concentrations, the fouling factor N (see Section 7.3) may become very high. To protect the anion-exchange resins in a demineralization plant, a special resin column can be installed at the head of the system to trap the organic matter. These resins are known as *organic scavengers.* Two kinds are available:

1) Highly porous strongly basic polystyrene or polyacrylic resins used in the chloride form for neutral or alkaline water, and regenerated with a mixture of sodium chloride and sodium hydroxide (e.g., Amberlite IRA900, Amberlite IRA958)
2) Weakly basic phenol – formaldehyde resin (e.g., Duolite A7) for water with high sulfate concentration, regenerated with sodium hydroxide (this resin is also capable of retaining certain detergents).

These scavenger resins have a limited efficiency, and their popularity decreased after the development of acrylic anion resins, which have excellent resistance to organic fouling. Furthermore, ultrafiltration is often superior for removing organic matter from water.

11.2. Treatment of Potable Water

Ion-exchange resins do not transform chemically or bacteriologically contaminated water into drinking water. Other processes are necessary, such as coagulation, filtration, chlorination, or ozonization. However, ion exchange can be used to reduce or remove excess hardness, alkalinity, nitrate or sulfate content, heavy metals, or radioactive substances from water destined for human or animal consumption. Dealkalization with carboxylic resins is frequently used in northern European breweries to prepare water that will make good beer. Many private houses, apartment buildings, or small communities have water softeners for domestic use.

Most ion-exchange resins, although totally insoluble, release minute amounts of residual organic raw materials into the treated water, particularly when they are new [41]. These impurities may not be allowed if the treated water (or solution in food-processing applications) is to be consumed by humans. Several countries now have regulations that must be observed by suppliers of potable water treatment equipment, including ion-exchange resins. In the United States [42], France [43], and Germany [44], these regulations specify the amount of impurities that may be released by the resin into the water. The composition of authorized resins is also described.

Removal of Nitrates [45]. In Europe and the United States, a trend toward higher nitrate concentration has been observed in both surface and groundwater. Nitrate levels higher than 50 mg/L endanger children's health. Any strongly basic resin in chloride form can remove nitrate from water. Unfortunately, the efficiency of standard strongly basic resins is only moderate because these resins also absorb sulfate and some bicarbonate, both of which are present in higher concentration than nitrate. This represents a waste of capacity and regenerant, and causes large variations in the quality of the treated water with regard to its sulfate (almost completely removed), bicarbonate (reduced), and chloride (increased) concentrations. These disadvantages can be partly alleviated by regenerating the resin with a mixture of sodium chloride, sulfate, and bicarbonate to reduce the variations in water quality [46].

Resins with a higher affinity for nitrates than for sulfates were developed in 1985 (Imac HP 555). They offer an advantage over conventional strongly basic resins when the water to be purified contains a high proportion of sulfate.

Carix Process. This was developed in Germany by the Kernforschungszentrum (Nuclear Research Center) in Karlsruhe [47]. It is unique in that it allows partial demineralization of water using carbonic acid ($CO_2 + H_2O$) as the sole regenerant. The salinity of the environment is not increased as is the case when a mineral acid and sodium hydroxide are used for regeneration. Moreover, in a mixed bed, the cation and anion resins do not require separation for regeneration, and the carbonic acid produced during service can be recovered and pressurized for use in the next regeneration sequence.

Two special acrylic resins are the key to the process: a weakly acidic cation exchanger (Amberlite IRC50) and a strongly basic anion exchanger (Amberlite IRA458). At the beginning of the service run, the cation resin is in its carboxylic acid form, whereas the anion resin is in the bicarbonate form. The principle of the Carix process can be described by the following reactions:

$$2R_A^+HCO_3^- + Ca^{2+}SO_4^{2-} \rightleftharpoons (R_A^+)_2SO_4^{2-} + Ca^{2+}(HCO_3^-)_2$$

$$2R_CCOO^-H^+ + Ca^{2+}(HCO_3^-)_2 \rightleftharpoons (R_CCOO^-)_2Ca^{2+} + 2CO_2 + 2H_2O$$

The combined mechanism of both resins produces the demineralization effect. The weakly acidic resin can exchange cations only if the water is alkaline. The alkalinity is produced by the anion resin, which is converted from the bicarbonate to the sulfate, nitrate, or chloride form.

Regeneration is carried out by a mechanism in exactly the opposite order. High carbon dioxide pressure displaces the second equation to the left; bicarbonate salts are formed in relatively high concentration and regenerate the anion resin.

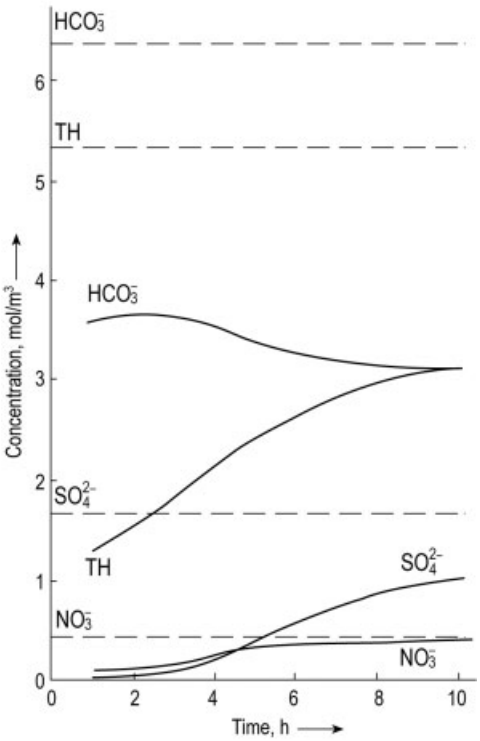

Figure 51. Carix process – profile of treated water in a partial demineralization test aimed at softening water
Cation : anion resin volume ratio = 2 : 1. Dashed lines indicate concentrations in raw water. Solid lines indicate concentrations in product water

Being reversible, the reactions are never complete, but the Carix process results in a 40 – 50 % reduction of raw water salinity. The system is particularly effective for reducing the concentration of hardness, bicarbonate, sulfate, and nitrate. Typical breakthrough curves are shown as a function of time in Figure 51. A schematic diagram of the first full-scale Carix installation is illustrated in Figure 52. The Carix process is not limited to the treatment of potable water. It is very economical in terms of regenerant usage and very environment-friendly. However, it brings about only a partial demineralization of water.

Disinfection of Water and Resins. Normal ion-exchange resins do not sterilize water. Moreover, they may promote microbial growth during shutdown periods. If a solution contaminated with microorganisms flows steadily through a resin bed, most of the insoluble contaminants are filtered out on the resin surface. However, as soon as flow stops, the immobilized silt provides an ideal feedstock for incoming bacteria. The bacteria proliferate, contaminating the treated solution when service is resumed. Both the solution and the resin bed must therefore be disinfected.

Influent water can normally be disinfected with chlorine or ozone. Unfortunately, this may often not be sufficient, because these oxidants are almost totally consumed by the silt and debris deposited on the bed surface and by the resin itself. The *resin bed* must therefore also be disinfected. This can be done at regular intervals by using chemical bactericides such as chlorine, quaternary ammonium salts, or peracetic acid. Resin manufacturers provide some guidance in this respect.

Water softeners are frequently disinfected with chlorine produced by electrolysis of the brine used for regeneration. This is common practice in Germany, where all domestic softeners are equipped with an appropriate device. Another method involves mixing a small percentage (usually 1 – 3 %) of a special resin containing precipitated silver with the actual softening resin. The precious metal diffuses within the resin bed during standby period and prevents the growth of bacteria. Unfortunately, some strains become progressively resistant to silver.

Triiodide resins, which are strongly basic anion exchangers loaded with a large amount of the triiodide ion, can disinfect treated water. They are used in India and other tropical countries.

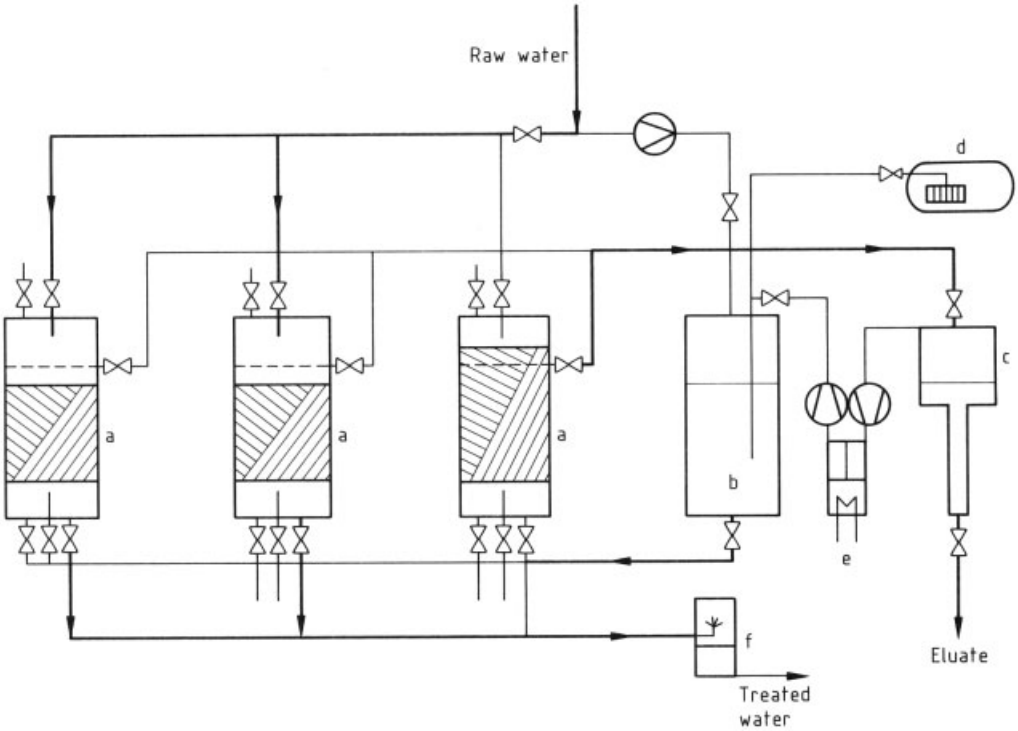

Figure 52. Simplified flow sheet of the large-scale Carix plant in Bad Rappenau (Germany)
a) Filter in operation – regeneration; b) Reaction tank; c) Vacuum degasifier; d) CO_2 storage tank; e) CO_2 vacuum – compressor unit; f) CO_2 degasifier
The first two columns are shown in production and the third in regeneration.

11.3. Treatment of Brackish Water

For many years, ion-exchange systems, such as those with SAC – WBA pairs (see Chap. 9), were used to demineralize brackish water containing up to 2500 mg (ca. 50 meq) of salt per liter. However, reverse osmosis has almost completely replaced ion exchange, which is now limited for economic reasons to water containing < 20 meq/L.

The Desal Process. Developed by Rohm and Haas in 1964 [48], the Desal system uses three columns: the first and third contain medium-basicity acrylic resins in the bicarbonate and hydroxide forms, respectively; the center column is filled with a carboxylic resin.

Exchange reactions for the three columns are:

$$R_A^+HCO_3^- + Na^+Cl^- \longrightarrow R_A^+Cl^- + Na^+HCO_3^-$$

$$R_C^-H^+ + Na^+HCO_3^- \longrightarrow R_C^-Na^+ + H_2O + CO_2$$

$$R_A^+OH^- + CO_2 \longrightarrow R_A^+HCO_3^-$$

During the process, the third column is converted to the bicarbonate form. Thus, after regeneration of the first column to the hydroxide form with a cheap alkali such as a calcium hydroxide suspension, and a normal regeneration of the cation unit with acid, the next cycle can proceed in the opposite direction, the third column now becoming the first.

This system has major limitations: sodium leakage from the center column is very high, so that desalination rarely exceeds 50 %. Bicarbonate is also lost from the system, unless the incoming water is strongly alkaline.

The Carbosoft Process. The softening of brackish water with a sulfonic acid resin in the sodium form is difficult if the influent contains a large proportion of sodium, due to its self-regenerating effect (see Section 7.2). Carboxylic

resins have a much better selectivity for multivalent ions than sulfonic resins. The Carbosoft process exploits this property. A weakly acidic resin in the sodium form selectively removes Ca^{2+} and Mg^{2+} ions from highly saline water with high efficiency and acceptably low leakage:

$$2\,RCOO^-Na^+ + Ca^{2+}Cl_2^- \longrightarrow (RCOO^-)_2Ca^{2+} + 2\,Na^+Cl^-$$

This system requires a two-step regeneration because the carboxylic resin cannot be regenerated with brine. The first step uses an acid:

$$(RCOO^-)_2Ca^{2+} + 2\,H^+Cl^- \longrightarrow 2\,RCOO^-H^+ + Ca^{2+}Cl_2^-$$

The resin is converted to the sodium form in a second step with dilute alkali:

$$RCOO^-H^+ + Na_2^+CO_3^{2-} \longrightarrow 2\,RCOO^-Na^+ + CO_2 + H_2O$$

Resins for the Carbosoft process must be able to withstand frequent sodium – hydrogen osmotic shock. In practice, only highly stable macroporous resins can be used. This process is widely employed to soften water injected as steam into oil wells to enhance the recovery of oil.

11.4. Processes Involving Sea Water

Sea water typically contains 35 g of sodium chloride per liter (i.e., 0.6 eq/L) and other minerals in lower concentration. This makes industrial-scale desalination by ion exchange totally impractical: only one bed volume of sea water would exhaust the resins, and more than one bed volume of treated water is needed for regenerant dilution and resin rinse. Therefore sea water desalination is performed exclusively by membrane methods (reverse osmosis and electrodialysis) or distillation. However, ion-exchange resins are well-suited to treat sea water in special cases.

Survival Kits. Many armed forces and airlines equip their soldiers and aircraft with emergency cartridges containing fine ion-exchange materials and silver salts to produce a few liters of potable water for immediate survival.

Softening of Sea Water. Sea water evaporators are subject to scaling with calcium sulfate. The Carbosoft process (Section 11.3) can be used to soften sea water.

Sea Water as a Regenerant. Because sea water contains essentially sodium chloride, it can be used to regenerate softening resins. However, the sodium chloride concentration of sea water (ca. 3.5 %) is lower than the optimum regenerant concentration (10 %), and the presence of other mineral salts is detrimental to hardness leakage and operating capacity.

11.5. Treatment of Condensates

Returned condensate from modern high-pressure boiler – turbine – condenser systems is not of suitable quality for recycling as boiler feedwater. It contains erosion and corrosion products, mainly insoluble iron oxides with traces of soluble salts of copper and other metals, together with salts arising from pinhole leaks in the condenser system which cause the entry of cooling water.

The problem of condenser leakage is particularly serious in coastal power stations where condensers are cooled by sea water. Boiler and turbine manufacturers keep up with the advances in ion exchange and demand increasingly strict standards for the quality of feedwater used in their equipment. The acceptable limit for sodium and chloride content is often expressed in parts per trillion or nanograms per liter.

As long as condenser leaks are small, raw condensates contain only very low concentrations of dissolved and suspended minerals, often a few micrograms per liter. During startup periods or with serious condenser leakage, concentrations may rise to 1 – 5 mg/L. The condensate is sometimes conditioned with 0.1 – 2 mg/L of ammonia or several milligrams of morpholine or other amines to raise the pH and reduce corrosion.

Special problems are associated with the filtration and demineralization of such condensates, and abundant literature exists on the subject [51], [52]. The processes fall into two main categories: one uses standard resins and plants similar to those used for demineralization; the

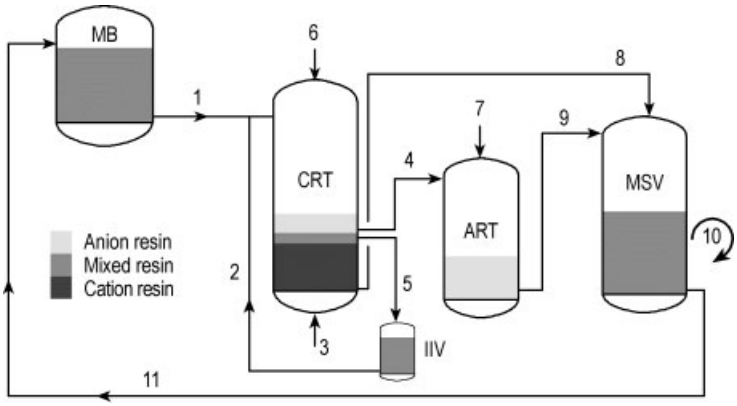

Figure 53. Regeneration with interface isolation
MB: mixed-bed unit; CRT: cation regeneration tower; ART: anion regeneration tower; MSV: mixing and storage vessel; IIV: interface isolation vessel
1) Exhausted resin transfer to regeneration station; 2) Transfer of interface resin into CRT; 3) Backwash for separation; 4) Anion resin transfer to ART; 5) Transfer of mixed interface to IIV; 6) Cation resin regeneration with acid; 7) Anion resin regeneration with caustic soda; 8) Cation resin transfer to MSV; 9) Anion resin transfer to MSV; 10) Air or nitrogen mixing of resins; 11) Return of mixed resins to mixed bed unit

other employs finely ground resins applied to precoated filters (see Section 11.5.2).

11.5.1. Conventional Resins

In the majority of plants, conventional resins in bead form are used in mixed beds (Section 10.2.4) which often operate with very high specific flow rates (100 – 130 BV/h) and very long cycles. For condensates conditioned with ammonia, the mixed bed is sometimes preceded by a cation exchanger which removes ammonia so that the mixed bed does not become saturated too quickly and can work at neutral pH where it is more efficient.

To ensure that ionic leakage is kept very low, especially at high pH,special techniques are available for the regeneration of mixed beds, mainly with a view to limiting cross-contamination between resins. Such processes include the following:

1) The *Seprex process* of Graver [53], in which the resins are separated by density differences in ca. 15 wt % sodium hydroxide solution.
2) The *Ammonex process* of Cochrane [54], in which an ammonia solution is circulated after regeneration of the anion exchanger to displace sodium from the contaminating cation resin.

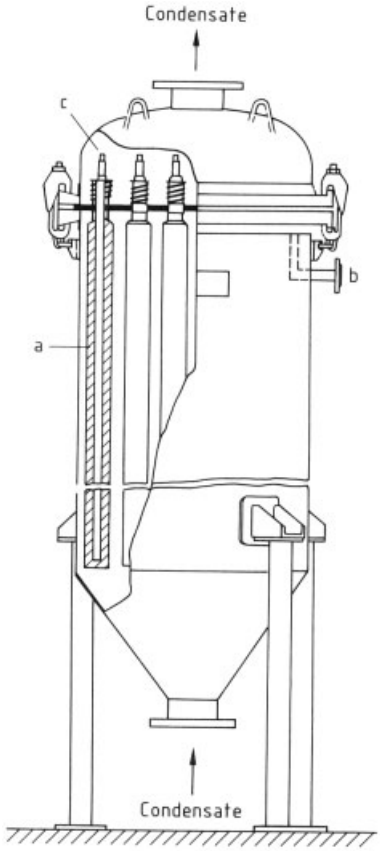

Figure 54. Candle filter
a) Filter candle covered with resin precoat; b) Emptying outlet for upper compartment; c) Candle outlet

3) The *Conesep process* of Kennicott, where separation is greatly improved with a conical column and monitored with a conductivity device.
4) The Interface Isolation technique, shown in Figure 53, makes sure that the resin that cannot be completely separated and thus fully regenerated is not returned to the mixed bed unit.

11.5.2. Powdered Resins

The process launched by Graver under the name *Powdex* [57], [69] uses completely regenerated ion-exchange resins that are ground into a powder with particle sizes between 20 and 200 µm. The cation-exchange powder in H^+ form (sometimes NH_4^+ form) and the anion-exchange powder in OH^- form are mixed with water to yield a large flaky mass which is applied to candle columns (Fig. 54) or plate columns (Fig. 55). About 1 kg dry weight of mixed exchanger is used per square meter of filter surface. Several manufacturers of powdered resins have devised ready-made mixtures that are easier to use [58].

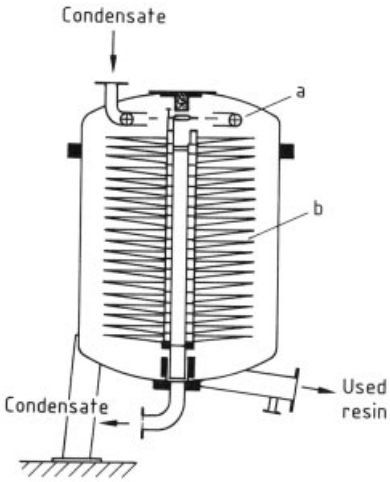

Figure 55. Plate filter
a) Distribution crown; b) Plate covered with resin cake

The advantage of powdered resins is that they filter suspended matter from the condensate very efficiently. Moreover, if radioactive substances are absorbed (e.g., in boiling water nuclear reactors, see Section 11.6), subsequent storage is easier because they are attached to a solid. The major disadvantage of powdered resins is their low demineralization capacity when condenser leakage is considerable.

Powdered resins are never regenerated. The caked layer of used resin is removed by a stream of compressed air, and a new layer is coated onto the surface. The criterion used to judge the end point is the head loss produced by compaction of the filtering layer from the effect of flow rate and the filtered matter: this head loss is normally limited to 180 kPa. In the case of condenser leakage, the end point occurs through exhaustion of the resin, whose operating exchange capacity is about 2 eq per kilogram of dry anion exchanger.

11.6. Water Treatment in the Nuclear Industry

In nuclear power plants, special problems occur because water circuits may contain radioactive elements. Water suspected of containing such products cannot be rejected to the public sewage system before being decontaminated. Similarly, spent regenerants of ion-exchange resins used in such circuits cannot be rejected either. Therefore, many nuclear circuits are treated with resins on a once-through basis: once exhausted, these resins are disposed of as nuclear waste (i.e., enclosed in cement, plastic, or glass as solid blocks; put in drums; and stored in underground nuclear dumps). Nuclear-grade resins must meet very strict purity specifications and are supplied in a highly regenerated form.

11.7. Production of Ultrapure Water [70]

In the semiconductor industry, the chips used for integrated circuits are washed with demineralized water and dried. Any mineral or organic impurity in the water may form a "bridge" between adjacent conductors which are separated by a fraction of a micrometer. This causes a short circuit and makes the part unusable. Therefore, extremely pure water must be used in this process. Typical water quality specifications are

Resistivity	18.2 MΩ · cm at 25 °C
(Conductivity	< 0.55 µS/cm at 25 °C)
TOC	< 3 µg/L
Microorganisms	< 2/mL
Suspended solids (< 1 µm)	< 10/mL

The production of such water requires the combination of several technologies, typically filtration, reverse osmosis, treatment with activated carbon, primary ion exchange, UV sterilization, microfiltration, final ion exchange, and ultrafiltration. Not only should the plant remove all impurities from the feedwater, but also none of its components must add any additional impurities to the treated water. Therefore, special materials with near-absolute insolubility must be used. The primary ion-exchange step often consists of two mixed-bed units in series. At the time of commissioning new resins, these are thoroughly cleaned by cycling.

In the final ion-exchange step, special resin grades of ultimate purity are used in nonregenerable "point-of-use" cartridges. These so-called semiconductor-grade resins are available from several producers and must comply with very strict specifications:

1) Extremely low content of leachable substances
2) Degree of regeneration close to 100 %
3) Absence of small resin fragments
4) Quick start-up rinse to specified resistivity and TOC value

12. Special Applications of Ion Exchange

Ever since their discovery, ion-exchange resins have been used primarily for water treatment. In 2000, 65 – 70 % of the resins produced were used for the softening or demineralization of water. The properties of ion-exchange resins were soon discovered to be advantageous in many areas other than water treatment. Since the end of the 1940s, resins have been used to treat sugar syrup or to extract precious metals. The number of such "special" applications has been constantly increasing because ion-exchange resins provide an efficient and elegant method for solving a large number of problems in chemical engineering. In many purification processes, the principles involved are similar to those described in previous chapters concerned with water treatment. However, specific properties of the solutions to be treated will affect the selection of resins: the feed solutions are often hot, colored, or viscous. Consequently, particle size distribution of the resin is adjusted to give acceptable pressure drops, and the porosity of the polymer is designed to prevent resin fouling or enhance the removal of color bodies or other large molecules from solution. Resins with uniform particle size (see Section 3.5) which have always been required for some applications, such as chromatographic separation, are now routinely available and have improved the efficiency of many processes.

Additionally, regulations now govern the use of ion exchange resins in many applications, particularly in the food processing industry. Many resins are authorized for the treatment of beverages and food, but regulations differ geographically.

There are also cases, such as catalysis, separation, or diffusion, where resins are used in a very different way, requiring specific columns and reactors. An exhaustive list of actual and potential applications cannot be given in this article. Table 9 lists in alphabetical order several uses of ion-exchange resins in the food, chemical, and pharmaceutical industries, and in hydrometallurgy. A few key examples are discussed in more detail. Some principal manufacturers of ion-exchange resins are listed in Table 11.

Table 11. Principal producers of ion-exchange resins

Company	Location	Principal trademarks
Bayer AG	Leverkusen, FRG	Lewatit
Dow Chemical Company	Midland, Michigan, USA	Dowex, Kastel
Mitsubishi Chemical	Tokyo, Japan	Diaion, Relite
Purolite Company	Bala Cynwyd, Pennsylvania, USA	Purolite
Rohm and Haas Company	Philadelphia, Pennsylvania, USA	Amberlite, Duolite, Imac
Sybron Corporation	Birmingham, New Jersey, USA	Ionac

12.1. Processing Steps

The uses of ion-exchange resins for various processing steps are described below. Resin manufacturers can provide more detailed information about specific processes and the corresponding ion exchange and adsorbent materials.

Table 9. Special uses of ion-exchange resins [a]

Product	Application[b]	Recommended resins [c], [d]
Acetic acid	Purification	Amberlite 252 – IRA67
Adipic acid	Catalyst recovery	Duolite C264NS H
Alkaloids	Separation & purification	Amberlite IR120, CR1320 K, IRC76, IRA402, IRA92
Alkylphenols	Catalysis	Amberlyst 15dry, 35dry, 36dry
Amino acids	Separation	Amberlite IRC50, 252, IRA67, IRA910, Duolite A7
Ammonium nitrate	Recovery & concentration	Duolite C264 – Amberlite IRA67
Anthocyanines	Recovery	Amberlite XAD16HP
Antibiotics	Purification (several processes)	Amberlite IRC50, XAD16HP, Amberjet 1500
Apple juice	Removal of hydroxymethylfurfural	Amberlite XAD16HP
Bisphenol A	Catalysis	Amberlyst 31wet
Borates	Removal from solution	Amberlite IRA743
Brine	Calcium, magnesium, strontium, barium removal	Duolite C467, Amberlite IRC748
Cadmium	Removal from waste	Duolite GT73, Amberlite IRC748
Caprolactam	Purification	Amberlite IRA402 – IR120
Carobs	Sugar recovery	Amberlite IR120 – IRA402
Casein	Precipitation	Amberlite 252
Catalysis (acidic)	Many processes	Amberlyst 15, 16, 31, 35, 36, 38
Cephalosporin	Extraction & purification	Amberlite IRA67, XAD1600
Chlorinated solvents	Removal from waste	Amberlite XAD4
Cholesterol	Drug for hypercholesterolhemia	Cholestyramine (Duolite AP143)
Chromates	Recovery from cooling circuits	Amberlite IRA96
Chromic acid	Iron removal	Amberlite 200C, Duolite C264NS
Chromic acid	Recovery & concentration	Amberlite 252 – IRA96, IRA92
Citric acid	Purification	Amberlite 252 – Duolite A561
Citrus juices	De-bittering	Amberlite XAD16HP
Cobalt	Recovery	Amberlite IRC76
Copper	Recovery & concentration	Amberlite IRC76, 252, IRC748, Duolite A7
Cyanides	Removal from waste	Amberlite IRA402, IRA958
Dates	Sugar recovery	Amberlite 252 – Duolite A561 – Amberlite IRA96
Electroplating	Rinse water recycling	Amberlite 252 – IRA96 – IRA410
Enzymes	Immobilization	Duolite A568, Amberlite IRA958
Esters	Esterification catalysis	Amberlyst 15wet, 36wet, Amberjet 1200 H
Formaldehyde	Acid removal	Amberlite IRA96, IRA92, IRA67
Fructose	Chromatographic separation	Amberlite CR1320 Ca
Fruit juices	Acid removal	Duolite A7, Amberlite IRA67
Fruit juices	Removal of fungicides	Amberlite XAD16HP
Gasoline	Octane enhancers (MTBE, ETBE, TAME)	Amberlyst 35wet, 15wet
Gelatin	Demineralization	Amberlite IR120 – IRA67, IRA92
Glucose	Demineralization	Amberlite 252 – IRA96, IRA92, IRA95
Glucose – fructose	Chromatographic separation	Amberlite CR1320 Ca
Glycerine	Demineralization & color removal	Amberlite IR120 – IRA96 – 200/IRA900
Glycol	Purification	Amberlite 252 – IRA96 – IRA900
Glyoxal	Nitrate removal	Amberlite IRA67
Gold	Recovery from plating solutions	Amberlite IRA410
Grape most	Rectification & sugar recovery	Amberlite SR1L – IRA92 – IRA900 – XAD16
Hydrochloric acid	Purification (iron and zinc removal)	Amberlite IRA402
Hydroculture	Resin as fertilizer	Lewatit HD5
Hydrogen peroxide	Purification	Amberlite XAD4, Amberlyst 15wet
Isomerase	Demineralization	Amberlite 252 – Duolite A561
Lactic acid	Demineralization	Imac HP1110 – Duolite A561
Lactose	Calcium removal	Imac HP336
Lactose	Demineralization	Imac HP1110 – Duolite A561
Lactose	Hydrolysis	Amberlyst 31wet
Lysine	Extraction & purification	Duolite C20N/2014

Table 9. (Continued)

Product	Application[b]	Recommended resins [c], [d]
Lysozyme	Recovery from egg white	Amberlite IRC50
Malic acid	Recovery from apple juice	Duolite A7
Mercury	Recovery or removal from waste	Duolite GT73
Methanol	Ammonia & sodium removal	Amberlite 252
Milk	Sodium removal	Imac HP1110
Molasse	Betain recovery	Amberlite CR1320 K, CR1310K
Molasse	Sugar recovery	Amberlite CR1320 K
Monosodium glutamate	Purification	Duolite C20N/2014, Duolite A7
Nickel	Iron and cobalt removal	Amberjet 4400
Nickel	Lead removal	Amberlite IRA96
Nickel	Recovery from waste	Amberlite IRC748, Amberlyst 35wet
Oxygen	Removal of traces from water	Amberlyst ER206
Pentaerythritol	Catalysis	Amberlite IRA958, IRA900
Phenol	Acid removal from phenol	Duolite A561, Amberlyst A21
Phenol	Purification	Amberlyst 16wet
Phenol	Removal from waste	Amberlite XAD4
Phosphoric acid	Purification	Amberlite 200C, 252
Pickling baths	Regeneration (Fe & Zn removal)	Amberlite IRA402
Platinum	Recovery	Amberlite IRA96, IRA402
Polyols	Demineralization (sugar industry)	Amberlite 252H, Amberlite IRA95, Amberlite IRA92
Polyols	Separation (sugar industry)	Amberlite CR1320 Ca, Amberlite CR1310 Ca
Semiconductor chips	Final rinse	Amberlite UP6040
Silica gels	Removal of sodium	Amberlite IRC84SP, IR120
Silver	Recovery	Amberlite IRA402, IRA67
Silver	Recovery	Amberlite IRA67, IRA402
Sorbitol	Demineralization	Amberlite 252 – IRA96 – 252/IRA900 – IRC76
Streptomycin	Purification	Amberlite IRC50, IRA404
Sucrose	Hydrolysis into glucose & fructose	Amberlyst 31wet
Sugar	Alkalization	Amberlite IRA958, IRA900
Sugar	Calcium removal (Gryllus process)	Amberlite 252
Sugar	Calcium removal (NRS process)	Amberlite SR1L
Sugar	Color removal from cane sugar	Amberlite IRA958 – IRA900
Sugar	Hydrogenation catalyst recovery	Imac HP336
Sugar	Quentin process	Imac C16NS
Sugar	Sucrose inversion	Amberlyst 31wet
Sugars	Chromatographic separation	Amberlite CR1320 Ca
Sweeteners	Final polishing (purification of traces)	Amberlite 252/IRA910, XAD16HP
Tartaric acid	Purification	Imac HP1110 – Duolite A561
Uranium	Mining (recovery from leach liquors)	Amberjet 4400, Amberlite IRA910U
Vanadium	Recovery	Duolite C264NS
Vitamins	Extraction	Amberlite XAD1180, XAD16, IRC50
Vitamins	Purification	Amberlite 252, IRA96
Whey	Calcium removal	Imac HP336
Whey	Demineralization	Imac HP336 – HP1110 – HP661
Xylitol	Purification	Amberlite IRA900
Zinc	Recovery	Amberlite IRA402, IRC748

[a] All listed applications are commercial [b] For food-processing applications, the prospective user must check whether the suggested resin meets local regulations [c] For other producers, see Table 10 [d] Dashes between resin names indicate that several resins are used in series; commas mean that several options are available

12.1.1. Purification

Ion exchange is used for the removal of impurities from an aqueous or organic solution. The term impurities covers here a large spectrum of chemical species.

Deacidification. A solution containing acids is passed through a strongly basic anion exchanger in OH^- form or a weakly basic exchanger in free base form, depending on whether all acids or only strong acids are to be removed.

Since this process is a neutralization reaction, the product of reaction with a strongly basic resin

Table 10. Equivalent resins [a,b]

Rohm and Haas	Bayer (Lewatit)	Dow (Dowex)	Mitsubishi (Diaion)	Purolite	Sybron
Weakly acidic cation resins					
Amberlite IRC50			WK10	C115	
Amberlite IRC76	CNP80				SR10
Amberlite IRC86	CNP80	MAC3, CCR2	WK20, WK40	C104, C105	CCP
Amberlite IRC86RF	CNP80WS	Upcore MAC3		C105FL	
Amberlite IRC86SB		MAC3LB		C105DL	
Imac HP333					
Imac HP336	CNP LF			C107E	
Strongly acidic cation resins					
Gel types					
Amberjet 1200 H		Monosphere C600 H, Marathon C H		PFC100 H	Impact CS398UPS
Amberjet 1200 Na	Monoplus S100	Monosphere C600, Marathon C		PFC100	
Amberjet 1500 H		HGRW			Impact CS399UPS
Amberlite IR100 Na		HCRS(E)		C120E	
Amberlite IR120 H	S100 H	HCRS H		C100MB H	
Amberlite IR120 Na	S100	HCRS	SK1B	C100	C249
Amberlite IR122 Na		HGR	SK110	C100x10	C399
Amberlite SR1L	S100LF			C100E, PF100E	
Imac HP1110 Na					
Macroporous					
Amberlite 200C Na					
Amberlite 252 H		MSC H		C150 H	CFP110
Amberlite 252 Na	SP112	MSC Na, 88	PK220	C150	
Amberlite 252RF H	Monoplus SP112	Marathon MSC		C150FL	
Ambersep 252 H	Monoplus SP112	Marathon MSC		C150TL	
Imac C16NS					
Weakly basic anion resins					
Amberlite IRA67	VPOC1072		WA10	A845, A847	
Amberlite IRA67RF				A845FL	
Amberlite IRA95			WA30		A329
Amberlite IRA96	MP64	MWA1	WA20	A100	A328
Amberlite IRA96RF	Monoplus MP64	Monosphere WB500		A100FL	
Amberlite IRA96SB		MWA1LB		A100DL	
Strongly basic anion resins					
Gel types					
Amberjet 4200 Cl	Monoplus M500	Marathon A, Monosphere A625		PFA400	Impact AG1P UPS
Amberjet 4400 Cl		Marathon ALB			Impact AG1 UPS
Amberjet 4600 Cl	Monoplus M600	Marathon A2, Monosphere A2 500		PFA300	Impact AG2 UPS
Amberlite IRA402 Cl	M500/M511	SBRP	SA10A	A400	ASB1P
Amberlite IRA402 OH				A400MB OH	
Amberlite IRA404 Cl			SA11A	A420S, A444	
Amberlite IRA410 Cl	M600, M610	SAR	SA20A	A200, A300	ASB2
Amberlite IRA458 Cl	VPOC1071			A850	
Amberlite IRA458RF Cl				A850FL	
Amberlite IRA478RF Cl	VPOC1073			A870	
Macroporous					
Amberlite IRA900 Cl	MP500	MSA1	PA308, PA312	A500, A500PS	A641
Amberlite IRA900RF Cl	Monoplus MP500	Marathon MSA		A500FL	
Amberlite IRA910 Cl	MP600	MSA2	PA412, PA416	A510	A651
Amberlite IRA958 Cl	VPOC1074			A860	
Ambersep 900 SO4		Marathon MSA		A500TL	A642
Imac HP555	SN36			A520E	SR7
Mixed resins (ready for use)					
Amberlite MB6113				MB500 IND	NM65

Table 10. (Continued)

Amberlite MB20	SM92	MB50		MB400	
Amberlite MB9L				MB46	
Nuclear-grade resins					
Amberlite IRN77	S100KR H	HGR NG	SKN1	NRW100	NC10
Amberlite IRN78	M500KR OH	SBR NG	SAN1	NRW400, NRW600	NA38
Amberlite IRN97 H		Monosphere 500C NG			
Amberlite IRN150		MR3LC NG	SMN1	NRW37	NM60
Amberlite IRN160		Monosphere MR500LC NG			
Semiconductor-grade resins					
Amberjet UP1400	Ultrapure 1212	650CUPW	SKNUP		
Amberjet UP4000		550AUPW	SANUP		
Amberjet UP6040	Ultrapure 1293	MR450UPW		EXMB	NM201SG
Amberjet UP6150		MR3UPW			
Inert spacer resins					
Amberlite RF14	IN42	IF59		IP4	
Ambersep 359		Monosphere 600BB		IP5	
Chelating resins					
Amberlite IRC748	TP207, TP208		CR10, CR11	S930	SR5
Duolite C467	TP260			S940	SR12
Duolite GT73	TP214			S920	SR4
Polymeric adsorbents					
Amberlite XAD16			HP20		
Amberlite XAD4			HP10		

[a] Only the most current resins are listed; [b] Correspondence of resins is only approximate. For detailed advice, consult the resin manufacturers

is water, and with a weakly basic resin the acid is taken up as a whole, so that the solution is effectively purified:

$$RN(CH_3)_2 + H^+Cl^- \longrightarrow [RNH(CH_3)_2]^+Cl^- \text{ or } RN(CH_3)_2 \cdot HCl$$

Examples: Removal of formic acid from formaldehyde, deacidification of fruit juice, and removal of mineral acids from alcohols.

Demineralization. All cations and anions may be removed (*total demineralization*) or only some of them, generally the strongly dissociated ions (*partial demineralization*). This is achieved by passing a solution successively through a cation-exchange resin in H^+ form and an anion-exchange resin either in OH^- or free base form. Non-ionized substances in the solution are largely unaffected by this treatment. This type of demineralization is sometimes called deashing.
Examples: Treatment of sugar syrups, glucose, antibiotics, glycerin, cheese whey, alcohols, polyols (e.g., sorbitol), and glycols.

In a variation of this process, organic acids are purified by using strongly acidic and weakly basic resins. Inorganic salts are removed, whereas weakly dissociated organic acids are not taken up by the anion-exchange resins. If the organic acid has a pK value lower than that of the resin, it is taken up during the first part of the run but then displaced by the stronger mineral acids, so that losses of the valuable organic acid are negligible or minimal. As weakly basic resins with different structures and active groups have different pK values, proper choice of resin will greatly enhance separation.
Examples: Demineralization of citric and lactic acids.

Selective Removal of Impurities. Harmful ions must be removed, while other constituents of the solution are affected as little as possible. Various sorts of selective resins are available.
Examples: The commonest applications are

1) Selective removal of toxic metals from industrial effluents by using complexing (chelating) resins (see Section 2.2.3) such as Amberlite IRC748 for transition metals and

Duolite GT73 for mercury, cadmium and other heavy metals
2) Nitrate removal from drinking water by using the Cl^- or HCO_3^- form of anion-exchange resins (see Section 11.2)
3) Calcium removal from saturated brine with an aminophosphonic resin (e.g., Duolite C467) that can remove Ca^{2+} in a concentration of 10^{-4} g/L from brine containing $>10^2$ g/L of Na^+
4) Removal of borate from drinking water with a resin having glucamino groups (e.g., Amberlite IRA743)

Decolorizing. differs from selective removal in that it normally combines ion exchange with adsorption. Resins have partially replaced bone char or activated carbon for this application, as they are easier to handle and to regenerate. Decolorizing is applied mainly to solutions of sugar or other organic compounds. Examples of the color bodies to be removed from solution are sugar degradation products, flavonoids, polyphenols, melanoidins, tannins, and anthocyanines. Because impurities responsible for the color of these solutions are often acidic, the resins used are normally highly porous anion exchangers, sometimes in combination with carbon, inorganic adsorbents, or nonionic resins.
Examples:

1) Strongly basic polystyrene materials in chloride form, strongly basic polyacrylic materials in chloride form are used for color removal from cane sugar, and regenerated with a sodium chloride brine
2) Weakly basic or nonionic phenol – formaldehyde resins are used for citric or tartaric acid
3) Acrylic weakly basic resins simultaneously decolorize and deacidify antibiotics
4) Aromatic polymeric adsorbents are used for the removal of polyphenols or anthocyanines

Adsorption of Impurities. True adsorption is not strictly speaking an ion-exchange process. However, adsorbent resins are so similar to ion-exchange resins that the process can be mentioned here. The adsorbent materials used are macroporous polystyrene such as the Amberlite XAD series or phenol – formaldehyde resins without active groups.
Examples: Current applications include the removal of impurities from effluents containing phenol derivatives, pesticides, chlorinated solvents, or dyes. Macroporous phenol – formaldehyde resins have also been used in cigarette filters to remove tar, aldehydes, and heavy nitriles.

12.1.2. Ion Substitution

All ion-exchange processes are ion substitutions. However, a distinction should be made between processes that remove ions while producing water or a pure solution of nonionic substances and those that leave the total ion concentration unchanged, merely replacing an undesirable ion by an acceptable one.

Softening. The principle of softening is the same as that used in water treatment: Ca^{2+}, Mg^{2+}, and possibly Fe^{2+} and Mn^{2+}, are replaced by Na^+ ions.
Example: The most widespread application is in the beet sugar industry, where the crude juice containing a high concentration of calcium is softened to prevent scaling in the evaporators. Variations on the basic softening process have been developed to save regenerant and reduce waste. One of the most efficient is the NRS process, in which regeneration of the cation-exchange resin is carried out with a mixture of the concentrated softened juice and a little caustic soda, so that no additional salt is required and the process produces no waste.

Desodation is a kind of reverse softening: the sodium in a solution is replaced by calcium, magnesium, or potassium, and the cation resin used is regenerated by salts of one or more of the latter metals.
Examples: Sodium removal from milk (Na – Ca, Na – K exchange) and the Quentin process in sugar refining (Na – Mg, K – Mg exchange).

Production of Organic Acids and Salts. Organic acids can be prepared from their salts simply by passage through a cation-exchange resin in H^+ form. This process has an excellent yield and is used widely to manufacture modified, water-soluble ethylenediaminetetraacetic acids and some amino acids from their sodium

salts. Conversely, an organic acid can be converted into its salt. One organic salt can also be converted into another.
Example: In the pharmaceutical industry, cephalosporin C is converted into its potassium salt by passage through a cation-exchange resin in K^+ form.

12.1.3. Recovery and Concentration

In recovery and concentration, the principle is the same as in selective purification, but the aim is different. Here, the objective is to recover a valuable substance from solution, so that the stage during which elution from the resin takes place is particularly important. In most cases, the substance recovered from the dilute solution is simultaneously concentrated during the regeneration stage.
Examples:

1) Recovery of traces of precious metals, most often in the form of metal complexes, in the metal plating industry (jewelry) and silver recovery from photographic baths
2) Isolation of pharmaceutical compounds during the manufacturing process (extraction and purification)
3) Recovery of ammonium nitrate in fertilizer plants
4) Recovery of sugar and amino acids from molasses in sugar refining by ion exclusion
5) Recovery and recycling of water by demineralizing the rinse water from surface treatment plants
6) Recovery of uranium or gold in the mining industry

12.1.4. Separation

A special type of ion-exchange resin (with fine, uniform particles and an accurately defined moisture content) is used in columns several meters tall for various industrial separation methods that are similar to those in chromatography. This technique allows separation of a mixture of ions, an electrolyte from a nonelectrolyte, and a mixture of nonelectrolytes (see also → Process-Scale Chromatography; → Basic Principles of Chromatography).

For separations involving nonelectrolytes, a special type of ion-exchange resin with fine, uniform particles and an accurately defined moisture content is used in columns several meters tall. Cation exchange resins with a mean diameter of 250 to 350 µm and a uniformity coefficient of less than 1.1 are used in many processes, in the appropriate ionic form, usually Ca, K, or Na. In the pharmaceutical field, chromatographic separations are carried out with polymeric adsorbents of similarly uniform particle size.

As shown in Tables 4 and 5, ions can be arranged in order of their relative affinity for a resin. For a sulfonated polystyrene exchange material, the following monovalent ions are listed as examples in order of increasing affinity:

$$Li^+ < H^+ < Na^+ < NH_4^+ < K^+ < Rb^+ < Cu^+ < Ag^+$$

Separation by Displacement. A small amount of the mixture to be separated is passed through a column (e.g., a mixture containing K^+ and Cs^+ is passed through a resin in Na^+ form). The ions taken up are then displaced by means of another ion with a greater affinity (e.g., Ag^+). The emergent liquor is thus fractionated because the silver ions, in turn, displace the sodium, potassium, and cesium ions.

Separation by selective displacement is a variation of the last process. A small amount of the mixture to be separated (e.g., Li^+ and Cs^+) is passed through a column containing a resin in a form with intermediate affinity (e.g., NH_4^+). This is followed by a solution containing NH_4^+ which displaces only the lithium. Lithium is therefore completely separated from cesium. Cesium can then be displaced by silver (Fig. 56).

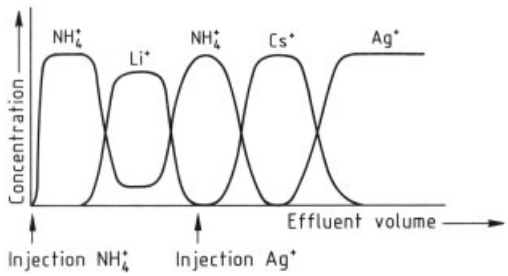

Figure 56. Separation of Li^+ and Cs^+ by selective displacement in a resin initially in NH_4^+ form

This technique is called displacement chromatography. It can be applied to the separation of amino acids which are taken up by an H^+ form of cation-exchange resin column and displaced with sodium hydroxide or ammonium hydroxide.

Separation by Elution. A column is prepared as above but, instead of ions being displaced by an ion with greater affinity than those taken up, they are now eluted with an ion of lower affinity. This is normally the ion with which the resin was initially loaded (e.g., H^+). The ions to be separated are displaced toward the bottom of the column by the eluant (HCl in the example shown in Fig. 57). The ions always emerge in order of their increasing affinity for the resin and are usually separated well.

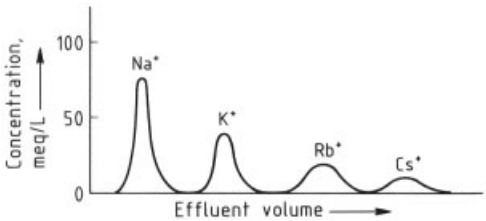

Figure 57. Separation of Na^+, K^+, Rb^+, and Cs^+ in a solution by elution with 0.1 N HCl through a resin initially in H^+ form

Separation by Ion Exclusion. An electrolyte can be separated from an electrolyte in an ion-exchange column by a sorption process (Fig. 58). The ion with which the resin column is loaded is normally the same as one of those in the electrolyte to be separated. For instance, an aqueous solution of a mixture of sodium chloride and glycerol can be passed through a column containing a cation-exchange resin in Na^+ form. Cation exchange does not take place because the column and the solution contain the same cation. The Cl^- anion cannot penetrate the resin because it encounters the Donnan potential which guarantees electrical neutrality within the resin (see Section 6.1). The electrolyte is thus excluded from the resin. The nonelectrolyte, on the other hand, can penetrate the resin by means of adsorption until its concentration is the same inside the beads as outside. When this equilibrium state has been reached, pure water is passed down the column and drives out the electrolyte more quickly than the glycerol, which must diffuse out of the beads. This process of alternating input of the mixture and displacement by water is repeated without the need to regenerate the resin, and successive fractions containing pure glycerol are obtained.

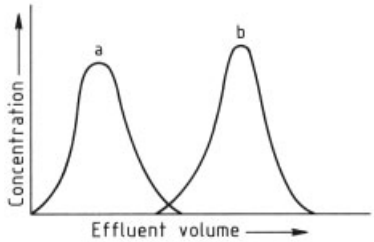

Figure 58. Separation of an electrolyte from an organic product by ion exclusion
a) Electrolyte; b) Organic compound

Examples: An exclusion process is used industrially to recover sugar from sugar beet molasses [59]. Figure 59 shows the successive fractions obtained, alternately rich in sugar and nonsugars. The same process can be applied to the separation of strong and weak electrolytes (e.g., hydrochloric acid from organic acids).

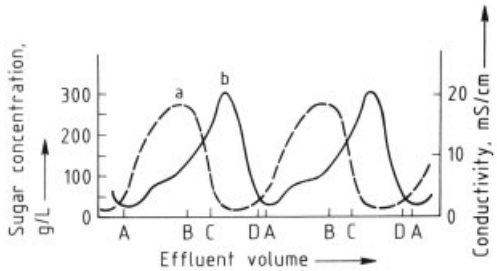

Figure 59. Recovery of sugar from molasses by ion exclusion
a) Conductivity; b) Sugar concentration
Fractions: AB nonsugars (residual molasses); BC recycled; CD sugar produced; DA recycled

Chromatography of Nonelectrolytes. Another property of resins becomes evident during separation by ion exclusion: if different nonelectrolytes are present in solution, the process not only separates out the mineral salts but also the nonelectrolytes from each other because of their different rate of uptake and release by the resin.

Example: Separation of glucose from fructose is carried out commercially with fine bead sul-

fonic resins, generally in the calcium form [60]. A typical elution profile is shown in Figure 60.

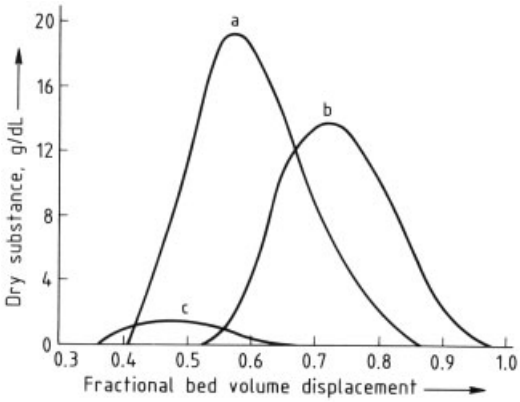

Figure 60. Separation of glucose from fructose in high-fructose corn syrup
a) Glucose; b) Fructose; c) Oligosaccharides

Separation by Acid Retardation. Somewhat surprisingly, an acid can be separated from its salts by using a column containing a strongly basic anion-exchange resin of suitable porosity and particle size [61]. This occurs because, at high concentration, the acid crosses the Donnan potential barrier (Donnan invasion) and is taken up by the resin, whereas the salts are excluded from it. The acid is thus "retarded", and the salts are allowed through. Pure fractions can be obtained at the output by passing concentrated electrolyte solution and water alternately down the column.
Example: Acid retardation is now used industrially for purifying and recycling acids that have been used for metal pickling. An elution curve is shown in Figure 61.

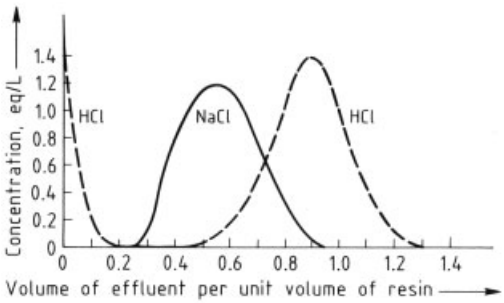

Figure 61. Acid retardation by a strong anion resin

12.1.5. Diffusion

An ion-exchange resin already loaded with ions releases them slowly into a solution. This property is used in two special applications.

Hydroponics. (Soilless Cultivation). A resin loaded with fertilizing elements can be used for the hydroculture of plants instead of liquid fertilizers [62]. The elements are released gradually into the water as needed by the plant (Fig. 62).

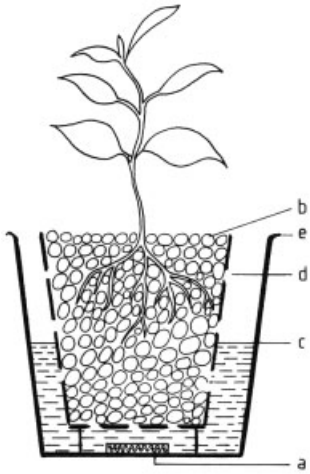

Figure 62. Hydroculture
a) Hydroponic resin; b) Beads of blown clay; c) Water level; d) Inner perforated pot; e) Outer pot

Delayed-Action Drugs. Instead of administering a pure drug directly, it can be absorbed onto a suitable resin and taken in this form. The active substance is then released more slowly in the stomach.

12.1.6. Catalysis (→ Heterogeneous Catalysis and and Solid Catalysts, Chap. 3.1.5.)

Acid or Alkaline Catalysis [63]. Because standard ion-exchange resins are insoluble acids or bases, they can be used with advantage in many organic reactions where an acidic or basic catalyst is required. Special macroporous sulfonic resins are used, for example, in esterification (e.g., the production of butyl acetate), synthesis (e.g., of methyl *tert*-butyl ether), or the hydrolysis of sucrose to glucose and fructose.

Catalysis with ion-exchange resins has the following advantages over the use of sulfuric acid:

1) A higher local concentration of H^+ (or OH^-) ions
2) No corrosion
3) Possibility of use in continuous processes
4) Fewer secondary reactions
5) Easy separation from the reaction medium (by simple filtration)

Catalyst Carrier. Some resins can be loaded with a metallic catalyst (palladium, silver, nickel, etc.), which is otherwise difficult to use because of its high solubility.

Enzyme Immobilization. Some ion-exchange resins can be loaded with enzymes to enable enzymatic reactions to take place continuously. Glucoisomerase or β-galactosidase, for example, can be taken up by weakly basic phenol – formaldehyde resins (see also → Immobilized Biocatalysts).

12.1.7. Dehydration

Ion-exchange resins are strongly hydrophilic. When dried, they tend to take up their quota of swelling water. This property is exploited in the two following applications.

Drying of Gases or Solvents. Previously dried sulfonic resins can be used to lower the humidity of air, gases, or organic solvents. The resins take up their own weight of water.

Dehydration of Alcohols. The affinity for water is so great that sulfonic resins can be used to dehydrate alcohols, (e.g., in converting ethylene glycol into dioxane or cyclohexanol into cyclohexene.

12.1.8. Coalescence on Oleophilic Resins

An oleophilic resin can be obtained by loading a resin permanently with an organic substance consisting of long-chain molecules [64]. In practice, a sulfonic resin loaded with an organic cation is used (e.g., an aromatic quaternary ammonium ion):

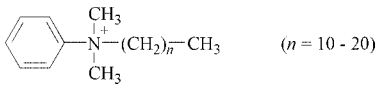

or a pyridinium ion:

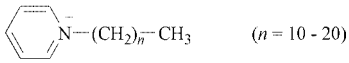

The cationic part of the oleophilic molecule is taken up almost irreversibly by the cation-exchange resin. An oleophilic resin bead of this type is illustrated in Figure 63.

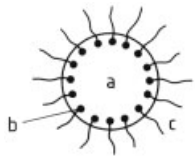

Figure 63. Schematic representation of an oleophilic resin bead
a) Sulfonic resin bead; b) Ionic end of oleophilic molecule; c) Oleophilic end of molecule

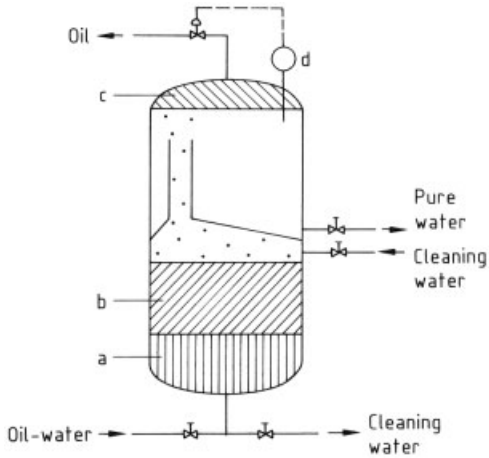

Figure 64. Elf coalescing deoiler
a) Oil – water mixture; b) Oleophilic resin; c) Decanted oil; d) Level controller operating the oil outlet valve

Water containing oils or fats passes through a column containing an oleophilic resin. The oil particles are taken up by the resin beads in the form of microscopic droplets. The droplets coalesce and grow in size as more oil is taken up. Beyond a certain size, the drops become detached from the resin and are then decanted in the normal way. This process has been developed on

an industrial scale as the Elf – ANVAR process [65]. A typical coalescing deoiling unit is shown in Figure 64.

12.1.9. Liquid Ion Exchangers
(→ Liquid – Liquid Extraction)

Liquid ion exchangers are high molecular mass, water-insoluble acids and bases that are soluble in oil and in organic solvents. The compounds have a hydrophobic structure and an ionogenic group. They are usually phosphonic acids and amines substituted with long-chain ($C_8 - C_{24}$) aliphatic groups. They are used for liquid – liquid extraction of electrolytes from aqueous solutions.

An example of cation exchange is the following reaction with a polyalkyl phosphonic acid:

$$(R_3PO_4H)_{org} + M^+ \rightleftharpoons (R_3PO_4M)_{org} + H^+$$

An example of anion exchange is the following reaction with a polyalkylamine in salt form:

$$(R_3NHA)_{org} + B^- \rightleftharpoons (R_3NHB)_{org} + A^-$$

where M^+ represents a metal cation, A^- and B^- represent two different anions, R is an alkyl substituent, and the subscript org denotes the organic phase. Recovery of the extracted component from the organic phase does not require distillation as is often the case in processes based on solubility differences. Instead, the liquid exchanger is "stripped" in a second liquid – liquid extraction, similar to the regeneration of solid ion exchangers. The regenerant (stripping agent) is an acid, base, or salt in aqueous solution. In practice, the liquid ion exchanger is used as a solution in an organic solvent (frequently kerosene), with an organic : aqueous phase ratio of between 1 : 1 and 1 : 4. In the exchange process, both immiscible phases are thoroughly mixed mechanically in conventional liquid – liquid extraction equipment (mixer – settlers, centrifugal contactors, and pulsed and nonpulsed extraction columns).

The advantages of liquid ion exchangers are the relative simplicity of their production, the possibility to adjust their concentration in the organic phase according to the concentration of the aqueous solution, and the possibility of continuous countercurrent operation. Their main disadvantage is the inevitable losses of solvent and ion exchanger during the extraction process.

The main industrial application of liquid ion exchangers is hydrometallurgy. Large plants are in operation for the recovery of uranium from sulfuric acid leach liquors. The uranyl sulfate anion is extracted with a dialkylamine and subsequently stripped with sodium carbonate, resulting in a water-soluble uranyl carbonate complex which can be precipitated (e.g., by addition of sodium hydroxide) to produce a "yellow cake" of uranium oxide.

Examples: Amberlite LA-1 and LA-2 liquid anion exchangers Rohm and Haas, USA.

12.1.10. Ion-Exchange Membranes
(→ Membranes and Membrane Separations)

Ion-exchange membranes are materials with ion-exchange properties that can be used as a separation wall between two solutions. They show pronounced differences in permeability toward counterions, co-ions, and neutral molecules. The membranes are permselective. Cationic membranes are permeable to cations and impermeable to anions. Anionic membranes are permeable to anions and impermeable to cations.

Ion-exchange membranes are classified as hetero- or homogeneous:

Heterogeneous membranes are made of finely ground ion-exchange resin particles embedded in an inert porous binder [polyethylene, poly(vinyl chloride)] in the form of flat sheets. The resin is usually sulfonated polystyrene for cationic membranes and polystyrene with quaternary ammonium groups for anionic membranes. To increase dimensional stability, the mixture is often reinforced with a woven fabric.

In *homogeneous membranes* the ion-exchange groups are grafted directly to the membrane polymer structure. Their manufacture is similar to that of bead-form ion-exchange resins. They are somewhat less stable than heterogeneous membranes, but have a more uniform structure. In the absence of electric current, the membrane is slightly permeable to nonelectrolytes and impermeable to electrolytes because the Donnan potential prevents co-ions from passing through it. Furthermore, the principle of electroneutrality makes it impossible

for counterions to concentrate on one side of the membrane. When an electric potential is applied, however, cations are attracted by the cathode and anions by the anode. In the case of a cation-exchange membrane, only cations can wander through it.

An example of application is the *electrolysis of brine* using a cation-exchange membrane (chlor-alkali process) to produce chlorine and caustic soda (→ Chlorine, Chap. 7.). Sodium chloride and water are introduced into the anode chamber where oxidation of the chloride ions to chlorine takes place. Water in the cathode chamber is reduced to hydrogen and hydroxyl ions. Sodium ions wander through the cation-exchange membrane into the cathode compartment and combine with the hydroxyl ions to form sodium hydroxide. Another important example is *electrodialysis,* where both cation- and anion-exchange membranes are used for demineralization of water (Fig. 65). If all three chambers are filled with sodium chloride solution, sodium passes through the cation-exchange membrane and produces NaOH in the cathode chamber, chloride goes through the anion-exchange membrane and produces HCl and the central chamber is progressively demineralized.

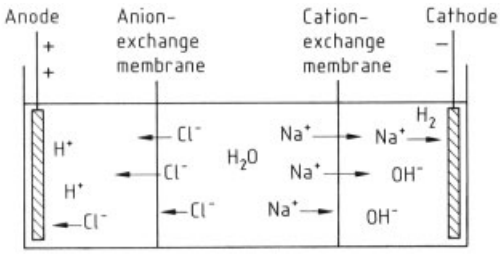

Figure 65. Electrodialysis for demineralization of water

In industrial practice, a large number of alternating cation- and anion-exchange membranes are combined for demineralization of water, especially brackish or sea water. Electrodialysis is also used for the demineralization of cheese whey to recover valuable proteins and lactose.

Membrane electrodes operating according to the principle of permselectivity can be used for analytical purposes. Ion-specific membranes are now available for measuring.

Important ion-exchange membrane manufacturers are Asahi Chemical (Japan), Asahi Glass (Japan), Du Pont (USA), Ionics (USA), Rhône-Poulenc (France), Sybron (USA), and Tokuyama Soda (Japan).

12.2. Technical Considerations

For special applications in softening, purification, demineralization, or recovery, the plants are often similar to those used in water treatment. Counterflow regeneration is sometimes used to improve regeneration efficiency and reduce leakage. Continuous systems (see Section 10.3) can be employed, e.g., in the softening of sugar syrup or the treatment of rinse water in electroplating. However, many special applications require special operating conditions.

Flow Rates. Concentrated solutions must frequently be dealt with, which requires relatively large amounts of resin. The specific flow rate is often low: 0.2 – 20 BV/h instead of the 10 – 100 BV/h used in water treatment.

Regenerant Recycling. Regenerant consumption can be reduced by recycling. The tails, which are relatively uncontaminated, can be used at the head of the next regeneration stage.

Merry-go-round Systems. To increase the treatment capacity of a plant and improve the quality of the treated liquid, a so-called merry-go-round system can be used. This involves passing the solution through several identical columns in series (at least two, but usually more), so that the one at the head is exhausted as completely as possible. When this state of exhaustion is reached, the next column is placed at the head of the series, and the one just regenerated is used for polishing while the exhausted column is regenerated. The cycle shifts by one column as each column at the head of the chain is exhausted. In this way, the product is obtained in the most concentrated form possible (Fig. 66).

Sweetening Off and On. In most of the purification processes described in Section 12.1.1, the liquid to be treated must be removed from the resin before the latter is regenerated. This involves inserting a rinsing stage before regeneration, a stage generally known as *sweetening*

off because of its similarity to the treatment of sugar syrup. In the same way, after this stage and a rinse of the water from the exchangers, the water filling the column is replaced by the liquid to be treated, a stage known as *sweetening on*.

The complete cycle thus consists of sweetening on, the service stage, sweetening off, decompaction, regeneration (or elution), and rinsing. Decompaction is sometimes carried out prior to sweetening off, with the liquid to be treated.

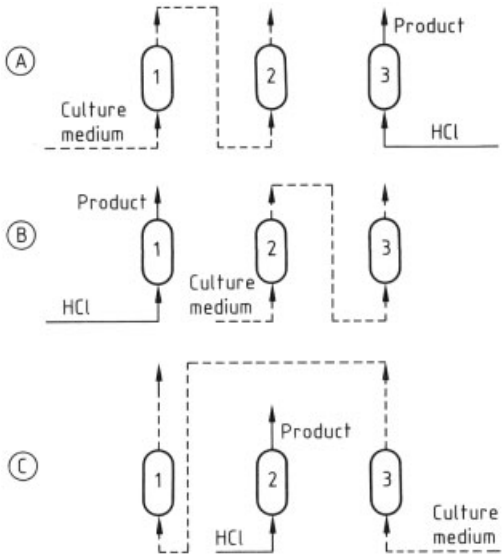

Figure 66. Extraction of streptomycin with weakly acidic resins (three columns in series)

Multistage Regeneration. The resin must sometimes be regenerated in several steps. In the first step, the substances taken up in one ionic form are removed or recovered. In the next step, the resin is converted into another ionic form more suitable for the service stage.

In the selective uptake of metals, for example, a carboxylic or chelating resin (see Section 2.2.3) is often used in Na^+ form in order not to lower the pH of the solution to be treated. However, the resin must be regenerated with acid. The H^+ form must then be converted to the Na^+ form with sodium hydroxide:

Service:

$$2\,RCOONa + ZnSO_4 \longrightarrow (RCOO)_2Zn + Na_2SO_4$$

Regeneration:

$$(RCOO)_2Zn + H_2SO_4 \longrightarrow 2\,RCOOH + ZnSO_4$$

Conversion to Na^+ form:

$$RCOOH + NaOH \longrightarrow RCOONa + H_2O$$

Temperature Control. Some viscous liquids, particularly concentrated organic solutions (e.g., heavy sugar syrups), must be decolorized when hot (70 – 80 °C) so that they remain fluid and do not crystallize. Other liquids, such as light sugar syrups, must be demineralized when cool (12 °C) so that they are not hydrolyzed when they come in contact with the cation exchanger in H^+ form. Temperature control thus plays a major role in many applications.

Special Plants. Some applications require special types of plant, e.g.,

1) Very long columns up to 10 m tall for sugar chromatography
2) Columns in the shape of truncated cones widening at the top to enable resins to swell during service or regeneration
3) Fluidized-bed systems, sometimes using high-density resins to treat liquids containing suspended matter

Simulated moving beds (SMBs) are progressively replacing the tall columns of the first industrial plants. Whilst the old plants had to be operated in a sequential, discontinuous way, alternating the feed solution to be separated and an eluant, the new SMB technology allows the separation of two or more compounds in a quasi-continuous, stepwise manner: the columns (typically 30 of them) are arranged in a carousel. The major system of this type is called ISEP [71] and is provided by the U.S. company AST (Advanced Separation Technologies). The feed and elution solutions are connected to a stationary upper distributor fitted with typically 20 ports, and the raffinate and the extract are connected to a lower stationary connector fitted with the same number of ports. The columns themselves are mounted on a rotating frame. The carousel rotates continuously at a speed of 0.1 to 1.5 revolutions per hour, and the ports are thus successively connected to all columns. A simplified scheme is shown in Fig. 67 with only eight resin

columns and six inlet and outlet ports. Figure 68 shows the principle of operation in two dimensions.

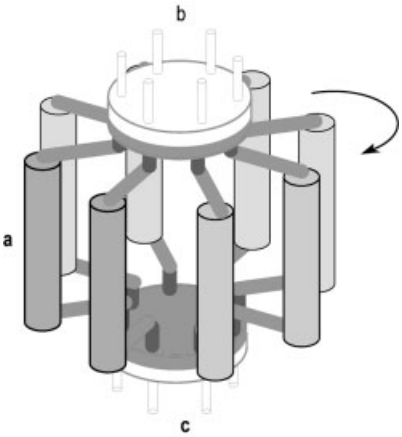

Figure 67. ISEP carousel (simplified)
a) Rotating resin columns; b) Stationary inlet ports; c) Stationary outlet ports

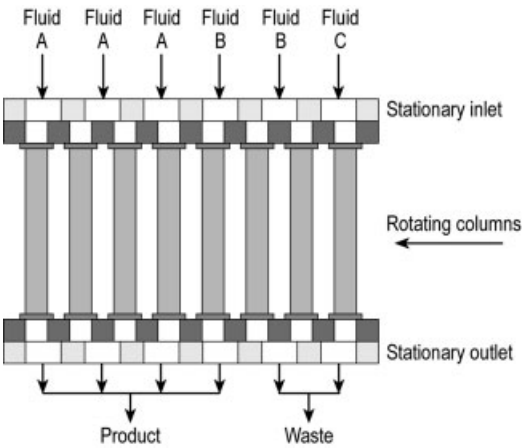

Figure 68. ISEP: two-dimensional view

Simulated moving beds can be used for purification of fermentation broth, sugar syrup deashing, decolorization of various solutions, separation of metals and other applications.

13. References

1. H. S. Thompson: "On the adsorbent power of soils," *J. R. Agric. Soc. Engl.* **11** (1850) 68.
2. J. T. Way, *J. R. Agric. Soc. Engl.* **11** (1850) 313.
3. R. Gans, *Jb. Preuss. Geol. Landesamt* **26** (1905) 126.
4. O. Liebknecht, US 2 191 060, 1940.
5. P. Smit, US 2 191 063, 1940.
6. B. A. Adams, E. L. Holmes, *J. Soc. Chem. Ind. London* **54** (1935) 1.
7. G. F. D'Alelio, US 2 366 007, 1945.
8. C. H. Mc Burney, US 2 591 573, 1952.
9. The Permutit Co. Ltd., GB 849 112, 860 695, 1960 (J. R. Millar).
10. Farben Fabriken Bayer., DEI 045 102, 1957.
11. Rohm and Haas Co., GB 932 125, 932 126, 1959.
12. R. E. Anderson: "Fundamentals of column ion-exchange in fully-ionised system," *18th Annual Liberty Bell Corrosion Course (USA),* 1980.
13. S. Fisher, G. Otten: "Sloughage of organic materials from field-decross-linked sulfonic acid cation exchange resins," *Proceedings of the 42nd International Water Conference,* Pittsburgh, Pennsylvania, 1981.
14. B. Hoffman, M. Kasahara, M. Gavaghan: "The effects of chlorine on mixed beds in ultrapure water systems," *Sixth annual semiconductor pure water conference,* 1987.
15. D. C. Auerswald: "Effects of cation resin leachables on condensate polisher system performance," *Ion Exchange for Industry* , Ellis Horwood Ltd, Chichester 1988, p. 11.
16. R. Kunin: Thermal stability of anion exchange resins. Amber-Hi-Lites **139**, 1974 (published by Rohm and Haas, Philadelphia).
17. R. E. Anderson: "Estimation of ion exchange process limits by selectivity calculations," *AIChE Symp. Ser.* **152** (1973) no. 71, 236.
18. R. E. Anderson: "Basics of column ion exchange with strong acid and strong base resins," *Third Annual Convention Am. FESAAC* , Mexico, 1982.
19. F. G. Helfferich: *Ion Exchange,* McGraw Hill, New York 1964.
20. F. G. Helfferich: *Ion exchange Kinetics – Evolution of a Theory, Mass Transfer and Kinetics of Ion Exchange,* NATO ASI Series, Series E N° 71, Martins Nijhoff, The Hague 1983, 157.
21. J. T. McNulty, M. Eumann, C. A. Bevan, V. C. Tan: "Anion exchange resin kinetic testing: an indispensable diagnostic tool for condensate polisher troubleshooting," *Proceedings of the 47th International Water Conference,* Pittsburg, Pennsylvania, 1986.

22. N. J. Ray, M. Ball, A. Coates: "Kinetic performance tests of anion exchange resins," *Proceedings of the 42nd International Water Conference,* Pittsburgh, Pennsylvania, 1981.
23. B. J. Hoffman, J. C. Konen, J. T. McNulty: "The Significance of Ion Exchange Equilibria and Kinetics in the Production of High Purity Water," *Ultrapure Water,* October 1987, 14.
24. J. J. Wolff: "Une technique moderne d'utilisation des échangeurs d'ions: la régénération à contre-courant," *L'Eau et l'Industrie* (F) **10** (1976) 67.
25. E. W. Jackson, J. H. Smith: "Make up treatment – Countercurrent regeneration experience in the United Kingdom," *Proceedings of the 38th International Water Conference,* Pittsburgh, Pennsylvania, 1977.
26. A. D. Prunac, E. Zaganiaris: "Development of Acrylobed: a new anion exchange resin system for water demineralization," *Proceedings of the 44th International Water Conference,* Pittsburgh, Pennsylvania, 1983.
27. Farbenfabriken Bayer AG, DE 1 442 689, 1963 (F. Martinda, G. Siegers).
28. Rohm and Haas Company, EP 0 142 359 A2, 1985 (A. D. Prunac, J. Dobias).
29. Degrémont, Brev. Franç., 7 834 555, 1978 (C. Barraqué, J. Burriat, J. M. Rovel).
30. B. V. Esmil, GB 1 539 161, 1977 (S. Evans, J. H. Hoostens, A. J. P. Van Eck).
31. Pintsch-Bamag AG, DE 1 807 694, 1968 (W. Lux).
32. G. J. Crits, W. Y. McIntire, D. R. Zimmermann: "Co-counter-current regeneration in ion exchange systems," *Proceedings of the 44th International Water Conference,* Pittsburgh, Pennsylvania, 1983.
33. C. J. Brown, C. J. Fletcher: "Water deionization by Recoflo short bed ion exchange," *Proceedings of the 47th International Water Conference,* Pittsburgh, Pennsylvania, 1986.
34. I. M. Abrams, F. de Dardel, J. G. Grantham: *Six years operating experience with Triobed,* PWT 8202A, Duolite International 1982.
35. M. A. Sadler, J. C. Bates, D. J. Gardner, A. D. Thomas: "A study of the Triobed condensate polishing plant system at Doel Power Station," *Nucl. Energy* , Br. Nucl. Energy Soc. 18 (1979) no.6, 389.
36. M. A. Sadler, J. C. Bates, G. R. Mills: "Anwendung einer "Triobett"-Kondensatreinigungsanlage im Kraftwerk Fawley, einschließlich des Betriebes in der Ammonium-Form," *VGB Kraftwerkstechnik* **61** (1981) no. 3, 221.
37. M. J. Slater: "Recent industrial-scale applications of continuous resin ion exchange systems," *J. Sep. Process Technol.* **2** (1981) no. 3, 2.
38. ANON: *Continuous ion exchange by the Asahi process,* Degrémont brochure 1297F, 1969.
39. K. H. Walter, M. Böse, K. Marquardt: "Kontinuierlich arbeitende Vollentsalzungsanlage," *Energie (D)* 10 Oct (1975) 245.
40. I. R. Higgins, R. S. Chopra: *Symposium of Ion Exchange in the Process Industries,* Soc. Chem. Ind. London 1969.
41. F. de Dardel: "Résines échangeuses d'ions et eau potable," *L'Eau, I'Industrie, les Nuisances Paris* **112** (1987) Sept 45.
42. Code of Federal Regulations, *Food and Drugs,* National Archives of the United States, vol. 21, Sect. 173.25, 1981.
43. Ministère des affaires sociales. Circulaire du 27 mai 1987 relative à l'emploi des résines échangeuses de cations pour le traitement des eaux destinées à la consommation humaine. ASE 87/34, 1987 69.
44. Empfehlung XXIV, "*Ionenaustauscher und sorptiv wirkende Polymere"* des Bundesgesundheitsamtes. 168. Mitteilung, Bundesgesundheitsblatt 28, 1985.
45. G. Solt: "Nitrate removal: a compromise solution," *Water Quality International* no. 1, 1987, p 29.
46. J. P. van der Hoek, A. Klapwijk: "Nitraatverwijdering Grondwater," *PT Procestechniek Netherlands* **40** (1985) no. 5, 22.
47. K. Hagen, W. Höll, W. Kretzschmar: "The Carix process for removing nitrate sulphate and hardness from water," *Aqua* **5** (1986) 275.
48. R. Kunin, B. Vassiliou: "Novel deionisation technology based on the use of weak electrolyte ion exchange resins," *Ind. Eng. Chem. Process Des. Dev.* **3** (1964) 404.
49. G. Boari, L. Liberti, R. Passino, *Ion Exch. Membr.* **2** (1975) 127.
50. D. E. Weiss et al.: "Thermal regeneration of ion exchange resins," *Austral. J. Chem.* **19** (1966) 561.
51. F. de Dardel: "Traitement de condensats et préparation d'eau ultrapure," *L'Eau et l'Industrie Paris* **53** (1981) 67.
52. I. M. Abrams: "New requirements for ion exchange in condensate polishing," *Proceedings of the 37th International Water Conference,* Pittsburgh, Pennsylvania, 1976.

53. C. W. Wages, C. W. Smith: "Operating experience of a deep-bed condensate polishing system," *Proceedings of the 38th International Water Conference,* Pittsburgh, Pennsylvania, 1977.
54. C. M. Kennedy, P. S. Meyers, G. J. Crits: "Experience with high rate ammoniated mixed beds for condensate polishing at Cipsco, Coffeen station," *Proceedings of the 28th International Water Conference,* Pittsburgh, Pennsylvania, 1967.
55. P. W. Renouf, J. H. Smith: "Operating experience with a Tripol ammonia cycle condensate polisher," *Proceedings of the 42nd International Water Conference,* Pittsburgh, Pennsylvania, 1981.
56. F. Martinola, US 4 400 278, 1981.
57. C. S. Parks: "Operating experience with two new filter precoat materials," *Proceedings of the 39th International Water Conference,* Pittsburgh, Pennsylvania, 1978.
58. G. Fené, F. de Dardel: "Technologie des microrésines prémélangées. Utilisation en traitement de condensats de centrales nucléaires," *L'Eau, I'Industrie, les Nuisances Paris* **72** (1983), 38.
59. H. G. Schneider, J. Mikule: "The recovery of sugar from beet molasses by the Pfeifer und Langen exclusion process," *18th meeting of the American Soc. of Sugar Beet Technologists,* San Diego 1974.
60. Roquette Frères. Brev. Franç., FR 79 10563, 1979 (F. Devos et al.).
61. M. J. Hatch, J. A. Dillon: "Acid retardation. A simple physical method for separation of strong acids from their salts," *I&EC Proc. Des. and Eq.* **2/4** (1963) 253.
62. Product profile: Duolite MB 1243. Effluent and Water Treatment J. May 1980.
63. A. R. Pitochelli: *Ion Exchange Catalysis and Matrix Effects,* Rohm and Haas Brochure, IE-214, 1988.
64. A. Abadie: *Etude de quelques perturbations des transferts d'ions. Thèse de doctorat.* Université Paul Sabatier, Toulouse, France. 1976.
65. *Oily Effluent Purification by Coalescence,* Elf-Aquitaine, Paris 1979.
66. F. de Dardel, "Some Principles of Ion Exchange Plant Design," *Ultrapure Water* **16** (1999) no. 8, 26.
67. R. C. Hunt, "Conductivity and pH Measurements in High Purity Water," *Ultrapure Water* **3** (1986) no. 5, 50.
68. R. N. R. Robinson, L. A. Chapple: "Operation of World's Largest Tripol Condensate Polishing Vessels at Stanwell Power Station, Queensland (Australia)," *Ion Exchange Developments and Applications*, proceedings of IEX '96, SCI, Cambridge, p. 27.
69. J. M. Ragosta: "Improvements in Powdered Ion Exchange Technology for Industrial Applications," *Ion Exchange Developments and Applications*, proceedings of IEX '96, SCI, Cambridge, p. 75.
70. J. Keary, A. D. Mortimer: "Ultrapure Water for the Manufacture of Semiconductor Devices", *Ion Exchange Developments and Applications*, proceedings of IEX '96, SCI, Cambridge, p. 35.
71. S. Weiss,: "ISEP Continuous Contactor for Continuous Ion Exchange," *Ion Exchange Advances*, proceedings of IEX '92, SCI, Elsevier Science Publishers, p. 256.

Ion Implantation → Thin Films
Ion Plating → Thin Films
Ionomers → Specialty Plastics
Iridium → Platinum Group Metals and Compounds

Iron

Franz Oeters, Berlin, Germany (Chap. 1, 3, 2.2, 2.2 and 2.4)

Manfred Ottow, Berlin, Germany (Chap. 1, 3, 2.2, 2.3 and 2.4)

Heinrich Meiler, Lurgi GmbH, Frankfurt, Germany (Chap. 2.1.1)

Hans Bodo Lüngen, VDEh, Düsseldorf, Germany (Chap. 2.1.2 and 2.5)

Manfred Koltermann, formerly Hoesch Stahl AG, Dortmund, Germany (Chap. 2.1.3)

Andreas Buhr, Almatis GmbH, Frankfurt, Germany (Chap. 2.1.3)

Jun-ichiro Yagi, Research Institute of Mineral Dressing and Metallurgy, Tohoku University, Sendai, Japan (Chap. 2.5)

Lothar Formanek, Lurgi GmbH, Frankfurt, Germany (Chap. 2.6)

Fritz Rose, Lurgi GmbH, Frankfurt, Germany (Chap. 2.6)

Jürgen Flickenschild, Deutsche Voest-Alpine Industrieanlagen GmbH, Düsseldorf, Germany (Chap. 2.7)

Rolf Hauk, Deutsche Voest-Alpine Industrieanlagen GmbH, Düsseldorf, Germany (Chap. 2.7)

Rolf Steffen, VDEh, Düsseldorf, Germany (Chap. 2.7)

Reiner Skroch, Lurgi GmbH, Frankfurt, Germany (Chap. 4)

Gernot Mayer-Schwinning, Lurgi GmbH, Frankfurt, Germany (Chap. 4)

Heinz-Lothar Bünnagel, Wirtschaftsvereinigung Stahl, Düsseldorf, Germany (Chap. 5)

Hans-Georg Hoff, Wirtschaftsvereinigung Stahl, Düsseldorf, Germany (Chap. 5)

1. Properties 2
2. Production 4
2.1. Raw and Auxiliary Materials . 5
2.1.1. Ores 5
2.1.1.1. Ore Deposits 6
2.1.1.2. Supply of Iron Ore 7
2.1.1.3. Most Important Iron-Producing Countries 8
2.1.1.4. Beneficiation of Iron Ore 10
2.1.1.5. Agglomeration 15
2.1.2. Reducing Agents 21
2.1.3. Refractory Materials 25
2.2. Thermodynamic Principles . . . 27
2.2.1. Thermodynamic Principles of the Reduction Reactions 27
2.2.2. Mass and Heat Exchanges of the Reduction Reaction with Carbon as the Primary Fuel 35
2.2.3. Thermodynamic Principles of Metal – Slag Reactions 38
2.3. Kinetic Principles 42
2.3.1. Reduction in the Solid State . . . 42
2.3.1.1. Time Dependence of Individual Processes During Reduction . . . 43
2.3.1.2. Rate Laws of Complex Reduction Reactions 50
2.3.2. Reduction in the Liquid State . . 57
2.3.2.1. General 57
2.3.2.2. Microkinetics 58
2.3.2.2.1. Numerical Data for the Calculation of Microkinetics . . . 60
2.3.2.2.2. Macrokinetics 61
2.3.3. Kinetics of Post-Combustion Heat Transfer 62
2.4. Classification of Reduction Processes 66
2.4.1. Limiting Conditions for the Reduction in Shaft Furnaces . . . 67
2.4.2. Limiting Conditions for the Reduction in Rotary Kilns 70
2.4.3. Limiting Conditions for the Reduction in Fluidized Beds . . . 73
2.4.4. Limiting Conditions for Smelting Reduction Processes 74
2.5. Blast Furnace Process 76
2.5.1. Outline of a Blast Furnace Construction 81
2.5.2. Metallurgy of the Blast Furnace . 91
2.5.3. Effective Utilization of Energy . . 94

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_461.pub2

2.5.4. Blast Furnace Productivity Criteria 107
2.5.5. Use of Blast Furnace Products . . 109
2.5.6. Process Control 111
2.5.7. Hot-Metal Desulfurization 120
2.6. Direct Reduction Processes . . . 126
2.6.1. Shaft Furnace Processes for Direct Reduction 127
2.6.2. Retort Processes 133
2.6.3. Fluidized-Bed Processes 134
2.6.4. Rotary Kiln Processes 137
2.7. Smelting Reduction Processes . 141
2.7.1. Characteristics of Coals for the Smelting Reduction Processes . . 142
2.7.2. Classification of Smelting Reduction Processes and Earlier Developments 144
2.7.2.1. Processes Using Electrical Energy 144
2.7.2.2. Coal – Coke-Bed Melter – Gasifiers 145
2.7.2.3. Converter-Type Iron Bath Reactors 148
2.7.3. State of the Art of Process Developments 150
2.7.3.1. Melter – Gasifiers 150
2.7.3.2. Iron Bath Reactors 151
2.7.3.3. Rotary Hearth Direct Reduction – Submerged Arc Furnace Combinations 154
2.7.3.4. Special Developments for Processing of Steelworks Dusts . 155
2.7.4. Conclusions 156
3. Production of Pure and High-Purity Iron 156
4. Aspects of Environmental Protection 159
4.1. Air Pollution Control 160
4.2. Prevention of Water Pollution . 163
4.3. Noise Reduction 163
4.4. Waste Management 164
5. Economic Aspects 164
6. Acknowledgement 170
7. References 170

Abstract

The article first evaluates the general features of iron and its naturally occurring oxides. The production processes from ores are described and the producer countries listed. Bedded deposits enriched in iron account for ca. nine-tenths of potential iron reserves. Benefication steps include crushing, concentration, flotation, and/or roasting. The blast furnace is the dominant type of equipment for the production of iron metal. Part of the heat needed for the production of iron is used to heat up the raw materials to be processed and part of it is required for endothermic chemical reactions. Carbon and hydrocarbon carriers, chiefly in the form of coke, are gasified to the primary reducing agents (CO, H_2) in blast furnaces. The use of fossil reducing agents in iron making is subject to careful re-evaluation, and development of new reduction techniques is expected. Gaseous reducing agents other than those formed in the reduction reactor itself are derived from natural gas, oil, or coal. It is quite likely that coal will come to play an increasingly important role as a source for production of reducing gas. Gas generation on the basis of nuclear energy is expected to remain economic for many years, although some effort is currently being directed toward the use of a plasma arc for converting carbon sources into reducing gas.

The world steel industry is the principal consumer of refractories. The refractory-unit consumption for steel production has constantly decreased over the last decades due to new steelmaking technologies. The trend towards high-performance refractories is ongoing due to increasingly severe process conditions. The refractory materials used for the blast-furnace process and the hot-metal transport are discussed.

Finally, the thermodynamic and kinetic principles of ore reduction are evaluated in detail. This will provide an understanding of the process technology as well as a critical evaluation of the process with respect to cost effectiveness, product quality, and efficient utilization of the reducing agents.

1. Properties

Physical Properties. The physical properties of iron are strongly dependent on its degree of purity. In the evaluation of physical property data from the literature it is, therefore, necessary to know and to take into account the amount and the nature of the impurities. The following distinctions are made:

technically pure iron, containing 99.5 – 99.9 % Fe;
pure iron, containing 99.9 – 99.99 % Fe;
high-purity iron, containing 99.99 % Fe.

Frequently used physical material constants are summarized in Tables 1 and 2. The majority of the data pertains to technically pure iron; however, some data valid for pure iron are also included. For cases where the required accuracy is particularly high the original reference should be consulted [1 – 6].

Table 1. Crystallographic, thermal and mechanical properties of technically pure iron

Crystal structures and lattice constant a	
α-Fe body centered cubic	a = 0.2866 nm (25 °C)
γ-Fe face centered cubic	a = 0.3647 nm (910 °C)
δ-Fe body centered cubic	a = 0.2932 nm (1400 °C)
Transition points and enthalpies	
Curie point	760 °C; ΔH = 2.76 kJ/mol
α-Fe ↔ γ-Fe	910 °C; ΔH = 0.92 kJ/mol
γ-Fe ↔ δ-Fe	1400 °C; ΔH = 1.17 kJ/mol
δ-Fe ↔ melt	1539 °C; ΔH = 15.28 kJ/mol
Melt ↔ vapor	2880 °C; ΔH = 340.4 kJ/mol
Standard entropy	27.2 J/mol K
Molar heats as a function of temperature	
α-Fe (273 < T < 1033)	C_p = 17.50 + 24.77 T J mol^{-1} K^{-1}
α-Fe (1033 < T < 1183)	C_p = 37.68 T J mol^{-1} K^{-1}
γ-Fe (1183 < T < 1673)	C_p = 7.70 + 19.51 T J mol^{-1} K^{-1}
δ-Fe (1973 < T < 1812)	C_p = 43.96 T J mol^{-1} K^{-1}
Fe_L (1812 < T < 1873)	C_p = 41.87 T J mol^{-1} K^{-1}
C_p (0 – 1500 °C)	0.681 J g^{-1} K^{-1}
Thermal conductivity at 100 °C	72.44 W m^{-1} K^{-1}
Density	
at 20 °C	7.87 g/cm^3
at 1550 °C (liquid)	7.08 g/cm^3
Coefficient of thermal expansion (0 – 800 °C)	14.5×10^{-6} K^{-1}
Mechanical properties (measured for normalized samples containing 0.005 % C, 0.005 % O_2 and 4 to 24 grains per mm^2)	
Brinell hardness	44 – 55 HB
Lower yield point	90 – 140 N/mm^2
Tensile strength	220 – 280 N/mm^2
Elongation at fracture (L_0= 5 d_0)	40 – 60 %
Modulus of elasticity	210×10^3 N/mm^2
Shear modulus	82×10^3 N/mm^2
Vapor pressure (1812 K < T < 3153 K); p (mbar): $\log p = -19710\, T^{-1}\ 1.27 \log T + 13.39$	
Surface tension (1550 °C)	1.78 N/m
Kinematic viscosity (approximately linear relationship between those points)	
at 1450 °C (undercooled liquid)	6.7 mPa s
at 1750 °C	4.4 mPa s

Chemical Properties [7, 8]. Iron [*7439-89-6*], Fe, atomic number 26, A_r 55.847, belongs to group 8 of the periodic table and to the first row of the transition elements. The electron configuration of the lowest energy state is characterized by $3d^6\ 4s^2$. Although the energies corresponding to the 3*d* and 4*s* orbital functions of the neutral atom are similar, the *d* orbital function is more stable than the *s* function whenever the atom acquires a charge. Thus, in ionic compounds and in complexes with non-π-bonding ligands iron acquires an oxidation state of 2 or higher. The highest possible oxidation state of 8 (the sum of the 4*s* and 3*d* electrons) on the other hand, is not observed.

The acid character of iron oxides increases with increasing oxidation state. The bond in iron halides is more covalent the higher the oxidation state of iron. Accordingly, the hydrolysis increases with increasing valency of iron.

At room temperature metallic iron is stable in dry air and in water which does not contain carbon dioxide and air. Although the equilibrium of the oxidation reaction of iron in air is shifted toward the formation of the oxide, the rate of the reaction is drastically decreased by the formation of a protective layer of iron oxide. In fact, the reaction is virtually entirely suppressed. The so-called passivation of iron by concentrated nitric acid and strong oxidizers, e.g., dichromate containing acids, is likewise caused by the formation of a protective oxide layer. Iron is not dissolved by these acids because of the formation of a dense oxide layer, although it is readily dissolved in dilute mineral acids in accordance with its position in the galvanic series (E_0 = − 0.409 V). Iron is dissolved by hot concentrated sodium hydroxide and precipitates as iron hydroxide, $Fe_2O_3 \cdot nH_2O$, in the presence of oxygen.

In the presence of air-containing water and moist air iron is transformed into iron (III) oxyhydrate (rust) (→ Corrosion). The layer that is formed on the iron during this process is not dense but flakes off, in contrast to the aforementioned oxide layer, and a fresh iron surface is therefore exposed to corrosion.

The oxides of iron (→ Pigments, Inorganic) are very important for the production of iron. There are three oxides of iron: hematite (Fe_2O_3), magnetite (Fe_3O_4), and

Table 2. Electrical and magnetic properties of technically pure iron

Temperature θ, °C	20	200	400	600	800	1000
Specific electrical resistivity ϱ, μ Ω cm	10	20	40	70	108	115
Degree of purity	technically pure		pure		ultra pure	
Coercive force, A/m	80		8 – 40		1.3	
Remanence, T	1.1		0.8		ca. 0.085	
Saturation polarization $4\pi J_\infty$, T	0.5 – 1		1 – 2		8.8 – 28	
Hysteresis loss, mJ/cm^3	0.5		0.1		0.002 – 0.01	

wustite ($Fe_{1-y}O$), with $y = 0.045 - 0.135$. The oxygen ions form an almost dense close-packed structures in all three of the oxides.

Hematite [*1309-37-1*], the oxide of trivalent iron, exhibits two commercially important modifications. The stable form, α-hematite, has the crystal structure of corundum. The oxygen ions form a hexagonal close-packed structure and the iron (III) ions occupy 2/3 of the octahedral interstices. In γ-hematite the oxygen ions form a cubic close-packed structure. The iron ions occupy 1/8 of the tetrahedral and 5/12 of the octahedral interstices. When γ-hematite is heated to a temperature of ca. 500 °C, it is irreversibly transformed into α-hematite.

In *magnetite* [*1317-61-9*] 1/3 of the iron ions are divalent, 2/3 trivalent. As in the case of γ-hematite, the oxygen ions form a cubic close-packed structure. The distribution of the cations over the interstices corresponds to that of the inverse spinel structure, i.e., half of the iron (III) ions occupy 1/8 of the tetrahedral interstices and the other half of the iron (III) ions, together with the iron (II) ions, occupy 6/12 of the octahedral interstices. The occupation of the interstices differs only slightly from that in γ-hematite, corresponding to the difference in the iron: oxygen ratio. The properties of both oxides are similar, as a result of the similar crystal structure. They are both ferrimagnetic, optically isotropic and have good electric conductivities compared to α-hematite.

Wustite [*1345-25-1*] is the oxide of divalent iron; however, it should be noted that the iron: oxygen ratio does not correspond to the stoichiometric value of unity. The oxygen ions form a cubic close-packed sublattice. The iron ions occupy the octahedral interstices but leave a certain number unoccupied. The overall lattice, therefore, corresponds to a defect NaCl type lattice. Wustite is electrically neutral because a portion of the Fe^{2+} ions has been replaced by Fe^{3+} ions. Wustite is thermodynamically stable at temperatures > 571 °C [9]. Below this temperature wustite disproportionates into magnetite and iron (eutectoid reaction). The rate of this disproportionation decreases rapidly with decreasing temperature. At room temperature wustite is, therefore, stable for an unlimited period of time.

(Commercially important iron compounds → Iron Compounds.)

2. Production

In the terminology a distinction is made between *primary reduction* (iron production), which is the subject of the present chapter, and *refining* (→ Steel). The production of iron can take place in the *blast furnace* which produces pig iron, or by means of other processes which yield equivalents such as sponge iron. Steel production refers to the processing step that involves the refining and the adjustment of certain solute contents in the iron (steel), irrespective of whether the raw material is crude iron, scrap, or sponge iron. This separation of processes is useful because the properties of steel (hot formability, strength) are acquired during the process of steelmaking and steel differs from raw materials such as pig iron and sponge iron with respect to these properties.

The most important metallurgical processing units for the manufacture of large quantities of liquid crude steel from ore are:

- for ore preparation
  - sintering plants
- for the reduction of oxide ores
  - blast furnaces
- for refining and alloying of steel and for scrap melting
  - converters operated with pure oxygen
  - open hearth furnaces
  - electric arc furnaces

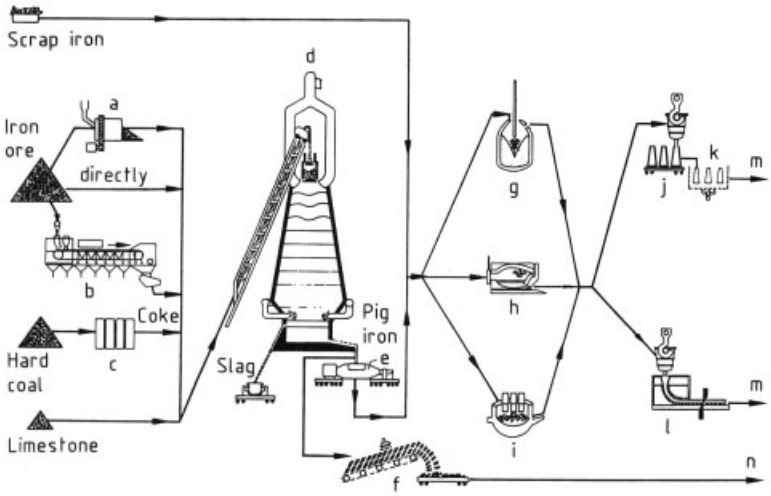

Figure 1. Schematic flow sheet of steel production
a) Pelletizing drum; b) Sintering grate; c) Coking plant; d) Blast furnace; e) Torpedo ladle; f) Pig-casting plant; g) Basic oxygen converter; h) Siemens – Martin furnace; i) Electric arc furnace; j) Continuous casting; k) Soak pit; l) Ingot casting; m) To rolling mill; n) To shipping

The majority of the crude steel is processed to semifinished products by ingot casting or, increasingly, by continuous casting. A schematic representation of this important processing route from iron to steel is shown in Figure 1.

In addition to the blast furnace process, oxygen converter, open hearth furnace, and electric arc furnace, there are some other steelmaking processes. Although these are less important in terms of the relative quantity of steel produced, their production in absolute terms is significant. The quantity of raw steel produced annually in these units (30 Mt/a) is equivalent to the annual production of all other metals combined.

2.1. Raw and Auxiliary Materials

The *raw materials* required for iron production are:

iron ore and scrap iron
solid, liquid, or gaseous reducing agents
solid, liquid, or gaseous fuels to supply the heat needed for reaction and melting.

In the theoretical analysis of the reactions, certain *auxiliary materials* appear to a limited extent or not at all, even though they are crucial from the process-engineering standpoint:

slag-forming agents such as silica and silicates; sources of calcium and magnesium such as limestone, dolomite, olivine; and coke – ash constituents;
refractory materials such as bricks made of silica, magnesite, dolomite chromite, or carbon.

2.1.1. Ores

Iron is the fourth most abundant element in the Earth's crust (4.7 wt %). In relation to the total mass of the Earth, iron is probably the most abundant element.

Iron is found in high concentration in ore deposits, where it occurs mainly as oxide. Table 3 lists the iron minerals of greatest industrial importance (along with some accompanying gangue minerals) and their relevant properties.

The only minerals of worldwide importance are hematite (specularite, Fe_2O_3), magnetite (Fe_3O_4) and limonite (FeOOH). Siderite ($FeCO_3$) finds limited use on a local basis. Other ores such as chamosite (an iron magnesium aluminosilicate) or pyrite are virtually not important for iron production any more.

Table 3. Important minerals in iron ores

Mineral	Chemical formula	Density, g/cm^3	Hardness (Mohs)	Specific susceptibility (order of magnitude)
Iron (for comparison)	Fe	7.88	4 – 5	
Magnetite	Fe_3O_4	5.2	5.5	10^{-1} – 1
Specularite ("hematite")	Fe_2O_3	5.2 – 5.3	6.5	10^{-4}
Limonite				
Needle ore	α-FeOOH	4.3	5 – 5.5	10^{-6} – 10^{-5}
Ruby mica	β-FeOOH	4.0	5	10^{-6} – 10^{-5}
Siderite	$FeCO_3$	3.7 – 3.9	4 – 4.5	10^{-6}
Pyrite	FeS_2	5.0 – 5.2	6 – 6.5	10^{-6}
Phyrrhotite	FeS	4.6	4	10^{-3} – 10^{-4}
Chalcopyrite	$CuFeS_2$	4.1 – 4.3	3.5 – 4	10^{-5} – 10^{-6}
Apatite	$Ca_5F(PO_4)_3$	3.2	5	-10^{-6}
Vivianite	$Fe_3(PO_4)_2 \cdot 8\,H_2O$	2.6 – 2.77	3	
Quartz	SiO_2	2.65	7	-10^{-6}
Orthoclase	$KAlSi_3O_8$	2.55	6	10^{-6}
Plagioclase	$NaAlSi_3O_8/CaAl_2Si_2O_8$	2.6 – 2.8	6 – 6.5	10^{-6}
Kaolinite, dickite	$Al_4(OH)_8Si_4O_{10}$	2.6	1	
Muskovite	$KAl_2(OH,F)_2\,[AlSi_3O_{10}]$	2.6 – 2.8	2 – 2.5	10^{-2} – 10^{-5}
Pyrolusite	MnO_2	5	1 – 6	10^{-6}
Calcite	$CaCO_3$	2.6 – 2.8	3	-10^{-6}
Ilmenite	$FeTiO_3$	4.5 – 5	5 – 6	10^{-4}

2.1.1.1. Ore Deposits

Of the variety of classification systems that have been proposed for iron-ore deposits, the one from the United Nations Survey is used here [10]. The criterion is the physical appearance of the deposit; the main types are bedded (A), massive (B), residual (C), byproduct (D) and other (E).

A Bedded Deposits
A-1 *Iron Formation*
Lake Superior Type
Chiefly Precambrian; primarily as sedimentary deposits, heavily metamorphized (itabirite, taconite, jaspilite or quartz-banded ore); enriched by weathering processes (Mesabi, Minas Gerais, Carajas, Venezuela, Labrador and Quebec, Krivoi Rog)
Algoma Type
Primarily as thinly-banded quartzite with interlayers of iron ore; enriched as limonite – hematite weathering ore
A-2 *Ironstone Formation*
Mainly marine sedimentary deposits of minette type; Clinton, Wabana; chiefly mesozoic, with limonite as the most important mineral
A-3 *Other Iron-Ore Sediments*
Examples: clastic deposits, ferruginous sandstone, and iron shale; also alluvial deposits and unconsolidated clastic sediments

B Massive Deposits
B-1 *Bilbao Type*
Deposits from weathering of siderite rock, with limonite and hematite in the weathering zone
B-2 *Magnitnaya Type*
Contact-metasomatic replacement deposits
B-3 *Kiruna Type*
Magnetite intrusion, usually with apatite
B-4 *Taberg Type*
Syngenetic titanomagnetite bodies and deposits of finely-intergrown ilmenite ores

C Residual Deposits
C-1 Laterite and iron-ore caps as a result of weathering of underlying iron-bearing rocks
C-2 Fluvial deposits and bog ore
C-3 Other residual ores, e.g., caps of sulfide ore deposits (gossan, "iron hat"), unconsolidated clastic deposits

D Byproducts
Recoverable values in byproduct iron oxide

E Other Types of Deposits
e.g., lode-type deposits (Siegerland)

Bedded deposits are stratigraphic members enriched in iron by a factor of ca. 4 – 12 above average. They occur in all geologic ages; including ores which originate from them, they account for ca. nine-tenths of potential iron reserves. Names applied to Precambrian metamorphic ore beds are itabirite, taconite, Lake Superior type, quartz-banded ore, jaspilite, magnetite quartzite, hematite quartzite, etc. More recent formations are known as minette (Jurassic), Salzgitter type, and other local names. In general, the term bedded deposits is used for iron-mineral enrichments in which ore minerals are more or less closely laminated with quartz, jasper or carbonate rock and the iron-bearing sediments lie conformally to the under- and overlying beds of igneous or metasedimentary rock. Thus, iron formations can be chemical rocks, clastic sediments, or rock beds that have been replaced by iron mineral during diagenesis or in their subsequent history.

Massive iron-ore deposits are ore bodies of irregular shape, discordantly embedded in the enclosing rock. They include (1) replacement deposits in carbonate rocks (Bilbao type) with siderite, hematite or limonite (the last two in the weathering zone above the water table); (2) contact-metasomatic deposits in the contact region of acidic intrusive rocks with magnetite and, to a lesser extent, hematite, siderite, pyrite, pyrrhotite and copper sulfide (Magnitnaya type); (3) magnetite intrusions in acidic magmatic rock (Kiruna type); and (4) massive or intergrown titanomagnetite or ilmenite concentrations in basic rock (Taberg type).

The most important type of *residual deposit* is laterite, which occurs as a cap overlying a wide variety of rocks. It is the product of weathering in tropical and subtropical climatic regions. The iron content of these ores is usually low (limonite, hematite), the alumina content high, and undesirable accessory constituents such as chromium, phosphorus, and nickel are frequently present (e.g., at Vogelsberg, Germany). Thus, laterite has largely ceased to be an important iron ore except where the lateritic caps of other iron ore deposits have secondary enrichments of up to 69 % iron. On the other hand, nickel-bearing laterites are the most important nickel ores of the future (Cuba, New Caledonia, the Philippines). Extended fluvial deposits, such as the Robe River deposit in Western Australia, should also be mentioned. Bog ores, in contrast, form deposits of small extent that were formerly locally important. They are precipitation products from iron-containing solutions. The weathering cap of sulfide deposits (gossan or "iron hat"), which are usually separated from the primary sulfides by an indistinct boundary, are of some importance.

2.1.1.2. Supply of Iron Ore

Until about the 1950s, ironworks were mainly supplied from their own deposits; ore and concentrate were transported over short distances. Since that time a fundamental change has occurred as a result of (1) the greater demand because of increased production, (2) the drop in the cost of overseas transport because of larger ship capacities, and (3) the need for higher productivity because of continually rising labor and energy costs. The iron ore demand of the most important steel-producing regions is essentially met by a few large ore-producing regions.

Medium-sized and small ironworks located at deposit are now found only in case of large deposits (northeast India, the Ukraine and other areas of the Former Soviet Union, northeast China). Table 4, listing world iron-ore reserves, also shows, from the tonnage standpoint, that only a few countries play important roles in ore supply. The short descriptions below are limited to the most important deposit areas and supplier countries.

Table 4. World iron-ore reserves [11]

	Reserves, $t \times 10^6$	Iron content, %
Former Soviet Union	110 750	25.4
China	42 000	30.0
Brazil	34 540	56.9
Canada	26 417	31.6
United States	25 400	20.7
Australia	17 781	60.0
India	13 500	61.5
South Africa	6 300	59.1
France	4064	40.2
Sweden	3 353	59.1
Venezuela	2 337	54.4
Liberia	1 668	39.5
Other	2 235	28.4
World total	290 235	35.4

Production figures for iron ore, ore concentrate, and ore agglomerate appear in Table 5.

Table 5. World iron-ore production, ore, concentrate and agglomerate (1975 – 1993) (t × 10^6) [12, 13]

	1975	1980	1986	1993
North America	132.0	127.1	82.8	96.0
Canada	46.9	48.8	36.1	32.3
Mexico	5.1	7.6	7.3	8.0
United States	80.1	70.7	39.4	35.7
South America	134.3	146.1	164.7	189.8
Brazil	89.9	114.7	132.0	169.4
Chile	11.0	8.6	7.0	7.0
Peru	7.8	5.7	5.0	5.2
Venezuela	24.8	16.1	19.1	15.5
Europe	355.5	335.3	313.6	184.2
Austria	3.8	3.2	3.1	1.4
France	49.6	29.0	12.4	3.5
Germany	3.3	1.9	0.7	0.1
Norway	4.1	3.9	3.7	2.2
Spain	7.6	9.2	6.1	2.5
Sweden	30.9	27.2	20.5	18.7
Former Soviet Union	232.8	244.7	249.9	154.0
United Kingdom	4.5	0.9	0.3	
Former Yugoslavia	5.2	4.5	6.7	0.3
Asia	122.7	122.8	151.9	303.8
China	65.0	68.1	90.0	224.7
India	41.4	41.9	47.8	56.0
North Korea	9.4	8.0	8.0	10.0
Turkey	2.4	2.6	4.0	5.1
Africa	57.7	60.8	55.8	44.2
Algeria	3.2	3.5	3.4	2.6
Liberia	24.0	18.2	15.3	
Mauretania	8.7	8.9	8.9	9.2
South Africa	12.3	26.3	24.5	29.4
Oceania	99.9	99.2	92.5	123.7
Australia	97.6	95.5	90.0	121.4
New Zealand	2.3	3.6	2.4	2.3
World total	902.0	891.3	861.3	941.7

It is striking that the most important supply sources are type A deposits (bedded), and among these the metamorphic deposits of Itabira type (under that or some other local name), chiefly of Precambrian age.

Especially important are the enrichment or weathering zones of these deposits, where the mobilization of silica and alumina has produced rich ores that can be forwarded to ironworks without preliminary beneficiation. Table 5 also shows that ores expensive to produce and those whose beneficiation does not yield concentrates with at least ca. 60 wt % iron have lost much of their importance. In fact, most of them will no longer be in production after a few years. Examples are the Mesozoic deposits in France, Belgium and Luxembourg (minettes), in the United Kingdom (home ore) and in the Federal Republic of Germany (Salzgitter, Siegerland, Lahn – Dill, Oberpfalz).

The important producer countries rank as follows by tonnage output (1993 production, 10^6 t):

1)	China	224.7
2)	Brazil	159.4
3)	Former Soviet Union	154.0
4)	Australia	121.4
5)	India	56.0
6)	United States	55.7
7)	Canada	32.3
8)	South Africa	29.4
9)	Sweden	18.7
10)	Venezuela	17.5

2.1.1.3. Most Important Iron-Producing Countries [10 – 12, 14, 15]

Brazil. Brazil is currently (ca. 2000) the biggest producer of ore for *overseas shipment.* Two ore districts are now in production. The "iron quadrangle" ("Quadrilatero ferrifero") in Minas Gerais state is the more important. It sup-

plies several local ironworks, but most of the product is exported.

The ores occur in three successive beds, Precambrian in age and strongly metamorphized, called (from shallowest to deepest) Rio das Velhas, Minas, and Italcolomi [10]. The ore is thought to consist of marine sedimentary deposits. The ore beds contain 35 wt % – 60 wt % iron.

Various ore types are distinguished:

(1) hematite enrichments (lump to powder ore), > 64 wt % iron; (2) silica-rich hematite ore, 60 – 64 wt % iron; (3) itabirite, 35 – 60 wt % iron, occurring in compact, soft and powder forms and typical of the Minas series; and (4) Canga, (60 – 68 wt % iron), a conglomerate-like ore that often consists of rounded fragments in a limonite matrix. Capacity is ca. 130×10^6 t/a of ore shipped; about a tenth is shipped as lump ore, 24×10^6 t/a as pellets, and the remainder as sinter fine ore ("sinter feed").

The second large ore district is Carajas, with a production target of 35×10^6 t/a, 10 – 20 % of this being lump ore. The Carajas ore also lies within Precambrian beds and has been enriched by secondary processes to 60 – 67 wt % iron.

Venezuela [10, 15]. The most important iron-ore district is the Imataca belt, including the well-known Cerro Bolivar, El Pao, and San Isidro deposits. The belt is part of the Precambrian Guyana Shield. The ores occur in a sequence of metamorphic rocks with gneiss, shales, and taconite-like iron-banded quartzite. Secondary processes have created rich ore deposits, partly through the precipitation of iron hydroxide in cavities. At the surface, a hematite-rich laterite cap, 1 – 50 m thick has formed; this is still the main source of ore. In 1987, ore output was 17.2×10^6 t, of which 5.5×10^6 t was consumed domestically.

Lump ore, sintered fine ore and pellet feed containing 62 – 67 wt % iron, 0.6 – 6 wt % silica, 1–2 wt % alumina and 2 – 5 wt % loss on ignition are produced. To meet the needs of direct reduction units (Midrex, HyL, see Section 2.6) at the Matanza ironworks, a 6 600 000 t/a pellet plant has been erected.

Canada. The Labrador geosyncline is the most important ore district. Precambrian deposits of the Lake Superior type are regarded genetically as continental shelf deposits that have undergone various degrees of metamorphosis. There are weathered rich ores containing over 60 % Fe and having high limonite contents; weathered, relatively coarse-grained magnetite hematite ores containing 20 – 50 % quartz; and unweathered banded ore.

United States—Lake Superior. Most of the known iron-ore reserves in the United States are found in the Lake Superior area, in Minnesota, Wisconsin, and Michigan. About three-quarters of U.S. ore production comes from this region; the output goes to ironworks south of the Great Lakes (Ohio, Pennsylvania). The deposits occur in Lower and Middle Precambrian strata. Iron oxide, carbonate, silicate and sulfides are thinly banded with gel quartz and other gangue minerals. The thickness is between 15 and 300 m. Fine-grained magnetite and hematite are the principal ore minerals.

In weathering zones locally enriched lump ores exist containing ca. 60 % Fe; these consist of hematite and limonite. Mining is limited to these rich ores and magnetite-bearing beds, which are relatively easy to beneficiate by magnetic separation.

Australia. The large iron-ore reserves of this continent are concentrated in the north of Western Australia, in the Pilbara district. The Precambrian banded itabirite ores contain magnetite, quartz, carbonate, and stilpnomelane. The near-surface region includes hematite and limonite weathering ores. Besides these in situ enrichments, conglomerate and residual ores are also found; the latter have a high limonite content and occur chiefly in old Tertiary riverbeds.

India. In India the most important ore deposits are enclosed in the strata of a Precambrian geosyncline, which runs parallel to the east coast over a significant part of the subcontinent. The ore is mainly of itabirite type; lateritic weathering has produced rich ore with a high alumina content in the gangue, which extends to a considerable depth. Locally, quartzbanded lean ore and similar primary ores have also been exploited. Hematite ore containing > 60 wt % iron is worked; it is partly lump ore and partly sinter fine ore. Hematite – limonite ore also occurs. The lateritic hard-ore cap is often underlain by

blue dust. This ore consists of medium- to fine-grained primary specularite, martite and, locally, primary magnetite as weathering residue, and contains only small amounts of quartz, kaoline minerals, and gibbsite as gangue.

China. China has many isolated ore deposits, most of which form the basis for local iron and steel industries. The largest concentration of reserves, however, is in the provinces of Liaoning (northeast China) and Hebei (eastern central China). The ores are of itabirite type, mainly with hematite – quartzite ores containing 50 wt % iron. Large rich ore reserves are obviously nonexisting in China. Of the ores now being mined, it is known that beneficiation is often difficult and costly. Exceptions are the various magnetite deposits of Magnitnaya type. Some ores contain undesirable secondary constituents such as tin, fluorspar and finely intergrown apatite. Despite the potential reserves, cost considerations may cause China to import iron-rich ore; the recently observed construction of steelworks at coastal sites is an indication of this policy.

Former Soviet Union. The Former Soviet Union has more than 20 iron-ore regions. The most important reserves are in the regions of Kursk, western Siberia, and the Ukrainian crystalline shield. The first two areas account for ca. 30 % each, the last for some 10 % of the Former Soviet Union's total reserves [10]. The most important production region is the Ukrainian Shield, with the Krivoi Rog syncline as its center ("Krivbas"). The Precambrian strata here contain both magnetite – quartzite and hematite – martite – quartzite ores.

In the area of the *Kursk Magnetic Anomaly,* three basins are distinguished: Belgorod in the west, Kursk-Orel in the center, and Staryi Oskol in the east. Near-surface rich ore with hematite and martite has been exploited for some time. The horizontal quartzite – magnetite beds are of increasing economic importance. From the geological standpoint, the iron-bearing beds form part of the Precambrian basement of the Central Russian shield.

In the *western Siberian* ore province, three districts exist. However, only the *Kuznetsk – Sayan* area has acquired economic importance. It consists of a large number of contact-metasomatic magnetite ore bodies of slight and moderate thickness, which also contain iron sulfide and zinc sulfide. The deposits are thought to have originated in the time of the Caledonian orogenesis.

The *Gornyi Altai* district has limited reserves in Clinton-type deposits (sedimentary deposits with partial secondary enrichment), which date from the Variscan orogenesis.

The western Siberian district, in a narrower sense, has considerable reserves of minette-type sedimentary ore deposits that have not been much explored; these had their genesis as marine coastal deposits in the Upper Cretaceous and the Eocene.

Sweden. For a long time, Sweden was the most important ore supplier for the Central European and, in part, the Western European iron industry. Quartz-banded magnetite – hematite ores of the Lake Superior type were mined in central Sweden. However, the chief source of ore for export is northern Lapland, in a district centered around Kiruna and Malmberget, where Kiruna type magnetite deposits (massive intrusions in granite-like enclosing rock, with a relatively high apatite content) occur. Reserves are more than 2×10^9 t, with ca. 60 wt % iron and 0.04 – 5 wt % phosphorus. The phosphorus content is reduced to 0.04 wt % by beneficiation, because hardly any Bessemer pig iron is produced in western Europe any more.

2.1.1.4. Beneficiation of Iron Ore

The purpose of beneficiation is to render the ore more suitable for transport and for various reduction and smelting processes in the production of pig iron or steel.

The only processing steps used for ores mined from rich deposits (rich ores) are crushing and screening. Such ores come mainly from the weathering zones of itabirite deposits. At some mines, washing is combined with the screening of fines ($< 6 - 8$ mm) for sintering; very fine material containing kaolin is thus separated, and the alumina content is reduced. Low-grade ore is *concentrated* to lower transport costs and bring about a chemical composition suitable for the subsequent reduction step. In this enrichment process, the gangue is reduced and, in many

cases, its composition improved. Current reduction and smelting processes require a basic slag ($CaO : SiO_2 > 1$), but the gangue, of nearly every type of iron ore, is "acidic", that is, it contains an excess of silica. Beneficiation lowers the silica level to 5 – 8 wt % in the concentrate.

Harmful components present in the ore must be removed. Such components include phosphorus, arsenic, chromium, copper, vanadium, alkali, sulfur and titanium; these either make metallurgical processing more difficult or degrade the iron and steel quality. A small quantity of titanium (< 0.5 %) is, however, added to the burden at some ironworks, because it has been found to increase the life of the blast-furnace brickwork.

The basic methods of beneficiation for iron ores are usually the same as for nonferrous ores and industrial minerals. Naturally some peculiarities exist because of the mineralogy and the type of intergrowth.

Crushing. Conventional equipment is employed for crushing (→ Size Reduction). Primary crushing takes place in *gyratory crushers* (up to 3500 t/h) or *jaw crushers.* Mobile versions of the latter are now employed directly at the mines. *Cone crushers* are common for the second and third stages of crushing. They are however, increasingly replaced by *semiautogenous grinding* (SAG). Here steel balls, 100 mm or greater in diameter are added in a quantity of up to 6 vol % to aid the comminuting action. *Autogenous mills* are tumbling mills of large diameter and relatively short length, roughly 0.4 – 1.2 times the diameter. *Fine grinding* takes place in *rod* and *ball mills,* usually in a closed circuit with classification by hydrocyclones. Both dry and wet comminution are practiced; the wet process is more common when this step is followed by concentration.

Another option is autogenous or semiautogenous primary crushing followed by fine grinding in a *pebble mill;* a tumbling mill which does not contain steel balls but lumps of ore of ca. 75 – 30 mm in diameter. This is possible only with hard ore that yields enough "pebbles" in the first crushing stage. This process has the advantage that no grinding balls are needed, but the energy requirement for grinding often increases if the conditions in the mill are not optimal.

Vibration screens with elliptical or linear oscillations have been widely adopted for screening. Multilevel screens of the Mogensen design are increasingly used.

Concentration. The concentration of iron ore is generally based on physical differences between the iron-bearing and gangue minerals (see Table 3). *Flotation* (→ Flotation) and *electrostatic concentration* (→ Electrostatic Separation) utilize physicochemical differences. The iron minerals have a higher specific gravity than the gangue minerals, so in many cases *gravity* (→ Gravity Concentration) or *centrifugal concentration* (→ Centrifuges, Filtering, → Centrifuges, Sedimenting, → Hydrocyclones) is possible. The *dense-media separation* (→ Dense-Medium Separation) process for coarse particles (≥ 3 mm) in drum or tank separators is also used. For medium-sized material, ca. 15 – 0.5 mm, cyclone separators or the cylindrical "Dyna Whirlpool" can be employed. Both use a centrifugal force field for separation. Gravity separators for fines (roughly 2 – 0.075 mm) include the Humphreys' spiral concentrator and the Reichert cone separator.

The Reichert cone separator has been adopted in the past two decades mainly for the beneficiation of placer ore, but it is used more and more for fine-grained weathering iron ores.

The *magnetic concentration* (→ Magnetic Separation) of magnetite ore takes place in drum-type separators with a low-intensity magnetic field. Dry separators are fed from the top and discharge the nonmagnetics, in wet separation the ferromagnetics are picked up from the feed pulp (underfeed). Ferrite-based permanent magnets have almost completely replaced ALNICO and electromagnets. The magnetic flux density is up to 0.15 T. Wet separators include drums up to 1500 mm in diameter and 3600 mm wide, with throughput capacities of 50 – 200 t/h. The drums, with up to five poles, are arranged in appropriately designed tanks for cocurrent or countercurrent operation. In addition, *high-intensity wet* magnetic separators have been employed since ca. 1970 in the concentration of weakly magnetic ore. The Jones design is most common. Here the separating elements are arranged in a ring which rotates carousel-fashion through a strong field generated by electromagnets. The magnetic flux density in the separation zone is 0.1 – 0.5 T; because of the tip configuration of the separating elements, it is concentrated at the tips to much higher values with a strong

convergence of the field lines. The carousel design permits semicontinuous operation.

High-intensity dry magnetic separators are generally not used in beneficiation any longer. Electrostatic drum type separators are employed only for the recleaning of the concentrate, where they remove mainly quartz-bearing gangue. The process is restricted to grain sizes in the range of 1 – 0.05 mm.

Flotation has been used for iron-ore beneficiation since ca. 1950. The first two systems in Michigan were built for "direct" flotation (values in froth) of fine-grained specularite ore. Tall oil and diesel oil were added as collectors. "Hot flotation" enabled concentrates to float with only a few percent of silica. The rougher concentrate in pulp form was heated to $> 90\,^{\circ}C$, then recleaned in several stages after cooling. "Indirect" or "reverse" flotation is most commonly used today; the gangue goes into the froth and the iron ore into the underflow. Conditions are controlled so as to prevent the ore from flotation. As depressants for iron ore, starches or dextrin are used in conjunction with a basic pH. The collectors employed for gangue depend on the gangue minerals: oleic acid or oleates (e.g., for apatite) or amine salts (for quartz and feldspar). In general, the goal is first to enrich the ore as much as possible by less expensive means (gravitiy, magnetic separation), using relatively expensive flotation only for the final beneficiation stage. In the case of rich ores, however, flotation can also be employed as the sole concentration process when only a little gangue is to be removed.

A process of limited importance is *magnetizing roasting* followed by low-intensity magnetic separation. The iron minerals are thermally converted to magnetite with solid, liquid or gaseous reducing agents. The process is used in shaft furnaces in China and rotary kilns in the Former Soviet Union.

A *thermal* beneficiation process is the roasting of siderite ore. The purpose of this process is to save weight for transport and to relieve the blast furnace of the carbon dioxide load. Shaft furnaces or sintering machines are employed. In the latter case, agglomeration is combined with calcination.

Figure 2 presents a general flowsheet for the beneficiation of iron ore. Currently (ca. 2000) almost all the crude ore comes from large-scale open pit mines with capacities between 10 000 and 100 000 t/d. Only in Swedish Lapland, the Former Soviet Union and China are high-capacity underground mines operated. Depending on the iron content and other requirements, the ore is processed into lumps, sinter or pellets; Table 6 lists the fractions of these products in blast-furnace burdens in several areas. Table 7 gives some analyses of ores that are of commercial importance.

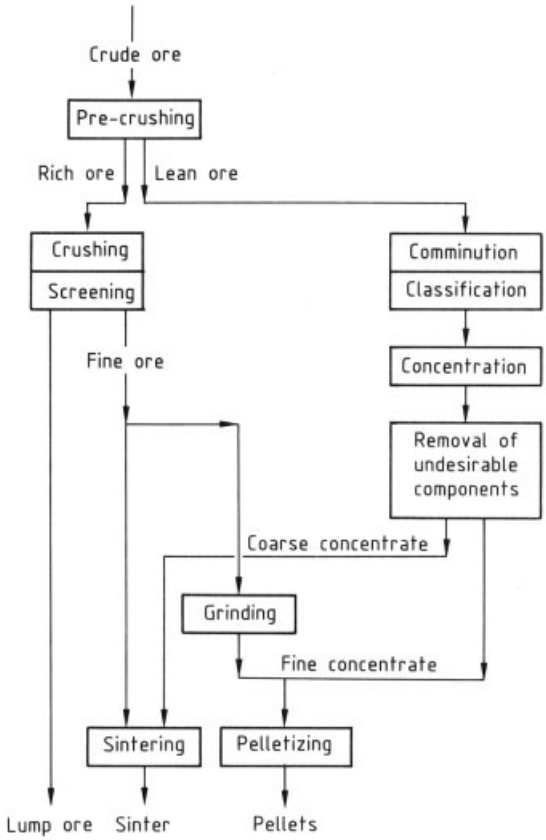

Figure 2. Schematic diagram of iron-ore processing

The beneficiation of an Indian laterite ore is shown schematically in Figure 3. The ore is a decomposition product containing limonite and hematite values along with kaolin, quartz and alumina minerals as gangue. By washing in a drum and then removing the finest fraction, the iron content is elevated by ca. 3 % units and the Al_2O_3 content is lowered from 3 – 4 % to < 1.7 wt %.

Table 6. Lump ore and agglomerate in blast-furnace charge, 1986, wt % [15]

	Germany	Western Europe	Japan	United States	Total
Lump ore	9.8	12.1	16.8	5.3	12.5
Sinter	64.4	70.5	73.6	22.5	62.4
Pellets	25.8	17.4	9.6	72.2	25.1
Total	100.0	100.0	100.0	100.0	100

Table 7. Analyses of iron ores [16, 17]

Source	Fe_{tot}	SiO_2	Al_2O_3	CaO	MgO	Ignition loss
Lump ores						
Hamersley Iron	65.0	2.9	1.6	0.1	0.1	2.7
Mt. Newman Mining	65.7	3.4	1.2	0.1	0.1	1.5
CVRD-Itabira	67.7	1.1	1.0	0.0	0.0	0.4
CVRD-Carajas	67.6	0.7	1.1	0.1	0.0	1.3
Ferteco Feijao	67.3	1.8	1.0	0.1	0.0	0.9
MBR-Aguas Claras	68.4	0.6	0.7	0.1	0.1	0.8
Chowgule	60.1	3.5	2.8	0.0	0.0	not known
Bellary-Hospet	66.9	2.1	1.9	0.1	0.1	0.8
SNIM-Mauretania	63.1	6.8	1.3	0.0	0.0	not known
ISCOR-Sishen	66.5	3.1	1.1	0.1	0.1	0.6
ISCOR-Thabazimbi	64.0	6.0	1.0	0.2	0.4	1.2
Cerro Bolivar	64.2	1.3	1.0	0.1	0.1	5.5
Fine ores						
Hamersley Iron	62.5	4.4	2.6	0.1	0.1	3.2
Mt. Newman	62.6	5.5	2.4	0.0	0.0	4.7
Robe River	57.2	5.6	2.7	0.1	0.1	9.2
CVRD-Sinter Fines	64.4	5.0	1.2	0.1	0.1	1.5
CVRD-Carajas	67.9	0.6	0.7	0.1	0.1	1.4
Ferteco-Fabrica	64.3	3.1	1.9	0.1	0.1	2.7
Romeral	65.4	4.4	0.9	0.9	1.0	0.8
Mifergui-Nimba	67.5	2.3	1.0	0.1	0.0	1.0
Sesa Goa	60.6	3.5	4.2	0.1	0.1	3.8
NMDC-Bailadila	65.4	2.1	1.9	0.0	0.0	not known
Carol Lake	65.8	4.5	0.1	0.5	0.4	0.3
Bong Mining Co.	64.7	7.5	0.3	0.2	0.3	0.3
LAMCO	64.7	4.7	1.0	0.0	0.0	1.7
LKAB – Kiruna KBF	69.7	1.2	0.2	0.6	0.5	0.2
LKAB – Kiruna KDF	62.0	4.6	0.8	3.6	1.2	0.6
Co. Andaluza de Minas	55.0	4.4	1.1	4.3	0.4	not known
ISCOR – SISHEN	65.5	3.7	1.6	0.1	0.1	0.7
C.V.G. – El Pao	64.5	3.7	1.4	0.1	0.1	1.9
Pellets						
Eveleth Mines	65.6	4.9	0.1	0.8	0.4	
Kiruna	66.1	2.5	0.3	0.3	2.0	
Malmberget	66.7	2.2	0.5	0.2	1.6	
CVRD	65.3	2.8	0.5	3.0	0.1	
Ferteco	65.0	3.3	0.9	2.5	0.0	
Samarco	65.1	2.6	1.2	2.2	0.2	
Kudremukh	66.1	2.8	0.3	2.3	0.1	
Carol Lake	65.6	4.9	0.3	0.5	0.3	
Cartier Mining Co.	65.2	5.2	0.5	0.4	0.3	
Wabush Lake	65.8	3.0	0.4	0.1	0.0	
Bong Mining Co.	64.1	7.0	0.3	0.3	0.3	
Las Encinas	66.8	1.5	0.7	1.6	0.8	
Empire CCI.	65.3	5.6	0.4	0.2	0.3	
Empire CCI.	59.5	5.3	0.4	7.0	1.9	
Minntac	65.6	5.4	0.2	0.2	0.3	

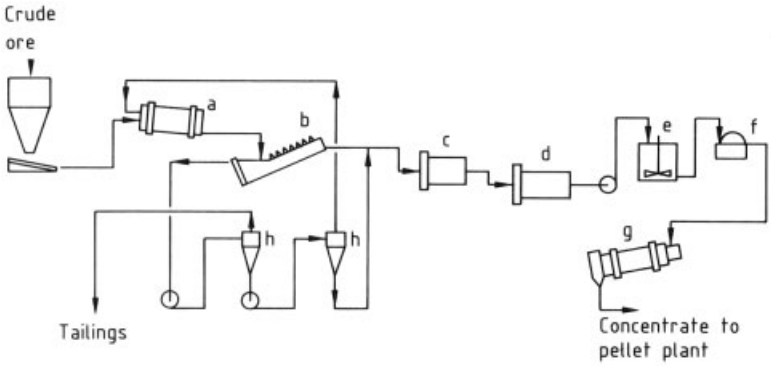

Figure 3. Flowsheet of Pale beneficiation plant, India
a) Washing drum; b) Spiral classifier; c) Rod mill; d) Ball mills; e) Slurry tank; f) Disk filter; g) Rotary dryer; h) Hydrocyclones

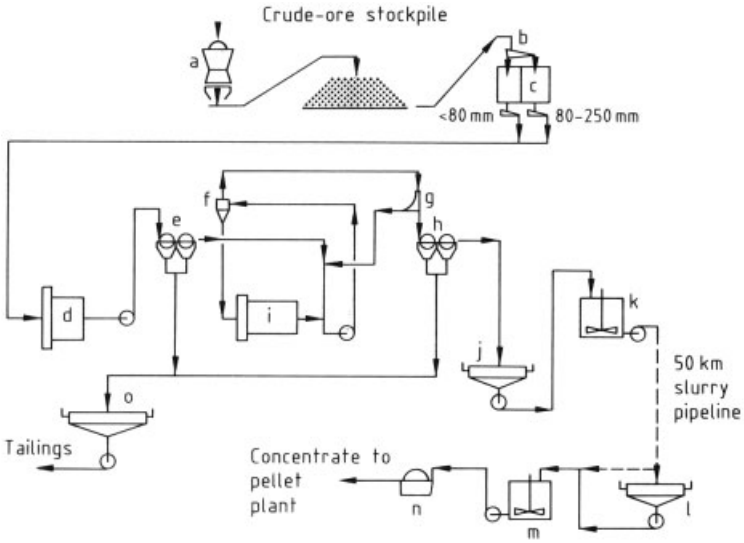

Figure 4. Flowsheet of Pena Colorada beneficiation plant, Mexico
a) Primary crusher; b) Crude-ore screen; c) Crude-ore bins; d) Autogenous mill; e) Primary magnetic separation; f) Hydrocyclones; g) Sieve bend; h) Secondary magnetic separation; i) Ball mill; j) Concentrate thickener; k) Slurry tank; l) Concentrate thickener; m) Slurry tanks; n) Disk filter; o) Tailings thickener

This simple way of improving the composition of iron-rich fine ore is often used in India and Brazil. A basically similar process with washing troughs or log washers was formerly used for high-alumina ores in many other regions (Salzgitter, Ilsede, Mesabi Range). If much of the alumina present is associated with limonite, a rod mill is employed instead of a washing drum. Applying 3 – 4 kWh/t on comminution frequently results in satisfactory removal of alumina. Figure 4 illustrates the beneficiation of a magnetite ore from a Magnitnaya-type deposit. Secondary and tertiary crushing and primary grinding are replaced by an autogenous mill. The concentrate has ca. 69 wt % iron; with a particle-size distribution of ca. 80 % $<$ 0.045 mm, which is fine enough for pipeline transportation and pelletizing. Numerous beneficiation plants for magnetite ore from other deposit types have been designed in a similar way, for example in the Former Soviet Union (Krivoi Rog, Kursk, Olenogorsk), in the Lake Superior area, and in northern Sweden. Older installations still employ conventional comminution (three stages of crushing and two to three stages of grinding). For hematite ore, the first grinding stage is followed by gravity separation or, in the case of fine grained ore, further grinding is followed by high-

intensity wet magnetic separation or flotation. In many cases, however, the principal iron minerals in such ores from weathering zones include not only hematite but martite (which is only a little less magnetizable than magnetite) with residual magnetite. In this case, complex flowsheets must be adopted, including low- and high-intensity magnetic separation, flotation, classification and slurry treatment for dewatering of the concentrate. This is also the case if undesirable constituents are present (e.g., apatite).

Finally, the *selective dispersion* process should be mentioned. In one Michigan plant, this operation is a preliminary to flotation. After grinding to the liberation size, sodium hydroxide and starch are added to the ore pulp (sodium hydroxide to disperse aluminous and limonitic constituents, starch to flocculate the specularite) and the flocculated agglomerates are separated from the dispersed minerals in a hydroseparator. The downstream flotation removes coarse quartz which is discharged in a conventional manner, with amine salts as collectors.

2.1.1.5. Agglomeration

In present-day smelting technology, very fine material is undesirable in the reduction equipment. The blast furnace is the dominant type of equipment for the production of iron metal. Only in exceptional cases (small furnaces with capacities up to ca. 100 – 150 t/d of pig iron) is unscreened crude ore used (after primary crushing if necessary). Modern high-performance blast furnaces require physical and metallurgical preparation of the burden. Formerly agglomeration was used in a purely physical way, as a process of size enlargement aimed at improving the permeability of the blast-furnace burden; however, developments in recent years have shifted part of the slagging into the fine-ore or concentrate sintering step. On the other hand, the requirements for physical and chemical quality of the blast furnace burden have become more stringent, and mining companies that supply ore to several ironworks (as the large ones generally do) now prepare a variety of concentrate or fine ore qualities from the same or similar crude ores.

As a rule, blast-furnace burden contains lump ore, sinter, and pellets. *Pellets* are made from both concentrate and iron-rich fine ore (see Fig. 1). If the beneficiation process yields very fine-grained concentrates (e.g., < 0.1 mm), these are usually converted to pellets at the mine and transported in this form. *Sinter* is generally produced at the ironworks. The varying loads imposed on material during transport to the blast furnace mean that pellets must satisfy stringent mechanical strength requirements. Pellets are also nearly always made from one well-defined ore or concentrate, whereas sinter is produced from predesigned mixtures of ores and additives. There are exceptions, such as pellets made at ironworks from mixed ore, or sinter made from a single ore or concentrate quality where an ironworks uses ore from a single deposit. Blast furnaces are also sometimes operated with lump ore alone, sinter alone, or pellets alone. Sponge iron plants are generally fed with a single iron source.

Sinter and pellets differ in the way the particles are bound into the agglomerate. In pellets, the ore particles are connected chiefly by bridges of ore. When magnetite is oxidized at temperatures > 900 °C during hardening of pellets, these bridges arise through mobilization of lattice constituents during recrystallization into the hematite lattice. In the case of hematite ores, mobilization is less pronounced, because the crystal lattice structure does not change. The formation of iron oxide bridges is supported by additional slag bridges; it must be remembered that the quantity of slag in many pellet qualities is only 3 – 4 wt % (corresponding to 6 – 7 vol %).

Sinter consists of a calcium ferrite matrix, with siliceous and aluminous constituents dissolved and primary hematite embedded in it. Whereas pellets are thoroughly oxidized (a maximum of 0.5 wt % FeO), sinter contains 4 – 6 wt % FeO, which is linked to silica and alumina in the matrix.

Pellets are burned at a temperature below the melting points of the constituents. Their strength gain is adequate only when the reaction temperature is controlled to $\pm$ 10 °C over long time periods. For this reason, hot gas is used to supply heat; it can be generated by the combustion of fuel gas, oil, or even coal. In sintering, on the other hand, the constituents of the mixture are melted. This happens at temperatures so high that the reaction itself goes to completion within minutes, but in a thin layer that continuously advances through the material bed cocurrently to the gas. The fuel is coke breeze, or occasionally

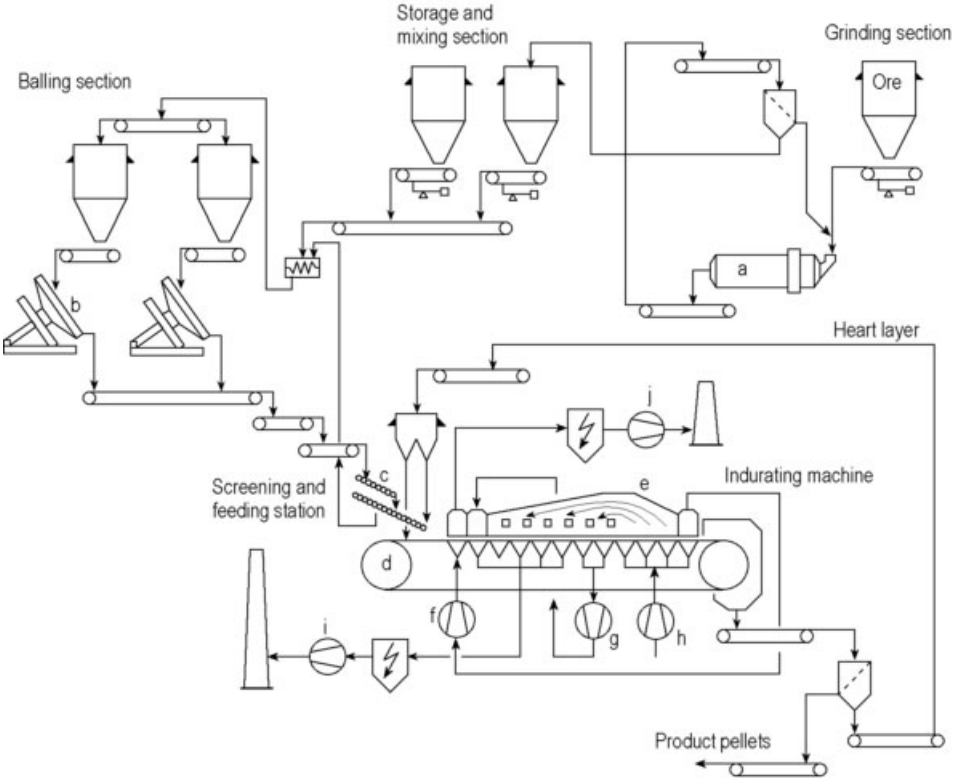

Figure 5. Flowsheet of Lurgi – Dravo traveling-grate pelletizing process
a) Grinding mill; b) Pelletizing disks; c) Roller feeders; d) Travelling grate; e) Firing hood; f) Updraft drying fan; g) Recuperation fan; h) Cooling air fan; i) Waste-gas fan; j) Hood exhaust fan
DRI = direct-reduced iron

non-gassing coal fines, evenly distributed in the sinter mixture.

Pelletizing. Iron-ore pellets are made in two steps. The concentrate to be pelleted must have a sufficiently fine particle-size distribution (65 – 85 wt % $<$ 0.045 mm) and a specific surface area of at least 1600 cm^2/g as measured by a permeability method (Blaine, Fisher Subsieve Sizer).

Spherical *green pellets* (green balls) of appropriate size (ca. 9 – 16 mm in diameter) are formed by a pelletizing disk or drum. The particles are bound together by capillary forces acting through water bridges between them. This means that green pellets can be obtained only at a narrow range of moisture content. If the water content in the fine ore (which is usually recovered as a filter cake) is too high, water-absorbent materials such as bentonite can be used to correct it within narrow limits. If the water content is low (ideal case) it is useful to hold it in the filter cake ca. 0.5 – 1 % lower than the green pellet moisture content to suppress the premature formation of undersize pellets ("micropellets") in the granulator.

Prior to (or during) the formation of green pellets, substances are added to the fines to render them stable against the stresses that occur in the hardening process. In addition to bentonite, whose active constituent is strongly water-absorbent montmorillonite, organic binders such as Peridur have come into use. The latter are advantageous in that they can be burned off during hardening, whereas bentonite elevates the silica and alumina content of the pellets.

Purchasers usually prescribe the gangue composition of pellets to fit the operating requirements of the particular blast furnace and the chemical composition of the other iron sources to be used. This composition, which also governs the chemical and metallurgical properties of the pellets, is controlled with additives, the most im-

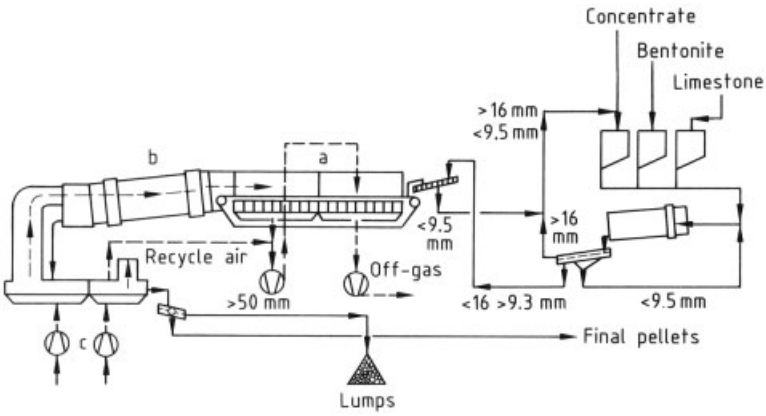

Figure 6. Flowsheet of combined traveling grate/rotary kiln ("grate kiln") process
a) Traveling grate; b) Rotary kiln; c) Annular coolers

portant being slaked lime, limestone, dolomite and olivine. These are added in a grain size corresponding to that of the fines.

The second operation is the *thermal hardening* of the pellets at 1200 – 1320 °C. Nonthermal hardening is employed only in a few unimportant cases. Three distinct heat hardening systems exist: the traveling grate (see Fig. 5), the combination of preheating traveling grate and rotary kiln (grate – kiln) (see Fig. 6), and the shaft furnace. The shaft furnace is mainly suitable for magnetite concentrates and low capacities (up to ca. 500 000 t/a) and has lost much of its market share in the past 15 years. The traveling grate and the grate – kiln combination each share ca. 50 % of the market.

The "basicity" of the final pellets, expressed as the ratio $(CaO + MgO) : (SiO_2 + Al_2O_3)$, is $<$ 0.1 for "acid pellets" or up to 1.3 for "basic pellets". Each range of values offers certain advantages and disadvantages with respect to mechanical strength and metallurgical qualities.

After a number of years of declining pellet production and the shutdown of several plants, the western world had, at the end of 1987, a pellet capacity of roughly 163×10^6 t/a (100 units using the three hardening processes).

According to present information, the Former Soviet Union has 20 traveling-grate plants and 5 grate – kiln plants. The first group has a total reaction area of ca. 6500 m^2, which might correspond to a pellet capacity of ca. 50×10^6 t/a. Grate – kiln plants add an estimated capacity of ca. 10^7 t/a.

While the Former Soviet Union has a large number of pellet plants, up to now this has not been the case for China.

For the assessment of the mechanical and metallurgical properties of burned pellets, a number of parameters are determined in more-or-less standardized tests. The most important are:

1) Cold Compression Strength. Individual pellets 10 – 12 mm in diameter are compressed to failure between two plane or slightly concave faces of a hydraulic press, and the crushing force is measured (values found in practice are usually $>$ 3000 N per pellet).
2) Abrasion Resistance. The abrasion resistance is determined in the ISO tumble test. Under standardized conditions, pellets are tumbled in a drum and the $>$ 6.3 mm and $<$ 0.6 mm fractions are determined. The values should be at least 95 % and at most 4 %, respectively.
3) Low-Temperature Disintegration Test. This test indicates how the pellets will behave, for example, in the blast furnace in the critical range of temperatures around 500 °C. In an electrically heated rotary kiln, pellets are heated to 500 $\pm$ 10 °C and held at this temperature for 60 min under reducing conditions. After cooling, the pellets are screened to determine the $>$ 6.3 mm and $<$ 0.5 mm fractions. The degree of reduction is determined. The values for the $>$ 6.3 mm fraction are usually between 60 and 80 wt %; those for $<$ 0.5 mm usually 10 – 20 wt %.

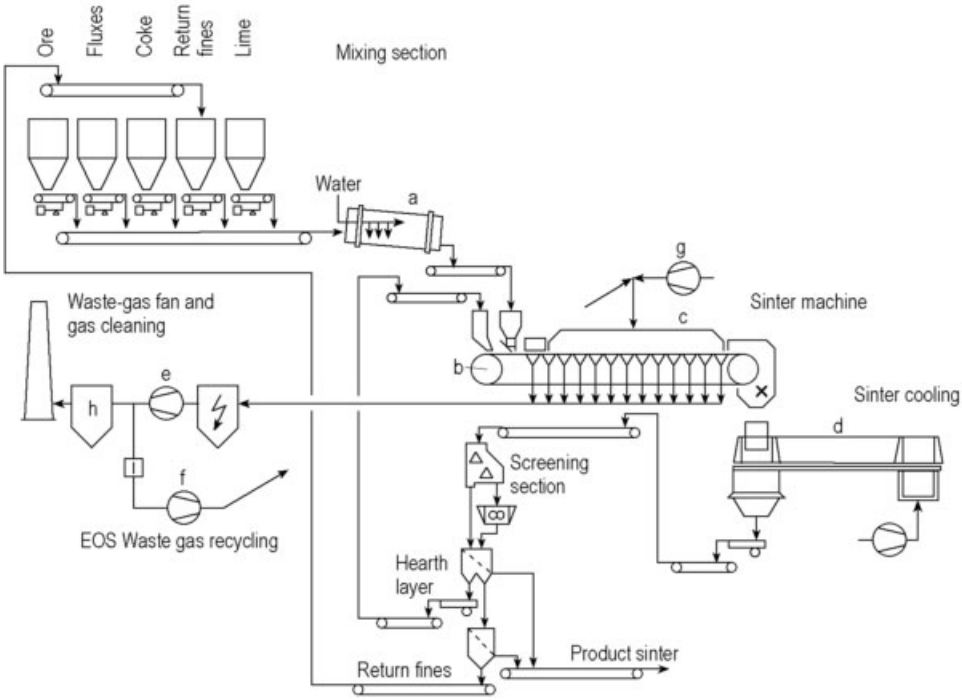

Figure 7. Flowsheet of sintering
a) Mixing drum; b) Travelling grate; c) Gas distribution hood (EOS only); d) Annular cooler; e) Waste-gas fan; f) EOS recycling gas fan; g) Fresh air blower (EOS only); h) Waste-gas final cleaning
EOS = Emission-optimized sintering

4) Free Swelling. In a vertical tubular furnace, 18 pellets are reduced with a carbon monoxide – hydrogen mixture at 1000 °C. The weight loss during reduction yields the degree of reduction. The volume change in the pellets indicates the swelling, which should be 20 % at maximum. The compression strength measured in a hydraulic press should be > 500 N per pellet.
5) Reduction under Load (RUL). In an upright retort, pellets are heated to 1050 °C, placed under a mechanical load of 5 N/cm^2, and exposed to a reducing atmosphere. The pressure drop of the reducing gas across the bed and the reduction rate are determined. A low pressure drop indicates pellets stable in reduction. The pressure drop should not exceed 100 Pa.

The reduction rate at 40 % reduction is an important parameter of reducibility; dR/dt_{40} should be at least 1 % per minute.

Special test methods have been devised for pellets used in various direct-reduction processes (HyL reduction test in an upright retort, Midrex test, and testing for the SL/RN process in an electrically heated laboratory rotary kiln). In the test for the SL/RN process, pellets and reducing agent (usually coal) are investigated at the same time. A number of different or modified test procedures are employed in individual countries or at individual ironworks.

Sintering. Iron-ore sinters are virtually all produced in traveling-grate installations (see Fig. 7). The rotary kiln and pan-sintering processes formerly in use are no longer important. Likewise, cooling of the hot sinter at the last section of the grate has become obsolete. Current standard is reflected by the straight-grate "Dwight – Lloyd process". Straight-grate and annular coolers are in use. Quite some effort is being made to reduce the level of solid and gaseous contaminants in the sintering off-gas (dust, SO_2, NO_x). A new development recently introduced by Lurgi in industrial plants is the Emission Optimized Sintering (EOS) process. In the EOS process approximately half of the off-gas of the sintering grate is recirculated

back to the gas distribution hood, where it is mixed with fresh air and used as oxygen carrier and heat transportation medium in the sinter bed. The advantages are reduction of the final waste-gas volume (and thus smaller waste-gas cleaning facilities) and substantial reduction of pollutants (dust, SO_2, NO_x, dioxins, furans) in the waste gas. Fuel consumption is lowered by about 5 kg per tonne of sinter, whereas the consumption of electrical energy is increased by about 1.5 kW h/t [18]. Sintering units are nearly always operated on site at the ironworks. As a rule, several ores are processed in one unit; a "monosinter" from just one ore is the exception. Along with ore and concentrate, the mixture for sintering also contains recycle material from the ironworks, provided no buildup of undesirable or detrimental substances results. Undersize sinter from the screening step is also recycled. The basicity of the sinter has been substantially increased in recent years. Originally sintering was considered only as a means to deal with fine ore with respect to improved handling and permeability of the blast-furnace burden ("physical burden processing"). The overall chemical composition of the blast furnace feed (including coke ash) was adjusted by addition of suitable materials in the form of lumps (limestone, dolomite, manganese ore if necessary). Thus, the basicity of sinter and lump ore generally was < 1 (exception minette ore with high calcite content of the gangue).

Today these additives are generally added to the sinter mix as fines (< 5 mm) and sometimes precalcined (burnt lime instead of limestone). The basicity of the sinter is generally higher than that of the blast furnace slag to compensate (1) for the "acid" coke ash, (2) the gangue of the lump ore, and (3) slag components of the iron-ore pellets. Increasing the basicity also improves the quality of the basic sinter as well as the heat economy. Many plants today produce sinter with a basicity of > 1.5. The slag analysis is controlled by addition of limestone or burnt lime, dolomite (to control CaO and MgO), olivine (to control MgO and SiO_2), and quartz sand. The most common fuel is coke breeze. To ignite the mixture gas or oil heating is applied. The total heat consumption for an iron-rich basic sinter is between 1.3 and 1.7 MJ per tonne of final sinter. When low-iron ore mixtures (e.g., those containing minettes) or ores containing high levels of alumina (2 – 4 wt % Al_2O_3) are sintered, the heat consumption is higher (up to 2.5 MJ/t). Analyses of sinter from German ironworks fall in the following approximate ranges:

Fe_{tot}	55.5 – 58 wt %
Fe(II)	4.5 – 6 wt %
SiO_2	5 – 6 wt %
CaO	9 – 11 wt %
MgO	1.2 – 3.2 wt %
Al_2O_3	1.2 – 2 wt %
$CaO : SiO_2$	1.4 – 1.9

Ironworks in the *Federal Republic of Germany* use imported ores almost exclusively; some heat-consumption and analysis figures for other European ironworks differ from the ones cited, because low-iron domestic ores still are processed there. In contrast, sintering plants in *Japan* operate under conditions similar to those in Germany. The basicity of the sinter in Japan, however, is nearly always > 1.8.

In the *United States,* the fraction of sinter in the burden is only 22.5 wt % (1986; see Table 6), much lower than in Europe and Japan. The fraction of pellets, on the other hand, is very high (72.2 wt %). Thus, in the United States improvement in pellet quality, above all in metallurgical properties, have been the focus of attention, with the result that basic pellets with $CaO : SiO_2$ ratio of ca. 1.5 have been sought.

Few details are known about current sintering technology in the *Former Soviet Union* and *China*. Rich fine ores are rare in both countries, and agglomerates are made chiefly from concentrates. In some works, blast furnaces are operated on 100 % sinter. Since the basicity of the sinter in these cases must be adapted to the requirements for blast-furnace operation, the basicity values are often too low for a sinter of truly good quality. One remedy might be to pelletize part of the concentrate at a low basicity and raise the basicity of the sinter to 1.6 – 1.8.

A variety of test methods are used for sinter quality control; some of the techniques are similar to the ones for pellets. The methods include:

1) Shatter Test. The shatter test has been standardized only in Japan. At Lurgi, 150 kg of sinter (grain size > 8 mm) is dropped three times from a height of 2 m onto a steel plate; then the > 20 mm, 20 – 8 mm, 8 – 2 mm and < 2 mm fractions are determined.
2) Tumble Test. The shatter test is often replaced by the tumble test, which follows the same guidelines as described above for pellets. The

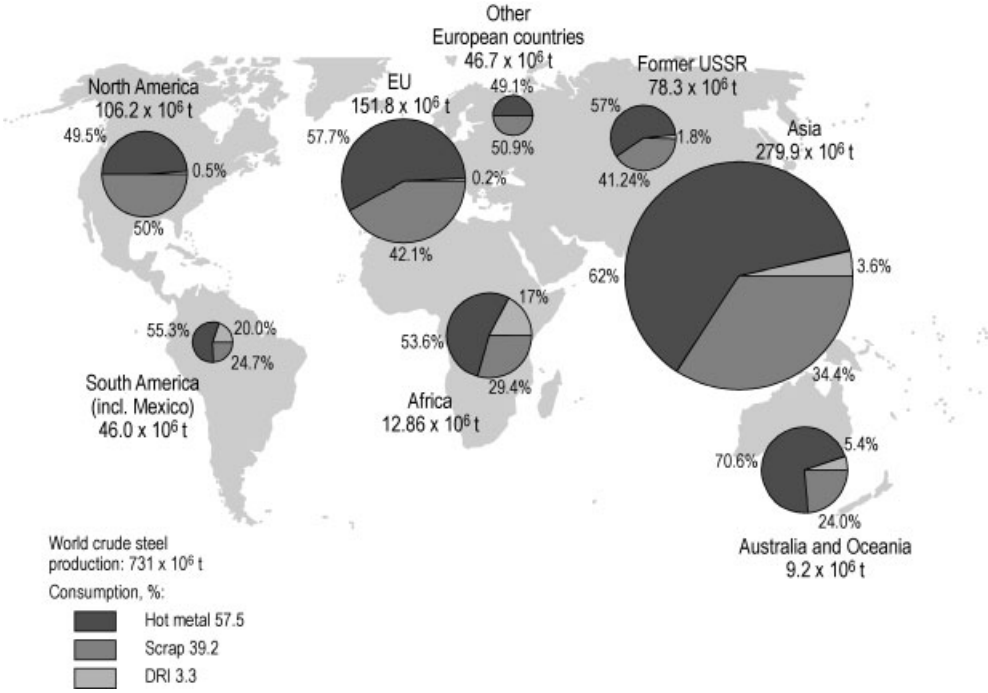

Figure 8. Consumption of hot metal, scrap, and direct-reduced iron (DRI) for the production of crude steel in 1994 (average charge 1152 kg/t crude steel)

fraction < 6.3 mm should be at least 70 wt %; at German ironworks it is between 70 and 85 wt %.

3) Low-Temperature Disintegration (LTD). This test is done according to ISO 13930 known also as the dynamic disintegration test and mainly used for pellets and lump ores. A sample of 500 g (10 – 12.5 mm for pellets) is reduced at 500 °C in a rotating drum (150 mm in diameter and 540 mm in length). The gas mixture consists of 20 vol % CO, 20 vol % CO_2, 2 vol % H_2 and 58 vol % N_2. Reaction time is 60 min. The LTD index is the fraction > 6.3 mm.
4) Reduction Degradation Index (RDI). The reduction degradation index is now regarded as the most important quality criterion for sinter. Within ISO two test procedures exist to evaluate the RDI value, both known as static disintegration test. In ISO 4696–1 500 g of sinter (10 – 12.5 mm) in a stationary bed is reduced for 60 min at 500 °C by a gas mixture comprising 20 vol % CO, 20 vol % CO_2, 2 vol % H_2 and 58 vol % N_2. After cooling the sample is placed in a drum (130 mm in diameter, 200 mm in length, with two lifters) and subjected to 300 revolutions. The RDI index is > 3.15 mm. This test is normally used in Europe. In Germany, the value of > 3.15 mm is required to be > 70 % for sinter. In ISO 4696–2 500 g of sinter (16 – 20 mm) in a stationary bed is reduced for 30 min at 550 °C by a gas mixture comprising 30 vol % CO and 70 vol % N_2. After cooling the sample is placed in a drum (130 mm in diameter, 200 mm in length, with two lifters) and subjected to 900 revolutions. The RDI index is the fraction < 2.8 mm. This test is mainly used in Japan. The RDI index for sinter in Japan is between 30 and 40 wt %.
5) Relative Reducibility. Ironworks use very different methods for the relative reducibility test. A sample is reduced at 900 °C using a carbon monoxide – nitrogen mixture similar to RDI ISO 4996–2. The index reported is the degree of reduction at the end of the test. In addition, a test as set forth in ISO 4695 is carried out to determine the reduction rate at 40 % reduction, which is usually ca. 1.5 % per minute.

2.1.2. Reducing Agents [19, 20]

A considerable amount of heat input is required in the production of iron, because essentially all the processes involved are pyrometallurgic. Part of the heat is used to heat up the raw materials to be processed and part of it is required for endothermic chemical reactions. The majority of this energy is obtained by gasification of fossil fuels to reducing gas. Approximately 40 % of the annual output of steel is derived from scrap in *electric arc furnaces,* which pose a correspondingly high demand for thermoelectric energy. In 1995 the percentage shares of the different steel production methods were as follows: 60 % oxygen steel, 33 % electric furnace steel, and 7 % open hearth steel. The proportion of the obsolete open hearth steelmaking process is largely a result of the slow pace of change in Eastern Europe. In 1994 hot metal had a share of 57.5 % of metals used for world crude steel production and the percentage of scrap was 39.2 % (see Fig. 8) [21].

Reducing Agents. Carbon and hydrocarbon carriers, chiefly in the form of coke, are gasified to the primary reducing agents (CO, H_2) in blast furnaces. Coke can be replaced partly by carbon and by hydrocarbon carriers injected via the tuyeres. Examples of hydrocarbon carriers are coal, oil, natural gas, coke oven gas, or plastics wastes. But for physical reasons a blast furnace cannot be operated without coke. As iron ores change their chemical and physical properties from top to the hearth of the furnace according to reduction, softening and melting, the coke remains the only solid material below the smelting zone of the iron bearing materials. On the one hand, it forms a strong grid which supports part of the weight of the overlying burden and on the other hand, it constitutes a highly permeable "trickle column" (see Section 2.5) which allows an adequate drainage of the molten material's slag and hot metal and a good distribution of the gas ascending from the raceway at the tuyeres.

The coke plays a very important role in the so-called "cohesive zone", i.e., the region where the softening and melting iron-bearing materials are believed to form impermeable layers, separated by "coke windows". This is why a blast furnace operation without coke is not possible for physical reasons. In the blast furnace hearth the coke forms a cone, the so-called "dead man", which should be permeable for hot metal, slag and the furnace gas (see Section 2.5).

Requirements on Coke Properties. It is assumed that all the coke produced in a coke oven plant of an integrated iron and steel works will be used for the production of hot metal only. Concerning the grain size of the produced coke, the distribution of the size fractions to be used are, for example, 75 % "bell coke" (coarse coke) in the range of 30 to 80 mm, 17 % "nut coke" (small coke) in the range of 10 to 30 mm and 8 % "coke breeze" (coke fines) < 10 mm in size [22]. Coke quality requirements can be derived from the functions of coke in the blast furnace in conjunction with the mode of blast furnace operation. Table 8 lists the requirements on chemical composition and physical properties of coke for the operation of big blast furnaces with low coke rates, high productivities, and the production of hot metal with uniform chemical composition and temperature [23].

Table 8. Requirements on coke properties in Germany (Source VDEh Blast Furnace Committee)

Chemical composition	
Ash	< 9.0 wf % *
Volatile matter	< 0.75 wf %
Sulfur	< 0.7 wf %
Alkali	< 0.2 wf ($Na_2O + K_2O$)
Chlorine	< 0.05 wf %
Phosphorus	< 0.03 wf %
Zinc	< 0.003 wf %
Physical properties	
CSR	< 10 mm > 65 %
CRI	max. 23 %
Grain size (before drumming at delivery)	
> 100 mm	0 %
> 80 mm	max. 10 %
< 40 mm	max. 18 %
< 20 mm	max. 3 %

* wf = water free.

The ash content of the coke in itself is rarely a major problem for furnace operation, but nevertheless high coke ash can mean higher slag volumes and consequently higher energy consumption and lower furnace productivity. The main problems with coke ash are related to tramp elements. Especially alkalis and zinc have highly detrimental effects because they disturb the regularity of the furnace operation and entail a higher consumption of reducing agents as well

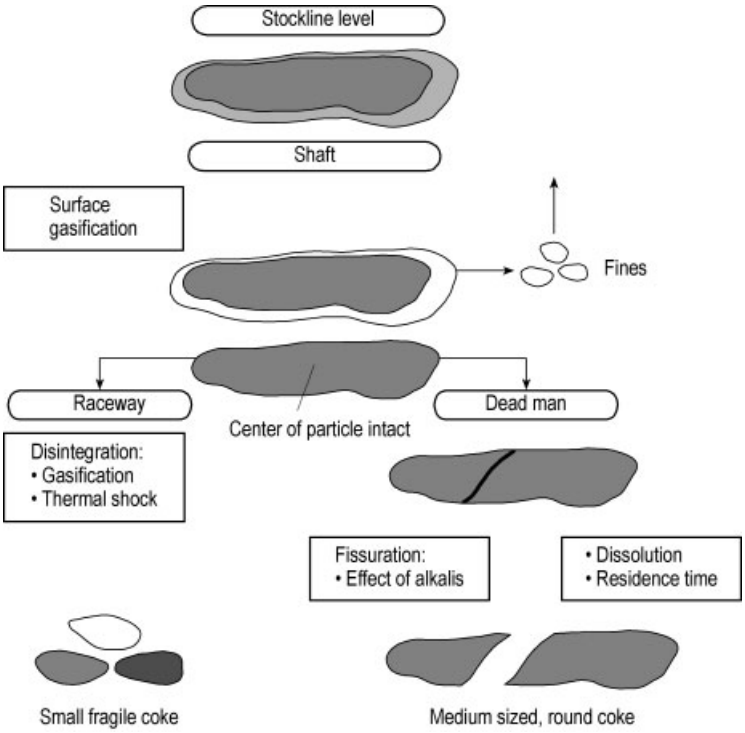

Figure 9. Schematic representation of the coke degradation in a blast furnace

as a loss in furnace productivity and in consistency of the quality of the hot metal.

Sulfur can be eliminated from the hot metal down to the required values partly by discharge with the blast furnace slag and partly by a separate desulfurization step prior to the steel shop, but its removal can become extremely costly. The use of sulfur-rich coke breeze in the sinter plant leads to higher SO_2 concentrations in the waste gas and to higher costs for waste gas cleaning. Both sulfur and alkalis should be discharged from the blast furnace via the slag. However, it is difficult to remove sulfur and alkalis simultaneously, because sulfur can only be chemically bound in a basic slag and removed from hot metal whilst alkalis can only be bound in an acid slag.

Coke moisture and chlorine content can be a problem for an efficient furnace operation, because both increase the dew point of the top gas. Thus, they either impose a thermally inefficient operation with higher top gas temperatures, or, if the top gas temperatures become too low, they cause condensation and corrosion in the gas lines and dedusting equipment. To maintain a good permeability of the blast furnace an excellent coke strength is important. Figure 9 shows a schematic representation of the coke degradation in a blast furnace [24]. The coke is attacked by mechanical load, by alkalis, and by gasification due to the Boudouard reaction. To guarantee the countercurrent principle of the process, the lumpy bell coke should change its grain size and grain form as little as possible up to the temperature of 1500 °C. Especially in the operation mode with low coke addition and high injection rates, the longer residence time of the coke in the very high temperature region and the supplementary abrasion by injected coal particles can enhance the coke degradation in the furnace bosh and at tuyere level [25]. Therefore, the coke must be stabilized, fissure free, and must have a cold strength I_{40} (IRSID-test) of more than 50 % plus 40 mm after tumbling.

In the hearth of the blast furnace the coke cone (dead man) must posses a sufficient permeability. Tapped volume of hot metal and slag and tapping time are closely related to coke size. Coke particles that are too small entail the formation of an impermeable inert central core in

the blast furnace, which deviates gas and liquid flow towards the walls of bosh and hearth. This can lead to disastrously rough furnace operation, wear of the refractory lining and breakouts [26].

High temperature properties of coke are generally expressed in terms of the CSR value (**C**oke **S**trength after **R**eaction with CO_2) and the CRI value (**C**oke **R**eactivity **I**ndex). Both values are measured in one single test procedure under gasification of a coke sample with CO_2 at 1100 °C. The CRI value is evaluated by analyzing the loss of coke weight and the CSR value expressed by the grain size portion > 10 mm after mechanical stress.

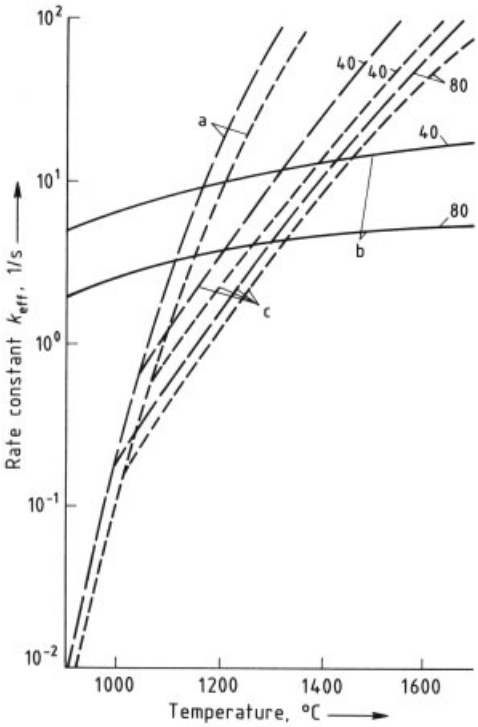

Figure 10. Limiting curves for the effective rate constant in the Boudouard reaction shown for two types of coke; – — coke type a; – – – coke type b (the curve parameter is the grain diameter in mm)
a) Chemical surface reaction; b) Boundary layer diffusion; c) Pore diffusion

Depending on coke temperature and particle size, the principal determinant in the rate of the Boudouard reaction

$$C + CO_2 \rightleftharpoons 2\,CO \tag{1}$$

is the varying resistance to reactivity. In the case of blast-furnace coke with a particle size between 40 and 80 mm, and at low temperature (< 1050 °C), the limiting factor is the resistance of the interfacial reaction. For particles < 8 mm in diameter the resistance of the interfacial reaction is rate-limiting up to ca. 1200 °C. Between 1050 and ca. 1350 °C the resistance of pore diffusion becomes the rate determining factor in the Boudouard reaction. Above 1350 °C the rate of reaction is governed by the diffusion of carbon dioxide through the gaseous layers adhering to the coke particles. The corresponding relationships are illustrated in detail in Figure 10.

Carbon conversion r_e in mol m^{-3} s^{-1} may be computed on the basis of the rate constant k_{eff} derived from Figure 10 [19]

$$r_e = \frac{dn_c}{dt} = k_{eff}\left(n^0_{CO_2} - n^{eq}_{CO_2}\right) \tag{2}$$

where $n^0_{CO_2}$ is the CO_2 concentration in the gas phase (mol/cm^3) and $n^{eq}_{CO_2}$ the CO_2 concentration in equilibrium with C and CO (Boudouard reaction).

Furthermore

$$k_{eff} = k_m \cdot m_C \cdot \eta \tag{3}$$

where m_C is the mass of carbon (g) per cubic meter of charge and η the pore utilization factor according to Thiele. Within the temperature range in which the interfacial reaction is rate determining, the rate constant is

$$k_m = H_C \cdot e^{-360/RT}\,g^{-1}s^{-1} \tag{4}$$

where H_C is the reactivity factor. Moreover, because $n = 1$

$$k_{eff} = m_C \cdot H_C \cdot e^{-360/RT}\,s^{-1} \tag{5}$$

The activation energy of 360 kJ/mol is uniformly applicable with respect to any fuel [27]. Depending on reactivity and charge size, the product $m_C \times H_C$ may range from 0.4×10^{13} s^{-1} to 0.4×10^{15} s^{-1}.

Reducing Gas. Assuming that iron oxide is present in the *solid state,* industrial-scale reduction of iron ore with carbon monoxide or hydrogen takes place according to

$$FeO_n + n\,(CO, H_2) \rightarrow Fe + n\,(CO_2, H_2O) \tag{6}$$

If iron oxide is present as a *liquid,* a second reaction is also possible:

$$FeO_n + nC \rightarrow Fe + nCO \quad (7)$$

Carbon monoxide is produced in the blast furnace from the exothermic reaction of oxygen in the hot blast air with hot coke and other reducing agents such as injected coal or oil

$$C + 1/2\,O_2 \rightarrow CO \quad (8)$$

Carbon monoxide is also the product of the endothermic Boudouard reaction of coke or coal carbon with CO_2 in the gas phase

$$C + CO_2 \rightleftharpoons 2\,CO \quad (1)$$

Carbon dioxide required by Equation (1) is formed during the reduction of iron oxide according to Equation (6). In some cases, *coal* is utilized as the reducing agent, which means that the gases actually responsible for reduction arise both from coal pyrolysis and from coal gasification inside the reduction reactor.

Coal pyrolysis may lead to varying quantities of hydrogen and hydrocarbons, which in turn serve partly as reducing agent and partly as energy carrier. The majority of the reducing gas in such a case originates from the solid carbon contained in the coal, where ore reduction and the Boudouard reaction are linked through Equations (1) and (6).

Gaseous reducing agents other than those formed in the reduction reactor itself are derived from natural gas, oil, or coal. Stoichiometrically, the large number of processes that are actually suitable for gas production, is limited significantly because a low content of steam and carbon dioxide as well as a specific hydrogen : carbon monoxide ratio is required in the reducing gas.

An appropriate gas for reduction above the decomposition temperature of carbon monoxide is that produced by steam reforming of natural gas. This gas has a composition of 73 % hydrogen, 13 % carbon monoxide, 1 % water, 8 % carbon dioxide and 5 % methane:

$$CH_4 + H_2O \rightarrow CO + 3\,H_2$$

In the *Midrex process* a 1.5 : 1 ratio of hydrogen and carbon monoxide is required. This ratio is achieved by catalytic reaction of natural gas with the CO_2 of the Midrex top gas in a reformer, which results in a better utilization of input energy:

$$CH_4 + CO_2 \rightarrow 2\,CO + 2\,H_2$$

The reducing gas produced in the reformer has a composition of approximately 56 % H_2, 36 % CO, 3 % CO_2, and 2 % CH_4. Reducing gas should contain as little steam and carbon dioxide as possible. Even relatively small amounts of these substances drastically limit the utility of the gas, because according to Equation (6) the reduction of iron oxide itself produces carbon dioxide and water. Their presence in the reduction gas would therefore shift the equilibrium to the side of iron oxide. At the temperatures normally employed for converting iron oxide to iron, the ratio $CO : CO_2$ or $H_2 : H_2O$ must not fall significantly below 2.3. In addition, reduction gas should contain as little undecomposed hydrocarbons as possible.

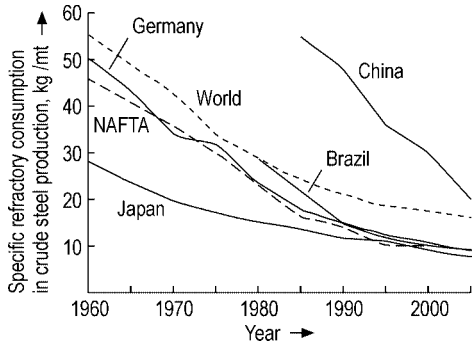

Figure 11. Steel industry – trend in specific refractory consumption

The use of fossil reducing agents in iron making is subject to careful reevaluation, and development of new reduction techniques is expected. It is quite likely that coal will come to play an increasingly important role as a source for production of reducing gas (see Section 2.7). This can be accomplished by partially combusting coal with oxygen in an iron bath for the production of hot metal in smelting reduction processes. The resulting hot gas is then used to reduce iron oxide in a separate vessel (see the equations above). Then the reduced iron is itself transferred to the iron bath where it is melted. Gas generation on the basis of nuclear energy is expected to remain economic for many years, although some effort is currently being directed toward the use of a

plasma arc for converting carbon sources into reducing gas.

2.1.3. Refractory Materials [27 – 29]

Refractory materials must meet the following requirements:

1) Mechanical stability under load at high temperatures (up to and beyond 1800 °C, depending on quality),
2) Resistance to slag and dust at high temperature,
3) Specific thermal properties, such as thermal conductivity, specific heat, and resistance to sudden temperature changes, depending on application.

These properties can be tested in the laboratory under conditions similar to the service conditions. Some methods are specified to internationally standards (PRE, i.e., Fédération Européenne de Produits Refractaires and ISO recommended test methods) [27]. Refractories are classified by chemical-mineralogical criteria and by process requirements. The basic distinction is between shaped products (bricks) and unshaped ones (mixtures, mortars, castables). Another basis for classification is the type of binding. Here a distinction must be made between fired products (firing temperatures 800 – 1800 °C) and unfired, chemically bonded refractories (using phosphate binders, resin binders, refractory cements, water glass, magnesium sulfate).

In the iron and steel industry in Germany, the classification of the "Steel and Iron Material Data Sheets" (Stahl – Eisen-Werkstoffblätter) is used [30].

Figure 11 shows the consumption of refractory materials (kg per tonne of crude steel) for the different regions worldwide. The steep decline is mainly accounted for by new steelmaking technologies (oxygen-injection process replacing the Siemens – Martin furnace, introduction of continuous casting) and by improvements in refractory quality [31 – 33].

This trend is also clear from Table 9, which lists the output of the main qualities of refractory bricks in Japan. The 1970 and 2003 figures show the decrease in fireclay brick and silicate brick production and the increase in high-grade refractories (magnesia-carbon, high-alumina castables).

The world steel industry is the principal consumer of refractories (ca. 60 %).

Pig-iron production includes (1) blast furnaces with hot blast stoves, and (2) torpedo ladles, with a molding capacity of 150 – 600 t of molten pig iron, as transport containers from the blast furnace to the steelworks. This sector mainly uses the following refractory qualities: carbon and graphite bricks, silicon carbide bricks, materials containing more than 50 wt % alumina, fireclay, and silica.

In steelworks the most important qualities are magnesia – carbon bricks, bricks and mixtures based on magnesia and dolomite, and high-alumina castables and bricks. Synthetic-alumina-based refractories have become more important. The trend towards high-performance refractories is ongoing due to increasingly severe process conditions.

Blast Furnaces. Molten pig iron and slag are present in the lower part of the blast furnace (bottom and hearth). Repairs cannot be made to this part of the furnace during operation. Chemical and mechanical attacks on the refractory material occur here. Graphite and carbon bricks are the most important structural refractories for this portion of the blast furnace. Carbon bricks have a higher cold-crushing strength than graphite bricks and those containing some graphite (part-graphite and semi-graphite bricks). Carbon bricks have a lower porosity than graphite bricks. Microporous carbon and part graphite bricks with additions of alumina and silicon have a higher resistance to alkali and pig iron [35 – 38].

The current state of the art is the "ceramic cup", which comprises either sintered mullite bricks or high-quality fireclay bricks (48 % Al_2O_3) placed in front of the carbon brick on the bottom, and either large prefabricated panels of corundum concrete or small SiAlON bonded corundum bricks in front of the carbon brick on the wall. This type of lining protects the carbon bricks and prolongs their service life greatly [39, 40]. In the other zones of the blast furnace (bosh, belly, stack), a variety of brick qualities are used. These include bricks containing over 50 wt % alumina, corundum bricks, and silicon carbide bricks with various binders. Hard fire-

Table 9. Production of refractories in Japan in 10^3 t [34]

	1970	1980	1990	2000	2003
Bricks:					
Fireclay	1800	829	325	132	86
High alumina	191	173	155	105	80
Silica	260	66	7	3	1
Chrome-magnesia	284	278	126	77	41
Magnesia-carbon		52	123	94	76
Doloma	251	104	13	6	3
Zircon	61	91	52	19	13
Silicon carbide		53	27	15	12
Insulants	88	55	27	11	5
Others	75	44	35	35	40
Unshaped products (mixtures)	570	907	836	782	734

Table 10. Properties of blast-furnace lining bricks

	Carbon	Microporous carbon (10 % Al_2O_3, 15 % Si)	Graphite	Mullite (70 % Al_2O_3)	Corundum concrete (90 % Al_2O_3)	Corundum brick SiAlON bonded (75 % Al_2O_3)
Bulk density, g/cm^3	1.53 – 1.57	1.63 – 1.73	1.55 – 1.60	2.45	3.30	3.20
Open porosity, %	14 – 18	14 – 18	25 – 28	15 – 18	9 – 12	13 – 15
Microporosity > 1 µm, %	8 – 11	≤ 2				
Ash, %	< 7	35 – 40	< 1.0			
Cold compression strength, N/mm^2	30 – 40	60 – 90	15 – 25	70	60	150
Thermal conductivity, W m^{-1} K^{-1}, at						
20 °C	3 – 8	17 – 20	130	2	5	5
500 °C	5 – 12	19 – 22	95	1.8	4	4
1000 °C	8 – 14	20 – 25	65	1.7	3	3

clay bricks are usually employed in the stack. For intermediate repairs of the stack, shotcreting materials are used. Here the material is wet mixed, pumped, and sprayed onto the surface of the wall. Furnaces often differ greatly in their refractory linings. The properties of some refractory structural materials are given in Table 10.

Runner and Taphole Mixtures. Tar-free unshaped products containing > 50 % alumina, silicon carbide, and carbon are used in cast, vibrated or rammed form. A large number of very different mixes are in use, including mixtures with spinel ($MgAl_2O_4$)-containing material for use at the metal level and products with up to 80 % SiC at the slag level.

Hot Blast Stoves. The standard lining consists of silica bricks in the dome (temperatures up to 1500 °C), bricks with >50 wt % alumina in the combustion chamber, and silica and fireclay bricks (depending on thermal stress) in the checkerwork shaft. Magnesia bricks have also been tried in place of silica bricks in the upper part of the checkerwork shaft.

Torpedo Ladles [41, 42]. In the lining design, the following factors relating to stresses on the refractory must be considered: temperature of pig iron (1450 – 1520 °C), distance transported, slag composition, and metallurgical reactions such as desiliconization, dephosphorization and desulfurization in the ladle.

Lining bricks mostly contain > 50 % Al_2O_3 (andalusite and bauxite bricks). If the slag has a $CaO : SiO_2$ ratio >1, unfired pitch-bonded dolomite bricks can also be used successfully. The increase in metallurgical reactions in torpedo ladles has led to new types of lining, which have been adopted in an attempt to extend the service life. This can range from 100 000 to 400 000 t pig iron transported. In zone lining,

resin-bonded corundum – carbon – silicon carbide bricks are used in the slag zone (especially in Japan), with resin-bonded andalusite – carbon bricks in the bottom. Linings differ from plant to plant, depending on service conditions.

2.2. Thermodynamic Principles

The following sections deal with the thermodynamic and kinetic principles of ore reduction and serve to provide an understanding of the process technology as well as a critical evaluation of the process with respect to cost effectiveness, product quality, and efficient utilization of the reducing agents.

2.2.1. Thermodynamic Principles of the Reduction Reactions [19, 43, 212]

The principal reactions that occur in the reduction unit are described by

$$\mathrm{FeO}_n \rightarrow \mathrm{Fe} + n\mathrm{O} \tag{9}$$

$$n\mathrm{O} + n\,(\mathrm{C,H_2,CO}) \rightarrow n\,(\mathrm{C,H_2,CO})\,\mathrm{O} \tag{10}$$

Richardson Diagram (Fig. 12). Taking into account that three oxides of iron exist, Equation (9) can be divided into the following steps

$$6\,\mathrm{Fe_2O_3} \rightarrow 4\,\mathrm{Fe_3O_4} + \mathrm{O_2} \tag{11}$$

$$4\,\mathrm{Fe_3O_4} \rightarrow 12\,\mathrm{Fe} + 8\,\mathrm{O_2} \tag{12}$$

$$4\,\mathrm{Fe_3O_4} \rightarrow 12\,\mathrm{FeO} + 2\,\mathrm{O_2} \tag{13}$$

$$12\,\mathrm{FeO} \rightarrow 12\,\mathrm{Fe} + 6\,\mathrm{O_2} \tag{14}$$

The transformation of 6 mol hematite into 12 mol iron requires an elimination of 9 mol O_2 in total. From these, 1/9 is separated between hematite and magnetite, 8/9 between magnetite and iron, 2/9 between magnetite and wustite, and 6/9 between wustite and iron. In other words, 1/3 of the oxygen is separated during the reduction of the higher oxides to wustite and 2/3 during the reduction of wustite to iron.

Oxidation of the reducing agent (Eq. 10) occurs according to the following reactions

$$2\,\mathrm{CO} + \mathrm{O_2} \rightarrow 2\,\mathrm{CO_2} \tag{15}$$

$$2\,\mathrm{C} + \mathrm{O_2} \rightarrow 2\,\mathrm{CO} \tag{16}$$

$$2\,\mathrm{H_2} + \mathrm{O_2} \rightarrow 2\,\mathrm{H_2O} \tag{17}$$

Equations (11) to (14) represent the dissociation reactions of the oxides. Assuming that the activities of the solid phases all equal unity, their equilibrium constants can be represented in the form

$$K_p = p_{\mathrm{O_2}} \tag{18}$$

Whenever the reactants or products are present in the form of liquid or solid solutions, the equilibrium constant of reaction (11), for example, is written as

$$K_p = \frac{a^4_{\mathrm{Fe_3O_4}}}{a^6_{\mathrm{Fe_2O_3}}} p_{\mathrm{O_2}} \tag{19}$$

where a is the thermodynamic activity.

The oxidation of *carbon monoxide* to carbon dioxide is described by Equation (15), the equilibrium constant of which is given by

$$\frac{1}{K_p} = \frac{p^2_{\mathrm{CO}}}{p^2_{\mathrm{CO_2}}} p_{\mathrm{O_2}} \tag{20}$$

Similar expressions are valid for the oxidation of the reducing agents hydrogen and carbon.

For a given temperature the partial pressure of oxygen for reaction (15) is determined by the ratio of carbon monoxide to carbon dioxide (see Eq. 20). The reduction of ore becomes thermodynamically possible by choosing the carbon monoxide : carbon dioxide ratio such that the oxygen partial pressure of the reducing agent is less than the dissociation oxygen pressure of the iron oxide. Similar considerations are valid for hydrogen as the reducing agent.

For *carbon oxidation* the ratio of the partial pressures of oxygen and carbon monoxide depends, for a given temperature, on the total pressure because of a change in the number of moles of the gaseous species during the reaction

$$2\,\mathrm{C} + \mathrm{O_2} \rightarrow 2\,\mathrm{CO}$$

For a given temperature and total pressure the oxygen partial pressure is fixed, provided the gas does not contain other components.

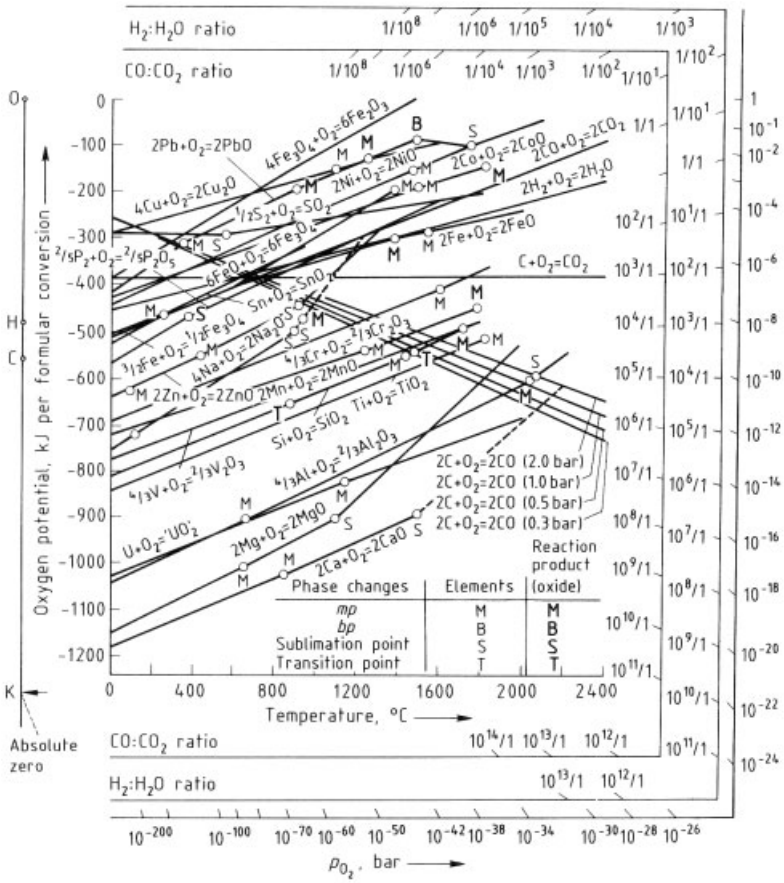

Figure 12. Temperature dependence and oxygen potentials of oxide systems, according to RICHARDSON and JEFFES [43]

The relationship between the free energy of a chemical reaction ΔG and the equilibrium constant K_p is given by

$$\Delta G = -RT \ln K_p \tag{21}$$

Combination with Equation (18) gives the free energy of the dissociation reaction of iron oxide

$$\Delta G = -RT \ln p_{O_2} \tag{22}$$

for the reducing agents the following expressions are valid

$$\Delta G_{CO} = -RT \ln \frac{p^2_{CO_2}}{p^2_{CO}} \cdot \frac{1}{p_{O_2}} \tag{23}$$

$$\Delta G_{H_2} = -RT \ln \frac{p^2_{H_2O}}{p^2_{H_2O}} \cdot \frac{1}{p_{O_2}} \tag{24}$$

$$\Delta G_{C} = -RT \ln p^2_{CO} \cdot \frac{1}{p_{O_2}} \tag{25}$$

Thus, in the case of iron oxides the equilibrium constant depends solely on the oxygen partial pressure. The same holds true for the reducing agents provided the ratios $CO : CO_2$ and $H_2 : H_2O$ are constant. Hence, the expression $-RT \ln K_p$ is called the oxygen potential (unit kJ/mol).

The oxygen potentials of the *dissociation reactions* of the most important oxides and those of the *oxidation reactions* of the reducing agents important in the context of this discussion are shown as a function of temperature in Figure 12 in the form of a Richardson diagram. The scales surrounding the diagram are used to evaluate the value of the oxygen partial pressure, or instead, the respective $CO : CO_2$ and the $H_2 : H_2O$ ratio. The points of the scales have to be connected with point O in the case of oxygen partial pressure, with point C in the case of $CO : CO_2$ ratio,

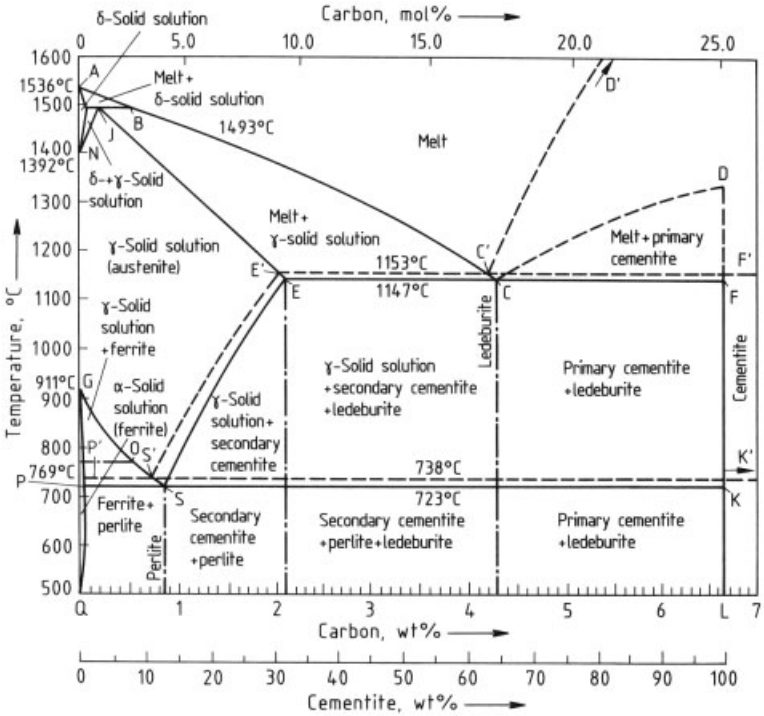

Figure 13. Phase diagram iron – cementite [43], broken lines refer to the stable system, full lines to the metastable system

and with point H in the case of $H_2 : H_2O$ ratio, respectively. The intersection of these lines with the free energy lines of the single metal – metal oxide system at the temperature of interest yield the equilibrium pO_2, $CO : CO_2$ and $H_2 : H_2O$ ratio of the respective reaction at the selected temperature. Considering for example the $H_2 : H_2O$ ratio, this means that any gas with a $H_2 : H_2O$ ratio in excess of the equilibrium ratio of the selected oxide can be used for reduction and vice versa.

With the aid of this diagram it can easily be determined whether a certain reaction can occur at a given temperature.

For example, iron oxide (FeO) can be reduced with chromium to produce chromic oxide (Cr_2O_3) and iron. On the other hand, silicon dioxide (SiO_2) cannot be reduced with either iron or chromium because at all temperatures the free energy of formation of silicon dioxide is more negative than that of iron oxide or chromic oxide.

It is particularly important that the free energy line pertaining to the reaction

$$2\,C + O_2 \rightarrow 2\,CO \qquad (8)$$

has a negative slope (opposite to that of all dissociation reactions). At temperatures > 1200 °C, the free energy line corresponding to reaction (8) intersects the lines pertaining to chromic oxide and manganese oxide, whereas the line corresponding to silicon dioxide is intersected at 1450 °C. These temperatures may be reached in the lower part of the blast furnace. Thus, it depends on its mode of operation to what extent these oxides are reduced as well.

When considering reduction processes based on carbon, it is important to recognize that the free energy lines of the reactions

$$2\,C + O_2 \rightarrow 2\,CO \qquad (8)$$

$$C + O_2 \rightarrow CO_2 \qquad (26)$$

intersect each other at ca. 700 °C if the pressure is 1 bar. At this point in the diagram $CO : CO_2 = 1$. As can be seen on the free energy line of reaction (8), at higher temperature and 1 bar $CO : CO_2 > 1$, i.e., carbon monoxide formation is favored. At lower temperatures $CO : CO_2 < 1$, i.e., carbon dioxide formation is favored.

To understand the properties of the product formed as a result of the reduction, (hot metal or sponge iron) the iron – cementite (Fe_3C) phase diagram is used (see Fig. 13). As may be seen from the phase diagram, the melting point of iron is lowered from 1540 °C to 1150 °C when 4.3 % carbon is dissolved in the iron. This lowering of the melting point by ca. 400 °C is very important for the blast furnace process. For reduction processes in which solid metallic iron is produced and the reducing agents are carbon or carbon monoxide, the solubility of carbon in α- and γ-iron is important. This control of the carbon content during the refining process subsequent to the reduction step is important to produce the required steel properties. The maximum solubility of carbon in α-iron is 0.03 wt % and 2.1 wt % in γ-iron.

Baur – Glaessner Diagram. This diagram (Fig. 14) may be derived from the Richardson diagram (Fig. 12). In the Baur – Glaessner diagram the stability areas of the iron oxides and iron are depicted as a function of temperature and the composition of the carbon monoxide – carbon dioxide mixtures. For the purpose of an easy evaluation of the process parameters this diagram is preferred to that depicted in Figure 12 because in Figure 14 the gas composition is expressed in volume percent rather than in terms of partial pressure. In addition to the stability areas of the iron oxides the course of the Boudouard reaction

$$2\,CO \rightleftharpoons CO_2 + C \qquad (1)$$

which is obtained by combining Equations (8) and (26) is also indicated in Figure 14. The position of the curve is derived from the data in Figure 12. At low temperature and high concentration carbon monoxide (underneath the Boudouard curve) disproportionates into carbon dioxide and carbon. At high temperature and high concentration carbon dioxide reacts with carbon to produce carbon monoxide.

According to thermodynamic considerations the reduction of iron oxides with carbon monoxide to metallic iron should not be possible below 700 °C (see Fig. 14). Carbon monoxide should disproportionate instead into carbon and carbon dioxide because the free energy of this reaction is higher than that of the reduction reaction FeO + CO → Fe + CO_2. However, nucleation of solid carbon which is necessary for the disproportionation reaction is suppressed. Instead, the reduction reaction is favored and occurs in spite of its lower thermodynamic possibility. Above the Boudouard curve carbon monoxide is regenerated via the reaction of carbon with the carbon dioxide produced in the reduction process. In this region carbon is the final reducing agent. Each formula conversion results in a doubling of the number of moles in the gas phase.

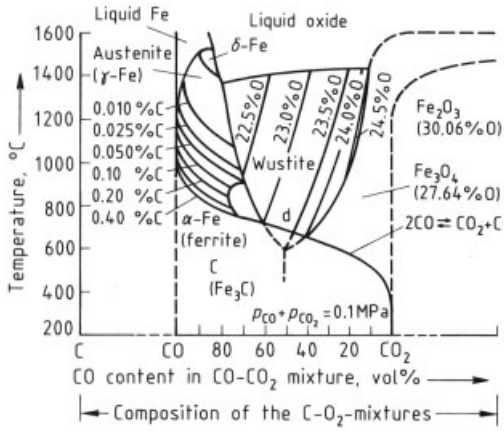

Figure 14. Baur – Glaessner diagram with Boudouard curve [49]

The change in the number of gas molecules as a result of the reaction implies that equilibrium depends on total pressure. The Boudouard curve in Figure 14 is shifted upwards when the pressure is increased.

Lines of equal carbon concentration in γ-iron are also indicated in Figure 14. The curve for the highest carbon concentration corresponds to the saturation boundary ES in the iron – cementite diagram (Fig. 13). On the Boudouard curve in Figure 14 iron is saturated with carbon because the carbon activity, a_C, equals unity on this curve. That is, the carbon content on the Boudouard curve corresponds to the line ES. The carbon content of iron decreases when, for a given temperature, the gas composition corresponding to the Boudouard equilibrium, ($p_{CO,B}$ and $p_{CO_2,B}$) changes into a gas with a lower carbon monoxide content ($p_{CO,l}$ and $p_{CO_2,l}$). The following expression gives the resulting carbon content of the iron, $c_{C,l}$ as obtained from the car-

bon content at saturation, $c_{C,s}$ assuming a constant activity coefficient:

$$c_{C,1} = \frac{p^2_{CO,1}}{p_{CO_2,1}} \frac{p_{CO_2,B}}{p^2_{CO,B}} c_{C,s} \quad (27)$$

Reduction with Gas Mixtures. The extended Baur – Glaessner diagram (Fig. 15) is important for reduction processes based on mixtures of carbon monoxide and hydrogen containing larger or smaller quantities of carbon dioxide and steam. Below 573 °C magnetite is directly reduced to metallic iron. However, the required hydrogen content in the gas mixture in question is higher than the carbon monoxide content. Above 573 °C magnetite is always reduced to wustite. The concentrations of reducing agents in the gas mixture required for the conversion wustite – iron are a function of temperature. The required hydrogen content of the reducing gas decreases with increasing temperature whereas the required carbon monoxide content increases.

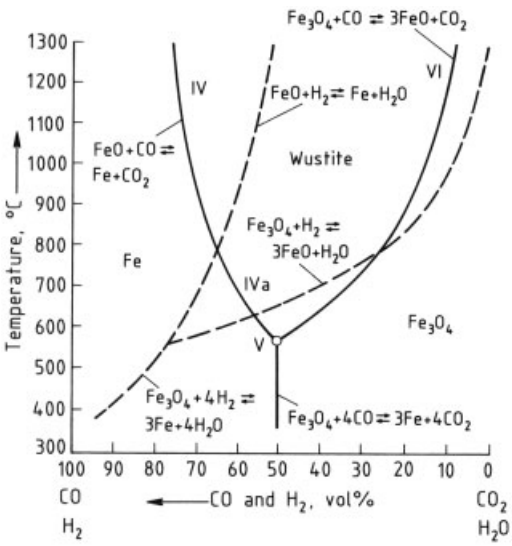

Figure 15. Baur – Glaessner diagram for carbon monoxide – carbon dioxide and hydrogen – water vapor mixtures [43]

From a more technological point of view it might be useful to consider only one Baur – Glaessner curve for the four-component gas mixture of CO, H_2, CO_2, and H_2O, and to use the (CO + H_2) or the (CO_2 + H_2O) content as the abscissa of the diagram. Baur – Glaessner curves for this case can be calculated [213]. This is described as follows for the equilibrium curve Fe/FeO.

With the volumetric fraction of the gas components in the mixture denoted by x, the following definitions are introduced:

$$x_{CO_2} + x_{CO} = k_C$$

$$x_{H_2O} + x_{H_2} = k_H$$

$$x_{H_2O} + x_{CO_2} = Y$$

k_C and k_H are the respective total contents of the elements carbon and hydrogen in the gas. Y represents the new abscissa in the Baur – Glaessner diagram. As the mole number does not change during reduction, k_C and k_H are constant for a given reducing gas and independent of the CO : CO_2 or H_2 : H_2O ratio in the gas.

Further:

$$\frac{x_{CO_2}}{x_{CO}} = K_{Fe,CO_2}$$

and

$$\frac{x_{H_2O}}{x_{H_2}} = K_{Fe,H_2O}$$

where K_{Fe,CO_2} and K_{Fe,H_2O} are the respective equilibrium constants for reduction of wustite with CO and H_2 (see Eq. 30). It follows that

$$\frac{x_{CO_2}}{k_C - x_{CO_2}} = K_{Fe,CO_2}$$

and

$$\frac{x_{H_2O}}{k_H - x_{H_2O}} = K_{Fe,H_2O}$$

These two equations are resolved with respect to x_{CO_2} and x_{H_2O}, respectively, and inserted into the above equation for Y

$$Y = k_C \frac{K_{Fe,CO_2}}{1+K_{Fe,CO_2}} + k_H \frac{K_{Fe,H_2O}}{1+K_{Fe,H_2O}}$$

This is the equation for the temperature-dependent Fe : FeO equilibrium curve in the four-component system. The equations for the FeO : Fe_3O_4 and Fe : Fe_3O_4 equilibria are of the same type, only with different equilibrium constants.

Iron – Oxygen Phase Diagram. The stability areas of the various phases are depicted in Figure 16. Information contained in this diagram has been partially incorporated in Figures 12 and 14. In addition, details may be obtained relating to the dissociation of hematite into magnetite and oxygen, the liquid phases that may be formed and the relationships between iron and wustite. It should be stressed that the phases depicted in the phase diagram are not in equilibrium with air but only with an atmosphere of distinct composition, the oxygen partial pressure of which may be obtained from Figure 17. The composition of the reducing gas corresponding to the given oxygen partial pressure may be obtained from the Richardson diagram.

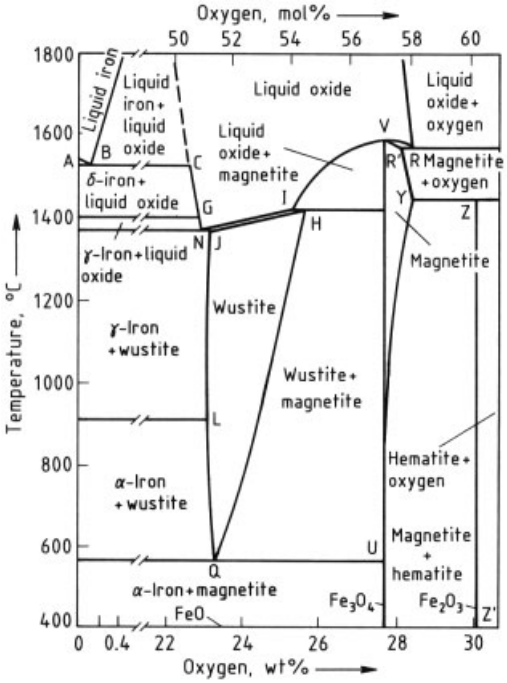

Figure 16. Phase diagram for iron – oxygen, according to DARKEN and GURRY [43]

Lines of equal iron activity are also indicated in Figure 17. The iron activity of the reaction

$$2\,Fe + O_2 \rightarrow 2\,FeO$$

is calculated from the oxygen potential obtained from the Richardson diagram

$$RT \ln K_{FeO,O_2} = RT \ln \frac{a_{FeO}^2}{a_{Fe}^2 \cdot p_{O_2}} \tag{28}$$

assuming $a_{FeO} = 1$.

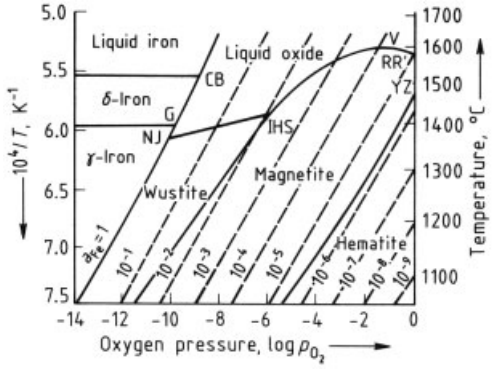

Figure 17. Oxygen pressure in equilibrium with iron and its oxides as a function of temperature, according to DARKEN and GURRY [43]

Reduction of Oxides in Solid Solution. As already mentioned (see Eqs. 19 and 28), the oxygen partial pressure p_{O_2} depends on the activities of the solid phases. In stable mixed oxides the activities of the iron oxides are less than unity. For example, if iron oxide (FeO) is to be reduced to iron from fayalite ($2\,FeO \cdot SiO_2$), then $a_{FeO} < 1$ in Equation (28). Thus, p_{O_2} is smaller than for the reduction of pure iron oxide. Hence, the carbon monoxide content of the reducing gas used for the reduction of fayalite must be higher than that of the gas used for the reduction of wustite, as may be seen from Equation (20a).

$$p_{O_2} = \frac{1}{K_p} \left(\frac{p_{CO_2}}{p_{CO}} \right)^2 \tag{20a}$$

Similar considerations apply to hydrogen – steam mixtures.

The thermodynamic relationships of the reduction of mixed oxides have not yet been studied as extensively as those pertaining to the pure oxides. Thus a Richardson diagram cannot be presented for the reduction of mixed oxides.

For industrial reduction processes the interactions of silicon dioxide and calcium oxide with the iron oxides are important. The solid phases expected to be formed during reduction and the minimum carbon monoxide content of the reducing gas that is required to produce metallic

iron are shown in Figures 18 and 19. Examples of the possible reduction paths are presented in the diagrams.

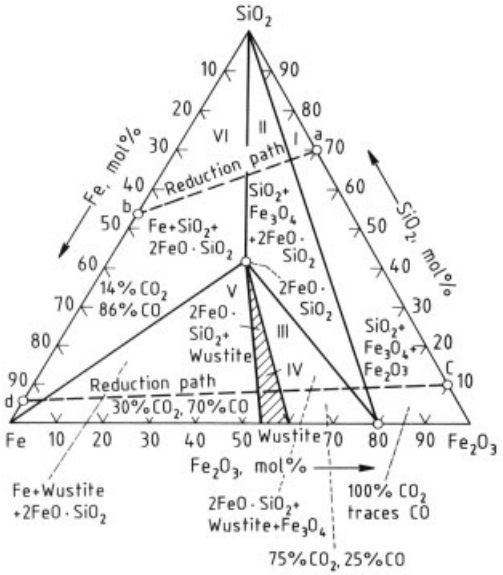

Figure 18. System Fe – Fe_2O_3– SiO_2 at 900 °C, showing examples of reduction paths [43]

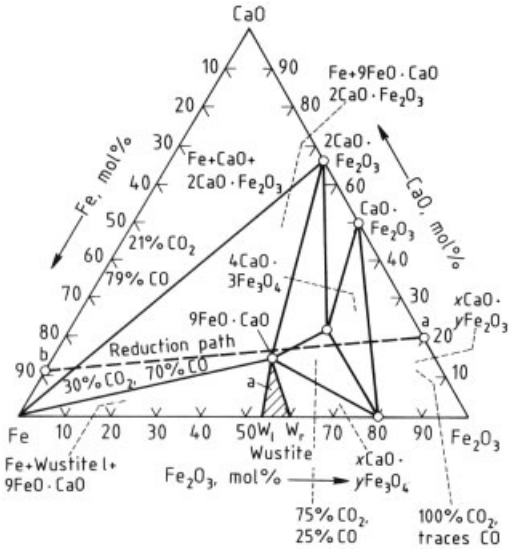

Figure 19. System Fe – Fe_2O_3 – CaO at 900 °C, showing example of reduction path [43]

Thermodynamic Considerations Concerning the Extraction Processes. From a thermodynamic viewpoint control of the oxygen potential inside the reactor is the only requirement to ensure that the reactions to produce iron and steel proceed. This requirement is more than fulfilled in the blast furnace, because except iron, also silicon, manganese, and phosphorus are reduced to a small degree, and the produced liquid iron also dissolves carbon until the limit of saturation. The oxygen potential of the hot metal is, therefore, much lower than required for iron production alone. In the subsequent oxygen converter, the accompanying elements in the hot metal must be separated by oxidation for the production of steel. The crude steel itself has finally, to be deoxidized. On the way from ore to steel, the attained final oxygen potential is thus, first falling below and then, increasing above the desired value. This circumstance has led to proposed processes in which this circuitous path is avoided. Two alternatives are depicted in Figure 20.

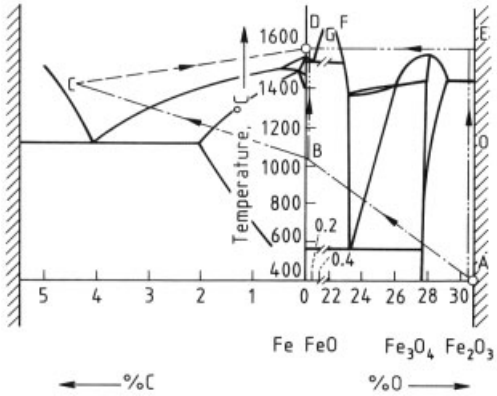

Figure 20. Representation of some reduction processes on the iron – iron oxide and iron – cementite phase diagrams [45]

The iron – iron oxide and iron – cementite phase diagrams, sharing a common temperature axis, are shown in Figure 20. The curve ABCD represents the conventional path to produce steel. AB is reduction to iron, BC carbon dissolution, and CD separation of carbon – and other accompanying elements, which is not shown here – in the oxygen converter. Alterna-

tives avoiding a too strong reduction as in the blast furnace are also shown. The path ABD represents, in principle, the reduction sequence in the direct reduction processes. The oxide oxygen is removed from the solid phase to low residual levels while the temperature increases (point B). Subsequently, the reduced material is melted (curve BD) and refined to steel. The curve AED represents a processing sequence firstly proposed by Ketorp [45], presently known as smelting reduction. According to this sequence, the oxide is melted and in the liquid state reduced with carbon until the system reaches the two-liquid region. Here, iron containing ca. 0.1 % oxygen but no carbon, silicon, phosphorus, or manganese is in equilibrium with a liquid slag still containing a few percent of iron oxide. After separation of the liquid iron, the content of dissolved oxygen is decreased by means of carbon or metallic deoxidizers.

Thermodynamic Boundary Conditions of the Reduction Stoichiometry. It follows from the discussion of the thermodynamic principles that the stoichiometric boundary conditions which determine the mass balances pertaining to the reduction processes, differ from those based on the simple reaction

$$FeO_n + n\,CO \rightarrow Fe + n\,CO_2 \tag{6}$$

In addition, the mass balances are more complicated as a result of the existence of the three oxides of iron.

For example, by making use of the Baur – Glaessner diagram (Fig. 15) it is found that the reaction

$$FeO + 3\,CO \rightarrow Fe + 2\,CO + CO_2 \tag{6a}$$

describing the reduction of wustite with carbon monoxide at the usual reduction temperature of 900 °C, can proceed only until a carbon dioxide content of 31 % in the $CO - CO_2$ gas mixture, as at this content the equilibrium

$$K_{FeO} = \sqrt{\frac{K_P}{K_{FeO,O_2}}} = \frac{p_{CO_2}}{p_{CO}} \tag{29}$$

(see Eqs. 20 and 28 with a_{Fe} and a_{FeO} = 1) is attained. That means that only a little less than one-third of the total amount of CO can be utilized. The remaining gas can, however, be used to reduce Fe_2O_3 to FeO, if the process is carried out in countercurrent flow, i.e., if carbon monoxide or – in the case of the blast furnace – a carbon monoxide – nitrogen mixture is directed upwards from bottom to top through a burden of iron oxide in a shaft. Under these conditions FeO is reduced to Fe in the lower part and Fe_2O_3 is reduced to FeO in the upper part of the burden. The reduction of Fe_2O_3 via Fe_3O_4 to FeO is not restricted by thermodynamic limits because the CO_2 content in the $CO_2 - CO$ mixture in equilibrium with Fe_3O_4 is 82 % at 900 °C and 74 % at 700 °C and in equilibrium with Fe_2O_3 is 100 % (see Fig. 15). According to the 100 % CO_2 in equilibrium with Fe_2O_3, a separate equilibrium line for this oxide is not shown. The amount of oxygen separated from 1 mol iron in the hematite – wustite reduction stage is half of that separated from the iron in the wustite – iron reduction stage (see Section page 33). Since the maximum possible carbon monoxide utilization at FeO reduction is 31 % at 900 °C due to the aforementioned thermodynamic condition, a further amount of $\frac{1}{2} \cdot 31\,\%$ = 15.5 % carbon monoxide can be utilized in the hematite → magnetite → wustite reduction stage. The total possible utilization is, therefore 31 + 15.5 = 46.5 %. This is the theoretical optimum if equilibrium is attained. The same considerations hold principally if $CO - H_2$ mixtures are used. Only the equilibrium gas composition is different and depends on the $CO : H_2$ ratio (see Fig. 15 and following text).

The described restrictions also hold when iron oxide is completely reduced by gas as it is the case in the direct reduction processes which use natural gas as the primary energy source (see Section 2.6). The off-gas of these processes consisting of a $CO - H_2 - CO_2 - H_2O$ gas mixture may, however, be used partially for combustion in order to supply the heat for the endothermic conversion of natural gas into the reducing gas ($CO + H_2$), and partially for the reforming reaction itself:

$$CH_4 + CO_2 \rightarrow 2\,CO + H_2$$

and

$$CH_4 + H_2O \rightarrow CO + 2\,H_2$$

The total gas utilization in that process type is, therefore, high.

If carbon is the primary fuel for the reduction process as for example in the blast furnace,

the gas utilization may be improved above that of pure gas reduction processes, because part on the wustite is reduced by the so-called direct reduction:

$$FeO + C \rightarrow Fe + CO \quad \text{(Eq. 7 with } n = 1\text{)}$$

This decreases the carbon monoxide volume per mole of iron required for reduction of the remaining wustite according to Equation (6a). Moreover, the percentage of carbon monoxide required for reduction of the higher oxides increases above one-third of the entire requirement, and hence, gas utilization rises (for details see Section 2.2.2).

2.2.2. Mass and Heat Exchanges of the Reduction Reaction with Carbon as the Primary Fuel

Two-Stage Reduction. The most important primary energy source for iron oxide reduction is coal. The following considerations are thus restricted to carbon, in order to simplify the fundamental relations. The reduction reaction itself proceeds always with gas, as only a gas can contact the entire – outer and inner – surface of the ore, and hence, enables the fastest reduction. The reduction gas carbon monoxide is produced by gasification of carbon. Gasification proceeds entirely or partially according to the Boudouard reaction (Eq. 1). Together with gas reduction (Eq. 6a) it yields the direct reduction reaction:

$C + CO_2 \rightarrow 2\,CO$	Boudouard reaction	(1)
$CO + FeO \rightarrow CO_2 + Fe$	Reduction reaction	(6a)
$C + FeO \rightarrow CO + Fe$	Direct reduction	(30)

Direct reduction is carried out at temperatures $> 800\ ^\circ C$, usually at ca. 1200 °C, because at this temperature the Boudouard equilibrium lies on the CO side and the product of direct reduction is iron, saturated with liquid carbon.

Carbon monoxide produced by the direct reduction reaction can be used to reduce further amounts of iron oxide according to Equation (6a), as mentioned. Therefore, reduction with coal is usually carried out in two stages, namely direct reduction above 800 °C and prereduction or indirect reduction, below 800 °C. This temperature can vary in a certain range depending on the coal reactivity.

Direct reduction is a strongly endothermic process due to the endothermic character of the Boudouard reaction, whereas indirect reduction is weakly exothermic. For this reason the two-stage reduction with coal needs supply of heat. This is possible in principally three ways:

1) Combustion of an additional amount of carbon with oxygen (or air) to carbon monoxide
2) Post-combustion of carbon monoxide to carbon dioxide
3) Supply of electrical energy

Combustion of an Additional Amount of Carbon with Oxygen (or Air) to Carbon Monoxide. This method is the most important. It is applied in the blast furnace and in smelting reduction processes.

The combustion of additional carbon with oxygen (or air) gives a certain amount of carbon monoxide as the combustion product. This amount together with the carbon monoxide generated by the direct reduction reaction gives a total amount of carbon monoxide which is then used for reduction of the remaining FeO and the higher oxides Fe_3O_4 and Fe_2O_3 in the prereduction stage. The total amount of produced carbon monoxide per tonne of iron must be the same as that consumed in the prereduction stage. In the blast furnace this occurs automatically because both stages are united within one process. In two-stage prereduction and smelting reduction processes this condition must be fulfilled in order to obtain the optimum operating point of the process in total. Figure 21 [214] shows this for the case that direct reduction is carried out with pure carbon and oxygen as the combustion medium; and that prereduction proceeds in a countercurrent shaft under the condition that equilibrium is attained at the position in the shaft where FeO reduction starts. The respective volumes of carbon monoxide consumption and production per tonne of iron are plotted against the degree of prereduction. Production and consumption are equal where the lines intersect. This is the case at 74 % prereduction. The point designates optimum operating conditions and at the same time the limit between the prereduction and the smelting reduction stage of the process. In the upper part of the figure, the carbon monoxide utilization at the end of reduction, i.e., at the top of the prereduction shaft, is shown.

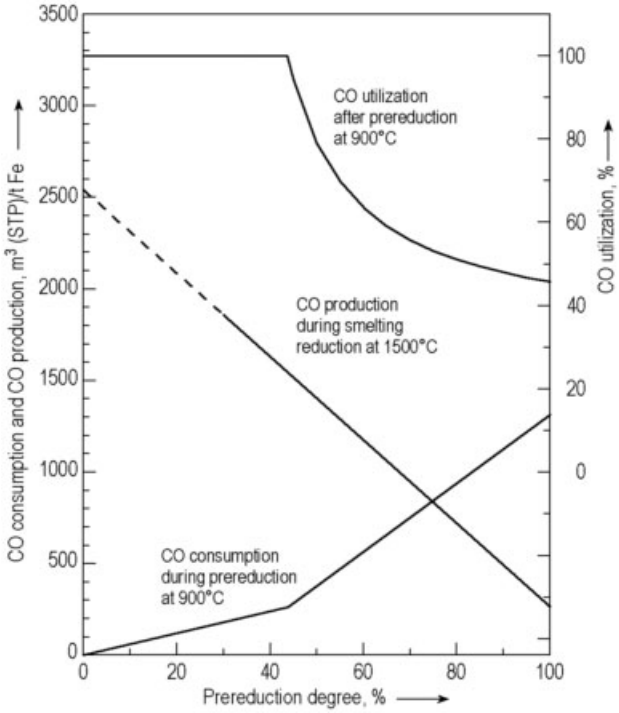

Figure 21. Optimum prereduction degree of a combined pre- and smelting reduction process with carbon. Prereduction at 900 °C [214]

The lowest utilization occurs if prereduction is 100 %, i.e., if the FeO is entirely reduced in the shaft. If FeO is partially reduced in the smelting reduction stage, gas utilization improves. Below prereduction degrees of 43 % gas utilization is 100 %, i.e., the off-gas consists to 100 % of CO_2. But in this case less gas would be required for prereduction than is actually generated in the smelting reduction stage at this degree of prereduction. The process is no longer balanced and excess carbon monoxide is generated. The way to overcome this difficulty is smelting reduction with application of post-combustion.

Post-combustion. Post-combustion of carbon monoxide to carbon dioxide and transfer of the thus generated heat to the direct reduction stage is the second method to balance the heat requirements in the direct reduction stage.

This method has been used for a long time in the rotary kiln direct reduction process [215]. Post-combustion with oxygen is also used in smelting reduction processes [216] (see page 137). The carbon monoxide, generated from the direct reduction reaction (30) is post-combusted above the melt with oxygen blown in from the top, and the combustion heat is transferred back to the melt. Instead by combustion of additional carbon to carbon monoxide, the heat requirement for direct reduction is supplied here by combustion of the produced carbon monoxide to carbon dioxide

$$CO + \frac{1}{2}\,O_2 \rightarrow CO_2$$

By this method the total enthalpy of CO_2 formation from carbon and oxygen is released, while in the first case it is only the enthalpy of CO formation. Therefore, carbon consumption decreases and hence, the amount of gas per tonne of iron produced in the smelting reduction stage. As a consequence, the intersection point of the curves of gas production and gas consumption in Figure 21 is shifted to lower prereduction degrees and gas utilization becomes higher. However, this advantage can only be exploited, if the post-combusted gas produced in the smelting reduction stage consists exclusively of carbon monoxide. Otherwise, the known thermodynamic limit prevents the reduction of wustite

to iron. It follows, that the CO_2 part in the post-combusted gas must be washed out before the gas enters the prereduction stage. The situation changes, when instead of wustite magnetite and hematite are reduced in the prereduction stage, i.e., if the prereduction degree is below 35 %. The thermodynamic limit of 33 % CO_2 in the CO – CO_2 mixture valid for the wustite – iron equilibrium at 800 °C vanishes, and only the limit of the magnetite – wustite equilibrium (76 % CO_2 in the CO – CO_2 mixture at 800 °C) remains (see Fig. 14). The degree of post-combustion must be ca. 52 % to cover the heat demand for smelting reduction of the 35 % prereduced ore. Since 52 % CO_2 is less than 76 % CO_2, that gas is able to reduce magnetite without CO_2 washing. Gas utilization at the top of the prereduction stage is 82 % in that case. That means an almost complete gas utilization and a correspondingly low fuel consumption can be attained by this operation.

Electrical Energy as the Heat Supplier. If electrical energy supplies the demand of heat for direct reduction, carbon is only consumed for the direct reduction reaction. No carbon combustion for heat generation as in the blast furnace is required. Carbon consumption and the corresponding gas production is, therefore, less than in the blast furnace. This way of operation was practiced in the so-called Plasmamelt process [217] (see Section 2.7.1).

Rist Diagram. The Rist diagram (see Fig. 22) is a useful tool for the evaluation of the complex problems related to the subject of mass balances and thermodynamic limits [218]. For a given reduction process, the diagram shows the exchange of oxygen between iron oxide and the reducing agent. The degree of oxidation of carbon is indicated on the abscissa by the molar ratio oxygen : carbon. The positive part of the ordinate gives the degree of oxidation of iron in terms of the molar ratio oxygen : iron. Here hematite is represented by $FeO_{1.5}$, magnetite by $FeO_{1.33}$ and wustite by $FeO_{1.05}$. The negative part of the ordinate indicates the molar ratio of oxygen from other sources to iron. In the blast furnace such "other source" is, for example, the hot blast used for the oxidation of coke, in smelting reduction processes it is the oxygen blown in for combustion of carbon to carbon monoxide.

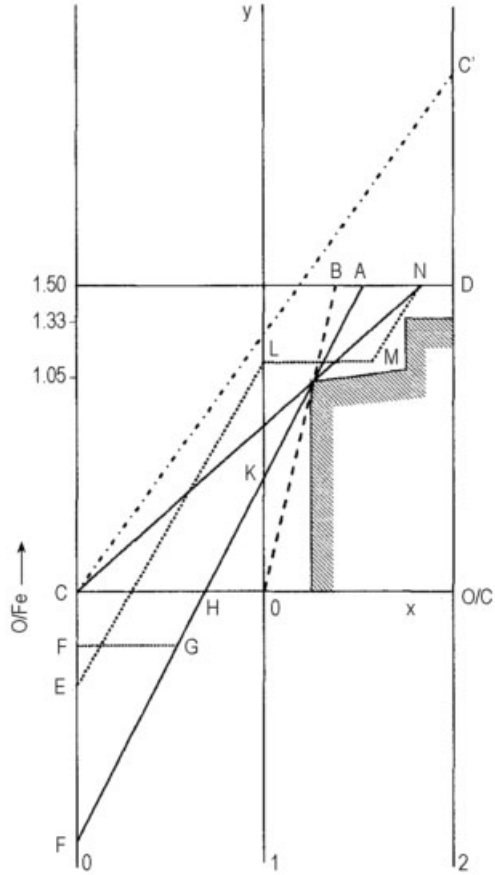

Figure 22. Representation of various reduction processes in the Rist diagram [45]

The blast furnace process is represented by the line FA in Figure 22. The oxidation of carbon to produce the process heat is represented by GH, carbon is oxidized from S to G. The section FG represents the reduction of oxides such as SiO_2 and MnO.

The direct reduction of iron oxide takes place on the section HK, after which carbon is completely oxidized to carbon monoxide. Indirect reduction proceeds along the line KA which touches point W (wustite) whenever the gas utilization is optimum. Point A is then located about halfway between O/C = 1 and O/C = 2, corresponding to a gas utilization of 50 %. In practice the gas utilization is closer to 40 %. The line FA is then steeper and no longer touches point W.

The slope (C/Fe) of a process line in the Rist diagram is an indication of the specific carbon or reducing agent demand of the process. For the blast furnace process the specific carbon demand is ca. 1.9 mol carbon per mole of iron.

The process represented by the line CC' is the rotary kiln direct reduction process, here for the case of hematite reduction. As no carbon is oxidized as a heat supplier before reduction and no SiO_2 and MnO, etc., is reduced, the line starts at point C. About 75 % of the iron is reduced by direct reduction (line CL). A CO – CO_2 mixture with 80 % CO leaves the charge and is combusted to 100 % CO_2 (point C'). The heat released from this oxidation satisfies the energy requirement of the reduction reaction. According to the degree of direct reduction, the specific carbon demand is 0.8 – 1.5 mol carbon per mole of iron for hematite and 0.8 – 1.33 mol carbon per mole of iron for magnetite. The line CN represents the case if the heat for direct reduction is supplied by electrical energy. Gas utilization is limited to 82 % because the curve contacts the wustite – iron equilibrium point W. The portion of direct reduction is somewhat higher than 50 %. Reduction of SiO_2 and MnO is not considered. Line ELMN presents the case of a combined pre- and smelting reduction process with 52 % post-combustion of the off-gas from the smelting reduction reactor. Section LM represents the post-combustion process. Prereduction is 25 %, gas utilization 82 %. It is assumed in Figure 22 that no oxygen is employed for the combustion of carbon to carbon monoxide. Heat is supplied only by the post-combustion reaction. In practice, usually part of the oxygen blown in is used for the combustion of carbon to carbon monoxide. Line 0B describes a process with 100 % reduction by gas according to

$$3\ (CO+H_2) + Fe_2O_3 \rightarrow 3\ (CO_2+H_2O) + Fe$$

Because of the wustite – iron equilibrium, gas utilization is low here.

2.2.3. Thermodynamic Principles of Metal – Slag Reactions [43, 46]

In reduction processes in which liquid iron is produced the slag plays an important role. It should

1) absorb those components contained in the raw material that are undesirable in iron,
2) be easily separable from the metal because of its low viscosity, low density and the low vapor pressure of its components,
3) be largely inert with respect to refractory materials although the slag consists primarily of metal oxides like many refractory substances,
4) have properties suitable to make it a saleable product.

Individual Reactions. The most important metal – slag reactions discussed here are:

$$(FeO) \rightleftharpoons [Fe] + [O] \tag{31}$$

$$(SiO_2) \rightleftharpoons [Si] + 2\,[O] \tag{25}$$

$$(MnO) \rightleftharpoons [Mn] + [O] \tag{33}$$

$$(3\,CaO{\cdot}P_2O_5) \rightleftharpoons 3\,(CaO) + 2\,[P] + 5\,[O] \tag{34}$$

$$[S] + (CaO) \rightleftharpoons (CaS) + [O] \tag{35}$$

$$[C] + [O] \rightleftharpoons \{CO\} \tag{36}$$

(Round brackets refer to the slag phase, square brackets to the metal phase and curly brackets to the gaseous phase.)

The equilibrium of each of these six reactions depends on the oxygen activity in the metal bath. The oxygen activity in the lower part of the blast furnace – for which these considerations are valid – is for all practical purposes constant to the extent that equilibrium is established. The oxygen activity in this equilibrium is, however, much higher than its activity in the equilibrium of reaction (36). This is easily to understand if it is taken into account that the reduction reactions (31) – (35) with coke (according to Eq. 36) start with high oxygen potential in the initially FeO rich slag while liquid metal and slag flow down through the burden of solid coke. The reactions cease when the liquid phases come below the plane of the tuyeres, since below that plane no gas phase exists any more. The absence of the gas phase prevents a further progress of reaction (36), and thus, the oxygen activity becomes constant. Equilibrium with carbon and carbon monoxide is not yet attained. In the hearth no more than exchanges among reactions (31) –

(35) are possible until the slag – metal equilibrium with all the slag and metal components and with one constant oxygen activity is established.

The free enthalpies of formation for the reactions (31) – (36) are expressed as follows:

$$-\Delta G_{FeO} = RT \ln \frac{a_{[Fe]}}{a_{(FeO)}} a_{[O]}$$

$$-\frac{1}{2}\Delta G_{SiO_2} = RT \ln \frac{a_{[Si]}}{a_{(SiO_2)}^{1/2}} a_{[O]}$$

$$-\Delta G_{MnO} = RT \ln \frac{a_{[Mn]}}{a_{(MnO)}} a_{[O]}$$

$$-\frac{1}{5}\Delta G_{3\,CaO\cdot P_2O_5} = RT \ln \frac{a_{(CaO)}^{3/5} a_{[P]}^{2/5}}{a_{(3\,CaO\cdot P_2O_5)}} a_{[O]}$$

$$-\Delta G_{CaS} = RT \ln \frac{a_{(CaS)}}{a_{(CaO)} a_{[S]}} a_{[O]}$$

$$-\Delta G_{CO} = RT \ln \frac{a_{[C]}}{p_{CO}} a_{[O]}$$

From these equations it can be derived that the reduction of components less noble than iron (Mn, Si, and P) into the hot metal, is enhanced as long as their activities in the metal phase are much lower than unity. Sulfur is mainly taken up by the slag as long as the slag has a high activity of CaO and the oxygen activity $a_{[O]}$ is low. Besides the above-mentioned elements, reduction of lime and alumina as important slag components is also possible in principal. However, their free enthalpies of formation are so large, that Ca and Al contents present in the hot metal remain below the limit of analytical proof. Thus, these reduction reactions can be neglected.

When considering reactions (31) to (36) individually the following should be noted.

Equation (31): For a given temperature the equilibrium *(FeO)-content of the slag* depends on the activity of iron oxide that, in turn, depends on the other slag components. This occurs because the iron activity in the hot metal equals unity while the oxygen content [O] is also fixed. Since the line $2\,Fe + O_2 \rightleftharpoons 2\,FeO$ in the Richardson diagram lies well above the line $2\,C + O_2 \rightleftharpoons 2\,CO$, the iron oxide reduction equilibrium is shifted to the far right, resulting in low (FeO)-contents in the slag.

Equation (32): The *reduction of silica* is of particular interest when discussing metal – slag reactions. This reaction is very sensitive to changes in temperature and, hence, the oxygen potential, as may be seen from the Richardson diagram. The extent of the silica reduction is used to judge the extent to which the other reactions have proceeded. This is a consequence of the circumstance that the silicon equilibrium is neither shifted entirely to the silicon [Si] side nor entirely to the silicon dioxide (SiO_2) side and because the (SiO_2) contents of the slag are certainly high, but SiO_2 is only weakly reduced due to its high free energy of formation. Silicon contents are usually above 0.2 % and mostly below 1 %, and in this range can easily and precisely be determined by the usual analytical methods. For blast furnace control purposes the [Si] and (SiO_2) contents represent important control quantities that describe the conditions in the lower part of the furnace.

Equation (33): The conditions that determine the *manganese content* [Mn] of the hot metal follow from considerations similar to those that determine the silicon content. However, under the conditions prevailing in the blast furnace the manganese equilibrium is shifted more to the metal side, because manganese is more noble than silicon, as may be seen from the Richardson diagram.

Equation (34): The *reduction of apatite* is complete and phosphorus is transferred entirely to the hot metal.

Equation (35): Because the *sulfur content* of the hot metal should be as low as possible, it is desirable that the reaction proceed in the direction of calcium sulfide (CaS) formation. The desulfurization increases with increasing lime activity in the slag and sulfur activity in the hot metal and with decreasing calcium sulfide activity in the slag and oxygen activity in the hot metal.

Equation (36): It is necessary that the reduction reaction with carbon proceeds to a certain degree in the stack of the blast furnace in order to attain sufficiently low oxygen activities in the hot metal for good desulfurization of hot metal. This is mainly achieved by temperature adjustment. High temperature favors the reducing capability of carbon according to equation (36) as can be derived from the Richardson diagram.

Equilibria. The equilibria of the aforementioned reactions are related to one another by the constant oxygen activity $a_{[O]}$. In order to handle the equilibria practically, it is useful to express the activities by the products of concentrations

and the activity coefficients. As far as the metal phase is concerned, the activity coefficients of its components are almost constant over the entire range of concentration, because at the low actual concentrations, the thermodynamic interrelation between the single elements can be neglected, which means that Henry's law is valid. This is not the case with carbon, but for this element $a_{[C]} \cong 1$ can be assumed as an approximation. The activity coefficients of the slag are strongly concentration dependent. For a current production run, however, the slag composition changes very little so that the activity coefficients are also constant. It must be taken into account, however, that with every change of slag composition the activity coefficients change. For blast furnace type slags a good documentation of slag activities exists [63]. Making use of the concept of activity coefficients in the case of the slag components and using concentrations instead of activities for the metal components except carbon, the following equilibrium constants can be written:

1) For the reduction of silica

$$K^*_{\mathrm{Si}} = \frac{[\mathrm{Si}][\mathrm{O}]^2}{f_{\mathrm{SiO_2}}(\mathrm{SiO_2})} \text{ or :} \qquad K^*_{\mathrm{Si}} \cdot f_{\mathrm{SiO_2}} = \frac{[\mathrm{Si}][\mathrm{O}]^2}{(\mathrm{SiO_2})} = K_{\mathrm{Si}} \tag{37}$$

2) For the desulfurization reaction

$$K^*_{\mathrm{S}} = \frac{[\mathrm{S}]}{(\mathrm{S})} \cdot \frac{f_{\mathrm{CaO}}(\mathrm{CaO})}{[\mathrm{O}]} \text{ or :} \qquad \frac{K^*_{\mathrm{S}}}{f_{\mathrm{CaO}}} = \frac{[\mathrm{S}]}{(\mathrm{S})} \cdot \frac{(\mathrm{CaO})}{(\mathrm{O})} = K_{\mathrm{S}} \tag{38}$$

3) For the reduction of manganese oxide

$$K^*_{\mathrm{Mn}} = \frac{[\mathrm{Mn}]}{f_{\mathrm{MnO}}(\mathrm{MnO})}[\mathrm{O}] \text{ or :} \qquad K^*_{\mathrm{Mn}} \cdot f_{\mathrm{MnO}} = \frac{[\mathrm{Mn}]}{(\mathrm{MnO})}[\mathrm{O}] = K_{\mathrm{Mn}} \tag{39}$$

The oxygen concentration in the equilibria (37), (38), and (39) is the same. Therefore, each two of the three equations can be combined by eliminating the oxygen concentration. Including the approximately constant SiO_2 and CaO contents of the slag in the constants and exchanging (MnO) for (Mn) then yields:

$$\frac{1}{K^{1/2}_{\mathrm{Si}} \cdot K_{\mathrm{S}}} \cdot \frac{(\mathrm{CaO})}{(\mathrm{SiO_2})^{1/2}} = K_{\mathrm{Si,S}} = \frac{1}{[\mathrm{Si}]^{1/2}} \cdot \frac{(\mathrm{S})}{[\mathrm{S}]} \tag{40}$$

$$\frac{K^{1/2}_{\mathrm{Si}}}{K_{\mathrm{Mn}}}(\mathrm{SiO_2})^{1/2} = K_{\mathrm{Si,Mn}} = [\mathrm{Si}]^{1/2} \cdot \frac{(\mathrm{Mn})}{[\mathrm{Mn}]} \tag{41}$$

$$\frac{(\mathrm{CaO})}{K_{\mathrm{Mn}} \cdot K_{\mathrm{S}}} = K_{\mathrm{Mn,S}} = \frac{(\mathrm{S})}{[\mathrm{S}]} \cdot \frac{(\mathrm{Mn})}{[\mathrm{Mn}]} \tag{42}$$

The three constants are approximate values that describe the relations between the silicon reduction, the manganese reduction, and the desulfurization reaction. They were introduced by Oelsen and coworkers [47]. The values of these constants are mainly dependent on the slag composition [47] especially the so-called basicity, but otherwise comparatively constant. Thus, they proved as a useful tool to predict the behavior of the blast furnace.

All the aforementioned constants depend on the slag basicity. The basicity is the ratio between the sum of the basic slag components and the sum of the acid components. Because the effect of various basic oxides on the basicity is different, their concentrations are expressed in terms of equivalent concentrations by making use of an equivalency factor; the latter equals unity for calcium oxide. Similar factors are used for the acid components in which case silica is the reference oxide.

The basicity B may be expressed by a series of complicated expressions [43], as well as by equations depending on accuracy and range of B required

$$B = \frac{(\%\mathrm{CaO}) + (\%\mathrm{MgO})}{(\%\mathrm{SiO_2}) + (\%\mathrm{Al_2O_3})} \tag{43}$$

or

$$B = \frac{(\%\mathrm{CaO}) + 0.7\,(\%\mathrm{MgO})}{0.93\,(\%\mathrm{SiO_2}) + 0.18\,(\%\mathrm{Al_2O_3})} \tag{44}$$

and, finally, when neglecting low magnesia and alumina contents

$$B = \frac{(\%\mathrm{CaO})}{(\%\mathrm{SiO_2})} \tag{45}$$

The constants $K^*_{\mathrm{Si,S}}$, $K_{\mathrm{Mn,Si}}$ and $K_{\mathrm{Fe,Si}}$ are shown in Figure 23 as a function of the basicity and for various alumina contents. $K^*_{\mathrm{Si,S}}$ is the reciprocal value of $K_{\mathrm{Si,S}}$ and $K_{\mathrm{Fe,Si}}$ is the constant that describes the reduction of iron oxide from the slag. It is seen from the diagram that, for a given basicity, the constants are hardly affected by the replacement of calcium oxide by magnesium oxide.

Activities. On the basis of many laboratory and blast furnace plant experiments it has been

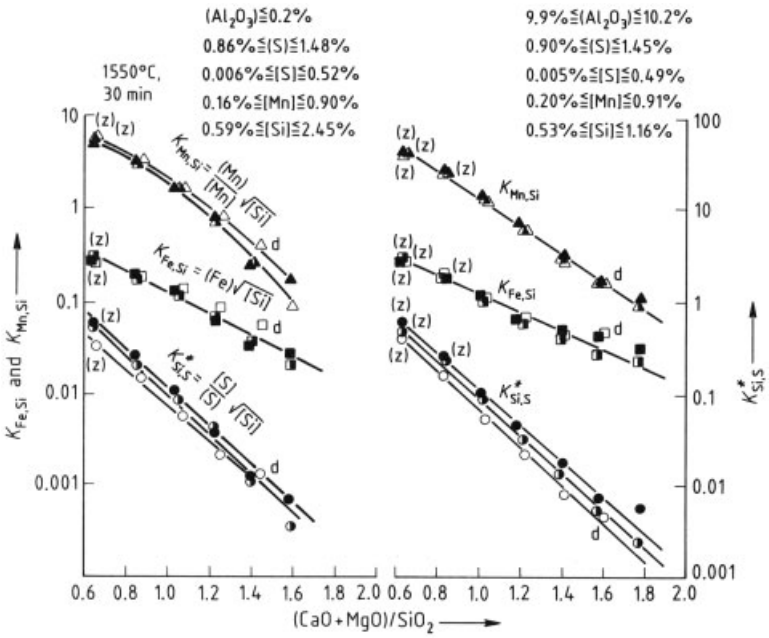

Figure 23. Equilibrium constants $K_{\mathrm{Mn,Si}}$, $K_{\mathrm{Fe,Si}}$ and $K^*_{\mathrm{Si,S}}$ as a function of the basicity (CaO + MgO)/(SiO_2) with the Al_2O_3 content and the MgO/CaO ratio as parameters [47]
(MgO)/(CaO):
1) = 0.0 for empty symbols; 2) = 0.5 for half-filled symbols; 3) = 1.0 for filled symbols.
(z) = slightly viscous liquid; d = heavy viscous liquid

established that the metal : slag distribution ratios can be accurately evaluated with the aforementioned constants. To obtain a better insight into the basic interactions between the various components in the metal and the slag, information is needed concerning their activities and how these are affected. In general, the relation between the concentration of a component and its activity a is characterized by the activity coefficient f. Whenever the activity is affected by other components, which also form a dilute solution with the solvent, the activity coefficient is denoted by $f^{\,x}$. The activity of sulfur is given by

$$a_{\mathrm{S}} = f^x_{\mathrm{S}}\,[\mathrm{S}] \tag{46}$$

The values of f^x_{S} may be obtained from Figure 24. The influence of carbon and silicon on the sulfur activity is extremely pronounced. Because their concentrations in the hot metal are relatively high, the sulfur activity is increased, facilitating desulfurization.

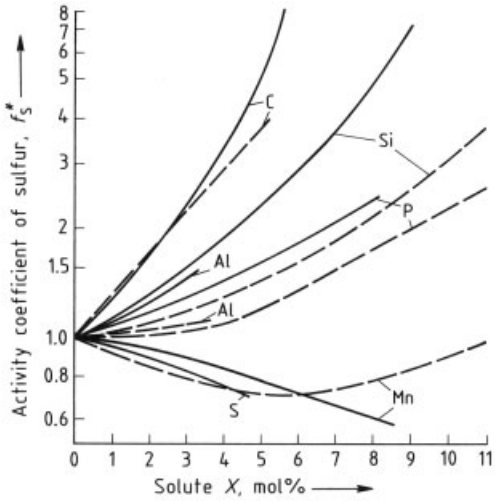

Figure 24. Activity coefficient of sulfur as a function of the concentration of other solutes [43]
——— Values after CHIPMAN; - - - - - - - Values after SCHENCK

To represent the sulfur activity in hot metal, it is useful to express the effect of all the other

solutes in terms of an *equivalent carbon content* Δc_C. The activity coefficient of sulfur can then be depicted as a function of the carbon content with the silicon content as the parameter (see Fig. 25). With the aid of this diagram it is easy to evaluate the activity coefficient of sulfur $f_S^{(Si,C)}$, important in the production of hot metal, at various temperatures. The carbon content c_C in Figure 25 is obtained from the actual carbon content $c_{C,t}$ as follows:

$$c_C = c_{C,t} + \Delta c_C \quad (47)$$

with

$$\Delta c_C = 0.043\% \, [P] - 0.025\% \, [Mn] - 0.028\% \, [S] \quad (48)$$

At equilibrium the chemical potential of sulfur in the metal and the slag is equal. The sulfur in the slag is present as calcium sulfide (CaS) for the oxygen potential prevailing inside the blast furnace. The solubility limit of calcium sulfide in lime – silica slags is ca. 2 % sulfur. Hence, the sulfur activity in the slag reaches its maximum value at these low sulfur contents. The activity is shown in Figure 26 as a function of the sulfur content of lime – silica slags with the basicity B as the parameter.

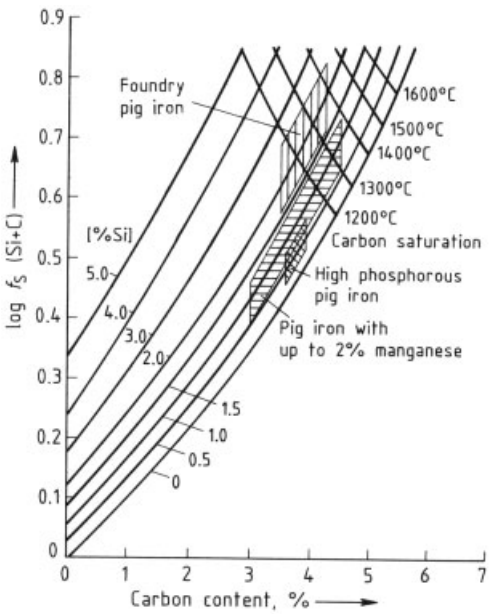

Figure 25. Activity coefficient of sulfur as a function of the carbon content [43] (isotherms for 1200 to 1600 °C)

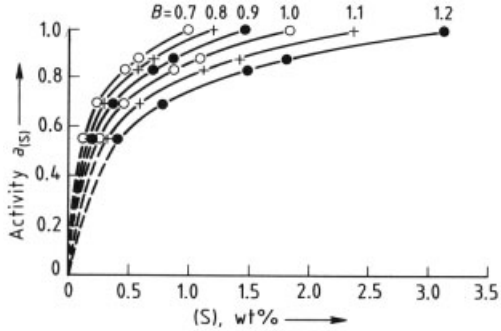

Figure 26. Activity of calcium sulfide as a function of the sulfur content of lime – silica slags with the basicity B as the parameter [43]

Phase Diagrams. Essentially, blast furnace slags contain four components: calcium oxide, silicon dioxide, magnesium oxide, and aluminum oxide. However, ternary systems are easier to visualize and for this reason blast furnace slags are frequently represented by a quasi-ternary section of the $CaO - SiO_2 - MgO - Al_2O_3$ diagram at a constant alumina content of, for example, 10 % which is typical for blast furnace slags (Fig. 27). This type of representation is useful because the alumina content is essentially determined by the gangue constituents of the ore and, hence, remains constant. The magnesia content, on the other hand, changes with changing dolomite additions to the burden.

2.3. Kinetic Principles [19, 43, 48, 212]

In the reduction processes which are commercially important at present, reduction in the solid and in the liquid state can be distinguished. Both processes are heterogeneous reactions between several phases. Knowledge of the reduction in the solid state is rather extended, whereas that on reduction in the liquid stage is still meager.

2.3.1. Reduction in the Solid State

During reduction in the solid state, the gas composition is generally a function of space and

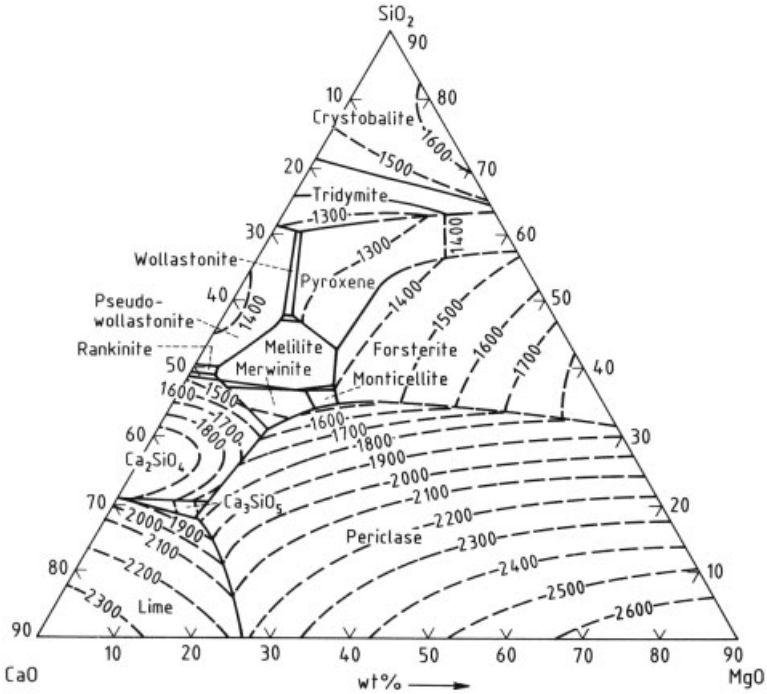

Figure 27. Phase relationships and liquidus temperatures in the system CaO – MgO – SiO_2 – Al_2O_3 containing 10 % Al_2O_3 [46]

time within the ore particle, and depends on the overall location in the *continuous reactor*. In a *batch reactor* the gas composition is also time dependent. The number, spatial arrangement, structure, and chemical composition of the solid phases change as the reduction progresses. The time dependencies of the various stages that constitute the reduction process are known. Three distinct phenomena have to be recognized:

1) Reducing gas components (CO, H_2) are transported to the solid oxide phase via *transport phenomena,* the oxidized gas components (CO_2, H_2O) are transported away from the reaction front.
2) The reduction reaction itself proceeds essentially at the *phase boundary* between gas and solid oxide.
3) As a result of the *removal of oxygen* from the surface of the solid phases and the accompanying transport phenomena, new solid phases such as lower iron oxides or metallic iron are being formed.

2.3.1.1. Time Dependence of Individual Processes During Reduction

The overall reduction process is characterized by a series of individual processes representing various resistances to the overall reaction rate. The potential difference that drives the reduction process causes matter to flow through the resistances which are usually in series with each other. Hence, the major portion of the potential difference is expended at the region of highest resistance. Generally spoken the higher the single resistance, the more rate determining it is for the entire process. Because the resistances corresponding to the various mechanisms change according to the choice of the experimental conditions and with the type of ore, the single resistances may be differently distributed over the locations of the ore. For a long time this circumstance was not properly recognized and, consequently, has led to contrasting opinions concerning the reaction step that determines the overall reaction rate. It was not until the 1960s that these phenomena were summarized and clearly presented [19, 212]

Fundamentals. The individual mechanisms important for the reduction kinetics are best described with a model (see Fig. 28). For a discussion of the fundamental laws a partial knowledge of the model is a prerequisite.

Kinetic reaction steps are described with flux equations. A flux equation expresses the fact that a flux (mol m^2 s^{-1}) is caused by a concentration difference as the driving force and related to it by a flux coefficient. Transport phenomena are primarily expressed by Fick's first law

$$j = -D\,\frac{\mathrm{d}c}{\mathrm{d}x} \tag{49}$$

and Fick's second law

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D\,\frac{\partial c}{\partial x}\right) = D\,\frac{\partial^2 c}{\partial x^2} \tag{50}$$

If the concentration field is independent of time, the gradient dc/dx in Equation (49) can be exchanged by

$$-\frac{\mathrm{d}c}{\mathrm{d}x} = \frac{c_2 - c_1}{\delta} \tag{51}$$

with $c_2 > c_1$. The quantities c_2 and c_1 are concentration values at fixed points, for example at a point of the inner surface of the ore and a point at the outer surface; δ is the distance between the locations of c_2 and c_1, in the present example it would be the capillary length. Alternately, c_2 and c_1 may also be the concentrations at the ore surface and in the bulk gas phase respectively. Then δ is the thickness of the diffusion gas boundary layer. Introducing Equation (51) into (49) yields

$$j = \frac{D}{\delta}\,(c_2 - c_1) \tag{52}$$

The ratio D/δ is sometimes termed mass transfer coefficient β. Equation (52) then changes into

$$j = \beta\,(c_2 - c_1) \tag{53}$$

Equation (53) has the advantage that it can formally express a flux caused by any concentration difference. It is often simple to find experimentally a relation between a flux and the appertaining driving concentration difference, and then to express the coefficient between both as a mass transfer coefficient. A comparison with Equation (1) shows that equations of the type of Equation (53) can also be applied to chemical reactions at the phase boundary. Instead of β, the rate constant k is then used, but the mathematical form of Equation (53) remains the same. It is the practical applicability, which makes Equation (53) so extremely suitable for the description of kinetic phenomena like those of iron ore reduction.

For the description of gas diffusion it is more practical to express the gas concentration in terms of the partial pressure p instead of the concentration c. Both values are interrelated by

$$c = \frac{p}{RT} \tag{54}$$

Although Equations (51) and (53) are valid for a constant concentration field, i.e., for a steady state, they are also applicable to situations where the concentration difference $c_2 - c_1$ changes during the process from an initially maximum value to zero, i.e., when equilibrium is attained. In contrast to a really time-dependent concentration field these changes are slow. The process is in a so-called quasi-steady state. This state can be treated in the same way as a steady state with a constant concentration field. The criterion for its validity is that capacity effects may be neglected.

The rate equations described contain quantities related to the particular geometry (mass exchange area A) and to material properties (the kinetic coefficients). In the following some definitions are given:

1) The *degree of reduction* $\Re$ is the ratio between the quantity of oxygen removed O_r and the quantity of oxygen initially present O_i:

$$\Re = \frac{O_r}{O_i} \tag{55}$$

Frequently the degree of reduction is expressed in percent (= 100 $\Re$).

2) *Reduction Rate.* When discussing the reduction kinetics, the rate of reduction, $\dot{\Re}$, is defined by

$$\dot{\Re} = \frac{\mathrm{d}\Re}{\mathrm{d}t} \tag{56}$$

3) *Effect of Diffusional Flows.* Because the reduction of ore by a gas does not result in a change in the number of moles in the gaseous phase (e.g., 1 mol CO gives 1 mol CO_2), the *total pressure* inside the pores of the ore does not change. The transport of matter occurs exclusively via the diffusion brought about by the difference in *partial pressure* of the gas components. The partial pressure of the

inert gas components present in the reducing gas is the same everywhere inside the ore particle. Hence, no macroscopic transport of these components takes place. Although during reduction with carbon monoxide inert gas is transported from the outside to the interior (corresponding to the direction of the CO flow toward the center) the carbon dioxide flow transports inert gas from the interior to the outside, the two effects cancelling each other. Only when hydrogen is used as a reducing agent is there a small net flow of inert gas from the interior to the exterior. This occurs because water vapor molecules have a significantly larger cross-sectional area than hydrogen molecules. Thus, the partial pressure of the inert gas is slightly lower in the interior of the ore than in the surroundings.

The flux of gaseous species diffusing inward, $j = 1/A \cdot \mathrm{d}n/\mathrm{d}t$, is related to the flux of the gaseous species diffusing outward via the stoichiometry of the reduction reaction:

$$j_1 = -j_2 \tag{57}$$

Expressing both fluxes by Equation (49) and denoting the respective diffusion coefficients of gas 1 in gas 2 and vice versa with D_{12} and D_{21}, yields:

$$D_{12}\,\frac{\mathrm{d}p_1}{\mathrm{d}x} = -D_{21}\,\frac{\mathrm{d}p_2}{\mathrm{d}x} \tag{58}$$

The sum of the partial pressures is constant ($p_1 + p_2 = p_{\mathrm{tot}}$), hence:

$$\frac{\mathrm{d}p_1}{\mathrm{d}x} = -\frac{\mathrm{d}p_2}{\mathrm{d}x} \tag{59}$$

Thus:

$$D_{12} = D_{21}$$

When considering the effect of the diffusion on the reduction it is, therefore, irrelevant whether the diffusional flow of the reducing gas into the ore or whether the outward flow of the oxidized gas from the interior to the surface of ore is considered. This is possible because the diffusion coefficients for the two counterflowing gas streams are equal.

Model Representations. The phenomena taking place in the *gaseous phase* and depicted in Figure 28 can be modeled as follows. In a first step, the reducing gas diffuses from the main flow through the gaseous diffusion boundary layer to the ore particle (step a in Figure 28). The transport within the ore particle is governed by diffusion inside the macropores (b) and the micropores (c). The distinction between macro- and micropores is determined by the mean free path of the gas molecules. Whenever the mean free path is large with respect to the pore diameter, the number of collisions between the gas molecules is small with respect to those between the gas molecules and the pore walls. In this case gas transport no longer obeys the laws of diffusion valid for free space, which is also valid for macropores with some geometric restrictions, but is determined by Knudsen diffusion, for which the diffusion coefficient follows a different law. Within the square denoted by (d), Within the square denoted by (d), the phase boundary reaction of oxygen separation from the solid ore crystal by the reducing gas takes place at the ore – gas interface. As a consequence of oxygen separation the ore becomes supersaturated with iron, leading to the precipitation of iron at a site where nucleation is favored. This site is different from that of oxygen removal, so that solid state diffusion of iron to the nucleation site is a further reaction step. The oxidized reducing gas (here water vapor) is transported to the exterior by the flow of gas in the opposite direction to the reducing gas and along the same path (steps f and g).

In principle, each of the various stages constituting the transport and conversion of matter, as depicted in Figure 28, can be rate determining. It may also be possible that several resistances are of about equal magnitude and thus together determine the reaction rate. Depending on the specific circumstances, an ore may be dense and yield a dense or a porous reaction product upon reduction; also, reduction may take place at high or low temperatures. Diffusion is only slightly affected by temperature, whereas the phase boundary reactions in general require high activation energies and are therefore strongly dependent on temperature. The result is that the rate-determining step of the overall reaction changes with varying temperature (cf. for example the Boudouard reaction rate constants in Fig. 10). A detailed discussion follows on the separate stages of the reduction process and for the case where they represent the largest resistance.

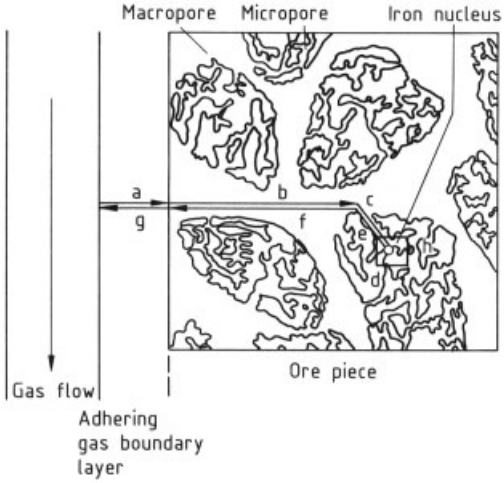

Figure 28. Schematic representation of the reduction mechanism in porous iron ores [19]
a) Diffusion of hydrogen through the boundary layer; b) Diffusion of hydrogen through the macropores of the ore; c) Diffusion of hydrogen through the micropores of the ore; d) Interfacial reaction; e) Diffusion of water vapor through micropores; f) Diffusion of water vapor through macropores; g) Diffusion of water vapor through the boundary layer; h) Migration of Fe^{2+} and $2e^-$ to the iron nucleation site

Gaseous Transport.

Diffusion through the Outer Gas Boundary Layer. The first transport resistance is that of transport of gas through the adjacent gaseous boundary layer (see Fig. 28). The flow of matter through this layer is determined by Fick's first law with the partial pressure difference as the driving force (Eqs. 49 and 54). If the partial pressure gradient $-\mathrm{d}p/\mathrm{d}x$ is replaced by $\Delta p/\delta_\mathrm{N}$ (Δp = pressure difference, δ_N = thickness of the Nernst's diffusion boundary layer) and the ratio D/δ_N replaced by the mass transfer coefficient β, Equation (49) takes the form

$$j = \frac{\beta\,\Delta\mathrm{p}}{RT} \tag{60}$$

The mass transfer coefficient β is inversely proportional to the thickness δ_N of the boundary layer. The latter decreases with increasing gas flow rate.

Reduction experiments in the laboratory have shown that the reduction rate of a Fe_2O_3 layer in carbon monoxide at 1000 °C is a function of the gas flow rate (see Fig. 29). The rate of reduction is expressed in terms of the degree of reduction attained after a certain time from the beginning.

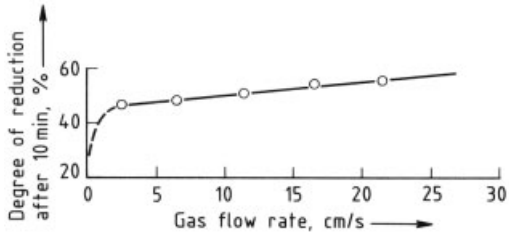

Figure 29. Relation between gas flow rate and reduction rate of iron(III)oxide in carbon monoxide at 1000 °C [43]

In the reduction of an array of particles (e.g., a burden) the boundary layer around each particle is determined by the individual fluid dynamic conditions. The mass transfer coefficient β in the burden can be calculated from the experimentally determined heat transfer coefficient which is easier to evaluate. Such calculation is possible on the basis of the physical similarity of heat transfer and mass transfer. For the conditions prevailing in the blast furnace the expression for β takes the form

$$\beta = \mathrm{C}\,\frac{u^{0.7}}{d^{0.3}} \tag{61}$$

where C is a constant, u the superficial gas velocity, and d the particle diameter.

For constant superficial velocity the mass transfer coefficient decreases slightly with increasing particle diameter.

The inverse of the mass transfer coefficient may be termed specific resistance. Comparison of this resistance, as defined by Equation (61), with the resistances of the other partial reaction steps of iron ore reduction kinetics (which are described below), shows that under the fluid dynamic conditions of packed beds in the blast furnace, gas transport through the outer boundary layer requires a partial pressure difference Δp_g as the driving force, which is only about 2 % of the required total partial pressure difference Δp_tot for all the reaction steps in total. That means, the transport resistance of the outer gas boundary layer may almost be neglected.

Diffusion in Macro- and Micropores. The second transport mechanism shown in Figure 28 is the gaseous diffusion in the macro- and micropores. The diffusion law (Eq. 35) governs the diffusion in *macropores.* However, the diffusion coefficient D has to be corrected in two

ways. The gas diffusion in a porous solid takes place only in the volume occupied by the pores which is a fraction γ of the total volume. Therefore, D for the bulk gas phase is multiplied by the porosity γ to give the pore diffusion coefficient D_{P}. In addition, the course of the pores in relation to the direction of the flow has to be taken into account. This is achieved by the introduction of the *tortuosity factor* ξ which gives D_{P} the expression:

$$D_{\mathrm{P}} = D\gamma\xi \tag{62}$$

For short time spans, i.e., from a few seconds to several minutes, depending on temperature and stage of reduction, the diffusion may be assumed to take place in the steady state and the flux j assumed to be constant. In that case Equation (53) takes the form

$$j = \frac{D_{\mathrm{P}}}{RT} \cdot \frac{(p_{\mathrm{a}} - p_{\mathrm{P}})}{x} \tag{63}$$

Here, p_{a} is the partial pressure of the reducing gas component at the interface between the ore and the boundary layer and p_{P} the partial pressure at the interface in the solid where transport by diffusion ceases and the chemical reaction occurs; x is the linear distance between the locations of p_{a} and p_{P}.

If the resistances representing the outer boundary layer and the phase boundary reaction at the ore interface are negligible, the partial pressure p_{a} may be replaced by p_0, the partial pressure of the reducing gas component in the bulk phase, while p_{P} may be replaced by p^*, the equilibrium partial pressure

$$j = \frac{D_{\mathrm{P}}}{RT} \cdot \frac{(p_0 - p^*)}{x} \tag{64}$$

where p^* is dependent on the temperature and on the reduction step considered (Fe_2O_3 to Fe_3O_4, Fe_3O_4 to FeO, or FeO to Fe). The corresponding values may be obtained from the Richardson diagram (see Fig. 12). Regarding technical problems the dependence of the flux j on temperature and total pressure is important. In Equation (65) $(p_0 - p^*)$ increases with increasing pressure and p^* depends on the temperature as may be seen from Figure 12. The pore diffusion coefficient changes with pressure and temperature – as may be determined experimentally or from the kinetic gas theory – according to:

$$D_{\mathrm{P}} \sim \frac{1}{P} \tag{65}$$

$$D_{\mathrm{P}} \sim T^n, 1.75 < n < 2 \tag{66}$$

As the pressure dependence of $(p_0 - p^*)$ is the inverse of that of D_{P} the diffusional flux j is independent of pressure.

The flow of matter in *micropores* is governed by the same expression valid for the flow in macropores (Eq. 63). However, in this case D_{P} must be replaced by the Knudsen diffusion coefficient D_{K}. For the component a this coefficient D_{K} may be calculated from the expression [50]

$$D_{\mathrm{K,a}} = \frac{d}{6}\left(\frac{8\,RT}{\pi M_{\mathrm{a}}}\right)^{1/2} \tag{67}$$

Here, d is the pore radius which is less than the mean free path λ of the gas molecules. As D_{K} is independent of pressure but $(p_0 - p^*)$ depends on pressure, the flow of matter caused by Knudsen diffusion increases with increasing total pressure, provided $d < \lambda$ despite the increase in pressure. The temperature dependence of the diffusion in micropores is lower than that of macropore diffusion.

Interfacial Reaction. The interfacial reaction phenomenon shown in Figure 28 has been studied in several respects. The results have been published in many individual publications; an extensive review of these complex phenomena is also available [43]. A simplified version is discussed here. The oxides of iron, hematite ($FeO_{1.5}$), magnetite ($FeO_{1.33}$), and wustite ($FeO_{\approx 1.08}$) have a close-packed structure of oxygen ions (see page 46). The tetrahedral and octahedral interstices of this structure are occupied with Fe^{2+} and Fe^{3+} ions. The transformation of hematite to magnetite requires a redistribution of the oxygen ions from the hexagonal to the face-centered cubic structure. During transformation of magnetite to wustite the oxygen ions are not rearranged, only the iron : oxygen ratio changes.

Reduction of Hematite to Magnetite. During reduction of hematite oxygen is removed at the gas – hematite interface from the hematite lattice according to

$$\mathrm{O^{2-} \rightarrow O + 2\,e^-}$$

The iron ions react with the free electrons according to:

$$\mathrm{2\,Fe^{3+} + 2\,e^- \rightarrow 2\,Fe^{2+}}$$

Whenever at the surface of the oxide the ratio $Fe^{3+}:Fe^{2+} = 2$ is reached, the hexagonal oxygen lattice transforms into the cubic magnetite lattice and hematite becomes covered with a thin layer of magnetite. The gas does not have direct access anymore to the hematite phase so that further oxygen removal occurs from the magnetite lattice (see Fig. 30).

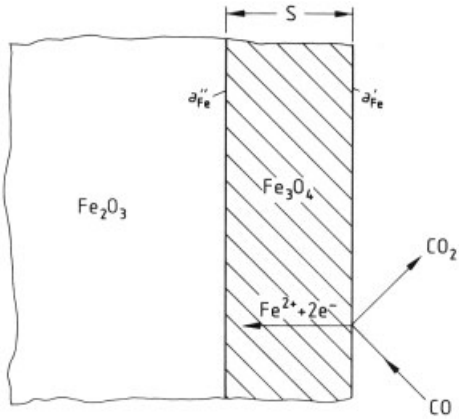

Figure 30. Reduction of hematite to magnetite

If the composition of the reducing gas is in equilibrium with magnetite (see Fig. 15), the removal of oxygen can proceed only to a level corresponding to an iron activity that is determined by the oxygen partial pressure at the gas – magnetite interface, according to Figure 17. This iron activity is higher than that at the hematite – magnetite interface (see Fig. 17). This difference in iron activity, which is opposite to that in oxygen activity, leads to diffusional flows of iron and oxygen ions. Experiments show that the differences in activity in magnetite and wustite result in relatively high diffusional flows of iron ions and electrons and in a small flow of oxygen ions. On the other hand, in hematite the transport of matter is primarily the result of the diffusion of oxygen ions. Accordingly, the reduction process depicted in Figure 30 proceeds in such a manner that the removal of oxygen at the magnetite – gas interface leads to an iron activity a'_{Fe}. Iron ions and electrons diffuse through the magnetite layer and react at the hematite – magnetite interface – where the iron activity equals a''_{Fe} – according to:

$$4\,Fe_2O_3 + Fe^{2+} + 2\,e^- \rightarrow 3\,Fe_3O_4 \qquad (68)$$

As a result the interface moves further into the hematite phase.

If reduction proceeds perpendicular to the *c*-axis of the hematite, the reduction rate is determined by the diffusion rate of the iron ions through the magnetite layer previously formed. For high rates of reaction (68) the activities a''_{Fe} and a'_{Fe} at the interfaces are fixed. Hence, the driving force $(a'_{Fe} - a''_{Fe})$ remains constant. However, the diffusion length s through the growing magnetite layer which is the thickness of the layer, increases during the reaction, and hence, the concentration gradient $(a'_{Fe} - a''_{Fe})/s$ decreases. The flux is expressed by Equation (52). Considering that here the flux j is proportional to the layer growth rate ds/dt and that δ is equal to the layer thickness s, gives $ds/dt \sim (c'_{Fe} - c''_{Fe})/s$. The growth rate ds/dt is indirectly proportional to the thickness itself. Integration of this equation leads to a parabolic rate law for the layer growth

$$s = k \cdot t^{1/2} \qquad (69)$$

where k is a constant and t is the time. If the reduction of hematite occurs parallel to the *c*-axis, a parabolic rate law is not observed experimentally. It is assumed that in that case the interfacial reaction (68) becomes rate determining.

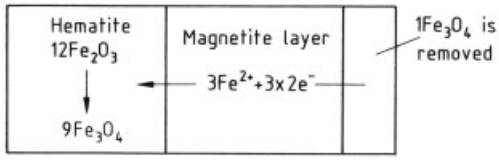

Figure 31. Geometric and stoichiometric relationships during the hematite reduction

The geometric and stoichiometric conditions during the reaction of 1 mol magnetite at the magnetite – gas interface are shown in Figure 31. The 3 mol iron ions produced, together with the electrons, react with 12 mol hematite to produce 9 mol magnetite at the hematite – magnetite interface. As 12 mol hematite contain 24 mol iron, but 9 mol magnetite contain 27 mol iron, 1 mol magnetite vanishes at the magnetite – gas interface at the same time. Because the spatial requirement of magnetite exceeds that of the previously present hematite, technical reduction processes frequently give rise to unde-

sirable structural disorders. On the other hand, these lead to an increase in the interfacial area exposed to the gas and, hence, an increase in the reduction rate.

Reduction of Magnetite to Wustite. This reaction proceeds similar to that of hematite to magnetite. The removal of oxygen from the wustite surface liberates iron ions that subsequently diffuse through the wustite layer, that grows at the expense of magnetite, and react with magnetite to form wustite. The removal of oxygen from the wustite phase, in which the oxygen content varies between 24.5 % and ca. 22 %, takes place in a similar fashion. The iron activity at the surface increases as dictated by the removal of oxygen. By diffusion of iron ions an activity is eventually established in the interior of the wustite phase that corresponds to the oxygen partial pressure (see Fig. 17).

Reduction of Wustite to Metallic Iron. This may give rise to two limiting cases concerning the morphology of the phases, as illustrated in Figure 32.

The spherical precipitations inside the wustite crystal and the lens-shaped precipitation of iron on a wustite surface are depicted in Figure 32 A and B; upon further reduction this usually leads to the formation of porous sponge iron. In this case oxygen is removed *directly* from wustite by direct contact between the gas and the wustite phase. Pores are formed at the sites from which the oxygen was removed; the pore walls are subsequently covered with metallic iron. The iron ions and electrons that are liberated as a result of the oxygen removal diffuse from the removal site to the metallic iron and condense there to form metallic iron. Iron ions and electrons do not recombine at the removal site of oxygen to produce metallic iron because nucleation is impeded there.

The reduction mechanism in the presence of a protective iron layer on the wustite surface is illustrated in Figure 32 C. In this case oxygen must diffuse through the iron layer to react with the reducing gas. The permeability $c_O D_O$ of the iron for oxygen is small because of the low solubility c_O of oxygen in iron. Thus, this morphology of the formation of iron leads to a slow reduction rate compared to the mechanism illustrated in Figure 32 A and B.

The conditions that play a role in the mechanism of iron formation are still unclear in many respects. However, it was established experimentally that the reduction of wustite with a high concentration of vacant iron sites, i.e., with a high oxygen content according to Figure 14, generally produces a porous iron layer. At lower concentrations of vacant iron sites the formation of a protective iron layer was observed. Generally, protective iron layers are formed around the wustite grains whenever magnetite ore is reduced but porous iron is formed whenever

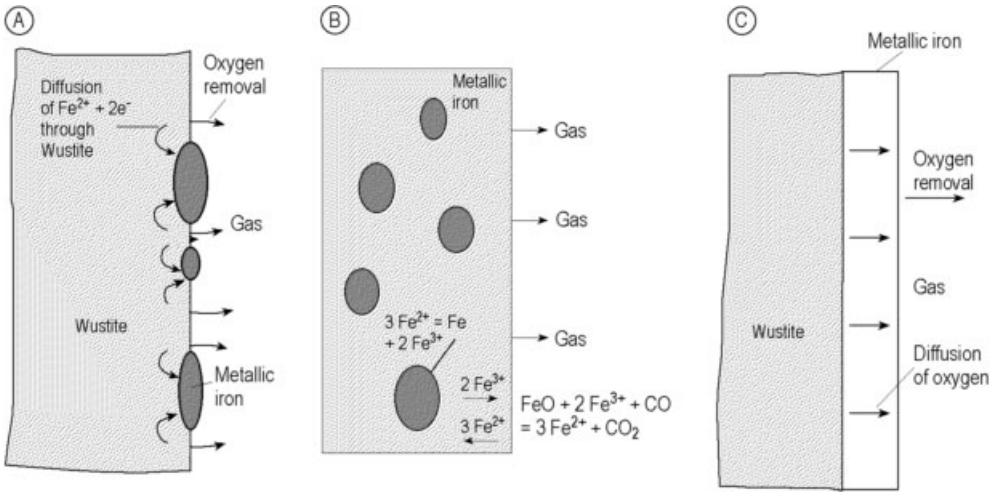

Figure 32. Morphology and location of the phases during the reduction of wustite to iron
A) Precipitation of iron on the surface; B) Iron precipitation within the oxide body; C) Reduction in the presence of a protective iron layer

hematite ore is reduced. In the latter case the reduction rate is higher because reducing gas has direct access to the oxygen lattice. Hence, magnetite ore is usually oxidized to hematite during sintering or pelletizing.

Under special conditions iron whiskers may be formed on the wustite surface. This may result in a significant change in volume of the ore particles, particularly when the grain structure of the particles was already weakened during the hematite – magnetite reduction stage. This form of iron precipitation is, therefore, undesirable in technical reduction processes.

Rate Laws. When discussing the rate laws that govern phenomena belonging to the group of "reduction reaction," important in technical processes, a few simplified assumptions can be made on the basis of the technical boundary conditions.

1) As during the reduction step hematite ($FeO_{1.5}$) $\rightarrow$ wustite ($FeO_{1.05}$) only ca. 1/3 but during the step wustite ($FeO_{1.05}$) $\rightarrow$ iron (Fe) ca. 2/3 of the total removable oxygen is eliminated, and as moreover the rate of the second of these two steps of the overall reaction is slower than that of the first, the latter step is solely taken into account. The error of that simplification is small.
2) For the ores processed in industrial reduction processes it can be assumed that porous iron, as illustrated in Figure 32 A, is formed. This means that the reducing gas has access to a portion of the oxide surface during the entire reduction process.
   The rate of oxygen removal from the interface is, analogue to Equation (2), given by

$$\frac{\mathrm{d}n_\mathrm{O}}{\mathrm{d}t} = \frac{k'}{RT} A\,(p_\mathrm{P} - p*) \tag{70}$$

   where p_P is the partial pressure of the reducing gas at the interface and k' is the rate constant in m/s. However, the oxygen removal as given by this expression is only formally correct. In the case of Figure 32 A for example, the reaction surface area A is only part of the total surface area that is exposed to the gas [49]. Moreover, in the reduction of porous oxides the effective surface depends on the porosity. It was found that the activation energy of oxygen removal prior to the formation of metallic iron is 170 kJ/mol and decreases to ca. 60 kJ/mol after the formation of metallic iron. This means that for the technically important case of reduction accompanied by the formation of iron the temperature dependence on the rate of the interfacial reaction is relatively small. The pressure dependence of the flux of material $\mathrm{d}n_\mathrm{O}/\mathrm{d}t$ follows from the pressure dependence of the term $(p_\mathrm{P} - p^*) = \Delta p$. The linear pressure dependence $\Delta p_2 = \Delta p_1 \cdot P_2/P_1$ is masked by side effects at higher pressures.

2.3.1.2. Rate Laws of Complex Reduction Reactions

Dense Ore Particles, Porous Iron Layers.

Gas Diffusion is Rate Determining. When discussing the technically important rate laws describing reduction processes, it is assumed that the ore particles may be approximated by a sphere.

During reduction of dense ores a layer of porous iron normally surrounds a core of unreduced ore. Whenever the diffusion of the reducing gas component i through the reduced layer is rate determining, Equation (49) with $j = 4\,\pi\,r^2 \cdot \mathrm{d}n_i/\mathrm{d}t$ and substituting p/RT for c and $\mathrm{d}r$ for $\mathrm{d}x$ takes the following form:

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = -\frac{4\pi r^2 D_{\mathrm{P},i}}{RT}\,\frac{\mathrm{d}p_i}{\mathrm{d}r} \tag{71}$$

where $r_0 > r > 0$ is the radius of the sphere. If p_0 and p_P are the partial pressures of the reducing gas in the bulk gas phase and at the gas – wustite interface near the unreduced core, respectively, and r' is the radius of the unreduced core, then integration of Equation (71) over the radius r between the limits r_0 and r' as well as over the partial pressure between the limits p_0 and p_P gives

$$\frac{\mathrm{d}n_i}{\mathrm{d}t}\left(\frac{1}{r_0} - \frac{1}{r'}\right) = \frac{4\pi D_{\mathrm{P},i}}{RT}\,(p_0 - p_\mathrm{P}) \tag{72}$$

Because of the stoichiometric relationship (CO, H_2) + O = (CO_2, H_2O) the number of moles of reducing gas $\mathrm{d}n_i$ diffusing into the ore particle equals the number of moles $\mathrm{d}n_\mathrm{O}$ of oxygen removed and diffusing out

$$\frac{\mathrm{d}n_i}{\mathrm{d}t} = -\frac{\mathrm{d}n_\mathrm{O}}{\mathrm{d}t}$$

The number of moles of oxygen diffusing out in time $\mathrm{d}t$ brings about a displacement of the reacting interface over a distance $\mathrm{d}r'$ given by

$$\frac{\mathrm{d}n_\mathrm{O}}{\mathrm{d}t} = 4\,\pi\,(r')^2 d_\mathrm{O}\,\frac{\mathrm{d}r'}{\mathrm{d}t} \tag{73}$$

where d_O is the oxygen content per unit volume of the ore (mol/cm^3).

Instead of the magnitude of the flux of oxygen, the reduction rate is usually employed. According to Equation (55) the following relation exists between the degree of reduction $\Re$, the initial volume V_0 and the volume V' of the unreduced spherical core of the ore particle:

$$\Re = 1-\frac{V'}{V_0} = 1-\frac{(r')^3}{(r_0)^3} \tag{74}$$

Substitution of Equations (73) and (74) into Equation (72), gives the following expression for the reduction rate $\dot{\Re}$

$$\dot{\Re} = \frac{\mathrm{d}\Re}{\mathrm{d}t} = \frac{3\,D_{\mathrm{P},i}\,(p_0-p_\mathrm{P})}{r_\mathrm{O}^2 d_\mathrm{O}\left[(1-\Re)^{-1/3}-1\right]RT} \tag{75}$$

In the case assumed here that solely pore diffusion is rate determining, the equilibrium partial pressure is established at the surface of the unreduced core, i.e. $p_\mathrm{P} \cong p^*$. Then, the reduction rate increases with increasing diffusion coefficient and concentration difference of the reducing gas and decreases with the square of the radius of the sphere (see Eq. 75). The effect of the degree of reduction on the reduction rate is particularly noticeable for low and high values of $\Re$.

Phase Boundary Reaction is Rate Determining. The following expression is valid when the phase boundary reaction is slow and, hence, rate determining

$$\dot{n}_\mathrm{O} = \frac{k'}{RT} A\,(p_\mathrm{P}-p*) \tag{70}$$

Here, $\dot{n}_\mathrm{O} = \mathrm{d}n_\mathrm{O}/\mathrm{d}t$, $A = 4\pi r'^2$. As $\dot{\Re} = \dot{\mathrm{n}}_\mathrm{O}/\mathrm{n}_\mathrm{O}^0$, where n_O^0 is the initial number of moles of oxygen present in the ore particle, and as

$$n_\mathrm{O}^0 = d_\mathrm{O}\cdot V_0 = d_\mathrm{O}\,\frac{4}{3}\pi r_0^3$$

Equation (70) yields with Equation (74)

$$\dot{\Re} = \frac{3k'(1-\Re)^{2/3}\,(p_\mathrm{P}-p*)}{RT\cdot d_\mathrm{O} r_0} \tag{76}$$

In the case assumed here that solely the phase boundary reaction is rate determining, the outer pressure p_0 is established in front of the reaction surface, i.e. $p_\mathrm{P} \cong p_0$.

Gas Diffusion and Phase Boundary Reaction have Similar Rates. In cases where the resistances representing the interfacial reaction and the pore diffusion are about equal in magnitude, the partial pressure of the reducing gas at the interface p_P is between p_0 and $p_0 > p_\mathrm{P} > p^*$.

Elimination of p_P from Equations (75) and (76) gives an expression for the reduction rate $\dot{\Re}$ in terms of the resistances representing diffusion and phase boundary reaction

$$\dot{\Re} = \frac{3\,D_\mathrm{P}\,k'(1-\Re)^{2/3}(p_0-p*)}{\left\{d_\mathrm{O} r_0 D_\mathrm{P} + r_0^2 d_\mathrm{O} k'\left[(1-\Re)^{1/3}-(1-\Re)^{2/3}\right]\right\}RT} \tag{77}$$

Integration yields a direct relation between the degree of reduction $\Re$ and the reaction time t

$$\frac{k'}{r_0 d_\mathrm{O} RT}\,(p_0-p*)\,t = \left[1-(1-\Re)^{1/3}\right] + \frac{k' r_0}{D_\mathrm{P}}\left[\frac{1}{2}-\frac{\Re}{3}-\frac{(1-\Re)^{2/3}}{2}\right] \tag{78}$$

Whenever the *phase boundary reaction* is rate determining, $D_\mathrm{P} >> RT\,k\,r_0$. In this case and under a given set of conditions $k'\,(p_0 - p^*) = k_\mathrm{c}$ and Equation (78) takes the form as suggested by McKewan [51]

$$d_\mathrm{O} r_0\left[1-(1-\Re)^{1/3}\right] = k_\mathrm{c}\,t \tag{79}$$

Whenever *gaseous diffusion* is rate determining, $D_\mathrm{P} << k'\cdot r_0$. In this case the first term on the right in Equation (78) may be neglected and Equation (78) takes the form

$$d_\mathrm{O} r_0^2\left[\frac{1}{2}-\frac{\Re}{3}-\frac{(1-\Re)^{2/3}}{2}\right] = \frac{D_\mathrm{P}\,(p_0-p*)}{RT}\,t \tag{80}$$

In the early stages of reduction the porous iron layer is thin and, hence, the path for the diffusing gas is short. Thus, initially the reduction rate is determined by the phase boundary reaction. In the later stages the gas diffusion becomes rate determining because of the increase in length of the diffusion path; this was experimentally verified.

For the experimental verification, reference is made to Equations (79) and (80). For a given degree of reduction $\Re$, a given ore and a fixed composition of the reducing gas the following expressions may be derived from Equations (79) and (80)

$$\log t = \log K_1 + \log r_0 \tag{81}$$

$$\log t = \log K_2 + 2 \log r_0 \quad (82)$$

The experimentally observed relation between log t and log r_0 is shown in Figure 33 for the reduction of hematite pellets in hydrogen at 750 °C and pellet radii between 8.5 and 17.5 mm. According to Equations (81) and (82) the slope of the lines should indicate whether the reduction time is primarily determined by the interfacial reaction or by gaseous diffusion. For low degrees of reduction the slope is closer to one whereas for higher degrees of reduction it is closer to two. From these observations it is evident that in the beginning the reduction rate is primarily determined by the interfacial reaction whereas in the later stages pore diffusion becomes rate determining.

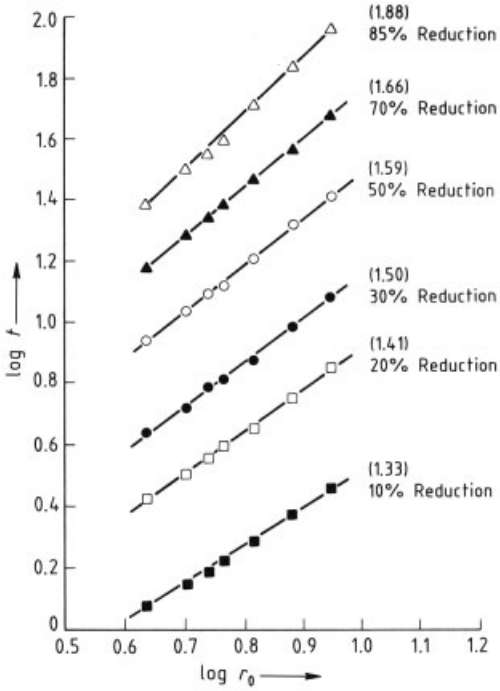

Figure 33. Reduction time t (min) needed to attain a given degree of reduction as a function of the pellet radius r_0 (mm) [19]

Porous Ore Particles, Porous Iron Layers. As mentioned in the beginning of this section, the mathematical expressions so far discussed are valid for the reduction of dense particles on which porous iron layers are formed during the reduction. The mathematical treatment of the reduction, particularly of porous iron oxide particles, is due to L. von Bogdandy and H. G. Riecke [52] as based on concepts proposed by E. W. Thiele. The basic premise is that the reducing gas diffuses into the spherical ore body and that the reaction occurs on the entire internal surface area of the sphere. Here the phase boundary reaction rate is defined by a volumetric reaction rate constant $k_V = k' A/V$ (s^{-1}), where A/V is the internal surface area per volume of the ore particle. The spherical ore body is divided into infinitely small shells of volume $4\pi r^2 dr$, where r is the radial coordinate. The reaction rate dv is the number of moles of oxygen removed from the ore per unit of time in this shell. As the removal of oxygen generates a corresponding decrease of the reducing gas concentration in the volume element, the concentration of this gas is higher at the position $r + dr$ than at the position r. The concentration difference generates an inward-directed diffusion flow of the reducing gas, for instance CO, and an outward flow of the reaction product, for instance CO_2. The difference of the diffusion flow $d\dot{n}_{r+dr}$ at the position $r + dr$ and $d\dot{n}_r$ at the position r must be equal to the consumption of gas by the phase boundary reaction dv. The phase boundary reaction is expressed as

$$dv = 4\pi r^2 dr \frac{k_V}{RT}(p_r - p*) = 4\pi r^2 dr \frac{k_V}{RT}\Delta p_r \quad (83)$$

where p_r is the partial pressure of the reducing gas in the volume element at radius r. The pressure difference $p_r - p*$ can be expressed as Δp_r . The difference of the diffusional flow between $r + dr$ and r is obtained by derivation of the diffusion equation with respect to r. For the diffusing gas CO, for example, the diffusional flow is expressed as

$$\dot{n}_{CO} = 4\pi r^2 \frac{D_P}{RT}\frac{(p_{r+dr} - p_r)}{dr} = 4\pi r^2 \frac{D_P}{RT}\frac{d(\Delta p_r)}{dr} \quad (84)$$

Derivation with respect to r yields

$$d\dot{n}_{CO} = 8\pi r dr \frac{D_P}{RT}\frac{d(\Delta p_r)}{dr} + 4\pi r^2 \frac{D_P}{RT}\frac{d^2(\Delta p_r)}{dr^2} dr \quad (85)$$

As $d\dot{n}_{CO}$ in moles per time is equal to dv in moles per time, Equations (83) and (85) must be equal. That yields

$$\frac{k_V}{D_P}(\Delta p_r) = \frac{2}{r}\cdot\frac{d(\Delta p_r)}{dr} + \frac{d^2(\Delta p_r)}{dr^2} \quad (86)$$

This equation contains the second derivative of the partial pressure difference because not only the partial pressure gradient for diffusional flow as in the previously described case of an inert porous iron layer, but also an additional change of partial pressure due to the chemical reaction in the pore must be considered.

The boundary conditions for the solution of Equation (86) are found as follows. The pressure difference Δp_r at the outer surface r_0 is Δp_0. In the core at $r = 0$, Δp_r has a minimum. Thus $\Delta p_r/dr$ is zero at this point. With these two boundary conditions, Equation (86) can be solved to give

$$\frac{\Delta p_r}{\Delta p_0} = \frac{p_r - p^*}{p_0 - p^*} = \frac{r_0}{r}\cdot\frac{\sinh\left[\sqrt{k_V D_P}\cdot r\right]}{\sinh\left[\sqrt{k_V/D_P}\cdot r_0\right]} \quad (87)$$

The equation describes the dimensionless pressure difference $(p_r - p^*)/(p_0 - p^*)$ as a function of the dimensionless radius r/r_0 of the ore sphere. The ore radius r_0 and the ratio between the reaction rate constant and the pore diffusion coefficient k_V/D_P are the parameters. The functional relationship between $(p_r - p^*)/(p_0 - p^*)$ and r for various values of the parameters is shown in Figure 34. The results are discussed later.

Before discussing Figure 34, a more general consideration is useful. To this end the dimensionless parameter $k_V\, r_0^2/D_P$, which combines the influences of r_0 and k_V/D_P is introduced. According to the value of this parameter Equation (87) has two limiting cases, namely $k_V\, r_0^2/D_P >> 1$ or $k_V\, r_0^2/D_P << 1$. In the first case, the phase boundary reaction is fast and/or the particle radius is large. As a consequence of the fast reaction, reduction starts at the surface. The CO partial pressure p_0 falls steeply to the equilibrium value p^* within a small layer of thickness Δr. The particle behaves like a dense ore. During the reduction process the layer and with it the steep partial pressure gradient of the reducing gas shifts gradually from the outer surface into the ore body. This is schematically shown in Figure 35. In the second case, pore diffusion is fast in comparison to the phase boundary reaction, and/or the particle radius is small. In this case the gas is distributed infinitely fast and uniformly over the entire particle. At any point of the particle, the partial pressure is p_0, and hence, the driving force has its maximum Δp_0. This latter case represents the maximum possible reduction rate v_{max}. It follows from Equation (83) that

$$v_{max} = \frac{4}{3}\pi r_0^3 \frac{k_V}{RT}\Delta p_0 \quad (88)$$

For the general case the actual reaction rate v is obtained by inserting Δp_r from Equation (87) into (83) and then integrating Equation (83) between $r = r_0$ and $r = 0$. This yields, if the reduction rate is dimensionlessly expressed by

$$\dot{\Re} = \frac{v}{\frac{4}{3}\pi r_0^3 \cdot d_O} \quad (89)$$

and if $\varphi = r_0(k_V/D_P)^{\frac{1}{2}}$ is inserted in Equation (87):

$$\dot{\Re} = \frac{v}{\frac{4}{3}\pi r_0^3 \cdot d_O} = \frac{3\, D_P}{r_0^2 \cdot d_O\, RT}\Delta p_0 \left[\coth\varphi - \frac{1}{\varphi}\right] \quad (90)$$

The reduction rate can be expressed in terms of the maximum reduction rate as $v/v_{max} = \eta$. In this way, η expresses the effectivity of the reduction. Then, it follows from Equations (88) and (90)

$$\eta = \frac{3}{\varphi}\left(\frac{1}{\varphi} - \coth\varphi\right) \quad (91)$$

From Equation (91) it follows that

$$\lim_{\varphi\to 0}\eta = 1$$

and hence

$$\dot{\Re} = \frac{k_V}{d_O\, RT}\Delta p_0 \quad (92)$$

and that

$$\lim_{\varphi\to\infty}\eta = \frac{3}{\varphi}$$

and hence

$$\dot{\Re} = \frac{3\cdot(k_V D_P)^{1/2}}{r_0 d_O\, RT}\Delta p_0 \quad (92a)$$

Thus, at the maximum reduction rate ($\varphi \to 0$) the rate is independent of the particle radius and depends on the phase boundary reaction rate only. As mentioned before, reduction takes place uniformly over the entire ore body as at each point of the ore the maximum partial pressure difference Δp_0 is present. At $\varphi \to \infty$ the rate is inversely proportional to the particle radius and

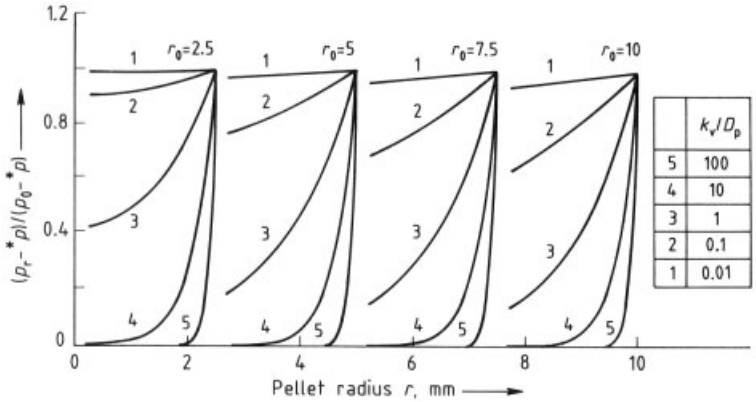

Figure 34. Partial pressure of the reducing gas in the interior of the porous ore sphere during the formation of porous iron

dependent on the square root of the product of the phase boundary reaction rate constant and the pore diffusion coefficient. Reduction takes place in a small reaction layer moving into the ore body. Reduction is determined by diffusion in the pores and by the phase boundary reaction rate constant k_V in the reaction layer.

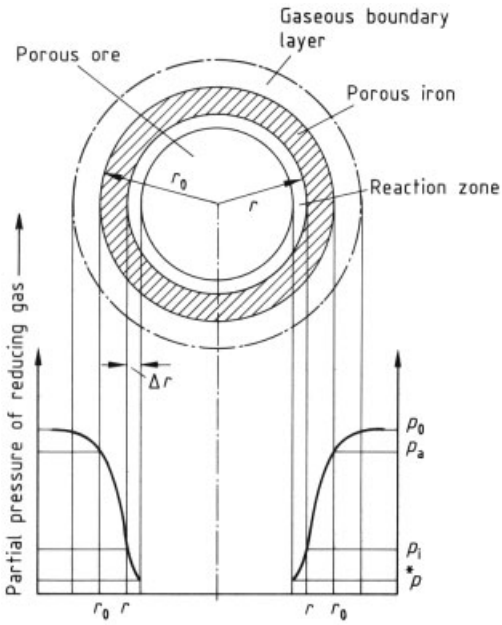

Figure 35. Partial pressure profile inside a porous ore sphere during formation of a reduction zone Δr

Between the limiting cases $\varphi = 0$ and $\varphi = \infty$ the dimensionless concentration gradient $(p_r - p^*)/(p_0 - p^*)$ in the ore is the more steep the larger the parameter k_V/D_P and the larger the particle radius r_0. This is shown in Figure 35.

Synopsis. A summary of the effects of various phenomena concerning the reduction rate, as obtained from the partially overlapping models of McKewan and von Bogdandy et al., is presented in Table 11. The material constants to be used in approximate calculations are:

k' = interfacial reaction rate constant,
A = activation energy of the interfacial reaction,
D_i = diffusion coefficient of the reducing gas component i,
γ = ore porosity,
ξ = tortuosity factor,
a_i = internal surface area of the ore.

The values of k' and A as determined by various investigators may be obtained from Figure 36 (lines 1 – 21).

Particular attention should be paid to the following.

- The *reaction rate constant* for the reduction of the homogeneous wustite phase, i.e., without metallic iron formation, is significantly lower than that observed when metallic iron has been produced.
- The *activation energy* for the reduction in the presence of iron is significantly lower than that observed prior to the formation of metallic iron.

Table 11. Dependence of the reduction rate on various experimental parameters (k' = rate constant of interfacial reaction; [52, 53]. A = activation energy of interfacial reaction; D_i = diffusion coefficients of reducing gases; γ = ore porosity, ξ = tortuosity factor)

Conditions	Functional Relationship between $\Re$ and				Remarks	Model of
	Radius of ore sphere, r_0	Difference in partial pressure of reducing gas, $p_0 - p^*$	Total pressure, p_{tot}	Temperature, T		
$r_0 k' < D_P$	$\sim r_0^{-1}$	$\sim (p_0 - p^*)^a$	$\sim p^0$	$\sim \exp - A/RT$	low temperature[c] higher temperature, r_0 small, or r_0 large but only at initial reduction stage [d]	McKewan [52]
$r_0 k' > D_P$	$\sim r_0^{-2}$	$\sim (p_0 - p^*)^a$	$\sim p^{0a}$ $\sim p^{0b}$	$\sim T^{-1.7}$	higher temperature, r_0 large [d]	
$r_0^2 k_V < D_P$	$\sim r_0^0$	$\sim (p_0 - p^*)^a$	$\sim p^0$	$\sim \exp - A/RT$	lower temperature, r_0 small [d]	von Bogdandy, Riecke [53], reaction and diffusion in porous oxide
$r_0^2 k_V > D_P$	$\sim r_0^{-1}$	$\sim (p_0 - p^*)^a$	$\sim p^{0\ a}$ $\sim p^{0b}$	$\sim \exp - A/2RT$	higher temperature, r_0 large [d]	

[a] For the case of gas diffusion primarily in macropores, mean free path $\lambda < d$ (pore diameter), i.e., high pressure.
[b] $\lambda > d$, i.e., low total pressure.
[c] No good agreement with experiment.
[d] Good agreement with experiment.

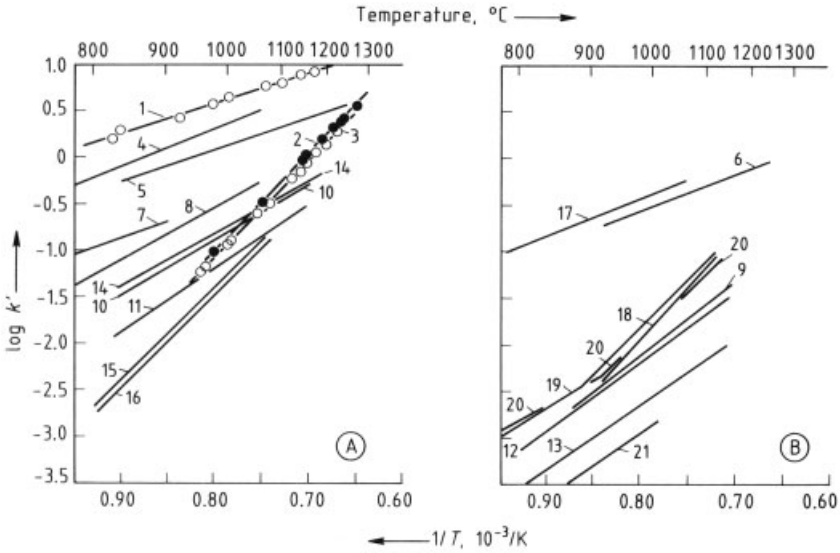

Figure 36. A) Temperature dependence of the rate constant k' (cm/s) for the reduction of wustite; B) activation energy of the reduction reaction [49]

– The *reaction rate constants* for the *reduction* with *hydrogen* are higher than those for carbon monoxide.

The presence of metallic iron is responsible for an increase in the reduction rate, as a result of the catalytic effect of iron, causing a decrease in the temperature dependence of the rate. The values shown in Figure 36 are valid for the reduction of pure wustite. The reaction rates for wustite doped with other ions may be significantly different.

The *diffusion coefficients* D_i may be calculated from the expression [50]:

$$D_{12} = D_{21} \simeq 1.84 \cdot 10^{-1} \frac{T^{3/2}}{P\sigma_{12}^2} \left(\frac{M_1 + M_2}{M_1 M_2} \right)^{1/2} \quad (93)$$

where

$$\sigma_{12} = (\sigma_1 + \sigma_2)/2$$

is the average collision diameter of the gas molecules in mm; P the total gas pressure in bar; M_1, M_2 the molecular mass of the gas components. For details of calculation reference is made to [50].

The various ores and agglomerates differ considerably with respect to porosity, tortuosity factor and internal surface area, even prior to the reduction. Calculations are complicated by the fact that these quantities change during reduction in an unpredictable manner. Therefore, the values shown in Table 12 serve as guidelines only. The change in porosity and internal surface area during the reduction of pellets has been measured (see Fig. 37).

Table 12. Values of porosity γ and inner surface area a_i for various ores

Ore	γ, %	a_i, cm^2/cm^3
Kiruna, magnetite	1.3	0.38×10^3
Steep Rock	15.3	1.8×10^3
Mesabi Range, hematite	25 – 40	
Sinter (micropores <14 µm)	1 – 18	
Pellets	20 – 35	2×10^3

Hematite pellets (curve marked "unreduced" in Fig. 37) were prereduced to wustite at 750 °C. Using this wustite as a starting material, the reduction was interrupted at several predetermined degrees of reduction and porosity and internal surface areas were measured with a mercury porosimeter. The porosity increases from 25 % (hematite) to 35 % (wustite) and then to 50 % at a reduction degree of ca. 50 % (50 % iron, 50 % wustite), primarily because of an increase in the number of pores with radii $r < 5.10^{-5}$ cm that were hardly present in the starting material. As a result the internal surface area increases by about a factor of ten.

The mean free path λ of the reducing gas molecules is of the same order of magnitude as the radii of those pores that provide the major contribution to the internal surface area of the pellets (see Table 13). At normal pressure diffusion in these pores takes place via Knudsen diffusion.

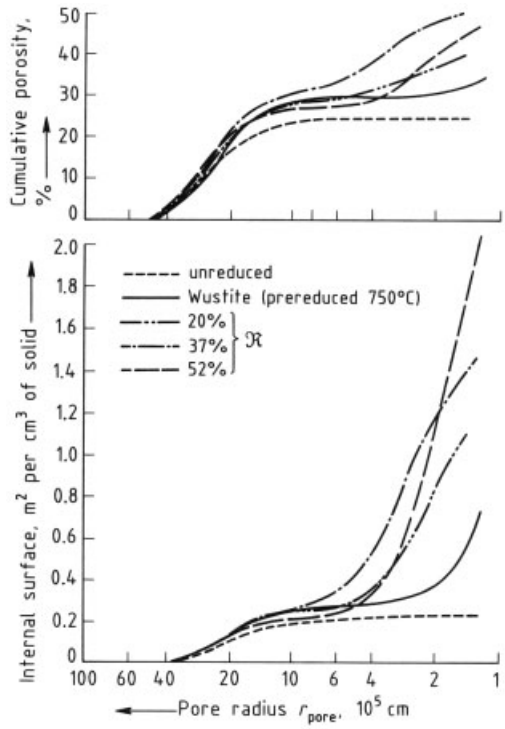

Figure 37. Internal surface area and porosity of pellets as a function of the pore radius [45]

The evaluation of the *tortuosity factor* ξ can only be made indirectly via the determination of the pore diffusion coefficient D_{P}. However, ξ is usually used to calculate D_{P} from the diffusion coefficient D in the bulk gas phase. Therefore, the value of ξ is only of interest in modeling pore diffusion. The calculated values of ξ vary between 0.06 and 0.4 [19]. As a first approximation the following expression is valid where D_i has been calculated from Equation (93).

$$D_{\mathrm{P}} \simeq 0.2 D_i \quad (94)$$

Additional Formulations. R. Jeschar [54] proposed a dimensionless formulation of Equation (78) which is expressed here as follows:

$$\frac{(p_0 - p*)}{RT \cdot d_{\mathrm{O}}} Fo = \frac{1}{Da} \left[1 - (1 - \Re)^{1/3} \right] - \frac{1}{2} \left[1 - (1 - \Re)^{2/3} \right] - \frac{1}{3} \left(1 - \frac{1}{Sh} \right) \Re \quad (95)$$

Table 13. Mean free path $\overline{\lambda}\times 10^5$ (cm) in binary gas mixtures at a total pressure of 0.1 MPa

Temperature, K	Mean free path $\overline{\lambda}$ in mixtures of			
	$CO-N_2$	H_2-N_2	H_2-H_2O	$CO-CO_2$
773	2.64	3.34	3.57	2.63
973	3.58	4.44	4.76	3.71
1173	4.52	5.62	6.01	4.91
1373	5.53	6.84	7.32	6.22

with

$$Fo = \frac{D_P \cdot t}{r_0^2} \text{ (Fourier number)}$$

$$Da = \frac{k\prime \cdot r_0}{D_P} \text{ (Damköhler number)}$$

$$Sh = \frac{\beta \cdot r_0}{D_P} \text{ (Sherwood number)}$$

The term 1/3 $\Re/Sh$ expresses the diffusion in the outer gas boundary layer, which was already proven to have a negligible influence. It is not part of Equation (78). Except that term, (78) and (95) are in agreement. The advantage of using dimensionless quantities is that the number of process-relevant variables is reduced. In the present case, the five variables D_P, t, r_0, k', and β are reduced to the three really independent dimensionless variables Da, Fo, and Sh.

In the model for reduction of porous ores, Equation (92a) describes the limiting case that the phase boundary reaction is fast in comparison to pore diffusion. In this case reduction takes place mainly in a small reaction layer, moving gradually from the ore surface to the core during reduction. The layer has no sharp limits, but in practice the outer part of the ore body behind the moving layer consist almost exclusively of porous iron, and the inner part in front of the layer is unreduced ore. Under these circumstances Equation (78) is applicable, if the condition of an infinitely thin reaction layer as assumed in Equation (78) is replaced by the assumption of a layer of finite thickness. This was done by G. Bitsianes and coworker [55]. The ore model is shown in Figure 35 with a layer of thickness Δr, where the concentration shows a course as shown in the lower part of Figure 35. Equation (78) is altered by taking into account the layer thickness Δr and the inner surface of the layer. The original Equation (78) is expressed as

$$\frac{(p_0-p*)}{RT\cdot d_O}t = \frac{r_0}{k'}\left[1-(1-\Re)^{1/3}\right] + \frac{r_0^2}{6\,D_P}\left[1+2\,(1-\Re)-3(1-\Re)^{2/3}\right] \tag{78a}$$

while Bitsianes et al. published the following expression

$$\frac{\Delta c}{d_O}t = \frac{r_0}{k'\sigma\,\eta}\left[1-(1-\Re)^{1/3}\right] + \frac{r_0^2}{6\,D_P}\left[1+2\,(1-\Re)-3(1-\Re)^{2/3}\right] + \frac{r_0}{3\,\beta}\Re \tag{96}$$

The last term takes into account the mass transport through the outer gas boundary layer which is negligible as mentioned. The rest differs from Equation (78a) by the product $\sigma\eta$ in the first term on the right-hand side. This term takes into account the finite thickness of the reduction layer; η is the reduction effectivity v/v_{max}; σ is the internal surface area available in the reaction zone in addition to the area of the spherical shell expressed in terms of the total shell area:

$$\sigma = \Delta r \cdot a_i\left(1+\frac{\Delta r}{r}\right)+1 \tag{97}$$

where Δr is the thickness of the reaction layer; Δr must be taken from practical experience.

2.3.2. Reduction in the Liquid State

2.3.2.1. General

In the liquid state, the reduction of iron oxide is essentially limited to the reduction of FeO, because above 1400 °C, Fe_2O_3 has an oxygen partial pressure of decomposition of more than 1 bar, and also Fe_3O_4 is thermodynamically stable only under high $CO_2/(CO + CO_2)$ ratios which are normally not present in reduction processes for the liquid state. Thus, consideration is confined to the reduction of FeO from liquid FeO-containing slags. This reduction is an important process in the blast furnace and in smelting reduction reactors. The primary reducing agent can be either solid carbon (in the form of coke or

char), carbon dissolved in liquid iron, or carbon monoxide. However – in the liquid state just as in the solid state – the reduction itself is only possible with gas, preferably carbon monoxide. A direct reduction of FeO from a liquid slag by carbon is not possible, neither with solid carbon nor with carbon dissolved in liquid iron. This is because a gas layer of a carbon monoxide – carbon dioxide mixture forms immediately between the surface of the carbon containing phase and the surface of the FeO-containing slag in the moment when both surfaces come into contact. The reason is the great difference in oxygen potential between slag and the carbon-containing phase. Due to the gas layer, carbon monoxide can be the only direct reducing agent. The role of carbon is the production of carbon monoxide from carbon dioxide according to the Boudouard reaction. In combination with the FeO reduction ($CO + FeO \rightarrow Fe + CO_2$) the overall reaction equation $C + FeO \rightarrow Fe + CO$ results. An experimental study and a formal treatment of the kinetics of FeO reduction from a liquid slag requires, therefore, a separate consideration of the Boudouard kinetics and of the reduction kinetics under the conditions of the liquid state.

Laboratory experiments in which FeO is reduced from a slag by solid carbon or liquid carbon containing iron, give vigorous CO gas bubble evolution and thus, an uncontrolled reduction process, from which rate constants for the single steps cannot be obtained. Such studies give a qualitative impression of the behavior of the entire reduction system. But the results cannot be transferred to large industrial vessels as the vessel sizes, the degrees of bubble – melt dispersion, and the stirring conditions are different.

2.3.2.2. Microkinetics

In processes where liquid hot metal is produced, the reduction reaction proceeds formally according to Equation (30):

$$C + (FeO) \rightleftharpoons CO + [Fe] \tag{30}$$

Four different phases participate in this reaction, namely (1) liquid iron oxide containing slag, (2) solid carbon or carbon-saturated iron, (3) carbon monoxide, and (4) metallic iron as the reaction product. The product precipitates either in the solid or in the liquid state depending on the reaction temperature.

The existence of four different phases necessitates a reaction in three steps at different locations (see Fig. 38). On the left-hand side of Figure 38 a carbon-saturated iron droplet is shown, which may also be a carbon particle. The droplet is surrounded by the aforementioned gas layer which on its outer side adjoins the iron-oxide containing slag. (If coke is present instead of a carbon-containing droplet as is the case in the blast furnace, an irregularly shaped gas layer exists between the coke particle and the liquid slag).

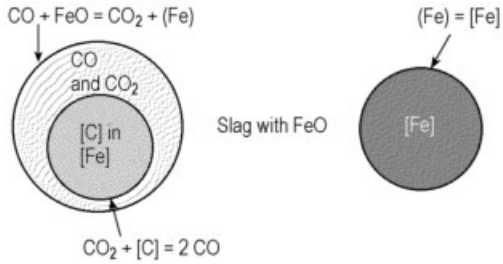

Figure 38. Schematic representation of the reduction process in the slag

The following reactions occur:

At the interface between carbon-saturated iron droplet and gas:

$$CO_2 + C \rightleftharpoons 2\,CO \tag{98}$$

At the interface between gas and slag:

$$CO + FeO \rightleftharpoons CO_2 + (Fe) \tag{99}$$

where (Fe) denotes iron dissolved in the slag. In reality, iron exists in the slag in the form of di- and trivalent ions, Fe^{2+} and Fe^{3+}. Reaction (99) can, therefore, also be expressed as

$$CO + FeO + 2\,Fe^{3+} \rightleftharpoons CO_2 + 3\,Fe^{2+} \tag{100}$$

Both expressions are equivalent. The thermodynamic activity $a_{(Fe)}$ of dissolved iron in the slag is given by the equilibrium of the reaction $3\,Fe^{2+} \rightleftharpoons (Fe) + 2\,Fe^{3+}$. Hence, $a_{(Fe)}$ can be expressed as the temperature-dependent value:

$$a_{(Fe)} = K_{Fe^{2+}/Fe^{3+}} \frac{a^3_{Fe^{2+}}}{a^2_{Fe^{3+}}} \tag{101}$$

The $(a_{Fe^{2+}})^3/(a_{Fe^{3+}})^2$ ratio is a function of the $(\%\,Fe^{2+})/(\%\,Fe^{2+} + \%\,Fe^{3+})$ value in the slag.

If $a_{(\mathrm{Fe})} = 1$ in Equation (101), thermodynamic equilibrium exists between slag and liquid pure iron.

In the gas layer between droplet and slag surface (cf. Fig. 38 left), CO_2 and CO move countercurrent to one another.

The result of Equations (98) and (99) is the separation of 1 mol oxygen from the slag and 1 mol carbon from the carbon-containing droplet which form 1 mol of carbon monoxide, and the transfer of 2 mol trivalent iron and 1 mol FeO into 3 mol bivalent iron, which is equivalent to the formation of one mole dissolved iron in the slag. Because of the production of carbon monoxide, the gas layer around the iron droplet grows. At a certain critical layer size a bubble separates. This process takes place continuously so that carbon monoxide bubbles are periodically evolved.

As a consequence of Equation (99) or (100) the iron activity in the slag increases and eventually becomes unity, or the slag becomes supersaturated with dissolved iron. That means $a_{(\mathrm{Fe})} \geq 1$. The reaction product precipitates from the slag phase as metallic iron. This is similar to the situation in the reduction of solid wustite (see Figs. 32 A and B). Precipitation occurs at a different location than the one of oxygen separation from the slag. This is shown on the right-hand side of Figure 38. The reaction is expressed as

$$(\mathrm{Fe}) = [\mathrm{Fe}] \tag{102}$$

where [Fe] is metallic iron.

Iron precipitation may also take place at the gas – slag interface. Both possibilities correspond to those shown in Figures 32 A and B for the case of solid oxide.

Kinetic treatment of the reaction sequence presented in Figure 38 is carried out in two steps [56, 57]. The first one is a description of the microkinetics at the single bubble and droplet as shown in Figure 38. The second step is a summation of the microkinetic steps over the entire population of droplets and bubbles present in the reactor. Calculations of this type are known for conditions of smelting reduction. In this process carbon-saturated iron droplets are suspended in a slag in which the reduction occurs. Simultaneously, the slag dissolves iron oxide from the charged ore. Reduction rate of iron oxide and charging rate of ore are equivalent. Thus, in the steady state a constant iron oxide content exists in the slag.

According to Figure 38 the single microkinetic steps are:

1) Phase boundary reaction at the interface between carbon-saturated iron droplet and gas (Boudouard reaction), Equation (98)
2) CO and CO_2 transport in the surrounding gas layer
3) Phase boundary reaction at the interface between gas and slag, Equation (99)
4) Transport of iron oxide in the slag
5) Phase boundary reaction of iron precipitation, Equation (102)

Steps 2 and 5 may be neglected as it can be shown that they are fast [56, 57]. Thus, steps 1, 3, and 4 remain.

The forward reaction of step 1 is expressed as

$$-\left(\frac{\mathrm{d}n_{\mathrm{CO_2}}}{\mathrm{d}t}\right)_{\mathrm{G-M}} = A_{\mathrm{G-M}} k_{\mathrm{G-M}} \frac{P_{\mathrm{CO_2},i}}{RT} a_{[\mathrm{C}]} = A_{\mathrm{G-M}} k_{\mathrm{G-M}} \frac{P_{\mathrm{tot}} - P_{\mathrm{CO},i}}{RT} a_{[\mathrm{C}]} \tag{103}$$

where $A_{\mathrm{G-M}}$ is the gas – metal interface area, $k_{\mathrm{G-M}}$ the reaction rate constant, $P_{\mathrm{CO_2},i}$ and $P_{\mathrm{CO_2},i}$ the respective CO_2 and CO partial pressures at the interface, and $P_{\mathrm{tot}} = P_{\mathrm{CO_2},i} + P_{\mathrm{CO},i}$. P_{tot} is usually 1 bar. The back reaction needs not to be considered because the equilibrium lies almost completely on the side of carbon monoxide at the reaction temperature of 1300 – 1500 °C.

The phase boundary reaction at the gas – slag interface (step 3) is expressed as

$$\left(\frac{\mathrm{d}n_{\mathrm{CO_2}}}{\mathrm{d}t}\right)_{\mathrm{G-S}} = A_{\mathrm{G-S}} k_{\mathrm{G-S}} \left(\frac{P_{\mathrm{CO},i}}{RT} a_{\mathrm{FeO},i} - \frac{a_{(\mathrm{Fe})} P_{\mathrm{CO_2},i}}{K \cdot RT}\right) \tag{104}$$

Here, $A_{\mathrm{G-S}}$ and $k_{\mathrm{G-S}}$ are the reaction interface area and rate constant, respectively. The terms in parentheses express the difference in the driving force between the forward and the back reaction. $a_{\mathrm{FeO,i}}$ and $a_{(\mathrm{Fe})}$ are the respective iron oxide and iron activities in the slag at the interface, and K is the equilibrium constant

$$K = \frac{P_{\mathrm{CO_2},i} a_{(\mathrm{Fe})}}{P_{\mathrm{CO},i} a_{\mathrm{FeO},i}} \tag{105}$$

$a_{(\mathrm{Fe})}$ may be taken as unity over the entire slag as the precipitation reaction is fast.

Iron oxide transport in the slag (step 4) is expressed as

$$\left(\frac{\mathrm{d}n_{\mathrm{FeO}}}{\mathrm{d}t}\right)_{\mathrm{G-S}} = A_{\mathrm{G-S}}\beta_{\mathrm{G-S}}\left[c_{\mathrm{FeO}} - c_{\mathrm{FeO},i}\right] \tag{106}$$

where $\beta_{\mathrm{G-S}}$ is the mass transfer coefficient of iron oxide in the slag, and c_{FeO} and $c_{\mathrm{FeO,i}}$ the respective iron oxide concentrations in the bulk slag and at the gas – slag interface.

The molar carbon dioxide production after Equation (103), the molar carbon dioxide consumption after Equation (104), and the molar iron oxide flux after Equation (106) must be equal. With this continuity condition, Equations (103), (104), and (106) can be combined to yield a flux equation of the overall microkinetics. Attention must be paid in this context to the fact, that the bubble volume around the carbon-containing iron droplet increases gradually with periodical separation of a carbon monoxide bubble, which then rises in the slag. The interface area $A_{\mathrm{G-S}}$ oscillates therefore. It is useful to make calculations with a time average of the interface area $A_{\mathrm{G-S}}$. If the value α^2 is introduced, where α is the ratio of the droplet to the bubble radius, the bubble interface area $A_{\mathrm{G-S}}$ may be expressed by the droplet surface area $A_{\mathrm{G-M}}$ as $A_{\mathrm{G-S}} = \alpha^2 \cdot A_{\mathrm{G-M}}$. α is given by the critical diameter for bubble separation which depends on the carbon monoxide evolution rate and on the slag surface properties, primarily on the surface tension. With these considerations, it follows from Equations (103), (104), and (106):

$$\left(\frac{\mathrm{d}n_{\mathrm{FeO}}}{\mathrm{d}t}\right)_{\mathrm{G-S}} = A_{\mathrm{G-M}} \cdot \left[\frac{c_{\mathrm{FeO}}}{\frac{1}{K\,k_{\mathrm{G-M}}{}^{a}[\mathrm{C}]^{b}} + \frac{1}{\alpha^2 k_{\mathrm{G-S}}{}^{b}} + \frac{1}{\alpha^2 \beta_{\mathrm{G-S}}}}\right] \tag{107}$$

with

$$b = \frac{P_{\mathrm{CO},i}\, f_{\mathrm{FeO}} V_{\mathrm{S}}}{RT\, n_{\mathrm{tot}}}$$

Here, f_{FeO} is the activity coefficient of FeO in the slag, V_{S} the slag volume, and n_{tot} the number of moles in the slag.

2.3.2.2.1. Numerical Data for the Calculation of Microkinetics

First, the mass transfer coefficient in the slag $\beta_{\mathrm{G-S}}$ must be determined. The expression for spherical cap bubbles can be used as approximation in this case. The mass transfer coefficient is expressed as [58]

$$\beta_{\mathrm{G-S}} = 0.80 \cdot r_{\mathrm{eq}}^{-1/4} D^{1/2} g^{1/4} \tag{108}$$

where r_{eq} is the radius of the sphere of equal volume, g the acceleration due to gravity , and D the diffusion coefficient of FeO in liquid slags expressed by [57 – 63]

$$\ln D_{\mathrm{FeO}} = -14.335 - \frac{10716}{T} \tag{109}$$

Equation (109) is valid for low FeO-containing slags with, for example, 40 % CaO, 30 % SiO_2, 12 % Al_2O_3, 15 % MgO, and 1.5 – 5 % FeO. It is used up to ca. 30 % FeO. The rate constant of the Boudouard reaction $k_{\mathrm{G-M}}$ was frequently measured for pure iron – carbon alloys as well as for iron – carbon alloys with third elements dissolved in it. Sulfur is of special importance for the practice of iron ore reduction due to its strong effect on the reduction rate. The following expression by Sain and Belton [64] can be used for pure iron carbon alloys

$$k_{\mathrm{G-M}} = \exp\left(-\frac{11700}{T} + \ln T - 0.68\right) \tag{110}$$

with k in m/s. The influence of sulfur on the rate constant of Equation (110) is expressed by an extension of that equation as follows [65]:

$$k_{\mathrm{G-M}} = \exp\left(-\frac{11700}{T} + \ln T - 0.68\right) \cdot \frac{1+\kappa \cdot A \cdot Y_{\mathrm{S}}}{1+A \cdot Y_{\mathrm{S}}} \tag{111}$$

where κ is the part of the interface not covered with sulfur (= 0.014); A the adsorption coefficient of sulfur: log A = 3600/T + 0.57. Y_{S} the sulfur content in the iron melt (mass %). Equation (111) is valid for sulfur contents in liquid iron of 0 – 1.0 mass %. For 1773 K and 1873 K Equation (110) yields $k_{\mathrm{G-M}}$ = 1.22 m/s and $k_{\mathrm{G-M}}$ = 1.84 m/s, respectively. Equation (111) yields for 1873 K and 1 mass % sulfur: $k_{\mathrm{G-M}}$ = 0.032 m/s. That means that under otherwise equal conditions the reaction rate in the second case is roughly 50 times slower than in sulfur-free iron.

For the rate constant $k_{\mathrm{G-S}}$ of iron oxide reduction from a liquid iron oxide containing slag by carbon monoxide, Equation (112) was obtained from measurements between 1593 and 1723 K [66]

$$k_{\mathrm{G-S}} = T \cdot \exp\left(-\frac{16618}{T} - 0.11033\right) \tag{112}$$

with k in m/s. For 1773 K and 1873 K values of k_{G-S} = 0.135 m/s and k_{G-S} = 0.235 m/s are calculated.

The equilibrium constant of Equation (105) is expressed by [67]

$$\ln K = \frac{5946}{T} - 4.98 \tag{113}$$

Raoult's activity coefficient f_{FeO} is expressed by [68]

$$\begin{aligned} 8.314T \cdot \ln f_{FeO} = & -41840X_{SiO_2}^2 - 31380X_{CaO}^2 \\ & -41100X_{Al_2O_3}^2 + 33470X_{MgO}^2 \\ & +60670X_{SiO_2}X_{CaO} + 44770X_{SiO_2}X_{Al_2O_3} \\ & +58570X_{SiO_2}X_{MgO} + 82430X_{CaO}X_{Al_2O_3} \\ & +102510X_{CaO}X_{MgO} + 63600X_{Al_2O_3}X_{MgO} \\ & -8540 + 7.142T \end{aligned} \tag{114}$$

For the aforementioned slag, this yields f_{FeO} = 4.3. Raoult's activity coefficient of carbon in liquid iron can be expressed by [62]

$$f_{[C]} = \frac{1155 \cdot \exp\left(9.9\, X_C + 11.89\, X_S\right)}{T} \tag{115}$$

where X in Equations (114) and (115) denotes mole fractions.

With the presented data, the numerical values of the single reaction resistances in the denominator of Equation (107) can be determined. Table 14 shows the influence of sulfur on the rate constant k_{G-M} and the reciprocal resistance $K \cdot k_{G-M} \cdot b \cdot [a]_C$ for $[a]_C$= 1. The other two reciprocal resistances in Equation (107) $k_{G-S} \cdot b$ and β_{G-S}, both for α= 1. The mass transfer coefficient β depends on the droplet diameter. Two values, one for the smallest and the other for the largest class of droplet diameter are, therefore, presented. On the gas – slag interface the reaction rate constant $k_{G-S} \cdot b$ is a little smaller than the mass transfer coefficient β_{G-S}, which means that the main resistance at that interface lies in the phase boundary reaction.

The resistance on the gas – metal side is in the same order of magnitude as that on the gas – slag side only as long as the sulfur content is extremely low. Above [S] = 0.005 % already, the overall resistance lays mainly on the gas – metal side. α values higher than 1 increase the gas – slag reaction interface area and thus, the rate at that interface considerably, as the interface area is proportional to α^2. At very low sulfur contents, where the gas – metal and the gas – slag resistance are in the same order of magnitude, growing α will, therefore, accelerate the overall rate. When the sulfur contents are higher, the resistance at the gas – metal interface solely determines the overall rate, and hence, increasing α has no more an influence. Under these circumstances the gas – slag resistance is so small in comparison with the gas – metal resistance, that thermodynamic equilibrium establishes at the gas – slag interface.

Table 14. Numerical values of the single rate constant at 1500 °C in Equation (107) [57]

Sulfur content in the melt, mass%	k_{G-M}, m/s	$Kk_{G-M}b$, m/s
0.000	1.22	1.28×10^{-4}
0.001	0.88	9.22×10^{-5}
0.002	0.69	7.22×10^{-5}
0.005	0.42	4.41×10^{-5}
0.010	0.26	2.70×10^{-5}
0.020	0.15	1.59×10^{-5}
0.050	0.075	7.84×10^{-6}
0.100	0.047	4.89×10^{-6}
0.200	0.032	3.36×10^{-6}
0.500	0.023	2.43×10^{-6}

$\beta_{G-S,d_{min}}$ = 3.4 $\times 10^{-4}$ m/s; $\beta_{G-S,d_{max}}$ = 1.88 $\times 10^{-4}$ m/s; $k_{G-M}b$ = 0.72 $\times 10^{-4}$ m/s; d_{min} = 0.38 mm; d_{max} = 4.08 mm

2.3.2.2.2. Macrokinetics

Equation (107) describes the microkinetic reaction sequence at a single pair of reaction sites as shown in Figure 38. To describe the overall reduction rate, called macrokinetics, the number of such sites must be known, primarily the number of carbon-containing iron droplets generated per unit time. Iron droplets are ejected from the iron melt into the slag under the influence of rising bubbles of the stirring gas. The ejected droplets stay in the slag for a certain residence time depending on their sinking rates and on the conditions of slag motion. Together with the droplet evolution rate this yields the number of reaction sites present in the slag. Further, the droplets size distribution must be known. An example is given in Table 15.

Table 15. Droplet size distribution [70]

Droplet class j	Droplet diameter d_i, m	Volume portion v_i ,%
1	0.38×10^{-3}	6
2	0.75×10^{-3}	20
3	1.50×10^{-3}	30
4	3.56×10^{-3}	26
5	4.08×10^{-3}	18

The macrokinetic reaction rate $\dot{n}_{\mathrm{FeO}}$ which is reasonably related to the tonnage of the melt and thus, has the unity mol t^{-1} s^{-1}, is then given by amount of reduction $n_{\mathrm{FeO,D},i}$ at a single site in the mole (where i designates the droplet size class) and the number of droplets ejected into the melt per time and per tonne of melt. That yields

$$\dot{n}_{\mathrm{FeO}} = \sum\nolimits_i \frac{\dot{M}_{\mathrm{D}}\sigma_i}{\frac{1}{6}\pi d_{\mathrm{D},i}^3 \varrho_{\mathrm{D}}} \cdot n_{\mathrm{FeO,D},i} \tag{116}$$

where

$$n_{\mathrm{FeO,D},i} = \int_0^{t_{\mathrm{R},i}} \frac{\mathrm{d}n_{\mathrm{FeO}}}{\mathrm{d}t}\mathrm{d}t$$

$\mathrm{d}n_{\mathrm{FeO}}/\mathrm{d}t$ is defined in Equation (107) and $t_{\mathrm{R},i}$ is the droplets residence time in class i. In Equation (116) $\dot{M}_{\mathrm{D}}$ is the droplet formation rate in kg t^{-1} s^{-1}, σ_i the portion of droplets in class i, $d_{\mathrm{D},i}$ the diameter of droplets in class i, and ϱ_{D} the droplet density. Inserting Equation (107) into (116) yields

$$\dot{n}_{\mathrm{FeO}} = 6\,\dot{M}_{\mathrm{D}} \frac{c_{\mathrm{FeO}}}{\varrho_{\mathrm{D}}} \sum\nolimits_i \frac{\sigma_i}{d_{\mathrm{D},i}} \int_0^{t_{\mathrm{R},i}} \beta_{\mathrm{tot},i}\mathrm{d}t \tag{117}$$

with

$$\beta_{\mathrm{tot}} = \frac{1}{\frac{1}{K\,k_{\mathrm{G-M}}a_{[\mathrm{C}]b}} + \frac{1}{\alpha^2 k_{\mathrm{G-S}}b} + \frac{1}{\alpha^2\beta_{\mathrm{G-S}}}} \tag{118}$$

β_{tot} in Equations (116) and (117) is the overall microkinetic mass transfer coefficient which is identical with the reciprocal of the denominator in the brackets of Equation (107). β_{tot} is a function of the droplet carbon content, which varies with time.

Numerical calculation requires knowledge of the data for droplet behavior in Equation (117), i.e., drop formation rate and residence time. This knowledge is still meager in the case of iron oxide smelting reduction processes, but some details are found in the literature [68, 69].

Calculations with Equation (117) for a smelting reduction process [70] are shown in Figure 39. The figure shows the overall reduction rate of FeO as a function of the sulfur content in the iron droplet for a drop formation rate of 5 kg t^{-1} s^{-1} (in practice, the lower limit of this rate), for α = 1, and for the size distribution given in Table 15. The FeO content of the slag is the parameter. The reduction rate decreases strongly with rising sulfur content in the melt because the rate constant $k_{\mathrm{G-M}}$ depends on the melt sulfur content (see Table 14). The influence of α is weak under the conditions chosen. Growing FeO contents in the slag increase the driving force (see Eq. 117) and thus accelerate the reduction rate. The maximum possible reduction rate for smelting reduction conditions – which is limited by the gas throughput – lies between 3 and 5 mol Fe per tonne and second [71]. This rate requires 20 % FeO in the slag at 0.1 % S in the melt, and 30 % FeO in the slag at 0.2 % S in the melt. The necessary FeO contents are inversely proportional to the drop formation rate. Sulfur contents of 0.2 % are reported for smelting reduction conditions [72]. Published slag iron oxide contents in industrial smelting reduction units under steady-state process conditions and at 0.2 % S in the melt are 2 – 5 % [71]. This requires drop formation rates of the order of 50 kg t^{-1} s^{-1}.

2.3.3. Kinetics of Post-Combustion Heat Transfer

In smelting reduction processes post-combustion of process gases with transfer of the post-combustion heat back to the melt is one of the process steps (see Section 2.2.2).

The post-combusted gas has an oxidizing action on the liquid iron below it. Reoxidation of iron must, however, be avoided. Therefore, heat is preferably transferred between gas and melt with the slag as the transfer medium. Heat transfer is mainly accomplished by two methods. The first utilizes slag droplets which are ejected into the gas space and which, after a short residence time fall back into the melt. During their residence in the gas space the droplets are heated up by the gas and return to the melt with a higher temperature than before. The heat content corresponding to this higher temperature is the heat transferred. The second route of heat transfer utilizes post-combustion in a foamy slag. Heat is transferred from rising bubbles of the hot gas to the slag foam and from there by bath circulation to the melt. In the following, the mechanism of heat transfer according to the first route is described as an example [71].

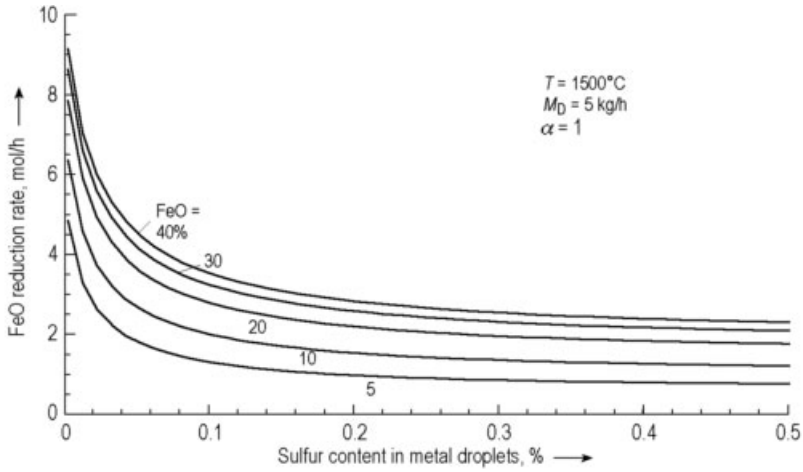

Figure 39. FeO reduction rate at various droplet sulfur contents and at T = 1500 °C. FeO content of slag: 1.5 mass %; droplet formation rate: 10 kg t^{-1} s^{-1}

Post-combustion with Heat Transfer via Slag Droplets. For calculation of the post-combustion reaction and the accompanied heat transfer [71], the stoichiometry and the thermodynamics of the combustion reaction, the heat transfer process, and the heat flow balance must be considered. It is assumed that $\dot{m}$ mol CO evolve from the melt per unit of time and that $\dot{n}$ mol O_2 per unit of time are blown into the gas space above the melt for post-combustion. The value 2 $\dot{n}/\dot{m}$ is the theoretical post-combustion degree

$$PC_{\text{theor}} = 2\,\dot{n}\,/\,\dot{m} \tag{119}$$

Complete combustion is theoretically attained, if $\dot{n}/\dot{m}$ = 0.5. The real degree of combustion can be lower, especially at high temperature, as the combustion reaction is limited by thermal dissociation of CO_2. The degree of dissociation α is expressed as

$$\alpha = \frac{2x_{O_2}}{2x_{O_2}+x_{CO_2}} \tag{120}$$

where x is the molar fraction of the respective component in the gas.

Combustion in the gas space proceeds infinitely fast in comparison to heat transfer. Further, as an approximation it is assumed that the gas is completely mixed. Under these conditions the combustion reaction can be written as:

$$\begin{aligned}\dot{\mathrm{m}}\ \mathrm{CO}\,(\text{at}\,T_{\mathrm{M}}) + \dot{\mathrm{n}}\ \mathrm{O_2}\,(\text{at}\,298\,\mathrm{K}) &= \dot{\mathrm{q}}\ \mathrm{CO_2}\,(\text{at}\,T_{\mathrm{G}}) + \\ (\dot{\mathrm{m}} - \dot{\mathrm{q}}\ \mathrm{CO}\,(\text{at}\,T_{\mathrm{G}}) + \left(\dot{\mathrm{n}} - 1/2\,\dot{\mathrm{q}}\right)\mathrm{CO_2}\,(\text{at}\,T_{\mathrm{G}})\end{aligned} \tag{121}$$

where T_{M} is the bath temperature, T_{G} the gas temperature, and $\dot{\mathrm{q}}$ the number of moles CO_2 generated per unit of time.

The desired gas temperature follows from the heat balance in Equation (121). However, the reaction is incomplete due to CO_2 dissociation. The degree of dissociation depends on the gas temperature. Thus, there are two unknown quantities: the real gas temperature and the degree of dissociation. They have to be calculated by two equations, namely the above heat balance and the equilibrium of CO_2 dissociation. These considerations are valid for the adiabatic case. In reality, part of the post-combustion heat is transferred to the melt, which lowers the gas temperature. Therefore, a third equation describing the heat transfer rate is required in order to completely describe post-combustion with heat transfer. From Equation (121) and from the definitions given in Equations (119) and (120) the fractions of the gas components and the degree of dissociation can be calculated [70]. The obtained data can then be inserted into the equilibrium constant of CO_2 dissociation

$$K_{CO_2} = \frac{x_{CO}\cdot x_{O_2}^{1/2}}{x_{CO_2}} \tag{122}$$

yielding the following expression for the equilibrium constant as a function of the theoretical post-combustion degree and the degree of dissociation:

$$K_{CO_2} = \left[\frac{\alpha}{PC_{\text{theor}}+\alpha}\right]^{1/2} \cdot \frac{PC_{\text{theor}}-2(1-\alpha)}{2(1-\alpha)} \tag{123}$$

The equilibrium constant is a function of the gas temperature. That means, Equation (123) represents one of the two required relations between gas temperature and degree of CO_2 dissociation. A second relation between gas temperature and degree of dissociation follows from the heat flow balance derived from Eq. 121. This balance states that the difference between the enthalpy fluxes after and before post-combustion must be equal to the heat flow which is transferred per unit time from the post-combustion gas to the melt:

$$\begin{aligned}\dot{Q}_t = &-\dot{m}\cdot\left\{\left(H_{T_\mathrm{G}}-H_{T_\mathrm{M}}\right)_\mathrm{CO}\right\}\\ &-\dot{m}\cdot\left\{PC_\mathrm{theor}\,(1-\alpha)\left[\Delta\mathrm{H}^0_{\mathrm{CO}_2}-\Delta\mathrm{H}^0_\mathrm{CO}+\right.\right.\\ &\left.\left.\left(H_{T_\mathrm{G}}-H_{298}\right)_{\mathrm{CO}_2}-\left(H_{T_\mathrm{G}}-H_{298}\right)_\mathrm{CO}\right]\right\}\\ &-\dot{m}\cdot\left\{\tfrac{1}{2}PC_\mathrm{theor}\alpha\left(H_{T_\mathrm{G}}-H_{298}\right)_{\mathrm{O}_2}\right\}\end{aligned} \tag{124}$$

where ΔH^0 denotes enthalpies of formation and H_T enthalpies at temperature T (T_M = melt temperature, T_G = gas temperature). The first line describes the increase of enthalpy of CO when its temperature rises from T_M to T_G. The second line is the combustion heat at gas temperature, and the third the enthalpy required for heating the combustion oxygen from room to gas temperature (gas temperature always means the temperature after combustion). Equation (124) describes the heat balance of post-combustion for a given CO evolution rate from the melt $\dot{m}$ and for a given theoretical post-combustion degree PC_theor. With known heat flow $\dot{Q}_t$, Equations (123) and (124) together allow to determine the degree of dissociation α and the gas temperature T_G. The heat flow $\dot{Q}_t$ is expressed by

$$\dot{Q}_t = h_\mathrm{glob}\cdot A\cdot\left(T_\mathrm{G}-T_\mathrm{M}\right) \tag{125}$$

where A is the effective gas – melt interface area and h_glob the global heat transfer coefficient between gas and melt.

The product $h_\mathrm{glob}\cdot A$ follows from the mechanism of heat transfer via the slag droplets ejected into the gas space. The droplets have a certain size distribution and must, therefore, be divided into a number of classes i. Each droplet in the same size class is heated up in the same way. The heat flow to a single droplet in class i is given as

$$\dot{Q}_{\mathrm{D},i} = \pi\, d^2_{\mathrm{D},i}\cdot h_i\left(T_\mathrm{G}-T_\mathrm{D}\right) \tag{126}$$

where $d_{\mathrm{D},i}$ is the droplet diameter, h_i the heat transfer coefficient in class i involving convective and radiative components, T_G the gas temperature, and T_D the droplet temperature. The heat transfer coefficient depends on temperature. As an approximation, this dependence is neglected in the following. Details for determination of the heat transfer coefficient at a single sphere in a surrounding fluid can be found in the literature [50, 70].

The heat flow $\dot{Q}_{\mathrm{D},i}$ is identical with the heat taken up by the droplet. It is given as

$$\dot{Q}_{\mathrm{D},i} = \frac{1}{6}\pi\, d_i^3\cdot C_\mathrm{P}\,\varrho_\mathrm{D}\,\frac{\mathrm{d}T_{\mathrm{D},i}}{\mathrm{d}t} \tag{127}$$

where C_P is the heat capacity, and ϱ the density. Equating (126) and (127) yields after integration with $T_\mathrm{D} = T_\mathrm{M}$ at $t = 0$

$$\frac{T_\mathrm{D}-T_\mathrm{M}}{T_\mathrm{G}-T_\mathrm{M}} = 1-\exp\left(-\frac{6h}{C_\mathrm{P}\,\varrho d_{\mathrm{D},i}}t\right) \tag{128}$$

The total heat transfer from gas to melt via the droplets is given by

$$\dot{Q}_t = C_\mathrm{P}\sum\nolimits_i \dot{M}_{\mathrm{D},i}\sigma_i\left(T_{\mathrm{D},i,\mathrm{final}}-T_\mathrm{M}\right) \tag{129}$$

$T_{\mathrm{D},i,\mathrm{final}}$ is the temperature which the droplets have in the moment when they fall back into the melt, i.e., at their residence time t_r. Its value is obtained from Equation (128) with $t = t_\mathrm{r}$. This yields

$$\dot{Q}_t = C_\mathrm{P}\sum\nolimits_i \dot{M}_{\mathrm{D},i}\sigma_i\eta_i\left(T_\mathrm{G}-T_\mathrm{M}\right) \tag{130}$$

where σ_i is the fraction of droplets in class i, and η_i the efficiency of heat transfer via droplets of class i:

$$\eta_i = 1-\exp\left(-\frac{6h_i}{C_\mathrm{P}\,\varrho d_{\mathrm{D},i}}t_{\mathrm{r},i}\right) \tag{131}$$

Equation (130) in connection with (131) allows to determine the heat flow $\dot{Q}_t$ which is identical with $\dot{Q}_t$ in Equation (124). Thus, the post-combustion reaction with heat transfer can be calculated, if the droplet size distribution is known. This distribution function obeys the Rosin – Rammler – Sperling function R_V, which gives the mass percentage on a sieve with a mesh diameter d:

$$R_\mathrm{v} = 100\exp\left[-\left(\frac{d}{d'}\right)^n\right] \tag{132}$$

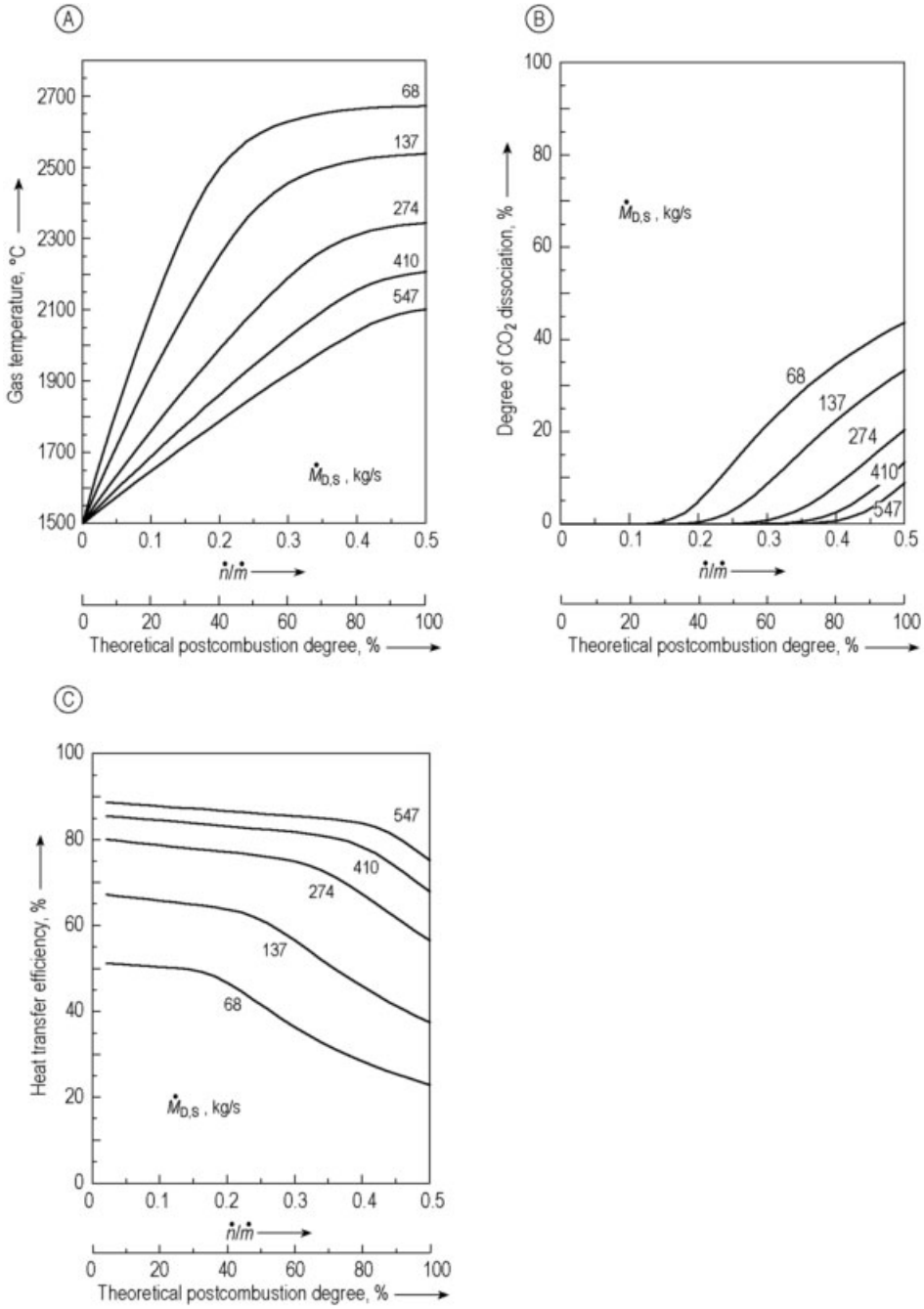

Figure 40. A) Gas temperature as a function of the theoretical post-combustion degree with the drop formation rate as the parameter;
B) Degree of CO_2 dissociation as a function of the theoretical post-combustion degree with the drop formation rate as the parameter;
C) Heat transfer efficiency as a function of the theoretical post-combustion degree with the drop formation rate as the parameter

The parameters d' and n were $d' = 3$ mm, $n = 1.3$ [70].

Figures 40 A – C show for various total slag drop formation rates $\dot{M}_{\mathrm{D,S}}$ the gas temperature, the degree of CO_2 dissociation, and the global heat transfer efficiency $\dot{Q}_t / \dot{Q}_{\mathrm{ideal}}$ (with $\dot{Q}_{\mathrm{ideal}}$ the heat transfer rate if $T_{\mathrm{G}} = T_{\mathrm{M}}$), as functions of the theoretical post-combustion degree PC_{theor}. It can be seen that the temperature increases with rising post-combustion degree. At low drop formation rates, where $\dot{M}_{\mathrm{D,S}}$ is almost zero, temperatures of up to 2800 °C are found. At high post-combustion degrees the temperature curves are getting increasingly horizontal, because the CO_2 dissociation degree becomes remarkable. The degree of dissociation reaches up to 50 %, and combustion then almost ceases. The temperatures remain lower at high than at low drop formation rates, because heat transfer is stronger in the former case. Simultaneously, CO_2 dissociation decreases. Below 1700 °C, dissociation is nearly zero at all post-combustion degrees. Heat transfer efficiencies for drop formation rates of 400 – 600 kg/s are ca. 80 %. At high post-combustion degrees where temperatures of 2000 °C are attained, the heat transfer efficiency decreases due to CO_2 dissociation.

2.4. Classification of Reduction Processes

For the pyrometallurgical reduction in the solid state iron ore must be heated to 450 – 1100 °C, for the reduction from an oxide melt temperatures of about 1400 °C are required. The reduction is a heterogeneous reaction, solid oxide is reduced by a gas, liquid oxide by a gas or a solid.

The choice of a particular reduction process usually also determines the subsequent steelmaking process. The iron produced by the reduction can have almost any combination of properties as shown in lists I and II:

I
1) solid
   - low in gangue content
   - high in gangue content
2) liquid

II
1) low in carbon content
2) high in carbon content

Many proposals have been advanced to achieve the reduction of iron on a technological scale.

Table 16. Classification of the technologically important reduction processes

Shaft furnace processes	
Blast furnace process	hot metal
Midrex process	solid sponge iron
HyL-III process	solid sponge iron
Rotary kiln processes	
SL/RN process	solid sponge iron
Krupp Codir process	solid sponge iron
ACCAR process	solid sponge iron
Fluidized-bed processes	
Tisco process	briquetted iron powder
Fior process	briquetted iron powder
Iron carbide process	briquetted iron powder
Circored process	briquetted iron powder
Discontinuous processes	
HyL-I process	solid sponge iron
Smelting reduction processes	
HIsmelt process	hot metal
DIOS process	hot metal
AISE initiative	hot metal
Electro-reduction processes	
Jupiter process	hot metal
Plasma-smelt process	hot metal

The *continuous processes* can be grouped according to the direction of the relative movement of the heat and the reducing gas flows versus the flow of the ore. The most important group of processes in which both heat and reducing gas flow countercurrently to the ore are the shaft furnaces and some fluidized-bed processes. Processes with countercurrent heat flow and crossflow of reducing gases use rotary kilns. Here, the reducing gas is liberated from the solid carbon contained in the burden and flows perpendicular to the direction of flow of the ore through the burden. Crossflow of heat and reducing gas is characteristic for most fluidized beds. Smelting reduction processes are mostly carried out in bath reactors. With few exceptions these processes usually constitute the final reduction stage of a two-stage process, the first stage being prereduction in the solid state. In smelting reduction the process gases are frequently post-combusted. The so-called *electro-reduction processes* occupy a special place in the family of smelting reduction processes because with them, reduction is possible with energy utilization near

100 %. Reduction processes presently used or under development and expected to be in use in the future on a large technological scale are summarized in Table 16.

2.4.1. Limiting Conditions for the Reduction in Shaft Furnaces

Gas Throughput. A prerequisite for the smooth operation of a shaft furnace is a good gas permeability of the burden. This requires that the various burden constituents have an average grain diameter that is as large as possible. There is, however, an upper limit to the grain diameter because the rate of heating and reduction decreases with increasing grain diameter. The choice of an optimum particle size must be based on a compromise.

The pressure drop of the gas characterizes the gas permeability of the burden in a shaft furnace. For turbulent flow through a burden consisting of equal grains the pressure drop ΔP is given by

$$\Delta \mathrm{P} = \frac{\mathrm{C_D}\,\varrho\,(1-\varepsilon)\,\mathrm{u}^2\mathrm{L}}{\varepsilon^3 \Phi \mathrm{d}} \tag{133}$$

where C_D is the drag coefficient which is a function of the Reynolds number; ϱ the density of the gas, ε the void fraction in the burden of randomly distributed particles, u the superficial gas velocity; L is the height of the burden; Φ the shape factor of the ore grains (0.6 to 1 for spheres) and d the grain diameter.

For a burden consisting of spheres of equal size, the random void fraction ε cannot be < 0.42 [50]. The effect of the void fraction on the pressure drop is, therefore, small. This is valid for all particle sizes as the void fraction of uniformly sized grains is always independent of the size itself. For a burden consisting of grains of various sizes the void fraction is in any case < 0.42, which means that the pressure drop increases. For small values of ε the effect of the cubic term in Equation (133) is large but the term $(1 - \varepsilon)$ has little influence. Moreover, the drag coefficient is not only a function of the Reynolds number but also of the grain size distribution. It is particularly sensitive to the presence of fines.

The relationship between drag coefficient C_D, Reynolds number Re and void percentage ε is given by the expanded version of the Karman – Kozeny equation [74]

$$C_\mathrm{D} = \left(\frac{160}{Re} + 3.1\,Re^{0.1}\right)\left(\frac{\varepsilon_\mathrm{k}}{\varepsilon_\mathrm{m}}\right)^{0.75} \tag{134}$$

where ε_k is the void fraction of the grain size fraction with the smallest grain diameter, ε_m is the void fraction of the mixture and

$$Re = \frac{\varrho u d \Phi}{\eta\,(1-\varepsilon)} \tag{135}$$

where η is the kinematic viscosity of the gas. The *pressure drop* in a burden consisting of spheres of various sizes can be calculated from Equation (133) by making use of Equation (134) for C_D and by substituting an equivalent grain diameter d_e for the grain diameter d, where d_e is given by

$$d_\mathrm{e} = \frac{1}{\sum\limits_{i=1}^{n} \frac{c_i}{d_i}} \tag{136}$$

d_i being the grain diameter of the grain size fraction i and c_i the volume percentage of fraction i.

The effects of the grain size and the relative quantity of fine grains under conditions approaching those in the blast furnace are shown in Figure 41. Depending on the ratio between the diameters of large and small grains, a fine grain fraction of even 10 % causes a pressure drop that is 4 to 7 times that of a burden consisting of equal sized grains.

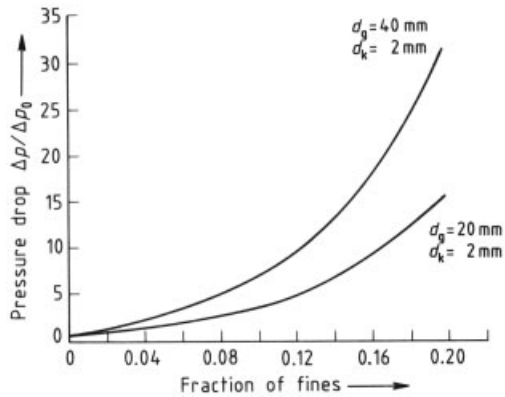

Figure 41. Relative pressure drop in burdens consisting of various grain sizes as a function of the fraction of fines [74] Δp = pressure drop in the presence of fines; Δp_0 = pressure drop without fines (diameter of shaft furnace = 7.0 m; quantity of cold blast = 90 000 m^3(STP)/h; $\varepsilon_\mathrm{k} = \varepsilon_\mathrm{g} = 0.38$ Properties of the air: temperature θ = 1000 °C, density ϱ = 0.265×10^{-3} g/cm^3, kinematic viscosity ν = 1.792 cm^2/s

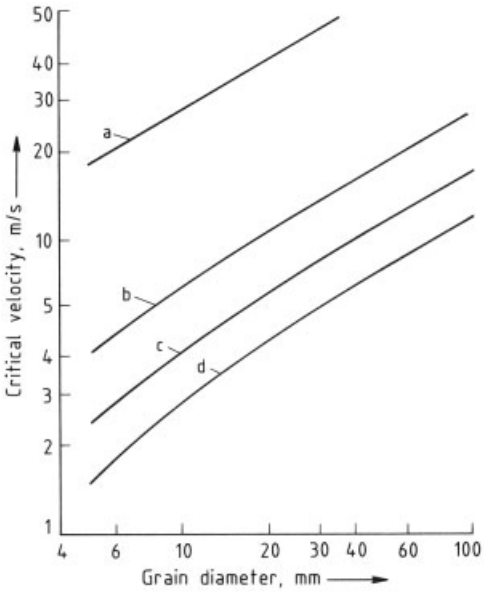

Figure 42. Stability limits of burden layers at 800 °C; burden consists of spheres of equal size; $\varepsilon = 0.4$ [19]
Bulk density of the burden grains: a) $\varrho_S = 5.0$ g/cm^3; b) $\varrho_S = 5.0$ g/cm^3; c) $\varrho_S = 2.0$ g/cm^3; d) $\varrho_S = 1.0$ g/cm; Flowing gas: a) Hydrogen; b, c, d) Carbon monoxide – nitrogen mixtures

Under practical conditions the fines often preferentially collect in regions where the resistance is high and, hence, the gas flow rate is low. This causes a further increase in the resistance. Finally, this condition may lead to the so-called "channeling", a situation in which the gas flows at a high rate through individual channels and is hardly being utilized, while it no longer flows through the rest of the burden. Thus, the burden destined for shaft reduction furnaces is carefully prepared by screening, comminution of large grains and agglomeration of fines to give a narrow particle size distribution. Abrasion strength, crushing strength, and swelling behavior have to meet stringent requirements.

A burden consisting of grains of a given diameter possesses a maximum limit of the gas flow rate through it (stability limit of the stationary burden). Above that limit the burden changes into a fluidized bed. The limit is determined by the requirement, that the pressure drop along the burden height remains lower than the weight of the burden. The pressure drop may be calculated from Equation (133) by using the value for C_D from Equation (134) for a burden consisting of equal grains and *Re* from Equation (135). Then, the thus calculated pressure drop per unit of burden height $\Delta p/L$ is set equal to the weight of the burden per unit of volume $(1-\varepsilon)\varrho_S\, g$; where ϱ is the bulk density of the burden, and g the gravitational acceleration. The obtained equation is resolved with respect to the gas velocity u, which is the desired critical velocity u_{crit} [214].

$$u_{crit} = \frac{0.54\left(\varrho_S g \varepsilon^3\right)^{0.526}}{[(1-\varepsilon)\,\eta]^{0.0526}} \frac{(d\Phi)^{0.579}}{\varrho_G^{0.473}} \tag{137}$$

Here, ϱ_G is the gas density.

The calculated critical flow rate for burdens consisting of grains of equal size is shown in Figure 42 as a function of the grain diameter d, the type of gas, and the density of the solid. Dense iron ore has a density of ca. 5 g/cm^3, coke ca. 0.9 g/cm^3. The maximum superficial velocity in the blast furnace is ca. 3 m/s and the Reynolds numbers are between 1000 and 3000 (see Fig. 43).

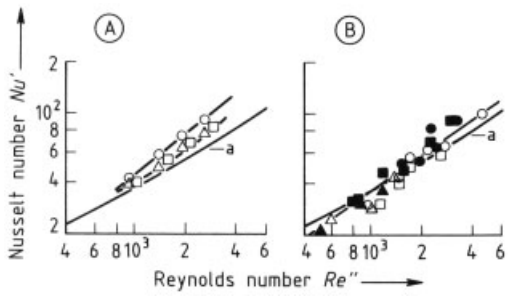

Figure 43. Nusselt number Nu' as a function of Reynolds number Re'' [75] A) For sinter (o), ore (□) and coke (△)
A) Relation valid for spheres [74]
B) For the following pellet burdens

Pellet	Diameter, cm	Porosity ε
△	1.1	0.528
▲		0.458
□	2.5	0.464
■		0.381
○	3.0	0.498
⬢		0.435

To increase the gas throughput of a shaft furnace, despite the fact that the stability limit has been reached, the gas pressure should be increased. The mass gas flow rate does not increase in proportion to the increase in pressure because the critical flow rate is lower at higher pressures.

When the pressure is increased from 0.1 MPa to 0.4 MPa, the process remains operating at the stability limit, and the mass throughput of the gas is only doubled.

Heat Transfer. The experimentally determined heat transfer between gas and burden is shown in Figure 43 A for sinter, ore, and coke and in Figure 43 B for pellets [75]. The heat transfer is depicted in the form of a relation between the expanded Nusselt number

$$N\,u\prime = \frac{\alpha_{\Sigma} d}{\lambda} \frac{\varepsilon}{(1-\varepsilon)}$$

and the expanded Reynolds number

$$Re'' = Re \frac{1}{1-\varepsilon}$$

Here, α_{Σ} is the overall heat transfer coefficient which reflects transition of heat to the total volume of a solid grain contained in the burden, i.e., it includes heat conduction in the solid. α_{Σ} increases with decreasing grain diameter and porosity and with increasing Reynolds number [75]. Curve a in Figure 43 describes heat transfer to solid spheres [74]. It can be seen from Figure 43 that the heat transfer between gas and sinter is the best and better than to solid spheres, while the heat transfer to ore, coke, and pellets roughly agrees with that to solid spheres and does not show any difference among the different types of burden. The Nusselt number, i.e., the heat transfer coefficient increases with increasing Reynolds number, i.e., increasing superficial gas velocity and decreasing grain diameter.

The high heat transfer coefficient for sinter in Figure 43 A result from the lower density of sinter in comparison with the other materials (see Fig. 44). Lower density means that less heat is taken up per mass unit. This yields an apparently higher heat conductivity in the solid and hence a higher α_{Σ}.

To minimize the time for *heating the burden* it must be kept in mind that the heat flux

$$Q = \alpha_{\Sigma} F\,\Delta \mathrm{T}$$

depends not only on the heat transfer coefficient α_{Σ} but also on the area per volume F available for the transfer of heat. This area increases with decreasing grain radius according to $F \sim 1/r$. The temperature difference ΔT at the gas – solid interface increases with increasing thermal conductivity of the solid because the accumulation of heat on the surface of the solid decreases with higher heat conductivity. The heat conductivities of various burden components are shown in Figure 45. The heat conductivity of coke is relatively poor and in the blast furnace its grains are relatively large. Hence, in the blast furnace coke is heated at a slower rate than ore. In general, this is desirable in order to achieve the suppression of the coke gasification according to the Boudouard reaction.

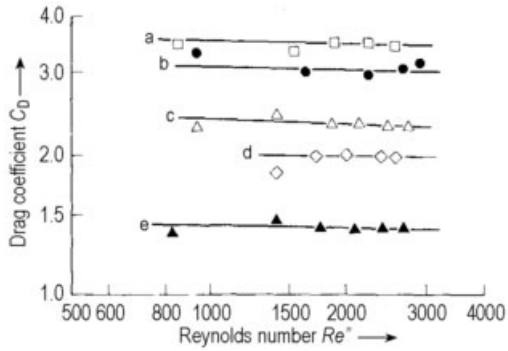

Figure 44. Relation between drag coefficient C_{D} and the Reynolds number Re'' for burdens consisting of various materials;
Average grain diameter d_{m}, cm: a) Smooth pellets d_{m} = 1.98; b) Sinter d_{m} = 1.76; c) Sinter d_{m}=1.72; d) Ore d_{m} = 2.18; e) Coke d_{m} = 1.79

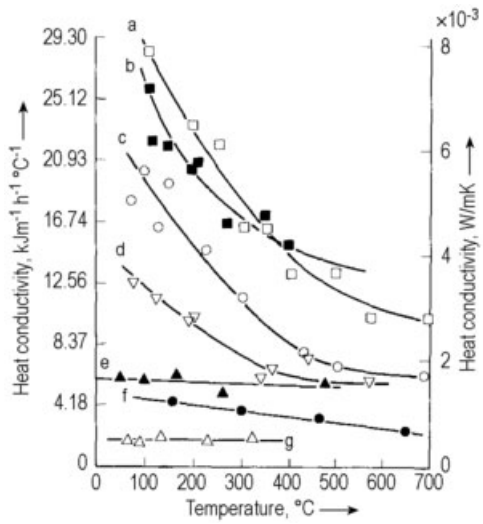

Figure 45. Heat conductivity of materials charged to the blast furnace [76]
a) Kiruna ore pellets; b) Wabana ore; c) Kiruna ore; d) Itabira ore pellets (hardened at 1200 °C); e) Coke (ϱ_{s} = 1.0; ash content 2.4 %); f) Goa ore; g) Coke (ϱ_{s} = 0.8, ash content 12 %)

From the temperature and Reynolds number distribution over the dimensionless burden height x/H (x = distance from the tuyere plane, H = burden height measured from the tuyere plane) (Fig. 46), it may be seen that a difference in temperature distribution in the center of the burden only slightly affects the Reynolds number and that the Reynolds numbers in the upper part of the shaft increase because the viscosity of the gas decreases with decreasing temperature. The high Reynolds numbers cause a rapid heating in the preheat zone of the furnace (see Figs. 43, 44, 45).

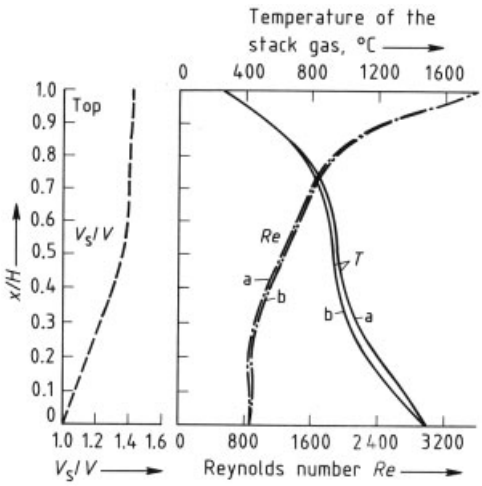

Figure 46. Profiles of the quantity of stack gas (left), its temperature and Reynolds number (right) in the column of a blast furnace burden [77]
x = distance to tuyere plane; H = burden height from the tuyere plane; V = quantity of blast = 90 000 m^3(STP)/h; V_S = quantity of stack gas; $Re = Re_0\ \eta_0/\eta$ with $Re_0 = u_0\ d/\nu_S$; η_0 = gas viscosity at 25°C; η = gas viscosity at shaft gas temperature; d = 21 mm; $\nu_0 = 1.33\times10^{-5}$ m^2/s; $u_0 = u_0{}^*/\varepsilon$; $\varepsilon = 0.33$; $u_0{}^* = V_S/A$; A = area of stack

Mass Transfer. The mass transfer between the gaseous phases and the ore has been discussed extensively in Section 2.3.1.2. Here, attention is drawn once more to the desirability of rapidly increasing the temperature to minimize the time needed for the exchange of matter. In addition, the burden should consist of grains with a radius so small that the reduction rate becomes independent of the grain radius.

2.4.2. Limiting Conditions for the Reduction in Rotary Kilns

The ore – coal mixture in a rotary kiln is usually heated countercurrently by hot combustion gases. The fuel consists primarily of carbon monoxide that is evolved from the burden, partly of hydrocarbons axially injected from the exit end. Carbon monoxide is combusted with air that is injected via tuyeres mounted in the drum. In the first section of the furnace (about one-third of the total volume in small furnaces, more than half of the total volume in large furnaces) the burden is heated to ca. 800 °C. Here, no oxygen is removed from the ore, except whenever high volatile coals are used. At higher temperatures carbon dioxide from the heating gas reacts with carbon to produce carbon monoxide according to the Boudouard reaction. Carbon monoxide reacts with the oxygen in the ore to regenerate carbon dioxide which, in turn, reacts with carbon to produce carbon monoxide, etc. During the Boudouard reaction the number of moles in the gas doubles, resulting in an outward gas flow from the coal – ore burden inside the kiln. The gas leaving the burden is a mixture of carbon monoxide and carbon dioxide with a $CO:CO_2$ ratio of roughly 80 : 20 and a certain amount of hydrogen and water vapor. The exact gas composition depends on the relative reactivities of the coal and the ore, whereby the carbon monoxide content grows with increasing coal reactivity. Figure 47 shows as an example the development of the gas and burden temperature and the degree of reduction for some ores in a 14 m rotary kiln [19]. It can be seen that the preheating zone where the burden has still a temperature below 800 °C, extends over ca. 40 % and the reduction zone over ca. 60 % of the kiln length. The final burden temperature is ca. 1100 °C.

The overall reduction reaction can approximately be written as

$$n\,C + FeO_n \rightarrow n\,CO + Fe \quad \Delta H = +164.1\ kJ/mol\ Fe \quad (138)$$

although the reduction proceeds via the gaseous phase.

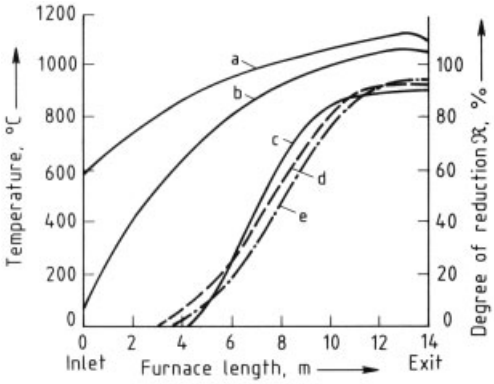

Figure 47. Degree of reduction and temperature profiles in rotary kilns [19]
Temperature: (a) in the furnace, (b) in the burden; Degree of reduction:(c) of Venezuela ore, (d) of Itabira ore, (e) of Bomi-Hill ore

The overall reaction, Equation (138), proceeds inside the burden and is strongly endothermic. The developed carbon monoxide is combusted over the burden where the required heat is liberated. (This is the oldest application of post-combustion in metallurgical processes). The heat is transferred from the gas to the burden directly and through the hot wall which serves as heat reservoir. The heat is always regenerated and the transfer occurs to a considerable degree by back-radiation from the hot wall when the wall rotates under the burden. The degree of reoxidation of the reduced iron by the oxidizing flame gases is only small because a layer of burning carbon monoxide at the burden surface shields the iron from direct contact with the gas. Because heat transfer in this zone is rather intense, heat transfer is not the rate determining step here. Rather, the rate of the reduction reaction limits the productivity in the reduction zone.

The reduction rate v_1 and the coal gasification rate v_2 may be approximated by [79]:

$$v_1 = -\frac{\mathrm{d}O}{\mathrm{d}t} k_1 O_\mathrm{V} \left(c - *c_1\right) \tag{139}$$

$$v_2 = -\frac{\mathrm{d}C}{\mathrm{d}t} = k_2 M_\mathrm{C} \left(*c_2 - c\right) \tag{140}$$

Here, O_v and M_c are the quantities of oxygen in the ore and carbon present at time t, both with respect to the ore volume, $*c_1$ is the carbon dioxide concentration in equilibrium with the iron oxide present (Fig. 14), $*c_2$ is the carbon dioxide concentration at the prevailing temperature and in equilibrium with carbon according to the Boudouard reaction (Fig. 14), c is the actual carbon dioxide concentration at a given location inside the burden. The reaction rate constants k_1 and k_2 correspond to the volumetric rate constants k_v and k_eff in Sections 2.3.1.1 and 2.3.1.2.

In the steady state of the process, $v_2 = v_1$. Then, from Equations (139) and (140) an expression for the overall reaction rate can be derived:

$$v = v_1 = v_2 = M_\mathrm{C} \left(*c_2 - *c_1\right) \frac{1}{\frac{M_\mathrm{C}}{O_\mathrm{V}} \cdot \frac{1}{k_1} + \frac{1}{k_2}} \tag{141}$$

The rate increases when the term $(*c_2 - *c_1)$ increases. According to Figure 14, this term is maximum at ca. 1000 °C. Furthermore, v_2 is affected by k_1 as well as by k_2. As mentioned above, the steady-state gas composition in the burden lies the more on the CO side, the higher the ratio k_2/k_1. In the Baur – Glaessner diagram depicted in Figure 48 measured gas compositions of the burden for the reduction with reducing agents of different reactivity k_2 are shown as a function of temperature. One agent was low temperature lignite coke, the other coke breeze. Curves a and b show the measured gas compositions. For steady state production conditions, v yields directly the production rate λ in kg Fe per cubic meter burden volume and time.

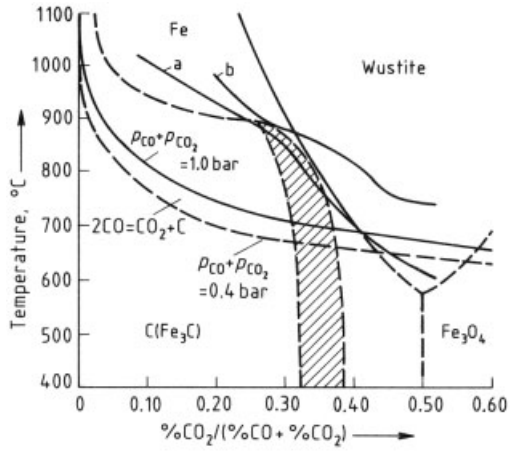

Figure 48. Gas composition inside the burden of a rotary kiln as a function of the temperature projected on a Baur – Glaessner diagram [19]
a) Reduction with low temperature lignite coke; b) Reduction with coke breeze

Whenever low temperature lignite coke with a value of k_2 that is in excess of that for coke breeze is used, the reduction at 1000 °C proceeds with higher CO content in the gas than in the case of coke breeze.

Factors Influencing the Kiln Productivity. The reduction reaction in the rotary kiln proceeds preferably above ca. 800 °C, depending on the char reactivity. Therefore, the burden initially passes a preheating zone in the kiln where the burden is heated from room temperature to 800 °C. In this zone heat is transferred to the burden from the combusted furnace gas coming from the reduction zone. Heat is transferred mainly by convection in the gas space and in the burden, and by regeneration via the rotating furnace wall, whereas radiation is weak below 800 °C, which is in contrast to the reduction zone. Due to the absence of noticeable radiation, heat transfer per unit of kiln length is less intense in the preheating than in the reduction zone and is, thus, rate determining in the former. The total kiln length is the sum of the preheating and the reduction zone length. Under steady-state conditions the consumption of heat per time for preheating the burden to the reduction temperature must be equivalent to the iron production in the reduction zone. The production P in t Fe/h can therefore be expressed either by the preheating or by the reduction process. With a few simplifications, this can be expressed as follows:

1) By the preheating process:

$$P = \frac{1}{Q}\bar{S}_{\mathrm{P}} l_{\mathrm{P}} \alpha_{\mathrm{P}} \left(\Delta\,\bar{\vartheta}\right) \tag{142}$$

2) By the reduction process:

$$P = \lambda l_{\mathrm{R}} \frac{\pi}{4} d^2 \bar{\vartheta}_{\mathrm{R}} \tag{143}$$

In Equation (142), Q is the heat demand in kJ per tonne Fe between 25 and 800 °C in the preheating zone, $\bar{S}_{\mathrm{P}}$ the average length of the secant which the burden surface forms with the inner kiln wall surface in the preheating zone in a cross section, l_{P} the length of the preheating zone. α_{P} the overall heat transfer coefficient in kJ m^{-2} h^{-1} K^{-1} between heating gas and burden in the preheating zone, $(\Delta\bar{\vartheta})_{\mathrm{P}}$ the average temperature difference between gas and bulk burden in the preheating zone. $\bar{S}_{\mathrm{P}} l_{\mathrm{P}}$ is the heat transfer area for preheating. The heat flux $\alpha_{\mathrm{P}}(\Delta\bar{\vartheta})_{\mathrm{P}}$ is estimated as being 50 kW/m^2 for a gas temperature of 500 °C and 230 kW/m^2 for 1000 °C [19]. In Equation (143), λ is the volume-specific reduction rate in t Fe m^{-3}h^{-1} in the reduction zone, l_{R} the length of the reduction zone, d the inner kiln diameter, and $\bar{\vartheta}_{\mathrm{R}}$ the average filling degree in the reduction zone. The secant $\bar{S}_{\mathrm{P}}$ in Equation (142) can be expressed by the product of the inner kiln diameter and a function of the filling degree $\bar{\varphi}_{\mathrm{P}}$: $\bar{S}_{\mathrm{P}} = d \cdot f(\bar{\varphi}_{\mathrm{P}})$.

Taking this into account and equating Equations (142) and (143) for the steady-state production, yields the length of the reduction zone as a function of the kiln length and the main characteristic process parameters used in Equations (142) and (143):

$$l_{\mathrm{R}} = \frac{l_{\mathrm{klin}}}{1 + \frac{\bar{\varphi}_{\mathrm{R}}}{f(\bar{\varphi}_{\mathrm{P}})} \frac{d}{C}} \tag{144}$$

with

$$C = \frac{\pi}{4} \frac{\bar{\alpha}_{\mathrm{P}} \left(\Delta\,\bar{\vartheta}\right)_{\mathrm{P}}}{Q \cdot \lambda} \tag{145}$$

C contains the characteristic process data. It has the dimension of a length and is, multiplied by $4/\pi$, the ratio of the volume of the reduction zone to the area of the preheating zone, which establishes, whenever the process runs in the steady-state. The factor C mainly depends on the burden properties, especially the reactivities of coal and ore, and on the conditions of post-combustion and heat transfer above the burden. Inserting Equation (144) in (143) with $\pi/4 \cdot d^2 \cdot l_{\mathrm{kiln}} = V_{\mathrm{kiln}}$ yields the volume-specific production rate

$$\left(\frac{P}{V}\right)_{\mathrm{klin}} = \lambda \cdot \frac{1}{\frac{1}{\bar{\varphi}_{\mathrm{R}}} + \frac{d}{C} \frac{1}{f(\bar{\varphi}_{\mathrm{P}})}} \tag{146}$$

as a function of the kiln length, of the factor C, and of the filling degrees in the preheating and in the reduction zone. The important consequence of Equation (146) is that the volume-specific production rate decreases with increasing kiln diameter d. The reason is that the heat demand is volume-specific, i.e., proportional to d^3, while heat transfer grows only with d^2. As a consequence, the fraction of the preheating zone length in the total kiln length increases with increasing d. For very large d only preheating is

rate determining, whereas for very small d only reduction is rate determining.

The following consideration illustrates the requirement of the heat transfer area [212]. Preheating of a typical charge for the production of sponge iron (1430 kg Fe_2O_3, 500 kg C, 100 kg gangue, 100 kg moisture, all per tonne Fe) to a temperature of 800 °C requires a heat quantity Q of $1650 \cdot 10^3$ kJ/t Fe. To transfer this heat to the charge, the latter must have a surface area $A_P = \bar{S}_P l_P$ (cf. Eq. 142) in the preheating zone of

$$A_P = \frac{P \cdot Q}{24q} \tag{147}$$

where q is the average heat flux (kJ m^{-2} h^{-1}). Division of A_P by the production rate P yields

$$a_P = \frac{A_P}{P} = \frac{Q}{24q} \tag{148}$$

where a_P is the specific heat transfer area per tonne of iron and day. Figure 49 shows the relation between the average heat flux and the necessary specific heat transfer area.

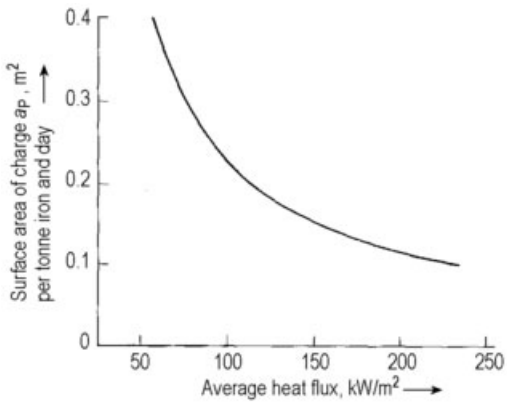

Figure 49. Required surface area of the burden in the preheating zone, a_P (m^2 per tonne Fe and day)) for the production of 1 t sponge iron per hour, as a function of the heat flux [212]. Burden per tonne of iron: 1430 kg Fe_2O_3, 500 kg C, 100 kg gangue, 100 kg water

The overall heat flow for a certain production rate in the rotary kiln is expressed by [79]

$$\dot{Q} = m\bar{c}_P \frac{dT_m}{dt} + \Delta H \cdot v \tag{149}$$

where m is the mass of the burden, $\bar{c}_P$ the average heat capacity of the burden, T_m the average burden temperature, ΔH the reaction enthalpy of the reduction reaction (143), v the reduction rate. As the reaction enthalpy is almost completely consumed by the coal gasification reaction, the value of ΔH for that reaction can be used and correspondingly the rate v_2 of that reaction.

2.4.3. Limiting Conditions for the Reduction in Fluidized Beds

The *reduction rate* of very small ore particles with diameters between 50 and 150 μm is significantly higher than that of the ore particles that are used in shaft furnaces or in rotary kilns. The reduction rate of very small particles in fluidized-bed reactors (→ Fluidized-Bed Reactors) is shown in Figure 50. The flow diagram of fluidized beds for iron powders fluidized with carbon monoxide or hydrogen is shown in Figure 51.

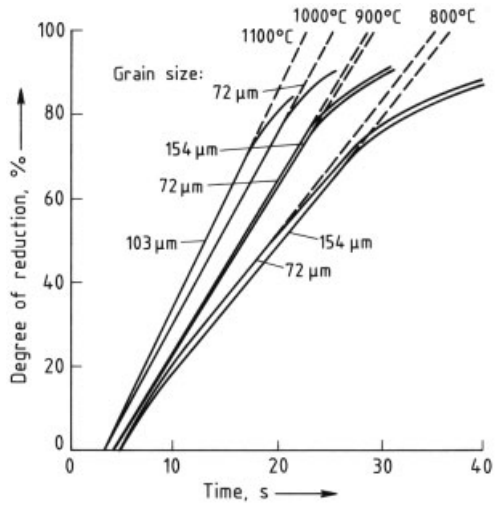

Figure 50. Degree of reduction as a function of time for various temperatures and particle diameters during reduction in hydrogen [19]

The *rate of heat transfer* from gas to solid is also high in a fluidized bed. This is the result of the large effective interfacial area between the phases but also of the high heat transfer coefficients in a fluidized bed [19].

Fluidized beds have a series of disadvantages, which prevented their broad utilization in the past:

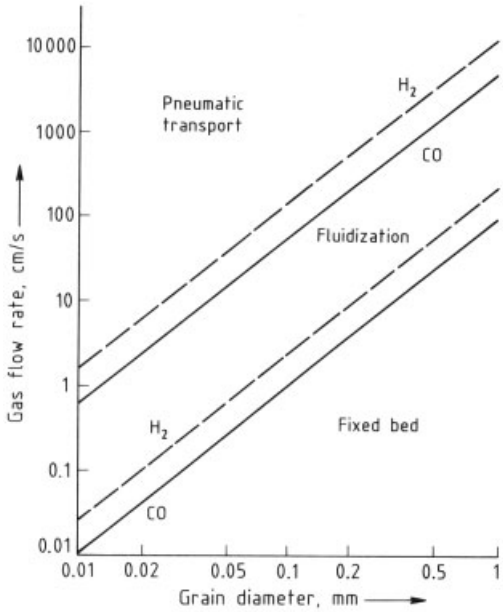

Figure 51. Flow diagram of a fluidized bed for iron ore powder ($\varrho = 5$ g/cm^3) at 700 °C [19]

1) Almost completely reduced material, which requires a reducing gas with a high percentage of a reducing agent for further reduction, is continually mixed with unreduced ore that does not require as high a percentage of reducing agent. This results in poor gas utilization and requires the recirculation of large quantities of gas. In technical installations fluidized-bed reactors are, therefore, put in series; this leads to a certain degree of countercurrent reaction.
2) As can be seen from Figure 51, each grain size requires a critical gas flow rate for a stable fluidized bed. In practice, the charge consists of a mixture of particles of different sizes. Therefore, at a certain gas flow rate, particles smaller than those corresponding to the critical flow rate are blown out, and larger particles fall down. The size distribution function must be small for a stable fluidized bed.
3) The tendency of freshly reduced iron particles to sintering, so-called *sticking,* is troublesome. As soon as the sintering process has started in any given location, it spreads to the entire fluidized bed in a few seconds. The result is a uniformly porous body that attaches itself to the perforated bottom plate of the reactor. The gas flows through a few channels inside the body and the reduction is greatly impeded.

The first two disadvantages, i.e., poor gas utilization and instability of the fluidized bed can be overcome by use of the circulating fluidized bed described in Section 2.6.3. The third is more or less successfully governed by the addition of fine inert material. An example is precipitation of carbon soot on the ore particles. As a result of these successful developments, fluidized beds gained increasing interest in recent years.

The reduction rate in the fluidized bed is determined by the rate of the interfacial reaction, particularly at low temperature. The rate of the interfacial reaction increases with increasing total pressure (see page 46). Increased pressure also results in an increase in the normally low heat capacity of the hydrogen and, hence, in a rapid heating of the ore by the reducing gas.

2.4.4. Limiting Conditions for Smelting Reduction Processes

Smelting reduction processes are usually operated in connection with prereduction of the iron ore in the solid state. Limits of energy consumption and productivity follow from a consideration of the entire two-stage process of pre- and smelting reduction.

Energy Consumption. The energy consumption of smelting reduction processes is composed of two parts. First, the oxygen separated from the reduced iron oxide is bonded to carbon monoxide, which corresponds to a certain amount of carbon as part of the fuel and hence, the energy consumption. This is the chemical part of the energy consumption. Second, energy is expended to satisfy the heat demand of the endothermic reduction reaction, which is the thermal part of the energy consumption. That part can be supplied (see Section 2.2.2) (1) by electrical energy, (2) by combustion of an additional amount of carbon to carbon monoxide, and (3) by post-combustion of CO and H_2 to CO_2 and H_2O.

The energy is almost completely consumed in the smelting reduction stage, whereas prereduction of solid ore with ($CO + H_2$) from the

Table 17. Process data for various procedures of smelting reduction [216]

Characteristic	Heat supply			Heat supply with fossil energy			
	with electrical energy	without gas washing and recirculation	without gas washing and recirculation	with gas washing and recirculation	with gas washing and recirculation	without gas washing and recirculation	without gas washing and recirculation
Fuel	carbon	carbon	coal *	coal *	coal *	coal *	coal *
Post-combustion degree	0	0	0	10	20	10	50
Gas amount in smelting reduction stage, m^3 (STP)/t Fe	320	842	1120	711	711	1122	733
Prereduction degree, %	47	75	90.5	91	81	81	30
Gas utilization, %	88	53	49	100	100	48	83
Fuel consumption, kg/t Fe	213	493	531	351	351	532	361
Total energy consumption, GJ/t Fe	10.4	16.2	17.6	11.6 **	11.6 **	17.6	12.0

* Composition: 82.8 % C, 6.2 % H, 8.0 % O, 2.0 % N 1.9 % S.
** Without energy consumption for gas washing.

smelting reduction stage is thermally nearly neutral. Energy consumption values calculated for off-gas temperatures of 1500 °C are shown in Table 17 for the three aformentioned ways of energy supply [216]. Prereduction is assumed to proceed countercurrently. For better comparison heat losses and heat consumption for slag formation and oxygen production are not considered. Taking these amounts in account, the total energy consumption can be estimated to be 20 to 40 % higher. It is further assumed that the conditions of Figure 22 are fulfilled. This yields the optimal operating point of the system.

Operation without Post-combustion. Row 2 in Table 17 shows the values when electrical energy is used for heat supply. Under this condition, carbon serves only for reduction. The degree of prereduction is 47 %, and the reduction gas is almost completely utilized. Operation is practically autothermic. The procedure is applied in the plasma-smelt process.

Rows 3 and 4 show results when additional carbon is combusted to carbon monoxide in the smelting reduction stage to supply the heat demand for reduction. As the optimal gas utilization is ca. 50 %, the fuel consumption is higher than if electrical energy is used. Coal lowers the gas utilization in comparison to pure carbon or coke due to its hydrogen content (row 4).

Operation with Post-combustion. A solution of the problem of limited gas utilization, if carbon or coal is the only heat supplier, is post-combustion of the process gas of smelting reduction (cf. Section 2.2.2). Rows 5 to 8 show results when the process gas is post-combusted. In rows 5 and 6 post-combustion of the process gas amounts to 10 and 20 %, respectively. CO_2 must be washed out from the gas in this case, in order to make it usable for prereduction. If CO_2 is not removed (row 7), the same situation in respect of gas utilization and total energy consumption exists as in processes without post-combustion (compare row 7 with row 4). If post-combustion is increased to 50 % and more (row 8), the smelting reduction process becomes autothermic for the reduction of wustite, and the thermodynamic limit of FeO – Fe gas reduction vanishes (cf. Section 2.2.2). Gas washing is no longer necessary. Gas utilization becomes 83 %, which is almost the same as in the case of reduction with electrical energy.

Productivity. It is known from the blast furnace and from converter processes that productivity is mainly determined by the maximum possible gas throughput. The iron-bath reactor for smelting reduction may be compared in this respect with a steel converter [80]. Application of this fact to smelting reduction vessels together with plant experience yields a specific gas throughput of an iron-bath reactor for smelting reduction operated under atmospheric pressure of 3 – 5 mol m^{-3} s^{-1} [71, 81]. The productivity of a process is generally defined as

$$P = \frac{\text{maximum volume} - \text{specific gas throughput}}{\text{amount of gas/kg fuel} \times \text{fuel consumption}} \qquad (150)$$

A coal consumption of 600 kg coal/t hm (hm = hot metal) = 570 kg coal per tonne Fe for a smelting reduction process [82] then yields with a volume-specific gas throughput of 3 mol m^{-3}

s^{-1} for a 100 t smelter [81] and 0.087 kmol gas per kilogram coal a productivity of 5.2 t Fe m^{-3} d^{-1} [83].

Influence of Pressure and Post-combustion on Productivity. Productivities are influenced by various factors. One of them is *pressure* (cf. Section 2.4.1). Application of pressure enables a higher volume-specific gas throughput and, thus, an increased productivity. The increase is approximately proportional to the square root of the pressure. Another factor is *decrease in fuel consumption*. It reduces the gas production per tonne Fe and, thus, also increases the productivity. For example, an almost autothermic procedure with post-combustion as that in row 8 of Table 17 has a gas throughput which is 0.65 times that of row 3 in the same table without post-combustion. That would increase the productivity according to Equation (150) from 5.2 to 5.2/0.65 = 8.0 t Fe m^{-3} d^{-1}. In this way post-combustion increases the productivity by lowering the coal consumption.

The total productivity P_{tot} of a two-stage pre- and smelting reduction process is composed of the productivities of smelting reduction P_{SR} and of prereduction P_{PR} as follows [83]

$$\frac{1}{P_{\mathrm{tot}}} = \frac{1}{P_{\mathrm{SR}}} + \frac{1}{P_{\mathrm{PR}}} \tag{151}$$

Data of P_{PR} are known for countercurrent shaft furnaces (see Table 18) [83].

Table 18. Productivities of various iron ore direct reduction processes carried out in shaft furnaces with countercurrent flow of gas and ore [83]

Process type	Top pressure, bar	Productivity, t Fe $m^{-3}d^{-1}$
Midrex [84]	1	8.4
Sumitomo [85]	2	9.6
Nippon Steel (pilot plant) [86]	5	13.2

In connection with the mentioned values for smelting reduction of 5.2 and 8.0 t Fe m^{-3} d^{-1} without and with post-combustion, respectively, the data given in Table 19 follow from Equation (151). The value 3.4 t m^{-3} d^{-1} of total productivity for 2 bar pressure and without post-combustion may be compared with the productivity of a blast furnace with increased pressure. It is higher than that of the blast furnace, which amounts to 2.4 t m^{-3} d^{-1} [87]. The difference is mainly due to the fact that in the prereduction stage the space required by the coke in the blast furnace vanishes.

Table 19. Total productivities of two-stage pre- and smelting reduction processes [83]

P_{SR}, t Fe m^{-3} d^{-1}	P_{PR}, t Fe m^{-3} d^{-1}	P, t Fe m^{-3} d^{-1}
5.2	8.4	3.2
5.2	9.6	3.4
5.2	13.2	3.7
8.0	8.4	4.1
8.0	9.6	4.4
8.0	13.2	5.0

Influence of Reactor Size on Productivity. The productivity of an iron-bath reactor depends on the reactor size. With increasing bath height the running length of the gas through the melt and, hence, also the resistance of gas throughput rises. The influence of bath height h on the maximum possible gas throughput $\dot{V}_{\max}$ can approximately be expressed by

$$\dot{V}_{\max} = c_1 \frac{d^2}{h^n} \tag{152}$$

[83], where d is the diameter of the cylindrically assumed vessel, n an unknown exponent and c_1 a proportionality constant. In order to link gas throughput with productivity, it is required to relate the volume flow to the reactor volume $V_{\mathrm{R}} \cong d^2\, h$. For the special case that d/h remains constant, i.e., $d/h = c_2$, it then follows

$$\frac{\dot{V}_{\max}}{V_{\mathrm{R}}} = c_1 \cdot h^{-(1+n)} = \mathrm{const} \cdot V_{\mathrm{R}}^{-\frac{1+n}{3}} \tag{153}$$

According to this equation, the maximum volume-specific gas throughput decreases with growing reactor volume V_{R}, if $n > -1$. The minimum value of n, which is physically possible, is zero. In that case, the resistance to gas throughput does not increase with the bath height. The productivity decreases only because the surface-to-volume ratio becomes smaller at scaling-up. If $n > 0$, the increasing resistance of the bath height to gas throughput acts as an additional factor that decreases the productivity.

2.5. Blast Furnace Process

A schematic drawing of a modern large blast furnace, here for example blast furnace Schwelgern no. 2 of Thyssen Krupp Stahl AG [88], is shown

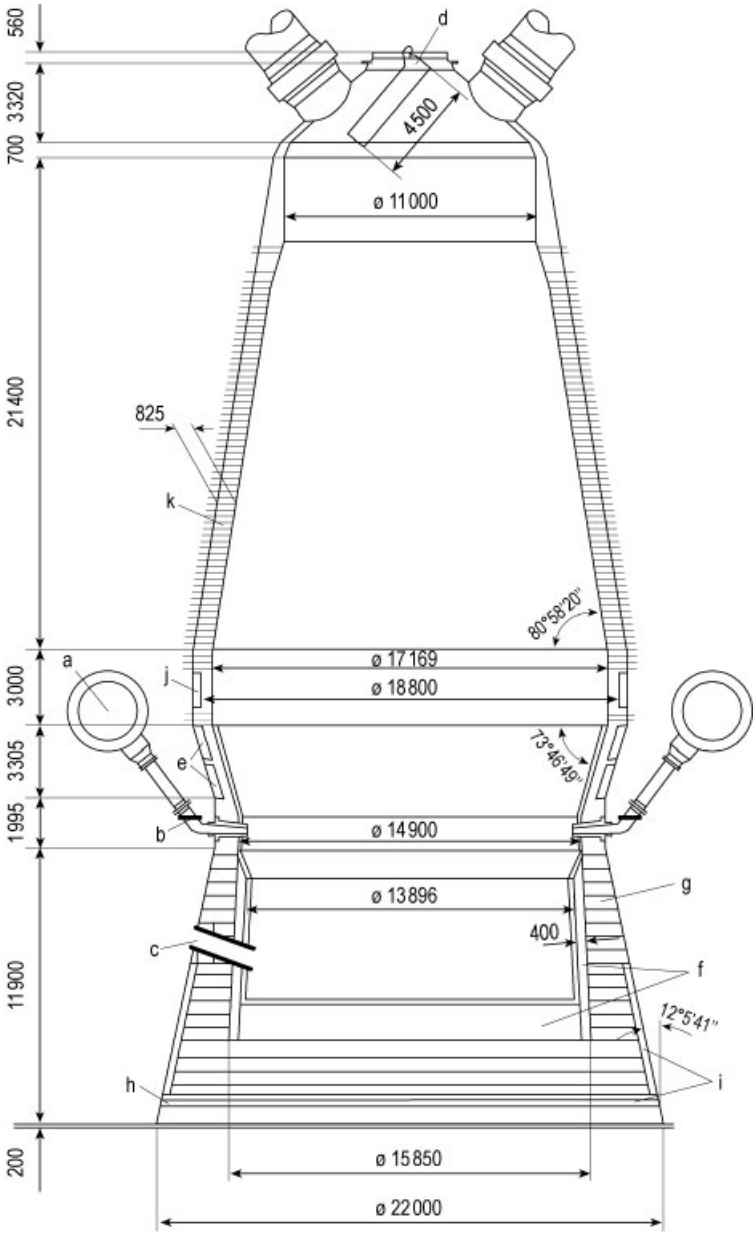

Figure 52. Profile of a typical blast furnace (Schwelgern no. 2, Thyssen Krupp Stahl AG, total volume 6020 m^3, all dimension in mm) (Original figure kindly supplied by Thyssen Krupp Stahl AG)
a) Bustle pipe; b) Tuyere; c) Tap hole; d) Paul Wurth top; e) Stave; f) Ceramic and chamotte lining; g) Carbon blocks; h) Bottom cooling; i) Graphite; j) Copper stave; k) Cooling boxes

in Figure 52. A blast furnace is principally divided into the following areas: upper and lower shaft, bosh, belly, and hearth (see Fig. 53). The definitions of the furnace volumes, which are important to calculate specific productivity data are as follows:

Working Volume (W.V.) = volume from 1 m below the lower flange of the adjustable throat armor in low position or from 1 m below the lower flange of the open lower bell to the tuyere level; with rotating chute, the distance in question is to be measured from 1 m below

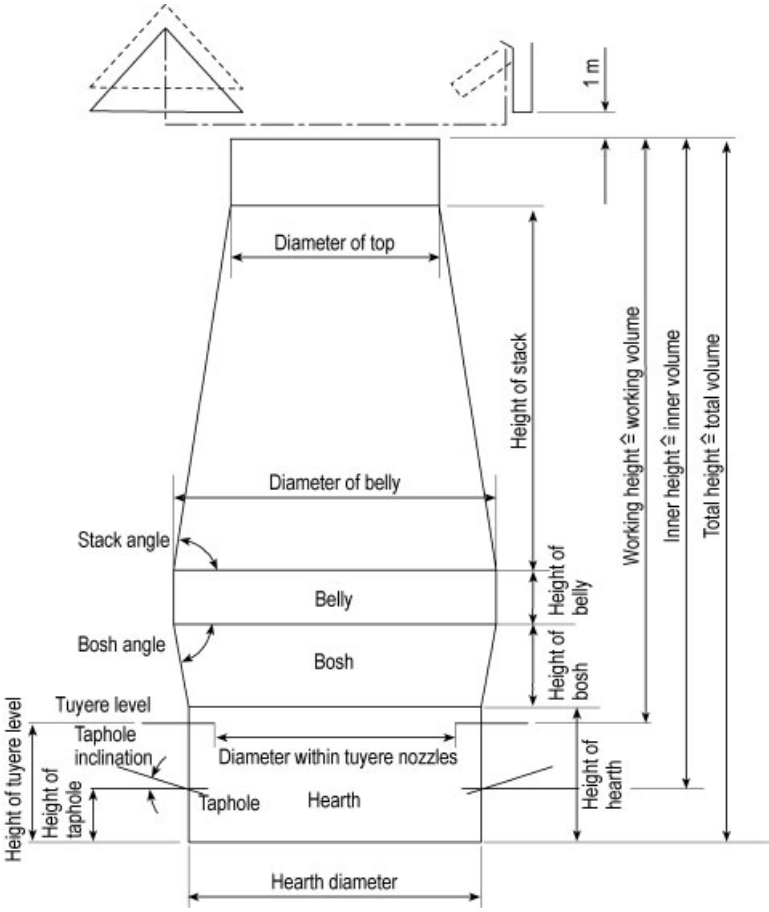

Figure 53. Blast furnace, sectional drawing

the lower flange of the vertically positioned chute to the tuyere level.

Inner Volume (I.V.) = volume of top area as defined for the working volume but to the taphole level (taphole center outside).

The blast furnace operates according to the countercurrent principle (see Fig. 54) [89]. The process principle makes the blast furnace dependent on lumpy materials to maintain a gas-permeable stock column. For this reason, fine grained iron ores are agglomerated, i.e., sintered or pelletized, before being charged into the blast furnace.

The solid materials, such as coke, ores, and fluxes are charged at the furnace top in a defined distribution mode. The burden column in the stack region therefore consists of a layered structure of ores and coke. In addition to its task to generate heat and reducing gas the coke has the important task to build up a supporting pillar for the solid burden column. The supporting pillar must remain permeable for the furnace gas when melting separates the formed iron and the other constituents of the ores, the gangue, from one another. This supporting pillar is called *cohesive zone*. Under the cohesive zone, only coke remains as solid material. It builds up a coke cone known as *dead man* (see Section 2.1.2). The oxygen content of the hot blast blown in through the tuyeres reacts with the carbon of the coke to produce a reducing gas, which ascends in the stack and extracts oxygen from the ores. The gasification of coke carbon with oxygen in front of the tuyeres leads to a high temperature of up to 2200 °C (*flame temperature*). The high temperature area in front of the tuyeres is called *raceway*. Auxiliary reducing agents, such as oil, coal, natural gas, coke oven gas, or plastics wastes, are injected via the tuyeres in order to save coke.

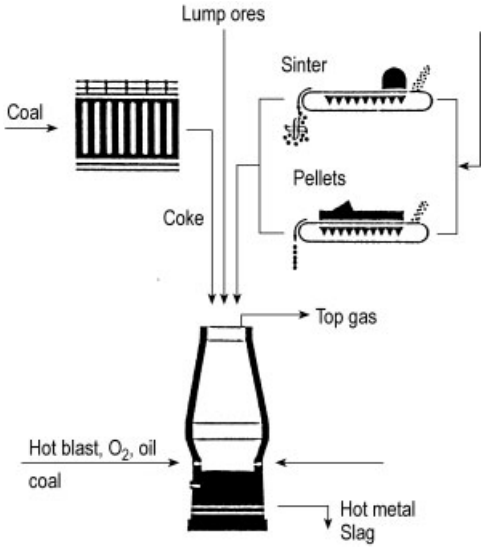

Figure 54. Operation mode of the blast furnace

Components of the slag formed from the gangue are also reduced and dissolved in the hot metal together with carbon coming from the coke. Slag and hot metal are tapped via tapholes in the hearth. Surplus gas is drawn off at the furnace top, cleaned, and used in the iron and steel works for heating purposes.

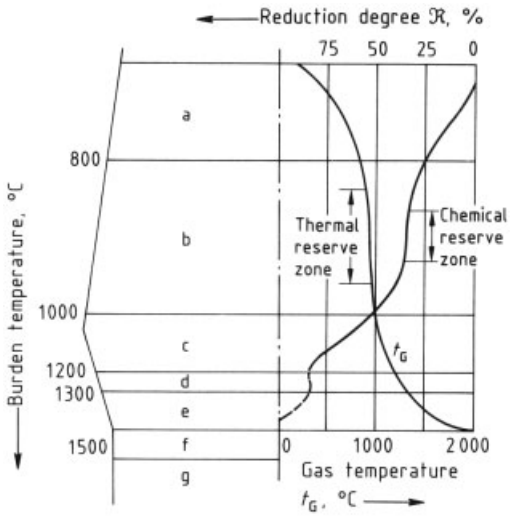

Figure 55. Reaction zones and profiles of temperatures and reduction degree in the blast furnace [97]
a) Indirect reduction of Fe_2O_3 and Fe_3O_4; b) Indirect reduction of Fe_xO; c) Direct reduction of wustite and Boudouard reaction; d) Melting of iron and slag; e) Reduction of metalloid and FeO in the slag; f) Combustion in the front of the tuyeres; g) Hearth

The bed structure inside the blast furnace was clarified by the study of the commercial size blast furnace [90 – 96]. When the high temperature gas ascends from the raceway through the coke bed in the lower part of the blast furnace (see Fig. 55), it is cooled rapidly to ca. 1500 °C by (1) heat exchange with the coke bed and with molten materials coming down countercurrently, and (2) by the endothermic reaction of the *direct reduction* of molten iron oxide according to the overall reaction

$$C + FeO \rightarrow CO + Fe$$

This overall reaction includes not only the actual reduction of FeO with solid carbon but also the sum of simultaneous reactions for the gaseous reduction of FeO and the Boudouard reaction. When the temperature of the ascending gas and of the burden decreases below 1000 °C, the Boudouard reaction does not proceed any further because of the low coke reactivity. In the *stack region,* the iron oxides are exclusively reduced by carbon monoxide thereby forming carbon dioxide. In addition, the higher iron oxides such as hematite and magnetite are reduced to wustite in the *upper part* of the stack as follows:

$$3\,Fe_2O_3 + CO \rightarrow 2\,Fe_3O_4 + CO_2$$

$$Fe_3O_4 + CO \rightarrow 3\,FeO + CO_2$$

These reactions are often called *indirect reduction* (together with the reduction reaction of $FeO + CO \rightleftharpoons Fe + CO_2$) because the coke present in the stack of a blast furnace does not play a role in the reduction mechanism. The reduction step from hematite to magnetite proceeds quickly and does not play a role in the overall rate law. However, attention has to be paid to this low temperature reaction with respect to reduction disintegration of iron ores, especially of the sinter, which is caused by the lattice rearrangement occurring during the reaction. In the upper part of the stack region, the burden is heated to 800 – 1000 °C. Below the region, considerably wide *thermal* and *chemical reserve zones* appear where the temperature of gas and solids is kept substantially unchanged at 800 – 1000 °C and no reduction of FeO occurs because thermal and chemical equilibrium conditions are almost established between solid phase (Fe – FeO) and gas phase ($CO - CO_2$).

The gas leaves the top of the furnace through four uptakes. The top gas has *chemical energy*

because approximately only half of its carbon monoxide content is utilized during the reduction processes in the upper part of the blast furnace. However, its sensible heat is rather low because of its low temperature of 100 – 230 °C, which depends on the operating conditions. After the top gas is cleaned, ca. 25 – 30 % is usually used to preheat the blast in the hot blast stoves operated on the basis of the regenerative principle, and 15 – 20 % is used to operate turboblowers to introduce the compressed blast to the blast furnace at the required blast pressure. The remainder (50 – 60 %) is used for heating in other sections of the plant. At blast furnaces operated with elevated top pressure, energy is recovered via a top gas recovery turbine (TRT), generating approximately one-third of the electricity needed for blast compression.

Above the tuyere zone heat and mass transfer (for heating up the burden and for the reduction and Boudouard reactions) occur between the ascending gas and the descending burden.

In the *lower part* the *dead man* (see Fig. 56), extending from the hearth bottom to the belly, has a unique and very important function. The dead man is formed by the charge materials' distribution at the top (for example, coke center charging) and by the gasification of the descending coke in the periphery at the tuyere zone. The oxygen injected with the hot blast is almost completely consumed within the raceway. As a result, the coke located in the center of the furnace is not gasified and forms the dead man. The coke forming the dead man is continually renewed.

The hot gas leaving the raceway ascends primarily into the permeable coke bed which constitutes the dead man, and gives its heat to both the coke bed and the molten material dripping down through the dead man. Then the gas ascends through the coke windows in the *cohesive zone.* Here all burden materials except coke soften and are finally melted down. Since the cohesive layers have a high gas-flow resistance, the gas ascends exclusively through the coke windows.

The molten iron and slag produced in that region trickle down through the coke bed to the hearth. During this process, the molten iron which is in contact with coke, absorbs carbon up to the thermodynamic limit.

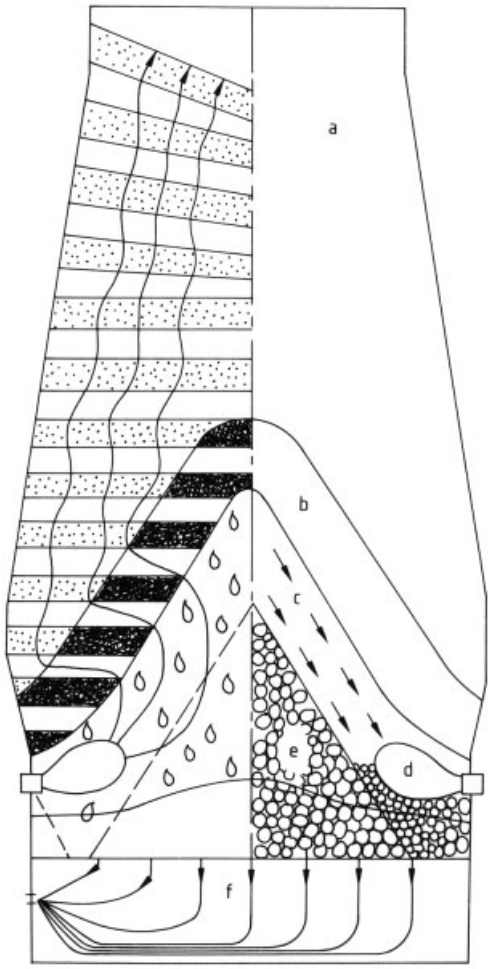

Figure 56. Schematic of blast furnace (internal situation) a) Stack; b) Cohesive zone; c) Dripping zone; d) Raceway; e) Dead man; f) Hearth

Hot metal and slag accumulate in the *hearth.* The slag floats on top of the metal bath because of its lower density. Therefore, the metal droplets or streams must come down through the slag layer.

The lower hearth is operated batchwise because hot metal and slag were tapped intermittently at intervals of 40 to 50 min. Hot metal and slag are tapped from the furnace through the same taphole. In most cases first the hot metal flows out and the slag appears later. To empty the hearth early tapping of the slag is a prerequisite that prevents rising of the slag level upward in the furnace which could lead to a deterioration in furnace operation.

Large blast furnaces are equipped with up to four tapholes. Here a continuous and overlapping tapping is state of the art. At the blast furnace Schwelgern no. 2 of Thyssen Krupp Stahl AG, for example, tapping technique was converted to continuous operation. At first a twin hole technology was applied which was superseded by the use of three holes [98]. Two tapholes are opened alternately as soon as one hole is empty. The third taphole has its own time characteristics and is opened in each case after a tap-to-tap time of 50 min. The slag is easily separated from the hot metal by a skimmer in a runner system.

The tapped *hot metal* is transported to the steel shop in *refractory-lined ladles* or in a few cases cast into small molds by using a pig-casting machine to produce cold pig iron of ca. 5.0 kg. The slag is either cast directly in a slag pit for solidification, cast in a slag ladle for transportation to the slag pit, or – as common practice today in many blast furnaces – granulated at the blast furnace directly in a granulation device using, for example, high pressure water.

2.5.1. Outline of a Blast Furnace Construction

A large blast furnace has a working volume of up to 5000 m^3, a furnace height of up to 36 m, a height of ca. 130 m including the exhaust pipes for the top gas, and a hearth diameter of ca. 15 m. The furnace must endure high mechanical and thermal loads because of the large weight of the structure and burden column and the high temperature of the hot blast (ca. 1300 °C) and of the burden in the lower part of the furnace (ca. 1350 °C). In addition the gas pressure inside the furnace may be as high as 6.5 bar.

Blast furnaces are today entirely encased in a steel shell (*mantle plate*). They are equipped with a refractory lining inside the mantle plate and with external cooling systems such as cooling boxes and/or stave coolers to cool the lining and the shell.

Blast furnaces are constructed according to two different principles illustrated in Figure 57, which represents the construction principle of the 1960s. In a *free-standing* type of construction (Fig. 57 A), the furnace is self-supporting. The weight of the top-charging gear, the top-gas uptakes and the top of the skip incline (or charging conveyer-belt gallery) is supported by a steel construction surrounding the furnace. In a *shaft-support* ring-type construction (most U.S. blast furnace constructions), the furnace shaft and the entire construction are supported by a mantel ring, i.e., a ring surrounding the furnace and attached at the belly.

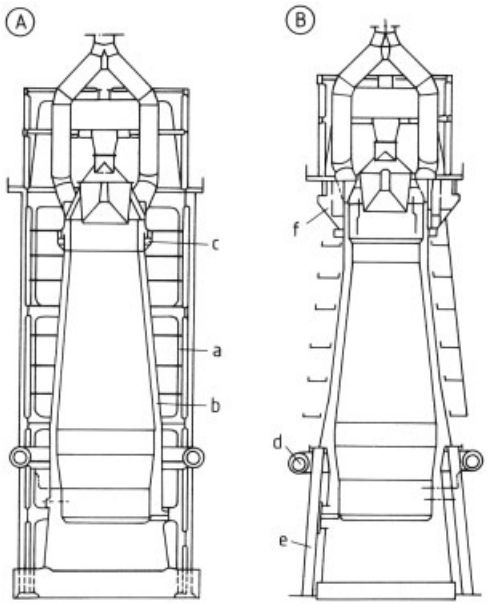

Figure 57. Blast furnace support structure [99]
A) Free standing construction; B) U.S. construction
a) Frame; b) Continuous armor; c) Compensation; d) Stack support ring; e) Ring support; f) Support for the top

Figure 58 shows a furnace with its top construction, the hot stoves, and the tap gas cleaning devices, taking blast furnace Schwelgern no. 1 of Thyssen Krupp Stahl as an example. For this furnace, the charge materials are conveyed to the top by a conveyer belt. For older blast furnaces the charge materials are transported to the top by skip cars along a skip incline. Figure 59 gives an example of a modern cast house layout as required for modern blast furnaces, showing full cast house dedusting equipment. The dust sources that need extraction are the taphole, the main trough, the slag skimmer, the hot metal and slag troughs as well as the hot metal and slag transfer points. If the trough system is fully extracted by means of hoods arranged above the troughs, volume flows of up to 1×10^6 m^3

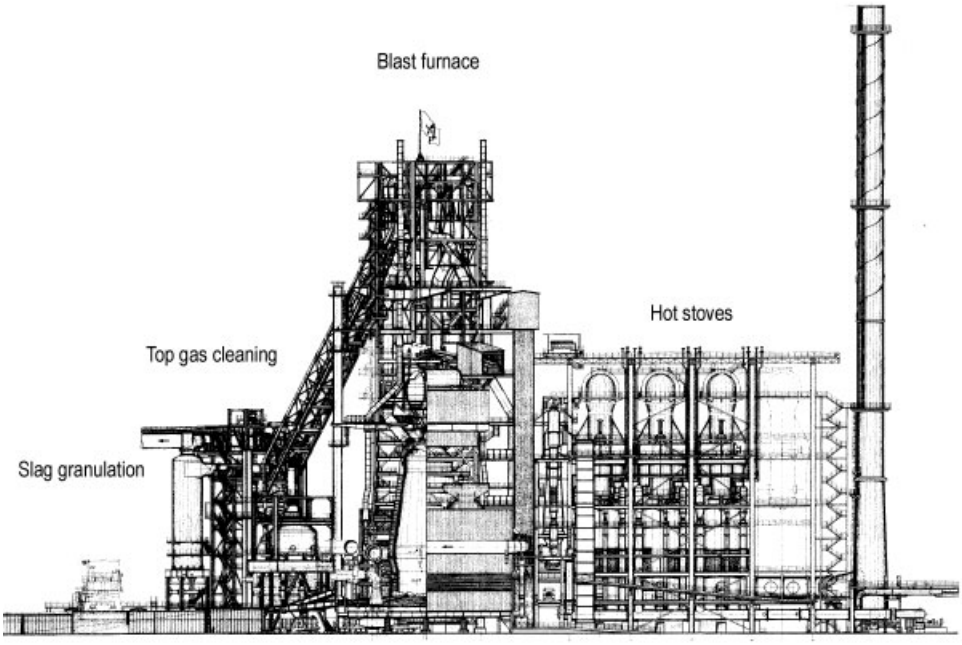

Figure 58. Blast furnace with top construction, hot stoves, and gas cleaning devices (blast furnace Schwelgern no. 1, Thyssen Krupp Stahl AG, original figure kindly supplied by Thyssen Krupp Stahl AG)

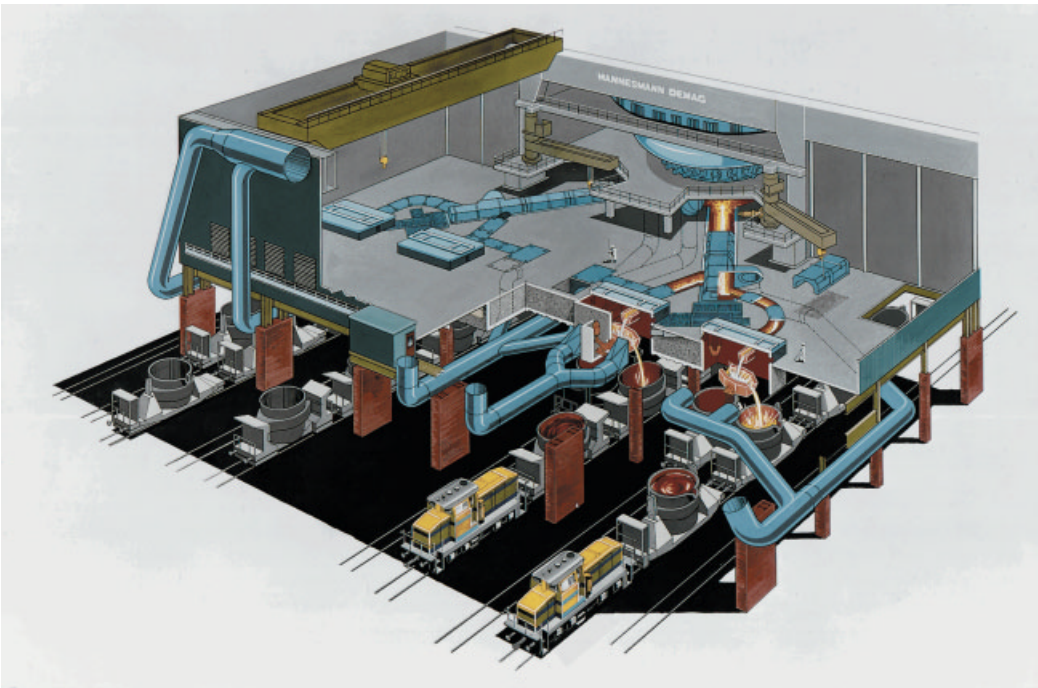

Figure 59. Cast house of blast furnace B, Hüttenwerke Krupp Mannesmann

(STP)/h or 4800 m^3 (STP) t hot metal are attained and cleaned in electrostatic precipitators or bag filters.

Large blast furnaces can produce up to ca. 13000 t hot metal per day. These furnaces require up to ca. 20000 t/d of ferrous burden materials and ca. 4550 t/d of coke, if operated with injection of auxiliary reductants and a coke rate of 350 kg/t hot metal. The sinter, which is the most effective ferrous burden material for producing hot metal in the blast furnace, is produced in the sinter plant. The sinter produced is screened and fed directly to the furnace by a belt conveyer, through weighing devices. The sinter may also be stored in intermediate hoppers to provide a buffer whenever production rate fluctuates. The fines created during transportation or storage are separated and recycled to the sinter plant. Coke and fluxes are also added from hoppers. The fines created during coke screening are also recycled to the sinter plant or charged as nut coke (10 – 30 mm) mixed with the sinter to the blast furnace periphery. The separation of the fines is very important to maintain the gas permeability of the burden column (see Section 2.4.1).

Charging. The charging equipment has two functions, particularly for large furnaces: the burden distribution and gas sealing. Appropriate *burden distribution* with respect to both ore : coke ratio and particle diameter over the cross section of the stack at the top of the burden bed is important for optimum gas permeability distribution. The latter provides a better reduction of iron ores in the stack region and higher efficiency of gas utilization. Modern charging equipment possesses a gas-tight sealing to prevent gas from escaping during the charging of the furnace. Gas sealing is nowadays possible even for high top-gas pressure of up to ca. 3.8 bar.

Particular attention should be paid to the materials and structure of the charging equipment because ore and coke are very abrasive. The failure of small and sometimes inaccessible parts leads to production disruptions, which in turn causes downstream installations such as the steel plant and the rolling mills to cease operation.

Furnaces with low Top-Gas Pressure. Older furnaces working with a low top-gas pressure are exclusively equipped with a *dual bell-and-hopper arrangement* for gas sealing. In the U.S. McKee type bell-and-hopper arrangement (Fig. 60 A) presently used, the upper bell with the feed hopper can be rotated, therefore the material fed to the hopper with a conveyer belt or a skip hoist can be distributed uniformly in circumferential direction. As soon as the hopper is filled, the small bell is lowered and the material is charged to the large bell. Then the small bell is closed. Subsequently, the large bell is opened. The material falls into the stack where the stock height is monitored with two stockline indicators.

Furnaces with High Top-Gas Pressure. A dual bell-and hopper arrangement is seldom used for modern blast furnaces working with very high top-gas pressure. Gas-tight sealing is not easy because of the abrasive nature of the dusty top gas which acts as a sand blaster when it passes with high velocity through a narrow leak. The large bell (up to 100 t) is expensive and easily damaged. Hence, charging equipment with up to four bells, with a combination of bells and sealing valves, and with sealing valves have been proposed in only the last 20 years. In the *four bell arrangement* of NKK (formerly Nippon Kokan KK) illustrated in Figure 60 B, the small and the intermediate bells are for gas sealing whereas the large bell is for the distribution of burden. The small bell and the hopper can rotate. By lowering the small bell, the charge falls on the intermediate bell. After closing the small bell, space I is pressurized with cleaned blast furnace gas to a pressure equal to that in the top of the burden bed. The charge is then discharged to the large bell by lowering the intermediate bell. After closing the intermediate bell the excess pressure inside space I is released and the large bell is lowered.

A dual-chamber sealing arrangement by DEMAG is shown in Figure 60 C. Two chambers, each of which can be sealed with two sealing valves, serve as a gas gate preventing the top gas from escaping. The charge enters a movable hopper through two parallel hoppers. The former is equipped with an eccentrically located opening in the bottom. The distribution of the charge over the small bell can be controlled with the movable hopper which influences the burden distribution at the top of the furnace considerably. The most innovative top charging device is the bell-less top of Paul Wurth company (see Fig. 60 D), using a rotating chute. The first Paul

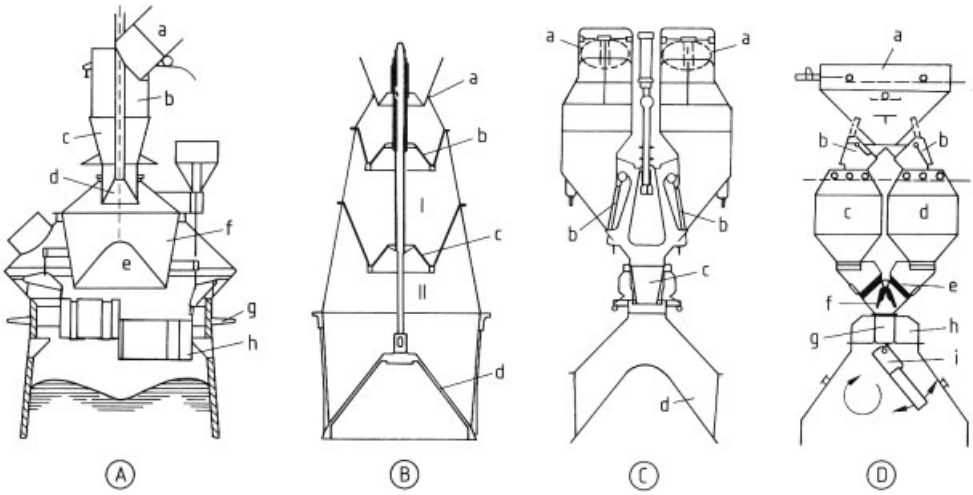

Figure 60. Various charging devices;
A) McKee charging mechanism and movable armor [99]
a) Skip car; b) Charging funnel; c) Movable funnel; d) Small bell; e) Large bell; f) Charging pan; g) Compensator; h) Movable armor
B) Four bell charging mechanism of NKK [100]
a) Movable funnel; b) Small bell; c) Intermediate bell; d) Large bell
C) Dual chamber charging mechanism, DEMAG construction
a) Upper closing valves; b) Lower closing valves; c) Movable funnel; d) Large bell
D) Charging mechanism with chute for material distribution, Paul Wurth construction [101]
a) Receiving hopper; b) Upper sealing valves; c) d) Material bin; e) Material flow regulating gate valve; f) Lower sealing valves; g) Central filling pipe; h) Feeder spout chute drive; i) Rotating distribute chute

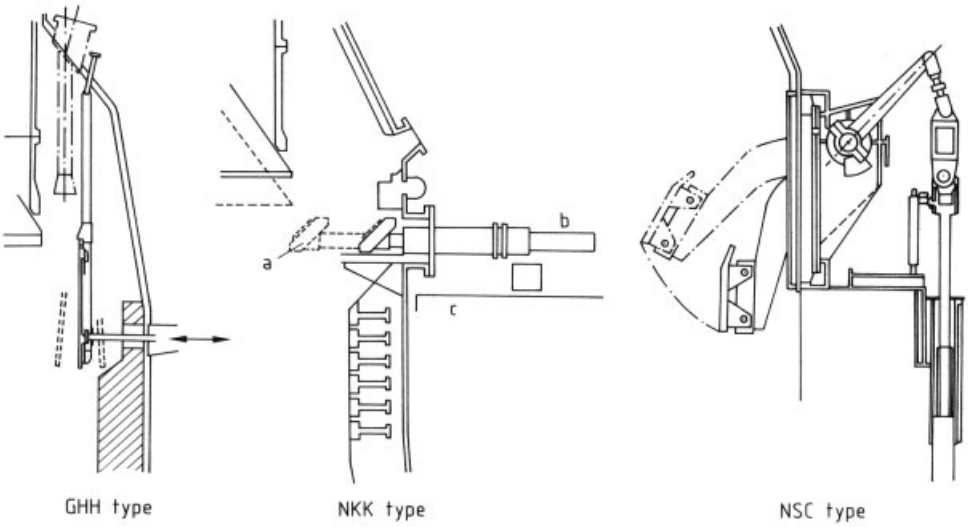

Figure 61. Movable armors [104] (Courtesy of The Iron and Steel Institute of Japan)
a) Stone box; b) Cylinder; c) Platform

Wurth top closing device was installed at blast furnace Hamborn 4 of Thyssen Stahl AG (today Thyssen Krupp Stahl AG) in January 1972 and developed by Thyssen to operational practice [102]. Up to now, worldwide, 186 blast furnaces have been equipped with this system [103]. This device has at least two hoppers with gas-tight sealing. Therefore, one can be discharging into

the furnace without losing the pressure at the top of the furnace while the other is being loaded. The large bell has been replaced by a chute that can be rotated in order to change its inclination angle. The speed of rotation and the angle of the chute can be independently regulated for almost any desired burden distribution, whereas the flow rate of the materials can be controlled by the valves. In addition, the Paul Wurth top has the advantages of lower height and less weight. For a large blast furnace, the charging device requires only a 30 t crane whereas a large bell and hopper arrangement must be equipped with a 120 t crane. It is also too difficult to obtain a suitable distribution with a bell-and-hopper arrangement. Consequently, most large furnaces employ movable *stockline armors* (see Fig. 61) or rotating chutes for the correct distribution of burden materials.

Lining and Cooling. The lining of a blast furnace is subject to a variety of demands as represented in Figure 62. Accordingly, it is constructed as shown in Figure 65. The throat area and the stack are primarily subject to *mechanical impact* and *friction* from burden materials (Figure 62). The lower stack and the belly are subject to *chemical attack* by alkalis. The alkaline compounds deposited from the gas phase on the cooler part of the stack – particularly on the cooler lining – originate from burden materials which evaporate in the hotter bosh region. They form low melting compounds with the lining components. The lower part of the furnace is subject to *wear* by the hot metal and the slag.

Below the tuyeres the lining consists principally of *carbon bricks.* They are preferred to fireclay bricks because of superior thermal properties. However, they are unstable in the oxidizing atmospheres and dissolve easily in the hot metal. Above the bosh, *silicon carbide* bricks are used in the lower one-third part of the stack. They contribute to effective cooling because of their high thermal conductivity. The upper part of the stack is usually lined with *fireclay* which is sufficiently resistant to abrasion.

The amount of cooling water required for a large blast furnace is ca. 9×10^3 m^3/h. The vaporization loss during a cooling operation is ca. 200 m^3 of fresh water per day. About 18 % of the cooling water is used for the cooling of the bosh and stack, 16 % for tuyere cooling, 49 % for the cooling of the hearth, and the remaining 17 % for cooling of the hot blast valves in the hot blast stoves and of the blast furnace motor units for charging device, dust collector, mud gun, inclined runner, air conditioner, and other equipments.

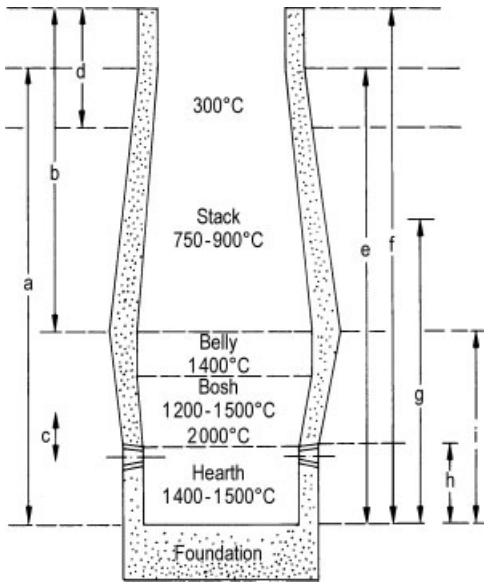

Figure 62. Wear mechanism of the refractory lining of a blast furnace [105]
a) Temperature changes and thermomechanical stress (hearth); b) Abrasion by burden; c) Abrasion by coke particles; d) Impact wear; e) Attack by alkalis; f) Attack by zinc deposit and by CO; g) Attack by slag; h) Attack by hot metal erosion and dissolution in hot metal; i) Attack by oxidation (carbon bricks)

The stack is equipped with *cooling plates* or *stave coolers* inside the steel mantle down to the belly. The coolers are grouped together. Figure 63 shows the so-called 4th generation cast iron stave of Nippon Steel Corporation as well as the "New Cooling Device" (NCD) of Kawasaki Steel Corporation [106, 107] and the durability of cast iron staves of the first three generations. As can be seen in this figure, the staves of the third generation, that were equipped with additional cooling serpentines in the end of the 1970s, have – after wear of the refractory material – a service life of more than ten years. Since 1985 the fourth generation cast iron staves are in use for the lower shaft and bosh area. These staves have a thickness of only 500 mm including the refractory inserts. The experts of

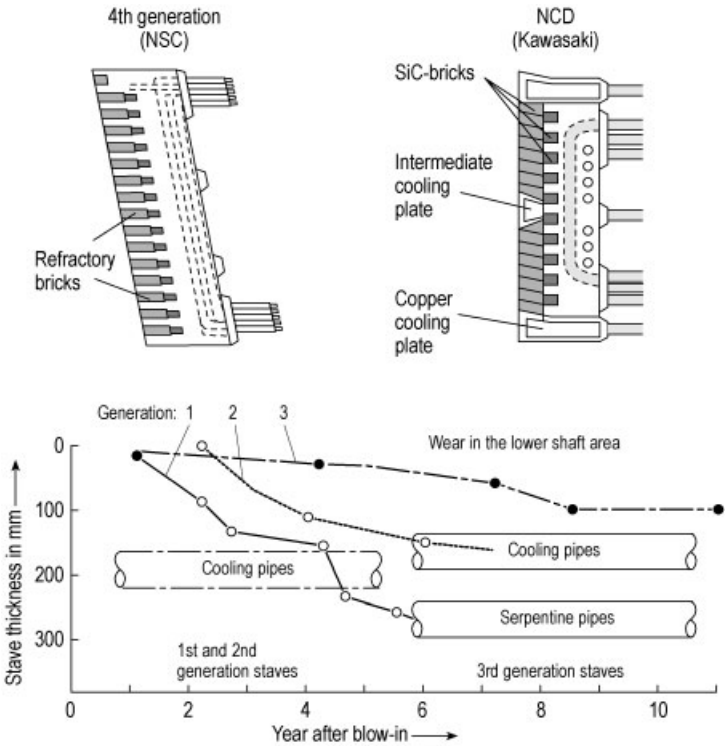

Figure 63. Development status of Japanese cast iron staves

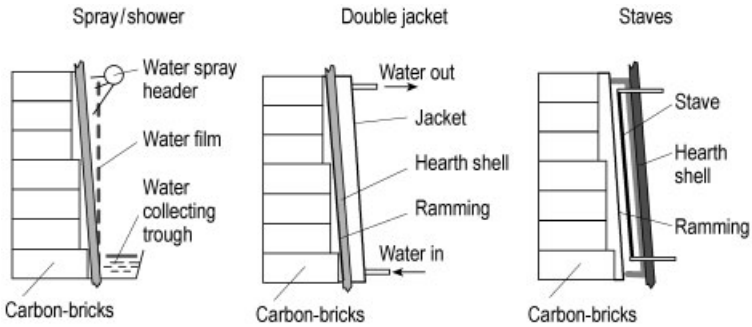

Figure 64. Cooling systems of blast furnace hearth

Kawasaki Steel Corporation aim at a blast furnace campaign life of up to 30 years. For the second campaign of Chiba No. 6 blast furnace – the first campaign lasted over 20 years and ended in March 1998 after a production of 60×10^6 t hot metal – Kawasaki installed in the lower shaft the NCD cast iron stave, which is supported by horizontal copper boxes and has a thickness of 800 mm including refractory lining. Figure 64 shows schematically the different cooling systems of the hearth walls. The spray cooling (left) requires a periodical cleaning of the shell but enables a complete control of the cooled area. The double jacket cooling (middle) and cast iron stave cooling (right) guarantee a dry shell and are more flexible regarding cooling intensity but are difficult to repair [105].

Copper staves developed by SMS GHH and Thyssen Krupp Stahl AG [108] have come into use at numerous blast furnaces for cooling the lower shaft and belly area. Their installation is also planned for cooling the hearth walls. There is a significant difference between the consistency of heat dispersion of nodular cast iron

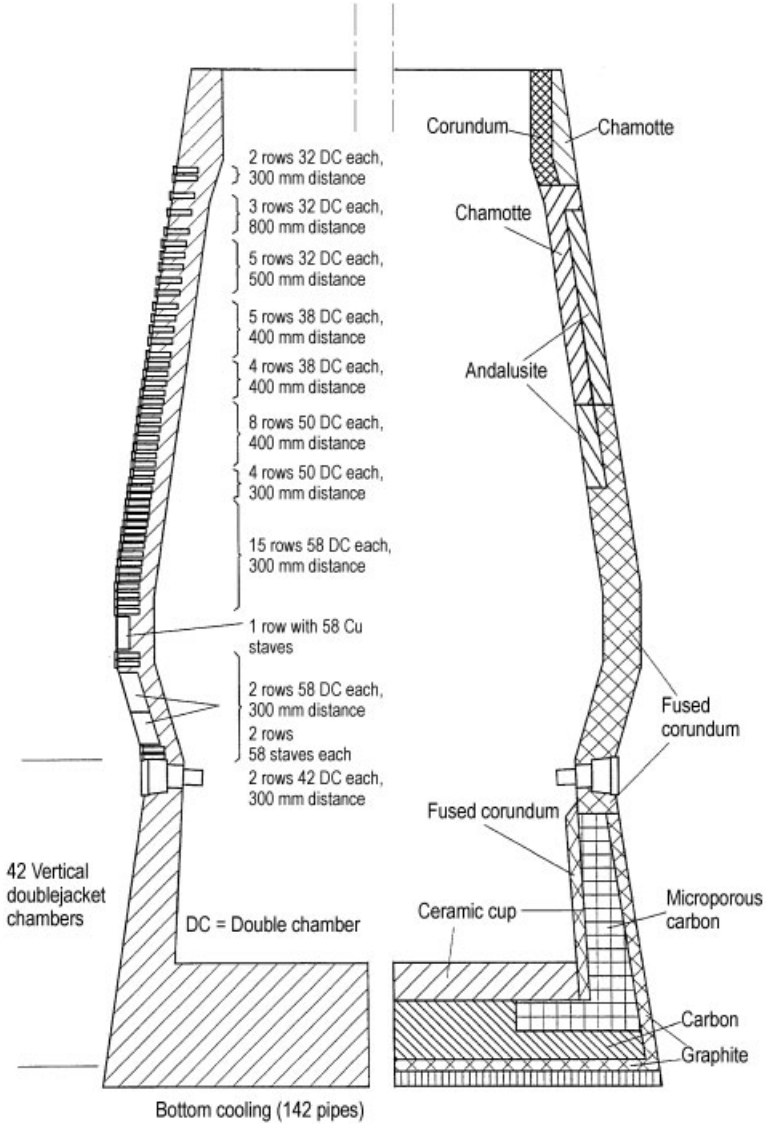

Figure 65. Lining and cooling of a modern large blast furnace (Schwelgern no. 2, Thyssen Krupp Stahl AG), hearth diameter 14.9 m, total volume 6020 m^3 (original figure courtesy of Thyssen Krupp Stahl AG)

staves and of copper staves. Due to the low thermal expansion coefficient of the copper staves, these can be longer than cast iron staves. Owing to the formation of stable protective layers by the burden materials, the use of expensive refractory materials to increase the service life of the copper staves can be dispensed with. These protective layers reduce the heat throughput and the wear of the staves to more or less zero [21]. Furthermore, additional furnace volume is gained by this design. With respect to the service life, wear of about 3 mm was measured on the experimental copper staves of the blast furnace Hamborn 9 at Thyssen Krupp Stahl after more than 9 years in operation. The staves were installed in 1979. Worldwide, copper staves are already installed in approximately 35 blast furnaces. The tuyeres are made of copper cast with water cooling. The bottom of the furnace is cooled by air thereby holding the surface temperature of the brick at 200 – 400 °C.

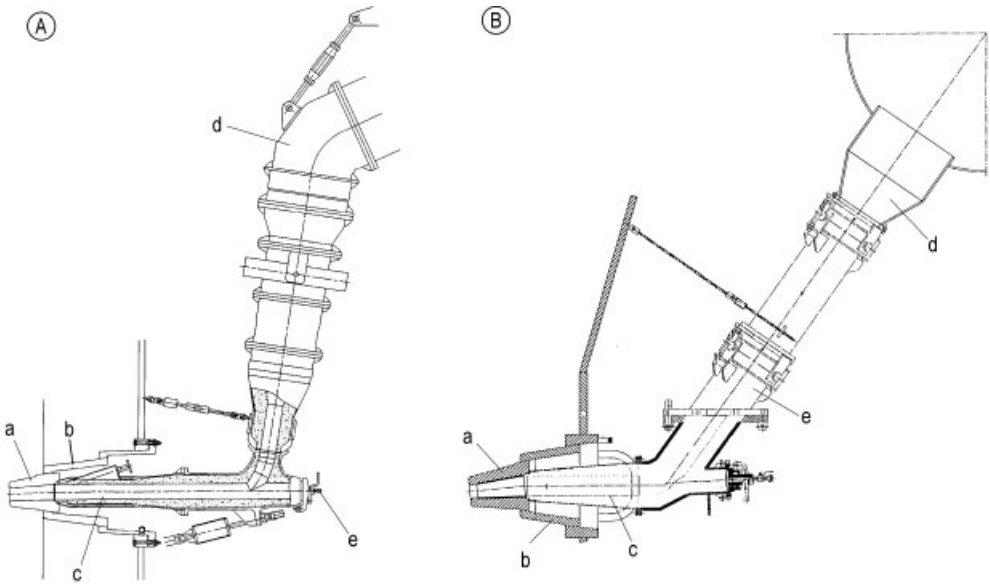

Figure 66. Two sets of tuyere – blowpipe assembly [110] (Courtesy of Thyssen Krupp Stahl AG (A) and Nippon Steel Corp. (B))
a) Tuyere; b) Tuyere cooler; c) Blowpipe with ceramic lining; d) Goose neck; e) Peep site

The cooling system is very important to avoid serious damage to the furnace steel jacket, particularly in the bosh and belly which are subject to a heavy thermal load. Thus the furnace life is extended.

The construction of the refractory lining and the cooling arrangement of furnace Schwelgern no. 2 of Thyssen Krupp Stahl that was blown in 1993 is shown in Figure 65. Particularly noteworthy is the cooling system which consists of copper cooling plates in the wall of the shaft, one row of copper staves in the belly as well as two rows of copper plates, cast iron staves in the bosh, double jacket cooling of the hearth wall, tuyere cooling, and hearth bottom cooling with water. These arrangements suppress thermal stresses in the steel jacket.

The further development of cooling systems depends highly on the development of refractory lining materials with high thermal conductivity. Although high thermal conductivity lining increases the heat losses – thereby requiring a corresponding increase in cooling capacity – it decreases the wear of the lining by allowing a more uniform temperature distribution. The heat loss of the modern large furnaces is less than that experienced in older furnaces because of a smaller surface area : volume ratio. The hot blast pipe and the tuyeres in the new furnaces shown in Figure 66 are also refractory lined because the hot blast temperature is as high as 1350 °C.

Blast. Forty years ago the hot blast temperature was ca. 800 °C. It has been increased up to 1350 °C at present. This has been made possible through the development of high performance hot stoves in the 1960s. Blast temperatures of 1300 to 1350 °C at dome temperatures of 1550 °C were achieved via constructive improvements (hot stoves with external combustion shaft), through heating with rich gas and through the use of silica checker work instead of chamotte checker work. The development of injection technologies was the main prerequisite to use such high blast temperatures because the injectants have a cooling effect on the flame temperature which increases with increasing H_2 content in the injectants. This makes it possible to keep the flame temperatures on a suitable level of 2200 °C, when using hot blast temperatures exceeding 1200 °C. The hot blast temperature could be increased because the gas permeability was increased by elimination of fines and the beneficiation of iron ore which led to lower gangue content, narrow temperature range for

cohesion, and lower reduction disintegration and swelling.

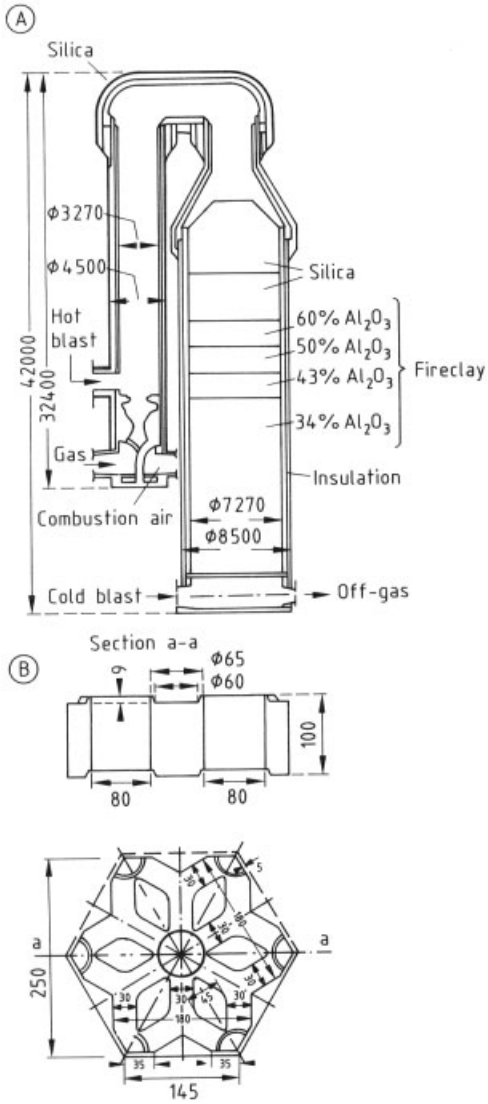

Figure 67. A) Hot blast stove with external combustion chamber [99]
B) A design of checker-building shape for blast temperature up to 1350 °C

The higher blast-temperature increases the production efficiency of the furnace. In addition, the larger part of the blast furnace top gas is utilized for preheating the blast. This results in a decrease in coke rate (coke consumption per tonne of hot metal) because the thermal energy that would otherwise have been lost from the process is returned to the blast furnace in the form of sensible heat. The hot blast stoves for blast furnaces work exclusively on the basis of the regeneration principle that was applied to preheating the blast air for blast furnace operation by E. A. Cowper in 1860. Most blast furnaces are equipped with three or four hot blast stoves. After cleaning, the top gas is used for preheating the blast by combustion in a combustion chamber. The top gas is sometimes enriched with natural or coke oven gas to improve its calorific value. The waste heat of the hot blast stoves can also be used to preheat combustion media. A waste gas heat exchanger, for example, is installed in the main waste gas collector in which waste gas is cooled from a maximum of 300 °C to 123 °C [109]. The heat recovered in this way is conveyed via a thermal oil system to the hot blast stoves and there transferred to the combustion air and gas. The combustion media are heated to around 180 °C. The oil circulation to the heat exchangers of those stoves that are on blast is shut off. The hot gas obtained by combustion is directed from the top to the bottom through a checker column where the thermal energy is stored. After switching to wind, the cold blast is then directed through the column of checkers. Depending on the temperature, the hot blast coming from a hot blast stove is mixed with a certain amount of cold blast or with a lower temperature blast generated from another stove where the temperature has already been decreased.

This mixing is done to regulate the temperature of the hot blast or to keep it constant. A hot blast stove with an *external* combustion chamber to generate high blast temperature is shown in Figure 67. High temperature hot blast stoves with an *internal* combustion chamber are also still in use.

The required size of the hot blast stove is calculated from the following equation:

$$B\Delta \mathrm{T_B C_p\,(B)} = \frac{\mathrm{G\,(LHV)}\,\eta}{100}$$

$$= z\,\varepsilon\, m_{\mathrm{st}} C_p\,(\mathrm{s})\,\Delta \mathrm{T_s n} \qquad (154)$$

where B is the quantity of air [m^3 (STP)/s]; ΔT_{B} the temperature difference between hot and cold blast air (K); C_p (B) the average specific heat of air at standard conditions [J m^{-3} (STP) K^{-1}]; G the quantity of heating gas [m^3 (STP)/s]; LHV the lower heating value of the off-

gas [J/m^3 (STP)]; η the efficiency of the hot blast stove, (%); z the number of hot blast stoves; ε the utilization factor of the checker brick; m_{st} the mass of the checker brick (kg); c_p(s) the average specific heat of the checker brick (J kg^{-1} K^{-1}); ΔT_s the average temperature difference of the checker brick from the beginning to the end of the blast cycle (K); and n the number of cycles per stove and second.

The quantity of air B depends stoichiometrically on the coke rate. When a blast temperature is specified, the output power of the stove required and the value of B can be read from Figure 68 for a specific coke rate which is the independent variable. The thermal efficiency η is ca. 90 %, taking into account the heat loss which includes the heat loss in the bustle pipe. The utilization factor ε of the checker brick takes into account that the center of the brick has a lower temperature than its surface. The value of ε is 0.9 for a blasting period of 0.5 h and an equivalent brick thickness of 30 mm. The temperature variation ΔT_s at the surface of the brick must not exceed 130 K in order to generate high blast temperature.

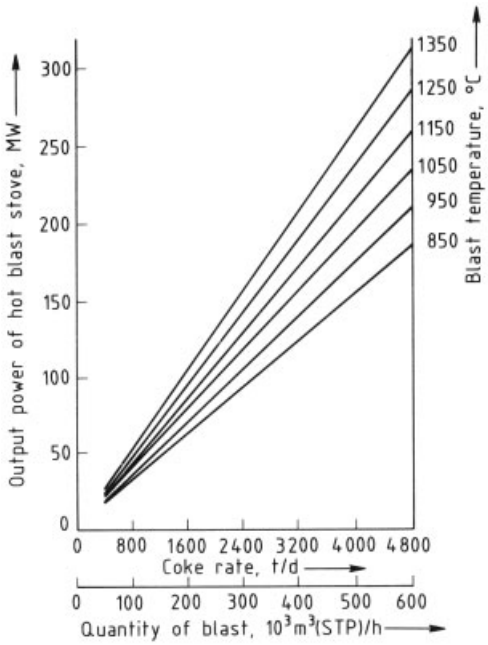

Figure 68. Output power required for a hot blast stove vs. coke rate with temperature as a parameter [111]

To estimate the mass of the checker brick from Figure 68, the power density is assumed to be 35 kW/t. Then, the required checker mass is found to be 7500 t for a coke consumption of 4000 t/d and a blast temperature of 1350 °C. Figure 67 A shows a checker brick column which is economically attractive because of the smaller required brick quantity, its stability to mechanical and thermal load stability and the resulting uniform flow conditions.

Gas Cleaning. The blast furnace generates 1400 to 1700 m^3 (STP) dry top gas per tonne hot metal, depending on the operation mode. The uncleaned top gas contains up to 30 g dust/m^3 (STP). The dust content must be decreased to below 10 mg/m^3 (STP) due to the works' own specifications in order to use the gas in the gas turbines for the blast, to protect burner equipments of other plants in the iron and steel works where the gas is utilized, and due to environmental regulations. The volume of the top gas to be cleaned for a large blast furnace with a production of over 11000 t hot metal per day amounts to more than 16 $\times 10^6$ m^3 (STP) per day.

Because of the cost associated with gas cleaning installations, it has become commonplace for large blast furnaces to operate the compressors for the blast with electric motors instead of gas turbines. The top gas is used for heating purposes. The exhaust gas from the boiler contains 10 – 50 mg dust per cubic meter (STP). It falls within the legal limit for dust emission. Figure 69 shows a schematic of the initial part of the top-gas cleaning system. At first, a dust catcher reduces the dust content of the top gas from 10 – 30 g/m^3 (STP) to 5 – 10 g/m^3 (STP). The dust content is further lowered to < 5 g/m^3 (STP) by the cyclones. Pressure losses occur in the dust catcher and the group of cyclones as well as in the venturi scrubber. In high-top-pressure blast furnaces, in which higher pressure losses are admissible, the top gas cleaning system is often a combination of dust catcher, cyclones, and venturi or ring slot washer. After cleaning, the pressure of the top gas can be about 2 bar higher than that of the works gas system and this excess pressure can be utilized in a turbine for power generation.

The amount of dust and sludges removed by the gas cleaning system are between 8 and 17 kg/t hot metal depending on the operation

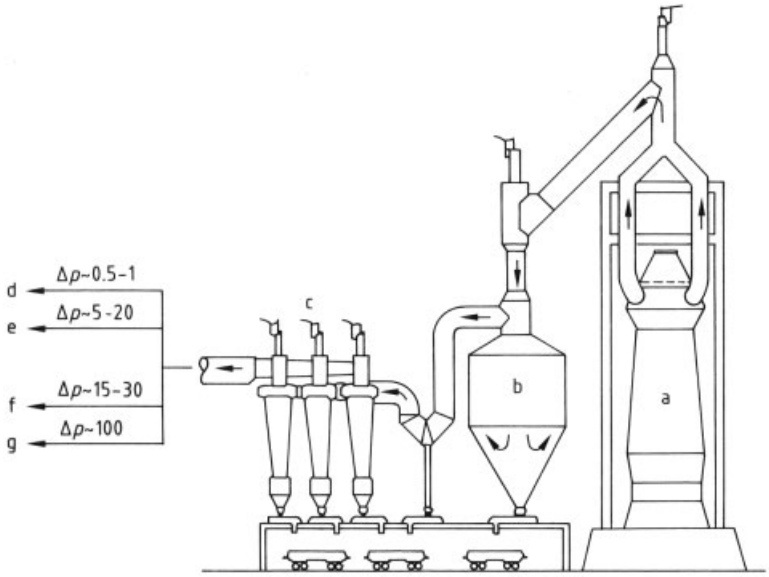

Figure 69. Top gas cleaning system [99]
a) Blast furnace; b) Dust catcher; c) Cyclone; d) To electric precipitator; e) To hurdle scrubber; f) To bag filter; g) To venturi scrubber

mode of the furnace. The process water produced by wet cleaning of the top gas is treated in sedimentation tanks. The blow down is often added to the slag granulation water. The excess water from slag granulation is then conveyed to a purification plant.

2.5.2. Metallurgy of the Blast Furnace

The hot metal produced in the blast furnace is saturated with carbon and is in the liquid state. Depending on burden composition and operation technique of the blast furnace, the molten hot metal produced contains varying amounts of other elements such as silicon, manganese, phosphorus and sulfur. The blast furnace produces hot metal, not steel.

A blast furnace is operated under the following three major prerequisites:

1) *Lowest energy consumption,* particularly the lowest coke rate;
2) *High productivity*, i.e., a high production rate per unit volume of the furnace;
3) *Composition* and *temperature* of the liquid hot metal produced must meet the requirements of the steel shop and must be constant.

Mass and Heat Balances. Table 20 shows the simplified mass and heat balances of the reactions that proceed sequentially during the descent of the burden in the blast furnace. All figures in the table are based on 1 mol Fe_2O_3, 1/4 mol $CaCO_3$ as a flux, and k mol of coke as reductant.

Hematite is first reduced via magnetite to wustite by the gas. Because the heat requirement of the reactions are small, the heat content of the gas is sufficient for heating the burden (see Fig. 70). The gas composition prevailing in the upper part of the furnace (below 1000 °C) has the reducing potential for hematite and magnetite as depicted in Figure 71. Only a portion x of the wustite is reduced to iron because of kinetic limitation. The deviation of the gas composition from the Boudouard equilibrium *above* 650 °C is due to the low reactivity of coke. Thus, carbon dioxide is not adequately converted to carbon monoxide. Deviation *below* 650 °C is caused by suppression of the carbon deposition reaction, $2\ CO \rightleftharpoons CO_2 + C$. For coke saving, the former deviation from equilibrium is desirable, however, the latter undesirable.

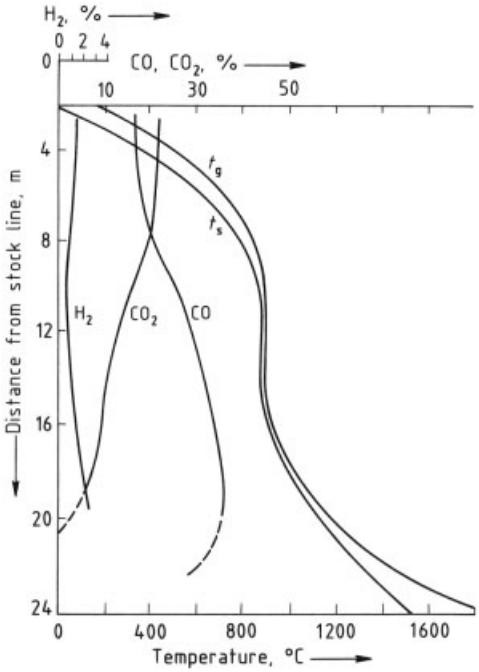

Figure 70. Schematic profile of temperature and gas composition in the longitudinal direction of the blast furnace

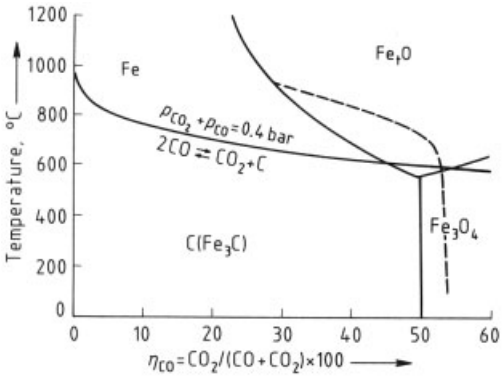

Figure 71. Variation of temperature and gas composition in a typical operation of the blast furnace superimposed on the reduction equilibrium diagram of iron oxide

During descent of the burden, limestone starts to decompose at ca. 700 °C. This reaction requires heat and releases carbon dioxide, therefore the gas is cooled and its carbon dioxide content increases. Consequently, the reduction rate of iron oxide decreases and the value of x in Table 20 decreases. A lower value of x corresponds to a higher coke rate according to the balance shown in Table 20. The wustite that has not yet been indirectly reduced by the gas to iron must be reduced directly with carbon. This reaction requires a great amount of heat. For simplicity it has been assumed in Table 20 that direct reduction of wustite proceeds at 1100 °C. This zone of the furnace is critical for the completion of the entire process. The amount of oxygen which *must* still be removed from the ore in this part of the furnace depends on the degree of reduction achieved in its upper part. Therefore the residual amount of oxygen depends on the reducibility of the ore. On the other hand, the amount of oxygen which *can* be removed from the ore in this zone depends on the enthalpy that is delivered by the gas at 1100 °C. This enthalpy, in turn, depends on the amount of coke that is available for gasification in front of the tuyeres.

If *excess heat* is available, a higher top-gas temperature is found. However, the utilization degree of the gas η_{CO} decreases because the excess heat is supplied by gasification of coke to carbon monoxide. This phenomenon can be controlled by decreasing the coke rate of the burden materials fed at the top of the bed. However, the action will only be effective after 4 – 6 h when the burden charged at new rates comes to the tuyere zone. An immediate effect is obtained if the heat input to the lower part of the furnace is decreased by increasing the moisture content or decreasing the temperature of the blast.

It is more difficult to compensate an *enthalpy shortage* in the lower part of the furnace. An increase in the coke rate is effective only after several hours. It is often not possible to increase the temperature of the blast because the blast temperature is usually at its maximum for economic reasons. The maximum blast temperature is limited by the capacity of the stoves and also by the maximum allowable flame temperature of the blast furnace. A short-term solution to compensate an enthalpy shortage is to inject reductants such as hydrocarbons, natural gas, oil, or pulverized coal together with an oxygen enrichment of the blast.

Upon further descent of the burden, the enthalpy $\Delta H_{E,\,S}$ needed to melt iron and slag must be supplied. For simplicity, they are assumed to melt down at 1300 °C (see item 6 in Table 20).

The various heat sinks are compensated by the equal amount of heat supplied from the gasi-

Table 20. Simplified mass and heat balances in the blast furnace as a countercurrent reactor [112]: increase (+), decrease (−)

	C	Fe_2O_3	FeO	Fe	$CaCO_3$	CaO	CO	CO_2	O_2	N_2	Heat [a]
1) Charge consisting of 1 mol Fe_2O_3, 1/4 mol $CaCO_3$, k mol carbon	$+k$	+1			+1/4						
2) Indirect reduction hematite → wustite $Fe_2O_3 + CO \rightarrow 2\,FeO + CO_2$		−1	+2				−1	+1			negligible
3) Indirect reduction wustite → iron x (2 FeO + 2 CO → 2 Fe + 2 CO_2)			$-2x$	$+2x$			$-2x$	$+2x$			negligible
4) Decomposition of limestone 1/4 $CaCO_3$ → 1/4 CaO + 1/4 CO_2					−1/4	+1/4		+1/4			ΔH_k for 1 $CaCO_3$ 67.2 kJ at > 700 °C
5) Direct reduction of remaining wustite ($1-x$) (2 FeO + 2 C → 2 Fe + 2 CO)	$-2(1-x)$		$-2(1-x)$	$+2(1-x)$			$+2(1-x)$				ΔH_{CO} for 2 FeO 315 ($1-x$) kJ at 1100 °C (heat of reaction)
6) Melting of iron and slag											ΔH_{es} for 2 Fe: 42 kJ at 1300 °C
7) Gasification of carbon y (C + 1/2 O_2 + 2 N_2 → CO + 2 N_2)	$-y$						y		$-y/2$		ΔH_C for 1 C −111 kJ
8) Hot blast (5/2) y mol									$+y/2$	$+2y$	ΔH_L for 5/2 mol air $-0.084\,(t_B - t_R)$ kJ [b]
Input at top	$+k$	+1			+1/4						
Input at bottom									$+y/2$	$+2y$	
Output at top			0				$y + 1 - 4x$	$1.25 + 2x$		$+2y$	
Output at bottom	$k - 21 - x - y = 0$	0	0	+2	0	+1/4					

[a] Heat of reaction not based on reaction temperature.
[b] t_B = Hot blast temperature. t_R = Temperature at which the direct reduction is assumed to proceed.

fication of the coke and from the hot blast as described in items 7 and 8 in Table 20. This requirement almost determines the coke rate for various conditions. This is an approximate calculation because the coke requirement for the reduction of molten wustite, silica, and manganese oxide as well as that for the carbon dissolved in the metal has not been taken into account. To consider the heat and mass balances, the furnace is divided into two parts by a boundary line which is drawn in the thermal and chemical reserve zone. In this zone, the temperature is almost constant (ca. 900 °C) and no reduction of iron ores occurs.

The coke rate is determined by the enthalpy requirement of the lower furnace. The quantity of carbon y (mol of carbon per moles of hematite to be reduced) needed to supply the heat is given by Equation (155) (for symbols see Table 20):

$$y = -\Delta H_{E,S} + \frac{(1-x)\,\Delta H_{CO}}{\Delta H_C + \Delta H_L} \tag{155}$$

k, the total amount of carbon required, is given by (see Table 20):

$$k = y - 2x + 2 \tag{156}$$

The amount of carbon required for direct reduction is:

$$k - y = -2x + 2$$

Substitution of the figures listed in Table 20 gives the carbon rate as a function of the hot blast temperature t_B and the amount of wustite reduced indirectly x, assuming that the direct reduction proceeds primarily at t_R = 1100 °C

$$k = \frac{357}{111 + 0.084(t_B - 1100)} - \left(\frac{2 + 315}{111 + 0.084(t_B - 1100)}\right) x + 2$$

The coke rate Q is then given by:

$$Q = 118k$$

on the basis of a carbon content of the coke of 85 % and on iron content of the hot metal of 93.5 %.

The *fraction of indirect reduction* r_{CO} indicates how much oxygen has been removed from the ore by indirect reduction. Together with the fraction of direct reduction r_C, the sum should be equal to unity:

$$r_{CO} + r_C = 1 \tag{157}$$

For reduction of hematite, the following relation is valid (Table 20, items 2 and 3):

$$r_{CO} = \frac{O_{removed}}{O_{total}} = 1 + \frac{2\,x}{3} \tag{158}$$

On the basis of the *balance* shown in Table 20, it is not possible to assign a value to x (amount of wustite indirectly reduced). The maximum value of x can be determined thermodynamically. If the value of x is given, then r_{CO} is obtained from Equation (158). In fact, the value of x is determined by the rate of the indirect reduction of iron ore and by the reactivity of the coke which, in turn, determines the degree to which the Boudouard reaction proceeds. The value of x can be evaluated from the composition of the top gas if the operational data are available. The gas utilization factor is defined by Equation (159):

$$\eta_{CO} = \frac{\%CO_2}{\%CO_2 + \%CO} \tag{159}$$

Substitution of the values in Table 20 provides the following Equation:

$$\eta_{CO} = \frac{(1+2x)}{2\,(1-x) + y} \tag{160}$$

Since y depends only on the enthalpies and x for a given blast temperature (see Eq. 110), the value of x can be determined if the decomposition of limestone is neglected. In practice, r_{CO} is between 0.45 and 0.65 and the corresponding values of x are between 0.175 and 0.475.

The *amount of air* l required per mole of hematite and the *amount of top gas* generated g_G are obtained from Table 20:

$$l = \frac{5y}{2} \text{ (amount of air)} \tag{161}$$

$$g_G = 3y - 2x + 2 \text{ (amount of top gas)} \tag{162}$$

These equations are approximate expressions and should be improved for more accurate calculations – particularly when the calculated results are to be used for furnace control to keep the composition and temperature of the hot metal constant. For this purpose it is necessary to formulate the heat and mass balances on hydrogen from water vapor and hydrocarbons and on the solutes in the iron (e.g., carbon, silicon, manganese, and phosphorus).

2.5.3. Effective Utilization of Energy

The high thermal efficiency of the blast furnace (Table 21) directly leads to a low coke rate. Heat losses through the wall in this example amount to ca. 4 %. Additionally, 2 – 5 % of thermal energy is lost in the form of top-gas enthalpy. These values decrease with increasing furnace size so that the total loss of thermal energy is <6 %. Compared to the low *thermal energy loss* of the blast furnace, the *chemical energy* of the top gas is rather large and amounts to ca. half of the thermal energy used effectively in the furnace.

Table 21. An example of enthalpy balance for the modern blast furnace (no. 6 BF at Chiba works of Kawasaki Steel Corp.)

Enthalpy	$\times 10^3$ kJ/ t hot metal	%
Input		
Raw materials	519	3
Coke	14 412	83
Blast	1 411	8
Electricity	1 093	6
Total	17 439	100
Output		
Top gas	5389 (chemical: 4966, thermal: 423)	31
Hot metal	9 697	56
Slag	908	5
Dust	255	1
TRT's recovery [a]	536	3
Wall, heat loss	653	4
Total	17 439	100

[a] TRT: Top-gas pressure recovery turbine.

The thermal energy supplied by the hot blast (Table 21) should be partly subtracted from the total energy required for the blast furnace operation because the blast is primarily heated by the top gas. Even in modern blast furnaces ca. 30 % of the primary energy supplied by coke leaves the process in the form of carbon monoxide and hydrogen in the top gas. A heat loss of 3 – 7 % is caused by *slag enthalpy.* With these

considerations in mind, lowering of the energy requirement can be accomplished as follows:

1) Use the chemical energy of the top gas as efficiently as possible in the process system,
2) Decrease the specific slag volume,
3) Prepare the burden (operation of the stack), i.e., carry out dehydration and decarbonization outside the blast furnace by utilizing inexpensive energy sources.

Additionally, efforts exist to replace metallurgical coke by other inexpensive energy sources because of high cost of coke. Table 22 shows the calculated energy balance of a blast furnace which is fitted well to newer operational results [72]. These calculations were done to assess whether carbon and hydrocarbon carriers are used in the blast furnace as fuel or as chemical raw material. Table 22 indicates that about 70 % of the energy supplied by the coke is converted into energy that is chemically bonded in the hot metal. 14 % equates to the low calorific value of the top gas which is inevitably produced in the blast furnace process and which is used in other parts of the integrated iron and steel works. Only 13 % of the energy converts into measurable heat of the hot metal and the slag. In this case only about 3 % is lost through the heat loss of the furnace walls or through the top gas.

Similar conditions prevail if coke replacement materials are injected via the tuyeres. The utilization ratio of reducing agents and energy used for the blast furnace process thus amounts to 97 %, which is an outstanding value. The carbon and hydrocarbon carriers used in the blast furnace should be classed as "chemical raw materials" and not as "fuels" as their components are mainly involved in the chemical reactions of the iron ore reduction [114].

Blast Furnace Operation and Physical Properties of Burden Materials. The heat required to produce 1 t of hot metal substantially depends on the physical properties of the iron ores. In modern blast furnace operation, processed ores such as sinter and pellets are usually fed to the furnace to conduct stable operation for pursuing higher productivity and lower coke rate. Physical properties which significantly affect the blast furnace operation are reducibility and reduction disintegration. *Higher reducibility* contributes to the increase in indirect reduction

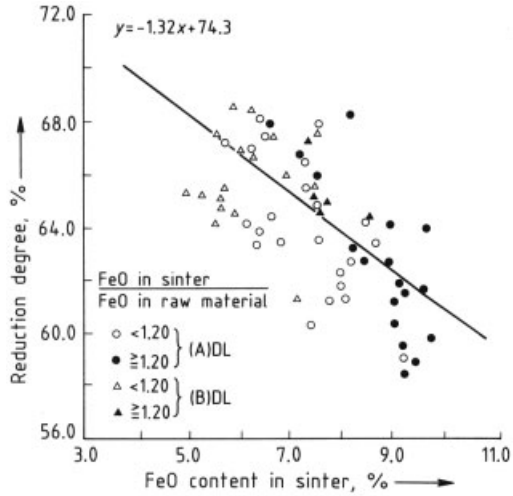

Figure 72. Influence of FeO content in sinter on the reducibility [115] (Courtesy of The Iron and Steel Institute of Japan)

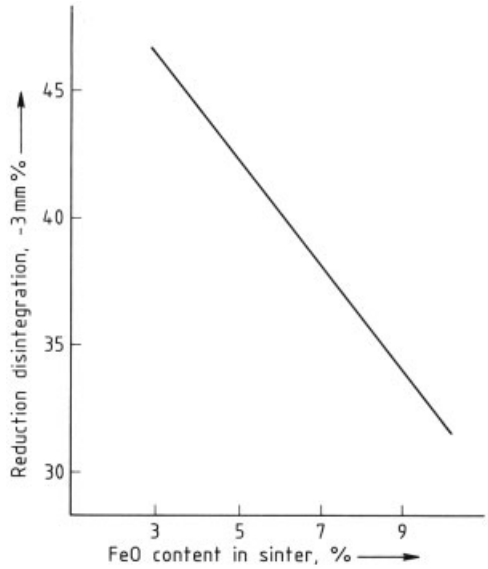

Figure 73. Relation between FeO content in sinter and reduction disintegration degree [116] (Courtesy of The Iron and Steel Institute of Japan)

in the stack region thereby increasing the utilization factor η_{CO} and decreasing the coke rate because of a decrease in direct reduction. *Reduction disintegration* directly affects the gas permeability of the burden column and may cause flow maldistribution which negatively affects the reduction of iron ore and gas utilization. There-

Table 22. Total balance of the blast furnace

	H_i, GJ/t hot metal	Base LHV_C, %	End distribution of enthalpies, %
1. Supplied enthalpies			
Coke, LHV, m_c incl. $C._{dissol}$*	12.6647	100.00	
Hot blast	1.6220	12.81	
Sum	14.2867	112.81	
2. Reaction enthalpies			
Reduction of Fe_2O_3 – FeO	6.7698	53.45	
Heating value $C_{dissol.}$	1.5922	12.57	
Reduction SiO_2, MnO	0.1597	1.26	
Sum	8.5217	67.29	67.29
3. Enthalpy of the top gas (latent enthalpy)			
Heating value CO	3.3596	26.53	26.53
less the preheating of the blast			−12.81
Sum		26.53	13.72
4. Enthalpy of the hot metal and slag (sensible enthalpy)			
Hot metal	1.0531	8.31	
Slag	0.6465	5.10	
Sum	1.6996	13.42	13.42
5. Enthalpy of the top gas (sensible enthalpy)			
CO	0.0537	0.42	
CO_2	0.0878	0.69	
N_2	0.1662	1.31	
Sum	0.3077	2.43	2.43
6. Losses in the furnace	0.1480	1.17	1.17
Sum 2 – 6	14.0366	110.83	98.03
7. Variance	0.250	1.97	1.97

* LHV = lower calorific value; $C_{dissol.}$ = carbon dissolution.

fore, the physical properties of the sinter should be controlled appropriately. A lower FeO content in the sinter enhances the reducibility, however, a higher content improves the reduction disintegration as shown in Figures 72 and 73.

A low gangue content of the burden materials also contributes significantly to the coke savings by decreasing the slag volume.

Chemical Energy of the Top Gas. Burden materials with a high reducibility and a low reduction disintegration provide a low coke rate, but a high gas utilization factor η_{CO}. A better burden preparation allows the use of higher blast temperatures which results in the use of a larger portion of the top gas for heating the blast. These circumstances significantly decrease the waste energy that leaves the blast furnace process in the form of off-gas. Figures 74 and 75 illustrate the average decrease in the coke consumption and the total reductant consumption, respectively, of the blast furnaces operated in Germany and in Japan [21]. They also show the effective replacing of coke by oil and pulverized coal. A productivity increase of a blast furnace is coupled with an increase in gas throughput and a reduction of the specific gas volume. A reduction of the specific gas volume can also be obtained by reducing the consumption of the specific reducing agents.

The productivity increases of the blast furnaces in Germany and in Japan took place simultaneously with the decrease in reductant rate. It is noticeable that in Japan a high capacity level was attained much earlier than in Germany. This is due to the setting up of the Japanese steel industry in the 1960s using large blast furnaces and, at the same time, the latest plant technology. However, the closer the reducing agents consumption gets to the process technology minimum, the flatter the curve becomes. Additional reserves for increasing productivity and decreasing reducing agents consumption could be exploited by changing the structure of the reducing agents in conjunction with the use of higher amounts of oxygen as well as improvements in burden qual-

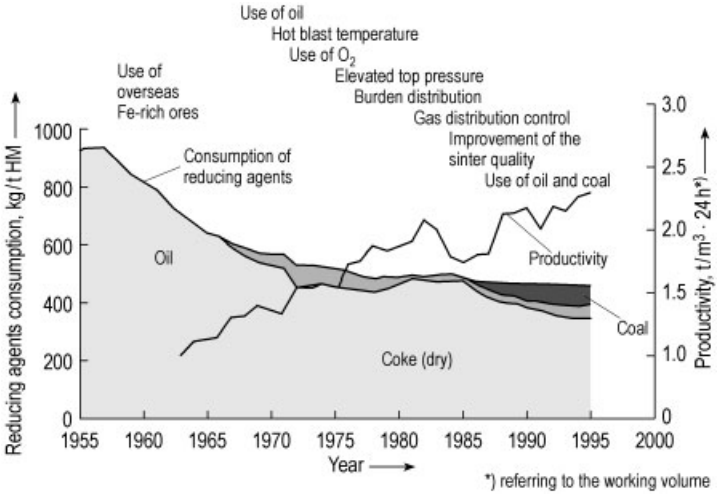

Figure 74. Development of the average reducing agents consumption and the average productivity of blast furnaces in Germany

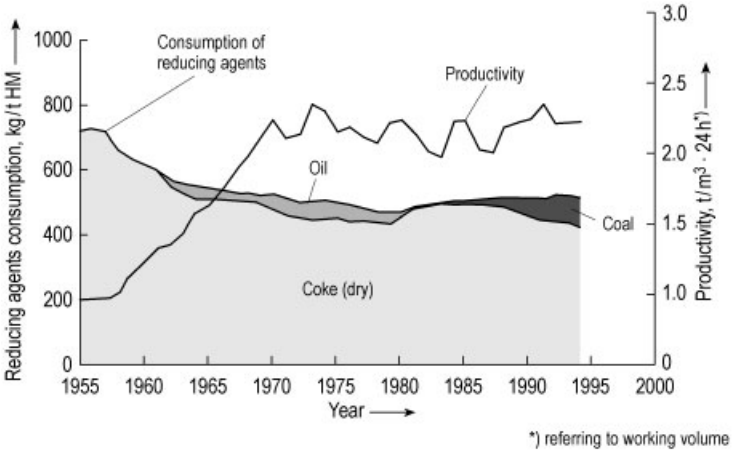

Figure 75. Development of the average reducing agents consumption and the average productivity of blast furnaces in Japan

ity. The theoretical limits of the coke rate for a fully oxidized burden is approximately 400 kg/t hot metal, for the use of 60 % prereduced burden approximately 350 kg/t hot metal and for the use of 100 % prereduced burden lower than 250 kg/t hot metal. When 100 % prereduced burden is used, the blast furnace is operated as a melting furnace like a cupola that produces cast iron. In this case, a gas with a very high calorific value is produced because the blast furnace works as an iron and gas producer. Prereduced burden such as hot briquetted iron (HBI) from direct reduction plants is only used in a few blast furnaces in the world in small amounts to increase productivity. Compared to sinter and pellets, HBI generally is an extremely costly burden material for blast furnaces.

In spite of the successful efforts to decrease the coke rate and to use auxiliary reductants since the 1950s, the problem to what limit the coke rate can be lowered is still not solved. Sufficient coke must be present in front of the tuyeres to form a dead man (Section 2.5) through which the gaseous products can ascend. In addition, the coke consumed in the dead man must be replaced. If the quantity of coke required for these portions of the process is not available, then the characteristics of the process change and it can no longer be called a blast furnace process.

Figure 76 gives a schematic illustration of the processes occurring in the raceway in front of the

tuyeres [118]. According to direct observations, coke particles are entrained in the turbulent air flow immediately in front of the tuyere opening and fly on circular trajectories within the raceway.

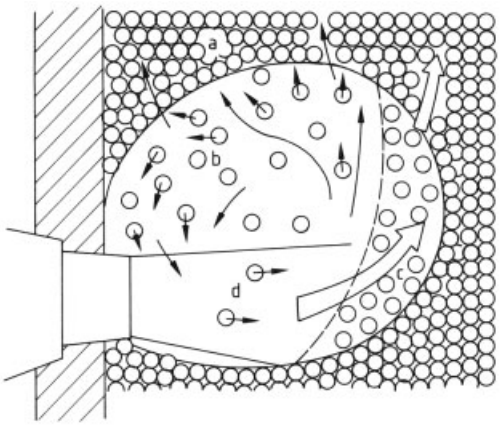

Figure 76. Movement of coke particles in front of the tuyere and location of gasification zone (raceway)
a) Packed zone; b) Coke fluidizing zone; c) Coke slow moving zone; d) Blast jet

The coke particles are primarily gasified in the raceway. According to the gaseous composition distribution in the raceway shown in Figure 77, coke burns with the gas at first forming carbon dioxide, and then carbon dioxide is reduced to carbon monoxide in an area near the coke-filled bed. A typical gas composition change from the tuyere via the raceway to the packed bed calculated by a two-dimensional model [119] is shown in Figure 78. From this figure the gasification reaction has almost gone to completion in the raceway region, therefore a very small quantity of CO_2 is left for the coke packed bed around the raceway.

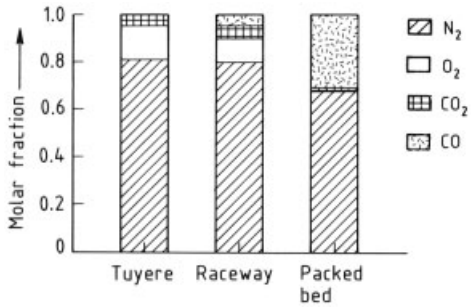

Figure 78. Change of gas composition from the tuyere to the packed bed

Moreover, the direct observations indicate that a shower of iron and slag droplets descends in front of the tuyeres, together with coke particles falling through the air flow. In addition, in no other part of the raceway roof was the burden observed to descend. On the basis of these observations it is assumed that the molten iron and slag trickles down in the space between the in-

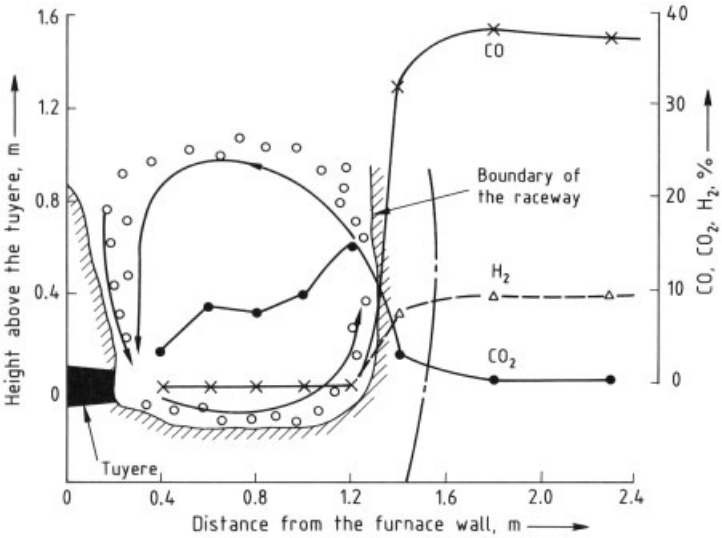

Figure 77. Combustion of coke in the raceway [117] (Courtesy of The Iron and Steel Institute of Japan)

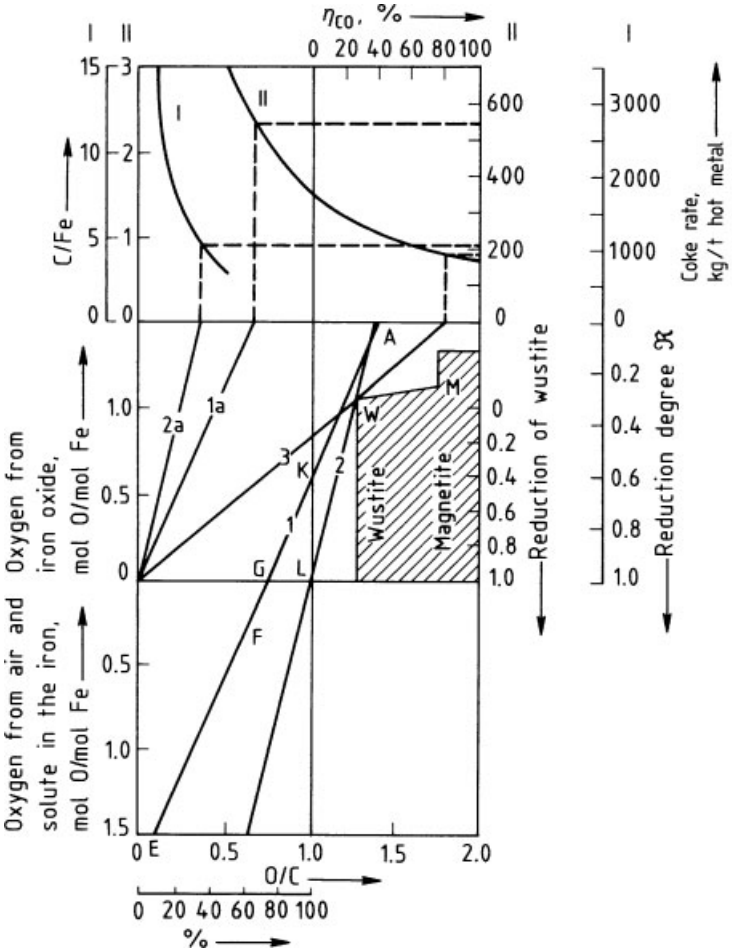

Figure 79. Extended Rist diagram [111]
Line 1: normal blast furnace operation (standard); Lines 2 and 3: hypothetical operation (Line 2: coke for reduction = 0, line 3: coke for heating = 0)

dividual tuyeres and within the dead man. This phenomenon has been proven by dissection of blast furnaces [90, 120].

Fluid flow studies have shown that the adequate quantity of gas cannot flow through the cohesive layers or through the layers which hold up the molten materials unless a sufficient volume of coke bed (which has a high permeability) is present inside the furnace.

Basic Concept to Improve Blast Furnace Performance. The following technologies or their combinations have been principally approved to decrease the coke rate and to increase the productivity of the blast furnace:

1) Pretreatment of iron ores improves the reducibility and permeability and hence, increases the gas utilization η_{CO}. Examples are *sintering* or *pelletizing* of the ore and *screening* to separate the fines.
2) Injection of *auxiliary reductants* such as hydrocarbons, oil, and pulverized coal replaces coke for heating and reduction purposes.
3) Oxygen enrichment of the blast together with reductant injection decreases the amount of blast; this decreases the amount of top gas per tonne of coke and increases the furnace efficiency.
4) *Increase in the blast temperature* reduces the amount of coke required for heating.

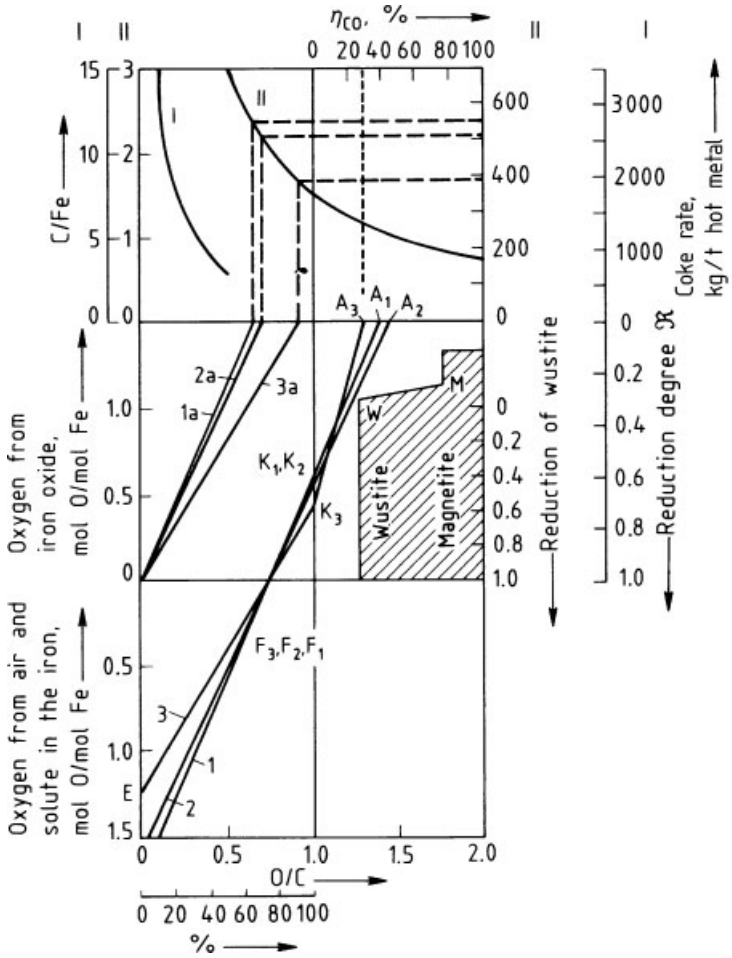

Figure 80. Effect of measures to decrease the coke rate and to increase the productivity of the blast furnace estimated by the Rist diagram [111]
Line 1: normal blast furnace operation (standard); Line 2: improvement in gas permeability and reducibility; Line 3: improvement in indirect reduction (K_3) by injecting reducing gas into the stack

5) *High pressure operation* allows the introduction of a larger amount of blast per unit time.
6) *Reducing gas injection* in the lower part of the stack lowers the amount of coke required for reduction.
7) The use of *partially reduced ore* results in a decrease in the coke rate and an increase in the productivity of the furnace.

The effect of these various technologies on the coke rate and their limits of applicability can be summarized with the aid of the extended Rist diagrams shown in Figures 79, 80 and 82.

Table 23 lists quantitative results obtained from operational data concerning the technologies applied to decrease the coke rate on the basis of the Rist diagram and to increase the productivity.

Curve 1 in Figure 79 represents the conditions in a blast furnace without additional technologies. Heat is liberated during the oxidation of carbon (coke for heating) by the hot blast (E to F). The segment FG represents the reduction of elements dissolved in iron and segment GK corresponds to direct reduction. The indirect reduction is represented by the segment AK. The diagram allows the determination of the amount of carbon required per mole of iron (C/Fe) and the amount of coke (85 % C) required per tonne of hot metal (93.4 % Fe). For this purpose, the operating line 1 is shifted to the line 1 a, which

Table 23. Effect of various parameters on the coke rate of the blast furnace, based on 1 t of hot metal [112, 113]

Parameter	Range	Change	Effect on coke rate kg/t hot metal
Slag volume, kg/t	250 – 350	−100	−15 to 25
Partially reduced pellets,[a] % of burden	0 – 30	+10	−1.5 to 2.0
Fully reduced pellets,[b] kg/t	0 – 20	+10	−2 to 3
Scrap, kg/t		+10	−4
Gas velocity in stack, [c] m/s	2.5 – 5.0	−0.1	−2.5 to 3.0
Hot metal silicon, %	0.5 – 1.5	−0.1	−4 to 7
Fines in burden, % < 5 mm	0 – 7	−1	−4 to 7
Crushed ore, % of burden			
8 – 40 to 8 – 30 mm	0 – 40	40	−10 to 13
8 – 30 to 10 – 25 mm	0 – 40	40	−5 to 7
Coke ash content, %	6 – 12	−1	−5 to 10
CSR value of coke, %	55 – 65	+1	−0.66 [113]
Pellet content of burden, %	0 – 100	+10	−5 to 10
Sinter content of burden, %	0 – 100	+10	−5 to 10
Blast moisture content, [d] g/m^3		−10	−6 to 10
Blast temperature, °C	900 – 1350	+100	−8 to 20
Oil injection, kg/t	50 – 130	+1	−1.1 [113]
Blast oxygen enrichment, %		+1	+2 to 3 % [e]
Top gas pressure, kPa	10 – 250	+10	+1 to 2 % [e]

[a] With 68 % Fe, 40 % metallic iron.
[b] Fully metallized.
[c] With respect to empty stack (superficial velocity).
[d] Air at standard conditions.
[e] Increase in melting rate.

goes through the origin. The intersection of 1 a with the abscissa in the upper part of the diagram together with curves I or II gives the carbon consumption (C/Fe) on the ordinates I or II on the left side, while the coke consumption can be read from the ordinate on the right. The lower abscissa (0 – 100 %) is valid for the segment between E and K. It shows the coke consumption for heating (EF), for reduction (GK), and for the reduction of minor constituents in the burden (FG). For the blast furnace operation in accordance with line 1, the coke rate is 550 kg per tonne of hot metal which is equivalent to a C : Fe ratio of 2.3. The consumption ratio is 60 % for heating, 25 % for direct reduction, and 15 % for the reduction of the minor constituents in the burden.

Lines 2 and 3 are theoretical operating lines for obtaining the limits of the coke rate. *Line 2* (no coke used for reduction) is based on the assumption that the reactivity of the coke is so poor that the Boudouard reaction and the direct reduction are suppressed completely. However, it is also assumed that a gaseous composition is reached which corresponds to the reduction equilibrium from wustite to iron (point W). In this case the entire quantity of carbon monoxide required for the reduction must be generated by the gasification of coke. This condition causes the extremely high coke rate (line 2 a) of ca. 1000 kg per tonne of hot metal, a lower gas utilization η_{CO} as well as higher temperatures of top gas and molten iron.

Line 3 is based on the assumption that no coke is required for heating and that the total heat required is supplied by a hot inert gas. In that case the coke rate would be ca. 200 kg per tonne of hot metal while the gas utilization would be 80 %.

Line 2 in Figure 80 shows the decrease of the coke rate caused by the improvement of gas permeability and reducibility compared with the standard case (line 1). Line 2 gives an improvement of 5 % in gas utilization thereby more closely approaching the equilibrium point W in the wustite field. In addition, the degree of direct reduction (K_2) decreases, which leads to a coke saving of 30 kg per tonne of hot metal (tHM). The reducibility of various burdens are shown in Figure 81, which were measured according to Japanese industrial standard.

It is evident from Figure 80 that improvement in the physical properties of burdens does not result in a large decrease in the coke rate but

can probably contribute much to the stable operation of the blast furnace. A lower coke rate (i.e., a lower slope of the operating line in Fig. 80) could only be achieved if it were possible to increase the degree of direct reduction together with the gas utilization. This method would lead to an increase in the heat requirement that must be supplied by a hotter blast if coke is to be saved. However, the hot blast temperatures have already reached their technological limit.

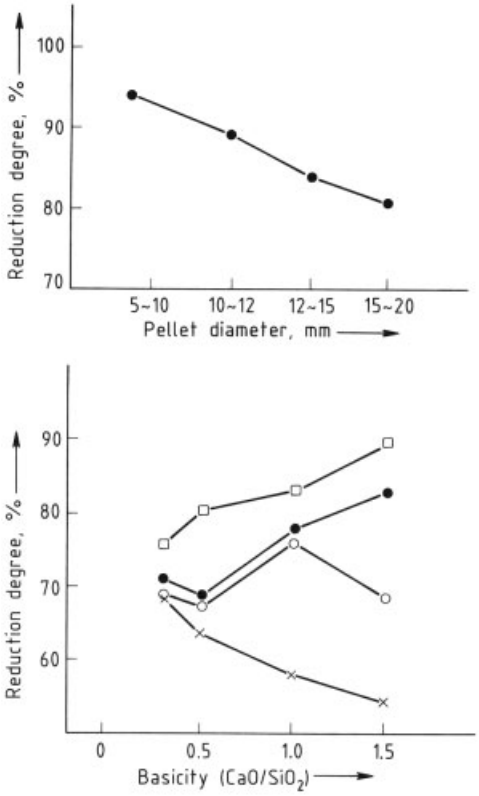

Figure 81. Influence of basicity and pellet diameter on reduction degree [122] (Courtesy of The Iron and Steel Institute of Japan)
—□— 1200 °C; —•— 1250 °C; —○— 1300 °C; —×— 1350 °C

Reductants injection through the tuyeres (e.g., pulverized coal or hydrocarbons) can supply a part of the heat requirement. Line 2 in Figure 82 is an example of an oil injection rate of 60 kg/t hot metal and a coke rate of 390 kg/t hot metal (gas utilization ca. 45 %, indirect reduction ca. 75 %). At very high injection rates, the blast is enriched with oxygen or oxygen is injected separately to increase gasification efficiency of the injected reductants. The best figure achieved up to now using oil was an oil injection rate of 155 kg/t hot metal at a coke rate of 283 kg/t hot metal (total reductant rate 438 kg/t hot metal) at blast furnace A of Hüttenwerke Krupp Mannesmann [123], (Fig. 83). A replacement ratio of coke by oil of more than 1.1 was achieved. Figure 84 shows the replacement of coke by injected coal at three European blast furnaces using different coal grades [124]. The total reductant rate achieved at coal rates of up to 200 kg/t hot metal was in the range of 480 to 500 kg/t hot metal.

The highest coal injection rate at a large blast furnace producing basic hot metal was 266 kg/t hot metal achieved at Fukuyama no. 3 blast furnace of NKK Corporation in June 1998 at a coke rate of 289 kg/t hot metal (total reductant rate 555 kg/t hot metal) [125].

In Figure 82 there is a slight break in the line at K_2 which indicates that the quantity of reducing gas is larger as a result of the oil addition than in the case of carbon gasification alone.

Auxiliary injectants can replace coke only up to a certain limit. In the raceway the hydrocarbons are only gasfied to carbon monoxide and hydrogen, liberating less heat than coke per equal quantity of oxygen.

Natural gas

$$CH_4 + (1/2)\,O_2 \rightarrow CO + 2\,H_2 \quad \Delta H_{298} = -36\,\mathrm{kJ/mol} \tag{163}$$

Oil

$$-CH_2- + (1/2)\,O_2 \rightarrow CO + H_2 \quad \Delta H_{298} = -90\,\mathrm{kJ/mol} \tag{164}$$

Coke

$$C + (1/2)\,O_2 \rightarrow CO \quad \Delta H_{298} = -111\,\mathrm{kJ/mol} \tag{165}$$

The higher H_2 content of auxiliary injectants compared to coke leads to a lowering of the flame temperature as hydrogen supplies gas, but contributes no energy to the heating. The cooling effect is therefore lowest for coal and highest for gas (Fig. 85) [126]. An undue lowering of the flame temperature and an incomplete gasification of the auxiliary reductants have a negative effect on the operation of the blast furnace, on

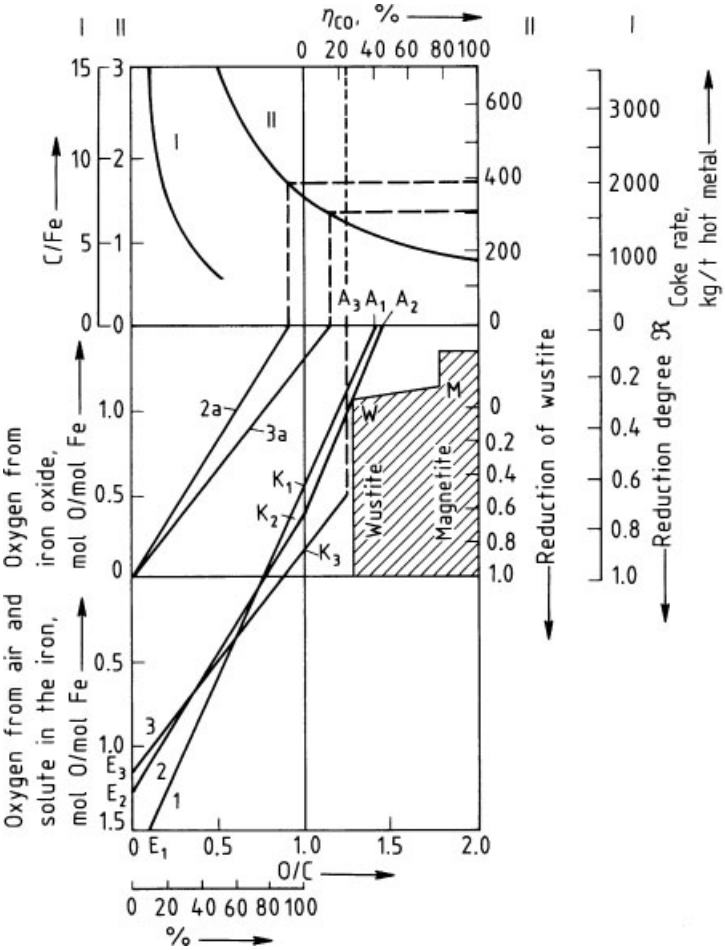

Figure 82. Effect of measures to decrease the coke ratio and to increase the productivity of the blast furnace estimated by the Rist diagram [111]
Line 1: blast furnace operation (standard); Line 2; decrease in the coke ratio through injection of inexpensive energy (best values achieved with Japanese furnaces); Line 3: operation with prereduced ore

total reductant rate of the blast furnace, and consequently on the replacement ratio of coke by injection materials. Thus a given quantity of heating coke can only be replaced by a larger quantity of hydrocarbons if the same amount of heat has to be generated. In other words, the replacement ratio, defined by the ratio hydrocarbons : coke is > 1. However, twice to three times the amount of reducing gas including hydrogen is produced per mole of combustion air (see Eqs. 118 and 119) compared to the combustion of coke (see Eq. 120).

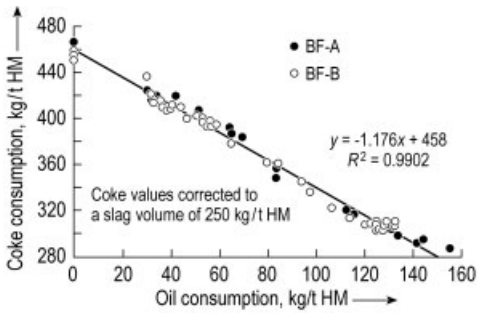

Figure 83. Ratio of coke to oil consumption in Hüttenwerke Krupp Mannesmann blast furnaces A and B

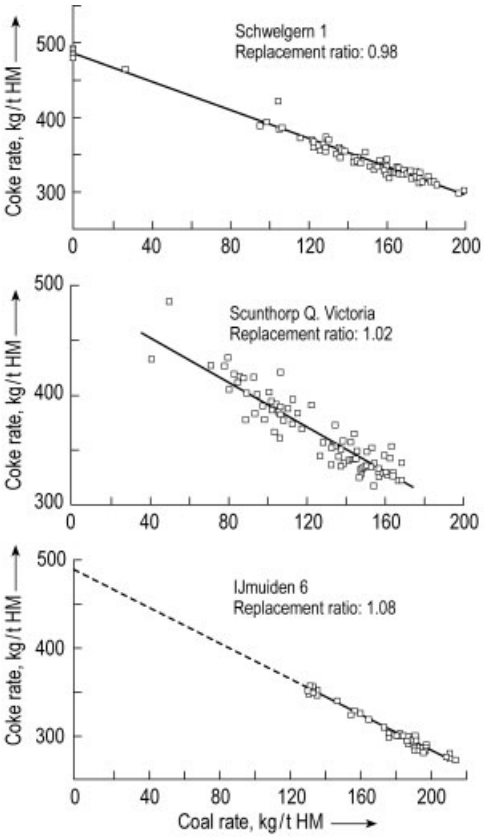

Figure 84. Achieved coke : coal replacement ratios

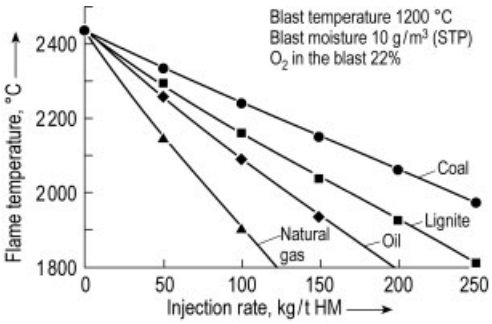

Figure 85. Change in flame temperature through injection of auxiliary fuels

The addition of hydrocarbons increases the degree of indirect reduction because the reduction rate of hydrogen is higher than that of carbon monoxide. Thus, the degree of direct reduction and its enthalpy requirement decrease. Consequently, the overall effect of hydrocarbon injection on the replacement ratio increases, thereby giving an overall hydrocarbon : coke replacement ratio <1. The enthalpy deficiency which suppresses further increase of indirect reduction becomes increasingly noticeable whenever more coke is replaced by increasingly larger amounts of hydrocarbons. The advantage of the lower reductant cost therefore becomes gradually less. However, a more serious problem is the decrease in the furnace productivity. This must be counteracted by process measures such as reducing the amount of steam, increasing the blast temperature and enrichment of the blast with oxygen or injection of oxygen separately to guarantee a flame temperature of approximately 2200 °C even at very high injection rates.

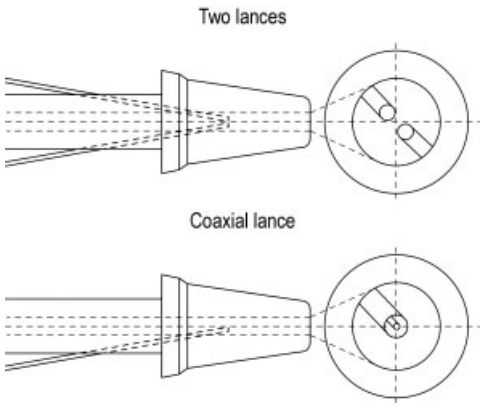

Figure 86. Injection lances at the Schwelgern blast furnace no. 1 of Thyssen Krupp Stahl AG

Figure 86 shows two lance designs which are in use for injection [127]. Double lances, i.e., two single lances mounted symmetrically, cause the oil or coal jets, which are supplied by the lances, to meet violently and therefore also lead to an improved mixing with the oxygen content of the blast. In coaxial lances, coal is injected through the inner and oxygen through the outer tube. An increase in the amount of oxygen added to improve the oxygen to carbon ratio may be achieved by enlarging the ring slot. A comparative evaluation of single, double and coaxial lances was carried out by temperature measurements and analyses of gas in the tuyere at Thyssen Krupp Stahl AG blast furnace Schwelgern no. 1 (Fig. 87). This evaluation showed that a markedly improved gasification behavior of

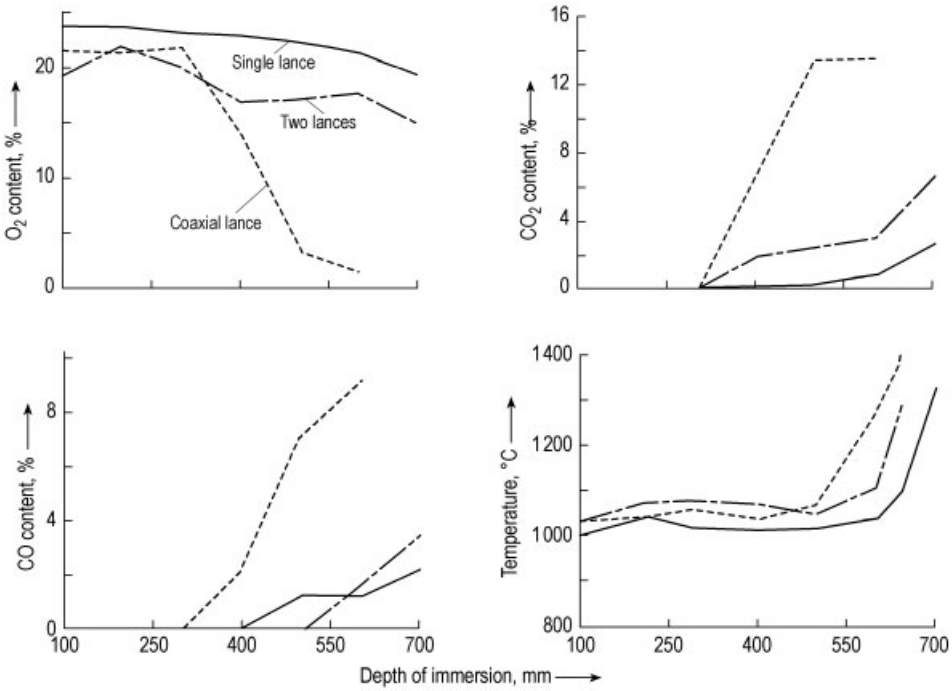

Figure 87. Distribution of temperature and composition of gas in the tuyeres of blast furnace Schwelgern no. 1, depending on lance design

the coal is observed when coaxial lances are used. In principle, coaxial lances conform to parallel flow burners, with the disadvantage of an incomplete mixing of the gasification media and large gasification lengths. For this reason an optimized lance design must be used so that the impulse of the oxygen and/or pulverized coal is deliberately utilized. Consequently the mixing of the gasification media is improved, the ignition time is reduced and the gasification increased.

The cost advantage of injecting reductants should take into account that using oil may result in an increase of sulfur input and using coal may result in an increase of chlorine and alkali input into the blast furnace, depending on the grade of purity of the injected reductant. Higher sulfur inputs are associated with higher desulfurization costs, higher alkali inputs are associated with blast furnace disturbances and higher chlorine inputs can damage the wet gas cleaning system. Pulverized coal is more difficult to inject into the blast furnace than oil because it is in the solid not in the fluid state and it contains ash. However, oil or natural gas injection is also conducted depending on local energy conditions all over the world.

Line 3 in Figure 80 shows a break which represents the operation of *injecting a reducing gas* into the stack to promote indirect reduction. This leads to the heating and reducing coke requirement of 390 kg/t (line 3 a). However, inherent to the low coke rate is a low gas utilization η_{CO} of only 30 %. The break in line 3 at K_3 is the result of an increase in the quantity of the gas at the point of injection. This technology has not been accepted for widespread application because of the poor gas utilization and the large amount of top gas. However, the reducing gas injection in the stack region is an important technology for the operation of the blast furnace with pure oxygen [128].

Line 3 in Figure 82 represents a blast furnace operation using *prereduced* ore with a metallization degree of 50 %. This procedure leads to a decrease in the coke rate to 310 kg/t. In this case, gas utilization is only ca. 30 %. Therefore, such an operation is only recommended when the blast furnace capacity is insufficient to satisfy a sudden increase in hot metal demand. The

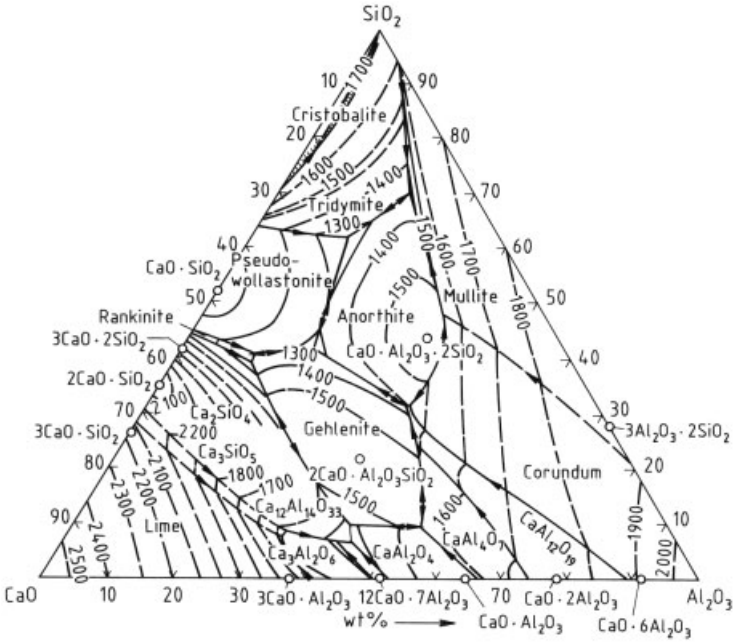

Figure 88. Phase diagram of the system CaO – Al_2O_3 – SiO_2 [129]

productivity of the blast furnace increases significantly by using prereduced material although it is more expensive.

Slag Practices. The slag must meet the following requirements:

1) It should absorb all unreduced non-volatile components of the burden and remove them from the blast furnace.
2) It should be a liquid of low viscosity.
3) It should be able to absorb the sulfur primarily contained in the reductants and it should contain as little iron oxide as possible to increase the yield of hot metal.
4) The slag volume should be as low as possible without impairing the desulfurization.
5) The temperature range where the burden components become cohesive should be narrow to ensure better permeability of the burden column.
6) Finally, the slag should be converted to salable material.

These requirements are in part complementary and in part mutually exclusive. It is therefore necessary to state priorities.

About 95 % of the slag consists of SiO_2, CaO, MgO, and Al_2O_3. The requirement of *low viscosity* can be met by a variety of components in this quaternary system. Ignoring the presence of MgO, the phase diagram of the ternary system CaO – Al_2O_3– SiO_2 illustrated in Figure 88 shows a low melting temperature region which is parallel to the CaO – SiO_2 binary for low Al_2O_3 content. This region extends from high SiO_2 content to the saturation isotherm for 2 CaO · SiO_2 and then for essentially constant CaO content toward high Al_2O_3 content. The MgO content of the slag does not substantially affect the relative position of the low melting temperature region and only affects the absolute values of the melting temperatures. The compositions of blast furnace slags as encountered under various operating conditions are shown in Figure 89.

As shown in Section 2.2.2, the desulfurization of the hot metal increases with slag basicity, i.e., with increasing CaO and/or MgO content. *Region 1* in Figure 89 can therefore be used only for processing low sulfur burden (e.g., in Brazil where very low sulfur content charcoal is often used as reductant). Because the gangue constituents usually form a low basicity slag, region 1 largely represents the slag composition without addition of fluxes. The furnace can be operated at a relatively low temperature because

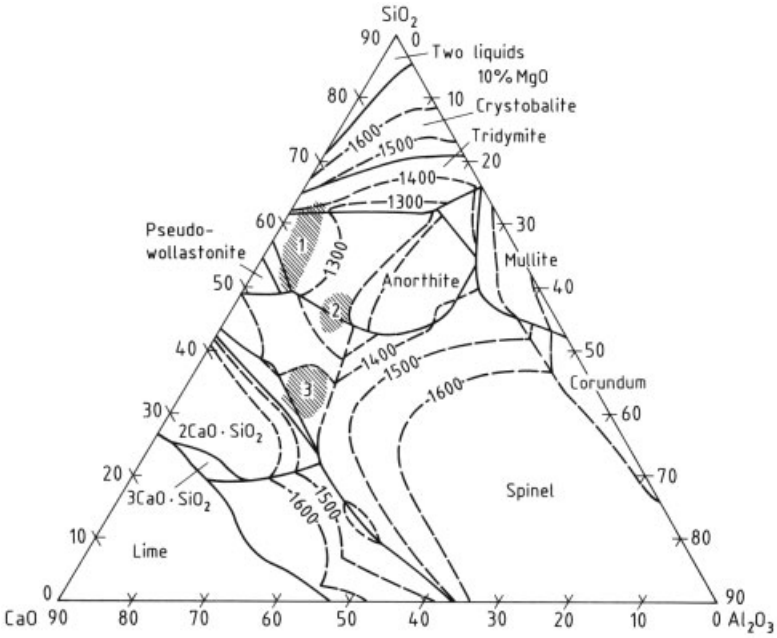

Figure 89. Composition of typical blast furnace slags on the CaO – Al_2O_3 – SiO_2 phase diagram (see Figure 88) [111] (for explanation of regions 1, 2, and 3 see text)

of the low melting points. *Region 2* is reached for low iron content burden with *acid gangue* constituents. This mode of operation prevails, for example, in Salzgitter and requires extensive desulfurization of the hot metal outside of the blast furnace. The attainment of a basicity that would result in adequate desulfurization within the furnace would require a large lime addition which leads to a high slag volume and consequently to higher coke rate. *Region 3* represents the worldwide preferred slag compositions for *large blast furnaces.* In this case, depending on the alumina content, dolomite must be added to satisfy the required MgO contents (see Table 24).

Table 24. Optimum composition of blast furnace slag in % [129]

Al_2O_3	CaO	MgO	SiO_2
5	43	16	36
10	44	14	32
15	44	12.5	28.5
20	45	11	24
25	48	8	19
30	56	5	9
35	54	4	7

Slags with higher basicities *B*, as shown in Table 24, would favor *optimum softening conditions.* The softening and melting range of the gangue constituents is ca. 80 – 130 °C for $B = 0.5$ and ca. 20 – 50 °C for $B = 2.0$. Because of the higher melting temperature of the highly basic slag and of the extra energy required due to the larger quantity of flux addition, the slag basicity is maintained at ca. 1.2.

2.5.4. Blast Furnace Productivity Criteria

The production rate *P* (t/d) of a blast furnace has been linked in various ways with the design data of the blast furnace and with metallurgical properties of the burden to give the specific furnace productivity. To date, no consensus has been reached to define a parameter which characterizes the furnace size and can be used in the calculation of the productivity. The characteristic parameters proposed up to now are (1) the hearth diameter, (2) the hearth bustle area that determines the depth of penetration of the blast in front of the tuyeres, (3) the square of the hearth diameter, (4) the hearth area, and (5) the working volume of the furnace. In Germany the working

volume is mostly used as a characteristic parameter for the furnace size, whereas in Japan the inner volume of the furnace is often used.

To compare various furnaces, the *coke consumption* with respect to the hearth diameter is often used. The coke consumption is a criterion (typical parameter) to express the productivity in the furnace which is operated in a given mode. The amount of coke consumed in the furnace is determined stoichiometrically by the constant blast volume blown. If the quantity of blast exceeds the limit, unfavorable phenomena such as hanging of the burden or variations in hot metal composition occur. The following relations have been proposed between hearth diameter and coke consumption:

$$\begin{aligned} D &= \sqrt{4Q/\pi R} \\ Q &= 30.763\times1.829\,(D-1.829-2L_{\mathrm{p}}) \\ Q &= \gamma D^2 \\ \log Q &= 2.59\times\log D+0.73 \\ Q &= 9.06D^{2.22} \end{aligned} \tag{166}$$

where γ is the productivity factor, D is the hearth diameter and R a constant. The production of hot metal P is related to coke consumption per day Q (t coke per day) by the coke rate K (t of coke consumed for producing 1 t of hot metal) and is indirectly evaluated from the quantity of blast blown or the coke consumed:

$$P = Q/K \tag{167}$$

Blast furnaces which are well operated have a specific hot metal production rate $p = P/A$ of 65 t m^{-2} d^{-1} at a coke rate of 470 kg/t HM, a slag volume of 300 kg/t HM, a superficial gas velocity at the top of 3 m/s and a blast temperature of 1100 °C. In Japan specific production rates of 80 t m^{-2} d^{-1} together with slag volume of 180 kg/t HM have been achieved. Productivities based on the hearth area should be interpreted carefully, because the hearth diameter increases during the course of a furnace campaign by the wear of the lining. Therefore, the productivity should be evaluated by using the actual hearth area.

In Figure 90 the coke consumption Q is shown as a function of the hearth diameter for a productivity factor γ between 10 and 25. The daily production rate of some blast furnaces, as a function of the hearth diameter is shown in Figure 91. Curves for the specific hot metal productivity p are also shown for comparison. In practice it has been observed that the specific iron productivity does not increase proportionally for extremely high coke rates because for hard driven furnaces the coke rate K in Equation (167) increases with decreasing gas utilization.

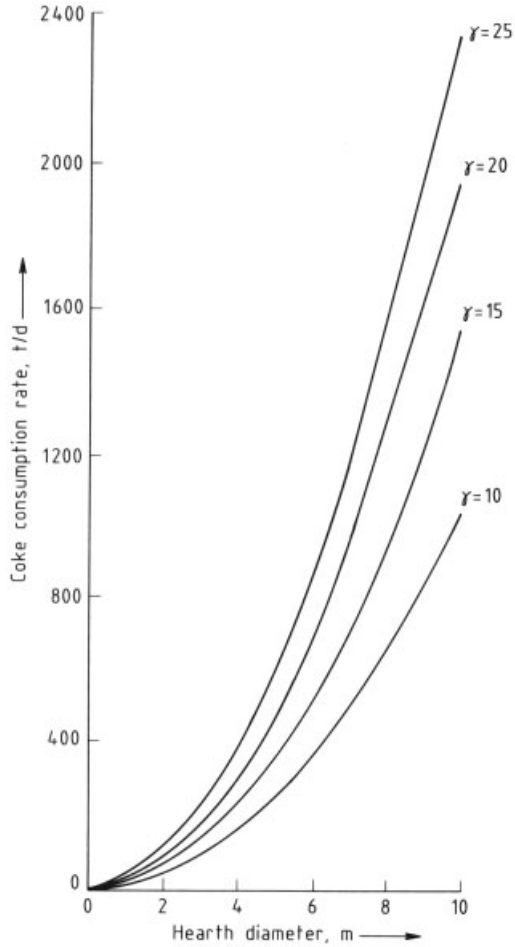

Figure 90. Change of coke consumption with the hearth diameter D and productivity factor γ [131]

Formulae for the *calculation of the productivity* based on empirical observation of operating furnaces have been derived with the aid of regression analysis [132]. These formulae are used to compare various furnaces or used for the design of new furnaces. A regression formula on productivity is given for the blast furnace which produces hot metal for steelmaking (Eq. 168):

$$P = A \frac{\left[\begin{array}{c} 924\,[1 + 4.5\,(O_2 - 2.20)/ \\ 100 + \{l + (m/100)\,(P_t - n)/100\}] \end{array}\right]}{\left[\begin{array}{c} 489 - 0.10\,(t_b - 1132) + 10\,(\mathrm{Ash} - 10.8) \\ + 0.2\,(\mathrm{SV} - 304) - 0.9\,[(\mathrm{SR} - \mathrm{SFPe}) \\ - 71.7] - 0.5\,(\mathrm{OPe} - 9.1) \end{array}\right]}$$

where P is the productivity (t d^{-1} m^{-3}); A a constant (0.85 – 1.15); O_2 the oxygen enrichment of the blast (%); P_t the top-gas pressure (gf/cm^2 = gram force per square centimeter = 98.0665 Pa); l, m, n are constants given in Table 25, t the blast temperature (°C); Ash the ash content in coke (%); SV the slag volume, (kg/t hot metal); SR the sinter ratio (%); SFPe the self fluxing pellet ratio (%), OPe the acid pellet ratio (%).

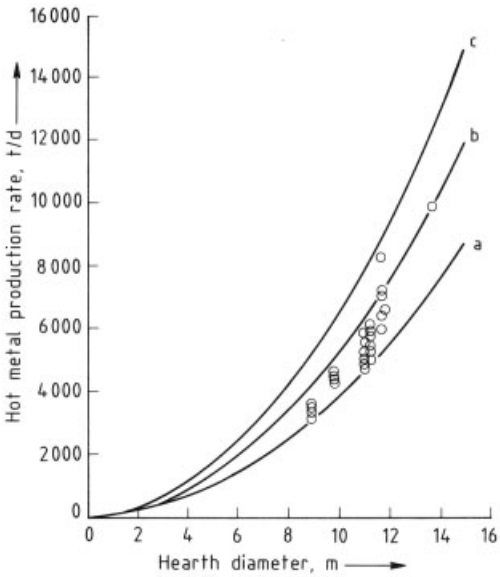

Figure 91. Relationship between hot metal production rate and the hearth diameter. The curves are for different specific hot metal production rates [132]
a) 50 tm^{-3} d^{-1}; b) 68 tm^{-3} d^{-1}; c) 85 tm^{-3} d^{-1}

The disadvantage of this formula is that it is entirely empirical. It does not predict productivity increases with technological improvements. Important operational data for some large blast furnaces are summarized in Table 26.

2.5.5. Use of Blast Furnace Products

Hot metal, slag and top gas are the products of the blast furnace.

Hot Metal. Most of the hot metal is processed to *steel* by means of the LD (Linz Donawitz) oxygen top-blowing processes. The hot metal used for the LD steelmaking process is known as *basic* hot metal. It is characterized by low phosphorus and low sulfur contents. *High phosphorus* hot metal (Thomas iron) can also be refined to steel by means of the oxygen steelmaking (LD/AC) process. Other types of hot metal are further processed in foundries to *cast iron* or are used as alloys in the steel shop or in the foundry because of their high alloy content (e.g., manganese, silicon). The composition of the common types of hot metal are listed in Table 27. The Bessemer (Thomas) and open hearth processes have been replaced by the oxygen top-blowing processes.

Slags. Depending on the conditions during solidification, the slags produced by the blast furnace are further processed to:

1) Air-cooled slag (solidification in slag pots or pouring pads),
2) Granulated slag (spraying of the liquid slag with water),
3) Pumice slag (foaming with water), or
4) Slag wool (blowing of the liquid slag with air or steam)

Granulated slag can be in the form of pebbles or can have a porous structure depending on the granulation process. The different structure affects the hydraulic-bonding properties, the grindability, and the strength. Granulated slag is used as a material for making blast furnace cement by mixing it with Portland cement after pulverization.

Pumice slag is used in porous form as an insulation and filler material in heavy concrete and as a replacement for other bulk additions.

Air-cooled slag is either ground, such as granulated slag, or crushed and screened to be used as ballast, aggregate, or for road construction.

Slag wool is used in the form of pads for thermal insulation.

The relative quantities of the various products made from blast furnace slag used in Japan and

Table 25. Values of constants in Equation (168) [133] (Courtesy of the Iron and Steel Institute of Japan)

	l	m	n
$P_t \leq 1000$	0	2.0	500
$1000 < P_t \leq 1500$	0.10	1.5	1000
$1500 < P_t \leq 2000$	0.175	1.2	1500
$2000 < P_t$	0.235	1.0	2000

Table 26. Characteristic data for modern large blast furnaces

	Japan		Germany	France
	No. 2 Oita NSC	No. 1 Kakogawa Kobelco	No. 2 Schwelgern Thyssen Krupp Stahl AG	No. 4 Dunkerque SOLLAC
Year of construction	1976	1970	1993	1973
Blown in date after last relining, month/year	12/88	1/88	10/93	11/87
Hearth diameter, m	14.9	14.1	14.9	14.0
Inner volume, m^3 (see Fig. 53	5245	4550	5513	4333
Number of tuyeres	40	40	42	40
Number of tapholes	4	4	4	4
Bell and hopper arrangement [a]	2 B + V	4 B	PW	PW
Top gas pressure, bar	3.5	3.0	3.8	3.1
Hot blast stoves (number-type) [b]	4/ECS	4/ECS	3/ECC	4/ECC
Blast blower	electr.	electr.	electr.	electr.
Gas cleaning [c]	DC/2VS	C BF	DC C VS	DC VS C
Production, t/d	11 800	9000	11 573	8640
Productivity on hearth area, t m^{-2} d^{-1}	67.7	57.7	66.4	56.1
Net burden weight, kg/tHM	–	1570	1596	1614
Slag volume, kg/tHM	307	260	262	290
Coke rate (dry), kg/tHM	331	315	341	309
Pulverized coal injection (dry), kg/tHM [d]	142	230	131	179
Blast temperature, °C	1196	1195	1170	1221
Blast volume (standard conditions), m^3	980	1010	924	914
Oxygen enrichment, %	2.6	2.9	3.7	3.5

[a] B = bell; V = valves; PW = Paul Wurth top without a bell.
[b] ECS = external combustion stack; ECC = external combustion chamber.
[c] DC = dust catcher; VS = venturi scrubber; BF = bag filter; C = cyclone.
[d] tHM = tonne of hot metal.

Table 27. Average composition of various kinds of hot metal (wt %) [130, 134]

	C	Si	Mn	P	S
Basic LD hot metal	4.4 – 4.7	0.4 – 0.7	0.1 – 0.3	0.06 – 0.1	<0.06
Thomas (Bessemer) iron	3.5 – 3.9	ca. 0.3	ca. 0.8	ca. 1.8	<0.055
Hematite hot metal	3.5 – 4.2	1.5 – 3.5	0.7 – 1.0	<0.12	<0.04
Foundry iron I	3.5 – 4.5	1.5 – 3.5	1.0	0.5 – 0.7	<0.04
Foundry iron III	3.5 – 4.5	1.5 – 3.5	1.0	0.7 – 1.0	<0.06
Foundry iron IV A	3.5 – 4.5	1.5 – 3.5	0.7	1.0 – 1.4	<0.06
Foundry iron IV B	3.5 – 4.5	1.5 – 3.5	0.7	1.4 – 2.0	<0.06
Specialty and malleable cast iron	3.4 – 4.3	0.3 – 4.0	0.2 – 1.0	0.05 – 0.10	<0.05
Specialty metal (high carbon)	4.0 – 4.8	<2.5	0.3 – 0.6	0.08 – 0.15	<0.05
Specialty metal for the production of nodular cast iron	2.8 – 4.3	<2.0	<0.2	<0.06	<0.02
Siegerland specialty metal	2.8 – 3.4	2 – 3.5	2 – 4	<0.1	<0.05
Spiegel iron	4.0 – 5.0	<1.0	6 – 30	0.1 – 0.15	<0.04
Ferro manganese (75 %)	6.0 – 7.0	<1.0	70 – 80	<0.25	<0.03
Ferro silicon	1.4 – 2.2	8 – 13	0.5 – 0.7	<0.15	<0.04
Charcoal hot metal	3.6 – 4.2	0.25 – 2.75	<1.0	ca. 0.03	ca. 0.015

Germany for the various application are listed in Tables 28 [135] and 29 [136].

Table 28. Slag production and use in Japan in 1995 [135]

	10^3 t	%
Slag production from blast furnace	22 678	
Slag use		
Road construction	4 741	20.9
Cement	14 788	65.2
Earthwork materials	1 544	6.8
Aggregate	1 605	7.1

Table 29. Slag production and use in Germany in 1997 [136]

	10^6 t	%
Production		
Basic hot metal	7.4	99.2
Other slags	0.06	0.8
Total	7.46	100.0
Use		
Raw slag	0.06	0.8
Air-cooled slag	3.36	45.04
Slag lime	0.06	0.8
Granulated slag for cement	3.46	46.38
Granulated slag for other uses	0.27	3.62
In-plant use	0.11	1.47
Intermediate storage	0.14	1.89
Total	7.46	100.0

In 1997 the blast furnace slag produced in Germany amounted to 7.46×10^6 t [136]. Blast furnace slags were mainly used as mineral material mixtures for road construction (ca. 44 %) and as granulated blast furnace slag – which has excellent hydraulic properties – for the production of blast furnace slag cement (ca. 46 %) (see Table 29). If the granulated blast furnace slag replaces lime, clay, or marl in the cement industry this results in a decrease in energy consumption in cement production and in a decrease in CO_2 emissions.

Top Gas. The top gas consists mainly of N_2, CO_2, CO, and H_2. The utilization degree of the top gas is usually ca. 50 % therefore it is a lean gas whose calorific value is between 2850 and 3560 kJ/m^3 (STP). This value is very low compared to those of coke oven gas and LD gas. Almost all the top gas is used in the iron and steel plant itself which has been extensively discussed (see Figs. 7 and 70).

2.5.6. Process Control

Fourty years ago, the blast furnace process was operated completely by manual control. Relatively few measuring devices were available for process monitoring purposes. Since the 1950s the productivity of a typical furnace has increased by ca. 20 times. This improvement in operation increased the importance of the process control. Modern blast furnaces are therefore equipped with a sophisticated system of measuring and controlling devices, and are directly linked to the process computer (Fig. 92). The latter is often incorporated into the total control system of the whole plant via a larger business computer.

The computer control of the blast furnace process for the optimization of the quality and costs started ca. twenty years ago. Research in this area is still being carried out. Much information can be obtained about the process by measuring the in- and outflows of mass and energy and also about many phenomena taking place inside the furnace. However, the basic problem is to convert these information into control variables to serve the control circuit. The algorithm to evaluate the control variables from the measured values are based on various mathematical models.

Static and Dynamic Models Based on Statistics. In these models the blast furnace is considered to be a black box. The effect of certain deliberate changes in the input variables (e.g., coke addition, pulverized coal injection, moisture content of the blast, back pressure at the top) on productivity, hot metal composition, top gas utilization, etc. is evaluated. With the aid of statistical models, this leads to regression expressions containing the desired algorithms for the determination of the control variables. These statistical blast furnace models that have already been applied successfully do not, in principle, require any knowledge concerning the phenomena occurring inside the blast furnace. An example of such a regression expression is the productivity formula (see Eq. 168) which should only be used for control purposes based on the values of a particular furnace and not based on values collected from various furnace operations. Although it is relatively simple to derive such an expression representing a blast furnace model from operational data, its application is limited.

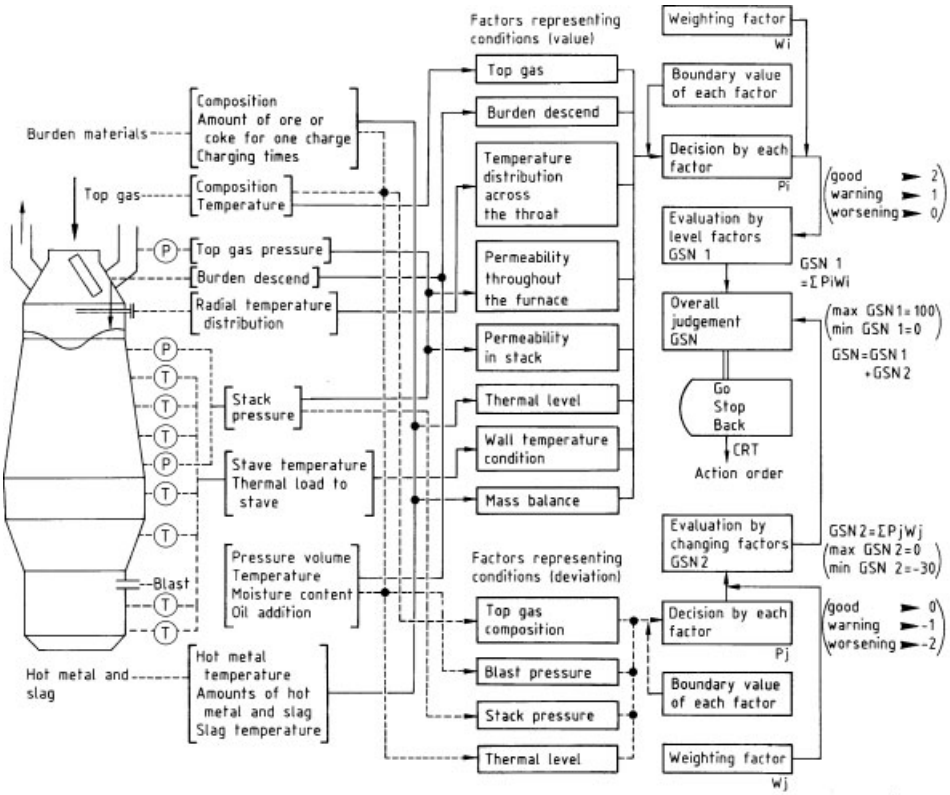

Figure 92. A control system of blast furnace developed by Kawasaki Steel Corp. [137] (Courtesy of The Iron and Steel Institute of Japan)

If such a model is applied to new operating conditions, then its predictions deviate from reality, the more the new operating conditions deviate from those described by the statistical data. Optimization of the operating conditions becomes stringently conservative when applying statistical models.

Static models based on statistics allow predictions concerning the steady state of the furnace only. *Dynamic models* based on statistics allow predictions concerning the behavior during a transient when passing from one steady state condition to another.

Models Based on Overall Balances. Models derived by taking overall balances represent the internal phenomena of the blast furnace on the basis of two-stage heat and mass balances. The procedure is principally similar to that described in Section 2.5.2 in which the application of the Rist diagram and the heat and mass transfer phenomena (Table 20) were discussed. However, the models are more exact and detailed. A disadvantage of these blast furnace models is that they are unable to predict time-dependent phenomena in the furnace, although data concerning phenomena occurring in the blast furnace are evaluated. For the evaluation of the control variables it is therefore necessary to make assumptions concerning the *thermal equilibrium* between gas and solids in the thermal reserve zone and the *chemical equilibrium* between gaseous components (CO and CO_2) and solid component (iron and wustite), in the chemical reserve zone. The assumption corresponds to point W in the Rist diagram.

In the overall balance model, a direct relationship exists between input and output variables. This is a definite difference to the statistical model. The overall balance model has only one apparent degree of freedom which is the degree of indirect reduction. The existence of this

apparent degree of freedom is caused by a lack of knowledge and will disappear when the internal phenomena of the blast furnace can be measured exactly and incorporated in the model. Under the fixed technological boundary condition the function of the blast furnace is *determined* by these phenomena in combination with certain pretreatment of ore and coke as well as a given hot metal composition.

Kinetic-Dynamic Model. The kinetic-dynamic model in its most extensive form is based on knowledge of all phenomena that constitute the entire blast furnace process. The phenomena include thermodynamic equilibrium and the kinetic relationships that govern the momentum, heat and mass transfer. The model is also based on the knowledge of the behavior of burden materials (e.g., size distribution at the top of the furnace, descending movement, softening and melting), coke gasification, and gas permeability. The model can simulate the behavior of a blast furnace for any given situation with a large computer. All the data required for the model can be determined in the laboratory and essentially need not be corrected with data based on actual operating conditions. The model thus constructed has proved to be advantageous for the study of new processing conditions and the predictions of the operating results without performing costly experiments with the blast furnace. Its disadvantage is the high expenditure required to gain complete information on all the important phenomena occurring in the furnace. For routine blast furnace control, the model must be simplified to reduce computer loads.

The most important kinetic phenomena for constructing the mathematical model by taking differential balances around a control volume are the chemical reactions (e.g., reduction of iron oxide, coke gasification, water – gas shift reaction, and reduction of silica). The reduction rate of *iron oxide* by carbon monoxide or hydrogen can be expressed in terms of the three-interface unreacted-core shrinking model which is represented in Figure 93. Rate and equilibrium constants needed for the rate equation are given in reference [138]. For *silica reduction,* a mechanism via the intermediate formation of SiO gas has been proposed [139]. A rate equation has been derived for the transfer of silicon from the SiO gas to the hot metal in accordance with the mechanism as described by Equation (169).

$$\frac{\mathrm{d}\,[\%\mathrm{Si}]}{\mathrm{d}t} = k_{\mathrm{c}}\,(A/M)\,P_{\mathrm{SiO}} \tag{169}$$

where k_{c} designates the reaction rate constant, A/M the specific area and P_{SiO} the partial pressure of SiO.

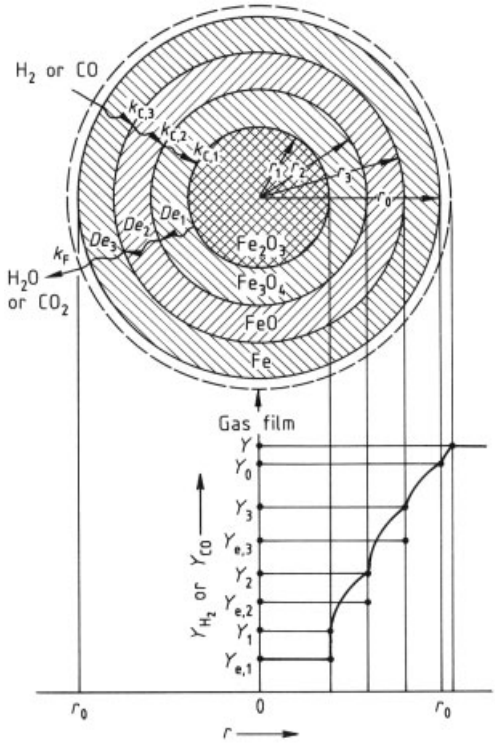

Figure 93. Three-interface unreacted core model [138] (Courtesy of The Iron and Steel Institute of Japan)
De_i = intraparticle diffusivity in i-th phase, k_{ci} = rate constant of i-th chemical reaction; k_{F} = mass transfer coefficient through gas film; r = radial coordinate; Y = molar fraction of gas

This mechanism has been found to give excellent simulation results when it is incorporated into the mathematical model.

The *heat transfer* between the different phases in a packed bed is usually expressed in the form of convective heat transfer. The estimation of the heat transfer coefficient is mainly based on the Ranz Equation [140]:

$$N\,u = 2.0 + 0.6(9\,Re_{\mathrm{P}})^{1/2}(Pr)^{1/3} \tag{170}$$

where *Nu, Re_{p}*, and *Pr* designate Nusselt, particle Reynolds, and Prandtl numbers.

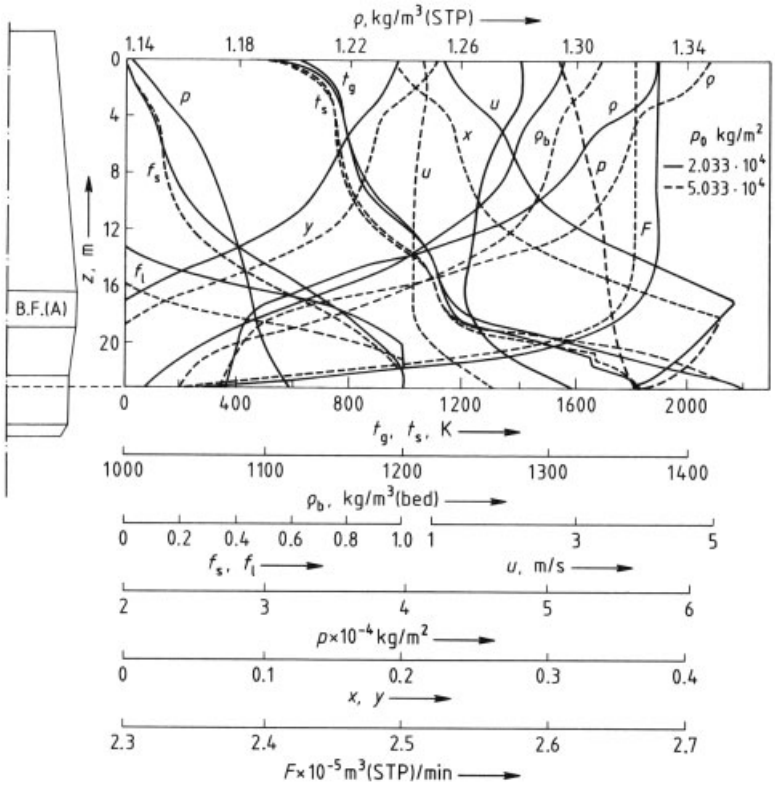

Figure 94. Longitudinal distributions of process variables in blast furnace (A) for the case of high-pressure operations [141] (Courtesy of The Iron and Steel Institute of Japan)
t_g, t_s= temperatures of gas and solid; f_s= fractional reduction of iron oxide; f_1= fractional decomposition of limestone; u = superficial gas velocity; p = gas pressure; p_0= top gas pressure; x, y = molar fraction of CO and CO_2; F = volume flow rate of gas; ϱ = gas density; ϱ_b= bulk density of burden bed

However, this equation sometimes gives an inappropriate value. The rate of heat exchange in a packed bed is currently being studied. *Flow maldistribution* of gas in a packed bed can be estimated by using the Ergun equation, a *potential flow mechanism* is often applied to the solid flow [141].

On the basis of the principal concepts mentioned above, *one-dimensional* [142], or *two-dimensional* [141, 143, 144] mathematical models were derived. On the other hand, dynamic changes in the internal situation of a blast furnace were also simulated by an *unsteady-state* mathematical model [145, 146].

Even if a supercomputer is available for the simulation computation of a whole blast furnace, it is still quite difficult to obtain detailed results which include all the phenomena occurring in a blast furnace. Mathematical models presenting some specified phenomena in a blast furnace are effective when detailed information is requested. Such models are developed for analyzing burden distribution at the top of a blast furnace [147], gas flow and heat transfer in the cohesive zone [148], raceway phenomena [119], fluid flow in the hearth [149] and so on.

Some examples of the simulation results are shown in Figures 94, 95 and 96. Figure 94 depicts a one-dimensional distribution of process variables such as temperature, pressure, concentration, and fractional reduction for two different top-gas pressure.

Temperature and gas-concentration distributions reveal the existence of thermal and chemical inactive zones and characteristic distribution of CO concentration which shows maximum value at the position where the CO_2 concentration is zero. The simulation results agree

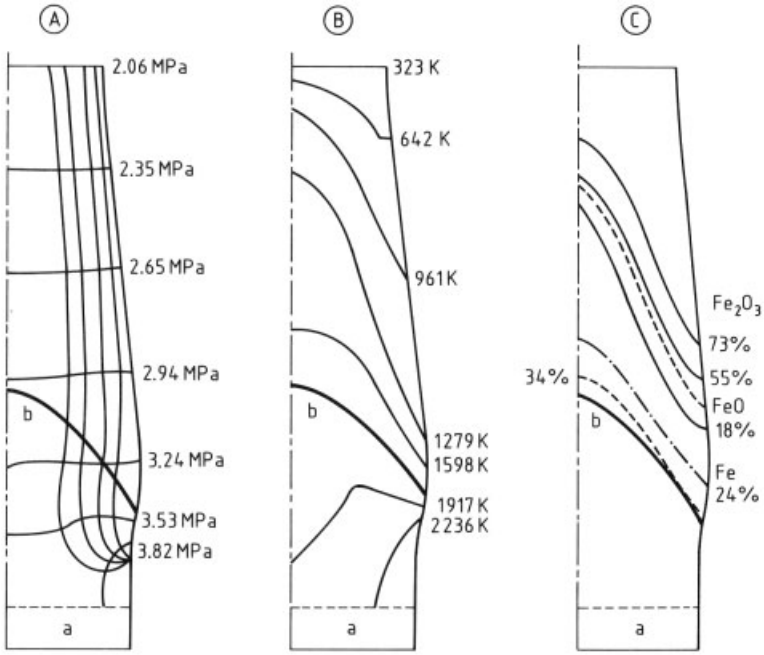

Figure 95. Two-dimensional simulation results by finite difference method [143] (Courtesy of the Iron and Steel Institute of Japan)
A) Streamlines and isobars; B) Isotherms of solid; C) Reduction of ore
a) Liquid; b) Fusion zone

well with the principal behaviors of the process variables previously shown in Figures 55 and 70.

Figure 95 shows the streamline of gas, as well as the radial distribution of pressure, temperature and reduction degree, which show a considerable nonuniformity. One of the important problems is how to control or utilize the nonuniformity to improve productivity.

Figure 96 shows computed radial profiles of pressure, gas velocity, and temperatures of gas and solids. In this computation, the layer-by-layer structure of the packed beds and the radial particle size distribution were considered, which markedly affect the gas permeability and then gas flow and temperature distributions.

Complete measurement of the internal state of a blast furnace is very difficult because of its huge size. Simulation technology in combination with operating and some measured data is very important to understand the behavior of the blast furnace.

Application of the Models. The purpose of the blast furnace control during continuous operation is the increase in productivity by stabilizing the furnace operation and the production of high quality pig iron of appropriate temperature and composition. The control variables are (1) ore : coke ratio and burden distribution at the top of the furnace, (2) flow rate and temperature of blast, water vapor addition, and oxygen injection, pulverized coal, oil and fine ore injection.

Originally, attention was paid to the *thermal condition* in the *lower part* of the blast furnace for the stable operation. The so-called *Wu model* was proposed by the Institute de Recherches Sidérurgiques (IRSID) in France. In the model, the heat input into the lower part of the furnace is calculated from blast volume and composition, tuyere injection, and the off-gas analysis. The amount of heat *Wu* is that required for the melting of iron and slag, the reduction of the alloying elements, and compensation for heat loss. The value of *Wu* can be calculated by subtracting the heat consumed by direct reduction of the iron oxides from the total heat supplied to the lower part of the furnace. By maintaining *Wu* at a constant set value, pig iron of a constant composition can be produced.

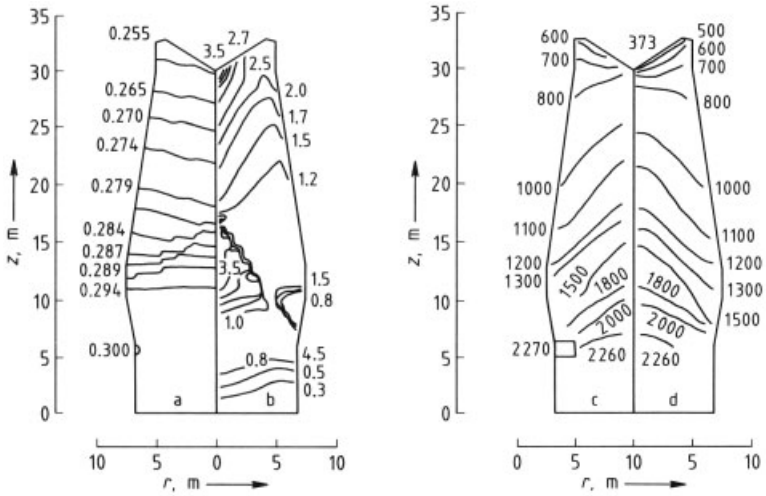

Figure 96. Two-dimensional simulation result by finite element method [141] (Courtesy of The Iron and Steel Institute of Japan)
a) Isobars (P in MPa); b) Contour lines of average mass velocity of gas (kg m^{-2} s^{-1}); c) Isotherms of gas (T_g in K); d) Isotherms of solid (T_s in K)

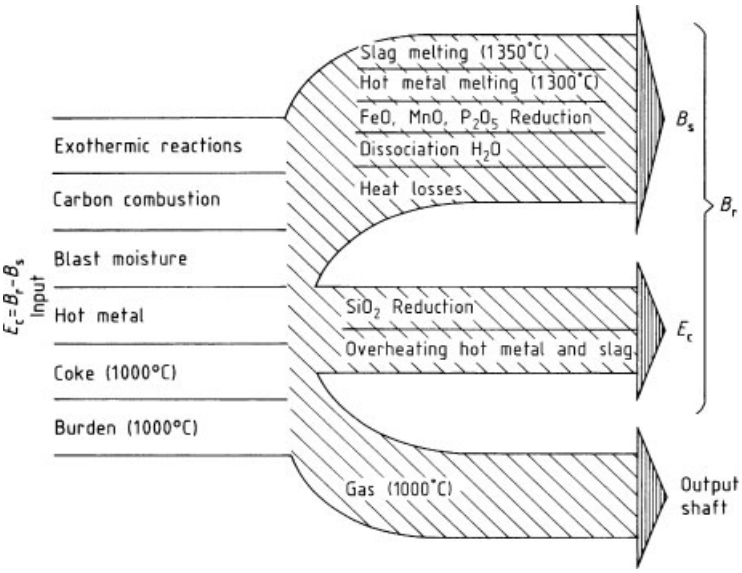

Figure 97. Heat balance for calculating value of E_c-heat required for the reduction of silicon and for the overheating of iron and slag [150] (Courtesy of the Institution of Mining and Metallurgy)

At the Centre du Recherches Métallurgiques (CRM) in Belgium, a somewhat different model of the same basic type was developed for automatic regulation of the thermal condition of the furnace. In this model, the heat input per tonne of hot metal into the high-temperature zone of the furnace was calculated from the temperature and composition of the top gas, the blast volume, blast temperature, and blast composition. The value obtained is called the heat requirement. The term E_c was defined as the difference between the actual (B_r) and the standard heat requirements (B_s). The latter quantity represents the amount of heat needed to produce a pig iron at 1300 °C containing 0 % Si.

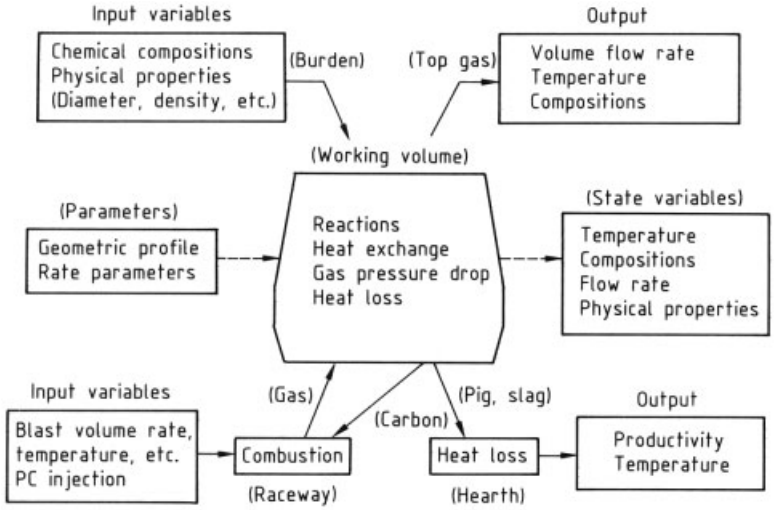

Figure 98. Schematic diagram of a dynamic simulation model of blast furnace [146] (Courtesy of The Iron and Steel Institute of Japan)

The value of E_c corresponds to the heat required for silica reduction and for overheating slag and metal. The heat balance to calculate E_c is shown in Figure 97.

A *one-dimensional dynamic model* was applied to the *unsteady-state* blast furnace operation such as the shut down, the blow-in, and the blow-out together with a *two-dimensional gas flow* model. A schematic diagram is shown in Figure 98. Figure 99 shows the simulation results for the blow-in operation of the blast furnace according to the scheduled increase in blast volume and temperature as shown in Figure 100. The initial condition given to the mathematical model was obtained from the actual filling condition of coke and burden before the start of the blow-in. The results confirm the effectiveness of the model application to the planning of the blow-in operation of the blast furnace. Another successful example for the application of a dynamic model to the actual operation has been reported for the blow-out of the blast furnace. The content of the blast furnace was blown out with the decrease of stock level from the top to the upper bosh under the schedule predicted by the mathematical model. Actual changes of the operation conditions are shown in Figure 101 in comparison with the scheduled ones. The actual results are in reasonable agreement with the schedule (see Table 30). Some improvements to the mathematical model are necessary for obtaining a satisfactory agreement. However, principally such a kinetic-dynamic model is found to be effective not only for steady-state but also for unsteady-state operations.

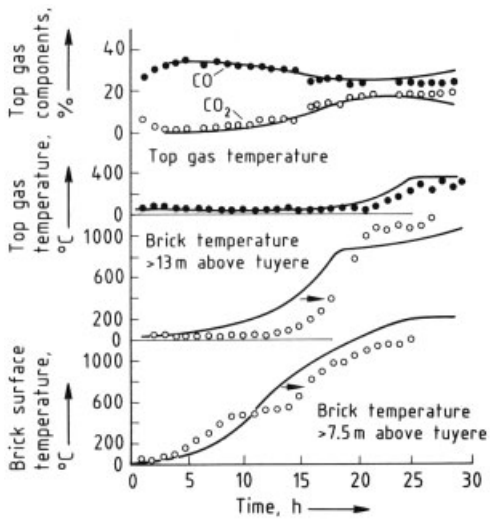

Figure 99. Comparison between measured and simulated changes of process variables in blow-in operation of Kokura no. 2 blast furnace, Sumitomo Metal Industries, Ltd [146] (Courtesy of The Iron and Steel Institute of Japan)
—— Predicted; • o Measured

Table 30. Results of the empty blow-out operation of the Kokura no. 2 blast furnace in Sumitomo Metal Industries [151] (Courtesy of The Iron and Steel Institute of Japan)

	Actual results	Schedule	Difference, %
Blow-out operation time, h	23.3	24.5	−5
Accumulated amount of blast volume, $m^3 \times 10^3$ (STP)	1380	1675	−18
Accumulated amount of water spray, t	694	716	−3
Hot metal produced during blow-out, t	880	998	−12
Amount of total coke consumption, t	550	690	−20
Coke consumption per blast volume, $t/m^3 \times 10^3$ (STP)	0.40	0.41	−2

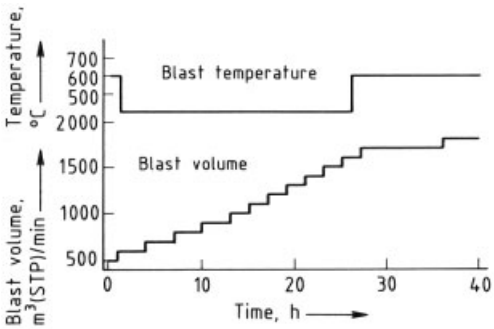

Figure 100. Schedule of the blow-in operation for the Kokura no. 2 blast furnace based on a dynamic model [146] (Courtesy of The Iron and Steel Institute of Japan)

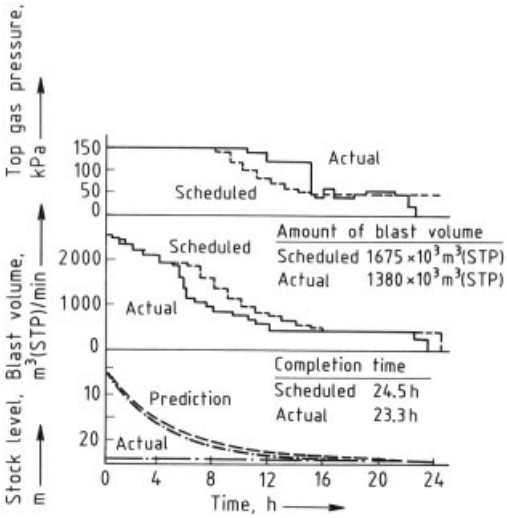

Figure 101. Results of the empty blow-out operation of the Kokura no. 2 blast furnace [151] (Courtesy of The Iron and Steel Institute of Japan)

The most precise mathematical model of a blast furnace to date is the *two-dimensional steady-state* kinetic model developed by Sugiyama et al. [152] (Nippon Steel Corp.). It includes models for operational design, burden distribution, gas, solid and liquid flow, for the reaction rate of seven chemical reactions, for heat transfer and a total combination model. When operational conditions and a blast furnace profile are given, this model provides (1) streamlines of gas, liquid and solids, (2) pressure distribution, (3) reduction degree of iron oxides, (4) gas, solid and liquid temperature, and (5) location of the cohesive zone.

An integrated process computer system in the entire ironmaking department is being developed for increasing flexibility and productivity in the ironmaking systems which consist essentially of coke ovens, sintering machines, hot stoves, and blast furnaces.

The three-layer hierarchical system shown in Figure 102 is used for integrated control. In the system, *microcomputers* can accept the operating data directly from the plant and can control it. A *process computer* which eventually controls the overall ironmaking system is installed to provide sufficient information to the operator for realizing effective operation. Finally, a central *business computer* is connected to the process computers for optimal control of the overall iron and steel production system. Chiba works of Kawasaki Steel Corp. constructed this three layers system for two sintering plants and achieved effective operational control.

Although great progress has been made in the computer control of blast furnaces by introducing mathematical models, control is still incomplete. Recently, artificial intelligence has been introduced into blast-furnace control. The principal strategy of this technology is based on operational data [153]. Therefore, the question how to accumulate the operating data is most important. Some trials have been made to couple the artificial intelligence (Expert system) with a mathematical model.

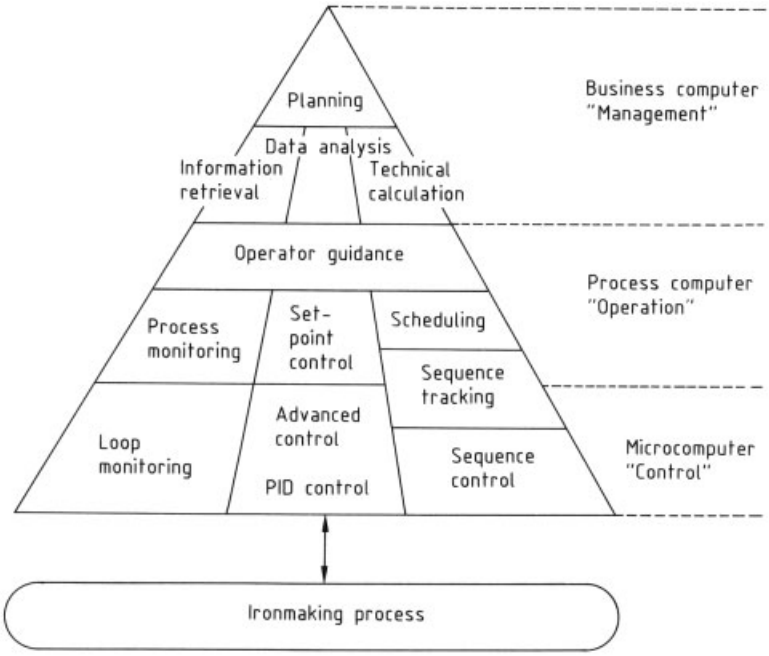

Figure 102. Functional hierarchy of the ironmaking information system [152] (Courtesy of The Iron and Steel Institute of Japan)

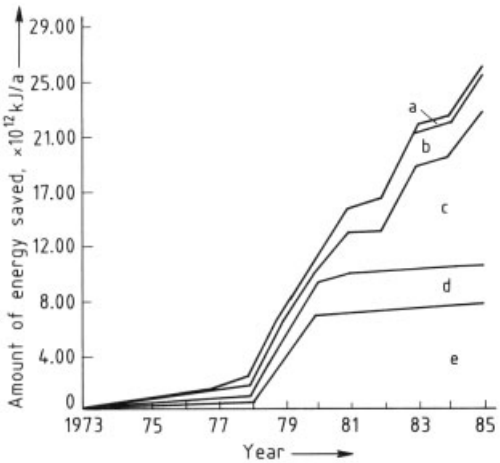

Figure 103. Chronological change in the amount of energy saved by main waste heat recovery process in Nippon Steel Corp. [154]. (Courtesy of The Japan Iron and Steel Federation)
a) Recovery of sensible heat from coke oven gas; b) Recovery of waste heat from sinter cooler; c) Coke dry quenching; d) Recovery of sensible heat from hot stove; e) Top gas pressure recovery turbine

Energy Recovery. After the oil crisis, energy cost increased drastically. Therefore in high-energy consuming industries, energy saving has been examined from various points of view. Energy saving and energy recovery are still conducted in each ironmaking process. In the blast furnace, a lot of energy is saved by using top-pressure recovery turbines. Figure 103 illustrates the development of energy savings in the ironmaking department of NSC between 1973 and 1985.

Energy recovery methods and energy network possibilities interconnecting the coking plant, sintering plant and blast furnace areas are illustrated in Figure 104 by the example of Thyssen Krupp Stahl AG [155]. Of decisive importance for the economic viability of investments aimed at energy recovery are the price of the primary energy to be replaced, the distance between supplier and consumer, the supply : demand ratio, and the extent of the necessary plant engineering.

The coke oven gas produced in the coking plant meets the requirement for gaseous fuel in the sintering plants, the rolling mill furnaces, in the coke plant, and partly of the power station – while blast furnace gas is partly used as an underfiring gas in the coking plant – for heating the blast furnace hot stoves and in the power plant.

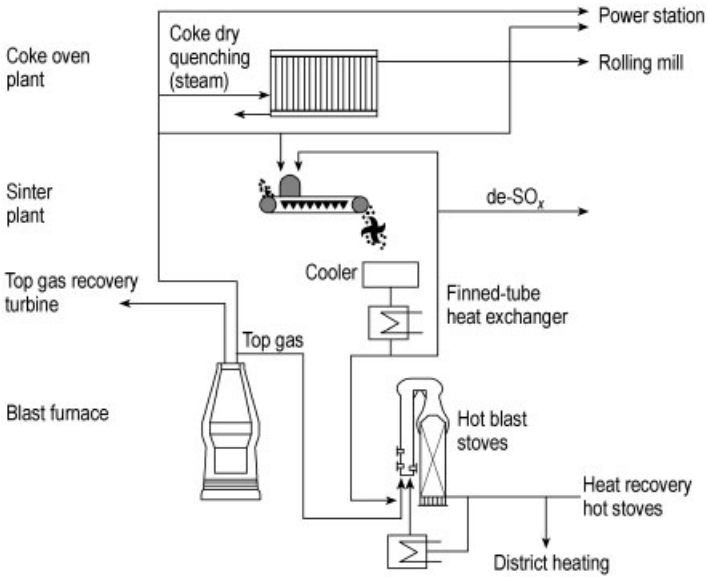

Figure 104. Example for energy consumption and recovery at coke plant, sinter plant, and blast furnace

At the coke dry quenching facility (if available) preheated feed water from the power plant is converted into high grade steam and this, in turn, is used in a power plant turbine to generate electricity. It has to be noted, that in Germany it is today uneconomical to operate a coke dry quenching installation because the credits earned from the recovered energy are insufficient to cover the coke dry quenching installation's operating and capital costs. The dry coke quenching facilities are built because statutory requirements in Germany demand a low-emission quenching process, even if it was found that the emissions of a modern wet quenching system approximate the emission levels of a coke dry quenching installation [156].

In sintering plants the sinter cooler offers the simplest and most abundant potential for heat recovery. The recovered heat is used in the form of hot combustion air for use in the ignition furnaces and/or for heat supply to a hood downstream of the ignition furnace of the sinter plant. Such recovery systems are used in many sinter plants, for example, in Japan and Europe. As shown in the example of Figure 104, part of the heat can be recovered from air by means of thermal oil circulating in a finned-tube heat exchanger between the blower and stack and the combustion media of the hot blast stoves of the blast furnace can thereby be preheated. The combustion media of the hot blast stoves' are also preheated by the hot stoves waste gas by means of thermal oil. The flue gas' waste heat from the hot blast stoves can also be supplied to a municipal district heating network.

A further significant potential for energy recovery is the use of top pressure energy recovery turbines to exploit the pressure difference of blast furnaces operating with elevated top pressure. The use of top pressure energy recovery turbines at blast furnaces is also gaining in importance because of the steady rise in electricity prices. By this energy recovery approximately one-third of electricity required for blast compression can be recovered.

2.5.7. Hot-Metal Desulfurization

The sulfur content of the hot metal produced depends on the sulfur content of the reductants, the slag basicity, the slag volume, the silicon content of the hot metal, and the alkali load of the blast furnace. The sulfur transfer mechanism in the blast furnace is shown schematically in Figure 105. After the hot metal is separated from slag, the former is desulfurized by gas – metal and slag – metal reactions. However, in the upper part of the cohesive zone, the sulfur compounds present in the gas phase are absorbed partly by solid iron ores.

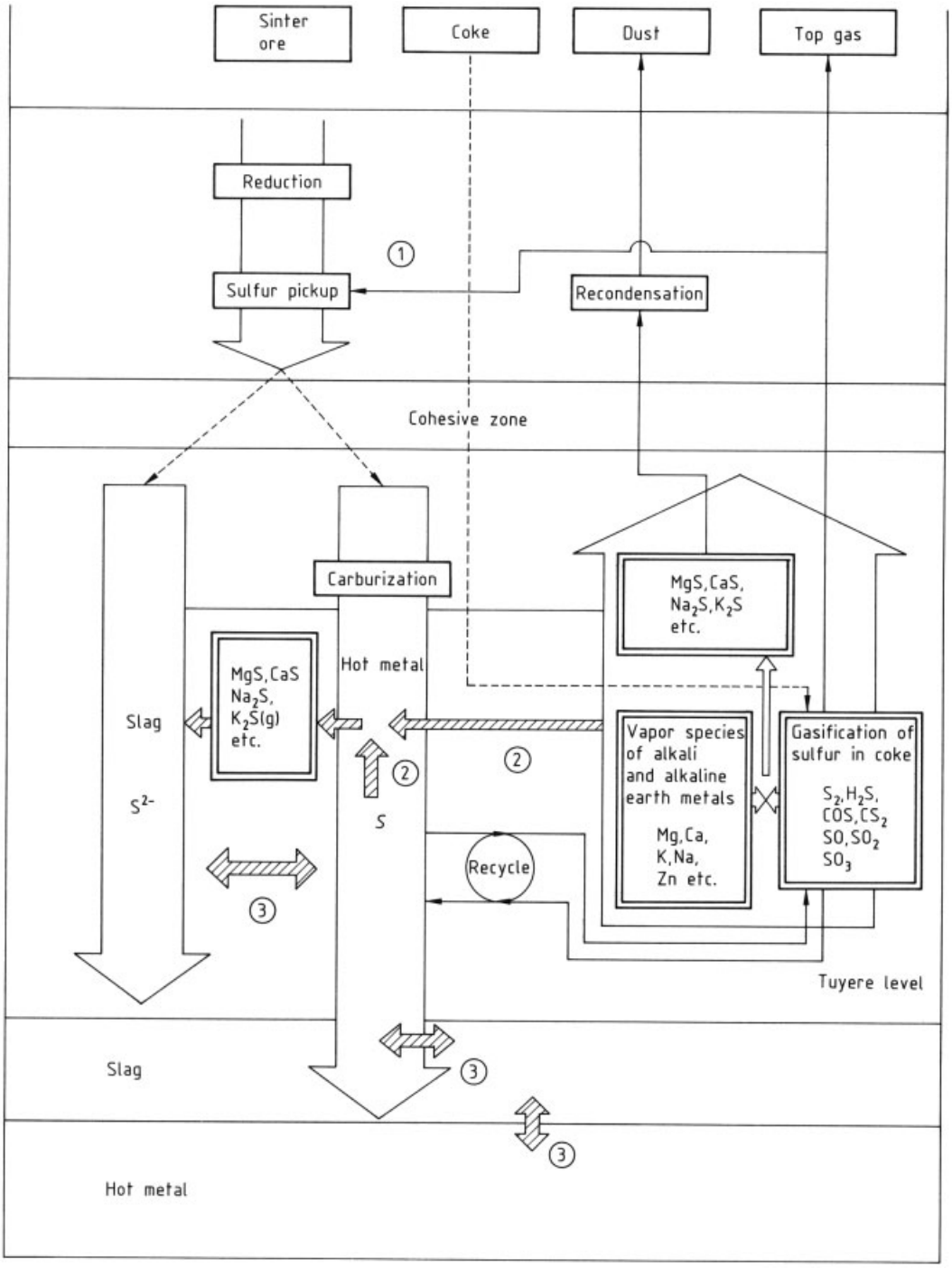

Figure 105. Schematic representation of sulfur transfer routes in the blast furnace [160] (Courtesy of The Iron and Steel Institute of Japan)

The steelmaking process requires a sulfur content of the hot metal of $< 0.025 - 0.035\,\%$, because desulfurization is difficult in the steelmaking process. Sulfur can be removed in the blast furnace, between the blast furnace and the steel shop, and in the steel shop. The cost is reported to be the lowest for desulfurization between the blast furnace and the steel shop [157].

Desulfurization in the *blast furnace* requires an increase in the limestone additions. This leads to the additional costs associated with the limestone itself and the increased coke rate because more heat is needed for melting the slag and for decomposing the limestone. In addition to this, the carbon dioxide generated from limestone must be compensated for. Limestone addition to the blast furnace can be avoided in certain limits if basicity and chemical composition of the sinter introduces all slag-forming components necessary for the burden. This saves coke consumption in the blast furnace because decompostion of limestone is transferred to the sinter plant.

Desulfurization in the *steel-producing vessel* requires a large slag volume even for highly basic slags because of the unfavorable sulfur distribution ratio under oxidizing conditions (see Section 2.2.2). Iron losses to the slag increase because of the large slag volume.

At present, it is advantageous to desulfurize at least a portion of the hot metal during *transportation to the steel shop.* After mixing with hot metal that is not desulfurized outside the blast furnace, an input sulfur content of $\leq 0.02\,\%$ can be delivered to the steel shop.

Desulfurizing Fluxes include *calcium compounds* which are in the solid state or half melting conditions at the hot metal temperature (e.g., CaO, CaC_2, $Ca(CN)_2$, and CaF_2), *alkaline compounds* in the liquid state (e.g., KOH, NaCl, and NaF), and *magnesium and its alloys* in the gaseous state. Their desulfurization ability is remarkably affected by treatment conditions such as (1) mixing, (2) atmosphere (oxidizing or reducing), (3) composition of hot metal, and (4) property and amount of blast furnace slag. Figure 106 shows the equilibrium sulfur content obtained by each desulfurizing flux. Actual desulfurization curves for the *bottom blowing* method with nitrogen are shown in Figure 107.

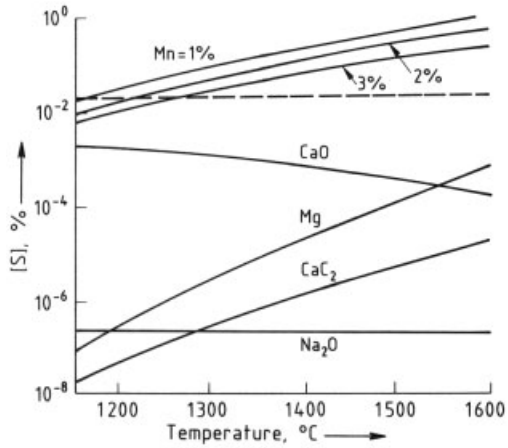

Figure 106. Temperature dependence of equilibrium sulfur content of carbon-saturated hot metal [158] (Courtesy of The Iron and Steel Institute of Japan)

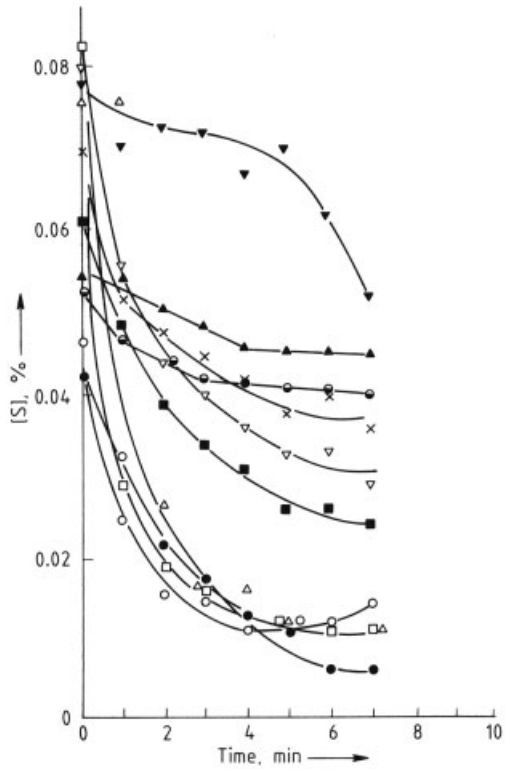

Figure 107. Changes of sulfur content with time in bottom-blowing ladle desulfurization process for variable fluxes (amount of flux addition 10 kg/t hm) [158] (Courtesy of the Iron and Steel Institute of Japan)
● CaC_2; ▲ CaO; ■ $Ca(N)_2$; □ Na_2CO_3; △ NaOH; □ KOH; x NaCl; ▽ NaF; ▼ Ca F_2; ⬢ Na_3AlF_6

According to these figures, desulfurization degree, method, cost, and workability should be considered in addition to desulfurizing ability when selecting of a desulfurizing flux.

The equation for the desulfurization reaction with *soda ash* [159] and of concurrent reactions are as follows (Eq. 171 – 173):

$$(Na_2CO_3) + [S] + [Si] \rightarrow (Na_2S) + (SiO_2) + CO \quad (171)$$

$$(Na_2CO_3) + (FeS) + [Si] \rightarrow (Na_2S) + (SiO_2) + [Fe] + CO \quad (172)$$

$$(Na_2CO_3) + [Si] \rightarrow (Na_2O{\cdot}SiO_2) + C \quad (173)$$

Furthermore, Na_2CO_3 reacts easily with SiO_2 from the refractories forming $2Na_2O \cdot SiO_2$, $Na_2O \cdot SiO_2$, and $Na_2O \cdot 2SiO_2$. The vaporization loss of Na_2CO_3 is ca. 6 % at 1250 °C. This value increases to 35 % at 1350 °C, therefore the effect of desulfurization decreases sharply at temperatures >1300 °C. The disadvantages of desulfurization with soda ash are its corrosive action to the vessel lining and the environmental problems related to discarding slags which contain high alkaline contents. Environmental problems arise from sodium oxide-containing fumes and from the reaction of water with the sodium sulfide contained in the slag which produces hydrogen sulfide and sodium hydroxide.

The desulfurization reactions of hot metal with *calcium carbide* are as follows:

$$\langle CaC_2\rangle + [S] \rightarrow (CaS) + 2\,[C] \quad (174)$$

$$\langle CaC_2\rangle + (1/2)\,\{O_2\} \rightarrow \langle CaO\rangle + 2\,[C] \quad (175)$$

$$\langle CaC_2\rangle + (1/2)\,[O] \rightarrow \langle CaO\rangle + 2\,[C] \quad (176)$$

$$C + 1/2\,O_2 \rightarrow CO \quad (177)$$

As calcium carbide is easily oxidized (see Eqs. 175 and 176) the desulfurization should take place under controlled atmosphere. The reaction product CaS which has a high melting temperature of 2450 °C forms the solid slag. The diffusion of sulfur in the hot metal to the boundary between hot metal and calcium carbide is the rate-determining step of the desulfurization reaction. In the sulfur-transfer mechanism, mixing strongly affects the degree of desulfurization. When calcium carbide particles are injected with an inert carrier gas, a calcium sulfide layer grows at the outer surface of the particle as the reaction proceeds. The reactivity of each particle decreases with increasing thickness of this layer. Thus, the efficiency of this flux is relatively low. Its advantage is that a very low residual sulfur content of the hot metal can be achieved irrespective of its initial sulfur content.

When *lime* is used, the following desulfurization reactions proceed:

$$4\,CaO + 2\,[S] + [Si] \rightarrow 2\,CaS + 2\,CaO{\cdot}SiO_2 \quad (178)$$

$$2\,CaO + 2\,[S] + [Si] \rightarrow 2\,CaS + SiO_2 \quad (179)$$

The reaction given by Equation (179) occurs at 1300 °C even if the silicon content of the hot metal is only 0.05 %. If the silicon in the hot metal is oxidized to silicon dioxide in this process, it reacts with lime to form $2CaO \cdot SiO_2$. This reaction decreases the amount of lime available for desulfurization. Therefore, oxidation of silica should be prevented by using the inert or reducing atmospheres or airtight vessels.

Magnesium is sometimes used as a desulfurizing flux because it has a very strong affinity for sulfur. Equations (180), (181), and (182) represent the desulfurization by magnesium and the oxidation of magnesium which occurs in the desulfurization process.

$$\{Mg\} + [S] \rightarrow MgS \quad (180)$$

$$\{Mg\} + [O] \rightarrow MgO \quad (181)$$

$$\{Mg\} + (1/2)\,O_2 \rightarrow MgO \quad (182)$$

The vaporization temperature of magnesium is relatively low (1107 °C). Therefore, magnesium evaporates upon contact with the hot metal and the desulfurization efficiency decreases. It is therefore necessary to control the vaporization (see below).

Desulfurization Processes. To produce low-sulfur or ultra-low sulfur steel, a number of desulfurization processes have been developed. Typical processes for desulfurization *outside* the blast furnace are listed in Table 31. In each process, acceleration of the desulfurization reaction is achieved by increasing the contact area between the desulfurizing fluxes and the hot metal. This is carried out by employing improved stir-

ring techniques. The other requirements for the processes are as follows: (1) inexpensive desulfurizing fluxes, (2) high reproducibility with regard to the degree of desulfurization, (3) small decrease of the hot metal temperature, (4) low iron loss to the slag, and (5) an easy and inexpensive disposal of the slag.

Table 31. Typical desulfurization processes outside the blast furnace [104] (Courtesy of The Iron and Steel Institute of Japan)

Method	Process
Soda-ash paving	soda-ash paving process
	reladling process
	ladle with siphon process
Shaking ladle	shaking ladle process
	DM converter process
	rotating drum process
Stirrer	Demag-Östberg process
	Rheinstahl process
	KR process
Injection	(ladle process)
	(torpedo-car process)
Gas bubbling	bottom blowing process
	top blowing process
Gas-lift mixing reactor	GMR process
Desulfurization with magnesium	injection process
	plunging bell process
Continuous desulfurization in the blast-furnace runner	turbulator process
	powder injection process
	paddle-type stirrer process
	electromagnetic stirrer process

Desulfurization in the *blast-furnace runner* is not widely accepted at present despite the favorable surface : volume ratio and the low degree of technical expenditure required. The reasons are the corrosive attack of the slag on the runner material, the hazardous exposure of the workers to alkaline fume, and the poor reproducibility with respect to the degree of desulfurization.

The soda-ash paving method is frequently used for desulfurization in the *transport ladle* because of its simplicity. The transport ladle is used to transport the hot metal from the blast furnace to the hot-metal mixer located in the steel shop. In this method, the desulfurizing flux is placed in the ladle prior to filling it with hot metal. Good mixing is achieved by formation of carbon dioxide and by the falling hot-metal stream. The slag is removed from the filled ladle. Figure 108 illustrates the relationship between the final degree of desulfurization and the amount of soda ash added. The final desulfurization degree increases with increasing soda-ash addition up to a value of 80 %, which is obtained at 8 kg "soda-ash addition" per tonne of hot metal. After that the desulfurization degree in most cases reaches a plateau level.

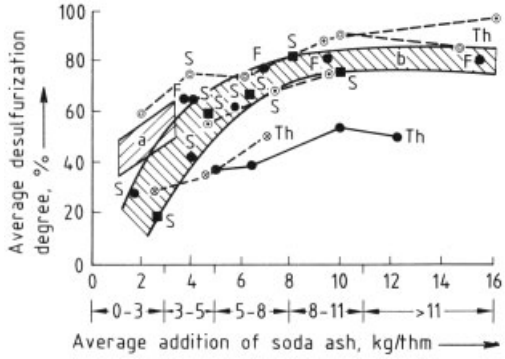

Figure 108. Effect of the amount of soda ash addition on the desulfurization of hot metal in various desulfurization processes [161] (Courtesy of Verlag Stahleisen mbH)
○ Blast-furnace runner; ■ Transportation ladle; ⊗ Mixer – ladle; ⊙ Mixer – ladle-ladle; ○○ Torpedo car ladle, ladle with siphon
S = pig iron for steelmaking; Th = Thomas pig iron; F = pig iron for foundry
a) Ladle with siphon; b) furnace-runner and transportation ladle

Efforts to improve the mass transfer of the sulfur by moving the vessel containing the hot metal involved trials with *rotating drums* (Kalling – Domnarvet process [162]) and *shaking* or *rotating ladles.* In the shaking ladle, the ladle axis rotates around a circular path at ca. 70 rpm. By reversing the direction of the movement, the mixing effect can be enhanced (Kobe Steel process [160]).

Figure 109 shows desulfurization processes with mechanical stirrers. The stirrer proposed by Östberg (see Fig. 109 A) has a *hollow* stirrer which discharges the *hot metal* through horizontal tubes by centrifugal forces and absorbs hot metal from the bottom of the vertical tube. The simple stirrer proposed by Kraemer in Rheinstahl Hüttenwerk (Fig. 109 B) is made of refractory. This stirrer mixes *desulfurizing fluxes* into the hot metal. The stirrer developed by Nippon Steel (Fig. 109 C) is an impeller which produces eddy flow of the *hot metal.* Results of plant experiments for this stirrer are shown in Figure 110. According to this figure the process is highly effective for desulfurization. The stirrer method is widely used in North America, Eu-

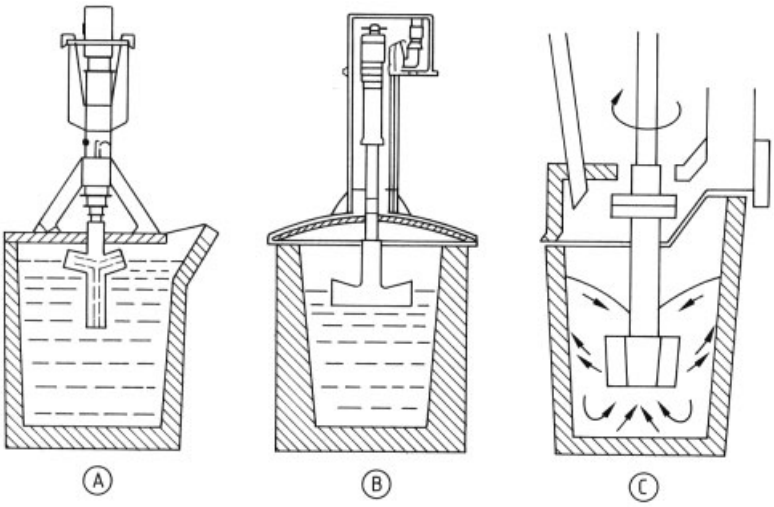

Figure 109. Stirrers for desulfurization of hot metal [163] (Courtesy of The Iron and Steel Institute of Japan)
A) Demag-Östberg process; B) Rheinstahl process; C) KR process

rope, and Japan for the production of low-sulfur steel or ultra-low sulfur steel.

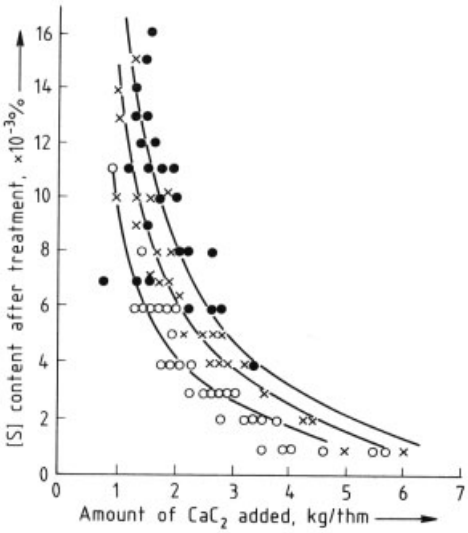

Figure 110. Relation between amount of CaC_2 added and [S] content after treatment [163] (Courtesy of The Iron and Steel Institute of Japan)
[S] Content before treatment (1350 °C): ○ 0.02 %; × 0.03 %; ● 0.04 %

The *injection method* of hot metal desulfurization is widely accepted for the desulfurization in a ladle or a torpedo-car. In this method the flux powder of desulfurization is injected with a gas stream. As flux powder CaC_2, $Ca(CN)_2$, CaO and soda ash are widely used. Magnesium is also used. The injection lance is better for dipping into the hot metal bath. The dipping depth should be 1/2 to 3/4 of the bath depth. NSC [164] and TSAG [165] developed the respective process for desulfurization by using the injection method.

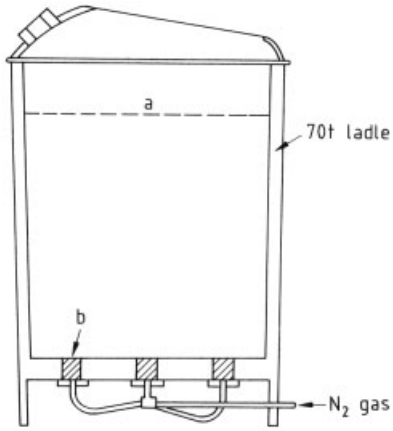

Figure 111. Desulfurization ladle with porous plugs [166] (Courtesy of The Iron and Steel Institute of Japan)
a) Surface of hot metal; b) Porous plugs

In the *gas-bubbling* method, the desulfurizing flux is at first added to the surface of the hot metal and then mixing is started by the injection

of gas into the hot metal. The injection is made through the lance dipped into the hot metal (top-blowing method) or through the porous plug at the bottom (bottom-blowing method). Figure 111 represents schematically the bottom-blowing desulfurization ladle. Ultra-low sulfur steel containing 0.001 – 0.002 % S can only be produced by this method if the second addition of calcium carbide is made after the slag of the first treatment is removed.

The *gas-lift mixing reactor* method (GMR method) was proposed by Kobe Steel. The basic principle of this method is to use the buoyancy force of the gas injected at the bottom of the riser as driving force for a circulatory motion of the hot metal as illustrated in Figure 112. This method removes sulfur in hot metal to $<$ 0.002 % by adding 5 kg of CaC_2 per tonne of hot metal.

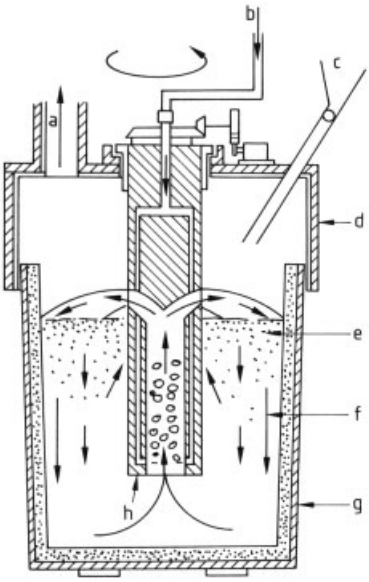

Figure 112. Schematic diagram of the gas lift mixing reactor for desulfurization [167] (Courtesy of The Iron and Steel Institute of Japan)
a) Dust collecting duct; b) Compressed nitrogen; c) Inlet of flux; d) Cover for dust catch; e) Desulfurizing flux; f) Hot metal; g) Ladle; h) Main part of gas-lift mixing reactor

If magnesium is used as a desulfurizing flux, magnesium – coke [168], a plunging bell [168 – 170] injection of a magnesium – aluminum alloy, surface-coated magnesium, or a mixture of CaO and Mg with gas [171 – 173] are used to suppress the vaporization loss of magnesium.

Since some problems remain unsolved in magnesium desulfurization (such as violent scattering of hot metal and resulfurization which requires a treatment time that is too long), the operation techniques must be improved. The advantages of this method in comparison to other methods are the very small amounts of flux required, ca. 1 kg/t, and the easy disposal of the slag.

It is very difficult to compare the various desulfurization processes for the pretreatment of hot metal, because many factors such as facility cost, operation cost, amount and cost of desulfurizing flux, degree of desulfurization, workability, and iron loss must be considered. Figure 113 shows the relationship between degree of desulfurization and operation cost. To reduce the cost of desulfurization, effective and inexpensive desulfurization fluxes and desulfurization processes requiring less flux must be developed.

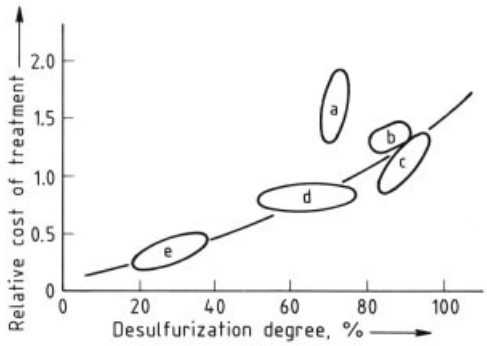

Figure 113. Relation between desulfurization degree and relative cost of treatment [164] (Courtesy of Dept. of Materials Science and Engineering, McMaster University, Hamilton, Canada)
Sulfur content before treatment: 0.030 %
a) Mag-coke process; b) Porous plug process; c) KR process; d) Top blowing process; e) Soda-ash pouring process

2.6. Direct Reduction Processes

The distinctive features of the individual processes have already been described in Section 2.4.

The name "direct reduction" is misleading, because these reduction processes do not consti-

tute a "more direct" route to steel than does the blast furnace. Nor does the term have anything to do with direct reduction in the blast furnace. Despite all these objections, however, the term has become accepted in international usage.

Table 32 lists the plants for direct reduction of iron ore that were in operation or under construction in 1995, together with their locations and the processes employed.

Table 33 contains technical data for the direct reduction processes described in this chapter and certain other processes. The stated capacities should be regarded as approximate, as some of the processes have been further developed. The annual plant outputs often vary from the original design capacity. In particular, the demand of the client (i.e., of the electric steel mill), a low scrap price, natural gas supply, or a shortage of spare parts may lower the annual output.

In most direct reduction processes, up to 92 – 95 % of the iron oxide is reduced to metallic iron. The degree of metallization $Fe_{metallic}/Fe_{total}$ in % is an important variable for characterizing the direct-reduced iron (DRI) produced by the process in question.

2.6.1. Shaft Furnace Processes for Direct Reduction

Wiberg – Söderfors Process. The oldest shaft furnace process for direct reduction of iron ore, the Swedish Wiberg process, is based on a patent issued in 1918. Up to the end of the 1950s, it was technologically improved and adapted to changed conditions on the ore and fuel markets. Nowadays, the Wiberg process is only of historical significance.

The development of the process was promoted above all by the growing shortage of very low-sulfur charcoal pig iron, the raw material for swedish high-quality steel. In addition, it was intended that coke would only be used for the reduction process, the process heat being supplied by cheap electricity available because of the local conditions.

In the Wiberg process (Fig. 114) the iron ore is preheated in the top section of the furnace (a) to the reduction temperature of ca. 950 °C by hot gases generated by combustion of part of the so-called excess gas. In the zone immediately below, the ore is reduced to wustite with ca. one-third of the reducing gas volume. The reduction to metallic iron takes place in the bottom section of the furnace at ca. 950 °C with ca. 1300 m^3 (STP) gas/t DRI.

Two thirds of the reducing gas are withdrawn by means of a hot gas fan (e) and regenerated in the carburetor (b) over a hot layer of coke. The energy required for the endothermic reactions

$$CO_2 + C \rightarrow 2\,CO$$

$$H_2O + C \rightarrow CO + H_2$$

is supplied by resistance-heating electrodes.

The gas, which is contaminated with sulfur from the coke, is passed through dolomite in a shaft furnace for desulfurization (c). The reducing gas contains 95 % carbon monoxide and hydrogen in a ratio of ca. 3 : 1.

The Wiberg process, which was developed for the special conditions of the Swedish iron industry, was a pioneering technology. The high cost of coke and electrical energy to generate the reducing gas is the reason why this process is no longer in use.

The principle of generating the reducing gas by a chemical reaction between the off-gas of the reduction shaft (the top gas) and natural gas has been applied in two processes: the Purofer and the Midrex process.

Purofer Process. This process (Fig. 115) was developed at Hüttenwerke Oberhausen (later Thyssen Niederrhein); the ICEM process developed by the Romanian Research Institute for Metallurgy is similar to it. The ore (preferably pelletized) passes vertically downward through the shaft furnace in countercurrent to the hot reducing gas. The gas is used for heating the charge and for reduction. The reducing gas is generated catalytically in regenerative-type gas generators. After cooling and purification, the top gas is used partly for heating the gas generators and partly for chemical conversion of the natural gas at 900 – 1000 °C to reducing gas according to the equations

$$CH_4 + CO_2 \rightarrow 2\,CO + 2\,H_2$$

$$CH_4 + H_2O \rightarrow CO + 3\,H_2$$

Table 32. Direct reduction plants worldwide as of December 31, 1997 [175]

Plant	Location	Capacity, 10^6 t/a	Modules	Start-up	Status *
Midrex process					
Georgetown Steel	Georgetown, SC, USA	0.40	1	71	O
Ispat HSW	Hamburg, Germany	0.40	1	71	O
Sidbec-Dosco 1	Contrecoeur, Quebec, Canada	0.40	1	73	O
SIDERCA	Campana, Argentina	0.40	1	76	O
Sidbec-Dosco 2	Contrecoeur, Quebec, Canada	0.60	1	77	O
SIDOR I	Matanzas, Venezuela	0.35	1	77	O
ACINDAR	Villa Constitucion, Argentina	0.60	1	78	O
Qatar Steel Co.	Mesaieed, Qatar	0.40	1	78	O
SIDOR II	Matanzas, Venezuela	1.27	3	79	O
Caribbean Ispat Ltd.	Point Lisas, Trinidad/Tobago	0.84	2	80 – 82	O
Delta Steel	Warri, Nigeria	0.51	1	82	I
Hadeed I	Al-Jubail, Saudi Arabia	0.80	2	82 – 83	O
OEMK	Stary Oskol, Russia	1.67	4	83 – 88	O
Amsteel Mills	Labuan Island, Malaysia	0.65	1	84	O
Khouzestan Steel Co.	Ahwaz, Iran	1.20	3	85 – 92	O
ANSDK I	El Dikheila, Egypt	0.72	1	86	O
LISCO I	Misurata, Libya	1.10	2	89 – 90	O
Essar Steel I & II	Hazira, India	0.88	2	90	O
OPCO	Puerto Ordaz, Venezuela	1.00	1	90	O
VENPRECAR	Matanzas, Venezuela	0.66	1	90	O
Essar Steel III	Hazira, India	0.44	1	92	O
Hadeed II	Al-Jabail, Saudi Arabia	0.65	1	92	O
NISCO	Mobarakeh, Iran	3.20	5	92 – 94	O
Ispat Industries	Raigad, India	1.00	1	94	O
ANSDK II	El Dikheila, Egypt	0.80	1	97	O
IMEXSA	Lázaro Cárdenas, Mexico	1.20	1	97	O
LISCO II	Misurata, Libya	0.65	1	97	O
Tuscaloosa Steel/Mobile DRI	Mobile, AL, USA	0.80	2	97 – 98	O/C
American Iron Reduction	Convent, LA, USA	1.20	1	98	C
COMISIGUA	Matanzas, Venezuela	1.00	1	98	C
Hanbo Steel	Asan Bay, South Korea	0.80	1	98	C
Ispat DR3	Point Lisas, Trinidad/Tobago	1.36	1	98	C
Saldanha Steel	Saldanha Bay, South Africa	0.804	1	98	C
		28.754	49		
HyL III process					
Hylsa 2M5	Monterrey, Mexico	0.25	1	79	O
Hylsa 3M5	Monterrey, Mexico	0.50	1	83	O
IMEXSA	Lázaro Cárdenas, Mexico	2.00	4	88 – 90	O
Grasim	Raigad, India	0.75	1	93	O
PT Krakatau Steel	Kota Baja, Indonesia	1.35	2	93 – 94	O
PSSB	Kemaman, Malaysia	1.20	2	93	O
Usiba	Salvador, Bahia, Brazil	0.31	1	94	O
Hylsa 2P5	Puebla, Mexico	0.61	1	95	O
Hadeed III	Al-Jubail, Saudi Arabia	1.10	1	98	C
Hylsa 4M	Monterrey, Mexico	0.675	1	98	C
Lebedinsky GOK	Gupkin, Russia	0.90	1	98	C
Posven	Matanzas, Venezuela	1.50	2	99	C
		11.145	18		
HyL I process					
Tamsa	Veracruz, Mexico	0.28	1	67	I
SIDOR I	Matanzas, Venezuela	0.36	1	76	O
PT Krakatau Steel	Kota Baja, Indonesia	1.68	3	78 – 82	O/I
SIDOR II	Matanzas, Venezuela	1.70	3	80 – 81	O
ASCO	Ahwaz, Iran	1.03	3	93 – 95	O
		5.05	11		
FINMET process					
BHP	Port Hedland, Australia	2.50	4	98	C
Orinoco Iron	Matanzas, Venezuela	2.20	4	99 – 00	C
		5.05	8		

Table 32. Continued

Plant	Location	Capacity, 10^6 t/a	Modules	Start-up	Status *
SL/RN process					
Piratini	Charquedas, Brazil	0.06	1	73	I
SIIL	Paloncha, India	0.06	2	80 – 85	O
Siderperu	Chimbote, Peru	0.10	3	80	O
ISCOR	Vanderbijlpark, South Africa	0.72	4	84	O
BSIL I	Chandil, India	0.15	1	89	O
Prakash Industries I	Champa, India	0.30	2	93 – 96	O
Nova Iron & Steel	Bilaspur, India	0.15	1	94	O
		1.54	14		
Iron Carbide process					
Nucor Iron Carbide	Point Lisas, Trinidad/Tobago	0.30	1	94	O
Qualitech	Corpus Christi, TX, USA	0.66	2	98	C
		0.96	3		
Jindal process					
Jindal Strips	Raigarh, India	0.50	5	93 – 96	O
Monnet Ispat I	Raipur, India	0.10	1	93	O
Jindal Strips	Raigarh, India	0.10	1	98	C
Monnet Ispat II	Raipur, India	0.10	1	98	C
		0.80	8		
DRC process					
Scaw Metals I	Germiston, South Africa	0.18	2	83 – 89	O
Tianjin Iron & Steel	Tianjin, China	0.30	2	97	O
Scaw Metals II	Germiston, South Africa	0.15	1	97	O
		0.63	5		
GHAEM process					
Esfahan Steel	Esfahan, Iran	0.60	1	96	O
CODIR process					
Dunswart	Benoni, South Africa	0.15	1	73	O
Sunflag	Bhandara, India	0.15	1	89	O
Goldstar	Mallividu, India	0.22	2	92	O
		0.52	4		
Circored process					
Cliffs & Associates, Ltd.	Point Lisas, Trinidad/Tobago	0.50	1	98	C
Fior process					
Fior de Venezuela	Matanzas, Venezuela	0.40	1	76	O
OSIL process					
OSIL	Keonjhar, India	0.10	1	93	O
Lloyd's Metal I	Ghugus, India	0.15	1	95	O
		0.25	2		
Purofer process					
ASCO	Ahwza, Iran	0.33	1	77	O
SIIL process					
Bellary Steels	Bellary, India	0.06	2	92 – 93	O
HEG	Borai, India	0.06	2	92	O
Kumar Met.	Nalgonda, India	0.06	2	93	O
Raipur Alloys	Raipur, India	0.06	2	93	O
Tamilnadu Sponge	Salem, India	0.03	1	93	O
Aceros Arequipa	Pisco, Peru	0.06	2	96	O
		0.33	11		
TISCO process					
Ipitata Sponge I	Joda, India	0.12	1	86	O
Ipitata Sponge II	Joda, India	0.12	1	98	C
		0.24	2		
DAV process					
Davsteel	Cullinan, South Africa	0.04	1	85	O
Kinglor-Metor process					
No. 3 Mining Enterprises	Maymyo, Burma	0.04	2	81 – 84	O

Note: This list does not include plants that are inoperable or have been dismantled.
* Status code: O = Operating; I = Idle; C = Construction.

Table 33. Typical consumption figures for direct reduction processes

Process	Reductant	Consumption, GJ/t DRI	Electrical energy, kWh/t DRI	Metallization, %
Midrex	natural gas	10.0	100	92 – 94
HyL I	natural gas	18.9		85
HyL III	natural gas	10.5		90 – 93
Purofer	natural gas, oil	14.3	93	95
Nippon steel	natural gas	10.5		90 – 92
Fior	natural gas	16.0		90
SL/RN	coal	14.8	80	93

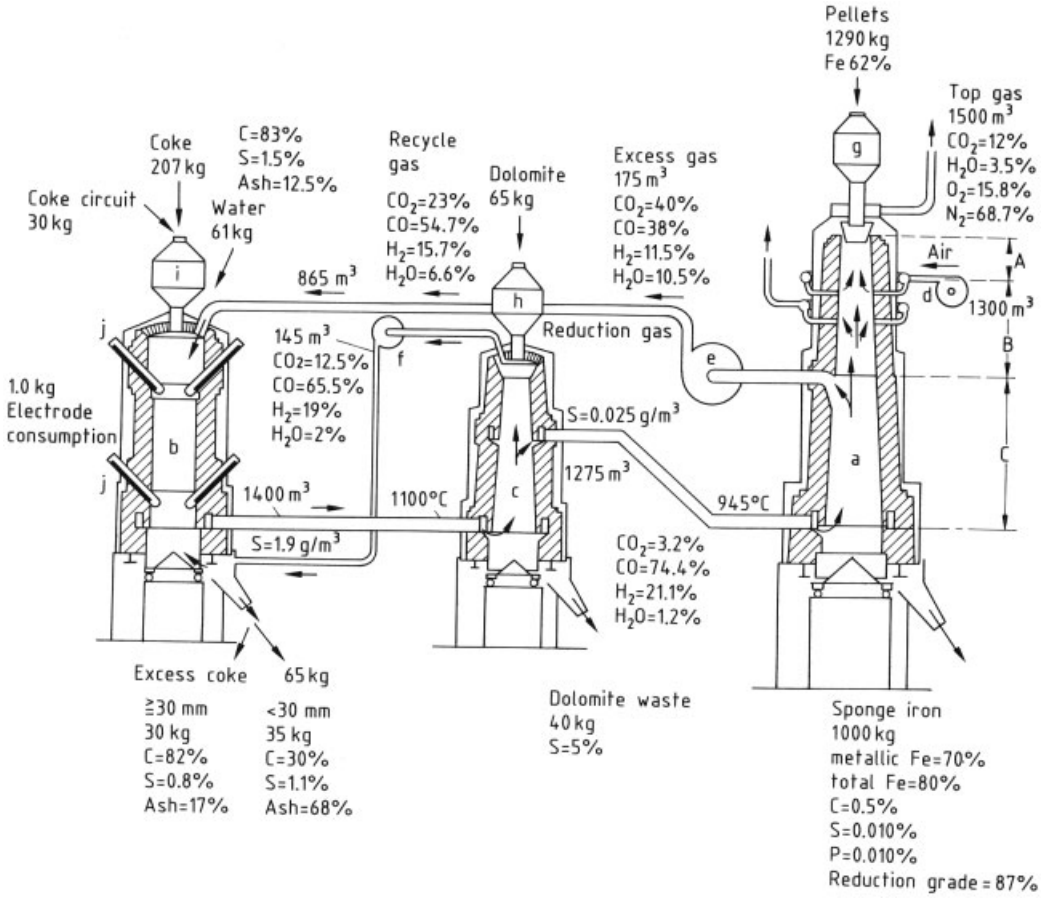

Figure 114. Flow Sheet and material balance (based on 1 t sponge iron) of the Wiberg process. All volumetric values of the gases refer to standard conditions [174]
a) Reduction shaft; b) Carburetor; c) Desulfurization shaft; d, e, f) Fans; g, h, i) Charging bins for pellets, dolomite, and coke; j) Electrodes
A) Preheating zone; B) Pre-reduction zone; C) Main reduction zone

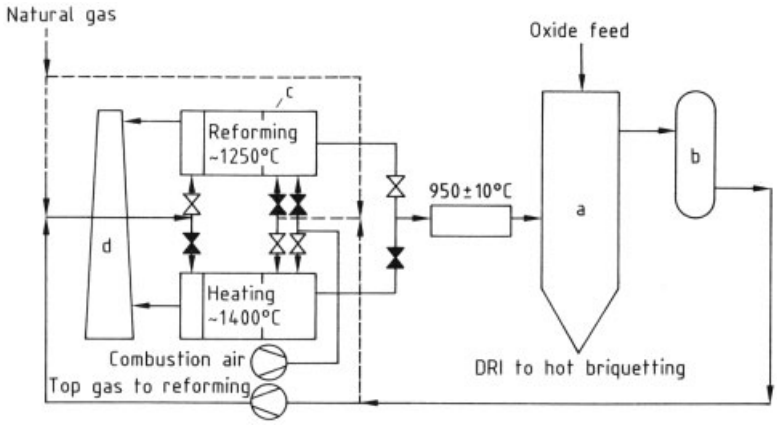

Figure 115. Schematic of the Purofer process
a) Shaft furnace; b) Top-gas scrubber; c) Gas reformer; d) Stack

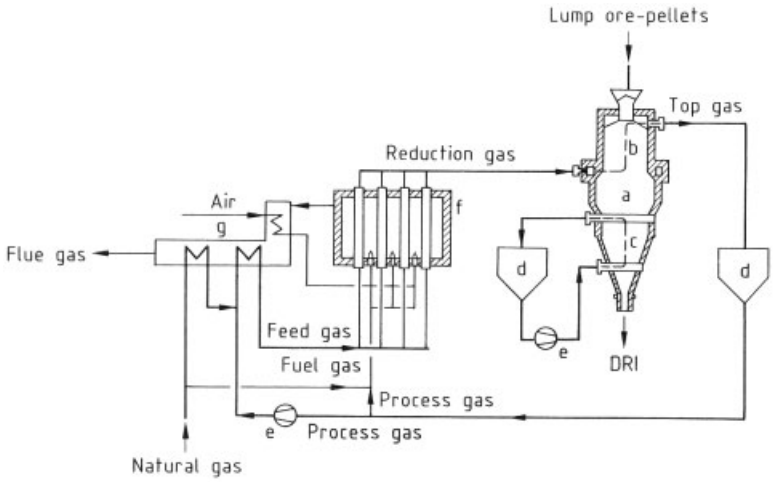

Figure 116. Schematic of the Midrex process
a) Shaft furnace; b) Reduction zone; c) Cooling zone; d) Scrubber; e) Gas compressor; f) Gas reformer; g) Recuperator

The carbon from methane reforming that is deposited on the catalyst of the regenerator during reducing-gas production is burned off during heating-up. This regenerative phase causes slight variations in reducing gas temperature and composition. In the Purofer process, the DRI is discharged hot and can be transported in containers either directly to an electric steel mill or to a hot briquetting plant.

In addition to the pilot plant in Oberhausen (capacity 500 t/d), one plant was built in Brazil (but with a Texaco heavy-oil gasifier as gas generator), which is now dismantled. The only plant in operation in Iran is designed for operation with natural gas.

Midrex Process. The most successful gas-based direct reduction process is the Midrex process developed by Midland-Ross Corp. of Toledo, Ohio (see Fig. 116) [175].

In this process, natural gas continually undergoes catalytic conversion to hydrogen and carbon monoxide (ratio ca. 1.5 : 1) at ca. 900 °C using part (ca. 2/3) of the top gas, which is cleaned and cooled to lower its water-vapor content. Reduction is accomplished in the cylindrical part of the shaft furnace at 780 – 900 °C, depending on the temperature at which sintering ("sticking") of the reduced ore or the pellets occurs. In the conical discharge section of the shaft furnace, the DRI is cooled to ca. 45 °C by recirculated cooling gas. Here the carbon content of the

product can also be adjusted to between 1.2 and 2.5 %. There are also Midrex plants featuring hot discharge followed by hot briquetting.

One-third of the top gas is used to heat the gas reformer, whereas the hot flue gases serve to preheat the combustion air and the feed gas mixture prior to reforming.

With increasing heat recovery in the recuperator and reduction temperatures of ca. 900 °C, energy consumption per tonne of DRI may fall to 9.6 GJ in the Midrex process.

The reducing gas can also be generated by *steam reforming* of natural gas. The first shaft furnace direct-reduction process to employ steam reforming was the *Armco process.* Here natural gas undergoes continuous catalytic conversion with excess steam according to the reaction:

$$CH_4 + x\,H_2O \rightarrow CO + 3\,H_2 + (x-1)\,H_2O$$

After removal of the water vapor, the top gas is used as cooling gas.

Only one commercial plant was ever built because of the difficulty of achieving complete mixing of the cooling gases entering the lower part of the shaft with the hot top gases from the reduction zone, the low reducing potential of the gases, and a range of technical problems.

Nippon Steel (NSC) Process. The reducing gas containing over 95 % carbon monoxide and hydrogen is generated by steam reforming with only a slight excess of water vapor and addition of top gas freed from water and carbon dioxide (see Fig. 117). Like the Purofer process, the NSC process has a hot discharge facility, from which DRI can be filled into containers and taken either directly to a steel mill or to a hot briquetting plant. Only one plant in Malaysia with a design capacity of 1000 t/d was built which was later modified to a HyL plant.

HyL-III Process. Hojalata y Lamina S.A. (HyL) of Mexico has also developed a shaft-furnace process based on steam reforming of natural gas (HyL-III process, Fig. 118). This process represents a further development of HyL's retort process (see Section 2.6.2). Because of the large quantity of excess steam produced by continuous catalytic conversion of the natural gas, the generated gas must first be cooled to remove the excess water vapor. After addition of water- and carbon dioxide-free top gas and indirect heating to ca. 850 °C, the reducing gas is fed into the reduction shaft. In the bottom section of the shaft furnace, the DRI is cooled down to $<$ 50 °C as in the Midrex process. HyL-III plants can also be operated with hot discharge and hot briquetting.

The degree of metallization of the DRI produced by the various shaft furnace processes

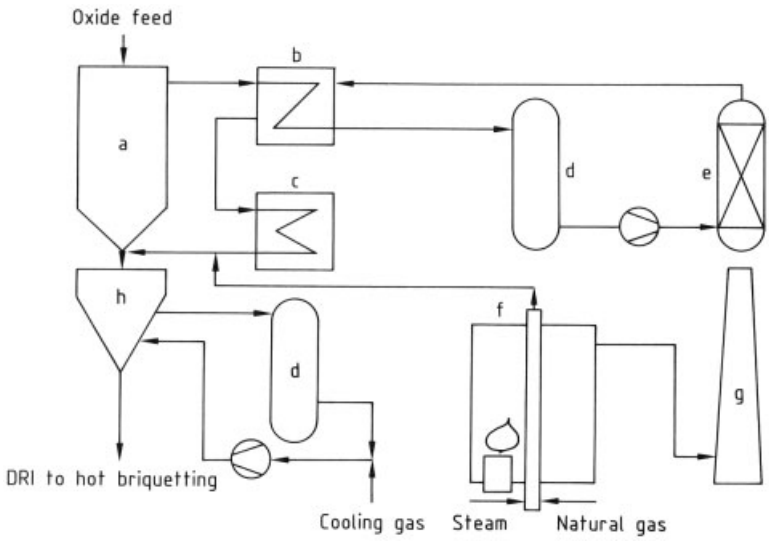

Figure 117. Schematic of the Nippon Steel process
a) Shaft furnace; b) Heat exchanger; c) Gas heater; d) Dust remover; e) CO_2 scrubber; f) Reformer; g) Stack; h) Cooling zone

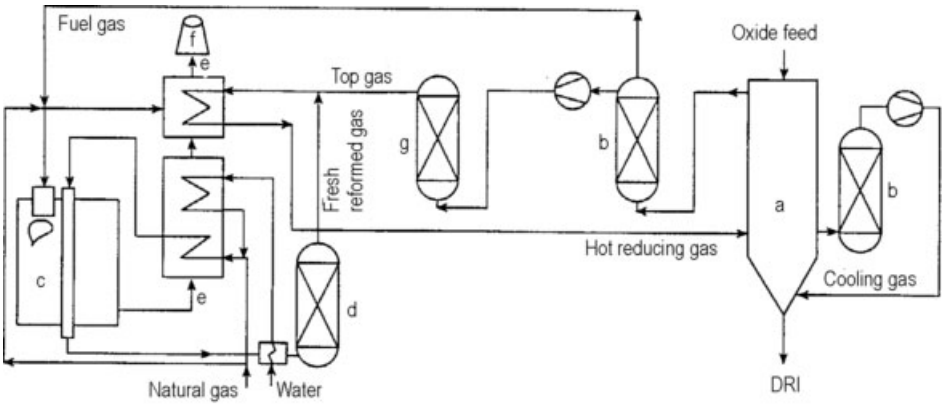

Figure 118. Schematic of the HyL-III process
a) Shaft furnace; b) Gas scrubber; c) Reformer; d) Cooler; e) Recuperator; f) Stack; g) CO_2 scrubber

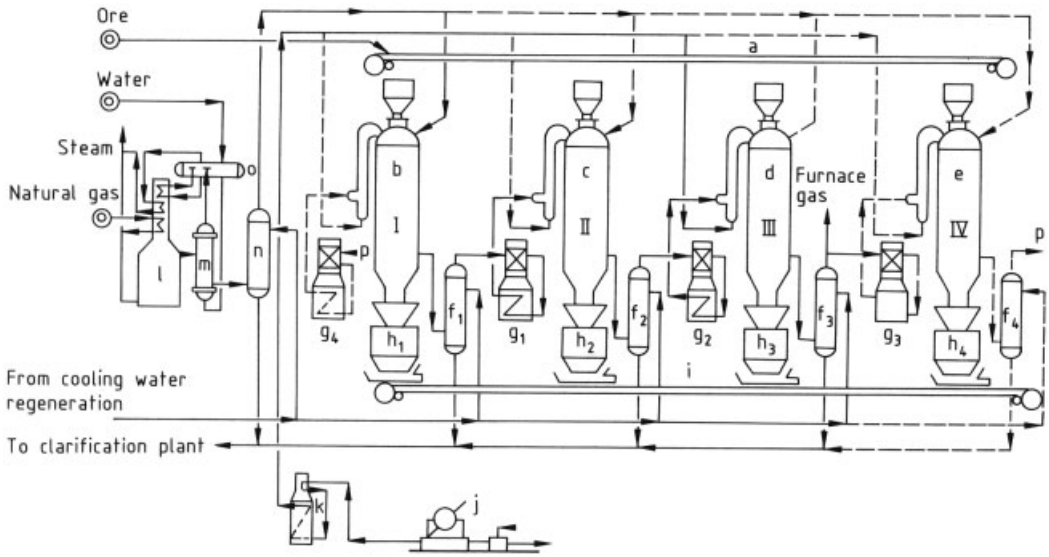

Figure 119. Flow sheet of the HyL I process
a) Ore conveyor belt; b) Retort I at cooling stage; c) Retort II at final reduction stage; d) Retort III at pre-reduction stage; e) Retort IV at discharging and charging stage; $f_1 - f_4$) Cooler; $g_1 - g_4$) Gas preheating; $h_1 - h_4$) Intermediate container for sponge iron; i) Conveyor belt for sponge iron; j) Air compressor; k) Air preheating; l) Reduction gas generator; m) Superheated steam generator; n) Cooler; o) Saturated steam generator; p) Connection between cooler f_4 and preheater g_4

ranges from 92 – 95 %. The carbon content is usually 1.5 – 2 %.

2.6.2. Retort Processes

HyL Process (Hyl-I Process). A number of plants employing this process were built until 1975 before the shaft-furnace processes became established. The development of this process began in 1951 at Hojalata y Lamina S.A., Monterrey (Mexico). It operates with four reactors simultaneously, each one being at a different point in the reduction cycle at a particular time. Fig. 119 shows the stage at which fully reduced DRI is cooled with cold reducing gas in retort I. At this point, the carbon content of the DRI is adjusted to ca. 2 %. After cooling and water-vapor condensation, the off-gas from this retort is reheated to reduction temperature and passed through retort II, where pre-reduced material is fully reduced. The gas is then recooled to remove any newly formed water vapor, reheated

and fed to retort III containing freshly charged ore, which is then pre-reduced. The off-gas from this stage is cooled and freed from water vapor; it is then available for use as fuel gas for steam raising. Retort IV is simultaneously emptied from DRI and charged with ore.

As lump ore or pellets are stationary in the retorts during reduction, the risk of sticking is high. Sophisticated removal devices are required for emptying the retorts.

The degree of metallization of the DRI is often lower at the bottom of the retorts than in the upper layers. Thus the target average degree of metallization is often only ca. 86 %. As gas generation is carried out with a large excess of steam and cooling and reheating of each retort is very costly, an energy consumption of ca. 21 GJ per tonne DRI must be taken into account.

Höganäs Process. This process has been in operation since 1911 and has undergone no significant technical modifications since the 1950s, nor has its use become any more widespread. In this process, iron ore is reduced with coal in hermetically sealed crucibles with indirect heating (by combustion of the carbon monoxide escaping from the crucibles). The sulfur content of the DRI is minimized by adding limestone or dolomite to the charge mix and rapidly cooling the product. It is thus suitable for use in powder metallurgy or stainless-steel production. The capacity of plants operating on the Höganäs process is relatively low, ca. 35 000 t/a.

Kinglor – Metor Process. This process occupies a category between the shaft furnace and the retort processes. The development of this process (Fig. 120) by Co. Kinglor – Metor S.p.a., Mineraria & Metallurgica, Italy, began in 1971 but was based on the Echeverria process which had been in operation in Spain since 1957 [176]. Like the rotary kiln processes (see Section 2.6.4) and in contrast to the gas-based direct reduction processes described so far, it uses a solid carbon-containing reductant, which is charged together with the ore (6 – 25 mm size) and, where applicable, limestone or dolomite is used for desulfurization. The heat for the strongly endothermic reaction

$$FeO_x + x\,C \rightarrow Fe + x\,Co$$

which in this case takes place at ca. 1050 °C, is supplied through the wall of the shaft furnace, which must therefore have a high thermal conductivity. Silicon carbide satisfies the requirements as to thermal conductivity, heat resistance, and abrasion resistance. The shaft heating burners are designed for gaseous and liquid fuels. The reduced material is indirectly cooled with gas in the bottom zone of the shaft. Commercial plants consist of a group of six reduction chambers, each with a capacity of ca. 20 t/d DRI.

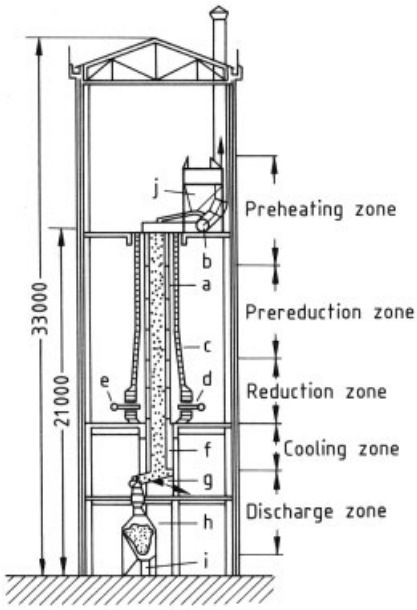

Figure 120. Shaft furnace of the Kinglor-Metor process (dimensions in mm)
a) Silicon carbide shaft; b) Gas discharge pipe; c) Fireclay outer wall; d) Burner; e) Gas or oil pipe; f) Air cooling; g) Extruding screw discharge; h) Hopper; i) Sponge iron discharge; j) Charging hopper

2.6.3. Fluidized-Bed Processes

Fluidized-bed processes for direct reduction exploit the lower cost of fine ores compared with lump ores and pellets. However, the disadvantages of these processes are that (1) the fine ores have a tendency towards sticking above ca. 700 °C and that (2) below 600 °C the susceptibility to reoxidation of the DRI is so high that it must be stored under inert gas or hot-briquetted.

Fluidized-bed processes that operated with pure hydrogen were the *Novalfer process* and

the *H-Iron process;* however, they did not attain any major technological significance. In the *HIB-process,* reduction was carried out in several stages with a gas containing ca. 85 % hydrogen. The ore of 0 – 2 mm size underwent only ca. 65 % metallization and was then hot-briquetted. The sole commercial plant has been converted by Midrex into a shaft furnace plant with hot discharge – hot briquetting.

Fior (Fluidized Iron Ore Reduction) Process. Development of the Fior process (Fig. 121) started in the 1950s at Esso Research and Engineering in collaboration with Arthur D. Little. Following the building and operation of a 300 t/d pilot plant in Dartmouth, Nova Scotia, from 1965 to 1969, a 400 000 t/a commercial plant operating on this process was built in Puerto Ordaz, Venezuela.

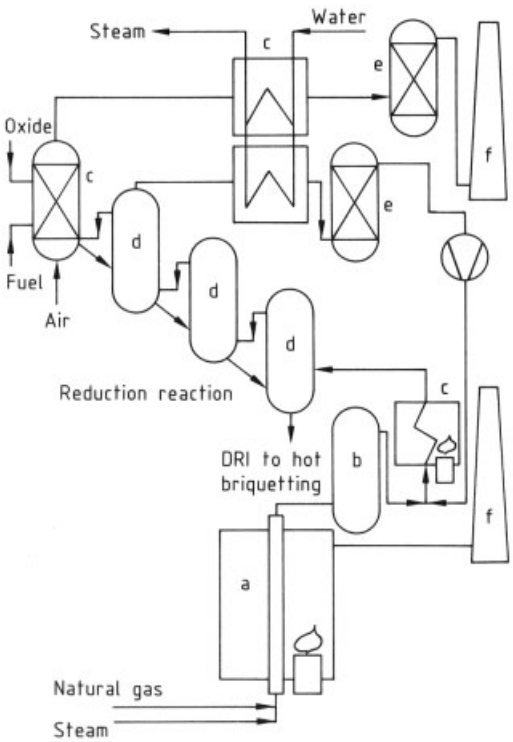

Figure 121. Schematic of the Fior process
a) Reformer; b) Gas cooler; c) Heater; d) Fluidized-bed reactor; e) Scrubber; f) Stack

After drying and preheating to ca. 880 °C, the fine ore (mostly < 5 mm) is metallized to a degree of ca. 92 % in a three-stage fluidized bed. The reducing gas is generated by steam reforming of natural gas in continuously operating catalytic reformers. The reduced fine ore is hot-briquetted, the briquettes being cooled and shipped to the end-users without further passivation.

The *Finmet process* is a further development of the Fior process, taking advantage of the experience gained in more than ten years of Fior plant operation.

Iron Carbide Process. In the Iron Carbide process developed by Stephens (Iron Carbide Holding Ltd.) (see Fig. 122) fine-grained iron oxide (0.1 – 1 mm) is reduced by a gas containing hydrogen and methane to give iron carbide [210]

$$3\,Fe_2O_3 + 5\,H_2 + 2\,CH_4 \rightarrow 2\,Fe_3C + 9\,H_2O$$

A 300 000 t/a iron carbide plant is operated in Trinidad by a joint venture of NUCOR and SAMITRI.

The iron oxide is heated to 700 °C (Fig. 122 a) and then fed to the fluidized-bed reactor (b), where reduction of the oxide and formation of iron carbide occurs at ca. 570 °C and ca. 4 bar.

The fluidized-bed reactor is equipped with baffles which effect movement of the solids from the inlet to the outlet. The relatively low reduction temperature required to form iron carbide resulted in a retention time of about 16 h in the first 900 t/d plant built in Trinidad. Due to a high gas : solids ratio of about 10 000 m^3 (STP) per tonne oxide, much larger gas volumes must be handled than in shaft furnace processes.

The fines carried over with the reduction gases are partially recovered in a cyclone (c) and recycled to the reactor. After heat exchange with the cold reduction gas (d) the off-gases are scrubbed, cooled for condensation of water (e), compressed, and recycled to the process via heat exchangers and a gas heater (h). Make-up gas is hydrogen. The reduction gas consists of 60 % CH_4, 34 % H_2, 2 % CO, 1 % H_2O, and minor amounts of CO_2 and N_2. The iron carbide formed in NUCOR's plant contains 90 % total iron (90 % as Fe_3C, 8 % as FeO, 2 % as metallic iron), 6.2 % C, and 2 % gangue.

Iron carbide is categorized as a nonpyrophoric material, therefore handling and ship-

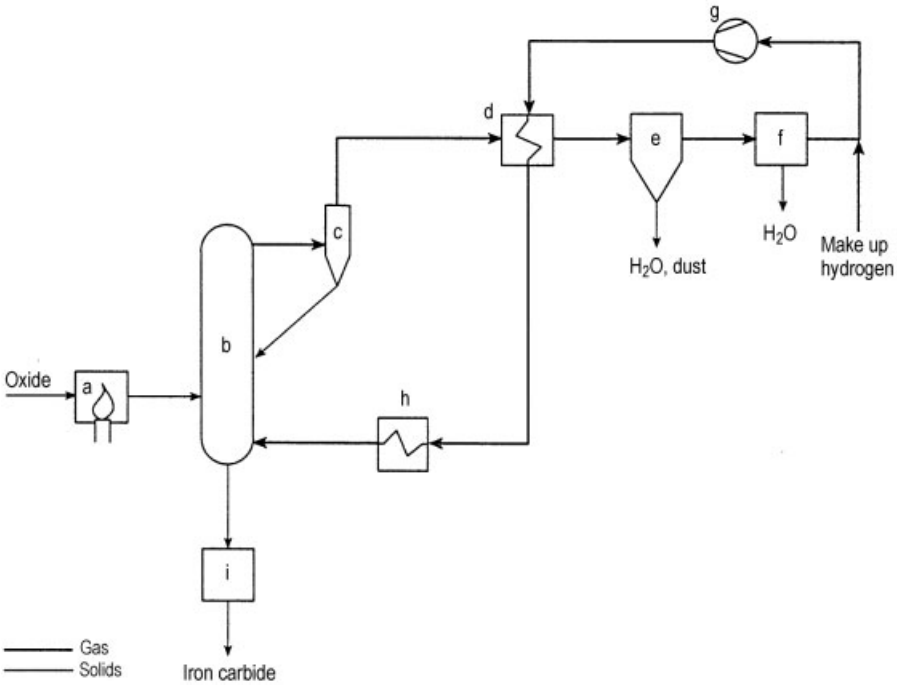

Figure 122. Flow sheet of the Iron Carbide process
a) Oxide heater; b) Reduction reactor; c) Cyclone; d) Heat exchanger; e) Scrubber/cooler; f) Aftercooler; g) Compressor; h) Gas heater; i) Product cooling

ping do not require briquetting or other treatment to avoid reoxidation.

Circored Process. In the Circored process (see Fig. 123) developed by Lurgi [211] small-size iron oxide ores (0.1 – 1 mm) are reduced by hydrogen according to

$$Fe_2O_3 + 3\,H_2 \rightarrow 2\,Fe + 3\,H_2O$$

The ore is preheated to ca. 900 °C and first prereduced to about 65 % metallization in a circulating fluidized-bed reactor (CFB) within 20 min.

Final reduction to 90 – 93 % metallization is accomplished at lower gas velocities in a subsequent bubbling fluidized-bed reactor (FB) with several compartments. This step requires up to 4 h, depending on the desired metallization and the reducibility of the oxide. The reactors are operated at temperatures generally below 650 °C to avoid sticking. The off-gases from the final reduction step in the FB and make-up hydrogen are passed into the CFB reactor. Most of the solids in the off-gases from the CFB are recovered in the recycle cyclone. The gases are then led through a heat exchanger, dedusted, cooled for condensation of water and compressed to the plant pressure of ca. 4 bar. After heat exchange with the hot off-gases from the CFB the reduction gas is heated to 750 °C before being reintroduced to the two reduction reactors.

Before entering the CFB, hot make-up hydrogen is used to transport the reduced fines from the FB to a briquetting plant and to heat them up. The reduced DRI fines are hot-briquetted, if the product is intended for export. It can also be charged directly into smelting furnaces.

A 500 000 t/a Circored plant is currently being started up in Trinidad.

The use of hydrogen as a single reductant source leads to the fact that no carbon is present in the Circored product. By introduction of hydrocarbons into the product stream after the final reduction stage the Circored process can also be applied for production of a carbon containing hot briquetted iron (HBI) with ca. 1 % carbon.

Circofer Process. The Circofer process (see Fig. 124) is another fluidized-bed process developed by Lurgi Metallurgie. This process also

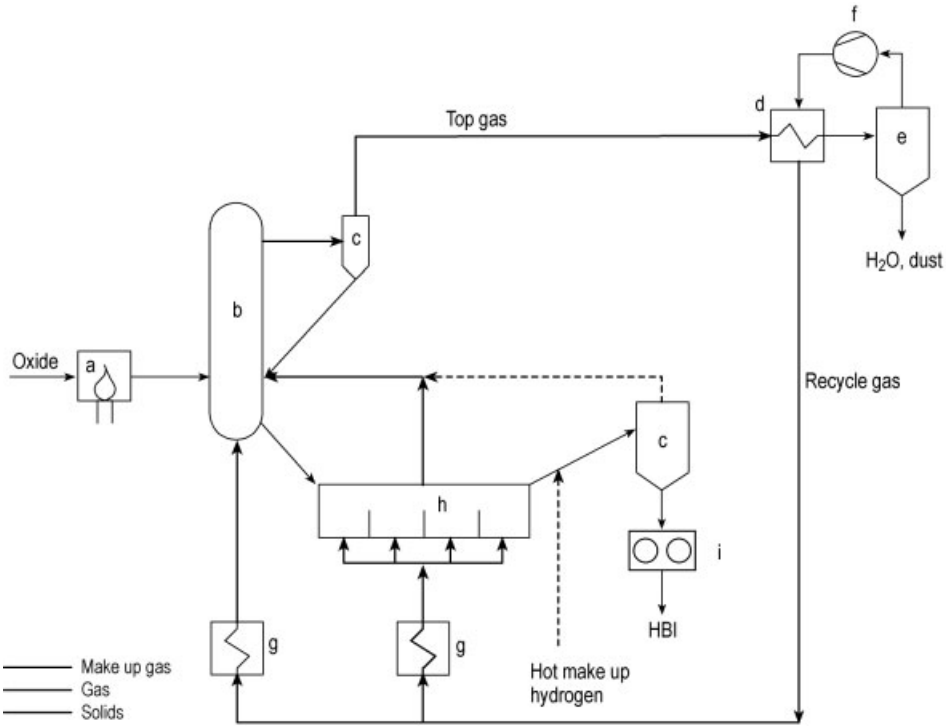

Figure 123. Flow sheet of the Circored process
a) Oxide heater; b) Circulating fluidized bed; c) Cyclone; d) Heat exchanger; e) Scrubber/cooler; f) Compressor; g) Gas heaters; h) Fluidized bed; i) Briquetting machine

operates with a CFB prereduction and a final FB reduction step and is based on coal as the reductant and basic energy source. Fine coal is partially burned in a heat generator and the semi-coke produced is used both as a heat carrier to transfer heat to the CFB and to prevent sticking of the iron particles at the reduction temperature of > 900 °C. Final reduction is accomplished at lower temperature in the FB with recycled off-gases from the CFB. Gas treatment is similar to that in the Circored process. In the Circofer process, however, a CO_2 absorber step had to be included, because reduction is mainly effected with CO containing gases. A 5 t/d pilot plant is operating at Lurgi's R & D center in Frankfurt.

2.6.4. Rotary Kiln Processes

Rotary kiln processes for direct reduction of iron oxides are based on *coal* as reductant. Typical features of rotary kiln processes are their high flexibility with regard to feedstocks and their capability for economic production of even small quantities of DRI. All non-caking coals ranging from lignite through bituminous coals up to anthracite or coke breeze are suitable as reductants. In addition to the pellets and lump ores that are the main iron-bearing materials, iron sands, and ilmenite concentrates are also industrially used.

Of the various rotary kiln processes [177], the has attained the greatest technological significance, and most of the DRI currently produced in rotary kilns worldwide is manufactured by the SL/RN process (see Table 33). Therefore, the principle of the rotary kiln processes will be explained with reference to this process. The fundamentals of the reduction mechanism are dealt with in Section 2.4.

The SL/RN Process. The SL/RN process was developed in 1964 out of the combination of two separate processes: the Stelco – Lurgi (SL) process for producing DRI from *high-grade ores* and the Republic Steel – National Lead (RN) process for beneficiation of *low-grade ores.*

In the SL/RN process (Fig. 125), reduction of iron oxide is carried out in a refractory-lined

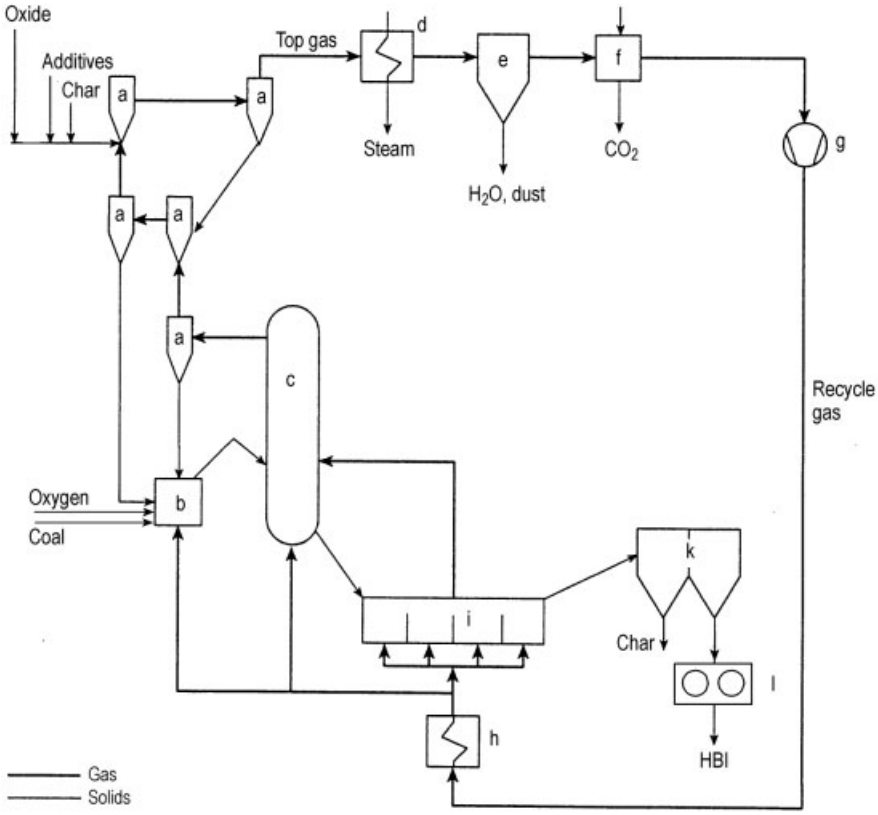

Figure 124. Flow sheet of the Circofer process
a) Preheating cyclone system; b) Heat generator; c) Circulating fluidized bed; d) Boiler; e) Scrubber/cooler; f) CO_2 removal; g) Compressor; h) Gas heater; i) Fluidized bed; k) Magnetic separation; l) Briquetting machine

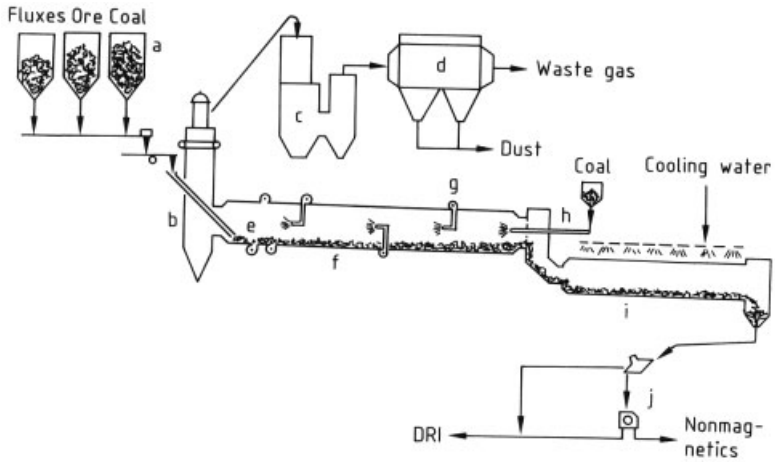

Figure 125. Lurgi SL/RN process
a) Bins for raw materials; b) Post-combustion chamber; c) Waste-heat boiler; d) Gas cleaning; e) Underbed air injection; f) Rotary kiln; g) Air tubes; h) Injection coal; i) Rotary cooler; j) Product separation

rotary kiln (f) inclined from the feed to the discharge end. The slope of the kiln is ca. 1.5 – 2.5 % depending on the feed materials.

The kiln is equipped at both ends with special seals between the rotating kiln and the fixed feed and discharge heads. These seals essentially prevent inleakage of air and thus reoxidation. A number of fans are installed on the shell of the kiln; these supply the necessary process air via so-called air tubes or via air injection nozzles in the feed zone of the rotary kiln.

The *internal kiln temperature* is continuously recorded by thermocouples installed along the entire length of the kiln and projecting into the kiln freeboard. The *power supply* to the shell fans and the transmission of the data recorded by the individual thermocouples is accomplished by means of slip rings mounted on the kiln shell. Devices for taking samples of material during operation are also mounted on the kiln shell.

The *central burner* installed at the kiln discharge end is used to dry out the kiln refractory lining during the initial startup and to heat up the system after a shutdown. Under normal operating conditions, only air without fuel is fed through the central burner. However, if low-reactivity coals (anthracite) are used, it may be necessary to continuously fire a small amount of supplementary fuel via the central burner to satisfy the overall heat requirement.

Fresh coal, recycle carbon (char), ore or pellets, and dolomite or limestone as desulfurizing agent are charged in a predetermined ratio at the kiln feed end by a gastight charging system.

Up to ca. 35 % of the total fresh coal required is fed into the kiln from the discharge end with a special pneumatic injection system (h). This improves the utilization of the energy contained in the coal volatiles and prevents depletion of carbon in the material bed, especially in the case of very long kilns.

The charge passes through the kiln in ca. 8 – 12 h. The retention time mainly depends on the slope, degree of filling, and the rotational speed of the kiln. For each combination of iron-bearing material and reductant, a minimum retention time is necessary to achieve the desired degree of metallization; this is determined with laboratory tests. For example, the necessary retention time for a combination of "porous" pellets and highly reactive lignite is low, whereas that for "compact" lump ores and low-reactive anthracite is high.

In simplified terms, the rotary kiln can be divided into a preheating and a reduction zone (see Fig. 126).

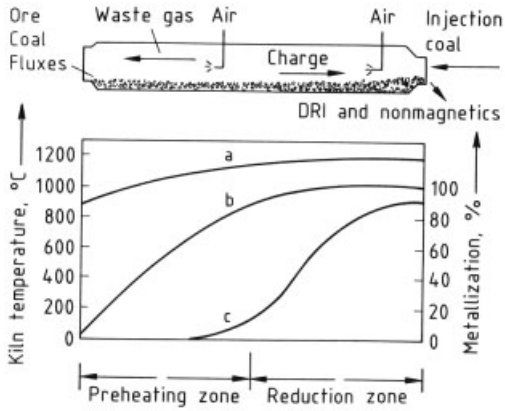

Figure 126. Temperature and degree of metallization as a function of the location in the kiln
a) Gas temperature; b) Charge temperature; c) Metallization

In the *preheating zone,* the feed materials are dried, and coal volatiles escape, and the charge is heated to reduction temperature (ca. 950 – 1100 °C).

An overview of the coals and ores used in SL/RN plants is given in Table 34. The usual grain sizes of the feed materials are < 15 mm for the reductant, < 25 mm for the iron-bearing materials and < 3 mm for the desulfurizing agent.

The charge is heated through exchange of heat with the kiln waste gases flowing countercurrently to the charge. The energy is supplied by the sensible heat of the waste gases from the reduction zone, by the partial combustion of the coal volatiles, and by direct burning of carbon in the preheating zone.

If high-volatile coals are used, injection of air into the charge can bring about partial combustion of the volatiles within the charge, which accelerates preheating.

In the preheating zone, a certain degree of prereduction is accomplished through the coal volatiles. However, the greater part of the reduction occurs in the *reduction zone* at near-constant temperature according to

$$Fe_xO_y + CO \rightarrow Fe_xO_{y-1} + CO_2$$

Table 34. Iron ores and coals processed in SL/RN plants

Plant	City	Country	Iron ore					Coal				
			Type	Analysis wt %, dry				Type	Analysis wt %, dry			
				Fe	SiO_2	CaO	Others		Fixed carbon	Volatile matter	Ash	Sulfur
Highveld	Witbank	South Africa	Mapochs lump ore	54	2.1		13.2 TiO_2 1.7 V_2O_5	Greenside	54	31	15	0.6
NZS	Glenbrook	New Zealand	iron sand	58	1.1	0.2	8 TiO_2	Huntly	51	43	6	0.3
Piratini	Charqueadas	Brazil	Itabira pellets	67	2.4	1.6		Charqueadas	40	25	35	0.4
Siderperu	Chimbote	Peru	Marcona pellets	66	2.2	1.0		Coke Breeze	81	3	18	0.7
SIIL	Paloncha	India	Bayaram lump ore	63	4.5	0.1		Singareni	44	31	25	0.3
Iscor	Vanderbijlpark	South Africa	Sishen lump ore	66	2.5	0.2		Van Dyksdrift	59	26	14	0.6
BSIL	Chandil	India	Bihar lump ore	67.4	1.5	0.1		Ray Bachara	51	24	25	0.7
Prakash	Champa	India	Orissa lump ore	66.5	1.8			Rajgamnar	52	27	21	0.8
Nova	Dagori	India	Orissa lump ore	66.5	1.8			Dhilwadi	49	28	23	0.7
Westralian Sands	Capel	Western Australia	ilmenite concentrate	27	1.0		58 TiO_2	Collie	57	37	6	0.8
TIWEST	Chandala	Western Australia	ilmenite concentrate	24	0.7		61 TiO_2	Collie	57	37	6	0.8

* Prereduction only.

The carbon dioxide formed during reduction reacts within the charge with fixed carbon according to the Boudouard reaction

$$C + CO_2 \rightarrow 2\,CO$$

and forms the carbon monoxide required for the reduction process.

The necessary heat in the reduction zone is supplied through combustion of the surplus carbon monoxide from reduction, of the volatiles in the injected coal, and of fixed carbon.

The desired temperature profile is established by controlled addition of combustion air along the length of the kiln.

Inside the kiln, an *oxidizing atmosphere* above the charge and a *reducing atmosphere* within the charge prevail in close proximity (see Fig. 127). Accordingly, the reactions occurring above the charge in the kiln freeboard are exothermic and those taking place in the material bed are endothermic. Separation of the two atmospheres is ensured by the gas generated in the material bed, which also prevents reoxidation of particles of DRI at the charge surface.

After leaving the rotary kiln, the DRI is cooled to ca. 100 °C together with the excess coal in an indirectly cooled rotary cooler. The DRI is separated from the nonmagnetic kiln discharge material by *screening* and *magnetic separation.*

The > 3 mm fraction of the *magnetic portion* can be directly used for steel production, the < 3 mm fraction is normally briquetted before further use.

The > 3 mm fraction of the *nonmagnetic portion* is mostly char and can be recycled to the process. The < 3 mm fraction, containing ash, calcined desulfurizing agent, and char, is disposed of.

In particular cases, e.g., if the ash content of the coal is very low, it is advantageous to charge the entire kiln discharge material directly into a downstream electric furnace without cooling and separation. This brings considerable savings in electric energy for the melting down process.

The waste gases leave the kiln at the feed end at ca. 850 – 1000 °C. They pass through a dust-settling chamber to remove coarse dust and enter an afterburning chamber, where combustible gases and carbon particles are burnt off. The waste gases are then cooled, cleaned and discharged to the atmosphere. Alternatively, the sensible heat of the gases leaving the afterburning chamber may be recovered in a downstream steam boiler for generation of process steam or combined for production of steam and electric power.

Since 1984, a four-strand SL/RN plant with a capacity of 600 000 t/a DRI and waste-heat recovery has been in operation at the Vanderbijl-

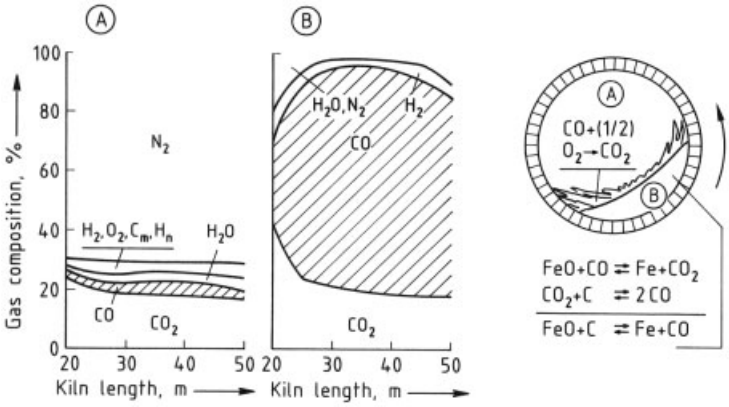

Figure 127. Gas composition in the kiln freeboard (A) and kiln charge (B) measured in an industrial kiln of 50 m length

park Works of ISCOR in South Africa. Approximately 1.9 t of process steam (1.6 MPa, 260 °C) per tonne of DRI are generated for the steel mill. In 1993 a second 150 000 t/a SL/RN plant with waste-heat recovery for power generation was installed in Champa, Maharashtra, India. This plant was extended by an identical unit in 1996.

Taking into account a variety of reducing agents, the energy requirement (without char recycling) referred to as the lower heating value of the coals, is ca. 19.3 GJ/t DRI for lignite briquettes (A), ca. 18.4 GJ for bituminous coal (B), and ca. 17.8 GJ for a mixture consisting of 70 % anthracite (C) and 30 % bituminous coal (B) [178]. The analyses of these coals are shown in Table 35.

If a combination C : B = 50 : 50 with recycled char is used, the energy requirement is only 14.8 GJ/t DRI [178].

Further typical consumption figures are given below:

Ore or pellets (Fe_{tot} 67 %), t	1.42 – 1.44
Fixed carbon, kg	380 – 475
Electric power, kWh	70 – 90
Water, m^3	1.5 – 2.5
Manpower, man hours	0.4 – 0.6
Repair and maintenance, $	8 – 10

In addition to DRI production, SL/RN rotary kilns are commercially used for *prereduction* of iron ore at Highveld Steel in Witbank, South Africa (13 units with an ore capacity of 200 000 t/a each, 30 – 35 % metallization) and for reduction of *ilmenite concentrates* at Westralian Sands Ltd, Capel, Western Australia (two units, total ilmenite throughput ca. 480 000 t/a) and at TIWEST, Chandala, Western Australia (one unit, ilmenite throughput ca. 300 000 t/a).

2.7. Smelting Reduction Processes

In smelting reduction processes hot metal is produced directly with coal, particularly noncoking coal.

The main reasons for developing smelting reduction processes (since 1970) was to find processes which produce liquid iron economically at low capacity and use coal directly. The processes should be more flexible regarding raw material and operation and should cause less pollution than a blast furnace including coke making.

The ore, fine ore, or lump ore is *prereduced* with the gas produced in the melting vessel. The prereduced ore (sponge iron) is melted down in a melting vessel where melting energy is generated by oxidation of coal. Some processes use also electrical energy in the melting stage.

The principles of the *prereduction stage* (shaft furnace or fluidized bed) are given in Sections 2.2 and 2.3 and the thermodynamic principles for the final *melter stage* are presented in Section 2.3.

Simulations with computer models were performed for different smelting reduction process configurations [179 – 181] to find an optimum process and to estimate the influence of post-combustion and prereduction on coal consumption. (Post-combustion is defined as the partial oxidation of a gas previously generated in the melting vessel, whereby the energy produced

Table 35. Coal analyses

Coal	A	B	C
Fixed carbon, %	44	59	88
Volatile matter, %	51	27	8
Ash, %	5	14	4
Lower heating value (dry), GJ/kg	23.3	27.0	32.7

by this combustion is transferred to the bath with high efficiency.) Economical and technical comparison can be found in the literature [182 – 184].

2.7.1. Characteristics of Coals for the Smelting Reduction Processes

The most important property of the coal for the smelting reduction processes is its volatile content, because this determines the gasification temperature of the reaction

$$\text{Coal} + \text{Oxygen} \rightleftharpoons \text{Carbon monoxide} + \text{Hydrogen}$$

Coals with *low volatile* content generate a high temperature when gasified with oxygen, the released energy being used for sponge iron melting. Gasification of coals with a *high volatile* content such as lignite results in a low temperature because the volatile hydrocarbons must be cracked before gasification can occur.

Figure 128 shows the calculated adiabatic gasification temperature (gasification to $CO + H_2$) as a function of the H : C and O : C molar ratio of the coal. The higher the gasification temperature the better the coal is suited for smelting reduction process, i.e., the lower is the coal consumption. The classification of coals by the aforementioned molar ratio is more accurate than by volatile content because coals with the same volatile content can show remarkable differences in their gasification temperature. The coal rank classification shown in Figure 128 gives only a rough estimation.

For *converter-type melters* the degree of post-combustion must be increased when using high-volatile coal, but there are some limitations with respect to the movement of the bath. For processes with countercurrent flow of coal and gas, besides energetic limitation, problems with tar formation can also occur.

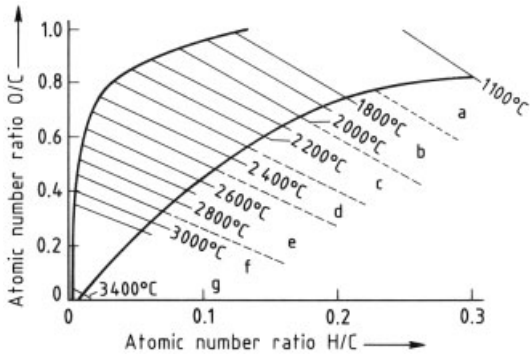

Figure 128. Adiabatic gasification temperatures ($H_2 + CO$) for coals in dependence on the atomic number ratio
a) Lignite; b) Subbituminous coal; c) High volatile bituminous coal; d) Medium volatile bituminous coal; e) Low volatile bituminous coal; f) Anthracite; g) Coke

The *ash content* of the coal is less critical than its volatile content. The mainly acidic ashes of the coal must be compensated for by additives so that basic slags can be formed. Thus, processes which require a high slag basicity are more sensitive to the ash content. The additives (limestone, dolomite) enter the melter stage in most cases in the calcined form. Otherwise the influence of the ash content on the coal consumption is too high and extra energy for calcining the additives and for the endothermic Boudouard reaction is needed.

The *moisture content* of the coal (ca. 10 %) decreases the gasification temperature drastically because of the endothermic water – gas reaction. Hence, most of the processes use dried coal.

Other properties of the coal such as the melting behavior of the ash, the swelling index, or the reactivity are of little or no concern. The Hardgrove index influences the grinding cost when grinding is necessary (converter processes).

The *sulfur content* of the coal should be of course as low as possible (≤ 1 %). However, the tolerable range must be evaluated for each smelting reduction process separately.

Table 36. History of smelting reduction processes from 1970 to 1990 [184]

Process	Ore type	Fuel	Prereduction (reduction degree)	Final reduction	Development stage	References
Char or coke bed melter – gasifier						
KR/Corex	lump	lump coal	shaft furnace (90 %)	melter	1000 t/d	[186 – 189]
Sumitomo SC	lump	coke (30 %) fine coal	shaft furnace (60 %)	cupola	8 t/d	[190, 191]
Kawasaki XR	fine	lump coal	circulating fluidized bed (40 – 80 %)	fluidized bed	10 t/d	[192, 193]
Converter with post-combustion						
HIsmelt (CRA)	fine	fine coal	fluidized bed	horizontal cylindrical converter (C-injection)	10 t	[194 – 196]
MIP and melting cyclone	fine	fine coal	hot and melting cyclone (30 %)	horizontal cylindrical converter (C-injection)	20 t	[197]
NKK	fine	fine coal	fluidized bed (15 %)	converter (C-top blowing)	5 t	[198, 199]
NSC	fine	fine coal	circulating fluidized bed (30 %)	converter (C-top blowing)	5 t	[200]
Kobe Steel	lump	fine coal, CH_4	shaft furnace	converter (C-injection)	0.5 t	[183]
Without post-combustion						
Coin (Krupp)	fine	fine coal	circulating fluidized bed (>90 %)	converter (C-injection)	3 t	[201 – 203]
CBF (British Steel/Hoogovens)	lump	fine coal	shaft furnace (>90 %)	two-stage converter (C-top blowing)		[183]
Process with electrical energy						
Elred (ASEA)	fine	fine coal	circulating fluidized bed	d.c. arc furnace	8 t/h melting, 0.5 t/h red.	[204, 205]
Inred (Boliden)	fine	fine coal	melting chamber	arc furnace	8 t/h	[206, 207]
Combismelt (Lurgi)	lump	lump coal fine coal	rotary kiln	submerged arc furnace		see Section 2.6.4
Plasmasmelt (SKF)	fine	coke fine coal	fluidized bed	shaft furnace	8 t/h melting	[208, 209]

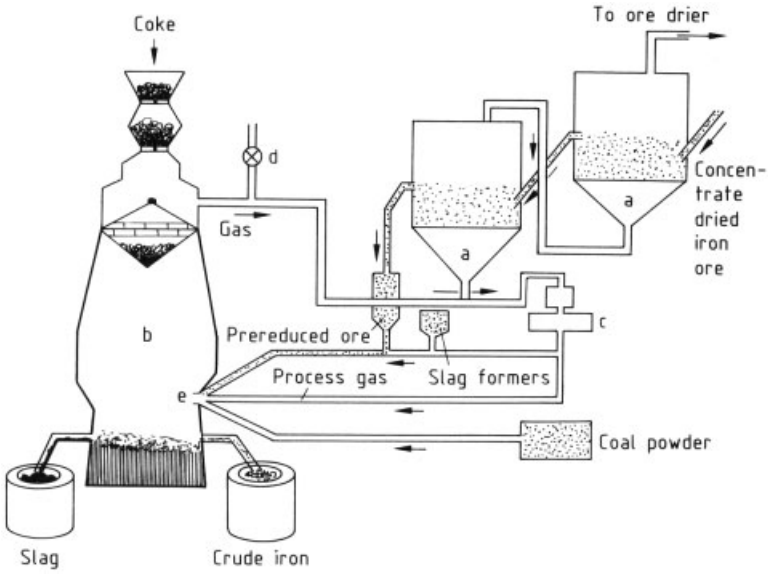

Figure 129. The Plasmasmelt process
a) Prereduction; b) Shaft furnace; c) Compressor; d) Pressure control; e) Plasma generator

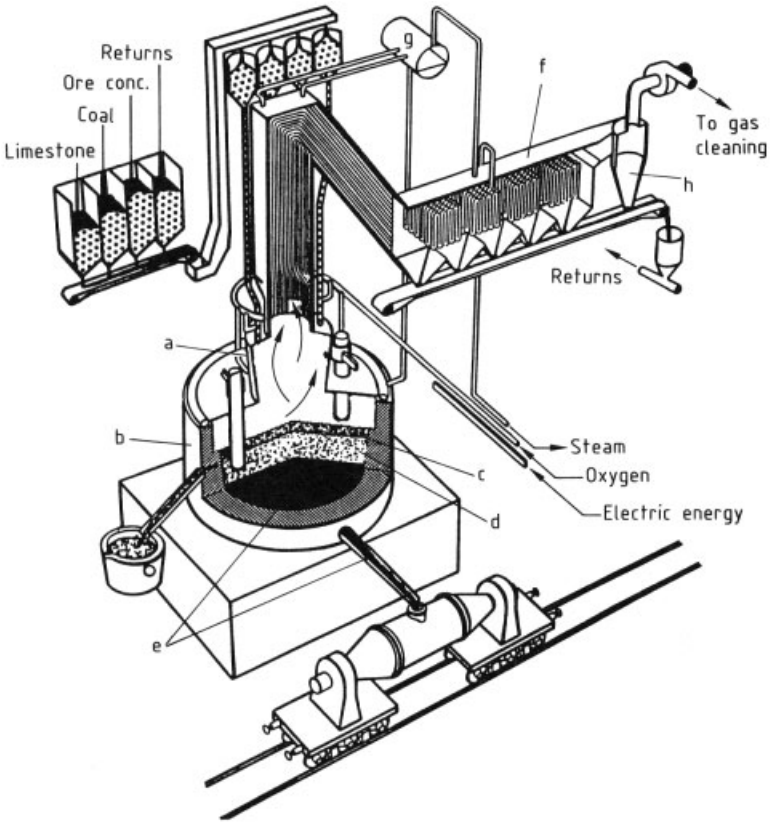

Figure 130. The Inred process
a) Flash-smelting chamber; b) Submerged-arc furnace; c) Coke bed with sponge iron and slag; d) Slag; e) Iron; f) Waste-heat boiler; g) Boiler drum; h) Cyclone

2.7.2. Classification of Smelting Reduction Processes and Earlier Developments

The first developments in smelting reduction on a laboratory scale failed in the 1960s because the conversion of coal and fine ore in a melting vessel was tried without prereduction [185]. The next development was in Sweden in the 1970s, where the energy was introduced into the melting vessel in the form of *electrical energy.* Further developments were the *converter processes* with and without post-combustion. Processes which are closest to the blast furnace are those with a char-or coke-bed melter – gasifier (see Table 36).

Processes which only modify the blast furnace technique such as the oxygen – coal blast furnace of BSC and NKK or the plasmaheated blast furnace Pirogas [183] are not discussed here.

In the case of processes using fine ore generally only the reduction reactor is developed. The integration of the fine-ore reduction into the process and therefore directly into the smelting reactor has not yet been realized in any of the existing pilot plants.

2.7.2.1. Processes Using Electrical Energy

These processes were developed in Sweden and consist of a prereduction stage of the fine ore and a final reduction which use electrical energy. The electrical energy is introduced via plasma burners (Plasmasmelt), electric arc (Elred), or electric resistance heating (Inred). The only process which has a technical application is the Plasmasmelt process in a modified form (Plasmadust and Plasmachrome). Because of the high electricity consumption these processes can only be consid-

ered for iron production in countries with cheap electricity.

The *Plasmasmelt process* [203, 208] was developed by SKF Steel Engineering AB. The flow sheet is shown in Figure 129. Fine ore is prereduced with a reduction gas generated in the smelting reduction stage. The prereduced material (reduction degree 50 – 60 %) is injected together with powdered coal into a coke-filled shaft furnace (melter) provided with plasma generators. Recycled process gas is used as injection gas. The final reduction and melting occur inside the coke column which also protects the refractory. A proportion of the spent reduction off-gases is combusted to dry and preheat the ore fines.

The Plasmasmelt process requires a low coke intake (consumption 50 – 100 kg /t hot metal), 1100 kWh of electricity and 200 kg coal per ton hot metal. The prereduction stage was developed separately in a 1 t/h pilot plant. Integrated operation of melting and reduction has not yet been tested.

A melter with a capacity of 8 t/h is integrated in a modified process (Plasma dust) where waste oxides from steelworks are processed in Southern Sweden.

The *Inred process* [206, 207] was developed by Boliden and tested from 1982 – 1984 in a pilot plant with a capacity of 8 t/h. Fine ore, coal, and fluxes are fed together with oxygen in a *flash-melting chamber* giving temperatures of 1900 °C. Superheated liquid FeO together with some coal char fall into the *electric furnace with submerged arc,* where final reduction takes place (see Fig. 130). Because of the high FeO content in the slag (7 – 8 %) a sulfur content of 0.15 – 0.2 % must be expected for a commercial plant.

In the *Elred process* [204, 205] fine ore is prereduced to 60 – 70 % in a single-stage fast-fluidized bed at 950 – 1000 °C before being melted down in an electric arc furnace (hollow Söderberg electrode). The reducing gas is produced by partial combustion of coal with air within the fluidized bed. The *prereduction pilot plant* was built by Lurgi with a capacity of 500 kg/h and the *electric arc furnace* by ASEA (capacity 8 t/h) (see Fig. 131). Up to now this process has found no technical application for iron production.

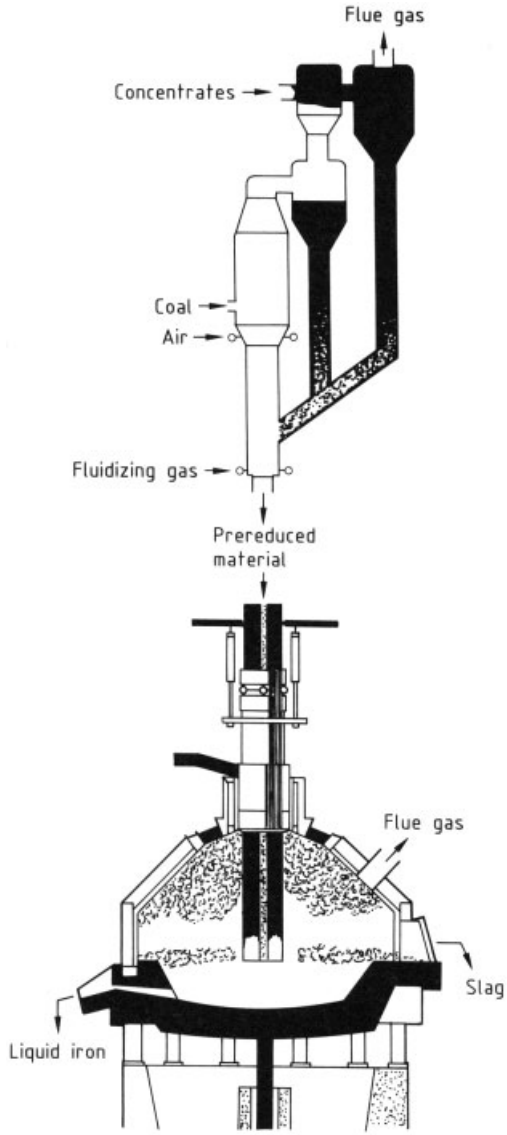

Figure 131. The Elred process

The *Combismelt process* combines the Lurgi prereduction rotary kiln and a submerged electric arc furnace (DEMAG). The electric energy is generated by waste heat and off-gas (see also Section 2.6).

2.7.2.2. Coal – Coke-Bed Melter – Gasifiers

In these processes the melting stage is a fixed bed or fluidized bed of charcoal or coke. Prere-

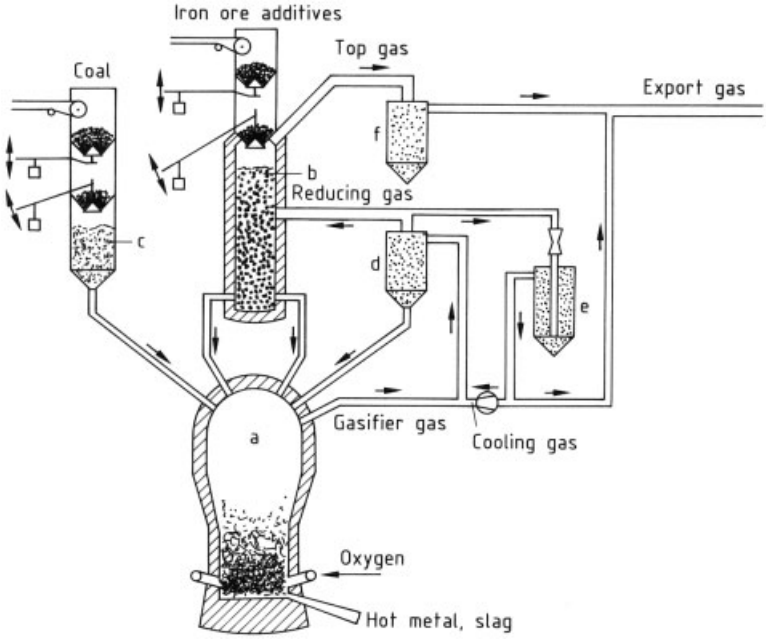

Figure 132. The Corex process, basic flow sheet with export gas
a) Melter – gasifier; b) Reduction shaft furnace; c) Coal feed system; d) Hot dust cyclone; e) Cooling-gas scrubber; f) Top-gas scrubber

duction is performed in a shaft furnace (Corex, Sumitomo) or fluidized bed (Kawasaki).

The *Corex Process* (formerly called KR process) is the most advanced smelting reduction process [186 – 189]. This process was developed by Deutsche Voest-Alpine Industrieanlagenbau (formerly Korf Engineering) and its parent company Voest-Alpine Industrieanlagenbau (VAI) in Austria to a 60 000 t/a pilot plant from 1981 to 1987 in Kehl, Germany.

The Corex process (Fig. 132) separates the iron-ore reduction and melting steps into two reactors:

1) Generation of reducing gas and liberation of energy from coal for melting occur in the *melter – gasifier.*
2) Reduction of iron ore occurs in a *shaft furnace.*

The Corex process is designed to operate under elevated pressure, up to 0.5 MPa.

Coal is fed into the gasifier (a) where it comes into contact with a reducing gas atmosphere at a temperature of ca. 1000 – 1100 °C. Instantaneous drying and degasification of the coal particles occur in this upper portion of the melter – gasifier. Reducing gas is generated in a fluidized bed, by the partial oxidation of coal. At first carbon is oxidized to carbon dioxide. Then, carbon dioxide reacts with free carbon to form carbon monoxide. The gas temperature in the fluidized bed is in the range of 1600 – 1700 °C. The temperature conditions in the freeboard zone above the fluidized bed guarantee production of a high-quality reducing gas which contains ca. 65 – 70 % carbon monoxide, 20 – 25 % hydrogen, and 2 – 4 % carbon dioxide. The remaining constituents are methane, nitrogen and steam.

After leaving the melter – gasifier, the hot gas is mixed with cooling gas to obtain a temperature suitable for direct reduction, ca. 850 – 900 °C. The gas is then cleaned in hot cyclones (d) and fed to the shaft furnace as reducing gas. A small amount of the cleaned gas is converted to cooling gas in a gas cooler. The fines captured in the hot cyclone are reinjected into the melter – gasifier.

The reducing gas is fed into the reduction furnace and countercurrently ascends through the iron burden.

The iron ore, charged into the shaft furnace through a lock hopper system, descends by gravity. The reduction reaction in the shaft furnace,

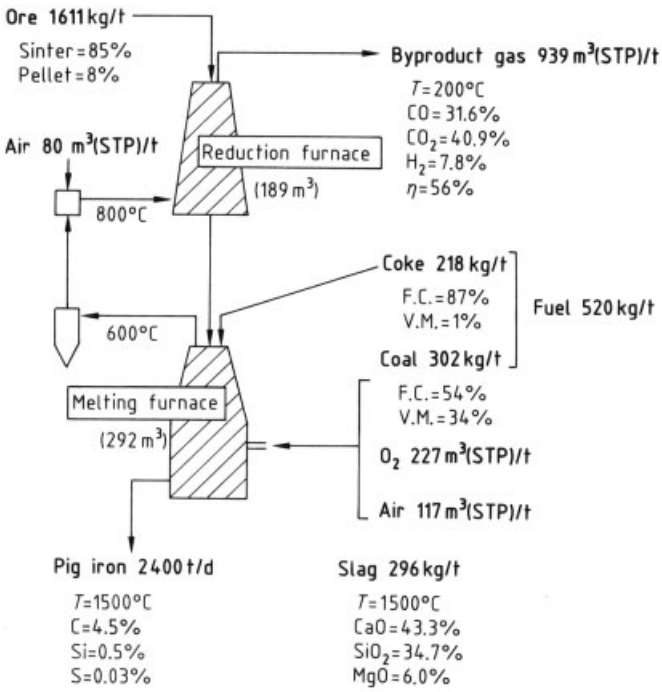

Figure 133. Expected material flow sheet for a Sumitomo SC commercial plant (F.C. = fixed carbon; V.M. = volatile matter)

using gas with ca. 70 % carbon monoxide and 25 % hydrogen is exothermic, leading to temperatures in the burden which are above the reducing gas temperature. Because of carbon monoxide decomposition, carbon deposits will form on the iron and act as a lubricant. Formation of Fe_3C occurs as well and cluster formation is avoided. The top gas is cleaned and cooled in a scrubber, and is then available for various purposes. Metallization of the DRI averages 90 – 95 %, and its carbon content is in the range of 3 – 5 %, depending on the raw material used and operating conditions.

The continuous transfer of the hot DRI (temperature 800 – 900 °C) from the reduction furnace to the melter – gasifier is carried out by a special controllable transport system which discharges into connected downcomers. The velocity of the descending sponge iron particles is lowered in the fluidized bed, further reduction is completed, heating and melting occurs. Hot metal and slag drop to the bottom of the melter –gasifier. Analogous to the practice in the blast furnace hot metal and slag are discharged by conventional tapping procedures. Tapping is carried out every 2.5 – 3 h on an average. For a good desulfurization a slag basicity >1 is required.

Temperature and silica content of the hot metal can be adjusted by various process parameters, i.e., coal particle size, height of the fluidized bed, system pressure, and rate of hot metal production.

The melter – gasifier may generate a certain amount of surplus gas depending on the coal selected. This surplus gas becomes part of the cooling gas stream, and can be either used separately or mixed with the top gas from the shaft furnace. When using a high volatile bituminous coal, the resulting gas mixture (export gas) approximates 1800 m^3 (STP)/t hot metal with a lower heating value of ca. 8000 kJ/m^3 (STP).

According to the composition and quality of the coal, the specific oxygen consumption is ca. 450 – 550 m^3 (STP)/t hot metal. The energy for the oxygen production can be supplied from about one third of the export gas of the Corex plant. The coal consumption depends on the coal quality and is ca. 0.5 – 0.7 t fixed carbon per ton hot metal.

If the export gas cannot be used economically, it is converted to reducing gas by removal of carbon dioxide and recycled. This process route offers the advantage of a lower coal and oxygen consumption.

The *Sumitomo SC process* [190, 191] uses a combination of two shaft furnaces (Fig. 133), the reduction furnace which produces DRI from lump ore and a *coke-bed melting furnace* which generates the reducing gas. Oxygen, air, and pulverized coal are blown into the coke bed. Low-grade coke is charged at the top of the melting furnace which corresponds to the lower part of a blast furnace or a cupola.

The gas coming out of the melting furnace with a temperature of 600 °C is cleaned in a cyclone, then it must be heated up by partial combustion with air to the reducing temperature of 800 °C.

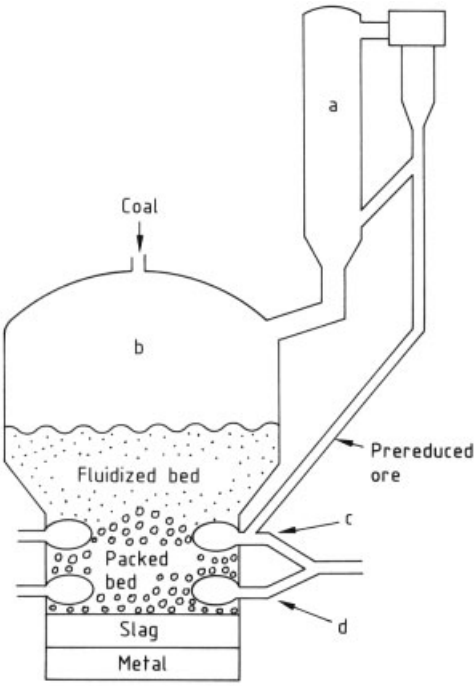

Figure 134. The Kawasaki XR process
a) Fluidized bed prereduction furnace; b) Coke-bed melter; c) Upper tuyere; d) Lower tuyere

The *melting furnace* was developed in a pilot plant with a capacity of 8 t/d in 1982. In 1984 the *reduction furnace* was integrated. Expected consumption figures (see Fig. 133) for a commercial plant are given on the basis of the result obtained by the pilot plant. The expected coke consumption of 218 kg/t hot metal (in the pilot plant much higher) is nearly as high as that of a modern blast furnace with a high coal-injection rate. So the goal of a cokeless process is not achieved.

Kawasaki developed a smelting reduction process for ferrochrome production on the basis of a *coke-bed melter* with two-stage tuyeres, the *XR process* [192, 193]. Powdered ore was prereduced in a fluidized-bed reactor and injected into the furnace through the upper tuyere. For iron production, this process was modified to feed coal from the top as schematically shown in Figure 134.

Small-size degassed coal forms a fluidized bed in the upper part of the furnace and the lumpy char forms a packed bed in the lower part. The pilot plant for the XR process with a capacity of 10 t/d was started in 1987. It is planned to reduce the fines in a circulating fluidized bed to a reduction degree of 80 %.

2.7.2.3. Converter-Type Iron Bath Reactors

The converter-type melters (see Table 37) are under development from the experience of the oxygen LD-BOF converter for steelmaking (→ Steel, Chap. 3.3.1.1). Coal is injected with or without oxygen in a liquid iron bath.

In most of the processes the carbon monoxide – hydrogen gas mixture coming out of the melt is partially combusted (post-combusted) to generate heat (up to 1900 °C) which is transferred to the bath. Processes using fine ore prereduce the ore in fluidized beds to a reduction degree of 15 – 90 %. In most cases the process pressure is low (ca. 50 kPa). At present only the melting stage is developed for most of the processes. *Integrated operation* of melting and reduction has not been tested. Apart from the latter the most critical areas of these processes are refractory lifetime, gas-handling system (accretions), and foaming and splashing of the melt. Because of the high FeO-content in the slag the sulfur content of the hot metal is higher and the silicon content lower than in the blast furnace process.

Processes with Post-combustion. The most advanced converter process has been developed

by Klöckner, Germany, and CRA, Australia, since 1981 [194 – 196]. After pretests a pilot 10 t-converter started operation in 1984 with a capacity of 4 t/h where especially post-combustion was investigated. The prereduction stage was not tested in this plant. In 1987 Klöckner was bought out and CRA formed a joint venture with Midrex to further develop the process under the name *HIsmelt.* The flow-sheet of the HIsmelt process in its first stage of development is shown in Figure 135. Coal is injected from the bottom into the melt. The energy for melting and final reduction is generated by partial combustion of the converter gas with preheated air. The converter off-gas is used for prereduction of the fine ore. Because of the high oxidation degree of the gas it is only possible to reduce the ore to iron(II) oxide which is separated in a cyclone and fed to the converter. The main problem of this process is whether the high post-combustion degrees, up to 65 % depending on coal quality, and reduction degree can be maintained for continuous operation and whether the huge amount of hot gas can be handled.

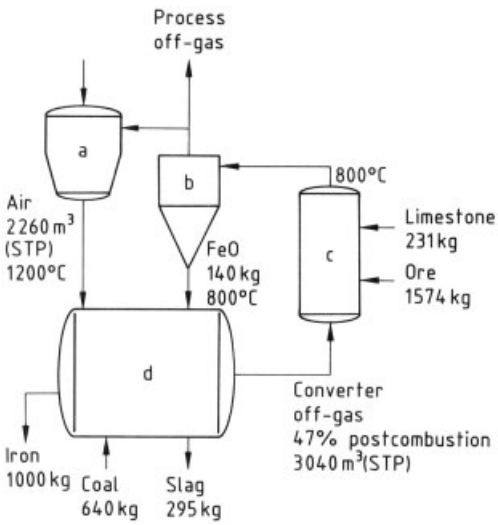

Figure 135. Flowsheet of HIsmelt process for the case 47 % post-combustion and reduction to FeO
a) Air preheaters; b) Cyclone; c) Prereducer; d) HIsmelt converter

The NKK process [198, 199] is another converter type process which also operates with a high post-combustion degree (see Fig. 136). Coal and oxygen are blown in from the top to the bath. The fine ore is prereduced only to a low degree.

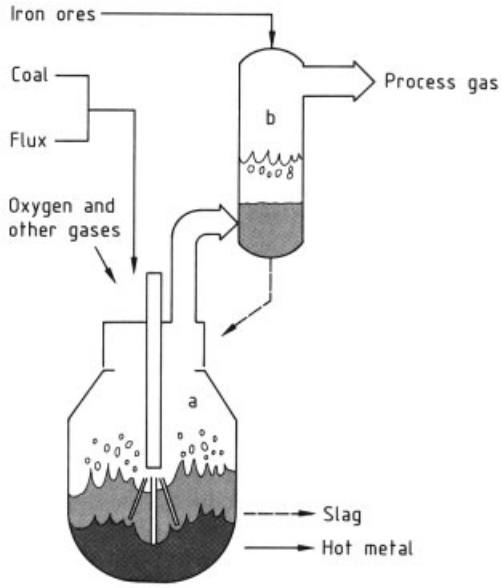

Figure 136. The NKK's smelting reduction process for iron-making
a) Smelting reduction furnace; b) Preheating and prereduction furnace

Since 1986 tests have been performed in a 5 t-converter. In 1988 a prereduction stage was integrated in the pilot plant with a capacity of 100 t/d.

Processes without Post-combustion. The *Coin process* [201 – 203] developed by Krupp operates without post-combustion and with a high reduction degree. In the first published flow sheets the prereduction stage was a shaft furnace. The latest one (see Fig. 137) shows a *multistage prereduction* of fine ore which was tested separately in a pilot plant. The prereduction stage is a combination of entrained-bed and fluidized-bed reactors.

The converter pilot plant was a 3 t-converter. To achieve a high degree of metallization a high reduction temperature is needed. Sticking of the ore is prevented by carbon monoxide decomposition in a prereduction stage at 500 – 600 °C. Oxygen and degassed coal is blown into the bath from the bottom. As no post-combustion delivers energy, the coal (especially high volatile coal) must be preheated and charred by inject-

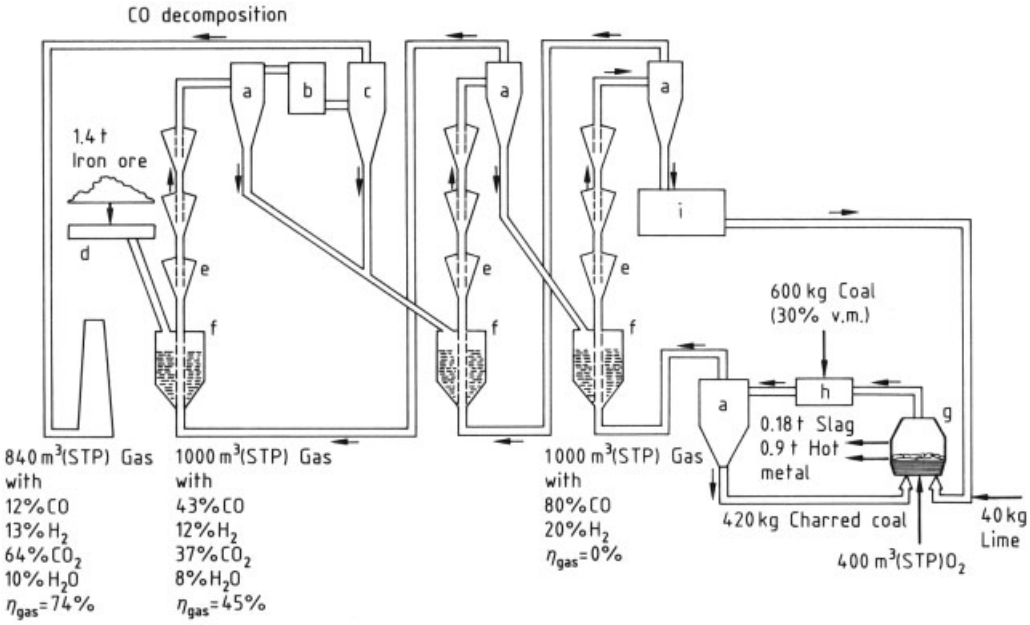

Figure 137. Coin concept of a coal-based smelting reduction process for iron ore fines
a) Hot cyclone; b) Gas cooler; c) Bag filter; d) Screw conveyor; e) Whirl box; f) Pneumatic conveyor; g) Melter gasifier; h) Charring; i) 1 t directly reduced iron and 85 kg carbon

ing in the converter off-gas and recycling via a cyclone. The process is very complex and integration of the multistage prereduction, the coal devolatilizer, and the converter is not easy.

2.7.3. State of the Art of Process Developments

The current situation of smelting reduction processes is characterized by further developments of melter – gasifier and iron bath reactors. Furthermore, a new third group of processes is emerging: the combination of rotary hearth furnace (RHF) and submerged arc furnace (SAF). In 1999 only the Corex process and a first RHF/SAF process are in industrial operation.

2.7.3.1. Melter – Gasifiers

The *Corex Process* (see Fig. 132) – after producing hot metal in the first operational C-1000 plant in Pretoria from 1988 to 1998 – is now operating in four industrial C-2000 plants, producing 2000 t/d hot metal per plant [219].

The Corex plant of POSCO (Korea) in Pohang started production in 1995. A hot metal production of 95 t/h and consumption figures for coal of 995 kg/t hot metal, for oxygen of 527 m^3/t hot metal as well as an availability of 94 % are achieved as representative average values.

The first hot metal of modul no. 1 of the two Jindal Corex plants at Jindal Vijayanagar Steel (India) was tapped in October 1998. The hot metal is refined in a basic oxygen furnace (BOF) steelmaking shop.

The Corex plant of Saldanha Steel (South Africa) started production in 1999. The production is based on 100 % lump ore charge; the annual capacity is 650 000 t. The hot metal is charged together with DRI into a Conarc furnace which combines the possibilities of a converter (BOF) with those of a electric arc furnace (EAF).

An economical necessity of the Corex plant operation is the utilization of the surplus gas to get a financial credit in respect of the production costs. The export gas of the Saldanha plant feeds a Midrex direct reduction plant with a production capacity of 800 000 t/a of DRI. In the case of the Jindal Corex plants the export gas is used

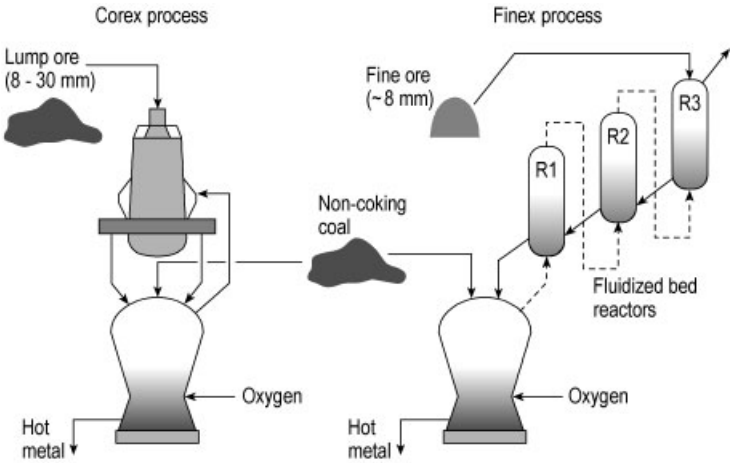

Figure 138. Concept of the Finex process compared to Corex

for power generation as well as for the energy supply of a pelletizing plant.

Further developments of the Corex process have been started by POSCO and VAI since 1992, aiming at the direct use of fine ores.

Fines injection into the Corex process is an important feature of a waste-free integrated plant concept without a sintering plant. Operational results are available for fine ore injection of about 5 % which has not negatively affected the Corex operation [220]. Further tests are directed to a fine ore addition of 20 to 30 %. Not only utilization of purchased fine ores and fines separated from the Corex plant, but also the recycling of iron-containing fines from steel works is considered.

The *Finex* development [221] will allow the use of 100 % fine ore charge. A three-stage fluidized-bed reactor is combined with the melter – gasifier (see Fig. 138). Fine ore is preheated (R3), prereduced to 20 to 50 % (R2), and then reduced to about 90 % (R1) before being charged into the melter – gasifier in hot condition. A Finex pilot plant for a daily production of 150 t was commissioned in 1999; this pilot plant is connected directly to the existing Corex C-2000 plant of POSCO. So a part of the reducing gas generated in the melter – gasifier can supply the fluidized-bed system of the Finex pilot plant. Test campaigns are scheduled to be performed until end of 2000 in order to acquire the operational and engineering data for commercialization [222].

2.7.3.2. Iron Bath Reactors

From the iron bath reactors mentioned in Section 2.7.2.3 only the HIsmelt and the NKK process (now DIOS process) have been further developed; in addition new processes have emerged. For all processes mentioned a commercial status has not been reached yet.

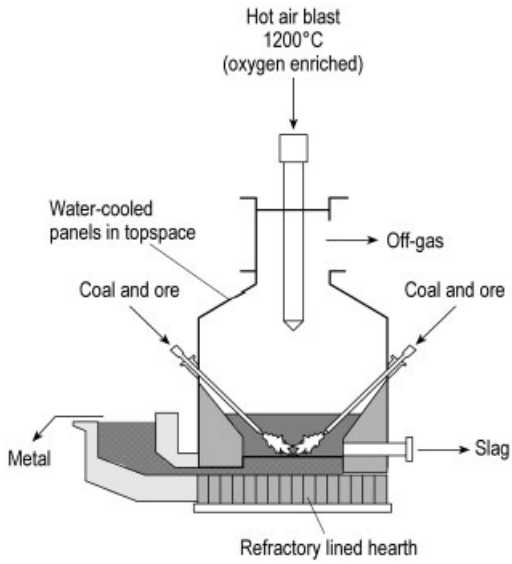

Figure 139. New HIsmelt vertical reaction vessel

HIsmelt. In 1997 Rio Tinto became the parent Company of HIsmelt Corp., Kwinana (Australia). The HIsmelt process was changed from

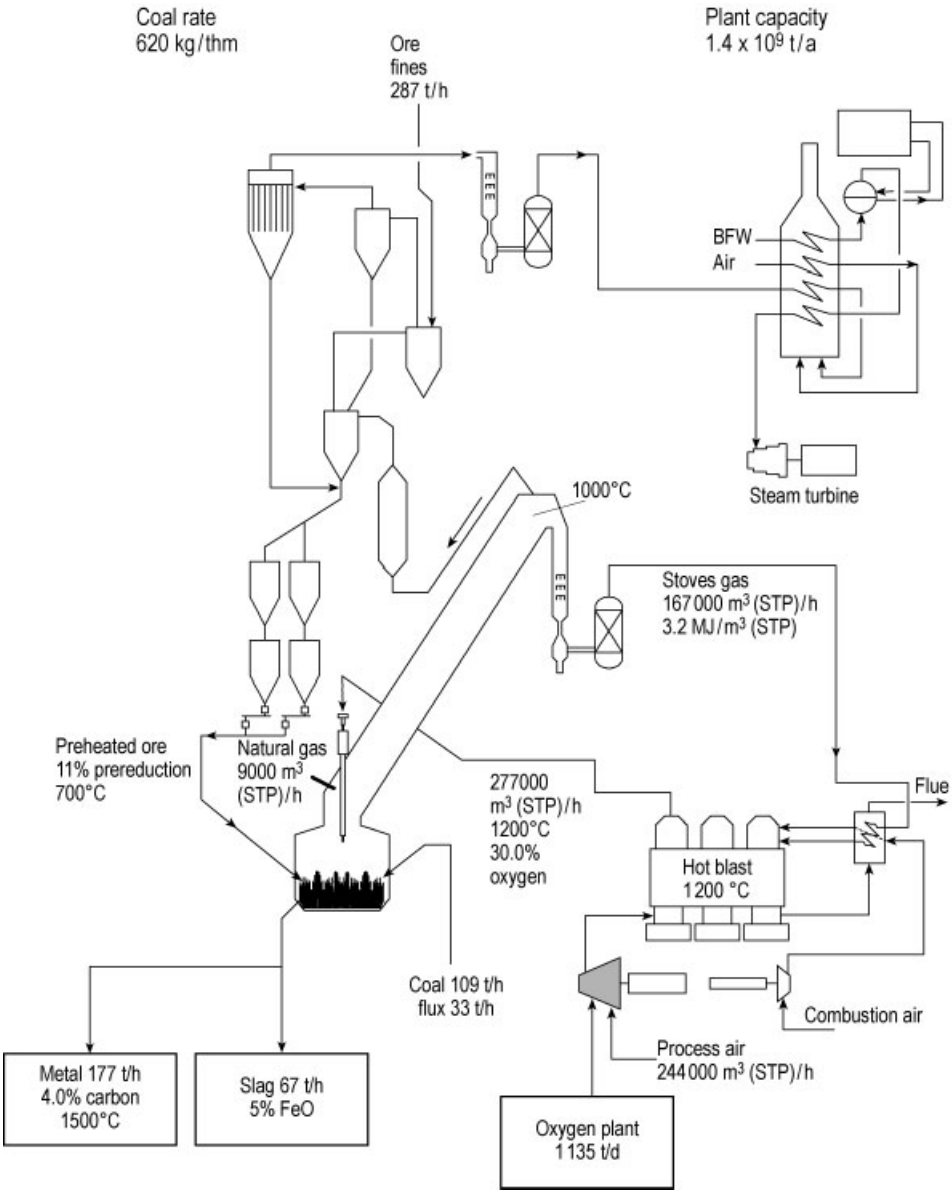

Figure 140. Lay-out of an industrial HIsmelt plant (8 m diameter vertical smelting reduction vessel)

a drum-like horizontal reactor to a vertical vessel (see Fig. 139), which was commissioned in 1997. Fourteen test campaigns have been conducted, involving direct injection of fine ore, steel plant reverts, and coal into the metal bath of the vertical water cooled smelting reduction vessel with 2.7 m inner diameter. Preheated air at 1200 °C, enriched with oxygen (27 or 30 % O_2), is injected through a water-cooled lance for post-combustion of CO and H_2 coming from the bath. The turbulent nature of the process is very effective at transferring this energy back to the metal bath. The off-gas is used to fire the stoves which preheat the hot air blast [223, 224]. Between August 1997 and April 1999 the plant was operated for 128 days with the longest continuous run being a 38 days campaign. With a degree of post-combustion of 60 to 80 % and cold iron

ore fines as ferrous feed, coal rates of around 880 kg/t hot metal have been recorded. Refractory wear rates less than 1 kg/t hot metal have been achieved. Next step is a scale-up to a 6 m diameter smelter for the production of 500 000 t/a. In Figure 140 a plant lay-out for producing 1.4×10^6 t hot metal per year is given, including fine ore charging, prereduction in a circulating fluidized bed and smelting in a 8 m diameter vessel.

DIOS. The NKK process (see Fig. 136) was the basis of the Japanese joint research program DIOS (Direct Iron Ore Smelting Reduction), which was promoted by the Japan Iron & Steel Federation in collaboration with the Center for Coal Utilization from 1988 to 1997. The pilot plant at Keihin Works of NKK had a 500 t/d nominal capacity. Test campaigns with this pilot plant were carried out from 1993 to 1996. The process flow sheet of the DIOS plant is illustrated in Figure 141. The plant consisted of an in-bath smelting reduction vessel and two bubbling fluidized-bed type preheating and prereduction furnaces. The smelting vessel was equipped with an oxygen blowing lance and several bottom tuyeres for bath stirring [225]. Water cooled panels at the vessel enable a high degree of post-combustion in the foaming slag. A minimum coal consumption of 780 kg/t hot metal was achieved for a low volatility coal and of 980 kg/t for a high volatility coal, with prereduction degrees of about 25 % at 750 °C and an oxygen consumption of about 750 m^3/t hot metal. The results were used to perform a feasibility study and the conceptual design of a commercial plant for a production of 6000 t/d. However, it is expected that the situation in Japan with high-performance hot metal production in blast furnaces will not allow commercializing a DIOS plant before 2010. Therefore the activities focus on the commercialization abroad to maintain the technologies in the meantime.

CCF. The cyclone converter furnace (CCF) process is characterized by melting and prereduction of fine ore in a cyclone on top of a converter-type reactor. The prereduced liquid iron oxide drips into the carbon-containing melt, where final reduction takes place. The emerging CO is post-combusted with oxygen, partly in the converter and partly in the cyclone. The CCF process was initially a development project of the European steel producers Hoogovens, British Steel, and Ilva, sponsored by the European Coal and Steel Community (ECSC). Later two different developments were undertaken by CSM, the Italian Steel Research Center in Rome, and by Hoogovens (Netherlands). In the middle of the 1990s CSM performed some trials with a small pilot plant (5 t/h) in Taranto; the process was then named *CleanSmelt*.

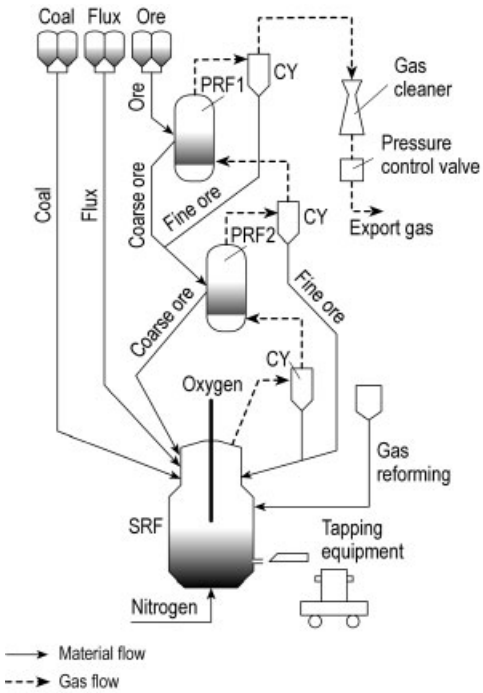

Figure 141. Process flow sheet of the DIOS pilot plant
CY= Cylone; PRF = Prereduction furnace; SRF = Smelting reduction vessel

Hoogovens performed tests with the cyclone part of the CCF process (20 t/h) [226] and took over the know-how of a former direct smelting project in a converter-type vessel of the "Direct Steelmaking Program" undertaken in a collaborative research by the American Iron and Steel Institute and the Department of Energy (AISI/DOE) in the USA [227]. The CCF process scheme is illustrated in Figure 142 [228]. The consumption figures calculated for producing 1 t of hot metal are given as follows: 1540 kg fine ore, 750 kg coal, 200 kg limestone and 720 m^3 oxygen. Hoogovens performed a feasibility

study for building a demonstration plant with a capacity of 700 000 t/a, until now no decision has been made for the investment.

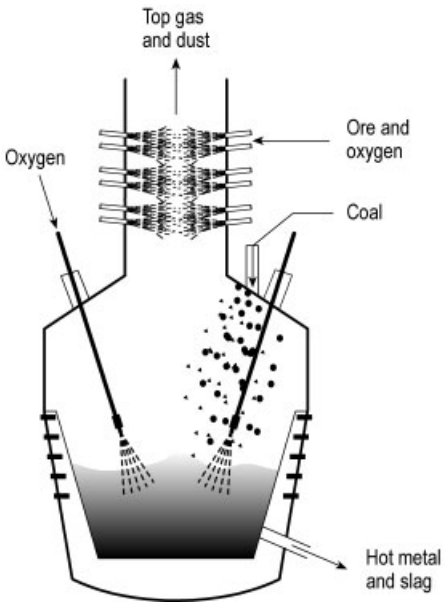

Figure 142. CCF process scheme

AusIron. In 1981 the Ausmelt Ltd. was formed in Australia to develop and market the technology of CSIRO and of the Siromelt process (→ Copper, Chap. 5.6). The Ausmelt technology is applied to commercial production of nonferrous metals. The Ausmelt *AusIron* technology includes a bath smelting system, which has been developed by pilot tests since the beginning of the 1990s. The heart of the technology is a top submerged lance which supplies fine coal, air, and oxygen to a slag bath reactor where iron oxide is reduced. The smelting of iron bearing titania residues has been performed at Ausmelt's Melbourne pilot facility [229].

Romelt. The Romelt process was developed by the Moscow Institute of Steel and Alloys in Russia in the late 1970s. It is a single stage smelting reduction process producing liquid iron by reducing ferrous materials without preparation. A pilot plant (0.3×10^6 t/a) has operated intermittently at the Novolipetsk Iron and Steel Works since 1985, licensees for the process are Nippon Steel Corp. (Japan), Kaiser Intern. (USA) and SAIL-Romelt (India) [230, 231]. In the Romelt process the liquid iron is produced in a molten slag bath reactor using noncoking coal and oxygen. The process requires both post-combustion of about 70 % of the primary gases from the bath and high heat transfer efficiency. Oxygen is introduced into the rectangular reactor at two different levels through water cooled tuyeres. The lower level tuyeres, submerged to the slag, provide oxygen enriched blast for gasification of carbon and bath agitation while the upper row of tuyeres supply pure oxygen achieving a high degree of post-combustion. Metal and slag are individually discharged through opposite ends of the reactor. Calculated specific consumption figures per ton of hot metal for a commercial plant are 1.23 t of coal, 0.53 t of limestone, 800 to 900 m^3 O_2, 215 m^3 compressed air, 0.056 t fuel oil and 170 kWh electric power related to a productivity of 37 t hot metal per hour. Studies by Tata Steel are available about the economics of a commercial Romelt plant at several locations in India.

2.7.3.3. Rotary Hearth Direct Reduction – Submerged Arc Furnace Combinations

Production of high quality flat steel grades in electric arc steelmaking shops, based mainly on scrap, needs addition of some virgin iron to the EAF charge. In this context a new route to produce hot metal or pig iron is emerging: Coal-based direct reduction of green pellets on a rotary hearth furnace and subsequent melting of the hot DRI pellets in a submerged arc furnace. Compared to the Combismelt process (Section 2.7.2.1) in this new line the rotary kiln is replaced by a rotary hearth furnace which has a much more higher specific productivity. However, RHF units also have a limited capacity of only about 0.7×10^6 t/a.

The first industrial plant working since 1999 in this way is the *Iron Dynamics Process* of Steel Dynamics Inc. in Butler (IN, USA). Green pellets, i. e., pellets which are not hardened by firing, are prepared by mixing of iron ore concentrate and fine coal. The DRI from these dried green pellets produced on the RHF has a carbon level of 3.5 to 4 %. In the SAF the DRI is molten down to hot metal and slag by resistance heating with electric energy using submerged electrodes

in the liquid slag. The production capacity of the plant is 500 000 t/a [232].

Processes following this route have also been developed by other engineering companies. Mannesmann Demag/Demag Italimpianti (now SMS Demag) has presented a process based on the Inmetco rotary hearth furnace (see Fig. 143). The new process line is named *Redsmelt* [233].

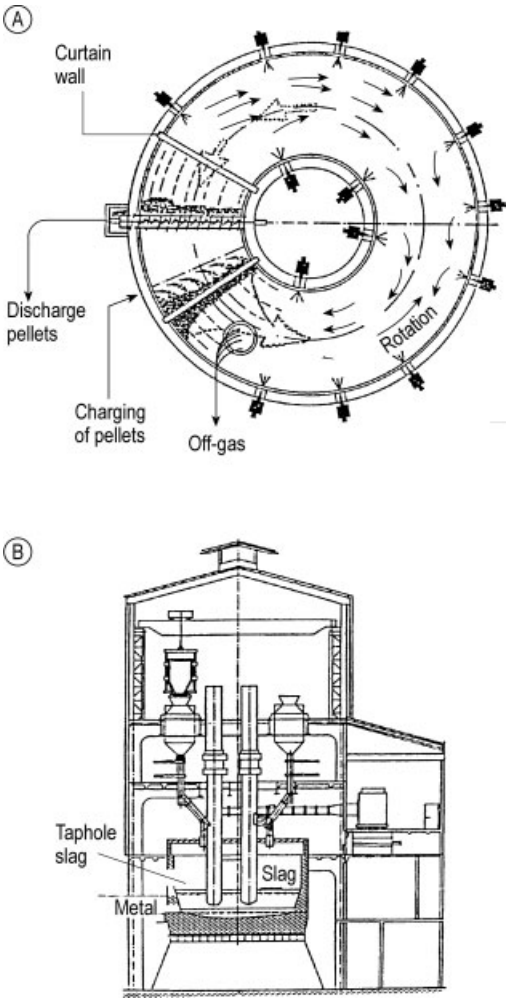

Figure 143. Core elements of the Redsmelt process a) Rotary hearth furnace; b) Submerged arc furnace

The plant construction company Midrex started development of the process line *Fastmelt*, based on the Fastmet rotary hearth furnace. A further development has been started by Kobe Steel in 1997 with its *ITmk3 process* (Iron Technology Mark 3), developed together with Midrex Corp. Based again on the Fastmet rotary hearth furnace, this process aims at a complete separation of liquid metal and liquid slag during reduction [234]. The new process uses the same raw materials and reductants as Fastmelt, but the reduction mechanism and products are completely different. The reduction occurs at higher temperatures (1350 – 1500 °C) and shorter residence times than required for Fastmelt. The heating of the gangue and ash components leads to softening and subsequent initiation of the coalescence of slag droplet. Hollow iron shells are formed, followed by forming iron droplets completely separated from the slag. Iron nuggets with about 1.8 % C were produced with an iron content of 97 to 99 %. Pilot plant operation is being planned to confirm successful transfer of the experimental knowledge to the ITmk3 process. The *Sidcomet* process is a development of Sidmar (Belgium) to produce hot metal from DRI using an RHF. This process is in the demonstration stage [235]. It is based on the Comet RHF process developed by CRM and Sidmar, working with a mixture of iron ore, limestone, and fine coal. It was decided to build two production plants with a total capacity of 1.4×10^6 t/a.

2.7.3.4. Special Developments for Processing of Steelworks Dusts

Steelworks dusts contain relatively high contents of iron oxides and also nonferrous elements such as zinc and lead which could be enriched and recovered. Some of the aforementioned smelting reduction processes have been successfully tested for the recycling of steelworks dusts, for example HIsmelt, DIOS, Romelt, or the AISI/DOE direct steelmaking development in the USA.

New process developments are announced by some Japanese steel producing companies. Based on experience with its coke bed reactor (see Section 2.7.2.2) [193], Kawasaki Steel has developed a new smelting reduction process with a coke packed bed for simultaneous recovery of zinc and iron from EAF dust. A commercial plant to recover chromium and nickel with a dust treatment capacity of 230 t/d was started up in 1994 and has been successful operating since.

In a second stage the technology was applied to recycling of EAF dust. The key technologies were established through pilot plant tests with a scale of 10 t/d in 1996. The process uses two-stage tuyeres and employs direct feed of the raw material, without agglomeration. Fine raw materials are injected through the upper tuyere, heat is supplied from the lower tuyere. Thus, a high temperature and intensively reducing conditions prevail in the region between the tuyeres. The raw materials are fused immediately in a high temperature raceway. The concept of the process is illustrated in Figure 144. The coke rate was 1260 kg/t of dust. The Zn content of the recovered materials was stable above 60 % for EAF dust injected with about 27 % Zn; the contents of zinc in hot metal and slag were 0.005 and 0.1 % [236].

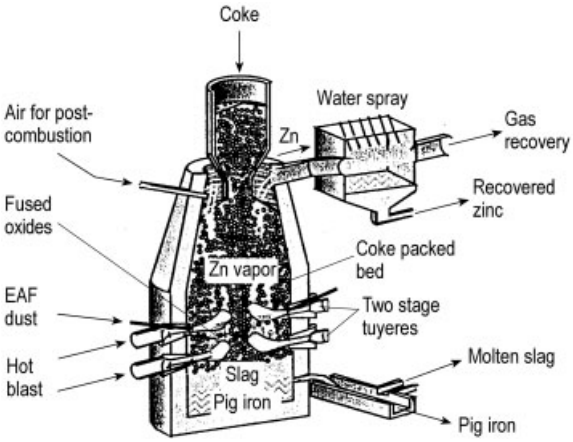

Figure 144. Kawasaki Steel concept of the smelting reduction process for EAF dust recycling

Daido Steel has developed a new dust and slag treatment process called *DSM* (Daido Special Method of Dust Slag Melting). In the DSM process EAF dust is melted together with reduced slag with a special burner to get oxidized slag and zinc concentrate. The oxidized slag obtained is used as road construction material; the secondary Zn concentrate is sold as raw material for Zn smelting. The process scheme is shown in Figure 145. A demonstration plant for a melting rate of 9 t/h started its operation in the Chita Plant of Daido Steel in 1996. Since 1999 100 % of the EAF dust of this plant is treated by the DSM process with a treatment quantity of 1500 t dust and 300 t reduced slag per month. The secondary dust of the DSM process contains more than 50 % zinc whereas about 2 % Zn and < 0.02 % Pb remain in the DSM slag [237].

2.7.4. Conclusions

Various attempts have been made for more than 30 years to develop a smelting reduction process that produces hot metal by direct processing of *fine ore* and *noncoking coal.* However, no process development that meets these requirements could be transferred to industrial application. The only important industrial process so far is the Corex process, reducing lumpy iron oxides in a shaft furnace in a first stage and processing the hot DRI in a melter – gasifier in a second stage to produce hot metal. The production of Corex plants reach about 800 000 t/a.

Direct fine ore processing with coal is enabled by the iron bath process developments. Some of these processes seem to be technical feasible (for example HIsmelt, DIOS, CCF), but their production costs for hot metal do not allow to finance such an investment if the costs have to be compared with those of existing high performance blast furnaces.

A new process line, tailored for electric steelmaking shops producing high-quality flat products, is emerging. These processes produce hot metal by combination of ore reduction with coal on a rotary hearth furnace with subsequent melting of the produced DRI in a submerged arc furnace (Iron Dynamics, Fastmelt, Redsmelt, Sidcomet).

Another new field of application for smelting reduction processes is the recycling and processing of iron-containing steelworks wastes.

3. Production of Pure and High-Purity Iron

Pure iron is required in large amounts in powder metallurgy and magnet technology. *High-purity* iron (definition, see page 158) is almost exclusively used in basic research. In general, the production of iron of a higher degree of purity does not start directly with oxidic ores, but with conventionally produced iron. However, special requirements with regard to

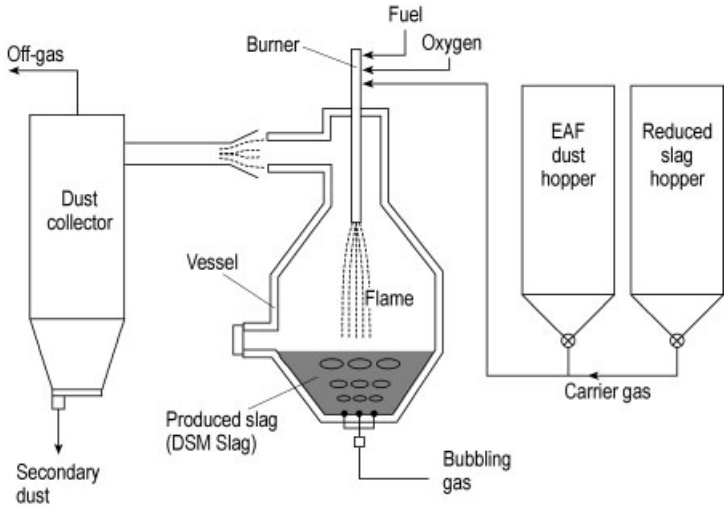

Figure 145. DSM process of Daido Steel in the Chita Plant

the limiting concentrations of certain impurities (e.g., copper) have to be met by the production method.

Production of Pure Iron. Pure iron is currently produced by four different processes:

1) Vacuum treatment of liquid steel
2) Electrolysis of aqueous electrolytes, usually using iron anodes that are produced by vacuum treatment of steel (see above)
3) Conversion of iron to iron carbonyls and subsequent dissociation of these compounds (→ Iron Compounds, Chap. 3.5)
4) Conversion of iron to iron chloride, subsequent purification and reduction with hydrogen.

Attempts were made as early as the turn of the century to recover "a material with the chemical purity and uniformity of early puddle iron" [238]. It was hoped that the superior corrosion resistance of puddle iron would again be achieved. The American Rolling Mill Comp., Middletown, Ohio (USA) was able to produce iron of high purity, known as *Armco Iron,* by using the conventional Siemens – Martin process (→ Steel). In the literature, the name Armco Iron is often synonymous with low-carbon iron that is produced pyrometallurgically and has a very low content of other accompanying substances. However, iron of this type is offered today by many producers and is made by the aftertreatment of well purified steel under vacuum conditions.

Depending on the origin of the liquid starting material, the *vacuum treatment* of liquid steel (→ Steel) used for the production of pure iron can remove (1) dissolved nitrogen and hydrogen (starting material: electrolytic iron), (2) carbon by reaction with oxygen (starting material: Siemens – Martin electrolytic steel or basic oxygen steel), and (3) occasionally even oxygen by reaction with carbon or hydrogen. The treatment also promotes separation of oxides of certain companion elements suspended in the liquid metal. These companion elements have a higher oxygen affinity than iron and were oxidized with oxygen in the preceding purification step. Companion substances such as copper, that are not oxidizable in the presence of metallic iron, cannot be removed by vacuum treatment. Therefore, the starting material used for this production process should not contain such substances.

The technically realizable pressure decrease is sufficient to remove gases dissolved in iron, such as nitrogen and hydrogen. Oxygen, however, can only be abstracted by using the following reactions:

$$[\mathrm{O}] + [\mathrm{C}] \rightarrow [\mathrm{CO}] \text{ or } [\mathrm{O}] + 2\,[\mathrm{H}] \rightarrow \{\mathrm{H_2O}\}$$

and by removing the resulting reaction products, carbon monoxide and water vapor. The purpose of vacuum treatment is usually decarburization.

Since the equilibrium constant

$$K_{\mathrm{C}} = \frac{[\mathrm{C}]\,[\mathrm{O}]}{p_{\mathrm{CO}}} \text{ (Vacher–Hamilton)}$$

Table 37. Analyses of pure and high-purity iron [240] (impurity content in ppm)

	Vacuumtreated, or Armco iron	Carbonyl iron	Electrolytic iron	High-purity iron of Alloy Corp.
Al	200 – 300	nd *	<200 – 800	1
Si	100	na **	50 – 500	1
Be	nd *	nd *	nd *	nd *
Co	600 – 800	na **	800 – 1000	<0.5
Cr	200 – 800	na **	<200 – 1000	3
Cu	4000 – 4600	na **	80 – 100	0.2
Mn	1600 – 6000	na **	~1000	0.1
Mo	200 – 400	na **	~200	<0.6
Ni	1600 – 2000	na **	<500 – 2200	0.2
Sb	200	nd *	nd *	nd *
Sn	600 – 900	nd *	~100	<1
Ti	50 – 100	nd *	nd *	<1
V	nd *	nd *	<300	<1
As	600 – 1200	na **	nd *	nd *
C	1000 – 1500	~70	200 – 600	nd *
H	~30	nd *	nd *	nd *
N	300 – 400	nd *	nd *	nd *
O	6600 – 7000	<1000	~2000	nd *
P	300 – 500	na **	100 – 500	nd *
S	1700 – 2500	na **	<400	nd *
Mg	nd *	nd *		0.4

* Not determined.
** Not detectable analytically.

at p_{CO} = 1 bar and 1620 °C is 0.0025 (see Fig. 146), the carbon content cannot be lowered below 0.05 % even at oxygen concentrations of 0.05 %. However, Figure 146 also shows that decarburization at a carbon monoxide pressure (total pressure or partial pressure) of 0.1 bar and with comparatively low oxygen concentrations leads to lower carbon contents. The influence of the temperature on K_C is comparatively small. Table 37 gives analytical values for pure iron produced by vacuum treatment.

Electrolytic iron, with a residual concentration of typical impurities shown in Table 37, can be produced with an expenditure of energy of 2 kWh/kg of iron (process of Husqvarna Vapenfabriks AB, Sweden) [241]. Low-carbon soft iron is employed as the anode material. The electrolytes used are, e.g., ammonium iron(II) oxalate, iron(II) chloride and iron(II) sulfate, with supporting electrolytes and possibly organic wetting agents, brighteners, and depolarizers as additives (→ Electrochemical and Chemical Deposition) [242]. The sole commercial application for electrolysis is in refining or coating. At present, the production of pure iron by electrolysis of an ore leachate in cells with inert anodes is not economical.

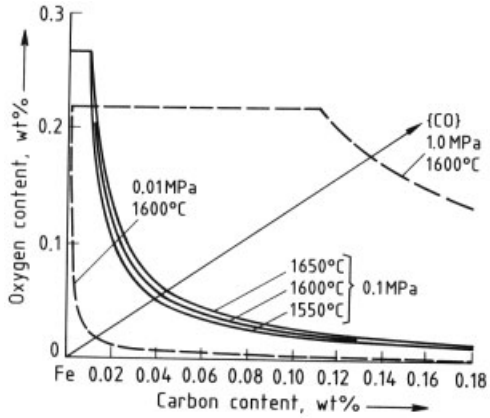

Figure 146. Equilibrium between oxygen and carbon, both dissolved in liquid iron

Electrolytically refined iron can be separated as a brittle, coherent cathode coating and marketed in pieces of ca. 10 mm in diameter. Further processing usually involves melting the iron in vacuum induction furnaces and possibly alloying. However, this iron can also be deposited on the cathodes in such a way that it breaks down to a powder when subjected to slight mechanical stress. This powder is an excellent starting material for sintered products. The special prop-

erties of the powder (such as high compressibility or high porosity at high values of breaking elongation) can easily be adjusted to definite requirements by varying the process parameters of electrolysis. The procedure for the production of electrolytic iron powder is as follows. After washing and careful grinding (with as little mechanical deformation of the powder particles as possible), the cathode deposits are annealed in a hydrogen atmosphere (reduction) to make the material ductile and to remove oxygen absorbed during the preliminary treatments. The powder is subsequently sized with sizing screens. The particle-size distribution desired for a particular purpose is obtained by mixing the required amounts of different particle fractions.

The concentrations of impurities contained in *carbonyl iron* are usually below the analytical detection limit, which of course depends on the analytical method employed (see Table 37). For production of carbonyl iron see → Iron Compounds, Chap. 3.2. Only high-purity iron contains still less impurities, which are detectable with extremely sensitive physical analytical methods.

Another large-scale process involves the use of *iron(II) chloride,* which is formed by dissolving sponge or scrap iron in hydrochloric acid. Iron(II) chloride is then purified, dried, and reduced with hydrogen at elevated temperature [243]. Hydrogen, in turn, is released during the dissolution of scrap iron and the hydrogen chloride formed during reduction is reused for dissolution. The iron obtained by using this method is almost as pure as high-purity iron.

Production of High-Purity Iron. High-purity iron is usually required in small quantities only and can be produced by *zone melting.* The starting product is either electrolytic iron or iron (protective gas is usually argon or argon – hydrogen mixtures). The zone purification process alters the ductility and the brittleness of the iron at very low temperatures as well as the oxygen solubility [244].

The *sublimation process* of Alloyd Electronics (United States) involves the chlorination of electrolytic iron, sublimation of the vaporized chloride ($FeCl_3$), and subsequent reduction of the sublimate—taken from the sublimation zone with the lowest degree of impurity—with hydrogen to give iron (analytical values are presented below) [245].

The use of chemical analyses is only conditionally suitable for a *comparison of the degrees of purity* of various types of high-purity iron because the concentrations come close to the detection limit of even the most sensitive methods. For this reason, the ratio of the specific electrical resistance ϱ at 300 K and at 4.2 K is frequently used as the criterion for purity. This ratio increases with even the slightest increase in the total amount of impurities. The degree of purity of pure and high-purity iron expressed as the ratio of the specific resistance at 300 K and at 4.2 K ($\varrho_{300}/\varrho_{402}$) is as follows [245]:

Iron produced in the Siemens – Martin furnace	
rolled	18
annealed, 300 °C	28
Vacuum treated, conventionally produced soft iron	29
Electrolytic iron, annealed	190
Electrolytic iron	
from carbonyl iron	207
from highly pure iron salt solution	240
Zone-melted iron	189
Alloyd iron	383

4. Aspects of Environmental Protection

In the recent past, the rapidly increasing environmental consciousness has led to considerably tightened environmental standards in many countries. In the Federal Republic of Germany, for instance, a limiting value for dust emissions of 150 mg/m^3 (STP) was stipulated as early as 1974 in the TA Luft for the prevention of air pollution. The amendment to the TA Luft in 1986 further reduced dust emissions in waste gases to 50 mg/m^3 (STP) [246]. This limit applies to *total* dust emissions. Lower limits have been set for *individual* components, such as heavy metals, e.g., 0.2 mg/m^3 (STP) for cadmium, mercury, and thallium; 1 mg/m^3 (STP) for arsenic, cobalt, nickel, selenium, and tellurium; and 5 mg/m^3 (STP) for antimony, lead, chromium, manganese, vanadium, and copper as well as cyanide and fluoride. Urged by stringent requirements, the steel industry has made substantial investments in environmental technology in recent years [247]. In the Federal Republic of Ger-

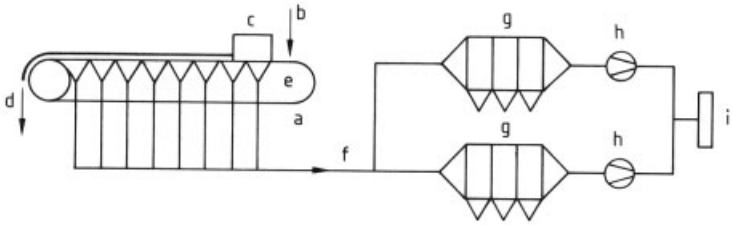

Figure 147. Sinter plant waste-gas dedusting
a) Sintering machine; b) Feed point; c) Igniter; d) Sinter; e) Wind boxes; f) Manifold; g) Electrostatic precipitator; h) Fan; i) Stack

many, for instance, from 1975 to 1984 ca. € 1.2×10^9 (10 % of the steel industry's resources for new investments) were spent on environmental protection. The lowering of dust emissions from 10 kg per tonne of raw steel in 1960 to the present value of 1.4 kg per tonne of crude steel serves as an example for the success of new techniques in controlling air pollution [248].

The endeavors towards pollution control are not limited to the *air pollution control,* but also include *water pollution control, reduction of noise,* and *waste management.* The division of capital expenditure between the individual areas in the Federal Republic of Germany (1975 – 1984) is as follows:

Air pollution control	67.7 %
Water pollution control	21.4 %
Noise reduction	9.5 %
Waste management	1.4 %
Total	€ 1204×10^6

It can be seen that the investment in air-pollution control represents two thirds of the total capital expenditure.

Apart from the *direct measures* taken to control pollution, the *saving* or *recovery* of *energy* also has a favorable effect on the environment because the corresponding amounts of energy can only be produced at the expense of the environment. The most important examples of these measures are (1) the reduction of the specific fuel consumption during sintering by half since the 1950s, (2) the expansion of blast-furnace top gas in turbines and its utilization for heating, (3) dry coke quenching, and (4) CO gas recovery (→ Steel).

Pollution-control techniques being currently employed in pig-iron production and in related auxiliary plants will be described with a series of examples. However, in view of the diversity of processes available, a complete survey cannot be presented here.

4.1. Air Pollution Control

Virtually all areas of ironworks make use of gas dedusting techniques. Apart from mechanical and filtering dust separators and scrubbers, electrostatic precipitators are frequently used today [249 – 251].

Sintering Plants. In the past, dust separation from off-gases in iron-ore sintering plants was carried out with cyclones. Today's demands are essentially fulfilled by three-field horizontal dry electrostatic precipitators, which are operated on the suction side of the sinter waste-gas fan. The waste-gas flow rate of a sinter belt with a reaction area of e.g., 400 m^2 is in the range of 1.8×10^6 m^3/h. As shown in Figure 147, the waste-gas dedusting in that case is achieved in two parallel electrostatic precipitators.

The electrical resistance of sinter dust, which in most cases is very high, frequently impedes electrostatic dust precipitation [252]. Increasing the humidity of the gas, by adding water either to the off-gas or to the sinter belt, can improve dust separation by *reducing the electrical resistance* of the dust. Dust emission can be reduced by 20 – 50 % of the starting values by conditioning the waste gas with sulfur trioxide [251]. In this process sulfur dioxide, obtained, e.g., by combustion of sulfur, is oxidized to sulfur trioxide and added to the waste gas in front of the electrostatic filter. A level of 10 – 30 ppm of sulfur trioxide in the waste gas is sufficient to reduce the electrical resistance of dust and thus cause the improvement in dedusting efficiency mentioned above. The required amount of sulfur trioxide is so low that only ca. 33 kg/h of sulfur

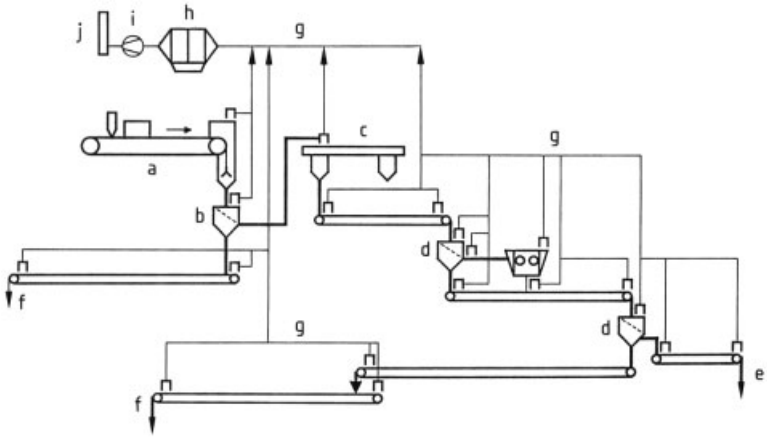

Figure 148. Sinter plant room dedusting
a) Sintering machine; b) Hot screening; c) Cooler; d) Cold screening and crushing; e) Sinter product; f) Return fines; g) Suction system; h) Electrostatic precipitator; i) Fan; j) Stack

are needed to produce e.g., 20 ppm of sulfur trioxide for the conditioning of the waste gas of a 400 m^2 sinter belt. Extra sulfur trioxide emission does not occur because it is bound to the dust by chemisorption (sulfate formation). The effect of this process depends on the basicity of the sinter feed and on the composition of the raw material.

Simple housings at feeding points, at the sinter discharge end, at screening machines and at numerous transport devices cannot effectively prevent the escape of dust into the surroundings. For this reason, *suction hoods* are installed that are connected by means of branched ducts to a central dedusting system, which is often an electrostatic precipitator. The schematic of a so-called room dedusting system of this type is presented in Figure 148. In many cases, the dust-laden off-gas from the sinter cooler is passed through the room dedusting plant, however, it may sometimes be led to a separate electrostatic precipitator.

Blast Furnace. A blast furnace with an open top is not used any longer. Dedusting systems have allowed the cleaning of top-gas to dust loading values of < 10 mg/m^3 (STP), and thus have permitted the utilization of the energy contained in this gas in turbines and for the heating of hot blast stoves. Apart from the direct protection of the environment brought about by top-gas purification compared to an open top, an indirect effect is achieved by saving energy which would have to be produced at the expense of the environment.

The top-gas leaving the blast furnace (a) is loaded with 10 – 40 g/m^3 (STP) of dust. The volumetric flow rate of a blast furnace that has, e.g., an output of 5000 t of hot metal per day is ca. 420 000 m^3/h (STP). Figure 149 shows the components of the top-gas purification plant of a blast furnace that is operated at a *slight overpressure.* The larger particles of dust are first separated in a gravity dust catcher (b), i.e., a settling chamber with flow deflection, and subsequently in a cyclone (c). Further cleaning of the blast-furnace gas occurs in a single- or multiple-stage scrubber (d) and in a wet electrostatic precipitator (e), which carries out both fine purification as well as removal of droplets. If the top-gas volumetric flow rates are very high, several purification plants are installed in parallel.

The top-gases from blast furnaces that are operated at *higher pressures* can be freed from dust to the desired degree of purity by using a pressure difference of ca. 20 kPa in high performance scrubbers without an electrostatic precipitator. Another effective method of top-gas dust separation is dry dust removal in electrostatic precipitators. It has the advantage of involving low pressure losses and consequently has a lower energy requirement for gas transport and, in addition, water and sludge treatments are no longer necessary. Dry top-gas purification represents the latest technological development.

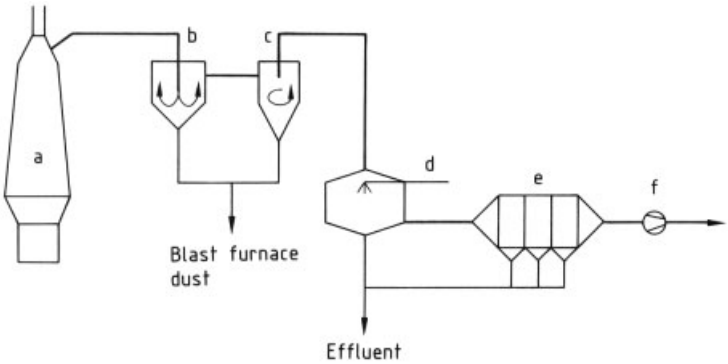

Figure 149. Top-gas cleaning
a) Blast furnace; b) Gravity dust catcher; c) Cyclone; d) Scrubber; e) Wet electrostatic precipitator; f) Fan

In the area of raw material handling a room dedusting plant prevents dust emissions. The burden bunkers and conveyors are connected by a suction system to a bag filter or an electrostatic precipitator.

The *smoke* that arises during the tapping of the blast furnace represents a minor environmental problem as far as emission protection is concerned. However, it is a nuisance for the operating personnel and makes working conditions on the casting platforms very difficult. *Cast house dedusting* is being carried out for some years now to improve working conditions. In fact, even older blast furnaces are being equipped with dedusting systems [253, 254]. The enclosing of the skimmer, iron, and slag runners and their points of transfer with easily removable hoods having fireproof linings is characteristic for a casting house dedusting plant. The enclosed areas are connected to suction lines, which are preferably installed below the casting platform. The region of the tapping hole cannot be enclosed because it has to remain accessible to the drilling and plugging machines. In this case, smoke is usually caught with a suction hood placed near the tapping hole and then led to the precipitator. The flow diagram of the main, hot metal, and slag runners on a casting platform as well as the suction system is shown in Figure 150. The hatched parts of the diagram represent the enclosed areas. It can be seen that each suction point has a damper. When the plant is in operation, only the suction points at which smoke is to be trapped are opened. For instance, if no slag is discharged in the first phase of tapping, all the dampers on the slag runner remain closed.

In blast furnace tapping, the flow volume that has to be trapped, dedusted, and transported is 500 000 – 800 000 m^3/h. The power consumption of a casting house dedusting system can reach 0.8 MW, hence, modern plants are automatically controlled. Control dampers in the suction lines and variable-speed blowers draw off as little air as necessary. For example, a minimum amount of air is withdrawn during tapping pauses and maximum amounts at the start and termination of tapping. Bag filters or electrostatic precipitators are employed as dust collectors. The advantages of electrostatic precipitators are the substantially lower pressure loss and the fact that they can be operated during tapping pauses with minimal power input.

Dry Coke Quenching. As mentioned above, dry quenching of coke also has a positive effect on the environment, because of the fact that, in contrast to wet quenching, the waste heat of the coke is utilized and no cooling clouds are produced. The heat is removed from the coke with the help of recirculating gas and used for steam generation in a waste-heat boiler. In an ironworks in the Federal Republic of Germany, two 70 t/h dry coke quenching plants produce 36 t/h of steam each, which is then used as a fuel substitute in a neighboring power plant [255]. The fuel saved in this way contributes to environmental protection. An equivalent of > 20 MW of electrical power is produced.

Direct Reduction Plants. In direct reduction plants the waste gases leaving the furnace are freed from dust in an electrostatic precipitator, bag filter or scrubber, depending on the re-

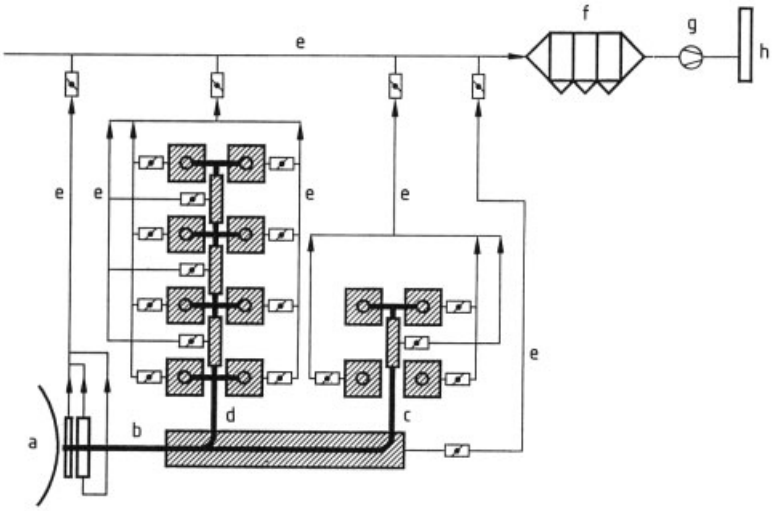

Figure 150. Cast house dedusting
a) Blast furnace, area of tapping hole; b) Main runner with skimmer; c) Hot metal runner with transfer points; d) Slag runner with transfer points; e) Suction lines; f) Electrostatic precipitator; g) Fan; h) Stack

quirements in the operational areas. Electrostatic precipitators or bag filters are used for room dedusting plants.

Regarding the emissions of the dust components mentioned above, cadmium, lead, and possibly manganese can be of relevance in the making of pig iron, however, the limiting values for these metals are generally adhered to.

4.2. Prevention of Water Pollution

In the course of pig iron production, modern ironworks cause practically neither water pollution, nor warming up of waters. Only small amounts of water are required in *sintering plants* for rolling the burden and, if necessary, for sinter cooling or gas conditioning. The water used is released into the atmosphere in the gaseous state through the waste gas stack. Generally, *direct reduction plants* do not produce wastewater.

Blast furnaces, on the other hand, have a considerable water requirement for cooling. In the past, up to 60 m^3 of water were consumed per tonne of pig iron [256]. Another water-consuming process pertaining to the blast furnace is wet top-gas purification. It is becoming increasingly rare today that the wastewater of this process is released into the atmosphere. In fact, the introduction of a closed water circulation with the appropriate water conditioning plant [257, 258] for top-gas purification and blast furnace cooling contributes to the marked decrease in the water requirement for steelmaking, which is today ca. 30 m^3 per tonne of raw steel [248].

Blast-furnace cooling with a closed circulation requires water-cooling plants. Usually the circulation water is evaporated in the blast furnace coolers and condensed and cooled in cooling towers.

The essential components of a scrubber/precipitator effluent for top-gas purification are shown in Figure 151. The clarification of the effluent can be carried out in a thickener and the dewatering of the sludge in a filter press. However, other dewatering machines, such as drum or rotary disk filters may be applied.

4.3. Noise Reduction

As in other industrial plants, numerous sources of noise exist in ironworks, which contribute to the total noise emission. Many of these noise sources are not specific to the production of pig iron. These include fans, compressors, pumps and ore-dressing plants. The noise abatement measures consist of, e.g., the use of low-vibration motors, the enclosing of machines or other apparatus, or even entire parts of the plant, such as the sinter waste-gas fans [259].

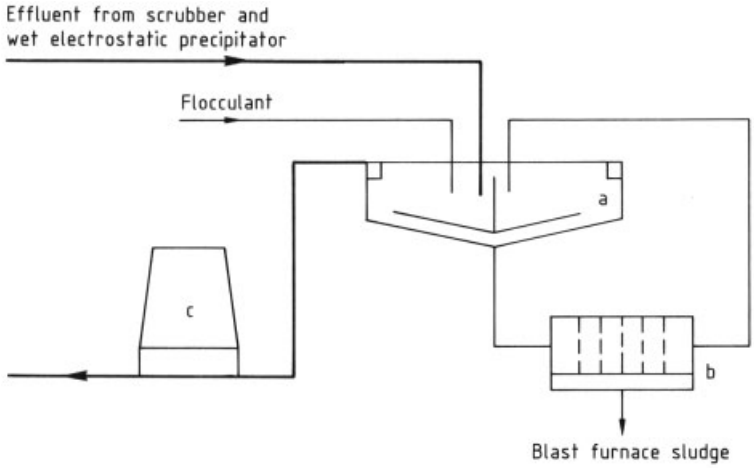

Figure 151. Effluent treatment for top-gas cleaning
a) Thickener; b) Filter press; c) Cooling tower

Special measures are required to dampen the feeding and expansion noises at hot blast stoves. Considerable noise reduction can be achieved here by covering the valves and the expansion line with a sound-absorbing material and by using sound absorbers or perforated disks in the blowoff line [260]. Emission values of ca. 40 dB(A) can be achieved at a distance of 1000 m from a blast furnace or sintering plant if noise damping measures are used consistently [261].

4.4. Waste Management

Blast furnace slag is quantitatively the most important waste material created in the production of pig iron. This material is completely utilized. About two thirds of the total amount is employed in road construction. The remainder is used by the cement industry for the production of blast furnace cement [262].

The recycling of the *dry dust* obtained in top-gas purification also causes no problems because it can be recycled to the sintering plant. However, *blast furnace sludge* is for the most part still being dumped, in amounts of 130 000 t/a in Germany [263]. The reason for this is the high content of lead, zinc and alkali found in the sludge. If the sludge were recirculated, these substances would accumulate in the blast furnace. Today, various processes are available which reduce the environmental impact caused by these dumps by allowing the recycling of blast furnace sludge and other waste containing lead and zinc.

One way of expelling the unwanted accompanying substances is the reprocessing of dusts and sludge in rotary kilns [264 – 267]. The product thus obtained consists essentially of metallic iron, which can, e.g., be briquetted and returned to the pig iron production [268].

Most of the *dust* separated in the waste-gas precipitators of sintering plants can usually be led back into the process. Only the dust arising in the last field of the electrostatic precipitator must be removed from the circulation because of alkali enrichment.

The *prime goal* is always the avoidance or minimizing of residual substances. For example, the specific amount of slag has decreased from 800 kg/t at the beginning of the 1950s to 300 – 350 kg/t of pig iron today. In a modern steel plant, ca. 90 % of the total amount of waste products is recycled.

5. Economic Aspects

Hot Metal. The conventional and by far the most common approach to iron production involves reducing and smelting the iron ore in a blast furnace together with coke, the immediate product being hot metal. Hot metal, in turn, is the principal pre-product for the production of commercial crude steel. Production of a tonne of

Table 38. Hot metal production and fraction of world output, by country

Country	Production in 1000 t			% of world output		
	1998	1999	2000[a]	1998	1999	2000
Germany	30 162	27 934	30 845	5.8	5.2	5.4
Belgium/Luxembourg	8618	8430	8472	1.7	1.6	1.5
France	13 458	13 852	13 916	2.6	2.6	2.4
United Kingdom	12 746	12 139	10 859	2.5	2.2	1.9
Italy	10 771	10 615	11 161	2.1	2.0	1.9
Netherlands	5562	5307	4970	1.1	1.0	0.9
Portugal	387	389	380	0.1	0.1	0.1
Spain	4236	4146	4135	0.8	0.8	0.7
Finland	2920	2954	2983	0.6	0.5	0.5
Austria	4021	3913	4318	0.8	0.7	0.8
Sweden	3149	3212	3146	0.6	0.6	0.5
European Union	96 030	92 891	95 185	17.8	17.2	16.6
Norway[b]	70	70	70	0.0	0.0	0.0
Switzerland[b]	80	80	80	0.0	0.0	0.0
Turkey	5286	5181	5333	1.0	1.0	0.9
F.R. Yugoslavia	850	139	598	0.2	0.0	0.1
Eastern Europe	21 091	17 711	19 823	3.9	3.3	3.4
Bulgaria	1390	1152	1220	0.3	0.2	0.2
Poland	6179	5233	6491	1.1	1.0	1.1
Romania	4525	3006	2985	0.8	0.6	0.5
Czech Republik	4982	4023	4621	0.9	0.7	0.8
Slovakia	2756	2987	3166	0.5	0.6	0.6
Hungary	1259	1310	1340	0.2	0.2	0.2
Former U.S.S.R.	58 107	65 438	73 540	10.8	12.1	12.8
Russia	34 736	40 033	44 536	6.4	7.4	7.7
Ukraine	20 777	21 937	24 993	3.9	4.1	4.3
Kazakhstan	2594	3468	4011	0.5	0.6	0.7
North America	61 699	59 947	61 496	11.4	11.1	10.7
United States	48 230	46 268	47 709	8.9	8.6	8.3
Canada	8937	8857	8904	1.7	1.6	1.5
Mexico	4532	4822	4883	0.8	0.9	0.8
South America	28 852	28 074	31 625	5.4	5.2	5.5
Argentina	2122	1985	2182	0.4	0.4	0.4
Brazil	25 132	24 549	27 723	4.7	4.5	4.8
Chile	993	1030	1026	0.2	0.2	0.2
Columbia	256	264	285	0.0	0.0	0.0
Peru	283	191	327	0.1	0.0	0.1
Venezuela				0.0	0.0	0.0
Africa	8805	8065	8400	1.6	1.5	1.5
South Africa	6893	6100	4580	1.3	1.1	0.8
Egypt	1357	1300	1400	0.3	0.2	0.2
Middle East	2117	2112	2202	0.4	0.4	0.4
Iran	2117	2112	2202	0.4	0.4	0.4
Asia	247 667	253 568	269 032	46.0	46.9	46.8
Japan	74 981	74 520	81 071	13.9	13.8	14.1
China (Peoples Rep.)	118 521	125 392	130 535	22.0	23.2	22.7
India[c]	20 194	20 139	21 321	3.7	3.7	3.7
North Korea	250	250	250	0.0	0.0	0.0
South Korea	23 299	23 329	24 937	4.3	4.3	4.3
Taiwan	9374	8890	9618	1.7	1.6	1.7
Australia[c]	7723	7047	7049	1.4	1.3	1.2
New Zealand	609	620	603	0.1	0.1	0.1
World	538 986	540 943	575 036	100.0	100.0	100.0

[a] Preliminary, in some cases estrimates. Source IISI.
[b] estimation
[c] incl. ferralloys.

Table 39. Iron ore production and fraction of world output, reported by country

Country	Production, 10^6 t						% of World Output					
	1995	1996	1997	1998	1999	2000	1995	1996	1997	1998	1999	2000
France	1.5	1.3	0.7	0.0	0.0		0.1	0.1	0.1	0.0	0.0	0.0
United Kingdom	0.0	0.0	0.0	–	–		0.0	0.0	0.0			
F.R. Germany	0.0	0.0	0.0	–	–		0.0	0.0	0.0			
Spain	2.0	1.2	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0
Sweden	21.7	21.3	21.9	20.9	18.9	20.6	2.0	2.1	2.1	2.1	1.9	2.0
Austria	2.1	1.9	1.3	1.3	1.3	1.3	0.2	0.2	0.1	0.1	0.1	0.1
European Union (15)	27.1	25.7	23.9	22.2	20.2	21.9	2.6	2.5	2.2	2.2	2.0	2.1
Turkey	5.5	6.4	6.2	5.5	4.5	4.4	0.5	0.6	0.6	0.6	0.4	0.4
Norway	1.7	1.3	0.6	0.5	0.5	0.6	0.2	0.1	0.1	0.1	0.0	0.1
											0.0	0.0
United States	63.2	61.1	62.7	62.6	57.7	63.0	6.0	5.9	5.9	6.3	5.7	6.0
Canada	38.2	36.0	38.5	38.6	32.4	37.8	3.6	3.5	3.6	3.9	3.2	3.6
Brazil	175.0	170.0	184.0	192.7	186.6	197.7	16.5	16.4	17.3	19.3	18.5	18.8
Venezuela	19.4	18.7	18.7	17.5	13.2	17.6	1.8	1.8	1.8	1.8	1.3	1.7
Mexico	10.0	10.2	10.5	10.6	11.0	12.7	0.9	1.0	1.0	1.1	1.1	1.2
Peru	6.1	4.8	4.6	5.0	4.2	4.2	0.6	0.5	0.4	0.5	0.4	0.4
Chile	7.9	8.5	8.1	8.3	7.7	7.9	0.7	0.8	0.8	0.8	0.8	0.8
Columbia	0.6	0.6	0.4	0.4	0.4	0.6	0.4	0.1	0.0	0.0	0.0	0.1
South Africa	31.2	29.7	32.0	32.4	29.0	32.0	2.9	2.9	3.0	3.2	2.9	3.0
Zimbabwe	0.3	0.4	0.5	0.5	0.5	0.5	0.5	0.0	0.0	0.1	0.0	0.0
Mauretania	11.6	11.4	11.6	11.5	10.5	11.3	1.1	1.1	1.1	1.2	1.0	1.1
Algeria	2.1	2.3	2.3	1.8	1.8	1.8	0.2	0.2	0.2	0.2	0.2	0.2
Tunisia	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Morocco	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Nigeria	0.5	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Egypt	2.3	2.3	2.3	2.7	3.0	2.5	2.2	0.2	0.2	0.3	0.3	0.2
India	66.8	66.7	69.4	71.7	70.2	76.3	6.3	6.5	6.5	7.2	6.9	7.3
Iran	3.0	4.3	4.6	5.6	6.4	6.4	0.3	0.4	0.4	0.6	0.6	0.6
South Korea	0.6	0.5	0.5	0.5	0.4	0.4	0.1	0.0	0.0	0.1	0.0	0.0
Malaysia	0.3	0.3	0.3	0.4	0.4	0.3	0.0	0.0	0.0	0.0	0.0	0.0
Australia	142.8	142.6	160.0	181.2	151.9	172.9	13.4	13.8	15.1	18.1	15.0	16.5
New Zealand	2.6	2.3	2.5	2.3	2.6	3.0	0.2	0.2	0.2	0.2	0.3	0.3
Russia	78.2	72.1	70.8	72.3	81.9	87.0	7.4	7.0	6.7	7.2	8.1	8.3
Kazakhstan	14.9	12.9	12.6	8.7	9.1	10.0	1.4	1.2	1.2	0.9	0.9	1.0
Ukraine	50.4	47.6	53.0	50.8	47.1	55.0	4.7	4.6	5.0	5.1	4.7	5.2
China (Peoples Rep.)	261.9	252.3	268.6	222.2	237.2	224.0	24.7	24.4	25.3	22.2	23.5	21.3
North Korea	1.0	1.0	1.0	1.0	1.0	1.0	0.1	0.1	0.1	0.1	0.1	0.1
Romania	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0		0.0	0.0
Bulgaria	0.5	0.9	0.9	0.3	0.2	0.2	0.0	0.1	0.1	0.0	0.0	0.0
Former CSFR	1.0	1.0	0.4	0.5	0.5	0.5	0.1	0.1	0.0	0.1	0.0	0.0
World total	1062.1	1033.5	1063.0	999.6	1011.1	1050.6	100	100	100	100	100	100

in some cases estimates
IISI, Statistical Yearbook 2001, p. 96 – 97; Statistisches Jahrbuch Stahl, WV 2001, p. 378.

steel requires ca. 700 kg of hot metal in addition to scrap and flux.

World production of hot metal in 2000 reached a level of 576 $\times$ 10^6 t, iron ore production a level of ca. 1051 $\times$ 10^6 t (see Tables 38 and 39) [269]. The European Community alone produced 95.2 $\times$ 10^6 t of hot metal in 2000, using 134.5 $\times$ 10^6 t of iron ore together with 40 $\times$ 10^6 t of coke (blast furnace coke plus sinter coke breeze). The major cost factors in hot metal production are related to raw materials (49 %), reductants (30 %), and processing (21 %) [270].

In view of the massive requirements for raw material and energy, and investment costs, the potential for savings through optimization of scale and exploitation of more economical

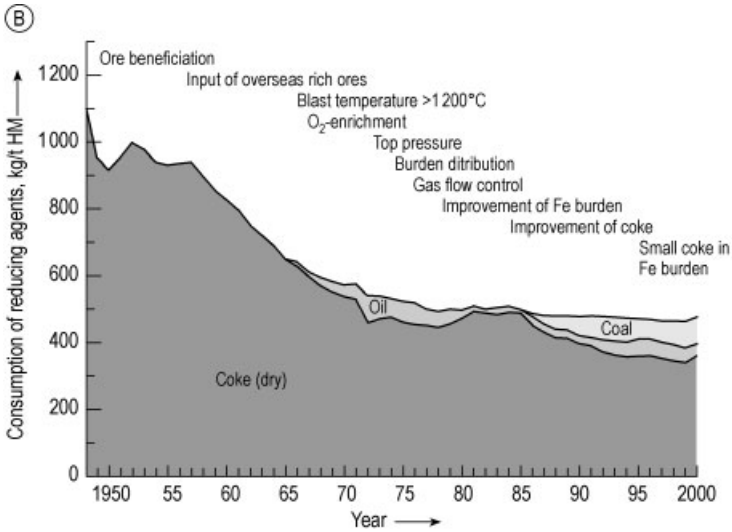

Figure 152. Evolution of blast furnace burden composition (A) and reductant rate (B) in the Federal Republic of Germany since 1950 for basic hot metal

sources of raw material is extremely attractive. Environmental protection measures have also come to play an increasingly important economic role in hot metal production in recent years.

Careful organization and quality control on the input side of hot metal production is also important with respect to the properties of the principal end product – steel – and consequently on the market picture as it relates to steel vs. alternative materials. Economic factors and quality are therefore closely linked.

Several recent developments deserve special mention as determinants of the competitiveness of steel, especially from the standpoint of cost reduction:

1) Increased emphasis on major sources of ore displaying a high iron content, a consideration which helps to minimize fluctuations in ore quality.
2) Significant improvement in the facilities for *handling* and *transporting* ore in increasingly large unit amounts. Intercontinental marine ore transport today involves units as large as 350 000 dwt (i.e., dead weight), representing a three-fold increase in maximum load in the course of 30 years.
3) Steady improvement in the *preparation of the furnace charge* in terms of ore classification, sintering, and pelletizing. For example, the use of lump ore has decreased in the Federal Republic of Germany from 35 % in 1970 to 10 % ca. 2000. Sinter input, on the other hand, has increased to 60 %, and pelletized ore – virtually unknown before 1965 – now represents 30 % of the charge (see Fig. 152).
4) Significant decrease in *reductant consumption* and more variety in the nature of the injected reductants employed for reduction.

Reducing agent consumption has dropped by a total of ca. 20 % over the last 30 years, now amounting to only 475 kg per tonne of hot metal. Coke continues to constitute the major source of reductant (360 kg/t), although since about 1986 inexpensive coal has increasingly been used as a coke substitute (81 kg/t HM as average in Germany). Oil injection currently accounts for ca. 33 kg/t HM.

By far the largest part of the hot metal output goes into *steel production*.

In Europe essentially *no open market* exists for basic hot metal, because the steel industry is effectively self-supporting. Iron of this type is purchased only rarely and in times of shortage, as when demand is particularly strong or some industrial mishap causes supply problems. By contrast, there is a lively trade in such products as *foundry iron, hematite iron,* and *spheroidal pig iron*.

Table 40. Development of import and export of pig iron (worldwide) 1000 t

Country	1994	1995	1996	1997	1998	1999	2000
Import:							
World	10 685	12 428	10 402	12 838	13 352	12 636	12 946
FR Germany	387	415	248	300	358	272	257
Belgium/Luxembourg	246	103	70	91	194	204	336
Denmark	90	85	102	100	134	121	148
France	183	202	181	171	208	222	288
United Kingdom	165	178	215	201	225	156	134
Ireland	1	10		3	4	3	5
Italy	1203	1637	1352	1310	1712	1254	1104
Netherlands	256	199	311	283	408	203	255
Greece	35	17	14	13	11	8	5
Portugal	14	24	21	23	27	19	24
Spain	232	316	373	255	350	325	388
European Union (12)	2813	3185	2887	2750	3629	2787	2942
Finland	29	74	67	34	30	22	33
Austria	24	25	19	17	25	23	29
Sweden	46	40	39	36	89	48	78
European Union (15)	2912	3323	3011	2837	3772	2879	3082
Switzerland	23	45	31	22	36	28	31
Czech Rep.	62	70	31	31	42	62	66
Poland			40	31	26	22	12
Hungary	21	24	32	32	46	26	25
United States	2442	2364	2658	3150	5145	4992	4970
Canada	64	19	33	130	124	168	68
Japan	1601	2613	792	1051	184	143	851
Taiwan	1198	1090	850	1112	955	952	847
South Korea	1764	2142	2427	3692	2481	2752	2272
others	598	738	498	749	539	611	723
Export:							
World	6579	9430	8493	13 000	13 390	12 505	15 122
FR Germany	162	118	92	127	65	84	117
Belgium/Luxembourg	106	90	85	89	139	137	187
France	2	16		48	8	18	0
United Kingdom	2	5	5	4	2		1
Italy	3	5	8	35	32	21	27
Netherlands	59	32	8	81	67	107	99
Portugal		2	2	2	3	3	2
Spain	1		1	1	1	1	2
European Union (12)	334	267	202	389	317	371	435
Finland		5	6	6	3	4	8
Austria			2	1		1	1
Sweden	30	13	64	42	16	2	28
European Union (15)	365	286	274	437	337	378	471
Czech Rep.	198	193	168	228	161	42	33
Russia	2604	2206	2078	2305	2469	2796	3563
Ukraine	1657	712	399	536	1267	1865	1555
Australia			22			3	3
Japan	1201	500	954	250	2311	1734	224
Canada	79	129	110	117	153	117	109
United States	56	54	60	86	87	83	72
Brazil				2563	3212	3030	3809
China (People. Rep.)		4892	3711	5600	2434	1652	3333
others	419	459	719	878	960	806	1951

Amtliche Außenhandelsstatistik

The largest exporter of pig iron, and by a wide margin, are *Russia* and *Brazil* at 3.6 to 3.8 $\times$ 10^6 t/a (see Table 40) [271].

DRI (Direct Reduced Iron). The direct reduction process (see Section 2.6) provides the market with another form of reduced iron: DRI, available in significant quantity since the 1970s

Table 41. Direct reduced iron production, 1995 – 2000 (million metric tons)

Country	1995	1996	1997	1998	1999	2000
FR Germany	0.41	0.37	0.47	0.45	0.40	0.46
Sweden	0.12	0.13	0.13	0.12	0.12	0.13
European Union (15)	**0.53**	**0.50**	**0.60**	**0.57**	**0.52**	**0.59**
Russia	1.68	1.51	1.73	1.73	1.88	1.92
Canada	1.00	1.42	1.39	1.24	0.92	1.13
Mexico	3.70	3.79	4.44	5.58	6.07	5.95
United States	0.42	0.45	0.51	1.60	1.67	1.56
NAFTA	**5.12**	**5.66**	**6.34**	**8.43**	**8.66**	**8.63**
Argentina	1.33	1.42	1.50	1.54	0.99	1.42
Brazil	0.29	0.34	0.32	0.34	0.40	0.41
Peru	0.00	0.02	0.07	0.06	0.05	0.08
Trinidad and Tobago	1.08	1.01	1.14	1.02	1.38	1.51
Venezuela	5.10	5.73	5.64	5.45	4.78	6.41
Central and South America	**7.79**	**8.52**	**8.67**	**8.40**	**7.60**	**9.83**
Egypt	0.85	0.83	1.19	1.61	1.67	1.53
Libya	0.96	0.86	1.01	1.00	1.30	1.49
Nigeria	0.04	0.01	0.00	0.00	0.00	0.00
South Africa	0.95	0.90	1.12	1.07	1.26	1.53
Africa	**2.80**	**2.60**	**3.31**	**3.68**	**4.23**	**4.54**
Iran	3.30	3.78	4.33	3.51	4.03	4.54
Qatar	0.62	0.63	0.56	0.71	0.67	0.62
Saudi Arabia	2.13	2.30	2.11	2.27	2.34	3.06
Middle East	**6.05**	**6.71**	**7.00**	**6.48**	**7.04**	**8.22**
PR China	0.00	0.00	0.07	0.08	0.11	0.05
India	4.27	4.79	5.12	5.12	5.29	5.50
Indonesia	1.71	1.72	1.53	1.57	1.74	1.82
Malaysia	1.18	1.17	1.49	0.88	0.95	1.26
Myanmar	0.02	0.04	0.04	0.04	0.03	0.04
Asia	**7.17**	**7.72**	**8.25**	**7.69**	**8.12**	**8.67**
Australia					0.28	0.56
World	**31.15**	**33.21**	**35.90**	**36.98**	**38.33**	**42.96**

and produced from lump ore or pellets, mainly with gas as reductant. The current world output of DRI is 43×10^6 t/a (see Table 41, [272]).

This production method represents a significantly different route to hot metal production when natural gas is available in abundance and for a low price as in the case of Indonesia, Malaysia, the Middle East, Mexico, and Venezuela. Indeed, DRI can often be produced for \$ 120/t under appropriate conditions, and most of the direct reduction facilities currently in operation are located in the regions cited [273].

DRI is devoid of elemental silicon and phosphorus, and it contains only little carbon – i.e., it lacks any source of energy for the subsequent conversion to steel. In this sense it is comparable to steel scrap, and it is highly priced as a charge for electric furnaces where these can be operated economically.

Alternative Production Methods. So far, no other alternative routes to produce liquid hot metal than the Corex process have achieved any significance. The Corex method utilizes a direct reduction shaft and a melter – gasifier (see Section 2.7) which was developed in Germany and first introduced in South Africa in 1988. In 2001 in total four units operated in India, South Africa and South Korea. One important advantage of the process is that the expensive construction of coking and sintering facilities can be avoided. The Corex process is dependent on lumpy ferrous burden materials (lump ores and/or pellets).

Size of the Production Facility. One way of classifying the various manufacturing methods is in terms of dimension (scale, capacity). For example, *blast furnaces* today may be as large as 14.9 m in diameter, with a working volume of 4789 m^3 (Schwelgen no. 2, Thyssen Krupp Stahl). Yearly the hot metal production reaches over 4×10^6 t. Local and company factors may well dictate smaller units.

For technical reasons, *direct reduction* facilities are much smaller, with a typical maximum annual output of 1.3×10^6 t DRI/modul. The *smelting reduction* Corex units are currently producing 0.6 to 0.8×10^6 t hot metal per annum. The economy of a Corex unit is linked with gas credits for the export gas produced at higher quantities.

Lower investment costs are responsible for the fact that smaller steel mills in particular (i.e., those with an annual output $< 1 \times 10^6$ t) tend to utilize a direct-reduction or smelting-reduction iron process in contrast to the blast furnaces at larger integrated steel mills, where the *lower reductant consumption* associated with a conventional blast furnace is the decisive factor [274].

The principal economic factors addressed here with respect to iron intended as a precursor to steel – performance of the production unit, investment costs, energy consumption, advantages of scale – all suggest that the blast furnace will continue to retain its primacy over the long term. Only where unusual local conditions favor other sources (through cheap sources of energy or limited demand) is the otherwise efficient and economical blast furnace likely to be displaced.

6. Acknowledgement

Chapters 1, 3 and Sections 2.1.2, 2.2 – 2.5 are based on the corresponding sections in Ullmann's, 4th ed. written by Manfred Ottow

7. References

1. Gmelin-Durrer: *Metallurgie des Eisens*, 4th ed., Springer Verlag, Berlin – Heidelberg 1964.
2. *Hütte, Taschenbuch für Eisenhüttenleute*. Akademischer Verein Hütte (ed.) 5th ed., Verlag Wilhelm Ernst u. Sohn, Berlin 1961.
3. R. C. Weast: CRC *Handbook of Chemistry and Physics,* 62nd ed., CRC Press, Inc., Boca Prota 1981.
4. K. Daeves: *Werkstoffhandbuch Stahl und Eisen*, 4th ed., Verlag Stahleisen mbH, Düsseldorf 1965.
5. O. Kubaschewski, E. L. Evans: *Metallurgische Thermochemie*, VEB Verlag Technik, Berlin 1959.
6. E. Plöckinger, H. Straube: *Die Edelstahlerzeugung*, 2nd ed., Springer Verlag, Wien 1965.
7. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry* 4th ed., John Wiley and Sons, New York 1980.
8. A. F. Holleman, E. Wiberg: *Lehrbuch der anorganischen Chemie*, 91rst – 100th ed., Walter de Gruyter, Berlin – New York 1985.
9. W. Stannek: *Beitrag zur Thermodynamik und Fehlstellenstruktur des Wüstits,* Dissertation Technische Universität Berlin 1973.
10. United Nations, Survey of World Iron Ore Resources, New York 1970.
11. *Skillings Mining Review* **75** (1986) no. 39, 4.
12. *Iron Ore 1986*, American Iron Ore Association, Cleveland, Ohio 1987.
13. United Nations Conference on Trade and Development: *Review of Iron Ore Statistics, GF.94 – 53 266,* 1994.
14. *Skillings Mining Review* **77** (1988) no. 17, 4 – 5, 17.
15. K. J. Pieper, *Handel und Verbrauchsentwicklung von Eisenerzpellets in der westlichen Welt*, lecture at VDEh Düsseldorf, June 10 1987.
16. R. C. Anderson et al., *Skillings Mining Review* **76** (1987) no. 33, 6 – 9.
17. Exploration und Bergbau GmbH, personal communication 1988.
18. H. J. Werz et al.: "Environmental Protection in Iron Ore Sintering by Waste Gas Recirculation," *Stahl Eisen* **115** (1995) 120 – 126.
19. L. v. Bogdandy, H. J. Engell: *Die Reduktion der Eisenerze*, Springer Verlag, Berlin/Verlag Stahleisen mbH, Düsseldorf 1967.
20. Energie- und Betriebswirtschaftsstelle des Vereins Deutscher Eisenhüttenleute (eds.): *Anhaltszahlen für die Wärmewirtschaft in Eisenhüttenwerken*, 6th ed., Verlag Stahleisen mbH, Düsseldorf 1968.
21. H. Hille, W. Lanzer, H. B. Lüngen, R. Sieger: "Technical and economical examination of the development of blast furnace capacities," *MPT International* 1997, no. 4, 66 – 80.
22. G. Winzer, H. B. Lüngen: "Aspects on coke quality – a brief survey," *Proc. Eur. Coke Committee, Linz, Austria*, May 30, 1995.
23. Evaluation by VDEh Blast Furnace Committee 1995.
24. Evolution de la composition granulometrique du coke entre la cokerie et le creuset haut fourneau; ECSC Research Project 7210 – AA/211; 7210 – AA/315, 1995.

25. H. B. Lüngen, A. Poos: "Injection of coal into the blast furnace – ECSC synthesis report," *Cokemaking International* **8** (1996) 14 – 31.
26. H. Bertling, W. Lanzer, H. B. Lüngen: "Coke quality and behavior of coke in the blast furnace," *Cah. Inf. Tech. Rev. Metall.* **96** (1993) no. 3, 295 – 307.
27. Fédération Européenne de Produits Refractaires *PRE Refractory Materials;* Recommendations 1985, PRE Administrative Secretariat, Rue de Colonies 18 – 24, Bte 17, B-1000 Brussels.
28. M. Koltermann: "Feuerfeste Baustoffe 1980 bis 1990 – Rückblick und Prognosen," *Stahl Eisen* **111** (1991) no. 8, 41 – 48.
29. *Feuerfestbau-Stoffe, Konstruktion, Ausführung*, Vulkan-Verlag, Essen 1987. *Refractory Engineering – Materials, Design, Construction*, 2nd edition, Vulkan-Verlag, Essen 2005.
30. *Stahleisen-Werkstoffblätter* 912 bis 917, Taschenbuch der Stahleisen Werkstoffblätter, 10. Auflage, Stahlinstitut VDEh, Düsseldorf 2001.
31. S. Kataoka, *"Refractories for Steelmaking in Japan,"* vol. 1, UNITECR Proceedings, Kyoto1995, pp. 1 – 27.
32. H. Nishio: "Steelmaking Refractory Trends in Japan", Proceedings UNITECR 2003 Congress, Osaka, October 2003, pp. 1–4.
33. A. Buhr: "Refractories for Steel Secondary Metallurgy", *CN-Refractories*, Vol. **6** (1999), No. 3, pp. 19–30.
34. "Statistics," *Taikabutsu Overseas* **24** (2004) no. 3, 154.
35. Institut für Gesteinshüttenkunde der RWTH Aachen, *Feuerfeste Werkstoffe für die Herstellung und den Transport von Roheisen*, 31. Internationales Feuerfestkolloquium, Aachen 10. – 11. Oktober 1988.
36. H. Nakamura: "Carbon Containing Bricks for Blast Furnaces," *Taikabutsu Overseas* **15** (1995) 13 – 18.
37. W. Kowalski, H.B. Lüngen, K.P. Stricker: "Hochofenhaltbarkeit: Stand, Entwicklung und Maßnahmen zur Verlängerung von Ofenreisen", *Stahl und Eisen* **119** (1999), Nr. 4, pp. 119–128.
38. P. Eisen: "Zustellung und Inbetriebnahme des Hochofens B der Hüttenwerke Krupp Mannesmann", *Stahl und Eisen* **121** (2001), Nr. 9, pp. 51–57.
39. M. Peters, P. Schmöle, H.P. Rüther, C. Leuermann: "Blast furnace relining strategies for campaign life of more than 20 years", *Stahl und Eisen* **123** (2003), Nr. 1, pp. 21–26.
40. J. Janz, D. Ringel, D. Lücke, R. Vollmer, H. Hille, G. Rassel: "Einsatz von Cu-Staves im Hochofengestell auf die Feuerfestkonstruktion", *Stahl und Eisen* **122** (2002), Nr. 6, pp. 31–37.
41. M. Koltermann, "Torpedo ladle Refractories in West-Germany," *Taikabutsu Overseas* **5** (1985) no. 2, 35 – 40.
42. H. Nishio, A. Matsuo: "Al_2O_3 – SiC – C Bricks for Torpedo Car," *Taikabutsu Overseas* **15** (1995) 39 – 46.
43. Verein Deutscher Eisenhüttenleute (ed.), Grundlagen des Hochofenverfahrens, Verlag Stahleisen mbH, Düsseldorf 1973.
44. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances*, Springer Verlag, Berlin – Heidelberg – New York/Verlag Stahleisen mbH, Düsseldorf 1973.
45. S. Eketorp in *Proceedings of the Meeting* "Alternative Routes to Steel," London 1971, p. 14, The Iron and Steel Institute, London 1971.
46. Gmelin-Durrer: *Metallurgie des Eisens*, 4th ed. vol**s. 4 a, 4 b**, Springer Verlag, Berlin – Heidelberg 1972.
47. W. Oelsen, H. G. Schubert, *Arch. Eisenhüttenw.* **40** (1969) 1 – 4.
48. Gmelin-Durrer: *Metallurgie des Eisens*, 4th ed. vol**s. 2 a, 2 b**, Springer Verlag, Berlin – Heidelberg 1968.
49. D. Neuschütz, N. Towhidi, *Arch. Eisenhüttenwes.* **41** (1970) 421 – 426.
50. R. B. Bird, W. E. Stuart, E. N. Lightfoot: *Transport Phenomena*, John Wiley & Sons, New York 1960.
51. W. M. McKewan, *Trans. Metall. Soc. AIME* **212** (1958) 791.
52. L. v. Bogdandy, H.-G. Riecke, *Arch. Eisenhüttenw.* **29** (1958) 603 – 609.
53. L. v. Bogdandy, H. P. Schulz, B. Würzner, I. N. Stranski, *Arch. Eisenhüttenw.* **34** (1963) 401 – 409.
54. R. Jeschar, *Arch. Eisenhüttenw.* **38** (1967) 663 – 667.
55. D. Papanastassiou, G. Bitsianes, *Metall. Trans.* **4** (1973) 477 – 486.
56. A. Paul, B. Deo, N. Sathyamurthy, "Kinetic Model for Reduction of Iron Oxide in Molten Slags by Iron-Carbon Melt," *Steel. res.* **65** (1994) 414 – 420.
57. M. Sedlmeier: *Reaktionsabläufe bei der Reduktion von Eisenoxid in Schlacken durch kohlenstoffhaltige Eisentropfen*, Diploma Thesis, Technische Universität Berlin, Berlin 1996.

58. M. H. J. Baird, J. F. Davidson, "Gas Absorption by Large Rising Bubbles," *Chem. Engg. Sci.* **17** (1962) 87 – 93.
59. R. F. Johnston, R. A. Stark, J. Taylor, "Diffusion in Liquid Slags," *Ironmaking Steelmaking* **4** (1974) 220 – 226.
60. F. Oeters, "Zur Kinetik des Frischens unter besonderer Berücksichtigung des Sauerstoffaufblasverfahrens," *Arch. Eisenhüttenwes.* **37** (1966) 209 – 219.
61. Gmelin, Durrer (eds.): *Metallurgie des Eisens*, 4th ed., vol. **1a**, Springer-Verlag, Berlin–Heidelberg–New York 1964.
62. Gmelin, Durrer (eds.): *Metallurgie des Eisens*, 4th ed., vol. **5a**, Springer-Verlag, Berlin–Heidelberg–New York 1978.
63. Verein Deutscher Eisenhüttenleute (ed.): *Slag Atlas*, Verlag Stahleisen, Düsseldorf 1995.
64. D. R. Sain, G. R. Belton, "Interfacial Reaction Kinetics in the Decarburization of Liquid Iron by Carbon Dioxide," *Mettalurg. Trans.* **79B** (1976) 235 – 244.
65. D. R. Sain, G. R. Belton, "The Influence of Sulphur on Interfacial Reaction Kinetics in the Decarburization oif Liquid Iron by Carbon Dioxide," *Mettalurg. Trans.* **9B** (1978) 403 – 407.
66. T. Nagasaka, Y. Iguchi, S. Ban-ya: "Rate of Reduction of Liquid Iron Oxide with CO," *Proc. 5th Int. Steel Congr. April 6–9, 1986; Process Technol. Proceedings*, Vol. **6, Book 3**, AIME, 1986, pp. 669 – 678.
67. E. T. Turkdogan: *Fundamentals of Steelmaking*, The Institute of Materials, London 1996.
68. S. Ban-ya: "Mathematical Expression of Slag–Metal Reactions in Steelmaking Process by Quadratic Formalism Based on the Regular Solution Model," *ISIJ International* **32** (1992) 2 – 11.
F. Oeters: *Metallurgy of Steelmaking*, Verlag Stahleisen, Düsseldorf 1989.
69. P. Grassmann: *Physikalische Grundlagen der Verfahrenstechnik*, 2nd ed., Sauerländer-Verlag, Frankfurt a. M. 1961.
70. L. Zhang, F. Oeters, "A model of Post-Combustion in Iron-Bath Reactors," *Steel res.* **62** (1991) 107 – 115.
71. T. Ibaraki et al., "Experimental Operation of Smelting Reduction with a 100 mt Smelter. Part I Operation and the Slag in the Smelter. Part II Post-Combustion and Dust," *Iron Steelmaker* 1995 March, 83 – 98.
72. N. Tokumitsu et al., "Development of Smelting Reduction Process Using an Iron Bath," *Process Technol Conf. Proc.* (1982) 99 – 107.
73. R. Steffen, H. B. Lüngen: "Stand der Reduktion von Eisenerzen zu Eisenschwamm" *Stahl Eisen* **108** (1988) no. 7, 339 – 342.
74. R. Jeschar, *Arch. Eisenhüttenw.* **35** (1964) 91 – 108, 512 – 526.
75. H. P. Beer, H. Krainer, *Tech. Mitt. Krupp Forschungsber.* **24** (1966) 2, 25 – 47.
76. H. P. Beer, H. Krainer, *Arch. Eisenhüttenw.* **37** (1966) 109 – 113.
77. H. Hansen, *Arch. Eisenhüttenw.* **34** (1963) 151 – 157.
78. B. Rollinger, *Iron Steelmaker* **2** (1975) 1, 10 – 16.
79. H. Serbent, M. Zeilhofer, *Vortrag beim Europ. Treffen für Chemische Technik*, ACHEMA 1967, Frankfurt/M.
80. P. J. Kreyger., "Zusammenhang zwischen Konvertervolumen und Sauerstoffblasgeschwindigkeit," *Stahl Eisen* **96** (1976) 957 – 960.
81. F. Höfer et al., "Scrap Melting with Cost-Effective Energies," *Proceedings of the Sixth International Iron and Steel Congress 1990, Nagoya*, ISIJ, vol. 4, Steelmaking II, pp. 1 – 10.
82. VOEST-Alpine Industrieanlagen Bau (ed.): The COREX Process, Linz, Austria.
83. F. Oeters: "Fundamentals of In-Bath Smelting with Post-Combustion," in J. K. Brimacombe et al. (eds.): *Proc. Savard-Lee Int. Sympos. on Bath Smelting*, The Minerals, Metals & Materials Society, Warrendale, PA 1982.
84. Hamburger Stahlwerke GmbH, personal communication.
85. M. Hatano, T. Miyazaki, Y. Kamei: "New Ironmaking Process by Use of Pulverized Coal and Oxygen," *Proc. Process Technol. Conf.*, Vol. **6**, Iron Steel Soc., AIME 1986, pp. 1049 – 1055.
86. J. Muroki, N. Otsuki, K. Haichisuka, Y. Hara, "Nippon Steel Direct Reduction Process," Nippon Steel Techn. Report No. 12, Dec. 1978, pp. 114 – 127.
87. H. Bonnekamp et al., "Erfahrungen beim Betrieb eines Hochofens hoher Leistung," *Stahl Eisen* **103** (1983) 67 – 73.
88. European Blast Furnace Committee: Constructional Features of Sinter Plants and Blast Furnaces, VDEh, January 1999.
89. K. H. Peters, H. B. Lüngen: "The dependence of the core of a blast furnace on the supply of industrial gases," European air separation day in Cologne, Germany, October 29, 1992.

90. K. Kanbara et al., *Trans. Iron Steel Inst. Jpn.* **17** (1977) 371 – 380.
91. Y. Shimomura et al., *Trans. Iron Steel Inst. Jpn.* **17** (1977) 381 – 390.
92. M. Sasaki et al., *Trans. Iron Steel Inst. Jpn.* **17** (1977) 391 – 400.
93. K. Kojima et al., *Trans. Iron Steel Inst. Jpn.* **17** (1977) 401 – 409.
94. K. Sasaki et al., *Tetsu to Hagane*, **62** (1976) 580 – 591.
95. H. Bonnekamp, K. Engel, W. Fix, and K. Grebe, *Stahl Eisen*, **103** (1983) 75 – 82.
96. J. J. Bosley, N. B. Melcher, M. M. Harris, *J. Met.* **12** (1959, Sept.) 610 – 615.
97. Verein Deutscher Eisenhüttenleute (ed.): *Der Hochofenprozeß: Das kinetisch-dynamische Simulationsmodell des Vereins Deutscher Eisenhüttenleute*, Verlag Stahleisen mbH, Düsseldorf 1973.
98. H. B. Lüngen, H.-P. Rüther: "Casting practice at German blast furnaces," *Proc. Eur. Blast Committee, Linz, Austria*, April 25, 1996.
99. Gmelin-Durrer: *Metallurgie des Eisen,* 4th ed. vols, 2 a, b, Verlag Chemie, Weinheim 1964.
100. M. Higuchi, M. Iizuka, T. Shibuya: *Developments in Ironmaking Practice*, The Iron and Steel Institute, London 1973 p. 98.
101. G. Heynert, E. Legille, in [100] , p. 109.
102. G. Heynert, K. H. Peters, G. Ringkloff, *Stahl Eisen* **93** (1973) no. 12, 505 – 517.
103. Reference list of Paul Wurth Company, Nov. 1998.
104. Iron and Steel Inst. Japan: *Handbook of Iron and Steel*, vol. **2**, Maruzen Publishing Co., Tokyo 1979, p. 295.
105. W. Kowalski, H. B. Lüngen, K. P. Stricker: "Blast Furnace Lifetime: State of the Art and Practice for Prolonging Blast furnace Campaigns," *Cokemaking International* **11** (1999) no. 2, 20 – 30.
106. S. Mochizuki, N. Morishita, M. Higuchi, J. Mijawa: "Ofenreisen und Neuzustellungen des Hochofens 1 des Hüttenwerkes Oita der Nippon Steel Corp.," *Stahl Eisen* **116** (1996) no. 12, 83 – 90.
107. E. Akimoto et al.: "Short term revamping techniques of blast furnace," Kawasaki Technical Report no. 29, Nov. 1993, pp. 38 – 44.
108. P. Heinrich, H. Hille: "Der Kupferstave – Ein wesentlicher Beitrag zur Wirtschaftlichkeit des Hochofenverfahrens," Fachausschußbericht Nr. 1.031 des Vereins Deutscher Eisenhüttenleute (VDEh) über die 127. Vollsitzung des Hochofenausschusses, Bericht Nr. 1, Juli 1998.
109. B. Held, F. Reinitzhuber, *Stahl Eisen* **106** (1986) no. 22, 1238 – 1248.
110. [104] , p. 306.
111. *Ullmann*, 4th ed., **10**, 370.
112. M. Manes, J. S. Maikay, *J. Metal.* **14** (1962) 308 – 314.
113. VDEh Blast Furnace Committee 1995.
114. R. Jeschar, G. Dombrowski: "Summary evaluation and assessment of carbon and hydrocarbon raw materials for iron ore reduction," *1998 ICSTI/Ironmaking Conf. Proc.*, 431 – 441.
115. [104], p. 110.
116. [104], p. 120.
117. M. Nakamura, T. Sugiyama, T. Uno, Y. Hara, S. Kondo *Tetsu to Hagane* **63** (1977) 28 – 36.
118. G. Heynert, P. Ischebeck, W. V. Spee, *Stahl Eisen* **81** (1961) 1 – 12.
119. M. Hatano, K. Kurita, T. Yanaka: *Int'l Blast Furnace Hearth* and Raceway Symposium, Newcastle AIMM 1981, p. 4.1 – 4.10.
120. K. Takeda et al., *Proc. Ironmaking Conf.* **46** (1987) 191 – 198.
121. H. Lackmann: *4. Dozentenseminar des Vereins Deutscher Eisenhüttenleute*, Gelsenkirchen 1973.
122. [104], pp. 159, 160.
123. W. Hartig, K. Langner, H. B. Lüngen, K. P. Stricker: "Measures for increasing the productivity of blast furnaces," *MPT International* 1996, no. 4, 42 – 51.
124. K. H. Peters, H. Mohnkern, H. B. Lüngen: "Ways and means for low coke blast furnace operation," *Proc. 1st Int. Congr. Sci. Technol. Ironmaking, Sendai* 1994, 493 – 499.
125. "NKK′s Fukuyama blast furnace sets world record for PCI rate," *Skillings Mining Review*, Sept. 19, 1998, p. 6.
126. H.-U. Lindenberg et al.: "Rohstoffliches Recycling von Kunststoffen als Reduktionsmittel im Hochofen," *Stahl Eisen* **116** (1996) no. 8, 89 – 93.
127. K. H. Peters, E. Beppler, B. Korthas, M. Peters: "Effect of high levels of coal injection on blast furnace operating," *Proc. 2nd EIC* 1991, 247 – 263.
128. Y. Ohno et al., *7th Process tech. Conf. Proceedings*, Toronto 1988, pp. 195 – 201.
129. [99], vols. 4 a, b.
130. [99], vols. 4 a, b.
131. H. Schenk, *Stahl Eisen*, **83** (1963) 1683 – 1690.

132. H. Wysocki, *Vorlesung Entwerfen von Hüttenwerksanlagen*. Techn. Univ. Berlin 1975.
133. [104], pp. 278 – 300.
134. European Blast Furnace Committee and VDEh Blast Furnace Committee, Data Exchange of Sinter Plants and Blast Furnaces 1997/1998.
135. *Properties and Effective Uses of Iron and Steel Slag*, Nippon Slag Association, Tokyo 1996.
136. *Jahrbuch Stahl 1999*, Verlag Stahleisen mbH, Düsseldorf 1999, p. 229.
137. M. Saino: *116th and 117th Nishiyama Memorial Seminar organized by Iron Steel Institute*, Japan 1987, pp. 237 – 278.
138. Y. Hara, M. Sakawa, S. Kondo, *Tetsu to Hagane* **62** (1976) 315 – 323.
139. Iron and Steel Inst. Japan: *Blast Furnace Phenomena and Modeling*, Elsevier, London 1987, pp. 121 – 127, pp. 132 – 133.
140. W. E. Ranz, *Chem. Eng. Prog.* **48** (1952) 247 – 253.
141. J. Kudoh, J. Yagi, *Tetsu to Hagane* **73** (1987) 2020 – 2027.
142. J. Yagi, I. Muchi, *Trans. Iron Steel Inst. Jpn.* **10** (1970) 392 – 405.
143. M. Hatano, K. Kurita, *Tetsu to Hagane* **66** (1980) 1898 – 1907.
144. M. Kuwabara et al., *Proc. Joint Symp. of Iron Steel Inst. Jpn. and Australasian Institute* Mining and Metallurgy, Tokyo 1983, pp. 193 – 204.
145. S. Taguchi et al., *Tetsu to Hagane* **68** (1982) 2302 – 2310.
146. M. Hatano, K. Kurita, H. Yamaoka, T. Yokoi, *Trans. Iron Steel Inst. Jpn.* **25** (1985) 933 – 940.
147. Y. Okuno et al., *Tetsu to Hagane* **73** (1987) 91 –98.
148. T. Sugiyama, J. Yagi, Y. Omori: *Proc. 3rd Int'l Iron Steel Congr.*, 1978 Chicago, pp. 479 – 488.
149. J. Ohno et al., *International Blast Furnace Hearth & Raceway Symposium*, Newcastle, Australia 1981, pp. 10 – 12.
150. A. Poos: *Heat and Mass Transfer in Process Metallurgy*, A. W. D. Hills, 1967, pp. 1 – 38.
151. M. Hatano, K. Kurita, H. Yamaoka, T. Yokoi, in [100], pp. 941 – 948.
152. T. Sugiyama et al., *Curr. Adv. Mater. Proc. (in Japanese)*, **1** (1988) 22 – 25.
153. J. Yagi, T. Akiyama, *Curr. Adv. Mater. Proc. (in Japanese)*, **2** (1989) 2.
154. A. Nakagawa, *Tekkokai* (1986) no. 6, 5 – 9, issued by The Japan Iron and Steel Federation.
155. K. H. Peters, H. B. Lüngen: "Economical hot metal production in the blast furnace as basis for crude steel production in an integrated iron and steel works," *Proc. 6th Int. Iron Steel Congr.*, 1990, Nagoya, ISIJ, Preprint vol. 5, pp. 302 – 310.
156. D. Ameling, H. Baer, H. Bertling, H. B. Lüngen: "Cokemaking Technology 2000 – State-of-the-art-technology and new structures," *MPT International* 1998, no. 5, 46 – 56.
157. M. Wahlster, H. P. Schulz, J. Thein, *Stahl Eisen* **89** (1969) 478 – 486.
158. [104], p. 450.
159. W. Domalski, K. Fabian, D. Nolle, *Stahl Eisen* **88** (1968) 906 – 919.
160. [104], p. 443.
161. H. P. Schulz, *Stahl Eisen* **89** (1969) 249 – 262.
162. S. Eketorp, B. Kalling, *Giesserei* **46** (1959) 905 – 912.
163. [104], p. 446.
164. H. Kajioka in W. K. Lu (ed.): *Symposium on External Desulphurizations of Hot Metal*, McMaster Univ., Hamilton, Canada, 1975, p. 15–1 – p. 15–35
165. W. Meichsner, K. H. Peters, W. Ullrich, H. Knahl, *J. Met.* **26** (1974) no. 4, 55 – 58.
166. [104], p. 441.
167. [104], p. 448.
168. H. Nashiwa, S. Yamaguchi, M. Takano, M. Iwami, *Sumitomo Kinzoku* **27** (1975) 207 – 211.
169. W. H. Duquette: *Open Hearth Conf. Proc.*, **56** TMS-AIME, (1973) 79.
170. P. F. Potocic, K. G. Leewis, in [162] , pp. 3–1 – p. 3–19.
171. W. A. Voronova, *Steel USSR* **4** (1974) 261.
172. L. G. Nelson, *Proc. Ironmaking Conf.* **34** (1975) 451 – 459.
173. P. Koros, R. G. Petrushka, in [168], pp. 7–1 – p. 7–25.
174. *Ullmann* 4th ed., **10**, 395.
175. Midrex Corporation (ed.): *World Direct Reduction Statistics 1997*, Midrex Direct Reduction Corporation, Charlotte, NC 1998.
176. R. Ferrari, F. Colanti, *Iron Steel Eng.* **53** (1975) 57 – 60.
177. R. L. Stephenson, R. M. Smailer: "Direct Reduced Iron, Iron and Steel" *AIME Annu. Meet. Proc. Sess.* (Warrendale, PA) 1980.
178. L. Formanek, H. Eichberger: *Direktreduktion im Drehrohrofen*, European Coal and Steel Community (ECSC) research project; Document no. 7210-BA/101, 1985.
179. F. Oeters, A. Saatci, *Proc. Tech. Proc. ISS* **6** (1986) 1021.

180. F. Oeters, A. Saatci: *Stahl Eisen Report*, "Mass and Heat Balances," Verlag Stahleisen GmbH, Düsseldorf 1987.
181. R. J. Fruehan, K. Ito, B. Ozturk: "Analysis of bath smelting processes for producing iron", *Steel Res.* **60** (1989) no. 3/4, 129 – 137.
182. J. O. Edström: "Alternative Ironmaking Processes," *European Ironmaking Congress*, Aachen 1986.
183. R. B. Smith, M. J. Corbett: "Coal based iron making process, *7th Process Techn. Conf. Proc., AIME*, *Toronto* 1988, 147 – 178.
184. F. Oeters, R. Steffen: "Entwicklungslinien der Schmelzreduktion und des Einschmelzens*," 4. Kohle-Stahl-Kolloquium*, TU Berlin 1989.
185. S. Eketorp: "Smelting reduction," Process Ironmaking Conference," *Metallurg. Soc. AIME* **27** (1968) 36 – 39.
186. G. Papst, R. Hauk, W. Kepplinger, E. Ottenschläger: "The Corex process," *Metal. Plant. Technol. 9* (1986) no. 6, 24, 26, 28, 29.
187. R. Hauk, J. Flickenschild, E. Ottenschläger: "KR-Process hot metal production on the basis of coals," *6th Process Techn. Conf. Proc.*, AIME, Washington 1986, pp. 1031 – 1039.
188. H. Feichtner, W. Maschlanka, F. Helten: "The COREX-process," *Skilling's Min. Rev. 78* (1989) no. 2, 20 – 27.
189. H. van der Merwe, W. Kepplinger, R. Hauk, B. Vuletic: "Operation results in the first commercial COREX-plant at ISCOR Pretoria Works," *International Congr. New developments in metallurgical processing Proc.*, vol. **1**, VDEh, Düsseldorf 1989.
190. M. Hatano, T. Miyazaki, H. Yamaoka, Y. Kamei: "New Ironmaking process by use of pulverized coal and oxygen," *6th Process Techn. Conf., Proc.*, AIME, Washington 1986, pp. 1049 – 1055.
191. T. Miyazaki, H. Yamaoka, Y. Kamei, F. Nakamura: "A new ironmaking process consisted of shaft type reduction, furnace and cupola type melting furnace," *Trans. Iron Steel Inst. Jpn.* **27** (1987) 618 – 625.
192. T. Hamada et al.: "Development of Kawasaki smelting reduction process for the reduction of iron and ferroalloys," *Proceedings of the pyrometallurgy 1987* London 1987, pp. 435 – 459.
193. H. Itaya et al.: "Development of the XR-process," *7th Process Techn. Conf. Proc.*, AIME, Toronto, 1988, pp. 209 – 216.
194. L. von Bogdandy, K. Brotzmann, K. Schäfer, H.-G. Geck: *Stahl Eisen* **104** (1984) no. 22, 1143 – 1148.
195. K. Brotzmann: "New concept and methods for iron and steel production," *70th Steelmaking Conference Proceedings*, ISS-AIME; Pittsburgh 1987, pp. 3 – 12.
196. J. A. Innes, J. P. Moodie, I. D. Webb, K. Brotzmann: "Direct smelting of iron ore in a liquid iron bath – the HIsmelt process," *Process Techn. Conf. Proc.*, AIME, Toronto 1988, pp. 225 – 231.
197. J. Barin, M. Lemperle, M. Modigell: "Smelting reduction of iron ore," *Steel Res.* **60** (1989) no. 3/4, 120 – 121 .
198. K. Iwasa et al.: *Curr. Adv. Mat. Proc.* **1** (1988) no. 4, 1085.
199. K. Iwasaki, H. Kawata, K. Yamada, T. Kitagaw: "Integrated test plant work on smelting reduction ironmaking process," *International Congr. New Developments in Metallurgical Processing Proc.*, vol. 1, VDEh, Düsseldorf 1989.
200. N. Tokumitsu et al.: "Development of smelting reduction process using an iron bath," *7th Process Techn. Conf. Proc.*, AIME Toronto 1988, pp. 99 – 107.
201. J. Hartwig, D. Neuschütz, D. Radke, F. Seelig: "Entwicklung eines Einschmelzverfahrens mit kombinierter Direktreduktion unter Einsatz beliebiger Feinkohlen," *Stahl Eisen* **100** (1980), 535 – 543.
202. J. Hartwig, D. Neuschütz, D. Radke, W. D. Röpke: "The Krupp COIN Process Combined with Direct Reduction," *UNECE Seminar*, Noordwijkerhout, Netherlands, May 1983.
203. D. Neuschütz, T. Hoster: "Concept and present state of a coal-based smelting reduction process for iron ore fines," *Steel Res.* **60** (1989) no. 3/4, 113 – 119.
204. P. Collin, H. Stickler: "Ironmaking alternative paves the way to cheaper steel," *Iron Steel Int.* **53** (1980) 81 – 86.
205. H. Stickler: "Varianten des Elred-Verfahrens," *Stahl Eisen* **104** (1984) 539 – 541.
206. H.. Elvander, R. A. Westman: *Iron Steel Eng.* **59** (1982) no. 4, 57 – 60.
207. H. Elvander, G. Omberg: "Das Inred-Iron-Verfahren für die Roheisenerzeugung," *Stahl Eisen* **104** (1984) 864 – 866.
208. H. G. Herlitz, B. Johansson, S. O. Santen: "A new family of reduction processes based on plasma technology," *Iron Steel Eng*, **61** (1984) 39 – 44.
209. J. Feinmann (ed.): *Plasma Technology in metallurgical processing*, Iron and Steel Society, Warrendale 1987, p. 111.

210. R. Garraway, *Iron Steelmaker* **23** (1996) June, 27 – 30.
211. M. Hirsch, R. Husain, P. Weber, H. Eichberger, International Conference "Prereduced Products in Europe," Milan, Italy, September 23 – 24, 1996.
212. L v. Bogdandy, H.-J. Engell: *The Reduction of Iron Ores*, Springer-Verlag and Verlag Stahleisen, Berlin – Heidelberg – New York – Düsseldorf 1971.
213. L. Zhang: *"Mathematische Modelle der Schmelzreduktion von Eisenerz und des Vorheizens von Schrott im Gegenstrom*," Thesis Technische Universität Berlin, Berlin 1994.
214. F. Oeters, A. Saatçi: *Mass and Heat Balances During the Reduction of Iron Ores*, Verlag Stahleisen, Düsseldorf 1987.
215. "*Present state and development potential of the SL/RN process*" in: 9th Seminar on the Economic Problems Related to the Creation of Steel Plants, Using Direct Reduction. Noordwijkerhout Nederland, May 6 – 10, 1983, Report R 26.
216. F. Oeters, R. Steffen: "Entwicklunglinien der Schmelzreduktion," *Stahl Eisen* **109** (1989) 728 – 742.
217. S. Santén, *Jernkontorets Ann.* 1980, no. 185 – 189.
218. A. Rist, N. Meysson, *Rev. Met.* **61** (1964) 121 – 145.
219. A. F. Eberle, E. Schiefermair, H. Freydorfer, D. Siuka: "Hot metal and DRI production with Corex," IISI TECHCO-31, Technical Exchange Session, Praha, April 26/27, 1999, International Iron and Steel Institute (IISI), Brussels 1999.
220. J. L. Schenk, F. Wallner, W. L. Kepplinger, M. K. Shin, M. Y. Cho, I. O. Lee: "Technology for Increased Portion of Fine Ore in Corex Process," Scanmet I, Proceedings 1st Int. Conf. on Process Development in Iron and Steelmaking, vol. 1, Lulea, June 7/8, 1999, pp. 275 – 291.
221. S. Joo, H. G. Kim, I. O. Lee: "Finex – a new Process for Production of Hot Metal from Fine Ore and Coal," Scanmet I, Proceedings 1st Int. Conf. on Process Development in Iron and Steelmaking, vol. 1, Lulea, June 7/8, 1999, pp. 293 – 300.
222. H.-K. Lee, S. Joo, C.-O. Kang: "The Corex Process for Hot Metal Production without Coke," Metec Congress 99, Proceedings Int. Conf. on New Developments in Metallurgical Process Technology, Düsseldorf, June 13/15, 1999, pp. 36 – 41.
223. R. J. Dry, C. P. Bates, D. P. Price: "HIsmelt – Competitive Hot Metal from Ore Fines and Steel Plant Wastes," Metec Congress 99, Proceedings Int. Conf. on New Developments in Metallurgical Process Technology, Düsseldorf, June 13/15, 1999, pp. 58 – 65.
224. The HIsmelt Process. Kwinana Vertical SRV. Web Site by Rio Tinto Technology - R & TD Perth. June 1999.
225. T. Kitagawa, K. Iwasaki, M. Kawakami, K. Takahashi: "The DIOS process – the Advanced Smelting Technology for the 21st Century," Metec Congress 99, Proceedings Int. Conf. on New Developments in Metallurgical Process Technology, Düsseldorf, June 13/15, 1999, pp. 42 – 49.
226. H. K. A. Meijer et al.: "Cyclone Converter Furnace," *Steel Times* **224** (1996) 396 – 397.
227. E. Aukrust: "Results of the AISI/DOE Direct Steelmaking Program," Proceedings Savard/Lee Symposium on Bath Smelting, Montreal, 1992, 591 – 610.
228. "The CCF process," Company brochure of Hoogovens Steel Primary Products, IJmuiden, February 1998.
229. J. Sofra, E. Mounsey: "Recent Pilot Scale Development of the AusIron Direct Ironmaking Process," Scanmet I, Proceedings 1st Int. Conf. on Process Development in Iron and Steelmaking, vol.1, Lulea, June 7/8, 1999, pp. 301 – 324.
230. A. Chatterjee: "Technical Evaluation of the Romelt Process for Possible Application under Indian Conditions," Scanmet I, Proceedings 1st Int. Conf. on Process Development in Iron and Steelmaking, vol. 1, Lulea, June 7/8, 1999, pp. 325 – 335.
231. P. Sandhier, A. N. Siderov, *Iron & Steel Review* 1998, Juli, 61 – 68.
232. M. Millet: "CSP start-up and production results serving automotive," Metec Congress 99, Int. Conf. on New Developments in Metallurgical Process Technology, Düsseldorf, June 13/15, 1999 (Paper distributed during conference).
233. R. L. Schultz, P. Fontana, H. Lehmkühler: "Redsmelt: An Alternative Ironmaking Technology to Improve Economics and Environmental Compatibility for Pig Iron Production," Metec Congress 99, Proceedings Int. Conf. on New Developments in Metallurgical Process Technology, Düsseldorf, June 13/15, 1999, pp. 50 – 57.

234. G. E. Hoffman, K. J. Shoop, Y. Tanigaki, K. Tokuda: "The Current Status of ITmk3 Technology Development," Report at the Advanced Technology Symposium of I.S.S. "The Production and Use of Scrap Substitutes – Part III", Port of Spain, Feb. 28/March 3, 1999.
235. *ARBED News*, 1999, Jan., 2 – 4.
236. T. Matsui, Y. Hara, N. Ishiwata, H. Itaya: "Smelting Reduction Process with a Coke Packed Bed for Electric Arc Furnance Dust Recycling," Proceedings 10th Japan – Germany Seminar, The Iron and Steel Institute of Japan, Tokyo, May 18/19, 1999, 31 – 38.
237. K. Mori, M. Shinkai, T. Okamoto: "Development of a New Dust and Slag Treatment Technology," Proceedings 10th Japan – Germany Seminar, The Iron and Steel Institute of Japan, Tokyo, May 18/19, 1999, 47 – 54.
238. F. Stein, *Stahl Eisen* **49** (1929) 1269 – 1270.
239. Akademischer Verein Hütte (ed.): *Hütte, Taschenbuch für Eisenhüttenleute*, 5th ed., Verlag Wilhelm Ernst u. Sohn, Berlin 1961.
240. H. Remy: *Lehrbuch der anorganischen Chemie,* 11th ed., vol. 2, Akad. Verlagsges. Geest & Portig KG, Leipzig 1961.
241. I. Ljungberg, *Stahl Eisen* **74** (1954) 279 – 285.
242. H. W. Dener, J. Elze: *Handbuch der Galvanotechnik,* vol. 2, Carl Hanser Verlag, München, 1966.
243. *Chem. Eng. News* **46** (1968) 10, 17.
244. H. Schildknecht: *Zonenschmelzen*, Verlag Chemie GmbH, Weinheim 1964, p. 167.
245. M. Badiali, R. Bakisch, N. Kirschenbaum, in "Reinststoffe in Wissenschaft und Technik," Symposium Nov. 1961 Dresden, Akademie Verlag Berlin 1963.
246. *Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft)* 27. Febr. 1986.
247. D. Eickelpasch et al., *Stahl Eisen* **100** (1980) no. 6, 260 – 270.
248. J. A. Philipp et al., *Stahl Eisen* **107** (1987) no. 11, 507 – 514.
249. H. J. White: *Industrial Electrostatic Precipitation*, Addison-Wesley Publishing Company, Inc., Reading Massachusetts 1963.
250. G. Mayer-Schwinning, R. Rennhack, *Chem.-Ing.-Techn.* **52** (1980) no. 5, 375 – 383.
251. S. Oglesby Jr., G. B. Nichols: *Electrostatic Precipitation (Pollution engineering and technology; 8)*, Marcel Dekker, New York 1978.
252. R. Bothe, *Stahl Eisen* **88** (1968) no. 25, 1414 – 1422.
253. H. Kahnwald, *Stahl Eisen* **104** (1984) no. 7, 351 – 356.
254. P. O. Spawn, T. J. Maslany, *JAPC* **31** (1981) 1060.
255. B. Bussmann et al., *Stahl Eisen* **105** (1985) no. 3, 121 – 130.
256. R. Görgen, W. Theobald, *Stahl Eisen* **97** (1977) no. 14, 657 – 664.
257. W. Theobald, G. Schnegelsberg, *Berghüttenmann Monatsh.* **116** (1971) no. 9, 328 – 340.
258. G. Metcalf, G. Eddy: *Wastewater Engineering*, Mc Graw-Hill Book Company, Boston 1979.
259. L. L. Beranek: *Noise Reduction*, McGraw Hill Book Company, London, 1960.
260. K. Althoff, *Stahl Eisen* **107** (1987) no. 22, 1071 – 1075.
261. H. U. Haering et al., *Stahl Eisen* **102** (1982) no. 15/16, 749 – 754.
262. J. A. Philipp, H. Maas, *Stahl Eisen* **104** (1984) no. 8, 403 – 407.
263. E. Mertins, *Erzmetall* **39** (1986) no. 7/8, 399 – 404.
264. H. Serbent, *Techn. Mitt.* **71** (1978) no. 11/12, 569 – 575.
265. G. Meyer et al., *Stahl Eisen* **96** (1976) no. 24, 1228 – 1233.
266. H. Maczek, R. Kola, *J. Met.* **32** (1980) no. 1, 53 – 58.
267. G. Kossek et al., *Erzmetall* **32** (1979) no. 3, 135 – 139.
268. W. Kaas et al., *Stahl Eisen* **98** (1978) no. 24, 1277 – 1281.
269. IISI Steel Statistics, year book 2001, pp 3, 4, 96, 97.
Statistisches Jahrbuch Stahl, WV 2001, p. 297, p. 318, p. 378.
270. M. Peters, P. Schmöle, *Stahl Eisen* **122** (2002) no. 6, forthcoming.
271. Amtliche Außenhandelsstatistik WV Stahl
272. World Steel in Figures, IISI, 2001, Edition, p. 18.
273. H. B. Lüngen, K. Mülheims, R. Steffen, *Stahl Eisen* **121** (2001) no. 5, p. 35 – 47.
274. H. B. Lüngen, R. Steffen, *Stahl Eisen* **118** (1998) no. 3, p. 87 – 93.

Related Articles

Ferroalloys, Magnetic Materials, and Steel are separate keywords.

Iron Compounds

Iron Cyanides → Cyano Compounds, Inorganic; Iron Blue Pigments, Iron Oxides → Pigments, Inorganic

Egon Wildermuth, Leverkusen, Federal Republic of Germany (Chap. 1)

Hans Stark, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2)

Gabriele Friedrich, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 3)

Franz Ludwig Ebenhöch, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 3)

Brigitte Kühborth, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 3.7)

Jack Silver, University of Essex, Colchester, United Kingdom (Chaps. 4 and 5)

Rafael Rituper, Keramchemie GmbH, Siershahn, Federal Republic of Germany (Chap. 6)

1. Iron Sulfates 1
1.1. Iron(II) Sulfate 1
1.2. Iron(III) Sulfate 2
2. Iron Chlorides 3
2.1. Iron(III) Chloride 3
2.2. Iron(II) Chloride 5
3. Iron Pentacarbonyl 6
3.1. Properties 6
3.2. Production 7
3.3. Quality Specifications and Analysis 8
3.4. Safety, Storage, and Transportation 9
3.5. Uses 10
3.6. Economic Aspects 11
3.7. Toxicology and Occupational Health 12
4. Iron Compounds, Miscellaneous . . 12
5. Iron Compounds for the Treatment of Anemia 13
6. Regeneration of Iron-Containing Pickling Baths 14
6.1. Sulfuric Acid Pickling Solutions . . 15
6.1.1. Crystallization 15
6.1.2. Electrolysis 16
6.2. Hydrochloric Acid Pickling Solutions 16
6.3. Nitric and Hydrofluoric Acid Pickling Solutions 18
7. References 20

1. Iron Sulfates

1.1. Iron(II) Sulfate

Properties. Iron(II) sulfate heptahydrate [*7782-63-0*], ferrous sulfate, copperas, green vitriol, $FeSO_4 \cdot 7\,H_2O$, M_r 278.02, *mp* 64 °C, ϱ 1.898 g/cm^3, crystallizes from aqueous solution as green, monoclinic crystals, which are readily soluble in water and glycol but insoluble in alcohol, acetone, and methyl acetate. On being heated to 60 – 70 °C, 3 mol of water is driven off to form the tetrahydrate [*20908-72-9*], $FeSO_4 \cdot 4\,H_2O$. In the absence of air, at ca. 300 °C, iron(II) sulfate monohydrate, a white powder, is formed. When heated to ca. 260 °C in the presence of air, the monohydrate is oxidized to iron(III) sulfate. Decomposition of iron(II) sulfate begins at about 480 °C. Iron(II) sulfate is efflorescent in dry air. In moist air, the surface of the crystals becomes covered with brownish basic iron(III) sulfate. Aqueous solutions of iron(II) sulfate are slightly acidic due to hydrolysis. The rate of oxidation by air increases in alkaline solution and on raising the temperature. Iron(II) sulfate reduces copper(II) ions to copper(I) ions and gold or silver ions to the corresponding metal. With univalent cations (e.g., K, Rb, Cs, NH_4, and Tl) and divalent cations (e.g., Cd, Zn, Mn, and Mg), double salts (alums) are readily formed; an example is Mohr's salt, iron(II) ammonium sulfate, $FeSO_4(NH_4)SO_4 \cdot 6\,H_2O$. This green salt is soluble in water, easily purified, and more stable to oxidation than iron(II) sulfate heptahydrate.

Production. Iron(II) sulfate is produced by dissolving scrap iron in dilute sulfuric acid or by oxidizing moist pyrites in air.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_591

Iron(II) sulfate is also obtained as a byproduct from steel pickling (see Chap. 6) and from the production of titanium dioxide (sulfate process, with ilmenite as raw material). It is recovered from pickling liquors in a relatively pure form. The purity of iron(II) sulfate from the production of titanium dioxide depends on the source of the ilmenite. A typical composition is 87 – 90 % $FeSO_4 \cdot 7\,H_2O$, 6 – 7 % $MgSO_4 \cdot 7\,H_2O$, 0.6 – 0.7 % $TiOSO_4$, 0.2 % $MnSO_4 \cdot 7\,H_2O$, 3 – 6 % water, and depending upon the type of ilmenite used, traces of nickel, chromium(III), vanadium, copper, zirconium, and others.

Uses. Iron(II) sulfate is used for the preparation of other iron compounds. It is employed to a lesser extent in writing inks, wood preservatives, desulfurizing coal gas, and the production of Turnbull's blue, Prussian blue, and other iron pigments; as an etchant for aluminum; for process engraving and lithography; and as an additive to fodder.

As sensitivity to environmental protection increased, other uses had to be found for iron(II) sulfate, because the above mentioned applications did not consume the amounts available as a byproduct from industrial processes. Large amounts of iron(II) sulfate are used for the clarification of community effluents. With an additional clarification step, up to 90 % of the phosphate in the effluent can be eliminated. Sludge formed in clarification tanks may be used as fertilizer [1], [2]. Iron(II)sulfate can also be recycled. After drying to the monohydrate, it is decomposed in a calciner:

$$4\,FeSO_4 \longrightarrow 2\,Fe_2O_3 + 4\,SO_2 + O_2$$

Sulfur dioxide is converted by the contact process to sulfuric acid [3]. The conversion of iron(II) sulfate to gypsum and iron(II) chloride by treatment with calcium chloride has also been proposed. The iron(II) chloride is then decomposed to iron(III) oxide and hydrochloric acid [4]. To convert surplus iron(II) sulfate to a product that can be disposed on a dump site, it can be neutralized with one equivalent of calcium hydroxide. The conversion is reportedly 97 % [5]. To regulate the setting characteristics of cement, gypsum can be replaced by iron(II) sulfate [6]. As an additive to cement, iron(II) sulfate can reduce the content of water-soluble chromates substantially [7]. Iron(II) sulfate solutions can be used to eliminate chlorine in waste gases [8]. Effluents containing cyanides and chromates may be decontaminated with iron(II) sulfate [9]. Iron(II) sulfate can be used to combat chlorosis, a disease of vines [6]. It is also used to treat alkaline soil and to destroy moss [6].

Toxicology. If swallowed, iron(II) sulfate can cause disturbances of the gastrointestinal tract. Ingestion of large quantities by children may cause vomiting, hematemesis, hepatic damage, and peripheral vascular collapse [10].

The acute oral toxicity (LD_{50}) of iron(II) sulfate in the rat is 1389 – 2778 mg/kg, and the in rabbit, 2778 mg/kg [11].

1.2. Iron(III) Sulfate

Properties. Anhydrous iron(III) sulfate [*10028-22-5*], ferric sulfate, $Fe_2(SO_4)_3$, M_r 399.88, is a yellowish-white solid. It dissolves in water, hydrolyzing at the same time, to yield a brownish solution. Basic iron(III) sulfate precipitates when the solution is boiled. With alkali metal (M) sulfates, iron(III) sulfate forms alums, $M^I Fe(SO_4)_2 \cdot 12\,H_2O$. Alums are almost colorless but normally have a violet tinge.

Production. Iron(III) sulfate is prepared by treating iron(II) sulfate with boiling concentrated sulfuric acid, or by evaporating a mixture of iron(III) oxide and sulfuric acid. Iron(III) sulfate solutions are produced industrially by injecting chlorine gas into an iron(II) sulfate solution [12]. The solution thus obtained contains a mixture of iron(III) sulfate and iron(III) chloride.

Uses. Ferric sulfate is used to prepare alums and iron oxide pigments, and as a coagulant for the treatment of liquid effluents. Ammonium ferric sulfate is used for tanning.

Solutions of iron(III) compounds are used to reduce the volume of sludge from effluent treatment plants. By using a mixture of iron(III) sulfate, iron(III) chloride, and calcium hydroxide it is possible to achieve a dry-matter content of the sludge, after filtration, of at least 35 % [13].

2. Iron Chlorides

Of the halides of iron, only iron(II) chloride and iron(III) chloride have become important commercially.

2.1. Iron(III) Chloride

Iron(III) chloride [*7705-08-0*], ferric chloride, is a byproduct of some metallurgical and chemical processes, such as the chlorinating decomposition of iron-bearing oxide ores. Its ready availability and the cheap feeds used to produce it (iron and chlorine) have made iron(III) chloride, especially its aqueous solution, a significant feed stock for many industries, in particular, water treatment and the production of iron oxides or other iron compounds.

Properties. Iron(III) chloride, $FeCl_3$, M_r 162.206, sublimation temperature ca. 305 °C, *bp* 332 °C [14], ΔH_f^0 (25 °C) – 399.67 kJ/mol [14], ϱ (25 °C) 2.89 g/cm^3 [15], is an almost black crystalline solid.

It is dimeric (Fe_2Cl_6) in the gas phase up to about 400 °C, dissociates at higher temperature, and is monomeric at 750 °C in the presence of excess chlorine. If chlorine is not present in excess, the compound decomposes to iron(II) chloride and chlorine above 200 °C [16].

Anhydrous iron(III) chloride has a hexagonal layer structure and forms dark, almost black, microcrystalline platelets with a metallic luster. The strongly hygroscopic crystals form a series of hydrates in humid air ($FeCl_3 \cdot x\,H_2O$, $x = 6$, 3.5, 2.5, 2) [17] and deliquesce on absorbing more moisture. The crystalline commercial product is iron(III) chloride hexahydrate [*10025-77-1*], $FeCl_3 \cdot 6\,H_2O$, M_r 270.30, *mp* ca. 37 °C. It forms yellow crystals and occurs as $[FeCl_2(H_2O)_4]^+Cl^- \cdot 2\,H_2O$.

Iron(III) chloride is very soluble in water: 430 g of $FeCl_3$ per kilogram of solution at 0 °C, 480 g/kg at 20 °C, and 743 g/kg at 40 °C [18]. The enthalpy of solution of anhydrous iron(III) chloride is 95 kJ/mol for the formation of a 40 % solution at 20 °C. The aqueous solution is strongly acidic. If a solution of iron(III) chloride is diluted and neutralized slowly with alkali, $[Fe(H_2O)_6]^{3+}$ is deprotonated to give yellow $[Fe(OH)(H_2O)_5]^{2+}$ and dimeric $[Fe_2(OH)_2(H_2O)_8]^{4+}$. Brown, colloidal β-FeOOH is formed upon further neutralization; at higher temperature, iron hydroxide precipitates [19].

Iron(III) chloride is readily soluble in liquids having donor properties, such as alcohols, ketones, ethers, nitriles, amines, and liquid sulfur dioxide, but only sparingly soluble in nonpolar solvents such as benzene and hexane. The reaction of iron(III) chloride with hydrochloric acid or chlorides yields the tetrachloroferrates $[FeCl_4]^-$ and $[FeCl_4(H_2O)_2]^-$. Because complexes are also formed with donor solvents, iron(III) chloride can be extracted from aqueous solutions in the presence of hydrochloric acid by using ether [20].

Iron(III) chloride is a strong oxidizing agent. Thus, many metals (for example, Fe, Cu, Ni, Pd, Pt, Mn, Pb, and Sn) are dissolved by aqueous iron(III) chloride, with the formation of iron(II) chloride; magnesium dissolves with evolution of hydrogen. Iron(III) chloride solution liberates carbon dioxide from alkali-metal carbonates. When iron(III) chloride is heated in air, iron(III) oxide and chlorine are produced. Above 200 °C, iron(III) chloride is rapidly reduced by hydrogen to iron. It is also reduced by coal.

Production. Anhydrous iron(III) chloride is produced industrially by reaction of dry chlorine with scrap iron at 500 – 700 °C [21]. The process is known as direct chlorination. Preheated scrap is charged into the top of a water-cooled iron shaft furnace, while chlorine (or a mixture of chlorine and nitrogen) is admitted at the bottom. To prevent formation of iron(II) chloride, a 10 – 30 % excess of chlorine is used. Iron scrap is added continuously at a rate depending on the reaction temperature and the amount of chlorine remaining in the exit gas. The iron(III) chloride vapor produced passes through water- or air-cooled condensation chambers. The crystals obtained are removed continuously from the walls by vibration or knocking. They are then ground, screened, and packaged with exclusion of moisture. Chlorine in the waste gas is removed either by scrubbing with iron(II) chloride solution, to give iron(III) chloride, or by reacting with sodium hydroxide solution, to yield sodium hypochlorite and sodium chloride.

In another process, carried out in a reactor with an acid-resistant liner, iron scrap and dry

chlorine gas react in a eutectic melt of iron(III) chloride and potassium or sodium chloride (e.g., 70 % $FeCl_3$ and 30 % KCl) [22]. First, iron scrap is dissolved in the melt at 600 °C and oxidized to iron(II) chloride by iron(III) chloride. Iron(II) chloride then reacts with chlorine to yield iron(III) chloride, which sublimes and is collected in cooled condensation chambers.

The advantages of direct chlorination include higher space – time yields, lower capital cost, and less waste. The melt process gives a purer product because impurities remain largely in the melt; however, for this reason the melt must be changed and disposed of frequently.

Iron(III) chloride solutions are prepared by dissolving iron in hydrochloric acid and oxidizing the resulting iron(II) chloride with chlorine. In a continuous, closed-cycle process, iron(III) chloride solution is reduced with iron and the resulting iron(II) chloride solution is reoxidized with chlorine. Iron(III) chloride solution is also produced by dissolving iron(III) oxide in hydrochloric acid. Solid iron(III) chloride hexahydrate is crystallized by cooling a hot concentrated solution.

Quality Specifications and Analysis. Commercial forms of iron(III) chloride include sublimed anhydrous iron(III) chloride ($\geq$ 99 % $FeCl_3$), iron(III) chloride hexahydrate (ca. 60 % $FeCl_3$), and iron(III) chloride solution (ca. 40 % $FeCl_3$). Iron(III) chloride for potable water treatment must comply with DIN 19 602 (February 1987).

Trivalent iron is determined by iodometric titration with sodium thiosulfate; chloride ion is determined by argentometry.

Occupational Health. Iron(III) chloride irritates the skin and, especially, eyes. Therefore, workers handling it must wear protective glasses, facial protection, and rubber gloves. The TLV for iron(III) chloride is that for soluble iron salts: 1 mg of iron per cubic meter.

Storage and Transportation. Anhydrous iron(III) chloride can be stored in standard steel containers. It is shipped airtight in sheet-iron drums or plastic barrels, and loose in tank trucks. Moist iron(III) chloride and iron(III) chloride solutions attack ordinary metals. The following materials are suitable for containers to hold iron chloride solutions: rubberized steel; plastics such as polyethylene, poly(vinyl chloride), and polytetrafluoroethylene; fiber-reinforced plastics; glass; stoneware; porcelain; and enamel. The only suitable metallic materials are titanium, tantalum, and Hastelloy C.

The transport classification for anhydrous iron(III) chloride is UN No. 1773, RID – ADR Class 8.11 c, and for iron(III) chloride solution, UN No. 2582, RID – ADR Class 8.5 c.

Iron(III) chloride spilled by accident must be collected and forwarded to a sewage plant for disposal. Contaminated soil must be cleaned with a large amount of water and neutralized with lime. Iron(III) chloride is a Class 1 hazard to water quality.

Uses. Anhydrous iron(III) chloride is used in organic chemistry as a chlorinating agent for aliphatic hydrocarbons and for aromatic compounds. It also acts as a catalyst in Friedel – Crafts syntheses and condensation reactions. Iron(III) chloride is also used occasionally as an oxidizing agent [23], [24].

Because solutions of iron(III) chloride dissolve metals such as copper and zinc without troublesome evolution of hydrogen, they are used for the etching or surface treatment of metals (e.g., in the manufacture of electronic printed circuits, copperplate rotogravure printing, textile printing rolls), and for making signs. Iron(III) chloride solutions are also used in leaching the copper ore chalcopyrite (→ Copper, Chap. 5.10.).

Iron(III) chloride is used as a starting material in the preparation of other iron compounds, particularly oxides and hydroxides. However, most iron(III) oxide pigments obtained by hydrolysis of iron(III) chloride at high temperature have weak colors and a bluish cast, so these processes have not achieved commercial importance.

The largest use of iron chloride is in the form of dilute solutions that are employed as flocculating and precipitating agents in water treatment [25]. In processing surface waters into potable and process water, 5 – 40 g of iron(III) chloride is added to 1 m^3 of crude water at pH 6 – 7. The precipitated iron(III) hydroxide adsorbs finely divided solids and colloids (e.g., clays and humic acids). If decarbonization or the precipitation of heavy metals is to be carried out at the

same time, slaked lime is added to adjust the pH to 9 – 11.

In municipal and industrial wastewater treatment, iron(III) chloride is particularly effective because it precipitates heavy metals and sulfides, whereas contaminants such as oils and polymers, which are difficult to degrade, are adsorbed on the iron hydroxide flocs. Phosphates are also lowered to a level of ca. 1 mg of phosphorus per liter of water in the clarifier discharge. Preliminary treatment with iron(III) chloride can increase the capacity of overloaded clarifier plants.

Sludge conditioning with iron(III) chloride and lime improves the dewatering of filter sludge, so that drier sludges, better suited to disposal or incineration, are obtained.

Consumption of iron(III) chloride in the United States in 1994 was as follows [26]:

Sewage and wastewater treatment	65 %
water treatment	25 %
Catalyst, etchant, and other	10 %

Solutions of iron chlorosulfate, $FeClSO_4$, obtained by chlorinating iron(II) sulfate, are also used extensively as flocculants in water treatment [28].

Economic Aspects. Of the few anhydrous iron(III) chloride producers worldwide, the most important in Europe is BASF. Most of the iron(III) chloride is obtained as a byproduct when iron-bearing raw materials are chlorinated or when steels are pickled with hydrochloric acid (see Chap. 6); the major commercial form is the 40 % solution.

Outputs (by year, as 100 % $FeCl_3$) and major producers are

Europe (1994)	estimated 150 000 t (Solvay, Atochem, BASF, SIDRA)
USA (1994) [26]	200 000 t (Du Pont, PVS, Imperial West Chem.)
Japan (1987) [27]	322 000 t

The most important European producers of iron chlorosulfate are Kronos Titan and Thann et Mulhouse SA.

2.2. Iron(II) Chloride

Iron(II) chloride [*7758-94-3*], ferrous chloride, is much less important industrially than iron(III) chloride. The etching of steel with hydrochloric acid yields large amounts of iron(II) chloride, which is chlorinated in solution or thermally decomposed to iron oxide and hydrochloric acid. A solution of iron(II) chloride is also obtained as a byproduct from the chloride process for the production of titanium dioxide.

Properties. Iron(II) chloride, $FeCl_2$, M_r 126.753, *mp* 677 °C, *bp* 1074 °C, ΔH_f^0 (25 °C) − 342.1 kJ/mol, ϱ (25 °C) 3.16 g/cm^3, forms pale green, rhombohedral crystals which are readily soluble in water, alcohol, and acetone, but only sparingly soluble in ether and benzene. The compound occurs naturally as the mineral lawrencite. It forms several hydrates, $FeCl_2 \cdot x\,H_2O$ ($x = 1, 2, 4, 6$).

When heated in air, iron(II) chloride forms iron(III) chloride and iron oxides. Iron(II) chloride solution undergoes virtually no hydrolysis; it is oxidized readily by air.

Production. Anhydrous iron(II) chloride is produced from iron filings heated to red heat in a stream of hydrogen chloride or from iron and chlorine at 700 °C. The product is obtained as a sublimable white mass. Iron(II) chloride can also be prepared by passing hydrogen over heated iron(III) chloride.

Solutions of iron(II) chloride are prepared by dissolving iron in hydrochloric acid or by reducing iron(III) chloride solutions with iron. When a hot saturated solution is cooled to 0 °C, pale green, monoclinic crystals of the hexahydrate are obtained; cooling to room temperature yields the blue-green tetrahydrate. If the solution is concentrated at 90 °C, the green dihydrate is obtained; above 120 °C or on vacuum evaporation, the dihydrate is converted to the monohydrate, $FeCl_2 \cdot H_2O$. In the absence of air, the monohydrate is converted to anhydrous iron(II) chloride at 230 °C [29].

Quality Specifications. Iron(II) chloride is sold as an anhydrous powder (≥96 % $FeCl_2$), as the dihydrate (>75 % $FeCl_2$), as the tetrahydrate (>60 % $FeCl_2$), and as an aqueous solution (ca. 30 % $FeCl_2$).

Transportation. Solid iron(II) chloride is shipped in sheet-iron drums, plastic bags, or glass containers, and its solution is transported in rubberized containers or plastic barrels.

Uses. Iron(II) chloride is used as a reducing agent and for the production of other iron compounds. Increasingly, iron(II) chloride solution serves as a precipitating and flocculating agent with reducing properties for use in water treatment; it is especially effective with wastewater containing chromate, for example.

Pure iron(II) chloride solution is an important starting material for the preparation of acicular goethite, α-FeOOH, and lepidocrocite, γ-FeOOH, which are further processed to γ-Fe_2O_3 magnetic pigments. Pure iron(II) chloride is also employed in the preparation of pure iron.

3. Iron Pentacarbonyl

Metal carbonyls are complexes in which carbon monoxide is coordinated to the central metal atom (→ Organometallic Compounds and Homogeneous Catalysis). Of the iron carbonyls, only iron pentacarbonyl [*13463-40-6*], $Fe(CO)_5$, is economically significant.

M. Berthelot [30] and, independently, L. Mond and L. Quincke [31] discovered iron pentacarbonyl in 1891 [32]. Only after BASF had developed high-pressure technology for ammonia synthesis was its industrial production feasible. In 1925, a large plant was built in Germany by BASF to make iron pentacarbonyl, which had proved to be a very effective antiknock agent for Otto engines. When the compound was replaced shortly afterwards by lead alkyls, which were beneficial to motors, the production of iron powder from iron pentacarbonyl came to the fore.

In 1940, GAF began producing the pentacarbonyl in the United States under license from BASF. A BASF plant was moved to the former Soviet Union in 1945. Plants built in England (1944) and France (1952) have ceased production. In addition to iron powder, BASF has been producing pure iron oxide from the carbonyl since about 1930.

3.1. Properties

Physical Properties. Iron pentacarbonyl, M_r 195.9, is a clear, yellow, mobile liquid with an indistinctive odor. The molecule has a trigonal-bipyramidal structure [33].

Endothermic decomposition to iron and carbon monoxide begins at ca. 60 °C. Although decomposition becomes marked at the boiling point, iron pentacarbonyl can be distilled at atmospheric pressure with slight losses.

Some physical properties of iron pentacarbonyl are as follows:

mp	− 20 °C
bp	103 °C
Critical temperature	285 – 288 °C
Critical pressure	2.90 MPa
Heat of fusion	69.4 kJ/kg
Heat of vaporization	190 kJ/kg
Specific heat (20 °C)	1.2 kJ kg^{-1} K^{-1}
Viscosity (20 °C)	76 mPa · s
Surface tension (20 °C)	0.022 N/m
Coefficient of linear expansion (20 °C)	0.00125
Refractive index (n_D^{22})	1.518
Thermal conductivity	0.139 W m^{-1} K^{-1}
Enthalpy of dissociation (vapor)	994 kJ/kg
Heat of combustion (to CO_2 and Fe_2O_3)	− 8200 kJ/kg
Density, g/cm^3	
at 0 °C	1.495
at 20 °C	1.457
at 40 °C	1.419
at 60 °C	1.380
Vapor pressure, kPa	
at 20 °C	3.49
at 60 °C	21.33
at 100 °C	87.58

Iron pentacarbonyl is completely miscible with petroleum ether, hexane, benzene, pentanol and higher alcohols, ethyl ether, acetone, acetic acid, and ethyl acetate. It is partially miscible with paraffin oil and lower alcohols up to butanol [35].

Water solubility data for iron pentacarbonyl are contradictory; a value of 50 – 100 mg/L can be assumed [36]. The solubility of water in iron pentacarbonyl is 200 – 400 mg/kg.

The most complete documentation on iron pentacarbonyl is given in [37]; for earlier literature, see [38].

Chemical Properties. Iron pentacarbonyl is an easily combustible substance. It does not react with water or with weak or dilute acids. With

concentrated acids, the corresponding iron salts are formed with the evolution of carbon monoxide and hydrogen. Reactions with halogens yield iron halides. Iron pentacarbonyl also reduces organic compounds; for example, nitrobenzene is reduced to aniline; ketones to alcohols; and indigo to indigo white. The Hieber base reaction yields iron carbonyl hydride or its salts [41–45]:

$$Fe(CO)_5 + 4\,NaOH \longrightarrow Na_2Fe(CO)_4 + Na_2CO_3 + 2\,H_2O$$

The salt $Na_2Fe(CO)_4$ is a strong reducing agent.

Visible-light photolysis of pure iron pentacarbonyl or its solutions yields diiron nonacarbonyl, which precipitates as golden hexagonal platelets [46], [47]:

$$2\,Fe(CO)_5 \xrightarrow{h\nu} Fe_2(CO)_9 + CO$$

The trinuclear iron carbonyl $Fe_3(CO)_{12}$ is also known [48–53].

As a highly reactive and readily available compound, iron pentacarbonyl is used for the preparation of many complexes, but these are not very significant industrially [39].

The most important industrial reactions of iron pentacarbonyl are thermal decomposition in the absence of air, yielding iron powder and carbon monoxide, and combustion to iron(III) oxide, Fe_2O_3; (see Section 3.5).

3.2. Production

Although nickel carbonyl can be obtained from nickel and carbon monoxide at atmospheric pressure and moderate temperature, the production of iron pentacarbonyl requires a pressure of 5 – 30 MPa, a temperature of 150 – 200 °C, and the presence of reactive iron.

Values reported for the reaction equilibrium are not in good agreement with those calculated [54–57]. Table 1 lists experimental data for the equilibrium constant along with values calculated from the Nernst approximation [57].

Even at high temperature and pressure, massive iron reacts sluggishly with carbon monoxide, so iron sponge, with its greater surface area, is used as starting material [58]. Better yields are said to be obtained with iron quenched and granulated from the melt and containing 2 – 4 % sulfur [59], which has a catalytic action [60]. For the reaction kinetics, see [61].

The exothermic reaction

$$Fe + 5\,CO \longrightarrow Fe(CO)_5(l), \quad \Delta H = -226.92\,kJ/mol$$

is carried out in high-pressure batch equipment with a vertical reactor charged with iron. The gas, circulated by a pump, is preheated and admitted to the reactor. It leaves the reactor hot and loaded with iron pentacarbonyl, which is condensed in a heat exchanger and allowed to expand into the unpressurized purification system under low pressure. To maintain the pressure in the system, the carbon monoxide used is replaced with fresh gas. Most of this is obtained when the carbonyl is decomposed to iron powder. This closed operating cycle removes carbonyl from the reaction equilibrium [62].

Carbon monoxide pressure, temperature, and flow rate are controlled carefully throughout the batch to avoid spontaneous decomposition of carbon monoxide, which would result in a sharp temperature increase and the deposition of carbon black [63].

The liquid carbonyl contains some lubricating oil (from the pump), water, and iron dust. Depending on the starting material, volatile carbonyls of nickel, chromium, molybdenum, and tungsten may also be present. These are removed by distillation.

The batch process is rather expensive and requires inert gas purging at the beginning and end of each batch. Nonetheless, other approaches have not been implemented. In a wet process, carbon monoxide was reacted with solutions of iron(II) salts in aqueous ammonia at 11.5 MPa and 80 °C, giving iron pentacarbonyl in yields of 40 – 50 % [64], [65]. A solution of iron(II) chloride in methanol containing a sulfur compound and Mn powder is said to form iron pentacarbonyl on reaction with carbon monoxide [66]. Another process uses three fluidized beds in series, for iron ore (e.g., hematite) reduction, high-pressure carbonylation, and carbonyl decomposition; however, the process has not been adopted commercially [67], [68].

3.3. Quality Specifications and Analysis

Commercial pure iron pentacarbonyl is brought to high purity by double distillation. The levels

Table 1. Equilibrium data for the formation of iron pentacarbonyl

T, °C	p_{CO}, bar	$Fe(CO)_5$, vol% (at equilibrium)	$K = \frac{(p_{CO})^5}{p_{Fe(CO)_5}}$ Observed	Calculated
60	0.69	22.4	0.86	0.063
80	1.47	21.8	20.1	3.39
160	38.83	9.0	2.4×10^7	1.0×10^5
200	132.68	5.7	5.5×10^9	1.26×10^5

of contaminants, metals that form volatile carbonyls, and sulfur are at or below the threshold of detectability [36].

To test for trace contaminants, a sample is solidified by cooling, treated with excess bromine, and brought slowly to room temperature. After the reaction stops, the residue is dissolved in dilute nitric acid. Techniques suitable for analyzing the solution (e.g., for Cr, Ni, Mo, or Pb) are atomic absorption spectroscopy and emission spectroscopy with inductively coupled plasma. Iron pentacarbonyl can be analyzed in a similar fashion after being dissolved in concentrated nitric acid.

Iron pentacarbonyl in gases, including air, can be oxidized with a hydrogen peroxide–methanol mixture and analyzed as iron hydroxide [36]. Bromine water and concentrated nitric acid are also suitable oxidizing agents. The exposure of individual workers to iron pentacarbonyl is measured by a personal air sampler, consisting of a charcoal-filled tube and a small battery operated gas pump. The iron pentacarbonyl is catalytically oxidized on the activated charcoal to give iron oxide, which can be determined as iron after combustion of the charcoal. The average content of iron pentacarbonyl in the air is then calculated.

Indoor air can be continuously monitored by radiometry with ionization detectors. Iron pentacarbonyl is atomized to an aerosol in a measuring chamber at 150 °C; the aerosol diminishes the ionization current generated by a radioactive source. The change in current is a measure of the carbonyl concentration. The detection threshold is ca. 0.1 ppm [69].

Gas chromatography with an electron-capture detector has been used to determine iron pentacarbonyl and nickel tetracarbonyl in synthesis gas. The detection limit for nickel tetracarbonyl is 0.001 ppm, and that for iron pentacarbonyl somewhat lower [70].

3.4. Safety, Storage, and Transportation

Safety. Two potential hazards must be considered in relation to iron pentacarbonyl: toxicity and combustion. These can be controlled only if the product is handled in sealed equipment [36]. The fire hazard can be summarized as follows:

Flash point	below − 15 °C
Ignition point	<65 °C
Lower explosion limit	2.6–4.5 vol %

On substances with a high surface area, such as activated charcoal, ignition is possible even at room temperature.

Because the heat of combustion of iron pentacarbonyl is only one-fifth that of fuel oil, iron pentacarbonyl fires are easy to extinguish. The best extinguishing agent is water, which removes heat and forms an airtight cover. Water that has been used to extinguish iron pentacarbonyl fires must not be discharged into natural waters.

Iron pentacarbonyl burns completely to iron(III) oxide and carbon dioxide only when it is atomized with air. Open combustion (e.g., in a dish) does not go to completion, and carbon monoxide is formed.

The best protection against the occurrence of iron pentacarbonyl concentrations that might present a danger of poisoning or explosion is continuous air monitoring. Carbon monoxide alarms respond only at much higher, toxic concentrations.

Storage. At industrial plants, iron pentacarbonyl is stored and processed in the absence of air in metallic (usually steel) equipment. Suitable seals and fittings are recommended by the carbonyl producer [36].

In unpressurized tanks, the carbonyl is covered with a protective gas (CO, N_2, Ar, CO_2). Tanks are connected to a gas reservoir so that the pressure does not drop below atmospheric even when the tank is cooled. When the product is stored in pressurized vessels, the liquid is withdrawn through a submerged tube, not through valves in the tank bottom. Tanks and pumps are installed in liquid-tight collecting basins.

Cleaning of Equipment and Disposal. Process equipment can be cleaned before opening by flushing with steam and then condensing the iron pentacarbonyl and water.

Carbonyl residues are collected and reused whenever possible. Because of the density difference, the product can easily be separated from water.

Laboratory carbonyl residues can be collected in a bottle partly filled with water. Glassware is washed several times with a small amount of acetone, which is poured into water for phase separation.

Small amounts of iron pentacarbonyl can be disposed of by burning in a sheet-metal tub. The presence of carbon monoxide in the off-gas must be taken into account.

Transportation. Because of fire and toxicity hazards, stringent regulations apply to the transportation of iron pentacarbonyl. Containers must be airtight and stable.

In Germany, the same regulations apply as in the case of nickel carbonyl. Both substances are assigned to Class 6.1, No. 3 under RID and GGVE for rail shipment, and under ADR and GGVS for road shipment. These regulations allow pressurized containers up to 250 L rated at 1 MPa. They are subject to inspection as pressure vessels (e.g., by TÜV) before being placed in service and at five-year intervals thereafter. The contents are limited to 1 kg of iron pentacarbonyl per liter of capacity. Every pressurized container must bear a nameplate with the following information: name of substance, owner, empty weight, year of first inspection and year of most recent inspection, inspector's stamp, maximum filling weight, and rated pressure (GGVE, Section 604).

The largest producer of iron pentacarbonyl, BASF, sells it in 20- and 250-kg pressurized containers. Each container has two valves or a double valve, so that liquid can be withdrawn through a submerged tube.

Also used as nonreturnable packaging in Europe are 1-L pressurized aluminum bottles, with a rated pressure of 1 MPa and a filling weight of 1 kg.

For international shipment by sea, iron pentacarbonyl is classified as follows: UN No. 1994, Class 6.1, IMDG Code. Air shipment and mailing are forbidden.

Similar regulations now apply in the United States. Although iron pentacarbonyl is a Poison B, because of the risk of inhalation it must be packaged in gas containers under pressure as prescribed for a Poison A under DOT Hazardous Materials Regulations. The containers hold 300 lb of carbonyl.

3.5. Uses

Carbonyl Iron Powder. By far the largest use of iron pentacarbonyl is for the production of carbonyl iron powder.

The equilibrium reaction

$$Fe(CO)_5(g) \rightleftharpoons Fe + 5\,CO, \quad \Delta H = +194.7\,\text{kJ/mol}$$

goes from left to right at standard pressure above 200 °C.

The iron deposited on hot surfaces is brittle, hard, and not commercially usable. In hot liquids such as paraffin oil, finely dispersed iron is obtained [71]. The only process that has become industrially important is decomposition in cavity decomposers, which yields spherical iron particles of 1 – 10 µm diameter [72].

The decomposer is a pipe, externally heated by electricity or gas, into which iron pentacarbonyl vapor is admitted from the top. The carbonyl decomposes not on the wall, which is heated to > 300 °C, but on iron particles circulating in the gas stream. These particles grow to 3 – 8 µm and are then discharged along with the carbon monoxide generated. Carbon and oxygen produced by the decomposition of carbon monoxide are incorporated in the iron. Addition

of ammonia reduces the carbon and oxygen content but leads to incorporation of nitrogen in the iron particles.

The cavity decomposer, invented by BASF in 1924, has not changed fundamentally. A variety of practices have been adopted to control the product, above all its particle size. Fine particles are obtained by admitting oil vapor into the decomposer [73], by diluting the carbonyl vapor with recycled carbon monoxide gas [74], or by applying a temperature gradient from the top to the bottom of the decomposer [75]. Fine particles result if the rate of flow of iron pentacarbonyl is high; larger ones, if the rate is low [76]. Inlet velocity is also a factor [77].

Carbon, nitrogen, and oxygen can be largely eliminated from the iron powder by treatment with hydrogen at ca. 450 °C [78]. The spheres, which were previously very hard (Vickers hardness 800 – 900), become soft (100 – 150 Vickers) and can be deformed by pressure. Hard and soft types are marketed for diverse applications [79].

Carbonyl iron powder is made into magnetic cores for electronic components [80], [81]. Carbonyl iron powder is an important starting material for powder metallurgy (see → Powder Metallurgy and Sintered Materials and → Sintered Steel and Iron: pure iron and iron alloys are made into parts by pressing [82], [83] and by metal injection molding [84–87]. Carbonyl iron powder is also employed for the fortification of foods such as white bread and in the production of iron-containing pharmaceuticals. It is used as a reducing agent in organic chemistry, and is incorporated in rubber or plastic sheets that are used for microwave attenuation [88–90]. For preparation, use, particle-size distribution, and behavior on heat treatment, see [91–95].

Carbonyl Iron Oxide. Only BASF converts iron pentacarbonyl to iron oxides. The processes are harmless to the environment, because the only byproduct is carbon dioxide, and no pollution of the wastewater with salts occurs, unlike when iron oxides are precipitated from solutions of iron salts.

Red iron oxide, in a finely divided form similar to carbon black, is obtained by atomizing iron pentacarbonyl and burning it in an excess of air [96]. By varying temperature and residence time in the reactor, transparent or highly transparent red pigments can be produced [97–99]. The highly transparent pigments are nearly X-ray amorphous and have BET surface areas of 80 – 160 m^2/g. The less transparent grades have surface areas of 10 – 35 m^2/g and clearly show a hematite structure. The iron oxide is used to make high-quality ferrites and similar ceramic materials. Because the content of other metals and electrolytes is extremely low and little energy is needed for dispersion, the oxide is particularly suitable for use as a lightfast, UV-blocking red pigment for paints, wood varnishes, fiber dyeing, and color printing.

Water injection [100] and passage through an intense centrifugal force field [101] during production improve both separation and compactibility.

Air oxidation of iron pentacarbonyl to magnetite is possible at higher temperature [102]. At 100 – 400 °C, magnetic black Fe_3O_4 or brown γ-Fe_2O_3 can be obtained, as desired. The carbon monoxide formed undergoes virtually no oxidation if the reaction takes place in a mechanically agitated fluidized bed of product oxide [103]. Finely dipersed γ-Fe_2O_3 pigments and Fe_3O_4 pigments (particle size 2 – 200 nm) are produced by BASF for use in aqueous suspension as contrasting agents in nuclear magnetic resonance tomography [104].

To obtain mixed oxides with iron as the primary component, iron pentacarbonyl is atomized together with a compound of the other metal (in liquid or dissolved form) that is converted to the oxide at high temperature. For example, by atomizing the carbonyl with an aqueous solution of chromic acid, iron chromium oxide can be obtained for use as a catalyst or as a brown pigment [105], [106].

Other Uses. Polycrystalline iron whiskers can be made by decomposing iron pentacarbonyl in a magnetic field [107–117]. The whiskers are remarkably strong and said to be suitable for making composites or for catalysts. Another process for producing iron whiskers decomposes iron pentacarbonyl in an empty space decomposer above 360 °C [118].

Novel effect pigments ranging from bright yellow to brilliant red are prepared in a fluidized bed of aluminum powder, which is coated with iron oxide when iron pentacarbonyl vapor and air are admitted [119], [120]. Novel

brilliant pigments are produced by coating aluminum flakes with nanometer films of hematite deposited from iron pentacarbonyl and, optionally, a second layer of colorless metal oxide [121]. Other pigments are made on a mica base [122]. These pigments are lightfast and corrosion resistant and are therefore suitable for use in automobile paints.

Carriers for the toner in photocopiers are obtained by coating fine spherical particles of iron or glass [123], respirable plastics [124], or porous silica-containing material [125]. Colored single-component toner powders are made of dispersions of carbonyl iron powder, reflecting pigments or dyes, and other substances in a binder [126], [127].

Suspensions of finely divided iron [128] or its alloys [129] are prepared by decomposing iron pentacarbonyl in solution. This process can be used to produce magnetic liquids [130].

If iron pentacarbonyl is decomposed on, for example, gold ore, the iron deposits preferentially on the metal, which could then be recovered by magnetic separation [131].

Iron pentacarbonyl is not an important industrial catalyst [132]; for its use as a catalyst in organic chemistry, see [133]. Iron pentacarbonyl has been described as a catalyst for the hydrogenation of coal [134], [135]. The catalysis of partial subsurface combustion of heavy crude oil by iron pentacarbonyl, with the aim of lowering the viscosity, has been reported [136], [137].

The complex $Na_2Fe(CO)_4$ formed in solution by the Hieber base reaction [138], recommended long ago for reducing vat dyes [139], has met with renewed interest. For example, chlorate can be removed from diaphragm-cell caustic soda with this product [140]. The use of $Na_2Fe(CO)_4$ to reduce organic compounds is described in [141]. A fuel cell operating with iron pentacarbonyl and an aqueous alkali metal hydroxide of pH < 9 has been proposed [142].

If iron carbonyl is added during thermal degradation of hydrocarbons to carbon fibers, subsequent high-temperature treatment gives a higher yield of better graphitized, more conductive fibers [143].

Pyrite films are obtained from iron pentacarbonyl and sulfur or hydrogen sulfide by chemical vapor deposition. The films are photoactive and can be used to make solar cells [144].

3.6. Economic Aspects

Three companies produce iron pentacarbonyl: BASF (Federal Republic of Germany) with a capacity of over 9000 t/a, GAF (Huntsville, Alabama) with an estimated 1500 – 2000 t/a, and a plant in the former Soviet Union.

These plants also produce carbonyl iron powder, with capacities of ca. 1500 t/a (BASF) and 500 t/a (GAF). The BASF plant has an iron oxide capacity of ca. 1000 t/a.

Consumption of the carbonyl for other purposes is insignificant compared to inplant consumption.

3.7. Toxicology and Occupational Health

Based upon the limited data that is available from laboratory animal studies, iron pentacarbonyl must be classified as highly toxic.

The first laboratory study demonstrated that a 45.5-min inhalative exposure to 0.025 vol % (250 ppm) of iron pentacarbonyl is fatal to rabbits [145]. Studies by BASF showed that 30-min exposure to 40 ppm of iron pentacarbonyl is already lethal to rabbits [146]. Cats reacted with considerably less sensitivity and survived exposure to 300 ppm. In guinea pigs, rats, and mice, fatalitities began after exposure to 140 ppm. After 30-min exposure, LC_{50} values of 2.190 mg/m^3 (275 ppm) for mice and 910 mg/m^3 (115 ppm) for rats were found in later acute toxicity studies [147]. The 4-h median lethal concentration was determined to be 10 ppm in rats [148].

The oral LD_{50} of iron pentacarbonyl is 0.012 mL/kg (0.018 mg/kg) in rabbits and 0.22 mL/kg (0.033 mg/kg) in guinea pigs [149]. After percutaneous application, the LD_{50} in rabbits is 0.24 mL/kg [149]. In BASF experiments, the following LD_{50} values were determined for acute oral toxicity: rabbit, 20 mg/kg; rat, 25 mg/kg; mouse, 100 mg/kg; cat, 100 mg/kg [146]. Iron pentacarbonyl was found to be nonmutagenic in the Ames test. No skin and eye irritation was observed in OECD 404 and OECD 405 tests.

Depending on carbonyl concentration, lethargy, respiratory symptoms, and lung edema

were found in rats in further inhalation experiments [150]. No data are available on the chronic toxicity of iron pentacarbonyl. Only one six-month feeding study with rats, in which 13 mg/d did not lead to any toxic effects, has been reported [151]. The clinical symptomatology of iron pentacarbonyl intoxication is similar to that of nickel carbonyl poisoning. It is marked by immediate onset of disorientation and vomiting. Fever, coughing, and difficulty in breathing occur after 12 – 36 h. Fatalities usually occur 4 – 11 d after exposure to a lethal dose. Pathologic lesions are found in the lungs and in the vascular and nervous systems [152]. Although iron pentacarbonyl and nickel carbonyl intoxications have similar symptoms, iron pentacarbonyl is probably less toxic (see Table 2) [153]. Furthermore, the use of iron pentacarbonyl should result in less exposure than nickel carbonyl because of its lower vapor pressure.

Table 2. Acute rat inhalation toxicity

Substance	LC_{50}, mg/m^3	Exposure time, min
$Ni(CO)_4$	240	30
$Fe(CO)_5$	910	30

Safety precautions are, however, required when working with iron pentacarbonyl to prevent oral, inhalatory, and dermal exposure, because of the relatively high acute toxicity. The MAK value for iron pentacarbonyl is 0.1 ppm. The TLV-TWA is 0.1 ppm (0.8 mg/m^3), as Fe.

4. Iron Compounds, Miscellaneous

Iron Acetates. When scrap iron reacts with acetic acid, it dissolves to form acetate salts [154]. If the initial black solution is concentrated to 12 %, solid *iron acetate* [*2140-52-5*] can be obtained. This salt is a mixture of iron(II) and iron(III) oxidation states of indefinite proportions best formulated as $Fe_x(CH_3COO)_y$. It is used as a catalyst for acetylation and carbonylation reactions.

Pure *iron(II) acetate* [*3094-87-9*], $Fe(C_2H_3O_2)_2$, is colorless. It is used in textile dyeing as an iron base for dark brown, dark blue, and black colors.

Iron(III) acetate [*1834-30-6*], $Fe(C_2H_3O_2)_3$, is prepared industrially by reacting scrap iron with acetic acid, followed by oxidation of the solution with air; it is very sensitive to light. Iron(III) acetate is used for dyeing and printing textiles, as a mordant in dyeing, as a catalyst in organic oxidation reactions, and in dyeing chamois leather.

Basic iron(III) acetate [*10450-55-2*], $Fe(OH)(C_2H_3O_2)_2$, is obtained by boiling a solution of iron(III) acetate and allowing the hydroxy salt to precipitate. This brownish-red compound is used as a mordant in dyeing, for weighting silk and felt, as a conservation additive for timber, and as a colorant for leather.

Iron(II) carbonate [*563-71-3*], $FeCO_3$, forms as a white precipitate when solutions of alkali-metal carbonate salts are added to solutions of iron(II) salts. It is used as a supplement in animal diets and as a flame retardant [154].

Iron Citrates. *Iron citrate* [*2338-05-8*] is a complex of indefinite formula, containing both iron(II) and iron(III). *Iron(II) citrate* [*23383-11-1*] and *iron(III) citrate* [*28633-45-6*] are also of indefinite stoichiometry. *Iron(III) citrate* [*3522-50-7*] (1 : 1 Fe – citric acid) is also known. Iron citrates are readily soluble in hot water and exhibit complex solution chemistry. These compounds have been used as supplements to animal diets and to soil.

Iron(III) ammonium citrate [*1185-57-5*] also has an indefinite formula. The brown hydrated form [*1332-98-5*] contains 16.5 – 18.5 % iron, ca. 9 % ammonia, and 65 % citric acid, whereas the green hydrated form [*1333-00-2*] contains 14.5 – 16 % iron, ca. 7.5 % ammonia, and 75 % citric acid [154]. Iron ammonium citrates are very soluble in water but insoluble in ethanol. They are used as iron additives in food for human consumption (e.g., bread and milk) and for the treatment of iron deficiency in small animals and cattle. They are sensitive to light and are used in light-sensitive paper [154].

Iron Halides. The iron(II) halides (fluoride, bromide, and iodide) are used as catalysts for fluorination, bromination, and iodination in organic reactions. *Iron(II) iodide* [*7783-86-0*] is used as a source of iron and iodine in veterinary medicine.

Iron(III) fluoride [*7783-50-8*] is used as a catalyst in organic reactions. *Iron(III) bromide* [*10031-26-2*] is used in the catalytic bromination of aromatic compounds.

Iron Fumarates. *Iron(II) fumarate* [*141-01-5*], $Fe(C_4H_2O_4)$, is prepared by adding a hot aqueous solution of sodium fumarate to a hot aqueous solution of iron(II) sulfate. It is only slightly soluble in water and can be separated by filtration as a red-brown powder. It is used in human and veterinary medicine in iron supplements. *Iron(III) fumarate* [*52118-11-3*], $Fe_2(C_4H_2O_4)_3$ has also been investigated as a food additive in veterinary studies.

Iron Nitrates. *Iron(II) nitrate hexahydrate* [*14013-86-6*], $Fe(NO_3)_2 \cdot 6\,H_2O$, *mp* 60.5 °C, forms green, rhombohedral crystals. It is prepared by dissolving iron in cold nitric acid ($d < 1.034$ g/cm^3). With increasing density, a greater proportion of iron is oxidized to iron(III). Iron(II) nitrate is used as a catalyst for reductions.

Iron(III) nitrate nonahydrate [*7782-61-8*], $Fe(NO_3)_3 \cdot 9\,H_2O$, forms colorless or pale violet monoclinic crystals. It is made by dissolving iron in nitric acid ($d > 1.115$ g/cm^3).

Iron(III) nitrate hexahydrate [*13476-08-9*], $Fe(NO_3)_3 \cdot 6\,H_2O$, forms colorless, cubic crystals (*mp* 35 °C). Iron(III) nitrate is used as a mordant, in tanning, and as a catalyst for oxidation reactions.

Iron(II) phthalocyanine [*132-16-1*] is synthesized from 1,2-dicyanobenzene and an iron(II) complex in refluxing 1-chloronaphthalene and is purified by sublimation at 450 °C. This green compound is insoluble in most solvents. Because it can exist in many oxidation states, iron(II) phthalocyanine is used as a catalyst for a variety of chemical and electrochemical redox reactions. It is also an important pigment [155].

5. Iron Compounds for the Treatment of Anemia [156], [157]

Sufficient iron in the diet of humans and animals is essential for tissue growth. Iron is found at the active site of many important proteins in the human body. These include hemoglobin (oxygen transport), myoglobin (storage of oxygen), cytochrome c oxidase (converts oxygen to water), cytochrome P450 (hydroxylation of poisonous or unwanted chemicals), and other cytochromes such as cytochrome c (part of the electrontransport system). In addition, iron is necessary for the biosynthesis of iron – sulfur proteins such as rubedoxin.

Although a sufficient amount of iron can usually be found in the diet, the level of absorption of this element from food is generally low. Therefore, the supply of iron can become critical in a variety of conditions. Iron-deficiency anemia is commonly encountered in pregnant women and may also be a problem in the newborn, especially in animal species such as the pig. Some diseases (e.g., rheumatoid arthritis, hemolytic diseases, and cancer) result in poor distribution of iron in the body and thus lead to chronic anemia.

More than 34 iron-containing preparations are used as hematinics for the treatment of iron-deficiency anemia [157]. Oral iron therapy is usually the preferred method of treatment unless good reasons exist for using another route. The daily oral dosage of elemental iron should be 100 – 200 mg. Iron(II) salts are more commonly used than iron(III) salts because they are assumed to be more soluble in the pH range 3 – 7. Solubility is essential if the compound is to permeate cell membranes. The rate of regeneration of hemoglobin appears to be independent of the salt used, provided sufficient iron is given. Factors affecting the choice of preparation are the incidence of side effects and the cost. If side effects occur, an alternative iron preparation or a reduced dosage can be tried. The main iron compounds used in iron supplements are iron(II) fumarate [*141-01-5*], iron(II) gluconate [*299-29-6*], iron(II) sulfate, iron(II) succinate [*10030-90-7*], iron(III) chloride solution, and iron(III) ammonium citrate [*1185-57-5*]. In addition, iron dextran [*9004-66-4*] and iron sorbital [*36134-87-9*] are used in injectable hematinics. Iron(II) sulfate is often preferred because it is the cheapest form and is at least as effective as any other. A number of oral preparations contain ascorbic acid (vitamin C) to stabilize the iron(II) state. However, the therapeutic advantage is minimal and the cost may be increased.

Folic acid is used in conjunction with an iron salt for the prevention of anemia in pregnant women.

Iron(II) complexes have a major disadvantage, however, in that they are sensitive to oxidation, especially in aqueous environments. This can be hindered initially by protectively coating iron(II) compounds in tablet form. However, when they dissolve in the gut, oxidation still occurs to give insoluble iron(III) salts (containing hydroxide). These frequently cause irritation and gastrointestinal distress. Such side effects are often severe, because large doses are neccesary to ensure that enough iron is absorbed. Nausea and epigastric pain may occur and are said to be dose-related, although the relationship between altered bowel behavior (constipation) and dosage is unclear [157].

Soluble iron(III) complexes are thus preferred (because oxidation is not a problem). Unfortunately, such complexes are likely to be charged (which is unsatisfactory for passive membrane diffusion). Even if these complexes were neutral, they may still be toxic.

The ideal properties of an iron(III) chelation complex for treating iron-deficiency anemia are: (1) solubility in the pH range 6 – 9, (2) a ligand with a high affinity for iron(III), (3) a neutral iron(III) complex, and (4) nontoxicity. Furthermore, after absorption into the body the complex must enter into an equilibrium with transferrin (the iron-transport protein) so that iron is available for utilization in metabolic pathways. This means that the chelator must be metabolized rapidly to ensure that iron is freshly available.

Iron(III) maltol [*33725-54-1*], [FeL_3] (maltol(HL) = 3-hydroxy-2-methyl-4*H*-pyran-4-one) and similar complexes [157][158][159][160][161] have been shown to possess the desirable properties outlined above for the treatment of iron-deficiency-related diseases.

Another iron-related disease in humans is severe tissue iron overload. This may occur during the treatment of anemia, mainly as a result of a large number of blood transfusions. This condition is treated by injections of desferrioxamine mesylate, which is a good chelating agent for iron(III), and is also used to treat cases of accidental iron poisoning.

6. Regeneration of Iron-Containing Pickling Baths

In the Federal Republic of Germany in 1987, the output of rolled steel products from base, high-grade, and stainless steels averaged 2.0×10^6 t per month. Approximately 60 % of this is pickled at least once.

Pickling removes the oxide layer that adheres to the steel surface. The oxide layer arises as scale in heat-treatment processes or as rust through the corrosive action of water, with its load of dissolved susbstances, or through humid atmospheric corrosion.

Whereas rust consists largely of Fe_2O_3, scale is made up of the three iron oxides (wustite FeO, magnetite Fe_3O_4, and hematite Fe_2O_3) in a ratio that depends on steel composition, annealing conditions (temperature, heating time, furnace atmosphere), final temperature of rolling, and rate of cooling after rolling.

The "picklability" of a steel and the acid consumption for pickling also depend on many factors: adherence of scale, composition of the steel, mode and magnitude of mechanical working, type and composition of pickling solution, and pickling conditions. The selection of pickling acid is dictated by the required surface quality and by economic factors. The most important pickling acids for iron and steel are sulfuric and hydrochloric. Phosphoric, nitric, and hydrofluoric acids are used for special purposes and stainless steels.

Iron(III) ions that dissolve during pickling are reduced by metallic iron, so that the pickling acid contains primarily iron (II) salts. The quantity of iron dissolved during pickling is called chemical pickling loss; it accounts for 0.2 – 1.2 wt % of the iron or steel being pickled and accounts for approximately three-quarters of the weight loss during pickling.

If iron salts are removed from the pickling liquor, the unconsumed pickling solution can be returned to the pickling unit after the addition of an amount of acid equivalent to the quantity of salts precipitated.

The acid consumed in pickling can also be completely regenerated from the iron salts. The pickling liquor recycled between the pickling tanks and the regeneration unit is thus not consumed, apart from evaporation losses.

6.1. Sulfuric Acid Pickling Solutions

When steel is pickled with sulfuric acid, the scale oxides dissolve to give iron sulfates. The rate of dissolution of the three oxides in sulfuric acid increases in the order Fe_2O_3, Fe_3O_4, FeO. The reaction of sulfuric acid with metallic iron to give iron(II) sulfate and hydrogen is inhibited by pickling additives (inhibitors).

Two processes are employed to regenerate sulfuric acid pickling liquor:

1) Crystallization of iron(II) sulfate heptahydrate
2) Complete regeneration by electrolysis

6.1.1. Crystallization

In industrial plants, spent pickling solution is usually cooled by using brine. The cooling effect depends strongly on the cleanness of the cooling surface, on which salt crust can easily form. These systems, therefore, require very careful upkeep. Frequent shutdowns must be expected.

Vacuum Crystallization (Fig. 1). A vacuum equal to the vapor pressure of water at the desired final temperature is maintained in the crystallization vessel.

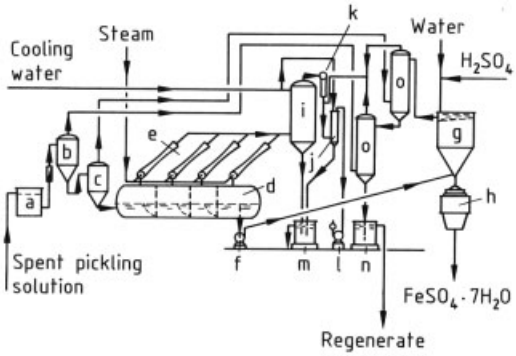

Figure 1. Vacuum crystallization (Keramchemie/Lurgi)
a) Suction tank; b) Precooler; c) Precooler; d) Crystallizer; e) Steam-jet compressor; f) Salt-slurry pump; g) Thickener; h) Centrifuge; i) Direct-contact condenser; j) Auxiliary condenser; k) Steam-jet venting device; l) Vacuum pump; m) Water receiver; n) Pickling solution receiver; o) Acid direct-contact condensers

Part of the pickling solution is pumped continuously from the pickling tank to the vacuum crystallizer, where it is cooled evaporatively to 5 °C and then to 0 °C to crystallize iron(II) sulfate. Water vapor flows directly, or after compression in a high-performance steam-jet apparatus, to a direct-contact condenser. The salt–liquid mixture is removed continuously by a salt-slurry pump and separated in a centrifuge. After the mother liquor has been freshened with concentrated sulfuric acid and water, it is returned to the pickling tank. The vacuum in the evaporator is maintained by a vacuum pump connected to the direct-contact condenser.

A modified design omits the steam-jet apparatus; therefore, only electrical energy is needed for its operation, which may be an economic advantage because in this case no fuel is required for steam generation.

Cyclone Crystallization (Fig. 2). In the cyclone crystallizer (b), fine droplets of spent pickling solution are sprayed countercurrent to a stream of air. In this way, rapid evaporation and cooling are accomplished. If the iron content of the pickling solution is to be kept low, a second cooling stage is added. Indirect cooling with water or brine takes place in the cooling vessel (c). A blower injects air to prevent the crystals from settling and forming a crust on the vessel wall. Iron(II) sulfate crystals are separated in the centrifuge (d); the filtrate (regenerate) is returned to the pickling plant.

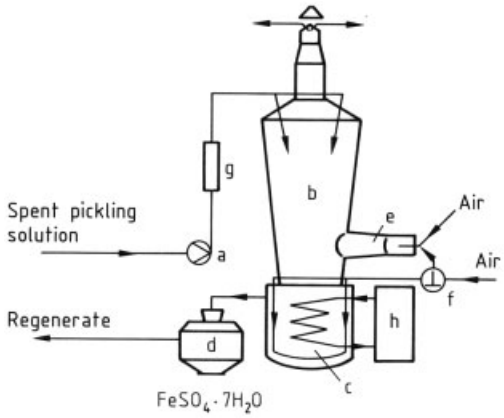

Figure 2. Cyclone crystallizer (Andritz-Ruthner)
a) Acid pump; b) Cyclone crystallizer; c) Cooling vessel; d) Centrifuge; e) Vent; f) Blower; g) Pickling liquor tank; h) Cooling system

The iron(II) heptahydrate can be used in the production of cyanoferrates, as a flocculant in

wastewater treatment, in the production of gas adsorbents, as a pesticide, as a chemical fertilizer, and in the production of iron-oxide pigments.

The heptahydrate can also be converted to the monohydrate. A continuous apparatus for the preparation of very pure monohydrate is built by Mannesmann Anlagenbau/Messo-Chemietechnik.

The heptahydrate is dehydrated to the monohydrate at a well-defined acid concentration at ca. 90 °C. After separation of the crystalline monohydrate in a centrifuge, iron sulfate-poor mother liquor is returned to the reactor. The water vapor resulting from the dehydration is removed from the cycle.

6.1.2. Electrolysis

An electrolytic process has been developed [162]. Cathode and anode compartments are separated by a diaphragm (Fig. 3). Metallic iron is deposited at the cathode; sulfuric acid and oxygen are produced at the anode. Electrolysis is facilitated by addition of an electrolyte, which is not consumed and does not affect the pickling process, but allows iron to be recovered most effectively from the acidic solution. Sulfuric acid either flows directly from the anode compartment back to the pickling tank or is detained temporarily in a holding tank.

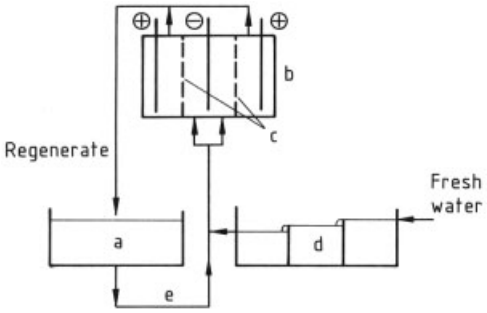

Figure 3. Regeneration by electrolysis
a) Pickling tank; b) Electrolysis cell; c) Diaphragms; d) Cascade rinsing; e) Spent pickling solution

Evaporation and electrochemical decomposition result in a continual loss of water. Using multiple countercurrent cascade rinsing after pickling allows the quantity of water to be adjusted so that no wastewater is produced. The unit is marketed by Keramchemie.

6.2. Hydrochloric Acid Pickling Solutions

Because of its technical advantages, hydrochloric acid has largely replaced sulfuric acid in pickling baths, especially in large plants.

Hydrogen is produced when iron is dissolved. Iron(III) chloride, formed by dissolution of Fe_3O_4 and Fe_2O_3, is reduced by the hydrogen to yield $FeCl_2$, so that virtually the only chloride present in the pickling liquor is iron(II) chloride. Spent pickling solution can be regenerated completely by thermal decomposition. Iron chloride is converted to iron oxide and hydrochloric acid:

$$4\,FeCl_2 + 4\,H_2O + O_2 \longrightarrow 2\,Fe_2O_3 + 8\,HCl$$

Gas or oil is used as fuel. The heat consumption is ca. 3 MJ for the simultaneous regeneration of 1 L of pickling solution and 1 L of rinse liquor.

In large-scale operations, the two processes that have found widespread application are 1) the fluidized-bed process and 2) the spray-roasting process.

Each is employed in more than 100 industrial plants. The efficiency of the acid recovery is ca. 99 %.

Fluidized-Bed Process. In the Lurgi–Keramchemie fluidized-bed regeneration process [163] (Fig. 4), spent pickling solution is led via the settling tank and the venturi loop into the fluidized bed in the reactor. The bed consists of granulated iron oxide. Residual acid and water are evaporated at 850 °C, and iron chloride is converted to iron oxide and hydrogen chloride. Growth and new formation of iron oxide grains in the fluidized bed are controlled so that a dust-free granulated product is obtained, with a particle diameter of 1 – 2 mm and a bulk density of ca. 3.5 g/cm^3. The granular product is discharged continuously at the bottom of the reactor and transported by a vibrating cooling chute and a vibrating spiral conveyor to the storage bin. In smaller units, the granular product can also be discharged directly into shipping containers.

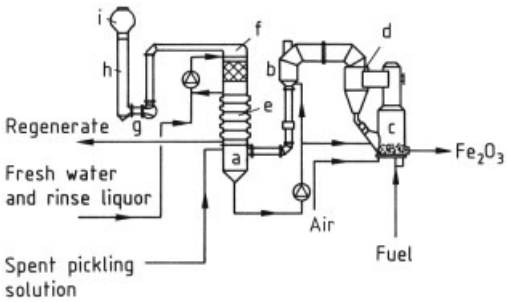

Figure 4. Fluidized-bed process for regeneration of hydrochloric acid pickling solutions
a) Separating tank; b) Venturi scrubber; c) Reactor; d) Cyclone; e) Absorber; f) Scrubbing stage; g) Off-gas blower; h) Stack; i) Mist collector

The hot gases from the reactor contain hydrogen chloride and a small amount of iron oxide dust, which is collected in the cyclone and returned to the fluidized bed. In the venturi scrubber, the off-gas is then cooled to ca. 100 °C. The thermal energy of the off-gases is used to concentrate the pickling solution by evaporation before it is fed to the reactor. Fine dust particles in the gas stream are removed by scrubbing.

From the venturi scrubber, the cooled gas stream goes to the absorber, where hydrogen chloride is absorbed adiabatically with rinse liquor and fresh water. The hydrochloric acid thus produced has a concentration of ca. 18 %. It is recycled to the pickling unit or held in a storage tank. After passing through a scrubbing stage and a mist collector, the off-gas is virtually free of hydrochloric acid and is released to the atmosphere.

Spray-Roast Process. The spray-roast process is often employed for the recovery of metal oxides from metal chloride solutions. It can also be used for the regeneration of iron chloride-laden hydrochloric acid solution [164]. The pure spray-roast oxide obtained in this way is used in the ferrite industry (150 000 t in 1987).

As shown in Figure 5, spent pickling solution goes to the venturi scrubber where hot gases concentrate it by evaporation of water. In the next step, the highly concentrated pickling liquor is injected into the spray-roast reactor at a pressure of 0.3 – 0.5 MPa as a fine spray either cocurrent or countercurrent to hot combustion gases. The cocurrent arrangement is common for pickling acids that contain zinc and lead, because their chlorides have high vapor pressures and are separated after a shorter retention time.

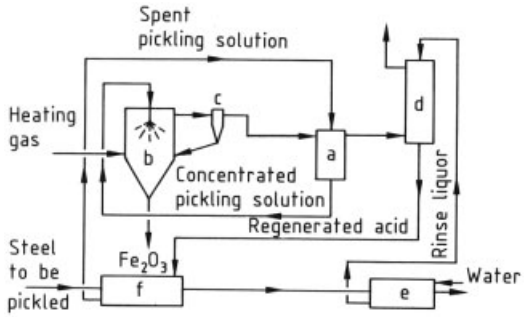

Figure 5. Spray-roast process for regeneration of hydrochloric acid pickling solutions
a) Venturi scrubber; b) Spray-roast reactor; c) Cyclone; d) Absorber; e) Rinse tanks; f) Pickling tanks

The temperature of the roasting gas outlet from the spray tower is about 400 °C. The gases proceed to the cyclone. Fine iron oxide powder, together with the oxide collected in the cyclone, is discharged continuously (bulk density 0.3 – 0.4 g/cm^3).

When some silicon-rich steels are pickled, the silicon dioxide content of the pickling solution does not permit the iron(III) oxide product to be used in the ferrite industry. The quantity of silica can be reduced in an additional precleaning step [165], in which silica is selectively removed from the pickling solution.

After passisng through the cyclone, the roasting gases are cooled and scrubbed in the venturi scrubber before going to the absorber, where hydrogen chloride is absorbed to produce 18 – 20 % hydrochloric acid. The acidic iron chloride solution, which is generated when the pickled steel is rinsed, is used as the absorption liquor. Hydrochloric acid is recycled to the pickling tank. Residual gases, with a low hydrochloric acid content, are discharged to the atmosphere.

Spray-roast plants have unit capacities of 0.3 – 25 m^3/h of pickling acid and a similar amount of rinse liquor.

The Fe_2O_3 generated in the thermal decomposition process can be used as raw material in various industries. The most important options are the production of magnetic materials (e.g., soft and hard ferrites), the production of iron powder for the fabrication of sintered parts and welding electrodes, and use as an additive in the

manufacture of magnetic tapes, abrasives, tiles, glass, cosmetics, and pigments.

6.3. Nitric and Hydrofluoric Acid Pickling Solutions

Stainless and acid-resistant steels are usually pickled with a mixture of nitric and hydrofluoric acids. Nitric acid serves as an oxidizing agent, whereas hydrofluoric acid forms complexes with metal ions. In contrast to the pickling of carbon steel, the permissible iron concentration in a pickling solution for special steels is limited by the tendency of iron and the alloy components to form fluoride complexes and thus reduce the concentration of fluoride ions in solution. Severe inhibition of the pickling process occurs at a total metal concentration of 50 g/L, so that pickling solutions must be topped up with fresh hydrofluoric acid, and eventually neutralized and discarded, even though only about half the acid content has been consumed.

When the pickling bath is neutralized with calcium hydroxide, the resulting metal hydroxides and calcium fluoride form a sludge that can be disposed of without treatment, but nitrates remain in solution and contribute to eutrophication in receiving waters.

The following regeneration processes have been devised:

1) Liquid – liquid extraction
2) Fluoride crystallization
3) Bipolar membrane process
4) Retardation process

Liquid – Liquid Extraction. Liquid – liquid extraction [166] makes use of the fact that nitric and hydrofluoric acids form adducts with tributyl phosphate that are soluble in hydrocarbon solvents. Addition of sulfuric acid also renders extractable acid anions bound in the form of metal salts. Around 93 % of the nitric acid and 60 % of the hydrofluoric acid can be recovered. The flow sheet for the process is shown in Figure 6.

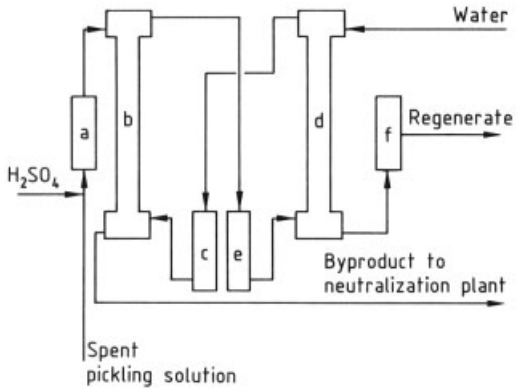

Figure 6. Regeneration of HF – HNO_3 pickling liquors by liquid – liquid extraction
a) Heat exchanger; b) Extraction column; c) Storage tank; d) Reextraction column; e) Storage tank; f) Adsorber

After sulfuric acid has been added to the spent pickling solution and the mixture has been cooled, it is directed to the top of an extraction tower; the organic phase enters at the bottom. The aqueous raffinate, containing sulfuric acid, is discharged at the bottom. From the top of the tower, the organic phase, laden with extracted acid, goes to the foot of a re-extraction tower. The water inlet at the top washes out the acids. The organic phase discharged at the top goes back to the extraction tower. In an adsorber, traces of tributyl phosphate are removed from the aqueous regenerate before it is returned to the pickling tank. The tributyl phosphate is returned to the tank (c).

Fluoride - Crystallization. In this process, the recovery of nitric acid is ca. 90 % (Fig. 7) and that of hydrofluoric acid, 55 %, despite the fact that metals precipitate as fluorides, so that corresponding amounts of hydrofluoric acid are consumed. The filter residue can be employed as a feedstock for the production of hydrofluoric acid or for the recovery of nickel.

Spent pickling solution is led through a particulate collector and a flow-through evaporator, where it is evaporated within minutes to about half its volume by electrical heating. The water vapor, which contains part of the nitric and hydrofluoric acids, is condensed by indirect cooling. The sparingly soluble fluorides of iron and of the alloy constituents precipitate from the concentrated, supersaturated solution in crystallization tanks and are removed periodically by

filtration. The filtrate is led to a storage vessel where it is combined with condensate from the condenser. After addition of fresh nitric and hydrofluoric acid, the liquor is returned to the pickling tank.

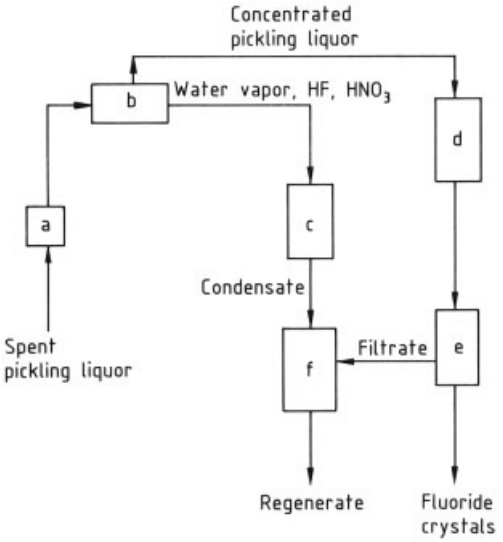

Figure 7. Regeneration of HNO_3–HF pickling solutions by fluoride crystallization
a) Particle collector; b) Flow-through evaporator; c) Condenser; d) Crystallization tanks; e) Filter; f) Storage tank

Bipolar Membrane Process. In the bipolar membrane process [167] (Fig. 8), the free acids are first recovered in an electrodialysis cell and then returned to the pickling tank. Metal ions are then precipitated as metal hydroxides by neutralization with potassium hydroxide. Metal hydroxide sludge is collected in the downstream filter press. The filtered salt solution (potassium fluoride and potassium nitrate) goes to a bipolar membrane module where HF, HNO_3, and KOH are separated from the potassium salts. The acid mixture is returned to the pickling tank, whereas potassium hydroxide is reused for neutralization. The dilute salt solution (KF, KNO_3) is desalted by reverse osmosis. Part of it is used as wash liquor for the filter press, and the rest is recycled to the membrane unit to obtain the highest possible yield of acids and potassium hydroxide. The residue from the filter can be reused as raw material in iron and steel works, depending on its content of metal hydroxides.

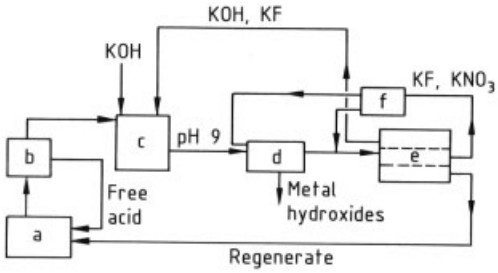

Figure 8. Regeneration of HNO_3–HF pickling solutions in a bipolar membrane cell
a) Pickling tank; b) Electrodialysis cell; c) Neutralization; d) Filter press; e) Bipolar membrane module (Aquatech cell); f) Reverse osmosis

Retardation Process. The retardation process can generally be employed for separating strong inorganic acids from their salts. By loading a strongly basic ion-exchange resin with spent pickling solution and then eluting with water, separation into a high-salt fraction and a subsequent high-acid fraction is achieved. The flow sheet for the process is shown in Figure 9.

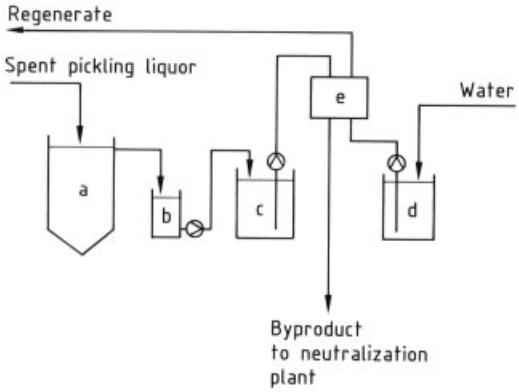

Figure 9. Regeneration of HNO_3–HF pickling solutions by acid retardation
a) Sedimentation tank; b) Buffer tank; c) Pickling-liquor metering tank; d) Water metering tank; e) Resin bed

Acid retardation can probably be accounted for by diffusion of undissociated acid into the ion-exchange resin grains so that upon elution with water, it must diffuse back out; the result is that elution of acid is retarded [168].

For nitric and hydrofluoric acid pickling solutions, most of the free acids can be recovered in this way. Partial regeneration of pickling solutions reduces disposal costs.

7. References

1. K. Mudrak, G. Stobbe: "Anwendung der Simultanfällung zur Phosphateliminierung," *Wasser Luft Betr.* **18** (1974) no. 5, 289 – 293.
2. K. I. Dahlquist, L. Hall, L. Bergmann: "Eliminierung von Phosphaten mit zweiwertigem Eisensulfat in der Kläranlage Käppala (S)," *Wasser, Luft Betr.* **20** (1976) no. 5, 107 – 112.
3. A. Heitmann, P. Reher, *Chem. Ing. Tech.* **46** (1974) 592 – 593.
4. Standard Messo, DE-OS 30 30 964 A1, 1980 (H. D. Kutta).
5. Enteco Impianti SpA, DE-OS 29 37 131, 1979 (L. Piccolo, A. Paolinelli, A. Rovere).
6. K. H. Linder et al.: *Materialien,* "Rückstände aus der Titan Dioxid Produktion," vol. 2/76, Umweltbundesamt, Berlin 1982, pp. 96 ff.
7. Aktieselskabet Aalborg Portland Cement Fabrik, EP 01 60 747, 1981 (P. L. Rasmussen).
8. J. Latorezai, I. Nagy, I. Kalo, J. Mihaleczku, HU 24 384, 1979.
9. M. Munemori, T. Aoki, Y. Inove, in R. B. Pojasek (ed.): *Toxic and Hazardous Waste Disposal,* "Simultaneous removal of hazardous metals from waste water and disposal of the resultant sludge," Ann Arbor Science Publishers, Ann Arbor 1980, pp. 97 – 105.
10. *Dangerous Prop. Ind. Mater. Rep.* **7** (1987) no. 1, 55 – 60.
11. W. S. Spector: *Handbook of Toxicology,* **vol. 1**, W. B. Saunders Co., Philadelphia London 1956.
12. Kronos Titan GmbH, DE 27 10 969, 1977 (A. Hartmann, A. Kulling, D. Schinkitz, E. Klein).
13. Kronos Titan GmbH, DE 30 30 558, 1980 (A. Hartmann, D. Schinkitz).
14. JANAF, *Thermochemical Tables,* NSRDS-NBS 37, 2nd ed., National Bureau of Standards, Washington DC, 1971.
15. R. C. Weast et al.: *Handbook of Chemistry and Physics,* 68th ed., CRC Press, Boca Raton, Florida, 1987 – 1988.
16. H. Schäfer, *Z. Anorg. Allg. Chem.* **266** (1951) 269 – 274.
17. *Gmelin,* System no. 59 "Eisen", part B, pp. 239 – 241.
18. R. K. Freier: *Daten für Anorg. und Org. Verb.,* **vols. 1 + 2**, "Wässrige Lösungen," Walter de Gruyter, Berlin – New York 1978.
19. W. Feitknecht, R. Giovanoli, W. Michaelis, M. Müller, *Helv. Chim. Acta* **56 Fasc. 8** (1973) no. 293, 2847 – 2856.
20. D. E. Chalkey, R. J. P. Williams, *J. Chem. Soc.* 1955, 1920 – 1926.
21. BASF, DE 1592 222, 1967 (K. Opp et al.).
22. BASF, DE 830 787, 1948 (H. Schlecht).
23. *Houben-Weyl,* "Halogenverbindungen," V/3.
24. *Synthetica Merck,* **vol. I,** Merck AG, Darmstadt 1969, pp. 182 – 189.
25. R. Klute, *Entsorgungspraxis* **11** (1987) 536 – 548.
26. *Chem. Mark. Rep.,* Oct. 28, 1985, p. 50.
27. *Jpn Chem. Week,* Sept. 9 1987, p. 2.
28. G. von Hagel, *Korresp. Abwasser* **33** (1969) no. 10, 908 – 915.
29. DOW, US 1 938 461, 1932 (C. F. Prutton).
30. M. Berthelot, *C. R. Hebd. Seances Acad. Sci.* **112** (1891) 1343.
31. L. Mond, L. Quincke, *J. Chem. Soc.* 1891, 604.
32. W. A. Herrmann, *Chem. Unserer Zeit* **22** (1988) no. 4, 113 – 122.
33. R. Boese, D. Blaeser, *Z. Kristallogr.* **193** (1990) 289 – 290.
34. D. Braga, F. Grepioni, A. G. Orpen, *Organometallics,* **12** (1993) 1481 – 1483.
35. Produkt-Merkblatt GAF 1977.
36. F. L. Ebenhöch: *Pentacarbonyl Iron,* Brochure, BASF Aktiengesellschaft, Ludwigshafen 1988.
37. *Gmelin, Eisenorganische Verbindungen,* **B 3,** pp. 1 – 247.
38. *Gmelin, Eisen, Teil B,* **no. 59,** pp. 486 – 498.
39. *Ullmann,* 4th ed., **10,** 417 – 418.
40. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* VCH Verlagsgesellschaft, Weinheim 1988.
41. W. Hieber, F. Leutert, *Ber. Dtsch. Chem. Ges.* **64** (1931) 2831.
42. W. Hieber, F. Leutert, *Naturwissenschaften* **19** (1931) 360.
43. W. Hieber, F. Leutert, *Z. Anorg. Allg. Chem.* **204** (1932) 145.
44. W. Hieber, *Z. Anorg. Allg. Chem.* **204** (1932) 165.
45. Ch. Elschenbroich, A. Salzer: *Organometallchemie,* Teubner, Stuttgart 1986, pp. 243 – 245.
46. J. Dewar, H. O. Jones, *Proc. R. Soc. (London)* **Ser. A 76** (1905) 558 – 577.
47. E. Speyer, H. Wolf, *Chem. Ber.* **60** (1927) 1424.
48. J. Dewar, H. O. Jones, *Proc. R. Soc. London Ser. A* **79** (1907) 66.
49. H. Freundlich, E. J. Cuy, *Chem. Ber.* **56** (1923) 2264.

50. P. Chini, *Inorg. Chim. Acta Rev.* 2 (1968) 31 – 51.
51. W. Hieber, G. Brendel, *Z. Anorg. Allg. Chem.* **289** (1957) 324.
52. BASF, DE 928 044, 1953.
53. BASF, DE 948 058, 1952.
54. A. Mittasch, *Angew. Chem.* **41** (1928) 827.
55. H. Pichler, H. Walenda, *Brennst. Chem.* **21** (1940) 134.
56. L. W. Ross, F. H. Haynie, R. F. Hochmann, *J. Chem. Eng. Data* **9** (1964) 339 – 340.
57. H. E. Charlton, J. H. Oxley, *AIChEJ* **11** (1965) no. 1, 79 – 84.
58. BASF, DE 499 296, 1924 (A. Mittasch, M. Müller-Cunradi, A. Pross).
59. BASF, DE 634 283, 1934 (L. Schlecht, H. Naumann).
60. W. Hieber, O. Geisenberger, *Z. Anorg. Allg. Chem.* **262** (1950) 15.
61. E. M. Vigdorchik, R. A. Shvartsman, P. P. Shukvostov: *Issled. V. Obl. Metall. Nikelya i Kobalta,* Leningrad 1983, pp. 125 – 132.
62. BASF, DE 485 886, 1925 (A. Mittasch, C. Müller, L. Schlecht).
63. BASF, DE 442 718, 1925 (M. Müller-Cunradi).
64. W. Reppe, *Justus Liebigs Ann. Chem.* **582** (1953) 116.
65. BASF, DE 753 618, 1940.
66. Montecatini, IT 728 074, 1966.
67. IT 887 928, 1971(C. Dufour-Berte, E. Pasero).
68. C. Dufour-Berte, E. Pasero, *Chim. Ind. (Milan)* **49** (1967) no. 4, 347 – 354.
69. Mine Safety Appliances Comp., DE 1 086 460, 1957 (R. A. Morris, R. Heine-Geldern).
70. W. Schäfer, *Fresenius Z. Anal. Chem.* **335** (1989) 785 – 790.
71. BASF, DE 493 874, 1926 (A. Mittasch, C. Müller, E. Linckh).
72. BASF, DE 500 692, 1924 (A. Mittasch, W. Schubardt, C. Müller).
73. GAF, US 2 612 440, 1950 (G. O. Altmann).
74. GAF, US 2 914 537, 1957 (D. J. Randall).
75. V. G. Syrkin, I. S. Tolmasky, SU, 1 186 398, 1964.
76. BASF, US 2 851 347, 1956 (L. Schlecht, E. östereicher, F. Bergmann).
77. BASF, DE 3 428 121, 1984; US 4 652 305, 1985 (F. L. Ebenhöch, R. Schlegel).
78. BASF, DE 528 463, 1927 (W. Meiser, W. Schubardt, O. Kramer).
79. BASF Broschüre, *"Carbonyleisenpulver",* Ludwigshafen 1994.
80. BASF, US 1 840 286, 1926 (E. Hochheim).
81. C. Heck: *Magnetische Werkstoffe und ihre technische Verwendung,* Heidelberg 1975, pp. 79 – 81, 397 – 415.
82. F. Duftschmid, L. Schlecht, W. Schubardt, *Stahl Eisen* **52** (1932) 845 – 849.
83. F. V. Lenel: *Powder Metallurgy,* MPIF Princeton, NJ 1980.
84. L. F. Pease, *Metal Powder Report* **43** (1988) no. 4, 242 – 254.
85. M. Blömacher, D. Weinand, *Metal Powder Report* **43** (1988) no. 5, 328 – 330.
86. M. Blömacher, D. Weinand, M. Schwarz, E. Langer: "Powder Injection Molding" in *Advances in Powder Metallurgy & Particulate Materials,* **vol. 5,** Metal Powder Industries Federation, Princeton, N. J. 1993.
87. E.-M. Langer, M. Schwarz, H. Wohlfrom, M. Blömacher, D. Weinand, *Prakt. Metallogr.* **33** 1996 5.
88. Whittaker Corp., US 4 173 018, 1979 (M. Dawson, L. Saffredial).
89. The Dow Corp., US 4 414 339, 1982 (J. Sole, R. R. Harris).
90. H. Dominik, E. Eckert, *Nachrtechn. Z.* **41** (1988) 280 – 283.
91. A. Mittasch, *Angew. Chem.* **41** (1928) 827 – 833.
92. G. Böhm, (BASF) World Steel & Metalworking (1982/83).
93. F. L. Ebenhöch, *Prog. Powder Metall.* **42** (1986) 133 – 140.
94. F. L. Ebenhöch, *Metal Powder Report* **42** (1987) no. 1, 12 – 14.
95. GAF, *Metal Powder Report* **43** (1988) no. 5, 338 –340.
96. BASF, DE 422 269, 1924 (A. Mittasch).
97. BASF, DE 2 210 279, 1972 (F. L. Ebenhöch, K. P. Hansen, H. Stark).
98. BASF, DE 2 344 196, 1973 (W. Ostertag et al.).
99. W. Ostertag, F. L. Ebenhöch, K. Bittler, G. Wunsch, *Defazet Aktuell* (1979) no. 12, 434 – 435.
100. BASF, DE 1 116 643, 1959 (H. Klippel).
101. BASF, DE 2 619 084, 1976 (W. Ostertag et al.).
102. BASF, DE 830 946, 1949 (L. Schlecht, G. Trageser).
103. BASF, DE 3 208 325, 1982 (M. Appl, F. L. Ebenhöch, R. Schlegel, E. Völkl).
104. BASF, WO-A 96/06891, 1994.
105. BASF, DE 1 953 518, 1969 (F. L. Ebenhöch et al.).
106. BASF, DP 2 517 713, 1974 (W. Ostertag, G. Wunsch, F. Ebenhöch, E. Völkl, G. Bock).
107. US 3 441 408, 1964 (H. J. Schladitz).
108. US 3 570 829, 1971 (H. J. Schladitz).
109. H. J. Schladitz, *Z. Metall.* **59** (1968) 18.

110. H. J. Schladitz, W. A. Jesser, D. S. Lashmore, *J. Appl. Phys.* **48** (1977) 478.
111. Klöckner-Werke AG, US 3 955 962, 1975 (W. Dawihl, W. Eicke).
112. Klöckner-Werke AG, US 4 002 464, 1975 (W. Dawihl, W. Eicke).
113. Klöckner-Werke AG, DE-OS 2 603 951, 1976 (H. Schön, R. Gustke).
114. Fa. W. M. Müller, DE 3 206 838, 1982 (F. Unterreithmeier).
115. W. Dawihl, W. Eicke, *Powder Metall. Int.* **3** (1971) 75.
116. L. E. Mürr, O. T. Inal, *J. Appl. Phys.* **42** (1971) 3887.
117. H. G. F. Wilsdorf, O. T. Inal, L. E. Mürr, *Z. Metallk.* **69** (1978) no. 11, 701 – 705.
118. BASF, US 5 085 690, 1992.
119. BASF, DE 3 003 352, 1981 (W. Ostertag, K. Bittler, G. Bock).
120. W. Ostertag, N. Mronga, P. Hauser, *Farbe Lack* **93** (1987) no. 12, 973.
121. BASF, EP 05 580 022, 1992.
122. BASF, US 4 344 987, 1981 (W. Ostertag, K. Bittler, G. Bock).
123. Xerox Corp., US 4 238 558, 1980 (R. F. Ziolo).
124. Xerox Corp., US 4 245 026, 1981 (R. F. Ziolo).
125. Xerox Corp., GB 1 577 257, 1980 (R. F. Ziolo).
126. BASF, US 4 803 143, 1989.
127. Océ , US 4 443 527, 1984.
128. Xerox Corp., US 4 252 671, 1981 (T. W. Smith).
129. Hitachi Maxell Ltd., JP-Kokai 58/137202, 1983.
130. J. M. Ginder: *Encyclopedia of Applied Physics,* **vol. 16,** VCH Publishers, New York 1996.
131. Hazen Research Inc., US 4 229 209, 1980 (J. K. Kindig, R. L. Turner).
132. *Ullmann,* 4th ed., **10,** 417 – 418.
133. H. Alper in T. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls,* **vol. 2,** Wiley-Interscience, New York 1977, pp. 545 – 593.
134. T. Suzuki, O. Yamada, Y. Takahashi, Y. Watanabe, *Fuel Process Technol.* **10** (1985) no. 1, 33 – 43.
135. Pentanyl Corp., US 4 451 351, 1984 (C. R. Porter, H. D. Kaesz).
136. D. Racz, I. Lorincz, B. Toth, A. Kassay, HU 15 890, 1975.
137. P. Kolocsai, E. Danoczy, K. Heberger, D. Racz, HU 35 768, 1983.
138. BASF, DE 441 179, 1925 (M. Müller-Cunradi).
139. BASF, DE 486 596, 1927 (A. Schneevoigt).
140. BASF, DE 3 707 713, 1988 (W. Kochanek, B. Leutner, D. Schläfer).
141. J. P. Collmann, *Acc. Chem. Res.* **8** (1975) 342 – 347.
142. BASF, DE 3 837 309, 1988.
143. Asahi Chem. Ind., JP 86/225 320 – 225 328, 1985 (K. Nahamura, Y. Komatsu).
144. H. Tributsch et al., *Mater. Res. Bull.* **21** (1986) 1481 – 1487.
145. H. W. Armit, *J. Hyg.* **8** (1908) 565.
146. Letter from BASF Corporation to USEPA submitting enclosed follow-up information concerning enclosed reports and studies on iron pentacarbonyl with attachments (1991) NTIS Order No.: NTIS/OTS0529732.
147. F. W. Sundermann et al., *Arch. Ind. Health* **19** (1959) 11.
148. Initial Submission: an Acute Inhalation Toxicity Study of Iron Pentacarbonyl in the Rat (Final Report) with Attachments and Cover Letter dated 022792 (1992), NTIS Order No.: NTIS/OTS0535889.
149. W. B. Deichmann, H. W. Gerarde: *Toxicology of Drugs and Chemicals,* Academic Press, New York –London 1969, p. 335.
150. J. C. Gage, *Br. J. Ind. Med.* **27** (1970) 1 – 18.
151. P. V. Sacks, D. N. Houchin, *Am. J. Clin. Nutr.* **31** (1978) 566 – 573.
152. R. S. Brief et al., *Am. Ind. Hyg. Assoc. J.* **28** (1967) 21.
153. H. E. Stokinger in: *Patty,* **vol. II A**, pp. 1797 – 1799.
154. *Kirk-Othmer,* **13,** 764 – 788.
155. *Ullmann,* 4th ed., **10,** 426 – 428.
156. A. B. Prasad (ed.): *British National Formulary,* British Medical Association and The Pharmaceutical Society of Great Britain, Norwich 1986, pp. 282 – 288.
157. M. T. Ahmet, C. S. Frampton, J. Silver, *J. Chem. Soc., Dalton Trans.* 1988, 1159.
158. R. C. Hider, G. Kontoghiorghes, J. Silver, M. A. Stockham, GB-A 2 117 766, 1983.
159. R. C. Hider, G. Kontoghiorghes, M. A. Stockham, EP-A 107 458, 1984.
160. R. C. Hider, H. Kontoghiorghes, J. Silver, M. A. Stockham, EP-A 138 421, 1985.
161. R. C. Hider, G. Kontoghiorghes, J. Silver, M. A. Stockham, EP-A 138 420, 1985.
162. Keramchemie GmbH, DE 3 206 538, 1986 (H. J. Heimhard, H. J. Simon).
163. H. J. Heimhard, G. Hitzemann, *Stahl Eisen* **105** (1985) 1222 – 1228.
164. G. Diez, P. Nieder, *Ind. Anz.* **87** (1965) no. 56, 1303 – 1311.
165. Andritz-Ruthner Division, AT 380 675, 1984 (H. Krivanec, B. Wasserbauer, D. Gausriegler).

166. H. Mühlberg, J. Mensler, P. Björklund, *Stahl Eisen* **95** (1975) 639 – 642.
167. Company brochure, Allied Signal Inc. USA, 1987.
168. R. Rituper, *Metall* **43** (1989) no. 9, 854 – 858.

Isobutanol → Butanols
Isobutene Polymers → Polyolefins
Isobutylene → Butenes

Isocyanates, Organic

CHRISTIAN SIX, Bayer AG, Dormagen, Germany

FRANK RICHTER, Bayer AG, Leverkusen, Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Production 9
4.1. Phosgenation of Free Amines 10
4.2. Other Phosgenation Procedures 11
4.3. Non-Phosgene Processes 13
5. Environmental Protection 14
6. Quality Specifications and Analysis . 15
7. Storage and Transportation 15
8. Uses 16
9. Economic Aspects 17
10. Toxicology and Occupational Health . 18
11. References 19

1. Introduction

Organic isocyanates are esters of isocyanic acid with the formula R–N=C=O. The first synthesis of an organic isocyanate was reported by A. WURTZ in 1848. Prominent chemists, including A. W. VON HOFMANN and T. CURTIUS, investigated the chemistry of organic isocyanates in the 1800s. The commercially important synthesis of isocyanates from amines and phosgene was discovered by W. HENTSCHEL in 1884 [1]. Diisocyanates became commercially important in the 1930s when the addition polymerization of difunctional isocyanates and polyols to produce polyurethanes was discovered by O. BAYER and coworkers at the I.G. Farben laboratories in Leverkusen [2]. More than 270 isocyanates were synthesized between 1934 and 1949 [3]. Today, 4,4′-diphenylmethane diisocyanate (MDI), toluene diisocyanate (TDI), hexamethylene diisocyanate (HDI), isophorone diisocyanate (IPDI), hydrogenated MDI [bis(4-isocyanatocyclohexyl)methane, H_{12}MDI], and derivatives thereof are of major practical importance.

2. Physical Properties

Organic isocyanates are colorless liquids or low-melting solids; they are normally purified by distillation. Excessive heating should be avoided because thermal degradation, sometimes accompanied by evolution of carbon dioxide and carbonization of the residue, and formation of oligomeric products may occur due to the high reactivity of the isocyanate group. The more volatile isocyanates are strong lacrimators with high inhalation toxicity; exposure should therefore be avoided. Polymeric MDI products are dark liquids; the viscosity depends on the diisocyanate content. Some physical properties of commercially available mono- and diisocyanates are listed in Tables 1, 2, and 3. The boiling and melting points of a variety of diisocyanates have been tabulated [3].

3. Chemical Properties

For general information, see [4–6].

Reaction with Nucleophiles. The reactivity of isocyanates is evident from the electronic resonance structures (Eq. 1).

$$\mathrm{R{-}N{=}C{=}O} \longleftrightarrow \mathrm{R{-}N^{-}{-}\overset{+}{C}{=}O} \longleftrightarrow \mathrm{R{-}N{=}\overset{+}{C}{-}O^{-}} \quad (1)$$

As a result, nucleophilic reagents rapidly attack the electrophilic carbon atom of the isocyanate group. Aromatic isocyanates react more rapidly than their aliphatic congeners, primarily for electronic reasons. The rate of reaction is further increased by electron-withdrawing substituents on the aromatic ring and decreased by electron-donating substituents. Steric hindrance also influences the rate of reaction. Aliphatic isocyanates react in the rate order primary > secondary > tertiary. Aromatic isocyanates with substituents adjacent to the isocyanate group also react more slowly because of steric hindrance. Thus, 2,4-toluene

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_611

Table 1. Selection of industrially important monoisocyanates

Compound (abbreviation)	Chemical structure	0CAS registry number	*mp*, °C	*bp*, °C	Major producers/suppliers
Methyl isocyanate (MI)	CH_3NCO	[*624-83-9*]	−17	38	Carbolabs, DuPont, Rhône-Poulenc
Isopropyl isocyanate (i-PrI)	$(CH_3)_2CHNCO$	[*1795-48-8*]	< −75	75	Bayer, SNPE/Tolochimie
n-Butyl isocyanate (n-BI)	$CH_3(CH_2)_3NCO$	[*111-36-4*]	−75	116	Bayer, SNPE
Hexadecyl isocyanate/ octadecyl isocyanate (stearyl isocyanate; e.g., 10/90 mix)	$CH_3(CH_2)_{15}NCO$/ $CH_3(CH_2)_{17}NCO$	[*1943-84-6*]/ [*112-96-9*]	23	352	Bayer, Mitsubishi, SNPE
Cyclohexyl isocyanate (c-HI)	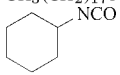	[*3173-53-3*]	−80	171	Bayer, Paushak, SNPE
Phenyl isocyanate (PI)	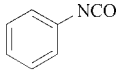	[*103-71-9*]	−32	165	Bayer, SNPE, United Phosphorus
1-Naphthyl isocyanate (“α-naphthyl”)	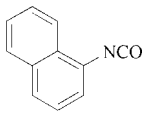	[*86-84-0*]	4	277	Bayer
o-Tolyl isocyanate (o-TI)	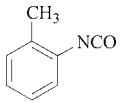	[*614-68-6*]	−18	188	Bayer
m-Tolyl isocyanate (m-TI)	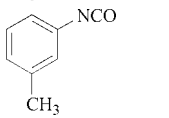	[*621-29-4*]	−75	189	Bayer, SNPE
p-Tolyl isocyanate (p-TI)	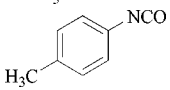	[*622-58-2*]	−25	189	SNPE/Tolochimie
3-Chlorophenyl isocyanate (m-CPI)	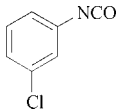	[*2909-38-8*]	−5	205	Bayer
4-Chlorophenyl isocyanate (p-CPI)	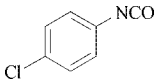	[*104-12-1*]	30	206	Bayer, SNPE

Table 1. (Continued)

3-Chloro-4-tolyl isocyanate (3-Cl-4-TI)	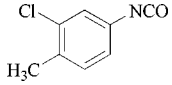	[*28479-22-3*]	23	225	Bayer, SNPE/Tolochimie
2,4-Dichlorophenyl isocyanate (2,4-DCPI)	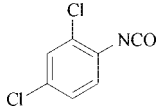	[*2612-57-9*]	58	233	SNPE/Tolochimie
3,4-Dichlorophenyl isocyanate (3,4-DCPI)	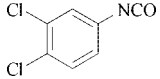	[*102-36-3*]	42	228	Bayer, Prochrom, SNPE/Tolochimie, United Phosphorus
3,5-Dichlorophenyl isocyanate (3,5-DCPI)	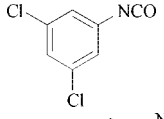	[*34893-92-0*]	29	243	Bayer, SNPE/Tolochimie
p-Isopropylphenyl isocyanate (p-IPPI)	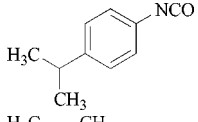	[*31027-31-3*]	< -60	220	Bayer, SNPE/Tolochimie
2,6-Diisopropylphenyl isocyanate (DIPPI)	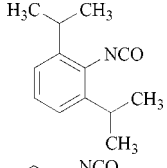	[*28178-42-9*]	−80	250	Aliachem, Bayer
α,α,α-Trifluoro-3-tolyl isocyanate (3-TFMPI)	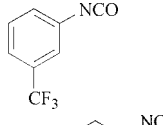	[*329-01-1*]	−30	174	Bayer, SNPE
p-(Trifluoromethoxy)phenyl isocyanate (4-TFMOPI)	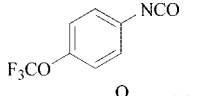	[*35037-73-1*]	−29	78 (0.3 kPa)	Bayer
p-Toluenesulfonyl isocyanate (4-TSI)	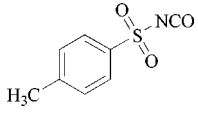	[*4083-64-1*]	5	270	Baxenden, Bayer, Nippon Soda, Vanchem

Table 2. Selection of industrially important aromatic diisocyanates.

Compound (abbreviation)	Structure	CAS registry number	*mp*, °C	*bp*, °C	Major producers/suppliers
Toluene diisocyanate (TDI)∗	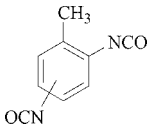	[*26471-62-5*], 2,4-TDI: [*584-84-9*], 2,6-TDI: [*91-08-7*]	13	246 – 247	BASF, Bayer-Shell Isocyanates, Cydsa, Dow, EniChem, Formosa, Huntsman, Korea Fine Chemicals, Lyondell, Mitsui, Narmada Chematur, Nippon PU, Oriental, Pronor, Rio Tercero, Bayer-Shell Isocyanates, Sodaso, Sumitomo, Takeda, Zachem
Methyl-enediphenyl diisocyanate (MDI)	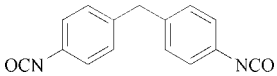	4,4′-MDI [*101-68-8*], 2,4′-MDI: [*5873-54-1*], 2,2′-MDI: [*2536-05-2*]	38 – 40	ca. 300	BASF, Bayer-Shell Isocyanates, Borsodchem, Dow, EniChem, Huntsman, Mitsui, Nippon PU, Bayer-Shell Isocyanates, Sumitomo, Takeda
Polymeric MDI (PMDI)	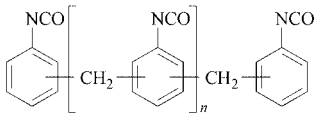	[*9016-87-9*]	∗∗	∗∗	
1,5-Naphthylene diisocyanate (NDI)	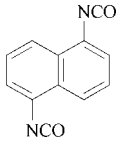	[*3173-72-6*]	127	155 (0.4 kPa)	Bayer, Mitsui
p-Phenylene diisocyanate (PPDI)	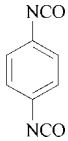	[*104-49-4*]	96 – 99	260	Akzo, DuPont
3,3′-Di-methylbiphenyl-4,4′-diyl diisocyanate (TODI)	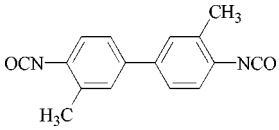	[*91-97-4*]	71	165 (0.1 kPa)	Nippon Soda

∗ TDI-80 isomer ratio: 2,4-TDI/2,6-TDI = 80/20; other commercially important TDI isomer mixtures are obtained by multistage crystallization: TDI-100, TDI-65. ∗∗ Properties depend on composition of polyisocyanate.

Table 3. Selection of industrially important aliphatic diisocyanates.

Compound (abbreviation)	Structure	CAS registry number	*mp*, °C	*bp*, °C	Major producers/suppliers
Hexamethylene diisocyanate (HDI)		[*822-\01106-6*]	−67	255	BASF, Bayer, Lyondell, Mitsui Toatsu, Nanyo Kasei, Rhodia, Takeda
3-Isocyanatomethyl-3,5,5-trimethylcyclohexyl isocyanate (IPDI)	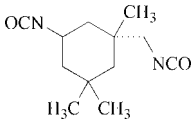	[*4098-71-9*]	−60	158 (1.3 kPa)	Bayer, Degussa, Rhodia
1,3-Bis(isocyanatomethyl)benzene (XDI)	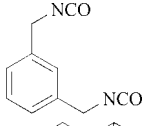	[*3634-83-1*]	−7	130 (0.3 kPa)	Takeda
Bis(isocyanatomethyl)-bicyclo[2.2.1]heptane, (NBDI)	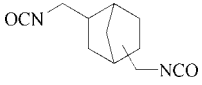	[*74091-64-8*], 2,5-NBDI: [*135540-95-3*], 2,6-NBDI: [*135540-96-4*]	< −30	159 (0.8 kPa)	Mitsui
2-Heptyl-3,4-bis(9-isocyanatononyl)-1-pentyl-cyclohexane (DDI)	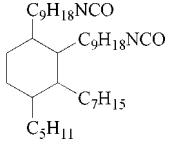	[*68239-06-5*]	-48	290	Bayer
4,4'-Methylenedicyclohexyl diisocyanate ($H_{12}MDI$)[a]	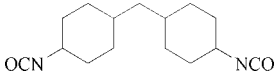	[*5124-30-1*]	ca. 25	179 (0.1 kPa)[b]	Bayer

Table 3. (Continued)

Compound (abbreviation)	Structure	CAS registry number	*mp*, °C	*bp*, °C	Major producers/suppliers
1,3-Bis(1-isocyanato-1-methylethyl)benzene (*m*-TMXDI)	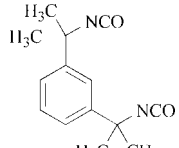	[*2778-42-9*]	-10	150 (0.4 kPa)	Cytec
1,6-Diisocyanato-2,2,4-trimethyl-hexane (TMDI)	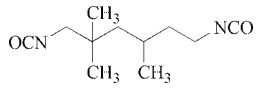	[*16938-22-0*] isomer mix		149	Degussa
1,6-Diisocyanato-2,4,4-trimethyl-hexane	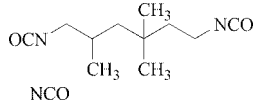	[*15646-96-5*]			
1,4-Diisocyanatocyclohexane (*trans*-CHDI)	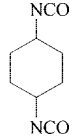	[*2556-36-7*]		123 (1.6 kPa)	DuPont
1,3-Bis(isocyanatomethyl)cyclohexane (H_6XDI)[c]	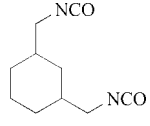	[*38661-72-2*]		98 (0.1 kPa)	Takeda

[a] Mixture of stereoisomers. [b] Decomposition when distilled under normal pressure. [c] Mixture of *cis* and *trans* stereoisomers.

diisocyanate reacts with nucleophiles preferentially at the 4-position. Steric hindrance in the attacking nucleophile similarly influences reaction rate: alcohols react in the order primary > secondary > tertiary. The reactivity of an attacking nucleophile is proportional to its nucleophilicity: aliphatic amines > aromatic amines > alcohols > thiophenols > phenols. The reactivity of the slower reacting phenols and thiophenols can be increased considerably by using the corresponding anions, but side-reactions may occur (e.g. trimerization; see page 8).

Commercially, the most important nucleophilic reaction of isocyanates gives addition or condensation polymers. Reaction with polyether or polyester polyols produces polyurethanes (→Polyurethanes); reaction with polyamines gives polyureas; reaction with dicarboxylic acids yields polyamides (→Polyamides); and reaction with carboxylic acid dianhydrides results in polyimides (→Polyimides).

The reaction of isocyanates with alcohols to form carbamates is catalyzed by tertiary amines and a variety of metal-containing catalysts, most prominently organotin compounds. The carbamates dissociate on heating (Eq. 2).

$$R^1N{=}C{=}O + R^2OH \rightleftarrows R^1NHCOOR^2 \quad (2)$$

The thermal stability of polyurethanes is therefore limited by the reversibility of this reaction. Use is made of this thermal instability in so-called blocked isocyanates, which contain the rather reactive NCO group in a deactivated form. Adducts of isocyanates with phenols, ketoximes, malonic esters, N heterocycles, etc. are examples of these blocked isocyanates. They are preferably used in coatings [7], [22].

The reaction of isocyanates with water gives a carbamic acid as an unstable intermediate, which undergoes decarboxylation to produce carbon dioxide and the corresponding amine (Eq. 3); the latter immediately reacts with more isocyanate to yield a symmetrically disubstituted urea (Eq. 4).

$$RN{=}C{=}O + H_2O \longrightarrow [RNHCOOH] \longrightarrow RNH_2 + CO_2 \quad (3)$$

$$RN{=}C{=}O + RNH_2 \longrightarrow RNHCONHR \quad (4)$$

Following the discussions about the role of chlorofluorocarbons (CFCs) in the depletion of the ozone layer in the early 1990s, this reaction with water is nowadays made use of in the production of flexible polyurethane foam from toluene diisocyanate and polyether polyols. The carbon dioxide generated then acts as a blowing agent [8], that is, as a CFC substitute.

Use can also be made of the NCO – water reaction in the manufacture of biuret-type aliphatic polyisocyanates. The final product, a biuret, is the addition product of isocyanate to urea (Eq. 5).

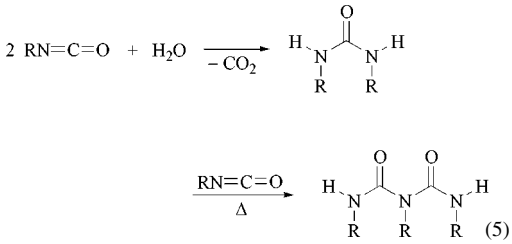

4-Toluenesulfonyl isocyanate undergoes very fast reactions with nucleophiles. This isocyanate is used as a water scavenger. Thus, 4-toluenesulfonamide is formed (Eq. 6), which does not undergo reaction with a second molecule of isocyanate.

$$MeC_6H_4SO_2N{=}C{=}O + H_2O \longrightarrow MeC_6H_4SO_2NH_2 + CO_2 \quad (6)$$

The reaction of isocyanates with amines (Eq. 7) to give disubstituted ureas is very fast and quantitative. For example, the reaction with excess di-*n*-butylamine is used for quantitative determination of isocyanate content (see Chap. 6) [9].

$$R^1N{=}C{=}O + R^2_2NH \longrightarrow R^1NHCONR^2_2 \quad (7)$$

The rate of reaction of isocyanates with amines increases with increasing basicity of the amine. Sterically hindered or electronically deactivated aromatic diamines are used as chain extenders in segmented polyurethane elastomers to achieve suitable reaction rates [7]. The only application of amine-type reaction partners for the isocyanate component in coatings applications, where the use of aromatic amines (and isocyanates) is in most cases less favourable due to discoloration of the final product, are aspartic acid esters (Eq. 8, adducts of diamines to maleic acid esters) [10].

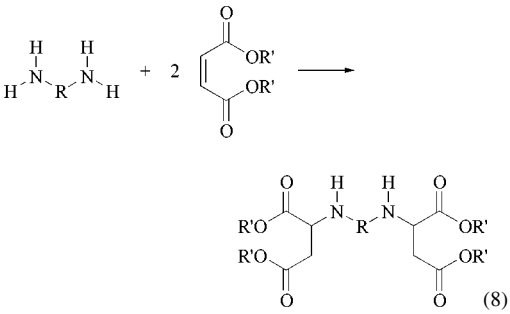

(8)

Carboxylic acids react with isocyanates to give mixed anhydride intermediates, which are decarboxylated to give amides at elevated temperature (Eq. 9).

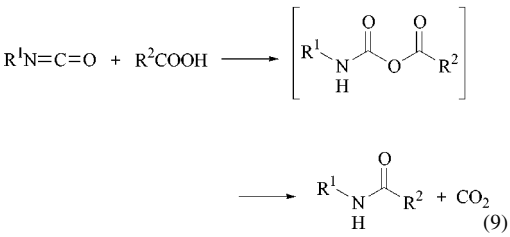

(9)

Base catalysis is required to avoid formation of disubstituted urea byproducts. The reaction of suitable carboxylic acid anhydrides, such as phthalic, trimellitic, pyromellitic, and benzophenonetetracarboxylic acid dianhydride, with isocyanates to give imides is catalyzed by nucleophiles [11].

Hydrogen halides form carbamic acid halides with isocyanates, which are thermally labile (Eq. 10).

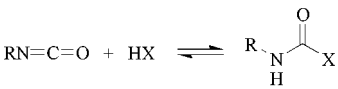

X = halogen (10)

Carbamic acid chlorides (X = Cl) are the thermally labile intermediates in the production of isocyanates from amines and phosgene (see Section 4.1).

Organic isocyanates undergo addition reactions with molecules containing active hydrogen atoms (e.g., Michael-type substrates, Eq. 11).

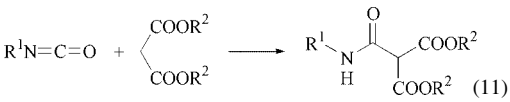

(11)

Insertion reactions occur with a variety of substrates containing metal – oxygen, metal – sulfur, or metal – nitrogen bonds. This reaction is observed with elements of Groups 1, 2, 12, and 13 [6].

Cycloaddition Reactions. Isocyanates undergo cycloaddition, in most cases across the carbon – nitrogen bond, with a variety of substrates containing double or triple bonds. Dimerization (Eq. 12) and trimerization (Eq. 13) are also examples of this type of reaction. Additions across the C=O bond are rare. An example is the formation of iminooxadiazinedione trimers (Eq. 14) [12], [13].

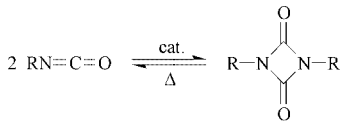

uretdione (12)

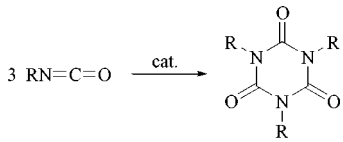

isocyanurate (13)

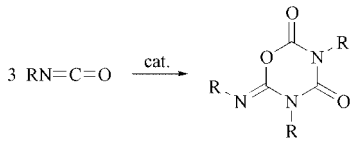

iminooxadiazinedione (14)

Slow dimerization occurs during storage [14] and may be accelerated at elevated temperature. Aromatic isocyanates dimerize and trimerize more readily than aliphatic isocyanates. Dimerization is catalyzed by pyridines and phosphines. Strong bases such as tertiary amines, hydroxides, or carboxylates catalyze trimerization to isocyanurates. Iminooxadiazinediones are side products in most known trimerization reactions. (Poly)fluorides, however, lead to a significant higher content of iminooxadiazinedione in the mixture of trimers [13]. The trimerization of PMDI is used in the manufacture of rigid polyurethane foams with higher thermal stability and better combustibility characteristics [9]. The cycloaddition of isocyanates with carbodiimides is used in the production of carbodiimide-modified liquid MDI products (see Chap. 8).

The conversion of isocyanates to carbodiimides (Eq. 15) is catalyzed by phospholene oxides and other species containing a P=O group [15].

$$2\,RN{=}C{=}O \longrightarrow RN{=}C{=}NR + CO_2 \qquad (15)$$

Isocyanates also undergo anionic homopolymerization. The reaction is preferred at low temperature and is to some extent reversible, provided the polymer has not aged. Catalysts used include sodium cyanide in DMF and aqueous solutions of alkali metal salts [16] but several other "dimerization" and "trimerization" catalysts described above also convert isocyanates at low temperature to 1-nylons (Eq. 16).

(16)

1-Nylon type isocyanate homopolymers have not found commercial use so far. Reviews on isocyanate derivatives are available [17], [18].

4. Production

For general information, see [3], [8], [19–21].

Isocyanates in general can be prepared in many ways. Besides some well-known laboratory-scale methods, among which are the Curtius, Hofmann, and Lossen rearrangements via nitrene intermediates, a number of special syntheses are available [6], [23], [73], [74]. Most of these methods, however, are not satisfactory for large-scale operation. The reaction of phosgene with amines or amine salts and the thermal cleavage of urethanes, preformed from the appropriate amine, urea, and an alcohol, are used commercially on a large scale [24], [25].

The production chain to polyurethanes using TDI and/or MDI is portrayed in Figure 1.

Toluene is the primary raw material for industrial TDI manufacture. The classical dinitration of toluene with mixed acid produces a mixture of the 2,4- and 2,6-dinitro isomers in a ratio of 80:20. Catalytic reduction of these derivatives under hydrogen pressure leads to the corresponding diamines TDA [26–31], which are subsequently treated with phosgene to give TDI [32], [33].

The production of PMDI involves condensation of aniline with formaldehyde in the presence of hydrochloric acid to give oligomeric di- and polyamines that are phosgenated without further need of purification (→Aniline, Chap. 6.1.). The percentage distribution of the homologues and isomers of MDA depends on the ratio of aniline to formaldehyde, the acid concentration, and the reaction conditions [34]. Monomeric MDI is obtained from PMDI by continuous thin-film distillation. The residual crude oligomeric isocyanate product (PMDI) is used mainly in the production of rigid polyurethane or polyisocyanurate foams that usually find use in insulation, construction, or automotive applications. Yields in the production of PMDI – MDI are quantitative. The production of aromatic amines used in the manufacture of TDI and PMDI is described in [22], [23].

The most common aliphatic diisocyanates are hexamethylene diisocyanate (HDI), isophorone diisocyanate (IPDI), and hydrogenated MDI (H_{12}MDI). They are industrially made from the corresponding amines (HDA, IPDA, and H_{12}MDA) by reaction with phosgene or by the carbamate route (reaction of the amine with urea and alcohol). Hydrogenated MDI is produced by catalytic hydrogenation of methylenedianiline (MDA) on a rhodium or ruthenium catalyst to yield a mixture of *trans*,*trans*-, *cis*,*trans*-, and *cis*,*cis*-H_{12}MDA, which is then phosgenated (Eq. 17).

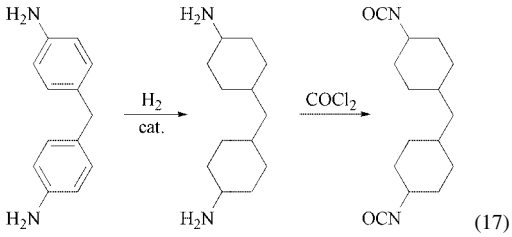

(17)

For conversion to pure *trans*,*trans*-H_{12}MDA, see [35], [36]. To obtain a liquid diisocyanate product in the subsequent phosgenation, the amount of *trans*,*trans*-H_{12}MDA, which is solid in the pure state, has to be reduced to ca. 20 % (selective hydrogenation and/or crystallization).

4.1. Phosgenation of Free Amines

The reaction of free amines with phosgene is by far the most efficient and economically attrac-

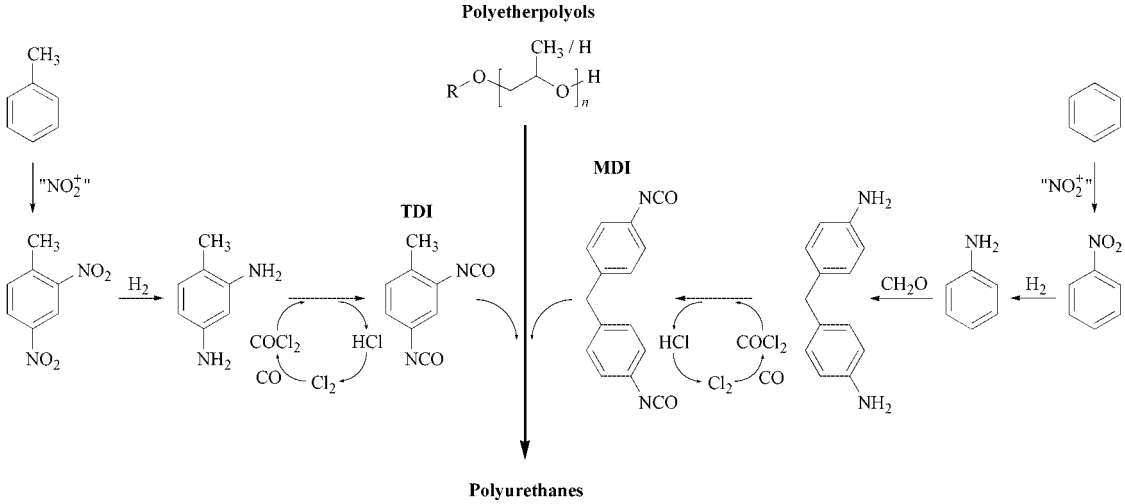

Figure 1. Industrial route to TDI and MDI (for the sake of clarity, only one isocyanate isomer is shown)

tive method for large-scale production of isocyanates. Depending on the boiling point and the basicity of the free amine, phosgenation can be carried out in the liquid or gas phase.

Solvents used in the phosgenation of free amines include hydrocarbons such as toluene, xylene, decalin (decahydronaphthalene), chlorobenzene, and *ortho*-dichlorobenzene. Chlorinated hydrocarbons are preferred because of their higher polarity, which aids in dissolution of the amine hydrochlorides. The overall reaction is shown in Equation (18).

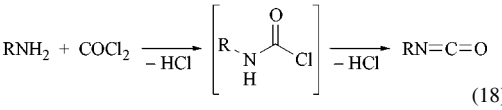

(18)

For the phosgenation of aromatic (poly)amines such as TDA or PMDA in a batch process, a 10 – 20 wt % solution of the amine is prepared in monochlorobenzene or *ortho*-dichlorobenzene and is thoroughly mixed with excess liquid phosgene or a solution of phosgene (100 – 500 % phosgene excess) in the same solvent at ambient temperature. The reaction mixture is heated and excess phosgene and HCl byproduct are purged by a nitrogen flush. Phosgene can be regarded not only as a reactant but also as a solvent. The amount of excess phosgene necessary to achieve optimum isocyanate yield depends on the number of amine groups and, in case of polyamines, the differences in reactivity between these groups: the more the reactivities of the different amino groups of a compound differ, the more residence time, dilution, and phosgene excess is usually needed. As soon as evolution of hydrogen chloride is complete and a homogeneous solution is obtained, the solvent is removed by distillation, and the pure isocyanate is obtained by fractional distillation, crystallization, or sublimation.

In continuous isocyanate production processes the same raw materials are mixed to produce a slurry of the corresponding carbamic acid chloride. This slurry is pumped into a reactor or a series of reactors where the reaction to give the free isocyanate is completed [32]. The temperature in the first reactor is usually in the range of the dissociation temperature of the carbamic acid chloride (< 80 – 100 °C), and the subsequent reaction is conducted above 150 °C to complete the evolution of hydrogen chloride and formation of isocyanate. Hydrogen chloride and excess phosgene are vented at the higher temperature to prevent recombination of hydrogen chloride with the isocyanate. Phosgene and hydrogen chloride are recovered [37], [38] (see Chap. 5). The waste gas, containing small amounts of phosgene, is hydrolyzed with water on charcoal, whereby traces of organic solvents are also adsorbed. The isocyanate raw product is purified by fractional vacuum distillation; the stripped solvents are reused. Usually the distillation residue is incinerated after polymerization

to give solid particles [39–41] or hydrolyzed to recover part of the amines [42–46].

In the case of PMDI the byproducts formed during phosgenation may remain in the product [47]. Though these byproducts influence the reactivity of the crude polyisocyanate, because hydrolyzable chlorine compounds (e.g., *N*-chloroformylchloroformamidines; see Eq. 22) deactivate the amine catalysts used in polyurethane production. The coproduct MDI is recovered by thin film evaporation. In this manner, a mixture consisting of ca. 99 % 4,4′-MDI and 1 % 2,4′-MDI is obtained. For some applications, production of MDI with higher amounts of the 2,4′ isomer is preferred. This is best accomplished by conducting the aniline – formaldehyde condensation at higher temperature, which favors formation of the 2,4′ and 2,2′ isomers.

For the production of aliphatic isocyanates, continuous free amine phosgenation processes in liquid phase are unsuitable because of the slower reaction rates observed for aliphatic amine hydrochlorides and side reactions resulting from the higher basicity of aliphatic amines. Two-phase phosgenation with an inert solvent and an aqueous phase containing a base (as the hydrogen chloride scavenger) is feasible for aliphatic amines. In general, aliphatic diisocyanates are produced either by phosgenation of a slurry of the corresponding amine hydrochlorides (see Section 4.2) or of the carbamate salts obtained by the reaction of the aliphatic diamines with carbon dioxide (see Section 4.2), through phosgenation in the gas phase [48–54], or by non-phosgene routes (see Section 4.3). Exceptions are longer chain aliphatic isocyanates, such as octadecyl isocyanate, 1,12-dodecane diisocyanate, and the diisocyanate obtained from dimer acid (→Dicarboxylic Acids, Aliphatic, Chap. 3.4.2.). Longer alkyl chains increase the solubility of the corresponding hydrochlorides in hydrocarbon solvents. Thus, the reaction rates necessary for continuous phosgenation of the free amine can be obtained.

As discussed earlier the phosgenation reaction is typically conducted in highly diluted solutions, since in most cases the solubility of the amine is low and the formation of urea can be suppressed by dilution. Disubstituted ureas are formed as byproducts by reaction of the isocyanate with unconverted amine. In the case of an aliphatic amine, the urea reacts further with phosgene to form allophanoyl chlorides (Eq. 19), which can be thermally converted to isocyanates (Eq. 20) [55].

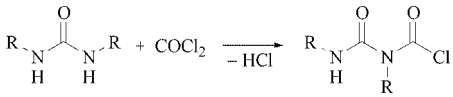

(19)

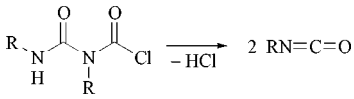

(20)

For aromatic amines, the urea byproducts can be dehydrated by phosgene to form carbodiimides (Eq. 21), which then react with excess phosgene to give *N*-chloroformylchloroformamidines (Eq. 22); cycloadducts of carbodiimides with isocyanates are also formed (Eq. 23).

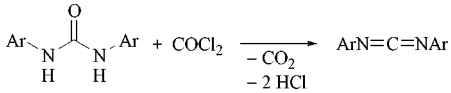

(21)

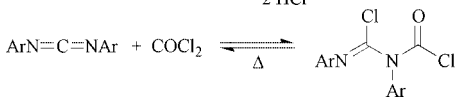

(22)

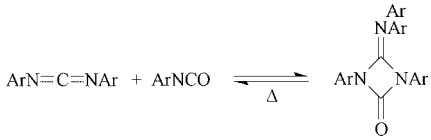

(23)

Cycloadducts with 4-membered rings can be cleaved by thermal treatment.

4.2. Other Phosgenation Procedures

Phosgenation of Amine Hydrochlorides. One method used to avoid side reactions is the phosgenation of amine hydrochlorides, rather than the free amine, exemplified by the conversion of isophorone diamine dihydrochloride to isophorone diisocyanate [56]. Isophorone diamine in decahydronaphthalene solution flows downward against a hydrogen chloride gas stream. The fine slurry of amine salts obtained is heated to 150 °C and treated with phosgene in a series of agitated autoclaves to give a clear solution of isophorone diisocyanate. Excess phosgene is removed by purging with nitrogen. Isophorone diisocyanate is isolated by stripping off the solvent and fractional vacuum distillation of the residue.

In the phosgenation of aliphatic diamine dihydrochlorides, the reaction has to be conducted below 150 °C to avoid the formation of ω-chloro monoisocyanates because of cleavage of the NCO group by hydrogen chloride [3].

Phosgenation of Carbamate Salts. Carbamate salts can be used instead of aliphatic amine hydrochlorides to achieve faster reaction [57–59]. For example, hexamethylenediamine and *ortho*-dichlorobenzene are introduced into a cooled reaction vessel and stirred at 5 – 10 °C under 0.5 atm CO_2 pressure to form a suspension of the carbamate salt. This suspension is subsequently transferred into a reactor containing a phosgene solution. Similarly, tetramethylene diisocyanate can be prepared. The recovery of phosgene and hydrogen chloride and workup of the reaction mixture are similar to other phosgenation processes.

Phosgenation of Ureas. Phosgenation of disubstituted ureas [60], [61] is used in the production of arenesulfonyl isocyanates. Direct phosgenation of sulfonamides proceeds in polar solvents only above 200 °C, whereby significant amounts of arenesulfonyl chlorides are formed as byproducts. In contrast, 1-arenesulfonyl-3-alkylureas undergo rapid reaction with phosgene at ca. 150 °C [62] . In the commercial production of arenesulfonyl isocyanates, reaction of the sulfonamides is conducted in the presence of a catalytic amount of *n*-butyl isocyanate (Eq. 24) to achieve rapid rates at 150 °C [63]. The rate-determining step is the intermediate formation of the urea. The *n*-butyl isocyanate catalyst is regenerated in the phosgenation of the urea (Eq. 25).

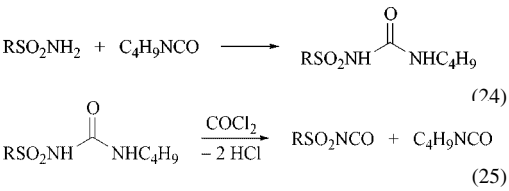

(24)

(25)

The reaction of disubstituted ureas (Eq. 26) with phosgene can also be used to synthesize aliphatic isocyanates which contain groups such as esters or ethers that are sensitive to prolonged heating with phosgene. The lower boiling aliphatic isocyanate R^2NCO is continuously removed from the reaction mixture (Eq. 27) [64].

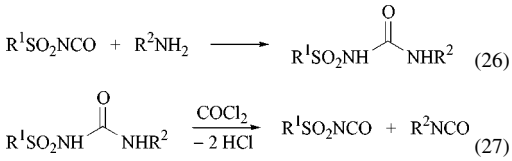

(26)

(27)

Arenesulfonyl isocyanates are used as intermediates for pharmaceuticals and herbicides, and as water scavengers in the formulation of coating systems.

4.3. Non-Phosgene Processes

On a laboratory scale, liquid trichloromethyl chloroformate (diphosgene) [65], solid bis(trichloromethyl) carbonate (triphosgene) [66], or solid di-*tert*-butyl tricarbonate ("pyrocarbonate") [67–69] can be used as substitutes for phosgene to synthesize isocyanates from amines. Because of their lack of industrial importance in commodity isocyanate production, the synthesis and utilization of diphosgene [70], triphosgene [71], [72], and pyrocarbonate [73], [74] are not discussed here. However, since the phosgene oligomers still have the potential to liberate phosgene, they should not be regarded as true phosgene substitutes.

In recent years several methods have been developed or investigated for the industrial production of isocyanates that do not require the use of the highly toxic phosgene. The most promising approaches are depicted in Figure 2 together with the classical phosgenation route (path 1):

- Reductive carbonylation of nitro compounds (paths 2 and 3) [75–78]
- Dehydrogenation/degradation of formamides (paths 4 and 5) [79], [80]
- Carboalkoxylation/carboaroxylation of amines (path 6) [81]
- Oxidative carbonylation of amines (path 7) [82], [83]
- Cleavage/thermolysis of carbamates (path 8)

The most important aliphatic diisocyanates, HDI and IPDI, could be manufactured without phosgene by the urea route (see Fig. 2, paths 6 and 8) [84–89].

Reductive carbonylation of nitro compounds, especially nitroaromatic compounds, was the subject of thorough industrial research from

1962 until the early 1990s due to the demand at that time for a new, phosgene-free method for the production of isocyanates and the discussions on the chlorine cycle in industry. The target was a low-cost one-step synthesis starting from the corresponding nitro precursors (Fig. 2, path 2). Arising problems favored an alternative two-step reaction via urethanes (see Fig. 2, paths 3 and 8), which seemed to represent a feasible technical method to significantly reduce the costs of isocyanate production. All announcements referring thereto have been shown to be invalid, simply because the abundant observations claimed in numerous patents and other publications led to an inadequate and optimistic evaluation, although the chemistry was very poorly understood at that time. After 1969 research on homogeneous reductive carbonylation of nitro compounds with compounds of ruthenium, rhodium, and palladium as catalysts increased in academic laboratories. The most recent period is characterized by in-depth studies in academia on the mechanism of activation and catalysis and the nature of the catalytic species, and a declining interest in industry, at least with regard to large-scale manufacture of diisocyanates typically used in the polyurethane industry [75], [76].

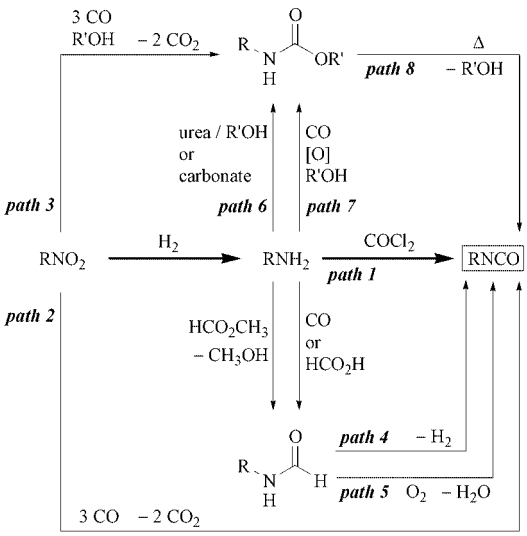

Figure 2. Non-phosgene routes to isocyanates in comparison to classic phosgenation route (path 1)

The route to PMDI starting with nitrobenzene has the additional complication that unreactive methyl phenylcarbamate has to undergo the condensation reaction with formaldehyde. Another process for PMDI manufacture is based on the reaction of aniline (or oligomeric amines derived from aniline) and formaldehyde with dimethyl carbonate [90], or with methanol, carbon monoxide, and oxygen [91], [92]. These processes also involve thermolysis of the methyl carbamates as the final step (Fig. 2, path 8). Reaction with unconverted carbamate cannot be prevented because the reactive isocyanates are formed at high temperature. Carbodiimides are also formed as byproducts.

Enichem recently announced that a pilot plant was planned for the phosgene-free two-step manufacture of TDI from TDA using dimethyl carbonate (DMC) to catalytically produce the urethane at 160 – 170 °C (100 % conversion of TDA, 95 % selectivity for urethane), followed by cracking of the urethane in the gas phase at 400 – 450 °C (Fig. 2, paths 6 and 8). It was possible to achieve high conversions of urethane, and all of the methanol could be recycled [93], [94]. ICI disclosed a process in which organic isocyanates (e.g., TDI) were prepared via the corresponding carbamates in high yield and selectivity by treating an amine (e.g., TDA) with DMC in the presence of a catalyst (e.g., lead octanoate) and an organic solvent (chlorobenzene) [95]. High yield and selectivity of the reaction of amines with organic carbonates in the presence of a metal alkanoate catalyst (Ti, Zr, Mn, Fe, Co, Ni, Cu, Zn, Sn, Pb, Bi, Cd) were achieved if the reaction was carried out with DMC, e.g., in the presence of a lead 2-ethylhexanoate catalyst, producing 2,4-bis(methoxycarbonylamino)toluene (2,4-TDU) from TDA [96]. The dicarbamate of bis(1,1,1,3,3,3-hexafluoropropyl) carbonate and MDA could also be prepared in high yield and selectivity. The carbamate could be readily thermolyzed into the corresponding isocyanate MDI [97].

Insufficient overall yield or selectivity was the key problem of the carboalkoxylation/carboaroxylation pathway in the past. Recent developments may have overcome these disadvantages. The only remaining roadblock may be the availability of the necessary carbonates at attractive prices, especially if made without the use of phosgene.

Cytec produces tetramethylxylylene diisocyanate (*m*-TMXDI) by addition of cyanic

acid and hydrogen chloride to 4- or 3-diisopropenylbenzene (Eq. 28) [98].

In the second step of this reaction, carbamic acid chloride (H_2NCOCl) is formed as a byproduct and reacts with diisopropenylbenzene to form the intermediate dichloride and simultaneously regenerate isocyanic acid. The isocyanic acid required is produced in situ by air oxidation of hydrocyanic acid [99].

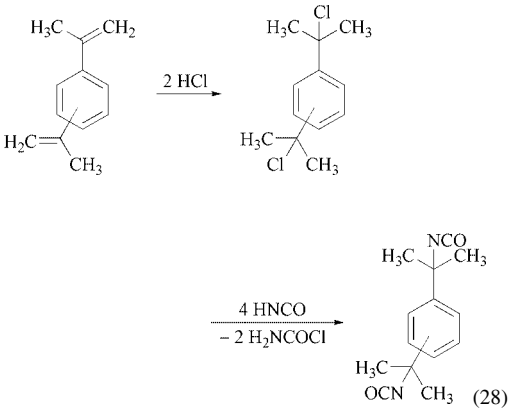

(28)

Du Pont has developed a process for the production of methyl isocyanate based on the oxidative dehydrogenation of *N*-methylformamide (Fig. 2, path 4) [79], [80], but examples referring to commercially important isocyanates making use of this route are rare and show very small overall yields.

To the present day non-phosgene routes developed for TDI and MDI/PMDI have not been adopted commercially, because the manufacture of difunctional aromatic isocyanates still suffers from the formation of nonvaluable residual byproducts, insufficient catalyst stability, selectivity, efficiency and recovery that rendered the alternative processes economically unacceptable. Nevertheless, attempts are still underway to find routes to diisocyanates that are not based on phosgene.

5. Environmentgal Protection

Hydrogen chloride, the byproduct in the production of organic isocyanates, is part of the chlorine cycle in the chemical industry. In modern phosgenation plants all of the hydrogen chloride produced is separated from excess phosgene and solvent and generally passed on to neighboring plants or companies via pipeline which oxidize the hydrogen chloride to chlorine by electrolysis. The chlorine is reused in the production of phosgene, and total HCl recycling thus results. The alternative method is manufacturing with an external supply of chlorine and HCl delivery to an 1,2-dichloroethane/PVC producer. A third option is to absorb the hydrogen chloride gas in water and selling the acid or using it in another process.

Excess phosgene from isocyanate production is separated from hydrogen chloride and recycled. Ammonia or sodium hydroxide scrubbers are available in the case of accidental phosgene release.

The inert solvents used in the production of organic isocyanates are removed by distillation and recycled. Distillation residues and preliminary cuts containing volatile impurities are incinerated. In particular, trace amounts of the volatile phenyl isocyanate formed in MDI production, resulting from trace amounts of aniline in the starting amines, must be removed because phenyl isocyanate is a strong lacrimator.

6. Quality Specifications and Analysis

The use of organic isocyanates as monomers and as intermediates for pharmaceutical and agricultural chemicals requires a high degree of purity. This is usually achieved by fractional distillation. Solid isocyanates are sometimes purified by crystallization or sublimation.

Polymeric MDI (PMDI) is an undistilled (liquid) product. The composition of PMDI varies, the main component being 40 – 60 % 4,4′-MDI; the remainder is composed of other isomers of MDI, triisocyanates, and higher molecular mass oligomers. Important product variables are functionality and acidity. The 4,4′-MDI content varies because some of this coproduct is removed and sold separately, depending on market demand. Samples containing higher amounts of 2,4′-MDI are also produced for specialty elastomer applications. TDI is produced mainly as an 80:20 mixture of the 2,4 and 2,6 isomers. Pure 2,4-TDI and a 65:35 mixture of 2,4- and 2,6-TDI are also available after multiple crystallization steps.

The commodity isocyanates TDI and MDI are required in high quality by the polymer industry. Several phosgenation byproducts are sources of acidic or hydrolyzable chlorine, which strongly influences the catalytic generation of polyurethane foam. Organic isocyanates are usually characterized by their NCO content, determined analytically by reaction of a sample of the isocyanate with excess di-*n*-butylamine in an inert solvent. Unconverted amine is determined by back titration with hydrochloric acid [9]. Total and hydrolyzable chlorides are determined for all isocyanates. Traces of metals such as iron, tin, and zinc, which act as catalysts in polyurethane reactions, are determined by AAS. GC, HPLC, FTIR, and NMR techniques are also used to determine isomer distribution in TDI, MDI, and H_{12}MDI, and to determine the composition of PMDI. Analytical methods used in the determination of organic isocyanates and derived polyurethanes are summarized in [9].

7. Storage and Transportation

The highly reactive organic isocyanates are hazardous materials, and handling of these chemicals is regulated in all industrial countries. Some crystallization of 4,4′-MDI from PMDI may occur on prolonged storage at low temperature. The same is true for the hydrogenated derivative due to separation of solid *trans,trans*-H_{12}MDI. Exposure to atmospheric moisture must be prevented. The reaction of isocyanates with water leads to the formation of carbon dioxide, which results in a build-up of pressure and the formation of urea. The disaster in Bhopal, India, involving massive release of highly volatile and toxic methyl isocyanate was caused by the addition of water to the methyl isocyanate storage tank [101].

Isocyanates can undergo slow cyclodimerization on storage [14]. To minimize dimerization MDI, for example, must be stored under refrigeration. The formation of dimers in MDI causes turbidity on melting. Specific storage recommendations for individual isocyanates are available from the producers. Thermal degradation of isocyanates occurs on heating. This reaction is exothermic, and above 175 °C a runaway reaction can occur with charring and evolution of carbon dioxide. In view of the heat sensitivity of isocyanates, solid isocyanates such as MDI should be melted with caution following the supplier's recommendations. Isocyanates are also flammable and their vapors are toxic (see Chap. 10).

Isocyanates are transported in railroad tank cars, tank trucks, tanks in ships, containers, and drums. They are stored in steel tanks and processed in steel equipment. For long-term storage stainless steel is recommended. To avoid contamination by atmospheric moisture, a dry air or inert gas blanket is essential.

Decontamination of spilled isocyanates and disposal of isocyanate waste are best conducted by using aqueous ammonia (3 – 8 % concentrated ammonia solution in 90 – 95 % water with 0.2 – 5 % liquid detergent) or aqueous sodium carbonate (5 – 10 % sodium carbonate in 90 – 95 % water and 0.2 – 5 % liquid detergent). An alcoholic solution (50 % ethanol, isopropyl alcohol, or butanol; 45 % water; and 5 % concentrated ammonia) may be preferred because of the low miscibility of isocyanates with water.

8. Uses [7], [8]

Monofunctional organic isocyanates (e.g., methyl, *n*-butyl, phenyl, and halogenated phenyl isocyanates) are used as intermediates in the production of carbamate and urea insecticides and fungicides. *n*-Butyl isocyanate is also employed in the manufacture of sulfonyl urea antidiabetic drugs. Arenesulfonyl isocyanates are used in the production of sulfonylurea herbicides and as drying agents for coating systems.

Di- and polyfunctional isocyanates are used as monomers for addition and condensation polymers. For example, the major consumption of the aromatic isocyanates TDI, MDI, and PMDI is in the production of polyurethanes. Water-blown flexible polyurethane foams are produced from TDI and trifunctional polyols; rigid polyurethane and polyisocyanurate foams are made from PMDI and polyester polyols. The open-cell flexible foams are used mainly in bedding and seating; the closed-cell rigid foams are used as insulation in building and construction. Also PMDI is used as a wood chip binder for particle board, and the coproduct MDI is used as a monomer in the production

Table 4. Worldwide TDI and MDI nameplate production capacity (1999/2000) in 10^3 t/a. The total worldwide production capacity of aliphatic diisocyanates and derivatives thereof in 1999/2000 was estimated at 110×10^3 t/a.

Company	TDI	Company	MDI
Bayer	310[a]	Bayer	736[a]
Lyondell[e]	258[b,c]	Huntsman	500[f]
Dow Chemical	210[c]	BASF	433[b]
BASF	160[b,c]	Dow	386[g,h]
Mitsui Toatsu	128[c]	Nippon Polyurethanes	100[f]
EniChem	118[b]	EniChem	80[b]
Korea Fine Chemicals	45[c]	Mitsui Toatsu	75[i]
Huntsman	40[c]	Borsodchem	52[b]
Nan Ya	30[c]	Bayer-Shell Isocyanates	36[a,d]
Takeda Chemical	30[c]		
Bayer-Shell Isocyanates	30[a,d]		
Total (approx.)	**1400**[a]	**Total (approx.)**	**2500**[a]

[a] Source: Bayer AG, BU PU; key date July 01, 2000.
[b] Source: data kindly provided by producer; key date December 31, 1999.
[c] *European Chemical News*, 18 September (2000) 20.
[d] Source: data kindly provided by producer; key date July 01, 2000.
[e] Lyondell is putting Rhodia's European TDI capacities on the market.
[f] Source: *Urethanes Techn.* **16** (1999) 14.
[g] Source: *Chemical Market Reporter* **257** (2000) 29.
[h] Source: *European Chemical News*, 30 November (1998) 20.
[i] Source: *Chemical Week*, April 8 (1998) 37.

of linear thermoplastic and cross-linked thermoset polyurethane elastomers, primarily in the automotive industry. To achieve faster reaction rates and better green strength in reaction injection molding (RIM), aromatic diamine extenders are used to produce polyurethane elastomers. Amine-terminated polyethers have also been used as comonomers to yield polyureas. Spandex-type elastomeric fibers are made from MDI. The aliphatic diisocyanates HDI, IPDI, and H_{12}MDI are used in the formulation of light-stable polyurethane coatings.

Diisocyanates are often used in modified form. In reaction injection molding, liquid diisocyanates are preferred. Liquid modified MDI is made by conversion of some isocyanate groups to carbodiimide groups [44]. The latter undergo cycloaddition with unconverted MDI to form a cycloadduct (Eq. 29). The small amount of cycloadduct depresses the melting point of MDI to give a product that is liquid at room temperature. The cycloadduct dissociates at elevated temperature, thereby maintaining difunctionality in most applications.

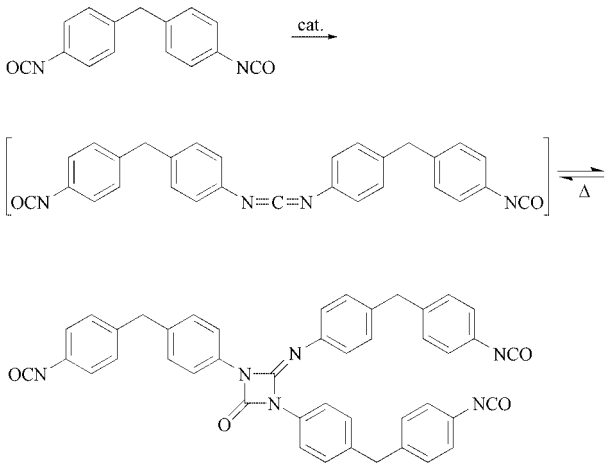

(29)

9. Economic Aspects

Bayer, Pronor, and SNPE/Tolochimie have a pre-eminent position in the monoisocyanates business (Table 1).

The most important aromatic and aliphatic di- or polyisocyanates are listed in Table 2 and Table 3.

Selected MDI and TDI production capacities in 1999/2000 are given in Table 4.

Organic isocyanates are sold by their chemical names or under trade names. The latter are used mainly for the commodity isocyanates. Formulated products and systems containing isocyanates are also sold under a variety of trade names. Some trade names of isocyanates and formulated isocyanates are listed in Table 5.

Table 5. Selected trade names of isocyanates (and/or modifications) for PUR

Trade Name	Producer/Supplier	Product
Basonate	BASF	HDI
Bayhydur	Bayer	HDI (water dispersible polyisocyanate)
Caradate	Bayer-Shell Isocyanates	TDI, MDI – PMDI
Coronate	Nippon PU	TDI, MDI – PMDI
Cosmonate	Mitsui Chem.	TDI, MDI – PMDI, NDI, NBDI
Desmodur	Bayer	TDI, MDI – PMDI, HDI, IPDI, H_{12}MDI, NDI
Enimont	EniChem	MDI – PMDI
Isonate	Dow	TDI, MDI
Lupranat/e	BASF	TDI, MDI – PMDI
Luxate	Lyondell	HDI, IPDI
Millionate	Nippon PU	TDI, MDI – PMDI
Mondur	Bayer	TDI, MDI – PMDI, NDI
PAPI	Dow	PMDI
Rhodocat	Rhodia	HDI (water dispersible polyisocyanate)
Rubinate	Huntsman	MDI – PMDI
Scuranate	Lyondell	TDI
Sumidur	Sumitomo	TDI, MDI – PMDI
Suprasec	Huntsman	MDI – PMDI
Takenate	Takeda	TDI
Tedimon	EniChem	TDI
Tectrade	Huntsman	MDI – PMDI
Tolonate	Rhodia	HDI, IPDI
Vestanat/e	Degussa	IPDI
Voranate	Dow	TDI, MDI – PMDI

In 1998 the worldwide consumption of plastics was estimated at 150×10^6 t/a, of which PU-based materials accounted for ca. 5 % (7.5×10^6t/a). About 30 % of PU was used in the area of furniture/mattresses; automotive and construction applications accounted for 15 % each. Technical insulations accounted for 10 %. The remaining 30 % was divided between coatings, adhesives, TPU, shoes, artificial leather, elastomers, fibers, integral-skin foams, and electronic encapsulation. 80 % of all PU materials could be classified as foam (flexible, rigid, and semi-rigid integral-skin). Total consumption of diisocyanates in 2000 exceeded 3.4×10^6 t (Table 6). The average global growth rate was expected to be ca. 5 % for the PU raw materials TDI, MDI, and polyols.

After years of providing polyurethane precursor materials such as MDI, TDI, or polyols, polyurethane producers appear to be moving towards becoming fully integrated polyurethane providers, with acquisition of ICI's urethane business by Huntsman and, more recently, Bayer's acquisition of Lyondell's polyols business. The companies are making their moves as the polyurethane industry undergoes an extended period of growth [102].

Table 6. Regional demand for the diisocyanates TDI and MDI (1999) in10^3 t/a. (Source: International Isocyanate Institute statistics, May, 2000).

Region	TDI	MDI
NAFTA	335	764
Europe	298	854
Japan	75	165
Asia-Pacific (without Japan)	323	340
Latin America	110	58
Rest of the World	121	84
Total demand	1262	2265
Total consumption	1333[a]	2309[a]

[a] Bayer, PU Business Group, May, 2000.

10. Toxicology and Occupational Health

Isocyanates are classified as dangerous substances (EC Guidelines). Both monomeric and polymeric MDI are labeled "harmful", TDI is labeled "very toxic", most others are labeled "toxic" and should therefore be handled with care.

All diisocyanates are very reactive chemicals that are potentially hazardous to humans. They are harmful by inhalation of vapors and aerosol mists, irritating to eyes, respiratory system, and skin, and some may cause sensitization by inhalation and skin contact. The potential for exposure depends on the vapor pressure.

Table 7. Acute toxicity of diisocyanates in rats (according to MSDS)

Isocyanate	Type	LD_{50} (oral), mg/kg	LC_{50} (inhalation), mg/m^3	Concentration of saturated vapor in air at 25 °C, mg/m^3	Maximum allowable concentration, mg/m^3
HDI	1,6-HDI monomer	746	120 – 350[a]	95	0.035
IPDI	monomer	5490	40[b]	3.65[c]	0.09
H_{12}MDI	4,4'-H_{12}MDI monomer	>11 000 (>10 000[d])	434[b]	0.22	0.054
TDI	TDI-80	5800	610[e]	255	0.07
MDI-PMDI	4,4'-MDI monomer	>2500	490[b]	0.09	0.05
NDI	1,5-NDI	>15 000	>2075[f]		0.09

[a] 4-h exposure. [b] Aerosol, 4-h exposure. [c] 20 °C. [d] LD_{50} (dermal), rabbit. [e] Aerosol, 1-h exposure. [f] 1-h exposure.

Volatile short-chain monomeric aliphatic isocyanates (e.g., methyl isocyanate) must be handled with particular caution. The more volatile isocyanates are strong lachrymators, but their primary physiological effect is respiratory irritation [103]. Therefore isocyanates must be handled with procedures designed to minimize exposure.

For TDI, OSHA has set a PEL of 0.02 ppm as a ceiling concentration and a STEL of 0.02 ppm as a 15-min TWA. For MDI, OSHA guidelines are a PEL of 0.02 ppm as a ceiling concentration. Up to late 1998, OSHA guidelines have not been established for major aliphatic diisocyanates (HDI, IPDI, H_{12}MDI) [104]. Data on the acute toxicity of several diisocyanates are listed in Table 7.

There are contradictory results on the mutagenicity of TDI and MDI. According to the Salmonella/microsome test [105] in a suitable matrix for isocyanates, MDI gives negative findings while TDI gives positive findings due to the instability of TDI leading to TDA, a known mutagen [106].

11. References

1. W. Hentschel, *Chem. Ber.* **17** (1884) 1284.
2. O. Bayer, *Angew. Chem.* **59** (1947) 257.
3. W. Siefken, *Justus Liebigs Ann. Chem.* **562** (1949) 75.
4. M. Szycher: *Szycher's Handbook of Polyurethanes,* CRC Press, Boca Raton 1999, pp. 4-6 – 4-13.
5. H. Ulrich: *Chemistry and Technology of Isocyanates,* Wiley, Chichester 1996, pp. 42 – 133, 347 – 356, 395 – 426.
6. S. Ozaki, *Chem. Rev.* **72** (1972) 457.
7. D. A. Wicks, Z. W. Wicks, *Prog. Org. Coat.* **36** (1999) 148.
8. G. Oertel: *Polyurethane Handbook,* 2nd ed., Hanser Publishers, Munich 1993.
9. D. J. David, H. B. Staley: *Analytical Chemistry of Polyurethanes,* Wiley-Interscience, New York 1969, pp. 85 – 112.
10. C. Zwiener, L. Schmalstieg, M. Sonntag, *Europ. Coat. J.* **10** (1992) 588.
11. P. S. Carleton, W. J. Farrissey, J. Rose, *J. Appl. Polym. Sci.* **16** (1972) 2983.
12. K. H. Slotta, R. Tschesche, *Ber. dtsch. Chem. Ges.* **60** (1927) 295.
13. F. Richter, H. Mertes, *Farbe Lack* **106** (2000) 60.
14. R. B. Wilson, Y. S. Chen, I. C. Paul, D. Y. Curtin, *J. Am. Chem. Soc.* **105** (1983) 1672.
15. *Houben-Weyl,* **E-4,** pp. 883 – 912.
16. V. E. Shashoua, W. Sweeny, R. F. Tietz, *J. Am. Chem. Soc.* **82** (1959) 866.
17. H. J. Laas, R. Halpaap, J. Pedain, *J. prakt. Chem.* **336** (1994) 185.
18. H. Ulrich: "Isocyanate-Derived Polymers", *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **Vol. 8,** J. Wiley & Sons, New York 1987, pp. 448 – 462.
19. H. Ulrich: *Chemistry and Technology of Isocyanates,* Wiley, Chichester 1996.
20. M. Szycher: *Szycher's Handbook of Polyurethanes,* CRC Press, Boca Raton 1999, pp. 4-20 – 4-40.
21. *Houben-Weyl,* **E-4,** pp. 738 – 784.
22. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 4th ed., VCH, Weinheim 1994, pp. 408 – 413
23. H. J. Twitchett, *Chem. Soc. Rev.* **3** (1974). 209.
24. BASF, DE-OS 3 314 788 1984 (F. Merger, F. Towae, H. Hellbach, W. Koehler, G. Isbarn).
25. BASF, DE-OS 3 339 300, 1985 (O. Mattner, F. Merger, F. Towae).
26. Bayer, DE-OS 3 315 191, 1984 (D. Becher, H. Witt, H. Dallmeyer, S. Humburger, W. Schneider, H. Stein).

27. Bayer, EP-A 634 391, 1995 (U. J. Zarnack, F. Pohl, D. Grenner, H. Hetzel, H. Judat).
28. Bayer, EP-A 757 034, 1997 (K. Biskup, B. Keggenhoff).
29. Bayer, WO 99/19292, 1999 (H. Beckhaus, R. Langer).
30. BASF, DE-OS 19 857 409, 2000 (M. Sander, U. Penzel, H. V. Schwarz, E. Ströfer, D. Stützer, J. Müller, M. Maurer).
31. Air Products and Chemicals, US 6 005 143, 1999 (R. M. Machado, D. J. Parrillo, R. P. Boehme, R. R. Broekhuis).
32. Bayer, DE 3 736 988, 1989 (G. Zaby, H. Judat, S. Humburger, S. De Vos, R. W. Eckermann).
33. BASF, DE-OS 3 403 204, 1985 (R. Ohlinger, H. Schnez, L. Pfannenstiel, B. Blumenberg, H. J. Raabe).
34. Bayer, DE-OS 3 407 494, 1985 (B. Keggenhoff, E. Mählmann, W. Eifler, G. Ellendt).
35. Upjohn, US 4 020 104, 1977 (R. H. Richter).
36. Upjohn, US 4 026 943, 1977 (R. H. Richter).
37. BASF, WO 99/54289, 1999 (U. Penzel, V. Scharr, D. Starosta, H. Bösel, E. Ströfer, J. Pfeffinger, F. Poplow, J. Dosch, H. V. Schwarz, F. Naumann, P. Van Den Abeel, J. Jacobs, F. Nevejans, W. Van Pee).
38. E. I. Du Pont De Nemours, WO 97/24320, 1997 (F. J. Freire, B. A. Kaiser, V. M. Felix, D. T. Mah, J. A. I. Trainham, C. G. Law, J. S. Newman, D. J. Eames).
39. Bayer Hispania Industrial, DE-OS 4 142 769, 1993 (J. C. Arribas, C. D. Arribas, S. V. Rodriguez).
40. Bayer, EP-A 626 368, 1994 (H.-J. Benedix, W. Motika).
41. Bayer, EP-A 699 659, 1996 (H. Hetzel, D. Grenner, W. Ebner, K. Biskup).
42. Olin, US 4 091 009, 1978 (J. R. Cassata).
43. Olin, US 4 137 266, 1979 (J. R. Cassata).
44. BASF, WO 99/65868, 1999 (H. V. Schwarz, M. Maurer, M. Sander, E. Strofer, R. Leuthold, B. Hantel, S. Richter, J. Therre).
45. Mitsui Chemicals, JP 2000 136 264, 2000 (M. Iwama, S. Ogawa).
46. Dow Chemical, WO 00/68180, 2000 (D. S. Treybig, K. J. Hock, V. A. Kent, S. A. Dai).
47. Bayer, EP 567 881, 1993 (C. König, M. Gallus, R. Wüllner, F. M. Richter, R. Uchdorf, W. C. Forester).
48. Bayer, DE-OS 3 714 439, 1988 (H. G. Frosch, H. Grave, H. Stutz, E. Waldau, P. Fuhrmann).
49. Bayer, EP-A 676 392, 1995 (E. Bischof, P. Breidenbach, J. Dahmer, A. Flink, A. Molnar, H. Stutz).
50. Bayer, EP-A 749 958, 1996 (E. Bischof, J. Dahmer, A. Flink, W. Krohn, A. Molnar).
51. Bayer, EP-A 570 799, 1993 (K. Biskup, C. König, E. Waldau).
52. Rhône-Poulenc Chimie, EP-A 593 334, 1994 (F. Joulak, D. Revelant, P. Vacus).
53. Rhône-Poulenc Chimie, EP-A 699 657, 1996 (J. Armand, D. Revelant, H. Chiarelli, P. Vacus).
54. Rhône-Poulenc Chimie, US 5 679 839, 1997 (J. Armand, H. Chiarelli, D. Revelant, P. Vacus).
55. Upjohn, US 3 275 669, 1966 (H. Ulrich, J. N. Tilley, A. A. Sayigh).
56. Scholven-Chemie, DE 1 202 785, 1965 (K. Schmitt, F. Gude, K. Rindtorff, Jr., J. Disteldorf).
57. Farbenfabriken Bayer, DE 947 471, 1956 (J. Pfirschke).
58. Monsanto, US 5 233 010, 1993 (W. D. McGhee, M. K. Stern, T. E. Waldman).
59. Monsanto, WO 95/18098, 1995 (W. D. McGhee, M. D. Paster, D. P. Riley, K. W. Ruettimann, A. J. Solodar, T. E. Waldman).
60. H. Ulrich, A. A. R. Sayigh, *Angew. Chem. Intern. Ed. Engl.* **5** (1966) 724.
61. SNPE, EP-A 723 955, 1996 (P. Hussenet, P. Le Goff, G. Sennyey).
62. H. Ulrich, B. Tucker, A. A. R. Sayigh, *J. Org. Chem.* **31** (1966) 2658.
63. Upjohn, US 3 371 114, 1968 (A. A. R. Sayigh, H. Ulrich).
64. Upjohn, US 3 410 887; 1968 (A. A. R. Sayigh, H. Ulrich).
65. K. Kurita, Y. Iwakura, *Org. Synth.* **59** (1980) 195.
66. H. Eckert, B. Forster, *Angew. Chem.* **99** (1987) 922.
67. B. M. Pope, Y. Yamamoto, D. S. Tarbell, *Org. Synth.* **53** (1973) 1851.
68. B. M. Pope, Y. Yamamoto, D. S. Tarbell, *Org. Synth.* **57** (1977) 45.
69. Mitsubishi Gas Chemical, JP 06 100 501, 1994 (M. Tsunoda, I. Hagiwara, N. Okuda).
70. Nihon Tokushu Noyaku Seizo, JP 7 589 328, 1975 (J. Saito, T. Tamura, N. Morishima).
71. Nippon Kayaku, JP 10 007 623, 1998 (S. Hayashi).
72. L. Cotarca, P. Delogu, A. Nardelli, V. Sunjic, *Synthesis* (1996) 553.
73. H. W. I. Peerlings, E. W. Meijer, *Tetrahedron Lett.* **40** (1999) 1021.
74. R. M. Versteegen, R. P. Sijbesma, E. W. Meijer, *Angew. Chem. Int. Ed.* **38** (1999) 2917.

75. F. Paul, *Coord. Chem. Rev.* **203** (2000) 269.
76. R. Sundermann, H. J. Scholl in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **Vol. 2,** VCH, Weinheim 1996, pp. 1072 – 1080.
77. S. Cenini, F. Ragaini: *Catalytic Reductive Carbonylation of Organic Nitro Compounds,* Kluwer, Dordrecht 1997.
78. F. Ragaini, S. Cenini, *Chim. Ind. (Milan)* **78** (1996) 421.
79. E. I. du Pont de Nemours, US 4 469 640, 1984 (P. F. Carcia, G. E. Heinsohn,V. N. M. Rao).
80. E. I. du Pont de Nemours, US 4 683 329, 1987 (V. N. M. Rao).
81. Bayer, EP-A 520 273, 1992 (P. Heitkämper, T. Schieb, S. Wershofen).
82. H. Ulrich: *Chemistry and Technology of Isocyanates,* Wiley, Chichester 1996, pp. 375 – 379.
83. A. Klausener, J.-J. Jentsch in B. Cornils, W. A. Herrmann (Eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **Vol. 1,** VCH, Weinheim 1996, p. 182.
84. BASF, DE-OS 3 314 790, 1984 (H. Hellbach, F. Merger, F. Towae).
85. BASF, US 4 713 476, 1987 (F. Merger, F. Towae).
86. BASF, US 4 851 565, 1989 (F. Merger, F. Towae).
87. BASF, EP 566 925, 1993 (A. Otterbach, H. V. Schwarz, F. Merger, W. Schwarz, E. Brandt, P. Magnussen, O. Mattner).
88. BASF, DE-OS 19 627 552, 1998 (G. Laqua, U. Schoner, E. Ströfer, A. Otterbach, J. Pfeffinger).
89. BASF, DE-OS 19 820 114, 1999 (J. Pfeffinger, P. Pfab, E. Ströfer).
90. Dow Chemical, US 4 268 683, 1981 (A. E. Gurgiolo).
91. S. Fukuoka, M. Chono, M. Kohno, *J. Org. Chem.* **49** (1984) 1458.
92. S. Fukuoka, M. Chono, M. Kohno, *CHEMTECH* **14** (1984) 670.
93. P. Cesti, *Chim. Ind. (Milan)* **81** (1999) 324.
94. EP-A 881 213, 1998 (A. Bosetti, E. Cauchi, V. Carletti, P. Cesti).
95. Imperial Chemical Industries, WO 98/56758, 1998 (R. C. Smith, J. C. B. Hunns).
96. Imperial Chemical Industries, WO 98/55451, 1998 (R. C. Smith).
97. Huntsman ICI Chemicals, WO 99/47493, 1999 (J. Bosman).
98. K. A. Henderson, Jr., V. A. Alexanian, *Org. Prep. Proced. Int.* **18** (1986) 149 – 155.
99. American Cyanamid, US 4 364 913, 1982 (D. S. Katz, K. E. Olson).
100. Akzo, AU 523 314, 1982 (P. Hentschel, H. Zengel, M. Bergfeld).
101. T. D. J. D'Silva, A. Lopes, R. L. Jones, S. Singhawangcha, J. K. Chan, *J. Org. Chem.* **51** (1986) 3781.
102. R. Brown, *CMR focus* (2000) June 19, FR12.
103. R. N. Shiotsuka, D. L. Warren, A. T. Halliburton, D. W. Sturdivant, *Inhalation Toxicol.* **12** (2000) 605.
104. H. Chinn, *CEH Marketing Research Report - Diisocyanates and Polyisocyanates,* SRI International, September, 1998.
105. D. M. Maron, B. N. Ames, *Mutation Res.* **113** (1983) 173.
106. K. Seel, U. Walber, B. Herbold, R. Kopp, *Mutation Res.* **438** (1999) 109.

Isocyanuric Acid → Cyanuric Acid and Cyanuric Chloride
Isoleucine → Amino Acids
Isophorone → Ketones

Isoprene

Hans Martin Weitz, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2–7)

Eckhard Loser, Bayer AG, Wuppertal, Federal Republic of Germany (Chap. 8)

1. Introduction 1
2. Properties 2
3. Production 3
3.1. Synthetic Methods 3
3.2. Dehydrogenation Procedures 6
3.3. Recovery from C_5 Cracking Fractions 7
4. Quality Specifications 10
5. Storage and Transportation 11
6. Uses 11
6.1. Polymer Synthesis 11
6.2. Terpene Synthesis 12
7. Economic Aspects 13
8. Toxicity and Occupational Health . . 14
9. References 17

1. Introduction

Isoprene [*78-79-5*] (**1**), C_5H_8, M_r 68.118, is named according to IUPAC rules as 2-methyl-1,3-butadiene. Its isomers include the commercially less important compounds *cis*- and *trans*-1,3-pentadiene (**2**) [*504-60-9*] (piperylene).

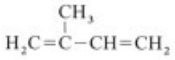

1

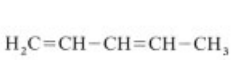

2

Free isoprene has been observed in nature only in very low concentration [1], [2]. The compound is present in roasted coffee and in the gas phase of tobacco smoke [3], and can be regarded as a precursor of polycyclic aromatics [4]. Isoprene is the structural unit of countless natural products (the terpenes, which include natural rubber and camphor) as well as biologically important substances such as vitamin A and the steroid sex hormones (cf. the "isoprene rule" of O. Wallach and L. Ruzicka [5]). It is now known, however, that the biosynthesis of rubber and other natural products containing the isoprene skeleton proceeds not via isoprene itself, but rather via mevalonic acid [*150-97-0*] (**3**), 3,5-dihydroxy-3-methylpentanoic acid [6].

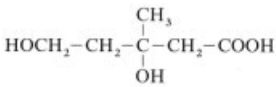

3

Isoprene was first synthesized in 1860 by C. E. Williams by the pyrolysis of natural rubber [7]. The reverse reaction–polymerization of isoprene to poly(*cis*-1,4-isoprene), with a structure corresponding to that of natural rubber–was long the subject of intensive effort [8]. The first successful attempts were reported in 1954 and 1955 by the Goodrich Gulf (Al–Ti Ziegler catalyst [9]) and Firestone companies (alkyl lithium catalyst [10]).

Isoprene was commonly prepared on a laboratory scale by thermolysis of turpentine oil (the so-called isoprene lamp of C. Harries [11]). Pyrolysis of dipentene [*5989-27-5*], limonene, was even used in the United States early in the Second World War as a commercial source of isoprene (Bibb process [12]). The reverse of this reaction–synthesis of terpenes from isoprene–has been discussed in a vast number of scientific publications, but has so far achieved little commercial significance.

Isoprene itself was commercially unimportant until after the Second World War because it could not be offered at a price that was sufficiently attractive for its principal potential market, the manufacture of synthetic rubber. This situation changed with the development of improved methods for obtaining isoprene from petrochemical sources, as well as of polymerization techniques for the generation of poly(*cis*-1,4-isoprene), a synthetic rubber whose valuable material characteristics have been widely confirmed.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a14_627

2. Properties

Physical Properties. Under normal conditions, isoprene is a colorless, volatile liquid. Important properties of isoprene are(see also [15]):

mp (101.3 kPa)	−145.95 °C
bp (101.3 kPa)	34.059 °C
Critical data [13]	
Temperature	483.3 K
Pressure	3.74 MPa
Volume	266 cm^3/mol
Density (293 K)	0.68095 g/cm^3
Viscosity (293 K)	0.216 mPa · s
Surface tension (293 K)	18.22 mN/m
Refractive index n_D^{20}	1.42194
Vapor pressure [14] at − 20 °C	9.8 kPa
0 °C	26.4 kPa
20 °C	60.7 kPa
34.059 °C	101.3 kPa
40 °C	123.8 kPa
60 °C	229.1 kPa
80 °C	392.1 kPa
100 °C	629.5 kPa
Specific heat	
Vapor (298 K)	102.69 J mol^{-1} K^{-1}
Liquid (298 K)	151.07 J mol^{-1} K^{-1}
Heat of formation	
Vapor (298 K)	75.75 kJ/mol
Liquid (298 K)	49.36 kJ/mol
Free energy of formation	
Vapor (298 K)	146.23 kJ/mol
Liquid (298 K)	145.57 kJ/mol
Entropy of formation	
Vapor (298 K)	314.56 J mol^{-1} K^{-1}
Liquid (298 K)	228.3 J mol^{-1} K^{-1}
Heat of combustion (298 K)	−3186.58 kJ/mol
Heat of hydrogenation (298 K)	−230.20 kJ/mol
Heat of fusion (− 145.95 °C)	4.88 kJ/mol
Heat of vaporization (at 25 °C)	26.39 kJ/mol
at 34.059 °C	25.87 kJ/mol
Heat of polymerization (298 K)	− 75 kJ/mol
Entropy of polymerization (298 K)	− 101 J mol^{-1} K^{-1}
Constants for the Antoine equation $\log p = A - B/(t + C)$, p in hPa, t in °C [14]	A 6.05329
	B 1092.997
	C 236.002
Flash point	−48 °C
Ignition temperature	220 °C
Explosive limits in air (total pressure 13 kPa, 25 °C)	
Upper limit	7 – 9.7 vol%
	200 – 275 g/m^3
Lower limit	1 – 1.5 vol%
	28 – 40 g/m^3
Maximal explosion pressure	0.66 MPa

Spectra are to be found in the literature: IR [16], UV [17], NMR [18], Raman [19].

The *solubility* of isoprene in water at 20 °C is 0.029 mol%. It is miscible in all proportions with organic solvents such as ethanol, diethyl ether, acetone, and benzene. Regarding solubility in high-boiling, polar solvents, see chap.4. The Bunsen absorption coefficient α for oxygen in isoprene is 0.4065 m^3 m^{-3} $mbar^{-1}$ at 0 °C and 0.4557 at 20 °C.

Binary *azeotropic mixtures* of isoprene are listed in Table 1. The azeotrope of isoprene with *n*-pentane has acquired commercial significance in the isolation of pure isoprene . For information regarding ternary azeotropes of isoprene see [19].

Table 1. Binary azeotropes of isoprene [19]

Component	*bp*, °C	Isoprene content, wt %
Methanol	29.57	94.8
n-Pentane	33.6	72.5
Carbon disulfide	<34.15	<93
Methyl formate	22.5	50
Bromoethane	32	>65
Ethanol	32.65	97
Dimethyl sulfide	32.5	65
Acetone	30.5	80
Propylene oxide	31.6	40
Ethyl formate	<32.5	>76
Isopropyl nitrite	33.5	72
Methylal	32.8	70
Diethyl ether	33.2	52
Perfluorotriethylamine	30.2	82
Acetonitrile	33.5 – 33.6	97.5
Isopropylamine		>72.4

Data for vapor – liquid equilibria of binary and ternary mixtures of isoprene and other organic compounds are provided in [20] and [21]. Liquid – liquid equilibrium data for binary and multicomponent systems containing isoprene are listed in [22]. Activity coefficients at infinite dilution are given in [23]. Curves describing the temperature dependence of various physical quantities may be found in [24] and [25].

Chemical Properties. Spectroscopic analysis has revealed that most isoprene molecules in the ground state at 50 °C are in the *s-trans* conformation. Only 15 % of the material has the *s-cis* form, which is higher in energy by 6.3 kJ/mol.

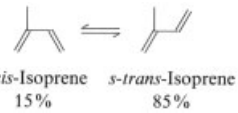

Equilibrium concentrations for the isomer system isoprene – pentadiene (a total of seven isomers) are presented graphically in [26] for temperatures between 200 and 1600 K. Reaction enthalpies for the corresponding isomerizations are provided in [27].

Isoprene exhibits the typical characteristics of a conjugated diene. Its methyl group makes the compound react more readily than butadiene with electrophiles and Diels – Alder dienophiles. Moreover, the reactivity of isoprene is more varied than that of butadiene, which only possesses two types of hydrogen atoms in contrast to the four of isoprene. Isoprene can participate in a wide range of reactions (including substitution, addition, ring formation, complexation, and telomerization reactions). For a summary of the general reactivity of isoprene see [19].

The only chemical reaction of isoprene of practical importance is its conversion to terpenes. Isoprene is mainly used for polymer synthesis (Chap. 6).

3. Production

3.1. Synthetic Methods

Laboratory methods for the synthesis of isoprene from a wide variety of starting materials are summarized in [19].

Industrial syntheses of isoprene utilize the following four principles in creating the C_5 skeleton:

$C_1 + C_4 \longrightarrow C_5$
$C_2 + C_3 \longrightarrow C_5$
$C_3 + C_3 \longrightarrow C_6 \longrightarrow C_5 + C_1$
$C_4 + C_4 \longrightarrow C_8 \longrightarrow C_5 + C_3$

$C_1 + C_4 \longrightarrow C_5$. *Isobutene* [*115-11-7*] is the C_4 building block in several syntheses of this type (Fig. 1). The C_1 component may be formaldehyde [*50-00-0*] (reactions 1, 2, and 3 in Fig. 1), one of its derivatives, or methanol [*67-56-1*] [28]. The C_1 molecule carbon monoxide also plays a role in synthesis; this route (4 in Fig. 1) involves hydroformylation of isobutene to 3-methylbutanal [30].

The most frequently used synthetic procedure is the acid-catalyzed addition of formaldehyde to isobutene (Prins reaction, route 1 in Fig. 1). This produces a 1,3-dioxane, which is then cleaved in the gas phase (at 200 – 300 °C) over an acid catalyst (e.g., H_3PO_4 on a suitable carrier) to give isoprene, with recovery of half of the reacted formaldehyde. This reaction was first suggested for the synthesis of isoprene in 1938 [31], and it became widely known through the work of the Institut Français Pétrol [32], [33]. The method became the subject of further development at several locations as a result of the ready availability of the starting materials [34–36]. Individual procedures differ primarily in terms of catalysts employed and engineering details. Production facilities of this type are currently in operation in Japan (cf. Chap. 7).

The ultimate goal of developmental efforts in the synthesis of isoprene from isobutene and formaldehyde has been simplification and increased competitiveness. Takeda Chemical in Japan has proposed a one-step gas-phase procedure (cf. Fig. 1, route 2), in which isobutene and formaldehyde are passed over oxide catalysts at 300 °C [36]. Much preliminary work toward this approach was contributed by British Hydrocarbon Chemicals.

An important aspect of the above syntheses with respect to their economic viability is the problem of byproduct formation from formaldehyde. Efforts have therefore been made to use formaldehyde derivatives as a way of suppressing resin buildup. For example, both Idemitsu and Sun Oil have suggested the use of methylal [*109-87-5*] (**4**), dimethoxymethane [37], [38]. Marathon Oil has proposed a two-step process based on monochlorodimethyl ether [*107-30-2*] (**5**) [39], [40], while Sumitomo has discussed the use of dioxolane [*646-06-0*] (**6**), 1,3-dioxacyclopentane [41].

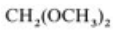

4

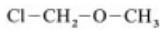

5

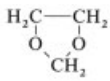

6

Consideration has also been given to integrating formaldehyde production directly into the process, i.e., by mixing methanol and oxygen directly with isobutene [42]; a modification of this approach is the use of methyl *tert*-butyl ether (MTBE) and oxygen [43].

Japanese patent applications from Kuraray, Nippon Zeon, and Sumitomo Chemical use *tert*-

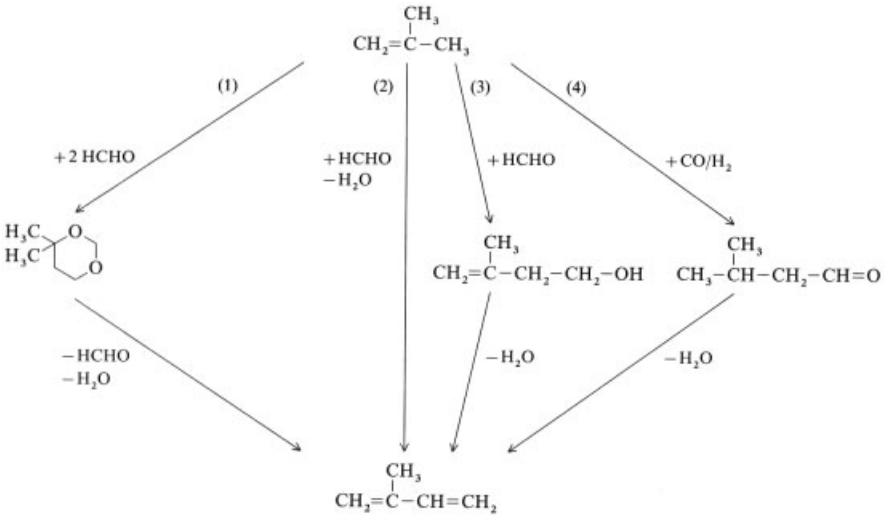

Figure 1. Synthesis of isoprene from isobutene

butyl alcohol [*75-65-0*] as a C_4 building block in isoprene synthesis, while a number of former Soviet publications describe a variety of methods based on cleavage of 1,3-dioxane. Probably the greatest disadvantage of the one-step gas-phase process is the low space – time yield of the catalysts, together with their rapid deactivation under the prevailing reaction conditions [44–46]. It is not known whether any of the above-described variants have been carried to production scale.

In addition to the Prins reaction with all its modifications, there is one other economically interesting route to the synthesis of isoprene from isobutene and formaldehyde via 2-methyl-1-buten-4-ol (**7**) (Fig. 1, reaction 3) [47]:

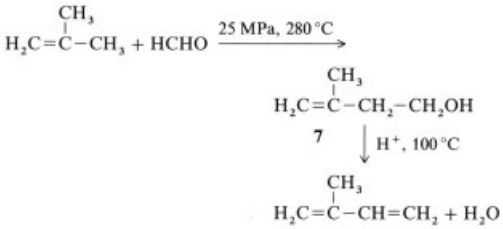

One advantage here is that the synthesis (ene reaction) of the intermediate **7** is a purely thermal process occurring in the liquid phase, which permits use of an aqueous formaldehyde solution [48]. Dehydration of **7** is conducted in aqueous NaCl/HCl solution [49] that can be easily regenerated (low catalyst cost).

2-Butene [*107-01-7*] is the C_4 building block for an entirely different synthesis of isoprene based on the $C_1 + C_4 \rightarrow C_5$ principle via the intermediate 2-methylbutanal [*96-17-3*] (Fig. 2). 2-Methylbutanal is readily prepared by rhodium-catalyzed hydroformylation of 2-butene [30], [50], [51]. An alternative route to this C_5 aldehyde is the catalytic hydrogenation of 2-ethylacrolein which is easily prepared from *n*-butyraldehyde and formaldehyde [52], [62].

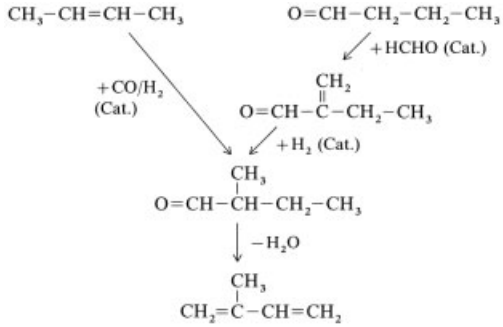

Figure 2. Synthesis of isoprene via 2-methylbutanal.

Catalytic dehydration of 2-methylbutanal to isoprene has been the subject of numerous studies [53]. A patent issued to Erdölchemie [54] describes the use of a β-phosphate catalyst, which is also discussed in most of the subsequent publications dealing with this reaction [30], [50], [55–57]. Other patents for the dehy-

dration of 2-methylbutanal describe the use of magnesium ammonium phosphate [58], molecular sieve [59], as well as zeolite catalysts [60]. Nevertheless, all the catalysts cited are subject to rapid loss of activity [30], or else display low activity from the outset. Zeolites that have been silanized or doped with small amounts of cesium display longer lifetimes. For example, a zeolite containing 0.4 % cesium showed no decrease in yield (51 %) or isoprene selectivity (88 %) after 120 h of use [61].

Cleavage of 3-methylbutanal over a boron phosphate catalyst (Fig. 1, route 4) gave poorer results than were observed with the 2-isomer [30], [62]. At the present time there are still no industrial facilities for the synthesis of isoprene from *n*-butenes or butanal via 2-methylbutanal.

$C_2 + C_3 \longrightarrow C_5$. Codimerization of ethylene and propene in the presence of triethylaluminum leads to 2-methyl-1-butene, which can be dehydrogenated to isoprene. However, this method has not been carried beyond the experimental stage [63], [64].

By contrast, another isoprene synthesis based on C_2 and C_3 starting materials (acetylene and acetone) was carried to production scale by SNAM in Italy, whose manufacturing facilities in Ravenna had been in operation for several years [65–68]; the plant is, however, not currently in operation. A plant based on this principle has been built in the Republic of South Africa (see Chap. 7). The isoprene produced by this method is extremely pure but the process is relatively expensive.

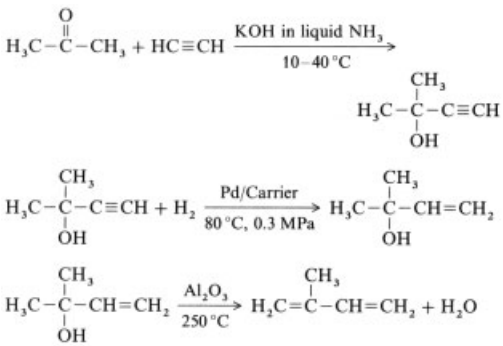

$C_3 + C_3 \longrightarrow C_6 \longrightarrow C_5 + C_1$. The first step of the familiar Goodyear – Scientific Design isoprene synthesis consists of the dimerization of propene to 2-methyl-1-pentene. This is then isomerized to 2-methyl-2-pentene, which is subsequently cracked with loss of methane to give isoprene [64], [69], [70].

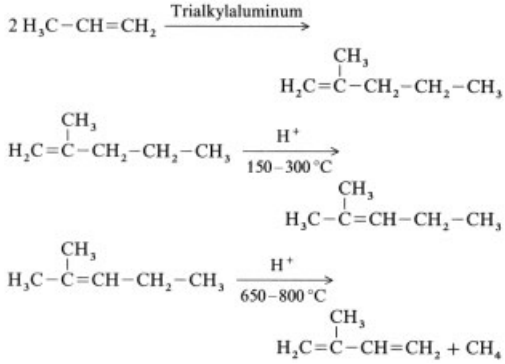

The first commercial synthesis of isoprene was based upon this principle (Beaumont, Texas). Production was discontinued after a fire in 1975 because it was no longer competitive due to sharp increases in the price of propene.

$C_4 + C_4 \longrightarrow C_8 \longrightarrow C_5 + C_3$. Olefin metathesis represents a very interesting approach to isoprene manufacture from readily available petrochemicals. Thus, a butene fraction containing 2-butene and isobutene yields 2-methyl-2-butene and propene:

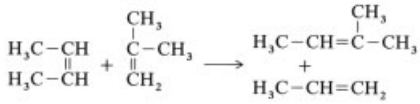

2-Methyl-2-butene can then be dehydrogenated to isoprene by standard methods (cf. Section 3.2) [71–75]. However, one disadvantage of the metathesis reaction is that all of the olefins in the reactor can react with one another yielding a broad spectrum of byproducts. This is especially true if the starting material is a technical-grade butene mixture containing 1-butene, which represents the only economically feasible starting material [76].

3.2. Dehydrogenation Procedures

The production of isoprene by dehydrogenation of isopentane or the isopentenes (methyl-

butenes) has been the subject of many investigations [77], [78]. Such reactions closely resemble the analogous dehydrogenations of *n*-butane and the *n*-butenes to butadiene (→ Butadiene, Chap. 4.1.).

One-step dehydrogenation of isopentane to isoprene can be carried out according to the Houdry – Catadiene procedure (Cr_2O_3/Al_2O_3 catalyst, ca. 600 °C, and ca. 7 kPa) with a yield of 52 % [79–81]. Isoprene is prepared commercially in this way in the former Soviet Union [82]. A two-step process for the dehydrogenation of isopentane is described in [83].

Isoprene production plants based on the dehydrogenation of methylbutenes exist in the United States and the Netherlands but are not currently in operation. The requisite starting material comes largely from cat-cracker off-gases [84], [85]. Methylbutenes are extracted in the form of semiesters from the appropriate distillation fraction of these off-gases by dissolution in sulfuric acid, followed by back extraction with paraffinic hydrocarbons (Sinclair procedure). As far as is known, dehydrogenation is effected with a Shell catalyst (Fe_2O_3– K_2CO_3– Cr_2O_3) at 600 °C and dilution of the methylbutenes with steam; the yield is 85 % [86]. A Sr – Ni – phosphate catalyst (Dow Type S catalyst, a modification of the well-known Dow Type B catalyst) has also been reported to be suitable for this reaction [87].

Oxidative dehydrogenation of isopentane [88–90] and the methylbutenes [91–95] has also been widely investigated. The behavior of oxide catalysts in the oxidative dehydrogenation of the methylbutenes is described in [45]. The yield and selectivity for isoprene are reported to be 60 % and 95 %, respectively.

Although many publications and patents discuss the preparation of isoprene by dehydrogenation, little significant progress is apparent. Two former Soviet reports will be mentioned as examples. The oxidative dehydrogenation of methylbutenes to isoprene with aluminum phosphate catalysts is described in [96]; the reactivity of the olefins decreases in the order 3-methyl-1-butene, 2-methyl-2-butene, and 2-methyl-1-butene. Oxidative dehydrogenation of methylbutenes to isoprene can also be achieved with a silicate catalyst doped with alkaline-earth oxides or the oxides of nickel, iron, or cobalt [97].

Conversion of isopentane and methylbutenes to isoprene by co-oxidation is claimed in [98].

3.3. Recovery from C_5 Cracking Fractions

The C_5 cracking fractions obtained as a byproduct in the pyrolysis of hydrocarbons to ethylene (→ Ethylene) have come to play a major role in isoprene manufacture [99]. Workup of the cracking products from naphtha cleavage yields so-called " crack gasoline", which contains only low concentrations of isoprene and other C_5 hydrocarbons, the major constituents being C_6– C_8 aromatics. Nevertheless, this material may still be distilled into C_5 and aromatic fractions.

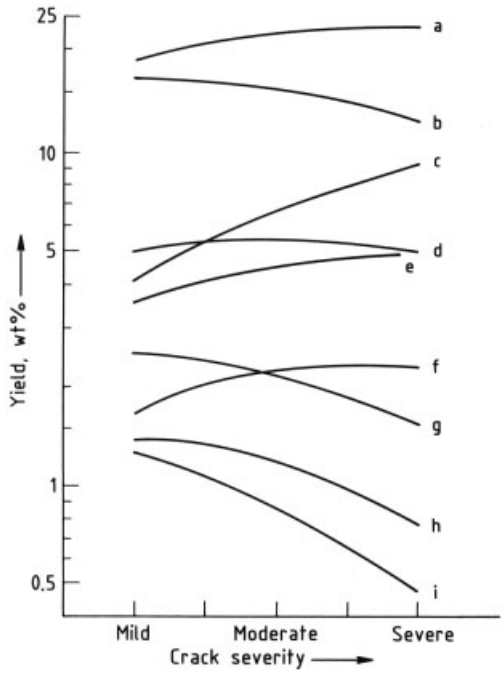

Figure 3. Influence of cracking severity on the product distribution [102]
a) Ethylene; b) Propene; c) Benzene; d) Butadiene; e) Toluene; f) Cyclopentadiene; g) Isobutene; h) Isoprene; i) Piperylenes

The yield of isoprene is typically 2 – 5 wt % based on ethylene, although it may be increased by starting with a heavier raw material such as gas oil. Table 2 demonstrates the influence of the nature of the cracker feed on the C_5 product distribution. Figure 3 shows that the yield of isoprene decreases with increasing crack severity

Table 2. Composition (in weight percent) of the C_5 fraction from the pyrolysis of hydrocarbons [100] and boiling points of the individual components

Component	*bp* at 101.3 kPa, °C	Starting material					
		1	2	3	4	5	6
Isopentane	27.85	1.99	11.02	0.99	24.1	7.8	0.8
n-Pentane	36.07	0.6	2.54	2.72	26.1	14.3	20.8
2-Methyl-1-butene	31.16	4.95	5.21	3.80	11.8	2.9	0.8
2-Methyl-2-butene	38.57	2.61	2.13	3.01	11.8	2.4	0.8
1-Pentene	29.97	24.10	3.84	4.08	4.2	2.9	3.1
trans-2-Pentene	36.35	4.74	2.14	2.57	4.2	5.8	
cis-2-Pentene	36.94	3.47	1.54	1.36	4.2	5.8	1.6
trans-1,3-Pentadiene	42.03	14.32	4.95	6.84	8.8	7.3	10.2
cis-1,3-Pentadiene	44.07	8.00	22.13	3.65	8.8	4.7	5.5
Isoprene	34.06	6.71	6.65	4.34.	13.7	16.1	13.3
Cyclopentadiene	42.50	16.46	2.40	17.77	7.5	20.4	37.0
Cyclopentene	44.24	12.50	5.76	6.16	1.6	5.9	5.5

1) Paraffin wax (*mp* 43.5 °C). 2) Low octane gasoline (high severity cracking). 3) Sulfur-containing petroleum (high severity cracking). 4) Naphtha. 5) Romaskino straight-run gasoline (boiling range ≈ 40 – 160/180 °C). 6) Synthetic gasoline from Fischer – Tropsch synthesis (boiling range ≈40 – 150 °C).

(temperature and residence time). Nevertheless, this overall decrease is accompanied by an increase in isoprene concentration in the C_5 crack fraction (cf. Table 3). Table 3 also provides the product distribution in a typical methylbutene dehydrogenation fraction. Based on energy consumption, recovery of isoprene from crack fractions is considerably more efficient than synthesizing the compound chemically [101].

Increasingly heavier raw materials will be used in Europe and the United States for ethylene production in crackers [100]. The quantity of the isoprene byproduct will therefore also increase. In some crackers only an aromatic fraction is recovered from crack gasoline. The residue (i.e., the C_5 hydrocarbons) is often combined with the C_4 fraction and recycled with the cracker feed.

Pure isoprene cannot be isolated from the C_5 crack fraction simply by distillation, because the numerous components of the mixture differ only slightly in boiling point (cf. Table 2), and isoprene forms an azeotropic mixture with *n*-pentane (cf. Table 1). For the separation of cyclopentadiene from the C_5 crack fraction by dimerization, see → Cyclopentadiene and Cyclopentene, Chap. 1.3.

Distillative enrichment of isoprene in C_5 fractions to a concentration of 25 – 50 % is of some commercial interest because it reduces the transportation and separation costs associated with the isolation of pure isoprene. Examples of special isoprene enrichment processes are described in [103].

Goodyear has developed a process for separating isoprene from C_5 crack fractions in the form of its azeotrope with *n*-pentane [104], [105]. This procedure was carried to production scale in France, although that facility is no longer in operation. The resulting mixture is free of components that might adversely affect the polymerization of isoprene, and is thus suitable for polymerization processes that employ a solution of isoprene in an inert hydrocarbon.

"Chemical" isolation procedures involving copper(I) compounds have also been considered for the recovery of isoprene from C_5 hydrocarbon mixtures [106]. In contrast to the situation with butadiene (→ Butadiene) this process has never been adopted commercially for isoprene.

Selective organic solvents are used in "physical" processes for separating isoprene by extractive distillation with or without liquid – liquid extraction. The principles involved are similar to

Table 3. Typical composition (in weight percent) of C_5 crack fractions as a function of cracking severity

Component	Crack severity			
	Mild	Normal	Severe	Methylbutene dehydrogenation fractions
Pentanes	26.70	34.36	14.12	65 – 75
n-Pentenes	5.15	6.60	4.21	65 – 75
Methylbutenes	6.95	9.17	5.76	65 – 75
Pentadienes	10.90	9.40	19.46	traces
Isoprene	14.20	18.00	23.25	25 – 35
Cyclopentane and cyclopentene	4.40	3.30	5.76	traces
Cyclopentadiene (including dimer)	16.70	15.97	24.91	traces
2-Butyne	0.40	0.4	1.03	traces
Other hydrocarbons	balance	balance	balance	

Table 4. Relative volatilities ∗ of hydrocarbons to be separated from isoprene in various solvents, based on [112]

Hydrocarbon	*bp*, °C	Isoprene solvent ∗∗						
		ACN	DMF	NMP	MOPN	DMAC	DMSO	γ-BL
Isopentane	27.9	2.92	3.02	3.00	2.62	2.77	2.72	2.79
n-Pentane	36.1	2.37	2.39	2.40	2.27	2.20	2.07	2.13
3-Methyl-1-butene	20.1	2.58	2.63	2.65	2.49	2.53	2.51	2.56
1-Pentene	30.0	1.89	1.90	1.88	1.78	1.82	1.73	1.75
2-Methyl-1-butene	31.2	1.71	1.73	1.75	1.62	1.67	1.62	1.61
trans-2-Pentene	36.3	1.56	1.55	1.56	1.46	1.50	1.45	1.45
cis-2-Pentene	36.9	1.49	1.54	1.53	1.35	1.47	1.41	1.40
2-Methyl-2-butene	38.1	1.38	1.39	1.40	1.32	1.34	1.29	1.28
Isoprene	34.1	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
trans-1,3-Pentadiene	42.0	0.77	0.76	0.78	0.77	0.77	0.77	0.77
cis-1,3-Pentadiene	44.2	0.70	0.70	0.71	0.70	0.72	0.70	0.70
Cyclopentadiene	41.3	0.62	0.62	0.62	0.63	0.63	0.62	0.63
2-Butyne	26.7	0.96	0.90	0.997	0.97	0.99	1.08	1.08
3-Methyl-1-butyne	26.3	1.04	0.89	0.95	1.10	0.93	0.95	0.94
2-Methyl-1-buten-3-yne	32.5	0.62	0.49	0.48	0.67	0.51	0.55	0.53
1-Pentyne	40.2	0.58	0.55	0.54	0.64	0.54	0.55	0.56
2-Pentyne	56.1	0.42	0.44	0.67	0.43	0.46	0.40	0.41

∗ The figures indicate the ratio of the volatility of the specified hydrocarbon to that of isoprene measured at 50 °C in a mixture containing 12.5 wt % of the hydrocarbon, 12.5 wt % of isoprene, and 75 wt % of solvent.

∗∗ ACN = acetonitrile; DMF = dimethylformamide; NMP = *N*-methylpyrrolidone; MOPN = β-methoxypropionitrile; DMAC = dimethylacetamide; DMSO = dimethyl sulfoxide; γ-BL = γ-butyrolactone

those developed for separation of butadiene from C_4 crack fractions (→ Butadiene, Chap. 4.2.). The most appropriate solvents for production-scale separation of isoprene from other hydrocarbons include *N*-methylpyrrolidone [107], [108], dimethylformamide [109], [110], and acetonitrile [111]. The relative volatilities of the hydrocarbons in various polar solvents (i.e., the selectivities of the solvents for the isolation of isoprene from C_5 hydrocarbon mixtures) are listed in Table 4. Figure 4 shows the relationship between the overall concentration and the relative volatilities of 2-methyl-2-butene – isoprene mixtures in various solvents; these data are of considerable importance for estimating the size of a proposed production facility. 2-Methyl-2-butene is the key low-boiling component under the conditions prevailing in this extractive distillation procedure. Additional data relevant to selective solvents for separation of isoprene are presented in [113]. Isolation of isoprene concentrates from C_5 hydrocarbon mixtures by extrac-

tion with selective organic solvents is discussed in [114].

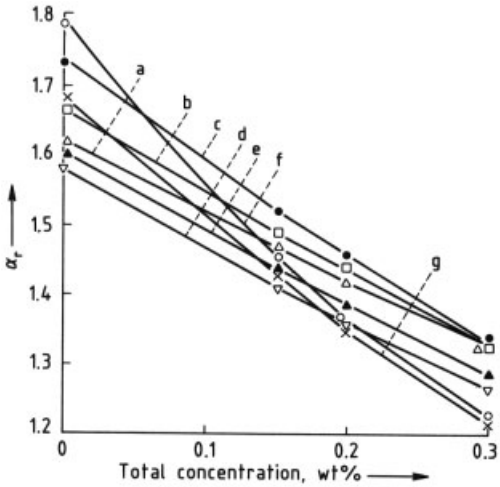

Figure 4. Relationship between the total concentration and the relative volatilities α_r of 2-methyl-2-butene – isoprene mixtures in various solvents [112]
a) Acetonitrile; b) Dimethylformamide; c) *N*-Methylpyrrolidone; d) β-Methoxypropionitrile; e) Dimethylacetamide; f) Dimethyl sulfoxide; g) γ-Butyrolactone

Figure 5 is a flow diagram for the isolation of isoprene from a C_5 cracking fraction. The first step in the separation is a combination of a liquid – liquid extraction with an extractive distillation, an approach particularly suited to C_5 fractions with a low isoprene concentration. Preconcentration of isoprene occurs in extractor (a). Crude isoprene is removed as a sidestream from the extractive distillation column (b), after which it is freed from piperylenes and cyclopentadiene in a second extractive distillation step incorporating columns (c) and (d). Final purification of isoprene to polymerization grade occurs in distillation columns (e) and (f) [115]. C_5 fractions that contain cyclopentadiene can be used in this process because the cyclopentadiene is removed together with the piperylenes.

Incorporation of relevant equipment in this process allows the isolation of cyclopentadiene or the commercially valuable cyclopentene (→ Cyclopentadiene and Cyclopentene, Chap. 2.3.). Introduction of a supplementary column permits recovery of pure piperylenes.

The high quality specifications that are required for isoprene used in the preparation of poly(*cis*-1,4-isoprene) are met with the aid of modern separation facilities. It should be noted, however, that suggestions have also been made for the chemical removal of specific impurities (e.g., cyclopentadiene, acetylene) [116].

Economic comparison of the various techniques for separating isoprene from C_5 cracking fractions is difficult because a large number of criteria must be considered and each must be assigned a proper weighting which depends on the location of the plant. Various selective solvents are compared in [112] with respect to their suitability for isoprene separation. Energy consumption data are analyzed in [117]. A detailed cost analysis for one specific solvent is provided in [118].

4. Quality Specifications

Isoprene is used primarily for the synthesis of polymers, including poly(*cis*-1,4-isoprene) and block polymers containing styrene (SIS polymers, see Chap. 6) [119].

Table 5 provides typical specifications for isoprene used as a monomer for the production of poly(*cis*-1,4-isoprene). Purity specifications for polymerization with Ziegler catalysts are less rigorous than for polymerization with alkyl lithium catalysts. The latter are also used for preparing SIS polymers.

Table 5. Monomer specifications for isoprene used in the production of poly(*cis*-1,4-isoprene) [120]

Component	Alkyl lithium catalyst	Ziegler catalyst (Al/Ti)
Isoprene, wt %	99.5	99.5
Monoolefins, wt %	0.4	0.5
1,3-Pentadiene, ppm	<80	<200
Cyclopentadiene, ppm	< 1	< 1
Acetylene, ppm	< 3	< 50
Allene, ppm	< 5	< 50
Carbonyl compounds, ppm	< 5	< 10
Sulfur (as H_2S), ppm	< 5	< 5
Water, ppm	< 5	< 5

Isoprene for use as a monomer in the manufacture of butyl rubber (isobutene – isoprene rubber, IIR) may contain higher levels of the common impurities, as indicated in Table 6. Table 6 also contains information about technical-grade isoprene.

In most polymerization processes, isoprene is used in dilution with other inert hydrocarbons.

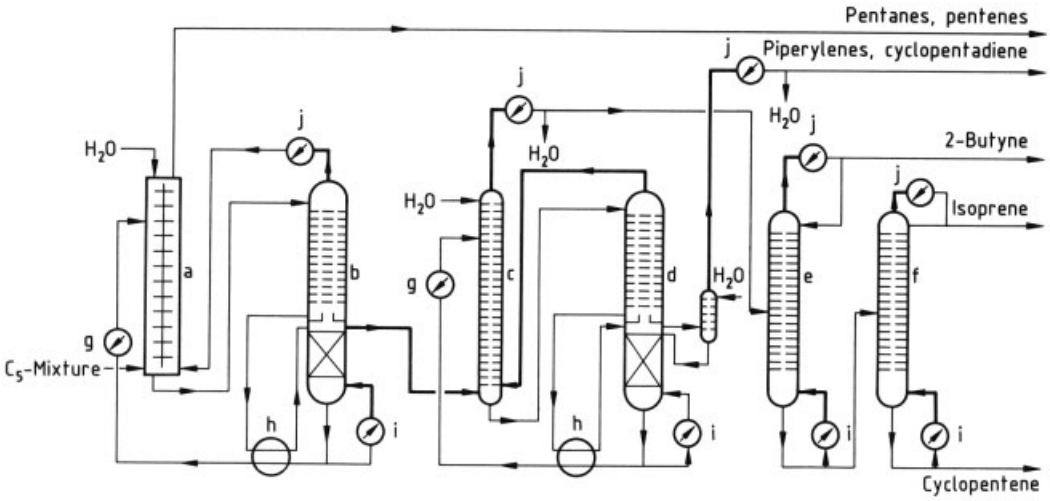

Figure 5. Recovery of isoprene from C_5 hydrocarbon mixtures by a combination of extraction and extractive distillation [107]
a) Liquid – liquid extraction column; b) – d) Columns for extractive distillation; e) Butyne column; f) Isoprene column; g) Cooler; h) Heat exchanger; i) Reboiler; j) Condenser
—— Vapor phase; —— Liquid phase

In the production of IR and SIS polymers, it may therefore be more economical to use an isoprene product (isoprene concentrate) in which isoprene of the required purity (content of cyclopentadiene, acetylene, etc.) is diluted with hydrocarbons that do not interfere with polymerization. Thus, in the synthesis of isoprene rubber a starting material containing 60 – 85 % isoprene is sufficient, whereas an isoprene content of 80 – 90 % is required for SIS polymers. For the synthesis of butyl rubber, however, the isoprene cannot be diluted with other hydrocarbons.

Table 6. Specifications for technical-grade isoprene [121] and isoprene intended for use in the production of butyl rubber (IIR) [122]

Component	Technical-grade isoprene	Isoprene for IIR
Isoprene, wt %	99	>92
Diolefins (total), wt %		>95
C_3 olefins, wt %	1	
Cyclopentadiene, wt %	0.05	<1
α-Acetylene, wt %		<0.35
Peroxides (as H_2O_2), ppm	100	<10
Carbonyl compounds (as acetone), ppm	150	<500
Sulfur (as H_2S), ppm	5	<100
Distillation residue, wt %		<1.5

The purity of isoprene is normally verified by gas chromatography. Data for sulfur, peroxides, carbonyls, distillation residues, and inhibitors are determined by the usual standard analytical methods, preferably those based on ASTM or ISO/DIN recommendations.

5. Storage and Transportation

Table 7. Dimerization of isoprene as a function of temperature [123]

Temperature, °C	Dimerized isoprene, wt %/h
20	0.000017
40	0.00019
60	0.0021
80	0.023
100	0.25

Appropriate regulations must be observed during construction and operation of tank storage facilities for isoprene (e.g., the VbF rules in the Federal Republic of Germany). Reference [123] should be consulted regarding special problems, such as prevention of the formation of peroxides (→ Butadiene, Chap. 5.). Other conditions being equal, the rate of reaction of oxygen with isoprene is significantly greater than that with butadiene. Data regarding the rate of thermal dimerization of isoprene are given in Table 7; this reaction cannot be suppressed by

inhibitors. Thermal dimerization at low temperature is a Diels – Alder reaction, producing only cyclohexene derivatives, while at higher temperatures cyclooctadienes are also formed [124]. Transport regulations are summarized in Table 8 [125].

Table 8. Guidelines for the labeling and transport of isoprene

Means of transport	Region	Guideline
Road	Europe	ADR class 3, no. 2 a
	Germany	GGVS class 3, no. 2 a
	USA	DOT regulations
Rail	Europe	RID class 3, no. 2 a
	Germany	GGVE class 3, no. 2 a
	USA	DOT regulations
Inland shipping	Europe	cat. K 1 n
	Germany	ADNR class 3, no. 1 a
Sea freight	Germany	GGV See class 3.1, UN no. 1218
	International	IMDG Code class 3.1, UN no. 1218
Air freight	International	IATA class 3, UN no. 1218

6. Uses

6.1. Polymer Synthesis

Most of the isoprene that is produced is utilized for the synthesis of poly(*cis*-1,4-isoprene) (isoprene rubber, IR), a material (particularly the titanium type) that closely resembles natural rubber in both structure and properties (→ Rubber,3. Synthetic). This IR is largely used for the production of vehicle tires (→ Tires). By contrast, poly(*trans*-1,4-isoprene) has the properties of guttapercha or balata, and has found no major commercial application. Most of the trans polymer is limited to the manufacture of cable insulation and golf balls [124, p. 499], [126].

The second largest market for isoprene is the manufacture of styrene – isoprene – styrene (SIS) block copolymers (→ Thermoplastic Elastomers) [119], [120, p. 837], [124, p. 532], [127]; 0.79 – 0.88 t of isoprene are required per tonne of polymer. These products are especially useful as thermoplastic rubbers and as pressure-sensitive or thermosetting adhesives.

Smaller amounts of isoprene are used in the production of butyl rubber (isobutene – isoprene rubber, IIR), a copolymerizate with isobutene (→ Rubber,3. Synthetic) [128]. The isoprene content of butyl rubber is 0.5 – 3.0 mol%. The distinctive features of IIR include its low gas permeability, leading to its use in the construction of hoses and as a liner in tubeless tires.

Finally, mention should also be made of the hydrocarbon resins (petroleum resins), which result from the copolymerization of isoprene from cyclopentadiene-free C_5 crack fractions with other unsaturated C_5 compounds (→ Resins, Synthetic) [129].

6.2. Terpene Synthesis

The chemical reactions of isoprene are very varied [19], but they have limited industrial significance. Only terpene synthesis is important and is the subject of intensive investigation.

C_5+C_3+C_2 Routes. In 1972, Rhodia Incorporated (United States) began to develop a route for the industrial synthesis of C_{10} terpenes and their derivatives from isoprene (C_5), acetone (C_3), and acetylene (C_2) (Fig. 6). The isoprene is first reacted with hydrochloric acid to give prenyl chloride which is then converted to dehydrolinalool in two steps [7], [130]. Rhodia used the dehydrolinalool as a starting compound for the industrial synthesis of most of the compounds shown in Figure 6, but the plant was decommissioned. Kuraray in Japan, however, produces the compounds shown in Figure 6, squalane, and other substances from isoprene [131].

C_5+C_5 Routes. Rhodia has also used a *Grignard synthesis* to produce the C_{10} terpene alcohol lavandulol [*1845-51-8*] from two molecules of isoprene [132].

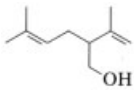

Lavandulol

Isoprene can react in a variety of ways to form terpenes via oligomerization or telomerization (→ Terpenes) [133–137]. The syntheses of myrcene and *N*,*N*-diethylnerylamine (C_{10} terpenes) are described below as examples of

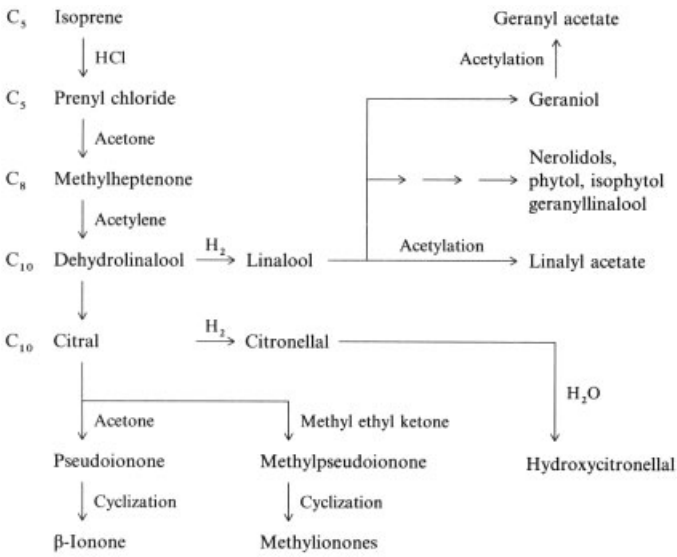

Figure 6. Synthesis of terpenes from isoprene via the $C_5+C_3+C_2$ route

scientifically and industrially important linear oligomerization and telomerization reactions, respectively.

All naturally occurring terpenes are composed of isoprene units arranged in a "head-to-tail" fashion [133–137]. Terpenes synthesized from isoprene must also possess this structure and their double bonds must be correctly positioned [138], [139]. On account of this problem C_{10} terpenes and their derivatives are not generally synthesized from isoprene but by other routes [47], [138].

Variation of the quantitative composition of a Pd-complex catalyst can influence the structure of dimethyloctadienes synthesized from isoprene [140].

Myrcene [*123-35-3*] has been synthesized by *oligomerization* of isoprene with a sodium/dialkylamine catalyst by TAKABE et al. [141] (see also [142–144]).

Myrcene

Nissan Chemical Industries have developed a route for the industrial production of myrcene and have announced that it will be implemented industrially [145].

Telomerization of isoprene proceeds with a variety of compounds, e.g., ammonia or amines [146] (see also [135], [136]). Interestingly, reactions occur in which the regioselectivity can be highly modified by varying the amount and concentration of Brönsted and Lewis acids added to the catalyst [147], [148].

Isoprene can react with diethylamine to form *N,N*-diethylnerylamine [*40137-00-6*] [149]; this telomerization reaction is catalyzed by butyl lithium. *N,N*-Diethylnerylamine can be further reacted to give linalool, geraniol, nerol, hydroxycitronellal, citronellal, and menthol (Fig. 7) [149–152].

The Takasago Perfumery Company in Japan has developed an industrial synthesis for enantiomerically pure L-menthol [*2216-51-5*] from optically active citronellal. The reaction involves the asymmetric allylamine – enamine isomerization of *N,N*-diethylgeranylamine or *N,N*-diethylnerylamine with a Rh – BINAP [2,2′-bis(diphenylphosphino)-1,1′-binaphthyl] catalyst [153–156]. Geranylamine is obtained by addition of diethylamine to myrcene [22]. The synthesis of nerylamine is described above [149]. Takasago produces 1500 t/a of L-menthol by this route.

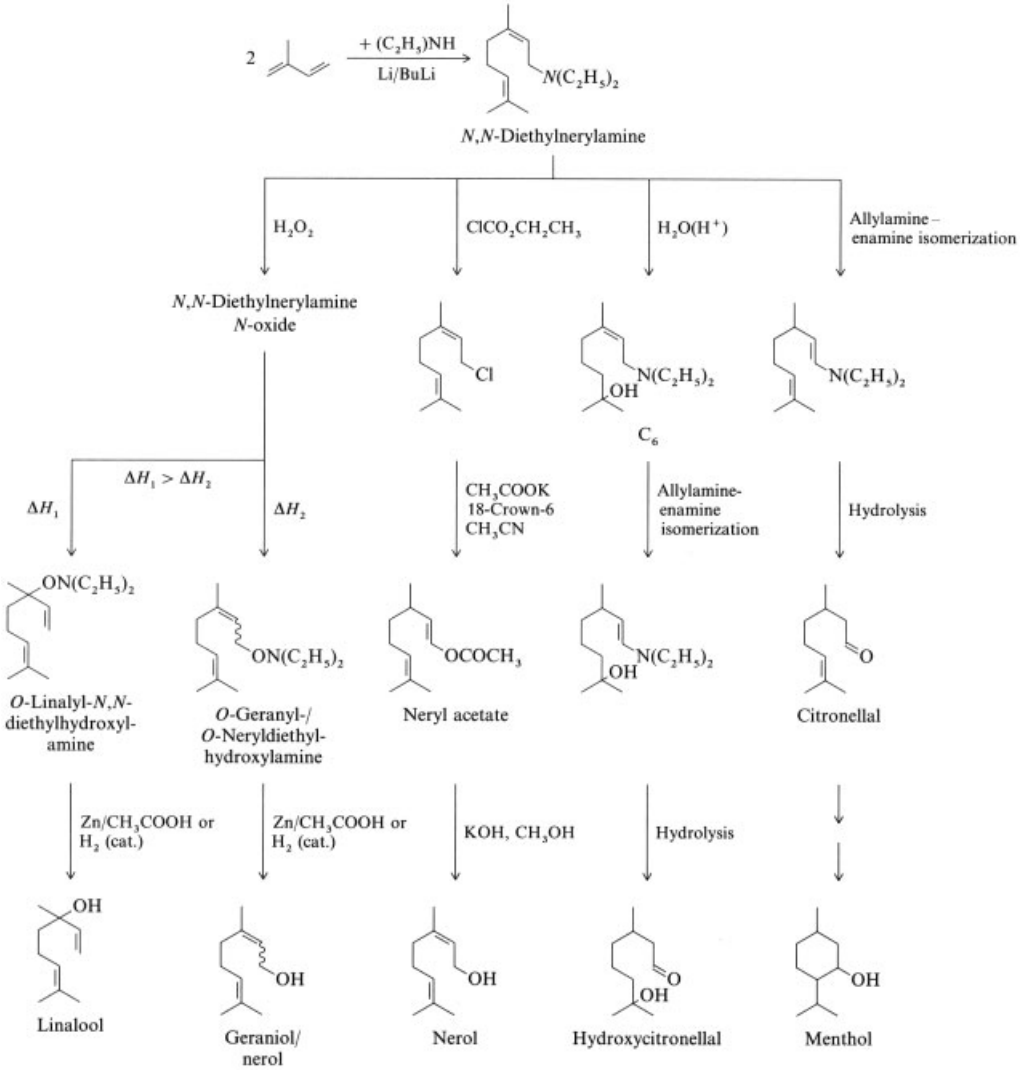

Figure 7. Synthesis of terpenes from isoprene via nerylamine

7. Economic Aspects

Isoprene is used mainly as a starting material for the manufacture of synthetic rubber, the market for the compound therefore depends heavily upon the widely varying price of natural rubber (cf. [157] and corresponding publications from previous years). As noted in Section 3.3, extraction from C_5 crack fractions is the most economical route to isoprene [101]. Nevertheless, a key factor in the calculation is the location of the separation facility, since separation is only profitable if C_5 fractions (possibly enriched in isoprene) from several cracking plants can be transported economically to a separator of sufficient capacity [158]. Utilization of the large quantities of residual C_5 hydrocarbons is also a major consideration.

The total production of C_5 diolefins in Western Europe in 1987 was 83 000 t [159]. It is of some interest that more (dimeric) cyclopentadiene was produced (44 000 t) than isoprene (23 000 t). The remainder (15 000 t) was accounted for by piperylenes. As shown in Figure 8, an increasing demand for cyclopentadiene relative to isoprene is also anticipated in the

United States in the coming years. The growing interest in cyclopentadiene may be attributed to the latest developments in the field of metathetic polymerization of the compound [161]. Combining this technology with the reaction injection molding method opens the way to very economical injection castings with good mechanical properties.

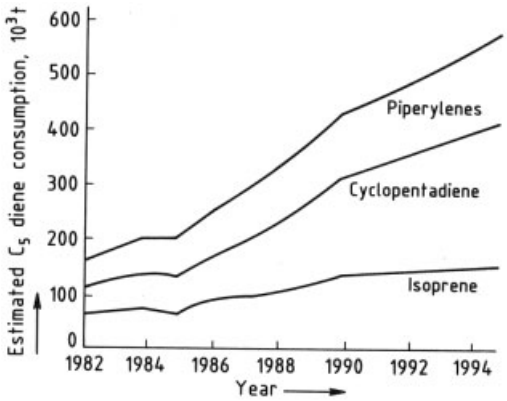

Figure 8. Estimated demand for C_5 dienes in the United States [160]

Table 9 provides an overview of the known production facilities for pure isoprene and isoprene concentrates. The isoprene concentrates are used either directly or, if necessary, after additional purification. Table 10 summarizes production capacities and actual production of polyisoprene (IR), butyl rubber (IIR), and synthetic rubber for the years 1987 and 1989, subdivided according to country. Corresponding totals for the period 1979 – 1989 are given in Table 11. The obvious decrease in the production of polyisoprene is largely a result of declines in the United States. Consumption data for isoprene used in the manufacture of SIS copolymers in 1986 follow [119]:

United States 25 000 t
Western Europe18 000 t
Japan 250 t

8. Toxicity and Occupational Health

Acute Toxicity. At high levels of exposure isoprene is an anesthetic in animals, finally resulting in paralysis and death. Single 2-h inhalation exposures of mice to 56 000 mg isoprene per cubic meter of air did not produce toxic effects. Levels of 98 000 – 126 000 mg/m^3 resulted in deep narcosis; death occurred after 2 h of exposure to 140 000 mg/m^3. The LC_{50} for a 2-h exposure was 180 000 mg/m^3 in the rat. The threshold concentration for irritative effects in cats is reported as 800 mg/m^3. An oral LD_{50} of 2100 mg liquid isoprene per kilogram of body weight and an intraperitoneal LD_{50} of 1400 mg/kg in the male rat have been determined. A single dermal exposure of rats to a 1-mL dose of liquid isoprene per kilogram body weight did not cause adverse symptoms or mortality [164], [165].

Toxicity after Repeated Exposure. Fischer 344 rats and B6C3F1 mice were exposed by inhalation for two weeks to atmospheric isoprene concentrations of 0, 438, 875, 1750, 3500, and 7000 ppm. The rats did not show any exposure-related changes (survival, clinical symptoms, body weight, biochemistry, hematology, macro- and micropathology). Mice at the 7000 ppm level only showed lower body weight gain, but the mice in all groups suffered from anemia, testicular atrophy, olfactory epithelial degeneration, and epithelial hyperplastic changes in the stomach. As in the case of 1,3-butadiene, mice therefore seem to be more susceptible than rats [166].

Repeated dermal exposure (2 applications of 500 µL on each of 5 consecutive days) to the ear of the rabbit caused only limited, reversible irritation [165]. No data on long-term inhalation exposure are available.

Mutagenicity. Isoprene and its monoepoxides do not induce point mutations in the Ames test. The dioxide of isoprene (2-methyl-1,2,3,4-diepoxybutane) was found to be mutagenic in this test system [167–170]. Due to some similarities with 1,3-butadiene in the formation of mutagenic metabolites the possible carcinogenic effects of isoprene need further evaluation.

Inhalative exposure of B6C3F1 mice to isoprene levels from 438 to 7000 ppm in air for 6 h/d on 12 days resulted in elevated frequencies of sister chromatid exchange indicating cytogenetic effects [171].

Reproduction. No data on reproductive or teratogenic effects are available.

Table 9. Production facilities for isoprene

Country	Company	Location	Capacity as 100 wt % isoprene, t/a [a]	Product [b]	Process [c]
United States	Goodyear	Beaumont, Texas	84 000	P	refining of isoprene concentrate (extractive distillation?)
	Arco Lyondell	Channelview, Texas	36 000	C	extractive distillation (ACN) of C_5 fractions
	Chevron	Cedar Bayou, Texas	9 000	C	extractive distillation (?)
	Dow	Freeport, Texas	16 000	C	extractive distillation (?)
		Plaquemine, Louisiana	16 000	C	extractive distillation (?)
	Exxon	Baton Rouge, Louisiana	14 000	C	extractive distillation (ACN) of C_5 fractions
		Baytown, Texas	14 000	C	extractive distillation (ACN) of C_5 fractions
	Shell	Dear Park, Texas	40 000	C	extractive distillation (ACN) of C_5 fractions
Netherlands	Shell	Pernis	25 000	C	extractive distillation (ACN) of C_5 fractions
Republic of South Africa	Sentrachem (Karbochem)	New Castle/ Natal	45 000	P	synthesis from acetone and acetylene, plant not yet on line
Japan	Nippon Zeon	Mizushima	45 000	P	extractive distillation (DMF) of C_5 fractions
	Japan Synthetic Rubber	Kashima	30 000	P	extractive distillation (ACN) of C_5 fractions
	Kuraray	Kashima	30 000	P	synthesis from isobutene and formaldehyde
former Soviet Union		several locations	ca. 800 000	C	dehydrogenation of isopentane extractive distillation of C_5 fractions synthesis from isobutene + formaldehyde
Standby Capacities					
United States	Shell	Port Neches, Texas	45 000	P	dehydrogenation of isopentene (plant bought from BF Goodrich in 1982 and modified for production of pure isoprene)
Netherlands	Shell	Pernis	50 000	C	dehydrogenation of isopentene
Italy	Enichem (ANIC)	Ravenna	30 000	P	synthesis from acetone and acetylene

[a] Capacities in 1987.
[b] C = concentrate; P = pure isoprene.
[c] ACN = acetonitrile; DMF = dimethylformamide.

Metabolism. Rodent mitochondrial fractions transformed isoprene to its monoepoxides 3,4-epoxy-3-methyl-1-butene and 3,4-epoxy-2-methyl-1-butene, leading finally to the corresponding diols. Minor oxidation of the more stable metabolite (3,4-epoxy-2-methyl-1-butene) to a diepoxide has been described [172], [173]. Isoprene is metabolized significantly in the respiratory tract. Up to atmospheric isoprene concentrations of about 300 ppm, metabolism is directly proportional to the exposure concentration [174], [175]; a saturation effect occurs at levels above 300 – 500 ppm. Mice metabolize isoprene at a higher rate than rats [174].

Isoprene is synthesized endogenously. The synthesis rate in unexposed mice and rats has been calculated to be $0.4\,\mu mol\,h^{-1}\,kg^{-1}$ and $1.9\,\mu mol\,h^{-1}\,kg^{-1}$, respectively [174].

Effects in Humans. Isoprene is narcotic in humans at very high exposure levels. It may also cause irritation of the skin, eye, mucous membranes, and respiratory tract.

Hygienic Standards. No TLV or MAK values have been established for isoprene. A threshold of 40 mg/m^3 has been set in the former Soviet Union.

Table 10. Production capacities (in 10^3 t/a) for polyisoprene (IR), butyl rubber (IIR), and total synthetic rubber [162] ∗

Country	IR		IIR		Synthetic rubber	
	1987	1989	1987	1989	1987	1989
Belgium			85	85	150	150
Germany (FRG)					591	592
France			48	48	719	719
Great Britain			60	60	485	485
Netherlands		40			719	719
	40					
Western Europe	*40*	*40*	*193*	*193*	*2970*	*3044*
Canada			120	120	284	284
USA			219	219	2 733	2 796
	68	68				
North America	*68*	*68*	*339*	*339*	*3 017*	*3 080*
Brazil					350	427
		13				
Latin America		*13*			*594*	*689*
Japan			75	75	1 587	1 627
	67	67				
Republic of South Africa					136	136
	45	45				
Countries with market economics			607	607	8 844	9 195
	220	233				
Countries with centrally planned economics	1 045	1 045	120	210	3 579	3 860
World Total	**1 265**	**1 278**	**727**	**817**	**12 423**	**13 055**

∗ Year-end data, 1989 being based on available projections.

Table 11. Production and production capacity (in 10^3 t) for polyisoprene (IR), butyl rubber (IIR), and total synthetic rubber [163] ∗

		1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
IR	capacity	281	289	289	297	397	289	285	255	225	220	213
IR	production	242	219	168	129	120	146	110	132	113		
IIR	capacity	493	498	446	533	539	583	580	583	585	607	615
IIR	production	457	395	426	370	372						
Synthetic rubber	capacity	7975	8144	8379	8878	8944	8710	8639	8772	8730	8844	9108
Synthetic rubber	production	5830	6167	5347	5281	4623	4854	5397	5211	5486		

∗ Excluding countries with centrally planned economies.

9. References

1. W. C. Meuly, *Am. Perfum. Cosmet.* **85** (1970) 123.
2. Y. Yokouchi, Y. Ambe: *J. Geophys. Res. D. Atmos.* **93** (1984) no. 04, 3751; *Chem. Abstr.* **109** (1988) 26 887 p.
3. T. Dalhamn, R. Rylander, *Arch. Environ. Health* **20** (1970). J. L. Egle, Jr., B. J. Gochberg, *Am. Ind. Hyg. Assoc. J.* 1975 (May) 369.
4. E. Gil-Av, J. Shabtai, *Nature* **197** (1963) 1065.
5. L. Ruzicka, *Experientia* **9** (1953) 357.
6. F. Lynen, U. Henning, *Angew. Chem.* **72** (1960) 820.
7. C. G. Williams, *Proc. Roy. Soc.* **10** (1860) 516; *Philos. Trans. R. Soc. London* 1860, 241.
8. F. Hofmann, *Chem. Ztg.* **60** (1936) 693.
9. Goodrich Gulf, US 3 114 734, 1954 (S. E. Horne).
10. Firestone Tire and Rubber Comp., US 3 208 988, 1955 (L. E. Forman et al.).
11. C. Harries, K. Gottlob, *Justus Liebigs Ann. Chem.* **383** (1911) 228.
12. Newsport Ind., US 2 386 537, 1945 (C. H. Bibb).
13. K. H. Simmrock, R. Janowsky, A. Ohnsorge: "Critical Data of Pure Substances" in D. Behrens, R. Eckermann (eds.): *Chemistry*

Data Series, DECHEMA, Frankfurt 1986, vol. II/1, p. 289.
14. T. Boublik, V. Fried, E. Hála: *The Vapour Pressure of Pure Substances,* Elsevier Science Publ., Amsterdam, New York 1984, p. 329.
15. *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds,* American Petroleum Research Project 44, Carnegie Press, Pittsburg 1953, p. 361.
16. *The Aldrich Library of FT-IR Spectra,* I/1 ed., 30 D Milwaukee, Wisconsin 1985.
17. L. C. Jones, L. W. Taylor, *Anal. Chem.* **27** (1955) 228.
18. Sadtler Research Laboratories: *Nuclear Magnetic Resonance Spectra,* Spectrum no. 3434 M.
19. W. J. Bailey in E. C. Leonard (ed.): *Vinyl and Diene Monomers,* **vol. 2,** Wiley, New York 1971, Chap. 5.
20. J. Gmehling, U. Onken, W. Arlt, B. Kolbe: "Vapor – Liquid Equilibrium Data Collection," in D. Behrens, R. Eckermann (eds.): *Chemistry Data Series*, DECHEMA, Frankfurt, vol. I/6 a, 1980, p. 38; vol. I/6 c, 1983, p. 40.
21. *Landolt – Börnstein,* NS IV/3, 1975, 367.
22. J. M. Sørensen, W. Arlt: "Liquid – Liquid Equilibrium Data Collection," in D. Behrens, R. Eckermann (eds.): *Chemistry Data Series,* DECHEMA, Frankfurt vol. V/1, 1979, p. 582; vol. V/3, 1980, p. 472.
23. D. Tiegs et al.: "Activity Coefficients at Infinite Dilution," in D. Behrens, R. Eckermann (eds.): *Chemistry Data Series,* DECHEMA, Frankfurt 1986, vol. IX/1 and 2, p. 880.
24. R. W. Gallant: *Physical Properties of Hydrocarbons,* Gulf Publ. Comp., Houston 1968, 157 – 166.
25. C. L. Yaws, *Chem. Eng. (N.Y.)* **83** (1976) no. 5, 107.
26. J. E. Kilpatrick et al., *J. Res. Nat. Bur. Stand.* **42** (1949) 225.
27. F. M. Fraser, E. J. Prosen, *J. Res. Nat. Bur. Stand.* **54** (1955) 143.
28. T. Reis, *Chem. Process Eng. (London)* **53** (1972) no. 2, 68.
29. C. A. Cociasu, A. Dumitrescu, *Rev. Chim. (Bucharest)* **23** (1972) 652.
30. D. Forster, J. P. Sluka, A. Vavere, *CHEMTECH* **16** (1986) no. 12, 746.
31. I.G. Farbenind., DE 741 152, 1938 (W. Friedrichsen).
32. *Hydrocarbon Process.* **50** (1971) no. 11, 167.
33. F. Coussemant, M. Hellin, *Erdöl Kohle Erdgas Petrochem.* **15** (1962) no. 5, 274, 348. M. Hellin, M. Davidson, *Inf. Chim.* 1980 no. 206, 163, 181.
34. W. Swodenk, W. Schwerdtel, P. Losacker, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **23** (1970) no. 10, 641. Z. Kriefalussy, *Chem. Ing. Tech.* **55** (1983) no. 12, 1965.
35. *Hydrocarbon Process.* **50** (1971) no. 11, 168.
36. K. Naito, K. Ogino, *Pet. Petrochem. Int.* **13** (1973) no. 11, 44.
37. Idemitsu Petrochem., JP 1 90 82, 1965.
38. H. J. Peterson, J. O. Turner, *Hydrocarbon Process.* **53** (1974) no. 7, 121.
39. *Hydrocarbon Process.* **50** (1971) no. 11, 171.
40. D. W. Hall, F. L. Dormish, E. Hurley, Jr., *Ind. Eng. Chem. Prod. Res. Dev.* **9** (1970) no. 2, 234.
41. Sumitomo Chem., DE-OS 2 004 355, 1970 (K. Takagi et al.).
42. Sumitomo Chem., DE-OS 1 816 739, 1968 (Y. Watanabe et al.).
43. Sumitomo Chem., DE-OS 1 941 949, 1969 (Y. Watanabe et al.).
44. A. Mitsutani, *Am. Chem. Soc., Centennial Meeting,* New York 1976.
45. A. Mitsutani, S. Kumano, *Chem. Econ. Eng. Rev.* **3** (1971) no. 3, 35.
46. M. Ai, *J. Catal.* **106** (1987) 280.
47. H. Pommer, A. Nürrenbach, *Pure Appl. Chem.* **43** (1975) 527.
48. BASF, DE 1 275 049, 1967 (H. Müller, H. Overwien, H. Pommer). P. R. Stapp, *Ind. Eng. Chem. Prod. Res. Dev.* **15** (1976) no. 3 189.
49. BASF, DE-OS 2 031 921, 1970 (H. Müller).
50. Monsanto Co., EP 80 449, 1981 (D. Forster, G. E. Barker). D. A. Bolshakov et al., *Neftekhimiya* **23** (1983) no. 2, 183; *Chem. Abstr.* **98** (1983) 217 599 k.
51. Goodyear Tire & Rubber Co., US 4 484 006, 1983 (H. R. Menapace).
52. BASF, EP 58 927, 1982 (F. Merger, H. J. Förster). BASF, EP 92 097, 1983 (G. Dümbgen et al.).
53. A. V. Irodov et al., *2H. Org. Khim.* **18** (1982) no. 7, 1401; *Chem. Abstr.* **97** (1982) 111 060 s.
54. Erdölchemie GmbH, DE 2 163 396, 1971 (H. Fischer, G. Schnüchel).
55. J. B. Moffat, *Rev. Chem. Intermed.* **8** (1987) no. 1, 1.
56. Goodyear Tire & Rubber Co., US 4 524 233, 1984 (W.-L. Hsu et al.); US 4 587 372, 1985 (W.-L. Hsu); US 4 628 140, 1986 (L. G. Wideman).
57. Monsanto Co., US 4 547 614, 1988 (A. Vavere).

58. International Synthetic Rubber Co., GB-A 2 093 060, 1981 (D. E. Timm).
59. Union Carbide, US 4 734 538, 1986 (G. L. O'Connor, St. W. Kaiser, J. M. McCain).
60. Union Carbide, EP-A 272 662, 1987 (J. H. McCain, St. W. Kaiser, G. L. O'Connor). BASF, EP 162 385, 1984 (W. Hölderich et al.). W. Hölderich, *Pure Appl. Chem.* **58** (1986) no. 10, 1383.
61. BASF, EP-A 219 042, 1985 (W. Hölderich, P. Hettinger, F. Merger).
62. D. A. Bolshakov et al. *Prom-St. Sint. Kauch.* **1980** no. 8, 2; *Chem. Abstr.* **94** (1981) 84549 u.
63. H. Ando et al., *Sekiyu Gakkaishi* 1969, 356, 362.
64. C. A. Cocoiasu, *Rev. Chim. (Bucharest)* **23** (1972) 722.
65. A. Fovorski, *Zh. Russ. Fiz.-Khim. O-va.* **32** (1902) 356, 652.
66. Bayer, DE 246 241, 1910.
67. *Hydrocarbon Process.* **54** (1975) no. 11, 154.
68. A. Heath, *Chem. Eng. (N.Y.)* **80** (1973) no. 22, 448.
69. R. Landau, G. S. Schaffel, A. C. Deprez, *Erdöl,* Kohle, Erdgas, Petrochem. 16 (1963) no. 7, 754.
70. F. Andreas, K. Gröbe: *Propylenchemie,* Akademie-Verlag, Berlin 1969, Chap. 14.
71. *Chem. Eng. News* **48** (1970) no. 10, 60.
72. *Oil Gas J.* **69** (1971) no. 16, 17.
73. R. L. Banks, R. B. Regier, *Ind. Eng. Chem. Prod. Res. Dev.* **10** (1971) no. 1, 46.
74. R. L. Banks in B. E. Leach (ed.): *Appl. Industr. Catalysis* **2,** Academic Press, New York 1984, p. 218.
75. S. Warwel, G. Pütz, *Chem. Ztg.* **112** (1988) no. 11, 15.
76. K. L. Anderson, T. D. Brown, *Hydrocarbon Process.* **55** (1976) no. 8, 119.
77. *Eur. Chem. News* **18** (1970) no. 452, 30; no. 453, 36.
78. T. Reis, *Chem. Process Eng. (London)* **53** (1972) no. 3, 66.
79. A. A. Digiacomo, J. B. Maerker, J. W. Schall, *Chem. Eng. Prog.* **57** (1961) no. 5, 35.
80. R. G. Craig, J. M. Duffalo, *Chem. Eng. Prog.* **75** (1979) no. 2, 62.
81. *Hydrocarbon Process.* **65** (1986) no. 4, 106.
82. I. V. Garmonov, paper no. 1 presented at International Symposium on Isoprene Rubber, Moscow 1972.
83. Y. A. Gorin, A. A. Wasilew, A. N. Makashina, *Khim. Promst. (Moskau)* 1958, no. 1, 1.
84. R. L. Foster, D. K. Wunderlich, S. H. Patinkin, R. A. Sanford, *Pet. Refin.* **39** (1960) no. 11, 229.
85. *Pet. Refin.* **40** (1961) no. 11, 257.
86. *Hydrocarbon Process.* **52** (1973) no. 11, 140.
87. R. A. Stowe, J. L. Mallory, I. J. Martin, J. V. Bishop, *Prepr. Div. Pet. Chem. Am. Chem. Soc.* **17** (1972) B 19; *Erdöl, Kohle, Erdgas, Petrochem. Brennst. Chem.* **27** (1974) 94.
88. O. D. Sterligov, I. M. Arinich, T. V. Dudukina, *Izv. Akad. Nauk SSSR, Ser. Khim.* 1971, 2833.
89. Petro-Tex, US 4 658 074, 1965 (L. Bajars).
90. Phillips Petroleum, US 3 851 009, 1971 (R. S. Cichowski).
91. A. Farkas, *Hydrocarbon Process.* **49** (1970) no. 7, 121.
92. E. Nishikawa, T. Ueki, Y. Morita, *Bull. Jpn. Pet. Inst.* **12** (1970) 84.
93. Petrotex, US 3 278 626, 1962 (L. Bajars).
94. H. Pichler, H. Schulz, A. Zeinel Deen, *Erdöl, Kohle, Erdgas, Petrochem. Brennst. Chem.* **29** (1976) no. 9, 418.
95. M. Sittig: *Diolefins Manufacture and Derivatives,* Noeyes Dev. Corp., Park Ridge, N.J. 1968, p. 2.
96. S. M. Zulfugarova, D. B. Tagiev, *Neftekhimiya* **26** (1986) no. 2, 189; *Chem. Abstr.* **106** (1987) 101 699 x.
97. K. E. Tokhadze et al., *Neftekhimiya* **23** (1983) no. 9, 511; *Chem. Abstr.* **99** (1983) 174 240 y.
98. Halcon, DE-AS 1 568 764, 1966 (J. Kollar).
99. D. L. Schultz, *Rubber World* **183** (1980) no. 10, 87.
100. G. Zimmermann, D. Zschummel, S. Brändel, *Chem. Tech. (Leipzig)* **21** (1969) 97.
101. M. J. Rhoad: *Proceedings,* 15th Annual Meeting, International Institute of Synthetic Rubber Producers, Inc., Kyoto 1974; *Rubber Ind. (London)* **9** (1975) no. 12, 68; *Rubber News* **14** (Jan. 23, 1978).
102. Y. Wada, *Chem. Econ. Eng. Rev.* **6** (1974) no. 8, 36. H. Kuper, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 3, 119.
103. J. W. Bennet, V. J. Guerico, *Chem. Eng.* (*London*) **1971**, 315. Esso, US 2 768 224, 1953 (L. B. Page, J. W. Rector).
104. *Eur. Chem. News* **25** (1974) no. 642, 12, 27.
105. Goodyear, DE-OS 2 444 232, 1974 (H. M. Lybarger).
106. E. R. Gilliland: "Chemicals for Synthetic Rubbers" in: *The Science of Petroleum,* V/II, Oxford University Press, London 1953, p. 32.
107. H. Kröper, H. M. Weitz, *Oil Gas J.* **65** (1967) no. 2, 98.
108. U. Block, B. Hegner, D. Wolf, *Chem. Ing. Tech.* **47** (1975) 35.

109. R. W. Swanson, J. A. Gerster, *J. Chem. Eng. Data* **7** (1962) 132.
110. S. Takao, *Oil Gas J.* **77** (1979) no. 40, 81.
111. T. Reis, *Chem. Process Eng. (London)* **53** (1972) no. 1, 34.
112. B. A. Sarajew et al., *Khim. Promst. (Moskau)* 1975, 815.
113. M. Enomoto, *Bull. Jpn. Pet. Inst.* **13** (1971) 169.
114. G. H. Dale, W. C. MacCarthy, *World Pet. Congr. Proc. 7th* **4** (1967) 55.
115. R. E. Lynn, J. C. Healy, *Chem. Eng. Prog.* **57** (1961) 46. H. Scherb, *Chem. Ing. Tech.* **44** (1972) no. 11, 704.
116. Jap. Synthetic Rubber, DE 2 105 523, 1970 (T. Chikatsu et al.).
117. M. J. Rhoad, *J. Inst. Rubber Ind.* **9** (1975) 68.
118. S. J. Pawlow et al., *Khim. Promst. (Moskau)* 1971, no. 4, 256.
119. *Isoprene, Chemical Economics Handbook,* SRI International, Menlo Park, Calif. 1988.
120. J. Witte in *Houben-Weyl,* **vol. E 20/2,** 1987, 823.
121. C. E. Morrell in G. S. Whitby: *Synthetic Rubber,* J. Wiley & Sons, New York 1954, p. 84.
122. J. Pech, *Österr. Chem. Ztg.* **68** (1967) 77.
123. *Enjay, Storage and Handling of Liquefied Olefins and Diolefins,* company leaflet, New York 1962.
124. M. L. Senyek in H. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* **vol. 8,** J. Wiley & Sons, New York 1987, p. 494.
125. G. Hommel et al.: *Handbuch der gefährlichen Güter,* 3rd ed., Springer, Berlin – Heidelberg – New York 1987, Merkblatt 359.
126. *Polyisoprene Elastomers, Chemical Economics Handbook,* SRI International, Menlo Park, California 1988.
127. *Styrenic Thermoplastics, Elastomers, Chemical Economics Handbook,* SRI International, Menlo Park, California 1986.
128. E. N. Kresee, R. H. Schatze, H. C. Wang in H. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* **vol. 8,** J. Wiley & Sons, New York 1987, p. 433.
129. Exxon Chemical International Marketing Inc., *Polymers for Adhesives and Sealants,* company leaflet, Kraainem, Belgium 1989.
130. *Chem. Week* **107** (1970) no. 21, 109; *Chem. Mark. Rep.,* 29th Apr. (1974) 7, 18. S. Wiseman, B. T. Bawden, *Chem. Eng. Progr.* 1977, no. 5, 60.
131. *Chem. Age (London)* **111** (1975) no. 2399, 8. *Chem. Week* **121** (1977) no. 3, 17. *Jpn. Chem. Week* (1978) 6th June, 6. *Chem. Ind.* **34** (1982) no. 10, 252.
132. J. Sibeud, *Perfum. Flavor.* **1** (June/July 1978) 7. Rhodia Inc., US 3 819 733, US 3 856 867, 1972 (H. Ramsden).
133. H. Morikawa, S. Kitazome, *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1988) no. 4, 254.
134. A. J. Chalk, S. A. Magennis, *Ann. N.Y. Acad. Sci.* **333** (1980) 386.
135. W. Keim, A. Behr, M. Röper, "Alkene and Alkyne Oligomerization, Cooligomerization and Telomerization Reactions," in G. Wilkinson et al. (eds.): *Comprehensive Organometallic Chemistry,* vol. 8, Pergamon Press, Oxford – New York 1982, pp. 399, 445.
136. A. Behr, *Aspects Homogeneous Catal.* **5** (1984) 48.
137. W. Gaube, R. Berger, *Mitteilungsbl. Chem. Ges. D.D.R.* **34** (1987) no. 6, 122.
138. W. Hoffmann, *Chem. Ztg.* **97** (1973) no. 1, 23.
139. A. D. Josey, *J. Org. Chem.* **39** (1974) no. 2, 139.
140. K. Takahashi, G. Hata, A. Miyake, *Bull. Chem. Soc. Jpn.* **46** (1973) no. 2, 600.
141. K. Takabe et al., *Synthesis* 1977, no. 5, 307.
142. Nissan Chem. Ind., DE-OS 2 542 798, 1974 (K. Takabe et al.).
143. J. Barton et al., Collect. Czech. Chem. Comm. **48** (1983) no. 8, 2361.
144. Takasago Perfumery Co., Jp-KK 81/5206, 1976 (S. Agutagawa).
145. A. Mitsutani, *Hydrocarbon Process.* **58** (1979) no. 6, 56-C. *Chem. Eng.* **85** (1978) no. 2, 144.
146. W. Keim, M. Röper, *J. Org. Chem.* **46** (1981) no. 18, 3702.
147. W. Keim, M. Röper, M. Schieren, *J. Mol. Catal.* **20** (1983) 139.
148. M. Röper, R. He, M. Schieren, *J. Mol. Catal.* **31** (1985) 335.
149. K. Takabe, K. Katagiri, J. Tanaka, *Chem. Lett.* 1977, no. 9, 1025.
150. Nissan Chem. Ind., DE-OS 2 634 309, 1975 (J. Tanaka et al.).
151. Nissan Chem. Ind., DE-OS 2 720 839, 1976 (A. Murata et al.).
152. Takasago Perfumery Co., JA-AS 83/26 735, 1977 (H. Kumobayashi). *Chem. Abstr.* **90** (1979) 23 327 j Takasago Perfumery Co. GB-A 2 088 355, 1980 (S. Akatugawa, T. Taketomi).
153. K. Tani et al., *J. Chem. Soc. Chem. Commun.* 1982, no. 11, 600. K. Tani et al., *J. Am. Chem. Soc.* **106** (1984) no. 18, 5208. H. Takaya et al., *J. Am. Chem. Soc.* **109** (1987) no. 5, 1596.

154. Takasago Perfumery Co., EP-A 0 170 470, 1984 (H. Kumobayashi et al.).
155. G. W. Parshall, W. A. Nugent, *CHEMTECH* **18** (1988) no. 6, 376.
156. R. Noyori, M. Kitamura, "Enantioselective Catalysis with Metal Complexes," in R. Scheffold (ed.): *Modern Synthetic Methods,* vol. 5, Springer-Verlag, Berlin – Heidelberg – New York 1989, p. 115.
157. K. Möbius, *Gummi, Fasern, Kunstst.* **42** (1989) no. 1, 6.
158. M. Rosenzweig, *Chem. Eng. (N.Y.)* **79** (1972) no. 8, 52 F, 52 M.
159. *Eur. Chem. News* **52** (1989) no. 1364, 7.
160. *Hydrocarbon Process.* **67** (1988) no. 2, 17.
161. S. Warwel, *Chem. Ztg.* **103** (1987) no. 5, 243.
162. *Rubber Statistical Bulletin,* **43** (1988/89) no. 1, 45.
163. *Worldwide Rubber Statistics,* Internat. Inst. of Synth. Rubber Producers, Inc., Houston, Texas 1989.
164. H. W. Gerarde in G. D. Clayton, F. E. Clayton (eds.): *Patty's Toxicology,* 2nd rev. ed., vol. II, J. Wiley & Sons, New York – London 1963, p. 1205.
165. Unpublished data, Bayer AG.
166. R. Melnick: "Inhalation Toxicology Studies of Isoprene in F344 Rats and B6C3F1 Mice," *Abstract, Int. Symp. Toxicology, Carcinogenesis & Human Health Aspects of 1,3-Butadiene,* NIEHS, Research Triangle Park, N.C., April 1988.
167. K. Mortelmans et al.: "Salmonella Mutagenicity Tests: II. Results From the Testing of 270 Chemicals," *Environ. Mutagen.* **8, Supplement 7** (1986) 1 –119.
168. C. de Meester in I. Gut, M. Cikrt, G. L. Plaa (eds.): *Industrial and Environmental Xenobiotics,* Springer Verlag, Berlin 1981, pp. 195 – 203.
169. A. Kushi et al.: "Mutagenicity of Gaseous Nitrogen Oxides and Olefins on Salmonella TA102 and TA104," *Mutat. Res.* **147** (1985) 263 – 264.
170. P. G. Gervasi et al.: "Mutagenicity and Chemical Reactivity of Epoxide Intermediates of the Isoprene Metabolism and Other Structurally Related Compounds," *Mutat. Res.* **156** (1985) 77 – 82.
171. R. R. Tice et al.: "Chloroprene and isoprene: cytogenetic studies in mice," *Mutagenesis* **3** (1988) 141 –146.
172. M. Del Monte et al.: "Isoprene Metabolism by Liver Microsomal Mono-oxygenases," *Xenobiotica* **15** (1985) 591 – 597.
173. V. Longo et al.: "Hepatic Microsomal Metabolism of Isoprene in Various Rodents," *Toxicol. Lett.* **29** (1985) 33 – 37.
174. H. Peter et al.: "Pharmacokinetics of Isoprene in Mice and Rats," *Toxicol. Lett.* **36** (1987) 9 – 14.
175. A. R. Dahl et al.: "The Fate of Isoprene Inhaled by Rats: Comparison to Butadiene," *Toxicology Appl. Pharmacol.* **89** (1987) 237 – 248.

Isoprene Polymers → Rubber, 3. Synthetic
Isopropanolamine → Ethanolamines and Propanolamines
Isopropyl Alcohol → Propanols
Isoquinoline → Quinoline and Isoquinoline
Isotope Separation → Isotopes, Natural

Isotopes, Natural

Mark A. Miller, BWXT Y-12, L.L.C., Oak Ridge, Tennessee 37831, United States

A. Andrew Carey, CTI, Inc., Knoxville, Tennessee 37932, United States

1. Fundamentals 2
1.1. Definitions and Nomenclature . . . 2
1.2. Survey of Natural Nuclides 2
1.3. Origins of Nuclides 12
1.4. Discovery of Isotopes 14
1.5. Determination of Isotopic Masses and Abundance 16
1.6. Atomic Mass Scales 17
1.7. Variations in Abundance 18
2. Properties of Isotopes 18
2.1. Physical Differences 19
2.1.1. Mass Effects 19
2.1.2. Nuclear Effects 21
2.2. Chemical Differences 21
2.2.1. Equilibrium Effects 21
2.2.2. Kinetic Effects 22
2.2.3. Bonding Effects 23
2.3. Isotope Effects in Nature 24
3. Uses of Isotopes 26
3.1. Nuclear Energy 26
3.2. Research 27
3.2.1. Nuclear Magnetic Resonance 27
3.2.2. Electron Spin Resonance 29
3.2.3. Mössbauer Spectroscopy 29
3.2.4. Isotopic Substitution 30
3.2.5. Cosmology 31
3.2.6. Nuclear Science 31
3.2.7. Materials Science 31
3.2.8. Ecological Research 32
3.2.9. Isotope Chronostratigraphy 34
3.2.10. Anthropology 35
3.3. Medical Uses 35
3.3.1. Magnetic Resonance Imaging 35
3.3.2. Stable Isotopes in Metabolic Studies 37
3.3.3. Stable Isotopes in Medical Diagnosis 38
3.3.4. Stable Isotopes as Radioisotope Precursors 38
3.4. Industrial Uses 38
4. Isotope Separation 39
4.1. The Separation Factor 39
4.1.1. Continuous Processes and Cascades 40
4.1.2. Discontinuous Processes 43
4.2. Separation Processes 44
4.2.1. Distillation 44
4.2.2. Diffusion 48
4.2.3. Gas Centrifugation 52
4.2.4. Nozzle Separation 55
4.2.5. Electromagnetic Separation 56
4.2.6. Exchange Reactions 59
4.2.7. Photochemical Separation 64
4.2.8. Electrolysis 64
4.2.9. Electromigration 65
4.2.10. Other Separation Processes 66
5. Availibility 66
6. References 67

1. Fundamentals

1.1. Definitions and Nomenclature

The term *nuclide* is used to indicate a species of atom characterized by the constituents of its nucleus; i.e., the number of protons Z (the atomic number) and the number of neutrons N. The total number of protons and neutrons, referred to nonspecifically as *nucleons,* is the mass number $A = Z + N$. The term *isotope* is used to define any of two or more nuclides that have the same atomic number Z but different atomic mass numbers A. Thus, isotopes of an element differ in the number of neutrons each contains. The definitions indicate that all isotopes are nuclides but all nuclides are not necessarily isotopes. Correct usage dictates that an element having only one naturally occurring species (e.g., fluorine) should be referred to as mononuclidic; monoisotopic is a contradictory term and should not be used.

The nomenclature used to identify nuclides employs the atomic mass number of the species as a superscript on the left of the chemical symbol of the element (e.g., ^{20}Ne, ^{196}Hg, and ^{202}Hg). In some publications, the mass number is written as a superscript to the right of the sym-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_001

bol, but this position is properly reserved for the charge on an ionic species or for the indication of an excited state. The atomic number can be written as a subscript on the left of the symbol (e.g., $^{20}_{10}$Ne, $^{196}_{82}$Hg and $^{202}_{82}$Hg). The isotopes of hydrogen are probably the best known and most widely used nuclides, and have therefore been given individual names: ^{2}H is called *deuterium* and is abbreviated D; ^{3}H is called *tritium* and is abbreviated T. For consistency, the suggestion has been made that ^{1}H be called *protium* and that the term hydrogen should be used only to refer to the mixture of isotopes, but this has not been generally accepted. The term hydrogen is used in this chapter to refer to both the element and its most abundant isotope.

The presence of a variable number of neutrons in the nucleus of isotopes can affect the chemical reactivity of compounds, especially those containing isotopes of the lightest element, hydrogen. In general, however, isotopic effects are small compared to reactivity difference between elements. More prominent differences are noted in physical and nuclear properties among isotopes of the same element, where mass and magnetic properties play an important role.

1.2. Survey of Natural Nuclides

Natural nuclides are those present in the Earth's crust and atmosphere. At least one nuclide has been discovered on Earth for each element with atomic numbers 1 to 92, with the exception of technetium ($_{43}$Tc) and promethium ($_{61}$Pm). Isotopes of both these elements, however, have been identified in the spectral signature of some stars and have been produced artificially in the laboratory [1]. A minute quantity of an isotope of plutonium $^{244}_{94}$Pu has also been found in ores excavated in California [2]. The 91 naturally occurring elements consist of 327 nuclides. The existence of isotopes is common over much of the periodic table. Only 19 of the elements found on Earth are mononuclidic. Two other elements, bismuth and thorium, have only one naturally occurring stable isotope, but also have other isotopes that are members of radioactive decay series. Conversely, tin has ten isotopes, and 16 other elements have seven or more isotopes. Table 1 lists naturally occurring nuclides by atomic number and atomic mass number. Experimentally determined atomic masses and the abundance of each nuclide as a percentage for each element are also included.

The naturally occurring nuclides can be divided into two categories, stable and radioactive. This classification is, however, arbitrary in some cases. The group of 287 stable nuclides also includes isotopes of at least 17 elements that undergo radioactive decay but whose half-lifes are long with respect to geologic time (generally nuclides with half-lifes of ca. 10^9 a and longer are considered stable). A qualifier is used here because the radioactive decay of some nuclides, predicted theoretically, is confirmed experimentally from time to time. These species usually have extremely long half-lifes, which makes detection of nuclear reactions difficult (e.g., ^{82}Se, $t_{1/2}$ ca. 10^{20} a; see Section 3.1). At present, 14 nuclides of this type have been identified. The term stable might prove to be a misnomer because certain theories of nuclear physics suggest that the proton may be unstable to decay, with a half-life of $10^{30} - 10^{32}$ a [3]. Experimental results have not yet shown this to be true, but with events expected approximately once every hundred years, some time may be required for this conjecture to be proved or disproved.

The group of radioactive nuclides includes all the naturally occurring species with atomic number higher than 83 (bismuth), with the exception of one isotope of thorium (^{232}Th) and two isotopes of uranium (^{235}U and ^{238}U), whose halflives qualify them as stable. Of these 40 high-mass radioactive species, 39 are daughter nuclides in the decay series of thorium and uranium. The plutonium isotope ^{244}Pu goes through several decay steps and is eventually transformed into ^{232}Th. Although 20 nuclides with lower atomic numbers also undergo radioactive decay, 18 of these are regarded as stable. The two exceptions ^{3}H and ^{14}C have been identified in trace amounts in the Earth's atmosphere [4]. These radioactive species have relatively short half-lifes (12.26 and 5730 a, respectively) and are thought to be generated continually in small quantities by solar-related phenomena in the upper atmosphere. Table 2 gives information about naturally occurring radioactive isotopes, including half-life, decay mode, decay energy, and the product of nuclear reactions for each species.

Table 1. Naturally occurring nuclides

Element	Relative atomic mass	Isotope	*Z*	*N*	Abundance %	Mass
Hydrogen	1.00797	^{1}H	1	0	99.985	1.0078
		^{2}H	1	1	0.015	2.0140
		^{3}H [a]	1	2	trace	3.0160
Helium	4.0026	^{3}He	2	1	0.00014	3.0160
		^{4}He	2	2	99.99986	4.0026
Lithium	6.9409	^{6}Li	3	3	7.5	6.0151
		^{7}Li	3	4	92.5	7.0160
Beryllium		^{9}Be	4	5	100	9.0121
Boron	10.8100	^{10}B	5	5	19.8	10.0129
		^{11}B	5	6	80.2	11.0093
Carbon	12.0111	^{12}C	6	6	98.90	12.0000
		^{13}C	6	7	1.10	13.0034
		^{14}C [a]	6	8	trace	14.0032
Nitrogen	14.0067	^{14}N	7	7	99.63	14.0031
		^{15}N	7	8	0.37	15.0001
Oxygen	15.9994	^{16}O	8	8	99.762	15.9949
		^{17}O	8	9	0.038	16.9991
		^{18}O	8	10	0.200	17.9992
Fluorine		^{19}F	9	10	100	18.9984
Neon	20.179	^{20}Ne	10	10	90.51	19.9924
		^{21}Ne	10	11	0.27	20.9938
		^{22}Ne	10	12	9.22	21.9914
Sodium		^{23}Na	11	12	100	22.9898
Magnesium	24.305	^{24}Mg	12	12	78.99	23.9850
		^{25}Mg	12	13	10.00	24.9858
		^{26}Mg	12	14	11.01	25.9826
Aluminum		^{27}Al	13	14	100	26.9815
Silicon	28.0855	^{28}Si	14	14	92.23	27.9769
		^{29}Si	14	15	4.67	28.9765
		^{30}Si	14	16	3.10	29.9738
Phosphorus		^{31}P	15	16	100	30.9738
Sulfur	32.066	^{32}S	16	16	95.02	31.9721
		^{33}S	16	17	0.75	32.9715
		^{34}S	16	18	4.21	33.9679
		^{36}S	16	20	0.02	35.9671
Chlorine	35.453	^{35}Cl	17	18	75.77	34.9688
		^{37}Cl	17	20	24.23	36.9659
Argon	39.948	^{36}Ar	18	18	0.337	35.9676
		^{38}Ar	18	20	0.063	37.9627
		^{40}Ar	18	22	99.60	39.9624
Potassium	39.0983	^{39}K	19	20	93.2581	38.9637
		^{40}K [a]	19	21	0.0117	39.9640
		^{41}K	19	22	6.7302	40.9618
Calcium	40.078	^{40}Ca	20	20	96.941	39.9626
		^{42}Ca	20	22	0.647	41.9586
		^{43}Ca	20	23	0.135	42.9588
		^{44}Ca	20	24	2.086	43.9555
		^{46}Ca	20	26	0.004	45.9537
		^{48}Ca	20	28	0.187	47.9525
Scandium		^{45}Sc	21	24	100	44.9599
Titanium	47.88	^{46}Ti	22	24	8.0	45.9526
		^{47}Ti	22	25	7.3	46.9518
		^{48}Ti	22	26	73.8	47.9480

Table 1. (Continued)

Element	Relative atomic mass	Isotope	Z	N	Abundance %	Mass
		^{49}Ti	22	27	5.5	48.9479
		^{50}Ti	22	28	5.4	49.9448
Vanadium	50.9415	^{50}V [a]	23	27	0.25	49.9472
		^{51}V	23	28	99.75	50.9440
Chromium	51.996	^{50}Cr	24	26	4.35	49.9460
		^{52}Cr	24	28	83.79	51.9405
		^{53}Cr	24	29	9.50	52.9406
		^{54}Cr	24	30	2.36	53.9389
Manganese		^{55}Mn	25	30	100	54.9380
Iron	55.847	^{54}Fe	26	28	5.8	53.9396
		^{56}Fe	26	30	91.72	55.9349
		^{57}Fe	26	31	2.2	56.9354
		^{58}Fe	26	32	0.28	57.9333
Cobalt		^{59}Co	27	32	100	58.9332
Nickel	58.67	^{58}Ni	28	30	68.27	57.9353
		^{60}Ni	28	32	26.10	59.9308
		^{61}Ni	28	33	1.13	60.9310
		^{62}Ni	28	34	3.59	61.9283
		^{64}Ni	28	36	0.91	63.9280
Copper	63.546	^{63}Cu	29	34	69.17	62.9396
		^{65}Cu	29	36	30.83	64.9278
Zinc	65.39	^{64}Zn	30	34	48.6	63.9291
		^{66}Zn	30	36	27.9	65.9260
		^{67}Zn	30	37	4.1	66.9721
		^{68}Zn	30	38	18.8	67.9248
		^{70}Zn	30	40	0.6	69.9252
Gallium	69.723	^{69}Ga	31	38	60.1	68.9256
		^{71}Ga	31	40	39.9	70.9247
Germanium	72.59	^{70}Ge	32	38	20.5	69.9243
		^{72}Ge	32	40	27.4	71.9221
		^{73}Ge	32	41	7.8	72.9235
		^{74}Ge	32	42	36.5	73.9212
		^{76}Ge	32	44	7.8	75.9214
Arsenic		^{75}As	33	42	100	74.9216
Selenium	78.96	^{74}Se	34	40	0.9	73.9225
		^{76}Se	34	42	9.0	75.9192
		^{77}Se	34	43	7.6	76.9199
		^{78}Se	34	44	23.5	77.9173
		^{80}Se	34	46	49.6	79.9165
		^{82}Se [a]	34	48	9.4	81.9167
Bromine	79.904	^{79}Br	35	44	50.69	78.9183
		^{81}Br	35	46	49.31	80.9163
Krypton	83.80	^{78}Kr	36	42	0.35	77.9204
		^{80}Kr	36	44	2.25	79.9164
		^{82}Kr	36	46	11.6	81.9135
		^{83}Kr	36	47	11.5	82.9141
		^{84}Kr	36	48	57.0	83.9115
		^{86}Kr	36	50	17.3	85.9106
Rubidium	85.4678	^{85}Rb	37	48	72.17	84.9118
		^{87}Rb [a]	37	50	27.83	86.9092
Strontium	87.62	^{84}Sr	38	46	0.56	83.9134
		^{86}Sr	38	48	9.86	85.9093
		^{87}Sr	38	49	7.00	86.9089
		^{88}Sr	38	50	82.58	87.9056
Yttrium		^{89}Y	39	50	100	88.9059
Zirconium	91.224	^{90}Zr	40	50	51.45	89.9047

Table 1. (Continued)

Element	Relative atomic mass	Isotope	Z	N	Abundance %	Mass
		^{91}Zr	40	51	11.27	90.9056
		^{92}Zr	40	52	17.17	91.9050
		^{94}Zr	40	54	17.33	93.9063
		^{96}Zr	40	56	2.78	95.9083
Niobium		^{93}Nb	41	52	100	92.9064
Molybdenum	95.94	^{92}Mo	42	50	14.84	91.9068
		^{94}Mo	42	52	9.25	93.9051
		^{95}Mo	42	53	15.92	94.9058
		^{96}Mo	42	54	16.68	95.9047
		^{97}Mo	42	55	9.55	96.9060
		^{98}Mo	42	57	24.13	97.9054
		^{100}Mo	42	58	9.63	99.9075
Ruthenium	101.07	^{96}Ru	44	52	5.52	95.9076
		^{98}Ru	44	54	1.88	97.9053
		^{99}Ru	44	55	12.7	98.9059
		^{100}Ru	44	56	12.6	99.9042
		^{101}Ru	44	57	17.0	100.9056
		^{102}Ru	44	58	31.6	101.9043
		^{104}Ru	44	60	18.7	103.9054
Rhodium		^{103}Rh	45	58	100	102.9055
Palladium	106.42	^{102}Pd	46	56	1.02	101.9056
		^{104}Pd	46	58	11.14	103.9040
		^{105}Pd	46	59	22.33	104.9051
		^{106}Pd	46	60	27.33	105.9035
		^{108}Pd	46	62	26.46	107.9039
		^{110}Pd	46	64	11.72	109.9052
Silver	107.8682	^{107}Ag	47	60	51.84	106.9051
		^{109}Ag	47	62	48.16	108.9048
Cadmium	112.41	^{106}Cd	48	58	1.25	105.9065
		^{108}Cd	48	60	0.89	107.9042
		^{110}Cd	48	62	12.49	109.9030
		^{111}Cd	48	63	12.80	110.9042
		^{112}Cd	48	64	24.13	111.9028
		^{113}Cd	48	65	12.22	112.9044
		^{114}Cd	48	66	28.73	113.9034
		^{116}Cd	48	68	7.49	115.9048
Indium	114.82	^{113}In	49	64	4.3	112.9041
		^{115}In	49	66	95.7	114.9038
Tin	118.710	^{112}Sn	50	62	1.0	111.9048
		^{114}Sn	50	64	0.7	113.9028
		^{115}Sn	50	65	0.4	114.9033
		^{116}Sn	50	66	14.7	115.9017
		^{117}Sn	50	67	7.7	116.9030
		^{118}Sn	50	68	24.3	117.9016
		^{119}Sn	50	69	8.6	118.9033
		^{120}Sn	50	70	32.4	119.9022
		^{122}Sn	50	72	4.6	121.9034
		^{124}Sn	50	74	5.6	123.9053
Antimony	121.75	^{121}Sb	51	70	57.3	120.9038
		^{123}Sb	51	72	42.7	122.9041
Tellurium	127.60	^{120}Te	52	68	0.096	119.9040
		^{122}Te	52	70	2.60	121.9030
		^{123}Te [a]	52	71	0.903	122.9043
		^{124}Te	52	72	4.816	123.9028
		^{125}Te	52	73	7.14	124.9044
		^{126}Te	52	74	18.95	125.9033

Table 1. (Continued)

Element	Relative atomic mass	Isotope	*Z*	*N*	Abundance %	Mass
		^{128}Te	52	76	31.69	127.9045
		^{130}Te	52	78	33.80	129.9062
Iodine		^{127}I	53	74	100	126.9045
Xenon	131.29	^{124}Xe	54	70	0.10	123.9059
		^{126}Xe	54	72	0.09	125.9043
		^{128}Xe	54	74	1.91	127.9035
		^{129}Xe	54	75	26.4	128.9048
		^{130}Xe	54	76	4.1	130.9035
		^{131}Xe	54	77	21.2	130.9051
		^{132}Xe	54	78	26.9	131.9041
		^{134}Xe	54	80	10.4	133.9054
		^{136}Xe	54	82	8.9	135.9072
Cesium		^{133}Cs	55	78	100	132.9054
Barium	137.33	^{130}Ba	56	74	0.106	129.9063
		^{132}Ba	56	76	0.101	131.9050
		^{134}Ba	56	78	2.417	133.9045
		^{135}Ba	56	79	6.592	134.9057
		^{136}Ba	56	80	7.854	135.9045
		^{137}Ba	56	81	11.23	136.9058
		^{138}Ba	56	82	71.70	137.9052
Lanthanum	138.9055	^{138}La	57	81	0.09	137.9071
		^{139}La	57	82	99.91	138.9063
Cerium	140.12	^{136}Ce	58	78	0.19	135.9071
		^{138}Ce	58	80	0.25	137.9060
		^{140}Ce	58	82	88.48	139.9054
		^{142}Ce	58	84	11.08	141.9092
Praseodymium		^{141}Pr	59	82	100	140.9076
Neodymium	144.24	^{142}Nd	60	82	27.13	141.9077
	^{143}Nd	60	83	12.18	142.9098	
		^{144}Nd	60	84	23.80	143.9101
		^{145}Nd	60	85	8.30	144.9126
		^{146}Nd	60	86	17.19	145.9131
		^{148}Nd	60	88	5.76	147.9169
		^{150}Nd	60	90	5.64	149.9209
Samarium	150.36	^{144}Sm	62	82	3.1	143.9112
		^{147}Sm [a]	62	85	15.0	146.9149
		^{148}Sm [a]	62	86	11.3	147.9148
		^{149}Sm [a]	62	87	13.8	148.9172
		^{150}Sm	62	88	7.4	149.9173
		^{152}Sm	62	90	26.7	151.9197
		^{154}Sm	62	92	22.7	153.9222
Europium	151.96	^{151}Eu	63	88	47.8	150.9198
		^{153}Eu	63	90	52.2	152.9212
Gadolinium	157.25	^{152}Gd [a]	64	88	0.20	151.9198
		^{154}Gd	64	90	2.18	153.9209
		^{155}Gd	64	91	14.80	154.9226
		^{156}Gd	64	92	20.47	155.9221
		^{157}Gd	64	93	15.65	156.9240
		^{158}Gd	64	94	24.84	157.9241
		^{160}Gd	64	96	21.86	159.9270
Terbium		^{159}Tb	65	94	100	159.9254
Dysprosium	162.50	^{156}Dy	66	90	0.06	155.9253
		^{158}Dy	66	92	0.10	157.9244
		^{160}Dy	66	94	2.34	159.9252
		^{161}Dy	66	95	18.9	160.9269
		^{162}Dy	66	96	25.5	161.9268

Table 1. (Continued)

Element	Relative atomic mass	Isotope	*Z*	*N*	Abundance %	Mass
		^{163}Dy	66	97	24.9	162.9287
		^{164}Dy	66	98	28.2	163.9292
Holmium		^{165}Ho	67	98	100	164.9304
Erbium	167.26	^{162}Er	68	94	0.14	161.9288
		^{164}Er	68	96	1.61	163.9292
		^{166}Er	68	98	33.6	165.9303
		^{167}Er	68	99	22.95	166.9320
		^{168}Er	68	100	26.8	167.9324
		^{170}Er	68	102	14.9	169.9355
Thulium		^{169}Tm	69	100	100	168.9342
Ytterbium	173.04	^{168}Yb	70	98	0.13	167.9339
		^{170}Yb	70	100	3.05	169.9348
		^{171}Yb	70	101	14.3	170.9363
		^{172}Yb	70	102	21.9	171.9364
		^{173}Yb	70	103	16.12	172.9382
		^{174}Yb	70	104	31.8	173.9390
		^{176}Yb	70	106	12.7	175.9426
Lutetium	174.967	^{175}Lu	71	104	97.4	174.9408
		^{176}Lu [a]	71	105	2.66	175.9427
Hafnium	178.49	^{174}Hf [a]	72	102	0.16	173.9400
		^{176}Hf	72	104	5.2	175.9414
		^{177}Hf	72	105	18.6	176.9432
		^{178}Hf	72	106	27.1	177.9437
		^{179}Hf	72	107	13.74	178.9458
		^{180}Hf	72	108	35.2	179.9465
Tantalum	180.9419	^{180}Ta [a]	73	107	0.012	179.9475
		^{181}Ta	73	108	99.988	180.9480
Tungsten	183.85	^{180}W	74	106	0.13	179.9467
		^{182}W	74	108	26.3	181.9482
		^{183}W	74	109	14.3	182.9502
		^{184}W	74	110	30.67	183.9505
		^{186}W	74	112	28.6	185.9544
Rhenium	186.207	^{185}Re	75	110	37.4	184.9530
		^{187}Re	75	112	62.6	186.9557
Osmium	190.2	^{184}Os	76	108	0.02	183.9525
		^{186}Os	76	110	1.58	185.9538
		^{187}Os	76	111	1.6	186.9557
		^{188}Os	76	112	13.3	187.9559
		^{189}Os	76	113	16.1	188.9581
		^{190}Os	76	114	26.4	189.9584
		^{192}Os	76	116	41.0	191.9615
Iridium	192.2	^{191}Ir	77	114	37.3	190.9606
		^{193}Ir	77	116	62.7	192.9629
Platinum	195.08	^{190}Pt [a]	78	112	0.01	189.9599
		^{192}Pt [a]	78	114	0.79	191.9610
		^{194}Pt	78	116	32.9	193.9626
		^{195}Pt	78	117	33.8	194.9648
		^{196}Pt	78	118	25.3	195.9649
		^{198}Pt	78	120	7.2	197.9679
Gold		^{197}Au	79	118	100	196.9665
Mercury	200.59	^{196}Hg	80	116	0.15	195.9658
		^{198}Hg	80	118	10.1	197.9667
		^{199}Hg	80	119	17.0	198.9682
		^{200}Hg	80	120	23.1	199.9683
		^{201}Hg	80	121	13.2	200.9703
		^{202}Hg	80	122	29.65	201.9706

Table 1. (Continued)

Element	Relative atomic mass	Isotope	Z	N	Abundance %	Mass
		^{204}Hg	80	124	6.8	203.9735
Thallium	204.383	^{203}Tl	81	122	29.52	202.9723
		^{205}Tl	81	124	70.48	204.9744
		^{206}Tl [b]	81	125		
		^{207}Tl [b]	81	126		
		^{208}Tl [b]	81	127		
		^{210}Tl [b]	81	129		
Lead	207.2	^{204}Pb	82	122	1.4	203.9730
		^{206}Pb	82	124	24.1	205.9744
		^{207}Pb	82	125	22.1	206.9759
		^{208}Pb	82	126	52.4	207.9766
		^{210}Pb [b]	82	128		
		^{211}Pb [b]	82	129		
		^{212}Pb [b]	82	130		
		^{214}Pb [b]	82	132		
Bismuth	208.9804	^{209}Bi	83	126	100	208.9804
		^{210}Bi [b]	83	127		
		^{211}Bi [b]	83	128		
		^{212}Bi [b]	83	129		
		^{214}Bi [b]	83	131		
Polonium		^{211}Po [b]	84	127		
		^{212}Po [b]	84	128		
		^{214}Po [b]	84	130		
		^{215}Po [b]	84	131		
		^{216}Pb [b]	84	132		
		^{218}Pb [b]	84	134		
Astatine		^{215}At [b]	85	130		
		^{218}At [b]	85	133		
Radon		^{219}Rn [b]	86	133		
		^{220}Rn [b]	86	134		
		^{222}Rn [b]	86	136		
Francium		^{223}Fr [b]	87	136		
Radium		^{223}Ra [b]	88	135		
		^{224}Ra [b]	88	136		
		^{226}Ra [b]	88	138		
		^{228}Ra [b]	88	140		
Actinium		^{227}Ac [b]	89	138		
		^{228}Ac [b]	89	139		
Thorium		^{227}Th [b]	90	137		
		^{228}Th [b]	90	138		
		^{230}Th [b]	90	140		
		^{231}Th [b]	90	141		
		^{232}Th [a]	90	142	100	232.0380
		^{234}Th [b]	90	144		
Protactinium		^{231}Pa [b]	91	140		
		^{234}Pa [b]	91	143		
Uranium	238.03	^{234}U [b]	92	142	0.005	234.0409
		^{235}U [a]	92	143	0.720	235.0439
		^{238}U [a]	92	146	99.275	238.0508
Plutonium		^{244}Pu [a]	94	150	trace	244.0642

[a] Naturally occurring radioactive nuclide (see Table 2).
[b] Member of radioactive decay series (see Table 2).

Table 2. Decay of naturally occurring radioactive nuclides

Isotope	Half-life	Decay mode ∗	Product	Decay energy, MeV
^{3}H	12.26 a	β^-	^{3}He	0.018
^{14}C	5730 a	β^-	^{14}N	0.156
^{40}K	1.28×10^{9} a	β^-	^{40}Ca	1.35
		β^+	^{40}Ar	1.505
		EC	^{40}Ar	
^{50}V	6×10^{15} a	EC	^{50}Ti	
^{82}Se	$>10^{20}$ a	double β^-	^{82}Kr	
^{87}Rb	4.9×10^{10} a	β^-	^{87}Sr	0.274
^{115}In	4.4×10^{11} a	β^-	^{115}Sn	0.496
^{123}Te	1.2×10^{13} a	EC	^{123}Sb	0.06
^{138}La	1.06×10^{11} a	EC (66 %)	^{138}Ba	1.75
		β^- (34 %)	^{138}Ce	1.04
^{144}Nd	5×10^{15} a	α	^{140}Ce	1.8
^{147}Sm	1.06×10^{11} a	α	^{143}Nd	2.314
^{148}Sm	1.2×10^{13} a	α	^{144}Nd	2.001
^{149}Sm	4×10^{14} a	α	^{145}Nd	1.9
^{152}Gd	1.1×10^{14} a	α	^{148}Sm	2.24
^{176}Lu	3×10^{10} a	β^-	^{176}Hf	1.02
^{174}Hf	2×10^{15} a	α	^{170}Y	2.55
^{180}Ta	$>10^{13}$ a	β^-	^{180}W	
		EC	^{180}Hf	
^{187}Re	7×10^{10} a	β^-	^{187}Os	<0.01
^{190}Pt	6×10^{11} a	α	^{186}Os	3.18
^{192}Pt	10^{15} a	α	^{188}Os	2.6
^{244}Pu	8×10^{7} a	α	^{240}U	4.66
Thorium series				
^{232}Th	1.4×10^{10} a	α	^{228}Ra	4.08
^{228}Ra	6.7 a	β^-	^{228}Ac	0.055
^{228}Ac	6.13 h	β^-	^{228}Th	2.14
^{228}Th	1.91 a	α	^{224}Ra	5.521
^{224}Ra	3.64 d	α	^{220}Rn	5.787
^{220}Rn	52 s	α	^{216}Po	6.405
^{216}Po	0.16 s	α	^{212}Pb	6.906
^{212}Pb	10.6 h	β^-	^{212}Bi	0.58
^{212}Bi	60.5 min	β^-	^{212}Po	2.25
		α	^{208}Ti	6.206
^{212}Po	3×10^{-7} s	α	^{208}Pb	
^{208}Tl	3.1 min	β^-	^{208}Pb	4.994
Actinium series				
^{235}U	7.13×10^{8} a	α	^{231}Th	4.681
^{231}Th	25.6 h	β^-	^{231}Pa	0.381
^{231}Pa	3.43×10^{4} a	α	^{227}Ac	5.148
^{227}Ac	21.8 a	β^-	^{227}Th	0.043
		α	^{223}Fr	5.04
^{227}Th	18.4 d	α	^{223}Ra	6.145
^{223}Fr	21 min	β^-	^{223}Ra	6.145
^{223}Ra	11.7 d	α	^{219}Rn	5.977
^{219}Rn	3.92 s	α	^{215}Po	6.944
^{215}Po	1.83×10^{-3} s	α	^{211}Pb	7.524
		β^-	^{215}At	
^{211}Pb	36.1 min	β^-	^{211}Bi	1.37
^{215}At	1.0×10^{-4} s	α	^{211}Bi	8.16
^{211}Bi	2.16 min	α	^{207}Tl	6.75
		β^-	^{211}Po	0.60
^{207}Tl	4.8 min	β^-	^{207}Pb	1.44
^{211}Po	0.52 s	α	^{207}Pb	7.592
Uranium series				
^{238}U	4.51×10^{9} a	α	^{234}Th	4.268
^{234}Th	24.1 d	β^-	^{234}Pa	0.263
^{234}Pa	1.18 min	β^-	^{234}U	2.23
^{234}U	2.48×10^{5} a	α	^{230}Th	4.856
^{230}Th	8×10^{4} a	α	^{226}Ra	4.767
^{226}Ra	1.62×10^{3} a	α	^{222}Rn	5.06
^{222}Rn	3.82 d	α	^{218}Po	6.68
^{218}Po	3.05 min	α	^{214}Pb	6.111
		β^-	^{218}At	0.278
^{214}Pb	26.8 min	β^-	^{214}Bi	1.04
^{218}At	2 s	α	^{214}Bi	6.816
^{214}Bi	19.7 min	β^-	^{214}Po	3.28
		α	^{210}Tl	5.616
^{214}Po	1.6×10^{-4} s	α	^{210}Pb	7.835
^{210}Tl	1.32 min	β^-	^{210}Pb	5.50
^{210}Pb	19.4 a	β^-	^{210}Bi	0.061
^{210}Bi	5.0 d	α	^{206}Tl	5.044
		β^-	^{210}Po	1.16
^{206}Tl	4.2 min	β^-	^{206}Pb	1.524
^{210}Pb	138.4 d	α	^{206}Pb	3.72

∗ EC = electron capture

Figure 1 shows a graph of the natural nuclides with the number of neutrons plotted against the number of protons; included in this figure are lines of constant mass. Note that the nuclides are generally grouped along the $N = Z$ line at low atomic numbers and have $N > Z$ as the atomic number increases. The stability of nuclides is a function of several factors, including the number of protons and the number of neutrons, the binding energy holding the nucleus together, and the availability of pathways for transforming one nucleus into another. Certain numbers of protons or neutrons appear to be particularly stable. In the case of protons, $_{20}$Ca and $_{50}$Sn have more isotopes (six and ten, respectively) than other elements in the vicinity of each of these species. The *nuclear shell theory* [5] explains this by invoking subshells that are filled sequentially as atomic number and atomic mass number increase, much as electron shells are filled. Protons and neutrons have separate subshells, but both have the same numerical configuration. Completely filled subshells are said to contain *magic numbers* of nucleons (2, 8, 20, 28, 50, 82, and 126) and confer added stability on the nucleus,

just as completely filled electron shells result in low chemical reactivity for the noble gases. The binding energy of a nucleus can be envisioned as a part of the mass that is converted to energy when a nucleus is formed from its constituent nucleons. The relative stability of a nuclide is found by calculating the binding energy per nucleon and comparing it with that of its neighboring species. For spontaneous decay to transform one nuclide into another, the product must be more stable than its precursor. Thus, $^{82}_{34}$Se does not decay to $^{82}_{35}$Br (by electron emission), $^{82}_{33}$As (by electron capture or positron emission), or $^{78}_{32}$Ge (by alpha emission), instead, it undergoes the rare double-beta decay to form $^{82}_{36}$Kr.

Several general features may be gleaned from Table 1 and Figure 1:

1) All but one ($^{9}_{4}$Be) of the mononuclidic elements have odd atomic numbers. Other odd-numbered elements have only two naturally occurring stable isotopes. Table 1 lists three isotopes for hydrogen and potassium, but ^{3}H and ^{40}K are radioactive and present in very small amounts.
2) Elements with $Z < 22$ are mononuclidic or have one predominant isotope; six are mononuclidic; five are at least 99 % abundant (^{12}C is 98.9 % abundant); six are at least 90 % abundant; three are at least 75 % abundant.
3) For elements with odd atomic numbers and $Z < 22$, the only nuclide, or the predominant isotope, has $N = Z + 1$, with hydrogen and nitrogen being exceptions.
4) For elements with even atomic numbers and $Z < 22$, the most abundant isotope has $N = Z$ (e.g., nitrogen). Argon is an exception.
5) At higher atomic numbers ($Z > 21$), the likelihood of a larger number of isotopes and a broader distribution of relative abundances increases. All of the elements with even atomic numbers have at least four isotopes. The occurrence of seven or more isotopes is common above niobium ($Z > 41$). Elements with odd atomic numbers that are not mononuclidic tend to have a more balanced distribution of isotopic abundances: copper, thallium, and rubidium, ca. 70 : 30; gallium, antimony, rhenium, and iridium, ca. 60 : 40; bromine, silver, and europium, ca. 50 : 50. Vanadium, lanthanum, lutetium, and tantalum do not fit this pattern. They each have one isotope that is at least 99 % abundant (97 % for Lu), but the less abundant species in these cases is radioactive. Indium is unusual in that its most abundant isotope (^{115}In, 95.7 %) is radioactive.
6) Figure 1 shows that as atomic number increases, the stable nuclides are those that are neutron rich (i.e., $N > Z$). In fact, no naturally occurring nuclides heavier than ^{3}He exist in which the number of protons is greater than the number of neutrons. The ratio of neutrons to protons is close to 1 : 1 at low atomic numbers but gradually increases to ca. 1.5 : 1 for the heaviest natural nuclides. One theory of nuclear structure suggests that as the number of positively charged protons increases, more neutrons are needed to separate positive charges.

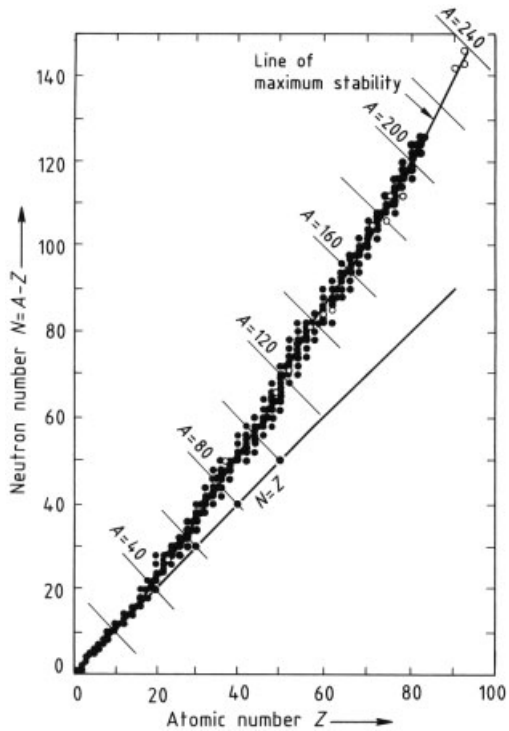

Figure 1. Chart of natural nuclides
● Stable nuclides; ○ Radioactive nuclides with half-life $>10^8$ a

Table 1 lists the abundance of the nuclides found on Earth. By way of contrast, Figures 2 and 3 depict the relative abundances of elements and nuclides in the Solar System, expressed as number of atoms per 10^6 atoms of silicon [6].

The data are obtained from investigations of meteorites and the solar atmosphere. Values reflect the relative amounts of nuclides in the nebula from which the Solar System evolved, some 4×10^9 years ago. In some cases, these values do not correspond to the abundances found on Earth, especially for volatile elements and compounds. A striking example of this is argon, whose most abundant isotope on Earth is ^{40}Ar. The most abundant isotope (>99 %) in the Solar System is ^{36}Ar. The isotope ^{40}Ar found on Earth is almost exclusively a result of the decay of ^{40}K since the Earth was formed; the original argon, rich in ^{36}Ar, was lost through outgassing from the newly condensed matter that eventually became the Earth.

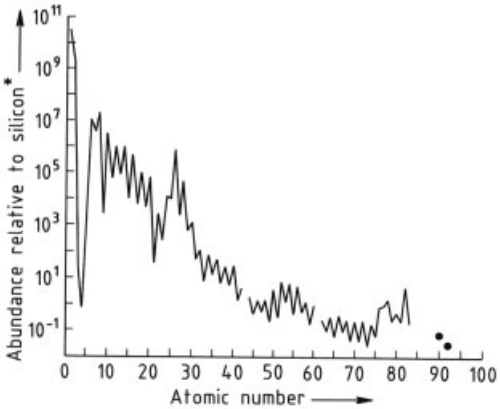

Figure 2. Abundance versus atomic number
∗ Number of atoms per 10^6 atoms of silicon

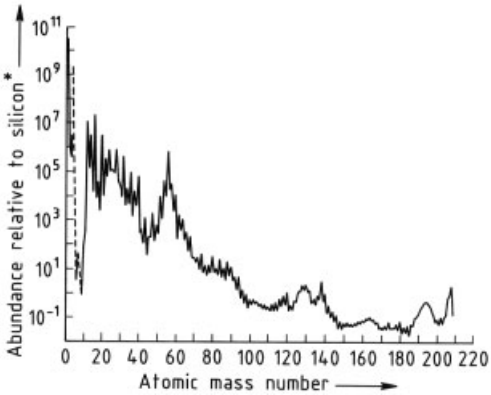

Figure 3. Abundance versus atomic mass number
∗ Number of atoms per 10^6 atoms of silicon

Figures 2 and 3, which plot relative abundance versus atomic number and atomic mass number, respectively, also show some general trends:

1) As atomic number increases, a near exponential decrease in abundance occurs from hydrogen through most of the second transition series (until ca. $Z = 42$ and $A = 100$). After that point, the decrease is more gradual and contains some peaks of relatively high abundance. Deuterium, beryllium, lithium, and boron do not follow the trend and are rare compared to other nuclides of similar mass.
2) Elements with even atomic numbers are more abundant than those with odd numbers in any given area of Figure 2 (except for Be).
3) Up to atomic number 40 (Ca), nuclides with a mass number that is a multiple of 4 are more abundant than their neighbors: $^{12}_{6}$C, $^{16}_{8}$O, $^{20}_{10}$Ne, $^{24}_{12}$Mg, $^{28}_{14}$Si, $^{32}_{16}$S, $^{36}_{18}$Ar, and $^{40}_{20}$Ca (again, except for Be).
4) Positions of even atomic mass number may be occupied by more than one nuclide. This is common above $A = 36$ and includes some mass numbers that correspond to three stable species, (e.g., $A = 96$). No stable species heavier than $^{14}_{7}$N has both an odd number of protons and an odd number of neutrons.
5) Positions of odd atomic mass numbers (even Z, odd N, and vice versa) are occupied by only one stable nuclide with the exception of $A = 113$ (isotopes of In and Cd). Other odd atomic mass number positions occupied by two species (87, 115, 123, and 187) include one radioactive nuclide. The odd atomic number species of each pair decays into an even atomic number species by electron emission (except $^{123}_{52}$Te, which decays by positron emission to $^{123}_{51}$Sb).
6) A large abundance peak occurs in the first transition series, centered at ^{56}Fe, where the abundance is about three orders of magnitude greater than that expected from the general trend. Other areas of locally higher abundance appear in Figure 3 at higher mass number.

1.3. Origins of Nuclides

Explanations for the trends in abundance of nuclides are found in the mechanisms of their creation. The origin of matter begins with the origin of the universe. One widely held theory of cosmological evolution is the big bang, proposed by ALPHER, BETHE, and GAMOW in 1948 [7], which postulates that matter was created from a suddenly expanding singularity. Later revisions of the original theory postulate that the initial form of matter was a mixture of leptons, including electrons, and quarks. The latter combined to form protons and neutrons. A neutron that is not confined in a nucleus decays into a proton, an electron, and an antineutrino with a half-life of ca. 11 min. Thus, the building blocks of the elements were created in the first hours of the universe. Conditions appear to have been appropriate for some helium to form in the initial cosmological fireball, but the majority of matter was in the form of hydrogen. This theory accounts for more than 99.9 % of the atoms that currently exist in the universe.

Nucleogenesis is the science that attempts to explain the origin of all nuclides found in the universe. For this purpose, processes that occur in the interior of stars must be examined. A comprehensive theory of nucleogenesis was presented by BURBIDGE, BURBIDGE, FOWLER, and HOYLE in 1957 [8]. This work is the basis for later discussions of elemental origin and abundance [9]. The theory postulates variations in the density of hydrogen in the primitive universe that allowed gravitational attraction to cause accretion of hydrogen and helium. As matter contracts under the force of gravity, the temperature rises until nuclear reactions begin in the core of what now can be called a star. The life cycle of a star involves a series of different types of nuclear processes, some of which are determined by the initial mass of materials that condense to form the star. As a star evolves, various processes become important as density and temperature change, and reactions that have already occurred provide material for subsequent reactions. The processes may be examined in three groups: (1) exothermic reactions, (2) neutron capture processes, and (3) miscellaneous processes both inside and outside of stars [10].

The initial reaction that occurs as the temperature of the contracting mass of material increases is hydrogen burning, an exothermic reaction. When the core of the star reaches ca. 10^7 K, the fusion of four hydrogen nuclei to form a helium nucleus becomes possible. Other nuclei, especially carbon and nitrogen, may catalyze the reaction. During this period, thermal pressure of radiation from the reaction halts further collapse of the star.

After a certain amount of helium has been produced, the thermal pressure is insufficient to offset gravitational attraction and contraction of the core is again possible, which increases the density and causes further heating. Above ca. 2×10^8 K, three helium nuclei may combine to form a carbon nucleus. This exothermic reaction involves the creation of a very short-lived intermediate ^{8}Be from two helium nuclei, which may then react with another helium nucleus to produce ^{12}C. Subsequent additions of helium nuclei result in the synthesis of oxygen, neon, and magnesium. As before, the stellar core eventually reaches the point of further gravitational collapse.

At ca. 5×10^8 K, carbon burning becomes possible. Two carbon nuclei may combine to form a magnesium nucleus, or to give lighter nuclei, hydrogen, and helium. Hydrogen and helium can participate in subsequent reactions to generate other nuclei in this mass range. Nuclear reactions similar to carbon burning may proceed in some stars that use other nuclei as fuels. After carbon is depleted, oxygen burning and then silicon burning are possible under appropriate conditions.

The next stage in stellar evolution depends on the mass of the star. If the stellar mass is less than about 1.5 solar masses (one solar mass is ca. 2×10^{30} kg), continued contraction of the core may occur, accompanied by increased heating. Above ca. 10^9 K, an exothermic process can occur that generates ^{16}O and ^{4}He from ^{20}Ne. This is the source of helium nuclei for the *alpha process,* which is responsible for the creation of nuclides having atomic mass numbers that are multiples of 4. This process stops at ^{40}Ca because the next member of the series, ^{44}Ti$*$ (the asterisk denotes an excited state), is unstable to decay by electron capture.

If the star is very massive (1.5 – 3.5 solar masses), its sequence is different from that described above. Hydrogen is consumed much more quickly, and helium reactions begin much

earlier. These conditions lead to instability, and the star disintegrates in a supernova explosion. Before this cataclysmic event, however, the temperature of the core rises above 3×10^9 K and many different nuclear reactions involving neutrons, protons, and gamma rays occur in great numbers. In a short time, a statistical equilibrium is established among the various nuclei and nuclear particles. Especially prominent as a result of this series of exothermic reactions, termed the *e-process,* are the nuclei of the elements with atomic numbers 22 – 29 (the first-row transition metals). Their abundances are determined by the magnitude of the binding energy associated with the various nuclei. The most stable nucleus with the highest binding energy per nucleon is ^{56}Fe, which is also the most abundant nuclide in this mass range (see Fig. 2).

Other nuclear reactions are associated with conditions immediately preceding a supernova explosion and are believed to be responsible for the production of some isotopes of the heavier elements. Proton capture may explain formation of the lighter isotopes of some heavier elements (e.g., ^{74}Se, ^{120}Te, and ^{196}Hg). These isotopes, which have a higher ratio of protons to neutrons than other isotopes, are usually present in only small amounts compared to the abundance of isotopes produced by means other than proton capture.

Neutron absorption processes are thought to be responsible for creation of the majority of heavy nuclides. Rapid neutron capture, the *r-process,* may also occur in unstable massive stars about to disintegrate in a supernova explosion. In a matter of seconds a large number of neutrons may be added to a nucleus. Absorption continues until the neutron capture cross section of the resulting nucleus diminishes to the point at which the unstable species has time to decay (after absorbing as many as 100 neutrons). Numerous beta emissions give rise to neutron-rich heavier isotopes such as ^{82}Se, ^{134}Xe, and ^{198}Pt. The lower mass parts of the double-peaked abundance maxima (see Fig. 3) near $A = 80$, 130, and 194 are the results of unstable nuclei with neutron quantities near the magic numbers undergoing these types of reaction. In addition, some of the lighter isotopes (^{36}S and ^{48}Ca) that have no other obvious method of creation are thought to be products of the r-process.

The more common neutron absorption process is slow neutron capture, the *s-process.* These reactions occur in the outer atmosphere of less massive stars over much longer time periods, hundreds or thousands of years for the s-process as opposed to seconds for the r-process. Again, neutron capture cross sections play an important role in determining which nuclides have higher relative abundance. In this case, because beta emission may occur before further neutron absorption, the most abundant species formed are those with magic numbers of neutrons. Thus, the higher mass parts of the double-peaked maxima (e.g., $^{90}_{40}$Zr, $^{138}_{56}$Ba, and $^{208}_{82}$Pb) are a result of the s-process.

The existence of some nuclides cannot be explained by the various processes that occur within stars. These are the nuclides with atomic numbers lower than carbon and, as shown in Figure 2, abnormally low abundance; i.e., ^{2}H, ^{3}He, ^{6}Li, ^{7}Li, ^{9}Be, ^{10}B, and ^{11}B. Although ^{2}H and ^{3}He nuclei are intermediate products of hydrogen burning, they are almost completely consumed in this process. None of the other nuclei result from the processes described above. In addition, except for ^{3}He, none of these nuclei are stable above ca. 5×10^6 K, which is well below the temperature at which nuclear synthesis begins in the core of stars. Initially, the unknown method of synthesis of these nuclides was termed the *x-process.* It is now believed that several different reactions, all occurring exterior to stars, are responsible for their creation.

The majority of ^{2}H and ^{3}He is thought to have been formed, along with the vastly greater amount of ^{1}H, early in the universe. Boron isotopes and ^{7}Li are predominantly the products of proton capture reactions in interstellar space as a result of the high temperature generated by supernova shock waves. Other nuclides in this group appear to be almost exclusively the products of interstellar spallation reactions. Such reactions may result when cosmic rays, a variety of atomic particles traveling at relativistic speeds, collide with other atoms in the interstellar gas. If the incident particle is of low mass (^{1}H or ^{4}He), it can enter the nucleus of a higher mass target atom. Several nucleons may be ejected immediately from the target, leaving a nucleus in a highly excited state. The energy of this species is redistributed, and a number of nucleons are emitted (commonly, 10 – 20), individually or in

clusters. In this manner, ^{6}Li and ^{9}Be, and possibly some ^{2}H and ^{3}He, are created.

The above discussions concern naturally occurring stable nuclides. All nuclides with atomic numbers higher than 83 are radioactive. The occurrence of these nuclides in nature is explained by the decay series of their long-lived precursors, and their abundance is determined by the half-lifes of the various species from which they derive. The precursors of the three known radioactive series, ^{232}Th for the thorium series, ^{238}U for the uranium series, and ^{235}U for the actinium series, are created by rapid neutron capture processes during supernova explosions. The half-life of ^{232}Th (1.4×10^{10} a) is commensurate with the age of the universe. The isotopes of uranium were probably synthesized in many supernova explosions starting around 10^{10} years ago, before the formation of the Solar System or the Earth (ca. $4.5-5\times10^{9}$ years ago). Studies of the formation and decay of these nuclides also give an approximate age to the Milky Way galaxy just slightly less than the age of the universe (ca. $1.5-2\times10^{10}$ years).

Thus, through a variety of processes, the synthesis of naturally occurring nuclides can be explained and their relative abundances accounted for. Some of the theoretical computations of these processes can be made quite accurately, especially in situations where statistical equilibrium applies. However, the great number of possible reactions and the inability to determine precise reaction rates or conditions, especially during supernova explosions, make computation difficult.

1.4. Discovery of Isotopes

The discovery of isotopes resulted from the study of radioactivity and radioactive elements. By the early 1900s, more than 30 different radioactive species were known. The chemical properties of these species, however, suggested that they should occupy no more than 12 positions in the periodic table. The chemical equivalence of groups of radioactive species such as ionium, thorium, and radiothorium (now known to be ^{230}Th, ^{232}Th, and ^{228}Th); radium and mesothorium-1 (^{226}Ra and ^{228}Ra); and ordinary lead (with an atomic mass of 207.2) and radium-D (^{208}Pb) was well established. Frederick Soddy generalized that all species occupying the same position in the periodic table are chemically identical and inseparable by chemical means and proposed that such species be called isotopes [11].

At that time, the atomic mass of an element was thought to be the major determinant of its chemical properties. Investigations showed that species with different atomic masses could have the same chemical properties. Conversely, other groups of radioelements were discovered that had essentially the same atomic mass but different properties (e.g., radium-B, radium-C, and radium-C' are now known to be ^{214}Pb, ^{214}Bi, and ^{214}Po, respectively). Such species are called *isobars* (→ Radionuclides).

In 1913, schemes were proposed for the various radioactive decay series in which emission of an alpha particle (a helium nucleus) moves a radionuclide two places in the periodic chart to a lower group, with an atomic mass number four units less, and emission of a beta particle (an electron) results in a change to the next highest group, with no change in the mass number. Emissions from radioactive species were also known to be charged particles, the alpha particle carrying two positive charges and the beta particle a single negative charge. The combination of these findings led to the idea that a relationship existed between the charge of the nucleus and its position in the periodic table, later to be quantified by the atomic number. The idea was first formally published in 1913 by A. van den Broek in connection with work on characteristic X-ray scattering of light elements. This determines the number of electrons in an atom, which is equal to the positive nuclear charge [11, p. 61]. A change to the next lowest group was later discovered to result from decay involving emission of a positron (the antiparticle of the electron) with the same mass and spin as the electron but a positive electrical charge. Evidence for the chemical equivalence of isotopes provided another fact in support of the theory.

Definite proof for the existence of isotopes was supplied by studies of the atomic mass of lead [12]. In 1913, radium-G, the final product in the uranium decay series, was found to be chemically identical to ordinary lead. The same conclusion was reached for the final products of the thorium and actinium series (thorium-D and actinium-D), but the differences in atomic

mass were recognized, even though the same element was being produced by radioactive decay. In 1914, SODDY and HYMAN showed that lead obtained from thorite, a mineral containing very little uranium, had an atomic mass almost one unit higher than ordinary lead. In the same year, LEMBERT and RICHARDS determined that the atomic mass of lead obtained from several uranium ores was almost one unit less than that of ordinary lead. These results were confirmed in independent studies and provided powerful support for the group displacement law and the theory of radioactive decay. More important, the results showed that a nonradioactive element could exist in several isotopic forms, all having identical chemical properties.

At the same time that the above studies were being conducted with high-mass elements, work in a seemingly unrelated field with lighter elements showed the phenomenon of isotopy to be universal. In 1912, J. J. THOMPSON was working in the area of *positive-ray analysis,* the study of charged particles produced by passing an electric arc through a gas at low pressure. A beam of positively charged molecules or atoms was known to be deflected from a straight line into a parabolic curve by an electric or magnetic field. The curve is determined by the charge-to-mass ratio *m/e* of the particles. In his analysis of the positive rays produced from a discharge tube containing neon, THOMPSON found not only the expected curve corresponding to an *m/e* of 20, but also a curve indicating a particle with an *m/e* of 22. The first explanation for this unexpected result was contamination by carbon dioxide (mass 44) with a double positive charge, but the curve still remained when the carbon dioxide was removed. THOMPSON concluded that neon must be a mixture of two gases with masses of 20 and 22. Attempts to separate the two gases were undertaken by F. W. ASTON [13]. Fractional distillation proved to be ineffective, but repeated diffusion through clay tubes gave two fractions whose atomic masses were higher and lower than that of ordinary neon. Positive-ray analysis of these fractions showed curves at *m/e* values of 20 and 22, but with appreciable differences in intensity compared to ordinary neon. The emission spectra of both fractions, however, were identical to that of ordinary neon. The implication was that neon consists of two isotopes of masses 20 and 22.

In 1919, ASTON resumed his research on neon. He constructed an apparatus, the *mass spectrograph,* which focused the particles produced in a discharge tube into lines called a *mass spectrum* with much higher accuracy than had been previously achieved. The mass spectrum of neon confirmed the presence of two components with atomic masses 20 and 22. Moreover, the ratio of the lighter to the heavier component was found to be approximately 10 : 1, indicating a mean atomic mass of 20.2. This was in excellent agreement with the known atomic mass of neon.

In 1920, ASTON examined the mass spectra of nineteen elements and discovered that nine of these were composed of two or more isotopes. Each isotope was found to have a mass that was very close to an integer. For those elements that exhibited isotopy, the relative abundances and masses of the isotopes combined to give atomic masses close to the known values. The elements with atomic masses close to integers, such as helium, carbon, nitrogen, oxygen, and fluorine, were found to be mononuclidic. (Later work showed that some of these elements do have more than one isotope, but that one of these is present in 99 % or higher abundance.) These discoveries led ASTON to formulate the "whole-number rule," which states that all atomic masses are very close to integers and fractional elemental masses are a result of the presence of isotopes, each of which has an approximately integral mass and contributes to the elemental mass in proportion to its relative abundance.

The universality of the whole-number rule was confirmed by experiments performed by A. J. DEMPSTER at the same time. DEMPSTER's version of the mass spectrograph allowed determination of the relative proportions as well as the masses of isotopes. DEMPSTER showed that lithium, magnesium, potassium, calcium, and zinc are composed of isotopes of approximately integral atomic masses. Thus, by 1921, the existence of isotopes was widely accepted. The fact that each nuclide has an approximately integral mass allows the idea that each atomic nucleus contains a definite number of neutrons and protons which each have a mass that is close to unity on the atomic mass scale.

1.5. Determination of Isotopic Masses and Abundance

The mass spectrographs developed by ASTON and DEMPSTER were the first instruments to discern ions having different mass-to-charge ratios with any degree of accuracy. Improvements in design and electronics afforded greater accuracy in determining atomic mass and abundance. Initially, the equipment was used almost exclusively to study isotopes, but since the late 1940s many other applications have been found. Modern equipment (the *mass spectrometer*) is used to identify the components of complex organic mixtures, elucidate molecular structures and chemical reaction pathways, and analyze the composition of solid-state samples [14]. High-resolution mass spectrometers have the ability to separate ions whose masses differ by less than one part in 30 000, and other specialized equipment can analyze extremely small samples of gas (as little as 10^{-12} mol of material may be required for analysis).

DEMPSTER's spectrometer used a method known as *directional focusing,* in which ions are accelerated by an electric potential and then forced by a magnetic field to travel in a circular path to impinge on a detector. If the position of the detector is fixed and the magnetic field strength is constant, the magnitude of the electric potential may be varied, causing ions with a particular *m/e* ratio to be collected by the detector. The mass can be calculated if the charge on the ions is assumed to be constant. Absolute values of the various factors need not be known if the apparatus is calibrated with samples of known mass.

Another type of mass spectrometer employs velocity focusing. Ionized particles are subjected simultaneously to opposing magnetic and electric fields. Those ions unaffected by the combined fields (i.e., having a characteristic velocity) pass into the next portion of the analyzer, where they are subjected to another magnetic field, deflected into circular paths determined by the *m/e* ratio of each particle, and impinge on different parts of the detector.

Modern spectrometers employ both types of focusing (double focusing). A schematic of a double-focusing mass spectrometer is shown in Figure 4. These take advantage of the line sharpness available from directional focusing and the simple relationship between particle mass and position of detection of the velocity focusing technique.

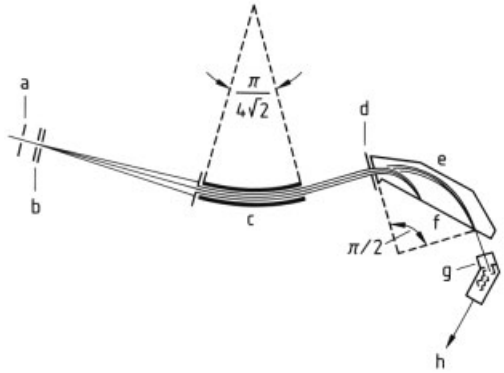

Figure 4. Double-focusing mass spectrometer
a) Sample electrode (or inlet gas system) and ion source; b) Accelerating electrostatic slits; c) Electrostatic analyzer; d) Intermediate image slit and beam monitor electrodes; e) Magnetic analyzer; f) Focal plane; g) Auxiliary collector and electron multiplier; h) Amplification and recording

Other types of instrument dispense with the magnetic field. The *time-of-flight* mass spectrometer generates ions in very small pulses, all with the same kinetic energy. The masses are determined by measuring the time required for them to strike the detector. *Radio-frequency* (rf) mass spectrometers employ a charged grid with alternating steady and variable electric potentials to selectively accelerate ions with an *m/e* that are in phase with the varying potential. The mass of the ions is proportional to the frequency and the grid spacing. The *quadrupole* mass spectrometer also uses variable radio frequency to separate ions. Equal but opposite electrical potentials, with both d.c. and rf components, are applied to the two pairs of conducting rods that make up the analyzer portion of the instrument. Ions injected along the axis of the region formed by the poles are made to oscillate by alternating the rf portion of the electric potential. Those ions having a *m/e* ratio in resonance with the frequency of the alternating potential travel the length of the analyzer and into the detector portion of the apparatus.

Double-focusing mass spectrometers have been adapted to determine the percent abundance of isotopes. The *isotope ratio spectrometer* employs two detectors, and ion currents from two ion beams (e.g., $^{10}BF_3$ and $^{11}BF_3$) are mea-

sured simultaneously. Determining the ratio of the intensities of the ion currents then gives the isotopic ratio. This method eliminates virtually all variables that can arise from internal and external sources. It is also used in analysis of tracer materials for isotopic labeling, medical research, and radioisotopic dating studies.

The relatively great difference in physical properties between the isotopes of hydrogen allows physical measurements to be used in determining the isotopic abundance of hydrogen and hydrogen-containing compounds [15]. Methods include density by the flotation method, near- and middle-infrared absorption spectroscopy, and refractive index.

1.6. Atomic Mass Scales

Atomic masses have always been of great importance, and values have been updated and refined continually since the early 1800s [16]. In 1874, the American Chemical Society published a table of atomic masses for 64 elements; all values were integers. Presently, atomic masses are given to a precision of one part in 10^4 for most elements and one part in 10^6 for some. The atomic masses of all but ten elements (Ti, Zn, Ge, Se, Mo, Sn, Sb, Te, Hg, and Tl) are determined partially or entirely by mass spectroscopy. Masses of the remaining elements are still determined by chemical means, and all except antimony have a large number of isotopes: Ti, Zn and Ge have five; Se and Tl have six; Mo and Hg have seven; Te has eight; Sn has ten; and Sb has two.

Before 1962, the atomic mass scale used in mass spectroscopy employed ^{16}O (99.76 % abundance) as the standard. It was assigned a mass of exactly 16.0000. At the same time, atomic masses determined by chemical techniques were based on atmospheric oxygen having an atomic mass of 16.0000. Since ordinary oxygen has naturally occurring isotopes of mass numbers 16, 17, and 18, the chemical standard had a mass of 16.0045 on the spectroscopic scale. In 1962, the International Union of Pure and Applied Chemistry and the International Union of Pure and Applied Physics adopted a new scale based on a value of 12.0000 for the atomic mass of ^{12}C.

Atomic masses of the 19 elements with only one nuclide (see Table 1) can be determined most precisely. Those elements with one predominant isotope (H, He, N, O, Ar, La, Ta, and U) may also be measured very precisely. Atomic masses of elements with a large number of isotopes and a broad distribution of abundance are the most difficult to measure. Mass spectroscopic methods rely on determining both the mass and the abundance of all isotopes of an element and combining these values to give a weighted average. Most often, chemical methods are used in conjunction with mass spectroscopic methods for such elements.

1.7. Variations in Abundance

For many elements, the atomic mass cannot be more precisely determined because of variations in isotopic abundance of the naturally occurring materials. One reason for uncertainty is exemplified by lead. Because three of the isotopes of lead (^{206}Pb, ^{207}Pb, and ^{208}Pb) are final products of the three radioactive decay series, the atomic mass of any specific sample of lead depends on the particular mixture of radioactive precursors in the ore from which it comes. Although the atomic mass of a specific sample can be determined quite precisely, the tabulated atomic mass must take into consideration the variation of almost two mass units that may occur in samples of different origin. Thirty other elements either have isotopes that are radioactive or are the products of radioactive decay, and are subject to the same uncertainty.

A different situation involves hydrogen, an element with one overwhelmingly abundant isotope whose atomic mass cannot be given more precisely. The significant difference in masses of deuterium and protium results in differences in the physical properties of their compounds. The variation in properties may allow enrichment of one isotope to occur through natural processes. In addition, the method used to prepare the element may result in a change in the isotopic ratio and, therefore, the atomic mass of hydrogen. Electrolysis of water breaks down D_2O and HDO more slowly than H_2O, and therefore gives hydrogen with a lower atomic mass than that obtained by other methods.

The atomic mass of sulfur can vary by as much as 0.01 mass units from source to source. Samples of naturally occurring boron are known

to vary by as much as 0.07 % in atomic mass. Other elements in this category include lithium, carbon, oxygen, argon, and copper.

At present, tables of atomic mass are footnoted to indicate those elements that may deviate from published values because of radioactive decay, natural or artificial enrichment activities, and place of origin.

Variation in the natural abundance of isotopes is an important part of geochemistry [17]. Determination of the relative differences in abundance among particular samples goes hand in hand with the study of chemical reactions in geological and biological systems. Variations in isotopes in natural settings can be significant because of the large quantities of material and long periods of time involved. In addition to the study of isotope ratios on Earth, the study of meterorites allows determination of the abundance of nuclides in the early Solar System and provides evidence for their origin (see Section 1.3). Rocks gathered from the Moon show it to have extremely low amounts of carbon and nitrogen. Oxygen-containing minerals taken from its surface have a much narrower range of isotopic ratios than found on Earth. Interplanetary probes have gathered information about the isotopic ratios of the atmospheres of Venus and Mars. Both planets have carbon and oxygen isotopic ratios close to that of Earth, whereas Mars has a large enrichment of ^{15}N (ca. 75 % greater than Earth) and Venus has an enormous (100-fold) relative enrichment of deuterium. Such studies aid in determination of the geological history of the Solar System.

2. Properties of Isotopes

Variation in the number of neutrons in the nucleus can result in different properties for isotopes of an element; these differences are known as *isotope effects.* Physical effects are noticed in the two general areas of mass effects and nuclear effects. Chemical differences are manifest in reactivity effects, including equilibrium and rate-determining (kinetic) effects as well as chemical bonding and structural effects. Physical differences among isotopes are generally the most pronounced, whereas chemical differences are more subtle. The effects of isotopy are particularly great for the lightest elements, which have larger relative mass differences between isotopes than the heavier elements. Although the differences that result from isotope effects are usually very small, they can be identified readily with the sophisticated techniques and sensitive instrumentation currently available and are useful in some applications.

2.1. Physical Differences

2.1.1. Mass Effects

Table 3. Physical properties of H_2, HD, and D_2

Property	H_2	HD	D_2
M_r	2.01588	3.0238	4.0318
mp, K	14.01		18.55
Heat of fusion, J/mol	117.1	159.4	196.6
bp, K	20.39	22.13	23.67
Heat of vaporization, J/mol	903	1074	1225
Density (liquid), g/cm^3	0.070		0.169
Triple point			
Temperature, K	13.96	16.0	18.73
Pressure, Pa	7300	12 800	17 700
Critical point			
Temperature, K	33.24	35.91	38.35
Pressure, kPa	1341	1534	1720
Volume, mL/mol	65.5	62.3	60.3
Dissociation energy, kJ/mol	431.6	435.1	439.2
Zero-point energy, kJ/mol	25.96	22.53	18.44

Because the existence of isotopes results from a difference in the number of neutrons in the nucleus, differences in mass-related phenomena among isotopically different elements and molecules are to be expected. Such phenomena include density, viscosity, thermal conductivity, diffusivity, and surface tension. The vibrational energy of a chemical bond depends on the masses of the atoms that make up the bond because the frequency of vibration is inversely proportional to the square of the mass. Isotopic differences change the bond vibrational energies and, therefore, the zero-point energies; this leads to variations in bond dissociation energy and enthalpies of fusion and vaporization. These variations result in changes in the melting and boiling points of isotopically different elements and compounds. For example, molecular deuterium (D_2) boils 3.12 °C higher than H_2, and heavy water (D_2O) boils 1.43 °C higher than H_2O. Table 3 presents the physical properties of H_2, D_2, and HD. Table 4 compares H_2O

and D_2O, which show large isotope effects because of the difference of a factor of two in the masses of hydrogen and deuterium and lesser effects because of the differences between ^{16}O and ^{18}O.

Table 4. Physical properties of H_2O and D_2O

Property	H_2O	D_2O
M_r	18.016	20.028
Heat of formation (25 °C), kJ/mol	285.83	294.61
mp, °C	0.0	3.813
Heat of fusion, kJ/mol	6.008	6.280
bp, °C	100.0	101.43
Heat of vaporization, kJ/mol	44.01	45.396
Vapor pressure, Pa		
20 °C	2340	2030
100 °C	1013	961.8
Critical Point		
Temperature, °C	374.2	371.5
Pressure, kPa	22 072	22 080
Density, g/cm^3	0.314	0.363
Heat capacity (25 °C), J mol^{-1} K^{-1}	76.022	84.725
Density (20 °C), g/cm^3	0.9982	1.105
Temperature of maximum density, °C	3.98	11.23
Viscosity (20 °C), mPa · s	1.005	1.260
Dipole moment, D	1.86	1.87
Refractive index, n_D^{25}	1.3325	1.32796
Solubility of NaCl (25 °C), g/100 g H_2O	35.92	30.56
Conductivity of 0.01 M KCl (25 °C), cm^2 Ω^{-1} mol^{-1}	143.3	117
Ion product, (mol/L)^2	1.008×10^{-14}	1.38×10^{-15}
Dielectric constant (25 °C)	78.25	78.54

Absorption spectra of molecules exhibit isotope effects because they result from transitions between vibrational and rotational energy levels in molecules (coupled with a nuclear spin component). Vibrational energy levels depend on the masses of the nuclides involved in a bond, while rotational energy levels are related to the center of mass of a molecule. Both quantities change with isotope substitution. Figure 5 shows part of the infrared spectrum of hydrogen chloride, with 12 pairs of rotational bands superimposed on a single vibrational band. In each pair of rotational bands, the transition at higher energy is related to ^{37}Cl and at lower energy to ^{35}Cl. Measurements of the *isotopic line shift* can be made very accurately in the microwave region of the spectrum and are a basis for determining isotopic ratio and abundance [18].

Electronic energy levels are also affected by isotopic differences. Optical emission spectra are dependent on the reduced mass of the species involved. The wave number, ν (the reciprocal wavelength) of the light emitted is given by

$$\bar{\nu}=R_\infty Z^2 \frac{\mu}{m_e}\left[\frac{1}{n_1^2}-\frac{1}{n_2^2}\right] \quad (1)$$

where R is the Rydberg constant, Z is the atomic number, m_e is the mass of the electron, n_1 and n_2 are the main quantum numbers of the lower and higher electron orbitals of the transition, and μ is the reduced mass, given by

$$1/\mu = 1/m_e + 1/m_n \quad (2)$$

where m_n is the mass of the nucleus. Thus, the wave number of the emission increases with increasing mass of the nucleus, i.e., emission occurs at higher energy. The isotopes of hydrogen give the largest line shifts; the H_a emission line is 0.179 nm higher for D_2 than H_2. This difference was predicted theoretically and confirmed experimentally by Urey in 1932, providing the first evidence for the existence of deuterium [19]. Differences in reduced mass become less significant with heavier elements. However, changes in nuclear size and shape also affect transition energies, and significant effects from these changes can be found between isotopes of heavier elements. For ^{235}U and ^{238}U, the isotopic shift is 0.025 nm (ca. 14 % of the hydrogen-deuterium shift) even though the reduced mass difference of these uranium isotopes is only ca. 1 % that of the hydrogen isotopes. An element with many isotopes may exhibit more complicated isotope effects in its spectrum. Figure 6 shows the hyperfine splitting of one of the resonance lines in the emission spectrum of mercury, with ten peaks at 40 °C and five at 15 °C. These multiple emission lines are attributed to the seven isotopes that make up the element, with certain isotopes having more than one emission in the region shown. Overlap of peaks occurs at lower temperature.

Mass-related effects are also found in the behavior of high-temperature oxide superconductors [20]; e.g., $La_{1.85}Sr_{0.15}CuO_4$, which has a transition temperature T_c to the superconducting state of ca. 37 K, and $YBa_2Cu_3O_7$, which has a T_c of ca. 92 K. According to theory, the T_c of a material depends on the frequency of lattice vibrations. Thus, an isotope effect is expected

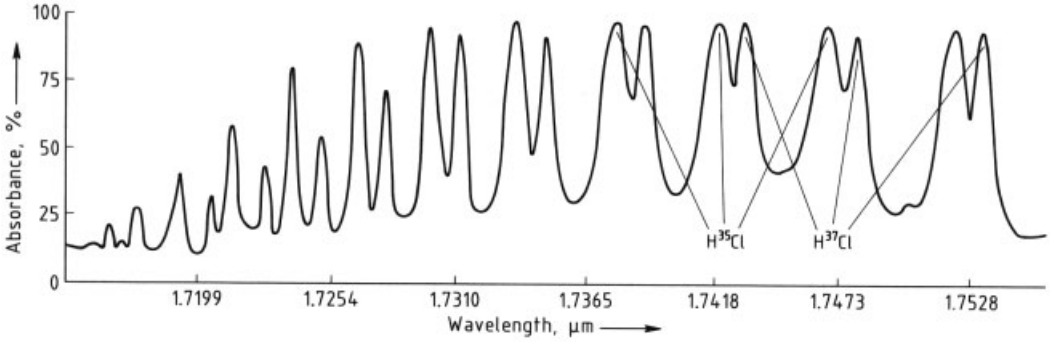

Figure 5. Portion of the infrared absorption spectrum of HCl

because the vibrational frequency is related to the mass of the nuclei that form the lattice. Evidence [21] has confirmed that high-temperature oxide species exhibit an isotope effect when ^{18}O is substituted for ^{16}O, the T_c is lowered by ca. 1 K in both cases (this may be reversed by resubstitution of ^{16}O). However, the shift in T_c calculated theoretically is much higher than the experimentally determined value; it is twice as great when calculated for the lanthanum compound and five times greater for the yttrium species.

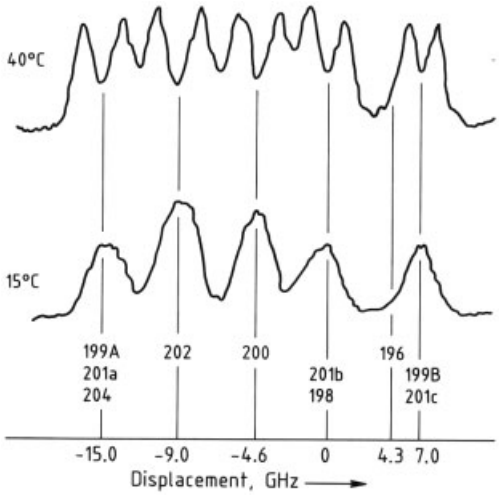

Figure 6. Portion of the emission spectrum of mercury

2.1.2. Nuclear Effects

In addition to its mass, other properties of the nucleus are affected by the number of neutrons. The stability of the nucleus depends on the nucleon composition. Nuclear theory suggests that protons pair with protons and neutrons with neutrons. The most stable nuclei are those with even numbers of both nucleons. Unpaired protons or neutrons lead to lower stability and may result in radioactive decay, which produces a more stable nuclear configuration. An example is ^{40}K, a radioactive species with 19 protons and 21 neutrons. This isotope decays by negative beta (β^-) emission to form ^{40}Ca, with 20 protons and 20 neutrons; the unpaired neutron is converted to a proton, an electron, and a neutrino. The species also decays by either positive beta (β^+, positron) emission or orbital electron capture to form ^{40}Ar, with 18 protons and 22 neutrons; the unpaired proton is converted to a neutron, a positron, and a neutrino in the first case, and to a neutron and a neutrino in the second case. Pairing of nucleons also determines nuclear spin I. Nonzero nuclear spin results in a magnetic moment for a nucleus which may interact with electric and magnetic fields. This property is exploited in chemical analysis by nuclear magnetic resonance (see Section 3.1).

The total number of nucleons has another effect on the absorption spectra of some molecules in addition to the isotope shift caused by mass differences. Nuclei with an odd atomic mass number are governed by Fermi – Dirac quantum statistics and are referred to as *fermions,* whereas nuclei with an even atomic mass number obey Bose – Einstein quantum statistics and are known as *bosons* [22]. The different quantum statistical treatments determine which specific rotational transitions of a diatomic molecule are allowed and which are forbidden. Thus, the infrared spectra of $^{16}O_2$ and $^{17}O_2$ show different

relative intensities for the rotational transition lines. The nuclear spin can affect other properties, as exemplified by the fact that the boson ^{4}He becomes a superfluid at low temperature, whereas the fermion ^{3}He does not exhibit superfluidity. In the hydrogen molecule, each of the two nuclei (protons) has a spin of 1/2. The spins may be paired or unpaired, resulting in molecules of different energies with measurable property differences. When the spins are paired ($I = 0$), the species is known as *para hydrogen.* When the spins are unpaired ($I = 1$), the molecule is known as *ortho hydrogen.* Under normal conditions, the ratio of ortho to para is 3 : 1. Although not strictly an isotope effect, the ortho and para molecules do show differences in physical properties.

2.2. Chemical Differences

2.2.1. Equilibrium Effects

In a chemical reaction involving molecules with a common element (e.g., AX and BX), the common element X may be exchanged between the molecules. If the common element consists of isotopes (X and X∗), these may be involved in the exchange reaction:

$$\mathrm{AX} + \mathrm{BX}* \rightleftharpoons \mathrm{AX}* + \mathrm{BX} \tag{3}$$

The equilibrium constant K for the reaction is approximated by the equation

$$K = \frac{[\mathrm{AX}^*]\,[\mathrm{BX}]}{[\mathrm{AX}]\,[\mathrm{BX}^*]} \tag{4}$$

If the equilibrium constant for this reaction deviates from unity, a noticeable isotope effect occurs. A value higher than 1 favors the forward reaction, and the isotope X∗ is preferentially associated with the AX∗ species; a value lower than 1 favors the reverse reaction and results in enrichment of the X∗ isotope in the form of the BX∗ molecule.

Statistical thermodynamics provides the explanation for this phenomenon and allows calculation of the magnitude of the effect [4, app. B]. The equilibrium constant may also be written in terms of partition functions:

$$K = \frac{Q_{\mathrm{AX}^*} Q_{\mathrm{BX}}}{Q_{\mathrm{AX}} Q_{\mathrm{BX}^*}} \tag{5}$$

where Q is the grand partition function for each molecule and includes all possible energy states of that molecule. These energy states may be represented by the individual partition functions for the various forms of energy: translation (Q_{tr}), rotation (Q_{rot}), vibration (Q_{vib}), electron movement (Q_{el}), and nuclear spin (Q_{nsp}). The grand partition function is, therefore,

$$Q = Q_{\mathrm{tr}} Q_{\mathrm{rot}} Q_{\mathrm{vib}} Q_{\mathrm{el}} Q_{\mathrm{nsp}} \tag{6}$$

All these partition functions are important in explaining the physical differences among isotopes, but only the first three influence the chemical equilibrium isotope effect. The equation describing the translational energy of a molecule includes terms for the mass and the molar volume of the molecule, both of which are affected by isotopic differences. Equations for the rotational and vibrational energies both include a term for the reduced mass of the molecule. Thus, the mass of the molecule enters into the partition functions for all three modes of molecular motion. The largest isotope effect is found in the vibrational energy levels, which are important in the transition states of chemical reactions. Temperature is also a factor in the translational energy partition function, which indicates that isotopic exchange reactions are temperature dependent. Thus, the mass difference between isotopes may alter the exchange reaction to favor the association of one isotope with a particular compound in a given chemical reaction.

An example of the isotope effect is the exchange reaction of oxygen isotopes between carbonate ions and water:

$$\mathrm{C}^{16}\mathrm{O}_3^{2-} + \mathrm{H}_2^{18}\mathrm{O} \rightleftharpoons \mathrm{C}^{18}\mathrm{O}^{16}\mathrm{O}_2^{2-} + \mathrm{H}_2^{16}\mathrm{O} \tag{7}$$

The equilibrium constant K for this reaction is 1.112 at 0 °C. The isotope effect is, therefore, 11.2 % in this case and results in the enrichment of ^{18}O in the carbonate that precipitates from solution. Since the equilibrium constant is temperature dependent, the ratio of ^{18}O to ^{16}O is also a function of temperature. Thus, the analysis of precipitated carbonates can be used to obtain information on geological temperature. Isotope effects are common in biological systems and may be quite significant for the lighter elements, especially hydrogen. Certain algae and bacteria

prefer deuterium to hydrogen and show enrichment factors up to 4. Other microorganisms survive in environments where the isotopic abundance differs markedly from natural abundance; this is not the case for higher organisms. Some reactions, such as the exchange of hydrogen isotopes between water and hydrogen sulfide (see Section 4.2.6), exhibit sufficiently large isotope effects for them to be used in industrial processes to enrich isotopes. Note that even when these examples are cited as systems in which the effects are quite large, they are still small by normal chemical reaction criteria; for most elements, isotope effects are small enough to be neglected in everyday experimental or commercial systems. Fortunately, this means that isotopic tracers can be used in chemical systems without undue complications.

2.2.2. Kinetic Effects

In a simple chemical reaction, the rate of reaction is proportional to the concentration of reactants. For isotopically substituted species [23], these reactions might be

$$\begin{aligned} &\mathrm{A+BC \longrightarrow AB+C} \\ &\mathrm{A+B{*}C \longrightarrow AB{*}+C} \end{aligned} \tag{8}$$

and the expressions for the reactions rates are

$$\begin{aligned} \frac{\mathrm{d\,[A]}}{\mathrm{d}t} &= K\,[\mathrm{A}]\,[\mathrm{BC}] \\ \frac{\mathrm{d\,[A]}}{\mathrm{d}t} &= K'\,[\mathrm{A}]\,[\mathrm{B^*C}] \end{aligned} \tag{9}$$

where K and K' are rate constants for the reactions. Each reaction is assumed to take place by way of a transition state known as the activated complex:

$$\begin{aligned} &\mathrm{A+BC \rightleftharpoons ABC^{\#} \longrightarrow AB+C} \\ &\mathrm{A+B{*}C \rightleftharpoons AB{*}C^{\#} \rightleftharpoons AB{*}+C} \end{aligned} \tag{10}$$

where # indicates a short-lived transition state. The activated complex is a combination of the reacting species, and possesses its own bonding scheme and vibrational frequencies. The transition state lies at an energy above that of the reactants; the difference is the activation energy. Once formed, the complex may decompose either into the products or back to the reactants. Activated complex theory states that an equilibrium is established between the reactants and the activated complex and that the rate constant for the overall reactions is proportional to the equilibrium constants for the formation of the activated complex. The rate constants for the isotopically substituted species are

$$\begin{aligned} k &= (k_{\mathrm{B}}\,T/h)\,K^{\#} \\ k' &= (k_{\mathrm{B}}\,T/h)\,K'^{\#} \end{aligned} \tag{11}$$

where k_{B} is the Boltzmann constant, $K^{\#}$ and $K'^{\#}$ are equilibrium constants for the formation of the transition states, h is Planck's constant, and $k_{\mathrm{B}}\,T/h$ is essentially the average lifetime of the activated complex. The equilibrium constant for the activated complex may be represented as a ratio of grand partition functions for the reactants and, in these cases, the activated complexes. The form of the partition function for the activated complex is slightly different from those of the reactants because of restrictions in its degrees of freedom. The motion, and thus the available energy states, of the activated complex are limited to the reaction coordinate, i.e., by the nature of the reacting species and how they combine in the transition state. Nevertheless, three of the four terms involved in the grand partition function depend on vibrational frequencies, which are mass dependent. For a kinetic isotope effect to exist, the vibrational frequencies of the activated complex must differ from those of the reactants and must involve motion of the isotopically different atoms B or B∗. If these conditions are not met, only an equilibrium isotope effect is seen.

Kinetic isotope effects usually result in a reduction in reaction rate when heavier nuclides are substituted for lighter ones. Kinetic effects tend to be larger than equilibrium effects because the difference in activation energies of the reactions is usually greater than the difference in the enthalpies of the reactions for isotopically substituted molecules.

The kinetic isotope effect has proved useful in elucidating reaction mechanisms and identifying the rate-determining step for some organic reactions. The mechanism for electrophilic aromatic substitution was shown to be a two-step reaction involving an intermediate. Other uses could take into consideration the slower reactions that result from substitution of heavier isotopes for lighter ones. Deuteration (replacing

hydrogen by deuterium) of lubricating oils has been suggested to reduce the rate of oxidation of these materials. The effect is also of interest in biology. The fact that higher organisms cannot survive in environments of unusual isotopic abundances is due to the reduction in rates of some metabolically important reactions.

2.2.3. Bonding Effects

Variations in isotopic composition also influence chemical bonding and crystal structure [17, p. 15]. Although these effects are mass related to some extent (because vibrational bond frequencies are involved), the ionic nature of the counterion in some compounds appears to be the most important factor in determining the preference for one isotope over another in a specific bonding situation; the degree of order in the crystal structure plays a similar role in bulk solids. For example, the small, highly charged Si^{4+} cation has been found to be associated with a relatively higher ^{18}O concentration compared to other ionic species. Conversely, the larger Fe^{2+} cation is always found with a lower ^{18}O concentration than most other cations. Sulfur follows the same pattern, with the heavier ^{34}S associating more strongly with lighter, smaller cations such as Zn^{2+} as opposed to heavier cations such as Pb^{2+}. Crystal structure is of secondary importance to chemical bonding considerations, but definite trends can be cited in which the heavier of two isotopes is associated preferentially with more closely packed and well-ordered structures. Large effects may be noticed in specific cases, such as diamond and graphite, where the more ordered diamond crystal structure incorporates a higher fraction of ^{13}C than graphite.

2.3. Isotope Effects in Nature

The science of geochemistry has placed much emphasis on natural variations in isotopic abundance and the processes that cause fractionation of isotopes in geological and biological systems [18]. Many of the reaction mechanisms in these systems have been elucidated by studying the enrichment of a particular isotope and by theoretical calculations of the isotope effects associated with the system. Several generalizations can be made with respect to the natural variations of relative isotopic abundance. First, significant fractionation occurs only when relative mass differences between isotopes are large. Elements up to a mass number of ca. 40 exhibit measurable abundance variations, with only isolated examples at higher atomic mass. Second, equilibrium isotope effects are most important in geological systems, whereas kinetic effects predominate in biological systems, particularly in oxidation – reduction reactions. Third, the elements with the widest variations in isotopic composition are those that form compounds which are stable over a wide temperature range, because both equilibrium processes and kinetically controlled chemical reactions are temperature dependent. Also, elements that form compounds in different phases (solid, liquid, and gaseous) exhibit greater isotopic variations; in general, heavier isotopes exist in higher oxidation states in the solid state.

In geochemical studies, relative abundance is presented as the ratio of the amount of one isotope in a specific compound to the amount of that isotope in another compound. Measurement of absolute isotope abundance, a more difficult task, is very important because these values form the basis for calculations of the relative differences on a universal scale for each element [17, p. 23]. Values of absolute abundance are given as δ, in parts per thousand (‰), which compares the isotopic ratio of a sample with that of an accepted standard:

$$\delta\,(‰) = \frac{R_{\text{sample}} - R_{\text{standard}}}{R_{\text{standard}}} \times 1000 \qquad (12)$$

where R is the ratio of the abundance of the less common to the more common isotope. The primary standards, given in Table 5 for the more common elements, have been widely accepted. Although an individual laboratory may use its own internal standards for experimental work, values published in the literature are converted to the scale of one of these primary standards. In some cases (e.g., PDB for determining carbon isotope ratios), samples of the actual standard are no longer available and secondary standards are used based on the known ratio of the primary standard and produced by the National Bureau of Standards in the United States. The ranges of δ values given in the following examples are all referenced to primary standards.

Table 5. Worldwide standards in use for isotopic composition of some elements

Standard	Source	Isotope ratio	Value, ppm
SMOW (Standard Mean Ocean Water)	International Atomic Energy Agency	D/H	155.76
		$^{18}O/^{16}O$	2 005.20
		$^{17}O/^{16}O$	373
PDB	belemnite from Pee Dee formation, South Carolina	$^{13}C/^{12}C$	11 237.2
		$^{18}O/^{16}O$	2 067.1
		$^{17}O/^{16}O$	379
Air nitrogen	atmosphere	$^{15}N/^{14}N$	3 676.5
CDT	troilite from Canyon Diablo meteorite	$^{34}S/^{32}S$	45 004.5

Hydrogen shows large isotope effects. The range of D/H ratios is broadest in water trapped in meteorites, with δ values from +50 to -450 ‰ relative to Standard Mean Ocean Water (SMOW). Water on Earth varies between 0 and -80 ‰. Vapor-pressure and freezing-point differences between ordinary water and heavy water result in an increase of deuterium concentration in ice in the liquid – solid system, and in water in the vapor – liquid system. Equilibrium exchange reactions lead to significant enrichment of deuterium in the water vapor – methane, water vapor – hydrogen, and water vapor – hydrogen sulfide systems. Hydrous minerals also show isotope effects; δ values for sedimentary and igneous rocks on Earth lie in the range of -20 to -150 ‰. These systems are very complicated in their behavior. More than one type of chemically reactive site is usually present in these materials, and the effects of temperature and the presence or absence of water can contribute differently to the overall isotope fractionation. The isotope fractionation factor is defined as the D/H ratio in one substance divided by the D/H ratio in another. Figure 7 shows how the D/H isotope fractionation factor α^{D} changes with temperature in hydrated-mineral – water systems. Some clays and minerals act as semipermeable membranes to the passage of water, an effect known as *ultrafiltration,* and can cause enrichment of deuterium in the residual water because of their preference for absorbing the heavier isotope. Fractionation can also occur between the water in the hydration sphere of a mineral and the bulk water surrounding it. Biological systems are also involved in isotope enrichment. Hydrogen and methane produced as metabolic byproducts in some bacteria are enriched in the light isotope of hydrogen by a factor of up to 6, compared with that expected in the normal hydrogen – water system.

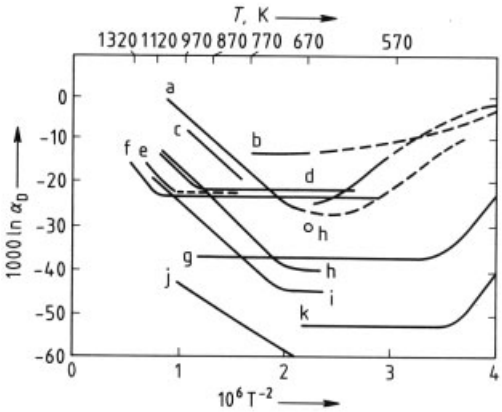

Figure 7. Isotope fractionation vs. temperature for some minerals
a) Muscovite; b) Serpentine; c) Phlogopite; d) Tremolite; e) Hornblende (B); f) Hornblende (A); g) Epidote – Chinozoisite; h) Actinolite; i) Biotite; j) Zoisite; k) Boehmite

Oxygen is also fractionated by equilibrium effects in many of these systems. Because of the smaller mass difference between oxygen isotopes, the range of δ values is smaller than that of hydrogen. In meteoric waters, the range for $^{18}O/^{16}O$ is + 10 to – 40 ‰ (relative to SMOW). The heavy isotopes are concentrated in two-phase water systems and hydration sphere – bulk water systems in the same manner as deuterium. Abundance ratios in rocks vary between + 40 and -5 ‰. Bond strengths play an important role in the fractionation of ^{18}O between minerals, as in the previous examples of silicon and iron. When a small quantity of water is in contact with a large amount of rock, the ^{18}O concentration in the water increases. However, when a small quantity of rock is in contact with a large

amount of water, the ^{18}O concentration of the rock increases. Kinetic isotope effects are evident in other oxygen-containing systems. The concentration of ^{18}O in the atmosphere cannot be explained by a simple equilibrium between oxygen in air and water. Photosynthesis results in enrichment of ^{18}O in the oxygen produced, when compared to the water from which it is derived. Respiration by plants and animals returns oxygen to the atmosphere with a higher ^{18}O concentration than that taken in.

Carbon is found on Earth in two major reservoirs: organic matter and sedimentary carbonates. In organic compounds, fractionation results in the enrichment of ^{12}C, with $^{13}C/^{12}C$ ratios ranging from -5 to -35 ‰, relative to the Belemnitella (PDB) standard; the figures are higher in this range for marine organisms and lower for sedimentary organic materials (e.g., petroleum). Kinetic isotope effects are also found in photosynthesis. The lighter isotope of carbon, ^{12}C, is concentrated in organic matter, and thus depleted in atmospheric carbon dioxide, in a two-step process. The uptake and diffusion of carbon dioxide are subject to a reversible equilibrium effect which produces a small enrichment of ^{12}C. The first carbon-dioxide-fixing carboxylation reaction of the photosynthesis sequence is an irreversible reaction with a large enrichment factor. As is typical in biological systems, the reaction rate is faster for molecules with lighter isotopes than for those with heavier isotopes. The geological form of carbon is enriched in ^{13}C mainly as a result of the equilibrium isotope effects in the $CaCO_3-CO_2-H_2O$ system:

$$Ca^{2+} + 2\,HCO_3^- \rightleftharpoons CaCO_3 + H_2O + CO_2 \qquad (13)$$

Some enrichment occurs as a result of the hydration and dehydration of carbon dioxide, but the major fractionation reaction is the precipitation of calcium carbonate; δ values between +5 and -15 ‰ are found. Kinetic effects are also seen in supersaturated solutions when equilibration does not take place between solid and solution before precipitation. Interactions between the organic and carbonate reservoirs allow a wide variety of isotope ratios to be found in nature.

Most of the nitrogen (>99 %) found on Earth is in the atmosphere in the form of molecular nitrogen (N_2) and has an essentially constant composition. Many biochemical reactions involving nitrogen (e.g., fixation, nitrification, and denitrification) have a small fractionation factor which results in an increase in the concentration of ^{15}N in organic materials. Nitrogen associated with natural gas deposits has been shown to have been depleted in ^{15}N by diffusional migration, the depletion increasing with increasing migration distance. The widest range of abundance ratios for nitrogen on earth is found in natural gas samples, with δ values from +30 to -25 ‰.

Kinetic isotope effects are significant in the absorption of boron compounds by clay minerals; ^{10}B is enriched by this process. Samples of sulfur compounds have been found to vary by as much as 150 ‰. Sulfur is enriched in ^{32}S by bacterial reduction of sulfates to hydrogen sulfide, for which kinetic effects are prominent. Ultrafiltration also results in fractionation because ^{32}S is absorbed into clays at a higher concentration than the heavier isotopes.

Natural exchange reactions and chemical bonding effects are found in a range of metallic sulfides; as with oxygen, the heavier isotopes of sulfur are associated preferentially with smaller, lighter cations. Selenium is also fractionated by bacterial reduction of selenates because compounds containing ^{76}Se are reduced faster than those containing ^{82}Se. Silicon shows very small effects (a range of ca. 6 ‰) since it is almost exclusively associated with oxygen in silicates and does not undergo oxidation–reduction reactions. Also, no gaseous and few liquid silicon compounds occur naturally. Chlorine is fractionated by equilibrium processes, mainly diffusion of liquid and gaseous compounds, resulting in enrichment of ^{35}Cl as a function of distance. Calcium shows very small equilibrium effects and no differences between geological and biological systems.

3. Uses of Isotopes

The differences in physical and chemical properties of individual isotopes are the basis for many applications. Isotopes of naturally occuring radioactive elements were first exploited for military purposes. Later, the principles of atomic energy were used to generate electricity and provide power plants for ships. Uses were also

found for a variety of isotopes in research. The discovery of methods for monitoring nuclear magnetic properties provided powerful tools for molecular analysis. The availability of quantities of separated or enriched isotopes allowed isotopic substitution studies of structures and mechanisms in chemistry. Physicists used the specific properties of individual isotopes to study nuclear phenomena and to detect cosmological events. In addition to research, applications have broadened to include medical and industrial uses. Radioactive isotopes are employed in the treatment of diseases, most notably cancer. Stable isotopes are used as precursors of the radioactive agents or as tracers in diagnosis. Many industrial uses of isotopes are in the developmental or limited-use stage. Promising techniques include isotopic tracing of products and waste materials. Environmentally, isotopes may be used to track the movement of pollutants and in hydrogeological investigations.

3.1. Nuclear Energy (→ Nuclear Technology)

In the late 1930s researchers discovered that naturally occurring uranium may undergo radioactive decay by fission, either spontaneously or induced by neutron capture. Fission is a nuclear transformation in which the nucleus splits into two relatively massive nuclei while emitting one or more neutrons. Fission is accompanied by release of a large amount of energy. All nuclides with mass number higher than 230 undergo fission reactions with half-lives that decrease with increasing mass number. Approximately 250 different species have been identified as fission products of uranium, and the average energy released is 195 MeV per atom. Self-sustaining chain reactions can be initiated if at least one neutron produced in a fission reaction induces decay of another atom.

The first use of fissile isotopes was in the construction of atomic bombs in the Manhattan Project during World War II [24]. Shortly thereafter, methods were developed to use the tremendous heat and pressure developed by atomic explosions as triggers for fusion, or thermonuclear, devices. Fusion reactions involve combination of light nuclei to form heavier species with release of even larger quantities of energy per atom than fission reactions. Specific isotopes are also important in fusion reactions. One of the most accessible fusion reaction (i.e., requiring the lowest temperature and pressure) is combination of deuterium and tritium. Tritium occurs only in trace amounts on Earth, but may be synthesized from ^{6}Li by neutron capture ($^{6}\mathrm{Li} + \mathrm{n} \rightarrow {}^{4}\mathrm{He} + {}^{3}\mathrm{H}$). A ^{6}Li-containing target is irradiated in a nuclear reactor or an accelerator to produce tritium, which is then extracted from the target. Lithium deuteride $^{6}\mathrm{Li}^{2}\mathrm{H}$ has been suggested as a fuel for fusion weapons as well [25]. A fission trigger creates an intense neutron flux that converts the Li to tritium, which fuses with the deuterium.

Fission reactions in nuclear weapons are uncontrolled, but by limiting the extent of the reaction this energy source can be harnessed in nuclear reactors for useful purposes such as generating electricity or providing high neutron fluxes for research and the production of valuable radioactive isotopes. Here again, the nuclear properties of specific nuclides are exploited. To sustain a controlled chain reaction, the energy of neutrons produced by fission must be reduced through collisions with other nuclei, a process called moderation. A good moderator has a mass close to that of the neutron to maximize energy loss per collision and a very small neutron capture cross section to minimize unwanted nuclear reactions. Deuterium is by far the best moderator; more than 80 times better than hydrogen and 30 times better than ^{12}C or ^{18}O [26]. Heavy water ($^{2}\mathrm{H}_2\mathrm{O}$) is almost as good as deuterium and has the distinct advantage of being a nonflammable liquid. Heavy water is also an excellent neutron reflector, and thus decreases the number of neutrons that escape the reactor core without participating in fission reactions. For fine control of reactor operations, neutron-absorbing isotopes are used to adjust the number of moderated neutrons available for inducing fission. Species with high thermal neutron capture cross sections include ^{10}B and ^{113}Cd. Other isotopes with very small cross sections are used to clad the nuclear fuel to protect it from the harsh environment in the reactor. Aluminum, a mononuclidic element, is widely used for this purpose.

3.2. Research

3.2.1. Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) spectroscopy was discovered in 1946 by PURCELL and BLOCH [27]. In addition to the charge on the nucleus, about half of the naturally occurring nuclei possess spin angular momentum I as a result of the composition of the nucleus. Those nuclei with an even number of both protons and neutrons have a nuclear spin of zero. Other even atomic mass number nuclei with an odd number of both protons and neutrons (the only stable nuclides in this category are ^{2}H, ^{6}Li, ^{10}B, and ^{14}N) have integral spins (1, 2, 3, etc.). Nuclei with odd atomic mass numbers have half-integral spins (1/2, 3/2, 5/2, etc.). The multiplicity of the spin is determined by the *nuclear quadropole moment,* a measure of the charge distribution in the nucleus, and increases as the charge distribution deviates from spherical symmetry. The spinning charged nucleus generates a magnetic field and, therefore, has a magnetic moment m which is different for each nuclide. Table 6 lists some nuclides with nonzero nuclear spin and includes the magnetic moments of such species.

Nuclear magnetic resonance spectroscopy (see → Nuclear Magnetic Resonance and Electron Spin Spectroscopy) is based on the fact that the magnetic field generated by the nucleus is influenced by an external magnetic field. A distribution of allowed orientations with respect to the external field occurs in a group of the same nuclei. Changes in the distribution can be induced by an external radio-frequency energy input and monitored. The frequency causing a change in the distribution is called the *resonance frequency.* Because of the different magnetic moments possessed by different nuclides, the combination of magnetic field strength and radio-frequency energy may be varied to influence and monitor specific nuclides. Most often, frequency is held constant and magnetic field strength is varied to produce an NMR spectrum. The usefulness of this technique comes from its ability to detect subtle changes in the magnetic moment of a nucleus, and thus the resonance frequency, that are caused by the local electron environment in which the nucleus is found and its interaction with other nuclei in the vicinity. The difference between the resonance of any particular nucleus and that of a nucleus in a standard reference compound is referred to as the *chemical shift. Coupling,* interaction of the magnetic moments of two or more nuclei, is transmitted through the electrons that make up the chemical bonds between atoms and results in splitting the single resonance into more than one signal. The magnitude of splitting is called the *coupling constant.* The multiplicity of splitting is a result of the number of equivalent nuclei that interact with a given nucleus and their spins. A particular resonance signal may be split repeatedly by nonequivalent nuclei.

Table 6. Some stable nuclides with nonzero nuclear spin

Nuclide	Spin (I)	Magnetic moment, μ_N*
^{1}H	1/2	+2.793
^{2}H	1	+0.857
^{6}Li	1	+0.822
^{7}Li	3/2	+3.256
^{9}Be	3/2	−1.178
^{10}B	3	+1.801
^{11}B	3/2	+2.689
^{13}C	1/2	+0.702
^{14}N	1	+0.404
^{15}N	1/2	−0.283
^{17}O	5/2	−1.884
^{19}F	1	+2.629
^{23}Na	3/2	+2.218
^{25}Mg	5/2	−0.855
^{27}Al	5/2	+3.642
^{29}Si	1/2	−0.555
^{31}P	1/2	+1.132
^{33}S	3/2	+0.644
^{35}Cl	3/2	+0.822
^{45}Sc	7/2	+4.756
^{51}V	7/2	+5.151
^{55}Mn	5/2	+3.469
^{69}Ga	3/2	+2.017
^{73}Ge	9/2	−0.879
^{81}Br	3/2	+2.271
^{89}Y	1/2	−0.137
^{105}Pd	5/2	−0.642
^{107}Ag	1/2	−0.113
^{183}W	1/2	+0.118
^{195}Pt	1/2	+0.609
^{197}Au	3/2	+0.146
^{207}Pb	1/2	+0.593
^{235}U	7/2	−0.35

* The magnetic moment is given in nuclear magnetons $\mu_N = eh/2m_p$, where m_p is the proton rest mass

The first and still most widely used application for NMR spectroscopy is to investigate the spectra of ^{1}H in organic chemical analysis. The different proton environments in a molecule can be determined from the chemical shifts ob-

tained in a spectrum (e.g., protons of a methyl group, amino protons, etc). The splitting patterns and coupling constants give information about the arrangement of different environments, i.e., what groups are adjacent to each other. By determining the area of various signals, the ratio of protons in various environments can be determined. The wealth of information available from a NMR spectrum makes it a powerful tool for determining the presence of known materials and identifying the structure of newly synthesized compounds. An example of the latter is confirmation of the structure of dodecahedrane, which was first synthesized in 1985 in a 28-step procedure [28]. The NMR spectrum gave a single sharp peak indicating that the hydrogen atoms at each of the 20 vertices of the molecule are all in equivalent environments, expected for a symmetrical structure.

In addition to qualitative uses, ^{1}H NMR is also valuable for quantitative analysis because of the direct relationship between resonance peak area and relative proton abundance. For example, the purity of epoxy monomers can be calculated by comparing the area of the signal from the methylene protons of the glycidyl group with that of the signal from the proton attached directly to the adjacent carbon on the oxirane ring [29]. Similarly, the chain length of oligomeric compounds may be determined by measuring the ratio of protons in the repeating units to those of the end groups.

Nuclear magnetic resonance spectroscopy is applied more universally than merely to compounds containing hydrogen, and many other nuclei are used routinely with this technique. Mononuclidic elements with nuclear spin 1/2 and large magnetic moments are the species of choice to avoid isotopic interference, simplify spectra, and provide large separations in chemical shifts for different nuclear environments. Much of the chemistry of ^{31}P and ^{19}F has been determined with the help of NMR. Evidence for the *Berry mechanism of pseudorotation* in molecules with trigonal – bipyramidal configuration, important in the chemistry of phosphorus, was obtained from NMR spectra [30].

Other nuclei that are often used for NMR spectroscopy include ^{11}B, ^{13}C, ^{29}Si, and ^{195}Pt. Such species introduce complexities into the experimental technique because of low natural abundance, small magnetic moments, or large quadropole effects, but they can give valuable chemical information not obtainable by other means. Although ^{13}C NMR is an important adjunct to ^{1}H NMR in organic chemical analysis, it suffers from both the low abundance (1.11 %) and the small magnetic moment (ca. 25 % that of the proton) of ^{13}C.

3.2.2. Electron Spin Resonance

An unpaired electron possessing a spin of 1/2 has a magnetic moment that responds to a magnetic field in the same way as a nucleus (see Section 3.2.1). Electron spin resonance (ESR; sometimes called electron paramagnetic resonance, EPR) is a spectroscopic method that detects the transitions between energy levels created by interaction of the electron spin with a magnetic field. Stable radical species, certain electronic energy states, and ions and compounds of transition metals and rare-earth elements with unpaired electrons can be analyzed by this technique. Electron spin resonance was first employed in the analysis of organic radical species, but its use has been extended to include quantitative determination of ions such as vanadium(IV), manganese(II), copper(II), chromium(III), gadolinium(III), iron(III), and titanium(III) [14, p. 236].

Isotopes play a role in ESR spectroscopy because the signal from an unpaired electron is affected by the magnetic moments of nuclei contained in the radical molecule or associated with the metal ion containing the unpaired electron. The magnetic moment of an atom with a nonzero nuclear spin couples with the magnetic moment of the electron, called *hyperfine interaction,* to split the resonance signal of the electron into more than one peak. The value of the nuclear spin and the number of interacting nuclei determine the multiplicity of ESR signal splitting. Analysis of hyperfine splitting aids in determining the molecular structure and the nature of the unpaired electron, e.g., whether it is localized on one atom or delocalized over part or all of the molecule.

3.2.3. Mössbauer Spectroscopy

Nuclear gamma-ray resonance spectroscopy, better known as *Mössbauer spectroscopy,* is a

radiochemical analysis technique that relies on the ability of certain nuclides to absorb and emit gamma rays without loss of energy when tightly bound in a crystal lattice [5, p. 185]. Gamma-ray photons originate from a source outside the sample and are absorbed by nuclides in the sample that have the same electronic configuration, and thus the same transition energies, as the source; the two nuclides are then said to be in *resonance.* The absorbed photons are then reemitted, and the apparatus detects both the reduction in number of photons from the source and the reemitted photons. A different electronic configuration of the nuclide in the crystal results in different transition energies and does not allow resonance. The energy of gamma radiation from the source may be altered by *doppler shifting* of the source (i.e., by moving it toward or away from the sample to increase or decrease the photon energy, respectively). A scan over a range of velocities reveals the particular point at which resonance is established. The difference in energy between the source and the target is called the *isomer shift,* and is measured in cm/s relative to a standard such as stainless steel. Signal splitting arises when the quadrupole moment of the nucleus being examined is nonzero, because the resulting electric field gradient interacts with the electron cloud around the nucleus. Thus, splitting provides information about the chemical bonding, oxidation state, crystal lattice, and other features that affect electron distribution.

The Mössbauer effect has been observed in ca. 50 nuclides, with the most widely used isotopes being ^{57}Fe (2.1 % abundance) and ^{119}Sn (8.6 % abundance). The ^{57}Fe is produced from radioactive ^{57}Co ($t_{1/2} = 270$ d) which decays by electron capture to an excited metastable state ^{57m}Fe. The excited state is transformed to either the ground state of ^{57}Fe or an isomeric metastable form of ^{57m}Fe by emission of gamma photons. The isomeric form also decays to the ground state with the emission of gamma rays that are the source for Mössbauer spectroscopy of ^{57}Fe compounds. An example of the use of this technique is the investigation of iron cyanates, which showed that Prussian blue, formed from Fe^{3+} and $Fe(CN)_6^{4-}$, and Turnbull's blue, from Fe^{2+} and $Fe(CN)_6^{3-}$, both have the composition $Fe_4[Fe(CN]_6]_3$ [31]. In addition, high-spin iron(III) with five unpaired d-electrons and low-spin Fe(II) with six paired d-electrons were each identified in the compound, indicating that the individual iron atoms have distinct electron environments.

3.2.4. Isotopic Substitution

Property differences among isotopes of an element can be used in conjunction with other analytical techniques to determine chemical mechanisms, molecular structure, and various phenomena of interest. Because the isotopes of most elements do not differ significantly in their chemical properties, isotopes may be interchanged in a molecule and physical differences (e.g., mass or nuclear spin) can be exploited to highlight specific features by NMR, microwave, mass spectroscopy, or other analytical methods. This technique is especially valuable when applied to elements with one overwhelmingly abundant isotope.

Isotopic substitution can be used to determine chemical mechanisms. The mechanism of alkaline hydrolysis of esters has been elucidated by synthesizing materials with ^{18}O in specific sites (*isotopic labeling*) and following the reactions involving those sites [32]. By labeling the alkoxy oxygen of an ester, cleavage of the ester to the alcohol was shown to occur between that oxygen and the acyl group. Analyses in these investigations were made by mass spectroscopy and identification of the species with which the higher mass oxygen isotope was associated.

Structural determination is also aided by isotopic substitution. Isotopes introduced into specific sites allow identification of the atomic arrangement in molecules. One of the most widespread methods employs deuterium substitution in ^{1}H-NMR studies. In these cases, the desired result is obtained not by monitoring deuterium but by seeing what effect the absence of a proton at selected sites has on the ^{1}H spectra. Specific resonances disappear or are reduced in intensity, and signal splitting multiplicities and coupling constants change as a result of the presence of deuterium. Other spectroscopic investigations also benefit from isotopic substitution. An example is the first reported isolation of the primary ozonide of ethylene [33]. The short-lived species was generated in the cavity of a microwave spec-

trometer at low temperature and was identified as the initial product of the combination of ozone and ethylene, but its structure could not be determined unambiguously. By substituting ethylene that contained deuterium and ^{13}C for ordinary ethylene and repeating the experiments, additional information obtained from changes in the mass of the constituent atoms allowed calculation of bond angles and lengths. From this information, the position of the atoms could be determined. Other types of reactions and chemical systems studied include molecular rearrangements, enzyme-catalyzed reactions, and solvolytic nucleophilic reactions [34].

3.2.5. Cosmology

Many other areas of science in addition to chemistry take advantage of the unique properties of specific isotopes. In the field of cosmology the investigation of solar neutrinos produced in fusion reactions in the sun relies on neutrino interactions with certain isotopes. One of the first studies involved the use of chlorine, in the form of a large quantity of carbon tetrachloride. Neutrinos induce a nuclear reaction in ^{37}Cl which converts it to ^{37}Ar, whose β emission can be readily monitored. These studies showed that the number of events caused by neutrinos was only about one-third of that predicted by the best theoretical models of the sun [35]. The ^{37}Cl detector, however, is sensitive only to neutrinos with energy above a certain threshold and not to those produced by the fusion of two hydrogen nuclei, the most common solar nuclear reaction. More recent experiments have been conducted with other isotopes sensitive to a wider spectrum of neutrino energies [34]. The GALLAX detector, operated between 1991 and 1997, used 100 t of gallium chloride containing about 12 t of ^{71}Ga. Reaction of this isotope with neutrinos produces radioactive ^{71}Ge, which was extracted as the volatile germanium chloride and converted to the hydride, GeH_4, and its decay measured. A typical run lasted six months and detected 25 neutrino interactions [36]. The Sudbury Neutrino Observatory used 1000 t of heavy water, 99.92 % isotopically enriched deuterium oxide. This detector was sensitive to the low-energy neutrinos resulting from the production of ^{8}B in the sun. The information from these and other experiments reinforced the initial conclusion that the neutrino flux is less than predicted by the standard solar model. A number of theories have been put forth to explain the results, but none has yet received widespread acceptance.

3.2.6. Nuclear Science

Nuclear physicists have employed specific isotopes to make the first observations of an extremely rare form of radioactive decay, the simultaneous emission of two beta particles. Double-beta decay was predicted in the 1930s as a method for decay of unstable nuclides that were prohibited from undergoing simple beta decay because of binding-energy considerations (→ Radionuclides, Chap. 1.3.). A number of very sensitive, long-term experiments have been conducted to study this phenomenon [37]. The first attempts to identify double-beta decay involved geochemical analysis of tellurium- and selenium-containing ores to measure the amounts and isotopic ratios of noble gases trapped in the ores. The nuclides ^{130}Xe, ^{128}Xe, and ^{82}Kr are the daughter products of double-beta decay of ^{130}Te, ^{128}Te, and ^{82}Se, respectively. The results gave evidence for this type of nuclear decay and estimated the half-lives to be 10^{19} a or greater. However, these studies could not determine which of the theoretical decay modes was involved. Attempts to observe the phenomenon in the laboratory have employed ^{76}Ge, ^{100}Mo, and ^{136}Xe. These experiments have been able to detect the decay process that produces two neutrinos but only estimate a lower limit for the neutrinoless mode. The latter is of particular interest to physicists since, if it actually exists, it will help to establish a mass for the neutrino.

3.2.7. Materials Science

Positron spectroscopy has been applied with great success to materials science [38]. The technique is based on the fact that certain radioactive species decay by conversion of a proton to a neutron accompanied by emission of a positron, the anti-matter counterpart of the electron. Positron lifetimes are very short (ca. 10^{-10} s). When a positron encounters an electron both particles are annihilated and energy

is produced, usually in the form of two γ-rays. Positron spectroscopy is the detection and measurement of these γ-rays. This decay mode is typical of unstable nuclides that are proton-rich. Such species can be produced in cyclotrons, in which stable isotope targets are bombarded with charged particles, protons, or deuterons, or in nuclear reactors or other devices, where neutrons are the bombarding species. The process of adding particles to stable nuclides is called activation (→ Radionuclides, Chap. 3.). Stable-isotope precursors of positron-emitting species for materials science uses include ^{63}Cu, used to produce ^{64}Cu (n, γ reaction); ^{79}Br to produce ^{79}Kr (p, n reaction); ^{58}Ni to produce ^{58}Co (p, n); and ^{22}Ne to produce ^{22}Na (p, n).

Attributes of positron spectroscopy that make it a useful analysis technique include its ability to provide depth-profiling and its nondestructive nature. When a positron is introduced into a material it has a tendency to localize at regions of relative negative potential such as vacancies and voids. Upon annihilation, the photons produced easily escape from the material, and their signature is directly related to local environment around the positron when it reacted with an electron. Early uses of the method were limited to investigation of simple metals such as copper and aluminum for detection of point defects in metal lattices. The technique has been extended to study the effects of annealing, work hardening, and radiation damage in metals and alloys, as well as investigating free volume effects in polymers. With the advent of slow positron sources, positron spectroscopy has become a valuable tool in study of silicon-based electronic devices, especially the various methods for their manufacture and the interactions at the interfaces between the different materials that constitute such devices.

3.2.8. Ecological Research

Stable isotope analysis can be applied to ecological research to investigate processes within and between organisms and between organisms and their environment [39–42]. Isotope ratios were first used in the 1970s by geochemists in analysis of fossil plant matter to elucidate climate and climatic changes during the period when the plants lived. Hydrogen isotope ratios were analyzed in peat samples and in annual growth rings of treas and correlated to temperature. The initial studies included the simplifying assumption that the hydrogen isotope ratios in plants were only affected by the ratios in the water available, which is a function of temperature. Several sources of error were identified in these studies, including the fact that cellulose contains hydroxyl hydrogen atoms that are easily exchangeable. Hence, the measured hydrogen ratio is influenced by the last water to which the cellulose was exposed, which may have been in sample preparation. In addition, different substances within the same plant may have different isotope ratios as a result of the specific biological processes they undergo. Whole plant sampling gives an overall ratio that is affected by the relative amounts of the various constituents. The procedures now used for climatic studies involving plant matter minimize these errors and better reflect environmental differences. Recognition of the possible artifacts affecting climatic studies were a catalyst for investigation of a wide range of physiological processes in plants, by using stable isotope analysis. The greater availability of isotope-ratio mass spectrometers also made such studies more common.

Carbon isotope ratios have been used to study the differences in photosynthesis pathways of plants. The primary carboxylase enzymes involved in fixation of carbon dioxide from the air by the C_3 and C_4 pathways have different kinetic isotope effects and thus result in different δ (^{13}C) values in plant matter. The C_3 pathway typically results in a value of about -27 ‰ and the C_4 pathway -12 ‰. δ (^{13}C) values in soil are also affected when C_3 crops replace C_4 crops (or vice versa), whereby different types of soil and different particle sizes exhibiting different replacement times. Fine clay appears to be the "youngest" part of a soil mixture.

Studies of carbon isotope ratios have also been used to measure the efficiency of water use by plants. The rate of CO_2 assimilation by plants is proportional to the rate of water loss, since both are a function of the diffusion of gases through plant membranes such as leaves. Since diffusion of $^{13}CO_2$ is slower than that of $^{12}CO_2$ because of the greater mass of the former, a carbon isotope ratio in the plant that is closer to that of the atmosphere indicates less diffusion of CO_2 into the plant and implies less water loss, i.e.,

greater water efficiency. The same techniques have been used to study the response of plants to various environmental stresses such as water availability and salinity, light flux, and temperature.

While carbon dioxide is essentially the only source of carbon for photosynthesis in plants, there are three potential sources of oxygen: carbon dioxide, molecular oxygen from air, and water. Plants grown in atmospheres containing CO_2 with greatly different ^{18}O content, but the same water, did not show significant differences in oxygen isotope ratios. Effects that may arise from molecular oxygen during photorespiration appear to be lost during subsequent steps in the photosynthesis. These results, coupled with other more direct experiments, indicate that water is the only significant source of differences in oxygen isotope ratios, and that the δ (^{18}O) value in plants is consistently about +27 ‰ greater that of water. Mechanistic studies point to the carbonyl hydration reaction at the three-carbon-sugar level as the overriding step in photosynthesis that determines the oxygen isotope ratio. The same degree of enrichment also occurs in the simplest chemical system analogous to this reaction, a δ (^{18}O) of +27 ‰ is found in the carbonyl oxygen of acetone in equilibrium with water.

The nitrogen cycle in terrestrial systems is complex, starting with N_2 from the atmosphere which is reduced (fixed) to ammonia by microorganisms in the soil or in the commercial production of fertilizers involving the Haber Process. Ammonia or the ammonium ion can be directly assimilated by plants or oxidized to nitrate by other soil bacteria and absorbed in that form. Studies of nitrogen-fixing plants have used analysis of variations in the natural isotope abundance of nitrogen very effectively. Other experimental work has employed materials artificially depleted in ^{15}N, including fertilizers. Assessment of the effects of different methods of soil preparation showed that nitrogen uptake from fertilizer was approximately the same in all cases, but the total nitrogen was greater in the plants grown in soil that had been disturbed the least in preparation. This suggested that microbial activity converting nitrogen to a form usable by the plants was greater when the soil was not tilled. Depleted fertilizer was also used in a study of the ability of various plant residues, when returned to the soil, to provide nitrogen to the next generation of crops. Soybeans were found to be a good source of nitrogen that could be utilized by plants, whereas corn was not.

In many cases it is essential to analyze the isotopic composition of more than one element to draw the correct conclusions. Studies of the generation and destruction of methane by microorganisms in sediments included analysis of both carbon and hydrogen isotopes. Because of a kinetic isotope effect, the methane produced is depleted in ^{13}C. The oxidation of methane also has an isotope-fractionation factor, in this case leaving the CH_4 enriched in ^{13}C. In one study, δ (^{13}C) values of methane were measured as a function of depth in sediment samples to determine where the various processes were occurring. The results indicated that methane production occurred throughout the sediment, while oxidation occurred only in a narrow zone. Analysis of the δ (^{2}H) values confirmed that the differences in ^{13}C concentration were a result of the oxidation of methane and not an artifact of the different diffusion rates through the soil expected for the various isotopic combinations of carbon and hydrogen making up the methane.

Strontium isotopic analysis has been used to study large ecological systems. Analysis of trees in a northern New Mexico watershed allowed determination of the percentages of the biomass that were derived from weathering of the bedrock versus the amount derived from atmospheric input. An unexpected benefit from this study was recognition that the method could also be used to estimate the depth of the rooting zone of the trees.

Environmental studies have been performed by using analysis of sulfur and nitrogen isotopes. Oxides of sulfur and nitrogen are emitted from fossil fuel burning power plants and other air pollution sources in large quantities and affect large areas in industrialized countries. Studies were carried out, beginning in the 1970s, of the effect of SO_2 on ecosystems in areas around natural gas refineries in Canada. A number of factors made these sites ideal for such investigations. The SO_2 contained in the natural gas that was removed during the refining process had a δ (^{34}S) that was significantly different from the background values in surrounding soil (as much as +15 ‰). There were few other potential environmental pollutants, such as heavy

metals, and the refinery locations were separated such that each acted as a point source. Initial studies looked at the distribution of sulfur in the surrounding areas and found the highest sulfur and, particularly ^{34}S, concentrations downwind from the refineries, and the lowest concentration upwind. The composition of plant life was also different, but the cause and effect relationship to the presence of SO_2 could only be inferred. To examine the accumulation of sulfur in plants and the metabolic effects, SO_2 enriched in ^{34}S has been used. Plants exposed to high levels of labeled SO_2 show increased levels of sulfates, sulfur-containing amino acids, and other sulfur containing compounds. δ (^{34}S) values in all of these compounds were found to be higher than in control samples, and the indicated atmospheric sulfur as the source.

Methane is a very efficient greenhouse gas, and its concentration has been increasing at about 1 % per year in the atmosphere. Analyses of ice core samples suggest this increase began in the 1700s and accelerated in the twentieth century. Methane has a residence time in the atmosphere of about nine years and is chiefly removed by reaction with hydroxyl radicals. There is great interest in determining the contributions of the numerous sources of the gas. Atmospheric methane has a δ (^{13}C) of -47.7 ‰. Analysis of the δ (^{13}C) values of its sources, allows estimations of the importance of their contributions to the approximately 600×10^6 of methane entering the atmosphere annually. Wetland and tundra soils are estimated to contribute about 25 % of the annual methane production. Natural gas production and fossil fuel burning contribute another 25 % and rice cultivation a similar amount. Physiological processes in ruminants and biomass burning are other major sources. These estimates include rather large uncertainties because of the range of δ (^{13}C) values found for each of the source categories, including seasonal variations in the same source.

Another isotope that is valuable for environmental studies is ^{14}C, a naturally occurring radioactive isotope that is produced in trace amounts in the atmosphere by cosmic rays interacting with ^{14}N (p, n reaction). It enters the biosphere from the air and has been extensively used in dating of materials. Because of its short half-life ($t_{1/2} = 5730$ years) ^{14}C is essentially absent from long-sequestered sources of carbon like fossil fuels, and these are known as "dead" carbon. Measurement of ^{14}C in the atmosphere during the 20th century shows a decrease in δ (^{14}C) from the early 1900s to the mid 1950s, as the burning of fossil fuels added large amounts of CO_2 to air that contained little of the radioactive isotope. A large increase in the ^{14}C concentration of the atmosphere occurred in the 1950s as a result of atmospheric testing of nuclear weapons. This source has caused problems with interpretations of radiocarbon dating experiments but has allowed a much greater understanding of the carbon cycle in the Earth. Bomb-produced ^{14}C has been used to verify residence times of gases in various layers of the atmosphere and to demonstrate that the mixing between surface and deep ocean waters is much slower than originally thought, taking hundreds to thousands of years.

3.2.9. Isotope Chronostratigraphy

A specialized area of geological research that is of potential commercial interest is isotope chronostratigraphy [43], the determination of when a layer of sedimentary rock was formed by analysis of the carbon and oxygen isotope ratios. The basis of this technology is studies of samples from the Deep Sea Drilling Project. Extensive analysis of these samples by a number of methods has given a remarkably detailed description of the oceanographic and biological history of the Earth during the Tertiary era, 65 to 2.8 million years ago. Values of δ (^{13}C) and δ (^{18}O) in carbonates from fossilized marine animals (benthic and planktonic foraminifera and coccoliths) from the DSDP samples have been correlated to the overall timeline. Variations in these values over time reflect changes in ocean chemistry. Parameters such as temperature, water depth, salinity and sources of carbon dioxide affect the δ values of both carbon and oxygen in marine carbonates, but not necessarily to the same extent for each element. For example, the oxygen ratio is affected to a much greater degree than the carbon ratio by temperature.

Extrapolation of the isotope ratio information from the well-characterized samples of the DSDP to other samples from exploratory drilling sites is of great interest to oil companies. Values of δ (^{13}C) and δ (^{18}O) and trends in their

changes versus drilling depth from new wells may be compared with the features of known petroleum-bearing layers to identify the potential for new wells to find untapped reserves. Isotope chronostratigraphy appears particularly suited to investigation of deep-water sites where formations resemble the DSDP drilling sites. Advances in the methodology include more sophisticated mathematical treatments to replace the original empirical analysis of results and the integration of isotope ratio information with other analysis techniques such as biostratigraphy and paleomagnetic stratigraphy. These innovations hold promise for broadening the applicability of the technology to a wide variety of sedimentary formations.

3.2.10. Anthropology

Isotope analysis of anthropological remains is one of many methods available for the study of diets of prehistoric peoples [44]. Information on food acquisition, consumption and distribution is typically very important in understanding ancient societies and their evolution. Carbon and nitrogen have been the most extensively analyzed elements in such studies, but others have been used as well. Collagen, the organic matrix in bone, contains carbon and nitrogen, and collagen residues are typically associated with skeletal remains. δ (^{13}C), and δ (^{15}N) values can be used to indicate aspects of ancient diets, especially differences in consumption of C_3 plants (rice, potatoes, wheat, etc.) versus C_4 plants (maize, sorghum, millet, etc.), and marine versus terrestrial food sources. Hydroxyapatite, the inorganic portion of bone that makes up the majority of the mass, contains carbon as carbonate, but its usefulness is complicated by exchange reactions with carbonate in groundwater. The technique of stable isotopic analysis has been applied to many surveys throughout the world, and the results have been helpful in determination of diet, changes in agricultural practices, food distribution, and domestication of animals.

Strontium is also useful in dietary investigations [45] as it is present in most rocks in small quantities and makes its way into the food chain as a substitute for calcium in certain minerals. The isotopic composition of strontium varies with rock type and geological age because one of its isotopes, ^{87}Sr, is a decay product of the naturally occurring radioactive isotope ^{87}Rb ($t_{1/2} = 4.9 \times 10^{10}$ years). Analysis of Sr isotopes has been extensively used in geological studies to date rock formations, and a large body of information is available about the isotopic distribution of Sr in many locations. A pilot study of strontium isotopic analysis in ecological investigations involved analysis of bones and teeth of prehistoric humans. The ^{87}Sr/^{86}Sr ratio in remains of two persons found in an ancient coastal village near what is now Los Angeles, California, were consistent with the strontium ratio in seawater and indicated a diet high in marine food. Remains from an inland village inhabited by people from the same tribe were found to have a lower ^{87}Sr/^{86}Sr ratio than the coastal remains, but higher than that of the soil in the area of village. The results were consistent with the known behavior of the tribe: inland villagers typically spent part of the year inland but moved to the coast during other parts of the year, where they ate fish. Carbon isotope analyses was consistent with the strontium results: a high-marine-food diet in the coastal villagers and lower amounts of marine food in the diet of the upland villagers. An interesting specific application of the strontium analysis technique would be to identify the remains of Christopher Columbus and his son, Don Diego, who are said to be buried together in the Cathedral of Santo Domingo in the Dominican Republic. The elder Columbus was a resident of Genoa, Italy, during his youth, while his son lived in Seville, Spain, as an adolescent. The strontium ratio in the soil from the two locations is sufficiently different that isotopic analysis of the teeth and bones of the remains should distinguish between these two historical figures.

3.3. Medical Uses

One of the fastest growing uses of isotopes is in medicine. New techniques exploit the special properties of particular isotopes for diagnostic work and employ stable isotopes as target materials to produce radioactive species for both diagnosis and treatment.

3.3.1. Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a medical diagnostic method that exploits the magnetic properties of certain nuclei [47] in the same way as NMR is used in chemical analysis (see Section 3.2.1). This technique is able to generate cross sections and three-dimensional images of the human body and to follow the progress of physiological functions that involve magnetically active nuclei. In NMR, the transitions between spin states of magnetic nuclei are monitored, and the frequency necessary to induce transitions as well as the intensity of the transitions are determined. Conversely, MRI monitors the rate of *relaxation* between spin states and the intensity of the transitions. Relaxation is the process by which the nucleus undergoes a transition from a higher energy spin state to a lower one. The energy difference between spin states depends on the strength of the applied magnetic field but is typically very small (ca. 0.01 eV), and is thus lower than the average thermal energy of atoms and molecules at room temperature. Therefore, the population of the two energy states can easily become saturated and equal numbers of nuclei in both states result in equal amounts of absorption and emission and, thus, no measurable signal. However, nuclei can interact with their surroundings and relax to the lower energy spin state by two mechanisms. In *spin – lattice relaxation,* the component of magnetization in the z direction (aligned with the external magnetic field) is returned to its equilibrium value; it is characterized by a relaxation time T_1. Relaxation occurs when nuclei interact with oscillating fields produced by neighboring magnetic nuclei which have the same frequency and phase as the excited nuclei. Many factors are involved in spin – lattice relaxation, including viscosity, temperature, and concentration. Interaction with the unpaired electrons of paramagnetic molecules, such as dissolved molecular oxygen in the sample, is a very effective mechanism for inducing spin – lattice relaxation. Nuclei with a quadropole moment may relax by interaction with their own electrons. Relaxation in the xy plane, perpendicular to the applied field, is called *spin – spin relaxation* and is characterized by a relaxation time T_2. Spin – spin relaxation is a function of the phase coherence of the excited nuclei and may be affected by the viscosity of the medium and spin – spin exchange reactions between excited nuclei.

Magnetic resonance imaging usually monitors protons, which are found in most biological substances. The specific environment of the nucleus, the molecule of which it is a part, and the molecules in its immediate vicinity determine the relaxation times T_1 and T_2. The number of protons of a particular type determines the intensity of the signal. With sophisticated computer control and signal analysis, MRI spectrometers collect data simultaneously from two perpendicular directions to generate cross-sectional images of the body. Combining successive cross sections develops a three-dimensional picture. The differences in relaxation times and intensities for the hydrogen nuclei in various tissues give remarkable definition to the images. Among others, tumors, aneurisms, and herniated disks can be diagnosed with this technique.

Although hydrogen is an excellent nucleus for providing well-defined images of soft tissue structures in the body, its ubiquity makes it less suitable for the study of physiological functions. For investigations of this type, the best technique is to introduce isotopically substituted compounds into the body and follow their pathways, concentration sites, and chemical reactivity. Ideally, a compound used in this manner will incorporate a magnetically active isotope in place of a nonactive isotope in a biologically important compound; for example, ^{13}C is frequently used to replace ^{12}C, and ^{17}O to replace ^{16}O. The metabolism of glucose in the brain of animals has been studied by following the chemical reactions of ^{13}C-labeled glucose [48]. Other nuclides, including ^{2}H, ^{7}Li, ^{23}Na, ^{35}Cl, and ^{87}Rb, are used in high magnetic fields ($>$10 T), rather than the 1.5 – 5-T fields routinely used for ^{1}H. The higher fields result in greater splitting and first-order spectra, which allow much more information to be obtained about metabolic functions [49]. For very precise measurements, the equilibrium and kinetic isotope effects of the substituted species (see Section 2.2) should be insignificant, or data should be corrected to reflect the different reactivity of the labeled compound. In addition to the direct study of metabolism with isotopically labeled materials, secondary effects may be used to obtain information. Changes in relaxation times are induced by isotopic composition changes of the

medium in which the targeted nucleus is found. For example, water incorporating ^{17}O induces relaxation in the hydrogen nucleus at a different rate than water containing ^{16}O. Advances in MRI include the use of contrast-enhancing agents, such as organic chelates of gadolinium, especially ^{153}Gd, to provide increased definition.

3.3.2. Stable Isotopes in Metabolic Studies

Stable-isotope analysis has been used in the study of human and animal metabolism. Studies of carbon isotope ratios in animals [50] shows a correlation between the overall δ (^{13}C) value in the body and the δ (^{13}C) value in the diet, with a slight enrichment of ^{13}C in the animal. As might be expected from similar findings in plants, individual tissue types differed in carbon isotope ratios. In gerbils, hair was enriched, and fat depleted, in ^{13}C from the diet, and the results were somewhat diet-dependent. Changes in δ (^{13}C) values in body tissues were noted when diets were altered in ^{13}C content (by switching from C_3 to C_4 plants). The tissue replacement half-lives could then be estimated: they range from 6.4 d for liver to 27.6 d for muscle.

Human metabolism of inorganic nutrients, including Fe, Zn, Mg, Cu, Se, and Ca, has been studied by using enriched stable isotopes [51]. Introduction of the labeled species can be intrinsic, through biological incorporation in food, or extrinsic, by mixing with food before consumption. Iron and zinc have most often been incorporated intrinsically. Peas, spinach, and wheat have been labeled during plant growth. Milk, eggs, and meat have been labeled by injecting animals with solutions of enriched isotopes. The uptake of nutrients may be calculated from the difference between the administered amount and the amount found in analysis of excreted materials. Other studies have measured isotopic abundance directly in blood plasma or hemoglobin. A double stable-isotope technique, analogous to that first developed with radioactive tracers, has been used in which two different enriched isotopes of the same element are simultaneously administered, orally and intravenously. Studies of calcium absorption have used oral dosing with ^{44}Ca-enriched foods and injection of ^{42}Ca-enriched solution. The ratio of the isotopes in plasma or urine after 24 h is measured to determine the metabolism of the element. Double-isotope studies with ^{67}Zn and ^{70}Zn have been performed to measure the normal loss of this nutrient from the body and allow more accurate calculation of absorption values.

Whole-body calorimetry, developed more than 100 years ago, measures energy requirements by enclosing a human subject for hours or days in a room that monitors heat loss and oxygen consumption. While allowing very accurate measurements, it hardly simulates a person's typical routine. In the 1950s another method for obtaining the same information was developed, the doubly labeled water technique which employs $^{2}H_2^{18}O$. Both isotopes reach equilibrium rapidly in the body, the deuterium through exchange reactions with water, and the oxygen with both water and carbon dioxide. Test subjects drink the labeled water or it is injected into the blood stream. The disappearance of the isotopes from the body is monitored by analysis of urine, blood, or saliva samples, during which time the person may goes about normal activities. Analysis of the deuterium concentration over time gives the water turnover rate. Analysis of the ^{18}O concentration gives the sum of the water and carbon dioxide turnover rates and, by difference, the carbon dioxide value. The CO_2 value is used in calorimetric calculations to determine the energy expenditure over the study period. Several assumptions must be made, including the relative quantities of fat and carbohydrates consumed and the amount of water lost through perspiration, but the precision of the method is estimated to be better than 5 %. A comparison with whole-body calorimetry over a 12-d period found them to agree within 2 %. The doubly labeled water method has been used in a large number of diverse biological investigations [52]. Specific studies in humans include metabolic correlation of obesity and activity, the effects of total intravenous feeding on adults, and metabolic processes of premature infants. In other species, it was one of the technique used in a field study of Galapagos marine iguanas that showed how these animals regulated osmotic balance in their bodies despite large intake of sea water. It also has been used to compare the energetic cost of flight in various types of birds, and to understand the behavioral characteristics of desert rodents that allows them to survive with-

out winter hibernation. Sufficient data has been gathered to allow predictive models to be created to describe energy and material flow though habitats and estimate field metabolic rates and food requirements for various species.

3.3.3. Stable Isotopes in Medical Diagnosis

A simple, accurate, noninvasive test that uses stable isotopes to identify a bacteria that causes stomach and intestinal ulcers has been approved by the U.S. FDA [53]. *Helicobacter pylori* infects epithelial cells throughout the digestive system in humans, producing ammonia and the ammonium ions to regulate the pH in its immediate vicinity and protect itself against normal stomach acidity. The most important enzyme in the bacteria's protective mechanism is urease, which hydrolyzes urea to form ammonia and carbon dioxide. The test uses ^{13}C-enriched urea, swallowed by the patient, and measurement of the ^{13}C content of carbon dioxide from a breath test taken about 30 min after ingestion of the labeled urea. Elevated levels of $^{13}CO_2$ in the exhaled gases, as compared to reference breath sample taken before the test, indicate infection by *H. pylori*. The test has been shown to be sensitive and specific for the bacteria and can be used in initial diagnosis and to determine if subsequent treatment is effective. The FDA has approved several companies to produce the labeled urea.

3.3.4. Stable Isotopes as Radioisotope Precursors

In addition to their use directly in diagnosis, stable isotopes are more prevalent as precursors to radioactive species used in both diagnosis and treatment. Positron emission tomography (PET) is a medical technique analogous to positron spectroscopy (see Section 3.2.7) that can produce three-dimensional images of the body or specific organs and can also be used to study physiological functions. Positron-emitting species for PET are typically low atomic mass elements that may be substituted for stable nuclides in biologically active compounds. These are routinely produced in cyclotrons by charged particle activation (→ Radionuclides) and include ^{11}C, ^{13}N, ^{15}O and ^{18}F. An extensive area of chemistry has developed around production of PET-useful compounds incorporating these short-half life species. Neutron activation of stable isotopes occurs in the same manner as charged particle activation and produces neutron-rich nuclides that often decay by beta emission. Fission reactions may be regarded as a special type of neutron activation because they are generally induced by neutron capture. Many useful radioactive isotopes are produced in fission reactions, including ^{90}Sr, ^{99}Tc, ^{131}I, ^{133}Xe, ^{137}Cs, and ^{144}Ce. One of the newest medical diagnostic and treatment techniques combines radiochemistry with genetic manipulation. Radioactive nuclides have been attached to monoclonal antibodies created by recombinant DNA techniques to be specific for antigens contained on the surface of cancer cells [54], [55]. The therapeutic agent is thus carried preferentially to the desired site. The technique has proven most effective in treatment of lymphomas and leukemia as opposed to solid tumors.

Great advantage may accrue from using pure or enriched isotopes as target materials in activation processes. The target isotope is often an element that contains several other isotopes as well. Using the natural isotopic mixture may create unwanted radioisotopes in addition to the desired product, and if the desired product is of low natural abundance, the desired product may be only a small fraction of the activated material. For example, iron radioisotopes ^{55}Fe and ^{59}Fe are both useful materials produced by neutron capture from naturally occurring iron isotopes ^{54}Fe and ^{58}Fe, respectively, but have different applications. Because the natural abundance of ^{54}Fe is much greater than that of Fe (5.8 vs. 0.3 %), activation of a normal sample of iron gives nine times as much ^{55}Fe as ^{59}Fe. However, enrichment of the starting material in the target of interest by electromagnetic separation can be performed to alter the final product ratio. By using iron containing 97 % ^{54}Fe, the activation product contains almost exclusively ^{55}Fe. Conversely, if the ^{58}Fe is enriched to 85 %, the product contains less than 5 % ^{55}Fe. Another benefit of preenrichment of target material is the reduction in time and amount of radiation, i.e., energy, needed to produce a given yield of the desired product.

3.4. Industrial Uses

Few applications have been found for specific isotopes in industrial settings. Radioactive species have been used in the construction industry for site characterization and the investigation of construction materials and techniques, including the detection of reinforcing rods and flaws in concrete [59]. In general, however, the use of separated or enriched stable isotopes for industrial purposed is limited by availibility and cost. The separative methods employed for commercial sale of isotopes produce milligram and gram quantities of materials, and the repetitive processing required to obtain significant enrichment is energy-intensive and therefore expensive. Nevertheless, potential markets have been identified for large amounts of enriched stable isotopes if they become cheaper.

Enrichment of one of the isotopes of mercury, ^{196}Hg, has been shown to increase the efficiency of fluorescent lighting by up to 5 % [60]. However, no cost effective method of implementing this technology as been developed. Tagging of products with enriched isotopes has been suggested as an intrinsic method of identifying their origin and to thwart counterfeiting [61]. A subtle change in the isotopic composition of one or more of the constituent materials would serve to unequivocally identify a component made with such material. This method of labeling would be consistent throughout the entire component and could not be obliterated short of destroying the part. Isotopically altered inks have been proposed for use in legal documents, securities, and other printed matter, even paper currency. These incorporate an element such as chromium in altered abundance that could easily be determined by isotope-ratio mass spectrometry to allow detection of forgeries. Isotopic enrichment of silicon has been postulated as a means for increasing the efficiency of integrated circuits. ^{28}Si has greater thermal dissipation efficiency than natural silicon and may allow finer circuitry to be produced on existing equipment. Zinc is used as a corrosion-resistant additive in water-circulation systems of nuclear power plants. Exposed to high neutron fluxes, it eventually becomes a radioactive waste material. Depletion of ^{64}Zn from the natural occurring material results in much lower radioactivity and a significantly shorter half-life in the waste material.

The priority pollutants listed by the U.S. EPA are now commercially available in isotopically altered form. These labeled materials can be added to ordinary (unaltered) materials to allow them to be traced and identify their origin.

4. Isotope Separation [15], [61]

4.1. The Separation Factor

Certain differences in physical and chemical properties of isotopes (see Chap. 2) can be used as a basis for their separation. The behavior of atoms and molecules in a gravitational field, or of ions in an electric or magnetic field, depends directly on the mass and varies individually for isotopes. Statistical differences arising from variations in mass are found for the mean velocity of atoms and molecules at a given temperature, the vapor pressure, and the diffusivity. Indirect dependence is found for properties related to vibrational frequencies and the zero-point energy of atoms and molecules; melting and boiling points, as well as equilibrium and kinetic isotope effects, fall into this category.

For any given method of isotope separation, the amount of enrichment that can be expected is given by the *elementary separation factor* q_0. This quantity is specific to the process employed and the conditions under which separation is performed. Derivation of the separation factor may be generalized by considering a material containing an isotope of interest with a mole fraction N (i.e., the number of moles of the target isotope divided by the total number of moles present). After a short time, any enrichment process operating on a quantity B of material will result in a portion of that material δB which is enriched in the target isotope, and a portion $B-\delta B$ which is depleted in that same isotope. The mole fraction of the target isotope in the enriched portion is defined by N' and the mole fraction in the depleted fraction by N''. For a binary mixture, the elementary separation factor is then defined by the equation

$$q_0 = \frac{N'/(1-N')}{N''/(1-N'')} = \frac{R'}{R''} \tag{14}$$

where R is the relative abundance of the isotope, R' the abundance in the enriched portion, and

R'' the abundance in the depleted portion. For an enrichment process that occurs under conditions of thermodynamic and kinetic equilibrium, the separation factor is always equal to the elementary separation factor even after the process has been halted. For a process exhibiting a kinetic isotope effect, the equality holds only while the process continues. The system comes to equilibrium after the process has been halted, and the separation factor may be determined by the amount of material that has undergone the particular transfer or transformation involved.

In most cases, the elementary separation factor is small, typically close to 1.0. If a process is allowed to operate on a quantity of material, the enriched fraction will have only a slightly higher concentration of the target isotope than the depleted fraction. To achieve a large increase in concentration, the elementary separation effect must be multiplied by subjecting the enriched fraction to another separation and repeating this sequence until the concentration of the target isotope reaches the desired level. Multiplication of the elementary separation effect may be accomplished in one of two ways: by a discontinuous process acting on a single batch of material or by a continuous process involving many interconnected steps.

4.1.1. Continuous Processes and Cascades

In a continuous process, material is introduced into and removed from various parts of the enrichment operation at constant rates. A continuous process usually employs many individual separation units, called *elements,* usually identical and capable of enriching the target isotope to an extent determined by the elementary separation factor. To handle large quantities of material, many individual elements may be connected in parallel to form a *stage.* To multiply the elementary separation effect, stages are arranged in series to form a *cascade.* In a cascade, the enriched output from one stage, called the *product* or *heads,* is used as feed for the next stage. The process acts on this material, with an initially higher target isotope concentration, to enrich the desired isotope again by the elementary separation factor. Because the product from a stage contains only part of the feed, each stage in a cascade has fewer elements than the preceding stages.

The depleted material removed from each stage, called the *tails,* can be discarded or recycled, depending on the economics of the process and the substances involved. Recycling involves adding the tails from one stage to the feed of a stage upstream in the cascade (i.e., a point of lower enrichment), creating a *countercurrent cascade.* Recycling the tail to the stage immediately upstream creates a symmetrical cascade, whereas recycling two or more stages upstream is said to be asymmetrical. Another method used to maximize the quantity of enriched product from a given amount of initial feed material is to process the tails of the first stage of the cascade. This fraction is of lower concentration than the initial feed and is generally of no use for enrichment. However, it does contain some of the target isotope and thus may be used as feed for a series of stages operating at lower concentrations, in a *stripping section,* to yield a product at concentrations useful in the enrichment section of the cascade. The utility and design of a stripping section is determined by the cost of the feed material compared to the cost of the total enrichment process. Figure 8 illustrates a simple symmetrical countercurrent cascade with a stripping section.

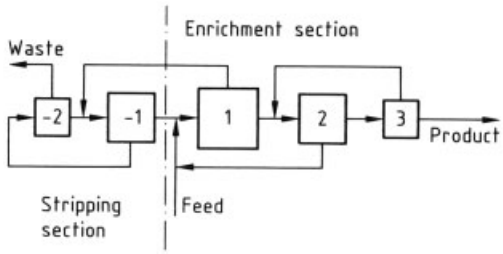

Figure 8. Symmetrical countercurrent cascade

The enrichment that occurs in an element of the cascade depends on several factors. The cascade has a feed with flow rate L, with a mole fraction of target isotope N and output in two forms: an enriched product with flow rate L' at mole fraction N', and a depleted tail with flow rate L'' at mole fraction N''. The portion of the feed that is removed as product is called the *cut* θ, and it determines the flow rates of the product and tails streams:

$$L' = \theta L$$
$$L'' = (1-\theta)\, L \tag{15}$$

Conservation of matter requires that

$$L = L' + L''$$
$$LN = L'N' + L''N'' \tag{16}$$

Several equations define the terms needed to establish relationships among the various streams and their isotopic concentrations. The ratio of relative abundance in the product and tail streams is called the *separation factor q,* which, if no remixing of the streams occurs, is given by (cf. Eq. 17):

$$q = \frac{N'/(1-N')}{N''/(1-N'')} = \frac{R'}{R''} \tag{17}$$

The *separation gain g* is defined as

$$g = q - 1 = (R'/R'') - 1 \tag{18}$$

The *enrichment factor* α is the ratio of relative abundances of the target isotope in the product and feed streams:

$$\alpha = \frac{N'/(1-N')}{N/(1-N)} = \frac{R'}{R} \tag{19}$$

The *enrichment gain* ε is defined as

$$\varepsilon = \alpha - 1 = (R'/R) - 1 \tag{20}$$

The separation factor and separation gain give a measure of how the process partitions the target isotope between the product and the tails streams, whereas the enrichment factor and enrichment gain measure to what extent the process increases the concentration of the target isotope in the product relative to the feed. Note that all of these equations include the concentration of the target isotope. At low concentrations, this becomes insignificant and the following approximations may be made

$$q = N'/N''$$
$$\alpha = N'/N \tag{21}$$

Applying this simplification and performing a series of substitutions allows the enrichment factor to be expressed in terms of the separation factor and cut:

$$\alpha = \frac{q}{1+\theta\,(q-1)} \tag{22}$$

Figure 9 shows the dependence of the enrichment factor on cut for different values of separation factor. For very large cuts, the enrichment factor tends toward unity; for small cuts, the enrichment factor approaches the separation factor. The cut may thus be optimized to provide the desired combination of quantity of product and degree of enrichment from an individual unit. Because a stage consists of more than one element, the same mathematical treatment applies but with the flow rates multiplied by the number of elements.

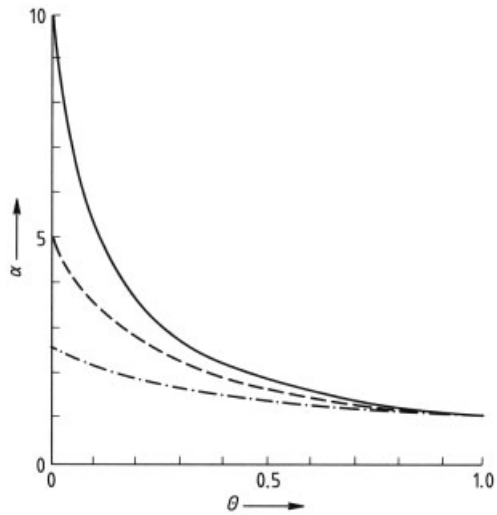

Figure 9. Enrichment factor α as a function of cut θ
—· $q = 10$; – – – $q = 5$; – · – $q = 2.5$

In a cascade, a number of stages are connected in series, and in a situation analogous to the individual separation unit, a cascade has feed with a flow rate F and mole fraction N_F of the target isotope, and outputs in the form of product with flow rate P and mole fraction N_P, and tails with flow rate W and mole fraction N_W. These flow rates and concentrations are the external parameters of the cascade; any four external parameters may be used to define the system. Usually, the product flow rate and the concentrations of feed, product, and waste streams are set, which determines the other two flow rates. In a symmetrical countercurrent cascade, the flow rates and concentrations between stages, called the *internal parameters,* are more

complex because the feed for any one stage is a mixture of the product from the stage immediately upstream and the tails of the stage immediately downstream. Other internal parameters are the number of stages and the cut for each stage. All of the internal flow rates can be described as functions of the product flow rate and the cuts of individual stages. The concentration gradient along the cascade is a function of the separation gain and external concentration parameters, and approaches a maximum value given by

$$\frac{\mathrm{d}N}{\mathrm{d}s}=gN\,(1-N) \tag{23}$$

where s represents a given stage.

The separative performance of a single stage, and of a cascade as a whole, is greatest when the product flow rate is zero. This is essentially a steady-state situation in which no new feed is introduced and all the effort of the process goes into achieving the maximum concentration gradient. If enriched material is to be removed from the cascade, a net transport f_1 of material through the cascade must occur, which results in a decrease in the concentration gradient. The net transport is at a maximum $f_{\max}$ when the product flow rate is equal to the feed flow rate. In this case, there is no tails flow and the concentration gradient is zero; i.e., no separation of isotopes occurs. This also implies that the value of $f_{\max}$ with respect to the tails flow rate must be the same as the maximum value of the concentration gradient multiplied by the tails flow rate L''; i.e.,

$$f_{\max}=gN\,(1-n)\,L'' \tag{24}$$

Thus, the net transport and concentration must be optimized to give an efficient yield of enriched product from the cascade. The *separative work E* of a stage, which is the product of the concentration gradient and the net transport with respect to the tails flow rate, is defined as

$$E=\frac{\mathrm{d}N}{\mathrm{d}s}\cdot fL''=\left(N'-N\right)\left(N-N''\right)L'' \tag{25}$$

The most efficient operation of a stage occurs when the value of E is maximized. This is true when the net transport and the concentration are both one-half of their maximum values:

$$E_{\max}=1/4g^2N^2\,(1-\mathrm{N})^2\,L'' \tag{26}$$

The separative work of a stage is therefore concentration dependent. However, the *separative power* δU, obtained by dividing the separative work by the concentration terms, is an intrinsic property of a stage. Its maximum value $\delta U_{\max}$ is given by

$$\delta U_{\max}=\frac{E}{N^2\,(1-N)^2}=\frac{1}{4}g^2L'' \tag{27}$$

Two important special cases of cascades can be considered. In the *ideal countercurrent cascade,* streams (product and tails) that mix to form the feed of a given stage have the same isotopic abundance. For an ideal cascade, the relationship between enrichment factor and separation factor is

$$\alpha=\sqrt{q} \tag{28}$$

For any stage s, the cut is a function of the enrichment factor and the isotopic abundance of the target isotope in the feed:

$$\theta=\frac{1+\alpha R_{\mathrm{s}}}{(1+\alpha)\,(1+R_{\mathrm{s}})} \tag{29}$$

The number of stages S needed to permit a cascade to operate ideally is a function of the relative isotopic abundance of the initial feed R_0, the desired product abundance R_{p}, and the enrichment factor:

$$S+1=\frac{\ln\,(R_{\mathrm{p}}/R_0)}{\ln\alpha} \tag{30}$$

When the abundances in the feed and product are low, the relative abundances may be replaced by the mole fractions (N_0 and N_{p}) without significantly altering the results of the calculations.

Given a small value for the enrichment factor, typical for most separation processes, the number of stages required to establish an ideal cascade can be very large. For example, if enrichment of ^{235}U from its natural abundance of 0.71 % to 3.0 % is desired and a process is employed with a separation factor of 1.004 (typical for gaseous diffusion), the number of stages required in an ideal cascade for this enrichment is more than 2100.

The concentration gradient in each stage of an ideal cascade is given by

$$\frac{dN}{\mathrm{d}s}=\varepsilon N\,(1-N)=\frac{1}{2}gN\,(1-N) \tag{31}$$

This is the concentration gradient for optimum separative work (see Eq. 32), and by setting the net transport to its optimum value, one can see that an ideal cascade is the most efficient design for isotope separation. The separative power of this type of cascade is simply the separative power of the individual stages multiplied by the number of stages n; i.e.,

$$\Delta U = \delta U_{\max} n \tag{32}$$

The second case is that of the *square cascade* in which the flow rates of the feed streams of all stages are maintained at a constant value (the exception is the last stage because no recycling of a tails stream occurs at this point). This implies that the product and tails streams that mix to form the feed stream for a given stage usually have different isotopic abundances. The enrichment factor thus varies from stage to stage. The cut for stages in a square cascade can be arranged in two ways, either a constant value of θ for all stages or an arrangement in which the value is the same for alternate stages (i.e., all odd-numbered stages have one value θ_1, and all even-numbered stages have another value θ_2). In both cases the values are constrained by the equation

$$\theta_1 + \theta_2 = 1 + (P/F) \tag{33}$$

where P and F are the flow rates of the product and feed, respectively.

The constant cut situation ($\theta_1 = \theta_2$) is more common and is used in square cascades for the enrichment of various isotopes by distillation, thermal diffusion, and centrifugation. The maximum concentration gradient for a square cascade is given by

$$\frac{\mathrm{d}N}{\mathrm{d}s} = gN(1-N) - \frac{2P}{L}(N_{\mathrm{p}} - N) \tag{34}$$

and therefore, a lower value results for the separative power than for the ideal cascade when the net transport for the square cascade is optimized. In general, square cascades are less efficient than ideal cascades because some of the separative work done in each stage is lost when the product stream is mixed with a less highly enriched tails stream to form the feed for another stage. Despite this loss of separative work, square cascades are often preferred to ideal cascades as the throughput can be significantly greater and because capital costs are often lower. The total separative power of a square cascade is optimized around a mean value for the separative power of the individual stages; i.e.,

$$\Delta U = \delta U_{\mathrm{avg}} n \tag{35}$$

4.1.2. Discontinuous Processes

In a discontinuous process, a quantity of material is subjected to an enrichment process which increases the concentration of the desired isotope in the form or phase that remains in the apparatus and removes part of the depleted form or phase. The process is halted after a period of time, and the enriched material is removed. Material may be added to the apparatus continuously during the enrichment process, but the term discontinuous is applied because the process must be halted to recover the product. An example is *Rayleigh distillation,* which is used to enrich deuterium in the form of heavy water.

A rigorous mathematical treatment is not given here, but in principle, vapor above the liquid in the distillation apparatus is in equilibrium with the liquid and has a higher concentration of the more volatile species. The process begins with material having a relative abundance R and quantity Q of the more volatile isotope. The relative abundance of the isotope in the vapor phase is R', with quantity Q'; the abundance in the liquid phase R'', with quantity Q''. The separation factor is then given as in Equation (17) and the enrichment factor as in Equation (19). If a portion of the vapor is continuously removed, the concentration of the less volatile species in the liquid phase increases as material relatively rich in the more volatile species vaporizes to maintain the equilibrium. The separative power of an element employing such a discontinuous process is determined ideally by the separation gain and the cut:

$$\delta U = \frac{G}{2} g_0^2 \frac{1-\theta}{\theta} [\ln(1-\theta)]^2 \tag{36}$$

where G is the feed flow rate in moles per unit time. (The separative power per batch is simply U/G.) For Rayleigh distillation, the cut is optimized at ca. 0.8, which implies that much material is required to achieve a significant enrichment of deuterium.

This enrichment process is complicated by the presence of the target isotope in more than one form. At low concentration, deuterium exists predominantly as HDO, but as the concentration increases, D_2O dominates. Additional complexity is introduced by the fact that the volatility difference between H_2O and the deuterium-containing species decreases with increasing temperature; above ca. 230 °C, the volatility of D_2O is actually greater than that of H_2O.

4.2. Separation Processes

4.2.1. Distillation

Fractional Distillation. Distillation is one of the oldest methods of isotope separation. ASTON used this method unsuccessfully in 1919 in an attempt to isolate the isotopes of neon. UREY and others used distillation to isolate the first enriched fraction of deuterium from liquid hydrogen in 1932.

Distillation is a mass-dependent process because the difference in mass between isotopes results in differences in the reduced mass of molecules containing isotopes of an element. This, in turn, results in higher vibrational frequency for the lighter species, higher zero-point energy, and lower heat of vaporization. The last property is reflected in higher volatility of the lighter species at low temperature (e.g., <100 K). A simplified treatment, which assumes the molecules to be simple harmonic oscillators, gives the average energy of a molecular species as

$$E_{\mathrm{avg}} = \frac{1}{2}h\nu + \frac{h\nu}{e^{h\nu/kT} - 1} \tag{37}$$

As temperature increases, the second term of the expression dominates and tends to equalize the energies of the molecules. Because volatility is a measure of the number of molecules that have energy in excess of the heat of vaporization, this quantity also tends to be equalized at high temperature. Volatility is also affected by the electromagnetic interaction between molecules; this may change the relative volatilities of isotopically substituted species (e.g., $^{11}BF_3$ is always more volatile than $^{10}BF_3$ for this reason). Dipole – dipole interactions may also be important for some isotopically substituted species.

Table 7. Vapor pressure ratios for some isotopic systems

System	Temperature, °C	v_1/v_2
$^3He/^4He$	−271.0	7.0
H_2/HD	−259.4	3.61
	−252.9	1.73
H_2O/HDO	40.0	1.059
	100.0	1.026
H_2O/D_2O	0.0	1.254
	100.0	1.052
	240.0	0.9965
NH_3/ND_3	− 77.7	1.26
	− 33.6	1.11
CH_4/CH_3D	−182.5	1.0016
	−161.9	0.9965
C_6H_6/C_6D_6	57.0	0.973
$^{10}BF_3/^{11}BF_3$	−100.0	0.9925
$^{12}CO/^{13}CO$	−205.7	1.011
	−195.8	1.0071
$^{14}N_2/^{15}N_2$	−209.9	1.012
	−195.8	1.008
$^{14}N_2/^{14}N^{15}N$	−208.3	1.0062
	−195.5	1.0040
$^{14}NH_3/^{15}NH_3$	− 77.7	1.006
	− 33.6	1.0025
$^{14}N^{16}O/^{15}N^{16}O$	−163.6	1.032
	−151.5	1.0262
$^{16}O_2/^{16}O^{18}O$	−183.0	1.0054
$H_2^{16}O/H_2^{18}O$	0.0	1.010
	77.0	1.0065
	100.0	1.005
$^{14}N^{16}O/^{14}N^{18}O$	−163.5	1.0465
	−151.4	1.0368
$^{20}Ne/^{22}Ne$	−248.6	1.043
$^{36}Ar/^{40}Ar$	−189.2	1.002

For a pure substance, the volatility v is equal to the vapor pressure p. In an ideal mixture, the volatility of any one component is independent of other components and is defined as the ratio of the partial pressure p_{p} of the component in the vapor phase to the mole fraction of the component in the liquid phase N. The partial pressure is also given by the mole fraction n in the vapor phase multiplied by the total pressure p_{T}; therefore,

$$v = p_{\mathrm{p}}/N = P_{\mathrm{T}}\,(n/N) \tag{38}$$

A two-component system has two such equations:

$$\begin{aligned} v_1 &= p_{\mathrm{T}}\,(n_1/N_1) \\ v_2 &= p_{\mathrm{T}}\,(n_2/N_2) \end{aligned} \tag{39}$$

and the ratio of the two volatilities can be written as

$$v_1/v_2 = \frac{n_1/n_2}{N_1/N_2} \tag{40}$$

The numerator of this equation is the ratio of relative abundance of the components in the vapor phase, and the denominator is the ratio of abundance in the liquid phase. If the components are isotopes or isotopically substituted species, the ratio of volatilities is, according to the definition in Equation (17), the elementary separation factor, i.e.,

$$q_0 = v_1/v_2 \tag{41}$$

Most systems, however, are not ideal, and the volatility of individual components varies with composition, voltaility being proportional to the vapor pressure of the pure substance. Equation (41) remains valid, but the values of v_1 and v_2 must be determined experimentally to find the actual elementary separation factor. Table 7 lists vapor pressure ratios for a number of isotopically substituted systems at various temperatures.

A simple distillation column, with single boiler and condenser, uses gravity to establish a flow between the vapor rising in the column from the boiler at the bottom and the liquid falling from the condenser at the top. The more volatile species concentrates in the vapor phase and the less volatile in the liquid phase. Such an apparatus enriches an isotopic mixture only to the extent allowed by the elementary separation factor. To multiply the effect, many simple columns can be arranged in parallel and series to form a cascade. A more efficient method involves fractionating columns in which intimate contact occurs between liquid and vapor phases at many points throughout the column and continual exchange of isotopic species between phases can occur. Figure 10 shows one arrangement of a fractionating column, called a *bubble-cap plate* column. Vapor traveling upward from the boiler passes through the caps and bubbles through the liquid lying on top of each plate. When it reaches the top of the column, the vapor condenses and is added to the liquid on the top plate. Liquid overflowing from one plate passes down the column to the plate immediately below. The top plate has the highest concentration of the more volatile species in both the liquid and the vapor phases, whereas the bottom plate has the highest concentration of the less volatile species, with a concentration gradient established along the column. Thus, the fractionation column can be envisoned as a cascade, with each plate being a stage. Ideally, an equilibrium is established between the vapor and liquid phases at each plate, with the vapor enriched in the more volatile component to the extent allowed by the elementary separation factor. Such an equilibrium situation leads to the concept of *theoretical plates*. In practice, with a continual flow of feed material into the boiler at the bottom of the column and enriched product removal at the top, only a fraction (but usually a large fraction) of the elementary separation factor is realized for each plate.

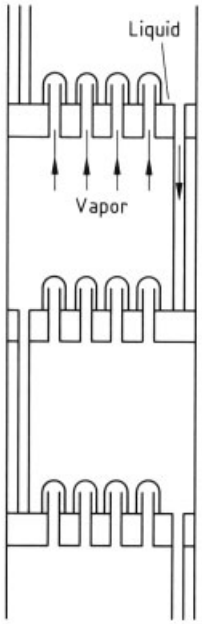

Figure 10. Bubble-cap column

Another arrangement of the fractionation column uses a packing material (beads, rings, or other material with a large surface area). Liquid condenses on the packing material and flows downward, in intimate contact with the vapor traveling upward at all times. This situation is equivalent to a cascade with continuous stages. A stage may be envisioned as the length of packed column over which the vapor attains the same concentration of the more volatile species as it would above one theoretical plate in a bubble-cap column. This length is known as the height equivalent of a theoretical plate (HETP).

Distillation was used for large-scale enrichment of heavy water in the Manhattan Project during World War II [62]. Large quantities of highly enriched D_2O were needed as a modera-

tor for nuclear reactors. Between January 1943 and October 1945, three plants were built and operated by Du Pont in the United States (Morgantown, West Virginia; Childersburg, Alabama; and Newport, Indiana) which produced more than 1 t of D_2O per month when operating at full capacity and a total of 20.7 t of D_2O during their overall operation. The distillation process enriched heavy water from a natural abundance of 0.0143 % (D_2O equivalent, actually present, almost exclusively, as HDO in normal water) to 88 – 92 % D_2O. This material was subjected to further enrichment to 99.8 % by electrolysis (see Section 4.2.8).

The distillation cascade at the Morgantown facility, the smallest of the three, consisted of eight stages: half of the columns in the first stage were comprised of *tunnel-cap plates* (similar to bubble-cap plates), whereas the remainder of the first stage and the second and third stages had bubble-cap plates. The last five stages were packed columns. The first stage contained five pairs of large columns in series, ca. 4.6 and 3.6 m in diameter and ca. 30 m tall, with 79 and 89 plates, respectively, in each column. The second stage was made up of a pair of slightly smaller columns with 71 and 82 plates; the third stage consisted of a single column, 1 m in diameter, with 71 plates. The remaining columns were packed with Raschig rings to a height of 22 m. The final column had an HETP of ca. 0.5 m (and thus had ca. 45 theoretical plates). Figure 11 shows a schematic of the distillation unit. The first-stage columns were operated at low pressure to take advantage of the greater difference in volatility between H_2O and HDO (or D_2O) at lower temperature. The large holdup capacity of the system (amount of water contained in the system at any given time) meant that it was slow to react to changes in process parameters and three months was usually required for a change in equilibrium.

The efficiency of distillation was low; ca. 2.8 % of the quantity of D_2O present in the feed material was actually recovered at the Morgantown facility, the most efficient of the three plants. This meant that for every 10^3 t of normal water processed, 4.2 kg of enriched heavy water was obtained. The 20.7 t of material from the three distillation facilities was produced at an average cost of \$ 0.46 per gram.

Distillation of liquid hydrogen was also examined by scientists and engineers connected with the Manhattan Project [62, p. 82]. This process was abandoned at the developmental stage in favor of water distillation because of the mature state of technology available for the latter process, and the lack of experience in operating equipment and handling large quantities of materials at temperature close to absolute zero. The efficiency of hydrogen distillation for the separation of deuterium is much greater than that of water distillation. Table 7 shows that at the respective triple points, the ratio of vapor pressure for H_2/HD is three times as great as the ratio for H_2O/D_2O (the difference is six times when the H_2O/HDO ratio is considered). The process is used on a laboratory or pilot-plant scale to enrich small quantities of deuterium.

The isotope ^{18}O is also obtained by distillation of water. This process is complicated because deuterium separation also occurs during water distillation and has a larger enrichment factor. To improve efficiency, water that has been depleted in deuterium is used as the feed for oxygen isotope separation, so the waste product from heavy-water enrichment is a good feedstock.

Another isotopic system examined was boron, in which ^{10}B and ^{11}B can be separated by distillation of BF_3 [63]. This separation is aided by the fact that fluorine is mononuclidic. The vapor pressure ratio of $^{10}BF_3$ to $^{11}BF_3$ is less than 1.0 at all temperatures; i.e., the heavier isotopic species is more volatile. Because ^{10}B, with 19.8 % natural abundance, is usually the desired species, this system lends itself to distillation by a discontinuous, batch process. In 1945, again in connection with the Manhattan Project, a large quantity (50 kg) of boron, enriched in the ^{10}B to more 90 %, was produced by distillation. The process used the dimethyl ether complex of boron trifluoride, $(CH_3)_2O \cdot BF_3$, to avoid the lower temperature needed if pure boron trifluoride is used. During the developmental phase of the project the vapor pressure ratio of the isotopically labeled complexes was found to be very close to 1.0 (again, the higher mass compound was the more volatile) and could not, in itself, account for the magnitude of the experimentally determined separation factor. Further experiments showed that an exchange reaction involving free $^{11}BF_3$ and the $(CH_3)_2O \cdot {}^{10}BF_3$

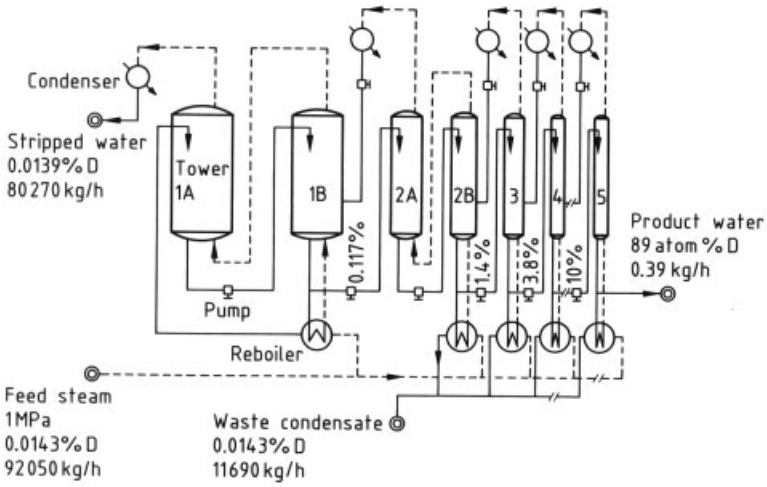

Figure 11. Heavy-water distillation unit at Morgantown

complex in the vapor phase exhibits an equilibrium isotope effect (see Section 2.2.2) that is responsible for the major portion of the separation of boron isotopes.

Helium isotopes have the highest ratio of vapor pressures listed in Table 7 and could be separated very efficiently by distillation. Problems arise because of the need to work at very low temperature. Isotopes of chlorine can be separated by distillation of carbon tetrachloride. Interestingly, the most volatile species in the isotopic mixture contains the light isotope of chlorine, ^{35}Cl, and the heavy isotope of carbon, ^{13}C.

Molecular Distillation. Under ordinary distillation conditions, an equilibrium is established between the vapor and liquid phases at each plate or on the surface of the packing material in a fractionation column. This is a dynamic situation in which atoms or molecules are continually exchanged between phases by evaporation and condensation, with the rates of these two phenomena being equal at equilibrium. Increased efficiency of distillation results if a nonequilibrium situation can be maintained in which the species evaporating from the liquid are removed from the system immediately by condensation on a nearby surface; this process is called *molecular distillation.* In this case, the composition of the material evaporating from the liquid is determined only by the relative evaporation rates of the components of the liquid phase. Langmuir's hypothesis states that at low temperature and, therefore, low pressure, the evaporation rate is independent of any material already present in the vapor phase. Thus, the evaporation rate is equal to the condensation rate under equilibrium conditions. The condensation rate is determined by assuming the vapor to be an ideal gas that follows the kinetic theory of gases. From this theory, the number of molecules (or atoms) and their velocity may be found, which leads to

$$R_{\mathrm{evap}} = \frac{p}{(2\pi MRT)^{1/2}} \tag{42}$$

where R_{evap} is the evaporation rate, p the partial pressure, M the molecular (or atomic) mass of the species, R the gas constant, and T the absolute temperature. Langmuir used this equation to determine the vapor pressure of pure metals at various temperatures. For isotope separation in a situation analogous to that developed for ordinary distillation in Equations (28) – (30), the elementary separation factor for a two-component system under ideal conditions of molecular distillation is given by

$$q_0 = \frac{R_{\mathrm{evap}_1}}{R_{\mathrm{evap}_2}} = \frac{V_1}{V_2} \cdot \frac{M_1^{1/2}}{M_2^{1/2}} \tag{43}$$

Thus, molecular distillation has a separation factor that is larger by a factor of the square root of the mass ratio than ordinary distillation under equilibrium conditions. The method has been

used successfully to separate isotopes of mercury on a laboratory scale [64] and as a means for separating lithium isotopes [65].

Some problems arise with the molecular distillation process that can severely limit its effectiveness. The elementary separation factor may be reduced from the ideal value if removal of the evaporating material is not quantitative. Atoms or molecules can return to the liquid phase after initial evaporation because of collisions with other atoms or molecules before condensing on the receiving surface or because of reevaporation from the receiving surface. These phenomena may change the ratio of evaporation rates or at least reduce the flow rate of the process. In addition, because boiling of the liquid phase does not occur, mixing of that phase to maintain a homogeneous distribution of components does not occur readily. The surface layer of the liquid becomes depleted in the most volatile species by evaporation more rapidly than diffusion through the liquid can replenish that component. In effect, the mole fraction of the most volatile species in the feed becomes lower than that of the original mixture and the enrichment gain suffers as a result. Apparatus and process conditions that allow mixing of the liquid and maximize removal of evaporated material are available. The best operating conditions for molecular distillation are very low pressure and relatively low temperature, but optimum combinations of these conditions usually restrict the number of isotopic systems that can be enriched by this process.

4.2.2. Diffusion

Gaseous Diffusion. The kinetic theory of gases also describes the process of isotope separation by diffusion through a porous barrier or membrane, which is called gaseous diffusion. If a mixture of gases in a container is at thermal equilibrium, the root mean square velocity of any of the species is

$$v_{\mathrm{rms}} = (3RT/M)^{1/2} \tag{44}$$

This indicates that the lighter molecules or atoms have a higher velocity. The number of collisions c, per unit area, with the walls of the container is given by

$$c = n' \, (RT/2\pi M)^{1/2} \tag{45}$$

where n' is the number density of the species in the container. If one wall is perforated with small holes and the pressure on the other side of the wall is such that essentially no molecules or atoms return to the original container, Equation (45) also gives the rate of *molecular effusion* w through the openings in the wall. (The term effusion is used when the mean free path of the molecules is greater than the diameter of the pores in the barrier). If the ideal gas law is used to substitute pressure for the number density, the rate of effusion is given by

$$w = p/\,(2\pi RT/M)^{1/2} \tag{46}$$

and is the basis for the *Knudsen cell,* which is used to determine vapor pressures. Note that this equation is the same as that derived for molecular distillation (Eq. 42). In a mixture of gases, the partial pressure of any component is given by the total pressure multiplied by the mole fraction of that component. For a two-component mixture, the mole fraction can be expressed as N'', and $1 - N''$, where N'' is the mole fraction of the lighter species. A ratio of the effusion rates for the two components gives

$$\frac{w_1}{w_2} = \frac{N''p}{(1-N'')\,p} \cdot \frac{(2\pi RTM_2)^{1/2}}{(2\pi RTM_1)^{1/2}} \tag{47}$$

If the difference $1 - N''$ is close to the original mole fraction in the container (i.e., N), the elementary separation factor for a gaseous diffusion process is simply the square root of the ratio of masses of the two species:

$$q_0 = (M_2/M_1)^{1/2} \tag{48}$$

which may be approximated by

$$q_0 \simeq 1 + \frac{M_2 - M_1}{M_2 + M_1} \tag{49}$$

In fact, the separation factor is only proportional to the mass ratio because of non-idealities in the process. The size of the openings and the mean free path of the molecules or atoms involved lower the separation efficiency. Although the efficiency increases as the size of the openings decreases and the mean free path increases, these quantities are restricted by the physical parameters of the barrier and of the effusing material. Also, the pressures that exist in large-scale

diffusion systems cause reductions in the concentration gradient across the membrane in two ways. A significant back pressure, on the side of the barrier receiving the gas mixture enriched in the lighter species, results in diffusion in the opposite direction. An increase in the pressure on the side of the barrier containing the feed causes viscous flow through the membrane that does not result in isotope separation. A particular membrane has a separation efficiency that depends on operating pressure and temperature and on properties (e.g., viscosity) of the gas. Surface effects lower the separation efficiency by reducing the mean free path of particles colliding with the container walls.

Gaseous diffusion has been used successfully for several isotopic systems. The largest application of the diffusion process is for the enrichment of the fissile isotope of uranium (^{235}U) for nuclear power plants, nuclear propulsion units for submarines and surface ships, and nuclear weapons. The compound used for gaseous diffusion is uranium hexafluoride (UF_6) which has several distinct advantages. It sublimes at 56.5 °C (101.3 MPa), has a relatively low viscosity (1.76 mPa · s at 101.3 kPa and 20 °C), and contains only the mononuclidic element fluorine in addition to uranium. Drawbacks to this compound are its small mean free path and its reactivity; uranium hexafluoride is a good fluorinating agent, which reacts with most organic compounds, with water, and with many metals. Construction materials must be selected with this chemistry in mind. Experimental diffusion apparatuses include fluorinated hydrocarbons and metals such as silver and aluminum, which form a protective fluoride surface layer. Water must be rigorously excluded because its reaction with uranium hexafluoride forms uranyl fluoride (UO_2F_2), a high-melting solid, and hydrofluoric acid, which presents severe corrosion problems.

The first diffusion plant for the separation of uranium isotopes was constructed in Oak Ridge, Tennessee, and became operational in 1945. More cascades were added to the Oak Ridge facility and additional plants were constructed in Paducah, Kentucky, and Portsmouth, Ohio, in the 1950s [14, p. 195]. The Oak Ridge and Portsmouth plants each have about 4000 stages in series, whereas the Paducah plant has two cascades in parallel, both with about 880 stages. The power requirement for operation of the plants at full capacity is 6000 MW, and the total separative work available is ca. 17 000 t SWU (separative work units) per year. At one time, operation of the three plants was coordinated, with the naturally occurring isotopic mixture of uranium processed at the Paducah facility to enrich ^{235}U from 0.0714 to 0.96 %. This material served as the feedstock for the other two plants. The Oak Ridge plant enriched the isotope to ca. 4 %, which could be used as fuel in nuclear power reactors. Some of this material was sent to the Portsmouth facility where it served as additional feed for enrichment to >96 % and was useful for nuclear propulsion reactors in submarines and for nuclear weapons. At present, the demand for enriched uranium is low, and the Oak Ridge plant has been shutdown. The other two facilities operate at reduced capacity. Diffusion plants for the enrichment of uranium are also located in Great Britain, France, Russia, and China.

Mass Diffusion. Mass diffusion also exploits the differences in molecular velocity of isotopically substituted species. This is essentially the process responsible for the isotopic fractionation of nitrogen-containing compounds in natural gas (see Section 2.3), which leads to separation of the lighter, ^{14}N-containing materials by diffusion through large quantities of methane. The mass diffusion technique allows a mixture of isotopically substituted compounds to diffuse through a third gaseous material, called the separating agent, as the agent flows through a porous barrier or screen [66]. *Sweep diffusion* is based on the same physical phenomenon but does not employ a barrier, it is the process actually involved in the natural enrichment of nitrogen compounds [67]. Both methods are based on the fact that a lighter species has a higher diffusion coefficient through the separating agent.

For example, if a container is partitioned into two areas by a porous screen and one side initially contains the isotopic mixture and the other side the separating agent, diffusion through the screen eventually results in an equilibrium in which the concentration of each species is the same in both areas. However, if the gases on the side initially containing the separating agent are examined some time before equilibrium is reached, the ratio of the lighter to heavier iso-

topic species is greater than their initial ratio (i.e., the lighter material is enriched) because the lighter species diffuses through the separating agent faster than the heavier material. The elementary separation factor for this simple process may be approximated by the ratio of the diffusion coefficients:

$$q_0 = D_1/D_2 \tag{50}$$

where the subscripts 1 and 2 represent the lighter and heavier species, respectively.

Certain properties must be considered when selecting a separating agent: selectivity of the agent for the species to be separated (also known as separability of the materials in the agent), high diffusion coefficients of the isotopic species in the agent, and easy separation of enriched material from the separating agent. No chemical interaction should occur between the separating agent and the isotopic species. The first two factors work at odds with each another. The selectivity of the separating agent increases with increasing molecular weight, but the diffusion coefficient decreases at the same time. Generally, a separating agent with a molecular weight close to that of the materials used for enrichment gives the best combination of properties.

Separation of the isotopic components from the agent is simplified by selecting an agent with a boiling point much higher than those of the isotopic materials. This aids in removing and recycling the separating agent by condensation from the agent – product mixture. When the molecular masses of the materials are very similar, the elementary separation factor may be approximated by

$$q_0 \simeq 1 + \frac{M_2 - M_1}{M_2 + M_1} \cdot \frac{M_0}{M_0 + (M_1 + M_2)/2} \tag{51}$$

where M_1 and M_2 are the masses of the lighter and heavier isotopically substituted species, respectively, and M_0 is the mass of the separating agent. This equation resembles that given for gaseous diffusion (Eq. 49) but includes a second mass factor contributed by the separating agent that is always less than 1.0. As a result the separation factor for mass diffusion is lower than that for gaseous diffusion for any given isotopic system. The advantage of the process is that it is less technically demanding and is amenable to multiplication of the elementary separation factor by establishing countercurrent flows in a mass diffusion apparatus. Unlike gaseous diffusion, which can only be operated on a large scale, this technique can be used on any scale.

A mass diffusion column, developed by Benedict and Boas in the early 1950s, is shown schematically in Figure 12. It comprises three concentric columns surrounded by a jacket containing cooling water.

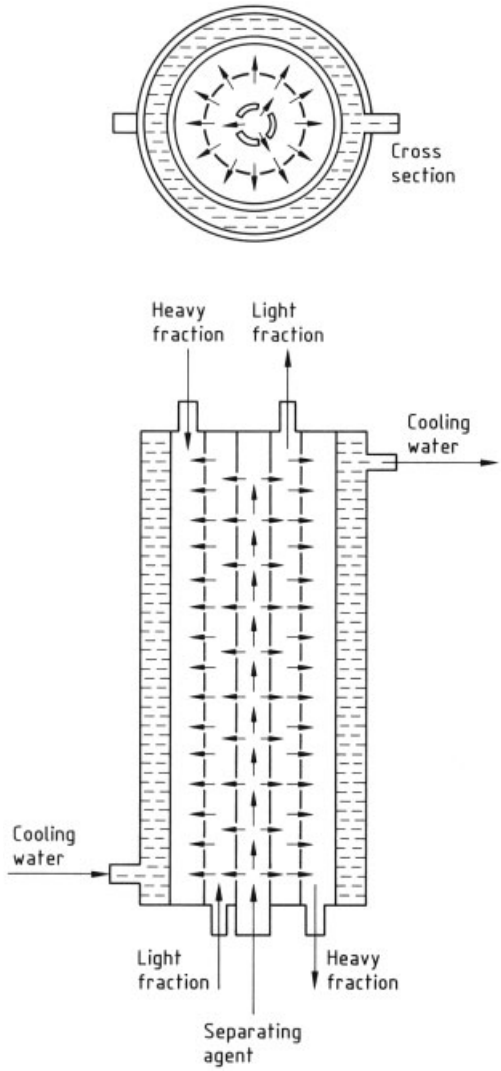

Figure 12. Mass diffusion column

The central column is perforated so that gases can pass into the next column. This second column contains some of the isotopically substituted mixture of gases to be separated. A porous

screen separates the second column from the third, which contains the remainder of the isotopic mixture. Countercurrent flow is created by causing the gas in the second column to rise toward the top of the apparatus and gas in the third column to flow to the bottom.

The gaseous separating agent, produced by an evaporator at the bottom, rises through the central column, diffuses through the holes into the second column, and then through the screen into the third column, and condenses on the cooled outer wall. The liquid separating agent flows to the bottom of the column and is returned to the evaporator to be recycled. Enrichment of the isotopic mixture is a result of the flow of separating agent from the central column to the outer wall, which carries the isotopic mixture of gases with it and establishes a radial concentration gradient. This induces back diffusion of the isotopic species toward the center. Because the diffusion coefficient of the lighter isotopic compound is greater than that of the heavier one, the former flows back toward the center at a higher rate. Thus, gas flowing upward in the second column is enriched in the lighter isotope and gas flowing downward in the third column is enriched in the heavier isotope. In this kind of diffusion column, separation occurs mainly at the screen where a large isotopic concentration gradient exists between the inner portion of the third column and the outer portion of the second column.

The sweep diffusion apparatus is essentially the same in design, with a central column for introducing the separating agent but no screen dividing the remainder of the apparatus into two columns. The flow of separating agent again creates a concentration gradient along the radius of the column and allows back diffusion of the isotopic mixture. Here, the isotopic gradient is essentially linear across the radius of the outer column, and the elementary separation factor can be calculated theoretically as

$$\ln q_0 = \frac{RT\phi_0}{P}\left[\frac{1}{D_1} - \frac{1}{D_2}\right] r_a \ln \frac{r_a}{r_b} \tag{52}$$

where R is the ideal gas constant, T is the absolute temperature, ϕ is the flow rate of the separating agent, p is the pressure, D_1 and D_2 are the diffusion coefficients of the light and heavy isotopic species, respectively, r_a is the radius of the central column, and r_b is the radius of the outer column. Countercurrent flows are established by differences in the density of the isotopic mixture in different parts of the column. The material is less dense near the center of the column because of the concentration gradient and enrichment of the lighter isotope there, and an upward flow is created. A higher density and a downward flow are created near the outer wall. Countercurrent circulation multiplies the elementary separation effect so that the total separation factor is proportional to the length of the column.

Operation of these types of diffusion column is complicated by several factors. Convection currents within the column may be established which mix gases that have different isotopic ratios, either radially or lengthwise in the column. This reduces the overall separation efficiency because reseparation of these mixtures is required and can be especially troublesome in the sweep diffusion method. The isotopic mixture of gases may be trapped in the separating agent when it condenses on the outer wall of the apparatus, or the isotopic compound may be soluble in the agent. Such trapped and dissolved material is unavailable for enrichment by back diffusion. If heavier isotopic species are trapped at the outer wall (where the concentration is highest) and recycled along with the separating agent, these species can contaminate the lighter fraction near the exit holes of the central column.

Mass diffusion was originally employed for the separation of hydrogen from methane by using steam as the separating agent. Later, the carbon isotopes ^{12}C and ^{13}C were separated in the same way. In 1934, HERTZ developed a mercury vapor diffusion pump to separate the isotopes of neon and achieved a fivefold enrichment of ^{22}Ne. The same type of apparatus has been used to enrich ^{18}O, ^{35}Ar, ^{15}N, and ^{13}C. Researchers in the former Soviet Union have used column-type mass diffusion for separating carbon isotopes, with xylene, water vapor, or $C_2H_2Cl_2$ as the separating agent, and for separating isotopes of neon and argon with water vapor as the agent.

Thermal Diffusion. Thermal diffusion relies on the difference in migration rates of isotopically substituted materials between regions of different temperature. The flow rate of ideal gases is inversely proportional to the square root of the mass at constant temperature (Eq. 42 and 46). If a temperature difference exists, the net

flow of any given component of the gas can be expressed by

$$\phi = \frac{P}{(2\pi RM)^{1/2}} \left[\frac{1}{T_1^{1/2}} - \frac{1}{T_2^{1/2}} \right] \tag{53}$$

where T_1 and T_2 are the lower and higher temperature, respectively, and M is mass of the component. According to this equation, the elementary separation factor for a two-component system should be the square root of the ratio of the masses of the components, the same as that for gaseous diffusion given by Equation (48). This, however, is an oversimplification that does not take into account all of the factors that influence the thermal diffusion of the components.

The complex theoretical treatment of the distribution of molecular velocities in a nonuniform state is not presented here, but it does result in derivation of the elementary separation factor in the following form:

$$\ln q_0 = a \ln (T_2/T_1) \tag{54}$$

where a is the thermal diffusion factor, which is a complex function of the molecular masses, the relative proportions of the components of the mixture, and specific integrals determined by intermolecular interactions of the components. The thermal diffusion factor can be calculated theoretically; for isotopic mixtures, to a first approximation it is

$$a = \frac{105\,(M_2 - M_1)}{108\,(M_2 + M_1)} \tag{55}$$

This value is derived by assuming perfectly elastic spherical particles having no attractive potential. The thermal diffusion factor is then determined experimentally, and a ratio R_T is used to compare the experimental to the theoretical value:

$$a_\mathrm{calc} = a_\mathrm{theo} \cdot R_\mathrm{T} \tag{56}$$

Values for R_T may range from zero to unity. For the $^{12}CH_4/^{13}CH_4$ system, R_T is 0.33; for H_2/D_2, it is 0.5; and for $^{20}Ne/^{22}Ne$, it is 0.65. For most isotopic systems, the separation factor for thermal diffusion is lower than that for either gaseous or mass diffusion. Advantages of this method are its simplicity and the fact that it is applicable to liquid and gaseous systems [68].

A simple thermal diffusion column was developed in 1938. It consisted of a long vertical tube with a conducting wire running along the axis. The tube and reservoirs attached to either end of it hold the mixture to be separated. To establish a temperature gradient across the tube, the outer wall is cooled while the wire is heated electrically. The temperature gradient not only allows elementary separation to take place across the width of the tube, but also causes countercurrent flow along its length. A gas stream enriched in the lighter isotopic species travels upward near the wire, while a stream enriched in the heavier species travels downward near the outer wall. Each stream is continually enriched by exchange, and the multiplication of the elementary separation factor by this process can be several thousand. Early experiments employed a column ca. 2.5 m long and 1 cm in diameter, with a temperature difference between the wire and the wall of ca. 1000 °C. This apparatus achieved threefold enrichment of the concentration of ^{22}Ne. Later devices used interconnected tubes, with a diameter of 8.5 mm and a total length of 13 m, to produce ^{79}Br at 85.5 % abundance by using HBr as the gaseous species [69]. Thermal diffusion is currently used to enrich ^{15}N and noble gases on a production scale.

4.2.3. Gas Centrifugation

The idea of separating isotopes by means of gravitational or centrifugal fields was first proposed by Aston and Lindeman in 1919 in connection with their work on the isotopes of neon. A discontinuous method of centrifugal separation was suggested by Mulliken in 1922. The first successful isotopic enrichment by a continuous process was demonstrated by Beams et al. in 1936 and used to separate isotopes of chlorine. An effort to develop a centrifugation process for the enrichment of ^{235}U as a replacement for the gaseous diffusion processes has been abandoned [70].

To calculate the elementary separation factor for the centrifugal process, consider an ideal gas of molecular mass M and density d [$d = (MP)/(RT)$, where P is pressure] contained in a cylinder (rotor) rotating at an angular veloc-

ity ω. A very thin annulus of gas at a radius r_1 in the rotor has a pressure p given by

$$p = p_0 \exp \frac{M\omega^2 r_1^2}{2RT} \tag{57}$$

where p_0 is the pressure at the axis of the rotor, R is the ideal gas constant, and T is the absolute temperature. The pressure at this annulus is developed by the centrifugal potential $1/2\,\omega^2\,r^2$. Because the effect is dependent on molecular mass, in a mixture of two compounds each would have a partial pressure given by Equation (57). The ratio of partial pressures defines the elementary separation factor for the process:

$$q_0 = \exp \frac{(M_2 - M_1)\,\omega^2 r^2}{2RT} \tag{58}$$

If separation takes place across the entire width of the cylinder, r is the radius of the rotor. Often, however, the effect takes place in a narrow annulus of gas near the wall of the rotor. The value of r then is the difference between the radii defining the inside and outside edges of the annulus, i.e., $r_2 - r_1$. Note also that separation depends on the absolute difference between the masses of the components and not the relative mass difference. When all other factors are constant, the separation factor is the same for a mixture of heavy isotopes as for a mixture of light isotopes having the same mass difference. The separation factor can be increased significantly by increasing the peripheral speed of the rotor and by lowering the temperature. As temperature increases, the average molecular velocity of the gas increases, diminishing the concentration gradient established by the centrifugal force.

Three types of centrifuge are used for isotope separation. The *evaporative centrifuge* consists of a rotor with a hollow shaft along its axis. The apparatus is evacuated, and a liquid isotopic mixture is introduced at the bottom of the rotor. When the rotor is spun, the liquid is dispersed in a thin layer over the outer wall. Heat is then applied to evaporate the liquid. Vapor diffuses toward the center of the rotor and is removed through the hollow shaft. Diffusion takes place under the influence of the centrifugal field; thus, material near the axis is enriched in the lighter isotopic species. This technique has been employed successfully to separate isotopes of chlorine by using carbon tetrachloride and isotopes of bromine by using ethyl bromide. Drawbacks of this process include the fact that it is discontinuous; as separation proceeds and the charge is depleted, the concentration of the lighter isotopic species removed at the center of the machine decreases. The efficiency of separation cannot exceed the elementary separation factor and only approaches that value at very slow withdrawal rates at the beginning of the withdrawal process. If only a portion of the liquid is evaporated from the outer wall, enrichment of the heavier species can be multiplied to some extent. This may be thought of as analogous to the batch distillation.

The first continuous separation process was achieved with the *cocurrent centrifuge.* In this apparatus, a gaseous mixture of the isotopic species is introduced at the bottom of the spinning rotor in two concentric streams, one near the wall of the cylinder and one closer to the axis. The streams flow upward and are removed at the top. The centrifugal field causes radial transport of the components of each stream. The heavier isotopic material in each stream is carried further toward the outer wall of the rotor because of its higher molecular mass. Opposing the transport of the components is diffusion toward the center, which is greater for the lighter component. These phenomena result in enrichment of the inner stream in the lighter isotopic species and the outer stream in the heavier species. As with the evaporative technique, enrichment by the cocurrent centrifuge is limited to the elementary separation factor, but methods have been proposed to multiply the effect. One technique divides the interior of the rotor into a series of vertically stacked compartments such that the inner part of one is connected to the outer part of the next higher compartment. This arrangement forces mixing of the inner and outer streams of gas, creating a situation analogous to column fractionation (see Section 4.2.1). Each compartment can be thought of as a plate of the column.

The most successful means of multiplying the elementary separation factor has been achieved with the *countercurrent centrifuge* (Fig. 13). The feed gas mixture is introduced into the rotor along the axis, near the middle, from a hollow central shaft. The product and tails are withdrawn by scoops at the top and bottom of the rotor, respectively. The entire rotor assembly is encased in an evacuated vessel to minimize drag

on the rotating cylinder. The bearings are critical to the apparatus because they must support the rotor and allow it to turn freely, but must damp out any vibrations that would disrupt the dynamic stability of the rapidly spinning rotor.

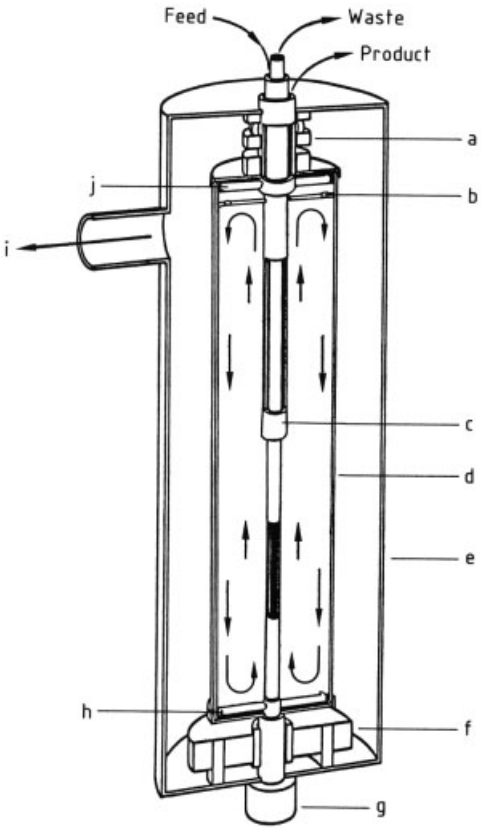

Figure 13. Countercurrent gas centrifuge
a) Magnetic bearing and damping assembly; b) Rotating baffle; c) Center post (stationary); d) Rotor; e) Casing; f) Electromagnetic motor; g) Bottom bearing; h) Bottom scoop (waste); i) To vacuum system

As with the countercurrent mass diffusion column, the elementary separation effect is multiplied by establishing countercurrent flows along the length of the rotor. A concentration gradient is thus established in the axial direction which, in a rotor with a large ratio of length to radius, is greater than the radial concentration gradient created in either the evaporative or the cocurrent design. Flows are induced in the gas by the action of stationary scoops or by creation of a thermal convection profile by heating one end of the rotor and cooling the other. The ideal flow profile would consist of a very thin layer of gas moving in one direction near the outer wall of the rotor and an equal stream in the other direction spread out over the remainder of the radius of the cylinder. The thermally induced flow profile approximates this situation most closely.

The concentration gradient in a countercurrent centrifuge can be determined by solving a complex partial differential equation with terms that depend on the flow rates and the flow profile. Including the physical properties of the specific isotopic materials and the parameters of the apparatus allows calculation of the theoretical maximum separative power of the countercurrent machine as

$$\delta U_{\max} = (\pi/2)\, D d g^2 Z \tag{59}$$

where D is the diffusion coefficient of the target isotopic species, d is the density, Z is the length of the rotor, and g is given by

$$g = \frac{(M_2 - M_1)\,\omega^2 r^2}{2RT} \tag{60}$$

(the definition of terms is the same as in Eq. 57). Thus, separative power is a function of the length of the rotor, the operating temperature, and most significantly, the fourth power of the peripheral velocity. Separation efficiency is given by the actual value of the separative power under operating conditions. This is proportional to the theoretical value and may be evaluated in four parts:

$$E = \delta U_{\mathrm{act}}/\delta U_{\max} = E_{\mathrm{c}} E_{\mathrm{i}} E_{\mathrm{e}} E_{\mathrm{f}} \tag{61}$$

where E_{c} is the circulation efficiency determined by the internal flow rates, which, consequently, gives the actual concentration gradient in the rotor; E_{i} is the nonideality efficiency, which allows for the fact that product and tail are withdrawn from an operating centrifuge; E_{e} is a factor derived by consideration of the turbulent flow and end effects in the rotor; and E_{f} is the flow efficiency that relates the actual flow profile to the ideal profile. Total efficiency factors of 0.75 and greater have been achieved in experimental situations. A distinct advantage of the countercurrent centrifuge is the large separation factor available from a single machine. This is a consequence of internal recirculation and results in a low throughput for each unit. Therefore, cascades of such machines require many fewer stages than other enrichment techniques (especially gaseous diffusion), and each stage is made

up of a large number of individual units. This allows a degree of flexibility in an enrichment facility because units can be added easily or isolated from each stage to increase or decrease the amount of material processed.

Engineering problems associated with the construction of practical centrifuges include construction materials for the rotor, dimensions of the machine, and the bearing assemblies necessary to support the rotating cylinder. Because the separative effect is affected most significantly by the peripheral velocity, much effort has been concentrated on developing materials and fabrication techniques for centrifuge rotors. The amount of gas in the machine at any time is small, and its mass contributes negligibly to the total mass of the rotating components. Therefore, the stress on the rotor is essentially a result of its own mass, and the maximum peripheral velocity is given by the square root of the ratio of the ultimate tensile strength of the rotor material to its density (i.e., the square root of the specific strength of the material). High-strength steels have good tensile properties, but their high densities lead to lower specific strengths than aluminum and titanium alloys. Composite materials (glass, carbon, or other fibers in a polymer matrix) have the highest specific strengths but are limited by problems in fabricating homogeneous materials that will be stressed equally over the entire rotor in a centrifugal field. Other material considerations include chemical interactions between the construction materials and the gas mixture used for separation. Additional properties of concern include the behavior of the materials for long periods of time under stress, including their fatigue and creep resistance.

The dimensions of the rotor also interact with the properties of the rotor material. When the frequency of rotation coincides with the natural frequency of the rotor material (the *critical frequency*), a resonance is established that results in axial vibration of the rotor. The critical frequencies (simply called *criticals*) are a function of the rotor length-to-diameter ratio and the modulus of elasticity of the rotor material. A machine that operates at speeds below the first critical is known as a *subcritical centrifuge,* whereas one operating above the first or any subsequent higher critical is termed *supercritical.* Operation of supercritical centrifuges involves the added problems of acceleration of the rotor through one or more criticals and maintenance of an operating speed between resonance frequencies.

Most of the work on the gas centrifuge has been directed toward enrichment of uranium isotopes; thus, specific information about equipment and capacities is classified by the governments involved. Experimental machines for this separation have been built and operated in the United States, Great Britain, France, Germany, Russia, the Netherlands, Italy, and Japan. A facility located in the Netherlands and owned by the joint Dutch – English – German consortium URENCO, with a capacity of about 300 t of separative work per year, is currently producing enriched uranium for nuclear power reactors. Another production facility is being operated in Japan. A demonstration plant was constructed at Oak Ridge, Tennessee, with 120 machines that operated for more than two years to provide data on reliability. Three production facilities with over 40 000 centrifuges were originally scheduled to be brought on line in the late 1980s but, because of the decrease in demand for enriched uranium, the program was canceled. Centrifugation has recently been transferred to the private sector for the separation of non-radioactive isotopes.

4.2.4. Nozzle Separation

Nozzle separation was developed in the mid-1950s [71]. The separative effect relies on the difference in diffusion rates and the effect of a centrifugal field on isotopically substituted species in an expanding stream of gas flowing at supersonic speed. The original apparatus employs a tube with a small nozzle from which the gas mixture to be separated flows at high velocity. The stream of gas is permitted to expand into an area adjacent to the nozzle where a pressure gradient in the gas is established transverse to the direction of flow–higher pressure at the center of the stream and lower pressure at its periphery. As described for centrifugal separation, the rate of diffusion of gases in a pressure gradient depends on the mass of the species involved; lighter species diffuse at a faster rate and become enriched in areas of lower pressure. Therefore, gas at the periphery of the expanding jet is enriched in the lighter isotopic species. The gas stream is then divided, and the two components–

one enriched in the heavier species, the other in the lighter species–are collected separately.

The elementary separation factor for the nozzle process is a complex function of several factors, including the masses of the isotopic species, the inlet pressure, and the expansion ratio of the gas. Unlike other enrichment techniques described here, the elementary separation factor in this case is also related directly to the cut taken in the collection portion of the device. Because of these complicated variables, the factor is determined experimentally for any given nozzle and collector configuration and operating conditions.

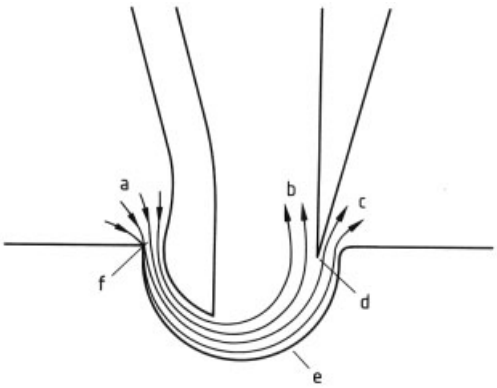

Figure 14. Nozzle separation apparatus
a) Feed gas, 5 mol% UF_6, 95 mol% He; b) Light fraction enriched in He and ^{236}U; c) Heavy fraction depleted in He and ^{236}U; d) Knife edge; e) Curved wall; f) Nozzle

The apparatus described above has several drawbacks because large pressure differences must be maintained between various parts of the device and the throughput is low. For the enrichment of uranium isotopes, a different design has been devised; it is shown in cross section in Figure 14. In this apparatus, an approximately semicircular groove is cut into a long piece of metal. Two blades extend into the groove to create the nozzle (actually a slit in this arrangement) and the collector. The first blade has a curved side with a smaller radius of curvature than the groove, whereas the second has a knife edge. Gas is forced through the slit formed by the groove and the first blade. Because of different radii of curvature of the two pieces, the gas expands in the space between the blade and the groove. The gas is made to expand at supersonic speed by maintaining a pressure difference across the nozzle. The gas stream follows the curved wall of the groove and is then divided into two components by the knife edge. The gas traveling along paths of smaller radii of curvature is enriched in ^{235}U, whereas that which follows the wall more closely is enriched in ^{238}U.

The separating effect in this nozzle process is mainly a result of the centrifugal forces imposed on the gas stream. Lighter isotopic species are more easily deflected into a curved path and, thus, are preferentially guided into the center section of the collection part of the device by the knife edge. A large proportion of helium is used in the feed gas. Because of its low atomic mass, helium increases the rate of gas flow and the centrifugal force on the uranium compounds. By diluting the feed gas mixture, helium also hinders diffusion and remixing of the isotopic species. Separation gains for this process can be high and usually lie between those of gaseous diffusion centrifugation.

Individual separation units have been combined in stages to form a cascade on a pilot-plant scale in Germany. Although the efficiency of the plant, determined mainly by power consumption, is such that it is not economically competitive with gaseous diffusion, the technique has certain advantages. The proces requires fewer stages to achieve the same degree of enrichment as gaseous diffusion, so the equilibrium time is less (ca. 10 – 100-fold) than that of gaseous diffusion. This means that operating conditions may be changed much more rapidly and varied over a much broader range in this process. Simply by changing the temperature and pressure, the capacity and, therefore, the power consumption can be varied by 50 % or more.

4.2.5. Electromagnetic Separation

Electromagnetic separation is based on the same principles as mass spectroscopy (see Section 1.4) [72]. Electromagnetic separators, such as the ones used at the Oak Ridge National Laboratory (see Fig. 15), consist of four major components: ion source, accelerating electric field, magnetic field, and collecting apparatus. All are housed in an ultrahigh-vacuum enclosure. Ionic species are affected by the electric and magnetic fields of the spectrometer according to the equation

$$r_0 = \frac{c}{B}\left[\frac{2MV}{e}\right]^{1/2} \tag{62}$$

where r_0 is the radius of curvature of the ion path, c is the speed of light, B is the strength of the magnetic field, M is the mass of the ion, V is the accelerating potential of the electric field, and e is the charge on the ion. Under the same conditions of operation, a change in mass of the ion results in a change in the radius of the path the ion follows. The change can be determined from Equation (62) as

$$\Delta r_0 = \frac{r_0}{2} \cdot \frac{\Delta M}{M} \tag{63}$$

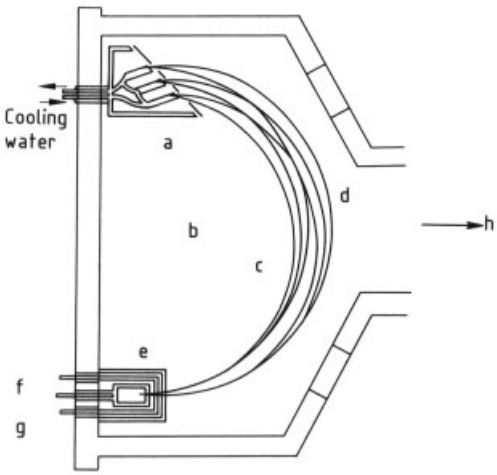

Figure 15. Electromagnetic separator at Harwell
a) Separated isotope collector; b) Vacuum chamber; c) Paths of light ions; d) Paths of heavy ions; e) Ion source; f) High positive voltage; g) High negative accelerating voltage; h) To vacuum pump

If the apparatus is a 180° focusing instrument like that shown in Figure 15, the difference x in the position at which the ions strike the collector is given by

$$x = r_0\,(\Delta M/M) \tag{64}$$

By constructing a large enough collection array and adjusting the electric and magnetic fields, adequate separation can be achieved even for ions with large masses and small mass differences. In theory, an electromagnetic separator should be able to effect the total separation of isotopes of any element; however, the process has practical limitations. The efficient separation power of a given instrument is defined as the largest $M/\Delta M$ ratio obtainable.

The ion source used for large-scale separators is usually a U-shaped tungsten or tantalum filament (the electron source), which is maintained at a negative potential. The source as a whole is maintained at a positive potential. A containing magnet is arranged so that electrons produced by the heated filament pass through the length of the ion source and impact gas molecules introduced into the ion source from a gas cylinder or heated crucible. Positive ions are produced in a complex process by electron collisions and extracted through the slit by a potential difference. The ions are initially accelerated to energies of ca. 60 keV and then decelerated to ca. 30 keV before being allowed to enter the magnetic field of the separator. This acceleration–deceleration sequence isolates the positive ion beam from the source, helps to focus the beam (primary focusing), and provides more uniformity in the energies of the ions. In practice, several different sources are needed if isotopes of a variety of elements are to be separated. The simplest sources are used for gases and very volatile materials; more sophisticated designs are required for higher temperature operation. The Oak Ridge separator has the capability for ion generation at 2000–2800 °C.

Source efficiency is the ratio of singly charged species of the appropriate nuclide leaving the source to the total amount of that species introduced into the source. Losses are primarily due to neutral species being carried out of the source with the ion beam. If the feed is a molecular compound, other ions may be generated by fragmentation to yield charged particles with masses very different from the target ions. For example, separation of uranium isotopes uses uranium tetrachloride (UCl_4) as the feed gas. In addition to U^+ (the desired species for collection), Cl^+, UCl_2^+, and multiply charged ions of these species are also produced. In this case, ca. 25 % of the input molecules are ionized and ca. 70 % of the ionized species are U^+.

Other problems arise because the output from the source is not a tightly focused beam but spreads out over an angle. As the density of the beam increases, the angle of spread becomes larger. Deflection of a diverging ion beam by a homogeneous magnetic field results in impact

on the collection array over a certain length of the array, rather than at a single point. If the beam divergence is defined by an angle $2a$, the maximum spread l of the impact position, called the *focal shift,* is given by:

$$l = r_0 a^2 \tag{65}$$

Large values of a, and hence large values of l, result in overlap of adjacent mass positions in the collection array. Secondary focusing is used to reduce the divergence. The ion beam is deflected into a curved path by a magnetic field in the *tank portion* of the spectrometer. *Shims,* electrically charged metallic strips inserted between the poles of the magnet, generate electric fields that affect the energy of the ions as they pass through the tank.

Two other problems that affect passage of the ion beam are interaction with residual species in the tank and space – charge interactions in the beam. Collisions of ions with molecules in the tank atmosphere can result in *small-angle scattering* and *charge exchange.* Scattering does not prevent ions from reaching the collector but does cause broadening of the beam. In charge exchange, particles lose their charge in a collision but may retain their kinetic energy; these neutral species are not affected by the magnetic field. Efficient evacuation of the tank can reduce these interactions to a large extent, but this also increases the magnetic and electrostatic interactions. This arises from the fact that two equally charged particles traveling along the same path generate magnetic fields, which attract one particle to the other, and electric fields, which have the opposite effect. Magnetic attraction is negligible, but electrostatic repulsion becomes significant as the ion current intensity increases. Repulsive forces tend to broaden the beam and, for production-scale separators, can result in total overlap of target areas by beams of ions with similar masses; isotopic separation is impossible in this case. One method for moderating the electrostatic interaction is to allow collisions with other species in the tank, which leads to formation of negative ions that can separate the positive charges. Such a technique does lead to losses by charge exchange, but an acceptable compromise is to maintain the tank, at ca. 1.3 mPa. Other methods of space – charge neutralization include placing a negatively charged grid or electron-producing device in the path of the ion beam.

The *collection array* is a series of individual pockets that gather and hold the isotopes. These are made of electrically conducting materials and are insulated from one another so that the ion current from each isotope can be monitored. A cover plate with a small slit may be placed in front of each pocket to monitor the focus of the beams. The pockets are deep, narrow boxes, arranged at an angle to the path of the beam so that ions impinge on one side of the pocket and are deflected downward into the box. The ions cannot be decelerated efficiently before they reach the collector, so all the energy of the species must be dissipated within the pocket. Ions that reach the collector may be lost through evaporation from the pocket and through sputtering caused by subsequent ions. Evaporation results from heating of the pocket due to the subsequent impact of ions. The pockets are cooled to prevent distillation of isotopes from one pocket to another and to prevent the pocket from melting. *Sputtering* is the process by which some of the material of the pocket is removed by impact of the ions. The two materials may become intermixed to the extent that another separation step is required to recover the product. Some ions may be sputtered back onto the cover plate or even the cover plate of an adjacent pocket. At large ion energies and throughputs, the life of a pocket may be limited to a few days. With isotopes that are gases under normal conditions, the ions are collected by bombardment of thin metal foils which trap the ions in their surface layers.

The separation efficiency of electromagnetic separators is usually in the 2 – 10 % range. Nevertheless, the technique is valuable because of its adaptability to virtually any element and the fact that small amounts of highly pure isotopes can be obtained. Most elements have been separated by the electromagnetic process, and it is currently used for species that cannot be produced by other methods. The elements are generally grouped by similarities in their properties and the operating conditions needed for separation; these roughly follow the grouping of the periodic table.

Certain gases (Ne, Ar, Kr, Xe, N, O, and CO_2, which is used to separate ^{12}C and ^{13}C) present few problems in source and tank behavior. Collection is limited by the amount of material that can be retained in the foil targets. Alkali and alkaline-earth metals (Li, K, Rb, Ca,

Sr, Ba, and Mg) are processed as the metal or metal halide. Most of these present corrosion problems for the source but are ionized easily, give high current densities, and may be separated in relatively large quantities. Metals (Cu, Ag, Be, Zn, Cd, In, Tl, Sb, Sn, Pb, Cr, Fe, Ni, and Ga) require a medium source temperature and usually employ a halide as the feed material. Problems can arise from plating of the metals on surfaces, which leads to shorting of electrical components. The more volatile elements must be collected in cooled pockets to avoid evaporation. Elements with very volatile chlorides (B, Si, Ti, Ge, Zr, Hf, Ta, Mo, and W) require a low source temperature and sometimes cooling of the source. The metals are refractory and present no problems in collection. Mercury presents special problems because a significant amount of ionized metal may be lost from the source. Gold pockets are often used as collectors to retain mercury by amalgamation. Some elements (e.g., S, Se, Te, Cl, and Br) provide certain problems in collection and require good cooling for efficient retention. The lanthanides (La, Ce, Nd, Sm, Eu, Gd, Dy, Er, Yb, and Lu) are introduced into the source as their chlorides. These materials require a high source temperature (700–900 °C) and may also interact severely with the filament material. Many of these elements have numerous isotopes with closely spaced $M/\delta M$ values, and especially narrow pockets must be used to ensure good separation. Platinum-group metals (Pt, Ru, Pd, and Ir) have very high melting and boiling points and require sources capable of withstanding extremely high temperature. The actinides (Th, U, and Pu) are also used as their chlorides. Sputtering is a problem in collecting these heavy elements. In addition to the separation of naturally occurring isotopes, this technique is well-suited to the separation of artificially produced radioactive species, especially those with short half-lifes.

Large electromagnetic separation production facilities are currently in operation at the Oak Ridge National Laboratory in the United States, at Harwell in Great Britain, and in Denmark, Germany, France, the Netherlands, Japan, South Africa, Russia, and Sweden.

4.2.6. Exchange Reactions

Chemical Exchange. Isotope separation by chemical exchange exploits the equilibrium isotope effect (see Section 2.2.1). The rates of exchange of isotopes between compounds often indicate a preference for association of a particular isotope with one compound rather than another, and this phenomenon may be exploited to enrich that isotope. As shown in Equation (4), the equilibrium constant for an exchange reaction is the ratio of the concentration of the reaction products to the concentration of the reactants. This concentration ratio also defines the elementary separation factor. The factor is a function of the masses of the isotopes involved and is therefore greater for elements with isotopes having the greatest relative mass differences. The isotopes of hydrogen are thus most easily separated by chemical exchange processes. Other isotopes enriched by this technique include those of boron, carbon, nitrogen, oxygen, and sulfur. Many different processes exist for enrichment by exchange reactions which depend on material properties and equilibrium constants. Materials may be brought into contact in the same phase (gas – gas and liquid – liquid processes), in different phases (gas – liquid processes), in distillation processes, at constant temperature (monothermal processes), and by reaction at different temperatures (dual-temperature processes).

Gas – gas reaction processes are used to separate a number of isotopic systems, including the enrichment of ^{13}C by exchange reactions of carbon monoxide and carbon dioxide, and of ^{15}N by exchange reactions involving nitrogen oxides. The most common separation employing this process is the H_2O – HD exchange reaction:

$$H_2O + HD \rightleftharpoons HDO + H_2 \tag{66}$$

The elementary separation factor (equilibrium constant) for this reaction ranges from >4 at 0 °C to ca. 1.2 at 750 °C. For large-scale production, a series of columns, each one a stage, is arranged to form a cascade. The columns, also known as *exchange towers,* are similar to those used for fractional distillation (see Section 4.2.1), with an evaporator at the bottom, a condenser at the top, and bubble-cap plates throughout the tower, connected by downcomer

tubes. In addition to these features are the devices that allow the exchange reaction to take place, catalyst beds between each pair of bubble-cap plates and an electrolytic cell to generate hydrogen.

The electrolytic cell produces hydrogen which is humidified by the evaporator. The gas – vapor mixture passes upward through the column and encounters the first catalyst bed (platinum or nickel alloys) where the exchange reaction enriches the deuterium concentration of the water vapor according to Equation (66). The mixture then passes through the first bubble-cap plate where some of the HDO in the vapor is replaced with H_2O. The new gas – vapor mixture is then directed to the next catalyst bed where the exchange reaction is repeated, passes through the next bubble-cap plate, and proceeds up the tower. Thus, water traveling downward from plate to plate is enriched in deuterium and hydrogen traveling upward is depleted. The water passes on to the refluxing unit of the next stage, whereas the hydrogen is added to the gas flow of the previous stage. A plant with this configuration was constructed in Canada in 1943 and operated until 1956, producing ca. 500 kg of 99.8 % D_2O per month [62, p. 9].

This single-phase separation process is rare. Most of the satisfactory exchange reaction enrichment processes involve two phases (gas and liquid) or two immiscible liquids. For such a system to be practical, several criteria must be met. The two compounds involved in the exchange reaction should each be present in only one of the phases, e.g., in the enrichment of ^{15}N by exchange between nitric acid (liquid) and nitric oxide (gas). The exchange should be rapid or capable of being catalyzed so that short exchange towers can achieve a high degree of enrichment. The equilibrium constant for the exchange should be as large as possible to reduce the number of stages required in the cascade. This implies that the isotope of interest should be weakly bonded in one of the compounds and strongly bonded in the other, a situation which is at odds with rapid exchange (an isotope weakly bonded in both compounds would have the fastest exchange rate). Finally, the system should be able to support total reflux, in the sense that the liquid phase is converted to the gas phase by chemical reaction, rather than boiling, in each stage to achieve maximum interaction of the two phases.

Separation of ^{15}N by means of the reaction

$$^{15}NO + H^{14}NO_3 \rightleftharpoons {}^{14}NO + H^{15}NO_3 \tag{67}$$

is a good example of a gas – liquid exchange process [73]. Separation can be accomplished in a packed column. The exchange reaction is rapid and is catalyzed by nitrous acid. An enrichment of ^{15}N from 0.36 to 99.6 % has been achieved by using two columns of 3.0- and 1.5-cm diameter. At the bottom of each column, nitric acid is converted to nitric oxide by reduction with sulfur dioxide. The gas flows upward through the packed bed and exchanges the heavier isotope, as shown in Equation (67). The gas reacts with oxygen and is absorbed into water at the top of the column to regenerate nitric acid, which flows downward in the column.

A liquid – liquid exchange process has been developed for the enrichment of ^{6}Li [74]. This process involves the exchange of lithium isotopes between a liquid lithium amalgam and a solution of lithium chloride in absolute ethanol. Amalgam is fed dropwise into the top of the 18-m column containing the solution and participates in the reaction

$$^{7}LiHg + {}^{6}LiCl \rightleftharpoons {}^{6}LiHg + {}^{7}LiCl \tag{68}$$

At the bottom of the column, the lithium amalgam is converted to lithium chloride by hydrogen chloride in alcohol solution. To avoid violent reactions with metallic lithium, a neutral solution must be maintained and water must be rigorously excluded. The actual separation factor was not determined in this system because each droplet of the amalgam did not follow the same path through the solution and it was not known whether equilibrium was achieved. The process operated in a discontinuous mode. Each batch required ca. 24 h for processing and was enriched in ^{6}Li by a factor of slightly more than 2.

As stated, one criterion for an acceptable exchange reaction separation process is the availability of a reflux method that allows the two phases or liquids to be readily interconverted, ideally in a quantitative manner. Not all isotopic systems with large equilibrium constants are amenable to the electrolytic or chemical oxidation and reduction reactions used with the sample systems discussed here. Another method, *exchange distillation,* may be employed in some

cases. In this technique, however, enrichment does not rely on the difference in vapor pressures of the isotopic species. Instead, the liquid being distilled dissociates in the vapor phase to give a material with different composition. The separation of boron isotopes uses such a process [63]. The liquid phase is the dimethyl ether adduct of boron trifluoride, which dissociates to the uncomplexed compound upon distillation. The exchange reaction is thus given by the equations

$$BF_3 \cdot O(CH_3)_{2(l)} \rightleftharpoons BF_{3(g)} + O(CH_3)_{2(g)} \tag{69}$$

$$^{10}BF_{3(g)} + {}^{11}BF_3 \cdot O(CH_3)_{2(l)} \rightleftharpoons {}^{11}BF_{3(g)} + {}^{10}BF_3 \cdot O(CH_3)_{2(l)} \tag{70}$$

The process operates in a packed fractionation column. The etherate is boiled in an evaporator at the bottom of the column, and the dissociated vapor passes upward through the column. The adduct is reformed in a condenser at the top of the column. The exchange reaction takes place between the liquid etherate and boron trifluoride gas on the surface of the packing material. Separation factors for columns used in this exchange are ca. 1.02. The process was first developed in 1944 in connection with the Manhattan Project to supply a large quantity of boron (50 kg) enriched to 90 % in ^{10}B. The process was classified by the U.S. government until 1957, and in the meantime, was developed in Sweden and Great Britain in the 1950s.

Another method is available to avoid problems associated with refluxing a two-phase system. The *dual-temperature process* employs two columns operating at different temperatures and, therefore, having different equilibrium constants for the exchange reaction [75]. In one column, one phase of the system is enriched at the expense of the second at a given temperature to provide the desired product. In the other column, the depleted phase from the first column is enriched by exchange with the first phase at a different temperature. The second phase from this column is then used as the feed for the first column. Figure 16 shows a dual-temperature arrangement for the production of heavy water by exchange with hydrogen sulfide gas. The two columns together constitute a stage, and several stages are arranged to form a cascade. The reaction for this process is given by

$$H_2O_{(l)} + HDS_{(g)} \rightleftharpoons HDO_{(l)} + H_2S_{(g)} \tag{71}$$

The choice of operating conditions for the process must take several factors into account. The temperatures of the two columns are selected so as to maximize the difference between the equilibrium constants. However, the upper temperature that can be used is limited by the increasing corrosivity of H_2S at high temperature. High-pressure operation is desirable to minimize the column cross-section, but formation of a solid hydrate of H_2S at low temperature and liquification at high temperature restrict the maximum pressure. A successful compromise is to operate the columns at 30 and 130 °C and 2 MPa pressure.

The equilibrium constant (elementary separation factor) for the reaction at 30 °C is 2.3, whereas at 130 °C it is 1.8. The upper (cold) column is operated at the lower temperature and fed with a flow of ordinary water (for the first stage in the cascade; successive stages are fed with product from the previous stage). Deuterium is exchanged from hydrogen sulfide to water as the gas bubbles through successive plates of the column. At the bottom of this column, some water is withdrawn as product and is passed to the next stage in the cascade. The remainder of the water is heated in a heat exchanger, mixed with additional water from the stage immediately downstream in the cascade, and passed to the hot column below. The depleted hydrogen sulfide, removed at the top of the cold column, is also heated and then introduced at the bottom of the hot column. As it flows upward through the plates of this column, hydrogen sulfide becomes enriched in deuterium at the expense of water. The enriched gas is then cooled and fed into the upper column to repeat the enrichment of water there. Water from the hot column, depleted in deuterium, goes to a heat exchanger where it is used to heat water entering this column and is then directed to the next column upstream in the cascade to make up part of the feedwater for the hot column of that stage. Thus, the refluxing requirement of the process is met by the two different equilibria of the hot and cold columns.

The enrichment factor α for this particular arrangement of columns and flows in the cascade depends on the equilibrium constants and the cut θ, the portion of water leaving the upper column that is directed to the next stage of the cascade:

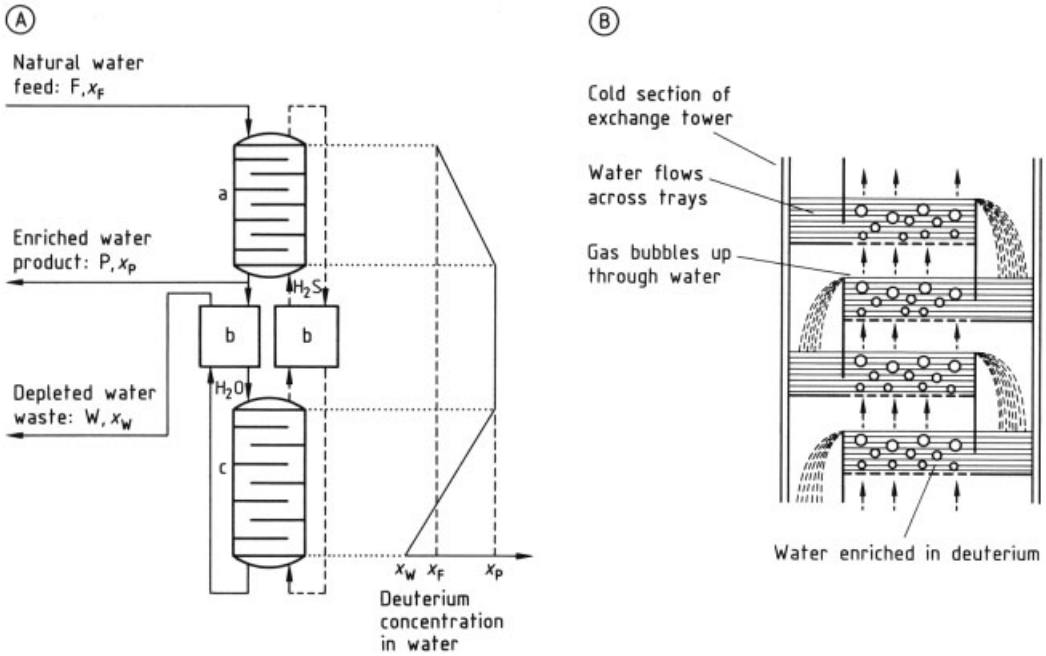

Figure 16. A) Dual-temperature exchange column; B) Cross section through cold column
a) Cold column, temperature 30 °C, $K = 2.3$; b) Heat exchanger; c) Hot column, temperature, 130 °C, $K = 1.8$

$$\alpha = 1 + \frac{K_1 - K_2}{K_1} \cdot \frac{1-\theta}{\theta} \tag{72}$$

where K_1 and K_2 are the equilibrium constants for the lower and higher temperatures, respectively. The enrichment factor can be large at small values of the cut. Dual-temperature chemical exchange on a production scale has been implemented only for the enrichment of deuterium. Other isotopic systems do not show the large equilibrium isotope effect that is found for protium and deuterium and, consequently, do not have large enough differences in equilibrium constants as a function of temperature to make the technique viable for isotope separation. In addition to the water – hydrogen sulfur process described, other systems that allow deuterium enrichment by the dual-temperature process include water – ammonia, water – hydrogen halides, water – hydrogen, and ammonia – hydrogen.

Ion-Exchange Processes. A different type of chemical exchange technique is employed in ion-exchange processes, sometimes referred to as *chromatographic separation* [76]. Instead of two phases moving countercurrent to one another, ion-exchange methods use columns in which one phase is stationary and a second phase moves through the column. The countercurrent aspect of separation is achieved even with one stationary phase because the species to be separated move through the column and interact with both phases continually.

An example of ion-exchange separation is the enrichment of ^{15}N by interaction of a dilute aqueous ammonia solution with a cationic exchange resin that incorporates sulfate groups on the surface of the resin particles. The resin is first washed with acid to saturate all of the sulfate groups with protons. A quantity of ammonia solution is added to the column and reacts with the resin:

$$\text{Resin-H} + NH_4OH \rightleftharpoons \text{Resin-}NH_4 + H_2O \tag{73}$$

Thus, a portion of the resin is in equilibrium with the solution. The length of solution in contact with the resin, called a *band,* is determined by the size of the charge applied; the front end of the band is sharply defined because of the large equilibrium constant for the reaction between

solution and resin (in this case ca. 10^9). To create flow through the column, a solution of sodium hydroxide is added to the column above the band of ammonia solution. Another reversible reaction between the resin and this solution takes place according to the equation

$$\text{Resin-NH}_4 + \text{NaOH} \rightleftharpoons \text{Resin-Na} + \text{NH}_4\text{OH} \tag{74}$$

The large equilibrium constant for this reaction (10^5) sharply defines the rear of the band. More hydroxide is added to move the band containing the ammonia solution down the column. As the front of the band encounters new resin, the exchange reaction continues, whereas at the rear of the band, ammonium ions are continually returned to the band through displacement by sodium hydroxide, keeping the band at constant size as it moves along the column. This method of separation is known as *displacement* because the target ions absorbed onto the resin are displaced by other ions.

The separation of isotopes occurs as a result of equilibrium isotope effects in Equations (73) and (74). The exchange of ammonium ions with protons on the column results in a small enrichment of ^{15}N in the bound species, whereas the release of ions by sodium hydroxide gives a much larger separation of the isotopes. The overall result of these reactions is that the ammonium ions at the rear of the band are enriched in the heavier isotope and the front of the band is depleted in that species. The elementary separation factor for the process is given by the product of the ratios of the concentrations of the two isotopic species in their various forms:

$$q_0 = \frac{[^{14}\text{NH}_4\text{OH}]\,[^{15}\text{NH}_4\text{-resin}]}{[^{15}\text{NH}_4\text{OH}]\,[^{14}\text{NH}_4\text{-resin}]} \tag{75}$$

Other methods of chromatographic separation are the *elution technique* and the *breakthrough technique.* In the former, the column is washed with a solution containing a compound of the isotopes to be separated, and some of the compound is allowed to be absorbed on the resin. Another solvent is then passed through the column, and the compound is desorbed into the new solvent. If the properties of the isotopically substituted species are such that their desorption equilibria differ, the front of the band containing the desorbed species is enriched in one isotope and the rear of the band in the other. In the breakthrough technique, a difference in the absorption equilibria of the isotopically substituted species is used in the same way. A solution containing two species is passed through the column, and the material that is absorbed least strongly is concentrated at the front of the column.

An advantage of ion-exchange methods for isotope enrichment is that they may be applied to virtually any isotopic system with the right combination of resin and solvents. Exchange processes have been used effectively to separate isotopes of lithium, boron, sodium, sulfur, calcium, and uranium. Gas chromatography has been used to enrich isotopes of hydrogen and neon. Gaseous isotopes are separated by their different rates of absorption and desorption on a suitable material, (e.g., palladium or charcoal). An interesting effect is achieved in the separation of hydrogen isotopes over activated alumina. Not only are H_2, HD, and D_2 collected individually, but the *ortho* and *para* isomers of both H_2 and D_2 are also separated. Another advantage of ion exchange is the relatively high density of the stationary phase (the resin particles can be manufactured with very high ratios of surface area to volume). Thus, the height equivalent of a theoretical plate may be very small (ca. 1 mm or less) compared to other processes such as distillation in a packed column (HETP = ca. 0.5 m); therefore, columns can easily be prepared with tens or hundreds of plates.

Disadvantages of the process include the fact that the process is discontinuous because one phase is stationary and material must be processed batchwise. Furthermore, only a limited number of plates are in contact with the band of material as it passes through the column, and as little as 6 % of the effective area of the exchange resin may be involved in the enrichment at any one time. These problems can be reduced somewhat by introducing a second charge before the first has left the column, a technique that can be repeated several times before the resin must be totally regenerated. A more serious disadvantage arises from the long time needed to reach equilibrium between solid and liquid phases. A compromise is necessary between the rapid collection of isotopes and the lower enrichment resulting from operating further from equilibrium conditions. If the mobile phase is a gas or if the exchange reactions of species in a mobile liquid phase have a relatively high equilibrium

constant, these difficulties are reduced. The ion-exchange process is particularly suited for the production of laboratory quantities of light isotopes.

4.2.7. Photochemical Separation

The difference in mass between isotopes results in differences in their optical spectra, both in the elemental form and in compounds containing them (see Section 2.1.1). Photochemical separation of isotopes takes advantage of the different energies of various transitions to ionize or otherwise excite one isotope selectively. Thus, the process holds great promise of achieving a very large elementary separation factor. Several criteria must be satisfied for this potential to be realized: (1) the isotopic species must have well-defined absorbance differences as a consequence of the isotopic effect; (2) the exciting energy must be limited to a range that affects only the species of interest; (3) the modes of energy exchange from the species of interest to other isotopes must be slow compared with the process used to remove the excited species from the mixture of isotopes; (4) the excitation energy must be intense enough to give a good yield; and (5) extraction of the excited species from the isotopic mixture should be rapid, selective, and quantitative.

The first successful photochemical isotope separation was achieved in 1958 by Gunning, who enriched ^{202}Hg from its natural abundance of ca. 30 % to more than 90 % [77]. The process involved excitation of the specific isotope by irradiation in the area of the 253.7-nm absorption band of mercury. The excitation energy was selectively filtered by passing it through mercury vapor that had been depleted in ^{202}Hg by another means. In the presence of oxygen, the excited species reacts to give $^{202}HgO_2$, with an oxidation rate orders of magnitude faster than that of the unexcited mercury isotopes. The oxide was stabilized by addition of butadiene and removed from the reaction mixture. Similar methods for separating other mercury isotopes are of interest because of their potential for use in fluorescent lighting (see Section 3.4).

Photochemical reactions have also been explored for separation of uranium isotopes. Ionization of metallic uranium in the vapor phase and dissociation of gaseous UCl_6 to form UCl_5^+ have been accomplished with laser excitation. The charged species are removed from the isotopic mixture by collection on a negatively charged plate. Laser sources appear to be ideal because they can be tuned with precision to narrow energy bands. Scaleup of this process is currently hampered by difficulties in handling large quantities of molten uranium and by secondary reactions that reduce the selectivity of the excitation process; nevertheless, it promises to allow uranium enrichment at a substantially lower cost than gaseous diffusion.

4.2.8. Electrolysis

Separation of isotopes by electrolysis is based on the difference in rates of production of an element by oxidation or reduction of an ionic species [78]. Many elements exhibit isotopic effects upon electrolysis, but hydrogen has by far the largest rate difference between isotopes. Separation effects have been reported for isotopes of nitrogen, oxygen, and chlorine (among others), but all have elementary separation factors close to 1.0, whereas the factor for hydrogen enrichment by electrolysis of water (containing an electrolyte such as potassium hydroxide or potassium carbonate) ranges from ca. 3 to >17, depending on temperature and other conditions.

No definitive theory can explain the difference in rates for different isotopes. With hydrogen, the need for a greater overvoltage (the voltage above the reduction potential actually necessary to generate the element at the cathode) for deuterium appears to make its rate of production slower than that of protium. Various experimental studies have cited different steps as rate determining, including discharge of the ion at the electrode, formation of the molecular species from the atoms, and exchange of isotopes between water molecules adsorbed onto the cathode surface and those in solution. Despite this ambiguity, the elementary separation factor can be determined experimentally by simply comparing the relative abundance of deuterium in the water with its relative abundance in the gas produced at the cathode. Several factors influence the value of the separation factor. The temperature of the electrode during electrolysis is a major influence; the separation factor increases

as temperature decreases. The ion current is important because it affects electrode temperature; high current densities may heat the electrode faster than the surrounding solution can dissipate energy. The cathode material affects the separation of hydrogen and the overvoltage needed to cause reduction in ways that are not completely understood. Some of the differences in separation factor can be attributed to the catalysis of exchange reactions by certain materials, which alters the isotopic abundance of the species reacting or formed at the electrode. In general, materials requiring a higher overvoltage give lower separation factors, but this is only a qualitative observation and many exceptions have been reported.

The use of electrolysis for the production of highly enriched deuterium from water at natural deuterium concentration is not economically viable. As an adjunct to hydrogen production plants for other purposes (e.g. fertilizer synthesis) the added expense of separating deuterium-containing species may be warranted. This type of arrangement yields only small quantities of heavy water; a 300 t/d ammonia plant would produce only 24 t of D_2O per year. Another practical use of electrolysis is the final enrichment of deuterium that has been highly enriched by another process. Production of heavy water for the Manhattan Project was accomplished in such a manner. The distillation plant at Morgantown (see Section 4.2.1) used electrolysis to enrich water from a deuterium content of 88 – 92 % to ca. 99.9 %. The process was reversed in Canadian and Indian facilities, with electrolysis producing enriched feed material for distillation. These separation operations are associated with ammonia production plants.

Most often, electrolysis is operated as a batch process. A charge of water is placed in a cell and electrolyzed until the material remaining in the cell has the desired deuterium abundance. This allows greater flexibility in accommodating feed materials of different isotopic abundances. Such facilities are generally easier to construct and operate but are more labor intensive and increase the possibility of material losses.

Continuous processes involve a cascade with a number of stages, depending on the desired final deuterium concentration, and may be of the recycling or nonrecycling type. In a nonrecycling system, hydrogen and oxygen (produced at the anode) are drawn off and used for other purposes. A recycling system burns the gases to form water, which is then added to an upstream stage. Such cascades are usually designed in a square configuration (see Section 4.1.2). The flow from stage to stage is in the form of steam that accompanies production of the gases. Instead of burning the hydrogen produced by electrolysis, the yield of the continuous process can be increased by incorporating a chemical exchange system in the recycling portion of the cascade. This recycling operation takes the hydrogen and steam, which accompany gas production (steam that has a higher deuterium concentration than the gas), from a separation unit and passes them through a scrubber column. A countercurrent flow is established with water coming from an earlier stage in the cascade. In the scrubber, water is enriched in deuterium at the expense of steam and passes to the next electrolytic cell in the cascade. Steam, depleted in deuterium, is condensed and fed into the exchange reaction chamber of the same stage along with hydrogen. Here, the equilibrium isotope effect enriches the deuterium concentration of water at the expense of hydrogen. The water is then returned to the electrolytic cell of the stage from which it came, whereas the hydrogen is used to make up part of the gas stream entering the scrubber column of the stage immediately upstream. Some heavy-water production facilities use a combination of recycling processes. An electrolytic plant in Norway employs scrubbers and exchange reaction chambers in the first section to help achieve a 2 % enrichment of deuterium. The remainder of the facility, which increases the concentration to ca. 15 %, employs burning of hydrogen for recycling.

Electrolysis also concentrates tritium in the liquid phase. The separation factor for enrichment of tritium is higher than that for deuterium and tritium may be present in measurable quantities in highly enriched heavy water.

4.2.9. Electromigration

Another phenomenon that occurs during electrolysis–electromigration–can be used to separate isotopes. Instead of relying on the difference in the rate of production of isotopes in their elemental form, electromigration depends

on the different mobilities of the ionic species as they move toward the electrodes of the cell. The separation effect is, theoretically, determined by the ratio of the migration rates of the light and heavy isotopes of an element, but it may also be affected by thermal gradients and back diffusion of ions in the solution or melt being electrolyzed. Elementary separation factors are similar to those for electrolysis, close to 1.0 for most elements. The process was successfully applied to the separation of ^{39}K and ^{41}K in the late 1940s, and to ^{6}Li and ^{7}Li thereafter. The apparatus for electromigration consists of a tube that holds a molten salt, a molten salt eutectic, or a solution of the salt containing the isotopes to be separated. Electrodes at either end of the tube are given a potential difference great enough to cause oxidation and reduction of ionic species in the tube. Migration toward the appropriate electrode enriches the lighter isotopic species in its vicinity and depletes it near the opposite electrode. Because the production of elemental species is not necessary, some method of regenerating the ionic species is employed at each electrode.

In the separation of potassium isotopes an aqueous solution of potassium chloride is electrolyzed [79]. The lighter ^{39}K is concentrated at the cathode; any metallic potassium produced by reduction is allowed to react with hydrochloric acid to maintain a constant total ionic concentration. Oxidation of chloride ions at the anode is reversed by addition of potassium hydroxide. Because potassium hydroxide also contains potassium isotopes in their natural abundance, collection of enriched ^{41}K is not possible. A porous filter isolates the electrodes and the solutions adjacent to them from one another and prevents contamination of the enriched ^{39}K near the cathode. In the separation of lithium isotopes, molten lithium chloride is used instead of an aqueous solution. Chlorine gas is added to reduce metallic lithium formed at the cathode. The desired product ^{7}Li is concentrated near the anode because ^{6}Li migrates toward the cathode at a higher rate. A buffer solution, a compound whose cation is not electrolyzed at the potential used (e.g., $PbCl_2$), is employed to isolate the lithium chloride melt from the anode. An abundance of 99.9 % for ^{7}Li has been achieved by this process. The technique is not economical for large-scale separation of isotopes but can provide small quantities of enriched isotopes for laboratory work.

4.2.10. Other Separation Processes

Several new technologies for enrichment of isotopes are in various stages of development [80], and not enough information is available to describe them in detail. *Vacuum arc centrifugation* (VAC) uses magnetic fields to rotate a beam of ions containing the isotopes to be separated. The separation efficiency is related to the difference in the masses of the isotopes and the rotation frequency of the plasma. VAC is said to be applicable to a wide range of isotopes and could produce them in kilogram quantities. The *solitron* also separates isotopes in an ion beam, in this case by imposing a traveling electric field wave on the beam. This device is expected to be essentially portable and produce a variety of isotopes for research purposes. The *SILEX process* is a proprietary laser separation process under development in Australia that appears to be based on excitation of molecular species containing the isotopes of interest.

5. Availibility

Stable isotopes are available from a large number of companies. A significant part of the market for stable isotopes includes isotopically labeled organic compounds, typically used in diagnostics and research. With the approval of medical diagnostic tests using labeled compounds, the production of isotopically labeled compounds will likely see a large increase. The most common isotopes for labeling are deuterium and ^{13}C, with ^{15}N and ^{18}O also readily available. Compounds such as amino acids, carbohydrates, nucleosides, and pharmaceuticals are marketed with partial or complete substitution. Noble gases and metal isotopes are other important categories of commercially available isotopically enriched materials. Most isotope suppliers have custom synthesis capabilities for manufacture of uniquely substituted compounds. Radioisotopes, produced from stable isotopes by neutron irradiation, are also available from commercial and government sources.

Canada and Russia supply the overwhelming majority of radioactive isotopes at this time.

The U.S. Department of Energy national laboratories produce and sell both stable enriched isotopes and radioisotopes to fill needs where private sector capability is nonexistent or is insufficient to meet U.S. demand. Brookhaven, Los Alamos and Oak Ridge National Laboratories have facilities for target irradiation and will lease the use of hot cells for separation of radioisotopes produced at the laboratories. DOE's New Brunswick Laboratory produces standards and references for thorium, uranium, and plutonium.

A partial list of companies offering enriched stable isotopes and/or radioisotopes for sale includes:

Cambridge Isotope Laboratories
www.isotope.com
Isotec, Inc.
www.isotec.com
Toronto Research Chemicals, Inc.
www.trc-canada.com
Medical Isotopes, Inc.
www.medicalisotopes.com
Trace Sciences International
www.isotopetrace.com
Omicron Biomedical Inc
www.omicronbio.com
Isonics, Corp.
www.isonics.com
Isoflex USA
www.isoflex.com
ChemSyn Laboratories
www.chemsyn.com
Canrus Enterprises Inc.
www.magi.com/~ppeinc/index.shtml
U.S. Department of Energy
www.ornl.gov/isotopes/catalog.htm
Michigan Memorial Phoenix Project Isotope
www.umich.edu/~mmpp/services/isotopes/isotope.html

Associated with the rise in the use of isotopes for research and medicine and their potential use in industry is an increase in the number of facilities for the analysis of isotopic composition. Many universities have isotope laboratories associated with research programs that employ stable or radioactive isotopes and will perform such analysis for other customers as well. Many of the suppliers of isotopes also market the capability to perform analyses.

6. References

1. C. R. Hammond in R. C. Weast (ed.): *Handbook of Chemistry and Physics,* 65th ed., CRC Press, Boca Raton, Fla., 1984, p. B5.
2. D. C. Hoffman, F. O. Lawrence, J. L. Mewherter, F. M. Rourke, *Nature (London)* **234** (1971) 132.
3. P. H. Abelson (ed.): *Science* **218** (1982) 274.
4. G. R. Choppin, J. Rydberg: *Nuclear Chemistry: Theory and Applications,* Pergamon Press, Oxford 1980, p. 222.
5. B. G. Harvey: *Introduction to Nuclear Physics and Chemistry,* 2nd ed., Prentice-Hall, Englewood Cliffs, N.J., 1969, p. 117.
6. A. G. W. Cameron, *Space Sci. Rev.* **15** (1970) 121.
7. R. Alpher, H. Bethe, G. Gamow, *Phys. Rev.* **73** (1948) 803.
8. E. M. Burbidge, G. R. Burbidge, W. A. Fowler, F. Hoyle, *Rev. Mod. Phys.* **29** (1957) 547.
9. L. H. Aller: *The Abundance of the Elements,* Wiley-Interscience, New York 1961, p. 20.
10. N. N. Greenwood, A. Earnshaw: *Chemistry of the Elements,* Pergamon Press, Oxford 1984, p. 10.
11. S. Glasstone: *Sourcebook on Atomic Energy,* 3d ed., Van Nostrand Reinholt, New York 1967, p. 160.
12. R. D. Russell, R. M. Farquhar: *Lead Isotopes in Geology,* Wiley-Interscience, New York 1960, p. 13.
13. F. W. Aston: *Mass Spectra and Isotopes,* 2nd ed., Edward Arnold & Co., London 1942, p. 35.
14. H. H. Willard, L. L. Merritt, J. A. Dean: *Instrumental Methods of Analysis,* 5th ed., D. Van Nostrand, New York 1974, p. 455.
15. S. Villani: *Isotope Separation,* American Nuclear Society, Hinsdale, Ill. 1976, p. 23.
16. N. N. Greenwood in I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* 2nd ed., **Vol. 1,** Wiley-Interscience, New York 1978, p. 453.
17. J. Hoefs: *Stable Isotope Geochemistry,* 3rd ed., Springer Verlag, New York 1987, p. 26.
18. G. T. Fraser, R. D. Suenram, F. J. Lovas, *J. Chem. Phys.* **86** (1987) no. 6, 3107.
19. G. M. Murphy in H. Craig, S. L. Miller, G. J. Wasserburg (eds): *Isotopic and Cosmic Chemistry,* North-Holland Publ., Amsterdam 1964, p. 1.

20. J. Bardeen, L. N. Cooper, J. R. Schrieffer, *Phys. Rev.* **106** (1957) 162.
21. H-C. zur Loye et al., *Science* **238** (1987) 1158.
22. S. M. Blinder: *Advanced Physical Chemistry,* Macmillan Co., New York 1969, p. 422.
23. R. E. Weston, Jr., H. A. Schwarz: *Chemical Kinetics,* Prentice-Hall, Englewood Cliffs, N.J., 1972, p. 109.
24. P. Wyden: *Day One: Before Hiroshima and After,* Simon & Schuster, New York 1984.
25. D. B. Hoisington: *Nucleonics Fundamentals,* McGraw-Hill, New York 1959, p. 318.
26. S. E. Liverhant: *Elementary Introduction to Nuclear Reactor Physics,* J. Wiley & Sons, New York 1960, p. 134.
27. J. B. Lambert et al.: *Organic Structural Analysis,* Macmillan, New York 1976, p. 3.
28. L. A. Paquette, R. J. Ternansky, D. W. Balogh, G. Kentgen, *J. Am. Chem. Soc.* **105** (1983) 5446.
29. J. G. Dorsey, G. F. Dorsey, A. C. Rutenberg, L. A. Green, *Anal. Chem.* **49** (1977) 1144.
30. G. M. Whitesides, H. L. Mitchell, *J. Am. Chem. Soc.* **91** (1969) 4115.
31. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 3rd ed., Wiley-Interscience, New York 1972, p. 856.
32. R. T. Morrison, R. N. Boyd: *Organic Chemistry,* 3rd ed., Allyn & Bacon, Boston 1973, p. 679.
33. C. W. Gillies, et al., *J. Am. Chem. Soc.* **113** (1991) (Mar. 27), no. 1, 2412 – 2421.
34. E. Buncel, W. H. Saunders, Jr., *Isotopes in Organic Chemistry, Vol. 8, Heavy Atom Isotope Effects,* Elsevier Science Publishers B.V., Amsterdam 1992, p. 133.
35. G. Friedlander, J. Weneser, *Science* **235** (1987) 760.
36. T. A. Kirsten, A. Faessler (eds.): *Progress in Particle and Nuclear Physics, vol. 40, Neutrinos in Astro, Particle and Nuclear Physics,* Elsevier Science BV, Amsterdam 1998, p. 85.
37. T. Kotani, H. Ejiri, E. Takasugi (eds.): *Nuclear Beta Decays and Neutrino,* World Scientific Publishing Co., Singapore 1986.
38. L. D. Hulett, Jr., et al. (eds.): *Advanced Materials Research, Vol. 3, Application of Positron Spectroscopy to Materials Science,* Scitec Publications Ltd, Switzerland 1994.
39. P. W. Rundel, J. R. Ehrelinger, K. A. Nagy (eds.): *Stable Isotopes in Ecological Research,* Springer-Verlag, New York 1989.
40. D. S. Schimel, *Theory and Application of Tracers,* Academic Press, Inc., San Diego 1993.
41. K. Lajtha, R. H. Michener, *Stable Isotopes in Ecology and Environmental Science,* Blackwell Scientific Publications, London 1994.
42. G. Shearer, D. H. Kohl in Reference [34], p. 133.
43. D. F. Williams, I. Lerche, W. E. Full, *Isotope Chronostratigraphy: Theory and Methods,* Academic Press, Inc., San Diego 1998.
44. M. J. Schoeninger in Reference [34], p. 191.
45. J. E. Ericson in P. W. Rundel, J. R. Ehrelinger and K. A. Nagy, p. 252.
46. L. Newman: *"An Evaluation of the Use of Sulfur Isotopes for Environmental Studies,"* A position paper prepared for Environmental Research and Technology as a part of the MARTEX Feasibility and Design Study, Rev. 2, Brookhaven National Laboratory, Jul. 1984.
47. A. W. Trivelpiece (ed.): *Science* **238** (1987) 888.
48. J. R. Alger et al.: *Science* **214** (1981) 660.
49. N. R. Bolo, T. F. Budinger, *Magn. Reson. Med.* **2** (1987) 591.
50. D. Halliday, G. N. Thompson in J. R. Jones (ed.): *Isotopes: Essential Chemistry and Applications II,* Royal Society of Chemistry, London 1988.
51. F. Mellon, B. Sandstrom (eds.): *Stable Isotopes in Human Nutrition,* Academic Press, London 1996.
52. K. A. Nagy in Reference [39], p. 270.
53. *Med. Sci. Bull.* **20** (1997) no. 2.
54. M.-L. Sautter-Bihl, H. Bihl, M. Wannemacher (eds.): *Systematic Radiotherapy with Monoclonal Antibodies,* Springer-Verlag, Berlin 1996.
55. S. E. Order (ed.): *NCI Monographs, 1987, Number 3, International Symposium on Labeled and Unlabeled Antibody in Cancer Diagnosis and Therapy,* U.S. Department of Health and Human Services 1987.
56. M. A. Chaudri in G. Gendreau (ed.): *9th Int. Conf. on Cyclotrons and their Applications,* Les Editions De Physique, Caen (France) 1981, p. 683.
57. Y. Wang (ed.): *Handbook of Radioactive Nuclides,* The Chemical Rubber Co., Cleveland, Ohio 1969, pp. 7 – 13.
58. L. J. Roth (ed.): *Isotopes in Experimental Pharmacology,* University of Chicago Press, Chicago 1965.

59. J. A. Forrester in E. Pohl (ed.): Proc. Intern. Colloq. Radioact. Isot. Constr. Ind. in; *Sci. J. Liepsig Inst. Constr. Ind.* (1960) translated by United States Atomic Energy Commission, Div. of Tech. Inform., (USAEC-tr-5632) p. 65.
60. J. Maya, M. W. Grossman, R. Lagushenko, J. F. Waymouth, *Science* **226** (1984) 435.
61. H. London (ed.): *Separation of Isotopes,* George Newnes, London 1961, p. 1.
62. G. M. Murphy, H. C. Urey, I. Kirshenbaum (eds.): *Production of Heavy Water,* McGraw-Hill, New York 1955, p. 53.
63. G. M. Murphy (ed.): *Separation of Boron Isotopes,* U.S. Atomic Energy Commission, Technical Information Service, Oak Ridge, Tenn, 1952, p. 69.
64. A. K. Brewer, S. Madorsky, *J. Res. Nat. Bur. Stand.* **38** (1947) no. 1, 129.
65. D. B. Trauger, J. J. Keyes, Jr., G. A. Kuiper, D. M. Lang, in J. Kistemaker, J. Bigeleisen, A. O. C. Nier (eds.): *Proc Int. Symp. Isot. Sep.* Amsterdam, Netherlands, April 23 – 27, 1957, Wiley-Interscience, New York 1958, p. 350.
66. M. Benedict, A. Boas, *Chem. Eng. Prog.* **47** (1951), 51, 111.
67. M. D. Cichelli, W. D. Weatherford, J. R. Bowman, *Chem. Eng. Prog.* **47** (1951) 63, 123.
68. P. H. Abelson, J. I. Hoover, in [65], p. 483.
69. H. U. Hostettler and K. Clusius, in [65], p. 419.
70. D. G. Avery, E. Davies: *Uranium Enrichment by Gas Centrifuge,* Mills & Boon, London, 1973.
71. E. W. Becker, K. Bier, W. Bier, R. Schutte, D. Seidel, *Angew. Chem., Int. Ed. Engl.* **6** (1967) 507.
72. R. H. M. V. Dawton, M. L. Smith, in J. Koch (ed.): *Electromagnetic Isotope Separators and Applications of Electromagnetically Enriched Isotopes,* North-Holland Publ., Amsterdam, 1958, p. 184.
73. T. I. Taylor and W. Spinel in [65], p. 158.
74. G. N. Lewis, R. T. MacDonald, *J. Am. Chem. Soc.* **58** (1936) 2519.
75. W. P. Bebbington, V. R. Thayer, *Chem. Eng. Prog.* **55** (1959) 70.
76. F. H. Spedding and J. E. Powell: *A Laboratory Method for Separating Nitrogen Isotopes by Ion Exchange,* U.S. Atomic Energy Commission Report ISC-611, Washington, D.C., 1955.
77. H. E. Gunning, *Can. J. Chem.* **36** (1958) 89.
78. M. P. Appleby, G. Ogden, *J. Chem. Soc.* (1936) 163.
79. A. K. Brewer, S. L. Madorsky, *J. Res. Nat. Bur. Stand.* **38** (1947) no. 2, 137.
80. S. J. Adelstein, F. J. Manning (eds.): *Isotopes for Medicine and the Life Science,* National Academy Press, Washington, D.C. 1995, p. 29.

Isotopes, Radioactive → Radionuclides
Itaconic Acid → Dicarboxylic Acids, Aliphatic
Jute → Cellulose
Kaolin → Clays
Karaya Gum → Polysaccharides

Ketenes

Raimund Miller, Lonza Inc., Fair Lawn, New Jersey 07410, United States

Claudio Abaecherli, Lonza AG, Visp, Switzerland

Adel Said, Lonza AG, Basel, Switzerland

Barry Jackson, Lonza AG, Visp, Switzerland

1. Introduction 1
2. Ketene 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.3. Production 3
2.4. Uses 4
3. Diketene 6
3.1. Physical Properties 6
3.2. Chemical Properties 6
3.3. Production 8
3.4. Storage, Handling, and Transportation 9
3.5. Uses 9
3.5.1. First-Generation Diketene Derivatives 9
3.5.2. Second-Generation Diketene Derivatives 11
4. Higher Ketenes 12
4.1. Properties 12
4.2. Production and Uses 12
5. Toxicology and Occupational Health 12
6. References 14

1. Introduction

Ketenes are described by the general formula $R^1R^2C{=}C{=}O$. Of commercial and industrial importance are mainly the lowest member of this class, ketene itself, $H_2C{=}C{=}O$, and its dimer, diketene. The latter has attained a prominent role in industrial synthetic chemistry. Ketene and diketene lend themselves to a number of useful conversions because of their multifunctionality.

The parent substance, monomeric ketene, was first prepared in 1907 by Wilsmore by immersing a glowing platinum wire in liquid acetic anhydride. Staudinger subsequently described its preparation by the reaction of bromoacetyl bromide [*598-21-0*] with zinc. The first industrial use of ketene appeared in the 1920s, when larger quantities of acetic anhydride were required for the production of cellulose acetate.

2. Ketene

2.1. Physical Properties [1–13]

Ketene [*463-51-4*], $CH_2{=}C{=}O$, C_2H_2O, M_r 42.02, is a colorless, toxic gas with a characteristic odor; it is unstable at room temperature and atmospheric pressure. Ketene is very soluble in acetone and lower esters; it is only slightly soluble in hydrocarbons and halogenated hydrocarbons.

Some important physical properties of ketene follow:

mp	− 150 °C
bp at 101.3 kPa	− 49 °C
Vapor pressure at − 120 °C	0.53 kPa
− 71 °C	19.5 kPa
20 °C	1.9985 MPa
30 °C	2.6355 MPa
40 °C	3.4137 MPa
50 °C	4.3506 MPa
90 °C	10.0253 MPa
Density	
gas	1.93 g/cm^3
liquid (− 60 °C)	0.650 g/cm^3
Vapor density (theoretical, air = 1)	1.45
Free energy of formation, ΔG_f^0	-46.9 ± 1.6 kJ
Enthalpy of formation, ΔH_f^0	− 47.7 kJ
Dipole moment	1.41 D

Crude ketene, as obtained from the cracking of acetic acid and used industrially, contains ca. 8 vol % acetic anhydride and 7 vol % inert gases (mainly carbon monoxide, carbon dioxide, ethylene, and methane). Ketene from the cracking of acetone also contains a stoichiometric amount of methane.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_063

2.2. Chemical Properties [1–13]

Ketene is an extremely reactive and unstable compound: it dimerizes to diketene and polymerizes very quickly. The 4-methyleneoxetan-2-one structure of diketene was established in 1952 by X-ray diffraction [15]. Under the conditions for the industrial production of diketene, less than 5 % of ketene dimerizes to 1,3-cyclobutanedione in its monoenolic form, which in turn is further acetylated. It can only be stored at liquid nitrogen temperature. To prevent polymerization, reactions with ketene are best conducted under reduced pressure. If this is not possible, the reaction vessel should be directly connected to the ketene source or compressor.

To avoid formation of acetic acid and acetic anhydride, reactions with ketene should be carried out in a dry atmosphere, but because it does not form peroxides (in contrast to the ketoketenes) ketene does not require an oxygen-free atmosphere.

The main reactive center in ketene is the carbonyl carbon atom, which is strongly electrophilic. Consequently, ketene reacts readily with nucleophiles. The electronic structure of ketene can be described by the following canonical forms:

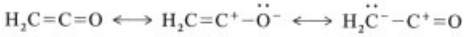

More modern interpretations use a molecular orbital approach [11].

Ketene preferentially undergoes addition reactions with substances containing polar or easily polarized single or multiple bonds. Addition usually takes place across the C=C double bond, and the oxygen atom is seldom involved. In the absence of other reaction partners, ketene adds a second molecule of ketene to form diketene. The main types of reactions that ketene participates in are acetylation, cycloaddition, ketene insertion, and polymerization.

Acetylation. Compounds with a reactive hydrogen atom are acetylated by ketene:

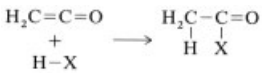

Ketene acetylates substances containing hydroxyl groups, e.g., water is acetylated to acetic acid, aliphatic alcohols to alkyl acetates, phenols to aryl acetates, carboxylic acids to anhydrides, enols to enol acetates, hydrogen peroxide to diacetyl peroxide, and hydroperoxides to peracetates. Acidic and basic catalysts strongly increase the reaction rate [16]. Metal salts are also effective catalysts [17].

Sulfur analogues react with ketene similarly: hydrogen sulfide is acetylated to thioacetic anhydride via thioacetic acid, and aliphatic thiols yield alkyl thioacetates. Hydrogen halides give the corresponding acetyl halides. The reaction with hydrogen sulfide requires a catalyst; other sulfur compounds do not necessarily require catalysis.

Ketene reacts with ammonia, primary and secondary amines, amides, imides, hydrazines, hydrazones, ureas, and azomethines to give the corresponding monoacetates; sometimes the di- and triacetates are also formed. With strong bases the reaction can be quite violent. For weakly basic or substituted NH groups, acids or amides are recommended as catalysts [18]. Ketene can selectively acetylate amino groups in the presence of hydroxyl groups. Examples are the acetylation of ammonia [1, p. 124] and the *N*-acetylation of ethanolamine in aqueous solution. Alcohols or even water can be used as solvents for such reactions.

Cycloadditions. Substances bearing a double bond (e.g., C=C, P=N, C=X, or N=X, where X = O, N, or S) can undergo [2 + 2] 1,2-cycloaddition with the ketene C=C bond. This synthetically useful reaction leads to versatile four-membered rings:

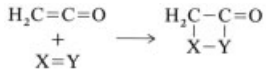

A catalyst is often necessary. Ketene itself does not react with nonpolar olefins (ketoketenes and halogenoketenes are more reactive), but it does react with 1,3-dienes and active olefins such as vinyl ethers, vinyl esters, enamines, alkoxyacetylenes, and allenes. Carbonyl compounds, especially those with electron-withdrawing groups, give 2-oxetanones (β-lactones). The dimerization of ketene to diketene falls into this category.

Optically active 2-oxetanones have been obtained with chiral tertiary amines as catalyst

in cycloadditions with electron-deficient aldehydes and ketones [19]. Electron-rich aldehydes give chiral β-lactones when a chiral Lewis acid is used [20]. Thietanones are obtained from thiocarbonyl compounds; nonenolizable imines (azomethines) give 2-azetidinones (β-lactams) [21], [22]. 1,3- and 1,4-cycloadditions are also known, though rare; an example follows (yield ca. 16 %):

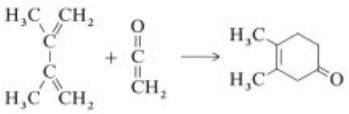

For a review of cycloadditions with ketenes see [1], [8], [11], [13], [23].

Recent work has shown that some 1,2-cycloadditions are in fact 1,4-cycloadditions to the C–O bond followed by a 1,3-Claisen rearrangement [24].

Ketene Insertion. Substances with strongly polarized or polarizable single bonds add to the ketene C=C double bond in a reaction known as ketene insertion:

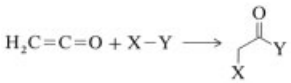

Thus, active acyl chlorides with an electron-withdrawing substituent (e.g., CCl_3 or CF_3) give the corresponding 3-ketoacyl chlorides, phosgene inserts two moles of ketene to yield acetone dicarboxylic acid dichloride [25], and α-haloethers give 3-oxa-substituted propionyl halides. Acetals and ortho-esters with Lewis acid catalysis give β-alkoxy carboxylic esters [26].

Other insertion reactions take place with chlorine (to give chloroacetyl chloride), bromine, thionyl chloride, and with mineral acids.

Polymerization of ketene with various catalysts can give either polyesters (**1**) or polyketones (**2**); higher ketenes also give polyacetals (**3**) [27].

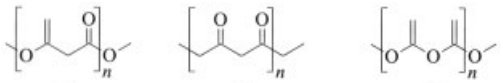

1 **2** **3**

Ketene equivalents, such as the stable trimethylsilylketene, have been described for synthetic use [28], [29].

Reduction of ketene with Pd/H_2 to give acetaldehyde and/or vinyl acetate allows the production of these intermediates directly from a syngas (CO/H_2) source [30].

2.3. Production

Ketenes are the internal anhydrides of the corresponding carboxylic acids, and as such can be made by removing a molecule of water from these acids, either directly or indirectly. Numerous methods have been used to convert a precursor to the corresponding ketene [8], [9], [11], [13].

Ketene is produced by *pyrolysis of acetic acid* [*64-19-7*], which, although endothermic, is quite an efficient process. Decomposition products formed in the pyrolysis of acetic acid to ketene include methane, hydrogen, carbon monoxide, carbon dioxide, and ethylene. Higher acids can also be cracked to the corresponding ketenes, although the efficiency of conversion decreases as the chain length increases. The corresponding carboxylic anhydrides give better results [31], [32]. Long-chain alkylketenes are made industrially from the corresponding fatty acid chlorides. Ketene generators and ketene reaction apparatus for both the laboratory and production scale have been reviewed [33].

High-quality acetic acid is evaporated and the vapor passed continuously through a radiant coil under reduced pressure at 740 – 760 °C. Triethyl phosphate [*78-40-0*] catalyst is injected into the acetic acid vapor. Ammonia is added to the gas mixture leaving the furnace to prevent ketene and water from recombining and to neutralize the catalyst. The gas mixture is chilled to < 100 °C to remove water, unconverted acetic acid, and the acetic anhydride formed as a liquid [34], [35]. The gaseous ketene is absorbed in the appropriate reaction medium or dimerized in a liquid ring pump (liquid seal pump) if diketene is the desired reaction product. The medium consists of a liquid substance or a solution of a compound that is to be treated with ketene (e.g., alcohols, amines, and thiols). Diketene itself is

the preferred dimerization medium for ketene. The crude ketene gas is at a pressure of 20–30 kPa before entering the dimerization unit. It can therefore only be used for reactions with higher boiling reactants that will not vaporize at this pressure. In the case of more volatile reactants or when a higher reaction temperature is required, the crude ketene is compressed to ambient pressure. This is best done in a liquid ring pump by using a liquid in which ketene is only sparingly soluble (e.g., decalin) [36].

A useful laboratory-scale synthesis of high-purity ketene has been described [37]. This method, which gives very pure ketene by *pyrolysis of diketene*, can also be employed on an industrial scale. Thus, a convenient and technically feasible process for producing pure ketene, uncontaminated by methane and other hydrocarbons and oxides, is available. Conversion of diketene to monomeric ketene is accomplished on an industrial scale by passing diketene vapor through a tube heated to 350–600 °C:

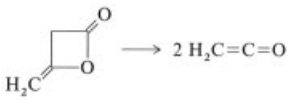

The pyrolysis of dimeric ketenes, whether in the β-lactone or cyclobutane-1,3-dione form, is a general method for obtaining pure ketenes in high yield.

Ketene can also be generated on an industrial scale by *thermal decomposition of acetone or acetic anhydride.* Nevertheless, the former is only interesting as a raw material when the price is low.

2.4. Uses

Ketene, the internal anhydride of acetic acid, is a powerful acetylating agent (see Section 2.2). It reacts generally at the C=C double bond with compounds containing an active hydrogen atom to give an acetyl derivative:

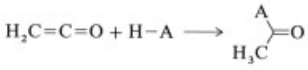

Production of Acetic Anhydride. Most of the ketene produced worldwide is used in the production of acetic anhydride (see also → Acetic Anhydride and Mixed Fatty Acid Anhydrides, Chap. 1.3.).

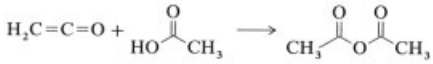

In the early 1980s, the Tennessee Eastman Company developed a new process for the coal-based production of acetic anhydride by carbonylation of methyl acetate [*79-20-9*] [38]. This process has economic advantages in times of high oil prices.

Production of Diketene. The second most important industrial use of ketene is the production of diketene by controlled dimerization (see Section 3.2).

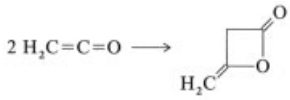

The incorporation of the acetoacetate moiety into pigments, pharmaceuticals, and agricultural chemicals is an important area of synthetic chemistry.

The diketene–acetone adduct is a nonlachrymatory liquid which can be safely handled at room temperature.

Production of Sorbic Acid. The third major use of ketene, albeit less important than the first two mentioned above, is in the production of sorbic acid [*110-44-1*] [39], [40], [41]. Reaction of ketene with crotonaldehyde [*123-73-9*] gives, depending on the conditions, either the β-lactone **4** or the polyester **5**. Both products can be converted, thermally or by acidic catalysis, to sorbic acid:

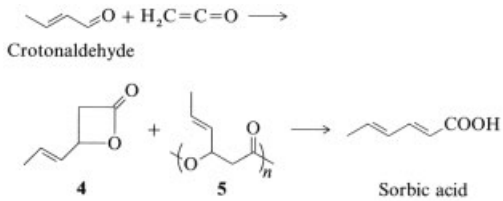

The corresponding reaction of ketene with formaldehyde, once an important use of ketene in the production of acrylates via propiolactone,

is no longer used on a commercial scale because the intermediate β-propiolactone [*57-57-8*] (2-oxetanone) is a suspected carcinogen.

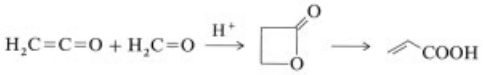

The most economical process for the production of acrylates is now based on the two-stage vapor-phase oxidation of propene to acrylic acid [42] (see also → Acrylic Acid, Chap. 1.3.).

Production of Chloroacetyl Chloride. Chlorine adds to ketene to form chloroacetyl chloride [*79-04-9*] [43]:

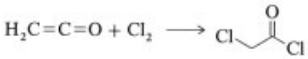

Chloroacetyl chloride has varied industrial uses, such as the production of herbicides of the alachlor type and local anesthetics of the lidocaine type.

4,4,4-Trihaloacetoacetates. In the presence of strong Lewis acids such as boron trifluoride, appropriately substituted acyl chlorides add to ketene to form the corresponding acetoacetyl chlorides, which can then be treated with an alcohol to form the corresponding haloacetoacetates.

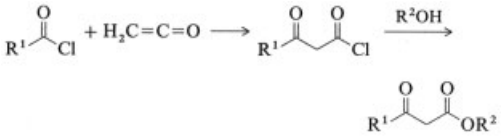

Of industrial significance are ethyl 4,4,4-trifluoroacetoacetate [*372-31-6*] for the production of herbicides and antimalarial agents, and ethyl 4,4,4-trichloroacetoacetate [*3702-98-5*] for the production of pharmaceuticals.

Isopropenyl Acetate and Acetylacetone. Acetone, like other carbonyl compounds containing active hydrogen atoms, undergoes enol acetylation on reaction with ketene; a strong acid catalyst is required. The isopropenyl acetate [*108-22-5*] formed [44], [45] is the starting material for the production of acetylacetone [*123-54-6*], which is an intermediate for the sulfa drug sulfamethazine [*57-68-1*].

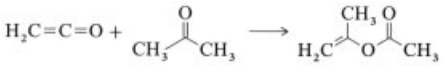

The reaction of acetone with ketene with Lewis acid catalysis gives 4,4-dimethyloxetan-2-one [*1823-52-5*], which is used to produce the food-additive calcium 3-hydroxy-3-methylbutyrate [*135236-72-5*] [46].

Chiral Lactones. Chloral and 1,1,1-trichloroacetone react with ketene to form the corresponding 4-trichloromethyloxetanones. In the presence of an optically active catalyst, chiral lactones are obtained [19], [47], [48].

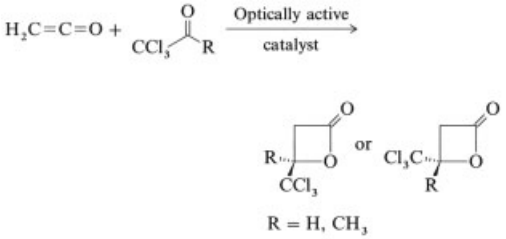

The synthesis of chiral β-lactones [49–51] has assumed greater importance due to pharmaceutical activity found for some β-lactone structures [52].

Cinnamic Esters. The insertion reaction of ketene with *p*-methoxybenzaldehyde [*2186-92-7*] is used to produce 2-ethylhexyl *p*-methoxycinnamate [*5466-77-3*], a sunscreen used in cosmetics [53] (see top of the next page).

3. Diketene

3.1. Physical Properties

Diketene [*674-82-8*], $C_4H_4O_2$, M_r 84.04, is a colorless liquid when pure; it possesses a pungent odor and is a strong lachrymator. The human odor perception threshold is 0.019 mg/m^3 [54].

Crude diketene obtained from the dimerization of ketene is often used without further purification; it is dark brown and contains up to 10 % higher ketene polymers.

Some physical properties of diketene are listed below [1–7]:

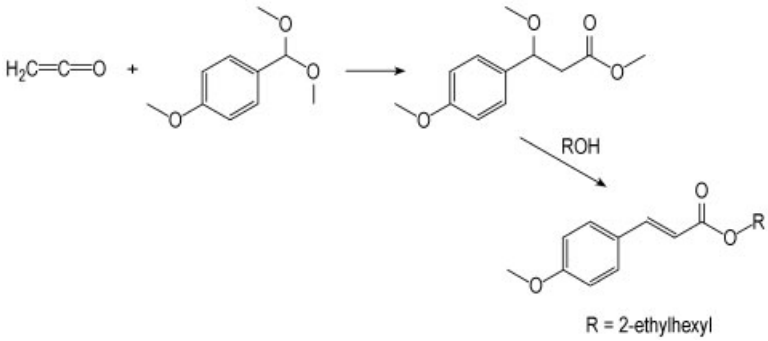

mp	− 7.5 °C
bp	
101.3 kPa	127 °C
13.3 kPa	69 – 71 °C
5.5 kPa	50.5 °C
3.1 kPa	38.5 °C
Vapor pressure (20 °C)	1.07 kPa
d_4^{20}	1.09471
d_4^{25}	1.088754
Vapor density (20 °C, theoretical, air = 1)	2.9
n_D^{20}	1.4379
Surface tension, mN/m	
20 °C	31.82
24 °C	30.37
25 °C	33.89
Critical pressure	5.47 MPa
Critical temperature	310 °C (583.16 K)
Viscosity	0.88 mPa · s
Flash point (open cup)	33 °C
Ignition temperature (air)	291 °C
Explosive limits in air, vol%	
lower	2 ∗
upper	11.7
Dipole moment	3.53 D
Dielectric constant	16 – 17
Standard enthalpy of formation ΔG_f^0	− 233.41 J

∗ A value of 0.2 % has also been reported [55].

Diketene is miscible with acetic anhydride, with many ketones and esters (e.g., acetone, methyl acetate, ethyl acetate, butyl acetate) with diethyl ether, tetrahydrofuran, chlorinated hydrocarbons, acetonitrile, and toluene.

Furthermore, contrary to most literature reports, it is slightly soluble in aliphatic hydrocarbons such as hexane (ca. 12.5 wt % at 23 °C), octane, cyclohexane, and methylcyclohexane. It is very slightly soluble in water, in which it decomposes slowly.

3.2. Chemical Properties

Ketene itself forms primarily the unsymmetrical lactone dimer with less than 5 % being converted to the symmetrical dimer, 1,3-cyclobutanedione [*15506-53-3*], which in turn is converted to the ketene trimer [*38425-52-4*] by enol acetylation [56]. This trimer can be used to synthesize the dye intermediate squaric acid [57].

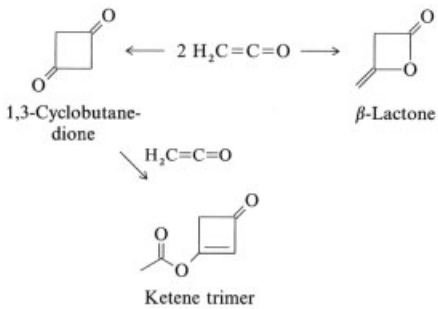

Diketene is a reactive, versatile compound, which cannot be stored indefinitely, even in the presence of stabilizing agents such as sulfates, borates, or sulfur [58]. On standing for several days or weeks, it turns yellow and then brown. Glass vessels are not recommended for storage because the basicity of glass promotes decomposition: stainless steel or aluminum is preferred.

Diketene reactions proceed mainly by ring opening and are highly exothermic. Traces of base or strong acid can lead to violent, almost explosive polymerization, and due care should be exercised when handling diketene.

Only the most important types of the large number of reactions of diketene are briefly discussed here. More detailed information can be found in [1–7], [12], [59].

Pyrolysis. Diketene is cleaved to two molecules of ketene by pyrolysis at 350 – 600 °C [34], and it is a convenient laboratory source of ketene (see Section 2.3). In a copper reactor at 550 °C and with admixture of nitrogen, decomposition can lead to allene and carbon dioxide [60].

Reactions at the Olefinic Double Bond (without Ring Opening) [61–63]. Hydrogen adds catalytically to the exocyclic olefinic double bond of diketene to form β-butyrolactone (4-methyl-2-oxetanone). This hydrogenation can be made enantioselective by using a chiral catalyst [64]. Other reactions include radical addition (e.g., with halogenated compounds, secondary alcohols, alkylthiols, dialkyl phosphites) to give 4-substituted oxetanones. Diketene can also undergo addition of carbenes and nitrenes, and photochemical [2 + 2] and thermal [4 + 2] cycloadditions with olefins and other unsaturated compounds (Diels – Alder) to give spirocyclic compounds.

Polymerization of diketene under free-radical conditions leads to a polymeric form with intact β-lactone rings. Ozonolysis to give the highly unstable malonic anhydride has been reported [65], [66].

Ring-Opening Reactions. In most reactions, the strained β-lactone ring is opened, and diketene appears to react as acetylketene or one of its tautomeric forms:

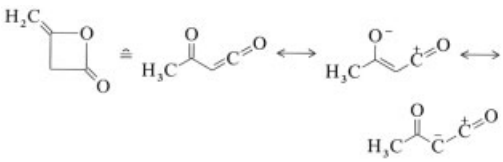

Nucleophiles normally add to the acyl carbon – oxygen bond leading to acetoacetylation:

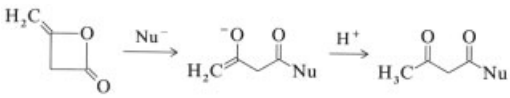

Nu = nucleophile

A variety of compounds with active hydrogen atoms react in this manner, e.g., alcohols, phenols, carboxylic acids, hydrogen halides, amines, carboxylic acids, amides, ureas, *N*-arylhydroxylamines, and sulfonamides. With the less reactive compounds, a catalyst is often required. When other functional groups are present (e.g., ureas, thioureas, hydrazines, hydroxylamines, amidines, urethanes, enamines, and amides), acetoacetylation can be followed by ring closure with either the keto group or the active methylene group of the acetoacetyl moiety. This leads to a variety of heterocyclic compounds [67], an example is:

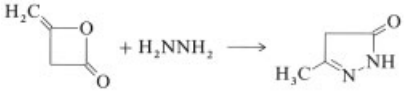

Water hydrolyzes diketene only very slowly, leading to the unstable acetoacetic acid, which further decomposes to acetone and carbon dioxide [68].

Electrophiles react first with the exocyclic olefinic double bond of diketene. Subsequent nucleophilic attack at the lactone carbonyl group gives 4-substituted acetoacetic acid derivatives. Thus, chlorine gives the versatile building block 4-chloroacetoacetyl chloride [69]. Bromine reacts analogously, and reaction with fluorine at − 78 °C gives the corresponding 4-fluoro derivatives [70].

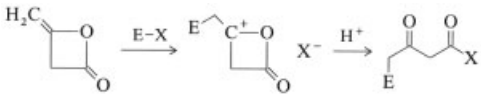

With *carbonyl compounds,* diketene gives 2,2-disubstituted 6-methyl-4*H*-1,3-dioxazin-4-ones (**6**):

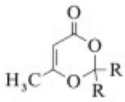

6

The use of specific Ti-catalysts produces chiral 4–alkyl-3–hydroxybutanoates [71]. The most common adduct, diketene – acetone (**6**, R = CH_3), 2,2,6-trimethyl-4*H*-1,3-dioxin-4-one, is often used as a convenient diketene replacement, because it reacts similarly to diketene, but is safer to handle and to transport [72], [73].

Analogously, compounds with C = N groups, such as carbodiimides and imines, lead to 1,3-oxazines:

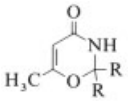

Diketene Polymers. Under various conditions, diketene dimerizes to give dehydroacetic acid [74]. This compound is frequently encountered as a byproduct in reactions of ketene and diketene.

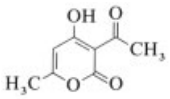

Dehydroacetic acid

With $HgCl_2$, ion-exchange resins, or alkoxides, a low molecular mass polyester of the form **1** (Section 3.1) is obtained. With BF_3 the formation of a polyketone of repeating $-(CH_2-CO-)$ units **2** has been reported [75], [76]. With vinyl monomers, diketene affords copolymers with useful properties for adhesives, unsupported films, and molded objects [77].

3.3. Production

Ketene possesses a considerable tendency to polymerize. Under controlled conditions and in the appropriate dimerization medium, ketene can be dimerized to diketene, with yields of up to 85 %, based on acetic acid (see also Section 2.2). The remaining 15 % is made up by higher ketene oligomers [56]. The preferred dimerization medium is diketene itself [78]. Crude diketene, containing 3 – 4 % acetic anhydride, is a mobile, dark brown liquid due to 8 – 10 % higher polymers. Diketene of > 99.5 % purity can be obtained by distillation [79]. Ultrapure diketene (99.99 %) can be obtained by crystallization [80].

Diketene is now generally consumed at the site of production, because of its extreme reactivity and hazardous properties. In many cases, the 1 : 1 adduct of diketene and acetone, 2,2,6-trimethyl-4*H*-1,3-dioxin-4-one [*5394-63-8*] (**6**, R = CH_3) can be used in place of diketene [72].

3.4. Storage, Handling, and Transportation

Pure diketene is surprisingly stable at temperatures up to ambient in the absence of acid or base catalysts. Liquid diketene is stable for several weeks at 0 °C, slow polymerization occurs above 15 °C. At elevated temperature, diketene undergoes slow decomposition and discoloration. Diketene can be stabilized with agents such as water [81], sulfur [82], alcohols, phenols, borate, and sulfate salts. In Japan, lactic acid is reportedly used. Diketene, stabilized with 1 % anhydrous copper sulfate, retains its strength for up to one year at 25 °C.

Nevertheless, due to the high reactivity of diketene and transportation restrictions in certain countries, diketene is mainly used at the site of production. Hazard classifications for transportation follow:

IMDG code: class 3.3
RID/ADR: class 3, no. 31 c
CFR 49: 172.101 flammable liquid
UN number: 2521

In the United States and Japan, diketene is shipped in refrigerated tank trucks. Care must be taken to keep shipping time to a minimum and to ensure that the diketene is used upon arrival. Due to the potential hazard of shipping diketene, however, this practice is being discontinued in the United States.

Most reactions with diketene are strongly exothermic; it is therefore generally added in a small stream or in portions to other reactants. A build-up of unconverted diketene is thereby avoided. Diketene should not be stored in glass bottles due to the alkalinity of glass; stainless steel or aluminum containers should be used.

The total annual worldwide production of diketene is estimated at ca. 130 000 t. The major producers are Lonza, Eastman Chemical, Clariant, Wacker, Daicel, and BASF [14].

3.5. Uses

Diketene is a powerful bactericidal agent [83] and can raise the octane rating of gasoline [84],

but is not used for these purposes in the Western hemisphere.

The main reactions and uses of diketene have been comprehensively reviewed [7], [12], [59]. The most important industrial uses and some newer developments are highlighted in the following sections.

3.5.1. First-Generation Diketene Derivatives

First- and second-generation diketene derivatives are depicted in Figure 1.

Acetoacetates. In the presence of tertiary amine catalysts, diketene reacts rapidly with lower alcohols. Industrially preferred catalysts include trimethylamine [*75-50-3*], triethylamine [*121-44-8*], and pyridine [*110-86-1*] [85]. Acidic catalysts such as sulfuric and sulfonic acids can also be used [86], [87]. The industrial acetoacetylation of alcohols is carried out continuously in a loop reactor at 50 – 80 °C without solvent. A comprehensive treatment of acetoacetate chemistry is given in [12], [88]. Methyl and ethyl acetoacetates [*141-97-9*] are the most widely used esters, finding use in the pharmaceutical, agricultural, and allied industries.

2-Chloro- and 4-Chloroacetoacetates. The most specific chlorinating agent for acetoacetates in the 2-position is sulfurylchloride. Alternatively, diketene can be treated with hydrogen chloride at low temperature in chlorinated hydrocarbon solvents to afford acetoacetyl chloride [89], which is then chlorinated in the 2-position in excellent yield to give 2-chloroacetoacetyl chloride. The latter can be further treated with alcohols or with aliphatic or aromatic amines to afford the corresponding 2-chloroacetoacetate derivatives [90].

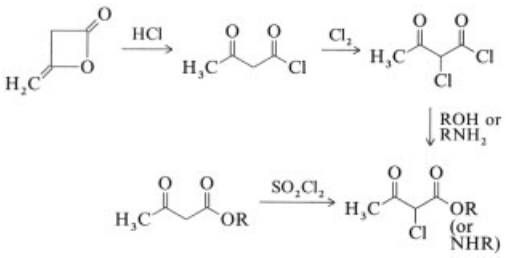

4-Chloroacetoacetates are conveniently prepared on an industrial scale by reaction of gaseous chlorine with diketene to form 4-chloroacetoacetyl chloride [91], which can react further with chlorine to afford 2,4-dichloroacetoacetyl chloride. Subsequent reaction with alcohols gives the corresponding 2,4-dichloroacetoacetates [92].

4-Chloroacetoacetates can be converted to the corresponding 4-alkoxyacetoacetates [93], useful starting materials for tetronic acid [94], [95]. Treatment of 2-chloro- and 2,4-dichloroacetoacetyl chlorides with water in the cold affords chloroacetone and 1,3-dichloroacetone, respectively [96].

6-Methyluracil [*626-48-2*] is produced by condensation of diketene with urea. It is used in the production of the analgesic mepirizole, the feed additive orotic acid, and 5-aminoorotic acid.

Diketene – acetone adduct [*5394-63-8*] can be regarded as a stabilized form of diketene. It can be safely transported and may conveniently be used instead of diketene in many reactions. For example, it reacts with alcohols and amines to yield acetoacetates and acetoacetamides, respectively.

Diketene – acetone adduct generates unstable acetylketene together with acetone when heated above 120 °C [97]. Subsequent condensation with isocyanates, arylcyanates, or substituted cyanamides gives access to a wide range of 1,3-oxazine derivatives [98]. Furthermore, diketene – acetone adduct can also be used for the preparation of a variety of compounds that are not accessible from diketene [72].

Dehydroacetic Acid. Controlled dimerization of diketene in the presence of tertiary amine catalysts produces dehydroacetic acid [*771-03-9*], DHA, 3-acetyl-4-hydroxy-6-methyl-2-pyrone [99]. Many other bases are also useful dimerization catalysts, e.g., sodium acetate, potassium sorbate, and sodium alkoxides or sodium phenoxides.

On standing at room temperature, diketene is slowly converted to DHA, which is frequently observed as a byproduct in reactions of diketene. In most cases, DHA can be easily removed from

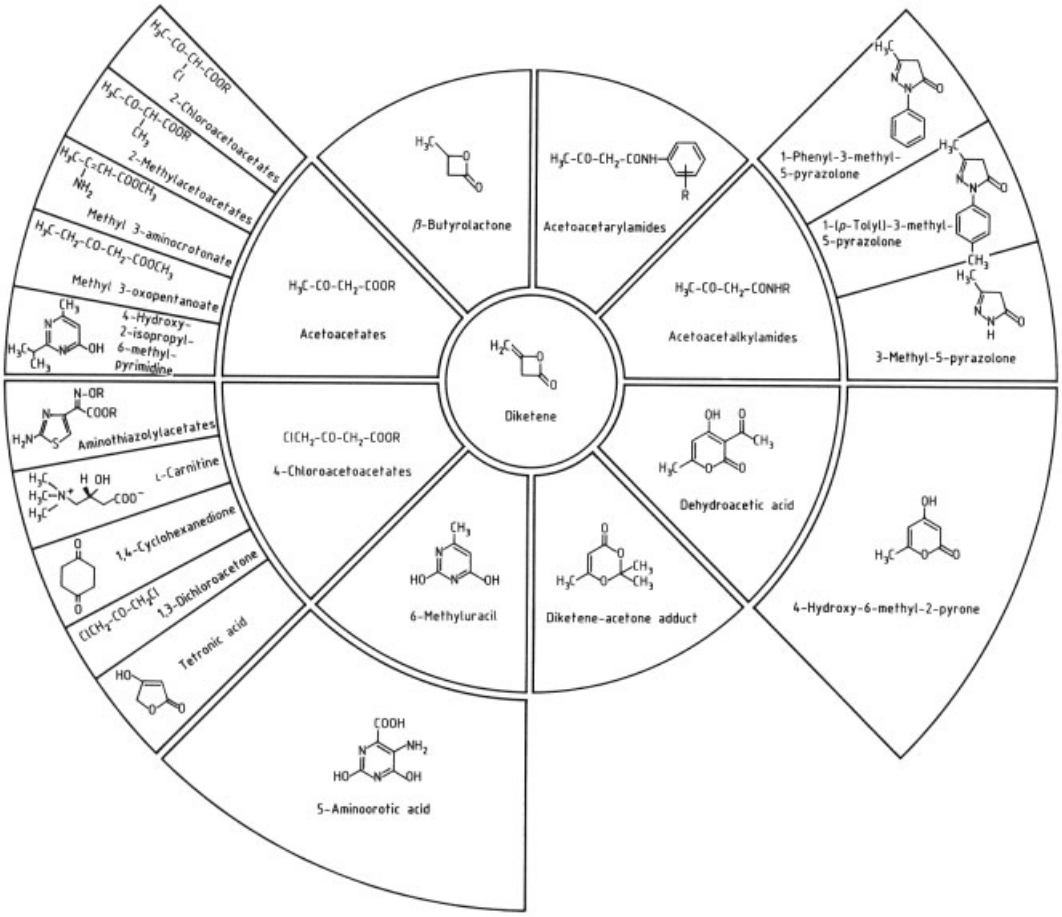

Figure 1. Diketene derivatives of the first generation (inner segments) and second generation (outer segments)

the reaction mixture by bicarbonate extraction or by precipitation.

Dehydroacetic acid or its sodium salt is used as a food and cosmetic preservative, as a starting material for the veterinary drug clopidol [*2971-90-6*] (Coyden), and in the production of 4-hydroxy-6-methyl-2-pyrone.

Acetoacetalkylamides and Acetoacetarylamides. Aqueous ammonia and alkylamines react readily with diketene to form acetoacetamide and acetoacetalkylamides, respectively, whereas arylamines react more slowly to afford acetoacetanilides. Arylhydrazines yield 3-methyl-1-aryl-5-pyrazolones. Sulfamic acid reacts to give, after cyclization, the sweetener Acesulfame [100].

Acetoacetamide and acetoacetalkylamides are mainly used as aqueous solutions. The main uses for the alkylamides are in the production of insecticides such as monocrotophos, dicrotophos, phosphamidon, and oxamyl. *N*-(2-Hydroxyethyl)-3-oxobutanamide [*24309-97-5*] is used in the production of olaquindox, a growth stimulant for pigs.

Acetoacetarylamides are mainly used in the production of high-quality organic pigments, particularly as azo coupling components, e.g., for a variety of yellow and orange pigments. The simplest compound of this series, acetoacetanilide [*102-01-2*], is also used in the production of the fungicides carboxin and methfuroxam, in addition to its use in the production of the pigment yellow 12.

β-Butyrolactone. Diketene undergoes catalytic hydrogenation to give β-butyrolactone, 4-methyl-2-oxetanone [*3068-88-0*] in ethyl acetate at 0 °C over a palladium catalyst [102]. β-Butyrolactone is used in the production of the analgesic bucetin. Enantioselective reduction to chiral β-butyrolactone has been reported [64].

3.5.2. Second-Generation Diketene Derivatives

A number of second-generation diketene derivatives (see Fig. 1) have already been mentioned in Section 3.5.1. A few are selectively described here.

3-Aminocrotonate esters, e.g., methyl 3-aminocrotonate [*14205-39-1*] (3-amino-2-butenoic acid, methyl ester) can be prepared by treatment of acetoacetates with aqueous ammonia [103]. They are mainly used in the production of a series of the dihydropyridine-type calcium antagonists.

2-Aminothiazol-4-yl Acetates. Diketene can be used in a number of ways for the preparation of substituted acetoacetamido side chains for β-lactam antibiotics, e.g., by conversion of the 4-chloroacetoacetamido side chain of 4-chloroacetoacetamidocephalosporanic acid with thiourea [104]. Alternatively, the aminothiazolylacetic acid side chains can be prepared separately for attachment to a variety of β-lactams [105] (e.g., semisynthetic cephalosporins of the third and fourth generation and of carbapenems [106]).

Acetoacetoxyethyl methacrylate [*21282-97-3*], AAEM, has a pendant acetoacetoxyethyl group which can modify the physical properties of acrylic polymers by decreasing solution viscosity and lowering the glass transition temperature. In addition, the acetoacetyl group contains an active methylene group and a carbonyl group, which can be used to modify or cross-link the polymer [107].

4. Higher Ketenes

4.1. Properties

Higher ketenes are divided into aldoketenes RCH=C=O and ketoketenes $R^1R^2C{=}C{=}O$ depending on whether one or two hydrogen atoms of ketene are substituted by an alkyl or aryl group. They are less important industrially than ketene or diketene.

Physical Properties. Methylketene is a gas at room temperature; most other aldo- and ketoketenes are yellow to reddish-brown liquids, some higher diarylketenes are solids. Some physical properties of the most common aldo- and ketoketenes are given in Table 1.

Chemical Properties. The chemical behavior of the higher ketenes is essentially the same as that of ketene [1] although there are some noticeable differences. Aldoketenes are extremely reactive, and it is very difficult to isolate the monomeric form. They are, therefore, generated and used in situ [1], [2], [9], [11].

Ketoketenes differ from ketene and aldoketenes in one important respect: they are oxidized by molecular oxygen even at temperatures well below 0 °C (− 80 °C has been reported for some aryl ketenes), forming hazardous peroxides that can explode at the slightest touch. Ketoketenes must therefore only be handled under an oxygen-free atmosphere. In all other reactions, including polymerization, ketoketenes are less reactive than aldoketenes or ketene itself: more forcing reaction conditions or a catalyst are often required. Ketoketenes, particularly diarylketenes, can be stored in oxygen- and catalyst-free solutions for longer periods, up to several months, due to their lower reactivity. Small amounts of aluminum trialkyl can be added to prevent peroxide formation [31].

4.2. Production and Uses

Various methods are available for the production of higher ketenes; the method used depends on the ketene [11], [108].

Dialkylmalonic acids are useful starting materials for the preparation of dialkylketenes. When dimethylmalonic acid is heated with

Table 1. Physical properties of higher ketenes

Name	CAS registry number	M_r	*bp* °C (kPa)	n_D^{20}
Methylketene	[*6004-44-0*]	56.06	− 56 (101.3)	
Ethylketene	[*20334-52-5*]	70.09	36 (101.3)	
Phenylketene	[*3496-32-0*]	118.14		
Dimethylketene	[*598-26-5*]	70.09	34 (101.3) (*mp* 98 °C)	
Methylethylketene	[*36854-53-2*]	84.12	− 26 (1.60)	
Diethylketene	[*24264-08-2*]	98.15	91 (101.3)	1.4126
Di-*n*-propylketene	[*58844-38-5*]	126.20	30 (1.47)	
Di-*n*-butylketene	[*36638-69-4*]	154.25	68 (1.40)	1.4308
Di-*tert*-butylketene	[*19824-34-1*]	154.25	73 (6.0)	1.4370
Methylphenylketene	[*3156-07-8*]	132.16	74 (1.6) (*mp* − 8 to − 5 °C)	
Ethylphenylketene	[*20452-67-9*]	146.19	42 (0.04)	
Diphenylketene	[*525-06-4*]	194.23	119 (0.47) (*mp* 9 – 10 °C)	1.615

acetic anhydride, the intermediate cyclic acylal of dimethylketene and dimethylmalonic acid [*595-46-0*] can be isolated [109]. This compound gives dimethylketene on pyrolysis at 130 °C.

Diphenylketene is made by removal of hydrogen chloride from diphenylacetyl chloride with triethylamine [110]. This method is generally useful for the in situ preparation of ketenes, whereby the transient ketene can be captured by a ketenophile [9], [11], [111].

Long-chain monoalkylketenes (aldoketenes) and their lactone dimers, the alkylketene dimers, have been used for paper sizing for many years. They are prepared by removal of hydrogen chloride from long-chain fatty acid chlorides, e.g., stearic acid chloride. The intermediate highly reactive ketene dimerizes to the lactone dimer (*mp* ca. 50 °C) in the presence of acidic catalysts [112]:

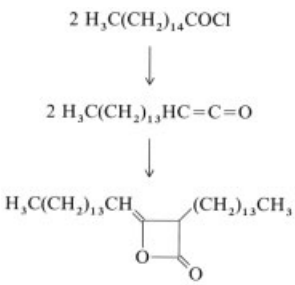

Emulsions of these sizing agents (so-called AKD's) form covalent bonds with the hydroxyl groups of cellulose below 100 °C, making fiber surfaces hydrophobic [113–115].

5. Toxicology and Occupational Health

Ketene is one of the most irritating gases to the respiratory tract, due to its great reactivity towards nucleophiles such as hydroxyl, amino, carboxyl, thiol, and phenolic groups, leading to acetylation of proteins. Its toxicity is considered to be of the same order of magnitude as that of phosgene [117]. Like the latter, it also causes latent damage to the respiratory tract (pulmonary edema) that may become acute only many hours after exposure.

Different, sometimes conflicting values, are found in the literature for inhalation toxicity [54], [117–119]. This could indicate a different sensitivity in different species, but is more likely to be due to the instability of ketene, which makes it difficult to maintain and measure a low concentration over a long time. The LCLo (mouse) is 23 ppm (30 min), 53 ppm (100 min) in the rat, rabbit, and guinea pig and 200 ppm (10 min) in the monkey; other sources report an LC_{100} of 200 ppm (5 min) in the rat. No toxic effects were reported with a concentration of 1 ppm for 7 h/d during 55 days. No human casualties have been reported, although irritation of the respiratory tract and of the eye due to unknown concentrations of ketene among workers is known. After inhalation of even small quan-

tities, a characteristic, long-lasting, unpleasant taste can be observed; smokers seem to be more sensitive to this phenomenon.

The LD_{50} of ketene (rat, oral) is 1300 mg/kg [119].

The TLV-TWA exposure limit for ketene is 0.5 ppm (0.9 mg/m^3), which is also the MAK value in the Federal Republic of Germany and in Switzerland. Regarding the maximal allowable instantaneous concentration, the TLV-STEL value is 1.5 ppm; the German MAK commission classifies ketene in category I with a limit twice the MAK (corresponding to 1 ppm) which is to be tolerated for no longer than 5 min, and no more than eight times a day [119–122]. No carcinogenic activity for ketene has been reported.

Diketene [117–119], [122], [123]. Although less acutely toxic than ketene, diketene is a strongly irritating agent and a powerful lachrimator. It is especially dangerous to the eye (in which it rapidly damages the corneal tissue) and to the respiratory tract; formation of pulmonary edema is possible for up to two days after exposure.

Symptoms of diketene exposure are a burning sensation in the eyes, nose, and throat, coughing, difficulty in respiration and ultimately loss of consciousness. A physician should be consulted immediately after inhalation of the vapor. Liquid diketene can be absorbed through the skin, where it causes itching and burning. No chronic toxic effects are known.

The acute oral toxicity (LD_{50}) is 560 mg/kg (rat) and 800 mg/kg (mouse). The reported LCLo is 20 000 ppm for an exposure of 1 h. The acute dermal toxicity LD_{50} in the rabbit has been given as 2.83 mL/kg and 6730 mg/kg.

No carcinogenic effects of diketene have been found on skin application in mice or subcutaneous injection and intragastric feeding in rats [124–126].

No permissible exposure limits have been set. Diketene has a pronounced "warning" effect due to its very low irritation threshold, and, when inhaled, it becomes unbearable before a lethal limit is reached.

Higher Ketenes. No toxicological data have been reported for higher ketenes, but it is prudent to consider them at least as hazardous and toxic as ketene itself.

6. References

General References

1. *Houben-Weyl,* **VII/4,** 53 – 447; **E15/2+3,** 2353 – 2881, 2933 – 3146.
2. S. Patai (ed.): *The Chemistry of Ketenes, Allenes and Related Compounds (The Chemistry of Functional Groups Series),* Wiley-Interscience, 1980.
3. *Beilstein,* **H1** 724; **EI1** 376 – 377; **EII1** 779; **EIII1** 2942 – 2947; **EIV1** 3418 – 3420.
4. *Beilstein,* **H7** 552; **EI7** 309; **EII7** 525; **EIII/IV17/5** 4297 – 4299; **EV17/9** 115 – 118.
5. *Ullmann,* 4th ed., **14,** 181 – 189.
6. *Kirk-Othmer*, 4th ed., **14** (1995) 954.
7. R. J. Clemens, *Chem. Rev.* **86** (1986) 241 – 318.
8. B. Snider, *Chem. Rev.* **88** (1988) 793 – 811.
9. D. C. Harrowven, S. T. Dennison: "Ketenes, their Cumulene Analogues and their S, Se and Te Analogues", in A. R. Katritzky, O. Meth-Cohn, C. W. Rees (eds.): *Comprehensive Organic Functional Group Transformations,* **vol. 3,** Elsevier, Oxford 1995, pp. 525 – 554, 733 – 856.
10. T. T. Tidwell: "Ketene Chemistry, the Second Golden Age", *Acc. Chem. Res.* **23** (1990) 273 – 279.
11. T. T. Tidwell: *Ketenes,* Wiley, New York 1995.
12. V. H. Agreda, J. R. Zoeller (eds.): *Acetic Acid and its Derivatives,* Marcel Dekker, New York 1993, pp. 161 – 171, 173 – 224.
13. J. A. Wyatt, R. W. Raynolds: "Ketene Cycloadditions", in *Organic Reactions,* **vol. 45,** Wiley, New York 1994, pp. 159 – 646.
14. W. K. Johnson, U. Fink, Y. Sakuma: "CEH Product Review, Ketene/Diketene", in *Chemical Economics Handbook,* SRI International, Menlo Park, CA, 1999, 669.5000 A-M.

Specific References

15. L. Katz, W. N. Lipscomb, *J. Org. Chem.* **17** (1952) 515.
16. C. D. Hurd, A. S. Roe, *J. Am. Chem. Soc.* **61** (1939) 3355 – 3360. A. A. Ponomarev, Y. B. Isaev, *Z. Obsc. Khim.* **22** (1952) 652 – 654; *Chem. Abstr.* **47** (1953) 2695 a.
17. BASF, EP 949 239, 1999(W. Aquila et al.).
18. R. E. Dunbar, G. C. White, *J. Org. Chem.* **23** (1958) 915 – 916.

19. H. Wynberg, E. G. J. Staring, *J. Am. Chem. Soc.* **104** (1982) 166 – 168.
20. H. W. Wang, D. Romo, "Methods for the synthesis of optically active β-lactones", *Tetrahedron* **55**(1999) no. 21, 6403 – 6434.
21. J. C. Sheehan, E. J. Corey, Organic Reactions IX 395 – 399.
22. G. I. Georg (ed.): *The Organic Chemistry of β-Lactams,* Wiley, New York 1993, pp. 295 – 368.
23. H. Ulrich: *Cycloaddition Reactions of Heterocumulenes,* 9th ed., Academic Press, London – New York 1967.
24. T. Machiguchi et al., *J. Am. Chem. Soc.* **121** (1999) 4771 – 4786.
25. Lonza AG, CH 543 474, 1970 (F. Broussard).
26. H. Pommerening, S. Pflaum, *Chim. Oggi* **14**, no. 3/427–31.
27. G. F. Pregaglia, M. Binaghi in N. M. B. Kaler (ed.): *Encyclopedia of Polymer Science and Technology,* "Ketene Polymers," 8th ed., Interscience, New York 1968, pp. 45 – 57.
28. T. Shiori et al., *J. Heterocyclic Chem.* **36** (1999) 1555 – 1563.
29. P. Kocienski, A Pommier, J.-M. Pons, *J. Chem. Soc. Perkin Trans. 1* **1998** 2105 – 2118.
30. Eastman Chemical, US 5 719 315, 5 731 456, 1988 (G. C. Tustin et al.).
31. G. P. Pregaglia, M. Binaghi, *Macromol. Synth.* **3** (1968) 152 – 165.
32. G. Sioli et al. *Chim. Ind.* (*Milan)* **53** (1971) 133 – 139.
33. H. Stage, *Chem. Ztg.* **97** (1973) 67 – 73.
34. Consortium für Elektrochem. Ind., DE 687 065, 1933 (J. Sixt).
35. Consortium für Elektrochem. Ind., DE 734 349, 1934.
36. Wacker, DE 1 079 623, 1963 (T. Altenschöpfer, H. Spes); DE 1 210 809, 1963; DE 1 203 248, 1964.
37. S. Andreades, H. D. Carlson, *Org. Synth. Coll.* **5** (1973) 679 – 684.
38. H. W. Coover, R. C. Hart, *Chem. Eng. Proc.* **78** (1982) 72 – 75; Halcon International, DE-OS 2 441 502, 1974 (C. Hewlett).
39. Wacker, US 3 759 988, 1973 (G. Kunstle, H. Spes).
40. Daicel, US 3 574 728, 1971 (I. Takasu, M. Higuchi, Y. Hijioka).
41. T. Hatori et al., *Chem. Commun.* **2000** 73 – 74.
42. *Kirk-Othmer,* 3rd ed., **1,** 337 – 338.
43. Monsanto, US 3 758 569; US 3 758 571, 1973 (D. Bissing, V. Gash).
44. H. Spes, *Chem. Ing. Tech.* **38** (1966) 955 – 962.
45. B. H. Gwynn, E. F. Degering, *J. Am. Chem. Soc.* **64** (1942) 2216 – 2218. H. J. Hagemeyer, D. C. Hull, *Ind. Eng. Chem.* **41** (1949) 2920 – 2924. Wacker, DE 1 121 605, 1959 (T. Altenschoepfer, E. Enk, F. Knoerr, H. Spes). R. N. Lacey, *Adv. Org. Chem.* **2** (1960) 213 – 263. Hoffmann-La Roche, BE 621 067, 1962 (A. Sturzenegger).
46. University of Iowa, WO 9 414 429, 1994(S. Nissen).
47. H. Wynberg, E. G. J. Staring, *PCT Int. Appl.* **WO 84/01 577** (1984); *NL Appl.* **82/4 070** (1982).
48. Lonza AG, unpublished results.
49. P. Stutte: "The Preparation and uses of Enantiomerically Pure β-Lactones", in A. N. Collins et al. (eds.): *Chirality in Industry,* Wiley, Chichester 1992, pp. 341 – 348.
50. A. Pommier, J. M. Pons, "Recent Advances in β-Lactone Chemistry", *Synthesis* **1993** , 441 – 459.
51. B. W. Dymock, P. J. Kocienski, J. M. Pons, *Synthesis* **1998** , 1655 – 1661.
52. C. Lowe, J. C. Vederas, "Naturally Occurring β-Lactones: Occurrence, Syntheses and Properties", *Org. Prep. Proced. Int.* **27** (1995) 305 – 346.
53. BASF, US 5 359 122, 1994 (M. Huellmann et al.).
54. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* Van Nostrand, New York 1977, pp. 258, 415.
55. G. Hommel: *Handbuch der gefährlichen Güter,* 4th ed., Merkblatt 350, Springer Verlag, Berlin 1987.
56. L. Tenud, M. Weilenmann, E. Dallwjk, *Helv. Chim. Acta* **60** (1977) 975 – 977.
57. Lonza, EP 442 431, 444 563, 1991 (B. Jackson, T. Scholl).
58. Lonza AG, CH 421 080, 1963 (H. Zima).
59. L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis* Wiley, Chichester 1995, 1938 – 1941.
60. Air Reduction Co. Inc., US 2 818 456, 1957 (R. T. Conley, T. F. Rutledge).
61. T. Kato, T. Chiba, *Yuki Gosei Kagaku Kyokaishi* **39** (1981) 733 – 746; *Chem. Abstr.* **95** (1981) 168 405.
62. T. Kato, *Acc. Chem. Res.* **7** (1974) 265.
63. Wacker, US 2 763 664, 1956 (J. Sixt).
64. Takasago, JP 11 049 787, 1999 (Y. Gonda et al.).
65. C. L. Perrin, T. Arrhenius, *J. Am. Chem. Soc.* **100** (1978) 5249 – 5251.

66. Lonza, EP 496 362, 1992 (C. O'Murchu).
67. T. Kato, *Lect. Heterocycl. Chem.* **6** (1982) 105 – 119.
68. J. M. Briody, D. P. N. Satchell, *J. Chem. Soc.* 1965, 3778 – 3785.
69. Lonza AG, CH 642 611, 1979 (M. Gross); US 4 473 508, 1984 (M. Gross).
70. Takasago, JP 11 322 736, 1999(S. Miyano et al.).
71. N. Ogumi, K. Tanaka, H. Ishida, *Synlett.* **1998**, 601 – 602.
72. R. J. Clemens: "The Chemistry of the Diketene Acetone Adduct," *Kodak Lab. Chem. Bull.* **55** (1984) no. 3.
73. M. F. Carrol, A. R. Bader, *J. Am. Chem. Soc.* **75** (1953) 5400 – 5402.
74. A. B. Steele, A. B. Boese, M. F. Dull, *J. Org. Chem.* **14** (1949) 460 – 469.
75. K. Yoshida, Y. Yamashita, *Makromol. Chem.* **100** (1967) 177.
76. W. R. Grace & Co., US 4 588 792, 1986 (Y. Okamoto, E. F. Hwang).
77. Du Pont, US 2 585 537, 1952 (D. D. Coffman).
78. Distillers Co. Ltd., DE 1 043 323, 1956 (R. N. Lacey). Consortium für Elektrochem. Ind., DE 700 218, 1937 (M. Mugdan, J. Sixt). Bayer, DE 832 440, 1949.
79. Lonza AG, GB 852 865, 1960 (W. Moser, R. Perren). BASF, US 3 865 846, 1975 (G. Schulz, Y. Matthias, W. Kasper).
80. Lonza AG, CH 423 754, 1967 (H. Keller).
81. Celanese, US 3 759 955, 1973 (M. L. Jacobs, B. W. Higdon).
82. Lonza AG, US 3 271 420, 1965 (H. Zima); GB 1 015 080, 1965 (H. Zima).
83. US Army, US 3 733 413, 1973 (D. R. Spiner, R. Hoffmann). Allied, FR 1 579 873, 1969 (A. C. Pierce).
84. Esso, GB 927 637, 1963.
85. I.G. Farbenindustrie, GE 717 652, 1939 (F. J. Pohl, W. Schmidt); US 2 351 366, 1944 (F. Pohl, W. Schmidt).
86. O. B. Boese Jr., *Ind. Eng. Chem.* **32** (1940) 16. Lonza AG, US 3 651 130, 1972 (O. Marti, W. Zimmerli, C. Zinstag, H. Keller); US 3 117 186, 1962; GB 903 613, 1962.
87. Standard Oil, US 2 103 505, 1934 (P. J. Wiezerich, A. H. Gleason); US 2 228 452, 1941 (A. H. Gleason).
88. *Kirk-Othmer,* 2nd ed., **vol. 1,** 1963, pp. 153 – 159. Lonza AG, Technical Brochure, "Acetoacetates," 1987.
89. C. D. Hurd, C. D. Kelso, *J. Am. Chem. Soc.* **62** (1940) 1548 – 1549.
90. Lonza AG, EP 61 657, 1984 (L. Tenud, R. Miller, B. Jackson).
91. Lonza AG, EP-A 28 709, 1981 (M. Gross); US 4 473 508, 1984 (M. Gross); Eastman Kodak, US 4 468 356, 1984 (D. E. van Sickle, G. C. Newland).
92. Lonza AG, US 3 950 412, 1976 (K. J. Boosen).
93. Lonza AG, EP 76 378, 1983 (C. Abächerli); EP 76 379, 1985 (E. Greth).
94. Lonza AG, CH 649 294, 1981 (R. Miller, L. Tenud); CH 649 996, 1981 (R. Miller, L. Tenud).
95. Lonza AG, CH 658 056, 1984 (T. Meul, L. Tenud, A. Huwiler).
96. Carbide and Carbon Chemicals Co., US 2 209 683, 1941 (A. B. Boese, Jr.).
97. D. M. Barry et al., *J. Org. Chem.* **62** (1997) 7114 – 7120.
98. G. Jäger, J. Wenzelburger, *Liebigs Ann. Chem.* 1976, 1689 – 1712.
99. Distillers Co. Ltd., US 2 849 456, 1958 (S. J. Branch).
100. R. Lipinski, G. Wolfhard, *Food Sci. Technol.* **48** (1991) 11 – 28.
101. Hoechst, WO 9 319 055, 1993 (G. Roscher et al.).
102. Wacker, US 2 763 664, 1956 (J. Sixt).
103. Armstrong Cork Co., US 4 046 803, 1977 (J. S. Heckles).
104. Takeda Chem. Ind. Ltd., US 4 379 924, 1983 (S. Terao et al.); US 4 421 912, 1983 (I. Ninamida et al.). Fujisawa Pharm. Co. Ltd., US 4 254 260, 1976 (T. Takaya et al.).
105. Lonza AG, US 4 391 979, 1983 (A. Huwiler, L. Tenud); EP-A 34 340, 1981.
106. Lonza AG, Technical Brochure, "Aminothiazolylacetic acid derivatives – The new antibiotic side chains," 1986.
107. Eastman Kodak Company, Technical Brochure, Publication No. N-319 (Oct. 1988).
108. H. Eck, H. Spes, *Method. Chim.* **5** (1975) 493 – 510.
109. H. Bestian, D. Günther, *Angew. Chem.* **75** (1963) 841 – 845.
110. E. C. Taylor, A. McKillop, G. W. Hawks, *Org. Synth. Coll.* **6** (1988) 549 – 551.
111. W. T. Brady, *Synthesis* 1971, 415 – 422.
112. Deutsche Hydrierwerke, DE 885 834, 1937. *Chem. Eng. News* **33** (1955) 1018.
113. C. A. Weisgerber, C. B. Hanford, *Tappi* **43** (1960) no. 12, 178 A.
114. J. W. Dawis, W. H. Roberson, *Tappi* **39** (1956) no. 1, 21.
115. D. H. Dumas, *TAPPI Papermakers Conf. [Proc.]* 1979, 67 – 72.

116. K. T. Hodgson, *Appita J.* **47** (1996) 402 – 406.
117. H. A. Wooster, C. C. Luskbaugh, C. E. Redemann, *Ind. Hyg. Toxicol.* **29** (1947) 56 – 57.
118. J. J. Gomer, *Zentralbl. Arbeitsmed. Arbeitsschutz* **1** (1951) 37 – 39, 58 – 63.
119. Registry of Toxic Effects of Chemical Substances 1981 – 1982, 595 (Nr. OA7700000); 930 (Nr. RQ8225000), NIOSH Cincinnati, OH.
120. DFG, *Maximale Arbeitsplatzkonzentration und biologische Arbeitsstofftoleranzwerte,* VCH Verlagsgesellschaft, Weinheim 1987.
121. SUVA, *Zulässige Werte am Arbeitsplatz,* Schweiz. Unfall Versicherungsanstalt, Luzern 1987.
122. *Documentation of the Threshold Limit Values and Biological Exposure Index,* 5th ed., American Conference of Governmental Industrial Hygienists, Cincinnati, OH 1986.
123. C. P. Carpenter, C. S. Weil, H. F. Smyth Jr., *Toxicol. Appl. Pharmacol.* **28** (1974) 313 – 319.
124. B. L. Van Duuren et al., *J. Natl. Cancer Inst.* **31** (1963) 41 – 55.
125. B. L. Van Duuren et al., *J. Natl. Cancer Inst.* **37** (1966) 825 – 838.
126. B. L. Van Duuren et al., *J. Natl. Cancer Inst.* **39** (1967) 1213 – 1216, 1217 – 1228.

Ketone Resins → Resins, Synthetic

Ketones

Acetone, 2-Butanone, Cyclohexanol and Cyclohexanone are separate keywords.

HARDO SIEGEL, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

MANFRED EGGERSDORFER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 2
2. Methyl Alkyl Ketones 3
2.1. 3-Methyl-2-butanone 3
2.2. 4-Methyl-2-pentanone 4
2.3. 3-Methyl-2-pentanone 5
2.4. 3,3-Dimethyl-2-butanone 6
2.5. 2-Heptanone 6
2.6. 5-Methyl-2-hexanone 7
2.7. 2-Octanone 7
2.8. Higher Methyl Alkyl Ketones 8
3. Dialkyl Ketones 8
3.1. 3-Pentanone 8
3.2. 2,4-Dimethyl-3-pentanone 9
3.3. 2,6-Dimethyl-4-heptanone 9
3.4. 3-Heptanone 10
3.5. 5-Methyl-3-heptanone 10
4. Cyclic Ketones 11
4.1. Cyclopentanone 11
4.2. Cycloheptanone 11
4.3. Higher Cyclic Ketones 11
5. Unsaturated Ketones 12
5.1. 3-Buten-2-one 12
5.2. 3-Methyl-3-buten-2-one 12
5.3. 4-Methyl-4-penten-2-one 13
5.4. 3,3,5-Trimethyl-2-cyclohexen-1-one . 14
6. Diketones 14
6.1. 2,4-Pentanedione 14
6.2. 2,3-Butanedione 16
6.3. Higher α-Diketones 16
7. Aromatic Ketones 17
7.1. Methyl Phenyl Ketone 17
7.2. 1-Phenyl-1-propanone 17
7.3. Diphenyl Ketone 17
7.4. 4,4′-Diphenoxydiphenyl Ketone 18
7.5. 1-Phenyl-2-propanone 18
7.6. 2-Hydroxy-1,2-diphenylethanone . . . 18
8. Toxicology and Occupational Health . 19
9. References 20

1. Introduction

Ketones are organic compounds that contain a carbonyl group and two aliphatic or aromatic substituents.

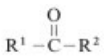

The substituents R^1 and R^2 may be saturated or unsaturated, linear or branched alkyl groups, or aromatic residues. They may be the same or different. Moreover, the alkyl or aryl substituents can also contain heteroatoms. More than one carbonyl group may be present in a particular molecule.

Chemical Properties. The alkyl and aryl groups of ketones act as electron donating groups toward the carbon atom of the carbonyl group. As a result, the reactivity of the carbonyl group is lower than that of, for example, aldehydes. This is also clear from a consideration of the limiting electronic structures with which ketones can be described:

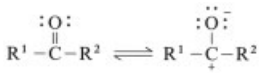

In addition, the reactivity of the carbonyl group is influenced considerably by steric effects, e.g., branching at the α-carbon atom.

Industrially, the most important reactions of ketones are hydrogenation, reductive condensation, reductive amination, oxidation to peroxides, ketal formation, and condensation reactions. The acidic character of the α-hydrogen atoms flanking the carbonyl group enables numerous substitution and addition reactions to take place, e.g., the aldol reaction.

A typical property of ketones is the ability to enolize, this is particularly important for β-diketones (see next page).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_077

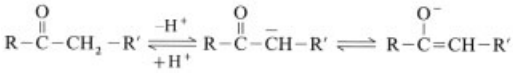

The carbon atom of the carbonyl group is sp^2 hybridized and so the substituents lie in the plane of the carbonyl group. Ketones are therefore prochiral, and enantioselective reactions at the carbonyl group can lead to chiral compounds.

Physical Properties. Aliphatic and cyclic ketones with ten carbon atoms or less are largely stable, colorless liquids that have a pleasant odor. They are miscible with organic solvents and are good solvents for paints, cellulose ethers, nitrocellulose, etc. Some are used as flavors and fragrances (→ Flavors and Fragrances).

Production. As a result of the high stability of ketones, they can be synthesized in a multitude of ways, however only a few processes are practised industrially. Reactions such as condensation of methyl alkyl ketones with compounds containing acidic C−H groups or the ketonization of carboxylic acids or aldehydes are important. The functionalization of an existing carbon skeleton (e.g., by oxidation of alcohols, olefins, or alkanes) plays an important role.

Economic Aspects. Although solvent-poor or solvent-free systems have been developed and solvent recycling is becoming more widespread, ketones have retained their economic importance because they are increasingly replacing other solvents (e.g., chlorinated hydrocarbons) for aromatic compounds.

2. Methyl Alkyl Ketones

For acetone, see → Acetone. For 2-butanone (methyl ethyl ketone), see → 2-Butanone.

2.1. 3-Methyl-2-butanone

3-Methyl-2-butanone [*563-80-4*], methyl isopropyl ketone, $CH_3COCH(CH_3)_2$, $C_5H_{10}O$, M_r 86.13.

Properties. 3-Methyl-2-butanone is a colorless liquid with a characteristic ketone odor. Some physical properties are listed below:

mp	−92 °C
bp	94–95 °C
d_4^{20}	0.803
n_D^{20}	1.3890
Flash point	6 °C
Solubility in water (20 °C)	6.52 wt %

3-Methyl-2-butanone undergoes condensation reactions preferentially at the methyl group. The tertiary carbon atom is easily oxidized and the resulting hydroperoxide (*mp* 26–27 °C) can be produced by base-catalyzed air oxidation [1].

The use of condensation reactions of the methyl group and reactions of the carbonyl group for synthesis have been described.

Production. 3-Methyl-2-butanone is produced industrially by the following reactions:

1) Condensation of 2-butanone [*78-93-3*], formaldehyde, and hydrogen in a single step over palladium [2], or in two steps by condensation on a strongly acidic cation exchanger, isolation of the 2-methyl-1-buten-3-one, followed by hydrogenation [3]:

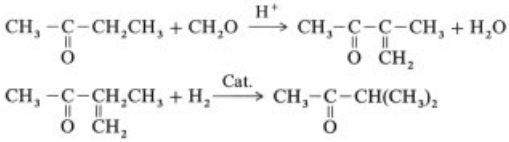

2) By reaction of isobutyraldehyde with acetaldehyde in the gas phase over manganese oxide – aluminum oxide at 450 °C [4]:

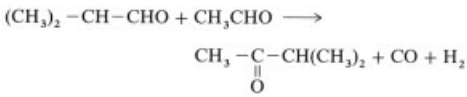

Isobutyric acid and acetic acid may be used in place of isobutyraldehyde and acetaldehyde [5].

3) By addition of water to isoprene at ca. 200 °C in the presence of phosphoric acid on silica [6]:

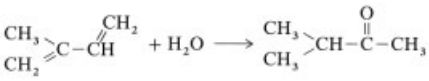

4) By oxidation of 2-methyl-2-butene with oxygen [7] or hydroperoxides in the presence of molybdenum naphthenate at 150 °C [8]:

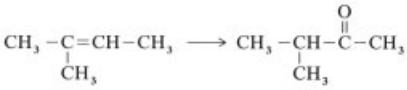

Storage and Transportation. 3-Methyl-2-butanone is available with a purity of > 98 wt %. The following regulations apply to storage and transportation. Labelling according to the Gefahrstoffverordnung (GefStoffV, Federal Republic of Germany) [9]: F; R phrase: 11; S phrases: 9 – 16 – 33. Hazard codes for transportation are as follows [10–12]:

IMDG Code: class 3.2
RID/ADR: class 3, number 36
CFR 49: 172.102 flammable liquid
VbF: A 1
UN number: 2397
EEC number: 5788

Containers of V2A steel are used for transportation and storage.

Uses. 3-Methyl-2-butanone is employed predominantly as an intermediate for the production of pharmaceuticals, herbicides [13], and dye precursors [14]. It is also used in the synthesis of rubber auxiliaries and for the selective extraction of rare earth elements [15].

2.2. 4-Methyl-2-pentanone

4-Methyl-2-pentanone [*108-10-1*], methyl isobutyl ketone, MIBK, $CH_3COCH_2CH(CH_3)_2$, $C_6H_{12}O$, M_r 100.16.

Properties. 4-Methyl-2-pentanone is a colorless liquid with a less pronounced ketone-like odor. Some physical properties are given below:

mp	– 80 °C
bp	116 °C
d_4^{20}	0.8004
n_D^{20}	1.3959
Enthalpy of evaporation	36.15 kJ/mol
Specific heat (20 °C)	1922 kJ/g
Enthalpy of combustion	3740 kJ/mol
Flash point (DIN 51 755)	14 °C
Ignition temperature	475 °C
Ignition class (VED)	G1
Explosion class (VED)	1
Solubility at 25 °C	
4-Methyl-2-pentanone in water	1.7 wt %
Water in 4-methyl-2-pentanone	1.9 wt %

4-Methyl-2-pentanone is poorly soluble in water but is miscible with common organic solvents. It forms an azeotrope with water and with a large number of solvents. For example, the azeotrope with water (*bp* 87.9 °C) contains 75.7 wt % 4-methyl-2-pentanone, and that with *n*-butanol (*bp* 114.4 °C) 70 wt % 4-methyl-2-pentanone.

4-Methyl-2-pentanone undergoes autoxidation to form a peroxide:

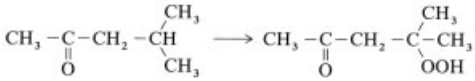

After evaporation of 4-methyl-2-pentanone – water mixtures in the presence of air, a dangerous, increasing concentration of peroxide in the aqueous phase has been reported [16]. The peroxide can also be obtained directly by oxidation with 50 % hydrogen peroxide in the presence of acids [17].

In condensation reactions with carbonyl compounds such as acetone, the α-methyl group usually reacts. Only formaldehyde is able to undergo condensation reactions with the α-methylene group [18]. The usual industrial reactions such as hydrogenation and reductive amination may be carried out with the keto group.

Production. 4-Methyl-2-pentanone can be produced industrially from acetone or isopropyl alcohol according to the following four processes:

1) In three steps from acetone via diacetone alcohol and mesityl oxide, with subsequent hydrogenation.
2) In two steps from acetone via mesityl oxide.
3) In one step from acetone and hydrogen.
4) As a byproduct in the dehydrogenation of isopropyl alcohol to acetone.

In the *three-step process* acetone undergoes alkali-catalyzed condensation in the first step to form diacetone alcohol. Dehydration of the latter to mesityl oxide takes place in the liquid phase at 90 – 130 °C over an acid catalyst, such as phosphoric or sulfuric acid, with high selectivity [19]. Selective hydrogenation of the carbon – carbon double bond in mesityl oxide to give 4-methyl-2-pentanone may be carried out in both the liquid phase and the gas phase with various catalysts, preferably palladium [20]. The selectivity is excellent.

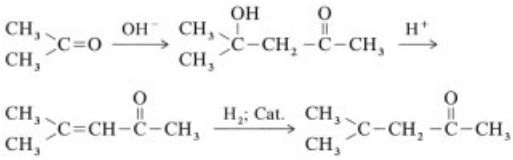

Dehydration and hydrogenation can also be combined into a single process step.

The *two-step process* is less important industrially. Acetone is converted in a single step to mesityl oxide, which is then hydrogenated to 4-methyl-2-pentanone in a second step. Catalysts such as copper chromite [21] or zirconium phosphate [22] are used for the condensation, and palladium on aluminum oxide for the hydrogenation. 4-Methyl-2-pentanol [*108-11-2*] is formed as a byproduct.

For the *single-step production* of 4-methyl-2-pentanone from acetone and hydrogen a combination of a cation exchanger with palladium is generally used as catalyst [23]. The acetone is fed with hydrogen over the palladium-charged catalyst at 130 °C and 0.5 – 5.0 MPa. A selectivity of > 95 % at a conversion of < 50 % is thereby attained. As a result of the increasing production of acetone by dehydrogenation of isopropyl alcohol, the production of 4-methyl-2-pentanone as a byproduct in this process is becoming more important [24] (see also → Acetone, Chap. 4.2.).

Storage and Transportation. 4-Methyl-2-pentanone is available with a purity of > 99 %. Typical specifications are given below:

Content	> 99 wt %
Water content (ASTM D 1364)	0.1 % max.
Hazen color no. (APHA, DIN 53 409)	10 max.
Nonvolatile residue (ASTM D 1353)	0.002 wt % max.
Boiling range (DIN 51 751)	114 – 117 °C
Acid content (as acetic acid)	0.002 wt % max.

Conditions that could lead to autoxidation (presence of air, light, heat, or heavy metals) or condensation (strongly acidic or basic media) should be avoided during storage. Containers of V2A steel are recommended for both storage and transportation.

Transportation is governed by the following regulations: Labelling according to GefStoffV [9]: F; R phrase: 11; S phrases: 9 – 16 – 23 – 33. Hazard codes for transportation are as follows [10]:

IMDG Code: class 3.2
RID/ADR: class 3, number 3 b
CFR 49: 172.101 flammable liquid
UN number: 1245

Uses. The major application of 4-methyl-2-pentanone is as a solvent for vinyl, epoxy, and acrylic resins, for natural resins, and for nitrocellulose. A further use is as a solvent for dyes in the printing industry. 4-Methyl-2-pentanone is also a versatile extracting agent, e.g., for the production of antibiotics, or the removal of paraffins from mineral oil for the production of lubricating oils. The importance of 4-methyl-2-pentanone as an intermediate for synthesis is relatively low. The most important product is 4-methyl-2-pentanol, obtained by hydrogenation of the ketone. 4-Methyl-2-pentanone peroxide has some importance as a polymerization initiator for ethylene [25] and for the hardening of unsaturated polyester resins [26].

Economic Aspects. In terms of production volume, 4-methyl-2-pentanone lies behind acetone and 2-butanone as the most important aliphatic ketone. In 1987 production in Europe was 70 000 t, in the United States 70 000 t, and in Japan 40 000 t. The total produced amounted to 180 000 t. In 1996 production in Western Europe was 100 000 t, in Japan 60 000 t, and in the Federal Republic of Germany 18 000 t. The total produced amounted to 380 000 t [118].

2.3. 3-Methyl-2-pentanone

3-Methyl-2-pentanone [*565-61-7*], methyl *sec*-butyl ketone, $CH_3COCH(CH_3)CH_2CH_3$, $C_6H_{12}O$, M_r 100.16.

Properties. 3-Methyl-2-pentanone is a colorless liquid with an odor reminiscent of peppermint. Some physical properties are listed below:

mp	− 83 °C
bp	116 °C
d_4^{20}	0.8130
n_D^{20}	1.4012
Flash point	12 °C
Solubility in water (20 °C)	2.26 wt %

Production. 2-Butanone undergoes base-catalyzed aldolization with acetaldehyde to form 4-hydroxy-3-methylpentan-2-one [27]. Acid-catalyzed dehydration to 3-methyl-3-penten-2-one [28] is followed by hydrogenation over a palladium catalyst to give 3-methyl-2-pentanone:

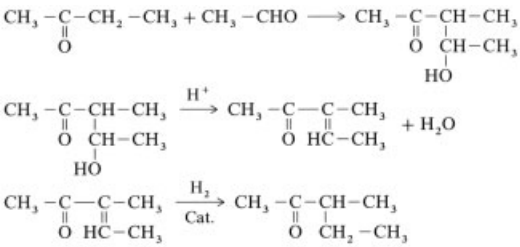

Uses. 3-Methyl-2-pentanone is used as a solvent. Its use as an intermediate for various syntheses has been reported, but the industrial importance is relatively low [29].

2.4. 3,3-Dimethyl-2-butanone

3,3-Dimethyl-2-butanone [*75-97-8*], *tert*-butyl methyl ketone, pinacolone, $CH_3COC(CH_3)_3$, $C_6H_{12}O$, M_r 100.16.

Properties. 3,3-Dimethyl-2-butanone is a colorless liquid with a light peppermint- or camphor-like odor.

mp	− 52.5 °C
bp	106.2 °C
d_4^{20}	0.801
n_D^{20}	1.3964
Flash point	23 °C
Solubility in water (15 °C)	2.44 wt %

3,3-Dimethyl-2-butanone is stable toward autoxidation; under standard conditions it does not form hydroperoxides.

Only the α-methyl group can undergo condensation reactions. At the carbonyl group the usual reactions (hydrogenation, reductive amination) can take place.

Production. 3,3-Dimethyl-2-butanone can be produced by the following routes:

Hydrolysis of 4,4,5-trimethyl-1,3-dioxane, the product of the Prins reaction of isoprene with formaldehyde [30]

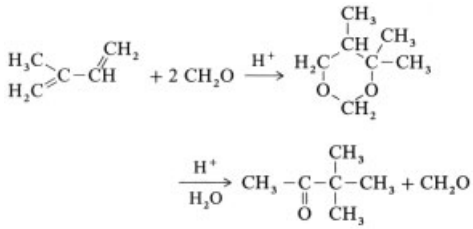

The reductive dimerization of acetone under the influence of an acid catalyst with simultaneous rearrangement (the pinacol rearrangement) [31]

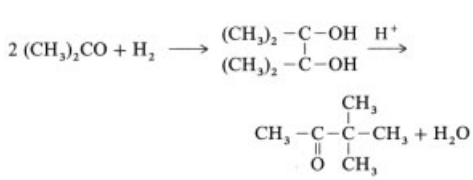

Ketonization of pivalic acid with acetic acid or acetone over thorium, zirconium, or cerium catalysts [32]

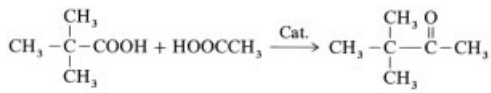

Oxidation of 2,3-dimethyl-2-butene with hydrogen peroxide [33]:

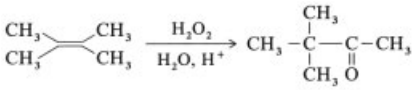

Uses. 3,3-Dimethyl-2-butanone is produced in large amounts for fungicides, e.g., Triademefon (Bayer) [34]. An important intermediate for this type of fungicide is triazolylpinacolone [*58905-32-1*]. A further use is the synthesis of herbicides, e.g., Metribuzin (Bayer) [35].

2.5. 2-Heptanone

2-Heptanone [*110-43-0*], methyl pentyl ketone, methyl amyl ketone,$CH_3COCH_2(CH_2)_3CH_3$, $C_7H_{14}O$, M_r 114.18.

Properties. 2-Heptanone is a colorless liquid with a fruity odor.

mp	−35 °C
bp	149–150 °C
d_4^{20}	0.820
n_D^{20}	1.4085
Flash point	47 °C

Production. 2-Heptanone is produced industrially by reductive condensation of acetone with butyraldehyde in one or two steps [36]:

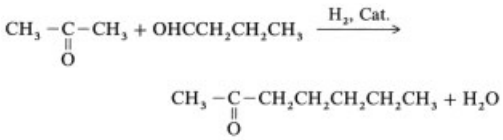

Storage and Transportation. Labelling according to GefStoffV (class II d) [9]: X_n; R phrases: 10–22; S phrase: 23. Hazard classifications are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.101/102 flammable liquid
UN number: 1110

Uses. 2-Heptanone is used as a solvent and as a synthetic building block.

2.6. 5-Methyl-2-hexanone

5-Methyl-2-hexanone [*110-12-3*], methyl isoamyl ketone, $CH_3COCH_2CH_2CH(CH_3)_2$, $C_7H_{14}O$, M_r 114.18.

Properties. 5-Methyl-2-hexanone is a colorless liquid with an odor similar to that of amyl acetate.

mp	−41 °C
bp	145 °C
d_4^{20}	0.888
n_D^{20}	1.4062
Flash point	41 °C

Production. 5-Methyl-2-hexanone is produced by condensation of acetone with isobutyraldehyde [37]. This reaction may be carried out in one or two steps, in the liquid or in the gas phase [38]:

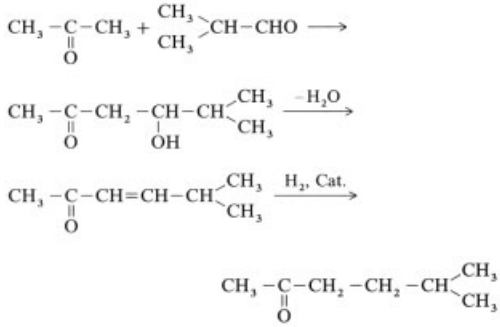

Storage and Transportation. Labelling according to GefStoffV [9]: R phrase: 10; S phrase: 23. Hazard codes for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c
CFR 49: 172.102 flammable liquid
UN number: 2302

Uses. 5-Methyl-2-hexanone is used as a solvent for the production of high-solids coatings, because it possesses a high solubility for alkyd resins, acrylic resins, cellulose esters, cellulose nitrate, and vinylic copolymers.

2.7. 2-Octanone

2-Octanone [*111-13-7*], methyl hexyl ketone, $CH_3COCH_2(CH_2)_4CH_3$, $C_8H_{16}O$, M_r 128.21.

Properties. 2-Octanone is a colorless liquid with an odor reminiscent of apples and a taste similar to that of camphor.

mp	−16 °C
bp	173 °C
d_4^{20}	0.819
n_D^{20}	1.4150
Flash point	62 °C

Production. 2-Octanone is produced industrially by two methods:

1) Reductive condensation of acetone with pentanal [39]:

$$CH_3-\underset{\underset{O}{\|}}{C}-CH_3 + OHCCH_2CH_2CH_2CH_3 \xrightarrow{H_2,\ Cat.} CH_3-\underset{\underset{O}{\|}}{C}-CH_2(CH_2)_4CH_3 + H_2O$$

2) Oxidation of 1-octene with hydrogen peroxide in the presence of a catalyst containing rhenium oxide or a special palladium–copper catalyst in the presence of nitriles [40]:

$$CH_2=CH(CH_2)_5CH_3 \xrightarrow{H_2O_2,\ Cat.} CH_3-\underset{\underset{O}{\|}}{C}-CH_2(CH_2)_4CH_3$$

Uses. 2-Octanone is used as an intermediate and in fragrance compositions.

2.8. Higher Methyl Alkyl Ketones

Properties. The higher straight-chain methyl alkyl ketones (C_9–C_{18}) are colorless liquids or solids, with characteristic odors. The melting point of methyl alkyl ketones with 12 or more carbon atoms lies at room temperature or above. The most important physical properties are listed in Table 1.

Production. Higher methyl *n*-alkyl ketones can be prepared by the following processes:

1) Catalytic gas-phase ketonization of acetic acid with fatty acids [5]:

$$CH_3COOH + RCOOH \longrightarrow CH_3COR + CO_2 + H_2O$$

2) Catalytic gas-phase condensation of acetaldehyde with an *n*-aldehyde [41], [42]:

$$CH_3CHO + RCHO \longrightarrow CH_3COR + CO + H_2$$

3) Catalytic dehydrogenation of secondary alcohols [43]:

$$CH_3-\underset{\underset{OH}{|}}{CH}-CH_2R \xrightarrow{-H_2} CH_3-\underset{\underset{O}{\|}}{C}-CH_2R$$

4) Direct catalytic oxidation of α-olefins [40], [44]:

$$CH_2=CH-CH_2R \xrightarrow[Cat.]{[O_2]} CH_3-\overset{\overset{O}{\|}}{C}-CH_2R$$

5) Oxidation of α-olefins with organic hydroperoxides in the presence of palladium complex catalysts [45]:

$$CH_2{=}CH(CH_2)_9CH_3 + (CH_3)_3COOH \longrightarrow CH_3CO(CH_2)_9CH_3 + (CH_3)_3COH$$

Uses. Higher methyl alkyl ketones are used as nonvolatile solvents for resins and rubber formulations, and also as plasticizers. Condensation products with hexahydropyrimidines [46] or 1,4-phenylenediamine [47] are used in the production of antioxidants and vulcanization accelerators.

3. Dialkyl Ketones

3.1. 3-Pentanone

3-Pentanone [*96-22-0*], diethyl ketone, $CH_3CH_2COCH_2CH_3$, $C_5H_{10}O$, M_r 86.13.

Table 1. Physical properties of higher linear methyl alkyl ketones

Total number of carbon atoms	M_r	*mp*, °C	*bp*, °C (kPa)	Flash point, °C
C_9	142.2	−8 to −9	194 (101.3)	160
C_{10}	156.3	2–3	210 (101.3)	
C_{11}	170.3	11–12	223 (101.3)	192
C_{12}	184.3	20–21	144 (14.5)	
C_{13}	198.3	27–28	120 (6.5)	225
C_{16}	240.4	43	165 (16)	
C_{17}	254.4	45–46	156 (4)	248
C_{18}	268.5	52	251–252 (133)	

Properties. 3-Pentanone is only slightly soluble in water, but miscible with organic solvents.

mp	−40 °C
bp	102 °C
d_4^{20}	0.814
n_D^{20}	1.3924
Flash point	12 °C
Ignition temperature	445 °C

Production. 3-Pentanone can be produced by catalytic ketonization of propionic acid over a thorium oxide or zirconium oxide catalyst at 350–380 °C [48]. Hydroformylation of ethylene in the presence of cobalt carbonyl complexes at 100 °C is a further route [49]:

$$2\,CH_2{=}CH_2 + CO + H_2 \xrightarrow{\text{Cat.}} CH_3CH_2{-}\underset{\underset{O}{\|}}{C}{-}CH_2CH_3$$

Storage and Transportation. The following regulations apply:Labelling according to GefStoffV [9]: F; R phrase: 11; S phrases: 9–16–33. Hazard classifications for transportation are as follows [10–12]:

IMDG Code: class 3.2
RID/ADR: class 3, number 3 b
CFR 49: 172.101 flammable liquid
VbF: A1
UN number: 1156
EEC number: 5789

Uses. 3-Pentanone is used as a solvent for paints and as a starting material for organic syntheses, e.g., for the production of trimethylphenol, a precursor of vitamin E. A further use is in fragrances.

3.2. 2,4-Dimethyl-3-pentanone

2,4-Dimethyl-3-pentanone [*565-80-0*], diisopropyl ketone, $(CH_3)_2CHCOCH(CH_3)_2$, $C_7H_{14}O$, M_r 114.19.

Properties. 2,4-Dimethyl-3-pentanone is a colorless liquid with a characteristic odor similar to that of camphor; it is insoluble in water but soluble in organic solvents.

mp	−69 °C
bp	125 °C
d_4^{20}	0.806
n_D^{20}	1.3990
Flash point	15 °C

Production. 2,4-Dimethyl-3-pentanone is preferably obtained by ketonization of isobutyric acid over a thorium oxide or zirconium oxide catalyst at 430 °C [50].

Storage and Transportation. Labelling according to GefStoffV [9]: F; R phrase: 11; S phrases: 16–23. Hazard classifications for transportation are as follows [10–12]:

RID/ADR: class 3, number 3 b
VbF: A1
EEC number: 5558

Uses. 2,4-Dimethyl-3-pentanone is used as a solvent and as an intermediate.

3.3. 2,6-Dimethyl-4-heptanone

2,6-Dimethyl-4-heptanone [*108-83-8*], diisobutyl ketone, $(CH_3)_2CHCH_2COCH_2CH(CH_3)_2$, $C_9H_{18}O$, M_r 142.24.

Properties. 2,6-Dimethyl-4-heptanone is a colorless liquid with a mild ketone-like odor.

mp	− 49 °C
bp	169 °C
d_4^{20}	0.808
n_D^{20}	1.4114
Flash point	60 °C
Solubility (20 °C)	
2,6-Dimethyl-4-heptanone in water	0.06 wt %
Water in 2,6-dimethyl-4-heptanone	0.45 wt %

Production. This ketone can be prepared by the dehydrogenative condensation of isopropyl alcohol with 4-methyl-2-pentanone as byproduct [51], [52]. A mixture of acetone and isopropyl alcohol may be used instead of the latter alone [53]. 2,6-Dimethyl-4-heptanone can also be produced by the reductive condensation of acetone with 4-methyl-2-pentanone [54]:

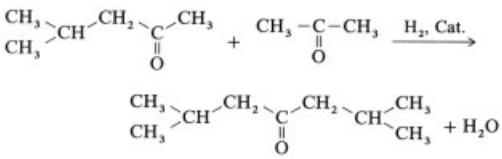

Storage and Transportation. Labelling according to GefStoffV [9]: X_i; R phrases: 10 – 37; S phrase: 24. Hazard classifications for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.101 flammable liquid
UN number: 1157

Uses. 2,6-Dimethyl-4-heptanone is used mainly as a solvent for paints, dyes, and adhesives, and as an extraction agent and solvent for recrystallization of organic compounds.

3.4. 3-Heptanone

3-Heptanone [*106-35-4*], butyl ethyl ketone, $CH_3CH_2COCH_2CH_2CH_2CH_3$, $C_7H_{14}O$, M_r 114.19.

Properties. 3-Heptanone is a colorless liquid with a green odor.

mp	− 39 °C
bp	146 – 149 °C
d_4^{20}	0.818
n_D^{20}	1.4085
Flash point	41 °C

Production. 3-Heptanone is produced industrially by reductive condensation of propanal with 2-butanone [54]:

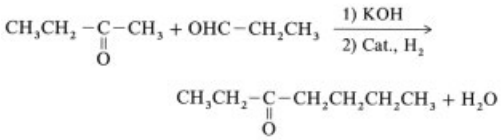

Storage and Transportation. Labelling according to GefStoffV [9]: X_n; R phrases: 10 –22; S phrase: 23. Hazard codes for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.101 flammable liquid
UN number: 1993 (nos)

Uses. 3-Heptanone is used as a perfume and fragrance, as a solvent for cellulose resins, nitrocellulose resins, and vinyl resins, and as a synthetic building block.

3.5. 5-Methyl-3-heptanone

5-Methyl-3-heptanone [*541-85-5*], amyl ethyl ketone, $CH_3CH_2COCH_2CH(CH_3)CH_2CH_3$, $C_8H_{16}O$, M_r 128.22.

Properties. 5-Methyl-3-heptanone is a colorless liquid with low solubility in water (0.3 wt %); water is poorly soluble in the ketone (0.9 wt %). 5-Methyl-3-heptanone is highly soluble in common organic solvents.

mp	< − 70 °C
bp	157 – 162 °C
d_4^{20}	0.823
n_D^{20}	1.4142
Flash point	43 °C

Production. 5-Methyl-3-heptanone is obtained, together with 2-butanone, by the gas-phase dehydrogenation of *sec*-butyl alcohol. The industrial product generally contains ca. 4 % of the isomeric 3,4-dimethyl-2-hexanone [*19550-10-8*].

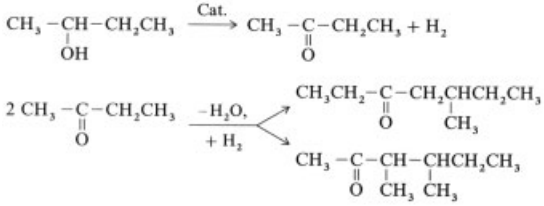

To increase the yield of 5-methyl-3-heptanone, the gas-phase hydrogenation of *sec*-butyl alcohol is carried out in the presence of noble metals on zinc oxide, thorium oxide, or titanium oxide supports [55]. Such catalysts promote the self-condensation of 2-butanone. 5-Methyl-3-heptanone is easily separated from 2-butanone by distillation because of the difference in boiling points.

5-Methyl-3-heptanone can also be obtained by base-catalyzed dimerization of 2-butanone, dehydration of the resulting 5-methyl-5-hydroxy-3-heptanone, and hydrogenation of the ensuing 5-methyl-4-hepten-3-one [56].

Storage and Transportation. Labelling according to GefStoffV [9]: X_i; R phrases: 10–36/37; S phrase: 23. Hazard codes for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.102 flammable liquid
UN number: 2271

Uses. 5-Methyl-3-heptanone is used predominantly as a high-boiling solvent, e.g., for varnish resins.

4. Cyclic Ketones

For cyclohexanone, see → Cyclohexanol and Cyclohexanone.

4.1. Cyclopentanone

Cyclopentanone [*120-92-3*], C_5H_8O, M_r 84.12.

Properties. Cyclopentanone is a colorless liquid with a pleasant, slightly pepperminty odor. It is soluble in water and miscible with common organic solvents.

mp	−51 °C
bp	130–131 °C
d_4^{20}	0.951
n_D^{20}	1.4359
Flash point	30 °C

Production. Cyclopentanone is produced by decarboxylative cyclization of adipic acid [57]. This reaction may be carried out in the liquid phase at 350 °C over zeolites with adipic acid esters [58]. A further production route is oxidation of cyclopentene with oxygen in the presence of palladium (II) chloride or copper (II) chloride as catalyst [59].

Storage and Transportation. Labelling according to GefStoffV [9]: X_i; R phrases: 10–36/38; S phrase: 23. Hazard codes for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.102 flammable liquid
UN number: 2245

Uses. Cyclopentanone is used in fragrances and as an intermediate in organic synthesis, e.g., for the synthesis of pharmaceuticals (such as cyclopentobarbital), jasmones, or cyclopentylamine. The latter is used for fungicides such as Pencycuron (Bayer). The use of cyclopentanone as a special solvent for polycarbonates has been reported [60].

4.2. Cycloheptanone

Cycloheptanone [*502-42-1*], $(CH_2)_6CO$, $C_7H_{12}O$, M_r 112.17.

Properties. Some physical properties are listed below:

mp	−21 °C
bp	179 °C
d_4^{20}	0.951
n_D^{20}	1.4611
Flash point	55 °C

Production. Cycloheptanone is produced by cyclization and decarboxylation of suberic acid (1,8-octanedioic acid) or suberic acid esters in the gas phase at 400–450 °C over alumina doped with zinc oxide or cerium oxide [61].

Uses. Cycloheptanone is a precursor for syntheses, in particular for pharmaceuticals (e.g., Bencyclane, a spasmolytic agent).

4.3. Higher Cyclic Ketones

Higher cyclic ketones include those compounds that are easily accessible from the products of olefin oligomerization over nickel catalysts (see also → Cyclododecanol, Cyclododecanone, and Laurolactam).

Properties. The higher cyclic ketones are colorless solids, each with a characteristic odor. Only cyclodecanone, used as a precursor for laurolactam, possesses significant importance; the others are used as fragrances or as intermediates. Physical properties are listed in Table 2.

5. Unsaturated Ketones

5.1. 3-Buten-2-one

3-Buten-2-one [*78-94-4*], methyl vinyl ketone, $CH_2{=}CHCOCH_3$, C_4H_6O, M_r 70.09.

Properties. 3-Buten-2-one is a colorless liquid with a penetrating odor. The compound is soluble in water and in organic solvents.

bp	81.4 °C
d_4^{20}	0.842
n_D^{20}	1.408
Flash point	−6 °C
Ignition temperature	370 °C

Pure 3-buten-2-one is stable only below 0 °C; it polymerizes rapidly at room temperature. It is stabilized for storage and transportation by the addition of hydroquinone and glacial acetic acid (0.1 wt % each).

A characteristic reaction of 3-buten-2-one is the addition of nucleophiles (e.g., amines [62] or alcohols) at the carbon–carbon double bond. Michael addition to 3-buten-2-one has been developed into a valuable preparative method in terpene and steroid synthesis; this specific reaction is known as Robinson annelation [63]. The dimerization to 2-acetyl-6-methyl-3,4-dihydro-2*H*-pyran is frequently observed as a side reaction in reactions with 3-buten-2-one.

Production. Two synthetic routes are industrially important [64]:

1) The Mannich reaction of diethylamine, acetone, and formaldehyde [65]:

$$CH_3COCH_3 + HCHO + (C_2H_5)_2NH \cdot HCl \xrightarrow{100\,°C} CH_3COCH_2CH_2N(C_2H_5)_2 \cdot HCl \xrightarrow{150-200\,°C} CH_3COCH{=}CH_2 + (C_2H_5)_2NH \cdot HCl$$

The diethylamine hydrochloride released by thermolysis of the Mannich base can be recycled for use in the reaction with acetone and formaldehyde.

2) Condensation of acetone with formaldehyde [66], [67] and subsequent elimination of water:

$$CH_3COCH_3 + HCHO \longrightarrow CH_3COCH_2CH_2OH \xrightarrow[-H_2O]{H^+} CH_3COCH{=}CH_2$$

The water can be eliminated from 4-hydroxy-2-butanone in the liquid or gas phase. Singlestep processes are also known [68].

Storage and Transportation. 3-Buten-2-one can only be stored in stabilized form at room temperature. Hazard codes for transportation are as follows:

IMDG Code: class 3.2
RID/ADR: class 3, number 3 b
CFR 49: 172.101/102 flammable liquid
UN number: 1251

Table 2. Physical properties of higher cyclic ketones

	CAS registry number	Empirical formula	*mp*, °C	*bp*, °C (kPa)	d_4^{20}
Cyclooctanone	[*502-49-8*]	$C_8H_{14}O$	42	74 (1.6)	0.959
Cyclodecanone	[*1502-06-3*]	$C_{10}H_{18}O$	29	107 (1.7)	0.958
Cyclododecanone	[*830-13-7*]	$C_{12}H_{22}O$	61	125 (1.6)	0.906
Cyclotetradecanone	[*3603-99-4*]	$C_{14}H_{26}O$	53	155 (1.6)	

Uses. The Diels – Alder adduct with cyclopentadiene, methyl (5-norbornen-2-yl) ketone, is used as a precursor for pharmaceuticals such as Akineton (Knoll). Derivatives of α-vinyl lactic acid nitrile (from 3-buten-2-one and hydrocyanic acid) are used as fungicides, e.g., Ronilan (BASF).

The most important product based on 3-buten-2-one is 3-methyl-2-penten-4-in-1-ol, a building block in an industrial synthesis of vitamin A.

5.2. 3-Methyl-3-buten-2-one

3-Methyl-3-buten-2-one [*814-78-8*], methyl isopropenyl ketone, C_5H_8O, M_r 84.12.

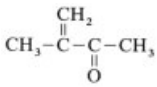

Properties. Methyl isopropenyl ketone is a water-clear, lachrymatory liquid with a pungent odor.

mp	− 54 °C
bp	98 °C
d_4^{20}	0.855
n_D^{20}	1.4236
Flash point	5 °C
Solubility in water (20 °C)	5 wt %

3-Methyl-3-buten-2-one shows the typical properties of an α,β-unsaturated ketone, such as the Michael addition of alcohols and amines, and hetero Diels – Alder reactions. It possesses only limited stability; even at low temperature some cyclodimerization occurs [69].

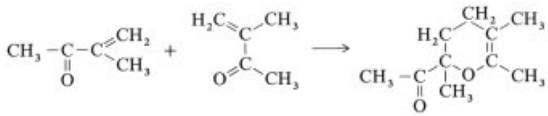

Under the influence of radical initiators or light, methyl isopropenyl ketone polymerizes to poly(methyl isopropyl ketone) [70]. It is able to undergo copolymerization with suitable monomers such as styrene, butadiene, acrylonitrile, or methacrylic acid esters [71].

Production. 3-Methyl-3-buten-2-one is produced by acid-catalyzed condensation of 2-butanone with formaldehyde in the liquid phase [72]. A further possibility is the methylolization of 2-butanone under alkaline conditions followed by elimination of water:

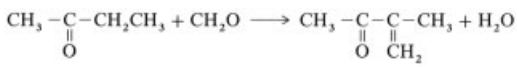

Storage and Transportation. Hazard codes for transportation are as follows:

IMDG Code: class 3.2
RID/ADR: class 3, number 3 b
CFR 49: 172.101 flammable liquid
UN number: 1246

Uses. 3-Methyl-3-buten-2-one is used predominantly as an intermediate in production of 3-methyl-2-butanone, together with a small amount for production of polymers.

5.3. 4-Methyl-4-penten-2-one

4-Methyl-4-penten-2-one [*141-79-7*], mesityl oxide, $(CH_3)_2C{=}CHCOCH_3$, $C_6H_{10}O$, M_r 98.15.

Properties. 4-Methyl-4-penten-2-one is a colorless, oily liquid with an odor reminiscent of peppermint.

mp	−53 °C
bp	129.5 °C
d_4^{20}	0.855
n_D^{20}	1.4458
Flash point	30 °C
Solubility in water (20 °C)	1.07 wt %

Mesityl oxide forms explosive peroxides on prolonged exposure to atmospheric oxygen. Mesityl oxide can be selectively hydrogenated at the olefinic double bond to give 4-methyl-2-pentanone.

Production. 4-Methyl-4-penten-2-one is the precursor of methyl isobutyl ketone (see Section 2.2) and can be produced from acetone in a single- or a two-step process (→ Acetone, Chap. 8.3.). Pure mesityl oxide is obtained by azeotropic distillation of the crude water-containing product. Subsequent purification by removal of the accompanying impurities, such as mesitylene and phorone [*504-20-1*], is effected by distillation [73].

Storage and Transportation. Labelling in accordance with GefStoffV [9]: X_n; R phrases: 10–20/21/22; S phrase: 25. Hazard classifications for transportation are as follows:

IMDG Code: class 3.1
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.101 flammable liquid
UN number: 1229

Uses. 4-Methyl-4-penten-2-one is used to a small extent as an intermediate in synthesis. Most of it is converted to 4-methyl-2-pentanone by selective hydrogenation. Although 4-methyl-2-pentanone has excellent solvent properties for nitrocellulose, cellulose esters and ethers, copolymers of vinyl chloride – vinyl acetate and other synthetic resins, as well as natural and synthetic rubber, its use as a solvent is limited as a result of its toxicity and chemical instability. The fraction of 4-methyl-4-penten-2-one permitted in solvent systems has been set at 5 vol% (Rule 66 of the Los Angeles, California, Air Pollution Control District).

Some of the numerous possible reactions that mesityl oxide may undergo have been used for the synthesis of products such as Budralazine, an antihypertensive agent [74].

5.4. 3,3,5-Trimethyl-2-cyclohexen-1-one

3,3,5-Trimethyl-2-cyclohexen-1-one [*78-59-1*], isophorone, α-isophorone, $C_9H_{14}O$, M_r 138.21.

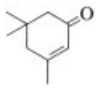

α-Isophorone

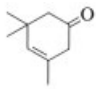

β-Isophorone

Properties. Isophorone is a colorless liquid with a camphor-like odor. It usually contains 1–3 % of the isomeric β-isophorone (3,5,5-trimethyl-3-cyclohexen-1-one). 3,3,5-Trimethyl-2-cyclohexen-1-one is miscible with aromatic and aliphatic hydrocarbons, aldehydes, ketones, alcohols, esters, and ethers.

mp	−8 °C
bp	215.5 °C
d_4^{20}	0.922
n_D^{20}	1.4775
Flash point	95 °C
Ignition temperature	470 °C
Solubility in water (20 °C)	1.2 wt %

Production. Isophorone is produced by condensation of acetone in the liquid phase at ca. 200 °C and 3.6 MPa in the presence of an aqueous potassium hydroxide solution (ca. 0.1 %). The process steps condensation, separation of unreacted acetone, and hydrolysis of byproducts can be carried out in a single reactor [75]. Reaction in the gas phase at 350 °C over calcium – aluminum oxide has also been reported [76].

Storage and Transportation. The following regulations apply for storage and shipping: Labelling according to GefStoffV [9]: X_i (Appendix I.2.1); R phrase: 36/37/38; S phrase: 26; VbF: class A III [11]. Hazard classifications for transportation are as follows [10]:

RID/ADR: class 3, number 32 c.

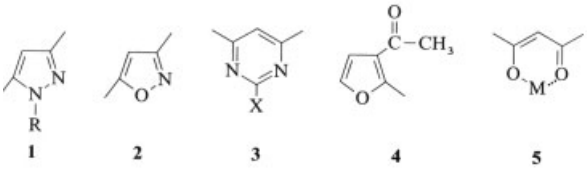

Uses. As a cyclic unsaturated ketone, isophorone can be aromatized to 3,5-dimethylphenol or 2,3,5-trimethylphenol [77].

Hydrogenation leads, depending on conditions, to 3,3,5-trimethylcyclohexanone or 3,3,5-trimethylcyclohexanol, from which trimethyladipic acid is obtained by oxidation [78]. Trimethyladipic acid is used for the production of 2,2,4-trimethylhexamethylenediamine, which is a precursor for polyamides and polyurethanes.

Addition of hydrogen cyanide to the olefinic double bond [79], followed by amination of the keto group in the presence of hydrogen [80], gives 3,5,5-trimethyl-3-aminomethylcyclohexylamine [*2855-13-2*], isophorone diamine. Reaction of the latter with phosgene gives isophorone diisocyanate [*4098-71-9*]. Both compounds are used in large quantities for the production of polymers, mainly polyurethanes. Worldwide production capacity for isophorone is ca. 50 000 t/a.

6. Diketones

6.1. 2,4-Pentanedione

2,4-Pentanedione [*123-54-6*], acetylacetone, $CH_3COCH_2COCH_3$, $C_5H_8O_2$, M_r 100.11.

Properties. 2,4-Pentanedione is a colorless liquid with a mild ketone-like odor; it is completely miscible with common organic solvents.

mp	−23 °C
bp	140.4 °C
d_4^{20}	0.975
n_D^{20}	1.4510
Flash point	34.5 °C
Ignition temperature	350 °C
Solubility (water, 20 °C)	16 wt %

The chemical properties of 2,4-pentanedione are determined by the keto–enol tautomerism that exists in all states of aggregation:

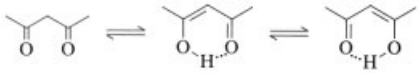

The equilibrium position is influenced by the state of the ketone (i.e., liquid, solid, gas), the temperature, and the solvent. For example, the liquid at 20 °C contains 18 % ketone and 82 % enol.

As a trifunctional compound with two carbonyl groups and an activated methylene group, 2,4-pentanedione is particularly suitable for the synthesis of heterocycles; these may be formed by reaction of both carbonyl groups or of one carbonyl group and the activated methylene group [81]. For example, reaction with hydrazine leads to 3,5-dimethyl pyrazoles (**1**); with hydroxylamine 3,5-dimethylisoxazole (**2**) is formed; with amidines, ureas, thioureas, or guanidines, 2,4-dimethylpyrimidines (**3**) are obtained; with chloroacetaldehyde 2-methyl-3-acetyl furan (**4**) is formed. The best known reaction of 2,4-pentanedione is the formation of metal acetylacetonates (**5**) [82] (see top of page).

Production. 2,4-Pentanedione is produced by thermal or metal-catalyzed rearrangement of isopropenyl acetate [83] (obtained from acetone and ketene [84]):

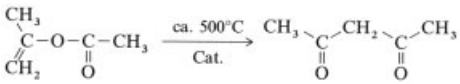

Isopropenyl acetate vapor is fed at atmospheric pressure through a V2A steel tube with an inner temperature of 520 °C. The hot reaction gases are quenched, condensed, and cooled to 20 °C, whereby the gaseous byproducts carbon monoxide, carbon dioxide, methane, and ketene are separated. The product is purified by fractional distillation.

Other industrially less important processes for the production of 2,4-pentanedione, include the Claisen ester condensation of ethyl acetate

with acetone using sodium ethoxide as condensation agent [85] and the acetylation of acetoacetic acid esters with acetic anhydride in the presence of magnesium salts [86].

Storage and Transportation. 2,4-Pentanedione is sold with a purity of > 99 %. In addition to small amounts of acetone, 2,4-pentanedione contains 0.1 % water (max.) and 0.2 % acetic acid.

The preferred material for storage containers is enamel, because in the presence of atmospheric oxygen and water, 2,4-pentanedione can form acetylacetonates with most metals.

Labelling according to GefStoffV [9]: X_n (Appendix IIc); R phrases: 10 – 22; S phrases: 21 – 23 – 24/25. Hazard classifications for transportation are as follows:

IMDG Code: class 3.3
RID/ADR: class 3, number 31 c [10]
CFR 49: 172.102 flammable liquid
UN number: 2310

Containers of aluminum or V2A steel are suitable for transportation, although coloring of the product may occur.

Uses. 2,4-Pentanedione is used as an intermediate, in particular for the synthesis of heterocyclic substances for biologically active compounds and dyes, and for the production of metal acetylacetonates. Furthermore, 2,4-pentanedione is used as a solvent, and as an absorption and extraction agent, particularly for the separation and purification of metal ions. Another application is for the purification of metal-containing wastewater and for corrosion protection. 2,4-Pentanedione is also used as a component of catalyst systems for polymerization, copolymerization, oligo- and dimerization.

Metal Acetylacetonates. The industrial importance of metal acetylacetonates is due to their stability (e.g., in air) and their solubility in organic solvents, whereby homogeneous catalytic or stoichiometric reactions of metal ions with organic compounds may be carried out.

Production. Metal acetylacetonates are usually produced by reaction of soluble metal salts with 2,4-pentanedione in aqueous or ethanolic solution. A base is required to neutralize and release the anion from the metal salt. Special production processes are used for some compounds.

Uses. Metal acetylacetonates are used predominantly for polymerization (e.g., olefins, styrene, methacrylates, acrylonitrile), for cross-linking (e.g., epoxy resins, olefins), or for the hardening of epoxy resins or unsaturated polyester resins.

Metal coatings formed with metal acetylacetonates (e.g., for solar cells [87]) have been described. A further application is their use as a stabilizer in poly(vinyl chloride) for minimizing changes in color [88].

Metal acetylacetonates can also catalyze reactions such as oxidation, isomerization, hydrogenation, dehydrogenation, esterification and transesterification [89]. A review of the uses of metal acetylacetonates has been published [90].

6.2. 2,3-Butanedione

2,3-Butanedione [*431-03-8*], diacetyl, $CH_3COCOCH_3$, $C_4H_6O_2$, M_r 86.09.

Properties. 2,3-Butanedione is a light, yellow – green liquid, which is readily soluble in all important organic solvents.

mp	− 2.5 °C
bp	88 °C
d_4^{20}	0.9831
$n_D^{18.5}$	1.3933
Flash point	26 °C
Solubility in water (15 °C)	20 wt %

In the presence of alkali 2,3-butanedione readily undergoes self condensation to form compounds with a quinonoid structure.

Production. 2,3-Butanedione can be obtained by oxidation of 2-butanone over a copper catalyst at 300 °C in a yield of ca. 60 % [91]. Production from glucose by fermentation has also been described [92], and also the acid-catalyzed condensation of 1-hydroxyacetone (obtained by dehydrogenation of 1,2-propanediol) with formaldehyde [93].

Dehydrogenation of 2,3-butanediol over a copper or silver catalyst in the presence of air is industrially important [94]. 3-Hydroxy-2-butanone (acetoin) is generally formed as a byproduct.

Storage and Transportation. Hazard codes for transportation are as follows:

IMDG Code: class 3.2
RID/ADR: class 3, number 3 b
CFR 49: 172.102 flammable liquid
UN number: 2346

Uses. 2,3-Butanedione is mainly important as a flavor component (buttery taste). It is used in low concentration in ice cream, baked goods, and margarine.

6.3. Higher α-Diketones

2,3-Pentanedione [*600-14-6*], $CH_3COCOCH_2CH_3$, $C_5H_8O_2$, M_r 100.12:

mp	−52 °C
bp	110–112 °C
Flash point	18 °C

2,3-Hexanedione [*3848-24-6*], $CH_3COCOCH_2CH_2CH_3$, $C_6H_{10}O_2$, M_r 114.5:

bp	129–130 °C
Flash point	28 °C
n_D^{20}	1.4137
d_4^{20}	0.934

Higher α-diketones are found in small amounts in various foods.

Production. Processes used include oxidation of the corresponding diols over a copper or silver catalyst [94], or selective oxidation of the methylene group in the corresponding methyl alkyl ketones with oxygen in the presence of catalysts such as nickel [95].

Uses. These compounds are used as aroma components in alcohol-free beverages and in baked goods.

7. Aromatic Ketones

7.1. Methyl Phenyl Ketone

Methyl phenyl ketone [*98-86-2*], acetophenone, $C_6H_5COCH_3$, C_8H_8O, M_r 120.15.

Properties. Methyl phenyl ketone is a colorless solid or liquid with a characteristic odor.

mp	20 °C
bp	202 °C
d_4^{20}	1.028
n_D^{20}	1.5325
Flash point	82 °C
Solubility in water (20 °C)	0.8 wt %

Methyl phenyl ketone is readily soluble in organic solvents. The usual condensation reactions can be carried out at the methyl group.

Production. Most methyl phenyl ketone originates from the Hock process for the production of phenol from isopropylbenzene (→ Phenol); it is isolated from the residue of this process. In addition, acetophenone can be obtained as a main product by selective decomposition of cumene hydroperoxide in the presence of copper catalysts at 100 °C [96]:

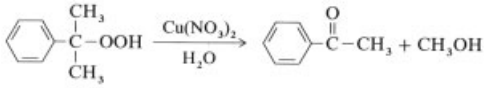

A second possibility is the oxidation of ethylbenzene with air or oxygen at 130 °C and 0.5 MPa. Catalysts used include cobalt salts or manganese salts of naphthenic or fatty acids. Conversion of ethylbenzene is limited to ca. 25 % to minimize the byproducts 1-phenylethanol and benzoic acid [97].

A third method is the Friedel–Crafts acetylation of benzene with acetic anhydride, but this is not of industrial importance.

Uses. Methyl phenyl ketone is mainly converted to resins (synthetic resins) by reaction with formaldehyde [98]. It is also used as a photoinitiator for special printing plates as well as for organic syntheses, particularly for pharmaceuticals. Certain derivatives, such as 1-phenylethanol (obtained by hydrogenation) and its acetate, are used as fragrances. Furthermore, methyl phenyl ketone is used for the synthesis of optical brighteners.

7.2. 1-Phenyl-1-propanone

1-Phenyl-1-propanone [*93-55-0*], ethyl phenyl ketone, propiophenone, $C_6H_5COCH_2CH_3$, $C_9H_{10}O$, M_r 134.12.

Properties. 1-Phenyl-1-propanone is a colorless liquid with a flowery odor, insoluble in water, readily soluble in organic solvents. Typical reactions can be carried out at the methylene group, the carbonyl group, and at the aromatic nucleus.

mp	18 °C
bp	218 °C
d_4^{20}	1.009
n_D^{20}	1.5258
Flash point	87 °C

Production. Propiophenone is produced by Friedel – Crafts acylation of benzene with propionic acid chloride in the presence of an equivalent amount of aluminum chloride [99]. Another industrial method is the catalytic ketonization of benzoic acid with propionic acid over a calcium acetate – aluminum oxide catalyst at 440 – 520 °C [100].

Uses. 1-Phenyl-1-propanone is used mainly as an intermediate for pharmaceuticals such as D-Propoxyphen [101], phenylpropanolamine, and Phenmetrazine [102].

7.3. Diphenyl Ketone

Diphenyl ketone [*119-61-9*], benzophenone, $C_6H_5COC_6H_5$, $C_{13}H_{10}O$, M_r 182.22.

Properties. The stable form of diphenyl ketone consists of colorless rhombic prismatic crystals. Several unstable forms with lower melting points also exist. Diphenyl ketone is insoluble in water and readily soluble in organic solvents.

mp	49 – 51 °C
bp	305 °C
d_4^{50}	1.0869
n_D^{45}	1.5975
Flash point	ca. 155 °C

Production. Benzophenone is usually produced by atmospheric oxidation of diphenylmethane in the presence of metal catalysts such as copper naphthenate [103], [104]. Other processes include Friedel – Crafts acylation of benzene with benzoyl chloride or of benzene with phosgene [105].

Storage. Benzophenone should be stored under cool, dry conditions. A temperature above 45 °C should be avoided (risk of caking).

Uses. Diphenyl ketone and substituted diphenyl ketones in particular are used mainly as photoinitiators for UV-curable printing inks and coatings. Diphenyl ketone is also used as an intermediate for pharmaceuticals and agricultural chemicals. It is employed in the perfume industry as a fixative and as a fragrance with a flowery note.

7.4. 4,4′-Diphenoxydiphenyl Ketone

4,4′-Diphenoxydiphenyl ketone [*14984-21-5*], 4,4′-diphenoxybenzophenone, $C_{25}H_{18}O_3$, M_r 366.42.

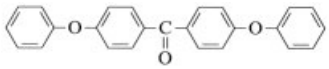

Properties. 4,4′-Diphenoxydiphenyl ketone, *mp* 148 °C, *bp* 270 °C, is a colorless solid, that is insoluble in water and soluble in organic solvents.

Production. 4,4′-Diphenoxydiphenyl ketone is produced from diphenyl ether and phosgene [106] in the presence of aluminum chloride as catalyst. It may also be prepared by the catalyzed atmospheric oxidation of 4,4′-diphenoxydiphenylmethane [107] or the reaction of diphenyl ether, tetrachloromethane, and aluminum chloride [108].

Uses. 4,4′-Diphenoxybenzophenone is used exclusively for the production of high-performance plastics, e.g., by Friedel – Crafts reaction with aromatic dicarboxylic acid chlorides, such as terephthalic acid dichloride [109].

The plastics display a series of industrially interesting properties, such as long-term heat resistance up to 270 °C, fire-extinguishing properties, and high chemical- and solvent resistance. The polymer is worked by injection molding or spinning.

7.5. 1-Phenyl-2-propanone

1-Phenyl-2-propanone [*103-79-7*], phenylacetone, benzyl methyl ketone, $C_6H_5CH_2COCH_3$, $C_9H_{10}O$, M_r 134.18.

Properties. 1-Phenyl-2-propanone is a colorless liquid with a pleasant odor; it is insoluble in water and soluble in organic solvents.

mp	− 15 °C
bp (1.73 kPa)	100 °C
d_4^{20}	1.015
n_D^{20}	1.5158

Production. 1-Phenyl-2-propanone is produced by catalytic ketonization of phenylacetic acid with acetic acid over an alumina-supported cerium oxide catalyst at 400 – 500 °C [110]. It may also be produced by the rearrangement of phenylpropylene oxide on zeolites [111] or oxidation of 2-phenylpropanol.

Uses. 1-Phenyl-2-propanone is used as an intermediate for the synthesis of pharmaceuticals, e.g., Prenylamine, a coronary vasodilator (Hoechst) [112].

7.6. 2-Hydroxy-1,2-diphenylethanone

2-Hydroxy-1,2-diphenylethanone [*579-44-2*], benzoin, α-hydroxybenzyl phenyl ketone, $C_6H_5COCH(OH)C_6H_5$, $C_{14}H_{12}O_2$, M_r 212.25.

Properties. Benzoin is a colorless solid with an odor similar to that of bitter almonds and camphor, *mp* 133 – 134 °C, *bp* 194 °C (1.6 kPa). It occurs in two enantiomeric forms that can be separated by reacting the acid phthalates with (+)-quinidine in methanol to form diastereomic salts [113]:

S (+)-Benzoin [*5928-67-6*], $[\alpha]_D^{12}$ +120.5° (1.2 g/dL in acetone)

R (−)-Benzoin [*5928-66-5*], $[\alpha]_D^{16}$ − 119.2° (2.8 g/dL in acetone)

Production. Benzoin is usually synthesized by reacting benzaldehyde in the presence of cyanide ions [114].

Uses. Benzoin is the starting compound for the synthesis of benzil [*134-81-6*]. Benzil derivatives are used as photoinitiators in polymerization reactions [115]. Benzoin and its alkyl ethers are also used for this purpose [116].

8. Toxicology and Occupational Health

The LD_{50} values of the ketones discussed in 2, 3, 4, 5, 6, 7 are listed in Table 3.

Table 3. LD_{50} values for ketones [117]

Ketone	Animal	LD_{50}, mg/kg
Methyl Alkyl Ketones		
3-Methyl-2-butanone	mouse (oral)	2 572
	rabbit (dermal)	6 350
4-Methyl-2-pentanone	rat (oral)	2 080
	rabbit (inhalation)	8 000
2-Heptanone	mouse (oral)	730
	rat (oral)	1 670
	rabbit (dermal)	12 600
5-Methyl-2-hexanone	rat (oral)	4 760
2-Octanone	mouse (oral)	3 824
Dialkyl Ketones		
3-Pentanone	rat (oral)	2 140
2,6-Dimethyl-4-heptanone	rat (oral)	5 750
	rabbit (dermal)	16 000
Cyclic Ketones		
Cyclopentanone	mouse (subcutaneous)	2 600
Unsaturated Ketones		
3-Buten-2-one	rat (oral)	31
	rat (inhalation, 4 h)	7 *
4-Methyl-4-penten-2-one	mouse (oral)	710
	mouse (inhalation, 2 h)	10 *
3,3,5-Trimethyl-2-cyclohexen-1-one	rat (oral)	2 330
	rabbit (oral)	1 500
Diketones		
2,4-Pentanedione	rat (oral)	1 000
Aromatic Ketones		
Methyl phenyl ketone	rat (oral)	815
1-Phenyl-1-propanone	rat (oral)	4 490
Diphenylketone	mouse (oral)	> 2 500
	rabbit (dermal)	3 500
4,4′-Diphenoxydiphenylketone	rat (oral)	> 5 000
	rabbit (dermal)	> 2 000
2-Hydroxy-1,2-diphenylethanone	rat (oral)	> 6 400

* Values are given in milligrams per cubic meter.

Methyl Alkyl and Dialkyl Ketones. 3-Methyl-2-butanone is moderately irritating to the skin and eye [116], its TLV-TWA value is 200 ppm or 705 mg/m^3. The MAK value for 4-methyl-2-pentanone is 100 ppm or 400 mg/m^3 [28]. The TLV-TWA value is 50 ppm or 205 mg/m^3. 2-Heptanone is classified as harmful to health. Inhalation of the vapor or aerosols containing the liquid should be avoided. Its TLV-TWA value is 50 ppm or 235 mg/m^3. 5-Methyl-2-hexanone has a TLV-TWA value of 50 ppm or 240 mg/m^3. The TLV-TWA value for 3-pentanone is 200 ppm or 705 mg/m^3. 2,6-Dimethyl-4-heptanone is registered according to GefStoffV as irritant; skin contact should therefore be avoided. The MAK value has been fixed at 290 mg/m^3 [28]; the TLV-TWA value is 25 ppm or 250 mg/m^3. 3-Heptanone has a TLV-TWA value of 50 ppm or 230 mg/m^3.

5-Methyl-3-heptanone is classified as an irritant. Inhalation of the vapor or aerosols containing the compound should be avoided. Its TLV-TWA value is 25 ppm or 130 mg/m^3.

Cyclic Ketones. Cyclopentanone is classified as an irritant (GefStoffV). Inhalation of the vapor or aerosols containing the liquid should be avoided.

Unsaturated Ketones. 3-Buten-2-one is a lachrymator and is highly toxic. Exposure of any kind should be avoided [16].

2-Methyl-2-penten-4-one (mesityl oxide) is regarded as harmful to health (class IId, " mindergiftig "), according to GefStoffV. It possesses a strong lachrymatory effect. Its MAK value is 100 mg/m^3 [28] and its TLV-TWA value is 15 ppm or 60 mg/m^3.

3,3,5-Trimethyl-2-cyclohexen-1-one (isophorone) is classified according to GefStoffV as an irritant (irritates the eyes, throat, and skin). Its MAK value is 28 mg/m^3 [28] and its TLV-TWA value is 5 ppm or 25 mg/m^3.

Diketones. 2,4-Pentanedione is classified as harmful to health (GefStoffV, " mindergiftig "). Contact with the skin and with the eye should be avoided.

Aromatic Ketones. Diphenyl ketone (benzophenone) has been employed for many years in the pharmaceutical and cosmetics industries and is regarded as essentially nontoxic. It is included in the FDA list of substances that may be added to foods. Benzoin is dangerous to health when inhaled or absorbed cutaneously.

9. References

1. H. R. Gersmann, A. F. Bickel, *J. Chem. Soc. B* 1971, 2230.
2. BASF, DE-OS 2 941 386, 1979 (G. Heilen, K. Halbritter, W. Gramlich).
3. Rheinpreussen AG, DP 1 198 814, 1963; DP 1 233 848, 1964 (J. Wöllner, F. Engelhardt).
4. Eastman Kodak, BP 1 194 058, 1967 (H. S. Young, J. W. Reynolds).
5. Shell, EP 85996, 1982 (F. Wattimena). BASF, DE-OS 2 758 113, 1977 (H. Fröhlich et al.).
6. Kuraray Co., JA 7 448 407, 1969.
7. Shell, GB 2 167 406, 1984 (B. L. Feringa).
8. Texaco, US 4 000 200, 1971 (J. K. Cox).
9. *Gefahrstoffverordnung (GefStoffV),* BGBl I, Ecomed Verlagsgesellschaft, Landsberg 1986, p. 224.
10. *Gefahrenvermerke für Transport (GGVS),* Verkehr-Verlag J. Fischer, Düsseldorf 1988.
11. *VBF-Richtlinien,* Carl Heymanns Verlag, Köln 1980.
12. *Arbeitsschutz Inland,* UVV 1, Berufsgenossenschaft der Chemischen Industrie, Jedermann Verlag, Heidelberg 1984.
13. American Cyanamid, EP 205 879, 1986 (M. Los, D. W. Loug, G. P. Withers).
14. A. E. Arbuzow, Yu P. Kitaev, *Zh. Obshch. Khim.* **27** (1957) 2328; *Chem. Abstr.* **52** (1958) 7281 h.
15. W. G. Gruzensky, *US Bur. Mines Rep. Invest.* 1961, no. 5910; *Chem. Abstr.* **56** (1962) 8362 a.
16. L. Bretherick, *Chem. Ind. (London)* 1972, no. 22, 863.
17. *Chem. Abstr.* **77** (1972) 100 652 x.
18. J. E. Dubois, R. Luft, *C. R. Hebd. Seances Acad. Sci.* **240** (1955) 1540.
19. *Hydrocarbon Process.* **48** (1969) no. 11, 205.
20. Scholven-Chemie, US 3 361 822, 1968; V. Macho, CS 131 903, 1969.
21. Showa Denko, JP 45-41566, 1970.
22. Tokuyama Soda, JP 45-28566, 1970.
23. Rheinpreussen AG, DT 1 260 454, 1966 (J. Wöllner, W. Neier). VEBA-Chemie, DT 1 643 044, 1967 (J. Disteldorf, W. Flakus, W. Hübel).
24. Showa Denko KK, JP 7 215 334, 1968.

25. BASF, DE-OS 1 943 698, 1969 (H. Groper et al.).
26. Japan Oil and Fats Co., JP 7023.414, 1966.
27. J. E. Dubois, R. Luft, *C. R. Hebd. Seances Acad. Sci.* **238** (1954) 485.
28. R. B. Wagner, *J. Am. Chem. Soc.* **71** (1949) 3214.
29. Zoecon Corp., DE-OS 2 123 147, 1970 (J. B. Sidall). Schering, DE-OS 2 262 402, 1972 (L. Nuesslein, F. Arndt).
30. Mobay, US 4 059 634, 1975 (D. N. Smith Jr.).
31. Phillips Petroleum, US 2 596 212, 1952 (F. E. Gondon). Sumitomo, JP 57142-938, 1981.
32. Daicel Chem. Ind., JP 7 197 237, 1981. Diamond Samrock Corp., DE-OS 2 737 511, 1977 (R. Cryberg, R. Bimber).
33. Bayer, DE 3 211 306, 1982 (G. Rauleder, H. Waldmann).
34. Bayer, DE 2 201 063, 1972 (W. Meisner et al.).
35. Bayer, BE 697 083, 1976.
36. BASF, DE 2 625 540, 1976 (A. Nissen et al.).
37. R. Heilmann et al., *Bull. Soc. Chim. Fr.* 1957, 112.
38. Labofina S.A., FR 2 094 808, 1970 (P. Camerman, J. Hanotien).
39. BASF, DE 2 625 540, 1976 (A. Nissen et al.).
40. Union Oil Comp., US 3 518 285, 1967 (D. M. Fenton, L. G. Wolgemuth); Catalytica Assoc., US 4 738 943, 1986 (J. Vasilevskis, P. L. Ridgway, E. R. Evitt).
41. BASF, DE-OS 2 758 113, 1977 (H. Fröhlich et al.).
42. Eastman Kodak, FR 1 533 651, 1966 (J. W. Reynolds et al.); FR 1 529 019, 1968 (J. W. Reynolds, H. S. Young).
43. Wacker-Chemie, DE-OS 2 358 254, 1973 (L. O. Sommer, F. Knörr); Union Oil Co., US 3 836 553, 1970 (D. M. Fenton).
44. Shell, GB 2 167 406, 1984 (B. L. Feringa).
45. Burmah Oil Trading Ltd., DE-OS 2 403 411, 1973 (P. Field, D. A. Lock).
46. Universal Oil Prod. Co., US 2 787 416, 1964 (H. A. Cyba).
47. Pennwalt Corp., DE-OS 2 414 002, 1973 (I. G. Popoff, P. G. Haines).
48. Daicel, JP 57197-237, 1981. BASF, DE-OS 2 111 722, 1971 (L. Schuster, L. Arnold).
49. BASF, DP 2 445 193, 1974 (N. v. Kutepow).
50. BASF, DP 2 758 113, 1977 (H. Fröhlich et al.).
51. K. Kawamoto et al., *Bull. Chem. Soc. Jpn.* **41** (1968) 932.
52. M. Araki, Y. Kotera, *Bull. Jpn. Pet. Inst.* **15** (1973) no. 1, 45.
53. V. N. Ipatieff, V. Haensel, *J. Org. Chem.* **7** (1942) 195.
54. BASF, DP 2 615 308, 1976 (A. Nissen et al.).
55. H. Miyake et al., *Bull. Jpn. Pet. Inst.* **16** (1974) no. 1.
56. W. Wayne, H. Adkins, *J. Am. Chem. Soc.* **62** (1940) 3401.
57. *Bulletin of the Academy of Sciences of the USSR Div. of Chem. Science* (1968) 609.
58. BASF, DE-OS 3 638 005, 1986 (H. Lermer, W. Hölderich, M. Schwarzmann).
59. Agency of Ind. Sci., DE-OS 3 305 000, 1982 (K. Takehira et al.).
60. *Plast. Eng.* **43** (1987) no. 10, 35.
61. Hüls, DE-OS 2 637 787, 1986 (W. Kleine-Homann).
62. S. I. Suminov, A. N. Kost, *Russ. Chem. Rev. (Engl. Transl.)* **38** (1969) no. 11, 884 – 899.
63. B. P. Mundy, *J. Chem. Educ.* **50** (1973) 110 – 113.
64. G. S. Mironov, M. I. Faberov, *Russ. Chem. Rev. (Engl. Transl.)* **33** (1964) no. 6, 311 – 319.
65. Bayer, DT 877 606, 1953 (G. S. Mironov et al.); *Khim Promst. (Moscow)* **10** (1969) 727 – 729.
66. Bayer, DT 222 551, 1909; Consortium F. Elektrochem. Ind., DT 730 117, 1943 (P. Halbig, A. Treibs); BASF, DT 1 269 121, 1968 (H. Pommer, H. Müller, H. Köhl).
67. BASF, DT 1 277 235, 1968 (H. Müller et al.).
68. Mitsubishi Chem., JP 4 828 211, 1971.
69. J. Colonge, J. Dreux, *C. R. Hebd. Seances Acad. Sci.* **228** (1949) 583.
70. C. S. Marvel et al., *J. Am. Chem. Soc.* **64** (1942) 93.
71. Rheinpreussen AG, DT 1 004 807, 1955 (W. Grimme, F. Engelhardt).
72. Rheinpreussen AG, DT 1 198 814, 1963; DT 1 233 848, 1964 (J. Wöllner, F. Engelhardt).
73. Esso, GB 1 116 037, 1968 (L. J. Sirois, F. J. Hermann, M. E. Oldweiler).
74. K. Keno et al., *Chem. Pharm. Bull.* **24** (1976) 1068.
75. Bergwerksges. Hibernia, DT 1 165 018, 1960 (K. Schmitt, W. Baron, J. Disteldorf).
76. Union Carbide Corp., EP 146927, 1983 (A. J. Papa, S. W. Kaiser).
77. Shell, US 2 369 197, 1945 (De Loss E. Winkler, S. A. Ballard). Shell, US 2 628 985, 1952 (De Loss E. Winkler, H. de V. Finch).
78. Bergwerksges. Hibernia, GB 915 510, 1963.
79. Scholven-Chemie, DT 1 240 854, 1964 (K. Schmitt et al.).
80. Hibernia-Chemie, DT 1 229 078, 1961 (K. Schmitt et al.).
81. A. Heikel, *Chem. Zentralbl.* **1** (1936) 2076.

82. *Houben Weyl,* Oxygen Compounds, part 2, vol. 6/2, 1963.
83. Wacker-Chemie, DT 1 001 249, 1955 (E. Enk, F. Büttner). Wacker-Chemie, DT 2 047 320, 1970 (G. Künstle et al.).
84. Wacker-Chemie, DT 1 121 605, 1959 (E. Enk et al.).
85. H. Adkins, J. L. Reiney, *Org. Synth. Coll.* **III** (1955) 16.
86. UCC, US 2 369 250, 1943 (W. H. Reeder, G. A. Lescisin).
87. Hitachi, DE 3 447 635, 1983 (H. Matsuyama et al.).
88. *Jpn Chem. Week* 11/1984, 3.
89. Merkblatt Wacker-Chemie, Acetonylacetone, München 1983.
90. J. R. Harwood: *Industrial Applications of Organometallic Compounds,* Reinhold Publishing Corp., New York 1963.
91. Mitsubishi Chemical Ind., JP 4 742 711, 1971.
92. *Nature* **157** (1946) 336.
93. DOW, US 2 799 707, 1956 (O. C. Dermer).
94. BASF, DP 2 831 229, 1978 (W. Saner et al.).
95. Mitsubishi Chem. Ind., JP 54 132 515, 1978.
96. Rhone-Poulenc, BE 815 578, 1973.
97. H. J. Sanders, H. F. Keag, H. S. McCulloug, *Ind. Eng. Chem.* **45** (1953) 2 – 14.
98. Hüls, DP 3 324 287, 1983 (J. Dörffel).
99. A. I. Vogel, *J. Chem. Soc.* 1948, 614.
100. Union Carbide Corp., US 4 172 097, 1979 (C. A. Smith, L. F. Theiling).
101. Eli Lilly, US 2 728 779, 1955 (A. Pohland).
102. Boehringer Ingelheim, US 2 835 669, 1958 (O. Thoma).
103. Universal Oil Products Co., US 3 642 906, 1969 (S. Kuhn).
104. Nippon Shokubai Kagaku, JP 59 219 248, 1983.
105. BASF, DT 403 507, 1920 (A. Mittasch, M. Müller-Cunradi).
106. DOW Chemical, US 3 845 015, 1974 (R. J. Thomas).
107. Nippon Shokubai Kagaku, JP 9 148 731, 1983.
108. N. C. Deno, W. L. Evans, *J. Am. Chem. Soc.,* **79** (1957) 5804.
109. Raychem, GB 1 471 171, 1974 (K. J. Dahl, V. Jansons).
110. Diamond Shamrock Corp., BE 857 956, 1976 (R. Cryberg, R. Bimber).
111. BASF, EP 162 387, 1984 (W. Hölderich et al.).
112. *Optical Resolution Procedures for Chemical Compounds,* vol. 3, Optical Resolution Information Center, Manhattan College Riverdale, New York 10471, 1984.
113. J. Solodar, *Tetrahedron Lett.* **2** (1971) 287 – 288. *Org. Synth. Coll.* **I** (1941) 94.
114. M. B. Rubin, *Top. Curr. Chem.* **129** (1985) 1.
115. H. J. Hagermann, *Prog. Org. Chem.* **13** (1985) 123.
116. *Sigma-Aldrich,* Library of Chemical Safety Data, Edition I, 1984.
117. *Registry of Toxic Effects of Chemical Substances 1985 – 86 Edition,* **Vol. 1,** US: Department of Health and Human Services.
118. K. Weissermehl, H.-J. Arpe, *Industrielle Organische Chemie,* Wiley-VCH Weinheim, 1998, 308.

Kieselguhr → Silica
Kopals → Resins, Natural
Krypton → Noble Gases
Krypton Compounds → Noble Gases
Kurchatovium → Transuranium Elements

Lactic Acid

Surinder P. Chahal, Croda Colloids Ltd., Widnes, United Kingdom
John N. Starr, NatureWorks LLC, Minnetonka, MN, USA

1. Introduction 1
2. Properties 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.2.1. Internal Esterification 2
2.2.2. Esterification 2
2.2.3. Other Reactions 3
3. Production 3
3.1. Lactic Acid Fermentation 3
3.2. Lactic Acid Purification 4
3.3. Synthesis of Lactic Acid 4
3.4. Construction Materials 5
4. Uses 5
4.1. Food Uses 5
4.2. Polymers 5
4.3. Industrial Uses 5
5. Quality Specifications 5
6. Analysis 5
7. Storage and Transportation 6
8. Legal Aspects 6
9. Economic Aspects 6
10. Environmental Protection 7
11. Toxicology and Occupational Health 7
12. References 7

Abstract

Lactic acid is a naturally occurring C_3 α-hydroxycarboxylic acid that is a product of anaerobic metabolism of sugars in virtually all living things. It is a colorless to slightly yellow, hydroscopic liquid commercially available in 50–90 wt% solutions. Lactic acid is made industrially by fermentation of carbohydrates or synthetically via the hydrolysis of lactonitrile. Production by fermentation has increased over the last decade as it produces the highly enantiopure L-lactic acid that is needed for food and polymer applications. Historically, foods have been the largest volume application, as lactic acid is a mild-tasting acidulant and food preservative. The use of lactic acid based polymers has increased significantly since ca. 2002. These polymers have the potential to develop into a million-tonne industry over the next 20 years. The commercial lactic acid processes use a variety of sugar for raw materials (sucrose from sugar beets and sugar cane, dextrose from corn), depending on price and availability of these sugars near the plant location. There are also a number of purification processes that are used commercially (lactate ester distillation, distillation, liquid–liquid extraction).

1. Introduction

Lactic acid, 2-hydroxypropionic acid, was discovered in 1780 by the Swedish chemist Scheele [1]. He isolated it from sour milk as an impure brownish syrup. The French chemist Braconnot [2] established that lactic acid could be made by fermentation processes. Industrial manufacture of lactic acid was established in 1881 in the USA. The history of industrial production of lactic acid has been written by Benninga [22]. The scientific literature on lactic acid is extensive due to its important role in metabolic pathways. Ref. [9] describes the properties and chemistry of lactic acid and is an excellent source of information on lactic acid.

2. Properties

2.1. Physical Properties

Lactic acid, 2-hydroxypropionic acid [*50-21-5*], $CH_3CH(OH)COOH$, M_r 90.08, is the simplest hydroxycarboxylic acid with an asymmetrical carbon atom. It occurs as a racemate (DL) and in two optically active forms, L-(+)-lactic acid [*79-33-4*] and D-(−)-lactic acid [*10326-41-7*].

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_097.pub2

L (+) Lactic acid D (−) Lactic acid

Pure and anhydrous racemic lactic acid is a white crystalline solid with a low melting point [3]. It is difficult to determine its melting point exactly because of the tremendous difficulty in producing anhydrous lactic acid; values range from 18 [3] to 33 °C [4]. Lactic acid undergoes internal esterification to produce lactoyllactic acid (see Section 2.2.1) [5]. If one of the optical isomers is present in excess in lactic acid, this isomer can be isolated by fractional crystallization of the freshly distilled acid from a mixture of equal parts of diethyl ether and diisopropyl ether [6, 7]. The melting point of the pure optical isomers is 52.7 – 52.8 °C. The boiling point of anhydrous lactic acid is approximately 125 – 140 °C at 27 kPa [8].

Because of the very hygroscopic nature of lactic acid, it is normally obtained as a concentrated solution, up to 90 wt %, which is colorless, virtually odorless, and contains a significant amount of lactoyllactic acid and other lactic acid oligomers. It is completely soluble in water, ethanol, diethyl ether, and other organic solvents which are miscible with water. It is virtually insoluble in benzene and chloroform. The distribution of lactic acid between water and some organic solvents can be found in [9].

Synthetically produced lactic acid has historically been optically inactive, i.e., a racemic mixture. Lactic acid produced by fermentation is optically active, and either isomer can be produced depending on the nature of the bacteria. Density [11, 12], viscosity [10, 12], refractive index [12], and conductivity [13] of aqueous lactic acid solutions of varying concentrations are given in Table 1. The dissociation constant K of lactic acid is 1.38×10^{-4} at 25 °C which corresponds to a pK of 3.86. The pH of a 10 wt % aqueous solution of lactic acid is 1.75 [12].

2.2. Chemical Properties

2.2.1. Internal Esterification

Lactic acid, being both an alcohol and an ester, can undergo intermolecular esterification to produce an ester, lactoyllactic acid [*617-57-2*].

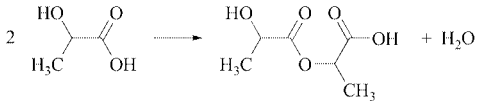

Lactic acid Lactoyllactic acid

Another bimolecular ester which can form is 3,6-dimethyl-1,4-dioxane-2,5-dione or lactide [*95-96-5*], a cyclic ester. Lactide has three different forms, L,L-lactide [*4511-42-6*], D,D-lactide [*13076-17-0*], and D,L-lactide [*13076-19-2*], depending on the chirality of the two lactic acid entities making up the lactide molecule.

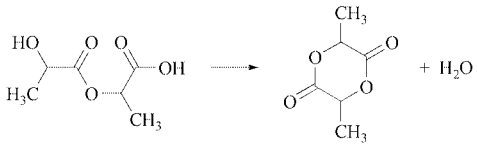

Lactoyllactic acid Lactide

On further condensation of lactoyllactic acid, polylactic acids can be formed. These are linear polyesters of undefined chain length. Identical chemical structures can also be produced by the ring-opening polymerization of lactide, and in this case the term polylactide is normally used.

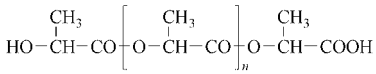

Polylactic acid

Commercially available aqueous lactic acid solutions contain varying amounts of lactoyllactic and polylactic acids, depending on the concentration and age of the solution. Solutions as weak as 6.5 wt % contain ca. 0.2 wt % lactoyllactic acid [12]; an 88 wt % lactic solution contains < 60 % free lactic acid [12], and 100 % lactic acid contains only 32 % of the free monomeric acid [16].

2.2.2. Esterification

Synthesis of the lactates of lower alcohols occurs relatively easily by direct reaction with the corresponding alcohol. The yield can be increased by removing the water formed by azeotropic distillation. Lactate esters can also be made by the reaction of alcohols with polylactic acid to produce minimal water [23].

Table 1. Physical data of aqueous lactic acid solutions at 25 °C

Concentration, wt %	Density, g/mL	Viscosity, mPa·s	Refractive index	Conductivity, mS/cm
6.29	1.0115	1.042	1.3390	3.670
25.02	1.0570	1.725	1.3586	3.823
54.94	1.1302	4.68	1.3909	1.530
88.60	1.2006	36.9	1.4244	0.0567

2.2.3. Other Reactions

Lactic acid has the potential to be a low cost feedstock for two important chemicals: propylene glycol (1,2-propanediol) and acrylic acid. Propylene glycol is produced by reducing lactic acid [24]. Propylene glycol production is over 500 000 t worldwide. Acrylic acid is produced by the dehydration of lactic acid [25 – 27]. Acrylic acid represents a market of more than 3 000 000 t worldwide.

Many more different reactions of lactic acid can be found in the literature. Lactic acid is also of great biological importance [9].

3. Production

Lactic acid is produced on an industrial scale by fermentation or a synthetic method. Since ca. 1995, all new lactic acid production capacity has employed the fermentation approach because it produces lactic acid with high chiral purity that is needed for the food and polymer markets.

3.1. Lactic Acid Fermentation [17 – 20]

The fermentation process requires carbohydrates, nutrients, and a microorganism to produce lactic acid via fermentation. The carbohydrates used in fermentation consist predominantly of hexoses or compounds which can be easily split into hexoses, e.g., dextrose, corn syrups, molasses, sugar beet juice, whey [17], as well as rice, wheat, corn, and potato starches. Lignocellulosics have the potential to become a low-cost source of carbohydrates, and the use of pentoses (xylose, arabinose) for lactic acid production may increase as lignocellulosic hydrolysis technology matures [28, 29].

The nutrients required by the microorganisms include soluble peptides and amino acids, phosphates and ammonium salts, and vitamins. In many cases, the peptides and amino acids are a complex nitrogen source such as yeast extract paste, corn steep liquor, corn gluten meal, malt sprouts, soy peptone, and meat peptone. Only a minimal amount of these complex nitrogen sources are used in order to simplify purification of the lactic acid.

During fermentation, the pH of the broth must be controlled between 5.0 and 6.5. Lime (calcium hydroxide), calcium carbonate, ammonium hydroxide, and sodium hydroxide are typically used to neutralize the lactic acid made in the broth to maintain constant pH. Thus, calcium lactate, ammonium lactate, or sodium lactate salts are formed in the fermentation broth.

Many different microorganisms make lactic acid and can be used in high volume production. *Lactobacillus*, *Bacillus*, and *Rhizopus* strains can be used in the conversion of carbohydrates to lactic acid. The Rhizopus fungal strains have the advantage that they do not require a complex nitrogen source, but tend to give lower yields of lactic acid than the bacterial strains. The bacterial strains perform anaerobic fermentations, and the *Rhizopus* strains aerobic fermentations.

Industrial-scale production of lactic acid requires large amounts of bacterial inoculum which must be cultured in an appropriate seed train. Contaminating microorganisms must be severely limited in the seed train and production vessels to achieve high yields and chiral purities.

Lactic acid yields are between 85 and 95 % based on fermentable sugars. Typical fermentation byproducts, such as formic acid and acetic acid, are found in concentrations of less than 0.5 wt %. "Homofermentive" bacterial strains are typically used as they make the least amount of byproducts.

Recent developments in lactic acid fermentation are moving towards the use of genetically engineered yeast strains to produce lactic acid under neutral or low-pH conditions [30 – 32].

These microorganisms have the advantage of using a defined media and the ability to produce lactic acid, not lactate salts, in the broth. The costs of the neutralizing agent (lime), the acidifying agent (sulfuric acid), and the disposal of the byproduct salt (gypsum) are virtually eliminated by low-pH fermentation.

3.2. Lactic Acid Purification

After fermentation, the lactic acid broth needs to be purified for its intended use. For certain food applications, purification may be minimal. For other applications, particularly as monomer for polylactide production, the purification requirements are considerable. The choices of nutrients and microorganism for fermentation have a significant role in purification decisions. The complex nitrogen source and byproducts from the microorganisms are sources of impurities that must be separated. Unit operations to remove residual proteins, amino acids, sugars, cations, anions, and other organic acids are needed.

In all cases, the microorganisms must be removed from the broth. A variety of technologies can be used, depending on which microorganism was used in the fermentation. Bacterial strains can be removed by flocculation under alkaline conditions [33] or ultrafiltration [34]. *Rhizopus* cell morphology must be controlled carefully in the fermenter in order to easily separate it on a filter.

The lactate salts formed in the fermentation must be converted to lactic acid. There are a number of methods to do this, but addition of sulfuric acid to calcium lactate is routinely used in large-scale production. This method forms a low-value byproduct salt, calcium sulfate or gypsum, which has low solubility in water.

Considerable effort has gone into using other processes that would eliminate the production of a byproduct salt, such as water-splitting electrodialysis [35 – 37] and the use of carbon dioxide in conjunction with liquid – liquid extraction [38].

For some food applications, the broth must be passed through activated carbon and ion exchange, and then concentrated to meet the needs of the customer.

Further purification is needed to produce lactic acid for pharmaceutical, polymer, and other applications. A number of large-scale purification processes are used for production.

A purification process of interest is liquid – liquid extraction with long-chain tertiary amines in the organic phase [21, 39 – 41]. In this case the lactic acid is extracted into the organic phase at a relatively low temperature, and then backextracted into water at high temperature to produce a high-quality lactic acid stream. The selectivity for lactic acid over residual sugars and proteins can be high due to acid – amine complexation.

Distillation of lactic acid is possible with proper operating conditions to minimize the formation of lactic acid oligomers that would limit yield [42 – 44]. The distillate of a lactate ester, such as ethyl lactate, can produce a high quality lactic acid stream but has the additional complexity of the esterification and hydrolysis reactions [45 – 47]. The ethyl lactate distillation route could be advantageous for companies also interested in selling ethyl lactate or other lactate esters.

There has also been considerable effort in the purification of lactic acid streams using membranes with electrodialysis [48, 49], and anion exchange [50 – 52], and these operations may find a role in future production plants.

3.3. Synthesis of Lactic Acid

Since the 1960s, racemic lactic acid has also been produced via synthetic routes. The present industrial synthesis of lactic acid is based on the reaction of acetaldehyde with hydrogen cyanide followed by the hydrolysis of the resultant lactonitrile:

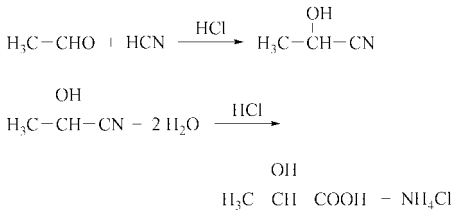

Musashino is the last major synthetic lactic acid producer. If a synthetic route were able to produce a chirally pure lactic acid through an enantioselective catalyst at a low cost, this synthetic route could replace some fermentation capacity.

3.4. Construction Materials

Lactic acid is corrosive toward metallic materials normally used in production and processing, especially at high temperatures. Hastelloy, titanium, zirconium, tantalum, niobium, chromium, and molybdenum are resistant to corrosion. The corrosive properties of lactic acid also depend on its concentration; thus, initial evaporation of lactic liquors can normally take place in 316 stainless steel evaporators, but final evaporation to 80 % lactic acid normally takes place in titanium-lined evaporators.

Lactic acid fermentations are carried out at pH 5.0 – 6.5 where corrosion is minimal, but stainless steel fermenters are mainly used for ease of sterilization.

4. Uses

Lactic acid and some of its derivatives (salts and esters) are used in three main applications: foods, polymers, and industrial.

4.1. Food Uses

The food industry is the traditional large consumer of lactic acid and lactate salts. It is mainly used as an acidulant and preservative. Lactic acid has a mild acid taste and does not overpower weaker aromatic flavors. Since it occurs naturally, it does not introduce a foreign element into the food, and its salts are very soluble, thus giving the possibility of partially replacing the acid in buffering systems. Lactic acid and its salts are used in a variety of beverages, candies, meat, and sauces.

Calcium lactate can be used in a variety of foods to make a calcium-enhanced product. Steroyl lactate and its sodium and calcium salts are used in bread.

4.2. Polymers

The fastest growing use for lactic acid is its use as a monomer for the production of polylactic acid or polylactide (PLA). NatureWorks LLC produces PLA via ring-opening polymerization of lactide [53]. Lactide is made by condensation of two lactic acid entities. Applications for PLA include containers for the food and beverage industries, films and rigid containers for packaging, and serviceware (cups, plates, utensils). The PLA polymer can also be spun into fibers and used in apparel, fiberfill (pillows, comforters), carpet, and nonwoven applications such as wipes.

4.3. Industrial Uses

There are a variety of industrial uses for lactic acid and its derivatives. It is used in metal plating, cosmetics, and the textile and leather industry. Lactate esters are used in the manufacture of paints and inks, electronics, and metal cleaning. Agricultural uses include animal feed.

5. Quality Specifications

Pharmaceutical- and edible-grade lactic acids are considered to be of most importance for external sale of lactic acids. Although quality specifications for pharmaceutical-grade lactic acid are laid down in the national pharmacopia in many countries, there is still no defined specification for edible-grade lactic acid. In the USA the specification for edible-grade lactic acid is laid down in the Food Chemicals Codex (FCC). Table 2 compares specifications for pharmaceutical-grade, FFC, and typical edible lactic acid.

6. Analysis

Quantitative analysis of lactic acid can be carried out by titration against sodium hydroxide with phenolphthalein as indicator. This method is only accurate for very weak lactic acid solutions. Stronger solutions contain lactoyllactic acid which must be previously hydrolyzed to lactic acid. This is done by boiling the lactic acid in a measured excess of sodium hydroxide for a few minutes and backtitrating with hydrochloric acid and phenolphthalein as indicator.

A HPLC method is often used in for quantitative analysis in both research and production environments. This method is accurate and can

Table 2. Specifications of lactic acid

	Pharmaceutical grade	FCC (July 1996)	Typical edible grade
Assay, %	88.0	95 – 105[a]	80
Identification		positive test for lactate	
Chloride, %	< 0.008	< 0.1	< 0.02
Citric, oxalic, phosphoric, or tartaric acid		passes test	
Cyanide, ppm		< 5	
Sulfate, %	< 0.02	< 0.25	< 0.05
Arsenic, ppm	< 4	N.A.	< 0.2
Heavy metals, ppm	< 33	< 10	< 10
Iron, ppm	< 10	< 10	< 10
Lead, ppm		< 5	
Ash, %	< 0.1	< 0.1	< 0.1
Sugars		passes test	
Calcium, %	< 0.02		

[a] Of specified concentration.

also quantitatively measure the presence of other entities, e.g., sugars and other organic acids.

To determine the L-(+)- and D-(−)-lactic acid content of a solution, enzymatic methods are used. The enzymes and analyzers are commercially available together with lactic acid solutions of very high individual isomeric purity. Chiral HPLC columns and GC columns are also available to determine chiral purity.

7. Storage and Transportation

Lactic acid is stored in glass, plastic, or stainless steel tanks and is transported and supplied in well-sealed plastic containers.

8. Legal Aspects

Lactic acid is permitted as a food acid by legislation in most countries. In Europe it must conform to the specific criteria laid down in EC Council Directive 96/77/EC and FCC. In the United States, it has GRAS (generally recognized as safe) status.

9. Economic Aspects

Worldwide capacity of lactic acid is about 450 000 t. Table 3 lists major lactic acid producers, plant locations, and capacities. Actual production data are not reported, as the producers are either privately owned or a subsidiary of a much larger company. There are a number of smaller (< 15 000 t) lactic acid producers in China and India. The rise in PLA production will have a significant effect on the lactic acid market. For instance, the scale of lactic acid production plants has increased dramatically. The 180 000 t/a. NatureWorks plant is three to four times larger than the next largest lactic acid plant. Large economies of scale are required to lower the cost of PLA to allow it to compete against incumbent polymers such as poly(ethylene terephthalate), polystyrene, and polypropylene. Large lactic acid plants with lower production costs due to beneficial economies of scale could significantly disrupt the current lactic acid supply and demand balance. PLA will soon become the single largest use for lactic acid, if it is not already.

The recent increase and volatility in oil price (> $ 60/barrel in 2005) has shown the need to move away from oil-based raw materials to the use of annually renewable raw materials where economically viable. PLA pricing is already competitive with many of the incumbent materials and will become more advantageous as oil production peaks in the next 15 – 30 years. The increasing demand for PLA will create a significantly larger lactic acid production base. The Chinese government has recently been supportive of the emerging PLA industry and has

Table 3. Major lactic acid producers

Company	Plant location	Approx. capacity, 10^3 t/a	Comments
NatureWorks LLC	Nebraska, USA	180	For PLA manufacture. Subsidiary of Cargill Inc.
Purac	Netherlands	30 – 50	Subsidiary of CSM
	Spain	20 – 30	
	Brazil	15 – 20	
PGLA-1	Nebraska, USA	35	Joint venture between Purac and Cargill Inc.
Galactic	Belgium	20	Subsidiary of Finesucre s.a.
Anhui BBCA & GALACTIC Lactic Acid Company Limited (B& G)	China	30	Joint venture between Galactic and BBCA Biochemical Co. Ltd.
Archer Daniels Midland (ADM)	Illinois, USA	20 – 30	
Musashino	Japan	15	Synthetic lactic acid.

envisioned an annual PLA capacity of millions of tonnes by 2025.

10. Environmental Protection

Lactic acid, a naturally occurring organic acid found in many different biological systems, is environmentally safe. Industrial fermentation of lactic acid produces a salt byproduct that is landfilled or incorporated into building materials or used in agriculture. It also produces an aqueous stream that can be treated in a typical wastewater treatment plant.

As more lactic acid goes into making PLA, there are considerable environmental benefits of using PLA over conventional oil-based polymers, including lower CO_2 emissions and wider choice of disposal options that includes composting.

11. Toxicology and Occupational Health

Lactic acid is a nontoxic naturally occurring edible acid used widely in the food industry as an acidulant. However, it is acidic and will cause discomfort if it should come into contact with eyes or broken skin. It can cause eye irritation and must be washed out immediately. The problem of oral toxicity should not arise unless a large quantity is swallowed (LD_{50}: 3730 mg/kg in rats).

12. References

1. C. W. Scheele, Om Mjolk: *Kgl.* Vetenskaps-Academiens nya Handlingar, 1, Stockholm 1780, pp. 116 –124.
2. H. Braconnot, *Ann. Chim. (Paris)* **86** (1813) 84 – 100.
3. F. Krafft, W. A. Dijes, *Ber. Deut. Chem. Ges.* **28** (1895) 2589 – 2597.
4. L. B. Lockwood, D. E. Yoder, H. Zienty, *Ann. N. Y. Acad. Sci.* **119** (1965) 854.
5. M. Brin, "Lactic acid — some definitions," *Ann. N. Y. Acad. Sci.* **119** (1965) 1084 – 1090.
6. H. Borsook, H. M. Huffman, Y. P. Liv: "The Preparations of Crystalline Lactic Acid," *J. Biol. Chem.* **102** (1933) 449 – 460.
7. H. A. Krebs, *Biochem. Prep.* **8** (1961) 75 – 79.
8. D. Vorlander, R. Walter, *Z. Phys. Chem.* **118** (1925) 1 – 30.
9. C. H. Holten: *Lactic Acid*, Verlag Chemie, Weinheim 1971.
10. A. Sepitka, *Prum. Potravin* **13** (1962) 385 – 388, 605 –609; **14** (1963) 45 – 49, 82 – 85.
11. R. A. Troupe, W. L. Aspy, P. R. Schrodt, *Ind. Eng. Chem.* **43** (1951) 1143 – 1146.
12. W. Vidler, *Personal communication to Stichting Ilra*, 1959.
13. A. W. Martin, H. V. Tartar, *J. Am. Chem. Soc.* **59** (1937) 2672 – 2675.
14. G. Saville, H. A. Gundry, *Trans. Faraday Soc.* **55** (1959) 2036 – 2038.
15. O. Meyerhof, *Biochem Z.* **129** (1922) 594 – 604.
16. S. Bezzi, L. Riccoboni, C. Sullam, *Reale arrad. d'Italia, Mem. cl Sci. Fis. Mat. nat.* **8** (1937) 181 – 200.
17. G. C. Cox, R. D. Macbean, *Aust. J. Dairy Technol.* **32** (1977) no. 1, 19 – 22.

18. M. A. Cousin, E. H. Marth, *J. Food Prot.* **40** (1977) no. 6, 406 – 410.
19. M. A. Cousin, E. H. Marth, *J. Food Prot.* **40** (1977) no. 7, 475 – 479.
20. E. S. Mil'ko, O. V. Sperehip, I. L. Rabotnova, *Z. Allg. Mikrobiol.* **6** (1967) no. 4, 297 – 301.
21. A. M. Baniel et al., US 4275234, 1981.
22. H. Benninga: *A History of Lactic Acid Making*, Kluwer Academic Publishers, Dordrecht, 1990.
23. E. I. DuPont de Nemours and Company, US 5264617, 1991 (L. D. Brake).
24. Wisconsin Alumni Research Foundation, US 6441241, 2001 (R. D. Cortright, J. A. Dumesic).
25. Battelle Memorial Institute, US 5252473, 1991 (P. C. Walkup, C. A. Rohrmann, R. T. Hallen, D. E. Eakin).
26. C. T. Lira, P. J. McCrackin, *Ind. Eng. Chem. Res.* **32** (1993) 2608–2613.
27. H. Danner, M. Urmos, M. Gartner, R. Braun, *Appl. Biochem. Biotechnol.* 70–72 (1998) 887–894.
28. Purac Biochem B.V., US Patent Application 20040203122, 2004 (R. Otto).
29. Cargill Dow LLC, WO 2004 099381, 2004 (V. Rajgarhia, K. Koivuranta, M. Penttilae, M. Ilmen, P. Suominen, A. Aristidou, C. Miller, S. Olson, L. Ruohonen).
30. Purac Biochem B.V., WO 2004 038032, 2002 (P. J. M. Baets, E. R. Coldenhoff).
31. A. Torang, A. Jonsson, G. Zacchi, *Appl. Biochem. Biotechnol.* **76** (1999) 143–157.
32. Cargill Dow Polymers LLC, US 6485947, 2002 (V. Rajgarhia et al.).
33. Purac Biochem B.V., WO 2004 038032, 2002 (P. J. M. Baets, E. R. Coldenhoff).
34. A. Torang, A. Jonsson, G. Zacchi, *Appl. Biochem. Biotechnol.* **76** (1999) 143–157.
35. Michigan Biotechnology Institute, EP-A 19900301828, 1989 (D. A. Glassner, R. Datta).
36. Roquette Freres, US 6280985, 2000 (J. J. Caboche, C. Fouache, J. C. Choque, P. Duflot, E. Dubois).
37. Archer-Daniels-Midland, US 6627061, 2001 (K. N. Mani).
38. Engenics, Inc., US 4698303, 1986, (R. B. Bailey, D. K. Joshi, S. L. Micharls, R. A. Wisdom).
39. J. Mizrahi, A. M. Eyal, R. Canari, B. Hazan, J. N. Starr, US Patent Application 20020102672, 1999.
40. Purac Biochem B.V., US 6,509,179, 1999, (B. I. Veldhuis-Stibos, J. Van Bruegel, W. J. Groot, B. M. Dierdorp).
41. Engenics, Inc., US 4698303, 1986, (R. B. Bailey, D. K. Joshi, S. L. Michaels, R. A. Wisdom).
42. Brussels Biotech, US 6489508, 1999 (F. Van Gansberghe, J. C. Bogaert, E. Malhaize, M. Van Gansberghe, F. Wolff).
43. Purac Biochem B.V., US 6630603, 2000 (J. Van Breugel, J. Van Krieken, A. Cerda Baro, J. M. Vidal Lancis, M. Camprubi Vila).
44. Cargill Dow LLC, WO 0138284, 1999 (G. Quarderer, L. A. Robbins, C. M. Ryan).
45. E. I. DuPont de Nemours and Company, US 5210296, 1991 (M. C. M. Cockrem, P. D. Johnson).
46. A. E. Staley Manufacturing Co., US 6291708, 1999 (M. C. M. Cockrem).
47. Cargill Inc., US 6160173 1997 (A. M. Eyal, P. R. Gruber, P. McWilliams, D. R. Witzke).
48. Lactascan ApS, US 6319382, 1997 (B. Norddahl).
49. Chronopol, Inc. US 5681728, 1995 (F. Miao).
50. UOP, US 5068418, 1989 (S. Kulprathipanja, A. R. Oroskar).
51. Reilly Industries, Inc., US 6114577, 1995 (F. H. Verhoff, M. Grendze).
52. The Regents of the University of California, US 5132456, 1991 (C. J. King, L. A. Tung).
53. Cargill Inc., US 5338822, 1992 (P. R. Gruber, J. J. Kolstad, E. S. Hall, R. S. Eichen Conn, C. M. Ryan).

Lactose and Derivatives

For Lactitol see → Sugar Alcohols.

Willem A. Roelfsema, Friesland Coberco Dairy Foods, Meppel, The Netherlands (Chap. 1, Section 2.2)

Ben F. M. Kuster, Technische Universiteit Eindhoven, Eindhoven, The Netherlands (Chap. 1, Section 2.2)

Michiel C. Heslinga, Solvay Pharmaceuticals, Weesp, The Netherlands (Section 2.1)

Hendrik Pluim, Solvay Pharmaceuticals, Weesp, The Netherlands (Section 2.1)

Marinus Verhage, Solvay Pharmaceuticals, Weesp, The Netherlands (Section 2.1)

1. Lactose 1
1.1. Properties 1
1.1.1. Physical Properties 1
1.1.2. Chemical Properties 2
1.2. Production 3
1.3. Quality Specifications 3
1.4. Uses 3
1.5. Economic Aspects 4
1.6. Producers 4
1.7. Physiological Properties 4
2. Derivatives 5
2.1. Lactulose 5
2.1.1. Properties 6
2.1.2. Production 6
2.1.3. Quality Specifications 6
2.1.4. Toxicology and Physiology 7
2.1.5. Medical and Pharmaceutical Uses . . 7
2.2. Miscellaneous Derivatives 7
2.2.1. Lactobionic Acid 7
2.2.2. Hydrolyzed Lactose Syrup 7
3. References 7

1. Lactose

Lactose, milk sugar, $C_{12}H_{22}O_{11}$, 4-*O*-β-D-galactopyranosyl-α-D-glucopyranose [*63-42-3*], M_r 342.30 (anhydrous); monohydrate [*64044-51-5*], M_r 360.31, is the principal carbohydrate constituent of milk of most mammals. Lactose concentration varies from 0 to 9 wt %, being zero in sea lions, 4.0 % in goats, 4.8 % in cows, and 6.7 wt % in humans [1].

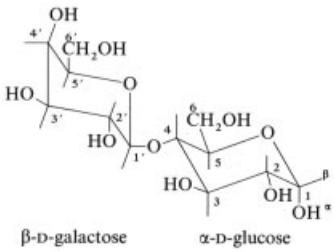

α-Lactose

Bartoletti discovered lactose in milk in 1633, and in 1780 Scheele identified its sugar character. The structure was confirmed by synthesis in 1942 by Haskins et al. [2].

1.1. Properties

Lactose exists in two anomeric forms: α- and β-lactose, differing only in the relative positions of the hydrogen and the hydroxyl group at the C^1-atom of the pyranosidic glucose-part. In aqueous solution lactose consists of 61.5 % of β-pyranose and 38.5 % of α-pyranose; the equilibrium between both forms is established by mutarotation ($K_\alpha + K_\beta = 1.74 \times 10^{-4}\ s^{-1}$ at 20 °C) [3]. Lactose usually crystallizes in the form of its α-lactose monohydrate. The anhydrous β-form can only be obtained at crystallization temperatures > 93 °C. Amorphous lactose, which is quite hygroscopic, is a mixture of α- and β-lactose; it can be obtained by rapid drying methods, e.g., spray drying.

1.1.1. Physical Properties

The physical properties are shown in Table 1 and Figures 1 and 2 [1], [4–9].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_107

Table 1. Physical properties of Lactose

Properties	α-Lactose monohydrate	β-Lactose
mp, °C	201 – 202 (decomp.)	253 (decomp.)
Specific rotation $[\alpha]_D^{20}$, 10^{-2} ° · cm^2/g	+89.4	+34.3
equilibrium value	+52.5	+52.5
Crystalline form	sphenoidal monoclinic ″tomahawk″ (Fig. 1)	monoclinic ″uneven-sided diamond″
a, nm	0.7982	1.081
b, nm	2.1652	1.334
c, nm	0.4824	0.484
β, °	109.78	91.15
Density (293 K), g/cm^3	1.54	1.59
Bulk density, g/cm^3		
poured	0.5 – 0.9	
tapped	0.75 – 1.1	
Enthalpy of dissolution, J/g	− 50.24	− 9.62
Enthalpy of combustion, J/g	− 16.1	− 16.5
Specific heat (298 K), J K^{-1} g^{-1}	1.251	1.193
Entropy (298 K), J K^{-1} mol^{-1}	415	386
Entropy of formation (298 K), J K^{-1} mol^{-1}	− 2453	− 2248
Enthalpy of formation (298 K), kJ/mol	− 2481	− 2233
Gibbs free energy (298 K), kJ/mol	− 1750	− 1564
Solubility, g/100 g		
in H_2O	see Fig. 2	see Fig. 2
in DMSO	59 (20 °C)	30 (20 °C)
	73 (80 °C)	43 (80 °C)
pK_a at 3 – 5 °C	13.8	13.5
Diffusion constant in H_2O, m^2/s	43×10^{-11}	

Properties of solutions	Concentration (g lactose per 100 grams water)				
	0	10	20	30	40
Density (20 °C), kg/m^3	998	1043	1082	1124	1173
Viscosity, mPa · s					
(20 °C)	1.00	1.38	2.04	3.42	7.01
(60 °C)	0.47	0.70	0.90	1.29	2.19
Refractive index (25 °C)	1.3325	1.3484	1.3659		

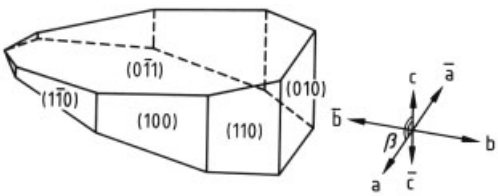

Figure 1. Crystal structure of α-lactose monohydrate

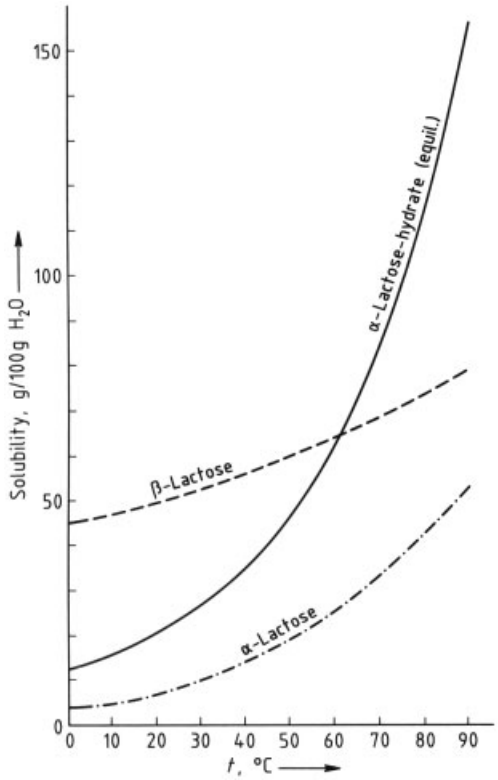

Figure 2. Initial solubilities of α-lactose and β-lactose and final solubility at mutarotation equilibrium of α-lactose-hydrate vs. temperature

1.1.2. Chemical Properties

Lactose has the same principle molecular structure as most other disaccharides and differs only in its configuration. Thus, it generally reacts in the same way as other sugars (→ Carbohydrates). Important points of attack are the reducing end of the glucose moiety, the various hydroxyl groups, and the glycosidic bond between the galactose and the glucose moiety [10].

Examples of reaction types are: (1) formation of anhydro derivatives under the influence of heat and radiation and, (2) formation of unsaturated heterocycles by heat in the presence of an acid [11], (3) oligomerization [12], (4) more or less extensive oxidation at various sites by chemical means or catalytically [11], (5) Maillard type reactions with amino compounds [13], (6) isomerization [14], (7) acetalation [15], (8) hydrogenation [16], (9)

alkaline degradation [13], (10) complexation of metal cations [17], (11) glycosidation [11], (12) etherification [13], (13) halogenation [13], (14) esterification [14], and (15) hydrolysis by enzymatic or acid catalysis [18], [19].

1.2. Production

Generally cheese (sweet) whey or casein (acid) whey (→ Cheese, Processed Cheese and Whey, Chap. 7.2.) is used as raw material for production of lactose (see Fig. 3). Today, permeates from the ultrafiltration of milk or sweet whey are also coming into consideration. With all types of raw materials precautions have to be taken regarding the minerals present; calcium salts may give rise to deposits in evaporators.

The liquid whey (ca. 6.5 % dry matter content) is clarified (a) and concentrated in falling-film evaporators (b) to 55 – 65 % total solids. Upon cooling (c), the greater part of the lactose crystallizes as lactose monohydrate. The crystal slurry is transferred to a line of decanters and centrifuges and washed several times. Thus, after drying, milling and fractionation, yellowish lactose (edible grade) of different particle sizes is obtained.

After the washing stage the raw product can be redissolved (i), treated with activated carbon (j) and recrystallized to refined or pharmaceutical grades of lactose. From 1000 kg liquid whey 32.5 kg lactose and 32.5 kg delactosed wheypowder are obtained.

Environmental Protection. To avoid pollution the dryer should be provided with a cyclone to remove the powder from the exhaust air. Fluid cleaning streams must be purified by simple aerobic or anaerobic digestion before entering a public sewage treatment.

1.3. Quality Specifications

Lactose is produced in various grades of purity (see Table 2).

Chemical analyses of lactose are carried out by standard methods, complying with U.S.P./NF, EP, and JP specifications. For details see the Pharmacopeias mentioned.

Table 2. Specifications of various grades of lactose

Lactose grade	Crude	Edible	Pharmaceutical
Lactose, monohydrate, %	98	min. 99.0	min. 99.7
Free moisture, %	0.4	max. 0.3	max. 0.2
Protein (N content × 6.38), %	1.0	max. 0.3	max. 0.1
Ash, %	0.5	max. 0.3	max. 0.1
Riboflavin, ppm	200	50	0

1.4. Uses

In the various uses of lactose the following functional properties are important [14], [18], [20], [21]:

1) free flowability,
2) flavor and taste enhancement,
3) nonhygroscopic, anti-caking, fat-coating properties,
4) adsorption and absorption capacities,
5) ease of solubility,
6) agglomerating and instantizing capacities,
7) slow fermentation plus acid regulation,
8) stability, long shelf life,
9) absence of odor,
10) good compressibility, and
11) moderate sweetness; the concentration-dependent sweetness of lactose is ca. 30 % relative to sucrose

Food quality, i.e., *edible lactose* is used for:

1) infant and baby food,
2) frozen desserts,
3) ice cream,
4) skim milk, condensed milk,
5) dietetic foods,
6) coffee creamers,
7) chocolate and sugar confectionery,
8) bakery products, meat products,
9) canned fruit and vegetables,
10) dry soups, sauces and spices,
11) binder for liquid foods, and
12) fermentation processes

Pharmaceutical-grade lactose is used for: tableting (wet granulation, direct compressing, retableting), capsule filling, hypodermics, pan coating, and as raw material for lactulose and lactitol production.

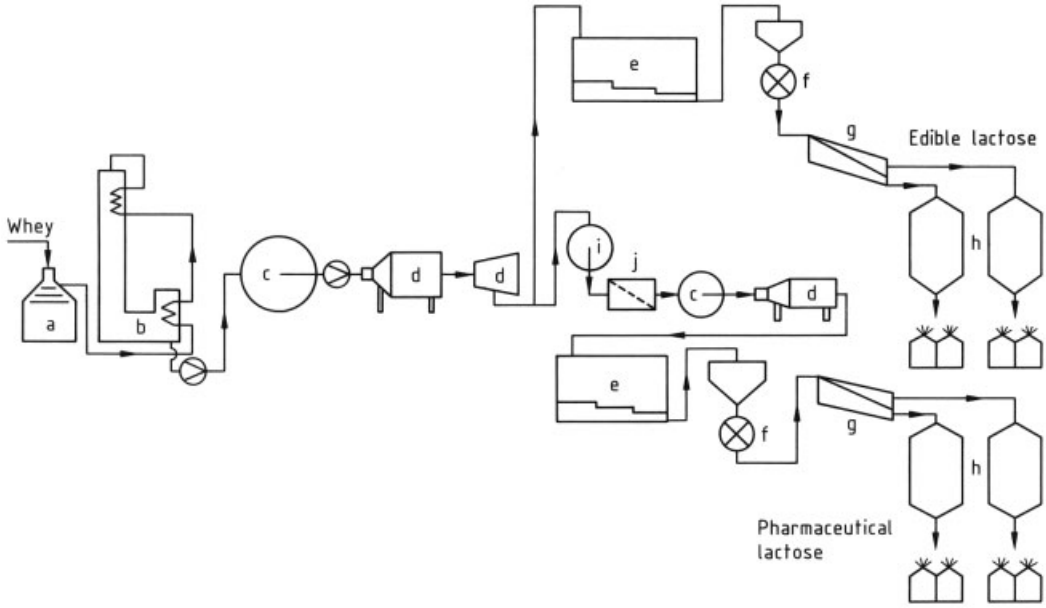

Figure 3. Production of lactose
a) Clarifier; b) Evaporator; c) Crystallization tank; d) Separator; e) Dryer; f) Mill; g) Sieve; h) Packing; i) Redissolving tank; j) Filter

1.5. Economic Aspects

Production of lactose is especially located in those areas which have a highly centralized cheese production such as Europe, Eastern and Central States of America and New Zealand. Production figures of 1996 are in 10^3 t/a:

World total	500
Europe	275
North America	180
Oceania	45

Areas of application can roughly be divided into pharmaceutics 100 000 t/a (20 %), baby and infant food 150 000 t/a (30 %), human food 190 000 t/a (38 %), fermentation and animal feed 60 000 t/a (12 %).

1.6. Producers

Important producers of lactose are:

Canada:	Saputo Ltd., Montreal
Finland:	Valio, Helsinki
France:	Besnier Bridel SA, Retiers
Germany:	Alpavit, Heising
	BMI, Landshut
	Meggle Milchindustrie GmbH & Co. KG, Wasserburg
	Milei GmbH, Stuttgart
Ireland:	Avonmore Creameries Ltd., Kilkenny
The Netherlands:	Borculo Whey Products, Borculo
	D.M.V. International, Veghel
	DOMO Food Ingredients, Beilen
New Zealand:	Lactose Company of New Zealand, Hawera
United States:	AMPC Inc., Ames IA
	Avonmore Ingredients Inc., Monroe WI
	Foremost Farms USA, Baraboo WI
	Land O'Lakes Inc., Minneapolis MN
	Leprino Foods, Denver CO
	Lynn Proteins Inc., Granton WI
	United Dairyman of Arizona, Tempe AZ
	Wisconsin Whey International Inc., Madison WI

1.7. Physiological Properties [20]

The metabolizable energy of α-lactose monohydrate amounts to -15.7 kJ/g, that of β-lactose to -16.0 kJ/g.

Lactose is less cariogenic than other natural mono- and disaccharides. This results from:

1) A slower rate of fermentation in the tooth-plaque,
2) A lower final acidity in the same environment, and
3) The lack of production of extra-cellular bacterial polysaccharides

Once lactose reaches the intestines it is enzymatically cleaved into glucose and galactose. The lactic acid formed by subsequent fermentation, and the resulting acid pH prevent the growth of undesirable microorganisms.

The intestinal absorption of calcium and other di and trivalent metal ions, such as magnesium, barium, zinc and iron is strongly improved by the presence of lactose. This can be explained by a decrease of the pH in the digestive tract, which increases the solubility and absorbability of the complexed metal ions.

Lactose has moderate laxative properties because its fermentation products increase the osmotic pressure in the large intestine, thus the loss of too much water is prevented by reabsorption.

In Asia, Africa, and South America large parts of the population have difficulties in absorbing lactose, whereas in Europe and North America, $< 8\,\%$ of the population have such difficulties. This so-called lactose intolerance, i.e., clinical phenomena such as flatulence and diarrhoea, is normally caused by a strongly reduced lactase activity in the mucosa of the small intestine, giving rise to malabsorption of lactose. This does not imply, however, that a malabsorber is totally lactose intolerant [17].

Generally lactase activity in adults is ca. 10 % of that of newborns.

2. Derivatives

2.1. Lactulose

Lactulose [*4618-18-2*], 4-*O*-β-D-galactopyranosyl-D-fructose, $C_{12}H_{22}O_{11}$, M_r 342.30, is a synthetic disaccharide in which galactose is linked to fructose. It was first synthesized in 1929 by isomerization of lactose [22].

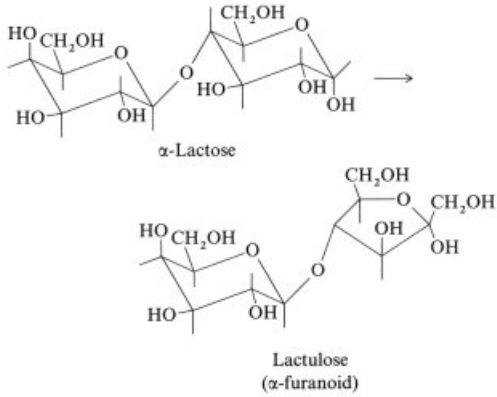

The first synthesis of lactulose was completely described by MONTGOMERY and HUDSON in 1930 [23]. They prepared lactulose from lactose in an alkaline solution of lime. Lactulose has not been detected in raw milk, but has been found in milk and milk products after heat treatment [24], [25].

Molecular Structure. The molecular structure of lactulose has been elucidated mainly by HICKS [26], [29].

The anomeric composition of lactose depends on the physical state. In solution the anomeric composition appears to be related to temperature [27] (Table 3).

Table 3. Variation in the percent distribution of anomeric forms of lactulose in water with temperature

	25 °C	58 °C	73 °C
ß-Pyranose	61.5 %	54.7 %	52.1 %
ß-Furanose	29.3 %	31.6 %	35.5 %
α-Furanose	7.6 %	11.2 %	12.4 %
Open ketoform	1.6 %	2.5 %	n.d.

In the solid state two crystal modifications have been described; lactulose anhydrate [29] and lactulose trihydrate [28]. Both modifications differ in anomeric composition as shown in Table 4.

Table 4. Percent distributions of anomeric forms in the crystal modifications of lactulose

	Anhydrate	Trihydrate
ß-Pyranose	16.0 %	
ß-Furanose	72.2 %	100 %
α-Furanose	11.8 %	

The coexistence of different anomers in the same crystal structure is often observed in carbohydrate chemistry, because the ability to form hydrogen bonds of comparable energy facilitates cocrystallization.

2.1.1. Properties

Lactulose anhydrate and trihydrate are white, odorless crystalline powders. Its most important properties are as follows:

mp anhydrate [30]	169 °C
mp trihydrate [31]	68 °C
Solubility in water (kg/kg) anhydrate [30]	3.2 (30 °C), 4.3 (60 °C), > 6.2 (90 °C)
Solubility in water (kg/kg) trihydrate [32]	1.3 (5 °C), 1.5 (10 °C), 2.1 (20 °C), 3.1 (30 °C)
Specific rotation $[\alpha]_D^{20}$ [30]	− 51.4 °
Degree of sweetness [33]	0.48 – 0.62 (sucrose = 1.0)
Concentration (% wt) related to refraction index [34]	0.97° Brix (sucrose index scale)

2.1.2. Production

Lactulose is available on the market both as a syrup and as a dry crystalline powder (anhydrate).

Lactulose can be prepared from an alkaline solution of lactose by the Lobry de Bruyn – Alberda van Ekenstein rearrangement (→ Carbohydrates, → Glucose and Glucose-Containing Syrups, Chap. 2.1.). The first synthesis was performed with lime [23]. Since then many catalysts, such as alkali hydroxides [36], [37], magnesium oxide [38], alkali and alkaline-earth aluminates [39] or alkaline-earth sulfite [40], alkali borate [42], [43], alkali tetraborate plus ion-exchange resin [44], anion exchange resin [45], and amines [46] are reported to give good yields or to minimize secondary reactions.

On a industrial scale, lactulose syrup is usually produced by one of two ways:

1) Common lactulose syrup is prepared by dissolution of lactose in water and addition of sodium hydroxide solution while heating. Because of inevitable degradation reactions the conversion has to be limited. The excess of lactose is removed by crystallization. The solution is deionized using ion exchange resins. A subsequent concentration step leads to a second crop of unreacted lactose.
2) Alternative processes use coreagents such as borate [41] and aluminate [47]. These lactulose complexing agents enable the conversion of lactose to lactulose in a high yield by the prevention of degradation reactions [48]. So higher conversion and selectivity can be reached. Borate or aluminate is subsequently recycled.

The resulting syrup is the common marketed lactulose solution, containing 47 – 53 wt % of lactulose and some other sugars such as lactose, epilactose, galactose, fructose and tagatose. The amount of these related sugars depends on the process used and optional additional purification steps.

Two different processes for obtaining crystalline lactulose anhydrate are used:

1) Further purification of common lactulose syrup by crystallization from alcohols and/or water followed by drying [49].
2) Solidification of an already pure lactulose syrup by well controlled evaporation and stirring, optionally using seed crystals [50].

Two completely different approaches to the synthesis of lactulose, (1) an enzymatic coupling of galactose and fructose using β-galactosidase and (2) an Amadori rearrangement (→ Carbohydrates) of *p*-tolyl-*N*-lactosylamine followed by acid hydrolysis, have been reported [51], [52]. These processes have not found industrial application because of their low yield and intricacy.

Producers. The main manufacturers of lactulose are Solvay (the Netherlands and Canada), Inalco (Italy), Leavosan (Austria), Morinaga (Japan) and Danipharm (Denmark) producing about 12 000 t/a for medical and pharmaceutical uses.

2.1.3. Quality Specifications

Lactulose, both the syrup and the crystalline anhydrate, is a pharmaceutical product and its specifications are laid down in, e.g., the U.S. Pharmacopeia, European Pharmacopoeia, and Japanese Pharmacopoeia.

2.1.4. Toxicology and Physiology

Toxicology. Lactulose administration has not been reported to cause adverse side effects nor does it accumulate in the recipient. Therefore its intake is considered to be safe.

In oral administration, the acute toxicity is at the same level as found for sucrose (male rat 25 g/kg, female rat 28 g/kg) [53], [54].

Physiology. The significance of lactulose was recognized by PETUELY in 1957 when he discovered that lactulose plays an important role in infant nutrition as a "Bifidus Factor" to increase the prevalence of *Bifidobacterium* in the intestinal tract [55].

Most orally administered di-, oligo- and polysaccharides are split by intestinal enzymes into monosaccharides which are absorbed and enter the peripheral circulation. Lactulose, however is not hydrolyzed in the human small intestine and is consequently not absorbed [55]. Most of it is transported to the large intestine, where it is metabolized to low molecular mass organic acids by lactobacilli (*Bifidobacterium*).

Lactobacilli suppress harmful intestinal bacteria, assist in digestion and absorption, produce vitamins, and stimulate immune responses. Lactulose is effective in promoting the proliferation of lactobacilli. Moreover, lactulose does not raise blood glucose levels or influence insulin secretion [56], [57]. Since it is not hydrolyzed in the mouth, lactulose does not promote dental caries [58].

2.1.5. Medical and Pharmaceutical Uses

The main application of lactulose is in humans for the treatment and prevention of portal systemic encephalopathy (PSE) and chronic constipation [59]. Lactulose is an effective laxative because it is cleaved in the colon by lactobacilli to low molecular mass organic acids. This leads to a lowering of the pH of the colon contents and via an increase in osmotic pressure to an increase of the colon contents.

The efficacy of lactulose in the treatment of PSE probably also results from its effect on increasing the number of lactobacilli in the colon. As lactobacilli lack urease and other ammonia-generating enzymes, this might result in a decreased faecal urease activity, decreased urea hydrolysis, and therefore decreased ammonia formation.

Consistent with this hypothesis is the fact that when patients with PSE are treated with lactulose a reduction of the elevated arterial and venous blood ammonia levels occurs and lactulose notably improves the disease.

Other therapeutic uses suggested for lactulose include the treatment of salmonellosis [60].

The discovery of the important role of lactulose in the growth of lactobacilli has led to several areas of application in medicine (therapeutics, diagnostics), human nutrition (food additive), and even veterinary science and nutrition. These areas are rapidly increasing in importance.

2.2. Miscellaneous Derivatives

2.2.1. Lactobionic Acid

Lactobionic acid [*96-82-2*] is used as a carrier for antibiotics and prepared on a small scale by electrochemical oxidation [61].

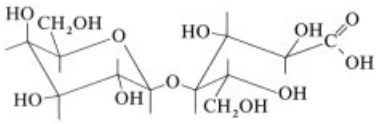

Lactobionic acid

2.2.2. Hydrolyzed Lactose Syrup

Lactose is hydrolyzed by enzymatic processes and yields a syrup after concentrating. By this procedure solubility is increased to 60 % total solids and sweetness becomes 0.70 relative to sucrose. Most hydrolysis procedures however are applied to deproteinized/demineralized permeates, (demineralized) whey and milk products [20].

3. References

1. P. Linko in G. G. Birch, K. J. Parker (eds.): *Nutritive Sweeteners,* Applied Science Publishers, London – New Jersey 1982, Chap. 6.

2. *Ullmann,* 4th ed. **24** 768 – 770.
3. P. A. Morrissey in P. F. Fox (ed.): *Developments in Dairy Chemistry,* **vol. 3,** Elsevier Applied Science Publishers, London – New York 1985, Chap. 1.
4. T. A. Nickerson in B. H. Webb, A. H. Johnson, J. A. Alford (eds.): *Fundamentals of Dairy Chemistry,* 2nd ed., The Avi Publishing Company Inc., Westport 1974, Chap. 6.
5. J. W. Noordik et al., *Z. Kristallogr.* **168** (1984) 59 – 65.
6. R. A. Visser, doctoral thesis, Nijmegen, The Netherlands (1983).
7. G. de Wit, A. P. G. Kieboom, H. van Bekkum, *Carbohydr. Res.* **127** (1984) 193 – 200.
8. P. Walstra, R. Jenness: *Dairy Chemistry and Physics,* John Wiley & Sons, New York 1984, Chap. 3.
9. T. J. Buma, G. A. Wiegers, *Neth. Milk Dairy J.* **21** (1967) 208 – 213.
10. L. A. W. Thelwall in P. F. Fox (ed.): *Developments in Dairy Chemistry,* **vol. 3,** Elsevier Applied Science Publishers, London – New York 1985, Chap. 2.
11. J. R. Clamp, L. Hough, J. L. Hickson, R. L. Whistler, *Adv. Carbohydr. Chem.* **16** (1961) 159 – 206.
12. B. W. Lew, US 3 974 138, 1976.
13. L. A. W. Thelwall in C. K. Lee (ed.): *Developments in Food Carbohydrate,* **vol. 2,** Applied Science Publishers London 1980, 275 – 326.
14. J. G. Zadow, *J. Dairy Sci.* **67** (1984) 2654 – 2679.
15. L. A. W. Thelwall, *J. Dairy Res.* **49** (1982) 713 – 724.
16. J. A. van Velthuijzen, *J. Agric. Food Chem.* **27** (1979) 680 – 686.
17. D. M. Paige, L. R. Davis in P. F. Fox (ed.): *Developments in Dairy Chemistry,* **vol. 3,** Elsevier Applied Science Publishers, London – New York 1985, Chap. 4.
18. P. G. Hobman, *J. Dairy Sci.* **67** (1984) 2630 – 2653.
19. R. R. Mahoney in P. F. Fox (ed.): *Developments in Dairy Chemistry,* **vol. 3,** Elsevier Applied Science Publishers, London – New York 1985, Chap. 3.
20. B. F. Pritzwald-Stegmann, *J. Soc. Dairy Technol.* **39** (1986) 91 – 97.
21. J. G. Zadow in *Milk, The Vital Force* (Proc. XXII International Dairy Congress, The Hague, 1986), D. Reidel Publishing Company, Dordrecht 1987, 737 – 748.
22. E. M. Montgomery, C. S. Hudson, *Science* **69** (1929) 556.
23. E. M. Montgomery, C. S. Hudson, *J. Am. Chem. Soc.* **52** (1930) 2101 – 2106.
24. S. Adachi, *Nature* **181** (1958) 840 – 841.
25. S. Adachi, S. Patton, *J. Dairy Sci.* **44** (1961) 1375 –1393.
26. P. E. Pfeffer, K. B. Hicks, W. L. Earl, *Carbohydr. Res.* **111** (1983) 181 – 194.
27. P. E. Pfeffer et al., *Carbohydr. Res.* **102** (1982), 11 – 22.
28. G. A. Jeffrey, *Carbohydr. Res.* **226**(1992) 29 – 42.
29. G. A. Jeffrey, R. A. Wood, P. E. Pfeffer, K. B. Hicks, *J. Am. Chem. Soc.* **105** (1983) 2128 – 2130.
30. B. J. Oosten, *Recl. Trav. Chim. Pays-Bas* **86** (1967) 675 – 676.
31. G. A. Jeffrey et al., *Carbohydr. Res.* **226** (1992) 29 – 42.
32. T. Mizota, *Carbohydr. Res.* **263** (1994) 163 – 166.
33. F. W. Parrish et al., *J. Food Sci.* **44** (1979) 813 – 815.
34. A. A. de Reijke, P. van Bemmel, H. W. Geluk, *J. Pharm. Sci.* **73** (1984) 1478 – 1479.
35. E. M. Montgomery, *J. Am. Chem. Soc.* **52** (1930) 2101 – 2106.
36. L. Gatzsche, H. Haenel, *Ernährungsforschung* **12** (1967) 641 – 647.
37. Morinaga Milk Industry Co., US 3 816 174, 1974 (T. Nagasawa et al.).
38. S. Galup, ES 397 810, 1974.
39. Kraftco Corporation, US 3 546 206, 1970 (J. H. Guth, M. Prospect, L. Tumerman).
40. Laevosan Gesellschaft, DE 2 002 385, 1970 (E. Nitsch, S. Mühlbock).
41. Duphar International Research B. V. U.S. 5 071 530, (1991) (R. E. Krumbholz et al.).
42. J. F. Mendicino, *J. Am. Chem. Soc.* **82** (1960) 4975 –4979.
43. Morinaga Milk Industry Co., JP 91/77, 1977 (T. Nagasawa et al.).
44. R. Carubelli, *Carbohydr. Res.* **2** (1966) 480 – 485.
45. H. Demaimay, C. Baron, *Lait* **575 – 576** (1978) 234 – 245.
46. F. W. Parrish, US 3 514 327, 1970.
47. Sirac SRL, US 4 957 564, 1990 (R. Carrobi et al.).
48. R. van de Berg et al., *Carbohydr. Res.* **253** (1994) 1 – 12.
49. F. Innocenti, EP 318 630, 1988.
50. Duphar International Research B. V., EP 480 519, 1991 (H. Pluim et al.).

51. M. Vaheri, V. Karppinen, *Acta Pharm. Technol.* **87** (1978) 75.
52. S. Adachi, *Chem. Ind. (London)* (1957) 956.
53. H. Okumura, T. Gomi, *Lab. Clin.* **7** (1973) 3517.
54. E. M. Boyd, I. Godi, M. Abel, *Toxicol. Appl. Pharmacol.* **7** (1965) 609 – 618.
55. F. Petuely, *Z. Kinderheilk.* **79** (1957) 174.
56. H. Ruttloff et al., *Nahrung* **11** (1967) 39 – 47.
57. K. Hoffman, D. A. A. Mossel, W. Konus, J. H. Kamer, *Klin. Wochschr.* **42** (1964) 126 – 130.
58. K. K. Mäkinen, J. Rekola, *J. Dent. Res.* **54** (1975) 1244.
59. U.S. Philips Corp., US 3 461 204, 1967 (J. Bircher).
60. K. Hoffmann, *Dtsch. Med. Wochenschr.* **100** (1975) 1429 – 1431.
61. S. K. Dutta, K. Sanat, US 4 137 397 (1979).

Lactulose → Lactose and Derivatives
Laminated Glass → Glass
Laminated Plastics → Reinforced Plastics
Laminated Wood → Wood

Lamps

Willem J. van den Hoek, Philips Lighting B.V., Central Development Lamps, Eindhoven, The Netherlands
Alexander G. Jack†, Philips Lighting B.V., Central Development Lamps, Eindhoven, The Netherlands
Gerardus M. J. F. Luijks, Philips Lighting B.V., Central Development Lamps, Eindhoven, The Netherlands

1. Introduction 1
2. Light Generation 4
2.1. Thermal Radiation from Solids . . . 4
2.2. Radiation from Gas Discharges . . . 5
2.3. Radiation from Semiconductors (LEDs) 6
2.4. Modeling of Lamps 10
3. Main Types of Lamps 10
4. Lamp Chemistry 17
4.1. General 17
4.2. Transport Chemistry 18
4.2.1. Principles 18
4.2.2. Incandescent Lamps 19
4.2.3. Halogen Incandescent Lamps 20
4.2.4. Metal Halide Lamps 21
5. Basic Lamp Materials 22
5.1. Metals 22
5.2. Glass 24
5.2.1. Production and Composition 24
5.2.2. Properties 25
5.2.3. Glasses for Lamp Types 26
5.2.4. Glass-to-Metal Seals 28
5.3. Polycrystalline Alumina 28
5.4. Fluorescent Powders 31
6. Special Lamp Materials 32
6.1. Getters 32
6.2. Amalgams 35
6.3. Emitters 36
6.4. Cements and Marking Paints 38
7. Coatings 38
8. Lamp Production 39
9. Economic Aspects and Uses 40
10. References 42

1. Introduction

It is more than one hundred years since the first viable incandescent lamp was produced industrially with a sufficiently long operating life. Most major advances in light sources are the result of the discovery and development of new materials. Two such examples are fluorescent powders and translucent ceramic envelopes which led to the development of fluorescent lamps and high-pressure sodium lamps, respectively. Worldwide research and development continuously improve lamp performance. Figure 1 shows the improvement over the years in the efficiency with which a lamp converts electrical power into light (i.e., luminous efficacy).

General. Light is the visible part of the electromagnetic spectrum and extends from about 380 to 760 nm. The response of the eye to the spectral stimulus is described by the *spectral luminous efficiency* $V(\lambda)$ which is depicted in Figure 2.

Types of Lamp. A light source is a device for converting energy into light. A radiation source is similar to a light source but is optimized to radiate in the IR ($>$ 760 nm) or the UV ($<$ 380 nm) region and is often used for industrial or medical applications.

Although a large number of ways exist to generate light, only a limited number of processes are applied in lamps. The main considerations for a practical light source concern conversion efficiency, color aspects, lamp life, and economic factors. In practice almost all light sources convert electrical energy into light and comprise incandescent lamps and gas discharge lamps. Conversion of chemical energy takes place in combustion flash lamps, flames, and candles.

In an *incandescent lamp,* electrical energy is transformed into thermal radiation by heating a coiled tungsten wire to temperatures in the range 2700 – 3300 K (see Section 2.1). The tungsten coil is enclosed in a transparent vessel of glass or fused silica. The medium in the vessel may be a vacuum or an inert gas (normal lamps), or an in-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_115

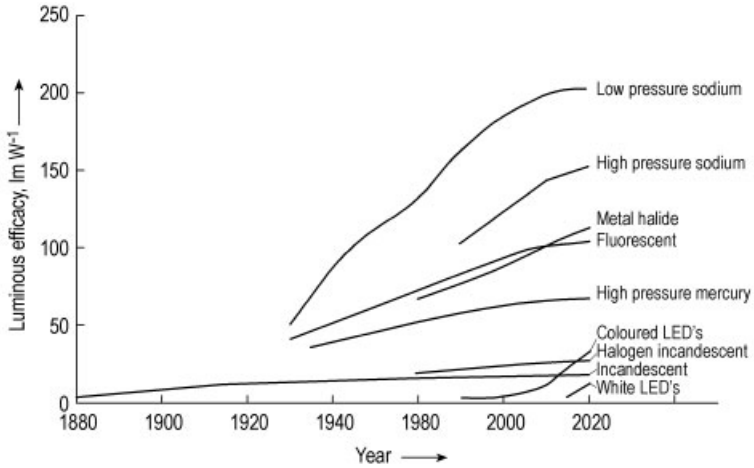

Figure 1. Improvement of luminous efficacy of various light sources during the history of electric lighting

ert gas with a small amount of gaseous halogens (halogen lamps).

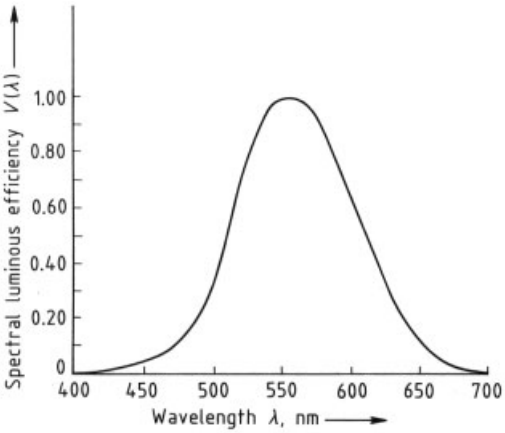

Figure 2. The spectral luminous efficiency $V(\lambda)$ as a function of wavelength λ

In *gas-discharge lamps,* electrical energy is converted into radiation by passage of an electrical current through an ionized gas enclosed in a vessel of glass, fused silica, or translucent alumina. An ionized gas is a mixture of free electrons, positive ions, and neutral species, which is usually referred to as a plasma. The majority of discharge lamps has interfaces between the plasma and the solid conductors in the form of electrodes. In the 1990s, also electrodeless lamps have been introduced in the market (see Chap. 3). Because of the nature of the voltage – current characteristics of discharge plasmas, stable operation of a discharge lamp requires a stabilizing element (or *ballast*) in the electrical circuitry. Furthermore an *ignitor* may be required for converting the neutral gas into a conducting plasma upon switching on the lamp. Discharge lamps are classified as high-pressure or low-pressure, depending on the rate of energy transfer in the discharge (see Section 2.2).

In most discharge lamps the light comes directly from the discharge plasma. In the low-pressure mercury discharge, however, the generated radiation is in the UV region (254 and 185 nm). This "invisible" radiation is converted into light by a fluorescent powder coating on the wall of the discharge vessel (see Section 5.4). In recent years, there has been a breakthrough in the efficiency of light-emitting semiconducting solid-state devices, so-called light emitting diodes (LEDs, see Section 2.3 and Chap. 3).

Quantification of Emitted Light [1]. In a light source the input power results in emission of radiation, which is divided over the spectral range from the ultraviolet to the infrared. The *spectral power distribution* $P(\lambda)$ (expressed in radiated watts per wavelength interval) determines important characteristics of the light source. The *light output* or *luminous flux* (expressed in lumen) is found by integrating the spectral power distribution $P(\lambda)$ multiplied by the spectral luminous efficiency $V(\lambda)$, over the wavelength range. Division of this total lumen

output by the (electrical) input power P_{in}, gives the *luminous efficacy* η in lumens per watt:

$$\eta = k \int_{\lambda} P(\lambda)\, V(\lambda)\, \mathrm{d}\lambda / P_{\mathrm{in}}$$

The coefficient k has a value of 683 lm/W.

Another important quantity related to the luminous flux is the luminance or brightness. The *luminance* gives the luminous flux per unit of solid angle per unit of surface from a given surface and is expressed in candela per square meter, where the candela (cd) is defined as 1 lm/sr (lumen per steradian).

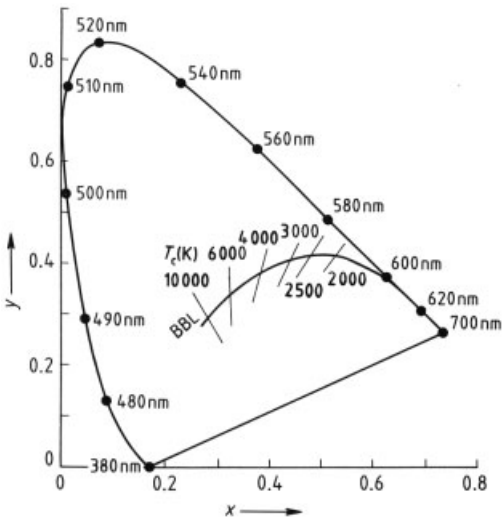

Figure 3. The CIE 1931 (x, y) chromaticity diagram
The blackbody locus (BBL) shows the chromaticity coordinates for black body radiators at the indicated temperatures.

Color. Besides the overall amount of light, the quality of the light, (i.e., color and color rendition) is also very important. Light quality is intimately related to color perception by the human eye. The color of light is indicated by the chromaticity coordinates x and y in the chromaticity diagram, see Figure 3. The (x, y) values are obtained from integrals over the spectral power distribution, weighted by the standardized spectral sensitivity functions, which describe human color perception [1] (→ Dyes, General Survey, Chap. 3.3.). The correlated *color temperature* of a light source corresponds to the temperature of a blackbody radiator with (approximately) the same chromaticity coordinates (cf. Section 2.1). The perception of colored pigments upon illumination with a particular light source is characterized by the *color rendering index* (CRI). This quantity is based on a comparison of the color renditions of the light source in question and a standard source (e.g., a blackbody source or daylight). The maximum value of the CRI is 100 and denotes excellent color rendition. Luminous efficacy and color-related properties are conflicting aspects in lamp design and a compromise always has to be made.

Lighting Systems. The great majority of lamp users are not primarily interested in lamps but require lighting, i.e., provision of the right light at the right place and at the right time. This necessitates a system for creating high-quality lighting at the lowest cost. A lighting system usually consists of (1) a lamp, (2) a ballast – ignitor combination, and (3) a luminaire (a device which directs the light to where it is needed).

The combination of lamp and luminaire is important for obtaining the light distribution required by the user and for minimizing unwanted effects such as glare and stray light. This calls for a coordinated design of lamp and luminaire as in the lighting system for sports arenas shown in Figure 4.

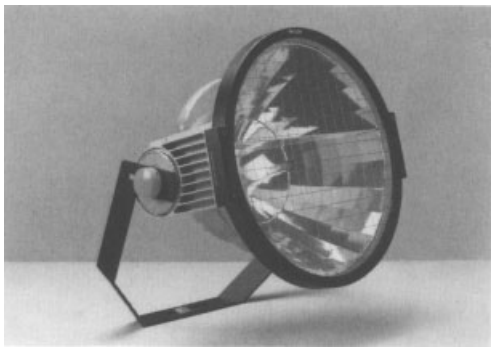

Figure 4. Lamp and luminaire for a lighting system for sports arenas

Ballasts stabilize the current through the discharge in gas discharge lamps. Traditionally, this function has been fulfilled by copper-iron inductive coils [2]. In the last decades the electronic operation of discharge lamps has gained increasing importance [3]. There are several stimuli towards the use of electronics, such as miniaturization, energy saving, control functions, etc. The interaction between a gas discharge and its sta-

bilizing device may be quite complex. To fully exploit the opportunities offered by electronics, an integral system design is required. For example, high-frequency electronic ballasts have been developed for fluorescent lamps. This type of ballast can be used for more sophisticated control functions such as the dimming of lamps according to the amount of daylight and automatic switching off of the lighting system in an empty room.

2. Light Generation

The generation of electromagnetic radiation in the UV, visible, and IR regions of the spectrum involves electrons. In incandescent light sources, the mobile free electrons in the tungsten wire move due to a voltage difference between the ends of the wire. The energy gained by the electrons is transferred to the ionic lattice of the metal by electron – lattice interactions (ohmic heating), which then gives rise to temperature increase coupled with the emission of radiation. Incandescent emission can be described satisfactorily starting from Planck's formula for blackbody radiation (Section 2.1).

For gas discharge lamps, more detailed considerations are required. In the conducting plasma the free electrons are accelerated by the voltage difference between the electrodes. The high-energy electrons collide with gas atoms, leading to excitation of the atoms and subsequent emission of radiation (Section 2.2).

With semiconductor sources (LEDs) the radiation is generated by electron – hole recombination (Section 2.3).

2.1. Thermal Radiation from Solids

The laws of physics underlying the generation of light from solid bodies were proposed around 1900 by PLANCK. Of central importance is blackbody radiation–the radiation present in any region of empty space in thermodynamic equilibrium. Consideration of blackbody radiation yields the maximum amount of radiation that can be emitted by hot solid (i.e., incandescent) bodies. The blackbody radiation $M_0(\lambda, T)$ emitted into a hemisphere per unit area of an ideal surface with an absolute temperature T in the wavelength interval d λ is given by

$$M_0\,(\lambda,T)\,\mathrm{d}\lambda = \frac{2\pi hc^2}{\lambda^5}\,\frac{\mathrm{d}\lambda}{\mathrm{e}^{(hc/\lambda kT)}-1}$$

where $M_0(\lambda, T)$ is expressed in watts per cubic meter, h is Planck's constant (6.63×10^{-34} J · s), k is Boltzmann's constant (1.38×10^{-23} J/K) and c is the velocity of light (3.0×10^{8} m/s). Differentiation of $M_0\,(\lambda, T)$ with respect to T gives the value of the wavelength $\lambda_{\max}$ at which $M_0(\lambda, T)$ has its maximum. This value is given by Wien's law, which states that

$$\lambda_{\max}T = 2.9\times10^{-3}\,\mathrm{m\cdot K}$$

These relationships are illustrated in Figure 5. The value of T at which $\lambda_{\max}$ coincides with the maximum of the spectral luminous efficiency (i.e., 550 nm) is approximately 5500 K, which is about the surface temperature of the sun. Integration of $M_0\,(\lambda, T)$ yields the total radiation $M_0\,(T)$ emitted into the hemisphere per unit area

$$M_0\,(T) = \frac{2\pi^5k^4}{15h^3c^2}T^4$$

which is known as the Stefan – Boltzmann law.

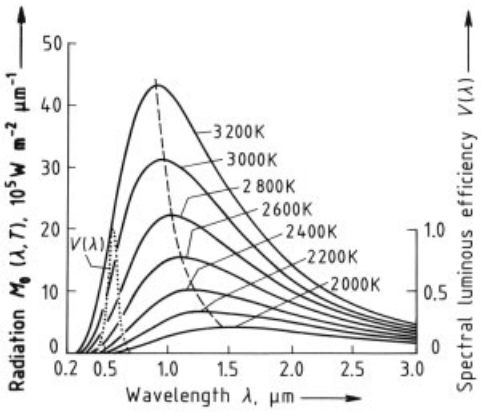

Figure 5. The blackbody radiation M emitted per micrometer of wavelength difference and per unit area into a total hemisphere at various temperatures
The broken line connects the maxima of the curves (Wien's law). The spectral luminous efficiency $V\,(\lambda)$ is also shown.

The thermal radiation $M\,(\lambda, T)$ from surfaces of real materials can be described as

$$M\,(\lambda,T) = \varepsilon\,(\lambda,T)\,M_0\,(\lambda,T)$$

where $\varepsilon(\lambda, T)$ is the emissivity or the emission coefficient of the surface and equals unity for a blackbody radiator. The emissivity equals the absorption coefficient of the radiating body (Kirchhoff's law). A radiator is called "grey" if ε is independent of λ and smaller than unity. If ε depends strongly on λ one speaks of a selective radiator.

The luminous efficacy $\eta(T)$ (expressed in lumen per watt) of a radiator can be written as

$$\eta(T) = \frac{\int_0^\infty V(\lambda)\,\varepsilon(\lambda,T)\,M_0(\lambda,T)\,\mathrm{d}\lambda}{\int_0^\infty \varepsilon(\lambda,T)\,M_0(\lambda,T)\,\mathrm{d}\lambda + \psi}.$$

The term ψ in the denominator represents the energy losses (due to heat conduction etc.) that occur in a real lamp. The function $\eta(T)$ is depicted in Figure 6 for a radiator with constant ε (i.e., a black or grey radiator) and no energy losses (i.e., $\psi = 0$). The form of the function $\varepsilon(\lambda, T)$ may strongly influence the behavior of $\eta(T)$. Metals generally have a decreasing emissivity in the infrared, thus enhancing the luminous efficacy; the luminous efficacy curve for a tungsten surface is also given in Figure 6. Attempts to replace tungsten in incandescent lamps by more selective materials have failed.

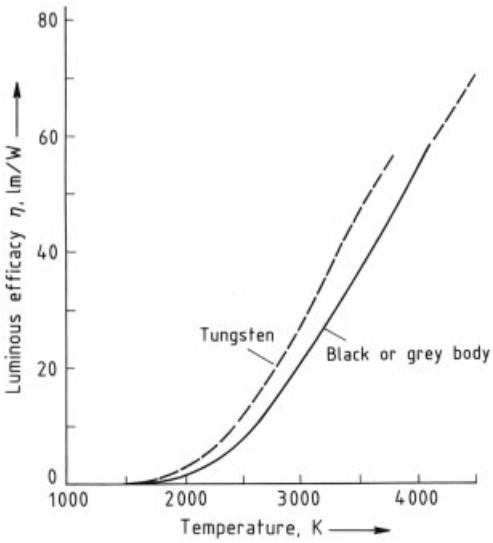

Figure 6. Luminous efficacy η as a function of temperature for a black or grey radiator and a tungsten radiator with a smooth surface

The power balance of a typical incandescent lamp is shown in Figure 7; only a relatively small percentage of the input power (8 %) is radiated in the visible wavelength range.

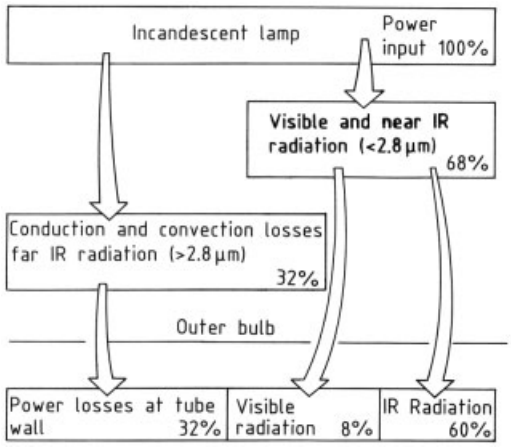

Figure 7. Power balance for an incandescent lamp

2.2. Radiation from Gas Discharges [2], [4]

A metal wire can be considered as a stable plasma consisting of mobile electrons in a lattice of positive metal ions. A gaseous plasma is not stable and requires continuous energy input to prevent it from decaying to a neutral gas. In the plasma column of a gas discharge there is continuous formation and decay of charged particles. The free electrons acquire energy from the electric field between the electrodes. Atoms and molecules are excited and ionized via nonelastic collisions involving the electrons. In the steady state, an equilibrium exists between ionization and the recombination of electrons and ions.

Two types of discharge can be distinguished and discussed in terms of electron temperature T_e and gas temperature T_g. The accelerated electrons can be characterized by a kinetic temperature T_e, often called the plasma temperature. The energy transfer in the discharge is then described as a heat transfer from the electron energy reservoir with temperature T_e to the kinetic energy reservoir of the gas with temperature T_g. Depending on the frequency and effectiveness of the collisional processes, low-pressure discharges occur when $T_e \gg T_g$ and high-pressure discharges when $T_e \geq T_g$. The emphasis on pressure comes from the importance of collisional

frequencies, which are directly related to the gas pressure.

In *high-pressure discharges* (for typical pressures, see Table 2), energy transfer is very effective due to the strong interactions of the various plasma constituents. Recombination of electrons and ions takes place mainly in the gas. These plasmas are close to thermodynamic equilibrium, which allows a physical description by assuming so-called local thermodynamical equilibrium conditions (LTE).

In *low-pressure discharges* (for typical pressures, see Table 2) the interactions between the plasma constituents are weak and recombination mainly takes place at the wall of the discharge vessel, so that energy transfer is small. These systems are very far from equilibrium. A physical description therefore requires a detailed consideration of the rate equations for the atomic and molecular processes involved. Table 2 gives a survey of some typical conditions in the important discharges for light production. Since all discharges are enclosed in vessels of transparent or translucent materials with a maximum working temperature of 1500 K, extremely large temperature gradients exist (of the order of 1000 K/mm), especially in high-pressure discharges.

The radiation emitted by discharges is mainly line radiation, originating from excited atoms. In some discharges, band radiation from excited molecules or continuous radiation from electron – ion recombination (e.g., in the xenon discharge) may contribute significantly. Figure 8 gives the atomic energy levels of sodium and mercury atoms. In low-pressure discharges the radiation is almost exclusively emitted by electronic transitions from the lowest excited states to the ground state (resonance radiation). For the low-pressure mercury discharge these are the 254 and 185 nm lines; for the low-pressure sodium discharge these are the Na-D lines at 589 and 589.6 nm. In high-pressure discharges a large proportion of the radiation comes from transitions involving higher energy levels; due to the high collision frequencies the emission lines display more line broadening than those from low-pressure discharges.

In order to understand the spectral power distribution from high-pressure lamps the importance of radiative transport should be emphasized. Consider a plasma column of cylindrical symmetry (Fig. 9). Emission and absorption of radiation take place in all parts of the plasma column. Thus, the radiation observed outside the discharge by a detector or the human eye represents an integral over the optical path in the plasma.

An important distinction can be made between optically "thin" and optically "thick" radiation. If absorption is small (i.e., optically thin), the observed emission contour $I(\lambda)$ resembles that depicted in Figure 9 A. If the absorption is large, the resulting line contour is strongly broadened and at strong absorptions the phenomenon of self-reversal takes place (Fig. 9 B). The spectral features of the high-pressure sodium lamp are largely due to line broadening and self-reversal. In this discharge the temperature at the discharge center is about 4000 K, while at a distance of only a few millimeters the temperature of the envelope wall is about 1500 K. In simplified terms, a core of "hot" emitting sodium vapor is surrounded by a "cold" mantle of absorbing sodium vapor. The phenomenon of self-reversal occurs because the radiation generated in the core is strongly absorbed in the outer mantle, especially close to the wall. Figure 10 shows the characteristic spectral power distributions of some gas discharge lamps.

The power balance of a typical fluorescent lamp is shown in Figure 11. The low-pressure mercury discharge is relatively effective in converting electrical power into radiation but overall only 25 – 30 % of the input power is radiated in the visible wavelength range. For the high-pressure sodium lamp about 30 % of the input power is converted into visible radiation (Fig. 12).

2.3. Radiation from Semiconductors (LEDs)

A semiconductor can have intentional impurities added, which result in excess electrons (*n*-type semiconductor), or the absence of electrons known as holes (*p*-type semiconductors). A $p-n$ junction occurs when a *p*-type and *n*-type semiconductor are brought together. LEDs are basically $p-n$ junctions (see Fig. 13). The $p-n$ junction acts as diode. When the junction is biased with enough voltage in the appropriate direction, current can flow whereby electrons are

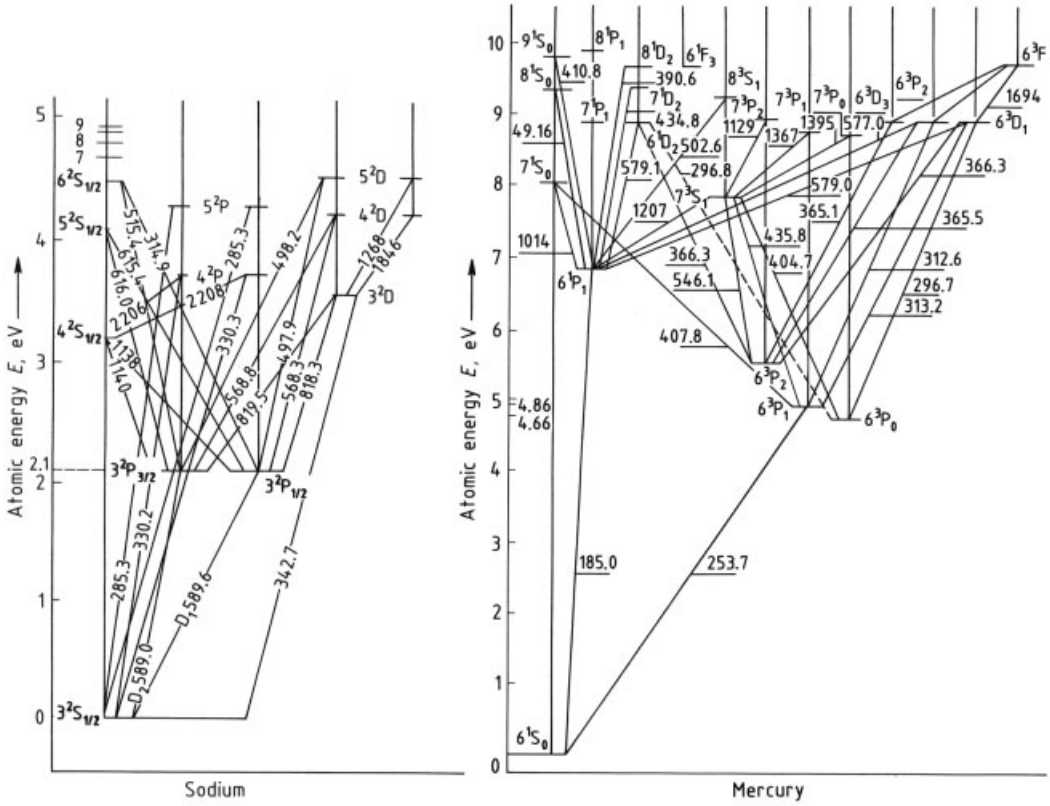

Figure 8. Atomic energy levels of sodium and mercury atoms
The wavelength is given in nanometers.

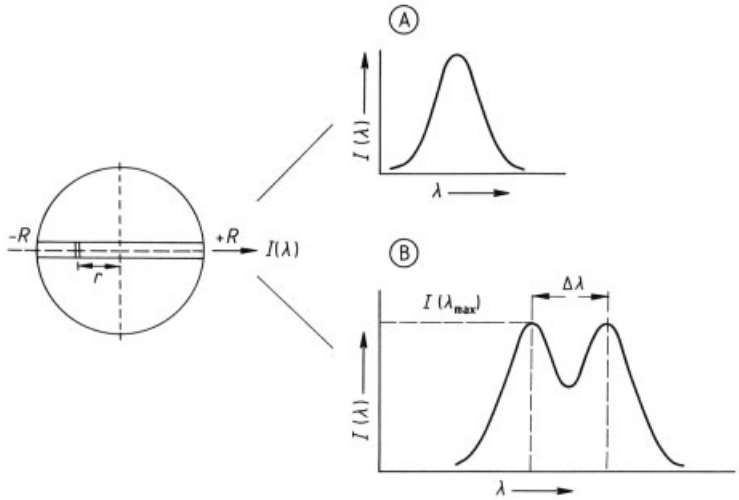

Figure 9. Radiative transport in a cylindrical plasma column
The observed emission $I(\lambda)$ is primarily an integral of the local emissions over the path from $-R$ to $+R$. The radiation generated in the volume element at coordinate r, however, is attenuated by absorption on its way from r to R. The actual expression for $I(\lambda)$, therefore, is an integral equation involving emission and absorption contributions. The optically thin (A) and optically thick (B) line contours are indicated; the latter shows the self-reversal phenomenon.

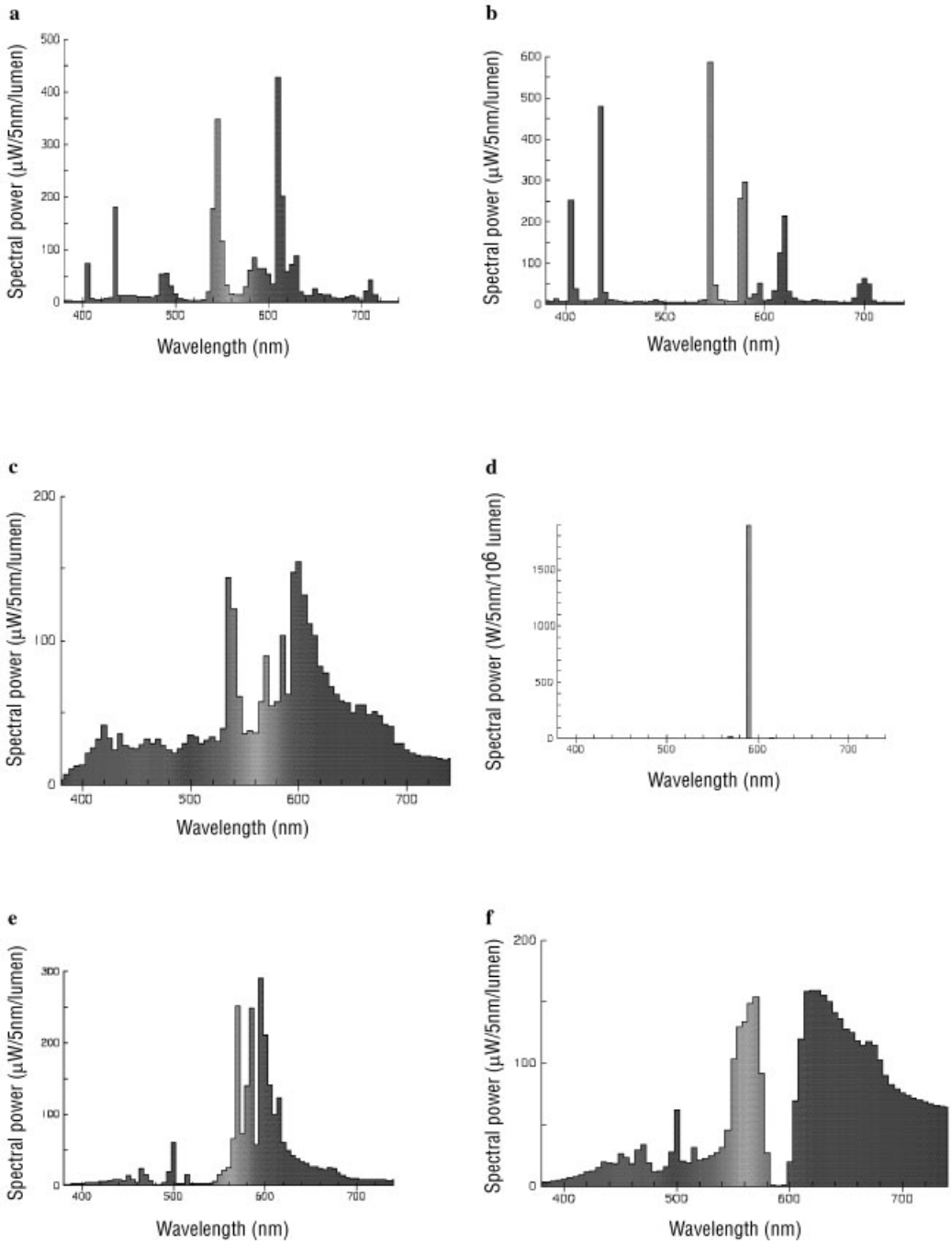

Figure 10. Characteristic spectral power distributions of discharge lamps. In all lamps, except for the fluorescent lamp, the spectrum is directly emitted by the discharge. Note the self-reversal of the Na-D lines (590 nm) in the spectra of the high-pressure sodium lamps (e, f). In the fluorescent lamp (a), the Hg resonance radiation at 185 and 254 nm from the discharge is converted to visible radiation by the fluorescent powder (or phosphor). The spectral distribution in the visible is thus determined by the particular phosphor or phosphor mix.

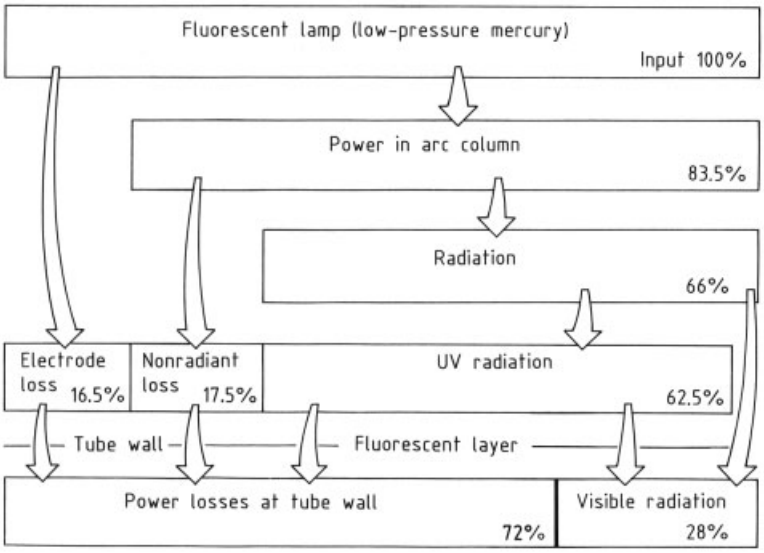

Figure 11. Power balance of a fluorescent lamp

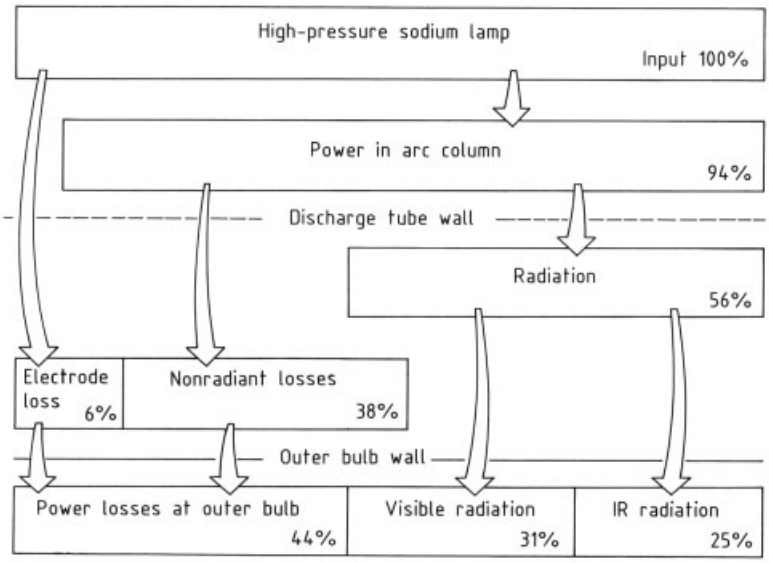

Figure 12. Power balance of a high-pressure sodium lamp

injected into the p-type region where they can recombine with holes. This recombination results in the release of energy in the form of a photon. The wavelength of the photon is determined by the semiconductor energy gap E_g ($\lambda = hc/E_g$). The semiconductor materials used for spectral line radiation at various wavelengths are summarized in Table 1. It can be seen that LEDs with relatively high performance can be fabricated throughout the visible spectrum.

Table 1. Some data on (commercially available) light emitting diodes

Material	Peak wavelength, nm	Typical luminous efficacy at 20 mA, lm/W
AlGaAs	650	8
AlGaInP	560 – 650	22 at 590 nm
		30 at 616 nm
		27 at 628 nm
InGaN	470 – 535	6 at 470 nm
		15 at 505 nm
		22 at 525 nm

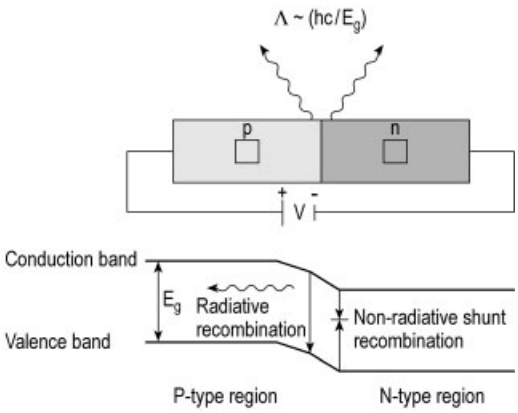

Figure 13. Schematic picture of light generation in a $p-n$ junction

White light (i.e., light with an x,y-point on the black body locus, cf. Fig. 3) can in principle be obtained by appropriate combination of blue, green, and red narrow-band emitting LEDs. An essential breakthrough in the generation of white light came in 1995 with the introduction of the blue LED (InGaN). On combination of the blue LED with a fluorescent powder (see Section 5.4) an all-solid-state white light source could be made (white LEDs).

In principle, an efficiency of 100 % for the energy conversion in LEDs can be reached. However, practical difficulties such as material defects, and the problem of extracting the light from the semiconductor chip, result at present in typical conversion efficiencies of the order of 10 %. Improvements in efficiencies are expected to continue as the technology for materials growth and processing continues to evolve.

2.4. Modeling of Lamps

The development of light sources has a long history as an empirical science. During the last decades, however, much effort has been spent in the development of lamp models involving numerical solution of the energy balance, set up in terms of the basic physical mechanisms. Computer programs are now available for modeling both discharge and incandescent lamps with various degrees of completeness, sophistication and accuracy. Although the use of such models is very important as a guide in the development of new and improved lamp types, it has supported rather than replaced the empirical methods used in lamp development.

3. Main Types of Lamps [5]

In this section emphasis is placed on the construction of the main types of light sources; Tables 2 and 3 list important properties. Materials and production processes are described in detail in Chapters 5, 6, 7. Up-to-date reviews on all types of lamps discussed here can be found in the Symposium Proceedings of the International Symposium on the Science and Technology of Light Sources (LS), Kyoto 1995 (LS-7) and Greifswald 1998 (LS-8).

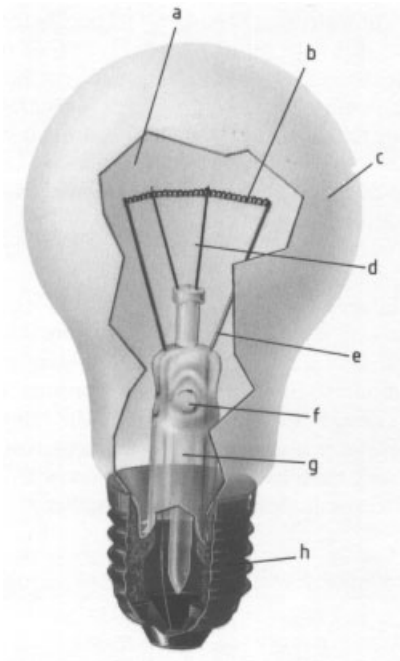

Figure 14. Schematic diagram of an incandescent lamp
a) Fill gas; b) Filament; c) Bulb; d) Support wires; e) Lead-in wires; f) Stem; g) Fuse; h) Lamp cap

Incandescent Lamps. The heart of the incandescent lamp is a tungsten filament that is coiled either singly or multiply and enclosed in a hermetically sealed glass envelope (Fig. 14). A special copper-clad (dumet) wire is used for the glass-to-metal seal. In order to prevent oxidation of the tungsten filament and tungsten transport via the water cycle (see Section 4.2), the impurity level in the glass vessel must be very low.

Table 2. Properties of typical high-pressure and low-pressure sodium and mercury discharge lamps

Property	High-pressure sodium	High-pressure mercury	Low-pressure sodium	Low-pressure mercury (fluorescent)
Lamp power, W	400	400	130	36
Discharge tube material	polycrystalline alumina	quartz glass	soda lime glass and borate layer	soda lime glass
Discharge tube inner diameter, mm	7.5	19	19	24
Electrode spacing, mm	82	72	1890	1120
Gas/vapor pressure in operating lamp				
Na, Pa	1×10^4		0.5	
Hg, Pa	8×10^4	4×10^5		1
Noble gas pressure (Ar, Kr, Ne, Xe), Pa	2×10^4	2×10^4	730	200
Lamp current, A	4.4	3.2	0.62	0.44
Lamp voltage, V	105	140	245	102
Discharge input power per unit length, W/m	4500	5500	60	30
Electron density at axis, m^{-3}	$10^{21}-10^{22}$	$10^{21}-10^{22}$	$10^{18}-10^{19}$	$10^{18}-10^{19}$
Gas temperature at axis, K	4000	6000	530	320
Electron temperature at axis, K	4000	6000	10 000	13 000

Table 3. Important lamp parameters

Lamp type	Power, W	Luminous flux, lm	Luminous efficacy, lm/W	Color temperature, K	Color rendering index	Lamp life, h
Incandescent						
general	10 – 1500	50 – 30 000	5 – 20	2500 – 3000	100	100 – 2000
halogen	10 – 2000	150 – 50 000	15 – 25	2800 – 3300	100	50 – 4000
Fluorescent	4 – 200	150 – 15 000	40 – 105	2500 – 6500	60 – 95	6000 – 20 000∗
High-pressure mercury	50 – 1000	1500 – 60 000	35 – 60	3000 – 6000	20 – 60	10 000 – 20 000
Metal halide	35 – 6000	1500 – 600 000	50 – 110	3000 – 7000	60 – 95	300 – 20 000
Low-pressure sodium	15 – 200	1500 – 35 000	100 – 200	1800		10 000 – 15 000
High-pressure sodium	35 – 1000	1500 – 130 000	50 – 150	2000 – 2500	20 – 80	10 000 – 30 000

∗With electrodeless lamps, lamp lives of more than 50 000 h are reached.

Substances known as getters are used to absorb traces of water (see Section 6.1). Usually the vessel is filled with an inert gas (Fig. 14, a) to decrease the evaporation of tungsten from the filament. A more diffuse light source is obtained by etching the inside of the bulb or coating it with a powder layer of TiO_2 (Chap. 7). Colored lamps can be made by using pigments in the powder layer or an exterior heat-resistant paint. The lamp cap is attached to the vessel by means of a cement.

Halogen Incandescent Lamps. In halogen incandescent lamps, the inert gas pressure is higher than in a normal incandescent lamp to further diminish the evaporation of tungsten. For a given design life this means that the lamp can operate at a higher tungsten filament temperature and thus has a higher luminous efficacy. Due to the higher pressures a smaller vessel is necessary. Use of fused silica (commonly called quartz or quartz glass) in place of glass allows the vessel to withstand higher temperatures and pressures. Halogens are used to keep the quartz vessel free from tungsten which evaporates from the filament (see Section 4.2.3). A typical halogen incandescent lamp is shown in Figure 15. Efficacies of halogen lamps can be increased by the use of coatings on the quartz surface, which reflect the infrared radiation back to the filament (see Chap. 7).

The tubular fluorescent lamp (Fig. 16) consists of a glass tube with electrodes sealed in at each end. These coiled tungsten electrodes are coated with a metal oxide mixture which improves electron emission from the electrodes

(see Section 6.3). The lamp is based on the gas discharge in a low-pressure mercury vapor (≈ 1 Pa) and rare gas (≈ 400 Pa) which primarily generates UV mercury radiation. The inner surface of the tube is coated with a fluorescent powder which converts the generated UV radiation into light (see Fig. 10 a). By varying the composition of the fluorescent powder, lamp types with different color appearance and color rendering properties can be made.

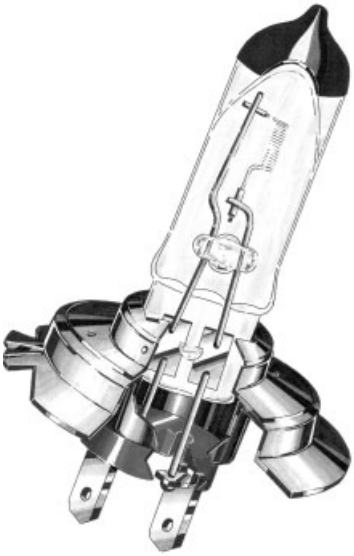

Figure 15. Typical halogen incandescent lamp: Automotive head lamp, including mounting construction

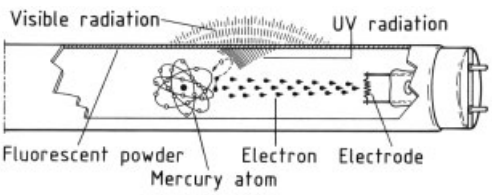

Figure 16. Schematic diagram of a fluorescent lamp

Fluorescent (Low-Pressure Mercury) Lamps. In compact fluorescent lamps, the relatively long tube must be folded because simple shortening of the tube leads to an inefficient gas discharge. Examples of compact fluorescent lamps are shown in Figure 17.

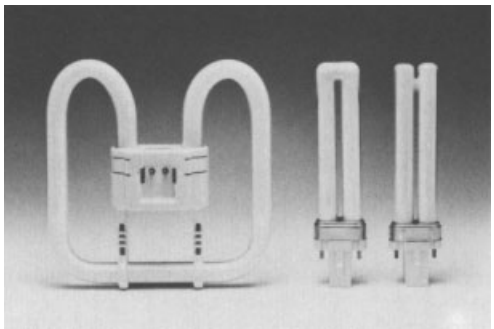

Figure 17. Examples of compact fluorescent lamps

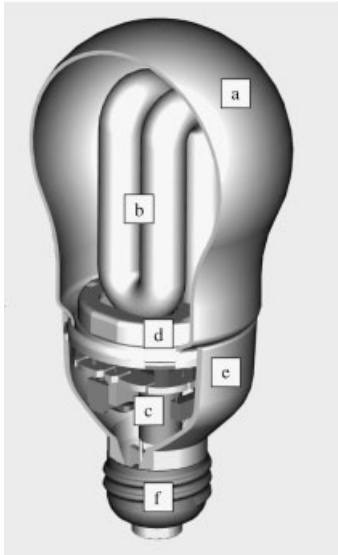

Figure 18. An open view of a typical compact fluorescent lamp
a) Outer bulb; b) Discharge tube; c) Ballast/starter circuitry; d) Mounting plate; e) Housing; f) Lamp cap
The narrow discharge tube is multiply folded and integrated into a single unit with the ballast and starter.

As in almost all gas discharge lamps, the lamp voltage of fluorescent lamps decreases as the current increases. This means that a current limiter, usually called a *ballast,* must be placed between the lamp and the mains supply. Also the mains voltage is usually insufficient to ignite the gas discharge so a *starter* or *ignitor* is required. In the regular case, where the ballast is a conventional Cu – Fe coil, this ignitor can be a *glow switch* [2], [3]. The glow switch is a small gas-discharge device, which generates a high-voltage induction peak from the ballast. The mechanism of peak-voltage generation is

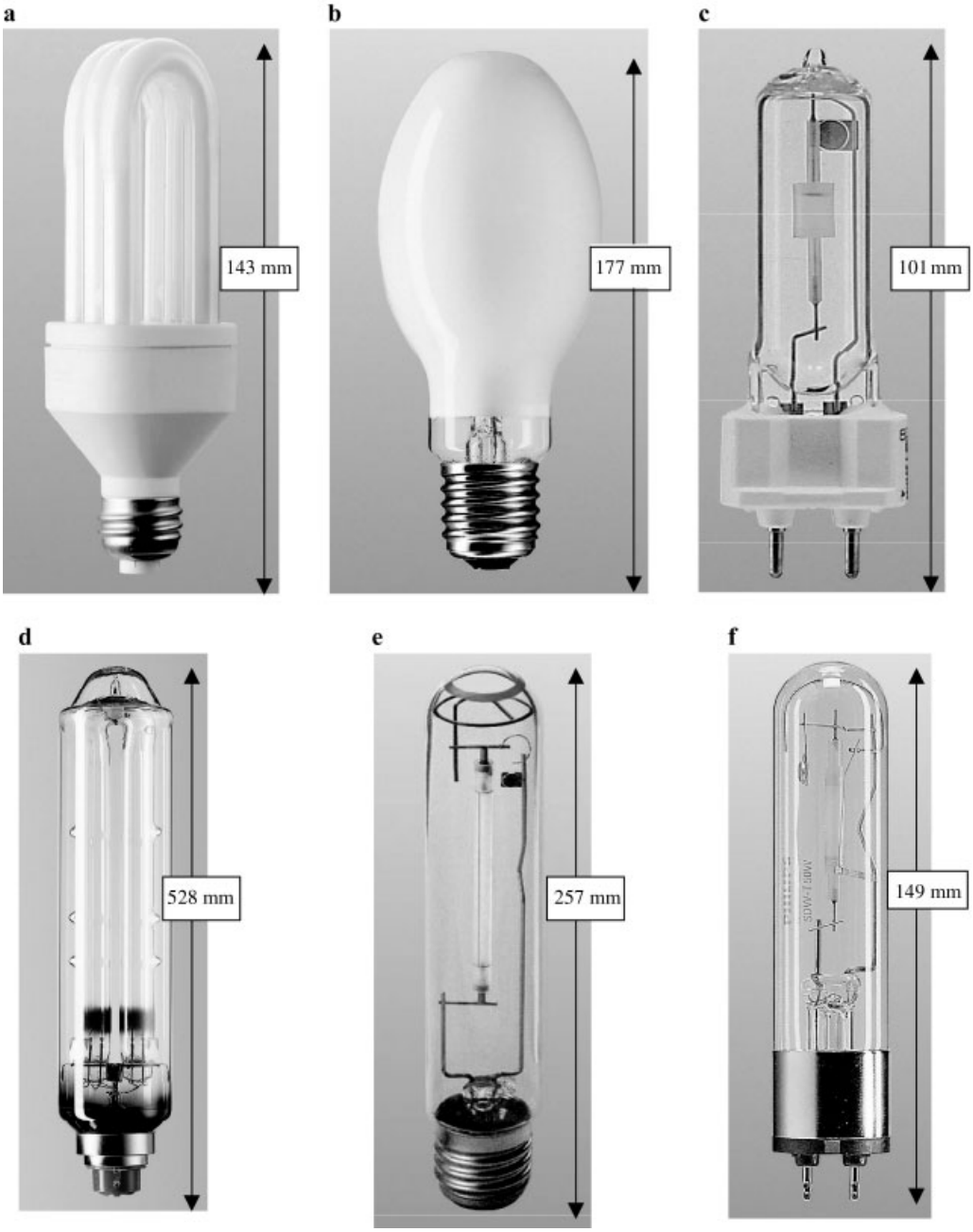

Figure 19. Examples of discharge lamps. Note the indicated longitudinal dimensions

interruption of a ballast short-circuit situation. This is caused by thermal closing of the glow-switch bimetal electrodes, as a result of its initial glow discharge. This causes an electrical breakdown in the gas and consequently the creation of a conductive discharge plasma in the lamp. However, with the increasing use of electronic ballasts, the ignition function is usually integrated in the ballast circuitry. The starter and the ballast are often separate from the lamp but may also be integrated into a single unit (see Figs. 18 and 19 a), which can be directly connected to the mains. In such an integral lamp the mercury vapor pressure is very high due to the high wall temperature and an amalgam is used to obtain the optimum mercury vapor pressure (see Section 6.2). A more recent development in low-pressure mercury fluorescent lamps is the electrodeless concept. All major manufacturers now offer fluorescent lamp systems which are based on (high-frequency) inductive coupling of electrical energy into the lamp. The advantage of these electrodeless lamp systems is that the lifetime can be significantly extended compared to lamps with electrodes, where the electrodes usually are the lifetime limiting factor. (see Section 6.3).

Low-Pressure Sodium Lamp. Like the fluorescent lamp, the low-pressure sodium lamp (Figs. 20 and 19 d) is based on a low-pressure discharge but in a sodium – rare gas mixture rather than mercury – rare gas. The fact that sodium is used in place of mercury has significant consequences. Since sodium radiates in the visible region (see Fig. 10 d) a fluorescent powder is not needed, but to obtain a sodium pressure of about 0.5 Pa the gas discharge tube must be at a temperature of about 260 °C. To minimize heat losses, the discharge tube (Fig. 20, a) is folded and placed in an evacuated outer bulb (b) with a getter to ensure a good vacuum throughout life (see Section 6.1). The inside of the outer bulb is coated with a film (typically indium oxide with a tin dope) which reflects in the near infrared and is transparent in the visible (see Chap. 7). At 260 °C sodium attacks normal soda lime glass so the inside of the gas discharge tube is coated with a thin film of sodium-resistant borate glass. The dimples (d) prevent the sodium from forming a film over the whole discharge tube surface.

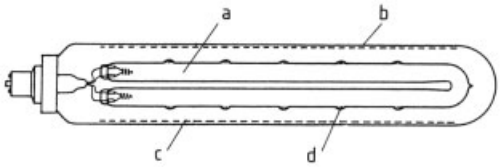

Figure 20. Schematic diagram of a low-pressure sodium lamp
a) Discharge tube; b) Outer bulb (evacuated); c) Indium oxide film; d) Dimples filled with sodium.
For typical dimensions see Table 2.

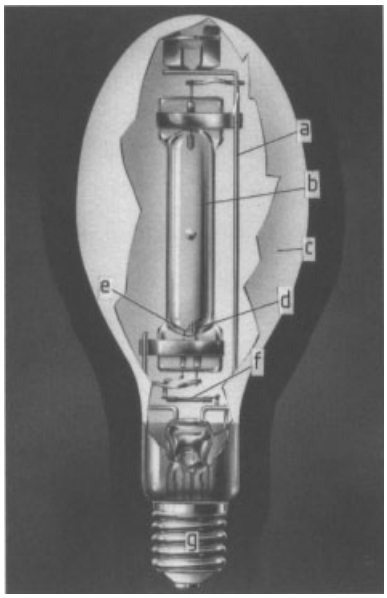

Figure 21. Schematic diagram of a high-pressure mercury lamp
a) Support for discharge tube; b) Discharge tube; c) Outer bulb; d) Main electrode; e) Auxiliary electrode; f) Resistor; g) Lamp cap

High-Pressure Mercury Lamp. The high-pressure mercury lamp (Figs. 21 and 19 b), is based on an arc discharge in mercury vapor with a typical pressure of a few hundred kilopascal. The presence of a rare gas at low pressure facilitates ignition. The discharge is compact with a high energy density. In order to withstand the high wall temperatures, the discharge vessel (Fig. 21 b) is made of quartz glass. The tungsten electrodes (d, e) are coated with an emitter material and are sealed into the quartz discharge vessel using molybdenum foil seals. This arc tube assembly is placed in a vacuum or gas-

filled outer bulb (c) which can be clear, frosted, or coated with a fluorescent powder. The high-pressure mercury discharge generates visible radiation and some UV radiation (Fig. 10 b). The latter can be converted to visible radiation by the fluorescent powder on the outer bulb to improve the color appearance and color rendering. A recent development based on the high-pressure mercury discharge is a 100 W lamp for LCD (liquid crystal display) projection applications (Fig. 22). This lamp combines an ultra-high mercury pressure (20 MPa) with a very short arc length (1 – 1.5 mm). The extreme power density in the short arc discharge leads to high arc temperatures and luminance (brightness). In addition a chemical tungsten transport cycle (cf. Section 4.2) is operative. This leads to perfect stability of the light output during lamp lives of more than 10 000 h.

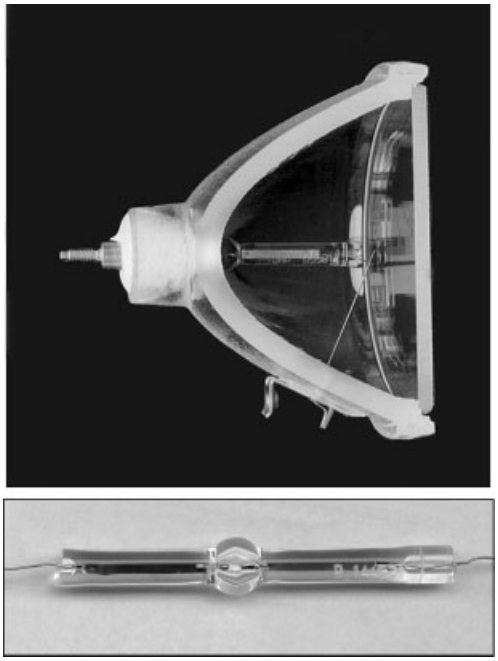

Figure 22. Reflector assembly and burner of the ultra-high pressure mercury lamp for LCD projection

High-Pressure Sodium Lamp. This lamp is based on a high-pressure sodium discharge where the sodium vapor pressure is typically 10 kPa. Normally, mercury is also present with a pressure of ca. 80 kPa. A rare gas facilitates ignition. Under operating conditions Na and Hg are present in the form of a liquid amalgam, the temperature of which determines their actual vapor pressures in the lamp. Due to the high pressure the yellow sodium-D lines are broadened (Fig. 10 e), producing a golden-yellow light rather than the monochromatic yellow produced by the low-pressure sodium lamp (Fig. 10 d). The production of this type of lamp (Figs. 23 and 19 e) only became possible after the development of a material that was resistant to the high sodium pressures at high temperatures ($\approx$ 1500 K). The arc tube consists of translucent polycrystalline alumina (PCA, see Section 5.3). The electrodes are sealed in the PCA tube, using a niobium tube and a sealing frit. The arc tube assembly is mounted in an outer bulb which is usually evacuated, but it can also be filled with an inert gas. The outer bulb contains a getter (see Section 6.1).

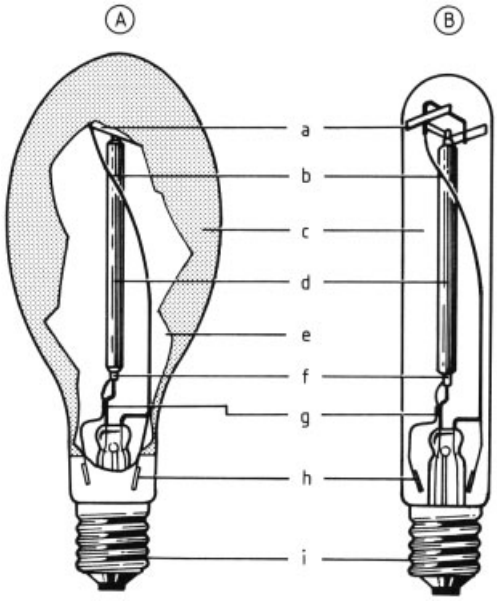

Figure 23. Schematic diagram of a diffuse ovoid (A) and a clear tubular (B) high-pressure sodium lamp
a) Support springs to maintain discharge tube alignment; b) Lead-in wire; c) Hard-glass outer bulb; d) Translucent polycrystalline alumina discharge tube; e) Inner diffusing coating; f) End cap of discharge tube; g) Lead-in wire, support; h) Getter rings for maintaining high vacuum; i) Screw base

Since the 1980s, several modifications of the basic high-pressure sodium lamp concept have been developed and commercialized. The use of a relatively high rare gas (xenon) pres-

sure (typically 200 kPa under operating conditions) has led to an efficacy increase of around 15 %. Furthermore, lamp types with increased sodium vapor pressure (up to about 100 kPa) are available. Under these conditions strong spectral line broadening occurs (cf. Fig. 10 f) leading to the emission of white light. Consequently, the color properties are greatly improved. Finally, environmental considerations have stimulated the development of high-pressure sodium lamps with reduced amounts of Hg. This has resulted in the development of lamp types operating under unsaturated vapor conditions . Recently, even mercury-free lamps types with only a sodium – xenon mixture have been introduced.

Metal Halide Lamp. The discharge vessels for metal halide lamps are made either from quartz glass or from polycrystalline alumina (PCA). The metal halide lamp with quartz vessel is basically similar in construction to a high-pressure mercury lamp, the major difference being that the discharge tube contains one or more metal halides in addition to mercury. The use of metal halides has two advantages:

1) The vapor pressure of the metal halide is usually higher than that of the metal.
2) The metal halides are chemically less aggressive toward quartz glass than the metals.

The halides are partially or wholly vaporized when the lamp reaches its normal operating temperature. The halide vapor dissociates in the hot central region of the arc into the halogen and the metal, and the vaporized metal radiates its characteristic spectrum. The tungsten halogen cycle also operates in some of these lamps. Various metal halide systems have been developed, examples being Na, In, and Tl iodides, Na and Sc iodides, Na and Tm iodides, Dy, Ho, and Tm iodides, and Sn iodides and bromides.

Low-power quartz metal halide lamps have found application in automotive headlighting (Fig. 24). The compact discharge tubes offer a significant improvement in optical performance as compared to halogen lamps. The color requirements for this application are met by the sodium – scandium iodide system. In general, it takes some time for metal halide lamps to arrive at full luminous intensity, due to the vaporization rate of the halide salts. To meet the severe requirements with respect to immediate availability of light upon switching, high-pressure xenon (600 kPa) is dosed. With appropriate ignition and run-up electrical power, a Xe discharge generates the initial light, which is then taken over by the Na – Sc discharge.

Figure 24. 35 W metal halide automotive headlight lamp; the arc length is of the order of 4 mm

A recent development is the so-called *ceramic* metal halide lamp (Fig. 19 c). Instead of a quartz glass tube, a discharge tube of polycrystalline alumina is used. PCA allows discharge vessels with higher wall temperatures than quartz glass. Therefore, higher metal halide vapor pressures can be reached, which result in improvements in efficacy and color properties. A typical metal halide system in these lamps consists of Na, Tl, and Dy iodides. Ceramic metal halide lamps became commercially available in 1993.

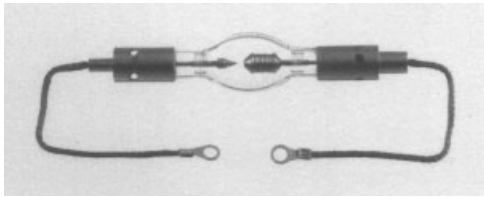

Figure 25. Xenon arc lamp

Other Lamp Types. Several other lamps are in use, mostly for specific applications. Some examples are discussed briefly below. An important category of lamps is based on Xe dicharges. The *xenon arc lamp* (Fig. 25) is based on the high-pressure xenon discharge while the *xenon flash lamp* (Fig. 26) is based on a low-pressure xenon discharge. A lamp of possible future importance is the Xe-excimer fluorescent lamp.

This lamp is based on a dielectric-barrier discharge, where excited Xe atoms recombine to Xe_2* excimer molecules, which emit UV radiation at 172 nm. Via an appropriate fluorescent powder visible light can be generated.

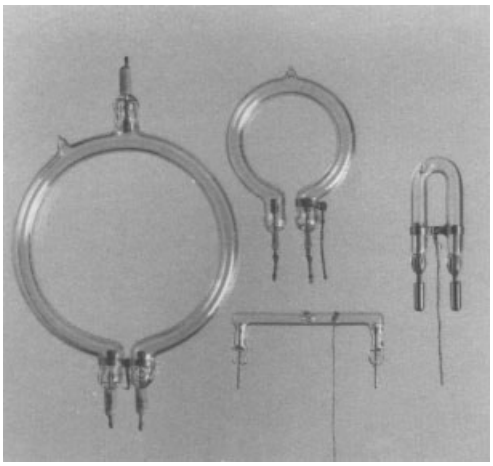

Figure 26. Xenon flash lamps

A *low-pressure neon discharge* with its specific red appearance has found application in automotive stoplights.

In the *super high-pressure mercury lamp*, the mercury vapor pressure and the wall loading are extremely high and water cooling of the discharge tube is necessary.

The *carbon arc lamp* and the *zirconium metal combustion flash lamp* are now virtually obsolete. The carbon arc is a high-pressure discharge in air between carbon electrodes. In combustion flash lamps, a light flash with a duration of about 30 ms is generated by the combustion of shreds of zirconium foil in oxygen at pressures of ca. $10^5 - 10^6$ Pa.

A recent discovery is the microwave-excited sulfur discharge lamp. The radiating species is the UV emitting S_2 molecule. Due to radiative transport the wavelength of the emitted radiation is strongly shifted to the visible region. The sulfur discharge itself is very efficient (≈ 170 lm/W). Due to the poor efficiency of microwave generation, the system efficiency is substantially lower.

Light Emitting Diodes. Schematic diagrams of typical LED packages are shown in Figure 27. Basically, the light-generating semiconductor chip is mounted on a carrier with metallic leads and protected by an epoxy dome, which also is instrumental in the outcoupling of the radiation. Typical power inputs per LED vary in the range 50 – 1000 mW. For luminous fluxes comparable to, e.g., incandescent lamps of several tens of watts, therefore, assemblies of LEDs are to be used. At present, the main applications of LEDs are in the signaling area. Due to the trends in efficacy improvement, thermal management, and LED-phosphor systems, together with aspects like long life and reliability, it is expected that LEDs will penetrate into various white-light applications. Present experimental results point into the direction of devices with increased efficacy values as compared to the present alternatives.

4. Lamp Chemistry [6]

4.1. General

The generation of light is mainly based on physical processes. Lamp production, however, depends primarily on the availability of appropriate materials. Light sources can be formulated as: "A variety of chemicals at high pressures and high temperatures during (many) thousands of hours"–a situation which has to be met by the combination of appropriate materials and control of chemical reactions over a long period.

Control of chemical reactions is a key factor in lamp development and engineering. Chemical reactions in the lamp system must be minimized during the required life-times. Chemically induced changes in, for instance, gas or plasma composition, transparency or translucency of the envelope, electron emissive power of the electrodes, or properties of the phosphors should be as small as possible. In this connection, materials processing during manufacturing is of vital importance.

Lamp chemistry covers a broad field since a variety of physicochemical phenomena take place in lamps [6] and belong to such different domains as plasma chemistry, photochemistry, high-temperature chemistry, and solid-state electrochemistry. Above all, however, phenomena related to solid – gas-phase transport chemistry play an important role (Section 4.2).

Chemical processes in lamps can be divided into two categories:

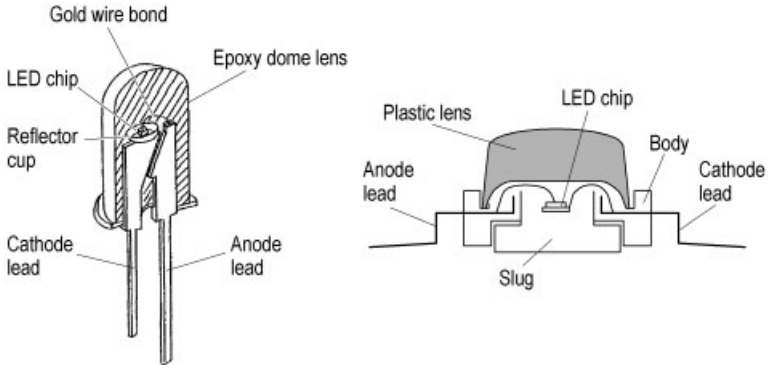

Figure 27. Schematic diagrams of LED packages

1) Processes which are essential for the working mechanism of the lamp. Most important examples here are the chemical transport cycles in halogen incandescent lamps, discussed in Section 4.2.3. Other examples are the dissociation reactions of salt vapors in metal halide lamps.
2) Processes related to lamp maintenance and lamp life. Of special importance in this connection is the presence of gaseous impurities (such as water or oxygen) which should be minimized by proper processing during lamp production. Impurities liberated during lamp burning can be handled by getters (see Section 6.1). Other lamp chemical aspects are discussed in the various sections on materials (Chap. 5).

4.2. Transport Chemistry [7]

4.2.1. Principles

In high-temperature light sources, transport of material via the gas phase affects lamp properties during lamp life both negatively and positively. Two types of transport mechanisms (physical and chemical) can be distinguished.

Physical transport, primarily by evaporation, relates to transport of a compound via its own vapor from a site with a high temperature to a site with a low temperature (e.g., evaporation of tungsten in incandescent lamps).

Chemical transport is a mechanism by which a nonvolatile compound (A) is transported to a site with a different temperature by means of a chemical reaction:

$$A\,(s) + B\,(g) \rightleftharpoons C\,(g)$$

At site I, compound A reacts with the gaseous transport agent B to form the gaseous compound C. The molecules C may reach a site II, for instance by diffusion or convection. If the circumstances at site II are such that C decomposes back into A and B, then A is transported from I to II and transport agent B can take part again in the transfer process. A transport cycle can be formed in which a small quantity of the gaseous transport agent B may cause continuous transport of the nonvolatile compound A. Depending on the reaction, A can be transported to a site with either a lower or a higher temperature.

The rate of chemical transport depends on various factors:

1) *The thermodynamics of the reaction.* Thermodynamic calculations yield the equilibrium vapor pressure of compound C at sites (temperatures) I and II. If substance A is present at both sites, then A will be transferred to the site where the equilibrium vapor pressure of C is lowest. The difference between the vapor pressure of compound C at sites I and II is a measure of the transport rate. The thermodynamics of the transport reaction therefore predict both direction and magnitude of the transport.
2) *Diffusion and convection.* Diffusion is usually slower than convection and offers a way of influencing transport rates.
3) *Kinetics.* The kinetics of the chemical reaction can be so slow that they determine the

rate of transport. For instance, the decomposition of compound C may be inhibited at low temperatures.

High-temperature light sources are complicated chemical systems because they always contain impurities such as oxygen, carbon, and hydrogen which react to form many gaseous compounds. Seldom is a single compound or reaction responsible for chemical transport phenomena. Fortunately in recent decades large thermodynamic data banks combined with powerful computers have become available. Predictions can be made of the direction and magnitude of transport even in complicated systems. Decisive is then the sum of the pressures of all compounds containing A (the element to be transported), each multiplied by a weighting factor equal to the number of atoms of A per molecule of the compound in question: SP (A). Transport will occur from a site with a high value of SP (A) to a site where this parameter is lower.

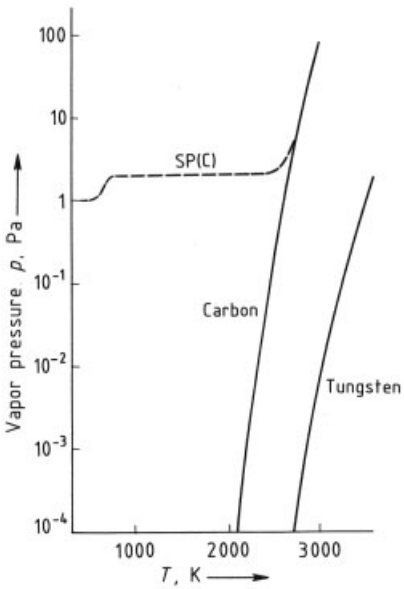

Figure 28. Vapor pressure of carbon and tungsten
The dashed line gives the summed carbon pressure for a carbon filament lamp containing 1 Pa of oxygen.

A simple example is provided by chemical transport in a carbon-filament incandescent lamp. The carbon vapor pressure curve is given in Figure 28. To keep the transport of carbon to the bulb wall low, the filament temperature should not exceed 2400 K, resulting in a rather inefficient lamp. Transport of carbon to the wall (and thus wall blackening) is strongly enhanced by oxygen impurities via the following chemical transport reaction:

$$C + CO_2 \rightleftharpoons 2\,CO$$

The dashed line in Figure 28 shows the SP (C) as a function of the temperature of carbon when 1 Pa of oxygen impurity is present in the light bulb. At 2300 K the vapor pressure of carbon is about 5×10^{-3} Pa whereas the SP (C) (mainly CO) is 2 Pa. At the cold bulb wall the major component is CO_2 and this leads to a SP (C) of 1 Pa. When the oxygen impurity is present, the difference in SP (C) between the filament and the bulb wall is 1 Pa whereas the corresponding figure for an absolutely clean lamp is 5×10^{-3} Pa. Oxygen impurities therefore strongly enhance wall blackening in this type of lamp. Fortunately, because the wall of a vacuum lamp is cold, the decomposition of CO is slow (kinetic limitation). This means the colder the bulb wall, the slower the blackening due to oxygen impurities. Usually phosphorus is used in lamp manufacture as an oxygen getter (see Section 6.1).

4.2.2. Incandescent Lamps

The vapor pressure curve of tungsten is also given in Figure 28. Depending on the inert gas pressure, filament temperatures in the range 2300 – 2800 K can be applied without running into wall-blackening problems caused by physical transport.

In 1913 Langmuir discovered that residual water vapor causes severe wall blackening by means of chemical transport. The reaction responsible for this is:

$$W + 2\,H_2O \rightleftharpoons WO_2 + 4\,H$$

At the filament, tungsten reacts with water vapor to form gaseous tungsten oxide and hydrogen atoms. At the bulb wall, WO_2 condenses and is reduced by the hydrogen atoms; water is formed and can take part again in the transport process. Essential in this transport cycle is that the hydrogen atoms do not recombine to form molecules because molecular hydrogen is not able to reduce tungsten oxides at the cold bulb wall. In vacuum lamps (residual pressure < 0.1 Pa) the

chance of recombination is small, but also in gas-filled lamps sufficient hydrogen atoms can reach the bulb wall to cause a noticeable reduction of tungsten oxides.

The main sources of water contamination are the glass wall and the coatings with a water-rich (silica) powder. Removal of water is usually not complete after manufacture of the lamp. Getters are used to absorb the residual water vapor in the lamps (see Section 6.1).

4.2.3. Halogen Incandescent Lamps

The efficacy and brightness of incandescent lamps can be improved by using higher filament temperatures but this is limited by wall blackening. In halogen incandescent lamps, this problem is circumvented by increasing the inert gas pressure and using halogens to transport tungsten from the bulb wall back to the filament. Although the principle of this method is fairly old, it was not until 1959 that lamps based on a regenerative iodine cycle were introduced. In 1966 these lamps were further improved by using bromine instead of iodine.

Condensation of tungsten on the bulb wall can be prevented if a gaseous tungsten compound is present which is stable at the temperatures that prevail close to the bulb wall. In lamps containing a small amount of iodine this compound is the oxide halide WO_2I_2 (i.e., besides iodine, traces of oxygen are necessary to make the cycle work). In Figure 29 the calculated vapor pressures of the tungsten-containing species for such a lamp are shown. The SP (W) is plotted as a function of the temperature for the situation that the lamp contains 1 Pa of oxygen and 500 Pa of iodine (pressures during operation). Striking is the minimum in the curve at about 2800 K. The filament temperature is usually above 3000 K and because of the steep temperature dependence of SP (W) in that region, tungsten is transported to sites with a temperature around 2800 K (i.e., to the cooler parts of the filament). The criterion for keeping the bulb wall clean is that the value of SP (W) at the bulb wall should be higher than at the filament (cf. Fig. 29). The consequence of the high value of SP (W) between 800 and 2000 K is that tungsten is transported from sites having these temperatures to sites where SP (W) is minimal (2800 K). Hence colder parts of the lamp such as the filament ends, filament supports, and lead-in wires are attacked. Severe thinning of these colder parts may cause end of life. To prevent this the WO_2I_2 pressure near the wall should not be too high. However, if the WO_2I_2 pressure is too low, blackening will occur. Thus, for a particular lamp design (i.e., bulb wall temperature) the iodine should be dosed within (often) narrow limits. Accurate dosing of iodine is one of the major problems in the production of iodine-cycle lamps.

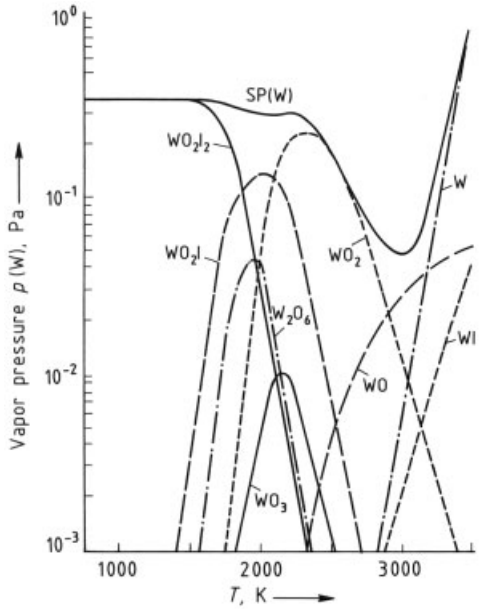

Figure 29. Calculated vapor pressures of various species in an iodine cycle lamp filled with 500 Pa iodine and 1 Pa (residual) oxygen
The bulb wall temperature is 750 K. The summed vapor pressure SP (W) is also indicated.

Tungsten bromides and oxide bromides are more stable than their iodide counterparts. Hence if the same amounts of iodine and bromine are dosed into separate identical lamps, the SP (W) will be higher in the bromine-containing lamps (Fig. 30) and thus increased attack of the colder parts will occur. This can be diminished by redesigning the lamp such that the bulb wall temperature is lower or by dosing less bromine. A further improvement can be obtained when hydrogen bromide is used instead of bromine (Fig. 30). Hydrogen bromide is a stable compound and only dissociates ap-

preciably near the filament. Thus, in the colder regions of the lamp the free bromine pressure is low and this causes a decrease in the value of SP (W) between 1400 and 2200 K; the attack of colder parts like filament supports is therefore much less. Dosing hydrogen bromide is again a tedious manufacturing task but can be circumvented by using bromohydrocarbon compounds such as mono-, di-, and tribromomethane. Since these compounds are sufficiently volatile and not aggressive at room temperature, they can be easily dosed into the bulb together with the inert gas. When the lamp is burned for the first time these compounds dissociate and carbon dissolves into the tungsten filament.

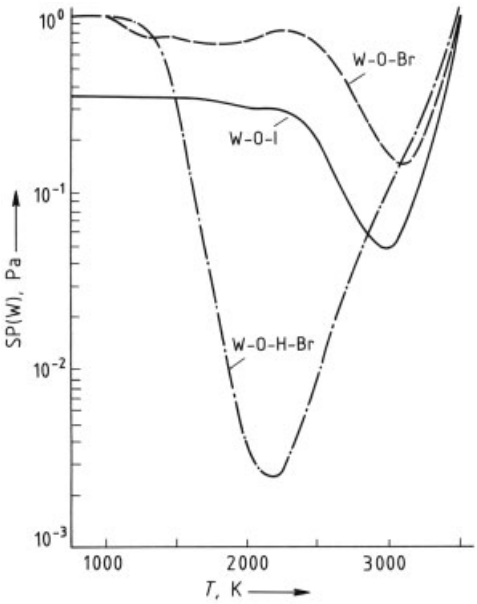

Figure 30. Summed tungsten pressure SP (W) as a function of the temperature of tungsten for iodine, bromine, and hydrogen bromide cycle lamps
All systems contain 1 Pa oxygen and 1 000 Pa (atomic) halogen.

Chlorine can be used in lamps whose bulb wall temperature is so low that the bromine and iodine cycle do not function (low wattage lamps). Thermodynamic calculations show that in fluorine-containing lamps tungsten can be transported exclusively to the hottest filament sites. However, fluorine compounds are aggressive toward the glass bulb.

4.2.4. Metal Halide Lamps

Quartz Metal Halide Lamps The color rendition and efficacy of high-pressure mercury discharge lamps can be greatly improved by adding small amounts of metal halides to the discharge vessel. In high-pressure mercury discharge lamps the rate at which electrode material (tungsten) is transported to the wall is very low due to the use of emitter material which keeps the maximum temperature of the electrode relatively low (< 2000 K). In metal halide lamps, emitters based on calcium and barium oxide cannot be used (cf. Section 6.3). Consequently the temperature of the electrode tip is usually higher than 2 500 K. This means that one of the problems during lamp life may be wall blackening.

In order to obtain good color properties, high vapor pressures of the metal halides are required: iodides are preferentially used in combination with a high wall temperature. The thermal loading of the bulb wall is limited by softening of the fused silica bulb or by reactions of the metal halides with fused silica. Depending on the desired life of the lamp, the wall temperature is between 1000 and 1300 K. As in halogen incandescent lamps, the bulb wall of metal iodide discharge lamps can be kept clear if tungsten oxide iodides are stable close to the bulb wall. Two types of additive mixtures are used:

1) *Mixtures containing iodides which react poorly with oxygen* (e.g., the iodides of sodium, thallium, indium, and tin). The oxides of these elements are not very stable, oxygen-containing iodide compounds including WO_2I_2 are present in the gas phase. By using a proper combination of additive dose and wall temperature (lamp design), wall blackening can be prevented. In analogy with halogen incandescent lamps, however, colder parts of the electrodes suffer from corrosion (see Section 4.2.3) and this may eventually lead to the electrodes breaking off.
2) *Mixtures containing iodides which react strongly with oxygen.* The advantages of the iodides of scandium and many of the rare-earth metals are their good color-rendering properties and the high luminous efficacy. The oxides of the rare-earth elements are very stable and therefore residual oxygen im-

purities react with the iodides leading to a situation in which there is not enough oxygen left to form WO_2I_2. Lamps containing these iodides consequently suffer from blackening and not from attack of the colder parts of the electrodes. A fortunate circumstance is that the tungsten deposit on the bulb wall does not form a closed, strongly light-absorbing thin layer. Tiny (sub)micron tungsten crystals are formed. As a consequence, the bulb wall is only partially covered with light-absorbing material. Therefore, by optimizing the arc tube design, the lumen loss rate due to wall-blackening can be kept so low, that effective burning times of more than 10 000 h are possible.

Another transport process which may occur in quartz metal halide lamps is transport of silica. This process occurs mainly in lamps containing iodides of category (2). The reaction of rare-earth iodides with silica usually yields a small amount of silicon iodides. These compounds act as transport agents and transport silica from sites with a high temperature to sites with a low temperature. As a result of this etching and deposition, the wall becomes frosted and scatters light.

Ceramic Metal Halide Lamps. Most mechanisms of the tungsten – metal halide chemistry, as discussed above, also apply to ceramic metal halide burners. In addition, corrosive interactions of the salts occur with the alumina wall material. The precise nature of these reactions is still under investigation, but their timescale is such that lamp lives well above 10 000 h are realized (see also Section 5.2.3).

5. Basic Lamp Materials

Materials used for lamp production comprise:

1) Metals, especially doped tungsten (Section 5.1)
2) Envelope materials (glasses and fused silica, Section 5.2, and translucent polycrystalline alumina, Section 5.3)
3) Fluorescent powders (Section 5.4)
4) Special materials such as getters, amalgams, emitters, cements, and marking paints (Chap. 6).

Material aspects of coatings used to modify spectral distribution or spatial distribution of light are discussed in Chapter 7 with special emphasis on interference and semiconductor optical filters.

5.1. Metals [8]

Metals play an important part in the construction of light sources. Table 4 gives a survey of the most important functions of metals applied in lamps.

Table 4. Functions of metals in lamps

Metal *	Lamp component
W (3680 K), W with additions (K, ThO_2, Re)	incandescent filaments
W, W with additions	electrodes
Mo (2883 K), Fe, Ni, Cr, Cu	glass – metal seals
Nb (2741 K)	ceramic – metal seals
Mo	construction elements
Na, Hg	discharge media
Zr, Al, Ba, Ni, Ta	getters
Al, Cu, Zn	lamp caps

* Melting points are given in parentheses.

Tungsten. Tungsten is used for incandescent filaments and electrodes, a successful alternative has not yet been found. The temperature range for incandescent applications varies from 2000 K in small vacuum lamps, through 2700 K in inert gas-filled lamps, up to 3300 K in halogen lamps. The temperature range for electrodes is 1200 – 2800 K. Tungsten has unique properties for these applications. Its melting point (3680 K) is the highest of all metals, its vaporization behavior is favorable, and it displays some selectivity in the emission coefficient (cf. Section 2.1). Moreover its mechanical properties at high temperatures satisfy the required conditions.

Tungsten wire is produced by powder metallurgy. Tungsten powder is compacted into rods (dimensions ca. $20 \times 20 \times 750\,mm^3$), which are sintered to densities some 10 % below the theoretical density ($19\,250\,kg/m^3$). The sintered rods are pressed to wire by rolling, swaging, and wiredrawing. Deformation of metal is always carried out above the brittle – ductile transition temperature to limit work-hardening. Tungsten, therefore, is deformed in the temperature range

1100 – 1900 K. Coiled filaments are prepared by winding tungsten wire on molybdenum mandrels. Dissolving the molybdenum mandrel selectively in nitric acid results in a filament ready to be mounted in the lamp.

Drawn tungsten wire has a fibrous microstructure (see Fig. 31 A), which is sufficiently ductile to accomodate the strains imposed in the coiling process. Upon heating to 1 500 – 2 000 K, recrystallization occurs, resulting in an equiaxed grain structure with an average grain size of the order of the wire diameter (Fig. 31 B). This structure is insufficiently stable with respect to creep (i.e., sliding along grain boundaries) at high temperature. Creep results in distortion and sag of the filament and consequently to short lamp lives.

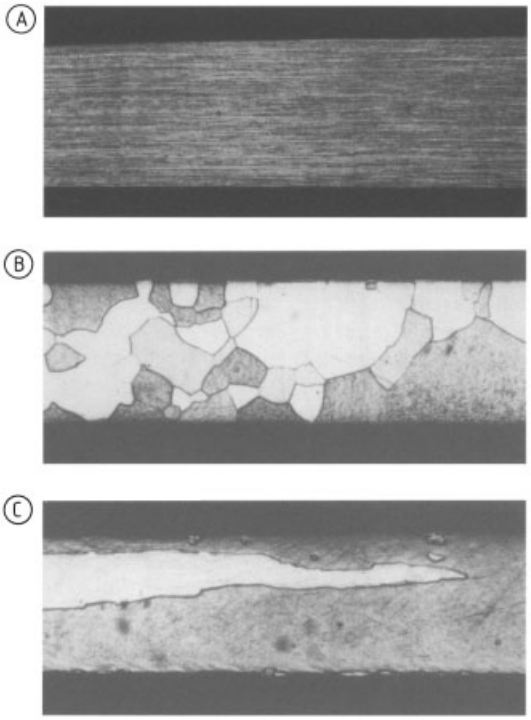

Figure 31. Tungsten wire
A) Fibrous structure of drawn tungsten wire; B) Equiaxed grain structure of recrystallized pure tungsten wire; C) Interlocking grain structure of recrystallized doped tungsten wire

Creep-resistant tungsten is obtained by doping with small amounts (ca. 100 ppm) of potassium (doped or nonsag tungsten). Potassium is added in the form of a potassium aluminosilicate during the preparation of tungsten powder. Potassium is almost insoluble in the tungsten lattice and therefore remains as potassium "bubbles" in the sintered ingot. The deformation processes, particularly wire drawing, produce rows of potassium bubbles that are aligned parallel to the wire axis. These bubbles act as barriers for grain-boundary movement, especially in the direction perpendicular to the wire axis. This forces grains to grow parallel to the wire axis during recrystallization. The result is an interlocking grain structure, with grain lengths many times the wire diameter (Fig. 31 C). This structure exhibits a sufficient resistance to sliding along grain boundaries (i.e., creep).

The recrystallization temperature of doped tungsten is higher than that of pure tungsten (2000 – 2500 K versus 1500 – 2000 K). Recrystallization is accompanied by embrittlement, which might give rise to difficulties in the handling of coils. Therefore, after the coiling procedure the filament is heat treated below the recrystallization temperature to relieve the stresses due to the coiling. Heating to the recrystallization temperature and thus the formation of the nonsag structure takes place during the processing of the lamp after the mounting of the filament. To ensure satisfactory mechanical properties for filament application close control of the microstructure, especially with regard to the dope distribution during all production stages is essential.

Other additives are used for special-purpose tungsten alloys. Thoriated tungsten (tungsten containing 0.75 – 3 wt % ThO_2) has a fine-grained recrystallization structure. The thoria particles located at grain boundaries impede boundary movements, thus increasing the recrystallization temperature and improving creep resistance. Thoriated tungsten is used for incandescent filaments in lamps with special requirements related to mechanical aspects such as vibration. The material also finds application as electrode material (e.g., in xenon arc lamps). At high temperatures, the thoria is reduced by tungsten, and the released thorium atoms act as an emitter (see Section 6.3). The use of thoria, however, is restricted because of radioactivity safety regulations.

Tungsten – rhenium alloys (3 % Re) have improved mechanical properties (toughness and ductility), furthermore they show improved cor-

rosion resistance, e.g., as high-pressure sodium lamp electrodes.

Molybdenum. The use of tungsten filaments always implies the application of molybdenum. Molybdenum is essential in coilmaking as mandrel material. Molybdenum wire is also used as filament support, while parts such as dimming caps in motorcar lamps are formed out of molybdenum sheet. Molybdenum wire and foils are used as lead-in wires with fused silica lamp envelopes. Fused silica has an extremely low thermal expansion coefficient, so that directly matched seals are not possible. Knife-edged molybdenum foil is sufficiently ductile to deform but not fracture under mechanical stress (see Section 5.2.4).

Niobium. Niobium is used in the ceramic-to-metal seals of high-pressure sodium lamps. Mostly a niobium – zirconium alloy (1 % Zr) is used because of improved ductility. The material is used in the form of deep-drawn or extruded tubes, or as wire (see also Section 5.3).

Other Metals. Composite wires with controlled thermal expansion are used as lead-in wires for glass envelopes (Dumet wires). These wires consist of a low-expansion core (e.g., iron or iron – nickel) and a high-expansion copper cladding. Such compositions are designed to match the expansion of glass (see Section 5.2.4).

5.2. Glass

The most important function of glass in lamps is to provide an envelope for the gas discharge or the glowing filament and its appropriate atmosphere. Soda lime glass was used in the first incandescent lamps and is still in use for most lamp types. The introduction of new discharge lamps, the trends towards miniaturization, and the continuous improvement of the luminous flux have increased the demands on the properties and dimensional accuracy of the glass.

Common characteristics for glasses are that they are hard and brittle, do not have melting points and can withstand relatively high temperatures without deformation. In addition to light, many glasses can transmit radiation in the near IR and UV regions. For a detailed description of the properties and production of glass, see → Glass.

No single glass can meet all the requirements of every type of lamp. The general requirements for the application of glass as lamp envelopes can be formulated as:

1) Good transmission at all wavelengths in the visible spectrum, and a controlled transmission in the ultraviolet region,
2) Adequate thermal shock resistance and adequate mechanical strength at high temperature during lamp operation,
3) Thermal stability (minimal evaporation and degassing) and impermeability to gases during lamp operation,
4) Suitability for making gastight seals to metals for the current feedthroughs,
5) Adequate electrical resistivity at high temperature during lamp operation,
6) Chemical inertness to the atmosphere in the lamp, and
7) Economic forming to the required shapes and the required tolerances.

5.2.1. Production and Composition

The first step in the manufacture of glass is the melting of the mixed raw materials in a furnace. The molten glass is delivered to glass-forming machines where lamp envelopes can be blown in the shape of bulbs or drawn in the shape of tubes. For some lamp types pressed glass parts are used. Close control of bulb and tube dimensions and wall thickness distribution are essential.

Most glasses used in lamps contain SiO_2 as the main component, the so-called glass former. These glasses may be classified into six main types according to their compositions (see Table 5).

Soda lime glass has a composition similar to the glass used for containers and window glass, but is designed to have optimum viscosity and temperature characteristics so that bulbs can be made and processed on high-speed machines. The advantages of soda lime glass are its low cost and easy processability. Disadvantages are its relatively low heat and thermal shock resistance, low chemical stability, and low electrical resistivity. For use in combination with soda

Table 5. Typical compositions and properties of main glass types used in lamps

Glass	Composition, wt %										Softening point, °C	Coefficient of thermal expansion α at 0–300 °C, 10^{-7} K^{-1}	Modulus of elasticity at 20 °C, GPa	Electrical resistivity at 350 °C, $\Omega \cdot$ m
	SiO_2	Na_2O	K_2O	B_2O_3	Al_2O_3	MgO	CaO	BaO	PbO	SrO				
Soda lime glass	72	16	1		2	4	5				700	94	72	10^3
Lead glass	63	7	7		2				21		635	93	61	10^5
Barium glass	67	8	5		4	2	2	9		3	675	93	61	10^5
Borosilicate glass	78	5		15	2						800	40	64	10^5
Aluminosilicate glass	61			1	16		10	12			1 025	45	88	10^9
Vycor	96			4							1 020	7.5	72	10^6
Quartz glass	>99.9										1 140	5.5	73	10^8

lime glass, a lead glass is available for those glass parts, that require higher electrical resistance (see page 26). Because of environmental reasons, however, a barium glass has been developed as a substitute, with almost similar properties.

Pure quartz glass (fused silica) has a superior resistance against heat and thermal shock, chemical stability, and electrical resistivity. The drawbacks of pure quartz glass are its high cost and high processing temperature.

Two types of high-silica glass have been developed which have many of the favorable properties of quartz glass but require lower processing temperatures: Vycor (containing a few percent B_2O_3) and doped quartz glass (made by the addition of a few hundred ppm of oxides like BaO, Al_2O_3, and K_2O to a pure SiO_2 melt).

In comparison with soda lime glass, borosilicate and aluminosilicate glass both have a relatively high thermal shock resistance because of their low thermal expansion coefficient, as well as a relatively high heat resistance and a high chemical stability. Aluminosilicate glass also has a very high electrical resistivity.

5.2.2. Properties

Important glass properties are listed in Table 5.

Mechanical Properties. Glasses are brittle substances with a relatively high modulus of elasticity and reasonable tensile strength, which are almost independent of composition. Characteristic for all kinds of glass is that the practical tensile strength (10 N/mm^2 for sustained stress, 20 N/mm^2 for dynamic or temporary stress) is much lower than the theoretical value (10 000 N/mm^2). This is due to surface defects and stress corrosion. Surface defects act as stress raisers and decrease the practical tensile strength with a factor up to 300. Stress corrosion causes glass to fracture under a sustained stress of 40 % of the mean instantaneous breaking stress. This factor is taken into consideration in the design of glass-to-glass and glass-to-metal seals in lamps. The resistance to thermal shock depends directly on shape, the strength of the glass, and on the reciprocal of the thermal expansion coefficient.

Thermal Properties. The maximum temperature glass can withstand without deformation and the temperature at which glass can be formed are dependent on the viscosity – temperature characteristics; the latter are often identified by specifying viscosity points. For instance, the softening point is the temperature at which the glass deforms under its own weight.

The thermal expansion coefficient is important for glass-to-glass and glass-to-metal seals. A thermal expansion mismatch leads to stresses after cooling to room temperature. With respect to thermal properties, the glasses used in lamps can be divided into three main groups:

1) Soft glasses (soda lime glasses and barium glass) with a thermal expansion coefficient of $(90-100)\times10^{-7}$ K^{-1}
2) Hard glasses (borosilicate and aluminosilicate glasses) with a thermal expansion coefficient of $(35-50)\times10^{-7}$ K^{-1}(The ‘official

boundary' between soft glass and hard glass is the softening point value of 750 °C)
3) High-silica glasses with a thermal expansion coefficient of $(5-7)\times 10^{-7}\ K^{-1}$

A favorable characteristic of glass is the temperature dependence of its thermal conductivity. At low temperatures the thermal conductivity is low, which makes local reshaping possible. The thermal conductivity increases at higher temperatures, enabling processing on high-speed machinery.

Optical Properties. Pure quartz glass has a wide transmission range from 0.18 to 4 µm. Glasses may be rendered colored, translucent, or opaque by the presence of dissolved, colloidal, or crystalline material. Oxides of iron, nickel, chromium, and manganese are some of the materials used to color glass. A controlled amount of iron and/or cerium is often required to absorb unwanted UV radiation. Sometimes special UV transmission demands are imposed on quartz glass (for instance lamps in solaria and photocopy lamps). The absorption edge can be shifted to higher wavelengths by adding titanium dioxide.

Electrical Properties. Glasses are insulators at room temperature. At higher temperatures, soda lime glass shows appreciable electrical conduction with Na^+ and K^+ as the charge carriers. Addition of PbO increases electrical resistance by hindering the mobility of Na^+ and K^+. In the environmentally friendly barium glass, the lead oxide is replaced by BaO and SrO, which fulfill the same function. Alkali-free glasses (aluminosilicate, quartz glass) have a high electrical resistance.

Chemical Properties. Resistance to the lamp atmosphere (sodium, mercury, halogen compounds) at high temperatures imposes special demands on glasses for lamps. Chemical stability of glass at elevated temperatures during lamp operation is necessary. Traces of water vapor and hydrogen must be eliminated. Manufacturing methods and subsequent thermal treatment are important factors. Quartz glass can be used up to 1400 K, although devitrification (i.e., crystallization) may occur at lower temperatures. The onset and the rate of devitrification is accelerated by the presence of impurities (especially alkali metal ions) in the glass or on the surface. The product of devitrification is cristobalite (a crystalline phase of SiO_2) that has a much higher expansion coefficient than quartz glass. This acts as a stress raiser and can cause lamp failure.

5.2.3. Glasses for Lamp Types

Incandescent Lamps. Glasses for the envelopes of incandescent lamps must be suitable for large-scale, high-speed production to keep down production costs. Impurities in the lamp atmosphere such as water vapor, oxygen, carbon dioxide, and hydrocarbons should not exceed levels of a few parts per million. However, soda lime glass is adequate when a good degassing process is provided in combination with a getter. The electrical resistivity of soda lime glass is insufficient for use in the feedthrough at the temperature of operation. The feedthrough and the components carrying the filament support wires must be made of barium glass or lead glass.

A special category comprises pressed glass lamps (special reflector lamps, sealed beam lamps) that are made by using peripheral sealing of reflector and lens pressings of borosilicate glass to enclose an accurately focussed filament assembly.

Tungsten Halogen Lamps. Soda lime glass cannot be used for tungsten halogen lamps due to their high wattage loading. Vycor, aluminosilicate, and quartz glass can be employed. Typical maximum operating temperatures of these glasses are:

aluminosilicate glass	950 K
doped quartz glass	1100 K
Vycor	1200 K
quartz glass	1400 K

With quartz glass there is a chance of crystallization from 1300 K upwards.

Fluorescent Lamps. The provision of drawn tubing (10 – 40 mm diameter) of extremely uniform cross section and wall thickness is vital for the production of fluorescent lamps. For normal fluorescent tubes, a soda lime glass

with a composition that closely resembles that of incandescent lamps is adequate. The composition (especially iron content) is controlled in such a way that radiation below 300 nm is absorbed.

Blackening of fluorescent lamps is a complicated interaction process between the mercury discharge, the fluorescent powder, and the glass; the mobility of sodium ions in the glass is an important factor. For lamps with a high wall loading (compact fluorescent lamps with narrow tubes), a special soda lime glass, lead glass, or barium glass with low alkali ion mobilities has to be used.

Germicidal lamps require high transmission at 254 nm and are produced using a special, highly transmitting sodium barium silicate glass.

Special lamps for use in solaria (low transmission below 280 nm required) and in therapeutic equipment require glass with special UV transmission properties. Sodium barium silicate glass containing titanium dioxide is used.

Low-Pressure Sodium Lamps. The discharge tube of a low-pressure sodium lamp must not be discolored by sodium at the operating temperature of 260 °C. Sodium-resistant borate glass (typical composition: 20 % B_2O_3, 50 % BaO, 5 % SiO_2, 10 % CaO, 5 % MgO, 10 % Al_2O_3) satisfies these demands. However, this glass cannot be formed in the traditional manner. A two-ply tube drawing process is used in which a closely controlled 80-µm thick layer of borate glass is continuously applied over the inside surface of a special soda lime tube.

High-Pressure Sodium Lamps. There are no glasses available that can withstand sodium at the operating temperatures of 800 – 1200 °C. For these discharge tubes, PCA is used (cf. Section 5.3).

High-Pressure Mercury Lamps. Quartz glass is used as the discharge tube material (operating temperature 800 °C). Evolution of hydrogen and/or water vapor from the envelope can be damaging especially because of the long lifetime of this lamp.

The outer bulbs of high-pressure mercury lamps for the higher wattages are made of borosilicate glass and sometimes aluminosilicate glass because of the required temperature and thermal shock resistance. For the lower wattages soda lime glass outer bulbs are used. These outer bulbs, of course, also have to block the transmission of UV radiation.

Metal Halide Lamps. The discharge tubes for metal halide lamps are made from high-quality quartz glass. The quartz glass must have a low hydroxyl content (< 5 ppm) so that evolution of hydrogen or water vapor does not occur during lamp operation. The halide cycle is highly sensitive to even the smallest amounts of water and hydrogen (cf. Section 4.2). Water vapor and hydrogen must be excluded during lamp manufacture. A typical metal halide discharge tube is shown in Figure 32. Quartz glass is permeable for Na^+ ions in the presence of an electric field. Such fields may arise in improperly designed lamps. Transport of Na^+ ions from inside the discharge tube to the outer bulb, gives rise to color drift problems due to changing salt composition. Furthermore, the formation of excess iodine in the discharge causes ignition problems [2]. Transport of Na^+ ions does not take place in 'ceramic' discharge vessels.

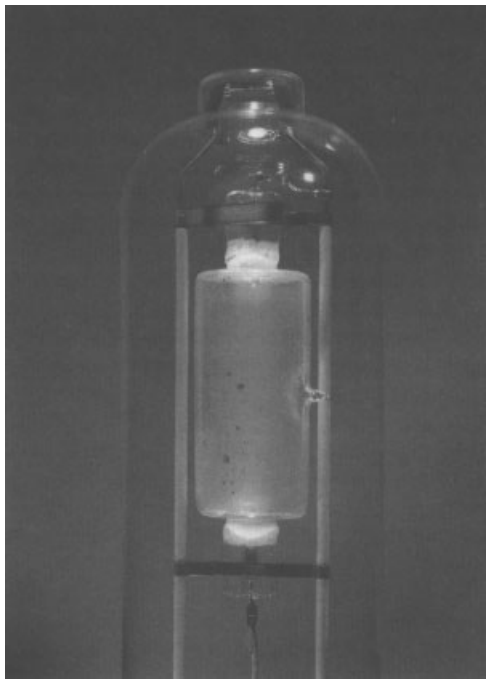

Figure 32. A 400-W metal halide discharge tube of quartz glass during the runup of the discharge.
Salt deposits are visible on the wall.

5.2.4. Glass-to-Metal Seals

Glass-to-metal seals are important for all lamp types. The electric lead-in wires are sealed in a stem glass or sealed directly in the bulb by a pinch seal. General requirements for a gastight glass-to-metal seal are:

1) The expansion of the stem glass should match the expansion of the metal from the sealing temperature to room temperature.
2) There should be a good adhesion between the glass and the metal.
3) The glass should retain its electrical resistance at high temperature.

For soft-glass seals, copper-clad wire (Dumet) and alloys of chromium, nickel, and iron are used, mostly sealed into the lead glass or barium glass stem. The wires are preoxidized to obtain a layer of Cu_2O or Cr_2O_3 that ensures good adhesion to the glass. For hard-glass seals (pure quartz glass, Vycor, and doped quartz glass), there is no metal that is both refractory enough and has a matching thermal expansion coefficient. Tungsten and molybdenum have much higher thermal expansion coefficients ($45\times10^{-7}\,K^{-1}$ and $54\times10^{-7}\,K^{-1}$, respectively) than quartz glass ($5.5\times10^{-7}\,K^{-1}$). Seals can be made by using an intermediate sealing glass with a thermal expansion coefficient between that of quartz glass and molybdenum or tungsten, but seals are usually made by pinching or shrink sealing the quartz glass tube over a thin molybdenum foil with a thickness to width ratio of 1 : 100 and a sharp foil edge. As the temperature is lowered after making the seal, the thermal expansion mismatch between the quartz glass and the molybdenum causes stresses to develop rapidly. As soon as the stresses reach the yield point of the molybdenum, they are relaxed by the plastic flow of the metal and further stress development does not occur. This is called the "Housekeeper" technique.

5.3. Polycrystalline Alumina [9]

Polycrystalline alumina (PCA), sintered to translucency, is used as the arc tube material in high-pressure sodium lamps (cf. Fig. 19 e, f), and in ceramic metal halide lamps (cf. 19 c). The first commercial high-pressure sodium lamp was produced in 1965. A complete discharge tube is shown in Figure 33. The tube material in high-pressure sodium lamps is exposed to sodium pressures from 5 to 100 kPa at temperatures ranging from 1000 to 1650 K. The development of a light-transmitting material for use under these conditions was a prerequisite for the realization of high-pressure sodium discharge lamps, since there are no glasses available that can withstand sodium at such operating temperatures. It was found that alumina powder of high (i.e., $>99.9\%$) purity can be sintered to translucency by addition of small amounts of sinter additives (e.g., ca. 0.2 wt % MgO); sintering is typically performed at temperatures of 2100 K.

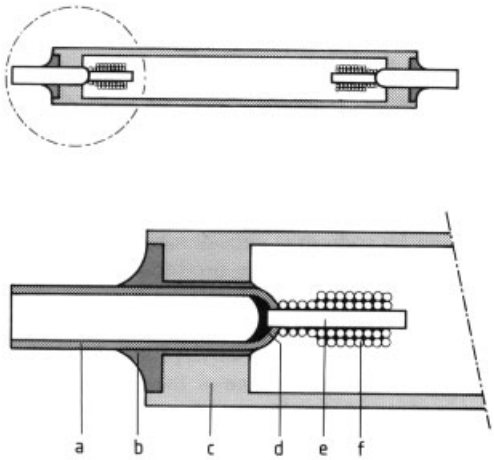

Figure 33. Typical discharge tube in a high-pressure sodium lamp
A niobium tube, carrying the tungsten electrode, is sealed into the alumina tube using a sealing ceramic
a) Niobium (containing 1 % Zr) current feedthrough; b) Sealing ceramic; c) Polycrystalline alumina tube (presintered polycrystalline alumina plugs and tube are sintered to a monolithic body); d) Brazing material (Ti) to connect niobium and tungsten (or, as an alternative, W – Nb laser welding is used); e) Tungsten rod; f) Tungsten windings with emitter

The use of PCA as tube material for ceramic metal halide lamps enables the operation of metal halide lamps at higher temperatures when compared to the quartz metal halide lamps. This implies that the vapor pressures of the radiating species can be higher than in case of quartz envelopes. Next to improving the color spread between lamps, this leads to improved color rendering and increased efficacies. The first com-

mercial ceramic metal halide lamp was introduced in 1993 (Fig. 19 c).

For PCA lamps a proper vacuum-tight connection between the current feedthrough and the arc tube material had to be developed, which also is chemically resistant to the used lamp filling (i.e., sodium and metal halides, respectively) at the elevated temperatures. The feedthrough region for the high-pressure sodium lamp is depicted in Figure 33.

Arc Tubes. Translucency, with high transmittance in the visible spectrum, and chemical inertness to the used lamp filling at high temperatures are not the only requirements for a ceramic arc tube material. Other requirements are:

1) Sufficient electrically insulating properties at high temperatures,
2) Impermeability for the gas filling enclosed by the arc tube,
3) Low sublimation rate,
4) Sufficient mechanical strength and thermal shock resistance, and
5) Economic feasibility.

Although many materials can be sintered to translucency (such as Al_2O_3, AlN, MgO, Y_2O_3, $Y_3Al_5O_{12}$, Gd_2O_3, Sc_2O_3, $LiAl_5O_8$, BeO, CaO, $3\,Al_2O_3 \cdot 2\,SiO_2$, $3\,Al_2O_3 \cdot 2\,GeO_2$, ThO_2), only a limited number meet the above requirements. At present the envelope material used by all ceramic discharge lamp manufacturers is alumina.

Two types of alumina have been employed as arc tubes: (1) transparent single-crystal alumina tubes grown from a melt, and (2) PCA tubes sintered to translucency at temperatures of about 2100 K from submicron high-purity alumina powders. Although the transparent monocrystalline arc tubes have a somewhat higher corrosion resistance to sodium and give rise to a higher luminous efficacy than PCA arc tubes, the anisotropic character of α-alumina causes considerable stresses leading to cracks and leakages upon the thermal shocks produced when the lamp is switched on and off. This, combined with the higher price of single-crystal alumina, has favored the use of PCA arc tubes.

Shaping of PCA arc tubes can be realized by conventional ceramic processing, such as:

1) Granules of alumina are pressed isostatically in a mold; the alumina granules contain sinter dope and binder and are made by spray drying to obtain proper powder flow characteristics.
2) A well-kneaded aqueous alumina paste containing dope, binder, and lubricant is extruded through an annular opening, the emerging tube being cut to the desired length.

After drying the tubes, well-fitting PCA plugs are inserted, and the assembly is presintered in air at temperatures up to 1700 K. During this process the binder is burned out. The decrease of the mechanical strength of the tubes resulting from binder removal is compensated for by bonding the primary particles to one another during presintering. In arc tubes manufactured by extrusion technology, the shape shown in Figure 33 is obtained by presintering the plugs to such a degree that they fit into the tube bore and shrink somewhat less than the tube during sintering of the assembly. Subsequent sintering to translucency is usually performed in a vacuum or in hydrogen. During (pre)sintering the dimensions of the initial pressed or extruded products decrease by some 20 %. Examples of PCA arc tubes are shown in Figure 34.

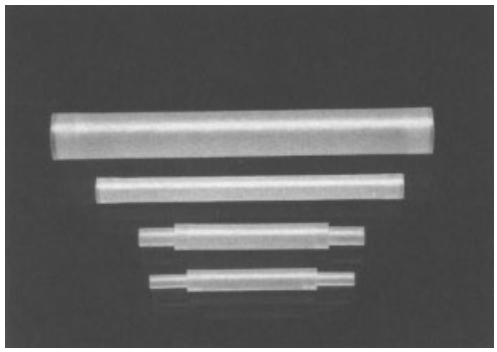

Figure 34. Examples of polycrystalline alumina arc tubes The length of the upper arc tube is 78 mm.

Microstructure. The effective removal of residual porosity is essential for obtaining translucent polycrystalline material. Typically, commercial PCA is sintered to more than 99.9 % of the theoretical density. Sinter dopes (such as $MgO, CaO, ZrO_2, HfO_2, CeO_2, Y_2O_3, La_2O_3$ or appropriate combinations) affect both the grain boundary mobility and the mobility of inclusions (pores). These parameters determine the

degree of inter- and intragranular pore entrapment which reduces light transmission due to scattering. The typical structure of a PCA surface is shown in Figure 35.

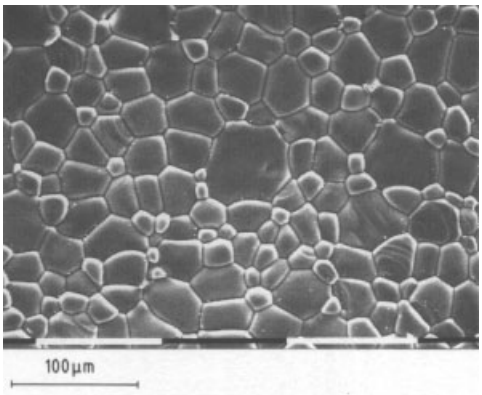

Figure 35. A scanning electron micrograph of the microstructure of a polycrystalline alumina arc tube surface

The final microstructure determines the optical, mechanical, and chemical properties of the end product. The optical quality of the material is determined by the spectral absorptivity (negligible for alumina in the visible spectrum), the spectral reflectivity (the normal single surface reflectivity is about 8 % for alumina), and the spectral transmittivity. Normal in-line transmission is very sensitive to the scattering properties of the alumina tube which depends predominantly on the residual porosity (distribution) and on the number of grain boundaries in the lightpath (in other words on the mean grain size after sintering). Obviously, optimum optical transparency is realized with single-crystal alumina (sapphire).

The mechanical properties of alumina, and the related thermal shock resistance, are determined primarily by the size distributions of the alumina grains and the residual pores. With increasing grain size the mechanical strength generally decreases, imposing an upper limit on the grain size. The compromise between the desired mechanical and optical properties is a mean alumina grain size of 25 – 40 µm.

High-purity alumina starting material (> 99.98 %) is required to minimize segregation effects (i.e., increased amounts of dopant and unintended impurities located at grain boundaries). When sinter dope levels exceed the solubility of the dope in alumina, precipitation of second phase particles is observed. Such segregation may negatively affect the optical and chemical properties of the PCA tube to a significant extent.

Corrosion. The chemical resistance of the PCA tube material to the lamp filling components is such that appropriate lamp lifetimes can be obtained. Two major types of attack on alumina are observed: (1) frontwise, or bulk alumina corrosion, which is independent on the presence of grain boundaries; and (2) corrosion which proceeds preferentially along the grain boundaries present in the PCA. The latter can result in leakage of the arc tube. Both the presence of sinter dopes and the conditions during sintering determine the chemical properties of the PCA.

In *high-pressure sodium lamps* alumina reacts with sodium to some extent. With sodium pressures up to 10 kPa, the reaction rate is so low, that it allows operational lifetimes from 10 000 to 30 000 h. With higher sodium pressures, the reduction of alumina by sodium is considerable. In addition to the solid-state reaction products ($NaAlO_2$ or $NaAl_5O_8$/$NaAl_{11}O_{17}$), aluminum has also been detected and condenses in the coldest part of the arc tube. The $NaAl_5O_8$/$NaAl_{11}O_{17}$ is formed in the high-temperature zone of the arc tube. Furthermore, emitter material sputtered from the electrode onto the alumina surface enhances the rate of sodium corrosion.

In the recently introduced *ceramic metal halide lamps*, corrosion reactions imply interaction of rare earth iodides with alumina. Many details thereof, however, are still unclear at the moment

Sealing. Niobium was chosen as a current-feedthrough material because its thermal expansion coefficient is very close to that of alumina. Furthermore, as far as high-pressure sodium lamps are concerned, niobium has sufficient chemical resistance to sodium at high temperatures, however it is also permeable to hydrogen; (hydrogen, which arises from moisture in arc tubes, has a negative influence on arc ignition, see Section 6.1). The feedthrough (Fig. 33) is sealed into the PCA tube by means of a sealing ceramic. The gastight seal must (1) show good adhesion to both PCA and niobium, (2) re-

sist the sodium vapor pressure at 900 – 1200 K, and (3) withstand thermal shocks occurring upon switching the lamp on and off. Typical sealant compositions (in weight percent) are: 5.6 MgO, 38.6 CaO, 8.7 BaO, 45.4 Al_2O_3, 1.7 B_2O_3 or 39.2 CaO, 9.3 SrO, 5.2 Y_2O_3, 46.3 Al_2O_3. The sealing ceramic is pressed in the shape of a ring which is positioned around the niobium feedthrough and heated to melting (typically between 1 500 and 1 600 K). The space between the feedthrough and the PCA tube is filled with the molten oxide mixture which must have a sufficiently low viscosity. Upon cooling, fine-grained crystalline phases are formed. The final composition of the seal depends on the initial sealant composition and the applied temperature – time schedule. An optimum seal closely matches the thermal expansion of both alumina and niobium.

With *ceramic metal halide lamps*, sealing is complicated by the fact that niobium reacts with the arc tube filling. Therefore, a more complex feedthrough construction is used, where direct contact between niobium and the arc tube filling is prevented. To realize this feedthrough construction a special end shape of the arc tube is required, in combination with both a specific sealing ceramic and a dedicated sealing process.

5.4. Fluorescent Powders [10]

Fluorescent powders (also called phosphors) for lighting applications are mainly used in combination with low-pressure mercury discharges. They absorb the UV radiation emitted at 185 and 254 nm and efficiently convert it into radiation of a longer wavelength, mostly visible radiation. In low-pressure mercury discharge lamps almost the whole of the emission spectrum is determined by the phosphor. For a detailed description of phosphors see → Luminescent Materials, Chap. 1.

Properties. A fluorescent powder consists of a host lattice doped with a few percent of an activator, which is responsible for the luminescent properties. The host lattices are always inorganic, oxygen-containing materials such as oxides, aluminates, phosphates, borates, or silicates. The activators are metal ions (e.g., Pb^{2+}, Mn^{2+}), in many cases rare-earth metal ions. A well-known phosphor is the halophosphate $Ca_5(PO_4)_3X : Sb^{3+}, Mn^{2+}$ where X = Cl, F. The UV photons are absorbed by the Sb^{3+} ions and the excitation energy is partly emitted in the blue region of the spectrum. The other part of the excitation energy is transferred to the Mn^{2+} ion, which has an orange luminescence; here Sb^{3+} acts as a sensitizer. In most cases the host lattice does not participate in the luminescent process; it does, however, influence the position of the energy levels of the activator ions and consequently the absorption and emission wavelengths.

In the lamp the phosphor layer is in direct contact with the discharge plasma. Therefore, the phosphor should be stable to short-wavelength UV irradiation as well as to ion and electron bombardment over prolonged exposure up to 10 000 h.

Table 6. Important fluorescent powders (phosphors) used in lamps

Material	Emission wavelength
General lighting phosphors	
Broad band emission	broad emission bands resulting in approximately white light
$Ca_5(PO_4)_3(Cl,F) : Sb^{3+}, Mn^{2+}$	
Narrow band emission	
$BaMgAl_{10}O_{17} : Eu^{2+}$	450 nm
$CeMgAl_{11}O_{19} : Tb^{3+}$	545 nm
$Y_2O_3 : Eu^{3+}$	612 nm
UV-emitting phosphors	
Narrow band emission	330 – 370 nm
$BaSi_2O_5 : Pb^{2+}$	
$YPO_4 : Ce^{3+}$	
$LaPO_4 : Ce^{3+}$	
$CeMgAl_{11}O_{19}$	

Types of Phosphors. The most important fluorescent powders are given in Table 6. A distinction can be made between phosphors for general lighting purposes and phosphors for industrial and medical purposes (UV-emitting phosphors). Until 1975 the general lighting market was dominated by the halophosphate lamps containing $Ca_5(PO_4)_3$ (Cl, F): Sb^{3+}, Mn^{2+}. The Sb^{3+} ions yield a blue luminescence, the Mn^{2+} ions an orange one. By adjusting the ratio of the blue to orange luminescence, white light can be obtained; the luminous efficacy of such lamps is 80 lm/W, but the color rendition is moderate and the CRI is only 50 – 60. A high luminous flux and a good color rendition can be obtained by blending three phosphors emitting in narrow wavelength intervals centered at 450, 545, and

612 nm, respectively. Narrow band emission is only found with rare-earth ions, notably Eu^{2+} emitting in the blue, Tb^{3+} in the green, and Eu^{3+} in the red.

The development of rare-earth based phosphors therefore was the crucial step in the development of low-pressure mercury discharge lamps with a high efficiency of 100 lm/W and a CRI of about 85. Recent trends are the application of $Sr_3(PO_4)_5Cl:Eu^{2+}$ instead of $BaMgAl_{10}O_{17}:Eu^{2+}$, while the green aluminate $CeMgAl_{11}O_{19}:Tb^{3+}$ is gradually being replaced by more efficient phosphors, such as $LaPO_4$: Ce^{3+}, Tb^{3+} and $GdMgB_5O_{10}:Ce^{3+}$, Tb^{3+}.

The Super de Luxe lamps, with near perfect color rendition, occupy a smaller but interesting market niche. The spectrum generated in these lamps is more or less continuous, extending from the blue to the deep red. Superior color rendition is obtained at the expense of the luminous flux. The phosphors applied here are based on combinations of those mentioned in Table 6. Lamps with a blend of $GdMgB_5O_{10}:Ce^{3+}$, Tb^{3+}, Mn^{2+}, $Sr_4Al_{14}O_{25}:Eu^{2+}$ and $Ca_5(PO_4)_3(Cl,F):Sb^{3+}$, Mn^{2+} reach CRI values of 95 along with a luminous efficacy of 65 lm/W.

The introduction of rare-earth based phosphors also resulted in a drastic improvement in lamp maintenance. Due to the chemical and thermal stability of these phosphors the wall load could be increased, resulting in a reduction of the tube diameter from 38 to 26 mm, and subsequently to 10 mm.

The luminous efficacy and the color rendition of a high-pressure mercury lamp is enhanced by applying an YVO_4: Eu^{3+} phosphor on the inner side of the outer bulb wall. This phosphor is highly efficient for the transformation of the Hg 254 and 365 nm line radiation into red light at 617 nm up to temperatures of about 300 °C. The UV light is absorbed by the vanadate groups, from which the energy is transferred to the Eu^{3+} activator.

Production and Application. Lamp phosphors are produced by solid-state reaction at elevated temperature in a controlled atmosphere. In view of the poor reactivity of the raw materials (i.e., oxides or carbonates), small quantities of reactive compounds such as halides are often added. The phosphors are prepared as fine-grained particles with a fairly narrow size distribution between 1 and 10 µm. Application on the lamp bulb wall takes place via a flow coating technique. At present coating suspensions are based either on water or on organic solvents (e.g., butyl acetate). The suspension is stabilized by adding chain-type polymers of high molecular mass such as cellulose derivatives. The phosphor powder should be applied as a thin layer of uniform thickness. This is achieved by careful design of the rheological properties of the suspension, combined with appropriate drying conditions. The polymer is removed in a subsequent baking step at about 600 °C in air. The phosphor must be able to withstand this "lehring" process.

Often a protective coating is applied to the glass bulb prior to the application of the phosphor layer. This protective layer consists of a highly refractive material such as alumina. Its primary function is the protection of the glass against discoloration by the discharge, resulting in an improved lumen maintenance of the lamp during life. By tuning the size of the alumina particles to the emission wavelength of the mercury discharge, the UV-reflectivity of the precoat layer can be enhanced. With the aid of such a reflective precoat the phosphor coating weight can be reduced considerably without affecting the output of the lamp.

6. Special Lamp Materials

6.1. Getters

As discussed in Section 4.1, long lifetimes and maintenance of the required lamp properties can only be obtained by controlling the chemistry of gaseous impurities. Impurities mainly consist of water, hydrogen, oxygen, nitrogen, carbon dioxide, methane, and other hydrocarbons. The source of these species may be residual gas remaining after pumping or contaminated lamp components. Lamps always contain impurities which are adsorbed at or dissolved in their solid components (e.g., water in glass). During burning, many of these impurities are gradually released into the gaseous phase. Although the total quantity of gaseous impurities may be very small ($< 10^{-6}$ mol) the effects on lamps can be severe. Notorious examples are:

1) Water vapor in incandescent lamps (cf. Section 4.2)
2) Hydrogen in gas discharge lamps, which increases ignition and reignition voltages
3) Oxygen-containing gases, which may corrode electrodes and filaments
4) Residual gases in vacuum incandescent lamps or outer bulbs, which cause arcing and may reduce thermal insulation

Lamp production processes are designed to reduce the level of impurities to a minimum within the limits of mass production. Production rates vary from 1000 to 6000 lamps per hour and the time available for pumping and degassing is about 1 min per lamp. However, slow degassing reactions (e.g., desorption of water from quartz) take hours. In most lamps this problem is overcome by using a getter. A getter can be regarded as a chemical pump that is built into the lamp and can absorb harmful impurities. The getters ensure that the proper gas composition is obtained and maintained during lamp life. Three types of getters are used:

1) Evaporation getters in which the getter is activated by fast evaporation and/or thermal decomposition (P_4, P_3N_5, TiP, Ba, $Ba + Pt/Al_2O_3$)
2) Zirconium-based getters (Zr + Al, Zr – Al alloys, Zr – Ni alloys)
3) Tantalum-based getters (Ta, Ta/W)

Evaporation Getters. Red phosphorus, P_4, is the most widespread phosphorus getter for incandescent lamps. The self-combustion in air and the high vapor pressure of red phosphorus limit the maximum process temperature during lamp production to 650 K. Higher temperatures are possible with more stable phosphorus compounds, such as phosphorus pentanitride, P_3N_5.

Phosphorus getters are dosed in the lamp by dipping the filament in a suspension of fine getter powder. After pumping and sealing off, the lamps are flashed for about 1 s several times. During the first flash the phosphorus getter evaporates and precipitates at the lamp bulb. During subsequent flashes the phosphorus layer reacts with the harmful gases.

Figure 36 displays the cleanup of gases in a clear vacuum incandescent lamp by red phosphorus. After pumping the residual pressure of 1 Pa is predominantly due to nitrogen. Without phosphorus, the pressure increases upon flashing to approximately 5 Pa due to the formation of carbon monoxide from carbon impurities in the tungsten filament. In lamps with a phosphorus getter, the pressure decreases to < 0.1 Pa. This low pressure prevents arcing.

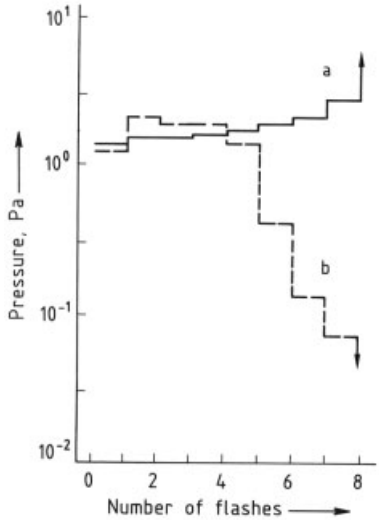

Figure 36. Pressure evolution in clear vacuum incandescent lamps during flashing without (a) and with (b) red phosphorus

The *titanium phosphide getter* works according to the same principle as phosphorus. TiP is used if higher thermal stability of the getter is required.

The *phosphorus pentanitride getter* decomposes during flashing into PN, N_2, and yellow P_2. Due to the formation of nitrogen, this getter can only be used in gas-filled incandescent lamps. The phosphorus is used to bind oxygen and water with formation of condensed phosphoric acid at bulb temperatures below 500 K. With higher bulb temperatures, a second getter (i.e., the zirconium – aluminum getter) has to be used (see below).

Barium getters are used in lamps to maintain the vacuum in the outer bulbs of low-pressure and high-pressure sodium lamps. Pressures of less than 0.1 Pa are required to prevent arcing and to ensure thermal insulation. Because of its high chemical activity, pure metallic barium cannot be used. Instead, a powder mixture of the intermetallic compound $BaAl_4$ and nickel is enclosed in a metallic ring and inserted in the outer bulb. After pumping and sealing off, the getters

are heated inductively to 1200 – 1300 K. This starts an exothermic reaction leading to the formation of a nickel – aluminum alloy in the ring and the evaporation of barium. Barium is deposited on the cold glass wall as a porous metallic film. This film reacts readily with most gases (e.g., H_2, N_2, O_2, H_2O, CO, and CO_2) apart from noble gases and methane.

After pumping off the outer bulbs of high-pressure sodium lamps, the pressure at 300 K is about 1 Pa and the residual gas composition is 70 % hydrogen and 30 % air. After deposition of the metallic barium film the pressure decreases to 0.1 Pa and more than 90 % of the gas consists of methane. Methane is formed as a byproduct when the barium film reacts with carbon-containing species and water or hydrogen. Figure 37 shows the evolution of the methane pressure in the outer bulbs of high-pressure sodium lamps. Starting at 0.1 Pa, the pressure rises to a maximum slightly below 1 Pa after 10 min. After 2 h in operation, the pressure has decreased to less than 0.1 Pa as a result of the thermal decomposition of methane at the hot discharge tube (temperatures > 1000 K).

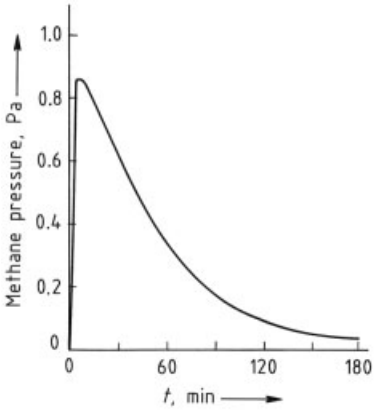

Figure 37. Evolution of the methane pressure in the outer bulbs of high-pressure sodium lamps

Since the niobium feedthroughs of the discharge tube are permeable to hydrogen, the barium film in the outer bulb also reacts with residual hydrogen from the discharge tube. The outer bulb pressure in lamps that have burned for a few hundred hours is approximately 3×10^{-2} Pa. The gas composition is 75 % helium, 15 % argon, and 10 % methane. Helium originates from air via permeation through the outer bulb wall. Barium is being increasingly replaced by zirconium-based getters.

The residual pressures in the outer bulbs of low-pressure sodium lamps after sealing and after evaporation of barium are the same as in high-pressure sodium lamps. Since the temperature of the discharge tube is about 550 K, methane is not decomposed thermally. The presence of methane diminishes the lamp efficacy due to thermal leakage. A methane decomposition catalyst comprising highly dispersed platinum on a porous alumina pellet was introduced in the 1980s. A single catalyst pellet mounted in the lamp decomposes the methane effectively at a temperature of 500 K. The carbon precipitates in the porous alumina pellet and hydrogen is captured by the barium getter. Within 15 h the methane pressure declines to below 10^{-2} Pa as shown in Figure 38. This application ensures excellent heat insulation during the complete life of 15 000 h of low-pressure sodium lamps.

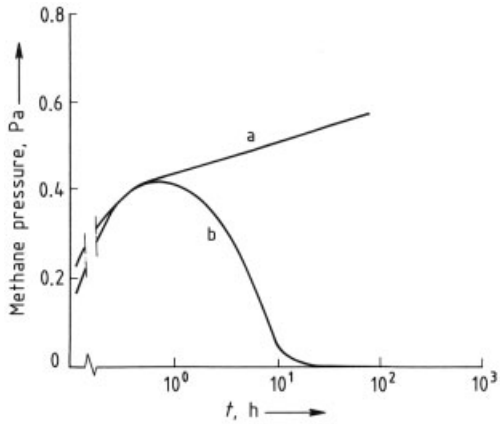

Figure 38. Evolution of the methane pressure in the outer bulbs of low-pressure sodium lamps without (a) and with (b) a decomposition catalyst (0.5 % Pt, Al_2O_3)

Zirconium-Based Getters. Zirconium has a very high chemical affinity for oxygen and nitrogen and also forms stable hydrides. At 600 – 650 K the changes in free energy for binding oxygen, nitrogen, and hydrogen are approximately 1000, 600, and 100 kJ/mol, respectively. However, surface layers of oxides and nitrides impede the absorption of hydrogen. At temperatures above 600 K, a metallic zirconium surface is formed because the surface oxide and nitride layers dissolve in the bulk zirconium. Conse-

quently, zirconium does not take up hydrogen below approximately 600 K.

In *gas-filled incandescent lamps,* a getter suspension comprising an 80 : 20 wt % mixture of zirconium and aluminum powder is applied at the stem or the lead wires. The temperature of the getter in an operating lamp is about 700 K and hydrogen is therefore absorbed. Hydrogen comes from water, which is always present in the bulb wall or its coating. As the wall temperature rises the water is released and consequently the water vapor cycle starts (cf. Section 4.2). Without the zirconium – aluminum getter, lamps with a designed life of 1000 h would only burn for a few hundred hours.

The gas-filled outer bulbs of *metal halide lamps* are provided with a hydrogen getter. Water vapor desorbed from the glass wall is converted into hydrogen and metal oxides at hot metal parts inside the outer bulb. Without a getter the hydrogen pressure may reach 25 Pa depending on the quality of the degassing process and the bulb temperature. After a few hundred hours the hydrogen pressure inside the discharge tube also increases because of a slow permeation of hydrogen through quartz. This results in increasing ignition and reignition voltages and ultimately in extinction of the lamp. To overcome these problems, the outer bulb getter has to maintain a hydrogen pressure of less than 5 Pa.

Since the gas filling consists of a mixture of noble gas and nitrogen, the getter must be selective for hydrogen. An alloy of zirconium and aluminum (84 : 16 wt %) or a zirconium – nickel compound (73 : 27 wt %) is commonly used. These getters are applied as porous pellets, rings, or coated strips. The zirconium – aluminum alloy operates in the temperature range 650 – 750 K where it is activated by prolonged exposure to hydrogen. Below 650 K the surface layer is impermeable to hydrogen and above 750 K nitrogen is gettered too. The zirconium – nickel compound absorbs hydrogen readily between 550 and 750 K. Nickel catalyzes the dissociation of hydrogen and thus facilitates hydrogen uptake. Figure 39 shows the influence of the getter on the evolution of the partial hydrogen pressure in the outer bulb of metal halide lamps. Both getters absorb the evolved hydrogen; the zirconium – nickel getter can reduce the hydrogen pressure to below 5 Pa within 10 h.

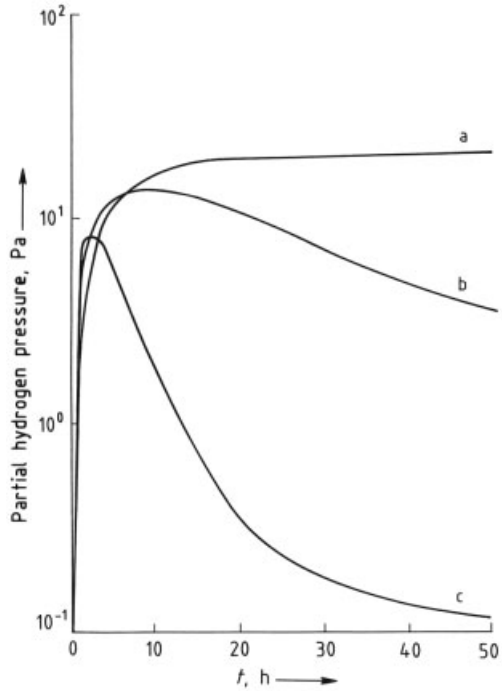

Figure 39. Partial hydrogen pressure inside the outer bulb of a high-pressure metal halide lamp as a function of the operating time of the lamp
a) Without a getter; b) With a Zr – Al getter at 350 °C; c) With a Zr – Ni getter at 250 °C

The zirconium – aluminum alloy is also used as a vacuum getter in evacuated outer bulbs of *high-pressure sodium lamps*. It either has to be activated by heat treatment (30 s at 1 200 K) or located at a position in the lamp where the temperature lies between 750 and 850 K. In contrast to the evaporated barium film, the activated zirconium – aluminum alloy does not block light transmission. This is particularly important for small lamps. The barium film, however, eliminates impurities much faster than the activated zirconium – aluminum getter.

Tantalum-Based Getters. Tantalum foil or tantalum – tungsten wire is sometimes used to getter oxygen in halogen incandescent lamps.

6.2. Amalgams [11]

Amalgams are employed in high-pressure sodium lamps and in fluorescent lamps. In high-pressure sodium lamps sodium amalgams are

used; the vaporized sodium determines the basic features of the discharge.

An interesting amalgam with "retrograde vaporization" properties is used in fluorescent lamps in which an optimum mercury vapor pressure (typically 0.5 – 1 Pa) must be maintained. Values outside this range lead to a significant decrease in light output. Usually, an excess of mercury is introduced into the lamp and thus the vapor pressure is determined by the coolest part of the tube. Fluorescent lamps are designed to operate at a temperature of ca. 40 °C. However, in certain situations the coolest part of the tube is far in excess of this. This is especially true for compact fluorescent lamps with integral control devices (see Fig. 18). By using an amalgam rather than mercury alone, the required mercury vapor pressure can be obtained at a higher wall temperature.

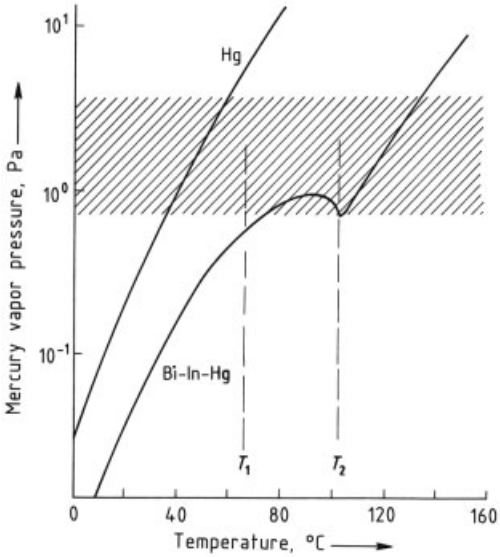

Figure 40. The mercury vapor pressure above pure mercury and above an Bi – In – Hg amalgam
The shaded area denotes the mercury vapor pressure range for optimal luminous flux ($\geq 90\,\%$ maximum value).

Figure 40 shows the vapor pressure of pure mercury as a function of temperature and also the mercury vapor pressure above a typical amalgam, namely Bi – In – Hg. Up to T_1 the amalgam is solid, between T_1 and T_2 the solid and liquid phases coexist, and above T_2 there is only the liquid phase. Between T_1 and T_2 the mercury vapor pressure is stabilized to a significant degree. This is because the mercury content of the liquid that coexists with the solid decreases with increasing temperature. This behavior counteracts the normal tendency of increasing mercury vapor pressure with increasing temperature.

Other amalgams, such as Pb – Bi – Sn – Hg and Pb – In – Sn – Hg are also in use, depending on the thermal balance of the lamp system.

6.3. Emitters

General. Most discharge lamps are equipped with electrodes coated with an emitter material, whose function is to improve the electron-emitting properties of the electrodes.

In gas-discharge lamps, current is mainly carried by thermionically emitted electrons. The number of electrons that can be emitted per unit time depends exponentially on the thermionic work function divided by the absolute temperature (Richardson's equation). Tungsten, which is commonly used as electrode material, has a relatively high work function. However, by using an emitter the work function can be reduced considerably, resulting in a much lower electrode temperature, which is beneficial for both lamp life and efficacy.

A common characteristic of emitter materials is that they contain one or more heavy elements of groups 1 and 2 (such as cesium and barium) or thorium. The emitters applied in gas discharge lamps usually contain barium, either in the form of barium oxide or barium-containing salts. The coverage of tungsten with a barium layer lowers the work function from approximately 4.5 eV to ca. 1.8 eV.

Emitters in Low-Pressure Discharge Lamps. The emitters used in fluorescent lamps and low-pressure sodium lamps consist of a mixture of calcium, strontium, and barium oxides. Since the oxides are not stable in air (they react with water and carbon dioxide), they are applied to the electrodes as their carbonates from a suspension. In order to obtain sufficient emitter material on the electrode, the electrode is built up from a main wire coiled with a thinner wire (Fig. 41). The space between the wires is filled with the suspension by dipping. If an extremely large amount of emitter is required, as in low-pressure sodium lamps, the emitter is applied by cataphoresis.

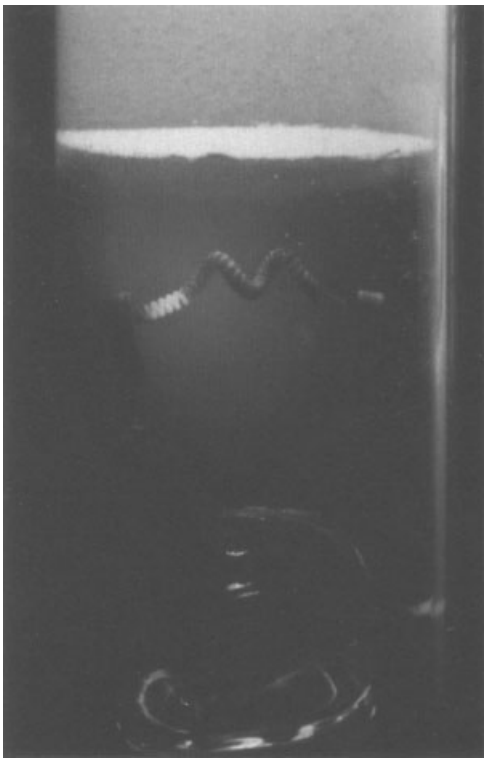

Figure 41. Electrode of an operating fluorescent lamp

Conversion of the carbonates to oxides takes place during the processing of the lamp by passing a current through the electrode to raise its temperature to approximately 1200 °C. During this activation process, a second important step takes place, which implies the formation of barium, by reactions of the type:

$$6\,BaO + W \longrightarrow Ba_3WO_6 + 3\,Ba$$

The role of barium is essential, it is the real emitter in the electron-emitting systems tungsten – barium and BaO – Ba. Furthermore, barium penetrates BaO making it an *n*-type semiconductor. During lamp operation the emitter mix is depleted from the coil, which finally gives rise to the end of life of the lamp.

Emitters in High-Pressure Discharge Lamps. The electrode temperatures in high-pressure discharge lamps are much higher than those in low-pressure discharge lamps; typical values are 1800 – 2000 K and 1500 K, respectively. A mixture of calcium, strontium, and barium oxides cannot be employed due to the high evaporation rate. Instead, barium salts (e.g., tungstates) are used, which have a sufficient Ba/BaO partial pressure for proper functioning of the electrode at these temperatures.

The electrodes in high-pressure mercury and high-pressure sodium discharge lamps consist generally of tungsten rod, coiled with a thinner tungsten wire. The space between the rod and the wire is filled with the emitter material (Fig. 42). Processing of the electrode – emitter combination takes place prior to lamp processing. To get the emitter material (e.g., compounds such as $Ba_3Y_2WO_9$ or Ba_2CaWO_6) into the electrode spaces, a suspension of the tungstates is used. After application of the suspension, the electrodes are heated to remove solvent and binder, and to ensure good fixation of the emitter material to the tungsten. Tantalates, aluminates, and other tungstates or mixtures can also be used as emitter material.

Figure 42. Electrode of an operating high-pressure mercury lamp

The performance of the emitter material depends more on processing than on the choice of the material. Suspension composition and heat

treatment strongly influence the performance of the emitter.

In metal halide lamps none of the abovementioned materials can be used because the halides react with barium oxide to form the volatile oxide halides. In these lamps thoriated tungsten is sometimes used and thorium serves as the emitter. For environmental reasons, there is a trend to eliminate the radioactive thorium in lamps.

6.4. Cements and Marking Paints

Capping Cements. The function of a cement is to attach the metal cap to the glass bulb and guarantee sufficient strength of the joint until the end of lamp life under a variety of atmospheric conditions. A suitable cement must have perfect adherence to cap metals and glasses, some permanent plasticity, and sufficient resistance to moisture and elevated temperatures.

A cement composition consists of adhesives, fillers, and additives. A typical thermosetting adhesive mixture consists of shellac and phenol – formaldehyde resin. Limestone is often used as a filler. Typical additives are colophony (rosin) and silicone resin. Filler and additives are added to obtain the required rheological properties. Silicone resin also improves resistance to atmospheric influences. The cement paste is obtained by mixing the resin – cement powder with low-boiling organic solvents (alcohols).

Marking Paints. Lamp brands, type designations, etc., are applied by means of an indelible marking paint. A typical marking paint is a lead borate glass combined with silver oxide. This mixture can be applied during lamp production and then heated under reducing conditions. Other marking paints are in use (pigmented stoving enamels).

7. Coatings

The radiant flux from the envelope of a light source often requires some modification with respect to spectral distribution or spatial distribution of the light. Such modifications are effected by inner or outer coatings on the lamp envelope.

Modification of the spectral distribution may involve improvement of the color balance of the lamp, an increase of the luminous efficacy, or the filtering of unwanted radiation. The two latter aspects are especially important for lamps in which a substantial part of the radiation is in the IR region rather than in the visible region. Thin film optical filters are used for this [12]; the most common types are interference filters and semiconductor filters.

Diffusive and reflective coatings modify the spatial light distribution. In addition, they may change the color appearance. Furthermore, important examples of coatings are the fluorescent powders, discussed in Section 5.4.

Interference Filters. The properties of an interference filter are determined by the thickness and the refractive index of different films which are deposited on top of one another. By varying the number and thicknesses of the layers, an infinite number of reflection spectra can be made. Generally, only two materials are used, one with a high refractive index (e.g., titanium dioxide, tantalum pentoxide, and zinc sulfide) and the other with a low refractive index (e.g., silicon dioxide and magnesium fluoride), which are stacked alternately. The filters can be deposited on a substrate by evaporation, sputtering, dipping, chemical vapor deposition, or plasma chemical vapor deposition (→ Thin Films). The performance of the filters depends strongly on the interplay between the materials used, the deposition technique, the geometry of the substrate, and the working environment (temperature and humidity).

An example of commercially available lamps with a dielectric multilayer film is an incandescent lamp provided with a ”cold” or cold-light” mirror. The inner side of a glass ellipsoid or paraboloid is coated with the cold-light filter and the light source is placed inside the reflector. The filter reflects visible radiation to illuminate an object that has to be protected from heat, and transmits the IR radiation to another site. A typical reflection spectrum of such a dichroic filter is shown in Figure 43. The amount of heat reduction depends strongly on the geometry of the system.

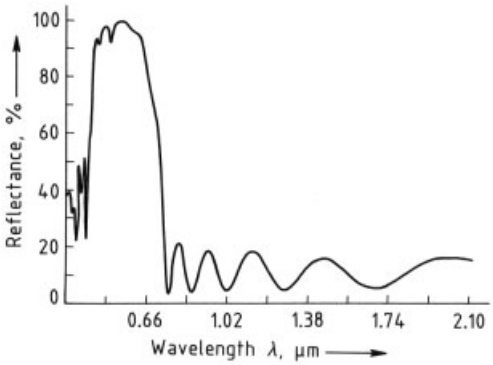

Figure 43. Reflection spectrum of a dielectric multilayer film ("cold light" filter)

Another example is an incandescent lamp with a "hot" mirror on the outer lamp surface. These filters are the reverse of the "cold" mirror. Visible radiation is transmitted, while the IR radiation is reflected. This reflected IR radiation is absorbed by the filament, thus diminishing the required power input. Commercially available lamps have an energy saving around 30 %. In addition, interference films can be used to make precise adjustments to the color appearance of the lamp.

Semiconductor Filters. In a semiconductor film, the refractive index can be influenced by adding free charge carriers to the material (doping). The doped semiconductor film behaves as a metal (i.e., a reflector) at long wavelengths and as a dielectric (i.e., a transparent material) at short wavelengths. By using suitable semiconductor and dope material with a properly chosen concentration of free charge carriers and film thickness, a semiconductor filter can be made which has a high transmittance in the visible part of the spectrum and a high reflectance in the infrared. The spectral characteristics are strongly dependent on the choice of the materials. Suitable combinations for IR reflection are indium oxide doped with tin (In_2O_3: Sn) and tin oxide doped with fluorine (SnO_2: F).

A semiconductor film is used as thermal insulation for the low-pressure sodium lamp which has an optimal luminous efficacy at a sodium vapor pressure of 0.4 Pa, corresponding to a temperature of 533 K. At this temperature the discharge tube radiates a large amount of energy. To save energy, the discharge tube is surrounded by an evacuated outer bulb that is provided with a doped indium oxide filter. The reflectance and transmittance spectra are shown in Figure 44. Compared to a lamp burning in free air, the luminous efficacy is approximately tripled.

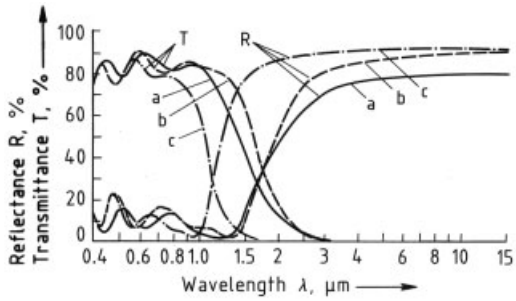

Figure 44. Reflectance (R) and transmittance (T) spectra for a semiconductor film
Doping level, N/cm^3 : a) SnO_2: F, 6×10^{20} ; b) In_2O_3: Sn, 5×10^{20} ; c) In_2O_3: Sn, 13×10^{20}

Diffusive and Color Coatings. Diffusive coatings of TiO_2, SiO_2, or Al_2O_3 powders, sometimes in combination with a colorant, are applied internally in incandescent light bulbs or in outer bulbs of gas discharge lamps. These materials are often mixed with pigments to change the color appearance. The layers are applied by suspension, vapor, or electrostatic coating. These techniques are also used with (transparent) inner color coatings. Typical materials here are prussian blue, cobalt oxide (blue), and nickel titanate (yellow). External coatings mainly consist of stoving enamels that harden at elevated temperature. A wide assortment of pigments are available which are basically not different from normal industrial pigments.

Reflective Coatings. Reflective coatings consist of aluminum, silver, or aluminum – copper (gold color). These coatings are typically applied by evaporation techniques. For more durable coatings sputtering is used.

8. Lamp Production

Nowadays, more than 15×10^9 lamps are manufactured annually. High-speed machinery is often used, as in the production of fluores-

cent lamps (worldwide production: $> 2.5 \times 10^9$ lamps per year).

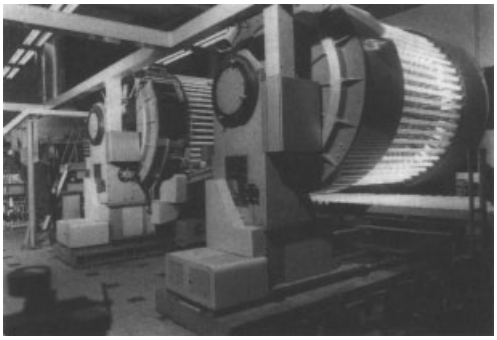

Figure 45. Large-scale production of fluorescent lamps

Usually the tungsten wire and coils for the electrodes and the fluorescent powders are made separately but the rest of the operation is often integrated in one plant. Glass tubing is continuously drawn from a furnace and cut to the required length. The tubes are coated on the inside with a fluorescent powder suspension and passed through an oven at a temperature close to the softening point of glass ($\approx 600\,^{\circ}$C); this burns away the organic compounds in the fluorescent powder suspension, improves powder adhesion, and eliminates all other impurities. The hot tubes, typically at 500 °C, are then automatically transferred to sealing – exhaust machines, which can produce more than 6000 lamps per hour (Fig. 45). Electrode assemblies which have been fabricated on side lines are sealed into the ends of the tube. Since the emitter on the electrode is sensitive to humidity, it is initially applied to the electrode coil in the form of a carbonate. Once the electrode assembly is sealed into the tube, the emitter is converted to oxides by controlled heating of the electrode coil. The tubes are now exhausted to remove all gas impurities. This is often carried out via a flushing technique in order to minimize the process time. After being filled with rare gas at the correct pressure (typically 400 Pa), the tubes are sealed. Often mercury is present in a sealed holder inside the lamp and this is only opened after the lamp has been sealed. This ensures accurate dosing of mercury in the lamp and minimal mercury contamination during the process. In the final stages the lamp caps are cemented onto the ends, product information is printed on the glass, and the correct functioning of the lamp is checked before it is automatically packed.

9. Economic Aspects and Uses [13]

An estimate of the world market for lamps is given in Table 7. The highest market numbers are for incandescent lamps. However, most of the artificial light in the world is generated by fluorescent lamps, because of their high luminous efficacy and their long life (cf. Table 3). The numbers in Table 7 illustrate the general growth of the lighting market. In particular there is a strong increase in the numbers for energy-efficient lamps such as compact fluorescent and high-pressure discharge lamps (HID).

Table 7. World market for lamps (in millions of lamps)

Type of lamp	1972[a]	1984[a]	1998[a]
Incandescent	4275	5285	11 700
Fluorescent	815	1200	2650
Compact fluorescent		25	420
High intensity discharges (HID)[c]	33	49	200

[a] Excluding the Peoples' Republic of China and the former Soviet Union. [b] Total world. [c] HID is a generic term for high-pressure discharge lamps.[b]

An overview of light and radiation sources related to the main application areas (indoor, outdoor, automotive, and special radiation) is given in Table 8.

Lamps often radiate in the ultraviolet (250 – 380 nm) and near infrared (760 – 2500 nm) as well as in the visible wavelength range. Some lamps can be especially designed to radiate predominately in a nonvisible wavelength range. The applications of lamps as radiation sources are diverse and include the curing of ink, paint and photoresist layers, document copying, photochemical processes, elimination of pollution via photooxidation, and heating (→ Photochemistry). Medical applications include the treatment of skin diseases, high bilirubin levels in babies, and sterilization. The cosmetic use of radiation sources for skin tanning is well known. In horticulture and agriculture they are used to promote photosynthesis in greenhouses, to help with stock raising, and in fishing, especially in fish farms.

Table 8. Overview of light sources and their application areas

	Lamps											
	General incandescent	Tungsten halogen	Fluorescent			High-pressure mercury	Metal halide		Low-pressure sodium	High-pressure sodium	Xenon lamps	LED
			Linear	Compact	Induction		Quartz	Ceramic				
Application area Indoor												
Industry					x	x				x		
Horticulture			x			x	x			x		
Shops/retail	x	x	x	x	x	x	x	x				
Offices		x	x	x								
Homes	x	x	x	x								
Sports			x	x	x	x	x			x		
Public spaces			x	x	x	x	x	x				
Education	x	x	x	x								
Hospitals	x	x		x								
Health care	x	x	x	x								
Projection		x				x	x				x	
Back lighting displays			x								x	x
Transportation	x	x	x									x
Outdoor												
Industry						x	x			x		
Residential areas	x		x	x	x	x		x		x		
Buildings and structures			x			x	x	x		x		
Traffic roads						x			x	x		
Sports							x			x		
Automotive	x	x					x					x
Special radiation												
Printing		x	x				x					
Copying		x	x				x					
Painting		x	x				x					
Photoetching							x				x	
Photochemical			x			x	x					
Photooxidation			x			x	x					
Heating/drying	x	x										
Skin diseases			x				x					
Hyperbilirubinemia		x	x									
Muscle injury	x	x										
Tanning			x				x					
Germicidal/sterilization			x			x	x					

10. References

1. G. Wyszecki, W. S. Stiles: *Colour Science,* 2nd ed., J. Wiley & Sons, New York 1982.
2. J. F. Waymouth: *Electric Discharge Lamps,* M.I.T. Press, Cambridge 1971.
3. C. H. Sturm, E. Klein : *Betriebsgeräte und Schaltungen für electrische Lampen,* Siemens Aktiengesellschaft, Berlin 1992.
4. M. H. Hirsch, H. J. Oskam: *Gaseous Electronics, Electrical Discharges,* **vol.1,** Academic Press, New York 1978. Y. P. Raizer: *Gas Discharge Physics,* Springer Verlag, Berlin 1991.
5. C. Meyer, H. Nienhuis: *Discharge Lamps,* Kluwer Technical Books, Deventer 1988. J. J. de Groot, J. A. J. M. van Vliet: *The High-Pressure Sodium Lamp,* Kluwer Technical Books, Deventer 1986. J. R. Coaton, A. M. Marsden: *Lamps and Lighting,* 4th ed., Edward Arnold, London 1997. J. W. Denneman: "Low-Pressure Sodium Discharge Lamps," *IEE Proc.* **128 A** (1981) 397 – 414. A. G. Jack, L. E. Vrenken: "Fluorescent Lamps and Low-Pressure Sodium Lamps," *IEE Proc.* **127 A** (1980) 149 – 157. J. F. Waymouth: *Electric Discharge Lamps,* M.I.T. Press, Cambridge 1971.
6. "High Temperature Metal Halide Chemistry," *Proc. Electrochem. Soc.* (Pennington, N.J.) **78-1** (1977); "High Temperature Lamp Chemistry" **85-2** (1985); "High Temperature Lamp Chemistry II" **88-4** (1988).
7. H. Schäfer: *Chemical Transport Reactions,* Academic Press, New York – London 1964.
8. *Metals Handbook,* 9th ed., Am. Soc. Metals, **vol. 3,** pp. 314, **vol. 7,** pp. 152, pp. 629.
9. W. D. Kingery, H. K. Bowen, D. R. Uhlmann: *Introduction to Ceramics,* Chaps. 10, 13, J. Wiley & Sons, New York 1976.
10. K. H. Butler: *Fluorescent Lamp Phosphors,* Pennsylvania State University Press, University Park 1980. B. M. J. Smets: "Phosphors Based on Rare-Earths, A New Era in Fluorescent Lighting," *Mater. Chem. Phys.* **16** (1987) 283 – 299. T. Jüstel, H. Nikol, C. Ronda : "New Developments in the Field of Luminescent Materials for Lighting and Displays," *Angew. Chem. Int. Ed.* **37** (1998) 3084 – 3103. C. Ronda: "Enlightenment on Luminescent Materials," in B. DiBartolo (ed.): *Spectroscopy and Dynamics of Collective Excitations in Solids,* Plenum Press, New York 1997, p. 339.
11. J. Bloem, A. Bouwknegt, G. A. Wesselink: "Some New Mercury Alloys for Use in Fluorescent Lamps," *J. Illum. Eng. Soc.* **6** (1977) 141 – 147.
12. H. Köstlin: "Application of Thin Semiconductor and Metal Films in Energy Technology," *Festkörperprobleme* **22** (1982) 229 – 254.
13. M. Koedam: "Materials Vital to the Lighting Industry," *Lighting Design and Application* 15 (Jan. 1985) 18 – 27. CIE Publication No. 77 (1988), "Electric Light Sources – State of the Art – 1987," published by the CIE, Wien. M. Koedam: "Lighting in the Nineties," *J. Light & Visible Environ.* **15** (1991) 79.

Acknowledgements. The present article is a revision of the original version of 1990. The following colleagues within Philips Lighting have contributed to this revision: W. Cottaar, G. Derra, W. van Erk, J. Eijsermans, M. Haverlag, T. Kappen, J. van Kemenade, R. Keijser, B. Kock, J. van Lieshout, B. de Maagt, P. Postma, C. Ronda, B. Smets, M. Stebler, J. Suijker. Thanks are due to D. Work for providing the facilities for the present work.

Laser Dyes

ERWIN THIEL, Universität Siegen, Siegen, Federal Republic of Germany

1. Introduction 1
1.1. History 1
1.2. Dye Laser Operation 2
1.3. Uses 5
2. Classes of Dyes 5
2.1. Polyphenyl and Heteroaromatic Compounds 6
2.2. Stilbenes 7
2.3. Coumarins 8
2.4. Xanthene Dyes 12
2.5. Pyrromethene Dyes 14
2.6. Methine Dyes 15
2.7. Oxazine Dyes 17
3. References 19

This contribution is based on the article of RODERICH RAUE in Ullmann's Encyclopedia of Industrial Chemistry, 5th ed., Vol. A 15.

1. Introduction

Without the invention of the laser (→ Lasers) and the exploitation of its fascinating properties, modern society would be inconceivable. Today laser devices are indispensable in worldwide data transfer, entertainment electronics, medicine, technology, the sciences, etc. Besides the optical resonator, the heart of the laser is the light-amplifying or active medium. Suitable active media are all quantum mechanical systems in which a higher energetic level can be more highly populated than a lower level. Dye molecules occupy an important position among the different families of active media for a number of reasons that are related to their photochemical and photophysical properties. The dye laser is a remarkable and highly versatile source of tunable coherent radiation. Its success and importance are demonstrated by its wide applicability. From basic sciences, such as physics and spectroscopy, to medicine and industry, the dye laser has proved to be an extremely flexible and useful tool. The spectral range of dye laser radiation extends from the UV (300 nm) [81] to the NIR (1800 nm) [82]. Single laser dyes have tuning ranges of up to 100 nm. Dye laser radiation is created in active volumes of 10^{-15} to 1 L. Pulse energies on the order of 1 kJ, ultrashort pulse lengths in the femtosecond range and terawatt peak power, average powers of 2 kW, and conversion efficiencies higher than 80 % can be attained. In focused dye laser beams, electrical fields comparable to interatomic strength have been produced. For detailed information, see the general references.

1.1. History

The first *stimulated emission of radiation* in solution was observed by SOROKIN and LANKARD in 1966 with an alcoholic solution of aluminum chloro phthalocyanine [*14154-42-8*], stimulated by a ruby laser [11]. Independently, SCHÄFER, SCHMIDT, and VOLZE found, in the same year, that cyanine dyes were suitable for generation of laser light [12]. One year later, both groups demonstrated that a flash lamp was a suitable light source for pumping a dye laser [13]. A significant advance in the generation of narrow-bandwidth laser light was made in 1967 by SOFFER and MCFARLAND with the introduction of the reflex grating as a frequency-selective device [14]. In 1969, SNAVELY and SCHÄFER showed that the increasing population of the dye in the triplet state, which interferes with laser emission, occurs to only a limited extent in the presence of oxygen [15]. This result was the precondition for the development of the continuous dye laser, which was realized in 1970 by PETERSON, TUCCIO, and SNAVELY in the laboratories of Eastman Kodak [16]. An argon ion laser was used to stimulate a solution of Rhodamine 6 G. In the late 1980s THIEL et al. showed that continuous dye laser action can be excited by pumping with a commercial 10-mW HeNe laser [83] and – as was desired since the early days of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_151

continuous dye lasers – by pumping with a specially designed noncoherent Xe arc lamp [84]. The development of continuous dye lasers led to the generation of ultrashort laser pulses of 10–30 ps (10^{-12} s), which in 1972 were shortened to 1.5 ps by Dienes [17].

In 1973, Drexhage obtained a significant increase in the efficiency of dye lasers by restricting the possibilities of the dye to undergo radiationless deactivation; this was achieved by using ring closure to increase the rigidity of the molecule [18].

1.2. Dye Laser Operation

Laser Dyes. The properties of laser dyes can be discussed on the basis of the Jablonski energy state diagram [85] (Fig. 1) which is usually accepted for the characterization of organic dyes. The singlet states S_0, S_1, S_2, etc. and triplet states T_1, T_2, etc. of a dye are broad, almost continuous bands because of the large number of vibrational and rotational states and due to interaction with molecules embedding the laser dye. The states S_2, S_3, ... and T_2, T_3, ... are summarized as S_n and T_n. In equilibrium at room temperature only the S_0 state is occupied. The population of the vibrational and rotational sublevels of the S_0 state is given by the Boltzmann distribution.

The dye molecules can be promoted into vibrational and rotational levels of excited electronic singlet states ($S_0 \rightarrow S_1$ or $S_0 \rightarrow S_n$ transition), by absorption of light. A typical absorption spectrum is shown in Figure 2. The excess vibrational energy and the energy difference between S_n and S_1 are given up on a picosecond timescale. The lifetime of the most laser dyes in S_1 state is on the nanosecond timescale. From the S_1 state, several processes can occur. The S_1 state can be deactivated to the ground state by emission of a fluorescence photon or by nonradiative internal conversion. For all UV and VIS laser dyes the quantum yield for fluorescence is quite high ($> 50\,\%$). The fluorescence quantum yield decreases with increasing laser wavelength and reaches 10^{-4} for dyes emitting above 1500 nm [82]. With a low probability, dye molecules can reach the T_1 state. The lifetime of molecules in the T_1 state is relatively long (microseconds to seconds) because the transition to the ground state (phosphorescence) is spin-forbidden.

In comparison to the vibration period the lifetimes of the S_1 and T_1 states are much longer. Hence, in these excited states a thermal equilibrium is reached. In a certain sense one the absorption of light can be regarded as the formation of a new molecule. This holds for the geometrical, dielectrical, optical, and chemical properties of the excited state. The energy of the absorbed light quantum (visible or UV) is comparable to the chemical bonding energy. Therefore the properties of the excited molecule differ strongly from those of the ground state.

For the laser efficiency the properties of the excited molecule are of decisive importance. By means of stimulated emission, molecules in the S_1 state can act as light amplifiers, a prerequisite for the lasing process (see Fig. 1). However, it has been assumed since the very beginning of dye lasers that molecules in the S_1 and T_1 states may disturb or prevent laser action due to their absorption properties. Here the absorption mainly of laser light but also of pump light must be taken into account.

Even very low absorption cross sections of the S_1 or T_1 state can reduce the laser efficiency. Consequently for efficient laser dyes in particular, the excited state properties are difficult to measure. The performance of the classical photolysis techniques is insufficient for investigating the excited dye molecules with the required precision. During the first 25 years after the invention of the dye laser, published information on the physical and chemical properties of the first excited singlet and triplet states of laser dyes was scarce and controversial. Consequently, dyes were selected by trial and error. In the early 1990s, a laser-scanning spectroscopic technique was developed [86] with which the molecular parameters important for dye laser action were measured for a number of laser dyes with high precision [87]. It turned out that the spin-allowed $S_1 \rightarrow S_n$ and $T_1 \rightarrow T_n$ absorptions are the reason for the fact that of the thousands of known dye structures, only very few molecules are efficient laser dyes or even laser active. In fact for the most efficient laser dyes the main $S_1 \rightarrow S_n$ and $T_1 \rightarrow T_n$ absorption band lies outside the spectral region of the main $S_0 \rightarrow S_1$ absorption or fluorescence.

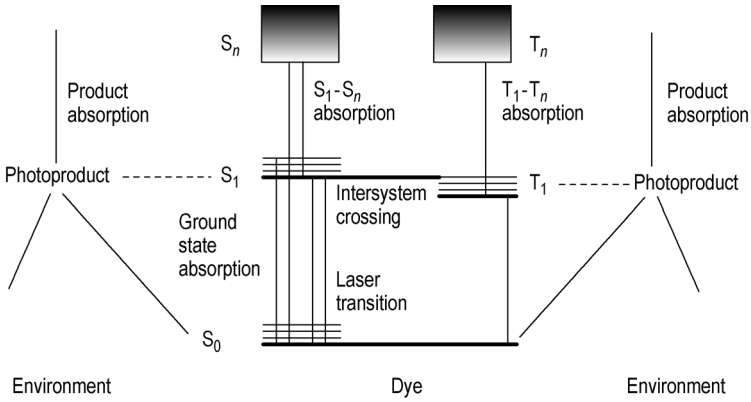

Figure 1. Extended energy level diagram
Processes which disturb lasing: excited state absorption, re-absorption and absorption of photoproducts
Processes that do not necessarily disturb the lasing process: intersystem crossing, photoreactions from S_1 and T_1 states

Besides the properties of the excited molecules, the absorption of the remaining ground-state molecules are of interest. These molecules can disturb the laser process by absorption (so-called re-absorption) of the created laser light. The importance of this process decreases with increasing narrowness of the ground state absorption band and increasing magnitude of the Stokes shift. Disturbances caused by re-absorption also can be reduced by using a lower dye concentration. However this again leads to less efficient absorption of the pump light. To absorb the pump light even at low dye concentrations, multichromophore dyes are used. In these dyes additional pump light absorbers (antennas) are covalently bound to the laser chromophore. The chromphores are separated by mesomeric structures. After the antennas are excited, the laser chromophore is excited by Förster energy transfer. It was shown that the laser efficiency of multichromophore dyes is generally reduced by a compulsive appearance of excited state absorption caused by the antennas [88].

To classify the importance of the dye properties discussed above, it is useful to distinguish between short- and long-pulse pumped dye lasers. When the temporal profile of the pump intensity is much shorter than the triplet lifetime, the triplet state absorption is of less importance. Hence, by applying a sufficient pump intensity the efficiency of the active medium is exclusively determined by the competition between stimulated emission and $S_1 \rightarrow S_n$ absorption. In practice, for good laser dyes at the maximum of the tuning range, the $S_1 \rightarrow S_n$ absorption reduces the lasing efficiency by $1-10\,\%$. At shorter wavelengths the tuning range is generally limited by re-absorption [89]. If the temporal profile of the pump intensity is comparable to or even longer than the triplet lifetime, then, besides the $S_1 \rightarrow S_n$ absorption, the population (accumulation) of molecules in the T_1 state will be of particular interest. Molecules in the T_1 state are removed from the pump–lasering cycle. These molecules can be compensated by an increased dye concentration. Therefore, this process does not disturb the laser process noticeably. However, molecules in the T_1 state may absorb pump or laser light and thus reduce laser efficiency or prevent lasing altogether. This disturbance can be reduced by decreasing the triplet state population. In all liquid solutions used as active dye laser media, the triplet state lifetime is determined by intermolecular quenching with dissolved molecular oxygen [90]. With the exception of the solvent ethylene glycol, which is used in continuous-wave dye lasers, the triplet lifetime in air-saturated solvents that are used in dye lasers is limited to values of less than 1 μs. A further reduction in the triplet lifetime has been achieved by using additional triplet quenchers, for example, unsaturated hydrocarbons such as cycloheptatriene or cyclooctatetraene as well as 2,2,6,6-tetramethylpiperidine or (+)-(*S*)-1-(9-anthryl)-2,2,2-trifluorethanol [90], [21], [22]. Lüttke et al. combined a laser dye with a triplet quencher

that is covalently bound to the dye but mesomerically isolated. The resulting quenching effect is quite pronounced when the connecting link between the fluorescent oxazole compound and *trans*-stilbene is a CH_2 group [23]. The organic quenchers tend to be unpleasant and difficult to use. So far in practice these agents are of minor importance.

In most continuous-wave dye laser systems, ethylene glycol is used as solvent. Owing to its high viscosity and its low oxygen solubility the triplet lifetime in this solvent is generally longer than 5 μs. Under these conditions even rhodamine 6G, the most popular and one of the most efficient laser dyes is laser inactive [91]. Only fast removal of the dye molecules in the triplet state will make laser action possible. This is achieved by using a fast-moving dye solution (jet technique).

Photodecomposition. The chemical properties of excited dye molecules are assumed to differ from those of the ground-state molecules. Given the intense excitation in the laser, intermolecular chemical reactions between dye molecules in the S_1 and T_1 states and reactants are of twofold interest. First, the excitation photon which initiates the photoreaction is lost to the laser process, and the photoproducts may absorb pump or laser light. Both processes reduce the laser efficiency, even that of a new laser solution. Second, possible irreversible photoreactions also lead to photodegradation and therefore to a light aging of the active medium. Radicals and singlet molecular oxygen are frequently discussed as reactants. To date, published information concerning the nature of the photodecomposition even of frequently used laser dyes is scarce and controversial. A review is given in [92]. In practical applications, the durability of the laser medium seems to be mainly determined by small amounts of impurities present in the dye or the solvent. It is generally absorption at the laser wavelength rather than the loss of dye that has the major degrading effect on laser output. In some cases the addition of agents (ascorbic acid, triplet quenchers, radical interceptors, etc.) or the removal of oxygen stabilizes the dye solution. So far in practice this technique is of minor importance.

Dye Laser Media. Besides the laser dye itself, the matrix which embeds the laser dye is of major importance not only for laser efficiency but also for the durability of the laser medium and the beam quality. In most dye laser devices the laser dye is dissolved in liquid solvents like ethanol, methanol, dimethyl sulfoxide, ethylene glycol, water, etc. or mixtures thereof. As discussed above the viscosity and the oxygen solubility of the solvent determines the triplet lifetime and to a certain extent the intersystem crossing rate constant the of the laser dye.

Water is a particularly suitable solvent for dye lasers, because at room temperature the temperature dependence of the refractive index is one order of magnitude lower than those of organic solvents. Around 0 °C the temperature dependence of the refractive index completely disappears. Hence, in water the inevitable nonuniform heating by the pump light does not lead to massive beam disturbances. A disadvantage is that many dyes form aggregates in aqueous solution. This can be prevented by the addition of surface-active compounds that can form micelles [24]. Aggregation of the dye can also be curtailed by addition of fluoroalcohols [25].

Intensive effort was devoted to impregnating organic dyes into various matrices, with the aim of achieving solid-state dye lasers [93]. Polymers, nanoporous hosts, silica gels, xerogels, alumina gels, ormosils, and composite glasses were used as matrices. Solid-state lasers have advantages over liquid dye lasers by being nonvolatile, nonflammable, nontoxic, compact, and mechanically stable. Problems such as heat dissipation and photodegradation, however, still remain to be solved for high-power applications. For long-pulse pumped and continuous-wave dye lasers an efficient triplet state quenching has also not yet been realized.

Excitation Sources. All practical applied dye lasers are pumped optically. Lasers such as Ar ion, Kr ion, nitrogen, Nd:YAG, Cu vapor, excimer, and semiconductor lasers, as well as noncoherent light sources such as Xe flash and continuous lamps are used as pump sources. Electronically pumped laser action of vaporized dye molecules is also possible [94] but has no practical relevance.

A current challenge is the electronic excitation of conducting polymers impregnated with a

laser dye. Although optical pumped laser action has already been achieved, so far no laser action could be obtained from electrically excited devices.

1.3. Uses

The dye laser is used in high-resolution spectroscopy, for example, LIF (laser-induced fluorescence), absorption, LIDAR (light detection and ranging), coherent and incoherent elastic and inelastic scattering methods, REMPI (resonance-enhanced multiphoton ionization), etc. [95]. Today these spectroscopic tools are widely used in biology, medicine, technology, fundamental science, etc. The high intensity and the narrow beam of laser radiation allow analytical measurements to be carried out on samples that are in a great distance from the measuring equipment. For example, a sodium concentration in the atmosphere at a height of 80 – 110 km of 10^9 atoms per cubic meter can be detected [27]. Similarly, one molecule of NO_2 in 10^7 molecules of air can be detected at a distance of 4 km [28].

The dye laser allows isotope separation with a high enrichment factor because of its very narrow linewidth. In a mixture of two chemically identical molecules which can be differentiated solely by isotopic composition, selective reaction of one type of molecule can be carried out if the absorption spectra, on the grounds of the isotopic shift, are so different that only one type of molecule absorbs laser light whereas the other type remains in the chemically inactive ground state. For enrichment of the isotope ^{37}Cl, iodine monochloride vapor in acetylene was excited with the continuous radiation of a Rhodamine 6 G dye laser. The selectively excited ^{37}ClI reacted with acetylene to form 1-chloro-2-iodoethylene [29]. By stepwise photochemical decomposition of symmetrical tetrazine molecules containing only ^{14}N or a mixture of ^{14}N and ^{15}N to hydrogen cyanide and nitrogen by irradiation with a dye laser (containing sodium fluorescein as active medium) pumped with an argon laser, nitrogen isotopes ^{14}N and ^{15}N could be separated [30]. Moreover, promising first attempts have been made for isotopic separation of ^{235}U and ^{238}U with a dye laser.

The ultra-short light pulses (femtosecond timescale) which can be produced by dye lasers are extensively used, for example, for the investigation of elementary chemical reactions (femtochemistry). Breaking and formation of chemical bonds have been observed in real time [96]. Nowadays, an exciting challenge is the time-resolved observation of electronic transitions. For this purpose light pulses on the attosecond time scale are necessary. Owing to the broad emission spectra of organic dyes, this aim should be reached in the near future.

Laser dyes include a large number of long-known fluorescent compounds such as Rhodamine 6 G. Other representatives are optical brighteners (→ Optical Brighteners) and fluorescent dyes (→ Fluorescent Dyes, Chap. 1.) from the large chemical companies (e.g., Bayer, CIBA-Geigy, Hoechst, Eastman Kodak). The range of laser dyes is completed by compounds that have resulted from university research. Special firms have overtaken the distribution of laser dyes: Lambda-Physik GmbH, Göttingen (Lambdachrome laser dyes); Exciton Chemical Co., Dayton, Ohio; Radiant Dyes Chemie, Wermelskirchen. An exception among the larger companies is Eastman Kodak, Rochester, New York, which offers laser dyes as Kodak Laboratory Chemicals.

2. Classes of Dyes

Since the invention of dye lasers, a considerable number of dye structures were found to be suitable for generating laser radiation. Some extensive systematic reviews are given [97–100]. Here only the most important dye classes are discussed.

A dye is generally understood to signify a compound that absorbs light in the visible region of the spectrum. The concept of laser dyes is considerably broader. The spectral range of laser emission is 300 – 1800 nm; that is, it embraces long-wavelength ultraviolet and the beginning of the infrared. The structures of the compounds used are also correspondingly diverse, their common feature being a high fluorescence quantum yield.

2.1. Polyphenyl and Heteroaromatic Compounds

The short-wavelength part of the spectrum is dominated by aromatic compounds, at least those not substituted with donor groups. Condensed aromatic ring systems, such as anthracene, pyrene, or perylene possess only moderate laser properties because of their excited-state absorption. This is not true for those systems formed by linear coupling of benzene rings.

In contrast to biphenyl *p*-terphenyl [*92-94-4*] is an efficient laser dye. When excited with a KrF excimer laser, the tuning range is 322 –352 nm. Substitution of *p*-terphenyl in the 2,2″-positions with methyl or trifluoromethyl groups leads to twisting of the terminal benzene rings about the long molecular axis as a result of steric hindrance; this produces a moderate hypsochromic shift, resulting in a tuning range of 308 – 352 nm [31].

p-Quaterphenyl [*135-70-6*] is tunable over the range 360 – 390 nm by using a nitrogen laser for excitation. Introduction of methyl groups into the terminal rings in sterically hindered positions also causes a hypsochromic shift and a reduction in intensity of the fluorescence band, again because of twisting. Excitation with an XeCl excimer laser gives a tuning range of 335 – 375 nm [32]. Substitution with a sulfonic acid group in the *p*-position of quaterphenyl renders the compound sufficiently soluble in ethylene glycol. It is commercially available as Polyphenyl 1 (**1**) (Lambda-Physik) for use in cw lasers. The tuning range attainable by stimulation with an argon ion laser is 362 – 412 nm. Polyphenyl 1 is produced by palladium-catalyzed dimerization of 4-bromobiphenyl-4′-sulfonic acid [33]:

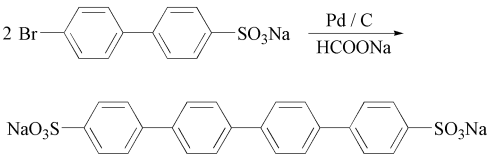

1

Substitution of quinquephenyl with methyl or *tert*-butyl groups leads similarly to effective laser dyes which, by stimulation with an XeCl excimer laser, are tunable from 355 to 400 nm [34].

The exchange of phenyl groups for heterocyclic groups leads to a shift of the absorption band to long wavelengths. Furan 1 (**2**) (Lambda-Physik) is produced by condensation of salicylaldehyde with bis(chloromethyl)biphenyl, followed by ring closure and subsequent sulfonation:

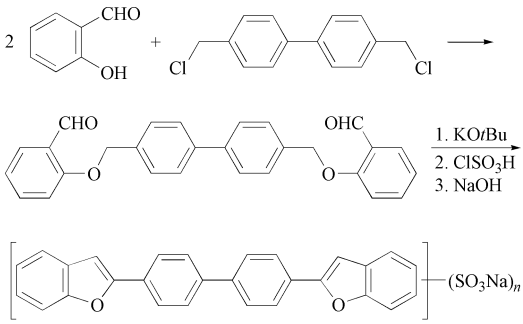

$n = 2, 3$ **2**

When dissolved in methanol and stimulated with an XeCl excimer laser, Furan 1 can be tuned from 410 to 435 nm [35].

Another heteroaromatic polyphenyl compound is 1,4-phenylenebis(5-phenyl-2-oxazole) [*1806-34-4*] (**3**) (POPOP). By coupling of two fluorescent compounds, dimethyl-POPOP (**3a**),

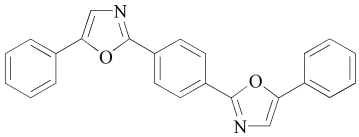

3

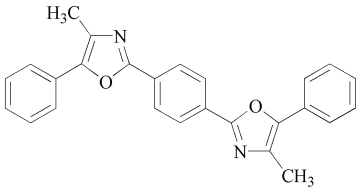

3a

and terphenyl via the methyl group SCHÄFER and LÜTTKE obtained bifluorophoric laser dyes e.g., 1-(4-methyl-5-phenyloxazolyl-2)-4-(4-terphenylylmethyl-5-phenyloxazolyl-2)-benzene (**4**) which, through energy transfer,

enables a greater part of the excitation energy to be used [36].

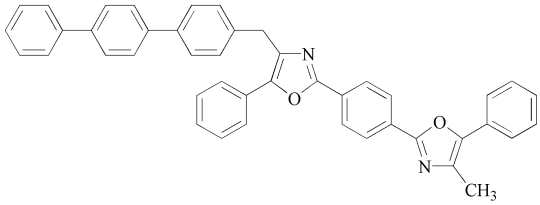

4

2.2. Stilbenes

4,4′-Diphenylstilbene [*26569-48-2*] is suitable for pulse operation with an XeCl excimer laser or a nitrogen laser; the tuning range is 395 – 415 nm. Introduction of sulfonic acid groups in the *para* positions gives a laser dye that has sufficient solubility in ethylene glycol to be suitable for cw operation; stimulation with an argon ion laser gives a tuning range of 395 – 435 nm [37]. This dye is available under the name Stilbene 1 (**5**) (Lambda-Physik). Its lightfastness is strongly dependent on the mode of excitation. With cw stimulation the input energy amounts to 200 Wh measured until the output of the dye laser energy has dropped to 50 % while in case of excimer stimulation it is only 10 Wh [38]. The photokinetics of laser dyes were investigated by Gauglitz with fluorescence spectroscopy; a rapid decline in performance occurs because of *cis* – *trans* rearrangement, followed by slow decomposition of the *cis* isomer [39].

The synthesis of Stilbene 1 can be carried out by Wittig – Horner reaction of the potassium salt of 4-carboxaldehydebiphenyl-4′-sulfonic acid with 4-diethoxyphosphonomethylbiphenyl-4′-sulfonic acid ethyl ester in dimethylformamide, in the presence of potassium *tert*-butoxide, and subsequent saponification [40]:

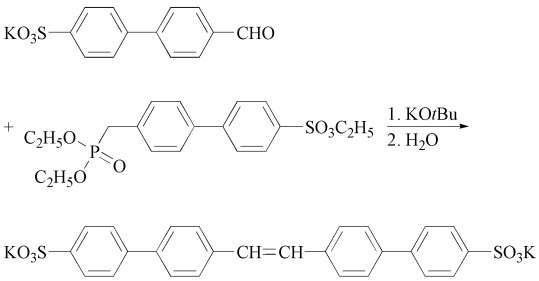

5

Majewski and Krasinski investigated the laser activity of the optical brightener **6** derived from 4,4′-diaminostilbene-2,2′-disulfonic acid. They used only compounds which had no further sulfonic acid groups [41].

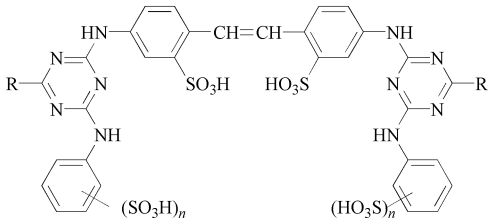

6

The brightener **6** with $R = OCH_3$ and $n = 0$ [*3426-43-5*] has a tunable range of 415 – 445 nm when stimulated with a nitrogen laser. The tuning range of the compounds **6** is increased considerably by the introduction of additional sulfonic acid groups into the phenyl substituents. The brightener with $R = N(C_2H_4OH)_2$ and $n = 1$ [*12224-02-1*] can be tuned from 417 to 464 nm [42]. Furthermore, these compounds are suitable for cw operation. The tuning range of the dye with $R = N(C_2H_4OH)_2$ and $n = 1$ is 410 – 480 nm [43].

Distyrylbiphenylsulfonic acids constitute a particularly useful class of laser dyes [44]. The laser dye Stilbene 3 (**7**) (Lambda-Physik) in methanol solution can be tuned over the ranges 412 – 443 nm and 408 – 457 nm by using an XeCl excimer laser and a nitrogen laser, respectively. In aqueous solution, the tuning curve is shifted to long wavelength, reaching 420 – 470 nm. Addition of a nonionic surfactant increases the laser performance by deaggregation [45]. Stilbene 3 (**7**) [*27344-41-8*] is particularly suitable for use in cw dye lasers.

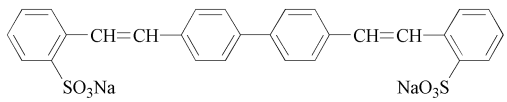

7

Stimulation with an argon ion laser gives a particularly broad tuning range of 403 – 493 nm

[43]; its lightfastness is strongly dependent on the mode of excitation. With cw stimulation the input energy amounts to 300 Wh measured until the output of the dye laser energy has dropped to 50 % while in case of excimer stimulation it is only 14 Wh [38].

2.3. Coumarins

A large number of highly effective laser dyes can be found among the coumarins. Their tuning range is 400 – 600 nm; with specific substitution this can reach 700 nm. Most coumarin laser dyes contain an amino group in the 7-position. 7-Hydroxycoumarins have also been described as laser dyes.

4-Methyl-7-hydroxycoumarin [*90-33-5*] in alkaline solution can be tuned from 420 to 520 nm by excitation with a flash lamp [46]. With a nitrogen laser, the tuning range is 370 – 580 nm; with continuous stimulation by an argon ion laser, it is 460 – 560 nm. This unusually broad tuning range results from the large number of possible mesomeric structures, including (**8a**, **8b**, and **8c**):

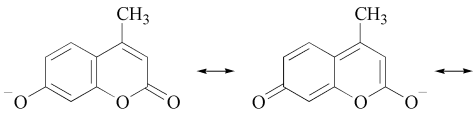

8a **8b**

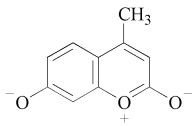

8c

Laser activity can also be observed in strongly acidic solution, due to the pyrylium structure **8c** [47]. The enolate anion at the 7-position of coumarin may also be replaced by the phenylsulfonylamino group [48].

The oldest aminocoumarin dye is 7-dimethylamino-4-methylcoumarin (**9**), which VON PECHMANN obtained by condensation of 3-dimethylaminophenol with ethyl acetoacetate in the presence of zinc chloride [49]:

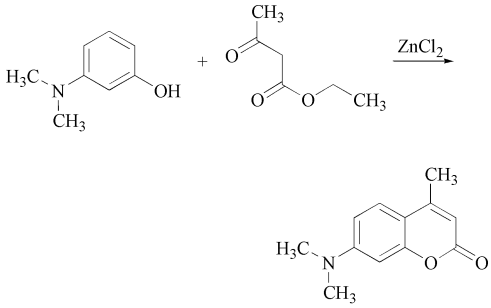

9

This compound is available as Coumarin 311 (Kodak) [46]. The parent compound 7-amino-4-methylcoumarin [*26093-31-2*] is available as Coumarin 120 (Kodak). This dye is tunable over 423 – 462 nm by stimulation with an XeCl excimer laser, and over 425 – 475 nm by continuous stimulation with an argon ion laser. The tuning range of 7-diethylamino-4-methylcoumarin [*91-44-1*] (Coumarin 1, Kodak) is 436 – 486 nm (nitrogen laser stimulation) and 450 – 500 nm (continuous stimulation with an argon ion laser). A disadvantage of this laser dye is the low photostability. Particularly when a flash lamp is used for excitation, the 4-methyl group undergoes photochemical oxidation to a carboxyl group. Even a small amount of photooxidation product can reduce the laser radiation because the absorption wavelength of the carboxylic acid derivative lies within the range of the laser wavelength of the 4-methyl compound. SCHIMITSCHEK was able to raise the photostability of coumarin compounds considerably by introduction of the CF_3 group in place of CH_3 [50]. 7-Dimethylamino-4-trifluoromethylcoumarin [*53518-14-2*] (Coumarin 152, Kodak) has a tuning range of 490 – 570 nm by stimulation with a XeCl excimer laser.

The sodium salt of 4-methyl-7-sulfomethylaminocoumarin (Coumarin 175, Kodak) is water-soluble and thus insensitive to formation of optical inhomogeneities when a flash lamp is used for excitation [51].

An increase in fluorescence quantum yield can be obtained by limiting the opportunities for radiationless deactivation. In Coumarin 1, loss in energy results from rotation of the diethylamino group; this rotation is cut off by ring closure with the phenyl group. This effect was called "rigidizing" by DREXHAGE. Using 3-

Table 1. Tuning ranges of some coumarin laser dyes

Dye	CAS registry number	Structure	Tuning range, nm	
			XeCl Excimer	cw Argon
Coumarin 102	[*41267-76-9*]	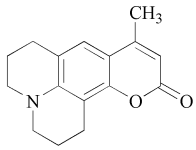	460 – 510	460 – 560
Coumarin 153	[*53518-18-6*]	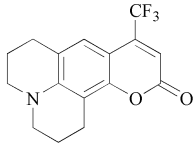	522 – 600	
Coumarin 6	[*38215-36-0*]	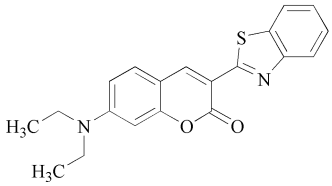	515 – 558	520 – 580
RDC 650	[*95523-61-8*]	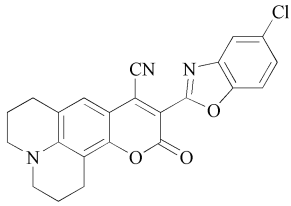	600 – 705	
Coumarin 7	[*27425-55-4*]	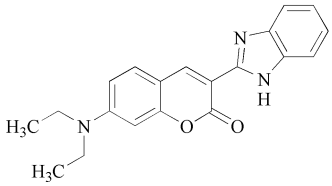		495 – 570
Coumarin 30	[*41044-12-6*]	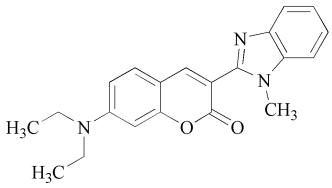		480 – 555 [a]

[a] cw Krypton laser.

hydroxyjulolidine, he produced a large number of highly effective coumarin laser dyes, including Coumarin 102 and Coumarin 153 (Kodak, Lambda-Physik) (see Table 1) [52].

Introduction of an N-heterocyclic group in the 3-position of the coumarin system shifts the tuning range of 7-aminocoumarin compounds to longer wavelengths. These dyes were first produced in 1958 for dyeing textiles (→ Methine Dyes and Pigments). Coumarin 6 (**10**, Kodak, Lambda-Physik) is obtained by reaction of 2-hydroxy-4-diethylaminobenzaldehyde with 2-cyanomethylbenzothiazole in alcohol by using piperidine as catalyst [53] (see also Table 1):

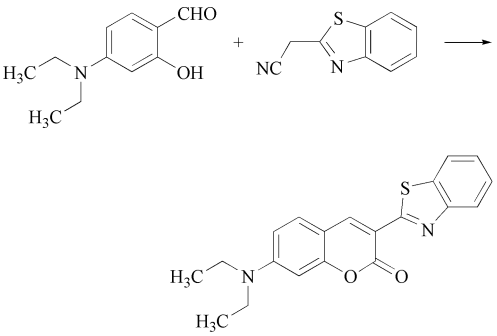

10

DREXHAGE et al. produced a continuous emission with coumarin 6 by dissolving it in aqueous solution by adding *N,N*-dimethyldodecylamine *N*-oxide, with cyclooctatetraene as triplet quencher, and using an argon ion laser for stimulation [54]. Coumarin 7 and Coumarin 30 contain the benzimidazolyl group in the 3-position of the coumarin (see Table 1).

Other efficient laser dyes have a benzoxazolyl group in the 3-position [55]. A cyano group can be substituted in the 4-position of these dyes by addition of sodium cyanide and subsequent oxidation with bromine [56]. By this means, the tuning range of coumarin dyes can be extended to 705 nm [57], e.g., RDC 650 (Radiant Dyes Chemie) (see Table 1).

Azacoumarin laser dyes (e.g., **11** [*57980-06-0*]) are obtained by treatment of 2-hydroxy-6-dialkylaminopyridine with acetoacetates [58]:

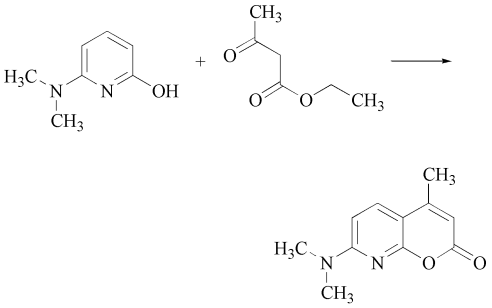

11

Table 2. Carbostyril laser dyes

Dye	Carbostyril 7	Carbostyril 3
CAS registry number	[*19840-99-4*]	[*58336-39-3*]
Structure	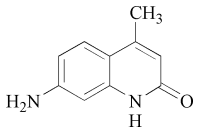	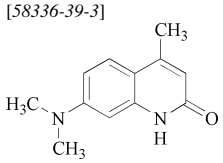
Tuning range, nm		
Flash lamp	408 – 420	420 – 440
cw argon		415 – 490

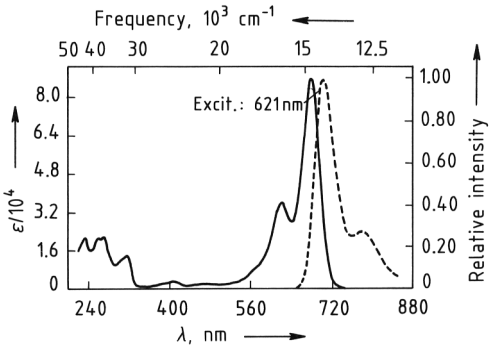

Figure 2. Absorption (—-) and fluorescence (- - - -) spectra of Rhodamine 800

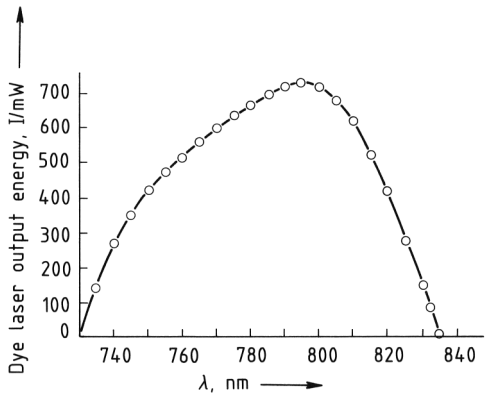

Figure 3. Tuning curve of Rhodamine 800

Table 3. Some xanthene laser dyes

Dye	CAS registry number	Structure	Tuning range, nm		
			XeCl Excimer	Flash lamp	cw Argon
Sodium fluorescein	[*518-47-8*]	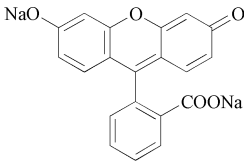	532 – 561	549 – 574	530 – 590
Rhodamine 6G	[*989-38-8*]	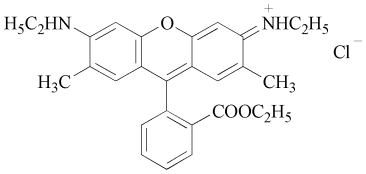	569 – 608	555 – 620	560 – 650
Rhodamine B	[*81-88-9*]	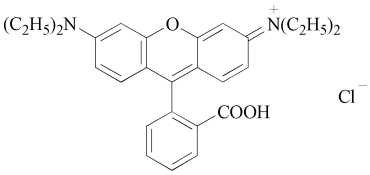	588 – 644	590 – 640	605 – 675
Rhodamine 101	[*64339-18-0*]	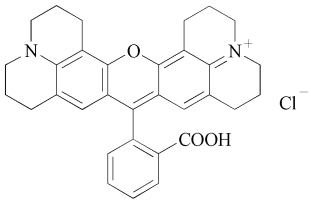	614 – 672	623 – 676 [a]	620 – 700
Rhodamine 630		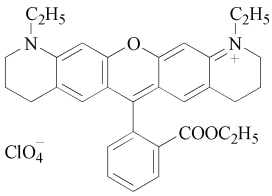	595 – 680 [36]	595 – 660 [37]	595 – 680 [38]
Sulforhodamine B	[*2609-88-3*]	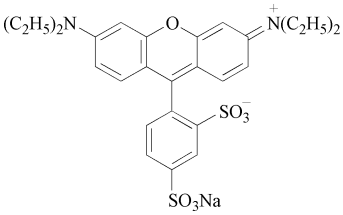	594 – 642		610 – 680

[a] With N_2 laser. [b] cw Krypton laser.

Table 3. (Continued)

Dye	CAS registry number	Structure	Tuning range, nm		
			XeCl Excimer	Flash lamp	cw Argon
Sulforhodamine 101	[*60311-02-6*]	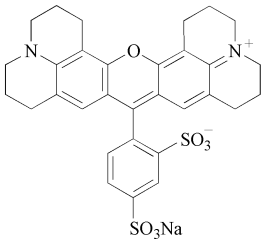	616 – 667		636 – 675
Rhodamine 700	[*63561-41-1*]	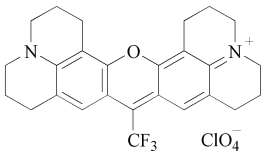	701 – 768	705 – 798	700 – 810 [b]

[a] With N_2 laser. [b] cw Krypton laser.

Further effective laser dyes are found in the range of the carbostyril compounds (see Table 2). By analogy to coumarins, they are obtained by reaction of 1,3-diaminobenzene and its alkyl derivatives with ethyl acetoacetate [59].

Chinolon 390 (**12**) is an aza analogue of the carbostyril dyes, prepared by treatment of 2,6-diaminopyridine with malonic ester and subsequent alkylation [60]:

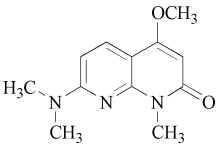

12

Nd-YAG 3rd 380 – 400 nm

2.4. Xanthene Dyes

Xanthene dyes result from insertion of an oxygen bridge in triphenylmethane dyes (→ Triarylmethane and Diarylmethane Dyes, Chap. 6.). In this way, the rotation of phenyl groups is hindered and the possibilities of radiationless deactivation are restricted. These dyes show strong fluorescence and were discovered relatively early as laser dyes (see Table 3). A ruby laser was originally used for stimulation, followed by a flash lamp, the latter is characterized by a steep increase in energy. The preferred dyes are sodium fluorescein and Rhodamine 6G (see Table 3) [61]. From the early days of dye lasers until the early 1990s, Rhodamine 6G remained the universally applicable laser dye with the highest performance. Today, only very few dyes (Rhodamine 630, see Table 3; Pyrromethene 597, see Section 2.6) reach a slightly higher laser efficiency in some laser systems.

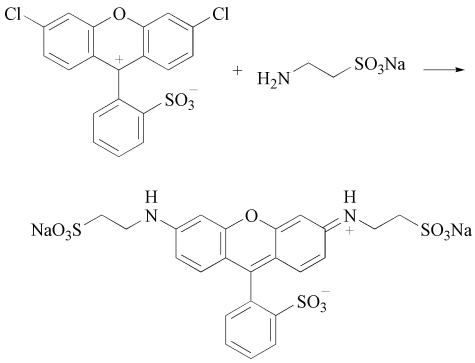

13

A mixture of two laser dyes can lead to an increase in radiation intensity. This is the case, for example, with a 1 : 1 mixture of Rhodamine B and Rhodamine 6 G [61].

The fluorescence quantum yield of Rhodamine B is temperature-dependent: at room temperature, it is ca. 40 %, and it can be increased to almost 100 % by reducing the temperature; in contrast, at the boiling point of ethanol it is only a few percent. This is an example of radiationless deactivation by thermal stimulation [62]. Coupling the two alkyl groups with the aromatic ring, e.g., in Rhodamine 101, strongly reduces the mobility of the amino group; the fluorescence quantum yield then reaches almost 100 % at room temperature [52] (see also Table 3).

Conversion of 3,6-dichloro-9-phenylxanthylium-2′-sulfonate with taurine [*107-35-7*] in aqueous solution leads to Rhodamine 504 (**13**, Radiant Dyes Chemie) whose tuning range with continuous excitation (550 – 622 nm) is shifted to shorter wavelengths compared to Rhodamine 6 G (Table 3) [63].

Sulforhodamine B [64] and Sulforhodamine 101 (Kodak) are suitable for continuous stimulation (see Table 3).

A cyanide group can be introduced into the 9-position of the xanthene ring system by addition of alkali cyanides and oxidation with bromine to give (**14**) (Rhodamine 800, Lambda-Physik, RDC 780, Radiant Dyes Chemie). The fluorescence is shifted 100 nm to longer wavelengths by this means. The absorption and fluorescence spectra of dye **14** in methanol are shown in Figure 2.

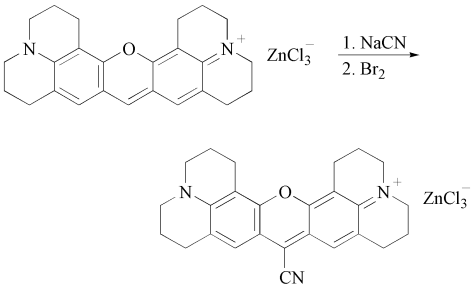

14

These dyes are characterized by a particularly broad tuning range; with a krypton ion laser, **14** can be tuned over 730 – 835 nm [65] (see Fig. 3).

A CF_3 group in the 9-position of the xanthene system produces a similar bathochromic effect. The tuning range of Rhodamine 700 (Lambda Physik) is 700 – 810 nm (see Table 3).

Table 4. Rhodamine and phyronine dyes with fixed amino groups

Dye	Structure	ϕ_{fl},[a] %	$\Delta\lambda$,[b] nm
Rhodamine 101 (acid)	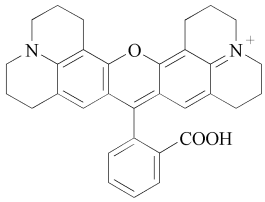	93	21[c]
Rhodamine 630	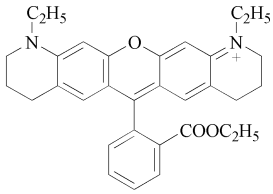	97	25[c]
Pyronine 105	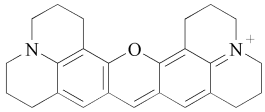		18[d]
KS 35	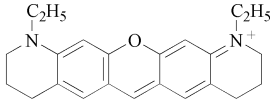	94	20[d]
Rhodamine 700 (acid)	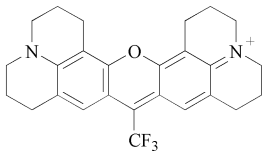	38	26[d]
JA 47 (acid)	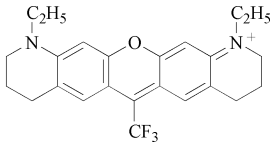	62	31[d]
Rhodamine 800 (acid)	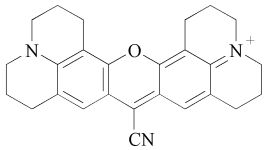	19	26[c]
CNR 1 (acid)	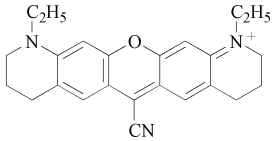	26	33[c]

[a] Fluorescence quantum yield. [b] Stokes Shift. [c] In ethylene glycol. [d] In ethanol.

As shown in Table 3 the fixation of the amino groups by the julolidine structural motif leads to a number of efficient laser dyes. Their laser efficiency can be further increased if the amino groups are embedded in a single six-membered ring instead of the julolidine backbone [103] (Table 4). this dye class is distinguished by a higher fluorescence quantum yield and – more important for the increased laser efficiency – by reduced re-absorption [101].

2.5. Pyrromethene Dyes

Pavlopoulos and Boyer [104], [105] established pyrromethene BF_2 complexes as laser media. Phosphoric acid at 140 °C converts ethyl-3,5-dimethylpyrrole-2,4-dicarboxylate to 2,4-dimethylpyrrole. Treatment with acetyl chloride at 40 °C converts the latter to 1,3,5,7,8-pentamethylpyrromethene hydrochloride, and subsequent treatment with triethylamine and boron trifluoride etherate affords 1,3,5,7,8-pentamethylpyrromethene BF_2 complex (Pyrromethene 546).

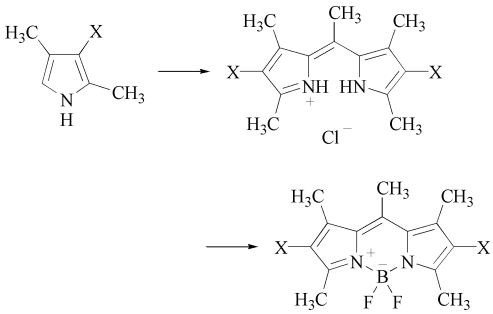

X = H, CH_2, etc.

Table 5. Pyrromethene laser dyes

Dye	Structure	Pump source	Tuning range, nm
Pyrromethene 546	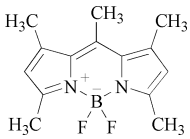	flash lamp	520 – 580
Pyrromethene 556	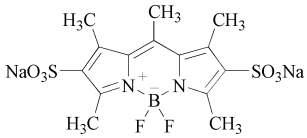	flash lamp/cw Ar	540 – 580/530 – 624
Pyrromethene 567	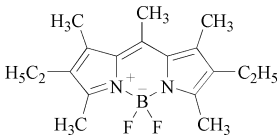	cw Ar/Nd:YAG ∗	552 – 608/549 – 592
Pyrromethene 580	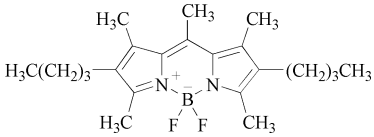	flash lamp	
Pyrromethene 597	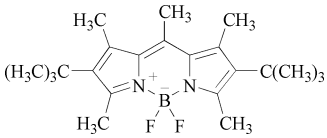	flash lamp	
Pyrromethene 650	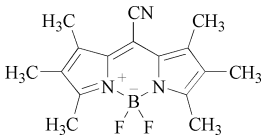	Nd:YAG ∗	

∗ Frequency-doubled.

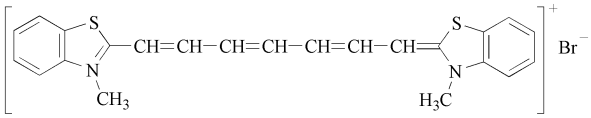

15

In similar manner, the Pyrromethene dyes shown in Table 5 are synthesized. These efficient and stable laser dyes [93], [100] can be used for pulsed and continuous operation.

2.6. Methine Dyes

SCHÄFER et al. irradiated a 10^{-4} M solution of dye **15** (3,3′-diethylthiatricarbocyanine bromide) in acetone with pulses from a 5-MW ruby laser and observed that the broad fluorescence band of spontaneous emission was reduced to a narrow, intense fluorescent band. This was the first observation of stimulated emission of a methine dye [12] (see Fig. 4).

The iodide salt of this dye is available commercially as DTTCI [*3071-70-3*] (Lambda Physik). Stimulation with an XeCl excimer laser gives a tuning range of 828 – 883 nm. Benzannelation of both terminal benzthiazole groups leads to a shift of the fluorescence band to longer wavelengths. The tuning range of 3,3′-diethyl-4,4′5,5′-dibenzothiacarbocyanine iodide [*23178-65-6*] is 899 – 975 nm. Further heptamethine dyes are listed in Table 6 [66].

These dyes can also give continuous radiation by stimulation with a krypton ion laser [67].

Insertion of a ring system in the polymethine chain raises the photostability of the dye considerably. A suitable intermediate is the dialdehyde **16**, which ZEMLICKA and ARNOLD obtained from reaction of cyclohexanone with phosphorus oxychloride and dimethylformamide by the Vilsmeier reaction [68]:

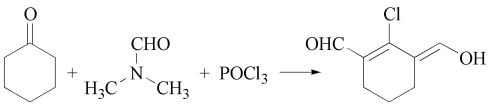

16

Condensation of **16** with 2,4-diphenyl-5,6-tetramethylenepyrylium perchlorate gives the infrared laser dye IR 5 [*61010-01-3*] (**17**, Lambda Physik):

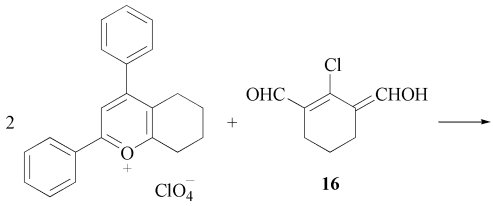

16

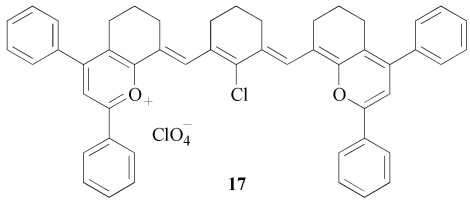

17

Stimulation with an Nd – YAG laser allows the dye to be tuned over 1180 – 1400 nm [69]. DREXHAGE used the same aldehyde for synthesis of the laser dye IR 26 [*76871-75-5*] (**18**, Lambda Physik), tunable from 1200 to 1320 nm (Nd – YAG laser) [70]:

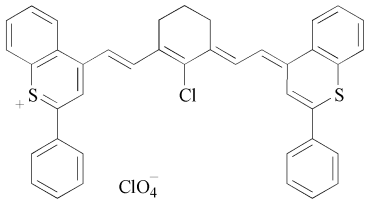

18

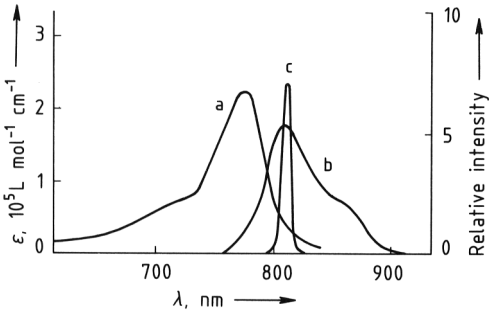

Figure 4. Absorption and fluorescence spectra of 3,3′-diethylthiatricarbocyanine bromide (**15**)
a) Absorption; b) Spontaneous emission; c) Stimulated emission

Table 6. Tuning ranges of some heptamethine (cyanine) laser dyes

Dye	CAS registry number	Structure	Tuning range, nm	
			XeCl Excimer	cw Argon
DTTCI	[*23178-65-6*]	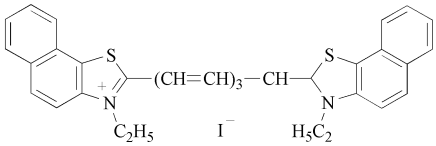	899 – 975	
DMOTC	[*19764-95-5*]	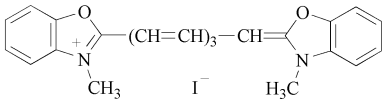	774 – 810	745 – 790
HITCI	[*19764-96-6*]	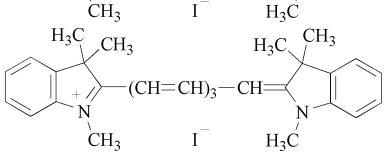	837 – 905	815 – 920
IR 125	[*3599-32-4*]	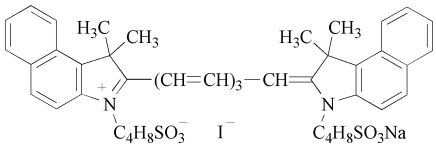	890 – 960	

Pyrylium dyes lacking a methylene chain (Zero methine) (→ Methine Dyes and Pigments) fluoresce and can be tuned from 500 to 550 nm when stimulated with a ruby laser or a flash lamp [71]. An example is the dye **19** [2,6-bis(4-ethoxyphenyl)-4-pentyloxyphenylpyrylium perchlorate] [*47776-83-0*].

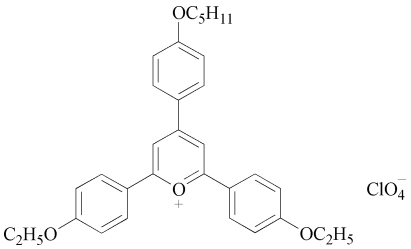

19

Hemicyanine (styryl) dyes can be prepared by reaction of 4-dialkylaminobenzaldehydes or the vinylogous cinnamaldehydes with quaternary N heterocycles that have a methyl or methylene group capable of undergoing reaction. For example, 4-dimethylaminocinnamaldehyde gives Pyridine 1 with quaternary 2-methylpyridine and Pyridine 2 with quaternary 4-methylpyridine, respectively (see Table 7) [72].

In the same way, Styryl 6 is obtained from 1,3,3-trimethyl-2-methyleneindoline, Styryl 8 from quaternary 2-methylbenzothiazole, and Styryl 11 (all Lambda Physik) from quaternary lepidin, respectively (see Table 7) [72].

Neutrocyanine dyes with aromatic [73] (**20**) or heterocyclic (**21**) [74] terminal groups are also suitable for producing laser radiation.

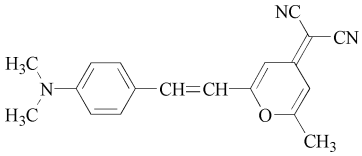

(**20**)

DCM
XeCl-Exc.: 632 – 690 nm
[*51325-91-8*] (Kodak)
4-Dicyanomethylene-2-methyl-6-(*p*-dimethylaminosyryl)-4*H*-pyran

Table 7. Tuning ranges of some styryl laser dyes

Dye	CAS registry number	Structure	Tuning range, cw argon laser, nm
Pyridine 1	[*87004-02-2*]	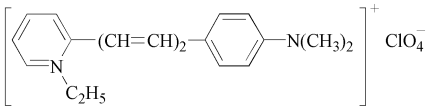	670 – 780 670 – 760 [a]
Pyridine 2	[*89846-21-9*]	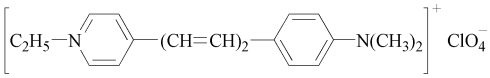	685 – 820 695 – 790 [a]
Styryl 6	[*29238-76-4*]	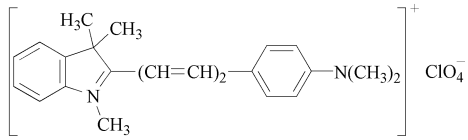	670 – 760 [b]
Styryl 8	[*87004-17-9*]	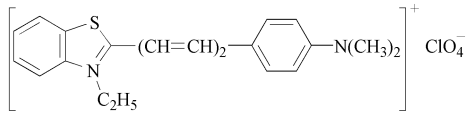	700 – 840
Styryl 11	[*82988-09-8*]	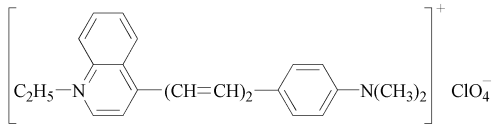	770 – 845

[a] XeCI excimer. [b] Nd-YAG 2nd.

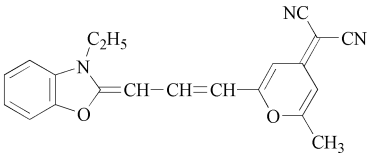

21

Peak 640 nm
[*50539-54-3*] (Kodak)
4-Dicyanomethylene-2-methyl-6-(3-ethyl-2-benzoxazolinylidenpropylene)-4*H*-pyran

Infrared Laser Dyes. According to the electron gas model [75] an increased number of conjugated double bounds should lead to a bathochromic shift of cyanine dyes. This concept was used for the synthesis of IR laser dyes by DREXHAGE et al. [82], [106], [107], [108], [109]. Laser action of dye molecules in the IR region is difficult to achieve. The compounds with a $S_0 \rightarrow S_1$ transition in the IR region show very small fluorescence quantum yields ($< 10^{-4}$) and generally a poor photostability. In [110] a review of IR laser dyes is given. The dye S 501 in *o*-dichlorobenzene, excited by picosecond pulses reaches the longest dye laser wavelength (tuning range between 1700 and 2000 nm). Remarkably, this dye solution could be used for many hours without a decrease in laser efficiency.

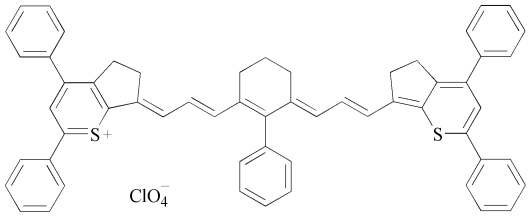

22

2.7. Oxazine Dyes

Replacement of the −CH= group in xanthene dyes with nitrogen leads to the oxazine dyes. According to the electron gas model of methine dyes (H. KUHN) [75] replacement of the central methine group by a group with greater electron affinity should produce a marked bathochromic shift. This is in fact the case. The tuning range of the oxazine dyes is 640 – 900 nm (for preparation of these dyes see → Azine Dyes). Two

Table 8. Tuning ranges of some oxazine laser dyes

Dye	CAS registry number	Structure	Tuning range, nm	
			XeCl Excimer	cw Argon
Oxazine 1	[*24796-94-9*]	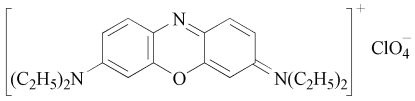	692 – 768	695 – 800 [a]
Oxazine 4	[*60786-96-1*]	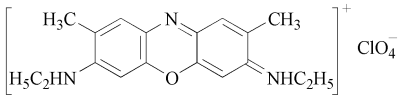	665 – 718	
Cresyl Violet 670	[*41830-80-2*]	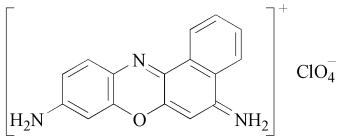	643 – 688	650 – 695
Nile Blue A	[*53340-16-2*]	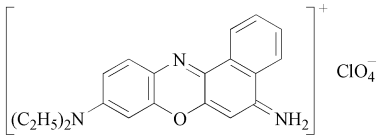	688 – 747	690 – 780
Oxazine 750	[*85256-40-2*]	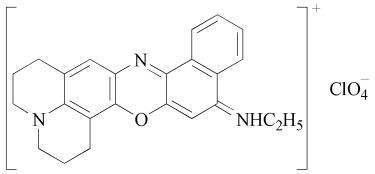		790 – 900 [a]
Phenoxazone 9	[*7385-67-3*]	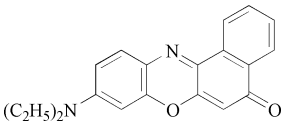		560 – 700 [b]

[a] With cw krypton laser. [b] With nitrogen laser (in ethanol).

efficient diaminophenoxazine dyes are Oxazine 1 and Oxazine 4 (Kodak, Lambda Physik) (see Table 8) [76].

The fluorescent quantum yield of Oxazine 4 increases from 63 to 89 % in *O*-deuterated alcohols. The factor responsible for this effect is apparently the reduced radiationless deactivation by N–H vibration because of a partial H – D exchange. Oxazine 1 possesses no NH group and does not show this effect. In the latter case, however, the fluorescent quantum yield increases with decreasing temperature, due to hindered rotation of the amino group [77].

Cresyl Violet 670 and Nile Blue A (Kodak, Lambda Physik) are examples of diaminobenzophenoxazine dyes (Table 8) [76]. An increase in laser performance is observed by addition of Rhodamine 6 G to a solution of Cresyl Violet 670. Energy transfer occurs because the fluorescence band of Rhodamine 6 G overlaps the absorption band of Cresyl Violet [78]. The laser beam with longest wavelength is obtained with the ring-closed dye Oxazine 750 (see Table 8) [79].

Nile Blue A phenoxazone (Phenoxazon 9) displays pronounced solvatochromism. On stimulation with a nitrogen laser, the tuning range in ether is 580 – 631 nm and in hexafluoroisopropanol 690 – 743 nm. By mixing solvents, the wavelength range stretches from 580 to 743 nm without interruption [80].

3. References

General References

1. F. P. Schäfer: "Principles of Dye Laser Operation," in F. P. Schäfer (ed.): *Dye Lasers* 3rd ed., Springer, Berlin1990.
2. K. H. Drexhage in F. P. Schäfer (ed.): *Topics in Applied Physics, "Structure and Properties of Laser Dyes,"* **vol. 1,** Dye Lasers, Springer Verlag, Berlin 1973.
3. M. Maeda: *Laser Dyes,* Academic Press, Tokyo 1984.
4. U. Brackmann, *Lambdachrome® Laser Dyes,* Lambda-Physik, Göttingen 1986.
5. U. Brackmann, *Lambdachrome® Laser Dyes,* 2nd ed., Lambda-Physik, Göttingen 1997.
6. Exciton Chemical Company, *Laser Dyes*, Dayton, Ohio (1983).
7. G. Jones II in F. J. Duarte, L. W. Hillmann (eds.): *Dye Laser Principles* , Academic Press, Boston 1990.
8. C. Jensen in F. J. Duarte (ed.): *High-Power Dye Lasers* , Springer, Berlin 1991.
9. M. Stucke (ed.): "Dye Lasers: 25 Years," *Topics Appl. Phys.* **70**(1979).
10. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996.

Specific References

11. P. P. Sorokin, J. R. Lankard, *IBM J. Res. Dev.* **10** (1966) 162 – 163.
12. F. P. Schäfer, W. Schmidt, J. Volze, *Appl. Phys. Lett.* **9** (1966) 306 – 309; F. P. Schäfer, *Max-Planck-Ges. Jahrb.* 1984, 46 – 67.
13. P. P. Sorokin, J. R. Lankard, *IBM J. Res. Dev.* **11** (1967) 148; W. Schmidt, F. P. Schäfer, *Z. Naturforsch. A* **22 A** (1957) 1563 – 1566.
14. B. H. Soffer, B. B. McFarland, *Appl. Phys. Lett.* **10** (1967) 266 – 267.
15. B. B. Snavely, F. P. Schäfer, *Phys. Lett. A* **28 A** (1969) 728 – 729.
16. O. G. Peterson, S. A. Tuccio, B. B. Snavely, *Appl. Phys. Lett.* **17** (1970) 245 – 247.
17. A. Dienes, E. P. Ippen, C. V. Shank, *IEEE J. Quantum Electron.* **QE-8** (1972) 388.
18. K. H. Drexhage, *Laser Focus* **9** (1973) no. 3, 35 – 39.
19. S. L. Shapiro: *Ultrashort Light Pulses,* Springer Verlag, Berlin 1977; F. P. Schäfer, *Max-Planck-Ges. Jahrb.* 1984, 46 – 67.
20. C. K. Rhodes: *Excimer Lasers,* Springer Verlag, Berlin 1978.
21. General Telephone and Electronics Lab., US 3 677 959, 1970 (C. Brecher, R. Pappalardo, H. Samelson).
22. General Telephone and Electronics Lab., US 3 681 252, 1970 (C. Brecher, A. Heller, R. Pappalardo).
23. Bodo Liphardt, Bernd Liphardt, W. Lüttke, *Opt. Commun.* **38** (1981) no. 3, 207 – 210. *Chem. Ber.* **115** (1982) 2997 – 3010.
24. Eastman Kodak, US 3 818 371, 1971, DE-OS 2 113 127, 1971 (A. H. Herz, W. C. McGolgin, J. S. Hayward, O. G. Peterson).
25. Eastman Kodak, DE-OS 2 226 334, 1971 (K. H. Drexhage).
26. W. Hartig, H. Walther, *Appl. Phys.* **1** (1973) 171 – 174.
27. M. R. Bowman, A. J. Gibson, M. C. W. Sandford, *Nature (London)* **221** (1969) 456 – 457.
28. K. W. Rothe, U. Brinkmann, H. Walther, *Appl. Phys.* **3** (1974) 115 – 119.
29. Max-Planck-Ges. z. Förd. d. Wiss., DE-OS 2 722 310, 1977 (F. P. Schäfer, M. Stuke).
30. U. Boesl, H. J. Neusser, E. W. Schlag, *Chem. Phys. Lett.* **61** (1979) no. 1, 57 – 61, 62 – 66.
31. W. Zapka, U. Brackmann, *Appl. Phys.* **20** (1979) 283 – 286; DE-OS 3 007 234, 1980 (W. Zapka et al.).
32. M. Rinke, H. Güsten, H. J. Ache, *J. Phys. Chem.* **90** (1986) 2661 – 2665.
33. W. Hüffer et al., *Opt. Commun.* **33** (1980) no. 1, 85 – 88. Bayer, DE-OS 2 938 132, 1979 (R. Raue et al.).
34. Lambda-Physik, DE-OS 3 225 741, 1982 (H. Gerhard, R. Vehrenkamp).
35. Hoechst, DE-OS 2 238 628, 1972 (W. Sahm, A. Horn); Bayer, DE-OS 2 843 850, 1978 (H. Telle, R. Schieder, R. Raue, U. Eckstein).
36. F. P. Schäfer, Z. Bor, W. Lüttke, B. Liphardt, *Chem. Phys. Lett.* **56** (1978) no. 3, 455 – 457; B. Liphardt, W. Lüttke, *Liebigs Ann. Chem.* 1981, 1118 – 1138; W. Lüttke, F. P. Schäfer, *Laser Optoelektronik* 1983, no. 2, 127 – 136.
37. W. Hüffer et al., *Opt. Commun.* **28** (1979) no. 3, 353 – 354.
38. V. S. Antonov, K. L. Hohla, *Appl. Phys. B* **B 30** (1983) 109 – 116; *Appl. Phys. B* **B 32** (1983) 9 – 14.
39. G. Gauglitz, R. Goes, W. Stooß, R. Raue, *Z. Naturforsch. A* **40 A** (1985) 317 – 323.
40. Bayer, DE-OS 2 841 519, 1978; US 4 301 091, 1979 (R. Schieder, H. Telle, R. Raue, W. Brinkwerth).
41. W. Majewski, J. Krasinski, *Opt. Commun.* **18** (1976) no. 3, 255 – 259.

42. Bayer, DE-OS 2 700 293, 1977 (U. Brinkmann, H. Telle, R. Raue); H. Telle et al., *Opt. Commun.* **24** (1978) no. 1, 33 – 37.
43. J. Kuhl, H. Telle, R. Schieder, U. Brinkmann, *Opt. Commun.* **24** (1978) no. 3, 251 – 254.
44. Bayer, DE-OS 2 700 292, 1977; US 4 274 062, 1978 (U. Brinkmann, H. Telle, R. Raue, C. W. Schellhammer).
45. H. Telle, U. Brinkmann, R. Raue, *Opt. Commun.* **24** (1978) no. 3, 248 – 250; R. K. Bauer, A. Balter, A. Kowalczyk, Ch. Jung, *Z. Naturforsch. A* **35 A** (1980) 1319 – 1324; M. Rinke, H. Güsten, *Ber. Bunsenges. Phys. Chem.* **90** (1986) 439 – 444.
46. Eastman Kodak, DE-OS 1 764 982, US 3 521 187, GB 1 183 999, 1967 (B. B. Snavely, C. G. Peterson, R. F. Reithel).
47. C. V. Shank, A. Dienes et al., *Appl. Phys. Lett.* **16** (1970) 405 – 407.
48. Bayer, DE-OS 2 702 237, 1977 (R. Raue, H. Harnisch, C. W. Schellhammer).
49. H. v. Pechmann, M. Schach, *Ber. Dtsch. Chem. Ges.* **32** (1899) 3690 – 3696.
50. E. J. Schimitschek, J. A. Trias, M. Tayer, J. E. Celto, *IEEE J. Quantum Electron.* **QE-9** (1973) 781 – 782; E. J. Schimitschek et al., *Opt. Commun.* **16** (1976) no. 3, 313 – 316.
51. K. H. Drexhage, G. R. Erikson, G. H. Hawks, G. A. Reynolds, *Opt. Commun.* **15** (1975) no. 3, 399 – 403.
52. G. A. Reynolds, K. H. Drexhage, *Opt. Commun.* **13** (1975) no. 3, 222 – 225; Eastman Kodak US 3 873 940, 1971 (K. H. Drexhage); DE-OS 2 238 050, 1972 (K. H. Drexhage, G. A. Reynolds).
53. Geigy, DE 1 098 125, 1958 (H. Häusermann, J. Voltz).
54. S. A. Tuccio, K. H. Drexhage, G. A. Reynolds, *Opt. Commun.* **7** (1973) no. 3, 248 – 252.
55. Bayer, DE-OS 2 411 969, 1974 (F. P. Schäfer, H. Harnisch, R. Raue).
56. P. Moeckli, *Dyes Pigm.* **1** (1980) 3 – 15.
57. Bayer, DE-OS 3 322 946, 1983 (R. Raue, H. Harnisch, K. H. Drexhage); R. Raue, H. Harnisch, K. H. Drexhage, *Heterocycles* **21** (1984) 167 – 190.
58. US-Department of Energy, US 4 103 256, 1976 (P. R. Hammond, R. L. Atkins, R. A. Henry, A. N. Fletcher).
59. A. N. Fletcher, *Appl. Phys.* **14** (1977) 295 – 302.
60. US Department of Energy, US 4 103 257, 1976 (P. R. Hammond, R. L. Atkins, R. A. Henry, A. N. Fletcher).
61. P. P. Sorokin, J. R. Lankard, E. C. Hammond, V. L. Moruzzi, *IBM J. Res. Developm.* **11** (1967) 130. W. Schmidt, F. P. Schäfer, *Z. Naturforsch. A* **22 A** (1967) 1563. Eastman Kodak, GB 1 255 399, 1969 (O. G. Peterson, B. B. Snavely).
62. K. H. Drexhage in F. P. Schäfer, (ed.): *Structure and Properties of Laser Dyes in Dye Lasers,* Springer Verlag, Berlin 1973, p. 148.
63. Bayer, DE-OS 2 529 903, 1975 (H. Walther, W. Hartig, A. Friedrich, R. Raue).
64. Bayer, DE-OS 2 421 607, 1974 (H. Walther, W. Hartig, A. Friedrich, R. Raue).
65. Bayer, DE-OS 3 322 945, 1983 (K. H. Drexhage, H. Harnisch, R. Raue); K. H. Drexhage, H. Harnisch, R. Raue, *Heterocycles* **21** (1984) 167 – 190.
66. Y. Miyazoe, M. Maeda, *Opto-electronics (London)* **2** (1970) no. 4, 227 – 233; O. Hildebrand, E. Göbel, DE-OS 2 431 966, 1974.
67. K. M. Romanek, O. Hildebrand, E. Göbel, DE-OS 2 648 398, 1976.
68. J. Zemlicka, Z. Arnold, *Collect. Czech. Chem. Commun.* **26** (1961) 2852.
69. G. A. Reynolds, K. H. Drexhage, *J. Org. Chem.* **42** (1977) 885 – 888; B. Kopainsky, W. Kaiser, K. H. Drexhage, *Opt. Commun.* **32** (1980) no. 3, 451 – 455; B. Kopainsky et al., *Appl. Phys. B* **B 29** (1982) 15 – 18.
70. W. Kranitzky et al. *Opt. Commun.* **36** (1981) no. 2, 149 – 152.
71. Eastman Kodak, US 3 753 146, DE-OS 2 109 040, 1970 (G. A. Reynolds, S. A. Tuccio, O. G. Peterson, D. P. Specht).
72. Eastman Kodak, DE-OS 2 260 210, 1971 (F. G. Webster, W. C. McColgin); K. Kato, *IEEE J. Quantum Electron.* **16** (1980) 1017 – 1018; J. Hoffnagle, L. Ph. Roesch, N. Schlumpf, A. Weis, *Opt. Commun.* **42** (1982) no. 4, 267 – 268.
73. Eastman Kodak, FR 2 156 723, 1971 (F. G. Webster, W. C. McColgin); V. S. Antonov, K. L. Hohla, *Appl. Phys. B* **B 32** (1983) 9 – 14.
74. Eastman Kodak, DE-OS 2 255 195, 1971 (W. C. McColgin, F. G. Webster).
75. H. Kuhn, *Chimia* **4** (1950) 203; *Z. electrochem. angew. physikal. Chemie* **53** (1949) 165. *Experientia* **9** (1953) 41.
76. Eastman Kodak, US 3 731 222, DE-OS 2 316 040, 1972 (K. H. Drexhage).
77. R. Sens, K. H. Drexhage, *J. Lumin.* **24/25** (1981) 709 – 712.
78. W. Schmidt, W. Appt, N. Wittekindt, *Z. Naturforsch. A* **27 A** (1972) 37 – 41.
79. G. D. Aumiller, *Opt. Commun.* **41** (1982) no. 2, 115.

80. D. Basting, DE-OS 2 534 713, 1975; D. Basting, D. Ouw, F. P. Schäfer, *Opt. Commun.* **18** (1976) no. 3, 260 – 262.
81. F. P. Schäfer: "Principles of Dye Laser Operation," in F. P. Schäfer (ed.): *Dye Lasers* 3rd ed., Springer, Berlin1990, p. 7.
82. H. J. Polland, T. Elsaesser, A. Seilmeier, W. Kaiser, M. Kussler, N. J. Marx, B. Sens, K. H. Drexhage, *Appl. Phys. B* **32** (1983) 53
83. E. Thiel, C. Zander, K. H. Drexhage, *Opt. Commun.* **60** (1986) 396.
84. E. Thiel, C. Zander, K. H. Drexhage, *Opt. Commun.* **62** (1987) 171.
85. S. P. McGlynn, T. Azumi, M. Kinoshita: *Molecular Spectroscopy of the Triplet State* , Prentice-Hall International, London, 1969, p. 2.
86. E. Thiel, K. H. Drexhage, *Chem. Phys. Letters* **199** (1992) 329.
87. R. Menzel, R. Bornemann, E. Thiel, *Phys. Chem. Chem. Phys.* **1** (1999) 2435, and references therein.
88. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996, p. 144.
89. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996, pp. 108, 135.
90. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996, p. 64, and references therein.
91. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996, p. 244.
92. G. Jones II in F. J. Duarte, L. W. Hillmann (eds.): *Dye Laser Principles* , Academic Press, Boston 1990, p. 287.
93. R. Reisfeld, E. Yariv, H. Minti, *Opt. Mater.* **8** (1997) 31, and references therein.
94. G. Marowsky, R. Cordray, F. K. Tittel, W. L. Wilson, C. B. Collins, *Appl. Phys. Lett.* **33** (1978) 59.
95. D. L. Andrews (ed.): *Applied Laser Spectroscopy* , VCH, Weinheim, 1992.
96. J. Mans, L. Wöste (eds.): *Femtochemie,* VCH Verlagsgesellschaft, Weinheim, 1995.
97. K. H. Drexhage: "Principles of Dye Laser Operation," in F. P. Schäfer (ed.): *Dye Lasers* 3rd ed., Springer, Berlin1990, p. 187.
98. G. Jones II in F. J. Duarte, L. W. Hillmann (eds.): *Dye Laser Principles,* Academic Press, Boston 1990, p. 433, and references therein.
99. Exciton Chemical Company, *Laser Dyes*, Dayton, Ohio (1983).
100. U. Brackmann, *Lambdachrome® Laser Dyes,* 2nd ed., Lambda-Physik, Göttingen 1997.
101. E. Thiel: *Eigenschaften angeregter Rhodamin-Farbstoffe und deren Wirkung im Farbstofflaser* , Shaker Verlag, Aachen1996, p. 161.
102. D. Peros, Dissertation, University Siegen (1991) p. 56.
103. Roche diagnostics GmbH, EP 96 24 97 A, 1999 (K. H. Drexhage et al.).
104. T. G. Pavlopoulos, J. H. Boyer, SPIE Vol. 1132, High Power Lasers and Laser Machining Technology, 1989.
105. Bo-Dekk Ventures, Ltd., US 5 189 029, 1993 (J. H. Boyer, L. R. Morgan).
106. B. Kopainsky, W. Kaiser, K. H. Drexhage, *Opt. Commun.* **32** (1980) 451.
107. W. Kranitzky, B. Kopainsky, W. Kaiser, K. H. Drexhage, G. A. Reynolds, *Opt. Commun.* **36** (1981) 149.
108. B. Kopainsky, P. Qui, W. Kaiser, B. Sens, K. H. Drexhage, *Appl. Phys. B* **29** (1982) 15.
109. A. Seilmeier, W. Kaiser, B. Sens, K. H. Drexhage, *Opt. Lett.* **8** (1983) 205.
110. M. Stucke (ed.): "Dye Lasers: 25 Years," *Topics Appl. Phys.* **70**(1979) 95.

Lasers

Holography and Laser Dyes are separate keywords.

JOHN H. KELLY, University of Rochester, Laboratory for Laser Energetics, Rochester, New York, United States

1. Introduction 1
2. Physical Principles 2
2.1. Interaction Between Electromagnetic Radiation and Matter; Amplification of Light 2
2.2. Properties of Light and Laser Light; Diffraction, Interference, and Coherence 3
2.3. Optical Resonators 4
2.3.1. Stable Resonators 4
2.3.2. Unstable Resonators 7
2.4. Short-Pulse Generation, Longitudinal Mode Locking 7
2.5. Q-Switching and Cavity (Resonator) Dumping 8
2.6. Nonlinear Optics 9
3. Types of Laser and Their Properties 10
3.1. Gas Lasers 10
3.2. Dye Lasers 11
3.3. Solid-State Lasers 11
3.4. Semiconductor Lasers 12
4. Uses 13
4.1. Spectroscopy 13
4.2. Remote Sensing 14
4.3. Laser-Assisted Chemical Processes 14
4.4. Metrology 15
4.5. Medical Applications 15
5. Safety 16
6. References 18

1. Introduction

A laser is a device that generates coherent electromagnetic radiation in the infrared, visible, or ultraviolet region of the spectrum. The electromagnetic radiation generated by a laser differs from that of conventional sources in several respects:

1) Laser radiation can be extremely monochromatic; i.e., it can be described as having nearly a single frequency.
2) Laser radiation can be extremely directional.
3) Laser radiation can be extremely coherent both spatially and temporally; i.e., interference effects may be observed between portions of the output both spatially and temporally separated.
4) Laser radiation exhibits the statistical properties of an amplitude-stabilized oscillator rather than those of narrow-band Gaussian noise.

A typical laser may exhibit all of these properties to varying degrees. Laser radiation, regardless of frequency, is generally referred to as laser light.

Lasers are light amplifiers that operate using the principle of *l*ight *a*mplification by *s*timulated *e*mission of *r*adiation (laser). The addition of sufficient correctly phased feedback around an optical or electronic amplifier will convert it into an oscillator. Although the vast majority of lasers are oscillators, laser amplifiers and oscillator-driven amplifier chain systems are not uncommon.

Stimulated emission of radiation may be described as follows. Atoms, ions, or molecules that have absorbed energy (been excited) can release their energy either radiatively or nonradiatively. Radiative decay may occur by the usually observed spontaneous emission or, alternatively, by stimulated emission. In spontaneous emission, light is emitted with random phase and direction and a frequency characterized by a line shape function. If light is incident on the excited species, they may be stimulated to emit their energy as additional light with the same frequency, phase, polarization, and direction as the incident light. The incident light appears amplified. This process is the converse of stimulated absorption in which light is attenuated. Whether stimulated emission or stimulated absorption predominates is determined by the ratio of the excited to the unexcited species. When the excited species out-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_165

number the unexcited species, an inversion is said to exist and net amplification occurs.

Feedback in electromagnetic systems in which the physical size of the components is large compared to a wavelength is provided by resonant structures. These are usually some form of reflective box or cavity. Due to the very short wavelength of coherent light, which renders it highly directional, only two opposing sides of the box are necessary. These reflective sides of the box (mirrors) provide the feedback which allows the amplifier to oscillate.

Finally, all that is needed to complete a useful device is a means of coupling out some of the light circulating within the oscillator. The most common method is to make one mirror partially reflecting. The portion of the light not reflected is transmitted out of the cavity. Other techniques include shaping the mirrors so that the non-perfect directionality of the light allows part of it to bypass the mirror. For lasers that have a pulsed rather than a continuous-wave (cw) output, a switch can be inserted into the cavity which allows the light to bypass the mirrors when it is actuated.

The laser can trace its historical origins to masers, devices that operate on the same principle in the microwave portion of the electromagnetic spectrum. The first maser was operated by C. H. Townes, J. P. Gordon, and H. Zeiger at Columbia University. A device was built in the Soviet Union shortly thereafter by N. G. Basov and A. M. Prokhorov. These groups shared the 1964 Nobel Prize in Physics for their work. In 1958, A. L. Schawlow and C. H. Townes published a landmark paper which extended the theory of masers to the visible spectrum. Although not published at the same time, similar work was done by the then Columbia University graduate student Gordon Gould. This work was the basis for awarding the fundamental laser patent, after much litigation, to Gould's company, Patlex. Theodore H. Maiman demonstrated the first successful laser, a flashlamp-pumped ruby device operating at 694 nm, in 1960 at the Hughes Research Laboratories. In the same year, Ali Javan at the Bell Telephone Laboratories demonstrated the first gas laser in an electrically excited mixture of helium and neon at 1.15 μm. In 1961 the ubiquitous 633-nm helium – neon laser was demonstrated. Subsequent to these early developments, a plethora of published accounts of laser action in a wide variety of materials appeared and the field of laser research blossomed.

Today laser performance spans an enormous range. Powers range from microwatts to tens of kilowatts (cw), and pulsed terawatts (10^{12} W). Lasers operate from the microwave portion of the spectrum to the soft X-ray region. Laser pulse widths range from a few femtoseconds (10^{-15} s) to continuous wave. Bandwidths as low as 1 kHz have been measured. Lasers range from miniature semiconductor devices to the enormous machines required for inertial-confinement fusion.

2. Physical Principles

The physical principles of laser operation are covered in detail in [1]. These principles are summarized in this chapter.

2.1. Interaction Between Electromagnetic Radiation and Matter; Amplification of Light

Consider the atomic system shown in Figure 1A. Atomic levels 1 and 2 are separated by an energy corresponding to $h\nu$, where h is Planck's constant and ν is the resonant frequency. Level 1 is assumed to be the ground state of the atom. The other levels are minimally populated thermally. The decay rate of level 3 to level 2 is assumed to be faster than that of level 2 to level 1. A photon of frequency ν can be absorbed by the system, inducing a transition from level 1 to level 2. Level 2 may then spontaneously emit a photon of frequency ν. If the atom has been excited to level 2, an induced transition of the atom from level 2 to level 1 may occur. In this case, the transition energy ($h\nu$) is added to the stimulating light with the same frequency, direction, and polarization.

Under normal circumstances, levels 1, 2, and 3 are thermally populated according to the Boltzmann distribution:

$$n_2/n_1 = \exp(-h\nu/kT)$$

where k is the Boltzmann constant, T is the temperature, and n_1, n_2 represent the number of

atoms in levels 1 and 2, respectively. A similar relation holds for level 3. The probability of a transition, either spontaneous or stimulated, is proportional to the product of the number of atoms in the initial state, the density of photons, and the radiative transition rate. This is essentially what is termed Fermi's golden rule. In most commonly encountered systems, the number of atoms in level 1 far exceeds the numbers in other levels; thus, absorption occurs. If the number of atoms in level 2 can be made to exceed the number in level 1, the stimulated emission rate will exceed the absorption rate and net amplification will occur. The system is then said to be inverted. Levels 2 and 1 are termed the upper and lower lasing levels, respectively.

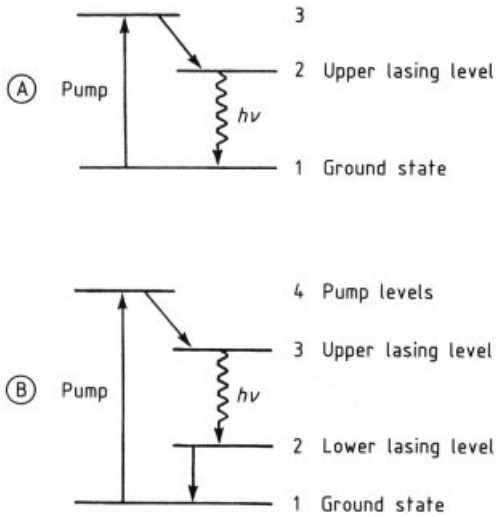

Figure 1. Three and four-level atomic systems
A) Three-level atomic system; B) Four-level atomic system

The number of atoms in level 2 can be made to exceed those in level 1 by various methods. Collisional excitation by electrons and optical pumping by flash-lamp light are two examples. In optical pumping, photons from an intense conventional source such as a flash lamp or an arc lamp are absorbed by the atomic medium and excite atoms from level 1 to level 3. Level 3 then decays, radiatively or nonradiatively, to level 2. When the number of atoms in level 2 exceeds that in level 1, gain is observed. Note that this is an inherently inefficient system because over half of the atoms must be removed from level 1 to achieve unity gain. A system of the type shown in Figure 1A is termed a *three-level system.*

A more efficient laser system is shown in Figure 1B. This is a *four-level system.* Level 1 is the ground state where the unexcited atom lies. Lasing action occurs between levels 3 and 2. Level 2 decays rapidly to the ground state. Level 3 is the upper lasing level and typically has a longer lifetime than other excited levels of the system. Finally, level 4 represents the pump bands from which atoms decay to level 3. Note that this system is very efficient because only enough atoms must be raised to level 3 to exceed the (usually small) thermal population of level 2.

Saturation of gain may occur in either three- or four-level systems when the stimulated emission rate exceeds the pumping rate. In the case of cw lasers, saturation intensity may be defined as the intensity required to reduce the inversion to one-half of its unsaturated value. For pulsed systems, in which pumping is assumed to be negligible for the duration of the laser pulse, it is more convenient to define a saturation fluence as the energy per unit area required to reduce the initial inversion to 1/e of its original value.

2.2. Properties of Light and Laser Light; Diffraction, Interference, and Coherence

Laser light offers the opportunity to easily observe effects due to diffraction and coherence that are not readily apparent with conventional light sources. Conventionally, light is thought of within the framework of geometrical optics; that is, light is considered to travel in rays. This model is valid only in the limit of infinitely short wavelength. Since the wavelength of green light is ca. 500 nm, this is usually a very good approximation. Diffraction is a direct manifestation of the nonzero wavelength of light. The classic example of diffraction of light is the pattern produced at a long distance (compared to the dimensions of the aperture) from a circular aperture illuminated by a monochromatic plane wave. The Airy pattern so produced is shown in Figure 2. The central maximum is surrounded by a series of rings of decreasing intensity, despite the fact that the pattern observed immediately after the aperture is a direct map of the aperture. A less obvious effect is the modification of the phase

of the light. In Figure 2 the phase within each bright ring is constant, but a phase change of π occurs between each of the rings.

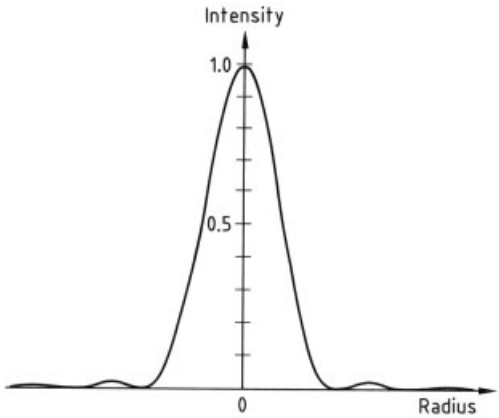

Figure 2. The Airy pattern
The plot shows the observed intensity along a diameter of the azimuthally symmetric pattern. The central spot contains 84 % of the energy transmitted by the aperture, and the maximum intensity of the first ring is 1.8 % of the central maximum.

A similar effect is observed for beams with other transverse intensity profiles. An important class of transverse intensity profiles are the Gaussian profiles, whose intensity is characterized by

$$I(r) = I_0 \exp(-r/w)^2$$

where I_0 is the on-axis intensity, r is the distance from the axis, and w characterizes the size of the beam. These are the so-called Gaussian beams produced by a large number of lasers. The propagation of a Gaussian beam with an initially flat phase profile is characterized by spreading of the width of the Gaussian profile and parabolic curving of the phase front of the beam in the direction of propagation. This is important in the discussion of resonators.

A second phenomenon that is readily observed with laser sources is interference. If a screen having two adjacent pinholes is illuminated by a laser source of sufficient coherence, alternating bands, or fringes, of light and dark are observed. The intensity in the dark regions may reach zero, whereas that in the bright regions may attain a maximum of four times the intensity observed when one pinhole is blocked. This phenomenon is a consequence of the wave nature of light, the fact that detectors (the human eye included) respond to power rather than to electric field, and the fact that detectors have a response time far longer than the typical frequency of light (10^{15} Hz). The bright regions are areas in which the phases of the light from the two pinholes are the same and the equal electric fields add. Since the intensity is proportional to the square of the electric field, the intensity is four times that of one pinhole. Dark regions occur where the phases of the light oppose and cancel, leaving a net intensity of zero. Observation of this effect is possible because the light waves from the two pinholes have a fixed phase relationship to each other. If the phase difference between the two beams were to fluctuate in a time period that is short compared to the response time of the detector, uniform intensity with no fringes would be observed. The fixed phase relationship between the light from the two pinholes is termed lateral coherence.

A similar effect can be obtained by delaying part of the laser output with respect to itself. Here, fringes of light and dark occur as a function of the delay. The phase relationship necessary to observe this effect is a fixed phase difference between an earlier and a later portion of the beam. This is termed temporal or longitudinal coherence.

Most conventional sources are essentially incoherent, making interference effects impossible to observe. Most laser sources, on the other hand, exhibit a high degree of both lateral and longitudinal coherence.

2.3. Optical Resonators

Optical resonators are the structures that provide the necessary feedback for laser oscillation. They may be divided into two categories; stable and unstable. Stable resonators are by far the more common.

2.3.1. Stable Resonators

The simplest resonator consists of two opposing mirrors. A resonator is termed stable, if a transverse intensity and phase profile can exist within the resonator that, after reflection from both mirrors and return to original location (i.e., after one *round trip*), remains unchanged. The

transverse intensity and phase of the beam in the resonator are termed its transverse mode structure. Whether or not a particular resonator is stable is determined by the mirror's size, curvature, and separation.

Transverse Mode Structure. Electromagnetic fields in free space are described, in general, by the scalar wave equation. If the approximation is made that the field or light is essentially all propagating within ca. 0.5 rad of the z axis (a very good approximation for most lasers), the wave equation may be simplified to the paraxial wave equation:

$$\nabla_T^2 u(x,y,z) - 2jk\frac{\partial u(x,y,z)}{\partial z} = 0$$

where u is the electric field, $j = \sqrt{-1}$, and $k = 2\pi/\lambda$

(λ is the wavelength). The solutions to this equation in rectangular coordinates are the Hermite–Gaussian modes. In cylindrical coordinates, the solutions are the Laguerre–Gaussian modes. The lowest order solution to this equation is the Gaussian beam mentioned earlier:

$$I(r,z) = \frac{I_0}{w^2(z)}\exp\left[-2r^2/w^2(z)\right]$$

where $w(z)$ is given by:

$$w(z) = w_0\left[1 + (z/w_0)\right]^{1/2}$$

and w_0 is the $1/e^2$ radius at the narrowest part of the beam, termed its waist. The phase and intensity associated with this beam are shown in Figure 3. At $z = 0$, this beam has a waist where the phase front is planar and the beamwidth is narrowest. The beam spreads in either direction from $z = 0$ and acquires a parabolic phase curvature. A stable cavity for a given Gaussian beam can be assembled by placing opposing mirrors in the beam whose curvatures match those of the beam's phase front at the z location. The mirror must be considerably larger than the size of the beam.

The more conventional problem is to determine whether a stable Gaussian beam can propagate between two mirrors of given radii of curvature R_1 and R_2, separated by a distance L. Not all combinations of curved mirrors and separations produce stable cavities. The solution may be determined with the aid of the resonator g parameters:

$$g = 1 - L/R_i i$$

where $i = 1,2$ and R_i is positive if the mirror is concave toward the interior of the cavity. The condition for stability is

$$0 < gg < 121$$

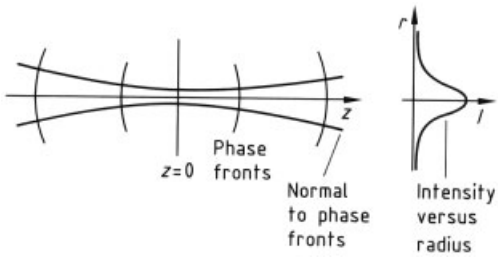

Figure 3. Gaussian beam

Several examples of stable cavities are shown in Figure 4. The *plane-parallel* or *Fabry–Perot cavity* (Fig. 4A), which has $g_1 = g_2 = 1$ and $R_1 = R_2 = \infty$, has only borderline stability. It is often used for solid-state lasers in which the lasing medium acts as a weak positive lens to enhance stability. If the curvatures of both mirrors are decreased equally until $1 > g_1 = g_2 > 0$, a *symmetric resonator* is formed (Fig. 4B). A *half-symmetric resonator* (Fig. 4C) is formed when one of the mirrors in a symmetric resonator is replaced by a plane mirror at $z = 0$ where the phase front is flat. If the radius of curvature of the two mirrors is decreased until $L = R$ ($g_1 = g_2 = 0$), a *symmetric confocal resonator* is formed (Fig. 4D). Because a tilt of either mirror shifts the resonator axis only slightly in the transverse direction, the confocal resonator is relatively insensitive to misalignment. The beam diameter per unit length of this cavity is the smallest of the stable resonators, making the confocal cavity unsuitable for high-power operation. If the radii of curvature of the two mirrors are further decreased until $L = 2R$ ($g_1 = g_2 = -1$), a *concentric resonator* (Fig. 4E) is formed. This resonator is again on the edge of stability. This cavity has a large beam size on the mirrors but a very small beam waist in the center. Care must be taken not to place easily damaged components near this waist. Finally, if one mirror in a concentric resonator is replaced by a plane mirror located at $z = 0$, a *hemispheri-*

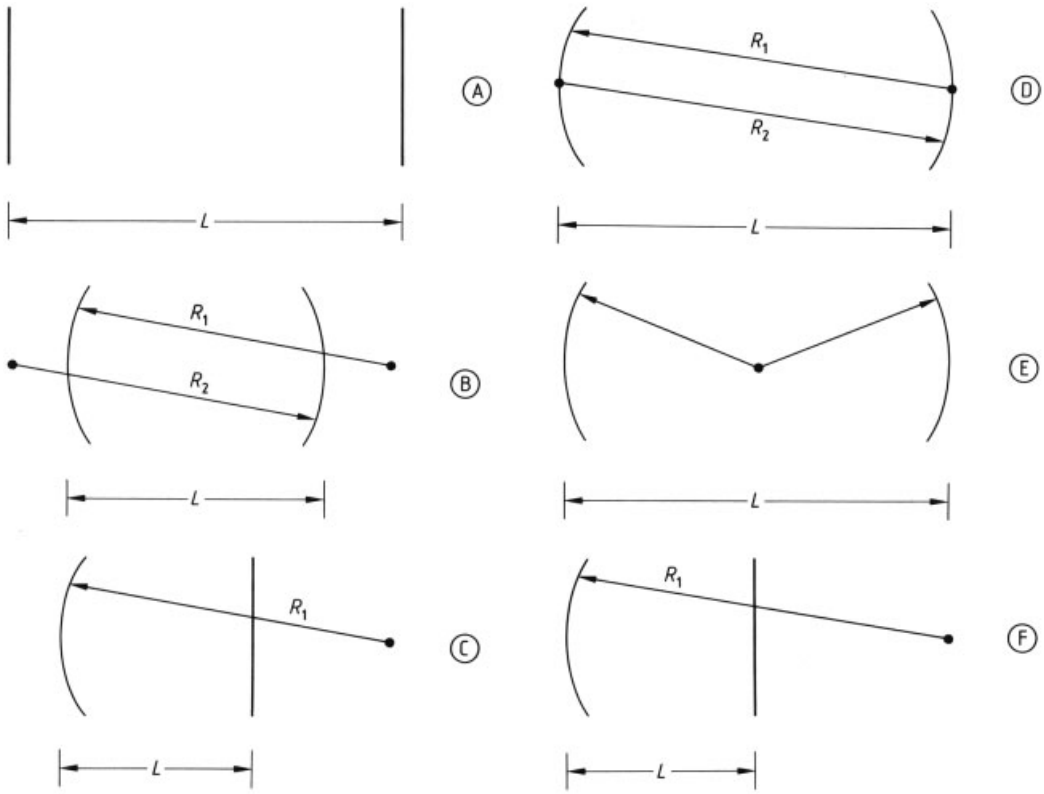

Figure 4. Stable resonators
A) The Fabry – Perot resonator has plane-parallel mirrors and is marginally stable $R_1 = R_2 = \infty$). B) The symmetric resonator has equal radii of curvature that are longer than the mirror separation ($R_1 = R_2 > L$). C) The half-symmetric resonator is a symmetric resonator with a plane mirror at the beam waist ($R_1 > L$, $R_2 = \infty$). D) The symmetric confocal resonator has radii equal to the mirror separation and is particularly stable with respect to mirror misadjustments ($R_1 = R_2 = L$). E) The concentric resonator has the sum of the radii equal to the mirror separation ($R_1 + R_2 = L$). F) The hemispherical confocal resonator is a confocal resonator with a plane mirror at the beam waist ($R_1 = 2L$, $R_2 = \infty$).

cal confocal resonator (Fig. 4F) is formed. This resonator is widely used in gas lasers.

The Gaussian beam is the lowest order solution to the paraxial wave equation. An infinite number of solutions exist for the equation, which are referred to as the *t*ransverse *e*lectromagnetic *m*odes (TEM). They are designated by $\mathrm{TEM}_{\mathrm{mn}}$ where m and n are order numbers. The three lowest order modes are shown in Figure 5. Any of these modes can propagate in any of the stable cavities described above. In fact, typical unrestricted laser output will be some combination of many of these modes. The higher order modes are characterized by having progressively larger fractions of their energy at distances further from the z axis. Advantage is often taken of this fact to limit oscillation to the lowest order (TEM_{00} or Gaussian) mode by inserting an aperture into the cavity to increase the losses for the higher order modes while leaving the lowest order mode largely unchanged.

Longitudinal Modes. The modes are restricted further to have an integral number of wavelengths in the resonator. Because the length of a typical laser resonator is of the order of a few centimeters and the wavelength of light is of the order of 5×10^{-5} cm, a large number of wavelengths exist in the resonator. Each frequency ν_q that satisfies the equation:

$$L = qc/2\nu_{\mathrm{q}}$$

where c is the speed of light and q is an integer, defines one longitudinal mode of the cavity.

The frequency separation of longitudinal modes is given by

$$\nu_{q+1} - \nu_q = c/2L$$

Since the frequency band over which the lasing medium exhibits gain (the gain bandwidth) often exceeds $c/2\,L$, many longitudinal modes will experience gain greater than unity and, therefore, oscillate in a typical laser.

Often, no fixed phase relationship exists among these longitudinal modes and an effective reduction of the temporal coherence results. If a fixed phase relationship is imposed between the longitudinal modes, the output of the laser is temporally modulated. Advantage is taken of this fact to produce temporally short pulses by the technique of mode locking.

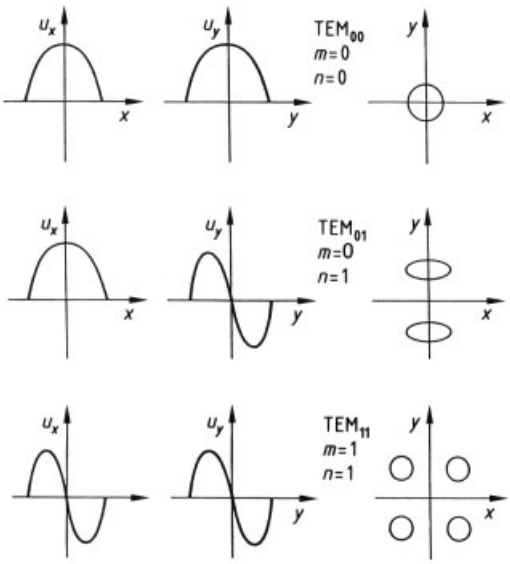

Figure 5. The three lowest order TEM modes in rectangular coordinates
The $u(x)$ and $u(y)$ functions for the electric field of each mode are also shown schematically. The intensity I is given by $I = [u(x) \cdot u(y)]$

2.3.2. Unstable Resonators

Consider the plane-parallel resonator shown in Figure 4A. If one of the mirrors is given a curvature that is convex toward the interior of the resonator, the simplest form of unstable resonator is formed. The remaining plane mirror is purposely kept small, and the output of the laser is the light leaking around that mirror. The transverse mode structure in this case is a so-called doughnut mode, an annular ring.

Unstable resonators have the advantages that their modes occupy large volumes compared to stable resonators and that they have very high transverse mode suppression, that is, they discriminate against higher order modes. They are particularly useful with high-gain and high-power laser media, such as carbon dioxide, in which large power output is desired and high gain is available to overcome diffraction losses.

2.4. Short-Pulse Generation, Longitudinal Mode Locking

Mode locking is a technique commonly applied to both pulsed and continuous lasers to produce pulses considerably shorter than the resonator round-trip time ($2\,L/c$). Typical mode-locked pulse widths range from 1 ns to < 1 ps.

For a laser with two longitudinal modes of equal oscillating amplitude, the electric field is given by

$$E(t) = \cos(\omega t) + \cos[(\omega + \Delta\omega)\,t]$$

where ω is the radian frequency.

If the output of this laser is monitored with a slow (compared to the frequency of light) detector, the observed intensity is given by

$$I(t) = 2\,[1 + \cos(\Delta\omega t)]$$

Thus, the output is amplitude-modulated at the difference in frequency between the two modes. The peak intensity is four times that of a single mode. The analysis may be extended to N longitudinal modes of equal amplitude locked together with zero phase difference. The intensity is given by

$$I(t) = \frac{\sin^2(N\Delta\omega t/2)}{\sin^2(\Delta\omega t/2)}$$

and is shown in Figure 6. The laser output is a periodic train of pulses of width approximately $\Delta\omega/(2\,\pi\,N)$, with period $2\,L/c$. The peak intensity is N times the intensity of one mode. For a set of modes with a Gaussian intensity distribution in frequency, the laser temporal output is a periodic train of Gaussian pulses.

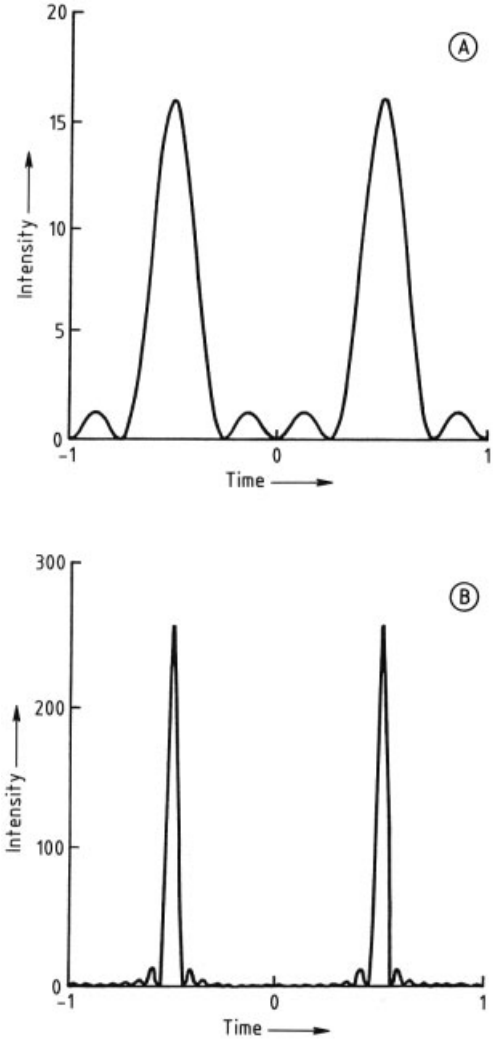

Figure 6. Intensity versus time for unit amplitude modes locked in phase.
A) Four modes locked in phase; B) Sixteen modes locked in phase

Various techniques can be used to mode-lock a laser in order to produce short pulses. The most common consists of inserting an amplitude modulator into the resonator. If the modulator is driven by a periodic signal which is synchronized with the pulse circulation time inside the cavity, the pulse will be transmitted and the intra-pulse intensity attenuated. An *acousto-optic modulator* uses standing sound waves to produce a spatially stationary, time-periodic phase grating in a transparent solid such as fused silica. Any light traversing the modulator when it has a nonzero acoustic field experiences a diffraction loss. An *electro-optic beam deflector* driven by a periodic signal deflects light out of the resonator when its driving voltage is nonzero. Another type of modulator is the *pockels cell – polarizer* combination. The pockels cell is driven by a periodic signal. When a voltage is applied to a pockels cell it alters the polarization state of the incident laser light, thereby increasing the loss at a polarizing component in the cavity.

2.5. Q-Switching and Cavity (Resonator) Dumping

In many applications, it is desirable to extract the majority of the energy stored in the lasing medium in a single giant pulse. Resonator Q-switching and resonator dumping are two techniques that allow extraction of most of the energy stored in a gain medium in a pulse of 5 – 500-ns duration. Furthermore, Q-switched lasers are often also mode-locked, thereby producing a burst of very short (0.1 – 1 ns), intense pulses.

Many laser media require a much longer time for the pump mechanism to build up a significant inversion than the time required for lasing to build up in the resonator. Consequently, as soon as the inversion has risen to the point where the gain exceeds the losses in the cavity, lasing action begins and clamps the inversion by gain saturation. The total energy stored in the laser medium is thus a function of the loss in the resonator structure. This loss includes all mechanisms for coupling energy out of the resonator, such as output coupling, diffraction losses around mirrors, absorption in internal components, and scattering. A resonator with high loss requires a proportionally higher inversion to initiate lasing. A commonly used measure of the losses in a resonator is the Q factor, which is inversely proportional to the power loss per round trip through the resonator with no gain. The lower the resonator Q, the greater is the inversion required for lasing.

Q-Switching is accomplished by switching the cavity from a low to a high Q value after a large inversion has been allowed to accumulate. The field in the resonator then builds up rapidly because the gain is high compared to

the loss. Pulse growth is limited by gain saturation, which ultimately reduces the gain in the resonator to less than the loss, causing the pulse to decay. Since pulse growth and pulse decay have different mechanisms, the Q-switched pulse shape will be temporally asymmetric with a typically faster rise than fall. The precise shape of the pulse and its asymmetry are functions of how high the gain medium is pumped above the threshold. The pulse is longest and nearly symmetric about its peak when the resonator is Q-switched only slightly above threshold. As the pumping increases, the pulse width becomes shorter and the asymmetry more pronounced.

Several techniques are used for Q-switching a resonator. One of the most common is to insert a pockels cell between crossed polarizers into the resonator. With no applied voltage on the cell, it has no effect on the polarization of the light in the resonator. The resonator effectively has a pair of crossed polarizers inside it, which prevent lasing. At the gain peak, the pockels cell is pulsed to the half-wave voltage which causes it to rotate the plane of polarization of incident light by 90°. This allows light to be transmitted. Lasing then builds up. A second technique is to insert a traveling-wave acousto-optic cell into the resonator. The traveling acoustic wave acts as phase grating which diffracts light out of the resonator. At gain peak, the acoustic drive to the cell is switched off, eliminating diffraction losses. A third technique is to insert a cell containing a saturable dye into the resonator. If the inversion is increased until lasing action begins despite absorption, the absorption saturates and the cell becomes low loss. Such Q-switching with a saturable dye requires careful optimization of dye concentration and resonator parameters. If the saturation of the absorption and its recovery are sufficiently rapid, the dye cell can aid in mode-locking. The dye acts as an amplitude modulator inside the cavity, transmitting the pulse and attenuating the lower intra-pulse intensities. Indeed, not achieving simultaneous Q-switching and mode-locking can be difficult with some dyes.

The Q-switching provides pulse widths in the range of 5 – 500 ns, whereas mode locking provides pulse widths from 1 ns to 0.1 ps. Often, pulses are desired in the range 1 – 10 ns. Cavity dumping is used in this case. Lasing is allowed to build up in a resonator with 100 % reflecting mirrors. At the time of maximum intracavity intensity, an internal switch (typically a pockels cell) is activated, causing all the propagating light in the cavity to be deflected out of the cavity by a polarizer. All of the light in the cavity is deflected out, or dumped, in one cavity round trip; thus, the pulse width is one cavity round-trip time. The pulse shape is nearly rectangular. Cavity dumping can be combined with Q-switching to provide short, high-intensity pulses.

2.6. Nonlinear Optics

The optical properties of materials are independent of the amplitude of the optical field for conventional sources. The advent of lasers allowed generation of optical fields greater than 1 kV/cm, which are necessary to observe nonlinear effects. At these higher field strengths, optical properties (e.g., refractive index) become functions of the field strength or some power of field strength. An infinite number of nonlinear optical interactions exist, corresponding to the infinite number of powers to which the electric field of a laser may be raised. Only those interactions that have become, or are likely to become, technologically important are described here. Further details are given in [2].

The most common nonlinear technique is frequency conversion. This allows generation of additional wavelengths from readily available lasers. Additional frequencies can be generated by harmonic generation, parametric oscillation, or Raman conversion. In *harmonic generation,* multiples of the fundamental laser frequency may be generated in an appropriately phase-matched material. Phase matching requires the fundamental and the harmonic to have the same velocity in the material. Most commonly, frequency doubling or tripling is performed. The intensities required for efficient frequency conversion with current materials can be very high even for lasers. Frequency conversion usually requires Q-switched and/or mode-locked lasers. To obtain sufficient intensity from continuous lasers, tight focus is usually necessary or the frequency-converting material must be placed inside the resonator to take advantage of the large fields there. In *parametric oscillation,* a strong pump beam incident on a nonlinear material is converted into a signal beam and an idler beam,

which have different frequencies. The sum of the frequencies of the signal and idler beams is equal to that of the pump. The advantage of parametric oscillation is that the frequency of the signal beam is continuously tunable. A phase-matching condition again exists among the three beams. Tuning is accomplished by changing the orientation of the nonisotropic nonlinear material. Because parametric oscillation requires a pump with a frequency higher than that of the signal beam, it is most often used to generate infrared or near-infrared light. In *Raman conversion,* the laser frequency is shifted by scattering from molecular vibrations in a material. In Stokes scattering, the laser frequency is decreased by the energy associated with the vibration; the molecule is left in an excited vibrational state. In the much rarer anti-Stokes scattering, the laser frequency is increased by the energy of a molecular vibration, and the initially vibrationally excited molecule is left in a lower state. Raman scattering may also occur from molecular rotations. The range of frequency shifts that can occur is limited by the range of bond strengths available in nature to $400-4000\,\mathrm{cm}^{-1}$. Larger shifts may be obtained by multiple Raman scattering.

Other nonlinear processes are usually encountered in a less beneficial manner. Two-photon absorption and self-focusing are two examples. In *two-photon absorption,* material absorption at twice the frequency of the laser causes significant absorption of laser light, despite the fact that the material may be transparent at the laser frequency. *Self-focusing* is a change in the refractive index of a material with increasing laser intensity. In the majority of materials, refractive index increases with intensity. This causes focusing of the beam, leading to higher intensity and eventual catastrophic damage to the material. These effects are usually seen at intensities of $4\,\mathrm{GW/cm}^2$ in meter-long paths of common optical materials (e.g., glass).

3. Types of Laser and Their Properties

Lasers are conveniently classified according to the type of gain medium used. These include gas, dye, solid state, semiconductor, and others.

3.1. Gas Lasers

Gas lasers can be either pulsed or continuous wave and include the highest power cw lasers. Typically, they are excited by an electrical discharge in the gas. An example is the ubiquitous *helium–neon* (He–Ne) laser. A longitudinal discharge is maintained in a low-pressure mixture of helium and neon. The mixture is predominantly helium with ca. 10 % neon by pressure. The discharge excites the helium, which collisionally transfers the excitation to neon. Lasing can occur at several wavelengths in neon. The highest gain and the first wavelength to lase is in the infrared at 3391.0 nm. The familiar red line is at 632.8 nm. This line is much lower in gain than the 3391-nm line; therefore, care must be taken to suppress the infrared line to achieve laser action at 632.8 nm. Other wavelenghts available from the He–Ne laser are 543.5 and 1523.0 nm.

A second important gas laser is the *carbon dioxide laser.* This molecular laser can be either discharge or radio-frequency (RF) excited. The lasing wavelength is ca. 10.6 µm. Lasing in carbon dioxide occurs between vibrational levels of the molecule rather than electronic levels. The 10.6-µm transition is between the first asymmetric stretch mode and the symmetric stretch mode of the molecule. The transition is broadened by the rotational modes of the molecule. The carbon dioxide laser can be very efficient; electrical to optical power efficiencies of 30 % can be achieved. Furthermore, because the gas can be circulated, very high continuous powers can be achieved, on the order of tens of kilowatts. Because of their high continuous powers, carbon dioxide lasers find wide application in processing of materials, for example, welding, cutting, and heat treatment.

A third important gas laser is the *argon-ion laser* in which the gain medium is singly ionized argon. Inversion is obtained by discharge. Lasing can be achieved at a number of wavelengths, the most important of which are 488.0 and 514.5 nm. Operation in the UV region at 363.8 and 351.1 nm can be achieved by using doubly ionized argon. Argon lasers have the advantage of little or no competition between the major groups of lines; hence, operation can occur at many wavelengths simultaneously. Selec-

tion of a single wavelength can be made by an intracavity prism or by wavelength-selective cavity mirrors. Continuous powers of 10 – 30 W can be obtained by operating at several wavelengths simultaneously.

Other important, but less widely encountered, gas lasers include *nitrogen excimer,* and *chemical* lasers. Further information is given in [3].

3.2. Dye Lasers

Of particular interest to chemists are lasers that are tunable in frequency. Considerable effort has gone into developing organic dyes that lase over a continuous range of frequencies (see also → Laser Dyes). Because of the short upper-level lifetimes characteristic of dyes, excitation is provided by another laser that is absorbed by an absorption band of shorter wavelength than the dye emission band. Short-pulse flash-lamp excitation may also be used. Typical pump lasers are nitrogen, argon ion, and frequency-doubled Nd – YAG (neodymium/yttrium – aluminum – garnet). Dyes are available that span wavelengths from 108.4 to 310.0 nm, but because of the limited tuning range of each dye, ten or more different dyes may be required to cover that range. Lasing in dyes can be continuous or pulsed. For continuous lasing, the dye is circulated through a jet that forms the gain region. Wavelength selection is obtained by using a grating as one of the end mirrors in the cavity or as an intracavity element. Continuous powers of a few tens of milliwatts can be achieved by using continuous laser pumping. Less power is to be expected from the wavelength extremes of any dye or from the more exotic dyes. In pulsed operation, up to 1 J of energy can be obtained, although 0.1 J is much more typical. Repetition rates of >100 Hz can be obtained at these energies. A consequence of the tunability of dye lasers is an extremely broad gain bandwidth. Pulses of <30 fs [4] have been generated by using colliding-pulse mode-locked ring dye lasers. A widely used dye is Rhodamine 640 in methanol – ethylene glycol solvent, tuning from 600 to 700 nm with argon-ion laser pumping.

3.3. Solid-State Lasers

Solid-state lasers are those having a solid active medium rather than a semiconductor (see Section 3.4). In most such lasers, the host medium is a crystal or a glass, and the lasing species is present either substitutionally or interstitially. In color center lasers, the lasing species is actually a type of crystal defect. Many examples of solid-state lasers are available; however, this section discusses only the technologically important Nd^{3+}-based lasers and the tunable Ti^{3+} – sapphire laser.

The neodymium ion, Nd^{3+}, is a rare-earth ion with a partially filled subshell of $4f$ electrons; however, a $5d$ electron participates in the chemical bonding of Nd^{3+} in the solid state. The inner $4f$ electrons are thus shielded to a large degree from the influence of the host, and transitions between $4f$ levels are similar to those of the free ion. Since electric dipole transitions among the $4f$ levels are forbidden by the parity selection rule, lifetimes are long and significant energy storage in the form of inversion is possible. The Nd^{3+} ion is a four-level laser system. Lasing occurs near 1.06 µm. Pumping is achieved by absorption of intense visible light which excites the $4f$ electron into levels lying above the $^4F_{3/2}$ upper laser level. The higher level excited state relaxes to the upper laser level by emission of phonons. This generates heat in the host and is a drawback. Heat in the host causes changes of polarization and phase in the laser beam. Typically, twice as much heat is generated as is energy stored in the form of inversion. Heat removal is facilitated by using a plate geometry for the lasing material, rather than the more familiar rod geometry.

Hosts for Nd^{3+} may be divided into the classes crystalline and glass. Crystalline hosts generally have thermal conductivity an order of magnitude higher than glasses. They also have similarly higher cross sections and, therefore, gains. The Nd^{3+} ion typically substitutes for one of the crystal ions; thus, each Nd^{3+} ion experiences practically the same local field environment. This leads to narrow absorption and emission bands. These attributes make crystalline hosts suitable for cw and quasi-cw applications at relatively high average powers. For example, the Nd^{3+} ion in the crystalline YAG host has achieved > 1 kW of cw power when pumped by

krypton arc lamps. Beam phase and polarization can be greatly reduced by using a plate geometry for the YAG.

Glass hosts can be cast in sizes much larger than those attainable by crystal growth. Glass laser disks of dimensions $30 \times 60 \times 4$ cm have been successfully made. Unlike most crystals, glass is optically and mechanically isotropic and can be cast with superb optical quality. In glass, the Nd^{3+} ion may occupy both substitutional and interstitial sites. Because of the random nature of the neodymium sites in glass, the absorption and emission bands are substantially broadened. For example, the emission bandwidth of Nd – YAG is ca. 1 nm, whereas that of Nd^{3+} in phosphate-based glass is 20 nm. Similarly, the emission cross section is reduced, making possible efficient pumping and storage of energy as inversion. The Nd – glass lasers are thus ideal high-peak-power lasers. Relatively small Nd – glass lasers are capable of achieving a large fraction of a terawatt peak power in pulses of a few picoseconds. Large inertial confinement fusion drivers have achieved more than 50 TW of peak power in a pulse of a few nanoseconds from a multibeam system.

The Ti^{3+} – sapphire laser is an example of a vibronic system that is tunable over 700 – 950 nm. The laser transition consists of the emission of a photon and a quantum of the crystal lattice vibration, a phonon. Energy may be partitioned continuously between phonon and photon, giving rise to the tunability of optical emission. The Ti^{3+} ion consists of an argon core and one $3d$ electron. The fivefold degenerate free-ion d-electron levels are split by the crystal field. The lasing transition is between two of these levels and is thus strongly coupled to the lattice vibrations via the crystal field.

The upper lasing level lifetime is relatively short, ca. $3\,\mu$s at room temperature. Pumping is therefore usually by another laser. Flash lamps, dye lasers near 500 nm, doubled Nd – YAG lasers at 532 nm, and argon-ion lasers operating on blue and green lines have been used. Continuous-wave lasing has been achieved with an argon-ion pump laser. Several hundred millijoules per pulse have been achieved with doubled Nd – YAG pumps. The usual geometry is to mount the crystal on a cold plate and have the pump and output beams collinear. Operation can be achieved readily on a single longitudinal mode. The pulsed output may be doubled to achieve tunability over most of the visible region.

3.4. Semiconductor Lasers

Lasing action can occur in semiconductor p – n junctions. The upper and lower states in this case are the conduction band and the valence band, or the conduction band and an impurity band that lies just above the valence band. Emission of a photon occurs when an electron in the conduction band combines with a vacancy (hole) in the valence band. This requires a partially filled conduction band and a partially empty valence band in a spatially localized region; p-type semiconductors have a partially empty valence band, whereas n-type semiconductors have a partially filled conduction band. When these two types of semiconductor material are brought together in a p – n junction in thermal equilibrium, the energies of the last-filled valence band state in the p-type material and the last-filled conduction band state in the n-type material are equal. Application of a forward bias to the junction raises the energy of the last-filled conduction band in the n-type material. Lasing action can occur in the region of the junction where the bands from the p- and n-type materials are adjacent. This localization can act as an emission guide in one of the spatial dimensions. Laser output is taken through the ends of the device (facets). Typical spontaneous emission times are of the order of a few nanoseconds, and the distance carriers diffuse before recombination can be fairly large (a few microns) for semiconductor-sized structures. This requires large carrier injection rates or, equivalently, junction current to maintain lasing. Structures more complicated than the simple p – n junction are required for cw lasing to confine the carriers to the active region. This is achieved by surrounding the active region or layer with cladding layers in the heterostructure laser diode. Because the cladding layers have a lower refractive index than the active layer, this construction confines the laser output more effectively in one spatial dimension. Further guiding can be obtained by burying the entire heterojunction, except for the facets, in a material of lower index. This is the so-called buried heterostructure diode laser.

Single laser diodes, because of their small size, are limited to modest cw powers of the order of several hundred milliwatts. Pulsed powers can be substantially higher. Diodes may be formed together by semiconductor processing techniques into centimeter-long bars or arrays of diodes. Kilowatts of power per square centimeter of array can be obtained in research devices. Furthermore, the output of individual diodes in these arrays may be phased so that a single mode can be extracted from the device.

Wavelengths of semiconductor lasers are determined by the band gap energy of the materials used to construct them. Al – Ga – As devices operate in the 700 – 900 nm range, whereas In – Ga – As and In – Ga – As – P devices operate at 1200 – 1600 nm. See also → Electroluminescent Materials and Devices, Chap. 5.

4. Uses

The uses of lasers are growing rapidly. This chapter describes those uses that are encountered most commonly.

4.1. Spectroscopy

The laser has revolutionized optical spectroscopy. The tunable, single-frequency dye laser found immediate application in ultrahigh-resolution atomic spectroscopy. Precision measurements of Zeeman, Stark, fine, and hyperfine splittings can be obtained. Because the line width of tunable dye lasers can be <1 MHz, Doppler line widths can easily be measured. Early work centered on atomic and molecular beams, in which thermal motion can be confined to a direction perpendicular to the direction of observation. This largely eliminates Doppler broadening in these observations.

Furthermore, laser techniques have been developed to eliminate the cumbersome molecular beam apparatus required by early experiments. An example of these techniques is saturation spectroscopy, in which the laser is used to label a velocity subset of the molecules under observation. A laser beam is split into two beams, one weaker than the other, that counterpropagate through the sample cell. The stronger of the two, the saturating beam, is amplitude-modulated, either electro-optically or with a chopper. This strong beam partially saturates the absorption seen by the weaker probe beam. By detecting just that part of the probe beam modulated at the same frequency as the saturating beam, only the "tagged" molecules that have interacted with both beams (i.e., those that have no velocity component along either direction of propagation) are observed. Numerous variations on this technique are possible. The same pump – probe beam technique may be used to modulate the polarizability of molecules of a sample between crossed polarizers. When the pump beam irradiates the sample, molecules alter the state of polarization of the probe beam, allowing it to pass the second polarizer. Another, similar technique uses counterpropagating beams tuned to a two-photon transition and detects fluorescence from the upper level of this transition.

Absorption spectroscopy can be performed with unprecedented sensitivity by using tunable lasers. The high collimation of lasers allows long sample paths. Instead of directly detecting the transmitted beam, relative absorption may be measured by detecting the acoustic wave (optoacoustic spectroscopy) or electrical characteristic change (optogalvanic spectroscopy) induced in the sample by a modulated beam. An increase in sensitivity by a factor of $10^5 - 10^6$ over direct techniques can be achieved by placing absorption samples inside the laser resonator and measuring laser output versus tuning frequency. Multiple passes of the light within the cavity effectively multiply the cell length. Ultimate sensitivities are achieved by monitoring fluorescence or photoionization induced by the illuminating beam.

Short-pulse lasers allow time-resolved spectroscopy. The large gain bandwidth of dye lasers also allows the generation of intense pulses that may be as short as a few hundred femtoseconds. Extremely short fluorescence lifetimes and relaxation processes can be investigated by using these lasers. Furthermore, coherent transient phenomena such as self-induced transparency and photon echoes, which yield insight into radiative damping, may be observed with short pulses.

4.2. Remote Sensing

Remote sensing of atmospheric constituents and pollutants became possible with the advent of the laser. The three most common processes used are Raman scattering, linear absorption, and fluorescence. In the majority of applications, a pulsed laser is used with time-gated detection to obtain range information. The pulse is transmitted via a telescope, and the return signal is detected by using either the same telescope or a physically adjacent optical system. By measuring the strength and spectrum of the return signal, data can be obtained on the type and quantity of the sensed species, the thermal conditions, and the velocity field at the remote location.

Raman scattering is particularly useful in detecting species such as water vapor and methane. Raman cross sections are typically small, and the relatively large concentration of these species helps improve the signal-to-noise ratio of the measurement. The first Stokes vibrational line is usually detected. The data are obtained as the typical vertical density profile.

If a laser is available that can be tuned to an absorption feature of the species of interest, the technique of differential absorption lidar (DIAL) may be used (lidar = *li*ght *d*etection *a*nd *r*anging). The return signal for DIAL is generated by Rayleigh scattering from the air molecules and Mie scattering from the atmospheric particles. The DIAL technique compares the magnitude of the return signal with the laser tuned on the absorption feature to the return signal with the laser tuned off the absorption feature. Since the scattering that generates the return signal changes little with the small change in laser frequency, the round-trip absorption from the transmitter to the remote region and back is obtained. By repeating the measurement for a slightly different distance, the actual remote absorption may be determined. This technique has been successfully applied to the detection and profiling of pollutants such as nitrogen dioxide, sulfur dioxide, and ozone.

Finally, if laser-induced excitation of a species is possible, its fluorescence can be detected. This technique is useful for the upper atmosphere where collisional deexcitation rates are reduced. The hydroxyl radical and alkali metals have been detected by using this technique.

4.3. Laser-Assisted Chemical Processes

Laser-assisted chemical processes have not found widespread use in the chemical industry. The reason for their limited application is the high cost of laser light. The few successful applications have occurred when no practicable alternative exists.

One successful application is in nuclear chemistry where lasers have replaced gaseous diffusion and centrifugation in isotope separation. Naturally occurring uranium contains ca. 0.7 % ^{235}U, with the remainder being largely ^{238}U. The isotope ^{235}U is fissionable, whereas ^{238}U inhibits fission by fast-neutron capture and gamma-ray emission. Enrichment of the isotopic abundance of ^{235}U to greater than 3 % is necessary for use in reactors.

A similar application exists in reducing the abundance of the isotope ^{91}Zr in zirconium used in reactor fuel rod cladding. The ^{91}Zr with 11 % natural abundance accounts for 75 % of the thermal neutron capture in the cladding. It is desirable to reduce the abundance of ^{91}Zr to < 3 %.

The process used for uranium enrichment is the atomic vapor laser isotope separation (AVLIS) technique. In this process, advantage is taken of the isotope shift in the absorption spectra of the two uranium isotopes to selectively excite and ionize atoms of ^{235}U, leaving ^{238}U unaffected. This requires that the isotope shift be well-resolved. Because the inner unfilled d- and f-electron shells are shielded by the outer, filled s-electron shell, transitions involving these electrons are spectrally narrow, and the 0.2-cm^{-1} isotope shifts in the spectral absorption lines can be observed. The transitions are spread throughout the visible region. The ionization energy of uranium is ca. 6 eV, and the energy of a green photon is approximately 2 eV. Three photons, each tuned to a specific transition of ^{235}U, are absorbed by ^{235}U atoms until they are ionized. The ^{235}U ions may then be separated from unchanged ^{238}U atoms electrostatically. Atomic uranium is generated by electron-beam vaporization of a uranium melt. Aperturing of the vapor from the melt creates an atomic beam to limit Doppler effects. The requisite three separately tuned photons are generated by dye lasers pumped with a copper-vapor laser. Large, high

repetition rate laser powers, of the order of 100–1000 W, are required for an economically viable facility.

4.4. Metrology

Metrology, in particular precision metrology, is synonymous with the field of optical interferometry which predated the invention of the laser. It was the laser's vastly improved coherence, both spatial and temporal, and its greater brightness that revolutionized metrology. Metrology may be somewhat arbitrarily divided into two broad areas: classical interferometry and holography. In either area, the interference of coherent light permits measurement.

The classic example in interferometry is the Mach–Zehnder interferometer. In this device, a laser beam is split into two beams of equal intensity by an optical beam splitter. The two beams are made parallel by means of a mirror in one beam, and a transparent object is introduced into the other. The two are then combined with another beam splitter. The resulting combined beam has light and dark areas, corresponding to areas in which multiples of a full and a half wave of optical path have been introduced into the sample beam, respectively. By counting these fringes, differences in path of a fraction of the wavelength of light can be measured. By replacing the transparent sample with an optically reflecting surface, surface deformations of the same magnitude may be measured. By counting the number of fringes that pass a point, displacement can be measured; by measuring the rate at which fringes pass, velocity can be measured.

This basic concept may be applied to a large number of measurements. If the reflecting surface in the above example is the wall of a bellows, extremely sensitive measurements of pressure may be made. Similarly, if the reflecting surface is a free diaphragm, an extremely sensitive microphone can be made. Most underwater acoustic sensors are now based on fiber optic implementations of this technique. Temperature can be determined if the thermal expansion coefficient of the transparent plate and the dependence of its refractive index on temperature are known. Advantage may be taken of electro- and magneto-optical effects in transparent materials to measure voltage, electric field, or magnetic field. These techniques are often combined with fiber optics. Sensitive measurements of magnetic fields are now made by measuring the change in optical length of a fiber optic that is coated with a magnetostrictive material. Making the fiber longer increases sensitivity. Any of these measurements can be combined with a pulsed laser to yield high-precision time-resolved data.

Holography (see also → Holography) is similar to interferometry in that the interference pattern between a reference beam and the coherent light reflected by an object is recorded, usually on film. The film is developed and illuminated with a reconstruction beam in the same direction as the original reference beam. The reconstruction beam is diffracted by the interference pattern recorded on the film. The resultant diffracted wave is the same as the original object wave. It contains all the phase information of the original beam and, when viewed with the eye, appears fully three-dimensional.

Holography is used in the measurement of the displacement of surfaces. If holography is performed with a pulsed laser, time-instantaneous images of moving objects may be obtained. Furthermore, if long exposure of a vibrating object is recorded in a hologram, the brightness of the hologram is reduced for those parts of the object that move. The resulting hologram shows a pattern of light and dark fringes corresponding to nodes and antinodes of the vibrational mode.

4.5. Medical Applications

Lasers found their first routine application in medicine in eye surgery, where other techniques were extremely difficult or simply not available.

In certain cases, usually as a result of injury to the eye and sometimes due to aging, the retina may become detached from the back of the eye, resulting in partial or total blindness. In retinal reattachment, a visible laser is focused through the lens of the eye onto the retina and an intense pulse of laser light tacks it back in place.

A similar application is the treatment of diabetic retinopathy. In this condition, small blood vessels on the retina proliferate, degenerate, and hemorrhage. The blood vessels obscure the retina, and blood mixes with the vitreous humor,

further obscuring vision. A low- to moderate-power argon-ion laser focused on the retina is used to destroy excess blood vessels. This technique makes use of the "bloodless" nature of the laser. Blood vessels are cauterized by the laser so that no posttreatment bleeding occurs.

A more recent application involves permanent correction of myopia and astigmatism. In laser keratomy, fine radial incisions are made in the cornea, which accounts for part of the static focusing power of the eye. These incisions flatten the corneal surface, thereby reducing its optical power and restoring sharp vision. The location of the incisions may be varied to correct azimuthally asymmetric visual aberrations. Either an argon-ion laser (with a suitable absorbing dye placed on the cornea) or an excimer laser (whose UV output is absorbed directly by the cornea) may be used. The laser can be used not only to perform surgery but also in a highly attenuated and unfocused mode to sense the shape of the cornea as the operation progresses.

The combination of fiber optics and lasers opens previously inaccessible regions to surgery; an example is laser angioplasty. A fiber optic is a thin (several microns) glass fiber that can transport laser light by total internal reflection. Similarly, a fiber or fiber bundle can transmit diagnostic fluorescence or an image from a remote area. In laser angioplasty, a catheter equipped with one or more optical fibers is inserted into an artery and routed up to a blockage. The laser, usually dye or excimer, is fired and the arteriosclerotic plaque is vaporized. Either the laser energy is absorbed directly by the plaque, or a metal ball is affixed to the end of the fiber and the laser is used to heat the ball to remove the plaque. The latter technique does not allow use of the spectral absorptivity differences between the plaque and healthy vessel tissue to avoid damage to healthy tissue. In more advanced systems, a separate low-power laser pulse is used to excite fluorescence in the artery location where the catheter ends. The fluorescence is carried out of the patient by another fiber and analyzed spectrally. Real-time computer processing of the fluorescence spectra can differentiate between healthy and abnormal arterial tissue, and the high-power laser can be controlled so as to prevent damage. This technique has found greatest success in the treatment of leg arteries. The small arteries that serve the heart are often convoluted, requiring catheters more flexible than those currently available. Rapid progress is expected as catheter technology develops.

Another application combining fiber optics and lasers is the reduction of kidney stones. The laser is again introduced via catherization. A very high-intensity laser pulse is used to rapidly heat and fragment the kidney stone. The resulting pieces are then passed naturally.

Lasers also aid in diagnosis. The prime diagnostic application is in flow cytology. Here, a laser illuminates a narrow glass capillary tube through which flows a stream of cells to be diagnosed. By treating the cell sample with a fluorescent dye that is preferentially absorbed by cancerous cells, a fluorescence detector located downstream of the laser identifies the presence of malignant cells. An alternative technique is to examine the diffraction pattern produced by the cells. The pattern may be analyzed by computer to identify misshapen cells, which are presumably abnormal.

5. Safety

Laser safety requires a knowledge of the hazards presented by a laser or laser system. Several excellent guides to laser safety have been published (e.g., the Laser Safety Guide by the Laser Institute of America) describing proper precautions. Much of what is presented here can be found in more detail in [5].

Hazards presented by lasers can be divided into three categories: ocular, skin, and associated hazards.

Ocular Hazards. The eye transmits light and focuses it onto the retina in the wavelength range 400 – 1400 nm. Because of the collimation of laser light, it is focused to a point on the retina. Depending on the state of dilation of the pupil, the intensity on the retina may be more than 10^4 times that incident on the eye. If sufficient energy enters the eye before the blink response (0.25 s), permanent retinal damage can occur. Direct viewing of the laser beam or a specular reflection presents the greatest danger. Viewing of diffuse reflections presents a lesser hazard because they are imaged onto a finite area on the retina, thereby reducing the intensity.

Blue and ultraviolet laser light emitted by certain types of laser (e.g., excimers) present additional hazards. Blue light in the region from 400 to 550 nm with an intensity too low to trigger the blink response may produce retinal damage. Near-ultraviolet radiation in the 300 – 400-nm region is absorbed by the lens of the eye, leading either to damage there or, in the longer term, to possible cataract formation. Hard ultraviolet radiation from 200 to 300 nm is absorbed by the cornea and can lead to similar damage there. The same considerations apply to the often substantial UV output from flashtubes used as pumps for certain lasers (e.g., solid-state and pulsed dye lasers).

Skin Hazards. The principal skin hazard is the possibility of burns with a laser of sufficient power. Particular care should be taken with lasers in the 300 – 400-nm region. Finally, a skin cancer hazard is associated with exposure to UV lasers in the region 200 – 300 nm.

Some relatively simple precautions greatly reduce the probability of an accident involving the eye or skin. The route of the laser beam and of its specular reflections must be known. Beam blocks or enclosures should be used where necessary. Adequate eyewear should be worn; the exposure should be estimated and the eyewear absorbance at the laser wavelength measured. Lasers should not be operated at eye level. A laser must not be boresighted under any circumstances, even if it is supposedly deenergized. Untrained personnel should not work with lasers.

Associated Hazards. Other hazards are associated with lasers. The greatest danger arises from the high voltage often used. Other hazards arise from gases (e.g., fluorine for certain excimers), chemicals (e.g., solvents in certain dye lasers), fires inadvertently ignited by high-power (a few watts) cw lasers, and cryogenic gas hazards. In cutting and drilling applications, gaseous and particulate products may represent a respiratory hazard. Most of these dangers can be handled in the same manner as in nonlaser applications; for example, adequate insulation and venting must be used when handling cryogenic gases for lasers. In the case of high-power cw lasers, beam dumps of adequate absorption with sufficient thermal mass should be used. The chemical safety data sheets for dyes and solvents used in any laser should be read carefully for proper handling procedures. Laser gas handling manifolds should be designed specifically for the type of gas involved. Adequate venting and particulate filtration should always be provided in materials-processing applications.

The high-voltage hazard deserves special mention because of the fatalities it has caused among laser workers. Commonly accepted safety procedures for working with high-voltage sections of a laser should always be followed. High voltages should be enclosed in interlocked enclosures. Energy storage capacitor circuits should be equipped with some form of "soft discharge" capability that occurs on power down and interlock violation. Manual grounding sticks should be used on high-voltage parts prior to servicing. Workers who are tired, hungry, medicated, agitated, or otherwise incapable of concentrating should never work on high voltage. No one should work alone on high voltage. Metallic objects such as rings and watchbands should not be worn. All floors are to be considered conductive. Workers should stand on an insulating dry mat and use only one hand when troubleshooting. Emergency procedures for aiding victims of electrical shock should be posted.

Laser Classification. All lasers are classified according to the potential hazard that they represent. The classification described here is due to the American National Standards Institute (ANSI), standard Z-236.1, "Guide for Safe Use of Lasers (1986)." A brief description of each classification follows:

Class 1	No hazard exists.
Class 2 (a)	Lasers present no hazard for direct viewing up to 1000 s.
Class 2 (b)	Lasers do not present a hazard for direct viewing because of the normal blink response.
Class 3 (a)	Lasers present an immediate hazard if viewed with collection optics.
Class 3 (b)	Lasers are hazardous if viewed directly. Diffuse reflections are usually not hazardous except with higher powered lasers.
Class 4	These are the most hazardous lasers. Diffuse reflection is hazardous. Fire and skin hazards exist.

6. References

1. Y. R. Shen: *The Principles of Nonlinear Optics,* John Wiley and Sons, New York 1984.
2. A. E. Siegman: *Lasers,* University Science Books, Mill Valley, California, 1986.
3. R. J. Pressley (ed.): *CRC Handbook of Lasers, with Selected Data on Optical Technology,* CRC Press, Cleveland, Ohio, 1971.
4. P. W. Smith, A. M. Weiner, *IEEE Circuits and Devices* **4** (1988) no. 3, 3.
5. J. F. Smith (ed.): *Laser Safety Guide,* 6th ed., Laser Institute of America, Toledo, Ohio, 1987.

Latex → Rubber, 2. Natural
Latex Technology → Polymerization Processes

Laxatives

Robert Engelhorn, Biberach, Federal Republic of Germany (Chap. 1 – Section 2.4.2)

Ernst Seeger, Biberach, Federal Republic of Germany (Chap. 1 – Section 2.4.2)

Jan H. Zwaving, University of Groningen, Groningen, The Netherlands (Section 2.4.3)

1. Introduction and History 1
2. Types of Laxatives 2
2.1. Emollients (Lubricants) 2
2.1.1. Paraffinum Liquidum 2
2.1.2. Glycerol 3
2.1.3. Dioctyl Sodium Sulfosuccinate 3
2.2. Bulking Agents 3
2.2.1. Agar 3
2.2.2. Linseed 3
2.2.3. Plantago Seed 4
2.2.4. Methylcellulose 4
2.2.5. Tragacanth 4
2.2.6. Wheat Bran 4
2.3. Osmotically Active Substances . . . 4
2.3.1. Sodium Sulfate 4
2.3.2. Magnesium Sulfate 5
2.3.3. Sugars and Sugar Alcohols 5
2.4. Antiabsorptive and Secretion-Stimulating Substances and Compounds that Selectively Increase the Rate of Colonic Peristalsis 5
2.4.1. Synthetic Diphenol Laxatives 5
2.4.2. Naturally Occurring Laxatives 7
2.4.3. Anthraquinone Laxatives 7
3. References 9

1. Introduction and History

In past centuries defecation induced by orally or rectally administered medications was widely applied by the medical profession. This practice, like phlebotomy, was intended to help cure ailments whose origin was often unknown.

Purgation was highly valued in earlier times, because knowledge of physiological and pathophysiological interactions was limited. Today however, medical science has a clear understanding of how, when, and where the use of laxatives is clinically indicated. Although this should have considerably reduced their use, laxatives are often important ingredients of pharmaceuticals which are designed to help achieve rather dubious goals (blood cleansing, alleged reduction in body weight). Their popularity is often further increased by reference to their "natural", "biological", or plant origin. The consumer is deliberately misled: the active ingredients are chemically clearly defined compounds – it is irrelevant whether they are produced synthetically or derived from plants.

A large proportion of laxatives or laxative-containing compounds are consumed without a sensible therapeutic goal and as uncontrolled self-medication. The dangers involved with such behavior, particularly after long-term use, (e.g., functional and morphological disturbances of the gastrointestinal tract; Na^+, K^+ and water loss; insufficient degradation and absorption of essential nutrients; secondary hyperaldosteronism) are often disregarded.

Mastication initiates the digestive breakdown of solid foods. Thorough chewing facilitates the chemical degradation of the nutrients. Saliva, constituents of gastric fluid and bile, pancreatic secretions, and enzymes from the intestinal epithelial cells cleave proteins, fats, and carbohydrates into absorbable compounds. Furthermore, water, vitamins, and minerals are absorbed from the intestinal lumen and enter the blood and lymph fluid. However, electrolytes and water are not merely absorbed, they can also be secreted back into the intestinal lumen by the action of various laxatives. The solid matter reaching the rectum consists of nondigestible material (cellulose, intestinal bacteria) and water.

Two vagus-dependent neuronal plexus (Plexus myentericus and Plexus submucosus) are predominantly responsible for the motor

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_183

control of the intestinal musculature. Sympathetic nerve fibers from the ganglion coeliacum innervate the musculature and the blood vessels vascularizing the intestine. The plexus are responsible for two types of coordinated motor activity. The first type – nonpropulsive segmentation contraction – is a result of localized constrictions and dilations of the intestinal lumen and leads to a mixing of the contents in the small and large intestines. The second type of motor activity – propulsive contractions – occurs several times a day and causes the fecal mass to be transported in the direction of the rectum. The intensity and frequency of the contractions depend on the degree of intestinal filling and the associated stimulation of intrinsic stretch receptors. From about two years of age onward, the cerebral cortex also influences the propulsive contractions. The initiation of defecation is then not dependent solely on autonomic control, but is also subject to conscious influences. If the physiological reflex resulting from the internal pressure stimulus is repeatedly suppressed, the reflex threshold is raised so that a greater internal pressure is required to initiate defecation. Since prolongation of the retention time of the feces in the rectum is associated with absorption of water, inspissation of the feces results; this is often the reason for obstipation. Emotional occurrences can accelerate defecation.

The individual defecatory rhythm of healthy adults varies considerably (1 – 4 times per day up to once every second or third day).

Obstipation can be caused by spastic contractions of the intestinal musculature with uncoordinated transport, by atonic colonic and rectal musculature, as well as by insufficient excitability of the defecatory reflex (dyschezia).

Obstipation can also be associated with other diseases, such as obstructions of the gastrointestinal tract, appendicitis, hemorrhoids, or anal fissures. Nonsteroidal anti-inflammatory agents, sedatives, analgesics, anticholinergics, opiates, antacids, and substances with an anticholinergic action in addition to their primary pharmacologic effect can all cause obstipation.

A very important factor in the development of obstipation are nutritional mistakes, especially when the diet contains insufficient indigestible material (e.g., cellulose) or when physical activity is restricted.

Therapy for obstipation depends on the underlying causes. A prerequisite for the short-term medically correct use of laxatives is to rule out beforehand those causes that cannot be cured or can only be cured by other therapies. The use of laxatives is thus limited to the following clinical indications:

1) Emptying of intestinal contents before surgical, endoscopic, or radiographic procedures
2) Facilitation of defecation in case of hernias and anal fissures by reducing the consistency of the stool; avoiding an increase in intraabdominal pressure when there is danger of apoplexia
3) Increased elimination of poisons taken orally

Finally, cases of chronic obstipation exist which do not improve even after a change in life style and diet. These cases require a continued careful treatment with laxatives.

Laxatives are administered to soften inspissated feces and to initiate the defecation. Laxatives can therefore be divided into four groups according to their mode of action.

1) Emollients (lubricants)
2) Bulking agents
3) Osmotically active substances
4) Antiabsorptive and secretion-stimulating substances and compounds that selectively stimulate the transport capacity of the colon

2. Types of Laxatives

2.1. Emollients (Lubricants)

2.1.1. Paraffinum Liquidum

Paraffinum liquidum (mineral oil) (DAB 9, Ph. Eur.) specific gravity 0.827 – 0.892 is still occasionally used. If taken with food, it can be absorbed to a small degree and thus lead to the formation of foreign body granulomas. Furthermore, its use can lead to hypovitaminosis of fat-soluble vitamins. For these reasons it should only be used therapeutically in case of organic solvent poisonings.

The single dose for adults is one tablespoon. Mineral oil acts after 8 – 12 h.

The simultaneous use of dioctyl sodium sulfosuccinate with paraffin can lead to an increased absorption of the latter. These two substances should therefore not be used together.

Trade Names. Obstinol Mild (Thiemann); as an additive to Salus-öl (Salushaus); Sanato-Lax (Hotz); Agarol (Goedecke); Sanato-Lax-forte (Hotz).

2.1.2. Glycerol (→ Glycerol)

Glycerol [*56-81-5*], 1,2,3-propanetriol, $C_3H_8O_3$, M_r 92 (DAB 9, Ph. Eur.), is a clear, colorless strongly hygroscopic fluid, which acts mainly because of its osmotic effect. It can soften hardened feces and also works as a lubricant. Furthermore it stimulates rectal peristaltis.

Glycerol can be used either in suppository or enema form. The dosage for babies and infants up to six years of age is 1.0 – 1.5 g, for adults 1 – 3 g. It acts within ca. 2 h.

Trade Names. Babylax (Dentinox); Glycilax (Engelhard); Microklist (Pharmacia); Bébégel (Plantier); Glysolax (Cooper).

2.1.3. Dioctyl Sodium Sulfosuccinate

Dioctyl sodium sulfosuccinate [*577-11-7*] (docusate sodium, DONSS), $C_{20}H_{37}NaO_7S$, M_r 444.5, (U.S.P. XXI, The Extra Pharmacopeia, 28th ed. Martindale, 1982) is a wax-like substance which is sparingly water-soluble (1.5 wt %). It is synthesized by reaction of ethylhexanol with maleic anhydride and introducing the sulfonic acid group with sodium hydrogensulfite [9].

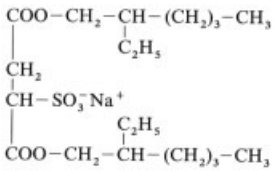

The compound has a softening effect and reduces surface tension. It therefore is supposed to impede excessive inspissation of the intestinal contents and acts as a lubricant.

In many European countries and in the United States it is available as a pure substance in 100 mg capsules or as a syrup. In the Federal Republic of Germany it is only available in combination with other laxatives.

Trade Names. Agaroletten (Goedecke); Florisan (Biotherax); Klyxenema (Ferring/Braun); Docusate Sodium (Lederle Laboratories); Doctynol (Astra); Regulex (Ayerst).

2.2. Bulking Agents

2.2.1. Agar

Agar [*9002-18-0*] (DAB 9, Ph. Eur.), is a constituent of cell walls from algae (→ Polysaccharides, Chap. 6.) which belong to the *Gelidium* species (mainly *Rhodophyceae* species). Approximately 90 % of it consists of polysaccharides with acidic groups. Their matrix is formed by D- and L-galactose units and always contains small amounts of galacturonic acid. The acidic groups are present partially as calcium, magnesium, sodium, and potassium salts.

Agar is produced by extraction of the algae with boiling water. The hot extract is filtered, concentrated and dried. Agar is tasteless, odorless and varies from colorless to slightly yellow. It is strongly hygroscopic and has a strong swelling capability. Agar is not metabolized in the gastro-intestinal tract.

The dosage is 4 – 16 g, one to two times per day. It acts within 8 – 12 h.

Trade Names. Agarol (Goedecke); Laxatan (Divapharma-Knufinke).

2.2.2. Linseed

Linseed, Lini Semen (DAB 9, Ph. Eur.) is the ripe dried seed of *Linum usitatissimum*. It contains 30 – 40 wt % fatty oils, 25 wt % protein, and 3 – 20 wt % mucilage. Through their swelling capability the linseed polysaccharides increase the volume of the stool by absorbing water, stimulate peristalsis and accelerate defecation. They are not absorbed. The therapeutic dose is 20 – 50 g. Linseeds act after 12 – 24 h.

2.2.3. Plantago Seed

Plantago seed, Psyllii Semen, (DAB 9) consists of the ripe seeds of *Plantago psyllium* and *Plantago indica.* Indian Plantago seed, *Plantaginis ovatae* semen, are the seeds of *Plantago ovata* and *Plantago ispaghula.* The seeds contain 10 – 12 % mucopolysaccharides which form two fractions, one is soluble in cold water, the other in hot water.

Mixed with an equal portion of water, the seeds form a mucilaginous mass. Not only the seeds, but also the seed coats are used (Plantaginis ovatae testa).

The dosage is between 2.5 and 5 g. Plantago seed is supposed to act within 12 – 24 h.

Trade Names. Agiolax (Madaus); Psyllium (Kneipp); Laxiplant (Schwabe); Naturacil (Mead Johnson & Comp).

2.2.4. Methylcellulose

Methylcellulose (methylcellulosum, DAB 9, Ph. Eur.) is a partially methylated derivative of cellulose (→ Cellulose Ethers, Chap. 3.). It is hydrophilic and nonabsorbable. With addition of water an opalescent, viscous colloidal solution is formed. This solution softens stool within 1 – 3 d. The commonly used therapeutic dosage is a total of 4 – 6 g per day for adults taken in 2 – 3 divided portions. The dosage for children is 1.0 – 1.5 g.

Trade Names. Laxariston-Granulat (Steiner); Methocel (Dow Chemical); Cellothyl (Warner – Chilcott); Cellucon (Medo-Chemicals).

2.2.5. Tragacanth (→ Polysaccharides, Chap. 8.)

Tragacanth [*9000-65-1*] (tragacantha, DAB 9, Ph. Eur.) is the hardened gum which is secreted by the trunk and branches of members of the *Astragulus* species, especially *Astragulus gummifer* as well as *A. microcephalus*. It is odorless and has a bland, slimy taste. Pulverized it is white or off-white.

Tragacanth is a complex mixture of polysaccharides. It is only partially water-soluble (15 – 40 %). The water-insoluble but strongly swellable main fraction is called bassorin, the water soluble fraction is called tragacanthin. One part tragacanth and ten parts water forms a mucoid gel.

The recommended dosage is one heaping teaspoon of the granulated material taken mornings and evening with plenty of water.

Trade Name. Normacol (Asche).

2.2.6. Wheat Bran

Wheat bran contains the seed husks which are separated from the wheat kernel during milling and are extracted from the flour by sifting. It contains predominantly (> 40 %) indigestible cellulose as well as 15 – 18 % protein and vitamin B_1 and E. Consumption of approximately 6 g of bran already leads to a noticeable increase in fecal volume. Innumerable commercial products contain bran.

Side Effects of Bulking Agents. Bulking agents merely represent an expensive alternative to a fiber-rich diet. As dried substances they require large amounts of fluids to be consumed simultaneously to prevent obstipations or the forming of conglomerates in the intestines.

2.3. Osmotically Active Substances

2.3.1. Sodium Sulfate

Sodium sulfate [*7757-82-6*] (Natrii sulfas decahydricus, Natrium sulfuricum, Glauber's salt), $Na_2SO_4 \cdot 10\,H_2O$ (Ph. Eur.), is absorbed poorly in the gastrointestinal tract and binds water. Thereby, an osmotically equivalent amount of fluid is retained in the intestinal lumen and prevents hardening of feces. Sodium sulfate should never be used as a hypertonic solution because of the danger of dehydration and thrombosis.

The best result is achieved with a solution of 20 g of the salt dissolved in 500 mL of water, which is almost an isotonic solution. It acts after 2 – 4 h. With prolonged use, sodium retention, edema and hypertension can occur. Therefore, Glauber's salt should only be used with care by hypertonics.

2.3.2. Magnesium Sulfate

Magnesium sulfate [*7487-88-9*] (Epsom salts) $MgSO_4 \cdot 7\,H_2O$, (Ph. Eur.), exerts its laxative effect in the same way as Glauber's salt. Inadequate magnesium ion excretion can occur leading to muscular weakness, reflex deficits and hypotension as side effects in pediatrics as well as with renal insufficiency.

2.3.3. Sugars and Sugar Alcohols (→ Sugar Alcohols, → Lactose and Derivatives)

Mannitol and sorbitol act similarly to the salinic laxatives and are scarcely absorbed.

Mannitol. D-Mannitol, [*69-65-8*] (DAB 9), $C_6H_{14}O_6$, M_r 182, *mp* 166 – 168 °C, is a sweet-tasting hexavalent alcohol which is widespread in nature, is easily soluble in water, and slightly soluble in ethanol. Mannitol is produced by catalytic hydrogenation of sucrose or invert sugar.

Sorbitol. Sorbitolum [*50-70-4*], D-glucitol, $C_6H_{14}O_6$, M_r 182,

mp 110 – 112 °C (anhydrous substance) (DAB 9, Ph. Eur.) cristallizes with 0.5 and 1 mol of water. It is a white, sweet-tasting compound which is highly soluble in water and occurs naturally in many plants.

Sorbitol is produced commercially by pressure hydrogenation of D-glucose [10].

Dosage: 5 – 15 g total daily taken in several divided doses.

Trade Name. Klysma-Sorbit (Pfrimmer).

Lactose (Lactosum) [*63-42-3*], 4-*O*-β-D-galactopyranosyl-α-D-glucopyranose monohydrate $C_{12}H_{22}O_{11} \cdot H_2O$ (DAB 9), M_r 360.3, *mp* 201 – 202 °C, is a white crystalline substance. It is easily soluble in water, almost insoluble in ethanol. Commercially, lactose is extracted from whey (see → Lactose and Derivatives, Chap. 1.2.). Lactose is hydrolyzed in the intestine. Its metabolites lower the pH value of the feces.

Lactulose [*4618-18-2*], 4-*O*-β-D-galactopyranosyl-D-fructose, $C_{12}H_{22}O_{11}$, M_r 342.30, *mp* 169 °C, is a sweet-tasting compound which is easily soluble in warm water [11]. Lactulose is a synthetic disaccharide consisting of D-galactose and D-fructose (see → Lactose and Derivatives, Chap. 2.1.). It is produced by alkaline epimerisation of lactose.

Lactulose passes through the stomach unaltered and is not hydrolyzed by the enzymes of the small intestine. Beginning in the large intestine the microflora cleaves it into short-chained fatty acids, especially lactic and acetic acids.

Lactic and acetic acids acidify the intestinal contents, retain water in the lumen and stimulate large intestinal peristalsis. Through the change in pH the intestinal bacterial flora is also altered. Lactulose can be used as a mild laxative.

10 – 35 g are recommended as a single dose. The substance acts within 8 – 10 h.

Trade Names. Bifiteral (Duphar); Laevilac (Atmos).

2.4. Antiabsorptive and Secretion-Stimulating Substances and Compounds that Selectively Increase the Rate of Colonic Peristalsis

These substances inhibit the absorption of fluid and sodium ions in the large intestine. Only castor oil is an exception because it is already active in the small intestine. Furthermore, they cause an influx of water into the intestinal lumen together with sodium, potassium, calcium and chloride ions. Thus, the feces are softened and their volume is increased. The resultant distension of the intestinal wall activates the defecation reflex. A direct influence on the visceral smooth muscles increasing peristalsis has to be taken into consideration as a further mode of action of this group of compounds.

2.4.1. Synthetic Diphenol Laxatives

These compounds act in the large intestine. Phenolphthalein, the oldest member of this group, was accidently discovered at the beginning of this century.

Phenolphthalein. Phenolphthalein [*77-09-8*], 3,3-bis-(4-hydroxyphenyl)-1-(3H)-isobenzofuranon, $C_{20}H_{14}O_4$, M_r 318, *mp* 258 – 260 °C, forms colorless crystals that are

almost insoluble in water but easily soluble in ethanol. It is produced by the condensation of phenol with phthalic anhydride in the presence of concentrated sulfuric acid [12].

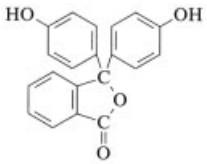

Approximately 15 % of any dose of phenolphthalein is absorbed by the organism and then excreted via urine. If the urine is alkaline a red coloration can occur. The rest (ca. 85 %) is excreted with the feces. The therapeutic dosage is 100 – 200 mg and leads to defecation within 6 – 10 h. Clinically, sporadic albuminuria and hemoglobinuria as well as allergic symptoms have been reported.

Trade Names. Agarol (Goedecke); Obstinol (Thiemann, Chefaro).

Bisacodyl [*603-50-9*], 4,4′-(2-pyridinyl methylene) bisphenol diacetate, bis-(*p*-acetoxyphenyl)-(2-pyridyl)-methane, $C_{22}H_{19}NO_4$, M_r 361, *mp* 136 – 138 °C, forms colorless crystals which are nearly insoluble in water but soluble in acetone, methanol and ethanol. It is produced by condensation of 2-pyridinealdehyde with phenol in the presence of dehydrating compounds. Subsequently, the phenolic hydroxyl groups are acetylated [13].

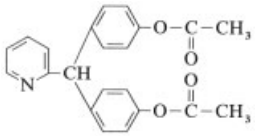

The compound is active after oral and rectal administration. Bisacodyl is available in 5.0 mg enteric coated tablets and in 10.0 mg suppositories. For improved colonic inspection it is also available in liquid form.

The clinical dosage is generally 10.0 mg. The tablets act after 6 – 8 h, suppositories within 15 – 30 min.

Bisacodyl has been used therapeutically since 1952. According to the common scientific opinion its mode of action is as follows: First, the two acetyl groups are split off in the small intestine. This is followed by (1) the absorption of the diphenol moiety, (2) enterohepatic cycling, (3) conjugation with glucuronic acid, and (4) the repeated excretion into the intestinal lumen with bile. From here the conjugates are transported to the large intestine. The microflora cleaves them to the active free diphenol, which stimulates the intestines. The long period of 6 – 8 h between oral administration and the drug's effect is explained by this long "metabolic detour" [14], [15].

However, studies in humans [16], using pH controlled tablets which release the substance in the large intestine, have shown that only small amounts of the cathartically active 4,4′-(2-pyridinyl methylene)bisphenol are absorbed. The same holds true for the administration of suppositories. No correlation between plasma concentrations and the effect of the substance could be found. It is postulated that the laxative effect is a result of interactions or direct contact of the active substance (bisphenol) with sensory nerve endings situated in the lower part of the alimentary tract. This produces increased peristaltic contractions of the colon and would explain the rapid onset of action after rectal administration.

Trade Name. Dulcolax (Dr. Karl Thomae); Contalax (Riker); Theralax (Beecham); Nourilax (Organon).

Sodium picosulfate [*10040-45-6*], 4,4′-(2-pyridinylmethylene)bisphenol, bis(hydrogensulfate) (ester), disodium salt, $C_{18}H_{15}NO_8S_2 \cdot 2\,Na$, M_r 481, *mp* 272 – 275 °C (decomp.), forms colorless crystals (recrystallized from ethanol) which are easily soluble in water.

The substance is synthesized by the reaction of 4,4′-(2-pyridinylmethylene)bisphenol with chlorosulfuric acid in the presence of pyridine. The sulfuric acid ester is converted to its sodium salt with sodium hydroxide [17].

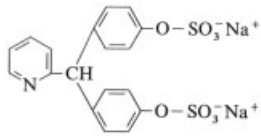

Sodium picosulfate can be much more easily produced by reaction of 4,4′-(2-pyridinylmethy-

lene)bisphenol with the easily available pyridine – sulfur trioxide adduct in dimethylformamide and subsequent addition of sodium carbonate [18].

Sodium picosulfate is available as 5 mg tablets as well as in an aqueous solution (7.5 mg/mL). The dosage for adults is 5 – 10 mg, for children four years or older approximately half this dosage. The substance acts in 5 – 8 h. The concurrent use of antibiotics can block the action of sodium picosulfate. It is only effective after oral not after rectal administration.

Sodium picosulfate is not absorbed to any appreciable extent in the gastrointestinal tract. Intestinal sulfatases are required to liberate the active diphenols by hydrolyzing both of the sulfate groups. In small children the sulfatases are generally not present.

Trade Name. Laxoberal (Biotherax).

2.4.2. Naturally Occurring Laxatives

Castor oil (Ricini oleum), (DAB 9, Ph. Eur.), is extracted from the seeds of *Ricinus communis* by cold pressing. It is a viscous fluid with a characteristic taste. The color varies from light yellow to colorless. It consists mainly of the triglyceride of ricinoleic acid.

The inactive triglyceride is hydrolyzed in the small intestine by pancreas lipases, forming glycerol and ricinoleic acid.

Ricinoleic acid [*141-22-0*], $C_{18}H_{34}O_3$, M_r 298.5 inhibits the reabsorption of fluid and electrolytes and stimulates peristalsis. Within a short time (2 – 4 h), defecation results because of the accumulation of fluids and stimulation of enteral motility. This effect can be so pronounced that it may take several days until intestinal filling is sufficient again to initiate the physiological defecation reflex. Therefore, castor oil should only be used in cases of acute obstipation or prior to diagnostic procedures, if necessary. Chronic use can lead to colics, dehydration, disturbances in electrolyte balance, malabsorption of vital substances as well as damage of the intestinal mucosa. Ricinoleic acid is absorbed and metabolized in the same way as other fatty acids.

The therapeutic dosage is between 15 and 60 mL. Because of its bad taste capsules should be favored.

2.4.3. Anthraquinone Laxatives

A number of laxatives of vegetable origin (aloes, cascara, frangula, rhubarb, senna) owe their activity to anthracene derivatives such as anthraquinones, anthrones, and dianthrones. This group of drugs is called anthraquinone laxatives, because substances of this type were the first to be isolated from these plants.

These natural products have been used as laxatives for thousands of years. At present the primary plant materials or their extracts are widely used, also in selfmedication. For reviews of many aspects of these laxatives see [19], [20].

Chemical Properties. In the plants *O*-glycosides of 1,8-dihydroxyanthraquinones occur together with their reduced forms, the anthrones and dimers of the latter, the dianthrones. Sometimes also *C*-glycosides are present.

A large variety of 1,8-dihydroxyanthracene derivatives occur because of substitution of the hydrogen at C-6 by a hydroxyl or methoxy group and/or methyl, hydroxymethyl or carboxyl substitution at C-3.

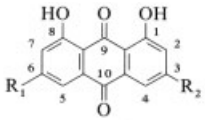

	R_1	R_2
Chrysophanol	$-H$	$-CH_3$
Physcion	$-OCH_3$	$-CH_3$
Emodin	$-OH$	$-CH_3$
Aloe-emodin	$-H$	$-CH_2OH$
Rhein	$-H$	$-COOH$

Mode of Action. Anthracene derivatives induce defecation by stimulating propulsive peristaltic activity of the colon through local irritation. *The absorption of electrolytes and water* from the colon is altered producing net fluid accumulation and laxation by (1) inhibiting intestinal Na^+-Ka^+-ATPase, (2) production of leaktight junctions in the mucosa of the colon, and (3) maybe by increasing the synthesis of prostaglandins and cyclic AMP [20], [21].

Following oral administration the anthracene glycosides are only slightly absorbed from the small intestine. Their sugar residue facilitates their transport to the site of action, the colon.

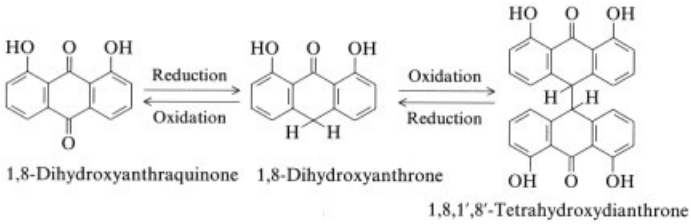

There most of the glycosides are hydrolyzed and reduced by colonic flora to anthrones, being the active form [22].

The free aglycons exhibit little therapeutic activity. Dianthrone glycosides are highly active, anthraquinone glycosides less active. Anthrone glycosides act more drastically causing discomforting griping [20], [21].

Raw Materials. *Aloes* (Ph. Eur. 3) consists of the dried juice of the leaves of *Aloe barbadensis* (Barbados aloes) containing 28 wt % of anthracene derivatives or from the leaves of *Aloe ferox* (Cape aloes) containing 18 % of anthracene derivatives. The main constituent is: barbaloin (aloin), i.e., aloe – emodin anthrone-10-*C*-glycoside

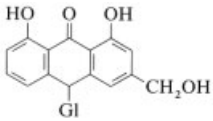

Barbaloin (aloe-emodinanthrone-10-*C*-glycoside)

In Cape aloes also aloinosides (*O*-glycosides of aloin) are present.

Standardized aloes dry extract (Ph. Eur. 3) is prepared from both types of aloes containing 20 wt % of anthracene derivatives.

Cascara (Ph. Eur. 3) consists of the dried bark of *Rhamnus purshianus* (Pacific coast of North America) containing 8 wt % of hydroxyanthracene glycosides. The main constituents are: cascarosides (C-10 isomers of 8-*O*-glycosides of aloin and 11-desoxyaloin) and *O*-glycosides derived from emodin, aloe – emodin and chrysophanol.

Frangula bark (Ph. Eur. 3) consists of the dried bark of *Rhamnus frangula* (Europe) containing 7 wt % of glucofrangulins. The main constituents are: glucofrangulin (emodin-6-*O*-rhamno-8-*O*-glycoside) and frangulin (emodin-6-*O*-rhamnoside).

Frangula bark dry extract standardized (Ph. Eur. 3) is prepared from Frangula bark, it contains not less than 15 wt % and not more than 30 wt % of glucorangulins.

Rhubarb (Ph. Eur. 3) consists of the dried rhizome of *Rheum palmatum* and *Rheum officinale* originally from China, but now also cultivated in Europe. It contains 2.2 wt % of hydroxyanthracene derivatives. Its main constituents are: *O*-glycosides of chrysophanol, physcion, emodin, aloe-emodin, rhein; dianthrone glycosides (i.e. sennosides A, B, C, D, E, F).

Senna leaf (Ph. Eur. 3) consists of the dried leaflets of *Cassia senna* (i.e., *C. acutifolia*) known as Alexandrian or Khartoum senna or of *Cassia angustifolia* known as Tinnevelly senna (India), containing 2.5 wt % of hydroxyanthracene glucosides.

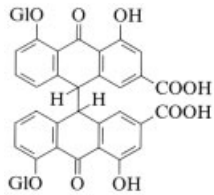

Sennoside A

Senna leaf dry extract standardized (Ph. Eur. 3) is prepared from Senna leaf, it contains not less than 5.5 wt % and not more than 8.0 wt % of hydroxyanthracene glucosides. *Senna pods Alexandrian* (Ph. Eur. 3) consists of the dried fruit of *Cassia senna*, they contain not less than 3.4 wt % of hydoxyanthracene glucosides. *Senna pods Tinnevelly* (Ph. Eur. 3) consist of the dried fruit of *Cassia angustifolia*, they contain not less than 2.2 wt % of hydroxyanthracene glucosides. The main constituents are dianthrone glycosides (sennosides A, B, C, D).

Chemical Analysis. For identification and determination the Bornträger test (or a modification) is used: anthraquinones produce a red color after addition of alkali. Dianthrones and anthrones give this test only after oxidation to anthraquinones [20], [21], [23]. HPLC methods for the determination of the different anthracene derivatives also exist [24].

Toxicology. With the exception of aloes which is very irritating, the anthraquinone laxatives are mild laxatives. However, they are habit-forming and long-term use may result in laxative dependance and loss of bowel function (cathartic colon). Long-term use can produce loss of electrolytes and melanosis coli, a reversible pigmentation of the colonic mucosa. Anthracene derivatives may be distributed in small amounts into the milk of nursing women. Therefore, prolonged or excessive use of these products should be avoided.

3. References

General References

1. L. L. Burton in L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* "Laxatives," 7th ed., Chap. 43, Macmillan Publ. Comp., New York 1985, pp. 994 – 1003.
2. J.-F. Riemann, *Dtsch. Apoth. Ztg.* **121** (1981) no. 26, 1350 – 1355.
3. H.-J. Ruoff in H. P. T. Ammon (ed.): *Arzneimittelneben- und Wechselwirkungen,* Chap. 43, Wiss. Verlagsgesellschaft, Stuttgart 1986, pp. 620 – 628.
4. K.-F. Sewing, *Dtsch. Apoth. Ztg.* **122** (1982) no. 5, 209 – 213. *DAZ-Fortbildung Pharmakologie* **6,** 39 – 43.
5. O. Tauber: "Laxantien und Therapie der Obstipation," *Dtsch. Apoth. Ztg.* **121** (1981) no. 26, 1345 – 1349.
6. F. Waldeck in R. F. Schmidt (ed.): *Thews, Physiologie des Menschen,* "Funktionen des Magen-Darmkanals," 22nd ed., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo 1985, pp. 642 – 667.
7. Zentralstelle des Schweizerischen Apothekervereins: *Index Nominum 1987, Arzneistoff-Synonymverzeichnis,* Société Suisse de Pharmacie, Zürich 1987.
8. E. R. Barnhart, *Physicians Desk Reference,* 41st ed., Medical Economics Company Inc., Oradell, USA 1987.

Specific References

9. Am. Cyanamid, US 2 028 091, 1936, US 2 176 423, 1939 (A. O. Jaeger).
10. E. Boye, *Chem. Ztg.* **82** (1958) 657.
11. E. M. Montgomery, E. Hudson, *J. Am. Chem. Soc.* **52** (1930) 2101.
12. A. v. Baeyer, *Justus Liebigs Ann. Chem.* **202** (1880) 68.
13. Dr. K. Thomae, DE 951 987, 1952 (E. Seeger, A. Kottler).
14. W. Forth, D. Henschler, W. Rummel: *Allgemeine und spezielle Pharmakologie und Toxikologie,* 5th ed., B.I. Wissenschaftsverlag Mannheim/Wien/Zürich 1987, pp. 351 – 358.
15. E. Mutschler: *Arzneimittelwirkungen,* 5th ed., Wissenschaftl. Verlags GmbH, Stuttgart 1986, pp. 503 – 507.
16. W. Roth, K. Beschke, *Arzneimittel. Forsch. Drug Res.* **38** (1988) no. 1, 570 – 574.
17. Istituto de Angeli, US 3 528 986, 1966; US 3 558 643, 1967 (G. Pala).
18. Dr. Karl Thomae, DE-OS 1 904 322, 1969 (E. Seeger, H. Machleidt).
19. J. W. Fairbairn (ed.): Symposium Proceedings. The Anthraquinone Laxatives, *Pharmacology* **14 (Supl. 1)** (1976) 1 – 105.
20. J. W. Fairbairn (ed.): Symposium Proceedings. Natural Anthraquinone Drugs, *Pharmacology* **20 (Supl. 1)** (1980) 1 – 134.
21. *Symposium Proceedings. Symposia of the Giovanni Lorenzini Foundation, "New trends in Pathophysiology and Therapy of the large Bowel,"* **vol. 17,** Elsevier, Amsterdam 1983.
22. M. Dreesen, J. Lemli: "Qualitative and Quantitative Interactions between Sennosides and some Human Intestinal Bacteria," *Pharm. Acta Helv.* **57** (1982) 350 – 352.
23. *European Pharmacopoeia,* 3rd ed.,Council of Europe, Strasbourg 1996.
24. L. Kabelitz, K. Reif, "Anthranoide in Sennesdrogen," *Dtsch. Apoth. Z.* **134** (1994) 5085 – 5088.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Lead

Standard Article

Charles A. Sutherland[1], Edward F. Milner[2], Robert C. Kerby[3], Herbert Teindl[4], Albert Melin[5], Hermann M. Bolt[6]

[1]Cominco Ltd., Trail, British Columbia, Canada
[2]Cominco Ltd., Trail, British Columbia, Canada
[3]Cominco Ltd., Trail, British Columbia, Canada
[4]Cominco Ltd., Trail, British Columbia, Canada
[5]Stolberg Ingenieur Beratung, Neuss, Federal Republic of Germany
[6]Institut für Arbeitsphysiologie an der Universität Dortmund, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a15_193.pub2
Article Online Posting Date: April 30, 2005

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Properties
3. Occurrence
4. Production
4.1. Ore Concentration
4.2. Smelting
4.2.1. Sintering Reduction Process
4.2.1.1. Downdraft Sintering
4.2.1.2. Updraft Sintering
4.2.1.3. Products of Sinter Roasting
4.2.2. Reduction in the Blast Furnace
4.2.2.1. Process Description
4.2.2.2. Lead Blast Furnaces
4.2.2.3. Products of the Lead Blast Furnace
4.2.2.4. Imperial Smelting Process
4.3. Roast Reaction Processes
4.4. Direct Smelting Reduction Processes
4.4.1. Air Flash Smelting
4.4.2. Oxygen Flash Smelting
4.4.2.1. Kivcet Process
4.4.2.2. Boliden Kaldo Process
4.4.2.3. Outokumpu Process
4.4.3. Air - Slag Bath Smelting, the Isasmelt Process
4.4.4. Oxygen - Slag Bath Smelting, QSL Process
4.5. Other Lead Processes
5. Refining of Lead Bullion
5.1. Pyrometallurgical Refining
5.1.1. Decoppering
5.1.2. Removal of Arsenic, Tin, and Antimony
5.1.3. Removal of Noble Metals
5.1.4. Dezincing
5.1.5. Debismuthizing

5.1.6. Final Refining and Casting of Lead
5.1.7. Processing of Intermediate Products of Pyrometallurgical Refining
5.2. Electrolytic Refining of Lead Bullion
6. Recovery of Secondary Lead from Scrap Materials
6.1. Battery Types and Composition
6.2. Battery Breaking and Processing Feed Preparation
6.3. Smelting of Battery Scrap Materials
6.4. Future Trends
7. Uses of Lead
8. Economic Aspects
8.1. Lead Concentrate Schedules
8.2. Lead Statistics
9. Toxicology and Occupational Health
9.1. Sources of Lead Exposure
9.1.1. Drinking Water
9.1.2. Atmosphere
9.1.3. Soil
9.1.4. Diet
9.1.5. Occupational Exposure
9.2. Absorption and Excretion
9.3. Effects
9.3.1. Acute Effects
9.3.2. Hematologic Effects
9.3.3. Neurological Effects
9.3.4. Renal Effects
9.3.5. Reproduction Effects
9.3.6. Carcinogenicity
9.4. Indices for Monitoring Lead Exposure
9.4.1. Lead in Blood
9.4.2. Teeth and Hair
10. Control of Lead Emissions
10.1. In-Plant Lead Emissions
10.2. Ambient Lead Emissions

[Top of Page]

1. Introduction

Charles A. Sutherland, Edward F. Milner, Albert Melin

The metal lead remains a major nonferrous commodity in modern industry as it has been from antiquity. In 1989 lead consumption in the nonsocialist world was estimated to be 4.5×10^6 t, which is well below the capacity of the primary producing and recycling plants. Concern about the toxicity of lead has brought about substitutions in some uses and proscriptions in others; consequently, the lead market is now dependent on the lead storage battery. Growth in the transportation field, principally in conventional internal combustion engine automobiles, has been the main factor in sustaining a modest growth in lead consumption of around 2.4 % in 1988. The consumption of lead in other use areas (including the drastically reduced gasoline additive market, pigments, ammunition, solders, plumbing, and shielding) amounts to less than that in the battery market. Recent discoveries in North America, Australia, South Africa, and China have brought economic lead ore reserves to a level comfortably above expected consumption.

History. The knowledge and use of lead goes back as far as 5000 B.C. to the ancient Egyptians. The metal was used extensively by the Phoenicians, Romans, Indians, and Chinese. The first lead mines in Spain were operated around 2000 B.C. The uses of the metal included weight standards, coinage (ultimately replaced by silver), statuary, sheathing and lining, trinkets, anchoring of iron rods, and making of seals. The Greeks operated lead mines at Laurium in the fifth century B.C. and the Romans mined in the Rio Tinto region of Spain in 300 B.C. The Romans used lead extensively for water piping; indeed latin " plumbum " for lead came from the word for a water spout, and from it has come the name plumber. They also used lead for soldering, in sheets for inscribed lettering, and for preparation of lead oxides. Lead uses increased in the Middle Ages with supporting mines appearing in England, Germany, Bohemia, and other parts of Europe. Lead roofing, lead " cames " in stained glass windows, guttering, drain pipes, and later type founding were additional uses that continue today. Lead received considerable attention from the alchemists. Older uses of lead as material for projectiles expanded considerably with the invention of firearms.

In the nineteenth century, the United States entered the lead production industry because of its major natural resources, and was followed by Australia and Canada. In the twentieth century the appearance of the automobile, chemical, and machine industries created large new uses for lead in gasoline antiknock additives, bearing and plumbing alloys, accumulator

batteries, and chemical equipment. Modern uses include glass making, sound attenuation, and radiation shielding.

[Top of Page]

2. Properties

Charles A. Sutherland, Edward F. Milner, Albert Melin

The element lead, Pb [7439-92-1], A_r 207.21, is a lustrous metal. When freshly cut, surfaces are bluish white but oxidize readily to the familiar dull grey color. Lead is located in group 14 of the periodic table. The element displays valences of + 2 and + 4 as indicated by its electronic configuration:

$$(1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10} 6s^2 6p^2)$$

This permits attainment of stability by donation of two or four electrons.

Lead Isotopes. Lead (atomic number 82) has four naturally occurring isotopes; mass numbers listed in order of abundance are 208 (51 – 53 %), 206 (23.5 – 27 %), 207 (20.5 – 23 %), and 204 (1.35 – 1.5 %). Variations in isotopic concentration can cause a difference of 0.15 in the normal atomic mass A_r of 207.21. Three of the natural isotopes are radioactive: ^{208}Pb is the end product of thorium decay, while ^{206}Pb and ^{207}Pb are formed from the uranium series via actinium and radium, respectively. A number of unstable radioactive isotopes exist with mass numbers of 200 – 203, 205, 209 – 214. Lead is a preferred shielding material against gamma rays and X-rays because of its high density and atomic number; if sufficiently pure, it does not become radioactive when used for shielding against neutron sources.

Physical and Mechanical Properties. The outstanding physical properties of lead are its low melting point, high density, softness, and malleability. The metal can readily be worked by rolling and extruding, and is slightly ductile. The physical and mechanical properties of lead follow:

Physical Properties:	
mp °C	327.4
Latent heat of fusion, J/g	23.4
bp °C	1741
Latent heat of vaporization, J/g	862
Vapor pressure, kPa	
980 °C	0.133
1160 °C	1.33
1420 °C	13.33
1500 °C	26.7
1600 °C	53.3
Density, g/cm^3	
20 °C	11.336
327.4 °C (solid)	11.005
327.4 °C (liquid)	10.686
400 °C	10.597
500 °C	10.447
Mean specific heat, J g^{-1} K^{-1}	
– 200 – 0 °C	0.1202
0 – 100 °C	0.131
0 – 200 °C	0.134
0 – 300 °C	0.136
Specific electrical resistance, Ω/cm	
20 °C	20.65
100 °C	27.02
200 °C	36.48
300 °C	47.94
Thermal conductivity, W m^{-1} K^{-1}	
– 100 °C	0.371

0 °C	0.355
100 °C	0.342
200 °C	0.329
300 °C	0.316
327.4 °C (solid)	0.313
327.4 °C (liquid)	0.155
Surface tension of liquid lead, mN/m	
327.4 °C	444
350 °C	442
400 °C	438
500 °C	431
600 °C	424
800 °C	410
Dynamic viscosity of liquid lead, mPa · s	
327.4 °C	2.75
350 °C	2.60
400 °C	2.34
450 °C	2.12
500 °C	1.96
550 °C	1.70
Magnetic susceptibility (20 °C), H m^{-1} kg^{-1}	-1.5×10^{-9}
Coefficient of linear expansion, K^{-1}	
at 20 °C	29.1×10^{-6}
mean 20 – 300 °C	31.3×10^{-6}
Mechanical Properties:	
Brinell hardness (20 °C)	2.5 – 3.0
Transverse contraction coefficient (20 °C)	0.45
Young's modulus, GPa	16.5
Elongation (in 5- cm gauge length), %	50 – 60
Velocity of sound in lead, m/s	1227
Tensile strength, kPa	
– 100 °C	42 000
20 °C	14 000
150 °C	5 000

Chemical Properties. Fresh cut or cast lead surfaces undergo oxidation and tarnish rapidly to form an insoluble protective layer of basic lead carbonate. Lead is attacked by water in the presence of oxygen, but again an insoluble carbonate or silicate protective layer is formed if dissolved carbonates or silicates are present in the water. This resistance of the metal to corrosion by air, water, and soil has been exploited in the historical uses of lead in plumbing and piping. Lead might be expected to dissolve in acids with liberation of hydrogen, but this is prevented by the high hydrogen overvoltage, and attack occurs only when the combination of oxidizing conditions and soluble salt species exists. Thus the metal resists hydrochloric acid corrosion to quite high concentrations, sulfuric acid to 60 °Bé (ca. 13.3 mol/L), and hydrofluoric acid totally; however, it dissolves readily in warm, dilute nitric acid to form soluble lead nitrate. Lead is also corroded by weak organic acids, such as acetic or tartaric acid, in the presence of oxygen. The protective coating formed by oxidation is removed by alkali with formation of soluble plumbites and plumbates.

Crystal Structure. Lead crystals are face-centered and cubic with the following dimensions:

Atomic radius, nm	0.175

Atoms per unit cell	4
Lattice constant (edge of unit cell), nm	0.49389

Electrochemical Properties. The electrochemical properties of lead follow:

Electrochemical equivalent, g A^{-1} h^{-1}	3.865
Ionic radius Pb^{2+}, nm	0.118
Pb^{4+}, nm	0.070
Metallic radius, nm	0.17502
Covalent radius (*sp* ^{3}), nm	0.144
Ionization potentials, eV	
1st	7.415
2nd	14.97
Standard oxidation potentials (25 °C), V	
$Pb \longrightarrow Pb^{2+} + 2e^-$	0.126
$Pb^{2+} + 2\,H_2O \longrightarrow PbO_2 + 4\,H_2 + 2\,e^-$	– 1.456
$Pb + 2\,OH^- \longrightarrow PbO + H_2O + 2\,e^-$	0.576
$Pb + 3\,OH^- \rightarrow HPbO_2^- + H_2O + 2\,e^-$	0.54

Lead forms two series of compounds corresponding to valence states of + 2 and + 4; lead (II) compounds are electrovalent (ionic), while those of lead (IV) are essentially covalent. The metal is amphoteric and forms lead salts, plumbites, and plumbates. When melted in air, lead oxidizes readily to its monoxide PbO (also called litharge or simply lead oxide); oxidation (drossing) can be promoted or inhibited by dissolved metallic impurities. Other lead oxides are the dioxide PbO_2 (actually sesquioxide Pb_2O_4) and red lead Pb_3O_4 (the orthoplumbate $Pb \cdot Pb_2O_4$). All the lead(III) halides are known, and are considerably more soluble in hot water than in cold; they also form double chlorides with alkali metals and a series of oxyhalides, the oxychlorides being the most familiar. Other common compounds of lead are the soluble nitrate and acetate; lead chromate (chrome yellow) used as a pigment along with red lead; lead borate used along with oxides in glass making; lead sulfide (galena), which is the principal naturally occurring form of the metal; and lead sulfate that is readily formed by sulfating a soluble lead salt and by natural oxidation of galena. Lead compounds are described in detail elswhere, ⟶ .

[Top of Page]

3. Occurrence

Charles A. Sutherland, Edward F. Milner, Albert Melin

Lead, like copper and zinc, is a markedly chalcophilic element. Its mean abundance in the Earth's crust is 16 g/t (0.0016 %). Enrichments occur only in the pneumatolytic and especially the hydrothermal sulfide phase in the form of vein (lode), impregnated, and particularly metasomatic deposits. Sedimentary enrichments are of little importance.

Lead Minerals. Galena or lead sulfide [1314-87-0], PbS, theoretically 86.6 % Pb, is by far the most important lead mineral and the sole primary one. Other common forms of naturally occurring lead are usually associated with or derived from galena deposits. They include mixed sulfides such as jamesonite, $Pb_4(Sb,Fe)_7S_{14}$ (ca. 40 % Pb) and boulangerite, $Pb_5Sb_4S_{11}$ (ca. 58 % Pb); anglesite, $PbSO_4$ (68.3 %) Pb), a product of galena oxidation; cerussite or white lead ore, $PbCO_3$ (77.5 % Pb), a decomposition product of galena; pyromorphite, also qualified as green, brown or variegated lead ore, $Pb_5(PO_4)_3Cl$ (76 % Pb); vanadinite, $Pb_5(VO_4)_3Cl$; wulfenite or yellow lead ore, $PbMoO_4$ (56 % Pb); crocoite, $PbCrO_4$; and phosgenite or horn lead, $Pb_2Cl_2CO_3$. Frequent impurities in lead minerals are zinc, copper, arsenic, tin, anitmony, silver, gold, and bismuth.

Deposits. The most important mineable lead ore is galena, which occurs chiefly in association with other minerals, mainly zinc ores. Lead deposits can be classified as follows:

1. Hydrothermal vein, impregnation, and replacement deposits (e.g., Broken Hill, Australia; Anvil Mine, Canada; and numerous deposits of the Cordilleran – Andean Orogeny);
2. Volcanogenic sedimentary deposits, some occurring as massive bodies of rich ore, others as impregnations (e.g., Sullivan Mine, Canada; Rammelsberg, Federal Republic of Germany; Mount Isa and Hellyer, Australia);
3. Hydrothermal or marine sedimentary deposits, typically in calcareous or dolomitic formations of the North American and European platforms (e.g., Pine Point, Canada; the New Lead Belt in the Missouri district, United States; Red Dog, Alaska).

In world mining output, mixed lead – zinc ore deposits are most important and account for ca. 70 % of total production of both metals. In second place are deposits that contain predominantly lead ores (ca. 20 % of total production), and the remainder (ca. 10 %) is obtained as a byproduct from zinc, copper – zinc, and other deposits.

An important byproduct of lead ore production is silver, which is incorporated into the lattice of galena, and is recovered with the lead when concentrates are smelted (⟶ Silver, Silver Compounds, and Silver Alloys). Some 70 % of world silver production comes from lead concentrates. Valuable amounts of gold are also contained in some lead deposits. Another byproduct from lead ore is copper, but the amount is a very small part of the total world copper output.

Reserves. Total world lead reserves were reported in 1987 as 75×10^6 t lead content (Table 1). Western countries owned about 73 % of the total.

Table 1. World lead ore reserves (millions of tonnes contained lead)

Country	Reserves [b]	Reserve base [c]
Australia	16	28
United States	11	22
Canada	8	15
Republic of South Africa	4	6
Mexico	3	4
Peru	2	3
former Yugoslavia	2	3
Other market economy countries	9	14
Centrally planned economies	20	30
Total	75	125

[a] U.S. Bureau of Mines, Division of Mineral Commodities, Jan. 1989.
[b] Estimated economically recoverable lead.
[c] Reserves augmented by marginally economic ores, some currently uneconomic but with potential, and also some provision made for determination of additional reserves in known ore bodies.

The United States has extensive lead deposits in the southeast of Missouri (New Lead Belt). These deposits and the new Red Dog lead – zinc orebody in Alaska are located in bedded marine-sedimentary or hydrothermal deposits. Further reserves are found in the Rocky Mountains (Idaho, Utah, Colorado) in the form of hydrothermal vein and replacement deposits. Canadian reserves are also substantial but of generally lower grade, and are located in the Appalachian region of New Brunswick, the Rocky Mountain trench, and the Yukon. North and South America, including the Latin American countries, hold about 35 % of known reserves.

Australia has the most substantial lead resources amounting to about 21 % of the world total; they are concentrated in Mount Isa (Queensland), Broken Hill (New South Wales), and Tasmania.

The remaining western reserves are found in many places in Europe, Africa, and Southern and Eastern Asia. The largest known European reserves are in former Yugoslavia.

Most Eastern Bloc reserves (ca. 27 % of world reserves) lie in the former Soviet Union (Kazakhstan and the Altai and Karatau regions). Only little information is available on these reserves and recent discoveries in the People's Republic of China.

Data on exploitable reserves are only valid in the short-term, because they are subject to long-term economic fluctuations. The growth of world reserves of lead ore through new discoveries can be expected to outstrip the increase in primary metal production. Recycling of lead, already a major source of the metal, particularly for battery making, can be expected to suppress production from primary sources.

[Top of Page]

4. Production

Charles A. Sutherland, Edward F. Milner, Albert Melin

4.1. Ore Concentration

The most common feed for lead production is sulfidic lead concentrate, which contains an average of 50 – 60 % lead (occasionally as high as 70 – 80 % or as low as 30 %). Oxide lead concentrates are of secondary importance.

Lead ores in deposits are usually intergrown with other minerals and with host rock. These crude ores are not smeltable and must first be beneficiated into concentrates of much higher lead content but with minimum loss of lead.

Ore beneficiation normally includes crushing, dense-medium separation, grinding, froth flotation, and drying of concentrate. Lead flotation is usually the first step in the separation of lead – zinc and lead – zinc – copper ores.

4.2. Smelting

The major process for the production of primary lead from a sulfide concentrate is the sinter oxidation – blast furnace reduction route. Older processes involving direct oxidation of lead sulfide to lead or the roast reaction between lead sulfide and the oxidation products lead oxide and lead sulfate are now of little importance. In the last two decades new oxygen

metallurgy processes featuring sulfide oxidation in a flash flame or by oxygen injection into a slag bath, followed by reduction of the lead oxide slag, have advanced to industrial application.

In the Imperial Smelting Furnace (ISF) process, lead is produced simultaneously with zinc. This process will not be considered here, for details see ⟶ Zinc.

4.2.1. *Sintering Reduction Process*

The sintering reduction process involves two steps: (1) a sintering oxidative roast to remove sulfur with production of PbO and (2) blast furnace reduction of the sinter product.

The objective of sintering lead concentrates is to remove as much sulfur as possible from the galena and the accompanying iron, zinc, and copper sulfides, while producing lump agglomerate of adequate strength combined with porosity for subsequent reduction in the blast furnace. The sulfur dioxide content of the sintering gases is often recovered and used in sulfuric acid production. The chemistry of the roast sintering process is complex because it involves many oxygen- and sulfur-containing lead compounds.

The Pb – S – O System. The Pb – S – O system was investigated comprehensively by KELLOGG and BASU [7] whose results were later confirmed by many investigators. WILLIS provides an excellent summary of the physical chemistry of the oxidation of lead sulfide [8]. Lead in sinter occurs mainly as lead monoxide or lead silicate. The temperature of the sinter charge must be high enough to attain the area of PbO predominance (i.e., > 950 °C, see Fig. 1).

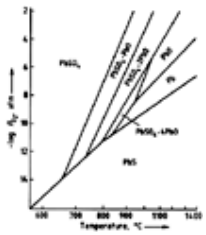

Figure 1. Phase relations in the Pb – S – O ternary system with air as the oxidizer and a partial sulfur dioxide pressure of 0.2 atm [9]

* 1 atm = 0.1013 MPa

Oxidation starts with the formation of lead sulfate, which is the stable reaction product at the oxygen and sulfur dioxide partial pressures and temperatures commonly used in sintering. Lead sulfate then reacts further with lead sulfide, decomposing to increasingly basic sulfates and ultimately to lead monoxide. Lead in dust carried out by the sintering gas is totally sulfated.

The composition of the sinter is important as regards its behavior in the blast furnace. The lead monoxide and lead silicate which account for most of the lead content are readily fusible and reducible. The integrity of the sinter thus depends on the presence and quantity of other constituents, typically silicates of zinc, iron, and calcium, and ferrites. These substances are much more refractory than the lead compounds and provide a framework that keeps the sinter lump from disintegrating after the lead compounds have been removed.

Sinter Roasting Techniques. The development of sinter roasting goes back to work by HUNTINGTON and HEBERLEIN (1896), who proposed that lead ores mixed with limestone chips be roasted in a converter with an air blast (blast roasting). Implementation of this idea became the basis of the modern sintering technique, which combines roasting with agglomeration of the charge in the sintering step.

A continuous sintering machine with a sintering bell was devised in 1905 – 1908 by DWIGHT and LLOYD. At about the same time, VON SCHLIPPENBACH introduced the circular sintering machine.

A distinction is made between *downdraft sinter roasting*, where the air required for roasting is drawn downwards through the sinter mixture, and *updraft sintering*, in which the air is blown upwards through the bed. Updraft sintering is preferred because it produces lower gas volumes of higher sulfur dioxide content that are suitable for sulfuric acid production.

4.2.1.1. *Downdraft Sintering*

The original downdraft process proposed by DWIGHT and LLOYD was the conventional technique for the sinter roasting of lead ores until about 1955. It has now been superceded by updraft sintering. In downdraft sintering the nodulized charge blend is fed continuously onto a traveling grate, and sulfides are ignited on the surface of the charge bed by burners. As the ignited charge moves away from the ignition zone, air is drawn downward through the charge. The sulfur is burned on a downward advancing front, and incipient fusion of solids agglomerates the nodules of desulfurized charge into a porous sinter product.

A characteristic of downdraft sintering of lead ores is the low sulfur dioxide concentration of the roasting gases (Fig. 2). Since off-gases containing 1 – 2 vol % sulfur dioxide occur at the end of the sintering machine and cannot generally be discharged into the atmosphere, recirculation of the gas or chemical scrubbing is necessary. Even with gas recirculation, the production of acid-strength gas causes problems due to the formation of liquid phases during sintering that reduce the gas permeability of the bed, and in particular to the metallic lead formed by the reaction of lead sulfide with lead sulfate. The molten lead flows downward through the sinter bed along with the gas stream and forms a very tight layer immediately above the grate. Part of the lead also drops through the grate into the wind boxes. Removal of the lead from the wind boxes is expensive and causes long downtimes. Water-bath quenching of wind-box lead and conveyor belt cleaning have been practised to overcome this problem. With downdraft sintering the maximum lead content of the sinter mix is limited to about 40 % to avoid excessive production of wind-box lead.

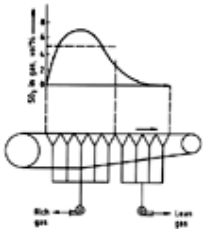

Figure 2. Variation in sulfur dioxide content of sintering gases along a downdraft sintering machine

4.2.1.2. *Updraft Sintering*

Principles. The difficulties encountered in downdraft sintering are avoided in updraft sintering in which sintering takes place in an ascending gas stream (Fig. 3). Roasting air is supplied to the machine from below through the wind boxes, and the gas flows upward through the sinter bed. To initiate roasting, a 2 – 3 cm layer of feed is placed on the grate and ignited in a downdraft ignition box (Fig. 3 A). The main sinter bed (d) is then distributed onto the burning ignition layer (b), the direction of gas flow is reversed (Fig. 3 B), and air is forced upward through the sinter bed. The main layer is ignited from the first layer, and sinter roasting advances upward. To prevent caking of the ignition layer on the grate bars during downdraft ignition, the grate is usually covered with a hearth layer of return sinter (c) before the ignition layer is placed.

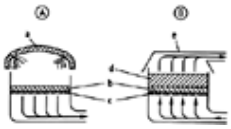

Figure 3. Updraft sintering

A) Downdraft ignition; B) Sinteringa) Ignition hood; b) Ignition layer; c) Hearth layer; d) Sinter bed; e) Exhaust hood

Since cold air flows into the sinter bed from below, any metallic lead that forms is immediately cooled and solidifies while still in the bed. This makes for good permeability of the bed, lower gas flow, and uniform sintering. Higher sulfur dioxide levels in the roast gases can thus be achieved, and sinter charges containing up to 50 % lead can be managed.

Plant Design. The development of updraft sintering began in 1950 – 1955 at the Port Pirie lead smelter in Australia and at Lurgi Gesellschaft für Chemie und Hüttenwesen, Federal Republic of Germany. The process has since been adopted worldwide. All constructed sintering plants have been designed for updraft sintering and a large number of downdraft plants have been reconstructed. Updraft sintering became an essential component of the Imperial Smelting zinc process because of the exacting requirements for satisfactory operation (i.e., low sulfur and high sinter hardness).

The flowsheet of a plant of modern design is shown in Figure 4. Developments have concerned not only increases in the size of updraft sintering machines, but also improvements in capacity, automation, and of working conditions. Dust-collection systems and gas-cleaning systems must comply with increasingly strict regulations on air and water pollution control.

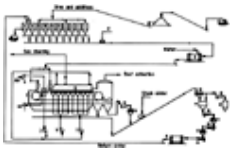

Figure 4. Flowsheet of an updraft sinter plant for lead ores

a) Return sinter bins; b) Lead concentrate bins; c) Flux and other charge material bins (silica, limestone, leach residues, etc.); d) Magnet; e) Mixing – nodulizing drum; f) Sinter machine; g) Sinter grate; h) Ignition air blower; i) Ventilation air blower; j) Ignition exhaust fan; k) Sintering air blower; l) Sinter gas recycle blower; m) Sinter screens; n) Roll crushers; o) Undersize sinter bin; p) Product sinter bin; q) Return fines cooling drum

An updraft sintering plant comprises many devices and machines, which employ a logical system of electrical interfaces. Precise control of process performance is difficult, primarily because lag times from entry of the feed to appearance of the product from the sinter strand are long and sampling of the feed and product sinter are difficult. Consistent feed preparation, either by means of large-scale bedding operations or very accurate weigh feeders for the concentrates and fluxes, can provide a high precision of feed-forward control.

Feed Mixture. A charge moisture content of 4.5 – 5.5 % is of utmost importance for achieving good bed permeability during combustion. The desired moisture content of the feed mixture is attained by adding water to the tumbling and mixing drum. Two drums usually operate continuously to achieve good control. If feed material moisture is too high, charge drying is required.

Crushed recycle sinter is required to control the fuel content of the charge and to enhance bed permeability. Cooling of recycle sinter is necessary for mechanical reasons and to minimize dust production. The main difficulties encountered in return sinter cooling are caused by irregular discharge of the sinter strand combined with the handling of moist ventilation gases produced in water quench cooling drums.

Because sinter quality and plant capacity are strongly affected by the physical quality of the recycled sinter, comminution of the returns is very important. The desired size distribution of the lump sinter that is to be reduced in the blast furnace is obtained by two-stage crushing and screening. The returns needed for the sinter charge are also broken down to the requisite particle size (normally 6 mm) by using roll crushers in a two-stage process.

Dust Collection. Dust collection is a major problem in updraft sintering plants. The plant comprises many machines and devices, some of which must transport, screen, and crush hot material; lead-containing dust is produced at all transfer points. Proper ventilation is essential to provide satisfactory in-plant hygiene, and efficient cleaning of exhaust gas is required to

control lead emissions. High volumes of ventilation air have to be handled at considerable extra cost. Dust recovered from baghouses and wet scrubbers must be recycled to the process.

Because sintering plants emit both dry and wet dust-laden gases, gas cleaning design has been constantly improved. For a detailed description of dust collection see ⟶ Dust Separation. The most recent plants employ a combination of wet scrubbing and dry baghouse gas cleaning. The cooling drum ventilation and the hot gases from the sintering machine are cleaned in water scrubbers. Discharge gas temperatures are too high for bag filters. The other dust-collection points (crushing and screening) are combined in a common baghouse. Dust recovered from baghouses is added to the thickener slime that is supplied as coolant to the cooling drum. The dust load in the process sulfur dioxide from the sintering machine is usually collected in high-temperature electrostatic precipitators, moistened, and added directly to the sinter mixture.

Recovery of Sulfur. For optimal operation of the sintering machine, some 15 – 20 $m^3\ m^{-2}\ min^{-1}$ of air (at STP) must be delivered along its whole length. The combined exhaust gases have an average sulfur dioxide content of 4 – 5 vol %. However, since peak sulfur dioxide levels of over 10 % occur at the front end of the sintering machine, a rich gas stream with a relatively high sulfur dioxide content can be separated and the balance tail gas is recycled to the front-end wind boxes. In this way, a rich gas containing 6 – 6.5 vol % sulfur dioxide can be delivered to a sulfuric acid plant. The technique involves a certain deliberate loss of capacity resulting from the injection of hot recycled lean gas at 300 – 350 °C.

Size and Capacity of Sintering Plant. Standard sizes have been adopted for updraft sintering machines. Size is defined chiefly by smelting capacity and by the quantity of sulfur to be combusted.

Roasting capacities of 1.3 – 1.5 t $d^{-1}\ m^{-2}$ are achieved with modern updraft sintering and are over 50 % higher than those of older downdraft systems. For normal lead sintering an annual lead production capacity of 70 000 – 100 000 t requires a sinter machine of 50 – 70 m^2 updraft area. Machines up to 90 m^2 are employed in Imperial Smelting process plants. These rates and sizes are material-dependent and must be confirmed by tests for each proposed charge. Data for the updraft sinter plant at Port Pirie, Australia are listed below:

Number of sinter machines	1
Roasting surface	
total, m^2	70.5
effective, m^2	62.3
Sinter production, t $m^{-2}\ d^{-1}$	20.1
Throughput, t/d	4500
Feed blend, %	
new materials	36
return sinter	64
Sulfur content, %	
concentrates	16 – 27
new materials	11
return sinter	2.5
total feed blend	5.5
product sinter	1.8
Lead in sinter, %	45
Thickness of ignition layer, mm	35
Total bed thickness, mm	300
Strand speed, m/min	1.42
Ignition method	downdraft
Ignition burners	
number	2
fuel	natural gas
usage, m^3/h	30
Wind boxes	
number	9
bed area, m^2	4.9 (no. 1)
	8.2 (nos. 2 – 9)
pressure, kPa	5.0 (no. 1)
	4.5 (no. 2)

	3.5 (no. 3)
	2.0 (nos. 4 – 9)
maximum air flow, $m^3 min^{-1} m^{-2}$	38.8
SO_2 concentration, vol %	4 – 6
Sulfuric acid production	yes

Sintering for the Imperial Smelting Process. Lead – zinc concentrates or mixtures of lead and zinc concentrates for the Imperial Smelting Furnace (ISF) process must be desulfurized and sintered before reduction in the zinc vaporizing blast furnace. The sinter must incorporate recycled intermediate products from the sinter plant and blast furnace and also the fluxes needed to obtain the desired slag balance. The sinter must be chemically uniform, well desulfurized, as well as hard and abrasion-resistant to avoid formation of fines in the furnace. It must also have a high porosity and a narrow particle-size range.

These requirements can be satisfied only by carefully controlled updraft sintering with feed sulfur held between 5 and 6.5 %, moisture at 4 – 7 % and return sinter sizing between 1 and 5 mm. For further details, see ⟶ Zinc.

4.2.1.3. *Products of Sinter Roasting*

Sinter. Product sinter should be agglomerated in large lumps that are strong enough to prevent decrepitation during handling. The sinter should have a cellular, porous structure without pockets of unroasted ore or very dense regions.

Sintering Gases. The sulfur dioxide concentration in the sintering gases is lower than that found in the roasting of zinc or copper ores, especially when high-lead concentrates containing as little as 14 – 16 % sulfur are processed. In the sintering of lead concentrates, the sulfur dioxide level fluctuates between 4 and 6 vol % under favorable conditions and with gas recycling. Release of the lean gases into the atmosphere is unacceptable, although this is still done in a few instances. The roasting gases contain a small amount of sulfur trioxide; their dust content is usually 2 – 5 g/m^3, but may reach 20 g/m^3 in exceptional cases. Extensive cleaning is always necessary if the gas is to be processed into sulfuric acid. Pollution control laws require that metal losses be kept to a minimum. The typical gas cleaning train consists of dry electrostatic precipitators, a wet scrubber, followed by wet electrostatic cleaning.

The cleaned sintering gases are generally forwarded to sulfuric acid production. They are sometimes processed into liquid sulfur dioxide after absorption in an organic solvent. In rare cases they are scrubbed with ammonia to produce ammonium sulfate and high-strength sulfur dioxide for recycle or liquid sulfur dioxide production.

Sulfuric acid derived from lead-sintering plants is produced as " black acid ". Its black color is due to carbon particles from the volatile organic constituents of flotation reagents. The color can be removed with hydrogen peroxide at additional expense.

Flue Dusts. Flue dusts from lead sintering plants contain 60 – 70 % lead, about 10 % sulfur, and varying amounts of zinc, cadmium, arsenic, antimony, and mercury. The lead is predominantly in the form of lead sulfate or basic lead sulfate, with a minor quantity of lead sulfide. Flue dusts are recycled to the sinter blend.

4.2.2. *Reduction in the Blast Furnace*

The second part of the roast – reduction process is carried out in a blast furnace. Where the lead content of the sinter (mainly lead oxide) is reduced to metallic lead, other metals such as copper, antimony, arsenic, and noble metals are also produced. Other constituents (*gangues*) contained in the sinter and the flux are carried off as silicate slag. In lead blast-furnace smelting, zinc remains in the slag, but in the Imperial Smelting process the furnace top is hot and gaseous zinc leaves the furnace along with the off-gases and is then condensed to liquid zinc metal.

The charge to the blast furnace comprises

1. Sinter, which incorporates the roasted concentrate and fluxes
2. Other oxygen-containing lead materials such as oxides and silicates
3. Metallurgical lump coke as the reducing and heating fuel

4.2.2.1. *Process Description*

The lead blast furnace is a countercurrent reactor in which a charge (sinter and coke) moves through a vertical shaft in countercurrent to the ascending gas flow. The descending charge successively passes through the preheating zone (to ca. 200 °C), the reaction or reduction zone (to ca. 900 °C), the melting zone (to ca. 1150 °C), and finally the combustion zone (tuyere zone). The liquid reaction products collect in the furnace crucible, which is located below the tuyere zone.

Principles. In the combustion zone, atmospheric oxygen blown in through the tuyeres reacts with the coke to form carbon dioxide with extensive release of heat. This hot gas ascends through a coke-rich layer (arch) and is reduced to carbon monoxide, the main reductant. In the melting zone and in the overlying reduction zone, heat is transferred from the gas and liquid slag is formed from the fluxes and sinter gangue. Most of the metal-forming reactions take place in the softening-melting zone, even though the residence time in this zone is relatively short. Before fusion, heterogeneous gas – solid reactions and heat transfer processes take place. Free moisture and water of crystallization are driven off in the preheating zone.

The rates of coke combustion and conversion of carbon dioxide to carbon monoxide (Boudouard reaction equilibrium) must

be in accordance with the rate of reduction in the reaction zone and with the rate of heat transfer. If the combustion rate exceeds the rate of heat transfer to the charge, the coke burns more rapidly than the charge melts causing the melting zone to rise in the furnace. If the combustion rate is lower than the rate of heat transfer, the coke burns so slowly that it prevents the charge from descending and slows down the performance of the furnace. If the rates of combustion and heat transport are in good agreement, the melting zone and thus the plastic region are narrow. The charge material then flows through at a uniform speed and gas channeling and blow holes are minimized.

For optimal heat and mass transport in the reaction zone, gas velocity should be high and the resistance to gas flow low. The physical and chemical characteristics of the sinter should ensure adequate solids integrity in the reaction zone. Finally, the blast furnace charge should be as physically homogeneous as possible.

Behavior of Sinter Components. The exothermic reduction of lead oxide by carbon, carbon monoxide, and hydrogen begins at relatively low temperature. Other oxygen-containing lead species (e.g., aluminates, silicates, and ferrites) require more intense reducing conditions. Lead sulfate and the basic lead sulfates are converted chiefly by reaction with silicates, but they can also react with sulfides to yield metallic lead. Lead sulfide partly reacts with sulfates to form metallic lead.

The noble metals are largely dissolved in the lead product (*bullion*), with small amounts distributed to sulfide matte and slag.

The copper content of the sinter is captured in the matte in sulfide form, provided sufficient sulfur is present. If, however, the sulfur level is insufficient, the copper is reduced and goes into solution in the lead bullion. A small portion of the copper is always incorporated into the slag in either oxide or sulfide form.

Zinc oxide is located in the slag, and to keep it in solution the slag must be rich in iron(II) oxide. Zinc levels in slag exceeding 18 – 20 % cause slag viscosity problems and some hearth accretions. Part of the zinc oxide is reduced to zinc vapor above the tuyere zone, but the vapor is immediately reoxidized to zinc oxide in the colder zone (zinc circulation in the blast furnace). Zinc sulfide, charged in the sinter or formed from metallic zinc, is notorious because it promotes the formation of accretions on the walls of the smelting shaft. If a matte is formed, it can dissolve a certain amount of the zinc sulfide, but then its melting point rises and the physical separation of the slag becomes more difficult.

Antimony and arsenic compounds are reduced to the respective metals which dissolve in the lead bullion. If the amount of arsenic is too large to be absorbed by the matte, and if copper, iron, and nickel are also present, metal arsenides (*speisses*) may be formed. Some arsenic is also volatilized and some is slagged. Tin is partly reduced into the lead bullion.

Separation of the fused products leaving the furnace hearth is based on their low mutual solubility and their differing densities (lead bullion about 10.5, matte 4.5 – 5.0, slag 3.5 – 3.8 g/cm^3).

4.2.2.2. *Lead Blast Furnaces*

Lead blast furnaces are classed in two groups: round (*mushroom furnaces*) and rectangular (*Rachette furnaces*). The round design is only used in special low-capacity operations. The Rachette furnace is the standard type of lead blast furnace, and often has rounded corners. The *chair-jacket furnace* is a modification of the Rachette design and has two rows of tuyeres, one above the other. Figure 5 shows two conventional blast-furnace profiles.

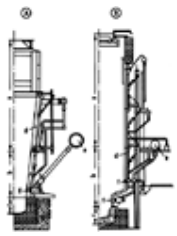

Figure 5. Blast furnace profiles

A) Single-level tuyere furnace; B) Double-level tuyere furnacea) Upper shaft (preheating); b) Lower shaft (melting, reduction, tuyere zone); c) Crucible (slag, lead); d) Shaft and tuyere water jackets; e) Blast air bustle pipe; f) Blast tuyeres

A typical lead blast furnace has a rectangular cross section 5 – 8.5 m above the tuyeres, a water-cooled melting zone, and a syphon tapper for continuous lead discharge. For a given charge, furnace capacity can be increased only by enlarging the cross section. The area of the tuyere zone is the controlling factor. The world's largest furnace, employed by the Broken Hill Associated Smelters, Port Pirie, Australia, has a tuyere zone length of 10 m. Structural and operating data for this furnace are listed below:

Number of furnaces	1
Type	Rachette type with chair jackets
Furnace feed	
sinter, t/d	1250
total, t/d	1415
coke as % of sinter	9.5
Bullion production, t/d	550
Blast air	

rate, m^3/min	510
pressure, kPa	120
Hearth area	
lower, m^2	17.13
upper, m^2	30.34
Number of tuyeres	
sides	81
ends	10
Tuyere diameter, mm	122
Nozzle diameter, mm	65
Elevation of bottom tuyeres above hearth, mm	742
Furnace coke	
fixed carbon, %	82
carbon combustion rate, t m^{-2} d^{-1}	3.3
Top gases, m^3 m^{-2} min^{-1} (STP)	16.2
Type of lead tapper	Asarco – Roy

The tuyere zone width can be increased only as long as the blast can penetrate from both sides, generally to around 1.5 m. The furnace can be widened if two offset rows of tuyeres are placed one above the other such that the center (“ dead zone ”) of the charge is brought to the melting point in the narrower bottom tuyere zone. Since the penetration depth of the blast in the furnace permits a maximum width of 1.8 m in the bottom tuyere zone, the top tuyere zone has a maximum width of 3.6 m. Widths of 1.5 and 3 m have been achieved in practice. The cooling jackets have the shape of a chair, which gives this configuration its name (see Fig. 5 B).

The furnace shaft was formerly constructed of refractory bricks, but in modern furnaces a series of supported water jackets form the shaft, which may be either vertical (Fig. 5 B) or taper toward the bottom (Fig. 5 A). The hearth is subjected to the highest temperature and is cooled by one or more rows of replaceable cooling jackets set one on top of another.

The jackets are cooled with water that is either pumped in a closed cycle or circulated by the thermosiphon principle (i.e., natural convection). Jacket cooling by evaporative cooling (steam generation) is practised in the former Soviet Union. The water jackets have openings for the tuyeres through which the blast enters. Combustion air must be at a pressure between 5 and 20 kPa. The resistance of the furnace depends not only on the quantity of blast injected but also on the permeability of the charge. Blowers whose delivery is variable but independent of the furnace back-pressure are most suitable. These requirements are best met by positive blowers with a variable speed, but equipment wear often dictates their replacement by centrifugal blowers.

Tests at several lead smelters have demonstrated that enrichment of the blast with oxygen increases throughput (ca. 5 % increase for each additional 1 % oxygen in the blast). When oxygen replaces part of the blast air, a higher combustion zone temperature is achieved, gas flow per unit of lead is reduced, and the furnace operates more smoothly. Benefits include increased production, a reduction in coke consumption, lower off-gas temperatures, and lower dust generation. Similar results have been achieved with preheated blast air. Whether oxygen enrichment or blast preheating is preferable depends only on economics. Oxygen enrichment has become more popular, especially since low-cost oxygen production has become increasingly common. However, as yet the degree of enrichment of the lead furnace blast is very modest: only 2 – 3 %.

A crucible located below the tuyeres serves to separate the liquid products and collect the lead. In large production units, lead is withdrawn continuously, and in smaller ones tapped at intervals. If the lead bullion has a high copper content, the continuous syphon tapper runs into accretion problems. It should be replaced by a discontinuous tapper placed just above the bottom of the crucible, so that it is always covered with molten lead to protect it against deposits. With the Asarco (Roy) continuous tapper the liquid phases are separated outside the furnace [10]. Lead and slag are discharged together into an external settler attached to the furnace, in which the liquid phases are separated. The Asarco tapper gives higher capacities with lower investment and maintenance costs, along with more hygienic workplace conditions. It has the drawback of a higher lead level in the slag. North American smelters employ the Asarco tapper; some European smelters use the Arents syphon tapper.

If a separate matte phase is formed, slag and matte can be separated in a heated launder, where the matte with its higher relative density settles and the slag floats. If the slag is to be discarded or kilned for zinc recovery, it is usually granulated. It may, however, also be transported in the molten state to a zinc fuming furnace.

The charging of blast furnaces is largely automated. Charging devices for open-top furnaces may be bucket elevators and skips or a variable discharge conveyor belt. Tipping or bottom dump charge buckets are used for the more common closed-top furnaces. Various alternating charging techniques are employed to achieve layers or blends of coke and sinter in the shaft. Blast-furnace top gases are drawn off the furnace top by an exhaust fan and passed through a baghouse for dust

removal.

4.2.2.3. *Products of the Lead Blast Furnace*

The lead blast furnace yields the following products:

1. *Lead bullion,* which must be further refined (see Chap. Refining of Lead Bullion).
2. *Slag*, which may contain large amounts of zinc and must then be further processed.
3. *Flue dusts and baghouse dusts,* which consist of fumes and dust particles carried over by the furnace top gases. Cadmium tends to accumulate in this dust, and sometimes a special cadmium separation leach is carried out before the dust is recycled to sintering.
4. *Top gas.* Blast furnace top gas usually contains 3 – 4 % carbon monoxide, which represents an unavoidable energy loss. With normal heat exchange between gas and solids in the shaft the top gas is too cold for practical heat recovery, so after dust removal it is released to atmosphere.
5. *Matte,* a sulfidic product of blast-furnace smelting, forms only if the burden contains enough sulfur. In general, provided the charge does not contain too much copper or sulfur, matte phase formation is avoided with the copper remaining in solution in the lead bullion and slag. Subsequent reverberatory melting of lead bullion drosses yields copper matte, lead metal, and slag. The copper matte can contain varying amounts of lead, iron, zinc, and noble metals. Copper matte is usually sold to special copper smelters, but processing of matte to copper sulfate or copper metal is becoming more common.
6. *Speiss* is only formed when the burden contains amounts of arsenic, nickel, cobalt, and antimony that are so large that they can no longer be taken up by the matte. Speiss is undesirable, because it dissolves large quantities of noble metals and also requires expensive refining.

Composition and Processing of Slag. The slag consists of gangue, fluxes, and the zinc contained in the charge. It also contains dissolved and suspended metals (lead, copper, silver), which are a potential economic loss. The chemical and physical properties of the slag have a considerable effect on furnace throughput, product separation, metal recovery, and coke consumption in the furnace. Lead blast furnace slag should have a low melting point, be adequately fluid (low viscosity) at the furnace bottom temperature [11], and have minimal density. Slags are compounded to attain these properties by fluxing to selected component ratios. Thus, in the ternary system CaO – FeO – SiO_2 almost all lead blast furnace slag compositions lie in an area that is bounded by weight ratios of 1 : 1 to 4 : 1 for SiO_2: CaO and 1 : 2 to 3 : 2 for SiO_2: FeO (Fig. 6). This is further illustrated in Figure 7 for a number of North American and Australian lead blast furnaces.

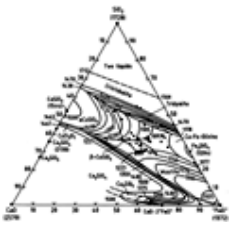

Figure 6. The FeO – SiO_2 – CaO ternary system, showing compositions of several lead blast furnace slags (•) [12]

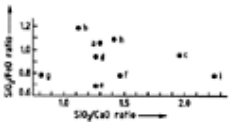

Figure 7. SiO_2 : FeO and SiO_2: CaO weight percentage ratios in blast furnace slags [13]

a) Asarco, East Helena; b) Asarco, El Paso; c) Asarco Glover; d) Amax; e) Brunswick M & S; f) Cominco; g) Mount Isa; h) BHAS, Port Pine; i) St. Joe Lead

Lead blast-furnace slag often contains sufficient zinc to make recovery profitable by rotary kiln or submerged combustion slag fuming. In both cases zinc oxide in the slag is reduced to metallic zinc (vapor) by carbon, oxidized by secondary air, and carried out in the furnace gases for recovery of zinc oxide dust.

Zinc recovery rotary kilns (Waelz kilns) are up to 95 m long with an internal diameter of up to 4.5 m and they are lined with refractory material. The granulated blast-furnace slag is mixed with other zinc intermediates (steel dusts, etc.) and solid fuel reductant; it travels down the kiln and is heated to reaction temperature by combustion gases from a burner at the discharge end. The stripped slag is usually close to its melting point at discharge, and normally contains 2 – 3 % zinc and <0.25 % lead. Production from a Waelz kiln is about 120 – 155 kg of zinc per day per cubic meter of furnace volume, declining with increasing furnace size. The captured dust contains roughly 75 % zinc and lead. Zinc recovery is 88 – 95 % and lead recovery is about 90 – 98 %.

In the *slag-fuming process* a mixture of coal dust and air is injected through tuyeres into a liquid blast-furnace slag at 1150 – 1250 °C in a water-jacketed furnace. The slag is delivered in liquid form directly from the lead blast furnace and may have up to 30 % of solid granulated slag or slag skulls added to it. The process is typically operated in batch mode in furnaces of 50 – 100 t capacity.

At the Plovdiv lead smelter in Bulgaria, slag is tapped continuously from the lead blast furnace through a slag syphon. It then flows through a preheating electric furnace (+ 100 °C) and enters the continuous slag fuming furnace. Heavy fuel oil is used as reductant rather than the more conventional pulverized coal.

Metal recoveries in slag fuming plants are about the same as in Waelz kilns. Slag fuming furnaces normally recover heat from the hot gases in waste heat boilers or air heat exchangers.

4.2.2.4. *Imperial Smelting Process*

In the Imperial Smelting process, lead and zinc are produced together in a single blast furnace operation. This is basically a zinc production process in which lead is also obtained. It is especially well-suited to processing lead – zinc mixed concentrates. The zinc to lead ratio in the sinter mixture is usually 2 : 1. The furnace design is similar to that of conventional lead blast furnaces of the chair-jacket type, but there is only one row of tuyeres. The top is closed by a double-bell system. For process principles and description, see ⟶ Zinc.

4.3. Roast Reaction Processes

Lead sulfide reacts with lead oxide or lead sulfate to form metallic lead and sulfur dioxide in the *roast reaction*:

$$PbS + 2\,PbO \rightarrow 3\,Pb + SO_2$$

$$PbS + PbSO_4 \rightarrow 2\,Pb + 2\,SO_2$$

A number of processes dating from the earliest times have been used to recover lead by partial air oxidation of lead sulfide. These processes are categorized as excess sulfur or excess oxygen, depending on whether the reaction series exhausts the oxide and sulfate or the sulfide species, respectively. High-grade lead concentrates are required in order to maximize the activities of the lead species and, therefore, lead recovery.

The principal commercial example of the excess sulfur type is the *Boliden electric furnace process.* Since this process contains elements of flash smelting, it is described in Section Direct Smelting Reduction Processes.

An example of the excess oxygen roast reaction chemistry is the *Scotch (Newnham) hearth process,* one of the earliest lead-making techniques. The old furnace design was improved and automated in the Bleiberger Bergwerks Union (Schlippenbach) rotary hearth [14]. This process produces a small quantity of high-lead (grey) slag which requires separate recovery treatment, and low-strength sulfur dioxide gas, unsuitable for production of sulfuric acid. The hearth processes are now antiquated and are responsible for a very small amount of lead production.

4.4. Direct Smelting Reduction Processes

The fundamental problems encountered in the sinter, oxidation – blast furnace method of lead production (e.g., high process and hygiene ventilation gas volumes, high maintenance costs for sintering machinery, and expensive coke fuel for reduction) have long been recognized. The quest for a simpler, more direct smelting system has been pursued for many years, and several new processes have now achieved industrial application.

Table 2 summarizes the key features of the most prominent new lead smelting processes. They can be classified into two principal categories: *flash smelting* of dried feed mixtures using air or oxygen, and *bath smelting* of undried feed mixtures also with air or oxygen. Oxidation of the feed produces a medium-strength sulfur dioxide gas, a variable amount of lead bullion, a slag rich in lead monoxide, and a lead sulfate flue dust which must be recirculated.

Table 2. Summary of direct smelting lead processes

Process	Smelting type	Source of oxygen	Special features
Boliden Electric furnace	dry charge flash smelting continuous	enriched air	flash oxidation in freeboard of electric furnace; slag reduction by coke breeze in electric furnace; lead bullion is high in sulfur requiring special treatment
Kivcet	dry charge flash smelting continuous	technical oxygen	an electric furnace freeboard is separated into an oxidation flash smelting shaft and a slag reduction section by a partition wall extending into the slag bath; majority of slag reduction is done by coke checker at bottom of smelting shaft
Boliden Kaldo	batch dry charge flash smelting into TBRC vessel	technical oxygen	batch oxidation is followed by slag reduction in same vessel using PbS followed by coke reductant
Outokumpu	dry charge flash smelting continuous	technical oxygen	slag reduction done in separate electric furnace employing powdered coal injection with nitrogen through lance
Isa Smelt	moist charge slag bath smelting continuous	air or enriched air	air is injected into slag batch by Sirosmelt lance; feed dropped into slag bath; slag reduction by coal in a similar lance-vessel unit
QSL	moist charge slag bath smelting continuous	technical oxygen	horizontal cylindrical reactor divided into oxidation and reduction sections by refractory wall; oxygen injected into lead-slag bath through nitrogen shrouded tuyeres; moist feed dropped into slag bath; slag reduction by submerged combustion of oxygen fuel in gas shrouded tuyeres

The strength of the sulfur dioxide gas produced depends on the degree of oxygen enrichment and the amount of auxiliary fuel required for autogenous operation of the oxidation step. With the use of oxygen, sulfur dioxide gas contents of 20 – 30 % are typical.

The amount of direct-smelt lead bullion produced during the oxidation step depends on the thermal balance. In the case of direct smelting of a very high-grade lead concentrate, up to 80 % of the lead in the charge can be obtained as direct bullion with the remaining 20 % as a slag rich in lead monoxide. With low-grade feed charges requiring auxiliary fuel, all lead can be oxidized to a high-lead slag, with no direct bullion production.

The second step in direct smelting processes is reduction of lead from the high-lead slag (25 –40 % Pb). Slag has been successfully reduced by submerged fuel combustion or injection, reduction assisted by an electric furnace, and a combination of these techniques.

4.4.1. *Air Flash Smelting*

Boliden Process [15]. Dry, high-grade galena concentrate (65 – 75 % Pb) is blown into an electric furnace with oxygen-enriched air. Limestone flux and coke breeze are also included in the feed charge. Suspended lead sulfide particles react to produce lead monoxide and lead metal, which collect in the furnace hearth along with the fused slag constituents. Desulfurization is not complete. The furnace products are lead bullion containing 2 – 3 % sulfur, slag containing 4 – 5 % lead, and flue dust. Desulfurization of bullion is completed by air blowing in a downstream converter.

The Boliden electric furnace process has operated successfully since 1963 at the Ronnskar Smelter in Skelleftehamn, Sweden. The 8000-kVA smelting furnace employs four Söderberg electrodes. Product slag is treated in a slag fuming furnace (see Section Products of the Lead Blast Furnace) along with copper slag for zinc and lead recovery. A section of the electric furnace is shown in Figure 8 and typical daily furnace charge and production figures are listed below:

Charge	
concentrate	280 t
limestone	40 t
iron oxide	7 t
coke breeze	7 t
recycle flue dust	140 t
drosses	15 t
oxygen enrichment	25 % O_2
Products	
slag	70 t (34 % CaO, 17 % Zn)
bullion	280 t

Figure 8. Boliden lead smelting electric furnace

a) Dryfeed hoppers; b) Dry feed-air injection ports; c) Söderberg electrodes; d) Slag bath; e) Lead bullion

Several drawbacks have prevented wider industrial application—high-grade lead concentrate is required, flue dust production is high, and bullion treatment is required to remove residual sulfur.

4.4.2. *Oxygen Flash Smelting*

The most prominent oxygen flash smelting processes for lead production are the Kivcet CS process developed in the former Soviet Union [16], the Kaldo process developed by Boliden in Sweden [17], and the Outokumpu process developed in Finland [18]. In all these processes, very complete desulfurization of the charge is effected through the production of a strongly oxidizing slag. The required oxygen potential can only be achieved with lead monoxide when its weight fraction is high (>35 % lead in the initial slag) [19]. The other major slag components are in ratios very similar to those found in blast-furnace slag.

4.4.2.1. *Kivcet Process*

The Kivcet process was developed at the Vniitsvetmet Institute, Ust Kamenogorsk, the former Soviet Union. Of the developed smelting processes, the Kivcet process is the most advanced, with industrial plants operating at the Ust Kamenogorsk Lead – Zinc Combine and at the Samim lead – zinc metallurgical complex in Portovesme, Sardinia. The Kivcet furnace at the Samim plant is a sophisticated structure, employing extensive water-cooling on the furnace side walls and in the smelting shaft. The general process flows are shown in Figure 9. Figure 10 shows a section of the smelting furnace at

Portovesme.

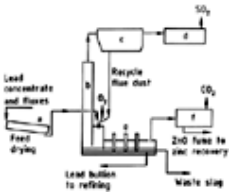

Figure 9. Kivcet lead smelting process

a) Feed drying; b) SO_2 gas cooler; c) SO_2 gas cleaning electrostatic precipitation; d) SO_2 gas scrubber; e) Flash smelting furnace; f) Electric furnace gas cleaning

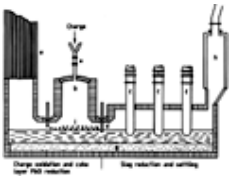

Figure 10. Kivcet lead smelting furnace (Portovesme, Italy)

a) Oxygen burner for dry feed; b) Flash smelting shaft; c) Furnace electrodes; d) Partition wall; e) Membrane wall SO_2 gas cooler; f) Slag bath; g) Lead bullion; h) Zinc fume combustor – cooler; i) PbO reduction coke layer

As shown in Figure 10, the oxidation section of the electric furnace freeboard is separated from the reduction section by a water-cooled, gastight wall (d) which dips into the furnace slag bath. Dried and blended charge components consisting of lead concentrate and other lead-bearing materials, required fluxes, recycle flue dust, and oxygen are fed through burners (a) at the top of the oxidation shaft (b). Combustion of charge constituents at a temperature up to 1400 °C and a shaft height of 3 – 5 m permits almost complete desulfurization before collection of the reaction products in the slag bath (f).

Reduction of the lead monoxide from charge combustion is accomplished mainly by using a layer of coke particles floating on the slag bath under the combustion shaft. The coke particles are also a component of the burner feed. Reduced slag and lead bullion flow under the submerged partition wall (d) into the reduction section of the electric furnace where reduction of lead monoxide is completed to produce a slag containing 2 – 3 % lead. If desired, additional coke breeze reductant and electrical energy can be used to fume zinc. In this mode of operation, product slag containing 3 – 4 % zinc and 1 – 2 % lead can be obtained. Zinc fumed in the electric furnace can theoretically be recovered as metal using a lead splash condenser, but a simpler approach is to burn the zinc vapor to zinc oxide in a combustion flue (h) and recover an oxidized flue dust for further zinc recovery processing.

Lead bullion is removed from the furnace by a syphon and slag is withdrawn continuously from a weir overflow.

Sulfur dioxide gas from the combustion shaft passes under a water-cooled partition wall before rising in an extended vertical wall boiler duct. Cooling of the sulfur dioxide gas – flue dust mixture to 450 °C is accomplished in the vertical boiler stack (e) before the gases enter an electrostatic precipitator for separation of the lead sulfate fumes. Flue dust production is moderate, about 10 – 20 % of the feed charge weight.

The Kivcet process can smelt oxidic charge materials such as zinc leach residues. Extra auxiliary fuel for adequate flame temperature and sufficient shaft height for adequate retention time must be provided to achieve the required degree of desulfurization.

4.4.2.2. *Boliden Kaldo Process*

The Boliden Kaldo process employs a top-blown rotary converter in a batch manner. A cycle starts with oxygen combustion of dry charge constituents into a preheated empty vessel. Figure 11shows the process flows. Charge combustion is continued until lead bullion and high-lead slag have accumulated to the limit of vessel weight or volume capacity. Charge combustion is then stopped and slag reduction commences. Lead sulfide concentrate is used to reduce a major portion of the lead monoxide in the slag, followed by coke breeze reduction to discharge slag composition. The vessel's slag and bullion contents are then poured into ladles, and the smelting cycle repeated.

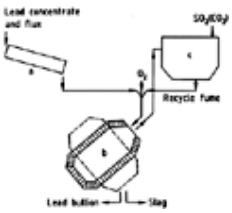

Figure 11. Boliden Kaldo process

a) Feed drying; b) Kaldo vessel; c) Gas cooling and cleaning

The Boliden Kaldo process can handle a variety of lead-containing charge materials such as secondary scrap and flue dusts. The cyclic production of sulfur dioxide – bearing gas and barren reduction gas introduces some complications to the gas handling and sulfur dioxide recovery system.

4.4.2.3. *Outokumpu Process*

Outokumpu Oy of Finland pioneered the air flash smelting of copper concentrates in the late 1940s, and this process has become standard worldwide. Experimental work on air flash smelting of lead concentrate in pilot plants was carried out in the 1960s. Little commercial interest was shown at that time, and pilot-plant testing on several types of lead concentrates was not resumed until the late 1970s, when oxygen rather than air was used for the flash smelt. Full-scale lead smelter

engineering studies have been carried out, but to date no production plants have been constructed.

The process scheme of the Outokumpu lead flash smelting process is shown in Figure 12. Lead concentrate and flux are dried in a rotary dryer (a), mixed with recycle flue dust, and then flash smelted with oxygen in the oxidation furnace smelting shaft (b). High-lead slag and lead bullion collected in the furnace hearth may be tapped separately to provide a direct lead bullion product. Alternatively, if direct lead bullion production is small, slag and bullion are tapped together into a separate electric slag reduction furnace (d). In the electric furnace, pulverized coal is injected with nitrogen into the slag bath to reduce lead monoxide.

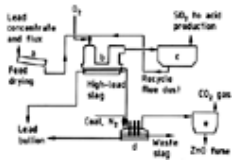

Figure 12. Outokumpu lead smelting process

a) Feed drying; b) Flash smelting furnace; c) SO_2 gas cooling and cleaning; d) Electric slag reduction furnace; e) Electric furnace gas cleaning

The pilot-plant equipment consisted of a rotary dryer, dry concentrate bins, pneumatic conveyors for concentrate and recycle flue dust, a flash furnace (15 m^3), a waste heat boiler (3 t/h steam), an electrostatic precipitator, and a 2 MVA electric furnace.

The pilot plant processed 3 – 5 t/h of concentrate ranging from 47 % to 75 % lead content. Bullion and slag were continuously tapped to the electric furnace for slag reduction. About 10 000 t of lead concentrate were processed.

4.4.3. *Air – Slag Bath Smelting, the Isasmelt Process*

The Isasmelt process (Fig. 13) consists of two stages and was developed by Mount Isa Mines and Csiro in Australia [20]. Both stages, the charge oxidation stage and the slag reduction stage, employ a stationary, cylindrical, refractory-lined reaction vessel. The process features the Sirosmelt lance, which is a steel pipe that is cooled by the high-velocity process air which it injects into the melt. In the oxidation vessel (b), air or oxygen-enriched air and supplementary oil fuel are injected through a Sirosmelt lance (c) into the center of the vessel. The blended feed materials (d), lead concentrate, fluxes, and moist recycle flue dust are dropped into the slag bath and agitated by the lance burner. Virtually all the charge lead content is oxidized to a high-lead monoxide slag. The slag is transferred to the reduction vessel (e), which employs a similar air –fuel lance (f), with additional lump coal charged separately as reductant.

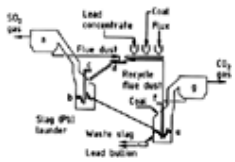

Figure 13. Isasmelt lead process

a) SO_2 gas cooling and cleaning; b) Oxidation; c) Lance for oil and air; d) Feed mixing; e) Reduction furnace; f) Lance for oil and air; g) Reduction gas cooling and cleaning

Starting in 1983, a 5 t/h concentrate feed demonstration plant was used to prove the smelting stage of the process. In the first year of operation, 21 000 t of lead concentrate were treated to produce high-lead monoxide slag, which was subsequently processed by sintering and blast furnace reduction. In 1985, construction of the second-stage reduction vessel demonstration plant was started. At the present time the installation of a full-scale production plant at Mount Isa is being studied.

A cross-section of the oxidation furnace is shown in Figure 13 (b). The furnace consists of a cylindrical steel shell, 2.7 m in diameter and 4.5 m high, lined with chrome magnesite bricks backed with an insulation lining. Typical analyses are given in Table 3. Operating conditions for the demonstration plant are as follows:

Lead concentrate rate	4 t/h
Smelting temperature	1 030 °C
Air: concentrate ratio	125 % of theoretical
Fume rate	10 % of lead in feed
Flux addition	12 % of concentrate rate
Oil requirement	25 kg per tonne of concentrate

Table 3. Typical analyses from the Isasmelt demonstration plant oxidation vessel

Component	Lead concentrate, %	Slag product, %
Pb	47.3	51.8
Zn	7.0	7.6
Fe	12.4	13.6

SiO_2	2.8	9.8
CaO	0.9	1.5
S	23.8	0.3

4.4.4. *Oxygen – Slag Bath Smelting, QSL Process*

The QSL Process is named after its inventors, P. E. QUENEAU and R. SCHUHMANN Jr., as well as the process developers Lurgi Chemie [21]. Development of the process progressed through bench-scale tests in 1974 – 1975, pilot-plant operation in 1976 – 1979, and a demonstration plant operation at Metallhüttenwerke, Berzelius, Duisburg, Federal Republic of Germany from 1981 to 1985. At present three commercial plants are under construction in Trail, Canada (Cominco Ltd.), the Federal Republic of Germany (Berzelius Binsfeldhammer Smelter), and the Peoples Republic of China. Operation of the Canadian plant is scheduled for late 1989.

Figure 14 shows the general scheme of a QSL process lead smelter. Lead concentrate, fluxes, auxiliary coal fuel, recycle flue dust, and other charge components are blended (b) in a moist condition to form a single feed charge to the QSL reactor. The reaction vessel is a horizontal refractory-lined cylinder that is divided by a partition wall into an oxidation section (c) and a slag reduction section (d). The working dimensions of the demonstration plant reactor (10 t concentrate per hour, 30 000 t of lead per year) were 2.5 m (diameter) by 22 m (length, 7 m of which is occupied by the oxidation zone). Nitrogen-shrouded tuyeres on the bottom of the reactor injected oxygen into the oxidation section. Similar shrouded tuyeres were employed for submerged combustion of oxygen – coal or oxygen –natural gas reduction mixtures. Blended feed was charged through ports at the top of the oxidation reaction section into the injection-agitated lead and slag bath.

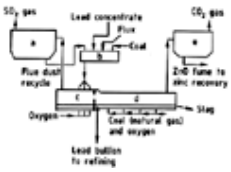

Figure 14. QSL lead smelting process

a) Oxidation gas cooling and cleaning; b) Feed mixing; c) Oxidation section; d) Slag reduction section; e) Reduction gas cooling and cleaning

The oxidation products in the QSL process are high-lead slag, sulfur dioxide gas, flue dust, and lead bullion. The amount of lead produced in the oxidation zone depends on the charge makeup and lead concentrate grade. Slag from the oxidation zone flows through an underflow port in the vessel partition wall into the slag reduction zone (d). Lead bullion is removed by a lead syphon at the side of the vessel. The smelting vessel is rotated by 90° during stoppages to bring the injection tuyere systems above the vessel slag level. Inspection of the refractory and tuyere maintenance can be carried out with the vessel in this standby position.

Figure 15 shows the QSL smelting plant in Trail, Canada. The reaction vessel has dimensions of 4.5 m (diameter) by 12.5 m (length) in the oxidation section, and 4.0 m (diameter) by 28 m (length) in the reduction section. Nominal plant capacity is 160 000 t/a of lead and 230 000 t/a of reduced slag. Reduction is performed with partially combusted natural gas.

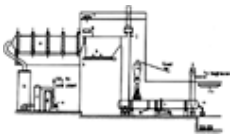

Figure 15. QSL lead smelting plant (Trail, Canada)

a) Electrostatic precipitator; b) Spray tower; c) Cooling tower; d) Stripper; e) SO_2 fan; f) Crane; g) Steam drum; h) Waste heat boiler; i) Emergency stack; j) Cooling channel; k) Feed bin; l) Reactor; m) Lead syphon; n) Slag tap

Three smelting vessel configuration options are shown in Figure 16.

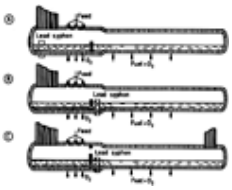

Figure 16. QSL smelting vessel configurations

A) Conventional, low arsenic – antimony bullion; B) Increased arsenic – antimony recovery in bullion; C) Recovery of zinc oxide fume

Figure 16 A. This is the simplest smelting configuration and is employed at the Duisburg demonstration plant. Reduction gas and oxidation gas flow into the same gas cooling system. Reduction and fuming of zinc in the reduction zone is controlled to limit zinc circulation in recycle flue dust. Lead bullion produced in the reduction zone flows countercurrent to oxidation slag through the partition wall underflow port. The product lead bullion is withdrawn from the oxidation end of the furnace. With this configuration the arsenic and antimony contents are minimized in the bullion and maximized in the reduced slag.

Figure 16 B. This differs from the previous option in that the product bullion syphon is located on the reduction side of the slag partition wall. Arsenic and antimony in bullion are maximized with this option.

Figure 16 C. This smelting vessel configuration has an auxiliary gas offtake at the reduction end of the vessel for withdrawing a zinc-bearing reduction fume. The lead syphon is located on the reduction side of the partition wall for

maximum arsenic and antimony deportment to bullion. This design is employed in the Cominco Trail smelter.

4.5. Other Lead Processes

Several other lead processes have been tested on small and pilot scales but none has been commercialized. The *Halkyn process* involves the fused-salt electrolysis of lead sulfide dissolved in molten lead chloride; it has been modified and improved, most by the U.S. Bureau of Mines [22]. Trouble-free fused-salt electrolysis is feasible only with high-quality lead concentrates; high contents of impurities (e.g., copper) cause the formation of barrier layers and cell complications.

Wet processes have not succeeded although many have been investigated in the chloride, nitrate, sulfate systems, and with selective alkylamine leaching of lead sulfate followed by carbonation of lead. An alternate leaching system employs ammoniacal ammonium sulfate, which also dissolves lead sulfate well and is cheaper than the amine solutions.

[Top of Page]

5. Refining of Lead Bullion

Charles A. Sutherland and Edward F. Milner

Production lead, known as bullion, contains numerous dissolved impurities. Molten lead is an excellent solvent for a variety of metals and compounds, and raw bullion commonly contains copper, iron, zinc, sulfur, arsenic, tin, antimony, bismuth, noble metals (silver, gold, platinum-group metals), oxygen, and occasionally nickel, cobalt, and tellurium. The lead content varies between about 90 and 99.9 % depending on the process and feed quality. The percentages of impurity contents for blast-furnace lead bullion treated in refining plants are as follows:

Cu	0.2 – 4.0
Fe	0 – 0.5
Zn	0 – 0.5
Ni	0 – 0.1
Co	0 – 0.1
As	0 – 2.0
Sb	0 – 6.0 (usually < 2.0)
Ag	0 – 2.0
Bi	0 – 6.0 (usually < 2.0)
Sn	0 – 4.0 (usually < 1.0)
S	typically 0.2 – 0.3
O	typically 0.1

The object of lead bullion refining is to produce pure commercial lead metal and to separate the valuable impurity metals into their most valuable marketable form. The two main refining routes used are pyrometallurgical and electrolytic. The latter is generally reserved for higher levels of impurities (especially bismuth).

Pyrometallurgical processes are generally operated in a batch manner; however, continuous operation is being applied to many of the process steps. The advantage of pyrometallurgical refining lies in the consecutive selective recovery of impurity metals.

In *electrolytic refining,* all the impurities are concentrated in the anode slime and must be separated by a relatively complicated process. On the other hand, electrolysis yields lead that is very low in bismuth (< 10 g of bismuth per tonne of lead), which cannot be economically produced by the pyrometallurgical route. Over 90 % of installed capacity for lead-bullion refining worldwide is based on pyrometallurgical methods.

5.1. Pyrometallurgical Refining

Principles. When metals are refined pyrometallurgically, impurities dissolved in the parent metal are enriched in a phase that is insoluble (or sparingly insoluble) in the parent metal. Enrichment can be achieved by adding reagents, by controlling the pressure and temperature, or by a combination of both techniques. Methods are classified as chemical, precipitation, and distillation processes.

Attention to the kinetics of the unit operations involved in lead refining is of the utmost importance. Kinetic factors can govern the degree of refining that is attainable when the system is far removed from the theoretical chemical or physical equilibrium. In other systems the chemical equilibrium is approached closely. A good review of the physical chemistry is given in [23].

Chemical Processes. A reagent added to the parent metal reacts with one or more impurities to form a compound that is insoluble in the parent metal. Occasionally, fluxes are also used, for example to cause oxides formed in the solid phase to go into a liquid slag. Reagents include oxygen (atmospheric oxygen or oxygen-evolving salts), sulfur, and chlorine. The refining step is then referred to as selective oxidation, sulfurization, or chlorination. Reagent selection is based on a comparison of the free energy of formation of the oxides, sulfides, or chlorides of the parent metal and of the impurities in question.

Precipitation Processes. Generally, impurities become less soluble in the host metal as the temperature falls, and can be rejected either directly from the melt (e.g., copper from the binary system copper – lead), or as a compound with the parent

metal (e.g., iron as $FeZn_3$ from iron – zinc solutions). Alternatively, suitable reagents can form insoluble compounds that can be skimmed off (e.g., bismuth from lead – bismuth solutions after addition of calcium and magnesium); this is really a chemical process that involves the formation of insoluble intermetallic compounds.

Distillation Processes. Distillation processes are employed when a large difference in vapor pressure exists between the parent metal and the impurity (e.g., in the separation of residual zinc from lead bullion after Parkes desilvering).

Process Steps. All of the above processes are used in the pyrometallurgical refining of lead bullion. A typical pyrorefining scheme is as follows:

1. *Removal of copper* is a two-stage operation. The first stage is primary decoppering by *drossing* (precipitation). The second is *selective sulfidizing* of copper by the Colcord process, which involves the addition of sulfur (a chemical process).
2. *Removal of arsenic, tin, and antimony* by selective oxidation with injected air or with oxidizing agents such as sodium nitrate in a caustic slag (Harris process).
3. *Removal of noble metals* is usually called desilvering or desilverizing and involves intermetallic precipitation with zinc by the Parkes process.
4. *Removal of zinc* remaining after desilvering is achieved by vacuum distillation.
5. *Debismuthizing.* If required to meet refined lead specifications, bismuth is removed by precipitation with alkali metals and/or alkaline-earth metals, for example with calcium and magnesium in the Kroll – Betterton process.
6. *Removal of alkali metals and alkaline-earth metals* remaining after debismuthizing is called final refining and is performed by selective oxidation with air under a salt slag cover.

5.1.1. *Decoppering*

Drossing (Liquation). Copper is soluble to several percent in hot lead bullion but to only 0.06 % at the freezing point. When the lead cools, copper and other impurities precipitate as a lead-rich *copper dross*. This process is always the first step in any lead refining operation. In the first phase of cooling to about 950 °C, nearly all dissolved iron and zinc are precipitated as oxides, magnetites, or spinels. The noble metals and other impurities with higher solubilities (e.g., copper, arsenic, antimony, tin, bismuth) remain dissolved in the lead.

In the next cooling stage, the copper compounds of sulfur, arsenic, antimony, and tin separate, along with considerable lead sulfide and entrained metallic lead. The large lead content must be separated by a dross retreatment process. Nickel and cobalt, if present, separate similarly to copper.

Fine decoppering may also be referred to as sulfur decoppering or sulfur drossing (Colcord process). After drossing, elemental sulfur is mixed into the molten lead bullion at the lowest possible temperature (320 °C) for selective sulfidization of copper. Lead(II) sulfide and copper (II) sulfide form stable phases in the ternary system Pb – Cu – S when liquid or gaseous sulfur is present. The copper sulfide can be produced by the reaction between copper – lead alloy and liquid or gaseous sulfur whenever there is a local excess of sulfur:

$$(\mathrm{Cu})_{\mathrm{Pb}} + \mathrm{S} \rightarrow \mathrm{CuS}$$

Copper(II) sulfide can decopper lead by the reaction

$$(\mathrm{Cu})_{\mathrm{Pb}} + \mathrm{CuS} \rightarrow \mathrm{Cu_2S}$$

Sulfur simultaneously reacts with lead, but much more slowly than with copper:

$$\mathrm{Pb} + \mathrm{S} \rightarrow \mathrm{PbS}$$

The lead corner of the Pb – Cu – S system is shown in Figure 17 and suggests that the copper level can in theory only be decreased to about 0.05 % by sulfur precipitation. With cooling, however, lead sulfide forms as well as copper (I) sulfide (Cu_2S). In fact copper levels of 0.001 – 0.002 % can be achieved if minor amounts of silver or tin are also present and attention is paid to good shear stirring and gradual addition of sulfur [23]. This is because silver and tin slow down the reaction between lead and sulfur. Without silver or tin, 0.01 % copper can be attained only by repeated sulfur additions and with prohibitive sulfiding of lead.

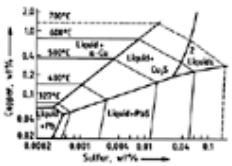

Figure 17. Lead corner of the Pb – Cu – S ternary system [23]

Process Engineering. Decoppering of lead bullion usually takes place immediately after the bullion is discharged from the blast furnace. In *batch decoppering,* drossing is carried out in kettles with capacities of up to 300 t. Powerful stirrers help separate the phases, allowing the formation of a " dry " dross containing little entrained lead metal. The dry dross is removed from the melt surface by skimming, usually with mechanically or pneumatically driven perforated scoops, or by suction. In industrial drossing, final copper contents of 0.06 – 0.12 % are reached. This level of copper would greatly increase the zinc requirement for Parkes process desilvering, and therefore fine (sulfur) decoppering is always necessary.

An important advance is *continuous decoppering*, which combines decoppering and reprocessing of the copper dross in a

single stage. Continuous drossing has several advantages over the batch process of dry stirring in kettles. Effective lead drossing by cooling is combined with the formation of matte and slag in a single process step. Handling of dusty copper dross is eliminated, which permits a major improvement in plant hygiene. Success in the development of a continuous drossing process has been achieved in Australia and the former Soviet Union by using a deep lead bath reverberatory furnace. The process is also employed at Cominco's lead smelter in Trail, Canada.

Figure 18 shows the furnace used by Broken Hill Associated Smelters (BHAS) in Port Pirie, Australia. The gas-fired furnace has an external bullion cooling launder (h) with immersible water-cooled plates (j). Hot bullion from the blast furnace is charged to the drossing furnace and is cooled by mixing with colder lead from the cooling launder. Dross formed rises to the heated surface and is melted to matte and slag. The third-generation furnace now used at Port Pirie has hearth dimensions of 6.15 m by 2.0 m and a depth of 1.7 m [24]. The process is best suited to refining lead with a fairly uniform impurity content. An important operating criterion is the ratio of copper to sulfur in the bullion. A high ratio causes the formation of high melting point speiss crusts in the furnace. Recent development has allowed addition of elemental sulfur to the furnace permitting satisfactory operation at bullion copper levels approaching three percent.

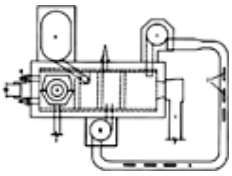

Figure 18. Plan view of the continuous drossing furnace of Broken Hill Associated Smelters at Port Pirie, Australia

a) Product drossed lead; b) Slag and matte tapping spout; c) Circulating pump pot; d) Gas burners; e) Charge port; f) Ventilation; g) Return launder pot; h) Cooling launder; i) Water-cooled flue; j) Water-cooled plates

In the Russian continuous drossing process, bullion cooling is achieved by indirect water-cooling through the furnace bottom. Copper drossing and cooling are finished in batch kettles, dross being recycled to the drossing furnace. Soda ash is added to the furnace to produce a low lead soda matte.

Sulfur Drossing. *Final decoppering* (sulfur drossing) is carried out *batchwise* in most plants in kettles fitted with a powerful stirrer. Fine sulfur is added in increments to the lead vortex at 330 – 340 °C, and the formed copper dross is skimmed, usually with the aid of sawdust. This dross is much lower in copper (and higher in sulfur) than primary decoppering dross and is returned to a convenient point in the smelting process. In industrial operation, about 1 kg of sulfur is added per kilogram of copper; the end copper content must be less than 50 g/t if possible. Several sulfur treatments are needed to achieve this.

A *continuous sulfur drossing system* has been developed by BHAS at Port Pirie [24] and has been operating since 1981. The process employs a series of small stirred kettles [23]. A continuous flow of bullion from the copper drossing furnace is cooled, treated with elemental sulfur, and mechanically drossed.

5.1.2. *Removal of Arsenic, Tin, and Antimony*

The removal of arsenic, tin, and antimony is called *softening* . Two processes are in use: selective oxidation with air in fired furnaces and the Harris process.

Oxidation in Air-Blown Furnaces. This method is based on the fact that tin, arsenic, and antimony react more readily with oxygen than lead, and are oxidized (in the order listed) before lead at 700 – 750 °C. Air is injected through vertical lances onto the surface of a lead bath in a rectangular oil- or gas-fired reverberatory furnace. The oxides of tin, arsenic, antimony, and lead form a fluid slag, which is periodically removed from the surface of the lead bath. A considerable amount of lead is oxidized and the slag generally contains 50 – 60 % lead monoxide. In batch reverberatory softening, segregation of slag skimmings into high arsenic – tin and high antimony lots is possible, but not usually practised.

The oxides float on the molten lead and are withdrawn as softening slag. Refining time depends on the amount of impurities in the charged bullion and also on the air-lancing method. Impurity oxidation rates in the order of 4 kg m^{-2} h^{-1} are typical.

Continuous bullion softening is operated by BHAS, Port Pirie, Australia. Drossed lead is pumped continuously into a softening reverberatory furnace fitted with an air-lance system similar to a batch softening furnace. The rate of bullion feed and the rate of arsenic and antimony oxidation are such that the antimony content is maintained at 0.02 – 0.05 %. Antimony oxidation rates of 40 kg m^{-2} h^{-1} are achieved, about an order of magnitude greater than in batch softening.

Harris Process. The Harris process employs molten sodium hydroxide as a medium for the oxidation of arsenic, tin, and antimony from lead bullion. Metallurgie Hoboken – Overpelt (MHO) Hoboken, Belgium [25] and the Norddeutsche Affinerie, Hamburg, Federal Republic of Germany [26], use this process. Harris refining is performed batchwise using a reaction cylinder holding molten sodium hydroxide installed in a lead kettle. Figure 19 shows one half of a dual kettle Harris reactor as employed by MHO at the Hoboken smelter. Bullion is pumped through the Harris reactor caustic charge until the required degree of impurity removal is achieved. A valve at the bottom of the caustic vessel is then closed allowing the loaded caustic to be discharged to a transport ladle for subsequent hydrometallurgical processing. Fresh molten caustic recovered from the salt processing plant is charged to the reactor by ladle for the next refining cycle.

Figure 19. Harris reactor at Metallurgie Hoboken – Overpelt S.A. [25]

One half of the dual reactor is shown.a) Lead bullion; b) Lead bullion pump; c) Molten caustic holding vessel; d) Molten caustic supply ladle; e) Reactor control system

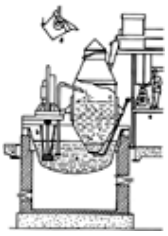

The Harris process is the preferred softening process for lead bullion containing high levels of impurities. The MHO plant at Hoboken can remove up to 400 t of arsenic, tin, and antimony from lead bullion per month.

Harris softening involves the oxidation of arsenic, tin, and antimony by air or by sodium nitrate added to the caustic bath. The oxidation products, in order of their formation, are sodium arsenate, sodium stannate, and sodium antimonate:

$$2\,As + 2\,NaNO_3 + 4\,NaOH \rightarrow 2\,Na_3AsO_4 + N_2 + 2\,H_2O$$

$$2\,Sb + 2\,NaNO_3 + 4\,NaOH \rightarrow 2\,Na_3SbO_4 + N_2 + 2\,H_2O$$

$$5\,Sn + 4\,NaNO_3 + 6\,NaOH \rightarrow 5\,Na_2SnO_3 + 2\,N_2 + 3\,H_2O$$

Tellurium is oxidized to sodium tellurate and telluride:

$$3\,Te + 6\,NaOH \rightarrow 2\,Na_2Te + Na_2TeO_3 + 3\,H_2O$$

$$xTe + Na_2Te \rightarrow Na_2Te_{x+1}$$

The arsenic salt is soluble in the molten caustic whereas the salts of tin and antimony form a solid suspension. Because of the reversibility of the oxidation chemistry (e.g., Na_3SbO_3 will oxidize arsenic and tin from bullion), a loaded caustic salt is produced that is high in arsenic and tin, followed by a salt containing primarily antimony. Table 4 gives the distribution of arsenic, tin, and antimony in the products of the Harris refining process at the Hoboken lead smelter. A flow sheet is given in Section Processing of Intermediate Products of Pyrometallurgical Refining (Fig. 24).

Table 4. Distribution of metals in products separated by the Harris process at Metallurgie Hoboken – Overpelt S.A. [23]

Product	Metal, wt %		
	As	Sn	Sb
Lead particles As-Sn	5.5	5.0	2.1
Lead particles Sb	0.2	0.1	4.0
Black antimonate **	0.5	2.5	3.8
Calcium stannate	2.2	91.8	0.2
Calcium arsenate	91.6	0.5	0.3
White antimonate		0.1	89.6

* For origin of products see Figure 24, p. 222.
** Contains more than 80 % of extracted Te, Se, and In.

5.1.3. *Removal of Noble Metals*

Noble metals and residual copper are removed from lead by the Parkes process in which zinc metal is added to the molten silver-bearing lead and the mixture is cooled (see also ⟶ Silver, Silver Compounds, and Silver Alloys). Silver – zinc mixed crystals precipitate out and rise to the bath surface as a crust (Parkes process crust). The intermetallic compounds have a higher melting point than lead and are virtually insoluble in zinc-saturated lead.

The metallurgical operations in the Parkes process are based on the lead corner of the ternary system Ag – Pb – Zn (Fig. 20). Most of the silver removal takes place in the liquid + ϵ-phase Ag – Zn crystals region; the last crop of crystals in the latter part of the second stage comes from the liquid + η-phase region. Silver removal is very efficient and depends on the quantity of zinc and the number of stages used, but it is usually not economical to use more than two stages.

Figure 20. Lead corner of the Pb – Ag – Zn system in rectangular coordinates [23]

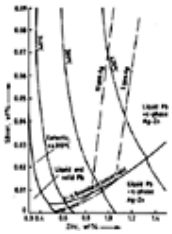

The isotherms (—– – —– –) indicate how much zinc can be dissolved in the lead. The cooling curves (— — —) indicate the course of desilvering in the batchwise two-stage process for final silver concentrations of 5 and 10 ppm.

Process Engineering of Desilvering. In the *batch process,* desilvering is performed in two stages in kettles with capacities of up to 300 t. In the first stage " lean crust " from the preceding charge is added along with other recycle zinc (see Section Dezincing) and stirred into new softened lead at 460 °C. The " rich crust " is skimmed off for silver recovery. In the second stage, zinc metal is added with stirring; more zinc – silver crystals form, and the bath is then cooled to near freezing at about 370 °C so that the last crystals separate. At this point, the silver content of the near-freezing lead is 5 – 10 ppm.

The *Williams continuous desilverizing process,* developed at Port Pirie, employs a deep kettle surrounded by heating jackets, which impose a vertical temperature gradient on the lead column from 600 °C at the top to the temperature of the ternary Ag – Pb – Zn eutectic at the bottom. The upper part of the column contains ports for the intake of lead, addition of zinc, and removal of the ternary alloy. A siphon for discharging the desilvered lead has its intake just above the bottom of the kettle.

The softened silver-bearing lead is brought to 650 °C in a conditioning furnace and then flows continuously into the desilvering kettle. The lead flows through the zinc layer in the upper part of the kettle, and becomes saturated with zinc. As it descends down the kettle, it is cooled along the imposed temperature gradient. When it approaches the freezing temperature, silver – zinc crystals precipitate and rise to the zinc layer at the lead surface. The desilvered lead contains about 3 ppm Ag and 0.6 % Zn, and leaves the kettle through the siphon.

The composition of a fresh zinc layer changes, becoming enriched in silver (and lead). When it approaches about 15 % silver, 20 % lead, it begins to freeze. The mushy alloy is then ladled out and replaced with fresh zinc. Residual copper and any gold are removed very efficiently along with the silver. Continuous desilvering produces only a single silver-zinc crust which can be sent directly to distillation—an advantage over two-stage desilvering. The modern batch operation, however, uses less zinc and has lower operating costs.

5.1.4. *Dezincing*

The desilvered lead still contains about 0.55 –0.6 % zinc which must be removed. Older methods used air and steam, caustic soda, or chlorine, all of which oxidized the zinc metal. These have been replaced by vacuum distillation of zinc from the desilvered lead.

The large difference in vapor pressure between the host lead and dissolved zinc permits the volatilization of about 95 % of the zinc from the desilvered lead when a vacuum of about 0.133 Pa is applied at about 600 °C. A condensing surface placed close to the surface of the lead allows kinetic control of the distillation. The zinc content of the vapor exceeds the mole fraction in the melt by over 500 % [23].

Process Engineering. *Batch dezincing* is performed in a kettle: a water-cooled bell is sealed vacuum-tight on the kettle and distillation is carried out at about 600 °C and a vacuum of about 0.133 Pa (measured at the pump). The lead is circulated by a stirring mechanism, also sealed vacuum-tight. Crystalline zinc precipitates on the bottom of the water-cooled vacuum bell and is struck off after the termination of the process. Adequate sealing was originally achieved by letting the bell dip into the lead bath (barometric seal); water-cooled gaskets are preferred.

At Port Pirie, *continuous dezincing* has also been developed. The dezincing apparatus (Fig. 21) consists of a thin-film evaporator with roughly 0.75 m^2 of evaporating surface. It can dezinc 16 – 37 t/h of Parkes lead, and 90 – 95 % of the zinc content is recovered on a water-cooled fin condenser. When about 3 t of zinc has been collected, the condenser is replaced. The deposited zinc is melted and returned to the desilvering operation (see Section Removal of Noble Metals).

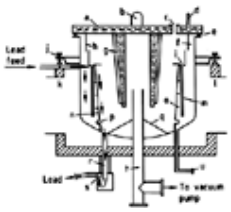

Figure 21. Continuous vacuum dezincing

a) Water-cooled cover; b) Hook; c) Window; d) Cooling water connections; e) Rubber gasket; f) Baffle; g) Cooling fins; h) Weir plate; i) Lead trough; j) Level control; k, l) Furnace setting gas entry and exit; m) Lead inlet to trough; n) Evaporation surface; o) Reinforcement; p) Flexible connection; q) False bottom; r) Lead outflow; s) Lead siphon; t) Vacuum connection; u) Drain

The lead is heated to 590 – 620 °C in a preheat furnace, before entering the vacuum kettle. After dezincing, it leaves the unit at 540 – 560 °C with a zinc content of about 0.03 %. The pressure inside the kettle is 3 – 7 Pa. The residual zinc in the lead bullion is removed during final refining (Section Final Refining and Casting of Lead).

5.1.5. *Debismuthizing*

Debismuthizing is the last step of lead-bullion refining (see also ⟶ Bismuth, Bismuth Alloys, and Bismuth Compounds – Crude Bismuth from Lead Concentrates). It is applied when the bismuth content of the ore is so high that it restricts the use of the soft lead smelted from it. When possible, bismuth-rich and bismuth-poor ores are smelted separately in order to refine

the minimum amount of lead bullion with the maximum bismuth content.

Bismuth is less reactive than lead, therefore it cannot be removed with oxygen, chlorine, or sulfur. It is customarily separated by addition of an alkali metal or alkaline-earth metal forming a bismuthide M_xBi, but this only permits a minimum bismuth content of 0.03 %, which is not adequate for many requirements. The removal of bismuth to <0.01 % can be achieved by the *Kroll – Betterton process* , which uses calcium and magnesium and also by the *Jollivet – Penarroya process*, which uses potassium and magnesium. Only the former remains in use; the metallurgical reactions take place in the lead corner of the corresponding quaternary system (Fig. 22). Bismuth separates as a crust of the double bismuthide, $CaMg_2Bi_2$.

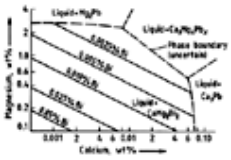

Figure 22. Lead corner liquidus of the quaternary Pb – Bi – Ca – Mg system [23]

The lowest attainable bismuth content is around 0.002 % at the quaternary eutectic point. Further addition of calcium and magnesium does not improve bismuth removal. The reagent requirement is given by the total calcium and magnesium needed for precipitation as the double bismuthide and for saturation of the lead.

Process Engineering. The Kroll – Betterton *batch debismuthizing process* is carried out in kettles. Magnesium and calcium (in lead) are stirred into the lead bath at a temperature of about 420 °C. The bath is then cooled to near the solidification point with slow stirring. The crust (enriched about 10 : 1 in bismuth) is removed in a similar manner to the Parkes process (see Section Removal of Noble Metals). The crust grade can be improved considerably by stirring into the next batch of lead. Both batch and continuous versions of the Kroll – Betterton refining are practiced. A procedure is also used for secondary, fine debismuthizing by antimony addition [27].

5.1.6. *Final Refining and Casting of Lead*

Refined lead contains arsenic, antimony, and tin at a level of 10 – 20 g/t; zinc corresponding to the final content in vacuum dezincing (ca. 30 g/t); and (if debismuthizing has been performed) calcium and magnesium corresponding to the saturation limit. These impurities are removed down to below 1 g/t by treatment with caustic soda and, in some instances, with sodium nitrate.

Most casting is done automatically in strip or (less often) bench form, in machines provided with automatic stacking devices. The casting temperature is around 400 °C. Immediately after admission to the chill mold, the thin oxide surface film is skimmed off. The cast lead has a coarsely crystalline surface; if final refining and casting techniques are correct, there are no temper colors.

5.1.7. *Processing of Intermediate Products of Pyrometallurgical Refining*

Copper Dross. Copper dross is a complex mixture of compounds (see Section Decoppering) and contains a substantial amount of sulfidic and metallic lead. Melting the dross in a retreatment furnace frees the metallic lead and forms a matte phase, consisting mainly of lead and copper sulfides, and a slag which contains metal oxides and often requires silicate fluxing. If arsenic is available and sulfur is deficient, a copper arsenide speiss product is formed. The slag is returned to the smelter, but the matte is a product for sale or further processing.

All constituents of the dross deport to the four possible liquated phases—lead, speiss, matte, and slag. If large amounts of arsenic are present, a speiss forms in which nickel, arsenic, and part of the copper are preferentially concentrated. If copper and arsenic levels in the dross are low, the matte and speiss must be enriched by a further treatment.

In dross retreatment, it is desirable to minimize the lead content of the final matte. If the iron and sulfur contents are known, the final lead content can be approximately estimated from Figure 23: % Fe is equated to % (Fe + Zn) in the lead-free matte and % S can be determined from the lead-free sum, Fe + Zn + Cu + S = 100 %. For example, for a lead-free matte containing 7 % Fe, 1 % Zn, 24 % S, and 68 % Cu, the lead content predicted from Figure 23 is 20 %. The final matte would then contain 20 % Pb, 5.6 % Fe, 0.8 % Zn, 19.2 % S, and 54.4 % Cu. This is a fair estimate only for low alkali mattes. Sodium or calcium additions lower lead solubility in the matte substantially; at 5 % sodium or calcium the solubility is reduced to less than 20 % of that in the same matte without alkali. This is the basis of the soda matte process used by Asarco (El Paso) and St. Joe Lead at Herculaneum, in which about 5 % soda ash is added to the dross during melting and produces a copper enriched matte.

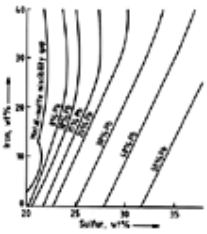

Figure 23. Solubility of lead in Cu – Fe – S mattes [23]

The recovered lead contains most of the noble metals. Matte, and particularly speiss if present, also hold significant amounts of precious metals. Fume from the furnace is enriched in indium, tin, arsenic, and zinc. Slag, matte and speiss are normally

granulated.

Hydrometallurgical methods for treating copper dross involve solubilizing the copper for electrowinning either by pressure oxidation leaching in sulfuric acid or by ammoniacal leaching and solvent extraction.

Softening Slags and Drosses. Softening slag and skimmings with variable contents of tin, arsenic, and antimony (Section Removal of Arsenic, Tin, and Antimony) are usually reduced to hard lead. An effort is made to slag off part of the arsenic during reduction. The process is normally carried out in a rotary furnace, but can also be done in a blast furnace.

In the short rotary furnace at Port Pirie, 7 t of softening slag are reduced batchwise with the addition of 5 % soda, 7 % charcoal, and 6 % " caustic slag " from subsequent refining. The charcoal is added in two portions. After the first portion is reacted, a lead bullion containing 2 – 3 % antimony is tapped off and recycled. After the second reduction, about 85 % of the feed antimony is tapped off as an alloy containing 25 – 30 % antimony.

Harris Salt Slags. Harris slags (Section Removal of Arsenic, Tin, and Antimony) are processed to recover a large portion of the valuable minor metals as black antimonate (Sn, Sb, Te, Se, In) and sodium antimonate (Sb); to precipitate the tin as calcium stannate and arsenic as calcium arsenate; and to recover caustic soda for reuse. The flowsheet of Metallurgie Hoboken – Overpelt is shown in Figure 24.

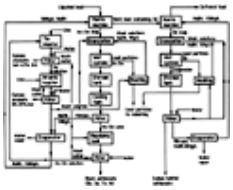

Figure 24. Treatment of Harris salt slags at Metallurgie Hoboken-Overpelt S.A. [25]

Silver – Zinc Rich Crusts for Desilverizing. The purpose of processing the silver – zinc-rich crusts (see Section Removal of Noble Metals) is to recover a malleable rich lead that contains all the noble metals. Some of the lead and as much as possible of the zinc must be removed; the latter is reused. Removal is achieved by expression and/or liquation with production of a dry or liquated crust, followed by zinc distillation.

Expression is commonly carried out in an air-driven Howard press submerged in a liquid lead bath, and the lead content of the compressed crust is brought to 50 – 60 %.

Liquation. Alternatively the crust may be melted at about 600 °C under a salt cover and separated into a lead-rich lower layer, which is returned to desilverizing, and a zinc-rich upper layer which passes to zinc removal.

Distillation allows the zinc content of the liquated rich crusts to be recovered in metallic form so that it can be reused for desilvering lead bullion (see Section Removal of Noble Metals). Distillation can be performed either at STP in the Faber du Faur furnace or under vacuum in the Leferrer furnace.

In the older *Faber du Faur furnace,* liquated zinc-rich crust is placed in inclined bottle-shaped graphite retorts and heated to 1000 – 1500 °C. Zinc condenses in a separate receiver in liquid form and is tapped off periodically. A certain amount of dross (zinc dust) is also formed; since this contains noble metals, it is recycled to the process.

A more modern method is the *Penarroya – Leferrer vacuum distillation* with radiation heating. This runs at a temperature (600 °C) well below the melting point of silver and considerable lead must be added to the crust to hold the silver in solution when most of the zinc has gone. The zinc condenses as a liquid. Low frequency induction furnaces are also used.

The distillation residue consists of silver containing other precious metals, zinc, lead, and copper. Zinc, lead, and copper are removed by cupellation (air oxidation), and the silver bullion is separated from gold by electrolysis (⟶ Silver, Silver Compounds, and Silver Alloys). The cathode silver can be further refined by chlorination.

Bismuth Drosses. Bismuth dross (see Section Dezincing) can be processed by either pyrometallurgical or electrolytic methods. In either case, the reagents calcium and magnesium are first removed, so that the subsequent operations are always performed on a pure lead – bismuth alloy.

Lead is removed by selective chlorination of the lead – bismuth alloy with chlorine gas, lead being converted to lead chloride with virtually no loss of bismuth. The rate of lead chlorination in lead – bismuth alloys with > 85 % bismuth is extremely high at 700 – 800 °C. In practice, chlorine is injected into an alloy already enriched to this level, and lean material is simultaneously added. Operation is thus continuous and material throughput is constant. The crude bismuth usually has a low foreign element content (copper, noble metals), which is removed by a process similar to that used for the refining of lead bullion.

In comparison with the pyrometallurgical technique, the electrolytic method has the advantage that the lead is deposited in pure form at the cathode (generally as fine lead). The anode slime from electrolytic lead – bismuth separation is subjected to reduction melting; the product is an alloy containing > 90 % bismuth along with lead, copper, and noble metals. This alloy can be processed pyrometallurgically (selective chlorination) or by bismuth electrolysis.

5.2. Electrolytic Refining of Lead Bullion

The electrolytic refining of lead bullion from soluble anodes has been practised for years in a number of large plants. Because the poor solubility of lead sulfate rules out sulfate electrolysis, the only candidate electrolytes are the few readily

soluble lead salts. In practice, solutions used have been restricted to the lead salts of fluosilicic acid (H_2SiF_6), fluoboric acid (HBF_4) and amidosulfuric (sulfamic) acid (H_2NSO_3H). The most important of these is the fluosilicate electrolyte.

Metals with a higher electrochemical potential than lead (silver, gold, copper, bismuth, antimony, arsenic, and germanium) remain essentially undissolved and accumulate in the anode slime, which is processed to recover them. Metals with a lower potential, such as iron, nickel, and zinc do go into solution, but the feed lead bullion contains such low levels of them that enrichment in the electrolyte is insignificant. The deposition potential of tin, – 0.140 V, is close to that of lead, – 0.126 V. Tin is therefore deposited along with lead from both fluosilicate and fluoborate electrolytes. The amidosulfate electrolyte, on the other hand, allows separation of lead and tin because of the low solubility of tin sulfamate.

The undissolved impurities, if they are present in large enough quantities, form a porous skeletal coating (slime). The slime adheres to the anode and retains its initial form, so that cells need not be regularly cleaned and deslimed. The mechanical strength of the slime depends on the content of foreign metals in the anode. Higher contents of bismuth and antimony enhance strength; increasing copper leads to a dense slime, so that the feed must be decoppered to <400 g/t copper. The production of adherent anode slimes requires a minimum antimony and bismuth content of 1 – 1.2 %. The cell voltage rises with increasing thickness of the slime coating and may reach up to twice the initial voltage; thus, energy consumption imposes an economic limitation. In addition, the increasing voltage can lead to dissolution of antimony, arsenic, and bismuth from the slime. The residence time of the anodes in the cell, is governed by this rise in cell voltage and thus by the concentration of impurities in the anode lead.

In industrial electrolytical lead refining, the *Betts process,* an aqueous solution of fluosilicic acid and lead fluosilicate ($PbSiF_6$) is electrolyzed. One liter of electrolyte contains 60 – 100 g lead as fluosilicate and 70 – 120 g free H_2SiF_6. Glue and/or other organic inhibitors (300 – 500 g per tonne of electrolytic lead) are added to obtain smooth deposits. Electrolyte losses result from the decomposition of H_2SiF_6 and entrainment in the anode slime.

The consumption of H_2SiF_6 is around 2 kg per tonne of electrolytic lead. The lead content of the electrolyte rises slowly because of the difference in current efficiency between anode and cathode. This increase is sometimes offset by precipitation of lead sulfate after addition of sulfuric acid. However, the lead content is usually held at the desired level by electrolysis in separate cells with insoluble graphite anodes. The following reactions take place:

Anode	$H_2O \longrightarrow 2\,H^+ + 1/2\,O_2 + 2\,e^-$
Cathode	$PbSiF_6 + 2\,H^+ + 2\,e^- \longrightarrow H_2SiF_6 + Pb$
Total	$PbSiF_6 + H_2O \longrightarrow H_2SiF_6 + Pb + 1/2\,O_2$

Cathode blanks are made of electrolytic lead and are 0.8 – 1 mm thick. They are fabricated either by casting over an inclined plane (older procedure) or with a water-cooled drum that dips into the lead bath. One or two cathode changes are generally required per anode. Current densities are 150 – 240 A/m^2 at the anode and 130 –220 A/m^2 at the cathode. Current efficiency is 85 – 95 % and energy consumption about 190 kWh per tonne of electrolyte lead. After electrolysis has been completed the anodes have about 30 – 40 % of their starting weight and are melted down into new anodes. The cathodes are thoroughly cleaned, melted down, refined with soda in a kettle to eliminate residual antimony, arsenic, and tin, and cast into ingots.

The size and weight of the electrodes used for lead electrolysis at the Cominco lead refinery in Trail, Canada are given in Table 5. Other important data follow [28]:

Lead electrofining data, Cominco Trail plant

Electrorefining cells	
Number	800
Dimensions (length×width×height), mm	2750×800×1250
Capacity, m^3	2.7
Material	polymer-containing concrete
Typical operating data (based on lead production of t/a 120 000)	
Anodes per cell	24
Cathodes per cell	25
Anode to anode spacing, mm	110
Current density, A/m^2	
Cathodes	130 – 220
	150 – 240

Anodes	
Current efficiency, %	80 – 90
Refining cycle, d	5
Cell voltage,	0.3 – 0.5
Power consumption for electrolysis, kW · h/t	168
Power consumption for other purposes, kW · h/t	50
Hydrofluosilicic acid consumption, kg/t	2
Aloes extract consumption, kg/t	0.17
Lignin sulfonate consumption, kg/t	0.24
Electrolyte temperature, °C	35 – 42
Electrolyte circulation to cell, L/min	20 – 25
Analyses	
Anode, wt %	1.0 – 1.2 Sb, 0.3 – 0.6 As, 0.04 – 0.07 Cu, 0.1 – 0.2 Bi, 0.1 – 0.6 Ag
Cathode, g/t	3 Cu, 8 Bi
Anode slime, wt %	40 – 50 Sb, 20 – 30 As, 10 – 20 Pb, 5 – 15 Bi, 5 – 15 Ag
Electrolyte, g/L	60 – 70 Pb, 80 – 100 H_2SiF_6

Table 5. Size and weight of anodes and cathodes used at the Cominco lead refinery

Immersed dimensions	Starter cathode	Plated cathode	Starter anode	Corroded anode
Length, mm	900	910	860	850
Width, mm	660	680	660	650
Thickness, mm	1	20	30	10
Area, m^2	1.2	1.25	1.2	1.15
Approximate weight, kg	6	140	200	60

Processing of Anode Slime. The composition of the anode slime depends directly on the impurities in the anode lead. In general, anode slime contains not only noble metals but also high levels of bismuth and antimony. Different compositions are seen at plants that have modified their operation for bismuth recovery. For example, at the Norddeutsche Affinerie, the anode lead consists of bismuth drosses and high-bismuth lead that has gone through the entire thermal refining process except for Kroll – Betterton debismuthizing.

No single processing method has been devised for such diverse intermediate products. By way of example, Figure 25presents the flowsheet of the Cominco refinery in Trail. The combination of proven pyrometallurgical processes yields product lead, dore (gold – silver), antimonial (hard) lead, and bismuth metal. The complexity of anode-slime processing places some restrictions on the inherently elegant electrolytic refining technique.

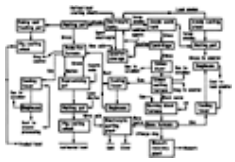

Figure 25. Flowsheet for electrolytic lead refining and slime treatment of Cominco Ltd. (Trail, Canada)

[Top of Page]

6. Recovery of Secondary Lead from Scrap Materials

Charles A. Sutherland and Edward F. Milner

The total world production of refined lead in 1987 was estimated to be 5.6×10^6 t. Of this approximately 60 % was primary lead from lead ores and concentrates; the remainder (40 %) was secondary lead produced from scrap materials. As the consumption of lead for sheet, pipe, and cable sheathing has declined and consumption for lead – acid storage batteries has increased, the secondary lead industry is primarily concerned with the processing of scrap batteries. In recent years the industry in the United States, and to a lesser extent in Europe, has experienced new, more stringent environmental regulations, and very low lead prices in the mid-1980 s. These factors have brought about the closure of a large number of small-scale secondary processing operations, and provided the impetus for the development of improved technology. Besides improving plant working conditions and reducing ambient lead emissions, an important objective of new environmental regulations is to force the efficient recycling of lead – acid batteries and prevent disposal in municipal waste sites. The potential generation of acetic acid and contamination of ground water with soluble lead acetate is the basis for this regulatory pressure. In the Federal Republic of Germany the efficiency of spent battery recycling is claimed to be 90 %.

6.1. Battery Types and Composition

Lead – acid storage batteries (⟶ Batteries – Battery Properties and Characteristics) are of three general types:

1. SLI (Starting, Lighting, Ignition) batteries, the bulk of which are automotive;
2. traction vehicle batteries; and
3. stationary batteries for supplying emergency power.

The key components in all these types are:

- a polypropylene containment box
- grids, connectors, and poles of lead alloy
- electrode paste: mixture of lead sulfate and lead oxides
- grid separators of poly(vinyl chloride) (PVC)
- sulfuric acid

The approximate composition of these materials on a weight percentage basis is:

Lead alloy components	21
Lead oxide	16
Lead sulfate	24.5
Water and sulfuric acid	24
Polypropylene	7.7
Poly(vinyl chloride)	3.8
Others	3.0

6.2. Battery Breaking and Processing Feed Preparation

Scrap battery preparation has progressed through four levels:

1. Acid drainage followed by pyrometallurgical processing of the entire battery and case. This blast furnace smelting approach became increasingly difficult with the change from bakelite to easily-fused polypropylene battery cases [27].
2. Separation of acid and case, both of which went to waste as gypsum (after neutralization with lime) and plastic garbage, respectively. The remaining mixture of poles, grids, grid paste, and PVC separators became secondary smelter feed.
3. Mechanical shredding of batteries followed by heavy medium separation to produce
   - Drainage acid for sale or limerock neutralization to produce gypsum as waste
   - Clean polypropylene for sale and recycle
   - A clean metallic fraction which can be smelted at low temperature
   - Battery paste mixture of $PbSO_4$ and PbO_x
   - Clean PVC plastic to waste or recycle
4. The same as (3) except that battery paste and drainage acid are reacted with sodium carbonate to produce dry crystals of sodium sulfate, a saleable product [28]. This system has several advantages. The sulfur content of the lead-bearing feed materials is largely eliminated, thus simplifying the subsequent smelting processes and reducing the amount of waste sulfur and lead-bearing slag. Figure 26 shows the flowsheet of the Tonolli CX process as an example.

Figure 26. CX scrap battery components recovery plant

6.3. Smelting of Battery Scrap Materials

The metallurgical equipment used for treating battery scrap consists of blast furnaces, reverberatory furnaces, rotary kilns, and short rotary furnaces.

Blast Furnace Treatment of Battery Scrap [29-32]. Typical blast furnace charge consists of cased batteries and battery plates with paste, limestone flux, scrap iron, metallurgical coke, and miscellaneous recycle drosses and slags.

Blast furnaces used for battery scrap smelting are smaller than primary lead industry blast furnaces. Typical hearth areas are in the range of 2 – 5 m^2. Smelting products are lead bullion containing most of the antimony content of the charge, iron matte containing most of the sulfur in the charge, slag, and flue dust which must be melted or agglomerated for recycle.

Plastic materials in the charge can, through fusion, impede the flow of material in the furnace shaft. This adverse effect is controlled by recirculating crushed slag in the furnace charge. Volatile organics in the off-gas must be combusted in an afterburner to prevent baghouse operating problems.

With advanced battery feed preparation systems (see Section Battery Breaking and Processing Feed Preparation), the role of the blast furnace in the industry is expected to decline in favor of reverberatory or rotary furnace reduction of lead carbonate – oxide sludge and metallic grid products.

Rotary Kiln Smelting of Secondary Lead. Continuous processing with long rotary kilns was developed by Preussag AG, Goslar, Federal Republic of Germany, and is practised in several other locations [33]. A typical equipment layout for this type of plant is shown in Figure 27. The kiln charge consists of a plastic-free mixture of sulfate, lead posts, grids, metallic iron, soda ash, coke breeze, and recycle flue dust. Products are lead bullion and a mixed matte – slag material for discard.

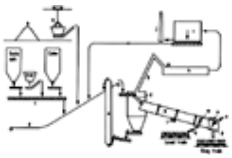

Figure 27. Rotary kiln scrap battery smelting plant

a) Buffer storage building; b) Bucket crane; c) Flux feed conveyor; d) Main feed conveyor; e) Bucket elevator; f) Feed chute; g) Vee duct; h) Balloon screw; i) Baghouse; j) Stack; k) Kiln; l) Kiln drive; m) Tap hole; n) Burner head; o) Burner; p) Slag chute; q) Lead launder

Short Rotary Furnace Smelting. The short rotary furnace is extensively used for smelting battery scrap materials [34]. Typical furnace dimensions are a diameter of 3.5 – 4.0 m and a length of 4 – 5 m. Reverse-flow firing is usually employed with the combustion burner and gas offtake at one end of the furnace. The charge components and general metallurgy are very similar to the rotary kiln process. The main difference is batch rather than continuous operation, which allows the use of different charges for the production of high- or low-antimony bullion products.

Reverberatory Furnace Smelting. The reverberatory furnace, although generally less productive than rotary furnaces, provides good control of the reduction potential required to produce low-antimony bullion and retain antimony in slag for subsequent reduction to high-antimony bullion. Reduction of desulfurized battery paste is also readily achieved at low temperature in reverberatory furnaces.

6.4. Future Trends

Increased environmental pressure is expected to continue to force innovation in secondary lead recovery. Slags produced by thermal reduction processes are classified as hazardous waste in the United States. Further electric furnace reduction of slag is now being practiced by the RSR Corporation in the United States to make slags acceptable for non hazardous disposal.

The lead – acid (SLI) battery industry is changing to sealed low-maintenance designs which require high-purity lead for the calcium – lead alloy grids. The secondary industry has been forced to produce an increasing percentage of high-purity lead compared to the grades of antimonial lead produced in the past. Removal of antimony by kettle – caustic softening is costly. The incentive to separate battery scrap into low-antimony paste and high-antimony metallic components for separate processing is thus expected to increase.

The search for nonthermal processes for high-purity lead recovery, particularly from desulfurized battery paste, is expected to intensify. Research work by the U.S. Bureau of Mines [35], and the RSR Corporation in the United States and Technologie Tonolli (Italy) on a fluosilicic acid leach – electrowin method of treating lead carbonate – oxide paste is expected to be used commercially in the early 1990 s.

[Top of Page]

7. Uses of Lead

Charles A. Sutherland and Edward F. Milner

The trend in end-use consumption of lead over the period 1977 – 1987 is illustrated in Table 6 for the United States, Japan, and Western Europe. A more thorough breakdown compiled by the Lead Development Association and the Lead – Zinc Study Group for presentation to the " Metals 2000 " Conference is drawn upon for the following analysis.

Table 6. Consumption of lead by end-use (%)

Use	United States		Japan		Europe *	
	1977	1987	1977	1987	1977	1987
Batteries	59.8	77.5	49.6	68.7	38.4	44.6
Cables	0.9	1.6	10.1	1.7	11.0	7.0
Semifinished	7.0	6.7	8.2	3.0	16.8	17.5
Pigments and chemicals **	21.0	7.6	15.0	16.2	25.4	22.3
Alloys	7.2	4.4	5.1	4.6	4.4	3.4
Other	4.1	2.2	12.0	5.8	4.0	5.2
Total	100.0	100.0	100.0	100.0	100.0	100.0
Total, 10^6 t	1.436	1.230	0.334	0.378	1.164	1.131

* Federal Republic of Germany, France, United Kingdom, Italy.
** Includes antiknock compounds.
Source: Metallstatistik 1977 – 1987.

Lead batteries were responsible for 60 % of refined lead consumption in the Western world in 1987 (ca. 2.5×10^6 t). The battery market is by far the largest end use of lead. It has shown a steady growth rate of 6 % from 1984 to 1989, corresponding to the increase in the number of automobiles worldwide. Automotive batteries for starting, lighting and ignition (SLI), traction batteries, and stationary batteries account for 75, 10, and 15 % of total battery lead consumption, respectively.

SLI batteries represent the predominant use of lead. In 1987 about 160×10^6 batteries were used in new cars and for replacements in a world car population of over 500×10^6. The number of vehicles is expected to increase, and so should lead use in batteries. The content of lead and lead compounds in battery walls, grids, and coatings has been steadily reduced to improve performance, minimize weight, and extend lifetimes. This has contributed to much better operating economics. The vehicle battery market is obviously vulnerable to substitution by other elemental couples and chemistry, but so far no competition appears likely to have a serious impact. Other factors that could affect future SLI battery use are increasing power requirements of cars and haulage vehicles, which would demand more powerful batteries and the possible use of higher voltages.

Traction batteries are used to power electric vehicles which have low running costs and are quiet and pollution-free. The vehicles are commonly used in short-range commercial delivery applications, in airports, and as fork lifts. The heavy weight and limited capacity of the conventional lead – acid battery seriously restrict vehicle speed and operating range between charges. It is unlikely that the lead battery will power a successful design of passenger vehicle.

Stationary batteries are used for standby and emergency power supply. Besides emergency lighting, special applications involve computer installations, process control systems, and emergency power in aircraft and trains. The market is growing. A new market is developing for small, sealed cells used in safety and security systems and personal computers.

Pigments and Lead Compounds. After batteries, the next largest use of about 14 % of refined lead is in glass making, ceramic glazes, plastic stabilizers, and paints. Lead glass has desirable optical, electrical, and radiation shielding properties, and lead crystal glass in particular is in great demand (⟶ Glass). Although ceramic glazes for tiles, tableware, and fine china have declined in application because of concern about lead's toxicity, they remain a significant part of the compound market. Lead compounds used for stabilizing PVC against temperature and UV light are finding intensive competition from other, more environmentally acceptable stabilizers. The lead pigment market is now very small.

Semifinished Products. Rolled and extruded lead, mainly sheet and pipe, account for about 8 % of total lead use. Piping was formerly the greater part of this sector, but toxicity concerns and substitution by plastics and copper have reduced the market considerably. The use of lead sheet for roofing, restoration of old buildings, sound insulation, radiation shielding, and chemically resistant linings is now greater than that for piping and is growing steadily.

Cable Sheathing. A major area of lead use was in the protection of electrical and telecommunications cables, but this market has declined drastically since the early 1960s due to substitution by aluminum and plastics. Nevertheless, this application still accounts for about 5 % of total lead use, and there now appears to be a return to thin lead sheathing as a total moisture barrier in association with plastic sheathing.

Lead Alloys. Solders, type metal, bearing alloys containing lead still account for about 4 % of the lead consumption (⟶). Type metal usage has nearly disappeared, but the bearing metal market is still reasonable in spite of some substitution. Uses of lead-base solder are growing, as electronic products find increasing markets which outweigh the trend to miniaturization and inroads made by tin.

Gasoline Additives. In 1972 the use of tetraethyl lead for improving the octane rating of gasoline was nearly 400 000 t/a. Since then, this sector has declined dramatically by 70 %, principally in the United States. The reason has been the proscription of lead antiknock compounds by environmental legislation. The tetraethyl lead market, now 3 % of lead use, can be expected to disappear with the eventual departure of older vehicle engines that require leaded gasoline.

Shot and Ammunition. Shot and ammunition account for about 2 % of lead consumption and represent a steady market. Lead shot is used in alloying steel and brass to free machining metals, and for shotgun cartridges.

Miscellaneous. Uses that do not fall into the above categories account for about 4 % of lead use, and include products featuring the high density of lead such as wheel weights, yacht keels, ornamental items, stained glass, and massive radiation shielding.

Much work is being done by lead concerns to identify and develop new applications for the metal. Promising new uses include load levelling battery banks used by utilities or by major power consumers to reduce high-cost peak power consumption; lead containers for disposal of radioactive waste; and asphalt stabilizers for retarding road paving deterioration.

[Top of Page]

8. Economic Aspects

Charles A. Sutherland, Edward F. Milner, Herbert Teindl

8.1. Lead Concentrate Schedules

Most lead smelters operate with a concentrate supply from associated mines. A few " custom " smelters (e.g., at Hoboken) purchase essentially all feed materials on the open market. In rare cases a smelter will agree to " toll " a concentrate, whereby the concentrate supplier pays for the processing of his material as part of the smelter's overall charge, and receives the recovered metal for his own use.

Lead concentrate from the beneficiation of lead – zinc or lead – zinc – copper ores is the primary feed for blast furnace and flash smelters (see Section Ore Concentration). Smelters employing the Imperial Smelting process generally use mixed lead – zinc concentrates. Conventional lead concentrates consist of galena in a sulfide matrix; the lead content is usually well over 60 %. The concentrate may also contain zinc and iron sulfides; gangues such as silicon dioxide, calcium oxide, magnesium oxide, and aluminum oxide; and a number of minor metal impurities that may be partially or totally recovered and which may also cause smelting or refining problems. The price assigned to a concentrate by a smelter reflects its own priorities and the suitability of the concentrate for the operation, as regards its valuable components and impurities. Smelters reject concentrates with elements that adversely affect metallurgy or refining practice. Examples are rejection of a high-sulfur (low-lead) concentrate by a down-draft sintering operation or of a bismuthic concentrate by a refinery without a capability for bismuth removal and recovery.

In compounding the aggregate feed and in selecting concentrates for purchase, lead smelter metallurgists and ore buyers balance the inputs of all valuable and harmful elements to create a charge that can be handled effectively and consistently within the limits of the operating processes, and with maximum profit. The pattern is set by the major concentrate supplies, usually from in-company sources: other raw materials are obtained to fit objectives of output, quality, cost, and timing.

A lead smelter's purchase schedule for lead concentrates comprises identifications and rules that govern payments for certain components, penalties levied on the presence of others, occasional bonus payments or penalties for desirable or undesirable combinations, and a treatment charge designed to cover smelting and refining costs and allow for profit. Most smelters can make available a general or " open " schedule set out in sufficient detail to cover a wide range of raw materials, but this is applied only to small lots. Major longterm purchases are normally made on the basis of a negotiated schedule. The schedule of purchase terms also covers sampling, assay of constituents and moisture, settlement of disputes, reference metal prices for determination of paid values, currency exchange, escalation of costs and prices, and payment terms.

Through its applied schedules, a lead smelter always pays for lead, silver, and gold, often for copper and zinc, sometimes for bismuth, and very rarely for antimony, platinum, palladium, or indium. Lead payment is offered on about 95 – 97 % of the contained lead at a defined lead price. A blast furnace or flash smelter prefers lead content > 60 %, but an ISP smelter accepts lower lead content, particularly with associated substantial copper value. Silver and gold are routinely recovered at high efficiency by lead refineries, and are paid for at current prices subject to minimum deduction from the assayed content, reflecting the minor losses, and to a treatment charge.

A payment may be offered for contained zinc if the smelter processes high-zinc charges by slag fuming but additional costs for fluxes and fuel are assessed against the payment. The ISP smelters routinely pay for zinc subject to minimum deduction and a treatment charge. Iron is a major component of lead smelting slag and is of particular value in fluxing zinc, but may be present in excess of requirement and would then incur a modest penalty.

Most blast furnace and ISP smelters pay for copper, which is usually recovered in drossing as matte, subject to minimum deduction and treatment charge. ISP smelters can handle higher copper inputs than other lead smelters.

Arsenic is always heavily penalized under lead smelter schedule terms from levels as low as 0.05 %; this reflects the uncompensated problems arsenic creates in refining, hygiene, and environmentally acceptable disposal. Antimony is usually neither paid for nor penalized.

Bismuth's treatment in a lead smelter schedule depends on whether the refinery can separate and recover the metal comprehensively (as in electrolytic refining) or by pyrometallurgical debismuthizing at high unit cost. More often it would be a reason for rejection of the concentrate. Certain smelters can recover tin, indium, cadmium, and thallium in the sintering, drossing, and refining circuits but these elements are never paid for. Thallium may be the cause of rejection by an electrolytic

refinery and tin by an ISP zinc – lead smelter.

Lime and silica are desirable fluxes in lead smelting and are usually assigned minor credits in a lead schedule. Blast furnace smelters may penalize alumina for its deleterious effect on slag; ISP smelters can take higher aluminum oxide but are concerned with magnesium oxide content.

Other materials that are treated in lead smelters for lead and other values are scrap lead, scrap batteries, lead ashes and sludges, and zinc smelter leach residues. Metal and battery scrap are the normal feeds for secondary smelters under specific purchase terms. Zinc residues contain substantial lead and silver values in addition to substantial undissolved zinc, and are handled at high cost in additional fuel, fluxes, and slag fuming furnace operation in joint lead – zinc smelters such as the one at Trail.

8.2. Lead Statistics

World mine production of lead has been steady at (3.4 – 3.5)×10^6 t/a of contained lead since the early 1970s. Production since 1925 in thousands of tonnes is as follows:

1925. 1501
1950. 1686
1960. 2376
1970. 3433
1980. 3470
1985. 3581
1986. 3378
1987. 3389
1988. 3400
1989. 3319
1990. 3340
1991. 3331

The estimates of world lead resources shown in Table 1 represent a comfortable reserve against expected consumption requirements for the future, and there are excellent prospects for further economic discoveries.

Production (primary and secondary) and consumption of refined lead in the last ten years are shown in Table 7, and by area and trading bloc for 1987 in Table 8. About one-third of refined lead production is from recycled materials.

Table 7. World production (primary and secondary) and consumption of refined lead

Year	Production, ×10^6 t	Consumption, ×10^6 t
1977	5.388	5.466
1985	5.634	5.460
1986	5.464	5.551
1987	5.631	5.623

Source: Metallstatistik 1977 – 1987.

Table 8. Lead statistics for most important lead producing and consuming countries, 1987

	Mine production (contained Pb), 10^3 t	Refined production (primary and secondary), 10^3 t	Refined consumption, 10^3 t
Federal Republic of Germany	24.5	340.4	344.6
France	2.2	245.5	207.5
Italy	12.2	168.3	244.0
United Kingdom	0.7	347.0	287.5
former Yugoslavia	82.0	124.6	129.4
Spain	83.2	122.7	105.8
Rest of Europe	156.1	265.4	311.8
Total Europe	360.9	1613.9	1630.6
Japan	27.9	338.5	377.9
Rest of Asia	111.8	244.2	407.3

Total Asia	139.7	582.7	785.2
Africa	214.6	159.9	111.4
United States	318.3	1027.9	1202.8
Canada	413.4	225.8	102.9
Mexico	177.1	185.1	99.6
Peru	192.0	70.8	21.9
Other South America	73.4	140.7	162.6
Total America	1174.2	1650.3	1589.8
Australia, Oceania	486.2	220.7	65.0
Total Western World	2375.6	4227.5	4182.0
former Soviet Union	510.0	780.0	775.0
Other Eastern Bloc	503.7	623.9	665.5
Total Eastern Bloc	1013.7	1403.9	1440.5
Total World	3389.3	5631.4	5622.5

Source: Metallstatistik 1977 – 1987.

Lead ranked first in production among nonferrous metals in the first quarter of the twentieth century, but now ranks fourth after aluminum, copper, and zinc. This is due more to consumption growth for the other metals than to decline in lead use. Lead has suffered the loss of some major markets in pigments, water piping, and lead alkyls caused by legislation reflecting concern for its toxicity and environmental impact, from substitutions by plastics and aluminum, and the decrease in requirement for type metal. These losses now appear to have stabilized. They have been more than offset by the growth in production of lead batteries for automobile and transport use. The lead battery represents the dominant use of production and recycled lead and is the mainstay of the lead industry.

Of very considerable concern to the industry has been the slump in lead price, which reflects but far transcends the worldwide industrial recession in the first half of the 1980s (Table 9). The price of lead reached its lowest point in recorded history in terms of the real value of money in 1985. The period from 1982 to 1986 saw dismally inadequate returns and outright losses, with the abandonment of many producing mines and severe retrenchment or closure of smelting plants. Since the mid 1980s the price of lead has recovered by over 50 %. This has encouraged many lead producers to modernize or replace old smelting plants by the new flash-smelting technologies that are better able to meet the stringent hygiene and environmental regulations that govern plant operation.

Table 9. Annual average producer price of lead in the United States

Year	Price (U.S. cents/kg)
1978	74.28
1979	116.20
1980	93.73
1981	80.57
1982	56.38
1983	47.86
1984	56.40
1985	42.10
1986	48.74
1987	79.20
1988	81.08

Source: Metallstatistik 1977 – 1987.

[Top of Page]

9. Toxicology and Occupational Health

Hermann M. Bolt

Lead is one of the seven metals used in antiquity by the Eurasian civilization. Its uses were varied. Physicians prescribed various forms of lead to heal ailments ranging from constipation to infectious diseases such as the plague. Lead was also used to preserve or sweeten wine, for water pipes, cosmetics, and ceramic glazes.

The symptoms of lead poisoning have been recognized for centuries. HIPPOCRATES described lead colic in 370 B.C. and PLINY reported poisoning among lead workers in about 50 A.D. In 1700 RAMAZZINI [36] noted the illness of potters who worked with lead. In 1933 KEHOE et al. [37] brought to light the important consideration that lead is very widely found in the environment. Their studies showed that although lead is not an essential element for the body, a measurable amount exists in all adult body tissues and fluids.

There is considerable literature on the effects of lead, ranging from overt signs and symptoms of lead poisoning to subtle effects such as reduced intelligence quotient scores, behavioral effects, and impaired hearing. The relevance of these effects and the definition of lead exposure levels are subjects of extensive debate.

9.1. Sources of Lead Exposure

Many sources of lead exposure affect the workplace and the general population.

9.1.1. *Drinking Water*

Drinking water can be a significant contribution to lead intake, especially for young children.

Lead can enter the aquatic system through fallout and surface runoff, from natural sources, or from lead-lined tanks, lead pipes, and lead solder joints.

The lead content of drinking water is controlled in the EC and North America by water quality standards, and has been subject to revision [39].

Lead concentrations in drinking water exceeding 300 µg/L were reported in the 1980s [40]. Such high contents can be attributed to very soft water, low pH, and high temperature; for example, samples of soft, hot, or stagnant “first draw” water exposed to lead or lead in pipes usually have a high lead content.

9.1.2. *Atmosphere*

Worldwide natural lead emissions to the atmosphere from natural sources have been estimated to be 24×10^3 t/a, compared to $(350 – 450) \times 10^3$ t/a from anthropogenic sources. By far the greatest contribution comes from automotive lead emissions. A determined effort has been made in most countries to reduce the amount of allowable lead in gasoline. The lowering of lead in gasoline that started in the USA has been cited by the U.S. Center for Disease Control (CDC) as the major reason for the 37 % decrease in childrens’ blood lead levels, based on a nationwide survey conducted between 1976 and 1980 [42]. Another view was that the effect is attributable mainly to socioeconomic improvements. Lead emissions from industrial sites such as smelters and metal recovery operations also contribute to atmospheric lead. However, fallout from these sources is localized and is estimated to contribute only about 10 % of the overall lead emission. Ambient lead levels in the proximity of primary and secondary smelters can be significant. At El Paso, Texas, mean lead levels in 1971 ranged from 0.46 to 2.72 µg/m^3 with peak values as high as 22 µg/m^3 [43]. Sampling data from former Yugoslavia recorded peak levels in excess of 200 µg/m^3 [44]. Typical levels near point source areas are 0.1 – 1 µg/m^3 and for rural areas ≤ 0.1 µg/m^3. Community air quality criteria are set in the range of 0.5 – 2.0 µg/m^3 with a view to lowering these levels.

9.1.3. *Soil*

Natural soil can contain from 2 to 200 ppm lead and can be greatly influenced by anthropogenic sources. There are few standards for lead in soil; however, many countries have guidelines with suggested maximum levels of ca. 500 – 1000 ppm.

Lead-bearing dust can be especially dangerous to very young children because of a high level of ingestion through their play and eating habits. The concentration of lead in street dust and surface soil can be very high, especially if located near a major freeway. Soil samples from areas of Vancouver, Canada taken near heavy traffic areas were similar to samples taken 1.6 km from a large primary lead smelter complex (1545 and 1662 ppm, respectively) [45]. A survey of street dust in 77 Midwestern U.S. cities showed a mean lead content of 1600 ppm in residential areas and 2400 ppm in industrial areas [46].

9.1.4. *Diet*

Lead (mainly organic compounds) can enter the human body through the skin, by inhalation, or by ingestion. Ingestion in food is generally the major source of lead intake. Lead in the diet can come from numerous sources such as uptake of lead into plants; deposition onto vegetables; from soil or water; and from lead solder in canned goods.

Absorption of lead from the gastrointestinal tract by adults is estimated to be about 5 – 10 %, but 25 – 55 % by young children [47]. Ingestion of lead-based paint by children is still considered the most frequent cause of lead intoxication in the USA [48]. The amount of lead absorbed through the digestive system is greatly influenced by stress, fasting, and the solubility of the lead compound; for example, lead acetate is 25 times more soluble in intestinal fluid than lead sulfide [49].

9.1.5. *Occupational Exposure*

Workers employed in primary and secondary lead smelters, as well as in the production of lead metal and lead compounds, are exposed to lead absorption hazards. Because of the fine particle size of lead fume and its solubility, it can readily enter the respiratory system and be absorbed in the blood stream.

The workplace lead limits in the EC and North America have been tightened to the current range of 0.05 – 0.1 mg/m^3. The MAK value and the OEL recommended by the E.U. Scientific Committee for Occupational Exposure Limits (SCOEL) is 0.1 mg/m^3 and the TLV-TWA value (recommended by ACHIH) is 0.05 mg/m^3. This has been strong motivation for the development of new process technology, improvement of engineering controls, and stricter use of personal protective equipment by industry to reduce worker exposure.

9.2. Absorption and Excretion

Absorption of lead into the body is greatly affected by the size of the lead-containing particle and its solubility. Retention through the respiratory system is 20 – 60 % [50]. Particles of 0.1 – 1 µm are small enough to reach the alveoli, where absorption into the blood stream can take place; particles larger than 1 – 2 µm may be deposited in the nasopharangeal area, trachea, or bronchi and leave by swallowing or expectoration. Thus, fine lead fume has a much greater influence on a worker’s blood lead value than coarser dust.

Gastrointestinal absorption of lead is discussed in Section Sources of Lead Exposure. Absorption through the skin is not considered to be important for inorganic lead compounds, but organic lead compounds such as tetraethyllead and lead naphthenate are readily absorbed across the skin barrier.

Approximately 90 % of ingested lead is not absorbed and eliminated in the feces. Lead absorbed into the blood stream deports to soft tissue and gradually accumulates in bone. Some is excreted in urine. With consistent intake of lead, a steady state between lead concentrations in blood, soft tissue, and bone is established [51].

9.3. Effects

9.3.1. *Acute Effects*

The total average daily lead intake for adults in Western countries up to the 1980s has been estimated to be 150 – 300 µg, which is reflected in a blood lead level of 10 – 25 µg/dL. Blood taken from “normal” persons in 16 countries in the 1960s showed an average of 17 µg/dL. In the 1990s this has been reduced in Western countries to 5 – 10 µg/dL, mainly due to the introduction of unleaded gasoline.

9.3.2. *Hematologic Effects*

One of the more deleterious effects of lead is on blood and blood-forming tissue. Lead can inhibit almost all steps in the biosynthesis of hemoglobin [52].

9.3.3. *Neurological Effects*

In the early part of the 1900s, excessive prolonged exposure to lead was reflected in brain disease and muscle weakness in lead workers and children. Such extreme exposures have long been eliminated. Nowadays, the leading toxic effect of lead on the CNS in adults is impairment of performance in neurobehavioral tests. Most authors agree that a long-term Pb level of 40 µg/dL probably represents a NOAEL (no observed adverse effect level) or LOAEL (lowest observed adverse effect level) in this respect.

Other endpoints of lead toxicity such as the PNS are relevant only for exposure levels which are consistently higher.

Of great concern are neurological effects on infants. Neurobehavioral deficits have been correlated with lead content in teeth [53], and blood lead levels in children which were formerly considered to be acceptable are now being reviewed in many countries.

9.3.4. *Renal Effects*

The kidney is the major route for the excretion of lead that has been absorbed into the body, and it is also one of the organs affected by lead toxicity [54]. In most situations renal dysfunction occurring as a result of chronic lead exposure is reversible provided serious tissue damage has not occurred. However, irreversible renal damage may result from a prolonged, very high lead body burden. Hypertension can also follow.

9.3.5. *Reproduction Effects*

Toxic levels of lead can cause sterility in men, and miscarriages and spontaneous abortions in women. The fetus is especially vulnerable because lead can cross the placental barrier. There is considerable uncertainty concerning impairment of reproductive function by lead. Men with blood lead levels exceeding 60 – 70 µg/dL may have reduced sperm count and mobility [55]. In females, however, it is relevant that cognitive deficits of the offspring are dose-dependently associated with lead exposure. The question of reversibility of such deficits is not yet satisfactorily resolved. On the basis of the present data no definite NOAEL can be deduced, which strongly calls for a minimization of exposure.

9.3.6. *Carcinogenicity*

Lead and its compounds are under close international scrutiny with respect to its potential as cancer-causing agents. The IARC has listed inorganic lead compounds in Group 2B (agents that are possibly carcinogenic to humans), in generalization of observations on lead acetate and phosphate [56]. The observed experimental carcinogenicity of Pb salts is, in the first instance, directed towards the kidneys as the target tissue. Most probably, these effects are to a great extent based on the renal toxicity of high doses of Pb. Hence an OEL based on avoiding functional CNS alterations is expected also to protect against PNS and renal toxicity, including possible renal cancer development.

Similar conclusions may probably be drawn with respect to other systemic toxicities, e.g., on heme biosynthesis and blood pressure, although there is a discrepancy in opinions, and more research is needed in this direction.

9.4. Indices for Monitoring Lead Exposure

Biological monitoring of workers provides an assessment of overall exposure to a chemical or compound, and sets guidelines for the evaluation of potential health hazards. Two common indices for evaluation of overall lead absorption are outlined below.

9.4.1. *Lead in Blood*

Lead in blood is generally considered the best indicator of the lead concentration in soft tissue and is an effective means of assessing exposure to lead. Some investigators have been able to predict increases in blood lead after low lead exposures with some accuracy [57]. Blood lead levels of 60 – 70 µg/dL in male workers were considered acceptable up to the 1990s [58], but some authorities such as the WHO have proposed lower limits, of 40 µg/dL for males and 30 µg/dL for females. These biological limit values have also been recommended for occupationally exposed persons in the EU by SCOEL [60].

9.4.2. *Teeth and Hair*

Lead content in teeth, particularly dentine, has been used to assess exposure and measure psychological deficits in children

but is not practical for second-growth teeth. Hair has also been suggested as a method of evaluating lead absorption, but it is almost impossible to discern between intrinsic lead and that absorbed on the surface from external sources.

[Top of Page]

10. Control of Lead Emissions

Robert C. Kerby

The adverse effects of lead ingestion on the health of humans and animals (Chap. Toxicology and Occupational Health) make control of lead emissions a major operating challenge in all primary, secondary, and processing production industries. This overall problem has a dual focus:

1. The effect of emissions on in-plant atmospheric contamination and lead absorption by production workers.
2. The effect of plant emissions on ambient air lead levels and accumulative leady dust fallout on areas around the production site.

This challenge has long been part of the lead production industry, but since the 1970s improved diagnosis of biological problems related to lead and the resulting increased government regulatory pressure have greatly increased the demand for reduction of overall emission. Improved engineering controls and total process revisions are the broad approaches that have been taken to achieve compliance. Much progress has been made worldwide; however, the degree of success largely depends on the regulatory standards which have been set or proposed. In the United States, the proposed levels of 50 $\mu g/m^3$ for in-plant air and an annual mean of 1.5 $\mu g/m^3$ for lead in air outside the plant boundaries represent perhaps the most difficult industrial challenge.

10.1. In-Plant Lead Emissions

Engineering Controls. Assuming major process gas flows are handled by efficient dust removal equipment and discharged in high stacks, in-plant lead levels are mainly influenced by fugitive process emissions. Well-engineered ventilation systems are required to control dust and fume emissions from hot sinter or molten slags. The provision of effective ventilation is costly in terms of energy usage and capital cost.

Personal Respirators. The use of personal respirators by plant workers is almost universal practise to reduce respiratory lead intake from the plant atmosphere. The use of respirators is often construed as evidence of inadequate engineering control of fugitive emissions. Certainly the elimination of respirator use for most production jobs other than special maintenance work is a desirable objective, but, it is not being achieved in most plants at present. Personal hygiene habits are increasingly recognized as critical to controlling lead intake by workers. Measures such as provision of clean clothing, compulsory washing, no smoking or eating on the job, and clean eating facilities are common.

Process Revisions. The new oxygen smelting processes described in Section Direct Smelting Reduction Processes are attractive because they have inherent advantages for lead emission control. The traditional sintering and blast furnace method of lead production is a most difficult engineering challenge for achieving good control of in-plant emissions. The sinter operation involves large recirculating flows of hot, dusty, abrasive sinter requiring large ventilation volumes. Blast furnace reduction can be an erratic operation often troubled by a hot top and blowing and accompanying fugitive emissions from the gas collection system.

The new oxygen smelting processes have the following favorable characteristics:

1. Inherent lower process gas volume with oxygen usage giving less cleaning and reduced emissions
2. Recirculating loads requiring separate ventilation are minimized
3. Equipment is relatively compact and well sealed

The introduction of these new pyrometallurgical processes over the next one to two decades should have a major beneficial impact on the in-plant and ambient lead emission problems of today's primary lead smelters.

10.2. Ambient Lead Emissions

Ambient lead emissions originate from residual lead after cleaning of process and ventilation gas and general fugitive lead emissions. The accumulation of ambient lead fallout on vegetation eaten by cattle or the buildup of an objectional concentration of lead in the top soil is the most serious aspect of ambient emissions. The use of high-efficiency baghouses for cleaning process and ventilation gases, combined with curtailment of fugitive emissions of dust and fume from process equipment, roads, and stock yards, have made major reductions in ambient emissions at most primary and secondary lead smelters.

[Top of Page]

References

General References

1. W. Hoffmann: *Lead and Lead Alloys, Properties and Technology*, 2nd ed., Engl. Transl., Springer Verlag, New York 1970.
2. R. C. Weast (ed.): *Handbook of Chemistry and Physics*, 67th ed., CRC Press, Boca Raton, FL, 1987.
3. J. F. Smith, R. R. Kubalak (eds.): " Lead " in *Metals Handbook,* 9th ed., vol. 2, Metals Park, OH 1979.
4. *AIME World Symp. Min. Metall. Lead Zinc Proc.* 1970, St. Louis, MO.

5. J. M. Cigan, T. S. Mackey, T. J. O'Keefe (eds.): *Lead-Zinc-Tin '80*, AIME, Warrendale, PA, 1980.
6. U. S. Bureau of Mines, Mineral Industry Survey Lead Reports, 1980 – 1988.

Specific References

7. H. H. Kellogg, S. K. Basu, *Trans. Metall. Soc. AIME* **218** (1960) 70 – 81.
8. G. M. Willis: " The Physical Chemistry of Lead Extraction," in [5]
9. G. W. Healey: "Sulfur Elimination in Updraft Lead Sintering–A Thermochemical Model," *J. Met.* **33** (1981) 30 – 37.
10. J. T. Roy, J. R. Stone, *Trans. Metall. Soc. AIME* **227** (1963) 177 – 179.
11. W. F. Caley, S. G. Whiteway, A. B. Neil, P. J. Dugdale: " Viscometer for Lead Blast Furnace Slags," 14th Annual Conference of Metallurgists, Canada, 1975 paper 113.
12. H. Winterhager, R. Kammel, *Z. Erzbergbau Metallhüttenwes.* **12** (1959) 479 – 486.
13. G. W. Toop, presentation to 22nd Annual Conference of Metallurgists, Edmonton, Alta., Canada, 1983.
14. H. Olaska, O. Notzold in[4]
15. H. I. Elvander: "The Boliden Lead Process," *Pyrometallurgical Processes in Nonferrous Metallurgy*, AIME Metallurgical Society Conference, vol. 39, Pittsburgh 1965, p. 225.
16. A. Perillo, A. Carminati, G. Carlini: " The Kivcet Lead Smelter at Portovesme, Commissioning and Operating Results," AIME Annual Meeting, Phoenix, AZ, 1988, paper no. A88 – 2.
17. S. Petersson: " Autogenous Smelting of Lead Concentrate in TBRC," paper presented at 106th AIME Annual Meeting, Atlanta, Georgia, 1977, TMS-AIME paper no. A77 – 11.
18. E. O. Nermes, T. T. Talonen: "Flash Smelting of Lead Concentrates," *Metall (Berlin)* **36** (1982) no. 9, 1007 –1009.
19. A. I. Taskinen, L. M. Toivonen, T. T. Talonen: " Thermodynamics of Direct Lead Smelting," in H. A. Fine, D. R. Gaskell (eds.): *Int. Symp. Metallurgical Slags and Fluxes,* vol. 2, 1984.
20. W. J. Errington, J. H. Fewings, V. P. Keran: " The Isasmelt Lead Smelting Process," Transactions of the Institute of Mining and Metallurgy, Section C, 870 300, vol. 96, no. 3, pp. C 1 – C 6.
21. P. Fisher: " The QSL-Process: A New Lead Smelting Route Ready for Commercialization," Advances in Sulfide Smelting, AIME Conference Proceedings, San Francisco, Nov. 1983, vol. 2, Technology and Practice, pp. 513 – 527.
22. M. Wong, F. Haver, R. Sandberg: " Ferric Chloride Leach-Electrolysis Process for Production of Lead," in [5]
23. T. R. A. Davey: " The Physical Chemistry of Lead Refining," in [5]
24. J. H. Fern, R. W. Shaw: " Copper Separation and Recovery at the BHAS Lead Smelter," MMIJ/Aus IMM Joint Sendai Symp., Port Pirie, S. Australia, 1983.
25. J. A. DeKeyser, W. E. Jespers: " The Harris Refinery of S.A. Metallurgie Hoboken-Overpelt N.V.," AIME, TMS Paper Selection A 81 – 10 1981.
26. K. Emicke, G. Holzapfel, E. Kniprath: "Lead Refining at Norddeutsche Affinerie," *Erzmetall* **24** (1971) 205 – 215
27. BHASP Ltd., AU 502 470 1979 (D. H. Ward, J. D. Iley).
28. D. L. Thomas, C. J. Krauss, R. C. Kerby, AIME, TMS Paper Selection A 81 – 8, 1981.
29. S. Bergsoe, N. Gram: "Lead Smelting, Refining, and Pollution," *Lead-Zinc-Tin '80 (Eighty), Proc. World Symp. Metall. Environ. Control* 1980, 1023 – 1030.
30. " CX Scrap Battery Compounds Recovery Plant," Commercial Literature, Tecnologie Tonolli, Milano, 1986.
31. J. D. Taylor, P. J. Moor: "Secondary Lead Smelting at Britannia Lead Company Limited," *Lead-Zinc-Tin '80 (Eighty), Proc. World Symp. Metall. Environ. Control* 1980, 1 003 – 1 022.
32. M. Koch, H. Niklas: "Processing of Lead-Acid Battery Scrap, the Varta Process," Productivity and Technology in the Metallurgical Industries, TMS Cologna, Sept. 1989, pp. 495 – 500.
33. R. C. Egan, M. V. Rao, K. D. Libsch: "Rotary Kiln Smelting of Secondary Lead," *Lead-Zinc-Tin '80 (Eighty), Proc. World Symp. Metall. Environ. Control* 1980, 935 – 973.
34. J. Godfroi: " Five Years " Utilization of the Short Rotary Furnace in the Secondary Smelting of Lead," *Lead-Zinc-Tin '80 (Eighty), Proc. World Symp. Metall. Environ. Control* **1980,** 974 – 984.
35. E. R. Cole, A. Y. Lee, D. L. Paulson: "Update on Recovering Lead from Scrap Batteries," *J. Met.* **37** (1985) 79 – 83.
36. D. Hunter: *The Disease of Occupations*, The English University Press, London 1957.
37. R. A. Kehoe, F. Thamann, J. Cholak: "On the Normal Absorption and Excretion of Lead," *J. Ind. Hyg.* **15** (1933).
38. R. Amavis, J. Smeets, *Sci. Total Environ.* **18** (1981) 293 – 300. Links
39. EPA: " Reducing Lead in Drinking Water: A Benefit Analysis," EPA Report no. EPA 230 – 09 – 86 – 019, 1986.
40. D. Worth *et al.*: *Environmental Lead*, Academic Press, New York 1981. R. F. Lacey, M. R. Moore, W. N. Richards, *Sci. Total Environ.* **41** (1985) 235 – 257. Links
41. B. E. Davies: *Lead in the Home Environment*, Science Reviews, Norchwood 1987.
42. NHANES II— "Chronological Trend in Blood Lead Levels Between 1976 and 1980," *N. Engl. J. Med.* **308** (1983) 1373 – 1377. Links
43. *Air Quality Criteria for Lead*, Office of Research and Development, U.S. Environmental Protection Agency, Washington, DC, 1977.
44. International Symposium on Environmental Lead Research, Dubrovnik, former Yugoslavia, *Arch. Ind. Hyg. Tox.* **26** (1975) Supplement.
45. N. Schmitt *et al.*: "Surface Soil as a Potential Source of Lead Exposure for Young Children," *CMA J.* **121** (1979) 1474 – 1478.
46. *Lead, Airborne Lead in Perspective*, National Academy of Sciences, National Research Council, Washington, D.C., 1972.

47. R. A. Kehoe: “The Metabolism of Lead in Man in Health and Disease, The Normal Metabolism of Lead,” *J. R. Inst. Pub. Health Hyg.* **24** (1961) 81 – 96.
48. J. J. Chisolm: “The Continuing Hazard of Lead Exposure and Its Effects in Children,” *Neurotoxicology* **5** (1984) 23 – 42. Links
49. K. Horiuchi: “Lead in the Environment and Its Effect on Man in Japan,” *Med. J. Osaka Univ.* **16** (1970) 1 – 28. Links
50. K. Nozaki: “Method for Studies on Inhaled Particles in Human Respiratory System and Retention of Lead Fume,” *Ind. Health* **4** (1966) 118 – 123. Links
51. S. B. Gross: “Kehoe Balance Experiments, Oral and Inhalation Lead Exposures in Human Subjects,” *J. Toxicol. Environ. Health* **8** (1979) 333 – 337. Links
52. S. Hernberg, J . Kikkanen, G. Mellin, H. Lilius: “Delta-Aminolevulinic Acid Dehydrase as a Measure of Lead Exposure,” *Arch. Environ. Health* **21** (1970) 140 – 145. Links
53. H. L. Needleman *et al.*: “Deficits in Psychologic and Classroom Performance of Children with Elevated Dentine Lead Levels,” *N. Engl. J. Med.* **300** (1979) 689 – 695. Links
54. R. A. Goyer: Current Topics in Pathology, vol. **55**, Springer Verlag, Heidelberg 1971.
55. I. Lancranjan *et al.*: “Reproductive Ability of Workmen Occupationally Exposed to Lead,” *Arch. Environ. Health* **30** (1975) 396 – 401. Links
56. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Overall Evaluation of Carcinogenicity: An Updating of IARC Monographs, vols. 1 – 42, Suppl. 7, 1987.
57. A. C. Chamberlain *et al.*: “Uptake of Lead by Inhalation of Motor Exhaust,” *Proc. R. Soc. London B* **192** (1975) 77 – 110. Links
58. R. R. Lauwerys: “Industrial Chemical Exposure: Guidelines for Biological Monitoring,” *Section 10* (1983) 29.
59. A. A. Lamola, T. Yamane: “Zinc Protoporphyrin in the Erythrocytes of Patients with Lead Intoxication and Iron Deficiency Anemia,” *Science (Washington, D.C.)* **186** (1974) 936.
60. European Commission, Directorate-General for Employment, Industrial Relations and Social Affairs, Unit V/F.5 (2002): Occupational Exposure Limits. Recommendations of the Scientific Committee for Occupational Exposure Limits to Chemical Agents. Luxemburg: Office for Official Publications of the European Communities.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Lead Alloys

R. David Prengaman, RSR Corporation, Dallas, Texas 75247, United States

1. Introduction 1
2. Properties of Lead and Lead Alloys . 1
3. Lead – Antimony Alloys 1
3.1. Binary Lead – Antimony Alloys 1
3.2. Lead – Antimony – Tin Alloys 4
4. Lead – Calcium Alloys 4
4.1. Binary Lead – Calcium Alloys 4
4.2. Lead – Calcium – Aluminum Alloys . . 6
4.3. Lead – Calcium – Tin Alloys 6
5. Lead – Tin Alloys 7
6. Lead – Copper Alloys 8
7. Lead – Silver Alloys 9
8. Lead – Tellurium Alloys 10
9. Low-Melting Alloys 10
10. Reactive Lead Alloys 10
11. Occupational Health and Safety 10
12. References 11

1. Introduction

Lead is one of the oldest known metals. Its use was known before 3000 BC. All early civilizations used lead extensively for ornamental and structural uses. The use of lead pipes for the transportation of water by the Romans, one of the early uses of lead, endures to the present.

The primary and most important mineral for recovery of lead is galena, PbS. Lead ores are found on every continent, usually in combination with zinc, copper, and silver. The United States is the leading lead-producing country followed by Canada and Australia. Total annual production of lead and lead alloys in 1987 was about 4.3×10^6 t. Recycling of lead scrap, mainly from batteries, lead sheet, and cable sheathing, provided 2.1×10^6 t.

About 40 % of lead is used as pure lead, lead oxides, or lead chemicals; the remainder is used in the form of lead alloys. The major uses of lead alloys are: lead acid batteries, ammunition, cable sheathing, building construction (sheets, pipes, and solders), bearings, gaskets, specialty castings, anodes, fusible alloys, shielding, and weights.

2. Properties of Lead and Lead Alloys

Lead is a heavy, soft, bluish gray metal which has a low melting point and a high boiling point. The density, malleability, lubricity, flexibility, and coefficient of thermal expansion are quite high. The elastic modulus, elastic limit, tensile and compression strength, hardness, and electrical conductivity are relatively low. Lead has excellent resistance to corrosion in a wide variety of media. Lead is easily alloyed with many other metals. Because lead alloys have low melting points, they can be cast into many shapes by using a variety of molding materials and casting processes.

Lead is very ductile and malleable, and can be fabricated into various shapes by rolling, extruding, forging, spinning, and hammering. The low tensile strength and very low creep strength of lead make it unsuitable for use without the addition of alloying elements. The major alloying elements used to strengthen lead are antimony, calcium, tin, copper, tellurium, arsenic, and silver. Minor alloying elements are selenium, sulfur, bismuth, cadmium, indium, aluminum, and strontium.

3. Lead – Antimony Alloys

3.1. Binary Lead – Antimony Alloys

Properties. Lead – antimony is the most widely used lead alloy. The phase diagram is shown in Figure 1 [3]. It contains a eutectic point at 11.1 wt % antimony and 252 °C. The solubility of antimony in lead decreases from 3.5 wt % at 252 °C to 0.25 wt % at 25 °C; therefore, the alloys are age hardenable. The grain structure of a typical cast lead – antimony alloy (see Fig. 2) consists of a lead matrix surrounded

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_237

by a network of antimony particles. The eutectic particles provide high-temperature strength. Addition of arsenic dramatically increases the rate of aging and final strength. Addition of tin increases the fluidity and, in combination with copper and arsenic, reduces the rate of oxidation of molten lead – antimony alloys [4].

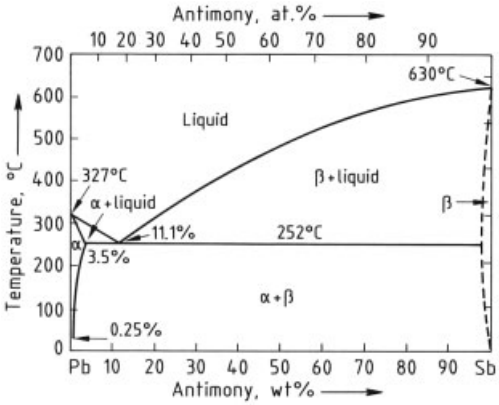

Figure 1. Phase diagram of the lead – antimony system [3]
Reproduced from ref. [3] by permission of the publisher.

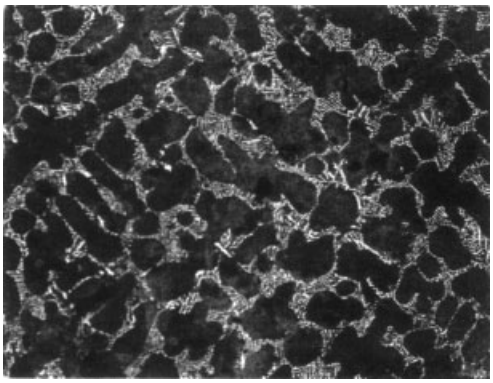

Figure 2. Grain structure of lead – antimony (5 wt %) alloy (magnification 160×)
White: antimony particles; Dark: lead matrix

The mechanical properties of lead – antimony alloys containing arsenic, tin, and copper are listed in Table 1. High-antimony alloys (> 3.5 wt % antimony) are strengthened primarily by the eutectic phase, whereas low-antimony alloys are strengthened by precipitation.

Lead – Acid Batteries (see also → Batteries, Chap. 4.). The major use of lead – antimony alloys is as grids, posts, and connectors for lead – acid batteries. The antimony eutectic prevents grid growth, retains the active material, and aids recovery from deep discharge. Deep-discharge industrial, load-leveling, and motive-power batteries utilize lead alloys containing 5 – 11 wt % antimony. Large tubular grids use 11 wt % antimony.

Antimony has one disadvantage for batteries. During use, antimony from the positive grid is oxidized, leaches into the electrolyte, and plates on the negative electrode. Antimony causes breakdown of water into hydrogen and oxygen during charging, leading to water loss. The amount of water lost is proportional to the amount of eutectic in the battery grid [5].

Automobile battery grids employ about 1 – 3 wt % antimony – lead alloys. Hybrid batteries use low antimony (1.6 – 2.5 wt %) alloys for the positive grids and nonantimony alloys for the negative grids to give reduced or no water loss. The posts and straps of virtually all lead – acid batteries are made of alloys containing about 3 wt % antimony.

Although low antimony content decreases water loss, it presents problems in casting thin grids. Alloys having less than 3.5 wt % antimony have reduced fluidity and increased volume shrinkage. The grain size increases with decreasing antimony content, as seen in Figure 3. The large columnar grains of low-antimony alloys are shown in Figure 4. To prevent formation of large grains, nucleants are added to lead – antimony alloys. Small amounts of sulfur, copper, or selenium serve as sites for growth of lead crystals during solidification and produce fine, rounded grain structures which are resistant to cracking [4–7]. By using nucleants, fine-grained structures, such as that shown in Figure 5, can be produced in cast alloys independent of the antimony content. The molten metal must be kept at a temperature high enough to assure complete solubility of the nucleants prior to casting the alloy. The United States uses primarily copper and sulfur as nucleants, while Europe and Asia use selenium. At very low antimony contents (1.0 – 1.6 wt %) selenium is used.

Ammunition. Lead shot is produced by dropping molten lead – antimony alloys containing 0.5 – 8 wt % antimony through holes in pans into water. The shot alloys contain arsenic in an amount equal to about 20 – 30 wt % of the

Table 1. Mechanical properties of lead – antimony alloys

Antimony content, wt %	Yield strength, MPa, after aging		Tensile strength, MPa	Elongation, %
	1 day	30 days	30 days	30 days
11	68.9	74.4	75.9	5
6	55.2	71.0	73.8	8
2.75	34.5	55.2	65.5	10
2	24.1	37.9	46.9	15
1	13.8	19.3	37.9	20
0	3.5	3.5	11.7	55

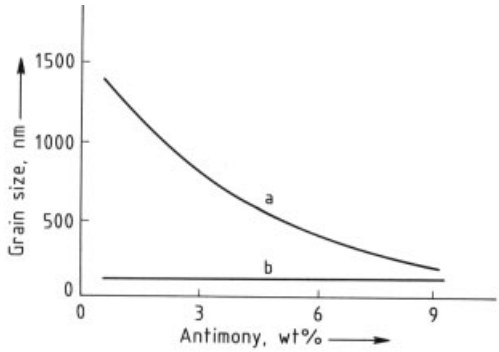

Figure 3. Grain size as a function of antimony content
a) Without grain refiners; b) With addition of selenium

Figure 4. The large columnar grain structure of lead – antimony (2.5 wt %) alloy without nucleants is prone to cracking (magnification 160×)
White: antimony; Dark: lead matrix

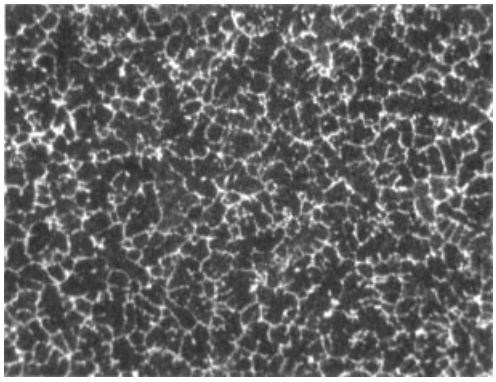

Figure 5. Fine, uniform, rounded grain structure of lead – antimony (2.5 wt %) with 0.015 wt % selenium as nucleant (magnification 160×)
White: antimony particles; Dark: lead matrix

antimony content. Arsenic permits the molten drop of lead to become round during freefall. Tin and cadmium produce elongated shot if present; therefore, they are restricted to less than 0.0005 wt %. Shot up to 5.8 mm in diameter is dropped from towers; larger shot must be cast in molds. Lead – antimony (0.5 – 3.0 wt %) alloys are used for cast or swagged bullets.

Cable Sheathing. Lead – antimony alloys containing 0.5 – 1.0 wt % antimony are used for cable sheathing as well as lead – copper and lead – tin – arsenic alloys. Antimony-containing alloys are used when higher strength and resistance to vibration is required. British standard BS-801 alloy B contains 0.85 wt % antimony. Lead – antimony alloys having 0.6 wt % or less antimony can be continuously extruded, are impervious to oil and moisture, and remain pliable indefinitely. Lead – antimony (1.0 wt %) is generally used as sleeves for cable splices.

Anodes. Lead – antimony (6 – 10 wt %) alloys containing 0.5 – 1.0 wt % arsenic have been used widely as anodes in copper, nickel, and chromium electrowinning and metal-plating processes. Lead – antimony anodes have high strength and develop a corrosion-resistant protective layer of lead dioxide during use. Lead – antimony anodes are resistant to passivation when the current is frequently interrupted.

Wrought Lead – Antimony. Wrought lead – antimony alloys are weaker and have higher creep rates than cast lead – antimony alloys, but have greater strength than lead – copper

alloys; they can be easily fabricated into most lead products. Because of their excellent resistance to corrosion by a variety of chemicals, lead – antimony (4 – 8 wt %) alloys are used for tank linings, pumps, valves, and heating and cooling coils, particularly where sulfuric acid or sulfate solutions are handled at elevated temperature. Arsenic is generally added to increase resistance to creep at elevated temperature.

3.2. Lead – Antimony – Tin Alloys

Lead – antimony – tin alloys are used for printing, bearings, solders, slush castings, and specialty castings. These alloys have low melting points, high hardness, and excellent high-temperature strength and fluidity.

Printing alloys generally contain more than 11 wt % antimony and 3 – 14 wt % tin. Typical alloys (see Table 2) have low melting points and high hardness for wear resistance in printing. Replication of mold detail is particularly important.

Table 2. Printing alloys

Process	Composition, wt %			Liquidus, °C	Solidus, °C	Brinell Hardness
	Sn	Sb	Pb			
Linotype	4	11.5	84.5	243	239	22
Electrotype	3	3	94	299	245	14
Stereotype	5	14	81	256	240	23
Monotype	7	16.5	76.5	275	240	26
Foundry type	14	24	62	318	240	32

Excellent antifriction properties and good hardness makes lead – antimony – tin alloys suitable for journal bearings, bushings, and sleeve bearings. The alloys contain 9 – 15 wt % antimony and 1 – 20 wt % tin. They also contain copper and arsenic, which improve compression, fatigue, and creep strength, important in bearings. Lead – antimony – tin bearing alloys are listed in ASTM B23-83 [8].

Specialty castings include belt buckles, trophies, casket trim, miniature figures, and holloware. Slush castings are produced by pouring an alloy into a mold, permitting a given thickness to solidify, and pouring out the remaining liquid to produce a hollow, light, high-detail casting. The near-eutectic lead alloy, containing 11 wt % antimony, 1 wt % tin, and 0 – 0.5 wt % arsenic, has a low melting point which permits silicon rubber molds to be used in spin-casting processes.

Low-antimony (2 – 5 wt %), low-tin (2 – 5 wt %) lead alloys are used for automobile body solder. Special lead – antimony alloys containing 1 – 4 wt % antimony are used for wheel-balancing weights, battery cable clamps, and collapsible tubes.

4. Lead – Calcium Alloys

Lead – calcium alloys are replacing lead – antimony alloys for many applications. Most U.S. original-equipment automotive batteries are constructed of lead – calcium, whereas most U.S. replacement batteries utilize lead – calcium alloys for the negative grids and lead – antimony alloys for the positive grids. Lead – calcium is used worldwide for standby power, submarine, and specialty sealed batteries. Lead – calcium alloy batteries do not require addition of water and can therefore be sealed. Lead – calcium alloy is used for electrowinning anodes, cable sheathing, sleeving, specialty boat keels, and lead alloy tapes.

4.1. Binary Lead – Calcium Alloys

The lead – calcium phase diagram is shown in Figure 6 [3]. The phase diagram is a peritectic with a peritectic temperature of 328.3 °C. The maximum solubility of calcium is 0.10 wt % at 328.3 °C and rapidly decreases to 0.02 wt % at 200 °C and 0.01 wt % at 25 °C. The large decrease in solubility with decreasing temperature permits precipitation strengthening of binary lead – calcium alloys by the compound Pb_3Ca [*12049-58-0*]. This compound is also formed in molten alloys containing more than 0.07 % calcium. Figure 7 shows the Pb_3Ca primary particles formed by solidification of an alloy containing 0.14 wt % calcium. The Pb_3Ca particles can act as nucleants and produce fine-grained microstructures.

Below 0.10 wt % calcium the alloys solidify into a cellular, dendritic grain structure (Fig. 8) [9]. Impurities segregate to the intercellular boundaries, causing them to be accented in

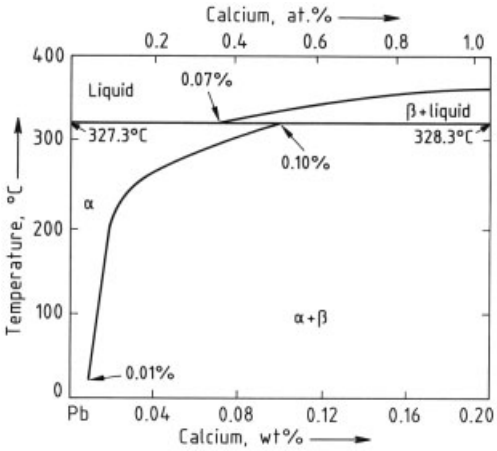

Figure 6. Phase diagram of the lead – calcium system [3] Reproduced from ref. [3] by permission of the publisher.

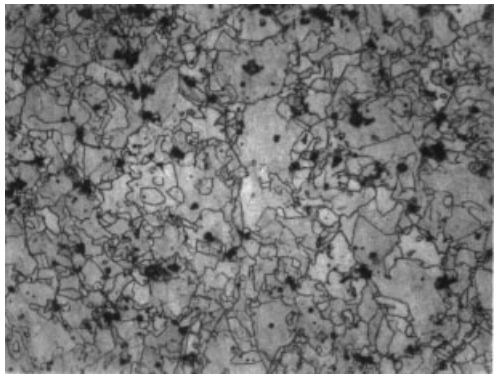

Figure 7. Grain structure of lead – calcium (0.14 wt %) alloy (magnification 160×); the dark spots are Pb_3Ca particles

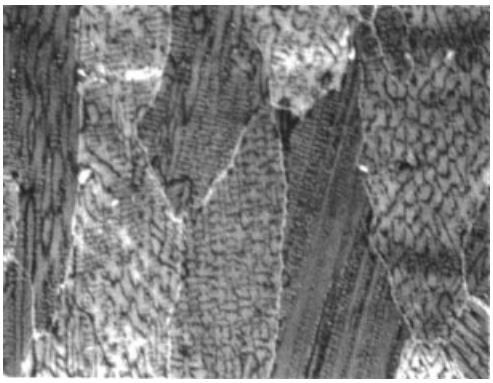

Figure 8. Grain structure of freshly cast lead – calcium (0.07 wt %) alloy (magnification 160×)

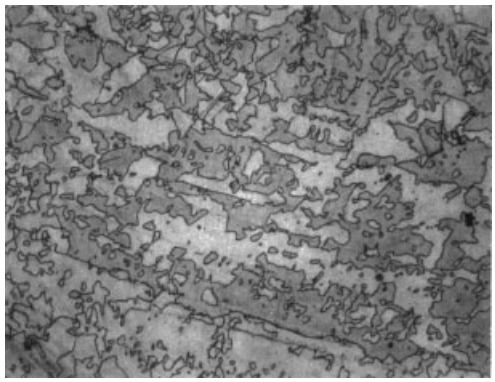

Figure 9. Grain structure of lead – calcium (0.07 wt %) alloy aged for seven days, showing the serrated grain boundaries (magnification 160×)

the microstructure. Reverse segregation causes some areas of the casting to contain 0.10 wt % calcium regardless of the original calcium content of the melt. Precipitation of calcium occurs by a discontinuous precipitation reaction [10]. In this reaction, grain boundaries move through the matrix, the original cast grain is destroyed, and the characteristic serrated grain boundaries of lead – calcium alloys seen in Figure 9 are produced [11]. The calcium content and the rate of grain boundary movement determine the final grain size and shape.

Lead – calcium (0.04 %) alloy, used as cable sheaths and sleeves for cable splices, is significantly stronger and has greater creep resistance than lead – copper (0.06 wt %) or lead – antimony (1 wt %). Table 3 compares the mechanical properties of lead – calcium to these alloys and to pure lead. The lead – calcium alloys have outstanding creep and fatigue resistance and relatively good ductility. Aging after production may cause a permanent set and difficulty in uncoiling; hence, lead – calcium is used primarily in straight lengths or as sleeves.

The main use of binary lead – calcium alloys is for the grids in large, stationary stand-by power batteries. These batteries use alloys containing 0.03 – 0.07 wt % calcium. The alloy has

sufficient strength for the application, but it is used principally because of its resistance to self-discharge and because it reduces water loss. The grids are large grained and resistant to growth.

Table 3. Mechanical properties of pure lead and lead alloys

Property	Composition, wt %			
	Pb, 100	Cu, 0.06	Sb, 1	Ca, 0.04
Tensile strength, MPa	12.2	17.4	20.9	27.9
Elongation, %	55	55	35	35
Fatigue strength, MPa	2.75	4.90	6.28	9.02

Reactivity of Lead-Calcium Alloys. Precise control of the calcium content is required to control the grain structure, corrosion resistance, and mechanical properties of lead – calcium alloys [12], [13]. Calcium reacts readily with air and other elements such as antimony, arsenic, and sulfur to produce oxides or intermetallic compounds. In these reactions, calcium is lost and suspended solids reduce fluidity and castibility. The very thin grids that are required for automotive batteries are difficult to cast from lead – calcium alloys.

A rapid method to determine the calcium content of lead alloys is a liquid-metal titration with lead – antimony (1 %) [14]. The end point is indicated by a gray oxide film pattern on the surface of a solidified sample of the metal when observed at a 45° angle to a light source. The basis for the titration is the reaction between calcium and antimony. The percentage of calcium in the sample can be calculated from the amount of antimony used. If additional calcium is needed in the alloy, the melt is sweetened with a lead – calcium (1 wt %) master alloy.

4.2. Lead – Calcium – Aluminum Alloys

Lead – calcium alloys can be protected against loss of calcium by addition of aluminum [15]. Aluminum provides a protective oxide skin on molten lead – calcium alloys. Even when scrap is remelted, calcium content is maintained by the presence of 0.02 wt % aluminum. Alloys without aluminum rapidly lose calcium, whereas those that contain 0.03 wt % aluminum exhibit negligible calcium losses, as shown in Figure 10 [16]. Even with less than optimum aluminum levels, the rate of oxidation is lower than that of aluminum-free alloys.

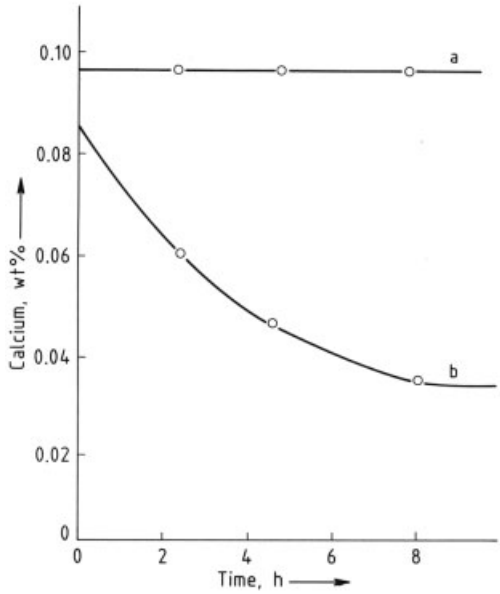

Figure 10. Effect of aluminum on loss of calcium
a) With 0.03 wt % aluminum; b) Without aluminum

Producing lead – calcium – aluminum alloys is difficult. Calcium and aluminum can be added simultaneously to lead using a calcium (73 wt %) – aluminum (27 wt %) master alloy [17]. With this method, the calcium and aluminum contents can be precisely controlled. Pressed pellets of metallic aluminum and metallic calcium are also used.

Lead alloys containing 0.09 – 0.15 wt % calcium and 0.015 – 0.03 wt % aluminum are used for the negative battery grids of virtually all lead – acid batteries in the United States and are also used in Japan, Canada, and Europe. If the molten alloy is held at too low a temperature, the aluminum precipitates from solution, rises to the surface of the molten alloy as finely divided aluminum particles, and enters the dross layer atop the melt.

4.3. Lead – Calcium – Tin Alloys

The addition of tin to lead – calcium alloys enhances the mechanical and electrochemical properties but reduces the rate of hardening

compared to binary alloys. The positive-grid alloys for maintenance-free lead – calcium batteries contain 0.3 wt % or more tin and also aluminum.

Cast lead – calcium – tin alloys usually contain 0.06 – 0.11 wt % calcium and 0.3 wt % tin. They have excellent fluidity, harden rapidly, have a fine grain structure, and are resistant to corrosion. Table 4 lists the mechanical properties of cast lead – calcium – tin alloys and other alloys.

Wrought lead – calcium – tin alloys contain more tin, have higher mechanical strength, exhibit greater stability, and are more creep resistant than the cast alloys. Rolled lead – calcium – tin alloy strip is used to produce automotive battery grids in a continuous process [18]. Table 5 lists the mechanical properties of rolled lead – calcium – tin alloys, compared with copper – lead and lead – antimony (6 wt %) alloys.

Lead Anodes. A major use for lead – calcium – tin alloys is lead anodes for electrowinning. The lead – calcium anodes form a hard, adherent lead dioxide layer during use, resist corrosion, and greatly reduce lead contamination of the cathode. Anodes produced from cast lead – calcium (0.03 – 0.09 wt %) alloys have a tendency to warp due to low mechanical strength and casting defects.

Wrought lead – calcium – tin anodes have replaced many cast lead – calcium anodes [19]. Superior mechanical properties, uniform grain structure, low corrosion rates, and lack of casting defects result in increased life for wrought lead – calcium – tin anodes compared to other lead alloy anodes.

5. Lead – Tin Alloys

The lead – tin phase diagram is shown in Figure 11 [3]. It exhibits a eutectic point at 61.9 wt % tin and 183 °C. The solid solubility of tin at 183 °C is 19 wt %, decreasing to 1.9 wt % at 25 °C.

Solders. The major use of lead – tin alloys is as solders for sealing and joining metals. Solders range in composition from 30 to 98 wt % lead, the remainder being tin. Lead – tin (2 wt %) solder is used to seal the side seams of steel (tin) cans. Addition of 0.5 wt % silver to the alloy significantly improves the creep strength which is essential in pressurized cans. Pre-tinned cans allow the use of low-tin solder. High-speed manufacturing of cans requires an alloy with a very narrow freezing range.

Lead – tin alloys containing 40 – 50 wt % tin are used for general purpose soldering. Low-melting lead – tin solder with 63 wt % tin is used for electronic soldering, particularly for printed circuit boards. Lead alloys with 15 – 30 wt % tin are used for soldering automobile radiators and other types of heat exchanger. A lead alloy containing 2.5 wt % tin and 0.5 wt % silver is used when high-temperature fatigue and creep strengths are required, for example, in soldered connections in heat exchangers. Solders with 35 – 40 wt % tin have a wide plastic range and are used for wiping lead joints. Sometimes, 0.5 – 2.0 wt % antimony is added to improve mechanical properties. Antimony contents of 0.2 – 0.5 wt % prevent brittle phase transformations at low temperatures.

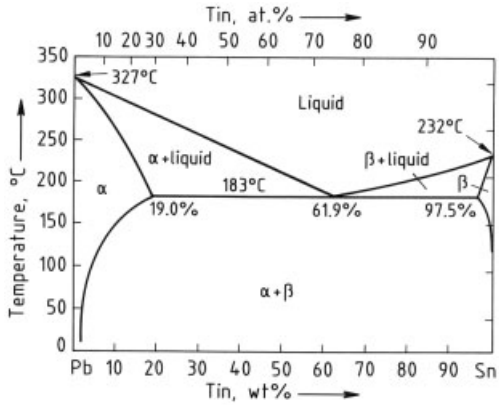

Figure 11. Phase diagram of the lead – tin system [3]
Reproduced from ref. [3] by permission of the publisher.

Corrosion Protection. Lead – tin alloys are used for corrosion resistant coatings on steel and copper. Terne steel is steel sheet coated with a lead alloy containing 15 – 20 wt % tin. This alloy contains sufficient tin to alloy with the surface of the steel. Terne sheets are produced flat or coiled, carrying different coating weights. These coated sheets are used for radio and television chassis, roofs, fuel tanks, air filters, oil filters, gaskets, metal furniture, gutters, and downspouts. Terne

Table 4. Mechanical properties of cast lead alloys

Property	Content of alloying components, wt %						
	Cu, 0.06	Ca, 0.06	Sn, 0.5	Ca, 0.04; Sn, 0.3	Ca, 0.08; Sn, 0.5	Ca, 0.08; Sn, 1.0	Ca, 0.08; Sb, 6
Tensile strength, MPa	17.4	34.8	41.8	46.0	48.8	59.7	59.2
Yield strength, MPa	9.0	24.3	27.9	34.8	38.2	46.0	55.8
Elongation, %	55	30	15	20	15	15	15
Time to failure at 20.9 MPa, h		10	30	50	100	850	>10 000

Table 5. Mechanical properties of rolled lead alloys

Property	Content of alloying components, wt %					
	Cu, 0.06	Ca, 0.06	Ca, 0.04; Sn, 0.5	Ca, 0.065; Sn, 0.7	Ca, 0.06; Sn, 1.3	Sb, 6
Tensile strength, MPa	17.4	32.8	48.8	62.8	69.6	30.6
Yield strength, MPa	9.0	25.1	46.0	59.2	66.2	19.5
Elongation, %	55	35	15	10	10	35
Time to failure at 20.9 MPa, h		7	50	500	1000	1.5

steel has good corrosion resistance, draws well, is relatively cheap, and offers a good base for painting.

A similar coating, containing 4 wt % tin, is applied to copper sheet and is used primarily for building flashings (ASTM B101-83) [8]. Other lead – tin alloys, usually with 50 wt % tin, are applied as coatings to steel and copper electronic components for corrosion protection, appearance, and ease of soldering. Lead – tin alloys can be built up easily to any desired thickness by electroplating from a fluoroborate solution. Electroplated coatings are not recommended for corrosion protection because they tend to be porous.

Tin is also used as an alloying element in lead – antimony alloys to improve fluidity and to prevent drossing; in lead – calcium alloys to improve mechanical properties and enhance electrochemical performance; in lead – arsenic alloys to maintain a stable composition; and as an additive to low-melting alloys.

Lead – tin (1.8 – 2.5 wt %) is used as a cable sheathing alloy (BS 801 alloy A and DIN 17 640). Tin is generally added to lead – arsenic cable alloys in small amounts. The arsenic alloys have excellent creep resistance and mechanical properties, but are unstable and lose arsenic readily by oxidation. The addition of small amounts of tin (0.10 – 0.20 wt %) eliminates arsenic loss. Lead alloys having 0.4 wt % tin and 0.15 % cadmium, which are used for cable sheathing, do not age harden, show excellent corrosion and creep resistance, and are very ductile.

6. Lead – Copper Alloys

Copper is an alloying element as well as an impurity in lead. The lead – copper system has a eutectic point at 0.06 % copper and 326 °C. In lead refining, the copper content can thus be reduced to about 0.08 % merely by cooling. Further refining requires chemical treatment. The solubility of copper in lead decreases to about 0.005 % at 0 °C.

High-copper lead alloys generally contain 60 – 70 % copper. The mechanical properties of sand-cast copper – lead alloys are shown in Table 6. The high-copper alloys are difficult to cast and are susceptible to extensive segregation. Cast lead – copper (60 – 70 wt %) alloys are used as bearings and bushings for high-temperature service. The cast alloys have been recently replaced by sintered copper powder products infiltrated with lead to produce more uniform distribution of the lead.

Only lead alloys containing copper below 0.08 % have practical applications. Lead sheet, pipe, cable sheathing, wire, and fabricated products are produced from lead – copper alloys with copper contents near the eutectic composition. Lead – copper alloys in the range 0.03 – 0.08 wt % copper are covered by many speci-

fications: ASTM B29-79 (- 84) [8], QQL 171 (United States), BS 334, HP2 Type 11 (Canada), DIN 1719, and AS 1812 (Australia). Lead–copper alloys are specified because of superior mechanical properties, creep resistance, corrosion resistance, and high-temperature stability.

Table 6. Mechanical properties of sand-cast copper–lead alloys

Property	Lead Content	
	30 wt %	40 wt %
Tensile strength, MPa	55.8–62.8	52.3–59.2
Elongation, % in 50.8 mm	6–8	6–8
Compressive strength, MPa	20.9	17.5
Brinell hardness	23–33	20–30

Lead–copper alloys are the primary material used in the continuous extrusion of cable coverings for the electrical power cable industry in the United States. Other alloys, containing tin and arsenic as well as copper, have also been developed for cable sheathing in the United States to provide higher fatigue strength.

Extruded or rolled lead–copper alloys contain a uniform dispersion of copper particles in a lead matrix. Because the solid solubility of copper in lead is very low, copper particles in the matrix remain stable up to near the melting point of lead, maintaining uniform grain size even at elevated temperature.

Copper-containing lead alloys undergo less corrosion in sulfuric acid or sulfate solutions than pure lead or other lead alloys. The uniformly dispersed copper particles give rise to local cells in which lead forms the anode and copper forms the cathode. Through this anodic corrosion of the lead, an insoluble film of lead sulfate forms on the surface of the lead, passivating it and preventing further corrosion. The film, if damaged, rapidly reforms.

Lead–copper alloys are also used as tank linings, tubes for acid-mist precipitators, steam heating pipes for sulfuric acid or chromate plating baths, and lead sheathing for roofs.

7. Lead–Silver Alloys

Silver readily forms alloys with lead. Lead is often used as a base metal solvent for silver-recovery processes. The lead–silver system is a simple eutectic with the eutectic point at 2.5 wt % silver and 304 °C. The solid solubility of silver in lead is 0.10 wt % at 304 °C, dropping to less than 0.02 wt % at 20 °C.

Lead–silver alloys show significant age hardening when quenched from elevated temperature. Because of the pronounced hardening which occurs with small amounts of silver, the content of silver as an impurity in pure lead is restricted to less than 0.0025 wt % in most specifications. Small additions of silver to lead produces high resistance to recrystallization and grain growth.

The major uses for lead–silver alloys are as anodes and high-temperature solders. Only lead–silver alloys containing less than 6 wt % silver are used commercially. Lead alloys containing 0.75–1.25 wt % silver are used as insoluble anodes in the electrowinning of zinc and manganese. Some zinc refineries have reduced the silver content of the anodes to as low as 0.25 wt %. Lead–calcium–silver anodes are also used for zinc electrowinning. Silver promotes the formation of very hard, dense, electrically conducting layers of lead dioxide on the surface of the anode.

Silver reduces the oxygen evolution potential at the anode, which reduces the rate of corrosion and decreases lead contamination of the cathode. Lead–antimony–silver alloy anodes are used for the production of thin copper foil for use in electronics. Lead–silver (2 wt %), lead–silver (1 wt %)–tin (1 wt %), and lead antimony (6 wt %)–silver (1–2 wt %) alloys are used as anodes in cathodic protection of steel pipes and structures in fresh, brackish, or sea water. The lead dioxide layer is not only conductive, but also resists decomposition in chloride environments. Silver-free alloys rapidly become passivated and scale badly in sea water.

Silver is also added to the positive grids of lead–acid batteries in small amounts (0.005–0.05 wt %) to reduce the rate of corrosion.

Lead–silver alloys are used extensively as soft solders; these contain 1–6 wt % silver. Lead–silver solders have a narrower freezing range and higher melting point (304 °C) than conventional solders. Solders containing 2.5 wt % silver or less are used either as binary alloys or combined with 0.5–2 wt % tin. Lead–silver solders have excellent corrosion resistance. The composition of lead–silver sol-

ders is listed in ASTM B-32-87 (solder alloys) [9].

8. Lead – Tellurium Alloys

Tellurium is often used in lead alloys when high mechanical strength at minimal alloy content is required. It is used for pipes and sheets, shielding for nuclear reactors, and cable sheathing. Lead alloys containing 0.035 – 0.10 wt % tellurium generally also contain copper in amounts of 0.03 – 0.08 wt %. The U.S. Federal Specifications QQL-201F and ASTM B-749-85 [8] cover the use of wrought lead – tellurium alloys. Lead – tellurium alloys for cable sheathing are specified by DIN 17 640.

Cold-rolled alloys of lead with 0.06 wt % tellurium often attain ultimate tensile strengths of 25 – 30 MPa. High mechanical strength, excellent creep resistance, and low levels of alloying elements have made lead – tellurium alloys the primary material for nuclear shielding for small reactors such as those aboard submarines. The alloy is self supporting and does not generate secondary radiation.

Wrought or extruded lead – tellurium (0.035 – 0.10 wt %) alloys produce extremely fine grains. The binary alloy is, however, susceptible to recrystallization. The addition of copper or silver reduces grain growth and retains the fine grain size. Since tellurium is a poison for sealed lead – acid batteries, the tellurium content of lead and lead alloys used for such purposes is usually restricted to less than 2 ppm.

9. Low-Melting Alloys

Lead alloys with large amounts of bismuth, tin, cadmium, and indium that melt at relatively low temperatures (10 – 183 °C) are known as fusible or low-melting alloys. The specifications of many of these alloys are listed in ASTM B 774-88 [8].

These alloys are used as fuses, sprinkler system alloys, foundry pattern alloys, molds, dies, punches, cores, and mandrels where the low-melting alloy is often melted out of a mold. The alloys are also used as solders, for the replication of human body parts, and as fillers for tube bending. Lead – indium alloys are often used to join metals to glass.

10. Reactive Lead Alloys

Strontium – lead alloys behave similarly to lead – calcium alloys in terms of mechanical properties and performance. Lead – strontium (0.08 – 0.20 wt %) alloys, also containing aluminum, have been used as battery grid alloys. They have excellent fluidity, harden rapidly, and have excellent resistance to corrosion. However, these alloys over-age rapidly, resulting in significant loss of mechanical strength within several days. The addition of 1 wt % or more of tin is required to resist over-aging. The high cost of lead – strontium alloys compared to lead – calcium alloys has restricted their use. Lead – strontium – tin alloys are used as anodes for copper electrowinning.

Lead – lithium and lead – lithium – tin alloys have been proposed as alloys for lead – acid battery grids because of very rapid aging and very high mechanical properties. These alloys are, however, susceptible to grain boundary corrosion. Lead – lithium alloys containing strontium, barium, and calcium have been used for bearings. The high melting point and the retention of mechanical properties and stability at high temperature make these alloys particularly attractive for bearings. The addition of aluminum to prevent oxidation permits casting at high temperatures without oxidation.

11. Occupational Health and Safety

Because of the toxicity of lead, special care must be taken when working with lead alloys. Lead and its inorganic compounds are neurotoxins which may produce peripheral neuropathy. For an overview of the effects of lead exposure, see Occupational Exposure to Lead, Appendix A (29CFR 1910.1025) [20].

Most lead alloys are first melted during processing. The melting or fusing of lead alloys may generate lead oxides, drosses, or fumes which can present a health hazard. The major routes for absorption of lead metal are inhalation and ingestion. Inhalation can be avoided by use of

ventilators and respirators where the exposure is above the permissible exposure limit (PEL) or the threshold limit value (TLV) specified by OSHA (United States) or other local, municipal, or federal regulations. These limits vary considerably by industry and country but generally consist of a PEL of ca. 50 – 200 µg/m^3 of lead in air, and lead levels of 50 – 100 µg/dL in whole blood.

Coveralls or other full-body clothing should be worn when working with lead alloys and properly laundered after use. Hard hats, safety glasses, safety boots, and other safety equipment should be worn as appropriate for the industrial environment where the lead alloys are used. Hands, face, neck, and arms should be washed before eating or smoking.

Lead – antimony or lead – arsenic alloys must not be mixed with lead – calcium(aluminum) alloys in the molten state. Addition of lead – calcium – aluminum alloys to lead – antimony alloys results in reaction of calcium or aluminum with the antimony and arsenic to form arsenides and antimonides. The dross containing the arsenides and antimonides floats to the surface of the molten lead alloy and may generate poisonous arsine or stibine if it becomes wet. Care must be taken to prevent mixing of calcium and antimony alloys and to ensure proper handling of drosses.

If the temperature of a molten lead – calcium(tin) – aluminum alloy is not kept sufficiently high, finely divided aluminum particles may precipitate and float to the top of the melt. They may become mixed with oxides of lead in the dross. The finely divided aluminum particles can react violently with the oxides in the dross if ignited. Ignition can occur if attempts are made to melt or "burn" the dross away from areas of buildup with a torch. The oxides in the dross can supply oxygen for the combustion of aluminum once ignited.

Despite the benefits of lead and lead alloys, the use of these materials is declining rapidly, due primarily to environmental health and safety factors. For many years, lead alloys have been the material of choice for many corrosive environments, but they are now being replaced by stainless steel, plastics, and exotic metals. The toxic nature of lead requires special precautions and handling not necessary with other materials. These requirements have reduced the usage of lead alloys.

12. References

General References

1. W. Hofmann: *Lead and Lead Alloys,* Springer Verlag, New York 1970.
2. J. F. Smith, R. R. Kubalak in *Metals Handbook,* 9th ed., **vol. 2,** ASM International, Metals Park, Ohio 1979, p. 493.

Specific References

3. M. Hansen: *Constitution of Binary Alloys,* McGraw Hill, New York 1958.
4. V. Heubner, A. Ueberschaer in *Proc. 6th Int. Conf. Lead, Lead Development Ass.,* London 1977.
5. R. D. Prengaman in *Proc. 8th Int. Conf. Lead, Lead Development Ass.,* London 1983, p. 69.
6. B. E. Kallup, D. Berndt in K. R. Bullock, D. Pavlov (eds.): *Advances in Lead Acid Batteries,* The Electrochemical Society, Pennington NJ 1984, p. 214.
7. R. D. Prengaman, *Bat. Man.,* Oct. (1983) 29.
8. *Annual Book of ASTM Standards,* Sec. 2, Vol. 02.04 ASTM, Philadelphia, PA 1988.
9. T. W. Caldwell, U.S. Sokolov, *J. Electrochem. Soc.* **123** (1976) 972.
10. R. D. Prengaman in *Proc. 7th Int. Conf. Lead, Lead Development Ass.,* London 1980, p. 34.
11. A. M. Howard, E. Willinganz, *Electrochem. Technol.* **6** (1968) 370.
12. G. W. Mao, J. G. Larson, P. Rao, *J. Electrochem. Soc.* **120** (1973) 11.
13. M. V. Rose, J. A. Young in *Proc. 5th Int. Lead Conf., Lead Development Ass.,* London 1974, p. 37.
14. G. M. Bouton, G. S. Phipps, *Trans Electrochem. Soc.* Preprint 92-13, 1974.
15. Lucas, US 3 920 473, 1974 (R. I. Sims).
16. R. D. Prengaman in K. Bullock, D. Pavlov (eds.): *Advances in Lead Acid Batteries,* The Electrochemical Soc., Pennington NJ 1984, p. 201.
17. RSR Corporation, US 4 439 398, 1984 (R. D. Prengaman).
18. General Motors, US 3 891 459, 1975 (C. P. McCartney, N. L. Williams).
19. R. D. Prengaman in D. L. Robinson, S. E. James (eds.): *Anodes for Electrowinning,* TMS-AIME, Warrendale PA 1984, p. 49.
20. OSHA, *US Federal Register,* November 14, 1978.

Lead Compounds

DODD S. CARR, International Lead Zinc Research Organization, Inc., Research Triangle Park, NC 27709-2036, United States

1. **Introduction** 1
2. **Lead Salts and Oxides** 1
2.1. **Lead Acetates** 1
2.2. **Lead Carbonates** 2
2.3. **Lead Halides** 3
2.4. **Lead Oxides** 3
2.5. **Lead Silicates** 5
2.6. **Lead Sulfates** 5
2.7. **Other Lead Salts** 5
3. **Organolead Compounds** 6
4. **Occupational Health** 7
5. **References** 8

1. Introduction

Lead is one of the oldest known metals, dating from about 3000 B.C. (→ Lead). The Romans used lead extensively for water pipes, while in ancient Egypt, lead compounds were employed for glazing pottery and ornamental objects, a purpose for which they are still used today. Now, the largest single use of lead compounds, mainly lead oxide, is for the manufacture of lead – acid storage batteries. Other industrially important uses of lead compounds range from antiknock additives in gasoline to lead crystal glasses and stabilizers for plastics during thermal processing.

2. Lead Salts and Oxides

Lead is an amphoteric element, which forms lead salts with acids as well as metal salts of plumbic acid with alkalies. Lead forms many salts, oxides, and organometallic compounds. Although lead has four electrons in its outer (or valence) shell, the usual valence of lead is + 2, rather than + 4, because of the reluctance of its two 6 *s* electrons to ionize. Also, the divalent plumbous ion differs from the other group 14 divalent ions, such as the stannous ion (Sn^{2+}), in that it does not have reducing properties.

Many inorganic lead compounds have two or more crystalline forms with different properties. As a result, the oxides and sulfide of divalent lead often appear colored because of their state of crystallization. Thus, pure, tetragonal α-PbO is red; orthorhombic β-PbO is yellow; and crystals of lead sulfide, PbS, have a black, metallic luster. On the other hand, the carbonates, sulfates, nitrates, and halides (except for the yellow iodide) are colorless.

Highly crystalline basic lead salts, both anhydrous and hydrated, are readily formed. For example, tetrabasic lead sulfate [*52732-72-6*], $4\,PbO \cdot PbSO_4$, and the tribasic lead sulfate hydrate [*12202-17-4*], $3\,PbO \cdot PbSO_4 \cdot H_2O$, are prepared by boiling suspensions of lead oxide and lead sulfate in water. Moreover, complex mixed salts, such as white lead, $2\,PbCO_3 \cdot Pb(OH)_2$, are readily formed.

Most lead compounds are derived from pig lead (refined metal),usually via conversion to lead monoxide (PbO), commonly known as litharge. In general, lead compounds may be formed by the reaction of a slurry of litharge and the appropriate acid, by the reaction of a solution of a lead salt with an acid, or by fusion of litharge with the appropriate metal oxide. In the latter case, for example, the ease with which lead monoxide combines with silicon dioxide to form a low-melting silicate has been utilized in the ceramics industry for producing glasses and glazes.

2.1. Lead Acetates

Lead reacts readily with acetic acid to form four different lead acetate compounds: anhydrous, basic, trihydrate, and tetraacetate. Some physical properties of the lead acetates are summarized in Table 1.

Anhydrous lead acetate, plumbous acetate, $Pb(C_2H_3O_2)_2$, is a white, crystalline solid that

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_249

Table 1. Physical properties of lead acetates [8]

Compound	CAS registry no.	M_r	*mp*, °C	ϱ, g/cm^3	Solubility, g/100 mL H_2O (t, °C)
$Pb(C_2H_3O_2)_2$	[*301-04-2*]	325.29	280	3.25	44.3 (20), 221 (50)
$2\,Pb(OH)_2 \cdot Pb(C_2H_3O_2)_2$	[*1335-32-6*]	807.69	75 (dec. 200)		6.25 (15), 25 (100)
$Pb(C_2H_3O_2)_2 \cdot 3\,H_2O$	[*6080-56-4*]	379.33	75 (dec. 200)	2.55	45.6 (15), 200 (100)
$Pb(C_2H_3O_2)_4$	[*546-67-8*]	443.38	175	2.228	dec.

is very soluble in water. It is made by dissolving lead monoxide or lead carbonate in concentrated acetic acid. Solutions of lead acetate are often used to prepare other lead salts.

Basic lead acetate, lead subacetate, $2\,Pb(OH)_2 \cdot Pb(C_2H_3O_2)_2$, is a heavy white powder that is very soluble in water and soluble in ethanol. It is prepared by dissolving lead monoxide in dilute acetic acid. Basic lead acetate is used for sugar analysis.

Lead acetate trihydrate, sugar of lead, $Pb(C_2H_3O_2)_2 \cdot 3\,H_2O$, is a white, monoclinic crystalline solid having a refractive index (along the β axis) of 1.567. It is very soluble in water, but insoluble in ethanol. Lead acetate trihydrate is prepared by dissolving lead monoxide in hot dilute acetic acid; large, long crystals separate upon cooling. Although the crystals are intensely sweet, they are poisonous. Lead acetate trihydrate is the usual commercial form of lead acetate and it is available in technical and reagent grades. It is used to make other lead compounds (for example, basic lead carbonate, lead chromate, and lead salts of higher fatty acids), as a mordant for cotton dyes, as a water repellant, and as a processing agent in the cosmetic, perfume, and toiletry industries.

Lead tetraacetate, plumbic acetate, $Pb(C_2H_3O_2)_4$, is a colorless, monoclinic crystalline solid that is soluble in chloroform and in hot acetic acid, but decomposes in cold water and in ethanol. Lead tetraacetate is prepared by adding warm, anhydrous, glacial acetic acid to red lead (Pb_3O_4) and subsequently cooling the solution. It is available in laboratory quantities as colorless to faintly pink crystals stored under glacial acetic acid. Lead tetraacetate is often used as an oxidizing agent in organic syntheses because it is highly selective in cleaving vicinal glycols. It also readily cleaves α-hydroxy acids, such as oxalic acid, at room temperature. Another use is for the introduction of acetoxy groups in organic molecules [9].

2.2. Lead Carbonates

Lead carbonate [*598-63-0*], $PbCO_3$, M_r 267.21, ϱ 6.6 g/cm^3, forms colorless orthorhombic crystals having refractive indices of 1.804, 2.076, and 2.078. It is slightly soluble in cold water (1.1 mg/L at 20 °C), is soluble in acids and alkalies, but is insoluble in aqueous ammonia and ethanol. Lead carbonate is produced by passing carbon dioxide into a cold dilute solution of lead acetate, or by shaking a suspension of a lead salt (less soluble than the carbonate) with ammonium carbonate at low temperature to avoid formation of basic lead carbonate. Lead carbonate is used in the catalytic polymerization of formaldehyde to give high molecular mass crystalline poly(oxymethylene) [10], to improve the bonding of polychloroprene to metals in wire-reinforced hoses [11], as a component of high-pressure lubricating greases [12], and as a lubricant and stabilizer for poly(vinyl chloride) [13].

Basic lead carbonate [*1319-46-6*], white lead, $2\,PbCO_3 \cdot Pb(OH)_2$, M_r 775.67, ϱ 6.14 g/cm^3, forms white hexagonal crystals that decompose at 400 °C. It is insoluble in water, slightly soluble in carbonated water, soluble in nitric acid, and insoluble in ethanol. Basic lead carbonate is made by reacting soluble lead acetate with carbon dioxide in the presence of air. It is no longer used as a white hiding pigment in paints because of its toxicity, but it has many other uses. For example, basic lead carbonate is used as a component of ceramic glazes, as a curing agent with peroxides to form improved polyethylene wire insulation [14], as a color-changing component of temperature-sensitive

inks [15], as a component of lubricating grease [16], and as a component of weighted nylon-reinforced fish nets made of poly(vinyl chloride) fibers [17].

2.3. Lead Halides

Lead combines with the halogens to form the corresponding fluoride, chloride, bromide, and iodide. Some physical properties of the lead halides are listed in Table 2.

Lead fluoride, lead difluoride, PbF_2, is a colorless, orthorhombic crystalline compound up to ca. 220 °C, above which it transforms into the cubic form. Lead fluoride dissolves in nitric acid, but is insoluble in acetone and aqueous ammonia. It is prepared by reacting lead hydroxide or lead carbonate with hydrofluoric acid, or by the reaction of lead nitrate with potassium fluoride. Lead fluoride is used in glass sealing disks for IR sensors [18]; in wear resistant automotive shock absorbers [19]; for the electroless deposition of lead [20]; as a flux for brazing of aluminum and its alloys [21]; in optical glass fibers for IR transmission [22]; and in thin-film batteries [23].

Lead chloride, lead dichloride, $PbCl_2$, forms white orthorhombic needles having refractive indices of 2.199, 2.217, and 2.260. It is slightly soluble in dilute hydrochloric acid and in aqueous ammonia, but is insoluble in ethanol. It is made by reacting lead monoxide or basic lead carbonate with hydrochloric acid, or by treating a solution of lead acetate with hydrochloric acid and allowing the precipitate to settle. Upon heating in air, lead chloride readily forms basic chlorides, for example, $PbCl_2 \cdot Pb(OH)_2$ [*15887-88-4*], which is known as Pattinson's lead white, an artist's pigment. Other uses of lead chloride include: a precursor of organolead compounds [24]; seawater-activated batteries [25]; expanding polymer-based mortar [26]; flux for soldering cast iron and cast brass [27]; sound-insulating rubber sealants [28]; corrosion inhibitor for galvanized steel [29]; and infrared-transmitting glasses for CO_2 lasers [30].

Lead bromide, lead dibromide, $PbBr_2$, forms white orthorhombic crystals that are slightly soluble in aqueous ammonia, soluble in acids and in aqueous potassium bromide (due to complex formation), but insoluble in ethanol. Lead bromide decomposes slowly when exposed to light and darkens due to the formation of lead. It is prepared by reaction of hydrobromic acid with lead monoxide or lead carbonate. Lead bromide is used industrially as a filler for flame-resistant polypropylene [31], in glass optical waveguides for infrared thermometers [32], and for catalysts for producing polyesters [33].

Lead iodide, lead diodide, PbI_2, forms a powder of yellow hexagonal crystals that dissolve in alkalies and in aqueous potassium iodide, but not in ethanol. Lead iodide is made by reacting a water-soluble lead compound with hydroiodic acid or with a soluble metal iodide. It is readily purified by recrystallization from water. Lead iodide is used in aerosols for cloud seeding to produce rain artificially [34]; for making high-contrast photographic images of laser radiation [35]; for high capacity cathodes in lithium batteries [36]; and in low-temperature thermographic copying materials [37].

2.4. Lead Oxides

There are two main oxides of lead, lead monoxide (PbO) and lead dioxide (PbO_2), in which it is divalent and tetravalent, respectively. In addition, lead forms a mixed oxide, Pb_3O_4, and a black oxide which normally contains 60–80 % lead monoxide, the remainder being finely divided metallic lead. The largest market for lead oxides is in the lead–acid battery industry, where lead monoxide is used to prepare the active paste for the electrode grids. The next largest market for lead oxides is the ceramics industry, which employs these compounds for the production of glasses, glazes, and vitreous enamels. Other major markets for lead oxides are the paint and rubber industries.

Another important outlet for lead oxide is for the production of lead salts, particularly those used as heat stabilizers for poly(vinyl chloride) resins, especially basic lead salts. In 1979, the U.S. plastics market consumed ca. 9500 t of lead salts [38]. Because of its electrical and electronic

Table 2. Physical properties of lead halides [8]

Compound	CAS registry no.	M_r	*mp*, °C	*bp*, °C	ϱ, g/cm^3	Solubility at 20 °C, g/100 mL H_2O
PbF_2	[*7783-46-2*]	245.20	855	1290	8.24	0.064
$PbCl_2$	[*7758-95-4*]	278.11	501	950	5.85	0.99
$PbBr_2$	[*10031-22-8*]	367.01	373	916	6.66	0.844
PbI_2	[*10101-63-0*]	461.01	402	954	6.16	0.063

properties, lead oxide is used in the production of capacitors, electrophotographic plates, transducers, and ferrimagnetic and ferromagnetic materials. Examples of the latter include mixed oxides, such as lead zirconate – lead titanate, and lead metaniobate.

Some physical properties of the lead oxides are given in Table 3.

Table 3. Physical properties of lead oxides [8]

Compound	CAS registry no.	M_r	Decomp., °C	ϱ, g/cm^3
PbO	[*1317-36-8*]	223.20	1472 (*bp*)	9.53
PbO_2	[*1309-60-0*]	239.20	290	9.375
Pb_2O_3	[*1314-27-8*]	462.40	370	
Pb_3O_4	[*1314-41-6*]	685.60	500	9.1

Lead monoxide, litharge, PbO, exists in a reddish alpha form up to 489 °C; it then transforms to a yellow beta form (massicot), which is stable at high temperatures. It has a water solubility of 17 mg/L at 20 °C, and is soluble in nitric acid, alkalies, lead acetate, ammonium chloride, and chlorides of calcium and strontium. In alkalies, it forms the plumbite ion, PbO_2^{2-}. Lead oxides are produced industrially by thermal processes in which lead is directly oxidized with air. In the ball mill process, metallic lead balls are tumbled in air to produce a "leady" oxide, which typically contains 20 – 35 % free lead. The Barton pot process oxidizes droplets of molten lead at ca. 430 °C to produce either litharge or leady litharge.

The principal use of lead monoxide is in the manufacture of pastes for the grids used in lead – acid batteries. It is also widely used in optical, electrical, and electronic glasses, as well as in glazes for fine tableware. To render the lead compounds insoluble in foods, the glazes and vitreous enamels are prepared from frits, in which the lead monoxide is converted to lead bisilicate, for example. Litharge is also used in rubber as a vulcanizing agent, in lead soaps employed as driers in varnishes, in high-temperature lubricants, as a neutralizing agent in organic syntheses, as a heat stabilizer in plastics, and as a starting material in the production of pigments [40]. The preparation, properties, and other uses of lead oxide are described in [41]. The consumption of litharge in the United States is summarized in Table 4 according to end-use.

Table 4. Litharge consumption in the United States [39]

	Litharge consumption, t			
Industry	1982	1983	1984	1985
Ceramics	30 980	36 782	37 960	65 413
Chrome pigments	6 591	5 973	4 367	3 794
Paints	3 052	3 256	3 635	3 397
Rubber	787	933	1 016	739
Other	10 267	9 596	9 920	10 590
Total	*51 677*	*56 540*	*56 898*	*83 933*

Lead dioxide, lead peroxide, plattnerite, PbO_2, is a brownish-black crystalline powder consisting of fine flakes in either the orthorhombic alpha or the tetragonal beta form. Lead dioxide decomposes to lead monoxide when heated above 290 °C. Lead dioxide is practically insoluble in water or alkaline solutions. It dissolves slowly in acetic acid or aqueous ammonium acetate, but more rapidly in hydrochloric acid and in a mixture of nitric acid and hydrogen peroxide. It is produced industrially by the oxidation of red lead (Pb_3O_4) with chlorine in an alkaline slurry. Lead dioxide is electrically conductive and is formed in situ as the active material of the positive plates in lead – acid batteries. Since it is a vigorous oxidizing agent when heated, lead dioxide is used in the manufacture of chemicals, dyes, matches, pyrotechnics, and liquid polysulfide polymers [42]. Other applications of lead

dioxide include antifriction compositions for plastic sliding bearings [43], ballistic modifers in high-energy propellants [44], electrodes for seawater electrolysis [45], filters for desulfurization of waste gases [46], and vulcanizing agents for butyl rubber puncture-sealing layers inside tires [47].

Lead sesquioxide, lead trioxide, Pb_2O_3, is an amorphous, orange-yellow powder. It is insoluble in cold water, it decomposes in hot water, and in acids it decomposes to the corresponding lead salts and lead dioxide. Lead sesquioxide can be prepared from lead dioxide by hydrothermal dissociation. It is used as a ballistic modifier for high-energy propellants [48], as a cathode material in lithium batteries [49], and as an additive to increase the shattering force of explosives [50].

Red lead, lead tetroxide, minium, Pb_3O_4, is a brilliant orange-red pigment. It is insoluble in water and ethanol, but dissolves in acetic acid or hot hydrochloric acid. Red lead is manufactured by heating lead monoxide in a reverberatory furnace in the presence of air at 450 – 500 °C until the desired composition is obtained. It is used as a pigment in anticorrosion paints for steel surfaces and in 1984 accounted for U.S. shipments of 11 236 t [51]. It is also used in lead oxide pastes for tubular lead – acid batteries [52], in ballistic modifiers for high-energy propellants [53], in ceramic glazes for porcelain [54], in lubricants for hot pressing metals [55], in radiation-shielding foam coatings in clinical X-ray exposures [56], and in rubber adhesives for roadway joints [57].

2.5. Lead Silicates

Lead silicates are produced commercially to specific PbO/SiO_2 ratios required by the glass and ceramic, paint, rubber, and plastics industries. Some physical properties of lead silicates are given in Table 5.

Lead monosilicate, lead pyrosilicate, $3\,PbO \cdot 2\,SiO_2$, is a white, trigonal crystalline powder. It is insoluble in water and comprises 85 wt % PbO and 15 wt % SiO_2. Commercial lead monosilicate is produced by dry roasting lead monoxide and silica in the mole ratio 3 : 2. It provides the most economical method of formulating lead-bearing glazes for the ceramics industry and it is also used in the glass industry as a source of PbO.

Lead bisilicate, $PbO \cdot 0.03\,Al_2O_3 \cdot 1.95\,SiO_2$, is a pale yellow powder. It is insoluble in water and has the composition of 65 wt % PbO, 1 wt % Al_2O_3, and 34 wt % SiO_2. It is available in granular form (10 mesh) and as a ground powder (325 mesh). Lead bisilicate was developed as a low-solubility source of lead in ceramic glazes for foodware, for which its low volatility and high viscosity are equally important.

Tribasic lead silicate, $3\,PbO \cdot SiO_2$, is a reddish-yellow powder. It is sparingly soluble in water and has the composition of 92 wt % PbO and 8 wt % SiO_2. It is available in granular form (10 mesh) and as a powder (325 mesh). Tribasic lead silicate has the lowest viscosity of the three commercial lead silicates and it is used primarily by glass and frit producers.

2.6. Lead Sulfates

Lead forms a normal and several basic sulfates. The most important compound is tribasic lead sulfate because of its extensive use in the plastics industry.

Tribasic lead sulfate [*12202-17-4*], $3\,PbO \cdot PbSO_4 \cdot H_2O$, M_r 890.93, ϱ 6.9 g/cm^3, refractive index 2.1, is a fine white powder. Its water solubility is 26.2 mg/L at 18 °C. Tribasic lead sulfate is made by boiling aqueous suspensions of lead oxide and lead sulfate. The anhydrous compound decomposes at 895 °C [59]. The addition of 2 – 7 % tribasic lead sulfate to flexible and rigid poly(vinyl chloride) provides efficient, long-term, economical heat stability. The compound is easily dispersible, has excellent electrical insulation properties, and is an effective activator for azodicarbonamide blowing agents for vinyl foams.

2.7. Other Lead Salts

Several salts of lead are used as heat stabilizers in the processing of poly(vinyl chloride),

Table 5. Physical properties of lead silicates [58]

Compound	CAS registry no.	M_r	*mp*, °C	ϱ, g/cm^3	n_D^{20}
Lead monosilicate	[*10099-76-0*]	294.85	700 – 784	6.50 – 6.65	2.00 – 2.02
Lead bisilicate	[*11120-22-2*]	343.37	788 – 816	4.60 – 4.65	1.72 – 1.74
Tribasic lead silicate	[*12397-06-7*]	729.63	705 – 733	7.52	2.20 – 2.24

alone or in combination with a basic lead salt, such as tribasic lead sulfate. These salts include dibasic lead phthalate [*17976-43-1*]; basic lead sulfate phthalate; and dibasic lead phosphite. In addition, dibasic lead stearate is sometimes employed as a combination internal lubricant and heat stabilizer in poly(vinyl chloride). Lead diamyl dithiocarbamate [*36501-84-5*] has been used as an antioxidant to retard the increase of viscosity and hardness of road asphalts at a loading of 2 – 6 wt % in paving trials in the United States, Canada, and Australia [60]. Further field trials are underway to test the commercial feasibility and environmental acceptability of lead-stabilized paving asphalts.

3. Organolead Compounds

Organolead compounds are those containing one or more lead atoms covalently bonded to at least one carbon atom. Unlike the inorganic compounds of lead, nearly all organolead compounds are derived from tetravalent lead. Table 6 gives examples of organolead compounds in which from one to four carbon atoms are directly attached to a lead atom.

Table 6. Typical organolead compounds [61]

Compound	M_r	*mp*, °C	Uses
$PhPb(OAc)_3$	461.43	102	Polyurethane foam catalyst [62]
$Bu_2Pb(OAc)_2$	439.50	103	Anthelmintic for tapeworms [63]
$Ph_3Pb(OAc)$	497.55	205	Toxicant for shipbottom paints [64]
Et_4Pb	323.45	−137	Anti-knock agent in gasoline [65]

Properties. In general, organolead compounds are well-characterized substances that are quite stable at room temperature. However, they are thermally the least stable of the organometallic compounds of group 14. Most organolead compounds are liable to severe decomposition on heating to 100 – 200 °C, but they are not explosive.

Stable organolead compounds are mostly derived from tetravalent lead and most fall within the four basic categories shown in Table 6, namely, $RPbX_3$, R_2PbX_2, R_3PbX, and R_4Pb (where R is an alkyl or aryl group; and X is a halogen, OH, or an acid radical). The R_2Pb species have limited stability and tend to form R_4Pb upon heating. The tetraorganolead compounds range from the liquid, volatile tetramethyllead [*75-74-1*], $(CH_3)_4Pb$, to the crystalline, nonvolatile tetraaryllead compounds. Some physical properties of tetramethyllead and tetraethyllead are given in Table 7.

Production. The best known industrial process for producing tetramethyllead and tetraethyllead is the reaction of the sodium – lead alloy PbNa with methyl chloride and ethyl chloride, respectively. For example, tetraethyllead [*78-00-2*] is produced according to the reaction:

$$4\,PbNa + 4\,C_2H_5Cl \rightarrow (C_2H_5)_4Pb + 4\,NaCl + 3\,Pb$$

Tetramethyllead (TML) [*75-74-1*] is also manufactured by an electrolytic Grignard reaction:

$$2\,CH_3MgCl + 2\,CH_3Cl + Pb \rightarrow (CH_3)_4Pb + 2\,MgCl_2$$

In a third lead antiknock production process, a chemically redistributed mixture of tetraalkyllead compounds results. Thus, in the presence of selected Lewis acid catalysts (for example, aluminum halides or boron trifluoride), a mixture of TML and TEL undergoes a redistribution reaction in which an equilibrium mixture of the five possible tetraalkyllead compounds is produced.

Economic Aspects. There are three major industrial producers of TEL and lead alkyl antiknock additives for gasoline: Ethyl Corporation, du Pont, and Associated Octel. Data from

Table 7. Physical properties of tetramethyllead and tetraethyllead

Compound	M_r	*mp*, °C	*bp*, °C (kPa)	ϱ, g/cm^3	n_D
Tetramethyllead	267.35	− 27.5	110 (101)	1.995	1.5120
Tetraethyllead	323.45	− 137	91 (2.53)	1.659	1.5195

the International Lead and Zinc Study Group in London indicate a decline of ca. 66 % in the world consumption of refined lead for gasoline additives between 1970 and 1986. Thus, from a peak consumption of 365 200 t in 1970 (excluding Eastern-bloc countries), environmental regulations have resulted in a decline to 124 000 t of refined lead for gasoline additives in 1986 [65]. The limits in 1988 on lead content of gasoline in various countries are given in Table 8. These limits were imposed as a means of meeting automotive emission standards. For catalyst-equipped cars, unleaded gasoline must be used to avoid poisoning of the catalytic converter.

Table 8. Limits on lead content of gasoline in 1988 [66]

Country	Maximum lead in gasoline, g/L
United States	0.026
Canada	0.291
Italy, France, Spain, Portugal	0.399
Other European Community countries	0.151
Australia	0.304 – 0.840
Austria, Sweden, Finland	0.151
Korea	0.301
Venezuela	0.840
Japan	0.000

Uses. Tetraethyllead is used worldwide in gasoline to prevent preignition knock in internal combustion engines prior to completion of the compression stroke. Accordingly, tetraethyllead is commonly called an antiknock agent. This application was for many years the single largest market for organometallic compounds. In addition to providing antiknock qualities to gasolines, organolead additives prevent valve seat wear in engines designed for leaded gasoline.

The preparation and properties of other organolead compounds are described in [67]. Thiomethyl triphenyllead is used as an antifungal agent [68], cotton preservative [69], and lubricant additive [70]; thiopropyl triphenyllead as a rodent repellant [71]; tributyllead acetate as a wood preservative [72] and cotton preservative [73]; tributyllead imidazole as a lubricant additive [74] and cotton preservative [75]; and tributyllead cyanamide and tributyllead cyanoguanidine as lubricant additives [76], [77].

4. Occupational Health

Lead is toxic in all forms, but to different degrees, depending upon the chemical nature and solubility of the lead compound. In general, organolead compounds are more toxic than inorganic lead salts if ingested, inhaled, or absorbed by humans. Children are more sensitive than adults to the effects of lead exposure. Such exposure may be acute (large single dose) or chronic (repeated low doses). Prolonged absorption of lead should be avoided because the toxic effects are many, cumulative, and severe.

The most common indications of the onset of lead poisoning (plumbism) are gastrointestinal, including nausea, diarrhea, anorexia, and constipation followed by colic. Since these symptoms of plumbism may be similar to those of other ailments, blood and urine tests should be made on the basis of prior exposure to lead, clinical symptoms, evidence of increased lead absorption, and significant biochemical changes. Lead absorption can affect hemoglobin synthesis, red blood cell survival time, and neuromuscular function. In addition, the toxic effects of lead in women may include decreased fertility, increased abortion, and neonatal morbidity. Encephalopathy is the most serious result of lead poisoning. It occurs frequently in children, when they have very high blood lead levels, but seldom in adults.

Sources of lead in the environment include soils near smelters, lead-pigmented paints, automobile exhaust emissions, food and water, and industrial exposure. Each of these sources is being reduced or eliminated through environmental and occupational safety regulations. For example, in the United States, the ambient airborne lead limit is 1.5 µg/m^3, measured over

a calendar quarter, which applies to the exterior plant environment and emissions. Moreover, lead-based paints are banned for residential use; leaded gasoline is being phased out; lower limits of lead in drinking water have been proposed (below 20 parts lead per billion parts water).

As a result of exposure monitoring, respiratory protection, protective work clothing and equipment, isolated dining areas, and medical surveillance, there is a very low incidence today of lead poisoning from industrial exposure. Lead-exposed employees in the United States, for example, must be removed from work if the average of their last three blood lead determinations is at or above 50 micrograms lead/100 g whole blood and the airborne lead level is at or above 30 $\mu g/m^3$ [78]. The permissible exposure limit (OSHA-PEL) of lead for industrial workers is 50 $\mu g/m^3$, averaged over an 8-h period. The TLV/TWA for lead is 0.15 mg/m^3, while the MAK value is 0.1 mg/m^3. Tetramethyl- and tetraethyllead have an MAK of 0.05 mg/m^3 (BAT-Wert 50 mg/l) [79].

5. References

General References

1. H. E. Brown: *Lead Oxide – Properties and Applications,* Int. Lead Zinc Res. Organization, Inc., Research Triangle Park, USA, 1985.
2. D. Greninger, V. Kollonitsch, C. H. Kline: *Lead Chemicals,* International Lead Zinc Research Organization, Inc., Research Triangle Park, NC, USA, 1975.
3. *Gmelin,* Syst. No. 47.
4. *Kirk-Othmer,* **14,** 160 ff.
5. H. Shapiro, F. W. Frey: *The Organic Compounds of Lead,* Interscience, New York 1968.
6. L. C. Willemsens: *Organolead Chemistry,* International Lead Zinc Research Organization, Inc., Research Triangle Park, NC, USA, 1964.
7. L. C. Willemsens, G. J. M. van der Kerk: *Investigations in the Field of Organolead Chemistry,* International Lead Zinc Research Organization, Inc., Research Triangle Park, NC, USA, 1965.

Specific References

8. R. C. Weast (ed.): *CRC Handbook of Chemistry and Physics,* 68th ed., CRC Press, Inc., Boca Raton, Florida 1987 – 1988.
9. D. Greninger, V. Kollonitsch, C. H. Kline: *Lead Chemicals,* International Lead Zinc Research Organization, Inc., Research Triangle Park, NC, 1975, p. 296.
10. Teijin Ltd., JP 18 963, 1960 (S. Futami); *Chem. Abstr.* **62** (1965) 16 413 a.
11. H. K. Porter Co., Inc., US 3 112 772, 1961 (R. E. Connor, J. C. Kitching); *Chem. Abstr.* **60** (1964) 4320 g.
12. Shell International Research, NL 6 511 631, 1965; *Chem. Abstr.* **65** (1966) 2046 h.
13. F. Kato, Y. Machino, JP 3 910, 1960; *Chem. Abstr.* **60** (1964) 2781 c.
14. General Electric Company, US 3 039 989, 1959 (W. O. Eastman); *Chem. Abstr.* **57** (1962) 8746 b.
15. Tempil Corporation, FR 1 335 076, 1962; *Chem. Abstr.* **60** (1964) 4309 a.
16. Shell International Research, NL 6 509 144, 1964; *Chem. Abstr.* **64** (1966) 15 661 b.
17. Rosendahls Spinnerier A/S, NO 105 799, 1962 (T. Ringvold); *Chem. Abstr.* **63** (1965) 4460 a.
18. Toshiba Corporation, JP-Kokai 60 169 155, 1984 (K. Kobayashi); *Chem. Abstr.* **104** (1986) 173 110 m.
19. Daido Metal Co., Ltd., US 4 312 772, 1979 (S. Mori); *Chem. Abstr.* **96** (1982) 167 214 s.
20. Riedel-de Haen AG, EP-A 167 949, 1986 (E. Scholz, H. J. Luttmann); *Chem. Abstr.* **104** (1986) 114 518 w.
21. Sumitomo Light Metal Industries, Ltd., JP-Kokai 6 109 995, 1984 (K. Ito et al.); *Chem. Abstr.* **104** (1986) 191 348 t.
22. Nippon Sheet Glass Co., Ltd., JP-Kokai 60 137 852, 1983 (J. Nishii, T. Yamagishi); *Chem. Abstr.* **104** (1986) 55 204 r.
23. Canon K.K., JP-Kokai 59 189 566, 1983; *Chem. Abstr.* **102** (1985) 175 002 d.
24. in [2]p. 173.
25. Chloride Group Ltd., GB 1 601 117, 1978 (S. Worrell, N. E. Bagshaw); *Chem. Abstr.* **96** (1982) 165 647 m.
26. Mordavia State University, SU 969 695, 1980 (V. P. Selyaev et al.); *Chem. Abstr.* **98** (1983) 180 456 h.
27. Council of Scientific and Industrial Research (India), IN 153 507, 1979 (S. Bhattacharya et al.); *Chem. Abstr.* **102** (1985) 118 196 w.
28. Matsushita Electric Works, Ltd., JP-Kokai 57 187 379, 1981; *Chem. Abstr.* **98** (1983) 199 981 x.

29. Agency of Industrial Sciences and Technology, JP-Kokai 57 188 663, 1981; *Chem. Abstr.* **98** (1983) 130 513 d.
30. Nippon Sheet Glass Co., Ltd., JP-Kokai 60 246 242, 1984 (J. Nishii, T. Yamagishi); *Chem. Abstr.* **104** (1986) 154 385 g.
31. Hooker Chemical Corporation, FR 2 039 700, 1969 (J. A. Peterson, H. W. Marciniak); *Chem. Abstr.* **75** (1971) 130 482 t.
32. Nippon Sheet Glass Co., Ltd., JP-Kokai 6 144 734, 1984; *Chem. Abstr.* **104** (1986) 229 306 t.
33. Mitsubishi Petrochemical Co. Ltd., JP 7 499 394, 1973 (Y. Fujita, C. Kaya); *Chem. Abstr.* **82** (1975) 113 096 x.
34. L. Kacetl, J. Pantoflicek, L. Sramek, F. Anyz, CS 121 444, 1965; *Chem. Abstr.* **67** (1967) 118 693 q.
35. Fuji Photo Film Co., Ltd., DE-OS 2 312 675, 1973 (E. Inoue et al.); *Chem. Abstr.* **79** (1973) 151 686 k.
36. Hitachi Maxell, Ltd., JP-Kokai 59 141 423, 1983; *Chem. Abstr.* **101** (1984) 233 249 g.
37. Tomsk Polytechnic Institute, SU 956 314, 1980 (M. A. Shustov, V. K. Zhuravlev, D. G. Kulagin); *Chem. Abstr.* **97** (1982) 205 799 s.
38. G. Smoluk, *Mod. Plast.* **56** (1979) no. 9, 74.
39. W. D. Woodbury, *Preprint from the 1985 Bureau of Mines Mineral Yearbook, Lead,* U.S. Department of the Interior, p. 19.
40. in [2] pp. 59 – 60.
41. H. E. Brown: *Lead Oxide Properties and Applications,* International Lead Zinc Research Organization, Inc., Research Triangle Park, NC, 1985.
42. in [2]p. 69.
43. N. M. Mamaev, S. M. Gubarev, V. G. Altynov, *Izv. Sev. Kavk. Nauchn. Tsentra Vyssh. Shk. Tekh. Nauki* (1980) no. 1, 49 – 52; *Chem. Abstr.* **93** (1980) 221 426 r.
44. United States Dept. of the Army, US 3 996 080, 1971 (J. S. Stack); *Chem. Abstr.* **86** (1977) 108 814 h.
45. Mitsubishi Heavy Industries, Ltd., JP 7 932 435, 1973 (K. Ueda); *Chem. Abstr.* **92** (1980) 66 846 m.
46. Mitsui Mining and Smelting Co., Ltd., JP 7 628 077, 1972 (A. Matsumoto, E. Isobe); *Chem. Abstr.* **87** (1977) 28 337 k.
47. Toyo Rubber Industry Co., Ltd., JP 8 030 543, 1976; *Chem. Abstr.* **94** (1981) 31 937 f.
48. in [37]
49. Societe des Accumulateurs Fixes et de Traction (SAFT), US 4 271 244, 1978 (J. P. Gabano, M. Broussely); *Chem. Abstr.* **95** (1981) 105 363 k.
50. V. Janda et al., CS 189 225, 1976; *Chem. Abstr.* **97** (1982) 75 012 s.
51. in [39] p. 18.
52. Chloride Group Ltd., GB 1 508 991, 1974 (K. Peters); *Chem. Abstr.* **90** (1979) 8891 v.
53. in [44]
54. H. Hayashi, JP-Kokai 8 015 969, 1978; *Chem. Abstr.* **93** (1980) 119 147 p.
55. S. T. Basyuk, A. M. Zinder, A. A. Savatyugin, V. M. Kochetkov, SU 540 907, 1974; *Chem. Abstr.* **87** (1977) 25 819 h.
56. Nippon Akuriru Kagaku K.K., JP-Kokai 7 796 636, 1976 (S. Kawamura, S. Watanabe, Y. Shikimori); *Chem. Abstr.* **88** (1978) 39 071 r.
57. Continental Gummi-Werke AG, BE 870 811, 1977 (R. Sohnemann); *Chem. Abstr.* **91** (1979) 22 198 b.
58. *Kirk-Othmer,* **14,** 173.
59. in [2] p. 254.
60. S. F. Radtke, D. S. Carr: "Lead Stabilized Asphalt," *Proc. Fifth Intl. Asphalt Conf.,* Australian Asphalt Pavement Association, Perth, Australia, Feb. 1983.
61. L. C. Willemsens, G. J. M. van der Kerk: *Investigations in the Field of Organolead Chemistry,* International Lead Zinc Research Organization, Inc., Research Triangle Park, N.C., USA, 1965.
62. International Lead Zinc Research Organization, Inc., US 3 324 054, 1967;US 3 417 113, 1968; US 3 547 481, 1970 (H. G. J. Overmars).
63. International Lead Zinc Research Organization, Inc., US 3 595 964, 1971 (G. Gras).
64. D. S. Carr: "Triphenyllead Acetate – Organolead Antifoulant Passes Tests," *Paint and Varn. Prod.* **64** (1974) no. 12, 19 – 23.
65. H. Shapiro, F. W. Frey: *The Organic Compounds of Lead,* Interscience, New York 1968, pp. 399 – 403.
66. D. R. Lynam: *The Past, Present and Future of Lead in Gasoline,* Ethyl Corporation, Baton Rouge, LA, USA, 1988.
67. in [61]
68. International Lead Zinc Research Organization, Inc., US 3 322 779, 1967;US 3 522 352, 1970 (M. C. Henry, A. W. Krebs).
69. International Lead Zinc Research Organization, Inc., US 3 420 700, 1969 (D. J. Donaldson, W. A. Guice, G. L. Drake, Jr., W. A. Reeves).

70. International Lead Zinc Research Organization, Inc., US 3 287 265, 1966 (W. L. Pearlstein).
71. International Lead Zinc Research Organization, Inc., US 3 683 090, 1972 (M. C. Henry, A. W. Krebs).
72. G. J. M. van der Kerk: "New Developments in Organolead Chemistry," *Ind. Eng. Chem.* **58** (1966) no. 10, 29 – 35.
73. International Lead Zinc Research Organization, Inc., US 3 420 701, 1969 (D. J. Donaldson, W. A. Guice, G. L. Drake, Jr., W. A. Reeves).
74. International Lead Zinc Research Organization, Inc., US 3 655 683, 1972 (L. C. Willemsens).
75. in [73]
76. International Lead Zinc Research Organization, Inc., US 3 527 704, 1970 (W. L. Pearlstein, H. A. Beatty).
77. International Lead Zinc Research Organization, Inc., US 3 436 414, 1969 (C. J. Worrell).
78. Federal Register (U.S.), vol. 43, Nov. 14, 1978, pp. 52 952 – 53 014, as corrected in vol. 44, April 6, 1979, pp. 20 680 – 20 681.
79. DFG, MAK- und BAT-Werte-Liste 2001, Wiley-VCH, Weinheim, 2001.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Leather

Standard Article

Günter Reich[1]

[1]Freiberg/Sachsen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a15_259.pub2
Article Online Posting Date: April 30, 2005

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Economic Significance of Leather
3. Unit Operations of Leather Manufacturing
3.1. The Main Processes
3.2. Chemicals
3.3. Reaction Vessels
3.4. Machines
4. The Raw Material: Hides and Skins
4.1. Production of Hides and Skins, Preservation
4.2. Topographical and Histological Structure and Chemical Composition of Animal Hides
5. Structure and Reactivity of Collagen
5.1. Occurrence and Function of Collagen
5.2. From α-Chains to Fiber Network: The Building Blocks of Collagen
5.3. Role of Water in Collagen
5.4. Biosynthesis of Collagen
6. Processes Preparing the Hides and Skins for Tanning
6.1. Soaking
6.2. Liming, Deliming, Bating, Degreasing
6.2.1. The Use of Enzymes in Leather Processing and in Collagen Upgrading in General
7. Tanning Processes
7.1. General Objectives and Results of Tannage
7.2. Chrome Tannage
7.3. Free-of-Chrome (FOC) Tannage
7.4. Vegetable Tannins and Tannage
7.5. Aromatic Syntans
7.6. Other Tannages including Modern Combination Tannages
8. Processes After Tanning
8.1. Retanning
8.2. Dyeing
8.3. Chemical Lubrication (Fatliquoring), Hydrophobization
8.4. Drying
8.5. Mechanical Softening (Staking, Milling)
9. Leather Finishing
10. Ecological Aspects of Leather Manufacturing
10.1. Ecotoxicology and Occupational Hygiene
10.2. Water and Wastewater
10.3. Solid Byproducts and Wastes
10.4. Air Pollution

11. Leather Properties and Testing
12. Leather Substitutes

[Top of Page]

1. Introduction

Leather is an ancient material and so is the profession of tanner, as shown by old findings of leather and leather goods in different cultures [1]. The tanner is mentioned in very early literature, for example, ARISTOTLE cites SOCRATES: “There are four professions to be the fundamentals of the citation of state: The architect, the weaver, the farmer and the tanner” and the bible mentions “Simon, who was a tanner” [2]. The reason for this importance of leather is clear: Humans needed protection against cold, wetness, and mechanical influences, and hides and skins were available from hunted animals (leather for shoes, garments, gloves, etc.). But leather and leather goods are not only manufactured as a protecting material but also as a fashionable material with aesthetic qualities, e.g., leather for fancy goods, handbags, book binding, and even carpetries, tapetries, etc. [3]. And this double function of leather is still maintained, as the interior of an automobile shows (Fig. 1).

Figure 1. The beauty of leather (example: leather in motor cars)

The many manufacturing processes, types, and properties of leather articles make definition of this material difficult. One possible definition is as follows: “Leather is a material manufactured from the collagen fiber network of hides and skins, endowed with the characteristics required for the intended purpose and produced by suitable chemical and physical processes”.

Thus, collagen is the raw material for the manufacture of leather. After a brief description of the industrial and economical significance of hides, skins, and leather and a survey of the unit operations of leather-making processes, the structure and reactivity of collagen are discussed, as well as the changes to which collagen is subjected during its transformation into leather and how the properties of the latter are influenced by collagen.

[Top of Page]

2. Economic Significance of Leather

Hides and skins, primarily those of cattle, sheep, goats, and pigs, are byproducts of human foodstuffs and wool. Collagen is thus a renewable raw material, the production of which grows in proportion to cattle stock, which in turn is growing in proportion to the earth's human population. The ratio of cattle stock to world human population between 1900 and 2000 remained relatively stable at 0.25 – 0.30.

To get an impression of the economical significance of leather production it is useful to study FAO sources [4], as summarized by BULJAHN [5] and SCHRÖER [6]. World leather production is growing slightly for the reasons mentioned above. Leather production in the developed countries is continuously decreasing, while it is increasing in developing countries, a general trend mainly due to different labor costs. For example, Germany today is an exporter of high-quality cattle hides but only a rather minor leather manufacturer. In contrast, China has recently become a country with a continuously growing leather industry and an major importer of hides, skins, and intermediate products (“wet blue”). Its light leather production rose from 79.5×10^6 m^2 in 1990 to 690.5×10^6 m^2 in 2002. In general, China, India, and Brazil have the elements required for rapid success, and the rest of the world will have to fight for market shares and special niches.

Worldwide, shoe leathers (uppers and linings) are the dominant leather types. World shoe production amounted in 2003 to about 12×10^9 pairs. China alone produces about 7×10^9 pairs, of which 4.2×10^9 pairs are for export. For comparision, Italy, the leading European shoe manufacturer, produces 312×10^6 pairs, of which 297×10^6 are for export [7].

Upholstery leather, especially automotive leather, is winning an increasing market share. In Germany, for example, more than 80 % of all leather types produced are upholstery leather (20×10^6 m^2 in 2001) because of the high quality of the available raw stock (Table 1).

Table 1. The economical importance of the global leather production in developing and developed countries 1990 : 2000 (according to FAO [4])

Type of leather/region	1990	2000	Change 1990 to 2000, %
Light leather, bovine, 10^6 m^2			
Light leather, bovine, 10^6 m^2			
world	872	1027	+18

developing countries	391	619	+57
developed countries	476	413	–13
Light leather, goat/sheep, 10^6 m^2			
world	342	425	+24
developing countries	202	305	+50
developed countries	104	102	–2
Heavy leather, 10^3 t			
world	381	494	+30
developing countries	183	307	+68
developed countries	198	187	–6

An estimated 7 × 10^6 t of hides and skins containing about 2.8 × 10^6 t of collagen are used yearly worldwide for leather production. For technical reasons only 50 % of the hide collagen is converted into leather. Therefore, the utilization of the remaining 50 % of the collagen remains an acute challenge. Some other important collagen-based materials are gelatin, leather boards, casings, medical products, and dog chews (Table 2) [8, 9].

Table 2. Approximate annual production of important collagen-based products (2000)

Collagen-based products	Estimated quantity/proportions
Light leather	1500×10^6 m^2
Bovine	875×10^6 m^2
Sheep/goat	345×10^6 m^2
Pig	250×10^6 m^2
Heavy leather	500×10^3 t
Light leather	100 %
Shoe leather	55 %
Garment leather	20 %
Furniture and automotive leather	15 %
Fancy goods	10 %
Leather board	100×10^3 t
Gelatines/glue	400×10^3 t
Casings (for sausage, etc.)	> 1 × 10^9 running meters of different diametres
Hydrolysates	> 200×10^3 t
Carcass meal (animal feed/fertilizer)	> 2×10^6 t
Collagen in cosmetics, medicine, and tissue engineering (exept gelatin)	3–5×10^3 t

The raw material is the main source of cost in leather manufacturing. The hides and skins account for more than 50 % of leather production costs, and hence leather prices depend on raw-hide prices and their fluctuation. For example, cow and bull hides in France cost, depending on mass class, 1.4 – 1.6 €/kg green weight (October 2004). Assuming an average yield of 0.15 m^2 grain leather/kg rawhide 1 m^2 of grain leather has raw hides costs of about 10 €. Grain leather, the former hair side, is the most valuable part of the hide.

Other important cost elements are chemicals and water, wages, manufacturing and overhead costs, and costs for energy and waste disposal (Fig. 2).

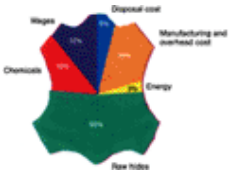

Figure 2. Approximate cost structure of leather manufacturing [10]

[Top of Page]

3. Unit Operations of Leather Manufacturing

Although today leather is a highly sophisticated product and manifold developments of chemicals and machines for leather making took place, mainly in the 1900s [11], the unit operations of leather making remain unchanged. The path from hides and skins of the living animals to finished leather ready for further processing is long and requires many unit operations, very different reaction vessels and machines, manifold chemicals, and a large amount of water.

Useful textbooks on leather manufacturing can be found in [12-18], and a brief introduction to leather manufacturing in [19].

3.1. The Main Processes

Leather manufacturing begins with animal husbandry: The better the hide quality, the better the leather. Two aspects are important here: First, meat is normally more valuable than hides and skins. Second, the hides are at risk of damage by insects and bacterial influences, mechanically caused grazes or flaws, brand marks, and other reasons. Therefore, tanners and farmers are in a permanent discussion of how to improve hide quality and to harmonize their different economical interests.

In the slaughterhouses a careful separation of the hides from the carcasses by flaying is also necessary to ensure high quality. Slaughtering defects like deep cuts, gouges, and even holes should be avoided.

For the time between flaying and processing in the tannery, hides and skins must be protected against microbial attack and putrefaction by preservation. In principle chilling and “instant manufacturing” (short-term preservation), (wet) salting, and drying are the main methods of preservation. The thus-prepared hides and skins arrive at the tanneries in a “green”, “salted”, or “dry” state.

In the tanneries processing begins, often after some storage in well-chilled rooms, with the cleaning of the hides and skins by soaking. It is followed by liming, bating, deliming, all associated with intensive rinsing or with “water-saving” washing. The corium, the leather-forming layer of the hide, must be cleaned from the epidermis along with hairs by liming, and the subcutis (“flesh”) must be removed by defleshing and possibly by degreasing (mainly sheep skins). Often splitting (“in the lime”) of the resulting corium into two or sometimes even three layers (grain and flesh splits) defines the thickness of the future leather, depending on its purpose. This operation takes place in the so-called beamhouse. The name refers to the old tanner's beam, a thick trunk used for manual defleshing a long time ago. The resulting product is called pelt.

It is followed by proper tanning, using inorganic tanning salts and/or natural and synthetic organic tannins. The most important inorganic tannins are basic chromium(III) sulfates. To a much smaller degree basic aluminum sulfates are used. Zirconium, iron, and titanium compounds are only occasionally used as tannins. Among the organic tannins the phenolic vegetable tannins have long been very important. Their substitutes since the 1920s are the so-called aromatic/phenolic syntans. More recent and of growing importance are the polymeric tannins, and also glutaraldehyde and its derivates.

Tanning takes place in the tanyard. In this unit operation the collagen is transformed into the “stabilized” leather matrix, and the pelt is converted to leather. In this state the leather is a wet intermediate product, called “wet blue” in the case of chrome tanned leather, and “wet white” in the case of free-of-chrome (FOC) tanned leather. The names come from the typical color of these intermediate products. The tanned intermediate product is partially mechanically dewatered by sammying (dewatering by feeding through a roller press) and its thickness regulated by shaving (adjustment of thickness by a rotating knife-cylinder). Some tanners split wet blue (“in the blue”) instead of pelts (“in the lime”), but in the latter case there are better possibilities to utilize the collagen waste that is unusable for leather.

During the following wet-end processes the leather is retanned to make the feel of (dry) leather fuller and softer, it is dyestuffed with “aniline” dyestuffs (azo dyestuffs) and chemically lubricated by fatliquoring with different fat emulsions. After washing, the leathers are again mechanically dewatered, and wrinkles and creases are removed by setting out (mechanical dewatering and stretching on a roller machine), often combined with sammying. The leather is then dried and mechanically softened by stretching, called staking, to avoid fibers sticking together. The resulting product is called crust leather.

By finishing the surface of the crust leather with polymer binders and lacquers it is upgraded to the final leather with the grain structure, color, and other desired properties depending on the application. Objectives are protection of the surface, covering defects, and last but not least realization of fashionable effects. The finish layers are thin films of different polymers, containing pigments and some other components, nowadays mainly applied as aqueous dispersions by roll coating and/or spraying (Fig. 3).

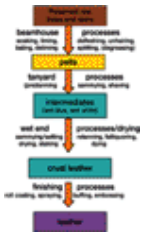

Figure 3. The unit operations of leather manufacturing

All the above processes need chemicals, which interact permanently or temporarily with the collagen.

3.2. Chemicals

To convert hide collagen into leather many different chemicals, from simple inorganic salts to highly sophisticated special organic compounds, are necessary. Some of them play only a temporary role but the most important substances — tannins

and fillers, dyestuffs, fats, and filling and finishing polymers — remain in or on the leather.

An impression of the economical significance of the typical “leather chemicals”, including the high water consumption, is given by approximate requirements to produce 1000 m^2 of chrome-tanned shoe upper leather (Table 3).

Table 3. Approximate demand for water and chemicals to produce 1000 m^2 of chrome-tanned grain leather (ca. 7.1 t of wet salted hide) and as byproduct 430 m^2 of split leather from cattle hide [20]

Substance	Quantity, kg (unless otherwise stated)
Water*	215 m^3
Inorganic salts, mainly NaCl*	570
Inorganic and organic acids*	30
Sodium sulfide*	175
Calcium hydroxide*	285
Enzymes*	20
Bactericides*	20
Tensides*	20
Chrome extract (Cr_2O_3)**	700 (175)
Vegetable tannins**	50
Aromatic syntans**	50
Polymers**	50
Resins**	10
Fatliquors**	150
Dyestuffs**	35
Polymer binder (finishing)**	30

*Temporarily applied.
**Permanently fixed on collagen in an amount of 85–98 %.

Relative to the world production of leather (see Table 1) one can see the large quantity and the major economic importance of the most important leather chemicals (Table 4).

Table 4. Quantities of the most important leather chemicals used for world leather production in 2000 (according to [10])

Substance	Quantity, 10^3 t/a (unless otherwise stated)
Water	320×10^6 m^3
Tensides	120
Lime hydrate	300
Sodium sulfide	200
Sodium chloride	270
Chrome extracts (Cr_2O_3)	1600 (400)
Vegetable tannins	300
Aromatic syntans	150
Glutaraldehyde	30
Resins	30
Polymers	150
Lubricating agents	400
Dyestuffs	90
Polymer binder	200

Leather chemicals (performance-type) had a worldwide value of about 3.5×10^9 € in 2002, with the following split between the various leather making processes:

Beamhouse	15 %
Tanning	28 %
Dying	12 %
Finishing	30 %
Commodities	15 %

Table 3 shows also another aspect: The temporarily used chemicals, mainly salts, the incompletely exhausted chemicals, and all the process and wash waters end up in the wastewater [20], cleaning of which in accordance with official regulations is technically demanding and expensive. To improve the processes and the chemicals is a permanent challenge for tanners and suppliers of chemicals (see also Chap. Ecological Aspects of Leather Manufacturing).

3.3. Reaction Vessels

The wet processes of leather manufacturing from the beamhouse up to the wet-end operations take place in different vessels. The interaction between the hides and skins and the chemicals dissolved in the float is a diffusion process and can be accelerated by movement of the hides and the floats. Therefore, the pits that were originally exclusively used are now very seldom found; instead, paddles, especially drums, mixers, and Y-machines ("segmented machines" or "three component vessels" in some variations) are the state of art today (Fig. 4).

Figure 4. Principles of different vessels used in leather manufacturing A) Drum; B) Paddle; C) Mixer

Nowadays, pits are used only occasionally in small tanneries and in special cases of vegetable tanning (ancient pit tannage).

Paddles have long been used. Today they are mainly used for manufacturing light skins. Movement of skins and the float improve the penetration of chemicals into the pelt. Also in fur manufacturing paddles are commonly used.

The invention of drums by the brothers DURIO at the end of the 1800s was a key innovation, that revolutionized the tanning processes. Drums have been the most important vessels ever since. They are advantageous because agitation of the float and the hides or pelts is associated with with tumbling and pumping effects. Both improve the penetration of the chemicals into the hides, pelts, or wet leather. Dry tumbling is done in the crust stage to soften the leather and bring out a fine grain of the leather. This process is called milling.

Drums are normally wooden vessels, but also stainless steel, which is better cleanable and is used especially for dry milling, and even plastic drums are known. They can be very large, up to 40 m^3 in volume. Loading and deloading is highly automated, the floats can be temporarily sampled, and the pH, temperature, and other parameters can be controlled. More recently, highly sophisticated drum constructions have been described [21].

An important alternative to the drum is the mixer or processor ("Challenge Cook"), introduced to tanneries about 1970. It is constructed like a concrete mixer. The advantages and working principles are similar to those of drums.

Y-section vessels (tanning machines, e.g., Dosemat) work almost like a drum but they contain three sections divided by perforated walls. The construction ensures a mild but effective treatment, especially in dying and fatliquoring [22].

3.4. Machines

The conversion of hides and skins to leather is mainly a chemical process with many single steps, but in nearly all stages machines are required. The first important processes are dehairing and defleshing of hides. The regulation of the thickness of the untanned or tanned pelts by splitting in the grain and the flesh splits, possibly with an additional middle split, are further steps.The thickness is is also regulated by shaving in the tanned state.

Another task is the dewatering of pelts or leather in the different states. Beside different kinds of presses, the most important dewatering machines are sammying machines.

The drying of leather is performed by hanging them on dryers in the open air or feeding them through a drying tunnel. The leather can also be fixed under a slight tension on frames and dried (toggling). But the most widely used drying device today is the vacuum dryer, especially for drying shoe upper leather.

Softening the leather is not only a question of fatliquoring but also of mechanical softening by staking and milling.

The finishing of leather needs coating machines; the most frequently used are roll coaters, spraying devices, printing machines, curtain coaters, and padding machinery. Feed embossing presses produce an artificial grain if desired.

In total there are more than 30 different types of machines in a modern tannery. They can be categorised in five groups (Table 5).

Table 5. Mechanical work in leather manufacture and the machinery used therefor (according to [23])

Machine	Function

Multiroller machines	
fleshing machines	removing the subcutis (flesh) from the corium by knife rollers
shaving machines	regulating the thickness of intermediate leather (e.g., wet blue) by knife rollers
sammying machines	mechanical dewatering of tanned hides before shaving
setting machines	mechanical dewatering of wet leather before thermal drying
buffing machines	modification of the leather surface by buffing the grain side (nubuck) or the flesh side (velour)
Splitting machines	
band knife splitting machine	regulating the tickness of hides, pelts, or leather by splitting into two or more layers
Staking machines	
vibration staking types	mechanical softening of the crust leather
arm staking types	
rotary staking types	
Drying machinery	
suspension (hang) dryer	thermal dewatering of wet leather to crust leather
toggle drying machines	
vacuum dryer	
paste dryer	
Secotherm dryer	
Finishing machinery	
padding machines	applying aqueous or organic solutions/dispersions of polymers, pigments, etc. to cover the leather surface, modifying the glance, etc.
spraying machines	
rollcoating machines	
curtain coating machines	
reverse coating machinery	
ironing, plating, embossing, glazing	producing an artificial grain
area measuring machines	measuring leather area

General trends are nowadays the application of the so-called feed-through principle wherever possible and the highly sophisticated use of automation and control, mostly computer-aided. It is not the intention of this chapter to go into details. For details, see [23-25].

[Top of Page]

4. The Raw Material: Hides and Skins

Cattle hides, sheep and goat skins, and to a smaller extent pig skins are the most important raw materials for leather production. The hides and skins of some other animals like horses, crocodiles, birds and fish are used for (special) leather manufacturing.

4.1. Production of Hides and Skins, Preservation

Depending on the average lifetime only a part of the mammalian livestock is annually slaughtered (“slaughtering rate”), and even after slaughtering not all hides and skins are rendered as leather raw material (“flaying quota”). Therefore the relationship between livestock and hide supply is somewhat complicated (Table 6).

Table 6. Livestock, amount of hides and skins (pieces and mass), hides and skins recovery rate in 2000 (according to [4])

Kind/region	Livestock in thousand heads	Hides and skins		Hide recovery rate, %
		10^6 pieces	10^3 t	
Bovine				
world	1 509 492	318.9	576.1*	21
developing countries	1 179 943	199.7	310.9*	17
developed countries	329 549	119.2	265.2*	36
Goats and kids				
world	723 385	345.2	246.7**	48
developing countries	690 741	327.9	233.9**	47
developed countries	32 644	17.3	12.8**	53
Sheep and lambs				
world	1 048 331	525.8	398.3**	50
developing countries	665 856	327.5	197.7**	50
developed countries	382 475	198.3	200.6**	51

*Wet salted weight.
**Dry weight.

From the leather-making point of view the hides and skins should be completely intact, and especially the grain layer should be undamaged. Unfortunately, this is rarely the case: Many defects during animal husbandry, cattle transport, slaughtering, and even preservation reduce the value of hides and skins. In general hide quality depends also on the breed, climatic conditions, kind of husbandry, sex, age, etc. Therefore, each hide is individual and all the tanners art is needed to produce a good and constant leather quality.

Instantly after slaughtering/flaying the hides are subject to postmortem changes due to autolytical reactions that are promoted by microbiological attack if not prevented by preservation, that is, the hides and skins must be cured. The oldest method is mild drying of the hides and skins, but this method is not as easy as it seems, as there is a danger of damaging the collagen by drying at too high a temperature or too rapidly. So today drying is rather rarely practiced, mainly in the case of skins.

The classical method of preservation is salting the hides. The dewatering effect of salting hinders bacterial growth. Salt, mainly NaCl, can be applied as solid salt or concentrated salt solutions (brines) in wet salting. The water content of the hides decreases from 70 to < 40 %, and to nearly the same degree salt is incorporated into the hides and must later be removed.

Another kind of preservation with rapidly growing significance is chilling with ice for a certain time during short-term storage and transport to the tanneries. The additional use of bactericides is employed. Sole application of bactericides for short-term curing is used, too, but not so often (Fig. 5).

Figure 5. Fresh hide processing (courtesy of Bader GmbH & Co., Göppingen, Germany)

Some other well-investigated preservation methods use KCl instead NaCl or irradiation with electron or γ beams, but mainly for economic reasons these methods haven't found their way into practical application yet: A radiation device for a large slaughterhouse would cost about 7×10^6 $ [26].

Pickling of skins is state of art in connection with wool production, mainly in New Zealand and Australia.

Finally the instant manufacturing of hides to wet blue in the slaughterhouses or in a plant close by is a very promising future technology.

4.2. Topographical and Histological Structure and Chemical Composition of Animal Hides

The hide's form corresponds to the flaying principle. Concerning leather making the several parts of the hide have different structure of the collagen fiber network and are therefore of different value and specially named (Fig. 6).

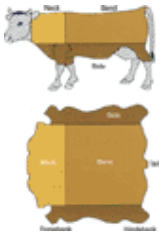

Figure 6. The topographical structure of bovine hide (courtesy of Bader GmbH & Co., Göppingen, Germany)

As a rule the central parts (croupons or bends) are of higher quality than the sides: the collagen fiber network there is more compact and denser than in the sides. To compare mechanical properties of leather it is necessary to define the place where the sample is to be taken. During the leather-making process the structural differences can be somewhat lessened, e.g., by selective filling during retanning, but not completely eliminated. Therefore, the specific structure of the hides must be taken into account during the processing of leather.

Hides and skins consist of three main layers: the epidermis, the dermis (corium or cutis), and the hypodermis (subcutis). Each of the layers can be divided into several sublayers (Fig. 7) [27].

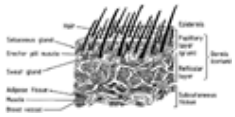

Figure 7. The histological structure of bovine hide

The epidermis is a cellular tissue containing the animal hair or the scales of fishes or reptiles. The hair is embedded in the hair follicles. The hair roots are located in the border layer between epidermis and corium. (Only the bristles of pig skins go through the whole corium.) After removing epidermis and hair prior to tannage by chemical dehairing one can see the characteristic grain structure, which is caused by the actual arrangement of the hair depending on the type of animal. The “grain” of the final leather surface is a typical feature of natural leather.

The main protein of hairs and epidermis is keratin. Typical for this protein is the high content of cystine, an amino acid with a disulfide bridge —S —S — which stabilizes the protein structure. When this bridge is destroyed by alkaline, reductive, oxidative, or even enzymatic treatment the hair dissolves or is damaged, so that dehairing is possible. Secondly, the small zone between the grain of corium and the epidermis, the basal membrane, is especially sensitive to the above-mentioned treatments. So it is possible to target this sensitive layer with milder treatments and destroy the hair roots only. The hair shafts remain intact and the hair can be used for certain applications like felt. The removal of hairs and epidermis by the liming process is discussed in Section Liming, Deliming, Bating, Degreasing.

The subcutis (“flesh”) consists mainly of water, fat, and collagen. It is mechanically removed prior to tanning by machine defleshing. The byproduct “fleshings” (glue stock) can be processed to industrial fat, glue, and collagen hydrolysates, but recently the conversion of the collagen hydrolysates and fat to biogas has been of growing importance.

After dehairing and defleshing the pelt remains, built from the relatively thick corium, consisting of two layers, the papillar and the reticular layer. Both layers contain the fibrous network of collagen, mainly of type I, and very small amounts of some other collagens, proteoglycans, elastin, some fat, blood cells, etc., and mainly water (> 70 %).

The papillar layer is less dense than the reticular layer, and the fibers are thinner. The density of the fiber network, the fiber diameter, and the weave angle depend on the topographical position on the hide. A *Histological Atlas* can be found in [28], and leather defects, with many microscopic images, are treated in [29].

[Top of Page]

5. Structure and Reactivity of Collagen

Collagen plays the key role in the leather making process. Therefore, it is necessary to discuss current knowledge of this important protein. Some introductory books and articles from the viewpoint of leather manufacturing are [30-36].

Collagen (from the Greek for glue-forming) is by far the most widely found protein in terms of quantity, and together with cellulose and lignin, one of the three quantitatively dominant biopolymers. This fact is due to its incidence: it is present throughout the entire animal kingdom with the exception of single-celled organisms.

Its significance derives from its function in living organisms. Mammals are a good example: as a component of the mesenchymal connective tissue, collagen is a primary component of the skeleton and the skin, and thus responsible for protecting and supporting the body; it is, however, also present in numerous other organs and, as part of the extracellular matrix, is also of great significance for cellular biology.

Once mocked in biochemical research as “boring” and being of interest at most to leather chemists for their own purposes, collagen now attracts wide attention in many scientific disciplines, and is in fact one of the most closely studied proteins. This can be attributed to the fact that the role played by collagen in cellular processes, in the growth of connective tissue, and also

in the incidence of certain clinical conditions ("collagen diseases") and for the natural ageing of the organism (a human is as old as his or her connective tissue) has increasingly been recognized.

5.1. Occurrence and Function of Collagen

Collagen is now known not to be a single substance, but rather a family of substances. The collagen types now known, 20 or even more by number, differ in some cases considerably in their chemical and structural characteristics, according to their function in the living organism. A common feature is their conformance (at least partially) to a triple helix ("triple-helix domains"), their content of hydroxyproline, an amino acid which is virtually specific to collagen, and their occurrence in the extracellular matrix. Collagens include the fibril-forming types I, II, III (80 – 90 % of all the collagen in the body), V, and XI. Type I is the dominant collagen type of hides and therefore the most important for leather making (Table 7). Other types control fibrils growth and are associated with the fibrils (types IX, XII, XIV), and there are also structures which form networks and filaments (IV, VI). The relationships between the structure of these different collagens and their significance within the living organism is now well documented [37].

Table 7. Data on the structure and reactivity of collagen which are relevant to leather making and wearing

Structural element	Dimensions	Significance for leather characteristics
Amino acids	3 × 1052 amino acids/mol, of which 381 are acidic (Asp, Glu), 268 basic (Arg, Lys, Hyl His), 492 hydroxyl (Hyp, Ser, Thr, Hyl) and ca. 3000 peptide groups	center of reactivity, including water bonding
Molecules	1.4 × 280 nm, ca. 300 000 Dalton	see fibrils
Microfibrils	∅ 40 nm, 5 molecules	metastable intermediate stage in fibril formation
Fibrils	∅ 100–200 nm, intrafibrillar and interfibrillar intervals of up to 20 nm (capillary region) and fibril surface areas of up to 30 m^2/g, depending on water content	location of intrafibrillar activity during conversion of the collagen to the leather matrix (including "cross-linking")
Elementary fibers (fibril strands)	∅ ca. 2 µm	location of interfibrillar activity during conversion of the collagen to the leather matrix (including "coating")
Fibers	∅ ca. 100 µm, intervals in the > 100 nm range (micropores)	largely responsible for garment comfort properties
Fiber network	∅ on the order of mm, woven, intervals in the order of µm (macropores)	largely responsible for mechanical characteristics

5.2. From α-Chains to Fiber Network: The Building Blocks of Collagen

Like keratin, elastin, and silk fibroin, type I collagen is a scleroprotein, consisting like all proteins of amino acids linked by —CONH— groups to form peptide chains. Of the three chains forming collagen, two are identical in the case of type I collagen, α1(I), and one, α2(I), differs in its amino acid composition: $[\alpha 1(I)]_2$ [α2(I)]. The three α-chains which combine to form the triple helix exhibit nonhelical, short-chain appended peptides (telopeptides) at their N- and C-terminal ends. The telopeptides are mobile, do not form helices, and undergo age-dependent cross-linking. They are sensitive to enzymatic attack.

Of the 19 different amino acids involved in the creation of collagen, a third are accounted for by glycine (gly), always in the position gly-xaa-yaa, and almost a further third by the cyclic amino acids proline (pro) and hydroxyproline (hyp). This is a requirement for the collagen-specific formation of a triple helix.

As already mentioned, hydroxyproline (hyp) is virtually specific to collagen, a fact which can be exploited for its quantitative detection. Hyp is also substantially responsible for the hydrothermal stability of native, fully hydrated collagen: the higher the hyp content of a certain type I collagen (e.g., there are significant differences between collagen of fish and mammals) the higher the hydrothermal stability of the collagen triple helices. Fully hydrated collagen in aqueous solution undergoes a helix – coil transition and is denatured at a defined denaturation temperature T_D. If present as a fiber network, as in the case of hide (and leather), denaturation is manifested by shrinkage to a third of the original length of the sample at the shrinkage temperature T_S. Both temperatures are a measure of the hydrothermal stability. The difference between T_D and T_S of mammalian collagen is constant (ca. 25 K).

The content of acidic and basic amino acids, amino acids containing OH groups, and peptide groups is considerable, and is decisive for the reactivity of the collagen, which is a requirement for its transformation into leather. Tanning reactions are discussed in Section General Objectives and Results of Tannage.

The α-chains of the collagen arrange to form higher structures that give rise to a proper structural hierarchy of the collagen. Three α-chains initially combine to form a three-stranded triple helix with a rise of 0.29 nm per amino acid residue and approximately 3.3 amino acid residues per turn. It is mostly assumed that the helix is stabilized by hydrogen bonds, with formation of intrahelical hydrogen bonds by the hyp residues [38]. Recently, an inductive effect of hyp was proposed as the cause of helix stability, and therefore the role of water in helix stabilization needs further discussion [39].

A collagen molecule (“tropocollagen”) soluble in acid and/or citrate buffer is thus formed the dimensions of which (300 × 1.4 nm) can be viewed by TEM or atomic force microscopy (AFM).

Fibrils are assembled from collagen molecules in vivo and in vitro (the process being triggered in the latter case by a change in pH and/or a rise in temperature). The Smith microfibrils constitute a metastable intermediate stage. The microfibrils in this case are successive layers of five molecules forming a “pentagon”. Partial sequences of the Smith microfibril have frequently been the subject of early computer modeling in tanning chemistry.

Since the late 1950s the creation of the fibrils has been well understood owing to the work of KÜHN [40]. Under the influence of their charge pattern, microfibrils are taken up laterally by each other, staggered by a quarter of their length up to an average thickness of 100 nm, and grow longitudinally apparently without limit and with formation of “gaps” between the molecules (see Fig. 10).

The fibrils are accessible to detailed characterization by TEM, SEM, and AFM. The cross-striation pattern typical of collagen with a periodicity of 64 – 67 nm (“D”) was established at an early stage by this means. As the assembly is dependent on the charge pattern, fibril formation can be influenced by modification of the charge pattern. It was thus demonstrated some time ago that transformations on the basic groups involving tungsten triphosphate or on the acidic groups of the collagen involving uranyl compounds cause the typical cross-striation to be highly resolved and more sharply expressed due to the “bringing into register” of the amino acids involved, that is, of their lateral chains - an observation of interest also for tanning chemistry (Fig. 8).

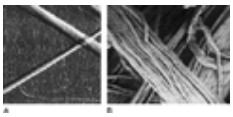

Figure 8. A) A single collagen fibril with significant cross striation (SPM image, courtesy of M. Tuckermann) B) Collagen fibril bundles (SEM image, courtesy of M. Tuckermann)

Using scanning probe microscopy (SPM) with a resolution on the nanometer scale in air and water and by micromanipulation of (tendon) collagen GUTSMANN concluded that the collagen fibril is an inhomogeneous structure composed of a relative hard shell and a softer, less dense core showing a tubelike mechanical behavior [41].

The fibrils continue to grow to form elementary fibers and fibers. The latter then form the fiber network with its great relevance to (mechanical) leather properties. The arrangement of the fibrils can be seen clearly by electron microscopy in cross sections of elementary fibers. The fibrils are bonded to each other not only by cross-linking but also by intertwining.

Finally, the fibers, with a wide range of fiber diameters, ramifications, and reticular networks with the widest conceivable variation in weave angle, grow out of the elementary fibers, as can be seen in a traditional view of a histological section provided by optical microscopy (Fig. 9).

Figure 9. Structural hierarchy of type I collagen in hides

Collagen therefore consists of structural elements which form a structural hierarchy. Up to and including the fibrils, this structure constitutes the microstructure of the collagen, and from the elementary fibers upwards, its macrostructure.

The essential structural data are known and can be exploited for quantitative calculation of material transformations on the collagen, e.g., penetration and binding of tanning materials (Table 7).

The structure of collagen described above is in principle the same for all hide and skin species employed in leather manufacture, in particular as far as the fibrillar microstructure and reactivity are concerned. By contrast, clear differences exist in macrostructure, according to hide and skin species, origin, age, etc. The same applies to noncollagenous constituents such as fat. To these are added topographical differences (“position”) in the hide with regard to weave density and angle of the collagen fibers; these differences also have a decisive bearing on the mechanical properties of the leather. The stratigraphical differences between the individual hide layers, grain membrane, papillary, and reticular layer are considerable.

5.3. Role of Water in Collagen

Water is an integral constituent of the collagen structure. At the molecular level of the collagen structure, the water content acts as nothing else than a “cross-linker” that stabilizes the helix structure to a greater or lesser extent, as discussions between different research groups show [39]. It also stabilizes the intra- and interfibrillar structure; further amounts of water are then responsible for the mobility of fibrils at the interfibrillar level. In other words, water acts both as a “tannin” and a “lubricating agent”. Some types of water are bound on collagen with different firmnesses. Finally, the freely moving intercollagenous bulk water is essential as a solvent for the chemical reactions on the collagen. Reduction or even removal of the different water types has far-reaching consequences (Table 8).

Table 8. The different types of water bound to collagen and their characteristics

Type of water	Amount, g/g collagen	Amount in leather,* %	Type of binding	Strength of binding	Effect when completely removed

H_2O type I (a and b)	a) 0.01 – 0.02	1.1	intra- and interhelical	very strongly bound	helix deformation
	b) 0.02 – 0.07	3.9			
H_2O type II (a, b, and c)	a) 0.07 – 0.3	15	intra- and interfibrillar	from a) to c) decreasingly strongly bound	capillary contraction: volume/area shrinkage; isometric: strain
	b) 0.3 – 0.6	0			
	c) 0.6 – 1.6	0			
H_2O type III	1.6 to > 2.0	0	capillary water ("bulk water")	loosely bound in the fibrous structure or even not bound to collagen	densification of the fiber network, pore contraction, lack of solvent for process chemicals

*Containing 65 % collagen and 15 % water.

Consideration of this "system character" (after BIENKIEWICZ [38]) of leather – water is crucial to an understanding of the tanning process and many properties of collagen and leather. Some important question are, e.g,. what happens to the water shell of collagen fibrils due to the reaction of collagen with tannins? Can tannins act as "lubricating agents" instead of water?

It should suffice here to point out that the thermal stability of collagen and leather is a function of its water content. The thermal stability T_S of water-free collagen is approximately 200 °C and is reduced somewhat by tanning. The hydrothermal stability of fully hydrated skin collagen is known to be only T_S = 60 °C and is increased by vegetable and aldehyde tanning by around 20 K and by chrome tanning by up to 50 K. The glass transition temperature T_G of (dry) collagen and leather lies between 50 °C for native collagen and 70 °C for tanned collagen. It characterizes the transition from the glass- to the rubberlike state, is important for the leather properties, and measured by dynamic mechanical thermo-analysis (DMTA) [42].

5.4. Biosynthesis of Collagen

How is collagen formed? Owing to its significance for cellular processes and the existence of numerous collagens with different functions, the biosynthesis of collagen has been a frequent subject of studies in the domain of cellular biology and biochemistry. A brief summary of the synthesis of type 1 collagen should therefore suffice. The genetic code for collagen is anchored at Position A 64 of the human genome. As the sections of the DNA responsible for collagen synthesis are known and sequence analysis of the nucleic bases is simpler than for the proteins, collagen sequences are in fact now decoded in the main via DNA sequences and stored in suitable databases. Synthesis takes place in the connective tissue cells, the fibroblasts, beginning with the single-stranded α-chains. The latter contain propeptides, both at the N- and the C-terminal ends, which give the strands their solubility. Proline and lysine are not hydroxylated initially. Only in the extracellular space are these propeptides cleaved off, a part of the proline (pro) and lysine (lys) is hydroxylated to hyp and hys, some of the hys residues are glycosylated, and the natural cross-linkages are developed and increase as the connective tissue ages and progressively reduce the solubility of collagen and its ability to swell [43]. Formation of the triple helix requires first cross-linking by disulfide bridges between the mobile telopeptides to ensure the right starting position, zipping the α-chains into a triple helix beginning at the C-terminal end of the chains. These processes are all controlled by enzymes (Fig. 10).

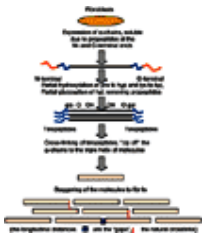

Figure 10. The biosynthesis of type I collagen fibrils (simplified overview)

[Top of Page]

6. Processes Preparing the Hides and Skins for Tanning

Before the actual tannage the hides and skins must be unhaired and defleshed to obtain the corium, the so-called pelt, consisting mainly of water and collagen. Trials have been made on unhairing the whole animal in the slaughterhouse, but for reasons of economics and hygiene, these were not successful till now [44]. Defleshing at the slaughterhouse is practicable but seldom used. So unhairing and defleshing are normally performed in the beamhouse of the tannery. In the case of skins rich of fat, e.g., sheep or pig skin, a chemical and/or in the case of pig skins a mechanical degreasing is necessary [45].

To start this processes a cleaning of the hides and skins is required, the task of soaking. The collagen structure of the pelt must "open up" to ensure the penetration of tannins and other chemicals into its fine structure and to result in soft leathers. This is, besides the dehairing effect, is also the task of liming and bating.

6.1. Soaking

The targets of soaking are manifold: Removal of dirt, dung, blood, and salts, mainly NaCl, in the case of hides preserved by

salting. Also soluble noncollagenous proteins such as albumins and globulins, and proteoglycans must be removed. In all cases complete rehydration of the hide collagen must be ensured. Especially in the case of dried skins it is not so easy to reach this aim. The addition of relatively small amounts of surfactants and proteolytic enzymes (trypsin from pancreas or bacterial proteases) enable shortening of the soaking time. Careful use of enzymes does not attack the collagen [46]. The same is promoted by mechanical movement and therefore the soaking process takes place in drums or very often in mixers. A small addition of bactericides is customary. It is necessary to wash and/or to rinse the soaked hides very intensively, and therefore water consumption is high.

Concerning collagen, soaking is a cleaning and rehydration procedure which will further be developed and completed in the following liming and bating operations.

6.2. Liming, Deliming, Bating, Degreasing

The objectives of these processes are manifold. Liming should achieve dehairing and removal the epidermis and to a certain extent degreasing to give a pure and clean pelt consisting mainly of corium type I collagen and water. This pelt collagen requires “opening up” of the fibril structure to ensure better penetration of tannins and other agents applied later in the leather manufacturing process. The bating process promotes and completes the described effects of liming. Deliming is necessary to remove the liming chemicals, mainly calcium compounds, from the pelt collagen and is normally combined with bating. Intensive washing and rinsing processes are necessary between all these stages.

As mentioned in connection with the histological hide structure, the target of liming is to weaken the basal membrane. An old method is the use of alkali and calcium hydroxide. This has a limited solubility and its saturated solution also a limited pH of around 12. Therefore, too strong swelling or even hydrolysis of collagen can safely be avoided. It is also cheap. Lime alone takes a long time to loosen the epidermis (6 d). Addition of sodium sulfide or sodium hydrogen sulfide drastically accelerates dehairing by reductive cleavage of the cystine bridge of keratin: cystine is reduced to cysteine and the sodium sulfide is oxidized to polysulfide.

Today the standard technology is the application of so-called hair-destroying liming with utilization of relatively large amounts of sulfide. Thus, the keratinous substances are degraded to a slippery mass which can be removed from the grain layer of the pelt by the drum movement alone. This method is quick and complete depilation is ensured. A disadvantage is the high COD load of the wastewater.

In a different technology, a mixture of lime and sulfide is applied to the flesh side only. Hydroxide and sulfide ions diffuse through the hide and arrive at the extremely sensitive basal membrane; thus, hair roots and the epidermis are loosened, but the hair remains completely intact. This so-called “painting” only works with thin skins (goat, sheep, calf) and up to now has been applied ito win undamaged wool (“pulling”).

In another hair-saving process, the keratin of the hair shafts is immunized by mild alkaline teatment. The conversion of cystine to the thioether lanthionine results in high resistance to sulfide, which can attack only the prekeratins at the hair roots. The hair is loosened and can be mechanically separated. Its quality is not so good, but a reduction of the wastewater load is thus achieved.

To avoid the use of sulfide, which can form H_2S under acidic conditions, and the lime load in the wastewater, many alternatives have been developed. A number of organic thio compounds have been used, but for several reasons (e.g., cost, smell) they have not found wider application. In oxidative dehairing with hydrogen peroxide [47, 48], possibly enzyme-aided [49], or sodium perborate [50], cleavage of the —S—S— bridge is caused by oxidation to a sulfonic acid group. However, since this process also has some disadvantages, it has been rather seldom used till now, but there are optimistic opinions about its future chances [51] (Fig. 11).

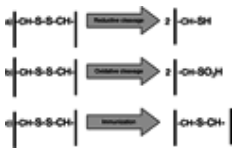

Figure 11. Cleavage of the cystine bridge of keratin during liming. a) Reductive (e.g., by Na_2S); b) Oxidative (e.g. by H_2O_2); c) Lanthionine formation by alkaline immunization

Especially in the last two decades much work has been done on enzymatic dehairing. Even centuries ago a kind of enzymatic unhairing was applied in the form of the so-called sweating process: the hides were simply hung in a warm room of high humidity. Bacteria rapidly grew and their enzymes loosened the epidermis. This process was very risky and is now not often applied, but the idea of using enzymes survived till now and some progress has been made [52]. However, it is nearly impossible to remove the hair or bristles completely, some hairs always remain in the grain, and so often an additional treatment with sulfide or organic thio compounds is necessary [53].

A more recent example of a substitute for lime is the use of water glass and NaOH [54].

It can be assumed, that the search for ecological better dehairing processes will continue as part of a general strategy of sustainable leather manufacturing [51].

The described operations are performed mostly in drums or mixers, possibly equiped with special appliances for hair recovery, e.g., sieves.

After liming, the hides are defleshed. Although the through-feed multiroller machines themselves are highly automated, defleshing is one of the most physically demanding jobs in the tannery (Fig. 12).

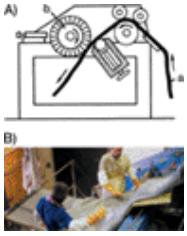

Figure 12. A) Working principle of defleshing machines as an example of multiroller machines in general (see Table 5) a) Hide; b) Knife cylinder B) Defleshing in practice (courtesy of Bader GmbH & Co., Göppingen, Germany)

Splitting machines split the limed hides in two or even more layers. Their operating principle using an endless band knife is used almost exclusively in tanneries (Fig. 13).

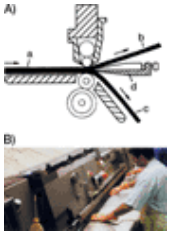

Figure 13. A) Working principle of band knife splitting machines a) Unsplitted hide; b) Grain layer; c) Flesh layer; d) Knife (see Table 5); B) Splitting of pelts ("in the green") in practice (courtesy of Bader GmbH & Co., Göppingen, Germany)

The collagen-related "opening up" effect of liming at first involves certain isolation of the collagen fibrils due to mild alkaline swelling [55].The effect is improved by application of certain enzymes to remove noncollagenic components such as proteoglycans [56]. This is an additional argument for using enzymes, even in liming. In the bating process it has long been the state of art. "Opening up" is therefore always a cleaning operation, too [57].

Alkaline liming also always causes a certain degree of desamidation of glutamine and asparagine to the corresponding amino acids with free carboxylic groups. This improves the later reaction of chrome tannins with collagen (see Section Chrome Tannage).

Bating has been an enzymatic process since its invention. Trypsin from pancreas glands has been used since 1908. Later versatile bacterial enzymes, expressed from strains producing enzymes with different optimal pH values, have increasingly been used in bating. They are produced in huge fermenters from bacterial strains genetically engineered for maximal yield. The object of bating is a further cleaning of the grain layer and opening up of the microstructure of collagen [58].

Deliming finally is necessary to remove the Ca^{2+} ions from the pelt. This is done using mild acids like lactic acid, ammonium salts — nowadays rather seldom for ecological reasons (wastewater load) — (di)carboxylic acids, and CO_2.

If degreasing is necessary, as in the case of sheep skin, the application of organic solvents is banned for ecological reasons, as is the use of alkylphenolethoxylates (APE), especially nonylphenolethoxylates. In the course of scientific investigations concerning the role of surface tension during the degreasing process and using computer aided product design a new degreasing agent with special chemical - physical properties and therefore better degreasing effect was developed [59, 60].

6.3. The Use of Enzymes in Leather Processing and in Collagen Upgrading in General

Dog excrement contains bacterial enzymes and was used by tanners for thousands of years for a kind of "bating". Later tryptic proteases were used, and enzymes are gaining growing importance in leather making and collagen utilization in general [61].

The enzymatic removal of dung from hides using mixtures of proteases, amylases, cellulases, hyaluronidase, and others has been intensively investigated [62, 63]. Proteolytic enzymes from different sources are available not only for the beamhouse processes [63], but also for cleaning wet blue [64]. The addition of enzymes to fatliquors is claimed to give softer leather [65]. By using enzymes it is possible to enhance the tanning properties of vegetable tannins by formation of reactive quinones by oxidoreductases in situ or in vitro [66]. Proteolytic enzymes (trypsin or pepsin) are used to hydrolyze leather wastes to recover valuable hydrolysates [67]. The generation of biogas is also an enzymatic process [68].

Transglutaminase (TGA) is used in the food industry as a cross-linker [69] (Fig. 14).

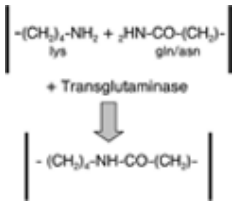

Figure 14. Cross-linking of collagen (proteins) by transglutaminase

TGA can also cross-link collagen hydrolysates and/or gelatins and casein in finishing [70], and the cross-linking of pelt collagen as a kind of tannage has been proposed [71] but also discussed somewhat critically [72]. Cross-linking of casein in

leather finishing by formaldehyde is also possible [73].

Given the above-mentioned developments, further growth of enzyme applications in leather making and collagen upgrading can be foreseen.

[Top of Page]

7. Tanning Processes

Tanning is only one stage in the transformation of collagen into leather, but it remains an important one. For a long time, “tanning” was virtually synonymous with “leather production”, as the tanner and leather maker were one and the same. Vegetable-tanned leathers, which dominated for thousands of years, owed their characteristics to the tanning material due to its presence in high quantities in the leather, for example, in a ratio of almost 1:1 to the hide substance in the case of sole leather. The characteristics of some other heavy vegetable-tanned leathers were determined mainly by the tanning process, too.

The characteristics of glacé leather, a very light and soft leather tanned with alum, salt, egg yolk, and wheat flour were generally regarded as being determined by this tanning process. The same applies to the characteristics of chamois leathers (wash leather), which were simultaneously tanned and softened by fish oils. The decisive role of tanning has also been emphasized for chrome leathers, as chrome tanning endowed the leathers with hitherto unknown characteristics. This form of tannage owed its breakthrough in part to the development of fatliquoring and retannage. In this case the tanning is still important but many other processes are also indispensable (see Chapters Processes After Tanning and Leather Finishing).

What is the current role of tanning in the manufacture of leather? Its first function is now the manufacture of an intermediate which is suitable for partial dehydration and subsequent adjustment of the thickness by shaving and/or splitting. It must be resistent to rapid putrefaction so that it can be stored and transported in the wet state, and can be equipped at a later stage with the desired leather properties in wet finishing. Retannage then completes this primary tanning effect and guarantees, in conjunction with fatliquoring, dyestuffing, and mechanical softening (by staking and milling), the manufacture of serviceable leather. The types of leather most important today, in particular all chrome tanned and free of chrome (FOC) types (generally tanned with glutaraldehyde) are therefore based on a two-stage tanning of this kind. Shoe upper leather, furniture leather, and automotive leather are relevant examples. The division of tanning into two stages simplifies recovery of the byproducts of thickness adjustment of the semifinished products, and allows unconventional pretannages to be used. The recently “rediscovered” sulfur tannage [74] and water glass tannage [54] serve as examples here.

To avoid confusion in the nomenclature the following scheme shown in Figure 15 is used here.

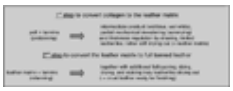

Figure 15. Tanning of light leather with chrome or glutaraldehyde and the like (FOC): a two-step process of pretannage and retannage. Only a combination of both steps truly converts collagen to (crust) leather.

In the following the objectives and results of tanning with respect to the changes in the limed and cleaned collagen are described in a more general way according to Reich [75]. The current theoretical basis of tanning is also briefly treated. Subsequently the most important tannins are explained in detail [76, 77].

7.1. General Objectives and Results of Tannage

Thanks to our knowledge of collagen, it is now possible to define the nature of tanning in the light of the changes brought about on the collagen. Despite all the differences in desired leather characteristics, material structure, and reactivity of the tanning materials employed, the tannages stabilize the collagen structure: tannage prevents the collagen microstructure from collapsing when the water is removed, as is the case when pelts dry to form parchmentlike, hard, and transparent material. This characteristic is accompanied by reduced swelling capacity, enhanced resistance to hydrolytic influences and to microbial and ultimately therefore enzymatic attack, and is often but not necessarily characterized by high hydrothermal stability.

The effects described are permanent in the sense that they do not change significantly even under repeated moistening and drying of the intermediate product thus obtained. A further characteristic is that the tanning materials remain in the leather. Therefore, the following definition of tanning is possible:

Tanning is an industrial process stage during leather manufacture in which the collagen fiber network of the hides and skins is permanently stabilized in its chemistry and structure such that it no longer collapses and adheres to itself when water is withdrawn from it, i.e., it dries out in a leatherlike manner. The intermediate product thus obtained exhibits reduced swelling capacity in the wet condition, is less susceptible to hydrolytic and enzymatic attack, and frequently exhibits also greater hydrothermal stability. All substances which provoke these effects and become part of the leather are tanning materials, irrespective of their chemical structure and the nature of their bond to the collagen.

The changes effected on the collagen by all tanning methods (though to different degrees) are described below in terms of their origin and the methods now available for their demonstration.

Prevention of the collagen structure from collapsing when the water is removed is identical to that what has been described since time immemorial as “leatherlike drying out”. The reason is the retention of air-filled voids between the collagen fibers and fibrils. This characteristic is due to the fact that the collagen fibrils do not adhere to each other during drying. They would be capable of doing so under the influence of the surface tension of the water contributing to their flexibility and the presence of a number of adhesive groups (the meaning of the term “collagen” is, after all, “glue-forming”). They are, however,

prevented from doing so when they are stiffened and/or their reactive groups (those with the potential to adhere/cross-link) are blocked by the tanning materials. This can occur both by cross-linking and by “distance enhancement” and/or mere coating of the surface of the internal structure elements of collagen, all of which involve blocking of the reactive groups (Fig. 16A).

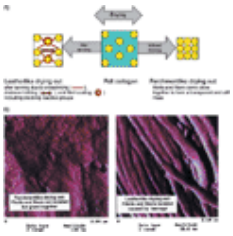

Figure 16. A) Different possibilities of tanning to ensure leatherlike instead of parchmentlike drying out of untanned pelt (schematic); B) The difference between leatherlike and parchmentlike drying out shown by scanning probe (atomic) microscopy (SPM/AFM; courtesy of M. Tuckermann [57])

Conventional microscopic images of dry leather produced by optical, scanning electron or especially scanning probe/atomic force microscopy (SPM/AFM) [78] reveal most impressively the isolation of fibrils and fibers (Fig. 16B).

Leatherlike drying out, which anyone can learn to assess subjectively as soft and warm handle, can also be defined numerically by a number of objective measures. All methods are based on the fact that isolation of the fibers and fibrils and the resulting porosity are measured directly or indirectly by way of the changes in characteristics caused by the porosity, e.g., lower apparent density of leather compared with parchment, and/or are visualized.

Tanning causes a reduced swelling capacity of collagen. Fully hydrated pelt exhibits a water content of over 200% in terms of collagen. The free water, not directly bound to the collagen (bulk water), accounts for the greater part and fills the voids between the collagen fibers and fibrils. The collagen is held together here by covalent cross-linking, electrostatic forces, and hydrogen bonds, which determine the water content and thus the degree of swelling. The natural cross-links increase with the age of the skins [43]: adult collagen thus swells notably less than young collagen. The same applies to the influence of cross-linking tanning: here, too, the degree of swelling and the water content are reduced. This does not become particularly apparent, however, until either the balance between acid and basic groups, which exists at the isoelectric point (IEP) and ensures minimum swelling, is disturbed, or the large number of intracollagenous hydrogen bonds with their stabilizing action is reduced. In both cases, the degree of swelling rises considerably since, due to structural destabilization of the collagen, more space has been created for further uptake of water. The former case applies to acid or alkali swelling; the latter is responsible for hydrotropic, lyotropic swelling. Tanning is capable of reducing both swelling effects.

Tanned collagen has increased hydrolytic resistance. As has already been stated, “collagen” means “glue-forming”, i.e., untanned collagen can be broken down by treatment with hot water and dissolved. This characteristic is eliminated to a greater or lesser degree by tanning, resulting in a complete resistance to hot or even boiling water in the case of chrome leather. The cross-linking action of the tanning materials is responsible for this characteristic. The resistance to dissolution is determined gravimetrically. The result reflects the thermal stability of the tannin – collagen bond. It is lower in the case of hydrogen bond formation than in the case of covalent bonding and highest when both kinds of bonding are combined (see below).

Ever since the effects of tannage were defined, the resistance to putrefaction, that is, the increased resistance to microbial and enzymatic degradation, has been most important. It has seldom been considered, however, that this is a feature of *wet* leather. Leather in this state is however a temporary exception. The improved resistance to microbial (i.e., enzymatic attack) is on the one hand a consequence of the fact that leather, in comparison to hide, contains less soluble substances which could serve as a potential culture medium for microorganisms. On the other hand, cross-linking, as evidenced by the reduced swelling (see above), brings about a delay in the attack, with the effect ultimately of sterically obstructing enzymatic attack. In addition, the tanning materials also act in part as enzyme inhibitors and/or biocides [79].

Owing to the assumed significance of leather's resistance to putrefaction, THOMAS and SEYMOUR-JONES introduced an enzymatic laboratory test as early as 1924. As collagen is resistant to proteases with the exception of collagenases, the rates of degradation of untanned collagen observed with trypsin, pepsin, or papain may be understood only if the test conditions are taken into consideration. These often involve pH or temperature conditions which bring about parallel hydrolytic degradation of the collagen. It is then only logical to conclude that this collagen, already damaged, is also degraded by proteolytic enzymes, i.e., the degradation can at best be delayed by tannage. In other words: resistance to enzyme attack reflects in principle only swelling capacity and hydrolytic resistance. In any case, any person skilled in the art is aware of the susceptibility of (wet) leather (e.g., wet blue) to microbial attack and of the decay of leather in soil.

The increased resistance to microbial attack is a notable characteristic of leather and can be influenced by tanning, but only in the progress of degradation, and not in its ultimate extent. Test methods must be considered critically; tests under practical conditions, geared for example to mold attack on wet blue and its prevention, are of greater relevance. It is regrettable that no reliable data exist for the susceptibility of leather to degradation in standardized composting tests.

The increase in hydrothermal stability, i.e,. the rise in the denaturing temperature of native collagen, has frequently been regarded as *the* characteristic indicating completion of tannage by cross-linking. It is undisputed that the shrinkage temperature T_S, which is based theron, is a quantity of interest in tanning theory. It was first measured by POWARNIN and interpreted in the context of tanning theory as early as 1910. It provides an indication of various types and grades of cross-linking, and its isometric measurement, the principles of which are summarized by TUCKERMANN et al. [80], yields considerable information on the modification which the collagen/leather structure has undergone [81]. Otherwise it provides little indication of the practical value of the leather, as (wet) leather is seldom subjected to thermal load during processing and in use. The thermal load during shaving is also lower than is often assumed. An attainment of as high a T_S value as possible

is therefore also not an objective in the development of new tanning methods, a fact already pointed out very early by GUSTAVSON [82].

Despite this imperative qualification, certain considerations from the perspective of tanning theory are appropriate, particularly since COVINGTON [83-85], RAMASAMI [86], and other scientists in the leather field attempted to formulate a new tanning theory encompassing all tanning processes on the basis of detailed thermodynamic discussions, having in mind the role of water in the collagen – water – tannin matrix, and using new analytical methods like EXAFS [87], computer modeling and model substances as tanning components].

Computer modeling is nowadays of growing importance in leather research, but there are some limitations. Type I collagen itself is too large for modeling as a whole. Therefore, collagen-like polypeptide chains and/or segments of the Smith microfibril with up to 10 – 12 amino acids per chain were used by BROWN et al., who were the first to use computer modeling in leather research [88] and have investigated structural aspects of the N- and C-terminal telopeptides by modeling [89, 90]. The second point concerns the complicated structure of most tannins. Here simplifications are used, as in the case of chrome tannins [91, 92], or model substances, as in the case of phenolic (vegetable) tannins [93, 94]. Fortunately, new paths for synthesis such substances are now available [95]. Up to now not it has not been so easy to perform modeling with consideration of water as integral component of all collagen – tannin reactions. This is mainly due to the required computer capacity and time. Initial attempts were made by FENNEN concerning the chrome – collagen interaction [92] and by SIGGEL and MOLNAR, who investigated the influence of pH changes on collagen swelling. Modeling with 32 triple helices of 46 amino acids per repeated strand in a large “water box” with ca. 215 000 atoms in total is on going at present [96].

Doubtless computer modeling will be an increasingly valuable and necessary tool in fundamental leather research and the product design of tannins.

Possible Interactions on Collagen. The reactivity of the collagen is determined by the functional groups present within it and their accessibility in the various regions of the collagen structure. The bound quantity of collagen is determined primarily by stoichiometric aspects, but is also influenced by available surfaces on the fibrils, elementary fibrils, and fibers, and in the voids between them (capillaries, pores). Another aspect concerns the tannins themselves: They must be able to react with the collagen independent of their special chemical nature in an essentially irreversible fashion (“water-resistant”), and react at least bifunctionally to ensure cross-linking. Their molecular size must also correspond to the distance between the respective reactive groups of the collagen, and their particle size in solution, which is often larger than the molecular size due to aggregation, should allow diffusion through the pores and capillaries (voids) of the collagen into its fine structure, that is, the ability to pass the “diffusion funnel”. This rule is valid for all kinds of penetration of chemicals in leather manufacturing (Fig. 17).

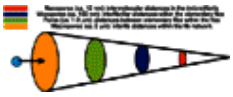

Figure 17. The “diffusion funnel” to be passed by all agents penetrating into the fine structure of collagen

Thanks to current knowledge about the structure and reactivity of collagen and tannins it is possible to summarize the fundamentals of the collagen – tannin interaction (Table 9).

Table 9. The reactivity of collagen with regard to the collagen – tannin interaction

Functionality	Type of bond	Typical tanning materials
Carboxyl groups: COOH	complex bonding	some metal salts, particular basic chromium(III) sulphates
Amino groups: NH_2	covalent bonding	aldehydes, diisocyanates, and other “reactive” (tanning) agents
Ammonium groups: NH_3^+	electrostatic bonding	anionic (tanning) agents, e. g. aromatic syntans
Peptide groups: —CO —NH —	hydrogen bonds	phenolic natural and synthetic tanning materials
Inner surface generally	hydrophobic, and van der Waals bonds	polymers, nonionic tensides and the like with longer alkyl residues, but also all organic tannins to a certain amount
Pores/capillaries	deposited “fillers”	various organic (e.g., some resins) and inorganic agents (e.g., clay, water glasss, sulfur)

The fact that we now know so much about the chemical side of the tannin – collagen interaction is partly due to the fact that a plethora of independent methods exist for demonstrating whether or not interactions have occurred. Their application also permits classification of new or future tanning materials and is a welcome basis for screening.

Despite the high level of our knowledge about tanning theory in general, a lot of questions are still unanswered in detail. When treating the main tannins in the following chapters, some additional remarks will be made in this direction.

7.2. Chrome Tannage

Since the early 1900s chrome tannage has become more and more the dominant tanning process. Today about 85 % of all light leathers are chrome-tanned. The reasons are clear: chrome tannage is a quick, safe, and cheap process, and the wet blue is storable and tradable product which can further be universally processed in many ways to meet special demands for leather properties.

The chrome tannins are nowadays commercially available basic chrome sulfates in which hydroxo (···OH —) and/or ol (—O —) bridges connect polynuclear complexes. The main reactions and terms are shown and explained in Figure 18.

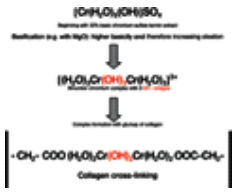

Figure 18. Basification of chrome tannins to form chromium complexes with the aspartic (asp) and/or glutamic acids (glu) of collagen and its cross-linking

In former times tanners produced their own chrome tanning extracts by reduction of chromium(VI) dichromates with thiosulfates, molasses, and/or other reducing chemicals. Today this is done much more effectively by the chemical industry, often using aldehydes as reducing agents. It is also possible to convert Cr^{IV} ions to Cr^{III} ions in situ in the pelt collagen. This was the principle of the first developed and nowadays very seldom applied chrome two-bath tannage.

To avoid a too quick binding of the chromium complexes at the carboxyl groups of collagen they are partially blocked by lowering the pH by the pickle process. The pelt is treated for several hours with dilute sulfuric or formic acid with addition of salt to prevent swelling of collagen. Tanning then begins at lower pH values, later raised by basification, which can be performed with magnesium oxide. Due to the ability of the chromium(III) ion to take up many different ligands into the Cr complex, one can modify the tannin properties ("masking"), e.g., to enhance penetration or to improve chrome exhaustion.

Characteristic of chrome-tanned leather is its high shrinkage temperature; normally it is boilfast (an empirical method to test the success of chrome tanning is comparison of the area of a piece of leather before and after boiling) with T_S > 100 °C.

Only Cr^{III} complexes have tanning properties and only the Cr^{VI} compounds are harmful, even toxic and carcinogenic. Hence, it must be ensured that in the tanning extracts, in the leather, and in the wastes no Cr^{VI} compounds are formed. The absence of Cr^{VI} was taken as absolutely secure knowledge for 100 years, but with the recent increasing accuracy of the analytical methods for leather one could find up to 100 ppm Cr^{VI}. This caused great activities, technological and analytical, to guarantee proper chrome leather quality [97]. Today it can be stated, that in commercial extracts there is no Cr^{VI} and that formation during tanning and wet-end processes can be definitely avoided [98].

In reality this chrome –e collagen interaction (described above) is a simplification. The dominant role of the COOH groups of glutamic and aspartic acids as ligands of bi- or polynuclear chrome complexes is certain [99, 100]. But what is the true nuclearity? What is the specific role of the sulfate ions (there is a distinct different tanning potential between chrome sulfates and chlorides)? What about the other more than 10 complexes of different size and charge detected in chrome tanning solutions, using sophisticated analytical methods [101] and what happens with the "water shell" of collagen (see Section Role of Water in Collagen)?

7.3. Free-of-Chrome (FOC) Tannage

Despite the manifold advantages of chrome tannage FOC tannage is gaining growing importance. One reason for this is that chrome-free solid wastes including worn-out leathers, especially upholstery leathers from automobiles, can be disposed off more easily than chrome tanning wastes, e.g., by incineration. So mainly ecological reasons and a certain pressure by the automobile industry are responsible for the enforced development of FOC leather.

Although there are some different possibilities to produce FOC leathers, the main technology is now a pretannage of the pelt with (modified) glutaraldehyde, mostly in combination with aromatic syntans [102]. After sammying and shaving, the resulting wet whites must be very intensively retanned and fatliquored (see Chap. Processes After Tanning).

Glutaraldehyde has been known as a useful tanning agent since 1960. It is reactive like formaldehyde but at a lower pH, gives sweatproof leather of light yellow color, and reacts very rapidly with collagen. Its vapor pressure is high. Therefore, modified glutaraldehydes, e.g., the acetal produced by reaction with methanol, alkaneglycols, glycerine, saccharides and the like are applied. Such modified glutaraldehydes have some advantages, but the problems with vapor pressure and leather yellowing are not completely solved. To avoid these disadvantages other products are under examination, e.g., reaction products of pyran derivatives with acetone [103].

The reaction between glutaraldehyde and collagen seems clear: cross-linking of the NH_2 groups of two lysine side chains by a covalent reaction with the COH groups of the aldehyde. However, glutaraldehyde reacts very rapidly with itself forming ringlike structures, so the question arises how these react with collagen.

As mentioned, the pretannage must "only" ensure thickness regulation. Recently experiments have been described to achieve this by nearly forgotten technologies using silicates (water glass) [54], sulfur [74], and even iron tannages [104]. Whether these methods will find a broader practical application remains an open question.

7.4. Vegetable Tannins and Tannage

Vegetable tannage is one of the oldest technologies. Many plants, mainly in specific parts like trunk wood, bark, leaves, or

fruits of tropical and subtropical plants contain a sufficient quantity of tannins that commercial exploitation is possible. The tannins are extracted and concentrated to liquid or powdery tannin extracts. They are very often chemically modified by sulfonation for a better solubility or by pH adjustment and the like to ensure better tanning properties [77] (Table 10).

Table 10. Selected information about some important vegetable tannins

Tannin (extract)	Part of plant	Chemical type	Main production region	Remarks
Mimosa (“wattle”)	bark	catechine, condensed	South African states, Brazil	most widely used vegetable tannin, mainly for retanning of light leathers, cultivation in plantations
Quebracho	wood	catechine, condensed	South american states	huge forests, wood winning regulated by “Natural Forestal Laws” also important, especially for tanning heavy leathers
Chestnut	wood	ellagtannin, hydrolyzable	Southern France	large forests, permanent recultivation “allround tannin” for light leather as well as heavy leather
Valonea	acorns	ellagtannin, hydrolyzable	Greece, Turkey	of minor importance, winning in oak forests
Tara	fruits (pods) of *Caesalpina spinosa*	gallotannin, hydrolyzable	Peru	growing importance for FOC leather manufacturing due to its lightfastness but minor amount
Sumach	leaves	gallotannin, hydrolyzable	Italy	of local importance only, gives very bright and lightfast leathers

In practise the tannin extracts of mimosa, quebracho, chestnut, and tara play a dominant role. They are used mainly as retanning and filling material in the wet-end processes in manufacturing of shoe, furniture, automotive, and similar “light” leathers. They are the main tannins in the tannage of sole- and insole leather and similar “heavy” leathers.

All vegetable tannins are pholyphenolic substances [105-107]. That is why they can react with collagen by hydrogen bonding (see Table 9). There are two classes: hydrolyzable and condensed tannins (Fig. 19).

Figure 19. A) Hydrolyzable tannins can be hydrolysed to smaller components, e.g., chebulinic acid to trigalloylglucose and chebulic acid; B) Condensed tannins contain condensation products of smaller units, e.g., dicatechin, the dimer of catechin

As polyphenols these tannins with an molecular weight of about 500 – 800 aggregate in concentrated solutions to colloidal particles with an apparent molecular weight of 10 000 or even more. Such particles can not diffuse into the fine structure of collagen (“diffusion funnel”, see Fig. 17). They remain on its surface and the tanning comes to an end shortly after beginning (“case hardening”). Therefore the “golden rule of tanning” is: Begin with low molecular tannins of small particle size and, when the collagen structure has been stabilized, continue with larger particles. This is a kind of two-step tannage, as mentioned in connection with pretanning.

Concerning the interaction between vegetable tannins and collagen the dominant role of hydrogen bonding is certain, but there are some questions open about the possibility of a partial amount of covalent (quinone) binding: A certain amount of tannin can be extracted from a vegetable-tanned leather without changing the shrinkage temperature T_S. The role of the water shell must be clarified, too. How does water/tannin exchange at the collagen surfaces take place?

7.5. Aromatic Syntans

Originally because of a shortage of vegetable tannins, synthetic tannins or syntans were developed with similar properties to vegetable tannins (exchange tannins/syntans) or with special properties, e.g., the ability to disperse aggregated vegetable tannins or to improve the dying process by shifting charges (auxiliary tannins/syntans) [77]. This basic innovation in 1913 is associated with the name of STIASNY [11].

The basic materials are phenols like phenol itself, naphthol, dihydroxydiphenylsulfone and aromatic substances like naphthalene. These are sulfonated and condensed with formaldehyde, often combined with urea to improve the lightfastness. The order and the extent of these reactions are variable [108]. As a rule, the more phenolic groups and less sulfonic acid groups in the tannin, the better the tanning power (exchange tanning) [109]. The auxiliary tannins are very often condensation products of naphthalenesulfonic acids. Typical exchange tannins are condensation products of unsulfonated

phenols or sulfones and aromatic or phenolic sulfonic acids (Fig. 20).

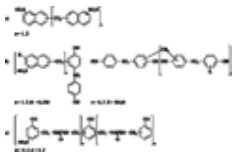

Figure 20. Approximate constitution of some aromatic syntans. a) A typical naphthalene-based auxiliary tannin; b) Two typical sulfone-based exchange tannins; c) A syntan that is lightfast due to the urea bridges

Depending on their content of phenolic and sulfo groups the reaction of aromatic syntans with collagen is mere hydrogen bonding like vegetable tannins or mostly electrostatic bonding between the NH^+ groups of collagen and the SO_3^- groups of the syntans.

7.6. Other Tannages including Modern Combination Tannages

Well known and also still produced to a small extent is chamois leather. It is made with a very old tannage using oils, especially fish oil with its high content of unsaturated double bonds. The fish oil is stacked into the pelt and oxidized in warm chambers with the result of polymerization and acrolein formulation [110]. After washing out the excess oil by alkaline soda solution, giving a useful fat emulsion (degras), the dried and stacked leather is very soft, alkali resistant, and has a high water-uptake ability. So window cleaning is a well known utilization for chamois leather.

Reactive tannins [108] react covalently with functional groups of collagen, mainly with the ∈-amino groups of lysine. This group includes the quinones, well investigated but without practical significance. More important are the aldehydes. The once used formaldehyde has largely vanished from the tanneries for ecotoxicological reasons [111], but glutaraldehyde plays an important role in the manufacturing of chrome-free (FOC) leathers, as mentioned above [102]. Some other chemicals are potential reactive tannins, for example, oxazolidine, trishydroxypropylphosphine (THP) and carbodiimides, but for economical and/or ecological reasons the play no role in the industrial of leather making.

Combination tannages aim to combine the advantages of one tannin with those of another or to compensate the disadvantage of one by using another type of tanning agent. Nowadays the combination of vegetable tannins with aluminum salts is being reinvestigated. In former times leathers of this type were well known as Dongola leathers but nearly forgotten with the invention of the chrome tannage. Such combinations give a T_S of >100 °C, and the individual tannins applied alone only 75 – 85 °C [112, 113]. A synergistic effect is also found for some other modern combination tannages such are vegetable tannage combined with aldehydes [114], oxazolidines [115], amino resins, and/or THP [116]. They are worth some attention from the practical and theoretical point of view, because they make it possible to produce FOC leather with a high, chrome-leather-like hydrothermal stability of T_S > 100 °C).

Especially in the case of mimosa – oxazolidine tanning the theoretical fundamentals are well investigated [115]. Oxazolidine reacts with the NH_2 group of collagen covalently; the phenolic groups of mimosa are bound at the COOH and/or CONH groups of collagen by hydrogen bonds. Additional oxazolidine cross-links the mimosa tannin. This is also the reason why the T_S is much higher (> 100°C) than that of the leather tanned with the single tannins (Fig. 21).

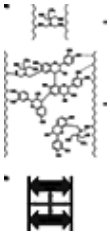

Figure 21. A) The synergistic effect of combination tannages using vegetable tannins and a tannin with double affinity. a) Collagen cross-linked by oxazolidine only; b) Collagen cross-linked by mimosa tannin and mimosa tannin additionally bound to collagen by oxazolidine [115]; B) The synergistic principle in general: A collagen-bound polyphenolic tannin (mainly mimosa) is additionally bound to collagen and/or cross-linked by an second tannin with double functionality (Al, Cr, and other complex-forming metal ions, aldehydes, oxazolidone, THP, etc. [113])

This is exactly what was postulated by Reich 12 years ago and named the synergistic effect of “substances with double affinity” (“doppelt affine Verbindungen”) [113].

[Top of Page]

8. Processes After Tanning

The tanned leather, partially mechanically dewatered by sammying, thickness-adjusted by shaving and possibly splitting, would become relatively stiff and hard on drying, not like untanned pelt which gives parchment, but too stiff for further processing to leather goods. Especially light leathers need additional retanning to partially fill the voids of the collagen to avoid collapse during drying. This filling should be selective, that is, it should cause an adaptation between the looser structure of the fiber network of the flanks and the denser structure of the butt.

Retanning often effects a remarkable softening of the dried leather, but in most cases an additional softening by fat application is necessary, nowadays mainly in the form of fatliquoring, which was originally developed with chrome tannage 100 years ago.

Dying is mostly performed after retanning and before fatliquoring. All three operations are performed in the same vessels, one after another, with washing and/or rinsing of the leather after the respective operation. Preferably used are the Y-

segmented vessels. Sometimes retannings and fatliquors are simultaneously applied in spray-dried form, using short floats (“compact methods”).

Subsequently the leathers are mechanically dewatered, similarly to the described sammying. But mostly it is combined with stretching to remove wrinkles and creases, the so-called setting out operation. Sammying and setting out are often performed by one machine, known as samm/setting operation.

Finally the leather is dried by very different methods and devices.

Even after the chemical lubrication by fatliquoring leathers need mechanical softening by staking, after-drying, and reconditioning.

Some leathers, mainly certain sport shoe uppers, military leathers, and the like, need hydrophobization to ensure the required watertightness. Also in this field several different technologies exist.

8.1. Retanning

As retanning chemicals vegetable tannins and aromatic syntans are often used. Their filling properties are good, but their equalizing property is poor. They react directly with the collagen and not void-specific; they are “overall” fillers. This is different with resins and polymers, which have no or only little affinity to collagen. They penetrate to where the largest space is, and therefore they are “selective” fillers [77].

Resins are condensation products of urea, melamine, and/or dicyandiamide with formaldehyde [117]. Nowadays they play a more minor role than they did 20 or 30 years ago, but a certain amount is still applied.

Now the polymers are much more important; they are mainly products of the polymerization of (meth)acrylates and/or maleic anhydride, which have long been well known in leather finishing (see Chap. Leather Finishing) and are available as a huge number of different products, and an end of the development is not foreseeable [118, 119]. Modern acrylic polymers act not only as filling tannins but also as lubricating substances, which can partially or completely replace fats. This relatively new product class is called lubricating tanning agents (LTA) [120]. It seems clear that from (meth)acrylate-based polymers further progress can be expected (Fig. 22).

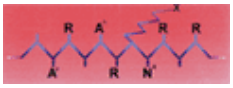

Figure 22. Principal structure of (meth)acrylate- and/or maleic anhydride-based (co)polymers used in retanning, lubrication, and hydrophobization (LTA; according to J.-P. Dix, BASF, Ludwigshafen, Germany)

In general, one can assume that the polymers will gain growing importance and open new paths for a sustainable ecofriendly leather manufacturing. It seems that polymers will be found some day from the beamhouse to the finishing department in all stages of leather production. Patent applications and papers confirm that opinion [121-123].

To choose the right retanning agent in leather manufacturing not only the specific properties, performance, and cost but also ecological aspects must be taken into account [124].

8.2. Dyeing

Leather dying has long been a well-investigated process [125, 126], but newer results are also of theoretical and practical importance [127, 128].

The dyestuffing process is performed mostly using the well known anionic aniline dyestuffs, acid, substantive (direct), 1:2 metal complex dyestuffs, as known from the textile industry (⟶ Azo Dyes – Leather Dyes; ⟶ Metal-Complex Dyes – Leather). Cationic dyestuffs are seldom used; reactives dyestuff do not yet play a role in leather dyestuffing worth mentioning, but this may change in the future [121].

The dyestuffs are available in liquid or powder form, and the desired color recipe can be set up computer aided.

The main dyestuffs in leather dyestuffing are doubtless the classical “Blacks” (Fig. 23).

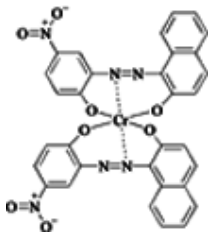

Figure 23. Typical black dye for leather dying (according to M. Schmitt, BASF, Ludwigshafen, Germany)

The penetration of the dyestuffs and the intensity and levelness of dyestuffing depend on the type of leather, especially on the kind of tannage and/or retannage. For example, in the case of the dominant chrome leather, not only reaction with collagen but also with the collagen-bound chromium complexes must be taken into account. Some ionic additives, e.g., some auxiliary tannins, influence the affinity of dyestuffs by charge modification and thus allow the dyestuffing process to be controlled. The influence of pH is also well known; acidification is used to fix dyestuffs on leather.

8.3. Chemical Lubrication (Fatliquoring), Hydrophobization

Fatliquoring is performed essentially for three reasons. First, the fat contributes to leatherlike, soft drying out of the leather matrix, a function which virtually and by definition lies within the area of tanning. A direct relationship does in fact exist between the type and intensity of tannage on the one hand and the need for fatliquoring on the other. Modern chrome-free tanned automotive leathers are a perfect example of how leather having undergone “meager” tanning can be given the necessary fullness and softness by intense fatliquoring. Second, the water uptake of the leather is reduced, and thus the leather becomes more or less hydrophobic. Third, the application of fat is also important for supplying the leather with sufficient tensile and tear strength. Fatliquoring significantly influences the physical properties of leather in general and these effects of fatliquoring have been manifold investigated [129-131].

Fatliquoring is performed with aqueous emulsions of different natural and/or synthetic oils and fats, made water soluble by well-known reactions [129] (Fig. 24).

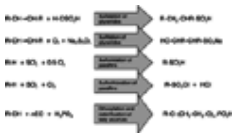

Figure 24. Important reactions to produce different raw stocks for fatliquoring (according to J.-P. Dix, BASF AG, Ludwigshafen)

The old-fashioned application of fat by stuffing, used mainly for heavier vegetable-tanned leathers, is nowadays only seldom practiced.

Leather is principally a hydrophilic substrate. Responsible for this is the porous structure of collagen and its hydrophilic groups. For some purposes, e.g., leather for certain sport shoes or military shoes, permanently waterproof leather is needed, and upholstery or garment leather should be hydrophobic, at least water repellent on the surface. In the latter cases easy-care and soil-release properties are desired [132].

There are some possibilities to ensure the described properties. The simplest is spray coating of finished leather with solutions of fluorine and/or silicone-containing polymers, which can protect the surface for a longer time. This procedure can be repeated later by the consumers themselves. The total water uptake is only minimally reduced by such a spray coating, especially when tested under dynamic conditions.

To ensure durable and effective waterproofing other processes are necessary. One possibility is to nearly completely fill the voids and pores with hydrophobic substances such as certain waxes. But in this case the advantage of leather, that is, breathability and ability to adsorb water vapor is lost (“closed hydrophobization”).

A second possibility is fixation of hydrophobic substances at the inner surfaces of the leather. It is advantageous to use substances which can react as ligands towards the bound chrome tannins inside the leather. Stearates are a well-known example. More recently, highly sophisticated polymers have become available. They are mostly acrylate-based products containing COOH functional groups able to bind to chrome complexes and long alkane chains, modified with silicone and/or fluorine derivatives. Often such polymers are said to be amphiphilic. They make the fibrils and fiber surfaces water repellent without closing the voids between (“open hydrophobization”).

It is easier to make waterproof chrome leather than waterproof FOC leather. This is an only partially solved problem and a challenge for the chemical industry [121, 122].

A third possibility of waterproofing is to introduce smaller amounts of water-swellable substances into the voids and pores of leather. Thus, normally the leather remains breathable and of high water vapor adsorption capacity. Only when liquid water penetrates the substances do swelling and pore closure occur, ensuring waterproofing to a certain degree: A hybrid between “open” and “closed” hydrophobization.

8.4. Drying

Drying is one of the key mechanical operations in the leather making process. Leather acquires its final texture, consistency, flexibility, and the drying conditions determine the area gain. Thus, one can find extended chapters on drying in all text books on leather manufacturing; a good overview of the scientific fundamentals of leather drying is given in [133].

The oldest method of leather drying is hang drying at the “open air”. Old tanneries are typically characterized by their high roofs. Hang drying is still applied today. The leathers hang, fixed by clamps, without any tension on belts passing drying channels with regulated temperature and humidity. Sometimes they even pass through the normal workshops (Fig. 25).

Figure 25. Hang drying in a leather factory (courtesy of Bader GmbH & Co., Göppingen, Germany)

It is a mild drying process in which the leather remains relatively soft but it loses a certain amount of area yield, which is an economical problem. Therefore, toggle drying was developed: The leathers are stretched and fixed in that state on wooden

frames by nails or preferably by clamps on wire frames. Depending on the tension before fixation the firmness and stiffness on the one hand and the area yield on the other can be regulated (Fig. 26).

Figure 26. Toggling drying in a leather factory (courtesy of Bader GmbH & Co., Göppingen, Germany)

A newer method is vacuum drying (Fig. 27) [134]. In special devices the leathers are placed with the grain side on heated steel plates and after closing the apparatus vacuum causes rapid dewatering. The advantages are time savings and a finer, smoother grain, but too intensive drying should be avoided, because otherwise the leather will be too hard and stiff, and will show area retention [135, 136]. Therefore, in many tanneries vacuum drying is used as a kind of predrying, completed by a subsequently hang or toggling drying. Not only the drying conditions, but also the “prehistory” of the leather, e.g., kind and intensity of fatliquoring, influence the leather properties after vacuum drying [137].

Figure 27. Vacuum drier “Cartigliano” in a leather factory (taken from *World Leather* 16 (2003) no. 2, 26)

To avoid losing yield and to ensure a fine grain paste drying was developed some decades ago. The leather was spread out on glass plates with the grain side fixed by special aqueous glue pastes, and the plates were introduced into drying tunnels. Another method was Secotherm drying. In this case the leathers were spread onto the outside of metallic hot water tanks. Both methods have the risk of giving hard leather, and they play only a minor role today, if at all.

Having collagen and the leather matrix in mind, it is obvious that drying is a very difficult process: All leathers have a limited hydrothermal stability, the drying temperature must be kept significantly below the shrinkage temperature T_S of the respective leather.

Water transport from the inner surfaces of the leather to the outer surface needs time; too rapid drying would eventually lead to an overdrying of the outside, followed by fiber condensation and thus to blockage of the diffusion process.

As shown in Section Role of Water in Collagen, the bound water is an integral component of the collagen structure. Water types I and II should not be removed by drying, otherwise the collagen structure collapses. That means as a rule that the water content of leather after drying should not be less than 30 % relative to collagen.

Tension during drying brings area gain but also harder leather: The collagen/leather fiber network is stretched as well as densified, so in the practice of leather drying it is necessary to find a reasonable compromise.

8.5. Mechanical Softening (Staking, Milling)

Even a properly tanned, retanned, chemically lubricated, and dried leather tends to be stiff due densification of the fiber network: the fibers are not mobile enough relative to each other. The task of staking, mostly performed after a conditioning of the dryed leather, is to ensure the mobility of the fibers and therefore soft leather handling.

There are different staking machines [23], e.g., the newer vibration or the older arm staking machines which achieve mechanical softening (Fig. 28).

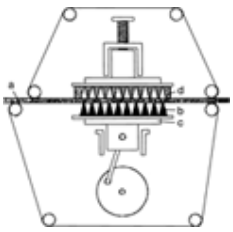

Figure 28. Working principle of a staking machine (modern vibration type)

Again the right water content before staking is very important. If it is too high, the staking seems easier and more effective, but after drying to normal water content during storage the leather will again become hard and stiff. On the other hand, if the leather is too dry before staking the staking effect is insufficient and there is even a risk of damaging the leather grain. Fatliquoring also of influences the staking effect [138].

Milling in drums improves also the softness of the leather and the grain of the leather will be better brought out (mill corn).

[Top of Page]

9. Leather Finishing

The general object of finishing leather is to upgrade the surface of the crusts according to the intended application and especially the grain quality. The leather should look as natural as possible, have the desired color, and should keep its leatherlike warm feel and the other physiological advantages (see Chap. Leather Properties and Testing). The finish should protect the surface against abrasion and other mechanical damages, against spots caused by dirt, water, oil, red wine, and the like [139]. To achieve this target under ecological and economical conditions is a task of the chemical industry [140].

One relatively simple way of upgrading crusts and also split leather is buffing on special buffing machines. The leather fiber network is roughened to fine short single fibers. Buffing the grain side containing the thinner fibers leads to nubuck leather, and buffing the flesh side of grain leather or split leather to suede leather. These velourlike surfaces can be treated by spraying with small quantities of special chemicals to protect the fibers, to improve the layout, and to make the surface water-repellent. This is of course a fine type of leather, but the demand for finished grain leather is much higher and therefore the finishing of leather plays the dominant role.

Finishing grain leather means building a thin layer of natural and/or synthetic polymers, softeners, pigments, dyestuffs, and some other additives on the grain side of the leather.

If the grain is intact and with a natural look, giving the most valuable leather, then only a very thin finish without pigments can be used. Such a leather is called aniline leather. If the grain is slightly defect, the color unequal, and/or the application need special protection of the grain side, a thicker layer is necessary. These layers are pigmented and the leather is called semianiline leather. In both cases the grain is unbuffed. If the grain defects are stronger, the grain side is slightly buffed, followed by a more intensive finishing. Before buffing some grain defects like small holes can be filled with polymer pastes (stucco), a useful and economically interesting upgrading of the grain layer. The result is corrected grain leather.

The difference in intensity of finishing between the above mentioned leather types can be expressed in the number of layers, the quantity applied in grams per square meter, or in the thickness of the finish layers in micrometers (Table 11).

Table 11. The main types of finished leathers and their characteristics

Type of finished leather	Characteristic
Aniline leather	full natural grain, no pigment finish, fully penetrated aniline dying, weight of dry finish layers: 10 – 20 g/m^2, thickness of dry finish layers: ca. 25 µm
Semi-aniline leather	full, light pebbled grain after dry milling, finely dispersed pigments as a light and transparent finish. Possibly slightly buffed grain, lightly pigmented base coat, and transparent top coat, weight of dry finish layers: 20–50 g/m^2, thickness of dry finish layers: ca.100 µm
Corrected grain leather (pigmented finish)	grain slightly or heavily buffed, base coat and top coat often embossed with a grain or fancy pattern, more heavily pigmented, weight of dry finish layers: 50 – 120 g/m^2, thickness of dry finish layers: ca. 150 µm
Heavily finished and/or coated split leather	artificial grain by embossing thick polymer layers, transfer coating or foil application, weight of dry finish layers: ca. 250 g/m^2, thickness of dry finish layers: ca. 0.15 mm
Nubuck leather	grain-side buffed, unfinished, aftertreatment for water and soil repellency, fiber appearance smooth
Suede (velour) leather	flesh-side buffed grain leather or buffed split, fiber appearance rougher

The general structure of the finish, built up by single layers, follows special demands and has its own nomenclature. First there is impregnation of the surface by the so-called impregnation coat. The objective is to guarantee a stable connection between the leather fibers and the polymer layers. These coats are mainly soft, well-penetrating acrylic aqueous polymer dispersions. The leather must be wettable without swelling, and in this context the wet-end processes of leather manufacturing need attention. The next layers are increasingly harder and pigmented, called “base coat”. They are aqueous acrylic and/or butadiene polymer dispersions. Finally the very hard and abrasion resistant top coat is applied. In former times organically dissolved products, e.g., nitrocellulose or polyurethanes, were used. Nowadays the top coats are also aqueous systems, mainly composed of polyurethane, additionally cross-linked to improve hardness and abrasion resistance (Fig. 29).

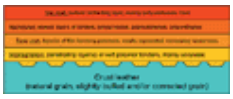

Figure 29. The principle buildup of leather finish layers

As mentioned above the main film-building polymers (“polymeric binders”) are polyacrylates, polybutadienes, and polyurethanes [141]. Polyacrylate dispersions [118] used in leather finishing since about 1935. The polyurethanes [142] started as (organically dissolved) finishing products ca. 1940 in the field of patent leather production and today also have a large share of (aqueous) binders. In the case of aniline leather casein is also used. Nitrocellulose plays a rather marginal role today. Both binders need an external plasticizer, e.g. modified castor oil, phosphate esters etc.

The polymers are to an increasing extent aqueous dispersions, but organic solvents are applied in some cases to ensure special properties. The polymers are synthesized in such a way that an external plasticizer is not necessary ("internal plasticization"). The dispersions are often foamed to reduce water consumption and to ensure thin layers using thickener and therefore safer application.

To ensure proper abrasion resistance, heat stability, and the like most polymer binders need cross-linking. Types of cross-linker include carbodiimides, isocyanates, and modified polyaziridines.

The finish systems are applied mainly by roll coater (base coat) and preferably by spraying (top coat; Fig. 30).

Figure 30. A modern finish unit (spray device "Spraystar" of Gemata S.p.A., Trissino (VI), Italy) in a leather factory (taken from World Leather **16** (2003) no 7, 46)

Other coating methods are available: printing, curtain coating, transfer finishing using release paper, and the like [23]. Between the application of the single layers very often embossing or ironing takes place to improve the connection between the polymers and the leather (impregnation) and between the polymer layers themselves.

Newer developments concern, for example, the application of high-solids systems such as reactive polyurethanes and/or polyacrylate polyurethanes, hardening by UV radiation, and powder coating. [140].

[Top of Page]

10. Ecological Aspects of Leather Manufacturing

Leather manufacturing was not environmentally friendly in the old days. Processing a biological substrate like hides and skins with dog excrement as enzyme source produced a terrible smell and therefore in ancient times the tanners were banished to live outside the village. Today a modern tannery is more ecofriendly, but major efforts are necessary to attain this.

Strict regulations have made intensive endeavours, expensive investment in water treatment, and flexibility necessary to survive as tannery. Indeed many tanneries in Europe and elsewhere closed in the last few decades. Earlier the regulations and laws in some developing countries were not so strict as in the developed countries, but nowadays such differences are more and more eliminated and it is coming to a worldwide harmonized level of ecological demands.

Therefore, "sustainable" leather manufacturing is needed, and in this connection the chemical industry plays an very important role. Very impressive examples for new thinking are given in [143]. For example, it is possible to produce a dry polymer-based FOC intermediate instead of the traditional wet white (LUPA process). Its low water content and weight is economically very advantageous and the best protection against putrefaction. It can be transported and stored without any restriction. It can be quickly and completely resoaked, which is impossible in the case of dryed wet blue (see Section Chrome Tannage) and processed to FOC or chrome leather. The shavings are free of chrome; their utilization causes no problems.

10.1. Ecotoxicology and Occupational Hygiene

Toxic substances are no longer applied in leather making: carcinogenic amines have been banished from dyestuffs and PCP from preservation, the Cr^{VI} risk can be avoided, and formaldehyde is nearly banished, too. The same is with NMP (*N*-methyl-2-pyrrolidon), in former times a component of polyurethane synthesis [144]. NPE (nonylphenol ethoxylate) APE (alkylphenol ethoxylate) have also largely vanished. The technological conditions exclude risks which are theoretical conceivable, e.g., erroneous acidification of sodium sulfide containing floats with release of hydrogen sulfide. There is a very useful document available called "clean leather technology" [145]. Working in a tannery today is a hard job as before, esepcially in the beamhouse, but not a dangerous profession.

10.2. Water and Wastewater

Leather manufacturing needs a lot of water to perform the wet (float) processes and the manifold washing operations. Not only the water consumption but also the cleaning of the wastewater causes high costs. For many years now tanners have therefore paid attention to minimizing water consumption. The success is shown by the following figures: Water consumption to manufacture 1t of bovine hides was more than 200 m^3 in 1950, about 150 m^3 in 1975, and less than 25 m^3 in 2000 (average values for tanneries with high technical level). Washing instead of rinsing, low float processes, and water/float recycling were some ways to reach this aim.

Besides the quantitative aspect the qualitative composition of the wastewater is most important. The wastewater of tanneries, with its high load of inorganic salts and organic matter of chemically very different kinds, sludge, and sediments, requires complicated measures (Table 12).

Table 12. Approximate composition of a typical wastewater (according to [20])

Quantity	Quantity of effluent load/1000 m^2 grain leather and, as byproduct, 430 m^2 split leather, causing 215 m^3 effluent	
	Load in mg/L	Total amount in kg

Suspended solids (SS)	116	25
Chemical oxygen demand (COD)	188	40
Biological oxygen demand (BOD_5)	68	15
Cr^{3+}	5	1
S^{2-}	7	1.5
Ammonium nitrogen (NH_3-N)	5	1
Total Kjeldahl nitrogen (TKN)	15	3
Cl^-	170	36
SO_4^{2-}	81	17

As in other industries, everything has been done to avoid or to minimize wastes. Salt-free preserving, substitution of ammonium salts in deliming, pickle-free chrome tanning, highly exhaustive chrome tannage, chrome recycling, and the like are well known steps in the right direction.

Cleaning the unavoidably highly loaded wastewater needs the combined application of mechanical, chemical, and biological efforts (Fig. 31).

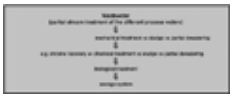

Figure 31. Simplified flow scheme of wastewater cleaning

Membrane processes are gaining in importance in wastewater treatment in the different operations of leather making [146], especially for desalination [147]. Membrane processes are also widely applied for the removal of salts, monomers etc. from syntans, dyestuffs, and the like [124].

Figure 32 shows a modern wastewater treatment plant of a large tannery.

Figure 32. A modern device of tannery wastewater cleaning (courtesy of VDL, Frankfurt/Main, Germany)

10.3. Solid Byproducts and Wastes

As mentioned above only ca. 50 % of the collagen is transformed into leather. The remaining 50 % is found in different solid wastes, for example, the trimmings and fleshings, unusable tanned or untanned splits and shavings, buffing dust, and so on [148]. There are some possibilities of utilization (Table 13), e.g., shavings can be manufactured into leather boards [149] or, after dechroming, converted by very different processes [36] to collagen hydrolysates and reusable chrome tannin. However, large quantities remain unused. Burning leather wastes has recently been a much discussed point, but the Cr^{VI} content of ashes when burning chrome leather must be considered.

Table 13. The main solid byproducts and/or wastes of manufacturing 1000 m^2 of grain leather and their practical and proposed utilization (partially according to [20])

Type of solid waste	Quantity/1000 m^2 "light leather", kg	Realised and/or proposed utilization
Fleshings/trimmings	2840	fat, hydrolysate, biogas, increasing burning
Splits from pelt or tanned intermediates, for leather usable	530 (pelt split), 370 (Cr split)	leather (suede, artificial grain)
Splits from pelt, for leather unusable	655	gelatin, casings, hydrolysates,
Splits from tanned intermediates, for leather unusable	525 (only in the case of splitting in the chrome)	see shavings

Shavings, chrome-tanned	700	hydrolysis: protein and Cr^{III} recovery, leather board
Shavings, FOC-tanned	700	fertilizer
Finished leather cuttings	210	biogas, burning

10.4. Air Pollution

Beside the normal and harmless typical smell of a tannery, the critical air pollution inside and outside the tanneries comes from the finish department, especially from the spraying machines (spray dust). Fortunately, the problem is eased by highly effective absorbers and by the fact that more and more aqueous systems are used for finishing. Buffing dust is filtered out and therefore no longer a danger.

A certain risk lies in the application of aldehydes, but the possibilities to avoid harmful effects have been mentioned above.

Some accidents happened occasionally due to H_2S, but only when a major mistake was made (acidification of beamhouse floats). In the USA since 1971 some 32 people died in such accidents.

Concerning the traditional chrome and FOC leather production there are no principle ecological advantages of the one or the other kind of tannage, except some special chromium aspects as mentioned above. That is not so surprising: The most critical process steps are beamhouse and wet-end operations [150, 151]. A very useful survey of environmental aspects of leather manufacturing is given in [152].

[Top of Page]

11. Leather Properties and Testing

The performance of leather has its origin in the physical and chemical structure of collagen, modified by the leather making process. The structure – properties relationship is thus comparatively well known [153].

Which of the manifold properties of leather are particularly important depends on its intended purpose. For example, for automotive leather there are more than 20 relevant parameters, tested in part by specially developed test methods. Therefore, this modern leather type is in a true sense a “high-tech leather”. Especially in the case of automotive leather, the very different climatic conditions which cars are exposed to worldwide must be taken into account, from arctic to tropical temperature, humidity, and sunshine (UV radiation) intensities. Leather should be heatfast, lightfast, nonshrinking, and the like, independent of these different environmental conditions. Another example is the fogging risk. When leather releases volatile substances (VOC) under the influence of heat then these substances can condense on the windscreen of the car and fog it. Nowadays only fogging-free leather has a chance to be used in cars, test methods are carefully developed, and the threshold for acceptable fogging values is very low [154].

More generally speaking and having collagen in mind, leather properties can be divided into three different categories: aesthetic, physiological, and physical/mechanical.

The esthetic properties comprise the appearance of leather, which should be as natural as possible, that means with a non “polymer-covered”, natural grain or a suede surface of nubuck or velour type. Leather should have a good warm handle, soft feel, and the smell should be leatherlike. It is possible to incorporate perfume-filled nanoparticles into the leather [155].

The physiological properties ensure the wearing comfort of leather goods which are in direct contact with the body: shoes, gloves, garments, and upholstery. Very important in this context is the permeability to water vapor (“breathability”) and especially the ability of leather to take up to about 30 % of its weight in water from the water vapor in the surrounding near-body near atmosphere without feeling wet (“water vapor adsorption ability”). Both properties are uniquely realized in leather and are due to the porous structure of collagen and its hydrophilic groups.

The good insulating behavior against heat and cold is also a welcome characteristic of leather and is also due to its porosity. Naturally, a porous and hydrophilic substrate cannot be waterproof. This would be a disadvantage for the utilization of leather in some cases. Fortunately, there are modern methods of leather impregnation that ensure a high degree of waterproofness without impairing the breathability and the ability to adsorb water vapor (see Section Chemical Lubrication (Fatliquoring), Hydrophobization).

The physical/mechanical properties of leather ensure its durability, resistance to mechanical influences, and thus guarantee a protective function. A very important property of leather is its unique stress – strain behavior. Weak forces easily cause a reversible tension of the leather to a certain point (netlike deformation, Fig. 33), then the resistance against more tension increases dramatically (elastic deformation). These properties are very important for ensuring wearing comfort, e.g., of shoes and gloves. It is a unique advantage of leather that these properties can be influenced by the technical conditions during the leather making process.

Figure 33. The unique stress – behavior of leather: “netlike” deformation of the leather fiber network at low stress, important for wear comfort of shoes and gloves [156].

Test methods for leather are standardized, partially at a national level (e.g., DIN in Germany), at an international level (e.g., EU standards), and/or by the scientific societies of leather chemists. The International Union of Leather Technologists and

Chemists Society (IULTCS) has long been very active in this field since. International commissions of experts are permanently developing and updating test methods for physical (IUP), chemical (IUC), and fastness (IUF) properties. There are almost 100 such methods, often identical with the ISO standards. An overview of leather testing is given in [157].

In the future new test methods will be developed, whereby growing importance will be attached to nondestructive test methods, e.g., based on acoustic emission [158].

Beside the test methods agreements about the tolerable data are necessary. They depend on the special purpose of the leather and have to be agreed upon between the leather manufacturer and the leather processor.

[Top of Page]

12. Leather Substitutes

See also ⟶

Since ancient times other materials have been used beside leather used for the same purpose, e.g., textiles for shoes and garments. Leather was preferred where it had the better properties for the specific article and/or where it is the more valuable material to meet the highest demands. Leather substitution must take into account both aspects. At the beginning of the machine age leather belts for power transmission dominated. Later rubber-coated textile belts came on the market, cheaper and of similar suitability. A real progress from the engineering point of view was brought by the V-belt and more recently by the direct drive. The result is that leather belts today are found only in technical museums. The same happened with other technical leathers once used, e.g., as so-called “textile leathers” in textile machinery.

An other example is sole leather. Apart from wooden soles of special shoes, leather was from the beginning of shoe making the only sole material used. This changed dramatically during the industrial and chemical revolution in the 1800s and 1900s century. Rubber soles showed a longer life than leather soles, and the different polymer-based soles (PP, PUR etc.) were foamed and therefore lighter. From the point of high productivity of industrial shoe manufacturing the direct molding of soles to shoe uppers was a further motive to substitute leather. But even today many elegant shoes are supplied with a leather sole.

The most important reason to develop leather substitutes was to meet the growing demand of materials for garments, upholstery, shoes, etc. Without substitutes it would be quite impossible to safeguard a high level of supply with all the necessary things for daily life. Thus, at the end of the 1800s the time of artificial leather had come. Nowadays the production of the different kinds of artificial leather is more than twice that of leather and, as a rule, there is no strong competition between both materials but a “peaceful coexistence” [159].

From the beginning the target for developing artificial leather was to achieve the optical and haptic properties of leather. Till now the construction principle has remained nearly unchanged: a textile base is coated with a special natural or synthetic polymer and its surface is then subjected to leatherlike finishing, as described in Chapter Leather Finishing (Fig. 34).

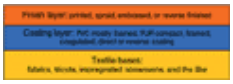

Figure 34. The principle structure of artificial leather

The different kinds of textile bases include fabrics, tricots, and impregnated nonwovens, based on natural and/or synthetic fibers, even microfibers. The polymer layer was originally made with linseed oil, later with organically dissolved nitrocellulose, and then mainly using PVC paste, mostly foamed in different ways, and finally by using organic- or water-based or high-solids (“reactive”) polyurethane. The coating can take place directly, using a doctor knife or (reverse) roll coating, or by transfer coating using release paper or endless steel bands (Sandwick principle). The finishing is similar to leather finishing, but printing dominates; also spraying is practiced. Embossing is possible by machines or, in the case of transfer finish, the structure is given by the (negative) grain structure of the transfer paper (Table 14).

Table 14. Artificial leather: main types, worldwide production, prices

Type of artificial leather	Approximate annual production world wide, 10^6 m^2	Average price, €/m^2
PVC compact artificial leather	1000	1
PVC foamed artificial leather	1500	2
PUR compact artificial leather	500	3
PUR microporous artificial leather, nonwoven base	30	8
PUR microporous artificial leather, fabric/tricot base	150	6
PUR microporous artificial leather without base	5	6
PUR impregnated microfiber artificial leather	40	25

Total	3220	
In comparison: leather	1700	40

Three more recent basic innovations should be mentioned here: The first concerns the production of porous, “breathable” polyurethane layers by the coagulation technique, the so-called “poromerics” (Corfam, Clarino, etc). The special PUR is dissolved in dimethylformamide and the substrate coated. Then the material is introduced into a water vapor saturated channel (precoagulation) and then passed through water baths in a countercurrent principle. The DMF is washed out, and the PUR loses its solubility and precipitates, building a microporous structure. The coagulation technique has been further modified in different ways [160].

The second innovation is the production of microfibers with a very fine titer (high yarn count), similar to collagen fibers in diameter. They are produced by the so-called matrix technology: a mixture of fiber-forming (e.g., polyester or polyamide) and matrix (e.g., polystyrene) polymer components is extruded to an endless fiber. After extraction by organic solvent (e.g., toluene) of the matrix the fibers are split off as microfibers with a diameter not achievable by direct spin extrusion. These fibers are processed to a nonwoven and impregnated by coagulation technology with a very leatherlike aspect. Its surface is of nubuck type (Alcantara) and can be finished similar to leather (Lorica). Products of this kind have been on the market for about 20 years now; they have the most leatherlike look and are also breathable. However, their adsorption capacity for water vapor is only about 10 % of that of leather.

Finally, a class of materials which are not a really artificial leather but is often combined with leather, are the so-called membrane composites. They are waterproof but permeable to water vapor. Combined with textiles in sports wear such membranes ensure comfort by waterproofness and breathability. Such membranes can also be found in sport shoes. A prominent example is Goretex, a PTFE-based film microporized by stretching. Other possibilities to achieve this aim are the incorporation of hydrophilic segments in PUR foil (Helsapor) or the application of a special coagulation technique.

In conclusion it can be said that artificial leather is no threat for leather; on the contrary, it is a welcome addition to meet the demand. Moreover, “real leather” has, beside its manifold advantages, the image of a unique and valuable material, manufactured from a renewable biomaterial. So the slogan is still right: “Leather, an old material with a bright future”!

[Top of Page]

References

1. G. A. Bravo, J. Trupke: *100 000 Jahre Leder*, Birkhäuser Verlag, Basel, Stuttgart 1970.
2. R. D. Higham: “The Tanner's Privilege – A Historical Review of the Tanner's Status in Different Societies”, *J. Soc. Leather Technol. Chem.* **83** (1999) 25.
3. G. Gall: *Leder im europäischen Kunsthandwerk*, Klienkhardt & Biermann, Braunschweig 1965.
4. Food and Agriculture Organisation (FAO): *World Statistical Compendium for Raw Hides and Skins, Leather and Footwear (1994–2002)*, FAO, Rome 2003.
5. J. Buljan: “Gegenwärtige und zukünftige Trends in der Lederindustrie”, *Leder & Häutemarkt* 2003, no. 4, 26.
6. T. Schröer: “Weltweite Entwicklung bei Rinderhäuten und Leder”, *Leder & Häutemarkt/Industrie du Cuir* 2004, no. 10, 35.
7. According to CTC report in *Leder & Häutemarkt/Industrie du Cuir* 2004, no. 10, 29.
8. G. Reich: “Die Nutzung von Kollagen außerhalb der Lederindustrie, Teil I: Allgemeines und Einsatzgebiete mit hoher wirtschaftlicher Bedeutung”, *Leder* **46** (1995) 2.
9. G. Reich: “Die Nutzung von Kollagen außerhalb der Lederindustrie, Teil II: Spezielle Einsatzgebiete und Kollagenhochveredlung”, *Leder* **46** (1995) 18.
10. T. Taeger: “Innovative Chemie für einen alten Rohstoff – Erste Schritte auf dem Weg zur nachhaltigen Lederherstellung”, *Leder & Häutemarkt* 2003, no. 9, 33.
11. G. Reich: “Die technologische Entwicklung der Lederherstellung – Betrachtungen zur Jahrtausendwende”, *Leder & Häutemarkt* 1999, no. 12, 25.
12. M. Bergmann, W. Graßmann: *Handbuch der Gerbereichemie und Lederfabrikation* (**6** Volumes), Springer Verlag, Wien 1932–1960.
13. F. Stather: *Gerbereichemie und Gerbereitechnologie*, 4th ed., Akademie Verlag, Berlin 1967.
14. H. Herfeld (ed.): *Bibliothek des Leders* (**11** Volumes), Umschau Verlag, Frankfurt/Main 1980–1990.
15. K. Bienkiewicz: *Physical Chemistry of Leather Making* Krieger Publishing Co., Malaba Florida (USA) 1983.
16. F. O'Flaherty: *Chemistry & Technology of Leather* (**4** Volumes), Reinhold Publishing Co., New York 1992.
17. T. C. Thorstensen: *Practical Leather Technology*. 4th ed., Krieger Publishing Co., Melbourne (USA) 1992.
18. E. Heidemann: *Fundamentals of Leather Manufacturing*, Roether Verlag, Darmstadt 1994.
19. R. Daniels: *Back to the Basics*, World Trades Publishing, Liverpool 2003.
20. J. Buljan, G. Reich, J. Ludvík: “Mass Balance in Leather Processing” *Leder & Häutemarkt* 1998, no. 4, 30; 1998, no. 5, 25.
21. “Tannery Drums – A Review”, *World Leather* **15** (2002) no. 3, 29. Council of Scientific and Industrial research (IN), WO 04-061136, 2003 (D. Lakshmanan, B. Thangaraj, S. Krishnan, N. Samivelu, P. G. Rao, Th. Ramasami).
22. “Convential vs. Three Component (Y) Vessels”, *World Leather* **15** (2002) no. 3, 33.
23. W. Landmann: *The Machines in the Tannery*, World Trades Publishing, Liverpool 2003.

24. H. Herfeld: “Rationalisierung der Lederherstellung durch Mechanisierung und Automatisierung – Gerbereimaschinen”, in H. Herfeld (ed.): *Bibliothek des Leders*, Vol. **7**, Umschau Verlag, Frankfurt/Main 1981.
25. A. C. Brill: *Gerbereimaschinen*, Roether Verlag, Darmstadt 1960.
26. D. G. Bailey: “Preservation of Hides and Skins” (44th John Arthur Wilson Memorial Lecture), *J. Am. Leather Chem. Assoc.* **98** (2003) 308.
27. W. Meyer: “Histologic Analogies of Bovine Hide, Pig Skin and the Human Integument”, *Proceedings 3rd Freiberg Collagen Symposium* 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK), Paper P I, 1–7.
28. The British Leather Manufacturer Research Association (ed.): *Hide, Skins and Leather under the Microscope*, Milton Park, Surrey 1957.
29. M. Dempsy: *Hide, Skin & Leather Defects: A Guide to their Microscopy*, Leather and Shoe Association (New Zealand), 1984.
30. K. H. Gustavson: *Structure and Reactivity of Collagen*, Academic Press, New York 1956.
31. G. Reich: *Kollagen – Eine Einführung in Ergebnisse, Methoden und Probleme der Kollagenforschung*, Steinkopff Verlag, Dresden 1966.
32. R. Fleischmajer, B. R. Olsen, K. Kühn (eds.): “Structure, Molecular Biology, and Pathology of Collagen”, *Ann. NY Acad. Sci.* **580**, 1990 Links
33. M. E. Nimni (ed.): *Collagen*, **4** Volumes, CRL Press, New York 1988; E. Heidemann: “The Chemistry of Tanning”, in *Collagen*, Vol. **3**, CRL Press, New York 1988, Chap. 2.
34. G. Reich: “Collagen Report: A Review about the Present State”, *Leder* **46** (1995) 192.
35. A. J. Bayley, R. G. Paul: “Collagen: A Not So Simple Protein”, *J. Soc. Leather Technol. Chem.* **82** (1998) 104.
36. J. Cot: “An Imaginary Journey to the Collagen Molecule for a Better Understanding of Leather Waste Treatments”, (45th J. A. Wilson Memorial Lecture), *J. Am. Leather Chem. Assoc.* **99** (2004) 322.
37. M. Van der Rest, R. Garonne: “Collagen Family of Proteins”, *FASEB J.* **5** (1991) 2814; Links R. Mayne, R. G. Brewton: “New Members of the Collagen Super Family”, *Curr. Opin. Cell Biol.* **5** (1993) 883. Links
38. K. J. Bienkiewicz: “Leather – Water: A System?” (31st Wilson Memorial Lecture), *J. Am. Leather Chem. Assoc.* **85** (1990) 305.
39. J. Engel: “Stabilization of the Collagen Triple Helix and Collagen Fibrils by Water: Pros and Contras”, *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper P II, 1–7.
40. K. Kühn, E. Zimmer: “Eigenschaften des Tropokollagenmoleküls und deren Bedeutung für die Fibrillenbildung”, *Z. Naturforsch. B* **16** (1961) 1670; K. Kühn: “Über den Ursprung des Querstreifungsmusters bei Kollagen”, *Leder* **11** (1960) 110.
41. T. Gutsmann: “Atomic Force Investigations into the Structure and Function of Collagen Fibrils”, *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper P III, 1–6.
42. S. Jeyapalina: “The Glass Transition Temperature of Leather and its Implications to the Leather Industry” *World Leather* **17** (2004) no. 5, 47.
43. A. J. Bailey, R. G. Paul: “The Mechanism and Consequence of Maturation and Ageing of Collagen”, *Proc. Indian Acad. Sci. (Chem. Sci.)* **111** (1999) no. 1, 57.
44. W. N. Marmer, R. L. Dudley, A. G. Gehring: “Rapid Oxidation Unhairing with Alkaline Peroxide”, *J. Am. Leather Chem. Assoc.* **98** (2003) 351.
45. A. Zissel: “Arbeiten der Wasserwerkstatt bei der Lederherstellung”, in H. Herfeld (ed.): *Bibliothek des Leders*, Vol. **2**, Umschau Verlag, Frankfurt/Main 1989.
46. T. Feigel: “The Role of Soaking Enzymes on Collagen Destruction in Bovine Hide”, *World Leather* **14** (2001) no. 6, 66.
47. B. Shi, X. Li, D. Sun: “Further Investigations of Oxidative Unhairing using Hydrogen Peroxide”, *J. Am. Leather Chem. Assoc.* **97** (2003) 185.
48. W. N. Marmer, R. L. Dudley, A. G. Gehring: “Rapid Oxidation Unhairing with Alkaline Hydrogen Peroxide” *J. Am. Leather Chem. Assoc.* **98** (2003) 351
49. B. Trommer, “Ein neuer katalytisch enzymatischer Oxidativäscher”, *Leder & Häutemarkt* **2002**, no. 4, 28.
50. W. N. Marmer, R. L. Dudley: “The Use of Oxidative Chemicals for the Removal of Hair from Cattle Hide in the Beamhouse”, *J. Am. Leather Chem. Assoc.* **99** (2004) 386.
51. F. Anders, W.Hopf: “Oxidative Enthaarung noch aktuell?”, *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper A XII, 1–12.
52. A. Crispim, M. Mota: “Unhairing with Enzyme”, *J. Soc. Leather Technol. Chem.* **87** (2003) 198.
53. BASF WO 04 038046, 2003 (H.-G. Lemaire, T. Taeger, G. Pabst, Ph. Lamalle, M. Breuer, B. Kröger, Th. Subkowsky); Novo Nordisk A/S WO 96 19590, 1995 (P. Lars).
54. K. H. Munz, S. Banaszak, G. Manzo, G. Grasso: “water glass in Leather Production”, *World Leather* **17** (2004) no. 5, 35.
55. J. Vaculík: “Die Faserstruktur in den einzelnen Stadien der Lederherstellung und ihr Zusammenhang mit den mechanischen Eigenschaften Teil I: Weiche bis Gerbung”, *Leder & Häutemarkt* 2001, no. 2, 25.
56. K. T. W. Alexander, B. M. Haines, M. P. Walker: “Influence of Proteoglycan Removal on “Opening Up” in Beamhouse”, *J. Am. Leather Chem. Assoc.* **81** (1986) 85.
57. G. Reich: “The Structural Changes of Collagen during the Leather Making Processes” (Atkin Memorial Lecture 1998), *J. Soc. Leather Technol. Chem.* **83** (1999) 63.
58. J. Christner: “Bating: Gaining Area while Retanning Cutting Values” *World Leather* **17** (2004) no. 6, 37.
59. G. Pabst, P. Lamalle, R. Boehn, G. Oetter, R. Erhardt, R. Strey, T. Scottmann, S. Engelskirchen “The Physico-Chemical Principles of Water-Based Degreasing of Animal Skins”, *J. Am. Leather Chem. Assoc* **99** (2004) 151.

60. BASF AG WO 04-063396, 2003 (G. Pabst, P. Lamalle, Th. Füngerlings, G. Oetter. R. Erhardt, G. Scherr, Th. Sottmann, S. Engelskirchen).
61. T. Keshavarz "Enzymes for Environmentally Friendly Processes – An Overview", *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper A VII, 1–9.
62. M. Tozan, A. D. Covington, C. S. Evans: "Studies on the Enzymatic Removal of Dung", *J. Am. Leather Chem. Assoc.* **97** (2002) 167.
63. C. S. Evans, A. D. Covington, F. Foroughi "New Enzymes for Use in the Beamhouse" *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper A VIII, 1–6.
64. L. Rasmussen: "Wet Blue Enzymes – A New Treatment for Area Gain" *World Leather* **15** (2002) no. 1, 44.
65. Lamberti, Spa, EP 1 460 137, 2003 (S. Peratello, M. G. Roaldo, Y. Galante, G. Li Bassi).
66. Novo Nordisk A/S, WO 93 12259 1992 (N. Sorensen, K. Ingvorsen, A. Maalgard).
67. M. M. Taylor, L. F. Cabeza, W. N. Marmer, E. M. Brown: "Use of Tryptic Enzyme Preparations in Treatment of Chrome Shavings" *J. Am. Leather Chem. Assoc.* **95** (2000) 243.
68. W. Bowden: "Gasification –Achieving Zero Waste", *J. Am. Leather Chem. Assoc.* **98** (2003) 20.
69. P. M. Nielsen: "Reactions and Potential Industrial Applications of Transglutaminase" Review of Literature and Patents *Food Biotechnol.* **9** (1995) 119. Links
70. M. M. Taylor, C.-K. Liu, W. N. Marmer. E. M. Brown: "Enzymatic Modification of Hydrolysis Products from Collagen using a Microbial Transglutaminase. III Preparation of Films with Improved Mechanical Properties", *J. Am. Leather Chem. Assoc.* **98** (2003) 435.
71. Fa. Trumpler GmbH & Co., WO 03 18662 1992 (T. Feigel, M. Palissa, R. Fuchsbauer, H.-P. German, M. Meyer).
72. R. J. Collighan, X. Li, J. Parry, M. Griffin, S. Clara: "Transglutaminases as Tanning Agents for Leather Industry", *J. Am. Leather Chem. Assoc.* **99** (2004) 289.
73. Novo Nordisk A/S, WO 94 13839, 1994 (L. Rasmussen, A. Millgard, B. Petersen, N. Sörensen).
74. J. Sagala: "Die Herstellung von Halbfabrikaten (Wet White) durch modifizierte Thiosulfatbehandlung" *Leder & Häutemarkt* **2000**, no. 9, 25.
75. G. Reich: "Die Theorie der Gerbung –gestern, heute morgen", *Leder & Häutemarkt* (2003) no. 1–2, 30.
76. K. H. Gustavson: *The Chemistry of Tanning Processes*, Academic Press, New York 1956.
77. K. Faber: "Gerbmittel, Gerbung, Nachgerbung" in H. Herfeld (ed.): *Bibliothek des Leders*, 3rd Ed., Vol. **3**, Umschau Verlag, Frankfurt/Main 1980.
78. G. Reich, J. Bradt, M. Mertig, W. Pompe, T. Taeger: "Scanning Probe Microscopy: A Useful Tool in Leather Research", *J. Soc. Leather Technol. Chem.* **82** (1998) 11.
79. R. Gayatri, R. Rajaram, T. Ramasami: "Inhibation of Collagenase by Cr(III): Its Relevance to Stabilization of Collagen", *Biochem. Biophys. Acta* **1524** (2000) 228.
80. M. Tuckermann, M. Mertig, W. Pompe, G. Reich: "Stress measurement – Ein neuartiges Verfahren zur Messung von Spannungszuständen im Leder in Abhängigkeit von dessen Struktur und Wassergehalt", *Leder & Häutemarkt* (2000) no. 5, 25.
81. A. D. Covington: "Crosslinking – What Crosslinking? Studies on the General Mechanism of Tanning", *Proceedings XXV. IULTCS-Congress Mexiko 2003* (on CD only)
82. K. H. Gustavson, *loc. Cit* **76** 3
83. A. D. Covington: "New Tannages for the New Millenium", (39th Wilson Memorial Lecture), *J. Am. Leather Chem. Assoc.* **93** (1998) 168.
84. A. D. Covington: "Theory and Mechanism of Tanning: Present Thinking and Future Implications for Industry", (Atkin Memorial Lecture 2000), *J. Soc. Leather Technol. Chem.* **85** (2001) 24.
85. A. D. Covington: "Chrome Tanning: Exploding the Perceived Myths, Preconceptions, and Received Wisdom" *J. Am. Leather Chem. Assoc.* **96** (2001) 467. A. D. Covington, L. Song: "Theory of Tanning: Current Thinking on the General Mechanism of Tanning", *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper P VI, 1–15.
86. T. Ramasami: "Approach Towards a Unified Theory of Tanning: Wilson's Dream" (42nd John Arthur Wilson Memorial Lecture), *J. Am. Leather Chem. Assoc.* **96** (2001) 290.
87. T. Reich, A. Rossberg, C. Hennig, G. Reich: "Characterisation of Chromium Complexes in Chrome Tannins, Leather, and Gelatine using Extended X-ray Absorption Finestructure (EXAFS) Spectroscopy" *J. Am. Leather Chem. Assoc.* **96** (2001) 133; A. D. Covington et al., "Extended X-ray Absorption Fine Structure Studies of the Role of Chromium in Leather Tanning" *Polyhedron* **20** (2001) 461. Links
88. J. M. Chen, S. H. Fearheller, E. Brown "Three-Dimensional Energy-Minimized Model for Calf Skin Type I Collagen", *J. Am. Leather Chem. Assoc.* **86** (1991) 475.
89. P. X. Qi, E. M. Brown: "Molecular Modeling of N-Terminal Telopeptides of Bovine Type I Collagen" *J. Am. Leather Chem. Assoc.* **97** (2002) 235.
90. E. M. Brown: "Potential Interaction of the C-Terminal Telopeptides of Bovine Type I Collagen", *J. Am. Leather Chem. Assoc.* **99** (2004) 376.
91. D. Buttar, R. Docherty, L. Tricker: "The Application of Computational Chemistry to the Study of Collagen Chemistry", *J. Am. Leather Chem. Assoc.* **92** (1997) 185.
92. J. Fennen: "Modelling of Tanning Processes", *J. Soc. Leather Technol. Chem.* **82** (1998) 51.
93. B. Madhan, N. N. Fathima, J. R. Rao, V. Subramanian, B. U. Nair: "Hydration in Matrix Stabilisation of Collagen Fibers: Relevance to Mechanistic Insight into Tanning", *J. Am. Leather Chem. Assoc.* **98** (2003) 263.
94. B. Madhan, V. Subramanian, J. R. Rao, V. Subramanian, B. U. Nair, T. Ramasami: "A Semi-Empirical Quantum Mechanical Modeling Study on the Interaction of Collagen-Like Peptides with Polyphenolic Molecules: An Attempt to

Gain Insights into Vegetable Tanning", *J. Am. Leather Chem. Assoc.* **98** (2003) 273.

95. H. R. Tang, A. D. Covington, R. A. Hancock: "Synthesis and Spectroscopic Characterisation of the Polygalloylesters of Polyols – Models for Gallotannins", *J. Soc. Leather Technol. Chem.* **87** (2003) 179.
96. L. Siggel and F. Molnar: "Molecular Modelling of Collagen Fibers" *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper P IV, 1–2.
97. D. Chaudhure, R. Chakraborti, P. Bhattacharya: "Further Studies on the Official Methods of Analysis for Determination of Hexavalent Chromium in Leather" *J. Am. Leather Chem. Assoc.* **98** (2003) 388.
98. D. Graf: "Formation of Cr(VI) Traces in Chrome Tanned Leathers: Causes, Prevention, and Latest Findings", *J. Am. Leather Chem. Assoc.* **96** (2001) 169.
99. J. R. Rao, B. U. Nair, Th. Ramasami: "Isolation and Characterisation of a Low Affinity Complex in Chrome Tanning Solution", *J. Soc. Leather Technol. Chem.* **81** (1997) 234.
100. J. R. Rao, R. Gayatry, R. Rqajaram, B. U. Nair, Th. Ramasami: "Chromium(III) Hydrolytic Oligomers: Their Relevance to Protein Binding", *Biochem. Biophys. Acta* **1472** (1999) 595.
101. N. T. Slabbert "Chrome(III) Complexes in Chrome(III) Sulfate Solutions", *Proceedings XIV IULTCS-Congress Barcelona*, 1975; comprehensive abstract by E. Heidemann, *Leder* **27** (1976) 24.
102. G. Wolf, M. Breth, J. Carle, G. Igle: "New Developments in Wet White Tanning Technology", *J. Am. Leather Chem. Assoc.* **96** (2001) 111.
103. BASF AG, WO 04067782, 2004 S. Hüffer, G. Scherr O. Reese).
104. M. Kleban: "Iron Tannage: An Alternative Chrome-Free Tanning System" *World Leather* **16** (2003) no. 5, 29.
105. E. Haslam: *Vegetable Tannins Revisited*, Cambridge University Press, Cambridge 1989.
106. Haslam, E. "Vegetable Tannins – Renaissance and Reappraisal", *J. Soc. Leather Technol. Chem.* **72** (1988) 45.
107. G. Reich: "Theorie und Praxis der organischen Gerbstoffe, Teil I: Allgemeine Grundlagen und natürliche organische Gerbstoffe", *Leder* **47** (1996) 74.
108. G. Reich: "Theorie und Praxis der organischen Gerbstoffe, Teil II: Synthetische organische Gerbstoffe", *Leder* **47** (1996) 157.
109. G. Reich: "Über den Zusammenhang zwischen Konstitution und Gerbwirkung aromatischer Syntane", *Leder* **12** (1961) 261.
110. H. Wachsmann: "Chamois Leather – Traditional and Today", *World Leather* **12** (1999) Part I (1) 68, Part II (6) 81; F. Elsinger: "Die Tranautoxidation und ihre Bedeutung für die Lederherstellung III", *Leder* **33** (1982) 125.
111. G. Wolf, S. Hüffer: "Formaldehyde in Leather – A Survey", *J. Am. Leather Chem. Assoc.* **97** (2002) 456.
112. N. T. Slabbert: "Mimosa-Al Tannages – Alternative to Chrome Tanning", *J. Am. Leather Chem. Assoc.* **76** (1981) 231.
113. G. Reich: "Zur Rolle des Aluminiums in chromfreien Gerbsystemen – Versuch einer Systematik", *Leder* **41** (1990) 46.
114. Z. Lu, X. Liao, B. Shi, "The Reaction of Vegetable Tannin-Aldehyde-Collagen: A Further Understanding of Vegetable Tannin-Aldehyde Combination Tannages", *J. Soc. Leather Technol. Chem.* **87** (2003) 173.
115. B. Shi, Y. He, H. Fan, A. D. Covington: "High Stability Organic Tanning using Plant Phenols Part The Mechanism of the Vegetable Tannin-Oxazolidin Tannage", *J. Soc. Leather Technol. Chem.* **83** (1999) 8.
116. A. D'Aquino, G. D. D'Elia, B. Naviglio, M. Seggiani, M. Tomaselli, S. Vitala "Synthetic Organic Tannage Based on Melamine Resin and THPS: Development of a Semi-industrial Scale Proces for High Quality Bovine Upper Leather" *J. Soc. Leather Technol. Chem.* **87** (2003) 189.
117. S. D. Gupta: "Amino Resins for Leather", *J. Soc. Leather Technol. Chem.* **63** (1979) 49.
118. G. Reich: "Polyakrylate in der Leder- und Kunstlederindustrie" *Leder* **46** (1995) 68.
119. B. Magerkurth, G. Wolf: "Recent Developments in Water-Soluble Polymers for the Post-Tanning Processes", *World Leather* **13** (2000) no. 1, 32.
120. A. El'Amman: "High Exhaust Acrylic Chemistry", *J. Am. Leather Chem. Assoc.* **98** (2003) 263.
121. S. Hüffer, T. Taeger: "Dry Pretanned Intermediaries: A Step to Sustainable Leather Manufacturing", *World Leather* **17** (2004) no. 6, 43.
122. BASF AG WO 04-038046, 2003 T. Taeger, G. Pabst, P. Lamalle, S. Hüffer, S. Schroeder).
123. BASF AG, WO 04-22790, 2003 (S. Hüffer, S. Schroeder, E. Wagner, T. Rädler, K. Vill).
124. M. Kleban: "Ecological Aspects of Retanning Agents", *J. Am. Leather Chem. Assoc.* **97** (2002) 8.
125. G. Otto: *Das Färben des Leders*, Roether Verlag, Darmstadt 1962.
126. K. Eitel: "Das Färben von Leder", in H. Herfeld (ed.): *Bibliothek des Leders*, Vol. **5**, Umschau Verlag, Frankfurt/Main 1987.
127. J. Berenguer, M. Dien, J. Roccas: "Formulation for the Properties Structure Relationship Studies on Leather Dystuffs", *J. Am. Leather Chem. Assoc.* **94** (1999) 43.
128. J. Berenguer, M. Dien, A. Navarro: "Comparison of the Properties of Black Dystuffs – New Developments", *J. Am. Leather Chem. Assoc.* **96** (2001) 79.
129. M. Hollstein: "Entfetten, "Fetten und Hydrophobieren bei der Lederherstellung" in H. Herfeld (ed.): *Bibliothek des Leders*, Vol. **4**, Umschau Verlag, Frankfurt/Main 1987.
130. Z. Zhang, H. Dai, J. Du: "The Effect of Fatliquor on the Physical Properties of Resultant Leathers", *J. Soc. Leather Technol. Chem.* **88** (2004) 110.
131. C.-K. Liu, N. P. Latona, G. L. DiMaio: "Physical Properties Studies for Leather Lubricated with Various Types of Fatliquors", *J. Am. Leather Chem. Assoc.* **97** (2002) 406.
132. M. Kleban, K. Markgraf, J. Reiners: "Waterproofing of Leather: State of Art and New Concepts", *J. Am. Leather Chem. Assoc.* **97** (2003) 487.

133. Ch.-K. Liu, N. P. Latona, J. Lee: “Effect of Drying Methods on Chrome Tanned Leather”, *J. Am. Leather Chem. Assoc.* **99** (2004) 205.
134. C.-K. Liu, G. DiMario: “Vacuum Drying Studies for Leather”, *J. Am. Leather Chem. Assoc.* **95** (2000) 102.
135. C.-K. Liu, G. L. DiMaio “Effect of Vacuum Drying Variables on the Mechanical Properties of Leather”, *J. Am. Leather Chem. Assoc.* **96** (2001) 243.
136. C.-K. Liu, N. P. Latona, G. L. DiMaio: “Area Retention Studies for Vacuum-Dried Leather”, *J. Am. Leather Chem. Assoc.* **97** (2002) 381.
137. C.-K. Liu, N. P. Latona, G. L. DiMaio “Effects of Fatliquor on Vacuum Drying of Leather”, *J. Am. Leather Chem. Assoc.* **97** (2002) 284; P. L. Kronick, A. R. Page “Recovery of Properties of Staked Leather on Storage”, *J. Am. Leather Chem. Assoc.* **91** (1996) 39.
138. C.-K. Liu, N. P. Latona, G. L. DiMaio, “Staking and its Interaction with Fatliquoring on Vacuum-Dried Leather”, *J. Am. Leather Chem. Assoc.* **98** (2003) 329.
139. R. Schubert: “Lederzurichtung, Oberflächenbehandlung des Leders” in H. Herfeld (ed.): *Bibliothek des Leders*, Vol. **6**, Umschau Verlag, Frankfurt/Main 1982.
140. E. Günther: “Innovationen und Trends bei der Oberflächenveredlung von Automobilleder – Stand und Tendenzen”, *Leder & Häutemarkt* **2002**, no. 2, 25.
141. W. Wenzel: “Polymeric Binders – Manufacturing Processes and Properties” *J. Am. Leather Chem. Assoc.* **98** (2003) 431.
142. G. Reich: “Polyurethane in der Leder- und Kunstlederindustrie. Teil Was sind, was können Polyurethane?” *Leder* **45** (1994) 100; Teil “Der heutige Stand der Technik und seine voraussichtliche Entwicklung” *Leder* **45** (1994), 156.
143. S. Hüffer, T. Taeger “Sustainable Leather Manufacturing – A Topic with Growing Importance”, *J. Am. Leather Chem. Assoc.* **99** (2004) 424.
144. W. Wenzel: “NMP-Free Polyurethane Products for the Leather Industry: A New Challenge for Finishing” *World Leather* **15** (2003) no. 8, 46.
145. Reference document about “Best Available Technologies in the Leather Industries” according art, 16, 2 IVU directive 96/61 EC (available http://www.umweltdaten.de/nfp-bat/zgerbereien.pdf)
146. S. Baumgarten, T. O. Buer, W. G. Scholz: “Erfahrungen mit Membranverfahren in unterschiedlichen Prozesssufen der Lederherstellung” *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper A IV, 1–21.
147. W. G. Scholz, P. Rougé, A. Bódalo: “Desalination of Mixed Tannery Effluent with Membrane Bioreactor and Reverse Osmosis Treatment”, *Proceedings 3rd Freiberg Collagen Symposium 2004 (Forschungsinstitut für Leder und Kunststoffbahnen, FILK)*, Paper A V, 1–18.
148. G. Reich “Die festen Abfälle und Reststoffe der Lederindustrie – woher, wohin?” *Leder* **44** (1993) 161.
149. A. J. Gish: “Leatherboard, a Practical Use of Tannery Offal” *J. Am. Leather Chem. Assoc.* **95** (2000) 49.
150. G. Reich: *Ökologische Aspekte wichtiger Gerbverfahren*, Forschungsgemeinschaft Leder e.V., Frankfurt/M. 2000.
151. G. Wolf, R. Wittlinger “Ökoeffizienz-Analyse zur Ledergerbung: Vergleich von Gerbverfahren mit und ohne Chromsalzen”, *Leder & Häutemarkt* 2002, no. 5, 31.
152. R. Daniel: *Back to the Basics: The Environment*, World Trades Publishing, Liverpool 2004.
153. G. Reich: “Die Eigenschaften und das Leistungsvermögen von Leder – Versuch einer Systematik auf der Grundlage von Struktur-Eigenschaftsbeziehungen”, *Leder & Häutemarkt* 2001, no. 5, 25.
154. M. Breitsamer, O. Gotz: “Emission from Automotive Leather – State of Art and a Critical Foresight”, *J. Am. Leather Chem. Assoc.* **99** (2004) 416
155. M. Kleban: “Leather with an Aroma”, *World Leather* **15** (2002) no. 2, 69; Bayer, DE 101 17 671 A1, 2000 (M. Kleban, J. Weisser, F. Koch, W. Schwaiger); Conziaricerca, WO 02-077295, 2001 (D. Zini).
156. W. Graßmann: “Physikalische Ledereigenschaften in Abhängigkeit von Hautstruktur und Gerbung”, *Leder* **1** (1950) 57.
157. M. K. Leafe: *Leather Technology Book*, Society of Leather Technologists and Chemists, London 1999.
158. C.-K. Liu, N. P. Latona: “Nondestructive Evaluation of Leather Based on Acoustic Emission”, *J. Am. Leather Chem. Assoc.* **99** (2004) 251.
159. G. Reich: “Leder und synthetische Austauschmaterialien heute – eine vergleichende Betrachtung”, *Leder & Häutemarkt* **1991**, no. 2, 1; no. 5, 6.
160. H. Träubel: *New Materials Permeable to Water Vapor*, Springer Verlag, Berlin, Heidelberg, New York 1999.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Leather Imitates

Claus Zürbig, Benecke-Kaliko AG, Hannover, Federal Republic of Germany

Hans-Hinrich Kruse, Benecke-Kaliko AG, Hannover, Federal Republic of Germany

1. Definition and History 1
2. Properties 1
3. Raw Materials 2
3.1. Carrier Materials 2
3.2. Coating Materials 3
3.3. Lacquers for Finishing 5
4. Production 6
4.1. Layer Composition 6
4.2. Coating 6
4.2.1. Coating Devices 7
4.2.2. Coating Processes 7
4.2.3. Finishing 9
5. Quality Specifications and Analysis 11
6. Economic Aspects and Uses 11
7. References 12

1. Definition and History

Leather imitates (leatherlike materials) are flexible sheets that have a surface structure similar to that of leather. They are usually composed of a polymeric coating and a carrier layer. Coated leather imitates with a textile carrier can be divided into three categories depending on the type of textile used: woven fabrics, knits, or nonwovens. Film leather imitates are compact or foamed plastic films with a leatherlike surface [7].

From the mid-1700s, leather imitates were produced in England and in Germany by coating textiles with air-drying boiled linseed oils. The slow-drying boiled oils were replaced by nitrocellulose coatings at the beginning of the twentieth century. These leather imitates were used for the interior décor of cars [8]. It was not until about 1935 that poly(vinyl chloride) (PVC) was introduced as a coating material; it is still used as a pasty mixture of PVC powder and a liquid plasticizer, which is spread on a textile carrier and hardened by heating. Foaming agents have been used since the 1950s to produce thick, soft coatings that have permitted the use of leather imitates in new areas. Even today, most leather imitates are coated with PVC.

Polyurethanes and other polymers for coating were introduced around 1960. Polyurethane products were mainly produced by using solvent-based coating systems with the coagulation method (see Section 4.2.2). Increasing demands on product properties resulted in the development of solvent-deficient, end-capped, high-solids polyurethane systems in the late 1970s [9], which can be applied by transfer coating.

In future, development will be controled by technical performance demands and environmental aspects and will lead to tailor-made products for each specific application. These factors already played an important role in the development of reactive high-solids polyurethane systems [10], [11] and thermoplastic polyolefins (TPO) as base materials for leather imitates.

2. Properties

A leather imitate, irrespective of its application, should have the appearance and feel of leather. This includes the grain, a dry feel, and a certain degree of softness. Chemical and physical properties are not uniform; they are determined by the structure and nature of the materials used, and depend on the intended application of the leather imitate. Important properties are resistance to abrasion, cleansing agents, common chemicals, light, and heat, as well as a certain extensibility. The type of carrier material selected is determined by the mechanical demands made on the leather imitate; depending on the application, it must also be rotproof, insensitive to moisture, or heat resistant. Leather imitates must be sewn, glued, welded, and thermally shaped during processing. Imitates used for apparel and footwear must simulate the wearing quality of leather; they should be permeable to water vapor and be

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_283

able to absorb moisture; their flexibility is often better than that of leather.

Leather consists of collagen fibrils that become increasingly compressed in the direction of the surface (→ Leather, Chap. 2.). Grained leather imitates, however, are usually multilayered (Fig. 1) and consist of at least one carrier material and one covering layer. The carrier is usually a textile or a plastic film which gives the leather imitate a certain firmness. To meet the specific demands, different constructions may be chosen.

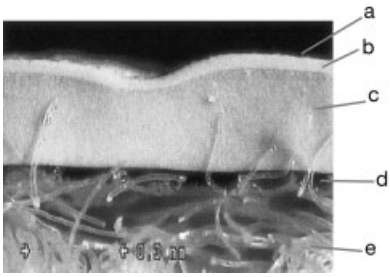

Figure 1. Construction of leather imitates
a) Lacquer; b) Dense top coat; c) Intermediate layer; d) Laminating adhesive; e) Textile carrier

Optical appearance, as well as the feel and touch of the composite, are defined by the transparent lacquer used as finish. By using different additives, these properties are adjusted to specific applications. Formulations with good resistance to light, scratching, and abrasion are used as top coats, since the physical, chemical, and mechanical properties of the surface are determined by this layer. Soft imitation leather contains a layer of foam below the covering layer, but this intermediate layer may also be dense. To avoid impression of the textile structure from the carrier at the surface, foamed constructions are used. By using suitable laminating adhesives, a strong compound of coating and carrier is achieved.

3. Raw Materials

A carrier-free foam, i.e., a *film leather imitate*, is made up of a covering layer and a foam layer. *Thin compact imitation leather* consists solely of a polymer film with a leatherlike grain.

Suede imitates are either composed of several layers, the covering layer being napped or provided with short fibers (Benella), or consist of a uniform nonwoven fabric of bundled microfibers (Alcantara or Lorica) [12]. Microfiber nonwovens simulate the appearance and feel of leather and have very similar wearing qualities.

Leather fiber materials consist of finely ground leather fibers and a binding agent; the leather fibers may also be mixed with textile or cellulose fibers.

3.1. Carrier Materials

The mechanical properties of a leather imitate, such as extensibility, tear resistance, and tear propagation, are mainly determined by the carrier. Woven fabrics, knits, or nonwovens can be used as carrier depending on the demands made on the finished product. These textiles consist of yarn or fibers made of cotton, other cellulose materials, polyamide, polyester, polypropylene, or mixtures of these substances. The yarn can be preshrunk or textured.

Woven fabrics are usually coated by the direct coating method (see Section 4.2.2) to produce leather imitates that are tear resistant and not very extensible. The types of weave commonly used are linen, twill, and satin; weights vary from 100 to 500 g/m^2. The cut edges of leather imitates made from woven fabrics tend to fray. The coating material penetrates into the fabric and the resistance to tear propagation of the leather imitate is poorer than that of the uncoated fabric. Fabric leather imitates are employed, for example, in cabriolet roofs.

Knits are used as carrier when softness and extensibility are required of the imitation leather. Warp or weft knits can be made in different ways; tricot and interlock knits are commonly used. Weights vary from 60 to 300 g/m^2. Undesirable penetration of the coating material into the knit can be avoided by using the transfer coating technique (see Section 4.2.2). Knitted carriers are very extensible because they consist of coiled threads; they are less elastic in the longitudinal direction than in the transverse direction. Projecting edges of the carrier easily roll into the coating mass during coating; this can be prevented by using knits with glued edges. Knitted leather imitates find wide application in upholstery and car seats.

Nonwovens. Many different leather imitates can be produced from nonwovens because of the large number of ways of combining fibers, fiber laying, and fixing methods (→ Nonwoven Fabrics). These products have leatherlike mechanical properties, such as low resiliency on extension. The binder content of splitting felt on a weight basis can be up to three times the fiber content; coating is performed by the direct method. Since needle-punch nonwovens with up to approximately 20 wt % binder readily absorb the coating material during direct coating, they are produced by the transfer coating technique. In the coating of nonwovens with PVC, attention must be paid to interaction between the binder and the coating, e.g., migration of plasticizer from the coating into the binder used in the nonwoven.

Microfiber nonwovens are now being used as new textile base materials for high-class leather imitates. Microfibers are fibers with a maximum linear density of 1 dtex; ultrafine or ultrasuperfine microfibers have values of around 0.09 – 0.13 dtex [13]. Since such fine fibers can not be spun and processed, special conjugated fibers and islands-in-a-sea fibers (i.e., the microfibers are extruded in removable, soluble matrix) with finenesses of greater than 1 dtex are used. After processing they are split into ultrafine fibers by treatment with solvents, alkali, or high-pressure water jets. The majority of conjugate fibers splittable by hydro-de-entanglement are produced in the Far East as staple fibers for processing on carding machines. Finishing (e.g., coagulation) and coloring leads to microfiber nonwovens with mechanical, optical, and feel and touch properties comparable to those of leather.

The weights of nonwovens used for coating lie between 30 and 600 g/m^2. The leather imitates are mainly used in the shoe industry.

Other carriers. Films made of acrylonitrile – butadiene – styrene copolymer (ABS), unplasticized PVC, polyethylene, and polypropylene are used mainly for the interior décor of cars. Similar to fabric carriers, films can also be covered with a coating or provided with a leather grain and possibly a foam layer on the reverse side. Small amounts of cardboard, paper, mats, and netting are also coated for certain applications.

3.2. Coating Materials

Poly(Vinyl Chloride). The coating material most often used for leather imitates is a plasticized (flexible) poly(vinyl chloride) [*9002-89-5*] [→ Poly(Vinyl Chloride)]. A thick viscous paste is produced by mixing 100 parts of PVC powder with 30 to 120 parts of a liquid plasticizer. This vinyl plastisol consists of a stable dispersion of PVC in the plasticizer. It is first coated onto the carrier and then fused by heating and gives a thermoplastic material. During this process, the plasticizer migrates into the swelling PVC particles, and most of the particle boundaries disappear. The plasticizer in plasticized PVC is relatively mobile because it is only bound by physical interactions. It can exude to the surface of the leather imitate as a sticky film, especially after heating, or it can be extracted from the coating by fats and plastics, such as polystyrene or certain polyurethanes. The plasticizers are generally esters of phthalic acid, aliphatic dicarboxylic acids, or fatty acids [14]. They influence the processing conditions and the properties of the final product. The migration tendency and the volatility of the plasticizer decrease with increasing length and branching of the hydrocarbon chain derived from the alcohol. However, the viscosity of the paste and the fusing temperature simultaneously increase, and the PVC paste becomes less workable. Straight-chain esters improve the low-temperature flexibility of the coating. Polymer plasticizers (oligoesters) have a very low migration tendency. To give the coating the desired properties, several plasticizers are often combined.

The amount of phthalate plasticizers emitted by evaporation from soft PVC products into interiors is relatively low (< 0.1 %) [15]. Larger amounts may be extracted by direct contact with materials which are miscible with the plasticizer (e.g., polyether-based polyurethane foams). In risk assessments for phthalate plasticizers, no toxicological, mutagenic, or environmental effects have been found [16–18]. Nevertheless, public discussion led to the development of so-called green PVC products containing nonaromatic plastisizers (e.g., esters of citric acid [19]).

The addition of metallic soaps, epoxides, organic phosphites, or organic sulfur – tin compounds confers stability to light and heat on plasticized PVC. Cadmium and lead salts have

now been largely replaced by barium, calcium, or zinc compounds for environmental reasons. Other additives are pigments, fillers (e.g., $CaCO_3$), flameproofing agents, and, if necessary, foaming agents, lubricants, and biocides.

The production of leather imitates with PVC pastes is economical and relatively simple. Coated surfaces can be heat-embossed with various grains. In the course of further processing, the products can be subjected to high-frequency, ultrasonic, or thermal welding and can also be glued. The leather imitates are mechanically insensitive, stable to alcohol, acids, and bases and, with suitable stabilization, very weatherproof.

Disadvantages of PVC coatings are the strong dependency of the modulus of elasticity on temperature and plasticizer mobility. The leather imitates must be lacquered to make sure that they feel dry and leatherlike. They can become brittle as a result of plasticizer migration.

Polyurethanes are the second most important class of coating materials. They do not require a plasticizer, and even small amounts produce tenacious, abrasion-resistant surfaces. Leather imitates constructed with polyurethane are mostly used for high-quality products. Polyurethanes for coating are made from aromatic or aliphatic diisocyanates. The polyol components also determine the polyurethane quality and are predominantly polyesters; polycarbonates are, however, sometimes employed (→ Polyurethanes).

One-component systems consist of solvent-based linear polyurethanes (solids content 20–30 %) or aqueous dispersions [20] (solids content 20–40 %). After spreading, film formation is carried out by evaporating the organic solvent or water in a drying tunnel. Polyurethane solutions used in the coagulation method (see page 9) contain a water-miscible solvent, frequently DMF. Addition of slow-reacting polyfunctional cross-linkers (e.g., aliphatic polyisocyanates) leads to coatings with increased stability against solvents or cleaning agents. Due to their low solids content one-component polyurethane systems are used for the preparation of thin films (20–100 μm) such as lacquers, top coats, and laminating adhesives [21].

Solvent-deficient, high-solids or solvent-free prepolymer systems are being increasingly employed for environmental reasons. The addition of a hardening agent, usually an isocyanate, causes cross-linking of the coating on drying and the loss of thermoplastic properties. The reactive groups of the hardener are often reversibly blocked in the process and can be thermally activated. Hence, the shelf-life of the coating material is controllable, and coatings derived from aqueous dispersions can be cross-linked above 100 °C after evaporation of the water. In the coating process the isocyanate groups are reformed by removal of the protecting group [e.g., 2-butanone oxime (**1**)] at elevated temperatures (Fig. 2) [22], [9].

The isocyanate end group then reacts with the chain extender (e.g., the cycloaliphatic diamine **2**). This leads to an increase in molecular mass and formation of the polyurethane film. By using cycloaliphatic diamines as chain extenders, poly(urea urethanes) with excellent physical and chemical properties are formed. The rate-determing deblocking reaction ($k_1 \ll k_2$) can not be catalyzed. Therefore, the rate of formation can only be increased by using higher temperatures. Residues of the protecting group remain in the product and result in high values in TVOC (total volatile organic compounds) measurements and a characteristic smell of the final product.

High-solid polyurethane formulations which are stable at room temperature without using reversible protection of the isocyanate groups have been developed in the mid 1970s [10], [11]. Handling is more demanding since contaminants (e.g., traces of water) strongly influence the pot life and reactivity of these systems. Coatings based on these raw materials contain very small amounts of volatile organic compounds, as is demanded for various applications. High-solid polyurethane systems are used to produce thick layers (0.2–0.6 mm) such as the intermediate coating.

Stabilizers [e.g., phenolic antioxidants or hindered amine light stabilizers (HALS)] increase the thermal and light stability of polyurethane coatings. Other additives are pigments, fillers, and flame-proofing and foaming agents.

Polyurethanes have the major advantage that they yield flexible, exceptionally tenacious, solvent-resistant coatings without a plasticizer. They are, however, sensitive to strong acids and bases. When polyurethanes are employed as coating materials, the raw materials used

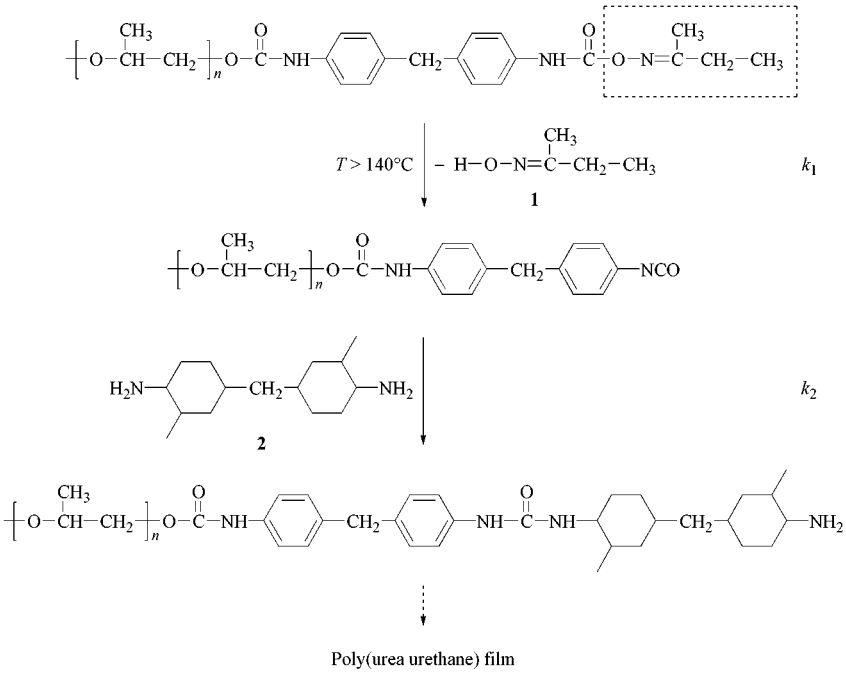

Figure 2. Chain extension of polyurethane prepolymers protected by 2-butanone oxime (**1**)

for their synthesis must be selected to comply with the requirements demanded of the coatings. Polyurethanes from aromatic diisocyanates are light-resistant and less expensive but display poor lightfastness. The more expensive aliphatic diisocyanates yield polyurethanes with a higher lightfastness. If the polyol component is a polyester, the polyurethane may be sensitive to hydrolysis. Its mechanical stability deteriorates under the influence of humidity and heat. Polycarbonates yield polyurethanes that are more resistant to hydrolysis. Modern polyurethane formulations combine good hydrolytic stablility with lightfastness, light and thermal resistance.

Other Raw Materials. Other Raw Materials. Ethylene – vinyl acetate copolymers and butadiene copolymers are also processed in solution or as dispersions. Polyamide coatings usually contain a plasticizer and are produced from alcoholic solutions. They are insufficiently stable to water and alcohol to be used in most applications. Unplasticized PVC, ABS, and thermoplastic polyolefins (TPO) can be employed for hot-melt coating or are first converted into films and then laminated on a carrier (e.g., sponge leather). TPO-based decorative foils are predominantly made by thermal lamination of extruded films on compact films or polyolefin foams as carrier. The TPO leather imitates are derived from textile carriers (e.g., nonwovens). To meet the technical demands on TPO foils, blends of a thermoplastic polyolefin matrix (e.g., polypropylene) and reinforcing elastomeric rubber components (e.g., EPDM) are used as raw materials.

3.3. Lacquers for Finishing

The surface of leather imitates, especially those made of plasticized PVC or TPO, must be lacquered to give them the feel of leather. Pigmented lacquers can be used to produce a leatherlike design; matt varnishes give the desired finish and dry feel. The lacquer also protects the surface: it improves the slip, acts as a plasticizer barrier, and reduces abrasion and soiling. Resistance to light and heat can also be increased.

The finishing lacquers are generally solutions of polyacrylate and PVC in organic solvents.

The lacquer film may be 3 – 20 µm thick. Plasticized PVC coatings can be lacquered with brittle acrylates because the lacquer film absorbs plasticizer from the PVC layer and becomes more flexible. Acrylate lacquers are very resistant to light and temperature, but cold forming of the leather imitate can cause whitening due to cracking and partial peeling of the lacquer film.

Polyurethane lacquers are generally applied as two-component systems. Depending on the degree of cross-linking of the polyurethane, the lacquer film either remains flexible or becomes hard, scratch-proof, and solvent resistant. The resistance of polyurethane lacquers to light and temperature is not as high as that of polyacrylate lacquers. Lacquers based on polyamide are stable to plasticizers, but are not sufficiently resistant to alcohol and water.

The solids content of solvent-containing lacquers usually does not exceed 15 %. Large amounts of highly volatile solvents are liberated on drying; contamination of the atmosphere should be prevented. Part of the organic solvent penetrates into the coating during lacquer application; this interferes with further processing and results in higher amounts of volatile organic compounds (VOC) in the end product. For these reasons, efforts have been made to develop aqueous lacquers for leather imitates. Appropriate dispersions consist of polyurethane or polyurethane – polyacrylate mixtures along with low amounts (< 10 %) of high boiling cosolvents (e.g., *N*-methylpyrrolidone). In new product developments, normally only aqueous lacquer systems are considered.

4. Production

4.1. Layer Composition

A thin primer layer is often applied to synthetic fabric carriers. Carriers can be bound to the coating with the help of a laminating coat of, for example, plasticized PVC. The overlying layers can vary in material, density, plasticizer content, stabilization, and filler content. When coating with plasticized PVC, a surface layer containing less plasticizer and more filler gives a dry, resistant surface. The underlying layer then contains more plasticizer to improve strength at low temperatures. A base layer of plasticized PVC can also be combined with a thin covering layer of a suitable polyurethane for the same reason.

Foamed leather imitates are produced in large quantities and contain a foam layer between the covering layer and the carrier. This foam layer may be 0.2 to 3 mm thick depending on the application; it is produced either as whipped foam before application to the carrier or by adding a chemical blowing agent which reacts during gelling of the coating (→ Foamed Plastics, Chap. 2.).

Foamed leather imitates without a carrier are mainly made of plasticized PVC and are produced in similar quantities to foamed leather imitates with a textile carrier. Special mechanical properties are not required because the films are subsequently laminated onto solid carriers.

Film leather imitates from calendered films are made up of one or more layers. In multilayer imitates, two or three thin films are laminated using heat and pressure to give a single film. The product can be embossed at the same time with a leatherlike grain. Films are often glued on the reverse side and are consequently provided with an adhesive primer. The application of a foam layer made of plasticized PVC, polyurethane or, less often, of polyethylene or polypropylene on the reverse side confers a certain softness to the film leather imitate.

4.2. Coating

Knife coating is the traditional and still most important method of coating. The coating material must be liquid or pasty. It is applied to the carrier and the carrier is then pulled under a knife or doctor blade which is placed at right angles to the direction of travel. The coating is subsequently hardened under heating radiators or in a drying tunnel.

An important processing parameter of the liquid coating materials is their viscosity. Most coating machines require viscous pastes to prevent the coating from flowing off the carrier during application or from penetrating too deep into the textile carrier. Therefore, solutions and thin dispersions are thickened with an appropriate agent. The paste is subjected to high shearing forces during application, and the viscosity should not increase too strongly as a result of

the dilatant flow behavior of the paste at a high shearing rate gradient [23].

Poly(vinyl chloride) pastes are made in a mixer by stirring the PVC powder and the solid additives into the liquid plasticizer. Reduced pressure facilitates the wetting of the PVC particles. Pigment additives must be distributed very uniformly; consequently, it is useful to disperse them in a plasticizer and mill them in a roll mill before addition to the paste. The pastes should be deaerated before application by stirring under vacuum to avoid air bubbles in the coat.

4.2.1. Coating Devices

The paste is applied with either a doctor (knife) or a roll. Different types of *doctor knives* can be employed (Fig. 3). Pointed doctor knives are used for coatings that are not penetrable and for small quantities. Doctors with wide profiles can be employed to produce coatings that are several millimeters thick. The wide nose with a sharp edge is designed to prevent formation of drops that fall into the wet coating and produce defects; such defects often occur in pastes exhibiting dilatant flow. The type of doctor coater chosen depends on the quantity of coating applied and on the type of carrier used.

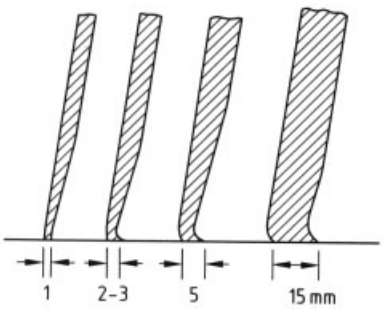

Figure 3. Types of doctor blade

In the case of floating knife coaters (Fig. 4A, B) and blanket coaters (Fig. 4C, D), the carrier is pressed firmly against the knife; this is required for the application of thin base coats on fabrics. A knife-over-roll coater (Fig. 4E), in which the space between doctor blade and carrier can be precisely adjusted, is used for the application of thicker layers up to ca. 1 mm.

The use of *applicator rolls* is of advantage for coating weights of 10 – 100 g/m^2. The coating material is transferred onto the carrier by the applicator roll, which is wetted with the paste. In direct roll coating, the applicator roll takes the paste out of a coating pan (Fig. 5A). The reverse roll coater is based on the indirect method (Fig. 5B). A metering roller removes surplus paste from the applicator roll which rotates counter to the direction of travel of the carrier. Other processes use a screen cylinder (Fig. 6A), a gravure roll (Fig. 6B), or other roller systems [24], [25].

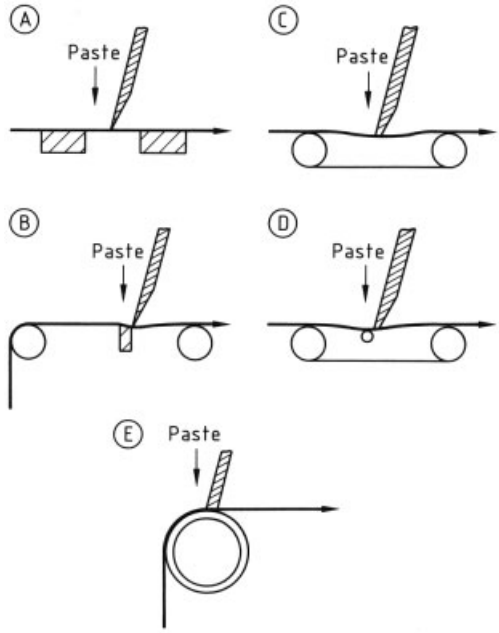

Figure 4. Knife coating systems
A), B) Floating knife coater; C), D) Blanket coater; E) Knife-over-roll coater

4.2.2. Coating Processes

Coating may be performed by either the direct or the transfer method.

Direct Coating. Direct application of the coating material to the carrier is suitable for woven fabrics, films, and, to a limited extent, for nonwovens. The carrier for the leather imitate is placed on a reeling off device (Fig. 7a) and then passed through the coating stations (b, d). A predryer (c) is positioned behind each coating station and may consist of a short drying tunnel,

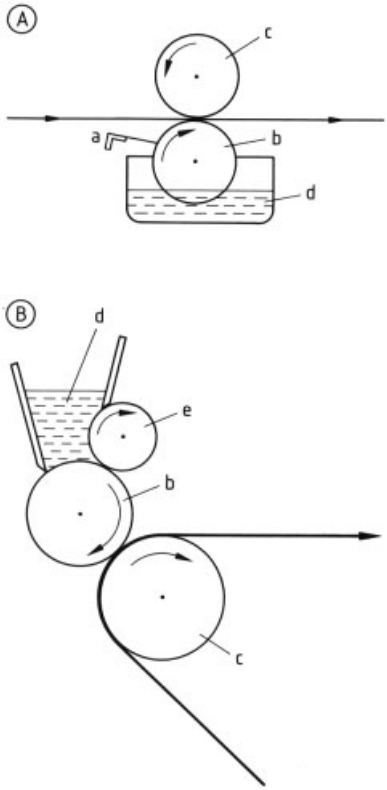

Figure 5. Direct roll coating (A) and indirect (reverse roll) coating (B)
a) Doctor blade; b) Applicator roll; c) Backing roll; d) Coating material; e) Metering roll

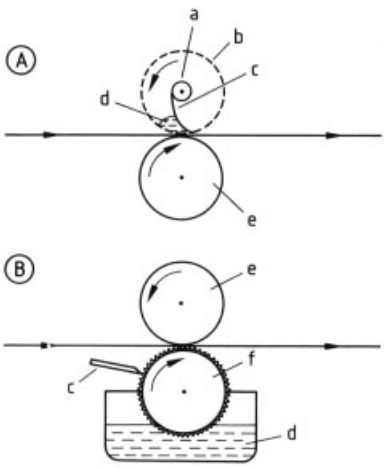

Figure 6. Coating with a screen cylinder (A) or a gravure roll (B)
a) Paste inlet; b) Screen cylinder; c) Doctor; d) Coating material; e) Backing roll; f) Gravure roll

IR radiators, or heated rollers. After predrying, known as pregelling in the case of plasticized PVC, the coating is hard enough to take another layer. In general, two to three coats may be applied. The coating attains its final firmness in a long main drying tunnel (e) in which a foam may also be produced if the layer already contains a chemical blowing agent. In this gas- or steam-heated tunnel, hot air of up to approximately 200 °C flows at right angles to the web (i.e., vertically). A guiding tenter can be used to avoid shrinking of the product in the tunnel. After passing through the main tunnel, the coated carrier is cooled (e.g., on water-filled rollers, f) and then wound in lengths of 100 – 1000 m on large reels.

The fabric used for direct coating must be dense and free of knots. It is usually washed and desized before processing. In the first coating station a thin base coat is applied in the form of a highly viscous paste. The paste should not pass through the carrier nor penetrate too deeply into it because this would result in a deterioration of the mechanical properties of the carrier. Pastes for foam and covering layers are applied in the following coating stations. The smooth surface of the coated material is lacquered (see Section 3.3) and embossed with a leatherlike grain in subsequent production steps.

Transfer Coating. In the transfer or reverse coating method, the coating material is spread as a paste on a release paper (Fig. 8a). As in direct coating, several coating stations with pregelling units can be used. The carrier for the leather imitate is placed on a reeling-off device and is applied to the coating after the last wet-coating step. Alternatively, an adhesive paste may be printed onto the pregelled coating. The textile or film carrier is then laminated to this layer. The reusable release paper is removed from the hardened coating after it has passed through the main drying tunnel. These release papers may be impregnated with silicone and must have a certain adhesiveness for the coating. The coating must not slip in the drying tunnel or on the rollers, but must also be easily detachable.

The reverse technique has many advantages over the direct method. The carrier for the leather imitate is stressed very little during production, so that light fabrics and, above all, knits can be used. The release papers can be provided with

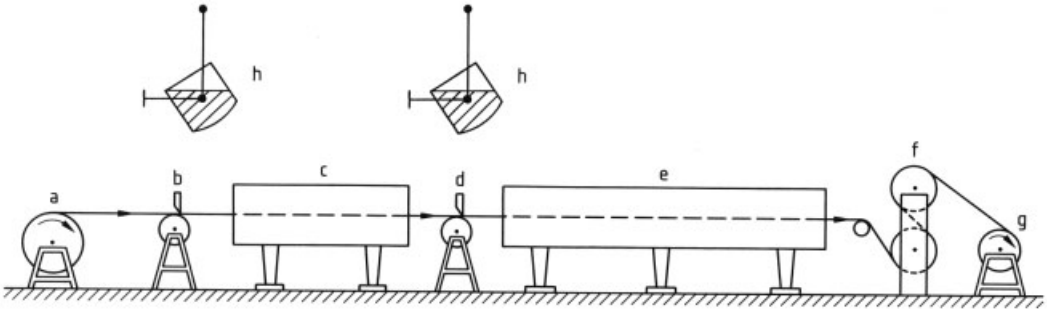

Figure 7. Direct coating line with two coating stations
a) Roll of carrier material; b) First doctor (knife); c) Predryer; d) Second doctor (knife); e) Main drying tunnel; f) Cooling rolls; g) Winding; h) Paste container

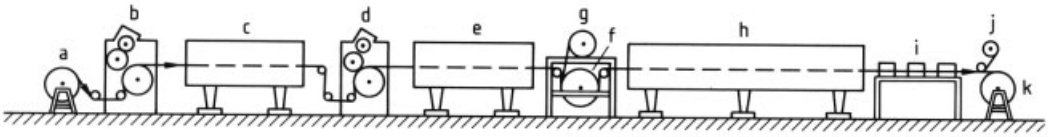

Figure 8. Transfer coating line
a) Release paper; b) First roll applicator; c) First predrier; d) Second roll applicator; e) Second predrier; f) Laminating device; g) Textile for film carrier; h) Main drying tunnel; i) Cooler; j) Winder for leather imitate; k) Winder for release paper

a negative profile of the leather grain. A leather imitate with the desired heat-stable grain is then obtained. Carrier-free foam films can only be produced by the reverse method. A disadvantage of this method is the additional paper cost.

Other Coating Methods. Thermoplastic polymers are also processed by using extrusion coating or the melt roll coating process [26].

In the *extrusion coating process,* a thin layer of the coating material is extruded through a slot die directly onto the carrier (Fig. 9). This method requires polymers with a narrow softening range such as low-density polyethylene. The *melt roll coating process* is suitable for a series of thermoplasts (e.g., PVC, polypropylene, polyethylene, and polystyrene) which are formed into a film on a calender and subsequently heat-laminated to the carrier on the last calender roll (Fig. 10). For the production of compact film leather imitates, single films are usually formed on a calender which are then thermally laminated in doubling and tripling machines and simultaneously provided with a leatherlike grain. The grained films can subsequently be coated with a foam layer or laminated with a foam sheet. This method produces leather imitates that are similar to the foam films made by the reverse coating technique.

Porous, vapor-permeable coatings (poromerics) can be made by the *coagulation method* (Fig. 11). The coating material, usually a polyurethane, is dissolved in a water-miscible solvent (e.g., dimethylformamide) and spread on a textile carrier. The coating hardens as the solvent is gradually displaced by water. The water-insoluble coating coagulates to give a porous film. This method is used, for example, to produce a moisture-storing base layer which can be covered with a compact covering layer by the reverse method. Leather imitates of this type are used in the shoe industry.

4.2.3. Finishing

Most leather imitates are printed and lacquered (see Section 3.3) to improve the finished surface of the material. The pigmented lacquers and transparent varnishes are usually applied by gravure printing. Other printing processes are also used, such as letterpress, flexography, or screen printing. The finish can be adjusted by using dull lacquers.

Embossing. Smooth thermoplastic coatings obtain their leatherlike grain in a final step by hot embossing (Fig. 12). The grain is usually pressed

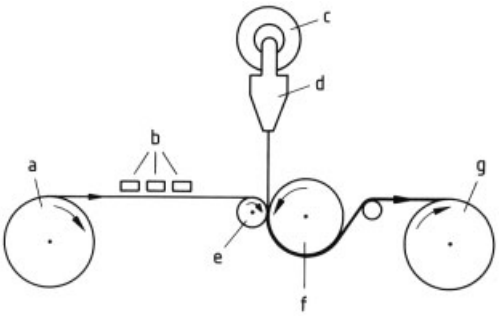

Figure 9. Extrusion coating
a) Reeling-off device for carrier material; b) IR radiators; c) Extruder; d) Coating die; e) Backing roll; f) Cooling cylinder; g) Winder

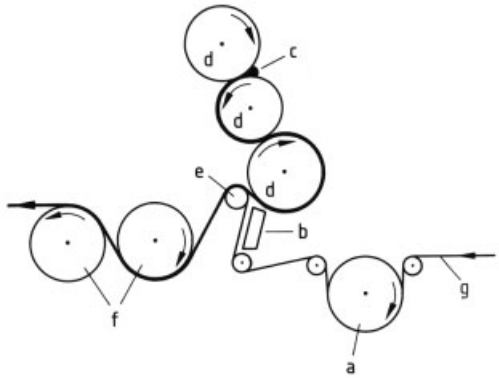

Figure 10. Melt roll coating
a) Preheating roll; b) Radiator; c) Polymer mass; d) Calender rolls; e) Pickup roll; f) Cooling rolls; g) Carrier material

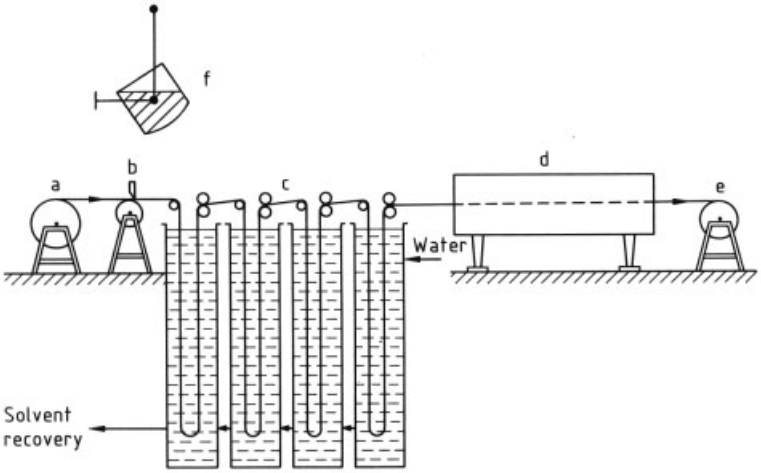

Figure 11. Coagulation coating
a) Reeling-off device for carrier material; b) Doctor (knife) coating; c) Baths for coagulation and washing out of solvent; d) Dryer; e) Winder; f) Paste container

into the softened surface with a roll. Before embossing the coating is heated to 120 – 220 °C, with heating rolls (a) and IR radiators (b), and is then pressed onto the engraved steel embossing roll (c) with a rubber roll (d). Cooling of the embossing roll improves the heat stability of the grain. However, the grain collapses if the leather imitate is heated to approximately the embossing temperature during further processing.

In the embossing of foamed layers, a gap must be set between the embossing roll and the rubber roll to avoid destruction of the foam layer.

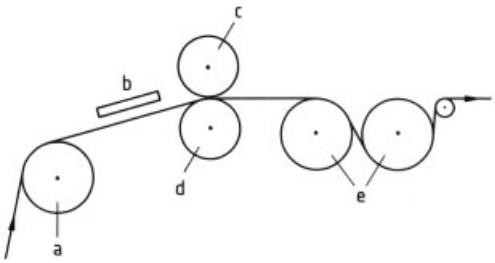

Figure 12. Embossing calender
a) Heating roll; b) IR radiator; c) Embossing roll; d) Backing roll; e) Cooling rolls

Embossing devices can be integrated in the coating plant. Embossing can be carried out immediately after coating especially in the case of foamless layers applied on a calender.

A natural leather grain can be reproduced almost perfectly on an embossing roll, which can be used to produce leather imitates with a very leatherlike surface.

Suede imitates are produced by various methods. After the coating process the covering layer can be roughened by sanding or flocked with short synthetic fibers (e.g., polyamide fibers). In addition, the covering layer can acquire a surface of short fibers made from the coating material (e.g., polyurethane) during the coating procedure.

Suede imitates based on nonwovens can be made from microfibers, which are bundled and encased in a polymer (e.g., polystyrene). After production the nonwoven fabric is hardened with coagulated polyurethane. The covering polymer is subsequently removed by dissolution, leaving the microfibers with a certain mobility and increasing the surface area of the impregnated nonwoven fabric.

5. Quality Specifications and Analysis

Varying demands are made on leather imitates depending on the intended application of the product. It is not possible to meet all these requirements at the same time. A large number of test methods have been developed to determine the physical and chemical properties. These methods are documented, e.g., in ASTM, DIN, and ISO specifications. Certain properties of leather imitates are important in all applications and are generally tested according to standard methods. These include measurements of the tear resistance and the tear propagation strength. Furthermore, it is important to determine the adherence between carrier and coating, the extensibility, and abrasion resistance. Aging resistance is tested by storing the leather imitate at very high or very low temperature in the presence of light or in the dark before mechanical testing. Polyurethane leather imitates, in particular, are subjected to damp, warm storage (hydrolysis test).

The *automobile industry* is an important purchaser of leather imitates and has developed its own quality criteria and tests [27]. This application requires especially heat-resistant, lightfast leather imitates, which do not embrittle in the cold and are abrasion resistant. Moreover, they must frequently have a flame retarding finish. Since the volatile components of leather imitates can cloud the windows (fogging effect); low-fogging leather imitates produced from low-fogging raw materials are being increasingly used.

The *shoe industry* has taken over many test methods (ASTM, DIN, ISO) from the leather field. Apart from the mechanical strength, abrasion resistance and cold flex temperature are of importance. In addition, the permeability to water vapor and the water storage capacity should be measured. Leather imitates used in *apparel applications* must be perspiration-resistant and fast to washing or dry cleaning.

6. Economic Aspects and Uses

The total production of leather imitates can only be estimated. In 2000 the production of textile and film leather imitates in Western Europe, with the exception of Spain, Portugal, and the Scandinavian countries, was approximately $500 \times 10^6 \mathrm{m}^2$. The automobile industry was the largest consumer with approximately $275 \times 10^6\,\mathrm{m}^2$(film leather imitates ca. $135 \times 10^6\,\mathrm{m}^2$, textile leather imitates ca. $140 \times 10^6\,\mathrm{m}^2$), whereas the shoe, and the furniture and bag industries consumed about $105 \times 10^6\,\mathrm{m}^2$ and $90 \times 10^6\,\mathrm{m}^2$, respectively. The consumption of leather imitates by the apparel industry amounted to approximately $50 \times 10^6\,\mathrm{m}^2$. The coating materials are mainly poly(vinyl chloride) (50 %) and polyurethane (20 %). The remaining 30 % are made from the raw materials described in Section 3.2. Leather imitates compete with leather in all fields of application, but also have their own areas of application. Leather imitates are produced specifically for their various applications, where they are equal or even superior to leather in many important properties. These include the mechani-

cal properties and the aging characteristics, their permeability to air and water vapor, and their ability to store moisture. The structure and finish of the grain of all types of animal skins can be perfectly reproduced on leather imitates.

The automobile industry uses leather imitates for seats and roofs. The inner lining of doors and the dash boards are often made of carrier-free foam films. Film leather imitates are laminated to a carrier material of wood fibers or plastic by vacuum thermoforming. The film is heated to 140 – 200 °C and formed under vacuum conditions. This processing method requires films with an especially temperature-resistant grain. They can be made either by the reverse method with negatively embossed release paper or by embossing the product at very high temperatures. Compact imitation leather films are often foamed on the reverse side for processing of automobile parts. Textile leather imitates that are not permanently heat deformed can be welded, glued, or sewn. Recent developments deal with the "in-mould decoration" technique where parts are obtained by placing the textile laminate in a mould followed by an injection moulding process step ("direct lamination").

The *shoe industry* mainly uses leather imitates for sport shoes and fashionable footwear. The trend is toward high-quality leather imitates based on polyurethane-coated nonwovens with highly resistant surfaces. The products have the wearing qualities of leather. They have a good storage capacity for moisture and the shoes adjust to the feet on wearing. Leather imitates that are permeable to water vapor are also suitable for lining shoes. Inexpensive shoes and sandals, where permeability to water vapor and moisture storage are not of vital importance, can be made from PVC leather imitates.

The *furniture industry* still uses large quantities of leather imitates, especially for office furniture. In the *bag sector*, leather imitates are required for suitcases, bags, and cases, as well as for covers and small leather articles.

The production of *apparel leather imitates* increased in the 1970s due to the introduction of crinkle leather and suede imitates but has since decreased dramatically. It is, however, still of importance for inexpensive outer garments.

7. References

General References

1. E. F. Hall (ed.): "Artificial Leathers – their Manufacture, Properties and Uses," *Proceedings of 1st Satra International Conference,* **vols. 1, 2,** Satra, Kettering, England 1971.
2. *Coated Fabrics Symposium,* American Association of Textile Chemists and Colorists, Newton, Massachusetts 1982; Wakefield, Massachusetts 1985.
3. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* **vol. 4,** W. A. Colom, Berlin 1976.
4. G. W. Becker, D. Braun (eds.): *Kunststoff Handbuch Polyvinylchlorid,* **vol. 1, vol. 2,** Carl Hanser Verlag, München Wien 1986.
5. G. W. Becker, D. Braun (eds.): *Kunststoff Handbuch Polyurethane,* **vol. 7,** Carl Hanser Verlag, München Wien 1983.
6. K. Weinmann: *Beschichten mit Lacken und Kunststoffen,* W. A. Colomb, Stuttgart 1967.

Specific References

7. DIN 16 922 Kunstleder.
8. W. Wittke, *Coating* (1988) no. 8, 278 – 282; no. 9, 340 – 345.
9. Bayer AG, DE 2 902 090, 1979 (E. Koenig, J. Perdain, H.-J. Koch, A. Schmitz, W. Thoma).
10. D.-I. Schütze, "Innovative Produkte für die Textilbeschichtung," *1st Symposium on Textile Coating & Laminating*, Gent, 1999.
11. Bayer AG, EP 825 209, 1998 (S. Groth, J. Pedain, L. Schmalstieg, D.-I. Schütze).
12. *Allgemeiner Vliesstoffreport* **1** (1988) 30 – 32.
13. D. Groitzsch "Ultrafine Microfiber Spunbond for Hygiene and Medical Application," *EDANA International Nonwovens Symposium*, Prague, 7 – 8 June, 2000.
14. S. Krämer, *Plastverarbeiter* **39** (1988) no. 8, 30 – 35; no. 9, 23 – 28.
15. D. Thölmann, "Emissionen von Phthalaten aus Weich PVC" *Proc. Achema-Konferenz*, München, 1996.
16. WHO International Series on Chemical Safety 131 DEHP, Geneva 1992.
17. J. A. McLachlan, *Science* **277** (1977) 462 – 463.
18. *BUA-Report* **212** (1999).
19. Morflex chemical Company, US 4 954 649, 1990 (E. H. Hull, E. P. Frappier):
20. Bayer AG, US 4 507 413, 1985 (W. Thoma, R. Langel, A. Schmitz, W. Schroer).

21. Produkte für die Textilbeschichtung, Anwendungstechnische Information, Bayer AG, 1994.
22. Bayer AG, DE 3 004 327, (E. Komarek).
23. J. L. Préat, R. Saffert, *Plastverarbeiter* **38** (1987) no. 6, 34 – 45.
24. M. Dominik, *Kunststoffe* **78** (1988) no. 10, 1000 – 1004.
25. W. Schaetzle, *Coating* (1987) no. 9, 314 – 316.
26. J. Brocks, *Melliand Textilber.* (1988) no. 1, 66 – 70.
27. G. F. Schmidt, W. Deininger, *Textilveredel.* **23** (1988) no. 5, 163 – 170.

Lecithin

HIROYUKI TANNO, Central Research Laboratories, Ajinomoto Co. Inc., Kawasaki, Japan

1. **Introduction** 1
2. **Production** 2
3. **Commercial Grades of Lecithin** 3
4. **Physical Properties** 4
5. **Chemical Properties** 4
6. **Uses** 5
7. **Quality Specifications and Analysis** . . . 6
8. **References** 6

1. Introduction

Lecithin was first isolated in 1850 by GOBLEY, who separated it from egg yolk; it was named lecithin after the greek word for egg yolk, *lekithos*.

Table 1. Composition of soybean lecithin

Component	Content, wt %
Phosphatidylcholine	19 – 21
Phosphatidylethanolamine	8 – 20
Inositol phosphatides	20 – 21
Other phosphatides	5 – 11
Soybean oil	33 – 35
Sterols	2 – 5
Carbohydrates, free	5
Water	1

Later, lecithin was identified as a mixture of many phosphatides. In the 1920s, a process which separates lecithin from crude soybean oil was developed, and today soybeans are still the main source for the mass production of lecithin. Generally, the term lecithin is used to describe soybean lecithin, which is composed mainly of soybean phosphatides and also contains soybean oil. The composition of soybean lecithin is given in Table 1. The uses of lecithin can be divided into foods, feeds, and general industrial uses. Lecithin is used widely in foods because it is both a natural emulsifier and a nutritious food in its own right (see Table 2) and since it gives rise to excellent flavors. In these applications, lecithin is mainly used as an emulsifier, and some treatment of lecithin (e.g., chemical modification and solvent extraction) is carried out. The physiological function of lecithin has been confirmed.

The term lecithin is also used in a chemical sense. In this case, lecithin refers to phosphatidylcholine (PC). The nomenclature of lecithin has been discussed by IUPAC and the International Union of Biochemistry (IUB). The chemical term for phosphatidylcholine was recommended to designate 1,2-diacylglycero-3-phosphorylcholine. The structures of the major phosphatides present in lecithin are shown in the following:

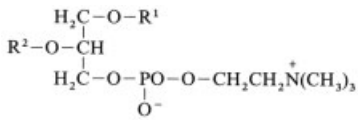

Phosphatidylcholine (PC)

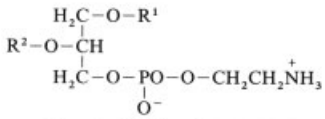

Phosphatidylethanolamine (PE)

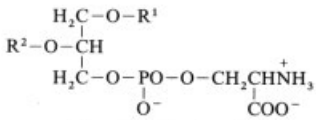

Phosphatidylserine (PS)

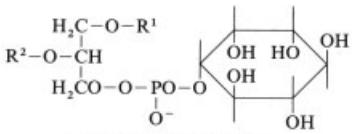

Phosphatidylinositol (PI)

where R^1 and R^2 represent fatty acid residues.

In this chapter, lecithin without special definition indicates soybean lecithin; the chemical term lecithin refers to phosphatidylcholine (PC).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_293

2. Production

Modern production of lecithin is based mainly on plant seeds, especially soybeans. The quality of lecithin products is determined by the production process and the quality of the raw material soybeans. The flow sheet for a lecithin production process is shown in Figure 1.

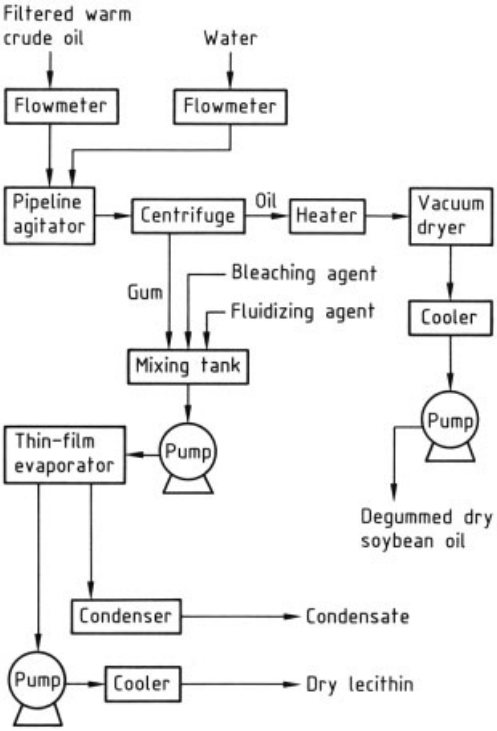

Figure 1. Flow sheet for the production of crude lecithin and degummed soybean oil

Seed Preparation. Soybean oil from mature soybeans containing few burnt or fragmented beans gives the highest quality lecithins. Fragmented beans, or beans that have been stored for a long time, result in reduced lecithin yield. Processing of immature beans results in a high chlorophyll content so that the color of the lecithin is dark. This color is difficult to remove by bleaching.

Degumming. Crude oil which contains lecithin is obtained from raw material seeds by compression. The crude oil contains about 2 – 3 wt % of phosphatide. Water (1 – 2.5 vol. %) is added to the crude oil to hydrate the phosphatides. The method that is generally used in the United States involves adding 2 vol % water. The water–oil mixture is then heated at 70 °C for 30 – 60 min in a pipeline agitator that ensures thorough mixing. The oil-insoluble lecithin fraction (a wet gum known as lecithin hydrate) is separated by centrifugation. The gum is then transferred to a holding tank to allow addition of fluidizing and bleaching agents, if required.

Table 2. Concentration of some minor components of soybean lecithin

Component	Concentration, µg/g
Tocopherol	1300
Biotin	0.42
Folic acid	0.60
Thiamin	0.12
Riboflavin	0.33
Pantothenic acid	5.59
Pyridoxine	0.29
Niacin	0.12

Bleaching. To obtain a lighter colored product, hydrogen peroxide and benzoyl peroxide are used to bleach the lecithin. This is carried out either by using a 0.3 – 1.5 % hydrogen peroxide solution instead of water for the degumming process, or by addition of peroxide to the holding tank. The color of a sample is determined using the Gardner method.

Drying. The bleached gum contains 25 – 50 % water. Since the high moisture content encourages the growth of microorganisms, it is necessary to dry the product. The method used is thin-film vacuum drying at a temperature of 80 – 105 °C and a pressure of 3.3 – 4 kPa. The drying is carried out within 1 – 2 min, because high temperatures promote the degeneration (darkening) of lecithin. After drying. it is necessary to cool the lecithin to 55 – 60 °C immediately, because of its instability at high temperature.

Molecular-Membrane Process. A molecular-membrane degumming process has been developed. The molecular mass of the triglycerides is ca. 900, while that of the phosphatides is ca. 800. However, in non-polar solvents such as hexane, the phosphatide molecules of lecithin aggregate to form micelles that have molecular masses of 20 000 – 30 000. In this way, lecithin

can be separated from the triglycerides by a membrane.

3. Commercial Grades of Lecithin

The National Soybean Processors Association (NSPA) trading rules classify commercial lecithins into six grades according to color and viscosity. Fluid lecithin and plastic lecithin are classified by viscosity and each of these is further classified into three classes according to color: natural (unbleached), bleached, and double-bleached. Of these six, only the two unbleached grades are permitted as food additives. The specifications for the six types of lecithin are listed in Table 3.

Table 3. Commercial soybean lecithin specifications

Analysis	Grade		
	Natural	Bleached	Double bleached
Fluid lecithin			
Acetone insoluble, % (min)	62	62	62
Moisture, % (max) [a]	1	1	1
Benzene insoluble, % (max)	0.3	0.3	0.3
Acid value (max)	32	32	32
Color (max) [b]	10	7	4
Viscosity at 25 °C, Pa · s (max)	15	15	15
Plastic lecithin			
Acetone insoluble, % (min)	65	65	65
Moisture, % (max) [a]	1	1	1
Benzene insoluble, % (max)	0.3	0.3	0.3
Acid value (max)	30	30	30
Color (max) [b]	10	7	4
Penetration, mm (max) [c]	22	22	22

[a] Determined by toluene distillation.
[b] Measured on a 5 % solution in mineral oil using the Gardner method.
[c] Using precision cone 73 525 and penetrometer 73 510.

The Food Chemical Codex (FCC) III classifies lecithin products as follows: 1) crude lecithin, 2) fluidized lecithin, 3) highly filtered lecithin, 4) compound lecithin, 5) chemically modified lecithin, 6) fractionated lecithin.

Crude Lecithin generally has high viscosity and is a brown fluid. The NSPA has standardized bleaching, the addition of minerals, and the moisture content.

Fluidized Lecithin is produced by addition of mineral salts, diluents, or fatty acids to crude lecithin. Fluidized lecithin can be handled more easily due to its lower viscosity. Although fluidized lecithins have a low content of acetone-insoluble material, they are graded as high class.

Highly Filtered Lecithin. High-grade lecithin products (e.g., for food) are made by removing the hexane-insoluble material by filtration.

Compound Lecithins. Compound lecithins are produced by blending lecithin with other components. These include oils, polysorbate, monoglyceride, emulsifiers, solvents, and plasticizers; they are generally blended with dry fluid lecithin.

Chemically Modified Lecithin. Lecithins are chemically modified by hydrogenation, hydroxylation, ethoxylation, halogenation, sulfonation, acylation, ozonization, and phosphorylation. Of the chemically modified lecithins, only hydroxylated lecithin is permitted as a food additive at present. Hydroxylation of lecithin increases the hydrophilicity and therefore also the tendency to form oil-in-water emulsions. Hydroxylated lecithins are produced by treating crude lecithin with water and hydrogen peroxide in lactic acid or acetic acid.

Fractionated Lecithin. The composition of crude lecithin, which contains many components, can be changed by fractionation with various solvents.

Acetone Fractionation. Most of the triglycerides and fatty acids can be separated from crude lecithin by acetone fractionation to give oil-free lecithin powders. Oil-free lecithin costs more than ten times as much as regular lecithin. Because it is oil-free, acetone-fractionated lecithin has greater hydrophilicity and increased emulsifying activity.

Alcohol-Extracted Lecithin. Lecithin can be enriched by alcohol extraction. Phosphatidylcholine (PC) is concentrated in the alcohol-soluble fraction. This fraction has increased emulsifying activity for the formation of oil-in-water emulsions. The alcohol-insoluble fraction is rich in the hydrophobic phosphatidylinositol (PI) and therefore favors the formation of water-in-oil emulsions. Phosphatidylethanolamine (PE) is evenly divided between the al-

cohol soluble and insoluble fractions. As alcohol fractionation can change the phospholipid composition, its emulsifying activity is greater than that of acetone-fractionated oil-free lecithin.

Other Fractionation Methods. Lecithins containing large amounts of phosphatidylcholine can be obtained by adsorption and distribution chromatography. For example, products which contain up to 95 % phosphatidylcholine are produced for pharmaceutical use by means of this method.

4. Physical Properties

There are many parameters which characterize the physical properties of lecithin. These are acetone-insoluble matter (AI), acid value (AV), moisture content, hexane-insoluble matter (HI), color, consistency, and clarity. Commercial lecithin is mainly composed of phospholipids (ca. 66 %); the other major components are nonphospholipid compounds such as triglycerides. Since phospholipids are insoluble in acetone, and triglyceride is soluble in acetone, the amount of *acetone-insoluble material* in lecithin corresponds approximately to the phospholipid content.

The *acid value* is measured in miligrams of KOH per gram of lecithin and is determined by acid titration. The acid value is the phospholipid index number as defined in AOCS OM Ja-6-55.

The moisture content of lecithin products is typically less than 1.5 %. The moisture content is determined using the Karl Fischer method.

The *hexane-insoluble matter* (HI) indicates the amount of residual fiber which originates from the production process. The HI measurement is the official Foods Chemicals Codex test method; the HI value must be suppressed to 0.1 %. A high HI level indicates a low grade of lecithin. High-clarity lecithin can be produced by oil filtration to give HI levels of nearly zero. Although the major constituents of lecithin are colorless, the color of commercial lecithin is red-brown or dark-red. The colored components include carotinoid pigments and pigments formed by reaction of amino compounds with phosphatide sugars (Maillard reaction). Phosphatidylethanolamine also contributes to the color through formation of aldol products. The bleaching method is treatment with hydrogen peroxide or benzoyl peroxide. The color is measured by the Gardner method.

Lecithin is classified as plastic or fluid according to its viscosity. The viscosity of fluid lecithin is decreased by addition of fatty acids. Commercial lecithin can be produced as a powder by treatment of the alcohol or acetone fractions. The acetone-insoluble grades consist of a mixture of isolated phospholipids which have either granular or powdered forms. The consistency of alcohol-fractionated lecithins depends on the phosphatidylcholine/phosphatidylethanolamine ratio, which in turn depends on the polarity of the alcohol used, the ratio of lecithin to alcohol, the temperature, and the extraction time.

Lecithin is soluble in oils and is hydrated in water, in which it disperses with formation of an emulsion. The solubility of crude lecithin and of the refined lecithin phospholipids is listed in Table 4.

Table 4. Solubitity of lecithin and its components

	Solubility in		
	Water	Alcohol	Acetone
Soybean lecithin	insoluble	soluble	insoluble
Phosphatidylcholine	soluble	readily soluble	sparingly soluble
Phosphatidylethanolamine	readily soluble	soluble	insoluble
Phosphatidylinositol	readily soluble	insoluble	insoluble

5. Chemical Properties

The chemical properties of lecithin are largely determined by the presence of phosphatides and the fatty acid residues which they contain. The production of modified lecithin has increased because of the need for lecithins with specific properties, such as emulsifiers and wetting agents. The modification methods are either physical, enzymatic, or chemical.

The chemical structure of lecithin remains unchanged upon physical modification. Nevertheless, acetone or alcohol fractionation allows the ratio PC/PE and the amount of phospholipid to be changed, leading to changes in the properties and giving new functions which are lacked by crude lecithin.

Table 5. Quality specifications for lecithin

Specification	Standard			
	FAO/WHO, 8431	FCC, 182 and 1400	EEC, E 332	Japan
Acid value	25 – 35	36	45	40
Benzene insoluble, %	0.3	0.3		0.3
Acetone insoluble, %	62	50	60	40
Peroxide value, meq/kg (max)	10		10	10
Heavy metals, ppm	40	40	40	20
Arsenic, ppm	3	3	3	4
Weight loss on drying, %	2	1.5	2	2

Lecithin can be hydrolyzed enzymatically. On an industrial scale, the enzyme used is phospholipase A 2, which selectively hydrolyzes the fatty acid in the 2 position of the phospholipid and results in the lyso form of lecithin.

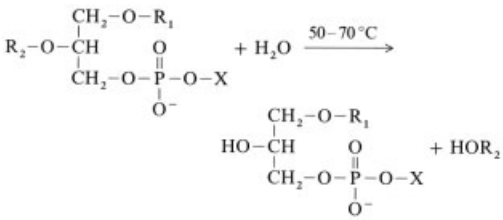

Lysolecithin has increased solubility in water and greater emulsifying activity for the formation of oil-in-water emulsions.

Chemical hydrolysis by acids or alkalis is non-specific and affords products that are unsuitably colored.

The reaction of lecithin with acetic anhydride leads to acetylation of the amino group of phosphatidylethanolamine. Acetylated lecithin more readily forms oil-in-water emulsions. The treatment of lecithin with lactic acid and highly concentrated hydrogen peroxide introduces hydroxy groups into the unsaturated fatty acids of the phospholipids. Hydroxylated lecithin more readily forms dispersions with cold water.

6. Uses

Major food applications of lecithin are in chocolate, confectionery, margarine, and baked products. Non-food uses of lecithin include textiles, insecticides, paints, and many others. Lecithin functions as follows: 1) as an emulsifier, 2) as an antispattering agent, 3) crystallization control, 4) viscosity modifier, 5) antisticking agent, 6) wetting agent, 7) dispersing agent, 8) release agent. Usually, the amount used accounts for 0.1 – 1.0 wt % of the final product.

Important food applications of lecithin include baked goods, in which it improves the volume, disperses fats, and acts as an antioxidant, increasing shelf life; chocolate, in which it reduces viscosity and prevents crystallization; its addition to instant foods as a wetting, dispersing, emulsifying, and stabilizing agent; and margarine, in which it acts as an antispattering agent and dispersant.

Non-food uses of lecithin include lubricating oil, fuel oil, leather, textiles, dyes, rubber, soaps, and cosmetics. Lecithin is used in pharmaceuticals as a biological emulsifier and as a source of phosphatidylcholine; in paints as a pigment dispersant and for the stabilization of emulsions; in the production of leather to improve the absorption of fat liquor; and for mosquito control by reduction of surface tension.

Another use is as an additive to animal feeds and as a milk replacer for calves. Lecithin is said to improve the physical properties of the feed and to promote adsorption of some components. Furthermore, because lecithin contains choline and inositol, it promotes the metabolism of oils, fats, and cholesterol in the liver.

Storage. Long-term storage of lecithin at high temperature in the presence of air leads to oxidation of the unsaturated fatty acids, resulting in an off flavor and a black coloration. Therefore, it is necessary to exclude air and to maintain low temperature ($< 20\,°C$) when lecithin is stored for a long time.

Since lecithin contains nitrogen-containing compounds, such as ethanolamine, serine, and B vitamins, water suspensions of lecithin encourage the growth of microorganisms, and therefore

require the addition of preservatives for long-term storage.

7. Quality Specifications and Analysis

Because lecithin compounds (including PC, PE, and PI) are too polar to be subjected to direct gas chromatographic analysis, liquid chromatographic methods are usually used for their analysis. For example, two-dimensional thin layer chromatography is the method most often used, and high performance liquid chromatographic methods are also available.

Some quality specifications for lecithin are listed in Table 5.

8. References

1. W. L. Erdarl et al., *J. Am. Oil Chem. Soc.* **50** (1973) 513.
2. F. B. Jungalwala et al., *Biochem. J.* **145** (1975) 517.
3. S. Chen et al., *J. Chromatogr.* **208** (1981) 339.
4. F. L. Phillips et al., *J. Am. Oil Chem. Soc.* **58** (1981) 590.
5. J. Holzl, *Biochem. Z.* **341** (1965) 168.
6. B. W. Nichols, *Biochem. Biophys. Acta* **70** (1963) 417.
7. S. Chen et al., *J. Chromatogr.* **276** (1983) 37.
8. C. G. Duck-Chong et al., *Lipids* **18** (1983) 387.
9. Van Nieuwenhuyzen, *J. Am. Oil Chem. Soc.* **58** (1981) 886.
10. W. J. Strittmatter et al., *Nature (London)* **282** (1979) 857.
11. A. R. Tall et al., *J. Med.* **299** (1978) 1232.
12. R. J. Havel, *Ann. N. Y. Acad. Sci.* **348** (1980) 16.
13. A. Kurode, *Food Chemical* **5** (1989) 73.
14. L. R. Juneja et al., *J. Am. Oil Chem. Soc.* **66** (1989) 714.
15. P. J. A. O'Doherty et al., *Lipids* **8** (1973) 249.
16. P. Tso et al., *Gastroenterology* **80** (1981) 60.
17. R. Hell et al., GB 1 215 868, 1970.

Leuco Esters of Vat Dyes

TOBY PINK, Imperial Chemical Industries, PLC, London, United Kingdom

1. **Introduction** 1
2. **Manufacturing Methods Using Sulfur Trioxide–Amine Addition Compounds** 1
2.1. **The Durand & Huguenin Process for Indigoid Dyes** 1
2.2. **The Scottish Dyes Process for Anthraquinone Dyes** 2
2.3. **The IG Farben (Hoechst) Process for Indanthrone Dyes** 2
2.4. **Miscellaneous Processes** 3
3. **The ICI Process Using Chlorosulfonate as Esterifying Agent** 3
4. **Properties** 3
5. **Uses** 4
6. **Economic Aspects** 4
7. **References** 4

1. Introduction

The leucoesters of vat dyes are formed by esterifying each hydroxyl group of the leuco-form of the parent dye with a di- or polybasic acid.

Their technical value lies in their greatly improved leveling properties compared with the parent leuco dye and their suitability for use on a wide variety of fibers.

The only leuco esters of vat dyes available at present are the sodium or potassium salts of the sulfuric acid esters; other esters (e.g., of phosphoric or 3-sulfobenzoic acid) have no practical application. BADER, SUNDER and VAUCHER developed a method of synthesizing the leuco disulfuric ester of indigo (patented for Durand & Huguenin in 1921 [2], [3]). They adapted VERLEY's method for sulfating phenols by using chlorosulfuric acid ($ClSO_3H$) in pyridine [4] to esterify isolated, dry leucoindigo:

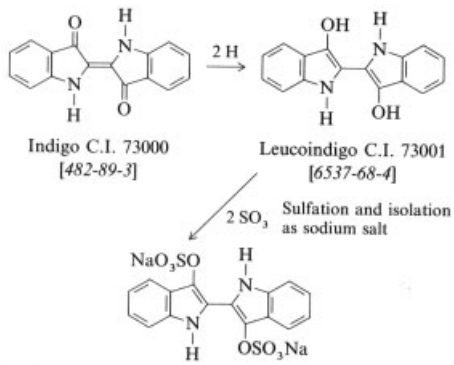

Indigosol O (Durand & Huguenin) C.I. 73002
Anthrasol O (Hoechst) [3875-70-5]

A more generally useful method that avoids isolation of the leuco dye was devised by chemists at Morton Sundour Fabrics (the parent company of Scottish Dyes) and patented in 1924 [2], [5]. This process was first used on a production scale in 1925 to prepare the leuco disulfuric ester of 16,17-dimethoxyviolanthrone as its disodium salt [2], [6] (→ Anthraquinone Dyes and Intermediates, Chap. 2.12.8.).

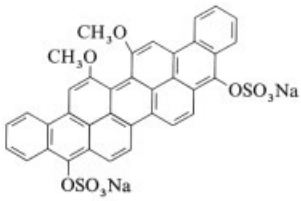

Soledon Jade Green X (Scottish Dyes, later ICI) C.I. 59826
Anthrasol Green IB (Hoechst) [2538-84-3]

The Scottish Dyes process is the basis of most manufacturing methods. Other, more recent, methods have limited or no commercial application. Leuco esters of vat dyes have declined significantly in commercial importance in recent years (see Chap. 6).

2. Manufacturing Methods Using Sulfur Trioxide–Amine Addition Compounds

2.1. The Durand & Huguenin Process for Indigoid Dyes

In the developed prototype process isolated, dry leucoindigo is esterified by using chlorosulfuric acid in dimethylaniline.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_299

Monochlorobenzene or excess dimethylaniline is used as solvent, and esterification is effected at 50 – 60 °C. The product (Indigosol O) is isolated as the disodium salt [7], [8]. This method, requiring isolation of the leuco vat dye, is useful primarily for indigoid and thioindigoid dyes.

2.2. The Scottish Dyes Process for Anthraquinone Dyes

The Scottish Dyes process avoids isolation of unesterified leuco forms which, for most anthraquinone and related dyes, are unstable and often very soluble. The overall reaction formally proceeds as follows using a *para*-quinone moiety:

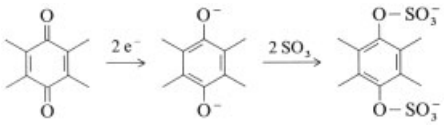

In this method, the dry, powdered vat dye is stirred into a cooled solution of chlorosulfuric acid in excess pyridine and reduced by addition of a metal powder (usually iron or copper). Esterification proceeds concurrently with reduction, but not necessarily at the same rate. Usual reaction temperatures are 40 – 60 °C. Pyridine acts as a solvent, a hydrogen chloride acceptor, and a sulfur trioxide carrier:

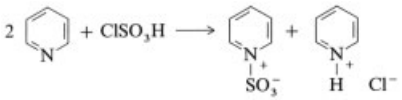

The reaction mixture is then poured into aqueous alkali and clarified; pyridine is removed and the product isolated as its sodium or, if the dye is relatively insoluble, its potassium salt. (Salts of lithium or ethanolamine have been proposed but not used to any extent) [9], [10].

The patent literature covers the use of many metals and alloys for reduction: iron, copper, aluminum, zinc, brass, and bronze.

Although widely applicable, this process does not work well with all vat dyes, particularly 1,4- and 1,5-disubstituted anthraquinone derivatives, because in these cases overreduction can occur, for example, with formation of anthranol esters [11] (loss of one hydroxyl group):

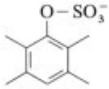

These products, which are usually not very soluble, do not regenerate on the fiber to the parent vat dye.

Some 1,4-disubstituted anthraquinone derivatives can be used satisfactorily. An example is

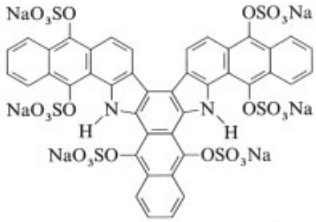

Anthrasol Brown IBR (Hoechst) C.I. 70801 [*23725-15-7*]
Soledon Dark Brown 3R (ICI)

made from Indanthren Brown BR (Anthraquinone Dyes and Intermediates).

2.3. The IG Farben (Hoechst) Process for Indanthrone Dyes

Direct esterification of indanthrone dyes produces a very insoluble diester rather than the desired tetraester. Hence, the following method (illustrated here by its most important application) was developed by IG Farben [10], [12], [13]:

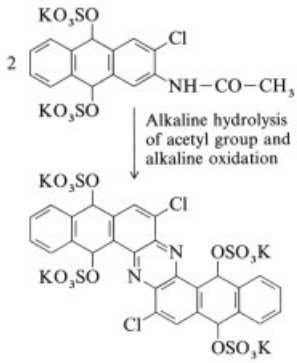

Indigosol (Anthrasol) Blue IBC (Hoechst)
Soledon Blue 2RC (ICI) C.I. 69826 [*2519-28-0*]

Lead(IV)oxide is a satisfactory and widely used oxidant, although other agents and anodic oxidation have been used [15], [16].

2.4. Miscellaneous Processes [17]

The ICI Dimethylformamide and Related Processes. Addition compounds of *N,N*-dialkylamides (particularly *N,N*-dimethylformamide) with sulfur trioxide can be used as the esterifying agents, with the amide itself (or a solvent such as dichloroethane) as the reaction medium [18].

The Durand & Huguenin Mixed Amine Process. Pyridine in the Scottish Dyes process can be replaced by a mixture of a strong tertiary alkylamine (e.g., *N,N*-diethylcyclohexylamine) and a weak tertiary amine (e.g., α-picoline). This combination enables esterification to proceed rapidly, whereas reduction is relatively slow, thereby minimizing overreduction [11], [19].

The American Cyanamid All-Aqueous Process. American Cyanamid proposed the use of addition compounds of sulfur trioxide and tertiary alkylamines such as *N*-ethylmorpholine as esterifying agents in an *aqueous* medium. If these agents are used in suitable excess, they react rapidly with the leuco vat dye in alkaline solution without excessive decomposition [16], [21].

3. The ICI Process Using Chlorosulfonate as Esterifying Agent

The ICI process is unusual because it does not employ an amine – sulfur trioxide addition compound as esterifying agent, but rather, a chlorosulfonate. Typically, methyltriethylammonium chlorosulfonate is used in a solvent such as acetonitrile. The chlorosulfonate ion is highly reactive and effects rapid esterification at low temperature [17], [20].

4. Properties

Stability. All leuco esters of vat dyes are to varying degrees, sensitive to heat, moisture, acid, atmospheric carbon dioxide, and light. They are, however, quite stable to alkaline hydrolysis.

Any local action of acids leads to hydrolysis of the ester groups, releasing more acid which can finally lead to total decomposition of the contents of a package. Various stabilizing agents are added during formulation to minimize this risk [22]. Urea, which also acts as a solubilizer, is particularly useful.

Drying is carried out under mildly alkaline conditions either at low temperature (ca. 50 °C) or by spray drying.

Storage should be in opaque, airtight packages away from sources of heat, including direct sunlight.

Solubility. The solubility of various leuco esters varies considerably. Sufficient solubility is particularly important in textile printing; thus, solubilizing agents such as urea, diethylene glycol, or thiodiglycol are usually incorporated in the print pastes.

Substantivity. The relative order of substantivity is generally the same as that of the sodium salts of the leuco forms of the parent vat dyes. However, the absolute values are always lower because of the presence of the sulfate group. This gives these dyes their most important property: improved leveling behavior which is required especially when pale shades are being dyed.

With the exception of Anthrasol Blue IBC, the anthraquinone dyes, as a rule, have distinctly higher substantivity than indigoid and thioindigoid dyes.

The best uptake of dye on the fiber is usually at 25 – 30 °C, but for a few anthraquinone dyes (e.g., Anthrasols Green IB and Brown IBR) the uptake is best at 50 – 70 °C.

Ease of Regeneration (Development). Dyes, once on the fiber, are normally regenerated with sulfuric acid and sodium nitrite. Most regenerate easily in a cold bath; a few require a temperature of 70 °C.

Lightfastness of the Developed Dye. Under standard test conditions on cellulose, most anthraquinone dyes are of very high fastness, 7 – 8 on the numerical scale. The indigoid and thioindigoid types generally have distinctly lower lightfastness.

5. Uses

Dyeing. Leuco esters of vat dyes are remarkably versatile and have been used to dye wool (perhaps the first application), silk, cellulose, polyester, cellulose – polyester blends, and even polyamide, although their lightfastness is relatively poor on the last. To date their main use has been for all-cellulose fabrics or cellulose – polyester blends. Dyeing of cellulose fabrics can be carried out by conventional exhaust and padding methods, with development usually by sulfuric acid and sodium nitrite. Cellulose – polyester blends are normally dyed by one of three methods: (1) the high-temperature one-bath method, (2) the Anthrasol – thermosol develop method and (3) the pad – thermosol cold-develop process [17], [23]. The last method is used when these dyes are applied in conjunction with disperse dyes.

Textile printing. The three steps involved in printing textiles are application of the dye in its soluble leuco form, dyeing onto the fiber, and finally regenerating (developing) the parent, insoluble vat dye. Dyeing must take place before regeneration.

Several methods are available for developing the printed dye, two of which are particularly important. These are the nitrite process using sodium nitrite (in or padded onto the print paste) with sulfuric acid as the oxidizing system, and the chlorate process using either sodium chlorate and ammonium thiocyanate (the sulfocyanide process) or ammonium chlorate [17].

The print pastes must be protected from sunlight and acidity to avoid premature development.

6. Economic Aspects

Development of modern textile application methods, together with associated improvements in the formulation of vat dyes, has reduced the need for their solubilized analogues. In addition, the development of reactive dyes with good color fastness has reduced the demand for indigoid and thioindigoid vat dyes and their solubilized equivalents. This, together with worldwide rationalization of the dye industry, has virtually elucidated manufacture in the Western world.

Although well over 60 of these dyes have been made, the core palette [7], [13] throughout the duration of their existance remained largely unchanged.

The products which survived for the longest time are the anthraquinone dyes with the highest light- and all-round fastness. Examples of these, already discussed, are Anthrasol Blue IBC, C.I. 69 826; Anthrasol Green IB, C.I. 59 826; and Anthrasol Brown IBR, C.I. 70 801.

7. References

General References

1. Colour Index, 3rd ed., vol. 3, pp. 3719 – 3837, vol. 4, pp. 4523 – 4533, Society of Dyersand Colourists, Bradford, England 1971.

Specific References

2. M. Bader, *Chem. Ztg.* **61** (1937) 741 – 744, 763 – 766; *Am. Dyest. Rep.* **27** (1938) 455 – 461, 467 – 471.
3. Durand & Huguenin, DE 424 981, 1921, (M. Bader, C. Sunder), *Friedländer,* **vol. 15,** p. 592.
4. A. Verley, *Bull. Soc. Chim. Fr.* **25** (1901) 46.
5. ICI Publication, M. R. Fox: "Dyemakers of Great Britain 1856 – 1976," p. 161, London 1987.
6. Scottish Dyes, GB 260 647, 1925 (J. Thomas, J. E. G. Harris, B. Wylam).
7. *Ullmann,* 3rd ed., **11,** 696 – 707.
8. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 2,** Academic Press, New York – London 1952, pp. 1046 – 1058.
9. F.I.A.T. Final Report 1313 II, p. 188 – 199.
10. B.I.O.S. Final Report 1493, p. 60 – 73.
11. H. Schenkel, *Chimia* **15** (1961) 203 – 207.
12. I.G. Farbenind., DE 470 809, 1926, *Friedländer,* **vol. 16,** p. 1316; DE 476 811, 1926 (K. Schirmacher, W. Schaich, A. Wolfram), *Friedländer,* **vol. 16,** pp. 1316 – 1317; DE 580 013, 1932 (J. Haller, G. Rösch), *Friedländer,* **vol. 20,** p. 1377.
13. H. A. Lubs: *The Chemistry of Synthetic Dyes and Pigments,* American Chemical Society, Reinhold Publ. Co., New York 1955, pp. 537 – 541.
14. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 5,** Academic Press, New York – London, 1971, pp. 188 – 189.
15. Han Yong Hui, *J. Electrochem. Soc.* **130** (1983) no. 1, 36 – 40.

16. Am. Cyanamid, US 2 396 582, 1944 (H. Z. Lecher, M. Scalera, E. M. Hardy).
17. Ullmann, 4th ed., **16,** 209 – 213.
18. ICI, GB 633 480, 1946 (S. Coffey, D. A. W. Fairweather); ICI, GB 633 483, 1946 (S. Coffey, D. A. W. Fairweather, D. E. Hathway, F. H. Slinger).
19. Durand & Huguenin, CH 292 804, 1951; *Chem. Abstr.* **48** (II) 8552 b, 1954.
20. ICI, GB 630 459, 1947 (D. A. W. Fairweather).
21. Am. Cyanamid, US 2 403 226, 1944 (H. Z. Lecher, M. Scalera, E. M. Hardy).
22. F.I.A.T. Final Report 1313, vol. 3, p. 58.
23. H. Walbrecht, G. Olpeter, *Int. Text. Bull.* (1981) Jan., 27 – 28, 33 – 34, 73 – 74, 79 – 82.

Levulinic Acid → Oxocarboxylic Acids
Light-Emitting Diodes → Electroluminescent Materials and Devices
Light-Sensitive Polymers → Imaging Technology

Lignin

Bodo Saake, Institut für Holzchemie, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Hamburg, Germany

Ralph Lehnen, Institut für Holzchemie, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Hamburg, Germany

1. Occurrence and Functions 1
2. Structure and Biosynthesis 2
3. Physical Properties 6
4. Chemical Properties 7
4.1. Kraft Pulping 7
4.2. Sulfite Pulping 7
4.3. Other Pulping Processes 8
4.4. Pulp Bleaching 8
4.5. Yellowing Reactions 9
5. Commercial Lignins 9
6. Analysis 10
6.1. Detection and Quantification 11
6.2. Isolation 11
6.3. Spectroscopic Methods 11
6.4. Degradation Techniques 12
6.5. Molar Mass Determination 13
7. Uses 13
8. Toxicology 14
9. References 14

1. Occurrence and Functions

Lignin is one of the three major constituents of vascular plants, the other two being cellulose and hemicelluloses. The name lignin is derived from the Latin word *lignum* meaning wood. After cellulose, lignin is the most abundant natural (terrestrial) organic polymer. Its content is higher in softwoods (27–33 %) than in hardwoods (18–25 %) and grasses (17–24 %). The highest amounts of lignin (35–40 %) occur in compression wood on the lower part of branches and leaning stems of conifers [1, 2]. Lignin does not occur in algae, lichens, or mosses [3], whereas the "lignins" of bark differ in their structure from typical wood lignins [4].

Lignin is a randomly branched polyphenol, made up of phenylpropane (C_9) units, rendering it discernible from the other two major wood components by its UV absorption maximum at 280 nm. Figure 1 shows a typical UV microscopic imaging profile on a microtome cut of an individual spruce wood fiber [5]. Novel instruments can as well provide color-coded 2D or 3D plots, showing the lignin distribution over the cross-section of fibers [6]. The highest lignin concentration (≈ 70 %) is found between adjacent cell walls (middle lamella) and at the cell corners, while it is much lower (≈ 20 %) across the secondary wall. However, due to the much larger volume of the secondary wall, most lignin (≈ 80 %) is located in the secondary wall of the wood cells.

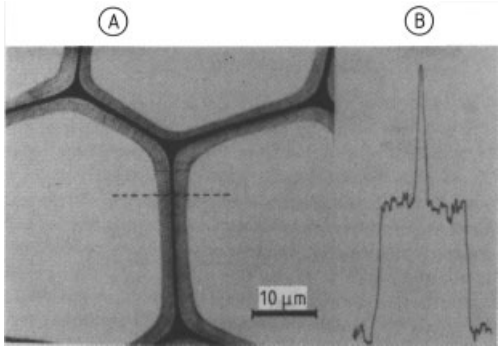

Figure 1. A) Black spruce earlywood
Cross section of tracheids of black spruce earlywood photographed in ultraviolet light ($\lambda = 240$ nm); B) Densitometer tracing across the tracheid wall showing the variation of lignin concentration along the dotted line [5]

In accordance with its distribution, lignin performs three important functions in the xylem tissue that are essential to the life of plants: Due

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_305.pub3

to its lipophilic character, lignin decreases the permeation of water across the cell walls, that consist of cellulose fibers and amorphous hemicelluloses, thus enabling the transport of aqueous solutions of nutrients and metabolites in the conducting xylem tissue. Secondly, lignin imparts rigidity to the cell walls and, in woody parts, together with hemicelluloses, functions as a binder between the cells generating a composite structure with outstanding strength and elasticity. Finally, lignified materials effectively resist attacks by microorganisms by impeding penetration of destructive enzymes into the cell walls.

The above mentioned functions of lignin are impressively demonstrated by the strength of the trunks of giant redwoods (*Sequoia*) in California, growing up to more than 100 m, supporting crown structures of several tons in weight for thousands of years, unparalleled by human constructions.

2. Structure and Biosynthesis

After cell growth has ceased, lignin is formed by a dehydrogenative polymerization of three *p*-hydroxycinnamyl alcohols (monolignols): i.e., *p*-coumaryl (**1**), coniferyl (**2**), and sinapyl alcohol (**3**) (Fig. 2), which are formed from D-glucose via shikimic acid [7]. It has to be mentioned that in lignin chemistry the designation of the carbon atoms does not follow the IUPAC nomenclature. The aromatic ring is numbered assigning the phenolic OH group to position 4. The propane side chain is labeled with Greek letters, starting from the benzylic carbon, termed α carbon. The phenylpropanoid monomers are oxidized to phenoxy radicals by hydrogen peroxide in the presence of peroxidase. The unpaired electron of the phenoxy radical is delocalized over the aromatic ring and the conjugated olefinic double bond of the side chain, the highest density of the unpaired electron being at the β-carbon atom, C-5, C-1, and O-4, as illustrated for coniferyl alcohol in Figure 3. Random recombination of two monomeric phenoxy radicals leads to dilignols, which are further oxidized to oligolignols and finally to the polymeric lignin. The most important structural units and their designation are depicted in Figure 4.

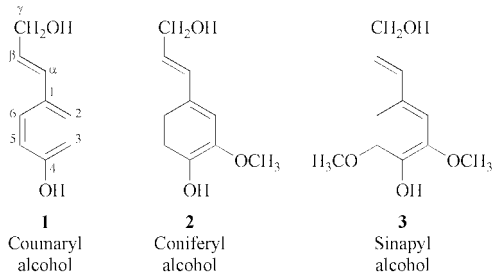

Figure 2. Monomeric lignin precursors

In order to prove this pathway Freudenberg and his group carried out tracer experiments with carbon-14 labeled D-glucose, phenylalanine, coniferyl alcohol, and D-coniferin [7]. In addition they performed in vitro model studies by dehydrogenation of coniferyl alcohol with peroxidase and hydrogen peroxide, obtaining polymers which were not identical but rather similar to lignin [7]. The resulting products were called dehydrogenation polymers (DHP) and have been intensively used as a lignin polymer model by many research groups.

The random polymerization by recombination of phenoxy radicals explains why lignin — in contrast to most other natural polymers such as proteins, polysaccharides, nucleic acids, and natural rubber — has an irregular structure and is optically inactive, though carbon atoms α and

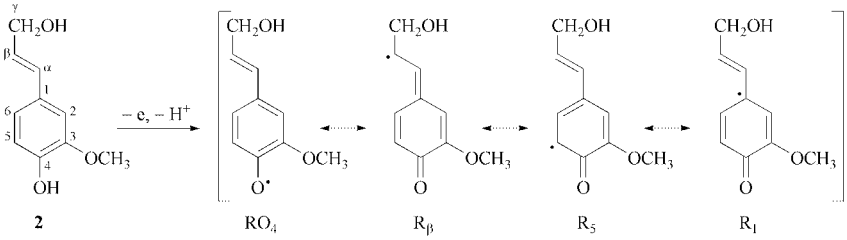

Figure 3. Dehydrogenation of coniferyl alcohol (**2**) yielding phenoxy radicals

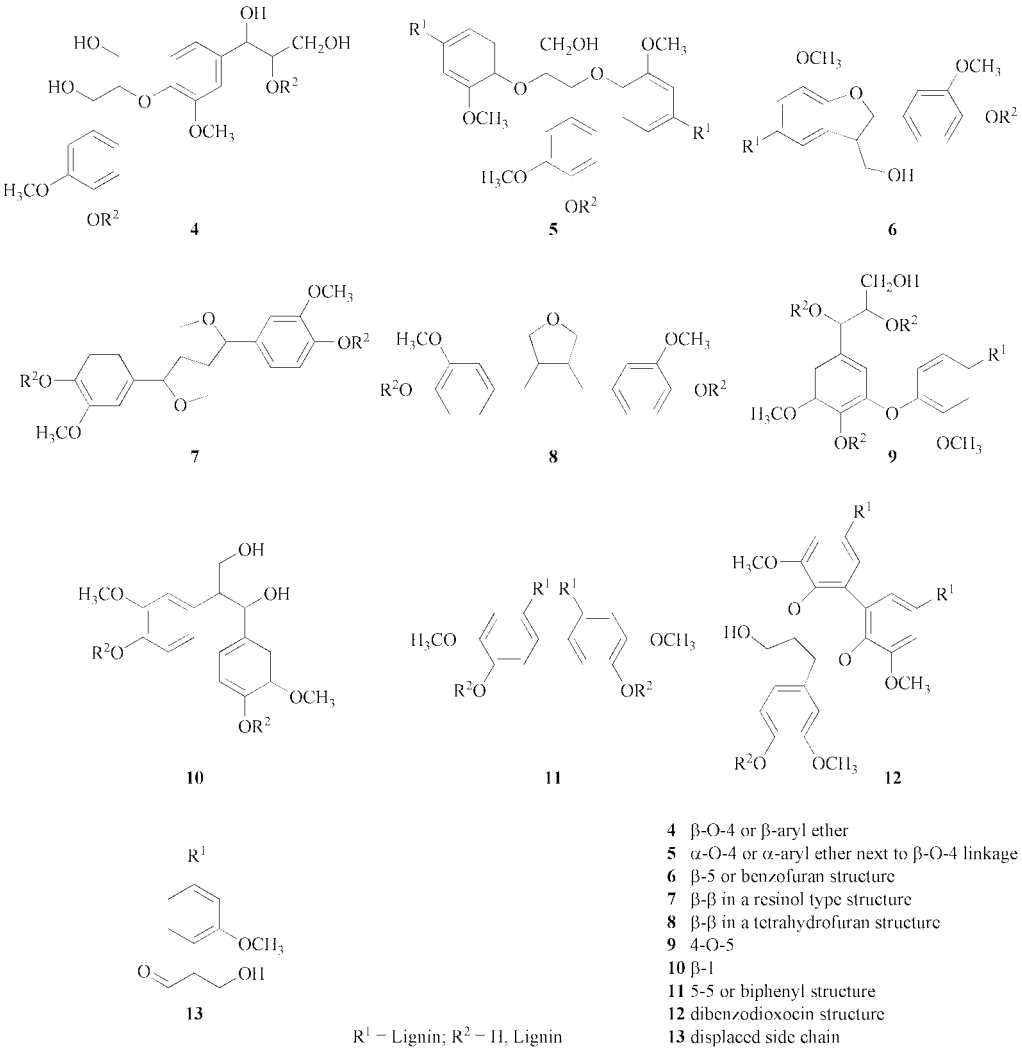

Figure 4. Most important structural units of lignin

β in the side chains of the phenylpropane structural units are asymmetric. However, it has to be kept in mind that the last polymerization step takes place within the secondary cell wall of differentiated wood cells in close contact with partly crystalline cellulose fibers, so that the aromatic nuclei in lignin may partly be aligned tangentially to the secondary wall [8]. Recently (in the early 2000s) a new hypothesis has evolved, postulating that the synthesis of lignin might be at least partly controlled by dirigent proteins. Such proteins are involved in the polymerization of lignans, a group of oligomeric products which belong to the extractives of lignocellulosic plants. This hypothesis is discussed rather intensely and summaries of the pros and cons have been published by the main protagonists [9, 10].

The random coupling reaction, which is still the most widely accepted concept, results in a three-dimensional, amorphous polymer without a regular structure or repeating unit. Accordingly, no definite lignin structure can be determined although several schemes have been proposed as a statistical structural model for various plant species. A first concept for spruce lignin was developed by Freudenberg [7] based on 18 C_9 units, of which 2.5 (units 2, 3, 5)

are coumaryl, one a syringyl (3,5-dimethoxy-4-hydroxyphenyl) (unit 16), and 14.5 guaiacyl (3-methoxy-4-hydroxyphenyl) units (Fig. 5). A structural scheme of beech lignin was proposed by NIMZ consisting of 25 structural units, of which six (5/6, 9/10, 24/25) are partly replaced by the dilignol units enclosed in brackets (Fig. 6). The concentrations of the latter units are below 4 %. The scheme shows a representative section of a ten to twenty times larger beech lignin molecule, in which ten different interunit bond types are randomly distributed. The relative amounts of interunit linkages for the two lignin structures are listed in Table 1 [11]. The data show as well that the calculated C_9 formulas are in good correlation with experimental data for isolated lignins. The most abundant interunit linkages in both schemes are the β-O-4 linkages. It is the concentration of these linkages in which lignins typically differ from other polyphenols occurring also in plant materials. During the lignification process the quinonemethide can also react with hydroxyl or carboxyl groups of polysaccharides forming covalent ether and ester bonds in the α-C position. These covalent bonds make it difficult to obtain lignin without carbohydrate impurities for both analytical and technical purposes. For this close association of lignin and carbohydrates the term lignin carbohydrate complex (LCC) has been coined [12].

Although the structures in Figures 5 and 6 do not include some new features such as the dibenzodioxocin structure (Fig. 3, **12**) discovered in the 1990s [15], they are still good models to describe the fundamental differences of softwood and hardwood lignins. According to the high proportion of guaiacyl units softwood lignins are called guaiacyl lignins (G-lignins), while hardwood lignins are termed guaiacyl–

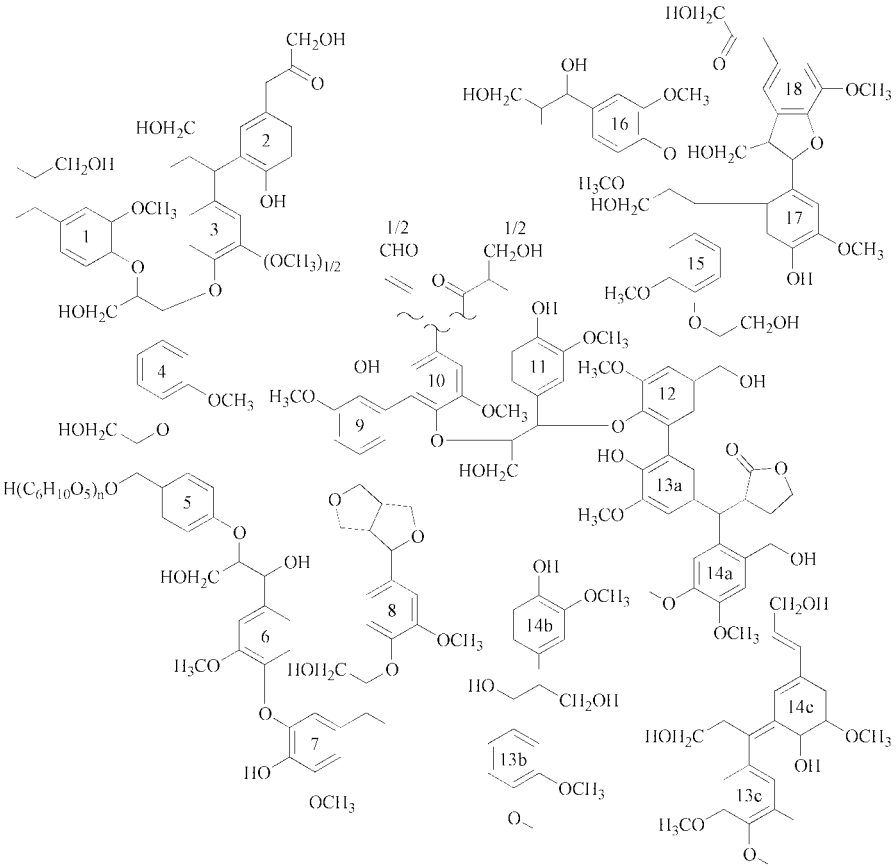

Figure 5. Constitution of spruce lignin after FREUDENBERG [7]

Table 1. Proportions (%) of interunit linkages, functional groups and elementary composition per C_9 unit of spruce and beech lignin according to structural schemes in Figures 5 and 6

Bond types, functional groups and elementary composition of C_9 unit	Spruce lignin	Beech lignin
β-O-4	39–48	32–37
α-O-4	11–16	28–32
β-5	6–10	8
β-β	7–10	6.4
5–5	7–9	2
4-O-5	6–7	2
β-1	2	16
α-5	7.2	
α-β		4
Aliphatic OH	92–98	88
Benzylic OH (α-OH)	18	4
Phenolic OH	29.4	16
Ketone groups	13.8	16
Aldehyde groups	2.8	4
Methoxy groups	91.7	136
C_9 formulae of structural schemes	$C_9H_{7.82}O_{2.4}(OCH_3)_{0.92}$ [7]	$C_9H_{7.16}O_{2.44}(OCH_3)_{1.36}$ [11]
C_9 formulae of MWLs*	$C_9H_{7.95}O_{2.4}(OCH_3)_{0.92}$ [13]	$C_9H_{7.10}O_{2.41}(OCH_3)_{1.36}$ [14]

* MWL (milled wood lignin), see Section 6.2

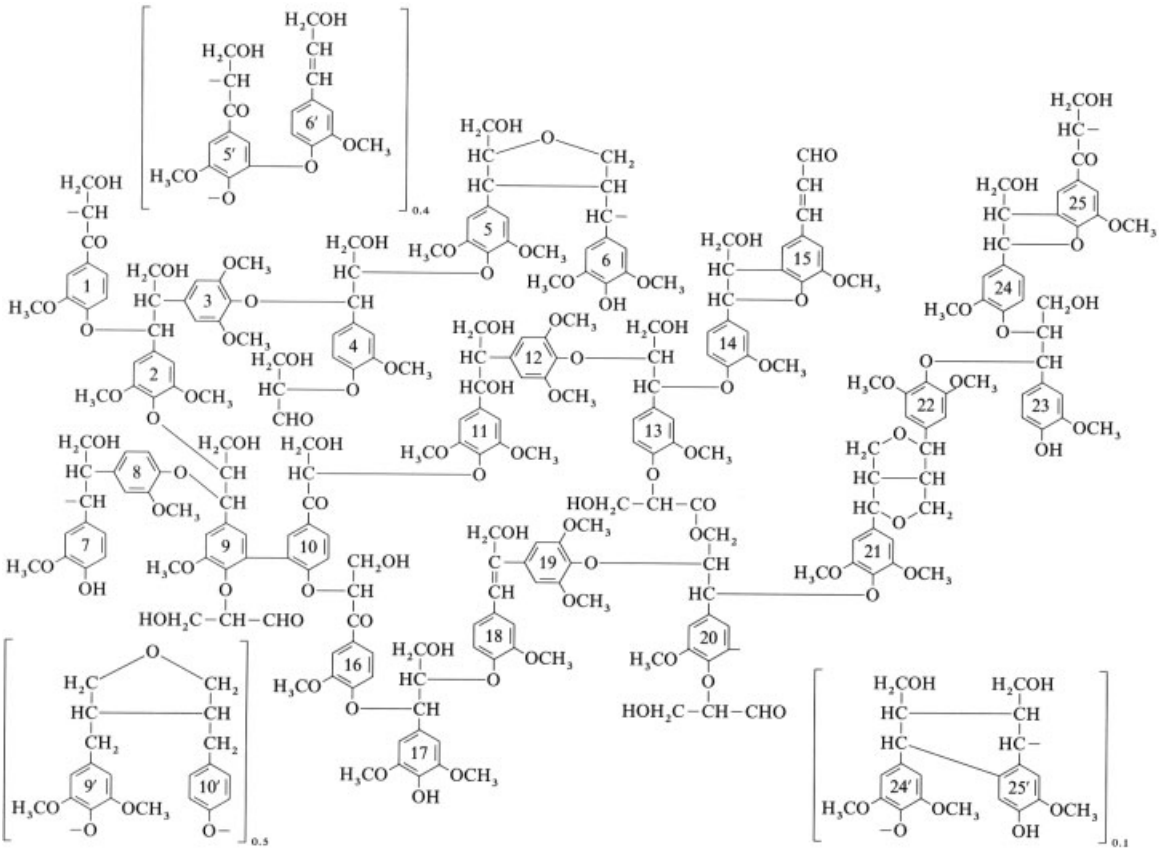

Figure 6. Constitutional scheme of beech lignin after Nimz [11]

syringyl lignins (GS-lignins). Grass lignins additionally contain *p*-hydroxyphenyl (coumaryl) units (GSH-lignins).

Accordingly, hardwood lignins have the highest methoxy values with 1.21–1.52 per C_9 unit, followed by softwood lignins with 0.80–1.01 and grass lignins with about 0.6–0.9 [16]. However, the low methoxy values of grass lignins may also be caused by substantial amounts of *p*-coumaric acid, linked to the γ-carbon atoms of grass lignins via ester bonds [17]. The methoxy group content of compression wood MWL from larch (*Larix leptolepis*) was found to be 0.2 per C_9 unit lower than that of normal wood MWL [18]. From the comparison of the carbon-13 NMR spectra of acetylated compression wood MWL with those of acetylated GSH- and H-DHPs (*p*-hydroxyphenyl dehydrogenation polymers), compression wood lignin was classified as a guaiacyl-*p*-hydroxyphenyl (GH)-lignin [19].

Variations in the methoxy group contents cause changes in the constitution of the lignins. As shown in Table 1 the proportions of α-O-4 and β-1 interunit linkages in beech lignin are significantly higher than those in spruce lignin, while the content of 5–5 and 4–O–5 linkages is lower, meaning that softwood lignins are more condensed than hardwood lignins and hardwoods may be more easily delignified than softwoods, agreeing with the differences in the chemical properties of hardwood and softwood lignins. Consequently, compression wood lignins have the highest level of condensation with the lowest solubility and degradability. A review with almost 700 references on the history, chemistry, and technology of lignin is given in [20].

3. Physical Properties

Solubility. As a branched polymer, native lignin is insoluble in all neutral solvents at room temperature. At temperatures above 100 °C or in alkaline solvents native lignin is partly dissolved as a result of degradation reactions. For isolated or technical lignin the solution properties vary depending on the isolation or production process and sample purity. Milled wood lignins are soluble in dioxane and acetone, containing 5–10 % water, as well as in dimethylformamide and pyridine. Lignosulfonates can be dissolved in water, while all isolated lignins are soluble in alkali. Most lignins can be solubilized in organic solvents such as dimethyl sulfoxide, dimethylformamide, dioxane, and pyridine at least for analytical purposes.

Molar Mass. The molar mass of lignin in situ is unknown and that of isolated lignins depends on the conditions of isolation. In general the molar mass determination of lignins with various methods show a large deviation (see Section 6.5). For spruce milled wood lignin an average molar mass (M) of 11 000 g/mol has been estimated by ultracentrifugation [21], corresponding to a weight-average degree of polymerization of approximately 60 C_9 units. For the most important commercial lignins from the kraft, soda, and sulfite pulping process molar masses in a range between 3 000 and 20 000 g/mol and polydispersities between 2 and 12 have been reported, while the intrinsic viscosities were in a range from 0.04 to 0.08 dL/g. The exponent a from the Mark–Houwink–Sakurada equation varies between 0.1 and 0.5, which is in the range between an Einstein sphere ($a = 0$) and a compact coil ($a = 0.5$) [22, 23].

Thermal Properties. Lignin is an amorphous thermoplastic polymer. The glass transition temperature T_g of lignin is strongly affected by moisture content and hydrogen bonding, a phenomenon of great importance for the wood processing industry. By dynamic mechanical thermal analysis the T_gs of native lignins in wood were calculated resulting in lower values in hardwoods (65-85 °C) than in softwoods (90-105 °C). For isolated lignins the T_g values are higher and show deviations according to the raw material, production process, and measurement procedure: MWL lignins: 110 to 160 °C, kraft lignins: 124–174 °C, steam explosion lignin: 113–139 °C [24]. The effect of moisture was demonstrated for an isolated periodate lignin which had a T_g of 195 °C when dry and a T_g of 90 °C when containing 27 % water [25]. Thermal properties are as well depending on molar mass. On fractionated kraft lignins the fraction of lowest molar mass (M 620 g/mol) had a T_g of only 32 °C while at highest molar mass (M 180 000 g/mol) T_g amounted to 173 °C. A similar tendency was observed for the temperature of

start of decomposition which increased with molar mass from 181 to 238 °C [26]. The calorific value of lignin depends on the purity of the sample. The dry matter of the black liquor from the kraft process has a value around 23.4 MJ/kg [1].

4. Chemical Properties

Besides ketone and aldehyde groups, the most reactive functional groups in lignin are the phenolic groups, the benzyl alcohol (α-OH) and noncyclic benzyl ether groups (α-O-4), which under acidic and alkaline conditions are prone to condensation reactions. From an industrial standpoint the reactions of lignin in pulping processes are most important since these reactions influence not only the pulping process itself but the structural features of technical lignins obtained as by products.

4.1. Kraft Pulping

The leading technical pulping process with sodium hydroxide and sodium sulfide is called kraft pulping. The word "kraft" originates from the German word "Kraft" (strength), meaning the strength of the obtained kraft pulps in comparison to that of sulfite pulps.

As a main reaction of the kraft process moieties with free phenolic groups are converted into quinonemethide groups (**15**), which add hydrogen sulfide ions at the α-carbon atoms (Fig. 7). The thus formed benzylthiolate anion (**16**) loses its β-phenolate anion in a neighboring displacement reaction. The units with free phenolic groups thus created may again form quinonemethides and add hydrogensulfide ions, if their α-carbon atoms are bonded to hydroxyl or noncyclic ether groups. At the high reaction temperature the sulfur from compound (**17**) is partly split off. Especially in the last stage of the kraft process carbon-carbon bonds can be formed between lignin units, a reaction which is referred to as "condensation". This results in structures which are very difficult to cleave in further process stages. Furthermore, the hydrogensulfide ions lead to demethylation reactions and subsequently to the formation of methyl mercaptan which resulted in former times in odor problems in the mills. An extensive review on the reactions and technology of pulping processes is given in [27].

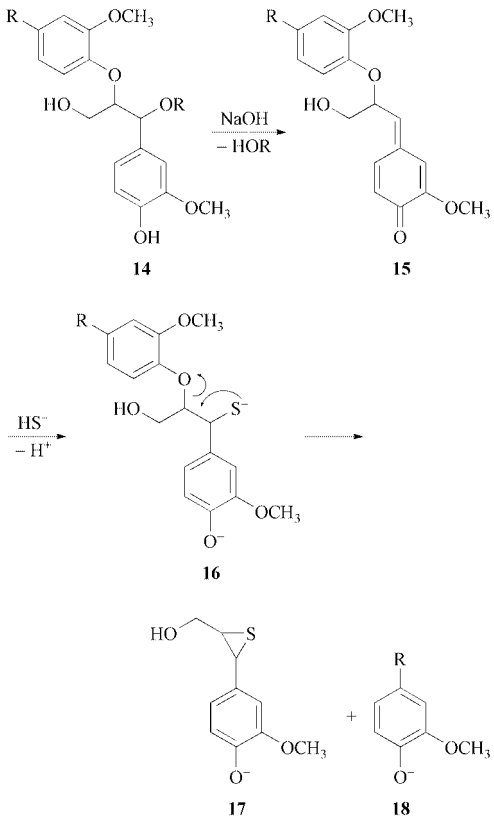

Figure 7. Mechanism of kraft cooking

4.2. Sulfite Pulping

In the technical sulfite pulping processes wood is usually reacted with calcium (pH ≈1-2) or magnesium sulfite (pH ≈3-5) at ca. 125 to 150 °C for 3 to 7 h. The three major reactions occurring simultaneously are sulfonation, hydrolysis, and condensation. The dissolution of lignin in aqueous solution is mainly caused by the introduction of hydrophilic sulfonic acid groups at the α-carbon atoms (Fig. 8). In the *sulfurous acid* process at pH 1–1.6 and in the *acid sulfite* process at pH 1.8–3.1, benzyl alcohol and benzyl ether groups in lignin form benzylium ions (**20**), that preferably add nucleophilic sulfite ions, but to some extent may also lead to condensation products (**21**) [28, 29]. The degree of condensation may be less in the *bisulfite process* at pH

4.5 and the *bisulfite-sulfite* process at pH 7. The β-O-4 bond is rather stable under acidic conditions. Some sulfite processes operate under neutral or alkaline conditions using sodium or ammonia as base. In these processes β-O-4 bonds can be cleaved as well depending on the reaction conditions. The spent sulfite liquor, besides degraded hemicelluloses contains lignosulfonic acids (**22**).

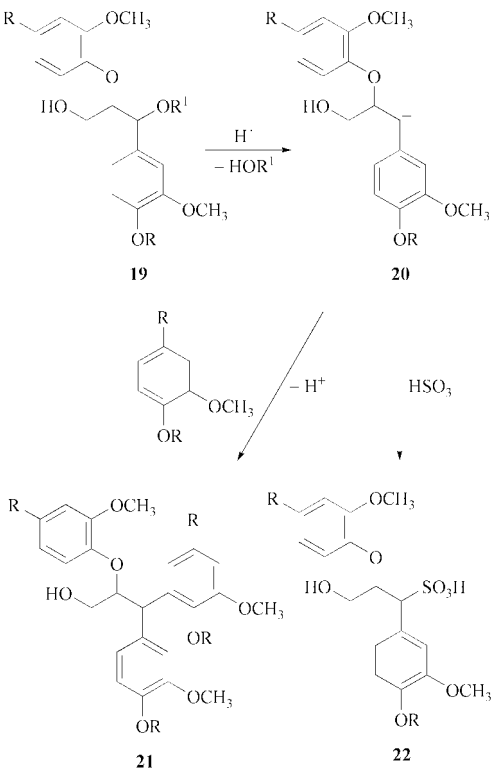

Figure 8. Mechanism of sulfite cooking

4.3. Other Pulping Processes

Especially in countries with small wood resources the soda process is used for the production of pulp from annual plants applying only sodium hydroxide as a reaction chemical. In this process only α-aryl ether bonds in phenolic lignin units are effectively cleaved [29]. The addition of the redox catalyst system anthrahydroquinone/anthraquinone can improve the delignification when raw materials with higher lignin content are processed. Anthrahydroquinone can catalyze the cleavage of lignin in a similar manner to sulfide, improving especially the cleavage of β-O-4 linkages [30]. It can be applied in all alkaline processes such as kraft or alkaline sulfite pulping.

A large number of pulping processes based on organic solvents have been investigated over the last decades and were summarized under the term "organosolv" processes. Some were based on organic solvents (especially methanol and ethanol) with and without addition of catalysts while others were applying organic acids (mainly formic and acetic acid) [31, 32]. Up to now none of these processes has been implemented in industry. Nevertheless, in many papers on new applications of lignin still organosolv lignins from former pilot plant trials are used as a raw material, especially when a sulfur-free lignin is required.

Steam explosion or steam refining is a process route for the production of fibers from wood or straw. These processes are often catalyzed autohydrolytically by organic acids liberated from the hemicelluloses of hardwoods or straw. During steam explosion a part of the lignin ($\approx$10–15 %) is rendered water-soluble while the rest can be extracted with dilute sodium hydroxide solution. During the steam treatment an acid-catalyzed hydrolysis of aryl ether linkages takes place while on the other hand homolytic cleavages occur. At high reaction severity a recondensation of lignin can occur as well. Accordingly steam explosion lignins have higher molar masses than sulfite, kraft, or organosolv lignins. Steam explosion lignins might become interesting in the future, because this process is the first step for saccharification of lignocelluloses for glucose or ethanol production, a process route which is currently under discussion as a source for bioenergy and fuel [31, 33] (the whole issue of [33] deals with steam explosion.)

4.4. Pulp Bleaching

Unbleached pulps still contains between 1.5 and 6 % of lignin. Since the residual lignin is intensively colored it has to be removed in order to produce pulps with high brightness and brightness stability. Previously chlorine and chlorine dioxide were major bleaching chemicals lead-

ing to toxic chlorinated phenols in the bleaching effluents (measured as adsorbable organic halogen (AOX). In order to reduce these effluents, chlorine (C) was almost completely and chlorine dioxide (D) partly replaced by alkaline oxygen (O), hydrogen peroxide (P), and ozone (Z), leading to *totally chlorine free* (TCF) or *elemental chlorine free* (ECF) bleached pulps. In the production of ECF pulps chlorine dioxide is still applied, but no gaseous chlorine. In the beginning of these developments it was often assumed that TCF bleaching should be the long-term goal for an environmental-friendly industry. However, it has been shown that a modern ECF sequence can be operated with very low AOX levels and accordingly ECF is now the dominating technology for modern pulp mills [34].

4.5. Yellowing Reactions

In the presence of alkali and oxygen or in daylight lignin shows an intense discoloring, which is observed by the yellowing reaction of paper, paper products, and wood. A major reaction pathway is the absorption of ultraviolet light by the α carbonyl group leading to a phenoxy radical which reacts subsequently with oxygen to form quinoid chromophores (Fig. 9).

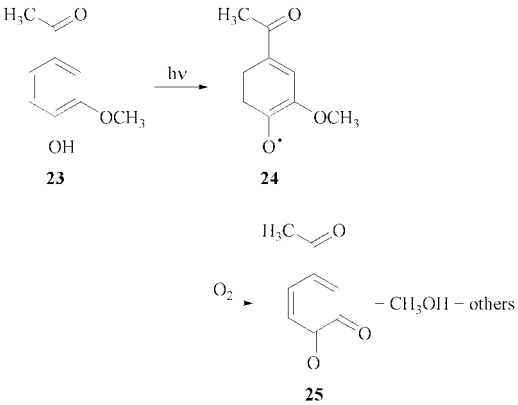

Figure 9. Main reaction pathway of photo yellowing

The photoyellowing is a rather complex process which was recently summarized in [35]. For applications of lignin the sensitivity to UV light has to be considered being a potential problem for product stability or a positive feature when UV absorbing properties are required.

5. Commercial Lignins

Up to now commercial lignins are exclusively obtained as byproducts from the chemical pulping industry. The kraft process is the dominating technology with about 89 % of the total production capacity while in 2000 the share of the sulfite process had declined to only 3.7 %. Nevertheless, lignosulfonates from sulfite processes are still of major importance for the application of lignin as industrial products. About 10 % of the total pulp production is based on nonwood plant materials like bast and leaf fibers, straw, sugar cane bagasse, or bamboo. These materials are easier to delignify and are mainly produced by the soda process. They are especially important raw materials in countries with small wood resources like India and China [36]. In former times these lignins were not commercially available. Recently a plant for lignin recovery from nonwood soda pulping has been installed which claims an annual production capacity of 10 000 t sulfur-free lignin [37] and accordingly soda lignins might become an alternative in the future. In 1998 some 50×10^6 t of lignin were produced in the western hemisphere as a byproduct of pulp, less than 2 % of this amount was isolated and sold. Kraft lignin, making up some 95 % of the totally produced lignin, is predominantly burnt to cover the energy demands of kraft mills and the recovery of caustic soda in the mill. Most of the commercially used kraft lignin is sulfonated to water-soluble lignosulfonates. Excluding the former Soviet Union, lignosulfonate production capacity in 1998 was estimated to be about 975 000 t/a with some new capacities under construction in South Africa [38].

The lignins are contained in the pulping liquor as a mixture with other wood degradation products with strongly fluctuating composition depending on the process and the raw material. Typical compositions of pulping liquors from hardwoods and softwoods obtained by the sulfite and kraft process are summarized in Tables 2 and 3 [39]. Kraft lignins are preferably obtained by precipitation from “black liquors” with mineral acids or carbon dioxide, whereas crude lignosulfonates are partly used directly as “spent

sulfite liquors". The isolation of lignosulfonates can be achieved by addition of excess lime, by treatment with a long-chain alkylamine and extraction, or by ultrafiltration. In order to produce high-purity lignosulfonates a chemical destruction of sugars, or fermentation treatments can be included into the refinement procedure [38]. The sulfur contents of lignosulfonates (4–8 %) are higher than those of kraft lignins (1–1.5 %). Due to the sulfonate groups the lignosulfonates show good solubility in water over the entire pH range but are insoluble in many organic solvents. Kraft lignins are not soluble in water but in alkali at pH > 10.5. They are as well soluble in many organic solvents. The molar masses of lignins are difficult to determine in a reliable way and often results in controversial discussion of data. Furthermore molar masses depend strongly on the pulping conditions. In most cases the molar masses of lignosulfonates are higher compared to kraft lignin and even values up to 60 000 g/mol have been reported [40], although lower data are found for most samples [22, 23]. For Kraft lignins the data are mostly below 10 000 g/mol [22, 23, 39].

Table 2. Example for the composition of spent sulfite liquors from softwood and hardwood [39]

Component	Percentage of total solids	
	Softwood	Hardwood
Lignosulfonate	55	42
Hexose sugars	14	5
Pentose sugars	6	20
Acetic and formic acid	4	9
Resin and extractives	2	1
Ash	10	10

Table 3. Example for the composition of kraft black liquors from softwood and hardwood [39]

Component	Percentage of total solids	
	Softwood	Hardwood
Kraft lignin	45	38
Xyloisosaccharinic acids	1	5
Glucoisosaccharinic acid	14	4
Hydroxy acid	7	15
Formic acid	6	6
Acetic acid	4	14
Resin and fatty acids	7	6
Turpentine	1	
Others	15	12

Kraft lignins contain larger amounts of phenolic hydroxyl, carboxyl, and catechol groups than lignosulfonates and are more likely to possess some unsaturated side chain structures. For certain application lignins can be further modified in order to improve the dispersing, complexing, or binding properties. These treatments can be sulfonation, sulfoalkylation, desulfonation, oxidation, carboxylation, amination, crosslinking, graft polymerization, and various combinations of the previous methods. Most of the kraft lignins are subjected to sulfonation either with sodium sulfite at 150–200 °C or oxidative with oxygen and sulfite. They can also be sulfomethylated with sulfide and formaldehyde at temperatures around 100 °C. Lignosulfonates are mostly obtained as magnesium, calcium, ammonium, or sodium salts [38 – 41].

Trade Names: A large variety of trade names and products exists. However, most of these products are marketed by two companies, Mead-Westvaco Speciality Chemicals and Borregaard-LignoTech, where the first company is dominating the kraft lignin and the latter the lignosulfonate market. Some trade names are listed below:

Unsulfonated kraft lignin	Indulin
Sulfonated kraft lignin	Polyfon, Reax
Lignosulfonates	Ameri-Bond, Borresperse, Borresol, Collex, Diwatex, Dustex, Dynasperse, Kelig, LignoBond, LignoSol, Marasperse, Norlig, Ufoxane, Ultrazine, Wafex, Wanin, Zewa
Soda lignin	Protobind, Biosurfact

6. Analysis

Due to its complicated structure the analysis of lignin is very difficult. For the pulp industry the analysis of lignin contents in pulps and raw materials is specified, while for technical lignins no well-defined industrial standards exist. For these reasons it is difficult to obtain reliable, detailed information on technical products. Due to this situation the following information can provide only a small insight into the field of lignin analysis. A review of most classical techniques for lignin analysis is given in [42].

6.1. Detection and Quantification

The easiest methods to detect lignin in plant materials are color reactions. However, those tests can be perturbed by some polyphenolic plant extractives. In the *Wiesner reaction* the material is treated with phloroglucinol in hydrochloric acid, leading to a purple conjugated quinone [43], formed between phloroglucinol and cinnamyl aldehyde groups in lignin. The reaction is positive with all lignins, but is weak with hardwood lignins with high contents of syringyl units.

For the *Mäule test* [44] lignified material is treated with a 1 % potassium permanganate solution, followed by washing, a treatment with 12 % hydrochloric acid, and a final moistening with aqueous ammonia. With hardwoods an intensive purple-red color is produced, while with softwood lignins the color is an indefinite brownish shade. The chromophores are probably chlorinated quinones.

The quantitative determination of lignin in plants most commonly used, the *Klason procedure*, is based on the observation that cellulose and hemicelluloses are hydrolyzed by concentrated sulfuric acid to soluble sugars while lignin is condensed to an insoluble cross-linked polymer which is recovered by filtration on a filter crucible and can be gravimetrically determined. This procedure is a standard test for raw materials in the pulp industry [45].

Another method for the quantitative determination of lignins in wood uses the solubility of wood in a mixture of acetyl bromide and acetic acid [46]. After removal of the reagent, the absorbance of the resulting solution is measured at 280 nm. A detailed overview on all analytical methods for the determination of lignin in raw materials as well as in pulps is given in [47]. A critical comparison of several methods including not only wood and straw but various agricultural residues was presented in [48].

6.2. Isolation

The complete isolation of unchanged lignin from wood is not possible, due to its intimate merging with cellulose and hemicelluloses in the secondary walls of wood cells. Lignin has to be degraded before extraction and accordingly the structure depends on the conditions of isolation. The most common procedure for structural and analytical purposes is production of *milled wood lignin* (MWL), also called *Björkman lignin* [49]. In the isolation procedure preextracted, dry wood meal is milled for, e.g., 30 days in a vibratory ball mill under a nonswelling solvent (toluene). In the subsequent extraction step with dioxane – water (9:1) up to 50 % of the lignin can be isolated from conifer wood which can be further purified in order to reduce carbohydrate impurities. Further methods for the isolation of lignins from pulps and pulping liquors are summarized in [42].

Recently it has been proposed to dissolve ball-milled cell wall material in dimethyl sulfoxide and tetrabutylammonium fluoride or *N*-methylimidazole. The product mixture of lignin and polysaccharides can be directly acetylated in the solution and allows further analyses, e.g., by 2D NMR techniques. Although this is no isolation procedure in the strict sense it can serve the same purposes for analytical techniques. A major advantage is that a higher lignin yield than in traditional isolation procedures is obtained [50].

For pulp fibers a method has been developed which is based on the enzymatic saccharification of polysaccharides and allows recovering of the undissolved lignin residue from the aqueous solution [51].

6.3. Spectroscopic Methods

UV Spectroscopy. The aromatic structure of lignin gives rise to a strong absorption maximum at 280 nm, while wood carbohydrates are transparent in the near-ultraviolet range. As, according to the Lambert – Beer law, the absorption of UV light by dissolved lignin is proportional to its concentration, UV absorption can be used for quantitative determination of lignin in solution or as a detection method in liquid chromatography [52]. In alkaline solution, the UV spectrum of lignin shows a bathochromic shift and hyperchromic effect, due to the formation of phenolate anions. An alkaline ionization difference spectrum ($\Delta\epsilon_i$) is obtained by subtracting the UV spectrum of the neutral solution from that of the alkaline solution. $\Delta\epsilon_i$ -Spectra are often used to obtain information on the amount of phenolic

hydroxyl groups. These spectra show as well specific absorption peaks for conjugated (carbonyl and olefinic) groups in lignin, which thus can be determined semiquantitatively [53, 54]. The UV absorption can be used also in the solid state for UV microscopy on microtome cuts (Fig. 1).

FTIR spectra of lignins are usually measured on finely ground solid samples dispersed in potassium bromide pellets. In solution the lignin spectra are influenced by strong interfering absorption bands of the solvents, i.e., semiquantitative work can be carried out better with KBr pellets using controlled pressing conditions and weight ratio of lignin to potassium bromide. The assignment of infrared absorption bands of lignins has been reviewed in [55] and [56].

NMR Spectroscopy. Of all physical methods, NMR spectroscopy provides by far the most complete information on the chemical structure of lignins. The assignment of some 40 signals in the ^{1}H broad-band decoupled carbon-13 NMR spectra of spruce and beech MWLs and DHPs as well as their acetates has been achieved in the early 1970s by comparison with the spectra of more than 50 lignin model compounds [57]. By now ^{1}H and ^{13}C shifts for a variety of model compounds have been determined and a large database is provided by the U.S. Dairy Forage Research Center in Madison, Wisconsin. This database is continuously updated and available free of charge in the internet [58]. Although solid-state NMR (^{13}C cross-polarization magic angle spinning (CP-MAS) NMR) spectra of lignin have been frequently recorded [59], the solution NMR techniques have by far the higher importance due to their higher resolution. One-dimensional high resolution ^{1}H or ^{13}C spectra of lignin can be obtained directly or after acetylation. The acetylation can improve the resolution of spectra and provide as well information on the primary, secondary, and phenolic OH groups from the carbonyl signals of the corresponding esters. For a good signal-to-noise ratio extreme long measurement times are required for ^{13}C spectra [60]. Here the short measurement times of ^{1}H spectra are advantageously, although the information is more limited. Besides the methoxy signals, a differentiation of phenolic and aliphatic hydroxyl groups, and information on the protons in aromatic and various aliphatic structures can be obtained. As an alternative ^{31}P and ^{19}F NMR spectra can be recorded with good sensitivity and short measurement times after labeling reactions with P and F containing reagents. The resulting spectra cannot provide information on the entire lignin structure, but on different hydroxyl, aldehyde, keto and quinone groups [61, 62].

The large number of signals in the carbon and proton spectra results in an intense overlapping of signals. This can be resolved by heteronuclear single- and multiple-bond shift correlation spectra which provide 2D plots for proton and carbon shifts. Under normal operation conditions 2D spectra are not suitable for quantification. However, various approaches have been developed in the early 2000s in order to enable a quantitative determination of lignin either by a calibration of the 2D spectra [63] or by a remodeling of the pulse sequences [64]. Also 3D spectra of lignin have been recorded [65, 66].

6.4. Degradation Techniques

Many methods for the determination of intermonomer bonds of lignin are based on analytical degradation procedures. The degradation by alkaline solutions of potassium permanganate was used intensively by Freudenberg's group in the 1950s and 1960s and further modified by Miksche and coworkers. Accordingly different variations of the method exist and are reviewed in [61]. The protocol consists of several reaction steps and result in mono- and dimeric aromatic carboxylic methyl esters, which can be analyzed by gas chromatography. This shows not only the proportion of hydroxyphenyl, guaiacyl, and syringyl units but allow calculation of the abundance of different lignin substructures on a molar basis. The procedure is applicable to all kind of lignins but has the disadvantage of a rather complicated multi-step protocol which requires an elaborated data treatment in order to conclude on the original lignin structure [68].

The thioacidolysis of lignin encompasses mainly the cleavage of β-O-4 linkages with boron trifluoride etherate and ethanethiol in anhydrous media. The monomeric degradation products can be analyzed by gas chromatography and normally between 1000 to 3000 µmol/g

of mono- and dimeric degradation products are obtained. The protocol is accordingly simpler and allows an evaluation of the amount of alkyl-aryl ethers in lignin samples. A disadvantage is the strong odor of the reagents and the fact that the method is less informative on degraded samples, which is sometimes the case with some technical lignins [68, 69].

Derivatization followed by reductive cleavage (DFRC) was introduced as a new degradation method for lignin analysis [70]. Lignin is derivatized with acetyl bromide, followed by a reductive cleavage with zinc dust in a mixture of dioxane, acetic acid, and water (5:4:1). The resulting products are then acetylated and analyzed by gas chromatography. In the DFRC procedure hydroxyphenyl, guaiacyl, and syringyl units linked by α- and β-aryl ethers are cleaved and transferred into 4-acetocinnamyl acetate, coniferyl diacetate, and sinapyl diacetate.

The pyrolysis of lignin or lignocellulosic materials, i.e., the thermal degradation in the absence of oxygen, leads to the formation of a large variety of volatile components, which contain information on the original structure. In analytical pyrolysis this degradation technique is combined online or offline with other analytical techniques and most importantly with mass spectrometry (Py-MS) or gas chromatography and mass spectrometry Py-GC/MS. In the GC or GC/MS analysis of pyrolysis degradation products more than 100 peaks are found. Accordingly, the evaluation requires a large database and is based on statistical evaluation methods. Information on various pyrolysis techniques is found, e.g., in [71, 72]. A general review on MS techniques in combination with all relevant lignin degradation techniques can be found in [73].

6.5. Molar Mass Determination

Size-exclusion chromatography (SEC) is the most common method to determine molar masses of lignins. Frequently, lignins were acetylated followed by SEC in THF using columns based on polystyrene – divinylbenzene [74]. However, in various round robin tests carried out by the World Energy Association and the European network Eurolignin this approach resulted in a tremendous variation of the calculated molar masses [75, 76]. This problem might be partly due to aggregation phenomena and partly due to interaction with the column material. Furthermore, the acetylation is not suitable for lignins containing carbohydrate impurities and for lignosulfonates, the most important technical lignins. Lignosulfonate samples have been recently investigated using a complex mixture of water, DMSO, sodium phosphate buffer, and sodium dodecyl sulfate [40]. A universal eluent system applicable to all technical lignins is sodium hydroxide, however, the handling is not so easy because most commercial prepacked columns for HPSEC cannot withstand the high alkalinity [74]. Recently DMSO – water – lithium bromide and dimethylacetamide – lithium chloride were applied as eluents, showing a good agreement with alkaline SEC [23]. The determination of absolute molar masses by light scattering is complicated by the fluorescence of lignin, while viscosimetric methods suffer from the low overall viscosity of the samples. All methods are affected by the impurities of technical samples especially by carbohydrates and ash. Recently matrix-assisted laser desorption ionization/time of flight (MALDI/TOF) was applied to lignins after a fractionation of the samples on SEC columns in order to obtain fractions with narrow polydispersity [77].

7. Uses

Most applications of lignin and lignosulfonates are based on their dispersing, binding, complexing, and emulsion-stabilizing properties. About 50 % of all *lignosulfonates* produced worldwide are used for concrete mixtures, usually in the form of calcium or sodium salts. Addition of 0.1-0.3 % lignosulfonates to cement retards the setting or hydration of concrete. The second most important application is as a binder for animal feed pellets, where mainly calcium and ammonium salts are applied in order to improve pellet durability and abrasive resistance. A maximum dosage of 4 % is possible in finished pellets. Primarily chrome and ferrochrome salts of lignosulfonates function in oil well drilling muds as mud thinners, clay conditioners, viscosity-control agents, and fluid-loss additives. Mud

Table 4. Data on the toxicity of different lignosulfonates [38]

Property	Calcium lignosulfonate	Sodium lignosulfonate	Ammonium lignosulfonate
Acute toxicity LD_{50}, mg/kg	> 2 000	> 10 000	> 10 000
Eye irritant	no	no	no
Skin irritant	no	no	no
Fish toxicity LC_{50}, mg/L	> 1 000	> 1 000	> 1 000
Bacteria toxicity EC_{10}, mg/L	5 000	5 000	343

systems conditioned with 0.2–0.5 % lignosulfonates, applied in the crude oil industry, perform well at high pressures and at temperatures of up to 175 °C. Another major market, especially for crude lignosulfonate spent liquors, is the dust control application for stabilizing insurfaced roads. In addition to these bulk applications a variety of speciality markets exist. Lignosulfonates are used for the granulation, complexation, or encapsulation of pesticides and as additives for gypsum boards in order to disperse the stucco. Lignosulfonates are used as dispersant in water-based paints and inks. Lignosulfonates and spent sulfite liquors are used for water treatment plants and paper machines to reduce deposits and slime formation or to complex metals such as zinc in cooling-water cycles. They are used in industrial cleaner formulations, for the complexation of nutrients in soil stabilization, for dyes or as additives in the brick industry. Modified lignosulfonates are additives in lead–acid batteries, which extend the service life of the product significantly. As further future uses the stabilization of enzyme formulations, biocide neutralization, and applications exploiting the antiviral and the chelating properties have been targeted [38].

Borregaard LignoTech still produces vanillin from softwood lignosulfonates, while all other producers closed down their production due to lower costs for vanillin production from petrochemical feedstocks. This was also the reason for stopping the production of dimethyl sulfoxide (DMSO) from kraft lignin by heating black liquor with sulfur to above 200 °C to obtain dimethyl sulfide (DMS), which upon oxidation with dinitrogen tetroxide (N_2O_4) yielded DMSO [78].

Up to now, much work has been carried out on new utilizations of lignin (see for instance [79]). Only few applications are economically feasible, taking into account that lignin is currently burnt, covering not only the energy demand of the pulp mills but providing a significant surplus of energy. Sometimes the argument occurs that the utilization of lignin as a chemical feedstock will increase at higher oil prices. However, one has to keep in mind that the economic burning value of lignin is increasing as well when the overall energy costs are rising. Accordingly, the changes for implementing new applications are higher when lignin is used not as a bulk chemical but as a specialty product using its specific properties.

8. Toxicology

Lignosulfonates are considered as nontoxic at the typical use levels and they are approved, e.g., by the U.S. Food and Drug Administration for a variety of applications in food packaging and food production. Examples are the use as a boiler-water additive for the production of steam, which comes into contact with food (21 CFR 173.310), or the approval of calcium lignosulfonates as a "specific usage additive" for the regulations on food additives for direct addition to food for human consumption (21 CFR 172.172). Some data for various lignosulfonates are listed in Table 4 [38]. However, one should keep in mind that the properties of a specific product are not only influenced by its lignosulfonate component but by its impurities.

9. References

1. D. Fengel, G. Wegener: *Wood: Chemistry, Ultrastructure, Reactions*, DeGruyter, Berlin-New York 1984.
2. K. V. Sarkanen, C. H. Ludwig (eds.): *Lignin, Occurrence, Formation, Structure*, Wiley-Interscience, New York 1971, pp. 45, 66, 75.
3. H. H. Nimz, R. Tutschek, *Holzforschung* **31** (1977) 101–106.
4. W. Zimmermann, H. H. Nimz, E. Seemüller, *Holzforschung* **39** (1985), 45–49;

C. Pascoal Neto, N. Cordeiro, A. Seca, F. Domingues, A. Gandini, D. Robert, *Holzforschung* **50** (1996) 563–568.
5. B. J. Fergus, A. R. Prokter, J. A. N. Scott, D. A. I. Goring, *Wood Sci. Technol.* **3** (1969) 117–138.
6. G. Koch, J. Puls, J. Bauch, *Holzforschung* **57** (2003) 339.
7. K. Freudenberg, A. C. Neish: *Constitution and Biosynthesis of Lignin*, Springer, Berlin-Heidelberg-New York 1968.
8. R. H. Atalla, *Wood Chem. Technol.* **7** (1987) 115.
9. L. B. Davin, N. G. Lewis, *Curr. Opin. Biotechnol.* **16** (2005) 407.
10. J. Ralph et al., *Phytochem. Rev.* **3** (2004) 29.
11. H. H. Nimz, *Angew. Chem.* **86** (1974), 336; *Angew. Chem. Int. Ed. Engl.* **13** (1974) 313.
12. T. Koshijima, T. Watanabe: *Association between Lignin and Carbohydrates in Wood and other Plant Tissues*, Springer, Berlin – Heidelberg 2003.
13. K. Freudenberg, J. M. Harkin, *Holzforschung* **18** (1964) 166.
14. K. Freudenberg, G. S. Sidhu, *Holzforschung* **15** (1961) 3.
15. P. Kanhunen, P. Rummakko, J. Sipila, G. Brunow, I. Kilpelainen *Tetrahatron Lett.* **36** (1995) 169–170.
16. K. V. Sarkanen, C. H. Ludwig: *Lignin, Occurrence, Formation, Structure*, Wiley-Interscience, New York 1971, pp. 64, 56, 75.
17. T. Higuchi, I. Kawamura, *Holzforschung* **20** (1966) 16.
18. S. Yasuda, A. Sakakibara, *Mokuzai Gakk.* **21** (1975) 363.
19. H. H. Nimz, D. Robert, O. Faix, M. Nemr, *Holzforschung* **35** (1981) 16.
20. J. L. McCarthy, A. Islam in W. G. Glasser, R. A. Northey, T. P. Schultz (eds.): "Lignin: Historical, Biological, and Materials Perspectives", *ACS Symp. Ser.* **742** (1999) 2–99.
21. A. Björkman, B. Person, *Svensk Papperstidn.* **60** (1957) 285.
22. W. G. Glasser, V. Davé, C. E. Frazier, *J. Wood Chem. Technol.* **13** (1993) 545.
23. O. Ringena, S. Lebioda, R. Lehnen, B. Saake, *J. Chromatogr. A* **1102** (2006) 154.
24. W. G. Glasser in W. G. Glasser, R. A. Northey, T. P. Schultz (eds.): "Lignin: Historical, Biological, and Materials Perspectives", *ACS Symp. Ser.* **742** (1999) 216–238.
25. D. A. I. Goring, in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignin, Occurrence, Formation, Structure*, Wiley-Interscience, New York 1971, p. 732.
26. H. Yoshida, R. Mörck, K. P. Kringstad, H. Hatakeyama, *Holzforschung* **41** (1987) 171.
27. H. Sixta, A. Potthast, A. W. Krotscheck in H. Sixta (ed.) *Handbook of Pulp*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 109–472.
28. J. Gierer, *Svensk Papperstidn.* **18** (1970) 571.
29. J. Gierer, *Wood Sci. Technol.* **14** (1980) 241.
30. J. Gierer, O. Lindberg, I. Noren, *Holzforschung* **33** (1979) 213.
31. R. A. Young, M. Akhtar (eds.): *Environmental Friendly Technologies for the Pulp and Paper Industry*, John Wiley & Sons, New York 1998, pp. 3–256.
32. "Lignin", in *Ullmann's*, 7th ed., Wiley-VCH Verlagsgesellschaft, Weinheim 2002.
33. G. Glasser, R. S. Wright, *Biomass Bioenergy* **14** (1998) 219.
34. H. Sixta (ed.): *Handbook of Pulp*, Wiley-VCH, Weinheim 2006.
35. C. Heitner, J. C. Scaiano (eds.): "Photochemistry of lignocellulosic materials", *ACS Symp. Ser.* **531** (1993) 1–223.
36. H. Sixta (ed.) *Handbook of Pulp*, vol. 1, Wiley-VCH, Weinheim 2006, pp. 8–9.
37. http://www.asianlignin.com.
38. J. D. Gargulak, S. E. Lebo in W. G. Glasser, R. A. Northey, T. P. Schultz (eds.): "Lignin: Historical, Biological, and Materials Perspectives", *ACS Symp. Ser.* **742** (1999) 304.
39. *Ullmann's*, 5th ed., A15, VCH Verlagsgesellschaft, Weinheim 1990.
40. G. E. Fredheim, S. M. Braaten, B. E. Christensen, *J. Wood Chem. Technol.* **3** (2003) 197.
41. W. G. Glasser, S. Sarkanen (eds.): "Lignin Properties and Materials", *ACS Symp. Ser.* **397** (1989).
42. S. Y. Lin, C. W. Dence (eds.): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992.
43. E. Adler, K. J. Björkquist, S. Häggroth, *Acta Chem. Scand.* **2** (1948) 93; E. Adler, L. R. Ellmer, *Acta Chem. Scand.* **2** (1948), 839.
44. C. Mäule, *Beitr. Wiss. Botanik* **4** (1900) 166.
45. Acid-Insoluble Lignin in Wood and Pulp, TAPPI T 222 OS-74, TAPPI Press, Atlanta.
46. D. B. Johnson, W. E. Moore, L. C. Zank, *Tappi* **44** (1961) 793.
47. S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 21–61.

48. R. S. Fukushima, R. D. Hatfield, *J. Agric. Food Chem.* **52** (2004) 3713.
49. A. Björkman, *Svensk Papperstidn.* **59** (1956), 477;
60 (1957) 158.
50. F. Lu, J. Ralph, *The Plant J.* **35** (2003) 535.
51. B. Hortling, E. Turunen, L. Sundquist, *Nord. Pulp. Pap. Res. J.* **7** (1992) 144.
52. S. Y. Lin in S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 217–232.
53. G. Aulin-Erdtman, *Svensk Papperstidn.* **55** (1952), 745.
54. O. Goldschmid, *J. Am. Chem. Soc.* **75** (1953) 3780.
55. H. L. Hergert in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignin, Occurrence, Formation, Structure*, Wiley-Interscience, New York 1971, p. 267.
56. O. Faix, *Holzforschung* **45** (1991) Suppl. 21.
57. H.-D. Lüdemanny, H. H. Nimz, *Biochem. Biophys. Res. Comm.* **52** (1973) 1162; *Angew. Makromol. Chem.* **175** (1974), 2393, 2409.
58. S. A. Ralph, J. Ralph, L. L. Landucci: *NMR Database of Lignin and Cell Wall Model Compounds*, November 2004. Available at URL http://ars.usda.gov/Services/docs.htm?docid=10491.
59. G. J. Leary, R. H. Newman in S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 146–160.
60. D. R. Robert, G. Brunow, *Holzforschung* **38** (1984) 85.
61. L. G. Akim et al., *Holzforschung* **55** (2001) 386.
62. R. M. Sevillano, G. Mortha, M. Barrelle, D. Lachenal, *Holzforschung* **55** (2001) 286.
63. L. Zhang, G. Gellerstedt in Proc. of Sixth European Workshop on Lignocellulosics and Pulp 2004, pp. 7–10.
64. H. Koskela, I. Kilpelainen, S. Heikkinen, *J. Magn. Res.* **174** (2005) 237.
65. T. M. Liitia, S. L. Maunu, B. Hortling, M. Toikka, *J. Agric. Food Chem.* **51** (2003) 2136.
66. J. Ralph, F. Lu, *Org. Biomol. Chem.* **2** (2004) 2714.
67. G. Gellerstedt in S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 322–333.
68. W. G. Glasser in W. G. Glasser, R. A. Northey, T. P. Schulz: "Classification of lignin according to chemical and molecular structure", *ACS Symp. Ser.* **742** (2000) 216–238.
69. C. Rolando, B. Monties, C. Lapierre in S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 334–349.
70. F. Lu, J. Ralph, *J. Agric. Food. Chem* **45** (1997) 2590.
71. D. Meier, O. Faix in: S. Y. Lin, C. W. Dence (eds): *Methods in Lignin Chemistry*, Springer Verlag, Berlin 1992, pp. 177–199.
72. P. Bocchini, G. C. Galletti, S. Camarero, A. T. Martinez, *J. Chromatogr.*, A **773** (1997) 227.
73. S. Reale, A. Di Tullio, N. Spreti, F. De Angelis, *Mass Spectrom. Rev.* **23** (2004) 87.
74. M. E. Himmel, J. Mlynár, S. Sarkanen in C. Wu (ed.) *Handbook of Size Exclusion Chromatography*, Marcel Dekker, New York, 1995, pp. 353–379.
75. T. H. Milne, H. L. Chum, F. Agblevor, D. K. Johnson, *Biomass Bioenergy* **2** (1992) 341.
76. A. Abächerli et al., in Proc. of the 7th International Lignin Institute Forum, Barcelona, ILI, Lausanne, 2005, p. 119.
77. A. Jacobs, O. Dahlman, *Nord. Pulp Pap. Res. J.* **15** (2000) 120.
78. D. W. Goheen in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignin, Occurrence, Formation, Structure*, Wiley-Interscience, New York 1971, p. 807.
79. T. Q. Hu (ed.): *Chemical Modification, Properties, and Usage of Lignin*, Kluwer Academic/Plenum Publishers, New York 2002.

Lime and Limestone

Tony Oates, Limetec Consultancy Services, Buxton, Derbyshire SK17 9AH, United Kingdom

1. Introduction . . . 1
2. Limestone . . . 3
2.1. Physical and Chemical Properties . 3
2.2. Formation and Occurrence . . . 5
2.3. Production . . . 5
2.4. Uses and Specifications . . . 6
3. Quicklime . . . 7
3.1. Physical and Chemical Properties . 7
3.2. Raw Materials . . . 9
3.2.1. Limestone . . . 9
3.2.2. Fuel . . . 10
3.3. Production . . . 11
3.3.1. Calcination . . . 11
3.3.2. Lime Kilns . . . 12
3.3.3. Quicklime Processing . . . 18
3.4. Uses and Specifications . . . 18
4. Hydrated and Slaked Lime . . . 19
4.1. Physical and Chemical Properties . 19
4.2. Raw Materials . . . 20
4.2.1. Raw Materials for Hydrated Lime . . 20
4.2.2. Raw Materials for Slaked Lime . . . 21
4.3. Production . . . 21
4.3.1. Normal Hydration Process . . . 21
4.3.2. Pressure Hydration Process . . . 22
4.3.3. Production of High Surface Area Hydrated Limes . . . 22
4.3.4. Normal Slaking Process . . . 22
4.3.5. Production of Ultrafine Milks of Lime 22
4.4. Uses and Specifications . . . 22
4.4.1. Uses . . . 22
4.4.2. Specifications . . . 25
5. Environmental Protection . . . 25
5.1. General . . . 25
5.2. Dust Emission . . . 25
5.3. Gaseous Emissions . . . 25
5.4. Noise and Vibration . . . 26
6. Physical Testing and Chemical Analysis . . . 26
6.1. Sampling and Sample Preparation . 26
6.2. Physical Testing . . . 27
6.3. Chemical Testing and Analysis . . . 27
7. Storage and Transportation . . . 27
8. Economic Aspects . . . 28
9. Toxicology and Occupational Health 29
9.1. Toxicology . . . 29
9.2. Occupational Health . . . 30
10. References . . . 31

1. Introduction

Limestone is a naturally occurring mineral that consists principally of calcium carbonate but may also contain magnesium carbonate as a secondary component. It is found in many forms and is classified in terms of its origin, chemical composition, structure, and geological formation. Limestone occurs widely throughout the world and is an essential raw material for many industries.

Quicklime is produced by the thermal decomposition of limestone. It consists mainly of calcium oxide. Its quality depends on many factors, including physical properties, degree of sintering, and chemical composition. As the most readily available and cost effective alkaline chemical, quicklime plays an essential part in a wide variety of industrial processes.

Hydrated lime and slaked lime are produced by reacting quicklime with water; they consist mainly of calcium hydroxide. In general, hydrated lime refers to a dry calcium hydroxide powder, while slaked lime refers to an aqueous suspension of calcium hydroxide particles in water. Hydrated and slaked lime are widely used in aqueous systems as low-cost alkaline chemical.

The term lime refers to quicklime and, less frequently, to hydrated or slaked lime. It is, however, sometimes used incorrectly to describe limestone, which is a frequent cause of confusion.

History. Limestone was probably used in the Stone Age. The first records relate to the Egyptian Second Dynasty (ca. 2800 B.C.), when it was employed in the construction of the Giza Pyramids. Marble, a highly crystalline form of limestone, was used by the Greeks shortly after this period for statues and the decoration of buildings. Limestone was widely used by the Romans for building roads.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_317

There is evidence of the widespread use of quicklime and hydrated lime for building by many civilisations by about 1000 B.C., including the Greeks, Egyptians, Romans, Incas, Mayas, Chinese, and Mogul Indians. The Romans employed hydraulic lime in many construction projects, including the Appian Way.

Lime was also well known to the Romans as a chemical reagent. In 350 B.C. XENOPHON referred to the use of lime for bleaching linen. The medical use of limewater was recorded by DIOSCORIDES in 75 A.D.

The burning of limestone in kilns was mentioned by CATO in 184 B.C. PLINY THE ELDER (ca. 70 A.D), in his "Chapters on Chemical Subjects" described the production, slaking, and uses of lime, and stressed the importance of chemical purity.

The use of limestone and lime in building spread throughout Europe in the 1400s.

In the 1700s, JOSEPH BLACK gave the first sound technical explanation of the calcination of limestone, including the evolution of gaseous carbon dioxide. LAVOISIER confirmed and developed BLACK's explanation. In 1766, DE RAMECOURT, published a detailed account of "The Art of the Lime Burner" which covered the design, operation, and economic aspects of limestone quarrying and lime burning.

Terminology. Because the quarrying of limestone and the production of quicklime and hydrated lime are long-established industries, they have generated many traditional expressions which frequently cause confusion. The following definitions cover the most common terms. A more comprehensive list has been published [5].

Air limes consist maionly of calcium oxide or hydroxide, which, when incorporated into a mortar mix, slowly harden by reacting with atmospheric carbon dioxide.

Air-slaked lime is produced by excessive exposure of quicklime to the atmosphere. It contains varying proportions of the oxides, hydroxides, and carbonates of calcium and magnesium.

Agricultural lime is a term which includes any limestone, quicklime, or hydrated lime product used to neutralize soil acidity.

Aragonite is one of the less abundant crystalline forms of calcium carbonate.

Available lime is an analytical term for the calcium oxide content of quicklime or hydrated lime that is able to react with sucrose under specified conditions.

Bituminous limestone – see carbonaceous limestone.

Carbonaceous limestone contains organic matter as an impurity. It is often dark gray and has a musty odor.

Carboniferous limestone was deposited in the Carboniferous period, 286 to 360 million years ago.

Calcitic limestone refers to a high-calcium limestone.

Calcite is the most abundant crystalline form of calcium carbonate.

Cement is produced by calcining limestone with materials containing silica, alumina, and iron oxide to produce a controlled blend of calcium silicates, aluminates, and ferrates in the form of a clinker. The clinker is then ground with gypsum and other materials to form cement.

Chalk is a naturally occurring limestone, which has been partially consolidated and therefore has a high porosity. It is generally relatively soft.

Chemical quality limestone is a high-calcium or dolomitic limestone with low levels of impurities, which meets the requirements of the chemical industry.

Dead-burned dolomite is a highly sintered form of dolomitic quicklime which is used primarily as a basic refractory.

Dolime refers to calcined dolomite.

Dolomite is strictly speaking the double carbonate containing 54 – 58 % of $CaCO_3$ and 40 – 44 % of $MgCO_3$. The term is, however, frequently used to describe dolomitic limestone.

Dolomitic limestone is generally understood to contain 20 – 44 % of $MgCO_3$.

Fine quicklime generally refers to screened products with a top size below 0.6 cm.

Free lime is an analytical term for the calcium oxide component of quicklime or hydrated lime. It excludes calcium oxide in $CaCO_3$, $Ca(OH)_2$, and calcium silicates.

Granular quicklime usually refers to screened products with a top size of 0.5 – 2.5 cm.

Ground quicklime refers to powdered products produced by milling.

Hard-burned quicklime is a sintered form of quicklime with low reactivity to water.

High-calcium limestone is a general term for limestone consisting mainly of $CaCO_3$ and having a low content of $MgCO_3$ (max. 2 – 5 %). Similarly, high-calcium quicklime contains mainly CaO and not more than 2 – 5 % MgO.

Hydrated lime is a dry powder produced by reacting quicklime with controlled amounts of water.

Hydraulic limestone is an impure carbonate containing considerable amounts of silica and alumina. Calcination of hydraulic limestone produces *hydraulic lime*, which, after mixing with water, has cementing (or hydraulic) properties and is capable of setting under water.

Iceland spar is a rare and very pure form of crystalline limestone. It is transparent and has been used in optical instruments.

Light-burned quicklime is quicklime that is lightly sintered and has a high reactivity to water.

Lime putty is a viscous dispersion of calcium hydroxide in water.

Lump quicklime usually refers to screened products with a top size above 6 cm.

Magnesian limestone is generally understood to be mainly $CaCO_3$ with 5 – 20 % of $MgCO_3$.

Marble is a highly crystalline carbonate rock which may be high-calcium or dolomitic limestone. It occurs in many colors with veined and mottled effects.

Marl is an impure, soft, earthy rock which contains clay and sand. If it contains more than 50 % $CaCO_3$, it is classified as a limestone.

Milk of lime is a fluid suspension of hydrated or slaked lime in water.

Neutralizing value is an analytical term for that proportion of limestone, quicklime, or hydrated lime (expressed as CaO) that is capable of reacting with hydrochloric acid under specified conditions. It includes the contribution of $CaCO_3$, CaO, $Ca(OH)_2$, and the acid soluble fraction of the calcium silicates, aluminates, and ferrates.

Pebble quicklime usually refers to screened products with a top size of 1.5 – 6 cm.

Precipitated calcium carbonate (PCC) is produced by blowing carbon dioxide into milk of lime, thereby precipitating finely divided calcium carbonate with a mean particle size in the range 0.02 to 0.2 µm.

Reactivity of quicklime is a measure of the rate at which it reacts with water. There are many reactivity tests (see Section 3.1).

Slaked lime generally describes an aqueous suspension of mainly calcium hydroxide. It includes milk of lime and lime putties. The term slaked lime is sometimes used to describe hydrated lime.

Soft-burned quicklime – see light-burned quicklime.

Solid-burned quicklime – see hard-burned quicklime.

Total lime is an analytical term for the total CaO plus MgO content of a limestone or lime, expressed in terms of CaO equivalent. It includes the carbonates, oxides, hydroxides, silicates, aluminates, and ferrates.

Type N or normal hydrated lime is defined in ASTM specification C-207 [2]. It is generally produced by hydrating high-calcium quicklime at ca. 100 °C.

Type S hydrated lime is also defined in ASTM specification C-207 [2]. It is produced by heating high-calcium or magnesian lime in an autoclave at ca. 180 °C. It may contain up to 8 % of unhydrated oxide.

Whiting is a finely powdered product produced by milling and classifying limestone (generally chalk). The nominal top size varies from 75 to 10 µm.

2. Limestone

2.1. Physical and Chemical Properties

The physical and chemical properties of limestones vary widely as a result of the origin of the deposit, its microstructure, and impurities. The information given below is typical of most commercially exploited deposits.

Color. Pure calcite is white. Chalk and marble are also generally white, although impurities in the latter can produce a variety of colors and patterns. Many limestones, however, contain carbonaceous material, which produces various shades of gray. Iron compounds can introduce a red color. Some impurities produce a surface coloration on weathering.

Odor. Any odor possessed by limestone generally arises from its content of carbonaceous material. The smell is musty or earthy.

Structure. All limestones are crystalline. The grain size increases with the amount of recrystallization that has occurred during formation of the deposit. Thus, shell limestones have a grain size of ca. 1 µm, marls and chalks from 2 to 5 µm, dense limestones from 5 to 250 µm, and marbles and calcite spar from 250 to 1000 µm. Calcite and dolomite have rhombohedral crystal structures, and aragonite is orthorhombic.

Porosity. The porosity of limestone decreases with increasing levels of compaction. Thus, marls have up to 50 % porosity, chalks 15 – 40 %, dense limestones 0.1 – 10 %, and calcite spar <1 %.

Density. For pure calcite, the density is 2.71 g/cm^3 at 20 °C; for aragonite and dolomite, it is 2.93 g/cm^3 and 2.86 g/cm^3, respectively. The porosity of limestone results in apparent densities of 1.5 – 2.3 g/cm^3 for chalk, up to 2.7 g/cm^3 for high-calcium limestone, and 2.7 – 2.9 g/cm^3 for dolomitic limestones.

Bulk Density. The bulk density depends primarily on the apparent density of the limestone and its particle size distribution. Crushed, screened limestone with a 2 : 1 size ratio generally has a bulk density of 1300 – 1450 kg/m^3. Crushed limestone, including the fines, has a bulk density of 1600 – 1750 kg/m^3.

Impurities. Magnesium carbonate is not generally regarded as an impurity. Impurities may be dispersed homogeneously as a result of their being present during the deposition process. Alternatively they may be present heterogeneously in features such as faults, bedding planes, or nodules. Silica and alumina, in the form of clay, silt, and sand, are commonly found as homogeneous impurities, but also occur as heterogeneous impurities. Similarly, iron can exist homogeneously as iron carbonate and heterogeneously as pyrite and limonite. Sulfur from sulfates and organic remains generally exists as a homogeneous impurity. Other trace impurities such as lead are often found in the vicinity of faults where mineralization has occurred. Typical levels of trace elements are given in Table 1.

Table 1. Typical ranges of impurities and trace elements in commercial limestones

Impurity/trace element	Typical range	Units
Silica (as SiO_2)	0.1 – 2.0	%
Alumina (as Al_2O_3)	0.04 – 1.5	%
Iron (as Fe_2O_3)	0.02 – 0.6	%
Sulfur (as $CaSO_4$)	0.01 – 0.5	%
Carbonaceous matter	0.01 – 0.5	%
Manganese (as MnO_2)	20 – 1000	mg/kg
Antimony	0.1 – 3	mg/kg
Arsenic	0.1 – 15	mg/kg
Boron	1 – 20	mg/kg
Cadmium	0.1 – 1.5	mg/kg
Chromium	3 – 15	mg/kg
Copper	1 – 30	mg/kg
Fluoride	5 – 3000	mg/kg
Lead	0.5 – 30	mg/kg
Mercury	0.02 – 0.1	mg/kg
Molybdenum	0.1 – 4	mg/kg
Nickel	0.5 – 15	mg/kg
Selenium	0.02 – 3	mg/kg
Silver	0.2 – 4	mg/kg
Tin	0.1 – 15	mg/kg
Vanadium	1 – 20	mg/kg
Zinc	3 – 500	mg/kg

Hardness. Pure calcite has a hardness of 3 Mohs. Naturally occurring limestones lie in the range 2 – 4 Mohs.

Strength. The compressive strength of limestone varies from 10 MPa for some marls and chalks to 200 MPa for some marbles.

Specific Heat. The specific heats of high-calcium limestone and dolomite at 20 °C are 817 and 880 J kg^{-1}K^{-1}, respectively. They increase with temperature [3].

Solubility. The solubility of calcite in distilled water free of carbon dioxide has been reported as ca. 15 mg/L at 20 °C [1]. The solubility in distilled water in equilibrium with atmospheric carbon dioxide is approximately double the above value. Dolomite has been reported as being somewhat more soluble than calcite [1].

Reaction with Acids. Calcium carbonate reacts with acids with evolution of carbon dioxide and heat (reported to be about 19 kJ/mol in hydrochloric acid [4]).

pH. Both limestone and dolomite are essentially neutral products, giving pH values of 8.0 – 9.2.

Thermal Dissociation. The heat of dissociation of calcium carbonate is 3180 kJ/kg of CaO relative to 25 °C, and 3010 kJ/kg relative to 900 °C (760 kcal/kg and 720 kcal/kg, respectively). The corresponding values for magnesium carbonate are 3010 kJ/kg MgO relative to 25 °C, and 2850 kJ/kg relative to 700 °C.

2.2. Formation and Occurrence

Limestone is widely distributed throughout the world in deposits of varying sizes and degrees of purity. Information regarding the deposits in a given country is generally best obtained by contacting the appropriate national geological society.

The process of limestone formation is believed to have started with the extraction of calcium salts from the earliest forms of igneous rocks by the combined effects of erosion by the weather and corrosion by dissolved acids, e.g., sulfurous acid and carbon dioxide.

Under certain conditions of concentration and temperature, calcium carbonate precipitated from the solutions of calcium salts. Deposits with such chemical origins are generally thin, of limited extent, and therefore of little commercial importance.

When the solutions of calcium salts drained into the sea, marine life extracted the dissolved calcium hydrogen carbonate component to build shells and skeletons of calcium carbonate. These gradually accumulated on the beds of shallow tropical seas to produce deposits, many of which were massive. The calcium carbonate became contaminated by waterborne clays and silts, and by airborne volcanic ash. The extent of contamination dependent on the distance from estuaries and volcanoes. In general the purest deposits originated from mid-ocean banks remote from land.

The limestone deposits became covered with other materials and were subjected to solvent action under high pressure and temperature. This consolidated the deposits and caused recrystallization to varying degrees. The sequence marl, chalk, limestone, and marble shows progressive consolidation and change of structure. Soft marls are porous and contain well-defined fossils. Marble is particularly dense, highly crystalline, and contains no discernible fossils. The grain size ranges from less than 5 µm for marls to over 250 µm for some marbles and calcite. The process of recrystallization is known as metamorphosis.

Under certain conditions, magnesium ions present in the seawater replaced alternate calcium ions in the calcite to form dolomite. This process is known as dolomitization.

2.3. Production

Most limestone is produced by open-cast quarrying. A small proportion (less than 5 %) is extracted by underground mining and a still smaller proportion (less than 1 %) as cut dimension-stone.

The first operation in open-cast quarrying is to remove the overburden (i.e., the soil, clay, and rock overlying the deposit). Various techniques and types of earth moving equipment are used for this.

The next stage is generally to drill the bedrock. Rotary and percussive drills are widely used to drill the holes. The spacing between holes and the burden (distance between the holes and the quarry face) is carefully controlled. The diameter of the holes varies from 5 – 25 cm, depending on the design of the blast.

The drill holes are then filled with controlled amounts of explosive. A mixture of ammonium nitrate and fuel oil (ANFO) is widely used; it is initiated with a high explosive. In a typical blast, 5000 – 50 000 t of stone is fragmented with about 140 g of explosive per 1 t limestone.

It is important to control the blast. Too little fragmentation produces oversize boulders which have to be broken using secondary blasting or a drop ball. Too much fragmentation produces an excessive amount of fine particles and increases the risk of throwing rock away from the quarry face. Optimum fragmentation is obtained by selection of the correct diameter and spacing of holes, time delay between holes, thickness of burden, and quantity of explosive.

Some soft rocks, e.g., chalk and marl, are broken by using heavy rippers. Other soft deposits,

as found in lakes or in the sea, are extracted with dredgers. Underground mining is used when the overburden is thick, or when the limestone deposit is overlain by other rocks. Most mines use the room and pillar technique. After blasting, the rock is loaded into dump trucks, generally by front-end shovels or hydraulic excavators and transported to the crushing and screening plant. In the crushing and screening plant, the larger lumps of rock are broken in a primary crusher to a size which suits the needs of the business and the characteristics of subsequent equipment. The rock is then screened into oversize (e.g., + 15 cm), large (e.g., − 15 + 5 cm), medium (e.g., − 5 + 0.6 cm), and fine (e.g., − 0.6 cm) fractions. The oversize stone is frequently recrushed and rescreened. The fine fraction is rich in impurities such as clay, and is often tipped. In some plants clay and the fine fractions are removed more efficiently by washing and screening.

Compression machines such as jaw, gyratory, and rolls crushers are generally selected when the amount of fine fraction must be minimized and when slabby lumps can be tolerated. Impact crushers such as hammer mills and impact breakers are selected when cubical lumps are required and increased fines production can be tolerated. Impact crushers have the advantage of being able to achieve a greater size reduction at lower capital cost. Although the major demand for limestone is in the − 2.5 + 0.1 cm size range, specialized requirements exist for very finely divided products, which are produced in a variety of mills, generally by dry grinding.

Dimension-stone production (for facing buildings and ornamental use) is a specialized process in which blocks are cut from the rock face with channelling machines or wire saws. The blocks are then cut, shaped, and, if required, polished to make the finished product.

2.4. Uses and Specifications

Construction. Limestone is the most widely used crushed rock, although it is generally outsold by sand and gravel.

The quality specifications for construction stone relate mainly to its physical properties. Thus the stone must be clean, strong, dense, durable, free from cracks, and have the required particle size distribution. Some specifications limit the amounts of organic matter, clay, or water-soluble components (e.g., alkali-metal salts and gypsum). Specifications for the stone in the top dressing of roads (particularly in the vicinity of junctions) sometimes require a high resistance to polishing; this excludes the use of limestone. Standards for the testing and size distribution of roadstone and construction stone are given in Section 6.2.

Cement. The production of cement is the main use for chalk, and a major use for dense, high-calcium limestone. Approximately 1.0 t of limestone is required for 1 t cement. For the estimated worldwide production of cement in 1994, the consumption of limestone was ca. 1420×10^6 t [5].

Because cement is a mixture of calcium silicates, aluminate, and ferrate, the presence of SiO_2, Al_2O_3, and Fe_2O_3 in the feedstone is acceptable, provided the level is uniform. When the composition of the limestone is variable, elaborate arrangements are made to produce a consistent chemical analysis by blending. The magnesium carbonate content, however, must be below 5 %. For further details of cement production, see → Cement and Concrete.

Quicklime. Approximately 1.8 t of limestone is required per 1.0 t of quicklime. Further details are given in Section 3.2.

Agriculture. Many crops grow best under neutral to slightly acid conditions (i.e., pH 6 – 7). Thus soils which are more acidic than pH 6 generally benefit from the application of limestone. The limestone also helps to replace the calcium and magnesium removed by crops. Other benefits include an increase in the supply of other chemical nutrients, an increase in the organic matter of the soil, an increase in beneficial soil microorganisms, improved soil tilth, an improved supply of trace elements, and an increase in the efficiency with which fertilizers are used by the crop [5].

It is important that the limestone used for agriculture is in the correct physical state; a top size of ca. 0.3 cm is generally required. The delivered cost per tonne of carbonate is an important factor. In some cases, the magnesium con-

tent of dolomite gives it an advantage over high calcium limestone.

Limestone is also used in animal feeds and poultry grit. In animal feeds, trace elements may be restricted (e.g., lead levels below 10 mg/kg).

Metal Refining. When limestone is used in metal refining it is initially converted to quicklime, which reacts with acidic oxides (e.g., SiO_2, Al_2O_3, and Fe_2O_3) to produce molten slags. Most is used in blast furnaces for the production of iron, where the slag typically contains 40–50 % CaO. Smaller amounts are used in the production of copper, lead, zinc, and antimony. The quality requirements are as for chemical-quality stone. Some magnesium carbonate is acceptable but not essential.

Flue Gas Desulfurization. A growing use of limestone is in the treatment of flue gas to remove sulfur dioxide. For the wet scrubbing process the limestone is finely ground to 90 % less than 45 µm and contacted with flue gases in a scrubber. The resulting calcium sulfate may be converted to salable gypsum, in which case the $MgCO_3$ content should not exceed 1 %. In the dry process finely divided limestone is injected into the furnace, where it is converted to quicklime, which then reacts with the sulfur dioxide.

Other Uses. There are many other uses of limestone, e.g., production of alumina, glass, wood pulp, paper, ceramics, mineral wool, fillers, whiting, and coal-mine dusts, neutralization of acids and construction of filter beds; they are described in [5].

3. Quicklime

3.1. Physical and Chemical Properties

Color. Most quicklimes are white. Impurities can result in gray, brown, or yellow tints. When quicklime is produced by solid fuel firing, a gray surface contamination is produced.

Odor. Quicklime has a slight earthy odor.

Porosity. The porosity of commercially produced quicklime can be as high as 55 %, if a highly porous limestone is lightly burned. Exposure to elevated temperature results in sintering (see Fig. 1) which can reduce the porosity to below 25 %. Dead-burned dolomite has a porosity of ca. 10 %.

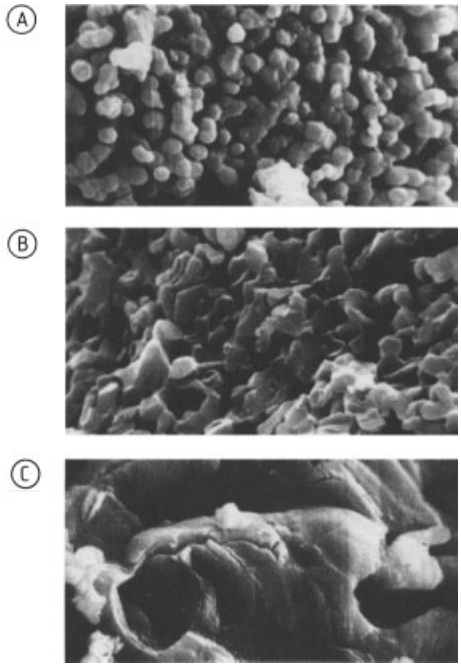

Figure 1. Scanning electron micrographs of quicklimes A) Apparent density 1.5 g/cm^3; B) Apparent density 1.9 g/cm^3; C) Apparent density 2.3 g/cm^3

Density. The true density of calcium oxide is ca. 3.3 g/cm^3. The apparent density of lumps can be as low as 1.4 g/cm^3 for highly porous quicklime. This can rise to over 2.3 g/cm^3 after sintering (Fig. 2).

Calcined dolomite is generally denser than high-calcium quicklime, if given the same heat treatment. Dead-burned dolomite has an apparent density of ca. 3.2 g/cm^3.

Bulk Density. The bulk density depends on the mean apparent density of the particles and on the voidage between them. The latter is related to the particle size distribution and the particle shape. Most screened commercial quicklime products have compacted bulk densities of 900–1200 kg/m^3. Inclusion of fines can increase the bulk density by 30 %.

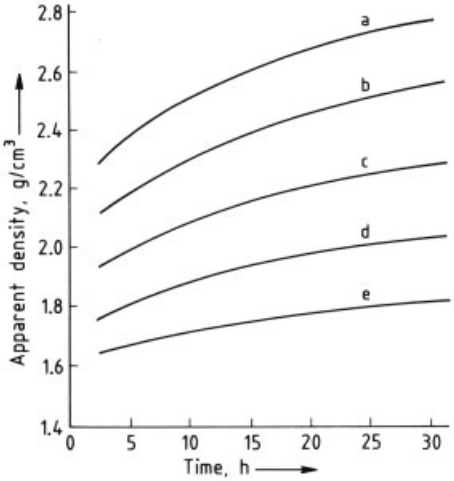

Figure 2. Variation of apparent density with temperature and time for a dense, high-calcium quicklime
a) At 1400 °C; b) 1300 °C; c) 1200 °C; d) 1100 °C; e) 1000 °C

Hardness. Most commercial quicklime products have a hardness of 2 – 3 Mohs. The value for dead-burned dolomite is in the range 3 to 5 Mohs.

Melting Point. The melting points for calcium oxide and magnesium oxide are 2580 °C and 2800 °C, respectively. The value for calcined dolomite is about 2400 °C [1].

Specific Heat. The mean specific heats for high-calcium and dolomitic quicklime at 20 °C are 760 and 830 J kg^{-1} K^{-1}, respectively. They increase with temperature [3].

Angle of Repose. The angle of repose for cubical, well-graded pebble quicklime is about 35°. This increases, however, as the fines content increases. For bunker design, valley angles of not less than 60 ° are recommended.

Heat of Hydration. The heat liberated by the reaction of quicklime with water is 1140 kJ/kg of CaO. The value for dolime is 880 kJ/kg of CaO · MgO.

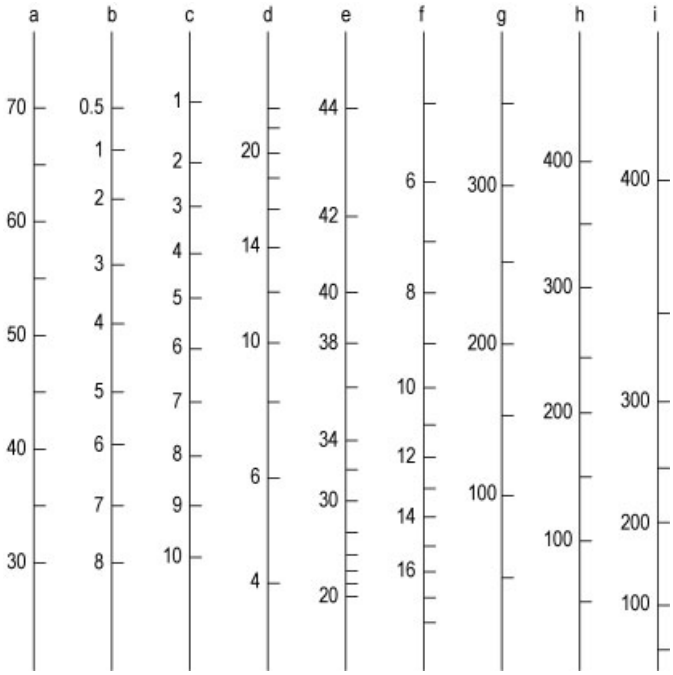

Figure 3. Comparisons between reactivity tests for a high-calcium quicklime
a) BS 6463 1987, °C after 2 min [7]; b) EN 459-2, t_{60} (min.) – time to reach 60 °C [8]; c) EN 459-2, t_u (min.) – time for 80 % slaking [8]; d) ASTM C 110, temperature rise (°C) after 30 sec [9]; e) ASTM C 110, maximum temperature rise (°C) [9]; f) ASTM C 110, time to maximum temperature (min.) [9]; g) Acid titration, (mL) at 3 min [10]; h) Acid titration, (mL) at 5 min [10]; i) Acid titration, (mL) at 10 min [10]

Reactivity to water may be measured by the rate of release of the heat of hydration [7–9] or by the rate at which an aqueous suspension reacts with hydrochloric acid [10]. The relationships between some reactivity tests are are shown in Figure 3. Reactivity is also related to the mean apparent density of quicklime produced from a given source stone (Fig. 4).

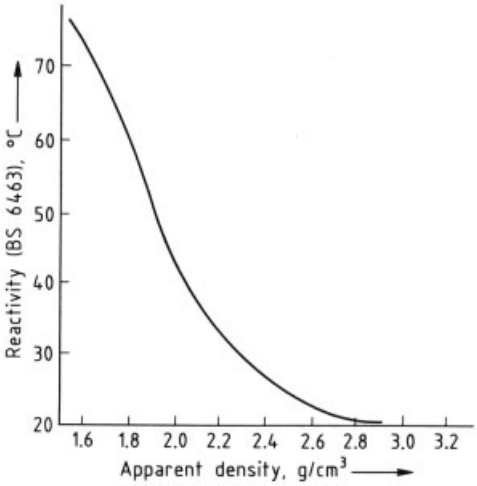

Figure 4. Relationship between reactivity and apparent density for a high-calcium quicklime

It should be noted that the reactivity of quicklime can be markedly affected by impurities in the water. For this reason, distilled water should be used as a reference standard.

The low reactivity of calcined dolomite probably arises mainly from the low solubility of $Mg(OH)_2$ in water. Sintering of MgO, which forms at lower temperatures than CaO may also be a factor.

Affinity for Water. Quicklime has a high affinity for water and is a more efficient desiccant than silica gel. Because of its high affinity for water vapor and (after partial hydration) for carbon dioxide, care should be taken to minimize exposure of quicklime to the atmosphere. Relatively small amounts of air slaking (less than 1.5 % of combined water) can reduce the reactivity significantly.

The reaction of quicklime with water is associated with an increase in volume by a factor of at least 2.5. This can cause the expansion of products that contain some lime which is not fully hydrated.

Reaction with Carbon Dioxide. In the absence of calcium hydroxide, quicklime only reacts with dry carbon dioxide above ca. 300 °C and below ca. 800 °C, depending on the carbon dioxide pressure. Partial recarbonation in the absence of calcium hydroxide can occur in lime kilns under certain abnormal conditions. It can produce an unexpectedly low reactivity and markedly affects the normal relationship between reactivity and mean apparent specific gravity.

Acid Neutralization. Quicklime is used extensively to neutralize acidic oxides, such as SiO_2, Al_2O_3, and Fe_2O_3, in high-temperature nonaqueous systems (e.g., in steel making). Temperatures are generally controlled to ensure that the resulting calcium and magnesium salts produce molten slags.

Reaction with Carbon. Above 1800 °C, carbon reduces calcium oxide to produce calcium carbide and carbon monoxide.

3.2. Raw Materials

3.2.1. Limestone

Size. The ratio of the top to bottom sizes in the stone fed to lime kilns should preferably be 2 : 1 and certainly no more than 3 : 1. This gives an open, porous bed which offers a low resistance to gas flow and heat transfer. The narrow size range also helps to limit sintering of the smaller sizes while the larger are still dissociating. Selection of the size is influenced by the cost and availability of the stone and by the limitations imposed by the design of the lime kiln.

Shape. While a cubical shape is generally preferred for lime kilns, the slabby shapes produced by jaw and roll crushers are acceptable, providing they do not lead to a packed bed with low porosity.

Strength and Abrasion Resistance. Both the stone and the quicklime should be sufficiently strong to resist the physical forces to which they are exposed in the kiln and the handling and storage system. Excessive breakage

in the kiln increases the resistance to gas flow and heat transfer. Breakage reduces the yield of granular products, which are usually at a premium, and increases the yield of fines, which are usually in surplus. Various empirical tests have been developed by kiln manufacturers to quantify the above factors.

Thermal Degradation. Some limestones, and particularly highly crystalline ones, are prone to break-down as a result of the heating, calcining, or thermal cycling processes within the kiln. Various empirical tests have been developed to quantify these effects.

Water. Some chalks are highly porous and may contain 20 % water, which reduces the temperature of the kiln exhaust gas and may cause condensation problems in dust collectors.

Impurities. When limestone is burned in a kiln, most of the impurities persist in the quicklime. Permissible levels of SiO_2, MgO, Al_2O_3, Fe_2O_3, S, Pb, and F depend on the quality specifications for the quicklime. Heterogeneous impurities (e.g., those arising from contamination with clay) tend to concentrate in the smaller sizes of quicklime. Removal of the fines (e.g., − 6 mm) by screening generally improves product quality. Much of the sodium and potassium compounds and some of the sulfur compounds are removed during calcination. Sulfur oxides, most of which may arise from the fuel, and volatile impurities, such as Pb and F, are reabsorbed onto the surface of the quicklime at the start of the calcining zone and concentrate in the fine fraction.

3.2.2. Fuel

In lime burning, the fuel is more than a source of heat. It interacts with the process, and the combustion products react with the quicklime. Many different fuels are used in lime kilns. The most common is coal. Coke, fuel oil, and natural gas are also widely used.

Selection of the correct fuel is important to the lime producer. Its cost per tonne of quicklime frequently represents 50 % of the total production cost. Some fuels cannot be used in certain kilns. Other fuels may markedly affect the heat usage, output, and product quality. Some require different refractory linings to be used.

The selection of a new fuel, e.g., a different coal, is often a matter of trial and error. The selection for a new kiln should be made with great care. Most kilns can be operated on more than one fuel, and this enables the operator to select the fuel or combination of fuels which gives the optimum economic performance.

The major factors relating to the performance and acceptability of fuel are discussed briefly below. Further details are available in [5].

Calorific Value. The calorific value coupled with the unit cost of the fuel and the kiln heat usage enables the cost of the fuel per unit of quicklime to be calculated. Some confusion arises between the gross (or upper) and the net (or lower) calorific values. The former includes the latent heat of condensation of the water produced by combustion; it is used in North America and the United Kingdom. The net value more logically excludes the latent heat; it is widely used in, for example, continental Europe and Japan.

Sulfur. Some important markets for quicklime, notably steelmaking, require low sulfur levels. Sulfur from the fuel is absorbed by quicklime in the cooler part of the calcining zone of lime kilns as calcium sulfate. In the more efficient kilns, e.g., shaft kilns, most of this sulfur is retained in the quicklime. In the less efficient units, e.g., some rotary kilns, much of the sulfur from the fuel may be eliminated from the kiln in the kiln gases by operating the calcining zone with low levels of excess air and high temperatures, and by limiting contact between the kiln gases and the quicklime in the cooler part of that zone.

Combustion Characteristics. The combustion characteristics of fuels vary markedly. Pulverized coal tends to burn with a short, hot, and highly (infrared) emissive flame. Natural gas and wood burn with longer, cooler flames which have lower emissivities. These differences can affect heat usage and many aspects of quality. In particular the relationship between the residual $CaCO_3$ content and the reactivity to water may be affected by the choice of fuel.

Particle Size. Solid fuels should be of the correct particle size. In some cases they need to have adequate strength. Their coking properties may be critical, as may be the amount of volatile matter. The ash generally contaminates the lime to some degree with silica, iron oxide, and alumina. It may cause bridging between particles, and can also combine with lime dust and with volatile alkalis (sodium and potassium) to form troublesome deposits.

Ash Fusion. The properties of the fuel ash can have a marked effect on the acceptability of the fuel. Key properties include the ash content of the fuel, the ash fusion temperature (which is affected by lime dust and by the concentration of oxygen and carbon monoxide in the atmosphere), and the level of alkalis.

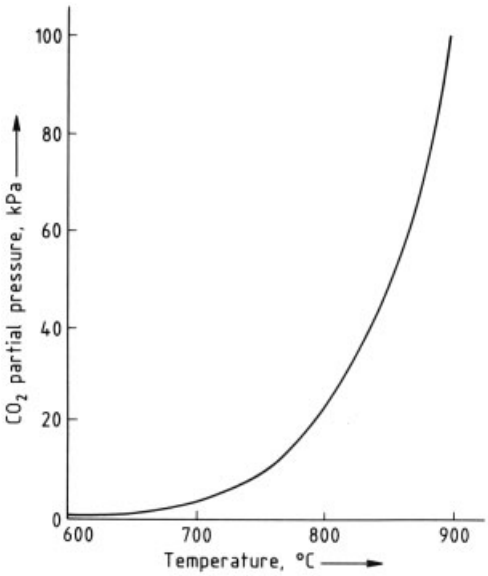

Figure 5. Variation of the CO_2 partial pressure above $CaCO_3$ with temperature

3.3. Production

3.3.1. Calcination

The chemical reaction for the decomposition of calcium carbonate by heat is:

$$CaCO_3 \longrightarrow CaO + CO_2$$

$$\Delta H = +3010\,\text{kJ/kg at } 900\,^\circ\text{C } (720\,\text{kcal/kg})$$

Figure 5 shows the variation of the partial pressure of carbon dioxide above calcium carbonate with temperature. The pressure is 101 kPa at ca. 900 °C. Thus, although surface calcination can proceed at lower temperatures, complete calcination only occurs above 900 °C.

The calcination mechanism is complex and involves several stages: heat transfer to the surface of the particle and through the outer layer of lime. The heat is absorbed by the chemical reaction at the lime – limestone interface. The carbon dioxide so generated migrates to the surface of the particle counter to the heat flow and then diffuses from the surface into the kiln gases.

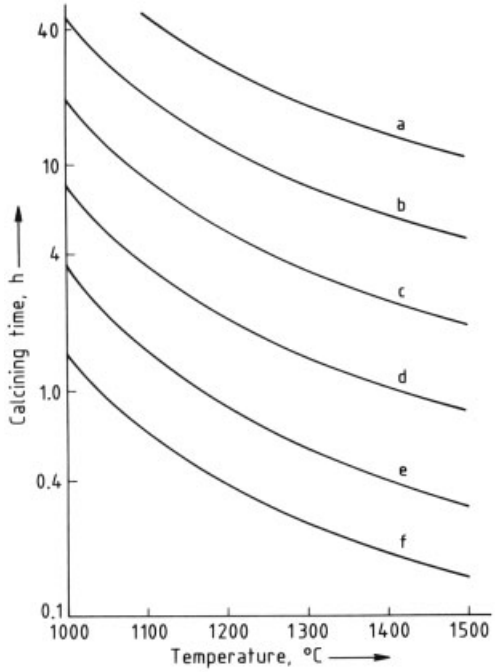

Figure 6. Calcining times for spheres of a dense, high-calcium limestone
a) 15 cm; b) 12.5 cm; c) 10 cm; d) 7.5 cm; e) 5 cm; f) 2.5 cm diameter

The rate-determining stage in the above process depends on the particle size, temperature, amount of calcination that has already taken place, and the composition of the kiln gases. Although attempts have been made to produce a mathematical model to account for observed effects in practical lime burning, none has proved applicable over wide ranges of the above variables.

Figures 6 and 7 summarize smoothed results for calcining spheres of UK carboniferous limestone. Although the results do not apply accurately to the random shapes encountered in practice or to other limestones, they serve as a useful guide to the relative effect of changing residence time or temperature.

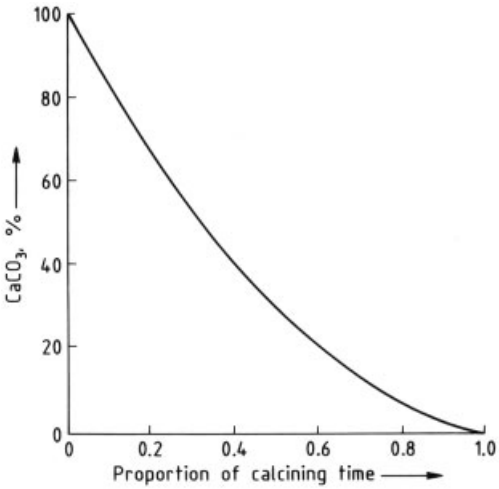

Figure 7. Progression of calcination of limestone spheres

Figure 6 shows the marked variation in the time for complete calcination caused by changes in particle size. This is one of the reasons why the size range of stone fed to lime kilns is generally in the ratio of 2 : 1. The other reasons are the effects on sintering and on the porosity of the stone bed, which must be sufficiently high to allow rapid gas flow and efficient heat transfer.

When a high-calcium limestone particle is decomposed at low temperature (950 °C or below), its external dimensions do not change significantly. Calcination of limestone with 2 % porosity theoretically produces quicklime with a porosity of 54 %, corresponding to an apparent density of 1.5 g/cm^3.

Prolonged heating of quicklime above 900 °C causes sintering, a reduction in porosity, and an increase in apparent density (see Fig. 2). This sintering process markedly reduces the reactivity of the quicklime to water (see Fig. 4).

In lime kilns, variations in heat distribution, temperature, and solids residence time further complicate the calcination process. As a result the quicklime particles discharged from the kiln can be grouped into three categories:

1) Particles that are not fully calcined (i.e., with a central $CaCO_3$ core), the lime layer of which has a low apparent density and a high reactivity to water
2) Particles that are just fully calcined and have a low apparent density and a high reactivity to water
3) Particles that have sintered to varying degrees and which have an increased apparent density and a reduced reactivity to water

The relative quantities of product in the above categories are influenced by the kiln design. In most kilns, the quicklime is exposed to gas temperatures of 1200 – 1300 °C just before it enters the cooling zone. Generally, such kilns only produce highly reactive quicklime from high-calcium limestone if the residual $CaCO_3$ content is relatively high. Some designs complete the calcination at a lower temperature (ca. 1100 °C), and produce a highly reactive quicklime with low levels of $CaCO_3$. In some kilns, changing the air – fuel ratio can affect the characteristics of the quicklime. A low ratio lengthens the flame, and hence the calcining zone, and reduces maximum gas temperatures. This leads to a higher reactivity for a given $CaCO_3$ content. It may, however, increase specific heat usage.

If the quicklime is exposed to kiln gases containing carbon dioxide at 600 – 800 °C, recarbonation occurs. This increases the $CaCO_3$ level and reduces the available lime content slightly. The most marked effect, however, is the reduction in the reactivity of the quicklime.

A small amount of recarbonation always occurs in the cooling zone of the kiln as a result of air slaking. The quantities involved are, however, small and the effect on the quality of the quicklime is negligible.

Any magnesium carbonate in the limestone decomposes at ca. 700 °C. Its heat of calcination is lower than that of calcium carbonate. The resulting magnesium oxide does not contribute significantly to the reactivity towards water.

3.3.2. Lime Kilns

Early lime kilns were constructed of stone and were generally built into the side of a hill [11]. An amount of fuel, originally wood, was placed on a hearth at the base of the kiln. Large stones

were placed above the fuel, followed by layers of increasingly smaller stone. The fuel was then lit and allowed to burn for a few days. After the fire had burned out and the lime had cooled, the kiln was drawn-down by hand from the hearth. The product often contained substantial amounts of both over- and under-burned lumps, and the thermal efficiencies were very low.

It was then recognized that a continuously operating lime kiln would be more productive and more thermally efficient. For the purpose of heat transfer, a kiln should be regarded as consisting of three zones (see Fig. 8).

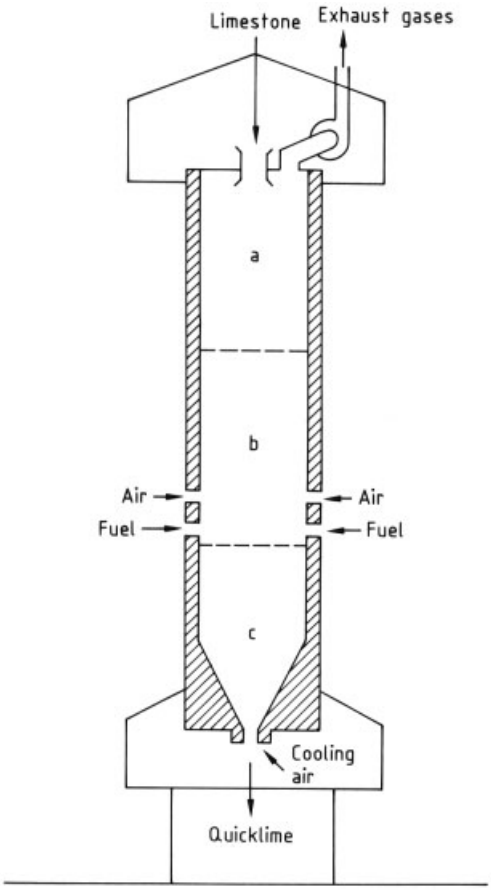

Figure 8. Schematic of a vertical shaft kiln.
a) Preheating zone; b) Calcining zone; c) Cooling zone

In the pre-heating zone, limestone is heated from ambient temperature to over 800 °C by the heat of the gases emerging from the calcining zone. In the calcining zone, the calcium carbonate is decomposed into calcium oxide, thereby absorbing some of the heat of combustion produced by burning the fuel and preheated air. In the cooling zone, quicklime is cooled by part or all of the combustion air, which in turn is preheated.

Subsequently, vertical shaft, mixed-feed kilns were developed in which either layers of stone were alternated with layers of fuel (initially wood or coal) or a blend of stone and fuel was charged to the top of the kiln. Calcined quicklime was drawn from the base of the kiln and additional fuel and stone were charged into the top of the kiln. The thermal efficiencies of these designs were better than those of the earlier kilns, but were still poor, owing to the incomplete combustion of the volatile matter. Use of coke overcame this problem and high thermal efficiencies were obtained. A large number of shaft kiln designs were developed [11]. Kiln outputs were often increased by the use of fans to increase the air and exhaust gas flow rates.

Most of the kilns in current use are based on either the vertical shaft or on the rotary design. There are a few other kilns based on different principles. All of these designs incorporate the concept of the three zones. In some kilns, they are incorporated into one unit, in others they exist as separate units (see Fig. 13).

Vertical Shaft Designs. Figure 8 shows a schematic of a vertical shaft kiln. The major problem with shaft kilns is obtaining uniform heat release across the shaft. Fuel injected at a wall usually does not penetrate more than 1 m into a packed bed. This limits the kiln width (or diameter) to 2 m. Increased fuel penetration can be achieved in larger shafts by:

1) Use of the mixed feed technique,
2) Use of central burners or lances,
3) Injecting the fuel via tuyères which penetrate ca. 1 m into the kiln,
4) Injection of the fuel underneath arches, or
5) Injection of air or recycled kiln gas above the fuel.

In general, vertical shaft kilns have relatively low specific heat usages because of efficient heat transfer between the gases and the packed bed. However, they retain most of the sulfur in the fuel so that low-sulfur quicklime can only be produced if a low-sulfur (and generally expensive) fuel is used to calcine a low-sulfur stone.

Older designs tend to produce quicklime with a low to moderate reactivity and a relatively high $CaCO_3$ content. Modern designs incorporate features which enable highly reactive lime to be produced with low $CaCO_3$ levels. Five designs of shaft kiln, which are used extensively throughout the world, are described below.

Mixed-Feed Shaft Kiln. Modern mixed-feed kilns use limestone with a top size in the range 5 to 15 cm and a size ratio of ca. 2 : 1. The most widely used fuel is a dense grade of coke with low reactivity and low ash content. The coke size is only slightly smaller than that of the stone so that it moves with it rather than trickling through the interstices. The stone and coke are mixed together and are charged into the kiln in such a way as to minimize segregation.

The net heat usage can be very low at ca. 4000 kJ/kg (950 kcal/kg). This advantage is offset by the high cost of coke compared to competitive fuels. Another advantage of the mixed-feed shaft kiln is that it produces kiln gas with a very high CO_2 content. For processes which can use both the quicklime and the CO_2 (e.g., the precipitated calcium carbonate process, the Solvay process, and the sugar beet process), this is an important factor in the overall economics.

The quality of the quicklime tends to be moderate, with the reactivity being considerably lower than that obtained by rotary kilns at the same $CaCO_3$ level. This, however, can be an advantage for certain uses. The retention of sulfur from the fuel is high.

Double-Inclined Shaft Kiln. The double-inclined shaft kiln (Fig. 9) is essentially rectangular in cross section but incorporates two inclined sections in the calcining zone. Opposite each inclined section, offset arches create spaces into which fuel and preheated combustion air are fired via three combustion chambers. Cooling air is drawn into the base of the kiln and is preheated. Part of the air is withdrawn and re-injected via the combustion chambers. The tortuous paths for both the gases and the burden, coupled with firing from both sides, ensures efficient distribution of heat. These kilns accept stone with a top size of 4 to 10 cm. They can produce reactive, low-carbonate quicklime at a net heat usage of about 4300 kJ/kg (1030 kcal/kg). A range of solid, liquid and gaseous fuels can be used, although they should be selected with care to avoid excessive buildups caused by fuel ash and calcium sulfate deposits.

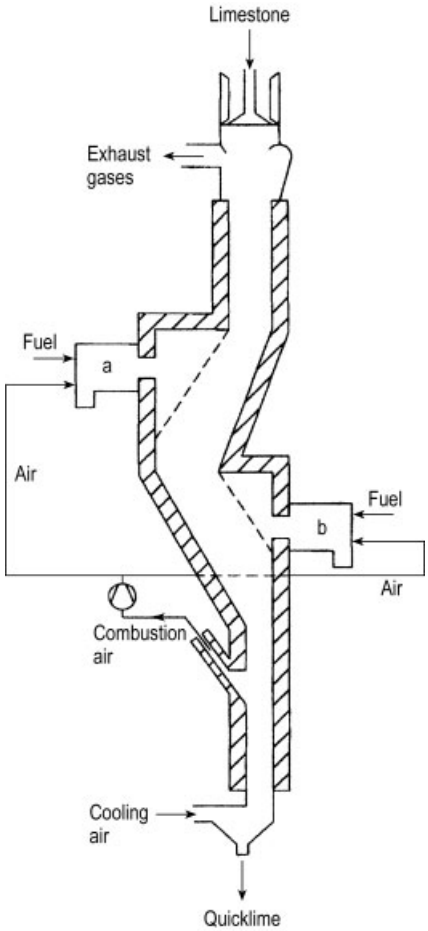

Figure 9. Schematic of a double-inclined shaft kiln
a) Upper burners; b) Lower burners

Multichamber Shaft Kiln. The multichamber shaft kiln is a development of the double-inclined kiln. It consists of four or six alternately inclined sections in the calcining zone, opposite each of which is an offset arch. The arches serve the same purpose as in the double-inclined shaft kiln. Cooling air is preheated by lime in the cooling zone, and is withdrawn, dedusted and re-injected via the combustion chambers. The temperature of the lower combustion chambers can be varied to control the reactivity of the quicklime over a wide range. The kiln can be fired with solid, liquid, and gaseous fuels

or a mixture thereof. Its net heat usage is about 4300 kJ/kg (1030 kcal/kg).

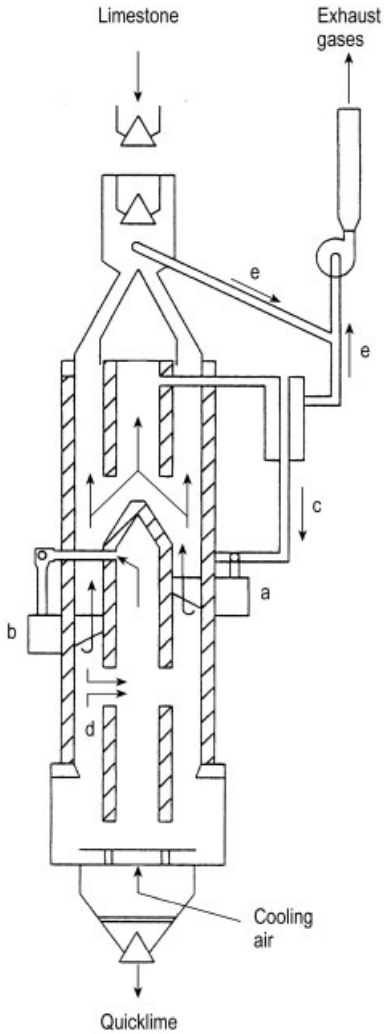

Figure 10. Schematic of an annular shaft kiln
a) Upper burners; b) Lower burners; c) Combustion air to upper burners; d) Combustion air to lower burners; e) Kiln gases

Annular Shaft Kiln. The major feature of the annular shaft kiln (Fig. 10) is a central cylinder which restricts the width of the annulus, and ensures good heat distribution. The central column also enables part of the combustion gases from the lower burners to be drawn down the shaft and to be injected back into the lower burner chamber. This recycle moderates the temperature at the lower burners and ensures that the final stages of calcination occur at low temperature. Both effects help to manufacture a product with a low $CaCO_3$ level and high reactivity.

The annular shaft kiln accepts stone with a top size of 5 – 11 cm. Use of a heat recuperator, in which 30 % of the kiln gases are used to preheat part of the combustion air, reduces the net heat usage to about 4180 kJ/kg (1000 kcal/kg). The kiln can be fired by gas, oil, or solid fuel.

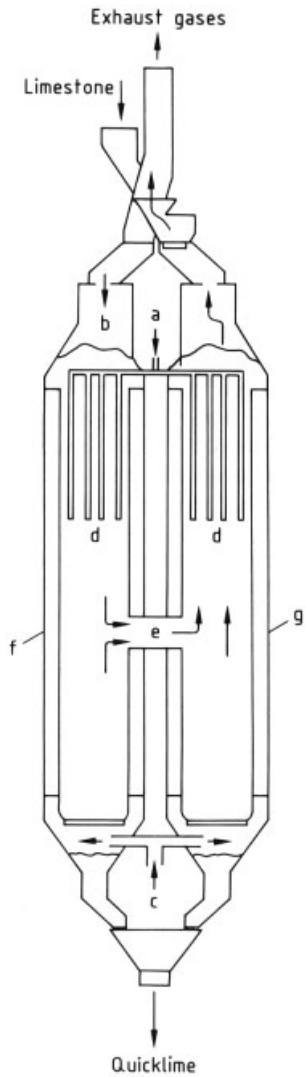

Figure 11. Schematic of a parallel-flow regenerative kiln a) Fuel; b) Combustion air; c) Cooling air; d) Lances; e) Cross-duct; f) Shaft 1; g) Shaft 2

Parallel-Flow Regenerative Kiln. The parallel-flow regenerative kiln (Figs. 11 and 12) consists of two or three interconnected vertical shafts. The following description relates to the more common two-shafted design. The operation consists of two equal stages, each of which lasts about 10 min.

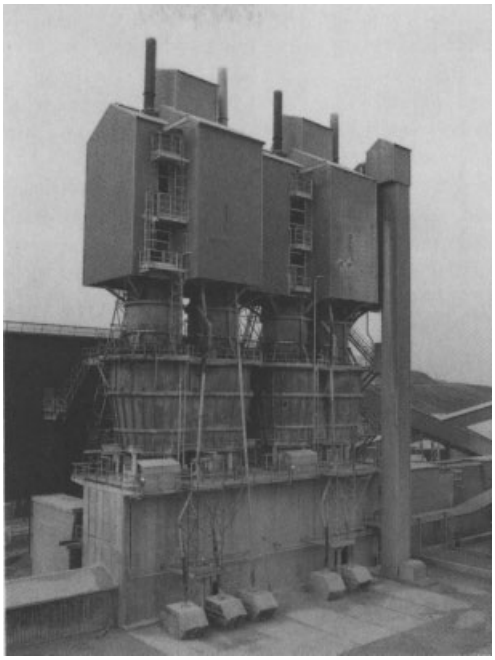

Figure 12. Two 300 t/d parallel-flow regenerative kilns

In the first stage, fuel is injected through lances into shaft 1 and burns with combustion air blown down the shaft. The heat released is partly absorbed by calcining the limestone in shaft 1. Air is blown into the base of each shaft to cool the lime. The air in shaft 1 mixes with the combustion gases, including the carbon dioxide from calcination. The mixture passes through the cross-duct into shaft 2, at about 1050 °C. In shaft 2, the gases from shaft 1 mix with the cooling air blown into shaft 2 and pass upwards. In so doing, they heat the stone in the pre-heating zone of that shaft.

During the second stage of the operation, the converse applies. The same amounts of fuel and combustion air are added to shaft 2. The combustion gases plus cooling air pass upwards in shaft 1, heating the stone in the pre-heating zone of that shaft.

The two key principles of the above operation are:

1) The stone-packed preheating zone in each shaft acts as a regenerative heat exchanger in addition to pre-heating the stone to calcining temperature. The surplus heat in the gases is transferred to the stone in shaft 2 during the first stage. It is then transferred from the stone to the combustion air in the second stage. As a result, the combustion air is preheated to about 800 °C. The net heat usage of the kiln is about 3770 kJ/kg (900 kcal/kg) of quicklime.
2) The calcination of the quicklime is completed at the level of the cross-duct at a moderate temperature of about 1100 °C. This favors the production of a highly reactive quicklime, which may, if required, be produced with a low $CaCO_3$ content. The kiln accepts stone with a top size of 5 – 12 cm, and it can be fired with gas, oil, or solid fuel.

Rotary Kilns. There are many designs and variants of the rotary kiln (Fig. 13). Most use a feedstone with a top size in the range 1 to 6 cm. They operate well on gaseous, liquid, or solid fuels.

Heat usages for rotary kilns are generally much higher than those of shaft kilns and their capital costs tend to be higher. These adverse factors are often offset by their ability to produce a high-quality quicklime with lower $CaCO_3$ and sulfur levels and high reactivity, when fired by less expensive fuels.

In the earliest rotary kilns, stone was charged into the elevated end of the rotating section (generally inclined at between 1.5 and 4°). It was preheated by the gases drawn from the calcining zone at the other end of the kiln, and then calcined as it moved towards and under the flame. The hot lime was then discharged into a pit to cool.

In later designs, the thermal efficiency of the rotary kiln is improved by:

1) Fitting lime coolers which preheat the combustion air,
2) Fitting raised sections of refractory (i.e., dams or mixers) in the calcining zone to improve heat transfer,

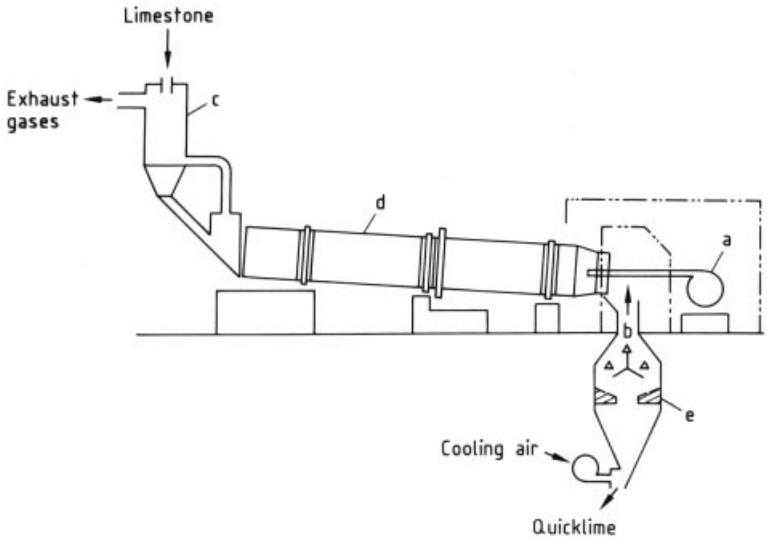

Figure 13. Schematic of a rotary kiln a) Burner; b) Combustion air; c) Pre-heater; d) Kiln; e) Cooler

3) Installing refractory trefoils, metal lifters, or similar devices in the pre-heating zone of the rotating section, or
4) Using a stone pre-heater, followed by a shorter rotary section.

Net heat usages of rotary kilns range from over 8370 kJ/kg (2000 kcal/kg) for the simplest gas-fired units, to about 5020 kJ/kg (1200 kcal/kg) for the more complex coal-fired installations.

One complication often associated with rotary kilns is the buildup of "rings" on the refractory material in the rotary section. They are produced by the combination of lime dust with clay, ash (if present), and sodium and potassium salts. They can be particularly troublesome in kilns fitted with preheaters and in coal-fired kilns. In the latter case, fine grinding of a well selected coal generally minimizes ring formation.

A particular advantage of rotary kilns without preheaters is that it is possible to eliminate most of the sulfur introduced with the fuel. This enables low-sulfur lime to be made with cheaper, high-sulfur fuels, subject to any limitations on sulfur dioxide emissions (see Section 5.4). Use of a preheater increases the contact between the combustion gases and the partially calcined stone, increases the absorption of sulfur dioxide, and reduces the ability to eliminate sulfur. Calcium sulfate deposits in the preheater can be troublesome.

Many lime producers have found that rotary kilns complement shaft kilns, because they use a smaller feedstone and produce quicklimes with different characteristics, which meet the requirements of certain customers.

Other Kilns. Many other designs of lime kiln have been developed, but few have gained wide acceptance. Recent noteworthy developments include the travelling grate (or CID) kiln and the top-shaped (or Chisaki) kiln, which accept limestone in the size ranges 1.5 – 4.5 cm and 0.5 – 4 cm, respectively [5]. A number of designs have been developed to calcine finely divided limestone. These include rotary kilns with suspension preheaters, flash calciners (to which a rotary kiln may be linked to facilitate control of the lime reactivity), and fluidized-bed kilns [5].

Dolime Kilns. Calcined dolomite is produced in both shaft and rotary kilns. Four qualities of calcined dolomite are produced: half-, light-, hard- and dead-burned. Half-burned dolomite ($CaCO_3 \cdot MgO$) is produced in small quantities in Germany using (it is believed) rotary kilns. Light-burned dolomite is produced in both rotary and shaft kilns. Hard-burned dolomite is generally produced in mixed-feed shaft kilns operating under reducing conditions. Two grades of dead-burned dolomite are produced. The high-purity grade, used for the manufacture of refractories, can be produced by direct

calcination of dolomite at up to 1800 °C in rotary or shaft kilns or by sintering pelletized, calcined dolomite at about 1800 °C, generally in rotary kilns. Fettling grade dolime is produced by heating a blend of finely divided calcined dolomite and iron oxide in a rotary kiln at 1400 – 1600 °C.

3.3.3. Quicklime Processing

The major demand is for the screened grades of quicklime in the range 4 to 0.6 cm. To obtain the maximum yield of those sizes, particularly from shaft kilns, it is generally necessary to crush the lime. Rolls and jaw crushers are often used because they minimize formation of fines. Surplus grades (generally − 0.6 cm) are often used as feed for the production of hydrated lime (see Chap. 4) or ground quicklime.

The production of finely ground quicklime has expanded markedly in recent years. As various grades are required, ranging from 30 to over 99 % passing 75 µm, they are generally produced in mills fitted with a variable speed classifier which returns coarse particles to the mill and controls the fineness of the finished product.

3.4. Uses and Specifications

Iron and Steel. A small amount of lime is used in the production of iron ore agglomerates from fines. The main advantage of adding 1 – 2 % of quicklime to the ore is a marked increase in the production capacity of the sinter strand. Ground quicklime is also used for the desulfurization of iron in the ladle before charging into the steelmaking furnace.

The major use of quicklime, however, is in the Basic Oxygen Steelmaking (BOS) Process. Its usage varies from about 30 – 50 kg/t of steel. The quicklime neutralizes the acidic oxides, SiO_2, Al_2O_3, and Fe_2O_3, to produce a basic molten slag. Correct formation of the slag is essential for the refining process.

A typical UK specification for steelmaking-quality quicklime is given in Table 2. The $CaCO_3$ content is limited to avoid excessive cooling of the melt through its decomposition to quicklime. The specifications for SiO_2 content and neutralizing value ensure that the quicklime has a high effective CaO content (1 % of SiO_2 neutralizes about 2.8 % of CaO). The grading and reactivity requirements ensure that the quicklime has an adequate surface area and reacts rapidly to produce the slag without excessive loss of fines from the vessel.

Table 2. Quicklime specification for BOS steelmaking

Specification	Criterion	Value
Sulfur	max. average	0.03 %
	maximum	0.04 %
Neutralizing value (CaO)	min. average	95.0 %
	at least 95 % greater than	93.0 %
Silica (SiO_2)	max. average	1.0 %
	maximum	1.5 %
Loss on ignition due to CO_2	max. average	2.0 %
	at least 95 % less than	2.5 %
	maximum	3.0 %
Reactivity (BS 6463)	minimum (at 2 minutes)	46 °C
Grading	nominal top size	38 mm
	nominal bottom size	5 mm
	max. amount passing 6 mm	5 %

Further information on the role of quicklime and the slag in removing phosphorus and sulfur from steel is given in the literature [13].

Calcined dolomite is also added to the BOS vessel to give a slag containing 6 – 8 % MgO, thereby reducing slag viscosity and attack of the basic refractory lining.

The electric arc steelmaking process uses quicklime to react with acidic oxides and produce a molten slag. The specification is similar to that for the BOS process.

Quicklime is also used in the argon oxygen decarburization (AOD) process, for which a −1.5 +0.6 cm product is required with less than 0.15 % C and 0.04 % S.

Calcium carbide (→ Calcium Carbide) is produced by reacting quicklime [14] with coke in furnaces which are heated electrically to 2000 °C. It is used to produce acetylene by reaction with water. The byproduct is an impure form of hydrated lime.

Aerated Concrete. Ground quicklime is used in the production of aerated concrete blocks, with densities of ca. 0.6 – 1.0 g/cm^3. The quicklime is mixed with an active form of silicon (either ground silica sand or pulverized fuel ash), sand, water, aluminum powder and,

depending on the quicklime quality, cement. The reaction of quicklime with aluminum powder generates hydrogen bubbles which cause the mix to rise. At the same time the quicklime reacts with the water, generating heat and causing the mix to set. Close control of the process results in a green set, which enables the semi-solidified mix to be removed from the mold and cut into blocks before autoclaving at elevated temperature and pressure.

The key aspect of the quicklime specification for this use is consistent reactivity (which includes a measure of the total free lime content). The product should have at least 90 % less than 75 µm and its MgO content should preferably be less than 2 %.

Soil Stabilization. Quicklime and hydrated lime can considerably increase the load carrying capacity of clay-containing soils. They do this by reacting with finely divided silica and alumina to produce calcium silicates and aluminates, which possess cementing properties.

Quicklime has the advantage over hydrated lime of drying out the soil. It does this by absorbing 30 % of its own weight of moisture and also by generating heat which accelerates evaporation.

Soil stabilization is used in road and rail construction to strengthen subgrades, thereby reducing construction depths. It may also be used to produce a sub-base in place of aggregate. It is used on clay-rich construction sites for placement and compaction of on-site material. In some countries lime piling is used to pin unstable soils and to provide support for building slabs.

The UK specification for quicklime used in soil stabilization is given in a Department of Transport publication [15].

Other Uses. Small quantities of quicklime are used in other processes [5], e.g., the production of glass, calcium aluminate cement, and organic chemicals. About 50 % of the total production, however, is converted to calcium hydroxide before use (see Section 4.4).

Specifications. Because of the dominance of the steel industry, a high proportion of producers select feedstone, kilns, and fuel which enable them to meet the Basic Oxygen Steelmaking specification (Table 2). Others operate to meet the requirements for building lime [8]. High reactivity is an advantage in soil stabilization. Consistent reactivity is essential in the aerated concrete process. The particle size requirements often depend on the customers' handling, conveying, and blending systems.

4. Hydrated and Slaked Lime

4.1. Physical and Chemical Properties

Color. Most hydrated limes are white. High levels of impurity can result in a gray or buff color.

Density. High-calcium hydrated lime has a density of about 2.24 g/cm^3. The values for partially and fully hydrated dolomitic lime are about 2.7 and 2.5 g/cm^3, respectively.

Bulk Density. The compacted bulk density is in the range 450 to 640 kg/m^3. In the as-poured state, it can be as low as 350 kg/m^3, owing to air entrainment.

Specific surface area may be measured by air permeability [7], or by surface absorption of nitrogen (BET surface area). The former method generally gives areas of 1000 – 2000 m^2/kg; the latter gives results which are some 20 to 30 times greater.

Angle of Repose. In the fluidized state, the angle of repose is 0°. In the compacted state, and particularly with 1 % or more of excess water, the angle of repose can be over 80°. With bunker aeration, valley angles of 70° have proved to be satisfactory.

Hardness. This is between 2 and 3 Mohs.

Specific Heat. The specific heat of calcium hydroxide rises from 1130 J kg^{-1} K^{-1} at 0 °C to 1550 J kg^{-1} K^{-1} at 400 °C [16]. That of dolomitic hydrate is believed to be about 5 % higher [1].

Solubility. This decreases from about 1.85 g $Ca(OH)_2$/L water at 0 °C to 0.71 g/L at 100 °C for commercial calcium hydroxide, depending on impurity levels [17].

Some inorganic compounds affect the solubility [18]. Calcium sulfate is of particular interest; a 2 g/L solution reduces the solubility to 0.06 g $Ca(OH)_2$/L. Organic compounds can increase the "solubility" of calcium hydroxide. Sugar has the greatest effect, as a result of the formation of calcium saccharate [18].

Magnesium hydroxide is only sparingly soluble in water (ca. 0.01 g/L) [19].

Carbon Dioxide. Calcium hydroxide reacts readily with carbon dioxide in the absence of water at temperatures below its dissociation point (ca. 540 °C). The reaction of quicklime with carbon dioxide below 500 °C only proceeds in the presence of some water vapor via formation of calcium hydroxide.

$$CaO + H_2O \longrightarrow Ca(OH)_2$$

$$Ca(OH)_2 + CO_2 \longrightarrow CaCO_3 + H_2O$$

Acid Neutralization. Hydrated lime, whether $Ca(OH)_2$, $Ca(OH)_2 \cdot MgO$, or $Ca(OH)_2 \cdot Mg(OH)_2$, reacts readily with acids and acidic gases. The rate of reaction depends in part on the particle size of the hydrated lime.

Silica and Alumina. Hydrated lime reacts with pozzolans (materials containing reactive silica and alumina) in the presence of water to produce hydrated calcium silicates and aluminates. The reactions may take months to proceed to completion at ambient temperatures (e.g., soil stabilization), but proceed within hours at elevated temperature and water vapor pressures (e.g., at 180 °C and a steam pressure of 1 MPa). This pozzolanic reaction is the basis of the strength generated by hydraulic quicklime.

pH. Because calcium hydroxide is a strong base, a concentration of 0.1 g $Ca(OH)_2$/L gives a pH of about 11.3 at 25 °C. A saturated solution at 25 °C, containing 1.8 g/L, gives a pH of 12.45. The pH of dolomitic hydrate is similar to that of calcium hydroxide.

Thermal Dissociation. Calcium hydroxide decomposes at about 550 °C to quicklime and water. The quicklime so produced has an exceptionally high reactivity to water.

Causticization. Hydrated lime reacts with soluble metal carbonates to produce insoluble calcium carbonate and the metal hydroxide. This reaction is used to produce caustic soda (see Section 4.4).

4.2. Raw Materials

4.2.1. Raw Materials for Hydrated Lime

Hydrated lime can be produced from any quicklime. In commercial practice, however, the quicklime quality should be within well-defined limits to enable the processing plant to be selected to produce salable products economically.

Reactivity. The reactivity of the quicklime should be compatible with the characteristics of the plant.

Thus highly reactive quicklime (over 55 °C, BS 6463) tends to produce a water-burned gritty hydrate with a particle size up to 3 mm. The grit is believed to be formed by very rapid hydration on the surface of a quicklime particle, leading to the production of a putty which bakes into a hard layer and holds the particle together. If the plant includes a mill which has the capacity to grind the product to substantially less than 75 µm, such grit does not present any quality problems.

Quicklime of intermediate reactivity (30–55 °C, BS 6463) generally hydrates well, providing it is given adequate residence time in the hydrator.

Quicklime of low reactivity (below 30 °C, BS 6463) also tends to produce a gritty hydrate, unless given a long residence time. This grit consists largely of particles of unreacted solid-burned lime with a coating of hydrate. This must be ground to substantially less than 75 µm and returned to the hydrator if a fully hydrated product is required. Any free lime eventually hydrates and expands. This can cause expansion in, for example, sand-lime bricks, and may result in cracking, surface defects, and unsoundness in, for instance, plaster.

Calcium Carbonate. The permissible $CaCO_3$ content of the feed quicklime depends on whether the plant is designed to reject a fraction rich in $CaCO_3$ and on the $CaCO_3$ specification for the finished hydrated lime. In some cases, with a hydrated lime specification of 2 % $CaCO_3$, feed quicklime $CaCO_3$ levels of up to

10 % can be accepted; in others the level must be below 2.6 %. The $CaCO_3$ is removed by screening or air classification.

Particle Size. The particle size of the quicklime fed to hydrators is generally below 5 cm and often as low as 0.6 cm, depending on the design of the plant.

Magnesium Oxide. As MgO is not readily hydrated, the normal hydration process does not convert it to the hydrate. Providing the MgO content is below about 2 %, the consequent unsoundness and expansion potential do not present problems. Higher MgO contents limit the markets into which the hydrated lime can be sold. Quicklimes with high MgO contents are often hydrated in the autoclave process, which produces a fully hydrated product with a high plasticity (see Section 4.3).

Other Impurities. Some applications limit the permissible levels of impurities in hydrated lime, which in turn place constraints on the impurities in the feed lime. These include lead, iron, silicon, and fluorine (see Section 4.4).

4.2.2. Raw Materials for Slaked Lime

In the slaking process, quicklime is reacted with a controlled excess of water to produce a milk or putty. The quicklime should be such that it gives the required quality of milk or putty and is compatible with the design of slaker used.

Reactivity. It is generally necessary to specify the reactivity to match the characteristics of the slaker and to obtain the required physical properties of the milk or putty.

Calcium Carbonate. Although many slakers remove grit rich in $CaCO_3$ by settling or filtration, their capacity to do so is limited. Therefore, the $CaCO_3$ content should also be limited. In other slakers, most or all of the $CaCO_3$ in the quicklime is included in the slaked lime; the $CaCO_3$ content should be compatible with the acceptable level in the slaked lime.

Particle Size. This is generally determined by the handling and processing equipment. In some cases a −2 cm quicklime is required. In others a ground quicklime is specified.

Magnesium Oxide. In applications where unsoundness or expansion potentials must be low, the MgO content in the quicklime should be below 2 %.

Water. The presence of impurities in water can have a marked effect on the reaction of quicklime with water in the reactivity test (see Section 6). However, water quality does not generally affect the industrial production of slaked lime.

Restrictions on other impurities in the quicklime are as for the production of dry hydrated lime.

4.3. Production

4.3.1. Normal Hydration Process

The normal hydration process, leading to the production of Type N hydrate, as defined in ASTM specification C-207 [2], is carried out at atmospheric pressure and ca. 100 °C. There are many designs of equipment, the choice of which depends on the reactivity and chemical purity of the quicklime, its particle size, and on the requirements for the finished product.

In a typical plant, the quicklime is mixed rapidly with about twice the stoichiometric amount of water in a premixer and is then passed into the main hydrating vessel, which is agitated and fitted with a weir at the discharge end to give an average solids residence time of 10 – 15 min.

As the hydration reaction proceeds, part of the water boils off vigorously, producing a partially fluidized bed. Coarse particles of gritty hydrate, unreacted lime, or unburned limestone are retained behind the weir, while fine particles of hydrate flow over it. In some plants a purge paddle close to the weir prevents excessive build up of coarse particles by lifting them over the weir. In others the coarse fraction is periodically rejected through a purge hole in the weir.

The raw hydrate is then conveyed along one or more tubes, which serve to complete the hydration and evaporate excess water. The hydrate is discharged at about 90 °C and contains less than 1 % excess water.

The raw hydrate is fed to an air classifier, in which a recycling air stream removes fine particles into an outer settling chamber and allows coarse particles to fall to the base of the inner feed chamber. The fine particles make up the finished product. The coarse particles may then be rejected as a stream rich in $CaCO_3$. Alternatively, they may be milled, with the product being recycled to the hydrator, added to the fines from the air classifier, or reclassified.

The steam generated in the hydrator, together with any air drawn into the process, may be dedusted in a wet scrubber before discharge to the atmosphere, in which case the milk of lime from the scrubber is fed to the premixer.

Some plants use aging silos to retain damp hydrate at elevated temperature for about 24 h to increase the extent of hydration. They are generally used when the feed quicklime has a low reactivity; with dolomitic hydrate; or when a product with a particularly low expansion potential is required.

4.3.2. Pressure Hydration Process

Type S hydrate, as defined in ASTM specification C-207 [2], is produced by reacting quicklime (generally dolomitic) and water under a steam pressure of up to 1 MPa and at a temperature of up to 180 °C. After hydration the product is dried, milled, and air classified.

4.3.3. Production of High Surface Area Hydrated Limes

Processes have been developed for the production of high-surface hydrated limes for injection into flue gases to remove acid gases [5]. One process hydrates the quicklime with an amine/glycol additive in a largely conventional plant [61]. Another uses a water/methanol mixture [62].

4.3.4. Normal Slaking Process

There are many ways of slaking quicklime, ranging from the batch process to sophisticated continuous slakers.

In the batch process, water and quicklime are added as required to maintain the milk at about 90 °C. With reactive quicklime, care should be taken to avoid local overheating which can lead to water-burning and a gritty hydrate (see Section 4.2.1). With cold water, this technique produces a viscous suspension with 35 – 40 % solids. Grit arising from uncalcined limestone and from unreacted and water-burned lime settles on the floor of the vessel.

There are various designs of continuous slakers, which produce either milk of lime or lime putty. Some feature automatic grit removal.

Slaking at ca. 90 °C produces the finely divided particles of calcium hydroxide required by most users. If the slaking temperature drops, e.g., to 70 °C, the resulting increase in particle size can reduce both the viscosity and reaction rate.

Aging the slaked lime for 30 min generally ensures complete hydration. Aging for a day generally improves the physical properties, e.g., higher viscosity and greater workability.

4.3.5. Production of Ultrafine Milks of Lime

Two processes have been developed for the production of ultrafine milks of lime [5]. Such milks dissolve very rapidly, can be processed more as liquids than suspensions, and can have beneficial rheological properties. They are produced by slaking at high shear rates [63] or by milling in a bead mill.

4.4. Uses and Specifications

4.4.1. Uses

In this Section, the term hydrated lime is used to describe both dry hydrate and slaked lime, unless otherwise specified.

There is little reliable information about the proportion of quicklime which is hydrated before use. The author's estimate for European countries is that it is roughly 50 %, of which about one-third is initially converted into dry hydrated lime and two-thirds are initially slaked to give milk of lime or lime putty.

Hydrated lime is used in a large number of processes. The more important ones are described briefly below. More detailed information is available in the literature [5].

The Solvay Process. A key step in the Solvay (or ammonia-soda) process for the production of soda ash and sodium hydrogen carbonate is the recovery of ammonia from an ammonium chloride solution by reaction with milk of lime.

$$2\,NH_4Cl + Ca(OH)_2 \longrightarrow 2\,NH_3 + CaCl_2 + H_2O$$

Causticization. Caustic soda can be produced by reacting sodium carbonate with milk of lime. Before the development of the electrolytic cell, this was the normal method of production of caustic soda.

$$Na_2CO_3 + Ca(OH)_2 \longrightarrow 2\,NaOH + CaCO_3$$

In addition to its use in the Bayer alumina process (see below) this reaction is used for the regeneration of caustic soda in wood pulp plants for the production of kraft or sulfate paper.

Nonferrous Metallurgy. Hydrated lime is used in the production of alumina by the Bayer process to regenerate sodium hydroxide from sodium carbonate solutions (see above). It is used in flotation processes to beneficiate copper ore and to extract gold and silver. It is an essential chemical for the extraction of uranium from gold slimes and for the recovery of nickel and tungsten after smelting.

Hydrated lime is used in the production of magnesia. In the seawater processes, a high-calcium hydrate is used to precipitate magnesium hydroxide, while in the Dow natural brine process, dolomitic hydrated lime is used.

Water Treatment. Hydrated lime is used to remove both carbonate (temporary) and non-carbonate (permanent) hardness. Calcium and magnesium hydrogen carbonates react with calcium hydroxide to produce insoluble calcium carbonate and magnesium hydroxide.

$$Ca(HCO_3)_2 + Ca(OH)_2 \longrightarrow 2\,CaCO_3 + 2\,H_2O$$

$$Mg(HCO_3)_2 + 2\,Ca(OH)_2 \longrightarrow Mg(OH)_2 + 2\,CaCO_3 + 2\,H_2O$$

Non-carbonate hardness caused by calcium and magnesium chlorides and sulfates can be removed by hydrated lime plus sodium carbonate.

$$CaSO_4 + Na_2CO_3 \longrightarrow CaCO_3 + Na_2SO_4$$

$$MgSO_4 + Ca(OH)_2 + Na_2CO_3 \longrightarrow Mg(OH)_2 + CaCO_3 + Na_2SO_4$$

Water can be purified by raising the pH to above 11 for 1 – 2 d, followed by recarbonation to pH 8 – 9. In addition to killing bacteria, this treatment removes temporary hardness. Acid water may be neutralized by the addition of lime. Specifications have been published for limes used in the United States for water treatment [14], [20]. Corresponding European standards are in preparation [48–50].

Sewage Treatment. Treatment of sewage sludge with hydrated lime in conjunction with ferrous sulfate is an effective way of removing solids and phosphorus compounds, and for destroying pathogens. It produces a sludge which dewaters well and can be used as a fertilizer.

Industrial Wastes. Hydrated lime is used widely to neutralize acid wastes and precipitate heavy metals in effluents from a wide range of industries. In some cases it is also used to assist with the clarification process.

Flue Gas Desulfurization. Of the many flue gas desulfurization processes, three use hydrated lime.

In the wet scrubbing process, milk of lime is sprayed through the gases. The product is a suspension of calcium sulfite, which may be oxidized to produce salable gypsum. Up to 95 % removal of sulfur dioxide has been reported, with over 90 % utilization of the hydrated lime [21].

In the dry scrubbing process, milk of lime is fed into a spray drier from which the product is a dry powder consisting of calcium sulfite and hydrated lime. Up to 85 % removal of sulfur dioxide and 40 – 60 % utilization of the hydrated lime have been obtained [21].

Sulfur dioxide may also be removed by injection of dry hydrate either into the boiler, or downstream from the boiler. The reaction products are collected with the fly ash for disposal. Only 50 – 60 % of the sulfur dioxide is removed, and the utilization of normal hydrated lime is

less than 40 % [21]. Hydrated lime products with higher specific surface areas than those of normal hydrated lime have been developed for this use.

Sugar Refining. In the refining of sugar beet, the crude solution of sugar is treated with milk of lime to precipitate calcium salts of organic and phosphoric acids. After filtering, the solution is neutralized with carbon dioxide, calcium carbonate is removed by filtration, and a purified sugar solution is produced.

As the process requires large quantities of both lime (about 250 kg/t of sugar) and carbon dioxide, producers of sugar from beet generally operate their own lime kilns as part of their process.

Because cane sugar is purer, its refinement requires much less hydrated lime (typically 5 kg/t).

Sand – Lime Bricks. There is a substantial market in Europe for sand – lime bricks. These are made by mixing hydrated lime with sand, followed by autoclaving for several hours under a steam pressure of ca. 1 MPa (180 °C). In some processes, it is essential that the hydrated lime is not expansive, because the expansion can occur in the autoclave and result in the production of oversize bricks.

Mortars. Hydrated lime is used in lime – cement – sand mortars in various proportions. The required compressive strength of the mortar is obtained by adjusting its composition [51]. The lime confers several benefits. It increases the plasticity and water retentivity of the wet mortar and increases the bond strength of the mortar to the masonry. The mortar is more flexible and less prone to cracking.

Soil Stabilization. Both dry hydrated lime and milk of lime are used in soil stabilization. Dry hydrate disperses readily when making a soil tilth. Milk of lime has advantages when the soil is dry; it can also be injected under pressure.

Agriculture. The advantages of adjusting soil pH with limestone are described in Section 2.4. Hydrated lime has the additional advantage of being quick acting.

Bleaching. Chlorine is reacted with hydrated lime to produce a powdered mixture of calcium hypochlorite and calcium chloride, commonly known as bleaching powder.

$$2\,Ca(OH)_2 + 2\,Cl_2 \longrightarrow Ca(OCl)_2 + CaCl_2 + 2\,H_2O$$

Calcium hypochlorite solution is used in bleaching wood pulp.

Precipitated calcium carbonate (PCC) is produced by blowing gases containing carbon dioxide through milk of lime. By controlling the conditions, a very finely divided calcium carbonate with a high reflectivity and a median particle size of 0.02 – 0.2 µm is produced. Some PCC's are coated with compounds which facilitate their blending with organic substancs (e.g., plastics and rubber).

Inorganic Salts and Bases. Hydrated lime is used in the production of mono-, di-, and tricalcium phosphates as well as other calcium salts. It is used in the Solvay process to produce calcium chloride as a byproduct. Lithium and barium hydroxides are made by causticization.

The brine purification process uses sodium carbonate and milk of lime to remove dissolved calcium and magnesium.

Other Uses. There are a wide range of other uses such as the production of calcium silicate insulation products, plasters, lubricants, pigments, soda lime, organic compounds and calcium salts. Hydrated lime is also used in the petroleum industry, glass manufacture, solid waste disposal, and leather tanning [5].

Table 3. Typical specification for high-calcium hydrated lime [8]

Specification	Criterion	Value
Carbon dioxide	maximum	12 % ∗
CaO + MgO	minimum	70 % ∗
MgO	maximum	5 % ∗
Residue on 180 µm	maximum	2 %
90 µm	maximum	7 %
Soundness		
expansion	maximum	20 mm
pat test	pops or pits	nil
Excess moisture	maximum	2 %

∗ Based on a water-free and bound-water-free product.

4.4.2. Specifications

In many countries, the most widely recognized specifications for dry hydrated lime relate to its use in building [2], [8]. These set minimum values for fineness, neutralizing value, and other properties. In practice, most dry hydrated limes meet these requirements with ease. A typical quality specification is given in Table 3.

Some end uses require especially low levels of impurities such as silica, iron oxide, magnesium oxide, and fluoride. Others require uniform surface areas.

Dry hydrated lime for use in the production of sand – lime bricks should have an expansion below 10 mm [22] and less than 4 % MgO (unless it is Type S).

5. Environmental Protection

5.1. General

The main environmental issues associated with the production of lime and limestone products are air pollution and energy consumption; the latter is described in Section 3.3.2. Dust emissions can arise from the crushing and handling of limestone and the burning, processing, and handling of quicklime and hydrated lime. In certain circumstances, the emission of sulfur dioxide and oxides of nitrogen can be an issue for rotary kilns, but that is rarely the case for shaft kilns. These and related topics are addressed in some detail in [23], [24].

5.2. Dust Emission

Standards. In many countries, emissions of dust to the atmosphere for new equipment are limited to 50 mg of solids per cubic meter (at 0 °C and 101 MPa).

Equipment. A variety of techniques are used to control dust emissions from stone processing operations. Dust suppression, which uses water with a wetting agent, causes dust to adhere to larger particles. It can be an effective and inexpensive method but can cause undesirable surface contamination of screened products with fines and increase downstream dust problems.

As the effluent gases from stone processing and local exhaust ventilation systems are generally at relatively low temperatures, bag filters or high-energy wet scrubbers are widely used. The dust collected by bag filters can often be sold as a filler for concrete production.

The temperature of the effluent gases from lime kilns is commonly 200 – 400 °C. In the past, high efficiency multi-cyclones were widely used to collect the dust. They had the advantage of being relatively inexpensive to install and operate. However, their efficiency is generally no longer adequate in the context of current emission requirements.

Electrostatic precipitators can handle gases at the above temperatures. Although their capital cost is high, their operating costs are moderate.

Gravel bed filters can be operated at high temperatures. They can be efficient but have higher capital and operating costs than electrostatic precipitators. Consequently, they are not widely used.

Bag filters are widely used when the gas temperatures are below 200 °C. They are lower in capital cost than precipitators, have a high collection efficiency, but have high operating costs. The reverse-jet method of cleaning the bags has generally been found to be the most effective.

The dust collected from lime kilns is generally high in $CaCO_3$ but also contains quicklime and possibly fuel ash. In some processes it is blended in a controlled manner into selected products; in others it is used for landfill

Wet collectors are not widely used with quicklime. They are, however, generally the preferred method of controlling dust from hydrating plants. The collected milk of lime is fed back to the process.

Dust collection from the handling, storage and loading of quicklime and hydrated lime generally involves bag filters. The collected dust usually has a high lime content and is sold.

5.3. Gaseous Emissions

In most lime kilns, *sulfur dioxide* emissions are acceptably low [e.g., < 500 mg SO_2/m^3 (STP)]. This is because the quicklime captures and retains a high proportion of the sulfur from the fuel and limestone. However, when low-sulfur quicklime is produced in rotary kilns, and hard-burned

calcium lime/dead-burned dolomite is produced in shaft or rotary kilns, most of the sulfur may be emitted with the exhaust gases. In such situations, there may be a need to abate the emission of sulfur dioxide. The only proven sulfur dioxide abatement technique for lime kilns is to use a fuel with a lower sulfur content. Sulfur dioxide abatement techniques used in other industries include the injection of calcium hydroxide into the exhaust gases. As yet, however, their application to lime kilns is unproven.

Most shaft kilns emit acceptably low concentrations of *nitrogen oxides* [e.g., < 500 mg NO_2/m^3 (STP)] when producing light- and medium-burned quicklime. However, when they are used to produce hard-burned calcium lime and dead-burned dolomite, higher emissions result. Emissions from rotary kilns producing soft-burned lime are generally acceptable, but increase to > 800 mg NO_2/m^3 (STP) when medium-, hard-, and dead-burned products are made. Currently (1999), there are no proven abatement techniques for lime kilns. Techniques used in other industries include low-NOx burners, selective noncatalytic reduction, and selective catalytic reduction, but their applicability to lime burning has yet to be established.

Significant emissions from lime kilns of *other substances*, such as volatile organic compounds, dioxins, and furans, are only likely to occur under exceptional circumstances [24]. There is no evidence of significant emissions of heavy metals.

5.4. Noise and Vibration

Blasting can produce powerful impulsive sound waves which can startle neighbors and cause buildings to vibrate. The sound level can be focused under certain atmospheric conditions, particularly temperature inversions. Advice has been published on good operating practice to mitigate this effect [25].

Mobile equipment, such as drills, mechanical shovels, and dump trucks, can produce unacceptable noise levels if not adequately silenced. A major problem with fixed plant is the sound produced by the impact of stone on metalwork. Appropriate control measures should be adopted [26].

Increasingly, exhaust fans associated with lime kilns are being designed to be more efficient and to generate more intense suction. These trends can increase the pure-note component generated within the fan and often necessitate the fitting of silencers. The problem of silencing the fans is complicated by the dust present in the gas; reactive silencers have proved to require less maintenance than absorptive designs.

The major source of vibration is blasting. The measurement and control of vibration is well understood, and there are established guidelines relating to complaint levels [27] and to the risk of damage to property [28].

6. Physical Testing and Chemical Analysis

6.1. Sampling and Sample Preparation

The most common cause of disagreement between laboratories arises from the failure to obtain representative samples. Sampling solids, and particularly granular solids made from naturally occurring minerals, is especially difficult. The problem is compounded when, as in the case of quicklime, the quality of the sample is affected by exposure to the atmosphere. In general, if quicklime contains 1 % or more of combined water, it has not been correctly handled, and its reactivity to water will have been reduced.

Guidance on the sampling of limestone, quicklime, and hydrated lime is given in BS 6463, Part 101 [7]. Similar advice is given in ASTM C 110 and C 50 [9], [30]. The sampling of aggregate and the reproducibility and repeatability of test results is described in BS 812 [31].

Samples should be taken from a falling stream whenever possible. Precautions should be taken to ensure that segregation within the stream does not lead to bias. Samples may also be taken from conveyor belts; precautions are required to obtain a representative sample. It is very difficult to obtain a representative sample from the body of a truck or from a stockpile. Careful checks should therefore be made to establish whether the sampling method is valid.

Each sample should consist of not less than ten increments. Each increment should be of an adequate volume (e.g., 10 L for lumps with a maximum size of 10 cm, decreasing to 2 L for

a maximum size of 1 cm, and to 1 L for powders). There are many mechanical aids to sampling ranging from augers to automatic samplers [32].

Some physical tests are performed on the sample as taken. For other physical tests and for chemical analysis, it is necessary to crush, subdivide and pulverize the product to less than 300 µm. This should be done on a separate sample in such a way as to avoid bias and, in the case of quicklime and hydrated lime, to avoid excessive exposure to the atmosphere.

All quicklime and hydrated lime samples should be stored in well sealed containers or bags. It is good practice to double-wrap samples which are to be stored for prolonged periods. Quicklime should not be stored in a desiccator because it removes water from some commonly used desiccants.

6.2. Physical Testing

Limestone. A series of CEN standard physical test methods for aggregates is being prepared. It consists of four standards (in a total of 32 parts) [52–55].

Quicklime and Hydrated Lime. Physical test methods for building lime are given in EN 459 [8]. Additional test methods (residue on slaking, specific surface area by air permeability, and workability of putty) are specified in BS 6463 Part 103 [7]. The corresponding US standard is ASTM C110 [9].

6.3. Chemical Testing and Analysis

Tests for the chemical properties of aggregates are given in EN 1744 [22].

The methods for chemical analysis of limestone, quicklime, and hydrated lime are closely related. The relevant standard for building limes is EN 459-2 [8], which adopts five methods from EN 196-2 [6]. The standard for products used in the preparation of water for human consumption, EN 12485 [12], describes a number of tests, including determinations of minor and trace elements. The British Standard BS 6463, Part 102 [7] specifies additional methods (including minor and trace element determinations) and provides cross-references to the EN standards. Test methods are specified in ASTM C 1271, 1301 and 1318 [56–58].

7. Storage and Transportation

Limestone. Screened sizes of limestone are stored in bunkers and in outside stockpiles. Fine grades are generally stored in sealed bunkers.

As most screened limestone is distributed to a large number of sites, generally within 40–80 km of the quarry, it is mainly transported in sheeted tipper road vehicles. Use of rail or water transport becomes more economic as the volume of business to a given site increases and as the distance becomes greater. It has been estimated that over 75 % of limestone is transported by road [1]. Fine grades of limestone are transported by air pressure discharge vehicles, generally by road, but suitable railroad cars are also available.

Quicklime. Quicklime is stored in enclosed bunkers with well-sealed discharge mechanisms to minimise the reaction with atmospheric water and carbon dioxide and to control dust emission. Storage capacities equivalent to 2 to 5 days' production are usually necessary.

Screened grades of quicklime (generally 0.5 cm and above) may be transported in tipper trucks. Effective sheeting is essential to minimize atmospheric attack on the quicklime and to control dust emission. Air pressure discharge vehicles are used to transport fine grades of quicklime (below 0.5 cm) and, increasingly, screened grades up to 2 cm. Where appropriate, use of rail or water transport can extend the economic delivery distance.

A small amount of quicklime is packaged. Plastic containers and sacks are used for amounts up to 50 kg. Intermediate bulk containers, both rigid and flexible, are being used increasingly for quantities up to 1 t. There is a small international market for packaged quicklime.

Where totally enclosed bunkers cannot be provided, quicklime should be stored on a concrete base, preferably in a separate bay within a building. It should be used as soon as possible after delivery. It can be kept longer by covering the pile with a close-fitting impermeable

Table 4. Main uses of limestone (including dolomite and chalk)

Use	USA (1994)	Germany (1996)	Japan (1994)	UK (1996)
Market share, %				
Construction and building *	69	35	30	73
Cement production	11	51	47	14
Agriculture	2	2	1	3
Flux stone **	3	5	11	3
Environmental protection	1	2	2	1
Other	14	5	9	6
Total tonnage, 10^6 t/a	788	65	208	108

* Mainly concrete and roadstone; excluding cement. ** Mainly for iron and steel production.

sheet. Open reception hoppers should be protected from the elements.

Contact of quicklime with flammable materials should be avoided, especially when there is a risk of water penetration. In the presence of organic matter, the heat of hydration has been known to cause fires.

Further details on the delivery, storage and handling of quicklime are often available from the supplier (see also [60]).

Hydrated Lime. Hydrated lime is stored in enclosed silos to minimize reaction with atmospheric carbon dioxide and to control dust emission.

Most hydrated lime is transported in air pressure discharge vehicles, both by road and rail. A substantial amount, however, is sold in paper sacks (generally 25 kg in the United Kingdom) Use of intermediate bulk containers holding up to 1 t is increasing. There is a significant international market for packaged hydrate.

The flow characteristics of hydrated lime are very variable. The angle of repose varies from 0 – 80° according to the amounts of entrained air and excess water, and other factors. Silos may need to be fitted with devices to cope with bridging of the powder above discharge points and to encourage flow (e.g., bin activators and air pads).

Hydrated lime in paper sacks is generally placed on pallets and stored under cover. By fitting impermeable slip sheets under the bags and shrink wrapping a cover over the loaded pallet, it may be stored out of doors for several months.

Further details on the delivery, storage and handling of hydrated lime are often available from the supplier (see also [60]).

Table 5. Sales of limestone in the United Kingdom in 1996 (in 10^6 t/a)

Use	Dense, high-calcium limestone	Dolomite	Chalk	Total
Construction	64	14	1	79
Cement	10	0	6	15
Agriculture	1	1	1	3
Others *	7	2	1	11
Total	82	17	9	108

* Limeburning, iron, steel, chemical and other uses.

8. Economic Aspects

Limestone is a major mineral product. Statistics on its production are not widely available and are often complicated by the inclusion of other crushed rocks. It is estimated that the worldwide production is over 4500×10^6 t/a. Table 4 gives tonnages for sales in the United States, Germany, Japan, and the United Kingdom, and the distribution of those sales between the major market segments. The high percentages of limestone used in construction and building in the United States and the United Kingdom reflects the relatively high availability of limestone in those countries when compared with Germany and Japan.

The physical and chemical qualities of limestones affect the applications for which they are used. Table 5 summarizes the quantities of dense, high-calcium limestone, dolomite, and chalk consumed in the UK in the main market segments [41]. Ex-works prices for limestone aggregate are in the range £ 2 to £ 5/t (1999). Chemical-quality limestone for lime burning, glass manufacture, and flue gas desulfurization commands prices in the range £ 5 to £ 10/t.

Ground limestone products (or whiting) typically sell for £ 25 to £250/t, depending on the fineness and quality. Precipitated calcium carbonate (PCC) is a specialist chemical which can command prices of £ 250 to £ 1000/t.

Quicklime and Hydrated Lime. Published estimates of the global production of lime products [56], [59] suggest a volume of ca. 120×10^6 t/a. The major producing countries are listed in Table 6. The published figures do not include the many captive producers and the multitude of small producers in the developing countries. It has been estimated that if their contribution were included the total would probably exceed 300×10^6 t/a.

Table 6. Estimated Lime Production in Various Countries (1994)

Country	Total production *
Australia	1 250
Austria	700
Belgium	1 750
Brazil	5 700
China **	19 500
Czech Republic	1 200
Denmark	130
Finland	300
France	3 100
Germany	8 500
Greece	500
Ireland	100
Italy	3 500
Japan	7 700
Mexico	6 500
New Zealand	110
Norway	80
Poland	2 500
Portugal	200
Romania	1 600
Sweden	470
Slovak Republic	730
Spain	1 000
South Africa	1 900
Sweden	500
Turkey	4 200
USA	17 400
UK	2 500
Former USSR	16 000
Others (by difference)	8 380
Estimated world total	118 000

* The above estimates are based on published values for open-market production [56], [57], [59].
** Estimates of over 100×10^6 t for China have been reported [44], but not substantiated.

Table 7 summarizes the use of lime products in the various market segments in nine countries [56–58].

The ex-works price of screened high-calcium quicklime products in Europe is generally in the range £ 40 to £ 60/t (1999). In the United States, the average ex-works price for screened high-calcium quicklime was $ 57/t (1998) [59]. Ground quicklime and hydrated lime prices are typically some 30 % higher than those of screened quicklime, reflecting the capital and operating costs of the additional processing stages (e.g., in the United States, the average 1998 price of high-calcium hydrated lime was $ 75/t). Specialist products, tailored to the needs of individual market segments or customers, also tend to command higher prices.

9. Toxicology and Occupational Health

9.1. Toxicology

Limestone is a practically nonharmful material, providing its content of crystalline silica is less than 1 %.

As an air-borne dust, its long term exposure limit (8-h TWA) is 10 mg/m^3 of total inhalable dust and 5 mg/m^3 of respirable dust [42]. Lower limits apply if the content of crystalline silica exceeds 1 %.

As with other nuisance dusts, limestone may irritate the eyes and cause discomfort. It can also cause discomfort as a result of its drying effect on the mouth and upper respiratory tract. It is nonirritant to the skin.

Limestone is used in the treatment of water for human consumption. In such applications, limits may be applied to the minor and trace element contents [50]. Calcium carbonate is a recognised ingredient of human and animal foodstuffs. It is used as an antacid. The therapeutic dose is 1 – 5 g. Chronic effects are unlikely to be encountered in industry. There are no occupational diseases known to be connected with the handling of limestone.

Quicklime and hydrated lime are alkaline in the presence of water (pH 12.4). In addition, when quicklime comes into contact with water, it generates a considerable amount of heat.

As airborne dusts, the long-term exposure limits (8-h TWA) for quicklime and hydrated lime are 2 and 5 mg/m^3, respectively [42]. The

Table 7. Percentage of lime products in the main market segments (1994)

Country	Iron and steel	Non-ferrous metals	Chemical industry	Other industry	Building materials [a]	Building trade [b]	Environmental protection [c]	Agriculture and food
Belgium	70		7	1	6	7	8	1
Germany	36		7	2	27	12	14	2
France	58	3	4				20	15
Italy	27		2			65	5	1
Japan	54	1	23	3	2	5	9	3
S. Africa	42	30	16	4		7	1	
Turkey	25	1	12	7	1	53	1	
USA	31	7	5	24		8	25	
UK	55		5	5	8	7	19	1

[a] Mainly aircrete, sand – lime bricks, other calcium silicate products and refractories.
[b] Mainly mortar, render, plaster, and drying/improvement/stabilization of soils.
[c] Mainly potable water, sewage, liquid and gaseous effluent treatment.

dusts are irritating to the respiratory tract and may cause inflammation. Contact of lime with the eyes can cause painful irritation and may result in serious damage unless immediate treatment is given.

Both quicklime and hydrated lime are classified as irritant and can cause "chemical burns" of the skin, when subject to abrasion in the presence of moisture, or perspiration. Under the EEC Dangerous Preparations Directive, both products are classified by risk phrases R38 (irritating to skin) and R41 (risk of serious damage to eyes). Prolonged and repeated contact may cause the skin to become dry and cracked, and may lead to dermatitis. Ingestion can cause corrosion and damage to the gastrointestinal tract.

Both quick- and hydrated lime are used in the treatment of drinking water. In such cases, limits may be applied to the minor and trace element contents [48], [49].

9.2. Occupational Health

Limestone. The quarrying industry in general has a poor safety record relative to general manufacturing [43]. In the United Kingdom, mobile equipment (shovels, dump trucks, etc.) is the major cause of deaths. The second largest cause of deaths and over half of the major injuries is stumbling, falling, and slipping. While the use of explosives is generally perceived to be a major hazard, it causes less than 5 % of the UK deaths and major injuries, presumably because it is well covered by legislation.

In limestone mines, rock falls are the major cause of deaths and serious injuries, and mobile equipment is the next most significant category. Special electrical precautions (e.g., high-integrity earthing and the use of low-voltage portable tools) are essential, as water percolates into most limestone mines [45].

Stone processing involves a number of mechanical operations. The largest category of injuries caused by these generally arises from inadequate guarding of machinery and from failure to electrically isolate equipment when carrying out maintenance and clearing blockages [44].

Many stone-processing operations produce high sound levels. Operators' exposure to noise is controlled by a combination of reduction and containment at source, excluding it from control rooms, and the use of remote cameras with monitors in control rooms. Effective personal hearing protection is still required when operators are required to enter noisy areas.

Ways of controlling in-plant dust arising from quarrying and stone processing are mentioned in Section 5.2.

Quicklime and Hydrated Lime. The lime industry generally has a lower accident rate than quarrying. Contributory factors are likely to be the reduced use of mobile equipment and the greater size of the works (there is a correlation between accident rate and the number of people employed on site [47]). Tripping and falling constitute the largest category of accidents.

Lime processing involves a number of mechanical operations. Adequate guarding of

moving machinery and effective isolation procedures when carrying out maintenance are essential.

Both quicklime and hydrated lime are dusty, alkaline products. Not surprisingly, a common injury arises from grit or dust entering the eye. It is advisable to wear eye protection at all times in lime plants. Techniques to control in-plant dust are mentioned in Section 5.2.

10. References

1. R. S. Boynton: *Chemistry and Technology of Lime and Limestone,* John Wiley, New York 1980.
2. ASTM C-207 Specification for Hydrated Lime for Masonry Purposes, American Society for Sampling and Testing.
3. J. Murray: "Specific Heat Data for Evaluation of Lime Kiln Performance", *Rock Prod.,* Aug. 1947, 148.
4. H. Bäckström, *J. Am. Chem. Soc.* **47** (1925) 2432, 2443.
5. J. A. H. Oates: *Lime and Limestone – Chemistry, and Technology, Production and Uses,* Wiley-VCH, Weinheim, 1998.
6. EN 196: Methods of Testing Cement.
7. BS 6463: Quicklime, Hydrated Lime and Natural Calcium Carbonate.
8. EN 459: Building Lime.
9. ASTM C 110: Methods of Physical Testing of Quicklime, Hydrated Lime and Limestone.
10. N. E. Rogers, *Cem. Lime Gravel,* June 1970, 149 – 153.
11. A. B. Searle: *Limestone and Its Products,* E. Benn, London 1935.
12. EN 12485: Chemicals Used for Treatment of Water Intended for Human Consumption – Calcium Carbonate, High-Calcium Lime and Half Burnt Dolomite – Test Methods (in preparation).
13. A. Jackson: *Oxygen Steelmaking for Steelmakers,* Butterworths, London 1969.
14. ASTM C 911: Specification for Quicklime, Hydrated Lime and Limestone for Chemical Uses.
15. Specification for Highway Works, Department of Transport, HMSO, London.
16. International Critical Tables, 5, McGraw Hill, New York 1929, pp. 98 – 99.
17. R. Haslam et al., *J. Am. Chem. Soc.* **46** (1924) 308.
18. A. Seidell: *Solubilities of Inorganic and Metal Organic Compounds,* Van Nostrand, Princeton 1965, pp. 309 –310.
19. N. Knibbs, *Lime and Magnesia,* E. Benn, London 1924, p. 71.
20. AWWA Standard B 202 – 83 American National Standards Institute.
21. D. H. Stowe, Lime for F. G. D., 6th International Lime Congress, 1986, International Lime Association.
22. EN 1744: Tests for Chemical Properties of Aggregates.
23. IPC Guidance note PG3/8(96): *Secretary of State's Guidance – Quarry Processes*, HMSO, London.
24. *Reference Document on Best Available Techniques in the Cement and Lime Industries* (in preparation), produced under the requirements of the European Union's Integrated Pollution Prevention and Control legislation.
25. BACMI Briefing No. 120, Feb. 1987, British Aggregate and Construction Materials Industries.
26. BS 5228, Parts 1 & 3 (1984): Noise Control on Construction and Open Sites.
27. BS 6472 (1984): Guide to Evaluation of Human Exposure to Vibration in Buildings (1 Hz to 80 Hz).
28. S. Stagg, D. Siskind, Bureau of Mines Information Circular, IC 9135, 1987, United States Department of the Interior, Washington.
29. ASTM C 1271: Methods for X-ray Spectrometric Analysis of Lime and Limestone.
30. ASTM C 50: Test Methods for Sampling, Inspection, Packing and Marking of Lime and Limestone Products.
31. BS 812 (1975): Testing Aggregate, Parts 1, 2 & 3.
32. BS 1017 (1977): Methods for Sampling of Coal and Coke, Part 1.
33. BS 822 (1983): Specification for Aggregates from Natural Sources for Concrete.
34. BS 63 (1987): Road Aggregates.
35. P. G. Roe, D. C. Webster, Road Research Laboratory, Scientific Report, 829, HMSO, London 1984.
36. BS 4987 (1988): Coated Macadam for Roads and Other Paved Areas.
37. ASTM C-25: Method for Chemical Analysis of Limestone, Quicklime and Hydrated Lime.
38. ASTM C 1301: Test Methods for Major and Trace Elements in Limestone and Lime by Inductively Coupled Plasma – Atomic

Emission Spectroscopy (ICP) and Atomic Absorption (AA).
39. ASTM C 1318: Test Method for Determination of Total Neutralising Capability and Dissolved Calcium and Magnesium Oxide in Lime for Flue Gas Desulfurisation (FGD).
40. Mineral Industry Survey – Crushed Stone – Annual Review 1994, the US Department of the Interior, Bureau of Mines.
41. U.K. Business Monitor, PA 1007, 1996, HMSO.
42. Occupational Exposure Limits, 1996, EH40/96, HSE Books, London.
43. Health and Safety Commission and Health and Safety Executive Reports from 1985/6 and 1986/7 and HSE "Quarries" Report for 1984/5, HMSO.
44. G. Chenxiang, "The Structural Characteristics and Rationality of China New Type Ordinary Lime Shaft Kiln," *Proc. International Lime Congress*, Berlin. 1994.
45. Health and Safety Executive Report "Mines"1985/6, HMSO.
46. J. M. Hobbs, *Safety in Quarries* (Part 1), private publication, 1988.
47. H. Krummhaar, "The Progress of Occupational Safety in the German Lime Industry", *Zement Kalk Gips* **1** (1989) 37 – 40.
48. EN 12518: Chemicals Used for Treatment of Water Intended for Human Consumption – High Calcium Lime (in preparation).
49. EN 1017: Chemicals Used for Treatment of Water Intended for Human Consumption – Half-Burnt Dolomite (in preparation).
50. EN 1018: Chemicals Used for Treatment of Water Intended for Human Consumption – Calcium Carbonate (in preparation).
51. EN 998: Specification for Mortar for Masonry.
52. EN 932: Tests for General Properties of Aggregates.
53. EN 993: Tests for Geometric Properties of Aggregates.
54. EN 1097: Tests for Mechanical and Physical Properties of Aggregates.
55. EN 1367: Tests for Thermal and Weathering Properties of Aggregates.
56. International Lime Association: Statistics, 1994.
57. European Minerals Yearbook, 1995.
58. Statistical Year Book, 1997, Quarry Products Association, London.
59. *Mineral Industry Survey – Lime – Annual Review 1998*, U.S. Department of the Interior, Bureau of Mines.
60. *Lime Handling, Application and Storage*, Bulletin 215, 5th ed., National Lime Association, 1988.
61. Lhoist Recherche et Developpement, US 5 173 279, 1992 (P. A. Dumont, R. Goffin)
62. Rheinische Kalksteinwerke, US 4 636 379, 1987 (H. Bestek et al.)
63. H. Becker, "Highly Reactive Calcium Hydroxide Suspensions for Industrial Chemical Processes – Production and Properties", *Chem. Ing. Tech.* **59** (1987) no. 3, 228 – 235.

Lining of Metals → Metals, Surface Treatment
Lipid-Lowering Agents → Cardiovascular Drugs
Lipsticks → Skin Cosmetics
Liquefied Natural Gas → Natural Gas

Liquefied Petroleum Gas

Stephen M. Thompson, Shell Internationale Petroleum Maatschappij B.V., Den Haag, The Netherlands
Gary Robertson, Shell Internationale Oil Products B.V., Den Haag, The Netherlands

1. Introduction 1
2. Properties 2
3. Production and Processing 2
3.1. Recontacting – Compression 3
3.2. Refrigeration 3
3.2.1. Low-Temperature Separation (LTS) . 4
3.2.2. Expander Plants 5
3.2.3. Combined Processes 6
3.3. Lean Oil Absorption 7
3.4. Adsorption 8
4. Purification 8
5. Storage and Transportation 10
6. Uses 12
7. Safety Aspects 12
8. Economic Aspects 12
9. References 14

1. Introduction

The term liquefied petroleum gas (LPG) describes hydrocarbon mixtures in which the main components are propane, butane, isobutane, propene, and butenes. Most commonly this term is applied to mixtures of propane and butane. These components and mixtures thereof are gaseous at normal temperature and pressure but can be liquefied by cooling, compression, or a combination of both processes.

Liquefied petroleum gas is produced from two distinct sources. The first is by *extraction* from crude oil and natural gas streams at or close to the point of production from the reservoir and contains propane and butane. The quantities of LPG in the well stream fluid vary greatly, depending on the nature of the reservoir. Production may be (1) of natural gas from a gas reservoir, (2) of gas and light liquid hydrocarbons from a gas condensate reservoir or (3) of crude oil and gas from a combined oil and gas field. The extent of recovery of LPG and heavier hydrocarbons from gas depends on the composition of the gas produced and the quality specifications of the gas to be transported to the consumer. Before storage or transportation of crude oil by tanker, its vapor pressure must be lowered so that it can safely be contained in a ship's tank. This reduction, known as stabilization, is achieved by the removal of LPG and lighter species to yield stabilized crude oil and additional natural gas (see Fig. 1). The LPG components in produced oil and gas are predominantly the saturated hydrocarbons propane, butane, and isobutane.

Second, LPG is obtained by *processing* of crude oil in refineries and in some instances as a byproduct from chemical plants. The remaining LPG in the stabilized crude oil at the refinery is separated in the crude oil fractionation column. The constituents of this LPG are propane, butane, and isobutane. Additional LPG is produced in refinery conversion processes such as catalytic reforming, thermal cracking, catalytic cracking, and hydrocracking (see Fig. 2). The composition of this LPG depends on the processing configuration at the refinery but typically includes both saturated (propane, butane, and isobutane) and unsaturated (propene and butenes) components.

Table 1. Physical properties of the main LPG components [4]

	Boiling point (101.3 kPa), °C	Vapor pressure (37.8 °C), kPa	Liquid density (at saturation pressure; 15.6 °C), kg/m^3	Gross calorific value (25 °C), kJ/kg
Propane	− 42.1	1310	506.0	50 014
Propene	− 47.7	1561	520.4	48 954∗
n-Butane	− 0.5	356	583.0	49 155
Isobutane	− 11.8	498	561.5	49 051
1-Butene	− 6.3	435	599.6	48 092
cis-2-Butene	3.7	314	625.4	47 941
trans-2-Butene	0.9	343	608.2	47 878
Isobutene	− 6.9	435	600.5	47 786

∗ Calorific value at 15 °C.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_347

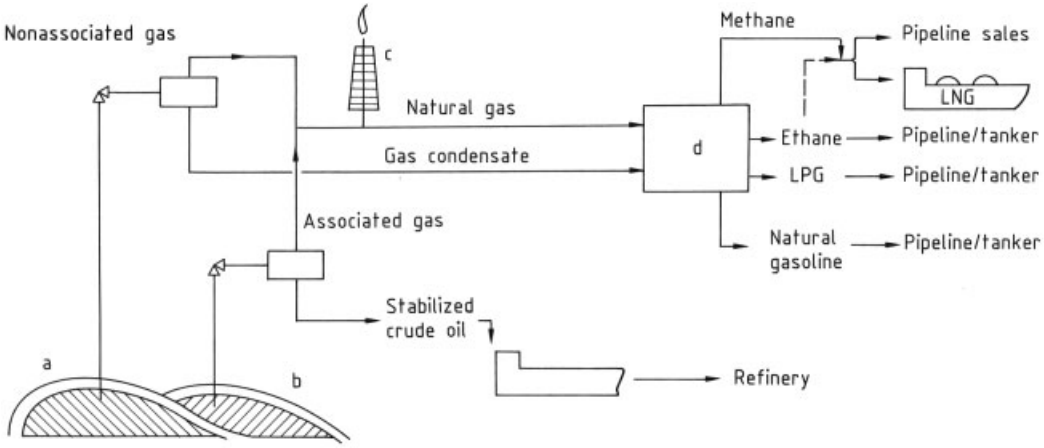

Figure 1. Oil – gas recovery scheme
a) Gas and gas condensate; b) Oil and gas; c) Vent-flare; d) Gas processing plant

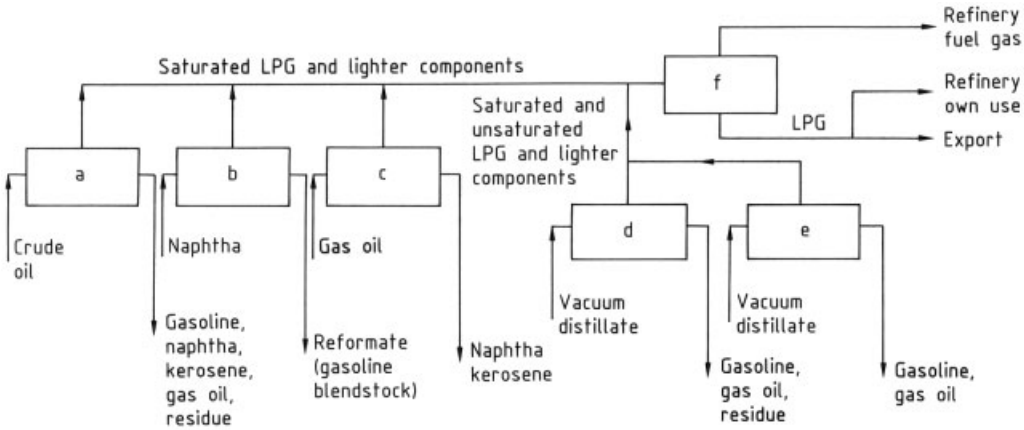

Figure 2. Sources of LPG in the refinery
a) Crude oil fraction; b) Catalytic reforming; c) Hydro-cracking; d) Thermal cracking; e) Catalytic cracking; f) LPG recovery

2. Properties

The physical properties of the main constituents of LPG are listed in Table 1. In addition to these components, other species may be present in trace quantities. Typically, sulfur compounds, water, and occasionally residual oils and tars may occur. Depending on the use of the LPG, these contaminants must be reduced to an acceptable level, commensurate with the applicable LPG specifications in the country of use.

3. Production and Processing

Recovery from Natural Gas. Natural gas consists largely of methane (→ Natural Gas) and smaller quantities of ethane, propane, butanes, and heavier hydrocarbons, together with varying amounts of water, carbon dioxide, nitrogen, sulfur compounds, and other nonhydrocarbons. Depending on the specifications of the gas supplied to the consumer, a proportion of the ethane and heavier component content is recovered in gas processing plants, which yield additional products such as ethane, LPG, and higher boiling hydrocarbons (natural gasoline). When no

incentive to recover ethane exists, only the LPG and natural gasoline are recovered from natural gas. Increasingly, however, ethane is considered a valuable product in the petrochemical industry, and plants are designed to recover not only LPG but also ethane. Potentially all of the ethane, LPG, and natural gasoline may be obtained as products from natural gas; however, some ethane must normally remain in the gas to meet sales specifications.

Recovery and Manufacture in the Refinery (→ Oil Refining). After crude oil has been stabilized for transportation to the refinery, small but significant amounts of LPG and lighter components (methane, ethane) remain in the oil. In the refinery, crude oil is first separated into a number of fractions by distillation (fractionation). Gas, which is the lightest fraction, is produced as the top product from the atmospheric fractionation column; it consists of LPG, ethane, and methane. Other fractions that contain the main refinery products are, in order of ascending atmospheric boiling points, naphtha, kerosene, gas oil, and residue.

In practice, the yield of each fraction does not match market demand. Consequently, the less valuable heavy fractions must be converted into lighter ones, and the lighter fractions must be chemically modified to improve their properties. Examples of these conversion processes are catalytic reforming, catalytic cracking, thermal cracking, hydrocracking, and alkylation.

The products of *thermal* and *catalytic* cracking include LPG and lighter components (ethane, ethylene, methane, hydrogen), both saturated and unsaturated. Conversely, during *catalytic reforming* and *hydrocracking* only saturated LPG and lighter components are formed. Generally, the products of these conversion processes are separated in a fractionation column to yield gas and the main products. Thus, gas streams consisting of LPG and lighter components are produced from direct fractionation of crude oil and from fractionation of the products of various refinery conversion processes.

These gas streams are used as fuel gas for refinery furnaces. However, LPG normally has a significantly higher economic value than fuel gas, and recovery of LPG is economically attractive. On the other hand, the most effective use of ethane in the refinery is as fuel gas because normally no incentive exists to recover ethane.

Depending on circumstances, various techniques may be used for the recovery of LPG which can be applied in gas plants or refineries. These methods are: (1) recontacting – compression, (2) refrigeration, (3) absorption, (4) adsorption, or (5) a combination thereof.

3.1. Recontacting – Compression (Fig. 3)

The overhead vapor product from the fractionator (a) is compressed to 1.2 – 1.5 MPa (c) and combined with the liquid top product, as shown in Figure 3. The combined stream is cooled to ca. 40 °C by air or water and directed to the separator (e). Because of the higher pressure, more LPG is obtained in the liquid phase of the separator than in the liquid phase in the overhead accumulator vessel (b) of the fractionator. Components heavier than butane are also recovered from the fractionator vapor top product by recontacting. The liquid phase from the separator passes through a deethanizer (f), from which the vapor product is recycled to the separator to improve LPG recovery. The vapor phase from the separator, which still contains some LPG, is used as fuel gas. The liquid product from the deethanizer is LPG, which may be further separated into propane and butane products in a depropanizer.

Recontacting – compression achieves ca. 75 % recovery of the LPG in the fractionator overheads. In gas processing plants, this technique is hardly used today because better recovery of LPG is almost always required.

3.2. Refrigeration

The recovery of LPGs from a gas stream can be accomplished by a variety of techniques. The most common technique applied in new LPG recovery plants is to refrigerate the gas stream to condense the LPG fraction. The recovered liquids are then fractionated to separate the LPG components.

Some of the most commonly employed processes are summerized below.

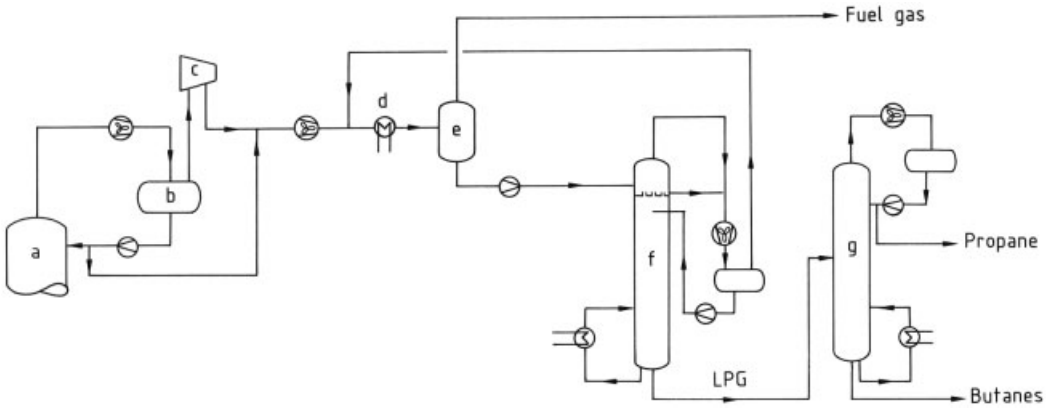

Figure 3. Recontacting – compression system
a) Fractionator; b) Overhead accumulator vessel; c) Compressor; d) Chiller (optional); e) Separator; f) Deethanizer; g) Depropanizer

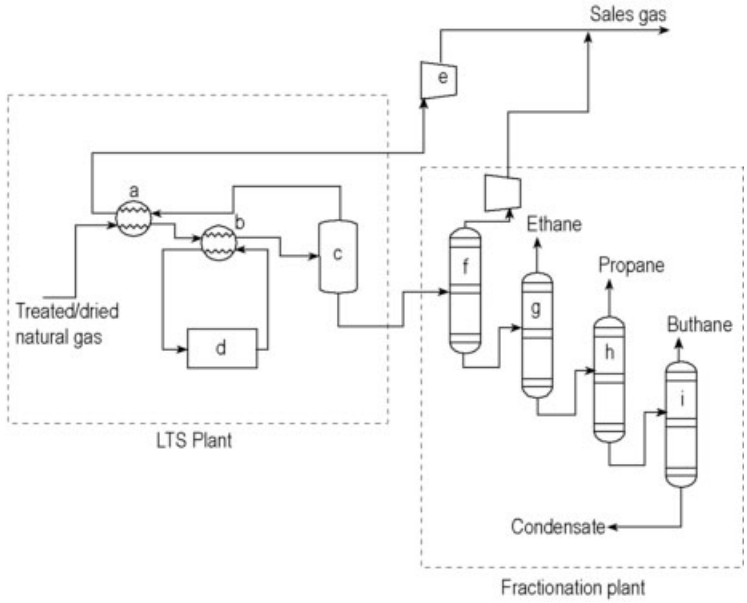

Figure 4. Schematic of an LPG plant
a) Feed gas – gas heat exchanger; b) Feed gas – refrigerant heat exchanger; c) HP separator; d) Refrigeration unit; e) Sales gas compressor; f) Demethanizer; g) Deethanizer; h) Depropanizer; i) Debutanizer

3.2.1. Low-Temperature Separation (LTS)

Refrigeration of an LPG containing gas stream can be accomplished by indirect heat transfer with an external refrigerant stream. In a conventional gas conditioning plant, refrigeration is normally provided by a closed-loop propane refrigeration circuit (see Fig. 4). Treated and dried feed gas to the plant is first contacted with cold LPG lean gas exiting the plant (a). The precooled gas is then further cooled by propane chilling to condense the LPG fraction in the feed gas (b). The condensed liquids are then recovered in the HP separator (c) and fed to a downstream fractionation plant. Cold from the lean gas exiting the HP separator is then used for cooling the feed gas.

The liquids from the HP separator are fed to a conventional fractionation plant comprising a demethanizer, deethanizer, depropanizer and

debutanizer for separation of the LPG components. If no extra ethane outlets are available, then the demethanizer and deethanizer column can be combined to a single column. The chief advantages of this type of process are its simplicity and low preassure drop.

Alternatively, chilling of the gas can be provided by a cascade refrigeration circuit. The refrigerant in this case is either a propane – ethane, a propane – ethylene, or a mixture of propane – ethane – methane – nitrogen (commonly called *mixture refrigerant*). A cascade ethane – propane circuit produces lower temperature than a single propane circuit. Accordingly, the method selected depends on the ethane and LPG recoveries desired. Recoveries of LPG $> 99\,\%$, corresponding to ca. 85 % ethane recovery, are achievable in practice by using mechanical refrigeration. Mechanical refrigeration is more usual in gas processing plants than in refineries because higher LPG recovery is normally required in a gas processing plant to meet the hydrocarbon dew point specification of the sales gas.

A cascade refrigeration plant is illustrated in Figure 4. Impurities such as water, carbon dioxide, and sulfur compounds are removed from the inlet gas, whereupon it is chilled against product gas and external refrigerant. Removal of water and carbon dioxide avoids possible freezing problems in the cold sections of the process. The liquid produced by chilling is fed to the demethanizer, where methane and some ethane are stripped. The vapor separated from separator (i) is used to chill the incoming feed before being commingled with the overhead vapor from the demethanizer. The liquid product from the demethanizer consists of ethane, LPG, and natural gasoline, from which LPG is separated by fractionation. The vapor phase from the demethanizer (mainly methane with some ethane) is heated against the incoming feed before being compressed for sale.

Liquid petroleum gas can also be recovered in a liquefied natural gas (LNG) plant, in which natural gas is fully liquefied by mechanical refrigeration (see Fig. 5). Generally, the external refrigerants employed are propane and a mixed refrigerant consisting of nitrogen – methane – ethane and propane. In the liquefaction process the partially condensed natural gas stream is passed to a distillation column (a) in which ethane, LPG, and the heavier natural gasoline are separated from natural gas. The natural gas (predominantly methane) leaves the distillation column overhead as a vapor product and is subsequently fully liquefied. The liquid phase from the column bottom is routed to a series of fractionation columns for separation into ethane, LPG (individually as propane and butane), and natural gasoline products. Some or all of the ethane and propane products may be used as refrigerants for the LNG plant, with any excess exported.

3.2.2. Expander Plants

Expander plants generate refrigeration by expansion of the feed gas. By dropping the pressure via a contol valve or an expander a sufficiently low temperature can be generated to condense the LPG fraction in the feed gas. The flow scheme for a conventional expander plant is shown in Figure 6.

Treated and dried gas to the plant is first cooled via indirect heat exchange with the cold lean gas exiting the plant (a). Any liquids formed during this first stage of cooling are then recovered in the HP separator (b) and fed to the bottom section of the demethanizer column (e). The overhead vapor from the HP separator is then expanded either via a rotating expander or via a control valve (commonly called a Joule – Thomson valve, d). The cold gas/liquid from the expansion is then fed to the top of the demethanizer. The demethanizer strips the lighter components (methane, nitrogen) from the liquids. The overhead vapor is subsequently fed to the feed gas – gas heat exchanger for recovery of cold before being compressed for export as sales gas. Compression power is minimized by coupling the first stage of the compression to the expander.

A recent development of the conventional expander process has been the gas subcooled process (see Fig. 7). This process is an enhanced expander process in which a rectifying section is added to the demethanizer column. The purpose in adding this section is to reduce the los of ethane and LPG in the overhead vapor by refluxing with cold liquedied feed gas. The reflux is obtained by taking a small portion of the HP separator gas which is then liquefied by chilling against the overhead vapor of the demethanizer.

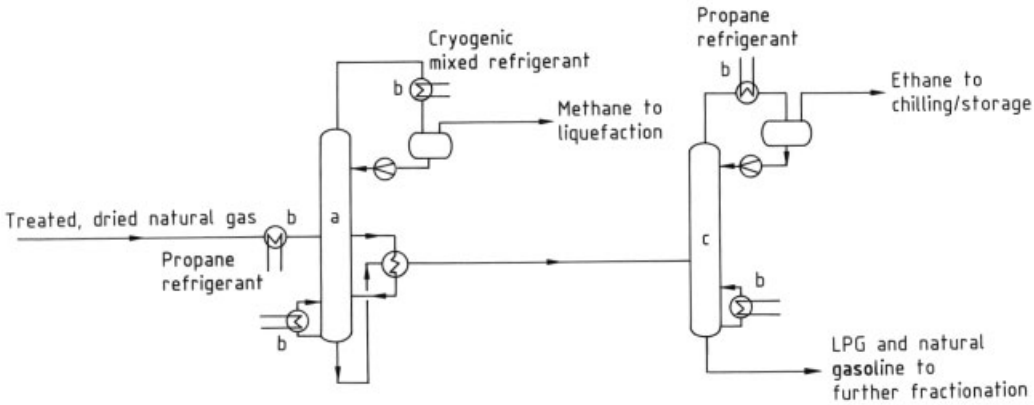

Figure 5. LPG recovery in an LNG plant
a) Distillation; b) Heat exchanger; c) Demethanizer

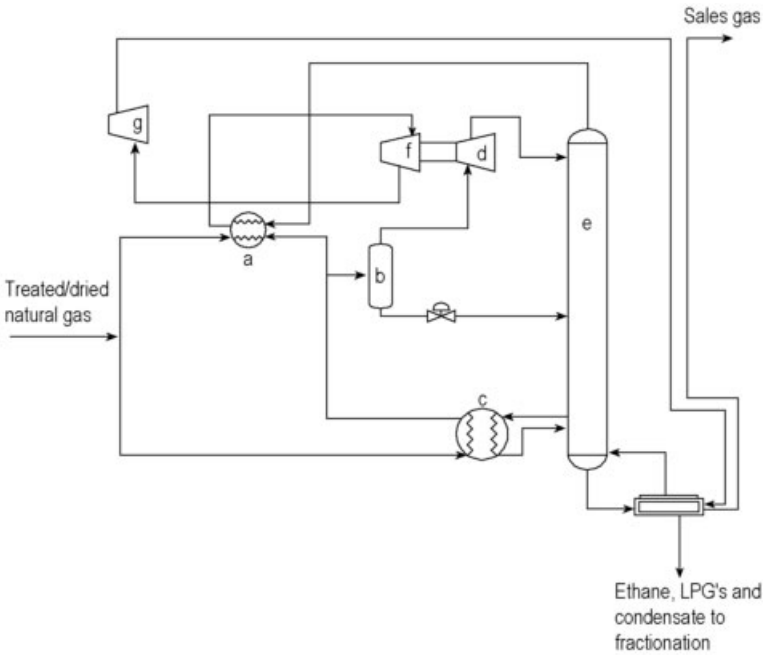

Figure 6. Schematic of a conventional expander plant
a) Feed gas – gas heat exchanger; b) HP separator; c) Demethanizer sideboiler; d) Turboexpander; e) Demethanizer; f) 1st Sales gas compressor; g) 2nd Sales gas compressor

The high-pressure cold liquid is then depressurized and fed to the top tray of the demethanizer column. The gas subcooled process is capable of recovering > 99 % of the LPGs in the feed gas and is also ideally suited for recovery of > 95 % of ethane in the feed.

3.2.3. Combined Processes

When high recoveries of LPG and/or ethane are desired, it is possible to assist the conventional expander process with external refrigeration. External propane or cascade refrigeration

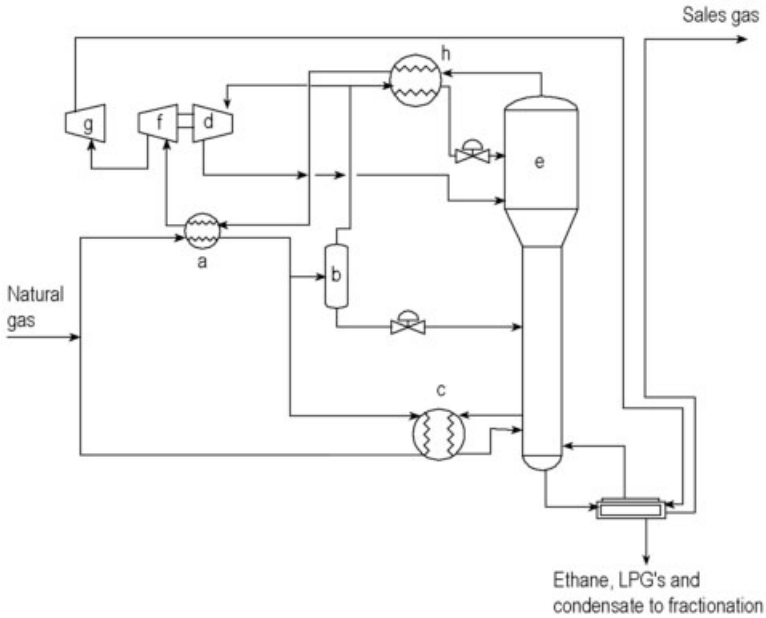

Figure 7. Schematic of the gas subcooled process
a) Feed gas – gas heat exchanger; b) HP separator; c) Demethanizer sideboiler; d) Turboexpander; e) Demethanizer; f) 1st Sales gas compressor; g) 2nd Sales gas compressor; h) Reflux exchanger

is applied to cool the feed gas entering the HP separator of the expander plant. The cold gas is then expanded and fed to a conventional demethanizer column.

With this type of process, more than 95 % ethane recovery and essentially 100 % recovery of LPG can be achieved.

3.3. Lean Oil Absorption

The lean oil absorption process employs a hydrocarbon oil to recover light components from either a natural gas or a refinery fuel gas stream. The process is used in refineries worldwide, as well as in natural gas processing plants, especially in the United States, where the vast majority of plants used the oil absorption process until the first expander plant started up in 1963. However, since 1963 expander plants have dominated natural gas processing.

Hydrocarbon oil may be bought in, or components of natural gasoline produced in the gas processing plant may be used. In the refinery, an available fluid such as naphtha is often employed. Despite the different terminology, these streams from the natural gas processing plant and oil refinery are very similar in terms of molecular mass. Typical absorption oils have a molecular mass of 100 – 180.

An example of a refinery lean oil absorption scheme is shown in Figure 8. The fractionator overhead vapor stream is compressed and combined with the liquid top product. The combined stream is then deethanized, and the vapor top product of the deethanizer is brought into contact countercurrently with the absorption oil in the absorber column. The oil absorbs propane and butane components from the gas, replacing them with fewer molecules of the absorber oil. The oil is taken from the main fractionation column (a), presaturated with vapor from the absorber column, chilled to 10 – 15 °C, and fed to the absorber column as lean oil. Presaturation and chilling of the lean oil improve LPG recovery in the absorber. The remaining vapor phase is directed to the fuel gas line. So-called rich oil from the bottom of the column, which contains absorbed LPG is recycled to the main fractionation column to complete the oil circuit. There the LPG is stripped from the rich oil and passed overhead to the deethanizer. The liquid product

from the deethanizer is routed to a fractionation column where the propane and butanes are separated. Typically, LPG recoveries of 85 – 90 % are attainable.

The process applied in natural gas processing plants is, in principle, the same as that used in the refinery, although LPG recoveries of around 98 % are often needed to ensure that gas supplied to the customer satisfies the hydrocarbon dew point specification. This high LPG recovery is achieved by cooling the lean oil to − 40 °C, usually by propane refrigerant. So that the absorber column can operate at this temperature, the natural gas feed must be dried to avoid problems caused by freezing.

3.4. Adsorption

In adsorption (→ Adsorption), gas molecules are reversibly bonded to the surface of a solid material, the adsorbent. Examples of adsorbents include silica gel, activated carbon, and alumina. Adsorption plants are generally used to obtain natural gasolines from natural gas to satisfy sales gas specifications for heavy hydrocarbon content. The LPG recovery is significantly lower than that achieved by refrigeration or lean oil absorption processes. A typical adsorption flow scheme is presented in Figure 9.

Three vessels are shown, each containing the solid adsorbent. The process is cyclical with, at any one time, one vessel on adsorption, one on regeneration cooling, and one on regeneration heating. The inlet gas passes through a bed (a) where heavy hydrocarbons are adsorbed. A slipstream of the inlet gas is compressed and used to cool another bed (b), which has previously been heated to remove adsorbed hydrocarbons. Once this bed is cooled, it is placed on standby ready to be put into adsorption service. Gas leaving this bed (b) is heated to ca. 300 °C in a furnace (e) and directed to another bed (c) to remove adsorbed hydrocarbons. Heavy hydrocarbons present in the regeneration gas leaving this bed (c) are recovered after cooling in a separator (g), from which the remaining gas is returned upstream of the initial bed (a). Liquid removed from the regeneration gas, consisting of natural gasolines and some LPG components, may be processed further by fractionation to yield separate LPG and natural gasoline products.

Whichever method is used to obtain LPG from natural gas or refinery fuel gas, LPG may be produced as a mixed product or as separate propane – propene, called C3 product, and butanes – butenes, called C4 product. The relative quantities of saturated and unsaturated species depend on the types of processes in the refinery that are responsible for LPG production (see Chap. 1). Further fractionation is sometimes employed to recover isobutane from the rest of the C4 product and propene from the C3 product. Isobutane is used in the refinery in the alkylation process, in which it reacts catalytically with either propene or butene to produce high-octane gasoline blendstocks. Propene can also be a valuable chemical feedstock. Thus, depending on the configuration of each refinery, the LPG pool may contain various blends of the saturated and unsaturated components that comprise LPG.

4. Purification

Whether produced in the refinery or a natural gas processing plant, LPG contains contaminants. Most common are water and sulfur compounds, including hydrogen sulfide, carbonyl sulfide, mercaptans, and elemental sulfur. Removal of these contaminants is necessary to satisfy locally applicable quality specifications, for example, DIN 51 622 in the Federal Republic of Germany, BS 4250 in the United Kingdom, and the Gas Processors Association (GPA) specification in the United States. As an example, Table 2 summarizes the requirements for different LPG products according to GPA specifications. Various processes are available to remove contaminants to the prescribed level, some of which are described below.

The first stage of drying LPG is achieved normally by separating *undissolved* water in a coalescer. If the LPG is commercial-grade *butane* which is stored at ambient temperature under its vapor pressure, this is sufficient to meet the water specification. However, if the butane is to be held in refrigerated storage at ambient pressure, further drying is necessary to remove *dissolved* water. This must be performed before the butane is chilled, otherwise problems caused by water freezing will ensue. Water levels in refrigerated LPG are typically less than 10 ppm (mol) to avoid freezing.

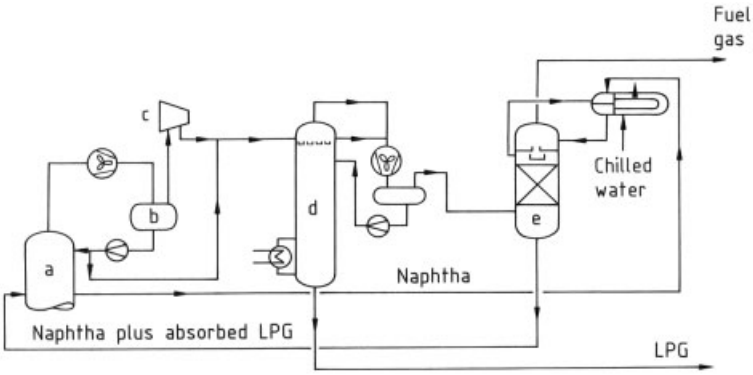

Figure 8. Oil absorption plant
a) Fractionator; b) Overhead accumulator vessel; c) Compressor; d) Deethanizer; e) Absorber column

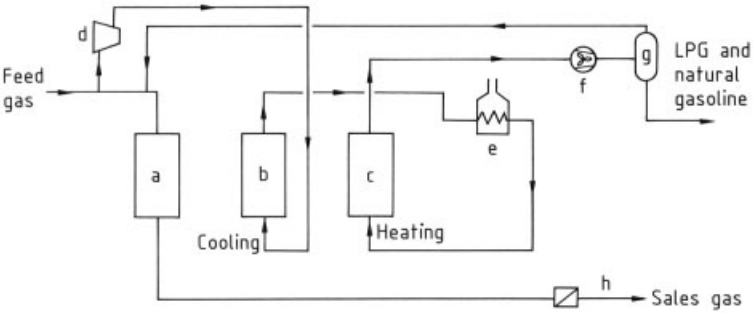

Figure 9. Schematic of adsorption plant
a) – c) Adsorbers; d) Compressor; e) Furnace; f) Cooler; g) Separator; h) Filter

Table 2. The LPG specifications of the U.S. Gas Processors Association [6], [7]

Product characteristics	Commercial propane	Commercial butane	Commercial propane – butane mix	Propane HD-5 [a]
Composition	mainly propane and propene	mainly butane and butene	mainly mixes of propane – propene and butane – butene	not less than 90 % propane; not more than 5 % propene
Vapor pressure (max.) at 100 °F, psig [b]	208	70	208	208
Temperature of volatile residue: at 95 % evaporation (max.), °F [b]	− 37	36	36	− 37
Butane and heavier, vol%	2.5			2.5
Pentane and heavier, vol%		2.0	2.0	
Residual matter, residue on evaporation of 100 mL (max.), mL	0.05			0.05
Oil stain observation	pass [c]			pass [c]
Corrosion, copper strip (max.)	no. 1	no. 1	no. 1	no. 1
Volatile sulfur (max.), grains/100 cu ft [e]	15	15	15	10
Moisture content	pass [d]			pass [d]
Free water content		none	none	

[a] HD-5 is the specification of propane for use as automotive fuel. [b] Conversion factors: 208 psig = 1434 kPa (g); 70 psig = 483 kPa (g); − 37 °F = − 38 °C; 36 °F = 2 °C. [c] An acceptable product shall not yield a persistent oil ring when 0.03 mL of solvent residue mixture is added to a filter paper in 0.1 increments and examined in daylight after 2 min as described in test method ASTM D 2158. [d] Passes the GPA Propane Dryness Test (cobalt bromide) or ASTM D 2713-70.

Irrespective of the method of storage, commercial-grade *propane* requires the removal of dissolved water to satisfy the specification.

Adsorptive Purification. Dehydration of LPG can be effected by using a solid adsorbent such as molecular sieve, alumina, silica gel, or calcium chloride. With the exception of calcium chloride, these are regenerative processes operating on a cyclical basis. Dried LPG is taken from the outlet of one of the beds and vaporized by heating. It is then passed through the bed which is to be regenerated, where it heats the bed, thereby desorbing the water. The outlet stream is cooled to condense water and LPG, and the LPG is separated from the water before being returned to the inlet of a bed which is on adsorption service. The regenerated bed is then cooled, filled with LPG, and returned to service, adsorbing water from the LPG.

Molecular sieve units can also handle the combined removal of water, hydrogen sulfide, carbonyl sulfide, and small amounts of mercaptans. Adsorption and regeneration occur in the same manner as for drying only, although water recovered during the regenerative cooling step, as well as the regeneration gas itself, is "sour" (i.e., it contains sulfur compounds) (→ Natural Gas). Molecular sieves tend to be favored when the quantity of contaminant to be removed is small.

Absorptive Purification. The LPG can be brought in contact with an *amine solution* in an absorber column primarily to remove relatively large amounts of hydrogen sulfide. However, amines are not particularly effective for the removal of mercaptans. The sour amine solution is regenerated continuously in a regenerator column at low pressure and high temperature before being recirculated to the absorber column. Amines used include monoethanolamine (MEA), diethanolamine (DEA), and diglycolamine (DGA). Shell's ADIP process employs diisopropanolamine for hydrogen sulfide removal which, with an added downstream mixer–settler system, also allows removal of carbonyl sulfide to low residual levels.

Caustic soda treatment is used primarily to remove mercaptans and smaller quantities (< 100 ppm) of hydrogen sulfide. In an absorber column, LPG is contacted countercurrently with sodium hydroxide solution. The sour caustic solution is then passed to a regenerator column where air stripping oxidizes the mercaptans to disulfides, which are removed from the regenerated sodium hydroxide solution. Regenerated caustic solution is recycled to the absorber. The removal of hydrogen sulfide by sodium hydroxide renders the latter nonregenerable and it must be discarded; hence, bulk hydrogen sulfide removal is normally performed upstream of the caustic unit. Variants of the regenerative caustic process are marketed with additives to enhance the process, for example, UOP's Merox system.

Depending on the type and quantity of contaminants present in the feed and product specifications, one or more of these processes may be employed for LPG treatment. For example, an amine unit for removal of hydrogen and carbonyl sulfides is frequently present in the refinery, followed by a caustic unit to remove the remaining hydrogen sulfide and mercaptans before final product drying with alumina (see Fig. 10). However, in case of particularly severe product specifications (e.g., if the LPG is used as chemical feedstock), further treatment may be necessary to remove carbonyl sulfide and mercaptans to very low levels. A "polishing" step is then incorporated, generally using adsorbents on a nonregenerative basis (e.g., Selexsorb).

5. Storage and Transportation

The storage of LPG may be categorized as follows: pressurized storage at ambient temperature; refrigerated storage at ambient pressure; and semirefrigerated, partially pressurized product storage. Further, LPG may be stored either above or below ground.

Aboveground Storage. The design pressure of aboveground *pressure storage* tanks is normally 1724 kPa for propane vessels and 860 kPa for butane vessels. These values correspond to the vapor pressures of both compounds at the maximum expected ambient temperature. For smaller inventories (up to ca. 100 t), the vessels are usually horizontally or vertically mounted cylinders with hemispherical heads. Larger inventories tend to be stored in spherical vessels.

The largest pressurized spheres hold ca. 1500 t; thus, for storage of larger quantities than

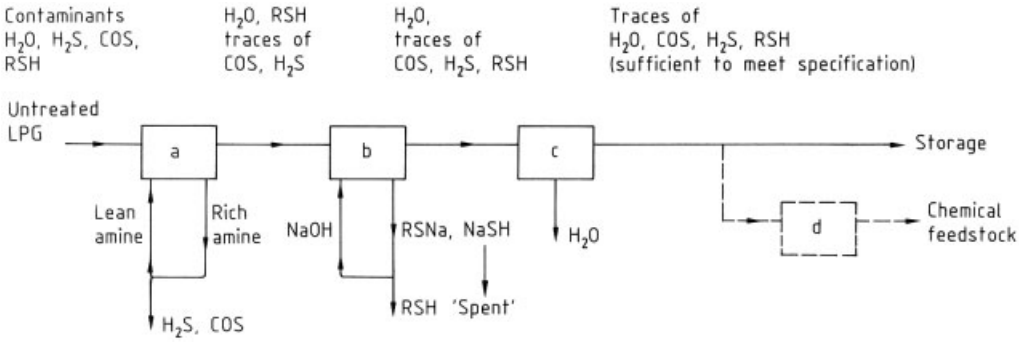

Figure 10. Refinery LPG product treating
a) Amine treatment; b) Caustic treatment; c) Drier; d) Polishing

this, either several spheres are needed or *refrigerated storage* must be used. This is designed to store the LPG product at its atmospheric boiling temperature (i.e., ca. $-42\,°C$ for propane and $-4\,°C$ for butane). The tank design pressure is typically ca. 110 kPa. Tanks are insulated; vapors liberated by heat leaking in are compressed, liquefied, and returned to the storage tank. Usually, the boil-off vapor is liquefied by the refrigeration system provided to chill the LPG product rundown from the producing plant. Sometimes, the aboveground storage facility is mounded with earth, as can be seen in Figure 11, which shows two 50 000-m^3 and two 35 000-m^3 tanks.

Figure 11. Fully refrigerated, mounded LPG storage tanks

Semirefrigerated storage can also be used: propane is stored at $-10\,°C$, at which its vapor pressure is ca. 320 kPa.

Underground Storage. The LPG may also be stored underground in salt or quarried caverns that can withstand its vapor pressure at subterranean temperature. These have been used in the United States, Canada, and Europe. In frozen-earth storage, the ground surrounding the excavation is frozen and an insulated roof is constructed over the cavity. Propane is more suitable for storing by this method because its boiling point at atmospheric pressure is considerably lower than the freezing temperature of the soil thus causing moisture in the soil to freeze. Refrigerated propane product at ca. $-42\,°C$ is loaded into the cavity where it is held at atmospheric pressure. Propane vapors liberated by heat leaking in are compressed, cooled, and returned to the cavity storage as liquid.

Transportation. Depending on the location of LPG-producing plants in relation to markets, LPG may be transported by pipeline, sea, road, or rail. Traditionally, LPG has been carried in pressurized ships as well as trucks and rail cars, except for bulk refrigerated deliveries to Japan. Pipelines have been used for inland transportation of large volumes, particularly in the United States. For transport to the consumer, LPG is delivered in cylinders and bulk vehicles of various sizes. The maximum capacity of a *pressurized* LPG ship is limited to about 2000 m^3 due to high manufacturing costs. *Semirefrigerated* ships usually hold 2500 – 5000 m^3, with *fully* refrigerated ships having capacities up to 100 000 m^3. The larger the LPG cargo, the more economical it is to transport fully refrigerated.

The LPG can be unloaded into refrigerated storage tanks or, after heating, into pressurized tanks which supply conventional pressurized distribution facilities. Alternatively, cargo may be transferred into smaller coastal tankers for delivery of refrigerated products to smaller terminals which cannot receive large oceangoing ships. In Figure 12 a fully refrigerated LPG ship is seen at a loading jetty.

6. Uses

The LPG supplied in cylinders or delivered to larger bulk tanks finds a market in domestic, commercial, and industrial sectors. In the domestic and commercial markets, LPG provides a clean fuel for cooking, heating, and automobiles. It is an especially attractive automotive fuel because it substantially lowers vehicle exhaust emissions. In the industrial market, LPG is suitable for many uses where high-quality fuel is required. The high costs of storage and handling normally limit LPG to those applications in which it can command a premium price.

Significant quantities of LPG are also used in the refining and petroleum industries. In the refinery, butane is used primarily in the production of gasoline; *n*-butane is added to increase the volatility and octane number of the fuel, whereas isobutane and butene are consumed by the alkylation process to provide high-octane alkylates for gasoline blending (→ Oil Refining). An important use of LPG as a chemical feedstock is in the manufacture of intermediates for polymers such as polyethylene, poly(vinyl chloride), and polypropylene, to name a few.

In the *pipeline gas industry*, LPG may be used in various ways. Propane can be injected close to the consumer when the pipeline is overloaded or, less often, earlier in the transmission system when there is a shortage of gas. Because of the effect on the calorific value and burning characteristics of the gas, the quantities of propane that may be used are limited. Alternatively, an air – propane mixture may be introduced in larger quantities, and this method of increasing gas supplies during peak demand periods has been employed for many years. Butane is not as attractive for this service because of its higher boiling point.

7. Safety Aspects

Throughout the process of LPG manufacture, storage, transportation, and consumption, high safety standards are necessary because LPG is a hazardous substance. It is flammable and, in the event of leakage, can release large amounts of flammable gas which may migrate considerable distances from the source of the leak before being ignited. Being heavier than air, escaping LPG readily accumulates at low points, including open and closed drains and sewers.

Generally, precautions must be taken to minimize the likelihood of uncontrolled leakage and eliminate ignition sources. When controlled leakage of LPG from storage or plant has to be allowed for, it must be piped to a safe location for disposal in a controlled manner, normally by flaring. To this end, equipment used in the storage and processing of LPG is fabricated according to procedures laid down in the appropriate codes. These specify, for example, construction materials; required wall thickness; inspection requirements; and protective devices such as safety relief valves, depressurizing facilities, water deluge systems, and emergency shutdown valves. Ignition sources are also kept out of plant or storage areas where the risk of LPG release exists. Accumulation of static electricity is prevented by ensuring that all equipment is earthed.

Occupational Health. LPG is an asphyxiant, causing suffocation by the exclusion of an oxygen atmosphere. It is, however, not a toxic substance. The precautions required center on making sure that people are not exposed to a suffocating atmosphere. For example on a manufacturing plant personnel must not enter a vessel unless that vessel is suitable for entry, i.e., it has been physically disconnected from other plant equipment and freed of flammable substances by purging with inert gas (normally nitrogen) followed by replacement of the inert gas with a breathable oxygen atmosphere.

8. Economic Aspects [8], [9]

The major producers and exporters of LPG are the Middle East, Canada, North Africa, and the

Figure 12. Refrigerated LPG ship

Far East. Reduced flaring of natural gas, particularly in the Middle East, is expected to contribute significantly to the quantities of LPG available for export. Prospects for international trade are also improved by the reduction in freight costs associated with the use of large-capacity refrigerated LPG ships.

The total world consumption of LPG in 1993 was 159×10^6 t, the main consumers being the United States, Japan, and Western Europe.

United States. In the United States, consumption in 1993 totaled 84×10^6 t. Historically, ca. 70 % of the LPG production in the United States has been supplied by gas processing plants, with oil refineries providing the remaining 30 %. The volumes produced internally have not quite satisfied local demand; therefore, some imports (ca. 10 % of the total demand) have been necessary. Most of the imports have come from Canada and the Middle East. Future production of LPG from natural gas is predicted to decrease as greater proportions of the gas are taken from wells containing less LPG. Consequently, LPG imports are expected to increase in the United States.

Japan. In 1993, LPG consumption in Japan was 20×10^6 t. Japan has little indigenous natural gas, which strongly influences the LPG market. The lack of natural gas has resulted in a public gas distribution system which is less extensive than in other industrialized countries. Consequently, on the demand side, the market for LPG is large with two-thirds of the total demand typically being from the domestic and industrial sectors. Another important feature of the Japanese market is the demand for clean burning fuels, for example, in the automotive and power generation sectors. On the supply side, the lack of natural gas means that the only sources of LPG produced locally are oil refineries. The LPG produced by refineries is insufficient to meet the demand, to such an extent that Japan was the largest single importer of LPG in 1993, importing 15×10^6 t.

Western Europe. In Western Europe, LPG consumption was 32×10^6 t in 1993. Traditionally, the LPG supply in Europe has come from refinery production, to a market restricted to some extent by the availability of pipeline natural gas. However, by the early 1980s increased demand had led to most Western European countries becoming net importers. In 1983 this was still true, although in the United Kingdom exports and imports were finely balanced because of the LPG produced from the British sector of the North Sea.

9. References

General References

1. Shell Briefing Service: *Liquefied Petroleum Gas,* Shell Int. Pet. Co., London 1979.
2. Royal Dutch/Shell: *The Petroleum Handbook,* Elsevier, Amsterdam 1983.
3. J. J. L. Ma, Ching-Shien Wu: *Natural Gas Liquids,* SRI International Report no. 135, Oct 1979.

Specific References

4. A. F. Williams, W. L. Lom: *Liquefied Petroleum Gases, Guide to Properties, Applications, and Uses,* 2nd ed., Ellis Horwood, Chichester 1982.
5. R. L. Rowell: "Hi-Cryo Process for Recovering Liquids from Refinery Fuel Gas," *Nat. Pet. Refin. Assoc. Tech. Pap.* 1982.
6. Gas Processors Suppliers Association: *Engineering Data Book,* GPSA, Tulsa, Oklahoma 1987.
7. Gas Processors Association: *LPG Specifications and Test Methods,* publication 2140 – 80.
8. G. Drayton: *The Market for LPG in the 1980's,* Economic Intelligence Report no. 96, 1981.
9. United Nations: *1983 Energy Statistics Yearbook.*

Liquid Crystals

IAN SAGE, BDH Ltd., Poole, United Kingdom

1. Introduction 1
2. Classification and Nomenclature of Liquid Crystal Phases 1
3. Thermotropic Calamitic Phases . . 2
3.1. Thermotropic Liquid Crystal Polymorphism 2
3.2. Structure – Phase Relationships . . . 5
3.3. The Nematic Phase 8
3.4. Physical Properties of Nematic Liquid Crystals 9
3.4.1. Refractive Indices 9
3.4.2. Dielectric Constants 10
3.4.3. Diamagnetic Constants 11
3.4.4. Elastic Constants 12
3.4.5. Light Scattering 13
3.4.6. Viscosities 13
3.4.7. Surface Alignment 14
3.5. Nematic Liquid Crystal Materials for Display Use 14
3.5.1. The Frederiks Transition 15
3.5.2. Electrically Controlled Birefringence 16
3.5.3. The Twisted Nematic Display 17
3.5.4. The Supertwisted Nematic Display . . 19
3.5.5. Active Backplane Liquid Crystal Displays 20
3.5.6. Guest – Host Displays 20
3.5.7. The Dynamic Scattering Display . . . 21
3.5.8. Polymer Dispersion Liquid Crystal Devices 22
3.6. Thermotropic Cholesteric Liquid Crystals 22
3.7. Blue Phases 25
3.8. Thermotropic Smectic Phases 25
4. Thermotropic Discotic Liquid Crystals 29
5. Polymer Liquid Crystals 29
6. Amphiphilic Liquid Crystal Phases 31
7. References 32

1. Introduction

For a century, compounds have been recognized which, on melting or addition of a solvent, pass through phases that have a degree of order lower than the crystalline solid, but higher than the normal isotropic liquid [7]. Compounds containing rod- or disk-shaped molecules frequently pass through such intermediate phases, termed *mesophases*, in which the molecules are free to move past each other but are aligned parallel to one another within macroscopic volumes of the fluid. Such materials are known as liquid crystals because they combine in a single phase the typical flow of a liquid with the anisotropy of properties found in noncubic crystals. All liquid crystal phases share this orientational order; different liquid crystal phases are distinguished by varying degrees of additional order.

Liquid crystal phases have been discovered in a great range of organic chemicals [8], including cellulose derivatives, peptides, surfactant systems, drugs, and dyes. Stimulus for the development of liquid crystals has, however, come largely from the discovery of electro-optic effects in a restricted range of liquid crystal phases and their exploitation in displays (→ Display Technology). This article reflects that emphasis and describes the technologically important liquid crystal phases in some detail, while providing a briefer account of other areas of liquid crystal science.

2. Classification and Nomenclature of Liquid Crystal Phases

The wide variety of chemical systems and phase types encompassed by the term liquid crystal has led to the use of a number of specific terms in the literature. Compounds which form liquid crystal phases are often termed *mesogens*. Liquid crystals formed on heating a crystalline solid are referred to as *thermotropic phases*. Such phases may also be obtained on supercooling an isotropic liquid below the melting point of the crystalline solid and are then described as *monotropic phases*. Monotropic transition temperatures are conventionally cited in parentheses. *Enantiotropic phases* exist as the thermodynamically stable phase between the crystalline

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_359

melting point and the *clearing point* (the temperature at which the liquid becomes clear due to a reversible transition to the isotropic state). In some cases a liquid crystal phase cannot be observed in a compound due to its failure to supercool adequately, but the properties of a *virtual liquid crystal phase* can be inferred from the behavior of the compound in solution in a liquid crystal solvent. Virtual transition temperatures are conventionally cited in square brackets.

Liquid crystalline behavior has been most thoroughly explored in compounds composed of rod- or lath-shaped molecules which form *calamitic phases*. This distinguishes them from the *discotic phases* formed by flat plate- or disk-like molecules. Calamitic liquid crystal materials form *nematic phases* having only orientational order, *cholesteric* or *chiral nematic phases* which have only a helical orientational ordering, and *smectic phases* in which statistically parallel molecules are arranged in layers. Discotic liquid crystals also form nematic and chiral nematic phases, as well as *columnar phases*: in the latter, disk-shaped molecules stack together to form rodlike entities.

Lyotropic liquid crystal phases are formed on addition of a solvent to a solid compound: lyotropic nematic, cholesteric, and smectic phases are known.

Both thermotropic and lyotropic *liquid crystal polymers* are known: such polymers may contain discotic or calamitic liquid crystal subunits in the main chain, as pendant groups, or in combination.

In addition to the common categories outlined above, less common liquid crystal phases occur. *Bowlic phases* derived from dome-shaped molecules have been discussed in the literature; *biaxial nematic phases* have been described [9–11], and new phases are sometimes named after the classes of compounds exhibiting them. In view of the rapid progress in liquid crystal science the reader must be prepared for the emergence of new phases and phenomena, with nomenclature to match.

3. Thermotropic Calamitic Phases

Liquid crystal phases derived from rodlike molecules form the best studied category of mesophases. The first liquid crystal material studied, cholesteryl benzoate (**1**) [*604-32-0*],

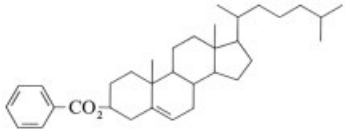

1

forms phases of this type. The relatively common occurrence of liquid crystal phases in, and ease of synthesis of, compounds in which two or more aromatic rings carrying terminal groups (**2**) are

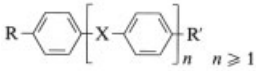

2

linked together led to these systems forming the exemplary phases on which liquid crystal science is based. A selection of compounds forming thermotropic calamitic phases is given in Table 1.

3.1. Thermotropic Liquid Crystal Polymorphism

Liquid crystal phases share the property of orientational ordering of anisotropic molecules, but a number of phases can be distinguished in which additional degrees of positional or orientational order exist [12]. The crystalline phase is designated as K and the isotropic liquid as I. The *nematic phase* (N), which usually forms a fluid of low viscosity, exhibits only long-range orientational order (Fig. 1 A). The long axes of the molecules tend to align parallel within macroscopic volumes of liquid.

Smectic phases (S), in which the molecules are additionally ordered in layers, form a range of very viscous fluids or waxy solidlike materials [13]. The least ordered smectic phase, denoted *smectic A* (S_A), has disordered layers in which the molecules can move freely in two dimensions (Fig. 1 B). Motion in the third dimension would imply hopping between layers and is restricted by an energy barrier. In the S_A phase, the average orientation of the molecular long axes is perpendicular to the layer planes, but this phase is subject to the same thermal disordering as the nematic phase. A corresponding phase in which

Table 1. Thermotropic calamitic liquid crystals

Structure	Name	CAS registry number	Empirical formula	Phase transition sequence, °C*
$CH_3O-C_6H_4-N(O)=N-C_6H_4-OCH_3$ **3**	bis(4-methoxyphenyl)diazene-1-oxide (paraazoxy anisole, PAA)	[*1562-94-3*]	$C_{14}H_{14}N_2O_3$	K 118.2 N 135 · 3 I
$CH_3O-C_6H_4-CH=N-C_6H_4-C_4H_9$ **4**	4-butyl-*N*-(4-methoxyphenyl methylene)benzenamine (methoxybenzylidene butylaniline, MBBA)	[*26227-73-6*]	$C_{18}H_{21}NO$	K 22 N 47 I
$CH_3O-C_6H_4-CO_2-C_6H_4-C_5H_{11}$ **5**	4-methoxybenzoic acid, 4-pentylphenyl ester (ME 105)	[*38444-13-2*]	$C_{19}H_{22}O_3$	K 29 N 42 I
$C_5H_{11}-C_6H_4-C_6H_4-CN$ **6**	4′-pentylbiphenyl-4-carbonitrile (pentyl cyanobiphenyl, K15, 5CB, PCB)	[*40817-08-1*]	$C_{18}H_{19}N$	K 22.5 N 35 I
$C_{16}H_{33}O-C_6H_3(NO_2)-C_6H_4-CO_2H$ **7**	3′-nitro-4′-(hexadecyloxy)-biphenyl-4-carboxylic acid	[*21351-70-2*]	$C_{29}H_{41}NO_5$	K 126.8 S_C 171 S_D 198.5 S_A 199.8 I
$C_5H_{11}OCOCH=CH-C_6H_4-N=CH-C_6H_4-C_6H_5$ **8**	3-[4-(4-biphenylylmethyleneamino) phenyl]-2-propenoic acid, pentyl ester		$C_{27}H_{27}NO_2$	K 92 S_E 101.5 S_B 168 S_A 204 I
$C_6H_5-C_6H_4-C_6H_4-C_6H_4-C_6H_5$ **9**	1,1′:4′,1″:4″,1‴:4‴,1⁗-quinquephenyl	[*3073-05-0*]	$C_{30}H_{22}$	K 380 N 431 I
$C_3H_7-C_6H_{10}-CO_2-C_6H_4-OC_4H_9$ **10**	*trans*-4-propylcyclohexanecarboxylic acid, 4-butoxyphenyl ester	[*67679-51-0*]	$C_{21}H_{32}O_3$	K 41.5 N 72.5 I
$C_3H_7-C_6H_{10}-CO_2-C_6H_{10}-C_5H_{11}$ **11**	*trans,trans*-4-propylcyclohexane-carboxylic acid, 4-pentylcyclohexyl ester	[*79284-92-7*]	$C_{21}H_{38}O_2$	K 24 S_B 51.5 N 54 I
$C_3H_7-C_6H_{10}-CO_2-C_6H_2(CN)_2-OC_4H_9$ **12**	*trans*-4-propylcyclohexanecarboxylic acid, 4-butoxy-2,3-dicyanophenyl ester	[*75941-67-2*]	$C_{22}H_{28}N_2O_3$	K 95 N (39) I

* K = crystalline phase; S = smectic phase; N = nematic phase; I = isotropic liquid.

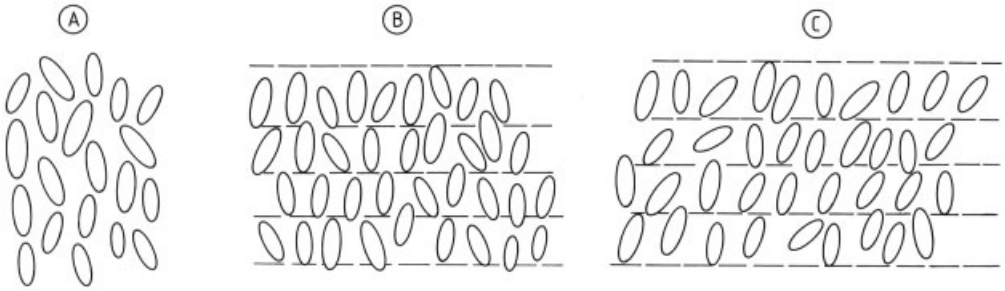

Figure 1. Schematic representation of molecular arrangement in the nematic (A), smectic A (B), and smectic C (C) phases

the molecules are tilted in the layers is known as the *smectic C phase* (S_C). The direction of molecular tilt is correlated between layers, and the S_C phase is formally biaxial (Fig. 1 C). Several subdivisions of S_A and S_C phases have been characterized.

In compounds and mixtures which exhibit liquid crystalline polymorphism, smectic phases usually occur at lower temperature than the less ordered nematic phase, but many exceptions to this rule are encountered. For example, the 4′-cyanobiphenyl ester of 4-(octyloxy)benzoic acid (**13**) [*61195-44-6*] shows the phase transition

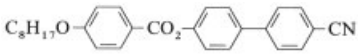

13

sequence K 97 N 120 S_A 201 N 240 I (the numbers indicate the phase transition temperature in degrees Celsius), thereby indicating a smectic phase that can be transformed to a nematic phase either by heating or by cooling. Examples of such *reentrant phase sequences* involving S_A and S_C phases are particularly common in compounds containing dipolar linking groups and terminal cyano or nitro groups.

Three phases exist in which hexatic (short-range hexagonal) molecular ordering exists in the layers (Fig. 2). In the hexatic smectic B phase (S_B^H) the molecules are statistically perpendicular to the layer planes and free to rotate about their long axes. There is, however, a short-range positional ordering of the molecules on a hexagonal net which possesses long range orientational order. The smectic I and F phases are tilted analogues of the hexatic B phase, in which the molecular tilt is directed toward either an apex or a face of the hexagonal net, respectively.

Strictly speaking, the remaining smectic phases (D, E, G, H, J, K) should not be regarded as liquid crystals; they possess three-dimensional long-range order and are more correctly classified as disordered solids. The crystal B phase (S_B^{Cr}) has true long-range order on a hexagonal net in each layer, with correlation of molecular position between layers. The tilted phases with three-dimensional ordering that correspond to the smectic I and F phases are denoted J and G, respectively. Finally, smectic phases are observed in which loss of complete rotational freedom of molecules about their long axes occurs. In these phases the molecules adopt a herringbone packing. An orthogonal phase (S_E) and tilted phases (S_K, S_H) are known. The smectic D phase occurs rarely, and its structure has not been elucidated fully but is based on a primitive cubic lattice. Other similar, but apparently distinct, cubic phases have been observed.

The identification and classification of liquid crystal phases rests on three techniques. Many phases show characteristic appearances (textures) when observed in a thin film in a polarizing microscope [13]. These textures are often sufficient to identify phases, especially when use is made of the known relative thermal stability of smectic phases [14]. When the sample temperature decreases, phases occur in a set order by selection from the sequence S_A, S_D, S_C, S_B^H, S_I, S_F, S_B^{Cr}, S_J, S_G, S_E. In case of doubt, recourse may be made to testing the miscibility of a phase with a known standard material. Miscibility over the complete composition range implies identity of the phases; however, different phases may also mix over all but a narrow region of the phase diagram, and phases of identical symmetry can fail to mix. Final clarification of a phase structure may be obtained by X-ray studies.

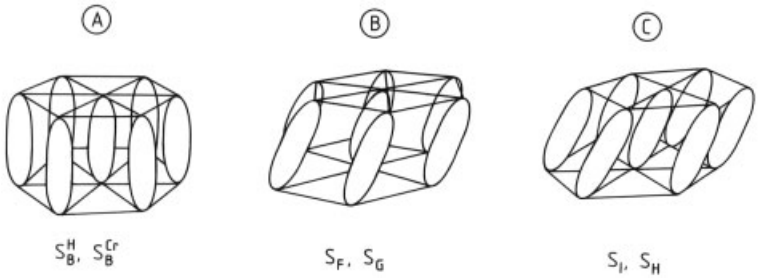

Figure 2. The hexagonal local order found in certain smectic liquid crystal phases
The molecules may be orthogonal to the layers [(A); S_B^H and S_B^{Cr}] or tilted toward a face [(B); S_F and S_G] or apex [(C); S_I and S_H] of the hexagonal net. In the hexatic phases, S_B^H, S_F and S_I, the positional order is short-range but the orientation of the hexatic net is maintained over long ranges. There is no correlation of molecular position between layers. In the crystal smectic S_B^{Cr}, S_G, and S_H phases the molecules lie on a lattice with long-range positional order in each layer. The molecular positions are also correlated between layers.

3.2. Structure – Phase Relationships

No predictively useful theory is available to relate molecular structure to liquid crystal phase type and stability, but experience accumulated over the past century has led to a number of useful empirical rules. Experienced workers can frequently predict liquid crystal phase stability and clearing points for novel systems by reference to analogous known materials. Prediction of phase type, particularly between smectic polymorphs, is much more difficult.

Relative ease of synthesis means that aromatic systems have been most extensively explored as liquid crystals. A very wide range of compounds conforming to structure **2** are known. At least two ring structures are normally necessary to stabilize a calamitic liquid crystal phase. The apparent contradiction of this principle by compounds such as 4-pentylbenzoic acid (**14**)

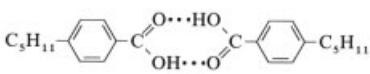

14
K 88 N 126.5 I

[*26311-45-5*], is explained by dimer formation, but a few genuinely acyclic mesogens are known e.g., *E,E*-2,4-undecadienoic acid (**15**) [*74267-92-8*].

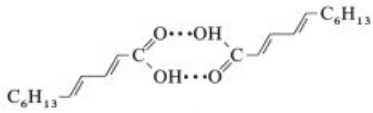

15
K 32 N 62.5 I

The central features of liquid crystalline compounds conforming to structure **2** include an elongated rigid molecular core substituted with terminal groups, which may be flexible or polar, and a high anisotropy of electric polarizability. The nature of the terminal groups can strongly affect the stability of the liquid crystal phases, and groups that extend the electronic conjugation of the core increase the transition temperature. An order may be compiled for the efficiency of terminal groups in promoting nematic phase stability:

$$C_6H_5 > CN > OCH_3 > NO_2 > Cl > CH_3 > F > CF_3 > H$$

On ascending a homologous series of mesogens, an alternation in transition temperatures is frequently observed. Odd-carbon alkyl chains give higher clearing points than even-carbon alkyl chains (e.g., $C_3H_7 > C_4H_9$); even-carbon alkoxy groups give higher clearing points than odd-carbon ones (e.g., $OC_2H_5 > OC_3H_7$). The underlying trend in clearing points varies with the nature of the molecular core: clearing points may rise gradually, rise initially and then fall, or fall initially and then rise. The temperature range of the liquid crystal phase is often largely

determined by the effect of terminal groups on the melting point of the system. In most homologous series, a nematic phase predominates in short-chain compounds and smectic phases are introduced in the higher homologues, with eventual exclusion of the nematic phase.

The central linking group in structure **2** also has an important effect on liquid crystal phase stability; a group that extends electronic conjugation between the rings significantly increases the clearing point. The group must also retain the rigidity and linearity of the core structure. Suitable linking groups include:

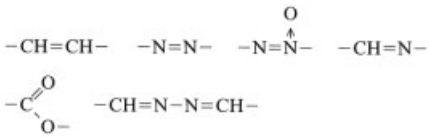

Groups having a poor effect on liquid crystal phase stability within this environment include:

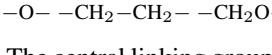

The central linking group was a frequent source of hydrolytic or photolytic instability in early liquid crystals used at room temperature. Exploitation of a single bond as a linking group led to the discovery of the cyanobiphenyls as the first highly stable room temperature nematic liquid crystals.

Early observations relating structure to phase stability, together with the success of the Maier–Saupe theory of nematic ordering (see Section 3.3) supported the view that electronic polarizability was an important contributor to liquid crystal phase stability. This theory was reinforced by the early synthesis of alicyclic systems with poor liquid crystal properties (cf. **16** [*1818-98-0*], **17**) but had long been contradicted

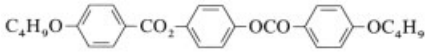

16
K 153 N 241 I

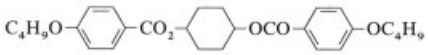

17
K 129 N 157 I

by the mesogenic properties of a range of derivatives of cholesterol and other sterols. Subsequent synthesis of structures in which replacement of an aromatic ring by an alicyclic ring improved liquid crystal stability, led to an appreciation of the true situation [15]: alicyclic ring structures form good structural units for liquid crystal molecules, but regions of high polarizability or polarity must not be separated by a linking group of low polarizability. These points are well illustrated by the N–I transition temperatures of the following sequence of compounds:

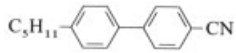

6
N 35 I

4′-pentylbiphenyl-4-carbonitrile [*40817-08-1*]

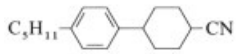

18
N–25 I

4-(pentylphenyl)cyclohexanecarbonitrile [*66227-29-0*]

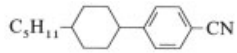

19
N 55 I

4-(4-pentylcyclohexyl)-benzonitrile (PCH5) [*62788-05-0*]

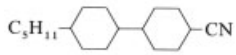

20
N 85 I

trans, trans-4-pentyl-1,1′-bicyclohexyl-4-carbonitrile (CCH5) [*77449-35-5*]

Another dramatic example is provided by the reversal of the linking group in compounds **21** and **22**:

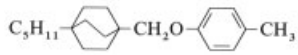

21
N 33.5 I

1-[(4-methylphenoxy)methyl]-4-pentylbicyclo(2.2.2)octane [*81472-49-3*]

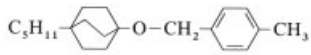

22
N–210 I

1-[(4-methylphenyl)methoxy]-4-pentylbicyclo(2.2.2)octane [*97129-82-3*]

However, the effect of isolating an ether group is not always as striking; for example, in the phenyldioxane derivatives the N – I transition temperatures are similar to those of their cyclohexane analogues–cf. **19** above and 4-(5-pentyl-1,3-dioxan-2-yl)benzonitrile (**23**) [*74800-62-7*].

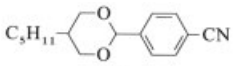

23
K 56 N (52) I

Subject to these restrictions, aromatic, alicyclic, and heterocyclic structures can be combined in liquid crystal molecules. The linearity of mesogenic molecules is compromised by lateral substitution. Large groups at the side of the molecule greatly lower the clearing point; smaller substituents have less effect. Fluorine occupies a unique position in that it substantially modifies the nature of the liquid crystal phase, while retaining a clearing point which is only modestly reduced relative to the unsubstituted parent compound. A lateral fluorine with the ring-fluorine bond directed toward the center of the molecule usually destabilizes smectic phases in favor of the nematic. Smectic phases are more likely to be tilted in the presence of such a substituent. When the ring-fluorine bond is directed away from the molecular core, however, smectic phase stability is little affected or enhanced. Halogens other than fluorine, together with short-chain alkyl groups, etc. give similar results, but with a much greater effect on the clearing point.

Compounds without strongly polar groups frequently show smectic phases over a wide temperature range. A nematic phase may then be introduced or extended by introduction of a polar group or by lateral substitution:

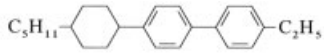

24
K 34 S 146 N 164 I

4-ethyl-4′-(4-pentylcyclohexyl)biphenyl (BCH 52) [*77702-36-4*]

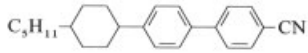

25
K 94 N 219 I

4′-(4-pentylcyclohexyl)biphenyl-4-carbonitrile (BCH 5) [*70711-80-7*]

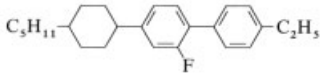

26
K 37 N 127 I

trans-4′-ethyl-2-fluoro-4-(4-pentylcyclohexyl)biphenyl (BCH 52 F) [*83171-55-5*]

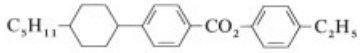

27
K 79 N 152 I

4-(4-pentylcyclohexyl)benzoic acid, 4-ethylphenyl ester [*91223-44-8*]

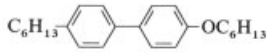

28
K 66 S 84 I

4-hexyl-4′-(hexyloxy)biphenyl [*56117-74-9*]

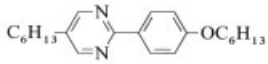

29
K 31 N 60.5 I

5-hexyl-2-[4-(hexyloxyl)phenyl]pyrimidine (PYP 606) [*51518-75-3*]

The choice of linking groups is important in determining smectic polymorphism. Schiff bases are frequently richer in smectic phases than their equivalent esters. Chain branching can also enhance smectic polymorphism, as well as lowering clearing points. An example of the effect of chain branching is demonstrated by

(±)-4′-octyl biphenyl-4-carboxylic acid, 4-(2-methylbutyl) phenyl ester (**30**) [*66424-45-1*]:

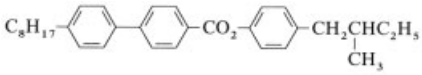

30

K 66 S_B (63) S_J 67 S_I 70 S_C 84 S_A 135 N 141 I

Smectic phases often predominate in liquid crystal materials having terminal dipoles with a strong component perpendicular to the molecular long axis. Examples are demonstrated by the hexyl ester of 4′-(decyloxy biphenyl)-4-carboxylic acid (**31**) [*62724-39-4*] and 1-(4′-methoxy biphenyl-4-yl)-1-pentanone (**32**).

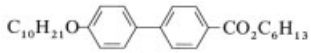

31

K 58 S_C 64 S_A 80 I

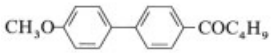

32

K 120.5 S_A (117.7) I

This brief discussion shows that rather subtle changes of molecular structure can have a radical effect on liquid crystal phase behavior. The existing basic knowledge of structure – property relations is adequate only when closely related materials are considered. This important area is the subject of review articles [15–17].

3.3. The Nematic Phase

The nematic phase is the least ordered of the liquid crystal phases, the molecules having their long axes arranged statistically parallel over large distances in the fluid (Fig. 1 A). The molecules do not have any positional order and can rotate freely about their long axes. Liquid crystal molecules with different terminal groups have no ordering of their disposition; that is, rotation of any molecule through 180° about its short axis leaves it in a state of the same energy. A vector drawn parallel to the preferred orientation direction of the molecular long axes is referred to as the *director*.

Although the molecules of a nematic liquid crystal align parallel to one another on average, considerable disorder occurs in this orientation in all real liquid crystal phases. This leads to the definition of an *order parameter* with the following properties:

1) The order parameter of the isotropic liquid is 0.
2) The (hypothetical) perfectly ordered liquid crystal with all molecular long axes parallel has an order parameter of 1.
3) The order parameter has the same symmetry properties as the nematic phase. In particular, the order parameter is unchanged if any molecule is rotated through 180° or if the angle it makes with the director is reversed.

The simplest, most commonly used order parameter is

$$S = 1/2\langle 3\cos^2\theta - 1\rangle$$

where θ is the angle between each long molecular axis and the director, and the angle brackets imply an average over a large number of molecules.

The Maier – Saupe theory of nematic ordering [18], [19] achieves semiquantitative agreement with experimental results when liquid crystal molecules are presumed to interact only via electrostatic dispersion forces. At first sight the approximations made in this approach appear unreasonable; in practice the neglected forces apparently make a large but almost isotropic contribution to the intermolecular potential and are not decisive in determining the stability of a liquid crystal phase [20]. The Maier – Saupe theory predicts that the order parameter decreases with increasing temperature T, but retains a finite value of ca. 0.42 at the nematic – isotropic transition $T_{\mathrm{N-I}}$, at which point it falls discontinuously to zero. The behavior may be approximated [4] by the equation

$$S \approx (1 - 0.98T/T_{\mathrm{N-I}})^{0.22}$$

where T is the absolute temperature.

An alternative approach to understanding the properties of the nematic phase is that of Onsager [6]. This theory seeks to avoid the worst assumptions of the Maier – Saupe approach by ascribing liquid crystalline ordering to repulsive interactions between molecules. Considering only these short-range forces, the Onsager

approach does not agree well with experimental results in regard to either S or density change at the clearing point.

Several attempts have been made to include both attractive and repulsive forces in a theory of nematic ordering, and computer simulations using increasingly realistic interaction potentials are now possible [21]. Even the most complex current calculations, however, fail to provide real insight into the changing liquid crystal behavior observed as molecular architecture is altered.

3.4. Physical Properties of Nematic Liquid Crystals

The static physical properties of liquid crystal materials resemble those of other organic liquids, but liquid crystal phases commonly show anisotropy in their properties [4], [22]; i.e., the magnitude of each physical property varies according to the relative geometry of the sample and the measuring equipment. Nematic liquid crystals are uniaxial, and each property is characterized by two principal values measured with the probe field parallel ($\|$) and perpendicular ($\perp$) to the director. The anisotropy of the physical property is defined as the difference between these two values.

3.4.1. Refractive Indices

The refractive indices (n) of nematic liquid crystals can be measured in a conventional instrument such as an Abbe refractometer with the director perpendicular to the direction of propagation of the light beam (Fig. 3). Such a sample geometry may be maintained by surface forces (see below). Under these circumstances an incident light beam is split into an ordinary and an extraordinary ray that are polarized perpendicular and parallel to the nematic director, respectively: these two rays are deflected through different angles appropriate to the corresponding principal refractive indices (n_0 and n_e) of the liquid crystal. Nematic liquid crystals are positive uniaxial crystals; i.e., $n_e > n_0$ and the birefringence Δn is positive:

$$\Delta n = n_e - n_0 > 0$$

On a molecular level, the refractive index of an organic medium is determined by the electronic polarizability. For a nematic liquid crystal, the polarizability both parallel and perpendicular to the molecular long axis must be considered. The contribution each molecule makes to the refractive indices of the phase is modified by the angle it makes to the director and thus to the order parameter. Rigorous interpretation of refractive index data is complicated by the anisotropy of the environment of each molecule. An extension of the Lorentz approach [4–6] leads to the prediction

$$\frac{n_e^2 - n_0^2}{\bar{n}^2 - 1} \sim S \tag{1}$$

$$\bar{n}^2 = \left(2n_e^2 + n_0^2\right)/3 \tag{2}$$

and is justified by the observation that the left hand side of Equation (1) is indeed closely proportional to S measured by independent means. The refractive indices of a nematic liquid crystal provide a means of measuring the order parameter S: Equation (1) is applicable if S can be obtained at a single temperature. Haller made the reasonable assumption that S tends to unity at zero Kelvin, and found [23] that linear extrapolation of refractive index data to this temperature was available from a plot (Fig. 4) of

$$\log\left(1 - T/T_{N-I}\right) \text{ vs. } \log\left[\left(n_e^2 - n_0^2\right)/\left(\bar{n}^2 - 1\right)\right] \tag{3}$$

This extrapolation therefore allows the proportionality constant of Equation (1) to be determined from refractive index measurements conducted at accessible temperatures. The constant number determined may be reapplied to the same data to obtain a value for S at each temperature of measurement. The denominator in the second expression of Equation (3) allows for the variation in density with temperature. Figure 5 shows the order function S as a function of temperature. For most applications, the birefringence of a nematic liquid crystal is the most important optical property because it defines the strength of interaction between the liquid crystal and light. Higher birefringence is observed in aromatic liquid crystals than in alicyclic ones and is further increased by π-bonded linking and terminal groups conjugated with the aromatic rings (Table 2). Birefringence

also increases in materials in which the clearing temperature is much higher than the measurement temperature. When low absolute refractive indices are required, suitable materials can be obtained by combining alicyclic liquid crystals with isotropic fluorinated dopants such as $C_7H_{15}CO_2CH_2(CF_2)_{10}H$ [24].

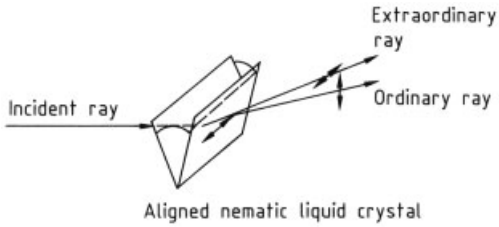

Figure 3. Double refraction in a sample of an aligned nematic liquid crystal

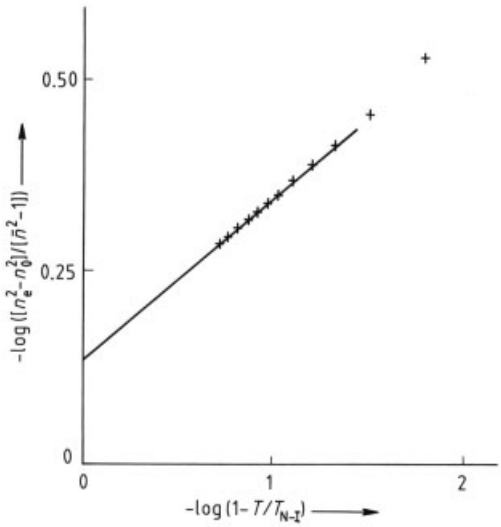

Figure 4. The Haller extrapolation for nematic mixture E7 consisting of two alkyl-4′-cyanobiphenyl-, one 4-alkoxy-4′-cyanobiphenyl- and one 4-alkyl-4′-cyanoterphenyl-molecules.

3.4.2. Dielectric Constants

The anisotropy of dielectric constants in the nematic liquid crystal phase mirrors that of the refractive indices; each material is characterized by two principal values parallel and normal to the director, denoted $\varepsilon_{\parallel}$ and $\varepsilon_{\perp}$, respectively. The Maier and Meier treatment of liquid crystal dielectric constants considers both the polarizability anisotropy of molecules $\Delta\alpha$ and a permanent dipole moment μ at an angle β to the long molecular axis [25]. As in the treatment of refractive indices, the internal field must be estimated, and the environment of the molecule is therefore assumed to be isotropic. The dielectric anisotropy obtained is

$$\Delta\varepsilon = \left[A \cdot \Delta\alpha - \frac{B\mu^2}{T}\left(1 - 3\cos^2\beta\right)\right] S \qquad (4)$$

where A and B are material-dependent constants. Although this expression has predictive usefulness for correlation of material properties with molecular structure, it fails to account for all the observed dielectric properties of liquid crystals. For a suitable value of β (ca. 55°), the second term in square brackets becomes zero, and any molecular dipole moment at this angle makes no contribution to the dielectric anisotropy. The result is a small positive dielectric anisotropy due to the positive polarizability anisotropy. Many liquid crystal molecules used in practice possess strong dipole moments almost parallel to the long molecular axis, produced by terminal substituents such as −CN and −NCS groups. In these cases, $\beta = 0$ and the positive dipolar term commonly dominates the polarizability in terms of their relative contribution to $\Delta\varepsilon$. Despite this, the dielectric anisotropy observed in terminal dipolar compounds is frequently much smaller (e.g., by a factor of two) than implied by Equation (4). Evidence has become compelling that a local antiparallel ordering of the dipole moments in these systems reduces their effective magnitude and their influence on dielectric properties [12], [26]. The degree of antiparallel correlation is a function of molecular structure and temperature, and depends on the composition of mixtures of liquid crystal materials. Consequently, the magnitude of this important effect cannot be predicted accurately. In a few cases, most notably esters, of the type **35**, the evidence indicates that a predominantly

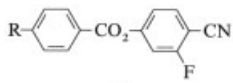

35

parallel ordering of the molecular dipoles exists [27]; these compounds are characterized by very high values of positive dielectric anisotropy. When the molecule has a lateral dipole, β is

Table 2. Birefringence and chemical structure of nematic liquid crystals

Structure	Name	CAS registry number	Empirical formula	Birefringence
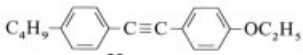 **33**	1-butyl-4-(4-ethoxy-phenylethynyl)benzene	[*85583-83-1*]	$C_{20}H_{22}O$	0.23
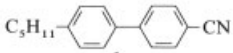 **6**	4′-pentylbiphenyl-4-carbonitrile	[*40817-08-1*]	$C_{18}H_{19}N$	0.21
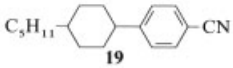 **19**	4-(4-pentylcyclohexyl)-benzonitrile	[*62788-05-0*]	$C_{18}H_{25}N$	0.16
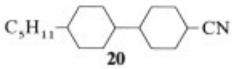 **20**	4′-pentyl-(1,1′-bicyclo-hexyl)-4-carbonitrile	[*77449-35-5*]	$C_{18}H_{31}N$	0.08
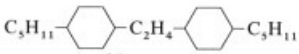 **34**	*trans,trans*-1,1′-(1,2-ethanediyl)bis-4-pentyl-cyclohexane	[*76988-15-3*]	$C_{24}H_{46}$	0.06

greater than 55° and the dipolar contribution to $\Delta\varepsilon$ is negative. However, even when $\beta = 90°$, a lateral dipole moment is only one-half as effective in inducing a negative $\Delta\varepsilon$ as an axial dipole is in inducing a positive one. Design and synthesis of liquid crystals with a very high negative dielectric anisotropy is difficult, the problems being compounded by the tendency of lateral substituent groups to distort the rodlike form of the liquid crystal molecule and adversely affect the thermal stability of the mesophases. Nevertheless, liquid crystal materials with a wide range of dielectric properties can be obtained by suitable selection of dipolar substituents (Table 3).

Dielectric relaxation leads to a change in the effective dielectric constant as the frequency is raised through a range where the molecular dipole can no longer contribute fully to the molecular response to a field. The relaxation commonly encountered reduces $\varepsilon_{\parallel}$ in a temperature- and material-dependent frequency range in the kilo- or megahertz region. The effect can be exploited to provide liquid crystal materials that have positive dielectric anisotropy at low frequencies but negative anisotropy above a crossover frequency [28]. Use of these two-frequency materials in display applications has been discussed but has never been widespread.

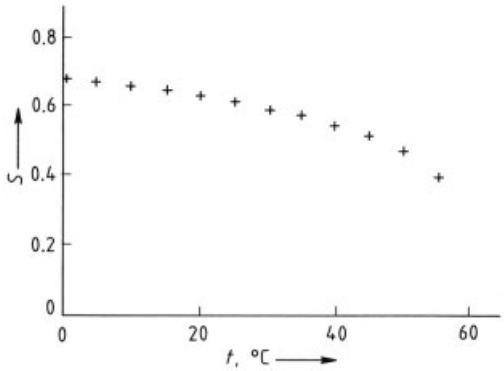

Figure 5. Order parameter vs. temperature for nematic mixture E7

3.4.3. Diamagnetic Constants

Common liquid crystal materials are diamagnetic and show a corresponding diamagnetic anisotropy $\Delta\chi$, which behaves in an analogous way to the dielectric anisotropy $\Delta\varepsilon$. The weak interaction of most liquid crystals with a magnetic field means that the effect is rarely used in practical applications. However, it is important to physicists because the small magnitude of the diamagnetic susceptibility means that the internal field problems relevant to dielectric properties do not arise and the theoretical treatment of

Table 3. Dielectric anisotropy ($\Delta\varepsilon$) of liquid crystal materials

Structure	Name	CAS registry number	Empirical formula	ΔE
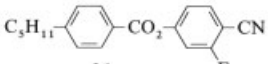 **36**	4-pentylbenzoic acid, 4-cyano-3-fluorophenyl ester (ME5NF)	[*86786-82-2*]	$C_{19}H_{18}FNO_2$	+ 50
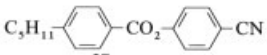 **37**	4-pentylbenzoic acid, 4-cyanophenyl ester	[*49763-64-6*]	$C_{19}H_{19}NO_2$	+ 20
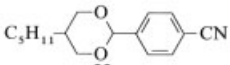 **23**	4-(5-pentyl-1,3-dioxan-2-yl)benzonitrile	[*74800-62-7*]	$C_{16}H_{21}NO_2$	+ 13
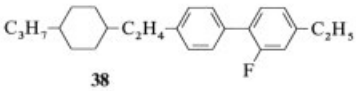 **38**	*trans*-4-ethyl-2-fluoro-4′-[2-(4-propylcyclohexyl)ethyl]biphenyl	[*95379-51-4*]	$C_{25}H_{33}F$	− 0.02
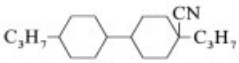 **39**	*cis,trans*-(4,4′-dipropyl)1,1′-bicyclohexyl-4-carbonitrile	[*88510-85-4*]	$C_{19}H_{33}N$	− 8
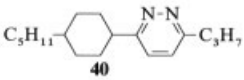 **40**	3-(4-pentylcyclohexyl)-6-propylpyridazine	[*84424-83-9*]	$C_{18}H_{30}N_2$	− 9
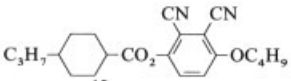 **12**	*trans*-4-propylcyclohexane carboxylic acid, 4-butoxy-2,3-dicyanophenyl ester	[*75941-67-2*]	$C_{22}H_{28}N_2O_3$	− 23

data is on a sure footing [5], [6]. Diamagnetic measurements are frequently used for the evaluation of order parameters:

$$S \sim \Delta\chi = (\chi_{\parallel} - \chi_{\perp})$$
$$\bar{\chi} = (\chi_{\parallel} + 2\chi_{\perp})/3$$

where $\chi_{\parallel}$ and $\chi_{\perp}$ are the diamagnetic susceptibilities measured with the magnetic field parallel and perpendicular to the nematic director, respectively. As in the case of order parameter measurements by using refractive index data (Section 3.4.1), an independent method is needed to establish the proportionality constant between S and $\Delta\chi$.

Most liquid crystal materials have a positive diamagnetic anisotropy; the 4-alkyl-4′-cyanobicyclohexanes (**21**) form a notable exception. The lack of practical applications means that little systematic optimization of diamagnetic properties has been attempted. Paramagnetic liquid crystals are known, but the anisotropy is apparently small [29].

3.4.4. Elastic Constants

Liquid crystals are unique among fluids in their elastic behavior. Although the definition of a fluid precludes the possibility of a static shear stress being sustained, a nematic liquid crystal can transmit a torque stress. The treatment of nematic liquid crystal elasticity elaborated by FRANK is based on the concept of a torque stress resulting in a corresponding strain [2–6]. Proportionality between stress and strain is assumed to hold for small deformations and is mediated by curvature elastic constants k having the units of force. All deformations of a liquid crystal sample can be expressed as a sum of three elementary distortions denoted splay, twist, and

bend. These distortions are opposed by the corresponding elastic constants k_{11}, k_{22}, and k_{33} (or k_1, k_2, and k_3, respectively) (Fig. 6). The elastic constants are of the order 10^{-11} N, leading to an elastic behavior considerably weaker than that observed in solids. Although the effect of the elastic constants can be detected, and their magnitude measured by standard means such as a torsion balance, a more convenient experimental technique is the observation of a Frederiks transition in a liquid crystal sample under an applied magnetic or electric field (see Section 3.5.1). For many samples, $k_{11} \approx k_{33} \approx 2\,k_{22}$

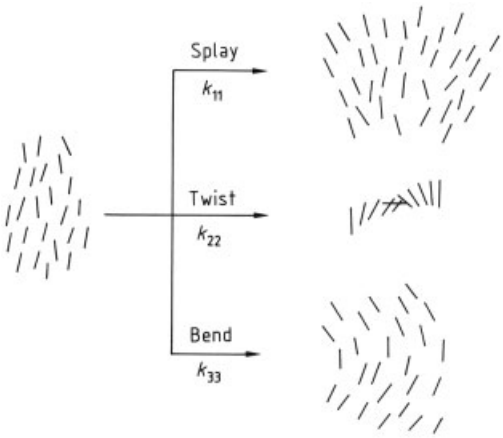

Figure 6. Elementary distortions of nematic liquid crystals

No satisfactory theory exists for predicting elastic constants from molecular structure. Information has been obtained by systematic measurement of a range of compounds, which allows an incomplete set of structural correlations to be formulated [30–34]. The ratio k_{33}/k_{11} is used most frequently as a measure of elastic properties both because of the body of data available and because it is important in determining the behavior of nematic liquid crystals in some practical applications. In molecules with long terminal alkyl substituents, k_{33}/k_{11} decreases. Terminal short chains and alkoxy substitution increase k_{33}/k_{11}. By contrast, evidence suggests that increasing the length : breadth ratio of the rigid core of liquid crystal molecule raises k_{33}/k_{11}. Aromatic and heterocyclic systems have lower k_{33}/k_{11} ratios than the corresponding alicyclic derivatives. The theoretical prediction that k_{ii} is proportional to S^2 is realized only approximately in practice. The proximity of a smectic phase alters the short-range order in a nematic liquid crystal and frequently reduces k_{33}/k_{11} [35–38]. Closer to an underlying smectic phase, a pretransitional increase in elastic constants is observed, especially k_{22} and k_{33}.

3.4.5. Light Scattering

An important consequence of the small magnitude of the elastic constants is that fluctuations in director orientation require little energy and can be thermally driven [5]. Such thermal fluctuations lead to changes in the effective refractive index experienced by a light beam propagating in the nematic liquid crystal. Fluctuations that occur at a short wavelength, comparable with the wavelength of light, lead to light scattering and opacity in the sample. Light scattering from nematic liquid crystals is several orders of magnitude stronger than from isotropic liquids and leads to the common description of the liquid crystal-to-isotropic transition temperature as the clearing point.

3.4.6. Viscosities

The viscous flow of nematic liquid crystals is complicated by two factors [4–6], namely, the anisotropy of the viscosities and the tendency for shear stress to rotate the director orientation in the sample and modify the experiment. Five independent viscosities are required to characterize a nematic liquid crystal fully; of these the three Miesowicz viscosities ($\eta_1 - \eta_3$) are the anisotropic analogues of ordinary fluid viscosities. The relative dispositions of the director and shear gradient allowing measurement of the Miesowicz viscosities are illustrated in Figure 7. The remaining two viscosities (γ_1 and η_{12}) have no direct analogue in isotropic liquids. Measurement of the viscosity of a nematic liquid crystal (η_{45}°) in a configuration intermediate between those shown in Figures 7 A and 7 B (i.e., with the director at 45° to the shear direction) leads to a contribution from a further term:

$$\eta_{45^\circ} = 1/2\,(\eta_1 + \eta_2) + 1/4\eta_{12}$$

The remaining rotational viscosity γ_1 opposes a pure rotation of the director, which can take place without any flow.

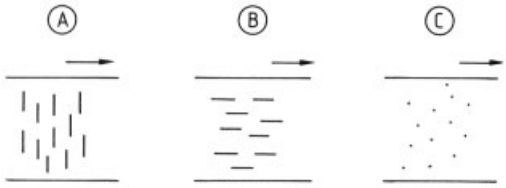

Figure 7. Shear and director geometry for measurement of the Miesowiscz viscosities η_1 (A), η_2 (B), and η_3 (C)

All these quantities can be measured by using a strong magnetic field to maintain uniform alignment, but the experiments are difficult and very few reliable sets of data are available for nematic liquid crystals.

In normal viscometry equipment (e.g., a capillary viscometer), the shear often results in an alignment of the director of nematic liquid crystal materials at a small angle to the direction of flow [22]. The flow viscosity measured under these circumstances approximates the Miesowicz viscosity η_2. As in measurement of η_2, therefore, the viscosity measured in the nematic liquid crystal phase is significantly lower than that of the isotropic liquid, and a discontinuity appears in the viscosity – temperature plot at the clearing point. The temperature dependence of viscosity otherwise approximates an Arrhenius behavior similar to an isotropic liquid [39].

3.4.7. Surface Alignment

The interaction of nematic liquid crystals with solid surfaces provides an important means of defining the director, and its control is crucial in almost all of the practical applications of liquid crystals. If a nematic liquid crystal is placed in contact with a glass surface, for example, the director lies at an undefined angle which is generally tilted out of the plane of the surface. Pretreatment of the glass, however, can determine unambiguously the director alignment at the surface. Rigorous cleaning or treatment with a surfactant leads to *homeotropic alignment* in which the director is perpendicular to the surface. Effective surfactants include silicones, quaternary ammonium salts, and metal complexes, which all have a long alkyl chain bonded to a polar head group. The head group is assumed to bond to the solid surface leaving the pendant alkyl group free to associate with the liquid crystal.

A wide range of treatments are also available to induce alignment parallel to a surface, sometimes termed *homogeneous or planar alignment*. Many polymers [e.g., polyamides, polyimides, and poly(vinyl alcohol)] are effective and can be deposited from an appropriate solvent.

In the absence of treatment, the director lies in an arbitrary direction in the surface plane. Mechanical rubbing of the polymer, however, imposes a preferred orientation direction on the surface. The director then lies along the rubbing direction and is tilted out of the surface plane by a small angle of ca. 1 – 5°. Rubbing bare glass with a fine abrasive, or oblique evaporation of a range of inorganic materials (e.g., magnesium fluoride or silicon monoxide) from a source ca. 30° out of the surface plane, leads to a zero tilt angle. Evaporation from a source ca. 5° out of the surface plane, however, gives a high tilt angle of ca. 20 – 30°. Because of the importance of surface alignment in the construction of displays, the technology has been explored in considerable depth [40], [41].

3.5. Nematic Liquid Crystal Materials for Display Use

The electro-optic display industry continues to motivate much of the development of liquid crystal materials worldwide. A range of physical properties must be manipulated simultaneously to optimize a fluid for each application. This is achieved by both the synthesis of new compounds and their incorporation into mixtures in new combinations. Since the invention of the first stable, colorless materials for displays in the 1970s, an unwillingness to accept new materials with photolytic, hydrolytic, or oxidative sensitivity has existed.

Materials for display use invariably consist of mixtures of liquid crystal compounds [42]. The first requirement for any such material is a nematic liquid crystal phase range extending beyond the anticipated working range of the device. For watch and calculator displays, a phase range from − 10 to + 60 °C is usual, whereas for industrial, automotive, and military applications a range from − 40 to + 90 °C is typical. No single compound exists that can meet these targets; mixtures extend the lower end of the phase

range by eutectic depression. The clearing point of a mixture T_{cl} does not suffer eutectic depression, but follows an additive relationship to the clearing points of the components T_i and their concentrations C_i:

$$T_{\mathrm{cl}} = \sum_i C_i T_i$$

The choice of components included in a mixture for a given application is based on the properties of available compounds. Account must be taken of the nonlinear variation of liquid crystal properties as the composition of a mixture changes, and mixture formulation usually relies heavily on measurement of the performance of trial samples in test devices.

In addition to their use in displays, nematic liquid crystals have been proposed for use in field sensors, thermal mapping, and nondestructive testing. Particular interest exists in third-order nonlinear optic processes in nematic liquid crystals that offer very high nonlinearity but slow speed [43], [44]. Both thermal and orientational nonlinearities exist.

3.5.1. The Frederiks Transition

Consider a thin sample of a nematic liquid crystal of positive dielectric anisotropy, contained between two parallel glass plates. Each plate is treated to induce planar alignment of the liquid crystal director along the same axis with zero tilt angle, and further carries a transparent electrically conducting layer beneath the aligning layer. The nematic liquid crystal adopts a uniform alignment throughout the thickness of the sample and is constrained by elastic forces that propagate from the surfaces (Fig. 8 A). If a voltage is applied between the plates, no torque on the liquid crystal molecules results, but the electric field tends to stabilize a tilted configuration of the director in the center of the cell because the tilt increases the apparent dielectric constant of the medium. In practice, no effect on director behavior is observed when the applied voltage increases until a critical value is reached at which the stabilization of tilt due to the increase in dielectric constant matches the elastic energy required to distort the director field. At this point a small tilt angle in the center of the cell, defined initially by a thermal fluctuation, is stabilized indefinitely by the applied voltage [45].

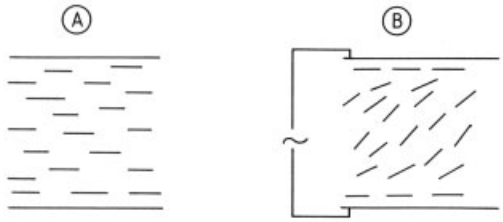

Figure 8. A Frederiks splay cell showing the off (A) and on (B) states

The initial deformation induced in the director orientation is a splay effect, and the onset of distortion is mediated by the elastic constant k_{11} (see Section 3.4.4). When the applied voltage is increased, the tilt angle at the center of the cell increases and the layers of liquid crystal closer to the surfaces also begin to adopt a significantly tilted orientation (Fig. 8 B). The deformation becomes a combination of splay and bend, and both k_{11} and k_{33} become significant. If the dielectric anisotropy of the liquid crystal is large, the variation in tilt through the thickness of the cell results in a corresponding change in the effective dielectric constant through the cell and a consequent redistribution of the electric field so that it is no longer uniform throughout the thickness. Both of these effects have consequences for the detailed form of the switching behavior in any particular liquid crystal material. In the cell described above, the tilt angle of the nematic director in the absence of an applied voltage is zero, and a degeneracy therefore exists between two states of equal and opposite tilt when a voltage above the threshold is applied. The direction of tilt is normally determined by a random thermal fluctuation, and different areas of the cell may tilt with the opposite sense. In this case, the uniformly tilted areas are separated by a defect known as a wall (Fig. 9).

At high voltage, the tilt of the director at the center of the cell effectively becomes 90°. The director at the aligning layers, by contrast, remains tied to the surface and consequently has a tilt of zero. Between these two extremes, the director adopts an equilibrium configuration determined by the conflicting elastic and dielectric forces. Under very high applied voltages the liquid crystal director becomes aligned perpendic-

ular to the surface except in a very thin boundary layer on each plate.

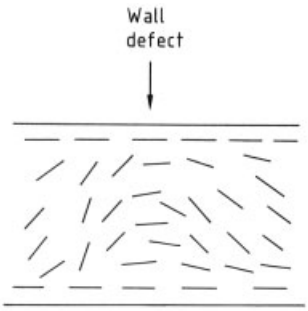

Figure 9. Wall defect separating domains of opposite tilt in a Frederiks cell

These phenomena may also be induced by an applied magnetic field, with only minor differences arising. In particular, the critical magnetic field is dependent on the thickness of the liquid crystal cell, whereas the critical voltage is not.

This field-induced realignment of a nematic liquid crystal is one of three types of Frederiks transition, the understanding of which is fundamental to the application of liquid crystals in electro-optic devices. Because the initial deformation is a splay, it is frequently referred to as the *splay* (*S*) *mode* transition. The remaining two Frederiks modes, the *bend* (*B*) and *twist* (*T*) *transitions,* are analogous to the splay transition but use different geometries of applied field and initial director configuration. The various types of Frederiks effect and the expressions for their threshold fields are summarized in Figure 10. Typical critical voltages (V_c) for common nematic materials are ca. 1 – 10 V, whereas in cells of the usual 10 – 100μm thickness, a threshold magnetic field (H_c) of $10^3 - 10^4$ gauss is common.

3.5.2. Electrically Controlled Birefringence

Consider a nematic liquid crystal cell in the configuration of the bend mode Frederiks transition (Fig. 10). The cell is placed between crossed polarizers. Because of the alignment of the nematic liquid crystal, the cell shows no birefringence and no light is passed. If a voltage higher than the Frederiks threshold is applied, the liquid crystal is reoriented. As a result, light propagates through the cell as ordinary and extraordinary rays that are polarized perpendicular and parallel to the direction of director tilt, respectively. The induced birefringence of the cell allows the transmission of light of intensity I:

$$I = I_0 \sin^2 (2\pi d \Delta n/\lambda) \tag{5}$$

where I_0 is the intensity of the incident light of wavelength λ and d is the cell thickness. Under a high applied voltage, Δn equals Δn of the liquid crystals. At intermediate voltage, a series of maxima and minima in the transmitted light intensity is observed. For practical materials and cell thicknesses, a $d \cdot \Delta n$ product of many times λ can be obtained, and a large number of such interference peaks and troughs is observed. One application of this technique is in measurement of the elastic constants of nematic liquid crystals. The first change in transmitted light intensity occurs at V_c. The constant k_{33} is obtained by substitution into the equation in Figure 10, $\Delta \varepsilon$ being measured independently; k_{11} and k_{33} can also be obtained from the switching curve above the threshold voltage. If a magnetic field is employed, changes in the geometry of the experiment may be used to measure the three elastic constants directly.

When mounted between polarizers, the Frederiks cell has the properties of an electro-optic modulator. To lower the degeneracy of the switched state and ensure that the tilt of the director is aligned correctly with the polarizers, surface treatment is required which induces a slightly tilted homeotropic alignment. This may be achieved by depositing a surfactant onto a rubbed polymer surface (see Section 3.4.7). Transmission of light in the field-on state is determined by Equation (5), with Δn replaced by the induced birefringence of the nematic liquid crystal. Transmission depends strongly on the wavelength of light λ, thickness of the cell d, and birefringence and, hence, temperature of the liquid crystal. In practice, if the cell is operated between a birefringence $d \cdot \Delta n$ of zero and $\lambda/2$, the device provides high contrast without undue sensitivity to temperature or wavelength.

The electro-optic effect described above is the basis of a class of liquid crystal displays known by various acronyms including the DAP (distortion of aligned phase) and ECB (electrically controlled birefringence) effect. This type of display has not been widely exploited because of the difficulties encountered in obtaining the

required alignment in a reproducible and stable manner, and in constructing devices with the necessary close tolerance on the liquid crystal layer thickness. Improvements in liquid crystal cell construction have, however, alleviated this problem, and the ECB display may find widespread application in the future.

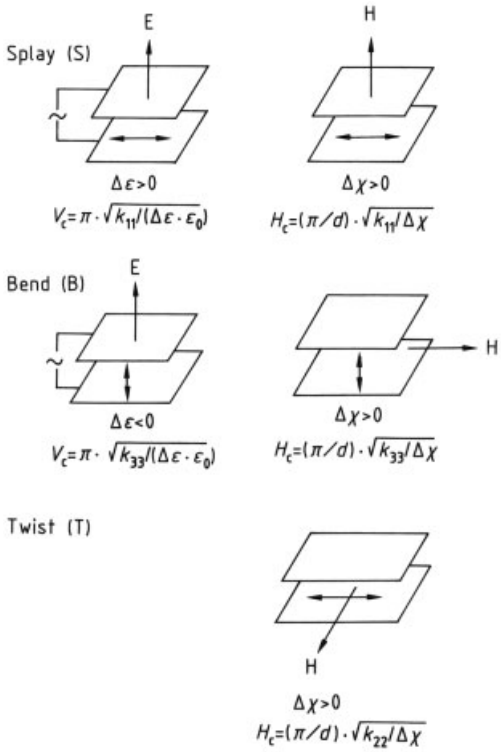

Figure 10. Geometries and threshold fields of Frederiks transitions
Double-headed arrows indicate the alignment of the liquid crystal molecules. V_c = critical voltage; k_{11}, k_{22}, k_{33} = elastic constants; ε = dielectric constant; H_c = critical magnetic field; d = cell thickness, χ = diamagnetic susceptibility.

3.5.3. The Twisted Nematic Display

Problems of sensitivity to temperature, wavelength, and cell thickness are common to all liquid crystal displays that rely on variable retardation of light to generate their optical effect. All these problems are avoided in the twisted nematic display, which is the dominant display in digital watches, calculators, and other applications involving devices of moderate complexity [46].

Principles. The twisted nematic display consists of two parallel glass plates ca. 6 – 10 µm apart, the gap being filled with nematic liquid crystal [47]. Each glass plate carries a transparent electrically conducting layer, usually of sputtered indium tin oxide, etched into a pattern defining the switchable areas of the display. A rubbed polymer layer is used to provide planar alignment of the liquid crystal director, with a small surface tilt angle. The directions of rubbing of the polymer layers on the two glass plates are mutually perpendicular, so the alignment directions on the plates are also at 90° to one another. The liquid crystal adopts a twisted conformation, its alignment making a 90° turn through the thickness of the cell (Fig. 11 A).

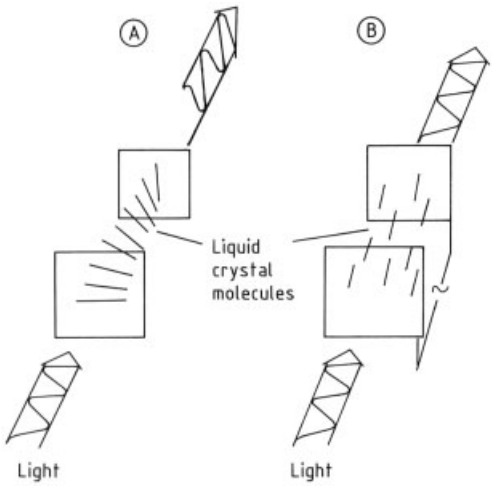

Figure 11. Twisted nematic display showing off (A) and on (B) states

The optical properties of the twisted nematic layer are unique [48]: when $d \cdot \Delta n$ is large (the Mauguin limit), plane-polarized light entering the twisted nematic cell with the electric vector either parallel or perpendicular to the rubbing direction is transmitted through the cell, but the plane of polarization is rotated through 90° (Fig. 11 A). This behavior can be closely approached by using practical materials and cell thicknesses. If the twisted nematic cell is placed between crossed polarizers, the whole assembly transmits light.

The optical behavior of the twisted nematic layer has been analyzed by Gooch and Tarry

[49]. In a practical twisted nematic device, emergent light is elliptically polarized, and perfect transmission is achieved only for certain values of the parameter U, where

$$U = 2d \cdot \Delta n / \lambda \tag{6}$$

The general expression for transmitted light intensity is

$$I = I_0 - \frac{I_0 \sin^2 \left[(\pi/2) \left(1+U^2\right)^{1/2} \right]}{1+U^2} \tag{7}$$

which implies ideal behavior for values of U equal to $\sqrt{3}$, $\sqrt{15}$, etc., or for values of $d \cdot \Delta n$ equal to 0.48 µm, 1.07 µm, etc., for light at the peak of the eye's sensitivity curve.

The denominator in Equation (7) accounts for the perfect behavior in the Mauguin limit and also means that successive variations in the transmitted intensity become rapidly less significant as U increases. For values of U greater than about unity or for values of $d \cdot \Delta n$ greater than 1 µm, the light-guiding properties of the cell are sufficiently independent of cell thickness and wavelength. Most twisted nematic cells are constructed with a $d \cdot \Delta n$ of ca. 1 µm or higher, but if defects lead to areas in the cell being thinner than specified, the display shows bands of coloration. Some displays intended for applications in which a wide viewing angle is important, are constructed with a $d \cdot \Delta n$ of ca. 0.48 µm and are referred to as "first minimum" displays. In this case, a good appearance depends on accurate control of the cell thickness.

When voltage is applied to a twisted nematic cell, reorientation of the liquid crystal director analogous to the Frederiks splay mode transition takes place (see Section 3.5.1). Under a high applied voltage, the liquid crystal director becomes perpendicular to the plane of the cell except in a very thin layer adjacent to the glass plates. Under these circumstances the twisted structure of the nematic layer ceases to exist, the cell no longer rotates the plane of polarization of light, and the activated area appears black between crossed polarizers (Fig. 11 B). The familiar black digits seen on liquid crystal displays are merely areas where crossed polarizing filters extinguish light falling on the display.

In practice, the properties of a twisted nematic display are modified by the finite voltage applied. A liquid crystal drive voltage approximately three times the Frederiks threshold is commonly applied; under these conditions the tilt angle at the center of the cell does not reach 90°, providing an angle-dependent contrast in the display. Switching is obtained at the lowest voltage in a single quadrant of the display determined by the alignment direction on each glass plate. Practical displays give good contrast over a wide cone in the preferred viewing quadrant.

The twisted nematic cell is a very successful device that embodies many of the advantages and disadvantages of liquid crystal displays. It is a passive device; i.e., it does not emit light but modulates the ambient light that falls on it. The display appears dull under poor lighting, but it does not suffer from "washout" even in direct sunlight. Like most of the devices discussed here, it is a field effect device which does not rely on the passage of a current for its operation. This and the low working voltage contribute to an extremely low power consumption which makes twisted nematic displays ideal for portable, battery-operated, and even solar-powered equipment.

Multiplex Addressing. Another difficulty that remains to be addressed is achieving a high information density on liquid crystal display. The twisted nematic display described above is constructed by making a separate electrical connection to each active area of the cell (pixel), thereby using a total of 24 connections to a simple 31/2-digit watch display. On the same basis, an 8-digit calculator display requires 65 connections and any form of graphics display would be impractical. The solution lies in multiplex addressing of the display, which reduces the number of connections needed at the expense of restricting the voltages that can be applied to both "on" and "off" segments of the cell.

The twisted nematic display, like other nematic liquid crystal displays, responds to the root mean square (RMS) voltage applied to it [50]. A multiplex scheme can be implemented for RMS-responding displays, by configuring both backplane and the front of the display so that the patterns on the two glass plates are topologically equivalent to the rows and columns of an $X-Y$ matrix (Fig. 12). The number of columns in the matrix is equal to the multiplex rate M, and for

a display with N pixels the necessary number of connections becomes $N/M + M$.

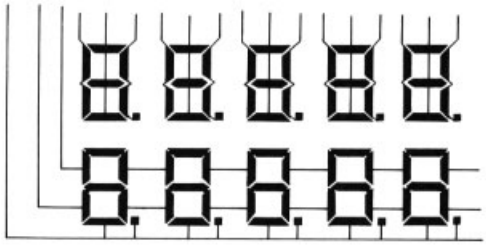

Figure 12. Electrode patterns allowing three-way multiplex drive of a liquid crystal display composed of seven-segment digits
Connections for applying the bias and data voltages are shown in the upper and lower rows, respectively.

Multiplexed displays are addressed by supplying information to the pixels on column at a time, and using the RMS response of the liquid crystal to retain the information while the other columns are driven in turn. The whole display is, therefore, driven in M such addressing time slots. A voltage pulse (bias voltage) is applied to each column of the matrix in turn, and information is written on the rows by applying a simultaneous data pulse either in phase ("off" pixels) or out of phase ("on" pixels) with the column voltage. These pulses also appear as error signals on pixels lying on columns to which no bias voltage is applied. Optimal discrimination between "on" and "off" pixels can be achieved for a particular ratio of data (V_{d}) to bias (V_{s}) voltage which is characterized by the select level S [50]:

$$S = V_{\mathrm{d}} / (V_{\mathrm{s}} + V_{\mathrm{d}}) = 1 + \sqrt{M} \tag{8}$$

The RMS voltages applied to both "off" and "on" pixels are determined by the following equations:

$$V_{\mathrm{OFF}} = \frac{V_{\mathrm{SUPPLY}}}{S} \left[\frac{M - 1 + (S-2)^2}{N} \right]^{1/2} \tag{9}$$

$$V_{\mathrm{ON}} / V_{\mathrm{OFF}} = \left[\frac{M - 1 + S^2}{M - 1 + (S-2)^2} \right]^{1/2} \tag{10}$$

When Equation (8) is satisfied, Equation (10) can be reduced to

$$V_{\mathrm{ON}} / V_{\mathrm{OFF}} = \left[\frac{M^{1/2} + 1}{M^{1/2} - 1} \right]^{1/2} \tag{11}$$

A RMS multiplexed drive scheme, therefore, applies a voltage to both "on" and "off" pixels in a ratio determined by the multiplex rate (M). The magnitude of the voltages is also determined by Equation (9): in practice an integral value of S is used, close to (but not greater than) the optimal value of Equation (8).

Multiplexed addressing is the norm for liquid crystal displays, except for applications in which a wide viewing angle and a wide operating temperature range are of paramount importance. Typical multiplex levels employed are 2-way for watches and clocks, 3-way for calculators, and 8-way for a single-line alphanumeric display. Graphics displays use multiplex levels up to 256 ways or even higher. The threshold voltage of any material used in a multiplexed liquid crystal display must be adjusted to be a little higher than V_{OFF} to avoid turning on the nonselected pixels ("crosstalk"). The threshold voltage of a material in a twisted nematic cell parallels the Frederiks threshold voltage, which may be varied by adjusting $\Delta\varepsilon$. Liquid crystal mixtures are formulated which contain liquid crystal components of both large positive and small $\Delta\varepsilon$ [51] and variation of their relative concentration allows the threshold voltage to be adjusted to any value above ca. 1 V. As the desired multiplex level increases, $V_{\mathrm{ON}}/V_{\mathrm{OFF}}$ decreases and approaches unity, and the display fails to generate adequate contrast in areas addressed by V_{ON}, given the above limitation on V_{OFF}. This can be rectified only by improving the steepness of the optical switching curve. Significant progress in this direction can be made by optimizing the dielectric and elastic properties of nematic liquid crystal mixtures used in twisted nematic displays [52], but the twisted nematic display remains poorly suited to multiplex drive at ratios above ca. 32 ways.

3.5.4. The Supertwisted Nematic Display

Invention of the supertwisted nematic display has alleviated the difficulties discussed in Section 3.5.3 [53], [54]. It relies on display cells similar in construction to the twisted nematic display, but by a combination of surface alignment and doping the liquid crystal with a chiral additive, the cell is made to contain a higher degree of twist between 180° and 270°. As the

twist angle increases, the switching curve becomes steeper, until at some point it is vertical. Further increase in the twist angle results in bistable behavior where, over a certain range of applied voltage, both a high tilt and a low tilt director configuration are stable [55]. A practical limit arises from the occurrence of a two-dimensional texture at high twist angles, which scatters light and prevents smooth switching of the display. The effect can be alleviated by using alignment methods to give high surface tilt. Special polymers (e.g., poly-1,4-phenylene and fluorinated polyimides) can provide tilt angles up to 10° or more, which is adequate for twist angles up to 270°.

With a high twist angle (180 – 270°), construction of displays that guide light in the manner of a twisted nematic cell is impractical, and a birefringence mode of operation is used. The resulting device, denoted the supertwisted nematic (STN) or supertwisted birefringence effect (SBE), provides much better contrast and viewing angle than a twisted nematic display at moderate to high multiplex rates. Disadvantages of the technology are a relatively slow response speed, the need to make cells with very accurately defined and uniform thickness, and undesirable coloration of the display. Most supertwist displays operate in a blue (on state) on yellow (off state) mode, both states having their own birefringent interference colors.

The color of the off state can be removed by an ingenious but expensive technique [56]; a second liquid crystal cell having the same magnitude but opposite sense of twist is laminated to the first. Electrical connections are made only to one cell. The second cell acts as a compensator plate, eliminating the optical retardation in the off state of the display. Less perfect compensation may be achieved with a birefringent plastic sheet. Other, less satisfactory routes to neutral, colored, supertwisted nematic displays include the use of very low $d \cdot \Delta n$ products [57] or dyes.

3.5.5. Active Backplane Liquid Crystal Displays

The limitations on display complexity encountered in using a nematic liquid crystal display can be circumvented by construction of the device on a substrate carrying an array of active switching elements (→ Display Technology, Chap. 4.2.1.). Of the wide range of devices available [3], field effect transistors have been most widely used. Liquid crystal color television sets represent the state of the art in this technology, and most use amorphous silicon field effect transistors to control a twisted nematic display less than 10 cm across the diagonal. Prototype displays with a 35-cm diagonal have, however, been demonstrated. Color is obtained by a fine pattern of red, green, and blue filters deposited on the pixels.

3.5.6. Guest – Host Displays

An important property of liquid crystal materials, which is exploited in nematic displays, is their ability to align rod-shaped solutes that do not themselves show mesophases. This guest – host effect is particularly useful when the solute is a dye molecule because many rodlike dyes have an anisotropic extinction coefficient. A solution of such a dichroic dye in a nematic liquid crystal shows strong absorption of light only when the liquid crystal director is aligned with the polarization axis of the light. The effect is quantified in terms of the order parameter of the dye:

$$S = \frac{A_{\parallel} - A_{\perp}}{A_{\parallel} + 2A_{\perp}}$$

where $A_{\parallel}$ and $A_{\perp}$ are the absorbances of a cell to the light polarized parallel or perpendicular to the director of a dyed nematic liquid crystal contained therein. Practical dyes with an order parameter of ca. 0.8 are available in a range of colors [58].

The *Heilmeier display* utilizes a Frederiks splay mode transition in a dichroic dyed nematic liquid crystal [59]. One polarizer is used, with its axis parallel to the alignment on the adjacent glass plate, giving a strong absorption of light by the dye (Fig. 13 A). Application of a voltage to the cell leads to tilting of the director, and the absorbance decreases as the dye transition moment approaches a direction perpendicular to the polarizer axis (Fig. 13 B). To obtain full contrast, the liquid crystal host and the dye must realign perpendicular to the plane of the cell throughout its thickness. This requires application to the

cell of higher voltage than the 3 – 5 V commonly used for twisted nematic cells. Because switching begins at the Frederiks threshold voltage, the slope of the optical switching curve is not steep and the Heilmeier display is not suitable for multiplex drive. The optical density of the display is proportional to its thickness so that more accurate spacing is required than for a twisted nematic display, although much less stringent than for supertwisted nematic displays. Advantages of the Heilmeier display include the elimination of one polarizer, a better angle of view than in the twisted nematic cell, and brighter appearance.

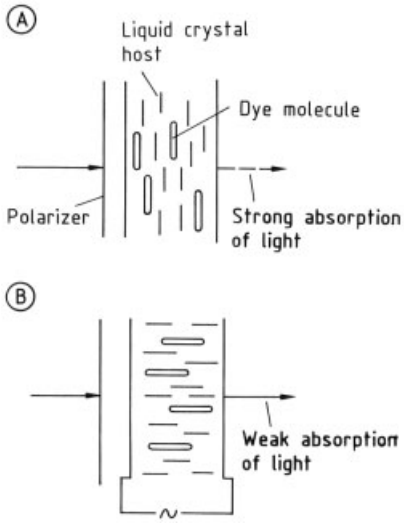

Figure 13. Heilmeier display cell in the off (A) and on (B) state

The second guest – host display is the *White – Taylor* or *phase change display* [60]. This display employs a dichroic dyed nematic liquid crystal which is doped with a chiral additive to induce a helical structure (i.e., a cholesteric phase) with a pitch length shorter than the thickness of the liquid crystal layer. The cell is treated to induce planar alignment, and the liquid crystal adopts a twisted structure with the helical axis perpendicular to the plane of the cell. The tightly twisted liquid crystal structure cannot guide the plane of polarization of light, which propagates through the helix as a combination of two elliptically polarized rays.

In the presence of a dye, both rays can be absorbed efficiently; a low birefringence and short pitch favor this. When a voltage is applied, the switching does not resemble a Frederiks transition, because the dominant elastic force is no longer imposed by surface forces but is associated with the pitch of the liquid crystal. The helix is progressively distorted and unwound, a process that is complete at the associated critical voltage V_c:

$$V_c = \frac{\pi^2 d}{P}\left[\frac{k_{22}}{\Delta\varepsilon\cdot\varepsilon_0}\right]^{1/2}$$

where P is the helical pitch length in micrometers. Application of a higher voltage results in realignment of surface layers of dyed liquid crystal, which may make a significant contribution to the full contrast. The White – Taylor display is used in applications where its almost hemispherical angle of view is required. The absence of a polarizer also results in a bright appearance. However, a high drive voltage is necessary, and a contrast as high as that of the twisted nematic cell is very difficult to achieve. Typical uses are in backlit displays for instrument and aerospace applications.

3.5.7. The Dynamic Scattering Display

The display modes described in Sections 3.5.1, 3.5.2, 3.5.3, 3.5.4, 3.5.5, and 3.5.6 are all based on field effects; optical properties are changed by application of a voltage, without the need for current to flow in the device. By contrast, the dynamic scattering device, which initiated the widespread development of liquid crystal materials and technology, relies on current flow through the cell. When a current passes through a nematic liquid crystal of negative dielectric anisotropy, the passage of ions causes a variety of realignment effects. This results in a range of two-dimensional periodic structures that are dependent on the magnitude and frequency of the applied voltage [5]. In nematic liquid crystal materials with a resistivity up to ca. $10^9\ \Omega\cdot$cm and at low frequency, and moderate applied voltage these structures give way to a chaotic, highly turbulent state that scatters light strongly [61]. In the Schiff 's base derivatives which provided the first practical room temperature nematic liquid crystals (e.g., **4** in Table 1), conductivity was mediated by impurities and degradation caused by traces of water and oxygen entering the device. The resulting displays were extremely unreliable and had a short working life. Later dynamic

scattering displays utilized highly purified liquid crystals in hermetically sealed cells; conductivity was provided by addition of an ionic dopant. These displays were capable of good performance and long lifetime when correctly fabricated. The appearance of the scattering state is very dependent on the lighting and viewing conditions, and the necessity to pass current through the cell meant that the power consumption of dynamic scattering displays could never be as low as that of a field effect device. After invention of the twisted nematic display and a variety of very stable nematic liquids of positive dielectric anisotropy, the dynamic scattering mode fell out of use.

3.5.8. Polymer Dispersion Liquid Crystal Devices

The displays described in Sections 3.5.1, 3.5.2, 3.5.3, 3.5.4, 3.5.5, 3.5.6, and 3.5.7 consist essentially of glass plates having a suitable gap between them filled with liquid crystal material. This limits the size and type of device that can be made. Plastic displays allow the production of very large devices and possibly electrically controlled windows. The technology has been called the *nematic curvilinear aligned phase* (NCAP) device [62]. In this display, small droplets of a nematic liquid crystal material are dispersed in a polymer matrix [e.g., poly(vinyl alcohol)] by precipitation or emulsification. Once formed, the droplets adopt an alignment determined by the nature of the polymer surface, shape of the droplets, and statistical factors. Uniform alignment geometry throughout the droplet is impossible for topological reasons. Light passing through the polymer film is scattered by the droplets, giving an opaque appearance. Application of a voltage to electrodes on opposite sides of the polymer sheet results in realignment of the droplets to bring the director parallel to the field. If the refractive index of the polymer matrix equals n_0 of the liquid crystal, no effective difference in refractive index exists, and therefore no scattering of light that passes from polymer to liquid crystal or vice versa. The film appears clear and transparent under these conditions. Disadvantages of the device are that a high voltage is required and it is not suited to multiplex drive.

3.6. Thermotropic Cholesteric Liquid Crystals

Addition of an optically active dopant to a nematic liquid crystal induces a helical structure in the director configuration (see Section 3.5.6). The nematic phase may be regarded as the special case in which the helical pitch is infinite. The chiral nematic (i.e., cholesteric) phase has distinctive physical properties, particularly when the pitch is short.

If a chiral nematic liquid crystal is placed between glass plates and subjected to shear, uniform alignment (Grandjean texture) is induced in which the helical axis is perpendicular to the plane of the cell. In this state, the liquid crystal has dramatic optical properties [2], [5], [6] which are highly wavelength dependent and center on the wavelength that matches the pitch P of the helix:

$$\lambda = \bar{n} P$$

where $\bar{n}$ is the average refractive index. Centered on this wavelength is a band of width $\Delta\lambda$,

$$\Delta\lambda = \Delta n \cdot \lambda / \bar{n}$$

in which incident light parallel to the helical axis undergoes Bragg-like scattering from the helical structure, resulting in selective reflection of a circularly polarized beam. Only one polarization is reflected under these circumstances: a cholesteric liquid crystal with a right-handed helix reflects only right-handed polarized incident light, and left-handed polarized light passes through without hindrance. Moreover, the light is reflected as a right-handed polarized ray, in contrast to ordinary surfaces which reverse the sense of circular polarization on reflection. If a cholesteric liquid crystal with a left-handed helix is used, the same phenomena occur with obvious changes of polarity. Efficient reflection of light can be obtained from layers with a thickness up to 10 µm.

If light is incident from a glancing angle, the behavior is different [63]; second and higher order reflections may be observed. The principal effect, however, is that selective reflection occurs at shorter wavelengths, as predicted by the Bragg condition

$$\lambda = \bar{n} P \cos\theta$$

where the angle of incidence θ is measured from the normal to the plane of the liquid crystal layer.

Light falling outside the reflection band passes through the liquid crystal layer but undergoes rotation of the plane of polarization, which becomes very large and changes sign [64]. Applications of cholesteric liquid crystals are, however, derived from the reflection effect. At suitable pitch lengths, visible light of a single color is reflected from the liquid crystal layer. On passing through the clearing point transition, this reflecting color disappears abruptly. On cooling, the cholesteric phase forms a "focal conic" texture (so named from the appearance of conic sector curves in the texture when viewed under a polarizing microscope) that scatters light without any pronounced wavelength peak, and coloration of the liquid crystal is very weak. Such a clearing point transition therefore forms the basis of an "irreversible" temperature indicator: the operating temperatures of mixtures can be adjusted by use of variable proportions of components with high and low clearing points. The focal conic texture reverts to the planar texture spontaneously over a period of hours or days, or can be rapidly realigned by shear and, therefore, is not a truly permanent record of thermal history.

Liquid crystal mixtures can be prepared which possess a cholesteric phase and, at a lower temperature, a smectic A phase. When such a smectic phase is approached by reducing the temperature of the chiral nematic phase, the helix unwinds. The enthalpy of the smectic–cholesteric transition is very small, and pretransitional effects on pitch can extend tens of degrees from the transition temperature. In a cholesteric liquid crystal with this arrangement of phases and a very short intrinsic pitch length, cooling extends the pitch so that visible light is reflected. Continued cooling results in successive reflection of each spectral color from violet to red, followed by IR light. This is the normal mode of operation of liquid crystal thermochromic devices used for temperature measurement and novelty applications. By varying the optical activity of the material the width of the color play band can be adjusted; the temperature at which thermochromism occurs can be varied by adding components with higher or lower S_A–cholesteric transition temperatures.

The most common procedure for the construction of practical thermochromic devices is to microencapsulate the liquid crystal in droplets with a diameter of ca. 10–20 µm, each having a protective coating of gelatin–gum arabic coacervate (→ Microencapsulation). This coating protects the liquid crystal from contamination by fats, plasticizers, etc., and minimizes its exposure to atmospheric oxygen which can affect stability. The capsules are mixed into a binder system that allows them to be printed onto paper or plastic sheets. Other methods include polymer dispersion and plastic lamination [64]. In devices, the brightness of reflection is inferior to that observed in thin, pressed films of liquid crystal due to imperfect alignment, but the dependence of color on viewing angle is somewhat less.

Liquid crystal phases were first recognized in cholesterol esters; the first thermochromic devices used such materials. Cholesterol derivatives have continued to provide a source of optically active liquid crystal phases. A difficulty in formulating practical mixtures of cholesterol derivatives, however, is their generally high melting point combined with their relative instability. Purification of the low-melting compounds such as cholesteryl oleyl carbonate is also difficult. Because cholesterol is only available in the naturally occurring optically active form, variation of the pitch of mixtures relies on the addition of derivatives such as cholesteryl halides with an opposite twist. Mesogens with the largely saturated cholesterol skeleton have a low birefringence, which means that the selective reflection band is narrow and the perceived brightness is poor. All of these problems have led to the development of nonsteroidal liquid crystals. Derivatives of branched alkyl chains are available in both the optically active and the racemic forms. They often have short pitch lengths, and the sign and magnitude of pitch can be correlated with molecular structure [64–66]. Variation of the mesogenic core can be used to vary the birefringence and phase range of the materials. Series of mixtures based on such compounds have been developed with a wider range of temperature than is available from cholesterol esters [64]. Steroid derivatives and nonsteroidal

Table 4. Cholesteric (**43–45**) and chiral (**c.46–48**) nematic liquid crystals

Structure	Name	CAS registry number	Empirical formula	Phase transition sequence, °C
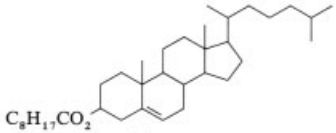 **43**	cholest-5-en-3-ol(3*β*), nonoic acid ester (cholesteryl nonanoate, CN)	[*1182-66-7*]	$C_{36}H_{62}$	K 80.5 S_A (77.5) Ch 92 I
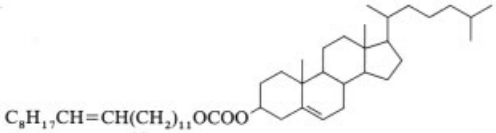 **44**	(*Z*)-cholest-5-en-3-ol-(3β)-9-octadecenyl carbonate (cholesteryl oleyl carbonate, COC)	[*17110-51-9*]	$C_{46}H_{80}O_3$	K 26 Ch 41 I
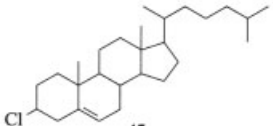 **45**	3-chlorocholest-5-ene(3β) (cholesteryl chloride, CCl)	[*910-31-6*]	$C_{27}H_{45}Cl$	K 97 Ch (62) I
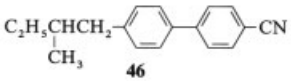 **46**	4′-(2-methylbutyl)biphenyl-4-carbonitrile (CB15)	[*63799-11-1*]	$C_{18}H_{19}N$	K 45 (− 54) Ch (− 30) I
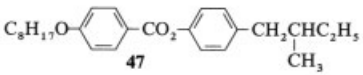 **47**	4-(octyloxy)benzoic acid, 4-(2-methylbutyl)phenyl ester (CE5)	[*69777-61-3*]	$C_{26}H_{36}O_3$	K 47.5 Ch (42) I
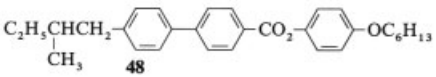 **48**	(+)-4′-(2-methylbutyl)biphenyl-4-carboxylic acid, 4-(hexyloxy)phenyl-ester (CE3)	[*62614-61-3*]	$C_{30}H_{36}O_3$	K 68.8 S_C 80.2 Ch 163.5 I

chiral nematics are now both exploited in devices; examples are listed in Table 4.

3.7. Blue Phases

In addition to the classic cholesteric phases described in Section 3.6, chiral nematic materials with a short pitch frequently pass through additional phases of short duration in the immediate vicinity of the clearing point [67]. Three separate blue phases have been identified, which are usually stable over an interval of no more than 1 – 2 °C. Blue phase I apparently consists of an array of twist disclination lines forming a body-centered cubic lattice. The phase arises from a local preference for a doubly twisted structure which cannot be sustained over long range in three dimensions for topological reasons. Blue phase II has a similar structure, but in this case a simple cubic lattice is formed. The structure of blue phase III, the "blue fog" phase, is unknown. Electro-optic effects have been demonstrated in the blue phases, but their narrow temperature ranges weigh heavily against their practical application.

3.8. Thermotropic Smectic Phases

Even in the least ordered smectic A (S_A) phase, the ordering of molecules into layers has fundamental consequences for the physical properties of the material. The only elastic distortion that is possible while maintaining a constant layer spacing is a bending of the layers (corresponding to a splay distortion of the director configuration); other distortions require the expenditure of energy akin to the compression of a liquid. This fact accounts for the divergence of k_{22} and k_{33} on approaching a smectic phase. Facile, reversible switching effects under an applied field, analogous to the Frederiks transition in nematic phases, are hardly observed in smectic materials. The smectic layers do not readily flow past one another. Flow in smectic phases is believed to occur by the movement of "plugs" of material through the bulk phase [5]. A very high apparent viscosity and strongly non-Newtonian behavior are observed in conventional viscometer geometries. The anisotropies of refractive indices, dielectric constants, and diamagnetic properties closely resemble those observed in nematic phases.

In view of its elastic behavior application of the S_A phase in electro-optic devices does not at first seem hopeful. Two different display technologies have nevertheless been demonstrated: thermally and electrically addressed devices.

Thermally Addressed Displays. Thermally addressed display devices utilize smectic material of positive dielectric anisotropy with a narrow nematic phase, enclosed in a display cell similar to that used for nematic liquid crystal displays [68–70]. In operation, an area of the cell is heated rapidly to a temperature above the clearing point of the liquid crystal. On cooling, the liquid crystal forms the natural microscopic structure of the smectic A phase which is the focal conic state. This texture contains a multiplicity of small domains in which director orientation varies rapidly; it therefore scatters light strongly. If an electric field is applied to the cell during heating and cooling, the liquid crystal aligns with the field during its passage through the nematic phase. Further cooling results in the formation of a uniform homeotropic smectic phase, and light scattering is not observed. Because of the high viscosity of the smectic phase, each texture remains indefinitely stable and information written on the display is stored without any refreshing. Application of the addressing voltage has no effect on areas of the display that are not heated. Separation of addressing into electrical and thermal processes circumvents the restrictions on display complexity imposed by RMS (root mean square) multiplexing. The information that can be written on the display is limited only by the resolution with which heat can be applied and the rate at which updating of the display is required.

Two variations of the thermally addressed smectic A display have been exploited. In the first, heating is performed by a laser focused onto the liquid crystal layer, the absorption of light usually being enhanced by inclusion of a suitable dye [71]. The second uses ohmic heating of rows of the display by means of stripe electrodes on one cell plate. On the other cell plate, column electrodes allow a voltage to be applied across the pixels which are to be driven into the clear state. Rows of the display are addressed in turn, and a small area device capable of video

rate addressing has been developed [70]. These devices have often been used in projection systems which take advantage of the high resolution to provide large, high-density data displays [68].

Electrically Addressed S_A Displays. An electrically addressed smectic A display has been demonstrated which also provides a high-density information storage device [72], [73]. The liquid crystal is doped to a resistivity of ca. $10^9\,\Omega\cdot$cm and is switched from a clear to a scattering state by dynamic scattering, which can be induced in S_A phases even if they have a high positive dielectric anisotropy. The process is reversed under dielectric forces, loosely analogous to a Frederiks transition. Discrimination between these two opposed switching processes is achieved by changing the addressing frequency, with scattering occurring at low frequencies. Such a device does not show an RMS response to applied voltage, and the characteristics of the display allow a very complex format to be achieved. The principal shortcoming is the high addressing voltages required (typically 50 – 80 V), which make this technology incompatible with the drive circuitry normally used for liquid crystal devices.

The S_C phase has the same degree of ordering as the S_A phase, but in the C phase the molecules in the layers are tilted and the phase is consequently biaxial. The macroscopic appearance of the S_C phase is very similar to that of a S_A phase, and the transition between these phases is frequently a very low-enthalpy process. The tilt angle in the C phase undergoes a pretransitional decrease as the transition temperature is approached. Systems in which S_C–N or S_C–I transitions occur, show a more abrupt collapse of tilt at the transition point. No significant applications presently exist for the common S_C phase, but the prediction and discovery of ferroelectric properties in the chiral C phase [74] and the demonstration of fast electro-optic devices exploiting the effect [75] have lead to extensive research.

Ferroelectric Displays. If an S_A phase is doped with a chiral additive or if its constituent molecules are optically active, a twisted structure does not result. In an S_C phase, a twisted structure results from helical ordering of the direction of molecular tilt in adjacent layers, and short-pitch S_C phases show selective reflection of light similar to that in cholesteric materials (see Section 3.6) [76]. The ferroelectric property of a chiral S_C phase, however, relies on the unwinding of this helix by surface forces or an applied field. Coupling between the chiral center of an S_C molecule and its environment then results in biaxial ordering, with a preferred orientation of transverse molecular dipole moments associated with COO−, C−Cl, C−CN groups etc. in a direction parallel to the layer and perpendicular to the tilt direction (Fig. 14). The resulting net dipole moment can be equated with the spontaneous polarization (P_s) of the ferroelectric phase. A chiral S_C phase can provide an electro-optic effect by aligning the material in the bookshelf geometry of Figure 15. The smectic layers and the spontaneous polarization vector are both perpendicular to the plane of the cell, and the molecules are constrained in the plane of the cell by surface forces. Two degenerate configurations then exist, with the molecular tilt to the left or the right of the layer normal; the P_s vector correspondingly points up or down from the plane of the cell. Switching from one configuration to the other can be accomplished by means of a short d.c. pulse, after which the states each have a limited bistability [75]. Each state remains stable over a much longer period than the switching time (i.e., for tens of milliseconds) in the absence of applied voltage pulses.

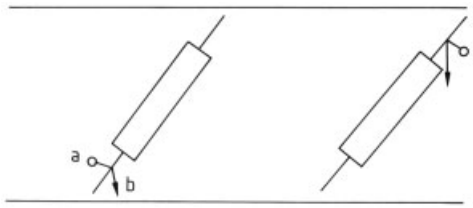

Figure 14. Microscopic ordering of chiral smectic C molecules
The combination of molecular tilt and steric influence (a) at the chiral center impose a biaxiality that orients the transverse dipole (b) in a unique direction without affecting the head – tail degeneracy of liquid crystal ordering.

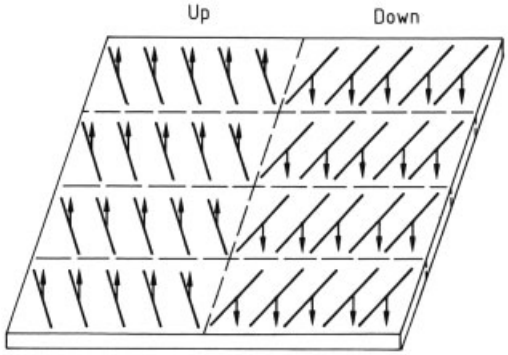

Figure 15. "Up" and "down" domains of ferroelectric liquid crystal in the bookshelf geometry

When a surface-stabilized ferroelectric device is switched, it behaves as a uniaxial birefringent crystal in which the direction of the optic axis is rotated through twice the S_C tilt angle in the plane of the cell. This switching may be converted to an electro-optic effect by laminating crossed polarizers onto the cell, with one polarizing axis aligned parallel to the liquid crystal alignment direction in one switched state. This state appears dark, and contrast is obtained against the other switched state where light is transmitted due to optical interference arising in the same manner as in the "on" state of the electrically controlled birefringence effect (Section 3.5.2). The optical contrast of the ferroelectric cell is maximized when the optic axis rotates through 45°, and fortuitously this is achieved by a S_C tilt angle (22.5°) similar to that observed in many materials.

Interest in the switching of chiral S_C phases has been especially great because it can be achieved on a timescale of tens of microseconds, as compared to the tens of milliseconds required for the switching of common nematic displays. The device does not respond to RMS fields, but to a voltage – time product in the addressing pulses, and the threshold behavior and limited bistability of the cell allow multiplex drive at unlimited levels using specially designed pulse sequences. Complex displays have been demonstrated which use this effect [77–82]. Because the chiral S_C cell switches between two stable states, the intermediate liquid crystal configurations cannot be sustained; the cell switches from black to white without any intermediate grey shades being available. This obvious disadvantage for graphic and video displays may be overcome by spatial or temporal division of the addressing of each pixel [82], [83].

Several remaining difficulties distinguish this technology from other liquid crystal devices. The first is the necessity for construction of a very thin (< ca. 2 μm) device to stabilize the required bookshelf geometry of the liquid crystal. Because the device operates in a birefringent mode, a high degree of uniformity is required, and fabrication of large displays to such a tolerance presents considerable problems. Alignment of smectic phases is also difficult, and the properties of higher temperature phases are usually exploited to facilitate this [84], [85]. Mixtures can be formulated with the phase sequence $S_C - S_A - Ch - I$, in which the cholesteric phase is of very long pitch, but the spontaneous polarization of the S_C phase is large. Cooling such a mixture from the isotropic phase in a thin cell leads successively to a planar aligned nematic phase, a planar S_A phase, and an aligned S_C phase.

Bookshelf geometry is a very idealized view of the true state of alignment in S_C devices. Real cells contain defects [86], and a chevron structure appears to exist in the liquid crystal layer. Elucidation of the true structures in these devices [87] and development of addressing techniques [88], [89], cell fabrication methods, and improved materials [71], [90–95], are major themes in liquid crystal technology.

A range of materials suitable for chiral smectic C devices is listed in Table 5. The first liquid crystal materials in which ferroelectric switching was demonstrated were the unstable, high melting point compounds decyloxybenzylideneamino 2-methylbutyl cinnamate (**49**, DOBAMBC) and hexyloxybenzylideneaminochloropropyl cinnamate (**50**, HOBACPC), but room temperature materials have since been developed. To allow molecular engineering of the different physical properties required, mixtures are usually formulated from a nonchiral S_C base mixture containing one or more dissolved chiral additives (e.g., **52**, **53** in Table 5) to impart spontaneous polarization. The magnitude of P_s depends on the strength of the lateral dipole in the dopant and on the efficiency of its coupling to the chiral center: values of ca. 100 – 300 nC/cm^2 are attainable. Materials with a high P_s and long helical pitch are known, and pairs of dopants exist in which the sign of P_s is

Table 5. Compounds used in chiral S_C devices

Structure	Name	CAS registry number	Empirical formula	Use
$C_{10}H_{21}O$–C$_6$H$_4$–CH=N–C$_6$H$_4$–CH=CHCO$_2$CH$_2$CH(CH$_3$)C$_2$H$_5$ **49**	3-[4-(4-decyloxyphenylmethylene-amino)phenyl]-2-propenoic acid, 2-methylbutyl ester (decyloxybenzylidene-amino 2-methylbutyl cinnamate, DOBAMBC)	[*55593-53-8*]	$C_{31}H_{43}NO_3$	ferroelectric S_C compound
$C_6H_{13}O$–C$_6$H$_4$–CH=N–C$_6$H$_4$–CH=CHCO$_2$CH$_2$CH(Cl)CH$_3$ **50**	3-[4-(4-hexyloxyphenylmethylene-amino)phenyl]-2-propenoic acid, 2-chloro-propyl ester (hexyloxybenzylideneamino chloropropyl cinnamate, HOBACPC)	[*60646-92-6*]	$C_{25}H_{30}ClNO_3$	ferroelectric S_C compound
CH$_3$CH(CH$_3$)–CH(Cl)–CO$_2$–C$_6$H$_4$–C$_6$H$_4$–OC$_8$H$_{17}$ **51**	(*S*)-2-chloro-3-methylbutanoic acid, 4′-(octyloxy)biphenyl-4-yl ester	[*100497-98-1*]	$C_{25}H_{33}ClO_3$	ferroelectric S_C compound
$C_8H_{17}O$–C$_6$H$_4$–C$_6$H$_4$–CO$_2$CH(CH$_3$)C$_6$H$_{13}$ **52**	4′-(octyloxy)biphenyl-4-carboxylic acid, 1-methylheptyl ester	[*62724-78-1*]	$C_{29}H_{42}O_3$	High P_s isotropic dopant
$C_8H_{17}O$–C$_6$H$_4$–C$_6$H$_4$–CO$_2$CH(CN)CH$_3$ **53**	4′-(octyloxy)biphenyl-4-carboxylic acid, 1-cyanoethyl ester		$C_{24}H_{29}NO_3$	high P_s isotropic dopant
$C_8H_{17}O$–C$_6$H$_4$–C$_6$H$_4$–CO$_2$–C$_6$H$_3$(F)–C$_5$H$_{11}$ **54**	4′-(octyloxy)biphenyl-4-carboxylic acid, 2-fluoro-4-pentylphenyl ester (MB805F)		$C_{32}H_{39}FO_3$	smectic C host
$C_{10}H_{21}O$–C$_6$H$_4$–CO$_2$–C$_6$H$_4$–OC$_6$H$_{13}$ **55**	4-(decyloxy)benzoic acid, 4-(hexyloxy)-phenyl ester	[*68162-09-4*]	$C_{29}H_{42}O_4$	smectic C host
$C_{10}H_{21}$–C$_4$H$_2$N$_2$–C$_6$H$_4$–OC$_8$H$_{17}$ **56**	5-decyl-2-(4-octyloxyphenyl)pyrimidine (PYP8010)	[*57202-62-7*]	$C_{28}H_{44}N_2O$	smectic C host

the same but the sense of helical twist is opposite.

A chiral, laterally dipolar S_A phase can show a form of induced ferroelectric behavior [96]. The ordinary S_A phase is uniaxial even when it is composed of chiral molecules, but application of an electric field to a planar aligned chiral S_A sample induces biaxial ordering of the lateral dipoles. This is coupled to the induction of a molecular tilt relative to the smectic layers, so that the optic axis of the phase rotates under the influence of the applied field. The tilt is proportional to the applied field; no threshold behavior or bistability occurs, so that the effect is not suitable for multiplex drive, but switching takes place on a comparable timescale (microseconds) to S_C switching. Induced tilt angles of $> 10°$ can be obtained close to an S_C–S_A transition temperature, but this value decreases at higher temperature. Potential applications of this device include fast shutters and modulators.

The remaining liquid crystal phases are relatively poorly explored with regard to their potential electro-optic applications. All chiral tilted smectic phases show switching behavior analogous to the chiral S_C phase. The bistability in these phases appears to be stronger and the switching considerably slower than in the S_C phase.

4. Thermotropic Discotic Liquid Crystals

The synthesis of thermotropic liquid crystals based on flat disclike (discotic) molecules was accomplished in 1977 [97]. Discotic nematic phases had previously been observed in mixtures obtained by pyrolysis of petroleum residues, but the structure was not clear. Discotic mesogens have been synthesized from a variety of core structures with four, six, or eight pendant chains (Table 6).

Discotic liquid crystals exhibit a polymorphism analogous to calamitic phases [9], [98]. Columnar phases contain stacks of discotic molecules; intermolecular distances within the stacks can be ordered or disordered, and the columns may be packed together in a hexagonal or a tetragonal array. The direction of molecular tilt within columns distinguishes different columnar tetragonal phases.

The discotic nematic phase is analogous to the corresponding calamitic phase, possessing only orientational ordering. The director is again defined along the unique (short) axis of a uniaxial phase. Because of the molecular form, a negative birefringence and dielectric anisotropy are more common in discotic phases and reentrant behavior also seems rather common. Discotic liquid crystals do not yet have practical applications, and all the mesophases are formed well above room temperature.

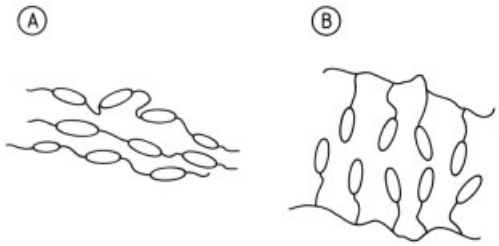

Figure 16. Illustrative structures of main-chain (A) and side-chain (B) liquid crystal polymers

5. Polymer Liquid Crystals

Polymers form both thermotropic and lyotropic liquid crystal phases, and the molecular architecture is subject to great variation [99]. Main-chain liquid crystal polymers incorporate mesogenic groups as part of the backbone, either alternating with flexible linking groups or forming a rigid rod polymer (see Fig. 16). In the latter, thermotropic liquid crystal phases are common only in copolymers whose structural irregularities serve to lower the melting point. Structural irregularities are introduced by copolymerization of bulky, flexible, or nonlinear molecules in a polymer composed largely of linear, rigid units. A typical example is the copolymerization of 2-hydroxynaphthalene-6-carboxylic acid with 4-hydroxybenzoic acid. Rigid rod homopolymers and regular copolymers may form lyotropic phases if a suitable solvent is available [100]. Many cellulose derivatives [101], [102], DNA fragments, and synthetic peptides [103], [104] also form lyotropic cholesteric phases in polar and semipolar solvents. The behavior of these lyotropic rod liquid crystals is qualitatively described by Flory's equation which, for rods with an axial ratio $x \gtrsim 10$, predicts liquid crystal

Table 6. Thermotropic discotic mesogens

Structure and phase transition sequence, °C	Name	Empirical formula
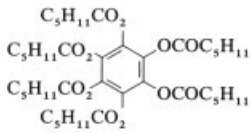 **57** K 68 D_{rd} 86 I	Hexanoic acid, benzenehexayl ester (benzene hexahexanoate) [*65201-69-6*]	$C_{42}H_{66}O_{12}$
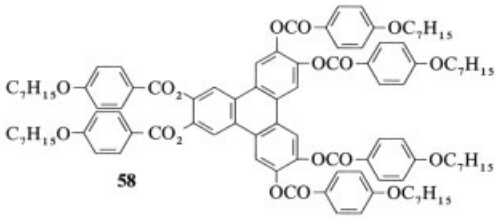 **58** K 168 N_D 253 I	4-heptyloxybenzoic acid, 2,3,6,7,10,11-triphenylenehexayl ester [*75747-38-5*]	$C_{102}H_{120}O_{18}$
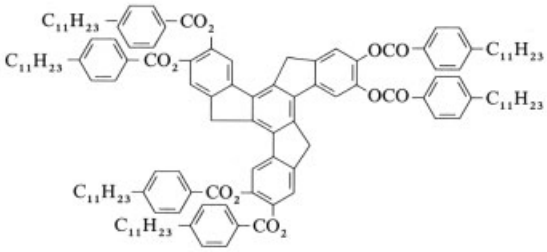 **59** K 90 D_{rd} 137 N_D 171 D_{rd} 274 N_D 297 I	4-undecylbenzoic acid, 2,3,7,8,12,13-[5,10,15*H*-diindeno-[1,2-a:1′,2′-c]fluorene] hexayl ester	$C_{135}H_{171}O_{12}$

phase formation above a critical volume concentration V_c, where

$$V_c = \frac{8}{x}\left(\frac{1-2}{x}\right)$$

In systems where flexible groups are interspersed with rigid ones, the phase behavior and stability are qualitatively comparable with those of the isolated mesogenic moieties. The technological importance of main-chain polymers arises from the ease with which alignment of the polymer backbone can be achieved under shear forces. Alignment of these materials in a magnetic or electric field is very difficult, but the flow alignment that results from molding or spinning the polymer in the liquid crystal phase is preserved in the solid polymer and imparts exceptional anisotropic mechanical properties. These phenomena were first exploited in lyotropic liquid crystal polymer systems [105], [106] such as the polyamide **60**.

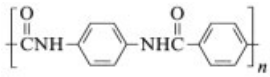

60

In solution in concentrated sulfuric acid, this polymer forms a nematic phase which on spinning yields fibers of very high tensile modulus and stiffness.

Thermotropic polymer liquid crystal systems have been extensively investigated [107] to avoid the inherent difficulties of lyotropic systems, which usually require aggressive solvents for their formation. Examples of thermotropic liquid crystal polymers follow:

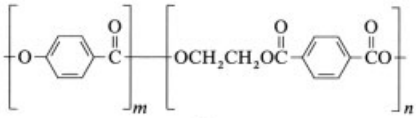

61

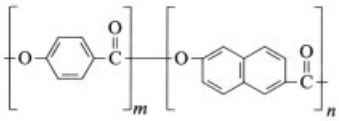

62

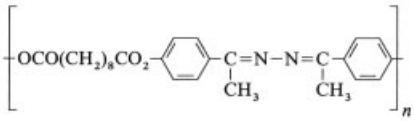

63

They show additional advantages for molding [108] because flow alignment can result in improved mold filling and self reinforcement of the solid. Moldings exhibit good dimensional tolerance and small coefficients of thermal expansion.

The alternative fundamental architecture available for liquid crystal polymer systems is the side-chain or comb polymer (e.g., **64**, **65**), in which a conventional flexible polymer chain is replaced by mesogenic groups via a flexible link-ing unit [99]:

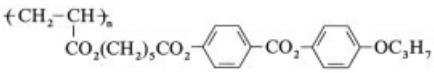

64

S_G 60 S_C 95 N 116 I

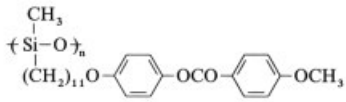

65

K 56 S_C 60 S_A 134 I

This concept has been extensively explored in polyacrylate and siloxane systems [109], and the resulting polymers show liquid crystalline phase behavior which frequently parallels that of the monomer. In polymers having pendant groups with strong terminal dipolar groups, however, smectic phase stability is enhanced, presumably due to cooperation among the polar association of mesogenic groups, the tendency of spacing groups to maintain the liquid crystal moities in parallel layers, and the effect of the liquid crystal groups on the flexibility of the polymer backbone.

Liquid crystal side-chain polymers show behavior similar to that exhibited by low molecular mass systems, but modified by the high viscosity typical of polymers. Polymers are now available which possess all of the liquid crystal phases exploited in low molecular mass systems in electrooptic devices. Alignment of the mesogenic groups can be achieved under electric or magnetic fields, but switching is normally very slow. Technological interest in side-chain polymer liquid crystals centers on systems that have benefits which are lacking in low molecular mass systems. In particular, smectic liquid crystal polymers offer the possibility of providing media for optical data storage with thermal and electrical addressing [110]. Use of a polymer provides a much more robust medium than low molecular mass materials. In an alternative approach a nematic polymer may be used, with data storage obtained when the polymer is cooled below the glass transition temperature, thereby freezing the order and alignment into place indefinitely. A similar technique can be applied to the production of specialized optical components [111]. Liquid crystal chiral smectic polymers may have potential as pyroelectric or piezoelectric elements, and polymer systems providing exotic materials such as liquid crystal elastomers may stimulate their own novel applications [112].

6. Amphiphilic Liquid Crystal Phases

A variety of amphiphilic surfactants associate with solvents to form liquid crystal phases in which the mesogenic units are aggregates of molecules [113]. At least four well-defined

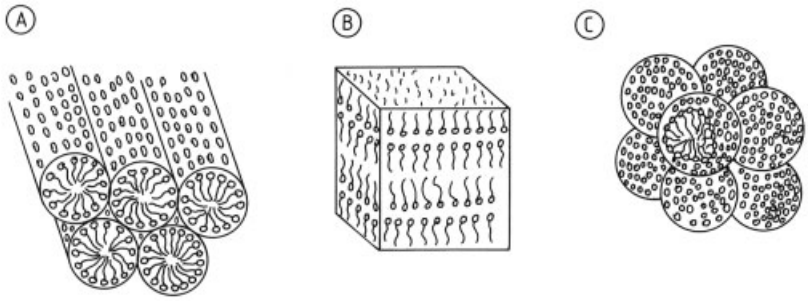

Figure 17. Illustrative structures of amphiphilic hexagonal (A), lamellar (B), and cubic (C) phases
The structures shown here are not the only ones in each class of phase.

classes of mesophase exist in these systems (nematic, hexagonal, lamellar, and cubic, Fig. 17), with the possibility of different phases in each class.

At concentrations and temperatures above a critical point, surfactants in solution aggregate in a head-to-head, tail-to-tail fashion to form either infinite two-dimensional surfaces or curved closed structures (*micelles*). In a polar medium, the hydrophilic heads of the surfactant associate with the solvent, whereas the hydrophobic tails interact to form the interior of the micelles. In nonpolar solvents, the situation can be reversed, the micelles having their hydrophobic groups on the outside (reversed-phase structures). Micelles can be spherical, rod-like, or discotic in shape; the shape and structure of any particular surfactant are determined by the relative sizes of the hydrophilic and hydrophobic groups. The shape of the micelle is also affected by solvent, temperature, and concentration.

Lyotropic nematic phases are formed by polydisperse rod- or disk-shaped micelles, usually only within a very restricted area of a phase diagram. Mixed solvent systems frequently favor nematic phase formation. Like thermotropic nematic phases, these materials can be aligned in a magnetic field. The conductivity of most systems precludes dielectric effects. The birefringence of the phases is very low, and no practical applications have been devised.

Rodlike micelles of essentially infinite length pack into hexagonal structures. Normal and reversed phases of this type are known which form viscous anisotropic liquids. Micelles also form isotropic solutions in which a variety of effects such as flow-induced birefringence and viscoelastic behavior may be observed [114]. In contrast, the limiting case of large discotic micelles implies the formation of infinite layers or *lamellae*. These layers form liquid crystal phases in which lamellar bilayers are separated by aqueous layers. The phases are viscous and birefringent.

The fourth category of mesophase formed in amphiphilic systems relates to *cubic phases*, which may form in two ways. The first is by the close packing of spherical normal or reverse micelles. Such micelles show a positive curvature of the surfactant bilayer, but when a net negative surface curvature exists, a cubic structure can form in which both the aqueous phase and the surfactant form infinite continuous structures. Being cubic in nature these phases are not birefringent.

Although lyotropic phases have not been given the same prominence as thermotropic phases, their importance should not be underestimated. The phases are crucial in the manufacture and mode of operation of soaps and detergents, and have an important role in areas such as food preparation. Amphiphilic bilayers also form model systems for cell membranes [115], [116]. Finally, the possibility remains that lyotropic liquid crystals may be exploited in electro-optic devices in the future, and in this respect the formation of mesophases including nematic phases in dye systems appears interesting [117].

7. References

General References

Research in liquid crystal science and technology is reported in the journals *Liquid*

Crystals, Molecular Crystals and Liquid Crystals, and *Proceedings of the Society for Information Display.*

1. G. W. Gray (ed.): *Thermotropic Liquid Crystals, Critical Reports on Applied Chemistry 22,* J. Wiley and Sons, Chichester 1987.
2. L. M. Blinov: *Electro-Optical and Magneto-Optical Properties of Liquid Crystals,* J. Wiley and Sons, Chichester 1983.
3. E. Kaneko: *Liquid Crystal TV Displays: Principles and Applications of Liquid Crystal Displays,* Advances in Optoelectronics, KTK Scientific Publishers, Tokyo 1987.
4. W. H. de Jeu: *Physical Properties of Liquid Crystalline Materials,* Liquid Crystal Monographs 1, Gordon and Breach, London 1980.
5. P. G. de Gennes: *The Physics of Liquid Crystals,* Clarendon Press, Oxford 1974.
6. G. Vertogen, W. H. de Jeu: "Thermotropic Liquid Crystals, Fundamentals," *Springer Series in Chemical Physics,* Springer Verlag, Berlin 1988.
7. D. Demus, H. Demus, H. Zaschke: *Flüssige Kristalle in Tabellen,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1974; vol. 2, 1984. C. B. McArdle (ed.): *Side Chain Liquid Crystal Polymers,* Blackie, Glasgow 1989.

Specific References

8. F. Reinitzer, *Monatsh. Chem.* **9** (1988) 421.
9. L. J. Yu, A. Saupe, *Phys. Rev. Lett.* **45** (1980) 1000.
10. J. Malthete, L. Liebert, A. M. Levelut, Y. Galerne, *C. R. Acad. Sci. Paris* **t 303** (1986) 1073.
11. S. Chanrasekhar, B. K. Sadashiva, B. R. Ratha, V. N. Raja, *Pramana* **30** (1988) L 491.
12. A. J. Leadbetter, Structural Classification of Liquid Crystals, in [1]
13. G. W. Gray, J. W. Goodby: *Smectic Liquid Crystals,* Leonard Hill, Glasgow 1984.
14. D. Demus, S. Diele, S. Grande, H. Sackmann, *Adv. Liq. Cryst.* **6** (1983) 1.
15. G. W. Gray, *Mol. Cryst. Liq. Cryst.* **63** (1981) 3.
16. K. J. Toyne, "Liquid Crystalline Behaviour in Relation to Molecular Structure," in [1]
17. P. Balkwill, D. Bishop, A. Pearson, I. Sage, *Mol. Cryst. Liq. Cryst.* **123** (1985) 1.
18. W. Maier, A. Saupe, *Z. Naturforsch. A* **14 A** (1959) 882.
19. W. Maier, A. Saupe, *Z. Naturforsch. A* **15 A** (1960) 287.
20. G. R. Luckhurst, C. Zannoni, *Nature (London)* **267** (1977) 412.
21. M. A. Cotter, *Mol. Cryst. Liq. Cryst.* **97** (1983) 29. A. V. Komolkin, Y. V. Molchanov, P. P. Yakutseni, *Liq. Cryst.* **6** (1989) 39.
22. M. G. Clark, *Mol. Cryst. Liq. Cryst.* **127** (1985) 1.
23. I. Haller, *Prog. Solid State Chem.* **10** (1975) 103.
24. I. Sage, D. Chaplin, *Electron. Lett.* **23** (1987) 1192.
25. W. Maier, G. Meier, *Z. Naturforsch. A* **16 A** (1961) 262.
26. W. H. de Jeu, *Philos. Trans. R. Soc. London Ser. A* **309** (1983) 217.
27. K. Toriyama, D. Dunmur, *Mol. Phys.* **56** (1985) 479.
28. M. Schadt et al., *Mol. Cryst. Liq. Cryst.* **94** (1983) 139.
29. S. Chandrasekhar, B. K. Sadashiva, B. S. Srikanta, *Mol. Cryst. Liq. Cryst.* **151** (1987) 93.
30. M. Bradshaw, D. G. McDonnell, E. P. Raynes, *Mol. Cryst. Liq. Cryst.* **70** (1981) 289.
31. M. Bradshaw, E. P. Raynes, *Mol. Cryst. Liq. Cryst.* **91** (1983) 145.
32. M. Bradshaw, E. P. Raynes, *Mol. Cryst. Liq. Cryst.* **72** (1981) 35.
33. H. Schad, M. Osman, *J. Chem. Phys.* **75** (1981) 880.
34. M. Schadt, P. Gerber, *Z. Naturforsch. A* **37 A** (1982) 165.
35. H. Schad, M. Osman, *J. Chem. Phys.* **79** (1983) 5710.
36. B. Scheuble, G. Baur, G. Meir, *Mol. Cryst. Liq. Cryst.* **68** (1981) 57.
37. M. Bradshaw, E. P. Raynes, *Mol. Cryst. Liq. Cryst.* **99** (1983) 107.
38. M. Bradshaw, E. P. Raynes, *J. Phys. (Orsay Fr.)* **45** (1984) 157.
39. M. Schadt, F. Muller, *Inst. Elect. Electron. Engrs., Trans Electron Devices* **25** (1978) 1125.
40. J. Cognard, *Mol. Cryst. Liq. Cryst.* **78** (1981) Suppl. 1.
41. J. Castellano, *Mol. Cryst. Liq. Cryst.* **94** (1983) 33.
42. I. Sage, "Materials Requirements for Nematic and Chiral Nematic Electro-optical Display" in [1]
43. I. C. Khoo, Y. R. Shen, *Opt. Eng.* **24** (1985) 579.
44. A. D. Lloyd, I. Janossy, H. A. MacKenzie, B. S. Wherrett, *Opt. Commun.* **61** (1987) 351.

45. V. Fredericks, V. Zolina, *Trans. Faraday Soc.* **29** (1933) 919.
46. M. Shadt, W. Helfrich, *Appl. Phys. Lett.* **18** (1971) 127.
47. B. Bahadur, *Mol. Cryst. Liq. Cryst.* **109** (1984) 1.
48. C. Mauguin, *Bull. Soc. Fr. Mineral. Crystallogr.* **34** (1911) 71.
49. C. Gooch, H. Tarry, *J. Phys. D.* **8** (1975) 1575.
50. P. Alt, P. Pleshko, *Inst. Elect. Electron. Engrs. Trans. Electron Devices* **21** (1974) 146.
51. K. Toriyama et al. in G. Sprokel (ed.): *The Physics and Chemistry of Liquid Crystal Devices,* Plenum Publishing, New York 1980.
52. B. Needham, *Philos. Trans. R. Soc. London A* **309** (1983) 179.
53. C. Waters, E. P. Raynes, V. Brimmel, *Mol. Cryst. Liq. Cryst.* **123** (1985) 303.
54. T. Scheffer, J. Nehring, *Appl. Phys. Lett.* **45** (1984) 1021.
55. R. Thurston, *Mol. Cryst. Liq. Cryst.* **122** (1985) 1.
56. K. Katoh et al., *Jpn. J. Appl. Phys.* **26** (1987) L 1784.
57. M. Schadt, F. Leenhouts, *Appl. Phys. Lett.* **50** (1987) 236.
58. A. V. Ivanshchenko, V. G. Rumyantsev, *Mol. Cryst. Liq. Cryst.* **150 A** (1987) 1.
59. G. Heilmeier, J. Castellano, L. Zanoni, *Mol. Cryst. Liq. Cryst.* **8** (1969) 293.
60. D. White, G. Taylor, *J. Appl. Phys.* **45** (1974) 4718.
61. G. Heilmeier, L. Zanoni, L. Barton, *Appl. Phys. Lett.* **13** (1968) 46.
62. J. Ferguson, *Proc. S.I.D. Conf.,* Orlando 1985.
63. D. W. Berreman, T. J. Scheffer, *Phys. Rev. Lett.* **25** (1970) 577.
64. D. G. McDonnell, Thermochromic Cholesteric Liquid Crystals in [1]
65. G. W. Gray, D. G. McDonnell, *Mol. Cryst. Liq. Cryst.* **34** (1977) 211.
66. J. W. Goodby, *Science* **231** (1986) 350.
67. H. Stegemeyer et al., *Wiss. Beitr. Martin Luther Univ. Halle Wittenberg* **52** (1986) 64.
68. F. J. Kahn, *Appl. Phys. Lett.* **22** (1973) 111.
69. A. G. Dewey, *IEEE Trans. Electron Devices* **24** (1977) 918. S. Le Berre, M. Hareng, R. Hehlen, J. N. Perbet, *Displays* (1984) 349.
70. M. Hareng et al., *Proc. Electron Displays Conf.* **1** (1984) 28.
71. D. Coates, Materials Requirements for Smectic Liquid Crystal Displays in [1]
72. D. Coates, W. A. Crossland, J. H. Morrissy, B. Needham, *J. Phys. D* **11** (1978) 2025.
73. W. A. Crossland, S. Cantor, *SID Symp. Digest* (1985) 124.
74. R. B. Meyer, L. Liebert, L. Strzelecki, P. Keller, *J. Phys. Lett. (Orsay Fr.)* **36** (1975) 69.
75. N. A. Clark, S. T. Lagerwall, *Appl. Phys. Lett.* **36** (1980) 899.
76. G. W. Gray, D. G. McDonnell, *Mol. Cryst. Liq. Cryst.* **37** (1986) 189.
77. M. F. Bone et al., *Displays* (1987) 115.
78. T. Harada, M. Taguchi, K. Iwasa, M. Kai, *SID Symp. Dig.* (1985) 131.
79. M. Geary, *SID Symp. Dig.* (1985) 128.
80. W. J. Hartmann, *Proc. Int. Disp. Res. Conf.* San Diego 1988, p. 191.
81. P. W. Ross, *Proc. Int. Disp. Res. Conf.* San Diego 1988, p. 185.
82. T. Leroux et al., *Proc. Int. Disp. Res. Conf.* San Diego 1988, p. 111.
83. S. Lagerwall, J. Wahl, N. A. Clark, *Proc. Int. Disp. Res. Conf.* San Diego 1985, p. 213.
84. J. S. Patel, T. M. Leslie, J. W. Goodby, *Ferroelectrics* **61** (1984) 137.
85. M. J. Bradshaw, V. Brimmel, E. P. Raynes, *Japan Display Conf.* Tokyo 1986.
86. Y. Ouchi, H. Takazoe, A. Fukuda, *Japn. J. Appl. Phys.* **26** (1987) 1.
87. N. Hiji, Y. Ouchi, H. Takazoe, A. Fukuda, *Japn. J. Appl. Phys.* **27** (1988) L 1.
88. J. Wahl, T. Matuszczyk, S. Lagerwall, *Mol. Cryst. Liq. Cryst.* **146** (1987) 143.
89. J. R. Hughes, F. C. Saunders, *Liquid Crystals* **3** (1988) 1401.
90. P. Keller, *Mol. Cryst. Liq. Cryst.* **102** (1984) 295.
91. J. W. Goodby, T. M. Leslie, *Mol. Cryst. Liq. Cryst.* **110** (1984) 175.
92. C. Bahr, G. Heppke, *Mol. Cryst. Liq. Cryst. Letters Sect.* **4** (1986) 31.
93. K. Terashima et al., *Mol. Cryst. Liq. Cryst.* **141** (1986) 237.
94. M. F. Bone et al., *Mol. Cryst. Liq. Cryst.* **164** (1988) 117.
95. M. J. Bradshaw et al., *Proc. Soc. Inf. Disp.* **29** (1988) 3.
96. G. Anderson et al., *Appl. Phys. Lett.* **51** (1987) 640.
97. S. Chandrasekhar, B. K. Sadashiva, K. A. Suresh, *Pramana* **9** (1977) 471.
98. C. Destrade et al., *Mol. Cryst. Liq. Cryst.* **106** (1984) 121.
99. H. Finkelmann, *Angew. Chem. Int. Ed. Engl.* **26** (1987) 816.
100. Y. Onogi, J. L. White, J. F. Fellers, *J. Polym. Sci. Polym. Phys. Ed.* **18** (1980) 663.
101. R. S. Werbowyj, D. G. Gray, *Mol. Cryst. Liq. Cryst.* **34** (1976) 97.

102. S. L. Tseng, A. Valente, D. G. Gray, *Macromolecules* **14** (1981) 715.
103. C. Robinson, *Tetrahedron* **13** (1961) 219.
104. C. Robinson, *Mol. Cryst.* **1** (1966) 467.
105. S. L. Kwolek, P. W. Morgan, J. R. Schaefgen, L. W. Gulrich, *Macromolecules* **10** (1977) 1390.
106. T. I. Bair, P. W. Morgan, F. L. Killian, *Macromolecules* **10** (1977) 1396.
107. A. Ciferri: "Rigid and Semirigid Chain Polymeric Mesogens" in A. Ciferri, W. R. Krigbaum, R. B. Meyer (eds.): *Polymer Liquid Crystals,* Academic Press, New York 1980.
108. M. K. Cox, *Mol. Cryst. Liq. Cryst.* **153** (1987) 415.
109. H. Finkelmann, "Liquid Crystal Polymers" in [1]
110. H. J. Coles, R. Simon: "Electro-Optic Effects in a Smectogenic Polysiloxane Side-Chain Liquid Crystal Polymer" in L. L. Chapoy (ed.): *Recent Advances in Liquid Crystal Polymers,* Elsevier, London 1985.
111. V. P. Shibaev, N. A. Plate, *Adv. Polym. Sci.* **60/61** (1984) 173.
112. H. Finkelmann, G. Rehage, *Adv. Polym. Sci.* **60/61** (1984) 99.
113. G. J. T. Tiddy, M. F. Walsh, *Stud. Phys. Theor. Chem.* **26** (1983) 151.
114. H. Hoffmann, G. Ebert, *Angew. Chem. Int. Ed. Engl.* **27** (1988) 902.
115. A. S. C. Lawrence, *Mol. Cryst. Liq. Cryst.* **7** (1969) 1.
116. D. M. Small (ed.): *The Physical Chemistry of Lipids-Handbook of Lipid Research,* vol. 4, Plenum Publishing, New York 1986.
117. R. A. Berg, B. A. Haxby, *Mol. Cryst. Liq. Cryst.* **12** (1970) 93.

Lithium and Lithium Compounds

Ulrich Wietelmann, Chemetall GmbH, Frankfurt, Federal Republic of Germany

Richard J. Bauer, Chemetall GmbH, Frankfurt, Federal Republic of Germany

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 3
3. Occurrence 4
3.1. Important Lithium Minerals 4
3.1.1. Lithium Aluminum Silicates 4
3.1.2. Micas 5
3.1.3. Lithium Phosphates 5
3.1.4. Other Lithium Ores 6
3.2. Reserves of Lithium Minerals 6
3.3. Lithium in Natural Brines 6
4. Production of Primary Lithium Compounds 7
4.1. Mining of Ore and Production of Concentrate 7
4.2. Ore Digestion and Production of Lithium Compounds 8
4.2.1. Acid Digestion 8
4.2.2. Alkali Digestion 10
4.2.3. Ion-Exchange Processes 10
4.3. Production of Lithium Carbonate from Brines 12
5. Lithium Metal and Lithium Alloys 12
5.1. Production of Lithium Metal 12
5.2. Uses of Lithium Metal 13
5.3. Lithium Alloys 15
6. Lithium Compounds 15
6.1. Inorganic Lithium Compounds and Lithium Salts 15
6.2. Organolithium Compounds, Lithium Alkylamides, and Lithium Alkoxides 20
7. Quality Specifications and Analysis 24
8. Toxicology and Occupational Health 25
9. Economic Aspects 25
10. References 25

1. Introduction

Lithium [*7439-93-2*] (from the Greek word for stone, *lithos*), Li, atomic number 3, A_r 6.941, is the lightest element that is solid at normal temperature. Its natural isotopes ^{6}Li, A_r 6.015, natural abundance 7.42 %, and ^{7}Li, A_r 7.016, natural abundance 92.58 %, are not radioactive. The artificial isotopes ^{8}Li, ^{9}Li, and ^{11}Li have half-lifes of 842, 178, and 8.7 ms, respectively (→ Radionuclides). The electronic configuration of the free atom is $1s^2\ 2s^1$. The atomic radius is 0.155 nm (in the metal), and the ionic radius is 0.086 nm [1].

Lithium is a member of group 1 of the periodic table and thus has a valence of + 1. In many respects (e.g., the strong alkalinity of lithium hydroxide solutions and the reactivity of the metal) it has the character of an alkali metal, but it also exemplifies the diagonal relationship in the periodic table in its resemblance to magnesium: the metal forms only the oxide on combustion (not the peroxide), and with nitrogen it forms the nitride directly. Lithium carbonate, fluoride, and phosphate are very slightly water soluble. In contrast to sodium chloride, lithium chloride is soluble in polar organic solvents such as methanol.

History. The lithium minerals petalite and spodumene were discovered between 1790 and 1800 by Jose de Andrada in an iron ore deposit on the Swedish island of Utö. In 1817 August Arfvedson discovered the element lithium in this petalite deposit after finding discrepancies in his analytical values. Berzelius named the element lithion to indicate its occurrence in a stone. In 1818 Davy and Brandé first prepared lithium metal in extremely small amounts by electrolysis of lithium oxide with a voltaic pile. At about the same time, Gmelin discovered the red flame coloration produced by lithium salts. In 1825 Berzelius determined the lithium content of various mineral sources. The first preparation of the metal in gram quantities was achieved in 1854 by Bunsen and Matthiessen by electrolysis of molten lithium chloride.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_393

For a long time, lithium and its compounds were of only minor interest (e.g., in pharmaceutical applications and as additives to glazes). Only after World War I did an appreciable demand for lithium develop, owing to its use in lithium-containing (0.04 %) lead-based alloys (Bahnmetall) for bearings. This was the reason for the first industrial production of lithium carbonate and lithium metal in 1923 based on the mineral zinnwaldite by the Hans-Heinrich works of the Metallgesellschaft in Langelsheim, Federal Republic of Germany. The availability of lithium carbonate in tonnage quantities made its industrial use possible, at first mainly in the ceramic industry. In 1929, the Maywood Chemical Works in the United States began production of lithium metal and lithium compounds based on German know-how. Lithium production in the United States increased substantially during World War II due to the use of lithium hydride as a convenient source of hydrogen, especially for small balloons carrying transmitter aerials for emergency use by ships and aircraft in distress and by lifeboats. This has been followed in the last 50 years by increasing use of lithium-based soaps in multipurpose lubricating greases for automotive and industrial use. A further upsurge in demand came from the nuclear industry: the isotope ^{6}Li yields tritium when subjected to a neutron flux (→ Nuclear Technology).

The world market for lithium and lithium compounds is now supplied mainly from production facilities in the United States and from the Salar de Atacama in Chile [Cyprus Foote Mineral, FMC Lithium Division, Sociedad Minera Salar de Atacama S.A. (MINSAL)], and from production by Chemetall GmbH (a subsidiary of Metallgesellschaft), Frankfurt/Main in Europe. In addition, China and the Soviet Union export some lithium carbonate and hydroxide and, to a lesser extent, lithium chloride and lithium metal. The small production facilities in Japan are for domestic use only. Between 1960 and 1964, lithium carbonate was produced by Quebec Lithium in Canada. Other minor producers of lithium carbonate and hydroxide, e.g., Maywood Chemical Works, Montecatini in Italy, Rhône-Poulenc in France, and Associated Lead Manufacturers in the United Kingdom, have all since ceased production for economic or technical reasons.

2. Properties

2.1. Physical Properties

Lithium, a silvery white metal, is the least dense of all known solids at room temperature, *mp* 180.54 °C, heat of fusion 431.4 J/g, *bp* 1342 °C. It has a very low Mohs hardness of 0.6 and is therefore very easy to form.

The density at 20 °C is 0.534 g/cm^3, and the density of the liquid between 200 and 1600 °C is given very accurately by the equation [2]

$$\varrho = 0.515 - 1.01 \times 10^{-4}\,(t - 200)$$

The vapor pressure at the melting point is calculated to be 1.88×10^{-8} Pa. Between 800 and 1400 °C, it is given by the relationship [2]

$$\log p = 10.4038 - 8283.1/T - 0.7081 \log T$$

The heat of vaporization is 22 705 J/g.

The dynamic viscosity at 180.5 °C is 0.599×10^{-3} Pa · s; between this temperature and 1000 °C, it follows the equation [2]

$$\log \eta = -1.5064 - 0.7368 \log T + 109.95/T$$

The surface tension σ at the melting point is 398 mN/m and, between 500 and 1600 °C, is given by [2]

$$\sigma = 0.16\,(3550 - T) - 95$$

The mean linear coefficient of thermal expansion α_{m} between 273 and 368 K is $56 \times 10^{-6}\,\mathrm{K}^{-1}$, and the mean coefficient of cubical expansion β_{m} between 291 and 453 K is $180 \times 10^{-6}\,\mathrm{K}^{-1}$. Lithium melts with a volume change of + 1.5 %. The volume of the molten metal between 185 and 235 °C is given by [3]

$$V_T = V_{298}\left[1 + 174.106 \times 10^{6}\,(T - 453)\right]$$

The specific heat c_p of lithium metal between 20 and 180 °C is $3.3 - 4.2\,\mathrm{J\,g^{-1}\,K^{-1}}$. For the molten metal at 450 °C, a value of $4.2\,\mathrm{J\,g^{-1}\,K^{-1}}$ is quoted. Between the melting point and 421.8 °C, the following equation is valid:

$$c_p = 1.083 - 2.00 \times 10^{-4} t$$

and from 421.8 to 895.3 °C:

$$c_p = 1.006 - 1.73 \times 10^{-5} t$$

The critical data are as follows: T_c = 3223 ± 600 K; p_c = 68.9 MPa; V_c = 66 ± 19 mL/mol; and ϱ_c = 0.120 ± 0.033 g/cm^3.

The thermal conductivity, in J s^{-1} m^{-1} K^{-1}, between 250 and 950 °C is given by

$$\lambda = 42.3 + 12.3 \times 10^{-3} t$$

The enthalpy, in J/g, between 500 and 1300 °C is given by [2]

$$\Delta H = -21.248 + 4.1902 \times t - 21.658/t$$

The ionization energy is 5.37 eV, and the electrode potential of the half cell ($Li \rightleftharpoons Li^+ + e^-$) is − 3.024 V.

Values in the range 0.23 – 0.31 nm have been quoted for the radius of the hydrated ion [4].

Metallic lithium has good electrical conductivity. The resistivity at 0 °C is 8.55 μΩ cm, and between 200 and 1000 °C it is given by

$$R = 18.33 + 3.339 \times 10^{-2} t - 6.795 \times 10^{-6} t^2$$

The electrical resistance increases with increasing density under pressure [3].

Between −180 °C and the melting point, lithium crystallizes in a body-centered cubic system (a = 0.35 nm at 20 °C). At ca. −190 °C, an allotropic transformation of α-lithium into a hexagonal modification with densest close packing occurs.

Very thin lithium films are reddish brown by transmitted light, a color similar to that observed with colloidal dispersions of lithium in inert liquids. Pure lithium vapor emits a red light. The arc spectrum of lithium between 607.82 and 230.22 nm exhibits more than 42 lines, the strongest of which is the red line at 670.8 nm; the orange line at 610.3 nm is also excited in a Bunsen flame. The red spectral line at 670.8 nm can be used to detect lithium chloride in amounts as low as 1.26 × 10^{-6} mg.

2.2. Chemical Properties

Lithium metal is very reactive, though considerably less so than other alkali metals. The presence of sodium as an impurity, even in amounts of 0.5 – 1 %, increases its reactivity, e.g., for the formation of lithium alkyls from lithium metal and organic halides.

A freshly cut surface of lithium metal has a silvery luster. At room temperature in dry air with a relative humidity of less than 1 %, the surface remains shiny for several days, although a very thin passive surface layer is formed that is hardly visible to the naked eye and consists mainly of lithium carbonate and oxygen-containing compounds [5]. Lithium metal can therefore be processed in dry air. However, in moist air a dull gray coating, consisting mainly of lithium nitride, lithium oxide, and lithium hydroxide, forms within a few seconds. If lithium ingots are allowed to remain in contact with air for some weeks, the reaction with atmospheric nitrogen extends into the interior of the metal with the formation of reddish brown lithium nitride and can lead to ignition. Even at room temperature dry nitrogen reacts slowly with lithium metal.

Protective gases for lithium metal include the noble gases, or pure sulfur hexafluoride up to 225 °C. Mineral oil is also suitable as a protective medium.

Lithium burns with a very luminous white flame, forming a dense white smoke consisting mainly of lithium oxide; the flame temperature is almost 1100 °C.

Burning lithium metal reacts with silicates (e.g., sand and concrete) and carbon dioxide. These substances are therefore unsuitable as fire-fighting agents, as is sodium hydrogen carbonate. Powdered limestone (calcium carbonate) and fire-extinguishing powders based on sodium chloride are suitable for fighting lithium fires. Another suitable material is Graphex (from the company CECA), a graphite – sulfate complex whose volume increases many times at high temperature, thereby effectively smothering the burning lithium.

Lithium reacts with water with formation of hydrogen, which ignites under normal conditions only if the metal is finely divided. Molten lithium reacts explosively with water.

Lithium reacts with hydrogen to form lithium hydride. This reaction is carried out on an industrial scale at 600 – 1000 °C. Lithium reacts with gaseous ammonia at elevated temperature to form lithium amide. The vigorous reaction

with halogens produces incandescence. Organic compounds containing active hydrogen or halogen usually react with lithium to form the corresponding organolithium derivative.

Lithium also reacts with boron, silicon, phosphorus, arsenic, antimony, and sulfur upon heating.

A suitable container material for molten lithium metal is low-carbon steel, up to ca. 700 °C. Because lithium reacts with carbides, carbon-steel vessels in contact with liquid lithium are destroyed by decarburization. Up to 1000 °C, niobium, tantalum, or molybdenum can be used; nickel is satisfactory up to 225 °C. Aluminum, calcium, magnesium, silver, platinum, and gold are attacked by molten lithium.

Lithium metal dissolves readily in liquid ammonia, the concentration of the saturated solution at −33.2 °C being 9.8 wt %. A saturated solution of lithium in ammonia has a density of 0.477 g/cm^3 at 19 °C, and is thus the lightest known liquid at this temperature. The concentrated solution has a bronze color, whereas dilute solutions show the blue color typical of alkali metals. Solutions of lithium in ammonia are used to produce lithium acetylide and for Birch reductions in organic chemistry.

3. Occurrence

Lithium is present in the upper 16 km of the earth's crust to the extent of ca. 0.006 %; it is the 27th most abundant element. Approximately 150 lithium minerals are known. Lithium enrichment in minerals took place in geological time by the fractional crystallization of molten magma. Lithium migrated to the low-viscosity molten phase which eventually solidified as pegmatite. Sometimes lithium ore deposits are also formed in the pneumatolytic phase, as shown by the frequent association of lithium minerals with pneumatolytically formed cassiterite. Typical examples of this are provided by the spodumene deposit at Manono in Zaire and at Kamativi in Zimbabwe. The lithium found in mineral water sources in magmatic zones was probably formed in a similar manner.

The four most important lithium minerals are amblygonite, spodumene, petalite, and lepidolite. Of these, spodumene and petalite are used directly in the form of ores or concentrates, especially in the glass industry, whereas spodumene concentrate is used in the production of lithium carbonate.

Natural brines with a high lithium chloride content, which are the end result of a natural leaching and evaporation process, are also important in the industrial production of lithium. Given the technology currently available, seawater, which has a lithium content of only ca. 0.17 ppm, cannot be regarded as a viable raw material for economic lithium production.

3.1. Important Lithium Minerals

Analytical data for the commercially important lithium minerals are given in Table 1.

3.1.1. Lithium Aluminum Silicates

Spodumene (triphane), $LiAlSi_2O_6$, has a theoretical Li_2O content of 8.03 %, but usually contains 6–7.5 % because of partial replacement of Li^+ by Na^+ and K^+. Spodumene has a monoclinic pyroxene structure consisting of $-SiO_3-$ chains linked together by aluminum ions, with the balance of the positive charge provided by lithium ions. Spodumene sometimes occurs in the form of large single crystals. Its color varies from white to gray-green to reddish. Very pure and transparent single crystals are regarded as semiprecious stones (kunzite). The high hardness value of spodumene (6.5–7 on the Mohs scale) leads to considerable abrasive effects when the material is worked. At ca. 1000 °C, naturally occurring α-spodumene undergoes an irreversible phase change to β-spodumene, whereby the volume increases greatly and the density decreases from 3.2 to ca. 2.4 g/cm^3. Spodumene ores consist of pure spodumene combined with quartz or feldspar; the Li_2O content ranges from less than 1 % to more than 5 %.

Petalite, $LiAlSi_4O_{10}$, has a theoretical Li_2O content of 4.9 % but usually contains only 3.5–4.5 %. It crystallizes in the form of monoclinic prisms and is white, gray, or reddish in color. It has a density of 2.4 g/cm^3 and a Mohs hardness of 6.5. Petalite decomposes at 1100 °C

Table 1. Typical analyses of lithium minerals, %

	Amblygonite	Spodumene concentrate	Spodumene, glassgrade	Petalite	Lepidolite
Li_2O	min. 7	min. 7	min. 4.8	min. 4	min. 3.5
K_2O	ca. 0.2	ca. 0.7	0.12	ca. 0.2	8 – 12
Na_2O	0.4 – 1.1	0.5 – 0.8	0.09	0.4 – 1.4	0.3 – 1.1
Al_2O_3	34 – 37	26 – 29	17 – 18	16 – 18	20 – 28
SiO_2	ca. 2.5	60 – 65	74 – 80	74 – 78	49 – 55
F	ca. 3.5	ca. 0.2	max. 0.1	ca. 0.1	4 – 6
Fe_2O_3	ca. 0.2	0.5 – 2	0.1 – 0.2	0.03 – 0.2	0.1 – 2
P_2O_5	44.5 – 50	ca. 0.5	max. 0.1	ca. 0.3	ca. 0.5
CaO	ca. 1	0.1 – 0.5	ca. 0.06	ca. 0.4	ca. 0.4
MgO	ca. 0.1	0.1 – 1	ca. 0.05	0.2 – 0.4	ca. 0.2

into β-spodumene and SiO_2. This transformation (and that of α-spodumene) is important in digestion of the ore.

Eucryptite, $LiAlSiO_4$, is a relatively rare lithium mineral that occurs mainly in combination with petalite, e.g., in Bikita, Zimbabwe. It has the highest Li_2O content (11.1 %) of the lithium aluminum silicates.

The three minerals are distinguished by their SiO_2 content:

Eucryptite: $Li_2O{\cdot}Al_2O_3{\cdot}2\,SiO_2$
Spodumene: $Li_2O{\cdot}Al_2O_3{\cdot}4\,SiO_2$
Petalite: $Li_2O{\cdot}Al_2O_3{\cdot}8\,SiO_2$

Thus, thermal transformation of petalite into β-spodumene and quartz can be regarded as a decomposition reaction.

3.1.2. Micas

The mica minerals form a subgroup of the aluminosilicates. They are sometimes formed by direct crystallization from molten magma but often owe their origin to pneumatolytic decomposition.

Lepidolite, $K[Li, Al]_3[Al, Si]_4O_{10}[F, OH]_2$, lithium mica, is the most important representative of this group. Depending on its origin and the weathering it has undergone, lepidolite can contain between 3.3 and 7.74 % Li_2O. Economic deposits contain 3 – 4 % Li_2O. The mineral is often pink to grayish violet and frequently contains 3 – 5 % rubidium and cesium oxides.

Zinnwaldite, which has the general formula $K[Li, Al, Fe]_3[Al, Si]_4O_{10}\,[F, OH]_2$, can be regarded as a variety of lepidolite with a high iron content. The Li_2O content varies with the actual composition of the mineral. The deposit in Zinnwald (Erzgebirge) contains 2 – 3 % Li_2O. Zinnwaldite is silvery to grayish brown, sometimes also tinged with red. The emerald green variety known as cryophyllite, found in Massachusetts, contains up to 5 % Li_2O.

3.1.3. Lithium Phosphates

Commercial amblygonite is a complex phosphate with the general formula $(Li, Na)Al(F, OH)PO_4$. Its principal constituents are pure amblygonite, $LiAlFPO_4$, with a theoretical Li_2O content of 10.1 %, and its hydroxyl substitution product, montebrasite, $LiAl(OH)PO_4$, with 10.3 % Li_2O. Amblygonite is milky white to gray in color, sometimes also with a brown or green tinge; it may have a sheen resembling a solidified stream of molten glass or a pearly luster. Deposits suitable for mining have an Li_2O content of 7 – 9 %, the highest of the commercially exploited lithium ores. However, amblygonite is never found in large accumulations but is widely distributed in small, usually lenticular deposits, often associated with other lithium minerals. Although amblygonite once was an important source of lithium, it is no longer regarded as a suitable source of raw material for today's large-scale lithium carbonate production plants.

Triphyline, $LiFePO_4$, and lithiophilite, $LiMnPO_4$, have a theoretical Li_2O content of 9.5 %. The known deposits yield a raw material

with 8 – 9 % Li_2O; however, they have not been commercially exploited because of their small size.

3.1.4. Other Lithium Ores

Lithiophorite, $[Li, Al]MnO_2(OH)_2$, is often found in veins of manganese ore.

Hectorite, which has the approximate composition $LiNaMg_8(Si_4O_{10})_3[F, OH]_6$ and contains ca. 1.2 % LiO_2, is the only lithium mineral that occurs in sedimentary rock and clay; it is classified in the montmorillonite group.

The unusual mineral zhabuyelite [6] was discovered recently in salt deposits in the Zha Buye lake in Tibet. It consists of colorless, transparent, monoclinic crystals of almost pure lithium carbonate.

3.2. Reserves of Lithium Minerals [7], [8]

The only mineral used as a raw material for the production of lithium compounds at present is spodumene, the largest known deposits of which are at Manono in Zaire. Reserves have been estimated to be over 2×10^6 t of lithium [9], [10]. New investigations by Lithium Australia into their own pegmatite deposits in Greenbushes (Western Australia) [11] have revealed reserves with 7×10^5 t of lithium, of which ca. 10 % is in the form of spodumene with more than 4 % Li_2O, ca. 50 % with more than 2.5 % Li_2O, and ca. 40 % with ca. 1.5 % Li_2O.

The U.S. producers Cyprus Foote Mineral and FMC Lithium Division base their operations on deposits in the "tin – spodumene belt" of North Carolina, which may contain as much as 2.8×10^6 t of lithium [9], [10]. Cyprus Foote Mineral ceased its spodumene operations in 1986. Since then it produces lithium only from brines.

The Canadian deposits at La Corne, which provided raw material for Quebec Lithium, are estimated to contain 10^5 t of recoverable lithium, with potential reserves of about twice this amount. Deposits belonging to Tantalum Mining Co. at Bernic Lake amount to ca. 60×10^3 t. The total known reserves in Canada are quoted as 6×10^5 t of lithium [12].

The deposits at Bikita in Zimbabwe consist mainly of petalite and lepidolite with some spodumene and eukryptite. They contain between 7.5×10^4 and 10^5 t of lithium. Russia has deposits of spodumene and lepidolite at Krivoj Rog and in Chita in Transbaikal. Other lithium minerals occur in the Altai mountains and on the Kola peninsula. No details are available concerning the extent of these reserves.

In northwest China, considerable reserves of spodumene containing ca. 1.5 % Li_2O have been reported in the southern slopes of the Altai mountains in the Koktohay – Aletai region. This provides raw material for lithium carbonate and hydroxide production in Urumchi. No details are available.

In addition to the principal reserves described above, a number of smaller ones exist in the United States, Canada, Australia, Zimbabwe, and China. Lithium minerals have also been found in Afghanistan, Argentina, Brazil, India, Madagascar, Mali, Mozambique, Namibia, and South Africa [13], [14].

The reserves in Europa are all of minor importance and so far have not proved worth mining. This is also true for the spodumene deposit that has recently been investigated in the Koralpe of Austria [15]. This contains 5×10^6 t of ore with an average of 1.7 % Li_2O. Spodumene has also been found in Lalin and near Caceres in Spain, Dublin in Ireland, and Tammela in Finland. Lepidolite deposits exist in France (Massif Central), Portugal, the island of Elba, Sweden, and central Finland. Petalite also occurs on Elba and at Utö in Sweden.

The confirmed lithium reserves based on pegmatite in the Western world are estimated at 3×10^6 t of lithium if recent figures from Greenbushes are included [16]; potential reserves amount to an additional 4×10^6 t [9].

Lithium ores are used directly or as flotation concentrate (sometimes mixed with other materials), e.g., in the glass industry or as slag formers in the foundry industry (see Chap. 9).

3.3. Lithium in Natural Brines

In the last decade, lithium-containing brines have become the most important raw materials for the production of lithium chemicals. Their industrial exploitation began at Searle's Lake,

California, in 1938 by Kerr-McGee Chemical, and in Clayton Valley, Nevada in 1966 by Foote Mineral (now Cyprus Foote Mineral).

In 1984, production of lithium carbonate from the brines of the Salar de Atacama in Chile was started by the Sociedad Chilena De Litio Ltda (SCL), now a 100 % affiliate of Cyprus Foote Mineral [17]. At the same Salar, Minsal S.A., a subsidiary of the Chilean mining company Sociedad Quimica y Minera de Chile S.A. (SQA, Soquimich) has also started its brine-based lithium carbonate production. In Argentina at the Salar del Hombre Muerto, FMC Corp. is building a new brine-based lithium extraction plant, due for start up in 1997.

Because of the extremely high solubility of lithium chloride in water, the basic concentration process for lithium-containing brines is natural evaporation such as occurs only in dry regions of the earth in particular geological formations [18]. The evaporation process leads first to increased lithium content in salt lakes, such as the Dead Sea or the Great Salt Lake. Further enrichment of lithium then takes place if the surface of the lake is completely dried out as in Searle's Lake and Clayton Valley. This enrichment also occurs in the last brines to be formed, which are enclosed in cavities. The surrounding salts can be practically lithium free or can contain up to 1500 ppm of lithium.

Important reserves of lithium-containing brines are found in western North and South America and in western China (Tibet and Qinghai Province) [6], [19], [20]. Their lithium contents vary greatly (see Table 2).

The reserves of lithium in these brines are sometimes considerable. It is estimated that the Salar de Atacama in Chile contains more than 1.1×10^6 t of lithium, Clayton Valley 7×10^5 t of lithium, and the Great Salt Lake 4.5×10^5 t of lithium. The quantities of lithium that can be recovered economically from these reserves are, however, much smaller. For the Clayton Valley they are estimated to be 4×10^4 t of lithium and for the Great Salt Lake, with coproduction of other salts, 2.6×10^5 t [9]. On the other hand, the amount of recoverable lithium from the Salar de Atacama, because of the higher concentration, is considerably larger (up to 8×10^5 t). A study by Amax Exploration of the recovery of 15×10^3 t/a of lithium carbonate as a coproduct with potassium chloride, potassium sulfate, and boric acid has been completed [21].

Not only is the lithium yield from brines reduced due to dilution caused by replacement of the withdrawn brine by fresh water, lithium is also lost through inclusion in salt crystals that form upon evaporation and in the sludge produced when magnesium is precipitated with lime. The verified reserves of lithium from salt lakes are estimated to be 1.8×10^6 t, and the potential reserves 18×10^6 t.

The recovery of lithium from material of low lithium content is also interesting, because of the importance of lithium to nuclear technology. An example is the proposed extraction from the Dead Sea by precipitation as lithium aluminate, $LiH(AlO_2)_2\cdot 5\,H_2O$ [22]. Also under investigation is the recovery of lithium from seawater [23] which, although it contains only 0.17 ppm of lithium (see Table 4), represents a total lithium reserve of 2.5×10^{11} t.

Another potential source of lithium is provided by subsurface brines associated with mineral oil. The brine lies beneath the oil layer and is often recovered at the same time as the oil. In Texas, the brines contain up to 692 mg of lithium per liter, but only rarely does the lithium content exceed 100 mg/L [24]. However, the high magnesium content presents a serious technical problem. For example, in the Smack-over brines in Arkansas the magnesium concentration is 20 times that of lithium. Given the other lithium resources available, industrial exploitation of the lithium in these brines does not appear to be feasable for economic reasons.

4. Production of Primary Lithium Compounds

4.1. Mining of Ore and Production of Concentrate

Lithium ores are recovered by quarrying or open-cut mining according to their location. Because the veins of ore are often only a few meters thick, the mining operation has to follow the veins, or open-pit mining must be used, which involves the removal of relatively large volumes of overburden. The ratio of overburden to ore is usually 3 : 1 – 5 : 1. An important exception is

Table 2. Typical analyses of brines, wt %

	Li^+	Na^+	K^+	Mg^{2+}	Ca^{2+}	SO_4^{2-}	Cl^-	Br^-
Clayton Valley (Nevada)	0.02	7.5	1.0	0.03	0.05	0.75	11.7	
Salar de Atacama (Chile)	0.15	7.6	1.8	0.96	0.03	1.78	16.00	trace
Salar de Uyuni (Bolivia)	0.024	9.42	0.51	0.44	0.05	0.72	15.91	trace
Salar del Hombre Muerto (Argentina)	0.06	9.8	0.6	0.09	0.05	0.82	15.9	trace
Salar del Rincón (Argentina)	0.05	10.8	0.7	0.4	0.12	0.7	16.7	trace
Zha Buye Lake (Tibet)	0.12	14.17	3.96	0.001	trace	4.35	15.63	0.6
Dachaidan, Qinghai (China)	0.02	10.6	0.4	1.3	0.04	2.25	18.7	7.0
Dead Sea (Israel/Jordan)	0.002	3.0	0.6	4.0	0.3	0.05	16.0	0.4
Great Salt Lake (Utah)	0.004	7.0	0.4	0.8	0.03	1.5	14.0	
Boneville (Utah)	0.007	9.4	0.6	0.4	0.12	0.5	16.0	
Seawater	0.000017	1.8	0.038	0.13	0.04	0.27	1.94	0.0004

the Greenbushes deposit in Australia, which has veins of ore up to 250 m thick.

If the lithium minerals are largely separated from the accompanying rock, preliminary sorting can be carried out by hand on a conveyor belt. In this way, it is possible to produce spodumene containing 4 – 5.5 % Li_2O from raw material containing 1 – 1.5 % Li_2O. A mechanical separation process, based on the difference in reflective properties between ore and gangue material, can also be used. This can give good results if the material is first suitably reduced in size.

The concentration of lithium ores by heavy-medium separation always entails large losses of lithium and has therefore not been used commercially.

A flotation process is employed on an industrial scale especially for spodumene [25]. The material is first ground to a grain size of less than 0.3 mm by using rod mills. Material finer than 0.075 mm is then removed (e.g., by hydrocyclone) because it would result in lower recoveries in the flotation process. Flotation agents such as anionic fatty acids in alkaline medium and sulfonated oils in acid medium are recommended. The use of cationic surfactants in hydrofluoric acid solution has also been proposed. The flotation process is carried out in a series of stages. Quartz, feldspar, and mica are concentrated in the foam, and spodumene is floated off. The spodumene concentrate contains 5.5 – 7.5 % Li_2O. It can be sieved into fractions with different lithium contents; the coarser fractions have the higher lithium contents. The yield of lithium from the raw mineral after concentration by flotation is 60 – 80 %, depending on the milling conditions used.

4.2. Ore Digestion and Production of Lithium Compounds

The digestion of lithium ore can be carried out with acid or alkali, or by an ion-exchange process using salts. Various processes are reviewed in [26]. Depending on the process and the digestion agent, the product obtained is lithium carbonate, lithium hydroxide, or lithium chloride.

Although amblygonite, triphyline, and lepidolite can be treated directly with the appropriate digestion agent, spodumene and petalite require pretreatment (decrepitation) to convert α-spodumene to β-spodumene, and petalite to β-spodumene and quartz. Decrepitation is carried out in rotary kilns at 1000 – 1100 °C. It results in a 30 % increase in the mineral's volume and a considerable improvement in its grindability.

4.2.1. Acid Digestion

The sulfuric acid digestion process can be used for all lithium ores. It is usually carried out at 250 – 400 °C, although 850 – 900 °C is recommended for amblygonite because this leads to the formation of insoluble aluminum phosphate which facilitates removal of the insoluble residue.

Sulfuric acid digestion gives a product that contains an insoluble residue, together with lithium sulfate which is leached from the crushed product with hot water. To remove alkaline-earth metals, aluminum, and iron, the solution is neutralized with lime and soda and then filtered. Lithium carbonate is precipitated by adding concentrated sodium carbonate solution. This is carried out near the boiling point, to take advantage of the decreasing solubility of lithium carbonate at higher temperature.

The sulfuric acid process gives good yields of lithium (ca. 90 %) and has a relatively favorable energy balance compared with other processes for the production of lithium compounds from ores. Digestion by hydrochloric acid or hydrogen chloride gas at 935 °C has been suggested for the production (by sublimation) of lithium chloride from lepidolite. This process has, however, little significance compared to the sulfuric acid process.

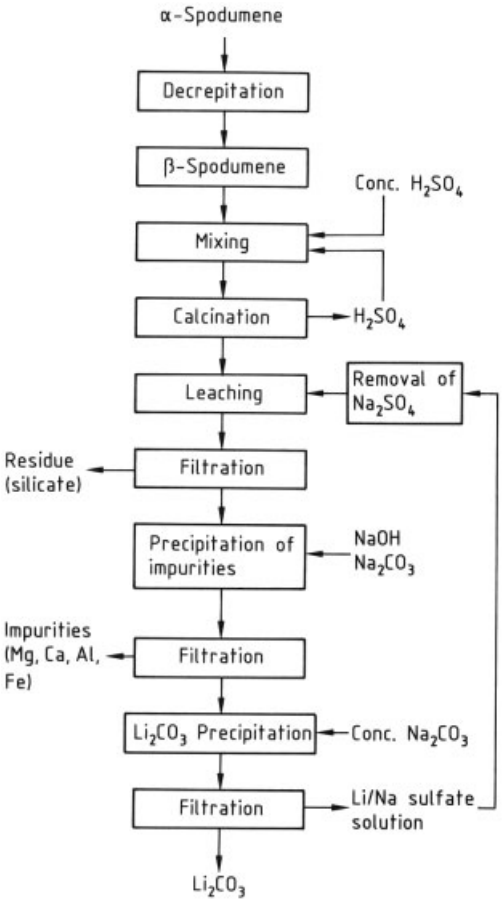

Figure 1. Flow sheet for lithium carbonate production from spodumene

The production of lithium carbonate from spodumene (petalite) by using sulfuric acid is described in detail in [25]. The flow sheet for the process is shown in Figure 1. The ore is first crushed to a particle size of 20 – 80 mm. Alternatively, flotation concentrate, which is already finely divided, is used. The crushed ore is fed into an oil-, coal-, or gas-fired rotary decrepitation kiln, which attains the required temperature (1050 – 1100 °C) over at least 10 m of its length. The spodumene takes ca. 4 h to pass through the rotary kiln, the dwell time in the decrepitation zone being 15 – 30 min, which results in practically complete transformation into β-spodumene.

The heat efficiency is improved by using a countercurrent flow of material and hot gas. The exhaust gases are used to preheat the ore, to heat the roasting furnace, and to concentrate the liquors by evaporation. The spodumene is quite hot as it leaves the rotary kiln, and sometimes a rotary cooler, which can also be used to preheat the combustion air, is necessary.

If the starting material is in lump form, the β-spodumene produced by decrepitation is crushed to a grain size of less than 0.1 mm. Sulfuric acid (93 %, in 35 % excess) is then added either by a mixer conveyor or by spraying onto a band conveyor fitted with a deflector plate. The digestion process, in which lithium in the ore is converted into soluble lithium sulfate, is carried out at ca. 250 °C in a lined rotary furnace or on a conveyor heated by exhaust gases, which carries the spodumene through a tunnel directly to the leaching stage. Digestion is usually complete in less than 1 h.

The lithium sulfate formed is leached out of the reaction product with hot water in Pachuca tanks agitated by compressed air or in stirred vessels. A continuous leaching process, in which the material and the water move countercurrently through several stirred vessels in series, has proved advantageous. After leaching, aluminum silicate residues are removed by a rotary vacuum filter, and the filtrate is treated with lime and soda to precipitate alkaline-earth metals and iron. After repeated filtration, the solution is neutralized with sulfuric acid. During neutralization and in the subsequent evaporation, which concentrates the solution to 200 – 250 g/L of lithium sulfate, aluminum hydroxide precipitates and is removed together with any purifying additives (e.g., activated charcoal).

The precipitation of lithium carbonate by concentrated sodium carbonate solution is carried out near the boiling point. Lithium carbonate is removed by filtration or centrifugation and washed with hot water to remove the adhering mother liquor (mainly sodium sulfate). It is then

dried with a vacuum plate dryer, a band dryer, or a vacuum screw conveyor dryer.

The particle size of lithium carbonate depends strongly on the concentration during precipitation, the temperature, and the rate of addition of sodium carbonate.

Mother liquor from lithium carbonate precipitation always contains ca. 1 g of lithium per liter and is therefore recycled. Before the liquor is returned to the leaching stage, sodium sulfate is removed by crystallization at low temperature. This is a valuable byproduct and improves the economic viability of the process.

The use of fluidized-bed furnaces has been proposed for the decrepitation process, enabling it to be carried out under mild conditions (see Fig. 2). Recycling of excess sulfuric acid from the dissolution process has also been proposed [27].

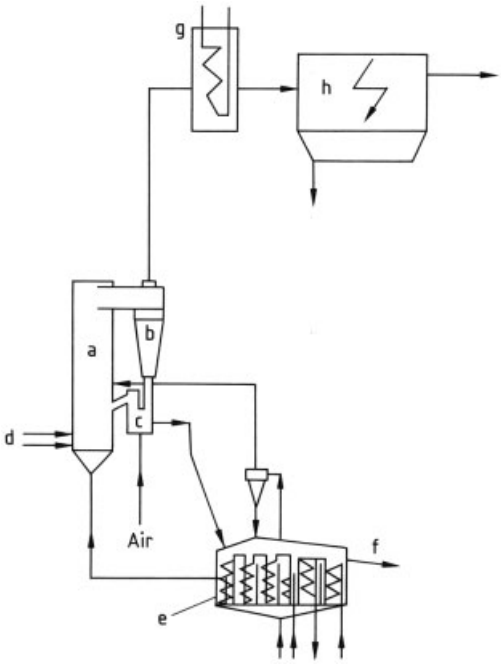

Figure 2. Schematic of a fluidized-bed reactor for spodumene decrepitation
a) Fluidized-bed reactor; b) Cyclone; c) Oversize return; d) Raw material feed and fuel supply; e) Countercurrent cyclone cooler; f) Product offtake; g) Heat exchanger (waste-heat boiler); h) Electrostatic precipitator

4.2.2. Alkali Digestion

Spodumene and lepidolite can also be digested by reaction with quicklime. Decrepitation of spodumene, calcination of limestone, and digestion are carried out in a single operation. Three parts of limestone are wet milled with one part of lepidolite, or 3.5 – 4 parts of limestone with one part of spodumene. The mixture is dried in a rotary furnace and then heated to the clinker point to give calcium orthosilicate and lithium aluminate together with excess lime.

The process is critically dependent on the temperature; lepidolite requires ca. 900 °C, and spodumene ca. 1040 °C. If the temperature is too high, a glassy clinker is formed, which severely hinders subsequent leaching of lithium hydroxide; too low a temperature leads to incomplete reaction. Control of the kiln is complicated because of caking and building up of the reaction mixture. In the case of lepidolite, the clinker obtained must be quenched immediately so that no reverse reaction occurs. Spodumene clinker may be cooled slowly, but quenching produces a clinker with better grindability.

After the clinker has been ground in ball mills, lithium hydroxide is leached out with water in several stages. The residues are then separated, and the solution is concentrated in multistage vacuum evaporators to crystallize lithium hydroxide monohydrate. The process has a relatively high energy requirement and gives a lithium yield appreciably below that of the sulfuric acid process. The lower yield is due to the problem of precisely controlling the kiln. For this reason, only the sulfuric acid process is used in the Western world. It is sometimes followed by the conversion of the lithium carbonate into lithium hydroxide. The production of lithium chloride by sublimation from spodumene mixed with calcium chloride and calcium oxide in a one-stage process with decrepitation at 1100 °C also has not become established.

4.2.3. Ion-Exchange Processes

Ion-exchange processes involve heating a lithium ore with a sodium or potassium salt. This leads to replacement of Li^+ ions in the ore by Na^+ or K^+ and formation of the soluble lithium salt of the anion used.

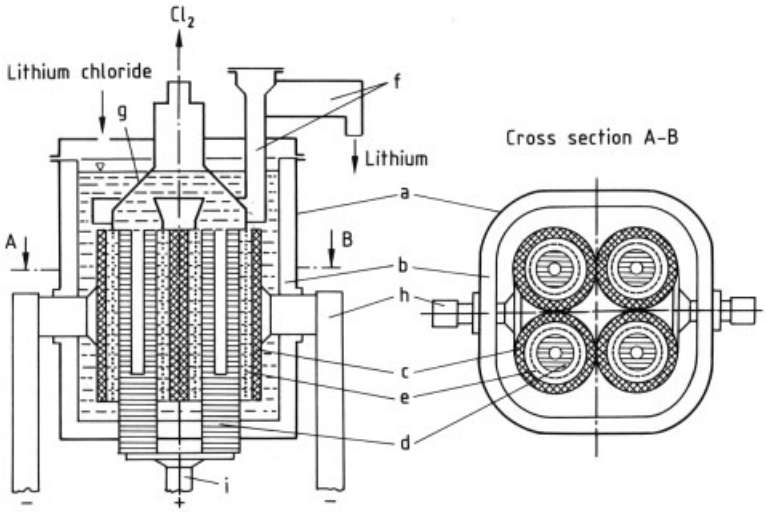

Figure 3. Schematic of an electrolytic cell for the production of lithium
a) Steel casing; b) Ceramic lining; c) Cathode with supporting arm; d) Anode; e) Diaphragm; f) Annular collection zone with riser pipe and overflow for lithium; g) Dome for chlorine collection and removal; h) Cathode conductor rail; i) Anode conductor rail

The digestion of lepidolite by potassium sulfate at 825 – 875 °C has attracted much interest. Lithium is leached out as lithium sulfate and then separated from excess potassium sulfate. After evaporation, potassium sulfate is crystallized, and lithium carbonate is precipitated by addition of sodium carbonate to the solution. The potassium sulfate is then reused in the dissolution stage [28]. However, the loss of potassium in the form of potassium aluminum silicate in the insoluble residues is far from negligible, and for this reason the process is no longer used. If lithium production from lepidolite were to be resumed, the sulfuric acid process would probably be favored over ion exchange, because it also produces aluminum sulfate as a byproduct and is therefore more economical.

The digestion of β-spodumene with sodium or potassium acetate is carried out at ca. 400 °C [29]. It produces lithium acetate from which lithium carbonate is precipitated by addition of sodium carbonate; the resulting sodium acetate is recycled. In theory, the process consumes an amount of sodium carbonate equivalent to the lithium carbonate produced, the sodium acetate acting only as a carrier and remaining in the system. However, so far, the process has apparently been operated only on a pilot-plant scale.

The Quebec process of the Quebec Lithium Corporation is a modification of the acetate process in which β-spodumene is treated with aqueous sodium carbonate solution (30 % excess) at ca. 200 °C and 5 MPa for at least 10 min. A slurry of sodium aluminum silicate and lithium carbonate is formed, and subsequent treatment with carbon dioxide at low temperature dissolves the lithium, forming the soluble bicarbonate. The insoluble residue is a sodium zeolite that can be marketed. Lithium carbonate is precipitated from the solution by heating to the boiling point. The carbon dioxide generated is used again for the dissolution of lithium carbonate [30].

This process was used on an industrial scale until 1966. It requires considerably more energy than the sulfuric acid process because of the repeated heating of large quantities of water. Furthermore, sodium carbonate is lost through formation of sodium hydrogen carbonate during dissolution of the lithium carbonate with carbon dioxide.

In a further development of the Quebec process, pressure digestion of β-spodumene is carried out at 250 °C with aqueous sodium chloride solution made alkaline by the addition of 0.1 – 0.2 % calcium hydroxide. Lithium carbonate can be precipitated from the resulting liquor

by addition of sodium carbonate. Alternatively, lithium chloride can be isolated by solvent extraction (e.g., with *n*-butanol) [31].

4.3. Production of Lithium Carbonate from Brines

The lithium in brines is at present extracted commercially only at Clayton Valley and the Salar de Atacama. In Clayton Valley, the brines are pumped from wells up to 100 m deep into a series of solar ponds, in which the lithium concentration increases continuously while the accompanying salts precipitate due to oversaturation. The production process requires the lithium concentration to be at least 5000 ppm. If the initial concentration is 200 ppm, the enrichment factor must be at least 30 to obtain solutions suitable for the production process, because lithium losses take place by inclusion in precipitated salts and by seepage of the solution through the solar ponds. This means that at least 95 % of the water in the original brine must be evaporated. Such a process can only be operated economically where solar energy is sufficient to provide high rates of evaporation, i.e., in hot, dry regions of the earth. Brines containing 100 ppm of lithium can be concentrated economically if the magnesium content is low. The concentrated brine is purified and then sodium carbonate solution is added at high temperature to precipitate lithium carbonate, which is separated and dried.

Brines containing less than 100 ppm of lithium (e.g., from Searle's Lake) require a different process to recover the lithium, e.g., by precipitating it first as lithium sodium phosphate, Li_2NaPO_4. This lithium concentrate, known as licon, which contains 20 – 21 % Li_2O is then treated with concentrated sulfuric acid, converting it to phosphoric acid and a solid mixture of lithium and sodium sulfates. This is dissolved, and lithium carbonate is precipitated from the solution by addition of sodium carbonate. The yield of lithium in this process is ca. 88 %, which is very good in view of the low starting concentration. The economic viability of the process depends on the recovery of associated byproducts such as potassium chloride, sodium sulfate, sodium carbonate, borax, phosphoric acid, and bromine [32]. Lithium production at Searle's Lake was abandoned in the '70s.

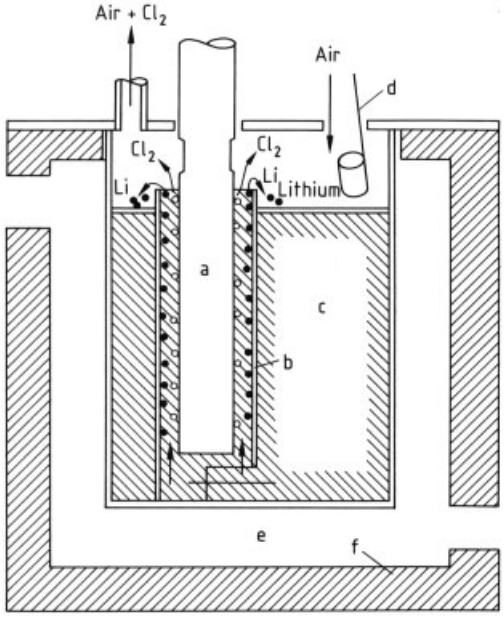

Figure 4. Schematic of an American lithium electrolysis cell
a) Graphite anode; b) Steel cathode; c) Melt; d) Ladle; e) Heating chamber; f) Insulation

5. Lithium Metal and Lithium Alloys

5.1. Production of Lithium Metal

Lithium metal is produced only by the electrolysis of molten lithium chloride, whose melting point of 614 °C is lowered by addition of potassium chloride. The lithium chloride – potassium chloride eutectic, with 44.3 % lithium chloride, melts at 352 °C. The salt mixtures used in industry contain 45 – 55 % lithium chloride, which allows electrolysis to be carried out at 400 – 460 °C.

The electrolytic cells most commonly employed (Fig. 3) resemble the Downs cell used for the production of sodium [33].

The cell and the cathode are constructed from steel, whereas the anode is made of graphite. Four cathode tubes are combined to form a unit, with each tube enclosing an anode cylinder. The anodes enter the electrolysis space from below, and the cathodes are supported by arms attached to the cell sides. The anode and cathode are separated by a wire-mesh steel diaphragm. The cell is equipped with an insulating inner lining, because both electrodes are introduced below the

surface of the melt. Lithium metal is formed at the cathode, rises to the top, and is collected under an argon atmosphere.

In the United States, lithium cells are sometimes operated without a diaphragm or ceramic insulating lining (Fig. 4) [34]. This greatly prolongs the lifetime of the cells but results in greater power consumption.

Another lithium electrolysis cell (Fig. 5) has neither ceramic insulation nor a diaphragm [35]. In this cell, a molten salt mixture containing the lithium metal formed rises continuously in the electrode space and is collected in an annular trough at the top of the cathode near the surface of the melt. It is then transferred through a siphon-like tube to a separation vessel connected to the cell, but isolated from the chlorine gas. Here, the lithium metal is separated from the electrolyte under an inert-gas atmosphere; the electrolyte is recycled, and the lithium metal is recovered.

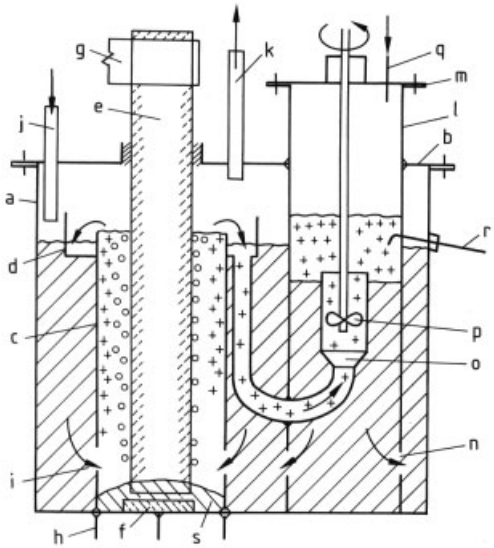

Figure 5. Schematic of a special electrolytic cell for production of lithium
a) Electrolysis cell; b) Cover plate; c) Cathode; d) Collecting trough; e) Graphite anode; f) Insulating material; g) Anode connection; h) Cathode connection; i) Openings in cathode wall; j) Charging pipe for lithium chloride; k) Chlorine outlet; l) Separating cylinder; m) Cover plate; n) Openings in wall of separating cylinder; o) Pipe connecting collecting trough to separating cylinder; p) Agitator; q) Inlet for inert gas; r) Outlet pipe for lithium; s) Solidified melt

The output of an individual cell is usually 200 – 300 kg of lithium metal per day. The decomposition voltage of lithium chloride is ca. 3.68 V, but in practice 6.0 – 6.8 V is usually required. The theoretical power consumption is 14.2 kW · h/kg of lithium, but the actual figure is 28 – 32 kW · h/kg of lithium. The current efficiency is ca. 90 %.

Lithium metal is shipped in steel drums or cans under an argon atmosphere or under a coating of oil. Lithium metal can be stored indefinitely in the absence of air.

Molten lithium can be cast in low-carbon steel molds without special precautions. Its low hardness enables it to be easily extruded or rolled. However, to maintain the silvery luster of lithium, the operation must be carried out in air of very low moisture content or under an inert-gas atmosphere.

5.2. Uses of Lithium Metal

The largest application of lithium metal is the production of organolithium compounds, principally butyllithium, and lithium hydride.

Lithium is used as a reducing agent in organic chemistry, as a solution in liquid ammonia for Birch reductions, and in the synthesis of vitamins [36].

A rapidly growing application of lithium metal is in sacrificial anodes in batteries. Lithium has a high electrochemical equivalent (3.86 A · h/g) and a low electrochemical potential (– 3.045 V). The main advantages of lithium batteries are high power output, extremely good storage properties (low rate of self-discharge), and relatively small power loss at low temperature (→ Batteries, Chap. 5.1.). Commercially produced primary cells have counterelectrodes made of manganese dioxide, chromium oxide (CrO_x), or graphite loaded with sulfur dioxide or thionyl chloride, and electrolytes consisting of lithium salts dissolved in organic solvents such as propylene carbonate or dimethoxyethane. They have cell voltages up to 3 V and energy densities up to 300 W · h/kg [37]. Secondary lithium cells, some based on aluminum – lithium alloys, are being investigated extensively [38], [39].

Lithium metal is becoming increasingly important as an alloying component (see Section 5.3); it is also used as a deoxidizing and desul-

furizing agent, especially for copper, nickel, and alloy steels.

Lithium and its compounds play a number of important roles in the nuclear industry, e.g., in the production of tritium, as a heat-exchange cooling medium, as a shielding material, and in the form of a molten salt mixture, as a solvent for other nuclear fuels.

These applications depend on the extremely large difference in neutron capture cross section and absorption distance between the two stable lithium isotopes (see Table 3) [40].

Table 3. Thermal neutron absorption cross section σ and absorption distance λ for lithium

Absorbent	σ, $cm^2 \times 10^{-24}$	λ, mm ∗
Natural isotopic mixture	71	3
^6Li	945	0.2
^7Li	0.0033	8000

∗ Distance at which radiation energy has decayed to 1/**e** of the original intensity.

Production of Tritium. The extremely large amount of energy released in thermonuclear reactions (see Table 4) is illustrated by the hydrogen bomb, in which tritium or its parent substance lithium in the form of $^6Li^2D$ reacts with deuterium nuclei.

Table 4. Thermonuclear reaction energies

Reaction	Energy, MeV
$^6Li(n, T)^4He$	4.8
$^7Li(n, nT)^4He$	2.5
$T(p, \gamma)^4He$	19.7
$T(D, n)^4He$	17.6
$T(T, 2n)^4He$	11.3

Several fusion reactions have been proposed for power generation (→ Nuclear Technology). Of these the reaction between deuterium and tritium

$$^2D + {}^3T \longrightarrow {}^4He + n + 17.6\,MeV$$

has the best prospect of success because it requires the lowest plasma temperature [41]. The tritium required for the reaction is produced from ^6Li by irradiation with thermal neutrons:

$$^6Li + n \longrightarrow {}^4He + {}^3T + 4.8\,MeV$$

The overall equation for the fusion reaction is [42]:

$$^6Li + {}^2D \longrightarrow 2\,{}^4He + 22.4\,MeV$$

This reaction takes place with loss of neutrons; therefore, neutron multiplication reactions (n, 2 n) are used to breed tritium, e.g., based on beryllium or using ^7Li as the starting material according to the equation

$$^7Li + n \longrightarrow {}^4He + {}^3T + n - 2.5\,MeV$$

Calculations based on this reaction show that 1 kg of natural lithium could produce 5×10^{13} J of energy.

Coolants. In addition to helium at ca. 5 MPa, the following coolants have been proposed for the cooling jacket of a fusion reactor: molten lithium metal, molten $LiF-BeF_2$, and molten lead containing 17 atom% lithium. In the latter materials the tritium breeding reaction can take place [41], [43]. An important parameter is the product of density and specific heat which shows that, compared with sodium as a coolant, a lower circulation rate of lithium metal (by a factor of 1.9) can be used. For $LiF-BeF_2$ this factor is 4.5 [44].

Solvents for Nuclear Fuels. Owing to the low neutron capture cross section of ^7Li, a mixture of ^7LiF (72 mol %), beryllium fluoride (16 mol %), thorium fluoride (12 mol %), and uranium tetrafluoride (< 1 mol %) has been proposed as a molten salt fuel for breeder reactors. In these reactors, which are envisaged as small power plants (e.g., in aircraft), the $LiF-BeF_2$ molten salt mixture acts as a solvent for the nuclear fuel [45].

Isotope Separation (→ Isotopes, Natural). The nuclear powers from time to time offer on the world market relatively large quantities of lithium hydroxide in which the natural ^6Li content is reduced by ca. 50 %. Separation of the lithium isotopes is carried out either by an isotope-exchange reaction between lithium amalgam and a lithium compound in solution [46] or by continuous ion-exchange chromatography of lithium salt solutions on special ion-exchange resins [47]. The isotopes ^6Li and ^7Li can be purchased with purities of 95 – 96 % and up to 99.9 %, respectively.

Lithium isotopes are supplied by the Isotopes Sales Department of the Oak Ridge National Laboratory and the US Energy Research and Development Administration.

5.3. Lithium Alloys

Lithium readily forms alloys with other metals such as lead, copper, silver, magnesium, boron, and aluminum [48]. Lithium often completely changes the properties of metals to which it is added, e.g., the hardness of aluminum and lead, or the ductility of magnesium.

The lead alloy Bahnmetall, which was developed in Germany in 1918 and contained 0.04 % lithium, is no longer important; it was used as a bearing metal for railroad cars until ca. 1955. A eutectic mixture of lead with 17 wt % lithium is used in nuclear fusion technology as a breeder and cooling medium.

The first aluminum – lithium alloy, containing 0.1 % lithium, was developed in 1924 in Germany under the name Skleron. In the United States in 1957, the alloy X-2020 (4.5 % Cu, 1.3 % Li, remainder Al) was used in the wings and horizontal tail surfaces of naval aircraft. During the 1980s, aluminum – lithium alloys have been developed continuously, especially by Alcoa, Alcan, and Pechiney. The low density of these materials, compared with the usual aluminum alloys, is of great interest to the aircraft industry. They will probably be used in the foreseeable future for military and large passenger aircraft, despite competition from carbon fiber-reinforced plastics.

Lithium is the only element, apart from beryllium, that simultaneously reduces the density of aluminum and increases its modulus of elasticity; furthermore, formation of the Al_3Li phase gives improved strength. Modern aluminum – lithium alloys are now based almost without exception on the system Al – Li – Cu – Mg – Zr [49]. The alloy 8090, for example, has the composition 2.5 % lithium, 1.3 % copper, 0.8 % magnesium, 0.12 % zirconium, and the remainder aluminum. Such alloys have a density about 10 % lower, and a modulus of elasticity ca. 10 % higher, than aluminum.

Aluminum – lithium alloys with the composition 50 mol % aluminum and 50 mol % lithium are used in primary thermal batteries in which the electrolyte (a mixture of salts) is melted before the battery is used.

6. Lithium Compounds [50]

6.1. Inorganic Lithium Compounds and Lithium Salts

Lithium amide [*7782-89-0*], $LiNH_2$, M_r 22.96, ϱ 1.18 g/cm^3, *mp* 373 °C, is a colorless compound that forms cubic crystals. It is converted to lithium imide, Li_2NH, above 400 °C, with loss of ammonia.

Production. Lithium amide is produced industrially by heating lithium metal or lithium hydride in a stream of ammonia. It also forms slowly when lithium metal dissolves in liquid ammonia.

Uses. Lithium amide is now used in many reactions in preference to sodium amide for safety reasons, and because it often shows greater selectivity. It is used as a catalyst for Claisen and aldol condensations; as an alkylation catalyst for amines, nitriles, ketones, and monosubstituted alkynes; and as a polymerization catalyst for acrylonitrile and methyl methacrylate.

Lithium acetate [*546-89-4*], $C_2H_3O_2Li$, M_r 65.99, is available commercially in the anhydrous form or as the dihydrate. Both are colorless powders that dissolve readily in water to give slightly alkaline solutions.

Production. Lithium acetate is produced by reacting lithium carbonate or lithium hydroxide with acetic acid.

Uses. Lithium acetate is used to a limited extent in pharmaceuticals. The technical grade is used as a component of catalysts for the production of polyesters [poly(ethylene terephthalate)] or as an additive to improve their physical properties. It is also added to organopolysiloxanes and acrylic polymers to prevent adhesion of ice.

Lithium benzoate [*553-54-8*], $C_7H_5LiO_2$, M_r 128.06, is used in the production of some synthetic polymers (e.g., to harden polyepoxides) [51].

Lithium citrate [*916-16-4*], $C_6H_5Li_3O_7$, M_r 209.92, is commercially available mainly as a di- or tetrahydrate and is used in the treatment of certain types of manic-depressive illness.

Lithium bromide [*7550-35-8*], LiBr, M_r 86.85, *mp* 547 °C, ϱ 3.463 g/cm^3, is a colorless, highly hygroscopic salt (→ Bromine Compounds, Chap. 2.1.2.). The saturated aqueous solution at 20 °C contains 59.8 % lithium bromide and, at 100 °C, 69.9 %. The vapor pressure of water above aqueous 52 % lithium bromide is 0.53 kPa at 25 °C, 4.98 kPa at 65.6 °C, and 48.1 kPa at 121 °C [52], [53].

Production. Lithium bromide is produced by treating an aqueous suspension of lithium carbonate with hydrobromic acid or by reacting lithium hydroxide with bromine. The latter reaction also produces lithium hypobromite, which is reduced to lithium bromide with formic acid or ammonia.

Uses. A 50 – 60 % aqueous solution of lithium bromide is used in air conditioners operating by the absorption principle. Solid lithium bromide is a component of oxidation and hydroformylation catalysts. It is also used for deprotonation and dehydration of organic compounds having acidic protons. Mixtures of lithium bromide and lithium carbonate in dimethylformamide are used to carry out mild dehydrohalogenation reactions [36], [54]. Lithium bromide is also used for the purification of steroids and prostaglandins, with which it reversibly forms complexes [55].

Lithium carbonate [*554-13-2*], Li_2CO_3, M_r 73.89, ϱ 2.11 g/cm^3, crystallizes as colorless monoclinic prisms that melt with slight decomposition at 732 °C. The technical product is a fine white powder with a bulk density of ca. 0.8 kg/L. Carbon dioxide dissociation pressures over lithium carbonate are as follows:

Temperature, °C	610	620	740	840	1000	1200
Pressure, kPa	0.1	0.3	1.9	3.8	9.2	30.4

Lithium carbonate is only slightly soluble in water; its solubility decreases with increasing temperature.

Temperature, °C	0	20	40	60	80	100
Solubility, g/100 g water	1.54	1.33	1.17	1.01	0.85	0.72

Lithium carbonate dissolves in aqueous carbon dioxide to form lithium hydrogen carbonate. Concentrations of 70 g/L of lithium carbonate at 20 °C and 0.1 MPa carbon dioxide pressure can be reached. Lithium hydrogen carbonate cannot be isolated. Aqueous solutions decompose on heating to form lithium carbonate and carbon dioxide. This reaction is used in the Quebec process (see Section 4.2.3) and for the purification of technical-grade lithium carbonate. Lithium carbonate reacts with acids stronger than carbonic acid to form the corresponding salts. Reaction with calcium hydroxide yields lithium hydroxide.

Production. Lithium carbonate is produced entirely from minerals or naturally occurring brines (see Chap. 4). The solution of lithium salts initially obtained (lithium sulfate, lithium chloride) is reacted with sodium carbonate solution and lithium carbonate is precipitated, or lithium hydrogen carbonate is formed first and then decomposed thermally.

Uses. Lithium carbonate is the starting material for the industrial production of all other lithium compounds and is itself used in industry in large quantities. It is an additive in glass, enamels, and ceramics [56], [57] (as are many lithium ores, e.g., spodumene and petalite). The lithium oxide generated by lithium carbonate is a very active flux for silica. It is increasingly replacing fluorspar (CaF_2) for environmental reasons. Its function is generally based on reduction of melting point (particularly important for enamels), reduction of viscosity of molten glass (leading to higher throughput and smaller wall thickness), increased surface tension (giving improved reflectivity to enamels and glazes), and improved chemical resistance. Lithium aluminum silicates with compositions in the range $LiAlSiO_4$–$LiAlSi_4O_{10}$ are produced that have extremely low or negative thermal expansion coefficients. These compounds and high-lithium glasses are used in heat-resistant ovenware.

The addition of lithium carbonate to cement or concrete leads to quicker setting. Amounts of 1 – 5 % are used to control setting times [58]. In quick-setting tile adhesives, lithium carbonate leads to faster adhesion of the tiles.

Lithium carbonate is also used as an additive to molten salt baths for the electrolytic production of aluminum. It reacts with aluminum trifluoride in the melt to form lithium fluoride. The concentration used is 2 – 5 %, preferably 3.5 %, of lithium fluoride. The following effects are produced: reduction of melting point (by

10 – 25 °C), density, and viscosity of the melt and an increase in electrical conductivity of ca. 10 %. This results in a reduced operating temperature, reduction in fluorine emission (up to 50 %), and savings in energy (2 – 3 %) and anode carbon (ca. 2 %). It is possible to operate existing electrolytic cells at higher loading (ca. 10 %), thus increasing capacity without additional capital investment [59–62].

Specially prepared high-purity lithium carbonate (conforming to the national pharmacopoeias, e.g., U.S.P., DAB, E.P., B.P., J.P.) is used increasingly for the treatment of manic-depressive conditions [63–65]; daily doses of ca. 150 – 500 mg of lithium carbonate are used. The successful treatment of herpes, alcoholism, and cancer by lithium carbonate assisted chemotherapy has also been reported [65].

Mixtures of lithium and potassium carbonate are used as electrolytes in molten carbonate fuel cells (MCFCs). The MCFCs operating at high temperatures (600 – 650 °C) are especially suited for modern coal-fired baseload electric utility plants and natural gas units. The high-quality waste heat can be used for industrial cogeneration or internal reforming [66].

Lithium chloride [*7447-41-8*], LiCl, M_r 42.40, *mp* 614 °C, *bp* 1382 °C, ϱ 2.07 g/cm^3, crystallizes in a cubic system. Its solubility in water is as follows:

Temperature, °C	20	40	60	80	100
Solubility, g/100 g water	83.2	89.8	98.4	111.9	128.3

Its solubility in ethanol at 25 °C is 2.5 g/100 g, and in methanol, 17 g/100 g.

Production. The known methods of producing lithium chloride from lithium ores have not been adopted commercially. At present, it is produced exclusively by reaction of lithium carbonate and hydrochloric acid, with special steel or nickel equipment because of the extreme corrosivity of lithium chloride. Concentration of the solution (e.g., in a vacuum evaporator) causes lithium chloride to crystallize. It is then separated from the mother liquor, dried, and packed in moisture-proof containers.

Uses. Lithium chloride is extremely hygroscopic; thus, like lithium bromide, it is used for drying gases and in air conditioners of the absorber type. Because molten lithium chloride readily dissolves metallic oxides, it is used as a component of fluxes and dipping baths for welding and brazing aluminum and light metal alloys. Most of the lithium chloride produced is used as starting material for the production of lithium metal by molten salt electrolysis.

Lithium cobalt(III) oxide [*12190-79-3*], $LiCoO_2$, M_r 97.87, ϱ 5.16 g/cm^3, can be used as a cathodic material in secondary lithium ion batteries [67]. The dark gray active material with a theoretical specific capacity of 137 mAh/g is prepared by sintering Li_2CO_3 with $CoCO_3$ at 900 °C. The first cells, of cylindrical design, have been commercialized by Sony [68].

Lithium hypochlorite [*13840-33-0*], LiOCl, M_r 58.39, is produced by reacting either chlorine with lithium hydroxide solution, or calcium hypochlorite (bleaching powder) with lithium sulfate. Unlike sodium hypochlorite, it may be obtained in solid form (e.g., by spray drying), although overheating must be avoided because oxygen is liberated above 100 °C.

Lithium hypochlorite is increasingly used as a disinfectant and bleaching agent for swimming pools and in food preparation areas (e.g., dairies and butcher shops). Its action in swimming pools is analogous to that of the cheaper calcium hypochlorite, but it does not increase the hardness of water and is completely soluble.

Lithium perchlorate [*7791-03-9*], $LiClO_4$, M_r 106.39, ϱ 2.428 g/cm^3, crystallizes from aqueous solution as the trihydrate (*mp* 95.1 °C), which is converted into the anhydrous compound at 130 – 150 °C via the dihydrate stage. Anhydrous lithium perchlorate melts at 236 °C and is stable up to 400 °C. Like lithium chloride, it is readily soluble in water (56.3 g/100 g water at 20 °C; 196.9 g/100 g water at 95.1 °C).

Lithium perchlorate is produced by reaction of lithium hydroxide or lithium carbonate with perchloric acid. It is used as a conducting salt in electrolytes for lithium batteries. Furthermore, solutions of $LiClO_4$ in polar aprotic solvents such as diethyl ether dramatically accelerate Diels – Alder reactions due to the Lewis acid character of the lithium ion [69].

Lithium tetrafluoroborate [*14283-07-9*], $LiBF_4$, M_r 93.74, ϱ 2.04 g/cm^3, and *lithium hexafluoroarsenate* [*29935-40-1*], $LiAsF_6$, M_r 195.85, are of potential interest as conducting salts for the electrolytes in lithium batteries.

Lithium hexafluorophosphate [*21324-35-3*], $LiPF_6$, M_r 151.90, 130 °C decomp., is highly soluble in aprotic solvents such as ethers and carbonates and decomposes slowly in water with formation of HF. Owing to its high solubility and electrochemical stability, it is the preferred conducting salt for secondary lithium batteries [70].

Lithium chromate [*14307-35-8*], Li_2CrO_4, M_r 129.89, is used as a corrosion inhibitor in air conditioners that contain lithium chloride or bromide solutions. It is made from lithium hydroxide and chromic acid.

Lithium fluoride [*7789-24-4*], LiF, M_r 25.94, ϱ 2.6 g/cm^3, *mp* 870 °C, is used as a flux for enamel and glass, and for brazing and welding light metals.

Lithium hydride [*7580-67-8*], LiH, M_r 7.95, ϱ 0.82 g/cm^3 (→ Hydrides, Chap. 2.2.), is used increasingly as a selective reducing agent for polar groups (e.g., esters, amides, nitriles) in organic compounds, although the complex hydrides are of much greater importance: e.g., lithium tetrahydridoaluminate [*16853-85-3*], lithium aluminum hydride, $LiAlH_4$ (see → Hydrides, Chap. 3.8., Chap. 3.9.), and lithium tetrahydridoborate [*16949-15-8*], lithium borohydride, $LiBH_4$ (see → Hydrides, Chap. 3.2., Chap. 3.3.). It can be used for the hydrogenolysis of Si – Cl bonds, for example, the production of monosilane (SiH_4) by the Sundermeyer process [71] and SiH-containing polymers [72].

Lithium hydroxide monohydrate [*1310-66-3*], $LiOH \cdot H_2O$, M_r 41.96, ϱ 1.83 g/cm^3, loses water on heating to form anhydrous *lithium hydroxide* [*1310-65-2*], LiOH, M_r 23.95, ϱ 1.46 g/cm^3, *mp* 462 °C.

Production. The production of lithium hydroxide monohydrate by digestion of spodumene with lime (see Section 4.2.2) has been replaced for economic reasons by the reaction between lithium carbonate and calcium hydroxide:

$$Li_2CO_3 + Ca(OH)_2 \rightarrow CaCO_3 + 2\,LiOH$$

This yields solutions containing up to 3.5 % lithium hydroxide. Attempts to work at higher concentration lead to loss of lithium carbonate in the calcium carbonate residue, because the solubility of lithium carbonate is reduced by the presence of lithium hydroxide. The reaction between these sparingly soluble starting materials requires intensive agitation and gives yields just over 95 % after many hours. The insoluble residue (mainly calcium carbonate) is removed, and lithium hydroxide monohydrate is crystallized from the mother liquor by evaporation, under vacuum if required. It is then separated and dried at 130 – 140 °C. Further heating to 150 – 180 °C produces calcined lithium hydroxide (LiOH). The removal of water can be accelerated by use of vacuum.

Other production processes that have been proposed include the reaction of lithium sulfate solution with barium hydroxide and the reaction of lithium carbonate with silicic acid in the molten state followed by hydrolysis of the lithium silicate formed. These processes are, however, of little more than academic interest.

The solubility of lithium hydroxide monohydrate in water is as follows:

Temperature, °C	20	40	60	80	100
Solubility, g/100 g water	21.6	22.0	23.1	25.6	29.6

Uses. Lithium hydroxide is used mainly for the production of lubricating greases. Over 50 % of the greases used today in motor vehicles, aircraft, and heavy machinery contain lithium stearates, mainly lithium 12-hydroxystearate. These greases have good viscosity properties up to 200 °C and are practically insoluble in water (→ Lubricants and Related Products).

Lithium hydroxide is also used in the production of dyes [73]. It has attained some importance as an additive to nickel – iron accumulators and, in its calcined form (98 % LiOH), as an absorbent for carbon dioxide in breathing apparatus, especially for crews of submarines and space vehicles. One gram of anhydrous lithium hydroxide absorbs 450 mL of carbon dioxide, so

that ca. 750 g of lithium hydroxide is sufficient to absorb the carbon dioxide generated by one person in one day.

Lithium hydroxide is also used in the building industry. As an additive to the mix water, it helps prevent the alkali – silica reaction, which is a cause of deterioration (cracking, disintegration) in portland cement concrete structures [74].

Lithium iodide [*10377-51-2*], LiI, M_r 133.29, ϱ 4.0 g/cm^3, is a yellowish crystalline powder that is prepared by reaction of lithium hydride with iodine in diethyl ether. It is used in organic syntheses to accelerate the aldol condensation of ketones with enolizable or nonenolizable aldehydes [75] and to cleave methyl esters and oxiranes [76]. Lithium iodide is also of interest for the manufacture of rapid solid ionic conductors [77]. *Lithium iodide trihydrate* [*7790-22-9*], $LiI \cdot 3\,H_2O$, M_r 187.34, ϱ 2.5 g/cm^3, is prepared by reaction of lithium hydroxide with hydriodic acid.

Lithium manganese oxide [*12057-17-9*], $LiMn_2O_4$, M_r 180.82, ϱ 4.28 g/cm^3 is a ternary spinel-type double oxide commonly prepared by sintering manganese oxide (MnO_2 or Mn_2O_3) with Li_2CO_3 at high temperatures. It is a potential cheap candidate for cathodes in high-power secondary lithium ion batteries. Practical specific capacities are in the range 90 – 120 mA h/g. [67], [78].

Lithium nitrate [*7790-69-4*], $LiNO_3$, M_r 68.95, ϱ 2.37 g/cm^3, is used, with other salts, to produce eutectic molten salt baths for vulcanization. A salt mixture consisting of lithium and potassium nitrates is available under the trade name Sabalith [79]. Unlike conventional nitrite-containing salt mixtures, Sabalith is not classified as a toxic substance, and it helps to reduce emissions of nitrous gases and carcinogenic nitrosamines.

Lithium nitride [*26134-62-3*], Li_3N, M_r 34.82, ϱ 1.38 g/cm^3, could attain industrial importance as a solid ionic conductor. It is also used together with magnesium boride and hexagonal boron nitride to prepare cubic boron nitride [80].

Lithium oxide [*12057-24-8*], Li_2O, M_r 29.89, ϱ 2.01 g/cm^3, is formed when lithium hydroxide is heated to 800 °C in vacuum. It is also obtained by the reaction of lithium metal with oxygen, which proceeds without formation of lithium peroxide.

Lithium peroxide [*12031-80-0*], Li_2O_2, M_r 45.88, ϱ 2.297 g/cm^3, is formed by reaction of lithium hydroxide solution with hydrogen peroxide; it can be isolated by careful evaporation in vacuum. It has some importance as a hardener for the formation of special polymers and for air purification in sealed spaces (submarines, breathing apparatus, space travel). One gram of Li_2O_2 produces 0.348 g of oxygen and absorbs 0.96 g of carbon dioxide.

Lithium double oxides are also known as lithium ceramics, and are used to a small extent in the ceramic and enamel industries. They include lithium aluminate [*12003-67-7*], $LiAlO_2$; lithium borate [*13453-69-5*], $LiBO_2$; lithium manganite [*12163-00-7*], Li_2MnO_3; lithium molybdate [*13568-40-6*], Li_2MoO_4; lithium silicate [*10102-24-6*], Li_2SiO_3; lithium titanate [*12031-82-2*], Li_2TiO_3; and lithium zirconate [*12031-83-3*], Li_2ZrO_3.

Lithium aluminate is of some importance as a construction material in fusion reactors [81] and fuel cells [82].

Lithium tetraborate [*12007-60-2*], $Li_2B_4O_7$, M_r 169.12, either pure or doped with beryllium or strontium as an internal standard, is used as a flux in X-ray spectroscopy and optical emission spectral analysis. It also has limited application in the manufacture of special lithium-containing, high-performance lubricating greases.

Lithium phosphate [*10377-52-3*], Li_3PO_4, M_r 115.79, ϱ 2.537 g/cm^3, is sparingly soluble in water (ca. 0.04 %). It is prepared by precipitation from solutions of lithium salts by addition of phosphates or phosphoric acid. It has some importance as a component of protective optical heat filters, in the corrosion protection of vessels for polyester production, and in the catalytic production of alkyl alcohol.

Lithium sulfate [*10377-48-7*], Li_2SO_4, M_r 109.94, ϱ 2.221 g/cm^3, *mp* 860 °C, is a colorless salt produced from lithium carbonate and sulfuric acid. It is used in the anhydrous or hydrated

form for the treatment of manic-depressive conditions [40]. It has also been considered for the production of quick-setting cements.

Other Lithium Salts. Lithium azide, borate, cyanide, molybdate, and salicylate, are prepared by reaction of lithium carbonate or lithium hydroxide with the corresponding acids. Lithium niobate and lithium tantalate single crystals are used as ferroelectric materials and also have acousto-optical properties that can be utilized in optical data transmission systems.

6.2. Organolithium Compounds, Lithium Alkylamides, and Lithium Alkoxides

The direct synthesis of organolithium compounds from metallic lithium and an alkyl or aryl halide was first described in 1930 [83]. The method is now used on an industrial scale. The compounds are usually prepared as solutions in hydrocarbons under a nitrogen atmosphere at 25 – 70 °C because they are thermally unstable. The presence of 0.5 – 2 % sodium accelerates the reaction. These precautions reduce the loss of yield by reactions of the type

$$RCl + RLi \rightarrow R_2 + LiCl$$

On completion of the reaction, excess lithium metal and the lithium chloride formed are removed by centrifugation, and the solution is clarified by filtration.

The lithium alkyls are oligomeric organometallic compounds with mainly covalent lithium – carbon bonds. They are oxidized by atmospheric oxygen to form the corresponding alkoxides. Compounds that contain active hydrogen, such as water, alcohols, phenols, acids, and primary and secondary amines, decompose lithium alkyls. For example, the reaction of water with butyllithium gives butane and lithium hydroxide, whereby the presence of oxygen may lead to spontaneous ignition. Commercially available 5 – 25 % solutions of butyllithium in hydrocarbons do not usually ignite spontaneously if they come in contact with air. However, if the contact involves a large surface area and higher concentration, spontaneous ignition is possible.

Because of their high reactivity, lithium alkyls are always stored and handled under nitrogen or noble gases. Carbon dioxide must not be used because it reacts exothermically with lithium alkyls. Suitable fire-extinguishing media are powdered limestone, powders containing sodium chloride, or sodium bicarbonate. Spillage of lithium alkyl solutions can be covered with powdered limestone and left to decompose in the air [84].

The most important lithium alkyls are *n*-butyllithium and *sec*-butyllithium.

***n*-Butyllithium** [*109-72-8*], *n*-C_4H_9Li, M_r 64.06, ϱ 0.765 g/cm^3, *mp* −76 °C, forms hexamers in the solid state and in hydrocarbon solutions. Neat butyllithium is a spontaneously flammable colorless and viscous liquid that can be distilled at 0.001 Pa and 80 – 90 °C. The commercially available neat *n*-butyllithium typically has a concentration of 90 % and a yellow color. *n*-Butyllithium is normally used as solution in hydrocarbon solvents. Those commercially available are:

Tech. *n*-hexane	15, 23, 90 %
Heptane	25 %
Cyclohexane	20, 24 %
Toluene	20 %

The hydrocarbon solutions normally do not ignite on contact with air, but as ignition cannot be ruled out and to avoid loss of active material, they must be handled under inert gas. The commercial solutions are thermally fairly stable: a 15 % (1.6 mol/L) solution in hexane decomposes at the rate of 0.06 and 0.25 % of the active material per month at 20 and 30 °C, respectively, forming 1-butene and lithium hydride [84]. In ethers, which are common solvents for reactions with organolithium compounds, butyllithium is much less stable. Donor solvents break down the butyllithium hexamers into more reactive tetra-, di-, and monomeric species. These species are more nucleophilic and tend to cleave polar bonds like the ether C – O bond. The half-life period $t_{\frac{1}{2}}$ of butyllithium at 25 °C in diethyl ether and 1,2-dimethoxyethane is 6 d and 10 min respectively. In THF $t_{\frac{1}{2}}$ is 23.5 h at 0 °C and 5 d at – 30 °C. Therefore low temperatures are normally required when ethers are used as cosolvents [85].

Production. Like all other commercially available organolithium compounds, *n*-butyllithium is manufactured from lithium metal and the corresponding organic halide. In principle, many aprotic solvents, including functionalized compounds like diethyl ether and unfunctionalized hydrocarbons, may be used. As alkyllithium compounds $C_nH_{2n+1}Li$ with $n \geq 4$ are miscible with hydrocarbons and these solutions are much more stable than those containing ethers or amines, commercial butyllithium synthesis is exclusively carried out in hydrocarbons at 30 to 60 °C.

$$C_4H_9Cl + 2\,Li \longrightarrow C_4H_9Li + LiCl$$

A satisfactory yield and a high rate of reaction are obtained only if the lithium metal is used as a fine dispersion containing about 0.5 – 2 % sodium. After completion of the reaction, the insoluble species (LiCl and excess lithium metal) are removed by filtration. An economic production requires the reprocessing of the insolubles to lithium metal (by electrolysis). The butyllithium is obtained as clear solutions containing small amounts of C_8 hydrocarbons.

Uses. Butyllithium is mainly used as initiator in the anionic polymerization of conjugated dienes [85–87]. In the first, stoichiometric step organolithium compounds carbolithiate the diene to give allyllithium-type species, which can undergo further insertion of excess monomeric diene to form so-called living polymers.

$$C_4H_9\text{–}Li + H_2C\text{=}CH\text{–}CR\text{=}CH_2 \longrightarrow C_4H_9\text{–}CH_2\text{-}CH\text{=}CR\text{-}CH_2\text{–}Li$$

The diene insertion can result in different stereoisomers: 1,4-addition gives either *cis*- or *trans*-polydienes, whereas 1,2-addition leads to *iso*- and *syndio*-polydienes containing vinyl groups. The resulting polymers exhibit a narrow molecular mass distribution, which can be adjusted by variation of the butyllithium concentration.

Polybutadiene rubber (BR) can be produced by Ti-, Co-, Ni-, or Nd-based Ziegler – Natta catalysts or by solution polymerisation with butyllithium (see also → Rubber, 3. Synthetic, Chap. 3.1.2.). In contrast to the high stereospecifity of the Ziegler catalysts (> 90 % *cis*-1,4 content) the lithium-BR microstructure is almost random (typically 30 – 50 % *cis*-1,4, 50 – 55 % *trans*-1,4 and 10 % 1,2-*iso* and *syndio*). Polar additives like ethers promote the formation of 1,2-addition products. Lithium-BR has an amorphous structure with very low glass temperatures, which is desired for typical rubber applications. Polyfunctional coupling agents such as $SiCl_4$ produce so-called star-polymers with excellent processability and abrasion resistance, used, for example, in the production of tires.

In contrast to butadiene, isoprene is polymerized by butyllithium with high stereoselectivity (90 – 92 % 1,4-*cis*) approaching that of the Ti-based Ziegler – Natta catalysts (98 % 1,4-*cis*). Synthetic polyisoprene rubbers (see also → Rubber, 3. Synthetic, Chap. 3.1.2.) are used in the same fields as natural rubber, which has exclusively a 1,4-*cis* microstructure.

Living polymers allow the preparation of block copolymers by sequential addition of different monomers. The most important products are styrene – butadiene rubbers (SBR), which are produced either in emulsion by radical-forming compounds or in solution, generally by using lithium-based initiators such as butyllithium. Depending on the reaction conditions, lithium-based SBR can either form random or block copolymers. The living polymers are coupled by terminating agents like 1,2-dibromoethane or $SiCl_4$ to give SBS three-block copolymers, which can be deformed at higher temperatures (thermoplastic elastomers). Thermoplastic elastomers containing 20 – 40 % styrene have a high tensile strength and are used for shoe soles, adhesion agents and as modifiers for bitumen and other thermoplastic elastomers like polystyrene. SBS-block copolymers with a higher content of styrene (about 70 %) are glass clear and have outstanding shock and impact strength. They are used for packaging foodstuffs or medical articles and for household articles like drinking beakers.

n-Butyllithium is also used in organic synthesis [88], [89] as a strong base for the deprotonation of hydrocarbons that are more acidic than butane, amines, and alcohols. An example is the manufacture of lithium diisopropylamide, which is an important reagent, mainly for the production of carbonyl or ester enolates:

$$C_4H_9Li + HN(C_3H_7)_2 \longrightarrow LiN(C_3H_7)_2 + C_4H_{10}$$

In aromatic chemistry metalation of the C – H bond is supported by heteroatomic sub-

stituents, and lithiation takes place in the *ortho* position (directed *ortho* metalation, DOM) [90]. As an example, 2,2-difluoro-1,3-benzodioxole is lithiated in the *ortho* position by *n*-butyllithium/TMEDA. The subsequent reaction with an electrophile yields an intermediate for the synthesis of the pyrrole fungicide Fludioxinil (Ciba Geigy) [91].

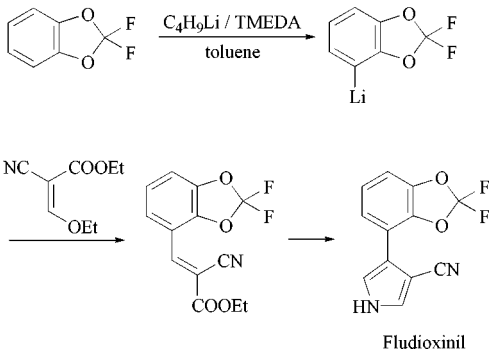

Butyllithium can be used for the preparation of other organolithium compounds by halogen/lithium exchange and for direct addition reactions to carbon – carbon, carbon – oxygen and carbon – nitrogen bonds. The most important reactions are summerized in Scheme 6 [89], [92], [93].

***sec*-Butyllithium** [*598-30-1*], *sec*-C_4H_9Li, M_r 64.06, is a spontaneously flammable, colorless, viscous liquid that forms tetramers in hydrocarbon solvents. It is commercially available only as 12 % solution in cyclohexane and methylcyclohexane. *sec*-Butyllithium is more nucleophilic and decomposes more readily than *n*-butyllithium. A 12 % solution in cyclohexane at 20 °C undergoes thermal decomposition to the extent of 1.4 % of the active material per month. The thermal decomposition yields butene gas and LiH; the latter is insoluble and responsible for the increasing turbidity of sec-butyllithium solutions on prolonged storage. Therefore solutions of *sec*-butyllithium must be kept cool during storage and transportation.

Uses. *sec*-Butyllithium is used as initiator for the solution polymerization of polydienes, especially styrene – butadiene block copolymers. In organic synthesis it is used when its higher reactivity compared with *n*-butyllithum is desirable.

***tert*-Butyllithium** [*594-19-4*], $(CH_3)_3CLi$, M_r 64.06, is a white crystalline substance which is pyrophoric in air and miscible with hydrocarbon solvents. The pure compound does not melt, but decompose above 140 °C. It can be sublimed at 100 Pa and 70 – 80 °C. In the solid state and in hydrocarbon solutions it forms tetramers. In hydrocarbon solution it is the most stable of the butyllithium isomers: at 20 °C a 15 % solution in pentane has a decomposition rate of only 0.015 % active material per month. Nevertheless it is the most reactive of the C_4H_9Li isomers, combining the highest basicity with reduced nucleophilic properties owing to steric effects. Ethers are more rapidly cleaved by *tert*-butyllithium than by its isomers.

tert-Butyllithium is produced in low-boiling hydrocarbons in a way similiar to *n*-butyllithium, but the yields obtained are considerably lower [92], [94].

The product is commercially available as 15 % solution in technical hexane and pentane.

tert-Butyllithium is very useful for introducing *tert*-butyl groups into organic and organometallic compounds. As example it can be applied for the synthesis of sterically hindered silane protecting groups:

$$t\mathrm{BuLi} + \mathrm{Me_2SiCl_2} \longrightarrow \mathrm{Me_2Si}t\mathrm{BuCl} + \mathrm{LiCl}$$

Because of the sterically demanding nature of the *tert*-butyl group, reactions with *tert*-butyllithium (e.g., deprotonations) can often be carried out with very high selectivity. *tert*-Butyllithium can also be used as convenient, homogeneous source of LiH. For example, trialkylborohydrides, which are inert towards commercial LiH, are converted to trialkylborohydrides with excellent yields [95]:

$$\mathrm{R_3B} + t\mathrm{BuLi} \longrightarrow \mathrm{Li[HBR_3]} + \mathrm{C_4H_8}$$

***n*-Hexyllithium** [*21369-64-2*], $C_6H_{13}Li$, M_r 92.11, ϱ 0.766 g/cm^3 is a pyrophoric colorless and viscous liquid in the neat state. It is commercially available as a 33 % solution in technical-grade hexane. Its reactivity resembles that of *n*-butyllithium and it can thus principally be used for the same applications. An advantage is that it does not form gaseous byproducts in deprotonation reactions and on hydrolysis, but liquid hexane instead.

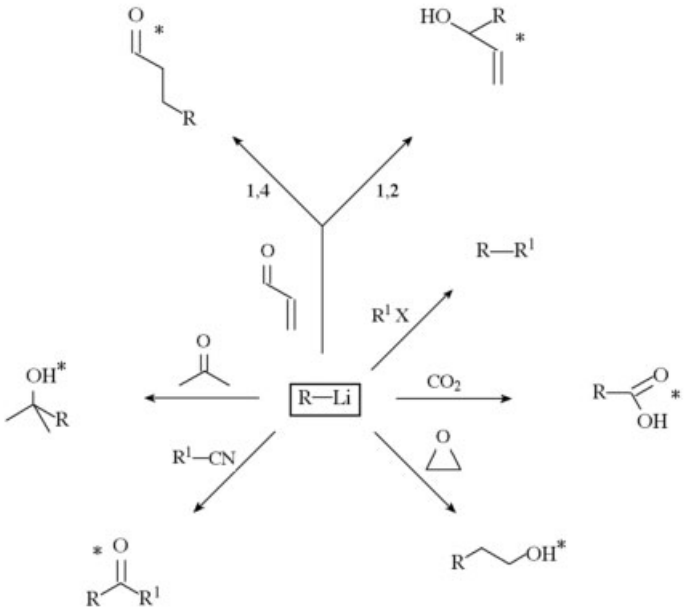

Figure 6. Reactions of *n*-butyllithium (R = Bu; asterisk denotes products isolated after hydrolysis)

Methyllithium [*917-54-4*], CH_3Li, M_r 21.98, forms colorless cubic crystals. It is virtually insoluble in aliphatic and aromatic hydrocarbons. In ethereal solvents it is more stable than butyllithium; thus it is produced and is commercially available in ethers or ether-containing solvent mixtures. Commercial products are 5 – 6 % solutions in diethyl ether and ca. 8 % solutions in diethoxymethane [96]. These solutions are produced both with low or stoichiometric lithium halide (LiBr, LiI) content. THF-containing products are not stable on prolonged storage. Methyllithium forms tetrameric aggregates in THF and diethyl ether.

Uses. Methyllithium is used for the introduction of the methyl group into unsaturated substrates. In order to modify the selectivity, the methyl group is often transferred to a transition metal, e.g., copper [97]. These bimetallic carbanionic bases perform conjugate 1,4-additions to α,β-unsaturated carbonyls. As a typical example the production of an intermediate in the synthesis of the androgen Mesterolone is shown [98]:

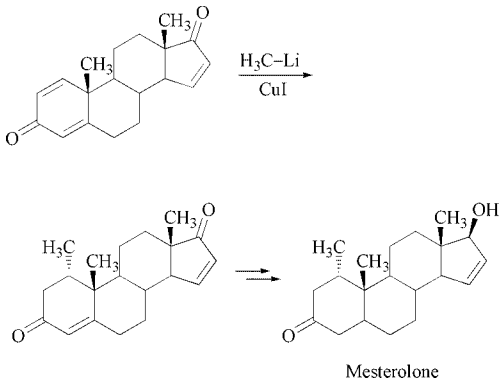

Phenyllithium [*591-51-5*], C_6H_5Li, M_r 84.05, forms colorless monoclinic crystals. Commercial products are a 20 % solution in a mixture of about 70 % cyclohexane and 30 % diethyl ether and a 17 % solution in dibutyl ether. The solutions are brown in color and must be stored at low temperatures to avoid thermal decomposition. Phenyllithium is used as a mild metalating agent in organic synthesis and can be used instead of Grignard reagents to introduce phenyl groups into organic compounds.

Lithium *tert*-butoxide [*1907-33-1*], $LiOC(CH_3)_3$, M_r 80.06, ϱ 0.89 g/cm^3, *mp* 283 °C is a colorless compound with good solubility in many organic solvents (e.g. 16 % in hexane). It is commercially available as min. 98 % powder and as a 20 % solution in THF. In organic synthesis it serves as a strong base. As one of the few volatile lithium compounds, it sublimes without decomposition at 110 °C in vacuum, and is therefore a suitable raw material for lithium doping in CVD processes.

Lithium methoxide [*865-34-9*], $LiOCH_3$, M_r 37.98, is a white solid that is soluble in methanol but virtually insoluble in aprotic organic solvents. It is commercially available as a 99 % fine crystalline solid and as 10 % solution in methanol. It can be used as base, especially as catalyst for transesterification reactions.

Lithium diisopropylamide (LDA) [*4111-54-0*], $LiN(CH(CH_3)_2)_2$, M_r 107.14 is a white pyrophoric powder. Freshly prepared, it is soluble in hydrocarbons (in hexane about 10 %), but it tends to precipitate irreversibly from solution as a polymer on heating or prolonged storage. In ethers the solubility is much higher, but with the exception of tetrahydropyran [99], LDA is decomposed at a rate depending on the ether, concentration, and temperature.

LDA can be conveniently prepared from butyllithium and diisopropylamine or purchased as a 2 mol/L (25 %) solution in THF and a mixture of various hydrocarbons. Although LDA is unstable in pure THF (at room temperature a 25 % solution loses about 1 % of its activity per day), the commercially available compositions containing only a limited amount of THF are satisfactory stable for technical applications.

Uses. LDA is a sterically hindered nonnucleophilic strong base used for selective deprotonations, especially for the production of kinetic enolates (i.e., the thermodynamically less favored isomer) and (hetero-)aromatic carbanions. In the production of the serum lipid regulating agent Gemfibrozil one key intermediate is formed by quenching an ester enolate with 1-bromo-3-chloropropane [100]:

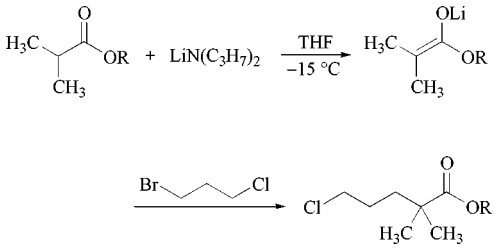

Lithium bis(trimethylsilyl)amide [*4039-32-1*], lithium hexamethyldisilazide (LHMDS), $LiN(Si(CH_3)_3)_2$, M_r 167.33, *mp* 106 °C, *bp* 115 °C/100 Pa, forms colorless crystals which are very soluble in many aprotic solvents like toluene, hexane, THF and *tert*-butyl methyl ether. This base is more stable, less basic, and less air sensitive than lithium diisopropylamide. It is commercially available as a 20 % solution in THF and as a 30 % solution in hexane and other solvents. Like LDA it is used as nonnucleophilic base to generate kinetic ketone and ester enolates. It is considerably more selective than LDA and undesired reductions (e.g., of nonenolizable ketones observed in the case of LDA) can be avoided by using LHMDS.

7. Quality Specifications and Analysis

Lithium metal [*7439-93-2*] of standard grade has a purity in excess of 99 % and contains ca. 0.5 % sodium. Battery grade lithium contains less than 200 ppm of sodium. Standard quality lithium is usually produced in the form of bricks or cylinders 0.1 – 1 kg in weight, rods with a diameter of 10 – 20 mm, or granules with a diameter of ca. 2.5 mm. Battery grade lithium is supplied as foil with a thickness down to 0.03 mm.

Lithium is detected qualitatively by its characteristic red flame coloration; this is also used for quantitative determination by flame emission spectroscopy, in which the emission line at 670.8 nm is measured. Analysis is carried out by using diffraction or refraction spectrometers. In the presence of interfering ions, especially other alkali metals, a calibration curve should be used to obtain adequate accuracy (± 1 %) [101].

Lithium can also be determined by gravimetric, titrimetric, and colorimetric methods,

after precipitation as the periodate, phosphate, zinc uranyl acetate, aluminate, or fluoride [102]. Lithium sulfate can be used as a primary standard. Lithium hydroxide and organolithium compounds can be determined easily by acidimetric methods. Indirect methods for the determination of lithium salts are also used, e.g., separation of the lithium by ion exchange followed by titration of the acid produced and, for lithium salts of organic acids, by titration with perchloric acid in nonaqueous media [103].

Analysis of lithium hydrides can be carried out by measuring the volume of gas evolved on hydrolysis. For low concentrations of lithium, ion-exchange chromatography is most suitable [104].

8. Toxicology and Occupational Health

The lithium ion is considerably more toxic than the sodium ion; for example, 5 g of lithium chloride can cause fatal poisoning. On the other hand, lithium salts have been widely used for over 30 years for the clinical treatment of manic-depressive illnesses. Therapeutic doses of 170 – 280 mg Li per day, mainly in the form of lithium carbonate, are administered over long periods. Lithium is an environmental trace element. Estimated background exposure is 2 mg/d and is believed to have beneficial effects [105].

In Europe and the United States, there are no special regulations concerning lithium in wastewaters. Lithium is included as a part of the allowed salt emissions.

No industrial disease caused by ingestion of lithium compounds has been reported thus far. Lithium hydroxide has caustic properties similar to those of other alkali-metal hydroxides; the caustic action of lithium hydride and lithium aluminum hydride is due to the lithium hydroxide formed on hydrolysis. Details of the toxicity and therapeutic possibilities of lithium are given in [63], [65]. The special safety precautions necessary when handling the often pyrophoric organolithium compounds have already been referred to (see Section 6.2).

9. Economic Aspects

The information available concerning world production of lithium ores is incomplete [106]. Including those amounts used for lithium carbonate production, it can be estimated at 6×10^5 t/a.

In 1992 the estimated consumption of lithium minerals in the Western world for direct use (e.g., in the glass, ceramics, and metallurgical industries) was 85.5×10^3 t/a, corresponding to about 12×10^3 t of lithium carbonate equivalents. A further $(150-200)\times10^3$ t was used in the production of lithium chemicals.

The bulk prices (FOT Amsterdam) in May 1995, in U.S. dollars per ton, were

Petalite (4.2 % Li_2O)	147
Spodumene concentrate (> 7.5 % Li_2O)	385
Glass-grade spodumene (5 % Li_2O)	175

The consumption of lithium compounds (calculated as lithium carbonate) in the Western world was ca. 3.2×10^4 t in 1989, with an annual rate of increase of ca. 3 – 4 %. Approximately 40 % of the lithium production was consumed by the United States, where, together with Chile, the main ore and brine deposits are to be found. The capacity available at present together with the projected expansion will ensure that the market remains satisfied. About 25 % of the lithium produced is used in the lubrication industry, 25 % in the aluminum industry, 25 % in the enamel, glass, and ceramic industries, and the rest for miscellaneous uses such as chemicals, pharmaceuticals, catalysts for the rubber industry, additives for welding materials, and absorbers in air conditioners.

Bulk prices in the United States in November 1996, in U.S. dollars per kg, were:

Lithium carbonate	4.25
Lithium hydroxide monohydrate	5.73
Metallic lithium	89.51
Lithium chloride	8.91
n-Butyllithium	47.18

10. References

1. M. F. C. Ladd, *Theor. Chim. Acta* **12** (1968) 333 – 336. W. Seelmann-Eggebert et al.: *Karlsruher Nuklidkarte*, 5th ed., 1981.
2. E. J. Cairns, F. A. Cafasso, V. A. Maroni: *Chemistry of Fusion Technology,* Plenum Publishing Corp., New York 1972, pp. 91 – 160.

3. *Landolt – Börnstein,* 6th ed., **IV, Part 2,** pp. 326 – 330.
4. *Gmelin,* System no. 20, Lithium, p. 150.
5. H. Roy, E. Geiss, *Microbeam Analysis 1981,* San Francisco Press, San Francisco 1981, pp. 303 – 308.
6. Xiang Jun: "Study of Zha Buyesaline/Lake Tibet," *Scientific papers on geology for the 27th Internat. Geol. Congr.* **4** (1985) 173 – 184, Geol. Publ. House, Beijing.
7. I. A. Kunasz, "Lithium Raw Materials," in S. J. Lefond (ed.): *Industrial Minerals and Rocks,* 5th ed., American Institute of Mining, Metallurgical and Petroleum Engineers, New York 1983, vol. 2, pp. 869 – 880.
8. D. London, D. M. Burt: *"Lithium Minerals in Pegmatites," Short Course Handb. Mineral. Assoc. Can.* **8** (1982) 99 – 133.
9. R. K. Evans: "Lithium Reserves and Resources," *Energy* **3** (1978) 379 – 385.
10. I. A. Kunasz: "Lithium how much?" *Foote prints* **44** (1980) no. 1, 23 – 27.
11. K. R. Suttill, *Eng. Min. J.* **188** (1987) 40 – 43.
12. Minister Energy, Mines, Resources Canada, "Lithium, an Imported Mineral Commodity," *Mineral Bulletin MR 212,* Canadian Government Publishing Centre, Ottawa 1986, p. 24.
13. G. Aubert: "Le Lithium," *Editions Bureau recherches géologiques et minières, Paris* **XV** (1963) no. 2, 23 – 93.
14. M. L. Deshpande: "Lithium Resources in India," *Indian Miner.* **32** (1978) no. 4, 41 – 47.
15. "Minerex explores for Lithium", *Ind. Miner.* **1984,** no. 8, 9.
16. D. I. Bleiwas, J. S. Coffman: "Lithium Availability–Market Economy Countries," Bureau of Mines, U.S. Dept. of Interior, Inf. Circular 9102, Washington D.C. 1986.
17. M. Coad: "Lithium Production in Chile's Salar de Atacama," *Ind. Miner.,* Oct. 1984, 27 – 33.
18. J. R. Davis, J. D. Vine: "Stratigraphic and Tectonic Setting of the Lithium Brine Field, Clayton Valley, Nev.," *Basin and Range Symp., Oct. 7 – 11, 1979,* Rocky Mountain Assoc. Geol. 1979, pp. 421 – 430.
19. "China Finds Lithium," *Min. Mag.* **155** (1986) no. 4, 303.
20. Qian Ziqiang, Xuan Zhiqiang: "Borate Minerals in Salt Lake Deposits at Chaidamn Basin, China," *Sixth Intern. Sympos. on Salt,* Salt Institute, vol. I, 1983, 185 – 194.
21. "Chile Aims at Top Producer Slot," *Metal Bull. Monthly,* Jan. 1988, p. 61.
22. D. Kaplan, *Isr. J. Chem.* **1** (1963) 115 – 120.
23. M. Steinberg, Vi-Duong Dang: "Preliminary Design and Analysis of a Process for the Extraction of Lithium from Seawater," *Lithium Resources and Requirements by the Year 2000, Geological Survey Professional Paper,* no. 1005, US Government Print Office, Washington 1976, pp. 79 – 88.
24. A. G. Collins: "Lithium Abundances in Oilfield Waters," in [23]pp. 116 – 123.
25. A. E. Schreck, *U.S. Bur. Mines, Inf. Civ.* 1961, no. 8053, 36 – 41.
26. G. G. Gabra, A. Torma: "La Metallurgie Extractive du Lithium," *L'ingénieur* (Montréal, école polytechnique), March, April 1977, 10 – 15.
27. Metallgesellschaft AG, *Verfahren zur Gewinnung von Lithiumcarbonat,* DE-OS 36 22 105 A 1, 1988 (P. Brödemann, G. Krüger, H. Heng).
28. G. von Girsewald: "Lithium," *Mitt. d. Metallgesellschaft* **2** (1929) 13 – 20.
29. J. Tixhon, *Ind. Chim. Belge Suppl.* **1** (1959) 752.
30. M. Archambault, US 3 112 171, 1963.
31. G. G. Gabra, A. E. Torma, C. A. Olivier, *Can. Metall. Q.* **14** (1975) no. 4, 355 – 358.
32. L. E. Rykken: "Lithium Production from Searles Valley" in [23]pp. 33 – 34.
33. W. Hinrichs, L. Lange in Winnacker, Küchler (eds.): *Chemische Technologie,* **4,** Carl Hanser Verlag, München – Wien 1986, p. 334.
34. R. Bauer, *Chem. Ing. Tech.* **44** (1972) 147 – 151. H. Lauck: "Galvanische Zellen mit organischen Elektrolyten," *BMFT-Forschungsbericht* T 74-02, 1974. F. J. Kruger, H. Lauck, E. Voss, *Chem. Ing. Tech.* **50** (1978) 284. J. O. Besenhard, G. Eichinger, *J. Electroanalyt. Chem.* **68** (1976) 1 – 18.
35. Metallgesellschaft AG, EP 0 217 438 A 1 (J. Müller, R. Bauer, B. Sermond, E. Dolling).
36. L. F. Fieser, M. Fieser: *Reagents for Organic Synthesis,* J. Wiley & Sons, New York 1967, pp. 54 – 56.
37. J. P. Gaban: *Lithium Batteries,* Academic Press, New York 1983.
38. K. M. Abraham in R. O. Bach (ed.): *Lithium–Current Applications in Science, Medicine and Technology*, J. Wiley & Sons, New York 1985, pp. 155 – 175.
39. M. Salomon, B. Scrosati, *Gaz. Chim. Ital.* **126** (1996) 415 – 427.
40. A. Klemm, *Angew. Chem.* **70** (1958) 21 – 24.
41. R. Klingelhöfer, *Chem. Ztg.* **111** (1987) 85.
42. C. M. Braams, *9th World Energy Conference,* Paper 4.1 – 11, London 1974, pp. 181 – 190.

43. H. U. Borgstedt, V. Coen, S. Malang: "Flüssige Brutstoffe", *Jahrestagung Kerntechnik 1987,* Deutsches Atomforum e. V., Bonn 1987.
44. W. Dänner, *Reaktortagung,* Hamburg 1972, Deutsches Atomforum e. V., Bonn 1972, pp. 747 – 750.
45. C. F. Baes, Jr., *J. Nucl. Mater.* **51** (1974) no. 1, 149 – 162.
46. Commissariat à l'Energie Atomique, DAS 1 244 750, 1960 (E. Saito, G. sur Yvette, G. Dirian, S.-O. Pallaiseau).
47. CEER, *Chem. Econ. Eng. Rev.* **15** (1983) 32.
48. P. D. Frost, *Batelle Tech. Rev.* **8** (1959) 3 – 8.
49. M. Peters, K. Welpman, *Metall* **39** (1985) 1141 – 1144.
50. W. A. Hart, O. F. Beumel, Jr.,: *Comprehensive Inorganic Chemistry,* vol. 12, Pergamon Press, Oxford – New York 1972, pp. 331 – 367.
51. Shell B.V., DE 1 914 004, 1969 (G. Rossa).
52. W. Pennington, *Refrig. Eng.* **63** (1955) 57 – 61.
53. D. S. Davis, *Chem. Eng.* **64** (1957) 292 – 294.
54. Sterling Drug Inc., US 885 776, 1978 (M. R. Bell, J. L. Herrmann, V. Akullian).
55. Searle, DE 3 207 470, 1981 (J. B. Hill, R. A. Erickson).
56. J. H. Fishwick: *Application of Lithium in Ceramics,* Cahners Publ. Comp., Boston, Mass. 1974.
57. B. R. Franklin, L. C. Klein: "Lithia in Container Glass," *Am. Ceram. Soc. Bull.* **62** (1983) 209. H. R. Carroll, J. I. Angelo: "Adding Lithium Can Improve Melting-Forming Performance," *Glass Ind.* Nov. 1983, 14 – 18.
58. Woellner Werke, DE-OS 27 25 524, 1978 (W. R. Semle). Armstrong World Industries, DE-OS 35 41 614, 1986 (J. O. Presley). Tokuyama Soda Co., JP 77 121 632, 1977 (T. Fukunaga, H. Iwakura, T. Uchiyama, H. Kono).
59. K. Grjotheim et al.: *Aluminium Electrolysis, the Chemistry of the Hall-Héroult Process,* Aluminium-Verlag, Düsseldorf 1977.
60. R. Pawlek: *übersicht über das Verhalten lithiumhaltiger Salze in der Tonerdeschmelzflußelektrolyse,* Schweizerische Aluminium AG, Zürich 1981.
61. H. Kvande: "The Optimum Bath Composition in Aluminium Electrolysis, does it exist?" *Erzmetall* **35** (1982) 597 – 604.
62. K. Grjotheim, H. Kvande, K. Matiasovsky: "Addition of LiF and MgF_2 to the Bath of the Hall-Héroult Process," *Light Metals,* March 1983, 397.
63. S. Gershon, B. Shopsin: *Lithium– its Role in Psychiatric Research and Treatment,* Plenum Press, New York, London 1973.
64. F. N. Johnson (ed.): *Lithium Research and Therapy,* Academic Press, London, New York 1974.
65. R. O. Bach (ed.): *Lithium–Current Applications in Science, Medicine and Technology,* John Wiley & Sons, New York 1985, pp. 337 – 407.
66. K. Kordesch, G. Simander: *Fuel Cells and their Application* , WILEY-VCH, Weinheim 1996.
67. K. Brandt, *Solid State Ionics* **69** (1994) 173 – 183.
68. K. Ozawa, *Solid State Ionics* **69** (1994) 212 – 221.
69. A. Loupy, B. Tchoubar: *Salt Effects in Organic and Organometallic Chemistry* , VCH Verlagsgesellschaft, Weinheim 1992.
70. J. M. Tarascon, D. Guymard, *Solid State Ionics* **69** (1994) 293 – 305.
71. W. Sundermeyer, O. Glemser, *Angew. Chem.* **70** (1958) 625.
72. Goldschmidt AG, EP 266 633, 1986 (K. Koerner, C. Weitemeyer, D. Wewers).
73. Hoechst AG, DE 3 443 305, 1986 (F. Meininger, L. Schläfer). Mitsubishi Chemical Industries Co. Ltd., EP 216 296, 1987 (Y. Narita, H. Takimoto, H. Sano).
74. D. Stark, B. Morgan, P. Okamoto, "Eliminating or Minimizing Alkali-Silica Reactivity," Strategic Highway Research Council SHRP-C-343, Washington, DC 1993.
75. R. G. Kelleher, M. A. McKervey, P. Vibuljan, *J. Chem. Soc. Chem. Commun.* 1980, 486.
76. J. E. McMurry, *Org. React.* **24** (1976) 187.
77. W. Weppner, W. Welzel, R. Kniep, A. Rabenau, *Angew. Chem.* **98** (1986) 1105 – 1106.
78. J. M. Tarascon et al., *J. Electrochem. Soc.* **138** (1991) 2859 – 2864.
79. Chemetall GmbH: "Sabalith, Nitrite-free Heat Transfer Salt," company brochure, Frankfurt, June 1994.
80. Showa Denko, JP 59 57 905, 1984. Showa Denko, JP 59 73 411, 1984.
81. J. Raeder: "Der Entwicklungsstand der Kernfusion," *Kerntechn.* **6** (1977) 253.
82. S. B. van der Molen, E. J. Siewers, H. den Uil, M. J. A. A. Goris: "Porous MCFC Component Development at ECN," *International Seminar Fuel Cell Technology and Applications,* Den Haag 1987.
83. K. Ziegler, H. Colonius, *Justus Liebigs Ann. Chem.* **476** (1930) 135.

84. Handling of Butyllithium, technical brochure, Chemetall GmbH, Frankfurt 1993.
85. F. Totter, P. Rittmeyer in M. Schlosser (ed.): *Organometallics in Synthesis,* John Wiley & Sons, New York 1994, pp. 167 – 194.
86. R. M. Young, R. P. Quirk, L. J. Fetters, *Adv. Polym. Sci.* **56** (1984) 1 – 90.
87. R. Taube, G. Sylvester in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* VCH Verlagsgesellschaft, Weinheim 1996, 280 – 318.
88. L. Brandsma, H. Verkruijsse, *Preparative Polar Organometallic Chemistry 1,* Springer Verlag, Berlin, 1986.
89. Chemetall GmbH, Broschure, *Butyllithium–Properties and Applications,* Frankfurt 1988, Federal Republic of Germany.
90. *Houben-Weyl*, Band E19d, 1993, pp. 369 – 433.
91. Ciba Geigy, EP 333 661, 1989; EP 333 658, 1989 (P. Ackermann, H.-R. Känel, B. Schaub).
92. B. J. Wakefield: "The Chemistry of Organolithium Compounds," Pergamon Press, Oxford 1974.
93. I. Coldham, *Contemp. Org. Synth.* **4** (1997) 136 – 163.
94. J. L. Wardell: "The Use of Organometallic Compounds in Organic Synthesis," in F. R. Hartley (ed.): *The Chemistry of the Metal – Carbon Bond,* J. Wiley & Sons, Chichester 1987.
95. H. C. Brown, G. W. Kramer, J. L. Hubbard, S. Krishnamurthy, *J. Organomet. Chem.* **188** (1980) 1 – 10.
96. Metallgesellschaft, DE 4 424 222, 1994 (U. Emmel, W. Weiß).
97. G. H. Posner: *An Introduction to Synthesis Using Organocopper Reagents,* Wiley, New York 1980.
98. Schering, DE 2 046 640, 1970 (R. Philipson, E. Kaspar).
99. Metallgesellschaft, DE 19 605 571, 1995 (U. Emmel, W. Weiß).
100. S. A. Omnichem, EP 575 303, 1992 (E. De Cock, W. Van Brussel, A. Mangelschots).
101. J. A. Dean: *Flame Photometry,* McGraw-Hill, New York 1960, pp. 153 – 160.
102. S. Kallmann: *The Alkali Metals, Treatise on Analytical Chemistry,* vol. I, part II, J. Wiley & Sons, New York 1961.
103. I. Gyenes: *Titration in nichtwässrigen Medien,* Ferdinand Enke Verlag, Stuttgart 1970.
104. J. Weiss: *Handbuch der Ionenchromatographie,* VCH Verlagsgesellschaft, Weinheim 1985.
105. G. N. Schrauzer: K.-F. Kippel: *Lithium in Biology and Medicine,* VCH Verlagsgesellschaft, Weinheim 1991.
106. D. J. McCracken, M. Haigh, *Minerals Review* **72** (1993) 105 – 106.

Local Anesthetics

Anesthetics, General is a separate keyword.

ROBERT RIPPEL, Hoechst Aktiengesellschaft (retired), Frankfurt, Federal Republic of Germany

1. Introduction 1
2. Benzoates 3
2.1. 4-Aminobenzoates 3
2.2. Basic Benzoates and Aminobenzoates 3
2.2.1. Basic Benzoates 3
2.2.2. Basic Aminobenzoates 3
3. Carboxylic Acid Amides 5
3.1. Carboxylic Acid Anilides 5
3.2. Amides of Heterocyclic Carboxylic Acids 6
3.3. Amides of Aliphatic Carboxylic Acids 7
4. Basic Ketones and Ethers 7
4.1. Basic Ketones 7
4.2. Basic Ethers 7
5. Nitrogen-Free Local Anesthetics . . 8
6. References 8

1. Introduction

Local anesthetics are compounds that reversibly inhibit the excitability of sensory nerve endings or the conductivity of neuronal pathways to pain or other sensory stimuli in a restricted area without causing permanent damage. This blockade prevents the perception of pain (analgesic action) or painlike sensations (anesthetic action); other sensations such as pressure can, however, still be felt. Local anesthetics are divided into various categories on the basis of regional differences in pain perception. *Topical anesthetics* are anesthetics for superficial sites, e.g., mucous membranes or wound surfaces. *Infiltrative anesthetics* are injected into the skin or subcutaneous tissue. Finally, *nerve block anesthetics* are injected to anesthetize specific neuronal pathways; lumbar and epidural anesthesia are special forms of this type of anesthetic.

Mechanisms of Action. The sensory nerve axons are insulated by a lipoprotein sheath of myelin which is interrupted at sites called the nodes of Ranvier to speed up conduction of the nervous impulse. The effect of the local anesthetics on the sensory nerve endings is not specific. However, different excitable structures have varying sensitivities. Several hypotheses have been proposed to explain the molecular mechanism of the local anesthetic action of clinically-used agents. All local anesthetics inhibit the rapid influx of sodium ions into the cell which normally occurs during passage of a nerve impulse. Some substances bind to a specific binding site in the Na^+ channel which can only be reached from the inner side of the nerve cell membrane. This group of anesthetics includes elongated quaternary ammonium compounds that bear a lipophilic ring. Other uncharged substances of variable chemical structure and low polarity are integrated into (i.e., infiltrate) the membrane, without binding to a specific receptor, and disturb the integrity of the Na^+ channel. Basic local anesthetics act by both of these mechanisms, but the predominant mechanism is the reaction with the binding site in the Na^+ channel. In order for the local anesthetics to reach the receptor, they first have to diffuse through the cell membrane in the nonprotonated state. After passing through the cell membrane the agents generally return to their active protonated state, dependent upon the pK value of the compound and the intracellular pH [4–6].

An ideal local anesthetic has to induce anesthesia sufficiently rapidly in a confined region for a sufficiently long time. Adverse side effects (such as excitation of the central nervous system, inhibition of transmission of nervous impulses to muscle and at the synapse, influences on heart function and the circulatory system, damage to nervous tissue, and irritation of other tissues) as well as hypersensitivity reactions and allergies should be reduced to a minimal level.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_415

Furthermore, a minimal absorption rate of the compound should be coupled with minimal toxicity. The active agent must be sufficiently stable and possess sufficient water solubility to allow preparation of a ready-for-use solution. The compound should not be too acidic or too alkaline, should be stable on sterilization, and should not decompose when stored for longer periods. Once the substance has been transported from the site of application into the bloodstream and organs, it should also be rapidly degraded into nontoxic metabolites and excreted.

Historical Aspects. The development of local anesthetics began in the second half of the 19th century with the discovery and use of the plant alkaloid cocaine. Initially cocaine was used as a topical anesthetic and nerve blocking agent. It is, however, now only used for special topical anesthesia of the eye, mainly due to its serious adverse side effects but also because of its low stability. Even before the structure of cocaine had been elucidated, efforts were made to modify cocaine and develop substances with a similar local anesthetic effect but without the drawbacks of cocaine. After various cocaine derivatives (e.g., psicaine, tropacaine, and encaine) had been prepared, the substituted benzoates of various alcohols, especially basic alcohols, were discovered. The major breakthrough occurred in 1904 when EINHORN and UHLFELDER synthesized procaine. With the discovery of procaine the systematic search for synthetic local anesthetics began. Procaine served as the first high-quality product and standard; all of its constituent parts were subsequently modified leading to the discovery of substances with local anesthetic action that displayed a variety of chemical structures. Simple relationships between structure and activity were also investigated [7–9].

Structure. The majority of the compounds now used as local anesthetics share three structural characteristics that will be illustrated with procaine:

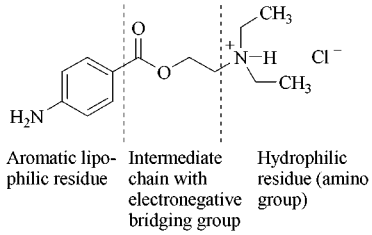

The *aromatic residue* (heterocyclic compounds are, however, also effective) confers a sufficient degree of lipid solubility to the molecule. In addition to −CO−O−, the *electronegative bridging group* can also be −CO−S−, −CO−NH−, −CO−, −O−, −NHCO−, or −NHCOO. It usually consists of 1 – 3 carbon atoms which may have alkyl substituents. The *amino group* often has two alkyl substituents (C_1–C_4), which can be joined together to form (substituted) piperidine or morpholine. Occasionally, only one alkyl group (C_1–C_6) is present. According to Löfgren's rule, the amino group represents the hydrophilic and the aromatic residue the lipophilic portion of the compound. The very small number of nitrogen-free local anesthetics (e.g., polydocanol) can also be incorporated into this model:

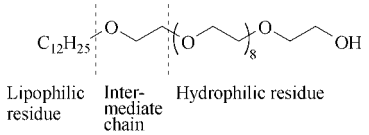

The above model is a prerequisite for a potential local anesthetic effect. However, many compounds fit this model, but have no or only minimal local anesthetic action. A useful local anesthetic is produced only when it has a favorable combination of important physicochemical characteristics such as solubility, rate of diffusion, basicity, degradability, and partition coefficient.

This article describes the most important local anesthetics but does not evaluate their absolute efficacy. Instead a review of the most commonly used or historically important compounds is given.

2. Benzoates

The majority of synthetic local anesthetics are basic esters of 4-aminobenzoic acid (or 2- or 3-aminobenzoic acid):

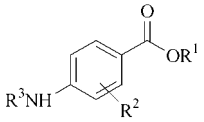

Active compounds were created by modification of the aminoalcohol R^1, insertion of a further substituent R^2 into the phenyl core, alkylation of the amino group $-NHR^3$, and even elimination of the $-NHR^3$ group.

2.1. 4-Aminobenzoates

In 1890 Ritsert had already recognized the good local anesthetic action of ethyl 4-aminobenzoate. Two 4-aminobenzoates which still have a certain degree of importance are benzocaine and butamben.

Benzocaine [*94-09-7*], ethyl 4-aminobenzoate, $C_9H_{15}NO_2$, M_r 165.2, *mp* 88 – 90 °C.

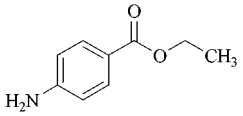

Benzocaine is produced by reduction of ethyl 4-nitrobenzoate with iron [10]. It is used as a surface anesthetic.
Trade Names. Anaesthesin (Ritsert, Germany); Hurricaine (Beutlich, USA).

Butamben [*94-25-7*], *n*-butyl 4-aminobenzoate, $C_{11}H_{15}NO_2$, M_r 193.2, *mp* 57 – 59 °C.

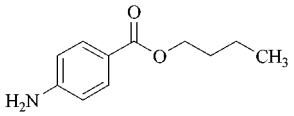

Butamben is produced by the esterification of 4-nitrobenzoyl chloride with *n*-butanol and reduction of the resulting nitroester. Butamben is available in the United States as a surface anesthetic in the form of a picrate.
Trade Name. Butesine (Abbott, USA).

2.2. Basic Benzoates and Aminobenzoates

While searching for compounds with a high water solubility, A. Einhorn and E. Uhlfelder in 1904 synthesized the diethylaminoethyl ester of 4-aminobenzoic acid, also known as procaine. Procaine's high water solubility, tolerance, and stability made it the standard anesthetic for many decades, even though it produced a relatively low level of topical anesthesia. Many similar substances with a greater potency and broader spectrum of use were subsequently marketed. Because of these new products other particularly effective agents are now available for special anesthesias. However for infiltration anesthesia, procaine remained the most widely used product for a long period of time.

2.2.1. Basic Benzoates

Piperocaine hydrochloride [*533-28-8*], 3-(2-methylpiperidino)propyl benzoate hydrochloride, $C_{16}H_{23}NO_2{\cdot}HCl$, M_r 297.83, *mp* 172 – 175 °C.

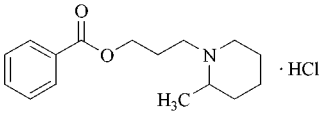

The compound is produced by esterifying benzoyl chloride with 3-chloropropanol and subsequent reaction of the ester with 2-methylpiperidine [11]. Piperocaine is used as a topical anesthetic and a nerve block agent.
Trade Name. Metycaine (Lilly, USA).

2.2.2. Basic Aminobenzoates

Procaine hydrochloride [*51-05-8*], 2-diethyl-aminoethyl 4-aminobenzoate monohydrochloride, $C_{13}H_{20}N_2O_2{\cdot}HCl$, M_r 272.77, *mp* 154 – 157 °C.

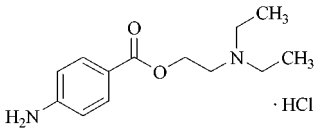

The compound is prepared by reacting benzocaine and 2-diethylaminoethanol in the presence of catalytic amounts of sodium and simultaneous removal of the nascent ethanol by distillation [12]. Procaine is employed as an infiltration anesthetic and nerve block agent.
Trade Names. Novocain (Hoechst, Germany); Novocain (Winthrop Pharmaceuticals, USA).

Chloroprocaine hydrochloride [*3858-89-7*], 2-diethylaminoethyl 4-amino-2-chlorobenzoate monohydrochloride, $C_{13}H_{19}ClN_2O_2{\cdot}HCl$, M_r 307.22, *mp* 176 – 178 °C.

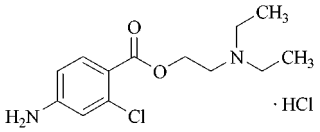

The compound is synthesized by esterification of 4-amino-2-chlorobenzoic acid with 2-diethylaminoethanol and subsequent transformation to the hydrochloride form [13]. Chloroprocaine is used as a nerve block agent and infiltration anesthetic.
Trade Name. Nesacaine (Astra Chemicals, USA).

Oxybuprocaine hydrochloride [*5987-82-6*], 2-diethylaminoethyl 4-amino-3-butoxybenzoate monohydrochloride, $C_{17}H_{28}N_2O_3{\cdot}HCl$, M_r 344.88, *mp* 157 – 160 °C.

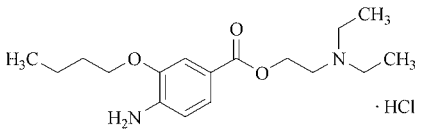

This compound is produced by transesterification of ethyl 3-butoxy-4-aminobenzoate with 2-diethylaminoethanol and subsequent transformation to the hydrochloride form [14]. Oxybuprocaine is used as a topical ocular, nasal, and otic (aural) anesthetic.
Trade Names. Novesine Wander (Wander Pharma, Germany); combined with sodium fluorescein in Fluren (Barnes Hind, USA).

Proxymetacaine hydrochloride [*5875-06-9*], 2-diethylaminoethyl 3-amino-4-propoxybenzoate monohydrochloride, $C_{16}H_{26}N_2O_3{\cdot}HCl$, M_r 330.85, *mp* 178 – 185 °C.

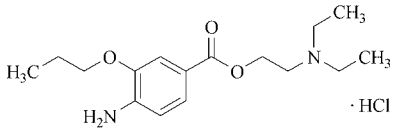

The starting material for the synthesis of this compound is 4-hydroxy-3-nitrobenzoic acid which is converted to 3-nitro-4-propoxybenzoic acid in a two-step process. The latter is esterified with 2-diethylaminoethyl chloride, the nitroester is reduced, and the compound converted to the hydrochloride form with hydrochloric acid [15]. Proxymetacaine is used as a surface anesthetic in ophthalmology.
Trade Names. Proparakain (Ursapharma, FRG); Ophthaine (Squibb, USA).

Tetracaine hydrochloride [*136-47-0*], 2-dimethylaminoethyl 4-butylaminobenzoate monohydrochloride, $C_{15}H_{24}N_2O_2{\cdot}HCl$, M_r 300.83, *mp* 134 – 148 °C (three polymorphic modifications 134, 139, 148 °C).

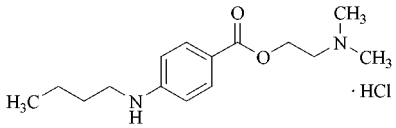

This compound is synthesized by esterification of 4-butylaminobenzoic acid with 2-dimethylaminoethanol [16]. It is used as a nerve block agent and as a topical anesthetic.
Trade Names. Pantocain, Gingicain (Hoechst, Germany); Pantocaine (Winthrop Pharmaceuticals, USA).

3. Carboxylic Acid Amides

Several carboxylic acid amides are local anesthetics, especially the anilides of basically substituted aliphatic fatty acids and several *N*-substituted pipecolic acids. Aromatic carboxylic acid amides are of less importance, with the exception of cinchocaine (a basic amide of a heterocyclic carboxylic acid) and oxetacaine (an aliphatic acid amide).

3.1. Carboxylic Acid Anilides

Prior to the discovery of procaine EINHORN and OPPENHEIMER synthesized a series of base-substituted anilides and tested their local anesthetic effects. Nirvanin (Hoechst) can be considered the parent molecule of this group of compounds:

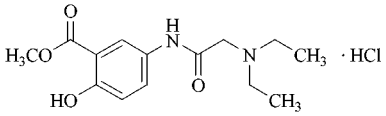

As a result of local adverse reactions this substance was only used for a short time therapeutically. It was not until 1937 that ERDTMANN and LÖFGREN resumed research into this group of compounds [17]. Several valuable derivatives were discovered which are described below. They produce rapid and long-lasting, deep local anesthesia with minimal toxicity, good tissue tolerance, and a high stability. Acid anilide compounds are consequently the most important and most widely used group of local anesthetics.

Lidocaine hydrochloride [*73-78-9*], 2-diethyl-amino-*N*-(2,6-dimethylphenyl)acetamide hydrochloride, $C_{14}H_{22}N_2O{\cdot}HCl$, M_r 270.82, *mp* 127–128 °C, as monohydrate [*6108-05-7*], M_r 288.82, *mp* 77–78 °C.

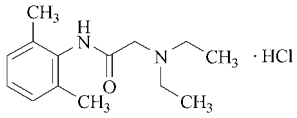

This compound is synthesized by reacting 2,6-dimethylaniline with chloroacetyl chloride in glacial acetic acid and adding sodium acetate. The product (chloroacetyl-2,6-dimethylanilide) is boiled with diethylamine in an inert solvent (e.g., benzene) [18]. The hydrochloride crystallizes with one molecule of water which can be removed by careful drying.

Lidocaine is used as an infiltration and surface anesthetic, as a nerve block agent, and as an antiarrhythmic agent after heart attacks.
Trade Name. Xylocain(e) (Astra Chemicals, Germany and USA).

Prilocaine hydrochloride [*1786-81-8*], *N*-(2-methylphenyl)-2-(propylamino)propanamidemonohydrochloride, $C_{13}H_{20}N_2O{\cdot}HCl$, M_r 256.77, *mp* 167–168 °C.

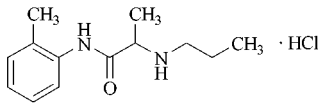

Prilocaine is synthesized by reacting 2-aminotoluene with 2-bromopropionyl chloride and then with *n*-propylamine [19].
Trade Names. Xylonest (Astra, Germany); Citanest (Astra, USA).

Mepivacaine hydrochloride [*16452-56-5*], *N*-(2,6-dimethylphenyl)-1-methyl-2-piperidinecarboxamide monohydrochloride, $C_{15}H_{22}N_2O{\cdot}HCl$, M_r 282.81, *mp* 262–264 °C.

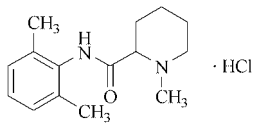

Mepivacaine is prepared either by reacting 2,6-dimethylaniline magnesium bromide with ethyl-1-methyl-2-piperidine carboxylate or by hydrogenation and subsequent *N*-methylation of pipecolic acid 2,6-xylidide [20], [21]. It is used as an infiltration anesthetic and nerve block agent.
Trade Names. Scandicain (Astra Chemicals, Germany); Carbocaine (Winthrop Pharmaceuticals, USA); Polocaine (Astra Chemicals, USA).

Bupivacaine hydrochloride [*18010-40-7*], 1-*n*-butyl-*N*-(2,6-dimethylphenyl)-2-piperidinecarboxamide monohydrochloride, $C_{18}H_{28}N_2O{\cdot}HCl$, M_r 324.9, *mp* 255–256 °C.

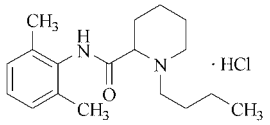

The compound is prepared by reacting 2,6-dimethylaniline with piperidine-2-carboxylic acid chloride hydrochloride. The resulting anilide is alkylated with 1-butylbromide and converted to the hydrochloride [22]. Bupivacaine is used as an infiltration anesthetic and as a nerve block agent especially for long duration blocks.
Trade Names. Carbostesin (Astra Chemicals, Germany); Marcaine (Winthrop Pharmaceuticals, USA).

Ropivacaine monohydrate [*132112–35–7*], (*S*)-(−)-1-propyl-*N*-(2,6-dimethylphenyl)-2-piperidinecarboxamide hydrochloride monohydrate, $C_{17}H_{26}N_2O{\cdot}HCl{\cdot}H_2O$, M_r 328.13, *mp* 269.5 – 270.6 °C, $[\alpha]_D^{20} = -7.28°$ ($c = 2$, H_2O).

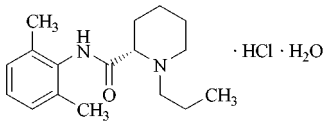

The compound is synthesized by reacting L-pipecolic acid with 2,6-xylidine and subsequent alkylation with 1-propylbromide [23]. Finally it is converted to the hydrochloride monohydrate [24].

Ropivacaine is an *N*-alkylpipecolic acid amide like mepivacaine and bupivacaine. However, in the case of ropivacaine not the racemate but only the (*S*)-enantiomer is used as a long-acting anesthetic. The indications are the same as for bupivacaine.
Trade Name. Naropin (Astra, Sweden).

Etidocaine hydrochloride [*36637-19-1*], *N*-(2,6-dimethylphenyl)-2-(ethylpropylamino) butanamide hydrochloride, $C_{17}H_{28}N_2O{\cdot}HCl$, M_r 312.9, *mp* 201 – 206 °C.

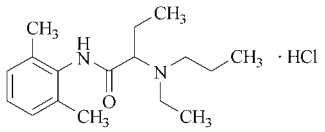

Etidocaine is synthesized by acylating 2,6-dimethylaniline with 2-bromobutyric acid. The bromine atom is replaced by iodine during boiling with potassium iodide in methanol. Reaction with *n*-propylamine yields the *n*-propylamino compound, which is then alkylated with diethyl sulfate [25]. Etidocaine is used as a long-acting infiltration anesthetic and nerve block agent.
Trade Names. Dur-Anest (Astra Chemicals, FRG); Duranest (Astra Chemicals, USA).

Butanilicaine hydrochloride [*6027-28-7*], *N*-(2-chloro-6-methylphenyl)-2-(butylamino) acetamide hydrochloride, $C_{13}H_{19}ClN_2O{\cdot}HCl$, M_r 291.2, *mp* 232 °C.

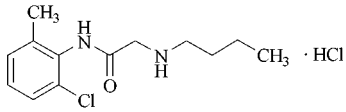

This compound is produced by reacting 2-chloro-6-methylaniline with chloroacetyl chloride. Subsequent reaction with excess *n*-butylamine yields the base which is converted to the hydrochloride form by reaction with hydrochloric acid [26]. Butanilicaine is used as an infiltration anesthetic and nerve block agent.
Trade Name. Hostacain (Hoechst, Germany).

Articaine hydrochloride [*23964-57-0*], 4-methyl-3-[2-(*n*-propylamino)propionamido]-2-thiophenecarboxylic acid methyl ester hydrochloride, $C_{13}H_{20}N_2O_3S{\cdot}HCl$, M_r 320.8, *mp* 175 – 176 °C.

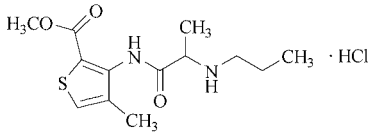

The compound is synthesized by reacting methyl 3-amino-4-methylthiophene-2-carboxylate with 2-bromopropionyl chloride and then with *n*-propylamine [27]. Articaine is used as an infiltration anesthetic and a nerve block agent.
Trade Name. Ultracain (Hoechst, Germany).

3.2. Amides of Heterocyclic Carboxylic Acids

Cinchocaine hydrochloride [*61-12-1*], 2-butoxy-*N*-[2-(diethylamino)ethyl]-4-quinolinecarboxamide monohydrochloride,

$C_{20}H_{29}N_3O_2 \cdot HCl$, M_r 379.93, *mp* 95 – 100 °C (decomp.).

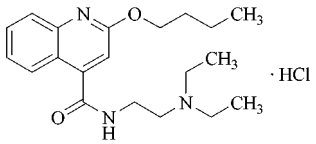

This compound is synthesized by rearrangement of *N*-acetylisatin with sodium hydroxide to form 2-hydroxyquinoline-4-carboxylic acid. Reaction with phosphorus pentachloride gives the acid chloride, whereby chlorine replaces the hydroxy group. The basic amide is then obtained on reaction with 2-diethylaminoethylamine. Addition of sodium butoxide gives the cinchocaine base [28], which is used as a topical anesthetic.
Trade Name. Nupercainal (Ciba-Geigy, USA).

3.3. Amides of Aliphatic Carboxylic Acids

Oxetacaine [*126-27-2*],2,2′-[(2-hydroxyethyl)- imino]bis[*N*-(1,1-dimethyl-2-phenylethyl)-*N*-methylacetamide], $C_{28}H_{41}N_3O_3$, M_r 467.65, *mp* 104 –104.5 °C.

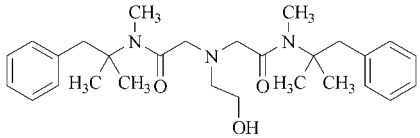

Oxetacaine is produced by reacting 1-phenyl-2,2-dimethylaminopropane with chloroacetyl chloride and then with 2-aminoethanol [29]. It is the most effective topical anesthetic in clinical use and is mainly employed for anesthesia of the gastric mucosa.
Trade Name. Tepilta suspension (Wyeth Pharma, Germany).

4. Basic Ketones and Ethers

In addition to the basic acid amides and acid anilides, extensive studies have also been performed on basic ketones and ethers. Effective local anesthetics were also found in these families of compounds; several examples are described below.

4.1. Basic Ketones

Propipocaine hydrochloride [*1155-49-3*], 1-(4-propoxyphenyl)-3-(1-piperidinyl)-1-propanone hydrochloride, $C_{17}H_{25}NO_2 \cdot HCl$, M_r 311.9, *mp* 166 °C.

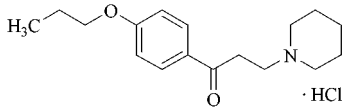

The compound is prepared by Mannich condensation of 4-propoxyacetophenone with formaldehyde and piperidine [30]. It is used as a topical anesthetic.
Trade Name. Anaesthesie (Ankerpharm, Germany).

Dyclonine hydrochloride [*536-43-6*], 1-(4-butoxyphenyl)-3-(1-piperidinyl)-1-propanone hydrochloride, $C_{18}H_{27}NO_2 \cdot HCl$, M_r 325.88, *mp* 175 – 176 °C.

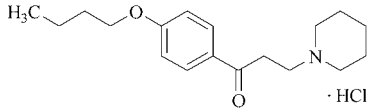

This compound is synthesized from 4-butoxyacetophenone analogously to Propipocaine [31]. It is used as a topical anesthetic.
Trade Name. Dyclone (Astra Chemicals, USA).

4.2. Basic Ethers

Pramocaine hydrochloride [*637-58-1*], 4-[3-(4-butoxyphenoxy)propyl]morpholine hydrochloride, $C_{17}H_{27}NO_3 \cdot HCl$, M_r 329.87, *mp* 181 – 183 °C.

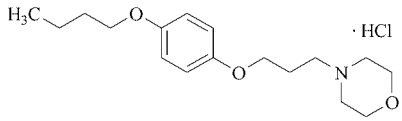

The compound is produced by alkylation of 4-butoxyphenol with 3-chloro-1-morpholinopropane [32]. It is used as a topical anesthetic.
Trade Name. Tronothane (Abott, USA).

Fomocaine hydrochloride [*56583-43-8*], 4-[3-(α-phenoxy-*p*-tolyl)propylmorpholine hydrochloride, $C_{20}H_{25}NO_2 \cdot HCl$, M_r 347.9, *mp* 171 – 173 °C.

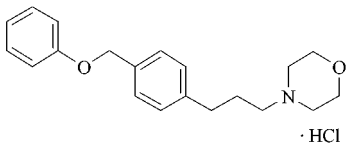

Fomocaine is produced by chloromethylation of phenylpropyl chloride, followed by synthesis of the phenolic ether and conversion of the ether to a base with morpholine [33]. It is used as a topical anesthetic.
Trade Name. Myko Posterine (Kade, Germany).

Quinisocaine hydrochloride [*2773-92-4*], 2-[(3-butyl-1-isoquinolinyl)oxy]-*N,N*-dimethylethanamine monohydrochloride, $C_{17}H_{24}N_2O \cdot HCl$, M_r 308.85, *mp* 144 – 148 °C.

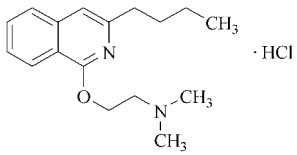

The parent substance for this compound is phthalaldehydic acid which is converted to 1,4-dihydroxy-3-butyl-3,4-dihydroisoquinoline by condensation with nitropentane, catalytic hydrogenation, and rearrangement with alkali. Reaction with phosphorus oxychloride results in the elimination of water and the production of 1-chloro-3-butyl-isoquinoline. The ether is obtained on addition of sodium 2-dimethylaminoethoxide [34]. It is used as a topical anesthetic.
Trade Names. Quotane (Smith, Kline & French, USA); Isochinol (Chemipharm, Germany).

5. Nitrogen-Free Local Anesthetics

The only nitrogen-free local anesthetics of any importance are alcohols.

Polydocanol [*3055-99-0*], poly(ethylene glycol)monododecyl ether, ca. $C_{30}H_{62}O_{10}$, $C_{12}H_{25}-O-(CH_2CH_2)_{8-9}-CH_2CH_2OH$ is synthesized by heating ethylene oxide with dodecanol at 180 °C in the presence of catalytic amounts of sodium hydroxide [35]. It is used as a topical anesthetic for phlebosclerosation.
Trade Name. Aethoxysklerol (Kreussler, Germany).

Benzyl Alcohol [*100-51-6*], C_7H_8O, M_r 108.14, *bp* 207.4 °C.

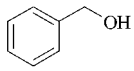

For production and properties see → Benzyl Alcohol. It is a constituent of various combination preparations used for topical anesthesia (prolongation of action).

6. References

General References

1. B. H. Takman, H. J. Adams in: *Burger's Medicinal Chemistry,* 4th ed., vol. 3, J. Wiley & Sons, New York 1981, p. 645.
2. G. Ehrhart, H. Ruschig: *Arzneimittel,* 2nd ed., **vol. 2,** Verlag Chemie, Weinheim 1972, p. 1.
3. H. Kilian: *Lokalanästhesie und Lokalanästhetika,* 2nd ed., Thieme Verlag, Stuttgart 1973.

Specific References

4. J. M. Ritchie, N. M. Green, "Local Anesthetics" in L. S. Goodman, A. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* 6th ed., Macmillan, New York – Toronto – London 1980, p. 300.
5. J. M. Ritchie, B. Ritchie, P. Greengard, *J. Pharmacol. Exp. Ther.* **150** (1965) 160.
6. H. C. Niesel: *Regionalanaesthesie, Lokalanaesthesie, regionale Schmerztherapie,* Georg Thieme Verlag, Stuttgart – New York 1994.
7. K. Miescher, *Helv. Chim. Acta* **15** (1932) 197.
8. N. Löfgren: *Studies on Local Anesthetics: Xylocaine, a New Synthetic Drug,* Hoegströms, Stockholm 1948.
9. J. Büchi, G. Cordes, X. Perlia, *Arzneim. Forsch.* **14** (1964) 161.
10. R. Adams, F. L. Cohen, *Org. Synth.* **8** (1928) 66.
11. S. M. McElrain, *J. Am. Chem. Soc.* **49** (1927) 2835.

12. Hoechst, DE 172 568 (1905); BIOS 116, Appendix 15, Proc. 1.
13. Wallace Tiernan, US 2 460 139 (1949).
14. J. Büchi et al., *Helv. Chim. Acta* **34** (1951) 1002.
15. R. O. Clinton et al., *J. Am. Chem. Soc.* **74** (1952) 592.
16. O. Eisleb, *Med. Chem. (Leverkusen, Ger.)* **2** (1934) 375.
17. H. Erdtmann, N. Löfgren, *Svensk. Kem. Tidskr.* **49** (1937) 163.
18. N. Löfgren, *Ark. Kemi Mineral. Geol.* **22 A** (1946) 30.
19. N. Löfgren, C. Tegnér, *Acta Chem. Scand.* **14** (1960) 486, 490.
20. B. af Ekenstam et al., *Acta Chem. Scand.* **11** (1957) 1183.
21. AB Bofors, US 4 110 331, 1978 (B. G. Petterson).
22. AB Bofors, DE 1 161 900 1955; 1 169 941, 1955 (B. af Ekenstam, B. P. H. Egnér).
23. Astra Läkemedel AB, US 4 695 576, 1985 (B. af Ekenstam, C. Bovin).
24. Astra Läkemedel AB, EP 0 239 710, 1986 (R. V. Sandberg).
25. Astra, US 3 812 147, 1974; 3 862 321, 1975 (H. Adams, G. Kronberg, B. H. Takman).
26. Hoechst, DE 942 993, 1952 (G. Erhart et al.); 1 005 075, 1955 (G. Erhart et al.); 1 009 633, 1953 (H. Ruschig, W. Aumüller, L. Stein, L. Ther).
27. Hoechst, US 3 855 243, 1974 (H. Ruschig, M. Schorr, R. Muschaweck, R. Rippel).
28. K. Miescher, *Helv. Chim. Acta* **15** (1932) 163.
29. M. Freed et al., *J. Org. Chem.* **26** (1961) 2378.
30. E. Profft, *Chem. Tech. (Berlin)* **3** (1951) 210; **4** (1952) 241; **5** (1953) 13; **10** (1958) 302.
31. H. J. Florestano, M. Bahler, *J. Am. Pharm. Assoc.* **45** (1956) 320.
32. H. B. Wright, M. B. Moore, *J. Am. Chem. Soc.* **76** (1954) 4396.
33. H. Oelschläger et al., *Arzneim. Forsch.* **9** (1959) 313; **18** (1968) 729; *Pharm. Acta Helv.* **46** (1971) 671.
34. J. W. Wilson et al., *J. Am. Chem. Soc.* **71** (1949) 937.
35. K. H. Siems, K. Soehring, *Arzneim. Forsch.* **10** (1960) 990.

Locust Bean Gum → Polysaccharides
Low-Density Polyethylene → Polyolefins

Lubricants and Lubrication

THORSTEN BARTELS, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Chaps. 9 and 16)

WOLFGANG BOCK, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Sections 10.1 and 10.2, Chap. 11)

JÜRGEN BRAUN, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chap. 6)

CHRISTIAN BUSCH, Fuchs Lubritech GmbH, Mannheim, Federal Republic of Germany (Chap. 15)

WOLFGANG BUSS, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Sections 13.1–13.4)

WILFRIED DRESEL, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chaps. 5 and 14)

CARMEN FREILER, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Chap. 12)

MANFRED HARPERSCHEID, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Chap. 8)

ROLF-PETER HECKLER, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chap. 14)

DIETRICH HÖRNER, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chap. 12)

FRANZ KUBICKI, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Section 13.1)

GEORG LINGG, Fuchs Petrolub AG, Queretaro, Mexico (Section 10.1)

ACHIM LOSCH, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Section 13.1)

ROLF LUTHER, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chaps. 7 and 19)

THEO MANG, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chaps. 1–4, 12, Sections 13.1–13.4, Chaps. 17, 18)

SIEGFRIED NOLL, Fuchs Petrolub AG, Mannheim, Federal Republic of Germany (Chap. 16)

JÜRGEN OMEIS, Fuchs Europe Schmierstoffe GmbH, Mannheim, Federal Republic of Germany (Chaps. 6 and 8)

1. Introduction 6
2. Lubricants in the Tribological System 7
2.1. Friction 8
2.1.1. Types of Friction 8
2.1.2. Friction and Lubrication Conditions 10
2.2. Wear 11
3. Rheology of Lubricants 11
3.1. Viscosity 11
3.2. Special Rheological Effects . . . 15
3.3. Viscosity Grades 16
4. Base Oils 18
4.1. Historical Review and Outlook . 19
4.2. Chemical Characterization of Mineral Base Oils 19
4.3. Refining 20
4.3.1. Distillation 20
4.3.2. Deasphalting 20
4.3.3. Traditional Refining Processes . . 21
4.3.3.1. Acid Refining 21
4.3.3.2. Solvent Extraction 21
4.3.4. Solvent Dewaxing 23
4.3.5. Finishing 24
4.4. Base Oil Manufacturing by Hydrogenation and Hydrocracking 24
4.4.1. Manufacturing Naphthenic Base Oils by Hydrogenation 25
4.4.2. Production of White Oils 25
4.4.3. Lube Hydrocracking 27
4.4.4. Catalytic Dewaxing 27
4.4.5. Wax Isomerization 28
4.4.6. Hybrid Lube Oil Processing 28
4.4.7. All-Hydrogen Route 29
4.4.8. Gas-to-Liquids Conversion Technology 29
4.5. Boiling and Evaporation Behavior of Base Oils 31

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_423

5. Synthetic Base Oils 32
5.1. Synthetic Hydrocarbons 32
5.1.1. Polyalphaolefins 33
5.1.2. Polyinternalolefins 33
5.1.3. Polybutenes 33
5.1.4. Alkylated Aromatics 35
5.1.5. Other Hydrocarbons 35
5.2. Halogenated Hydrocarbons . . . 35
5.3. Synthetic Esters 35
5.3.1. Esters of Carboxylic Acids 35
5.3.1.1. Dicarboxylic Acid Esters 36
5.3.1.2. Polyol Esters 36
5.3.1.3. Other Carboxylic Esters 36
5.3.1.4. Complex Esters 37
5.3.1.5. Fluorinated Carboxylic Acid Esters 37
5.3.2. Phosphate Esters 37
5.4. Polyalkylene Glycols 38
5.5. Other Polyethers 39
5.5.1. Perfluorinated Polyethers 39
5.5.2. Polyphenyl Ethers 39
5.5.3. Polysiloxanes (Silicone Oils) . . . 40
5.6. Other Synthetic Base Oils 42
5.7. Mixtures of Synthetic Lubricants 43
6. Additives 43
6.1. Antioxidants 43
6.1.1. Mechanism of Oxidation and Antioxidants 43
6.1.2. Compounds 44
6.2. Viscosity Modifiers 46
6.2.1. *VI* Improvement Mechanisms . . . 46
6.2.2. Structure and Chemistry of Viscosity Modifiers 46
6.3. Pour Point Depressants 46
6.4. Detergents and Dispersants . . . 48
6.4.1. Metal-Containing Compounds (Detergents) 48
6.4.2. Ashless Dispersants (AD) 49
6.5. Antifoam Agents 49
6.6. Demulsifiers 50
6.7. Dyes 51
6.8. Antiwear and Extreme Pressure Additives 51
6.9. Friction Modifiers 54
6.10. Corrosion Inhibitors 54
6.10.1. Antirust Additives 54
6.10.2. Metal Passivators 55
7. Lubricants in the Environment . 56
7.1. Current Situation 56
7.1.1. Economic Consequences and Substitution Potential 57
7.1.2. Agriculture, Economy, and Politics 58
7.2. Biodegradable Base Oils for Lubricants 58
7.2.1. Synthetic Esters 59
7.2.2. Polyalkylene Glycols 59
7.2.3. Polyalphaolefins 59
7.2.4. Relevant Properties of Biodegradable Base Oils 59
7.3. Additives 60
7.4. Products (Examples) 60
8. Lubricants for Internal Combustion Engines 62
8.1. Four-Stroke Engine Oils 62
8.1.1. General Overview 62
8.1.1.1. Fundamental Principles 62
8.1.1.2. Performance Specifications 63
8.1.1.3. Formulation of Engine Oils 63
8.1.1.4. Additives 64
8.1.2. Characterization and Testing . . . 65
8.1.2.1. Physical and Chemical Testing . . 65
8.1.2.2. Engine Testing 65
8.1.2.3. Passenger Car Engine Oils 66
8.1.2.4. Engine Oil for Commercial Vehicles 66
8.1.3. Classification by Specification . . 68
8.1.3.1. MIL Specifications 68
8.1.3.2. API and ILSAC Classification . . 68
8.1.3.3. ACEA Specifications 68
8.1.3.4. Manufacturers' Approvals 69
8.1.3.5. Future Trends 69
8.2. Two-Stroke Oils 71
8.2.1. Application and Characteristics . . 71
8.2.2. Classification 72
8.2.2.1. API Service Groups 72
8.2.2.2. JASO Classification 73
8.2.2.3. ISO Classification 73
8.2.3. Oils for Two-Stroke Outboard Engines 74
8.2.4. Environmentally Friendly Two-Stroke Oils 74
8.3. Tractor Oils 75
8.4. Gas Engine Oils 75
8.5. Marine Diesel Engine Oils 76
8.5.1. Low-Speed Crosshead Engines . . 76
8.5.2. Medium-Speed Engines 76
8.5.3. Lubricants 77
9. Gear Lubrication Oils 77
9.1. Introduction 77
9.2. Requirements of Gear Lubrication Oils 78
9.3. Tribology of Gears 78
9.3.1. Friction Conditions of Gear Types 78
9.3.2. Specific Gear and Transmission Failure 82
9.4. Gear Lubrication Oils for Motor Vehicles 83
9.4.1. Lubricants for Gear Drives in Commercial Vehicles 84

9.4.2. Lubricants for Gear Drives in Passenger Cars . . . 85
9.4.3. Lubricants for Automatic Transmissions and CVTs . . . 86
9.5. Multipurpose Lubricants in Vehicle Gears . . . 90
9.6. Gear Lubricants for Industrial Gears . . . 91
10. Compressor Oils . . . 93
10.1. Gas Compressor . . . 93
10.1.1. Displacement Compressors . . . 93
10.1.2. Dynamic Compressors . . . 95
10.1.3. Preparation of Compressed Air . . . 95
10.1.4. Oils for Compression of Other Gases . . . 95
10.1.5. Characteristics of Gas Compressor Oils . . . 95
10.1.6. Standards and Specifications of Compressor Oils . . . 96
10.2. Refrigerator Oils . . . 96
10.2.1. Introduction . . . 96
10.2.2. Minimum Requirements . . . 98
10.2.3. Classification . . . 99
10.2.4. Viscosity Selection . . . 104
11. Turbine Oils . . . 106
11.1. Demands on Turbine Oils – Characteristics . . . 106
11.2. Formulation . . . 107
11.3. Specifications . . . 108
11.4. Turbine Oil Circuits . . . 108
11.5. Monitoring and Maintenance of Turbine Oils . . . 108
11.6. Life of (Steam) Turbine Oils . . . 109
11.7. Gas Turbine Oils – Application and Requirements . . . 109
11.8. Fire-Resistant, Water-Free Fluids for Power Station Applications . . . 112
11.9. Lubricants for Water Turbines and Hydroelectric Plants . . . 112
12. Metalworking Fluids . . . 113
12.1. Mechanism of Action . . . 113
12.2. Water-Miscible Cutting Fluids . . . 114
12.2.1. Composition . . . 114
12.2.2. Corrosion Protection and Corrosion Test Methods . . . 114
12.2.3. Concentration of Water-Mixed Cutting Fluids . . . 114
12.2.4. Stability of Coolants . . . 115
12.2.5. Foaming Properties . . . 115
12.2.6. Preservation of Coolants with Biocides . . . 115
12.3. Neat Cutting Fluids . . . 116
12.3.1. Specifications . . . 116
12.3.2. Composition . . . 116
12.4. Application . . . 117
12.4.1. Machining with Geometrically Defined Cutting Edges . . . 117
12.4.2. Machining with Geometric Non-Defined Cutting Edges . . . 118
12.5. Storage . . . 120
12.6. Environmental Aspects . . . 120
12.7. New Trends in Coolant Technology . . . 120
13. Forming Lubricants . . . 122
13.1. Sheet Metal Working Lubricants 122
13.1.1. Deep Drawing . . . 122
13.1.2. Stretch Drawing and a Combination of Stretch and Deep Drawing 124
13.1.3. Shear Cutting . . . 124
13.1.4. Choice of Lubricants . . . 125
13.1.5. Sheet Metal Forming in Automobile Manufacturing . . . 126
13.2. Lubricants for Wire, Tube, and Profile Drawing . . . 126
13.2.1. Wire Drawing . . . 126
13.2.2. Profile Drawing . . . 128
13.2.3. Tube Drawing . . . 128
13.2.4. Hydroforming . . . 129
13.3. Lubricants for Rolling . . . 130
13.3.1. Rolling Steel Sheet . . . 131
13.3.2. Rolling Aluminum Sheet . . . 133
13.3.3. Rolling of Other Materials . . . 134
13.4. Lubricants for Solid Metal Forming . . . 134
14. Lubricating Greases . . . 137
14.1. Introduction . . . 137
14.2. Components of Greases . . . 138
14.2.1. Thickeners . . . 138
14.2.1.1. Simple Soaps . . . 138
14.2.1.2. Complex Soaps . . . 140
14.2.2. Other Ionic Organic Thickeners . 142
14.2.3. Nonionic Organic Thickeners . . . 142
14.2.4. Inorganic Thickeners . . . 143
14.2.5. Miscellaneous Thickeners . . . 143
14.2.6. Temporarily Thickened Fluids . . 143
14.3. Base Oils . . . 143
14.3.1. Mineral Oils . . . 144
14.3.2. Synthetic Base Oils . . . 144
14.4. Grease Structure . . . 144
14.5. Additives . . . 145
14.6. Manufacture of Greases . . . 146
14.6.1. Metal Soap-Based Greases . . . 146
14.6.2. Oligourea Greases . . . 147
14.6.3. Gel Greases . . . 147
14.7. Grease Rheology . . . 148
14.8. Performance . . . 149
14.8.1. Test Methods . . . 149

14.8.2. Analytical Methods 150
14.9. Applications 150
14.9.1. Roller Bearings 150
14.9.2. Cars, Trucks, Construction Vehicles 151
14.9.3. Steel Mills 153
14.9.4. Mining 153
14.9.5. Railroad, Railway 153
14.9.6. Gears 154
14.9.7. Food-Grade Applications 154
14.9.8. Textile Machines 154
14.9.9. Applications with Polymeric Materials 154
14.10. Ecology and the Environment . 154
14.11. Grease Tribology 155
15. Solid Lubricants 155
15.1. Classification 156
15.1.1. Class 1: Structural Lubricants . . . 156
15.1.2. Class 2: Mechanical Lubricants . . 156
15.1.3. Class 3: Soaps 158
15.1.4. Class 4: Chemically Active Lubricants 158
15.2. Characteristics 158
15.2.1. Crystal Structures of Lamellar Solid Lubricants 158
15.2.2. Heat Stability 159
15.2.3. Thermal Conductivity 159
15.2.4. Adsorbed Films 159
15.2.5. Chemical Stability 159
15.2.6. Particle Size 159
15.3. Products Containing Solid Lubricants 160
15.3.1. Powders 160
15.3.2. Dispersions and Suspensions . . . 160
15.3.3. Greases and Grease Pastes 160
15.3.4. Pastes 161
15.3.5. Dry-Film Lubricants 162
15.4. Industrial Uses of Products Containing Solid Lubricants 163
15.4.1. Screw Lubrication 163
15.4.2. Roller-Bearing Lubrication 163
15.4.3. Slide Bearing, Slide Guideway, and Slide Surface Lubrication 164
15.4.4. Chain Lubrication 164
15.4.5. Plastic and Elastomer Lubrication 165
16. Testing and Analysis 165
16.1. Base Oil Categories and Evaluation of Various Petroleum Base Oils 165
16.2. Laboratory Methods for Testing Lubricants 165
16.2.1. Density 165
16.2.2. Viscosity 166
16.2.3. Refractive Index 167
16.2.4. Structural Analyses 167
16.2.5. Flash Point 167
16.2.6. Surface Phenomena 167
16.2.7. Cloud Point, Pour Point 167
16.2.8. Aniline Point 168
16.2.9. Water Content 168
16.2.10. Ash Content 168
16.2.11. Acidity, Alkalinity 168
16.2.12. Aging Tests 168
16.2.13. Hydrolytic Stability 169
16.2.14. Corrosion Tests 169
16.2.15. Oil Compatibility of Seals and Insulating Materials 169
16.2.16. Evaporation Loss 169
16.2.17. Analysis of Grease 169
16.3. Mechanical – Dynamic Testing Methods for Lubricants 169
16.3.1. Tribological System Categories within Lubricant Tests 169
16.3.2. Standardized and Nonstandardized Test Methods for Lubricants 169
16.3.3. Common Mechanical – Dynamic Testers 170
17. Economic Aspects 175
18. Disposal of Used Lubricating Oils 177
18.1. Possible Uses of Waste Oil 177
18.2. Legislative Influences on Waste Oil Collection and Reconditioning 177
18.3. Re-refining 178
19. Toxicology and Occupational Health 182
19.1. Safety Aspects of Handling Lubricants (Working Materials) . . 182
19.1.1. Polycyclic Aromatic Hydrocarbons (PAK, PAH, PCA) 182
19.1.2. Nitrosamines in Cutting Fluids . . 183
19.2. Skin Problems Caused by Lubricants 183
20. References 184

The entire topic was coordinated by Theo Mang and Wilfried Dresl

Abbreviations

AAS atomic absorption spectrometry
ACEA Association des Constructeurs Européens d'Automobiles (Association of European Car Manufacturers)
AD ashless dispersant
AENOR Associatión Española de Normalización y Certificatión (Spanish Association for Normalization and Certification)
AFNOR Association Française de Normation (French Association for Normalization)
AGMA American Gear Manufacturers Association
ASTME American Society of Tool and Manufacturing Engineers (USA)
ATF automatic transmission fluid
ATIEL Association Technique de l'Industrie Européenne des Lubrifications (Technical Association of the European Lubricant Industry)
AW antiwear
AWT Almen – Wieland test
CAFE Californian Act for Fuel Emission
CCEL China Certification Committee for Environmental Labelling of products
CCMC Committee of Common Market Automobile Constructors (EU)
CCS cold cracking simulator
CEC Coordinating European Council for the development of performance tests for lubricants and engine oils
CETOP Comité Européen des Transmissions Oleohydrauliques et Pneumatiques (European Oil Hydraulic and Pneumatic Committee)
CFMS closed-field magnetron sputtering
CONCAWE The Oil Companies' Organization for the Conservation of Clean Air and Water in Europe
CRC Coordinating Research Council (EU)
CVT constantly variable transmission
DD detergent and dispersant
DGMK Deutsche Wissenschaftliche Gesellschaft für Erdöl, Erdgas und Kohle (German Scientific Society for Mineral Oil, Natual Gas and Coal)
DKA Deutscher Koordinationsausschuß im CEC
DLC diamond-like amorphous carbon
DSC differential scanning calorimetry
EBT electron beam texturing
ECP Environmental Choice Program (Canada)
EHD elastohydrodynamic
EHEDG European Hygienic Equipment Design Group
ELGI European Lubricating Grease Institute
EP extreme pressure
FCC fluid catalytic cracker
FM friction modifier
FTMS Federal Test Methods Standardization (USA)
FVA Forschungsvereinigung Antriebstechnik (German Research Association for Drive Technology)
FZG Forschungsstelle für Zahnräder und Getriebebau (German Research Center for Toothed Wheel and Gearing Engineering)
GfT Gesellschaft für Tribologie (German Society for Tribologie)
GOST Gossudarstwenny Obschtschessojusny Standart (former Soviet Union, Governmental Union Standard)
HC hydrocracked
HD heavy duty
HLB hydrophilic-lipophilic balance
HPDSC high-pressure differential scanning calorimetry
HRC Rockwell C hardness
HTHS high-temperature high shear
HVI high viscosity index
IBAD ion-beam-assisted deposition
ICP inductively coupled plasma atomic emission
IFP Institut Français du Pétrol (French Institute of Petroleum)
ILMA Independent Lubricant Manufacturers Association (USA)
ILSAC International Lubricant Standardization and Approval Committee
IP Institute of Petroleum (UK)
ISO VG ISO viscosity grade
JEA Japanese Environmental Association
JIS Japanese Industrial Standard
LDF long drain field test
LVFA low-velocity friction apparatus

LVI low viscosity index
MIL Military Standard (USA)
MLDW Mobil Lube Dewaxing
MQL minimum quantity lubrication
MRV mini rotary viscosimeter
MSDW Mobil Selective Dewaxing
N – D – M (refractive index) n-density-molecular mass
NLGI National Lubricating Grease Institute (USA)
OEM original equipment manufacturer
PCMO passenger car motor oil
PEP passive extreme pressure
PPD pour point depressant
RAL UZ Reichsausschuß für Lieferbedingungen und Gütersicherung, Umweltzeichen (German Imperial Committee for Quality Control and Labelling, Environmental Symbol)
RNT radionuclide technique
RVT Reichert Verschleiß Test (Reichert wear test)
SCDSC sealed capsule differential scanning calorimetry
SHPD super-high performance diesel
SRE standard reference elastomer
SRV Schwing – Reibverschleiß Gerät (translatorisches Oszillations-Prüfgerät) (translatory oscillation apparatus)
STLE Society of Tribologists and Lubrication Engineers
STOU Super Tractor Oil Universal
SUS Saybolt universal seconds
TAN total acid number
TBN total base number
TDA thermal deasphalting
TEI Thailand Environmental Institute
TEWL transepidermal water loss
TG thermogravimetry
TGL Technische Normen, Gütervorschriften und Lieferbedingungen (former GDR, Technical Standards, Quality Specifications and Terms of Delivery)
TOST turbine oil oxidation stability test
TRIP transformation induced plastics
TSSI temporary shear stability index
UEIL Union Européenne des Indepéndents en Lubrifiants (European Union of Independent Lubricant Manufacturers)
UHPD ultra-high performance diesel
UHVI ultra-high viscosity index
UNITI Bundesverband Mittelständischer Mineralölunternehmen (German Federal Association of Middle-class Oil Companies)
UTTO universal tractor transmission oil
VAMIL Regeling Willekeurig Afschrijving Milieu-Investeringen (Dutch regulation of discretion deductions for investments into the environment)
VCI vapor-phase (volatile) corrosion inhibitor
VDA Verband der Automobilindustrie (German Automobile Industry Association)
VDMA Verband Deutscher Maschinen- und Anlagenbau (German Engineering Federation)
VDS Volvo drain specification
VGB Technische Vereinigung der Großkraftwerkbetreiber (German Technical Association of Large Power Plant Operators)
VGO vacuum gas oil
VHVI very high viscosity index
VI viscosity index
VIE viscoisty index extended
VII viscosity index improver
VOC volatile organic compounds
VTC viscosity-temperature coefficient
WGK Wassergefährdungsklasse (Water Hazard Class, Water Pollution Class)
XHVI extra-high viscosity index
XRF X-ray fluorescence spectromety
ZAF zinc- and ash-free

1. Introduction

The most important function of lubricants is the reduction of friction and wear. Scientific research has shown that 0.4 % of gross domestic product could be saved in terms of energy in Western industrialized countries if current tribological knowledge, i.e., the science of friction, wear, and lubrication, was just applied to lubricated processes.

Apart from important applications in internal combustion engines, vehicle and industrial gearboxes, compressors, turbines, or hydraulic systems, there are a vast number of other applications which mostly require specifically tailored

lubricants. Between 5000 and 10 000 different lubricant formulations are necessary to satisfy more than 90 % of all lubricant applications.

Mineral oil components continue to form the quantitatively most important foundation of lubricants. Petrochemical components and increasingly derivatives of natural, harvestable raw materials from the oleo-chemical industry are finding increasing acceptance because of their environmental compatibility and some technical advantages

Lubricants today are classified into two major groups: automotive lubricants and industrial lubricants. Industrial lubricants can be sub-divided into industrial oils and industrial specialties; i.e., greases, metalworking lubricants, and solid lubricant films.

Process oils are often included in lubricant statistics. These are oils which are included as raw materials in processes, but above all as plasticizers for the rubber industry. Process oil's only link with lubricants is that they are mineral oil products resulting from the refining of base oils but they often distort lubricant consumption figures. They will not be covered in this book. However, to provide a degree of comparison, they have been included in the following lubricant statistics.

Apart from the most common lube oils, the many thousands of lubricant applications necessitate a diverse number of systems which is seldom equaled in other product groups.

The group next to oils are emulsions, which as oil-in-water emulsions are central to water-miscible cutting fluids (Chap. 12) and rolling emulsions. In these cases, the lubricant manufacturer normally supplies a concentrate which is mixed with water locally to form an emulsion. The concentration of these emulsions with water are generally between 1 and 10 %. The annual consumption of such emulsions in industrialized countries is about the same as all other lubricants together. From this point of view, the volumetric proportion of these products (as concentrates) is significantly under-rated in lubricant statistics.

The next group of lubricant systems are water-in-oil emulsions (invert emulsions). Their most important application is in metal forming. These products are supplied ready-to-use or as dilutable concentrates. Fire resistant HFB fluids are designed as water-in-oil emulsions too.

In some special cases, oil-in-oil emulsions are developed as lubricants and these are primarily used in the field of metalworking.

Water-based solutions in the form of non-dispersed systems are sometimes used in chip-forming metalworking operations.

Greases (Chap. 14) are complex systems consisting of base oils and thickeners based on soaps or other organic or inorganic substances. They are available in semiliquid form (semifluid greases) through to solid blocks (block greases). Special equipment is required for their production (grease-making plants). A group of products closely related to greases are pastes.

Solid lubricant suspensions normally contain solid lubricants in stable suspension in a fluid such as water or oil. These products are often used in forging and extrusion and other metalworking processes. Solid lubricant films can also be applied as suspensions in a carrier fluid which evaporates before the lubricant has to function.

Solid lubricant powders can be applied directly to specially-prepared surfaces.

In the case of dry-film lubricants (Chap. 15), solid lubricants are dispersed in resin matrices. Dry-film lubricants are formed when the solvent (principally water or hydrocarbons) evaporates.

Molten salts or glass powder are used for hot forming processes such as extrusion. These are normally supplied as dry powders and develop lubricity when they melt on the hot surface of the metal.

Polymer films are used when special surface protection is required in addition to lubricity (e.g., the pressing of stainless steel panels). Together with greases, these products are also used to some extent in the construction industry.

2. Lubricants in the Tribological System

Tribology (derived from the Greek tribein, or tribos meaning rubbing) is the science of friction, wear, and lubrication. Although the use of lubricants is as old as mankind, scientific focus on lubricants and lubrication technology is relatively new. The term tribology was first introduced in 1966 and has been used globally to describe this far-reaching field of activity since 1985. Even though efforts had been made since the 16th century to describe the whole phenomenon of fric-

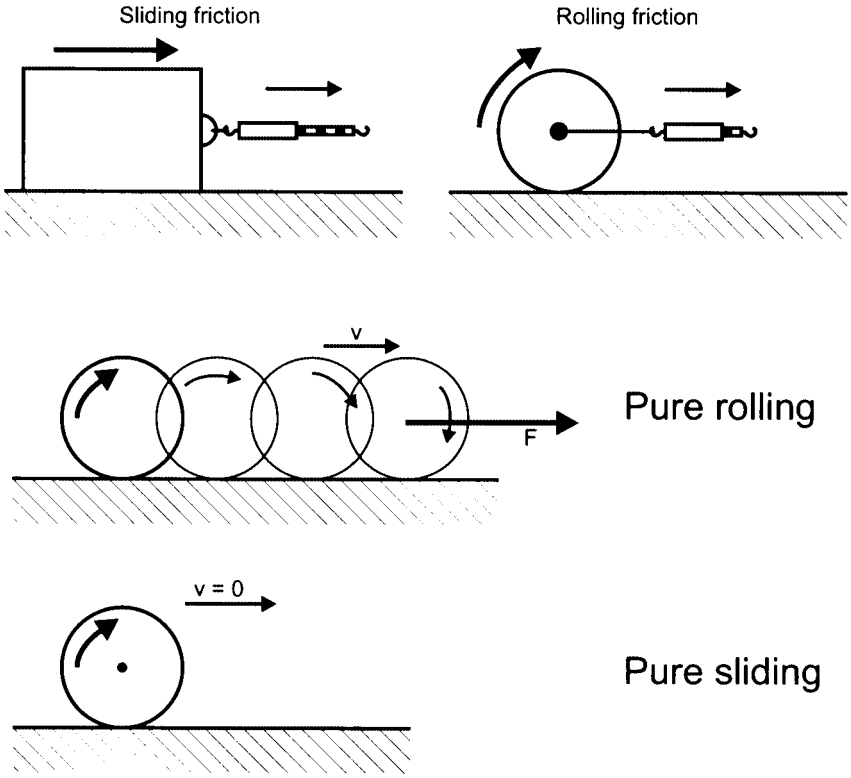

Figure 1. Sliding and rolling

tion scientifically (Leonardo da Vinci, Amontons, Coulomb), the work always concentrated on single aspects and lubricants were not even considered. Some research work performed up to the early 1970s totally ignored the chemical processes which take place in lubricated friction processes.

The tribological system (commonly referred to as the tribosystem) consists of four elements: the twocontacting partners, the interface between the two and the medium in the interface and the environment [52]. In lubricated bearings the lubricant is located in this gap. In plain bearings, the material pair are the shaft and the bearing shells; in combustion engines they are the piston rings and the cylinder wall or the camshaft lobes and the tappets and in metalworking, the tool and the workpiece.

The variables are the type of movement, the forces involved, temperature, speed, and duration of the stress. Tribometric parameters – such as friction, wear, and temperature data – can be gathered from the stress area. Tribological stress is the result of numerous criteria of surface and contact geometry, surface loading, or lubricant thickness. Tribological processes can occur in the contact area between two friction partners – which can be physical, physicochemical (e.g., adsorption, desorption), or chemical in nature (tribochemistry).

2.1. Friction

2.1.1. Types of Friction

Friction is the mechanical force which resists movement (dynamic or kinetic friction) or hinders movement (static friction) between sliding or rolling surfaces. These types of friction are also called external friction.

Internal friction results from the friction between lubricant molecules; this is described as viscosity (Chap. 3).

The causes of external friction are, above all, the microscopic contact points between two slid-

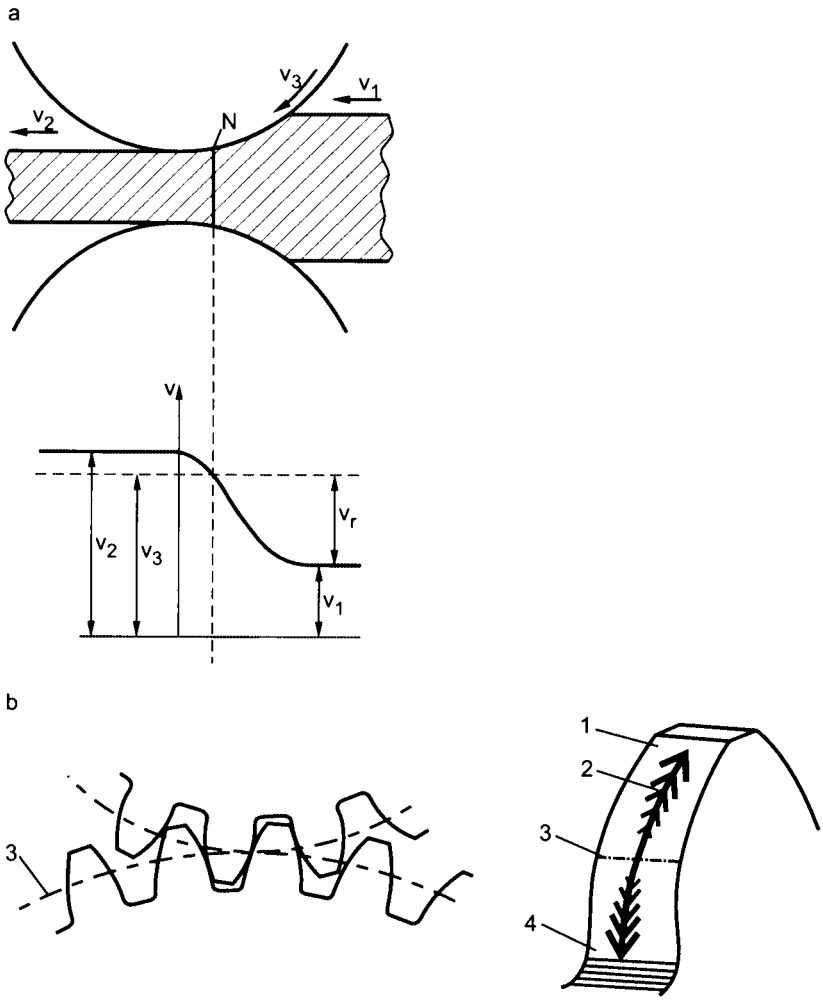

Figure 2. 'Wälzreibung', mixing of rolling and sliding motions.
a) Rolling in metal forming; v_1, initial speed of the sheet metal; v_2, final speed of the sheet metal; v_3, speed of the roller; v_r, speed difference in the roll gap (sliding part); N, neutral point (non-slip point, pure rolling)
b) Engagement of gear teeth, 1, 2, 4. high sliding/rolling ratio; 3. pitch circle (pure rolling, no slip)

ing surfaces; these cause adhesion, material deformation, and grooving. Energy which is lost as friction can be measured as heat and/or mechanical vibration. Lubricants should reduce or avoid the micro-contact which causes external friction.

Sliding friction occurs in a pure sliding motion with no rolling and no spin (Fig. 1).

The coefficient of friction μ is defined as the dimensionless ratio of the friction force F and the normal force N. The proportionality between normal force and frictional force is often given in dry and boundary friction conditions but not in fluid-film lubrication.

Rolling friction is the friction generated by rolling contact. In roller bearings, rolling friction mainly occurs between the rolling elements and the raceways, whereas sliding friction occurs between the rolling elements and the cage. The main cause of friction in roller bearings is sliding in the contact zones between the rolling elements and the raceways. It is also influenced by the geometry of the contacting surfaces and

the deformation of the contacting elements. In addition, sliding also occurs between the cage pockets and the rolling elements.

If rolling motion and sliding motion combine to any significant extent, as for gear tooth meshing, special terminology has been created. The word 'Wälzreibung' which is derived from 'Wälzen' (rolling, e.g., steel rolling) is used in Germany. Situations in which a high sliding/rolling ratio occur require totally different lubrication than does pure sliding. Figure 2 shows this 'rolling friction' during rolling and during gear meshing.

Static Friction. The static coefficient of friction is defined as the coefficient of friction corresponding to the maximum force that must be overcome to initiate macroscopic motion between two bodies (ASTM).

Kinetic Friction. Different from static friction, kinetic friction occurs under conditions of relative motion. ASTM defines the kinetic coefficient of friction as the coefficient under conditions of macroscopic relative motion of two bodies. The kinetic coefficient of friction (dynamic coefficient of friction) is usually somewhat smaller than the static coefficient of friction.

Stick – slip is a special form of friction which often results from very slow sliding movements when the friction partners are connected to a system which can vibrate. The process is influenced by the dependence of the coefficient of sliding friction on speed. This generally occurs when the static coefficient of friction (f_{stat}) is larger than the dynamic coefficient of friction (f_{dyn}). Stick – slip is normally encountered with machine tools which operate with slow feeds. Stick – slip can cause chatter marks on components.

2.1.2. Friction and Lubrication Conditions

In tribological systems, different forms of contact can exist between contacting partners.

Solid friction (dry friction) occurs when two solids have direct contact with each other without a separating layer. If conventional materials are involved, the coefficients of friction and wear rates are high. Lubrication technology attempts to eliminate this condition.

Boundary Friction. The contacting surfaces are covered with a molecular layer of a substance whose specific properties can significantly influence the friction and wear characteristics. Boundary friction layers are of great importance in practical applications when thick, long-lasting lubricant films to separate two surfaces are technically impossible. Boundary lubricating films are created from surface-active substances and their chemical reaction products. Adsorption, chemisorption, and tribochemical reactions also play significant roles.

Fluid Friction. In fluid friction, both surfaces are fully separated by a fluid lubricant film (full-film lubrication). This film is either formed hydrostatically or more commonly, hydrodynamically (Fig. 3). Liquid or fluid friction is caused by the frictional resistance, because of the rheological properties of fluids.

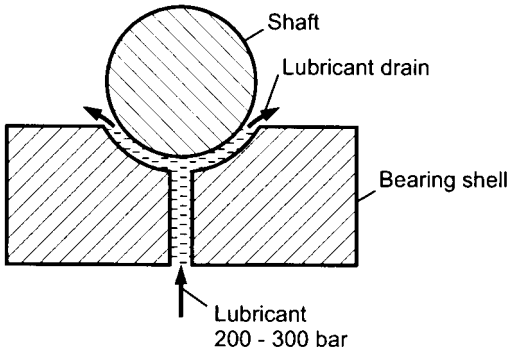

Figure 3. Hydrostatic lubrication as a form of fluid friction

If both surfaces are separated by a gas film, this is known as gas lubrication.

Mixed friction occurs when boundary friction combines with fluid friction. Machine elements which are normally hydrodynamically lubricated experience mixed friction when starting and stopping.

Solid lubricant friction occurs when solid lubricants are used (Chap. 15).

Stribeck Diagram. The friction or lubrication conditions between boundary and fluid friction are graphically illustrated by use of Stribeck diagrams (Fig. 4) [53]. These are based on the starting-up of a plain bearing whose shaft and bearing shells are, when stationary, separated only by a molecular lubricant layer. As the speed of revolution of the shaft increases (peripheral speed) a thicker hydrodynamic lubricant film is created what initially causes sporadic mixed friction but which, nevertheless, significantly reduces the coefficient of friction. As the speed continues to increase, a full, uninterrupted film is formed over the entire bearing faces; this sharply reduces the coefficient of friction. As speed increases, internal friction in the lubricating film adds to external friction. The curve passes a minimum coefficient of friction value and then increases, solely as a result of internal friction. The lubricant film thickness shown in Figure 4 depends on the friction and lubrication conditions including the surface roughness R.

Hydrodynamic Lubrication. In hydrodynamic lubrication the lubricant is pulled into the conical converging clearance by the rotation of the shaft. The created dynamic pressure carries the shaft.

On the basis of the Navier – Stokes theory of fluid mechanics, Reynolds created the basic formula for hydrodynamic lubrication in 1886. The application of the Reynolds' formula led to theoretical calculations on plain bearings. The only lubricant value was viscosity.

Elastohydrodynamic Lubrication (EHD Regime). Hydrodynamic calculation on lubricant films was extended to include the elastic deformation of contact faces (Hertzian contacts, Hertz's equations of elastic deformation) and the influence of pressure on viscosity (Chap. 3). This enables application of these elastohydrodynamic calculations to contact geometries other than that of plain bearings – e.g., those of roller bearings and gear teeth.

Figure 5 shows the elastic deformation of the ball and raceway of a ball bearing.

2.2. Wear (→ Abrasion and Erosion)

Wear is created by the processes of abrasion, adhesion, erosion, tribochemical reactions, and metal fatigue which are important to lubrication technology.

Types of Wear. There are several criteria which could be used to classify the different types of wear. For example, according to the types of (kinematic) friction which lead to wear (sliding wear, rolling wear, fretting wear), according to wear mechanisms (adhesive wear, abrasive wear, tribochemical wear) or according to the shape of the wear particle.

There is also highly specialized wear terminology for different lubricant applications; this is oriented to the geometry of the various bearing faces (e.g., clearance and crater wear in the field of chip-forming tools).

The Wear Process. Wear can be measured gravimetrically, volumetrically or in terms of area over a period of time or against increasing load. Uniformly decreasing wear which stabilizes at a very low level can be described as running-in wear. This can be controlled by the tribochemical reactions of the additives in the lubricants. Wear can occur at relatively constant speed and ultimately lead to the functional failure of the bearing. Wear at an increasing wear rate can lead to progressive wear. In time, the material damage caused by wear will also lead to the failure of a component.

Analysis of machine element failure as a result of wear and lubrication can be performed by use of practically oriented tests. Determination of the effect of different greases on the projected failure of roller bearings has become increasingly important.

3. Rheology of Lubricants

Consistency, flow properties, or viscosity in the case of oils, are key parameters to create lubrication efficiency and the application of lubricants.

3.1. Viscosity (→ Rheometry)

The measure of internal friction in a fluid is viscosity. Viscosity and its dimensions are best

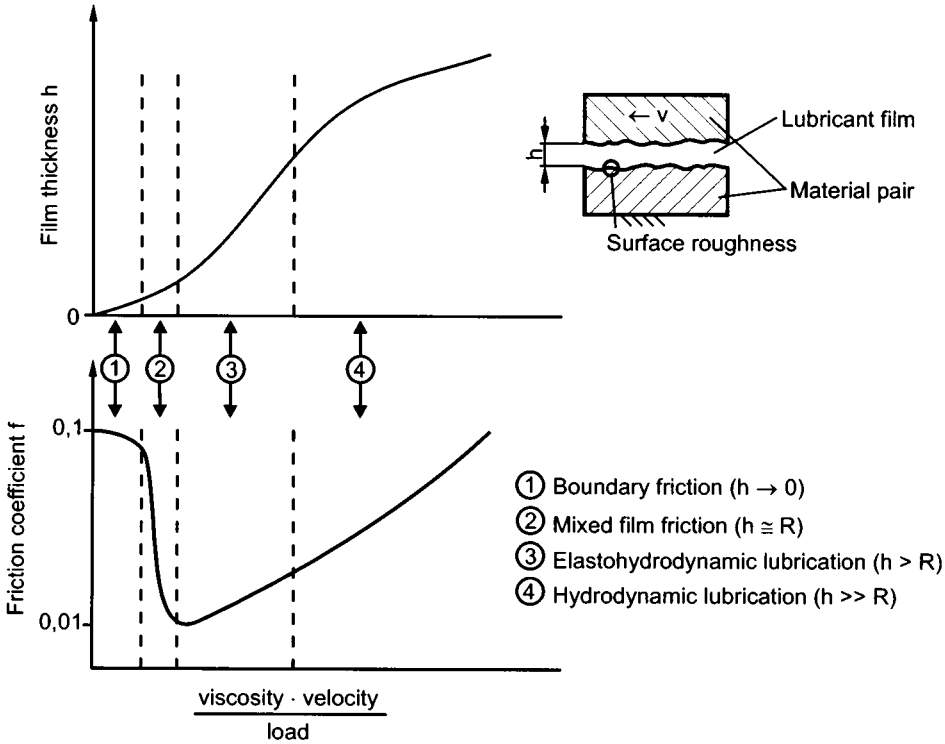

Figure 4. Stribeck graph according to H. CZICHOS and K.-H. HABIG [53]
a) Boundary friction ($h \to 0$); b) Mixed film friction ($h \approx R$); c) Elastohydrodynamic lubrication ($h > R$); d) Hydrodynamic lubrication ($h >> R$)

explained with a model of parallel layers of fluid which could be viewed molecularly. If this packet of fluid layers is sheared (τ), the individual fluid layers are displaced in the direction of the shearing force. The upper layers move more rapidly than the lower layers because molecular forces act to resist movement between the layers. These forces create resistance to shearing and this resistance is given the term dynamic viscosity. The difference in velocity between two given fluid layers, related to their linear displacement, is referred to as shear rate S. This velocity gradient is proportional to the shear stress (τ). The proportionality constant η is called dynamic viscosity and has the unit Pa · s.

$$\tau = \eta \cdot S$$

The laboratory determination of viscosity in run-out or capillary tubes is influenced by the weight of the fluid. The relationship between dynamic viscosity and specific gravity is referred to as kinematic viscosity ν. The following unit analysis applies:

$$\nu = \frac{\eta}{\varrho}\left[\frac{\mathrm{Pa{\cdot}s}}{\mathrm{kg{\cdot}m^{-3}}}\right] \text{ or } \left[\frac{\mathrm{N{\cdot}s{\cdot}m^3}}{\mathrm{m^2{\cdot}kg}}\right] \text{ or } \left[\frac{\mathrm{kg{\cdot}m{\cdot}s^{-2}{\cdot}s{\cdot}m^3}}{\mathrm{m^2{\cdot}kg}}\right] \text{ or } \left[\frac{\mathrm{m^2}}{\mathrm{s}}\right] \left[\frac{\mathrm{mm^2}}{\mathrm{s}}\right] \text{ or (Centistoke)} \tag{3.1}$$

Fluids which display the above proportionality constant between shear stress and shear rate are referred to as Newtonian fluids, i.e., the viscosity of Newtonian fluids is independent of shear rate (→ Fluid Mechanics). Deviations from this Newtonian behavior are sometimes referred to as structural viscosity. Those viscosities are named as apparent viscosities.

Influence of Temperature on Viscosity ($V-T$ Behavior). The viscosity of all oils used for lubrication purposes drops significantly

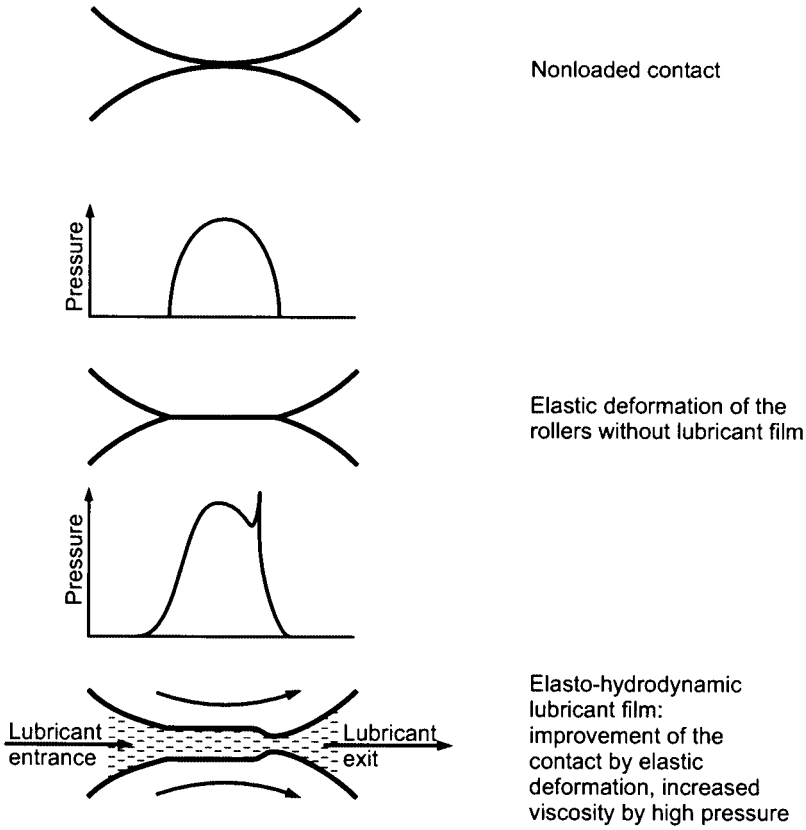

Figure 5. Improvement of hydrodynamic lubrication clearance between two rollers by Hertzian deformation (elastohydrodynamic contact, EHD contact), pressure distribution in the Hertzian contact

when their temperature increases. In linear systems, this $V-T$ behavior is hyperbolic and the practical differentiation necessary in practice is difficult to replicate and the interpolation between two measured viscosities is also problematic. For these reasons, $V-T$ behavior has been allocated to a function which results in a straight-line graph if suitable coordinates are selected (Ubbelohde – Walter equation).

$$\log\log(\nu + C) = K - m \cdot \log T \tag{3.2}$$

In this double-logarithmic formula, C and K are constants, T is temperature in Kelvin and m is the $V-T$ line slope. Figure 6 show the linear and Ubbelohde – Walter double-logarithmic $V-T$ curves of three oils with significantly differing $V-T$ lines (naphthenic oil, paraffinic HC-II / Group III oil / Chap. 4) and a natural vegetable (rapeseed) oil.

The constant C for mineral oils in the $V-T$ equation is between 0.6 and 0.9.

The constant C only plays a minor role in the viscosity calculation, larger differences are only apparent at very low temperatures.

The Vogel – Cameron equation is used for the rapid, computer-based calculation of dynamic viscosity. The Vogel – Cameron viscosity – temperature equation is:

$$\eta = A \cdot \exp\left(\frac{B}{T + C}\right) \tag{3.3}$$

where A, B, and C are constants and T is again temperature in Kelvin.

The m value in the Ubbelohde – Walter equation which represents the double-logarithmic $V-T$ graph slope is sometimes used to characterize the $V-T$ behavior of oils. The m value for lubricant base oils is between 4.5 and 1.1 [54]. The smaller values apply to oils which are less affected by temperature.

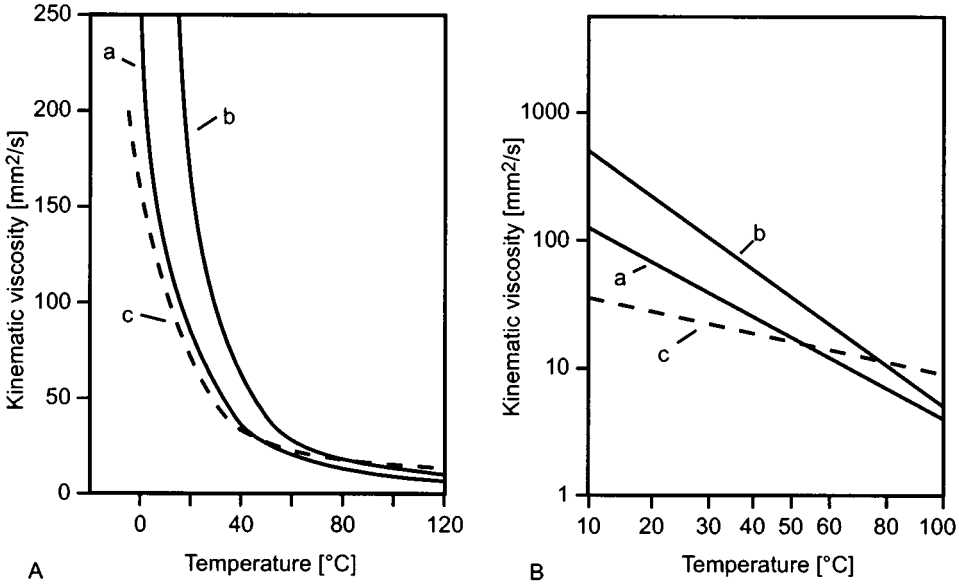

Figure 6. $V-T$ behavior of various oils
A) Linear; B) Double-logarithmic
a) Paraffinic base oil; b) Naphthenic base oil; c) Rapeseed oil

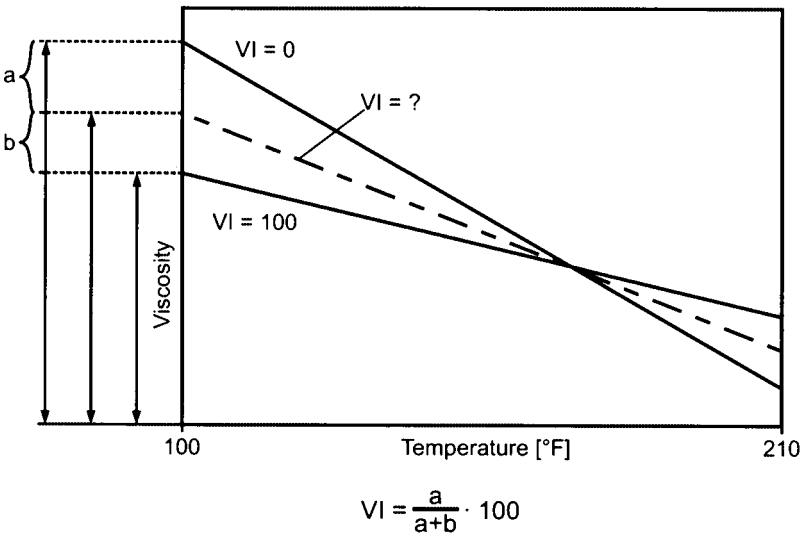

Figure 7. Graphical illustration of viscosity index (VI)

$$m = \frac{\log\log(\nu_1 + 0.8) - \log\log(\nu_2 + 0.8)}{\log T_2 - \log T_1} \qquad (3.4.)$$

The viscosity – temperature constant (VTC) was introduced to better differentiate $V-T$ behavior when the influence of temperature is low.

$$VTC = \frac{\nu_{40} - \nu_{100}}{\nu_{40}} \qquad (3.5)$$

The evaluation of $V-T$ behavior at low temperatures according to Ubbelohde – Walter lines or other straight-line $V-T$ graphs often leads to inaccuracies. These dependencies do not apply to

base oils which can suffer thickening caused by crystallization of some components (e.g., paraffins) at low temperatures or those whose polymer molecules simulate viscosity effects at low temperatures.

Viscosity Index. The viscosity index *VI* is far more important for the description of viscosity – temperature behavior than the *m* value or the viscosity – temperature constant. This value was first introduced in the USA in 1928. It was based on the then greatest ($VI = 100$) and the smallest ($VI = 0$) temperature dependence of US base oils.

VI is graphically illustrated in Figure 7. Values over 100 are calculated as *VIE*.

Viscosity index is also defined in the most international standards.

Table 1 shows various $V-T$ characteristics for a number of oils.

Viscosity – Pressure Dependency. The significance of viscosity – pressure dependency ($V-p$ behavior) was, and still is, underestimated for numerous lubrication applications. $V-p$ behavior has become a part of the calculation of elastohydrodynamic lubricant films.

$V-p$ behavior can be described by the formula:

$$\eta_{\mathrm{P}} = \eta_1 \cdot \mathrm{e}^{\alpha(p-p_1)} \tag{3.6}$$

where η_{P} is the dynamic viscosity at a pressure p, η_0 is the dynamic viscosity at 1 bar and α is the viscosity – pressure coefficient:

$$\alpha = \frac{1}{\eta_{\mathrm{P}}} \cdot \left(\frac{\mathrm{d}\eta_{\mathrm{P}}}{\mathrm{d}p}\right)_{\mathrm{T}} \tag{3.7}$$

$V-p$ dependence is defined by the chemical structure of the substances with the steric geometry of the molecules being of particular significance.

Figure 8 and 9 show viscosity – pressure relationship for a number of oils with different chemical structures.

α can increase significantly with falling temperature which has an exponential effect on viscosity.

Viscosity – Shear Rate Relationship. The viscosity of Newtonian fluids is a constant (proportionality factor) between the shear force τ and the shear rate S and not dependent on shear rate. Lubricants which display a dependence on shear rate are known as non-Newtonian or fluids with structural viscosity.

Oils containing polymers with specific additives or thickeners and mineral oils at low temperatures (long-chain paraffin effects) display such structure-viscous behavior. At normal application temperatures, most major lubricant base oils such as hydrocarbon oils (mineral oil raffinates or synthetic hydrocarbons), synthetic esters and natural fatty oils can withstand very high shearing forces (e.g., $10^9\ \mathrm{s}^{-1}$) as found in highly loaded machine elements (e.g., gearboxes) and are independent of shear rate.

Engine oils containing polymer $V-T$ improvers (Chap. 8) or polymer ash-free dispersants display structural – viscosity effects at low and high temperatures.

While the reduction in viscosity caused by the shearing of structure-viscous fluids is reversible, polymer-based oils can suffer a permanent reduction in viscosity. In these cases, the shearing forces lead to a mechanical change or reduction in the size of the polymer molecules so that their desired effects are minimized. These effects have been observed, in particular, with multigrade engine oils and high-*VI* hydraulic oils.

3.2. Special Rheological Effects

Apart from the above-described effects, lubricants are subject to further rheological phenomena and in particular, colloidal systems consisting of solid or fluid dispersions (solid dispersions or emulsions).

Even small mechanical loading such as vigorous stirring can cause a system to change completely, e.g., pasty systems can break down into low-viscosity systems (Fig. 10). If continuous mechanical load causes apparent viscosity to fall over time and the original viscosity is restored after a certain rest period, such fluids are called thixotropic. This effect is used for sheet forming lubricants (Chap. 13).

Time-based viscosity changes can also be caused by the separation of colloidal particles such as paraffins from mineral base oils, or additives during the storage of lubricants. This undesirable effect can occur during the cooling

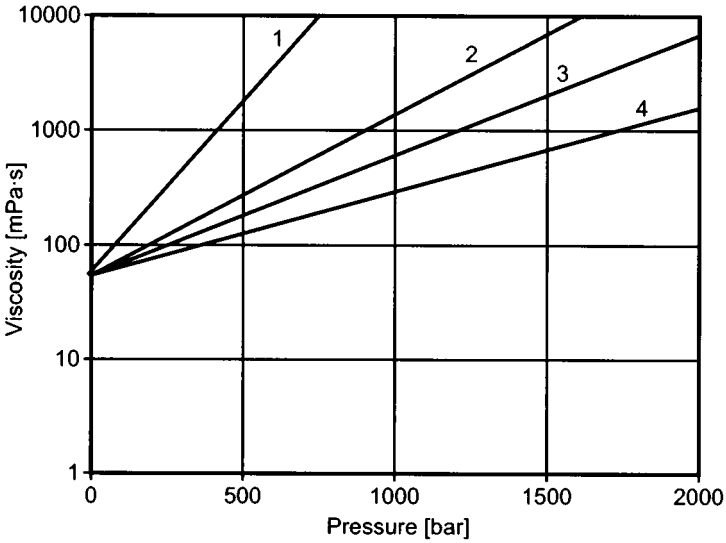

Figure 8. Viscosity – pressure relationship of various oils
1) Aromatic oil; 2) Naphthenic oil; 3) Paraffinic oil; 4) Biodegradable polyol ester

Table 1. Various $V-T$ characteristics for several oils [54]

	Kinematic viscosity mm^2/s		Viscosity index	*m* Constant	Viscosity temperature constant
	40 °C	100 °C	(*VI*)		(*VTC*)
Naphthenic spindle oil	30	4.24	40	4.05	0.847
Paraffinic spindle oil	30	5.23	105	3.68	0.819
Medium solvent extract	120	8.0	−50	4.51	0.939
Medium polyalkylene glycol	120	20.9	200	2.53	0.826
Medium silicone oil	120	50.0	424	1.14	0.583
Multigrade motor oil (SAE 10W-30)	70	11.1	165	2.82	0.841
Ester oil	30	5.81	140	3.40	0.806

phase following production or by long-term storage at low temperatures. This separation process can also be independent of temperature and be caused by the solubility status of additives in the base oils.

Greases. The rheological description of greases is also complex because of their complex system structure (fluids, soap thickeners, solid thickeners, additives, Chap. 14). The mathematical description of apparent viscosity has not gained general acceptance in practice.

At lower shear loads, greases behave similarly to structure-viscous substances but at high shear loads, like Newtonian base oils.

3.3. Viscosity Grades

To simplify the classification of lubricants according to their application, viscosity grades were introduced which are now internationally accepted. ISO viscosity grades apply to industrial lubricants while Society of Automitive Engineers (SAE) classifications apply to automotive engine and gear oils (see Fig. 11).

ISO Viscosity Grades. 18 viscosity grades are laid down in the ISO standard 3448. Over the range from 2 to 2500 mm^2/s, these are the international standard number series E6 rounded to whole numbers when the 6 numerals correspond to one power of ten (the first and fourth power of 10 are reduced). The viscosity grades were also

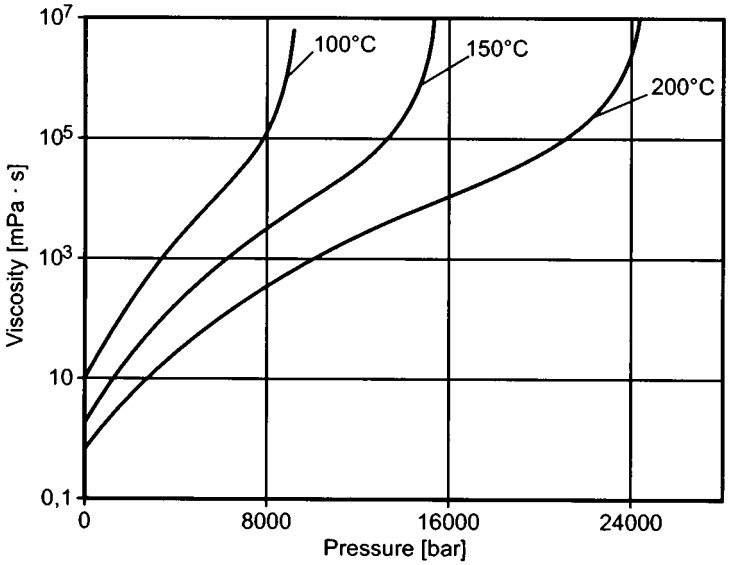

Figure 9. Increase of the viscosity – pressure coefficient with falling temperature according to H. HOLLAND [55]

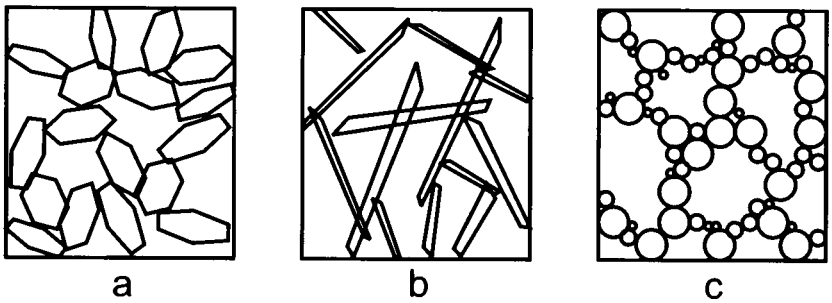

Figure 10. Various lubricant structures (a, b, c) with high sensitivity against shear stress

adopted into or added to national standards such as ASTM or DIN.

Viscosity grades are not used for all industrial lubricants. Particularly oils for chip-forming and chip-less metalworking processes are not classified in this way.

Apart from the viscosity grades, ISO 3448 defines tolerances as well as median viscosities.

Other Viscosity Grades.

Engine Oils. To define the viscosity of engine oils, two or three viscosity thresholds were selected to define flow properties at low temperatures and to define a minimum viscosity at high temperatures. Maximum viscosity at low temperature should ensure the rapid oil circulation to all lubrication points and permit a sufficiently higher cranking speed for starting and the minimum viscosity at 100 °C should ensure that adequate lubrication of the bearings occurs at high temperatures. The classification system was introduced by the SAE together with the American Society for Testing and Materials (ASTM) and is used throughout the world. Low temperature viscosity is measured as dynamic viscosity with a specially constructed rotational viscosimeter (cold cranking simulator) at low shear rates (Chaps. 8 and 11).

Automotive Gear Oils. Specific SAE viscosity grades have been created for automotive gearbox, axle and differential oils. Compared with engine oils, the low temperature behavior of these oils is more heavily weighted.

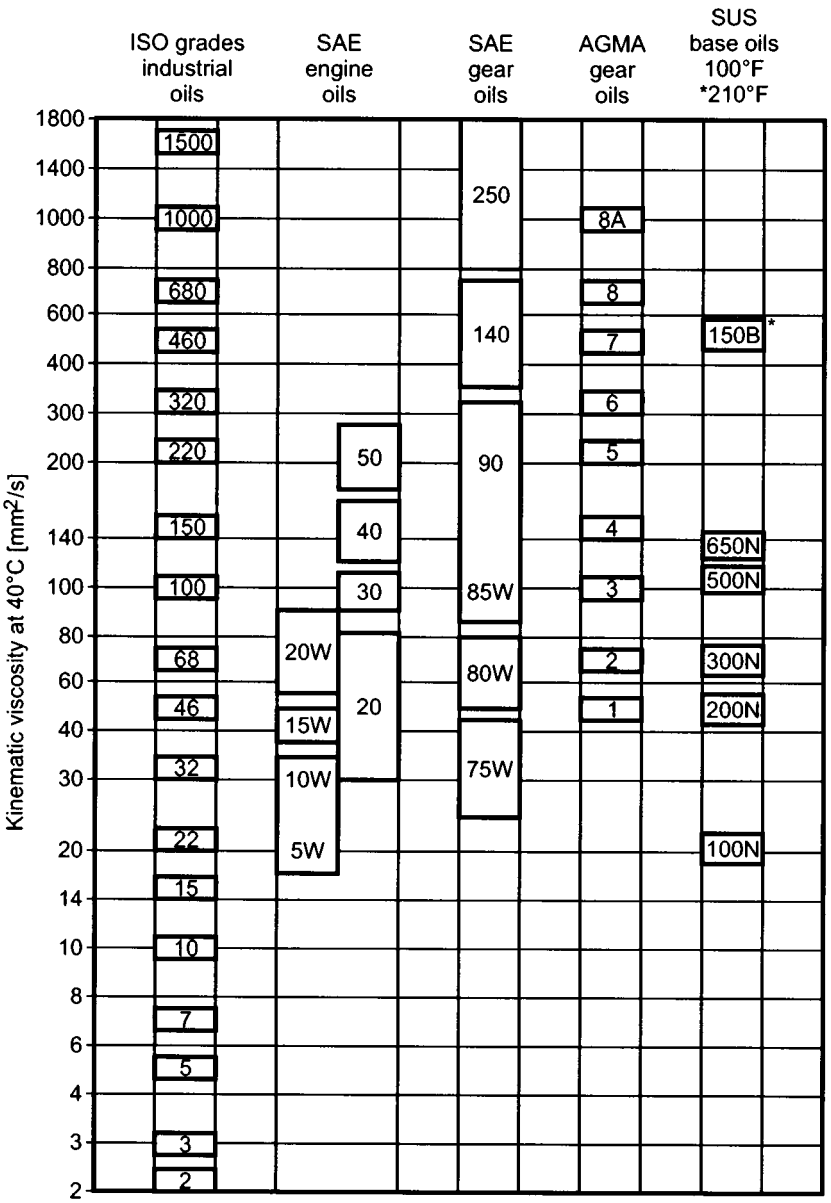

Figure 11. Comparison of viscosity grades for various applications

Industrial Gear Oils. AGMA (American Gear Manufacturers Association) defines nine viscosity ranges for industrial gear oils.

Viscosity Grades for Base Oils. Mineral base oils are traditionally classified according to Saybolt Universal Seconds (SUS). A 150 N base oil shows a viscosity of 150 SUS at 100 °F.

4. Base Oils

In terms of volume, base oils are the most important components of lubricants, they account for more than 95 % of lubricant formulations. There are lubricant families (e.g., some hydraulic and compressor oils) in which chemical additives

only account for 1 % while the remaining 99 % are base oils. On the other hand, other lubricants (e.g., some metalworking fluids, greases, or gear lubricants) can contain up to 30 % additives.

4.1. Historical Review and Outlook

Although the most important requirement of base oils in the 1950s was the correct viscosity and the absence of acidic components, base oils in the 1960s were downgraded to solvents or carriers for additives in the euphoria surrounding chemical additives. In the 1970s, there was a realization that some synthetic fluids with uniform basic chemical structures offered performance superior to that of mineral base oils. At that time, the considerably higher price of these products hindered their market acceptance. In the 1980s however, lower-price, quasi-synthetic hydrocracked oils were introduced in Western Europe which closely matched the properties of synthetic hydrocarbons (Shell, BP, Fuchs). In the 1990s, base oil developments were influenced by the ever-increasing demands on lubricant performance, and by environmental and health and safety criteria. This led to chemically pure oils such as hydrocracked products, polyalphaolefins and esters gaining acceptance. Natural fatty oils, particularly their oleochemical derivatives, have experienced a renaissance because of their technical characteristics but, above all, because of their rapid biodegradability.

The trend towards ever-greater performance and even better environmental compatibility will continue in the first decade of the new millennium. The significantly higher price of the new lubricants, which will be increasingly characterized by their base oils and less so by their chemical additives, will probably be accepted by users who will benefit from long product life and lower overall system costs. In Western Europe in 1998 [56], about 10 wt % of base oils were synthetic products (including Group 3 hydrocracked oils (Section 16.1). These, however, represented about 40 % of base oil value. In Germany, furthermore, ca. 5 % of lubricant base oils were rapidly biodegradable (natural and synthetic) esters in 1999. It is safe to assume that in Western Europe at least, synthetic and quasi-synthetic oils will in terms of value, have overtaken traditional mineral base oils by the year 2005.

4.2. Chemical Characterization of Mineral Base Oils

Crude oil generally consists of many thousands of single components and these are reflected in the processing of each fraction. Thus, mineral oil fractions have always been characterized by defining their technical properties or to identify and quantitatively determine groups of components with similar chemical character. Advanced physicochemical methods are, however, increasingly being used in routine testing.

Rough Chemical Characterization.
Viscosity – Gravity Constant (VGC). This value enables only rough chemical characterization of oils. Values near 0.800 indicate paraffinic character whereas values near to 1.000 point to a majority of aromatic structure (ASTM 2501-91)
Aniline Point. The aniline point is also used to characterize the hydrocarbon structure of mineral oils (see Section 16.2.8).

Carbon Distribution. The most important means of analysis for characterization of mineral oil hydrocarbons is the determination of carbon in terms of its three categories of chemical bond – aromatic (C_A), naphthenic (C_N), and paraffinic (C_P).

N – D – M analysis uses physicochemical data which are easy to obtain. These include refractive index, density, and molecular mass. Molecular mass can be determined by measuring the viscosity at different temperatures (e.g., ASTM D 2502). Carbon distribution is given in $\%C_A$, $\%C_N$ and $\%C_P$ (100 % in total). N – D – M analysis also determines the average total number of rings per molecule (R_T) and the breakdown into aromatic and naphthenic rings (R_N) per molecule ($R_N = R_T - R_A$).

Brandes [57] created a method of determining carbon distribution according to specific bands in the infrared spectrum (see Section 16.2.4). Exact determination of aromatic carbon content can be performed by NMR (ASTM 5292).

Hydrocarbon Composition. To determine the hydrocarbon composition of the base oil, the componenets are first separated by chromatography and the fractions are then subjected to advanced analytical procedures.

To differentiate mineral oils in the boiling range from 200 to 550 °C, high ionizing voltage mass spectrometry (ASTM D 2786-91) is used for saturated fractions and ASTM 3239-91 for aromatic fractions. Saturated fractions are separated into alkanes (0-ring), 1-ring, 2-ring, 3-ring, 4-ring and 5-ring naphthenes. The aromatic fractions are subdivided into seven classes: monaromatics, diaromatics, triaromatics, tetraaromatics, pentaaromatics, thiophenoaromatics, and unidentified aromatics.

Polycyclic Aromatics in Base Oils. Polycyclic aromatics (PAH, polycyclic aromatic hydrocarbons, or, in general, PCA, polycyclic aromatics) formed by the combustion of gasoline can accumulate in engine oils. They can also accumulate in quenching oils after long periods of heavy-duty use.

In traditional solvent refining processes, PAH largely remain in the extract. Nonsolvent extracted distillates contain PAH in line with their boiling point.

The carcinogenic characteristics of nonseverely treated distillates in the petroleum industry was established by the International Agency for Research on Cancer (IARC) in 1983 [58]. This led to considerable limitations in the manufacture and application of naphthenic base oils and the ending of the use of aromatic extracts as lube base oils.

PAH can be determined by HPLC with anthracenes or other aromatics as markers or by GC – MS after appropriate sample preparation.

The IP 346 method, adopted into national legislation in several countries, does not analyze PAH directly but an extract is obtained in dimethyl sulfoxide (DMSO), in which PAH accumulate. As a rule, the extract largely contains naphthenes or mono-aromatics. According to IP 346, DMSO extracts contain only 0.1 % PAH. After numerous skin-painting tests on mice, the carcinogenicity of petroleum products corresponds to the percentage of DMSO extract.

4.3. Refining (→ Oil Refining)

Lubricant refineries are divided into integrated and nonintegrated plants (Fig. 12). Integrated refineries are linked to primary crude oil refineries and are fed with vacuum distillate by pipeline. Nonintegrated refineries purchase vacuum distillate on the open market or buy atmospheric residues and perform their own vacuum distillation. Occasionally they perform vacuum distillation on crude oil.

4.3.1. Distillation

By way of fractional distillation, products are removed from crude oil which approximately meet the viscosity grades ultimately required. Often only four or five cuts suffice to fulfil lubricant requirements. The viscosity of the primary vacuum distillate is not identical to that of the finished base oils formed in hydrocracking processes.

After the corresponding separation of the lighter components from the crude oil by atmospheric distillation, the lubricant components are in the atmospheric residue. Figure 13 shows the yields of the various cuts in conventional lube oil refining with the corresponding boiling ranges of a typical lube crude. The atmospheric residue is now subjected to vacuum distillation to remove the components required for lubricants.

4.3.2. Deasphalting (→ Oil Refining, Chap. 3.6.5.1.)

The vacuum residue still contains highly viscous hydrocarbons which can supply valuable components for lube base oils. These highly viscous base oils, commonly known as brightstocks, are separated from the asphaltenes by extraction. Brightstocks are produced in lube oil refineries when the use of the asphaltene byproduct (hard asphalt) is worthwhile. The quality of the hard asphalt for the manufacture of high quality bitumen depends on the crude oil. Deasphalting is usually carried out with propane as solvent. Brightstocks can be manufactured with viscosities of more than 45 mm^2/s at 100 °C.

Figure 14 shows a flow sheet for the manufacture of suitable feeds for lube oil refining by distillation and propane deasphalting.

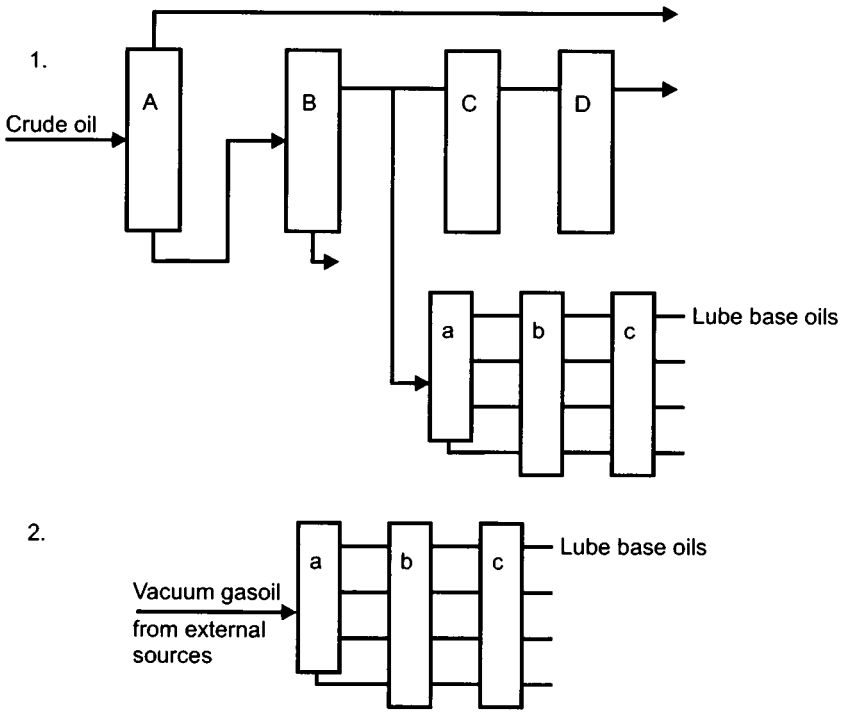

Figure 12. Integrated (1) and not integrated (2) lube refinery
A) Atmospheric distillation; B) Vacuum distillation; C), D) Processing of vacuum distillates for non-lube production
a) Fractionating vacuum distillation; b), c) Lube refining processes

4.3.3. Traditional Refining Processes

The vacuum distillates still contain components which can detrimentally affect aging, viscosity – temperature behavior and flowing characteristics, and components which are hazardous to health. To eliminate these compounds several refining methods were developed, of which solvent refining has become the most accepted method over the past few decades. New plants increasingly use hydrotreatment.

4.3.3.1. Acid Refining

Acid refining has become less popular because the acid sludge waste produced is difficult to dispose of and this method has been replaced by solvent extraction. Acid refining is still used to some extent for the re-refining of used lubricating oils (Chap. 18) and for the production of very light-colored technical or pharmaceutical white oils and petroleum sulfonates as byproducts.

When the distillates are treated with concentrated sulfuric acid or fuming sulfuric acid (oleum), substances which accelerate oil aging are removed. *Oleum treatment* (*wet refining*) not only readily removes reactive oil components such as olefins but also reduces the aromatic content, which in turn increases the viscosity index of the product. Reactions of oleum with saturated paraffinic structures lead to refining losses. Acid-refined oils require complex neutralization and absorption follow-up treatment to remove all traces of acid and undesirable byproducts.

4.3.3.2. Solvent Extraction

Solvent extraction (Fig. 15) creates base oils which are known as solvent raffinates or solvent neutrals (SN). Extraction processes using solvents create both a base oil and, after evaporation of the solvent, an aromatic-rich extract. The selectivity of the extraction media for aromatics is an important selection parameter.

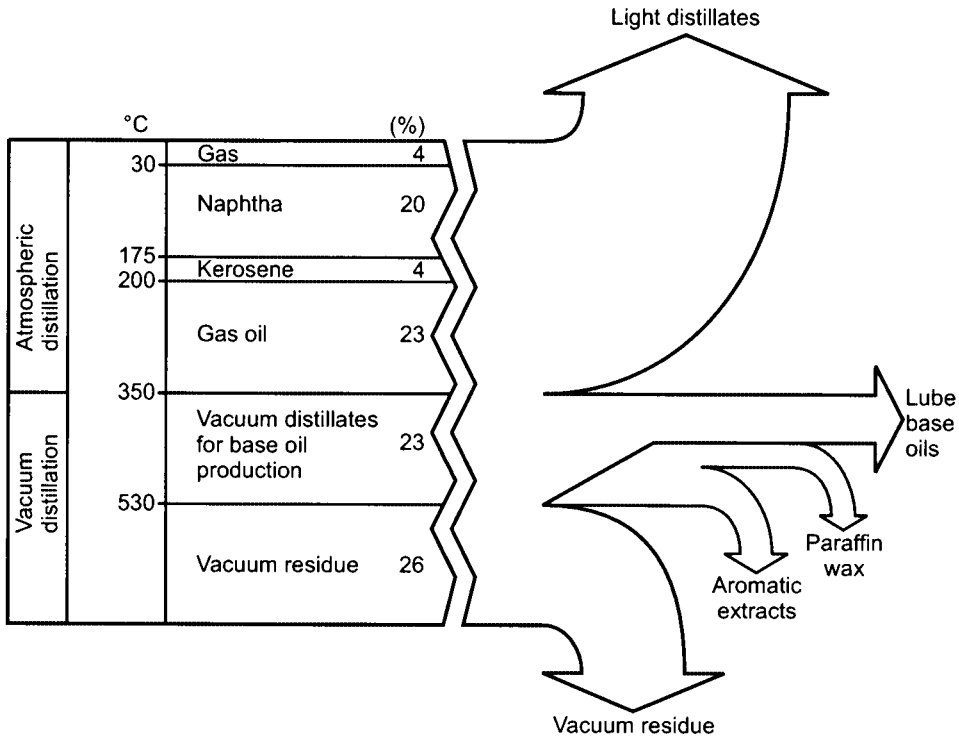

Figure 13. Yield of the various cuts in conventional lube oil refining of a typical lube crude

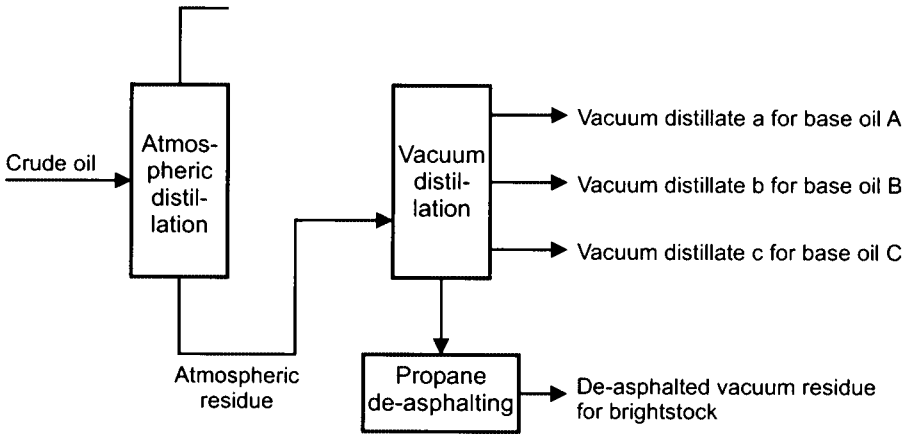

Figure 14. Flow sheet for the manufacture of suitable feeds for lube oil refining by distillation and propane deasphalting

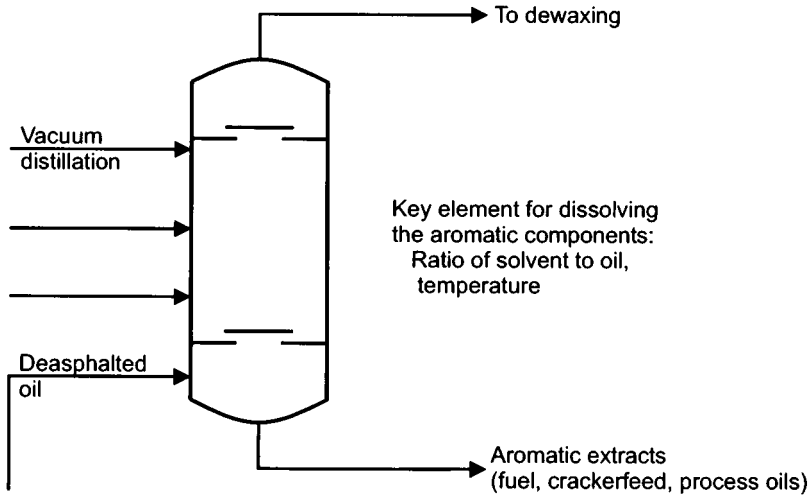

Figure 15. Solvent extraction

In particular the selectivity towards polycyclic aromatics has attracted attention because of the carcinogenicity of these compounds. Common extractants are furfural, *N*-methyl-2-pyrrolidone (NMP), and phenol. Sulfur dioxide should also be mentioned because of its historical importance as B. EDELEANU introduced extraction to petroleum technology in 1912. In 1999, sulfur dioxide was only used in very few refineries (one in Germany until 1999) for the refining of naphthenic distillates.

In recent years several furfural extraction plants have been converted into NMP plants, and even phenol plants have taken a back seat. NMP is a nontoxic solvent and can be used in a low solvent-to-oil ratio with high selectivity. This generates significant energy savings. NMP in new plants results in physically smaller units and thus lower capital expenditure.

Depending on the crude, the extract part of paraffinic oils can be 30 to 50 %. As the standard requirement of solvent neutral oils is a viscosity index of at least 95, the extraction severity is matched to this demand.

A higher proportion of aromatic and naphthenic hydrocarbons in the distillates requires greater extraction severity and thus a larger quantity of extract. The percentage share of the extract is a major economic factor in conventional lube refining. Aromatic extracts have a very high viscosity – pressure coefficient (Chap. 3) and in the past were often used in metal-forming operations, such as cold extrusion.

A 1-point increase in viscosity index as a result of greater extraction severity creates, on average, 1 % more extract. Some refineries produce so-called semiraffinates with *VI*s of between 70 and 80. It is cheaper to adjust the *VI* later with *VI* improvers than to produce 15 or 20 % more extract. These semiraffinates, however, have disadvantages resulting from the lower extraction severity – e.g., lower oxidation stability, possibly higher amounts of sulfur or higher amounts of polycyclic aromatics.

Solvent extraction is generally only economic to a minimum *VI* of 50 of the vacuum gas oil [59].

4.3.4. Solvent Dewaxing (→ Waxes, Chap. 4.2.3.1.)

In traditional refining processes, solvent extraction is followed by solvent dewaxing. Long-chain, high melting point paraffins (waxes) crystallize at low temperature and thus negatively affect the cold flow properties of lube oil distillates and lead to a high pour point. Their removal has, therefore, been an important consideration since the beginning of crude oil refining.

Dewaxing by crystallization of paraffins at low temperatures and separation by filtration are

the principal processes in traditional refining. Compared with catalytic dewaxing with hydrogen, urea dewaxing to separate *n*-paraffins is of relatively minor importance in lube oil refining. Crystallization methods involve mixing the solvent with the oil; this improves filtration, as a result of dilution, and promotes the growth of large crystal formations.

Solvents used for dewaxing are ketones and chlorinated hydrocarbons. *Ketone dewaxing* (dimethyl ketone or methyl ethyl ketone, MEK) is normally used for pour points down to −12 °C. For lower pour points, the *Di–Me* (dichloroethane – dichloromethane mixtures) method is used. This also enables the manufacture of hard and soft waxes. These waxes are only valuable products if their quality and processing is good and they can then be sold, e.g., as candle wax or for coatings. They are often used as cracker feed in fluid catalytic crackers (FCC) and the cracker yield is then assessed by use of special formulas.

Figure 16 illustrates the conventional solvent dewaxing process.

4.3.5. Finishing

A finishing stage often follows extraction and dewaxing. In the past, methods which employed absorbents were often used but these days predominantly hydrofinishing processes are used. Finishing should improve the color of the product and remove surface-active substances which can negatively affect the air-release or demulsifying properties of a lubricating oil.

In general, the hydrofinishing process is referred to as mild hydrotreating and focuses on improving color, odor, and ultraviolet stability. *Ferrofining* (BP) has achieved major economic importance as a finishing process. The process parameters do not generally lead to desulfurization. More severe processes with pressures in excess of 100 bar can bring about significant desulfurization and some dearomatization.

Additional finishing with absorbents (bleaching clays, bauxite) is sometimes used for the manufacture of refrigerator, transformer, or turbine oils. The subsequent filtration process in filter presses or other filtering equipment represents additional complexity and the disposal of the filter residue is an increasing problem.

Lube Crudes. The finished base oils have a high sulfur content, especially after conventional solvent refining. Hydrogenating processes almost fully eliminate sulfur but desulfurization consumes hydrogen.

The vacuum gase oils (VGO) are the direct feed for lube refining. Their *VI*, wax content, and sulfur content is especially important. A high *VI* leads to low extraction losses in solvent extraction and low hydrogen consumption during hydrogenation. High wax content increases production costs.

Table 2 shows the *VI* and wax content of VGO from various lube crudes.

Table 2. *VI* and wax content of vacuum gas oil (VGO) from various crudes

Crude	*VI* after solvent dewaxing	Wax in VGO (wt %)
Pennsylvania	100	0
Ordovician	85	13
Brent	65	19
Arab light	60	9
Arab heavy	55	9
Iranian light	55	16
Lagomedio	45	10
Urals	40	11
Iranian heavy	40	16
ANS	15	8

4.4. Base Oil Manufacturing by Hydrogenation and Hydrocracking

(→ Oil Refining, Chap. 3.3., → Oil Refining, Chap. 3.5.)

Hydrogenation and hydrocracking in the manufacture of lubricant base oils significantly influence the chemical structures of mineral oil molecules. Unstable molecules are chemically stabilized by the removal of the heteroatoms (sulfur, oxygen, nitrogen) and severe hydrogenation can convert aromatics into saturated naphthenic or paraffinic structures. In addition to the hydrogenation process, hydrocracking breaks down or cracks larger molecules into smaller ones. Larger molecular structures can re-form from small fragments. The principal process criteria are temperature, pressure, the catalyst and space velocity. If special conditions are met, a focal point of the process is the isomerization of paraffinic structures. Beside the saturation of

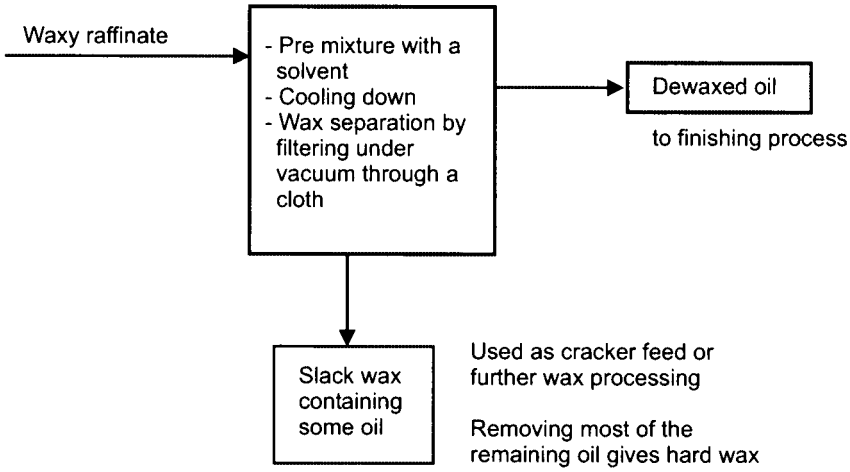

Figure 16. Conventional solvent dewaxing process

aromatics, opening of the naphthenes rings can occur.

Lubricant base oils can be much more easily tailored using these processes than is possible with simple solvent refining separation. The future of lube base oil production thus lies with hydrogenation and hydrocracking. An additional advantage of advanced hydrocracking is the lower dependence on the quality of the crude oil.

In addition to hydrocracking and hydroisomerization, the following typical chemical processes occur during hydrogenation (Table 3).

While the processes mentioned are generally replacing solvent extraction for the manufacture of base oils, there are hydrogenation processes which either crack long-chain paraffins with specific catalysts into light products and remove them from the base oil or convert them into isoparaffins with good low-temperature characteristics without a significant loss in yield. If the previously mentioned hydrogenation processes are combined with this type of catalytic hydrogenating dewaxing, this is called the 'all-hydrogen route'.

4.4.1. Manufacturing Naphthenic Base Oils by Hydrogenation

Only about 10 % of the petroleum base oils used in lubricants are naphthenic. Before 1980, naphthenic oils were significantly more important (USA 25 %). In 1983, a publication by the IARC defined a number of mineral oil products as carcinogenic, among them naphthenic oils. Although the IARC publication and subsequent legislation and classification requirements led to the disappearance of this type of oil from important markets, naphthenic base oils were still required for some lubricant applications (greases, metalworking oils, and refrigerator oils) because of their outstanding solubility and these are now manufactured with severe hydrogenation which produces noncarcinogenic base oils. Figure 17 shows a flowchart for the manufacture of naphthenic base oils. Severely hydrotreated Group C distillates are the most accepted naphthenic base oils in western markets.

4.4.2. Production of White Oils

White oils – light-colored, odorless, low-aromatic or aromatic-free mineral oil raffinates – are used in a number of applications as medical and pharmaceutical white oils or technical white oils in, above all, the food and beverage industries as food-grade lubricants. BASF Germany was the first to develop white oil manufacturing from the high-pressure hydrogenation technology used to liquefy coal. Up to that point, white oils were manufactured exclusively by acid treatment.

Table 3. Typical chemical processes during hydrogenation

Desulfurization	Mercaptans $R-SH \rightarrow RH + H_2S$ Disulfides $R-S-S-R \rightarrow RH + H_2S$ Sulfides $R-S-R \rightarrow RH + H_2S$ Thiophene 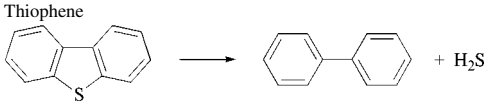
Denitrogenation	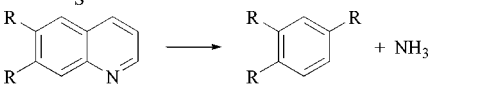
Polyaromatic saturation Viscosity Index: (a) −60; (b) 20 Pour point: (a) +50 °C; (b) +20 °C	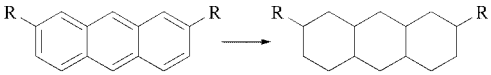 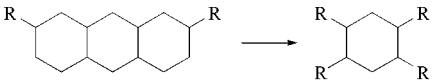
Naphthenic ring opening Viscosity Index: (a) 20; (b) ∼130 Pour point: (a) +20 °C ; (b) −10 °C	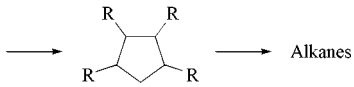
Isomerization Viscosity Index:(a) 125; (b) 125 Pour point: (a) +20 °C; (b) −40 °C	

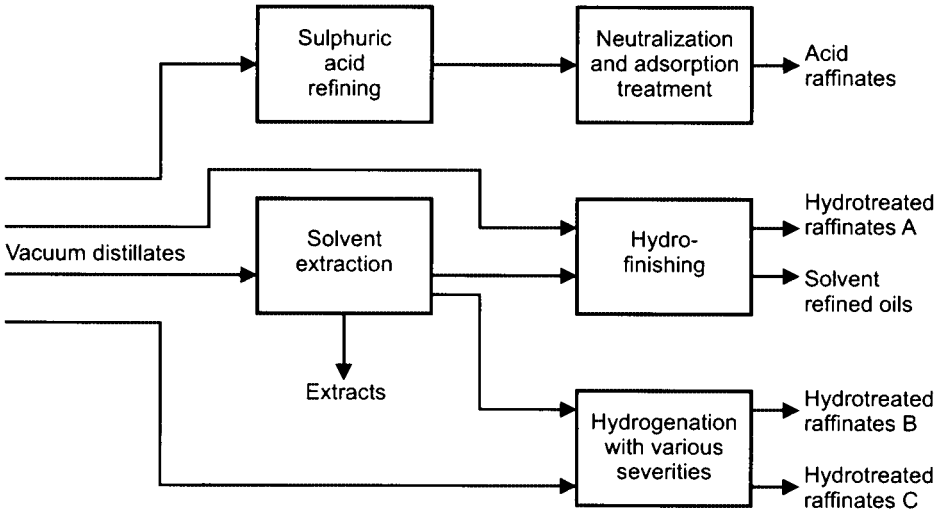

Figure 17. Manufacture of naphthenic base oils

The feeds for hydrogenated white oil manufacturing are generally naphthenic or paraffinic solvent-treated vacuum distillates or untreated vacuum distillates. In the BASF process hydrogenation takes place in two stages: the first at 300 – 380 °C and 80 – 200 bar with a NiMo catalyst and the second at 200 – 300 °C and 100 – 200 bar with an Ni catalyst. The first stage produces the technical qualities (according to the U.S. Food and Drug Administration) while the second stage produces the medical grades (e.g., German Pharmacopeia). Figure 18 shows a flow chart for white oil manufacture [59].

4.4.3. Lube Hydrocracking

The principal elements of the lube hydrocracking process are the cracking of low-*VI* components and the saturation of aromatics.

Hydrocracking can be performed with the principal objective of producing lubricating oils and to convert vacuum gas oil into high-grade fuels. Fuel hydrocracker residues are excellent feeds for the manufacture of lubricant base oils and hydrocrackers used to make olefin feeds for steam crackers can supply a premium quality feed for lube base oils. The severity has the most important effect on the quality of the base oil produced. High severity (e.g., 80 % light products) generates a high-*VI* and low-evaporation oil. While hydrocracking refineries for base oils with normal *VI*s and normal Noack evaporation (high viscosity index, HVI oils, Group II oils, Section 16.1) at moderate severity are being built in the USA and the Far East, based mainly on Chevron Technology, Western European hydrocracking plants were designed for low-evaporation base oils (Shell, Petit Couronne, France, 1972; Union Fuchs, Wesseling, Germany, 1986 and BP, Lavera, France). These produce Group III oils, both from waxes and from vacuum gas oils with hydrocracker residues.

Hydrocracked base oils differ from solvent-extracted oils by their extremely low aromatic content and their chemical purity, i.e., only traces of heteroatoms such as sulfur, nitrogen, or oxygen are present. Furthermore, Group III oils which were manufactured at high severity or from waxes have Noack evaporation characteristics which are about 50 % lower than equiviscous, solvent-extracted vacuum distillates. If catalytic dewaxing is performed, even lower pour points can be achieved than with SN oils.

Since 1995, more than half of all new base oil manufacturing plants have been built or are planned using hydrocracking technology.

The increasing number of fuel hydrocrackers, which are a response to increasing demands for low-sulfur oils (diesel fuels with < 50 ppm sulfur), represents a growing source of hydrocracked feeds. Fuel hydrocracker residue can be processed into high-quality, hydrocracked base oils by distillation and subsequent dewaxing (wax isomerization) at acceptable cost.

During the hydrocracking process, especially if vacuum gas oils are severely treated to create very high viscosity index (VHVI) oils, polycyclic aromatics can be formed along with saturated structures (aromatics to naphthenes or paraffins and isoparaffins) under some process conditions. These have to be removed by subsequent high-pressure hydrogenation or by extraction if hybrid methods are used. This subsequent hydrogenation can also be manipulated so that it also produces a significant increase in *VI*.

4.4.4. Catalytic Dewaxing

The most complex stage of traditional base oil refining is solvent dewaxing. In addition to the high capital expenditure and, above all, operating cost, the solvent-related limits to the achievable pour point are a disadvantage.

Thus, various methods have been developed which can remove unbranched and less branched paraffins, and some other petroleum components by catalytic and hydrogenation reactions or convert them into components which improve the low-temperature characteristics of the base oil.

The first technologies were based on the catalytic cracking. In 1979, Mobil introduced the *Mobil lube dewaxing process* (MLDW). The development of catalysts since then has led to strong hydroisomerization activity processes (*Mobil selective dewaxing*, MSDW).

While the MLDW process mainly involves the cracking of long-chain paraffins and the production of larger quantities of light byproducts and poor base oil yields, MSDW processes create high-yield, nonwaxy isoparaffins from the waxy parts of the feed.

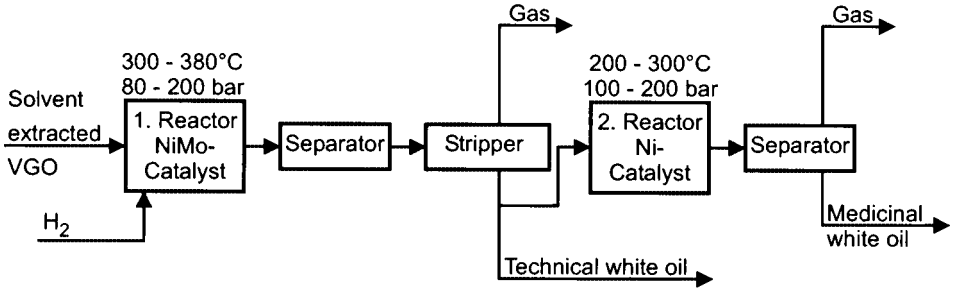

Figure 18. Flow chart for white oil manufacture (BASF)

At the beginning of the 1990s, Chevron successfully introduced its *isodewaxing* process. In 1999 it was the most important technology for catalytic dewaxing [59]. The process combines relatively high yields with low pour points and high *VI*s. The method uses an intermediate-pore silicoaluminumphosphate molecular sieve (e.g., Pt on SAPO-11) as catalyst. In terms of process economy, high base oil yields but also theformation of mainly high-value C_{5+} liquids as by-products are of significance.

Dewaxing catalysts must possess the right balance between hydrogenation activity and acid activity. Increasing hydrogenating properties usually leads to a reduction in paraffin isomerization and thus a rise of the pour point. Higher acid activity increases cracking and thus yield losses. Sulfur and nitrogen in the feed play important roles. Nitrogen is detrimental to acid activity and sulfur is catalyst poison.

Isodewaxing in principle enables manufacture of lubricant base oils with pour points below $-45\,^{\circ}C$, however, dewaxing to very low pour points (depending on process conditions and catalysts) leads to *VI* losses.

A major success of this new dewaxing process are the molecular sieve catalysts – zeolites with exactly defined mesh sizes. Although highly branched paraffins or polycyclic naphthenes or aromatics are not affected by the zeolites, long-chain less or unbranched molecules are isomerized by the catalysts [60–62].

Figure 19 shows the flow chart for catalytic dewaxing (hydrodewaxing) including a hydrofinishing reactor.

4.4.5. Wax Isomerization

VHVI oils have been manufactured from wax cuts from solvent extraction processes since the early 1970s. The conversion of long-chain normal paraffins or less branched paraffinic components into isomerized products with good low-temperature characteristics and high *VI*s succeeded, with suitable catalysts, in isomerizing hydrocracking processes. Because the input feed is already aromatic- and naphthene-free, the catalysts can be fully optimized to the conversion of the paraffinic material. At relatively high yield losses, the process creates high-quality but comparatively expensive hydrocracked oils (Shell XHVI, Exxon Exxyn, Mobil MWI-2 catalyst).

4.4.6. Hybrid Lube Oil Processing

The combination of traditional solvent refining with severe hydrotreating and hydrocracking processes is known as hybrid processing. To create higher *VI*s (> 105) and to reduce sulfur content, hydrotreating can follow solvent extraction (e.g., in the manufacture of turbine oils).

The combination of a mild furfural extraction with hydrocracking can produce hydrocracked oils by low-hydrocracking conversion in small reactors. The low hydrogen consumption adds to the economy of such processes. The introduction of a hydrocracking stage into a conventional solvent refinery offers the attractive possibilities of de-bottlenecking if the corresponding dewaxing capacity is available.

Catalytic dewaxing can, on the other hand, follow solvent refining. This presupposes that

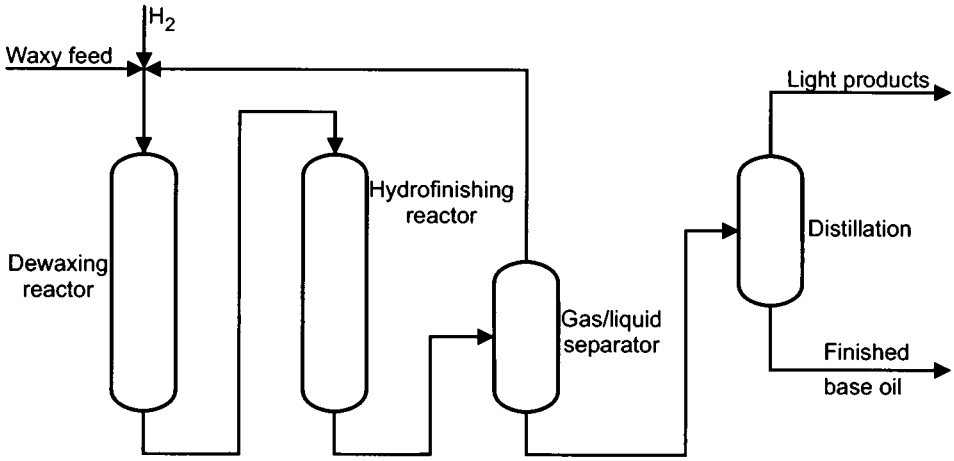

Figure 19. Flow chart for catalytic dewaxing (hydrodewaxing) including hydrofinishing

the catalytic dewaxing catalysts are not poisoned by the sulfur and nitrogen components in the solvent raffinate. In the Mobil processes, this is more likely with MLDW catalysts than with MSDW catalysts.

Several other high-paraffin components should be considered as future base oil feeds. These could include high-wax (> 70 %) natural distillates and Fischer – Tropsch waxes or synthetic fluids from natural gas.

The quality of low-quality lube feeds (VGO) can be improved by hydrogenation before solvent extraction (Hy-starting or Hystart).

4.4.7. All-Hydrogen Route

The production of base oils by hydrocracking and catalytic, solvent-free dewaxing is called the all-hydrogen route. Figure 20 shows a flow sheet for such a refinery used to manufacture VHVI oils. Depending on the severity of the hydrocracker, Group II HVI or Group III VHVI oils can be produced (HC-I and HC-II oils, Table 4). Leading examples for HC-I oils are the Chevron refinery in Richmond, California and the Conoco – Pennzoil refinery in Lake Charles (Exel Paralubes) [63]. Some HC-I refineries have been prepared to produce HC-II oils by the all-hydrogen route by increasing the severity of the hydrocracker. Another example of a hydrocracker refinery is the Petro-Canadas plant [64].

The use of a fuel hydrocracker for the production of base oils using the all-hydrogen route was first realized by the SK Corporation (Ulsan, Korea) in 1995. By recycling the hydrocracker bottom and the special integration of the fuel hydrocracker in the lube oil process, SK also developed a specific method (UCO process, Fig. 21) [65].

4.4.8. Gas-to-Liquids Conversion Technology

As a result of efforts to increase the value of natural gas in logistically favorable locations, the chemical liquefaction of natural gas (also the chemical reaction route) was developed (on the basis of the Fischer – Tropsch process; → Waxes, Chap. 5.). This process creates high-quality liquid products and paraffin wax. High-quality UHVI oils can then be obtained from natural gas by part oxidation, polymerization, and isomerization [66], [67].

The base oil market could undergo dramatic changes if Fischer – Tropsch waxes, which are part of gas-to-liquids technology, become more generally available. Shell has used such waxes from its Bintulu plant in Malaysia for its XHVI's. New technologies for smaller, efficient plants have been developed by Entech and Syntroleum in the last years.

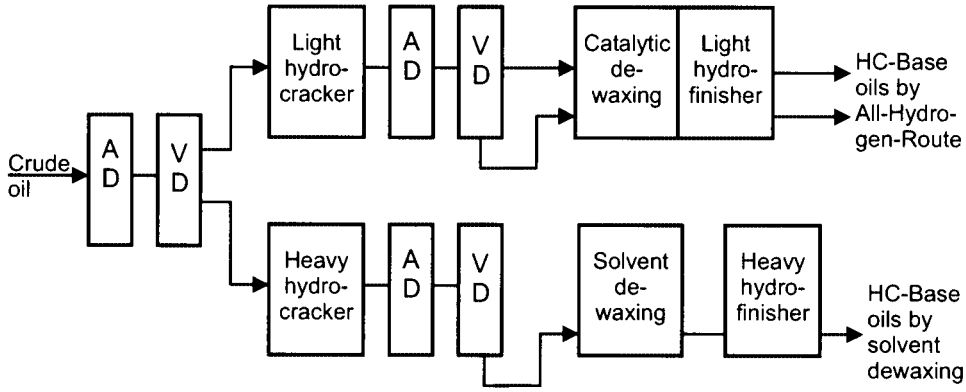

Figure 20. All-hydrogen route in Chevron's Richmond (California, USA) refinery

Table 4. API/ATIEL classification of base oils.

Group	Sulfur (wt %)		Saturates (wt %)	Viscosity index
I	> 0.03	and/or	< 90	80 – 120
I	> 0.03	and/or	< 90	80 – 120
II	≤ 0.03	and	≥ 90	80 – 120
III	0.03	and	90	> 120
IV	All polyalphaolefins (PAOs)			
V	All base oils not included in Groups I – IV			

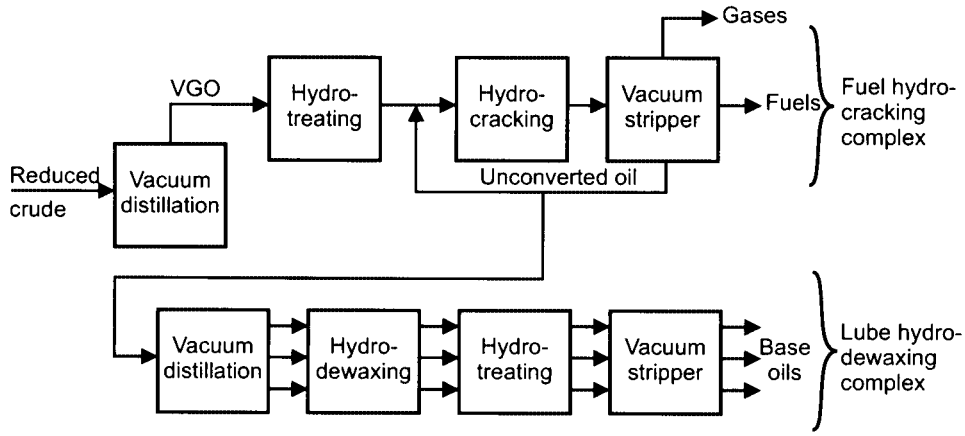

Figure 21. VHVI base oil production based on fuel hydrocracker residue (SK Corporation, Korea)

4.5. Boiling and Evaporation Behavior of Base Oils

The target of base oil distillation is the viscosity desired at 40 °C and 100 °C. The same distillation cut (same boiling point distribution) with different chemical structures leads to different viscosities. A highly naphthenic cut produces a higher viscosity than a paraffinic cut. In other words, equiviscous cuts of different chemical structures have different boiling-point distributions.

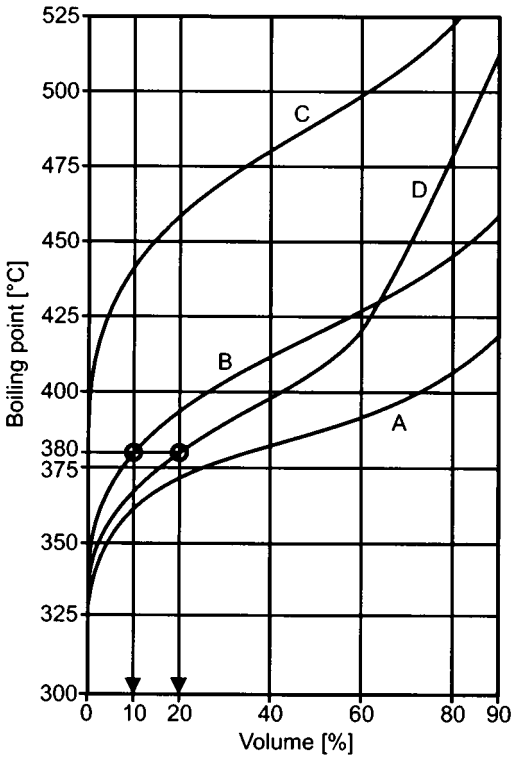

Figure 22. Distillation curves of three paraffinic solvent refined cuts
A) Light cut (low boiling); B) Medium light cut; C) Heavy cut (high boiling); D) Blend of A and C, equiviscous with B

In practice, ISO viscosity grades or other required viscosities are created by blending different cuts. If the boiling points of both cuts are too far apart, the flash point drops significantly and evaporation increases. Figure 22 shows the distillation curves of three paraffinic solvent refined cuts. One line is a blend of cut A with cut C which creates an equiviscous oil with cut B. The share with a boiling range up to 380 °C increases from 10 % (equiviscous cut) to 20 % (equiviscous blend).

Apart from the use of a variety of laboratory distillation apparatus to determine the boiling point distribution of a base oil, gas chromatographic methods of determining boiling point have gained popularity over the last few years. Both nonpolar packed and capillary gas chromatographic columns can be used. The injection port temperature is between 360 and 390 °C depending on the length of the column. The initial column temperature can be as low as − 50 °C, the final column temperature is between 360 and 390 °C and the programming rate is usually below 10 °C/min.

In recent years, the evaporation characteristics of lubricants have become increasingly important quality criteria. The reasons for this are the emissions created when a lubricant evaporates, and the accompanying change to the lubricant's composition. The topic of evaporation behavior has become increasingly important in line with the trend towards lower viscosity oils for most applications (energy-saving oils).

As a simple laboratory method, *Noack evaporation* (1 h at 250 °C) has become the established method of characterizing evaporation behavior in lubricant specifications with the evaporation losses being given in wt %. Gas chromatographic methods are also used but these can produce somewhat deviating results.

This led to a simulated Noack evaporation method using gas chromatography. In the USA, a gas chromatographic process was developed to determine engine oil volatility (ASTM D5480). In this case, evaporation takes place at 371 °C (700 °F) and the evaporating components are measured to ca. n-C_{22}. Although this process is easily reproducible, the values obtained are not comparable with those from the Noack method.

Figure 23 shows typical evaporation losses for well-cut paraffinic solvates; values for VHVI oils (HC-II oils, API Group III oils and typical ester base oils (Chap. 5) are shown for comparison.

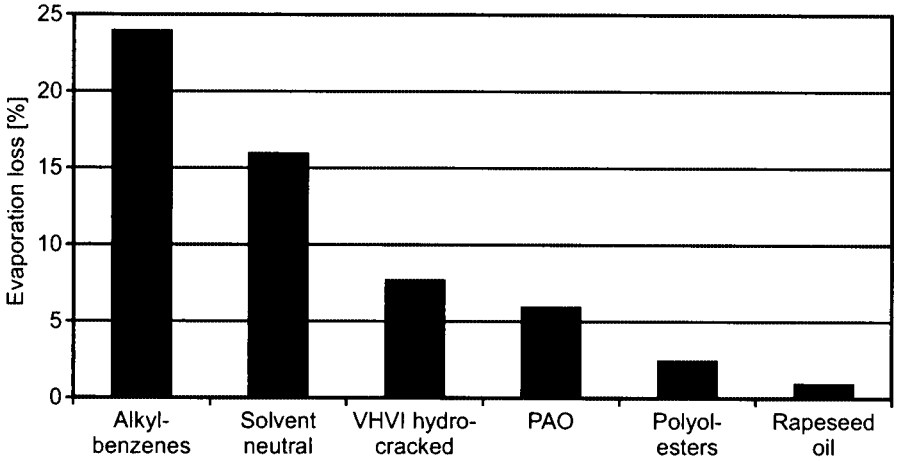

Figure 23. Typical evaporation losses (Noack evaporation) of various lube base oils (250 °C, 1 h)

5. Synthetic Base Oils

Classes of synthetic lubricants according to GUNDERSON and HART [68] and SHUBKIN et al. [69] are listed in Table 5.

Table 5. Classes of synthetic lubricant

GUNDERSON and HART (1962) [68]	SHUBKIN et al (1993) [69]
	Alkylated aromatics
Chlorofluorocarbon polymers	Chlorotrifluoroethylene
	Cycloaliphatics
	Dialkylcarbonates
Dibasic acid esters	Esters
Fluoro esters	
Neopentyl polyol esters	
	Perfluoroalkylpolyethers
Phosphate esters	Phosphate esters
	Phosphazenes
	Polyalphaolefins
	Polybutenes
Polyalkylene glycols	Polyalkylene glycols
Polyphenyl ethers	
	Silahydrocarbons
Silicate esters	
Siloxanes	Siloxanes

Although many of the synthetic base oils available today had been developed decades ago, their use on a large technical scale has increased only slowly because of their considerably higher cost [70], [71] (Table 6).

Table 6. Relative costs of synthetic base oils (relative to 1 for mineral oils in 1999)

Type	Relative cost
Alkylated aromatics	2–3
Polybutenes	3–5
Dibasic acid esters	7–15
Polyalkylene glycols	7–20
Polyalphaolefins	7–25
Neopentyl polyol esters	10–20
Siloxanes	40–200
Perfluoroalkylpolyethers	400–800

However, the consumption of polyalphaolefins, the most common synthetic lubricating oils, has increased enormously since the 1980s [72–74].

Synthetic base oils usually are prepared by reaction of a few defined chemical compounds – in many cases based on petroleum also – and tailored to their application by the right choice of reaction conditions.

Synthetic base oils have been classified according to both the production process and their chemical composition [71], [75]. From a chemical point of view the latter method is preferable.

5.1. Synthetic Hydrocarbons

Synthetic hydrocarbons were developed simultaneously in Germany and the United States. In Germany low-temperature performance and the

need to overcome the general shortage of petroleum base stocks was the driving force [76]. All synthetic hydrocarbons – and the other economically important synthetic lubricating oils – can be synthesized starting with ethylene.

5.1.1. Polyalphaolefins

The term polyalphaolefin (PAO) is derived from the source of this class of base oil, usually α-decene or a mixture of α-olefins containing, in general, a minimum of six and a maximum of twelve carbon atoms. Linear α-olefins were first used for the synthesis of lubricating oil by MONTGOMERY, GILBERT, and KLINE [77]. Polyalphaolefins have been described in detail in [74].

Oligomerization of α-olefins is generally carried out by boron trifluoride-based Friedel – Crafts oligomerization with alcohols as co-catalysts. Possible mechanisms have been described in detail in [78]. The boron trifluoride method has become predominant for the production of grades with lower viscosities, 2 – 10 mm^2/s. For higher viscosities, 40 and 100 mm^2/s, other catalysts must be used, and for the production of grades with viscosities of 10 – 25 mm^2/s, C-12 and C-14 olefins can be used; these yield mainly dimers [79]. In the next step the unsaturated olefins obtained are hydrogenated, using, e.g., nickel on kieselguhr or palladium on alumina as catalysts. In a third and final step the saturated oligomers are distilled. The oligomerization of α-decene or α-olefin mixtures always results in complex mixtures of isomers with more branching than expected. Because the rearrangements are intramolecular, the molecular masses of the products can be kept within a narrow range. A typical saturated α-decene trimer for example looks like a three-pointed star [80] (Fig. 24).

Polyalphaolefins therefore have several advantages – narrow boiling ranges, very low pour points, $VI > 135$ for all grades with kinematic viscosity $> 4\,mm^2/s$ at 100 °C. Their volatility is lower than that of all possible and equiviscous mineral oil grades and they contain only small quantities of unsaturated and polycyclic aromatic compounds and only traces of nitrogen, sulfur, or other impurities. Although in some oxidation tests without additives some mineral oils seem superior to polyalphaolefins, this is because of the presence of natural antioxidants in the mineral oils that have survived the refining process. The response of the synthetic products to antioxidants and their extreme pressure/antiwear synergists is better [81–83] than that of mineral oils. Their low polarity, on the other hand, leads to poorer solvency for very polar additives and can cause problems with seals. Therefore, PAOs tend to be used in combination with smaller amounts of (dicarboxylic acid) esters or solvent-refined mineral oils.

Polyalphaolefins have been used traditionally in aerospace and lifetime applications, but today they are used in a wider variety of applications and have gained additional importance because of the increasing need for them in automotive lubricants [84].

5.1.2. Polyinternalolefins

Polyinternalolefins (PIO) are rather similar to polyalphaolefins. Both kinds of hydrocarbon are prepared by the oligomerization of linear olefins. The difference is that polyinternalolefins are made from cracked paraffinic base stocks. Plants for the manufacture of polyinternalolefins and polyalphaolefins therefore look similar (Fig. 25). Internal olefins are more difficult to oligomerize and the resulting products have *VI*s 10 to 20 units lower than the *VI*s of equiviscous polyalphaolefins [85].

Mixtures of polyalphaolefins with polyethylenes with kinematic viscosities from 100 to 2000 mm^2/s at 100 °C are available commercially.

While the cationic polymerization of ethylene with aluminium chloride yields oils with *VI*s up to 120 and molecular masses from 400 to 2000, the polymerization of propylene gives oils with lower *VI*s and poor thermal stability. Both kinds of lubricating oil have lost importance, but copolymers of both base materials are said to have the potential for gaining importance again.

5.1.3. Polybutenes (→ Polyolefins, Chap. 4.)

Polybutenes (PBs) as synthetic base oils have been described in detail in [86]. They consist

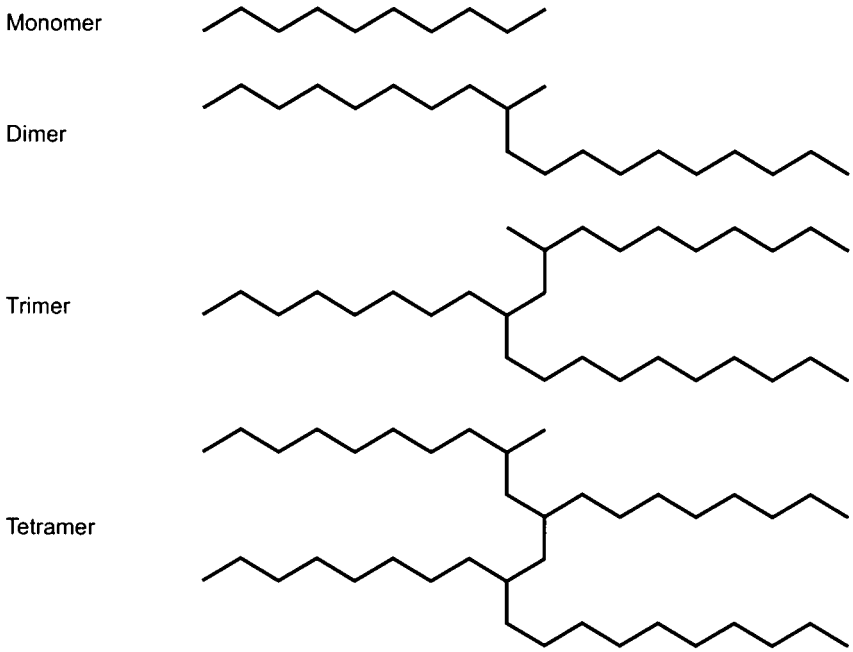

Figure 24. Molecular structures of typical polyalphaolefin oligomers

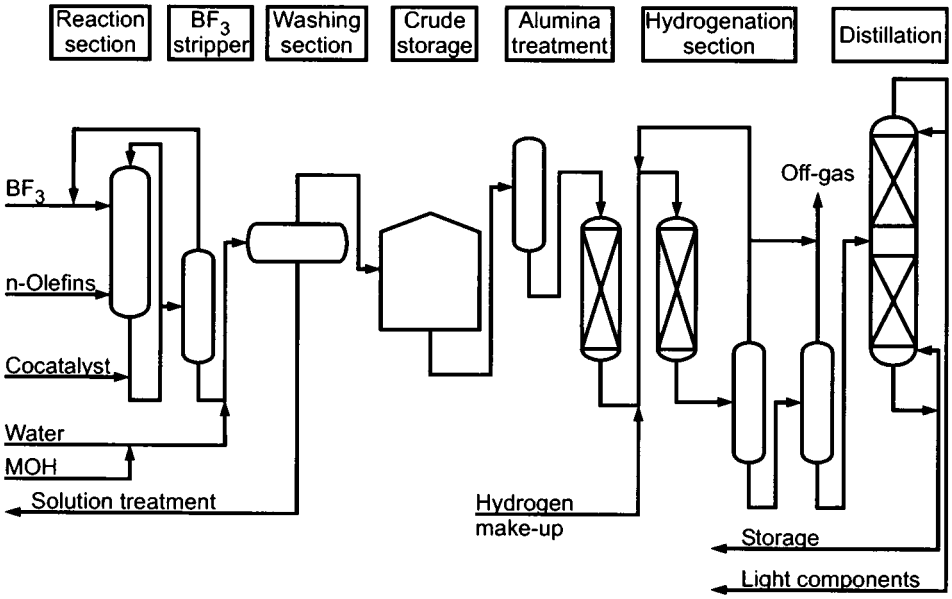

Figure 25. Schematic of a polyinternalolefin (polyalphaolefin) plant

mainly of isobutene and are, therefore, often also known as polyisobutenes (PIBs).

Polybutenes are produced by Lewis-acid catalyzed polymerization of the raw C_4 stream that beside isobutene contains the other two butenes and the butanes (→ Butenes). The process yields a copolymer with a backbone built mainly from isobutene units. The higher the molecular mass, the lower the content of other butenes, that is, the lower the molecular mass the more complex the structure. Polybutenes contain one double bond at the end of the carbon chain. Therefore, polybutenes are less resistant to oxidation than polyalphaolefins, polyinternalolefins, and alkylated aromatics – above 200 °C they begin to depolymerize and form gaseous products.

Polybutenes with molecular masses from about 300 to 6000 are important as *VI* improvers. They are used as components in two-stroke oils, gear and hydraulic oils, metal working lubricants, greases, compressor lubricants, and wire-rope protectives [87]. One of the options for future polybutene applications is the increased use as synergistic components in binary or ternary base oil systems [88].

5.1.4. Alkylated Aromatics

Alkylnaphthalenes as lubricants have been available in Germany since ca. 1930 [89], [90]. Low-cost dialkylbenzenes are produced as by-products in the manufacture of linear and branched monoalkylbenzenes that themselves are feedstocks for the manufacture of detergents. The Friedel – Crafts alkylation of benzene with olefins tends to yield polyalkylated products, because of the reactivity of the primary product. With excess benzene, a suitable olefin, e.g., propene, and the right choice of a catalyst the reaction can be controlled and specialized alkylbenzenes can be synthesized.

Lower-cost dialkylbenzenes are used in a wide variety of industrial areas, e.g., in transformer oils, where high resistance to gas evolution is important. Tailored grades, especially the linear products, although having poorer properties than polyalphaolefins, still are an option because of their excellent solvency, their suitability for low temperatures, and their compatibility with elastomers [91]. The most widespread use of these products is, however as refrigerator oils.

5.1.5. Other Hydrocarbons

Cyclic saturated or alicyclic hydrocarbons (naphthenes) occur naturally as components of mineral oils. Naphthenes do not yet have comparable synthetic analogs available on a technical scale.

5.2. Halogenated Hydrocarbons

Chlorinated hydrocarbons are very stable and not combustible. They were once widely used as insulating oils, heat-transfer fluids, and hydraulic fluids, but this use has been discontinued because of the environmental problems they cause.

Commercial liquid chlorofluorocarbons, all oligomers of chlorotrifluoroethylene (→ Fluorine Compounds, Organic, Chap. 4.7.), contain between two and twelve monomer units. They are suitable for use as lubricants mainly because of their extraordinary chemical stability, including against oxygen.

Their low corrosivity, the good low-temperature characteristics of the low-viscosity grades, and their lubricating properties are also advantageous. Disadvantages are their high volatility and their not ideal $V-T$ dependence. Their density naturally is relatively high (between 1.7 and 2 g/mL at 40 °C). Chlorofluorocarbons are used as lubricants in oxygen compressors, in pumps for mineral acids, halogens, and oxygen, and in mills or mixers for strong oxidizing agents. They are also suitable lubricants for turbine pumps in rocket engines and as base oils for non-flammable hydraulic fluids [92].

5.3. Synthetic Esters

5.3.1. Esters of Carboxylic Acids (→ Esters, Organic)

Synthetic ester lubricants for jet engines were developed in Germany during World War II [93].

Carboxylic groups, because of their strong dipole moments, reduce the volatility and increase the flash point of lubricating oils, and at the same time positively affect thermal stability, (the bonds of the COO group are thermally more stable than the C – C bond), solvency, lubricity, and biodegradability. On the other hand, however, they negatively affect the hydrolytic stability of a lubricant and the reactivity with metals or alloys that contain copper or lead [94].

5.3.1.1. Dicarboxylic Acid Esters [95]

Two types of ester molecule or their combinations have been found to be most suited as lubricant components:

1) Esters of branched primary alcohols with straight-chain dicarboxylic acids, and

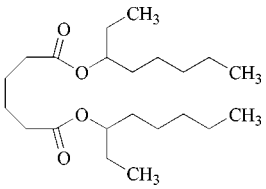

Di(2-ethylhexyl)adipate

2) Esters of straight-chain primary alcohols with branched dicarboxylic acids

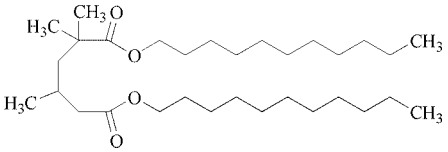

Diundecyl(2,2,4-trimethyl)adipate

The alcohols that are needed for the production of dicarboxylic acid ester oils can be obtained synthetically by the hydroformylation of olefins with carbon monoxide and hydrogen (→ Oxo Synthesis). The dicarboxylic acids can be obtained either by oxidation of vegetable oils, e.g., castor oil, as for azelaic and sebacic acid [96], or by cleavage of corresponding alicyclic hydrocarbons with oxygen, as for adipic acid [97].

Esters of straight-chain dicarboxylic acids have better $V-T$ characteristics than mineral oils, and higher VI values. The values decrease with increasing branching, but branching improves the low-temperature properties. Esters with little branching have the best properties, especially esters with methyl groups adjacent to the carboxyl group.

Dicarboxylic acid ester oils can improve the $V-T$ characteristics of automotive engine oils without having a negative effect on the low-temperature viscosity or the Newtonian flow behavior. Such oils are suitable for diesel engines also. Esters have also gained importance as lubricity additives in greases and in polyalphaolefins as additives for the improvement of the compatibility with elastomers.

5.3.1.2. Polyol Esters

Since the 1940s neopentyl glycol ((2,2-dimethyl)-1,3-propandiol), trimethylolethane and -propane (1,1,1-tris(hydroxymethyl)ethane and -propane), and pentaerythritol (2,2-bis(hydroxymethyl)-1,3-propanediol) have been used as the alcoholic components of esters [98]. Beside excellent thermal and oxidation stability, the esters have good $V-T$ and lubricating properties and excellent viscosity characteristics at low temperatures.

The mechanism of the polymerization of polyol esters during aging has been studied [99] and benzotriazole derivatives have been proposed as multifunctional additives [100]. Since the 1960s the polyol esters have gained importance as lubricants for high-temperature applications, for example in aircraft engines that are meant for speeds beyond Mach 2, and they have become even more important as biodegradable esters.

5.3.1.3. Other Carboxylic Esters

Esters of some aromatic polycarboxylic acids (**1**), mainly phthalic acid (**2**), trimellitic acid, and pyromellitic acid (**3**), with monofunctional alcohols are also used as base oils. The latter are suited as high temperature lubricants. The time dependence of the low-temperature viscosity of some grades is rather pronounced. Therefore their use in low temperature applications is problematic.

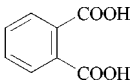

1

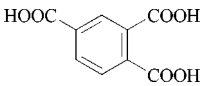

2

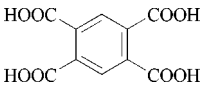

3

Monoesters of fatty acids and monofunctional alcohols are used in metal working. The sulfurized esters are excellent extreme pressure additives for all kinds of lubricant [101].

Dimeric acid esters, formed by dimerization of esters of oleic or tallow fatty acid, are used as engine oil and thickener components.

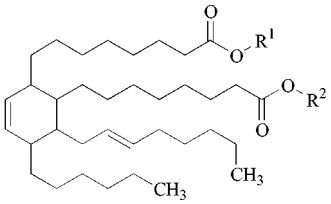

5.3.1.4. Complex Esters

Esters that contain straight or branched diols or polyalkylene glycols and straight or branched dicarboxylic acids as well as (mono)carboxylic acids and monofunctional alcohols have attracted interest. Usually first the diol is esterified with the dicarboxylic acid, then, depending on the desired product, the intermediate product is reacted with either a carboxylic acid or a monoalcohol. Two types of complex ester have become important:

A)

ROH · · · HOOCrCOOH · · · HORrROH
· · · HOOCrCOOH · · · HOR

B)

RCOOH · · · HORrROH · · · HOOCrCOOH
· · · HORrROH · · · HOOCR

The same scheme is valid when neopentane derived polyols [102] or glycerol [103] are involved.

Complex esters have higher molecular masses and higher viscosities than the common esters. Esters of type B are superior to type A with regard to flash point, pour point, and low-temperature viscosity. In type A polyalkylene glycols lead to lower pour points than aliphatic glycols.

High molecular mass complex esters of the type:

$$R \cdots (HOrOH \cdots HOOCrCOOH)_n \cdots R$$

have also been named polymer esters and find use mainly in applications where mineral oil and solvent resistance is required.

5.3.1.5. Fluorinated Carboxylic Acid Esters

Esters of fluorinated acids are readily hydrolyzed and the free acids have a strong oxidizing effect, therefore, only esters of nonfluorinated carboxylic acids with fluorinated alcohols can be used as lubricants. The thermal stability and the oxidation stability of fluorinated esters is better than that of the nonfluorinated analogs. Some types of rubber shrink in their presence and some antioxidants have a negative effect on their stability. Fluorinated esters were in the first place intended for high-temperature applications [104]. It can be assumed that their use will be limited because of competition from the fluorinated polyethers.

5.3.2. Phosphate Esters (→ Phosphorus Compounds, Organic, Chap. 10.) [105], [106]

Tertiary esters of phosphoric acid with alcohols or phenols were first prepared 150 years ago but were not introduced as antiwear additives before 1930. They have also gained importance as plasticizers, fire-resistant hydraulic fluids, compressor oils, and synthetic lubricants. They are usually divided into triaryl, trialkyl, and alkylaryl phosphates and are made by reaction of phosphoryl chloride with phenols or alcohols. Their properties range from low-viscosity fluids

to high-melting solids. With increasing molecular mass trialkyl phosphates change from water-soluble to insoluble liquids. The triaryl phosphates have higher viscosities and are insoluble in water. An aryl side-chain reduces the melting point. The properties of alkylaryl phosphates are between those of alkyl and aryl phosphates.

$$3\,ROH + POCl_3 \rightarrow OP(OR)_3 + 3\,HCl$$

The hydrolytic stability varies between good and poor. Alkylaryl phosphates are more susceptible to hydrolysis than trialkyl or triaryl phosphates. Phosphoric acid esters are less stable than carboxylic acid esters, but more stable than silicic or boric acid esters.

The thermal stability of triaryl phosphates is better than that of the alkyl compounds. With aminic antioxidants and rust inhibitors, triaryl phosphates can be used up to 175 °C, but alkyldiaryl phosphates only up to 120 °C. Branching of the alkyl radicals reduces the thermal stability, the effect becomes stronger with decreasing length. In general phosphoric acid esters are not corrosive. The high spontaneous ignition temperatures of up to 600 °C underline the good fire resistance of some products. Trialkyl and alkylaryl products have pour points down to − 65 °C, when they contain *VI* improvers.

The lubricating properties of the phosphates are excellent, particularly on steel. They can be blended with almost all lubricants and additives. Their dissolving power is, on the other hand, responsible for their incompatibility with rubbers, varnishes, and plastics. Nylon, epoxy and phenol – formaldehyde resins are stable.

5.4. Polyalkylene Glycols

(→ Polyoxyalkylenes) [108–110]

The first polyalkylene glycols suitable as lubricants were developed during World War II. The first patent was probably published in 1947 by Roberts and Fife [107].

Polyalkylene glycols are prepared by the reaction of epoxides, usually ethylene and propylene oxide, with compounds that contain active hydrogen, usually alcohols or water, in the presence of a basic catalyst, e.g., sodium or potassium hydroxide. Polymers with statistically distributed alkylene groups are made by use of a mixture of alkylene oxides. Separate addition leads to block copolymers. Because ethylene oxide is more reactive than propylene oxide, random copolymers tend to have the propylene oxide units at the chain ends.

Terpolymers with, for example, tetrahydrofuran have also been prepared. Pure tetrahydrofuran polymers can be obtained by polymerization of tetrahydrofuran in the presence of Friedel – Crafts catalysts. They are colorless oily or waxy substances of very low toxicity.

The miscibility of polyoxyalkylenes with water increases with the number of ethylene oxide units. In solution the water-soluble grades are practically nonflammable. The solubility in water decreases with increasing molecular mass and number of ether bonds. Solubility in hydrocarbons increases with the molecular mass. Polyalkylene glycols in general are soluble in aromatic hydrocarbons.

The molecular mass and viscosity of polyalkylene glycols can be significantly influenced during production and be adjusted within narrow limits. The possibility of engineering products in this way distinguishes them from many other lubricants.

Low-molecular-mass polyalkylene glycols containing more than 50 % propylene oxide have pour points that go down to − 65 °C. The lateral methyl groups are responsible for the disruption of crystallization. On the other hand pure high-molecular-mass polyethylene glycols are wax-like solids with pour points near + 4 °C.

Kinematic viscosities range from 8 to 100 000 mm^2/s at 40 °C. The *VI* values of polyalkylene glycols usually lie around 200. High-molecular-mass polyethylene glycols have *VI* values of up to 400.

Prolonged heating of polyalkylene glycols to above 150 °C leads to depolymerization. Traces of alkali or alkaline-earth metals promote the degradation. The decomposition can be prevented by addition of aminic antioxidants, to the extent that the oils can be used as heat-transfer fluids up to 250 °C.

The polar nature of polyalkylene glycols gives the products strong affinity for metals, and so the lubricating film remains intact even at high surface pressure, a property useful in lubricants and metal cutting fluids. Because swelling of elastomers decreases with increasing viscosity, polyalkylene glycols can be used with natu-

ral and with synthetic rubbers in hydraulic oils and brake fluids. Because of their low toxicity polyalkylene glycols find application in the food, pharmaceutical, tobacco, and cosmetics industries. Products with high ethylene oxide content are up to 80 % biodegradable. High viscosity water-soluble products are shear-stable liquid thickeners. Hydrophilic and hydrophobic fractions in block copolymers give them surfactant properties.

5.5. Other Polyethers

Alkylated aryl ethers, like alkyl ethers, have lower viscosities, lower *VI* values, lower pour points, and lower boiling points than the corresponding alkanes. Asymmetric substitution also reduces the pour point, but has an adverse effect on the $V-T$ relationship. Although such products have low viscosities, some have found use as components of lubricants used in the life sciences.

5.5.1. Perfluorinated Polyethers (→ Fluorine Compounds, Organic, Chap. 6.1.) [111–113]

Photochemical polymerization of tetrafluoroethylene in the presence of oxygen, followed by fluorination with elemental fluorine leads to products of type A (Eq. 5.1) in the case of tetrafluoroethylene. With hexafluoropropylene products of type B (Eq. 5.2) result. Anionic polymerization of hexafluoropropylene epoxide leads to products of type C (Eq. 5.3), and Lewis acid-catalyzed polymerization of 2,2,3,3-tetrafluorooxetane leads to products of type D (Eq. 5.4).

$$F_2C{=}CF_2 \xrightarrow{O_2/F} F_3C{-}O{-}\left[\overset{CF_3}{\overset{|}{CF}}{-}CF_2{-}O\right]_m{-}\left[CF_2{-}O\right]_n{-}CF_3 \quad (5.1)$$

$$F_2C{=}CF{-}CF_3 \xrightarrow{O_2/F} F_3C{-}O{-}\left[CF_3{-}CF_2{-}O\right]_m{-}\left[CF_2{-}O\right]_n{-}CF_3 \quad (5.2)$$

$$F_2\overset{O}{C}{-}CF{-}CF_3 \xrightarrow[(CsF)]{F} F_3C{-}CF_2{-}CF_2{-}O{-}\left[\overset{CF_3}{\overset{|}{CF}}{-}CF_2{-}O\right]_n{-}CF_2{-}CF_3 \quad (5.3)$$

$$H_2\overset{O-CF_2}{\overset{|}{C}}{-}\overset{|}{CF_2} \xrightarrow[(BF_3)]{F} F_3C{-}CF_2{-}CF_2{-}O{-}\left[CF_2{-}CF_2{-}CF_2{-}O\right]_n{-}CF_2{-}CF_3 \quad (5.4)$$

The density of the perfluorinated polyethers (PFPEs) is nearly twice that of hydrocarbons. They are immiscible with most of the other base oils and non-flammable under nearly all practical conditions. The more common types A and B have a good to very good $V-T$ and $V-p$ dependence, and low pour points [114]. The viscosity of linear PFPEs changes less with temperature and pressure than that of the non-linear variety; in air they are stable up to 400 °C. PFPEs are remarkably inert chemically and have an excellent hydrolytic stability.

The shear stability of PFPEs is better than that of other polymeric lubricants, but in the presence of steel and under boundary lubrication conditions PFPEs do not perform very well [115].

Perfluoropolyalkyl ethers have all the properties required by modern spacecraft: as lubricants and hydraulic fluids they resist to thermal and oxidative attack above 260 °C, possess good low-temperature flow characteristics, and are fire-resistant. They can also be used as power transmission and inert fluids and in transformers and generators as dielectrics with outstanding properties.

5.5.2. Polyphenyl Ethers

Polyphenyl ethers are formed by reaction of phenols and aryl halides [116]. Abbreviated formulas are in use. They contain the substitution position and the number of phenyl rings (ϕ) and ether bonds, for instance ppp5P4E stands for $\phi-O-\phi-O-\phi-O-\phi-O-\phi$ (p–p–p).

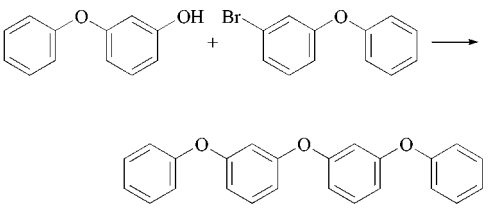

(5.5)

The aromatic groups increase the stability, but negatively affect the $V-T$ dependence of these polyethers. Alkyl groups lower the high melting points. *para* Derivatives have lower volatilities; for *ortho* products the volatilities are higher. Spontaneous ignition occurs between 550 and 600 °C and alkyl substitution reduces this by ca. 50 °C. With the usual elastomers swelling occurs.

The oxidation stability of polyphenyl ethers is only slightly lower than that of polyphenyls or tetraarylsilanes. Alkyl substituents reduce it. The thermal decomposition temperature is 465 °C. Short-chain substituents reduce it to below 380 °C and with higher alkyl groups it drops to below 350 °C, that is, to temperatures typical for aliphatic hydrocarbons. Trifluoromethyl groups decrease the thermal decomposition temperature to below 270 °C. Coke formation is low. It increases with alkyl substitution, particularly in the presence of methyl groups.

Polyphenyl ethers are the most radiation-resistant lubricants. At low temperatures radiation affects the viscosity more pronouncedly than at high temperatures. It increases viscosity, acidity, evaporation loss, corrosivity, and coke formation, but reduces flash and ignition point.

Between 200 and 300 °C the lubricating properties of polyphenyl ethers are reported to be comparable with those of mineral and ester oils and better than those of polysiloxanes and aromatic hydrocarbons. Alkylated polyphenyl ethers have better properties than the unsubstituted ethers. Polyphenyl ethers are suited as high-temperature or radiation-resistant hydraulic fluids and lubricants.

5.5.3. Polysiloxanes (Silicone Oils)

(→ Silicones, Chap. 3.)

Silicone oils suitable as lubricants are generally straight-chain polymers of the dimethylsiloxane (**4**) and phenylmethylsiloxane series (**5**). They have been described in detail [117], [118].

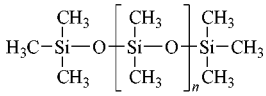

4

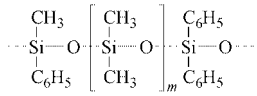

5

Methyl silicone oils are simply made from quartz and methanol in the end (Fig. 26), but commercial technology employs sophisticated steps, e.g., the so-called equilibration – treatment of a mixture of siloxanes of different molecular mass with strong acid or basic catalysts with the aim of achieving a narrow Gaussian molecular mass distribution [119].

Among the unique properties of silicone lubricating oils is their immiscibility with many organic fluids, the low temperature-dependence of their physical properties, and their physiological inertness. Silicone oils can be produced with very low pour points and high viscosity by increasing the asymmetry of the molecules, usually by partial replacement of the dimethylsilyl groups by phenylmethylsilyl groups. The density of silicone oils is in the region of that of water, those of dimethylsilicone oils are slightly below, those of phenylmethyl silicone oils slightly above.

The extremely low viscosity – temperature coefficients (*VTC*), < 0.6 for low-viscosity dimethylsilicone oils are mainly due to the extraordinary flexibility of their Si – O chains. Low to medium viscosity silicone oils have Newtonian behavior up to high shear rates, but with increasing viscosity the apparent viscosity decreases with increasing shear rate (pseudoplastic behavior). Compressibility and viscosity changes at high pressure depend strongly on the methyl-to-phenyl ratio of the silicone oils and are relatively large, again mainly because of the flexibility of their Si – O chains.

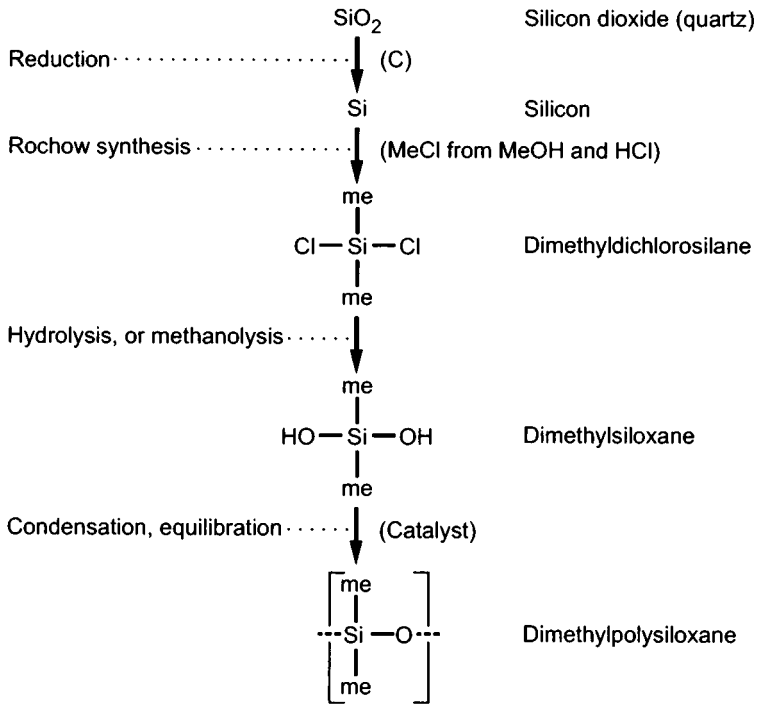

Figure 26. Preparation of dimethylsilicone oils

The thermal decomposition of silicone oils begins at ca. 300 °C. In general the decomposition products are not corrosive, but they lower the viscosity. Up to 200 °C the oxidation stability of silicone oils is superior to that of hydrocarbons, esters, and polyalkylene glycols. At higher temperature the formation of siloxyl and silyl radicals via cross-linking leads to the formation of polymer molecules and thus results in gels.

Some iron and cerium compounds are especially suitable as oxidation inhibitors for silicone oils. At elevated temperatures gel formation nevertheless occurs readily in the presence of selenium or tellurium. In the presence of chlorine even explosions are possible. Si – O bonds can also be broken by hydrolytic attack, but in contrast with silicic ester oils formation of silica or silicic acid gel does not occur.

The surface tension and the foaming tendency of silicone oils are much lower than those of mineral oils; silicone oils are, therefore, used as defoamers for the hydrocarbons.

Silicone oils can be found in all kinds of industrial and military installations. Dimethylsilicone oils are used as lubricants for bearings and gears with rolling friction. In sliding friction their performance depends on the metal pairs, they are used, for example, as lubricants for bronze or brass on aluminum. Silicone oils are among the best lubricants for plastic bearings, but in precision instrument applications spreading should be prevented by use of an epilamization agent. They are suitable lubricants for rubber parts, also, and also serve as switch and transformer oils. Silicone oils with higher phenyl substitution are mainly used for the lubrication of turbines, ball bearings, and all kinds of instrument, especially at high temperatures. Their radiation stability is also remarkable.

Silicone oils can be used as base oils for all kinds of lubricating *greases* [120].

5.6. Other Synthetic Base Oils

Of the following group of synthetic fluids that have been proposed as base oils for lubricants or that have been developed especially for that use, none has gained yet real economic or commercial importance in that area during the last 30 years, all still wait for their niche application, even though some have really astonishing properties.

Acetals have been investigated because of their high stability in alkaline media; they are, however, sensitive to acidic hydrolysis.

Adamantane derivatives, mainly the esters of 1,3-dihydroxy-5,7-dimethyladamantane, have been reported as being superior to the common dicarboxylic acid esters [122].

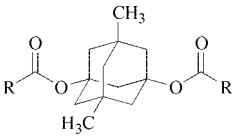

Aromatic amines, in particular their trialkylsilyl-substituted derivatives have been considered as lubricants because of their good heat resistance [121].

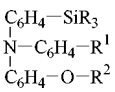

The performance of dialkylcarbonates, the diesters of carbonic acid, is similar to that of common dicarboxylic acid esters. They have been reported to be preferable to common ester with regard to lower toxicity and better seal compatibility [123]. In addition, no acid compounds are formed on decomposition.

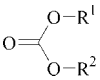

Heterocyclic Boron, Nitrogen, and Phosphorus Compounds. Triazine derivatives such as melamine and cyanuric acid form the nuclei for fire-resistant and high-temperature lubricants [121], [122]. Polycarboranesiloxanes (**6**) have been reported to be comparable with silicone oils, but can be used at temperatures above 220 °C [124]. The fluorine-containing phosphazene derivatives (**7**) that have been developed mainly for the use in fire-resistant high-temperature hydraulic fluids [125] and for use as lubricity additives in perfluorinated polyethers [126], [127].

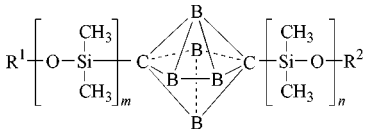

6

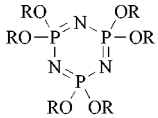

$R = C_6H_4F, C_6H_4{-}CF_3, \ldots$

7

Polyalkylene sulfides (polythioether oils) are polyalkylene glycols in which all C – O – C bonds are replaced by C – S – C bonds; they have good oxidation stability, and higher viscosities and lower pour points than the corresponding hydrocarbons [128].

Polyphenyl sulfides (polyphenyl thioethers, C-ethers) are polyphenyl ethers in which all C – O – C bonds are replaced by C – S – C bonds. This gives them lower pour points and better boundary lubrication properties, but reduced oxidative and thermal stability [129], [130].

Silahydrocarbons. Compared with their hydrocarbon analogs, silahydrocarbons, mainly tetraalkylsilanes have lower pour points, lower volatilities, higher *VI* values up to 155, and superior thermal stability [121, 131, 132]. They respond to antioxidants and antiwear additives in the same way as PAOs [133].

$R = CH_3, C_2H_5 \ldots C_{12}H_{25}$

Silicate esters, (ortho)silicic acid esters, have been known since the middle of the 19th century. Their properties differ strongly from those of siloxanes. Despite their good thermal stability and their low pour points their use is restricted to hydraulic, heat-transfer, and cooling fluids and to oils for automatic weapons because of their poor hydrolytic stability [134]. With the development of polysilicate clusters the shortcomings of the simple esters can, perhaps, be overcome [135].

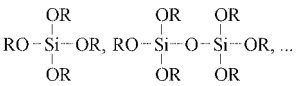

5.7. Mixtures of Synthetic Lubricants

Today few lubricants contain only one base oil, firstly because mixing two or more base oils with different properties often leads to a lubricant with the desired performance, and secondly because many of the more polar synthetic base oils serve as additives in less polar oils, e.g., esters in hydrocarbons and vice versa. In thickened systems even immiscible base fluids like hydrocarbons and complex esters, perfluorinated ethers, or polyalkylene glycols can be combined.

6. Additives

Base fluids – mineral oil and also synthetic products – generally cannot satisfy the requirements of high performance lubricants without using the benefit of modern additive technology.

Additives can be classified into types that

1) influence the physical, e.g., $V-T$ characteristics, demulsibility, low temperature properties, etc., and chemical properties, e.g., oxidation stability, of the *base fluids*
2) affect primarily the *metal surfaces* modifying their physicochemical properties, e.g., reduction of friction, increase of EP behavior, wear protection, corrosion inhibition, etc.

Additives are used at contentrations from a few ppm (antifoam agents) up to 20 wt % or even more. They can assist each other (synergism) or they can lead to antagonistic effects. Some additives are multifunctional products that decrease the possibility of additives interfering with each other.

Of course, there are some properties that cannot be influenced by additives, e.g., volatility, air release properties, thermal stability, thermal conductivity, compressibility, or boiling point..

Although well balanced and optimized additive systems can improve the performance of lubricants enormously, the formulation of high performance lubricants requires also excellent high quality base fluids.

6.1. Antioxidants

The function of a lubricant is mostly limited by the aging of the lubricant base stock. Typical characteristics of aged lubricants are discoloration and a characteristic burnt odor. In advanced stages the viscosity will begin to rise significantly and acidic oxidation products are formed, which in turn may induce corrosion and lubricant problems. This aging process can be delayed tremendously by the use of antioxidants.

6.1.1. Mechanism of Oxidation and Antioxidants

The aging of lubricants can be differentiated into two processes: the oxidation process by reaction of the lubricant molecules with oxygen and the thermal decomposition (cracking) at high temperatures. In practice the oxidative aging of the lubricant is the dominating process which influences significantly the lifetime of the lubricant. Caused by steadily increased power density and reduced lubricant volumes (higher load : oil ratio) as well as extended service life in the last years, the thermal stress on the lubricant molecules grows constantly.

The oxidation of hydrocarbons can be described by the well-known free radical mechanism via alkyl and peroxy radicals [136]. The main reaction steps are shown in Figure 27.

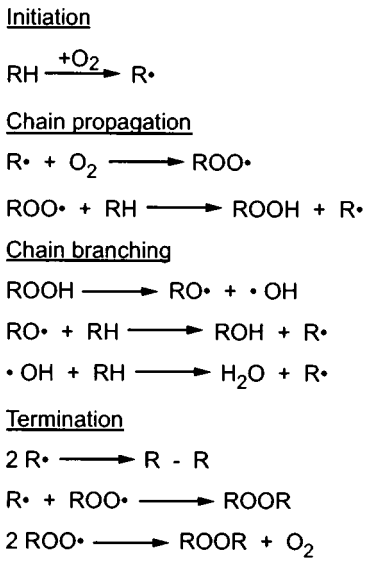

Figure 27. Mechanism of autoxidation

The difference in reactivity of the miscellaneous radicals formed during autoxidation explains why linear and saturated hydrocarbons exhibit a much higher oxidation stability compared to branched, aromatic and unsaturated hydrocarbons.

The typical oxidation products that will be formed by these oxidation processes include alkyl hydroperoxides, dialkyl peroxides, alcohols, aldehydes, ketones, carboxylic acids, and esters. By polycondensation processes high molecular mass oxidation products are formed. These products are responsible for the typical viscosity increase of aged oil. Further polycondensation and polymerization of these oxidation products lead finally to oil-insoluble polymers that can be observed as sludge and varnish-like deposits.

Because of the acidic character of most of the oxidation products the danger of corrosion is increased. Also the attack of alkyl peroxy radicals on the metal surface may be responsible for corrosive wear. Furthermore such dissolved metals can form salts which also precipitate as sludge.

Nonhydrocarbon-based lubricants may behave totally different. Thus the aging of polyalkylene glycols will lead to a decrease of the viscosity caused by the decomposition of the polymeric structure.

Normally refined mineral base stocks contain traces of nitrogen-, sulfur- and oxygen-containing heterocycles as well as mercaptans, thioethers, and disulfides that may act as so-called 'natural antioxidants' or as pro-oxidants that will accelerate the oxidation of the lubricant. Usually the antioxidants are dominating thus providing a relatively good oxidation stability of the base oil. On the other hand hydrocracked base stocks (API group II and III) and synthetic lubricants like PAOs (API group IV) require a well balanced antioxidant combination in order to exploit the advantages of that kind of lubricants [137–139].

Antioxidants can be classified in primary antioxidants (radical scavengers) and secondary antioxidants (peroxide decomposers).

Radical scavengers compete successfully with the lubricant molecules in the reaction with reactive radicals of the propagation process. They react preferably with the radical oxidation products forming resonance-stabilized radicals that are so unreactive that they will stop the propagation of the autoxidation. Peroxide decomposers convert hydroperoxides into non-radical products thus also preventing the chain propagation reaction.

In finding the optimum synergistic antioxidant and metal deactivator combination the maximum delay of oil oxidation is achieved.

6.1.2. Compounds

Phenolic Derivatives (→ Antioxidants, Chap. 4.1.). Sterically hindered mono-, di- and polynuclear phenol derivatives belong to the most effective antioxidants acting as radical scavengers and are used in many applications. Typically the phenols are substituted in the 2 and 6 positions with tertiary alkyl groups. The most common substituent is the tertiary butyl group. The most simple derivatives are 2,6-di-*tert*-butylphenol (2,6-DTB) (**8**) and butylated hydroxytoluene, 2,6-di-*tert*-butyl-4-methylphenol (BHT) (**9**). The advantage of polynuclear phenols like 4,4′-methylenebis(2,6-di-*tert*-butylphenol) (**10**) or types with high-molecular-mass substituents in the 4 position is the reduced volatility due to the higher molecular mass that makes these products suitable for high temperature applications.

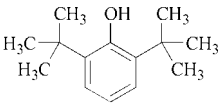

8

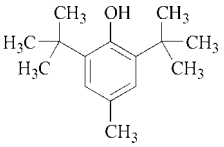

9

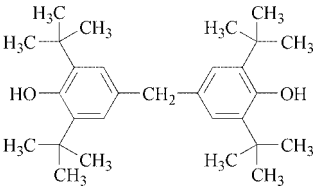

10

Aromatic Amines (→ Antioxidants, Chap. 4.2.). Oil-soluble secondary aromatic amines represent another important class of antioxidants that act as radical scavengers. Typical products are a large number of alkylated diphenylamines (**11**), *N*-phenyl-1-naphthylamine (PANA) (**12**) and the polymeric 2,2,4-trimethyldihydroquinoline (TMQ) (**13**). Because of its poor solubility in mineral oil the latter is commonly used in greases and polar lubricants.

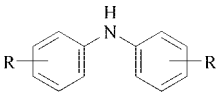

11

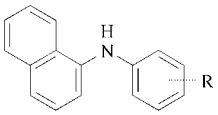

12

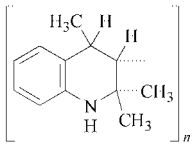

$n = 9\text{-}10$

13

Compounds Containing Sulfur and Phosphorus (→ Antioxidants, Chap. 4.3.; Section 6.8). The most famous representatives of this group of additives are the zinc dithiophosphates which mainly act as radical scavengers [140]. Because of their multifunctional properties (antioxidant, antiwear additive, extreme pressure additive, corrosion inhibitor) zinc dithiophosphates are still being used as standard additives.

In addition, a large number of ashless dithiophosphoric acid derivatives, the so-called O,O,S-triesters, exist. These are reaction products of dithiophosphoric acid with olefins, cyclopentadiene, norbornadiene, α-pinene, polybutene, unsaturated esters like acrylic acid esters, malenic acid esters and other chemicals with activated double bonds [140]. All this additives show antiwear and also antioxidant properties, although their performance as antioxidant is not really as good as that of the metal dithiophosphates.

Organosulfur compounds are typical peroxide decomposers. Zinc and methylenebis(dialkyldithiocarbamates) have been found to be highly efficient [140].

Organophosphorus Compounds. Triaryl- and trialkylphosphites are the most common types of this group. They are not only peroxide decomposers, but they also can limit photodegradation. Because of their relatively poor hydrolytic stability their application is restricted to sterically hindered derivatives.

Other Compounds. Surprisingly organocopper compounds in combination with peroxide decomposers can act as antioxidants although catalytic quantities of copper ions usually act as pro-oxidants [140]. Overbased phenates and salicylates of magnesium and calcium behave as antioxidants at higher temperatures. Many other compounds have been proposed as antioxidants but none of these could gain a real relevance.

Synergistic Mixtures. Commonly several types of antioxidation and combinations of them are used. Well known is the synergistic effect of aminic and phenolic antioxidants. The combination of radical scavengers and peroxide

decomposers is called heterosynergism. So the combination of phenolic antioxidants with phosphites is known as highly efficient especially in hydrotreated base stock [141]. As metals may act as catalysts in the oxidation process typical rust inhibitors and metal passivators are also used as synergistic compounds (see also Section 6.10).

6.2. Viscosity Modifiers

6.2.1. *VI* Improvement Mechanisms

In the simplest case, a desired viscosity index can be achieved by mixing fluids with corresponding viscosity index improvers (*VII*s). But for a number of technical and commercial reasons, viscosity modifiers are often used.

Viscosity modifiers are chain-like polymeric molecules whose solubility depends on chain length, structure and chemical composition [142], [143]. As a rule, the base oil solubility of these polymer chains deteriorates when the temperature falls and improves with increasing temperature so that an increase in viscosity induced by viscosity modifiers also increases the viscosity index.

The absolute increase in viscosity and the *VI* depends on the molecular mass and the concentration of viscosity modifiers in the formulation [144]. In practice and depending on the projected application, molecular masses of 10 000 to 250 000 are used. Concentrations are usually between 3 and 25 wt %. As a result of their high molecular mass, viscosity modifiers are always dissolved in a base fluid.

Apart from their thickening effect which is schematically illustrated in Figure 28A as a function of molecular mass, shear stability serves as a second characteristic [145]. According to Figure 28B, increasing molecular mass reduces shear stability if the polymer concentration remains constant.

The reason for this effect is mechanically- or thermally-induced chain degradation [146]. Depending on the type and duration of the load, a number of different molecular sizes are created. The resulting drop in viscosity is described by the Permanent Shear Stability Index

$$\frac{[\text{KV before shearing}] - [\text{KV after shearing}]}{[\text{KV before shearing}] - [\text{KV of the base oil}]}$$

$$\text{e.g.,}\quad \frac{11\text{ct} - 8\text{ct}}{11\text{ct} - 5\text{ct}}$$

where KV denotes the kinematic viscosity.

If the relaxation time for the polymer chains is short (10^{-6} s), the high-molecular molecules adopt a temporary alignment. This drop in viscosity is described by the temporary shear stability index (TSSI).

Both values are of great importance to automotive applications, especially to engine, gear and hydraulic oils because the specified characteristics of such an oil do not just apply to the fresh oil but should remain throughout the drain interval. As the reduction in molecular mass in practice is overshadowed by oxidation and other effects, a series of laboratory tests have been established to characterize shear stability [147].

6.2.2. Structure and Chemistry of Viscosity Modifiers

The most important monomers for viscosity modifiers are listed in Table 7.

It is also possible to sort these polymers according to their supramolecular chain structure, independently of the base monomers used (Fig. 29).

dispersing. Dispersing viscosity modifiers are in principle the link to ashless dispersants which are discussed in Section 6.4.2.

6.3. Pour Point Depressants

With the exception of polyalkylated naphthalenes, pour point depressants (PPD) are closely linked to a series of viscosity modifiers. The major difference between pour point depressants and viscosity modifiers is their application concentration and the selection of monomer building blocks [148].

Although paraffin crystallization cannot be suppressed, the crystalline lattice and thus the morphology of the paraffin crystals can be significantly altered by PPDs. While polyacrylates and ethylene vinyl acetate copolymers are used for crude oils and petroleum, special polyalkyl methylacrylates are normally used for mineral oil-based formulations. Due to co-crystallization of paraffinic components in the

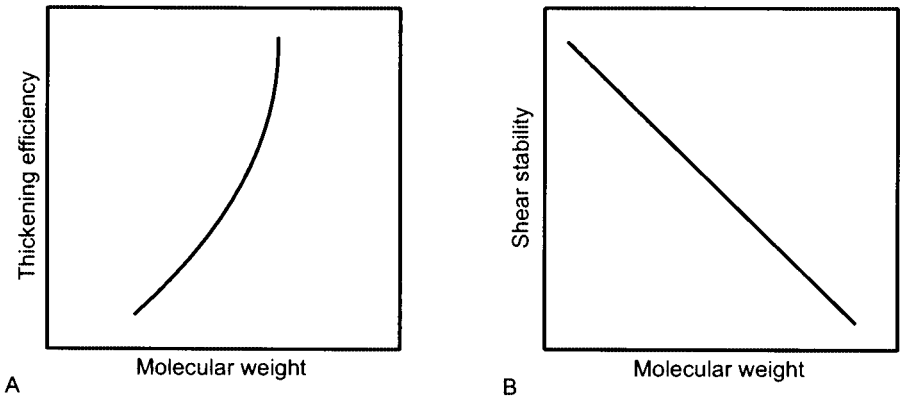

Figure 28. Thickening efficiency and shear stability of VMs

Table 7. Types of polymeric viscosity modifier

VM Type		Description	Main applications
OCP	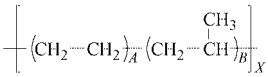	olefin copolymers	engine and hydraulic oils
PAMA	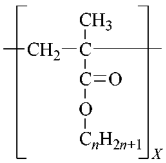	polyalkyl (meth)acrylates	gear and hydraulic oils
PIB	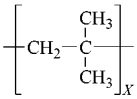	polyisobutene	gear oils, raw material for ashless dispersants
SIP	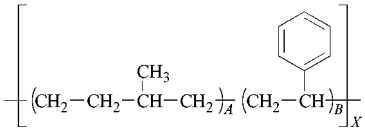	hydrogenated styrene – isoprene copolymers	engine oils
SBR	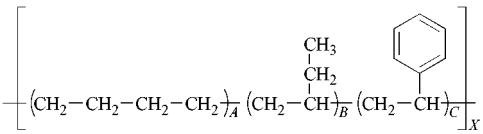	hydrogenated styrene – butadiene copolymers	engine oils

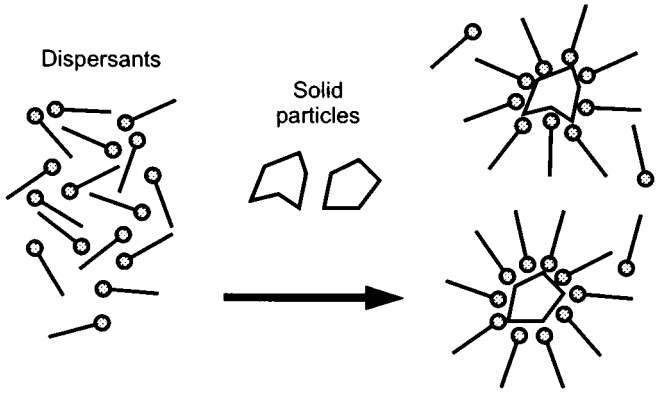

Figure 29. Viscosity modifiers, chain structures, and monomer make-up

base oil and the polymer chain, the crystal morphology is altered. Instead of the needle-like paraffin crystals which rapidly cause paraffin gelation, densely packed, round crystals are formed which hardly affect flowing properties even at temperatures below the pour point. The corresponding polyalkyl methacrylates-PPDS are in principle, comb-like and contain $C_{12} - C_{24}$ paraffinic side chains [149]. By determining the polymer solubility and crystallization of paraffinic components in the base oil, the optimum PPD-mix and concentration can be determined for every base oil mixture. However, as a wide range of base oils are used in practice, standard products are often used which cover almost the entire spectrum. In such cases, efficiency is controlled via PPD concentration.

6.4. Detergents and Dispersants

Detergents and dispersants (DD), often called heavy-duty (HD) additives have been indispensable for the development of modern engine oils for gasoline and diesel combustion motors. These lubricants are especially severely stressed due to the high engine temperatures and the additional influence of aggressive blow-by gases of the combustion process. DD additives keep oil-insoluble combustion products in suspension and also prevent resinous and asphalt-like oxidation products from agglomerating into solid particles. These additives prevent oil thickening, sludge and varnish deposition on metal surfaces and corrosive wear.

The first detergent additives used were *metal-containing compounds*, often with high alkaline reserve.

To meet the dramatically increased requirements of modern high-performance engine oils new *ashless dispersants* with improved dispersing properties have been developed. As these ashless compounds possess also cleansing properties in fact there is no real difference between detergents and dispersants. Thus, it seems to be more appropriate to speak of metal-containing and ashless DD or HD additives [150], [151].

Mechanism of Action. Detergents and dispersants are generally molecules having a large oleophilic hydrocarbon 'tail' and a polar hydrophilic head group. The tail section serves as a solubilizer in the base fluid, while the polar group is attracted to contaminants in the lubricant. Solid contaminants are enveloped forming micelles whereby the nonpolar tails prevent the adhesion of polar soot particles on metal surfaces as well as the agglomeration into larger particles (peptidization).

6.4.1. Metal-Containing Compounds (Detergents)

Phenates represent an important class of detergents which are synthesized by reaction of alkylated phenols with elemental sulfur or sulfur chloride followed by the neutralization with metal (calcium, magnesium, barium) oxides or hydroxides. Calcium phenates are currently the

most widely used types. Basic calcium phenates can be produced by using an excess of the metal base. Beside their good dispersant properties they also possess greater acid neutralization potential.

Salicylates are generally prepared by carboxylation of alkylated phenols with subsequent formation of divalent metal salts. Typically also these products are overbased by an excess of metal carbonate (calcium and magnesium) to form highly basic detergents that are stabilized by micelle formation. Salicylates exhibit additional antioxidant properties and have proven effective in diesel engine oil formulations.

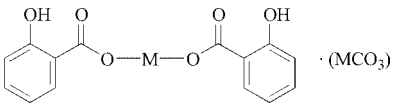

M = Ca, Mg

Thiophosphonates are produced by the reaction of polybutene (molecular mass from 500 to 1000) with phosphorus pentasulfide followed by hydrolysis and formation of metal (calcium, formerly also barium) salts.

The reaction products consist mainly of thiopyrophosphonates combined with thiophosphonates and phosphonates [152].

Sulfonates are metal salts of long-chain alkylarylsulfonic acids which can be divided into the petroleum and synthetic types. Beside their excellent anticorrosion properties (see Section 6.10.1) neutral and especially so-called overbased sulfonates with colloidally dispersed metal oxides or hydroxides (Fig. 30) additionally exhibit an excellent detergent and neutralization potential that makes them very cost-effective multifunctional DD additives for engine oils.

Calcium sulfonates are relatively cheap products with good general performance. Magnesium compounds have excellent anticorrosion properties but tend to form hard ash after thermal decomposition. Barium sulfonates are hardly used anymore due to toxicological concerns.

6.4.2. Ashless Dispersants (AD)

Ashless dispersants are generally derived from hydrocarbon polymers, usually from polybutenes with molecular masses of 500 to 3000 [153].

Polybutenes are thermally coupled with maleic anhydride (MA) to give alkylsuccinic anhydrides. In a further reaction step oligomeric aminoalkylenes are added to the anhydride to form thermally stable imides. By using suitable raw materials and reaction control, this two-stage, one-pot reaction can produce a variety of products. A sample reaction scheme is shown in Figure 31.

Novel developments are monofunctional and bifunctional polyisobutylene succinic anhydrides (PIBSAs). Apart from the bifunctional amines (triethylene tetramines, tetraethylene pentamines), alcohols such as pentaerythritols are used to improve the base character and the resulting interactions with elastomers.

The dispersion of oxidation, nitration and soot particles can be controlled by product mixture and the purity of the components.

Equally chlorine-free alternatives are low-molecular-mass, dispersant PAMAS. Contrary to hydrocarbon-based oligomers, the dispersing units are spread along the chains, whereby the type of dispersing groups can vary greatly [154]. Current research reports advantages of these type of AD with regard to soot dispersing and improved fuel efficiency characteristics [155].

6.5. Antifoam Agents

The foaming of lubricants is a very undesirable effect that can cause enhanced oxidation by the intensive mixing with air, cavitation damage as well as insufficient oil transport in circulation systems that can even lead to lack of lubrication. Beside negative constructive influences the foaming tendency is influenced by the surface tension of the base oil and especially by the presence of surface-active substances such as detergents, corrosion inhibitors and other ionic compounds.

Two types of foam can be discriminated:

Surface foam can be controlled by antifoam agents. Effective defoamers are usually not solu-

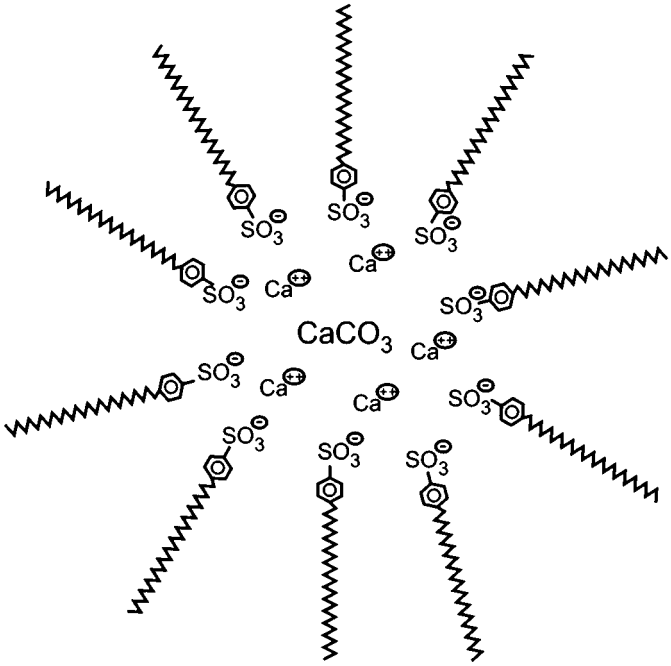

Figure 30. Model of the structure of overbased sulfonates

ble in the base oil and therefore have to be finely dispersed. The particle size of the dispersed defoamers should be smaller than 100 µm or even smaller than 10 µm [156].

The *inner foam* refers to finely dispersed air bubbles in the lubricant that can form very stable dispersions. Unfortunately the common defoamers that control the surface foam tend to stabilize the inner foam. Generally the air release properties of lubricants cannot be improved by additives. Lubricants which need excellent air release properties, e.g., turbine oils have to be formulated using specially selected base oils and additives.

Silicone Defoamers. Liquid silicones, especially linear and cyclic polydimethylsiloxanes, are the most efficient antifoam agents at very low concentrations of 1 to max. 100 mg/kg. To guarantee a stable dispersion silicones usually are predissolved in aromatic solvents.

Compared to other additives silicone defoamers have the disadvantage of being particularly easily carried out of the lubricant due to their insolubility and their strong affinity to polar metal surfaces.

Silicone-free defoamers are being used more and more in many applications, especially in metalworking processes. Cutting fluids and hydraulic fluids have to be silicone-free to guarantee the subsequent application of paints or lacquers on the workpieces.

The main representatives of silicone-free defoamers are special poly(ethylene glycols) and miscellaneous organic copolymers. Also tributylphosphate has been proposed as antifoam agent.

6.6. Demulsifiers

Most of the industrial oils in circulation systems (hydraulic, gear, turbine and compressor oils) require good or excellent demulsification properties to separate water from the lubricating system. Without demulsifiers lubricating oils can form relatively stable water-in-oil emulsions.

In principle, all surface-active substances are suitable demulsifiers. One of the first known types have been alkaline-earth metal salts of organic sulfonic acids particularly barium and calcium dinonylnaphthenesulfonates. Nowadays

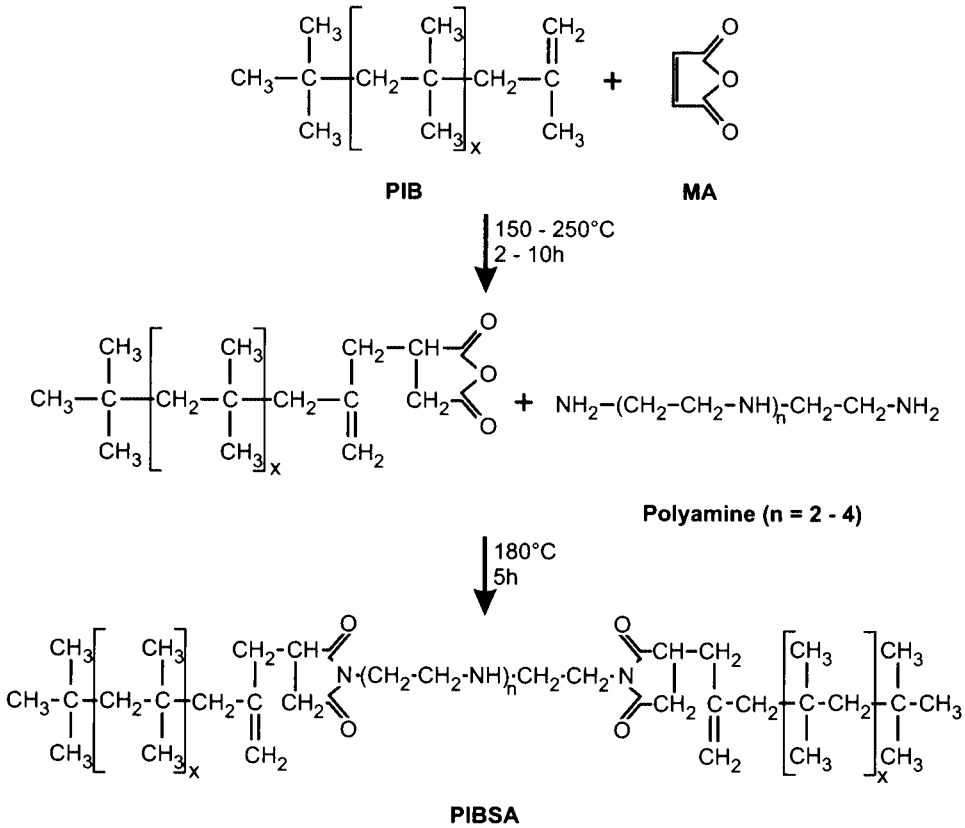

Figure 31. Reaction mechanism for the manufacture of ashless dispersants

special poly(ethylene glycols) and other ethoxylated substances are used.

Surprisingly the same class of chemical substances is used as emulsifiers. Here the molecular mass, the degree of ethoxylation and the treat rate are very important to guarantee the demulsifying properties.

6.7. Dyes

For marketing, identification or leak detection purposes some lubricants contain dyes [157]. Most of these substances are solids which often are suspended in mineral oil or dissolved in aromatic solvents to make their handling easier.

Generally oil-soluble azo dyes are used. Some of these products may be removed in the near future because they have to be labeled as potential carcinogenes.

Fluorescent dyes are typically used to detect leaks under UV light when the coloration of lubricants is not appreciated.

6.8. Antiwear and Extreme Pressure Additives

When two contacting parts of a machinery start to move and the hydrodynamic lubrication has not yet build up or in the case of severe stress and strong forces the lubricating system runs in the area of mixed friction. In this case antiwear (AW) and extreme pressure (EP) additives are necessary in any metalworking fluid, engine oil, hydraulic fluid or lubricating grease to prevent welding of the moving parts respectively to reduce wear.

Function of AW/EP Additives. Because of their polar structure AW/EP additives form

layers on the metal surface by adsorption or chemisorption that guarantees their immediate availability in the case of mixed friction conditions. When the hydrodynamic lubricating film is not yet or no longer valid, temperature will increase and the AW and EP additives can react with the metal surface forming tribochemical reaction layers (iron phosphides, sulfides, sulfates, oxides and carbides – depending on the chemistry of the additive) that will prevent direct contact between the sliding metals.

AW additives are mainly designed to reduce wear when the running system is exposed to moderate stress whereas EP additives are much more reactive and are used when the stress of the system is very high in order to prevent the welding of the moving parts that otherwise would lead to severe damage.

Phosphorus Compounds. Under condition of medium stress organic phosphorus compounds work excellent as antiwear additives.

Most of these additives are neutral and acidic phosphoric acid esters, their metal or amine salts or amides. As the acidic form of these compounds is the most reactive one, the reactivity decreases with the degree of neutralization.

Tricresylphosphate (TCP) is the most important neutral phosphoric acid triester. For toxicological reasons TCP should be free of *o*-cresol.

Amine salts of mixtures of mono- and dialkyl phosphoric acid partial esters are highly efficient FZG boosters. Some types exhibit additional anticorrosion properties.

Ethoxylated mono- and dialkylphosphoric acids are even more polar due to their hydrophilic structure what makes them more efficient.

Also *phosphites* could gain some signification. Because of the inherent hydrolytic instability of this chemical group in practice only sterically hindered derivatives like triarylphosphites and long chain trialkylphosphites are used for some applications.

Compounds Containing Sulfur and Phosphorus. The most important additives of this group are the zinc dialkyldithiophosphates (ZnDTP) [159]. They are synthesized by reaction of primary and secondary alcohols ($C_3 - C_{12}$) or alkylated phenols with phosphorus(V)sulfide followed by neutralization of the resulting dialkyldithiophosphoric acids with zinc oxide. Beside the neutral species also basic ZnDTPs can be obtained when the neutralization is done with an excess of zinc oxide.

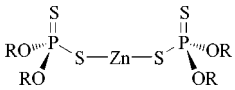

Zinc dialkyldithiophosphate/
zinc dialryldithiophosphate

ZnDTPs based on isopropanol or *n*-butanol are solids, whereas mixtures of short and long chain alcohols are liquid. The thermal and hydrolytic stability of ZnDTPs and thus their reactivity (AW/EP-performance) can be influenced by the structure of the alkyl groups. So the thermal stability increases with the chain length of the alkyl groups and their structure in the sequence secondary, primary and aromatic. By carefully directed alcohol composition the specific requirements of different applications can selectively be adjusted. Beside excellent antiwear and extreme pressure properties ZnDTPs are also efficient antioxidants and even metal passivators. This multifunctional properties makes them the widest spread cost effective additive group that is used nowadays in huge quantities in engine oils, shock absorber oils and hydraulic fluids.

Beside ZnDTPs also *ammonium, antimony, and molybdenum dialkyldithiophosphates* are known. The latter are effective antiwear additives with remarkable friction reducing properties and excellent antioxidant behavior.

Ashless dialkyldithiophosphoric acid-O,O,S-triesters have higher hydrolytic stability than the corresponding acids, but poorer antioxidant properties. Similar to the ZnDTPs the reactivity can be influenced by variation of the organic substituents.

Triphenylphosphorothionate (TPPT) is distinguished by its extraordinary thermal stability what makes it suited for high temperature lubrication.

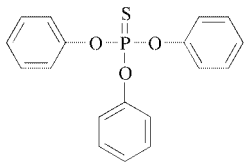

Compounds Containing Sulfur and Nitrogen.

Zinc-bis(diamyldithiocarbamate) and the ashless methylene-bis(di-*n*-butyldithiocarbamate) are highly effective EP additives and excellent antioxidants. Beside these main species also antimony and tungsten derivatives are known. Dithiocarbamates are predominantly used in lubricating greases and to some extent in gear oil formulations.

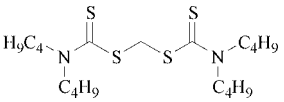

Dialkyl-2,5-dimercapto-1,3,4-thiadiazole (DMTD) derivatives (see Section 6.10) are usually classified as metal passivators and sulfur scavengers, but they also exhibit excellent EP properties. The use of thiadiazoles and dithiocarbamates is limited by their relatively high price.

Sulfur Compounds. From the early days of lubrication until now elementary sulfur has been dissolved directly in mineral oil (up to 1.5 %) to improve the EP properties of metal working fluids. Oil-soluble organic sulfur compounds, the so-called sulfur carriers, of the general formula $R-S_x-R$ offer improved solubility and exact control about the reactivity of the sulfur.

Inactive and active sulfur carriers can be differentiated. The inactive types with predominantly disulfide bridges ($x = 2$) possess relatively stable C – S bonds which will react at elevated temperatures only. The active forms with x between 3 and 5 (so-called pentasulfides) are much more reactive as the sulfur of the relatively labile polysulfide bridges can easily be made available even at low temperatures. Moreover numerous sulfur carriers with specific distribution of the different polysulfide bridges ($x = 1-5$) are used to cover the whole field of application with its varying stress requirements. The mechanism of action begins with physical adsorption followed by chemisorption and finally cleavage of the sulfur – carbon bond and reaction of sulfur with the metal surface (Fig. 32). Generally this reaction takes place at temperatures over 600 °C [158].

Active sulfur carriers are excellent EP additives. Because of the high reactivity with nonferrous metals active sulfur carriers cannot be used in the machining of nonferrous metals. Inactive sulfur carriers need higher temperatures to set the sulfur free. Therefore they are much more compatible with nonferrous metals and show to some extent antiwear properties.

The polarity and thus the affinity to the metal surface is determined by the organic substituents.

Sulfurized hydrocarbons are obtained by direct sulfurization of olefins (e.g., diisobutene, terpenes) with elementary sulfur in the presence of H_2S, their reaction with dichlorodisulfide or oxidation of mercaptans to disulfides. The first method will lead to products with high sulfur contents up to 45 %. The other will form inactive sulfur carriers that are predominantly used in gear oils and mild metalworking fluids (cutting fluids). Active pentasulfides are used for severe machining processes, e.g., broaching.

Sulfurized fatty oils and fatty acid esters are produced by sulfurization of the unsaturated raw materials with sulfur (dark colored products) or with sulfur in the presence of H_2S to obtain light colored products that are used predominantly in modern lubricants instead of the formerly used sperm oil products. The combination of sulfur with the friction reducing properties of the fatty raw materials leads to excellent EP additives with high load carrying capacity in the four ball testing machine. Sulfurized fatty acid methyl esters will lead to products with low viscosities that are used, e.g., for deep-hole drilling.

PEP Additives. Overbased sulfonates, especially calcium and sodium salts, can be used as highly efficient boosters in combination with active sulfur carriers to formulate metal working fluids with extremely high load carrying behavior. These overbased products are called PEPs (passive EP).

Chlorine Compounds. Dispite their excellent AW/EP properties, chlorine compounds are being increasingly replaced by other additives for environmental and toxicological reasons.

Solid Lubricating Compounds (see Chap. 15). Mainly finely ground powders of *graphite* and *molybdenum disulfide* and their dispersions are used as solid additives. Solid lubricants possess excellent emergency running

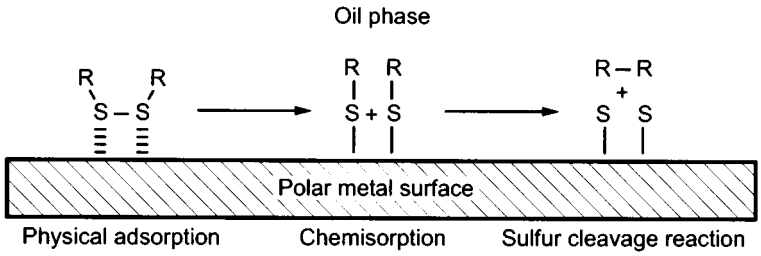

Figure 32. Mechanism of sulfur carriers under extreme pressure conditions

properties when the oil supply is breaking down. Also for extreme high temperature applications solid lubricating compounds are used due to their high thermal stability. Other compounds that are used especially in lubricating greases are *polytetrafluoroethylene, calcium hydroxide* and *zinc sulfide*.

6.9. Friction Modifiers

In the case of liquid (hydrodynamic) lubrication friction can be reduced only by the use of base oils with lower friction coefficients and lower viscosity respectively high viscosity indices. In the area of low slide velocities, moderately increased loads and low viscosities at higher temperatures the liquid lubrication can easily proceed to mixed friction conditions. In this case so-called friction modifiers have to be used to prevent stick – slip oscillations and noises by reducing frictional forces. They work at temperatures where AW and EP additives are not yet reactive. Therefore, friction modifiers (FM) can be regarded as mild AW or EP additives working at moderate temperatures and loads in the area of beginning mixed friction.

Friction modifiers can be classified into different groups regarding their function [160]: mechanically working FMs (solid lubricating compounds, e.g., molybdenum disulfide, graphite, PTFE, polyamide, polyimide, fluorinated graphite), adsorption layers forming FMs (e.g., long-chain carboxylic acids, fatty acid esters, ethers, alcohols, amines, amides, imides), tribochemical reaction layers forming FMs (saturated fatty acids, phosphoric and thiophosphoric acid esters, xanthates, sulfurized fatty acids), friction polymer-forming FMs (ethoxylated dicarboxylic acid monoesters, dialkyl phthalates, methacrylates, unsaturated fatty acids, sulfurized olefins) and organometallic compounds (molybdenum compounds like molybdenum dithiophosphates, molybdenum dithiocarbamates, their synergistic combination with ZnDTPs, copper containing organic compounds).

Typical applications of friction modifiers are modern fuel economy oils, slide way oils, automatic transmission fluids (ATF) that contain so-called anti-squawk additives, and lubricants for limited-slip axles that contain so-called antichatter additives.

6.10. Corrosion Inhibitors

Anticorrosion additives are molecules with long alkyl chains and polar groups that can be adsorbed on the metal surface forming densely packed, hydrophobic layers. The adsorption mechanism can base on a physical or chemical interaction of the polar anticorrosion additive with the metal surface [161].

Because of this high surface activity, anticorrosion additives compete with other polar additives like antiwear and extreme pressure additives for the metal surface and can therefore reduce their efficiency. Thus a FZG damage load stage of > 12 can be reduced down to a damage load stage of 8 if highly efficient and polar anticorrosion additives are added.

Corrosion inhibitors can be divided into two main groups: antirust additives for the protection of ferrous metals and metal passivators for nonferrous metals.

6.10.1. Antirust Additives

Sulfonates. *Petroleum sulfonates* (mahogany sulfonates) are byproducts at the pro-

duction of white oils by treatment with oleum. The resulting acid tar contains long-chain alkylarylsulfonic acids that can be neutralized with lyes. Sodium sulfonates with low molecular masses (below approx. 450) are typically used as low-priced emulsifiers and detergents with additional anticorrosion properties in water based metal working fluids, engine oils and rust preventatives. Sulfonates with higher molecular masses are highly efficient corrosion inhibitors especially when based on divalent cations like calcium, magnesium, and barium. The importance of the barium compounds decreasing constantly due to toxicological and ecotoxicological concerns.

Nowadays *synthetic alkylbenzene sulfonates* are used preferably in spite of a higher price level due to their higher and more constant quality. They are reaction products of specifically designed monoalkylbenzenesulfonic acids (typically C_{24} alkyl groups) and dialkylbenzenesulfonic acids (typically two C_{12} alkyl groups) with alkaline and alkaline-earth metal hydroxides.

A special group of synthetic sulfonates are the *dinonylnaphthenesulfonates* of which the neutral calcium and barium salts are distinguished by additional demulsifying properties and a good compatibility with EP additives.

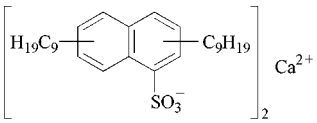

Beside the neutral or only slightly basic sulfonates *overbased sulfonates* with high alkaline reserve (TBN 100 to 400 mg KOH/g) play an important role especially in the formulation of engine oils. There they exhibit detergent properties and can neutralize acidic oxidation products.

Carboxylic Acid Derivatives. *Lanolin* (wool fat) and salts of the lanolin fatty acid mostly in combination with sulfonates have long been known as corrosion inhibitors in rust preventatives. *Oxidized paraffins* with their high content of hydroxy- and oxocarboxylic acids are still used for that purpose.

Zinc naphthenates are especially used in lubricating greases.

Alkylated succinic acids, their partial esters and half amides are known as highly efficient, not emulsifying antirust additives even at very low treat rates of 0.01 to 0.05 %. Therefore these additives are used preferably in turbine oils and hydraulic fluids. *4-Nonylphenoxyacetic acid* and derivatives have a similar performance.

Another widespread group are *amides and imides* of saturated and unsaturated fatty acids. The best known product of this type is *N-acylsarcosine* that shows a strong synergistic effect with *imidazoline derivatives*. Additionally these additives have good water-displacing properties.

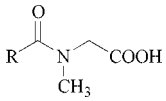

Amine Neutralized Alkylphosphoric Acid Partial Esters. Some special amine salts of mono- or dialkylphosphoric acid partial esters exhibit excellent anticorrosion properties in addition to their highly efficient antiwear properties [162]. Because of the well-known antagonism of anticorrosion and antiwear additives this behavior makes them one of the mostly used components in ashless industrial oils.

Vapor phase corrosion inhibitors (VCIs) for closed systems are substances with high affinity to metal surfaces and relatively high vapor pressure to guarantee their availability on parts that are not steadily in direct contact with the corrosion inhibited lubricant. The mostly used product group for this application are amines. *Morpholine*, *dicyclohexylamine*, and *diethanolamine* have proved to be highly efficient for that purpose. Because of toxicological concerns, that refer mainly to the nitrosamine forming potential of secondary amines, these products are being partly substituted by tertiary amines like *diethanolmethylamine* and similar products.

Another group of oil soluble VCIs are low molecular *carboxylic acids* (*n-C_8 to n-C_{10}*).

6.10.2. Metal Passivators

Metal passivators can be classified into three groups: film forming compounds, complex forming chelating agents, and sulfur scavengers.

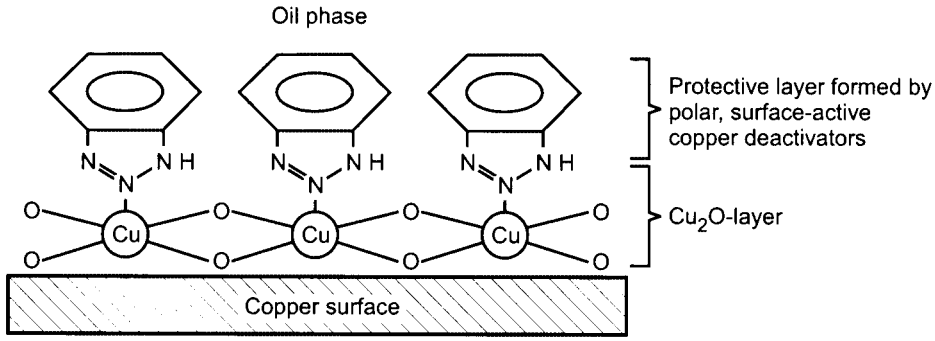

Figure 33. Function of benzotriazole as film forming agent

The fundamental function of the film forming types consists in the building of passivating protective layers on the nonferrous metal surface thus preventing the solubilization of metal ions that would work as pro-oxidants. The complex forming agents are able to build oil-soluble complexes with significantly reduced catalytic activity regarding the influence of nonferrous metal ions on the oxidative aging process of lubricants. Sulfur scavengers are even able to scavenge corrosive sulfur by integrating it into their molecular structure. Mechanism and function of these three types of metal passivator are shown in Figures 33, 34, and 35.

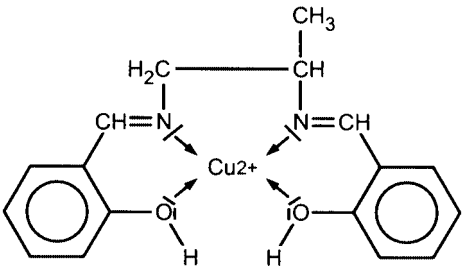

Figure 34. Function of *N,N′*-disalicylidene-propylenediamine as chelating agent

Metal passivators in combination with antioxidants show strong synergistic effects as they prevent the formation of copper ions respectively suppress their behavior as pro-oxidants. Thus these additives are used in nearly every formulation of modern lubricants.

The mostly used metal passivators are *benzotriazole* and *tolyltriazole* as well as their alkylated liquid derivatives with improved solubility. Typical treat rates of these film forming passivators are between 0.005 and 0.03 %.

Furthermore *2-mercaptobenzothiazole* (reduced importance) and especially *2,5-dimercapto-1,3,4-thiadiazole* derivatives are used as highly efficient film forming passivators. The later can also act as sulfur scavenger by building in sulfur into the alkyl – sulfur – thiadiazol bonds (see Fig. 35).

N,N′-disalicylidenealkylendiamines (where 'alkylen' is ethylene or propylene) belong to the group of chelating agents. Beside their EP and antiwear properties *zinc dialkyldithiophosphates* and *dialkyldithiocarbamates* have some metal-passivating properties (see Section 6.8).

7. Lubricants in the Environment

Lubricants and functional fluids are omnipresent due to their widespread use and they thus pollute the environment in small, widely-spread amounts and rarely in large, localized quantities.

Normally biodegradability of at least 60 % according to OECD 301 or 80 % according to CEC L-33-A-93 is considered as main objective criteria for 'Bio-Lubricants'. But the other criteria will get more and more importance in the near future.

7.1. Current Situation

Every year, about 40 – 50 % of the approximately 5×10^6 t of used lubricants in Europe end up polluting the environment. This is either technically desired (total-loss lubrication) or follows leaks, or other problems.

The environmental damage caused by mineral oil-based lubricants is largely caused by the

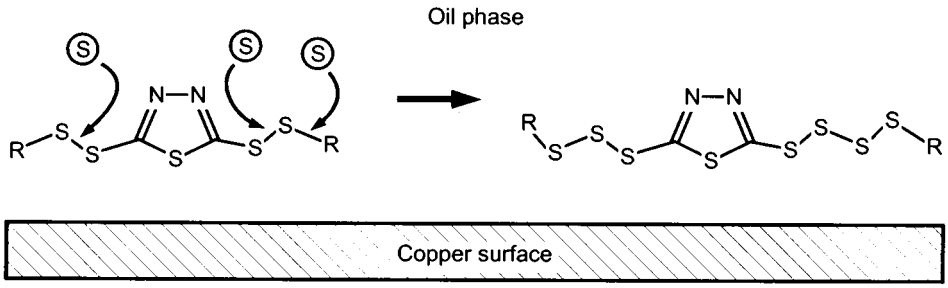

Figure 35. Function of dialkyl-2,5-dimercapto-1,3,4-thiadiazole as sulfur scavenger

approximately 40 % of lubricants which are not properly disposed of. This figures include total-loss applications, the residual oil in about 90 million oil cans and 20×10^6 oil filters, spillages during topping-up, leaks, drips from separated oil-line and hydraulic couplings, accident losses and all manner of emission losses (CONCAWE Report no. 5/96).

Even though mineral oil products can be relatively rapidly biodegraded by the microorganisms present in nature, these natural degrading systems are exhausted by the volume of the losses.

In the past, mineral oil-based products were used almost exclusively. Leakages and other causes allowed a part of these lubricants to pollute the environment. The problem has spurred society and politics into action, reflected by diverse legislation and recommendations as well as the issuing of environmental awards to products which cause less damage. The use of environmentally friendly materials and substances will play an increasingly important part in the scope and procedures of the EU Eco Audit.

Technically, and disregarding development and overall costs, more than 90 % of all lubricants could be made rapidly biodegradable. However, this change will only take place if the ecological benefits of using renewable raw materials is made clear and best of all, quantified in monetary terms.

7.1.1. Economic Consequences and Substitution Potential

In Germany the selection of rapidly biodegradable lubricants, especially those classified as 'not water pollutant' reduce the expense of oil spillages or disposal as well as the cost of complying with health and safety regulations to a minimum. Compared to mineral oil-based products, the selection of low-evaporation ester and vegetable oil-based lubricants can allow lower viscosities to be used. This enables faster machining and cuts drag-out losses. This, in turn, increases economy, increases productivity and generates product optimization.

The vast majority of rapidly biodegradable lubricants currently on the market are based on saturated or unsaturated ester oils. The acids required for esterification are mostly from vegetable oils. Compared to equivalent mineral oils, ester oils are significantly more expensive. This is a major barrier to the wider acceptance of such oils. But the overall cost of a lubricant results from the interaction between the following characteristics (in alphabetical order):

- Aging and temperature stability
- Disposal costs
- Machine compatibility (multifunctional oil)
- Machine investments
- Maintenance measures
- Oil change intervals
- Operating costs
- Reconditioning possibilities
- Reduction in storage costs
- Reduction in emission reduction measures
- Reduction in the cost of preventative health and safety measures
- Risk minimization and lower clean-up costs after spillages
- Simpler plant license procurement

Completely new mix of economic aspects can result from every new application. In particular, the use of fully saturated synthetic esters offers a number of opportunities. Pure vegetable oil for-

mulations are now only used for lower demands, for example total-loss chain saw or mold release oils and some types of greases.

7.1.2. Agriculture, Economy, and Politics

The world market price of vegetable oils is largely influenced by the price of soybeans. Since half of the world's soy production comes from the USA, vegetable oil prices are heavily dependent on the size of US soy harvest. On the German market, the price of rapeseed (in Hamburg, Rostock or Mannheim) is particularly significant.

Increasing sales of environmentally harmless lubricants could have a positive effect on prices and this, in turn, would make such lubricants even more interesting.

Until now, in Europe rapeseed and sunflower oil as lubricant base oils have been in the forefront because of the following reasons:

- They are abundantly available
- Their thermal oxidation stability are acceptable in some applications
- Their flowing properties are better than those of other vegetable oils

The disadvantages of the abovementioned vegetable oils are:

- The oxidation, hydrolytic and thermal stability of vegetable oils is not adequate to allow their use in circulating systems and
- High-performance additives for vegetable oils have not undergone sufficient development, especially those which are ecotoxicologically safe and ecologically harmless.

Apart from pure vegetable oils, oleochemical esters can also be used because their thermal-oxidation stability compared to rapeseed oil opens up a wider field of applications. However, these products are considerably more expensive, a factor which in turn, limits their use.

In 1999, nevertheless, 20 000 ha of rapeseed were cultivated in Germany only for application in lubricants. This corresponds to a volume of nearly 20 000 t/a from which an approximately similar quantity of lubricating or hydraulic oils are manufactured. It can be assumed that about 20 000 to 25 000 t of oleochemical esters are used for biodegradable lubricants per year so that biodegradable lubricants account for just over 4 % of all lubricants (Germany 1999).

In Germany as most important market for biodegradable lubricants today a number of political initiatives were started in recent years, of which some were included in two comprehensive reports which were submitted to the German parliament.

- Resolution dated June 16, 1994: Measures to promote the use of rapidly biodegradable lubricants and hydraulic fluids
- Resolution dated November 28, 1995: News from the Federal Ministry for Food, Agriculture and Forestry: 'Report on the use of rapidly biodegradable lubricants and hydraulic fluids and measures taken by the Government' (New Edition 1999)
- Regulation for introduction of environmental friendly lubricants to market (June 2000): This program supports the substitution of mineral based lubricants by ester based products.

The use of environmentally friendly fluids is mandatory in the public sector.

Research and Development on ecologically harmless fluids and the clarification of unanswered questions is sponsored through projects.

7.2. Biodegradable Base Oils for Lubricants

The most important biodegradable base oils are different types of ester oil, e.g.:

- Vegetable oil from harvestable raw materials, for example rapeseed or sunflower oil
- Semisaturated, transesterified ester oils with natural fatty acids, for example trimethyolepropane trioleate
- Fully saturated, synthetic esters based on chemical modified vegetable oils or mineral oil, for example, diisotridecyl adipate

Other in principle biodegradable base oils mostly are not considered as 'environmental friendly lubricants' because of their petrochemical origin:

- Poly(alkylene glycols) (PAG)
- Low viscosity polyalphaolefins (PAO 2)
- Some special types of synthetic hydrocarbon up to a viscosity of 6 mm^2/s at 100 °C

Natural Fats and Oils (Triglycerides). The worldwide market for natural oils amounts to 100×10^6 t/a. The dominant crops are soybean (Northern America), palm (SE Asia, Southern America) and rapeseed oil (Europe).

In Germany, nearly 900 000 t vegetable oils are used only for industrial applications – with increasing numbers in the last years. The main part today is transesterified to biodiesel (rapeseed methyl ester).

Apart from the abovementioned environmental aspects, the use of natural oils also has an agri-political benefit of enabling the agricultural industry to shift its production to technical raw materials. The lubricants industry has invested significant sums of money in developing and marketing biodegradable lubricants.

Natural fatty oils such as castor oil, palm oil, rapeseed oil, soybean oil, sunflower oil, lard, degras and sperm oil, i.e., triglycerides of more or less unsaturated fatty acids, have been used in lubricants for years. These base oils are biodegradable and, compared to mineral oils, have excellent tribological properties (low friction coefficient, good wear protection). Their range of use is limited by lower stability against thermal oxidative and hydrolytic stress and partly inferior cold flow properties. These limits can be improved gradually either with additives, or with the selection, cultivation or genetically modification of new types of plants. With new types of 'High oleic sunflower oils' (HOSO) with an amount of oleic acid of more than 90 % it seems possible to formulate oils for higher performance levels.

7.2.1. Synthetic Esters

The collective name 'synthetic esters' covers a broad range of chemicals with different qualities and prices. For the development of environmentally acceptable lubricants, esters have to be selected which fulfil the ecological requirements and have more favorable properties than natural fatty oils.

These properties are mainly thermal-oxidative resistance, better low temperature behavior and better resistance to hydrolysis.

At present, polyol esters such as trimethylolpropane esters (TMP esters) or glycerol trioleate dominate. The basis of these are mainly alcohols from petrochemical and oleochemical origin and fatty acids derived from natural oils.

In regard to hydrolytic stability, 'normal' polyol esters differ only slightly from rapeseed oil; the difference in oxidation resistance is much greater. Both characteristics are significantly improved with complex esters of polyols and saturated fatty acids. Normally an improvement in hydrolytic stability worsens the base fluid's biodegradability.

The most important chemical reactions to improve the properties of esters are transesterification, (selective) hydrogenation, ozonolysis, dimerization.

7.2.2. Polyalkylene Glycols (see Section 5.4)

Polyethylene glycols (PEG) are biodegradable. They are not miscible with mineral oils or esters, but are water-soluble which, at the present time, is considered to be a disadvantage. If PEG is released through a leak or an accident, it will migrate quickly in the ground or in the water.

The chemical industry is attempting to develop biodegradable, non-water-soluble PAGs. These could be alternatives to esters.

7.2.3. Polyalphaolefins (see Section 5.1.1)

Low viscous polyalphaolefins (PAO 2) are biodegradable. These base stocks have only limited application in the formulation of lubricants. The new types of biodegradable, higher viscosity synthetic hydrocarbons may have a greater influence in the near future.

7.2.4. Relevant Properties of Biodegradable Base Oils

Evaporation Loss. The evaporation loss of esters and vegetable oils is very low. The lower evaporation loss of ester oils has great advantages for example in regard to emissions from machine tools in metalworking and for the emissions, especially the particle emissions, of internal combustion engines.

Viscosity – Temperature Behavior. Using an ester instead of a mineral oil improves the viscosity – temperature behavior of a lubricant. The higher viscosity index (*VI*) of an ester results in a wider temperature range in application but with the recommended working viscosity. Also the higher *VI* can lead to polymer-free multigrade lubricants with improved shear stability.

Boundary Lubrication. Most vegetable oils, synthetic esters and glycols display excellent lubricity in boundary lubrication conditions. Because of their high degree of polarity these lubricants are superior to mineral oil based lubricants. This has been proven by a series of tests. Experimental investigations on twin disk test rigs showed that the friction coefficient of vegetable oils, synthetic esters and glycols is half that of mineral oils:

Figure 36 compares feiction coefficients of a mineral oil and a TMP ester of the same viscosity grade. As can be seen, the friction coefficient for this ester is half that of a mineral oil.

7.3. Additives (see also Chap. 6)

Extreme Pressure/Antiwear Additives. Most vegetable oils and synthetic esters used in lubricants display have a high degree of polarity. This characteristic results in better lubricity than mineral oils in boundary conditions. Sulfurized fatty materials are environmentally friendly EP additives. Sulfur as an additive in esters provides an AW effect. Sulfur carriers with 15 % total sulfur and 5 % active sulfur have proven effective in rapidly biodegradable esters and vegetable oils.

Corrosion Inhibitors. Vegetable oils and synthetic esters show a high polarity. This also applies to corrosion inhibitors and can result in a competitive reaction on the metal surface.

Special calcium sulfonates, succinic acid derivatives or ashless sulfonates can be used to provide corrosion prevention.

Antioxidants. Special phenolic and aminic materials are suitable antioxidants for the formulation of biodegradable lubricants.

7.4. Products (Examples)

Because of their physical properties, refined, chemically unchanged vegetable oils can be used as alternatives to petroleum products in a large number of lubricants. Compared to mineral oils, their performance in most applications is limited by their relatively poor aging resistance. The reasons for this are the unsaturated hydrocarbon chains in the natural fatty acids and the easy hydrolysis of the ester compounds.

Vegetable oils, at present, are primarily used for total-loss applications and products with relatively low technical specifications. At face value, using rapeseed oil as a basis for engine oils does not appear very promising – the present operating conditions of engine oils cannot be satisfied by chemically unmodified vegetable oils.

Over many years the development in the lubricants market had the main aim to formulate products for 'long oil drain intervals' or for lifetime. This was not possible with vegetable oil. But today the opposite development can be recognized in some applications – the substitution of circulating oils by total-loss lubricants:

- Example 1: Minimum Quantity Lubrication in metalworking
- Example 2: Die-casting and forming instead of cutting and grinding
- Example 3: Oil-refreshing systems for lubrication of four-stroke engines

These ideas are not primarily intended for use of chemically unchanged vegetable oils in new applications. The main idea behind all examples is the lower overall impact of lubricants and their degradation products on the environment by use of total-loss lubricants.

For four-stroke engines, a novel lubrication concept for diesel engines was developed which counteracts the relatively poor aging resistance of vegetable oils by continuous refreshing (Plantotronic system). This has permitted the use of rapeseed and sunflower oils for engine lubrication. In this lubrication system, used oil is gradually burned along with the fuel itself without negatively influencing emission values.

The concept can be seen as a way of continuously refreshing a vegetable-based engine oil or as a way of performing a continuous oil change. Combustion in the engine was only made possible by the use of low-additive vegetable oils –

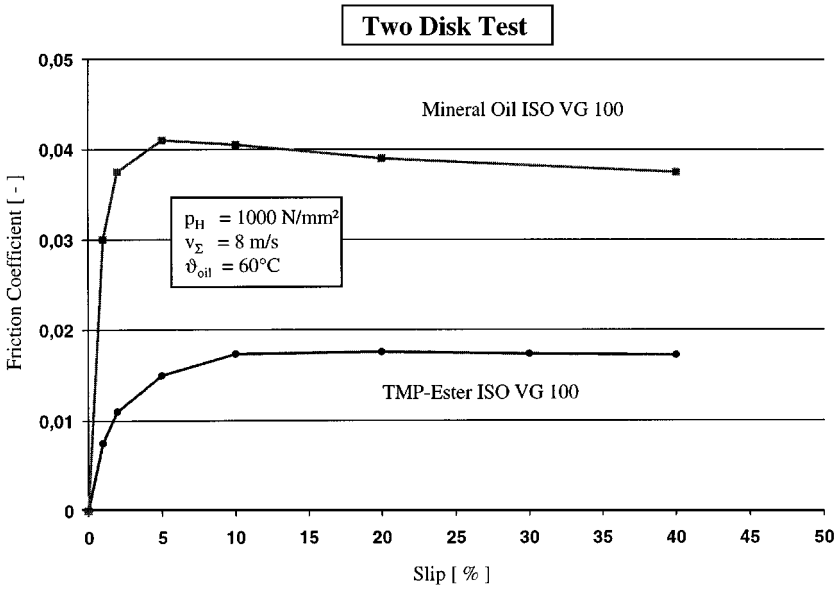

Figure 36. Friction coefficients of two base fluids as function of slip

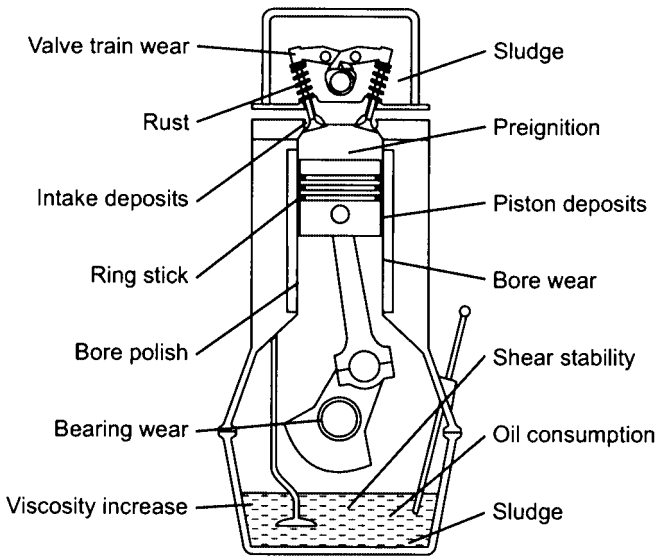

Figure 37. Scheme of lubrication points in an engine

burning conventional oils seems highly contentious considering the additives commonly used in such oils. The use of vegetable oils for engine lubrication depends on the progressive burning of used engine oil with the fuel. The advantage of the oil refreshing method is that chemically unmodified, low additivated vegetable oils can be used and that no waste oil has to be disposed of. Another advantage is the automated oil-change, which helps to reduce the maintenance costs especially for stationary engines (power plants).

8. Lubricants for Internal Combustion Engines

8.1. Four-Stroke Engine Oils

Technically and commercially, engine oils are the number 1 among lubricants and functional fluids; on the global lubricants market, they account for more than 60 %.

While demand in Europe is more or less stagnating and will decline slightly in the foreseeable future in spite of increasing vehicle registrations, there is still considerable growth potential in Southeast Asia and third world threshold countries (China, India, Korea etc.) [201].

8.1.1. General Overview

Historically the development of engine oils over the last 50 years has been focussed on the specifications issued by the international automobile industry. Starting with the first specifications published by the US Army (MIL specs), today there are three internationally recognized sets of minimum requirements. In Europe, these are ACEA, in the USA API, and ILSAC in Asia. Details of these specifications can be found in Section 8.1.3.

In principle, all specifications reflect the successive adaptation of oil qualities to developments in engine design. Back in the 1950s, monogrades dominated the engine oil market. The viscosity of these oils was matched to the prevailing ambient temperature and therefore had to be changed between summer and winter. The 1960s saw the development of mineral oil-based multigrades, i.e., combined summer and winter oils, initially high-viscosity types (SAE 20W-50) and later the present standard 15W-40. Base oil distillation and refining processes caught up with the new viscosity and performance requirements (see Chap. 4). This led to new qualities with comparatively low evaporation losses and optimized cold flowing properties. Semisynthetic, and since the 1920s, fully synthetic oils have dominated the premium-quality market, especially in the passenger car sector. Since the 1990s, a niche for environmentally friendly, biodegradable products has been created in the German-speaking market [202].

8.1.1.1. Fundamental Principles

Engine oils have to fulfil a wide range of functions in engines. The purely tribological task consists of guaranteeing the functional reliability of all friction points in all operating conditions (Fig. 37). In addition, engine oils have to perform a number of other functions. This begins with the sealing the cylinder and ends with the transport of sludge, soot, and abraded particles to the oil filter.

Starting with the tribological functions, the three classic sections of the Stribeck graph are satisfied, from hydrodynamic full lubrication to the elastohydrodynamic area in bearings to the boundary friction conditions at TDC and BDC [203]. All friction pairings and a whole series of parameters are covered. The temperatures encountered range from ambient in the Arctic (− 40 °C) to sump temperatures of 100 °C to peak values of over 300 °C under the piston crown.

During the combustion process, the engine oil helps to seal the piston and cylinder. At the same time, it should burn off at the cylinder wall without leaving any residue. The engine oil dissipates heat from the piston and thus cools it. The blow-by gases formed when fuel is burned and their reaction byproducts have to be neutralized and held in suspension. The same applies to the soot and sludge particles formed by incomplete combustion. The oil also transports dirt and any abraded particles to the oil filter and ensures its filterability. In addition, any water formed during the combustion process should be emulsified and even when higher water concentrations are present and when the phases separate as temperature falls, the oil should protect against corrosion.

Engine oils should reduce friction and wear during extreme, low-temperature start-ups as well as when the lubricating film is subject to high temperature and pressure in bearings and around the piston rings. Whereas the oil should still flow well and be pumpable without aeration at low temperatures (down to − 40 °C) to avoid metal-to-metal contact during cold start-ups, the lubricating film must perform satisfactorily in bearings and hydraulic tappets [204]. At low temperatures, additives must not precipitate and the oil must not gel. At the upper end of the

temperature scale, the oil must offer far-reaching resistance to thermal and mechanical aging. And finally, the stability of the lubricating film should not be diminished by fuel dilution of up to 10 %.

The viscosity of an engine oil is an indicator of how readily a load-carrying film can be formed at all lubrication points in an engine. The adequate and rapid circulation of the oil at low temperatures [205] must be achieved at cold start cranking speeds. On the other hand, viscosity must not fall too much at high temperatures so that adequate lubricating film stability is given at high thermal loads.

Corresponding threshold values were determined by the SAE and the ASTM as shown in Table 8.

According to this list, all viscosity grades can be described by their minimum kinematic viscosity at 100 °C. Additional dynamic viscosity thresholds apply to winter grades, which display the letter W. These values are determined in cold cranking simulators (CCS) or in minirotary viscometers (MRV). The dynamic viscosity value given by the CCS is a measure of flow properties [206] at low temperatures whereby the high shearing rate can mask paraffin crystallization. In the MRV, a so-called threshold viscosity of max. 60 000 mPa · s has been determined for 10 °C lower temperatures to ensure that the oil pump does not draw air.

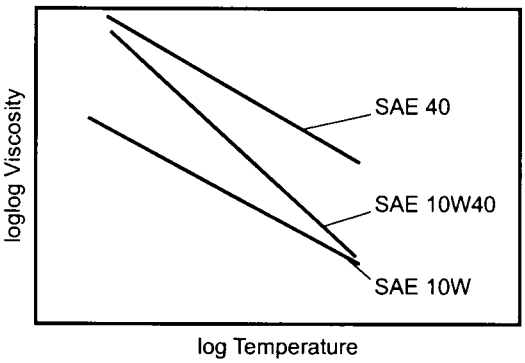

Figure 38. Comparison of monogrades and multigrades

The high-temperature high-shear viscosity is an additional criterion for evaluating lube film stability at high shear rates and high temperatures in summer grades. Modern engine oils are multigrade oils whose low temperature characteristics are indicated by the W and the high temperature viscosity by the number following the W. The $V-T$ behavior of multigrade oils is schematically sketched in Fig. 38.

8.1.1.2. Performance Specifications

The engine oil market is shifting away from conventional products to semisynthetic and synthetic formulations. Along with their higher price, the economic and ecological demands on these oil have also risen (see Chap. 4).

At present, the oil change intervals for cars are between 10 000 and 20 000 km and 30 000 and 60 000 km for trucks. In the future, these figures are expected to double. A retrospective view of specific oil performance shows that this development has been in progress for the last 50 years [207]. Figure 39 shows that oil consumption per energy unit has fallen eight times during this period.

As the established CEC engine tests do not allow comprehensive testing of all required oil properties, additionally in-house methods are now being used. A series of automobile manufacturer original manufacturer (OEM) specific, long-term trials, which have now been taken over for developing oils, are part of these engine tests. In total, these tests represent an enormous technical advance and financial expense because the engine tests themselves have been supplemented by radionuclide techniques [208]. The advantages of this technology lie in the on-line monitoring of wear in defined running conditions, from running-in to full-throttle operation as well as the selective examination of critical engine components.

8.1.1.3. Formulation of Engine Oils

Engine oils are complex mixtures of base oils and additives. Compared to other groups of lubricants, the base oils [209] play an important role (see Chaps. 4 and 5). Mixtures of base oils are selected which have the necessary viscosity and performance to correspond to a rough classification. The final products are then marketed as conventional, mineral oil-based, unconventional semisynthetic (hydrocracked) and synthetic engine oils. Precise international nomenclature divides base oils into five groups:

Table 8. Engine oil viscosity classification (SAE J 300 revised 4/1997).

SAE viscosity grade	Low temperature cranking viscosity (mPa · s) at temp. in °C (CCS)	Low temperature pumping viscosity (mPa · s) at temp. in °C (MRV)	Low shear-rate kinematic viscosity at 100 °C (mm^2/s)		High shear-rate viscosity at 150 °C and $10^6\ s^{-1}$ (mPa · s)
	Maximum	Maximum	Minimum	Maximum	Minimum
0 W	3250 at − 30	60 000 at − 40	3.8	–	–
5 W	3500 at − 25	60 000 at − 35	3.8	–	–
10 W	3500 at − 20	60 000 at − 30	4.1	–	–
15 W	3500 at − 15	60 000 at − 25	5.6	–	–
20 W	4500 at − 10	60 000 at − 20	5.6	–	–
25 W	6000 at − 5	60 000 at − 15	9.3	–	–
20	–	–	5.6	< 9.3	2.6
30	–	–	9.3	< 12.5	2.9
40	–	–	12.5	< 16.3	2.9*
40	–	–	12.5	< 16.3	3.7**
50	–	–	16.3	< 21.9	3.7
60	–	–	21.9	< 26.1	3.7

* For 0 W, 5 W, 10 W. ** For 15 W, 20 W, 25 W and Monogrades.

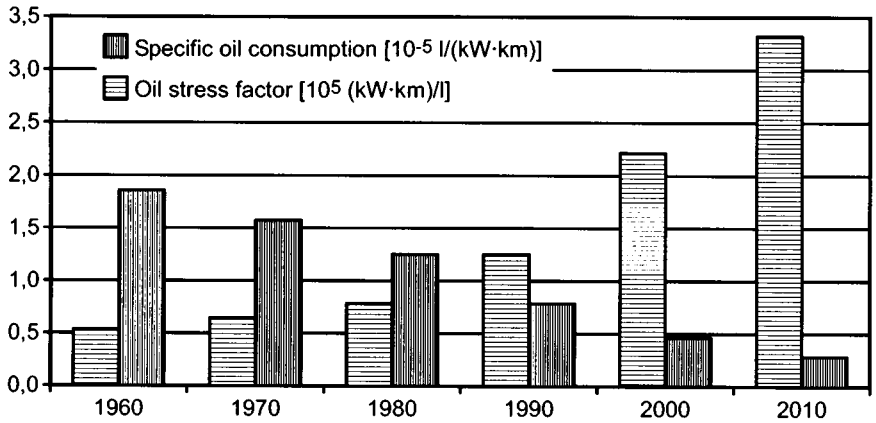

Figure 39. The history of oil stress

- Group 1: SN mineral oils with saturates < 80, 80 < VI < 120, S > 0.1
- Group 2: HC oils with saturates > 80, 80 < VI < 120, S < 0.1
- Group 3: HC oils with saturates > 80; VI > 120, S < 0.1
- Group 4: Polyalphaolefins
- Group 5: Esters and others

8.1.1.4. Additives

Depending on the base oil used and the required engine performance, engine oils can contain up to 30 different additives whose percentage content can range from 5 to 25 % in total. As a rule, performance additives make up the largest group.

Performance Additives. The following species of chemical components are summarised under the general term of performance additives (Table 9).

Table 9. Performance additives

Antioxidants	phenols and amines
Antiwear agents	metal dithiophosphates, carbamates
Detergents	Ca and Mg sulfonates, phenolates, salicylates
Dispersants	polyisobutene and ethylene-propene oligomers with nitrogen and/or oxygen containing functional groups
Friction modifiers	MoS compounds, alcohols, esters, fatty acid amides, etc.
Antimisting agents	silicones and acrylates

These additive groups are described in detail in Chapter 6. Particularly in the case of engine oils, the substance categories listed generally perform more than one function. ZnDTPs, for example, are primary antiwear additives but also have a secondary antioxidant character resulting from a specific decomposition mechanism [210]. Furthermore, complex formulations of a number of individual components typically display synergistic as well as antagonistic interactions which have to be matched to the application considered. The composition of the base oil components has an additional effect on these specific interactions. Considerable experience and know how is thus necessary to create an optimum formulation.

Viscosity Modifiers and Pour-Point Depressants (see Section 6.2). Viscosity modifiers can be divided into two groups, the nonpolar, nondispersing and the polar, dispersing group. The first group are only really needed to set the viscosity of multigrade oils. In an absolute concentration of 0.2 to 1.0 %, viscosity modifiers can generate a viscosity increase of between 50 and 200 % depending on chemical structure and base oil solubility. Due to special modification, dispersing viscosity improvers are often ashless dispersants with additional thickening effects. Furthermore, viscosity improvers and pour point depressants have an effect on low temperature behavior of a formulation (pour point, cold cranking simulator, mini rotary viscometer) and are essential parts of the high-temperature high-shear viscosity. In the USA, additional demands are made on low-temperature stability (gelation index) which cannot be achieved without the viscosity improvers and pour point depressants being matched to the base oil.

8.1.2. Characterization and Testing

8.1.2.1. Physical and Chemical Testing

The physicochemical properties of an engine oil can be determined in the laboratory with standard test methods (Section 16.3). This characterization mainly focuses on rheological test values and the previously shown SAE classification system.

Various viscosity tests are used to determine exact low- and high-temperature viscosities [211]. The viscosity thus determined is a characteristic of the engine oil at a defined engine state. At low temperatures (− 10 to − 40 °C), a mini rotary viscometer with a low shear gradient is used to determine the apparent viscosity and thus the oil's flowability in the area of the oil pump. In addition, maximum viscosity as the threshold of viscosity is determined in five graduated steps. The dynamic CCS viscosity, which is determined at − 5 to − 30 °C with a high shear gradient, is also an apparent viscosity which represents the tribological conditions at the crankshaft during cold starts. The maximum values laid down in SAE J 300 guarantee reliable oil circulation during the start-up phase. The rheological characteristics at higher thermal loads which occur during full-throttle operation are described by the dynamic viscosity at 150 °C and a shear rate of 10^{-6} s^{-1} or high-temperature high-shear (HTHS). The corresponding threshold values also guarantee an adequate lube film even in these conditions.

Apart from the rheological characteristics, the Noack evaporation test (see Section 16.3) is used to test the volatility of base oils and additives as well as foaming tendency and air release. The compatibility of high-additive oils with seal materials is tested on standard reference elastomers in static swelling and subsequent elongation tests [212]. The viscosity loss resulting from mechanical load is described in Section 6.2.

8.1.2.2. Engine Testing

Because realistic engine oil tests cannot be performed only over year-long field trials, a number of international committees have created methods of testing engine oils in defined test engines operated in reproducible and practically relevant conditions.

In Europe, the Coordinating European Council for the Development of Performance Tests for Lubricants and Fuels (CEC) is responsible for testing, approval and standardization [213]. Performance requirements are set-up in the form of Association des Constructeurs Europeen d'Automobiles (ACEA) limits which are decided together with the additive and lubricant industries. In the USA, this task is performed by

the automobile industry and the API. This institution lays down test procedures and limits. The Asian ILSAC has largely adopted the American specifications for automobiles.

In principle, the test procedures detailed in Section 8.1.3 focus on the following general performance criteria:

- oxidation [214] and thermal stability
- dispersion of soot and sludge particles
- protection against wear [215] and corrosion
- foaming and shear stability [216]

8.1.2.3. Passenger Car Engine Oils

Car engines include all gasoline and light diesel engines with direct or indirect injection. Relevant passenger car engine tests are listed in Table 10.

Table 10. Passenger car engine tests

Test engine	Test procedure	Test criteria
Peugeot XUD 11	CEC L-56-T-95	soot handling piston cleanliness
Peugeot TU3 H	CEC L-55-T-95	oxidation cleanliness ring sticking
Peugeot TU3 S	CEC L-38-A-94	cam and tappet wear
Sequence II D	ASTM STP B15 M P1	bearing corrosion
M 111 SL	CEC L-53-T-95	black sludge cam wear
Sequence IIIE	ASTM STR 315 M P2	oxidation wear cleanliness
Sequence VE		sludge piston cleanliness corrosion and wear
BMW M52		valve train air entrainment wear
VW T4		oil oxidation TBN depletion piston cleanliness
M111 FE	CEC L-54-T-96	fuel efficiency
VW-DI	P-VW 1452	piston cleanliness ring sticking
VW-TD	CEC L-46-T-93	piston cleanliness ring sticking

8.1.2.4. Engine Oil for Commercial Vehicles

Commercial vehicles include trucks, buses, tractors, harvesters, construction machines and stationary machinery powered by diesel engines. Apart from the prechamber diesels engines which have largely been superceded in Europe, the engines are usually highly turbocharged direct injection motors. Economic and ecological aspects along with high injection pressures, have improved combustion and thus reduce emissions. As an initiative of ACEA, oil change intervals have been extended up to 100 000 km for long haulage.

Long life and reliability are the criteria for the commercial vehicle sector. The HD oils have to match these requirements. Relevant heavy-duty engine tests are listed in Table 11.

Table 11. Heavy-duty engine tests.

Test engine	Test procedure	Test criteria
Caterpillar 1K/1N		piston cleanliness wear oil consumption
Cummins M11		valve train wear sludge
Mack T8		soot handling
Mack T9		cylinder wear
GM 6.2 Liter		valve train wear
OM 364LA	CEC L-42-T-99	piston cleanliness cylinder wear sludge oil consumption
OM 602A	CEC L-51-A-98	wear cleanliness oxidation oil consumption
OM 441LA	CEC L-52-T-97	piston cleanliness cylinder wear turbocharger deposits

HD oils are allocated to the following categories with increasing performance:

- heavy duty (HD)
- severe heavy duty (SHPD), and
- extreme heavy duty (XHD)

Independent of the viscosity grade and the base oils used, classic HD oils have a high reserve alkalinity and thus a higher content of alkaline-earth salts and organic acids [217]. Also regarding ashless dispersants, the oils are designed for at soot dispersing. Special viscosity improvers are used generally to avoid additional deposits.

Oils for vehicle fleets pose a particular challenge. As opposed to special products, these should simultaneously satisfy as many car and truck demands as possible. Possible piston cleanliness provided by high concentrations of

Table 12. Engine oil classification according to API SAE J 183

Gasoline engines	(Service classes)
API-SA	Regular engine oils possibly containing pour point improvers and/or foam inhibitors.
API-SB	Low-additive engine oils low-power gasoline engines. Include additives to combat aging, corrosion and wear. Issued in 1930.
API-SC	Engine oils for average operating conditions. Contain additives against coking, black sludge, aging, corrosion and wear. Fulfil the specifications issued by US automobile manufacturers for vehicles built between 1964 and 1967.
API-SD	Gasoline engine oils for more difficult operating conditions than API-SC. Fulfil the specifications issued by US automobile manufacturers for vehicles built between 1968 and 1971.
API-SE	Gasoline engine oils for very severe demands and highly-stressed operating conditions (stop and go traffic). Fulfil the specifications issued by US automobile manufacturers for vehicles built between 1971 and 1979. Covers API-SD; corresponds approximately to Ford M2C-9001-AA, GM 6136 M and MIL-L 46 152 A.
API-SF	Gasoline engine oils for very severe demands and highly-stressed operating conditions (stop and go traffic) and some trucks. Fulfil the specifications issued by US automobile manufacturers for vehicles built between 1980 and 1987. Surpasses API-SE with regard to oxidation stability, wear protection and sludge transportation. Corresponds to Ford SSM-2C-9011 A (M2C-153-B), GM 6048-M and MIL-L 46 152 B
API-SG	Engine oils for the severest of conditions. Include special oxidation stability and sludge formation tests. Fulfil the specifications issued by US automobile manufacturers for vehicles built between 1987 and 1993. Specifications similar to MIL-L 46 152 D
API-SH	Specification for engines oils built after 1993. API-SH must be tested according to the CMA's Code of Practice. API-SH largely corresponds to API-SG with additional demands regarding HTHS, evaporation losses (ASTM and Noack tests), filterability, foaming and flash point. Furthermore, API-SH corresponds to ILSAC GF-1 without the Fuel Economy test but with the difference that 15W-X multigrade oils are also permissible.
API-SJ	Supercedes API-SH. Greater demands regarding evaporation losses. Valid since 10/96.
Diesel engines	**(Commercial classes)**
API-CA	Engine oils for low-power gasoline and normally aspirated diesel engines run on low-sulfur fuels. Corresponds to MIL-L 2 104 A Suitable for engines built into the 1950s.
API-CB	Engine oils for low-to-medium power gasoline and normally aspirated diesel engines run on low-sulfur fuels. Corresponds to DEF 2101 D and MIL-L 2104 A Suppl. 1 (S1). Suitable for engines built from 1949 on. Offer protection against high-temperature deposits and bearing corrosion.
API-CC	Gasoline and diesel engine oils for average to difficult operating conditions. Corresponds to MIL-L 2 104 C Offer protection against black sludge, corrosion and high-temperature deposits. For engines built after 1961.
API-CD	Engine oils for heavy-duty, normally aspirated and turbocharged diesel engines. Covers MIL-L 45 199 B (S3) and corresponds to MIL-L 2 104 C Satisfies the requirements of Caterpillar Series 3.
API-CD II	Corresponds to API-CD. Additionally fulfils the requirements of US 2-stroke diesel engines. Increased protection against wear and deposits.
API-CE	Engine oils for heavy-duty and high-speed diesel engines with or without turbocharging subject to fluctuating loads. Greater protection against oil thickening and wear. Improved piston cleanliness. In addition to API-CD, Cummins NTC 400 and Mack EO-K/2 specifications must be fulfilled. For US engines built after 1983.
API-CF	Replaced API-CD for highly turbocharged diesel engines in 1994. High ash. Suitable for sulfur contents $> 0.5\,\%$.
API-CF-2	Only for 2-stroke diesel engines. Replaced API-CD II in 1994.
API-CF-4	Engine oil specification for high-speed, 4-stroke diesel engines since 1990. Meets the requirements of API-CE plus additional demands regarding oil consumption and piston cleanliness. Lower ash content.
API-CG-4	For heavy-duty truck engines. Complies with EPA's emission thresholds introduced in 1994. Replaced API-CF-4 in June 1994.
API-CH-4	Replaces API-CG-4. Suitable for sulfur contents $> 0.5\,\%$.
All engines	(Energy Conserving)
(API-EC I)	(min. 1.5 % less fuel consumption than an SAE 20W-30 reference oil in a 1982, 3.8 liter, Buick V6 gasoline engine. Sequence VI test)
(API-EC II)	Same as API-EC I but with minimum 2.7 % lower fuel consumption
API-EC	Replaces API-EC I and II. Only together with API-SJ. Cuts in fuel consumption: 0W-20, 5W-20 $> 1.4\,\%$, 0W-XX, 5W-XX $> 1.1\,\%$, 10W-XX, others $> 0.5\,\%$, Sequence VI A test: In a 1993, 4.6 liter Ford V8 engine. Reference oil 5W-30

over-based soaps is sacrificed because gasoline engines are prone to self-ignition if high proportions of metal detergents are present. As a result, other components are selected, such as the skillful use of unconventional base oils along with detergents, dispersants, *VI* improvers and antioxidants.

8.1.3. Classification by Specification

8.1.3.1. MIL Specifications

These specifications originate from the US Forces which set the minimum requirements for their engine oils. These are based on certain physical and chemical data along with some standardized engine tests. In the past, these classifications were also used in the civilian sector to define engine oil quality. In the 1980s, these specifications have become almost irrelevant for the West European market.

MIL-L-46152 A to MIL-L-46152 E. These military specifications have now been discarded. Engine oils which meet these specifications are suitable for use in US gasoline and diesel engines. MIL-L-46152 E (discarded in 1991) corresponds to API SG/CC.

MIL-L-2104 C classifies high-additive engine oils for US gasoline and normally aspirated and turbocharged diesel engines.

MIL-L-2104 D covers MIL-L-2104 C and requires an additional engine test in a highly-charged Detroit two-stroke diesel engine. In addition, Caterpillar TO-2 and Allison C-3 specifications are fulfilled.

MIL-L-2104 E similar in content to MIL-L-2 104 C The gasoline engine tests have been updated and include more stringent test procedures (Seq. III E / Seq. V E)

8.1.3.2. API and ILSAC Classification

The API together with the ASTM and the SAE have created a classification in which engine oils are classified according to the demands made on them, bearing in mind the varying conditions in which they are operated and the different engine designs in use (Table 12). The tests are standard engine tests. The API has defined a class for gasoline engines (S = service oils) and for diesel engines (C = commercial). Diesel engines in passenger cars are almost unheard of in the USA. In addition, a number of fuel economy stages has been determined (EC = energy conserving).

8.1.3.3. ACEA Specifications

As a result of persistent internal differences, the CCMC was disbanded and succeeded by the ACEA. CCMC specifications remained valid in the interim period. The new ACEA classifications (Table 13) came into force on January 1, 1996.

Table 13. Engine oil classification according to ACEA – Part I

Gasoline engines	
A1-96	Category for low-viscosity engine oils particularly low HTHS viscosity (< 3.5 mPa · s). This engine oil type was initially favored by Ford and Rover. Recommended viscosity grades are xW-30 and xW-20.
A2-96	Category for conventional and low-viscosity engine oils (no restriction on the permissible viscosity grades) with increased demands on the former CCMC G4. Out-performs API SH.
A3-96	Category for conventional and low-viscosity engine oils (no restriction on the permissible viscosity grades). Out-performs ACEA A2-96 with regard to Noack (evaporation losses), piston cleanliness and oxidation stability. Furthermore, out-performs API SH, CCMC G4 and G5.
Car diesel engines	(Light duty diesel engines)
B1-96	Category for low-viscosity engine oils with particularly low HTHS viscosity (corresponds to A1-96).
B2-96	Category for conventional and low-viscosity engine oils (no restriction on the permissible viscosity grades) with increased demands on the former CCMC PD 2.
B3-96	Category for conventional and low-viscosity engine oils (no restriction on the permissible viscosity grades). Out-performs ACEA B2-96 with regard to cam wear, piston cleanliness and viscosity stability in high soot conditions.
Truck diesel engines	(Heavy-duty diesel engines)
The new ACEA categories for truck diesel engines are largely based on Mercedes – Benz specification sheets.	
E1-96	Corresponds in general to the former CCMC D4.
E2-96	Largely based on Mercedes – Benz Sheet 228.1. Mack T8 test is also required.
E3-96	Largely based on Mercedes – Benz Sheet 228.3. Mack T8 test is also required.

The ACEA specifications were revised in 1996 and replaced by 1998 versions. Tests could be performed to 1996 specifications until March 1, 1999. The 1998 specifications became valid on March 1, 1998. 1996 specifications ceased to be valid on March 1, 2000. When ACEA specifications are quoted, the year must no longer be given.

Additional foaming tests were introduced for all categories and the elastomer tests were modified (Table 14).

Table 14. Engine oil classification according to ACEA – Part II

Gasoline engines	
A1-98	Category for Fuel Economy engine oils with especially low HTHS viscosity ($< 3.5\,\mathrm{mPa \cdot s}$). Recommended viscosity grades are xW-30 and xW-20. Compared to a 15W-40 reference oil, fuel savings of $\geq 2.5\,\%$ must be proven in a Mercedes test engine (M 111).
A2-98	Category for conventional and Fuel Economy engine oils. Replaces A2-96.
A3-98	Category for conventional and Fuel Economy engine oils with greater requirements than A2-98. Surpasses ACEA A2-98 with regard to Noack (evaporation losses), piston cleanliness and oxidation stability. Replaces A3-96.
Car diesel engines	(Light duty diesel engines)
B1-98	Category for Fuel Economy engine oils with especially low HTHS viscosity (same as A1-98). Recommended viscosity grades are xW-30 and xW-20. Compared to a 15W-40 reference oil, fuel savings of $\geq 2.5\,\%$ must be proven in a Mercedes test engine (M 111). Replaces B1-96.
B2-98	Category for conventional and Fuel Economy engine oils. Replaces B2-96.
B3-98	Category for conventional and Fuel Economy engine oils. Surpasses ACEA B2-98 with regard to cam wear, piston cleanliness and viscosity stability in high soot conditions. Replaces B3-96.
B4-98	New category for direct injection diesel engines (TDI).
Truck diesel engines	(Heavy-duty diesel engines)
E1-98	Corresponds to ACEA E1-96.
E2-98	Corresponds to ACEA E2-96.
E3-98	Corresponds to ACEA E3-96.
E4-98	Largely based on Mercedes – Benz Sheet 228.5. No OM 364 A engine test required but Mack T8 and T8E.
(E5xx)	(Planned category which should include all requirements which reduce wear caused by soot and deposits)

8.1.3.4. Manufacturers' Approvals

Apart from the specifications listed in Tables 13, 14, and 15, a number of manufacturers demand tests on their own engines (Table 15).

8.1.3.5. Future Trends

New generations of engines using optimized technologies advance the concept of tailor-made, special oils.

The continuing optimization of the combustion process to increase the efficiency of gasoline engines has led to the development of direct-injection gasoline engines (GDI engines) which may offer fuel savings of about 20 %. On the diesel side, direct injection with unit pumps or common rail technology using pressure of up to 3000 bar have become the norm. These designs which originate in truck engines, offer power increases of up to 50 % at almost constant fuel consumption. In the truck area, reducing exhaust emissions has the highest priority. The present thresholds of Euro 2 and Euro 3 (from 2001) will be easily surpassed with special exhaust recycling and catalytic converter systems. As these generations of engines will require low-sulfur diesel fuels (10 ppm), new demands will be made on the lubricants used.

Furthermore, the surface treatment of pistons and cylinders has improved to such an extent that topology-specific oil consumption is steadily falling. In addition, oil change intervals have been constantly increased. All in all, three factors will characterize engine oils of the future – fuel efficiency, long oil drain intervals, and low emissions.

Fuel Efficiency. As a result of strict limits to fuel consumption in the USA (CAFE = Californian Act for Fuel Emissions) and the proven fuel economy effect of low-viscosity engine oils [218], this topic is attracting attention in Europe and Asia. As a rule, engine-based savings can reach a theoretical $8 - 10\,\%$ [219]. As engine oils cannot totally eliminate frictional losses, saving potentials of 4 and 5 % present enormous challenges. However, values of between 3 and 4 % are already possible today.

Table 15. Manufacturers' approvals

VW-Norm	Application area
VW 50000	Low-viscosity oils for gasoline and normally-aspirated diesel engines. Only SAE 0W-XX, 5W-XX and SAE 10W-XX oils. XX $>$ 40 oils were not included after 10/91.
VW 50101	Multigrade engine oils without low-viscosity characteristics for gasoline and normally-aspirated diesel engines.
VW 50200	New specification for car gasoline engines with extended oil change intervals.
VW 50300	New specification for car gasoline engines with extended service intervals. Surpasses the requirements of 50200 (HTHS $\geq$ 2.9 mPa · s).
VW 50301	Planned specification for turbocharged car gasoline engines with extended service intervals (HTHS $>$ 3.5 mPa · s).
VW 50500	All-season engine oils for diesel engines with or without turbocharging.
VW 50501	All-season engine oils especially for Unit Pump diesel engines.
VW 50600	New specification for diesel engines with extended service intervals (HTHS $\geq$ 2.9 mPa · s).
VW 50601	Planned specification for Unit Pump diesel engines with extended service intervals.
MB Sheet	**Application area**
MB 226.0	Monograde engine oils for normally-aspirated diesel engines.
MB 227.0	Monograde engine oils for turbocharged diesel engines.
MB 227.1	Multigrade engine oils for turbocharged diesel engines.
MB 228.1	Multigrade engine oils for turbocharged diesel engines. Oil change intervals up to 30 000 km.
MB 228.3	SHPD engine oils for highly turbocharged diesel engines. Extended oil change intervals up to 45 000 km.
MB 228.5	UHPD (ultra high performance diesel) engine oils for highly turbocharged diesel engines. Extended oil change intervals in light-duty engines up to 45 000 km. Up to 160 000 km possible in heavy-duty diesel engines (with service interval displays).
MB 229.1	Engine oils for cars (gasoline and diesel engines). Introduced 05/97. More stringent than ACEA A2-96/A3-96 and B2-96/B3-96.
MB 229.3	Engine oils for cars with extended oil change intervals (30 000 km).
MAN in-house specification	**Application area**
MAN 270	Monograde engine oil turbocharged and non-turbocharged diesel engines.
MAN 271	Multigrade engine oil turbocharged and non-turbocharged diesel engines.
MAN QC13-017/M 3275	SHPD engine oils for all diesel engines with oil change intervals up to 45 000 km.
MAN M3277	UHPD engine oils for all diesel engines with oil change intervals up to 80 000 km.
MAN 3271	Oil for engines run on gas (natural gas, propane or butane).
Scania	**Application area**
ACEA E3	Oil change intervals up to 60 000 km.
LDF approval	ACEA E3 oil with special long drain field-test approval. Oil change intervals up to 120 000 km.
VDS	Volvo drain specification for extended oil change intervals (50 000 km).
VDS-2	Specified for Euro 2 engines (60 000 km).
(ACEAE4-98)	DAF recommends oils with the following viscosities; 10W-40, 10W-30 and 5W-40. Oil change intervals up to 100 000 km.
Oil Type 1	Normal quality (ACEA E1-96, E2-96)
Oil Type 1∗	Type 1 plus corrosion protection
Oil Type 2	Higher quality (SHPD, ACEA E3-96)

Reduction in car fuel consumption in urban conditions are achieved by lowering frictional losses during cold start-ups which simultaneously result in less wear and a lowering of viscosity in constant throttle conditions. There is a plateau-like optimum for the correlative fuel savings in engines with HTHS values between 2.5 and 2.9 mPa · s.

According to OEMs, HTHS values must not be lower than minimum 2.6 mPa · s in all manufacturer's approvals and new engine oil developments because of possible wear between critical material pairings.

Long Drain Intervals. Concerns about higher wear and thus shorter life caused by low-

viscosity oils run contrary to the trends of the new generation of engines which are designed to cope with even longer oil change intervals. As stated in manufacturer's specifications, oil change intervals for cars are currently 15 000 km for gasoline engines and 30 000 km for diesel engines. At present there are demands to double these figures without detracting from the fuel efficiency and wear performance of the oil. As a result of these diverging requirements, future engine oil specifications might contain a whole series of new OEM-specific engine tests.

The radionuclide technique, as a proven on-line tool, is experiencing a renaissance for examining the effects on wear in various operating conditions such as during running-in or to determine long-term stability.

Apart from long-term wear, very high demands are made on high oxidation stability and low evaporation losses. This strengthens the trend towards synthetic and unconventional oil as the basis for such high performance engine oils. A graphic correlation between evaporation tendency (Noack) and oil consumption is given in Figure 40. A technically realized milestone for fully synthetic engine oils based on present synthetic base oils is an evaporation loss of 5 to 6 %.

The suitability of extended oil change interval oils is tested in thermally stressed test engines which run hot and without oil top-up. The typical indicators of aging like viscosity and TBN are measured.

Low Emission. Heavy-duty engine oils already achieve drain intervals of 100 000 km. Because of the ever increasing number of trucks on roads, this is a useful contribution to improving environmental compatibility.

Apart from the CO, HC and SO_2 emissions which are seen to be caused by the fuel, particulate emissions play a significant role in HD engines. These particulates are a result of incomplete combustion and are a mixture of fuel- and lubricant-based components. As the oil-based particles are largely caused by highly volatile elements in the formulation, evaporation losses have a direct effect on reducing pollutants (see Fig. 40). Furthermore, it is assumed that sulfur compounds in diesel fuels will poison the catalytic converters of future Euro 3 and Euro 4 engines. The call for low sulfur (10 ppm) and most probably low phosphorus engine oils will not only lead to a different assessment of additives but also to the rejection of solvent neutral oils which, as a rule, contain between 0.2 and 1.0 % sulfur. Similar to passenger cars unconventional and synthetic base oils are preferred so that the trend towards low viscosity, fuel economy oils will spread definitively to the heavy-duty sector.

8.2. Two-Stroke Oils

8.2.1. Application and Characteristics [220–222]

Two-stroke engines are mostly used when high specific power, low weight and low price are key parameters. These engines are thus often used in motorcycles, boats (outboard engines), jet-skis, lawn mowers, chain saws and small vans, whereby the vast majority are found in motorcycles and boats.

Almost all two-stroke engines use total-loss lubrication, i.e., the oil is not circulated as in the case of four-stroke engines but added to the fuel. A large part is burnt in the combustion process, about one quarter is exhausted as unburnt oil mist. Simple engines as found in older mopeds still use the premix method, whrere a suitable two-stroke oil to is added manually the fuel tank at a ratio of about 1 : 20 to 1 : 100. More advanced designs use an automatic oil metering system. These either add a constant amount of oil to the fuel or add oil according to engine loading. Typical oil : fuel ratios are between 1 : 50 and 1 : 400.

In the majority of simple two-stroke, the engine breathes through classic carburetors. Contrary to four-stroke engines, the fresh fuel/air mixture in a two-stroke engine scavenges the cylinder after combustion. This simultaneous charging and emptying causes about 30 % of the fresh mixture to be exhausted without burning.

This disadvantage, along with the only partial burning of the oil, causes many two-stroke engines to generate comparatively high emissions. In highly populated areas with a large number of small motorcycles, such as in many Asian cities, this leads to severe odor, smoke and noise pollution.

In recent years, these typical disadvantages have been countered by some advances in two-

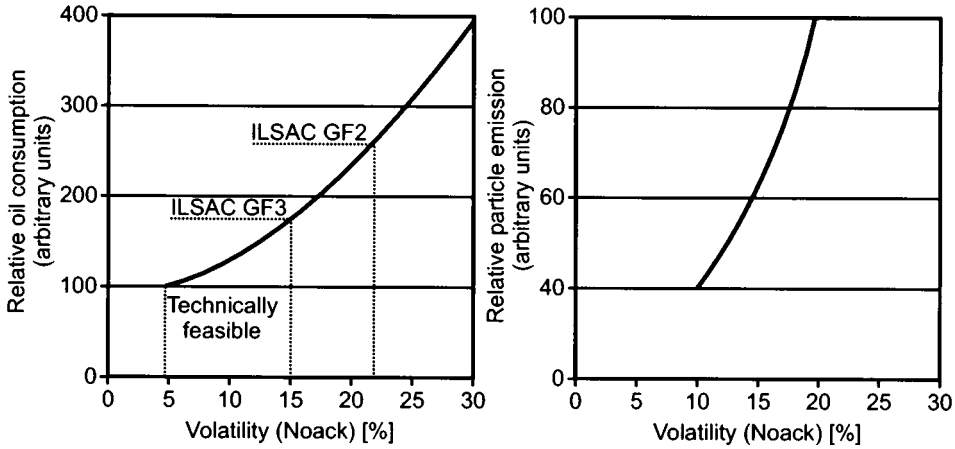

Figure 40. Oil consumption and relative oil generated particulate emissions versus evaporation loss

stroke technology. The development of direct or indirect fuel injection has led to significantly reduced emissions and improved fuel efficiency.

Today's engines require correspondingly high quality oils for reliable operation and long life. The principal criteria for the quality of two-stroke oils are:

- lubricity and antiwear properties
- cleaning function (detergent/dispersant properties)
- avoidance of deposits in the exhaust system
- low smoke
- spark plug cleanliness and the avoidance of preignition
- good fuel miscibility even at low temperatures
- corrosion protection
- good flowing properties

About 85 – 98 % of two-stroke oils are base oils with the rest consisting of various additives. In principle, all common base oils can be used, ranging from brightstocks, solvent neutral types to fully synthetic polyalphaolefins. Higher quality two-stroke lubricants often contain various synthetic esters and this is particularly the case for biodegradable oils which were specially developed for marine outboards.

The additives in two-stroke oils are usually matched to the requirements of the engine. As in the case of four-stroke engine oils, antiwear additives are included in two-stroke oils. Along with the most commonly used zinc dialkyldithiophosphates, nonash-forming types such as dithiophosphoric acid esters of alkyl and aryl esters or phosphoric acids are used.

The following DD additives are added to the oil: Alkali or alkaline-earth salts of sulfonates and/or phenolic compounds, and polybutene succinimides.

Furthermore, two-stroke oils contain small quantities of antioxidants, corrosion inhibitors, defoamers and flow improvers in addition to antiwear and DD additives.

Low-smoke two-stroke oils contain a significant amount of fully synthetic fluids, i.e., polybutenes (about 10 to 50 %). Apart from good lubricity compared to mineral oils, these fluids offer much cleaner combustion and significantly less coking [223], [224].

8.2.2. Classification

The basis for all classification systems are a series of laboratory and functional tests, the latter being bench tests performed on the latest generation of two-stroke engines.

8.2.2.1. API Service Groups

The API currently lists three categories (Table 16) which cover all engines from low-power

Table 16. API groups

API	Application	Test engine	Test criteria
TA	mopeds, lawn mowers, electricity generators, pumps	Yamaha CE 50 S (50 cm^3)	piston seizing, exhaust system deposits
TB	scooters, small motorcycles	Vespa 125 TS (125 cm^3)	preignition, power loss due to combustion chamber deposits
TC	high-performance motorcycles, chain saws	Yamaha Y 350 M2 (350 cm^3) Yamaha CE 50 S	preignition, power loss due to combustion chamber deposits piston seizing, ring sticking

lawn mowers to high-performance motorcycles. Engine tests are no longer performed because the specified test engines are no longer manufactured. In future, it is planned to replace the API groups with Japanese JASO and global ISO classifications. There are still a number of oils on the market with API classifications because this system was widely accepted in the past.

8.2.2.2. JASO Classification

Japanese Automotive Standards Organization (JASO), to which all major Japanese vehicle manufacturers belong, classifies two-stroke oils into three groups, FA, FB and FC (Table 17).

Table 17. JASO performance categories (Reference oil: JATRE 1 = 100)

Test criteria	JASO FA	JASO FB	JASO FC
Lubricity	> 90	> 95	> 95
Detergent effect	> 80	> 85	> 95
Exhaust smoke	> 40	> 45	> 85
Exhaust deposits	> 30	> 45	> 90

Table 18. Engine test criteria for JASO classifications

Test engine	Test criteria	Test parameters
Honda Dio AF 27	lubricity	piston ring wear, ring scuffing, piston seizing
Honda Dio AF 27	detergent effect	piston ring sticking as a result of lacquering, coking, deposits on piston and in combustion chamber
Suzuki SX 800 R	exhaust smoke	smoke particles
Suzuki SX 800 R	exhaust deposits	back pressure in exhaust system

All three categories use the same test engines and the corresponding performance category is allocated according to predetermined thresholds. The test results are determined in comparison to an exactly defined, high-performance reference oil (JATRE 1) and published as an Index relative to JATRE 1 (Table 18). The key test criteria are the lubricity and detergent effect of the oil as well as its tendency to create smoke and deposits in the exhaust system. The first specification for a low-smoke oil was created with the laying-down of JASO FC.

8.2.2.3. ISO Classification

In the mid-1990s, after JATRE 1 oils were tested in European engine tests, it became clear that JASO FC oils could no longer satisfy the latest demands of European two-stroke engines. An series of extended tests which satisfied all demands were thus developed in Europe. In addition to the testing of smoke, exhaust system deposits, lubricity and detergent effect according to JASO, a new category with a 3-h Honda Dio test to quantify improved piston cleanliness and detergent effect was added. The reference oil for all tests was JATRE 1. These new guidelines were created by CEC working parties in which European engine and lubricant manufacturers are represented.

The ISO now classifies two-stroke oils into three categories, ISO-L-EGB, -EGC and -EGD. A fourth category (-EGE) is currently being drafted with strong European representation.

The categories ISO-L-EGB and -EGC mirror the requirements of the JASO categories FB and FC while requiring additional proof of piston cleanliness. ISO-L-EGC and -EGD require proof of low smoke similarly to JASO FC. Table 19 shows all engine-based evaluation criteria.

Table 19. Summary of the ISO categories (Reference oil: JATRE 1 = 100)

Test criteria	ISO-L-EGB (incl. JASO FB)	ISO-L-EGC (incl. JASO FC)	ISO-L-EGD
Lubricity	> 95	> 95	> 95
Smoke	> 45	> 85	> 85
Exhaust deposits	> 45	> 90	> 90
Detergent effect	> 85 (1-h test)	> 95 (1-h test)	> 125 (3-h test)∗
Piston cleanliness	> 85 (1-h test)∗	> 90 (1-h test)∗	> 95 (3-h test)∗

∗ New requirements in addition to JASO FC.

8.2.3. Oils for Two-Stroke Outboard Engines

Neither API nor JASO or ISO classifications contain quality guidelines for outboard engine oils. These are usually oils whose formulation and characteristics have been matched to the engine technologies which have become established for powering boats. The main difference to other two-stroke oils lies in their additive chemistry. The additives in these oils are all nonash-forming because these engines all display a marked coking tendency in certain parts of the combustion chamber, such as the ring grooves. The performance categories of two-stroke oils for outboard engines were primarily developed by the American National Marine Manufacturers Association (NMMA). Back in 1975, the minimum requirements of such oils were incorporated into the TCW specification. In 1988, a far-reaching revision was issued with the title TCW 2. During the following years, problems were encountered with technologically advanced engines which were run on TCW or TCW 2 oils. This initiated a further tightening-up of minimum requirements, which was released as TCW 3. In 1997, the standards for oil quality were again increased to keep pace with continuing developments in engine technology. The new specification, TCW 3-R (R = recertified), now includes laboratory tests and tests on five different engines, three of which are outboards. Table 20 shows the tests which have to be performed on a newly developed outboard engine oil to achieve TCW 3-R classification. The costs generated by testing a TCW 3-R product development have never been so high. The engine tests alone generate costs of about $ 150 000 – 200 000.

Table 20. Test criteria for NMMA TCW 3-R

Engine Tests	
Test engine:	Test criteria:
Yamaha CE 50 S	lubricity (seizures)
Yamaha CE 50 S	power loss due to preignition
Mercury 15 HP (2 runs)	compression losses, ring jamming, piston cleanliness
OMC 40 HP	ring jamming, piston cleanliness
OMC 70 HP	ring jamming, piston cleanliness
Laboratory Tests	
Low-temperature viscosity	limited viscosity at −25 °C
Miscibility	mixing with fuel at −25 °C
Corrosion protection	standard rust tests compared to reference oil
Compatibility	stability after mixing with reference oil
Filterability	flow rate compared to pure fuel

8.2.4. Environmentally Friendly Two-Stroke Oils [225]

The ever increasing severity of environmental legislation is also effecting the development of two-stroke oils, especially outboard engine oils. Such ecologically optimized oils often have regionally differing classifications which reflect local environmental legislation and their biodegradability depends on varying minimum requirements. At the international level, the International Council of Marine Industry Associations (ICOMIA) has specified harmonized requirements [223]. In 1997, the ICOMIA Standard 27-97 was passed for environmentally friendly outboard engine oils. From a technical point of view, products thus labeled must fulfil at least TCW 3-R as well as offering very low algae, daphnia and fish toxicity and rapid biodegradability as defined by ISO and OECD standards. These oils are based on fully synthetic components with the base oils usually be-

ing rapidly biodegradable synthetic esters. By using correspondingly high quality esters, these products are the very best two-stroke oils and can even used for lubricating chain saws. The use of ester-based lubricants combines the highest technical performance with improved environmental compatibility.

8.3. Tractor Oils

Relatively newer generations of construction and agricultural machinery make differing demands on functional fluids. For reasons of simplified servicing but also because of the general wish to rationalize stocks, universal oils were developed which satisfy the various functional requirements of such machines.

These oils should guarantee long machinery life in all climatic conditions as well as extending service intervals and reducing down-time.

These days, two different oil technologies are used which are characterized by their application area. They are universal tractor transmission oils (UTTO) and super tractor oils universal (STOU).

The demands made on tractor oils have increased sharply in line with advances in vehicle technology and ease of operation.

Technical advances in tractors required a significant improvement in operating fluids as they did in the whole automotive area. Apart from big improvements in additive technology to cope with greater mechanical demands, the aging stability of the fluids has increased reflecting dramatically higher specific power outputs and lower oil volumes.

All-season use is now a standard requirement for tractor oils as it is in the automotive area. The result of this is that the viscosity grades and thus temperature ranges have been extended from SAE 15W-30 and 10W-30 to 15W-40, 10W-40 and 5W-40.

The performance of such universal oils as hydraulic fluids corresponds to, at least, HLP and HVLP levels because of the additives included to guarantee universal use.

The use of these products in vehicle gearboxes and wet brakes makes much greater demands on the fluid. Highly-stressed mechanical drive units make great demands on the wear protection and life of the lubricant. A special challenge to oil formulations is the use of these products in wet brakes which began back in the 1960s. Not least because of safety considerations, such oils must offer high thermal stability and balanced and stable friction characteristics in brakes. Fine-tuning these oils with special friction modifiers is one of the most difficult aspects of developing such oils.

While UTTO oils can satisfy the above-listed applications, their use as engine oils require vastly different additive packages. Apart from additives for friction and wear control, low-temperature stabilizers, oxidation, corrosion and foam inhibitors, engine oils additionally require significant quantities of detergents and dispersants. As a rule, normally aspirated engines require, at least, an API CE oil and if a turbocharger is fitted, the oil should be API CF or CF-4. In many cases, tractor engine manufacturers issue approvals for oils after they have successfully completed tests in the corresponding engines.

The latest generation of tractor oils not only include products with significantly higher performance than in the past but also oils which offer better environmental compatibility. Such products often contain rapeseed or sunflower base oils or synthetic, rapidly biodegradable esters.

Tractor manufacturers now issue their own, in-house fluid specifications which satisfy all specific requirements of the corresponding machinery. Table 21 lists several manufacturer specifications.

Table 21. Manufacturer specifications for tractor oils

Manufacturer	Oil type	Specification
Ford	STOU	ESN-M2C-159-C
Ford	UTTO	ESN-M2C-86-C
John Deere	STOU	JDM J 27
John Deere	UTTO	J 20 C, J 20 D
Massey Ferguson	STOU	CMS M1139, CMS M1144
Massey Ferguson	UTTO	CMS M1135, CMS M1143

8.4. Gas Engine Oils [226–228]

Gas has long been used to power vehicles and stationary engines. Such engines require a variety of lubricating oils, depending on the type of engine and the operating conditions.

There is currently no universal, harmonized specification for gas engine oils. The large variations in operating conditions between mobile and stationary engines generally require oils with different additive packages. A general differentiation is made between high-, medium-, and low-ash types which are recommended by the manufacturers in line with the designed use of the engine. As a rule, gas engine oils are subject to oxidation and nitration conditions which can accelerate the aging of the oil. Gas-powered cars normally use the same conventional engine oils as are used in gasoline-powered engines. Diesel engines usually need oils with a higher total base number and with more dispersants than gasoline engine oils. Typical qualities for car engines are API SH and/or ACEA A3 (gasoline engines) as well as API CG-4 and/or ACEA B3 (diesel engines). Similarly to the automotive sector, multigrade oils are used to cope with variable operating conditions and to guarantee reliable lubrication at low ambient temperatures. As the number of compressed natural gas (CNG) powered cars increases, there is increasing pressure to develop oils which are especially formulated for these applications. This could mark the beginning of a future, uniform specification.

Special multigrade oils have already been developed for use in heavy diesel engines, with CNG-powered buses being the major application. These have been tested and approved by various engine manufacturers. Examples of these approvals are Mercedes – Benz Sheet 226.9 or MAN M 3271. These oils were tested in bench tests as well as in realistic field trials.

Stationary gas engines can make significantly more complex demands on the oil and this has an effect on their development. While the common ACEA or API bench tests suffice for CNG-powered car engines, laboratory tests on gas engine oils are limited to initial oil screening. The real development takes place in field trials in which engines often have to run for years before they are evaluated. While particular attention is paid to sludge formation, valve train wear and low temperature flowing in car engines, other effects are important in stationary engines. Particularly important is controlling oil aging caused by oxidation and nitration but also preignition which is caused by high levels of ash-forming additives. This problem is most prevalent in two-stroke engines which generally need low-ash oils. Simply because of their long service lives, gas engine oils should protect against valve seat recession and spark plug fouling. As regards running on corrosive gases (caused by sulfur or halogen contamination), special attention must be paid to adequate corrosion protection. The development and application of gas engine oils normally takes place in close cooperation between the oil and engine manufacturers.

8.5. Marine Diesel Engine Oils [229]

These lubricants are heavily influenced by the type of fuel used and the design of the engines themselves. A number of similar engines are also used in stationary applications to generate electricity with conventional fuels or with steam power.

8.5.1. Low-Speed Crosshead Engines

Low speed diesel engines generating up to 1000 kW per cylinder at 50 to 120 rpm use the crosshead principle (large engines with over 900 mm bores and 3000 mm strokes and 20 000 kW outputs). In crosshead designs, the cylinder block and the crankcase are separate units. Sealing is provided by stuffing boxes and the liners are lubricated with cylinder oils by means of dosing devices. Depending on the bore and stroke of the cylinder, up to 16 dosing devices may be fitted.

8.5.2. Medium-Speed Engines

Medium-speed engines run at around 200 to 100 rpm. The overall design of these engines is roughly similar to that of vehicle internal combustion engines, they do not use the crosshead design and the crankshaft is connected to the pistons and thus the cylinders with connecting rods. In such designs, the same oil is used to lubricate the cylinder walls and the crankshaft bearings and forms part of a circulation system. These engines are sometimes called trunk piston engines.

These days, marine diesel engines are fuelled by the worst and heaviest crude oil fuel cuts.

These include vacuum residues, propane deasphalting residues, heavy solvent extracts from lube refining and other byproducts (→ Heating Oils, Chap. 2.).

8.5.3. Lubricants

Cylinder oils in crosshead engines are total-loss products which lubricate the sliding motion of the piston rings in the cylinder liner. To avoid the deposit of combustion residues, these oils must have good dispersant properties. In addition, they must be capable of neutralizing the corrosive acids which result from the high sulfur levels in the fuels. To satisfy these requirements, cylinder oils have a large proportion of over-based components (up to 30 % over-based calcium sulfonates or other over-based components). The alkalinity required for neutralization leads to TBNs of up to 100. When additives for such oils are being selected, special attention must be given to the good colloidal solubility of the additives in the base oil to avoid precipitation. The large surface areas of cylinders in crosshead engines requires the applied oil to disperse rapidly and reliably and this is achieved by good spreadability.

The stuffing boxes fitted to crosshead engines ensure that the *crankcase oils* are hardly contaminated by combustion chamber residues and these oils thus contain relatively few additives. Additives to combat thermal oxidation are essential and the neutralizing capacity can be relatively low with a TBN of 5. The contamination of crankcase oil with cylinder oil leads to increased emulsifying capacity which is undesirable. Water separation is therefore another vital characteristic of crankcase oils.

An increase in dispersant and detergent properties caused by the ingress of cylinder oil can also lead to wear problems as the effect of typical antiwear and EP additives such as zinc dialkyldithiophosphate is thereby diminished.

The type of oil used to lubricate *trunk piston engines* (combined crankcase and piston lubrication) is also largely determined by the sulfur content of the fuel. Oils with TBNs of 12 to 40 are common and the neutralizing capacity of the oils is very soon exhausted in some cases so that the TBN of used oils can be much lower.

During the development of such oils, tests are performed on one- or three-cylinder Bolnes engines but final results are only available from very time-consuming on-board trials lasting at least one year.

Significant restrictions on the sulfur content of marine diesel fuels will be imposed in the coming years. This will result in a lowering of the high proportion of over-based components necessary for such oils at present.

9. Gear Lubrication Oils

9.1. Introduction

Gear lubrication oils are of particular significance for the transmission. Apart from the important function of lubricating the sliding rolling contacts, the oils also fulfil the task of cooling and removing the friction heat generated in the sliding rolling contacts.

In general, the lubricant is a transmission component which can be relatively simply produced at low cost. However, in order to ensure the reliable and long-term operation of a transmission, the selection of the lubricant, in comparison to the selection of the other machine components, is of decisive significance during the transmission construction.

The following base oils are not only used for the lubrication of transmissions but also many other applications:

- mineral oils
- synthetic hydrocarbons (polyalphaolefins)
- poly(alkylene glycols) (homopolymers)
- esters (eco-friendly oils, mainly on synthetic basis)
- rapeseed oil, castor oil

Apart from the various base oils, the type and quantity of the additive has a significant influence on the function and service life of a transmission.

In principle, any type of lubricant, including engine oils, can be used in any transmission, thus ensuring for the moment the transmission's functionality.

However, the selection of a lubricant which is not adapted to the respective construction components and the operating and environmental conditions of a transmission can, given an

early failure of the transmission, lead to maximum consecutive damages up to the breakdown of complete systems. The resulting repair and standstill times of the system lead to unforeseen costs.

9.2. Requirements of Gear Lubrication Oils

In many areas of machine construction, the transmission plays an important role. Figs. 41 and 42 give a general survey over the various transmission types used today.

Each of these transmission types makes specific requirements on the lubricant which must be met to ensure the reliable function of the machines and plants. Thus, the heavily loaded lubricants in hypoid gears require high oxidation stability, together with a very good scuffing and scoring and wear load capacity, due to the high load of the tooth contacts. At the same time, the formation of a load-capable and separating lubricating film for sufficient lubrication and cooling of the sliding rolling contacts in hypoid gears requires a lubricant with an adequately high viscosity at operating temperature.

On the other hand, lubricants used in hydrodynamic gears, such as torque converters, hydrodynamic clutch couplings or retarders, do not need to have a good scuffing and scoring load capacity. However, they must have a high oxidation stability. Due to the viscosity-dependent losses, lubricants which are used in hydrodynamic gears therefore have a clearly lower viscosity at operating temperature in comparison to lubricants used in hypoid gears.

The abovementioned transmission types are used in machine and plant construction with varying degrees of frequency. Figure 43 presents the service lives currently required of transmissions according to the most important industries in the area of machine construction.

The heat development in a transmission has an influence on the temperature of the oil sump and the temperature in the oil tank. Heat accelerates the oil's aging process and, therefore, can cause a reduction of the oil's service life. The heat development and, thus, the oil temperature, is determined by the transmission type, the transferred power, the specific load, as well as by the switching periods – permanent or intermittent operation – and by the environmental conditions – use of the transmission in warm or cold climates, or in a mobile or stationary system.

In transmissions for which long service lives are required, such as paper machines or presses, the lubricant has to be changed according to the mechanical and thermal claim, in compliance with the oil producers' recommendations. Assuming that the average oil sump temperature in such transmissions is approx. 90 °C, the oil should be changed in intervals of 2500 operating hours. An increase of the oil sump temperature of 10 K leads to a 50 % reduction of the service life, reducing the temperature by the same value doubles the oil's service life.

Today, transmissions for which short or medium service lives are required, as for example in passenger cars or mobile hydraulic systems, are mainly filled with life-time oil fillings. The oil sump temperatures in these transmissions often rise up to 130 °C. Therefore, the use of synthetic lubricants is recommended in these applications. In general, the service life of synthetic lubricants is three times longer than the service life of mineral oils. However, the current prices for synthetic lubricants are also three times higher than those for mineral oils.

9.3. Tribology of Gears

There are significant differences between the tribology of gear drives [230] and the tribology of journal and roller bearings, since the lubrication conditions characterizing the sliding rolling contacts in toothed wheels differ from those in journal or roller bearings. As shown in Figure 44, the lubrication conditions in the tooth contacts are the most difficult.

9.3.1. Friction Conditions of Gear Types

Figure 45 gives a schematic overview of the gear types currently used [231], [232].

According to its specifically different geometric ratios and the operating conditions, each gear type is limited to a maximum transferable load.

The main key factors are the sliding rolling contact pressure and the specific load in the tooth contact and the gear's rotational speed, as well as

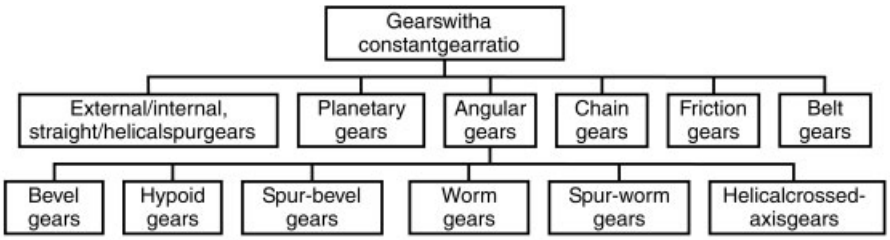

Figure 41. Gears with a constant gear ratio

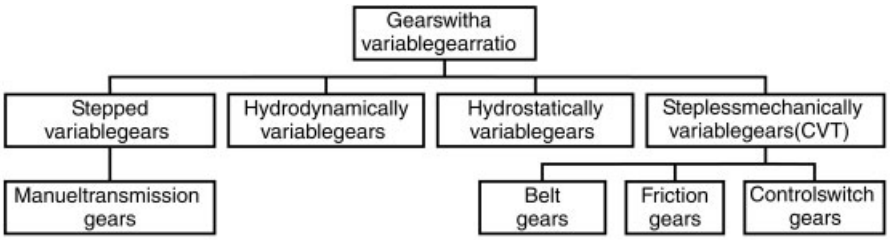

Figure 42. Gears with a variable gear ratio

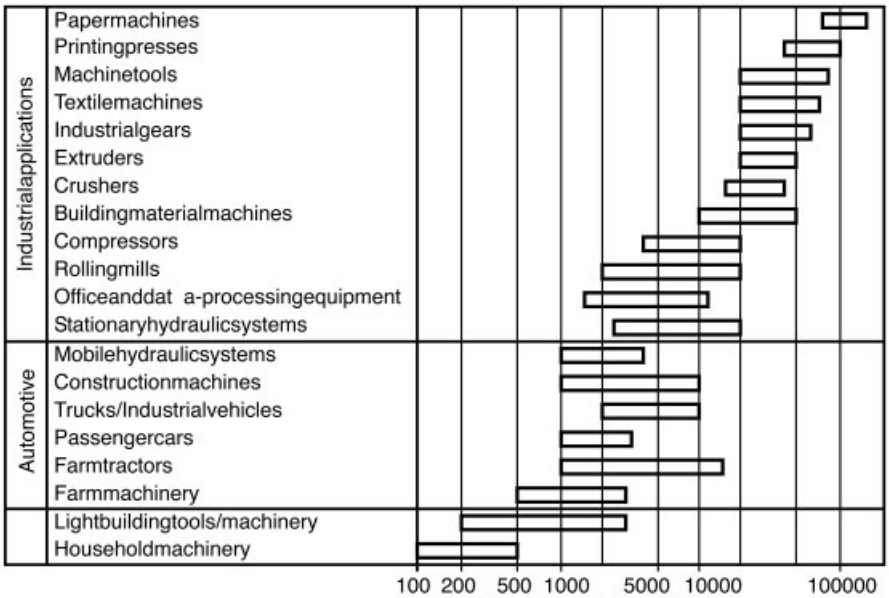

Figure 43. Guide values for the service lives required of gears

the currently effective rolling and sliding speed proportions in the tooth contact. The distribution of these factors during a tooth engagement is shown in Figure 46 on a nonprofile corrected spur gear.

The load and surface velocity of the tooth flanks are characterized by continuously changing conditions alongside the entire tooth engagement distance. The sum of the velocity of the tooth flank equals the effective hydrodynamic velocity, v_Σ. The difference of the velocity of the tooth flank equals the sliding velocity v_g. Together with the locally effective pressure or so-called sliding rolling contact pressure, this sliding speed causes a friction load flowing tangentially to the tooth flank and, thus, an increase in temperature in the sliding rolling contact. Only in the pitch point contact C, as shown in Fig-

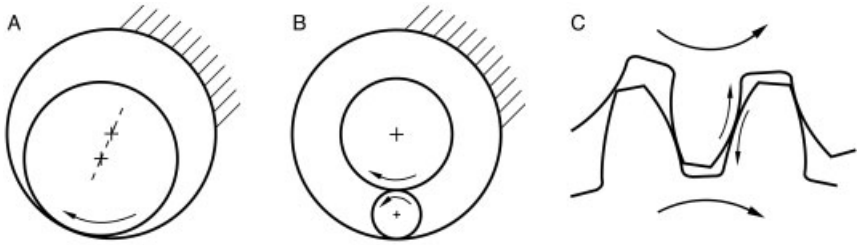

Figure 44. Schematic overview of the narrowing of the gap. (A) hydrodynamic journal bearings, constant sliding speed at good conformity and good lubrication conditions. (B) Cylindrical roller bearing, relatively high rolling speed at low sliding speed, poor conformity and medium lubrication conditions. (C) Tooth contact lubrication, poor conformity along the tooth engagement length at different sliding velocities in tooth-tip direction and under lubrication conditions inferior to A and B.

ure 46, a pure rolling movement can be observed without sliding speed in the whole tooth engagement.

For the toothing shown in Figure 46, a static load distribution alongside the tooth engagement distance can be assumed.

The tooth engagement is a vibrating system which, according to the geometry and the operational area, shows a dynamic load distribution which differs and sometimes clearly exceeds the static load distribution. In this vibrating system, the lubricant and, especially its operational viscosity have an absorbing effect.

Lubrication Film Generation in Tooth Contact. A significant key factor for the assessment of the lubrication generation and condition is the minimum lubricant film thickness. A sufficient minimum lubricant film thickness is necessary to avoid the metal surface contact of the tooth flanks and to lead the generated heat out of the contact zone by means of friction and absorption. With the minimum lubricant film thickness determined according to [234], [235], the current lubrication condition in the tooth contact can be assessed. The non-stationary sliding rolling contact conditions are the reasons for the differences in the lubrication conditions in each contact point of the tooth engagement distance.

In the pitch point C of a tooth engagement where – as in a roller bearing – a pure roll-off or rolling movement without slide speed proportions is generated, relatively optimal lubrication conditions can be assessed. The lubrication conditions during the tooth engagement and disengagement, however, are more unfavorable due to relatively high slide speed proportions.

Another factor with negative impact on the generation of a lubrication film in a tooth contact becomes effective during a new tooth engagement of each tooth when stripping off of the lubrication film by the tooth tip of the engaging toothed wheel. The lubrication film must be regenerated on this tooth tip with each new tooth engagement and breaks with each tooth disengagement.

When a gear is characterized by a more lubrication-unfavorable toothing geometry due to the relatively large slide – speed proportions along the tooth engagement length in this type of gear, for example in a hypoid or cylindrical worm gear, these gears are subject to significantly larger friction losses and higher sliding rolling contact temperatures than a simple spur gear.

Lubrication Conditions. To avoid tooth damage, a detailed observation and assessment of the lubrication conditions in the tooth contacts is particularly important.

Depending on the local load, pressure, temperature and based on the material values of the toothed wheels and the lubrication values of the lubricant, a lubricant film thickness flow over the tooth engagement distance of a toothing can be calculated [236]. Today, this method is generally used to calculate the pitch point C. Due to the fast progress made, a more complex observation of the tooth engagement is only a matter of time and is already partly carried out as a standard method by some companies.

The specific film thickness of the lubricant film is assessed according to the Eq. (10.1):

$$\lambda = h_{\min}/R_{\mathrm{a}} = 2h_{\min}/\left(R_{\mathrm{a1}} + R_{\mathrm{a2}}\right) \tag{10.1}$$

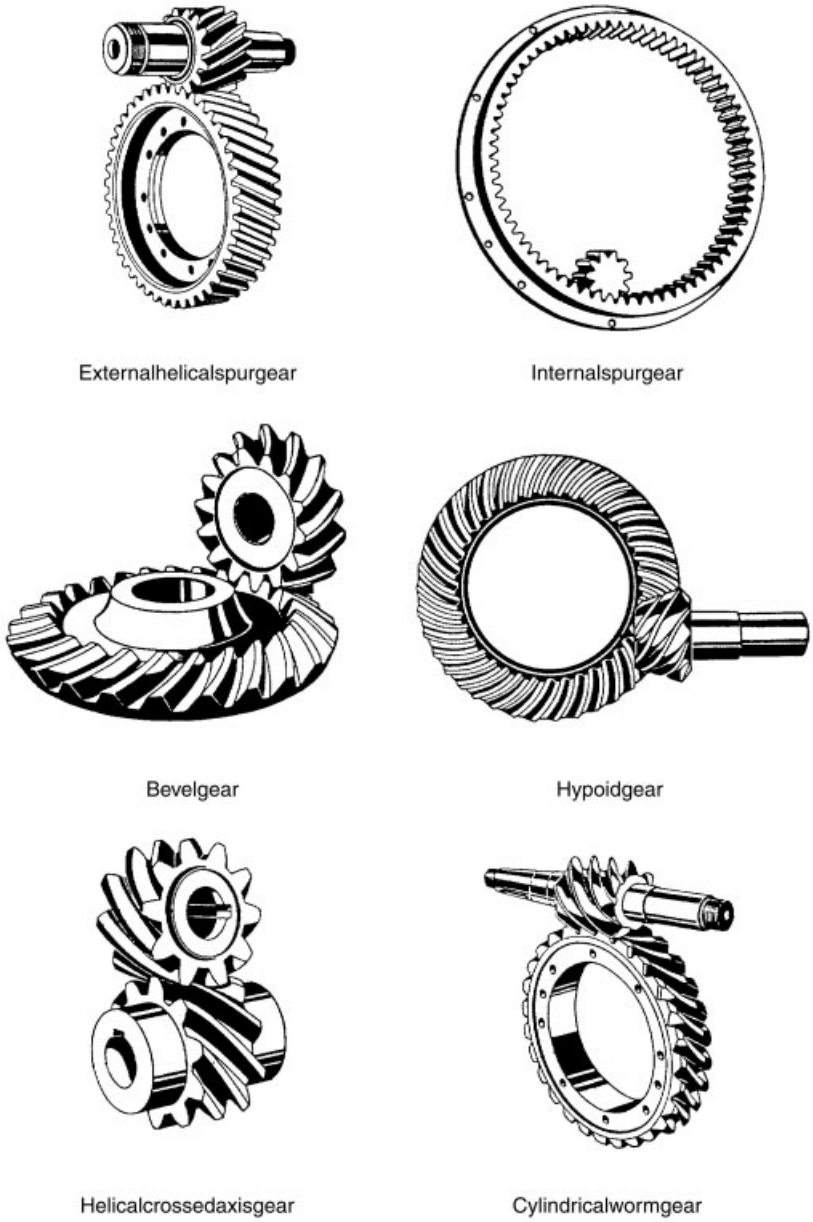

Figure 45. Important gear types

where λ is the specific film thickness according to [237], $h_{\min}$ is the minimum film thickness according to [233], and R_{a} is the arithmetic average mean roughness of the tooth flanks of pinion (R_{a1}) and wheel (R_{a2}) according to [238].

According to the evaluation of numerous observations, a differentiation between two ranges is made.

Specific Film Thickness $\lambda > 2$. In general, the tooth flank surfaces are separated by a sufficiently thick hydro- or elastohydrodynamic lubricant film. The viscosity of the applied lubricant is the decisive property of the lubricant. Only little wear occurs in the long term. Surface damage is not to be expected.

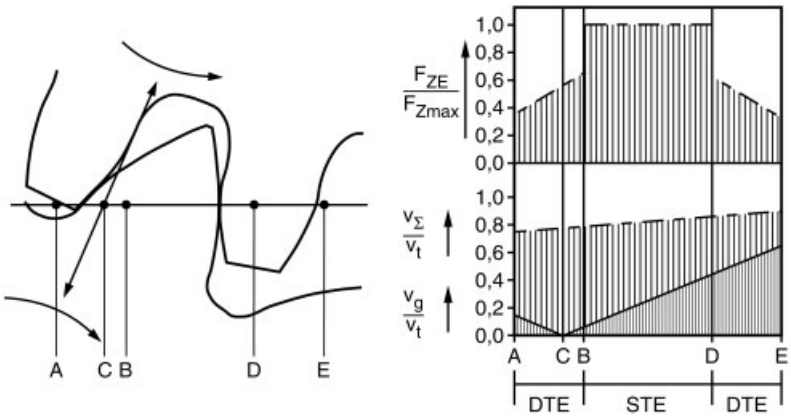

Figure 46. Load and speed distribution in tooth contact according to: A, beginning of tooth engagement; B, beginning of single-tooth engagement; C, pitch point, no sliding velocity; D, end of single-tooth engagement; E, end of tooth engagement; DTE, double tooth engagement; STE, single tooth engagement; F_{ZE}, tooth force within tooth engagement line; F_{Zmax}, maximum tooth force within tooth engagement line; v_Σ, hydrodynamic effective velocity; v_g, sliding velocity; v_t, circumferential velocity.

Specific Film Thickness $\lambda < 0.7$ (Borderline). Numerous industrial gears are operated in this range which is also called borderline lubrication. Falling below the critical value leads to an increased risk of surface damage as well as increased wear. In this range, the right lubricant in connection with a suitable lubricant additive system is particularly important.

If borderline lubrication is reached in the tooth contacts, the tooth flanks have to be protected against direct metal-to-metal contact in order to avoid tooth damage.

With sufficiently large film thickness rates ($\lambda > 2$), the effect of the lubricant viscosity is sufficient. In slowly running gears connected with low oil sump temperatures and low circumferential velocities, the lubrication conditions during borderline lubrication ($\lambda < 0.7$) is generally improved by using polar oil compounds, fatty acid or solid matter particles (molybdenum disulfide, graphite etc.). Due to the low circumferential speeds, the mentioned compounds can form physical deposits on the surfaces and develop a permanent protective layer.

In quickly running gears, in connection with relatively high oil sump temperatures and high circumferential velocities, EP additives improve the lubrication effect during borderline lubrication ($\lambda < 0.7$). Depending on the contact temperature, the EP additives begin a chemical reaction with the surface of the tooth flanks. This reaction leads to the generation of metallic soaps which serve as a slide layer, thus preventing metal contact.

9.3.2. Specific Gear and Transmission Failure

When constructing a gear, it must, first of all, be ensured that all sliding rolling contact points, especially the toothings, are supplied with sufficient quantities of lubricant in all operation modes. If this is the case, significant failures in the toothings, with the exception of corrosion or the like, will normally not occur in non-loaded gears. Failures in toothed wheels are always the result of excessive load, i.e., a too high local load of the respective material – lubricant combination.

Toothed wheel damage influenced by the lubricant is divided into three categories: Wear, scuffing and scoring, and fatigue (micropitting and pitting) [239].

Wear (→ Abrasion and Erosion). Continuous wear is predominantly observed at low circumferential velocities and during borderline lubrication, see [240].

The wear of a toothing can be significantly influenced by hardening the toothed wheels and selecting a suitable lubricant.

Scuffing and scoring of tooth flanks occurs predominantly at medium to high circumferrential velocities and in case-hardened toothings. According to [241], [242] the contacting surfaces weld together for a short time. The high relative velocity of the contact surfaces towards each other breaks this welded joints immediately, thus causing the typical scuffing and scoring phenomenon.

Scuffing is a very sudden damage which can already be caused by one single overload. Sudden scuffing can lead to the total destruction of tooth flanks.

Freshly produced toothed wheel surfaces without run-in are extremely subject to scuffing. By adding to the lubricant EP additives, the scuffing load capacity can be improved by a factor of 5 and more. The calculation of the scuffing load capacity is standardized according to [235], [243].

Fatigue.

Micropitting on tooth flanks can be observed in case-hardened toothings in practically all speed ranges. Tooth flanks with rough surfaces are especially subject to micropitting. Micropitting develop predominantly in the area of negative sliding velocities or in slip areas below the pitch circle as damage due to endurance failure according to [244].

A suitable lubricant with a positive influence on the micropitting must have a sufficiently high viscosity as well as a suitable lubricant additive system. When selecting a suitable additive system, the operational viscosity or temperature of a gear is of superior significance.

Pitting is an endurance failure which is observed on both quenched and tempered and case-hardened toothings for all types of gears and at all circumferrential speed ranges [245].

Pitting is observed predominantly on the tooth flank centers in the height of the pitch circle since this is where the highest load occurs (single-tooth engagement) or the largest stress-cycle amplitudes. Here, the material shows the quickest fatigue.

Recent examinations [235], [246] have shown that, in principle, an influence of the lubricant on the pitting load capacity is possible.

In general, an increase in the lubricant's viscosity has a service-life extending effect. However, it is not always possible and reasonable to increase the lubricant's viscosity in a gear (higher splash losses). Currently, lubricants with a lower viscosity are used more and more often.

Suitable lubricant additive systems exist – above all for lubricants with a low viscosity (ATF technologies) – which have a positive impact on the pitting load capacity of the toothings [246].

9.4. Gear Lubrication Oils for Motor Vehicles

Gear lubrication oils for motor vehicles have to meet the specific requirements of the gear drives (see Section 9.2). Mostly, the gearboxes which serve to change the gear ratio are manual transmissions with synchronization. They can also be automatic or semiautomatic transmissions or constantly variable transmissions (CVT). In automatics or CVTs, the lubricant is not only responsible for the gear drive lubrication but also for the function-related operation of wet clutches, wet brakes as well as torque converters and retarders.

Today, it is not possible to meet all working, operating and ambient requirements of the mentioned transmissions with only one lubricant. Currently, more and more tailor-made individual lubricant solutions are being developed for specific applications, special transmissions and transmission family types. Like the development of new engine oils, the current development of gear lubrication oils for motor vehicles has not yet come to an end and is mainly characterized by several factors:

- The requirements and specifications issued by the motor, transmission and vehicle manufacturers intend to considerably extend the oil-change intervals. Today, most of the manual transmissions and axle gears in passenger cars are operated with oil-fills for life (250 000 km). With respect to automatic transmissions in passenger cars, these long oil-change intervals will soon be achieved. For commercial vehicles, it is planned to achieve transmission service lives of 350 000 km, 500 000 km up to 1 000 000 km without oil-change.
- The use of mineral base oils is decreasing in the industrialized countries. Today, environmentally friendly, synthetic and ester-based

base oils are often used for reasons of environmental protection. Due to the better thermal oxidation stability, ester-based oils are used together with PAO and hydrocracked base oils.

Attempts are, furthermore, made to reduce lubricant toxicity, especially the chlorine content of commonly used additives.

Modern vehicles are expected to ensure an improved performance, a reduced noise emission and a reduced fuel consumption. This trend is expressed in the increasing use of lubricants with a low viscosity together with a reduced fill volume. Furthermore, multigrade oils are currently used on a large scale to considerably reduce the fuel consumption.

9.4.1. Lubricants for Gear Drives in Commercial Vehicles

Commercial vehicles are divided into so-called light trucks with a maximum load of 3 to 6 t and heavy trucks with a maximum load of more than 6 t. In 1999, the worldwide annual demand for gear lubrication oils for commercial vehicles amounted to approximately 647 000 t. In the same year, some 1.84×10^6 commercial vehicles were produced in Western and Eastern Europe, which corresponds to drive-line lubricant volume demand of 140 000 t of gear lubrication oils for the commercial vehicles' gearboxes and axle gears for the European market. In 1999, the number of manufactured light trucks amounted to some 1.505×10^6 vehicles.

Because the fill volume per transmission unit is two to six times larger and amounts to 10 to 30 L, heavy trucks are far more interesting for the oil manufacturers. At the same time, the oil development also confronts a bigger challenge in meeting the requirements made on the oils used in the commercial vehicles' large transmissions. In 1999, the number of manufactured heavy trucks amounted to approximately 335 000 in all European countries.

Tables 22 and 23 list the most important current (2000) specifications issued by the transmission manufacturers and automotive companies, split into manual transmission and rear axle gears for heavy trucks. With respect to the specifications, the modified new SAE classes are to be taken into account. Above all, the new classification SAE J2360 is to be mentioned which will replace the old MIL-PRF-2105E. In addition to this, the new SAE 'Automotive Gear Lubricant Viscosity Classification' is also to be taken into account.

Table 22. Specifications for lubrication oils for rear axle gears of commercial vehicles

Daimler – Chrysler 235.8	SAE 75W-90, extended drain, 250 000 km to 300 000 km
Volvo transmission oil specification 97312, STD 1273,12	SAE 75W-80/85 for transmissions made by Meritor extended drain, 400 000 km or 3 years
MAN 342 SL (synthetic, long distance)	extended drain, 320 000 km
Scania STO 1:0 (part synthetic, synthetic)	API-GL5+, extended drain, more than 120 000 km
RVI-TDL axle	SAE 75W-90, extended drain, 350 000 to 400 000 km or 2 to 3 years
ZF TE-ML 12 A-C (mineral, part synthetic)	extended drain, 120 000 km
ZF TE-ML 12 D (synthetic)	extended drain, 300 000 km

Table 23. Specifications for manual transmission gears of commercial vehicles

Daimler – Chrysler 235.11	SAE 75W-80/90, extended drain, 250 000 to 300 000 km
Volvo Transmission Oil Specification 97307, STD 1273,07	SAE 75W-90, extended drain, 400 000 km or 3 years
MAN 341 SL (synthetic, long distance)	for gears made by ZF and Eaton, extended drain, 320 000 km
MAN M 3343 mL/SL (mineral, synthetic, long distance)	total drive-line oils, extended drain, 180 000 km
RVI-TDL gear box	SAE 75W-90, extended drain, 350 000 to 400 000 km or 2 to 3 years
ZF TE-ML 02 A – C (mineral, part synthetic)	extended drain, 160 000 km
ZF TE-ML 02 D (synthetic)	extended drain, 300 000 km

This trend in the development of lubricants for commercial vehicles leads to a significant change in the current ratio between service fill and relubrication on the one hand and the factory fill on the other hand. The proportion of the service fill and relubrication of currently 75 % – with a factory fill proportion of 25 % – will be reduced to 20 % for the service fill and relubrication – with a factory fill proportion of 80 % during the next ten years.

All mentioned specifications constitute a challenge for the development of gear lubrication oils for commercial vehicles. Apart from the very high chemical and physical requirements

included in these specifications concerning the oxidation stability (test at 150 – 160 °C), corrosion, filterability, foam etc. (see Section 9.6) the mentioned specifications include a large number of mechanical – dynamic tests and require very high safety levels with respect to the scuffing and scoring, pitting and wear resistance of the toothings using different, standardized, nonstandardized and company-internal test procedures. This makes the development of new oils increasingly complicated and expensive. A particular obstacle in the development of oils for commercial vehicles is the improvement and optimization of the synchronization behavior during upshift or downshift operations.

Although many commercial vehicles of American manufacturers still do not use synchronized transmissions, almost all commercial vehicles in Europe are equipped with these transmission systems. The SSP 180 test bench has proven very successful in the oil development for a sufficient synchronization behavior, see also page 173.

The main problem in the oil development is to adjust the same friction or friction constancy during a required gear service life of approximately 100 000 gearshift operations with different material – lubricant pairs without significantly changing the other chemical – physical and dynamic – mechanical properties.

9.4.2. Lubricants for Gear Drives in Passenger Cars

The development of passenger cars is mainly influenced by the worldwide fierce competition of numerous producers. For the oil producers, this means to supply approved lubricants which have to be available in the same quality worldwide or at least in the production locations for gear-filling purposes.

A passenger car's sportiness requires a high performance and a high torque in identical or even smaller spaces for gearboxes. Due to the economy required at the same time, the lubricants are expected to optimize the total efficiency of the engine and total drive-line without having to be changed during the entire service life of more than 250 000 km.

The higher performance causes an increased heat development in the transmissions which accelerates the aging of the gear lubrication oils. The design of new vehicles leads to the development of carefully selected bodies, often with a low air resistance. This is often connected with an inferior air circulation and cooling of the transmissions. Additional capsules facilitate the reduction of driving-related noise to increase the driving comfort. At the same time, the poor heat dissipation leads to a heat stowing in the capsular spaces. Especially these factors increase the temperature in the vehicle's transmission and aggravate the operational conditions, thus facilitating the aging behavior of the oil.

New passenger cars include transmissions with a higher number of gear stages so that today passenger cars using gear drives with six stages are not unusual. To achieve this, all technical possibilities of dimensioning and material have to be exploited. This always has an effect on the corresponding lubricant when new lighter or more solid materials have to be used in the transmission as new friction partners with a friction and wear behavior so far unknown.

An increasing number of vehicles utilizes automatic transmissions and four-wheel drive systems. The necessary control elements and sensor systems consisting of electric components such as measuring units and sensors etc., are installed in the vehicles' transmissions. The compatibility of these sensible components with the oil leads to special requirements made on the lubricant and its development.

In Europe almost 90 % of all passenger cars are equipped with synchronized manual transmission systems. The variety of the friction material partners used in the synchronization of passenger cars is constantly increasing so that here more and more individual solutions are also required with respect to the lubricant. Another requirement on the gear lubrication oils for manual transmission systems, as well as for rear axle gears, is the very high scuffing resistance to ensure the protection of the spur gears' which have significantly more sliding proportions than the planetary gears' toothings normally used in automatic transmission systems.

All mentioned development trends are reflected in the increasingly stricter specifications and test procedures issued by leading automotive companies.

Several important specifications for lubrication oils for manual transmissions and transaxle

drive system gears for passenger cars are listed below:

VW, Seat, Skoda, Audi, TL 726, API GL 4 SAE 75W/SAE 80W (lifetime gearoil for manual transmissions and transaxles)
Daimler – Chrysler page 235.10
Volvo STD 1273,08
VW TL 52157 (75W 90, synthetic lifetime)
VW TL 52171 (75W, synthetic lifetime)
VW TL 52178 (75W, mineral lifetime)
Ford factory fill for life M2C 200C
BMW 602 00.0 (synthetic lifetime)

9.4.3. Lubricants for Automatic Transmissions and CVTs

The most important consumers of lubricants for automatic transmission and drive systems (ATFs) are off-highway vehicles and machines as well as commercial vehicles, mainly buses. In addition to this, ATFs are also used to fill power steering transmission systems in trucks, commercial vehicles and passenger cars.

90 % of all passenger cars on the North American and Asian vehicle market utilize automatic transmission systems. The total estimated ATF volume worldwide for the year 2001 amounts to approximately 992 000 t. The total volume breaks down as follows: North America (61 %), Asia Pacific (15.5 %), Europe (12.5 %), Latin America (8.8 %) and Middle East (2.2 %).

Most manufacturers of automatic transmission systems prescribe fluids for the application which meet the listed specifications:

Ford Mercon (revised September 1992)
Ford MerconV
GM Dexron II D
GM Dexron II E
GM Dexron III
GM Dexron IIIG
Allison C-4
Daimler – Chrysler 236.2
Daimler – Chrysler Mopar ATF3+, Mopar ATF 4+
Leyland E85

These specifications refer to the Asian, North American and European markets and to the business for service fill and relubrication of vehicles. Currently, the factory fill is subject to stricter specifications:

General Motors, Factory Fill GM 6418-M
Ford Factory Fill for Life, M2C 202B
Daimler – Chrysler Factory Fill for Life, MS 9602

European manufacturers of vehicles and, especially, transmission systems have issued particular specifications mainly for off-highway construction machines and commercial vehicles:

ZF TE-ML-14, Standard drain, 60 000 km
ZF TE-ML-14, Extended drain, 120 000 km
Voith G607, Standard drain, 60 000 km
Voith G1363, Extended drain, 120 000 km
Renk standard drain
Daimler – Chrysler 236.8
VW TL 52178

Worldwide, there are further specifications for off-highway construction machines and commercial vehicles with importance for the oil development:

Caterpillar TO-4
ZF-TE-ML 03
ZF-TE-ML 09
Komatsu
Komatsu – Dresser

ATF Lubricant Requirements for Hydrodynamic Transmissions. Automatic transmissions make use of the versatile possibilities of often series-connected hydrodynamic transmissions for the torque transfer and conversion. The hydrodynamic clutches facilitate an improved and smooth driving and operating behavior, especially with respect to heavy vehicles with a high inertia mass. A hydrodynamic transmission displays a constantly variable operation, whereby the resulting torque independently adapts to the respective load status by changing the rotational speed.

Apart from the hydrodynamic start-up clutches, hydrodynamic brakes, the so-called retarders and intarders are used, especially in commercial vehicles. They serve to limit the output speed when driving down long slopes in order to reduce the load on the vehicle's brakes or to go easy on them.

The fluid machines ensure the uncoupling of the drive-line and output torque and, thus, a constantly smooth and comfortable force transfer in

the drive shaft. They will, however, only work adequately with slip and are, thus, characterized by significant fluid losses. Apart from the larger weight, this constitutes another reason for the generally higher fuel consumption of vehicles with automatic transmissions in comparison to vehicles with manual transmissions.

During start-up procedures and permanent operation as well as during braking procedures, the lubricant as a force transfer medium in the hydrodynamic transmissions is subject to an extremely quick generation of heat and load despite installed radiator systems in comparison to gear drives. Due to the function-related fluid losses, hydrodynamic clutches, torque converters and brakes always require low-viscosity, mildly added gear lubrication oils with a high oxidation resistance. Short-term oil temperatures of more than 160 °C during the operation of a vehicle are not unusual. Especially for these fluid machines, a good viscosity – temperature behavior, corrosion protection, optimal foaming behavior and air separation properties are a must. With respect to the viscosity and $V-T$ behavior lubrication oils for automatic transmissions are very similar to engine oils.

To facilitate the distinction between ATFs and the engine oils, ATFs are always dyed with red color.

ATF Lubricant Requirements for Wet Clutches and Brakes. Like in the synchronization, the friction characteristics are of great importance for a comfortable and regular gearshift behavior during the transmission's entire service life in wet clutches and brakes as well. A wet clutch or brake consists of disk-plate fin blocks. The clutches and brakes are opened and closed hydraulically. Apart from the desired friction characteristics, the oil predominantly serves to ensure the cooling of the friction partners. Often the oil is injected from the interior into the disk-plate blocks via the rotating shaft.

To adjust a lubricant to a desired friction behavior or friction characteristic, a test in the machine or, at least, a test of the disk-plates fins used later in the machine is required. Therefore, test benches have been developed especially for this purpose which enable a more precise adjustment of the gearshift behavior to the selected material – lubricant combination.

A tester which demonstrates the different sliding – friction behavior of oils is the low-velocity friction apparatus (LVFA). Figure 47 gives an example of measured friction behavior for some lubricants as a function of the sliding speed.

Tests of original disk-plates can also be carried out using an SAE No. 2 tester or a DKA friction test bench. These testers work with given centrifugal masses which reduce the speed of the disk-plate blocks under defined test conditions. Static and dynamic friction coefficient and friction losses are determined using a high-resolution measuring technology.

According to the definition, the static friction is measured during the tests at low differential rotational speeds ($< 2\,\mathrm{min}^{-1}$) or towards the end of the gear-shifting operation.

To test the gearshift and friction behavior, especially in case of permanent slip of the friction partners, the transmission manufacturer ZF has developed another test bench. The so-called GK test bench or CSTCC (continuously slipping torque converter clutch) enables the simulation of almost all real operating conditions. Even slip-controlled clutches of hydrodynamic clutches and torque converters can be simulated using the GK test bench. Figure 48 shows possible operating conditions.

Gear Lubrication Oils for CVT Applications. Constantly variable transmissions in motor vehicles enable the operation of a combustion engine along certain preferred characteristic curves in the torque – rotational speed field. In contrast to all other vehicle transmissions, the CVT enables an ideal alignment of the supply field of a combustion engine with the request field of the vehicle.

Up until today, mainly three constantly variable transmission concepts have proven to be successful to meet these requirements.

B-CVT Belt Drives. For passenger cars with a low or medium drive power of up to 150 kW, the belt drive, mostly a 'Van Doorne' belt, has been very successful (Fig. 49). The belt embraces the cone pulleys which have been hydrostatically and axially applied to the drive and output shafts. Thus, the radii of the belt drive's course as well as the gear ratio can be varied constantly.

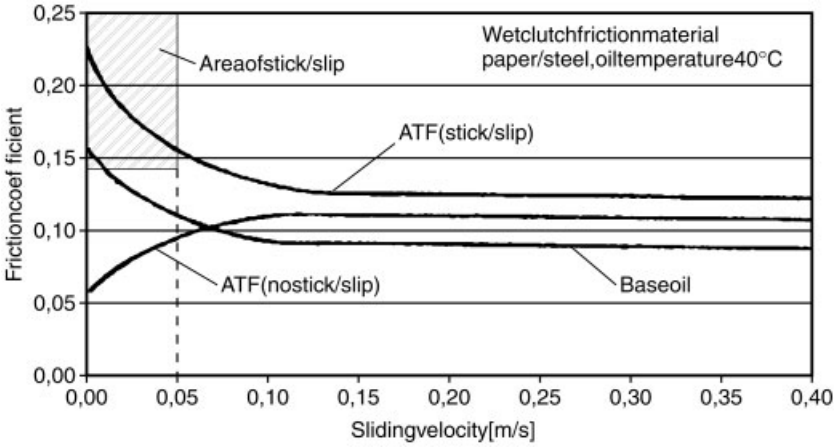

Figure 47. LVFA (low velocity friction apparatus)
Determination of the dependence of friction coefficient on sliding velocity to avoid stick – slip behavior

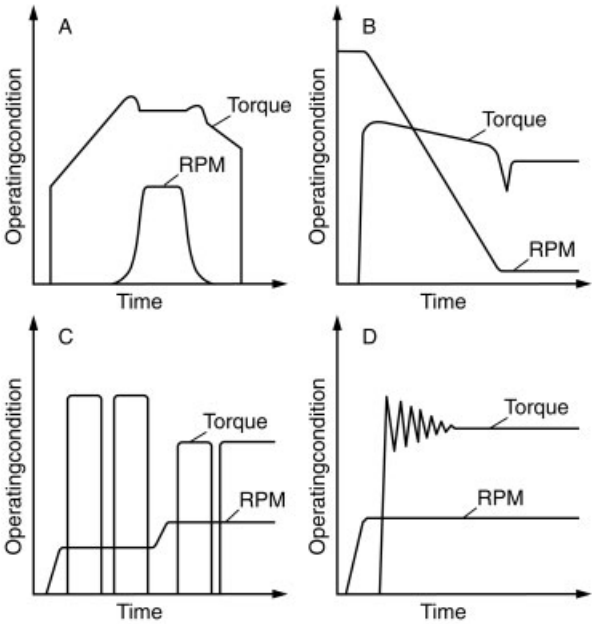

Figure 48. GK test bench. Test runs, variety of operating conditions
A) Opening and closing; B) DKA operation (limited); C) Similar to LVFA;. D) Static operation (friction vibrations)

For the instationary slide – roll contacts between the belt drive and the cone's contact surface, the lubricant is an extremely important construction component. The contact points are often subject to very high contact temperatures and mixed-lubrication conditions. Here, a very good wear protection of the oil against the extremely hardened surface of the cone pulleys is required. In addition, the cone's finely ground contact surfaces require from the lubricant a very high pitting load capacity. On the one hand, a slip of the belt drive due to insufficient contact pressure forces should be avoided by means of control engineering. This cannot be avoided totally. Accordingly, the lubricant also has to ensure a sufficient scuffing load capacity for these operating conditions.

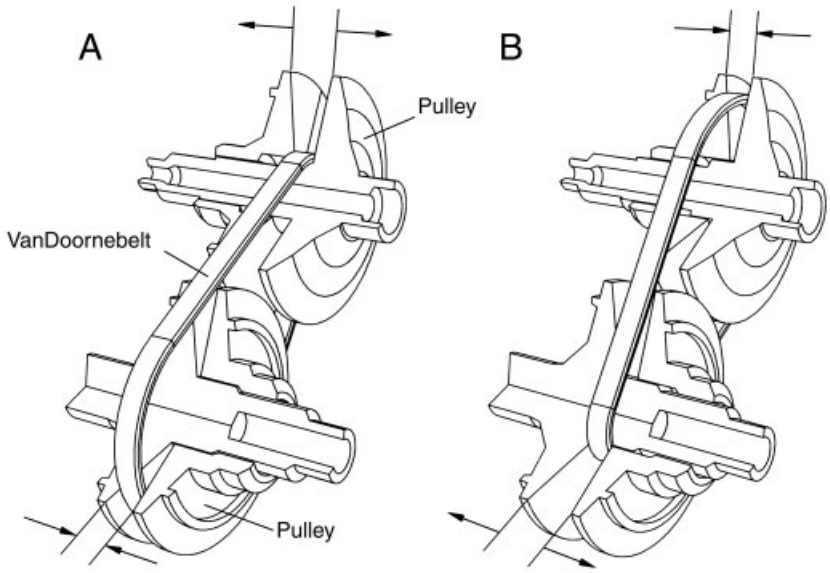

Figure 49. Adjusting device – principle of a B-CVT (Van Doorne Belt)
A) Gear ratio in position LOW; B) Gear ratio in position TOP

Nearly all belt drives of vehicles have a toothed wheel gear stage as well as a hydrodynamic start-up clutch. In order to increase the driving comfort and reduce the losses, slip-controlled clutches are used. Thus, the same requirements apply for both CVT oils and ATFs. Currently, CVTs are filled with ATF oils which are slightly modified lubricant variations or have been individually adjusted to the respective CVT. The viscosity, additives and base oils are very similar to ATFs, however, the so-called friction modifiers are of a greater importance.

T-CVT Traction Drives. The belt drive system's mechanics has reached its limits in power transmission at a driving power of slightly more than 150 kW, i.e., in mid-range passenger cars or limousines. Currently, constantly variable transmission concepts are tested in this respect which are based on so-called traction drives (Fig. 50).

The constantly variable adjustment of a traction drive takes place by tipping the transmission or idle wheels which, axially pressed together, roll on the semi or full-toroids' courses in a force-conclusive way. In traction drives as well, the lubricant is to be considered an important construction component and is as significant as the material, the surface treatment and the hardening of the rollers and toroids. Of all transmission types, semi and full-toroid transmissions stand out due to the highest surface pressure and circumferential speeds in the slide – roll contact points. Surface pressures of 4500 N/mm^2 and circumferential speeds of approximately 50 m/s in traction drives are not unusual.

In traction drives, the friction depends heavily on the oil, the material pair and the slip. Mainly those lubricants are used which enable a high transmission performance at the lowest slip possible, thus having a high friction coefficient.

Adequately added, naphthene-based hydraulic fluids have proven very successful for these applications. However, synthetic cycloaliphatic oils with a particularly high friction coefficient are even more suitable for these purposes. When using traction fluids in transmissions, e.g., in roller bearings and toothed wheels connected in series, the higher friction, however, leads to excessive and undesired overheating.

Another disadvantage of the traction fluids is a relatively low flash point of 130 to 150 °C. Therefore, the very high contact temperatures in

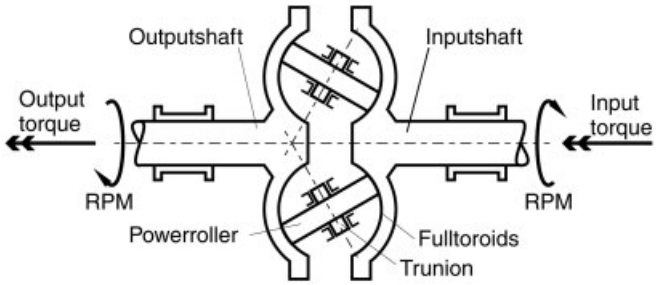

Figure 50. Adjusting device – principle of a T-CVT (traction drive)

the slide – roll contacts cause undesired evaporation losses.

Hydrostatic dynamic powershift drives are used in agriculture and in tracked vehicles with a usually very high drive power of more than 300 kW. In these transmission systems, planetary gear stages branch some of the drive power into a closed hydrostatic circulation system consisting of a controllable adjustment pump, mostly an axial piston pump, and a hydrostatic constant engine, mostly with inclined axles. The branching of the power and the control of the output speed depending on the volume flow takes place through the adjustment of the axial piston pump. These transmissions are always adjusted variably and under load in the entire adjustment range.

The requirements made on the lubricants used in this respect are limited to the toothings, roller bearings and hydraulic systems. For reasons of pumpability, the very good $V-T$ behavior of the low-viscosity hydraulic oils used is of great importance. In addition, these transmissions are to be protected against wear and corrosion using a suitable lubricant. ATFs are also often used.

9.5. Multipurpose Lubricants in Vehicle Gears

Multipurpose lubricants are used in agricultural and working machines such as tractors, harvesters, etc.

In these vehicles, the long-term and perfect function of the wet, force-conclusive additional clutches and brakes is to be ensured. The scuffing load capacity of the balancing hypoid transmissions is to be guaranteed using a suitable lubricant. Friction, hydraulic and wear requirements have to be met. In order to ensure the driving operation even at low temperatures, torque converters have to work adequately and safely, even under conditions of permanent slip of the wet clutches. Therefore, multipurpose oils have almost always a low viscosity and stand out due to a very good $V-T$ behavior. The presence of water and dirt has a significant impact on these oils, especially in respect of their foaming behavior.

This is aggravated by the fact that the mentioned requirements often have to manage with only one oil circulation system.

Against the background of these requirements it is easy to understand that major manufacturers of tractors and agricultural machines, such as Ford, John Deere and Massey Ferguson, have developed their own lubricant specifications for UTTOs and STOUs. The most important of these specifications are listed below:

UTTO Specifications
- Massey Ferguson M1127A
- Massey Ferguson M1127B
- Massey Ferguson M1135
- Massey Ferguson CMS-M1143 Limits
- Ford ESN-M2C-86C
- Ford ESN-M2C-134D
- John Deere J20C
- John Deere J20D
- MS 1207

STOU Specifications
- Massey Ferguson CMS-M1139
- Massey Ferguson CMS-M1144 Limits
- Ford ESN-M2C-159-C
- John Deere J27

9.6. Gear Lubricants for Industrial Gears

Industrial gear applications differ from vehicle transmissions mainly due to their larger variety and higher number of combinations of toothing types and sizes used (Fig. 45). Most important are worm gears, planetary gears and helical spur gears with crossed axis. Gears for industrial application are also applied in a larger variety of possible working and ambient conditions. They stand out due to much higher torque and performances, connected with clearly larger housing dimensions. At the same time, the gears size also requires larger lubricant quantities. The service life requirements on industrial gears are clearly higher than those made on vehicle transmissions (see Fig. 43).

In comparison to vehicle gear lubricants, industrial gear oils meet less requirements at one and the same time. With respect to industrial applications, the user has mainly to regularly and punctually change the oils used in accordance with the instructions of the oil producers. In addition, once-through lubrication is often applied. In this respect, the environmental compatibility of the lubricants used should especially be taken into account.

In opposite to vehicle gears, the type of lubrication in industrial gears can also be very different. Usually, vehicle gears are equipped with an oil immersion or injection lubrication system. According to their working conditions, industrial transmissions can, however, be lubricated manually by dropping or pouring, operating in an oil sump, through oil mist or – as well – through an oil injection system. Often larger oil lubrication systems are employed, e.g., in printing presses or paper machines, with fills of several hundreds of liters of lubricant. Figure 51 gives a schematic overview of a lubrication system commonly used today.

During the lubrication with an oil lubrication system, the total volume of the oil must not be selected in a too low quantity so that the air in the oil can be released. In this respect, the oil's air separation and foaming properties play a very important role since air is a bad lubricant. The purity of the lubricant during the operation of these systems is another central factor with respect to the gear's service life and, thus, the oil's filtration and filterability.

The viscosity-dependent pumpability of the oil, especially at cold temperatures as well as during the start-up of such systems, is also to be taken into account precisely. The wrong viscosity selection of a lubricant can lead to the standstill of the entire system. A guideline for the adequate viscosity selection as a function of load and speed for spur gears and worm gears in industrial applications is given in [247].

In comparison to the lubrication oils for vehicles transmissions, the worldwide number of specifications concerning the properties of the lubricants for industrial applications is quite small. Important specifications issued by the gear manufacturers and end consumers are listed below:

David Brown S1.53101 (Jan 1985)
U.S. Steel 224
DIN 51517 part 1 to part 3
Flender sheet A, 02.03.1999, Rev. 0

These specifications cover both simple mechanical – dynamic model test procedures and common component testers and test standards with roller bearings and toothed wheels. Apart from these specifications used worldwide, many transmission and system manufacturers are currently issuing their own, more sophisticated specifications for their industrial gears with increasing requirements.

***V – T* Behavior and Viscosity Index.** According to the ambient and operating conditions, the required $V-T$ behavior of gear lubrication oils constitutes a very important requirement during the application. Here, the base oil's properties according to the viscosity range is of great importance. Worldwide, the lubrication oils for industrial gears are subject to the ISO viscosity grade conversions. In the American region as well as in a large part of the automotive industry, the SAE gear viscosity numbers apply. In the American and, mainly, in the Asian regions, the viscosity ranges for AGMA lubricant are used most often.

Corrosion and Rust Protection. Especially corrosion and corrosion protection play a very important role with respect to lubrication oils for industrial gears. Due to the required long service lives of industrial gears, severe corrosion can facilitate an unexpectedly quick failure

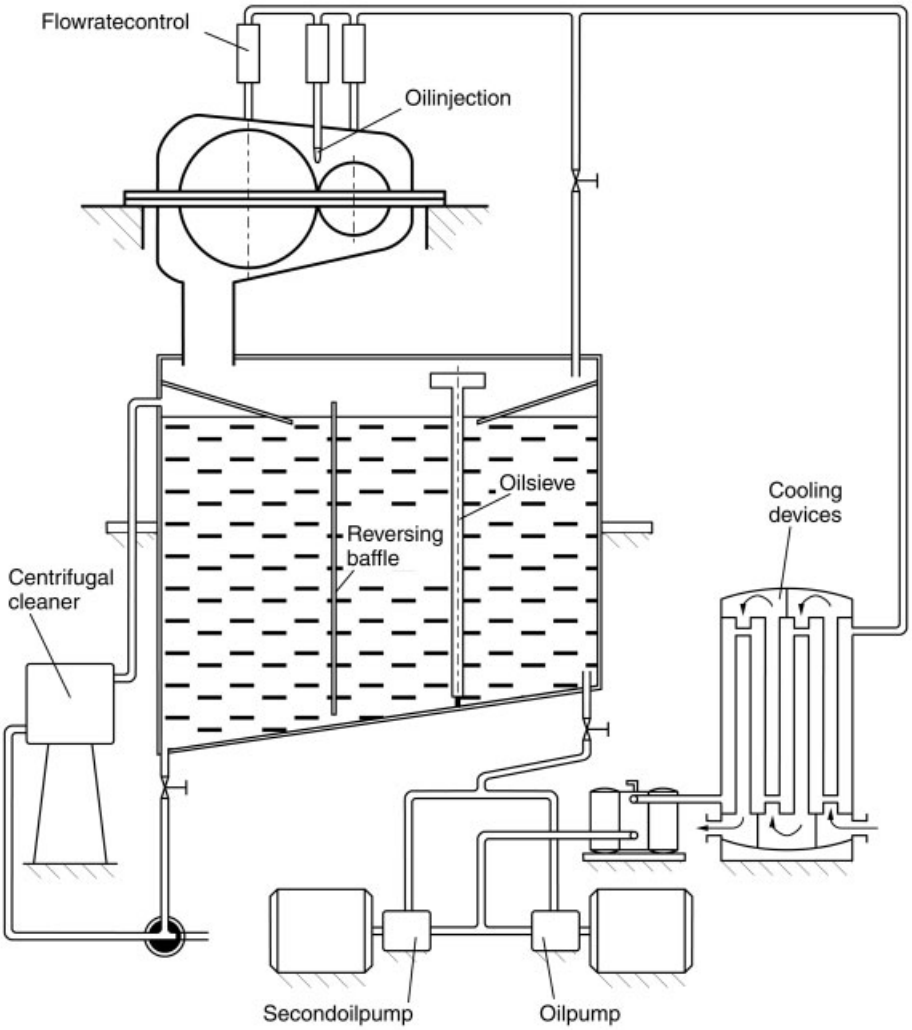

Figure 51. General layout of a lubrication system for large industrial gears

of bearings, toothed wheels and other important gear components. This has to be avoided by using a protecting lubricant combined with a suitable additive. Some industrial applications are aggravated by the presence of salt water in the gear, thus requiring special test methods.

Oxidation Stability. The aging behavior of a lubricant is connected with a change in viscosity and with an increase in the acid number. Today, sensitive gears are often tested on a regular basis. Should the acid number change noticeably, the oil producers recommend an oil change.

Flash Point and Pour Point. According to the application, the specifications almost always define applicable flash and pour points.

Demulsibility and Water Separation. Often the water cannot be prevented from penetrating the lubricant. Once the water proportions in the oil have exceeded a certain limit, the only possibility left is to change the oil or to separate oil and water. In large oil systems, the water will sink to the bottom of the oil tank due to the higher specific weight and can be released through a drain cock. This, however, requires a good water separation behavior of the lubricant.

Air Separation. Naturally, air is mixed into the oil in each gear. Since air is also a poor heat conductor, the air should be separated from the oil as quickly as possible in order to minimize the air's proportions in the oil.

Paint Compatibility. Because of the improved protection and extended service life, the interiors of large industrial gears are also coated with paint which can be attacked, softened and stripped off by oils. This applies especially for one-component paints whereas two-components paints are generally resistant against polyalkylene glycols.

Seal Compatibility. Like paints, seals often have organic components which can be attacked by lubricants. The risk of an attack rises considerably with an increase in the oil temperature. The elastomer compatibility tests of important specifications for the combination of lubricant and seal include both static and dynamic tests.

Foaming. Especially contamination cause the development of partly large quantities of foam in oils. This foam development can reach dimensions so that the foam escapes the gear through de-aerators, thus resulting in a contamination of the environment. In order to avoid this, only those oils should be used which in their fresh condition stand out due to a very low foaming tendency.

Environmental and Skin Compatibility. Today, especially the environmental and skin compatibility is considered very important by a growing number of gear manufacturers and system operators due to the effect on health and environment. Numerous specifications require compatibility tests.

Open gear drives are often found in the cement industry – the so-called milling gears – in rotary kilns in the iron and steel industry, in coal-burning plants or in open-cut mines. These open large gears are often lubricated with sprayable adhesion lubricants. Apart from the requirements made on the load capacity and wear protection, priority is also given to the adhesion properties, sprayability, pumpability and corrosion protection. In the mentioned systems, solid so-called compounded lubricants have proven to be successful as additives to such adhesion lubricants. Today, the health and environmental impact which can occur in oil mist and spray lubrication systems as well as the direct losses due to oil-containing waste air minimize careful oil separation processes. Open gears can also be lubricated manually or with the help of dosing pumps in regular intervals with high-viscosity oils.

10. Compressor Oils

10.1. Gas Compressor

10.1.1. Displacement Compressors

Reciprocating Piston Compressors. In reciprocating piston compressors the oil is also subject to enormous stress resulting from the high temperatures created when the medium is compressed (which can cause oxidation and lead to deposits) and in the case of air, the oxygen enrichment. The cleanest possible air or gas should be compressed because contaminants can accelerate oxidation and wear. In the case of drive unit lubrication, the lubrication of the bearings is of primary importance.

For reciprocating piston compressors, normally, lubricants based on mineral oil according to DIN 51 506 – VCL, VDL (or PAO- or diester-based lubricants) are used with viscosity grades of ISO VG 68 to ISO VG 150. Mobile compressors are often lubricated with monograde engine oil (SAE 20 – SAE 40) [250], [251]. Small to medium-sized piston compressors are used for pressures up to 10 bar.

Rotary Piston Compressors, Single Shaft, Rotary Vane Compressors. The pressure chambers of rotary piston compressors are cooled and lubricated by total-loss systems or by direct oil injection. The lubrication of rotary piston compressors is similar to the lubrication of the cylinders in reciprocating piston compressors insofar as the lubricant is subject to high outlet temperatures in both cases. In the case of oil-injected and oil cooled rotary piston compressors, a quantity of oil is continuously injected into the compressor chambers. The quantity of the oil is such that the outlet

Table 24. Overview of normally used compressor oils. The viscosity and the quality recommendations of the compressor manufacturers must be taken into consideration.

Viscosity classification	Type piston compressors[a]	Screw compressors oil-injected	Sliding vane compressors[a,b]	Turbo compressors (axial and radial)[c]
ISO VG 32		MO HC-oils PAO		TDL 32 TDL 32EP synth. oils
ISO VG 46		MO HC-oils PAO POE		TDL 32 TDL 32 EP synth. oils
ISO VG 68 (SAE 20W-20)	MO PAO diester	MO HC-oils PAO POE	MO diester HC-oils	
ISO VG 100 (SAE 30)	MO PAO diester		MO diester HC-oils	
ISO VG 150 (SAE 40)	MO PAO diester			

MO = mineral oil; PAO = polyalphaolefin; HC = hydrocracked oil; POE = biologically degradable polyol esters.
Diester, polyolester and PAO: for very hard working conditions, increase of service intervals is possible.
HC-oils: for medium and hard working conditions.
MO: for normal and medium working conditions.
Lubricants for roots-compressors: HL, CL, CLP; ISO VG 100-150, DIN 51 524, DIN 51 517.
Lubricants for vacuum pumps: ISO VG 68-150.
[a] Total-loss lubrication: HD-monograde motor oils HD 20W-20, HD 30, HD 40.
[b] For oil-injected compressors in mobile equipment (e.g., railways, buses): multigrade motor oils (e.g., 10 W 40, 15 W 40). For oil-injected compressors in stationary units: turbine oils according to DIN 51 515 TDL, air compressor oils according to DIN 51 506 VCL, VDL. For hard working conditions: monograde motor oils HD 20W-20, HD 30, MIL 2104 D.
[c] Turbine oils according to DIN 51 515 TDL or TDL-EP with extreme pressure additives.

temperature does not exceed 100 – 110 °C. At the same time, it seals the pistons against the housing and protects against wear. The cooling of the medium results in an increase in compression performance. Cooling and sealing increase the volumetric efficiency and thus the overall efficiency of the compressor. The oils used are normally VCL or VDL according to DIN 51 506 with an ISO VG between 68 and 150 or monograde SAE 20 – SAE 40 engine oils .

Screw Compressors. In lube-injected screw compressors, the oil has a lubricating, sealing and cooling function. The lubricant is injected into the pressure chamber between the rotors at about 3 – 4 bar. It then forms a hydrostatic and a hydrodynamic lubricating film. The oil therefore lubricates the meshing rotors and the plain and roller bearings which are part of the geared coupling. Furthermore, it seals the gaps between the rotor and the housing. It also helps absorb heat and dissipate this via radiators. The temperature of the compressed air of about 80 °C to 100 °C is adjusted by the quantity of oil injected. Downstream oil separators (normally cartridge filters) remove the oil from the air.

As the viscosity of the oil is of primary importance to elastohydrodynamic lubrication and thus for the mechanical stability of the film, it must be matched to start-up and normal running conditions. As a rule, ISO VG 46 lubricants cover most manufacturer's recommended viscosity thresholds of about 10 mm^2/s at operating temperature to about 500 mm^2/s when starting-up. This range also satisfies most applications in central Europe. Higher viscosity ISO VG 68 oils or synthetic ester-PAG or PAO-based lubricants are used in countries with high ambient temperatures. In recent years, lubricants based on hydrocracked oils have found increasing acceptance.

Roots Compressors. Recommended lubricants include DIN 51 517 CL and CLP or HD SAE oils in the viscosity grades ISO VG 68 and ISO VG 100 [250].

10.1.2. Dynamic Compressors

Turbo Compressors. The oils for turbo compressors lubricate bearings, radial shaft seals and possibly gears via a positive-feed circuit. In some cases, the bearings are lubricated with grease. Ideally, the same lubricant should be used for the compressor and its drive. Most often, DIN 51 515 TDL 32 and TDL 46 turbine oils or TDL-EP grades.

Turbo compressors are principally used for creating compressed air in mines and industrial manufacturing plants [249], [250].

10.1.3. Preparation of Compressed Air

The oil injected into oil-cooled screw and rotary piston compressors is always removed from the compressed air. The oil which is mixed with the highly compressed air is removed and collected in single or multistage downstream oil separators. Before the oil is re-circulated, it is filtered and cooled. Depending on the specific requirements, the compressed air may then pass a number of subsequent treatment stages such as refrigerant dryers or absorption dryers. Very low residual oil quantities in the air can be achieved by the fitting of a series of in-line oil separators [252–254].

10.1.4. Oils for Compression of Other Gases

Oxygen Compressors. Because of explosion hazards when oxygen is compressed, pressure chambers lubricants must be mineral oil-free. Water and water-based solutions such as glycerol can be used for cylinder lubrication. Mineral oil-based products may be used for compressor drives if it does not come into contact with the pressure chambers [250].

Acid Gas Compressors. Gases often contain acidic components such as SO_2 or NO_x. If standard compressor oils were used for such applications, the lubricating oil would soon become over-acidified. To counter this, lubricants are used for such applications which contain highly alkaline additives. These components can neutralize the acidic components in the gas. In these cases, it is recommended that monograde engine oils (20W-20, 30W, 40W) with high alkaline reserves (high TBN) are used [250].

Inert Gas Compressors. When inert gases are compressed, the same rules as for air compressors should be used [248], [250].

Hydrocarbon Compressors. Hydrocarbons such as ethane, propane etc. are easily soluble in mineral oil. This causes the viscosity of the lubricating oil to fall if mineral oil- based products are used. For this reason, higher viscosity mineral oils such as ISO VG 100 and ISO VG 150 must be used in piston compressors whose crankcases are subject to low inlet pressures ($1-3$ bar). In the case of screw compressors (high pressure; $10-15$ bar), ISO VG $68-100$ ester- or polyglycol-based lubricants with lower hydrocarbon solubility are recommended [248], [250].

Vacuum Pump Lubrication. Vacuum pumps are compressors whose inlet is connected to the chamber where the vacuum is created. VDL compressor oils can be used for low vacuums. Greater vacuums require synthetic oils with low vapor pressures (mostly synthetic ester oils). The lubricant selection must consider if the medium to be extracted is not air, but for example a refrigerant. In such cases, a compatible refrigerator oil can be used [250].

10.1.5. Characteristics of Gas Compressor Oils

Compressors whose chambers are lubricated pose particular safety problems if air or aggressive gases contact the lubricant. The selection of the most suitable lubricant depends on the type of compressor in question, the pressures involved, the outlet temperatures and the type of gas being compressed. Piston compressors which generate the highest pressures are particularly problematic. Turbo compressors which only have lubricated bearings and non-lubricated pressure chambers pose the least problems. Rotary and screw compressors with outlet pressures under 10 bar and correspondingly low outlet temperatures are examples of average compressor lubrication application. Table 24 shows an overview of normally used compressor oils.

10.1.6. Standards and Specifications of Compressor Oils

DIN 51 506 describes the classification and requirements of lubricating oils which are used in piston compressors with oil-lubricated pressure chambers (also for vacuum pumps). Lubricants for screw and oil-injected rotary vane and screw compressors are not included in DIN 51 506 [251]. Table 25 shows the classification of air compressor oils according to DIN 51 506. Tables 26, 27, and 28 contain the minimum requirements of air compressor oils according to DIN 51 506.

Table 25. Classification of air compressor oils according to DIN 51 506 – Table 1, September 1985

Maximum compressed air temperature, °C		
Lubricating oil category	For compressors on moving (mobile) equipment for brakes, signals and tippers	For compressors with storage tanks and pipe network systems
VDL	up to 220	up to 220
VC	up to 220	up to 160[a]
VCL		
VB	up to 140	up to 140

[a] Rotary multi-vane compressors designed for a once-through lubrication can be operated at compressor end temperatures of up to 180 °C using lubricating oils doped in the same manner as motor lubricants or doped compressor oils, provided that the requirements specified for VCL lubricating oils in table 2 are complied with.

According to this standard, such lubricants are pure mineral oils or mineral oils with additives to increase aging resistance and corrosion protection. The classification of the lubricants depends on the expected outlet temperatures and the general application. DIN 51 506 differentiates between lubricants for mobile applications and stationary applications with reservoirs. Principal differences between the listed groups: VB/VBL, VC/VCL as well as VDL are the use of oxidation and corrosion inhibitors, aging stability and residue formation and the quality of the base oils (cuts). The difference between group VB/VBL and group VC/VCL lies in the aging behavior (formation of Conradson coke after air-induced aging). Group VDL oils have to pass a more difficult aging test (formation of Conradson coke after air-induced aging in the presence of ferrous oxide). DIN 51 506 VDL oils display the best thermal and oxidation stability and form the least residues. The selection criteria in DIN 51 506 were adopted into the 1983 ISO/DP 6521 draft [251], [255]. Table 29 defines the requirements of air compressor oils for reciprocating piston compressors.

The selection criteria for screw compressor and piston compressor oils differ greatly. In oil-flooded rotary vane and screw compressors, the injected oil is constantly in contact with the 80 – 100 °C hot medium being compressed. The compressed medium and the oil are well mixed and the oil has to be separated by downstream separators and filters. This places special demands on the lubricant. On the whole, these include low foaming, excellent air release and good demulsification (separation) of condensed water. In addition of course, the specifications regarding wear protection (FZG $\geq$ 10, DIN 51 354), minimum formation of deposits, good corrosion protection etc. also apply. Most manufacturers of oil-injected rotary and screw compressors issue their own lubricant specifications. A draft of ISO/DP 6521 for oil-injected screw compressors has existed since 1983. Table 30 describes the requirements of air compressor oils for screw compressors.

10.2. Refrigerator Oils

10.2.1. Introduction

The longevity expected of refrigeration compressors is closely connected to the high quality which is required of refrigerator oils. The interaction with other substances which the refrigerator oil comes into contact with, and especially the extremely high and low temperatures, makes very specific demands on refrigerator oils.

The principal function of a compressor oil is to lubricate the pistons or rotors and to lubricate and seal the valves and, in some cases, the slip-ring seals. Furthermore, the refrigerator oil must dissipate heat away from hot compressor components and assist in sealing the compression chambers and valves. The refrigerator oil serves as a hydraulic control and functional fluid in refrigeration compressors [256], [257]. It is vital that refrigerator oil which reaches and collects in colder sections of the circuit in the form of oil vapor or oil mist or as a result of splashing must be returned to the compressor by mechanical means (oil separator) or via the refrigerant

Table 26. The minimum requirements of air compressor oils according to DIN 51 506 – VB, VBL

Grade VB is a pure mineral oil, Grade VB-L contains additives L to increase aging resistance and corrosion protection Lube Oil Group VB and VB-L									
Viscosity grade	ISO VG 22	ISO VG 32	ISO VG 46	ISO VG 88	ISO VG100	ISO VG150	ISO VG220	ISO VG320	ISO VG460
Kinematic viscosity (DIN 51 561 / 51562 part 1)									
min.	19.8	28.8	41.4	61.2	90	135	198	288	414
mm^2/s at 40 °C	to	to	to	to	to	to	to	to	to
max	24.2	35.2	50.6	74.8	110	165	242	352	506
mm^2/s at 100 °C	4.3	5.4	6.6	8.8	11	15	19	23	30
Flash point, °C (COC) min. (DIN ISO 2592)	175	175	195	195	205	210	225	225	255
Pour point, °C, max. (DIN ISO 3016)	−9	−9	−9	−9	−9	−3	0	0	0
Ash, wt % max. (DIN 51 575)	VB: max. 0.02 % oxide ash; VB-L: Sulf. ash to be stated by the supplier								
Water soluble acids (DIN 51 558 part 1)	neutral								
Neutralization number (acid), mg KOH/g max. (DIN 51 558 part 1)	VB: max. 0.15 mg KOH/g; VB-L: to be stated by the supplier								
Water, % (DIN ISO 3733)	0.1 max.	0.1 max	0.1 max	0.1 max	0.1 max	0.1 max	0.1 max	0.1 max	0.1 max
Aging characteristics 1.5% CRC max. after air aging (DIN 51 352 part 1)	2.0	2.0	2.0	2.0	2.0	2.5	2.5	2.5	2.5
% CRC max. after air/Fe_2O_3 aging (DIN 51352 part 2)	not required								
Kinematic viscosity (at 40 °C), max. of 20 % dist. residue (DIN 51 536 / 51 561 / 51 562 part 1)	not required								
Distillation residue % CRC max. of 20 % distillation residue (DIN 51 356 / 51 551)	not required								

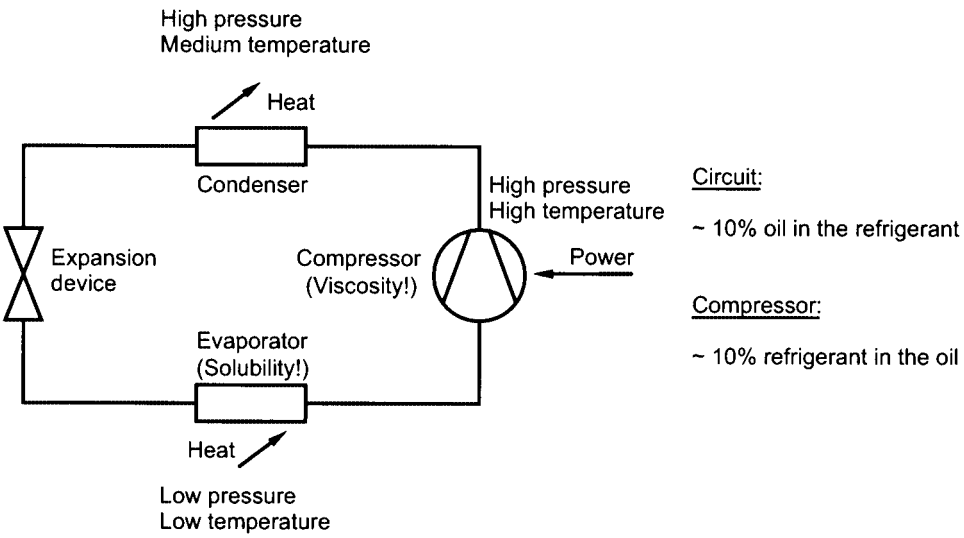

Figure 52. Principle of a vapor compression refrigeration cycle

Table 27. The minimum requirements of air compressor oils according to DIN 51 506 – VC, VCL

Grade VC is a pure mineral oil, Grades VC-L and VD-L contain additives L to increase aging resistance and corrosion protection					
Lube Oil Group	VC and VC-L				
Viscosity grade	ISO VG 32	ISO VG 46	ISO VG 68	ISO VG100	ISO VG150
Kinematic viscosity (DIN 51 561 / 51562 part 1) min.	28.8	41.4	61.2	90	135
mm^2/s at 40 °C	to	to	to	to	to
max	35.2	50.6	74.8	110	165
mm^2/s at 100 °C	5.4	6.6	8.8	11	15
Flash point, °C (COC) min. (DIN ISO 2592)	175	195	195	205	210
Pour point, °C, max. (DIN ISO 3016)	−9	−9	−9	−9	−3
Ash, wt % max. (DIN 51 575)	VC: max. 0.02 % oxide ash; VC-L: Sulf. ash to be stated by the supplier				
Water soluble acids (DIN 51 558 part 1)	neutral				
Neutralization number (acid), mg KOH/g max. (DIN 51 558 part 1)	VC: max. 0.15 mg KOH/g; VC-L: to be stated by the supplier				
Water, % (DIN ISO 3733)	0.1 max	0.1 max	0.1 max	0.1 max	0.1 max
Aging characteristics 1.5% CRC max. after air aging (DIN 51 352 part 1)	1.5	1.5	1.5	1.5	2.0
% CRC max. after air/Fe_2O_3 aging (DIN 51352 part 2)	not required				
Kinematic viscosity (at 40 °C), max. of 20 % dist. residue (DIN 51 536 / 51 561 / 51 562 part 1)	not required				
Distillation residue % CRC max. of 20 % distillation residue (DIN 51 356 / 51 551)	0.3	0.3	0.3	0.3	0.75

flow (refrigerant solubility) in all operating conditions. Figure 52 shows the principle of a vapor compression refrigeration cycle.

10.2.2. Minimum Requirements

The basic requirements of refrigerator oils are laid down in DIN 51 503-1. This standard defines the basic requirements of refrigerator oils according to the medium being compressed. The introduction of new chlorine-free, polar refrigerants such as HFC R 134a (to replace CFC R 12), necessitated a revision of DIN 51 503 which appeared as DIN 51 503-1 in November 1997 [258].

Refrigerator oils are classified in alphabetical groups according to the refrigerants being compressed:

KAA Refrigerator oils not soluble in ammonia
KAB Refrigerator oils soluble in ammonia
KC Refrigerator oils for partially- and fully-halogenated fluorinated and chlorinated hydrocarbons (CFC, HCFC)
KD Refrigerator oils for partially- and fully-fluorinated hydrocarbons (FC, HFC)
KE Refrigerator oils for hydrocarbon refrigerants such as propane or isobutane

The various refrigerants available are described in DIN 8960 and the ASHRAE standard (ANSI/ASHRAE 34-1992; American Society of Heating, Refrigerating and Air Conditioning Engineers) [259].

In addition to appearance, density (ISO 3675) and viscosity (DIN 51 550), several other properties are defined and determined:

Table 28. Minimum requirements of air compressor oils according to DIN 51 506 – VDL

Lube Oil Group	VDL				
Viscosity grade	ISO VG 32	ISO VG 46	ISO VG 68	ISO VG 100	ISO VG 150
Kinematic viscosity (DIN 51 561 / 51 562-1)					
cST at 40 °C	28.8 to 35.2	41.4 to 50.6	61.2 to 74.8	90 to 110	135 to 165
mm^2/s at 100 °C	5.4	6.6	8.8	11	15
Flash point, °C (COC), min. (DIN ISO 2592)	175	195	195	205	210
Pour point, °C, max. (DIN ISO 3016)	− 9	− 9	− 9	− 9	− 3
Ash, wt %, max. (DIN 51 575)	sulf. ash to be stated by the supplier				
Water soluble acids (DIN 51 558 part 1)	neutral				
Neutralization number (acid), mg KOH/g, max. (DIN 51 558 part 1)	to be stated by the supplier				
Water, % (DIN ISO 3733)	0.1 max.	0.1 max.	0.1 max.	0.1 max.	0.1 max.
Aging characteristics % CRC max. after air aging (DIN 51 352 part 1)	not required				
% CRC max. after air/Fe_2O_3 aging (DIN 51352 part 2)	2.5	2.5	2.5	3.0	3.0
Distillation residue % CRC max. of 20 % distillation residue (DIN 51 356 / 51 551)	0.3	0.3	0.3	0.3	0.6
Kinematic viscosity at 40 °C max. of 20 % distillation residue (DIN 51 536 / 51 561 / 51 562-1)	maximum of five times the value of the new oil				

U-tube flowing DIN 51 568
Flashpoint DIN-ISO 2592
Neutralization number DIN 51 558-3
Saponification number DIN 51 559
Oxide ash DIN EN 6245
Water content DIN 51 777-1
Pour point DIN-ISO 3016
Refrigerant miscibility E DIN 51 514
Refrigerant compatibility with R 134a ASHRAE 97/83

Table 31 gives an overview of the important refrigerants and suitable refrigerator oils on the market.

DIN 51503-2 describes criteria for evaluating used refrigerator oils [259].

10.2.3. Classification

Normally used refrigerator oils are listed in Table 32.

Dewaxed Naphthenic Refrigerator Oils. Naphthenic mineral oils are still the most significant group of oils for refrigerators using ammonia refrigerants along with CFC and HCFC (e.g., R 22). Naphthenic refrigerator oils generally display very low pour points, good cold flowing as well as high thermal and chemical stability. Selected cuts are normally used [256], [262].

Paraffinic Refrigerator Oils. Paraffinic refrigerator oils are ideal for use in R 11 and R 12 turbo compressors (ISO VG 68 and 100) because of their good $V-T$ behavior. These oils are not recommended for other compressors because of their generally inadequate solubility in refrigerants (miscibility gap, e.g., R 22). The boundaries between paraffinic and naphthenic oils are not rigid [256], [257].

Semisynthetic refrigerator oils are mixtures of highly-stable alkylbenzenes and highly-refined naphthenic mineral oils. The presence of alkylbenzenes greatly improves the solubility and thermal stability of the naphthenic components. The proportion of synthetic components is usually between 30 and 60 %. Semisynthetic oils are recommended for CFC/HCFC systems,

Table 29. The requirements of air compressor oils for reciprocating piston compressors – ISO/DP 6521 – Draft 1983

Mineral oil-based lubricants for reciprocating piston compressors

Category	ISO-L-DAA					ISO-L-DAB					Test Method
Viscosity grades	32	46	68	100	150	32	46	68	100	150	ISO 3448
Viscosity at 40 °C, $mm^2/s \pm 10\,\%$	32	46	68	100	150	32	46	68	100	150	ISO 3104 (IP71)
at 100 °C mm^2/s	to be stated					to be stated					
Pour point∗, °C max.	−9	−9	−9	−9	−9	−9	−9	−9	−9	−9	ISO 3016 (IP15)
Copper corrosion, max.	1b	1b	1b	1b	1b	1b	1b	1b	1b	1b	ISO 2160 (IP154)
Rust	no rust					no rust					ISO/DP 7120A (IP135A)
Emulsion characteristics											ISO/DP 6614 (ASTM D1401)
Temperature, °C			no requirement			54	54	54	82	82	
Time (min) to 3 mL Emulsion, max.						30	30	30	30	60	
Oxidation stability after aging at 200 °C											
Evapor. loss, %, max.	15	15	15	15	15			no requirement			ISO/DP 6617 Part 1
Increase in CCRC, %, max. after aging at 200 °C	1.5	1.5	1.5	2.0	2.0			no requirement			(DIN 51 352)
Evapor. loss, %, max.			not applicable			20	20	20	20	20	ISO/DP 6617 Part 2
Increase in CRC, %, max.			not applicable			2.5	2.5	3.0	3.0	3.0	(DIN 51 352)
Distill. residue (20 vol %) CRC, %, max.			not applicable			0.3	0.3	0.3	0.3	0.6	ISO/DP 6616 with ISO/DP 6615 and ISO 3104
Ratio of viscosity of residue to that of new oil, max.						5	5	5	5	5	

∗ When VG32 or VG46 oils are used in cold climate, pour points lower than − 9 °C are required.

Table 30. The requirements of air compressor oils for screw compressors – ISO/DP 6521 – Draft 1983

Mineral oil-based lubricants for rotary screw compressors											
Category	ISO-L-DAH					ISO-L-DAG					Test Method
Viscosity grades	32	46	68	100	150	32	46	68	100	150	ISO 3448
Viscosity at 40 °C, $mm^2/s \pm 10\,\%$	32	46	68	100	150	32	46	68	100	150	ISO 3104 (IP71)
Pour point∗, °C max.	−9	−9	−9	−9	−9	−9	−9	−9	−9	−9	ISO 3016 (IP15)
Copper corrosion, max.	1b	1b	1b	1b	1b	1b	1b	1b	1b	1b	ISO 2160 (IP154)
Rust	no rust					no rust					ISO/DP 7120A (IP135A)
Emulsion characteristics											
Temperature∗∗, °C	54	54	54	54	82	54	54	54	54	82	ISO/DP 6614 (ASTM D1401)
Time (min) to 3 mL emulsion, max.	30	30	30	30	30	30	30	30	30	30	
Foaming characteristics											ISO/DP 6247 (IP146)
Sequence I at 24 °C	300	300	300	300	300	300	300	300	300	300	
Tendency, mL, max.	nil					nil					
Stability, mL, max.											
Oxidation stability											
Evapor. loss, %, max.			to be decided					to be decided		to be established	
Increase in viscosity, %											
Sludge, wt %											
Increase in acidity, %											

∗ When VG32 or VG46 oils are used in cold climate, pour points lower than −9 °C are required.
∗∗ Required only in those applications where condensation of atmospheric moisture is a problem. Where this does not apply, oils with dispersant additives, which tend to have poor water separating properties, may be used satisfactorily.

Table 31. Classification of important refrigerants and refrigerator oils.

ASHRAE name	Trade name	Chemical name / formula	Refrigerator oil[a)]
Chlorine-free refrigerants and refrigerator oils			
R 134a	diverse	CH_2FCF_3	POE, PAG
R 507	Solkane 507, AZ 50	R 125/R 143a	POE
R 404 A	diverse	R 125/R 143a/ R 134a	POE
R 407 C	diverse	R 32/R 125/ R 134a	POE
R 410 A	Solkane 410, AZ 20	R 32/R 125	POE
R 600a/R 290	isobutane/propane	C_4H_{10} / C_3H_8	MO / AB
R 717	ammonia	NH_3	MO / PAO / AB
R 744	carbon dioxide	CO_2	synth. oil[b)]
Drop-in refrigerantns and refrigerator oils			
R 22	diverse	$CHClF_2$	MO / AB
R 401 A	MP 39	R 22/R 152a/ R 124	MO / AB
R 401 B	MP 66	R 22/R 152a/ R 124	MO / AB
R 402 A/B	HP 80/81	R 22/R 125/ R 290	MO / AB
R 403 A/B	69 S/L	R 22/R 218/ R 290	MO / AB
R 408 A	FX 10	R 22/R 143a/ R 125	MO / AB

[a] AB = alkylbenzene oil; MO = mineral oil; PAG = polyalkylene glycol; PAO = polyalphaolefins; POE = polyolester oil
[b] Development product; POE, PAG, ...

Table 32. Overview of normally used refrigerator oils. The viscosity recommendations of the compressor manufacturers must be taken into consideration

Compressor type	Refrigerant hydrocarbons[a] (e.g., R 290 R 600a)	Ammonia[b]	CFC, HCFC, (e.g., R 12, R 22)	CFC, HCF[c] (e.g., R 134a, R 404a)	Drop-In (e.g., R 402A, R 403A)
Hermetic compressors (e.g., piston compressors)	MO AB	–	MO AB (MO/AB)	POE	MO/AB
	ISO VG 15-32		ISO VG 15-32	ISO VG 10-32	ISO VG 32
Open piston compressors	MO AB PAO	MO AB PAO, PAG	MO AB MO/AB	POE	MO/AB
	ISO VG 46-100	ISO VG 32-68	ISO VG 32-68	ISO VG 32-68	ISO VG 32-68
Semihermetic compressors	MO AB PAO	–	MO AB MO/AB	POE	MO/AB
	ISO VG 46-100		ISO VG 32-68	ISO VG 32-68	ISO VG 32-68
Scroll compressors	MO AB	–	MO AB	POE	MO/AB
	ISO VG 46-100		ISO VG 32-68	ISO VG 32-68	ISO VG 32-68
Screw compressors	MO AB PAO, PAG	MO AB PAO, PAG	MO AB	POE	MO/AB
	ISO VG 68-220	ISO VG 32-68	ISO VG 68-150	ISO VG 100-150	ISO VG 68
Turbo compressors	MO PAO, PAG	–[d]	MO	POE	MO/AB
	ISO VG 68-100		ISO VG 68-100	ISO VG 68-150	ISO VG 68

[a] PAG-lubricants are partly soluble with hydrocarbons (low viscosity reduction); MO, AB and PAO are highly soluble with hydrocarbons (high viscosity reduction).
[b] MO, AB and PAO are not soluble with ammonia, PAG is (partly) soluble with ammonia, ISO VG 68 is used in piston compressors, up to ISO VG 220 is used in screw compressors, please use PAG compatible components, PAG is hygroscopic
[c] PAG-lubricants are used in R 134a air conditioning systems of cars and trucks; POE and PAG lubricants are hygroscopic
[d] Normally oil-free
MO = mineral oil; AB = alkylbenzene; MO/AB = mineral oil – alkylbenzene mixture; PAO = polyalphaolefin; PAG = polyalkylene glycol; POE = polyol ester.

low- temperature R 22 systems and for drop-in refrigerants such as 401 A/B, 402 A/B and R 22 mixtures [260], [263].

Fully Synthetic Refrigerator Oils.

Alkylbenzenes. Fully synthetic refrigerator oils are based on chemically and thermally highly stable alkylbenzenes. Alkylbenzenes have been used as refrigerator oils for a number of years. Carefully selected and specially treated alkylaromatics are used. A number of complex manufacturing stages ensure that the products are free of difficult-to-dissolve waxy substances and other contaminants such as sulfur. Alkylbenzene-based lubricants display excellent solubility in CFC (HCFC) refrigerants (e.g., R 22, R 502) and mixtures thereof at evaporation temperatures down to $-80\,^{\circ}C$ (e.g., R 22).

ISO VG 46 and 68 alkylbenzenes have proved particularly suitable for use in heavy-duty ammonia compressors with very high outlet temperatures. Compared to mineral oil-based refrigerator oils, alkylbenzenes form minimal amounts of coke and slash less when starting-up.

Alkylbenzenes are also increasingly being used in hermetically sealed and semisealed compressors. Alkylbenzenes are increasingly gaining acceptance for use with drop-in refrigerants such as R 401 A/B, R 402 A/B, R 22 mixtures and together with propane/isobutane [262], [263].

Polyalphaolefins are recommended for ammonia compressors because of their good thermal stability. The formation of oxidation products (coke) is avoided even at high compressor outlet temperatures. Compared to mineral oils and particularly in the case of screw compressors, the use of polyalphaolefins can reduce the amount of oil mist and oil vapor which collects in oil separators. The amount of oil in the refrig-

erant vapor can also be reduced to a minimum. As a result of their chemical structure, polyalphaolefins display good $V-T$ behavior (high VI) and thus good cold flowing characteristics. The low pour point and low viscosity of these products guarantee satisfactory oil return even at evaporation temperatures of – 50 °C.

Polyalphaolefins of the viscosity grade ISO VG 68 are generally used in screw and piston ammonia compressors [261].

Polyol Esters. Previously used refrigerator lubricants such as mineral oils, alkylbenzenes or polyalphaolefins are not or not sufficiently soluble in the new chlorine-free refrigerant mixtures such as R 134a, R 404 a and R 507. This led to the development of synthetic polyol ester-based refrigerator oils which are soluble in FC and HFC refrigerants according to DIN 8960. The chemical and thermal stability of these products is excellent.

Legislation passed in 1991 progressively banned the use of CFCs in all refrigeration units. Since January 1995, new equipment must be completely CFC-free. R 134a and R 22 should replace CFC R 12 in the long and short-term respectively. These substitute refrigerants are thus steadily gaining acceptance as is the demand for polyol ester-based, synthetic refrigerator oils. Ester oils are suitable for all refrigerant systems which use R 134a, R 404a, and FC and HFC mixtures. Products with the suitable viscosities (ISO VG 15–220) are available for industrial and household piston and screw compressors. The manufacturers' viscosity recommendations should be observed.

Similarly to all ester oils, saturated and high-purity polyol ester oils can hydrolyze if they come into contact with water in the compressor. It is therefore essential that these products are protected from water, and moisture in general, during storage and use. Ester oils are ultra-dried and are filled into air-tight metal drums with a water content of < 100 ppm in a nitrogen atmosphere [261], [264], [265].

In addition to their properties described in Section 5.3.1.2 polyol esters have the following specific characteristics:

- Excellent solubility in FC and HFC refrigerants
- Avoidance of oil build-up in the condenser/evaporator
- Very good thermal and chemical stability even in the presence of refrigerants
- Compatibility with all commonly-used sealing materials such as NBR, HNBR, EPDM and others
- The products are ultra-dried

Polyalkylene Glycols. Along with the introduction of R 134a refrigerants as a substitute for R 12 in the area of vehicle aircon systems, most of such compressors are designed to use polyalkylene glycol oils. These polyalkylene glycols are not always compatible and miscible with normal lubricants based on mineral oil, alkylbenzenes or esters. Polyalkylene glycols are naturally polar and thus miscible with R 134a. These polar characteristics make polyglycols very hygroscopic and this fact must be taken into consideration when handling these specialized lubricants. When filled, the water content of polyglycol oils should be < 1000 ppm (fresh oil < 500 ppm).

PAG-based refrigerator oils are ultra-dried before use [261].

Fully synthetic, polyglycol-based refrigerator oils (ISO VG 68, ISO VG 100) are also used for ammonia refrigerant systems.

In the past, ammonia system lubricants used relatively naphthenic mineral oils, alkylbenzenes or PAO. The problems of oil enrichment and oil deposits in such systems are well known. Polyalkylene glycols on the other hand, display excellent solubility in ammonia refrigerants. Such good solubility will lead to new system designs in the future including very low temperatures and dry evaporation. The carefully selected synthetic components display excellent $V-T$ behavior and high thermal stability. The water content of polyglycol oils must be kept low (about 300 – 500 ppm). Mixing or contamination with mineral oil must be avoided (formation of reaction products!) [266].

Other Synthetic Fluids. In the past, polysilicic acid, ester-based synthetic fluids were used for evaporation temperatures below – 120 °C. Products based on low-viscosity silicone oils (polydimethylsiloxane; PDMS) are also used.

Refrigerator Oils for CO_2. The natural refrigerant CO_2 is gaining acceptance among users. The properties of CO_2–oil mixtures are not fully known. In principle, the solubility of CO_2 in oil increases in the order

MO > AB > PAG > POE, i.e., at the same temperature and pressure, the most CO_2 is dissolved in ester oils and the least in mineral oils. Depending on the type of oil, carbon dioxide causes varying degrees of viscosity drops. In terms of mass %, viscosity decreases in the order AB > MO > POE > PAG. According to this, the viscosity of polyglycol- based oils fall the least. Suitable CO_2 refrigerator oils are currently being developed. Apart from solubility, thermal stability is also an important factor (outlet temperatures approx. 160 – 180 °C) [267], [268].

10.2.4. Viscosity Selection

The guidelines for the selection of refrigerator compressor lubricants are much the same as those applied to engineering in general, i.e., fast running machines permit the use of lower viscosities than slow running machinery. High bearing loads require higher viscosities than low bearing loads. In addition, refrigerator compressors require much lower viscosities than those calculated from the hydrodynamic lubrication theory.

In the past, many refrigerator systems were operated with chlorinated refrigerants. Chlorine compounds are excellent EP additives which protect against wear. Since the introduction of chlorine-free refrigerants, this function must be performed by the refrigerator oil.

The most important parameter for determining the lubricity of oils or oil – refrigerant mixtures is viscosity.

Mixture Concentration in Relationship to Temperature and Pressure. Figure 53 shows how much refrigerant is dissolved in a refrigerator oil when saturated at a defined operating condition (pressure, temperature). As saturation is time-related, the refrigerant concentration shown by the diagram is, as a rule, greater than the actual value. It can be viewed as the maximum concentration at any given operating condition. The viscosity which can be taken from the mixture concentration adds a safety margin to any bearing loading calculations.

Mixture Viscosity and Density in Relationship to Temperature, Pressure and Refrigerant Concentration. The precise refrigerant concentration in a system is given at a defined pressure and temperature as shown in Figure 53. Figure 54 can be used to read-off the kinematic viscosity of the oil – refrigerant mixture at a defined temperature, defined pressure and defined refrigerant concentration.

Care should be taken that the oil does not reach its maximum viscosity at problematic points in the oil return circuit (e.g., upward flows). It is also important that the conditions in the compressor crankcase do not approach those indicated by the falling left-hand part of the $V-T$ graph because the smallest temperature fluctuations can have a significant effect on viscosity in such conditions [261].

The density of an oil refrigerant mixture depends on the density – temperature behavior of the oil and the refrigerant.

Miscibility Gap, Solubility Threshold. R-refrigerants are among the most oil-soluble refrigerants. However, not all are miscible in refrigerator oils at all temperatures and in all concentrations. If, for example, a fully homogeneous oil – refrigerant mixture is cooled, a point is reached when the mixture separates into two phases. The size of the miscibility gap depends on the type of refrigerant and also to a large extent on the type of refrigerator oil. Refrigerant solubility is defined statically in E DIN 51 514. For normal applications, the miscibility gap (with alkylbenzenes) is not a problem for a series of refrigerants such as R 22. Some other refrigerants display pronounced miscibility thresholds. The miscibility gap is of great importance to the refrigerant circuit. If the oil – refrigerant ratio is within the miscibility gap, problems can occur caused by oil-enriched phases depositing in collectors, condensers, evaporators and the crankcase. Flooded evaporators require the largest possible quantity of refrigerant to be dissolved at evaporating temperatures without the phases separating.

Table 33. DIN 51 515 classification of turbine oils, Draft 1999

	'Normal' turbine oils – steam turbine oils	'High-temperature' turbine oils
Without EP	DIN 51 515-1	DIN 51 515-2
With EP (FZG, Load Stage min. 8)	DIN 51 515-1 Annex A	DIN 51 515-2 Annex A

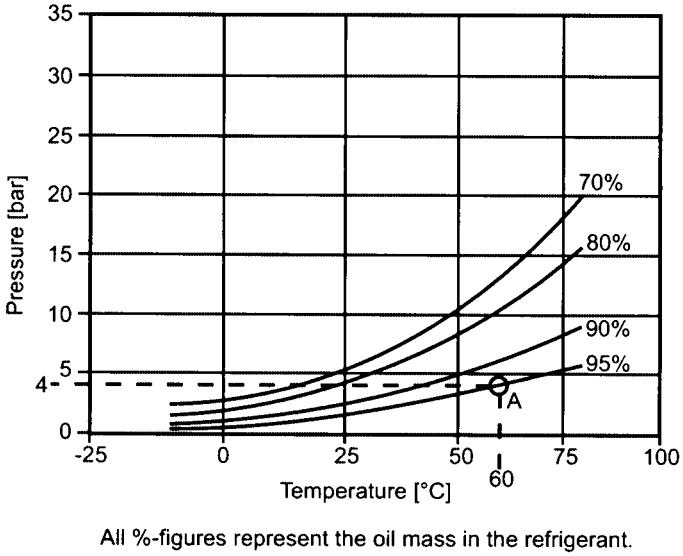

Figure 53. Mixture concentration in relationship to temperature and pressure (RENSIO Triton SE 55 – R 134a)

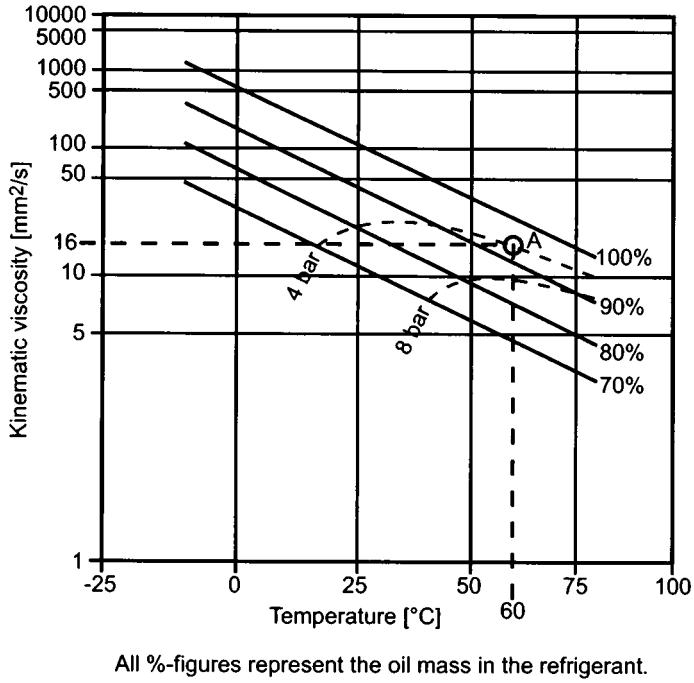

Figure 54. Mixture viscosity in relationship to temperature, pressure, and refrigerant concentration (RENSIO Triton SE 55 – R 134a)

Table 34. Requirements of steam turbine oils, DIN 51 515 Part 1, Draft 1999

Test	Limits				Testing according to[b]	Comparable ISO standards
Lubricating oil group	TD 32	TD 46	TD 68	TD 100		
ISO viscosity grade[a]	ISO VG 32	ISO VG 46	ISO VG 68	ISO VG 100	DIN 51 519	ISO 3448
Kinematic viscosity at 40 °C in mm^2/s min. max.	28.8 35.2	41.4 50.6	61.2 74.8	90.0 110	DIN 51 562-1 or DIN 51 562-2 or DIN EN ISO 3104	ISO 3104
Flash point (COC), °C min.	160	185	205	215	DIN ISO 2592	ISO 2592
Air release property[d] at 50 °C, minutes max.	5	5	6	no value specified	DIN 51 381	–
Density at 15 °C, g/mL max.	to be specified by supplier				DIN 51 757 or DIN EN ISO 3675	ISO 3675
Pour point, °C	≤ -6	≤ -6	≤ -6	≤ -6	DIN ISO 3016	ISO 3016
Neutralization number, mg KOH/g	to be specified by supplier				DIN 51 558 Part 1	ISO 6618
Ash (oxide ash), mass %	to be specified by supplier				DIN EN ISO 6245	ISO 6245
Water content, mg/kg max.	150	150	150	150	DIN 51 777-1	ISO/DIS 12937
Cleanliness level, min.	20/17/14	20/17/14	20/17/14	20/17/14	DIN ISO 5884 with DIN ISO 4406	ISO 5884 with ISO 4406
Water separation ability (after steam treatment), max. (seconds)	300	300	300	300	DIN 51 589 Part 1	–
Copper corrosion, max. corrosiveness (3 h at 100 °C)	2-100 A 3	2-100 A 3	2-100 A 3	2-100 A 3	DIN EN ISO 2160	ISO 2160
Corrosion protection against steel corrosiveness, max.	no rust 0 - A				DIN 51 585	ISO 7120
Aging behavior[c] (TOST)						
Time in hours to reach delta NZ of 2.0 mg KOH/g	2000	2000	1500	1000	DIN 51 587	ISO 4263
Foam:						ISO 6247
Seq. I at 24 °C, max. mL	450/0	450/0	450/0	450/0	–	
Seq. II at 93 °C, max. mL	100/0	100/0	100/0	100/0	–	
Seq. III at 24 °C after 93 °C, max. mL	450/0	450/0	450/0	450/0	–	

[a] Middle point viscosity at 40 °C in mm^2/s.
[b] The oil sample has to be stored without any contact to light before doing the test.
[c] The aging behavior test has to be done as a type test procedure because of the long testing time. TOST = Turbine Oxidation Stability Test.
[d] The test at a temperature of 25 °C has to be specified by the supplier when the customer needs values at low temperatures.
Annex A (normative) – for EP turbine oils.
If the turbine oil supplies also a turbine gear set, the oil has to reach a load stage of min. 8 according to DIN 51 345, part 1 and part 2 (FZG).

11. Turbine Oils

11.1. Demands on Turbine Oils – Characteristics

The demands on turbine oils are defined by the turbines themselves and their specific operating conditions. The oil in the lubricating and control circuits of steam and gas turbines has to fulfil the objectives [271], [272]:

- hydrodynamic lubrication of all bearings and the lubrication of gearboxes
- heat dissipation
- functional fluid for control and safety circuits
- avoidance of friction and wear on gear tooth flanks in turbine gearboxes and when the turbine is spooled-up

Apart from these mechanical – dynamic requirements, the following physical – chemical specifications also have to be fulfilled by turbine oils [271], [272]:

aging stability for long operating periods
hydrolytic stability (especially of the additives used)
corrosion protection even if water, steam and/or condensation is present
reliable separation of water vapor
rapid air release and low foaming
good filterability and purity

Table 35. Requirements of high-temperature turbine oils, DIN 51 515 Part 2, Draft 1999

Test	Limits		Testing according to[b]	Comparable ISO standards
Lubricating oil group	TG 32	TG 46		
ISO viscosity grade[a]	ISO VG 32	ISO VG 46	DIN 51 519	ISO 3448
Kinematic viscosity at 40 °C in mm^2/s				
min.	28.8	41.4	DIN 51 550 in conjunction	
max.	35.2	50.6	with DIN 51561 or DIN 51562-1	ISO 3104
Flash point (COC), °C, min.	180	200	DIN ISO 2592	ISO 2592
Air release property[d] at 50 °C, max. minutes	5	5	DIN 51 381	–
Density at 15 °C, g/mL, max.	to be specified by supplier		DIN 51 757	ISO 3675
Pour point, °C	≤ -6		DIN ISO 3016	ISO 3016
Neutralization number, mg KOH/g	to be specified by supplier		DIN 51 558-1	ISO/DIS 6618
Ash (oxide ash), mass %	to be specified by supplier		DIN EN 7	ISO 6245
Water content, mg/kg, max.	150	150	DIN 51777-1	ISO/DIS 12937
Cleanliness level, min.	17/14	17/14	DIN ISO 5884 with ISO 4406	ISO 5884 with ISO 4406 ISO 6247
Foam:				
Seq. I at 24 °C, max. mL	450/0	450/0	–	
Seq. II at 93 °C, max. mL	100/0	100/0	–	
Seq. III at 24 °C after 93 °C, max. mL	450/0	450/0	–	
Demulsibility, minutes	to be specified by supplier		DIN 51 599	ASTM-D 1401
Water separation ability (after steam treatment), seconds max.	300	300	DIN 51 589 Part 1	–
Copper corrosion corrosiveness, max.	2−125 A 3	2−125 A 3	DIN 51 759	ISO 2160
Corrosion protection against steel corrosiveness, max.	0 – A 5	0 – A 5	DIN 51 585	ISO/DIS 7120
Aging behavior[c]				
Time in hours to reach delta NZ of 2.0 mg KOH/g, min.	3.000	3.000	DIN 51 587	ISO/DIS 4263
RBOT, minutes	$\geq 800*$	$\geq 800*$	–	
Modified RBOT, % of time, minutes, in unmodified test	$\geq 85\,\%$	$\geq 85\,\%$	– –	ASTM-D 2272

∗ GE recommends only 450 min.
[a] Middle point viscosity at 40 °C in mm^2/s.
[b] The oil sample has to be stored without any contact to light before doing the test.
[c] The aging behavior test has to be done as a type test procedure because of the long testing time.
[d] The test at a temperature of 25 °C has to be specified by the supplier when the customer needs values at low temperatures.
Annex A (normative) – for EP turbine oils
If the turbine oil supplies also a turbine gear set, the oil has to reach a load stage of min. 8 according to DIN 51 345, part 1 and part 2 (FZG).

These stringent demands on steam and gas turbine oils are met with carefully selected base oils and the inclusion of special additives.

11.2. Formulation

Turbine oils used today contain special paraffinic base oils with good $V-T$ characteristics, antioxidants, and corrosion inhibitors. If geared turbines require a degree of load-bearing capacity (e.g., failure load stage FZG 9 – DIN 51 345), some mild EP additives are included.

These days, turbine base oils are created exclusively by extraction and hydration. Refining and subsequent selective high-pressure hydrogenation significantly determine and influence turbine oil characteristics such as oxidation stability, water separation, air release and foaming. This applies in particular to water separation and air release because these feature cannot be subsequently improved with additives. Steam turbine oils are generally made of special paraffinic base oil cuts [270–272].

A combination of phenolic and aminic antioxidants (synergies) are added to turbine oils to improve their oxidation stability [271]. To improve corrosion protection, non-emulsifiable corrosion protection agents and nonferrous metal passivators are used. These are not negatively affected by water or steam contamination during operation and remain suspended. If

standard turbine oils are used in geared turbines, mild, long life, temperature and oxidation stable EP additives (organophosphorus and/or sulfur compounds) are included. Furthermore, small quantities of silicone-free antifoams and pour point depressants are also used in turbine oils [271–273].

Moreover, these additives must not detrimentally affect the air release (very sensitive) of the oil. All additives should be free of ash (e.g., zinc-free). The purity of a turbine oil according to ISO 4406 should be 15/12 in the tank [274–277]. No circuits, wires, cables or insulation containing silicone should come into contact with turbine oils.

11.3. Specifications

Special paraffinic mineral oils with additives are normally used for gas and steam turbine lubricants. These serve to protect all shaft bearings in the turbine and generator as well as the gearbox in corresponding configurations. They can also be used as hydraulic fluids in control and safety systems. Fire-resistant, synthetic Type HFD-R fluids [272] are normally used in hydraulic systems which operate at pressures around 40 bar, if there are separate lubricating oil and control oil circuits (twin circuit systems).

DIN 51 515 (Lubricants and Control Fluids for Steam Turbines, issued 1986) is presently being revised. Following ISO 6743, Part 5, Family T (Turbines) – Classification of Turbine Oils [278], the latest draft of DIN 51 515, dated 1999, classifies turbine oils as in Table 33.

Table 36. ISO 6743-5 classification of turbine lubricants, Draft 1999

	Normal turbine oils	High-temperature turbine oils
Without EP	ISO-L – TSA (steam) ISO-L – TGA (gas)	ISO-L – TGB (gas)
With EP	ISO-L – TSE (steam) ISO-L – TGE (gas)	??*

* Under discussion.

The demands made by DIN 51 515-1 – Steam Turbine Oils [278] and DIN 51 515-2 – High-Temperature Turbine Oils [278] are shown in Tables 34 and 35.

ISO 6743-5 classifies turbine lubricants according to whether they are used in steam or gas turbines and if they contain EP agents (Table 36) [279], [280].

Synthetic fluids such as PAO and phosphoric acid ester are also listed (see Table 37) [279], [280].

Table 38 [274–277] shows the principal demands made by the world's leading turbine manufacturers. The main differences between these demands relate to the aging and/or oxidation tests.

11.4. Turbine Oil Circuits

The oil circuits play an especially important role in the lubrication of power station turbines. Steam turbines are normally fitted with pressurized oil circuits and separate lubricating and control circuits as well as separate lubricating and control oil tanks are used.

In normal operating conditions, the turbine shaft-driven main oil pump draws oil from the tank into the control and bearing supply circuits. Pressure in control circuits is normally between 10 and 40 bar [269–272]. Oil tank temperatures range from 40 – 60 °C. The velocity of the oil in the feed circuit is about 1.5 – 4.5 m/s [271]. Cooled and passing through reduction valves, the oil reaches the turbine, generator and possible gearbox bearings at a pressure of ca. 1 – 3 bar. The individual oil feeds return to the oil tank at atmospheric pressure. The lubricating oil circuit of a gas turbine is largely similar to that of a steam turbine.

Larger oil circuits are fitted with centrifugal by-pass filtering systems. These ensure that the finest of contaminants are removed along with aging byproducts and sludge.

Oil aging is also influenced by the frequency with which the oil is pumped through the circuit. If the oil is pumped too fast, excessive amounts of air are either dispersed or dissolved (problem: cavitation in bearings, premature aging etc.). Oil tank foaming can also occur but this generally collapses rapidly.

11.5. Monitoring and Maintenance of Turbine Oils

In normal circumstances, oil monitoring intervals of one year are perfectly acceptable. [270],

Table 37. Classification of lubricants for turbines, ISO 6743-5, Draft 1999.

Part 5: Family T (Turbines): Table – Classification of lubricants for turbines				
	General application	Composition and properties	Symbol ISO-L	Typical application
1)	Steam turbines, directly coupled or geared to the load, normal service	Highly refined petroleum oil with rust protection and oxidation stability	TSA	Power generation and industrial drives and their associated control systems, marine drives, where improved load carrying capacity is not required for the gearing
2)	Gas turbines, directly coupled or geared to the load, normal service		TGA	
3)	Steam turbines, directly coupled or geared to the load, high load-carrying capacity	Highly refined petroleum oil with rust protection, oxidation stability and enhanced load carrying capacity	TSE	Power generation and industrial drives and marine geared drives and their associated control systems where the gearing requires improved load-carrying capacity.
4)	Gas turbines, directly coupled or geared to the load, high load-carrying capacity		TGE	
5)	Gas turbines, directly coupled or geared to the load, higher-temperature service	Highly refined petroleum oil with rust protection and improved oxidation stability	TGB	Power generation and industrial drives and their associated control systems where high temperature resistance is required due to hot spot temperatures
6)	Other lubricants a) TSC – Synthetic steam turbine fluids with no specific fire-resistant properties (e.g., PAO) b) TSD – Synthetic steam turbine fluids based on phosphate esters with fire-resistant properties c) TGC – Synthetic gas turbine fluids with no specific fire-resistant properties (e.g., PAO) d) TGD – Synthetic gas turbine fluids based on phosphate esters with fire-resistant properties e) TCD – Synthetic fluids for control systems based on phosphate esters with fire-resistant properties			

[272], [274]. As a rule, these should be performed in the oil manufacturer's laboratories. In addition, a weekly visual inspection of the oil should be performed to spot contamination and impurities in the oil in good time. Filtering the oil with a centrifuge in a by-pass circuit is a reliable method.

11.6. Life of (Steam) Turbine Oils

Oil life of 100 000 h is not uncommon in large steam turbines [270], [271]. However, the antioxidant level in the oil can fall to about 20 – 40 % of the fresh oil (oxidation, aging). The life of turbine oil depends heavily on operating conditions such as temperature and pressure, oil circulation speed, filtering and the quality of maintenance and finally, the amount of oil topped-up (this helps maintain adequate additive levels).

The temperature of the oil in a turbine depends on the bearing loading, speed, bearing dimensions and the oil's flow rate. Radiated heat can also be an important parameter. The oil circulation factor, i.e., the ratio between flow volume per hour and tank volume should be between 8 and $12\,h^{-1}$ [271–273]. Such relatively low oil circulation factors ensure that gaseous, fluid and solid impurities can be efficiently separated while air and other gases can be released. Furthermore, low oil circulation factors reduce the thermal loads on an oil (with mineral oils, oxidation speed doubles when the temperature increases by 8 – 10 K).

11.7. Gas Turbine Oils – Application and Requirements

Gas turbine oils are used in stationary turbines. These produce either electricity or heat. The compressor fans generate pressures of up to 30 bar which vent into the combustion chambers where gas is injected [271]. Depending on the type involved, combustion temperatures of up to 1000 °C are reached (generally 800 – 900 °C) [271], [281]. Exhaust gas temperatures can reach about 400 – 500 °C.

Common mineral oil-based steam turbine oils are used for conventional gas turbines. However, since the temperature of some bearings in gas turbines is higher than in steam turbines, premature oil aging can be expected. Moreover, hot-spots can occur around some turbine bearings and localized temperatures can reach 200 – 280 °C [271] whereby the temperature of the oil in the tank remains at about 70 – 90 °C.

Table 38. Principal demands of turbine manufacturers

	Siemens TLV 901304 (Sept. 1996)	General Electric GEK 101941 A (Nov. 1999)	General Electric GEK 32568 E (May 1999)	ABB HTGD 90117 V0001Q	ABB HTGD 90117 V0001Q	Testing according to DIN ISO	Testing according to ASTM
Lubricating oil group	Steam and gas turbine oils[a]	Gas turbine oils with EP/AW additives with bearing ambients above 260 °C[b]	Gas turbine oils with bearing ambients above 260 °C[c]	Steam and gas turbine oils (with and without EP/AW additives), ISO VG 32/46[d]	Steam and gas turbine oils (with and without EP/AW additives), ISO VG 68[d]		
Kinematic viscosity (at 40 °C), mm^2(s						DIN 51 562-1	ASTM-D 445
ISO VG 32	28.8 – 35.2	28.8 – 35.2	28.8 – 35.2	VG 32: ± 10 %	VG 68: ± 10 %		
ISO VG 46	41.4 – 50.6			VG 46: ± 10 %			
Gravity, °API	–	29 – 33.5	29 – 33.5	–	–		ASTM-D 287
Air release property (at 50 °C), min	≤ 4	5 (max.)	5 (max.)	≤ 4	≤ 7	DIN 51 381	ASTM-D 3427
Neutralization Number, mg KOH/g						DIN 51 558-1	ASTM-D 974
without EP/AW additives	≤ 0.2	–	0.2 max.	0.3 max.	0.3 max.		
with EP/AW additives	≤ 0.3	0.2 max.	–	0.3 max.	0.3 max.		
Water content, mg/kg	≤ 100	–	–	–	–	DIN 51 777-1	ASTM-D 1744
Foaming (at 25 °C)		50/0 max.	50/0 max.	450/0	450/0	–	ASTM-D 892
Tendency, mL	≤ 400 (Seq. 1)	50/0 max.	50/0 max.	50/0	50/0		
Stability, s	≤ 450	50/0 max.	50/0 max.	450/0	450/0		
Water separability, s	≤ 300			≤ 300	≤ 300	DIN 51 589-1	–
Demulsibility, min	≤ 20			≤ 30	≤ 30	DIN 51 599	ASTM-D 1401
Density (at 15 °C), kg/m^3	≤ 900			≤ 900	≤ 900	DIN 51 757	ASTM-D 1298
Flash point				VG 32 and 46:	VG 68:	DIN ISO 2592	ASTM-D 92
ISO VG 32, °C	> 160	215 (min.)	215 (min.)	≥ 185	≥ 205		
ISO VG 46, °C	> 185						
Pour point, °C	≤ −6	−12 (max.)	−12 (max.)	< −6	< −6	ISO 3016	ASTM-D 97
Particle distribution (ISO class)	≤ 17/14			18/15	18/15	ISO 4406	–
Color	≤ 2	max. 2.0	max. 2.0	–	–	DIN ISO 2049	ASTM-D 1500
Copper corrosion, corrosiveness	≤ 2−100 A 3	1 B (max.)	1 B (max.)	≤ 2−100 A 3	≤ 2−100 A 3	DIN EN ISO 2160	ASTM-D 130

Table 38. (Continued)

	Siemens TLV 901304 (Sept. 1996)	General Electric GEK 101941 A (Nov. 1999)	General Electric GEK 32568 E (May 1999)	ABB HTGD 90117 V0001Q	ABB HTGD 90117 V0001Q	Testing according to DIN ISO	Testing according to ASTM
corrosion protection against steel, corrosiveness	$\leq$ 0 – b	0 – b (pass)	0 – b (pass)	free from corrosion when using artificial seawater		din 51 585	astm-d 665
aging behavior, increase of neutralization number in mg koh/g after h, tost test	$\leq$ 2.0 (after 2.500 h)	$\leq$ 2.0 (after 3.000 h)	$\leq$ 2.0 (after min. 3.000 h)	$<$ 2.0 (after 2.000 h) $*$	$<$ 2.0 (after 2.000 h) $*$	din 51 587	astm-d 943
additional requirements on turbine oils for use in gearboxes				without gear:[e] vg 32: 5, l.s. vg 46: 6, l.s.	without gear:[e] vg 68: 7, l.s.	din 51 354	astm-d 1947
FZG-Test: A/8,3/90 Failure load stage	$\geq$ 8	$\geq$ 8	–	with gear: VG 32: 9, L.S. VG 46: 9, L.S.	with gear: VG 68: 9, L.S.		
Carbon residue Ramsbottom, %		0.1 % (max.) (or equivalent)	0.1 % (max.) (or equivalent)	–	–	–	ASTM-D 524
Oxidation stability by rotating bomb, minutes		500 (min.)	500 (min.)	$>$ 300 min.	$>$ 300 min.	–	ASTM-D 2272
Oxidation stability by rotating bomb (modified) N2-blown RBOT		85 % (min.) of time in unmodified test	85 % (min.) of time in unmodified test	–	–	–	ASTM-D 2272
VI		95 (min.)	95 (min.)	$\geq$ 90	$\geq$ 90	–	ASTM-D 2270
Atomic emission spectroscopy							ASTM-D 4951
Zinc content		$<$ 5 ppm		$\leq$ 2 ppm	$\leq$ 2 ppm	–	

$*$ Neutralization number $<$ 1.8 mg KOH/g, sludge $<$ 0.4 % according to DP 7624.

Particle distribution (cleanliness levels) according to ISO 4406: between 16/13 and 15/12 depending on the manufacturer.

Base Oils:

[a] Mineral oils or synthetic oils with additives to improve corrosion protection and aging stability (additionally EP/AW additives in the case of gearbox lubrication). PAO accepted for steam turbines, PAO not accepted for gas turbines.

[b] Petroleum lubricating oil: Synthetic Hydrocarbons with greater high temperature, oxidation stability and R&O inhibitors, EP/AW additives.

[c] Petroleum lubricating oil: Synthetic Hydrocarbons with greater high temperature, oxidation stability and R&O inhibitors.

[d] Refined mineral oils with aging inhibitors and corrosion protection additives (no EP/AW additives).

Other important specifications (examples):

Westinghouse: I.L. 1250-5312 – Steam Turbines; 21 T 0591 – Gas Turbines; Solar: ES 9-224 – Gas Turbines.

[e] L.S. = Load Stage.

The temperature of the oil reaching a bearing is mostly between 50 – 55 °C and the exit temperature about 70 – 75 °C [271]. The volume of oil for a 40 – 60 MW generator (GE) is about 6000 – 7000 L and its life is between 20 000 and 30 000 h (in the case of a 40 – 60 MW Siemens, 14 000 L and 40 000 – 80 000 h [274], [277]. Semisynthetic turbine oils (multiple hydrated base oils) or fully synthetic turbine oils based on synthetic PAO's are recommended for these applications [271], [272], [276].

In civil and military aviation, gas turbines are used for propulsion. Because of the high temperatures encountered, special, low-viscosity (ISO VG 10, 22) synthetic oils based on saturated esters (e.g., neopolyesters) are used in these aircraft engines or turbines [281]. Aviation turbine oils containing special naphthenic base oils (ISO VG 15 – 32) with good low-temperature characteristics are also used [281].

11.8. Fire-Resistant, Water-Free Fluids for Power Station Applications

For safety reasons, fire-resistant fluids are used in control and governor circuits which are exposed to ignition and fire hazards. In power stations, this applies in particular to hydraulic systems in high temperature zones such as near to superheated steam pipes. The fire-resistant hydraulic fluids used in power stations are generally water-free, synthetic fluids based on phosphoric acid esters (type DFD-R HFD-R according to DIN 51 502 or ISO 6743-0, ISO VG 32–68) and offer the following features [272]:

- fire-resistance
- self-ignition temperature over 500 °C
- auto-oxidation stable to surface temperatures up to 300 °C
- good lubricity
- good protection against corrosion and wear
- good aging stability
- good demulsification
- low foaming
- good air release and low vapor pressure

Additives to improve oxidation stability (possibly foam inhibitors) as well as rust and corrosion inhibitors are sometimes used. According to the 7th Luxembourg Report, the maximum permissible temperature of HFD fluids in hydrodynamic systems is 150 °C. Continuous temperatures of 80 °C should not be exceeded in hydraulic systems. These synthetic fluids are generally used for control circuits, but in some special cases, also for the lubrication of plain bearings in turbines as well as other hydraulic circuits in steam and gas turbine installations. However, these systems must be designed for these fluids (HFDR-compatible elastomers, paint finishes and coatings). DIN 51 518 lists the minimum requirements which power station control circuit fluids have to fulfil. Further information can be found in guidelines and specifications relating to fire-resistant fluids.

11.9. Lubricants for Water Turbines and Hydroelectric Plants

Operators of hydroelectric power stations have to pay particular attention to the handling of water-polluting substances, e.g., the lubricants employed. The specific operating conditions of the hydroelectric plant need to be considered when selecting lubricants. The lubricants must display good water and air release, low foaming, good corrosion protection, FZG Wear protection > 12 in gearboxes, good aging resistance and compatibility with standard elastomers [272]. As there are no established standards for water turbine oils, the existing product specifications for general turbine oils are adopted as basic requirements. The viscosity of water turbine oils depends on the type and design of the turbine as well as its operating temperature and can range from 46 to 460 mm^2/s at 40 °C. Type TD and LTD lubricating and control oils according to DIN 51 515 are used. In most cases, the same oil can be used for bearings, gearboxes and control equipment. In many cases, the viscosity of these turbine and bearing oils is between 68 and 100 mm^2/s. Water turbine oils are subject to little thermal stress, and as oil tank volumes tend to be high, the life of water turbine oils is very long. In hydroelectric power stations, the oil sampling and analysis intervals can be correspondingly long. Particular care should be taken when sealing the turbine's lubricating oil circuit from possible water ingress. In recent years, rapidly biodegradable water turbine oils based on saturated esters have proven successful

in practice. In addition, Type HLP 46 hydraulic oils, Type HEES 46 rapidly biodegradable fluids and NLGI grade 2 and 3 greases are used in hydroelectric plants [272].

12. Metalworking Fluids [51]

The great flexibility and versatility of the different types of machining method are particularly significant in the metalworking industries. Although for some years now there has been a growing trend towards noncutting (forming) methods for workpiece quality reasons and to save material and process costs, this has not led to a recognizable decrease in the volume share of lubricants. This is also apparent from the machine tool statistics. The dramatic change predicted in the 1970s did not take place in the 1980s and 1990s.

Because of the particular significance of cooling for the cutting operation, this process is called cooling lubrication and the fluids used are called coolants. Apart from this term 'coolant' which is commonly used in general practice, there are also numerous other terms for specific applications such as, for example, cutting oils, grinding oils, reaming oils, deep hole drilling oils and honing oils.

The advantages of cutting fluids can be summarized briefly as follows: accelerated heat dissipation with increased tool service life to make higher cutting speeds possible; lubrication between tool, chip and workpiece with reduced tool wear and improvement of the quality of the workpiece surface finish; lubrication of sliding points outside the actual cutting zone between tool, workpiece and chips; and improved chip removal. Frequently less attention is paid to lubrication outside the tool – chip contact zone which, nevertheless, can be very important. Examples are margins of twist drill and reamers as well as the support and guide rails on deep hole drills and honing tools.

12.1. Mechanism of Action

When machining by cutting, a very reactive 'clean' workpiece material surface is continuously generated which can tend to react adhesively on the cutting surface of the tool in the workpiece material – tool contact zone. It must be assumed that this reactive surface not only tries to saturate the free valances of the tool material but that other available substances are also involved which can be bonded by physical or chemical absorption, or chemical bonding. The oxygen in the air plays a particular role. In numerous tests under vacuum it has been established that the tool wear is considerably reduced through the saturation of the surface opened up by cutting, compared to where a gas or gas mixture has been available at different partial pressures.

Coolant penetration into the contact surface [282] is via a network of capillaries which is linked to each other. The width of the capillary diameters is 10^{-3} to 10^{-4} mm.

As far as the saturation process of the newly formed surface is concerned, the surface is never completely covered by lubrication effective molecules; there is a clear gradient of surface activity in the direction of the tool tip, with little reaction at the outer range of the tool. High normal pressure also prevents lubricant transport when machining materials which are difficult to cut. The assessment of cutting tests with model substances has shown that 30 % saturation of the contact surface between chip and tool lead to a reduction in friction force by 75 % [282].

Cooling. In view of the findings relating to cooling and lubrication, it is impossible to differentiate the two. The changes in material properties caused by temperature are closely associated with the effect of friction. Cooling gains increasing significance for tool wear when the maximum cutting temperature approaches the softening point of the tool material.

On the one hand the cooling effect of the cutting fluid and the heat dissipated depends on its thermal properties, especially the heat capacity and the heat transfer coefficient; on the other hand the flow conditions and the heat transfer coefficient play a significant role. The heat transfer coefficient can be influenced considerably by the substances active at the interface and by evaporation.

The high specific heat and high heat transfer coefficient of water give water-miscible cutting fluids more favorable cooling properties than non-water-miscible oils.

12.2. Water-Miscible Cutting Fluids

By volume this is the most significant group of metalworking fluids in the USA and Europe (using as a reference the amount of concentrate from which the users produce the water-mixed finished products). In the Far East, especially in Japan, non-water-miscible products clearly outweigh water-miscible products. Here the trend to increase the share of water-miscible products is even more evident, whereas in Germany, a trend towards non-water-miscible products is more apparent, not least because of the stringent legislation covering storage and handling of lubricants of all kinds.

The question of the cutting performance of water-miscible cutting fluids, which has considerable significance, is not generally of primary importance in their selection, as is the case with non-water-miscible cutting fluids. The secondary demands put on these products are frequently so much the focal point of discussions that their actual main tasks, such as cooling and lubricating, are overshadowed. Of ever-increasing importance are the problems of health protection in the workplace, microbiology, hygiene and the disposal of used fluids.

12.2.1. Composition

The main ingredients of water-miscible can be summarized in the following groups [283].

1) Mineral oil hydrocarbons
2) Synthetic hydrocarbons, synthetic esters, fatty oils
3) Emulsifiers (alkali and alkanolamine soaps, naphthenates, sulfonates, sulfates, ethoxylates, sorbitan estes)
4) Corrosion inhibitors (alkali and alkanolamine soaps, naphthenates, amines, amides, boron compounds)
5) Stabilizers, coupling agents (glycols, alcohols)
6) Extreme pressure additives
7) Antiwear additives
8) Lubricating film improvers
9) Antifoam agents
10) Microbiocides
11) Complexing agents (EDTA derivatives)

Even if mineral oils are the major constituents, then part or even the total mineral oil content can be replaced by synthetic hydrocarbons such as polyalphaolefins or alkyl benzenes by fatty oils or even synthetic esters. In many cases the use of synthetic hydrocarbon is restricted to specific areas of use but synthetic esters or natural fat oils are in regular use.

Basically, suitable water-miscible cutting fluids can be produced both from paraffinic and naphthenic oils; naphthenic oils, however, have a higher solubility for additives and more favorable emulsifying properties. However, the use of naphthenic oils is declining because of their higher aromatic content compared with paraffin ic oils. This is considered critical, particularly with regard to health and safety at work.

12.2.2. Corrosion Protection and Corrosion Test Methods

Normally the corrosion protection afforded by a water-mixed cutting fluid is reduced as the degree of dilution increases. The quality of a product, as far as its corrosion protection properties are concerned, is, therefore, most frequently defined in laboratory application according to the concentration which still gives no corrosion or achieves a specific degree of corrosion. The most significant test methods are described in DIN 51 360, Part 1 (Herbert corrosion test) or IP 125 as well as in DIN 51 360, part 2 (chips/filter paper method). Similar testing methods are known under ASTM D-4627 and IP 287 whereas the French standard NF T 60-188 is more similar to the Herbert test.

12.2.3. Concentration of Water-Mixed Cutting Fluids

The possibility to produce water-mixed products from water-miscible coolants with varying concentrations for adaptation to different cutting processes gives great flexibility. Normally, water-mixed coolants with concentrations up to approx. 30 % can be applied without any particular problems. However, if this concentration range is exceeded then the stability has to be assessed, especially in the case of conventional emulsions. When working with hard water, lime soaps can arise when working far below this

concentration, which could lead to considerable irregularities in the course of machining; this depends to a great extent on the specific composition of the coolants. Often, the lower concentration limit is determined by the corrosion protection properties. In this respect, attention is drawn to the fact that concentrations below 1.5 % can only be checked with difficulty during operation because measurement of the concentration is a problem. Reducing the concentration by only a few tenths of a percentage point can already lead to a serious corrosion effect. In the case of water-miscible coolants, which contain microbiocides, the skin compatibility of these products must be observed when using particularly high concentrations; this also applies for other additives. Normally, cutting fluids are very often used for rough machining with a concentration of between 2 and 5 %, but very frequently at 4 %. In the case of special applications such as deep hole drilling, reaming or generally for particularly difficult cutting operations the use of considerably higher concentrations is possible.

Frequently a concentration of between 1.5 and 2.5 % is selected for grinding operations which are to be carried out with water-mixed coolants. In the case of high performance grinding (grinding with high chip volume) higher concentrations are also sometimes used.

12.2.4. Stability of Coolants

Stability is the most important property for the service life of a coolant. Apart from the general terms for stability, in coolant practice stability is also defined according to the different stresses experienced during service. For example, electrolytic stability, thermal and boiling stability and frequently also stability against microorganisms.

12.2.5. Foaming Properties

In all metalworking sectors as well as in the metal cutting field, the formation of stable foam on the processing fluids is an undesirable phenomenon. Especially coolant containing surfactants, and these include all emulsions, tend, to a greater or lesser extent, to develop surface foam, under appropriate agitation.

To prevent foaming, the selection of a lower foaming coolant is the first of the measures to be taken. Very low foaming products are especially available amongst the more conventional emulsions. One reason for this is that the surfactant molecules (in this case essentially emulsifier molecules and anionic corrosion inhibitors) are mainly bound on the oil – water interface, that is to say to the oil drop, and are not available to be sited on the foam lamellae.

The effect of the most significant group of defoamers for coolants, the silicones, is based on their particularly low surface tension and their insolubility in both oil and water.

12.2.6. Preservation of Coolants with Biocides

Biocides are used to reduce or avoid heavy microbial activity. Biocides should fulfill the following requirements:

- Effectiveness against as many types of microorganisms as possible (broad spectrum effective).
- Fast action: this has particular significance in the case of systems with high drag-out rates; products which act slowly are taken away with the coolant before they become effective.
- Long continuous effect: the biocide should not lose its activity through reaction with other components or contaminants.
- High thermal stability.
- Low influence on pH value of the coolant: different ways of assessing this exist. In many cases an alkaline biocide is desirable. In the case of coolants in use, which have clearly suffered a drop in pH value through microbial attack, it is expedient to use highly alkaline biocides for correcting purposes. Over and above this, alkaline biocides quickly form an essential part of the reserve alkalinity of coolants. A high alkali reserve at pH < 9 is generally favorable.
- Oil solubility: in many cases use in coolant concentrates (water-miscible coolant) is a basic requirement. The biocide is either dissolved or homogeneously dispersed in the concentrate. The use of biocides without solubility in the concentrate is restricted to the treatment of water-mixed coolants.

- No odor nuisance: no odor nuisance in workshops, both with regard to the biocide's own odor and by generation of odor by a reaction with microorganisms or components.
- Good skin compatibility: since biocides are generally based on a reaction with the proteins of the microorganisms, skin proteins are frequently subject to the risk of attack from these biocides. However, risk of dermatitis begins, as a general rule, in concentration ranges above the usual concentrations used in cutting fluids.
- Low toxicity: no toxic effects must occur when the biocides penetrate the body, either through the skin or through inhalation and swallowing of the coolant spray.
- No waste water hazard: biocides enter sewage/waste treatment plants through the waste water from the cutting fluids in such low concentrations that they can be biologically decomposed without the sludge flora being destroyed. The ability of biocides to be decomposed is necessary in biological clarification plants in order to avoid burdening public waters, especially through fish toxicity.
- High economy: the great expense of preservation of coolants mentioned at the beginning requires high economic efficiency in these products. Both the recommended concentration and the concentration applied in practice should be considered when comparing costs. For better practical handling, the dilution of the active substance should be selected by the biocide manufacturers so that the normal concentration used is between 0.1 and 0.2 % in water-mixed coolants.

The most significant groups of substances used as preservatives are:

- Aldehydes (formaldehydes)
- O-Formals
  alcohol + formaldehyde
  (semi formal of benzyl alcohol)
- N-Formals
  amine + formaldehyde (hexahydrotriazines)
  methylbisoxazolidines
  amide + formaldehyde (*N*-methylchloroacetamide)
- Phenolic derivatives (*o*-phenylphenol)
- Isothiazoles (benzoisothiazole)
  chloromethyl, methyl, octylisothiazoles
- Compounds derived from carbon disulfide (sodium-dimethyldithiocarbamate)
- Sodium pyrithione
- IPBC (3-iodine-propinylbutylcarbamate)
- Phenoxyethanol
- Phenoxypropanol

12.3. Neat Cutting Fluids

Neat cutting fluids are mainly products such as honing oils, deep hole drilling oils, grinding oils and broaching oils which are not mixed with water prior to use. Neat cutting fluids have the following advantages over water-miscible fluids: they offer longer tool life and better surface finishes for difficult operations performed at low cutting speeds, the maintenance of the oil is much less complicated and the life of the oil is significantly longer. The contamination of hydraulic and slideway oils with neat oils is less of a problem than with water-miscible products. Moreover, leaks of hydraulic oils and other machine lubricants are easier to cope with if compatible oils are used.

Their principal disadvantage compared to water-miscible fluids is the poorer cooling offered by oils, particularly in high speed cutting operations. An exception is high speed grinding where the superior lubricity of oil reduces the amount of heat generated. Another disadvantage is the combustibility of oils as well as the explosion risks of oil mist and vapor. As the viscosity of oils is higher than water-miscible metalworking fluids, drag-out losses on chips and components are also higher.

12.3.1. Specifications

Neat metalworking oils are classified according to ISO 6743/7, DIN 51 385, or DIN 51 520 on the basis of the additives used (Table 39).

12.3.2. Composition

Most metalworking fluids are based on petroleum hydrocarbons and contain additives to improve lubricity, wear protection and other additives to manipulate secondary characteristics

such as foaming, misting, corrosion protection, dispersion, oxidation stability and flushing. Of the petroleum hydrocarbon base oils, low-aromatic, paraffinic solvent extracted or hydrocracked products are preferred. Recent trends in Germany are heavily focused on ester based fluids.

Apart from sulfurized fatty oils and polysulfides, sulfur – phosphorus compounds and pure phosphorus compounds are often used as EP additives in advanced cutting oils. The use of zinc- and other heavy metal-based additives in cutting oils is declining due to environmental and wastewater considerations. Chlorinated paraffins are still used as universal EP agents throughout the world but combinations of sulfur additives and ester oils are being increasingly used in Germany and Western Europe as substitutes. The reasons for these developments are the considerably higher disposal costs for products containing chlorine in Germany.

Table 39. Classification of neat metalworking oils according to ISO 6743/7, DIN 51 385 or DIN 51 520

	DIN 51385 Code letters	ISO 6743/7 Code letters
Metalworking oils containing mineral oils without antifriction additives or EP agents	SNO	MHA
Metalworking oils containing antifriction additives	SNP	MHB
Metalworking oils containing EP agents, chemically inactive	SNPA	MHC
Metalworking oils containing EP agents, chemically active	SNPB	MHD
Metalworking oils containing antifriction additives and EP agents, chemically inactive	SNPC	MHE
Metalworking oils containing antifriction additives and EP agents, chemically active	SNPD	MHF

12.4. Application

12.4.1. Machining with Geometrically Defined Cutting Edges

Turning produces parts which are round in shape. Cylindrical and conical surfaces are generated from a rough cylindrical blank which rotates about its longitudinal axis and against a cutting tool. Most turning operations use single-point tools. Normally, a tool holder contains an indexable insert with multiple cutting edges. For turning operations, water-mixed metalworking fluids are used only when running automatic lathes; neat oils are often preferred.

Drilling. The most common shape found in a manufactured part is a circular hole, and many of these holes are produced by drilling. Since the chips are formed within the part, the flutes or grooves normally serve two purposes. In addition to providing a conduit for the removal of chips, they also allow the cutting fluid to reach the tool–workpiece interface. Most drills are made from HSS which demands the application of a cutting fluid [284].

Water-mixed metalworking fluids are used where high cooling properties are required. Recent developments show that this operation may be run also with Minimum Quantity Lubrication (MQL) or by using dry cutting technology. Both techniques depend on the material machined and the machine tool designed for this application.

Milling. A variety of milling processes is available where end milling, slab milling, and face milling are included. High metal removal rates are possible since the tools have multiple teeth and each tooth produces a chip. In most applications, the workpiece is fed into a rotating tool. The feed motion usually is perpendicular to the tool axis, and cutting occurs on the circumference of the tool.

When machining cast iron, dry milling has been state-of-the-art for many years. High tensile steels and aluminum alloys are lubricated with water-mixed metalworking fluids but recent developments show that MQL can in some cases provide further advantages from a cost perspective.

Gear Cutting. The most important machining operations are gear hobbing, gear shaping and gear shaving. The different forms of teeth and various types of machines with different tool kinematics very frequently lead to different oil recommendations by the machine tool maker. Nevertheless, in this case the users very clearly want to use one and the same product for several gear cutting applications. Today in large serial automobile gearbox production this has already been generally achieved and all hobbing and generating shaping operations are carried out with just one oil.

As far as tool cooling is concerned, the tool cutting edges only cut for a comparatively short time and sufficient time is available for cooling. As a consequence sufficient cooling is provided by neat cutting oils. Although today water-miscible coolants are available and can achieve good results, the majority of machine tools for hobbing and shaping are not designed for water-miscible products. The latest trends are towards dry machining and the basis for this is coated tools and chip transportation with compressed air.

Oils with viscosities of ca. 40 mm^2/s at 40 °C are particularly suitable for gear shaving with a higher module ($m > 2.5$); recommended for smaller modules are less viscous oils with viscosities between 15 and 25 mm^2/s at 40 °C.

Deep Hole Drilling. Current deep hole drilling techniques, previously employed for the most part only in the armaments industry, have been steadily modified and are now used in the automotive, shipbuilding, aircraft, machine and tool industries, in industrial plant construction, and for hydraulic and pneumatic equipment.

As one of the process parameters, the cutting fluid can make a considerable contribution to increased economic viability, assuming that all other influencing factors have been optimized. Its principal task is to cool and lubricate the cutting edges. Chemically active EP additives incorporated into the cutting fluid and polar effective substances help to reduce wear and friction.

In addition to the actual cooling and lubricating functions, the cutting fluid, which is fed to the tool at high pressure, must also carry away the swarf from the cutting point. The high-pressure supply system requires not only a carefully designed fluid circuit (with a recirculation rate no higher than 6 h^{-1}), but also for products with low foam and in particular low oil-mist characteristics.

Threading and Tapping. Frequently water-miscible lubricants are used in the production of external threads. The focus is clearly on neat cutting oils for tapping where the materials to be cut are difficult or most difficult (material groups 2 and 3, section 1); typical tapping and threading oils of this type are sulfur active oils with viscosities between 15 and 40 mm^2/s at 40 °C; the proportion of polar substances is also important. Tight drill holes and the low pitch of the internal threads call for less viscous oils because of the favorable flushing properties. Here again the risk of cutting edge build-up is great because of the generally low cutting speed, and damage to the cutting edge can result.

Frequently recommended where tapping is carried out and on difficult and most difficult-to-cut materials are thread cutting pastes (sometimes dilutable with water up to 1 : 5) or oils containing solvent. In particular, products with chlorinated hydrocarbons, more especially 1,1,1-trichloroethane as solvent, are easy to apply by brush or in drops. Today for environmental reasons such products have been replaced by low viscosity hydrocarbon solvents and/or other synthetic fluids.

Broaching. Far more than in many other cutting operations the main consideration in broaching is tool wear or tool life. The reason for this is the fact that the broaching tool is a very complicated part and, the tool is manufactured from one single part. The material is predominantly high speed steel; carbide metal as a tool material is only used for gray iron machining. In the case of broaching several teeth engage at the same time and very frequently the chip width is large. Chip removal can be very problematic which then generally also calls for relatively low viscosity oils.

Temperature sensitivity of the broaching tool combined with cutting from the solid workpiece which follows each tooth without any soft initial cut, call for relatively low cutting speeds. In broaching mostly neat cutting oils are used. At very high broaching speeds and high stock removal rates good results are achieved on free machining steels by using water-miscible coolants, more especially EP emulsions and synthetic products free of mineral oil with relatively high concentrations (10 to 35 %); in this case, compared with neat oils, the improvements of the cooling effect and of chip removal are important factors.

12.4.2. Machining with Geometric Non-Defined Cutting Edges

In all abrasives according to DIN 69111 the abrasive grains are bonded with each other to the base

material. The purpose of the bond is to secure the abrasive grains until they become dull and then release them. A differentiation is made between resinoid ceramic and metallic bonds because of the difference in their elasticity, temperature resistance, heat dissipation and dressing capacity. Their resistance to chemicals is particularly important as far as the coolants used are concerned.

Resin bonded grinding wheels are not stable to hydrolysis and can lose their tensile strength when water-mixed coolants are used [285].

Ceramic materials are highly resistant to chemicals and temperature but are brittle and glass-like in character. Metallic bonds are becoming more and more important as far as superabrasives are concerned because of their high resistance to wear.

Grinding Fluids. The grinding process, as a precision machining process, is used frequently for the final stage machining of workpieces. As a result, high demands are set on component accuracy. Particularly significant in this respect is the coolant.

Its first task is to reduce the friction between the abrasive material and surface of the workpiece in the contact zone and, as a result, lower the temperature in the grinding area. It is necessary to reduce the contact zone temperature to increase the tool life of the grinding wheel and prevent the peripheral zones of the workpiece from being affected in any way.

A further task of the coolant is, of course, to prevent thermal damage to either the workpiece or grinding wheel by dissipating the heat generated. Flushing the grinding wheel clean is an important criterion as far as tool life is concerned. Here the viscosity and quality of the base oil as well as the coolant additives play a decisive role.

Whether a water-miscible coolant or a straight oil is suitable for a process also depends, apart from the above-mentioned criteria, on the peripheral system. Here the machine tool equipment plays a particularly decisive role with regards to extraction, sealing, wiring and type of filter system.

Honing Oils. During honing, there is a very large contact zone between tool and workpiece with relatively low pressures. This means that the heat formation is relatively low compared with grinding so that the coolant does not to have to dissipate any great amount of heat.

An important criterion when honing is flushing the honing stone to remove the stock removed material and thus to ensure that the working surface retains its grip and is self sharpening.

As a consequence, very low viscosity oils between 2 and 10 mm^2/s at 40 °C are used when honing. These are applied, as a general rule, at low pressure but at a high flow rate through ring nozzles. To achieve the component accuracy required, the honing oil is kept at a constant temperature by chillers.

Apart from straight oils, water-mixed coolants are also used, especially when honing cylinder liners with diamond honing stones in the automobile industry. Reasons for this are to be found frequently in the process line, because previous machining is generally carried out with water-mixed fluids. This leads to reduced process costs when uniform media are used.

Lapping Oils. Unlike grinding and honing, lapping is a precision finishing process with non-bonded abrasive. In lapping, the workpiece surfaces are machined by the friction between the workpiece surfaces and the appropriate counter surface in the form of a working disc. The stock removal is by means of a lap (grit and liquid) which is applied between the workpiece and lapping wheel surface.

The purpose of lapping oil is to induce homogeneous mixing of the grain. However, lapping powder agglomeration must certainly be avoided. The lapping oil has an important task to perform in the operation, in that it lubricates properly to prevent cold welding between workpiece and working disk. The lubricating film must not be applied too thickly as otherwise hydrodynamics will prevent stock removal.

In the case of lapping oils a distinction is made between the very low viscous oils, which can only maintain the lapping powder in suspension for a short period of time, and the thixotropic, high viscous products, which suspend the lapping powder for days on end. The advantage of low viscous oils is the very good washing behavior, which is reflected in clean surfaces. Due to the lack of carrier capacity for the lapping powders these products are only used

on machines where the lapping medium tank is equipped with an agitator.

12.5. Storage

Whereas neat cutting oils can be stored for a number of years, water-miscible fluids can be stored for between 6 months and maximum 1 year. Water-miscible products should be stored between + 5 °C and + 40 °C. If the ambient storage temperature falls below 0 °C, the concentrate should be stirred (rolling the drum suffices) and warmed before use.

Containers should, preferably, be stored indoors. If drums are stored outdoors, they should be placed horizontally to avoid water and/or moisture seeping past the bung when the drum 'breathes'.

Ideally, cutting fluids should be constantly monitored. Most importantly, solid and liquid impurities such as chips, abrasion residue and tramp oils should be removed by various centrifuges and/or filters because only clean cutting fluids increase efficiency and reduce scrap. The condition of the emulsion in the tanks should be checked at least weekly, better daily. Such measures can significantly increase the life of the cutting fluid.

Water-miscible cutting fluids require much more monitoring than neat oils. Compared to neat cutting fluids, their change intervals are sometimes very short. The maintenance measures required for small tank capacities are often more complex than for larger central systems.

12.6. Environmental Aspects

Maintenance procedures allow cutting fluids to remain in use for long periods of time. Nevertheless, a point can come at which the level of solid and fluid contaminants has reached such a stage that the fluid has to be disposed of and the system refilled with fresh fluid.

This poses no particular problem in the case of neat cutting and grinding oils which may either be regenerated on-site and used for less demanding applications, used as heating oil or commercially disposed of.

The situation is totally different for water-miscible products which, containing an average of 3 % organic compounds, are water-polluting substances and cannot thus be discharged into rivers or water purification plants. In most cases, expensive treatment is thus necessary. The various separation processes available are not always ideal for the different types of cutting fluid in use. Moreover, the evaluation of the water phase remaining after separation depends on not just national legislation but also local waste water regulations.

Among these processes are:

- Electrolyte separation (e.g., salt and acid splitting)
- Splitting of emulsions with adsorbents (e.g., adsorption with amorphous silica and with metal hydroxides)
- Thermal methods (e.g., immersion heating, thin-film evaporation, incineration)
- Ultrafiltration

12.7. New Trends in Coolant Technology

Oil Instead of Emulsion. These days, the 'oil instead of emulsion' trend is seen as an answer to a number of problems. Not only are cost benefits realizable, but environmental, safety-at-work and technical performance of oils are superior. Another major aspect is the almost unlimited life of oils compared to the fluid change cycles of water-miscible fluids (6 weeks for individually-filled machines and 2 – 3 years max. for central systems). As regards pure machining performance, oils can satisfy more than 90 % of all machining operations.

Replacing emulsions with oils offers better lubricity, improved surface finishes and significantly longer tool life. A cost analysis performed in a gearbox plant generated the factor 2 as an average of all machining operations [287].

The only disadvantages of machining with oils are in processes which generate a lot of heat. With a four-fold reduction in cooling, tool and machining problems can occur. Critical operations include the manufacture of carbide drills from solid stock.

To perform these processes with oil, the viscosity has to be reduced to a minimum. This generates the first conflict between technical performance and safety-at-work. The evaporation of conventional oils based on paraffinic solvent

raffinates rises almost exponentially to falling viscosity. At the same time, the flash point decreases. This problem can only be solved by the use of base oils which combine the advantages of a high flash point with low evaporation at very low viscosities.

In response to these requirements, the first cutting oils based on hydrocracked oils with esters appeared at the end of the 1980s with pure ester oils established on the market in the early 1990s.

The trend towards machining with neat oils presented a range of cost-reduction possibilities. One analysis by a leading German machine tool manufacturer showed that an average of seven different lubricants are used in every machine tool. Apart form the leakage problem and the incompatibility of some lubricants, costs are also generated by the stocking of all necessary lubricants. Incorrect application of lubricants or grade confusion can cause considerable damage to machines or even production stoppages.

One possible solution is the introduction of 'multifunctional' products which satisfy a wide range of machining operations at the same time as fulfilling the demands of advanced machine tool lubrication.

A machine tool fluid family contains the same or compatible additives in lubricants of different viscosities.

The advantages are:

- the unavoidable leaks of hydraulic fluids and slideway oils have no negative effect on the cutting fluid
- manufacturing quality remains constant without complex analytical measures
- tramp oils function as cutting fluids and thus do not create additional costs
- higher process reliability, good machining results and reduced tool wear all serve to lower manufacturing costs
- universal application

Minimum Quantity Lubrication. The ever-changing legislation and an ever increasing awareness of the environment have led to a change in the production processes used up to now and especially in the production auxiliaries such as coolants.

The analysis of an automobile manufacturer [288] published early in the 1990s revealed that a considerable proportion of the process costs are caused by the application of metal working fluids. In this respect the costs for coolant concentrates themselves played an insignificant role. The actual costs are caused by system costs, personnel costs for the care and monitoring of materials, the high personnel and investment costs associated with splitting plants associated with water purification systems and last but not least, the costs for disposal.

This has led to much more attention being paid to the difference between dry machining and minimum quantity lubrication (in minimum quantity lubrication the coolant volume flow does not exceed 50 mL/h) [289]. The drastic reduction in the amount of coolants used as a result of these new technologies only offers a considerable potential for savings in process costs. However, dispensing with coolants also means that the primary functions of the coolants such as cooling, lubrication and chip transportation have to be replaced by other suitable technical solutions.

By means of a proportioning device a small amount of coolant spray (max. 50 mL/h) is applied to the machining point. Of all the dosing equipment on the market only two systems have proved successful [290]. The pressure systems have found widespread use, whereby differences can be made in this case between two strategies for preparing the mix.

All minimum quantity coolant systems have one thing in common. The coolant is dispersed into more or less fine drops which reach the machining point as an aerosol. However this causes the first problems with regards to toxicological and workplace hygiene. Although for some years now, in the case of conventional coolant flooding at the workplace, efforts have been made to minimize spray development by new technologies and low emission coolants, sprays are now being intentionally developed for minimum quantity lubrication which are released into the workshop atmosphere.

Today greases and oils as well as esters and fatty alcohols are in use, apart from conventional mineral oils and water-miscible coolants. Since minimum quantity lubrication is a matter of pure total loss lubrication and the coolant is frequently completely dispersed in the working area in the form of vapors and mist, particular attention should be paid to safety aspects. Ester oils and fatty alcohols with additives with

low toxicological values have proved particularly successful in this respect.

Natural oils and fats have the disadvantage of poor oxidative stability. Ester oils and fatty alcohols should be given preference in the case of processes where high heat is developed to avoid residues on workpieces and machines.

13. Forming Lubricants

According to DIN 8582 a distinction between six main groups of manufacturing processes is made: primary shaping, forming, separating, assembling; changing properties of materials and coating. The processes listed under forming (DIN 8583 forming under compressive conditions, DIN 8584 forming under combination of tensile and compressive conditions, DIN 8585 forming under tensile conditions. DIN 8586 forming under bending conditions, DIN 8587 forming under shearing conditions), all refer to noncutting production. Over and above this, forming technology also includes some subdivisions under severing operations (DIN 8588 severing operations, shear cutting) and jointing (DIN 8589 jointing by forming).

Apart from the differentiation according to DIN 8582, in practical applications the area of forming technology is normally divided into sheet-metal forming and extrusion forming. This conventional distinction is based on the fact that unlike extrusion, sheet metal forming workpieces are produced with nearly uniform metal thickness over the complete product. Additionally, extrusion forming uses much higher forces than are required when forming sheet metal [51].

13.1. Sheet Metal Working Lubricants

Sheet metal forming processes have not only gained considerable importance in the thin sheet metal processing sector but significant progress has also been made in materials which in earlier times could not be processed by forming. Especially noteworthy in this respect is the forming of stainless, austenitic and ferritic, sheet metal which initiated great activity in both the development of lubricants and materials for the tools.

During large scale production, for example in the production of parts for vehicle bodies, the greatest expense in development over recent years as far as lubricants are concerned was not for the actual lubrication but for fulfilling new secondary requirements of the lubricant. Advancements in production technologies made it necessary for lubricants to fulfill specific requirements for removability, new application methods, corrosion protection, behavior during sheet metal adhesive bonding, welding, temperature resistance, compatibility with special coatings along with evolving workplace health and safety concerns.

13.1.1. Deep Drawing

During deep drawing, hollow bodies are produced from metal blanks using punches, dies and blank holders. In no other forming operation are the friction and, as a result, the lubricating conditions so complex. In one and the same drawing operation, a particularly low coefficient of friction can be required in one area and a particularly high coefficient of friction in another.

The drawing force in the flange necessary to form the sheet metal is applied by the punch in the base of the cup and transferred from here through the wall into the flange (area C in Fig. 55). This transmission of force calls for the highest possible coefficient of friction on the punch edge: neither the punch nor the sheet metal blank should be lubricated in this area. This friction area requires lubrication in the boundary mode with a high coefficient of friction and antiwear behavior. This is possible with low viscosity drawing oils, polar substances or EP additives. Moreover, in this case the thickness of the lubricant film must not be excessive. It should be somewhere in the range of the sheet metal roughness R_a. A high coefficient of friction along with antiwear performance can also be achieved in some cases by lubricants with a high filler content. These filling materials or solid lubricants should be blunt in structure but not molybdenum disulfide or graphite.

Since the best possible lubrication and lowest coefficient of friction possible are required on the punch in the blank holder area of the sheet metal flange, a compromise frequently has to be made concerning the lubrication. It is often not practicable to lubricate one side of the sheet

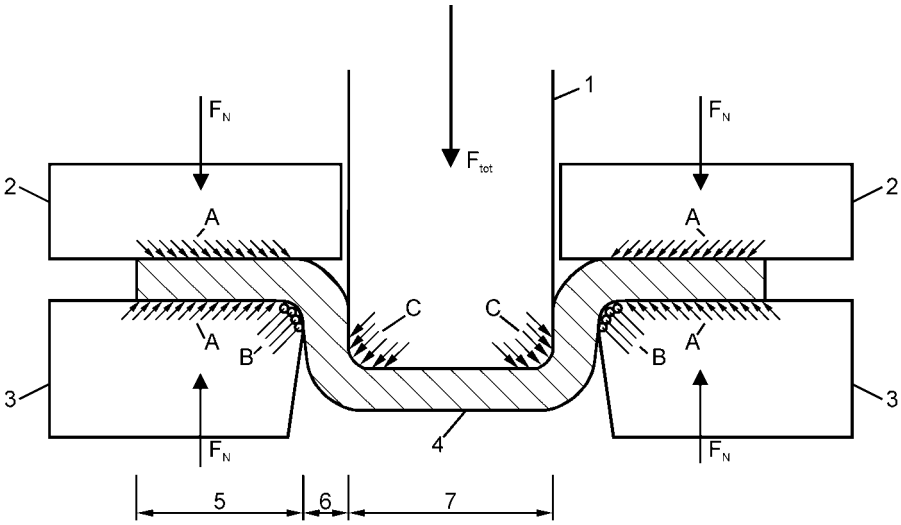

Figure 55. Friction areas when deep drawing a cupA) Friction area between sheet metal blank and holder and sheet blank die; B) Friction area between sheet metal blank and the die radius, C) Friction area between sheet metal blank and the punch edges1) Punch; 2) Blank holder; 3) Die; 4) Cup; 5) Flange area; 6) Cup wall; 7) Base of the cupF_{ges} = Total drawing force; F_N = Blank holder force

differently from the other, especially with small drawn parts.

The cup bottom and the connection to the wall have the least yield during deep drawing. The yield increases with frame height and, as a consequence, also the yield strength. If the drawing force transferred from the base of the cup is too high the sheet metal will tear at the point with the lowest strength resulting in bottom tearing, the most significant and most difficult deep drawing fault to be analyzed [291]. Here, the lubricant has the important task of reducing the considerable frictional component of the total drawing force in the region of the blank holder and the die radius.

The friction situation in the deformation area is explained by studying the friction in the region of the blank holder and the die radius (areas A and B in Fig. 55).

As a result of the tangential pressure stresses and radial tensile stresses in the flange, there is a tendency towards material compression on the outer rim. Because of this thickening of the sheet metal the blank holder can only transfer the blank holder force to a limited section of the sheet surface on the blank holder side, which leads to high surface pressures and difficult friction conditions. As a result of the lack of contact by the blank holder in the other areas, there is also the risk of fold or wrinkle formation. This wrinkling also creates special friction conditions since the backs of the wrinkles cause more or less line contact with the blank holder, producing local areas of high surface pressure.

However, in the case of high viscosity oils and drawing pastes, if the lubricant film applied is excessive, there is the risk that hydrostatic effects on the blank holder side of the flange can cause reduction of sheet metal contact with the blank holder resulting in wrinkling.

The friction conditions on the die are less complex than in the blank holder area. Here the surface pressure can be relatively accurately estimated and the coefficient of friction determined in deflection strip drawing tests. In this case the aim of lubrication is to achieve a minimum coefficient of friction with minimum wear. Even excessive lubrication of the sheet metal surface on the die side normally does not create a problem.

The geometrical design of the die has a significant influence on friction. If very small die radii are necessary for technical reasons, EP lubricants have proven to have a positive influence on the workpiece surface quality as well as die wear in spite of the high surface pressures occuring. However, this does not seem to bring about a reduction of the drawing force.

13.1.2. Stretch Drawing and a Combination of Stretch and Deep Drawing

In stretch drawing the blank sheet metal clamped firmly on both sides is stretched by tensile stress, generally by means of an upwards traveling stretch drawing punch. Here the thickness of the sheet metal is reduced unlike during deep drawing where the thickness of the sheet metal remains mostly unchanged. The goal is uniform deformation which is favored by the lowest possible friction over the entire stretch area. The reduction in sheet metal thickness leads to a considerable increase in surface area which has to be taken into consideration when choosing the type and amount of lubricant to be applied.

13.1.3. Shear Cutting

According to DIN 8580 shear cutting is a severing process in the separating group. Whereas *stamping* is a relatively old method of sheet metal working, *fineblanking* has gained particular significance in the last two decades. Lubrication has a greater significance in fineblanking than in normal stamping as far as the quality of the workpiece and tool wear are concerned. This normally requires higher performing lubricants.

Stamping. A significant characterization of stamping is that one third of the sheet metal thickness is a sheared edge and about two thirds a torn zone. An important stamping method is, for example, making holes for the production of pierced plate for protective grids and for many decorative applications or for the production of screens and filters. The sheared area of stamped parts is subdivided into three areas; the forming area, the cut area and torn area.

Efforts are being made to reduce the number of lubricants used for stamping. This reduction in types of lubricant also makes it necessary for the plants to select lubricants based on the most difficult operations, materials and sheet metal thicknesses. Today the majority of companies try to make due with a maximum of three lubricants in the production of stamped parts by combining blanking and forming operations.

The lubricants used for stamping are divided into 7 or 9 groups and can be described as follows:

- Group 1: oils, 50 to 180 mm^2/s, 40 °C; the oils are either completely or to a considerable extent active substances. Used as substances are EP materials (containing chlorine, sulfur and phosphorus), polar substances (P) such as synthetic esters, natural fatty oils or fatty oil derivatives. Group 1 is designated EP+++P+++.
- Group 2: oils, 30 to 110 mm^2/s, 40 °C; designation EP++P++.
- Group 3: oils, 30 to 110 mm^2/s, 40 °C; the oils are either completely or to a considerable extent polar substances, designation P+++.
- Group 4: oils, 30 to 110 mm^2/s, 40 °C; apart from considerable proportions of polar substances the oils contain lower concentrations of EP substances, designation EP+P+++.
- Group 5a: oils, 15 to 110 mm^2/s, 40 °C; the oils are mixtures of polar substances with mineral oil hydrocarbons, designation P++.
- Group 5b: oil, < 15 mm^2/s, 40 °C; these oils are mixtures of polar substances with mineral oil hydrocarbons, designation P+.
- Group 6a: water-miscible compounds. Mainly, these are products with a grease consistency, possibly also water-miscible oils with considerable concentrations of polar substances and medium to low EP levels, designation EP+P+++. The products are used as undiluted pastes or diluted as lean as 25 %. They are frequently diluted to a ratio of 1 : 3.
- Group 6b: water-miscible compounds. These are simple products without specific combinations of active substances. The products are used undiluted to 15 % but are frequently diluted with water at a ratio of 1 : 5.
- Group 7: corrosion protection oil. This is a surface protection oil which is already applied by sheet metal producers and is also to serve as a lubricant for further processing (mill oils, slushing oils, prelubes).

Most lubricants in Groups 1 to 5 may be diluted with petroleum hydrocarbon solvents. The solvent serves mainly to facilitate application. The difficult-to-cut austenitic materials with a sheet thickness up to 1.5 mm require a lubricant from Group 2 and lubricants for thicker sheet metals are in Group 1. Ferritic CrNi sheet metals require lubricants in Group 2 and for thick-

nesses in excess of 8 mm lubricants in Group 1 are selected. Depending upon thickness, the lubricants in the Groups 4, 2 or 1 are required for carbon steels. Lubricants in Group 5b are suitable for electrical quality sheet whereas the lubricants in Groups 3 to 5b are used for aluminum, non-ferritic metals and their alloys.

Fineblanking. The most significant difference between a part produced by stamping and a part produced by fineblanking is in the quality of the cut surface, which in many applications serves as gearwheels. Significantly higher shape accuracy is achieved by, among other things, a very small die clearance, which in fineblanking is typically only approx. 0.5 % of the sheet thickness. In stamping this clearance is approx. 10 to 15 % of the sheet thickness. The cutting surfaces have no fracture area and the workpieces have only an insignificant burr.

When fineblanking, unlike stamping, the sheet is clamped on all sides by a vee-ring. The specific tool structure is significant for lubrication. Lubricating chambers are formed by special chamfering on parts of the tool [292].

In fineblanking, in general, oils with higher concentrations of additives and oils with reactive EP substances are used because there is a strong tendency towards adhesive wear when fineblanking difficult-to-cut, thick materials with unfavorable structure. Oils with a high chlorine content are frequently used for difficult fineblanking operations [292].

Working with reactive EP substances, especially chlorine compounds, can lead to considerable corrosion problems on fineblanked parts, particularly on the cut surface. Because chlorine compounds cause environmental problems there is a clear trend towards chlorine-free products. The extensive use is of chlorine-free products is already well established in the area of low and medium thickness sheet metal.

Table 40. Different types of sheet-metal-forming lubricant

Lubricants	Main applications	Remarks
Neat lubricants		
Vanishing oils	stamping, drawing, pressing deep drawing of thin sheet metal punching and calibration	flash point 40 – 65 °C
Oils containing polar additives	lamella and disk plate stamping	low viscous applications some lubricants mineral oil free, biodegradable, compatible with prelubes
	embossing, stamping, drawing, pressing car body manufacturing, deep drawing	
Oils containing EP-additives (chlorine free)	severe stamping, pressing, deep drawing fineblanking drawing and ironing	lubricants can cause copper staining
Oils containing EP-additives (chlorine containing)	severe stamping, pressing, deep drawing	high carbon steel, stainless steel, some lubricants do not cause copper corrosion
	fine blanking drawing and ironing	
Drawing compounds without solid matter additives	severe drawing, deep drawing	low to high carbon steel
Drawing compounds containing solid matter and/or EP-additives	most difficult drawing, deep drawing	mostly pasty lubricants
Watermiscible lubricants		
Oils	drawing emulsions or solutions	synthetic, semisynthetic or conventional formulated lubricants
	secondary hydraulic fluid for hydroforming	
Drawing compounds without solid lubricants	severe drawing, deep drawing	sometimes pasty lubricants
Drawing compounds containing solid lubricants	most difficult drawing, deep drawing	often pasty lubricants

13.1.4. Choice of Lubricants

Before Forming. In this field of application a great number of different lubricants are available.

Table 40 shows different types of forming lubricants and their preferred applications.

After Forming. As mentioned before, steel sheet metal and strip are generally provided with a film of corrosion protective or prelube oil in the steel rolling mill. The range of products have not only proved successful as corrosion protective oil but also for the further processing in light to

medium severity drawing operations without an additional lubricant having to be applied.

The remaining film on the parts drawn in this way provides good corrosion protection.

If an additional lubricant (spot lubricant) is applied for forming this can be left in many cases on the workpiece after the operation to provide corrosion protection during intermediate storage or before further processing. However, in the majority of cases, it will be necessary to remove the lubricants immediately after processing or intermediate storage. This is done either by washing or, in the case of heat treatment, during the annealing operation wihtout previous cleaning.

13.1.5. Sheet Metal Forming in Automobile Manufacturing

The pressing of car body parts is one of the most important sheet metal drawing processes. The corrosion protection oil applied by the steel mills plays a part in every sheet metal forming operation and also makes up the largest proportion of lubricants used.

Prelubes. The idea to combine the corrosion protection properties of a corrosion preventive oil with the lubricity of a drawing oil led to the development of the prelubes. Prelubes have existed on the American market for more than 20 years, they were introduced in Europe in the early 1990s. Applied at the finishing lines of the steel mills, they serve as the forming lubricant in the press shops. The use of prelubes in steel mills reduces the number and quantity of spot lubricants for additional press shop oiling dramatically.

Among the many specifications for mill oils, the comprehensive standard of the Association of German Automobile Manufacturers (VDA) [293], [294] describes all relevant requirements of a mill-applied lubricant, defines all important properties and lists all test procedures.

Skin Passing. To achieve the desired quality, the final surface finish of sheet steel is often rolled-on wet. Water-based solutions are normally used for plain cold rolled steel but low-viscosity oils are used for zinc-coated sheet including electro-galvanized or hot dip galvanized. As a result of the increasing use of coated sheet steel in recent years, the compatibility of skin-pass oils is becoming an increasingly important factor. If the skin-pass and the subsequently applied corrosion protection oil are not compatible, the oil film can rapidly separate.

Additional Lubrication. Many drawing operations require additional spot lubrication which is applied in the press shop. These higher viscosity drawing lubricants lose out when prelubes are applied. Ultimately, they will become unnecessary if the lubricity of the mill applied oil on the panels is adequate. If in spite of the prelube, difficult deep drawing still requires the use of additional spot lubricants at critical points, these must be process-compatible. As such spot lubricants are again applied on top of the other oil layers, a thixotropic formulation is also advisable to avoid run-off.

13.2. Lubricants for Wire, Tube, and Profile Drawing

According to DIN 8584 these processes are classified under the manufacturing processes with forming under a combination of tensile and compressive conditions. Wire and rod drawing is applied to produce semifinished products with smooth surfaces and close tolerances for the most varied application purposes. Solid or hollow bodies are generally drawn through a stationary drawing tool (drawing die, drawing nozzle, drawing ring) to reduce the cross section or change the cross section geometrically.

13.2.1. Wire Drawing

In no other forming process the lubrication, the machine and tool technology are so closely associated with each other as in the case of wire drawing. Three different types of lubrication technology are applied:

- *Dry drawing*: (in wire drawing this term does not mean the absence of lubricant but the use of solid, not liquid lubricants) with dry drawing soaps in the drawing box. These lubricants only lubricate the drawing die and hardly ever have a cooling function.

– *Lubricated drawing with pasty or high viscosity lubricants*: which means drawing greases, drawing oils or drawing emulsions.
– *Wet drawing*: Wet drawing is generally done with oil-in-water emulsions, seldom with non-water-miscible oils, however, quite recently, also with special, highly viscous, degradable ester oils.

In wet drawing the task of the lubricant is to lubricate and cool the drawing tools and keep the machines clean. A further criterion for lubricant selection is the thickness of the wire.

Lubricants for Copper Wire Drawing. Water-mixed lubricants are used exclusively. Just in the first draft of coarse wire, a special lubricant of high viscosity is frequently used in the lubricated drawing or, alternatively, the concentrate used in a following wet drawing is used in a predrawing block to ensure more favorable lubricating conditions at the given low drawing speed. The lubricant used for this application must be compatible with the lubricant used afterwards.

Emulsions are the most important of all the lubricants in wet drawing. Apart from these, surfactant solutions, free of mineral oil and fatty oil, are significant. To a small extent other synthetic solutions are also used. The three groups of lubricants (see Table 41) can be characterized on the basis of the most important ingredients.

Table 41. Lubricants for wet drawing copper wire

Type of lubricant	Ingredients
Emulsions	hydrocarbons, mineral oil
	natural fatty oils
	synthetic esters
	nonionic surfactants
	anionic surfactants
	stabilizers, inhibitors
	antifoam agents
	other additives
Surfactant solutions	alkali soaps
	alkali salt of sulfated fatty oils
	nonionic surfactants
	other additives
Other synthetic solutions	polymers
	organic salts
	inorganic salts
	other additives

In practice these differ mainly through their degree of dispersion or their solubility properties in water.

The emulsions can vary from relatively coarsely dispersed emulsions with an average drop size of approx. 5 µm to the finely dispersed emulsions of an average drop size under 1 µm.

Lubricants for Steel Wire Drawing. The same basic technical requirements apply for wet drawing as is the case for copper wire drawing. The thicker the wire is for drawing, the greater the need for effective lubricating polar and EP substances. This applies for the emulsions used in the fine wire sector as well as for the drawing oil in the medium wire section.

Occasionally pasty drawing greases containing fillers are used in the lubricated drawing before wet drawing. The combination of semiplastic lubricant carrier layers with circulating oil lubrication is also a type of lubricated drawing [295]. However, the major portion of steel wire in coarse and medium draft is drawn with dry drawing soaps.

Dry drawing soaps are composed of alkali, alkaline-earth, and group 13 metallic soaps, sodium, calcium, and aluminum soaps are the most important. The soap content of the stearates can be between 20 to 80 %. Used as further substances in the content are natural and synthetic (under certain circumstances also chlorinated) waxes, polymers, inorganic solid substances such as lime, borax, soda, talcum, titanium oxide as well as molybdenum disulfide, graphite and sulfur. Generally alkaline-earth and group 13 metallic soaps are not dissolvable in water, but on the other hand sodium soaps are very soluble in water. This is why alkali stearates can be also applied by immersing the wire coil in the hot aqueous soap solution with subsequent drying, besides being used in the drawing box.

Dry drawing at high speed calls for lubricants with high melting soaps due to the high development of local heat. Generally, stearates have a higher melting point than oleates.

Nevertheless, technical drawing soaps are not only selected according to their melting point. More important are generally the lubricating properties, the matching grain size distribution, the flowability in the drawing box, the protection against wear for the drawing dies, the further processing of the semifinished products and the corrosion protection, etc.

In practice, it is the grain distribution and flowability of the dry drawing soap which are

decisive for lubrication. Coarse grains flow better than fine grains but are not so easily picked up by the wire. The degree can be determined by screen analysis. The resistance of a drawing soap to abrasion is also important. It depends upon the composition, production process and the milling and must not be too high so that the wire can take enough lubricant with it.

Lubricants for Aluminum Wire Drawing. To some extent, drawing compounds with a high content of drawing effective (polar) substances are still found in the non-slip single draft. Otherwise almost only very age-stable non-water drawing oils in a wide range of viscosities are used. For drawing coarse wire the oil viscosities can be between 300 and 1000 mm^2/s at 40 °C, in medium wire drawing between 50 and 100 mm^2/s and in finest and super-finest wire drawing oils with very low viscosity are frequently used, for example, 2 to 10 mm^2/s at 20 °C. The higher the viscosity, the more important preheating the oil becomes at start-up.

In the case of aluminum wire drawing very fine abrasion particles occur. Normal oil filtration, as it is used in other applications, faces great difficulties in the case of aluminum wire drawing so that the particles can normally only be removed by sedimentation at stand-still. Altogether in this way very long oil service lives (as a general rule several years) are achieved.

Lubricants for Wire Drawing from Other Materials.

Stainless Steel. For wire drawing of stainless steel, oxalate coating and also plastic films, for example from chlorinated polymers or lead layers are applied as a lubricant carrier. Lubrication in the drawing dies can be provided by dry drawing soaps or with EP oils (highly additivated, generally containing chlorine) [295].

Medium and fine wire is drawn with EP oils, finest wire also with water-miscible lubricants.

Nickel behaves in a similar way to stainless steel. Coarse wire with sodium stearate is drawn with non-water-miscible EP oils, which must be free of corrosive sulfur. During the annealing (> 800 °C in reducing atmospheres) the lubricant residue must not have a detrimental effect on the nickel surface.

Tungsten wire, which is required for the production of incandescent bulbs, is produced by a special method (→ Lamps, Chap. 5.1.). Aqueous dispersions of colloidal graphite [296] are used as lubricants. In the case of tungsten wire drawing the oxide layer serves as a lubricant carrier and improves the adhesion of the graphite [297], [298].

13.2.2. Profile Drawing

In the case of wire and rod drawing the preliminary materials will normally have a circular cross section. Where profile drawing is concerned, an extrusion molded, generally hot rolled initial material is available which already has a similar profile to that of the finished material.

Profile drawing is dominant in all kinds of steel working but has much less significance by volume in nonferrous metals.

The profiles are drawn on drawing benches, to some extent at > 150 m/min, with drawing carriages which are moved by chains, ropes, toothed racks or hydraulic plungers.

Lubricants for Profile Drawing. Dry drawing soaps in drawing boxes or even water soluble drawing soaps (applied by immersing the preliminary material in sodium stearate solution, for example) are used for lubrication. However, oil circulation lubrication has the greatest significance for completely flooding the intake side of the drawing tool – not least because of the easier insertion of the workpiece in the drawing hole during discontinuous working. The drawing oils usually have a viscosity between 50 and 120 mm^2/s.

Drawing oils with a high percentage of additives are used depending upon the degree of difficulty. In this case the same selection criteria are valid as in the case of wire drawing. Drawing oils containing chlorine are generally removed before intermediate annealing because of the development of corrosive gases. Special emulsions containing solid film lubricants and drawing grease are suitable for particularly polished drawn surfaces.

13.2.3. Tube Drawing

The great significance of tube drawing becomes clear when taking into account that about one

seventh of the steel produced worldwide is used for the production of seamless or welded tube. Moreover, the production of tube made of stainless steel, titanium, aluminum, and especially of copper has considerable significance, both commercially and technically.

Wherever possible, tube is produced by casting or forming methods in the required dimensions and tube quality. Precision tube with defined, closer tolerances, especially with defined surface finishes, can only be produced by cold forming. Cold drawing of tube, apart from other cold forming methods such as cold pilger rolling or cold rolling, is the last link in the production chain. As is the case with profile drawing, tube drawing is performed out on drawing benches. Long lengths of tube, especially copper tube, are drawn on drum drawing machines which in principle work in the same way as single wire drawing machines.

With regards to the tribological demands, the drawing method is the primary factor [299]. in the case of hollow drawing without an inner tool, only external friction occurs and as a result only lubrication from the outside is necessary. In the case of various mandrel drawing processes a considerable improvement in the inner surface finish and a defined wall thickness are achieved with a cross section reduction per draft of up to 50 %.

Lubricants for Drawing of Steel Pipes. Wet drawing with EP oils does play a role for less difficult forming operations, as is the case with profile drawing to produce blank drawing.

A special form of wet drawing is carried out with so-called reactive forming lubricants. These oils contain organic and inorganic derivatives of phosphoric acid applied initially in a heated immersion bath. For example, bundles of tubes are immersed after pickling or inert gas annealing for some 10 to 15 min in heated reactive oil at 40 to 70 °C where they are left a certain period of time before being drawn with the residual oil film. The phosphate layer remains as a rule on the pipes. Although the ferrous phosphate coating produced in this way cannot be stressed to the same degree, tribologically, as zinc phosphate coatings, or even saponified zinc phosphate coatings, the method has gained very strongly in importance for cost reasons. High alloyed materials as specified for rod and profile drawing are drawn with special highly viscous oils with chlorine additive (300 to 1000 mm^2/s, with a Cl content from 30 to 70 %). In this case the inner lubrication frequently turns out to be a problem. Apart from these, certain polymer and cellulose ether films have proved to be suitable.

Emulsion and water-mixed drawing greases have only certain significance when blank drawing with a low degree of deformation.

Lubricants for Drawing of Titanium Tubes. These are important as capillary tubes and for pipeline construction in the aircraft industry. Lubricants containing chlorine have good drawing properties which, however, cannot be used for applications in aircraft construction because of the risk of crack corrosion. The suitability of titanium for phosphatizing is considerably less than that of steel. Tribologically effective coating calls for multistage treatment with acid. In this case there is a risk of hydrogen embrittlement.

Lubricants for Drawing of Aluminum Tubes. The same working principles apply as for drawing aluminum wire. Strong polar forming oils are used mainly for this.

Lubricants for Drawing of Copper Tubes. Copper tube has gained outstanding significance in the refrigeration and air conditioning installation branch. Particularly rational drum drawing machines with high drawing speeds (up to 35 mm/s) are used with flying mandrels. The required quantity of lubricant is supplied by means of grease filler for inner lubricating. Oil circulation lubrication is used for the outside, if necessary, also with emulsions (frequently 30 to 40 %).

13.2.4. Hydroforming

The full name of 'internal high pressure forming' commonly is replaced by the term 'hydroforming'.

First applications using internal pressure are known since more than 100 years for the bending of tubes. In the 1960s the production of T-fittings became a typical application. Then exhaust components, motor cradles and other structural components for automobiles were manufactured. In

the past ten years the hydroforming technology grew considerably and it will gain more and more in importance. This is due to the general demands in saving energy and raw material cost, improved mechanical equipment and process control, replacement of complex structural components manufactured in short process chains (often with additional operations integrated in the hydroforming tools) and to the contributions of an intensive research [300].

Hydroforming is regularly made on tubular components where straight or preformed preproducts are expanded in a split tool by applying hydrostatic pressure [301]. The stages of this operation are as follows. The workpiece is placed into the die, the die halves close, the mandrels (rams or sealing plungers) move in a close position to the workpiece and the pressure fluid is filled into the workpiece while air escapes. The mandrels now seal perfectly and the internal high pressure is built up. The pressure increases and the workpiece is shifted and formed in axial and/or radial direction. Finally when the workpiece is already very close to the shape of the dies a calibration pressure of about 2000 to 2500 bar is applied by which the workpiece is accurately adapted to the contours of the dies.

Lubricants for Hydroforming. Lubricants are used in operations before hydroforming, e.g. for the bending of the tube and the preforming of the workpiece. The remaining lubricant sometimes suits also for the hydroforming process itself. Additional lubrication (sometimes called 'external lubrication') is necessary in many other cases depending on the shape and the geometry of the workpiece, especially when little radii of curvature and difficult formable materials have to be processed. Thus the range of lubricants is wide but intricate. It includes drawing oils and compounds as well as wax films and solid film lubricants [302]. In respect to the pressure conditions lubricants of high polarity in combination with antiwear or EP additives have to be used. The selection is preferably done in accordance with the application conditions. It is clear that the removability of the residues of the lubricant becomes more and more essential since this is the relevant parameter of the general process compatibility.

Another essential parameter is the secondary hydraulic fluid. This functional fluid is the pressure medium in hydroforming while the primary hydraulic fluid is used in bed and slide and to move the sealing plungers as well as the counterpressure cylinders. As secondary hydraulic fluids aqueous solutions or emulsions are in use. The solubility of the hydroforming lubricant in the secondary hydraulic fluid is one of the fundamental features of the hydroforming process.

Where oils are used as external lubricants two alternatives were successfully established in the practice depending both on the constructive conditions of the tank-side equipment and the degreasing facilities after hydroforming: the use of rather soluble and more often the use of insoluble (demulsifying) lubricants.

Additional operating steps are necessary when solid film lubricants are used. The particular application properties of this type of lubricants and the subsequent degreasing possibilities must be adequately arranged.

13.3. Lubricants for Rolling

Based on the quantitative production of semifinished goods, rolling is one of the most significant processes in forming technology. Cast slabs, blooms, billets, and wire can be processed to semifinished products through rolling.

The main forming takes place in the hot-forming temperature range, to exploit firstly, the advantages of lower rolling forces and, secondly, to obtain semifinished products with soft structures suitable for further processing. Semi-hot forming has only been used to an insignificant degree in the rolling sector. In contrast, cold rolling is a process with surprising significance, more especially in view of the increasing demands being put on forming accuracy and the quality of the surface finish.

Of all the products manufactured by rolling, in type and quantity the flat products have the greatest significance, in both the hot and cold rolling sectors. These processes are followed by the important hot rolling processes for tube, section and wire production and the cold rolling process for producing taps and spiral drills. Apart from this, teeth and thread are rolled – even on one-off work pieces – and there are a number of smooth rolling methods for surface finishing. Sheet and strip is made up to $> 90\,\%$ from ferrous metals, aluminum sheet accounts for only

approx. 5 to 8 %, with copper materials in the region of 2 to 3 %.

Wherever possible today, hot rolled strip is used instead of cold rolled sheet because of the lower price.

The use of high quality hot rolling oils in the forming process instead of water cooling has contributed to a large degree to improving the surface quality and the formability of hot strip.

In the past years a continuous further development of lubricants has occurred. Although the lubricants for the various processes are fundamentally different they have many common features for the technology and tribology of rolls, especially for the behavior of the lubricant in the roll gap.

Up to the 1930s, the standard roll stand for the cold rolling of sheet metal was the four-high reversing roll-stand. This was followed by the building of tandem roll stands, in the initial stages predominately three four-high roll stands, followed by four four-high roll stands in the 1950s and, finally, from the 1960s onwards five and six four-high roll tandem lines [303].

Today's five or six-stand tandems run at 1500 to 2000 m/min. Above all, modern roll stands can be run in all working phases at a higher speed with high mechanical safety. The reason for this is the enormous progress made in measuring, testing and control technology.

Unlike other forming processes the friction between tool (work roll) and workpiece (rolled material) when rolling must be neither too high nor too low, so that the friction necessary for feeding can be applied.

Figure 56 shows the rolling process as a simple model. The sheet to be rolled runs at the speed v_1 into the roll gap of the rolls turning at a peripheral speed v_3. During this operation the thickness of the sheet is reduced from s_1 to s_2.

Reducing the roll gap leads mainly to an extension of the rolling stock length but not to an increase in width. Accordingly, the initial speed of the sheet must be lower, and the final speed higher, than the peripheral speed of the rolls. As a result sliding friction occurs. The difference in speed between v_1 (run-in) and v_3 (rolls) given as a percentage, is called backward slip, and the difference between v_2 (run-out) and v_3 (rolls) is called forward slip.

In the contact area of the friction arc there is a point in which the speed of the rolls and sheet is the same, the neutral point N (no-slip point), at which there is no slip. As a result of varying friction, for example, caused by different lubrication, the N position within the friction arc can be moved. By improving the lubrication, N can be moved towards the roll gap exit. It is possible to reduce the pressure tension through an analog-running tensile stress, namely by involving reel tension. The tribological ratios are improved in general through this measure, which increases the degree of deformation efficiency so that the undesirable widening of the strip can be almost completely avoided.

Lower friction has the effect of lower roll wear and prevents the occurrence of heat stripes and similar rolling faults. However, the rolling operation is only possible with a minimum of friction, to safely avoid firstly, so-called refusals when starting the rolling operation, and secondly, to prevent roll slipping (and, as a result, further rolling faults such as, for example, uneven surface roughness, rubble marks, etc.) during the rolling operation.

13.3.1. Rolling Steel Sheet

Hot Rolling. Slab produced by means of continuous casting or continuous casting are produced on hot roll stands. Special hot rolling oils are used in the production of high quality strip on hot wide band lines.

These lubricants comprise thermally stable basic materials and are different from the usual emulsified rolling oils for cold rolling in that they cannot be mixed with water. Hot rolling oils are generally, together with the necessary cooling water, applied to the work rolls by means of a separate spray system and reduce the friction between work rolls and roll bite but also between work and back-up rolls. In this way the tribological relationships are improved in the long-term and lead, more especially, to reduced roll wear. Parallel to this, an effective reduction in thickness can be achieved by lowering the rolling force, in other words, the forming performance increases despite lower consumption of drive energy.

The most significant reason for the use of hot rolling oils is, however, the improved quality of the surface finish down the entire length of the strip.

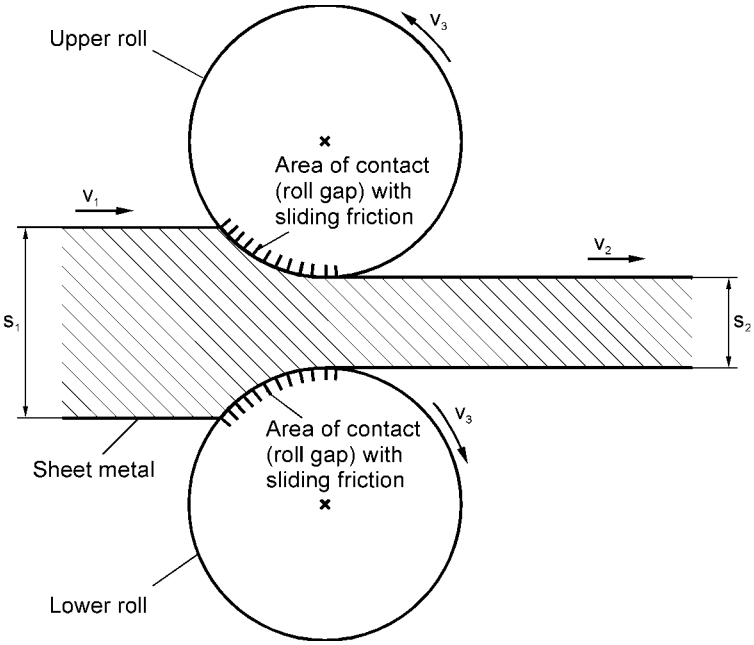

Figure 56. Principle of the rolling processv_1 is the initial and v_2 the final speed of the sheet, v_3 is the peripheral speed of the sheet, s_1 is the initial sheet metal thickness and s_2 the sheet metal thickness after rolling.

Hot rolling oils contain, besides mineral basic oils, polar and EP active substances, which form a wear-resistant 'protective skin' on the work rolls. The consumption of hot rolling oil for steel is between 15 and 80 g/t.

Cold Rolling. The final thicknesses, which are frequently between approx. 1.1 and 0.4 mm put highest demands on the cooling oil emulsions, particularly in this sector. The rolled strip widths are between 600 and 1800 mm.

The rolling speeds are between 600 to 2000 m/min, depending upon the design of the tandem line, and even higher in individual cases.

Rolling Emulsions. In rolling, the emulsion has to take over the very important function of heat transfer apart from its tribological task. The composition of this lubricant for rolling depends to a great extent on the circulation system and the other conditions and parameters for the rolling operation.

Rolling oil bases are mineral oil hydrocarbon substances or synthetic hydrocarbon material in combination with polar and EP active substances and more especially, a specifically matched emulsifier system. Rolling oil saponification numbers for rolling sheet are frequently between 30 and 130 mg KOH/g, which is equivalent to 15 to 65 % fat oil or ester components. Different synthetic esters are used, frequently polyol esters, but also vegetable and animal fats [304].

Free fatty acids indeed increase the lubricating effect but they must not be too high in proportion because of possible reaction with soluble soaps. The hydrolysis resistance of ester components must be high in order to keep the development of soaps low. The acid value (or neutralization number) of rolling emulsions is about 3 mg KOH/g, equivalent to approx. 1.5 % acids.

Rolling emulsions are mixed with fully softened water, generally between 1 and 3 %. The emulsion temperature is frequently between 35 and 55 °C. The consumption of precoat oil and emulsion concentrate is together between 0.25 and 1.00 kg/t sheet.

The emulsion stability is important for the rolling process. Stabile emulsions are generally used in high volume circulating systems, and metastable emulsions in smaller circulating systems. In the case of metastable emulsions one

advantage is the easier flotation of leakage oil, abrasion and dirt; stable emulsions are generally easier to control.

As a general rule, the emulsion is checked several times a day during operation and particular attention is paid to the degree of dispersion, the cleanness of the sheet and the risk of corrosion. The pH values are frequently between 6.0 and 7.0, the chloride content is not to exceed 20, the sulfate content 40 or the iron content 100 mg/L.

Re-rolling. In so-called skin pass rolling mills the strip must be re-rolled after annealing to improve the flatness and achieve the required material properties (forming changed tensile strength, tensile yield strength etc.). The reduction in thickness is only about 1 to 3 %. The strip is rolled dry or wet with special temper rolling oils, which, in particular, increase the strip cleanliness. Skin pass fluids have also proved to be successful for providing a uniform mat finish, because the surface of the temper rolling roll closes more slowly and does not pick up dirt or friction particles.

Skin pass fluids are synthetic aqueous solutions and mainly contain surfactants as well as organic and inorganic acid derivatives as corrosion protection components. Normally these are free of sodium nitrite and must be compatible with the subsequently applied corrosion protection oils and prelubes.

Finest Sheet Cold Rolling. Rolling is carried out with semistable high fat content emulsions by the so-called direct application of fat oil dispersions. In the case of direct application, natural fats such as palm oil or animal tallow are often prepared as a 10 to 20 % dispersion in completely softened water in large mixers. The drops of fat oil are approx. 50 µm in size. The fat and water are separated during the return run, treated and used again. Recovered fat can be used for rolling up to an acid number of approx. 20 mg KOH/g.

Where semistable emulsions are used initial concentrations of 3 to 6 % in fully softened water are employed. The size of the drops in such emulsions is between 10 and 50 µm.

Cold Rolling of High Alloy Steel Sheet. Higher alloyed steels, especially high grade stainless steels, are rolled on multiroll stands with a preference for lubricants which do not mix with water.

The rolling speeds, which depend upon the alloy and material thickness, can be up to 500 m/min, in which case, for example, high grade stainless steel hot strip be reduced by up to 85 % in eleven passes without intermediate annealing. Frequently, however, annealing, pickling and re-rolling have to take place in order to achieve the required strip quality.

The selection of the rolling oils has a decisive influence on the surface finish of the sheet. Mat surfaces are produced when the viscosity is too high even when using ground or polished work rolls. This is almost always undesirable in the case of high grade stainless steels. This is why in practice preference is given to mineral oils with viscosities of ca. 15 to 20 mm^2/s at 20 °C which have the closest possible boiling range [305].

13.3.2. Rolling Aluminum Sheet

Hot Rolling. Depending upon the alloy and required reduction in thickness the blocks are preheated to 450 to 580 °C before hot rolling. With a rolling speed between 180 and 300 m/min in the reversing process and appropriately higher speeds in tandem stands, the material is rolled with stable emulsions in an initial concentration between 2 to 6 % and an upstream temperature of 35 to 60 °C. The temperature in the last pass, depending upon the alloy, is still 280 to 230 °C.

The concentrates comprise mineral oil hydrocarbons or synthetic hydrocarbons with polar additives (ester, fatty acids, fatty alcohols), surfactants and further antiwear substances, where needed.

Fatty acids form aluminum soaps which can cause too high slip and tarnishing. This is why they are only to be used in low amounts.

Decisive for the lubrication and quality of the surface finish is the development of a uniform, fine coating of aluminum pick-up (roll coating) on the work rolls. If the coating is too thin, the friction conditions are unfavorable and frequently unstable. On the other hand a roll coating which is too thick causes uncontrolled release of parts of this coating which, in turn, leads to pick-up rolling and surface faults. To prevent this, excessive pick-up is removed from the work

rolls during the rolling process [306] by means of brush rolls. Exact control of the temperature of the work and back-up rolls is accomplished by means of specific emulsion dosing to obtain a favorable hot strip profile.

Cold Rolling. After soft annealing at temperatures from 370 to 430 °C, finish rolling is carried out on cold wide strip lines at speeds up to 2700 m/min. The sheet thickness is between 0.1 and 3.0 mm but the major amount of rolled strip is about 0.6 mm thick and achieved in two to seven passes.

Despite the advantages of better heat dissipation, lower fire risk and less expensive health protection, water-miscible rolling oils [307] have been unable to establish themselves for this application, because of, for example, patch forming as a result of residue moisture in the round coil, and hydrogen embrittlement of the work rolls can only be controlled to a deficient degree.

When making up strip rolling oils, the main attention has to be paid to the selection of the base oil. Particularly suitable are paraffinic hydrocarbons with a viscosity which is not too high, in order to avoid mat surface finishes. Consequently, closely cut base oils with flash points above 85 °C and a viscosity from 2 to 4 mm^2/s at 20 °C are mainly considered. The most popular additives are straight-chain alcohols, acids, and esters with chain lengths from 10 to 14 carbon atoms as well as oxidation inhibitors. The activation level seldom exceeds the 5 % mark.

In principle, foil rolling oils differ from strip rolling oils only by a lower additive content, occasionally also by lower viscosity and preferably by a very low aromatic content to ensure compliance with the stipulations for use in the food sector.

As is the case with hot rolling, to avoid spotting synthetic or barely-synthetic special lubricants and hydraulic oils are also used when cold rolling on the roll stand and in the rolling operations.

13.3.3. Rolling of Other Materials

Copper and its alloys are generally rolled in a hot roll process with water, i.e., without special lubricants, although occasionally with stable slightly fatty emulsions.

Generally stable emulsions with fatty oils and synthetic esters are also used for cold rolling. Under all circumstances, staining of the bright, mostly decorative surfaces by the emulsion or its contents has to be avoided. Consequently special attention must be paid to ensure the absence of substances containing sulfur [308].

Fatty oils are used for cold rolling of *titanium*.

Zinc is rolled in a semihot state at 200 °C, very often dry, or cold with a slight lubrication using low viscosity oils with polar substances or emulsions. By heating to between 120 to 150 °C when cold rolling, a large part of the lubricant is volatilized so that no further cleaning is necessary.

Tin and lead are generally cold rolled. Low viscous special lubricants are used for special alloys (solders and similar materials). Tungsten and molybdenum rolling has to be carried out at high temperatures under inert gas due to sensitivity to oxidization at high temperatures. Alternatively coating with glazes is also possible. Semi-hot or cold rolling is carried out with solid matter content or strongly polar oils with additives.

13.4. Lubricants for Solid Metal Forming

Under this heading, a large number of forming methods are included, most important are extrusion and forging.

Mainly processes are classified on material flow and the type of the tool used. However, there are frequently no clear differences between forging and extrusion.

Important processes include:

a) *Upsetting* is a process with compression between flat dies.
b) *Extrusion*. During extrusion processes, the workpiece is placed in a container and compressed by means of ram movement. Extruded products can be hollow or solid. According to the material flow, forward (direct), backward (reverse) and cross procedures are distinguished.
c) *Impression die forging* is the dominant hot forging method economically. This is also

reflected in the significance of this group of lubricants.

Closed die forging is a special form of impression die forging. Here the filling of the die is not supported by the development of a flash. There is no possibility for excessive material to escape, and vent holes are provided for vapor generated from the lubricant to escape.

d) *Open die forging* is different from impression die and closed die forging in that the metal is never completely enclosed as it is being shaped by the dies.

Cold Extrusion of Steel (Extrusion Oils).

Lubricating Oils. Of particular significance is the lubrication with EP oils, especially when working with wire without surface pretreatment of the blank metallic shearing surfaces. In the case of tools with extremely long service life, the forming is possible with the majority of highly active EP substances. An example of this is the production of inner square socket-head screws.

Apart from lubricating properties, the extrusion oils must have high thermal stability; the oil is overheated locally on the extrusion bodies and tools and temperatures over 90 °C can occur in the circulating system. The extrusion oil also has an essential cooling function, especially in the case of high speed automatic machines. No adhesive residue must remain as a result of the temperature stress as these could cause irregularities in operation.

The preferred viscosity range of extrusion oils is between 35 and 65 mm^2/s at 40 °C. Selection criteria for the initial viscosity are the workpiece temperatures, size of the extruded parts, machine pumping facilities and the specific criteria for the transfer between the working stages, as well as the thickening effects as a result of enrichment by abraded phosphate particles where phosphatized material is concerned.

EP additives containing sulfur and chlorine play a role in extrusion oils. Additives containing chlorine have lost their significance in Europe due to the environmental problem described earlier. Apart from the sulfurized fatty oils there are also other products in the polar additive sector such as thermally stable synthetic esters (also as base oils). Sulfurized mineral oils, sulfurized fatty oils and polysulfides play a special role as sulfur carriers. Zinc dialkyldithiophosphate smoothes the surface roughness of the workpieces [309]. Alkyl and arylphosphoric acid esters and even phosphoric acid partial esters are also used [310]. Because the thermal stress is considerable when using chlorinated products, particular attention must be paid to their stability and the risk of corrosion on machines and extruded parts. A degreasing of the parts directly after extrusion is recommended when using strongly chlorine containing oils.

For the most important areas of application, extrusion oils are classified by type and amount of additive into four classes [311]

- standard screws and high tension hexagonal bolts: oil with polar additives and EP additives on a phosphorus basis (also suitable for nonferrous metals)
- larger size bolts and hexagonal nuts produced from zinc phosphatized wire on multistage presses: oils with polar EP additives, copper active
- cylinder bolts with an inner hexagon recess or inner toothing made of wire on multistage presses (larger dimensions): oils with an active proportion of polar additives and EP additives on a sulfur basis, copper active
- high tensile rust and acid resistant steels subject to high deformation: maximum alloyed oils with a very high share of EP additives on a (chlorine) sulfur and phosphorus basis, copper active

Thermally stable, biodegradable ester oils with excellent tribological characteristics will gain more significance in the future.

Phosphate Coatings and Soap Lubricants. Phosphatization has remained the most important method of surface treatment for cold extrusion of steel until today and has undergone further technical development in conjunction with other lubricating systems. Particularly favorable is the ability of the layers of zinc phosphate to go along with large expanding surfaces without breaking away when extruding.

The high resistance to pressure, shearing and adhesion during extrusion of zinc phosphate coatings are considered as positive.

A fine crystalline coating approx. 5 to 7 mm is to be used for oil lubrication on multistage presses. Layers, which are too thick, lead to problems due to formation of layers of phosphate on the tools. This system is preferred for

parts with low piece weights such as screws, nuts, bolts or spark plug bodies produced on multistage presses.

Alkali soap in connection with a zinc phosphate coating has gained particular importance. In this case, not only the absorptive and chemically absorptive binding of the lubricants to the workpiece surface is neutralized but there is also considerable chemical conversion of the soaps with the zinc phosphate.In the aqueous solution the alkali soap reacts with the zinc phosphate coating according to the following formula:

$$Zn_3(PO_4)_2 + 6\,CH_3(CH_2)_nCOONa \rightarrow 3\,Zn[CH_3(CH_2)_nCOO]_2 + 2\,Na_3PO_4$$

zinc phosphate + sodium soap → zinc soap + sodium phosphate

Some 50 to 60 % of the phosphate coating can be converted depending on temperature, dwell time, concentration and the developing phosphate layer. The type of soap also plays a considerable role in this respect. Salts of stearic acid are more reactive than the salts of oleic and palmitic acid.

The zinc soap developed by the reaction leads to lower coefficients of friction when forming and the firm anchoring with the phosphate layer is particularly favorable for the lubricating operation.

The reactive soaps are supplied by the manufacturers in powder or flake form and mixed by the user with water to provide 2 to 10 % soap solutions, depending upon the degree of difficulty. The parts are soaped by immersion over a period of 2 to 6 min at bath temperatures from 60 to 85 °C [312–314].

Solid Lubricants. Although soap lubricants in conjunction with zinc phosphate coatings are preferred for medium to larger sized parts, structure-effective solid lubricants, especially molybdenum sulfide, are used for smaller parts.

As a general rule the solid lubricants are applied to a phosphate layer as dry powder lubricants by tumbling or from aqueous suspensions by immersion or spraying. Application by immersion follows in dipping baskets or rotating screen tubs [311], [315–318].

In a few cases, solid lubricants can be applied as suspensions in mineral oils, synthetic oils, or even solvents. In special cases, this method can solve wetting problems better than aqueous suspensions. In general, however, the aqueous dispersions offer considerable advantages as carrier fluids as a result of their environmental friendliness, especially when compared with organic solvents, and can also be better integrated in the wet treatment of workpieces.

Warm Extrusion and Forging. The yield stress decreases with the temperature increase in the workpiece material and as a result lower forming forces and higher degrees of workability can be achieved. This effect is utilized to some extent through the existing forming heat (up to 250 °C) even when cold extruding (without preheating the blanks), especially when working on high speed multistage presses.

Warm forming has the greatest significance for high alloyed steels and special materials. The method is applied when the parts cannot be cold formed or when it is possible to ensure more economical production by the warm process by less forming stages.

A phosphate layer is expedient at temperatures under 400 °C – as when cold extruding.

In the temperature range up to 350 °C zinc phosphate coatings serve as carrier layers; the most important fluids are mineral oil based oils with a high flash point, polyglycols and polybutenes; they can be applied as loss lubrication and at temperatures up to about 300 °C as circulating lubrication. Solid lubricants can also be suspended in the above-mentioned oils; in this case mainly MoS_2 and graphite. Aqueous MoS_2 and graphite suspensions are also suitable.

The lubricants mostly used in the temperature range 350 to 500 °C are graphite dispersions in water or organic carrier media. The decomposition of synthetic polymer oils is intentionally accepted, if necessary. In this case thermal resistance limits are set for MoS_2 lubricants.

In the temperature range 500 to 600 °C aqueous graphite dispersions are given preference and applied by spraying. Decomposition of the solid lubricants graphite and MoS_2 can be prevented when these are mixed with boric oxide powder [319].

At temperatures > 600 °C the lubricants and the application technology are similar to those applied for the hot forging process. In addition to graphite, zinc sulfide has also proved successful as a solid lubricant because of its particular re-

sistance to temperature in individual cases; it is also used in operations with graphite and water soluble organic carriers such as polyglycols.

Hot Forging. The hot workpiece material (in the case of steel approximately 1200 °C) flows into the cavities of the die during the forming operation. This is carried out to some extent at a high relative speed which depends, firstly, on the speed of the striking tool (hammer or forging press) and, secondly, on the die.

In hot forging of steel workpiece temperatures are between ca.1100 and 1200 °C and die temperatures preferably between 150 and 300 °C (maximum up to 500 °C). The temperature of the lubricant layer is between 725 and 775 °C at a die temperature of 250 °C and a workpiece temperature of 1200 °C.

Graphite preparations have gained outstanding significance as lubricants. Colloidal preparations in oil and in water as well as pasty graphite lubricants are used.

Graphite preparations are generally applied as water-based suspensions. The wetting problems given in the case of aqueous graphite suspensions, especially at high forging temperatures, can be solved by incorporating an organic phase with appropriate surface active substances, either as an emulsion phase or even as a real solution.

The propellant effect can be improved through inorganic salts (alkali carbonate and bicarbonate) or through organic substances in aqueous solution or by dispersion. Bonding agents frequently on silicate or borate basis are used to form surface films.

If the die temperature is clearly under 200 °C or even under 150 °C when the lubricant is applied it may well be that no film can be formed by the aqueous dispersion, depending on the available evaporation time. In such cases organic carriers can be used which either evaporate more quickly or form an oily film. When selecting synthetic polymer oils, care must be taken that no toxic monomers develop during the depolymerization process during contact with the hot forging. In individual cases pasty graphite preparations are still being used.

Aluminum Forging. Because of their low weight and excellent mechanical properties, parts produced by massive forming from aluminum and aluminum alloys are gaining increasing importance. Because of the major adhesion problem, the technology in the usual mechanical presses with high deformation rates has still not been completely developed. The most extensively used lubricants are water-based graphite dispersions [320–323].

14. Lubricating Greases

14.1. Introduction

Definition. Lubricating greases can be defined as solid to semifluid products that represent dispersion of a thickening agent in a liquid lubricant. They usually contain additional compounds to impart special properties, and usually the thickening agent is a metallic soap. It is not easy to classify greases as either liquids or solids – the transitions are fluid.

Greases in general contain from 65 to 95 wt % base oils, from 5 to 35 wt % thickeners and from 0 to 10 wt % additives.

History. It can be speculated that grease-like lubricants were already known to the Sumerians and used by them in their wheeled vehicles from ca 3500 to 2500 B.C. [324]; it has also been stated that as early as 1400 B.C. the Egyptians used greases made from olive oil or tallow and lime to lubricate the axles of their chariots [325], although only pig fat was reported for that purpose by ancient authors such as Dioscurides and Plinius Secundus [326]. Probably the first grease of the industrial age was patented by Partridge [327] in 1835; it was a calcium grease, also made from olive oil or tallow. Greases based on mineral oils and thickened by soaps were probably first proposed ca 1845 by Raecz [328], and a sodium grease made with tallow was patented by Little [329] in 1849.

Much general information on greases is given in two encyclopedic monographs, published in 1937 and 1954 [330], [331].

Classification. Greases are named after the industry in which they are used, e.g., steel mill greases, after the kind of application, e.g., wheel bearing greases, after their prevalent application temperature, e.g., low-temperature greases, or

after their range of application, e.g., multipurpose greases. The meaning of the latter name has changed with the years [332] and the other names do not say very much about the quality of the performance of the greases in question.

Even today, greases are named according to the consistency classes defined by the US National Lubricating Grease Institute (NLGI) in 1938 and in accordance with the cone-penetration method developed in 1925 [333] (Table 42).

Table 42. Classification of greases by NLGI numbers.

NLGI number	Appearance	ASTM worked penetration (1/10 mm)	Application
000		445–475	gear greases
00	Semifluid	400–430	
0		355–385	
1	Soft	310–340	
2	Creamy	265–295	greases for bearings
3		220–250	
4		175–205	
5		130–160	
6	Soap-like	85–115	block greases

Nowadays the performance of lubricating greases is described by standards such as ISO 6743-9 [334] or DIN 51825 [335] defining mainly consistency, upper and lower operating temperature, water resistance, and load-carrying capacity, and for automotive greases by ASTM D 4950 [336], which has been followed by the consideration of reference greases [337] and the introduction of the NLGI Certification Mark [338].

But to some extent greases are best judged simply by the chemical and physical properties of their base oils and the agents thickening them – naturally the consistency of greases increases with increasing thickener content, and with it some of their properties change. These properties at the same time best reflect the reasonable limits of practical applications.

14.2. Components of Greases

14.2.1. Thickeners

Thickeners not only transform liquid lubricants into consistent lubricants, they change the properties of the liquid lubricants also (see Section 14.8). Commercially important thickeners are listed in Table 43 [339].

14.2.1.1. Simple Soaps → Metallic Soaps

A maximum thickening effect is achieved with carboxylic acids having 18 carbon atoms. Thus, soaps are usually prepared both from vegetable-derived 12-hydroxystearic acid [340], [341] and from animal- or vegetable-derived stearic acid [342], or from their esters, usually their glycerides, and from the hydroxides of elements of the alkali and alkaline-earth metals. Soaps, by gelling their base oils give greases most of their unique properties. They are not only present as crystallites and dissolved molecules, but above all in a separate phase represented by agglomerates called fibrils or fibers (see Section 14.4).

Soap Anions. The carbon chain length of a carboxylic acid affects the solubility and surface properties of a soap. Longer and shorter carbon chains than 18 carbon atoms reduce its thickening capacity. Increasing the chain length increases the solubility in the base oil, shortening the chain length reduces it. A branched alkyl chain lowers the melting point of a soap and reduces its thickening effect. Unsaturated carboxylic acids are more soluble in mineral oils and also reduce the thickening effect and lower the dropping point. Their use is limited because of their lower oxidation stability. Hydroxyl groups increase the melting point and the thickening effect of a soap because of its increased polarity.

Soap Cations. The soap cations also are responsible for essential properties of soap greases. The cations govern the thickener yield, the dropping point, the water resistance, and, to some extent, the load-carrying capacity of a grease (see Section 14.8).

In 1996 greases based on simple soaps still accounted for more than 70 % of known world production [343]. Lithium soaps were the most important, with ca 50 %, then the calcium, sodium and aluminum soaps. The importance of the latter has steadily decreased during recent decades.

Table 43. Competitiveness of thickeners

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Sum
Lithium 12-hydroxystearate	2.5	1.0	2.0	1.5	2.0	2.0	2.5	1.5	2.5	2.0	1.0	3.0	2.0
Calcium 12-hydroxystearate	3.0	1.0	3.0	1.0	1.5	1.0	2.5	1.0	2.0	2.0	1.0	3.0	1.8
Lithium complexes	1.5	2.0	1.5	2.0	1.5	2.0	2.0	2.5	1.5	2.0	2.0	2.5	1.9
Calcium complexes	2.0	3.0	2.0	2.0	1.0	1.5	1.5	3.0	1.5	2.0	2.0	2.0	2.0
Aluminum complexes	2.0	2.0	2.0	2.5	1.5	2.0	2.0	2.5	2.0	2.0	2.0	2.5	2.1
Inorganic thickeners	1.5	1.0	1.0	3.0	3.0	1.0	3.0	1.0	3.0	3.0	2.5	3.0	2.2
Polyureas	1.0	1.5	1.5	2.5	2.0	1.5	2.5	2.0	3.0	3.0	1.0	2.0	2.0
Terephthalamates	1.5	1.5	1.5	1.0	2.5	1.5	2.0	1.0	2.5	2.0	1.0	2.0	1.7
Calcium – sulfonate complexes	2.0	3.0	2.0	2.0	1.0	2.0	1.5	3.0	1.0	1.0	2.0	1.5	1.8
Carbamate-like thickeners	2.0	1.5	2.0	2.0	2.0	2.0	2.5	1.5	2.0	2.0	1.0	2.0	1.9

I = high temperature, II = low temperature, III = aging, IV = compatibility, V = oil loss, VI = toxicity, VII = tackiness, VIII = flowability, IX = load, X = shear, XI = friction, XII = wear.
1.0 = excellent, 2.0 = average, 3.0 = poor

Lithium Soaps. Lithium soap-based greases are usually prepared by reacting lithium hydroxide as a powder or dissolved in water with 12-hydroxystearic acid or its glyceride in mineral oils or synthetic oils. The production temperature is between 160 and 250 °C, depending on the base oil and the type of reactor in use. The dropping point of a mineral oil-based NLGI 2 grease is usually between 185 and 195 °C. For such a multipurpose grease ca 6 wt % of soap is required with a naphthenic oil, ca 9 wt % with a paraffinic oil, and ca 12 wt % with a polyalphaolefin, and a kinematic viscosity of ca 100 mm^2/s at 40 °C. The thickening effect depends not only on the carbon distribution ($C_{aromatic} + C_{naphthenic} + C_{paraffinic}$) in a base oil, but also on its viscosity.

The fiber size in lithium 12-hydroxystearate greases is usually between 0.2×2 and $0.2 \times 20\,\mu m$. Good multipurpose properties, e.g., a high dropping point, good water resistance, good shear stability – which has been related to hydrogen-bonding of the hydroxyl groups – and good response to additives are the reasons why lithium 12-hydroxystearate-based greases have now been the most popular for more than half a century. They have found a wide range of applications from EP greases based on oils with kinematic viscosities of ca 200 to 1000 mm^2/s at 40 °C for heavy loads, through multipurpose greases based on mineral oils with kinematic viscosities from 60 to 120 mm^2/s at 40 °C for all kinds of bearings, greases made with diesters or polyalphaolefin oils and kinematic viscosities of 15 to 30 mm^2/s for high-speeds, to those with oil-insoluble polyalkylene glycols for gears. The lower and upper application temperature limit for a lithium soap thickened grease, as for all other greases, depends mainly on the physical properties of its base oil (see Section 14.3). Oil separation has been discussed as a criterion for both lower and upper application temperature limits [344]. In recent years attempts have been made to improve the structural stability of lithium soap-based greases by use of reactive polymers [345], [346].

Calcium Soaps. Greases based on calcium 12-hydroxystearates (anhydrous calcium soaps) are produced in the same way as lithium soap-based greases, but at temperatures between 120 and 160 °C. The size of their fibers lies between those for lithium soaps and hydrated calcium soaps. The greases can be used up to 120 °C. Their dropping point is between 130 and 150 °C, depending on their base oil. They usually have a very good corrosion resistance and good oxidation stability; when prepared from suitable base oils they are possibly the best low-temperature greases.

Calcium soaps mainly based on stearic, palmitic, or oleic acid still are called hydrated calcium soaps. For these greases the cost of raw materials, and the performance levels, are lowest. Calcium soap greases are prepared by neutralization of a slurry of calcium hydroxide in water with fatty acids or fats in mineral oil. Sta-

ble greases can only be obtained in the presence of some water, usually ca 10 wt % of the soap. The water content is usually adjusted in a second step in a stirring, or cooling vessel. The fiber size is usually ca $0.1 \times 1\,\mu m$. When the water is removed, the grease structure collapses. The dropping point of this type of grease is therefore only 90 to 110 °C and the upper application temperature limit is only ca 80 °C.

The greases have very good water resistance and good adhesive properties. Because their manufacture is rather demanding in relation to the performance obtained, their importance is declining rapidly.

Sodium Soaps. Compared with lithium and calcium 12-hydroxystearate greases the importance of greases based on sodium soaps is nowadays low, although in the form of semifluid products they still are of some interest for the lubrication of gears. Sodium greases made from fatty acids or fats have dropping points of ca 165 to 175 °C. Their upper temperature limit is ca 120 °C. Products are available with short and long fiber structures; the latter reaching up to $1 \times 100\,\mu m$ and are responsible to some extent for the rather high load carrying capability in gears. Although the greases have extraordinary good corrosion-preventive properties if only small amounts of water are present, their main disadvantage is the solubility of the sodium soaps in larger amounts of water; this leads first to gel formation, which dramatically increases the apparent viscosity, and later to breakdown of the whole structure.

Aluminum soap-based greases are usually produced with pre-manufactured aluminum soaps, usually aluminum stearate. They have dropping points of up to 120 °C, their upper temperature limit is 80 to 90 °C – above 90 °C they tend to gel. Typical of this soap is a particle size of $< 0.1 \times 0.1\,\mu m$, which is to some extent responsible for the low shear stability and the pronounced thixotropic behavior of the products. Aluminum greases are usually very transparent and smooth. They have good water resistance and adhesive properties, but have been widely replaced by lithium greases, partly because, for consistent products, aluminum greases must not be stirred during the finishing process but poured into pans and left for several hours to cool.

Cation Mixed Soaps M_1X/M_2X. Mixtures of soap greases with different cations, mainly lithium – calcium, calcium – sodium, and sodium – aluminum are called mixed-soap greases. Their properties largely depend on the relative proportions of the components. Lithium – calcium greases have better water resistance than pure lithium greases and often better shear stability. As long as the proportion of calcium soap does not exceed 20 wt % their dropping points, 170 to 180 °C, are close to that of the lithium soap alone, and their wear and friction performance are better than that of pure lithium grades [347]. Some calcium – lithium greases perform better than pure calcium 12-hydroxystearate greases.

Lithium – calcium greases have become quite popular as specialized multipurpose greases. Greases mainly based on sodium-aluminum stearate, have been used as substitutes for lithium greases. Cation mixed soap greases are usually manufactured by a one-step process, because the stability of mixtures of finished products are not always satisfactory.

Anion Mixed Soaps MX_1/MX_2. Because of the animal and vegetable origin of the acidic components most simple soap-based greases are already anion mixed soap based. Nevertheless the fine-tuning of multipurpose greases and specialized multipurpose greases is, especially for the comparatively pure 12-hydroxystearic acid, often achieved by substituting small amounts of the main acid by a complementary acid, e.g., behenic, naphthenic, or stearic acid.

14.2.1.2. Complex Soaps

Simple soaps can form complex soaps with salts of other acids. Examples are inorganic acids, e.g., boric and phosphoric acid, or with short-chain carboxylic acids, e.g., acetic acid, or with dicarboxylic acids, e.g., azelaic [348] and sebacic acids [349]. The addition of complementary salts always leads to an increase in the dropping point from ca 50 to 100 °C and to reduced oil separation, primarily because of the increased thickener concentration, and, thus, for the same reason, to reduced suitability for low-temperature application. Because of their superior properties complex soap-based greases have

become very popular and nowadays account for ca 20 % of all greases sold.

Lithium Complex Soaps. Apart from their higher upper-temperature limit, between 160 to 180 °C, some lithium complex soap-based greases perform similarly to the corresponding simple soap-based products. Of the many known formulations the most widespread is based on 12-hydroxystearic acid and azelaic acid, introduced in 1974 [350]. The first complex was based on 12-hydroxystearic acid and acetic acid and was patented as early as 1947 [351]. The lithium complex soaps with the best load-carrying capacities contain boric or phosphoric acid. The fiber size of such a complex soap does not differ much from that of the simple soap and it does not change significantly with common shear. In addition to azelaic and boric acids, other acids have also been examined systematically

Lithium 12-hydroxystearate
+
Lithium adipate
Lithium azelate
Lithium dimerate
Lithium sebacate
Lithium terephthalamate
...
Lithium borate
Lithium phosphate

With a share of about 10 %, lithium complex greases are the most widespread complex greases.

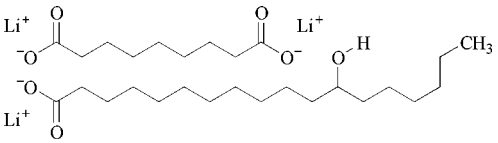

Calcium Complex Soaps. All commercial calcium complex greases contain acetic acid as a complementary acid. The complex was first described in 1940 [353]. Calcium complex greases have good shear stability and water resistance, a low oil separation, and good load-carrying capacity. Their upper temperature limit is 160 °C. Because of the formation of ketones, from about 120 °C pronounced hardening can occur. It has, nevertheless, been possible to retard hardening with the aid of polymeric structure-modifiers [354].

Calcium Sulfonate Complex Soaps. Competitive greases based on this complex were first offered in 1985 [355]. Initially they comprised in situ-produced overbased calcium sulfonates and the calcium salts of other sulfonates, of 12-hydroxystearic acid, and of boric acid [356]. The complex could be improved by replacing calcium borate by phosphate [357]. Greases containing calcium acetate instead are also possible [358]. The greases are extraordinarily corrosion protective and resistant to shear and have load-carrying capacities that can be matched only by greases of other soaps with very many additives. Their dropping points start at ca 220 °C, but their upper temperature limit is ca 160 °C. Some grades can, nevertheless, survive temperatures up to 250 °C for several hours. It can be expected that these greases will become even more important in the future, although, because of their high thickener content their low-temperature pumpability is limited.

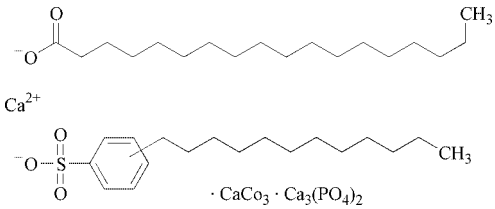

Aluminum Complex Soaps. Today only one of the possible aluminum complexes is still widely used. It comprises aluminum stearate and benzoate, and was first patented in 1952 [359]. Aluminum complex greases of that kind have good water resistance and good low-temperature properties. Their importance has decreased in recent years, although attempts have been made to make them more attractive again. This might be possible for food grade and biodegradable greases (see Section 14.10).

Other Complex Soaps *Sodium complex soap-based greases* have been used because of their suitability for high relative velocities, but like the simple soaps they lost importance because of their limited water resistance; *barium complex soaps* have been replaced almost as completely as the simple soaps.

Titanium complex greases were patented in 1993 [360]. They are based on 12-hydroxystearic and terephthalic acid. Of their properties [361] good load-carrying capability is most worthy of mention.

14.2.2. Other Ionic Organic Thickeners

Of many possible soap-like salts only the sodium and the calcium salts of stearylamidoterephthalic acid are used on an industrial scale.

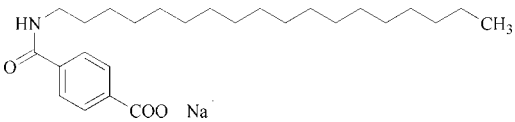

(14.1)

They were patented in 1954 [362] and proposed for multipurpose greases in 1957 [363]. Greases of this kind have dropping points of up to 300 °C and upper temperature limits of up to 180 °C. Although they have the thickening effect of simple soap greases they behave like complex greases; this makes them valuable wide-range greases. The thickeners are among the most expensive and are preferably used with synthetic base oils. Complex soaps comprising terephthalamate and benzoate have been described and complexes of aluminum stearate with terephthalamates have also been investigated [364].

14.2.3. Nonionic Organic Thickeners

Of a rather large number of theoretically suitable compounds, only the oligoureas, commonly called polyureas, have gained real industrial importance.

Diureas and Tetraureas. Oligourea thickeners were introduced in 1954 [365]. The reaction products of one molecule of di-4,4′-isocyanatophenylmethane (MDI) or other diisocyanates with two molecules of monoamines are called diureas.

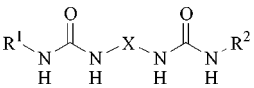

(14.2)

Tetraureas are the reaction products of two molecules diisocyanate with one molecule of diamine and two molecules of monoamine.

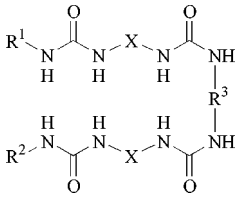

(14.3)

Depending on the performance required aliphatic or aromatic amines or mixtures of both are used. With excess diisocyanate three-dimensional structures are built along biuret-like bridges.

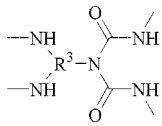

(14.4)

The upper temperature limit of oligourea greases is not determined so much by the stability of the thickener, the decomposition of which usually starts slightly below 250 °C, as by the stability of the base oil. These greases are, therefore superior to soap-based greases when the application temperatures exceed 180 °C. When an oligourea grease based on polyalkylene glycols is overheated, ideally, only gaseous products are formed. Although tetraureas also have some advantages [366], the trend in polyureas will probably be toward diureas [367]. Oligourea complex greases containing calcium acetate were introduced in 1974 [368]; others containing carbonate [369], [370] and other complementary salts followed and are still favored for some applications. Although soap-based greases cannot compete with oligourea greases at higher temperatures, below 180 °C lithium complexes, for example, have at least equal performance [371].

Carbamate-like thickeners are related to the oligoureas and simple soaps, and have properties intermediate between those of these groups.

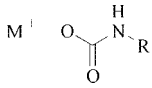

(14.5)

Other Nonionic Organic Thickeners. Polymerized perfluorinated hydrocarbons, usually micronized poly(tetrafluoroethylene)

(PTFE) powders, are the choice for greases that must function at temperatures above 220 °C, with an upper temperature limit of about 270 °C. For applications of this kind their liquid oligomers or preferably corresponding perfluoroalkylene ethers must be chosen as base oils [372–374]. Polymers such as polyamides or polyethylenes are mainly used as additives.

14.2.4. Inorganic Thickeners

To be suitable for use in lubricating oils inorganic thickeners must be treated with between 5 and 10 wt % reactive organic compounds. Without this treatment they would act like fillers, co-thickeners, and solid lubricants, that only at concentrations above ca 40 wt % yield grease-like pastes. Apart from these hydrophobic agents additional polar activators, e.g., acetone, ethanol, or, for safety reasons, propylene carbonate [375], are needed to accomplish gelation. They are used at a level of 10 wt % relative to the thickener. The thickeners themselves are stable up to temperatures of > 300 °C, but the resulting greases or gels are mainly used for applications up to 200 °C that do not require too pronounced shear stability. This is partially because of the diameter of the primary parts, only ca 0.05 µm. The tendency of inorganic thickened greases to hardening and oil separation during storage, and their sensitivity to polar additives can to some extent be overcome by addition of functionalized polymeric agents [376].

Clays. Bentonite-type aluminosilicates, mainly the smectites montmorillonite and hectorite [377], are the most important inorganic thickeners. They are usually treated with quaternary ammonium compounds, e.g., trimethyl stearyl ammonium chloride, and the activators mentioned above.

Highly dispersed silica is usually prepared by flame hydrolysis of silicon tetrachloride (→ Silica, Chap. 6.1.1.); it is made more suitable as a thickener by treating it with, e.g., silanes, silazanes, or siloxanes [378]. One advantage of these products is the low temperature-dependence of their consistency [379]. With suitable base oils and activators they form white to transparent gels that are used in food or medical applications.

14.2.5. Miscellaneous Thickeners

In principle inorganic and organic pigments of all kinds can be used as thickeners or co-thickeners or at least as fillers. The transition to lubricant additives is fluid. Only the inorganic materials soot and expanded graphite [380] and the organic phthalocyanines [381] are sometimes used on an industrial scale.

14.2.6. Temporarily Thickened Fluids

Under defined conditions liquids and suspensions of solids in liquids have significantly increased viscosities.

Some *liquid crystalline systems* are suitable as lubricants under the influence of changes in pressure [382], or temperature [383]. Some of the solutions that can form liquid crystals in limited concentration ranges [384] are comparable with consistent lubricants and some liquid crystals in concentrated point contacts are even superior [385].

Electrorheological or *electroviscous fluids*, i.e., suspensions of micronized highly polarizable and hydrophilic porous solids, are characterized by an enormous increase in apparent viscosity, when subject to electric fields. Originally silica in silicone oil was used with water as a promoter, later polyurethanes without a promoter in hydrocarbons. The first practical applications go back to 1942 [386]. electrorheological fluids are increasingly used in hydraulic valves, shock absorbers, and clutches [387], [388].

Magnetorheological fluids [389], suspensions of micronized transition elements, mainly ferrites, behave similarly in magnetic fields. Both electrorheological and magnetorheological fluids have also been named 'smart fluids'. They contain between 20 and 60 % solid particles which form more or less branched chains in the applied fields; they therefore act as Bingham plastics (see Section 14.7).

14.3. Base Oils

In principle all the lubricating oils already described in Chapters 4 and 5 can be used as base

oils for greases, but in general only oils with kinematic viscosities from 15 to 1500 mm^2/s at 40 °C are used. The oils with the lowest viscosities and the best low-temperature properties have the lowest operational temperatures at the highest speeds. The oils with the highest viscosities have the best performance at the lowest speeds and the highest loads, the lowest evaporation loss, the strongest adhesion and the best water or solvent resistance.

14.3.1. Mineral Oils (see Chap. 4)

Oils with low *VI*s usually require less thickener than those with high *VI*s and the same kinematic viscosity. The latter can be used over a wider temperature range and are, therefore, preferred nowadays. In principle the thickening effect (yield) depends on the difference between the solubility parameters of the base oil and thickener [390–393]. Because oil separation depends on the concentration of thickener, at a given temperature oil separation is most pronounced for greases based on low *VI* oils, e.g., aromatic oils or alkylbenzenes. Oil separation goes through a minimum with increasing kinematic viscosity for a given kind of base oil. This can easily be understood when only the dependence of particle interactions on the size of the involved particles is considered. Some of the specific properties of greases also depend on the ratio of the temperature and pressure coefficients of the viscosity of their base oils. If all this is taken into account, the effect of even a small change in base oil composition on the performance of a grease should not be underestimated [394].

14.3.2. Synthetic Base Oils (see Chap. 5)

Because of their cost, synthetic oils are used in greases only when the performance required cannot be achieved with mineral oils [395]. Even today greases of this kind represent less than 5 % of overall grease consumption.

Synthetic Hydrocarbons. Polyalphaolefins [396] are the ideal base oils for wide temperature-range greases. They are usually used in the same viscosity range as their mineral oil equivalents, but whereas the upper kinematic viscosity border of the former is ca 50 mm^2/s at 100 °C, polyalphaolefins and similar products are available to ca 2000 mm^2/s at 100 °C. In contrast with mineral oils they are preferably shrinking rather then swelling polymeric sealing materials. Usually, therefore, esters with softener properties must be added. Greases made from polyalphaolefins with standard viscosities are usually meant for lifetime lubrication, those made from mixtures with mineral oils (semisynthetic oils) and are used in specialized greases to increase lifetime.

Polybutenes and comparable polymers are usually applied to increase the viscosity of standard base oils; occasionally they are used as a base oil component, or even the only base oil.

Other Synthetic Base Oils. High molecular-mass esters made from diols or polyols and dicarboxylic acids with viscosities $> 2000\ mm^2/s$ at 40 °C are used as base oils in greases that must be solvent resistant.

In contrast with other greases silicone greases [397] are better distinguished according to the properties of their base oils [398]. The kinematic viscosities of silicone oils for greases range from ca 75 to ca 10 000 mm^2/s at 40 °C. Dimethylsilicone oils are physiologically inert; when thickened with highly dispersed silica they are mainly used as sealants or in applications where specific electrical and thermal conductivity is important.

Partially phenylated dimethylsilicone oils have very good low-temperature properties and are used with all types of thickener, depending on their (often military) application. Partially fluorinated dimethylsilicone oils are usually thickened with PTFE powders.

Apart from being suitable for temperatures up to 270 °C perfluorinated ethers thickened with PTFE powders are the only greases that can work in the presence of aggressive chemicals such as oxygen or chlorine. This is also their main advantage over the corresponding greases based on partially fluorinated silicone oils.

14.4. Grease Structure [399–402]

Greases have a more or less distinct lattice structure, soaps therefore form a framework of

string-like aggregates comprising micelles and crystallites that are called fibrils or fibers, the base oil is embedded in this framework. The oil is probably bound in three ways: firstly by molecular attraction between polar thickener and oil areas; secondly by capillary forces; and thirdly by mechanical occlusion. In soap-based greases small changes in the concentrations of the free acids and bases have a large effect on the shape and size of the aggregates. Even today it has not been definitively decided whether soap-based greases fit into the system of colloids (→ Colloids) or are of a more crystalline nature. Macroscopically the thickener particles are agglomerates of crystallites and micelles and dissolved components the size of which is limited by the method of homogenization. The nature of grease-like products formed with inorganic thickeners, pigments, or ferrites, is, however, not as difficult to describe. Here the laws of colloidal chemistry and physics can usually be applied fully [403].

The structures of liquid crystals are rather different (→ Liquid Crystals) [404]. Study of these has revealed much of interest in recent years [405].

14.5. Additives

Most of the additives already described in Chapter 6, can be used in greases also, and in the same way, although in general at clearly higher concentrations (Table 44). Interactions of thickeners and additives must therefore always be considered.

Table 44. Usual grease additive levels (%)

Antioxidants	0.10 – 1.00
Corrosion inhibitors	0.50 – 3.00
EP/AW additives	0.50 – 5.00
Metal deactivators	0.05 – 0.10
Solid lubricants (black)	1.50 – 3.00
Tackifiers	0.10 – 1.00

Structure Modifiers. Greases always contain compounds that affect the structure of the thickener itself to the same extent as they act as additives for the greases as a whole. These compounds are mainly water, glycerol, excess fatty acid, and excess alkali metal/alkaline-earth hydroxides. Interactions between structure modifiers and additives, e.g., that between EP additives and glycerol [406] or stearic acid [407], can severely affect the performance of a grease. Another structure modifier of greases that can easily be overlooked is the necessarily incorporated air. Though the air content must be limited, of course – often $< 3\,\%$ [408] – without air the transport of a grease through a long pipe would be almost impossible.

Antirust Additives (Corrosion Inhibitors). The most efficient corrosion inhibitor for a soap based grease is, when practicable, excess of the metal hydroxide. Sodium or calcium sulfonates, among the most powerful antirust additives, are also strong structure modifiers and can cause problems with most greases, especially greases made with inorganic thickeners. Imidazoline and sarcosine derivatives and alkenyl succinic acid esters are suitable inhibitors for soap-based greases; disodium sebacate is mainly used in gel greases. Carbodiimides have been recommended for use in ester-based greases, primarily to prevent hydrolysis [409].

Extreme Pressure and Antiwear Additives. Compounds containing heavy metals such as antimony or lead are, for reasons of toxicity, no longer acceptable as heavy-duty and antiwear additives, and even the use of zinc is disputed today, because of water pollution problems. Ashless additives, e.g., dithiocarbamates, dithiophosphates, and thiadiazole derivatives are slowly taking over [410] – the last also have corrosion-inhibiting properties [411] – and the importance of ashless alkylaminophosphorodithioates [412] as grease additives is also expected to increase.

Bismuth is one exception among the heavy metals. Additives containing this element have some of the properties of its toxic neighbors, but without their disadvantages [413], [414].

Solid Lubricants (see Chap. 15)**.** Because of their nature greases can contain these, or other additives that are insoluble in the base oil or base oil mixture, in rather high concentrations. For most applications an average particle size of the micronized solids of between 5 and 15 µm is acceptable.

Graphite and molybdenum disulfide are the oldest, the most common, and the by far best investigated solid lubricants for greases. Many manufacturers use them as mixtures, because they are believed to perform better. Fullerenes [415] have been tested successfully as solid lubricants [416], [417] but are still extremely expensive. The performance of molybdenum disulfide has been improved by the addition of zinc dithiophosphates [418] and by admixture of polyamide powders, urea resins, or salts of cyanuric acid derivatives. Admixture with PTFE powders does not give comparable results [419]. Molybdenum dithiophosphates and dithiocarbamates, rather than the disulfide, are of growing interest for joint greases [420].

Polytetrafluoroethylene, although rather expensive, is still the most common polymeric solid lubricant in greases. But high and ultrahigh molecular-mass polyolefin powders are becoming increasingly competitive.

Inorganic white solid lubricants, e.g., the hydroxides and diphosphates of some of the alkaline earth metals, improve the EP/AW properties of common greases and are especially suitable when the greases are subject to oscillations [421].

Friction Modifiers. In liquid lubricants carboxylic acids and their salts act as friction modifiers [422] and as additives for the boundary lubrication regime [423]. In greases that contain large amounts of these compounds, solid lubricants are sometimes called friction modifiers, instead [424], but sometimes polymeric additives also [425].

Some properties of greases, for example mechanical stability, can, to some extent, be improved only by use of polymeric additives; some properties, e.g., resistance against aggressive reagents cannot be improved by use of additives.

14.6. Manufacture of Greases

The properties of greases, especially metallic soap-based greases, depend not only on their composition but also, and to nearly the same extent, on the way in which the thickeners are prepared and dispersed.

14.6.1. Metal Soap-Based Greases

Batch Production with Preformed Metal Soaps. Greases can be prepared by dissolution or dispersion of preformed metallic spoaps in a suitable base oil with subsequent heating and cooling. But because of the higher costs of the thickener this method can be recommended only for highly sophisticated synthetic greases with precise chemistry or for functional base fluids that would react with the water or, even worse, with the steam generated during the neutralization process.

Batch Production with Metal Soaps Prepared In-situ. In general fatty acids or their glycerides or even their methyl esters are reacted with aqueous solutions or suspensions of the described metal hydroxides in part of the base oil. Each batch of a grease is produced by following a ten-point schedule.

1) dissolution or dispersion of the fatty acids in one third to two thirds of the base oil at temperatures up to 90 °C;
2) addition of the metal hydroxides as a solution or suspension in water;
3) heating to temperatures between 115 to 150 – or under pressure at 180 to 250 °C – depending on the type of reactor;
4) dehydration of the soap by heating to 180 to 200 °C or by pressure release;
5) crystallization during cooling to 150 to 130 °C combined with addition of more base oil;
6) addition of additives at temperatures below 80 °C;
7) (pre)filtration and homogenization with one of several possible devices;
8) adjustment of the specified consistency or flow pressure;
9) (end)filtration and deaeration; and
10) packing into containers

Different kinds of reactor or reactor combinations are used (Fig. 57). Single reactors are usually open vessels heated by gas, steam, or heating oil. Mixing is achieved with single action, double action, or counterrotating stirrers, usually designed as planetary stirrers, although turbines or dispenser disks can also be used.

Scrapers are always necessary, because of the need to move crystallizing material away from the walls to prevent a drop in heat transfer and degradation of the forming product. Sometimes the vessels can be used for evaporation or used under pressure. Since 1890, all in all the design has not changed and systematic optimization has not been reported [426]. During the reaction stage open reactors must not be heated much above 120 °C, because of the possibility of excessive foaming. Sometimes defoamers, for example low-viscosity dimethylsilicone oils, are added in small quantities during this stage.

Double or multiple reactors, the first being only a melting vessel, a Stratco contactor, or an autoclave – the last type of reactor can also be heated inductively – have some advantages in comparison with single and open reactors. The neutralization reaction is controlled more easily and the reaction time is shorter when the dehydration process is optimized [427]. The second vessel is either a finishing or a diluting vessel if several different grades are to be produced from one base grease. In each instance the main intention is to put the maximum thermal and mechanical energy into the reacting material in the minimum time. One further advantage of the multiple reactor process is that the first reactor never gets contaminated with additives. A third reactor is usually used for black, colored, or white greases.

When base oils such as silicone oils used for the in-situ production of soap greases, the thickeners are sometimes formed and dispersed in the presence of solvents which are subsequently removed by extraction or evaporation [428].

Continuous Production. Technology for the continuous manufacture of greases was first described in 1969 [429]. In principle a production unit consists of three elements only:

1) A tubular reactor – in which measured quantities of the base grease components are intensely mixed at pressures up to 2 MPa and temperatures up to 210 °C
2) A vacuum chamber in which dehydration of the product takes place
3) A finishing segment – usually a static mixer – in which oil and additives are added and homogenization takes place

The advantages of such a unit are its compact size (a few hundred liters only), its minimal energy consumption, and the uniformity of the product; the disadvantages are the difficulty of producing different ranges of products and the capacity of such units – far bigger than the need for common greases nowadays.

The finishing equipment comprises suitable pumps, filtration units, homogenization units (mills are still common although high-pressure injection homogenizers are preferable), deaeration units (vacuum pumps), and filling devices with geometries suitable for the types of containers to be filled – in a typical grease plant their range is usually from 400-g cartridges to containers holding 1 t.

14.6.2. Oligourea Greases

Oligoureas are usually prepared in a two-step process. In the first step the diisocyanate and the amine are separately dissolved or dispersed in appropriate amounts of the base oil. In the second step one of the two solutions or dispersions – usually that containing the diisocyanate – is added to the other portion [430]. The reaction is rather exothermic and, because of the toxicity of diisocyanates, the vessels must be sealed against the atmosphere during this step and subsequent heating. Because of the sudden and enormous increase in consistency very powerful stirring devices are necessary. To complete the reaction and for optimization of the thickener structure the greases are heated to 180 °C. Oligourea greases made in this way are preferably homogenized with kneaders or high-pressure injection homogenizers.

14.6.3. Gel Greases

For production of gel greases the raw materials are warmed to 60 to 80 °C only, to accelerate the activation and improve the gelation process. This is usually conducted in the same type of single reactor used for the one-step production of metallic soap-based greases. The consistency required of the product can best, and in many cases only, be achieved with the aid of high-pressure injection homogenizers or rotor – stator colloid mills [431], or, for polyester-based greases, with annular gear colloid mills [432].

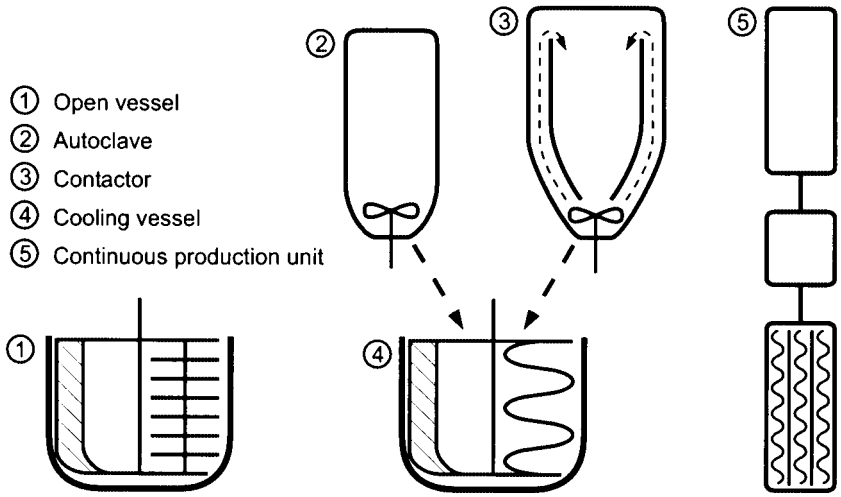

Figure 57. Reactors for the manufacture of greases

Table 45. Twenty-four properties of greases.

Twelve phenomena	Twenty-four properties
High temperature	Maximum thermal stability
	Minimum evaporation loss
	Maximum viscosity
Low temperature	No (regular) crystallization
	Minimum viscosity
Aging	Maximum oxidation resistance
	Resistance to changes in structure
Compatibility	No reaction with nonferrous metals
	Maximum corrosion inhibition
	Maximum polymer compatibility
	Immiscibility with foreign liquids
	Deflection of foreign solid matter
Oil loss	Optimum oil loss[a]
Toxicity	No toxicity
	Biodegradability
Tackiness	Optimum tackiness[a]
Flowability	Optimum relaxation[a]
	Maximum pumpability
Load	Optimum elasticity[a]
	Maximum lubricating film thickness
	Maximum emergency running properties
Shear	Maximum mechanical stability, or
	Optimum relaxation time[a]
Friction	Minimum, or optimum friction[a]
Wear	Minimum wear.

[a] According to application.

Table 46. Important ASTM test methods for lubricating greases.

ASTM D 0128-95	Analysis
ASTM D 1092-93	Apparent viscosity
ASTM D 1263-94	Leakage tendencies of automotive wheel bearing greases
ASTM D 1264-96	Water washout characteristics
ASTM D 1478-91	Low-temperature torque of ball bearing greases
ASTM D 1742-88	Oil separation during storage (air pressure method)
ASTM D 1743-94	Corrosion preventive properties
ASTM D 1831-88	Roll stability
ASTM D 2509-93	Load carrying capacity, Timken method
ASTM D 3337-94	Life and torque in small bearings
ASTM D 3527-95	Life performance of automotive wheel bearing greases
ASTM D 4049-86	Resistance to water spray

14.7. Grease Rheology (see Chap. 3)

Because of their complex rheological properties [433] greases have been described as both solid and liquid or as viscoelastic plastic solids [434]. Beyond the yield point, where flow occurs under stress, greases have an apparent viscosity which depends on shear rate, temperature, and shear time, and, to some extent, even on the mechanical pretreatment. This can be explained in simple terms, as, for example, in the NLGI Lubricating Grease Guide [435]. However, when the mathematical background of plastic flow according to

Bingham, pseudoplastic flow according to Ostwald, and liquid flow according to Newton is taken into account the subject [436], [437], a very sophisticated treatment is necessary [438]. Controlled stress rheometers seem to be the best suited devices for a better insight into grease rheology today [439] but will maybe be assisted by velocity imaging nNMR spectrometers in the near future [440].

It cannot surprise that the thickener has some influence on the rheological properties of a grease [441]. Of all the thickeners compared varieties of highly dispersed silica seem to respond best rheologically to both low and high temperatures [442]. For a series of military greases a quasi linear relationship was observed between apparent viscosity and low-temperature torque [443]. With perfluorinated polyethers thickened with PTFE it was possible to show correlation between rheological and lubrication properties [444].

New attempts to gain deeper insight into the dynamic behavior of greases include the examination of possible chaotic states [445], the observation of temperature-related changes with the aid of refractive index measurements [446], and the consideration of rheological wear [447] or the change of the storage modulus of a grease under stress [448].

14.8. Performance

The performance of a grease is, again, probably best judged in terms of the thickener [449]. From a more theoretical view there are twelve main phenomena which must be considered. Of these twelve, all but two – tackiness and toxicity – are related either to pressure or to temperature. The action or effect of pressure or temperature result in 24 requirements (Table 45). It is out of question for a real grease to satisfy all these requirements. The suitability of a grease for both high- and low-temperature applications is, for example, difficult to achieve, although the lower and upper temperature limits of a grease exceed those of its base oil, by analogy with Raoult's law. This means that the thickener acts as an impurity that, on the one hand, hinders crystallization and thus lowers the pour point of the oil yet, on the other hand, reduces its vapor pressure. The effect can be increased by use of the kinds of polymer employed as pour point depressants and viscosity index improvers in lubricating oils. The temperature-dependence of the apparent viscosity of a grease – at constant shear rates – is less pronounced than that of its base oil.

Table 47. Important DIN (ASTM) test methods for lubricating greases.

DIN 51350-4	(ASTM D 2596-97)
Testing in the Shell four-ball tester, determination of the welding load of consistent lubricants	
DIN 51350-5	(ASTM D 2266-91)
Testing in the Shell four-ball tester, determination of the wear parameters of consistent lubricants	
DIN 51801	(ASTM D 566-97), replaced by DIN ISO 2176
Determination of dropping point of greases	
DIN 51802	
Testing of roller bearing greases with regard to their corrosion-inhibiting properties, SKF-Emcor method	
DIN 51804-1	(ASTM D 217-97), replaced by DIN ISO 2137
Determination of cone penetration of greases with hollow cone and solid cone	
DIN 51804-2	(ASTM D 1403-97), replaced by DIN ISO 2137
Determination of cone penetration of greases with one-quarter cone	
DIN 51805	
Determination of yield pressure of lubricating greases, Kesternich-method	
DIN 51807-1	
Test for the behavior of greases in the presence of water, static test	
DIN 51808	(ASTM D 942-90)
Determination of oxidation stability of greases, oxygen method	
DIN 51810	
Determination of flow behaviour of greases in the rotary viscometer	
DIN 51811	(ASTM D 4048-86)
Testing of corrosive effects of greases on copper, copper strip test	
DIN 51817	
Determination of oil separation from greases under static conditions	
DIN 51821-2	
Test using the FAG roller bearing grease testing apparatus FE 9	

14.8.1. Test Methods

Many test methods are used today; all are meant to judge the single or combined and more or less complex properties of greases. The most important ASTM and DIN methods are described in Tables 46 and 47. Complete collections are published regularly [451], [452], the same is true for the French AFNOR, the English IP, the Japanese JIS, and some other national collections. A description of Russian GOST and former East German TGL methods is given in [450]. The devel-

opment of international standards (EN and ISO) is slowly proceeding.

The methods themselves are described in Chapter 16.

14.8.2. Analytical Methods

Elemental analysis of greases is nowadays performed by spectroscopic methods, e.g., X-ray fluorescence spectrometry (XRF), inductively coupled plasma atomic emission (ICP), or atomic absorption spectrometry (AAS) [453], with attention being directed mostly to methods of preparation [454], [455]. IR spectroscopy was introduced as a means of identifying greases and their components ca 40 years ago [456–458]. Its use has steadily grown since and been extended to questions of structure, development, and manufacture also [459]. NMR has also been used to investigate structural questions [460] and electron microscopy has enabled not only the scrutiny of soap fibers [461] but also study of thermal changes in greases [462].

The use of thermogravimetry (TG) is usually limited to investigations of base oils [463], but the other thermoanalytical method, differential scanning calorimetry (DSC) [464], [465], or a combination of both, is expected to become a valuable tool in the analysis of grease and antioxidant analysis.

Chromatographic methods, e.g., gas (GC) and high-pressure liquid chromatography (HPLC), are mainly used to identify the components of liquid or liquefied grease [466], [467].

14.9. Applications

The selection of a grease for a particular application is always a compromise between the demands of a customer and the circumstances the grease must face during its operational life – temperature, speed, load including centrifugal forces and vibrations, and relubrication intervals.

14.9.1. Roller Bearings

The servie life of roller bearings is connected with that of the grease used, especially under extreme conditions [468]. Sophisticated test equipment has been developed, mainly to ensure better lifetime predictability for the selected bearing together with the selected grease [469], [470]. The test rigs SKF R2F and later FAG FE 9 have found great acceptance in Germany [471], [472].

Clean production and low-noise greases are meeting reliability and lifetime requirements of roller bearings. The stringent requirements of high-precision bearings, for small bearings in video and audio applications, and bearings for military use have been established for years [473], [474].

In general greases used in roller bearings such as ball-, deep groove ball-, thrust ball-, spherical-, taper-, cylindrical- or needle roller bearings must have good working stability. This can be checked by the prolonged penetration test and the Shell-roller test (ASTM 1831) [475]. Conventional lithium greases in the NLGI class 2 are recommended for most types of bearing at working temperatures up to 120 °C. Greases of the NLGI class 1 are preferred for needle bearings. For bearings exposed to temperatures above 120 °C complex soap or polyurea thickened greases are preferable. For bearings that must operate under high load and/or low speed the base oil viscosity must be 200 mm/s^2 minimum at 40 °C. Greases for cold climates or for aerospace or military use have to ensure performance down to below −70 °C. Low-temperature performance can be checked with the low-temperature torque (ASTM 1263), the low-temperature penetration (AFNOR NF T 60-171) and the flow pressure (DIN 51805) tests [476–478]. Greases of that kind need base oils with sufficiently low pour points. In many military applications long-life properties are also required; these can be fulfilled by use of synthetic base oils only [479]

Because plain bearings are often exposed to moisture or water, calcium soap-thickened greases are recommended. When open housings are used in a dusty atmosphere frequent relubrication makes it possible to wash out the contaminated grease.

Relubrication Intervals. The lubrication interval t_f is based on the F_{10} value of a standard grease as agreed in DIN 51 825 [480] under normal environmental conditions with temperatures up to 70 °C and a mean bearing load of

$P/C < 0.1$. Figure 58 and Table 48 takes into consideration the type of bearing and the speed. For each 15 K temperature increase the relubrication interval is said to be reduced by 50 %. Severe working conditions reduce the lubrication interval t_f to a reduced lubrication interval t_{fq}.

$$t_{fq} = f_1 \times f_2 \times f_3 \times f_4 \times f_5 \times t_f \quad (16.1)$$

Table 48. Relationship between bearing type and correction factors for relubrication intervals (GFT Worksheet 3)

Bearing type		k_f
Deep grove ball bearings	single row	0.9–1.1
	double row	1.5
Angular contact ball bearings	single row	1.6
	double row	2
Spindle bearings	$\alpha = 15°$	0.75
	$\alpha = 25°$	0.9
Four point bearings		1.6
Self-aligning ball bearings		1.3–1.6
Thrust ball bearings		5–6
Angular contact thrust ball bearings	double row	1.4
Cylinder roller bearings	single row	3–3.5[a]
	double row	3.5
	full complement	25
Cylindrical roller thrust bearings		90
Needle roller bearings		3.5
Tapered roller bearings		4
Barrel roller bearings		10
Spherical roller bearings without lips (E-design)		7–9
Spherical roller bearings with center lip		9–12

[a] $k_f = 2$ for radial load or increasing thrust load and $k_f = 3$ for constant thrust load.

In relubrication it is usually impossible to remove the used grease. Consequently the relubrication interval t_{fq} must to be reduced by 30 to 50 %.

14.9.2. Cars, Trucks, Construction Vehicles

Most modern cars do not need relubrication, with the exception of door hinges, lock mechanisms, and battery poles. But among the approximately 30 hidden greases (Fig. 59) in a modern car [481], [482] only the constant-velocity (CV) joint greases are required in substantial quantities. Although improved conventional lithium greases containing molybdenum disulfide are still in use [483], lithium complex or polyurea greases are already preferred in some modern cars and this usage will increase in the future [484]. Most of the greases used in cars, for example the greases for CV joints, hub units, starters, alternators, seat adjustments, clutches release bearings, belt – pulley bearings, window levers and windshield wiper gears, are specified and approved by the large motor companies and developed in close cooperation with the grease manufacturers. For the same application, however, different motor companies have different grease specifications and approvals; for example, most European and U.S. motor companies prefer lithium complex greases in the front wheel bearings, whereas Japanese manufacturers prefer polyurea greases.

ASTM D 4950 (see below) describe the minimum requirements of current greases in automotive service – fill applications for passenger cars, trucks, and other vehicles operating under various service conditions [485].

Chassis grease classifications – intended use of chassis (L) classified greases:

LA Classification
Chassis components and universal joints under mild duty
- Frequent relubrication
- Noncritical applications

LB Classification
Chassis components and universal joints under mild to severe duty
- Prolonged relubrication intervals
- High loads
- Severe vibration
- Exposure to water or other contaminants

Intended use of wheel bearing (G) classified greases:

GA Classification
Service typical of wheel bearings operating under mild duty
- Frequent relubrication
- Non-critical applications

GB Classification
Service typical of wheel bearings operating under mild duty/moderate duty
- Normal urban, highway and off-highway service

GC Classification

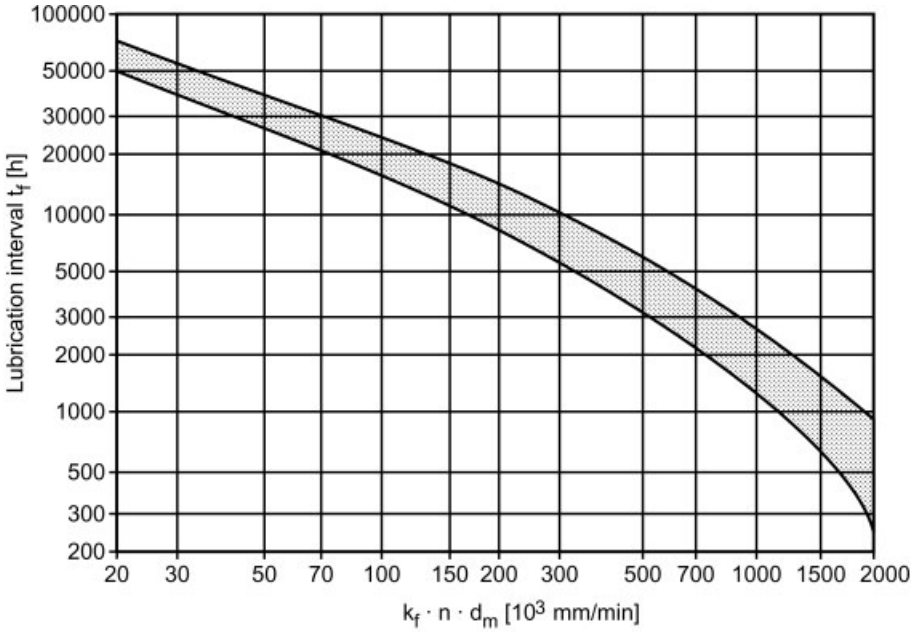

Figure 58. Re-lubrication intervals

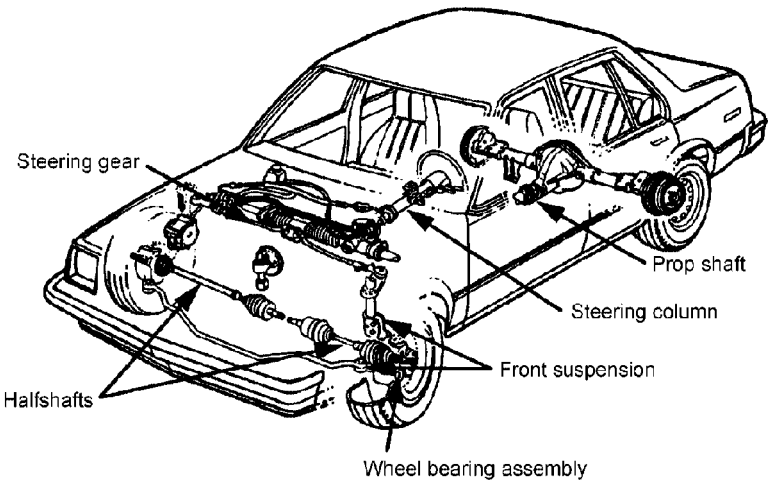

Figure 59. Greases hidden in a modern car

Service typical of wheel bearings operating under mild duty/severe duty
- High bearing temperatures
- Frequent stop and go service (buses, taxis, police)
- Severe braking service (trailer towing, heavy towing, mountain driving)

Lithium-based multipurpose greases have replaced several other greases for the relubrication of trucks and construction equipment. Conventional lithium soap based greases that require frequent relubrication are still in use for the wheel bearings of trucks and trailers. Modern trucks and trailers with prolonged oil-drain intervals require lithium complex greases with semisynthetic or fully synthetic base oils. Lithium greases containing black solid lubricants are recommended for 5th-wheel applica-

tions, for chassis points, and for plain bearings of construction equipment. Many trucks and buses and much construction equipment uses centralized lubricating systems, designed for semifluid greases of NLGI-class 00 or 000, for onboard relubrication. Other systems require greases of NLGI class 2. Lithium greases optimized for low temperature applications, good pumpability, and low oil separation are recommended.

14.9.3. Steel Mills

In Europe calcium complex and sometimes polyurea greases based on mineral oil are used for lubricating continuous casting equipment. The U.S. market prefers aluminum complex and polyurea greases and Japanese equipment manufacturers mainly recommend polyurea greases. Some roller bearing manufacturers equip continuous castings with double- or triple-sealed bearings that are greased for life, preferably with synthetic polyurea greases. When the sealing of the bearings is not perfect relubrication has a cleaning function. Contaminated grease that could result in limited lifetime is squeezed out.

Conventional EP lithium greases, calcium complex greases, calcium sulfonate complex greases, lithium complex greases, and aluminum complex greases, all based on mineral oils are used in hot rolling equipment. Most customers require a minimum base oil viscosity of 200 mm/s^2 at 40 °C [486]. The rotating shift system is typically used for hot-rolling lubrication. In general the bearings are not regreased during operation, only during maintenance.

Many different greases, e.g., conventional EP lithium greases, lithium complex greases, calcium complex greases, and calcium sulfonate complex greases, all based on mineral oil, are used in cold rolling equipment. Modern greases use a lithium – calcium mixed soap for improved water resistance.

14.9.4. Mining

Open pit mines which use wheel bucket excavators and belt systems to ship the spoil and coal are especially large grease consumers. Lithium or lithium – calcium mixed-soap-based grease with base oil viscosities > 350 mm/s^2 at 40 °C and an effective EP and AW additive package are used. Black solid lubricants are also recommended. Because the equipment is exposed to dust or water and mud, the sealing efficiency of the bearings determines their lifetime and must be supported by a grease.

Kilometers of wire rope are used on the excavators or drag lines and in the underground mines. During production of the wire ropes lubricants are applied to ensure corrosion protection and to minimize friction of the single wires when the rope stretches under load.

Lay-up lubricants based on wax resin are applied during manufacture to ensure lubrication of the individual wires and stands.

During operation in some countries the wire ropes are maintained by cleaning the surface and relubrication. In many countries bitumen-based products are still in use, but are being replaced by bitumen-free greases or even biodegradable lubricants.

Winch greases, preferable lithium complex greases with improved water-resistance, ensure prolonged winch lifetimes up to several years.

14.9.5. Railroad, Railway

Depending on the design of the driving system of the locomotive a gear oil or a gear grease is required. The poor sealing properties of these kind of gears leads usually to leakage; this is usually minimized by use of bitumen-based products. Modern greases are thickened with lithium or sodium soap and based on mineral oils of up to 2000 mm/s^2 at 40 °C. Traction motor gear greases are not described in terms of consistency; the apparent viscosity is usually checked with a Brookfield viscometer, for example with a number 3 spindle at 93 °C and at 4 rpm and giving results of 5000 to 10 000 mPa · s [487].

The axle bearings are greased with conventional lithium EP greases. The development of ready-to-build in axle boxes and the increasing speed of modern trains have led to improvement of the high-temperature performance and lifetime of greases.

Switch lubrication, wheel flange lubrication (mainly used in Europe), and rail track-side lubrication in curves, used mainly in the USA and Canada, cause environmental problems. For these applications biodegradable greases based

on esters have better wear-protection and consumption performance as conventional greases [488–490]. These greases must be sprayable because of the means of application.

14.9.6. Gears

Fibrous sodium soap greases of NLGI-class 0, 00, or even 000 consistency are used. Although the load-carrying capacity, antiwear properties, and adhesion of these greases is especially good, lithium-based greases and greases based on synthetic hydrocarbons or polyglycols are gaining in importance

Girth gear drives are a type of drive widely used for many large systems in the primary industries [491]. These open gears require an adhesive grease based on high-viscosity oils. Such greases are usually sprayed on the teeth by means of a centralized lubricating system.

14.9.7. Food-Grade Applications

Greases for lubricating machines used in food processing must fulfil specific requirements relating to food legislation, human health protection, taste, and odor. These properties must be approved by internationally accepted tests and organizations [492].

The FDA lists ingredients permitted for use in food-grade grease formulations (21 CFR §178.3570); white oils are listed in 21 CFR §178.3620. The USDA (United States Department of Agriculture) has approved the finished products. For white and synthetic oil-based products maximum contamination of 0.1 g/kg food is allowed. Since USDA stopped ist activities in 1999 a working group of ELGI, NLGI and the European Hygienic Equipment Design Group (EHEDG) was founded with the aim to install an ISO Standard for food grade lubricants.

14.9.8. Textile Machines

Because greases could contaminate textiles, they are formulated with white oils and thickened with water-soluble sodium soaps, and are thus removable by washing.

14.9.9. Applications with Polymeric Materials

There are three ways in which greases and polymeric materials can come into contact: polymers can be contained in greases as thickeners, as solid lubricants, and as additives; they can act as sealants; and they can form one or even both of the solid partners in friction couples. Although base oils are the primary consideration [493], when the interaction of greases with sealants is important [494], [495] – the influence of additives, and among these especially EP additives, must always be considered [496]. Even migration of soap molecules into a polymer can occur [497].

An early test method for the action of lubricants on polymeric materials was the pendulum according to Barker [498]. During the automation of this process a more meaningful friction and wear tester was developed [499], [500]. It proved also to be necessary to test the tendency of a polymer to form cracks under the influence of its lubricant and tension [501].

The lubrication of polymers or polymers and metals as friction couples is one field in which silicone greases and, better, fluorinated silicone greases, or, even better, greases based on perfluoropolyalkyl ethers are of some advantage. Although silicone greases made with highly dispersed silicic acid should be used as general lubricants and as sealants [502], silicone greases made with the usual thickeners are sometimes preferable to other synthetic greases. With calcium stearate or 12-hydroxystearate, for example, they are permitted in food applications [503].

14.10. Ecology and the Environment

Pre-industrial greases consisted of tallow or vegetable oils and their reaction products with lime; they were therefore, not merely ecologically compatible lubricants, but even rapidly biodegradable. When mineral oils took over, most of the thickeners and some of the additives remained ecologically acceptable. The ecological and environmental aspects of modern base oils have been described in detail in Chapter 7. Of all the vegetable oils that could

be chosen as base oils for greases, in Central Europe only rape seed oil has been introduced on a technical scale [504] and only for loss-lubrication applications. Initially calcium 12-hydroxystearate was chosen as the most suitable thickener [505]. Greases of that kind proved superior to lithium- and mineral oil-based general-purpose grades (Fig. 60). The limitations of vegetable oil-based greases soon led to the introduction of transesterified vegetable oils [506] and synthetic esters (see Chap. 5). Because of industrial demands greases based on calcium 12-hydroxystearate were soon replaced by lithium and lithium – calcium greases [507–509]. Polyurea greases [510], aluminum complex greases [511], and titanium complex greases [512] followed.

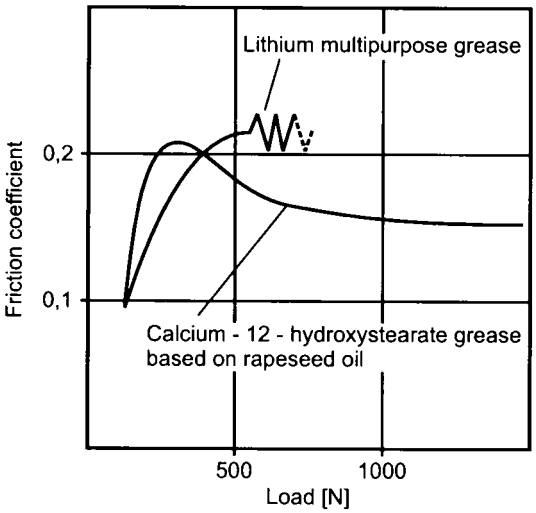

Figure 60. Friction – load behavior of rapeseed oil-based calcium 12-hydroxystearate greases

The biodegradability of greases essentially reflects the biodegradability of their base oils. In round robin tests greases based on rape seed oil and calcium 12-hydroxystearate achieved biodegradability close to 100 %. Methods enabling the use of emulsifiers led to significantly lower results [513].

The elimination of toxic components such as polycyclic hydrocarbons from base oils is almost complete. This is also true for chlorine substitution. The substitution of antimony and lead cations and nitrite anions in additives is not yet complete, because it is not easy to maintain the same performance level. In Germany the demand for environmentally harmless products [514] became an additional aspect of the performance of greases; changes in the demands for water pollution class (WGK) 0, that came into force in 1999 aggravated the situation. The most suitable antioxidants are only accepted in concentrations which are too low, hydrocarbons including polyalphaolefins, white oils, and low-viscosity polybutenes are not accepted and the most suitable soaps are also excluded.

Although experiences with biodegradable greases are promising [515], [516], some problems, mainly relating to legal aspects of waste management [517] and the balance between ecological requirements and economical possibilities [518], remain unsolved. Therefore the third and possibly most important aspect of applying greases in a manner that is both environmentally responsible and economical is to minimize the amount of lubricant necessary and to maximize lifetime by using the most inert materials [519], [520].

14.11. Grease Tribology

General aspects of friction and wear [521], and of lubricating oils and their viscosity [522], have been described in detail in Chapters 2 and 3. Recent tribological research on greases has dealt with the time-dependence of their behavior under stress as a fourth factor in addition to temperature, pressure, and thickener properties [523]. Behavior in EHL contacts, where the thickness of grease films can exceed that of their base oil under comparable conditions [524], [525], and the behavior under starved conditions, when the load is supposed to be carried by thickener particles deposited in the center of the contact [526], has also been of concern. Soap deterioration is another important subject [527], [528].

15. Solid Lubricants

Solid lubricants are required for lubrication under extreme conditions where the bearing surfaces in tribological contact must still be effectively separated. The life of lubricated machine parts depends on the functional and tribological design and optimization of lubricants as calculable functional elements. To make a systematic

choice of a suitable lubricant it is absolutely necessary to understand the relationship between friction, wear, and lubrication and the interaction between the elements of the tribological system and the specific properties of each element.

Special attention must be given to the different stress factors and the structure of the tribological system (see Chap. 2).

Products containing solid lubricants are often used to solve problems, particularly in boundary and mixed frictional states when high specific loads are applied to sliding surfaces and at very low hydrodynamically effective speeds but also in critical applications where, e.g., the lubricant must perform over a wide temperature range or under extreme temperature conditions, for example in aviation and in rocket technology.

Dry lubrication with solid lubricants is also required in nuclear reactors, in high vacuum applications, in aggressive environments, and in applications where contamination by lubricating oils or greases cannot be tolerated.

Table 49. Structural lubricants.

Name	Formula
Graphite	C
Graphite fluoride	$(CF_x)_n$
Molybdenum disulfide	MoS_2
Molybdenum diselenide	$MoSe_2$
Tungsten disulfide	WS_2
Tungsten diselenide	WSe_2
Niobium disulfide	NbS_2
Niobium diselenide	$NbSe_2$
Tantalum disulfide	TaS_2
Tantalum diselenide	$TaSe_2$
Titanium disulfide	TiS_2
Titanium diselenide	$TiSe_2$
Titanium telluride	$TiTe_2$
Cerium fluoride	CeF_3
Barium hydroxide	$Ba(OH)_2$
Cadmium chloride	$CdCl_2$
Cobalt chloride	$CoCl_2$
Zirconium chloride	$ZrCl_2$
Lead chloride	$PbCl_2$
Lead iodide	PbI_2
Boron nitride	BN
Silver sulfate	Ag_2SO_4
Borax	$Na_2B_4O_7$
Talc	$Mg_3(OH)_2Si_2O_{10}$
Mica	$KAl_2(Si_3Al)O_{10}(OH)_2$

15.1. Classification

A solid lubricant is often defined as any solid material which reduces friction and/or wear of contacting surfaces in relative motion. Various systems are used to classify the different types of solid lubricant. An arbitrary, but useful, classification is into structural lubricants, mechanical lubricants, soaps, and chemically active lubricants.

15.1.1. Class 1: Structural Lubricants

The most widely used solid lubricants are graphite and molybdenum disulfide. Their satisfactory lubricating properties are assumed to result from their layered lattice structures. In addition to these two substances, other solids are used, for example metal halides and sulfides, which have, in the main, inherent lubricant properties, a lamellar hexagonal crystal structure, and are usually anisotropic (Table 49).

In contrast to graphite and molybdenum disulfide these are not yet extensively used in industry.

15.1.2. Class 2: Mechanical Lubricants

There are different types of substances within this class whose lubricating effect is based on different physical and mechanical properties or special conditions. These lubricants are divided into self-lubricating substances, substances that need a supporting medium to create lubricating properties, and substances with indirect lubricating properties based on their hardness.

Self-Lubricating substances can be classified as organic compounds, metal films, chemical surface layers, and glasses.

Organic Compounds. The most important self-lubricating organic substances are liste below:

Linear polymers (thermoplastic types)
Poly(tetrafluoroethylene) (PTFE)
Fluoroethylene – propene (FEP) copolymers
Perfluoroalkoxy polymers (PFA)
Polyethylene (PE)
Polypropylene (PP)
Polyurethane (PU)
Polyamide
Polyacetals

Polyterephthalate
Polysiloxanes
Nylon
Cross-linked polymers (thermosetting types)
  Phenol – formaldehyde resin
  Urea – melamine – formaldehyde resin
  Epoxy resin
  Phenolic resin
  Unsaturated polyester resin
Polyimides

The sliding characteristics of polymers depend on their chemical nature and the mating partner. The permissible load is a function of heat dissipation. Temperature changes have little influence on the friction characteristics of polymers. Addition of MoS_2, graphite, and metal powder improves the frictional characteristics of the polymers and increases their hardness. Graphite also increases the elasticity module of PTFE.

Metal Films. Important self-lubricating metal films include:

Lead
Tin
Silver
Indium
Barium
Gold
Aluminum
Nickel (Ni and Ni – Cr alloys)
Copper (Cu and Cu alloys)
Zinc (Zn and Zn alloys)

Friction can be reduced by the coating of the body material surfaces with a thin film of a soft metal, because the friction depends on the shear strength of the soft metal film. The durability depends on the film hardness, homogeneity, and adhesion. The lubricating effect of soft metal layers is limited by their melting point.

Chemical Surface Layers (Conversion Films). In addition to the naturally occurring oxide films present on the surface of most base materials exposed to air, other solid lubricant films can be formed by chemical or electrochemical action on the metal surface.

Chemical surface coatings such as zinc, iron, or manganese phosphates behave similarly to soft metal coatings but consist not of a metal but of metal salts. Bonderization (phosphate treatment) creates a thin, microcrystalline, strongly adhering phosphate layer on the metal surface; this reduces the coefficient of friction and the danger of seizure during the running-in period. The lubricating efficiency of the layer, which is normally 2–5 µm thick, is based on its lower shear strength in comparison with the metal.

Glasses (→ Glass). The lubrication properties of glass depend on the composition. The coefficient of friction at a given temperature is a function of the viscosity, the thermal conductivity, the rate of shear, the area of shear, the capacity to dissolve different amounts of oxide from the surface of the material to be lubricated, and the contact angle between the glass and the material, because this determines the capacity of the glass to wet the material.

The importance of glasses used as lubricants is specially to be seen in metal-forming operations with operating temperatures up to ca. 1500 °C.

Substances with lubricating properties that need a supporting medium can be classified as inorganic compounds and metal powders.

Inorganic Compounds (Table 50). The coefficient of friction of natural oxide films on metals, which are usually approx. 10 nm thick, has been investigated. It is generally accepted that the oxide film reduces surface damage, makes sliding smooth, and often reduces friction.

Table 50. Inorganic compounds needing a supporting medium

Metal sulfides	ZnS, SnS_2, FeS, ...
Metal fluorides	CaF_2, LiF, ...
Metal phosphates	$Zn_2P_2O_7$, $Ca_3(PO_4)_2$, $Fe_2P_2O_7$, ...
Metal hydroxides	$Ca(OH)_2$, $Mg(OH)_2$, $Zn(OH)_2$, ...
Metal oxides	PbO, ZnO, FeO, Fe_2O_3, ...

Sulfides, fluorides, phosphates and hydroxides are claimed to act as a supporting agent or a catalyst by producing friction- and wear-reducing layers. Calcium hydroxide, for example, supports the production of a layer of Fe_3O_4 iron oxide on the rubbing surface of steel.

The process of formation of these layers depends on the chemical composition of steel and, in particular, on its surface chemistry.

Phosphate layers, by galvanic techniques, are mainly used as precoatings for dry-film application and as lubricant carriers in cold metal-forming processes. In addition to acting as a lubricant carrier, the phosphate coating can be

plastically deformed with the steel slug and, therefore, in conjunction with the lubricant, prevents metal – metal contact and thus reduces surface friction and wear.

The three main types of phosphating solutions contain zinc, iron, and manganese phosphates, and of these the zinc phosphate is probably the most widely used.

Metal Powders. In contrast to structural lubricants and self-lubricating mechanical lubricants, the lubricating properties of the other mechanical lubricants are mainly based on the supporting effect of a carrier substance or a binder. The main purpose of these substances, which include Pb, Sn, Zn, Cu, Ag, and In, is to improve the adhesion and cohesion properties of the non-self-lubricating mechanical lubricants.

Substances with Indirect Lubricating Properties Based on their Hardness (Physical Vapor Deposition, Chemical Vapor Deposition, and Diamond-like Carbon Layers). Vapor-deposited (VD) coatings of oxides, borides, nitrides, or silicates, for example TiN, TiNC, CrN, ZrN, and AlTiN, usually have quite high coefficients of friction but can prevent seizure even at high temperature and offer excellent wear resistance.

Another relatively new form of thin-film layer with proven good tribological properties is diamond-like amorphous carbon (DLC) (→ Thin Films, Chap. 2.3.). Hardness values for DLC range from 500 to 13 000 HV depending on the amount of hydrogenation, from 50 % to zero. In addition to hard coatings there are also soft physical vapor-deposition coatings containing MoS_2 or MoS_2 – metal composites. These films, normally deposited by various sputtering technologies like ion-beam-assisted deposition (IBAD) or closed-field-magnetron sputtering (CFMS), afford excellent tribological properties in terms of wear-resistance and low coefficient of friction. They are used for cutting and forming tools, ball bearings, and sliding contacts with steel and ceramics.

15.1.3. Class 3: Soaps

Soaps often give the lowest coefficients of friction obtainable with solid lubricants but in general cannot be used above their melting points or at high loads.

The main function of soaps in lubrication technology is in the preparation of greases (see Chap. 14). The main use of soaps as lubricants in their own right depends on their formation in situ on a metal surface, by the chemical attack of a fatty acid on the metal.

15.1.4. Class 4: Chemically Active Lubricants

This category includes EP and AW additives and a variety of chemicals which interact with the metal surface to produce a lubricating or protecting layer (see Chap. 6).

15.2. Characteristics

15.2.1. Crystal Structures of Lamellar Solid Lubricants

Graphite (→ Carbon, Chap. 3.1.). In the hexagonal graphite structure each carbon atom is linked to its nearest neighbors by sp^2 type σ bonds. The strength of the latter is increased above the normal single covalent C – C bond by the interaction of non-localized electrons associated with the $2p_z$ orbital of the planar carbon atoms. It is this extra bonding which gives graphite its thermal stability. The planes themselves are held together largely by van der Waals forces which are much weaker than the planar bonding forces and the strength of bonding to a substrate.

The low friction of graphite is not based solely on its crystal structure. It also depends on adsorbed films, particularly of water vapor, which provide surfaces with low cohesion. Thus, graphite develops its most favorable lubricating properties in the presence of moisture.

Molybdenum Disulfide (→ Molybdenum and Molybdenum Compounds, Chap. 7.3.). The crystal of molybdenum disulfide has a hexagonal structure with trigonal symmetry. Each Mo atom is surrounded by a trigonal prism of sulfur atoms.

The lubricating effect of MoS_2 is also based on its crystal structure. The weak bonds (van der Waals forces) between the sulfur layers enable

easy movement of the lamellae over each other, resulting in a low friction between the sliding parts. On the other hand, the ionic bond between Mo and S gives the layers a high strength so that they can resist the penetration of surface asperities.

15.2.2. Heat Stability

Heat stability is required if thermal decomposition of the solid lubricants is to be avoided, especially in the presence of oxygen-, carbon dioxide-, or water-containing atmospheres. Very often the decomposition products are either corrosive vapors or abrasive solids.

When heated in air, graphite forms carbon dioxide above 450 °C; this escapes without leaving any grinding particles behind. MoS_2 is attacked by oxygen even at 20 °C. In air, appreciable oxidation to molybdenum trioxide occurs above 400 °C and effects the lubricating properties and the adhesion of MoS_2 to metal surfaces. When oxidation can be excluded, MoS_2 retains its lubricating properties up to 800 °C. Lead monoxide is an excellent lubricant up to 700 °C. Metal-free phthalocyanines have been established as very efficient lubricants up to temperatures of 800–900 °C. Glass has proved to be an efficient lubricant at extreme loads and temperatures up to 1200 °C.

Other solids, for example metal oxides, sulfides or phosphates, also afford excellent heat stability. Without a supporting liquid, however, these solids provide no lubricating properties but build up separating layers only. The lubrication effect is, therefore, limited by the heat stability of the carrying substance.

The melting point of a solid lubricant can be regarded as a physical aspect of heat stability.

15.2.3. Thermal Conductivity

Frictional heat, developed by rubbing of material surfaces at the high spots, is also transmitted to the solid lubricant, which must be capable of dissipating this heat as quickly as possible – or local welding of the material might occur, despite the presence of the solid. This is particularly true of plastic solid lubricants, which all have poor thermal conductivity.

15.2.4. Adsorbed Films

Adsorbed film of gas or vapor might be attached to the external surface of the solid or chemically to an internal surface (intercalation). Physically or chemically adsorbed molecules on the external surface of the solid can influence:

- The sintering of the aggregates of the particles under the action of heat;
- The ease of breakdown of particles under shear stress; and
- The wear properties of the edges of lamellar solids.

The chemisorbed molecules within the particles can affect:

- The decomposition or oxidation temperature of the solid;
- The mechanical properties of the solid; and
- The conductivity (thermal and electrical) of the solid.

15.2.5. Chemical Stability

Chemical stability is closely linked with heat stability. When solid lubricants are to be used intermittently over long periods in relatively inaccessible places, tendency to corrosion can be serious. Corrosion can occur in three ways:

Interaction of the solid with the metal surface;
Action of heat and/or the atmosphere on the solid to give decomposition products, which then react with the metal surface; and
Action of heat and/or the atmosphere on material chemisorbed on to the solid lubricant producing material which has a corrosive action on the metal.

15.2.6. Particle Size

The smaller a nonporous particle, the more active it is. Although such particle activity is required in some applications there are other fields, for example many metal-forming operations, for which it is desirable to have a relatively large particle size range to give the required film properties. With graphite and molybdenum disulfide, relatively coarse particle sizes are advantageous because this results in optimum resistance to oxidation.

15.3. Products Containing Solid Lubricants

15.3.1. Powders

To ensure that solid lubricants in the form of powder give sufficient coverage in a tribological system.

The following requirements must be fulfilled

1) The level of adhesion between the lubricant film and the surface of the material must be great enough to ensure that this lubricant film adheres to this surface when it is subjected to friction.
2) The internal cohesion of the film must be sufficiently large that the film does not split when subjected to friction.
3) The adhesion between the particles and layers in the shearing direction should be as small as possible to keep the resistance to friction low.

These main requirements can be met only by self-lubricating dry lubricants especially by MoS_2 lattice; it is, as a result, the most commonly used. Other solid lubricants which are applied in powder form are organic compounds such as PTFE and graphite, although these fulfil the requirements listed above to a limited extent only.

Before the solid lubricant powder can be applied, the surface of the material must be thoroughly cleaned. Roughing the surface mechanically or with phosphates improves adhesion and therefore the lifetime. The lubricant can be effectively applied by simple rubbing with cloths, sponges, brushes, or polishing pads or polishing buffs, by applying it using suitable carriers, and coating by cathode evaporation in an ultra-high vacuum.

Solid Lubricants in Carrying Media. Most substances used as solid lubricants require a carrying medium, a bonding agent and/or pretreatment of the material surface, to help create, or improve, their adhesive and film-creating properties. The substances used as bonding agents are: organic binding agents, resins; inorganic binding agents, silicates; and oils, greases, water.

It is also advantageous to pretreat the surfaces by: degreasing; sandblasting, corroding, etching; phosphating, anodizing; and activating (e.g., low-pressure plasma).

15.3.2. Dispersions and Suspensions

Dispersions and suspensions in carrying liquids with low volatility are mainly used in areas in which, for tribological reasons, a dry lubricating film should be created, but where effective application of a powder is not possible for technical reasons. The same types of solid lubricant are used here in the same way as for powders.

Dispersions and suspensions of solid lubricants in water are usually used to coat mass elements for cold and hot forming. The most commonly used substances here are salts, special white solid lubricants, and graphite.

Dispersions and suspensions in oils also act as aids in forming techniques, and they are also used as additions in gear- and oil-lubricating systems. The solid lubricants used here in forming techniques are the same types as those used as dispersions and suspensions in water. MoS_2 plays a dominant role as an additive to lubricating gears and for general use in oil lubricating systems.

15.3.3. Greases and Grease Pastes

Addition of solid lubricants to greases is primarily intended to have a positive effect on their capacity to absorb pressure, and the ability to withstand wear and tear, and friction. The specific advantages of solid lubricants in comparison with oil-soluble, chemically reactive additives is that they react neutrally to many types of plastic and elastomer, perform well at high temperatures and have good safety reserves with regard to emergency running properties, which come into action when the layer of grease collapses. The use of black solid lubricants such as graphite and MoS_2 also have a beneficial effect on running-in processes and the capacity to withstand wear and tear in boundary and mixed friction areas. Where oscillating movements or vibration is involved, white solid lubricants have the advantage. These effects are illustrated in Figs. 61 and 62.

Solid lubricants in greases are used at concentrations of 1–3 %. When this proportion rises to

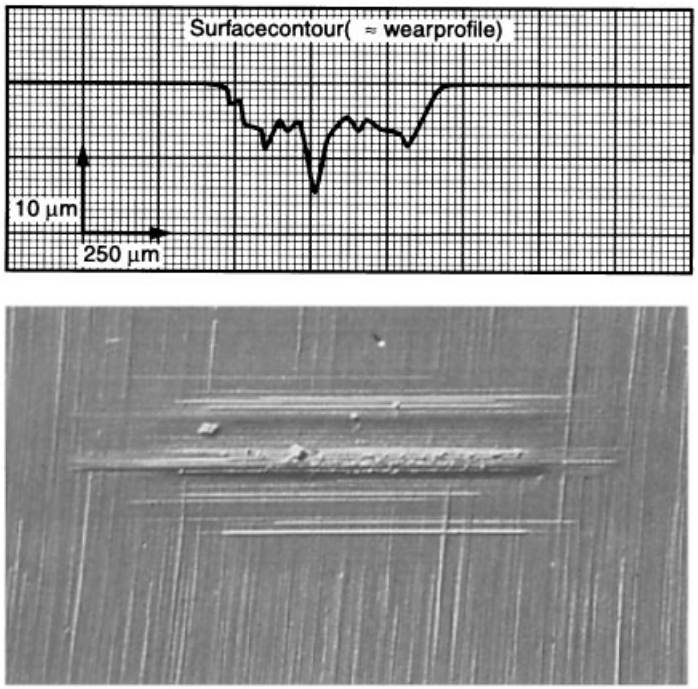

Figure 61. Lithium grease with molybdenum disulfide, showing the irregular pattern of the wear profile, the chipping of the material, and serious wear. The maximum wear depth exceeds 13 µm

over 10 % they are known as grease pastes, because the solid lubricant at this concentration has a noticeable thickening effect.

High-performance lubricating greases usually contain combinations of two or more substances, which have a synergistic effect. This applies both to the combination of graphite and MoS_2 and to mixtures of different inorganic solid lubricating substances, for example calcium hydroxide, zinc phosphate, or iron sulfide.

15.3.4. Pastes

Pastes are solid lubricants in a carrying oil. To achieve the correct consistency, the proportion of solid lubricant must be at least 40 %. To reduce the undesirable 'bleeding' effect, modern formulations often include a small amount of soap.

Black pastes usually contain graphite or MoS_2 and combinations of both. Because of their high load-bearing capacity, these pastes are particularly useful when movement with very slow relative speeds is involved, e.g., assembly processes, and running-in processes.

Pastes prepared from combinations of white solid lubricants are especially suitable for use where oscillating movements or vibration is involved. White pastes are particularly good at preventing the wear and tear which results from vibration and are proven to afford excellent protection against fretting corrosion.

Another area where pastes are traditionally used is for lubrication of screw connections. Special screws made of steels which are resistant to high temperatures, based on chrome/nickel alloys, are prone to seizure because an oxide layer is not created.

The pastes used in this particular application area are usually made up of combinations of solid lubricants and special metal powders. More recent developments show that the use of formulations which do not contain metal, and are based on white solid lubricants, perform well when used at high temperatures.

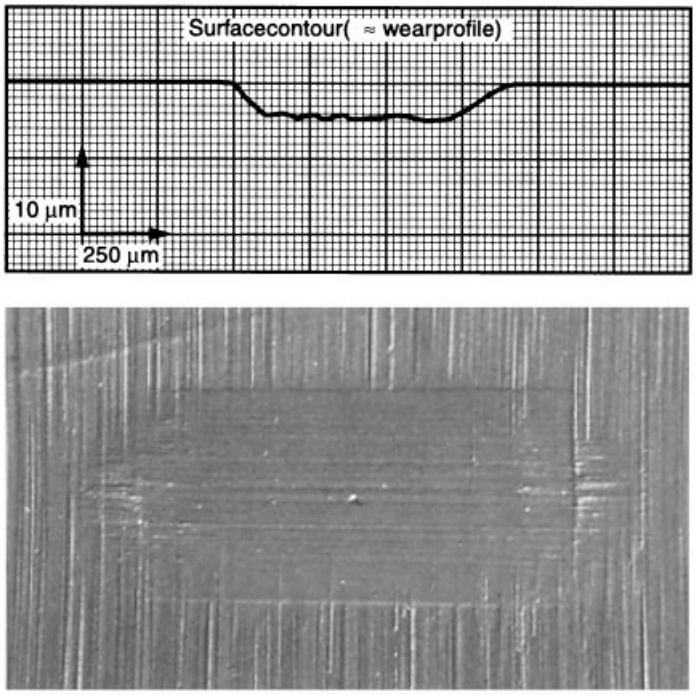

Figure 62. Lithium grease with white solid lubricants showing the smooth pattern of the wear profile. The maximum wear depth is only 5.9 µm, the wear pattern is uniform, and there is no chipping of material. Scanning electron photomicrographs (50 × magnification).

15.3.5. Dry-Film Lubricants

Lubricating varnishes or dry-film lubricants are suspensions of solid lubricants and other additives in a solution of inorganic or organic binding agents. The most commonly used solid lubricants are MoS_2, graphite and PTFE. The most important additives are pigments which protect against corrosion. The main types of binding agent used are organic resins (acrylic, phenol, epoxy, silicone, urethane), cellulose, and inorganic silicates and phosphates. Hydrocarbons or water are used as solvents. Dry-film lubricants can be used in a variety of ways; these depend primarily on the number, shape, and/or particular requirements with regard to partial coating. They are applied by dipping centrifuges and drums and by various types of spraying procedure. The hardening process depends on the type of binding system and happens at ambient temperature or in an enameling stove. After they have hardened lubricating varnishes form a highly adhesive, dry film of lubricant. The proportion of solid lubricants in the layer created in this way can be as high as 70 %.

Solid lubricants with a layer structure (MoS_2) have a floating effect in the wet film, whereby the layers arrange themselves horizontally as the film dries and settle on top of each other in individual layers. This is how a separating layer is created between the base unit and the opposing unit of the tribological system. In an ideal situation this layer is between 10 and 15 µm thick. When placed under pressure, the texture of this layer becomes compressed and creates an extremely smooth, shiny film surface. PTFE lubricating varnishes are usually applied in much thinner layers, e.g., 3 to 5 µm thick.

The performance of dry-film lubricants depends on:

1) Their composition, in particular the type of binding agent, the ratio of solid lubricants to the binding agent, the types of solid lubricants, and the film thickness;
2) The properties of the substrate to which the lubricating varnish is applied, in par-

ticular the cleanliness, roughness (porosity; Fig. 63), and hardness of the surface; and

3) Environmental (local) conditions, in particular the demands, placed on the lubricant (e.g., load), the type and speed of movement, the temperature, atmospheric pollution (humidity, dust, radiation), atmospheric pressure, and the surrounding media (gases, liquids).

The pretreatment of the substrate surface plays a particularly important role in the successful use of lubricating varnishes – as already mentioned, the roughness of the surface and the way in which the crystallites are arranged have a considerable impact on their lifetime. Various pretreatment procedures can be used to achieve the best possible surface. Sandblasting, phosphating, anodizing, atramentizing, etching, or corroding are the most suitable for metallic substrates. Plastic and elastomer surfaces can be pretreated by chemical and mechanical processes and by energetic activation procedures.

15.4. Industrial Uses of Products Containing Solid Lubricants

Because of the continuously increasing demands put on modern lubricating systems, almost all formulations contain a certain proportion of solid lubricants. Significant fields of application are the lubrication of machine elements and in process auxiliaries for various production technologies such as, for example, the production of glass or metalworking.

Solid lubricants play a significant role in metalworking, especially in forming. Their main task is to generate an effective separating layer between the tool and the surface of the workpiece and thus act against the adhesion forces between the materials. Solid lubricants, in combination with their carrier medium and other additives, also reduce friction or generate a defined coefficient of friction, and keep the process temperature constant. Through their wear protection function they increase the service life of tools, protect against corrosion, ensure a stipulated surface quality and the accuracy to gage of the workpieces.

Solid lubricants are used in the area of noncutting forming, especially cold and hot forging, pressure die-casting and drawing, and cold extrusion operations.

15.4.1. Screw Lubrication

The machine element most used is the screw. The selection of the most suitable screw lubricant depends upon the material or the surface of the screw and the thread partner, and on the conditions of use, for example temperature, humidity, and the influence of chemicals, radiation, and the constructive requirement according to a stipulated coefficient of friction or minimum pretensioning force, even after the screws have been tightened or loosened many times.

Special high-temperature pastes are used for the lubrication of screw connections at high working temperatures. These pastes comprise a basic oil which is mixed with different solid lubricants and can contain additional additives. In formulation or selection care must be taken to ensure that the components do not contain sulfur, zinc, lead, and aluminum, which can lead to damaging changes to the screw material, e.g., stress crack corrosion at high temperatures. They should also be free of health-hazardous nickel.

Particularly effective in high-temperature applications are greases and pastes combined with white solid lubricant prepared from metal hydroxides, phosphates, and sulfides.

For many screw lubrication applications, special dry-film lubricants are also available as an alternative to greases and pastes. The advantage of lubricating varnish is its simple application in immersion centrifuges or spray drums. The handling of coated screws is clean and suitable for robot systems.

Special areas of application for lubricating varnish and slide films are thread-cutting screws, thread-shaping screws, thread-grooving screws, all kinds of wood screws, and screws made of plastic.

15.4.2. Roller-Bearing Lubrication

For lubrication of roller bearings greases or pastes containing solid lubricants must always be used when reliable separation of the roller body and bearing ring at low speeds and high loads cannot be ensured by grease alone. Over and above this, greases or pastes containing solid

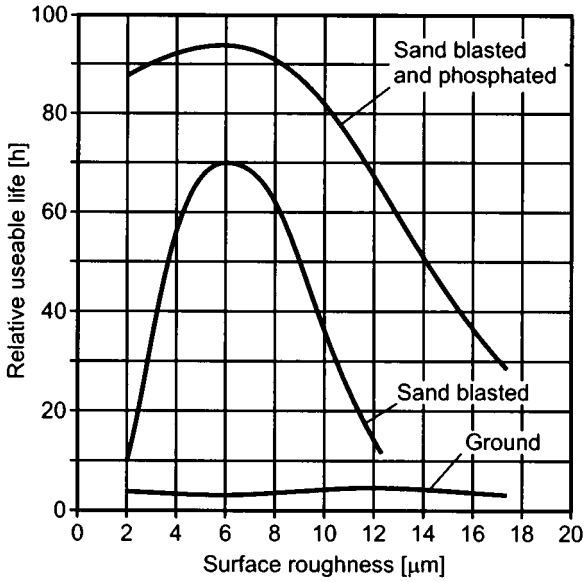

Figure 63. Relative lifetime of MoS_2 solid-film lubricants as a function of surface roughness

lubricants have good emergency running properties and afford high safety under shock and impact stress. Greases or pastes with white solid lubricants are preferable for use in roller bearings employed in oscillating operations. Lubricating varnishes are the best solution for applications in a vacuum, for example, in air and space travel, nuclear engineering, extreme temperatures, or for microsetting movements.

Special constructions also enable solid lubricants to be incorporated in the bearing material; this enables the production of maintenance-free bearings.

15.4.3. Slide Bearing, Slide Guideway, and Slide Surface Lubrication

Here, reliable separation of the material partners must be ensured at all times to comply with the primary demand put on the lubricant for reliable load transfer at lowest wear rates. Whereas, as a general rule, the operating point of a friction bearing is in the region of hydrodynamic lubrication and therefore the viscosity of the lubricant is the most important key value, slide guideways and other slide surfaces (for example, infeed wedges) are mainly in the boundary or mixed friction range. The best solution for these operational requirements is lubrication with PTFE belts and slide liners based on epoxy resin with the incorporation of solid lubricants such as MoS_2, graphite, or polyurethane. Pastes or lubricating varnishes containing solid lubricants are likewise suitable and offer themselves for the running-in phase and the warm-up and cool-down phases in normal operation. In this way the boundary and mixed friction zone can be safely worked through without wear occurring, even with pure oil lubrication, and emergency running properties can be achieved.

15.4.4. Chain Lubrication

When selecting a suitable lubricant special consideration must be given not only to the different chain-drive construction types, but also to the material pairs, the operating parameters, and the ambient conditions. For solid, highly viscous lubrication care must be taken to ensure adequate creep or penetration to get to the narrow gaps, for example, between the bolt and bush. The lubricant must also have an appropriate adhesion capacity so that it is not spun off at high chain speeds or high centrifugal forces at the turning point (pinion and wheel).

15.4.5. Plastic and Elastomer Lubrication

In plastic and elastomer lubrication the material compatibility plays a particularly important role. Swelling, shrinking, or embrittlement are the most serious changes which elastomers can experience when they come into contact with lubricants. Solid lubricants have limited influence on elasticity module, tensile strength, and hardness. On top of reducing friction and protection against wear, the separating effect of the lubricant also plays an important role. The majority of requirement profiles can be covered adequately by lubricating varnishes and films containing solid lubricants.

In plastic lubrication the special features of substances, i.e., interaction between the plastic and the lubricant, must also be taken into consideration. Particular attention must be paid to the tendency to develop stress cracking. Oils, greases and pastes containing solid lubricants, and also lubricating varnishes and slide films, can be used successfully for lubricating plastic and steel, and plastic – plastic pairs. The running-in process must also be given enough attention in the use of plastic materials, and as a general rule, can be positively influenced by solid lubricants.

16. Testing and Analysis

16.1. Base Oil Categories and Evaluation of Various Petroleum Base Oils

The American Petroleum Institute (API) and the Association Technique de l'Industrie Européenne des Lubrifiants (ATIEL) have classified base oils according to their chemical composition. Initially there were four groups; after the introduction of VHVI oils in Europe this was increased to five. The most important reason for these groups was the necessity to regulate base oil interchangeability for engine oils. (Chap. 8). The classification of petroleum base oils (Groups I to III) considers three parameters: saturates content, sulfur content, and viscosity index. Table 4 shows this classification. Accordingly, Group I oils are solvent extracted HVI oils (SN oils), Group II oils are hydrogenated or hydrocracked oils (as the sulfur content of < 300 ppm shows) and Group III products are VHVI oils manufactured by severe hydrocracking and or wax isomerization (VI > 120, sulfur < 300 ppm).

Table 51 shows typical data of various hydrocracked base oils (HC-oils), in comparison with solvent refined oils and polyalphaolefins.

HC-II oils are classified as very high viscosity index (VHVI) oils. HC-III oils are also known as extra or ultra high viscosity index (UHVI) oils. They are generally made by hydrocracking and isomerization slack waxes followed by solvent dewaxing.

Table 52 shows typical hydrocarbon compositions of HC oils in comparison with conventional solvent refined oils and polyalphaolefins [64].

HC-II and HC-III oils are being increasingly compared with synthetic polyalphaolefins (PAO); the close similarity of the performances of some hydrocracked oils and PAO, and their significantly lower cost, make them increasingly attractive for lubricant formulations.

Figure 64 illustrates the development of VHVI and UHVI oils (HC-II and HC-III oils) in Europe, Asia, and North America. Whereas Shell and BP in France and Union Fuchs in Germany were the only manufacturers of VHVI oil in the 1980s, by the mid-1990s similar activities had spread from Western Europe around the world [66].

16.2. Laboratory Methods for Testing Lubricants

The properties of lubricants are mostly determined by standardized test methods.

The most common laboratory test methods are listed in the following sections and an overview of the corresponding international standards is given e.g., AFNOR [529], ASTM [530], DIN [531], IP [532], and ISO [533].

16.2.1. Density

For the determination of density of lubricants different test equipment is in use such as hydrometer, pycnometer, hydrostatic balance or digital densimeter.

Table 51. Typical data for 4 mm^2/s base oils (HC oils) in comparison with conventional solvent refined oils and polyalphaolefins

	Solvent refined	HC-1	HC-2	HC-3	Polyalphaolefins
Viscosity, mm^2/s at 100 °C	4	4	4	4	4
Viscosity index	100	105	125	130	125
Volatility, Noack evaporation loss, wt %	23	18	14	13	12
Pour point, °C	−15	−15	−18	−20	−65
API group	I	II	III		IV

Table 52. Typical hydrocarbon composition of HC oils in comparison with conventional (4 mm^2/s) solvent refined oils and polyalphaolefins

	Solvent refined SN 100	HC-1	HC-2	HC-3	Polyalphaolefins
n- and Isoparaffins, %	25	30	55	75	96
Monocycloparaffins, %	20	35	24	15	4
Polycycloparaffins, %	30	34	20	10	–
Aromatics, %	24	0.5	0.3	0.1	–
Thiophenes, %	0.5	–	–	–	–

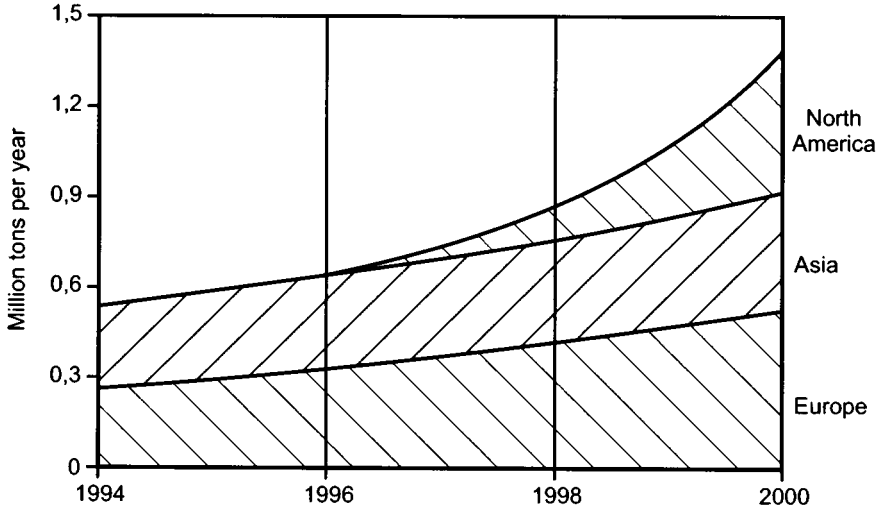

Figure 64. Development of VHVI and UHVI oils (HC-II and HC-III) in Europe, Asia and Northern America

16.2.2. Viscosity

For the determination of viscosity capillary viscometers, falling ball and rotary viscometers are used.

Capillary viscometers (DIN 51562/ASTM D 445) are generally used for most of the automotive and industrial lubricants for determination of the kinematic viscosity. The most frequently used viscometers are Ubbelohde for transparent liquids, and Cannon – Fenske viscometers (DIN 51366).

Rotary viscometers use the torque on a rotating shaft to measure a fluid's resistance to flow. The shear rates can be changed by modifying the speed of rotation, rotor dimensions and the gaps between rotor and stator.

The CCS apparent viscosity is measured in a range between 2.000 and 20 000 mPa · s at temperatures between −5 and −30 °C (ASTM D 2602/DIN 51377). The CCS apparent viscosity correlates excellently with low temperature engine cranking.

The MRV (ASTM D 3829) is applied to measure the borderline pumping temperature and is

related to the mechanism of pumpability of engine oils.

The Brookfield viscometer (ASTM D 2983/DIN 51398) is used for the determination of the low shear rate viscosity of automotive fluid lubricants in a temperature range between −5 °C to +40 °C.

The Ravenfield viscometer (ASTM D 4741) allows the determination of dynamic viscosity of engine oils at 150 °C using a high shear rate tapered/plaque viscometer.

16.2.3. Refractive Index

The refractive index (DIN 51423/ASTM D 1218) is used for the characterization of the structure of petroleum hydrocarbons like the density. Furthermore, it is mostly used as a property for fast identification of a product.

16.2.4. Structural Analyses

The amount (%) of aromatic and paraffinic hydrocarbons in a mineral oil can be determined by the carbon distribution method developed by Brandes [534]. The method determines the infrared absorbency at wavelengths of 1620 and 724 cm^{-1}, which correspond to the stretching frequencies of the aromatic and paraffinic hydrocarbons respectively. However, this method gives no information on isoparaffins and does not take into consideration sulfur content.

According to DIN 51378/ASTM D 2140 the kinematic viscosity at 40 °C, the density at 20 °C and the refractive index as well as the sulfur content are determined. From these characteristics the percentage of aromatic, naphthenic and paraffinic hydrocarbons is determined with or without sulfur correction.

For improved characterization chromatographic, gas chromatographic and NMR spectroscopy methods as well as element analyses of the narrow distillate fractions have to be carried out.

16.2.5. Flash Point

For the determination of the flash point closed and open cups are used. The open cup flash point is described according to DIN ISO 2592/ASTM D 92. The closed cup method is mainly used for lubricants with a flash point > 79 °C.

For lubricants and mixtures with solvents the flash point is determined by closed cup methods according to Abel – Pensky (DIN 51755) or Pensky – Martens (DIN 51758/ASTM D 93).

16.2.6. Surface Phenomena

During storage or application lubricants are in a continuous contact with air or even with water, which might lead to formation of foam or emulsion. For evaluation of the behavior of base oils and lubricants the following methods are applied.

Air Release. For the determination of air release behavior (DIN 51381/ASTM D 3427) compressed air is blown through the test oil which is heated to a defined temperature. After the airflow is stopped the time required for the air entrained in the oil to be released is determined. This is the time at which the density is narrowed to 0.002 g/mL of the density of the pure oil.

Demulsibility. For the determination of demulsibility (DIN ISO 6614, ASTM D 1401) 40 mL lubricant and 40 mL of distilled water are stirred at a defined temperature in a graduated cylinder. After stirring the time required for the separation of oil from the water is determined. If after 30 min standing the separation is not totally completed the volume of oil, water and emulsion is reported.

Foaming Characteristics. The foaming characteristics (ASTM D 892) are determined by blowing for 5 min a constant rate of air through the oil. The volume of the foam is measured immediately and after settling for 10 min.

16.2.7. Cloud Point, Pour Point

The pour point (ASTM D 97) is the lowest temperature expressed as a multiple of 3 °C at which the test fluid is observed to flow when it is cooled and examined under defined conditions.

The cloud point (ASTM D 2500) is the low temperature behavior where an oil starts to become cloudy when it is cooled under defined conditions.

16.2.8. Aniline Point

The aniline point (DIN 51775/ASTM D 611) is the temperature at which equal volumes of hydrocarbon oils and aniline separate in two phases. This value gives an indication on the structure of mineral oils. The higher the temperature of separation the more paraffinic is the oil.

16.2.9. Water Content

For water contents $\geq 0.05\,\%$ ASTM D 95 is used. This test procedure involves azeotropic reflux distillation of the oil with xylene with water separating from the reflux distillate.

For lower water contents between 50 and 1000 ppm the Karl Fischer method (DIN 51777/1/ASTM D 1744) is more accurate.

16.2.10. Ash Content

For unused oils the determination of the ash content (DIN EN 7/ASTM D 482) gives an indication of purity.

The content of acid constituents is characterized by the determination of sulfated ash from unused lubricating oils containing additives. Those additive concentrates contain a variety of metals. As this is a time consuming method it has more or less been replaced by determination of elements by atomic absorption, X-ray fluorescence or emission spectroscopy.

16.2.11. Acidity, Alkalinity

The neutralization number of base oils is determined by titration against a color indicator (ASTM D 974). When the oil contains additives such as dispersants or detergents, the end point of titration is difficult to observe, potentiometric titration is used.

Another method is the total acid number (TAN/ASTM D 664). The acidity of unused oils and fluids is normally derived from the type and concentration of specific additive material whereas the acidity of used oil is of interest to measure the degree of oxidation of the fluid.

The TBN characterizes the alkaline reserve. It is particularly used for engine oils where by acidic combustion products use up the alkaline reserve.

16.2.12. Aging Tests

Aging tests are carried out on base oils as well as fully formulated products to test the efficiency of additives. The different oxidation stability tests are the most common ways for testing the aging properties of lubricants. There is a big variety of standardized test methods. The majority of these tests is based on exposing the test fluids to oxygen or air at relatively high temperatures in presence of catalytic metals to increase oxidation rates and to reduce the testing period.

The turbine oxidation stability test (TOST) is used to evaluate the oxidation stability of inhibited steam turbine oils in presence of oxygen, water, copper and iron catalyst at 95 °C. The test is continued until the TAN measured reaches at least 2.0 mg KOH/g. The number of test hours required for the oil to reach 2.0 mg KOH/g is the 'oxidation lifetime' of an oil.

The oxidation stability by rotary bomb test (ASTM D 2272) oxidizes the oil at 150 °C in presence of water, metallic copper catalyst and oxygen at 620 kPa pressure. The time is registered to reach a specific pressure drop and this is an indication of the oxidation stability.

The aging test according to Baader (DIN 51554) is an oxidation test at atmospheric air and intermittent immersion of a copper spiral at a test temperature of 95 °C. After a given time the saponification number in mg KOH/g is measured.

In the oxygen bomb method a lubricating grease is oxidized in a bomb heated to 99 °C and filled with oxygen at 758 kPa overpressure. After a defined time the pressure drop is recorded. The degree of oxidation after a given period of time is determined by the corresponding decrease of oxygen pressure.

16.2.13. Hydrolytic Stability

The hydrolytic stability of additives is commonly tested in the beverage bottle method (ASTM D 2619). The method differentiates the relative stability of fluids in the presence of water under the conditions of this test.

16.2.14. Corrosion Tests

The detection of copper corrosion (ASTM D 130/DIN 51759, Part 1) describes a simple procedure for investigating the aggressiveness of the fluid towards a polished copper strip at a temperature for a time selected for the fluid under test.

The rust preventing characteristics of inhibited mineral oils (ASTM D 665/DIN 51585) in the presence of water is used for evaluating corrosion preventive properties of steam turbine oils and hydraulic oils.

For the evaluation of rust preventive properties of metal corrosion inhibitors under high humidity (ASTM D 1748/DIN 51359), steel panels are prepared to a prescribed service finish and dipped in a test oil.

16.2.15. Oil Compatibility of Seals and Insulating Materials

An essential requirement for lubricants is that they should be compatible with seal materials as they are frequently in prolonged contact at elevated temperatures. In all the test methods applied, well defined test specimens are immersed in the lubricant for a defined time at defined temperatures. After the test period the volume change, weight loss, and change of mechanical properties are evaluated.

16.2.16. Evaporation Loss

The volatility of oils is determined by evaporation loss expressed as a percentage by mass of nonadditive and additive-type lubricating oils (ASTM D 5800/DIN 51581, Part 1). The procedure according to Noack has been developed for motor oils but it can also be applied to other lubricants. A well defined air stream is directed over the oil which is heated for 1 h to 250 °C. The evaporation loss is established by weighting.

16.2.17. Analysis of Grease

The consistency of lubricating greases is measured by penetration of a standard cone (ASTM D 217/ISO 2137). Penetration is the depth in tenths of millimeters where a standard cone assembly sinks into the grease.

The dropping point of lubricating greases (ASTM D 566/ISO 2176) is the point at which a conventional soap-thickened grease passes from semisolid to a liquid state under the conditions of the test or the temperature at which a non soap-thickened grease rapidly separates oil.

16.3. Mechanical – Dynamic Testing Methods for Lubricants

Mechanical – dynamic lubricant testing has become an essential element in the development of modern lubricants. In this respect, standardized tribological, mechanical – dynamic testers and test methods play a decisive role in the development of lubricants. A variety of so-called house-internal methods complete the range of today's tribological, mechanical – dynamic tests.

16.3.1. Tribological System Categories within Lubricant Tests

Today's wide range of tribological, mechanical – dynamic testers and testing methods includes small laboratory instruments and field tests under real conditions. The various types of lubricant test are classified, for example, by the German DIN 50322 [535].

16.3.2. Standardized and Nonstandardized Test Methods for Lubricants

The most important test methods, used worldwide by many laboratories, are, above all, the ISO, CEC, ASTM, IP and DIN standards. Often, these standardized test methods will include overlapping among the procedures or even technically equivalent methods.

The common laboratory testers used in today's lubricant development and the tests performed with these testers [536, 536, 537], are listed in Table 53.

Table 53. Common laboratory testers and test standards.

Tester	Application	Test method
Roll stability test apparatus	standard test method for roll stability of lubricating grease	ASTM D 1831-94
SKF (EMCOR) Anti-rust rig	determination of rust prevention characteristics of lubricating greases	IP 220/93 DIN 51802
Water-washout apparatus	standard test method for determining the water washout characteristics of lubricating grease	ASTM D 1264-93 IP 215/93 DIN 51807 Part 2
ASTM Torque test apparatus	standard test method for low-temperature torque of ball bearing grease	ASTM D 1478-91
IP Low temperature torque apparatus	determination of low-temperature torque of lubricating grease	IP 186/93
ASTM Leakage testing apparatus	standard test method for leakage tendencies of automotive wheel bearing greases	ASTM D 1263-94
FAG Noise testing apparatus	noise inspection of lubricating greases (computer-aided)	FAG-QV3.102FB MGG 11
SKF Noise testing apparatus	noise inspection of lubricating greases (computer-aided)	SKF specification MVH 90 B
Sonic probe type apparatus	standard test method for sonic shear stability of polymer-containing oils	ASTM D 2603-91
Shell delimon rheometer	standard test method for determining the delivery resistance using the shell delimon rheometer	E-DIN 51816, part 1
Evaporation loss tester	standard test method for engine and hydraulic oils for the determination of the evaporation loss and sediments	WOLF specification
Incline tribometer	standard test method for determining friction coefficients	H. Schmidt Specification

16.3.3. Common Mechanical – Dynamic Testers

Four-Ball Apparatus. The four-ball apparatus is one of the oldest and best-known model test benches for liquid and solid lubricants. Because it furnishes very precise lubricant key values, it is used all over the world. In this apparatus, a roller-bearing ball rotates under pressure and at a constant speed on three fixed steel balls of the same type in an oil bath or lubricated with a solid lubricant (Fig. 65). The gradual increase in pressure enables the determination of weld loads, supplies key values concerning loadability, or enables allows the determination of the friction or start-up behavior in relation to the lubricant. In addition, the application of relatively small loads over a longer period of time enables the determination of key wear protection values and the friction values of lubricants. During these tests, the surface of the ball surface will produce wear impressions which when measured will lead to information on the effects of additives.

In the late 1990s several adapters have been developed for the four-ball apparatus to furnish information on the pitting load capacity and the shear stability of polymer-containing lubricants. Further modifications of the adapter testers enable the determination of friction torque and temperature in the immediate proximity of the outer ring of the roller bearings.

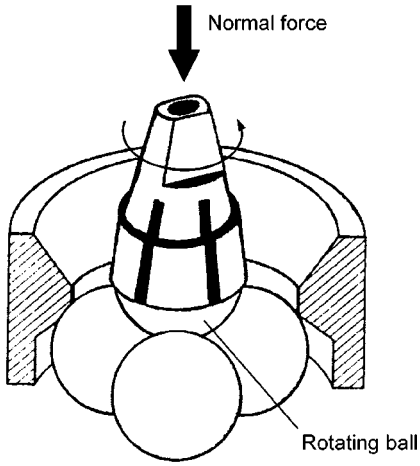

Figure 65. Diagram of the test using the four-ball, apparatus

As ever, today's numerous specifications for gear and hydraulic lubrication and for all types

of greases and pastes contain key values for the minimum requirements for lubricants.

Reichert Friction Wear Tester. The Reichert friction wear tester (Fig. 67) plays an important role in the determination of the wear rates of water-containing and non water-containing coolant lubricants.

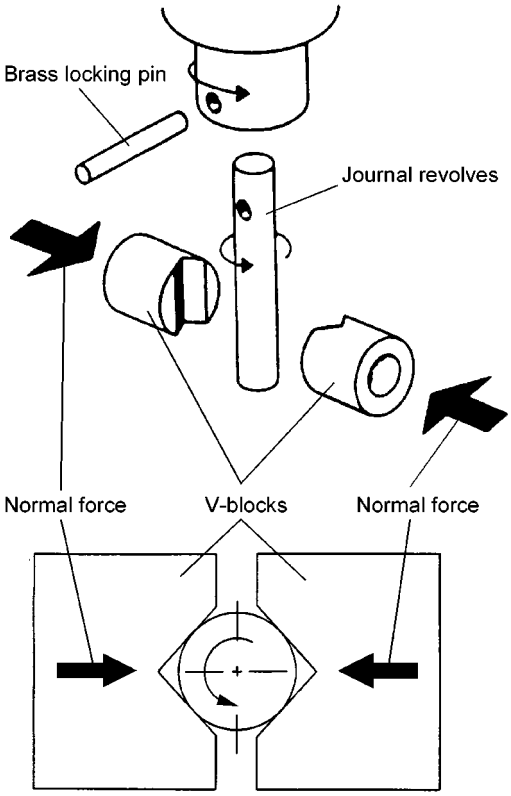

Figure 66. Reichert friction wear tester

By means of a double lever handle system, a firmly clamped cylindrical roll is pressed, under load, against a slip ring rotating cross-directionally to the test roll. Approximately the lower third of the test ring is dipped into the test fluid. After a walkway of 100 m at a constant rotational speed of the rotating slip ring, the elliptical wear mark produced on the test roll is measured.

Falex Tester. The Falex tester (Fig. 66) is used for testing lubrication oils and greases. It is used worldwide to determine the friction, load stage, and wear key values of a lubricant.

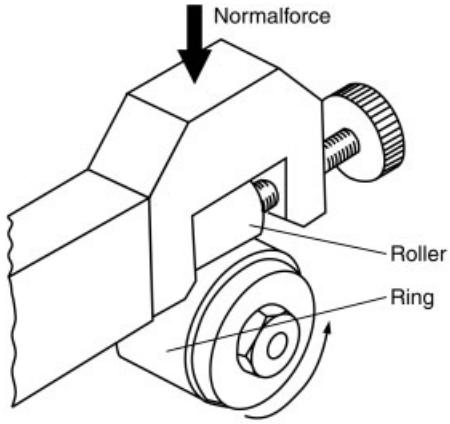

Figure 67. Falex teste

A soft steel shaft, either greased or in an oil bath, rotates under pressure between two hardened steel prisms. At constant rotational speed the test load is increased in three steps. If a lubricant withstands the applied load, the wear-loss of the softer shaft is measured.

Timken Tester. The Timken tester (Fig. 68) is particularly suitable for determination of the key values of greases, lubrication oils, and solid lubricants such as pastes. The apparatus tests the load stage and the wear behavior of the lubricants under mixed-lubrication conditions.

In the metallurgical and steel industry the Timken tester is still particularly important all over the world.

The lubricant test is conducted in a friction system comprising a cuboid block and a test ring. The tester is run with a lubrication oil circulation system, which can be adjusted variably, and a grease feeder with a feed volume of 45 g/min. The test force, which is gradually increased, is brought into the friction system via the test block. Individual weighing of the test block and the test enables determination of the wear loss in a load stage. The test period for each load stage is 10 min at a rotational speed of 800 min^{-1}.

Translatory Oscillation Apparatus (SRV). The translatory oscillation apparatus (Fig. 69) is used to test lubrication oils, grease, and solid lubricants. It is used worldwide to record the friction value, wear behavior, and load stage of a lubricant in key figures. The lubricant test is conducted under mixed-lubrication conditions.

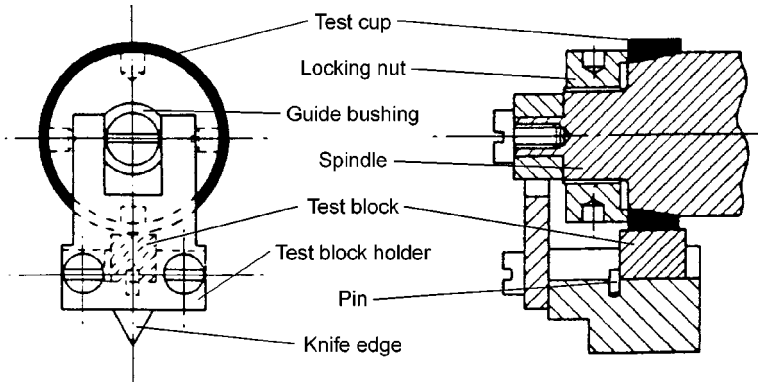

Figure 68. Timken tester

This, in particular, enables the testing of wear-protective and scuffing-preventing additive systems which are part of EP and AW active substances.

Test pieces made of roller bearing steel are inserted in the test chamber of the oscillation apparatus and wetted with the lubricant to be tested.

A great advantage of the tester is the possibility of performing systematic studies simply and quickly, with only small quantities of lubricant. These variations of testing conditions, which vary from the test standards, do, however, require sufficient experience with this tester and the interpretation of its results.

FZG Gear Test Rig. The FZG gear test rig (Fig. 70) is a so-called component test rig with a direct relationship to the machine or transmission system. The FZG gear test rig is one of the most important tribological test benches for gear lubrication and hydraulic oils worldwide. It is predominantly used to determine the maximum load of lubricants for all types of gear drive.

The test bench consists of a bed plate (a) with two firmly mounted gears which are coupled to each other through shafts. During operation the flanges of the load clutch (b) are connected with each other by means of screws. If these screws are loosened and one load clutch flange is locked using the lock pin (c), the second load clutch flange can be twisted by a lever arm with weight pieces, thus initializing a torque in the test bench. If the two flanges of the load clutch are rescrewed to each other and the load lever is removed, the initialized torque in the shafts is sustained. Accordingly, the forces also strain the toothing in the test and drive gears.

The FZG gear test is used for the load stage test (FZG test method A/8.3/90), micropitting test (GF-C/9(10)/8.3/90), pitting load capacity test (P-C/9(10)/8.3/90), wear load capacity test (ASTM D 4998 – 95), and gear efficiency test (VW-PV 1454).

FAG FE9 Roller Bearing Grease Tester. The FAG FE9 roller bearing grease tester (Fig. 71) has been developed for component tests for greases under realistic conditions. The service life of the grease is determined from the operating temperature, the load, and the rotational speed of the test bearing.

Test method A/1500/6000 has been standardized according to DIN 51821, part 2. The angular contact ball-bearing in the tester is filled with approximately 2 mL grease. The test-bearing is loaded with an axially aligned test force of 1500 N at a rotational speed of 6000 min^{-1}. The test temperature can be selected from within the range 120 – 200 °C (maximum) and is applied to the test bearings by means of heating elements. The high test temperatures in the greased test bearing facilitate oxidation of the lubricants. Failure of a lubricant in the test bearing results in a clear increase of the bearing friction.

FAG FE8 Roller Bearing Grease Tester. The FAG FE8 roller bearing grease tester (Fig. 72) is one of the most versatile component testers for lubricants. It tests the suitability of lubricants and lubrication oils with regard to the friction and long-term wear protection behavior

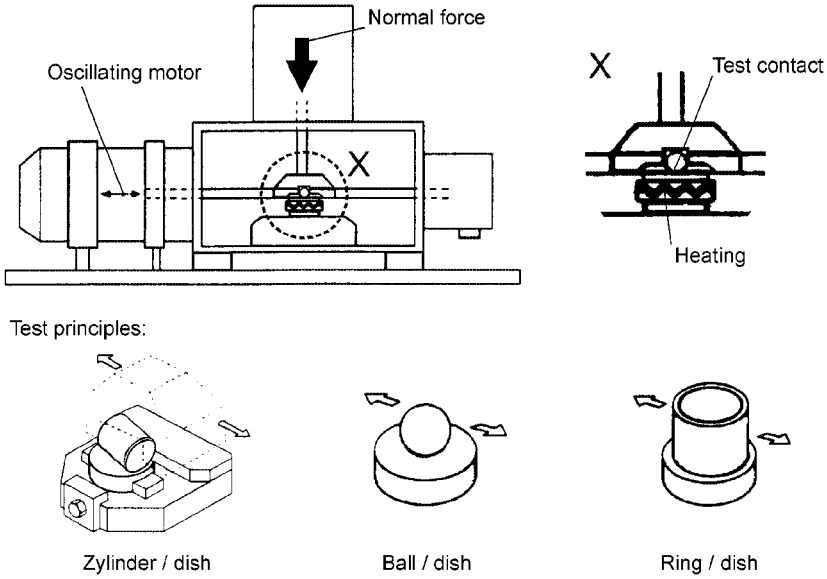

Figure 69. Diagram of the linear oscillation tester (SRV) and possible test sample geometries

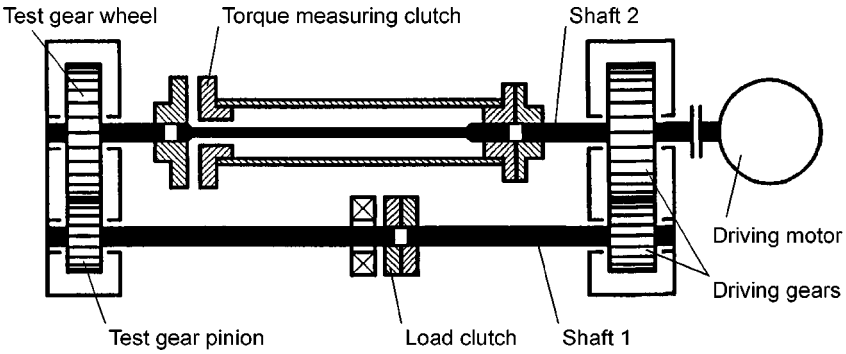

Figure 70. Diagram of the FZG gear test rig

of the lubricants in roller bearings for almost all applications.

Two test bearings of the test head of the FE8 are rigged, axially using disk springs, back to back. The test bearings are lubricated with grease or, by means of a circulation lubrication system, oil. The drive of the test bearings is effected through the inner races of the test bearing on the drive shaft. The test bearing's outer races are fixed in the test head box. By means of a wire rope the test head is fixed to a load cell which records the dynamic friction torque during the test run. An outer radiant heater, placed on the test head as a heating cap, enables tests at bearing temperatures up to 200 °C.

The bearing load is equivalent to 10, 20, 50, or 80 kN. The engine speed can be varied from 7.5 to 3000 min^{-1}.

Synchronizer Rig SSP 180. The synchronizer rig SSP 180 is a component tester developed to test the different types of synchronizer system normally used in the automotive industry and found in all kinds of manual transmission.

The durability to be determined depends not only on the synchronizer component system used but above all on the lubricant used. Nowadays numerous lubricants are perfectly suitable for use in one tribological system but cannot be used in another. Thus, this component test and this relatively new tester play an important role

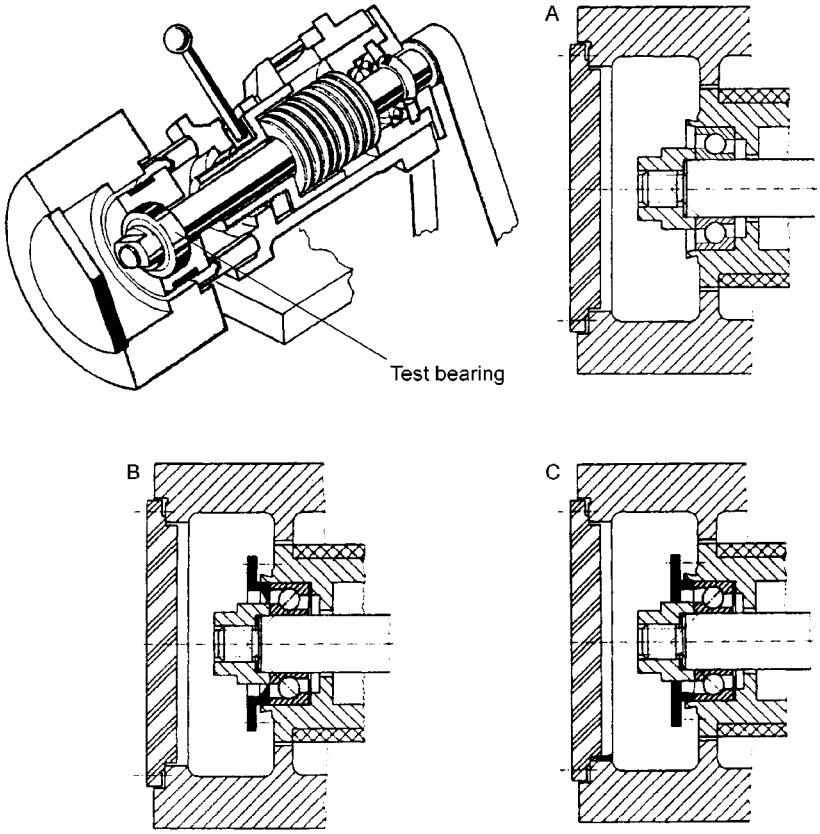

Figure 71. Test bench of the FAG-FE9 roller bearing grease tester. A) Open tester bearers; B) Covered tester bearings without grease reservoir; C) Covered tester bearings with grease reservoir

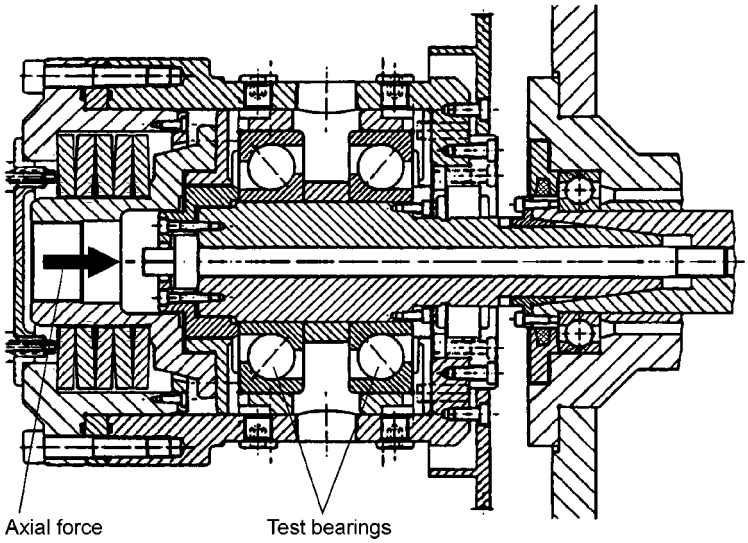

Figure 72. Test bench diagram of the FAG FE8 roller bearing grease tester

in the development and pre-selection or screening of new lubricants and additives for manual transmissions.

Diesel Injection Nozzle. This tester is, on the one hand, used in laboratories to test liquid lubricants and, on the other, enables a lubricant test with an aggregate. This test method has been developed to determine the change in viscosity of polymer-containing hydraulic and engine oils for in relation to the temperature.

For a specified number of run-through cycles, a sample (170 mL) of lubricant is subject to a shear stability test in the tester, which consists mainly of a two-cylinder pump and an injection nozzle set for calibration oil shear stability. Before and after the shear stress the kinematic viscosity is determined at 40 °C with hydraulic oils or 100 °C with engine oils.

17. Economic Aspects

In 1999, 37.3×10^6 t of lubricants were consumed worldwide (56 % automotive lubricants, 29 % industrial lubricants, 5 % marine oils, 10 % process oils). Of total industrial oils, 38 % were hydraulic oils, 7 % industrial gear oils, 33 % other industrial oils, 17 % metalworking fluids (including temporary corrosion preventives whose multipurpose function often includes lubrication) and 5 % greases.

Table 54 shows the 1999 per capita consumption for various regions. North America and Western Europe make up 40.6 % of world lubricant consumption even though these regions only account for 10.7 % of the world's population [538].

Table 54. 1999 Per capita consumption of lubricants (kg/a)

North America	31.5
Australia/Oceania	17.4
Western Europe	15.3
Central/Eastern Europe	11.9
Latin America	5.8
Asia	3.2
Africa	2.2
World	6.1

Since 1975, quantitative lubricant demand has significantly detached itself from gross national product and also from the number of registered vehicles. This quantitative view gives an inadequate impression of the significance of the lubricants business today. In almost all areas, products now have a longer life and offer greater performance, i.e., specific lubricant consumption has declined but specific revenues have increased noticeably. This is also confirmed by the volumetrically very important group of engine oils: the doubling of requirements with extended oil change intervals in recent years have quadrupled the cost of such oils.

As about 50 % of the lubricants sold worldwide end in and thus pollute the environment, every effort is made to minimize spillages and evaporation. An example is diesel engine particulate emissions, about a third of which are caused by engine oil evaporation.

A further incentive to reduce specific consumption is the ever-increasing cost of disposal or recycling of used lubricants. But this again creates new demands on lubricants because reduced leakage losses means less topping-up and less refreshing of the used oil. The new oils must therefore display good aging stability.

Bearing in mind the growth potential in Asia where per capita consumption of lubricants in some areas is extremely low (1999: India 1.1 kg, China 2.5 kg) and a continuing reduction in volumes or stagnation in Western industrialized countries, overall global growth is forecast. This has been estimated to be 6.5 % between 1997 and 2002 [539].

Worldwide, there are 1700 lubricant manufacturers ranging from large to small. On one hand there are vertically-integrated petroleum companies whose main business objective is the discovery, extraction and refining of crude oil. Lubricants account for only a very small part of their oil business. At present, there are about 200 such national and multinational oil companies engaged in manufacturing lubricants. In contrast, the 1500 independent lubricant companies view lubricants as their core business. While the large, integrated companies focus on high-volume lubricants such as engine, gear and hydraulic oils, many independent lube companies concentrate on specialties and niche business.

Less than 2 % of lubricant manufacturers produce more than 60 % of global lubricant volumes. Tables 55 and 56 show the world's largest manufacturers of lubricants and industrial lubricants.

Table 55. World's largest lubricant manufacturers (2000)

	Company	Country
1	Exxon Mobil	USA
2	Shell	UK / Netherlands
3	Chevron Texaco Caltex	USA
4	BP CASTROL	UK
5	SINOPEC / CNPC	China
6	Pennzoil Quaker State	USA
7	Total Fina Elf	France
8	Nippon Mitsubishi Oil / Fuji Kosan / Koa Oil	Japan
9	Lukoil	Russia
10	Valvoline	USA
11	SUN	USA
12	IDEMITSU	Japan
13	Fuchs	Germany
14	Indian Oil	India
15	AGIP / Petrogal	Italy
16	Repsol – YPF	Spain

Table 56. World's largest manufacturers of industrial lubricants (2000)

	Company	Country
1	Exxon Mobil	USA
2	Shell	UK / Netherlands
3	Sinopec / CNPC	China
4	Chevron Texaco Caltex	USA
5	BP Castrol	Germany

The independent lubricant manufacturers rarely operate base oil refineries. They buy their raw materials from the chemical and oleochemical industry and their mineral base oils from the large petroleum companies.

The production of simple lubricants normally involves blending processes but specialties often require the use of chemical processes such as saponification (in the case of greases), esterification (when manufacturing ester base oils or additives) or amidation (when manufacturing components for metalworking lubricants). Further manufacturing processes include drying, filtration, homogenizing, dispersion or distillation.

Towards the end of the 1990s, the petroleum industry was affected by a wave of mergers which will continue. These created new and larger lubricant structures at the merged companies. The principal reasons for these mergers were economic factors in crude oil extraction and refining which resulted in lower refining margins.

Grease Market. In 1995 37.1×10^6 t of lubricants were used worldwide. Of these 1.2×10^6 t were greases [541]. The NLGI has reported that 710 830 t of grease were produced in 1999 [542] and it must be considered that only ca 60 % of the worlds grease manufacturers contributed to this figure.

Conventional lithium greases (52.8 %) are the most used. Lithium complex greases account for 14.1 % and are followed by conventional calcium greases with 9.7 % (Table 57).

Table 57. Global share of thickeners (1999)

Thickener system	%
Conventional lithium soap	52.8
Lithium complex soap	14.1
Conventional calcium soap	9.7
Aluminum complex soap	4.7
Polyurea	4.3
Organophilic clay thickeners	3.5
Calcium complex soap	3.2
Sodium soap	2.5
Anhydrous calcium soap	2.0
Other metallic soap	1.2
Other thickeners	2.0

Although lithium soap based greases are the most used products worldwide, there are significant differences among local markets with regard to other grades of grease. In the US and Canadian market lithium complex greases (29.6 %) are the most used. Aluminum complex greases (6.8 %) are also above the global average. In Europe the calcium complex greases (9.4 %) are ahead of lithium complex grades (8 %). In Japan the polyurea greases (14.6 %) are behind conventional lithium greases (58.2 %); both figures are above the global figures (Table 58).

Table 58. Local share (%) of high-temperature thickeners (1999)

Thickeners	Global	USA	Europe	Japan
Lithium complex	14.1	29.6	8.6	0.2
Aluminum complex	4.7	8.0	4.7	2.7
Polyurea	4.3	7.6	1.7	14.5
Organophilic clay	3.5	6.4	2.2	0.5
Calcium complex	3.2	2.2	6.6	0.3

The main consumer market is the automotive market, with ca 50 %. The grease used are conventional lithium greases, conventional calcium greases, and often sodium greases for relubrication of trucks, construction, farming, and forestry equipment, and old-fashioned cars.

The industrial grease market is divided into a wide range of different and tailor-made greases. Big consumers are steel mills, mines, especially open-pit mines, railroad companies, motor companies, and manufacturers of all kind of machinery; they rely on many suppliers. Often these suppliers emphasize the performance of a wide variety of greases. Among industrial grease customers the roller bearing industry is very important; in addition to high-quality conventional lithium greases they increasingly need specific greases to fulfil their customers' requirements.

In Germany products used in switches have been completely replaced by ester based biodegradable greases. Trials with biodegradable wheel-flange greases in Austria, The Netherlands, and Germany [543] have been successfully completed. Testing in Canada of a biodegradable grease for rail track-side lubrication proved its superior load-carrying and wear-protection performance [544].

18. Disposal of Used Lubricating Oils

Figure 73 shows the fate of lubricants sold every year in Western Europe. Only 49 % are collectable and only 28 % are actually collected [545]. (The chart also includes process oils which are not lubricants.) To ameliorate this situation, the following objectives should be achieved: the intensive gathering of collectable oils and an improvement in the environmental compatibility of the lubricants.

18.1. Possible Uses of Waste Oil

Used lubricants represent a problem for the environment. Their ecologically-compatible use is therefore an important environmental protection measure.

The demand that new lubricants should be made of used products is based on the erroneous notion that re-refining can restore the original condition of a lubricant. In fact, lubricants lose value during use and re-refining, at best, can only restore the value of base oils. In the case of conventional mineral oils, this value is only slightly higher than that of fuels or heating oils. This is also the reason why re-refining is hardly economical without legislative provisions or subsidies. From a global competition point of view, other disposal options include the direct incineration of untreated waste oils, the simple pretreatment (cleaning) and alternative uses such as flux oils for bitumen or for the manufacture of secondary feeds for catalytic crackers and as blending stock for high-sulfur fuels.

18.2. Legislative Influences on Waste Oil Collection and Reconditioning

The EC Directive 87/101 contains a recommendation to all member countries concerning the regeneration of used oils insofar as economic, technical and organizational conditions allow. Emission thresholds for incineration plants (< 3 MW) make the burning of untreated used oils difficult. However, in some European countries, incineration in smaller incineration plants is still possible. Legislation permits incineration in high-temperature furnaces and by the cement manufacturing industry. In some countries, fuels and heating oils reclaimed from used lubricants are not taxed and are thus subsidized.

The polychlorinated biphenyl (PCB) problem which surfaced in Europe in 1983 significantly influenced European legislation on wastes, i.e., two categories of waste oil were created: Used oils which contain more than 0.2 % chlorine cannot be re-refined and are subject to expensive disposal procedures. This in turn has promoted the development of chlorine-free lubricants.

Re-refined used oils are subsidized in Italy. In Germany, the manufacturers of lubricants (including distributors) have transferred their legal requirement to properly dispose of waste oils to collection organizations. In 1999, these received about 90 per tonne from the lubricant consumers.

In the USA, state law on this subject differs. Since 1986, used oils have been classified as hazardous wastes in California and other states have since followed. In some states, the collector is

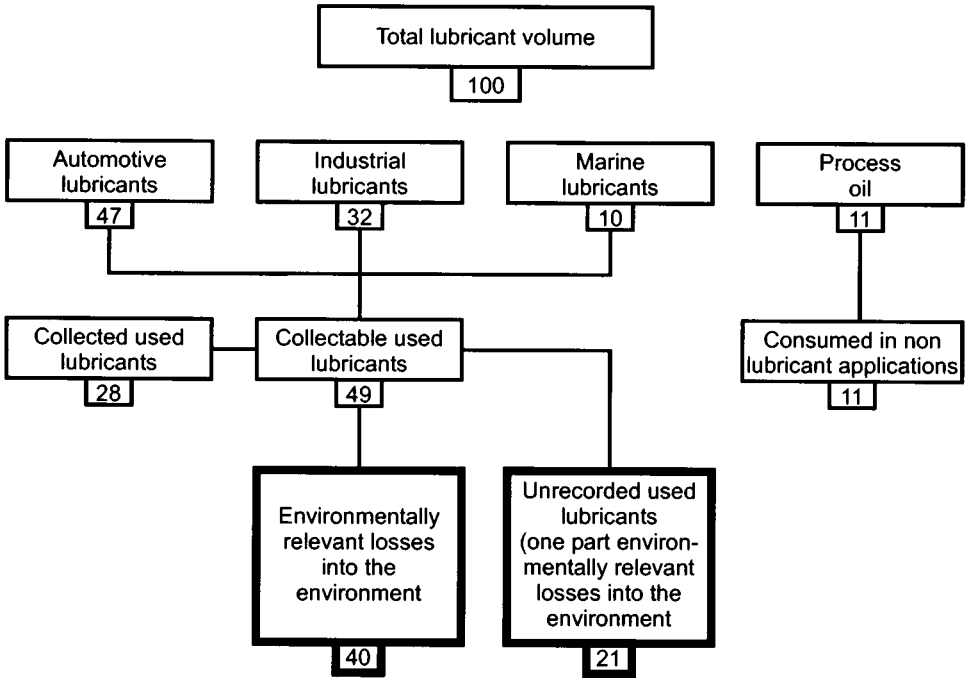

Figure 73. Lubricating oil supply, use and disposal in Western Europe (CONCAWE 1996) [545]

paid up to 20 cents per gallon by the oil user, in other states the collector has to pay.

Viewed globally, some extremely differing situations exist. While some countries do not regulate the collection and disposal of used oil and used oil is generally not collected, other countries can point to high collection and disposal rates (in 1996, 99 % of used oils were collected or properly incinerated in Germany; but only 60 % in the USA).

18.3. Re-refining

The re-refining of used oils to lube base oils started in 1935 [546]. The principal reasons why re-refining was unable to find acceptance were: high process costs and therefore high selling prices compared to relatively low virgin oil prices, inadequate removal of carcinogenic polycyclic aromatics, the negative image of such oils in most markets and the increasing complexity of base oil blends in engine and other lubricants. In Western Europe, only 7 % of base oil demand was satisfied by re-refined products in 1998.

Numerous re-refining technologies have been developed over the last 20 years. Many were patented but only few were suitable for large-scale application [547–552].

In general, the process stages, given below, are common to all the different methods.

1) Separation of larger solid impurities along with most of the water. This is normally achieved by sedimentation.
2) Separation of the volatile parts (fuel residues in engine oils, solvents and low boiling-point lubricant components), normally atmospheric distillation. The separated light hydrocarbons can usually be used in-house for energy generation.
3) Separation of the additives and aging by-products. This can occur by acid refining, solvent (propane) extraction, vacuum distillation or partly also by hydrogenation.

4) Finishing process to separate any remaining additives, aging byproducts and refining reaction products. This is normally done by hydrofinishing, with absorbents such as bleaching clay or mild, selective solvent extraction (e.g., furfural).

Sulfuric Acid Refining (Meinken). The sulfuric acid refining process was mostly developed by Meinken. Compared to older acid-based methods, various process stages reduce the amount of acidic sludge and used bleaching clay generated as well as increasing the lube oil yield.

Due to the acidic sludge problem, acid refining has largely been replaced by other methods. However, numerous such plants were still in operation in 1999. Figure 74 shows the flow sheet of a Meinken plant.

Propane Extraction Process (IFP, Snamprogetti). Of the principal extractive refining processes, the IFP (Institut Français du Pétrole) technology is worth mentioning. This process initially used propane extraction together with acid refining and later together with hydrofinishing. Propane extraction is also used by Snamprogetti (Italy) as the main refining step before and after vacuum distillation. Figure 75 shows the process with propane extraction [553], [554].

Mohawk Technology (CEP – Mohawk). The Mohawk Process (subsequently CEP – Mohawk) using high pressure hydrogenating was introduced in the USA at the end of the 1980s. The process begins with thin-film vacuum distillation of the waste oil (after flashing the light hydrocarbons and water). This is followed by hydrogenation of the distillate at 6900 kPa over a standard catalyst. Special steps realized catalyst life of 8 to 12 months, which was essential for the economy of the process.

A marked reduction in the amount of water which must be treated as effluent as well as the cheaper materials for construction (absence of corrosion) are further advantages. The Mohawk process which is based on the KTI process has been licensed for Evergreen Oil (USA and Canada).

KTI Process. The KTI (Kinetics Technology International) process combines vacuum distillation and hydrofinishing to remove most of the contamination and additives. The key to the process is the thin-film vacuum distillation to minimize thermal stress through mild temperatures not exceeding 250 °C.

The hydrofinisher removes sulfur, nitrogen and oxygen. The yield of finished base oils is high (82 % on a dry waste oil basis). Figure 76 shows the flow chart of this process.

PROP Process (Fig. 77). The PROP technology was developed by Phillips Petroleum Company. The key elements of the process are the chemical demetallization (mixing an aqueous solution of diammonium phosphate with heated base oils) and a hydrogenation process. A bed of clay is used to adsorb the remaining traces of contaminants to avoid poisoning of the Ni/Mo catalyst.

Safety Kleen Process. The Safety Kleen process uses atmospheric flashing for removing water and solvents, a vacuum fuel stripper, vacuum distillation with two thin-film evaporators, and a hydrotreater with fixed bed Ni/Mo catalysts. When using high severity the hydrotreater can reduce the content of polynuclear aromatics; it also removes higher boiling chlorinated paraffins.

Figure 78 shows a simplified block diagram. In 1998 the Safety Kleen process was used in the largest waste oil re-refinery in the world (East Chicago, Indiana, USA, plant capacity 250 000 t/a).

DEA Technology (Fig. 79). The best results with regard to the technical and environmental quality of the re-refined oil and the elimination of PAH are provided by a combination of thin film distillation followed by selective solvent extraction. In this process, the distillate from vacuum thin-film distillation towers equipment at the re-refinery (Dollbergen/Germany) are finally treated in a lube refinery solvent extraction plant followed by hydrofinishing (DEA, Hamburg/Germany). After this extraction process, the PAH content is lower than that of virgin solvent neutrals [554].

Other Re-refining Technologies. *Vaxon* (Enpotec fabrication facilities in Denmark)

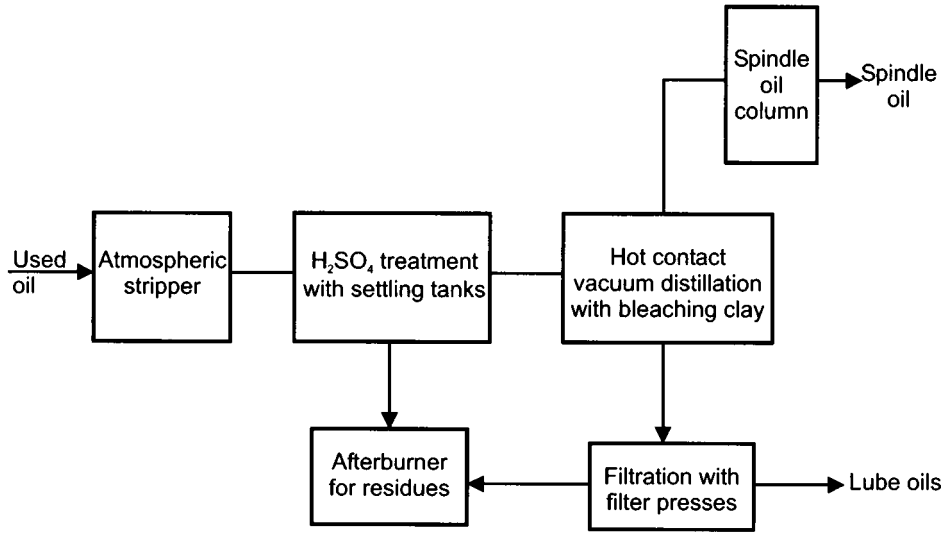

Figure 74. Flow sheet of sulfuric acid re-refining (Meinken process) [553]

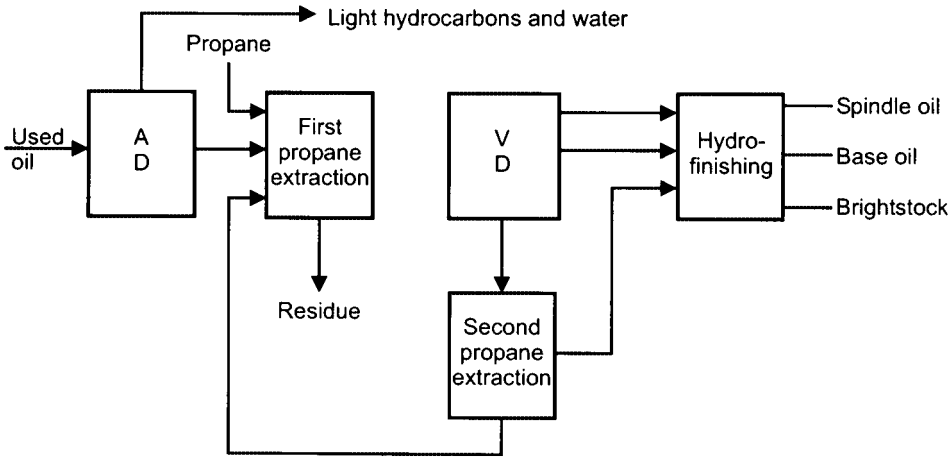

Figure 75. Re-refining by propane extraction (IFP, Snamprogetti)

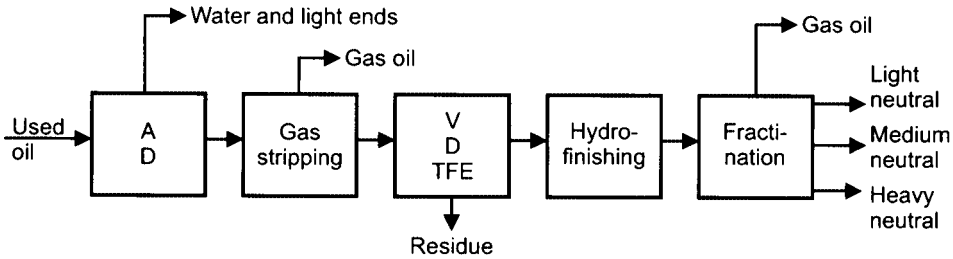

Figure 76. Flow chart of the KTI Process: thin film evaporator (TFE) with hydrotreatment

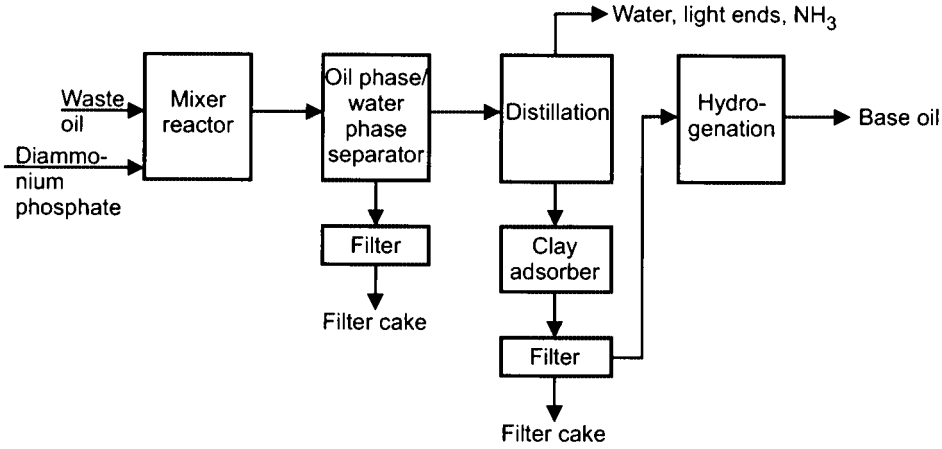

Figure 77. Flow chart of the PROP Process

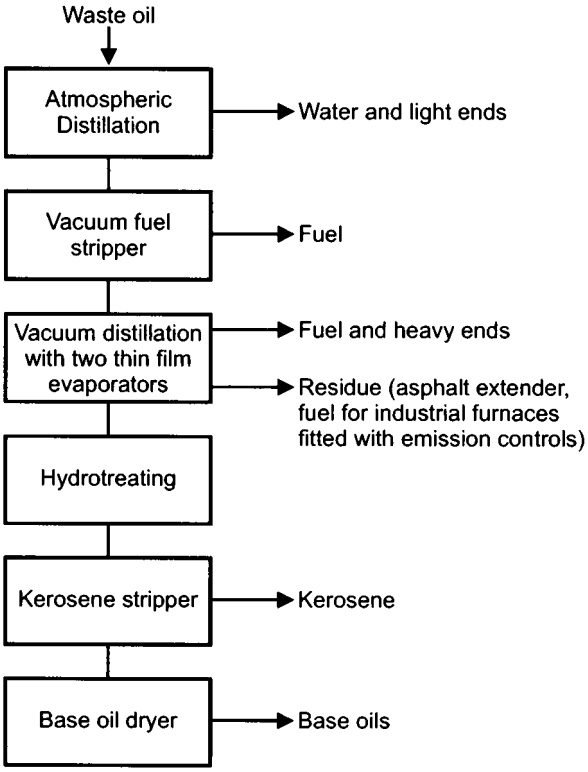

Figure 78. Safety Kleen process

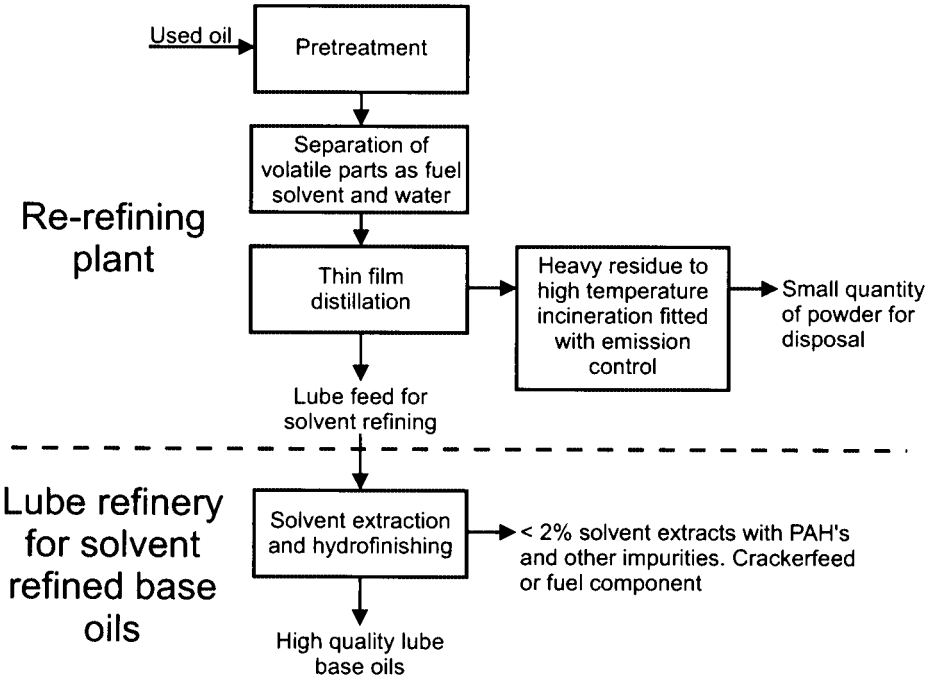

Figure 79. Introduction of selective solvent extraction in the re-refining process (DEA / Mineralöl-Raffinierie Dollbergen, Germany)

uses three or four vacuum cyclone evaporators and finishing treatment with chemicals for re-refining of lubricating oils.

The key step in the *ENTRA technology* is the special vacuum evaporation in a vacuum linear tubular reactor (single tube). After continuous evaporation by means of rapidly increasing temperature, vapor condensation is performed by fractional condensation. Complete dechlorination can be achieved with metallic sodium. Clay polishing is used as a finishing process.

The *thermal deasphalting* (TDA) process has been developed by Agip Petroli/Viscolube using the technology of PIQSA Ulibarri in Spain. The process is based on chemical treatment to facilitate subsequent deasphalting.

The deasphalting process is combined with high fractionating efficiency (TDA unit). Finishing can be performed by clay treatment or hydrofinishing.

19. Toxicology and Occupational Health

19.1. Safety Aspects of Handling Lubricants (Working Materials)

Lubricants are working materials when personnel contact with these substances via contact with the skin or clothes or by inhalation or swallowing. Toxicological, medical and industrial hygiene aspects thus have to be considered.

19.1.1. Polycyclic Aromatic Hydrocarbons (PAK, PAH, PCA)

The largest proportion of raw materials in metalworking lubricants are high boiling point petroleum cuts which contain very few polycyclic hydrocarbons (PAH).

Neat metalworking oils are in the forefront of the PAH discussion because these can directly contact a comparatively large number of people via the inhalation of oil mists. Efforts were already made in the past to evaluate the cancer risks posed by neat cutting oils by determining their concentration of polycyclic aromatic hydrocarbons.

19.1.2. Nitrosamines in Cutting Fluids

Discussions on this subject were triggered by the use of alkali metal nitrites as corrosion inhibitors. The most prominent was and is sodium nitrite, a widely available and cheap substance with good inhibitor properties. Reaction of secondary amines with nitrites can lead to the formation of nitrosamines, 80 % of which are carcinogenic. Sodium nitrite can be found, above all, in low-mineral oil (semisynthetic) or hydrocarbon-free (fully synthetic) solutions. These concentrations are so high that water-miscible application concentrations of 0.05 to 0.2 % can be found. If other, mostly organic inhibitors are used, significantly lower concentrations can be used. Apart from the use of sodium nitrite in water-miscible cutting fluid concentrates, it can also be directly added to water-miscible cutting fluids by consumers.

While the carcinogenic effect of some nitrosamines has been known since the mid-1950s, it was assumed that the reaction between nitrite and amines could not take place in cutting fluids because these were alkaline and the reaction requires an acidic environment. However, if cutting fluid mists are swallowed, nitrosamines can be formed in the acid milieu of the stomach.

High-performance liquid chromatography and other analytical methods were used to prove the presence of between 0.02 and 2.99 % of diethanol nitrosamine in cutting fluid concentrates of fully synthetic products. The pH of the products was between 9 and 11.

The nitrosamine discussions have led a number of cutting fluid manufacturers to develop nitrite-free cutting fluids. However, it must be remembered that in the past, a much larger proportion of water-miscible cutting fluids were free of nitrite.

19.2. Skin Problems Caused by Lubricants

In the metalworking industry, a large proportion of the problems are caused by contact with lubricants. Chip-forming machining operations with water-miscible and neat cutting oils are in the forefront because of the large number of contact possibilities.

Oil acne (particle acne) is one of the most common skin problems caused by contact with neat cutting and grinding oils. Usually it is not the oil itself but small particles such as metal fragments which are the cause. This is why this problem is often referred to as particle acne. Oil acne can appear wherever oil directly contacts the skin and this can include oil-stained clothing. Body areas particularly at risk are the lower and upper arms, the backs of hands, the face, thighs and waist.

Improvements in machine tools such as machine encapsulation, fume extraction, the avoidance of oil mists and general automation means that personnel have less contact with oil and the cases of oil acne have fallen significantly in recent years. The frequency of oil acne cases in factories is a measure of the hygiene standards in force and the personal cleanliness of machine room personnel.

Oil Eczema. The term ‘oil eczema’ encompasses several skin problems. These look like scaly or wet reddened areas, sometimes with a cracked surface.

Acute toxic eczema is caused by the direct effect of the substance on the skin and causes a particular appearance.

Degenerative eczema is the most important skin problem caused by water-miscible cutting fluids. Long-term contact between the cutting fluid and the skin produces signs of degeneration and a breakdown of the skin’s defenses. Key factors are the alkalinity of the fluid, the long-term wetness and specific product ingredients and particularly boundary-active substances from emulsifiers.

In microbiological eczemas, pathogenic and apathogenic germs play a part. There are considerable differences in opinion concerning the significance of microbiological eczemas in metalworking. One question is whether the large

number of germs in water-miscible cutting fluids promote this type of eczema. Some experience indicates that no hygienic risk is posed by germ numbers of 10^5 or 10^6 m/L if no particularly pathogenic germs are present.

Long-term degeneration, for example, caused by exposure to alkalis in water-miscible cutting fluids can ease the ingress of allergens into the skin and thus start the oversensitivity process.

Allergic skin diseases are relatively seldom in metalworking companies. However, the increasing number of chemically active substances in cutting fluids may increase the oversensitivity of skin. Such substances are often found in biocides, EP additives, and corrosion inhibitors.

20. References

General References

Important Books on Greases

1. E. N. Klemgard, *Lubricating Greases: Their Manufacture and Use,* Reinhold, New York 1937.
2. C. J. Boner, *Manufacture and Application of Lubricating Greases,* Reinhold, New York 1954.
3. C. J. Boner, *Modern Lubricating Greases,* Scientific Publications, Broseley 1976.
4. W. Mader, *Hinweise zur Anwendung von Schmierfetten,* Curt R. Vincentz, Hannover 1979.
5. G. Froishteter, K. K. Triliski, Y. L. Ishchuk, P. M. Stupak in G. V. Vinogradov (ed.): *Rheological and Thermophysical Properties of Grease,* Gordon and Breach, London 1989.
6. M. Ehrlich (ed.): *Lubricating Grease Guide,* 4th ed., NLGI, Kansas City 1996.
7. J. L. Ischuk, Composition, Structure and Properties of Plastic Lubricants, Kiew, 1996 (Russian).
8. A. T. Polishuk, A Brief History of Lubricating Greases, *Grease Technology*, Lima, 1998.
9. C. Balan (Ed.), The rheology of Lubricating grease, ELGI Publications, Amsterdam, 2000.
10. W. J. Bartz (ed.): *Schmierfette,* expert verlag, Renningen-Malmsheim 2000.

General References

Important Books on Solid Lubricants

11. A. Bartel, Festschmierstoffe – Aufbau und Anwendungsbereiche, VDI-Berichte 1966, 111, 63 – 66.
12. W. J. Bartz, J. Xu, Describing Wear Life of Bonded Solid Lubricants by Laboratory Methods, *Lubrication Engineering* **41** (1985) 607 – 613.
13. W. J. Bartz, R. Holinski, Synergistische Effekte bei der Anwendung von Festschmierstoffen in gebundener Form (Synergistic Effects with the Application of Bonded Solid Lubricants), *Tribologie + Schmierungstechnik* **33** (1986) no. 3, 140 – 146 **33** (1986) no. 4, 223 – 227.
14. W. J. Bartz, R. Holinski, J. Xu, Wear Life and Frictional Behavior of Bonded Solid Lubricants, *Lubrication Engineering* **42** (1986) no. 12, 762 – 769.
15. W. J. Bartz, J. Xu, Wear Behavior and Failure Mechanism of Bonded Solid Lubricants, *Lubrication Engineering* **43** (1987) no. 7, 514 – 521.
16. W. J. Bartz, R. Holinski, J. Xu, Tribological Behavior of Two- and Three-Component Bonded Solid Lubricants, *Wear* **115** (1987) 167 – 184.
17. W. J. Bartz, Einfluß des Einlaufs auf die Lebensdauer von gebundenen Festschmierstoffen (Gleitlacken), *Tribologie + Schmierungstechnik* **36** (1989) no. 5, 247 – 254.
18. K. G. Budinski, Control of Fretting Corrosion, *The Solid Films* **64** (1979) 359 – 363.
19. G. Dornhöfer, Einfluß des Einlaufs auf die Lebensdauer von gebundenen Festschmierstoffen (Gleitlacken), *Tribologie + Schmierungstechnik* **41** (1994) no. 2, 86 – 90.
20. M. Ferretti, Verfahren zur Erzeugung verschleißfester Oberflächen, *Tribologie + Schmierungstechnik* **43** (1996) no. 5, 269 – 271.
21. J. Gänsheimer, Optimierung von Schraubenverbindungen durch Festschmierstoffe, *Tribologie + Schmierungstechnik* **38** (1991) no. 1, 16 – 23.
22. J. Gänsheimer, J. Neunkirchner, O. Schanzer, R. Schneider, Lebensdaueruntersuchungen an Gleitlackschichten, *Tribologie + Schmierungstechnik* **40** (1993) no. 5, 294 – 299.
23. J. Gänsheimer, Leicht lösbar – Schäden vermeiden bei Hochtemperatur-Schraubenverbindungen mit Hilfe von Festschmierstoffen, *Maschinenmarkt* **98** (1992) no. 18, 94 – 101.
24. R. Holinski, The Influence of Boundary Layers on Friction, *Wear* **56** (1979) 147 – 154.

25. R. Holinski, Gleitlacke zur Unterstützung der Ölschmierung in Kraftfahrzeugen, *Tribologie + Schmierungstechnik* **38** (1991) no. 2, 82 – 85.
26. W. Hübner, T. Gradt, H. Börner, R. Döring, Tieftemperatur-Reibverhalten von Polymer-Werkstoffen, *Tribologie + Schmierungstechnik* **42** (1995) no. 5, 244 – 251.
27. A. Knappwost, D. Fuhrmann, Entwicklung von Mehrkomponenten-Schmierstoffen, *Tribologie + Schmierungstechnik* **36** (1989) no. 2, 81 – 86.
28. W. Läpple, Bedarfs- oder Alleinschmierstoff – Gleitlacke als Alternative zu herkömmlichen Schmierstoffen, *Industrie-Anzeiger* (1990) no. 9, 62 – 64.
29. W. Läpple, Bedarfs- oder Schmieren ohne Schmutz, *Industrie-Anzeiger* (1992) no. 4, 40 – 41.
30. W. Läpple, Schmierwirksame Lacke, *farbe + lack* **88** (1982) no. 2, 105 – 108.
31. L. C. Lipp, Solid Lubricants – Their Advantages and Limitations, *Lubrication Engineering* **32** (1976) no. 11, 574 – 584.
32. J. Neunkirchner, R. Schneider, Schwingungsverschleiß (Tribokorrosion), *Tribologie + Schmierungstechnik* **39** (1992) no. 3, 129 – 134.
33. J. Neunkirchner, R. Schneider, Optimierung von Tribosystemen bei schweren Gleitbeanspruchungen, *Tribologie + Schmierungstechnik* **40** (1993) no. 1, 16 – 22.
34. Product Review, Solid lubricants, *Industrial Lubrication and Tribology* **47** (1995) no. 6, 7 – 18.
35. A. Schröer, A. Savan, E. Pflüger, Reibungs- und Verschleissminderung durch Festschmierstoffe, Technische Berichte *T+S Tribologie und Schmierungstechnik*, Bericht Nr. 16, 1998, ISSN 0342-4288, 1352.
36. W. Seßler, Bestimmung der Reibungszahlen von gleitlackbeschichteten Schraubenverbindungen, *Tribologie + Schmierungstechnik* **35** (1988) no. 6, 319 – 325.
37. W. Seßler, Schwingungsverschleiß von Gleitlacken, *Tribologie + Schmierungstechnik* **37** (1990) no. 2, 91 – 96.
38. H. Winkler, Einsatzbedingungen für Tribo-System-Materialien, *Industrie-Anzeiger* (1996) no. 4, 37 – 39.
39. F. Wunsch, Möglichkeiten und Grenzen der Feststoffschmierung, *Tribologie + Schmierungstechnik* **29** (1982) no. 4, 138 – 144.
40. R. Zechel, R. Holinski, Feststoffschmierung – Wirkungsmechanismen und industrieller Einsatz Teil 1, *Tribologie + Schmierungstechnik* **40** (1990) no. 3, 168 – 175.
41. R. Zechel, R. Holinski, Feststoffschmierung – Wirkungsmechanismen und industrieller Einsatz Teil 2, *Tribologie + Schmierungstechnik* **40** (1990) no. 4, 241 – 247.
42. R. Zechel, R. Holinski, Feststoffschmierung – Wirkungsmechanismen und industrieller Einsatz Teil 3, *Tribologie + Schmierungstechnik* **40** (1990) no. 5, 300 – 305.
43. J. Gänsheimer, Schmierstoffe für oszillierende Bewegungen, *Der Zuliefermarkt*, Nov. 1988, Sonderteil in Hanser-Fachzeitschriften, Carl Hanser Verlag, München 1988, 1063.
44. J. Gänsheimer, Schmierstoffe für oszillierende Bewegungen – ein neuer Fortschritt in der Tribotechnik, *Konstruktion* **42** (1990) 391 – 395.
45. F. Wunsch, Festschmierstoffe, Theorie und Praxis, reprint from *Ingenieur Digest* **13** (1974) no. 12, **14** (1975) no. 1 – 3.
46. E. R. Braithwaite, *Solid Lubricants and Surfaces,* Pergamon Press, London 1964.
47. D. Klamann, *Schmierstoffe und verwandte Produkte: Herstellung, Eigenschaften, Anwendung,* Verlag Chemie, Weinheim 1982.
48. U. J. Möller, U. Boor, *Schmierstoffe im Betrieb,* VDI Verlag GmbH, Düsseldorf 1986.
49. O. Pigors, *Werkstoffe in der Tribotechnik,* Deutscher Verlag für Grundstoffindustrie, Leipzig-Stuttgart 1993, pp. 72 – 74.
50. U. Lippmann, Lebensdaueruntersuchungen an Gleitpaarungen mit Trockenschmierfilmen, Ingenieurarbeit an der TU Chemnitz, Fachbereich Maschinenbau II, Institut für Werkzeugmaschinen, Lehrstuhl Tribotechnik, 1992.
51. T. Mang, W. Diesel (eds.): *Lubricants and Lubrications,* Wiley-VCH, Weinheim, 2001
52. H. Czichos, *Basic Tribological Parameters, Friction, Lubrication and Wear Technology, ASTM Handbook,* **Vol. 18,**1992, p. 474.
53. H. Czichos, K.-H. Habig, *Tribologie Handbuch* , Vieweg, Wiesbaden 1992.
54. D. Klamann, *Lubricants and Related Products* , VCH Verlagsgesellschaft, Weinheim 1984.
55. H. Holland, Information of the Institut für Reibungstechnik und Maschinenkinetik,

Technische Universität Clausthal, Clausthal-Zellerfeld.
56. T. Mang, *Lubricants for the Next Century*, International Conference on Industrial Tribology, Calcutta, 1997, *63*, 639 – 653.
57. G. Brandes, *Brennstoff-Chem.* **37** (1956) 263 – 267.
58. IARC, *Classification of Mineral oils According to their Carcinogenicity,* **Vol. 33,** 1984.
59. S. J. Miller, M. A. Shippey, G. M. Masada, *Advances in Lube Base Oil Manufacture by Catalytic Hydroprocessing*, 1992 NPRA National Fuels and Lubricants Meeting, Houston, 1992.
60. S. M. Jacob, *Lube Base Oil Processing for the 21st Century*, 4th Annual Fuels and Lubes Asia Conference, Singapore, 1998.
61. J. Weitkamp, H. G. Karge, H. Pfeifer, W. Hölderich, *Zeolites and Related Microporous Materials: State of the Art 1994*, 10th International Zeolite Conference, Garmisch-Partenkirchen, 1994.
62. S. J. Miller, New Molecular Sieve Process for Lube Dewaxing by Wax Isomerization, *Microporous Materials* **2** (1994) 439 – 449.
63. E. S. Johnson, K. Thomas, *Excel Paralubes Base Oils for Tomorrow*, National Fuels and Lubricants Meeting, Houston. 1995.
64. S. C. Cohen, P. D. Mack, *HVI and VHVI Base Stocks*, The World Base Oils Conference, London, 1996.
65. W. S. Moon, Y. R. Cho, C. B. Yoon, Y. M. Park, *VHVI Base Oils from Fuels Hydrocracker Bottoms*.
66. P. Y. Min, *VHVI Base Oils: Supply and Demand*, 4th Annual Fuels and Lubes Asia Conference, Singapore, 1998.
67. T. Mang, Future Importance of Base Oils in Lubricants, *12th International Colloquium Tribology,* Stuttgart/Ostfildern, 2000.
68. R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants,* Reinhold, New York, 1962.
69. R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993.
70. A. Madhusudhana Rao, S. P. Srivastava, K. C. Mehta, Synthetic Lubricants in India – An Overview, *J. Synth. Lubr.* **4** (1987), 137 – 145.
71. W. J. Bartz, Comparison of Synthetic Fluids, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 319 – 336.
72. E. I. Williamson, Commercial Developments in Synthetic Lubricants – A European Overview*J. Synth. Lubr.* **2** (1985) 329 – 341.
73. A. Plagge, Gebrauchseigenschaften synthetischer Schmierstoffe und Arbeitsflüssigkeiten, *Tribologie + Schmierungstechnik* **32** (1985/1987) 270 – 278/148 – 156.
74. R. L. Shubkin, Polyalphaolefins, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 1 – 40.
75. W. J. Bartz, Comparison of Synthetic Fluids, *Lubrication Engineering* **48** (1992) 765 – 774; W. J. Bartz, Vergleich synthetischer Flüssigkeiten, *Mineralöltechnik* **38** (1993) 9.
76. E. Jantzen, The Origins of Synthetic Lubricants: The Work of Hermann Zorn in Germany Part 1 Basic Studies of Lubricants and the Polymerization of Olefins, *J. Synth. Lubr.* **12** (1995) 283 – 301.
77. US Patent 2 559 984, 1951 (C. W. Montgomery, W. I. Gilbert, R. E. Kline).
78. R. M. Mortier, S. T. Orszulik et al.: Synthetic base fluids, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants (2.)* Blackie, London, 1997, pp. 34 – 74.
79. G. Kumar, R. L. Shubkin, New Polyalphaolefin Fluids for Specialty Applications, *Lubrication Engineering* **49** (1993), 723 – 725.
80. D. Christakudis, Synthetische Schmieröle und ihre Bedeutung für die Herstellung von Motorenölen – Ein Überblick, *Schmierungstechnik* **17** (1986) 232 – 237.
81. U. Kristen, K. Müller, R. Schumacher, D. Chasan, Aschefreie EP/AW – Additive für PAO, *Ester und Polyalkylenglykole, 4*. International Colloquium, Esslingen, 1984, 56.1 – 8.
82. G. B. van der Waal, Properties and Applications of Ester Base Fluids and PAO's, *NLGI Spokesman* **53** (1989) 359 – 368.
83. V. M. Shkolnikov, O. N. Zvetkov, M. A. Chagina, G. V. Kolessova, Improvement of Antioxidation and Antiwear Properties of Polyalphaolefin Oils, *J. Synth. Lubr.* **7** (1990) 235 – 241.
84. R. Benda, J. Bullen, A. Plomer, Polyalphaolefins – Base Fluids for High-Performance Lubricants, *J. Synth. Lubr.* **13**(1996) 41 – 57.
85. G. Corsico. P. Pacor, M. Ciali, A. Zatta, N. Gatti, *PIO: A New Synthetic Basestock*, Presentation of Mixoil.
86. J. Fotheringham: Polybutenes, in R. L. Shubkin (ed.): *Synthetic Lubricants and*

High-Performance Functional Fluids Marcel Decker, New York, 1993, pp. 271 – 318.
87. B. Wilson, Polybutenes – The Multipurpose Base Oil and Additive, *Industrial Lubrication and Tribology* **46** (1994) 3 – 6.
88. J. Fotheringham, Hyvis/Napvis Polybutene, *Tribologie + Schmierungstechnik* **42** (1995) 92 – 96.
89. DE 565 249, 1930 (H. Zorn, M. Mueller-Conradi, W. Rosinski)
90. H. Dressler: Alkylated Aromatics, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 125 – 144.
91. E. D. Sowle, T. M. Lachocki, Linear Dialkylbenzenes as Synthetic Base Oils, *Lubrication Engineering* **52**(1996) 116 – 120.
92. L. J. Gschwender, D. Mattie, C. E. Snyder, W. M. Warner, Chlorotrifluoroethylene Oligomer Based Nonflammable Hydraulic Fluid. 1 Fluid, Additive, and Elastomer Development, *J. Synth. Lubr.* **9** (1992) 187 – 202.
93. E. Jantzen, The Origins of Synthetic Lubricants: The Work of Hermann Zorn in Germany Part 2 Esters and Additives for Synthetic Lubricants, *J. Synth. Lubr.* **12** (1995) 283 – 301.
94. M. Wildersohn, Esteröle – Struktur und chemisch – physikalische Eigenschaften, *Tribologie + Schmierungstechnik* **32** (1985) 70 – 78.
95. W. G. Dudek, A. H. Popkin, Dibasic Acid Esters, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants* Reinhold, New York, 1962, pp. 151 – 245.
96. P. W. Hurd, The Chemistry of Castor Oil and its Derivatives, *NLGI Spokesman* **60** (1996) 14 – 23.
97. A. Feßenbecker, I. Roehrs, A New Additive for the Hydrolytic and Oxidative Stability of Ester based Lubricants and Greases, *NLGI Spokesman* **61** (1997) 310 – 317.
98. T. G. Smith: Neopentyl Polyol Esters, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants* Reinhold, New York, 1962, pp. 388 – 401.
99. B. Koch, E. Jantzen, Thermal Oxidation Behaviour of Synthetic Oils: Basic Studies in the Polymerization of Polyol Ester Oils, *J. Synth. Lubr.* **4** (1987) 321 – 335.
100. Ren Tianhui, Xia Jian, Zheng Longzhen, Zhang Junyan, Liu Weimin, Xue Qunji, A Tribological Study of an S,P-containing Benzotriazole Derivative as an Additive in Pentaerythritol Ester, *J. Synth. Lubr.* **15** (1999) 263 – 269.
101. F. Bongardt, Einfluß der chemischen Struktur auf das Luftabscheidevermögen und die hydrolytische Stabilität von Estern, *Fat Sci. Technol.* **92** (1990) 473 – 478.
102. P. S. Venkataramani, S. L. Kaira, S. V. Raman, H. C. Srivastava, Synthesis Evaluation and Applications of Complex Esters as Lubricants: A Basic Study, *J. Synth. Lubr.* **5** (1988) 271 – 289.
103. H.-G. Schmidt, *Komplexester aus pflanzlichen Ölen*, 9th International Colloquium, Esslingen, Vol II, 1994, pp. 2.2-1 – 2.2-9.
104. C. M. Murphy: Fluoroesters, in R. C. Gunderson, A. W. Hart (eds.:) *Synthetic Lubricants,* Reinhold, New York, 1962, pp. 361 – 387.
105. R. E. Hatton: Phosphate Esters, in R. C. Gunderson, A. W. Hart (eds.:) *Synthetic Lubricants* Reinhold, New York, 1962, pp. 103 – 150.
106. M. P. Marino: Phosphate Esters, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 67 – 100.
107. US 2 425 755, 1947 (F. H. Roberts, H. R. Fife).
108. R. C. Gunderson, W. H. Millett, Polyglycols, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants* Reinhold, New York, 1962, pp. 61 – 102.
109. S. Kussi, Chemical, Physical and Technological Properties of Polyethers as Synthetic Lubricants, *J. Synth. Lubr.* **2** (1985), 63 – 84; S. Kussi, Eigenschaften von Basisflüssigkeiten für synthetische Schmierstoffe, *Tribologie + Schmierungstechnik* **33** (1986) 33 – 39.
110. P. L. Matlock, N. A. Clinton, Polyalkylene Glycols, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 101 – 123.
111. W. H. Gumprecht, PR-143 – A new class of high-temperature fluids, *ASLE Trans.* **9** (1965) 24 – 30.
112. W. Schwickerath, Perfluorether – Basisflüssigkeit für Hochtemperaturschmierstoffe, *Tribologie + Schmierungstechnik* **34** (1987), 22 – 38.
113. T. W. Del Pesco: Perfluoroalkylpolyethers, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 145 – 172.

114. C. Corti, P. Savelli, Perfluoropolyether Lubricants, *J. Synth. Lubr.* **9** (1992) 311 – 330.
115. D. J. Carré, Perfluoropolyalkylether Lubricants under Boundary Conditions: Iron Catalysis of Lubricant Degradation, *J. Synth. Lubr.* **6** (1990) 3 – 15.
116. C. L. Mahoney, E. R. Barnum: Polyphenyl Ethers, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants,* Reinhold, New York, 1962, pp. 402 – 463.
117. R. W. Awe, H. M. Schiefer: Silicones, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants,* Reinhold, New York, 1962, pp. 264 – 322.
118. D. H. Demby, S. J. Stoklosa, A. Gross: Silicones, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 183 – 203.
119. P. Huber, Siliconöle, *Tribologie + Schmierungstechnik* **32** (1985), 64 – 69; P. Huber, W. Kaiser, Silicone Fluids: Synthesis, Properties and Applications, *J. Synth. Lubr.* **3** (1986) 105 – 120.
120. M. Jungk, D. Hesse, Silicone Oil-Based Fluids as a Tool to Tailor High-Performance Lubricating Greases, *Eurogrease* (1998) January/February, 31 – 35.
121. G. Baum: Novel Synthetic Lubricants, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants,* Reinhold, New York, 1962, pp. 464 – 489.
122. V. Stepina, V. Vesely: *Lubricants and Special Fluids,* , Elsevier, Amsterdam, 1992, pp. 186 – 191.
123. G. Fisicaro, G. Gerbaz: Dialkylcarbonates, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids* Marcel Decker, New York, 1993, pp. 229 – 239.
124. *Gmelin*, Borverbindungen 6 (1974) 4.9.4 110.
125. R. E. Singler, M. J. Bieberich: Phosphazenes, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 215 – 228.
126. Q. Zhao et al., Tribological Study of Phosphazene-Type Additives in Perfluoropolyether Lubricant for Hard Disk Applications, *Lubrication Engineering* **55** (1999) 16 – 21.
127. H. J. Kang et al., The Use of Cyclic Phosphazene Additives to Enhance the Performance of the Head/Disk Interface, *Lubrication Engineering* **55** (1999) 22 – 27.
128. D. Klamann: *Lubricants and Related Products* , Verlag Chemie, Weinheim, 1984, pp. 116.
129. V. Stepina, V. Vesely, *Lubricants and Special Fluids* , Elsevier, Amsterdam, 1992, pp. 161 – 162.
130. A. R. Lansdown: *High Temperature Lubrication,* Mechanical Engineering Publications, London, 1994, p. 132.
131. F. A. Pettigrew, G. E. Nelson: Silahydrocarbons, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York, 1993, pp. 205 – 214.
132. C. E. Snyder Jr, L. J. Gschwender, C. Tamborski, G. J. Chen, D. R. Anderson, Synthesis and Characterization of Silahydrocarbons – A Class of Thermally Stable Wide-Liquid-Range Functional Fluids, *ASLE Transactions* **25** (1982) 299 – 308.
133. S. G. Thomas, D. G. Campbell, C. J. Hsu, Improvement of Silahydrocarbon Performance with Additives *J. Synth. Lubr.* **10** (1993) 195 – 212.
134. R. E. Hatton: Silicate Esters, in R. C. Gunderson, A. W. Hart (eds.): *Synthetic Lubricants,* Reinhold, New York, 1962, pp. 323 – 360.
135. V. Stepina, V. Vesely: *Lubricants and Special Fluids,* , Elsevier, Amsterdam, 1992, pp. 178 – 180.
136. M. Rasberger, Oxidative Degradation and Stabilisation of Mineral Oil Based Lubricants, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* Blackie Academic and Professional, 1992, pp. 83–93.
137. H. Rayn, Use of group II, group III base stocks in hydraulic and industrial applications, *Proc. the 5th Annual Fuels and Lubes Asia Conference 1999,* Fuels and Lubes Asia Publ., Manila.
138. C. A. Migdal, The Effect of Antioxidants on the Stability of Base Oils, *Proceedings of the 5th Annual Fuels and Lubes Asia Conference 1999*, Fuels and Lubes Asia Publ., Manila.
139. V. J. Gatto, M. A. Grina, Effects of Base Oil Type, Oxidation Test Conditions and Phenolic Antioxidant Structure on the Detection and Magnitude of Hindered Phenol/Diphenylamine Synergism, *Lubrication Engineering* 1999, January, 11–20.
140. M. Rasberger, Oxidative degradation and stabilisation of mineral oil based lubricants, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* Blackie

Academic and Professional, 1992, pp. 97 – 98, 104 – 107, 108, 102 – 103, and 99 – 100.
141. S. C. Cohen, A new synthetic screw compressor fluid based on two-stage hydrotreated base oils, *Synthetic Lubrication* **7** (1991), no 4, 267 – 279.
142. W. J. Bartz (ed.) in: *Additive für Schmierstoffe,* **Bd. 2**, Vincent Verlag, Hannover, 1986, p. 161ff.
143. J. Omeis, H. Pennewiß, *ACS Polymer Preprints***35** (1994) no. 2, 714.
144. U. F. Schödel, Automatic Transmission Fluids, *Int. Tribology Colloquium*, TAE Esslingen, vol. 3, 22.7 – 1, 1992.
145. H. Winter, K. Michaelis, B. M. O'Conor, *Mineralöltechnik* **8** (1986).
146. S. Bair, F. Queresti, W.O. Winer, *ASME/STLE Tribology Conference*, San Diego, 1992, Paper No. 92/3.
147. M. Mortimer, Laboratory shearing tests for viscosity index improvers, *Tribotest Journal* **2–4** (1996) no 6, 329.
148. W. Böttcher, H. Jost, SAE Technical Paper, 912410, 1991.
149. "Petroleum Additives – Pour Point Depressants," *Encycl. Polym. Sci. and Tech.* **9** (1968) 843 – 844.
150. J. H. Raddatz, Detergent-Dispersant Additive, ihre Herstellung, Anwendung und Wirkungsweise, in W. J. Bartz (ed.): *Additive für Schmierstoffe,* Expert Verlag, 1994, pp. 163 – 196.
151. N. J. H. Small, Dipersants, in J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive* Georg Thieme Verlag, Stuttgart, 1978, pp. 235 – 239.
152. W. T. Stewart, F. A. Stuart, Lubricating oil additives, in: *Advances in Petroleum Chemistry and Refining,* **vol. 7,** Chapter 1, Interscience, New York 1963, pp. 3 – 64.
153. EP 699 694, (J. Omeis, H. Pennewiss).
154. US 5 597 871, 1997 (C. Auschra, J. Omeis, H. Pennewiss).
155. K. Hedrich et al, *International Tribology Conference*, Nagasaki, 2000.
156. H. Hellberg, Schaumdämpfer, Farbstoffe und Geruchsüberdecker, in: J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive,* Georg Thieme Verlag, Stuttgart, 1978, p. 257.
157. *Colour Index International, Pigments and Solvent Dyes,* 1997 ed., The Society of Dyers and Colourists, Bradford, West Yorkshire, 1997.
158. E. S. Forbes, Antiwear an Extreme Pressure Additives for Lubricants, *Tribology* **3** (1970), 145.
159. C. G. A. von Eberan-Eberhorst, R. S. Hexter, A. C. Clark, B. O'Connor, R. H. Walsh, Aschegebende Extreme-Pressure- und Verschleißschutz-Additive, in: W. J. Bartz (ed.): *Additive für Schmierstoffe* Expert Verlag, 1994, pp. 53 – 83.
160. D. Christakudis, Friction Modifier und ihre Prüfung, in: W. J. Bartz (ed.): *Additive für Schmierstoffe,* Expert Verlag, 1994, pp. 134 – 162.
161. J. F. Ford, Lubricating oil additives–a chemist's eye view, *J. Inst. Petrol.* **54** (1968) 198.
162. W. F. Olszewski, D. D. Neiswender, Development of a Rear Axle Lubricant, SAE Paper 760, 1976, 326.
163. Administrative Regulation on the Classification of Substances Hazardous to Waters (in German "Verwaltungsvorschrift wassergefährdende Stoffe" – VwVwS), Germany 29.05.1999.
164. Anlagenverordnung, e. g. in Schiegl, Betrieblicher Umweltschutz, Juni 1992.
165. W. Backé, Bestandsaufnahme und Trends der Fluidtechnik, *Ölhydraulik und Pneumatik* **36** (1992) 16 – 19.
166. Bericht über den Einsatz biologisch schnell abbaubarer Schmierstoffe und Hydraulikflüssigkeiten und Maßnahmen der Bundesregierung.
167. Biodegradability of Two-Stroke Cycle Outboard Engine Oils in Water, CEC-L-33-A-93, London: Co-ordinating European Council, 1994.
168. BMFT-Project Emissionsarme Schmierstoffe 01 HZ 8821/5.
169. V. M. Cheng, A. A. Wessol, P. Baudouin, M. T. BenKinney, N. J. Novick Biodegradable and Nontoxic Hydraulic Oils, SAE Technical Paper Series, 42nd Earthmoving Industry Conference, Peoria, Illinois, April 9 – 10, 1991.
170. Collection and disposal of used lubricating oil, CONCAWE-Report no. 5/96.
171. H. W. Dresel, Schmierfette auf pflanzlicher Basis, Technische Akademie Esslingen, March 1990.
172. H. F. Eichenberger, Biodegradable Hydraulic Lubricants – an Overview of Current Developments in Central Europe, SAE Technical Paper Series, 42nd Earthmoving Industry Conference, Peoria, Illinois, April 9 – 10, 1991.

173. Gesetzesblatt für Baden-Württemberg: Verordnung des Umweltministeriums über Anlagen zum Umgang mit wassergefährdenden Stoffen und über Fachbetriebe (Anlagenverordnung-VAwS) vom 11.02.1994.
174. Grundlage für Umweltzeichenvergabe: Biologisch schnell abbaubare Hydraulikflüssigkeiten. RAL UZ 79, Deutsches Institut für Gütesicherung und Kennzeichnung e.V. (RAL), Bonn, Dezember 1995.
175. Grundlage für Umweltzeichenvergabe: Biologisch schnell abbaubare Kettenschmierstoffe für Motorsägen. RAL UZ 48, Deutsches Institut für Gütesicherung und Kennzeichnung e.V., Bonn, März 1988.
176. Grundlage für Umweltzeichenvergabe: Biologisch schnell abbaubare Schmierstoffe und Schalöle. RAL UZ 64, Deutsches Institut für Gütesicherung und Kennzeichnung e.V. (RAL), Bonn, Juni 1991.
177. J. Hübner, Das Problem der N-Nitrosamin-Bildung in wassergemischten Kühlschmierstoffen und die TRGS 611, *Praxis-Forum* 1994.
178. ICOMIA (International Council of Marine Industry Association) Standard 33 – 88 Lubricating Oil for Two-Stroke Cycle Outboard Motors – Ecologically Friendly, 1988.
179. H. Ihrig, M. Spilker, Umweltschonende Schmierstoffe in der Antriebs-Technik – Entwicklung und Ausblick, report at the Antriebstechnische Kolloquium ATK'93, 05.05.1993, published in symposium's report, 1993.
180. H. Ihrig, Erfahrungen mit Schmierstoffen und Hydraulikflüssigkeiten auf Pflanzenölbasis in Verlustschmierung und Langzeiteinsatz, *Mineralöltechnik* **8** (1990) 180 – 185.
181. H. Ihrig, Schmierstoffe und Umwelt, TAE-Esslingen, 1994.
182. H. Ihrig, Umweltverträgliche Schmierstoffe in den 90er Jahren, offprint from *Tribologie und Schmierungstechnik***3** (1992) 121 – 125.
183. R. Luther, Einsatz von Pflanzenölen in Motorenölen, in: Gülzower Fachgespräche – Statusseminar "Biologisch schnell abbaubare Schmier- und Verfahrensstoffe", 1999, pp. 125 – 141.
184. R. Luther, Motorenöle auf der Basis von Pflanzenölen – ein neuer Ansatz, *Fett/Lipid* **100** (1998) 7 – 12.
185. T. Mang, R. Turek, H. Ihrig, Rapsöl als Basis von Schmierstoffen, *RAPS–:Fachzeitschrift für Öl- und Eiweißpflanzen* **7** (1989) 192 – 194.
186. T. Mang, *Die Schmierung in der Metallbearbeitung,* Vogel-Buchverlag Würzburg 1983.
187. T. Mang, Legislative influences on the development, manufacture, sale and application of lubricants in the Federal Republic of Germany, Report on the CEC Symposium, 19. – 21.04.1990, Paris.
188. T. Mang, Lubricants and Legislation in the Federal Republic of Germany, *Erdöl und Kohle* **42** (1989) 400 – 407.
189. T. Mang, Rechtsrahmen und wirtschaftliche Verwendung umweltfreundlicher Schmierstoffe TAE-Esslingen, 1993.
190. T. Mang, Schmierstoffe und Funktionsflüssigkeiten auf Pflanzenölbasis, Erfahrungen eines Herstellers. Schriftenreihe des Bundesministers für Ernährung, Landwirtschaft und Forsten, Reihe A, Issue 391, 1990.
191. T. Mang, Schmierstoffe und Funktionsflüssigkeiten aus Pflanzenölen und deren Derivaten, DECHEMA Kolloquium, 05.03.1992, Frankfurt.
192. T. Mang, Umweltschonende und arbeitsfreundliche Schmierstoffe, *Tribologie + Schmierungstechnik* **38** (1991) 231 – 236.
193. U. J. Möller Entsorgung umweltschonender Flüssigkeiten, *Ölhydraulik + Pneumatik* **34** (1990) 6.
194. S. Odi-Owei, Tribological Properties of some vegetable oils and fats, *Lubricating Engineering* **45** (1988) 685 – 690.
195. E. Pelzer, Normung und Vergaberichtlinien des Umweltzeichens von umweltschonenden Hydraulikflüssigkeiten, Technische Akademie Esslingen, 12. – 15.02.1992.
196. Pflanzenöle als Schmierstoffe (Vegetable Oils as Lubricants – Working sheet): German Society for Tribology 1993.
197. S. D. Scott, Biodegradable Fluids for Axial Piston Pumps – Motors – Application Considerations, SAE Technical Paper Series, 42nd Earthmoving Industry Conference, Peoria, Illinois, April 9 – 10, 1991.
198. E. M. Stempfel, L. A. Schmid, Biodegradable Lubricating Greases, *57th Annial Meeting of NLGI*, Denver, Colorado, October 1990.
199. R. Turek, Schmierstoffe und Funktionsflüssigkeiten für Land-, Forst- und Baumaschinen, Technische Akademie Esslingen, 14 – 16.03.1990.

200. VDMA-Einheitsblätter 24568 und 24569 Biologisch schnell abbaubare Druckflüssigkeiten.
201. M. Fuchs, The World Lubricants Market–Current Situation and Outlook, TAE Esslingen, *Tribology 2000 Plus, 12th International Colloquium*, 11 – 13 January, 2000, vol. I , p. 9.
202. A. Hubmann, K. Reifer, W. Baumann, *Mineralöltechnik* (1994) 4.
203. J. Dearlove, W. K. Cheng, Simultaneous Piston Ring Friction and Oil Film Thickness Measurements in a Reciprocating Test Rig, SAE Paper 952470, 1995.
204. A. T. Riga, S. H. Roby, *Lubrication Engineering* **50** (1994) 441 – 417.
205. K. O. Kenderson, J. M. Sticking, The Effect of Shear Rate and Shear Stress on Viscosity Determinations at Low Temperature for Engine Oils, SAE Paper No. 902091, 1990.
206. T. W. Selby, Further Consideration of Low Temperature, Low Shear rheology Related to Engine Oil Pumpability, SAE Paper 852115, 1985.
207. T. Mang, International Conference on Tribology, Calcutta, 1997.
208. W. Macg, A. Gervé, IAVF Information Brochure, Karlsruhe, 2000.
209. R. J. Prince, Base oils from petroleum, in: R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* Blackie, London 1992.
210. M. Rasberger in *Chemistry and Technology of Lubricants,* Wiley, London, 1988, p. 83ff.
211. ASTM Standard Test Method D 4684 Determination of Yield Stress and Apparent Viscosity of Engine Oils at Low Temperature.
212. R. E. Davies, M. R. Draper, B. J. Lawrence, D. Park, A. M. Seeney, G. C. Smith, Lubricant Formulation Effects on Oil Seal Degradation, SAE Paper 952340, 1995.
213. C. van Eberan-Eberhorst, M. A. Cantab; *Mineralöltechnik* (2000) 4.
214. E. Choi, K. Akiyama, T. Astida, K. Kado, F. Ueda, H. Otira, Engine Testing Comparison Of The Relative Oxidation Stability Performance Of Two Engine Oils, SAE Paper 952530, 1995.
215. S. Jahanmir, Examination of wear mechanisms in automotive camshafts, *Wear* **108** (1986) 235 – 254.
216. M. J. Coritch, How Polymer Architecture Affects Permanent Viscosity Loss of Multigrade Lubricants, SAE Paper 982638.
217. W. van Dam, S. H. Broderick, R. L. Freerks, V. R. Small, W. W. Willis, TBN Retention–Are We Missing the Point? SAE Paper 972950, 1997.
218. M. T. Devlin, W. Y. Lam, T. F. McDonnell, Comparison of the Physical and Chemical Changes Occuring in Oils During Aging in Vehicle and Engine Fuel Economy Tests, SAE Paper 982504, 1998.
219. M. Völtz in *Biologisch schnell abbaubare Schmierstoffe und Arbeitsflüssigkeiten,* expert verlag, 1993, 136 – 157.
220. W. J. Bartz, *Handbuch der Betriebsstoffe für Kraftfahrzeuge,* **Teil 2,** Expert Verlag 1983.
221. C. von Eberan-Eberhorst, Lubrication of Two-stroke and Four-stroke Engines, SAE Technical Paper 937051, SAE, 1993.
222. E. G. Groff, B. L. Sheaffer, W. P. Johnson, Two-stroke Engines and Emissions, SAE Special Publication SP-1327, SAE, 1998.
223. H. Kawabe, Y. Konishi, Effect of Polybutene on Prevention of Clogging in the Exhaust System of Two-stroke Engines, SAE Technical Paper 93A085, SAE, 1993.
224. G. J. Souillard, F. van Quaethoven, R. B. Dyer, Polyisobutylene, a New Synthetic Material for Lubrication, SAE Technical Paper 710730, SAE, 1971.
225. International Council of Marine Industry Associations, Standard 27-97, Egham, UK, 1997.
226. W. B. Chamberlin, T. T. Curtis, D. M. Smith, Crankcase Lubricants for Natural Gas Transportation Applications, SAE Technical Paper 961920, SAE 1996.
227. A. Zareh, An overview of industrial gas engines and their lubrication–Part 1, *Lubrication Engineering* **52** (1996) 10.
228. A. Zareh, An overview of industrial gas engines and their lubrication–Part 2, *Lubrication Engineering* **52** (1996) 110.
229. B.H. Carter, M. Mortier, S.T. Orszulik (eds.): *Marine lubricants, in: Chemistry and Technology of Lubricants,* VCH Verlagsgesellschaft, Weinheim 1992.
230. GFT-Arbeitsblatt 3, Wälzlagerschmierung, Gesellschaft für Tribologie, 1993.
231. H. Jünemann, *Tribometrie–Prüf- und Meßtechnik,* Expert-Verlag, Böblingen 1997.
232. W. J. Bartz, *Lubrication of Gearing,* Expert-Verlag, Böblingen 1998.
233. D. Dowson, G. R. Higginson, *Elastohydrodynamic Lubrication,* 2nd ed., Pergamon Press, Oxford 1977.

234. B. J. Hamrock, D. Dowson, Isothermal elastohydrodynamic lubrication of point contacts, Part III–Fully Flooded Results, *ASME, Journal of Lubrication Technology* **99** (1977) 264 – 276.
235. DIN 3990, Tragfähigkeitsberechnung von Stirn- und Kegelrädern.
236. P. Oster, Beanspruchung der Zahnflanken unter Bedingungen der Elastohydrodynamik, Dissertation, TU München, 1982.
237. E. J. Wellauer, G. A. Holloway, Application of EHD Oil Film Theory to Industrial Gear Drives, *Trans. ASME, J. Engn. for Industry* **98** (1976) 626 – 634.
238. DIN 4768 Teil 1Messung von Oberflächen.
239. ISO DIS 10825, Gears–Wear and Damage to Gear Teeth.
240. H. J. Plewe, Untersuchungen über den Abriebverschleiß von geschmierten, langsamlaufenden Zahnrädern, Dissertation, TU München, 1980.
241. K. Michaelis, Die Integraltemperatur zur Beurteilung der Freßtragfähigkeit von Stirnradgetrieben, Dissertation, TU München, 1987.
242. K. Michaelis, Freßtragfähigkeit für Hochleistungs-Hypoidgetriebe-Schmierstoffe, *Mineralöltechnik* **23** (1978) 13.
243. ISO CD 13989, Gears–Calculation of Scuffing Load Capacity of Cylindrical, Bevel and Hypoid Gears.
244. G. Schönnenbeck, Einfluß der Schmierstoffe auf die Zahnflankenermüdung (Graufeckigkeit und Grübchenbildung) hauptsächlich im Umfangsgeschwindigkeitsbereich 1 bis 9 m/s, Dissertation, TU München, 1984.
245. W. Käser, Beitrag zur Grübchenbildung an gehärteten Zahnrädern. Einfluß von Härtetiefe und Schmierstoff auf die Flankentragfähigkeit, Dissertation, TU München, 1977.
246. FVA-Informationsblatt 2/IV, Einfluß des Schmierstoffes auf die Grübchenlebensdauer einsatzgehärteter Zahnräder im Einstufen- und im Lastkollektivversuch, 1997.
247. DIN 51509 Teil 1, Auswahl von Schmierstoffen für Zahnradgetriebe.
248. W.J.. Bartz, Betrieb, Wartung und Schmierung von Kompressoren, Lehrgang Nr. 19293/68.358, Technische Akademie Esslingen, Postfach 1265, D-73748 Ostfildern, Germany.
249. Fuchs DEA Schmierstoffe GmbH & Co. KG, Luftverdichteröle–Technische Information Produktentwicklung und Anwendungstechnik, Herr Tinger, Friesenheimer Str. 15, D-68169 Mannheim, Germany.
250. Schmierstoff-Tabelle für Verdichter, Kompressoren und Vakuumpumpen, 3rd edn, Technischer Dienst UNITI, Buchtstraße 10, D-2000 Hamburg 76, Germany, 1986.
251. DIN 51 506, Schmieröle VB und VC ohne Wirkstoffe und Schmieröl VDL: Einteilung und Anforderungen, September, 1985.
252. Mannesmann Demag Verdichter, Druckluft Rotationsverdichter, Firmenschrift, Johann-Sutter-Straße 6 – 8, D-79650 Schopfheim, Germany.
253. Comp Air Demag, Kolben- und Schraubenverdichter, company brochure, Argenthaler Straße 11, D-55469 Simmern/Hunsrück, Germany.
254. Boge Kompressoren, Schraubenkompressoren, company brochure, Lechtermannhof 26, D-33739 Bielefeld, Germany.
255. ISO 6743-3A, Compressors, 1987. ISO 6743-3B, Gas and refrigeration compressors, 1988.
256. R. Planck, *Handbuch der Kältetechnik,* Springer, Berlin/Göttingen/Heidelberg 1956.
257. H. Steimle, *Kältemaschinenöle,* Springer, Berlin/Göttingen/Heidelberg, 1950.
258. DIN 51 503-1, Kältemaschinenöle–Mindestanforderungen, August 1997, DIN 51 503-2, Kältemaschinenöle–gebrauchte Kältemaschinenöle, November 1998.
259. ASHRAE Standard ANST/ASHRAE 34, 1992.
260. Solkane Taschenbuch, Kälte- und Klimatechnik, 1st edn, 1997, Solvay Fluor u Derivate GmbH, Postfach 220, 30002 Hannover, Germany.
261. Fuchs Refrigerator Oils, Fuchs Petrolub AG, Mannheim, Germany, February 1999.
262. D. Klamann, *Schmierstoffe und verwandte Produkte,* Verlag Chemie, Weinheim 1982.
263. *Kälte Wärme Klima Taschenbuch 1998,* 31st ed., CF Müller Verlag, Heidelberg 1998.
264. Forschungsrat Kältetechnik e.V., Forschungsvorhaben, Erprobung von Kältemittelgemischen im Verdichterkreislauf, Frankfurt/Main, Germany, Juni 1996.
265. AFEAS–Alternative Fluorcarbons Environmental Acceptability Study, Production, Sales of Fluorcarbons through 1995, Washington D.C. 20005, USA.
266. Forschungsrat Kältetechnik e.V., “NH_3-lösliche” Kältemaschinenöle.

267. J. Reichelt, *Fahrzeugklimatizierung mit natürlichen Kältemitteln,* 1st ed., CF Müller, Heidelberg, 1996.
268. Kohlendioxid-Besonderheiten und Einsatzchancen als Kältemittel, Statusbericht des DKV Nr. 20, DKV, Stuttgart, Germany, November, 1998.
269. Turbine Lubrication, Society of Tribologists and Lubrication Engineers, 840 Dusse Highway, Parkridge, Illinois 60068-2376, 1983.
270. Arbeitskreis Schriftentumauswertung Schmierungstechnik, Dr. Ing. W. Bartz, Mühlhaldenstr. 91, 7306 Denkendorf: Dampfturbinenöle–Stand der Technik und Betriebsverhalten, E. Sommer, Deutsche Shell AG, Hamburg.
271. Maintenance Technology International Incorpored, Maintenance, *The Practical Handbook of Lubrication,* Section 17, Turbines and Turbine Oils, Cochrane, Alberta, Canada.
272. *Schmierstoffe und Steuerflüssigkeiten, VDEW Ölbuch,* 7th ed., **vol. 1,** VDEW Verlag, Frankfurt/Main 1996.
273. D. Klamann, Dampfturbinenöle, *Schmierstoffe und verwandte Produkte,* Verlag Chemie, Weinheim, 1982, Chapter 11.4.1.
274. Siemens AG, KWU-Energieerzeugung, Technische Liefervorschrift 901304: Turbinenöle (Lubricating, Control and Governor Oil), vom 20.09.96, Dr. Quick, Herr Sand
275. Lubricating Oil Recommendations with Antiwear Additives for Gas Turbines with Bearing Ambients above 500 °F (260 °C), GE Power Systems, Gas Turbine, GEK 101941A, General Electric Company, November 1999,
276. Schmier- und Steueröle für Turbinen, Spezifikation und Überwachung, ABB (Asea Brown Boveri), HTGD 90117 D, June 1994.
277. Lubricating Oil Recommendations for Gas Turbines with Bearing Ambients above 500 °F (260 °C), GE Power Systems, Gas Turbines, GEK 32568 E, General Electric Company, revised May 1999.
278. DIN 51 515, Turbinenflüssigkeiten (draft 1999), Protokolle der Arbeitskreissitzungen.
279. ISO 6743-5, Part 5: Family T (Turbines), Classification. 1988 (E).
280. ISO 8068, Petroleum products and lubricants–Petroleum lubricating oils for turbines (Categories ISO-L TSA and ISO-L TGA)–Specifications. 1988 (E).
281. E.V. Zaretsky, Tribology for Aerospace Applications, STLE Publication SP-37, USA, 1997, Library of Congress Catalog Number: 97-061830
282. J. a. Willimas, D. Tabor, The role of lubricants in machining, *Wear* **43** (1977) 275 – 292.
283. T. Mang, *Wassermischbare Kühlschmierstoffe für die Zerspanung,* Expert Verlag, Grafenau 1980.
284. J. E. Neely, R. R. Kibbe, *Modern Materials and Manufacturing Processes,* John Wiley & Sons, New York 1987.
285. H. J. Padberg: Entwicklung anwendungsorientierter Bindungssysteme für Schleifwerkzeuge, in E. Saljé, E. Westkämper (eds.): *Jahrbuch Schleifen, Honen, Läppen und Polieren,* Vulkan Verlag, Essen 1993.
286. F. Grünwald, G. Schilling, Werkstoffabtrag und Rauheit beim Planläppen metallischer Werkstoffe, *Feingerätetechnik* **28** (1979) no. 8, 345.
287. K. Weinert, *Dry Machining and Minimum Quantity Lubrication,* Springer Verlag, Berlin 1999.
288. Autorenkollektiv Kühlschmierstoff–eine ökologische Herausforderung and die Fertigungstechnik in Wettbewerbsfaktor Produktionstechnik, *Aaachener Werkzeugmaschinenkolloquium* 1993, VDI Verlag, Düsseldorf 1993.
289. T. Mang, C. Freiler, Trends bei der Kühlschmierung: neue Fluidfamilien für die spanende Fertigung, Sonderdruck aus *Spanende Fertigung*, Vulkan Verlag, Essen, Oktober 1997.
290. J. Ahrens, Ermittlung des Sprühverhaltens verschiedener MMKS-Systeme unter Verwendung biologisch abbaubarer Medien, Studienarbeit, Fachhochschule Mannheim, März 1998.
291. J. Grosch, Einfluß der Werkstoffeigenschaften unberuhigter Bleche auf Tiefziehverhalten in der Fertigung, *Mitteilung der DFBO* **6** (1968) 74 – 78.
292. H. Schenk, Prölss, *Schneidbarkeit von von Elektroblechen*, HFF-Bericht 6, Hannoversches Forschungsinstitut für Fertigungsverfahren e.V., 1980.
293. A. Losch, Prelubes for sheet metal forming, *Tribology 2000 Plus, 12th International Colloquium Tribology*, Ostfildern, January 11 – 13, 2000.
294. VDA-Prüfblatt 230–201 Prelubes, Verband der Automobilindustrie Frankfurt, Dec. 1998.

295. K. Masumoto, Advantages of 'Oil Draw' compared with 'Dry Draw' in the cold wire drawing process using continuous wire drawing machines (high and low carbon steel, including galvanized and stainless steel), *Lubric. Engineering* **34** (1978) 258 – 264.
296. M. Lee, D. G. Flom, Lubrication mechanisms in tungsten and molybdenum wire drawing. Effects of surface conditions on wetting and lubricant film formation, *Lubric. Engineering* **47** (1991) 135 – 140.
297. J. Doeland, A. Maaren, *Optimizing the drawing of tungsten wire by using some drawing theories (Die Optimierung des Ziehens von Wolframdraht mit Hilfe einiger Ziehtheorien)*, Bericht der Deutschen Gesellschaft für Metallkunde, Oberursel 1972.
298. L. Abermann, Encouraging trials using drawing cones cut out of Al_2O_3 (Versuch mit Drahtziehringen aus Al_2O_3 hoffnungsvoll), *Drahtwelt* **64** (1978) 280 – 281.
299. W. Huken, *Pretreatment of pipes, drawing methods and in this case applied devices and machines, VDEH-Reports* (*Vorbehandlung der Rohre zum Ziehen, Ziehverfahren und dafür eingesetzte Anlagen und Maschinen, VDEH-Berichtsreihe 'Herstellung von Rohren'*), Verein Deutscher Eisenhüttenleute, Verlag Stahleisen, Düsseldorf 195.
300. F. Drohmann in K. Lange (ed.): Internal high pressure forming, *Forming Technology, Handbook for Industry and Science,* **vol. 4: Special production processes, simulation, tool technology, production**, Chapter 11, Springer Verlag, Berlin 1993.
301. B. Engel, *Process strategy for internal high pressure forming* (*Verfahrensstrategie zum Innenhochdruck-Umformen*), Thesis, Technical University Darmstadt, Berichte aus Produktion und Umformtechnik, vol. 30, Shaker Verlag, Aachen 1996.
302. C. Busch, R. Schneider, Aspects of lubricant technology at hydroforming (Schmierstofftechnische Aspekte beim Innenhochdruckformen), Conference paper: *Industrial and Automotive Lubrication, 11th International Colloquium*, vol. 3: 2407.2421, Technische Akademie Esslingen 1998.
303. K. Peters, M. Bauer, *Manufacturing of steel sheet in manufacturing of thick strip, hot wide strip, and steel strip* (*Herstellung von Feinblech in Herstellung von Grobblech, Warmbreitband und Feinblech*), Verlag Stahleisen, Düsseldorf 1979.
304. P. Wehber, Circulating systems of coolants for the cold rolling of steel (Umlaufsysteme für Kühlschmierstoffe beim Kaltwalzen von Stahl), *3rd International Colloquium*, vol 1, Technical Academy Esslingen 1982.
305. D. Raabe, J. Adler, H. Oppermann, 20–rolls mill for high deformations and rolling speeds (20–rollen Walzgerüst für hohe Umformungen und Walzgeschwindigkeiten), *Stahl u. Eisen* **100** (1980) 683.
306. A. L. Moxon, Roll Conditioning – a backup system for control of pick-up when hot rolling aluminum, *Lubriation Engineering* **12** (1982) 743 – 748.
307. Aluminium Swiss, US 4 326 974.
308. H. M. Rue, *Lubricants for rolling copper and brass*, ASLE Non ferrous metals council, Boston 1971.
309. M. Moda, The effect of lubricant additives on heading socket screw bolt, *Lubriation Engineering* **8** (1978) 448.
310. G. H. Geiger, *Aufbau und Einsatz von Fließpreßölen*, VDI-Bericht 445, 1982, 41 – 45.
311. P. Zwez, Tendenzen der chemischen Oberflächenbehandlung beim Kaltmassivumformen, 6. Internationale Tagung Kaltumformen, Preprints, Düsseldorf 1980.
312. P. Hivart, J. P. Bricout, J. Oudin, J. Y. Dauphin, Seizure strenght imporvement for phosphate–stearate coatings in coald forging of steel, *Trans ASME, J. Tribol.* **114** (1992) no. 2, 248 – 252.
313. S. Adjerid, P. Hivart, J. P. Bricout, Some stressing and environmental parameters related to phosphate–stearate coating performance in cold forging of steel, *Tribol. Lett.* **2** (1996) no. 1, 55 – 70.
314. J. P. Bricout, P. Hivart, J. Oudin, Y. Ravalard, New testing procedure of zinc phosphate coatings involved in cold forging of cylindrical steel billets, *J. Mater. Process. Technol.* **24** (1990) 3 – 12.
315. Metallgesellschaft AG, AT 273 345, 1969.
316. D. Schmoeckel, Der Einfluß der Schmierung auf dei Kaltumformung von Stahl, *Schmiertechnik und Tribologie* **22** (1975) no. 1, 12 – 16.
317. H. Vetter, *Festschmierstoffe und ihre Anwendung in der Umformtechnik*, Veranstaltung an der Technischen Akademie Esslingen, Oktober 1978.
318. M. A. Zamzam, Wear resistance of agglomerated and dispersed solid lubricants in

aluminum, *Matr. Trans. JIM* **30** (1989) no. 7, 516 – 522.
319. F. Nonoyama, K. Kitamura, A. Danno, New lubricating method for warm forging of steel with boron oxide (B_2O_3), *Ann. CIRP* **42** (1993) no. 1, 353 – 356.
320. T. Udagawa, E. Kropp, T. Altan, Friction and metal flow in precision forging of aluminum alloys, *Advanced Technology and Plasticity* **1** (1990) 33 – 39.
321. H. Schoch, *Ein Beitrag zum Gesenkschmieden aushärtbarer Aluminiumlegierungen*, Fortschritts-Berichte VDI Reihe 2, Fertigungstechnik 22, VDI-Verlag Düsseldorf 1986.
322. G. Kershah, *Untersuchungen und Optimierungen von Schmiedeparametern beim Gesenkschmieden von Aluminiumwerkstoffen*, Thesis, Technical University of Darmstadt, 1987.
323. R. Seidel, H. Luig: Friction and wear processes in hot die forging, in H. Berns et al. (eds.): *New material experinces for tooling, Proc. Int. European Conf. Tooling Materials, Interlaken, Sept. 1992,* Mat Search Andelfingen 1992, 467 – 480.
324. D. Dowson, *History of Tribology,* Professional Engineering Publishing, London 1998, pp. 28 – 33.
325. W. F. Parish, Lubricants, *Encyclopedia Britannica,* 14th ed, **vol. 14,** 1929, p. 453.
326. R. König (ed.): *C. Plini Secundi Naturalis Historiae Liber XXVIII* Artemis, München 1988, §141.
327. GB 6945, 07.12. 1835.
328. J. Lewkowitsch, *Chemical Technology and Analysis of Oils, Fats and Waxes,* Vieweg, Braunschweig 1905, 2 Vols.
329. GB 12571, 16 April 1849.
330. E. N. Klemgard, *Lubricating Greases: Their Manufacture and Use,* Reinhold, New York 1937.
331. C. J. Boner, *Manufacture and Application of Lubricating Greases,* Reinhold, New York 1954.
332. R. H. Focke, Multi-Purpose Greases: A New Dimension, *NLGI Spokesman* **57** (1994) 430 – 434.
333. ASTM D 217, led to former DIN 51804 and to DIN ISO 2137 Erdölprodukte, Schmierfette und Vaseline, Bestimmung der Konuspenetration, Beuth, Berlin, 1985.
334. (Norm-Entwurf) ISO/DIS 6743-9, Schmierstoffe, Industrieöle und verwandte Erzeugnisse (Klasse L) Klassifikation Teil 9, Familie X (Fette), 1998.
335. DIN 51825, Schmierstoffe, Schmierfette K, Einteilung und Anforderungen, 1990.
336. ASTM D 4950, ASTM Annual Book of Standards, ASTM, West Cinshohocken, 1989.
337. J. Harris, History and Current Status of NLGI Reference Greases, *NLGI Spokesman* **59** (1996) no. 12, 18 – 22.
338. M. Ehrlich (ed.): *Lubricating Grease Guide,* 4th ed, NLGI, Kansas City 1996, 3.28 – 3.36.
339. W. H. Dresel, Einige Eigenschaften neuer Verdickersysteme, 11th International Colloquium Esslingen, 23 September 1998, Vol III, 2261 – 2264; W. H. Dresel, R.-P. Heckler, Some Aspects of Tomorrow's Greases, *NLGI Spokesman* **58** (1994) 17 – 24.
340. J. Bhatia, R. L. Dhawan, R. L. Sovani, Tailor-made Fatty Acids Based on Castor Oil for Processing of High Performance Greases, *NLGI Spokesman* **56** (1992) 265 – 272.
341. P. W. Hurd, The Chemistry of Castor Oil and its Derivatives, *NLGI Spokesman* **60** (1996) 14 – 23.
342. Beilstein E IV 2 1206.
343. J. A. Lurz, Grease Production Survey Report for the Calendar Years 1998, 1997, 1996, and 1995, NLGI, Kansas City, 1999.
344. R. Kühl, Ölabgabeverhalten bei tiefen und hohen Temperaturen–Einfluß auf den Temperatureinsatzbereich eines Schmierfettes in Wälzlagern, *GfT Tribologie-Fachtagung,* 1998, Conference Compendium 26/1 – 6.
345. H. F. George, C. R. Scharf, The Enhancement of grease structure through the use of functionalized polymer systems, *NLGI Spokesman* **59** (1996) no 11; H. F. George, K. J. Hole, C. R. Scharf, The Enhancement of grease structure through the use of functionalized polymer systems, *Eurogrease* (1997) Jan./Feb., 17 – 35.
346. D. Holweger, M. Volkholz, Polymer Esters–A Novel Component in High-Performance Greases, *ELGI* Preprint, 1997.
347. C. Lu, H. Shen, P. Jiang, R. Jiang, A study on anti-wear and anti-friction performance of lithium – calcium base greases, *NLGI Spokesman* **57** (1993) 65 – 69.
348. N. G. Janakiefski, The unique chemistry of azelaic acid, *NLGI Spokesman* **61** (1997) no. 5, 14 – 24.
349. P. W. Hurd, The chemistry of castor oil and its derivatives, *NLGI Spokesman* **60** (1996) 14 – 23.
350. Mobil Oil Co, US 3 806 650, 1974.

351. US 2 417 428, 1947 (L. W. McLennan).
352. Y. L. Ischuk, A. D. Stakhursky, Composition of lithium complex soap and its role in the formation of grease properties, *NLGI Spokesman* **59** (1995) 21 – 24.
353. US 2 197 263, 1940 (E. S. Carmichael, G. M. Hain).
354. W. H. Dresel, Moderne Schmierfette mit verlängerter Lebensdauer, *Tribologie und Schmierungstechnik* **36** (1989) 305 – 311.
355. US 4 560 489, 1985 (W. Blockhuis, R. Muir).
356. R. J. Muir, High Performance Calcium Sulfonate Complex Lubricating Grease, *NLGI Spokesman* **52** (1988) 140 – 146.
357. US 5 126 062, 1992 (J. F. Barnes).
358. J. Huang, Lubrication and seal, *Runhua Yu Mifeng* **1** 1996 37 – 44.
359. US 2 599 553, 1952 (B. W. Hotten).
360. US 3 587 351, 1993 (A. K. Bhatnagar et al.).
361. A. K. Bhatnagar et al., A new-generation high performance titanium complex grease, *NLGI Spokesman* **58** (1994) 25 – 30; A. K. Bhatnagar et al., Titanium complex grease: a product of high potential, *NLGI Spokesman* **59** (1995) no. 9, 12 – 18; A. K. Bhatnagar et al., Enhancing further performance properties of titanium complex grease, *NLGI Spokesman* **62** (1998) no. 6, 20 – 27; A. K. Bhatnagar et al., Ti-complex grease: A versatile grease for industry, in J. Bhatia (ed.): *Advances in Industrial Tribology* McGraw – Hill, New Delhi 1998, pp. 433 – 442.
362. US 2 820 012, 1954 (B.W. Hotten).
363. C. F. Carter, J. L. Dreher, B. W. Hotten, New synthetic thickener for multipurpose lubricating grease, *NLGI Spokesman* (1957) Feb., 10 – 15.
364. Q. S. Shen et al., The Research of N-Alkyl Terephthalamate Aluminum Complex EP Grease, *7. International Colloquium Esslingen*, 1990, vol. I, 4.7.1 – 4.7.10.
365. C. G. Brannen, L. C. Brimstrum, E. A. Swaken, Substituted ureas as grease thickeners, *NLGI Spokesman* **18** (1954) 8 – 13.
366. H. Li, L. Xie, A study of greases based on polyureas, *JSL* **8** (1991) 39 – 50.
367. T. Endo, Recent developments in diurea greases, *NLGI Spokesman* **57** (1993) 532 – 541; T. Endo, Current Trends in Diurea Greases in Japan, *Eurogrease* (1997) Nov./Dec., 25 – 40;
368. US 3 846 314, 1974 (C. F. Carter, J. L. Dreher).
369. US 3 920 571, 1975 (R. E. Crocker).
370. P. F. Vartanian, Polyurea complex greases–ten years later, *NLGI Spokesman* **49** (1985) June, 98 – 102.
371. J. Root, A comparative study of polyurea and lithium complex grease thickeners, *NLGI Spokesman* **58** (1995) 364 – 366.
372. W. Schwickerath, Perfluorether–Basisflüssigkeit für Hochtemperaturschmierstoffe, *Tribologie und Schmierungstechnik* **34** (1987) 22 – 38.
373. G. Caporiccio, L. Flabbi, G. Marchionni, G. T. Viola, The properties and applications of perfluoropolyether lubricants, *JSL* **6** (1989) 33 – 149.
374. F. Wunsch, W. Schwickerath, Perfluoralkylether, *Tribologie und Schmierungstechnik* **38** (1991) 130 – 136.
375. A. Chiavoni, M. D. Kieke, E. D. Magauran, W. W. Reichert, Effective utilization of organoclay dispersants, *NLGI Spokesman* **51** (1987) March, 453 – 460.
376. J. W. Harris, G. P. Newsoroff, The manufacture and properties of microgel thickened greases, *NLGI Spokesman* **56** (1992) 191 – 197.
377. A. F. Hollemann, E. Wiberg, *Lehrbuch der Anorganischen Chemie,* 91.–100. ed., de Gruyter, Berlin 1985, pp. 774 – 776.
378. D. G. Miller, W. F. Moll, Improving grease performance with surface-modified fumed silica, *NLGI Spokesman* **48** (1984) Feb., 390 – 394.
379. G. Schmidt, Gel-, bentonit- und polyharnstoff-schmierfette, *Tribologie und Schmierungstechnik* **31** (1984) 309 – 314.
380. Y. L. Ischuk, et al, Greases in which the dispersed phase is expanded graphite, *NLGI Spokesman* **61** (1996) no. 6, 22 – 30.
381. G. Meyer, D. Wöhrle, Phthalocyanine–ein System ungewöhnlicher Struktur und Eigenschaften, Teil 1, *Kontakte (Darmstadt)* **3** (1985) 38 – 47.
382. R. Eidenschink, A. M. Häger, Reduction of friction through presmectic lubricants, *Mol. Cryst. Liq. Cryst.* **304** (1997) 513 – 517.
383. R. Eidenschink, Liquid crystals in variable friction devices, *Angew. Chem.* **100** (1988) 1639 – 1940.
384. S. E. Friberg, S. Gunsel, A. J. Lockwood, A. J. Ward, Lyotropic liquid crystals, in G. Biresaw (ed.): *Tribology and the Liquid-Crystalline State,* **441**, ACS Symposium Series 1990 pp. 101 – 111.
385. M. Chiu, S. E. Friberg, H. S. Lee, W. O. Winer, S. H. Winoto, Film thickness and frictional behavior of some liquid crystals in concentrated point contacts, in G. Biresaw

(ed.): *Tribology and the Liquid-Crystalline State,* **441**, ACS Symposium Series 1990 pp. 113 – 125.
386. US 2 417 850, 1942 (W. M. Winslow); W. M. Winslow, *J. Appl. Phys.* **20** (1949) 1137 – 1140.
387. R. Tao, G. D. Roy (eds.): *Electrorheological Fluids,* World Scientific, Singapore 1995.
388. K. Havelka, F. Filisko (eds.): *Progress in Electrorheology,* Plenum, New York 1995.
389. K. Holmberg, M. Säynätjoki, Magnetic fluids in sealing and lubrication–a state of the art review, *JSL* **10** (1993) no. 2, 119 – 132.
390. J. H. Hildebrand, R. L. Scott, *Solubility of Non-Electrolytes,* 3rd ed, Reinhold, New York 1950.
391. A. Bondi, *Physical Chemistry of Lubricating Oils,* Reinhold, New York 1951, pp. 249 – 251.
392. R. N. Bolster, R. C. Little, Dependence of lithium grease dropping points on the solubility parameter of the oil component, *Ind. Eng. Chem. Prod. Res. Dev.* **5** (1966) 198 – 202.
393. A. Beerbower, Environmental capabilities of liquid lubricants, in Solid and Liquid Lubricants for Extreme Environments, ASLE Special Publication SP-15, 1984, pp. 58 – 69.
394. S. Beret, Impact of base oil changes on grease performance, *NLGI Spokesman* **57** (1993) 192 – 198.
395. W. J. Bartz, Comparison of synthetic fluids, inR. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York 1993, pp. 319 – 336.
396. R. L. Shubkin, Polyalphaolefins, in R. L. Shubkin (ed.): *Synthetic Lubricants and High-Performance Functional Fluids,* Marcel Decker, New York 1993, pp. 1 – 40.
397. W. Noll, *Chemie und Technologie der Silicone,* 2nd ed., Verlag Chemie, Weinheim 1968, pp. 368 – 370.
398. J. E. Barnes, J. H. Wright, Silicone greases and compounds: their components, properties and applications, *NLGI Spokesman* **53** (1989) 103 – 111.
399. G. R. Schultze, Wesen und Aufbau der Schmierfette, in C. Zerbe (ed.): *Mineralöle und verwandte Produkte,* Springer, Berlin 1952, pp. 974 – 980.
400. C. J. Boner, *Manufacture and Application of Lubricating Greases,* Reinhold, New York 1954, pp. 11 – 74.
401. G. Spengler, F. Wunsch, *Schmierung und Lagerung in der Feinwerktechnik,* VDI Verlag, Düsseldorf 1970, p. 51.
402. C. J. Boner, *Modern Lubricating Greases,* Scientific Publications, Broseley (Foreword 1976), 3.1 – 3.26.
403. D. Danow, N. D. Denkov, P. A. Kralchevsky, Chemical physics of colloid systems and interfaces, in S. Birdi (ed.): *Handbook of Surface and Colloid Chemistry,* CRC Press, Boca Raton 1997, pp. 333 – 494.
404. H.-D. Dörfler, *Grenzflächen- und Kolloidchemie,* VCH Verlagsgesellschaft, Weinheim 1994, pp. 238 – 245.
405. D. Demus, J. W. Goodby, G. W. Gray, H. W. Spiess, V. Vill (eds.): *Handbook of Liquid Crystals,* **4 vols.**, Wiley – VCH, Weinheim 1998.
406. Lu Chenqiao, Lu Peigang, Jiang Ronghua, Zhu Tingbin, A study of the effect of glycerol on extreme pressure properties of lithium base grease, *NLGI Spokesman* **55** (1991) 364 – 370.
407. L. Gengxin, L. Qinglian, B. Shaojun, L. Z. Zhong, Influence of stearic acid on lubricating greases containing zinc dialkyl dithiophosphate, *NLGI Spokesman* **59** (1995) no. 4, 15 – 18.
408. A. Gerling, H. Langer, U. Tiemann, Luftgehalt und rheologische Eigenschaften von Schmierfetten, *Tribologie und Schmierungstechnik* **36** (1989)312 – 315.
409. A. Feßenbecker, I. Roehrs, A new additive for the hydrolytic and oxidative stability of ester based lubricants and greases, *NLGI Spokesman* **61** (1997) no. 3, 10 – 17.
410. I. Röhrs, Schwermetallfreie Additive für Schmierfette, *8th International Colloquium, Esslingen*, 1992, vol. I 7.4, pp. 1 – 15; I. Röhrs, R. Whitehead, Heavy metal-free additives for lubricating grease, *NLGI Spokesman* **57** (1993) 335 – 341.
411. Y.-H. Luo, J.-C. Zhang, B.-Y. Zhong, The mechanism of copper-corrosion inhibition by thiadiazole derivatives, *Lubrication Engineering* **51** (1995) 293 – 296.
412. A. K. Bhatnagar, et al., Development of N-, P- and S-containing multifunctional additives for lubricants, *Lubrication Engineering* **53** (1997) 21 – 27.
413. O. Rohr, Bismuth: a new metallic but non-toxic replacement for lead as EP-additive in greases, *NLGI Spokesman* **57** (1993) 50 – 57; O. Rohr, Wismut: Metallischer, nichttoxischer Ersatz für Blei als Hochdruck-Additiv in Schmierfetten, *Tribologie und Schmierungstechnik* **41** (1994) 336 – 342.

414. A. K. Bhatnagar, et al., Synthetic metallic dialkyldithiocarbamates as antiwear and extreme-pressure additives for lubricating oils: role of metal on their effectiveness, *Lubrication Engineering* **51** (1995) 298 – 303.
415. W. Krätschmer, L. D. Lamb, K. Fostiropoulous, D. R. Huffmann, Solid C_{60}: a new form of carbon, *Nature* **347** (1990) 354 – 358.
416. W. H. Dresel, Grundlegende Aspekte zukunftsorientierter Schmierfette, *Tribologie und Schmierungstechnik* **40** (1993) 176 – 182; W. H. Dresel, R.-P. Heckler, Some aspects of tomorrow's greases, *NLGI Spokesman* **58** (1994) 17 – 24.
417. B. Bhushan, B. K. Gupta, Fullerene particles as an additive to liquid lubricants and greases for low friction and wear, *Lubrication Engineering* **50** (1994) 524 – 528.
418. T. J. Risdon, EP Additive Response in Greases containing MoS_2, *NLGI Spokesman* **63** (1998) no. 8, 10 – 19.
419. D. K. Landry, T. J. Risdon, Lubrication properties of molybdenum disulfide and ptfe in three common grease types, *NLGI Spokesman* **57** (1993) 523 – 531.
420. UK 2 28 771, 1993 (G. Fish).
421. J. Gänsheimer, Schmierstoffe für oszillierende Bewegungen–ein neuer Fortschritt in der Tribotechnik, *Konstruktion* **42** (1990) 391 – 395.
422. C. H. Bovington, Friction, wear and the role of additives in their control, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* 2nd ed., Blackie Academic and Professional, London 1997, pp. 332 – 346.
423. J. Crawford, A. Psaila, S. T. Orszulik, Miscellaneous additives and vegetable oils, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* 2nd ed., Blackie Academic and Professional, London 1997, pp. 181 – 187.
424. D. Christakudis, Friction Modifier und ihre Prüfung, in W. J. Bartz (ed.): *Additive für Schmierstoffe,* expert Renningen – Malmsheim, 1994, pp. 134 – 162.
425. W. H. Dresel, Moderne Schmierfette mit verlängerter Lebensdauer, *Tribologie und Schmierungstechnik* **36** (1989) 305 – 311.
426. R. L. Burkhalter, E. A. Brocco, A brief on grease kettles, *NLGI Spokesman* **62** (1998) no. 2, 26 – 32.
427. L. Floquet-Muhr, N. Midoux, G. Wild, Some guidelines for improving the dehydration process during grease manufacture, *Lubrication Engineering* **52** (1996) 559 – 562.
428. W. Noll, *Chemie und Technologie der Silicone,* 2nd ed., Verlag Chemie, Weinheim, 1968, pp. 368 – 370.
429. W. B. Green, A. C. Witte, Texaco's continuous grease manufacturing process, *NLGI Spokesman* **32** (1969) 133 – 136.
430. W. W. Bailey, Manufacture and properties of polyurea grease, *NLGI Spokesman* **49** (1985) Sept., 261 – 265.
431. D. E. Siptak, J. Williams, The use of organoclays in clay-based grease, *Eurogrease* (1993) Oct., 20 – 32.
432. M. Ficko, M. Schateva, The characteristics and applications of synthetic lubricating greases, *JSL* **3** (1986) 83 – 92.
433. G. Wassermann, From Heraklit to W S Blair, *Rheology* **91** (1991) 32 – 38.
434. G. Gow, Lubricating grease, in R. M. Mortier, S. T. Orszulik (eds.): *Chemistry and Technology of Lubricants,* 2nd ed., Blackie Academic and Professional, London 1997, pp. 307 – 319.
435. M. Ehrlich (ed.): *Lubricating Grease Guide,* 4th ed., NLGI, Kansas City 1996, pp. 3.06 – 3.07.
436. G. Froishteter, K. K. Triliski, Y. L. Ishchuk, P. M. Stupak, in G. V. Vinogradov (ed.): *Rheological and Thermophysical Properties of Grease,* 4th ed., Gordon & Breach, London 1989.
437. B. J. Hardy, C. J. Swartz, A rheological model for grease, *NLGI Spokesman* **54** (1990) 324 – 327.
438. J. Kempe, A. Müller, Schmierfette: Können rheologische Prüfungen das Verständnis des Gebrauchsverhaltens noch verbessern? *Tribologie und Schmierungstechnik* **45** (1998) 28 – 32..
439. P. Whittingstall, Controlled stress rheometry as a tool to measure grease structure and yield at various temperatures, *NLGI Spokesman* **61** (1997) no. 9, 12 – 23; R. Shah, P. Whittingstall, Yield stress studies on greases, *NLGI Spokesman* **62** (1998) no. 3, 8 – 23.
440. M. M. Britton, P. T. Callaghan, Nuclear Magnetic Resonance (NMR) Velocity Imaging of Lubricating Grease, in C. Balan (ed.): *The Rheology of Lubricating Grease,* Elgi Publications, Amsterdam 2000 pp. 67 – 78
441. G. Blettner, P. Vergne, Lubricating greases correlation between composition and rheology, *Eurogrease* (1998)Nov./Dec., 18 – 29.

442. P. A. Bessette, The influence of thickener type on the apparent viscosity, *NLGI Spokesman* **63** (1999)no. 10, 28 – 33.
443. R. Bradbury, A. Mistry, An evaluation of the relationship between the low temperature torque, the apparent viscosity and the low temperature cone penetration of lubricating greases ELGI, *Eurogrease* (1998)May/June, 22 – 32.
444. T. Rosemann, Correlation between bearing test rig performance and rheology based on a study of greases consisting of perfluorinated polyether fluids thickened with PTFE, *Eurogrease* (1998)Jan./Feb., 14 – 26.
445. W. Holweger, U. Jauch, B. Nägele, M. Wätzig, Stabilität und chaotisches Verhalten bei Schmierfetten, *Tribologie und Schmierungstechnik* **37** (1990) 71 – 75.
446. W. Holweger, N. Kleffner, D. Schach, Schmierfette: Struktur und Dynamik, *Tribologie und Schmierungstechnik* **44** (1997) 34 – 36; W. Holweger, N. Kleffner, D. Schach, Schmierfette bei geringem Schergefälle, *Tribologie und Schmierungstechnik* **44** (1997) 84 – 86.
447. E. Kuhn, Energy investigations of the rheological wear of lubricating greases, *Rheology* **92** (1992) 252 – 257.
448. E. Kuhn, Investigation of stressed greases by use of a rheometer, *Eurogrease* (1998)Nov./Dec., 9 – 11.
449. G. Schmidt, Chemie und Herstellung der Metallseifenfette am Beispiel der Metallseifenfette, in W. J. Bartz (ed.): *Schmierfette,* expert, Renningen-Malmsheim 2000, p. 55 – 63.
450. G. Jäger (ed.): *Schmierstoffe und ihre Prüfung im Labor,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1984.
451. ASTM (Annual Book of ASTM Standards), Vols 05.01 – 05.03, ASTM, West Cinshohocken, 1998.
452. DIN Taschenbuch 192, Schmierstoffe, Eigenschaften und Anforderungen, 4th edn, Beuth, Berlin, 1995; DIN Taschenbuch 203, Schmierstoffe, Prüfverfahren 1, 2nd edn, Beuth, Berlin, 1995; DIN Taschenbuch 248, Schmierstoffe, Prüfverfahren 2, Beuth, Berlin, 1996.
453. S. H. Kägler (ed.): *Neue Mineralölanalyse,* 2nd ed., **vol. 1: Spektroskopie,** Hüthig, Heidelberg 1987.
454. P. D. Robison, S. G. Salmon, J. R. Sieber, M. C. Williams, Elemental analysis of greases, *NLGI Spokesman* **56** (1993) 157 – 160.
455. M. L. Kieke, Microwave assisted digestion of zinc, phosphorus and molybdenum in analyses of lubricating greases, *NLGI Spokesman* **62** (1998) May, 29 – 35.
456. E. C. Milberger, Infrared studies of bentone grease – water systems, *NLGI Spokesman* **21** (1957) 11 – 15.
457. W. H. Bauer, D. B. Cox, S. E. Wiberley, Infrared studies of Greases, *NLGI Spokesman* **24** (1960) 328 – 333.
458. S. H. Kägler (ed.): *Neue Mineralölanalyse,* 2nd ed., **vol. 1: Spektroskopie,** Hüthig, Heidelberg 1987, pp. 589 – 590.
459. A. G. Izcue, Infrared spectroscopy in the development and manufacture of lubricating greases, *NLGI Spokesman* **52** (1988)Aug., 165 – 231.
460. M. Hermansson, M. Jansson, E. Johansson, Self diffusion of base oils in lubricating greases by NMR, *JSL* **13** (1996) 279 – 288.
461. M. D. Foster, M. R. Funk, L. Irwin, N. Riszo, Extracting, imaging and quantifying soap fibers in grease, *NLGI Spokesman* **60** (1996) 24 – 25.
462. C. Araki, H. Kanzani, T. Taguchi, A study on the termal degradation of lubricating greases, *NLGI Spokesman* **59** (1995) no. 8, 15 – 23.
463. H. Kopsch, E. Wedepohl, Zur Prüfung von Schmierstoffen mit thermoanalytischen Methoden, *Tribologie und Schmierungstechnik* **30** (1983) 132 – 138.
464. M. J. Pohlen, DSC – A valuable tool for the grease laboratory, *NLGI Spokesman* **62** (1998) no. 4, 11 – 16.
465. V. J. Gatto, M. A. Grina, Effects of base oil type, oxidation test conditions and phenolic antioxidant structure on the detection and magnitude of hindered phenole/diphenylamine synergism, *Lubrication Engineering* **55** (1999) 11 – 20.
466. T. A. Norris, Chromatography I and II, Lubrication 1979, 65, 1 – 12, 13 – 24.
467. W. Holweger, U. Jauch, Hochdruckflüssigkeitschromatographie, *Tribologie und Schmierungstechnik* **34** (1987) 352 – 355.
468. D. Loderer, Lifetime lubrication of rolling bearings under extreme conditions, *NLGI Spokesman* **60** (1996) no. 8, 8 – 13.
469. E. Kleinlein, Grease test system for improved life of ball and roller bearings, *Lubrication Engineering* **48** (1992) 916 – 922.
470. A. Kemble, Evaluation of industrial bearing grease performance, *Eurogrease* (1998) July/Aug., 10 – 25.

471. DIN 51 806, Test Using SKF-R2F Roller Bearing Grease Test Machine, 1988.
472. DDIN 51 821, Test Using the Roller Bearing Grease Testing Apparatus FE 9, 1990.
473. E. Beghini, Cleanliness and Its Performance to Bearing Performance, ELGI Preprint, 1992.
474. H. Komiya, Effect of contaminant in lubricant on noise of ball bearings, *NLGI Spokesman* **56** (1992) 94 – 100.
475. ASTM D 1831 Annual Book of ASTM Standards, Vol. 05.01, 1996.
476. ASTM D 1263 Annual Book of ASTM Standards, Vol, 05.01, 1996.
477. NF T 60-171 AFNOR, Vol. 7 2, Paris, 1988.
478. DIN 51 805, Testing of lubricants, determination of flow pressure of lubricating greases, Kesternich method, 1974.
479. V. Stepina, V. Vesely, Lubricants and Special Fluids, *Tribology Series* **23** (1992) 672 – 696.
480. DIN 51 825, Type K Lubricating Greases, 1990.
481. D. M. Jahn, Grease Applications – Automotive NLGI Grease Training Course, 1998.
482. W. H. Dresel, H. Glüsing, H. Jünemann, Hidden Uses of Greases in Cars, NLGI Preprint, 1987.
483. G. Fish, P. Ratcliffe, CV Joint Lubrication from the 50's to the Future, Institution of Mechanical Engineers, 1994.
484. J. Richter, Optimizing efficiency indices of constant-velocity joints with low friction lubricants, *Eurogrease* (1999) March/April, 34 – 39.
485. ASTM D4950-95, Standard Classification and Specification for Automotive Service Greases, 1995.
486. C. Mishra, O. Naruyan, D. Singh, Selection of Lubricants for Steel Plant Application: Practics in Tata Steel, *Greasetech India, II* **1** (1999) 16 – 26.
487. W. A. Christianson, Traction motor gear lubrication requirements and development, *Lubrication Engineering* **35** (1978) 303 – 308.
488. W.-D. Abel, Einige Aspekte zur Spurkranzschmierung bei der DB AG, *10th International Colloquium Esslingen*, 1996, Vol. I 4.4, 187 – 212.
489. R. L. Goyan, R. E. Melley, P. A. Wissner, W. C. Ong, Biodegradable lubricants, *Lubrication Engineering* **54** (1998) 10 – 17.
490. R. P. Heckler, Application Experience with Biodegradable Greases, *Greasetech India, II* **2** (1999) 5 – 9.
491. B. Biehl, Instandhaltung von Zahnkranzantrieben unter tribologischen Gesichtspunkten, *Zement-Kalk-Gips* **43** (1990) no. 11, 534 – 538.
492. L. Husfeld, Schmierstoffe, HACCP, Technik und Kosteneinsparung – ein Widerspruch ? *Food Technologie Magazin* (1998) Dec., 24 – 34.
493. K. Nagdi, Einwirkung von Mineralölen auf Gummi-Werkstoffe für Dichtungen, *Maschinenmarkt* **80** (1974) 118 – 121; K. Nagdi, Wechselbeziehungen zwischen synthetischen Flüssigkeiten und Dichtungsmaterialien, Arbeitskreis Schrifttumsauswertung Schmierungstechnik 1989, 17-89.
494. ASTM D 4289-91, Standard Test Method for Compatibility of Lubricating Grease with Elastomers, Annual Book of ASTM Standards, Vol. 05.03, 1991.
495. DIN 53538-1, Standard-Referenz-Elastomere, Acrylnitril-Butadien-Vulkanizat (NBR) peroxidvernetzt, zur Charakterizierung flüssiger Betriebsmittel hinsichtlich ihres Verhaltens gegen NBR, 1992; ISO 6072, Fluidtechnik, Hydraulik, Verträglichkeit zwischen Elastomerwerkstoffen und Fluiden, 1986.
496. S. Beret, T. J. Boersig, P. K. Wong, Elastomer compatibility – base oil and additive effects, *NLGI Spokesman* **60** (1996) no. 5, 21 – 25.
497. G. Dornhöfer, Die Wechselwirkung Schmierstoff-Kunststoff dargestellt an Hand von Beispielen aus der KFZ-Ausrüstung, in Gleitreibung und Gleitverschleiß bei Kunststoffen, Esslingen, 1997, Conference Compendium.
498. G. E. Barker, et al., The Comprehensive Laboratory Testing of Instrument Lubricants, *ASTM Bulletin* **3** (1946) 26 – 35.
499. M. Tillwich, W. Stehr, W. H. Dresel, Prüfgerät für Langzeitversuche zur Bestimmung tribologischer Werkstoff-Schmierstoff-Kenndaten, in Kunststofflagern, sowie Überprüfung ausgewählter Reibsysteme auf tribologische Reaktionen, *Tribologie,* **vol. 5**, Springer, Berlin 1982, pp. 125 – 185.
500. W. Stehr, Precision bearings and their tribological simulation, *JSL* **7** (1990) 297 – 309.
501. W. Kaufmann, Bestimmung der Chemikalienbeständigkeit von Kunststoffen unter mechanischer Spannung, *Kunststoffe* **65** (1975) 155 – 157.

502. W. Noll, *Chemie und Technologie der Silicone,* 2nd ed., Verlag Chemie, Weinheim 1968, pp. 408 – 410.
503. H. Wieczorek (ed.): *Kunststoffe im Lebensmittelverkehr,* 44th ed., Carl Heymanns, Köln 1993.
504. A. Hubmann, Rapsöl – ein alternatives Basisöl für Schmierstoffe, Mineralöltechnik, 1989, 34, 9; A. Hubmann, Chemie pflanzlicher Öle, *9th International Colloquium Esslingen*, 1994, vol I 2.1, 1 – 15.
505. W. H. Dresel, Schmierfette auf pflanzlicher Basis, in *Biologisch abbaubare Schmierstoffe und Arbeitsflüssigkeiten*, Esslingen, 1990, Conference compendium 6; W. H. Dresel, Schmierfette auf pflanzlicher Basis, inW. J. Bartz (ed.): *Biologisch schnell abbaubare Schmierstoffe und Arbeitsflüssigkeiten,* expert, Ehningen 1993, pp. 239 – 258; W. H. Dresel, Biologically degradable Lubricating Greases based on Industrial Crops, *Industrial Crops and Products* **2** (1994) 281 – 288.
506. H.-G. Schmidt, Modified Natural Base Oils – Their Properties and Use for Lubricating Greases, *ELGI Preprint*, 1993; H.-G. Schmidt, Komplexester aus pflanzlichen Ölen, *9th International Colloquium Esslingen*, 1994, vol I 2.2, 1 – 29.
507. E. M. Stempfel, L. A. Schmid, Biodegradable lubricating greases, development, application, trends, *NLGI Spokesman* **55** (1991) 313 – 321; E. M. Stempfel, L. A. Schmid, Biologisch abbaubare Schmierstoffe – Eine Standortbestimmung, speziell am Beispiel von Schmierfetten, *8th International Colloquium Esslingen*, 1992, vol I 12.1, 1 – 20; E. M. Stempfel, Practical experience with highly biodegradable lubricants, especially hydraulic oils and lubricating greases, *NLGI Spokesman* **62** (1998) 8 – 23.
508. D. Loderer, E. M. Stempfel, Lifetime lubrication of rolling bearings with rapidly biodegradable lubricating greases, *NLGI Spokesman* **61** (1997) no. 2, 16 – 20.
509. P. Legiza, K. Nahal, Some experience with biodegradable lubricants, *Synthetic Lubrication* **13** (1997) 347 – 360.
510. R. Hiss, A biodegradable vegetable polyurea grease, *NLGI Spokesman* **57** (1993) 188 – 191.
511. C. Kyriakopoulos, Aluminum complex greases formulated with biodegradable base oils, *NLGI Spokesman* **59** (1995) no. 2, 10 – 23.
512. A. K. Bhatnagar, et al., Titanium complex grease – some new findings, *NLGI Spokesman* **59** (1995) no. 2, 10 – 18; A. K. Bhatnagar, et al., Ecofriendly titanium complex grease, *NLGI Spokesman* **61** (1997) no. 8, 22 – 28.
513. L. Ulm, et al., Biodegradability of water-resistant greases, in *Proceedings of the International Conference on Tribology*, Gozd Martuljek, 1998, 104 – 114.
514. T. Mang, Environmentally harmless lubricants, *NLGI Spokesman* **57** (1993) no. 6, 9 – 15.
515. R.-P. Heckler, Application Experience with Biodegradable Greases, *Greasetech India, II* **2** (1999) 5 – 6.
516. E. Kleinlein, W. H. Dresel, N. Geheeb, R. Kühl, D. Sohn, Wälzlagerschmierung, *GfT-Arbeitsblatt* 3, Moers, 1993, 35; E. Kleinlein, W. H. Dresel, N. Geheeb, R. Kühl, D. Sohn, Rolling Bearing Lubrication, *GfT Work Sheet* 3, Moers, 1994, 35.
517. U. J. Möller, Entsorgungsmöglichkeiten für biologisch schnell abbaubare Schmierstoffe, *9th International Colloquium Esslingen*, 1994, Vol II 13.1, 1 – 9.
518. W. J. Bartz, Lubricants and the Environment, *Eurogrease* (1996) May/June, 15 – 52.
519. W. H. Dresel, Grundlegende Aspekte zukunftsorientierter Schmierfette, *Tribologie und Schmierungstechnik* **40** (1993) 176 – 182; W. H. Dresel, R.-P. Heckler, Some aspects of tomorrow's greases, *NLGI Spokesman* **58** (1994) 17 – 24.
520. G. Dornhöfer, Lebensdauerschmierung von Kfz-Aggregaten unter Berücksichtigung der Umweltschonung, *Tribologie und Schmierungstechnik* **41** (1994) 86 – 90.
521. F. P. Bowden, D. Tabor, *The Friction and Lubrication of Solids,* Oxford University Press, London 1954; F. P. Bowden, D. Tabor, *Reibung und Schmierung fester Körper,* Springer, Berlin 1959.
522. A. Bondi, *Physical Chemistry of Lubricating Oils,* Reinhold, New York 1951, pp. 19 – 100.
523. E. Kuhn, Tribologische Beurteilung des zeitabhängigen Verhaltens von Schmierfetten, *Tribologie und Schmierungstechnik* **43** (1996) 245 – 247; E. Kuhn, Lubricating grease as an active element of the tribological system, *Eurogrease* (1998) March/April, 13 – 19.
524. P. M. Cann, A. A. Lubrecht, An Analysis of the Mechanisms of Grease Lubrication in Rolling Element Bearings, *11th International Colloquium Esslingen*, 1998, Vol I 6.1, 561 – 569.
525. S. Hurley, P. M. Cann, Grease composition and film thickness in rolling contacts, *NLGI Spokesman* **63** (1999)no. 4, 12 – 22; P. M.

Cann, S. Hurley, Grease composition and film thickness in rolling contacts, *Eurogrease* (1999) May/June, 7 – 25.
526. P. M. Cann, Grease Lubricant Films in Rolling Contacts, *Eurogrease* (1997) Nov./Dec., 6 – 20.
527. S. Hosoya, Deterioration of lithium soap greases and functional life in ball bearings, *NLGI Spokesman* **53** (1989) 246 – 251.
528. P. M. Cann, The influence of temperature on the lubricating behavior of a lithium hydroxystearate grease, *Eurogrease* (1995) Jan./Feb., 25 – 32.
529. AFNOR Produits pétroliers et lubrifiants, Tome 1 and 2, L'Imprimerie Chirat, 42540 St-Just-La-Pendue 1992.
530. ASTM Volumes 5.01, 5.02, 5.03, American Society for Testing and Materials, 100 Barr Harbor Drive, West Conshohocken, PA 19428.
531. DIN Handbuch 2-1, 2-2, Beuth Verlag GmbH, Berlin, 1997.
532. *IP Standard Methods for Analysis and Testing of Petroleum and Related Products,* John Wiley & Sons, Chichester 1999.
533. DIN Handbuch 2-3, Beuth Verlag GmbH, Berlin, 1997.
534. G. Brandes, *Brennst. Chem.* **37** (1956) 263 – 267.
535. DIN 50322, Kategorien der Verschleißprüfung, Beuth Verlag, Berlin.
536. H. Jünemann, *Tribometrie – Prüf- und Meßtechnik,* Expert-Verlag, Böblingen 1997.
537. DIN-Taschenbuch 192, Schmierstoffe, Eigenschaften und Anforderungen, 4th edn, Beuth Verlag, Berlin, 1995.
538. M. Fuchs, The World Lubricants Market, *8th International Conference on Industrial Tribology*, Calcutta, 1997.
539. M. Fuchs, The Global and Regional Lubricants Markets, *Conference "Revitalizing the Lubricants Business in the 21st Century"* Houston 1998.
540. M. Fuchs, The World Lubricants Market, Current Situation and Outlook, *12th International Colloquium Tribology, Esslingen*, 2000.
541. M. Fuchs, The global lubricating grease market, *NLGI Spokesman* **60** (1996) no. 5, 25 – 40.
542. J. A. Lurz, Grease Production Survey Report for the Calendar Years 1999, 1998, 1997 and 1996, NLGI, Kansas City, 2000.
543. W.-D. Abel, Biologisch schnell abbaubare Weichenschmierfette (Prüfung und Anwendung), *11th International Colloquium, Esslingen*, 1998, vol. I 2.7, 225 – 236.
544. R. L. Goyan, R. E. Melley, P. A. Wissner, W. C. Ong, Biodegradable lubricants, *Lubricating Engineering* **54** (1998) July, 10 – 17.
545. Collection and Disposal of Used Lubricating Oil, *CONCAWE Report* **5** (1996).
546. C. Kajdas, Used Oil Re-refining: Overview of Current Technologies Used, *3rd European Congress on Re-refining*, Lyon, 1996.
547. D. J. McKeagan, Economics of Re-Refining Used Lubricants, *Lubrication Engineering* (1992) 418 – 423.
548. P. F. V. C. Oosterkamp, KTI Re-Refining Technology, UNIDO Workshop, Karachi, 1992.
549. D. W. Brinkman, Large Grassroots Lube Re-Refinery in Operation, *Oils and Gas Journal* (1991) 60 – 63.
550. D. Peel, Greening of America, *World Base Oils '96 Conference*, Elgin, 1996.
551. C. Schoen, Einrohrreaktorverfahren zurAufarbeitung von Altoel und fluessigen Abfallstoffen, Description of the ENTRA Technology, ENTRA Ingenieur- und Handels GmbH, Achern, 1996.
552. F. D. Giovanna, P. Tromeur, G. Cohen, Successful Re-Refining in Practice, *International Used Oil Conference*, Orlando, 1998.
553. Waste Oil Re-Refining, *Hydrocarbon Processing* **59** 1980 143.
554. D. Klamann, *Lubricants and Related Products,* VCH Verlagsgesellschaft, Weinheim 1984.

Luminescent Materials

Karl A. Franz, Riedel-de Haën AG, Seelze, Federal Republic of Germany (Chap. 1)

Wolfgang G. Kehr, Riedel-de Haën AG, Seelze, Federal Republic of Germany (Chap. 1)

Alfred Siggel, Riedel-de Haën AG, Seelze, Federal Republic of Germany (Chap. 1)

Jürgen Wieczoreck, Riedel-de Haën AG, Seelze, Federal Republic of Germany (Chap. 1)

Waldemar Adam, Institut für Organische Chemie der Universität Würzburg, Federal Republic of Germany (Chap. 2)

1. Inorganic Phosphors 1
1.1. Introduction 1
1.2. Luminescence of Crystalline Inorganic Phosphors 3
1.2.1. Luminescence Processes 3
1.2.2. Models of Luminescent Systems . 4
1.2.2.1. Energy Configuration Diagram . . 4
1.2.2.2. Energy Band Model 5
1.3. Preparation and Properties of Inorganic Phosphors 8
1.3.1. Sulfides and Selenides 8
1.3.1.1. Zinc and Cadmium Sulfides and Sulfoselenides 8
1.3.1.2. Alkaline-Earth Sulfides and Sulfoselenides 10
1.3.2. Oxysulfides 10
1.3.3. Oxygen-Dominant Phosphors . . . 11
1.3.3.1. Borates 11
1.3.3.2. Aluminates 11
1.3.3.3. Gallates 12
1.3.3.4. Silicates 12
1.3.3.5. Germanates 13
1.3.3.6. Halophosphates and Phosphates . 13
1.3.3.7. Oxides 14
1.3.3.8. Arsenates 15
1.3.3.9. Vanadates 15
1.3.3.10. Niobates and Tantalates 15
1.3.3.11. Sulfates 16
1.3.3.12. Tungstates and Molybdates 16
1.3.4. Halide Phosphors 16
1.4. Uses of Luminescent Pigments . 19
1.4.1. Lighting 19
1.4.2. X-ray Technology 19
1.4.3. Cathode-Ray Tubes 23
1.4.4. Product Coding 26
1.4.5. Safety and Accident Prevention . . 27
1.4.6. Dentistry 28
1.4.7. Other Uses 29
1.5. Testing of Industrial Phosphors 30
2. Chemiluminescence and Bioluminescence 31
2.1. Introduction 31
2.2. History 32
2.3. Mechanisms of Light Emission . 33
2.4. Efficiency of Light Emission . . . 34
2.5. Uses 35
2.5.1. Light Sources 35
2.5.2. Chemical Analysis 35
2.5.3. Biotechnology 37
2.5.4. Medicine 37
3. References 38

1. Inorganic Phosphors

1.1. Introduction

Inorganic luminescent materials are synthetically produced crystalline compounds that absorb energy acting on them and subsequently emit this absorbed energy as light, either immediately or over a longer period, in excess of thermal radiation. This light emission, for which, in contrast to thermal radiation, the exciting energy first does not contribute to the thermal energy of the compound, is known as luminescence. It arises from excited states in atoms or molecules which have a lifetime of at least 10^{-9} s. This means that between excitation and ligh emission there is a time span of 10^{-9} s. It is in this respect that luminescence differs from the other luminous effects that, like luminescence, are also not classified as thermal radiation; these include light scattering, the Raman effect, and Cherenkov radiation.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_519

Of the various possibilities for classifying inorganic luminescent materials, categorization according to families of substances is generally preferred, because it permits a clearer presentation of the whole field than the classification according to luminescent effects, such as, for example, the various decay times.

Luminescent inorganic compounds are also known as luminophors and phosphors. Depending on the nature of the exciting energy, the resulting luminescence is named as follows:

Low energy photons	photoluminescence
Cathode rays	cathodoluminescence
X-rays	X-ray luminescence
Ions (particles)	ionoluminescence
Mechanical forces	triboluminescence
Electric field strength	electroluminescence
Chemical reactions	chemiluminescence
Biochemical reactions	bioluminescence

Light emission occurring during excitation and up to ca. 10^{-8} s afterwards is called fluorescence, while the glow continuing longer than 10^{-8} s is known as afterglow or phosphorescence. The decay time is the time in which the brightness decreases to 1/10 or 1/e of the initial intensity, for hyperbolic and exponential decay, respectively.

Certain compounds store the absorbed energy over a longer period and only emit it as light under the influence of heat or IR radiation. This process is called stimulation. The opposite phenomenon, in which the light emission is decreased by the action of heat or IR radiation on luminescent materials, is known as quenching.

History. Various reports indicate that luminescent inorganic compounds were already known in antiquity. Titus Livius (Livy) reports in his book "History of the Romans" that in 186 B.C. the maenads in the bacchanalia had "glowing torches" in their hair, which they then threw into the Tiber. Jorrisson assumes that this was luminescence from calcium sulfide [9].

One of the first reports of the synthesis of a luminescent material dates from 1603, when Casciavolus attempted to obtain gold by heating certain minerals, but instead obtained a rock that emitted reddish light in the dark. Here barium sulfide was formed by reduction of rock containing barium sulfate.

In 1609 Brand discovered a substance that glows brightly in air and named it phosphorus. Since then inorganic luminescent pigments have been named phosphors and their luminescence, phosphorescence. In the 1600 s and 1700 s, many other researchers investigated this light phenomenon, including Goethe.

The systematic scientific investigation of inorganic phosphors only began, however, in the 1800 s with Becquerel, Verneuil, Lenard, and Stokes. Stokes published the first definition of fluorescence in 1852, and drew up the law stating that the exciting light always has a shorter wavelength than the emitted light.

In 1866 Sidot discovered the luminescence of zinc sulfide, which has been the basis of many important industrial phosphors.

In 1867 Bequerel postulated two types of afterglow based on his investigations of uranyl salts: one with exponential and the other with hyperbolic decay.

In 1895 Röntgen used "barium platinocyanate" to make X-rays visible. This discovery led to the first industrial breakthrough for inorganic phosphors. In 1887 Ver-neuil recognized the significance of heavy-metal impurities for the luminescence of inorganic compounds.

In the 1900 s inorganic phosphors gained increasing scientific and industrial interest. In 1903 and 1904 Lenard and Klatt reported the connection between activators and light emission [10]. They were also the first to use fluxes for crystallization and to discuss how they function. The first experiments on producing luminescence solely by the action of an applied electric field were carried out by Gudden and Pohl in 1920 [11], Lossew in 1923 [12], and Destriau in 1936 [13]. Urbach (1926) [14] and Randall and Wilken (1945) [15] traced the different time constants for the decay of the afterglow back to the differing depth of traps.

In 1938/1939 Seitz [16] and Mott [17] developed the electron configuration diagram, on the basis of which, phenomena such as Stokes' Law became understandable.

Using quantum theory, Riehl and Schön, among others, developed in 1939 a model for the luminescence of crystalline phosphors which is essentially still valid [18].

1.2. Luminescence of Crystalline Inorganic Phosphors

1.2.1. Luminescence Processes

The luminescence of inorganic phosphors is comprised of the following processes: 1) absorption and excitation, 2) energy tranfer, and 3) emission.

Excitation requires that energy acts on the phosphor that can be absorbed by it. The excitation mechanism differs, according to the type of energy involved.

Excitation by Photons. The dependence of the absorptivity of a phosphor upon the photon energy is given by the absorption spectrum. This can be divided into various regions such as basic lattice absorption, edge absorption, and defect absorption [8].

Even when all the absorbed photons obey Stokes' Law, the absorption spectrum does not convey what proportion of them contributes to the excitation of photoluminescence. Often, excitation only takes place with photons of higher energy. The absorption that is effective for excitation of luminescence is given by the excitation spectrum, which shows the dependence of the luminescence intensity on the photon energy. Strong luminescence occurs when the exciting photon has sufficient energy to promote an electron from the valence band to the conduction band (see Section 1.2.2.2).

Phosphors that do not effectively absorb photons of a specific energy can often be sensitized for this radiation by incorporation of certain ions with higher absorption for this radiation in the host crystal.

Excitation by Electrons. A portion of the electrons impinging on a phosphor are backscattered elastically with loss of energy. The backscattering factor η and the averaged atomic number $\bar{Z}$ of the atoms are related by [19]

$$\eta = \left(\ln\bar{Z} - 1.5\right)/6$$

For phosphors based on zinc sulfide ($Z_{\mathrm{Zn}} = 30, Z_{\mathrm{S}} = 16$; hence, $\bar{Z}_{\mathrm{ZnS}} = 23$) this gives a backscattering coefficient $\eta \approx 0.27$. This means that 27 % of the energy supplied is not available for luminescence. The effect of the electron energy on backscattering is described in [20].

The electrons that are not scattered elastically lose their energy on penetrating a phosphor crystal by inelastic scattering and formation of secondary electrons.

An expression for the specific energy loss $\mathrm{d}E/\mathrm{d}x$, i.e., the decrease in energy of the primary electron after traveling a distance x in the phosphor, is given by [21]

$$-\frac{\mathrm{d}E}{\mathrm{d}x} = \frac{2\pi N Z e^4}{E} \ln\left(\frac{E}{E_{\mathrm{i}}}\right)$$

where E is the electron energy, N the number of electrons per cm^3, Z the atomic number, e the electronic charge, and E_{i} the average ionization energy of all the electrons. A detailed presentation of this process in the case of cathode-ray excitation is given in [22, 23]. For primary electrons of higher energy (>30 keV), the penetration depth of an electron is given by the Thomson – Whidelington law as

$$\mathrm{E}^2 - E_x^2 = ax$$

where E is the initial energy of the primary electrons, E_x is the energy of the electron after traveling a distance x, and a is a constant. The total penetration depth R is given by $E_x = 0$ and $x = R$, thus: $E^2 = aR$. The range is therefore proportional to the square of the initial energy of the primary electrons.

The inelastically scattered electrons and the secondary electrons produce charge carriers in the phosphor, which then recombine with emission of luminescent radiation, either directly or after traveling in the lattice.

Excitation by X-Rays. As forms of electromagnetic radiation, X- and γ-rays follow the same absorption rules as photons. Because of the high energy of the secondary electrons formed on absorption, however, the excitation mechanism more closely resembles electron excitation than photon excitation. The absorption is given by

$$I = I_0 \mathrm{e}^{-\mu x}$$

where I_0 is the intensity of the incident X-rays, μ is the absorption of the radiation per centimeter, and x is the distance traveled in the material in centimeters.

In practice the mass extinction coefficient μ/ϱ (ϱ = density of the material) is used instead of the absorption coefficient μ.

The X-ray quanta absorbed by a phosphor mainly produce photoelectrons whose kinetic energy is at most equal to that of the absorbed X-ray quanta. In addition, Compton scattering and pair formation occur. The subsequent processes then continue as in electron excitation.

Energy Transfer. The excitation of phosphors does not always take place at the luminescence center but also at other locations in the crystal. A recombination and the associated emission require a transfer of the absorbed energy to the luminescence center. This can take place by the following mechanisms:

1) migration of excited electrons and electron holes,
2) migration of excitons (electron – hole pairs),
3) resonance between atoms with overlapping electric fields, and
4) reabsorption of a light quantum emitted from one activator center by another emission center.

The recombination processes are to a great extent independent of the nature of the excitation. The only parameters that are influenced by the type of excitation are the relative intensities of the various emission bands, the quantum yields, and the decay behavior. For photon excitation, the quantum yield can be up to 70 %, whereas it reaches a maximum of 25 % for electron excitation. The afterglow time and afterglow intensity are lower for electron excitation than for photon excitation.

The simplest case of reaction kinetics occurs when the excitation and emission occur in the same atom, molecule, or luminescence center. The recombination can then be treated as a first-order unimolecular reaction. The decrease in the number of excited electrons is given by:

$$\mathrm{d}n/n = -\lambda dt.$$

The number of excited electrons as a function of time is thus

$$n = n_0 \cdot e^{-\lambda t} \text{ where } \lambda = c/k$$

and the luminescence intensity is given by

$$I = I_0 \cdot e^{-\lambda t}.$$

The decay time is therefore independent of the number of other similarly excited atoms or molecules. If the recombination is delayed, e.g., by migration of excited electrons, luminescence takes place by a second-order bimolecular reaction. The probability of a luminescent recombination of the excited electron with the holes is then proportional to the product of the concentration n of electrons and the concentration p of holes. When $n = p$, $I = -k\,\mathrm{d}n/\mathrm{d}t = cnp = cn^2$ and

$$I = \frac{I_0}{(1+t/\tau)^2} \text{ where } \tau = \frac{k}{\sqrt{cI_0}}.$$

It follows that the lower the initial intensity is, and the further the decay has progressed, the slower the decay to the half value is.

This hyperbolic decay law is only of limited validity. If the excited electron is momentarily trapped before recombination, very complex interactions can arise.

1.2.2. Models of Luminescent Systems

In many cases, luminescence mechanisms are not understood in detail; therefore, simplified and thus generally applicable models are used for their interpretation.

1.2.2.1. Energy Configuration Diagram

This model, based on the energy level diagrams of atoms and molecules, is applicable to luminescence processes in which excitation and emission take place at the same luminescence center, as in the case of gases and organic molecules.

The energy relationships in a luminescence process are presented in a configurational coordinate diagram (Fig. 1). This illustrates the relationship between the potential energy E of the luminescence center (ordinate) and a space coordinate (abscissa), which gives the representative separation between the atom involved and its nearest neighbors or the deflection from its spatial equilibrium position.

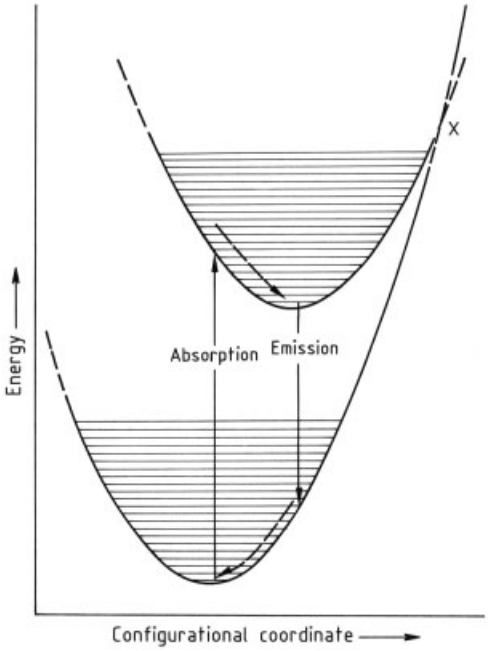

Figure 1. Configurational-coordinate diagram

The lower curve represents the ground state of the luminescence center with the vibrational levels, and the upper curve the excited state.

By absorbing exciting radiation the electrons are raised from the ground state to the excited state. These transitions take place so rapidly (10^{-9} s) that no displacement of the atomic nuclei occurs (Franck – Condon principle).

The space coordinate thus remains unchanged and the transitions can be represented by vertical lines. Because the excited system is not immediately in a state of equilibrium after absorption of energy, it first moves towards the lowest vibrational level with loss of energy to the lattice (lattice relaxation). From there the electrons can return to one of the vibrational levels of the ground state with emission of light and finally relax to the equilibrium position. The model explains the band form of the absorption and emission by the distribution over the vibrational levels and the Stokes' shift by means of nonradiative relaxation processes.

If an electron reaches the energy corresponding to the overlap point X of the curves for the excited and ground states by means of absorption and thermal activation, it can cross over into a vibrational level of the same energy in the ground state and subsequently lose the entire excitation energy nonradiatively by relaxation. Higher temperatures therefore reduce the luminescence quantum yield.

Since the excited electrons are captured by the luminescence center, this model does not account for phosphors in which photoconductivity occurs during the excitation.

Activated silicates, halides, halophosphates, uranyl compounds, and barium tetracyanoplatinate are examples of phosphors whose luminescence is explained by the configurational-coordinate diagram. The activator-free tungstates, molybdates and vanadates also belong to this class.

1.2.2.2. Energy Band Model

For many luminescent substances, excitation and emission do not take place within one atom or molecule. In these cases the energy band model is used.

In a crystal, the outer electrons can no longer be assigned to individual atoms but only to the whole crystal lattice. The interaction in the crystal leads to a splitting of the atomic orbitals into energy bands. At low temperatures, the highest completely occupied valence band is separated from the energetically lowest unoccupied conduction band by the forbidden zone, in which in ideal crystals, there are no allowed states. For real crystals, in which the periodic sequence of the crystal unit cells is distorted by lattice defects or impurity ions, discrete energy levels can develop in this forbidden zone. The energy levels near the valence band are termed donor levels, and those near the conduction band, acceptor levels or traps (Fig. 2, see next page).

By absorption of radiation with sufficient energy, an electron is promoted from the valence band to the conduction band (a), where it relaxes to the band edge with radiationless loss of energy to the lattice (d). The positive hole in the valence band is immediately filled by an electron from a donor level (b, c, i). In the case of fluorescence, a radiant transition of the electron from the conduction band to the positive hole formed in the activator level takes place (e). The wavelength of the emitted light is given by the energy difference between the band edge of the conduction band and the activator level.

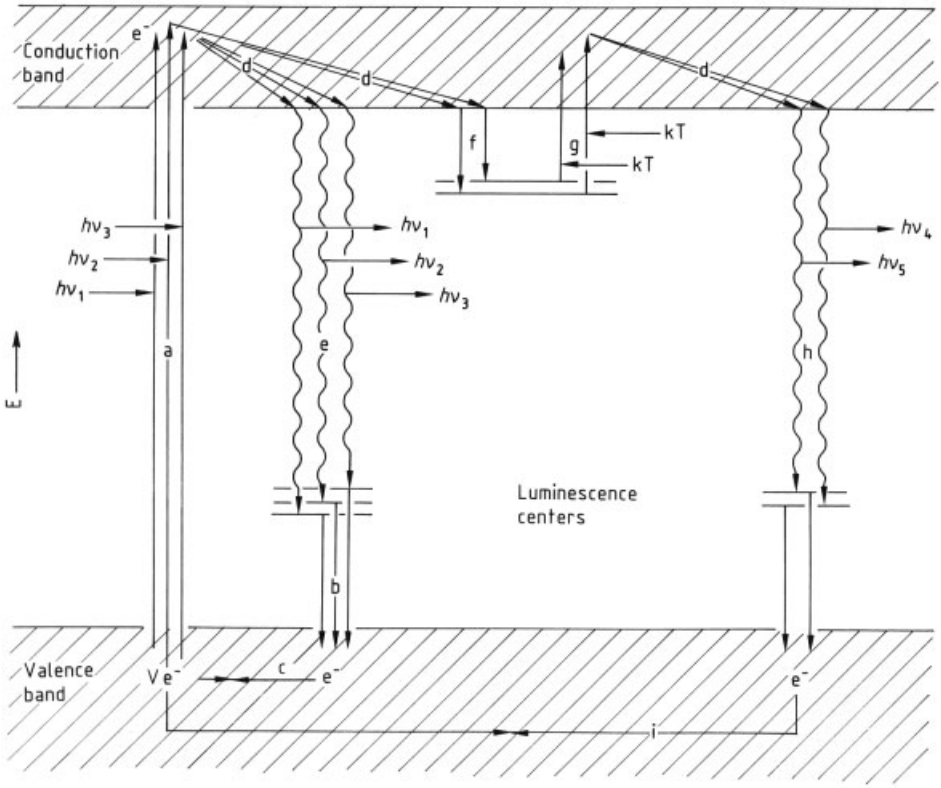

Figure 2. Energy level diagram for a phosphor

Before recombination, the electron can, however, also be captured nonradiatively by traps (f), from which it is only liberated by input of thermal energy or by absorption of long-wavelength radiation (g). Then, having been retarded, it recombines with the emission of light (h, phosphorescence, stimulated luminescence).

Unoccupied deep traps act as quenching centers and reduce the luminescence because the probability of a transition into the conduction band is lower than that for a radiationless recombination with an electron hole in the valence band. The causes of these types of quenching centers can be particular impurities in the crystal (killers) or lattice deformations brought about mechanically.

Most sulfides belong to the group of phosphors with delocalized luminescence processes, to which the band model can be applied.

Rare-Earth Activators. Phosphors activated by rare earth ions exhibit some peculiarities. In the energy level diagram of the rare earths, luminescence processes often correspond to electron transitions within the incompletely filled $4f$ shell, which is extensively shielded from the crystal field splitting that results from incorporation into a host lattice. Consequently, these phosphors have narrow band spectra, which are to a great extent independent of the nature of the host lattice. The sharp, discrete bands give rise to luminescence colors with high saturation. Because of the low interaction with the crystal lattice, the luminescence quantum yield of phosphors activated with rare earths is often high, and, compared with other phosphors, quenching occurs only at higher temperatures or higher activator concentrations. Europium, terbium, cerium, and neodymium are of industrial interest as activators. Ytterbium, terbium, and praseodymium are also used as sensitizers [24–26].

Multiphoton Phosphors. The particular electronic structure of the rare earths is such that they exhibit several sharp energy levels

when behaving as activators. This permits the energy from two or more photons in succession to be transmitted to one luminescence center. The excited electron thus reaches an even higher energy level.

When the electron returns from this higher level to the ground state, a photon with about twice the energy of the exciting photon is emitted. In this way it is possible to convert IR radiation into visible light. This effect is falsely termed Anti-Stokes' luminescence because no account is taken of the fact that the prerequisite for the emission of the high-energy light quanta is the previous excitation by two or more photons, whose combined energy is greater than that of the light quanta emitted, as Stokes' law requires. The low probability of a transition from the first metastable excited state to the doubly excited state means that the quantum yield is usually low. Alkali and alkaline-earth metal halides, and the halides, oxyhalides, and oxysulfides of yttrium and lanthanum are principally used as the base material, while Er^{3+}, Ho^{3+}, and Tm^{3+} serve as activators. The quantum yield is considerably increased by the addition of Yb^{3+} as sensitizer.

The luminescence processes can be best understood by means of the energy level diagram. Figure 3 shows the transitions for a phosphor with ytterbium as sensitizer and erbium as activator. When subjected to IR radiation, an electron from the ytterbium ground state $^2F_{7/2}$ is promoted to the excited $^2F_{5/2}$ state. The energy of the excited Yb^{3+} is transferred to Er^{3+} and promotes an electron from the ground state to the $^4I_{11/2}$ state. By further energy transfer, the electron reaches the metastable $^4F_{7/2}$ level. From here the electron first passes back to the level $^4S_{3/2}$ nonradiatively and then returns to the ground state $^4I_{15/2}$ with the emission of green light. For a specific phosphor, the emission spectra for single-photon excitation (e.g., UV) and multiphoton excitation are the same.

Some multiphoton phosphors for the conversion of IR into visible light are as follows: $BaYb_2F_8$:Er [27], $Cs_2Na(Er, Yb, Y)Cl_6$ [28], $NaLnF_4$:Yb, Er (Ln = Y, Gd, La) [29], LaF_3:Yb, Er [30], $BaYF_5$:Yb, Er [31], YOCl: Yb, Er [32], $(Y, Yb, Er)F_3$ [33], $CdF_2 \cdot YbF_3 \cdot ErF_3$ [34], and glass ceramics [35].

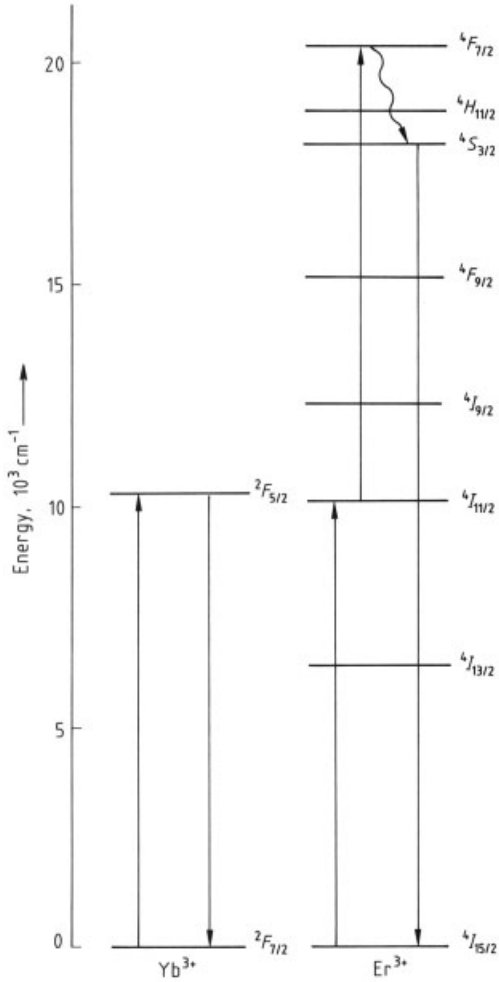

Figure 3. Multiphoton excitation, energy level diagram

Raw materials of very high purity (min. 99.99 %) are needed for the production of effective multiphoton phosphors since even traces of impurities lead to side reactions, which reduce the already low quantum yields still further.

The starting materials for halides are mostly mixtures of the corresponding oxides or oxalates, which are treated with very high purity hydrochloric or hydrofluoric acid.

The sensitizer content (Yb^{3+}) can be up to 0.5 mol % and the activator content (Er^{3+}, Tm^{3+}), up to 0.05 mol %. After washing and filtration the halides are then heated in N_2, Ar, HF or HCl to 600 – 1100 °C and fired for several hours at this temperature. It is advantageous to

add NH_4F or NH_4Cl to the firing mixture [27–34].

In many cases the multiphoton phosphors are used as single crystals, produced by the Bridgman process. Because of their relatively low efficiency, the industrial application of these phosphors is limited to use in IR radiation detectors (see Section 1.4.7).

1.3. Preparation and Properties of Inorganic Phosphors

1.3.1. Sulfides and Selenides

1.3.1.1. Zinc and Cadmium Sulfides and Sulfoselenides

The raw materials for the production of sulfide phosphors are high-purity zinc and cadmium sulfides, which are precipitated from purified salt solutions by hydrogen sulfide or ammonium sulfide [2,7,8,36]. The concentration of contaminants such as Fe, Co, or Ni must be below 1 % of the activator concentration. The $Zn_{1-x}Cd_xS$ can be produced by mixing precipitated zinc sulfide and cadmium sulfide. However, coprecipitation from mixed zinc – cadmium salt solutions is preferred because of the better homogeneity.

The most important activators for sulfide phosphors are copper and silver, followed by manganese, gold, rare earths, and zinc. The charge compensation of the host lattice is effected by coupled substitution with mono- or trivalent ions (e.g., Cl^- or Al^{3+}). In addition, disorders, such as unoccupied sulfur positions, can also contribute to charge compensation.

For the synthesis of phosphors, the sulfides are homogenized with readily decomposed compounds of the activators and coactivators in the presence of a flux and are fired in quartz crucibles at 800 – 1200 °C.

The luminescent properties can be influenced by the nature of the activators and coactivators, their concentrations, the composition of the flux, and the firing conditions. In addition, specific substitution of zinc or sulfur in the host lattice by cadmium or selenium is possible, which also influences the luminescent properties. The crystal lattice of zinc sulfide is also of influence. It is dimorphic and crystallizes below 1020 °C in the cubic zinc-blende structure and above that temperature in the hexagonal wurtzite lattice. When the zinc is replaced by cadmium, the transition temperature is lowered so that the hexagonal modification predominates. Substitution of sulfur by selenium, on the other hand, stabilizes the zinc blende lattice.

The $Zn_{1-x}Cd_xS$ phosphors have found the greatest industrial use.

Self-activation. Although pure substances do not normally luminesce, zinc sulfide that has been fired in the presence of a halogen luminesces bright blue [37,38]. The luminescence center for activator-free zinc sulfide is assumed to be a cation vacancy. The charge compensation occurs through exchange of S^{2-} by Cl^-.

Cation holes can also be created by coactivation with trivalent metal ions or by incorporation of oxygen [39]. The luminescence band of self-activated zinc sulfide with the zinc-blende structure exhibits a maximum at 470 nm. On transition to the wurtzite structure, the maximum shifts to shorter wavelengths. In the mixed crystals zinc sulfide – cadmium sulfide and zinc sulfide–zinc selenide, the maximum shifts to longer wavelengths with increasing cadmium or selenium concentration (Fig. 4).

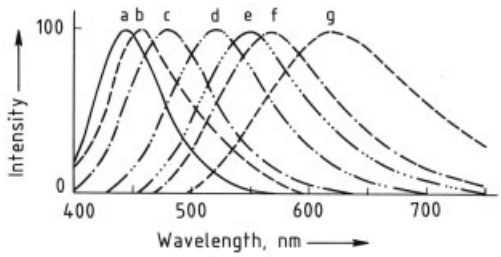

Figure 4. Spectra of self-activated $Zn_{1-x}Cd_xS$ phosphors a) $x = 0$, hexagonal; b) $x = 0$, cubic; c) $x = 0.1$; d) $x = 0.2$; e) $x = 0.3$; f) $x = 0.4$; g) $x = 0.5$

Silver Activation. Doping zinc sulfide with silver leads to the appearance of an intense emission band in the blue region of the spectrum at 440 nm, which has a short decay time.

Weak luminescence in the green (520 nm) and red regions can also occur. The blue band is assigned to recombination at substitutionally incorporated silver ions [40,41]. The red band is caused by luminescence processes in associates of silver ions occupying zinc positions

with neighboring sulfur vacancies [42], whereas the nature of the green luminescence center is as yet unknown [43].

Industrial silver-activated zinc sulfide phosphors use the intense blue emission exclusively. The ZnS:Ag phosphor for cathode ray tubes is obtained by firing zinc sulfide and silver nitrate at ca. 1000 °C in the presence of sodium chloride (coactivator Cl^-) [44]. The afterglow can be further reduced by addition of 10^{-3} – 10^{-4} % of nickel ions.

The substitution of zinc by cadmium in the ZnS:Ag phosphor leads to a shift of the emission maximum from the blue over to the green, yellow, red to the IR spectral region (Fig. 5) [8, 45–48].

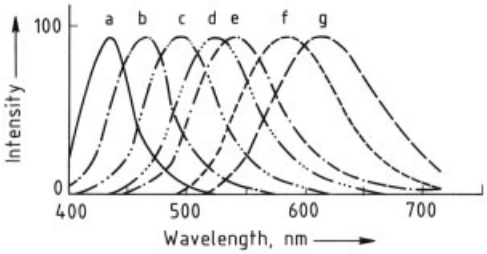

Figure 5. Spectra of silver-activated $Zn_{1-x}Cd_xS$ phosphors (10^{-2} wt % Ag)
a) $x = 0$; b) $x = 0.1$; c) $x = 0.2$; d) $x = 0.3$; e) $x = 0.4$; f) $x = 0.5$; g) $x = 0.6$

Copper Activation. Activation with copper causes an emission in zinc sulfide which consists of a blue (460 nm) and a green band (525 nm) in varying ratios, depending on the preparation. The green band is assigned to monovalent copper ions incorporated into the host lattice by substitution in the zinc positions, with halide ions acting as coactivators. The blue band is attributed to associates between Cu^+ ions occupying Zn^{2+} positions and interstitially incorporated Cu^+ ions [49–51]. The phosphor ZnS:Cu can also exhibit luminescence in the red or infrared spectral region, depending on the preparation. The red emission band is assigned to a Cu^+ ion associated with a sulfur vacancy [52].

Usually, green-emitting ZnS:Cu phosphors are used for industrial applications. The preparation consists of firing a mixture of ZnS, $Cu(NO_3)_2 \cdot 6\,H_2O$ (corresponding to 50 – 1000 ppm Cu in the phosphor), and 5 wt % NaCl at 900 – 1000 °C in covered fused silica crucibles [53].

The decay time increases with the copper concentration and can be increased to several hours by adding small quantities of cobalt (Fig. 6) (see Section 1.4.4 and 1.4.5). The decay time can be shortened by addition of nickel. It is thus possible to adapt the decay times of copper-doped zinc sulfides to various requirements. The distance between the valence and conduction bands is altered by partial substitution of Zn^{2+} by Cd^{2+}, and of S^{2-} by Se^{2-}. In this way the emission spectrum can be adjusted (Figs. 7 and 8, see next page).

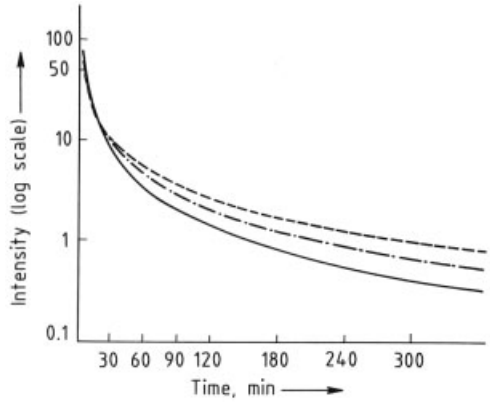

Figure 6. Decay curves of some commercial long-afterglow phosphors
—- LUMILUX Green N; –·–·LUMILUX Green N 5; - - - - LUMILUX Green N 10

Other Activation. Zinc sulfide forms a wide range of substitutionally mixed crystals with manganese sulfide. Manganese-activated zinc sulfide has an emission band in the yellow spectral region at 580 nm [54]. This band can be assigned to the inner transitions of the Mn^{2+} ions incorporated into zinc positions. At higher manganese concentrations, a spectral shift into the orange region takes place as a result of manganese – manganese interactions. At high manganese concentrations, excitation by means of an electric field is possible.

The activation of zinc sulfide with gold leads to luminescence in the yellow-green (550 nm) or blue (480 nm) spectral regions [43, 55–58], depending on the preparation, whereas a blue-white luminescing phosphor is formed on activation with phosphorus [59].

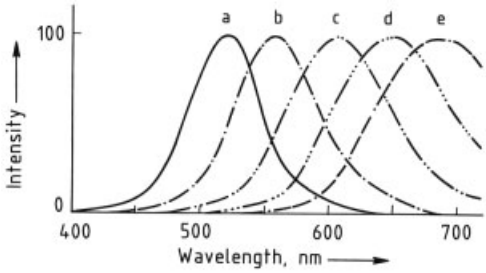

Figure 7. Spectra of copper-activated $Zn_{1-x}Cd_xS$ phosphors (5×10^{-3} wt % Cu)
a) $x = 0$; b) $x = 0.1$; c) $x = 0.2$; d) $x = 0.3$; e) $x = 0.4$

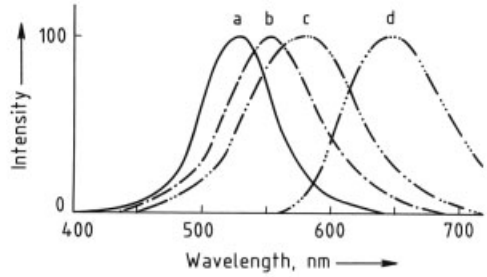

Figure 8. Spectra of copper-activated $ZnS_{1-x}Se_x$ phosphors (5×10^{-3} wt % Cu)
a) $x = 0$; b) $x = 0.2$; c) $x = 0.4$; d) $x = 1$

Multiple activation of zinc sulfide is also possible. Zinc – cadmium sulfide, doubly activated with silver and gold, which is used as a white-luminescing one-component phosphor for monochromic cathode-ray tubes [60], and the yellow-luminescing ZnS:Cu, Au, Al phosphor, whose emission color corresponds to that of Zn_{1-x} Cd_x S:Cu [58], are known.

1.3.1.2. Alkaline-Earth Sulfides and Sulfoselenides

Activated alkaline-earth metal sulfides have been known for a long time; their luminesence is very varied. Emission bands between the ultraviolet and near infrared can be realized by varying the activation [61–65]. They are produced by precipitation of sulfates or selenites from purified solutions, followed by reduction with Ar – H_2. The addition of activators, for example, copper nitrate, manganese sulfate, or bismuth nitrate, is followed by firing for 1 – 2 h. Alkaline-earth halides or alkali-metal sulfates are sometimes added as fluxes.

Among the long afterglow alkaline-earth sulfides, only (Ca, Sr)S:Bi^{3+} and CaS:Bi^{3+} still have any real importance because their respective blue and violet luminescence cannot yet be achieved with the less hydrolysis-sensitive zinc sulfide phosphors.

The alkaline-earth sulfides activated with rare earths are of greater importance. These phosphors are considered suitable for use in cathode-ray tubes because of the linear dependence of their brightness on current over a wide range [61–68]. The emission band of MgS with 0.004 mol % europium has a maximum at 600 nm, whose cathode-ray brightness reaches or surpasses that of $Y_2O_2S:Eu^{3+}$. On substitution of magnesium by calcium, a nonlinear shift of the emission maximum occurs ($Ca_{0.4}Mg_{0.6}S:Eu^{2+}$, $\lambda_{max} = 668$ nm, $CaS:Eu^{2+}$, $\lambda_{max} = 650$ nm) but at the same time there is a decrease in brightness [69].

$CaS:Ce^{3+}$ is a green-emitting phosphor. On activation with 10^{-4} mol % cerium, the emission maximum occurs at 540 nm. Greater activator concentrations lead to a red shift; substitution of calcium by strontium, on the other hand, leads to a blue shift; $SrS:Ce^{3+}$ (10^{-4} mol % cerium) has $\lambda_{max} = 483$ nm [70].

$MgS:Ce^{3+}$ (0.1 %) has two emission bands in the green and red spectral regions at 525 and 590 nm; $MgS:Sm^{3+}$ (0.1 %) has three emission bands at 575 nm (green), 610 (red), and 660 nm (red) [66].

Calcium or strontium sulfides, doubly activated with europium – samarium or cerium – samarium, can be stimulated by IR radiation (see Section 1.4.7), since Sm^{3+} acts as a trap through the transition $Sm^{3+} + e^- \rightarrow Sm^{2+}$ [67]. Emission occurs at europium or cerium and leads to orange-red ($SrS:Eu^{2+}, Sm^{3+}$) or green ($CaS:Ce^{3+}, Sm^{3+}$) luminescence.

1.3.2. Oxysulfides

Red-emitting $Y_2O_2S:Eu^{3+}$ phosphors were developed especially for excitation with cathode rays [71, 72]. Production generally starts from the mixed rare-earth oxides, which are fired with sodium carbonate and sulfur. The polysulfide formed acts as a flux and sulfurizing agent [73]. The addition of a small amount of terbium sensitizes the europium luminescence [74].

The main emission lines of $Y_2O_2S:Eu^{3+}$ occur at 565 and 627 nm. The intensity of the long-wavelength emission increases with the europium concentration, whereby the color of the emission shifts from orange to deep red. For use in color picture tubes, $Y_2O_2S:Eu^{3+}$ (see Section 1.4.3) contains ca. 4 wt % europium [75]; $Y_2O_2S:Eu^{3+}$ is unsuitable for use in lamps because of extensive thermal quenching of the luminescence above room temperature. Terbium in Y_2O_2S has main emission bands in the blue (489 nm) and green spectral regions (545 and 587 nm), whose intensity ratio depends on the terbium concentration. At low doping levels, $Y_2O_2S:Tb^{3+}$ luminesces blue-white, while at higher levels the color tends towards green. Although $Y_2O_2S:Tb^{3+}$ has been used for X-ray intensifying screens, $Gd_2O_2S:Tb^{3+}$, which also exhibits green luminescence, is predominantly used in X-ray technology because of its higher absorption. White luminescence for use in special cathode-ray tubes can be achieved with blends of Y_2O_2S:Tb and Y_2O_2S:Eu [76] or by double activation of Y_2O_2S with terbium and europium [77].

1.3.3. Oxygen-Dominant Phosphors

Oxygen-dominant phosphors are generally produced by solid-state reactions in which the components of the phosphor (mostly oxides) are intimately mixed and fired at 500 – 1500 °C. The desired phosphor is formed by solid-state reaction below the melting point of the compound during a firing period of 3 – 16 h.

The Tamman rule, which states that a measurable rate of change of position of the lattice components begins at two-thirds of the absolute melting point, gives an indication of the necessary firing temperature [78].

The activators are added in the form of oxides, oxalates, carbonates, or other compounds which readily decompose at higher temperatures.

The oxidation state of the activators can be influenced by the choice of the furnace atmosphere. As in the case of the sulfide phosphors, contamination by Fe, Co, or Ni greatly reduces the quantum yield.

As well as higher quantum yields, oxygen-dominant phosphors stand out because of their good temperature stability. For most substances the quantum yield only decreases noticeably above 250 °C.

1.3.3.1. Borates

Strontium fluoroborate $Sr_3B_{12}O_{20}F_2$: Eu^{2+} is used in photocopier lamps, black light lamps, and lamps for medical applications (λ_{max} = 366 nm).

It is produced by the addition of aqueous ammonia to an aqueous solution of $Eu(NO_3)_3$, $Sr(NO_3)_2$, and H_3BO_3 at 95 °C and pH 8 to precipitate europium-containing $SrB_6O_{10}\cdot 5\,H_2O$. The latter is mixed with SrF_2 and, after a preliminary reaction at 400 °C, is fired at 820 °C in air and then at 900 °C under N_2– H_2 (95/5) [79].

1.3.3.2. Aluminates

Yttrium aluminate $Y_3Al_5O_{12}:Ce^{3+}$ (YAG) is produced by precipitation of the hydroxides with NH_4OH from a solution of the nitrates and subsequent firing at 1450 °C [80]. To enable a lower reaction temperature of 1300 °C to be used, a freeze-dried powder is also used, which is produced from a solution of $(NH_4)_2Ce(NO_3)_6$, $Al_2(SO_4)_3\cdot 18\,H_2O$, and $Y_2(SO_4)_3$ [81].

The solubility of Ce^{3+} in $Y_3Al_5O_{12}$ depends on the atmosphere in which the mixture is fired. At 1450 °C in a hydrogen atmosphere, up to 6 mol % Ce^{3+} can be incorporated into the $Y_3Al_5O_{12}$ lattice. At higher concentrations, $CeAlO_3$ is formed as a second phase. At 1450 °C in air the solubility is only 2 mol %; higher concentrations lead to CeO_2 as the second phase. In each case the cerium is incorporated as Ce^{3+} [81].

When fired in a reducing atmosphere, $Y_3Al_5O_{12}$ exhibits a pronounced afterglow due to traps formed by oxygen vacancies [82]. Subsequent annealing in air diminishes this effect and leads to decay times of 200 – 300 ns; therefore, $Y_3Al_5O_{12}:Ce^{3+}$ is used in flying-spot scanner tubes. The emission maximum is at 550 nm. This phosphor is classified under P 46 (TEPAC) and KG (WTDS) (see Section 1.4.3).

Cerium magnesium aluminate (CAT) $Ce_{0.65}Tb_{0.35}MgAl_{11}O_{19}$ is produced by coprecipitation of the metal hydroxides from a so-

lution of the nitrates with NH_4OH and subsequent firing at 700 °C for 2 h followed by 1 h at 1500 °C [83]. Another possibility is to fire a mixture of Al_2O_3, $MgCO_3$, CeO_2, and Tb_4O_7 for 5 h at 1500 °C with small quantities of MgF_2 or AlF_3 as a crystallizing agent [83]. A strongly reducing atmosphere is necessary to ensure that the rare earths are present as Ce^{3+} and Tb^{3+}.

Cerium magnesium aluminate is a highly efficient green phosphor ($\lambda_{max.} = 541$ nm), which is used in trichromatic fluorescent lamps. The high quantum yield of 65 % is effected by the energy transfer from the sensitizer, Ce^{3+}, to the activator, Tb^{3+}.

Barium magnesium aluminate (BAM) $BaMg_2Al_{16}O_{27}$:Eu^{2+} is produced by firing a mixture of Al_2O_3, $BaCO_3$, $MgCO_3$, and Eu_2O_3 in the presence of a flux in a weakly reducing atmosphere at 1100 – 1200 °C [84]. It is used as the blue component ($\lambda_{max} = 447$ nm) in trichromatic fluorescent lamps.

Yttrium aluminum gallium garnet (YAGG) $Y_2Al_3Ga_2O_{12}$:Tb^{3+} [85] can be produced by mixing a stoichiometric quantity of the oxides (5 – 7 mol % terbium) with 20 % BaF_2 in ethanol. After drying, the mixture is fired for 2 h at 1500 °C. The flux is washed out with 20 % nitric acid. Yttrium aluminum gallium garnet is a green phosphor with high brightness for use in projection television tubes. It does not show any saturation even at high current densities of 160 μA/cm^2.

For the mixed aluminate – gallate, the attainable brightness shows a greater dependency on the firing temperature than is the case with pure aluminate or gallate. Differences of 100 °C in the firing temperature reduce the brightness by 25 –40 %.

1.3.3.3. Gallates

Strontium thiogallate is produced by firing a mixture of high-purity sulfides in a stream of hydrogen sulfide at 900 – 950 °C for 2 h; Pb^{2+} and Ce^{3+} (with Na^+ for charge compensation) are used as activators [86, 87].

$SrGa_2S_4$:Ce^{3+}, Na^+ with 4 mol % cerium ($\lambda_{max.} = 455$ nm, $\eta = 5.0$ %) and $SrGa_2S_4$:Pb^{2+} with 8 mol % lead ($\lambda_{max.} = 595$ nm, $\eta = 4.6$ %) are used in flying-spot scanner tubes (see Section 1.4.3) because of their short decay times (I_0/100) of 1.55 μs and 1.24 μs, respectively. Their low ageing upon electron beam bombardment is also noteworthy. Because of the relatively broad emission bands, a blend of 15 wt % $SrGa_2S_4$: Ce^{3+}, Na^+ with 85 wt % $SrGa_2S_4$: Pb^{2+} gives an extensive, regular emission throughout the visible spectrum (Fig. 9) [88].

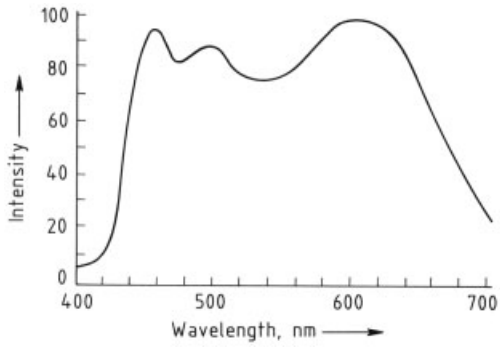

Figure 9. Spectrum of phosphor blend $SrGa_2S_4$:Ce + $SrGa_2S_4$:Pb (15/85)

1.3.3.4. Silicates

Zinc orthosilicate Zn_2SiO_4:Mn^{2+} can be produced by firing a mixture of ZnO, SiO_2, and $MnCO_3$ at 1200 °C for 2 h. Small quantities of MgF_2 or ZnF_2 accelerate the reaction [7].

$ZnSiO_4$:Mn is used as a green phosphor ($\lambda_{max.} = 525$ nm) in fluorescent lamps and cathode-ray tubes and is cited in the TEPAC list as P 1 and in the WTDS as GJ. Its ability to be excited by low-energy electrons (a few eV), its linear increase in brightness with electron beam intensity, and its high resistance to burn-in, even at high current density, are advantageous. The resistance to burn-in can be improved still further by employing a sol – gel process during the production. This involves reaction of a $[Zn(NH_3)_4](OH)_2$ solution with $Si(OC_2H_5)_4$ to give a gel, which is subsequently dried and fired [5]. The decay time of ca. 30 ms can be increased to the order of minutes by the addition of arsenic (a few ppm to 0.5 wt %) [89].

$ZnSiO_4$:Mn, As is standardized as P 39 in TEPAC and GR in WTDS.

By successive replacement of zinc by beryllium (up to 10 mol %), the red Mn^{2+} line at

610 nm becomes stronger and the line at 525 nm weaker, so that a shift of the emission color into the orange region occurs. Because of the toxicity of beryllium these phosphors are, however, generally no longer used.

Yttrium orthosilicate $Y_2SiO_5:Ce^{3+}$ is used as a highly resistant blue cathode-ray phosphor in projection television tubes ($\eta = 6\,\%$) [90]. It is classified as BH (WTDS) and P 47 (TEPAC). The emission maximum is between 400 nm (0.1 % Ce) and 460 nm (10 % Ce). The production involves firing a mixture of Y_2O_3, SiO_2, and $Ce(SO_4)_2 \cdot 4\,H_2O$ for 16 h at 1250 °C. Further firing for 4 h at 1160 °C in a reducing atmosphere is necessary to bring the cerium into the trivalent state. The addition of NaF or YF_3 improves crystallization. Doping with Tb^{3+} instead of Ce^{3+} leads to an efficient ($\eta = 7\,\%$) green phosphor, in which coactivation by Ce^{3+} can increase the luminescence by a further 10 % [91]. An yttrium orthosilicate doped with Ce, Tb, and Mn, is used in artificial teeth.

The phosphors obtained by solid-state reactions are not sufficiently resistant to ageing for use in fluorescent lamps. In a pseudo-sol – gel process an aqueous solution of (Y, Tb) $(NO_3)_3$ is treated with SiO_2 and heated to 1200 °C. After subsequent reductive firing under N_2/H_2 at 1700 – 1800 °C, a phosphor is obtained with a resistance to ageing comparable to that of the aluminates [92].

Barium disilicate $BaSi_2O_5$:Pb emits in the long-wavelength ultraviolet ($\lambda_{max} = 350$ nm). This phosphor is used in blacklight fluorescent lamps; the quantum yield is 75 %.

The phosphor is produced by stirring $BaCO_3$, BaF_2, $Pb(NO_3)_2$, and SiO_2 (fumed silica) to a slurry in methanol. After drying, the mixture is fired for 30 min at 1050 °C, with the heating up and cooling down between 800 and 1050 °C being carried out at a rate of 200 °C/h [93].

1.3.3.5. Germanates

Magnesium fluorogermanate, $3.5\,MgO \cdot 0.5\,MgF_2 \cdot GeO_2:Mn^{4+}$ is a brilliant red phosphor used in fluorescent lamps [94]. Because of its yellowish color, it strongly absorbs the blue mercury emission lines in these lamps, so that a saturated red light results. The emission maximum lies at 660 nm. The temperature stability of this phosphor is particularly high; therefore, it is used for high-pressure mercury lamps. With rising temperature the absorption edge of the host lattice (350 – 400 nm) shifts to longer wavelengths. Therefore, the brightness increases with increasing temperature up to 300 °C because the mercury line at 365 nm is then absorbed more effectively. The phosphor is produced by firing a stoichiometric mixture of MgO, MgF_2, and GeO_2 with 0.01 mol % $MnCl_2$ for 5 h at 1150 °C in an oxidizing atmosphere [94].

1.3.3.6. Halophosphates and Phosphates

The halophosphates are doubly activated phosphors [95], in which Sb^{3+} and Mn^{2+} function as sensitizer and activator, giving rise to two corresponding maxima in the emission spectrum. The antimony acts equally as sensitizer and activator. It transfers part of its excitation energy to manganese. By changing the activator concentration, the blue antimony emission can be adjusted in favor of the orange-yellow manganese emission in such a way that halophosphates can emit white light of varying color temperature.

Halophosphates crystallize in the apatite structure. The chemical composition can be expressed most clearly as $3\,Ca_3(PO_4)_2 \cdot Ca(F, Cl)_2:Sb^{3+}, Mn^{2+}$.

For many years halophosphates were quantitatively by far the most important group of phosphors for low-pressure mercury discharge lamps (fluorescent lamps). The number of studies on the optimization of the quantum yield of halophosphates is correspondingly large. The most important work on the properties and production technology has been reported, for example in [79, 96].

Halophosphates are produced by the solid-state reaction of high-purity $CaHPO_4$, $CaCO_3$, CaF_2, NH_4Cl, $MnCO_3$, and Sb_2O_3. Because of the volatility of antimony trioxide and to avoid the oxidation of Mn^{2+} and Sb^{3+}, the mixture is usually fired under nitrogen at 1150 – 1200 °C in covered crucibles.

The substitution of a small proportion of the calcium by cadmium gives a considerable improvement in the radiation stability and an improved quantum yield for halophosphates in fluorescent lamps. The formation of undesirable color centers is strongly suppressed by addition

of cadmium [97]. The cadmium presumably accelerates the reaction [98], thus minimizing side reactions during the firing process.

Because of the toxicity of cadmium, its use was abandoned in the 1980 s. Nevertheless, by improving the quality of the raw materials and changing the reaction stoichiometry, it was possible to reproduce the quantum yield of the halophosphates, even without the addition of cadmium [99]. The absorption of the exciting UV radiation is strongly dependent on the particle size of halophosphate phosphors and decreases rapidly for particles larger than 3 µm. The particle size is influenced to some extent by the particle size of raw materials (calcium hydrogen phosphate) in the firing mixture. On the other hand mechanical separation processes (sieves) are used to free the fired product from undesired coarse and fine particles.

The following phosphate phosphors have achieved industrial importance: $(Sr,Mg)_3(PO_4)_2$:Sn^{2+}; $LaPO_4$:Ce^{3+}, Tb^{3+}; $Zn_3(PO_4)_2$: Mn^{2+}; $Cd_5Cl(PO_4)_2$:Mn^{2+}; $Sr_3(PO_4)_2 \cdot SrCl_2$: Eu^{2+}; and $Ba_2P_2O_7$:Ti^{4+}. $(Sr,Mg)_3(PO_4)_2$:Sn^{2+} is used in high-pressure mercury lamps. Absorption takes place by Sn^{2+}. The emission band, with a half-band width of 150 nm, has a maximum at 630 nm. Since only the high temperature form ($\geq$ 1300 °C) exhibits good luminescence, it must be cooled rapidly in a reducing atmosphere after firing. The incorporation of small quantities of Al^{3+}, Mg^{2+}, Zn^{2+}, Ca^{2+}, or Cd^{2+} stabilizes the luminescent high-temperature form. In high-pressure mercury lamps, $(Sr,Mg)_3(PO_4)_2$: Sn^{2+} has been replaced by the more effective YVO_4:Eu^{3+}, Tb^{3+}.

The phosphor $LaPO_4$:Ce^{3+}, Tb^{3+} has assumed importance as the green component in trichromatic fluorescent lamps. The emission maximum is at 544 nm and the excitation maximum at 290 nm. It can be produced by solid-state reaction of La_2O_3, Tb_4O_7, $NH_4H_2PO_4$, and CeO_2 or CeF_4 in a reducing atmosphere at 1000 – 1250 °C. Although $LaPO_4$:Ce^{3+}, Tb^{3+} has a high quantum yield comparable with that of CAT (see Section 1.3.3.2), it is insufficiently stable to oxidation in air. Stabilization by large tetravalent cations (Th^{4+}, U^{4+}) [100] and by Sc^{3+} [101] has been claimed. $Zn_3(PO_4)_2$:Mn, which has an emission maximum at 635 nm, is used as a cathode-ray tube phosphor for data displays under the TEPAC number P 27 and WTDS specification RE. It is produced by solid-state reaction between ZnO, $MnCO_3$, and H_3PO_4 in an aqueous slurry. After drying, firing is carried out at 900 – 1150 °C for several hours in an open crucible [102]. $Cd_5Cl(PO_4)_3$:Mn is a yellow-orange-emitting CRT-phosphor for data displays with the WTDS designation LA. It is produced by solid-state reaction at 750 °C after the components $CdSO_4$, CdO, $MnCl_2$ and $NH_4H_2PO_4$ have undergone preliminary reaction in an aqueous medium [103]. 3 $Sr_3(PO_4)_2 \cdot SrCl_2$:$Eu^{2+}$ can be excited by radiation from the entire UV range. The excitation maximum lies at 375 nm and the emission maximum at 447 nm. Upon successive substitution of Sr^{2+} by Ca^{2+} and Ba^{2+}, the emission maximum shifts to 450 nm. In the preparation, $SrCl_2$ is used in excess to improve the crystallization [104]. The phosphor was used as the blue component in trichromatic fluorescent lamps [105] and in the first generation of compact lamps.

Blue-green emitting $Ba_2P_2O_7$:Ti (λ_{max} = 493 nm) is used as the blue component in mixtures with halophosphates for the adjustment of some cool-white light colors. This phosphor is prepared by firing $BaHPO_4$, TiO_2, and BaF_2 at 1050 °C [105–107].

1.3.3.7. Oxides

Industrially, the most important of this group is Y_2O_3:Eu^{3+}. The preparation is generally carried out by precipitating mixed oxalates from purified solutions of yttrium and europium nitrates. Firing the dried oxalates at ca. 600 °C is followed by crystallization firing at 1400 –1500 °C for several hours [108].

Yttrium oxide and europium oxide form a series of mixed crystals without vacancies because of the small difference between their ionic radii [109, 110]. Y_2O_3:Eu^{3+} shows an intense emission line at 611.5 nm in the red region, corresponding to the transition $^5D_0 \rightarrow {}^7F_2$ in europium. The luminescence of this red emission line increases with increasing Eu concentration up to ca. 10 mol %. The concentration in industrial lamp phosphors is ca. 3 mol %. Small traces of Tb enhance the Eu fluorescence of Y_2O_3:Eu^{3+}. Furthermore, this coactivation leads to a shift of the short-wavelength absorp-

tion edge of the host lattice from 220 (Y_2O_3) to 280 nm [111]. This allows excitation with short wavelength UV; excitation by electrons is also possible (Section 1.4.3). This phosphor for cathode-ray tubes is designated as RE (WTDS) and P 56 (TEPAC). Y_2O_3:Eu^{3+} shows no significant decrease in luminescence brightness up to 500 °C.

ZnO:Zn is a typical example of a self-activated phosphor. In the case of zinc oxide, it is an excess of zinc which enables the phosphor to luminesce. The production is carried out by thermal oxidation of crystallized zinc sulfide in air at ca. 400 °C. The green luminescence, with a broad maximum at 505 nm, has a very short decay time of 10^{-6} s. As a phosphor for cathode-ray tubes, ZnO:Zn is classified in the TEPAC list as P 24 and in the WTDS system as GE.

The compound β-Ga_2O_3:Dy, which has photoluminescence properties but no commercial significance, is described in [112].

1.3.3.8. Arsenates

Magnesium arsenate 6 MgO·As_2O_5:Mn^{4+} is a very brilliant red phosphor used in fluorescent lamps. Because of its yellow color, it absorbs the blue mercury lines from which the strong red of the fluorescent lamp light results. The emission maximum is at 655 nm. The phosphor is best produced from solutions of $MgCl_2$ and $MnCl_2$ by precipitation with pyroarsenic acid. The dried precipitate is fired for at least 8 h at 900 °C. A second firing at 1100 °C for 5 h is carried out in the presence of boric acid to improve the crystallization [113].

Magnesium arsenate has lost its industrial importance because of the toxicity of arsenic.

1.3.3.9. Vanadates

Of the vanadates activated with rare earths, YVO_4:Eu^{3+} [114–116] has found the greatest industrial application. Vanadates with other activators (YVO_4 with Tm, Tb, Ho, Er, Dy, Sm, or In; $GdVO_4$:Eu; $LuVO_4$:Eu) have attracted only scientific interest [117–119].

In YVO_4:Eu^{3+}, the europium emission lines at 619 and 614 nm, due to the transition $^5D_0 \rightarrow {}^7D_0$, are not quenched up to 300 °C. Although quenching occurs at higher temperatures, the main lines still retain 50 % of their intensity at 500 °C. Small quantities of terbium activate the red Eu^{3+} lines at higher temperatures [120].

The long-wavelength absorption edge of YVO_4:Eu^{3+} at 340 nm shifts to longer wavelength with rising temperature. The long-wavelength mercury lines of high-pressure mercury lamps can thus be used for excitation. The incorporation of Bi^{3+} sensitizes the Eu^{3+} emission [121] and results in a shift of the luminescence color towards orange, because the emission line at 565 nm, corresponding to the transition $^5D_0 \rightarrow {}^7F_1$, increases in intensity. YVO_4 forms a series of mixed crystals with the isostructural YPO_4. Successive substitution of VO_4^{3-} by PO_4^{3-} shifts the absorption edge of Y(VO_4, PO_4) at 340 nm to shorter wavelengths and increases the temperature at which thermal quenching occurs.

For the production of YVO_4:Eu, the mixed oxides of yttrium and europium are thoroughly ground with an alkali-metal hydroxide as flux and high-purity NH_4VO_3. In a two-stage firing process with intermediate grinding, the mixture is fired at 600 – 650 °C and then at 1100 – 1200 °C for several hours. Excess alkali-metal vanadate is washed out with dilute sodium hydroxide solution. The effect of the production conditions on the particle size of YVO_4:Eu^{3+} has been studied [122].

1.3.3.10. Niobates and Tantalates

The luminescence of rare-earth tantalates is described in [123]. Tantalates crystallize in three different structures, depending on the cation [124]. Yttrium tantalate and yttrium niobate – tantalate have good X-ray absorption and are used in X-ray intensifying screens because of their high conversion factor (Section 1.4.2.). Substitution of tantalum by small quantities of niobium considerably increases the blue fluorescence when excited by X-rays.

The incorporation of traces of terbium strengthens the green emission band for the so-called M' structure of the tantalate, which exists only below 1325 °C in the monoclinic structure (M' phase) [125]. The advantage of the X-ray excitable tantalates lies in the fact that the emis-

sion spectrum extends into the near UV. The films used with X-ray intensifying screens are particularly sensitive in this region.

The tantalates are produced by a solid-state reaction between Ta_2O_5, Nb_2O_5, and Y_2O_3 in a 14-hour preliminary firing at 1000 °C and a crystallization firing with lithium chloride or lithium sulfate as flux.

1.3.3.11. Sulfates

Without exception, the basic lattice of sulfate phosphors absorbs very short wavelength UV radiation. On excitation with X-rays or radiation from radioactive elements, a large proportion of the energy is stored in deep traps. For this reason, $CaSO_4$:Mn is used in solid-state dosimeters. Of the glowpeaks which can be selected by thermoluminescence, more than 50 % fail to appear at room temperature because of a self selection of the shallow traps. Other activators, such as lead or rare-earth ions (Dy^{3+}, Tm^{3+}, Sm^{3+}), stabilize the trapped electrons [126–128].

Photoluminescent sulfates are obtained by activation with ions that absorb short-wavelength radiation, for example, Ce^{3+}. Alkali-metal and alkaline-earth sulfates with Ce^{3+} emit between 300 and 400 nm. On additional manganese activation, the energy absorbed by Ce^{3+} is transferred to manganese with a shift of the emission into the green to red region. Sulfate phosphors can be produced relatively easily. Water-insoluble sulfates are thoroughly mixed with the activators and fired in covered crucibles below the melting point. In the case of activation by Ce^{3+} and Mn^{2+} the activator concentration is at least 0.5 mol % [129].

Water-soluble sulfates are obtained by evaporating a solution of the mixed salts to dryness.

The luminescence properties of sulfates activated by Ce^{3+} and Mn^{2+} are described in detail in [129].

1.3.3.12. Tungstates and Molybdates

Magnesium tungstate $MgWO_4$ and calcium tungstate $CaWO_4$ are commercially the most important self-activated phosphors. Magnesium tungstate is often used as a reference material for UV-excited phosphors because of its high quantum yield of 84 % for the conversion of the 50–270-nm radiation into visible light; it is cited as standard phosphor 1027 for quantum output by the NBS [130]. The temperature dependence is relatively low; the maximum yield is at 240 K, but is still as high as 50 % at 400 K.

The emission maximum after excitation in a fluorescent lamp is at 482 nm with a half-intensity width of 140 nm. Magnesium tungstate crystallizes in the wolframite structure [131]. On additional activation with rare-earth ions their typical emission also occurs. Calcium tungstate crystallizes with the scheelite structure. Magnesium tungstate is produced by heating an excess of MgO or basic magnesium carbonate with high-purity WO_3 or H_2WO_4 at 1000–1100 °C. Excess MgO is washed out with 10 % hydrochloric acid.

$MgWO_4$ is used in blends with other phosphors in high-voltage advertisement illumination tubes and as a fluorescent material in TLC plates. Because of its high effective absorption of X-rays, $CaWO_4$ is used in large quantities for the production of X-ray intensifying screens in combination with blue-sensitive films. $CaWO_4$:Pb is still used as the blue phosphor in fluorescent lamps.

The production of self-activated $CaWO_4$ is more difficult than that of $MgWO_4$. Small deviations from stoichiometry in the composition have a marked effect on the efficiency. Also, traces of Pb^{2+} shift the emission peak from 415 nm to 439 nm (0.0006 mol % Pb^{2+}). To bind excess CaO in the production process, a small proportion of sulfate is used. The precipitated $CaWO_4$ is washed until chloride-free, dried, and subjected to crystallization firing at 800 °C with 30 % flux. The formation of luminescent $CaWO_4$ from the precipitate is described in [132].

Self-activated molybdate phosphorus are described in [133]. The energy transfer that occurs in molybdates activated with Eu^{3+} is described in [114, 134].

1.3.4. Halide Phosphors

The longest-known luminescent halide is calcium fluoride. The term fluorescence is derived from the naturally occurring luminescent mineral fluorspar.

Table 1. Alkali-metal halide phosphors

Host crystal	Activator		ϱ, g/cm^3	*mp*, °C	Emission peak, nm	Use ∗
	Type	Conc., mol %				
LiI	Eu	$5-6\times10^{-2}$	3.49	449	440	S
NaI	Tl	10^{-1}	3.67	651	410	S
CsI	Tl	10^{-1}	4.51	624	565 – 600	S
CsI	Na	$4-6\times10^{-4}$	4.51	624	430	S
LiF	Mg	$10^{-8}-10^{-2}$	2.63	842	400	TLD
LiF	Mg, Ti	$10^{-3}-10^{-2}$	2.63	842	400	TLD
LiF	Mg, Na	$10^{-3}-10^{-2}$	2.63	842	400	TLD

∗ S = scintillator (see Section 1.4.7); TLD = thermoluminescent dosimeter (see Section 1.4.7).

Alkali-Metal Halides. Luminescent alkali-metal halides can be produced easily in high-purity and as large single crystals. They are therefore often used as model substances for the investigation of luminescence processes. The luminescence processes of the alkali-metal halides can be divided into: 1) the self-luminescence of the undoped crystals, 2) luminescence by lattice defects, and 3) sensitized luminescence.

Self-Luminescence. The action of UV light or ionizing radiation on pure alkali-metal halide crystals causes luminescence which is very intense, particularly at low temperature. The emission spectrum is characteristic for each individual compound. This fluorescence is comparable with the recombination luminescence which occurs upon capture of an electron by a V_K center (defect electron).

Luminescence of Lattice Defects. Many defect centers are known in the case of the alkali-metal halides, which are derived from electrons in anion vacancies (F centers, or color centers). Association of two or more F-centers gives new defect centers, which can each also take up an electron. These lattice defects act as luminescence centers, the emission spectra of which sometimes exhibit a large number of lines.

At certain temperatures phosphorescence can also be observed, which indicates strong relaxation in the excited state.

As well as the F-centers and F-aggregates (M, R, and N centers), there are also various other anion and cation defects which act as luminescence centers.

Foreign-Ion Luminescence. Through the incorporation of foreign ions (e.g., Tl^+, Ga^+, In^+) into the crystal lattice, further luminescence centers are formed. The emission spectra are characteristic for the individual foreign ions. The complicated luminescence mechanism is described in [135].

The alkali-metal halide phosphors are produced by firing the corresponding alkali-metal halide and the activator in platinum or fused-silica crucibles under an inert atmosphere.

Some industrially important alkali-metal halide phosphors are listed in Table 1.

NaI:Tl and CsI:Tl are used as detectors for X- and γ-rays because of their high absorption capacity for these types of radiation. Europium-activated lithium iodide is preferred for neutron detection because of its favorable capture cross section. The phosphor CsI:Na is particularly important industrially as an X-ray converter in optoelectronic image intensifiers (Section 1.4.2). It has a high effective quantum absorption for X-rays of energy 20 – 80 keV, which are used in medicine. Since the vapor pressure curves for CsI and NaI differ only slightly, it is possible to produce luminescent screens with good resolution by evaporation. The emission spectrum of CsI:Na also corresponds well to the spectral sensitivity of the photocathodes usually used for image intensifiers.

Alkaline-Earth Halides. Of the alkaline-earth halide phosphors, those doped with manganese or rare earths have been used industrially (Section 1.4.7) (e.g., CaF_2:Mn; CaF_2:Dy).

On excitation with high-energy radiation these compounds store part of the energy, which can subsequently be "called up" by stimulation. If the stimulation is achieved by heating, the resulting luminescence is termed thermoluminescence.

Thermoluminescent phosphors are used industrially in dosimeters. They are produced by coprecipitation of CaF_2 and an activator from a solution of the corresponding cations, or by the reaction of oxides or carbonates of the cations with hydrofluoric acid followed by firing. Moldings with high transparency for use in dosimeters can be produced by high-temperature pressing.

A further industrially important class of phosphors are the alkaline-earth fluorochlorides. The compound $Ba_{0.95}Eu_{0.05}FCl$ exhibits strong emission, with a peak at 375 nm, on X-ray excitation [136]. Substitution of Cl^- by Br^- shifts the peak to 385 nm. The Eu^{3+} acts as the luminescence center. The production is carried out by grinding BaF_2 with $BaCl_2$ or $BaBr_2$ in the molar ratio 1:1 with 0.04 – 0.07 mol % of a europium trihalide. After drying, the mixture is fired in a reducing atmosphere.

The Ba^{2+} can be completely or partially substituted by Sr^{2+} or Ca^{2+}, and Cl^-, by Br^- or I^-. Information on this and other production processes is given in [137–139]. These phosphors are used in X-ray intensifying screens (Section 1.4.2).

The $BaFBr:Eu^{2+}$ phosphors are gaining increasing industrial importance because of their storage effect. On irradiation with X-rays (or UV light) some of the Eu^{2+} ions are ionized to Eu^{3+} with loss of electrons to the conduction band. These electrons are captured by F^+ centers (anion vacancies) and can be liberated from these traps by the action of light (stimulation). The electrons return to the Eu^{3+} ions via the conduction band, converting them to Eu^{2+} [140–144].

The production of such storage phosphors is carried out by the method described for $BaFCl:Eu^{2+}$. These phosphors are used for storage panels in X-ray diagnostics (see Section 1.4.2).

Other Halide Phosphors. Information on the most important manganese-activated phosphors is summarized in Table 2. These fluorides are produced by the method given below for $(Zn, Mg)F_2:Mn^{2+}$.

MgF_2 and $ZnSO_4$ are made into a paste in the desired molar ratio and a solution of the activator is added. After drying, the mixture is fired in quartz trays above 1000 °C. These orange-emitting phosphors stand out because of their long afterglow time and are therefore still used in special radar tubes and oscilloscopes [7, 8, 145], despite their low stability towards burn-in compared with other cathode-ray phosphors.

Oxyhalides. The oxyhalides of yttrium, lanthanum, and gadolinium are good host lattices for activation with other rare-earth ions such as terbium, cerium, and thulium. The use of $LaOCl:Tb^{3+}$ as the green component in projection-television tubes has been discussed [146]. $LaOBr:Tb^{3+}$ and $LaOBr:Tm^{3+}$ exhibit high X-ray absorption, and, because of the good agreement between the emission spectrum and the spectral sensitivity of the X-ray films, they are used in X-ray intensifying screens [147].

The activator concentration (Tb, Tm) is 0.01 – 0.15 mol %. The phosphors are produced by mixing the corresponding oxides with an excess of ammonium bromide and subsequently firing at 400 – 500 °C. The reaction product is then subjected to crystallization firing with an alkali-metal bromide at 600 – 1100 °C [148]. The oxybromides are also formed by firing the oxides in a hydrogen bromide atmosphere. By coactivation, with ytterbium, the afterglow, which causes problems in medical applications, can be reduced. Partial substitution of lanthanum by gadolinium in $LaOBr:Ce^{3+}$ leads to an increase in the quantum yield upon electron excitation and an increase in the quenching temperature [149].

Table 2. Manganese-activated halide phosphors

Composition	WTDS Specification	Color location		Decay
		x	y	
$(Zn, Mg)F_2:Mn^{2+}$	LB	0.557	0.442	long
$KMg\ F_3:Mn^{2+}$	LC	0.573	0.426	very long
$MgF_2:Mn^{2+}$	LD	0.559	0.440	very long
$(Zn, Mg)F_2:Mn^{2+}$	LK	0.591	0.407	very long

1.4. Uses of Luminescent Pigments

1.4.1. Lighting

For a detailed description of the lamps discussed in this Section, see → Lamps.

Most inorganic phosphors are used in mercury vapor lamps. In these light sources, a plasma is produced from an inert gas and mercury vapor by electrical discharge. The plasma emits intense UV radiation, whose wavelength depends on the vapor pressure. The conversion of the UV radiation into visible light is effected by phosphors applied as a thin coating on the inside of the glass of the lamp. The numerous phosphors with various emission colors in industrial use make it possible to produce practically any desired light color or wavelength in the spectrum by means of the UV radiation from the gas discharge. Depending on the mode of operation of the gas discharge, there are two types of mercury discharge: low pressure and high pressure. The phosphor has a different function in the two types of discharge.

Fluorescent Lamps (Low-Pressure Mercury Lamps). In fluorescent lamps the phosphor is excited by energy-rich mercury resonance lines at 254 and 185 nm. For many years the most important phosphors were the halophosphates. Not only were fluorescent lamps for white light colors produced with these, but the so-called color-improved Deluxe Lamps were produced by means of mixtures with red phosphors. However, the maintenance of these halophosphate phosphors is not good. The quantum yield decreases by almost a third due to the radiation-induced decomposition of the halophosphate phosphor even after only 10 000 h of use. By lowering the diameter of the fluorescent tube from 38 to 26 mm, the specific radiation loading is also increased. In the search for more efficient phosphors with greater radiation-stability, blends of $Y_2O_3:Eu^{3+}$, $Ce_{0.65}Tb_{0.35}MgAl_{11}O_{19}$, and $BaMg_2Al_{16}O_{27}:Eu^{2+}$ were used from the 1970 s. With these trichromatic blends the quantum yield could not only be considerably improved but the color rendition of the fluorescent lamps became comparable to those of the earlier Deluxe fluorescent lamps because of the high proportion of red radiation.

High-Pressure Mercury Lamps. Phosphors for high-pressure mercury vapor lamps have to fill up the red part of the spectrum of these lamps, which otherwise emit intense radiation in the blue and green regions.

The lamp can reach an operating temperature of 150 – 300 °C, depending on the power rating. The phosphors must therefore have a very high quenching temperature. In addition they must be readily excitable in the long-wavelength UV. Many phosphors do not meet these criteria for use in high-pressure lamps. 3.5 $MgO{\cdot}0.5\ MgF_2{\cdot}GeO_2:Mn^{4+}$ and $YVO_4:Eu^{3+}$ have proved to be suitable. The excitation spectrum shifts towards longer wavelengths with increasing temperature for $YVO_4:Eu^{3+}$ and the quenching temperature is high. The magnesium fluorogermanate radiates in the deep red region; however, itis not as temperature stable as the vanadate.

UV Lamps. Special applications of UV phosphors include fluorescent lamps for tanning (sun studios), photocopier lamps, and lamps for photochemistry, medicine, and varnish harden-ing. Phosphors which are suitable here are $Ce(Mg,Ba)Al_{11}O_{19}$ ($\lambda_{max.} = 344$ nm), $BaSi_2O_5:Pb^{2+}$ (351 nm), $2\,SrO{\cdot}6\,B_2O_3{\cdot}SrF_2:Eu^{2+}$ (371 nm), $(Sr, Mg)_2P_2O_7:Eu^{2+}$ (392 nm), and $(Ba, Zn, Mg)_3Si_2O_7:Pb^{2+}$ (303 nm).

High-Voltage Advertisement Illumination Tubes. High voltage tubes with cold cathodes are widely distributed in illuminated advertisements in the form of letters, title strips, illuminated motifs, for lighting effects at exhibitions and show windows, and for façade illumination of department stores. The light colors of such tubes can be adjusted by colored glass, the filling gas (neon gives red; helium, white-pink; argon–mercury, blue), and varied precisely by the phosphors (Table 3, see next page) [150, 151].

1.4.2. X-ray Technology

By means of phosphors, the invisible shadow image which is formed on penetration of an object by X-rays is converted into a visible image which can then be examined. The phosphors

Table 3. Phosphors for high-voltage advertisement illumination tubes

Composition	Emission color
$Zn_2SiO_4:Mn^{2+}$	green
$CaWO_4:Pb^{2+}$	blue
$CaWO_4:Sm^{3+}$	pink
$6\,MgO{\cdot}As_2O_5:Mn^{4+}$	red
Halophosphates	all white-light colors
$Y_2O_3:Eu^{3+}$	red
$3\,Sr_3(PO_4)_2{\cdot}CaCl_2:Eu^{2+}$	blue
$3{,}5\,MgO.0{,}5\,MgF_2{\cdot}GeO_2:Mn^{4+}$	red
$MgWO_4$	blue
$Ba_2P_2O_7:Ti^{4+}$	blue-green
$Ce_{0.65}Tb_{0.35}MgAl_{11}O_{19}$	green
$CaSiO_3:Mn^{2+}, Pb^{2+}$	pink

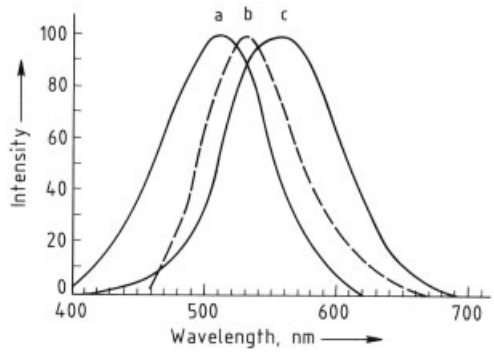

Figure 10. a) Sensitivity of human eye (night); b) Emission spectrum of (Zn, Cd)S:Ag fluoroscopic screen; c) Sensitivity of human eye (day)

must have a high effective quantum absorption for the energy-rich X-rays (20 – 60 keV). Thus, compounds of elements with high atomic numbers are principally used. The emission spectrum should correspond to the spectral sensitivity of the detector as well as possible. In the particle size distribution, a compromise between high brightness (coarse particles) and good image-transfer properties (fine particles) must be sought. The afterglow should be so low that on taking pictures of moving parts the image is not blurred. Different quality requirements exist for the various X-ray techniques. Therefore phosphors are used which are adapted to suit the particular application.

Fluoroscopic screens allow direct examination of X-ray images by the eye. Today a (Zn, Cd) S:Ag is used almost exclusively in which the molar ratio of ZnS:CdS is 60:40. The particle size lies between 20 and 40 µm. The emission spectrum corresponds well to the spectral sensitivity of the eye (Fig. 10).

The fluoroscopic screen consists of a substrate (cardboard, plastic sheet), which is first covered with a reflective layer of MgO or TiO_2. The phosphor layer is applied by pouring, after dispersion in a binding agent based on acetylcellulose. The density of the phosphor coating is 70 – 140 mg/cm^2. Gd_2O_2S:Tb is also used for the conversion of energy-rich X-rays (80 keV), for example, in luggage checking machines.

Intensifying Screens. Since X-ray films can only use 1 % of the incident X-rays for nucleation because of their low absorption, intensifying screens for X-ray photons are used in medical diagnosis for taking X-ray photographs. Here the X-rays are first converted into light which then irradiates the film. In most cases, two intensifying screens are used, the phosphor screen placed in front of the film (front screen) very often having a thinner coating than the rear screen.

Until recently, luminescent pigments with blue emission in combination with blue-sensitive films were used exclusively. Since the introduction of green-sensitive film material into X-ray technology, green-emitting phosphors are now being increasingly used. The most important phosphors are listed in Table 4.

Intensifying screens consist of cardboard or plastic sheets (e.g., polyester) as substrates. A reflective layer (e.g., TiO_2) is first applied to this and then the phosphor-binder layer. Gelatine, nitrocellulose, or polymethylmethacrylate are, for example, used as binders. The density of the phosphor coating is 10 – 100 mg/cm^2.

Storage panels (Fig. 11) are, in principle, constructed like intensifying screens. However, they differ in that a phosphor is used that, on excitation with X-rays, stores part of the absorbed energy in addition to producing spontaneous emission. The stored energy can later be read out by stimulation with light of a suitable wavelength, for example, with a finely focused laser beam. The possibility thus exists for computer digital radiography, in which information stored simultaneously (in parallel) is then read out serially [153]. Objects with large differences in contrast can be represented without loss with this technique.

Table 4. Phosphors for intensifying screens [124, 152].

X-ray phosphor	Absorption at 80 kV, %	Conversion into light, %	Density, g/cm^3	Emission maximum, nm
$CaWO_4$	26.7	4	6.8	425
$Y_2O_2S:Tb^{3+}$	27.0	18	5.0	418, 545
$Gd_2O_2S:Tb^{3+}$	37.7	15	7.44	545
$BaFCl:Eu^{2+}$	39.7	12	4.56	385
$LaOBr:Tb^{3+}$	41.5	13	6.28	437
$YTaO_4:Nb^{5+}$ ca.	41.0	–	7.56	356, 400

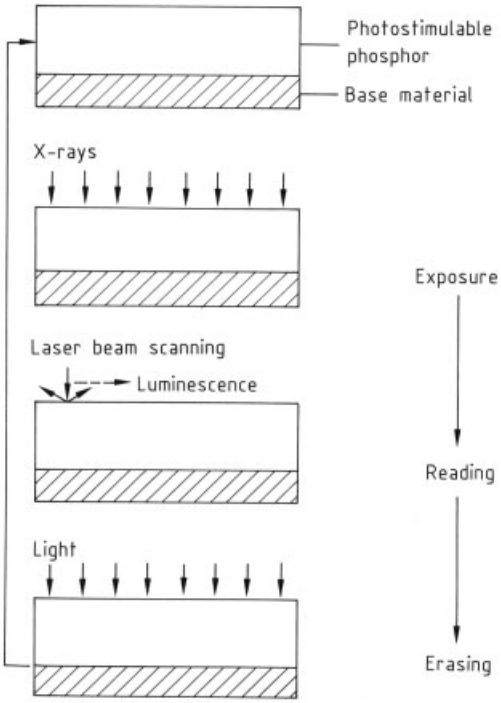

Figure 11. Storage panel – principle of exposure, reading, and erasure

$Y_2O_2S:Eu^{3+}$, Bi^{3+} ; $GdOBr:Sm^{3+}$; $LaOBr:Tb^{3+}$, Bi^{3+} ; and the barium fluorohalides are among the known storage phosphors. At present, $BaFBr:Eu^{2+}$ with various coactivators, mostly transition metals, is principally used in industry.

X-Ray Image Intensifiers. The use of an optoelectronic image intensifier is more efficient than direct examination with a fluoroscopic screen [154]. Further development of this principle led to the optoelectronic X-ray image intensifier (Fig. 12), [155], in which the X-ray phosphor screen (input screen) is in direct optical contact with a photocathode that converts the luminance distribution of the X-ray screen into an electron-density distribution. The electrons liberated are accelerated in an electric field between the photocathode and an anode (20 – 30 kV) and are focused by electron lenses onto a further phosphor screen (output screen), where conversion of the electron image to a visible image takes place. The two phosphor screens in the X-ray image intensifier differ in their construction, corresponding to their differing functions.

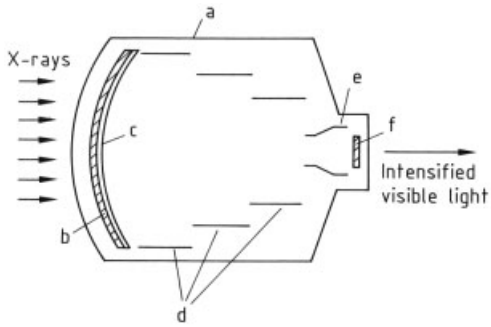

Figure 12. X-ray image intensifier
a) Glass envelope; b) Input conversion screen; c) Photocathode; d) Focusing electrodes; e) Anode; f) Output screen

Formerly, $(Zn, Cd)S:Ag^+$ was used for the *input screen* of these X-ray image intensifiers. This phosphor was sedimented onto the base material of the input screen (glass or metal sheet) from a suspension. Since ca. 1970, the input screens have mostly been made by condensing $CsI:Na^+$ onto thin aluminum sheets in vacuum. Cesium iodide has a higher mass extinction coefficient than (Zn, Cd)S. The evaporation coating process also gives a considerably higher packing density, which leads to higher resolution at coating densities of up to 120 mg/cm^2. The emission spectrum of $CsI:Na^+$ corresponds well with the sensitivity of the Cs_3Sb photocathode.

The *output screen* consists of a glass base material (sometimes also fiber optics), onto which fine particles of $(Zn, Cd)S:Ag^+$ are applied, usu-

Table 5. Phosphors for cathode-ray tubes

Composition	WTDS Specification	Emission color	CIE color coordinates x	CIE color coordinates y	Decay time
Oscilloscope tubes					
$ZnS:Ag^+$	BE	blue	0.136	0.148	medium
$ZnS:Cu^+$	GH	green	0.226	0.528	medium
$Zn_2SiO_4:Mn^{2+}$	GJ	yellow-green	0.218	0.712	medium
$ZnS:Cu^+$	GL	yellow-green	0.279	0.536	medium short
$(Ca,Mg)SiO_3:Ce^{3+}$	GJ	yellow-green	0.280	0.530	medium
$Zn_2SiO_4:Mn^{2+}, As^{3+}$	GR	yellow-green	0.223	0.668	medium short
$Gd_2O_2S:Tb^{3+}$	GY	yellow-green	0.333	0.556	medium
$Y_2O_2S:Tb^{3+}$	WB	white	0.253	0.312	medium
$KMgF_3:Mn^{2+}$	LC	orange	0.573	0.426	very long
$ZnS:Ag^+$	GB	blue	0.151	0.032	medium short
$(Zn, Cd)S:Cu^+$	GB	yellow-green	0.357	0.537	very long
$(Zn, Cd)S:Cu^+$	KB	fluorescence, blue-green phosphorescence, green	0.380	0.520	very long
Black-and-white television					
$ZnS:Ag^+ + (Zn, Cd)S:Ag^+$	WW	white	0.265	0.285	medium short
$ZnS:Ag^+ + (Zn, Cd)S:Cu^+$	WW	white	0.275	0.311	medium short
Projection tubes					
$Zn_2SiO_4:Mn^{2+}$	GJ	green	0.207	0.713	medium
$(Ca, Mg)SiO_3:Ti^{4+}$	BF	blue	0.164	0.108	medium
$Y_2O_3:Eu^{3+}$	RF	red	0.640	0.335	medium
Color television					
$(Zn, Cd)S:Cu^+$	X	green	0.345	0.598	medium
$ZnS:Cu^+$	X	green	0.290	0.611	medium
$ZnS:Ag^+$	X	blue	0.136	0.070	medium
$Y_2O_2:Eu^{3+}$	RF	red	0.640	0.335	medium
$Y_2O_2S:Eu^{3+}$	X	red	0.630	0.340	medium
Monitor tubes					
$ZnS : Cu^+ +$ $Zn_2SiO_4:Mn^{2+}, As^{3+}$	GW	yellow-green	0.238	0.568	medium to long
$(Zn, Mg)F_2:Mn^{2+} +$ $(Ca, Mg)SiO_3:Ce^{3+}$	YD	yellow-orange	0.541	0.456	long
$ZnS:Ag^+ + (Zn, Cd)S:Cu^+$	YC	white	0.330	0.330	medium to long
$(Zn, Cd)S:Cu^+$	KC	yellow	0.423	0.553	medium to long
$Y_2Si_2O_5:Ce^{3+}, Tb^{3+}$		white			very short
Penetration tubes					
$(Zn, Cd)S:Ag^+, Ni^{2+} +$	VB	green (15 kV)	0.398	0.546	medium short
$Y_2O_3:Eu^{3+}$		red (8 kV)	0.655	0.340	medium
$(Zn, Cd)S:Ag^+,Ni^{2+} +$	VC	green (12 kV)	0.414	0.515	medium short
$Y_2O_3:Eu^{3+}$		red (6 kV)	0.672	0.325	medium
$YVO_4:Eu^{3+} +$	VE	yellow-green (17 kV)	0.355	0.510	very long
$ZnS:Cu^+$		red-orange (8 kV)	0.604	0.360	medium
Radar tubes					
$(Zn, Mg)F_2:Mn^{2+}$	LK	orange	0.591	0.407	very long
$(Ca, Mg)SiO_3:Ce^{3+}$	GV	yellow-green	0.280	0.530	very long
$ZnS:Cu^+$	GF	green	0.240	0.520	long
$Zn_2SiO_4:Mn^{2+}, As^{3+}$	GR	yellow-green	0.223	0.698	long

ally by sedimentation, but also by brushing or spraying. Because of the optoelectronic diminution of the image (10:1 to 15:1) onto the relatively small output screen (20 – 40 mm diameter), the phosphor layer must have high resolution. Phosphors with an average particle size of 1.5 – 3 µm and a narrow particle size distribution are used. The thickness of the layer is ca. 5 – 10 µm, which corresponds to a coating density of 0.8 –1.6 mg/cm^2 of $(Zn, Cd)S:Ag^+$.

Computer Tomography. In computer tomography the patient to be examined lies within a circle of detectors (up to 1200) [156]. A con-

tinually irradiating X-ray tube rotates around the patient. The detectors measure the intensity of the X-rays which have passed through a particular layer of the body in each direction.

The detectors consist of a phosphor crystal which is combined with a photodiode. The electrical signals of the individual detectors are combined by computer to form an image.

In modern instruments, single crystals of $CdWO_4$ and $Bi_4Ge_3O_{12}$ ($\lambda_{max} = 480$ nm) are used, which achieve an absorption of up to 98 % in the short length required by the construction and have a high quantum yield. A fast decay is important for achieving a high signal-to-noise ratio [157].

1.4.3. Cathode-Ray Tubes

Phosphors are an important, quality-determining component of cathode-ray tubes (Fig. 13). In the phosphor screen, modulated electrons are converted into a visible image. Many methods exist for the production of phosphor screens [7]. Monochromatic tubes are predominantly produced by the sedimentation (settling) process. A suspension of the phosphor in alcohol or water, with addition of water glass solution, is placed in the tube and is deposited by sedimentation on the inner surface of the faceplate. There it forms a regular layer whose adhesion, both to itself and to the glass, is effected by the slowly precipitating silicic acid. The coagulation time of the water glass is adapted to the sedimentation rate of the phosphor by addition of electrolytes.

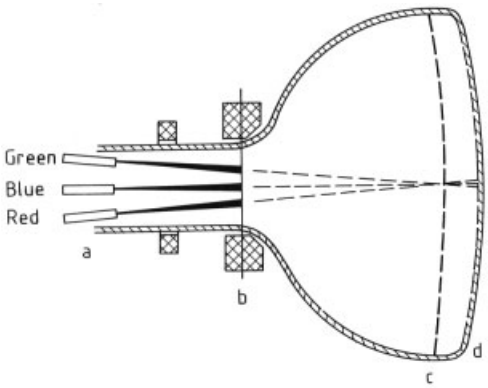

Figure 13. Schematic of a cathode-ray tube
a) Electron guns; b) Optoelectronic system; c) Shadow mask; d) Phosphor screen

The coating of color picture tubes is carried out in such a way that the three components red, green, and blue are applied consecutively as dispersions in a photosensitive resin. This is then exposed via a shadow mask so that the resin on the exposed areas becomes water-insoluble. The phosphor on the unexposed areas is then washed off with water.

For special tubes, other methods, such as electrophoresis or brushing, are also sometimes used. In most cases the phosphor layer is then subjected to an aluminum vapor deposition process. This causes an increase in the brightness by reflection of the backward-radiating luminescence and inhibits charge build-up in the phosphor layer.

Cathode-ray tubes have a very large number of applications which require varying phosphor qualities. Table 5 (see previous page) lists the most important cathode-ray luminescent pigments.

Since April 1982 the Worldwide Phosphor Type Designation System (WTDS) has existed, in which the optical data of the various phosphors for cathode-ray tubes are laid down [158]. It replaces several phosphor designation systems (e.g., TEPAC, Pro Electron) previously in use in various countries. As well as the chemical composition, WTDS records the fluorescence spectrum, the phosphorescence spectrum under the various usual operating conditions, the color coordinates x and y in the standard coordinate chromaticity diagram of the International Commission on Illumination [ICI or CIE (Commision International de l'Eclairage) system, 1931] [159], and the decay time. The phosphors are characterized by two capital letters, the first giving the position of the emission color in the Kelly Charts of Color Designation:

A: purple-red, purple, violet,
B: blue, greenish blue,
G: blue-green, green, yellow-green,
K: yellow-green
L: orange, yellow-orange,
R: red-orange, red, purple-red, pink, purple-pink,
W: white,
Y: green-yellow, yellow, orange-yellow.

The second letter is assigned to a particular product.

Television and Monitor Tubes. The main requirements for phosphors in television tubes

(Table 5) concern brightness, decay time, and color. While the brightness should generally be as high as possible, the decay time must be adapted to the image frequency to decrease flicker and poor image definition. The white luminescence required for black and white televisions is achieved with blends of $ZnS:Ag^+$ and $(Zn, Cd)S:Cu^+$, which give complementary blue and yellow emission. One-component phosphors which emit white light due to double activation have not been successful [59, 60].

In color picture tubes, the final color is formed by additive mixing of red, green, and blue light. Each of the three phosphors is excited by its own modulated electron beam and emits light of intensity corresponding to the particular degree of excitation. $(Zn, Cd)S:Ag^+$ and $(Zn, Cd)S:Cu^+$, Al^{3+} are used as the green-emitting component. Because of the toxicity of cadmium, however, they are now being replaced by $ZnS:Cu^+$, Al^{3+} and $ZnS:Cu^+$, Au^+, Al^{3+}. The phosphor $ZnS:Ag^+$ is exclusively used as the blue component. The most important red component is $Y_2O_2S:Eu^{3+}$ with a Eu concentration of 3 mol %; $Y_2O_3:Eu^{3+}$ is now less widely used for this purpose. The phosphors for color picture tubes have an average particle size of 8 –10 µm.

The requirements of phosphors for monitor tubes differ from those for entertainment tubes. Because of the high resolution required for monitor tubes, phosphors with smaller particle sizes are often needed. They are between 6 and 4 µm, for special applications even lower. For monitor tubes which reproduce slow movement only, phosphors with longer decay times are used to diminish image flickering. In $ZnS:Cu^+ + Zn_2SiO_4:Mn^{2+}$, As^{3+} and $(Ca, Mg)SiO_3: Ce^{3+} + (Zn, Mg)F_2:Mn^{2+}$, the decay times of the blends are prolonged by the second component (see Table 5). For the reproduction of faster movement, phosphors with shorter decay times are used, for example, $Y_2Si_2O_5:Ce^{3+}$, Tb^{3+} (white). For monochrome monitors with amber as the image color $Cd_5(PO_4)_3Cl:Mn^{2+}$ or a blend of $Y_2O_2S:Eu^{3+}$ and $(Zn, Cd)S:Cu^+$ is used.

The majority of phosphors for cathode-ray tubes are coated with an oxide, silicate, or phosphate before use to improve their processing properties and stability to burn-in. Pigmenting of $Y_2O_2S:Eu^{3+}$ with finely-divided Fe_2O_3 and of $ZnS:Ag^+$ with ultramarine or cobalt aluminate is also known. The incorporated pigment, which has the same body-color as the emission color of the phosphor, absorbs incident ambient light, effecting an increase in contrast [160].

A special case of pigmenting is the coating of one phosphor with another. This type of phosphor (onion-skin) is used in so-called penetration tubes. At lower acceleration voltages the luminescence of the outer phosphor dominates, and at higher voltages, that of the inner phosphor. In addition, phosphor blends are used, the components of which exhibit differing brightness–voltage dependencies. By using phosphors with differing decay times or emission colors, monitor tubes with differing afterglow times (variable persistence penetration CRT) or colors (multicolor penetration CRT) can be realized. Compared to three-electron-beam tubes, penetration tubes stand out because of their higher resolution and lower susceptibility to interference.

Oscilloscopes are electrical measuring instruments for the display and monitoring of rapidly changing electrical processes, for example, changes in voltage and current with time. An important part of these instruments is the display tube (Fig. 14). In this cathode-ray tube, a thermionic cathode emits an electron beam which is directed by the modulator and passes through two pairs of deflector plates placed at right angles to each other and then excites a phosphor screen. The operating voltage is 1 –15 kV. Because the energy of the electrons is low, phosphor screens without aluminum backing can also be used. The luminescent pigments are listed in Table 5. They must have a good efficiency for the particular electron energy. For tubes for the detection of fast processes, a short decay time is necessary. On the other hand, for the display of slow or periodic processes with low frequencies, phosphors with very long afterglows are used. The choice of emission color depends on the spectral sensitivity of the particular detector. For visual examination, green-emitting products are used, and for photographic examination, blue-emitting ones. The phosphor screens are mostly produced by sedimentation with water glass as the binding agent. A centrifugal process and cataphoresis are also used. For an operating voltage of 1 – 10 kV, a coating density of 2 – 8 mg/cm^2 is used.

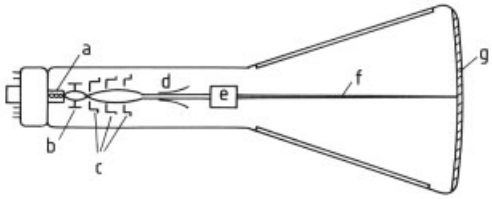

Figure 14. Oscilloscope tube
a) Cathode; b) Modulator; c) Electron lens; d) Y-plates; e) X-plates; f) Electron beam; g) Phosphor screen

Radar Tubes. In construction, a radar tube corresponds to the oscilloscope tube. It displays radar echos, i.e., the direction of and the transit time to and from a reflecting object.

Since the pulse duration is of the order of microseconds, the response time of the phosphor must be correspondingly short. On the other hand, the afterglow of the screen should be long enough for the radar image to be regenerated after one rotation of the radar antenna (of the order of seconds). These requirements are achieved by means of a combination of phosphors. In the *cascade screen*, two superimposed phosphor layers are used. The first consists of a blue, cathodoluminescent, fast-responding phosphor. The blue light of this phosphor excites the photoluminescent phosphor with yellow afterglow in the second layer; $ZnS:Ag^+$ (blue) and $(Zn, Cd)S:Cu^+$ (yellow) are an example of such a cascade. For the observation of moving objects, an even longer afterglow (of the order of minutes) is often desirable to obtain an impression of the direction and speed of the object from the direction and length of the afterglow trace. Suitable phosphors are listed in Table 5.

In *penetration tubes*, two phosphor layers, separated by a barrier layer, are applied to the phosphor screen. At lower acceleration voltages, only the first phosphor is excited and luminesces with its characteristic color and afterglow time. At higher voltages both phosphors luminesce. Since most electrons are then absorbed in the second layer, this determines the color and afterglow time. It is thus possible with only one electron beam to display simultaneously the inscription in one color with short afterglow and the radar signal in another with long afterglow by changing the voltage, as can be seen in the case of the phosphor VE in Table 5.

Flying-Spot Tubes. A rapidly moving spot of light (flying spot), obtained by focusing an electron beam onto the phosphor screen of a cathode-ray tube scans a photo or a film. The light which passes through or is reflected is then converted into signals in a photodiode behind or in front of the object, respectively. A short decay time (< 100 ns) and high spot brightness are important for fast scanning. In monochrome scanning tubes, the ZnO:Zn used initially has been replaced by $(Zn, Cd)S:Ag^+, Ni^{2+}$. In color scanning, blends of $Y_3Al_5O_{12}:Ce^{3+}$ with $Ca_2Al_2SiO_7:Ce^{3+}$ or $SrGa_2S_4:Ce^{3+}$ with $SrGa_2S_4:Pb^{2+}$ can be employed.

Optoelectronic Image Converter. The image converter is an optoelectronic system which converts an image invisible to the human eye first into an electron image and the latter into a visible one [161]. The image converter extends human vision into UV, IR, and X-ray regions. Figure 15 shows the principle of the converter. Light emitted from an object releases electrons in the photocathode in proportion to its intensity. The electrons are accelerated in the electron-optical system and focused onto a phosphor screen. Here the electron image is converted to an image that can be examined with the eye or further processed by means of other detectors such as film or television cameras.

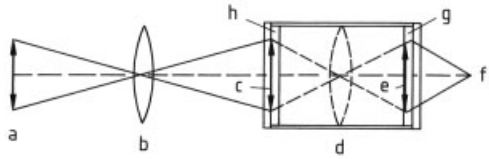

Figure 15. Optoelectronic image converter and intensifier, principle
a) Object; b) Imaging objective; c) Optical image; d) Optoelectronic imaging system; e) Electron image; f) Detector (eye or film); g) Phosphor screen; h) Photocathode

The quality of the system is determined to a great extent by the phosphor screen and the phosphor used. The phosphor should have as high a quantum yield as possible on excitation with 6 – 20 keV electrons and must be resistant to the alkali-metal vapors present during the production of the photocathode and to ion bombardment. The emission spectrum should be adapted to the spectral sensitivity of the particular detector. The requirements for high resolution and

Table 6. Phosphors for image converters

Phosphor	Typical particle-size distribution, µm			Emission peak, nm
	5 %≤	50 %≤	95 %≤	
(Zn, Cd)S:Ag^+	1.3	2.3	3.8	532
ZnS:Cu^+	1.3	2.5	5.0	525
Zn_2SiO_4:Mn^{2+}	1.5	3.5	9.0	526
Gd_2O_2S:Tb^{3+}	2.0	3.5	9.0	547
ZnS:Ag^+	1.1	2.5	6.0	446
CdS:Cu^+	2.6	9.4	21.0	845

good contrast lead to particular requirements for the particle size and particle-size distribution of the phosphor. The average particle size should not exceed 4 µm. The decay time should correspond to the classification medium to medium short. For special applications, luminescent pigments with the classification short and very short are required. Some important phosphors for image converters are listed in Table 6.

1.4.4. Product Coding

Phosphors are used extensively for coding of mass-produced goods or high value trademarked articles. For mass-produced goods, they serve to distinguish between particular types or, at a later stage, for the control of production batches. In the case of trade marked articles, coding with phosphors allows rapid recognition of imitations in the fight against product piracy. The phosphor is applied by means of transparent or colorless printing inks in the form of bar codes, numbers, or company emblems. Fluorescent fibers for packaging paper or fluorescent stick-on labels are further forms of coding. Phosphors are preferred that have clearly defined and narrow-banded emission, particularly under longwavelength UV light. Phosphors based on zinc sulfide, zinc cadmium sulfide, and various rare-earth compounds with white body color are usually used.

The readily excitable fluorescence or phosphorescence of phosphors can also be made use of in the automation, control, and regulation of automatic sorting processes, e.g., for mass-produced goods, in packaging or in quality control.

Since the 1960 s, stamp paper has been treated with fluorescent or phosphorescent phosphors. Since the stamp is stuck to the top right-hand corner of the letter, its recognition by the machine serves to establish the position of the letter and enables machine control of the correct positioning for postmarking and reading. The advantage of the phosphorescence of many inorganic phosphors lies in the fact that excitation and photoelectric scanning can be carried out separately. Thus the interfering blue emission of the optical brighteners present in most papers can be eliminated. By using fluorescing phosphors the blue emission of optical brighteners can be eliminated by inserting a blue filter (< 500 nm) with a sharp absorption edge. Fluorescent phosphors in stamps are preferably organic, while doped zinc sulfides with average particle sizes of 5 – 6 µm are used as afterglow phosphors.

The phosphor is either added to the paper pulp or applied as a surface coating of 1 – 2 g/m^2, which gives a readily detectable signal strength. Printing of fluorescent or afterglowing bar codes which allow, for example, sorting according to type of transport is increasing.

Some photoelectric scanning instruments combine excitation and photoelectric recognition into one system (luminescence scanners).

The coding of checks and check cards, passports, personal identity cards, securities, gift vouchers, entrance tickets, season tickets, and banknotes, with phosphors as invisible security markings is widely used. By this type of coding the phosphor acts as a passive protection against forgery and copying. With modern color copiers it is possible to produce forgeries which are virtually impossible for non-specialists to differentiate from the original. Using phosphors this differentiation is very simple. The original, with phosphor, fluoresces under UV light, the copy does not. Phosphors with white body color are generally used. The incorporation into the original is generally carried out by means of pulp coloration, embedding in plastic articles or us-

ing various printing processes, of which screen-printing is the most important because of the relatively coarse-grained nature of inorganic phosphors.

1.4.5. Safety and Accident Prevention

Inorganic luminescent pigments with long afterglows have long been used to make danger areas, escape routes, and switches visible.

For this application, phosphors must have the following properties:

1) excitability by daylight and artificial light,
2) high luminance during the first hour after excitation,
3) long afterglow time, i.e., decrease in luminance to 100 times that still detectable by the human eye (ca. 0.3 mcd/m^2) only after several hours,
4) an emission spectrum corresponding to the sensitivity curve of the eye, and
5) high chemical stability.

Zinc sulfide, zinc – cadmium sulfide and, with certain limitations, alkaline-earth sulfides form the basis of these phosphors. Table 7 lists commercial industrial phosphors. Copper-activated zinc sulfides are the most widely used phosphors for safety purposes. Their spectra correspond well with the sensitivity of the eye and it is excited by daylight and artificial light. At an illuminance of over 700 lx, complete excitation takes place within 30 s.

These zinc sulfides are sufficiently resistant to all media except strong mineral acids. A disadvantage is that they become grey through the action of high intensity UV light in the presence of water, due to photochemical reduction of Zn^{2+} to Zn^0. This effect can be reduced by coating with a UV-absorbing layer, but the outdoor application of these phosphors is limited. The phosphors are used in afterglow paints, plastics, ceramics and enamels.

The following basic rules should be observed in processing luminescent pigments:

1) The binder system must be colorless and transparent.
2) Additives based on heavy metals should not be used. This is particularly important at high processing temperatures.

Table 7. Properties of industrially important long-afterglow phosphors for accident prevention

Designation and luminescent color	Composition	Body color	Average particle size (FSSS)	Sieve analysis mesh size, µm	Sieve analysis residue, wt %	Typical luminance, mcd/m^2 3 min	10 min	30 min	60 min	100-fold visibility limit 0.32 mcd/m^2 in min.
LUMILUX green N	ZnS:Cu	yellow-green	40	125	≤1	128	30.4	7.9	3.3	370
LUMILUX green N-ST	ZnS:Cu	yellow-green	45	125	≤2	75	19.0	3.4	1.0	105
LUMILUX green N 5	ZnS:Cu, Co	yellow-green	40	125	≤1	123	32.2	9.5	4.4	580
LUMILUX green N 10	ZnS:Cu, Co	pale yellow-green	40	125	≤1	80	28.0	10.3	5.2	780
LUMILUX afterglow blend, green	ZnS:Cu, Co	yellow-green	40	125	≤1	126	31.3	8.9	3.9	460
LUMILUX green N-F	ZnS:Cu	yellow-green	25	100	≤0.2	135	29.7	7.2	2.9	310
LUMILUX green N-FF	ZnS:Cu	yellow-green	18	63	≤1	100	24.0	6.2	2.4	260
LUMILUX yellow N	(Zn, Cd)S:Cu	green-yellow	30	125	≤1	93.4	22.0	5.7	2.3	250
LUMILUX orange N	(Zn, Cd)S:Cu	yellow	30	125	≤1	65.0	15.7	4.2	1.8	220
LUMILUX red N	(Zn, Cd)S:Cu	orange	30	125	≤1	17.3	3.6	0.9	0.35	63
LUMILUX sea green	SrS:Bi	pale green	20	125	≤2	58.9	16.9	5.4	2.6	300
LUMILUX violet N	CaS:Bi	grey	20	125	≤2	15.3	4.1	1.8	0.64	105
LUMILUX blue N	(Sr, Ca)S:Bi	grey	20	125	≤2	31.5	8.9	2.9	1.3	190

3) Processes involving the action of large shearing forces on the pigments should be avoided (pressure-induced deterioration of luminescent pigments).

Markings for switches and escape routes, advice and warning signs, and special protective work clothes are produced increasingly from these afterglow phosphors.

In contrast to normal emergency illuminations, which are invisible or poorly visible in the presence of smoke, a low-light escape route system based on afterglowing pigments remains noticeable even in smoke [162–164]. The construction and action of such a system is shown in Figure 16.

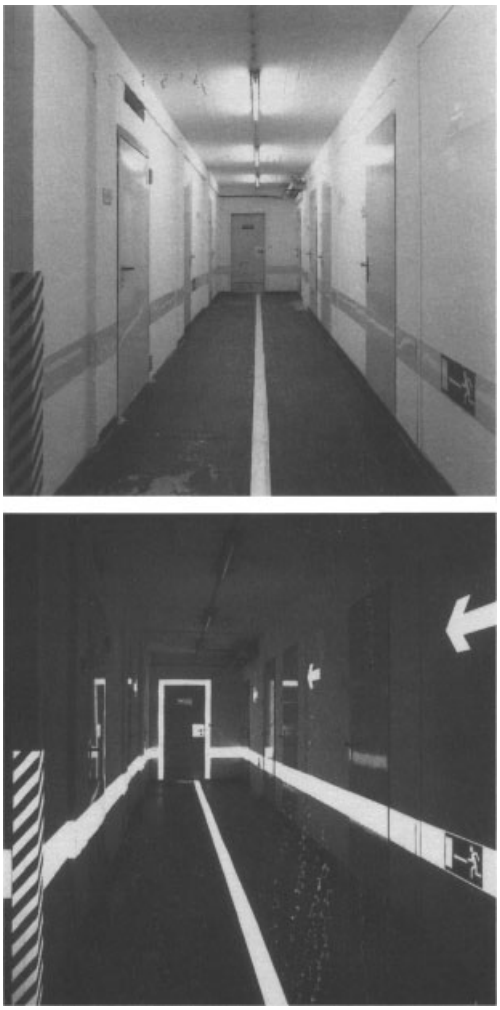

Figure 16. Low light level escape route system based on afterglowing LUMILUX$^{(R)}$ Green N phosphor

Guidelines for the measurement and characterization of long-afterglow products and systems are given in DIN 67 510.

1.4.6. Dentistry

Natural teeth exhibit blue-white fluorescence with a characteristic spectral distribution through the action of long-wavelength UV light [165]. The luminescence is attributed to certain components of collagen [166, 167].

For cosmetic reasons, the imitation of this phenomenon has long been attempted in the production of artificial teeth. Initially, uranyl salts were added to the ceramic paste. However, this also produced radioactivity in the artificial teeth [168]. Therefore, this technique has been abandoned, and the production of a fluorescence corresponding to that of natural teeth was attempted by the addition of rare earths or other activators [169]. The endeavors failed because the heat treatment necessary in the processing of the ceramics leads to uncontrolled changes in the intensity and spectrum of the emission. Later, luminescent pigments were developed which could be processed in the ceramic paste and gave the artificial teeth made from it blue-white fluorescence in the long-wavelength UV.

Yttrium silicates doped with cerium, terbium, and manganese give the best results [170]. A typical composition is $(Y_{0.937}Ce_{0.021}Tb_{0.033}Mn_{0.009})_2SiO_5$. The excitation maximum of these phosphors is in the range 325 – 370 nm (Fig. 17).

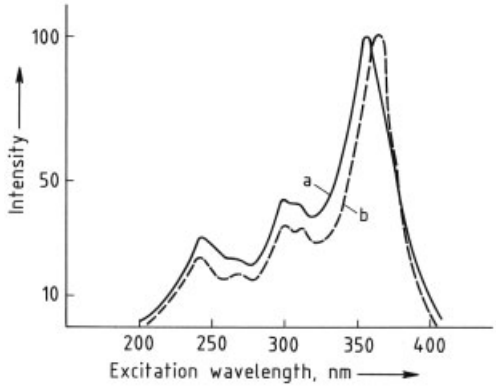

Figure 17. Excitation spectra of pigments for dental application
a) LUMILUX blue-white Z; b) LUMILUX yellow-white Z

The fluorescence color of the teeth can be varied by changing the concentration of activators in the phosphor, as the spectra of two commercial products show (Fig. 18). Addition of 0.3–0.5 wt % to the ceramic paste gives the artificial teeth the same fluorescence as natural teeth under the action of UV.

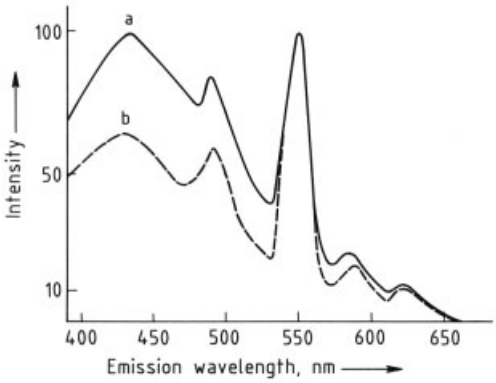

Figure 18. Emission spectra of pigments for dental application
a) LUMILUX blue-white; b) LUMILUX yellow-white

1.4.7. Other Uses

Green-fluorescing Zn_2SiO_4:Mn^{2+} and magnesium tungstate that fluoresces blue under short-wavelength UV are mixed with the sorbents (silica gel, aluminum oxide) of *TLC plates* in proportions of ca. 3 wt % to improve the contrast in the detection of UV-absorbing compounds. For this application the phosphors must be particularly fine grained and must not alter the pH value of the sorbent.

Infrared radiation can be detected by phosphors by various mechanisms. Upon excitation with UV radiation, ZnS:Cu^+, Co^{2+} emits an afterglow that is quenched by IR radiation, whereas ZnS:Cu^+, Pb^{2+} stores excitation energy and emits it spontaneously under the action of IR radiation.

The phosphors CaS:Sm^{3+}, Ce^{3+} and SrS:Sm^{3+}, Eu^{2+} only exhibit low fluorescence on excitation by UV radiation and daylight. Through the subsequent action of IR radiation these phosphors can be stimulated to emit green or orange-red light.

Infrared-stimulable phosphors are used as screens for the detection of IR radiation in IR measuring and testing techniques and for the adjustment of IR lasers, and IR light recorders. This method is cheaper than IR display devices.

Scintillation counters are used to detect energy-rich particles and γ-rays. They consist of a combination of a phosphor, usually a single crystal, with a photomultiplier or a photodiode as detector. When energy-rich particles impinge on the phosphor crystal, a flash of light is formed which is registered by the detector. It is important for the phosphors to have a high absorption for the exciting energy and a short decay time ($< 1\ \mu$s), to produce high counting rates at a good signal-to-noise ratio and proportionality between the absorbed energy and the number of photons emitted. Alkali-metal iodides (see Table 1), $Bi_4Ge_3O_{12}$, $CaWO_4$, $ZnWO_4$, $CdWO_4$, and ZnS:Ag^+, Ni^+ are typical phosphors for this application [170].

Thermoluminescence dosimeters (TLD) absorb ionizing radiation and accumulate the energy absorbed over a very long period. Heating the TLD results in emission of light proportional to the absorbed energy. For monitoring a person, an effective atomic number of the TLD is chosen to be as near as possible to that of human tissue; $Li_2B_4O_7$:Mn^{2+}, LiF (see Table 1), and also $CaSO_4$:Dy^{3+} and CaF_2:Mn^{2+} meet this requirement. For use as a TLD, the phosphor is encapsulated in a glass ampoule or is sealed in a heat-resistant synthetic resin.

Phosphors for radio-activation are used as mixtures with a radioactive compound whose radiation effects continuous luminescence of the phosphor. The radium compound used formerly in the production of radio-activated systems restricted their lifetime in spite of the very high luminescence yield and long half-life, because the exciting α-particles of the radium decompose the phosphor [171]. After the radiation safety regulations for α-emitters became stricter, β-emitters were used instead, even though the luminescence yield is less efficient for a given activity. Compounds with high β-activity were therefore used such as ^{3}H, ^{14}C, or ^{147}Pm compounds. In addition, β-radiation has the advantage that it causes less decomposition of the phosphor.

Tritium-labeled organic binders are used for the production of luminous products and paints. Because of the high cost of tritium, the phosphors must have a high efficiency.

Another radioactive luminescent system is the β-light, which consists of a gas-tight glass vessel that contains gaseous tritium and whose walls are coated with a fine-grained phosphor. The β-lights are mainly used in military applications for the illumination of instruments and scales. Clock faces and hands, escape route signs, and emergency exits are also made visible in darkness through the use of β-lights.

Because of health risks, particularly through the use of radioactive substances, their use is declining sharply and they are being replaced by long-afterglow products.

Fluorescent and afterglow products are used to attain particular effects in cinema films, on the stage and in works of art. These *effect pigments* are physiologically harmless phosphors, mainly based on zinc sulfides.

1.5. Testing of Industrial Phosphors

Luminescence color is determined by visual comparison, by calculation of the color coordinates x and y from the emission spectrum, or by means of a tri-stimulus colorimeter under the excitation definitive for the application (UV radiation, cathode rays, etc.).

Luminescence Brightness. A practical brightness measurement for UV-excitable phosphors is the comparison with internal reference phos-phors under excitation mainly with short- and long-wavelength (254 nm and 365 nm) UV lamps. The visible parts of the radiation from the mercury discharge must be removed with filters. Integral brightness values in mcd/m^2 are obtained with commercial brightness measuring instruments.

Values of the quantum yield are obtained by comparison measurements against standard phosphors of the NBS. The dependence of the quantum output on the excitation wavelength are obtained by comparison with phosphors whose quantum output is only slightly dependent on the excitation wavelength. Such phosphors are, for the blue emission region, sodium salicylate, for the blue-green, coronene ($C_{24}H_{12}$), and for the orange to orange-red, Lumogen Red T (BASF). Measurements of quantum yield are rarely carried out on industrial phosphors. The absolute measurement of the quantum yields is described in detail in [129, 172].

For long-afterglow phosphors, the brightness as a function of time after standardized excitation is measured. The excitation and measurement of the brightness decay and the characterization of products with long afterglow phosphors is laid down in DIN 67 510.

Phosphors for cathode-ray tubes, television screens, monitor screens, radar screens, and oscilloscopes are tested under electron excitation. Electron energy and density should be similar to the conditions of the tube in which the screen will be used. The phosphors are sedimented or brushed onto light-permeable screens and coated with an evaporated aluminum coating to dissipate charge. The luminescence brightness and color of the emitted light are measured with optical instruments such as photomultipliers or spectrophotometers.

For X-ray excitable phosphors, the intensifier effect is tested as a measure of the luminescence yield. A very simple process uses the exposure of blue- or green-sensitive films. The phosphor is excited by X-rays in a cell covered with a light-sensitive film. The blackening of the film is measured with a densitometer and, for given excitation conditions, is a measure of the efficiency of the phosphor.

Particle Size and Shape. The measurement of particle-size distribution is carried out with particle counting instruments (Coulter counter) or by laser or light-beam scattering (Cilas). The Fisher number is still widely used as a measure of the average particle diameter (Fisher sub sieve size). The average particle diameter is measured by means of the air permeability principle [173, 174].

The form and composition of phosphor crystals are assessed by scanning electron microscopy. Halophosphate phosphors have a lamellar form and $CaWO_4$ crystals are, for example, spherical.

For relatively coarse phosphors, such as $ZnS:Cu^+$, the particle-size distribution is determined from sieve residue curves.

Decay. The decay time requirements must be adhered to very precisely for cathode-ray tube phosphors. The measuring devices consist of fast excitation sources (flash lamps, lasers), photo-

multipliers with very low time constants, and an oscilloscope [171].

Sedimentation Behavior. The coating of cathode-ray tubes is usually carried out by sedimentation. The sedimentation rate in the sedimentation fluid is principally determined according to Stokes' law by the size, shape, density, and surface charge of the particles. The sedimentation characteristics of CRT phosphors and their blends under typical application conditions is an important criterion for their suitability, apart from the luminescence brightness and luminescence color.

The sedimentation density allows conclusions to be made on the degree of agglomeration of individual crystals at constant polarity and electrolyte concentration of the sedimentation fluid. A low density is an indication of extensive agglomerate formation.

Conductivity and pH of the Aqueous Extract. The conductivity and pH value of aqueous phosphor slurries are indications of hydrolysis processes, poor washing out of the flux from the phosphor, or the presence of acidic or alkaline coatings. Increases in the conductivity of the slurry caused by electrolytes is often a disadvantage in the production of image screens by means of the sedimentation process.

For phosphors that are applied by cataphoresis, the zeta potential is occasionally tested. In some cases, the specific surface area (BET) is also used for assessment.

For lamp phosphors, the definitive quality report is only given after tests in lamps, for which the most important criterion is the quantum yield throughout the lifetime.

2. Chemiluminescence and Bioluminescence

2.1. Introduction

Chemiluminescence is the emission of light from a chemical reaction. The energetics of this process are shown in Figure 19, which also gives for comparison the energy diagrams of typical nonluminescent thermal and photochemical reactions. For the nonluminescent thermal reaction (Fig. 19 A), the ground state reactant R_0 is activated by thermal energy to the transition state $R\neq$, which, after reorganization of the chemical bonds, affords the ground state product P_0; excess energy is dissipated as heat. This represents by far the most common chemical transformation.

In a photochemical process (Fig. 19 B), R_0 is promoted by absorption of light energy to the electronically excited reactant $R*$, which on reorganization of the chemical bonds leads to P_0, and excess energy is liberated as heat.

Although the chemiluminescent reaction (Fig. 19 C) has characteristics of both, it is unique and distinct in that thermal activation of R_0 leads to $R\neq$, which upon bond reorganization generates an electronically excited product $P*$, the latter affording P_0 with light emission.

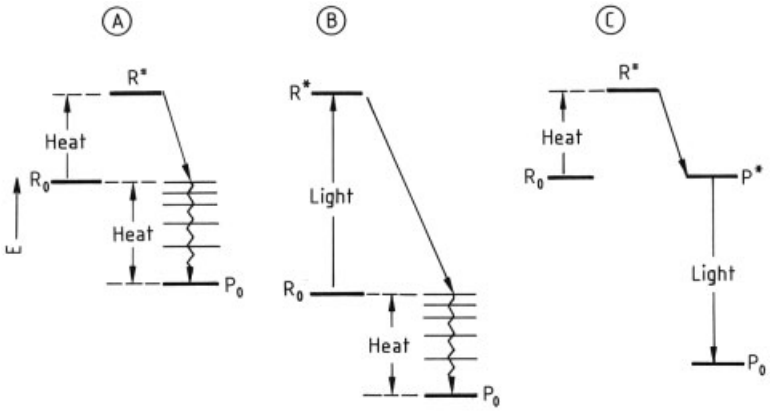

Figure 19. Energy level diagrams for typical chemical reactions
A) a thermal exothermic process; B) photochemical transformation; C) a chemiluminescent process

Bioluminescence is identical to chemiluminescence, except that light is emitted by biological matter, e.g., bacteria, fungi, plants, insects, and fish. A chemical reaction, promoted by enzymes, takes place in the living cell, generating an electronically excited product P∗, which manifests itself through characteristic light emission. A well-known example is the firefly.

Phenomena related to chemiluminescence and bioluminescence include anodoluminescence, candoluminescence, cathodoluminescence, electrochemiluminescence, piezoluminescence, pyroluminescence, radioluminescence, sonoluminescence, thermoluminescence, and triboluminescence.

2.2. History

Bioluminescence has been much longer known to humans than chemiluminescence. The Chinese reported the emission of "living light" ca. 5000 years ago. The first experimental inquires are credited to Aristotle, who recognized the heatless nature of bioluminescence over 2000 years ago. Scientific investigations of bioluminescence commenced with BOYLE, who demonstrated the requirement for oxygen in the emission exhibited by fungi associated with decaying plant matter (fox fire). NEWTON, FRAN-KLIN, DUBOIS, and others contributed decisively in the following period to the understanding of bioluminescence; but it was HARVEY [175] who developed the biochemical basis of modern research in this field. Contemporary research is summarized in [176, 177].

The history of chemiluminescence is much younger [178]. The first report of chemiluminescence in solution–the autoxidation of lophine (**1**)–was made by RADZISZEWSKI in 1871.

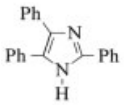

1

The earliest mention of singlet oxygen chemiluminescence dates to BLANCHETIERRE (1913), who observed a bright red emission when chlorine gas was passed through an alkaline hydrogen peroxide solution. DUFRAISSE (1926) is credited with the discovery of the singlet-oxygen chemiluminescence observed in the thermal decomposition of rubrene endoperoxide (**2**).

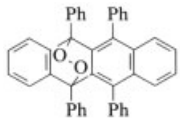

2

The autoxidation of luminol (**3**), leading to the emission of bluish light, was discovered by ALBRECHT (1928).

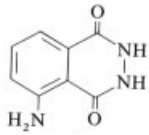

3

The perhydrolysis (reaction with H_2O_2) of lucigenin (**4**) was observed by GLEN and PLETSCH (1935), while that of oxalyl chloride was first reported by CHANDROSS (1963). The chemiluminescence of the related diaryl oxalate (**5**) is particularly significant because it came to commercial fruition in the form of chemical light sources (see Section 2.5.1).

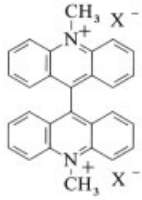

4

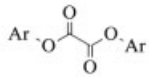

5

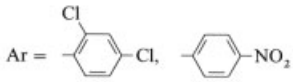

Radical-initiated autoxidation of alkyl-substituted aromatic hydrocarbons in solution was first shown by VASIL'EV and BELYAKOV (1967). The termination of peroxy radicals is responsible for the formation of the electronically excited product (Russell mechanism). In the presence of efficient fluorophors, e.g., 9,10-dibromoanthracene, the chemiluminescence is strongly enhanced.

Although dioxetanes (**6**) were postulated in the early 1900 s, KOPECKY and MUMFORD succeeded in isolating the first stable deriva-

tive, trimethyl-1,2-dioxetane, only in 1969. α-Peroxylactones (**7**) (1,2-dioxetanones), which in the mid-1960 s were proposed as intermediates in firefly bioluminescence, were synthesized for the first time by ADAM and LIU (1972). Dioxetane chemiluminescence is an active field of research [179] with promising medical applications [180].

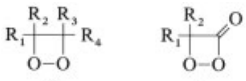

6 **7**

Chemically induced electron exchange chemiluminescence (CIEEL) was discovered by SCHUSTER (1977), who showed that diphenoyl peroxide (**8**) in the presence of easily oxidized electron donors leads to efficient chemiluminescence. The process is general and applies to a wide range of other cyclic peroxides, e.g., phthaloyl peroxide (**9**) and the peroxalate (**10**). Particularly significant is the fact that the α-peroxylactones (**7**) exhibitCIEEL, which implies that some forms of bioluminescence, such as that of the firefly, follow this mechanism.

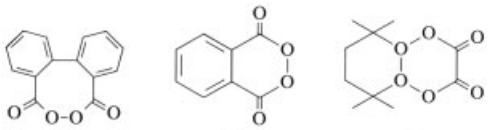

8 **9** **10**

The demonstration by CILENTO et al. (1977) that horse radish peroxidase-catalyzed autoxidation of enolizable carbonyl compounds leads efficiently to chemiluminescence opened up new pathways in the production of electronic excitation in cells. The biological consequences of this novel form of generation of excited state are far-reaching [180], because although the cell may use them beneficially to generate bioluminescence, the high-energy intermediates involved may induce cellular damage.

2.3. Mechanisms of Light Emission

With few exceptions, processes involving autoxidation or peroxides or both are predominant in chemi- and bioluminescence; for further details, see [176–179].

Three distinct mechanisms of light emission are known: direct, enhanced (energy transfer), and chemically induced (electron exchange) [179]. The chemiluminescent decomposition of α-peroxylactones (**7**) is given here as an example. For this case, in the direct mechanism an excited carbonyl derivative ($R_2C{=}O*$) is the emitting species (Eq. 1).

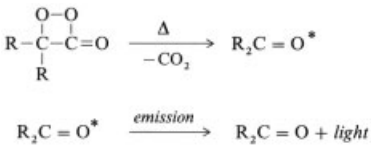

(1)

Firefly bioluminescence is an example of such direct emission [176].

In the enhanced mechanism, the first step is also the generation of $R_2C{=}O*$, but in the presence of an efficient fluorophor, e.g., 9,10-dibromoanthracene (DBA), the emitting species is the electronically excited DBA$*$ (Eq. 2). The DBA$*$ results from energy transfer between $R_2C{=}O*$, the energy donor, and DBA, the energy acceptor.

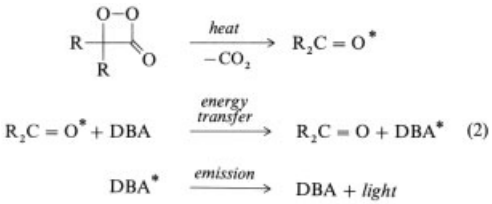

(2)

For efficient enhanced chemiluminescence, the fluorophor must be chosen such that the energy transfer is exothermic and its emission quantum yield high. By judicious choice of the fluorophor, the wavelength of the emitted light can be manipulated in this indirect form of chemiluminescence. In bioluminescence, the luminescent jelly fish *Aequorea* falls into this category [176], for which a photoprotein acts as fluorophor.

In the chemically induced mechanism (CIEEL) in addition to the chemiluminescent compound e.g., the α-peroxylactone (**7**), a fluorophor is necessary that is also a good electron donor, e.g., rubrene (Rub). As shown in Equation 3, the emitting species is electronically excited rubrene (Rub$*$).

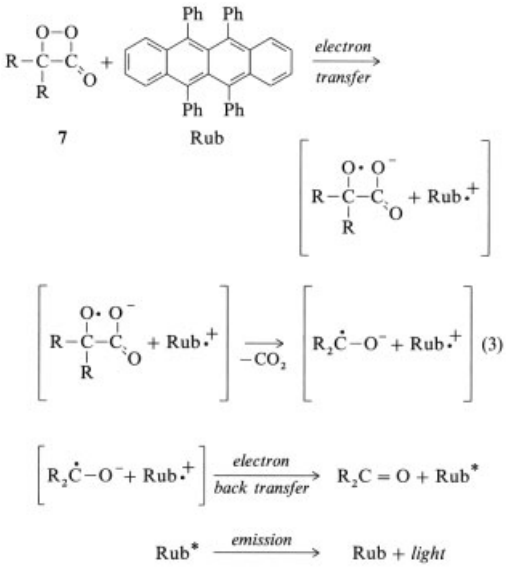

The Rub is energized by electron exchange, involving chemically induced electron transfer between **7** and rubrene to form a radical ion pair. Subsequent decarboxylation of the radical anion affords the ketyl radical, which transfers an electron back to the radical cation, generating Rub∗. The emitted light in this CIEEL process is characteristically that of Rub. This type of light emission mechanism operates in the firefly and the chemical light source Cyalume.

2.4. Efficiency of Light Emission

Consider the sequence of steps outlined in Equation 4 for a generalized chemiluminescent (bioluminescent) process.

(a) chemical reaction	R	$\rightarrow$ I	Φ_{chem}
(b) excitation	I	$\rightarrow$ P*	Φ_{exc}
(c) energy transfer	P* + F	$\rightarrow$ P + F*	Φ_{et}
(d) transmission	P*(or F*)	$\rightarrow$ P(or F) + light	Φ_{em}

(4)

In step (4 a) a chemical reaction (e.g., autoxidation, thermolysis, or electron exchange) transforms reactant R into a high-energy intermediate I that is capable of generating the necessary excited states for chemi- or bioluminescence with an efficiency Φ_{chem}. In the excitation step (4 b) the electronically excited product P∗ is formed, for which the efficiency is given by Φ_{exc}. In the presence of an appropriate fluorophor F, the electronically excited fluorophor F∗ is obtained by energy transfer (4 c), characterized by the efficiency Φ_{et}. Finally, in step (4 d) the observed emission occurs from P∗ or from F∗, depending on whether direct or enhanced (indirect) chemiluminescence (bioluminescence) applies, with an efficiency Φ_{em}. The quantum yield of chemiluminescence (bioluminescence) $\Phi_{CL(BL)}$, is given by the product of the efficiencies for the individual steps involved. Thus for direct emission

$$\Phi_{CL(BL)} = \Phi_{chem}\Phi_{exc}\Phi_{em} \tag{5}$$

and for enhanced emission

$$\Phi_{Cl(BL)} = \Phi_{chem}\Phi_{exc}\Phi_{et}\Phi_{em}\,. \tag{6}$$

Of these efficiencies, all but Φ_{exc} can in principle be determined experimentally by modern analytical methods. For example, conventional photometry is usually used to measure Φ_{CL} or Φ_{BL} [176, 177].

If the high-energy intermediate I, e.g., a dioxetane, can be isolated and purified, Φ_{chem} is necessarily unity. Otherwise Φ_{chem} must be independently determined by analytical techniques such as chromatography and/or spectroscopy. However, for very labile and short-lived transients, this is difficult, and large uncertainties in Φ_{CL} and Φ_{BL} can arise from this difficulty alone.

The Φ_{em} term is a characteristic photophysical property for P∗ (direct emission) or F∗ (enhanced emission), corresponding either to the fluorescence or phosphorescence efficiencies of the emitting species. In favorable cases these may be reported in the literature, but hardly ever under exactly the same conditions as required for the chemi- or bioluminescent experiment. When not available, standard fluorescence or phos-phorescence measurements must be performed to determine Φ_{em}.

The energy transfer efficiency Φ_{et} (only for the enhanced emission) is determined by means of Stern – Volmer kinetics, by determining the quantum yields Φ_{CL} or Φ_{BL} as a function of fluor-ophor concentration. The intercept of a double-reciprocal plot of Φ_{CL} or Φ_{BL} versus [F] provides these efficiencies at infinite fluorophor concentration, the conditions for 100 % efficient energy transfer, i.e., $\Phi_{et} = 1.0$. For further details see [179].

Once all efficiencies are experimentally determined, the excitation parameter Φ_{exc} can be calculated according to Equation 5 (direct emission) or Equation 6 (enhanced emission). This parameter is an inherent and characteristic property of a chemi- or bioluminescent species and takes values in the range 0 to 1 [176–179].

The firefly is one of the most efficient bioluminescent systems with $\Phi_{exc} = 0.88$. On the other hand, bacterial bioluminescence has $\Phi_{exc} = 0.1$. Among the chemical systems, the perhydrolysis of diaryl oxalates (enhanced chemiluminescence) is one of the most efficient, with $\Phi_{exc} \leq 0.3$. For comparison, the direct chemiluminescence arising from luminol autoxidation has $\Phi_{exc} = 0.01$. Nevertheless, even such inefficient chemiluminescent systems are profitably exploited in applied work.

2.5. Uses

Uses of bioluminescence include illumination, decoration, pleasure, decoy, bate, military purposes, scientific applications, medical use, and industrial needs.

Since almost all chemi- and bioluminescent systems involve autoxidation, one of the earliest analytical applications concerned monitoring the amount of oxygen in biological media or the enzymatic production of hydrogen peroxide by peroxidases. However, when it was realized that ATP is essential to release the bioluminescence of the firefly, a new era of scientific, medical and industrial activity and interest in biological and chemical light was initiated.

Applications in the gaseous, liquid, and solid states are known; details are given in [176–179].

2.5.1. Light Sources

One of the evident commercializations of chemiluminescence is its use for illumination. Cyalume lightsticks, pioneered by American Cyanamid, stand out. The diaryl oxalate/hydrogen peroxide/fluorophor chemiluminescent system is the source for this chemical light. Its color can be adjusted from blue to red, depending on the fluorophor used. Electron exchange chemiluminescence is usually the mechanism responsible for light production.

Because neither ignition nor electrical current is envolved in chemiluminescence, it is ideal for safety lights. Consequently it is used more and more as emergency lighting in camping, boating, and diving. It is also finding more frequent use as marking light for handicapped people, children, joggers, and animals. Chemical light sources are also used as bait in fishing and hunting and decoration in night performances. Besides the diaryl oxalate/hydrogen peroxide/fluorophor system, tetrakis(dimethylamino) ethylene – oxygen chemiluminescence also finds occasional use.

2.5.2. Chemical Analysis

Enormous progress has been made in the use of chemi- and bioluminescence in the qualitative and quantitative analysis of chemical and biological substances in clinical chemistry, food science, agricultural chemistry, soil science, cell physiology, microbiology, marine biology, medicine, veterinary medicine, and others. Particularly decisive for this development is the enormous sensitivity of modern photometric equipment. Thus, as few as 1000 photons can be detected with current single photon counters, which under optimal conditions correspond to a lower concentration limit of ca. 10^{-21} M that can be monitored in practical situations. The advantages of these analytical tools are high sensitivity, broad scope, high reproducibility, enormous selectivity, small sample size, low cost, convenience, speed of analysis (in optimal cases fractions of a second), low hazard, and minimal toxicity. Luminol, lucigenin, and diaryl oxalates are the most extensively used chemiluminescent systems. Of the many bioluminescent systems, the firefly, lumibacteria, and luminous jelly fish (*Aequorea* and *Obelia*) are used most frequently.

Nearly all chemiluminescent systems require molecular oxygen or hydrogen peroxide and a suitable catalyst. Thus, luminol requires cooxidants such as heme, peroxidase, peroxydisulfate, ferricyanide, or redox-active transition metals. Lucigenin requires transition metal ions as cooxidants, while for diaryl oxalates, alkaline conditions and an appropriate fluorophor are essential, e.g., polycyclic aromatic hydrocarbons (derivatives of pyrene, anthracene, tetracene, xanthenes, porphines).

For the bioluminescent systems, specific cofactors are needed in addition to the appropriate enzymes (luciferases or dioxygenases), and the luminescent substrate or luciferin. In the case of firefly bioluminescence, ATP serves this purpose, while bacterial bioluminescence depends on NADH (nicotinamide adenosine dinucleotide, reduced form) and FMN (flavin mononucleotide, oxidized form). *Aequorea* and *Obelia* bioluminescence utilizes Ca^{2+} ions and photoproteins

(aequorin and obelin), since their mechanism of light emission involves energy transfer. These cofactors in the chemi- and bioluminescent systems allow for selectivity and differentiation in chemical and biological analyses.

Since all the above-mentioned light-emitting systems require molecular oxygen or hydrogen peroxide, it is not surprising that these chemicals figured among the first in chemi- or bioluminescent assays. Luminol and lumibacteria permit the detection of ca. 10^{-13} M concentrations of molecular oxygen in aqueous solution. Such sensitivity allows the measurement of oxygen production at the cellular level, and numerous oxygen electrodes have been developed for this purpose. The concentrations of hydrogen peroxide that can be monitored by these luminescent probes are as low as 10^{-9} M. Hypohalites and chlorine gas also serve as oxidants and can therefore be analyzed by these luminescent methods.

Since metal ions serve an important role as cooxidants in luminol and lucigenin chemiluminescence, analytical techniques have been developed to determine low levels (parts per billion) of metal ions. These include, with the exception of Cd(II), Hg(I), T(IV), V(V), Zn(II), Zr(IV), essentially all redox-active transition metal ions. Some of the more common ones are: Ag(I), As(III), Bi(III), Co(II), Cr(II), Cu(II), Fe(II), Ir(IV), Mn(II), Mo(V), Ni(II), Os(IV, V), Pb(II), Pt(II), Ru(III, IV), Tl(III), and V(IV).

A specific analytical method for detecting and quantitating Ca^{2+} ions is the use of the photoproteins aequorin or obelin. This has enabled studies of the role of intercellular Ca^{2+} ions, allowing investigations of the nature of electrical and chemical stimuli in biological processes. Since the jelly fish photoproteins are also ultrasensitive to Sr^{2+} ions, a potential application is the detection of traces of radioactive strontium in nuclear fallout.

The chemiluminescence of luminol has also been used for inorganic anions. The scope is more limited, but common examples include CN^-, SCN^-, OBr^-, S^{2-}, $S_2O_3^{2-}$, and SO_3^{2-}. For this purpose redox titrations are employed, allowing detection levels of the order of parts per billion.

Light-monitoring analyses are useful in the qualitative and quantitative analysis of organic compounds, especially those of biochemical interest. Both chemiluminescent and bioluminescent systems are used for this purpose. Either the assay exploits the fact that bio- and chemiluminescence usually involve autoxidation reactions or that specific cofactors (e.g., H_2O_2, Ca^{2+}, ATP) are necessary. The combination of these two features offers broad scope and also selectivity in chemical and biological analysis.

Luminescent assays exist for alcohols (ethanol), polyhydroxy compounds (glycerol, mannitol, sorbitol, inositol), and sugars (glucose). Their ability to release hydrogen peroxide on autoxidation and detecting the light emission stimulated by the hydrogen peroxide form the chemical basis of this analysis. Also acids, e.g., ascorbic acid, uric acid, and amino acids (thyroxine), have been employed. The latter inhibit the luminol chemiluminescence and are analyzed in that way. Chemiluminescent assays have been reported for esters, including glycerides, lactates, malates, pyruvates, and phosphates. Long-chain aldehydes are monitored by bacterial bioluminescence, since they are required as luciferins for the light emission.

Fluorescent polycyclic aromatic hydrocarbons, e.g., anthracenes, pyrenes, tetracenes, and heteroaromatic compounds, are conveniently analyzed by electron-exchange chemiluminescence with diaryl oxalates. A bioluminescence assay also exists for TNT, which can be detected at concentrations as low as 10^{-18} M by means of an elaborate bacterial assay.

Bio- and chemiluminescent analyses have been reported for a wide range of steroids (testosterone, pregnanediol, androsterone, cortisol, estradiol), some antibiotics (tetracyclin), leukemia tumor drugs (methotrexate), and sedatives (barbiturates, cocaine, heroin). Organophosphorus nerve gases and insecticides have also successfully been analyzed.

Some of the relevant biomolecules include numerous vitamins (B_{12}, biotin, ascorbate, and

folate); however, the most significant are the nucleotides ATP, c AMP, FAD, FMN, c GMP, and NADH. Adenosine triphosphate is the essential component of the firefly bioluminescent system. It is one of the most sensitive luminescent assays, allowing the detection of as little as 10^{-17} mol ATP. Since a single bacterium possesses only ca. 10^{-18} mol of ATP, the firefly luciferin–luciferase assay permits detection of ca. 10 bacteria per milliliter of medium.

The nucleotides NADH and FMN are important in bacterial bioluminescence. When employed together with proteins, particularly enzymes, an analytical tool of wide scope is available for biotechnological and biomedical purposes. By means of firefly and bacterial bioluminescence, compounds ranging from the simple tripeptide glutathione to complex proteins such as insulin, albumin, immunoglobin, hemoglobin, hematin, myoglobin, have been assayed. Prominent applications concern enzymes such as oxidases, dioxygenases, peroxidases, catalases, dehydrogenases, kinases, hydroxylases, and phospholipases.

2.5.3. Biotechnology

The fact that ATP can be conveniently monitored by firefly bioluminescence, allowing the detection of trace quantities (ca. 10^{-17} mol per mL medium), constitutes the basis of most biotechnological applications. Thus, the detection of harmful bacteria in industry, e.g., foods (*Salmonella*), beverages, crops, soils, sediments, textiles, waste waters, and fresh water, engage this ATP assay based on the firefly luciferin–luciferase system. Such ultrasensitive detection methods allow rigorous control of environmental pollution.

Another application concerns the use of bacteria for the removal of specific pollutants in waste water recycling. A major problem is to recognize at an early stage when the viability of the specially cultured bacteria is endangered. By means of genetic engineering, the bacteria are equipped with genes for luminescence, permitting their function to be monitored. Diminished light-emission indicates that poisons are reducing the bacterial population, and immediate corrective measures can be taken.

The aging (autoxidation) of foods can be easily assessed by using lumibacteria to assay the resulting peroxidic products. This assay also allows the quality and efficiency of antioxidants for foods to be tested.

Promising opportunities concern monitoring biologically relevent processes by using proteins equipped with chemi- and bioluminescent agents. They should constitute efficient tools in biomass analysis, microbial assays, and germ warefare. Much progress along these lines is anticipated in the future.

2.5.4. Medicine

The quantitative analysis of ATP with the firefly luciferase – luciferin system also has medical applications. Thus, bacterial infections that increase the amount of ATP in the blood and urine can be detected very efficiently, conveniently, and at low cost. In fact, this ATP analysis was one of the earliest applications of bioluminescence.

The involvement of Ca^{2+} in heart disease and multiple sclerosis can be monitored with the jelly fish photoproteins aequorin and obelin.

A field of intensive activity in medicine concerns phagocytosis, in which macrophages employ intensive autoxidative activity to clean up infected cells, increasing (decreasing) the oxygen metabolites in the living system [177]. Disorders such as rheumatoid arthritis, lung diseases, muscular dystrophy, cystic fibrosis, ischaemic and myocardial infarction, carcinogenesis, diabetes, foetal and neonatal abnormalities, immune-based diseases, and others can be detected and monitored at an early stage.

The most important advances have been made in immunoassays. The high sensitivity rivals that of radioassays, but the problems involved with working with radioactive isotopes are avoided, and the cost is considerably lower.

A distinction is made between luminescent immunoassay (LIA), luminescent enzyme immunoassay (LEIA), luminescent enzyme multiplied immunoassay technique (LEMIT), and luminescent cofactor immunoassay (LCIA) [177].

Luminol and diaryl oxalate chemiluminescence and luciferin and bacterial bioluminescence are frequently used. Hormonal diseases, genetic disorders, endocrine problems, and nu-

merous other medical challenges have already been successfully dealt with.

3. References

General References

1. E. Newton Harvey: *A History of Luminescence (until 1900),* Academic Press, New York 1966.
2. N. Riehl: *Einführung in die Lumineszenz,* Verlag Karl Thiemig KG, München 1970.
3. E. J. Adirowitsch: *Einige Fragen zur Theorie der Lumineszenz der Kristalle,* Akademie Verlag, Berlin, 1953.
4. D. Curie: *Luminescence in Crystals,* Methuen & Co. Ltd., London 1963.
5. Felix Fritz: *Leuchtfarben: Geschichte, Herstellung, Eigenschaften,* Chem. Technischer Verlag Dr. G. Bodenbender, Berlin 1940.
6. P. Goldberg (ed.): *Luminescence of Inorganic Solids,* Academic Press, New York 1966.
7. W. Espe: *Werkstoffkunde der Hochvakuumtechnik,* vol. III, VEB Deutscher Verlag der Wissenschaften, Berlin 1961.
8. H. W. Leverenz: *An Introduction to Luminescence of Solids,* John Wiley and Sons, New York 1950.

Specific References

9. W. P. Jorrissen, *J. Chem. Educ.* **25** (1948) 685 – 686.
10. P. Lenard, V. Klatt, *Ann. Phys.* **15** (1904) 225 – 282, 425 – 484, 633 – 673.
11. B. Gudden, R. W. Pohl, *Z. Phys.* **2** (1920) 192.
12. O. W. Lossew, *Telegrafia i Telefonia* **18** (1923) 61.
13. G. Destriau, *J. Chem. Phys.* **33** (1936) 587.
14. F. Urbach, *Wien. Ber.* **II a 139** (1926) 473.
15. J. T. Randall, M. F. H. Wilkens, *Proc. R. Soc. London* **A 184** (1945) 347, 366, 390.
16. F. Seitz, *Trans. Faraday Soc.* **35** (1939) 74.
17. N. F. Mott, R. W. Gurney, *Trans. Faraday Soc.* **35** (1939) 69.
18. N. Riehl, M. Schön, *Z. Phys.* **114** (1939) 682 – 704.
19. S. G. Tomlin, *Proc. Phys. Soc. London* **82** (1963) 465 – 466.
20. J. E. Holliday, E. J. Sternglass, *J. Appl. Phys.* **28** (1957) 1189.
21. H. A. Bethe, *Ann. Phys. (Leipzig)* **5** (1930) 325 – 400.
22. V. E. Cosslett, R. N. Thomas, *Brit. J. Appl. Phys.* **15** (1964) 883.
23. V. E. Cosslett, R. N. Thomas, *Brit. J. Appl. Phys.* **15** (1964) 1283.
24. J. T. W. de Hair, *J. Solid State Chem.* **33** (1980) 33.
25. A. J. de Vries, G. Blasse, *Mater. Res. Bull.* **21** (1986) 683.
26. A. J. de Vries, G. Blasse, J. R. Pet, *Mater. Res. Bull.* **22** (1987) 1141.
27. L. Ozawa, *J. Electrochem. Soc.* **119** (1972) 1783.
28. Z. Mazurak, E. Lukowiak, B. Jezowska-Trzebiatowska, *J. Lumin.* **33** (1985) 159 – 166.
29. T. Kano, H. Yamamoto, Y. Otomo, *J. Electrochem. Soc.* **119** (1972) 1561 – 1564.
30. R. A. Hewes, J. F. Sarver, *Phys. Rev.* **182** (1969) 427.
31. H. J. Guggenheim, L. F. Johnson, *Appl. Phys. Lett.* **15** (1969) 51.
32. L. G. Van Uitert et al., *Appl. Phys. Lett.* **15** (1969) 53.
33. S. G. Parker, R. E. Johnson, *J. Electrochem. Soc.* **119** (1972) 610 – 613.
34. M. Greenblatt, E. Banks, *J. Electrochem. Soc.* **124** (1977) 409 – 413.
35. F. Auzel, D. Pecile, D. Morin, *J. Electrochem. Soc.* **122** (1975) 101 – 107.
36. R. Ward: *Preparation and Characteristics of Solid Luminescence Materials,* Cornell Symp. 1946, Wiley and Sons, New York/London 1948.
37. A. Schleede, *Angew. Chem.* **50** (1937) 908.
38. B. J. Verwey, F. A. Kröger, *Philips Tech. Rev.* **13** (1951) 90 – 95.
39. N. Riehl, H. Ortmann, Angew. Chem. u. Chem. Ingenieurtechnik, Monogr. 72, 1957.
40. K. Era, S. Shionoya, Y. Washizawa, *J. Phys. Chem. Solids* **29** (1968) 1827 – 1841.
41. Y. Uehara, *J. Chem. Phys.* **62** (1975) 2982 – 2994.
42. J. Dieleman, S. H. de Bruin, C. S. van Doorn, J. H. Haanstra, *Philips Res. Rep.* **19** (1964) 311 – 318.
43. N. R. J. Poolton, *J. Phys. Chem: Solid State Phys.* **20** (1987) 5867 – 76.
44. Kasei Optonix, JP (6295378) 8 795 378, 1987 (T. Hase).
45. S. Asano, *Sci. Light (Tokyo)* **4** (1955) 32.
46. W. van Gool, *Philips Res. Rep.* **13** (1958) 157 – 166.
47. A. Bril, H. A. Klasens, *Philips Res. Rep.* **10** (1955) 305.
48. W. Lehmann, *J. Electrochem. Soc.* **110** (1963) 754 –758.

49. S. Rothschild, *Trans. Faraday Soc.* **42** (1946) 635.
50. F. A. Kröger, J. E. Hellingman, N. W. Smit, *Physica (Amsterdam)* **15** (1949) 990 – 1018.
51. H. Blicks, N. Riehl, R. Sizmann, *Z. Phys.* **163** (1961) 594 – 604.
52. S. Shionoya in N. Riehl, H. Kallmann (eds.): *Int. Lumineszenzsymposium über die Physik und Chemie der Szintillatoren,* Thiemig-Verlag, München 1966.
53. H. Kawai, S. Kuboniwa, T. Hoshina, *Jap. J. Appl. Phys.* **13** (1974) 1593 – 1603.
54. F. A. Kröger, *Physica (Amsterdam)* **6** (1939) 369 –379.
55. F. A. Kröger, J. A. M. Dikhoff, *Physica (Amsterdam)* **16** (1950) 297 – 316.
56. F. A. Kröger, N. W. Smit, *Physica (Amsterdam)* **16** (1950) 317 – 328.
57. K. Kynev, V. Kuk, *Z. Naturforsch. A* **34 A** (1979) 262 – 264.
58. M. Tamatani, *Gold Bull.* **13** (1980) 98.
59. A. H. McKeag, P. W. Ranby, *J. Electrochem. Soc.* **96** (1949) 85 – 89.
60. F. A. Kröger, A. Bril, J. A. M. Dikhoff, *Philips Res. Rep.* **7** (1952) 241.
61. W. Lehmann, *J. Electrochem. Soc.* **117** (1970) 1389 –1393.
62. W. Lehmann, *J. Electrochem. Soc.* **118** (1971) 1164 –1166.
63. W. Lehmann, F. M. Ryan, *J. Electrochem. Soc.* **119** (1972) 275 – 277.
64. W. Lehmann, *J. Lumin.* **5** (1972) 87 – 107.
65. W. Lehmann, F. M. Ryan, *J. Electrochem. Soc.* **118** (1971) 477 – 482.
66. R. P. Rao, M. de Murcia, J. Gasiot, *Radiat. Proc. Dosim.* **6** (1984) 64.
67. K. Chakrabarti et al., *Symp. Lumin. Sci. Tech.* **88** (1988) 133.
68. R. P. Rao, *J. Electrochem. Soc.* **132** (1982) 2033 –2034.
69. H. Kasano, K. Megumi, H. Yamamoto, *J. Electrochem. Soc.* **131** (1984) 1953 – 1960.
70. F. Okamoto, K. Kato, *J. Electrochem. Soc.* **130** (1983) 432 – 437.
71. RCA, US 3 418 246, 1968 (M. R. Royce).
72. R. A. Brown, *J. Chem. Educ.* **56** (1979) 732 – 733.
73. O. Kanehisa, T. Kano, H. Yamamoto, *J. Electrochem. Soc.* **132** (1985) 2023 – 2027.
74. C. X. Guo, W. P. Zhang, C. S. Shi, *J. Lumin.* **24/25** (1981) 297 – 300.
75. H. Forest, *J. Electrochem. Soc.* **120** (1973) 695 – 697.
76. Sony Corp., JP 6 222 887, 1987 (S. Hashimoto).
77. Sony Corp., JP 6 222 888, 1987 (S. Hashimoto).
78. G. Tamman: *Lehrbuch der Metallkunde,* Voss, Leipzig, Hamburg 1932.
79. K. H. Butler: *Fluorescent Lamp Phosphors,* The Pennsylvania State University Press, London 1980.
80. T. Y. Tien et al., *J. Electrochem. Soc.* **120** (1973) 278 –281.
81. M. F. Yan, T. C. D. Huo, H. C. Ling, *J. Electrochem. Soc.* **134** (1987) 493 – 498.
82. E. F. Gibbons et al., *J. Electrochem. Soc.* **120** (1973) 835 – 837.
83. J. M. P. J. Verstegen, J. L. Sommerdijk, J. G. Verriet, *J. Lumin.* **6** (1973) 425 – 431.
84. J. M. P. J. Verstegen, A. L. N. Stevels, *J. Lumin.* **9** (1974) 406 – 414.
85. K. Ohno, T. Abe, *J. Electrochem. Soc.* **134** (1987) 2072 – 2076.
86. T. E. Peters, J. A. Baglio, *J. Electrochem. Soc.* **119** (1972) 230 – 236.
87. T. E. Peters, *J. Electrochem. Soc.* **119** (1972) 1720 –1723.
88. T. E. Peters, *J. Electrochem. Soc.* **122** (1975) 98 – 101.
89. in [7]p. 68.
90. A. H. Gomes de Mesquita, A. Bril, *Mater. Res. Bull.* **4** (1969) 643 – 650.
91. J. Shmulovich, G. W. Berkstresser, C. D. Brandle, A. Valentino, *J. Electrochem. Soc.* **135** (1988) 3141 –3151.
92. T. E. Peters, R. B. Hunt, R. G. Pappalardo, F. T. Taubner, *Proc. Electrochem. Soc.* **88** (1988) 199 – 203.
93. J. W. Gilliland, *J. Lumin.* **4** (1971) 345 – 356.
94. L. Thorington, *J. Opt. Soc. Am.* **40** (1950) 579.
95. H. G. Jenkins, A. H. McKeag, P. W. Ranby, *J. Electrochem. Soc.* **96** (1949) 1 – 12.
96. P. D. Johnson in P. Goldberg (ed.): *Luminescence of Inorganic Solids,* Academic Press New York, London 1966, pp. 287 – 336.
97. E. F. Apple, *J. Electrochem. Soc.* **110** (1963) 374 –380.
98. R. Knütter, *Techn.-wiss. Abh. Osram-Ges.* **11** (1973) 269 – 276.
99. U. Müller, R. Schirmer, *Techn.-wiss. Abh. Osram-Ges.* **12** (1986) 444.
100. F. Zwaschka, *Techn.-wiss. Abh. Osram-Ges.* **12** (1986) 470.
101. Kasei Optonix JP 63 264 694, 1988 (S. Fujino, F. Takahashi).
102. J. J. Brown, F. A. Hummel, *J. Electrochem. Soc.* **111** (1964) 1052 – 1057.
103. R. W. Wollentin, C. K. Lui Wei, R. Nagy, *J. Electrochem. Soc.* **99** (1952) 131 – 136.

104. F. C. Palilla, B. E. O'Reilly, *J. Electrochem. Soc.* **115** (1968) 1076 – 1081.
105. F. M. Ryan, *J. Lumin.* **24/25** (1981) 827 – 834.
106. S. T. Henderson, P. W. Ranby, *J. Electrochem. Soc.* **98** (1951) 479 – 482.
107. D. E. Harrison, *J. Electrochem. Soc.* **107** (1960) 217 –221.
108. Y. Kotera, T. Higashi, M. Sugai, A. Ueno, *J. Lumin.* **31/32** (1984) 709 – 711.
109. R. S. Roth, S. J. Schneider, *J. Res. Nat. Bur. Stand., Sect. A.* **64** (1960) 309.
110. H. Wulff, W. Kempfert, G. Herzog, *Wiss. Z. Ernst-Moritz-Arndt Univ. Greifsw., Math. Naturwiss. Reihe* **36** (1987) 30.
111. N. C. Chang, *J. Appl. Phys.* **34** (1963) 3500 – 3504.
112. W. C. Herbert, H. B. Minnier, J. J. Brown, *J. Electrochem. Soc.* **115** (1968) 104 – 105.
113. H. A. Klasens, *Philips Res. Rep.* **9** (1954) 377.
114. L. G. van Uitert, R. R. Soden, R. C. Linares, *J. Chem. Phys.* **36** (1962) 1793 – 1796.
115. L. G. van Uitert, R. C. Linares, R. R. Soden, A. A. Ballmann, *J. Chem. Phys.* **36** (1962) 702.
116. A. K. Levine, F. C. Palilla, *Appl. Phys. Lett.* **5** (1964) 118.
117. S. Faria, D. T. Palumbo, *J. Electrochem. Soc.* **116** (1969) 157 – 158.
118. S. Faria, C. W. Fritsch, *J. Electrochem. Soc.* **116** (1969) 155 – 157.
119. L. H. Brixner, E. Abramson, *J. Electrochem. Soc.* **112** (1965) 70 – 74.
120. S. Faria, E. J. Mehalchick, *J. Electrochem. Soc.* **121** (1974) 305 – 307.
121. R. K. Datta, *J. Electrochem. Soc.* **114** (1967) 1057 –1063.
122. P. Wacher, D. J. Bracco, F. C. Palilla, *Electrochem. Technol.* **5** (1967) 358 – 363.
123. G. Blasse, A. Bril, *J. Lumin.* **3** (1970) 109 – 131.
124. L. H. Brixner, H.-Y. Chen, *J. Electrochem. Soc.* **130** (1983) 2435 – 2443.
125. Du Pont de Nemours, JP 8 028 095, 1980.
126. K. H. Becker, A. Scharmann: *Einführung in die Festkörperdosimetrie,* Thieme Pocket Books, vol. 56, Verlag Karl Thiemig KG, München 1975.
127. J. R. Danby, K. Holliday, N. B. Manson, *J. Lumin.* **42** (1988) 83 – 88.
128. M. R. Mulla, S. H. Pawar, *J. Lumin.* **40/41** (1988) 179 – 180.
129. J. Rudolph, H. Ruffler, *Tech. Wiss. Abh. Osram Ges.* **7** (1958) 223.
130. A. Bril, W. Hoekstra, *Philips Res. Rep.* **16** (1961) 356.
131. D. ülkü, *Z. Krystallogr.* **124** (1967) 192 – 219.
132. P. R. Wonka in H. Ortmann (ed.): *Zur Physik und Chemie der Kristallphosphore,* vol. 2, Akademie-Verlag, Berlin (DDR) 1962.
133. F. A. Kröger: *Some Aspects of the Luminescence of Solids,* Elsevier, Amsterdam 1948.
134. J. J. Borchardt, *J. Chem. Phys.* **42** (1965) 3743.
135. E. Fischer in [2]pp. 151 – 181.
136. A. L. N. Stevels, F. Pingault, *Philips Res. Rep.* **30** (1975) 227 – 290.
137. Siemens, DE-OS 2 642 226, 1978 (H. Degenhardt).
138. Siemens, DE-OS 2 542 481, 1976 (A. L. N. Stevels).
139. Kasei Optonix, JP 1 576 241, 1977 (E. Mori).
140. K. Takahashi, K. Kohda, J. Miyahara, *J. Lumin.* **31/32** (1984) 266 – 268.
141. K. Takahashi, J. Miyahara, *J. Electrochem. Soc.* **132** (1985) 1492 – 1494.
142. H. v. Seggern, T. Voigt, K. Schwarzmichel, *Siemens Forsch. Entwicklungsber.* **17** (1988) 125 – 130.
143. D. M. de Leeuw, T. Kovats, S. P. Herko, *J. Electrochem. Soc.* **134** (1987) 491 – 493.
144. Y. Amemiya et al., International Conference on Biophysics and Synchrotron Radiation, Frascati, July 14 – 16, 1986.
145. R. J. Kurtz, *J. Electrochem. Soc.* **109** (1962) 18 – 24.
146. G. Blasse, A. Bril, *Philips Res. Rep.* **22** (1967) 481 – 504.
147. General Electric, US 3 795 814 1974 (J. G. Rabatin); US 4 070 583, 1978 (J. G. Rabatin).
148. General Electric, US 3 591 516, 1971 (J. G. Rabatin); US 3 607 770, 1971 (J. G. Rabatin).
149. D. M. de Leeuw, C. A. H. A. Mutsaers, H. Mulder, D. B. M. Klaassen, *J. Electrochem. Soc.* **135** (1988) 1009 – 1014.
150. H. Spangenberg: *Leuchtröhrenanlagen für Lichtreklame und moderne Beleuchtung,* Helios-Verlag, Berlin 1955.
151. H. Kullack: *Hochspannungs-Leuchtröhrenanlagen,* Verlag Technik, Berlin 1951.
152. H. Degenhardt, *Electromedica* **3** (1981) 154 – 158.
153. M. Sonoda, M. Takano, J. Miyahara, H. Kato, *Radiology* **148** (1983) 833 – 838.
154. L. H. J. F. Beckmann, A. J. Vermeulen, *Forth European Electro-Optics Conference Sira 1978,* Utrecht SPIE 164, 1978 p. 147 – 150.
155. C. W. Bates in D. A. Garrett, D. A. Bracher (eds.): *American Society for Testing and Materials,* Philadelphia, STP-716 1980, p. 45 – 65.

156. F. Roland, B. Ramm: *Das Röntgenbild: Theorie, Methode, Technik,* Thieme Verlag, Stuttgart 1980.
157. F. Williams in B. Di Bartolo (ed.): *Luminescence of Inorganic Solids,* Plenum Press, New York 1978.
158. P. Keller, *Proc. SID* **24** (1983) 323 – 328.
159. P. J. Bouma: *Farbe und Farbwahrnehmung,* Philips techn. Bibliothek, N. V. Philips, Eindhoven 1951.
160. K. Carl, J. A. M. Dickoff, W. Eckenbach, *Philips tech. Rev.* **40** (1982) 48.
161. F. Eckart: *Elektronenoptische Bildwandler und Röntgenbildverstärker,* Verlag Johann Ambrosius Barth, Leipzig 1962.
162. G. Krokeide in J. D. Sime (ed.): *Safety in the Built Environment,* E & FN Spon Ltd., London-New York 1988, pp. 134 – 146.
163. K. P. Günther, J. Mirow, H. E. Dolle, H. Dörre, *Brandschutz* **3** (1988).
164. G. M. B. Webber, P. J. Hallman in J. D. Sime (ed.): *Safety in the Built Environment,* E & FN Spon Ltd., London-New York 1988, pp. 147 – 157.
165. H. Stübel, *Pflügers Arch. Gesamte Physiol. Menschen Tiere* **141** (1911) 1 – 14.
166. D. Fujimoto, *Biochem. Biophys. Res. Commun.* **84** (1978) 52 – 57.
167. Y. Fukushima, T. Araki, M.-O. Yamada, *Cell. Mol. Biol.* **33** (1987) 725 – 736.
168. J. N. Weaver, *J. Dent. Res.* **55** (1976) 501.
169. D. R. Peplinski, W. T. Wozniak, J. B. Moser, *J. Dent. Res.* **59** (1980) 1501 – 1506.
170. H. Grassmann, H.-G. Moser, E. Lorenz, *J. Lumin.* **33** (1985) 109 – 113.
171. N. Riehl: *Physik. und technische Anwendungen der Lumineszenz,* Springer-Verlag, Berlin 1941.
172. A. Bril, W. van Meurs-Hoekstra, *Philips Res. Rep.* **19** (1964) 296 – 306.
173. E. L. Gooden, C. M. Smith, *Ind. Eng. Chem. Anal. Ed.* **12** (1940) 479 – 482.
174. P. C. Carman, *J. Soc. Chem. Ind. London* **57** (1938) 225 – 234.

General References

175. E. N. Harvey: *Living Light,* Hafner Publishing Company, New York 1965.
176. J. G. Burr (ed.): *Chemi- and Bioluminescence,* Marcel Dekker, New York 1985.
177. A. K. Campbell: *Chemiluminescence, Principles and Applications in Biology and Medicine,* Ellis Horwood Ltd., Chichester and Verlag Chemie, Weinheim 1988.
178. K.-D. Gundermann, F. McCapra: *Chemiluminescence in Organic Chemistry,* Springer Verlag, Berlin 1987.
179. W. Adam, G. Cilento (eds.): *Chemical and Biological Generation of Excited States,* Academic Press, New York 1982.
180. G. Cilento, W. Adam, *Photochem. Photobiol.* **48** (1988) 361 – 368.

Lutetium → Rare Earth Elements
Lysine → Amino Acids

Magnesium

Ketil Amundsen, Norsk Hydro ASA, Oslo, Norway
Terje Kr. Aune, Norsk Hydro ASA, Porsgrunn, Norway
Per Bakke, Norsk Hydro ASA, Porsgrunn, Norway
Hans R. Eklund, Norsk Hydro ASA, Oslo, Norway
Johanna Ö. Haagensen, Norsk Hydro ASA, Porsgrunn, Norway
Carlos Nicolas, Hydro Magnesium, Brussels, Belgium
Christian Rosenkilde, Norsk Hydro ASA, Porsgrunn, Norway
Sia Van den Bremt, Hydro Magnesium, Brussels, Belgium
Oddmund Wallevik , Norsk Hydro ASA, Porsgrunn, Norway

1. Introduction 1
2. Properties 1
3. Raw Materials 2
4. Production 3
4.1. Extraction by Electrolysis 3
4.1.1. Preparation of Magnesium Chloride Cell Feed 3
4.1.1.1. Chlorination of Magnesia and Magnesite 4
4.1.1.2. Dehydration of Aqueous Magnesium Chloride Solutions 5
4.1.1.3. Dehydration of Carnallite 9
4.1.2. Electrolysis 9
4.1.2.1. Technical Data 10
4.1.2.2. Electrolytic Cells 11
4.2. Thermic Reduction Processes . . . 14
4.3. Refining and Casting 16
4.4. Particulate Magnesium 19
5. Environmental Aspects 19
6. Quality Specifications 20
7. Uses 20
8. Economic Aspects 22
9. Toxicology and Occupational Health 24
10. Safety 24
11. References 25

1. Introduction

Magnesium [*7439-95-4*], Mg, A_r 24.3050 [58], is a silvery white metal; it has a valence of two, and its configuration is $1s^2, 2s^2, 2p^6, 3s^2$. The crystal structure is hexagonal close packed; lattice constants at 20 °C are $a = 0.32$ nm, $c = 0.52$ nm. The element has an atomic number of 12 and belongs to Group 2 of the periodic table. It occurs as three isotopes: ^{24}Mg (78.70 %), ^{25}Mg (10.13 %), and ^{26}Mg (11.17 %) [59].

History. Magnesium was recognized as an element by Black in 1755. Magnesium metal was first isolated by Davy in 1808, via electrolysis of anhydrous magnesium chloride with a mercury cathode. Bussy extracted the metal in 1828 by reducing fused magnesium chloride with potassium vapor. In 1833 Faraday electrolyzed dehydrated liquid magnesium chloride to form liquid magnesium and chlorine gas. The first industrial production of magnesium by electrolysis of molten carnallite began in 1886 in Hemelingen (Germany).

In 1940 L. M. Pidgeon pioneered the first industrial metallothermic magnesium extraction plant in Canada; it was based on early German patents in which dolomite was reduced with ferrosilicon under vacuum [1].

2. Properties

Physical properties of magnesium follow (for further data, see [2], [58], [59]):

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_559

mp	650 ± 2 °C	[3]
bp	1090 °C	[58]
Latent heat of fusion	0.37 MJ/kg	[4]
Latent heat of evaporation	5.25 MJ/kg	[5]
Heat of combustion	25.1 MJ/kg	[4]
Specific heat		[59]
at 20 °C	1025 J kg^{-1} K^{-1}	
at 600 °C	1330 J kg^{-1} K^{-1}	
Electrical resistivity		[2]
at 20 °C	4.45 μΩ·cm	
Thermal conductivity		[58]
at 25 °C	156 W m^{-1} K^{-1}	
Linear coefficient of thermal expansion		[3]
at 20 °C	25.2×10^{-6} K^{-1}	
at 20 – 300 °C	$(27-28) \times 10^{-6}$ K^{-1}	[5]
Density (solid)		[58]
at 20 °C	1.738 g/cm^3	
at 650 °C	1.65 g/cm^3	
Density above 650 °C (liquid), g/cm^3	$(1.834-2.647) \times 10^{-4}T$	[6]
Standard electrode potential versus hydrogen	− 2.340 V	[58]

Mechanical properties of magnesium are listed in Table 1. Its dynamic modulus of elasticity is 45 GPa, and its static modulus of elasticity 43 GPa. Pure magnesium is not used for commercial structural applications although the metal has a high damping capacity and is easily machined and formed into shapes by casting or hot forming processes [3].

Chemical Properties. Pure magnesium has a high resistance to corrosion because its galvanic activity is low. Contamination with heavy metals (usually copper, iron, and nickel), chlorides, and oxide or nitride inclusions combined with exposure to chloride-containing solutions on untreated surfaces, strongly promote corrosion due to enhanced galvanic activity. The metal is readily dissolved by most organic and inorganic acids. A protective layer of water-insoluble magnesium hydroxide is formed when magnesium is exposed to moist air or clean water at room temperature. Magnesium is resistant to alkali metal hydroxide solutions, hydrofluoric acid, fluorine, and fluorine compounds (including ammonium hydrogen difluoride) due to the formation of protective hydroxide and fluoride films.

The ability of magnesium to form stable protective oxide, chromate, phosphate, sulfate, and fluoride films is employed commercially in protective coating systems used as primers for paints or as final coatings (→ Magnesium Alloys, Chap. 5.3.). The behavior of magnesium towards various chemicals is listed in [8].

Magnesium burns in air with an intense white flame. The ignition temperature is 645 °C in dry air but decreases with increasing moisture content [5]. The rate of oxidation in moist air at different temperatures is illustrated in Figure 1 [7]. Burning magnesium reacts violently with water. Magnesium fires are extinguished with dry salts or fluxes (see Chap. 10). Magnesium reacts with gaseous chlorine to form $MgCl_2$ and with nitrogen at ca. 500 °C to form Mg_3N_2.

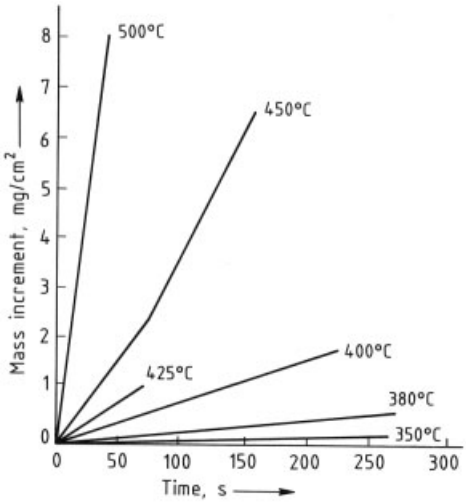

Figure 1. Rate of oxidation of magnesium in moist air [23]

3. Raw Materials [9], [10]

Magnesium is the eighth most abundant element in the earth's crust (average magnesium content, 2.1 %). Magnesium does not occur in nature in elemental form but in the form of compounds in seawater, minerals, brines, and rocks. The element is enriched in ultrabasic rocks.

The prime raw material sources for magnesium extraction are the minerals dolomite [*16389-88-1*] ($CaCO_3 \cdot MgCO_3$), magnesite [*12440-90-3*] ($MgCO_3$), and brucite [*1317-43-7*] ($Mg(OH)_2$); magnesium-rich salts such as carnallite [*1318-27-0*] ($MgCl_2 \cdot KCl \cdot 6\,H_2O$), kieserite [*14567-64-7*]

Table 1. Mechanical properties of magnesium at 20 °C

	Tensile strength, MPa	Tensile yield strength (0.2 %), MPa	Compressing yield strength (0.2 %), MPa	Elongation,∗ %	Brinell hardness∗∗
Sand cast, thickness 13 mm	90	21	21	2 – 6	30
Extrusion, thickness 13 mm	165 – 205	69 – 105	34 – 55	5 – 8	35
Hard rolled sheet	180 – 220	115 – 140	105 – 115	2 – 10	45 – 47
Annealed sheet	160 – 195	90 – 105	69 – 83	3 – 15	40 – 41

∗ In 50 mm.
∗∗ With 500-kg load, 10-mm-diameter ball.

Table 2. Composition of dolomites and magnesites (wt %)

Deposit	MgO	CaO	$Fe_2O_3 + Al_2O_3$	Acid insoluble ∗
Dolomites				
Sørfold, Norway	21.2	30.4	0.1	1.3
Addy, Washington	20 – 21.8	30 – 31.5		1.5
Marignac, France	19 – 20	32 – 34	0.3 – 0.5	0.2 – 0.4
Haley, Canada	21.3	30.7	0.1	0.15
Magnesites				
Liaoning, China	45 – 47	0.5 – 1.0	0.23 – 0.48	1.5 – 2.5
British Columbia, Canada	44 – 46	1.1 – 1.7	0.22 – 0.35	0.3 – 3.5
Queensland, Australia	44 – 47	0.7 – 1.7	0.06 – 0.19	0.5 – 1.0

∗ Mainly SiO_2.

($MgSO_4 \cdot H_2O$), bischofite [*13778-96-6*] ($MgCl_2 \cdot 6\,H_2O$), kainite [*1318-72-5*] ($KCl \cdot MgSO_4 \cdot 3\,H_2O$), and langbeinite [*14977-37-8*] ($K_2SO_4 \cdot 2\,MgSO_4$); and magnesium-rich brines and seawater. Magnesium-rich brines are obtained as byproducts from potash production or from surfacial or underground brine deposits. Analyses of some dolomites, magnesites, and brines used for magnesium extraction are given in Tables 2 and 3, respectively.

Magnesium silicates such as olivine [*1317-71-1*] ($MgFe(SiO_4)_2$), and forsterite [*15118-03-3*] (Mg_2SiO_4) have not been used so far for magnesium extraction. However, a new process has been developed by Noranda for leaching of magnesium from serpentinized material in rejects from asbestos production (asbestos tailings) [60].

Total world production of magnesium in 1999 was ca. 400 000 t, of which 54 % came from dolomite, 29 % from salt or brine deposits, 12 % from magnesite, and 5 % from seawater.

4. Production

Magnesium is produced commercially by electrolysis of magnesium chloride in chloride melts and by metallothermic reduction of magnesium oxide with silicon and/or aluminum. All extraction processes are followed by refining and casting.

4.1. Extraction by Electrolysis

The extraction of magnesium by electrolysis consists of two steps: (1) preparation of the magnesium chloride cell feed and (2) electrolysis.

4.1.1. Preparation of Magnesium Chloride Cell Feed

Cell feeds in industrial use are dehydrated $MgCl_2$ (s) or $MgCl_2$ (l); dehydrated carnallite $MgCl_2 \cdot KCl$ (s); and partially dehydrated $MgCl_2 \cdot 2\,H_2O$ (s). Most cell feeds contain 3 – 8 % alkali metal chlorides and minor impurities of C, SiO_2, MgO, SO_4^{2-}, B, and heavy-metal compounds. Alkalimetal chlorides accumulate in the electrolytic cell as major constituents of the electrolyte. Metallic and nonmetallic impurities are undesirable because they adversely affect the cell performance and the corrosion resistance of the magnesium produced.

Table 3. Composition of brines (wt %)

Brine	Mg^{2+}	Na^{+}	Ca^{2+}	Cl^{-}	SO_4^{2-}
Great Salt Lake ∗	1.1	7.6	0.02	14.1	2.0
Dead Sea	3.6	3.0	1.4	17.8	0.04
Seawater	0.13	1.08	0.04	1.94	0.27
Veendam, Holland ∗∗	8.2	0.15		24	0.12

∗ Composition is subject to seasonal and long-term variations. ∗∗ By solution mining.

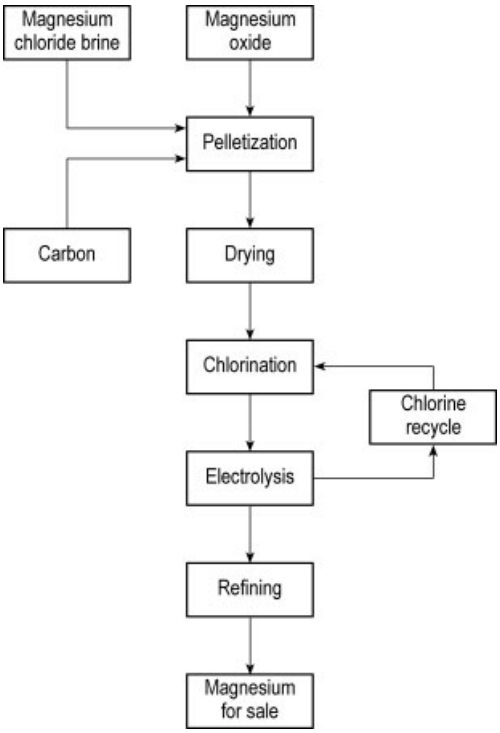

Figure 2. Flow sheet for production of magnesium via chlorination of magnesia (Norsk Hydro)

The use of pure dehydrated magnesium chloride allows coproduction of highly concentrated chlorine gas and electrolysis at high current efficiencies. Dehydrated carnallite (which contains only 50 % $MgCl_2$) leads to accumulation of significant tonnages of potassium-rich electrolyte in the electrolysis. A water content corresponding to $MgCl_2 \cdot 2\,H_2O$ results in high consumption of anode carbon, formation of sludge, and higher power consumption in the cells.

The two main routes to the production of dehydrated magnesium chloride cell feed are: (1) chlorination of magnesia [*1309-48-4*] (MgO) or magnesite [*12440-90-3*] ($MgCO_3$) in the presence of carbon or carbon monoxide and (2) dehydration of aqueous magnesium chloride solutions (brines) or hydrous carnallite ($MgCl_2 \cdot KCl \cdot 6\,H_2O$).

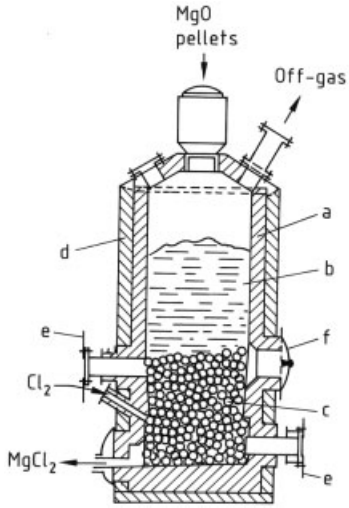

Figure 3. The IG chlorinator
a) Refractory material; b) Charge; c) Carbon resistors; d) Steel mantel; e) Carbon electrodes; f) Door for slag removal

4.1.1.1. Chlorination of Magnesia and Magnesite

IG Farben Process (see Fig. 2). The principal process used industrially to chlorinate magnesia is the IG Farben process in operation at Norsk Hydro's Porsgrunn plant in Norway until 2002. In this process, caustic magnesium oxide (lightly burned), produced from seawater and calcined dolomite (→ Magnesium Compounds, Chap. 4.2.2.), is mixed with char (low-ash coke from lignite) and magnesium chloride brine on a rotating disk to form pellets with a diameter of 5 – 10 mm. Hydrated magnesium oxide and magnesium oxychlorides act as binders. After drying, pellets containing ca. 50 % magnesium

oxide, 15 – 20 % magnesium chloride, 15 – 20 % water, 10 % carbon, and a balance of alkali metal chlorides are conveyed to the chlorinators (Fig. 3).

The lower third of the brick-lined cylindrical shaft furnace is filled with carbon blocks (c) that act as resistors and are heated by carbon electrodes (e). Chlorine produced during subsequent electrolysis of magnesium chloride is introduced in the resistor-filled zone. The charge (b) resting on the resistor bed reacts with chlorine at 1000 –1200 °C. The reaction mechanism is complex and involves several side reactions. The main reactions are:

$$2\,MgO + C + 2\,Cl_2 \longrightarrow 2\,MgCl_2 + CO_2$$

$$Cl_2 + C + H_2O \longrightarrow 2\,HCl + CO$$

$$MgO + 2\,HCl \longrightarrow MgCl_2 + H_2O$$

Molten magnesium chloride is collected in the bottom of the furnace from where it is tapped and transported to the electrolytic cells in sealed containers.

Off-gases at 100 – 200 °C containing air, CO, CO_2, traces of HCl, Cl_2, SO_2, H_2S, and chlorinated hydrocarbons are scrubbed, and the remaining contaminants are incinerated before release to the stack. The wash water is filtered and chemically treated in several stages before release to the sewer. Slag rich in magnesium silicate may be formed due to the presence of silica in the magnesia and must be removed from the furnace at intervals.

The magnesium yield is ca. 90 %, and the carbon consumption is 0.4 t per tonne of magnesium. Magnesium chloride solution added to the pelletization stage compensates for chlorine losses. The anhydrous $MgCl_2$ product typically contains less than 0.1 % MgO, 0.1 % C, 0.1 % SiO_2, and 20 ppm B. Each chlorinator has an equivalent annual magnesium production capacity of ca. 2000 t.

In the electrolysis-based magnesium plants in China, similar chlorinators are used, charged with magnesite and coke.

The MPLC process was developed by MPLC [20] for the chlorination of raw magnesite with the use of carbon monoxide gas (CO). In the late 1980s MagCan built a plant in Alberta, Canada, based on this process. The main reactor was brought into production, but other operational problems were experienced. The costs related to modifications and optimization were significant. These factors, combined with a soft market, led to the plant's closure in 1991.

The main features of the process are: Natural magnesite in lump form is charged at intervals to a reactor of the same general description as the IG shaft furnace (Fig. 3), but taller. Chlorine from the electrolysis is mixed with fresh makeup chlorine and carbon monoxide from a gas generator and fed into the lower section of the reactor, where it reacts with preheated magnesite resting on a bed of carbon resistors. The main reactions are:

$$MgCO_3\,(s) + CO\,(g) + Cl_2\,(g) \longrightarrow MgCl_2\,(l) + 2\,CO_2\,(g)$$

$$CO_2\,(g) + O_2\,(g) + 3\,C\,(s) \longrightarrow 4\,CO\,(g)$$

During the reaction, the evolution of CO_2 from the calcination maintains an open passage for chlorine and carbon monoxide to penetrate the lumps, while magnesium chloride simultaneously drains from the surface [19].

The process is in thermal balance at 1000 °C which is also the temperature in the reaction zone. Off-gases leaving the top of the reactor at 150 – 350 °C contain primarily carbon dioxide and small amounts of carbon monoxide, hydrogen chloride, and unconverted chlorine. Molten magnesium chloride is tapped from the reactors at 800 °C.

4.1.1.2. Dehydration of Aqueous Magnesium Chloride Solutions

Brines with a high content of magnesium chloride may be derived from various sources: From solution mining of salt deposits, e.g., bischofite ($MgCl_2 \cdot 6\,H_2O$); as a byproduct from the potassium industry; or produced by dissolving or leaching magnesium-bearing minerals in hydrochloric acid (magnesite in the Norsk Hydro process, asbestos tailings in the Magnola process, precipitated magnesium hydroxide in the Dow Chemical process). Naturally occuring dilute brines are concentrated by solar evaporation (Magcorp process). Unwanted impurities in the brines (e.g., Fe, Ni, SO_4^{2-}, B, Br_2) are removed prior to dehydration.

Norsk Hydro Process (see Fig. 4). The Norsk Hydro dehydration process was developed to produce anhydrous magnesium chloride in the Porsgrunn plant from brines with 33 % $MgCl_2$, a byproduct from the German potash industry. Brine-based magnesium production in Porsgrunn was shut down in 1991 due to increasing brine costs.

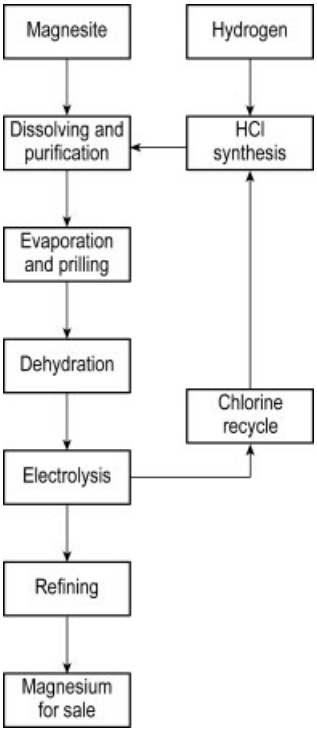

Figure 4. Flow sheet for production of magnesium via dehydration of magnesium chloride brine (Norsk Hydro)

In the Norsk Hydro magnesium plant in Bécancour, Québec, Canada, started up in 1989, $MgCl_2$ brine is produced by dissolving magnesite in hydrochloric acid. Impurities such as heavy metals are removed by precipitation and filtration [103], [104]. Purified brine is concentrated by evaporation to ca. 50 % magnesium chloride before prilling. The granulometry of the prills is kept within close tolerances, and their shape is controlled to optimize the subsequent dehydration step in fluidized beds, employing hot air at 150 – 180 °C and finally hydrogen chloride gas at 300 – 400 °C [21]. Water and magnesium chloride dust in off-gases from the hydrogen chloride dehydration step are absorbed in concentrated magnesium chloride solution [105]. Essentially dry hydrogen chloride gas is heated and returned to the hydrogen chloride dehydration step. Magnesium chloride prills containing < 0.1 % magnesium oxide are transported pneumatically to the electrolysis cells. The chlorine gas produced in the electrolysis is recycled by reaction with hydrogen in HCl synthesis furnaces to produce the hydrochloric acid used in the dissolving of magnesite. Energy is recovered from the HCl synthesis by raising steam [116].

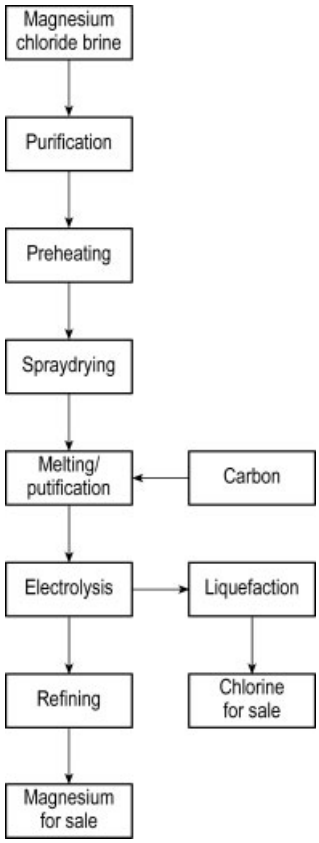

Figure 5. Flow sheet for production of magnesium via dehydration of brine (Magcorp)

Magcorp Process (see Fig. 5). The magnesium plant at the Great Salt Lake, Utah, was started up in 1972 and is now owned and operated by Magnesium Corporation of America (Magcorp). Great Salt Lake water is concentrated in solar evaporation ponds to precipitate potassium chloride, sodium chloride, and some carnallite ($MgCl_2 \cdot KCl \cdot 6\,H_2O$). The resulting

crude brine contains 7.5 % Mg, 4 % SO_4^{2-}, and < 1 % K and Na, and is desulfated with calcium chloride solution generated in the process. The boron content is decreased from 500 to 3 ppm by liquid extraction with kerosene. The brine is preheated before spray drying by direct contact with hot off-gases from the spray dryers. Each spray dryer is connected to a turbine fired with natural gas. Exhaust gases enter the top of the spray dryers at 515 °C. The energy efficiency of this system is ca. 80 %.

The spray-dried powder contains ca. 4 % water and 4 % magnesium oxide. It is melted in an electrically heated, brick-lined melt cell and treated with carbon and chlorine gas to lower the content of magnesium oxide, water, bromine, sulfate, aluminum, and heavy metals. In a second reactor vessel, impurities are further reduced with chlorine gas in the presence of carbon at 830 °C to an average magnesium oxide level of 0.05 %. Iron(III) chloride acts as a catalyst in these reactions.

Off-gases containing chlorine, hydrogen chloride, sulfur dioxide, magnesium chloride dust, and chlorinated hydrocarbons are scrubbed with milk of lime to form the calcium chloride used in the desulfatation stage. The molten magnesium chloride product is transported to the electrolytic cells in trucks and contains 95 % $MgCl_2$, < 0.1 % MgO, and 0.01 % SO_4^{2-}, the balance being alkali metal chlorides. The process is a net producer of chlorine [22].

Dow Chemical Process (see Fig. 6). Dow Chemical started magnesium production in Freeport, Texas, in 1941. Dow was for a long period of time the largest magnesium producer in the world. However, the Freeport magnesium operation was closed down in 1998. The process has been licensed to a prospective magnesium producer in Australia [61].

In the Dow process, hydrous magnesium chloride ($MgCl_2 \cdot 2\,H_2O$) is used as cell feed. Magnesium hydroxide slurry precipitated from seawater with sodium hydroxide or calcined dolomite is dewatered on rotary filters to 8 – 12 wt % magnesium, dispersed in magnesium chloride solution, and neutralized with hydrogen chloride. Sufficient sulfuric acid is added to precipitate excess calcium, which is then removed by filtration. Sulfate ions are adjusted by a proprietary process. The brine is further concentrated to 35 % magnesium chloride. The purified brine is fed to a dryer where it is concentrated further to 70 % magnesium chloride, which is a dry granular solid. Alternatively, the 35 % brine can be concentrated directly to 70 wt % solid particles in a spouting bed where the brine is sprayed on solidified particles in an airstream at 180 °C. Undersized particles are separated in cyclones and returned to the dryer.

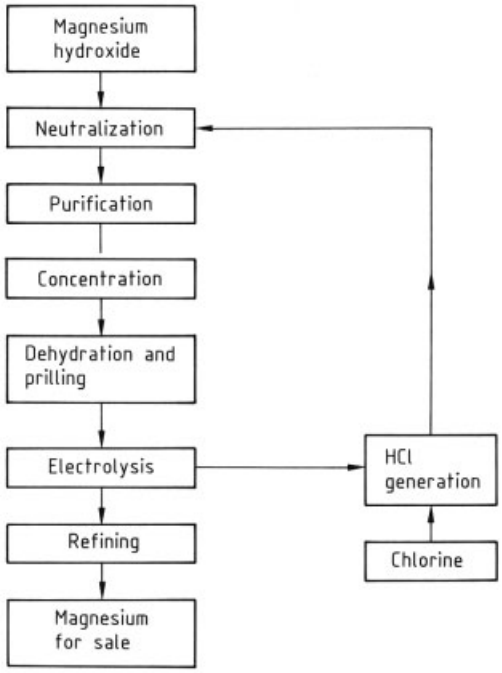

Figure 6. Flow sheet for production of magnesium from seawater (Dow Chemical)

The hydrogen chloride used for neutralization of the magnesium hydroxide is partly recovered from the anode gases and partly added to the hydrogen chloride furnaces as makeup chlorine in various forms.

Magnola Process. Noranda Inc., Canada, has developed this process, which is implemented in the Magnola Metallurgy magnesium plant started up in Danville, Quebec, Canada, during 2000 – 2001 [60]. The raw material is asbestos tailings, i.e., mining residues from the asbestos industry, containing up to 24 % Mg in serpentine minerals. The raw material may typically contain 40 % MgO, 38 % SiO_2 and 5 % iron compounds. The tailings are first subjected to a magnetic separation step to remove most of

the iron. The material is then leached with concentrated hydrochloric acid (33 % HCl), and the brine produced is neutralized in two steps with controlled pH to avoid the formation of silica gel while precipitating other impurities. Leaching residues and precipitated impurities are removed by filtration; the resulting purified brine contains 27 % $MgCl_2$.

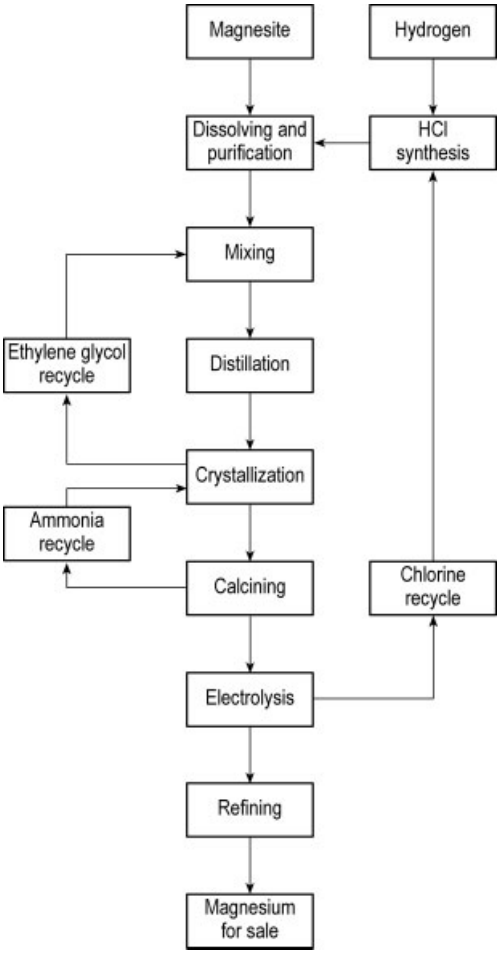

Figure 7. Flow sheet for production of magnesium via magnesium hexammoniate (QMC)

The brine is fed to a fluidized granulation and dryer bed, which produces free-flowing granules of hydrated magnesium chloride $MgCl_2 \cdot x\,H_2O$ with $x \approx 2$ and a content of ca. 2 % MgO. The final dehydration and chlorination are carried out in so-called super-chlorinators, in which the granules are melted in a bath of molten electrolyte recirculated from the electrolysis cells [62]. HCl gas is sparged into the super-chlorinators, reacts with MgO, and removes the water of hydration from the melt. The HCl gas is produced by reacting chlorine gas from the electrolysis with hydrogen in HCl synthesis furnaces, and the HCl-containing off-gas from the super-chlorinators is recycled as hydrochloric acid for leaching of the tailings.

QMC Process (see Fig. 7). Queensland Metals Corp. (QMC) of Queensland, Australia, has developed this process in cooperation with CSIRO of Australia, primarily for the exploitation of the huge cryptocrystalline magnesite deposits in Queensland (Kunwarara, Yaamba). This process is a further development of the Nalco process [18] [63].

Magnesite is dissolved in hydrochloric acid in two or more stages with controlled pH at each stage [64]. The resulting magnesium chloride brine is converted to anhydrous magnesium chloride cell feed for magnesium electrolysis by the following process steps [65]: The hydrous magnesium chloride brine is mixed with ethylene glycol to form an ethylene glycol magnesium chloride hydrate solution. This solution is subjected to distillation under vacuum to remove all or most of the water, resulting in a dehydrated magnesium chloride glycolate ($MgCl_2 \cdot 3\,HOCH_2CH_2OH$) dissolved in ethylene glycol. The dehydrated solution and anhydrous ammonia gas are introduced separately and simultaneously into a crystallizer containing ethylene glycol saturated with ammonia, whereby crystals of magnesium chloride hexammoniate ($MgCl_2 \cdot 6\,NH_3$) are precipitated. The hexammoniate precipitate is separated from the liquid phase and washed with methanol saturated with ammonia. Traces of methanol are removed by heating the crystals to a temperature of max. 120 °C. The hexammoniate crystals are then calcined in a fluidized bed at 400 °C or more to produce anhydrous magnesium chloride and release ammonia. The ammonia gas from the calciner is recycled to the crystallization stage, and the liquid streams from the crystallizer and the washings are processed to recover ethylene glycol, methanol, and ammonia for recycling. Salts other than $MgCl_2$ dissolved by the hydrochloric acid, e.g., $CaCl_2$, accumulate in the recycled ethylene glycol and can be removed by treat-

ment of the recycle stream. The hydrochloric acid for dissolving is produced by reacting chlorine from the magnesium electrolysis with hydrogen in HCl synthesis furnaces.

An alternative dissolving process has been proposed to avoid water in the dissolving steps by leaching magnesium from partially calcined magnesite with ammonium chloride dissolved in ethylene glycol and thereby reducing the energy consumption in the overall process [66].

4.1.1.3. Dehydration of Carnallite

Carnallite cell feed for magnesium production is used in the former Soviet Union (Russia, Ukraine, and Kazakhstan) and in Israel. Carnallite ($MgCl_2 \cdot KCl \cdot 6\,H_2O$) is obtained as a natural mineral, from solar evaporation of Dead Sea waters, or as a byproduct from the production of potassium chloride from complex salt deposits. Dehydration of carnallite is simpler than that of aqueous magnesium chloride because hydrolysis is weaker.

According to [23], impurities in the carnallite must not exceed the following levels: $CaSO_4$ 0.06 %, B 0.001 %, Fe 0.01 %, and SiO_2 0.01 %. The sodium chloride content may be 5 – 7 %.

Carnallite is dehydrated in two stages. The first stage is performed in a fluidized bed at 160 – 200 °C to give a water content of 3.7 – 4.8 %. The second stage, which yields molten dehydrated carnallite, is performed in a chlorinator at 700 – 750 °C. The chlorinator combines melting, chlorination, and settling – mixing in a single unit. The final product contains 49 – 51 % magnesium chloride and 0.6 – 1 % magnesium oxide. Efforts are being made to reduce the degree of hydrolysis during dehydration and to replace the chlorination stage by hydrogen chloride treatment in a fluidized bed. With carnallite as the raw material, about 4.5 t of excess electrolyte is produced as a byproduct per tonne of Mg; this byproduct with ca. 80 % KCl content is preferably used as a fertilizer.

4.1.2. Electrolysis [7]

Magnesium chloride is electrolyzed in a molten mixture with alkali metal chlorides at 660 – 800 °C, the operating window being limited on the low side by the freezing point of magnesium and on the high side by its rapidly increasing vapor pressure. The main electrochemical reaction is:

$$MgCl_2\,(l) \longrightarrow Mg\,(l) + Cl_2\,(g)$$

The electrolyte is contained in brick-lined vessels or unlined steel casings. For most electrolytes used, the magnesium metal collects on top of the electrolyte because of its lower density. Industrial electrolytic cells differ in electrode configuration, flow pattern of the electrolyte, and collection system for reaction products. Steel cathodes and graphite anodes are arranged vertically or at an angle. Electrodes may be monopolar, bipolar, or a combination of both, with interpolar gaps ranging from 3 to 12 cm and current densities from 2000 to 8000 A/m^2.

Electrolyte filled with chlorine gas adjacent to the anodes has a lower density than the bulk electrolyte, which results in circulation. The chlorine gas is separated from the circulating electrolyte, which then carries magnesium globules to the metal-collecting area. Salts in the cell feed of metals less noble than magnesium, e.g., alkali metal chlorides and calcium chloride, accumulate in the cell and have to be removed from time to time.

Alternative electrode materials have been the subject of research and development efforts, but not commercialized. A patent [36] covers inert anodes protected by a ceramic cover of doped and sintered Ti^{IV} and Ti^{III} compounds. Experiments have been carried out with CuAl bronze as an inert anode material [67]. Also anodes made of metal carbides, metal borides, metal nitrides, tantalum, tantalum oxides, and barium titanates have been tested [68]. Addition of H_2 to the anodes has been proposed so that the cells produce HCl rather than Cl_2 [69], [70]. Cathodes made of or coated with molybdenum have been proposed to avoid cathode wear [71].

Electrolysis of magnesium oxide dissolved in various chloride and fluoride electrolytes has been considered. Considerable work at the U.S. Bureau of Mines in the 1940s revealed the potential of and problems with the electrolysis of magnesium oxide [72]. This line of development has been pursued in several directions, with emphasis on fluorides [11], [12], rare

earth chloride or fluoride additions [73–76], and oxide conducting membranes [77]. Composite MgO – C anodes have been tested in fluoride, fluoride – chloride, and chloride electrolytes on a laboratory scale [13]. However, the solubility of magnesium oxide in all investigated electrolytes is low, and practical solutions for metal collection in industrial operation have not been found.

Electrolysis of magnesium chloride in aqueous solution liberates hydrogen, not magnesium, at the cathode. Attempts to electrolyze magnesium salts in organic solvents have not been successful.

4.1.2.1. Technical Data

The free energy and enthalpy of decomposition of $MgCl_2$ to Mg and Cl_2 at 1000 K are 482.7 and 596.7 kJ/mol, respectively. The corresponding decomposition potentials are 2.501 and 3.092 V, the former being the reversible decomposition potential and the latter the isothermal decomposition potential ($\Delta H/2F$). However, in a technical electrolyte where $MgCl_2$ is mixed with other alkaline earth metal chlorides and alkali metal chlorides, the reversible decomposition potential is higher due to the lower activity of $MgCl_2$. The activity varies from electrolyte to electrolyte. A typical $MgCl_2$ activity in a technical electrolyte is 0.02, yielding a reversible decomposition potential at 1000 K of 2.67 V. According to Faraday's law, 4.534×10^{-4} kg Mg is produced per A · h at a current efficiency of 1.

Theoretical energy consumption under thermoneutral conditions is 6.8 kW · h/kg Mg. In industrial cells, however, electrical energy consumption may range from 10 to 20 kW · h/kg. Correspondingly, 32 – 66 % of the energy input is lost as heat as a result of ohmic resistance in the electrodes and electrolyte, overvoltages on the electrodes and recombination of magnesium and chlorine. Heat generation is necessary to compensate for heat losses from the electrodes and the cell surface. Heat balances for cells are discussed in [7]. Current efficiencies normally range from 0.75 to 0.95. The main cause of reduced current efficiency is the recombination of dissolved chlorine and magnesium in the electrolyte.

The composition and main characteristics of electrolytes used in the electrolysis are listed in Table 4; they normally contain 8 – 20 wt % magnesium chloride. Other chlorides ($BaCl_2$ or LiCl) may be added to influence the density; lithium chloride also increases the conductivity of the melt. The density of the lithium – potassium chloride electrolyte is lower than that of liquid magnesium, permitting magnesium to collect at the bottom of the cell. The solubilities of magnesium and chlorine in commercial electrolytes are of the order of 0.005 wt % and 6×10^{-7} mol/cm^3, respectively. Solubilities increase with increasing temperature [24].

Comprehensive measurements of the characteristics of electrolyte compositions and the pure salts are reported in [25–31]. At 750 °C the standard decomposition voltages of the electrolyte components are: $MgCl_2$ 2.51 V, NaCl 3.22 V, KCl 3.27 V, LiCl 3.30 V, $CaCl_2$ 3.33 V, and $BaCl_2$ 3.40 V. Some sodium is always deposited into the magnesium metal phase, and the sodium content in the metal depends on the concentration of $MgCl_2$ and the cell design. With magnesium chloride contents below ca. 5 %, the current efficiency may be substantially reduced due to recombination of sodium and chlorine, which causes a temperature rise.

Impurities in the cell feed disturb cell performance by passivating the cathode surface and hindering the coalescence of metal globules or their free rise to the surface. Upper limits for impurities in cell feed, in parts per million, have been indicated in [7], [32], as shown in Table 5.

Table 5. Upper permissible limits for some impurities in cell feed (in ppm)

Impurity	Ref. [7]	Ref. [32]
MgO	5000	2000
C	n.d.∗	1500
H_2O	1000	1000
Mn	600	600
Fe	400	200
S	120	100
Si	n.d.	100
P	n.d.	50
B	20	10
Ti	80	n.d.

∗ n.d. = no data.

Small additions of fluoride, in the form of CaF_2 or NaF [7], [78], and amorphous carbon to

Table 4. Composition and properties of electrolytes

Electrolyte	Composition, wt %	*mp*, °C	Properties at 700 °C			
			Density, kg/m^3	Conductivity, S/m	Viscosity, mPa·s	Surface tension, mN/m
Potassium	5 – 12 $MgCl_2$ 70 – 78 KCl 12 – 18 NaCl	650	1600	183	1.35	104
Sodium – potassium	10 $MgCl_2$ 50 NaCl 40 KCl	625	1625	200	1.58	108
Sodium – calcium	8 – 16 $MgCl_2$ 30 – 40 $CaCl_2$ 35 – 45 NaCl 0 – 10 KCl	575	1780	200	2.22	110
Lithium – potassium	10 $MgCl_2$ 70 LiCl 20 KCl	550	1500	420	1.20	
Lithium – sodium	10 $MgCl_2$ 70 LiCl 20 NaCl	560	1521	488		
Sodium – barium	10 $MgCl_2$ 20 $BaCl_2$ 50 NaCl 20 KCl	686	1800	217	1.70	110
Magnesium		650	1580			

the electrolyte are reported to counteract these undesired effects.

Magnesium chloride and magnesium react with oxygen in air or with oxygen-containing compounds in the cell feed and the brick lining to form magnesium oxide, which has a low solubility in the melt. The oxide can form a layer on the cathode surface, resulting in higher electrical resistance. Magnesium oxide with entrapped electrolyte and magnesium globules form a sludge at the bottom of the cell which must be removed at intervals. The amount of sludge generated is significantly lower in sealed as opposed to open cell designs.

Theoretically, 2.917 t of chlorine is produced per tonne of magnesium. The chlorine yield from commercial cells varies, depending on design, but can be as high as 2.9 t of chlorine per tonne of magnesium at concentrations above 98 wt %. Commercial cells operating with yields of 2.7 – 2.85 t of chlorine are reported [79],[80]. The chlorine gas from the cells contains small amounts of air, dust from the electrolyte components, and minor amounts of chlorinated hydrocarbons, which require efficient systems for collection and destruction. Electrolytic magnesium contains typically $\geq$ 99.8 wt % magnesium.

4.1.2.2. Electrolytic Cells

IG Cell. The cell concept developed by IG Farben Industrie in Germany in the 1930s was used extensively in the Soviet Union, Norway, and the United States, but has now been replaced by new developments. Figure 8 illustrates the electrode configuration and the circulation of the electrolyte [32].

The brick-lined cell is divided into compartments by semisubmerged refractory partition walls (c). Cooled graphite anode plates (a) are installed and tightly sealed in the refractory cover of the cell. The semi walls on each side of the anodes separate the magnesium metal and the chlorine gas. Steel cathodes (b) are installed in the cathode compartments from above or through

the sidewalls. The metal is collected from each of several cathode compartments and chlorine from each anode compartment. Extensive cathode chamber ventilation causes oxygen and water in the ventilation gases to react with metal and electrolyte to form sludge. The current efficiency is typically 0.80 – 0.85, and the amperage is in the range 60 – 120 kA. The chlorine concentration may reach 90 – 95 wt % and is limited by air leaks through the semi walls. Anode consumption is 15 – 20 kg per tonne of magnesium. The power consumption per kilogram of magnesium is 16 – 18 kW·h when operated on molten magnesium chloride. For further details of the IG cell, see [33].

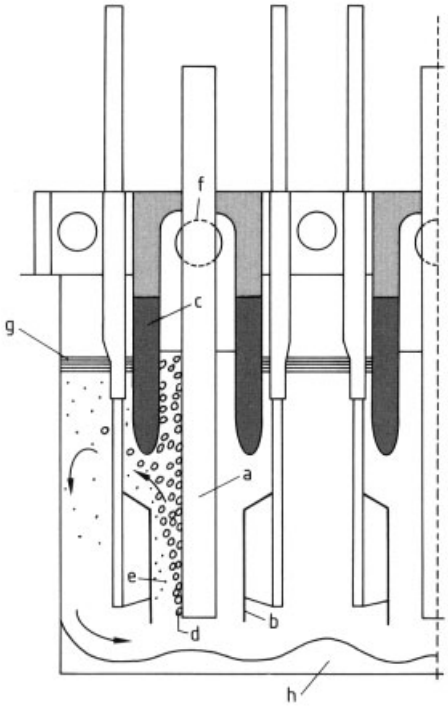

Figure 8. Electrode configuration in the IG cell
a) Graphite anode; b) Steel cathode; c) Semi wall; d) Chlorine bubbles; e) Magnesium globules; f) Chlorine outlet; g) Magnesium; h) Sludge

Norsk Hydro Cell. This cell has been in operation since 1978 and consists of a sealed, brick-lined body that is divided into two separate chambers for electrolysis and metal collection (Fig. 9). Densely packed, cooled graphite anode plates (b) are installed through the roof and double-acting steel plate cathodes (c) through the back wall. Chlorine (> 98 wt %) is collected from a central pipe in the anode compartment. Circulation of the electrolyte (h) is parallel to the electrodes bringing the magnesium metal to the collecting chamber from where it is extracted by vacuum and transported to the foundry. This cell operates on solid or liquid cell feed conveyed continuously or at intervals to the cell at 700 – 720 °C with a current load of 340 – 430 kA. Energy consumption is 13 – 13.5 kW·h per kilogram of magnesium and cell life exceeds five years [81].

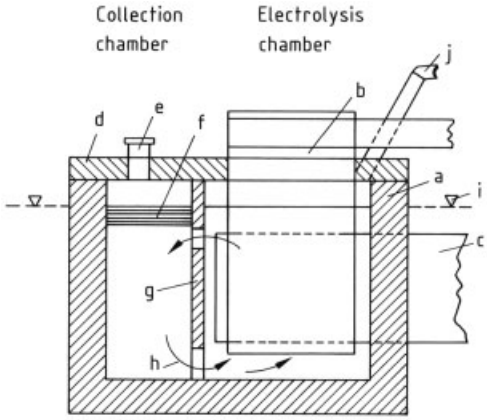

Figure 9. Norsk Hydro cell
a) Refractory material; b) Graphite anode; c) Steel cathode; d) Refractory cover; e) Metal outlet; f) Metal; g) Partition wall; h) Electrolyte flow; i) Electrolyte level; j) Chlorine outlet

VAMI and Titanium Institute Cells. Cell types developed by VAMI (Russian National Aluminium and Magnesium Institute), St.-Petersburg, and the Titanium Institute (State Research and Design Titanium Institute), Zaporozhye, Ukraine, are based on the same principles as the Norsk Hydro cell but operate at 150 – 180 kA and 700 – 740 °C. When operated on dehydrated molten magnesium chloride obtained after reduction of titanium(IV) chloride with magnesium (Kroll process, → Titanium, Titanium Alloys, and Titanium Compounds), the current efficiency is ca. 0.80, with an electric energy consumption of 13.5 kW·h per kilogram of magnesium [23], [79], [80].

A special feature is the hydraulic linking of the cells in a cell room, allowing electrolyte to flow from cell to cell in a "flow line". Cell feed is supplied to the first cell in the line together with

recycled electrolyte, and metal is separated and extracted from the last cell in the line. Significant improvements in performance and productivity are reported [79], [80].

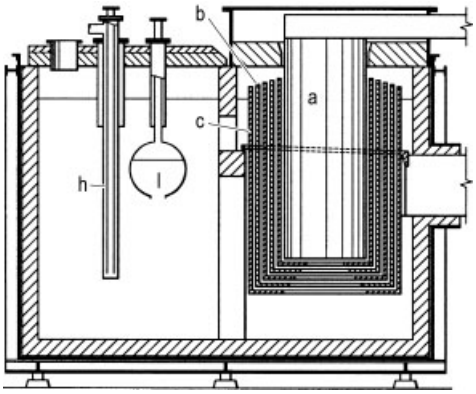

Figure 10. Cross-section view of the Alcan multipolar Ex series cell [83]
a) Anode; b) Bipole electrodes; c) Cathode; h) Heat exchanger; l) Gas ballast to control electrolyte level

Alcan Multipolar Cell. Alcan Aluminium Limited, Canada, was a magnesium producer in the 1950s, but ceased primary magnesium production in 1959 and has been engaged in technology development since then [34]. Magnesium electrolysis cell designs have been developed in cooperation with a titanium plant (Kroll process) in Japan. New multipolar cell designs have been in operation since 1983 [82]. In such cells a number of bipolar electrodes are placed between the anode and the cathode surfaces. The prime advantage of bipolar cells is the reduction of voltage losses, which in combination with narrow anode – cathode gaps results in an operating cell with very low energy consumption. Special provisions are made to prevent current leaks from bypassing the bipolar electrode system.

The latest Alcan multipolar cell design is the Ex series (Fig. 10) [83]. The average energy consumption of this cell, when operating on high-purity molten magnesium chloride from Kroll titanium production, is 10 – 11 kW · h/kg Mg. The cell operates at 90 – 150 kA with a cell life of about two years [82]. The latest installation of this cell was by Magnola at Danville, Québec, where the cell is connected to Magnola's superchlorinator [60]. The annual output is projected to exceed 2500 t of magnesium per cell.

Ishizuka Cell. The Ishizuka Research Institute, Japan, has developed a bipolar cell that has been used since 1983 for industrial magnesium production in the Japanese titanium industry [35].

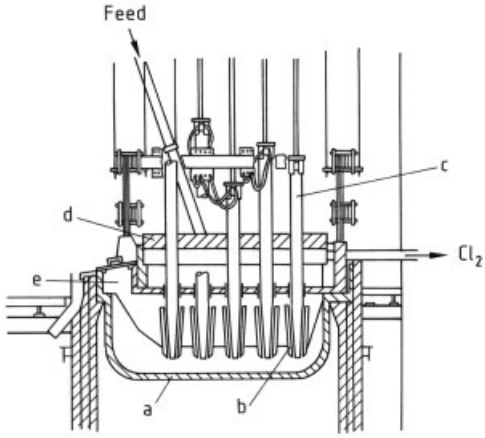

Figure 11. Dow cell
a) Steel mantel; b) Steel cathode; c) Graphite anode; d) Refractory cover; e) Magnesium collection compartment

Dow Cell. The Dow cell for hydrous cell feed ($MgCl_2 \cdot 2\,H_2O$), shown in Figure 11 was in operation at Freeport, Texas, until 1999. The technology was later licensed to an Australian project [61]. The graphite anodes can be lowered into the cells to compensate for their consumption. Densely packed cylindrical anodes (c) are led through openings in the refractory cell hood (d) and sealed to reduce air leaks. They are centered by insulating ceramic spacers. The development of efficient anode seals to withstand severe corrosive environment and high temperature has been a challenge. Conical cathodes (b) are welded to the steel shell (a) holding the electrolyte. Magnesium metal is conducted by the electrolyte flow to a collecting chamber (e) at the front of the cell. Water in the cell feed is partly flashed off, and partly reacts with $MgCl_2$ to form HCl and MgO. The MgO can react with carbon in the graphite to form CO and CO_2. Graphite consumption is probably 60 – 70 kg per tonne of magnesium. The Dow cell is operated at 700 °C with an estimated current load of 180 kA and current efficiency of ca. 0.8. In comparison with an anhydrous cell feed, the hydrous cell feed increases electric power consumption as well as the formation of sludge and dilute anode gases.

4.2. Thermic Reduction Processes [37], [38], [84]

Carbothermic reduction of magnesium oxide

$$MgO(s) + C(s) \longrightarrow Mg(g) + CO(g)$$

appears attractive but is not used industrially. The main problem is the high reaction temperature (1800 – 2000 °C) and the need for rapid cooling (quenching) of the reaction gases to suppress the reverse reaction back to magnesium oxide [14]. An industrial plant based on carbothermic reduction was operated in the United States during World War II, using natural gas as the quenching medium. The reaction product contained only 50 % Mg metal and had to be processed in a sublimation step to obtain pure Mg metal. In later experimental studies, quenching has been based on rapid adiabatic expansion of the gases [15] or cooling in liquid metal [84].

Calcium carbide was used as a reductant for MgO in England before and during World War II [85].

Industrial *metallothermic processes* for magnesium production are based on the reduction of magnesium oxide with silicon and/or aluminum. Magnesium oxide is provided in the form of calcined dolomite (MgO·CaO), sometimes enriched with calcined magnesite (MgO). The basic reaction of calcined dolomite with silicon is

$$2\,MgO \cdot CaO(s) + Si(s) \longrightarrow 2\,Mg(g) + Ca_2SiO_4(s)$$

This reaction is highly endothermic ($\Delta H \approx 230$ kJ/mol Mg). The Mg vapor pressure reaches 0.1 MPa at 1470 °C. However, in industrial operations, Si is used in the form of an Si-rich ferrosilicon alloy. As the reaction progresses, the amount of Si in the alloy is reduced. Spent FeSi typically contains 20 wt % Si. An Mg pressure of 0.1 MPa is reached at ca. 1700 °C when dolime and 20 % FeSi are in equilibrium. If the dolime is mixed with other oxides, such as Al_2O_3 and SiO_2, to enhance formation of a liquid slag, the temperature required to reach an Mg vapor pressure of 0.1 MPa may increase further. Industrial processes operate at reduced pressures to allow operation at lower temperatures (1200 – 1600 °C), thus reducing the wear on construction materials and suppressing the formation of gaseous species other than Mg.

When aluminum is used as the sole reductant, magnesium spinel will be formed [86]:

$$4\,MgO(s) + 2\,Al(l) \longrightarrow 3\,Mg(g) + MgAl_2O_4(s)$$

Aluminum is a stronger but more costly reductant than silicon ($\Delta H \approx 165$ kJ/mol Mg). An Mg vapor pressure of 0.1 MPa can be achieved at a temperature of ca. 1360 °C.

Reduction is carried out in batch mode due to vacuum operation. Work has been aimed at developing atmospheric processes, allowing continuous operation and collection of magnesium in liquid form [84], [86], [87].

Three principles are applied for the industrially reaction chamber:

1) Externally heated retorts producing 70 kg/d (Pidgeon)
2) Internally heated reactors producing 2 t/d (Bolzano/Brasmag)
3) Internally heated reactors with molten slag producing ca. 12 t/d (Magnetherm)

Pidgeon Process [39–42], [88]. The process developed by PIDGEON in the early 1940s is presently used by Timminco in Canada, in two Indian plants, and in a large number of plants in China, with an estimated total capacity of the order of 170 000 t/a in 1999 (see Fig. 12). Briquettes of crushed calcined dolomite and ferrosilicon fines in a typical ratio of 5 : 1 are loaded into tubular refractory steel retorts, heated externally to a reaction temperature of 1200 °C, and evacuated to 10 – 25 Pa. Magnesium vapor is condensed at the water-cooled end of the retorts in the form of "crowns", which are subsequently removed for remelting and casting into ingots.

Reaction kinetics depend on the thermal conductivity of the briquettes, the shrinking core pattern of the reaction zone, and the diffusion rate of magnesium vapor. Heat transfer to the charge limits the diameter of the retorts. Typical retorts are 3 m long and have a bore diameter of 300 mm; each retort yields around 70 kg of magnesium per day. Normally, a slight excess of silicon above the stoichiometric requirement is used. Recovery increases with increasing ferrosilicon concentrations. Consumption figures per tonne of magnesium in Chinese plants have been reported as follows [88]: 0.9 MWh electric energy, 10.5 t coal, 11 t dolomite, 1.1 t of 75 %

ferrosilicon, with an 82 % overall magnesium recovery and 6.2 t of slag produced.

Starting from a high-purity dolomite raw material (99.5 % $CaCO_3 \cdot MgCO_3$) makes it possible to produce 99.95 % or better magnesium in commercial quantities. Attempts are being made to improve the process, e.g., to increase capacity by increasing the retort diameter [89].

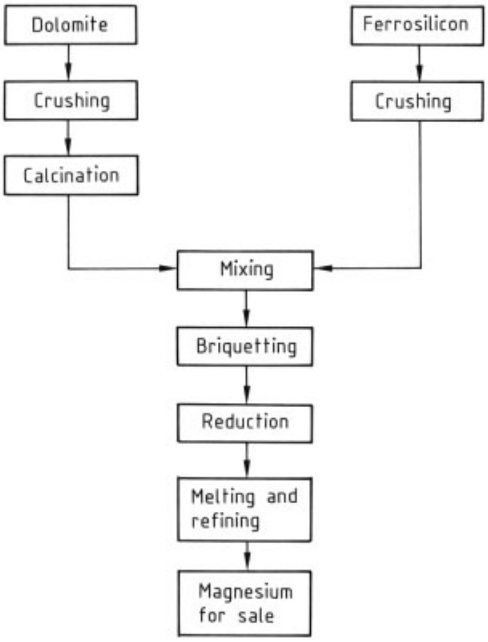

Figure 12. Flow sheet for the production of metallothermic magnesium (Pidgeon and Bolzano processes)

Bolzano Process [43], [44]. The Bolzano process was formerly operated in Italy, but now only in Brazil by Brasmag. It employs an internally heated, brick-lined cylindrical reactor (Fig. 13). The briquetted charge of homogeneously mixed calcined dolomite and ferrosilicon fines (d) is loaded on a charge support system (e). Internal electrical heating (f) is conducted to the charge through the charge support system. The process operates at 1200 °C and < 400 Pa. Magnesium vapor condenses inside the condensers, which are water-cooled to 400 – 500 °C. Each reactor has a production capacity of ca. 2 t of magnesium per 20 – 24 h reaction cycle. Production of 1 t of magnesium consumes 7 – 7.3 MW·h, 5 – 5.2 t calcined dolomite, and 0.7 t of silicon contained in 78 % ferrosilicon, with 81 % silicon recovery; 5 – 5.2 t of slag is produced per tonne of magnesium and sold for use in plaster and building bricks. Mechanization of charge handling and slag removal is a challenge in this process. Magnesium with a purity of 99.98 – 99.99 % is obtained.

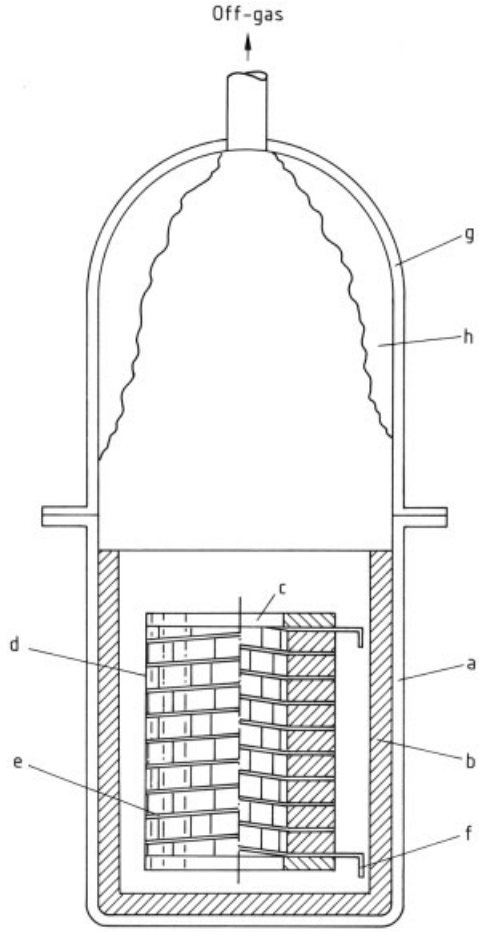

Figure 13. Bolzano reactor
a) Steel mantel; b) Refractory; c) Removable charge pack; d) Charge briquettes; e) Charge support; f) Electric heating connection; g) Removable condenser; h) Condensed magnesium

Magnetherm Process [46–48]. The Magnetherm process was developed by Pechiney Electrometallurgie in the early 1960s and has been used at the company's magnesium extraction plant in Marignac, France until 2001. The process is also operated by Magnohrom (Yugoslavia), and on a larger scale by North West Alloys (United States) until 2001. The total installed production capacity worldwide up to

2001 was 65 000 t/a. The overall equation for the process is

$$2\,MgO{\cdot}CaO\,(s) + Al_2O_3\,(s) + (Fe)\,Si\,(s) \longrightarrow Ca_2SiO_4{\cdot}Al_2O_3\,(l) + Fe\,(l) + Mg\,(g)$$

The dicalcium silicate slag formed in the reaction of calcined dolomite with ferrosilicon (Pidgeon and Bolzano processes) has a melting point of 2000 °C and is solid at the reaction temperature (1200 °C). The Magnetherm process, however, operates with a partly molten slag of the general composition $2\,CaO{\cdot}SiO_2{\cdot}nAl_2O_3$ at 1550 – 1700 °C. Typically, the slag may contain 55 % CaO, 25 % SiO_2, 14 % Al_2O_3, and 6 % MgO, some 40 % of which is in the form of solid Ca_2SiO_4. The slag is kept in an electrically heated, cylindrical, brick- and carbon-lined steel vessel (Fig. 14). The power input is conveyed through a water-cooled copper electrode (a) from the top through the slag (g) to the bottom graphite lining (b) of the reactor, which acts as a power outlet.

Coarse calcined dolomite and alumina screened to 3 – 30 mm, as well as ferrosilicon containing a minimum amount of fines, are added continuously to the slag under a vacuum of 3 – 6 kPa through charge holes in the water-cooled furnace roof. Magnesium vapor is condensed in a separate condenser system.

Calcined magnesite may be used as a supplement to the dolomite to adjust the slag composition. Calcined bauxite may be used alternatively to Al_2O_3. Aluminum metal may be used as a secondary reductant and Al source.

The main endothermic reaction takes place close to the surface of the slag. Deeper in the molten slag, the higher pressure slows down and finally stops the reaction at a silicon content in the bottom layer of ca. 20 %. Slag and 20 % ferrosilicon are tapped twice during a cycle of ca. 18 h, having produced more than 10 t of magnesium. Adjustment of the slag composition to higher magnesium oxide content and heating the surface of the slag with arc or plasma burners to 1900 – 1950 °C have been suggested to allow reduction to occur at atmospheric pressure [45], [90].

Production of 1 t of magnesium typically requires 4.6 t calcined dolomite, 0.6 t alumina, 0.85 t of 75 % ferrosilicon and 7 – 8 MW · h electric energy. About 5 t of slag is produced per tonne of magnesium and is used in the cement industry, for road building, and as a source of magnesium and silicon in agricultural soil. The ferrosilicon byproduct (yield 0.3 t per tonne of magnesium) is used in the metallurgical industry, in the production of heavy-liquid media for the separation of metals, or in drilling mud.

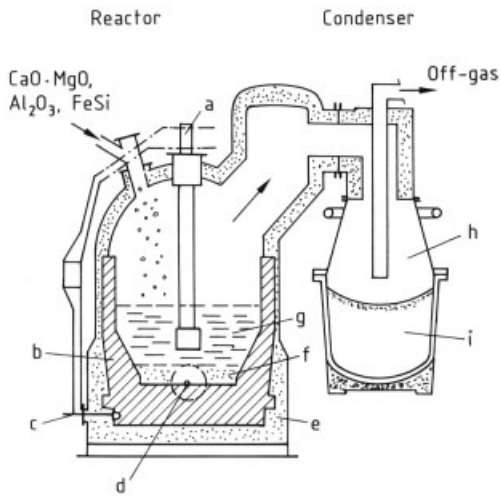

Figure 14. Magnetherm reactor
a) Copper electrode; b) Graphite lining; c) Bottom electrode; d) Taphole; e) Refractory; f) Low-grade FeSi; g) Slag; h) Condenser; i) Condenser crucible

4.3. Refining and Casting

All magnesium extraction processes are followed by refining to remove impurities and casting to convert the metal to ingots, billets, slabs, or granules [5].

Refining.

Nonmetallic impurities in the form of particulate oxides, nitrides, carbides, and chloride inclusions impair corrosion resistance as well as mechanical and surface properties. Most alkali metal chloride and magnesium chloride fluxes wet and coat the surface of the impurity particles. The higher density of the chlorides causes the impurities to sink to the bottom of the melt as a sludge. Fluxes are composed to maintain a minimum difference in density of 0.15 – 0.20 g/cm^3 relative to that of the metal.

Inspissating (thickening) agents such as magnesium oxide and calcium fluoride reduce

Table 6. Composition of fluxes used in melting and refining (wt %)

Purpose	$CaCl_2$	NaCl	KCl	$MgCl_2$	CaF_2	MgO
Melting	40	30	20	10		
Melting and refining	20	10	10	35	15	10
Refining	15	10	10	35	20	10

the fluidity of the sludge to prevent it from contaminating the metal after settling. Inspissating fluxes are also used to protect the metal surface during holding and casting because of their viscosity and stability (Table 6).

Metallic impurities are normally controlled by the choice of raw materials and extraction process parameters. The most common metal impurities in pure magnesium are iron, copper, nickel, manganese, zinc, aluminum, silicon, calcium, and sodium. Intermetallic particles of aluminum, manganese, iron, and silicon have also been identified.

Heavy metals reduce the corrosion resistance of magnesium while acting as cathodes in galvanic cells in the metal. Metallic impurities are also undesirable for certain applications in which magnesium is used as a reductant (Ti, Zr, U) or as a canning material for the fuel elements in nuclear reactors where the metal's neutron absorption properties are important.

The solubility of iron in pure magnesium is 0.018 wt % at 650 °C, increasing with temperature to ca. 0.04 wt % at 750°C [107]. A content of 200 – 400 ppm Fe is thus typical for commercial-grade Mg. The principles for removing iron by means of manganese and aluminum, to give a considerably more corrosion resistant alloy, was discovered more than 70 years ago, but the industrial importance of this was not recognized until the 1980s. In aluminum-free alloys, zirconium may be used to remove iron.

Hydrogen present in magnesium is partly due to pick-up from the moisture in air and in flue gas from gas or oil burners. $MgCl_2$ and other salt components are strongly hygroscopic and easily pick up water. Magnesium has a much higher solubility for hydrogen than aluminum, and typical contents in Mg and Mg alloys are 2 – 12 mg/kg. While inert gas purging removes hydrogen in molten Al or AlMg alloys fairly readily, this is not the case in molten Mg due to the unfavorable thermodynamics [108]. Due to the high vapor pressure of Mg, it is hardly an alternative to remove hydrogen by means of vacuum extraction. However, hydrogen in magnesium die cast parts is normally not considered a problem, as possible hydrogen porosity only adds to the porosity usually present anyway. For gravity cast magnesium alloys, hydrogen may be problematic, as for magnesium used as an alloying element in aluminum.

Equipment. The standard equipment in magnesium refineries consists of large, stationary, brick-lined furnaces or externally heated steel crucibles containing 10 – 25 t or 2 – 6 t of molten magnesium, respectively. The stationary furnaces can be heated electrically, with gas or oil burners from the top or side, or by submerged electrical heating systems.

Electrolytic magnesium recovered and processed in the molten state can be refined in continuous stationary furnaces with metal yields greater than 99.5 % and negligible flux consumption (Fig. 15) [49], [106].

Metallothermic magnesium crowns or condensates are recovered in the solid state. Standard practice involves batchwise melting and refining in steel crucibles and subsequent transfer to casting furnaces. The melting process and the high content of nonmetallic impurities require a flux consumption that is 5 – 10 % of the metal volume; metal recoveries are in the range 95 – 98 %.

Casting. Highly automated casting by pumping or static pressure is state of the art. Magnesium contracts by 4 % during solidification and 5 % during cooling to 20 °C, and hence is sensitive to surface cracks and shrink holes. Turbulent flow favors the formation of nonmetallic inclusions. Consequently, care is taken to control the solidification process and to reduce turbulence. At elevated temperature or in the case of turbulent metal flow, sulfur powder may be sprayed on the metal surface to sup-

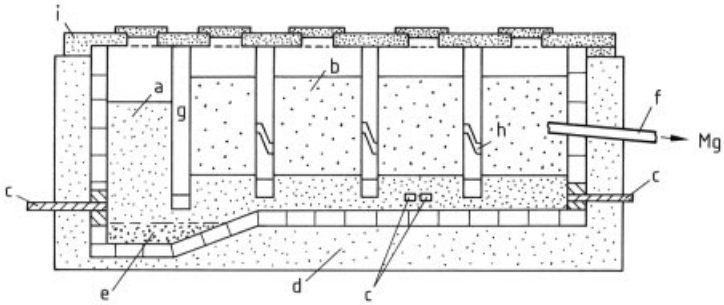

Figure 15. Norsk Hydro refining furnace [49]
a) Salt melt; b) Magnesium metal; c) Electrodes; d) Refractory; e) Sludge; f) Tapping spout; g) Partition wall; h) Metal port; i) Refractory cover

press oxidation by creating an SO_2-containing atmosphere.

The level of impurities in standard 99.8 % magnesium does not influence the solidification pattern. Pure magnesium solidifies at its melting point (650 °C). During holding and casting, the metal surface is protected against oxidation by a protective gas. Acting together with oxygen in the air, low concentrations of sulfur hexafluoride or sulfur dioxide in air form protective layers on the metal surface. Sulfur hexafluoride has been widely used during the last twenty years [50], [51], but due to the high global warming potential of SF_6 (GWP factor: 23 900), SO_2 is gradually being reintroduced [109]. Due to the negative impact of SO_2 on the working environment, efforts are made to find alternatives [110].

Pure magnesium rounds and T-bars and alloy billets are cast in multistrand continuous or semicontinuous direct-chill casting machines, also using state of the art hot-top and airslip technologies. The main safety concern to be managed by these technologies is associated with metal – water contact, which could result in hydrogen evolution and potential explosions.

To avoid discoloration and surface corrosion during transportation and handling, the metal is protected by suitable wrappings. The state of the art is to use dissicants inside the containers during transport in severe climatic conditions.

Recycling. Recyclability is important for the economic application of magnesium, especially for the rapidly growing application of die-cast parts made from magnesium alloys, e.g., in the automotive and electronics industries.

In typical magnesium die-casting operations, only around 50 – 70 % of the purchased ingot is present in finished castings, the remainder being production scrap of various qualities. The majority of the scrap (85 – 95 %) is normally sorted into so-called Class 1 scrap, suitable for remelting operations. Such clean scrap can be recovered to primary-metal quality [111]. The remelting requires less than 5 % of the energy required to produce primary metal.

Two principally different methods for treatment of scrap may be utilized [114]:

1) Flux-based metal protection and refining
2) Protective gas (SO_2/SF_6) for melt protection and settling, possibly assisted by argon purging or filtration (flux-free refining)

Flux-based protection and refining has been most widely used, and is probably still the most important in terms of tonnage. Recycling of magnesium by using salt fluxes has a potential to improve the recovery, but disadvantages are that the salt flux is corrosive to the surrounding equipment and environment (HCl fumes), and that it generates a sludge in the lower section of the crucible. It also creates a possibility of having salt inclusions in the cast product, and flux-free melting is thus preferred by die casters who recycle in house [115]. Chlorides may also combine with organic residues on the scrap to form highly toxic compounds, creating a need for comprehensive off-gas treatment systems.

The increasing use of recycling has generated an interest in metal cleanliness and various methods for determining the inclusion content have been developed [112].

Magnesium alloys from scrapped automobiles (end-of-life vehicles) as well as other post consumer scrap represent an increasing volume, and will be a technological challenge to the industry in the future.

4.4. Particulate Magnesium

Particulate magnesium comprises powders, granules, turnings, and raspings; particle sizes range from 0.04 to 5 mm, and particle shapes vary. Particulate products are used in chemical and metallurgical processes in which magnesium acts as a reagent, catalyst, or alloying element. Fine powders are used in pyrotechnics.

Particulate magnesium is considered inflammable and subject to specific labeling and packing regulations according to IMDG codes classes 4.1 and 4.3; UN No. 1869 and 1418; CFR 49: 172.01, flammable solid; RID/ADR class 4.1, number 13 b, class 4.3 number 1 d.

Flowability, uniformity of grain size, and defined area – volume ratios are important for injection into molten metal (steel, aluminum). Particulate magnesium is frequently blended with inert and active components (e.g., lime or alumina) or coated with salt or refractories. These products are used for desulfurization of iron. Blending and coating reduce the reactivity and increase the magnesium yield. Turnings and raspings commonly used for Grignard syntheses and pharmaceuticals are adjusted for reactivity (surface-to-weight ratio) and form of the particles according to customer specifications. Finely divided powder for pyrotechnics is subject to narrow tolerance limits, including the content of organic material.

Particulate magnesium is produced directly from the molten state by shotting, atomization, or granulation or by machining ingots, slabs, or billets, with subsequent grinding and screening. Airborne dust with a particle size below 0.15 mm is explosive. Safety precautions for machining and atomizing operations include prevention of explosion hazards in the processing and handling of fine dust, as well as design of the general plant layout to limit the consequences of accidental blowouts. Clean operation and personal hygiene are important.

Nearly spherical magnesium shots in the 0.2 – 2-mm size range are produced by nozzle spraying in a cooling tower with an inert atmosphere [52]. Atomizing is performed in a protective atmosphere of argon, helium, or methane. Magnesium may be atomized on a rotating disk or in a gas stream. During shotting and atomizing, precautions are taken to avoid the formation of layers of highly ignitable, pyrophoric substances.

Salt-coated granules were originally developed in the Soviet Union and are produced by mixing alkalimetal chloride salts and molten magnesium in a rotating perforated cup; the metal is disintegrated by centrifugal force with subsequent air cooling. Spheroidal particle sizes range from 0.2 to 2.5 mm, with a salt content of 10 % [23], [53]. Other methods include mixing molten mixtures of magnesium and alkalimetal chlorides and adding a stabilizing boron compound [54]. Various proprietary processes employ eutectic salt compounds with low hygroscopicity to produce particles with high flowability [55]. The salt and metal are separated by gentle crushing and screening, leaving granules with 5 – 10 % salt as coating. The production of salt-coated granules is safer than that of pure magnesium powder and can be integrated in normal foundry operations.

5. Environmental Aspects

Magnesium production processes require installations for cleaning off-gases, and most of them also for treatment of wastewater. Significant amounts of solid waste are deposited as landfill but may in some cases be used for agricultural purposes due to their content of magnesium or potassium. In production processes involving calcination of carbonate minerals or dehydration of hydroxides, dust is removed by conventional methods. The main emissions to air are carbon dioxide, nitrogen oxides, and sulfur dioxide from the combustion and calcination processes.

In *electrolytic processes*, carbon is used in some cases as a reductant in the preparation of anhydrous magnesium chloride, and carbon is always used as anode material in electrolysis. The presence of carbon, water, chlorine, or hydrogen chloride, even in low concentrations, at elevated temperature leads to formation of chlorinated

hydrocarbons and other organic compounds. Chloromethanes, chlorobenzenes, polychlorinated biphenyls (PCBs), chlorodibenzofurans, and chlorodibenzo-*p*-dioxins are highly toxic and are only slowly degraded in nature. Although such compounds only occur in liquid effluents from scrubber systems at low concentrations (a few micrograms per liter), collection and destruction are of great environmental importance.

Chlorinated organic compounds can be either in solution or present as small particles in the liquid effluents. Particles may be flocculated with appropriate cationic or nonionic agents. The agglomerates adhere to solid inorganic particles that are often present in the effluents (carbon and hydroxides or oxides of iron, aluminum, silicon, etc.). The particles are removed by filtration or centrifugation. Liquids may be treated with activated carbon.

Traces of chlorine and hydrochloric acid in off-gases are scrubbed, recycled, or neutralized by conventional methods. Where available, sea water is an efficient medium for scrubbing of acidic components such as chlorine, hydrogen chloride, and sulfur dioxide. Off-gases containing chlorinated hydrocarbons may be subjected to incineration or activated carbon treatment.

Solid byproducts from *metallothermic processes* are either used for industrial purposes or deposited as landfill.

Sludge from *foundry operations* contains entrapped metallic magnesium globules. When deposited as landfill, precautions must be taken to control the formation of hydrogen gas by the reaction of entrapped magnesium with water. Off-gases from foundry operations containing hydrogen chloride, sulfur dioxide, and magnesium oxide may be scrubbed by conventional methods to comply with local regulations.

6. Quality Specifications

Standard designations for unalloyed magnesium in the form of ingot or stick for remelting are listed in Tables 7 and 8. These figures correspond to ISO 8287 (2000-04-01), ASTM B92 (re-approved 1994) and EN 12421 (1998). Grade 9980 (99.80) is adequate for most uses and can be obtained by electrolysis by using controlled standard operations. Grades 9995 (99.95) and 9998 (99.98) are normally produced only by metallothermic processes. High-purity magnesium can also be produced by distillation of lower grades, but this is expensive [5], [23].

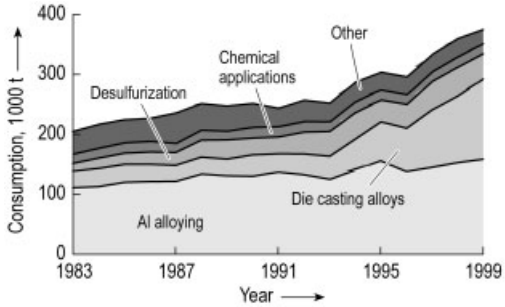

Figure 16. Consumption of magnesium by application

7. Uses

The uses and consumption of magnesium are summarized in Figure 16. The overall growth rate in the use of magnesium is estimated to be 3 – 5 % per annum.

Aluminum Alloys. The largest single use of magnesium is as an alloying element in aluminum alloys (→ Aluminum Alloys), which can contain 0.5 – 4.5 % Mg. Aluminum – magnesium alloys have improved ductility, enhanced resistance to saltwater corrosion, and improved cryogenic properties. User forms are ingots weighing 1 – 25 kg and sows and T-bars weighing 250 – 500 kg. The market growth of this application depends on the growth of aluminum and the recycling efficiency of aluminum scrap, particularly beverage cans.

Desulfurization and Nodularization. Magnesium is used as a hot-metal desulfurizing agent for the production of high-quality steels. Due to development of powder-injection technology, it is now possible to inject solid Mg-based reagents containing 50 – 80 % magnesium deep into a liquid iron bath, to produce steels with a sulfur content as low as 0.002 % [56], [57]. The magnesium consumption in the desulfurization process depends mainly on the sulfur to be removed, and is typically 0.5 kg of Mg per tonne of metal.

Table 7. Magnesium grades[a,b]

	ISO 8287, 99.5	EN 12421, 99.5	ISO 8287, 99.80-A	ISO 8287, 99.80-B	ASTM B92, 9980A	ASTM B92, 9980B	EN 12421, 99.80-A	EN 12421, 99.80-B
Al	0.1	0.1	0.05	0.05			0.05	0.05
Mn	0.1	0.1	0.05	0.05	0.10	0.10	0.05	0.05
Si	0.1	0.1	0.05	0.05			0.05	0.05
Fe	0.1	0.1	0.05	0.05			0.05	0.05
Cu	0.1	0.1	0.02	0.02	0.02	0.02	0.02	0.02
Ni	0.01	0.01	0.001	0.002	0.001	0.005	0.001	0.002
Pb			0.01	0.01	0.01	0.01	0.01	0.01
Sn			0.01	0.01	0.01	0.01	0.01	0.01
Na	0.01	0.01	0.003		0.006		0.003	
Ca	0.01	0.01	0.003				0.003	
Zn			0.05	0.05			0.05	0.05
AOE[d]	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Mg BD[e]	99.5	99.5	99.80	99.80	99.80	99.80	99.80	99.80

[a] Figures refer to maximum content in weight percent except where otherwise specified; [b] For complete information, refer to the standards; [c] For nuclear applications, refer to the standard; [d] Any other element, max; [e] Mg, by difference, min.

Table 8. Magnesium grades[a,b]

	ASTM B92, 9990A	ISO 8287, 99.95-A	ISO 8287, 99.95-B	ASTM B92, 9995A	EN 12421, 99.95-A	EN 12421, 99.95-B	ASTM B92, 9998A
Al	0.003	0.01	0.01	0.01	0.01	0.01	0.004
Mn	0.004	0.006	0.01	0.004	0.006	0.01	0.002
Si	0.005	0.006	0.01	0.005	0.006	0.01	0.003
Fe	0.04	0.003	0.005	0.003	0.003	0.005	0.002
Cu		0.005	0.005		0.005	0.005	0.0005
Ni	0.001	0.001	0.001	0.001	0.001	0.001	0.0005
Pb		0.005	0.005		0.005	0.005	0.001
Sn		0.005	0.005		0.005	0.005	
Na		0.003	–		0.003		
Ca		0.003	–		0.003		
Zn		0.005	0.01		0.005	0.01	
Ti				0.01			0.001
AOE[d]	0.01[c]	0.005	0.005	0.005[c]	0.005	0.005	0.005
Mg BD[e]	99.90	99.95	99.95	99.95	99.95	99.95	99.98

[a] Figures refer to maximum content in weight percent except where otherwise specified; [b] For complete information, refer to the standards; [c] For nuclear applications, refer to the standard; [d] Any other element, max; [e] Mg, by difference, min.

Magnesium is also used in the production of ductile iron, also known as nodular cast iron. Ordinary cast iron is very brittle at room temperature because it contains carbon in the form of flakes or laminar graphite. Addition of magnesium to liquid cast iron, either as pure solid magnesium or alloyed with ferrosilicon, modifies the laminar graphite into nodular graphite (globules), thereby inducing ductility (elongation in the range of 14 – 18 %), and making it suitable for several other industrial applications [91], [92].

Magnesium-based alloys represent the fastest growing area of magnesium application (see Fig. 16). They are normally delivered as alloyed ingots from the primary producers (→ Magnesium Alloys). According to the International Magnesium Association, over 90 % of shipments of magnesium alloys are consumed in the manufacture of high-pressure die castings, particularly for automotive applications; automotive die-castings make up 85 – 90 % of the total annual weight of die castings produced. Limits on fuel consumption and vehicle emissions imposed by CAFE (USA) and plans for such legislation in the EU have promoted the increasing use of magnesium in automobile production to reduce vehicle weight. Other growing applications include consumer electronics such as laptop computers, cell phones, and cameras, where magnesium's light weight, damping ca-

pacity, recyclability, and relative strength give it a clear advantage over alternative materials.

Chemical applications include the use of magnesium as a reagent in the production of titanium and zirconium by the Kroll process (→ Titanium, Titanium Alloys, and Titanium Compounds), the reduction of uranium fluoride to uranium metal, and the production of beryllium and hafnium. Titanium constitutes by far the largest user segment in this group, the magnesium being added in ingot form. Magnesium chloride from the Kroll process is electrolyzed to recover magnesium and chlorine. The normal magnesium requirement to compensate for losses is 0.1 t per tonne of titanium.

Magnesium raspings for the production of metallic uranium have a specified limited content of boron, cadmium, and rare-earth elements including gadolinium, samarium, europium, and dysprosium.

The use of particulate magnesium, raspings, or turnings in Grignard syntheses has been declining, due primarily to restricted use of tetraethyl lead.

Hydrogen Storage. Magnesium forms the hydride MgH_2, which is stable at room temperature (→ Hydrides, Chap. 2.6.). A hydrogen equilibrium pressure of 0.1 MPa is reached at ca. 290 °C [93]. At 25 °C the hydrogen pressure is less than 0.1 Pa. MgH_2 contains 7.66 wt % hydrogen and about 1.5 times more hydrogen than liquid hydrogen for a given volume. MgH_2 is therefore an interesting candidate as a hydrogen carrier, e.g., in onboard storage tanks for hydrogen for vehicles driven by fuel cells or hydrogen-combustion engines.

The reversibility of hydrogen absorption/desorption is poor for normal Mg powders, but improves greatly when the particle size is well below 1 μm and catalysts are added [94]. Of the theoretical amount of hydrogen 80 % (6 wt %) can be reversibly absorbed and desorbed at 300 °C within 10 min by using nanocrystalline Mg particles and various catalysts such as Pd and Fe [94].

Also Mg–Ni alloys close in composition to the compound Mg_2Ni can store significant amounts of hydrogen (3.6 wt %). The benefits of Mg_2Ni compared to pure Mg is that the temperature of absorption/desorption can be lowered from about 300 °C to about 200 °C [95].

Other Uses. The electronegative character of magnesium is used for corrosion abatement as sacrificial soil and hot-water-tank anodes. European restrictions on aluminum in drinking water have encouraged the use of magnesium anodes for hot-water tanks. Magnesium batteries for military purposes and rescue systems combine light weight and high energy content.

8. Economic Aspects

Since 1990 the use of magnesium has grown at a rate of 3–5 % per year, and this is expected to continue at a similar rate. The main reason for this forecast growth is the increasing use of magnesium in die-casting applications (a rise of 15 % per year is forecast over the next decade), particularly for use in the automotive industry. The main consumer regions in 2000 were North America (176.7×10^3 t), Western Europe (122.9×10^3 t), and the Pacific Rim (50.5×10^3 t), with the rest of the world (including China and the former Soviet states) consuming around 50×10^3 t. The optimistic predictions for growth have encouraged the development of a number of projects for new magnesium production facilities in, amongst others, Australia, the Netherlands, and the Republic of Congo. However, in 2001, in the face of increasing competition from Chinese production, three major producers, Pechiney in France, Northwest Alloys in the USA, and Norsk Hydro in Norway, were forced to stop primary production. Production capacities of main producers as of 2001are given in Table 9.

Several plants were closed also in the 1990s: the Dow Chemical plant closed in December 1998, and there were closures in Italy and Japan in the early 1990s. Production in the former Soviet Union has decreased in recent years, due to changes in titanium requirements. In contrast, China is now a major producer. Total world production of magnesium in 2000 was 405×10^3 t (see Table 10), and consumption was 400×10^3 t. Magnesium production is now divided fairly equally between the electrolytic and thermic processes.

Table 9. Main magnesium producers (2001)

Company	Installed production capacity, 10^3 t	Production process	Comments
Northwest Alloys, USA	45	thermic	Closed in October 2001
Norsk Hydro Canada, Inc.	45	electrolytic	Expanded to 48 in 2002
MagCorp, USA	43	electrolytic	
Norsk Hydro ASA, Norway	42	electrolytic	Closed primary prod. in March 2002
Dead Sea Magnesium, Israel	33	electrolytic	
Pechiney Electromet., France	20	thermic	Stopped prod. in 2nd quarter 2001
Timminco Metals, Canada	6	thermic	
Magnola Metallurgy, Canada	63	electrolytic	First commercial prod. in 2001
Brasmag, Brazil	10	thermic	
Magnohrom, Yugoslavia	5	thermic	
Avisma, Russia	25	electrolytic	
Ust-Kamenogorsk, Kazakhstan	20	electroytic	
Solikamsk, Russia	18	electrolytic	
Oriana, Ukraine	15	electrolytic	
China	170	thermic	Estimated capacity
	10	electrolytic	

The evolution of world production of primary magnesium from 1980 to 2000 is summarized in Table 10. Up to 1998 the United States was the largest producer (ca. 40 % of the world total), followed by China (20 %) and Canada (16 %). However, the picture changed somewhat with the closure of the Dow plant, and China is now by far the largest producer country. Most of the growth in the last decade has been in China, where a large number of plants of varying capacity using the Pidgeon process have started up.

Table 10. Evolution of world primary magnesium production (10^3 t/a)∗

Country	1980	1990	2000
Brazil		6.5	9
Canada	9.3	26.7	55
China	7	5.8	140∗∗
France	9.3	14.6	15
Italy	9.7	5.7	
Israel			33
Japan	9.3	12.8	
Norway	44.4	48.2	42
Poland	0.5		
Soviet Union/CIS total	76	88	45∗∗
United States	154.1	139.3	65∗∗
Yugoslavia	0.5	5.8	1
World total	320.1	353.5	405

∗ Data from the U.S. Geological Survey and the International Magnesium Association. ∗∗ Estimated.

There is no organised terminal market for magnesium; the major producers control their own distribution network. Prices (see Fig. 17), having risen to a high of over 230 c/lb at the end of 1995, have since dropped sharply, due in large part to downward pressure from Chinese producers.

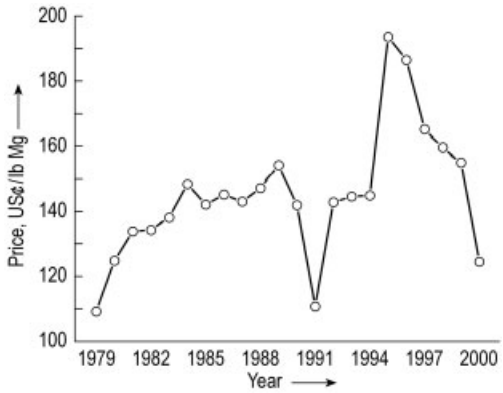

Figure 17. Price development for pure magnesium; Average year end US Spot Western Price in nominal terms (Source: Platts Metals Week)

The U.S. standard import duty on pure magnesium is 8 %. In addition there is an 305.56 % antidumping tariff on imports of Chinese pure magnesium into the States (one specific exporter has a lower tariff of 24.67 %). The European Union raised its antidumping duty on imports

of Chinese pure magnesium to 63.4 % in 2000; however, this duty will probably be dropped when Pechiney, the only EU producer of magnesium, announces the final closure of its plant in France.

9. Toxicology and Occupational Health

Magnesium is an essential element for human, animal, and plant life, and the toxicity of the metal is regarded as low. The toxicity of magnesium compounds is usually attributed to the anion [96]. Judging from the literature, the focus with respect to magnesium is on the effects of Mg deficiencies rather than the toxic properties of the metal. RTECS has no entries on magnesium, and no reports of serious poisoning of workers in industry have appeared in the literature. There is no evidence of lung injury as a result of Mg dust inhalation. Inhalation of freshly generated magnesium oxide, however, did experimentally induce metal fume fever in humans, but with a considerably milder response compared to that due to zinc oxide. No reports of industrial experience of this appear to have been made [96], [97]. Some investigators have reported a higher incidence of digestive disorders in magnesium plant workers and have suggested that a relationship may exist between magnesium absorption and gastroduodenal ulcers. Slivers of magnesium penetrating the skin or entering deep wounds can cause "aeroform tumours" or "gas gangrene", but reported cases have in no way paralleled those seen in animal experiments. Magnesium powder may cause irritation to the skin, mucous membranes, and upper respiratory tract [98].

Symptoms such as subjective feeling of heat and immense thirst, blurred vision, nausea and vomiting, CNS depression, hyporeflexia, hypothermia, hypotension, bradycardia (depressed pulse rate), respiratory depression, and even respiratory and cardiac arrests have been reported in connection with hypermagnesemia as a result of massive intakes of laxatives and antacids (i.e., magnesium sulfate and magnesium hydroxide) and inadvertently administered overdoses of magnesium compounds during medical treatment [99].

Humans are exposed to magnesium during the production, casting and machining of the metal and through dietary intake. The most serious hazard presented by magnesium is the danger of burns from molten metal.

10. Safety

Magnesium is for most end users a safe metal. Magnesium has a very high affinity for oxygen, and a magnesium oxide film is quickly formed on any freshly cut surface. This film prevents further oxidation of the metal. Ready-made magnesium parts, e.g., in the automotive industry, represent no hazard to the user.

There are some processes during production, manufacturing, and recycling of magnesium where special safety precautions must be taken. Molten magnesium will burn if exposed to air. Cover gases and fluxes are used during casting in magnesium foundries to prevent liquid magnesium from reacting with oxygen. A magnesium dust atmosphere may burn and explode if an ignition source is present [100]. Dust atmospheres are rare, but may be encountered during accidental events in foundries or recycling plants or during firefighting. Small magnesium chips will burn if ignited. They may release hydrogen if brought in contact with water, producing a potentially reactive atmosphere.

Special care must be taken in extinguishing magnesium fires. Recommendations are given in the National Fire Protection Association (NFPA) publications [101], [102]. Water is usually the wrong agent for magnesium fires because the metal reacts with water to form hydrogen, which will contribute to the fire. Furthermore, violent steam explosions can result if water enters molten metal. As an exception, where only small quantities of magnesium are involved, the fire may be extinguished by applying large amounts of water, provided no pools of molten magnesium are present. For larger fires, dry (water-free) extinguishing agents should be used, e.g. dry salt (sodium chloride) or magnesium foundry fluxes (mixed alkali and alkaline earth metal chlorides and fluorides).

Helium and argon are inert to magnesium and can be used to smother a magnesium fire. However, nitrogen, carbon dioxide, and halons are not inert, and will react with molten magnesium.

11. References

1. C. J. P. Ball: "The History of Magnesium", *J. Inst. Met.* **84** (1955/56) 399.
2. R. S. Busk: *Magnesium Products Design,* Marcel Dekker, New York 1987.
3. *Metals Handbook,* 10th ed., **vol. 2,** American Society for Metals, Ohio 1990.
4. *Comprehensive Inorganic Chemistry,* 1st ed., Pergamon Press, Oxford 1973.
5. E. F. Emley: *Principles of Magnesium Technology,* Pergamon Press, London 1966.
6. P. J. McGonigal et al., *J. Phys. Chem.* **66** (1962) 737 – 740.
7. Kh. L. Strelets: *Electrolytic Production of Magnesium,* Israel Program for Scientific Translations, Jerusalem 1977.
8. *Kirk-Othmer,* 4th ed., **vol. 15,** pp. 622 – 674.
9. Utah Deposits of Natural Resources, Bull 116, 1980.
10. A. N. Strahler: *Principles of Physical Geology,* Harper and Row Publ., New York 1977.
11. A. P. Lysenko et al., *Tsvetn. Met.* 1987, no. 8, 53 – 55.
12. General Motors, US 5 853 560, 1997 (R. A. Sharma).
13. J. C. Withers, R. O. Loufty, *Light Met. (Warrendale, Pa.)* 1986, no. 2, 1013 – 1017. A. P. Ratvik et al., *J. Electrochem. Soc.* **134** (1987) no. 2, 321 – 327.
14. D. A. Elkins et al.: *An Economic and Technical Evaluation of Magnesium Production Methods,* Part 2: Carbothermic, US Bureau of Mines RI 6946, 1967.
15. F. Hori, US 4 200 264, 1980.
16. Shell Int. Res., US 4 572 736-A, 1986.
17. A. M. Cameron et al.: "Carbothermic Production of Magnesium", *Pyrometallurg 87,* Institutions of Mining and Metallurgy, London 1987, pp. 195 – 222.
18. R. J. Allain et al.: *Proc. Annu. Meet. Int. Magnesium Assoc.* **34th,** 1979.
19. D. V. Pruttskov et al., *Tsvetn. Met.* 1986, no. 5, 52 – 56.
20. MPLC, US 4 269 816, 1981 (C. E. E. Shackleton).
21. Norsk Hydro, US 3 742 100, 1969 (Ø. Bøyum et al.).
22. R. D. Toomey, *Proc. Annu. Meet. Int. Magnesium Assoc.* **35th,** 1980.
23. A. N. Petrunko et al., *Proc. Annu. Meet. Int. Magnesium Assoc.* **32nd,** 1977.
24. E. Aarebrot et al., *Light Met. (Warrendale, Pa.)* 1977, no. 1, 491 – 512. T. Østvold, H. A. Øye, *Light Met. (Warrendale, Pa.)* 1980, 937 – 947.
25. K. Grjotheim et al., *Acta Chem. Scand.* **24** (1970) 489 – 509.
26. K. Grjotheim et al., *Trans. Faraday Soc.* **67** (1971) 640 – 648.
27. K. Grjotheim et al., *Acta Chem. Scand.* **26** (1972) 2050 – 2062.
28. D. Dumas et al., *Acta Chem. Scand.* **27** (1973) 319 – 328.
29. E. I. Savinkova, A. I. Orekhova, *Izv. Vyssh. Uchebn. Zaved., Tsvetn. Metall.* 1978, no. 3, 76 – 80.
30. J. B. Belavadi et al., *J. Appl. Electrochem.* **12** (1982) no. 5, 501 – 503.
31. J. B. Belavadi et al., *Adv. Electrometall. Proc. Symp.* 1983, 2/27 – 2/43.
32. K. Andreassen, *Erzmetall* **31** (1978) no. 7/8, 301 – 309.
33. N. Høy-Petersen, *J. Met.* **21** (1969) no. 4, 43 – 49.
34. O. G. Sivilotti, *Light Met. (Warrendale, Pa.)* 1988, 817 – 822.
35. H. Ishizuka, US 4 495 037, 1985; US 4 647 355, 1987.
36. Dow Chemical Co., US 4 448 654, 1984 (S. F. Spangenberg et al.).
37. K. S. Dean et al.: *An Economic and Technical Evaluation of Magnesium Production Methods,* Part 1: Silicothermic, US Bureau of Mines RI 6656, 1965.
38. M. C. Flemings et al.: *An Assessment of Magnesium Primary Production Technology*, US Department of Energy Report, Contract No. EX-76-A-01-2295, Feb. 1981.
39. L. M. Pidgeon et al., *Trans. Metall. Soc. AIME* **159** (1944) 315.
40. J. R. Wynnyckyj et al., *Metall. Soc. CIM Annu. Vol.* 1978, 73 – 81.
41. A. Froats, *Light Met. (Warrendale, Pa.)* 1980, 969 – 979.
42. J. R. Wynnyckyj et al., *Can. Metall. Q.* **30** (1991) no. 3, 139 – 143.
43. Soc. Ital. Magnesio, US 4 238 224, 1980 (C. Bettanini et al.); US 4 238 223, 1980 (C. Bettanini et al.); US 4 264 778, 1981 (S. E. Ravelli et al.).
44. Brasmag, ZA 8 704 237, 1987.
45. University of Manchester, WO 89/00 613, 1988 (A. M. Cameron).
46. C. Faure, J. Marchal, *Metals* **16** (1964) 721.
47. M. P. Lugagne, *Erzmetall* **31** (1978) 310 – 313.
48. R. A. Christini, *Light Met. (Warrendale, Pa.)* 1980, 981 – 995.
49. Norsk Hydro, US 4 385 931, 1983 (O. Wallevik, J. B. Rønhaug).

50. R. S. Busk, R. B. Jackson, *Proc. Annu. Meet. Int. Magnesium Assoc.* **35th,** 1980.
51. S. L. Couling, *Proc. Annu. Meet. Int. Magnesium Assoc.* **34th,** 1979.
52. Extramet, CA 1 175 618, 1984 (G. Bienvenue, B. Chaleat).
53. J. A. Barannik et al., SU 384 423, 1971.
54. Dow Chemical Co., US 4 186 000, 1980 (E. J. Shack, G. B. Cobel).
55. Norsk Hydro, US 4 421 551, 1983 (U. Müller).
56. N. A. Voronova: *Desulfurization of Hot Metal by Magnesium,* The International Magnesium Association and The Iron and Steel Society of the AIME, Dayton, OH, 1983.
57. S. K. Saxena: *Recent Trends in the Mass Desulphurization of Hot Metal in Integrated Steel Plants*, International Conference on Desulphurization and Inclusion Control, ABM, Volta Redonda, RJ, Brazil, Oct. 1997.
58. *ASM Speciality Handbook, Magnesium and Magnesium Alloys,* ASM International, Ohio 1999.
59. *Magnesium Taschenbuch,* Aluminium-Verlag, Düsseldorf 2000.
60. R. E. Brown, *Light Metal Age* (1998) no. 2, 60 – 63.
61. R. E. Brown, *Light Metal Age* (2000) no. 4, 54 – 55.
62. Noranda Metallurgy Inc., WO 95/31 401, 1995 (J. G. Peacey et al.).
63. Nalco Chemical Co., US 3 966 888, 1976 (D. G. Braithwaite, R. J. Allain).
64. CSIRO, AU 637 971, 1989 (D. Elsner et al.).
65. CSIRO, QMC, WO 9 511 859, 1993 (M. M. Hourn et al.).
66. Norsk Hydro, AU 711 888, 1997 (R. Schmidt et al.).
67. J. F. Moore et al., *Proc. TMS Symp. "Magnesium 2000"*, Nashville, TN 2000, 21 – 26.
68. C. Conradty, DE-OS 3 021 427, 1980 (D. H. Zöllner et al.).
69. General Motors , US 5 665 220, 1995 (R. A. Sharma).
70. Cellmag, WO 00/53 826, 1999 (G. Van Weert).
71. Norsk Hydro, US 4 784 742, 1986 (O. Wallevik et al.).
72. R. R. Lloyd et al., *AIME Technical Publication*, No. 1848 (1945).
73. R. A. Sharma, *JOM* (1996) no. 10, 39 – 43.
74. K. J. Cathro, R. L. Deutscher, R. A. Sharma, *J. Appl. Electrochem.* **27** (1997) 404 – 413.
75. General Motors, US 5 279 716, 1994 (R. A. Sharma).
76. General Motors, US 5 427 657, 1995 (R. A. Sharma).
77. D. E. Woolley, U. Pal, G. B. Kenney, *Proc. TMS Symp. "Magnesium Technology 2000"*, Nashville, TN, 2000, 35 – 36.
78. G. M. Rao, *J. Appl. Electrochem.* **16** (1986) 626 – 628.
79. G. Shekhovtsov et al., *Proc. TMS Symp. "Magnesium Technology 2000"*, Nashville, TN, 2000, 31 – 34.
80. Russian National Aluminum and Magnesium Institute (VAMI), RU 2 131 844, 1998 (A. N. Tatakin et al.).
81. O. Wallevik et al., *Proc. TMS Symp. "Magnesium Technology 2000"*, Nashville, TN, 2000, 13 – 16.
82. J. Christensen et al, *Proc. 1st Israeli Int. Conf. on Magnesium*, Dead Sea, Israel, 1997, 9 – 14.
83. Alcan, WO 96/33 297, 1995 (O. G. Sivilotti, M. Vandermeulen, J. Iseki).
84. A. M. Cameron, *Light Metals* (1997) 579 – 602.
85. G. L. Miller, *Vacuum,* **II** (1952) no. 1, 19 – 32.
86. M. W. Wadsley, *Proc. TMS Symp. "Magnesium Technology 2000"*, Nashville, TN, 2000, 65 – 70.
87. M. Marshall et al., *Proc. TMS Symp. "Magnesium Technology 2001"*, New Orleans, LO, 2001, 17 – 20.
88. Jing Chun Zang et al., *Proc. TMS Symp. "Magnesium Technology 2001"*, New Orleans, LO, 2001, 7 – 10.
89. Xiaoming Mei et al., *Proc. TMS Symp. "Magnesium Technology 2001"*, New Orleans, LO, 2001, 13 – 15.
90. Alcoa, US 5 383 953, 1995 (R. A. Christini).
91. I. Henych: *Some Metallurgical Aspects of Producing Ductile Iron and Holding Treated Melts,* George Fischer, Schaffhausen, Switzerland, 1986.
92. T. Skaland: "A Model for the Graphite Formation in Ductile Cast Iron", Dr. Ing. Thesis, MI-rapport 1992:9, NTH, Trondheim, Norway.
93. K. Zheng et al., *Int. J. of Hydrogen Energy* **24** (1999) 989 – 1004.
94. A. Zaluska et al., *J. Alloys and Compounds* **288** (1999) 217 – 225.
95. A. Zaluska et al., *J. Alloys and Compounds* **289** (1999) 197 – 206.
96. Meditext: Medical Management, Online Version, Micromedex, 1974 – 2000.
97. L. Permaggiani (ed.): *Encyclopedia of Occupational Health and Safety,* 3rd rev. ed., Publisher International Labour Office, Geneva 1983.

98. J. R. Erichsen Jones, *J. Exp. Biol.* **16** (1939) 425 – 437.
99. R. Jokinen, *Met. Ions in Biol. Syst.* **26** (1990) 15.
100. R. K. Eckhoff, *Dust Explosions in the Process Industries,* Butterworth-Heinemann, Oxford 1991.
101. *Fire Protection Handbook,* 18th ed., National Fire Protection Association (NFPA), Quincy, MA, 1997.
102. *Standard for the Storage, Handling, and Processing of Magnesium Solids and Powders, NFPA 480,* National Fire Protection Association, Quincy, MA, 1998.
103. Norsk Hydro, US 5 112 584, 1990 (G. T. Mejdell et al.).
104. Norsk Hydro, US 5 120 514, 1990 (K. W. Tveten et al.).
105. Norsk Hydro, US 3 779 870, 1970 (S. Gudmundsen et al.).
106. I. Barannik et al., *Proc. 1st Israeli Int. Conf. on Magnesium*, Dead Sea, Israel, 1997, 15 – 20.
107. A. A. Nayeb-Hashemi, J. B. Clark: *Phase Diagrams of Binary Magnesium Alloys,* ASM Intl., Metals Park , Ohio, 1988.
108. E. Øvrelid: "Hydrogen in Magnesium and Magnesium Alloys", Dr. Ing. Thesis, Institute of Metallurgy, NTNU, Trondheim, Norway, 1997.
109. W. Schubert, H. Gjestland in K. U. Kainer (ed.): *Magnesium Alloys and their Applications,* Wiley-VCH Weinheim 2000, pp. 761 – 766.
110. Int. Magnesium Assoc., *Technical Committee Report: Recommended Practices for the Conservation of Sulfur Hexafluoride in Magnesium Melting Operations*, 1998.
111. T. K. Aune et al., *Proc. NADCA 97*, NADCA, Rosemont, IL, USA.
112. P. Bakke, D. O. Karlsen, *SAE Technical Paper 970330*, SAE International, Warrendale, PA, USA.
113. H. Hu, A. Luo, *JOM* **48** (1996) no. 10, 47 – 51.
114. J. S. Waltrip, *SAE Technical Paper 900789*, SAE International, Warrendale, PA, USA.
115. R. Berkmortel et al., *Proc. Annu. Meet. Int. Magnesium Assoc.*, 57th, 2000.
116. H. Eklund et al., *Proc. TMS Symp. "Magnesium Technology 2002"*, Seattle, WA, 2002, 9 – 12.

Magnesium Alloys

Terje Kr. Aune, Norsk Hydro ASA, Porsgrunn, Norway
Haavard Gjestland, Norsk Hydro ASA, Porsgrunn, Norway
Johanna Øster Haagensen, Norsk Hydro ASA, Porsgrunn, Norway
Bjørn Kittilsen, Norsk Hydro ASA, Porsgrunn, Norway
Jan Ivar Skar, Norsk Hydro ASA, Porsgrunn, Norway
Håkon Westengen, Norsk Hydro ASA, Porsgrunn, Norway

1. Introduction 1
2. Commercial Alloys, Product Forms, and Uses 1
3. Metallurgy of Magnesium Alloys . . . 2
3.1. Magnesium – Aluminum Alloys 3
3.2. Grain-Refined Magnesium – Zirconium Alloys 5
3.3. New Developments 5
4. Production Methods for Magnesium Alloys 6
5. Manufacture of Articles from Magnesium Alloys 7
5.1. Shaping 7
5.2. Joining 8
5.3. Finishing 8
6. Corrosion and Corrosion Control . . . 9
6.1. Effect of Composition and Structure . 10
6.2. Effect of the Environment 11
6.3. Galvanic Corrosion 11
7. Environmental Aspects 12
8. Economic Aspects 13
9. References 13

1. Introduction

Most primary magnesium is used for metallurgical purposes: in alloying additions to aluminum and other metals, in magnesium-based alloy castings and wrought products, for desulfurization of iron and steel, for nodularization of ductile iron, as a reducing agent to produce metals from their oxides and chlorides, and in anodes for cathodic protection. Chemical uses include the production of magnesium compounds and as a catalyst in Grignard reactions. Its low density (ca. 1.8 kg/dm^3 in alloy form), coupled with favorable casting properties, makes magnesium the material of choice for lightweight components. Magnesium-based alloy castings and wrought products make up about 37 % of free-market magnesium shipments.

Information on the properties, production, fabrication, and use of magnesium alloys can be found in [4–9], [28–32]. A compilation of alloy properties and product design information is given in [2], [28], [30], [32]. Commercial magnesium alloy products are described in [3], [33], [34].

2. Commercial Alloys, Product Forms, and Uses

Designations for cast and wrought magnesium alloys, in accordance with ASTM B 275 [3], are based on their chemical composition. Letters represent the principal alloying elements:

A	aluminum	M	manganese
C	copper	Q	silver
E	rare earth metals (RE)	S	silicon
G	magnesium	T	tin
H	thorium	W	yttrium
K	zirconium	Z	zinc
L	lithium		

These are followed by respective weight percentages rounded to whole numbers and a serial letter that differentiates slight variations in compositions. Temper designations are in accordance with ASTM B 296 [3]:

F	as fabricated
O	annealed, recrystallized (wrought products only)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_581

H	strain hardened
	H1 strain hardened only
	H2 strain hardened and then partially annealed
	H3 strain hardened and then stabilized
T	thermally treated to produce stable tempers other than F, O, or H
	T4 solution heat treated
	T5 cooled and artificially aged
	T6 solution heat treated and artificially aged
	T7 solution heat treated and stabilized
	T8 solution heat treated, cold worked, and artificially aged

The compositions of a range of casting and wrought alloys are listed in Table 1. Alloys for die casting contain very small amounts of the heavy-metal impurities nickel, iron, and copper [3], [33], [34].

Table 3. Characteristics and typical uses of some magnesium alloys

Alloy	Characteristics and uses
Cast Alloys	
AM50A/AM60B	high-pressure die-casting alloy with outstanding ductility in the -F condition; used for automobile parts
AS41A/B	die-casting alloy with good creep properties up to 150 °C; used for automobile parts
AZ81A, AZ91C/E, ZE41A	general-purpose sand and permanent mold-casting alloy used for aircraft parts, machinery components, gearboxes
AZ91B/D	general-purpose die-casting alloy for automobile and computer parts, chain saws, sporting goods, cameras, projectors, and household equipment and appliances
EZ33A	high-temperature alloy for sand and permanent mold casting; excellent castability; creep resistant up to 250 °C and pressure-tight; used in aerospace and military applications
Wrought Alloys	
AZ31B/C	moderate strength general-purpose alloy used for luggage, ladders, hand trucks, bakery racks, and concrete tools
AZ61A, AZ80A, ZK60A	higher strength alloys used in batteries, military components, shelters, tent poles, and tennis racquets
HK31A	elevated-temperature alloy used in aerospace and military applications

Typical mechanical properties of cast and wrought alloys are listed in Table 2. The physical properties of magnesium alloys are in the following range (RT = room temperature):

Density (RT)	1.7 – 1.8 g/cm^3
Electrical resistivity (RT)	50 – 170 nΩ·m
Heat of fusion	280 – 373 kJ/kg
Melting range	445 – 650 °C
Specific heat (RT)	960 – 1050 J kg^{-1} K^{-1}
Linear coefficient of thermal expansion (RT)	$(26.5-27.3)\times 10^{-6}$ K^{-1}
Elastic modulus (RT)	45 GPa

A general description of alloy characteristics and typical uses is presented in Table 3. Equivalent alloy designations used in industrialized countries are given in [33], [34], [50].

3. Metallurgy of Magnesium Alloys

[1], [2]

Magnesium in its pure form is soft, mechanically weak, and hence not generally used for structural applications. By careful selection of alloying elements, both general-purpose alloys and alloys for special applications can be produced. As with other metallic alloy systems, a combination of well-known hardening mechanisms (solid solution hardening, particle dispersion hardening, work hardening, and grain boundary hardening; → High-Temperature Materials, Chap. 3.) determines the mechanical properties of magnesium alloys. Alloying additions influence other properties including reactivity of the melt, castability, and corrosion performance.

The alloying elements used with magnesium can be grouped into two categories :

1) Elements that actively influence the melt include beryllium (≤ 30 ppm), which lowers the rate of melt oxidation, and manganese (≤ 0.6 wt %), which reduces the iron content and hence the corrosion rate of the alloys. These additions are active in relatively small amounts and do not require extensive solubility in the melt.
2) Elements that modify the microstructure of the alloy by via the above-mentioned hardening mechanisms include those that influence castability. Except for grain-refining additions (carbon inoculation), which can be active in small amounts, these elements must be relatively soluble in liquid magnesium. Commercially interesting alloying elements include aluminum, cerium, copper, lanthanum, lithium, manganese, neodymium, silver, yttrium, zinc, and zirconium.

Table 1. Typical composition (wt %) of some cast and wrought magnesium alloys [3], [33], [34]

Alloy	Form ∗	Al	Zn	Mn	Si	RE	Zr	Th
AM50A	CD	5		> 0.26				
AM60B	CD	6		> 0.24				
AS21	CD	2		> 0.2	1			
AS41B	CD	4		> 0.3	1			
AZ31B	WB +	3	1					
				0.3				
	WS							
AZ61A	WF	6	1					
				0.2				
AZ80A	WB	8	0.5					
				0.2				
AZ81A	CS	7.5	0.7					
				0.2				
AZ91D	CD	9	0.7	> 0.15				
AZ91E	CS	9	0.7					
				0.2				
EZ33A	CS		3			3		
							0.8	
HK31A	WS							3
							0.7	
ZE41A	CS		4			1		
							0.7	
ZK60A	WB		6				> 0.45	

∗ CS = sand casting; CP = permanent mold casting; CD = die casting; WS = sheet or plate; WF = forging; WB = bar, rod, shape, tube, or wire

Table 2. Mechanical properties (at room temperature) of selected magnesium alloys [2]

Alloy	Form∗	Temper	Tensile strength, MPa	Yield strength, MPa	Elongation in 50 mm, %
AM50	CD	-F	200	110	10
AM60	CD	-F	220	130	8
AS41	CD	-F	210	140	6
AZ31	WB	-F	260	200	15
AZ61	WB	-F	310	230	16
AZ80	WF	-T6	345	250	5
AZ81	CS	-T4	276	85	15
AZ91	CD	-F	230	160	3
	CS	-T6	275	130	5
EZ33	CP	-T5	140	95	2
HK31	WF	-T5	260	195	21
ZE41	CS	-T5	205	140	5
ZK60	WB	-T5	360	295	12

∗ For definitions of abbreviations see Table 1

3.1. Magnesium – Aluminum Alloys

Aluminum is by far the most important alloying element for magnesium. Its maximum solid solubility in the Mg – Al system is 12.7 wt % at the eutectic temperature of 437 °C (Fig. 1). The eutectic contains 32 wt % aluminum, and it consists of α-Mg and $Mg_{17}Al_{12}$. Commercial alloys contain less than 10 wt % aluminum and, according to the equilibrium diagram, should solidify into a homogeneous matrix of magnesium with aluminum in solid solution. This is, however, not the case – a relatively large volume fraction of eutectic constituents is formed (Fig. 2). The solidification diagram shows a distinct plateau at the eutectic temperature (Fig. 3).

Scheil's equation describes freezing under nonequilibrium conditions where solute diffusion is negligible, so that a strongly segregated material results. The fraction solidified as a function of temperature for equilibrium and nonequilibrium freezing is illustrated in Figure 4 for two Mg – Al alloys. In practice, the Scheil equation

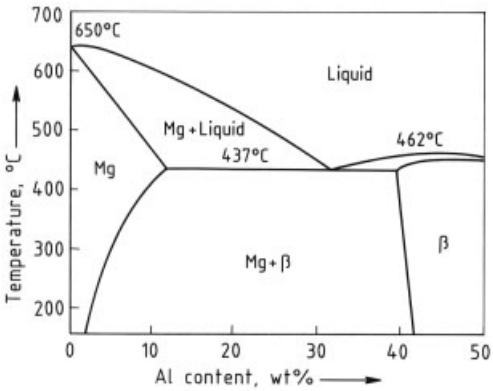

Figure 1. Phase diagram for the binary system Mg – Al

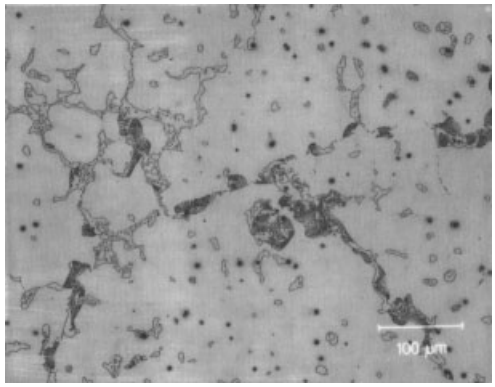

Figure 2. Microstructure of low-pressure die-cast AM100 alloy (× 200)

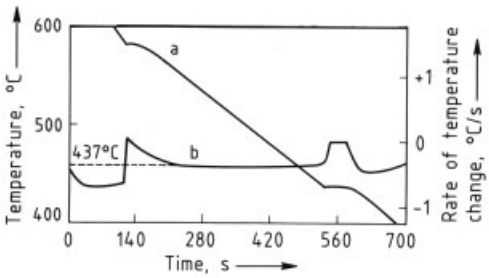

Figure 3. Solidification diagram showing the temperature variation during cooling of the Mg – Al alloy AM 100 A
a) Temperature; b) Rate of temperature change

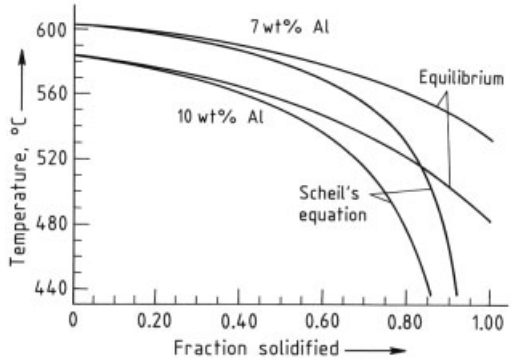

Figure 4. Fraction solidified as a function of temperature, calculated for equilibrium and nonequilibrium conditions for two Mg – Al alloys

provides a reasonable description of solidification during commercial casting processes.

The Mg – Al equilibrium diagram is typical of an age-hardenable system. After solution heat treatment at temperatures just below the eutectic temperature, the β-$Mg_{17}Al_{12}$ phase dissolves, and subsequent quenching gives a supersaturated solid solution. During artificial aging at 150 – 220 °C, a platelike precipitate forms and results in a significant hardening effect. A certain quench sensitivity is caused by a competing reaction, whereby the β-phase precipitates discontinuously along grain boundaries. This process occurs during cooling after casting or following solution heat treatment.

Commercial Mg – Al alloys usually contain a small amount of zinc (0 – 2 wt %), which strengthens them somewhat. Higher zinc contents increase the amount of eutectic constituents, lower the solidification temperature, and may cause hot shortness (cracking). Solution treatment of Mg – Al – Zn alloys increases their ductility. In the absence of solution heat treatment (as with high-pressure die-cast products), maximum ductility is obtained by using alloys of low aluminum and zinc content (e.g., AM60 alloy containing 6 wt % aluminum). Further improvement of ductility is possible with even less aluminum, but castability suffers somewhat.

High-temperature applications require improved creep properties, which are achieved by decreasing the aluminum content and thus the volume fraction of the low-melting eutectic. Fur-

ther improvements are possible by introducing elements that form finely dispersed particles in the matrix (dispersion hardening). In high-pressure die-cast parts that solidify rapidly, silicon addition causes the formation of finely dispersed MgSi particles; this is the basis of Mg – Al – Si alloys. Additions of rare earth elements and calcium have a similar effect on the microstructure and creep properties.

Manganese addition plays a crucial role in controlling the iron content by reducing the solubility of iron in the melt. Several high-purity alloys with greatly improved corrosion resistance are available commercially. Grain refining, when necessary, is performed by introducing carbon into the melt at > 750 °C (preferably ca. 780 °C).

3.2. Grain-Refined Magnesium – Zirconium Alloys

A series of high-performance magnesium-base alloys was developed after discovery of the extremely efficient grain-refining action of zirconium. This effect results from formation of a high density of tiny zirconium particles in the melt, which act as potent grain nuclei. An example of a zirconium-refined grain structure is shown in Figure 5. Unfortunately, elements such as aluminum and manganese lower the solubility of zirconium in the melt, thus precluding exploitation of this grain-refining mechanism in common aluminum-containing alloys.

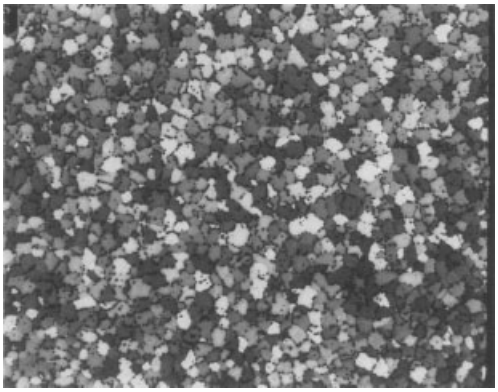

Figure 5. Structure of a zirconium grain-refined alloy RZ5 (× 50)

Mg – Zn – RE – Zr. Binary Mg – Zn alloys show inferior mechanical properties and castability, but the addition of zirconium for grain refinement and rare earth (RE) elements to reduce microporosity led to the development of alloys such as EZ33 and ZE41. These alloys have been used widely as sand castings in the T5 condition (cooled and artificially aged) for applications involving exposure to moderately elevated temperature (up to ca. 200 °C). The ZE63 alloy is a high-strength variant in which solution heat treatment in a hydrogen atmosphere is employed to remove some of the embrittling Mg – Zn – RE grain boundary phases. In extrusion, microporosity is of less concern and Mg – Zn – Zr alloys are used without RE additions.

Mg – Ag – RE – Zr. Addition of silver and neodymium-rich mischmetal, in addition to zirconium, promotes age hardening efficiently when the alloy is given a full T6 treatment (solution heat-treated and artificially aged). The precipitates formed are relatively stable; the alloy shows high mechanical strength and good creep properties at temperatures approaching 250 °C.

Mg – Y – RE – Zr. This alloy system for high-temperature application is based upon additions of yttrium and neodymium-rich misch metal. This alloy system was developed to meet the ever-increasing high-temperature performance requirements of the aerospace industry and may replace the Mg – Ag – RE – Zr and thorium-containing alloys. The alloys must be melted in a special inert atmosphere (argon) and develop their properties by T6 heat treatment.

3.3. New Developments [11], [29], [31]

Conventional alloy development has reached a level at which most of the foreseeable potential has been exploited. Alloy modifications and the development of new alloys for die casting are mainly linked to high-temperature applications. Further improvements are possible by combining existing alloys or modified alloys with new processing routes. Squeeze casting and semisolid alloys are examples of active fields of research. Microstructural refinement through rapid solidification offers great potential for improving the properties of both conventional al-

loys and new alloy compositions. The ability of magnesium alloys to wet a number of reinforcing fibers and particulates efficiently makes the material attractive as a matrix in metal-based composites.

4. Production Methods for Magnesium Alloys

Alloying and Grain Refining. The most commonly used magnesium alloys contain aluminum, zinc, and manganese; aluminum and zinc are added in ingot form, dissolving readily at normal melt temperature (ca. 700 °C). Manganese is added as the anhydrous chloride or as a hardener, for example, Al – 5 % Mn. The alloying elements are added at 720 – 750 °C; the alloy is homogenized and then cast.

High-purity magnesium alloys such as AZ91D, AZ91E, AM60B, and AM50A contain very low contents of the heavy-metal impurities nickel, iron, and copper, and hence have far better corrosion resistance than ordinary alloys. Iron is removed by saturating the melt with manganese at elevated temperature (700 – 750 °C) and then loweringthe temperature, thus precipitating the iron in the form of an AlMnFe intermetallic compound.

Mechanical properties of magnesium alloys improve as grain size is reduced. Carbon inoculation is used commonly to grain-refine aluminum-containing magnesium alloys, the inoculant being added at ca. 780 °C. Zirconium is frequently used to grain-refine alloys lacking aluminum. It is added as a Mg – 30 % Zr hardener above 750 °C. The mechanisms of grain refining are not fully understood, but several theories have been proposed. Both zirconium and magnesium have hexagonal crystal structures with similar lattice parameters, making zirconium an ideal nucleating particle for magnesium.

Casting of Billets and Slabs. Fabrication of wrought magnesium products starts with the production of extrusion billets, forging billets, and rolling slabs, which are usually cast by continuous or semicontinuous direct-chill cast processes. The methods are similar to those used for alloys of other metals such as aluminum (→ Aluminum Alloys, Chap. 3.2.) and copper. Multistrand casting with downspout tubes and floats has proved difficult for magnesium.

Shaped Casting. Magnesium alloys can be cast by all known processes, including die and sand casting. Remelting and recycling of magnesium alloys are economically viable.

High-Pressure Die Casting. Two methods are used for high-pressure die casting: the molten metal is injected at high velocity into a reusable metal mold (die) in cold-chamber (as for aluminum) or hot-chamber (as for zinc) machines. Because of their lower specific heat (i.e., heat content) and thus faster solidification, production rates of magnesium high-pressure die castings are usually 30 – 40 % faster than those of aluminum alloys. In hot chamber machines the production rates for medium to large components are similar to those for zinc. Parts weighing up to 20 kg are cast in 50- to 5000-t locking capacity cold-chamber machines at a rate of 20 – 60 shots per hour. For smaller and thin-walled parts, hot chamber machines of 50- to 1000-t capacity are common, with production rates up to 400 shots per hour. Molten magnesium alloy is automatically metered and transferred in closed systems [12], [35].

Because magnesium does not attack steel to the same degree as aluminum, dies used for casting magnesium have lifetimes up to several times longer than those for aluminum. Draft angles for magnesium are lower than for aluminum, allowing casting of holes within tolerance limits.

Low-Pressure Casting. Low-pressure casting of magnesium has been performed successfully [13]. Typical production rates of 12 – 15 shots per hour are achieved when casting a 12-kg part in steel dies. The process is also applicable to casting in sand molds. The properties of the castings can be improved by heat treatment, making this method useful for a wide range of applications.

Sand Casting. In sand casting, molds are formed by ramming conventional naturally bonded molding sand around a pattern. Chemical inhibitors such as inorganic fluorides or boric acid are added to the sand to suppress the tendency for molten magnesium to react with moisture. Sometimes dry sand molds or skin-dried green sand molds (i.e., surface-dried by torching) are used to reduce the oxidation tendency

further. The advent of chemically bonded dry silica sand has allowed closer tolerances and the production of more complex castings.

Melt Protection [46], [47]. Due to magnesium's high affinity for oxygen and high vapor pressure, the molten metal will oxidize and burn in air. This problem can be solved by using salt fluxes and/or protective gases mixed with the atmosphere above the melt. However, salt fluxes tend to lower the corrosion resistance due to salt contamination, and therefore gas protection is currently used by most die-casters. Various gases can protect magnesium from oxidation. The most commonly used gas for melt protection is sulfur hexafluoride SF_6. In die-casting operations a concentration of 0.2 % in dry air (sometimes mixed with CO_2) is used. SF_6 has recently been identified as a strong greenhouse gas and should therefore be replaced. A substitute is sulfur dioxide, which was used by the magnesium industry in the past. Although SO_2 is toxic at higher concentrations, several die-casters never converted to SF_6, and their experience shows that the gas can be applied without environmental problems.

Wrought Products. Wrought products are obtained by working a metal into the desired shape. Although the cold-forming properties of magnesium are very limited because of its hexagonal structure, magnesium alloys can be hot-worked readily by most of the well-known methods, such as extrusion, rolling, and forging.

Extrusion [36]. Extrusion presses designed for aluminum alloys can, after minor modifications of tooling, extrusion speed, and billet temperature, be used to extrude magnesium alloys. Compared with the 6000 series aluminum alloys, most magnesium alloys require a higher extrusion pressure (850 – 900 MPa) and thus a container of slightly smaller diameter. Magnesium alloys respond differently to extrusion speed. Alloys containing relatively large amounts of aluminum, such as AZ61A and AZ80A, are limited to maximum speeds of 5 and 3 m/min, respectively. At higher speed the temperature of the extrusion increases and hot cracking may occur. The AZ31B alloy, however, is extruded at speeds up to 20 m/min. As a general rule, an increase in the total amount of alloying elements results in lower extrusion speed. For some alloys (e.g., ZM21) an increase in extrusion speed results in lower mechanical properties. Therefore, selection of extrusion parameters is frequently a compromise between properties and productivity. Currently, AZ31B is the most popular moderate-strength extrusion alloy; AZ61A and AZ80A have higher mechanical properties and are often artificially aged for additional strength. The zirconium-containing alloys ZK40A and ZK60A are used when the yield strength is of prime importance. A new Mg – Zn – Cu – Mn alloy, ZCM711, is heat treatable to tensile yield strength values as high as 350 MPa in the T6 temper [14].

Rolling. Magnesium alloys rolled on conventional equipment are available commercially as flat sheet and plate, tooling plate, and tread plate [15]. Specialty items such as photoengraving sheet and plate, mechanical engraving plate, and electrochemical sheet are also available. Fine grain size and a good dispersion of any second phase are necessary for good rollability. Therefore, zirconium-containing alloys and the single-phase M1A alloy roll well.

Forging. In forging, the metal is worked with hammers and presses. Magnesium is one of the more forgeable metals. Of all magnesium product forms, forgings have the best combination of mechanical properties. They are also pressure-tight and have excellent machinability.

5. Manufacture of Articles from Magnesium Alloys

Magnesium products offer many advantages to the designer of light structures. In addition to the obvious weight-saving benefit, magnesium products provide manufacturing advantages such as excellent castability, good hot forming of wrought products, high weld efficiency, and unsurpassed machinability.

5.1. Shaping

Machining. Magnesium and magnesium alloys are the easiest of all metals to machine, permitting machining operations at extremely high speed. All standard machining operations such as turning, drilling, milling, tapping, threading, counterboring, reaming, sawing, and grinding

are commonly performed on magnesium parts. Deeper cuts can be made and higher feed rates used than with other metals. The life of high-speed steel cutting tools for magnesium equals that of carbide-tipped tools used for other metals. The excellent surface finish of machined magnesium is the result of its minimal tearing and dragging tendencies. The free cutting characteristic of magnesium produces well-broken chips that do not obstruct the cutting tool or machine. Standard machining operations produce extremely good tolerances. The power requirements to remove a given volume of metal are lower for magnesium than for any other metal; removable volumes are 0.024, 0.082, and $0.150\,cm^3\,W^{-1}\,min^{-1}$ for mild steel, aluminum, and magnesium, respectively.

A major difference with regard to other metals is that dry machining of magnesium is common practice. In especially difficult operations a coolant such as compressed air or mineral oil is sometimes used to assist in chip removal or to cool the workpiece. Some high-volume machining of magnesium automobile parts is performed with inhibited aqueous coolants, which are mandatory to retard the formation of hydrogen by reaction of magnesium with water.

When grinding or polishing magnesium, special care must be taken to avoid ignition of fine particles; fine dust should be collected in water-washed dust collectors.

Forming. Because of their hexagonal crystal structure, the cold forming of magnesium alloys is restricted to mild deformation and generous bend radii (bend radius : material thickness > 2). However, above 200 °C magnesium becomes very workable; formability at 200 – 300 °C improves to such an extent that more severe working is possible than with other metals at room temperature. (The workability of the cubic system metals is not greatly enhanced by increasing temperature.)

Working of metals at elevated temperature has several advantages over cold working. In some configurations, magnesium can be formed in one operation, thus avoiding repeated annealing and reforming, and reducing production time. Hot-formed parts are made to closer dimensional tolerances because of reduced springback.

5.2. Joining

Magnesium alloys are joined by all conventional methods including welding, brazing, soldering, adhesive bonding, riveting, and bolting.

Welding is carried out almost exclusively by the tungsten inert gas and the metal inert gas methods. In general, weld efficiencies are high.

In magnesium alloys the success or reliability of *adhesive bonding* depends greatly on the environment to which the parts are to be exposed. A conversion coating is required before joining. When exposed to severe environments such as salt and moisture, joints must be sealed. A combination of adhesive bonding and riveting is sometimes used.

Riveting is widely used for making mechanical joints. Standard rivet shapes are employed, spaced so that the distance between them does not exceed three times the rivet diameter. Aluminum 5000 series alloy rivets are preferred. When bolts, screws, and similar standard fasteners are used to join magnesium to other metals and materials, particular attention is given to joint design and the avoidance of galvanic corrosion.

5.3. Finishing

As with other structural metals, magnesium alloys are given a variety of surface treatments for decorative, functional, or protective purposes. Processes are available to meet almost any practical requirement, including chemical and electrochemical conversion coatings, wet painting, resin powder coating, and electroless nickel or electrolytic metal plating [5].

In some applications, relatively comprehensive surface treatments are specified. Choice of treatment depends upon the ultimate use of the product with regard to corrosion, wear, durability, and appearance.

Mechanical Preparation. The usual methods such as belt sanding, wheel grinding, polishing, brushing, buffing, blasting, and barrel or vibratory finishing are applied to magnesium products. Vibratory finishing with ceramic or resin-bonded chips is widely used in mass production for both deburring and smoothing purposes.

Grinding, polishing, and buffing methods used for aluminum and zinc are also used with magnesium. Because of its natural tendency to tarnish, polishing and buffing are used only to prepare magnesium alloy surfaces for plating or for clear lacquering.

Dry blasting of magnesium is usually avoided because most of the commonly used blasting media contaminate the surface with cathodic particles, resulting in deterioration of corrosion resistance. The least adverse effects are experienced with cut 5000 series aluminum alloy wire, which has a low iron and copper content. To remove surface contamination, blasting is followed by acid pickling which removes at least 0.05 mm of the prepared surface.

To minimize fire hazards, magnesium dust must not be allowed to accumulate in the working area, in suction ducts or pipes, or on operators' clothing. Well-grounded, explosion-proof equipment is used in conjunction with approved wet-type dust collectors.

Cleaning. The most common methods used for cleaning magnesium are alkaline cleaning and acid pickling (→ Metals, Surface Treatment). Because of the alkali resistance of most magnesium alloys, proprietary steel cleaners based on alkali metal hydroxides, carbonates, silicates, and pyrophosphates are normally suitable for magnesium. Acid pickling is used to remove surface contamination by sand particles, corrosion products, residues from blasting, chromate coatings, and burned-on lubricants.

Conversion Coatings. Conversion coatings react with the alloy surface to convert it into a component of the coating. Their primary function is to improve the adhesion of organic coatings and to prevent undercutting in corrosive environments at sites where the coating has become damaged. Chromate conversion coatings are most commonly used (→ Electrochemical and Chemical Deposition, Chap. 11.). Certain phosphate conversion coatings provide a satisfactory paint base for many corrosive environments [16] (→ Metals, Surface Treatment). Phosphate treatment has been used for many years as a base for organic coating of extruded and die-cast magnesium luggage frames [37]. Anodic coatings are used when a heavy-duty paint base or added protection against wear is required [17] (→ Electrochemical and Chemical Deposition, Chap. 12.2.2.).

Organic Coatings. Various types of organic coatings can be applied to magnesium alloys by conventional methods. However, consideration must be given to the alkaline nature and electrochemical properties of the metal in selecting primer vehicles and pigments; an alkali-resistant binder and suitable pigmentation must be used. To achieve optimum protection against corrosion, inclusions from blasting processes should be removed by acid pickling before further chromate and finishing steps are taken.

Use of a high-purity alloy is necessary to obtain maximum corrosion protection, especially where damage from stone chipping and scratches is likely to occur. The corrosion properties of painted high-purity alloy AZ91D and ordinary AZ91B are compared in Table 4.

Primers used on magnesium alloys should be based on alkali-resistant vehicles (e.g., vinyl, epoxy, polyvinylbutyral, acrylic, polyurethane, and vinyl epoxy). Zinc chromate and strontium chromate are commonly used pigments, whereas aluminum, zinc, lead, or iron pigments are avoided. For application above 200 °C, primer and paint vehicles based on silicone, epoxy–silicone, or polyimide are used. Cathodic (cataphoretic) electropainting is used on magnesium both as a primer and as a final finish (e.g., for automotive grills and small computer parts). For a detailed description of paints, see → Paints and Coatings.

Plating [5], [18]. Most cast or wrought magnesium alloys can be electroplated. The basic process is similar to the plating of aluminum in that both alloys are given a zinc immersion coating prior to plating (→ Electrochemical and Chemical Deposition, Chap. 5.3.6.).

6. Corrosion and Corrosion Control

(→ Corrosion)

Factors that affect the corrosion properties of magnesium are alloy composition, heavy-metal impurities, casting variables, metallographic structure, environment, surface condition, and contact with other materials. Thermodynamically, magnesium should react completely with

Table 4. Salt spray performance of painted die-cast AZ91B (standard) and AZ91D (high-purity) test panels (DIN 53 167)

Alloy	Paint system ∗	Exposure time, h	Scribe corrosion creepage, mm	Stone chip resistance rating no.∗∗
AZ91B	cathodic electropainting, 18 – 22 µm	240	7	
AZ91D	cathodic electropainting, 18 – 22 µm	240	0	
AZ91B	TiO_2 pigmented polyester powder, 80 – 120 µm	240	3	4
AZ91D	TiO_2 pigmented polyester powder, 80 – 120 µm	1000	0.5	1 – 2
AZ91D	zinc chromate-inhibited epoxy primer with TiO_2 pigmented polyester powder (80 – 120 µm) as top coat	1000	0	1

∗ All samples are vibratory finished and chrome pickled
∗∗ Rating no.: 1 = excellent; 4 = poor

oxygen and with water. The fact that it does neither is caused by passive behavior in many environments. In a corrosive environment, pitting or other forms of local corrosion occur as a result of film breakdown.

Magnesium alloys of suitable composition (see Section 6.1) and purity are corrosion resistant and are used successfully in a variety of applications. Automotive parts made of commercial high-purity die-cast AZ91D alloy located in the worst splash zone under a car illustrate the excellent corrosion resistance of magnesium [19].

6.1. Effect of Composition and Structure

Composition. The corrosion of magnesium alloys is commonly measured in salt solution either as continuous salt spray or in cyclic corrosion tests. For the latter, the samples are exposed to combinations of salt spray, high humidity, and drying conditions. These tests are designed to simulate real-life exposure.

Most of the elements present in, or added to, magnesium alloys have limited solid solubility in the alloy and therefore occur as precipitated phases. In virtually all cases, these phases are more noble (i.e., have a higher redox potential) than the matrix. Their influence on saltwater corrosion depends heavily on their potential relative to the matrix and their efficiency as cathodic sites, i.e., the ease with which they liberate hydrogen gas (overvoltage) [20].

The elements generally present in commercial magnesium alloys that influence saltwater corrosion can be classified as follows:

1) Generally benign or beneficial: aluminum, beryllium, manganese, rare earth metals, silicon, zinc, and zirconium
2) Moderately deleterious: silver
3) Severely deleterious: nickel (and cobalt), iron, and copper

The commercially important Mg – Al – Zn alloys used for die casting and sand casting have received intensive study [21], [22], resulting in the development of alloys with outstanding saltwater corrosion resistance. These alloys have a very low critical impurity content (Ni, Fe, Cu) and a controlled manganese content.

Structure. The size and distribution of the cathodic phases play an important role in corrosion and are influenced by process parameters and heat treatment. The influence of microstructure on the corrosion of die-cast AZ91D is described in [48]. Homogenized and artificially aged specimens of AZ91E (T6) show considerably lower corrosion than cast (F) and homogenized (T4) specimens [23]. Heat treatment influences mainly the distribution of the intermetallic β-phase ($Mg_{17}Al_{12}$) in the alloy. Aging to T6 temper causes precipitation of this phase as an almost continuous network of secondary particles along the grain boundaries. In the T4 condition the β-phase is fully dissolved. By air cooling from T4, only traces of β-phases can have the same effect as a full T6 treatment. Tolerance limits in cast magnesium alloys for the most important impurity elements (iron, copper, and nickel) are influenced by the cooling rate and the amount of manganese present [21], [22], [24], [38]. In the early stages of corrosion, filiform attack develops from an initiating pit adjacent to intermetallic particles [25], [26], and the

role of $Mg_{17}Al_{12}$ concentrated in grain boundaries can be clearly illustrated. Cold working of magnesium alloys (e.g., by stretching or bending) has no appreciable effect on corrosion rate [4].

Surface Contamination. Producers of magnesium have demonstrated the importance of high-purity alloys for structural applications. However, surface contamination from handling and mechanical treatment can destroy the corrosion resistance of high-purity alloys. This helps explain why ceramic blasting media containing iron oxide can be just as harmful to the corrosion properties of magnesium as steel grit.

6.2. Effect of the Environment

Atmosphere. A magnesium alloy surface exposed to a salt-free atmosphere develops a gray film consisting mainly of magnesium hydroxide that protects the metal from corrosion. Chlorides, sulfates, or other hydrophilic substances promote corrosion by destroying this film. Structural magnesium alloys are resistant to rural atmospheres and moderately resistant to industrial or mild marine atmospheres. The corrosion rate in marine atmospheres is significantly lower for the high-purity Mg – Al – Zn alloys.

Water. When magnesium is immersed in distilled water without the possibility of carbon dioxide absorption, the initial corrosion rate decreases rapidly to a very low value. A protective film of magnesium hydroxide forms on the surface. The solubility product of magnesium hydroxide in the solution is quickly reached, dissolution of the hydroxide is inhibited, and corrosion essentially stops. If the water is replenished, corrosion continues and increases on absorption of carbon dioxide due to dissolution of the protective film. Raising the temperature of distilled or natural water also increases the corrosion rate of magnesium alloys. Aluminum is beneficial as an alloying ingredient because it promotes the formation of protective hydrotalcite $[Mg_6Al_2(OH)_{16}CO_3 \cdot 4\,H_2O]$ films.

Acids. Magnesium is attacked by all acids except hydrofluoric or chromic acid. Passive films are formed in most concentrations of these acids, accounting for their use in many conversion coating processes.

Salt Solutions. Neutral solutions of salts of heavy metals such as nickel, iron, and copper are corrosive to magnesium alloys.

Other Media. The behavior of magnesium in contact with a large number of substances at room temperature is described in [2]. Being an alkaline-earth metal, magnesium is stable in most basic and neutral solutions. Its strong affinity for fluorine and the nature of the resulting passive film (mainly magnesium hydroxide with some magnesium fluoride) account for the extremely low corrosion of magnesium alloys in most fluoride-containing electrolytes. Magnesium is stable in the presence of most organic compounds. However, methanol, being slightly acidic, attacks magnesium rapidly unless enough water is present ($> 0.5\,\%$) to passivate the surface. Magnesium can be safely and effectively cleaned with chlorinated organic solvents.

6.3. Galvanic Corrosion [27], [49]

Two conditions must be satisfied for galvanic corrosion to occur: (1) dissimilar metal-to-metal contact and (2) bridging of the bimetal junction by a conductive solution (electrolyte). The electrochemical process that normally occurs at the galvanic couple containing magnesium is:

$$\text{Anode: Mg (metal)} \rightarrow Mg^{2+} + 2\,e^-$$
$$\text{Cathode: } 2\,H_2O + 2\,e^- \rightarrow H_2 + 2\,(OH)^-$$

In salt environments, the high solubility and acidic nature of the magnesium chloride formed at the anode can result in rapid penetration of magnesium alloys.

Proper protection against galvanic corrosion begins with good design [2], [4]. This includes good drainage to prevent entrapment of electrolyte. Furthermore selection of the most compatible metals, sealing of faying surfaces, small ratios of cathode to anode area, and use of alkali-resistant barrier coatings are important.

The closest approach to compatibility with magnesium is provided by aluminum alloys of

Table 5. Average annual shipments of primary magnesium according to major end use in free market countries ∗

End use	Average annual shipments, 10^3 t		Market distribution, %	
	1995	1999	1995	1999
Magnesium alloys				
Die castings	64.1	133.4	21.1	35.5
Gravity castings	1.8	2.0	0.6	0.5
Wrought products	4.2	4.1	1.4	1.1
Electrochemical	10.6	11.2	3.5	3.0
Subtotal	80.7	150.7	26.5	40.1
Pure magnesium				
Aluminum industry	157.1	159.8	51.7	42.6
Nodular iron	14.5	8.9	4.8	2.4
Desulfurization	36.3	41.7	11.9	11.1
Metal reduction	3.9	2.4	1.3	0.6
Chemical	6.5	5.2	2.1	1.4
Subtotal	218.3	218.0	71.9	58.1
Unclassified	5.0	6.8	1.6	1.8
Total	**304.0**	**375.5**		

∗ As reported by the International Magnesium Association, McLean, Virginia.

the 5000 – 6000 series, and nowadays fasteners made of AA6013 are available [39]. Tin, cadmium, and zinc platings on steel fasteners reduce galvanic action on magnesium in salt spray by 60 – 70 %, which is sufficient for many practical applications. Supplementary polymer coatings and inorganic sealers can reduce galvanic corrosion further [40], [41]. A combination of properly plated steel fasteners and aluminum washers has proven to be very efficient in preventing galvanic corrosion.

Additional resistance in the metallic or electrolytic portions of the galvanic cell circuit can reduce or eliminate galvanic current flow. Such resistance can be supplied by insulating materials such as nonmetallic bolts, insulating washers or tapes, or organic coatings.

The combination of a small magnesium anode and a large cathode area can lead to intense corrosion penetration of the magnesium. In painting a galvanic couple, the cathode or the entire couple must be coated. In no case should the magnesium alone be coated. Small areas of magnesium exposed by paint defects or scratches could be subjected to intense corrosion penetration.

7. Environmental Aspects

The Swedish Ecocycle Commission of 1997 [42] gave magnesium, together with three other metals, a positive evaluation with respect to sustainability criteria such as resource friendliness, recyclability, and toxicity.

Mining of magnesium-containing raw materials such as dolomite and magnesite affects limited areas, and there are no significant disposal or effluent problems. The minerals are transported to the plants by sea. Electrical energy is used in electrolysis. In addition some steam and fossil fuels are used in the process.

The main contribution to the equivalent total CO_2 emission from magnesium production is the use of SF_6 as a cover gas to protect liquid metal surfaces from oxidation in casting operation. SF_6 is nontoxic and odorless but has a strong Global Warming Potential (GWP = 23 900). There is a significant focus on reducing the use of SF_6 in the magnesium industry.

Solid waste consists mostly of inorganic salts and minerals from raw materials. Magnesium process scrap (class 1) is recycled. Typically 30 – 50 % of the shot weight from die casting of magnesium alloys is sorted into various scrap classes and recycled.

In spite of the relatively high demands for energy and natural resources in the production of light metals, the use of these materials in components for light vehicles has a net positive environmental effect compared with the materials and parts which they displace. A comparison of the environmental impact of magnesium and

competing automotive materials is described in terms of relative energy consumption throughout the lifetime of a passenger car [43].

8. Economic Aspects [1–3], [45]

About 40 % of the primary magnesium produced in the free-market countries is consumed in the form of magnesium alloys used in structural and electrochemical applications. For statistics covering world production and consumption of primary magnesium, see → Magnesium, Chap. 8.

The major end uses for magnesium and magnesium alloys are listed in Table 5. The use of magnesium alloys in aerospace and military applications has declined, but has been compensated by growth in commercial applications such as pressure die castings used in automobiles and computers. Electrochemical applications of magnesium alloys include anodes for cathodic protection, photoengraving plates, and seawater batteries (primary batteries in which seawater is the primary electrolyte, used in sonobuoys and torpedoes).

9. References

1. E. F. Emley: *Principles of Magnesium Technology,* Pergamon Press, London 1966.
2. R. S. Busk: *Magnesium Products Design,* Marcel Dekker Inc., New York 1987.
3. *1998 Annual Book of ASTM Standards* : "Non-ferrous Metal Products," vol. 2.02, ASTM, Philadelphia, PA,1998.
4. *Metals Handbook,*10th ed., vol. 2: "Properties and Selection: Non-Ferrous Alloys and Special-Purpose Materials," American Society for Metals, Metals Park, OH, 1990.
5. *Metals Handbook,*10th ed., vol. 5: "Surface Cleaning, Finishing and Coating," American Society for Metals, Metals Park, OH, 1995.
6. *Metals Handbook,*10th ed., vol. 6: "Welding, Brazing and Soldering," American Society for Metals, Metals Park, OH, 1993.
7. *Metals Handbook,*9th ed., vol. 13: "Corrosion," American Society for Metals, Metals Park, OH, 1987.
8. *Metals Handbook,*9th ed., vol. 14: "Forming and Forging," American Society for Metals, Metals Park, OH, 1988.
9. *Metals Handbook,*9th ed., vol. 15: "Casting," American Society for Metals, Metals Park, OH, 1988.
10. *Worldwide Guide to Equivalent Nonferrous Metals and Alloys,* 3rd ed., American Society for Metals International, Metals Park, OH, 1996.
11. S. C. Erickson: "A Process for the Injection Molding of Thixotropic Magnesium Alloy Parts," *Proc. 44th Annu. Magnesium Conf.* 1987, p. 39. G. A. Chadwick: "Squeeze Casting of Magnesium Alloys and Magnesium Based Metal Matrix Composites," *Proc. of London Conf.,* The Institute of Metals 1986, p. 75.
12. O. M. Hustoft, E. E. Estergaard: *Gravity Metering for Magnesium Cold Chamber Die Casting,* SDCE Int. Congress, Toronto 1987.
13. H. Westengen, O. Holta: "Low Pressure Permanent Mold Casting of Magnesium – Recent Developments," *SAE Tech. Pap. 880 509,* Detroit 1988.
14. W. Unsworth: "A New High Strength Wrought Magnesium Alloy," *Proc. 42nd Annual World Magnesium Conf.* 1982.
15. Dow Chem. Comp., *Magnesium Mill Products and Alloys,* Midland, MI, 1983.
16. J. E. Hillis, R. W. Murrey: "Finishing Alternatives for High Purity Magnesium Alloys," *SDCE Tech. Pap.* G-T87-003, Toronto 1987.
17. F. U. Scholten, M. H. Weidemann: "A New Wear and Corrosion Resistant Anodic Coating for Magnesium," *IMA World Materials Congress,* Chicago 1988.
18. A. L. Olsen: "Plating of Magnesium High Pressure Die Castings," *Trans. IMF* **58** (1980) 29.
19. D. Hawke, J. Davis, R. Fekete: "Field Corrosion Performance of Magnesium Powertrain Components in Light Truck," *SAE Tech. Pap. 890 206,* Detroit 1989.
20. J. D. Hanawalt, C. E. Nelson, J. A. Peloubet: "Corrosion Studies of Magnesium and its Alloys," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **147** (1942) 273 – 299.
21. J. E. Hillis: "The Effects of Heavy Metal Contamination on Magnesium Corrosion Performance," *SAE Tech. Pap. 830 523,* 1983.
22. K. N. Reichek, K. J. Clark, J. E. Hillis: "Controlling the Salt Water Corrosion Performance of Magnesium AS41 Alloy," *SAE Tech. Pap. 850 417,* 1985.
23. O. Lunder, J. E. Lein, T. Kr. Aune, K. Nisancioglu: " The Role of Mg17Al12-phase in the Corrosion of Mg Alloy AZ91," Paper

49: "*Corrosion '88,*" National Association of Corrosion Engineers, St. Louis 1988.
24. T. Kr. Aune, "Minimizing Base Metal Corrosion on Magnesium Products. The Effect of Element Distribution (Structure) on Corrosion Behavior," *Proc. World Magnesium Conf.,* 1983, p. 52.
25. T. Kr. Aune, O. Lunder, K. Nisancioglu: "In situ Microscopic Investigation of Corrosion on a Magnesium Alloy," *IMS Conf.,* Toronto, July 1988.
26. O. Lunder, T. Kr. Aune, K. Nisancioglu: "Early Stages of Corrosion of Magnesium Alloy AZ91 in Chloride Solutions," Abstract, National Association of Corrosion Engineers, 1989.
27. D. L. Hawke, J. E. Hillis, W. Unsworth: *Preventive Practice for Controlling the Galvanic Corrosion of Magnesium Alloys,* International Magnesium Association, McLean, VA, 1988.
28. *ASM Speciality Handbook, Magnesium and Magnesium Alloys,* ASM International, North America 1999.
29. B. L. Mordike, K. U. Kainer, "Magnesium Alloys and their Applications", *Proceedings from the Conference in Wolfsburg*, Germany, 1998.
30. K. U. Kainer, *Magnesium – Eigenschaften, Anwendungen, Potentiale,* Wiley-VCH, Weinheim 2000.
31. K. U. Kainer, "Magnesium Alloys and their Applications", *Proceedings from the Conference in Munich 2000,* Wiley-VCH, Weinheim 2000.
32. *Magnesium Taschenbuch,* Aluminium-Verlag, Düsseldorf 2000.
33. ISO 16220, Magnesium and Magnesium Alloys – Magnesium Alloy Ingots and Castings, 2000.
34. EN 1753:1997, Magnesium and Magnesium Alloys – Magnesium Alloy Ingots and Castings, 1997.
35. J. B. Rønhaug, N. Fantetti, "Gas Displacement Pump Metering System for Magnesium Cold Chamber Die Casting", *1996 CIM Conference in Montreal*, Canada.
36. B. Kittilsen, "Magnesium Alloys, Wrought", *The Encyclopedia of Advanced Materials,* Pergamon, Oxford 1994, pp. 1406 – 1414.
37. D. Hawke, D. L. Albright, "A Phosphate-Permanganate Conversion Coating for Magnesium", *Metal Finishing* (1995) no. 10, 34 – 38.
38. H. S. Tathgar, P. Bakke, Th. A. Engh, "Impurities in Magnesium and Magnesium Based Alloys and their Removal", *Int. Conf. on Magnesium Alloys and their Applications*, Munich, September 2000.
39. J. I. Skar, M. Walter, D. Albright, "Non-Chromate Conversion Coatings for Magnesium Die Castings", *Paper 970324, SAE*, Detroit 1997.
40. B. Reinhold, S. G. Klose, Ch. Strobl, "Die Silikatische Versieglung elektrolytisch verzinkter Oberflächen – ein Schutzsystem für Bauteile in Kontakt mit Magnesium", *Materials and Corrosion,* **50** (1999) 517 – 522.
41. J. I. Skar, "Corrosion and Corrosion Prevention of Magnesium Alloys", *Materials and Corrosion* **50** (1999) 2 – 6.
42. Strategi för kretsloppsanpassade material och varor, Kretsloppsdelegationens rapport 1997:14 Sweden
43. B. Kiefer, G. Deinzer, J. Ø. Haagensen, K. Saur, "Life Cycle Engineering Study of Automotive Structural Parts made of Steel and Magnesium", *Total Life Cycle Air, Sea and Land Mobility*, Graz, Dec. 1 – 3, 1998.
44. T. Mellerud, J. Ø. Haagensen, "Health Environment and Safety Challenges in Magnesium Production Seen in a Life Cycle Perspective", IMA Conference in Rome, June 1999.
45. R. L. Edgar, "Global Overview on Demand and Applications for Magnesium Alloys", *Int. Conf. on Magnesium Alloys and their Applications*, Munich, September 2000.
46. International Magnesium Association, Technical Committee Report, 1998, "Recommended Practices for Conservation of SF_6 in Magnesium".
47. H. Gjestland, W. Schubert, "Use of SO_2 as Protection gas in Magnesium Diecasting", *Int. Conf. on Magnesium Alloys and their Applications*, Munich, September 2000.
48. G. Song, A. Atrens, M. Dargusch, "Influence of Microstructure on the Corrosion of Diecast AZ91D", *Corrosion Science* **41** (1999) 249 – 273.
49. J. Senf, E. Broszeit, M. Gugau, C. Berger, "Corrosion and Galvanic Corrosion of Die Casted Magnesium Alloys", *Proceedings from the Magnesium Technology 2000 Symposium*, TMS, March 2000, pp. 137 – 142.
50. W. C. Mack (ed.), *Worldwide Guide to Equivalent Nonferrous Metals and Alloys, ASM Materials Data Series,* 3rd Edition, ASM International, North America 1996.

Magnesium Compounds

Margarete Seeger, Didier Werke AG, Wiesbaden, Federal Republic of Germany (Chaps. 1, 2, and Sections 4.1–4.4)

Walter Otto, Kali und Salz AG, Kassel, Federal Republic of Germany (Chaps. 3 and 5)

Wilhelm Flick, Hüls AG, Niederkassel, Federal Republic of Germany (Section 4.5)

Friedrich Bickelhaupt, Vrije Universiteit, Amsterdam, The Netherlands (Chap. 6)

Otto S. Akkerman, Vrije Universiteit, Amsterdam, The Netherlands (Chap. 6)

1. Introduction . . . 2
2. Magnesium Carbonate . . . 2
3. Magnesium Chloride . . . 3
3.1. Properties . . . 3
3.2. Raw Materials . . . 4
3.3. Production . . . 4
3.3.1. Precipitation of Magnesium Hydroxide from Seawater . . . 4
3.3.2. Recovery from Spent Liquors from the Potash Industry . . . 4
3.3.3. Evaporation of Seawater and Natural Brines . . . 6
3.3.4. Solution Mining . . . 7
3.3.5. Other Methods . . . 7
3.3.6. Production of Magnesium Chloride Hydrates . . . 8
3.3.7. Production of Anhydrous Magnesium Chloride . . . 8
3.3.8. Purification of Magnesium Chloride Solutions . . . 9
3.3.9. Materials of Construction for the Evaporation of Magnesium Chloride Solutions . . . 9
3.4. Environmental and Legal Aspects . . . 9
3.5. Quality Specifications and Analysis . . . 9
3.6. Transportation and Storage . . . 10
3.7. Uses . . . 10
3.8. Economic Aspects . . . 11
3.9. Toxicology and Occupational Health . . . 11
4. Magnesium Oxide and Hydroxide . . . 11
4.1. Properties . . . 12
4.2. Production of Magnesium Oxide . . . 12
4.2.1. Production from Magnesite . . . 12
4.2.2. Production from Seawater . . . 14
4.2.3. Production from Brines . . . 16
4.2.4. Miscellaneous Processes . . . 18
4.2.5. Burning of Caustic and Sintered Magnesia . . . 19
4.3. Uses . . . 21
4.3.1. Use of Sintered Magnesia as a Refractory . . . 21
4.3.2. Use of Sintered Magnesia as a Heat-Storage Medium . . . 22
4.3.3. Uses of Caustic Magnesia . . . 22
4.3.4. Uses of Magnesium Hydroxide . . . 23
4.4. Economic Aspects . . . 23
4.5. Fused Magnesium Oxide . . . 24
4.5.1. Properties . . . 24
4.5.2. Production . . . 24
4.5.3. Uses . . . 26
4.5.4. Economic Aspects . . . 27
5. Magnesium Sulfate . . . 27
5.1. Properties . . . 27
5.2. Occurrence and Raw Materials . . . 29
5.3. Production . . . 29
5.3.1. Production of Kieserite ($MgSO_4 \cdot H_2O$) . . . 29
5.3.2. Production of Epsom Salt . . . 30
5.3.3. Production of Anhydrous Magnesium Sulfate . . . 31
5.4. Quality Specifications and Analysis . . . 31
5.5. Transportation and Storage . . . 31
5.6. Uses . . . 31
5.7. Economic Aspects . . . 32
5.8. Toxicology and Occupational Health . . . 32
6. Organomagnesium Compounds . . . 32
6.1. Properties . . . 32
6.2. Production . . . 34
6.2.1. Grignard Reagents . . . 34
6.2.2. Diorganylmagnesium Compounds . . . 35
6.3. Uses . . . 35
7. References . . . 36

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a15_595

1. Introduction

Magnesium occurs in divalent form in all of its compounds. The bromide, iodide, sulfate, and nitrate salts are water soluble. Some of the water-soluble salts are highly hygroscopic and form crystals with a high water content. Magnesium fluoride, oxide, hydroxide, phosphate, and carbonate are sparingly soluble or insoluble in water.

The water-soluble magnesium salts tend to undergo hydrolysis and form basic salts due to the weakly basic character of magnesium hydroxide. This is responsible for the highly corrosive action of magnesium salt solutions; dehydration of magnesium salts containing water of crystallization without hydrolytic decomposition is also difficult to achieve.

The most important magnesium compounds are magnesium oxide, which is used for refractory magnesia bricks and as a heat-storage medium, and magnesium chloride which is electrolyzed to produce magnesium metal. Naturally occurring magnesium carbonate (magnesite) is mainly burned to give sintered magnesia for the production of magnesia bricks. Seawater is an important source of magnesium oxide and magnesium chloride. Calcined dolomite ($CaCO_3 \cdot MgCO_3$) is used as a raw material for the silicothermic production of magnesium metal (→ Magnesium). Basic magnesium carbonate is used for thermal insulation.

Anhydrous magnesium nitrate [*10377-60-3*] is difficult to produce (*mp* of the hexahydrate [*13446-18-9*] 89 °C). Concentrated magnesium nitrate solutions (e.g., from $MgCO_3$ and 60 % nitric acid) are used for the production of highly concentrated nitric acid (→ Nitric Acid, Nitrous Acid, and Nitrogen Oxides).

Naturally occurring magnesium silicates include asbestos (→ Asbestos) and talc (→ Talc). Materials for electrical insulation are also based on magnesium silicates (→ Insulation, Electric, Chap. 3.1.2.2., → Insulation, Electric, Chap. 3.1.2.3.). A series of refractory ceramics are based on magnesium (→ Refractory Ceramics).

Organomagnesium compounds are also known, the most important of these being the Grignard reagents which play an important role in synthetic organic chemistry.

Other magnesium compounds are dealt with elsewhere in the Encyclopedia: magnesium cyanide, → Cyanide Compounds, Inorganic, Chap. 2.2.; magnesium fluoride, → Fluorine Compounds, Inorganic, Chap. 4.4.; magnesium hydride, → Hydrides, Chap. 2.6.; magnesium ammonium phosphate, → Fertilizers, Chap. 4.4.4.

2. Magnesium Carbonate

Properties.

Anhydrous magnesium carbonate [*13717-00-5*], $MgCO_3$, M_r 84.33, occurs naturally as *magnesite* which is the starting material for the production of magnesia, MgO (see Section 4.2). Magnesium carbonate is also synthesized by reacting $Mg(OH)_2$ with carbon dioxide at high pressure [1]. Magnesium carbonate begins to decompose at ca. 400 °C, the reaction proceeds rapidly above 550 °C. The compound forms colorless or white to gray, trigonal crystals: lattice constant $a = 0.585$ nm, $\alpha = 103°20'$, $Z = 4$, cleavage parallel to $(10\bar{1}1)$.

Density	2.96 g/cm^3
Mohs hardness	4 – 4.5
Refractive index	n_0 1.700, n_ε 1.509
Enthalpy of formation (298 K)	− 1096 kJ/mol
Dissociation energy (298 K)	+ 1190 kJ/kg
Solubility in cold water	0.0106 wt %

Barringtonite [*5145-48-2*], $MgCO_3 \cdot 2\,H_2O$, M_r 120.35, forms colorless triclinic crystals, lattice constants $a = 0.9115$ nm, $b = 0.6202$ nm, $c = 0.6092$ nm; $\alpha = 94.00°$, $\beta = 95.53°$, $\gamma = 108.87°$; $Z = 4$; refractive index 1.458, 1.473, 1.501; density 2.83 g/cm^3.

The mineral *nesquehonite* [*14457-83-1*], $MgCO_3 \cdot 3\,H_2O$, M_r 138.37, occurs as colorless to white crystals. This carbonate can also be synthesized by carbonation of $Mg(OH)_2$ at atmospheric pressure. Density 1.84 g/cm^3, *mp* 165 °C (decomp.), solubility in cold water 0.129 wt %.

Lansfordite [*61042-72-6*], $MgCO_3 \cdot 5\,H_2O$, M_r 174.40, forms white monoclinic crystals, density 1.73 g/cm^3.

The natural mineral *hydromagnesite* [*12072-90-1*], $4\,MgCO_3 \cdot Mg(OH)_2 \cdot 4\,H_2O$, M_r 467.67, forms white monoclinic crystals,

density 2.25 g/cm^3 [2]. The synthetic product is known as magnesia alba.

Other $MgCO_3$-containing minerals include *huntite* [*19569-21-2*], $CaCO_3 \cdot 3\ MgCO_3$, M_r 353.05 [2], and *dolomite* [*16389-88-1*] $CaCO_3 \cdot MgCO_3$, M_r 184.41. Dolomite is a rock-forming mineral and is much more common than magnesite. It is burned to form sintered dolomite (for refractory bricks) and light-burned dolomite (used as a fluxing agent).

Uses. The main use of magnesite is the production of magnesia (Section 4.2). Small quantities are used as a filler. Plans have been proposed for producing magnesium metal from naturally occurring (cheap) magnesite [3]. The synthetic carbonates are used as fillers for paper, plastics, and rubber and as thickening agents for printing inks. Pharmaceutical-grade magnesium carbonate (magnesia alba) is also used in toothpaste and as an antacid. High-purity magnesium carbonate is employed in cosmetics.

3. Magnesium Chloride

Magnesium chloride [*7786-30-3*], $MgCl_2$, is the most common naturally occurring, water-soluble magnesium compound. It is a constituent of many salt lakes and natural brines [4], [5]; sea salt contains 17 wt % $MgCl_2$. Mineral salt deposits contain the important minerals carnallite [*1318-27-0*] $KCl \cdot MgCl_2 \cdot 6\,H_2O$, bischofite [*7791-18-6*] $MgCl_2 \cdot 6\,H_2O$, and occasionally the double salt tachhydrite in which the magnesium chloride is combined with calcium chloride ($CaCl_2 \cdot 2\,MgCl_2 \cdot 12\,H_2O$). Tachhydrite is found in salt deposits in Brazil and the Congo region.

Magnesium chloride is obtained from mineral salt deposits by processing the spent liquors remaining after extraction of potassium chloride and also by direct solution mining; it is also recovered from salt lakes and seawater.

3.1. Properties

Anhydrous Magnesium Chloride. Anhydrous magnesium chloride forms hexagonal crystals, its physical properties are as follows:

M_r	95.23	
mp	713 °C	
Refractive index (*n*)	1.675, 1.59	
d_4^{20}	2.41 g/cm^3	
Specific heat (c_p) between 0 and 713 °C (*mp*):		
$c_p = 0.78011 + 0.20607 \times 10^{-3} \times (t-20) - 0.10853 \times 10^{-6} \times (t-20)^2$ kJ kg^{-1} K^{-1} (t in °C)		
Specific electrical conductivity at		
700 °C	1.00 Ω^{-1} cm^{-1}	
750 °C	1.09 Ω^{-1} cm^{-1}	
800 °C	1.17 Ω^{-1} cm^{-1}	
Solubility in water at		
20 °C	35.3 wt %	54.6 g/100 g
40 °C	36.5 wt %	57.5 g/100 g
60 °C	37.9 wt %	61.0 g/100 g
80 °C	39.8 wt %	66.1 g/100 g
100 °C	42.2 wt %	73.0 g/100 g
Solubility in methanol at		
0 °C	15.5 g/100 g	
60 °C	20.4 g/100 g	
Solubility in ethanol at		
0 °C	3.61 g/100 g	
60 °C	15.89 g/100 g	

Properties of the Hydrates. Magnesium chloride forms five hydrates whose transformation temperatures and solubilities are given in Table 1 (see also Fig. 1). The most important is the hexahydrate, $MgCl_2 \cdot 6\,H_2O$, which forms monoclinic crystals and is present in mineral salt deposits:

M_r	203.31
mp	117 °C
Heat of fusion	34.3 kJ/mol
Density	1.57 g/cm^3
Refractive index (*n*)	1.495, 1.507, 1.528
Specific heat (c_p) at 48.6 °C	1.415 kJ kg^{-1} K^{-1},
100 °C	1.537 kJ kg^{-1} K^{-1}

Table 1. Transformation temperatures and corresponding solubilities of magnesium chloride hydrates

Hydrate transformation	Temperature, °C	Solubility, g $MgCl_2$/100 g H_2O
$MgCl_2$ solution→$MgCl_2 \cdot 12\,H_2O$	− 33.6	27.2
$MgCl_2 \cdot 12\,H_2O$→$MgCl_2 \cdot 8\,H_2O$	− 17.4	47.6
$MgCl_2 \cdot 8\,H_2O$→$MgCl_2 \cdot 6\,H_2O$	− 3.4	52.3
$MgCl_2 \cdot 6\,H_2O$→$MgCl_2 \cdot 4\,H_2O$	116.8	85.6
$MgCl_2 \cdot 4\,H_2O$→$MgCl_2 \cdot 2\,H_2O$	181.5	125.7

The penta- and hexahydrates can be dehydrated to the tetrahydrate stage almost without decomposition. On dehydration to the dihydrate,

however, hydrolytic decomposition starts. Dehydration of the dihydrate at 240 °C leads to extensive decomposition forming magnesium oxychlorides or oxide. To obtain anhydrous $MgCl_2$, the dihydrate must be dehydrated in an atmosphere of hydrogen chloride.

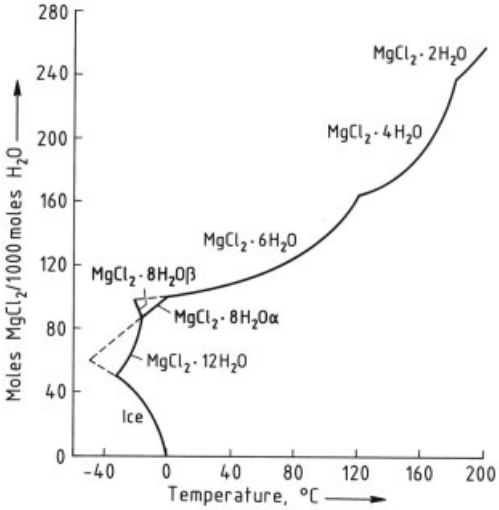

Figure 1. State diagram of the system $MgCl_2 \cdot n\,H_2O$

Chemical Properties. Magnesium chloride can be decomposed by oxygen to form magnesium oxide and chlorine [6] – the converse of the reaction between magnesium oxide and chlorine to form $MgCl_2$. With ammonia, $MgCl_2$ forms ammines that are analogous to the hydrates: $MgCl_2 \cdot NH_3$, $MgCl_2 \cdot 2\,NH_3$, and $MgCl_2 \cdot 6\,NH_3$. Corresponding addition compounds are formed with alcohols and with other organic compounds such as ethers, acetic acid, amines, ethyl acetate, pyridine, hexamethylenetetramine, and dioxane (e.g., $MgCl_2 \cdot 6\,CH_3COOH$).

A series of basic chlorides (magnesium oxychlorides) are formed with magnesium oxide and are of commercial importance as Sorel cement (see 3.8). The following oxychlorides are known: $MgCl_2 \cdot MgO$, $MgCl_2 \cdot MgO \cdot H_2O$, $2\,MgCl_2 \cdot MgO \cdot H_2O$, $MgCl_2 \cdot 3\,MgO \cdot 11\,H_2O$, and $MgCl_2 \cdot 3\,MgO \cdot 7\,H_2O$. The compounds $MgCl_2 \cdot MgO$ and $2\,MgCl_2 \cdot MgO \cdot H_2O$ are formed on dehydration of $MgCl_2 \cdot H_2O$ at 280–495 °C. Aqueous $MgCl_2$ solutions are weakly acidic due to hydrolysis. The electrolysis of aqueous $MgCl_2$ solutions produces magnesium hydroxide and chlorine, whereas electrolysis of anhydrous $MgCl_2$ yields magnesium metal and chlorine.

3.2. Raw Materials

Of the three main sources of raw materials for magnesium chloride production, the most important is *seawater*. Magnesium can be directly precipitated from seawater as magnesium hydroxide and converted to magnesium chloride by treatment with chlorine. The residues (bitterns) obtained after production of drinking water or sea salt from seawater can also be used. A second large source is the *spent liquor from the potash industry* arising from the processing of hard salts and carnallite-bearing raw materials. A third source is provided by the *salt lakes* in many parts of the world, the Dead Sea and the Great Salt Lake being particularly important.

3.3. Production

3.3.1. Precipitation of Magnesium Hydroxide from Seawater

If seawater cannot be preconcentrated by solar energy, magnesium chloride salts or solutions can only be produced on a large scale by precipitation as the hydroxide, followed by conversion to the chloride with hydrogen chloride.

The Dow Chemical process is described elsewhere (→ Magnesium). This process has been used to manufacture magnesium chloride from seawater in large quantities since 1941 in Freeport (Texas). It is an intermediate step in the production of magnesium metal.

3.3.2. Recovery from Spent Liquors from the Potash Industry

The spent liquors obtained after recovery of potassium chloride from carnallite-containing raw salts (→ Potassium Compounds) typically contain (in grams per liter): $MgCl_2$ 293, $MgSO_4$ 48, KCl 39, NaCl 26, and H_2O 890. They are used as a raw material for the recovery of magnesium chloride. For the Norsk Hydro process, see → Magnesium.

Production of Concentrated $MgCl_2$ Solutions by Kali und Salz AG (Germany). The spent liquor contains ca. 280 – 300 g/L $MgCl_2$ and is concentrated in a two-stage vacuum evaporator to ca. 460 g/L $MgCl_2$. (Vapor pressures of $MgCl_2$ solutions as a function of concentration and temperature are shown in Fig. 2).

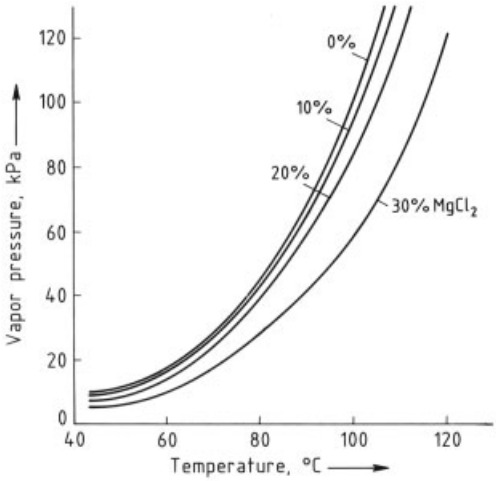

Figure 2. Vapor pressure of $MgCl_2$ solutions as a function of temperature

The solution in the first evaporator stage is heated by the vapor from the second stage to a steady temperature of ca. 55 °C. The second-stage liquor is maintained at 95 °C by heating with steam at 0.1 MPa and the hot, cloudy solution is clarified by removing the salts [carnallite, NaCl, $MgSO_4 \cdot 5/4\ H_2O$, and impurities such as $Fe(OH)_3$ and clay] by means of a rotary filter and centrifuge. On cooling carnallite crystallizes from the solution. The crystallized salts and sludge are sent to the potassium chloride plant where carnallite is decomposed together with carnallite containing crude potash salt. The carnallite is mixed with water or a low-$MgCl_2$ brine to produce a spent liquor containing 280 – 300 g/L of $MgCl_2$ and solid potassium chloride contaminated with sodium chloride and $MgSO_4 \cdot 5/4\ H_2O$ (→ Potassium Compounds). The cooled $MgCl_2$ solution can be purified by reheating and removing bromine with chlorine. In this case the debrominated solution is neutralized with lime, partly precipitating the sulfate ions in the form of gypsum. The precipitated sediment is removed. The highly concentrated 32 – 33 % $MgCl_2$ solution is sold as such (see Section 3.5) or can be processed further to obtain magnesium chloride hydrates (see Section 3.3.6).

Production of Concentrated $MgCl_2$ Solution by the VEB Kombinat Kali (former German Democratic Republic). The spent liquor from the manufacture of potassium salts had a $MgCl_2$ content of ca. 260 g/L. The solution was first mixed with synthetic carnallite from the vacuum cooling stage; the $MgCl_2$ from the carnallite dissolves leaving the undissolved potassium chloride behind. The increased $MgCl_2$ content (ca. 320 g/L) of the solution lowers the solubility of the alkali-metal salts causing the precipitation of potassium and sodium chlorides. After removal of these salts, the solution was pumped to the second of the two evaporation stages. The preconcentrated solution from the second evaporation stage was pumped to the first evaporation stage where $MgSO_4 \cdot 5/4\ H_2O$ precipitates together with some langbeinite, carnallite, and sodium chloride. On cooling this solution, (after removal of the salts) in multistage vacuum coolers, synthetic carnallite precipitates. The carnallite was filtered off and recycled to the first step of the process. All the evaporators were of the forced circulation type [7]. The first evaporator stage was operated at 120 °C and heated with steam. The second stage was heated to 80 °C by the vapor from the first stage. The vapors from the second stage were cooled with water.

In a variation of the process, the solution containing 320 g/L $MgCl_2$ was mixed with part of the product solution (containing 460 g/L $MgCl_2$). The increase in the $MgCl_2$ content causes more sodium chloride and carnallite to precipitate. Evaporation of the solution from the first stages then does not precipitate carnallite or langbeinite ($K_2SO_4 \cdot 2\ MgSO_4$). Only synthetic kieserite ($MgSO_4 \cdot 5/4\ H_2O$) and sodium chloride are precipitated and these are removed. The solution produced had a density of 1.340 g/cm^3 and contained 460 g/L $MgCl_2$, 25 g/L $MgSO_4$, 2 g/L KCl, 6 g/L NaCl, 2 g/L $CaSO_4$, and 4 g/L Br^-. If these solutions are used for the production of magnesium metal, they must be debrominated with chlorine (→ Bromine, Chap. 4.). If they are used for production of high-quality sintered magnesia, boron is also removed by boron-selective resins. The boron content decreases from 30 to 5 –10 mg/L [8].

Recovery from Raw Salts Containing Carnallite. Patents have been applied for in which pure carnallite or raw salts containing carnallite are decomposed with C_1–C_4 alcohols, causing the $MgCl_2$ to pass into the alcohol phase [9]. The alcoholic solution is concentrated by evaporation to an alcohol content of 10 %, and the hexahydrate is precipitated by addition of water. Alternatively, evaporation is continued until the solution contains 1 – 2 mol of alcohol per mole of $MgCl_2$; the alcoholate is then thermally decomposed in a hydrogen chloride atmosphere at 400 – 500 °C.

3.3.3. Evaporation of Seawater and Natural Brines

The direct production of $MgCl_2$ from seawater or brines by evaporation is only economical where the dilute solutions can be preconcentrated by solar evaporation (→ Sodium Chloride).

Recovery from the Dead Sea. Magnesium chloride-containing solutions are produced in large quantities by solar evaporation of brine from the Dead Sea. At the Dead Sea Works, carnallite precipitates in the final stage, and solutions containing 360 g/L $MgCl_2$, 110 g/L $CaCl_2$, 7 g/L NaCl, and 5 g/L KCl are obtained. Owing to the low vapor pressure of water, solar evaporation only occurs at very low atmospheric humidity levels. After two or three months of evaporation, a density of 1365 g/L is achieved and bischofite crystallizes [10]. Large quantities of $MgCl_2$-containing liquors of this type are now used by Dead Sea Periclase solely for the production of magnesium oxide by thermal decomposition.

In a process patented by the Dead Sea Works for the manufacture of magnesium chloride hexahydrate (bischofite), carnallite or a natural mixture of carnallite and sodium chloride is decomposed in its own water of crystallization in closed vessels under pressure at 167.5 °C. The potassium chloride remains together with the sodium chloride as a sediment and is removed. The residual solution contains up to 8.2 % potassium chloride and 42.3 % magnesium chloride. On cooling to 115 °C by flash evaporation, carnallite is precipitated and is fed back into the circulating system. The residual solution contains 1 % potassium chloride and consists essentially of bischofite [11].

Recovery from the Kara–Bogaz Gulf (Turkmenia). Following the isolation of the Kara-Bogaz Gulf from the rest of the Caspian Sea by a dam in 1979, the $MgCl_2$ content increased to 250 g/L. Research was carried out to find a new method of processing this brine [12]. The brines evaporated in summer contain 25 – 30 wt % $MgCl_2$. When they are cooled in winter, epsomite ($MgSO_4 \cdot 7\,H_2O$) precipitates and is used for producing schönite ($K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O$). The liquid phase is used for $MgCl_2$ production.

Recovery from Spent Liquors after Extraction of NaCl from Seawater. The spent liquors (sea bitterns) obtained after extraction of salt from seawater are evaporated in crystallizing pans by solar energy to a density of 1290 g/L. After removing the sodium chloride, the solution is cooled to 5 °C, which can be achieved in winter without energy consumption. Magnesium sulfate (epsomite) crystallizes with a few impurities and is removed. The bitterns are cooled to 0 °C; sodium sulfate (Glauber's salt) crystallizes and is removed. The solution is then debrominated with chlorine (→ Bromine, Chap. 4.) to exclude crystallization of magnesium bromide in subsequent processing. After neutralizing with potassium hydroxide, the liquor is heated to 125 °C. On cooling to 100 °C, a crude potassium-containing salt mixture separates that consists of langbeinite ($K_2SO_4 \cdot 2\,MgSO_4$) and kainite ($KCl \cdot MgSO_4 \cdot 3\,H_2O$), with some kieserite ($MgSO_4 \cdot H_2O$) and sodium chloride. On further cooling to 20 °C, carnallite crystallizes and can be treated with water to form a $MgCl_2$-containing solution and solid potassium chloride [13].

Recovery from the Great Salt Lake. For the National Lead Industries process, see → Magnesium. In the process operated by the Great Salt Lake Minerals Chemical Corporation, a 35-wt % $MgCl_2$ solution is first produced by solar evaporation, and this is then subjected to vacuum evaporation until bischofite ($MgCl_2 \cdot 6\,H_2O$) and hydrated $MgSO_4$ are

precipitated. Proposed further treatment consists of heating the crystals to 120 – 150 °C to redissolve the bischofite and allow removal of magnesium sulfate. The bischofite can then be crystallized [14]. Alternatively, the solution can be maintained at 120 °C until the magnesium sulfate content decreases to < 20 g/L as a result of crystallization of $MgSO_4 \cdot 5/4\,H_2O$. The bischofite is then crystallized by vacuum evaporation at 90 °C [15].

Recovery from Naturally Occurring Brines with High $CaCl_2$ Contents. Up to the end of World War II, the Dow Chemical Company used underground brines from Midland (Michigan) and Marysville for bromine production and also for the production of $MgCl_2$. The latter was electrolyzed to obtain magnesium. The alkali metal salts were removed by evaporation and crystallization. In an older process the calcium was then separated together with the magnesium as tachhydrite ($2\,MgCl_2 \cdot CaCl_2 \cdot 12\,H_2O$) which was converted into $MgCl_2 \cdot 6\,H_2O$. In a later process, the solution was reacted with an amount of magnesium hydroxide (produced from magnesite) equivalent to the calcium chloride content. The calcium was then removed as calcium carbonate by treating the solution with the carbon dioxide from the magnesite calcination, and the $MgCl_2$ was obtained from the solution.

Today, Dow obtains magnesium chloride for magnesium production only from seawater (→ Magnesium). The production of bischofite from Michigan brines has been discontinued.

3.3.4. Solution Mining

A 20-wt % $MgCl_2$ solution is produced in large quantities in the Netherlands from a carnallite salt deposit. It is used to produce concentrated $MgCl_2$ solutions and is converted to magnesium hydroxide (see Section 4.2.3) [16].

Solution Mining of Carnallite. Carnallite ($KCl \cdot MgCl \cdot 6\,H_2O$) deposits contain large amounts of magnesium sulfate as well as $MgCl_2$. Evaporation of solutions with high magnesium sulfate contents leads to precipitation of double salts containing sodium chloride and magnesium sulfate. In the processing of these solutions, most of the magnesium sulfate is precipitated as langbeinite ($K_2SO_4 \cdot 2\,MgSO_4$) [17], or as schönite ($K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O$) by precipitation with methanol [18]. In a process developed by the Mines de Potasse d'Alsace, a solution containing 320 g/L $MgCl_2$ can be obtained together with solid potassium chloride by solution mining. The solution used to extract the carnallite from the deposit contains a flotation agent for potassium chloride ($C_{12} - C_{20}$ fatty amines or their salts) and a gas-producing agent such as hydrogen peroxide. The potassium chloride floats and can be pumped out along with the $MgCl_2$ solution [19]. This production method appears promising, but it is not known whether large-scale development is planned.

3.3.5. Other Methods

Magnesium chloride can be produced from dolomite in combination with soda by the Solvay process. In the early part of World War II at a plant belonging to the Diamond Alkali Corporation in Painesville, Ohio, the dolomite was calcined in rotary furnaces, hydrated, and used to regenerate ammonia from the ammonium chloride solution produced in soda manufacture. Only the calcium hydroxide took part in the reaction. On treatment of the remaining solution with carbon dioxide, a magnesium chloride solution was obtained:

$$Mg(OH)_2 + CaCl_2 + CO_2 \rightarrow MgCl_2 + CaCO_3 + H_2O$$

This process is, however, no longer used.

In the Mathison Alkali Works, the calcium chloride solution from the soda process was treated with calcined dolomite while carbon dioxide was introduced. This process also produced a $MgCl_2$ solution but has been discontinued.

Calcined magnesite or magnesium hydroxide can also be used in the soda process to produce a $MgCl_2$ solution.

Processes are described in which magnesium chloride is produced from natural crude magnesite by leaching with hydrochloric acid, as well as the use of phosgene and crude hydrochloric acid with 27.5 % HCl and 1.2 % HF to treat magnesite and periclase [20–22].

3.3.6. Production of Magnesium Chloride Hydrates

Production of $MgCl_2 \cdot 6\,H_2O$ (Bischofite). In a process formerly used by Kali und Salz AG, end bittern from potash production was treated in a single-stage vacuum evaporator to produce a ca. 42-wt % $MgCl_2$ solution, and was further evaporated in pans with steam at 1.1 MPa. After neutralization with lime and removal of calcium sulfate, synthetic kieserite, and other impurities, the melt was fed onto rotating drum coolers or cooling bands to form a slab, which was broken up and marketed (see Section 3.5). Alternatively, after the initial single-stage evaporation, the solution can be prilled in a prilling tower or spray dried in a fluidized bed [23].

A very pure $MgCl_2$ hydrate with 4 – 6 molecules of water can be prepared in 99.9 % purity according to a patent of the VEB Kombinat Kali [24]. Carnallite is dehydrated at 180 – 200 °C and then cooled. The cooled salt is then treated with acetone, in which the $MgCl_2$ dissolves. After removal of the undissolved salt, the hydrate is precipitated from the acetone solution by adding the correct amount of water [24].

In a patent of the Mines de Potasse d'Alsace, $MgCl_2$ is precipitated from solution by addition of 10 – 25 % 1,4-dioxane at room temperature. At lower temperatures the yield of $MgCl_2$ is increased. The $MgCl_2$– dioxane complex separates as a voluminous precipitate that is easy to filter or centrifuge. It can be purified by washing with a little dioxane. The dioxane can be distilled off from the filtrate as an 80 % azeotrope and can be reused directly. The $MgCl_2$ hydrate can be obtained by washing with water. During drying the remaining dioxane (ca. 20 – 25 % of the total) is evaporated off and can also be reused [25].

Production of $MgCl_2 \cdot (2-3)H_2O$. In a process formerly used by Kali und Salz AG, a ca. 33-wt % $MgCl_2$ solution was heated to 110 – 130 °C and sprayed into a fluidized bed of magnesium chloride granules at 130 °C. The heating gas was at 400 – 600 °C. The average residence time of the granules is ca. 30 min. Fines and pulverized oversize granules were recirculated to the fluidized bed. The final product had a water content of ca. 30 % [23].

3.3.7. Production of Anhydrous Magnesium Chloride (See also → Magnesium, Chap. 4.1.1.)

The Nalco Chemical Company at its plant in Freeport, Texas, produces tetramethyl- and tetraethyllead. The Grignard reagents needed as intermediates for the process require a very pure source of magnesium. This is produced by the electrolysis of pure, anhydrous magnesium chloride, for which the following process has been developed.

Ethylene glycol is added with stirring to a 30 – 35-wt % $MgCl_2$ solution, the quantity added being sufficient to produce a 10 – 20-wt % $MgCl_2$ solution. All the water, including water of crystallization, is then removed by vacuum distillation. The distillate contains < 50 ppm ethylene glycol, and the residual solution contains < 25 ppm water together with all the other salts. Gaseous ammonia is then bubbled through the solution under slight pressure. The mixture is cooled to remove the evolved heat; in practice the temperature is allowed to rise to ca. 70 °C. When the glycol is saturated with ammonia, the hexaammine compound $MgCl_2 \cdot 6\,NH_3$ is formed. Towards the end of the reaction, the mixture is cooled to 15 – 30 °C, the $MgCl_2 \cdot 6\,NH_3$ precipitates and is filtered off. Any adhering solution is removed by washing with methanol. The ammonia and methanol are then removed in two stages: the methanol and some ammonia are first driven off at 74 °C. The rest of the ammonia is then removed in a calcination furnace at 400 °C [26].

As with the dehydration of carnallite, ammonium carnallite can be dehydrated almost without decomposition at 90 – 180 °C. It can then be decomposed at 180 – 400 °C to give ammonium chloride and anhydrous $MgCl_2$ [27], [28].

In the process developed by Esso Research, a 20 – 50-wt % solution of magnesium chloride reacts with ammonium chloride in a three-stage fluidized bed. Anhydrous magnesium chloride is produced at 350 – 550 °C [29].

In another process [30], ammonium carnallite containing $n\,NH_3$ or $n\,H_2O$ is decomposed at 800 °C. A product with a low magnesium oxide content is obtained. If the starting material contains water of crystallization, some is removed in a first stage, and further dehydration takes place

after adding ammonia. After the dehydration, the $MgCl_2 \cdot 6\,NH_3$ is thermally decomposed [31].

A process for the dehydration of magnesium chloride hydrates in a hydrogen chloride atmosphere is described in → Magnesium, Chap. 4.1.1.2..

3.3.8. Purification of Magnesium Chloride Solutions

Calcium ions can be removed from magnesium chloride solutions by adding the stoichiometric amount of sulfate ions [32]. Conversely, sulfate ions can be removed by adding calcium chloride solutions. Calcium and other metal ions can be extracted with dialkylphosphoric acid in a water-immiscible solvent. This process is used in the production of pure sintered magnesia from solutions of impure magnesites in hydrochloric acid [33]. These solutions can be treated with substances such as MgO, CaO, or $Mg(OH)_2$ to increase the pH, while simultaneously purging with air or oxygen at 80 – 90 °C. This removes impurities such as iron, aluminum, and manganese as hydrated oxides or hydroxides. This process is also used in the manufacture of pure sintered magnesia [34].

3.3.9. Materials of Construction for the Evaporation of Magnesium Chloride Solutions

At temperatures up to 110 °C, corrosion is usually prevented by the use of coatings, rubber linings, or plastic reinforced with glass fibers. Heater tubes and distillation trays are usually made of aluminum bronze ($AlBz_{5-7}$) as used in the potash industry, with a composition of 4 –8 % Al, 89 – 96 % Cu, and small amounts of Fe, Ni, Mn, As, Si, or Zn. Pumps are made of NiResist D 2.

Above 110 °C, and especially in the presence of chlorine and bromine (during debromination), glass and Teflon are suitable, with titanium being used for preheaters and pipework. According to [35] brick-lined steel vessels are used for evaporating $MgCl_2$ solutions and steam-heated tubes are made of Inconel alloy or stainless steel 347. Monel metal is also used for > 50 % $MgCl_2$ solutions.

3.4. Environmental and Legal Aspects

Magnesium chloride is a constituent of seawater. In the EEC Regulations (May 4th 1976) concerning disposal of hazardous substances in lakes and rivers, this substance is not mentioned. However, if large quantities are discharged into lakes or rivers, the concentration can be high enough to injure aquatic organisms and immediate dilution must be carried out. The WHO international standard for drinking water for magnesium is 30 mg/L Mg, or 125 mg/L if the sulfate concentration is < 250 mg/L and the chloride concentration is < 200 mg/L. In the EEC regulation 80/778 EWG the limit for drinking water is 50 mg/L Mg and 25 mg/L Cl^-. In the dehydration of magnesium chloride hydrates, hydrogen chloride occurs in the waste gases due to partial decomposition to magnesium oxide and the concentration must be kept below the permitted limit by use of gas purification equipment. The German TA Luft Regulations specify a maximum of 30 mg/m^3 HCl in air. Limits in Australia are 200 – 400, in the United Kingdom 150, and in Japan 80 – 700 mg/m^3.

3.5. Quality Specifications and Analysis

Quality Specifications. Magnesium chloride is available commercially in the form of solutions or solid hydrates. Solutions usually have a $MgCl_2$ content of 32 – 33 wt %. A typical solution with a $MgCl_2$ content of 32 – 33 wt % has the following approximate composition (in grams per liter);

$MgCl_2$	430 – 440
$MgSO_4$	12 – 25
KCl	2 – 3
NaCl	4 – 6
H_2O	ca. 870

For many applications, however, solutions produced by decomposition of carnallite (see Section 3.3.2) are used directly; due to transport costs this only applies when the user is close to the supplier.

The most important hydrate sold commercially is the hexahydrate (bischofite) which contains 43 – 47-wt % $MgCl_2$, ca. 51.5 wt % water,

ca. 2 wt % alkali metal chlorides, and, depending on its origin, some $MgSO_4$, $CaSO_4$, or $CaCl_2$. It is supplied in the form of broken pieces or flakes, or as blocks produced by melting and pouring into drums. Another hydrate in granular form contains ca. 2 – 3 mol of water per mole of $MgCl_2$ and corresponds to 67 wt % $MgCl_2$. Anhydrous magnesium chloride is mostly used directly by the producer for electrolytic magnesium production (→ Magnesium).

Analysis. For quality control the $MgCl_2$ solutions and the solid hydrates must be analyzed for the following ions: Mg^{2+}, Ca^{2+}, K^+, Na^+, Cl^-, SO_4^{2-}, and Br^-. In special cases, determination of boron, iron, and other ions may be required.

Magnesium and calcium are titrated complexometrically with ethylenediaminetetraacetic acid (EDTA) [36], [37]. Potassium and sodium are determined by flame photometry [38], or by ICP [39]. For the determination of sulfate ions, a titration with barium perchlorate solution in a 2-propanol–water mixture is carried out with thorin as the indicator [40] or it is analyzed by ICP. Bromide is determined gravimetrically by the method of D'Ans and Höfer [41].

The water content of the hydrates is determined by heating to 450 – 600 °C or by the lead oxide method [42]. Boron can be determined photometrically with azomethine-H or dianthrimide-1-amine [43], [44] and iron with bathophenanthroline (4,7-diphenyl-1,10-phenanthroline) [45]. Iron can also be determined by atomic absorption spectrophotometry [46] or by ICP.

3.6. Transportation and Storage

Magnesium chloride solutions can be transported in tank wagons. Owing to their corrosive properties, they should be stored in plastic containers reinforced with glass fiber, but iron vessels can also be used. When a 32 – 33-wt % $MgCl_2$ solution is cooled, $MgCl_2$ hydrates precipitate, gradually forming a sludge. At – 33.6 °C the residual solution containing 21.4 wt % $MgCl_2$ freezes as a eutectic mixture. The hexahydrate is transported in polyethylene sacks or big bags capable of holding > 50 kg. It is also supplied in the form of broken pieces or solid blocks produced by pouring molten salt in single-journey steel drums. Sealed vessels can be stored almost indefinitely, but, after opening, the material must be used immediately because of its hygroscopicity. All containers must be stored under dry conditions.

3.7. Uses

Anhydrous magnesium chloride is used chiefly for the electrolytic production of magnesium metal (→ Magnesium). Large quantities of magnesium chloride solutions are used to produce anhydrous magnesium chloride for this purpose.

An important area of application for the hexahydrate and the hydrate containing 67 % $MgCl_2$ is in the production of Sorel cement. This is made by mixing magnesium oxide with $MgCl_2$ solutions, and is mainly used as a binder for industrial floorings on account of its high elasticity and bending tensile strength properties [47]. This binder is also used in the manufacture of abrasive materials and grindstones, and for cementing rock faces during coal mining operations [48]. In combination with wood chippings, it can be used to make lightweight building slabs. Sorel cement mixtures are also used in water- and fire-resistant building materials [49]. They also serve as a binder for granulated fertilizers and in the production of salt licks for animals, where the magnesium also acts as a feed additive, preventing diseases such as grass tetany in cattle that are caused by magnesium deficiency [50].

Another application for $MgCl_2$ solutions and the hexahydrate is the manufacture of sintered magnesia and very pure $Mg(OH)_2$ (see Section 4.2.5). A high proportion of the solutions produced in Israel and the Netherlands are used for this purpose. Magnesium chloride hexahydrate and potassium chloride are used to produce artificial carnallite, which is used as a covering flux in the melting of light metal alloys. In the Quentin process used in the sugar industry, the alkali metal ions in the dilute sugar syrup are replaced by magnesium ions by treating it with an anion-exchange resin, thus reducing sugar losses in the molasses by improving sugar crystallization [51]. The resin is regenerated with $MgCl_2$ solution. In the mineral oil

industry, $MgCl_2$ solutions are used as drilling fluids in rocks with a high salt content. The low freezing point of the 21 % $MgCl_2$ solution ($< -30\,°C$) is utilized to prevent wet bulk materials from freezing to a solid mass. It is used as a de-icing agent in winter, as a ballast liquid for the tires of construction machinery and agricultural vehicles, for fire-extinguishing foams at low temperatures, and as brine for refrigeration. It is used in road surfaces combined with minerals and asphalt as a antifreeze. It is also used as a heat storage medium in solar heating installations, as an additive in the steel industry, as a coagulant in wastewater treatment, and in the treatment of combustion gases ($MgCl_2$ solution absorbs HCl and improves the removal of SO_2 by lime). In the plastics industry it is used in the manufacture of polymerization catalysts, e.g., in combination with $TiCl_4$ in Ziegler – Natta catalysts for the polymerization of $C_2 - C_{12}$ olefins. Addition of $MgCl_2$ also reduces losses of ammonia that is used as a fertilizer [52].

$MgCl_2$ is used in the production of moisture-proof salts. It is also used in the pharmaceutical and cosmetic industries, in food- and feedstuffs, and in bath salts. In addition it is used in the production of other Mg salts such as magnesium-chlorate.

3.8. Economic Aspects

The total worldwide production of $MgCl_2$ in 1996 was 1.5×10^6 t. In the world export markets ca. 300 000 t $MgCl_2$ was handled: ca. 200 000 t in the form of 33 % $MgCl_2$ solutions, and 100 000 t in the form of the hexahydrate. About half of the $MgCl_2$ solution was used for production of magnesium metal. In addition, approximately 180 000 t $MgCl_2$ from solution mining brines and 150 000 t $MgCl_2$ as 33 % solution were converted to magnesium oxide; these quantities were reported as material produced but not marketed. A further ca. 700 000 t magnesium chloride was not recorded in any statistics because it was used solely as an intermediate for the production of magnesium by electrolysis.

The most important countries for the production of $MgCl_2$ solutions and the hexahydrate are given below (excluding magnesium metal producers):

Germany (Kali und Salz AG)	byproduct solutions from the potash industry
Israel	Dead Sea brines
Netherlands	solution mining brines
Japan	sea bitterns from salt production
France	sea bitterns
United States	Great Salt Lake brines
Italy	sea bitterns
China	brines from various salt lakes
Australia	from Kunwarara magnesit deposit

Two thirds of the material appearing on the market originates from Germany.

3.9. Toxicology and Occupational Health

The handling of $MgCl_2$ hydrates, anhydrous $MgCl_2$, or $MgCl_2$ solutions on a laboratory scale does not present any personal hazards. German chemical regulations do not require any special marking. The LD_{50} (rat, oral) for $MgCl_2 \cdot 6\,H_2O$ is 8100 mg/kg. Eye protection and rubber gloves should be worn for handling anhydrous $MgCl_2$, $MgCl_2$ hydrates, or solutions. Prolonged contact with concentrated (ca. 35 wt %) solutions or with $MgCl_2$ dust can cause skin irritation, and possibly allergic skin reactions. In the event of skin problems, the affected area should be washed with copious soap and water, followed by treatment with a fat-containing skin cream. In case of contact with the eye, the eye should be washed with water and medical assistance should be obtained.

4. Magnesium Oxide and Hydroxide

Magnesium oxide [*1309-48-4*], MgO (magnesia, periclase), seldom occurs as a natural mineral; it is found in contact metamorphic limestone and dolomite, in volcanic ejecta, and in serpentine rocks. It does not form rocks or salt deposits because it is converted to magnesium hydroxide by the water vapor in the atmosphere. The primary sources of industrially produced magnesia are natural magnesite ($MgCO_3$), seawater, and natural and synthetic brines. Magnesium hydroxide is formed as an intermediate in the production of magnesia from seawater and brines. Magnesia is used in technical applications on account of its high melting point, chem-

ical resistance, high thermal conductivity, low electrical conductivity, and biological activity.

Technical-grade magnesia is mainly sintered to form *sintered magnesia* (also known as magnesia sinter, dead-burned magnesia, or sintered magnesite) which is used primarily as a refractory material and in the steel industry. Large quantities of calcined (decarbonated) material known as *caustic magnesia* or caustic-calcined magnesia, are produced for use in agriculture and the building industry. *Fused magnesia* is normally produced by melting caustic magnesia.

Magnesium hydroxide [*1309-42-8*], $Mg(OH)_2$ (brucite) is produced by precipitation of seawater and brines with calcium hydroxide (Section 4.2.2) or by precipitation of soluble magnesium salts. Natural deposits are rare and are not currently being mined.

4.1. Properties

Magnesium oxide [*1309-48-4*], MgO (magnesia, periclase [55]), M_r 40.31, is stable in an oxidizing atmosphere up to ca. 2300 °C and up to ca. 1700 °C in a reducing atmosphere.

Crystal Structure. Magnesia forms colorless, green, or brown, cubic hexoctahedral crystals, space group $O_h^5 - Fm\,3m$, ionic radius of Mg^{2+} 0.078 nm, $a = 0.421$ nm, $Z = 4$, cleavage parallel to (100).

Physical Properties

mp	ca. 2800 °C
Mohs hardness	5.5 – 6
Density	3.58 g/cm^3
Refractive index at 589 nm	1.736
Enthalpy of formation (298 K)	− 14 900 kJ/kg
Specific heat c_p	
20 – 200 °C	1.0 kJ kg^{-1} K^{-1}
20 – 1000 °C	1.2 kJ kg^{-1} K^{-1}
Thermal conductivity	
pure crystals	
300 °C	22 W m^{-1} K^{-1}
1000 °C	7 W m^{-1} K^{-1}
technical-grade	
300 °C	12 W m^{-1} K^{-1}
1000 °C	5 W m^{-1} K^{-1}
Linear thermal expansion coefficient	
room temperature	6.7×10^{-6} K^{-1}
1000 °C	14×10^{-6} K^{-1}
1800 °C	16×10^{-6} K^{-1}
Modulus of elasticity (sintered magnesia, bulk density 3.51 g/cm^3)	
25 °C	3×10^4 MPa
1000 °C	2.5×10^4 MPa
Dielectric constant (30 Hz – 1 MHz)	11
Electrical resistivity	
room temperature	1×10^{14} Ω · cm
500 °C	1×10^{11} Ω · cm
1000 °C	1×10^{7} Ω · cm
1300 °C	1×10^{5} Ω · cm

Magnesium hydroxide [*1309-42-8*], $Mg(OH)_2$ (brucite), M_r 58.33, starts to decompose at 380 °C.

Crystal Structure. Magnesium hydroxide forms ditrigonal scalenohedral crystals with a typical layer structure, space group $D_{3d}^3 - P\,\bar{3}m\,1$, $a = 0.313$ nm, $c = 0.474$ nm, $Z = 1$, cleavage parallel to (0001).

Physical Properties

Mohs hardness	2.5
Density	2.38 g/cm^3
Refractive index	n_ε 1.566, n_ε 1.581
Enthalpy of formation (298 K)	−15 800 kJ/kg
Dissociation energy (298 K)	+570 kJ/kg
Hydration energy	−386 kJ/kg
Solubility in cold water	0.0009 wt %

Toxicity. Magnesium oxide and hydroxide are not toxic. Precautions for handling dust must be observed when products with a small particle size are used.

4.2. Production of Magnesium Oxide

Caustic and sintered magnesia are generally obtained from magnesium carbonate or magnesium hydroxide. The production of the latter starting materials is described in Sections 4.2.1, 4.2.2, 4.2.3, 4.2.4. The burning and sintering processes are dealt with in Section 4.2.5. The production, properties, and uses of fused magnesia are described in Section 4.5.

4.2.1. Production from Magnesite [56]

The production of magnesium oxide from magnesite began over 100 years ago in Austria, after the development of processes for the manufacture of steel from high-phosphorus pig

iron in crucibles lined with sintered dolomite ($CaO \cdot MgO$).

Occurrence of Magnesite. Magnesia is produced from macro- and cryptocrystalline magnesite. Metamorphic *macrocrystalline magnesite* occurs in two Alpine mountain chains in varying degrees of purity. The rocks mainly consist of the isomorphous carbonate mineral $(Mg, Fe)CO_3$ (breunnerite), whose calcination results in the inclusion of 1 – 10 % Fe_2O_3 in the periclase lattice. In addition, they contain small amounts of dolomite ($CaCO_3 \cdot MgCO_3$), calcite ($CaCO_3$), phyllosilicates (e.g., talc, mica, chlorites, or clay minerals), graphitic particles, and sometimes quartz. Similar magnesite deposits are exploited in the former Czechoslovakia, in the Pyrenees, in Ural (Satka), South Manchuria (Liaoning), East Brazil, and Canada, both by open-cast and underground mining. The macrocrystalline rocks are always subjected to some form of processing before calcination.

In contrast to the macrocrystalline magnesites, which form extensive or lenticular deposits, *cryptocrystalline magnesite* deposits are irregular and often veined. The rocks are mostly snow white, even if they are not composed of high-purity $MgCO_3$. They are regarded as being formed by the alteration of ultrabasic rocks (e.g., serpentines) through the action of water containing carbon dioxide. Deposits of this type occur in former Yugoslavia, Greece (Chalkidike and Euboa), Turkey, India, Nepal, and South Africa. They contain chalcedony (finely divided SiO_2), various magnesium silicates, and dolomite as accompanying minerals. The deposits are exploited mainly by open-cast mining, but also by underground mining.

Sedimentary deposits of cryptocrystalline magnesite have been identified in the last 20 years. They are often associated with deposits of hydromagnesite and huntite. The sedimentary deposits are formed in semi-arid basins associated with strongly weathered ultrabasic rocks. The material washed into the basin dissolved and then recrystallized on the basin rim, mostly in the form of porous concretions. A more recent formation of this type has been found in Southwest Turkey [57], while an older deposit has been reported in North Greece [58], [59]. A large deposit of this type was discovered in Australia (Kunwarara) in 1985. The Queensland magnesia project has been planned for the production of sintered and fused magnesia [60], [61].

Magnesite Beneficiation. The insufficient purity of the magnesite and the stringent specifications for the uniformity of sintered magnesia mean that beneficiation of the magnesite is usually necessary. The steps used in the beneficiation process depend on the ore being processed and the required carbonate purity.

The highest quality magnesites, particularly those for refractory applications, are needed for a magnesia product with a high MgO content, a $CaO : SiO_2$ mass ratio of 2 – 3, and low contents of Fe_2O_3 and Al_2O_3. The presence of accompanying, low-melting minerals can adversely affect the properties of the sintered magnesia (cf. Table 2).

Preliminary sorting of the ore can be achieved by *selective mining*. The ore is then subjected to *selective coarse grinding* in impact crushers. The tailings can be separated from the ore by taking advantage of the difference in hardness; the ore is selectively crushed and removed by screening. Crushing is usually performed in several stages. The final particle size for calcination in shaft kilns is 20 – 200 mm, while a particle size of 2 – 60 mm is required for the more commonly used rotary kilns.

The difference in color between the cryptocrystalline magnesite ore and the tailings is usually so pronounced that *sorting* can be carried out by hand on a conveyor belt or with an automatic optical system.

The coarsely ground material is often *washed* and subjected to *wet desliming* to remove fine particles. Ore that is intended for direct caustic calcination is separated from fine material (often lamellar silicate particles) by screening.

When the magnesite is not intimately intergrown with its accompanying minerals, lump material for rotary kilns can be sorted by *gravity separation*. This is carried out in one or two stages in a dense-medium separator or cyclone to give particles of sizes of 10 – 40 or 4 – 60 mm, respectively. Since the differences in density are small (magnesite 2.95 – 3.05, dolomite 2.90, calcite 2.70, lamellar silicates 2.5 – 2.95, quartz 2.65, graphite 2.1 – 2.2 g/cm^3), separation depends on exact control of the density. Finely divided ferrosilicon is generally used as the dense medium; the impurities are removed

Table 2. Minerals accompanying periclase in sintered magnesia

Mineral	Formula	Abbreviation [a]	C: S [b]	*mp*, °C
Forsterite [*15118-03-3*]	$2\,MgO \cdot SiO_2$	M_2S	0	1890
Monticellite [*14567-83-0*]	$CaO \cdot MgO \cdot SiO_2$	CMS	0.94	1495
Merwinite [*13596-18-4*]	$3\,CaO \cdot MgO \cdot 2\,SiO_2$	C_3MS_2	1.40	1575 (decomp.)
Dicalcium silicate [c] [*10034-77-2*]	$2\,CaO \cdot SiO_2$	C_2S	1.87	2130
Tricalcium silicate [*12168-85-3*]	$3\,CaO \cdot SiO_2$	C_3S	> 1.87	1900 (decomp.)
Magnesioferrite [*12023-16-4*]	$MgO \cdot Fe_2O_3$	MF		1750
Dicalcium ferrite [*12013-62-6*]	$2\,CaO \cdot Fe_2O_3$	C_2F	> 1.87	1435

[a] Abbreviations as used in the cement industry: M = MgO, C = CaO, S = SiO_2, F = Fe_2O_3.
[b] Mass ratio of CaO to SiO_2 (dissolution of small amounts of CaO in the MgO is neglected).
[c] Several modifications.

as floats. The throughput in modern plants is 50 – 150 t/h.

If the degree of intergrowth of the magnesite with its accessories lies in the range 0.03 – 1 mm, separation can only be achieved after thorough grinding of the ore (particle size below ca. 0.2 mm) to expose the accompanying minerals. The most effective separation and enrichment process for this material is *flotation*, which exploits the differing wettabilities of the various minerals (→ Flotation). The process must be optimized for each deposit.

The magnesite is finely ground in rod or ball mills and should have as narrow a particle size distribution as possible (ca. 0.09 mm for coarsely crystalline magnesite). This is achieved by combining milling with gravity classification and classification cyclones. Flotation is usually preceded by desliming to remove the finest particles, which would otherwise cause problems in the flotation step.

The removal of talc, $Mg_3(OH)_2Si_4O_{10}$, is straightforward. Upon mixing with air in the presence of weak collectors – frothers (e.g., pine oil and primary aliphatic fatty amines), the talc particles float to the surface on bubbles and are removed by skimming.

The main problem in flotation is the separation of magnesite from dolomite, calcite, and silicates other than talc. Separation is achieved by *direct magnesite flotation* which entails the use of nonspecific, anionic collecting agents (that also act as frothers) and selective depressing agents for dolomite, calcite, and silicates. The flotation cells have a capacity of 0.5 – 3 m^3. The operational conditions in the flotation cells must be optimized (temperature, pH value, reagent dosing). In large-scale flotation plants, several series of cells are run in parallel. The typical throughput of a series is 20 t/h of magnesite, corresponding to 60 m^3 of suspension with a density of 1.25 g/cm^3. The magnesite concentrate is removed as a foam, then thickened, and brought to a residual water content of 8 – 12 % in centrifuges or filters. The product is then briquetted and burned (see Section 4.2.5). The waste slurries from the desliming and flotation stages are treated in thickeners, filters, and settling basins.

An important sorting process for high-iron sintered magnesia is *magnetic separation* (→ Magnetic Separation). The isomorphic inclusions of iron in the magnesite crystal cannot be removed by mechanical separation. After sintering and subsequent cooling, they are converted to magnetic magnesioferrite ($MgO \cdot Fe_2O_3$), either in the periclase lattice or at its crystal boundaries (see Fig. 3B). Calcite and dolomite usually do not contain iron in the carbonate lattice; the calcination of large crystals gives large pieces of very pure, usually only lightly sintered, nonmagnetic CaO or CaO and MgO. These oxides can be separated, together with nonmagnetic silicates, from the sintered iron-rich magnesia in magnetic drum separators.

Recently developed magnets are reported to be able to remove pieces of sintered magnesia with a somewhat higher iron content from low-iron sinter [60].

Table 3 lists typical analyses of sintered magnesia produced from natural magnesite (types 1 – 4, 7).

4.2.2. Production from Seawater [62], [63]

Few deposits of pure or easily accessible magnesite were known up to the mid-1980 s. In the

Table 3. Typical analyses of sintered magnesia

Type of magnesia*	Composition, wt %						Bulk density**, g/cm^3	Theoretical density, g/cm^3
	MgO	CaO	SiO_2	Fe_2O_3	Al_2O_3	B_2O_3		
1	91.5	1.5	4.5	1.1	1.4	0.01	3.27	3.57
2	89.0	2.4	0.5	7.8	0.3		3.26	3.64
3	88.0	2.5	0.6	8.6	0.3	<0.01	3.38	3.65
4	91.3	1.9	0.5	6.0	0.3	<0.01	3.40	3.63
5	98.0	0.7	0.7	0.2	<0.2	0.30	3.28	3.57
6	97.2	1.9	0.5	≤ 0.2	0.2	0.05	3.43	3.57
7	96.4	2.0	1.1	0.4	<0.1	<0.01	3.43	3.57
8	97.2	2.0	0.4	0.3	≤ 0.1	<0.01	3.42	3.57
9	98.5	0.7	0.15	0.5	<0.1	0.015	3.45	3.58
10	99.0	0.7	0.1	<0.1	<0.1	<0.01	3.43	3.58

* 1) Lump sintered magnesia with a high silicate (and iron) content
2) High-iron sintered magnesia produced from flotation concentrate
3) Lump high-iron sintered magnesia
4) Lump high-iron sintered magnesia
5) Sintered magnesia produced from seawater type 11 (C : S = 1 : 1)
6) Sintered magnesia produced from seawater type 31 (C : S = 3 : 1)
7) Lump low-iron sintered magnesia
8) Low-iron sintered magnesia with large crystals (LC) produced from flotation concentrate
9) Sintered magnesia produced from $MgCl_2$ brine
10) Sintered magnesia produced from pyrohydrolyzed $MgCl_2$

** According to DIN 51 065, part 2.

1960s the increasing demand for sintered magnesia as a refractory lining material in the steel industry stimulated new developments in methods for production of magnesia from seawater. This method was first introduced on a small scale in 1865 on the French Mediterranean coast. Since the 1930s it has been employed on a large scale in the United States and England (Hartlepool).

The composition of seawater is given in Table 4 [64]. Calculations show that 470 m^3 of seawater are required to produce 1 t of MgO, in practice 600 m^3 are needed. The process is based on the precipitation of magnesium hydroxide (solubility in water 0.0009 wt %) by addition of calcium hydroxide (solubility 0.185 wt %):

$$Mg^{2+} + 2\,Cl^- + Ca(OH)_2 \longrightarrow Mg(OH)_2\downarrow + Ca^{2+} + 2\,Cl^-$$

Freshwater from rivers contaminated with suspended material must not flow into the seawater used for magnesia production. Furthermore, the seawater to be extracted should not be mixed with already extracted seawater, this requires a suitable topography and tidal currents that run parallel to the coast. A supply of freshwater (>40 m^3 per tonne MgO) is required to wash the $Mg(OH)_2$ and to produce the milk of lime. High-purity limestone or dolomite deposits should be available in the vicinity; they are calcined and slaked to provide $Ca(OH)_2$ as the precipitating agent and should therefore contain minimal quantities of elements that form insoluble carbonates, sulfates, etc. The $MgCO_3$ in dolomite is converted to $Mg(OH)_2$, and thus provides a further source of magnesia.

Production of Precipitating Agent. The limestone or dolomite is thoroughly washed, crushed, and calcined at >1400 °C in a rotary or shaft kiln. The $CaCO_3$ must be completely decarbonated to CaO (<0.2 % CO_2), but should be lightly burned, and should contain minimal amounts of sulfur and other contaminants. In order to achieve rapid, complete reaction of the seawater, the calcined lime (quicklime) or dolomite (dolime) is ground, slaked to form $Ca(OH)_2$ and then converted into milk of lime or dolomitic slaked lime (containing ca. 20 % solids).

The seawater is pumped to the plant and suspended material is removed. Seawater contains dissolved carbon dioxide and must therefore be

Table 4. Composition of seawater [64]

Cation	Concentration		Anion	Concentration	
	g/kg	mmol/kg		g/kg	mmol/kg
Sodium	10.47	455.0	chloride	18.97	535.1
Potassium	0.38	9.7	bromide	0.065	0.81
Magnesium	1.28	52.5	sulfate	2.65	27.6
Calcium	0.41	10.2	bicarbonate	0.14	2.35
Strontium	0.013	0.15	borate	0.027	0.44

treated with sulfuric acid to prevent the subsequent formation of insoluble calcium carbonate on addition of $Ca(OH)_2$:

$$Ca(HCO_3)_2 + H_2SO_4 \xrightarrow{pH4} CaSO_4\downarrow + 2\,H_2O + 2\,CO_2$$

The freshwater is also partially decarbonated.

To ensure maximum removal of carbon dioxide, the water is pumped through a trickling installation into the reactor containing the milk of lime. The water leaving the reactor is pumped back into the sea; it has a low magnesium content, a high calcium content, and a pH of 10.5. In modern plants using the overliming process (see below), the extracted water is first pumped into settling basins where the excess lime is diluted with fresh seawater to precipitate more $Mg(OH)_2$ and to lower the pH to more ecologically acceptable values.

*Precipitation.*The aqueous suspension of finely precipitated $Mg(OH)_2$ is seeded with larger $Mg(OH)_2$ crystals in large, flat settling tanks to convert it into a thick, dense slurry for filtering. The slurry is removed from the bottom of the tank and fed into tanks where it is washed with decarbonated seawater and decarbonated fresh water to remove most of the alkali metal, chloride, and other ions. The water content of the $Mg(OH)_2$ slurry is then reduced to $< 50\,\%$ by filtration (vacuum disk or drum filter). The product is used as a starting material for caustic and sintered magnesia (see Section 4.2.5).

Unless specially treated, caustic and sintered magnesia produced from seawater usually contain ca. 0.2 % B_2O_3 and small amounts of CaO, SiO_2, Al_2O_3, and Fe_2O_3 derived from the limestone or wastes in the seawater. The contaminants and the $CaO{:}SiO_2$ (C:S) ratio have a pronounced effect on the behavior of the magnesia during sintering and in its subsequent applications. The C:S ratio is therefore adjusted to the desired value (3 – 4 or more for the high-quality sinters used in oxygen metallurgy). The B_2O_3 content of the magnesia is also generally lowered to ca. 0.05 % by using a 5 – 12 % excess of lime for precipitation (overliming); this increases the pH to 12 and minimizes the adsorption of boron. Addition of small quantities of Na_2CO_3 prior to sintering further decreases the B_2O_3 content to 0.03 % due to the formation of volatile sodium borate during sintering. Prolonged sintering at the maximum temperature decreases the B_2O_3 content. Ion exchange can also be used to lower the boron content of the seawater and thus of the sintered product.

Typical compositions of the two most common types of magnesia sinter derived from seawater are given in Table 3 (types 5 and 6).

4.2.3. Production from Brines

Precipitation of Magnesium Hydroxide. Large marine and terrestrial salt deposits were formed in various geological periods. Fossil mineral salt deposits have been mainly mined for their potassium salts: magnesium salts or magnesium brines are obtained as waste or as a byproduct.

Magnesium salts are formed in arid areas by evaporation of brines in flat basins, their composition often differs from those obtained from seawater. Magnesia is obtained from magnesium (potassium) brines of varying origin:

1) Synthetic brines produced during solution mining (e.g., Veendam, Netherlands)
2) Spent liquors obtained in salt production from natural brines (e.g., Dead Sea, potash

production; Laguma de Rey, Mexico, sodium sulfate production)
3) Spent liquors obtained in salt production from salt deposits (e.g., Teutschenthal, Germany)

In the Netherlands (Veendam, Ost-Groningen), purified freshwater is forced into an anticline at a depth of 1500 m to dissolve the bischofite ($MgCl_2 \cdot 6\ H_2O$) and carnallite ($KCl \cdot MgCl_2 \cdot 6\ H_2O$) [65]. The brine is then pumped to the surface and contains 20 wt % $MgCl_2$ and small amounts of alkali chlorides and $MgSO_4$. After removal of the sulfate, the magnesium is precipitated as $Mg(OH)_2$ with dolomitic slaked lime (see Section 4.2.1), thickened, and then sintered. The advantage of this deposit is that the bicarbonate and boron contents of the resulting brines are much lower than those of seawater and do not have to be removed by complex methods.

Pyrohydrolysis of Magnesium Chloride [66], [67]. In pyrohydrolysis, metal (M) halides (especially chlorides) are thermally decomposed in the presence of superheated steam at 300–1000 °C to give a pure, fine metal oxide precipitate:

$$M(II)Cl_2 + n\ H_2O \rightarrow MO + 2\ HCl + (n-1)H_2O$$

Pyrohydrolysis was originally used to recover hydrochloric acid in the pickling of steel but it can also be employed for the large-scale production of metal oxides. Magnesium oxide is produced from natural or synthetic $MgCl_2$ brines. Complete hydrolysis of $MgCl_2$ by conventional methods is not possible because evaporation leads to precipitation of bischofite ($MgCl_2 \cdot 6\ H_2O$) which, on further heating, melts or dissolves in its own water of crystallization. Complete pyrohydrolysis of $MgCl_2$ is achieved in a spray reactor without hydrolysis of $CaCl_2$ and alkali chlorides (Aman process).

The *Aman process* has been used since 1973 in an industrial plant with a capacity of 70 000 t/a near the Dead Sea. The starting material for this process is the spent liquor (brine) obtained after extraction of KCl from a magnesium- and alkali-chloride-rich natural brine. The compositions of these brines are given in Table 5.

The brine is purified to remove bromide and traces of boron and then fed via steel pipes into the spray nozzles of the reactor. It is sprayed into the cylindrical, externally insulated reactor at ca. 600 °C. The reactor is lined with refractories, it also has a conical base and tangentially positioned burners which ensure rotating circulation of the reactants. The exhaust gases leave the top of the reactor at 450 °C. The water evaporates from the atomized brine droplets leaving a perforated chloride crust which reacts with the steam to form MgO and HCl. The hollow spherical MgO agglomerates are removed from the bottom of the reactor via a gas lock, together with nonpyrolyzed, water-soluble potassium, sodium, and calcium salts.

Table 5. Composition (in grams per liter) of brines used in the production of magnesia by the spray roasting method

Component	Natural brine	Brine for pyrohydrolysis after KCl extraction
Mg^{2+}	41	102
Na^{+}	40	2
Ca^{2+}	17	9
K^{+}	7.5	2
Cl^{-}	215	319
Br^{-}	5	6
SO_4^{2-}	0.7	
Total	326	440
$MgCl_2$	163	400

The crude product is washed with water and hydrated in a stirred tank, and then concentrated in a thickener. The resulting slurry is difficult to filter and is washed and dewatered in a two-stage vacuum drum filter. The calcined product typically contains ≥ 99.5 wt % MgO, < 1 wt % CaO, ≤ 0.05 wt % SiO_2, ≤ 0.05 wt % Fe_2O_3, ≤ 0.005 wt % Al_2O_3, and ≤ 0.01 % B_2O_3; its specific surface area is 2–50 m^2/g, the loose bulk density ranges from 0.8 to 0.2 g/cm^3.

Most of the MgO dust in the hydrogen chloride exhaust gas is removed with a cyclone. The hot gas is then used to heat the fresh $MgCl_2$ solution in a wash recuperator where any remaining MgO dust is collected. The HCl gas is scrubbed with water in an absorption column. After removal of the combustion gases and steam, 20 % hydrochloric acid is obtained which is used to dissolve local phosphate deposits; this makes the process economically viable.

Pyrohydrolysis of Magnesium-Containing Waste. The principle of the above method can

also be used for the pyrohydrolysis of chloride solutions obtained from impure magnesium minerals such as magnesite, tailings, dust, asbestos, talc, and other magnesium silicate waste [68]. The hydrochloric acid byproduct is recycled within the process (former Yugoslavia, Nepal, Austria, and Germany).

The magnesium is dissolved with hydrochloric acid in a two-stage process. Neutralizing agents (filter dust or crude MgO), oxidizing agents (air, chlorine gas), and sulfate ions are then added. Heavy-metal hydroxides, calcium sulfate, and other contaminants precipitate out and are filtered off to leave a $MgCl_2$ solution that is concentrated in a vacuum evaporator and sprayed into the reactor.

4.2.4. Miscellaneous Processes

Sulmag II Process [69]. The Sulmag II process was developed by Sulzer for producing light-burned caustic magnesia in a gas suspension kiln from low-magnesite ores. Dissolved magnesium chloride is obtained by selective extraction with recycled NH_4Cl solution:

$$2\,NH_4Cl + MgO + H_2O + \text{Contaminants} \longrightarrow 2\,NH_4OH + MgCl_2 + \text{Tailings}$$

After filtration and washing, needle-shaped crystals of nesquehonite are precipitated out in the reactor and filtered off:

$$MgCl_2 + (NH_4)_2CO_3 + 3\,H_2O \longrightarrow MgCO_3 \cdot 3\,H_2O\downarrow + 2\,NH_4Cl$$

Caustic magnesia with a high specific surface area is obtained by directly heating the nesquehonite with oil or gas and briquetting while still hot. The product is then sintered. This process is reported to produce significant energy savings and is environmentally favorable due to the almost complete recycling. A pilot plant has been in operation in Switzerland for several years and a plant designed to produce 30 000 t/a of sintered magnesia is under construction for Tamilnadu Magnesite in India.

Other Processes. Processes have been described in which dolomite is decarbonated to form a mixture of CaO and MgO and the CaO is then chemically extracted. Patents also describe the partial decarbonation of dolomite, slaking of the MgO, and recarbonation. The product consists of large nesquehonite crystals ($MgCO_3 \cdot 3\,H_2O$) which are separated from the microcrystalline $CaCO_3$ by flotation. The accessory calcite and dolomite can also be extracted from magnesite rocks because they are readily soluble in acid.

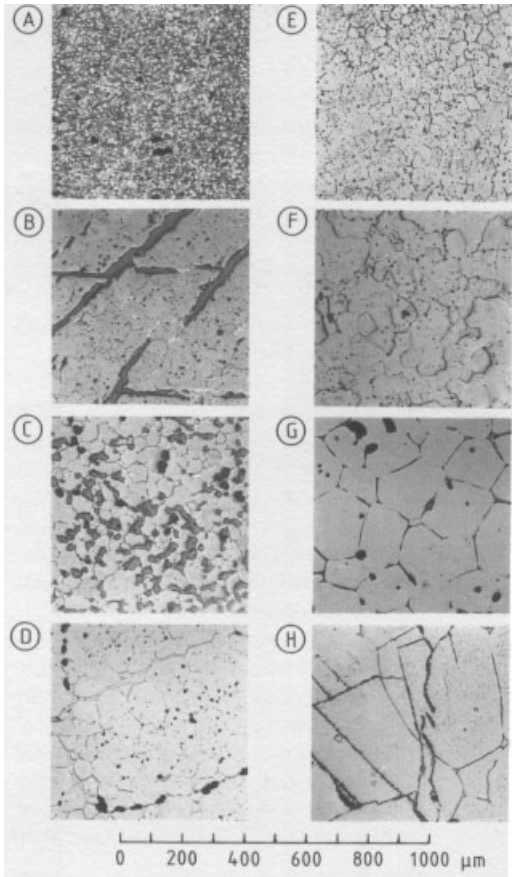

Figure 3. Polished sections of sintered dolomite (A), sintered magnesia (B – G), and fused magnesia (H)
A) Sintered dolomite: white areas are MgO, gray areas CaO, and black areas pores; B) High-iron magnesia sintered in lumps: white domains are $2\,CaO \cdot Fe_2O_3$ and $MgO \cdot Fe_2O_3$, dark gray areas are pores that run parallel to the carbonate cleavage planes (type 4 in Table 3); C) Briquetted high-iron sintered magnesia with a high porosity (type 2 in Table 3); D) Low-iron magnesia sintered in lumps (type 7 in Table 3); E) Seawater magnesia "31" (type 6 in Table 3); F) Sintered magnesia from brine (type 9 in Table 3); G) Low-iron C sintered magnesia (type 8 in Table 3); H) Fused magnesia with periclase cleavage planes

Magnesia can also be prepared from Epsom salt ($MgSO_4 \cdot 7\ H_2O$) or kieserite ($MgSO_4 \cdot H_2O$) but this technique is not used industrially at present.

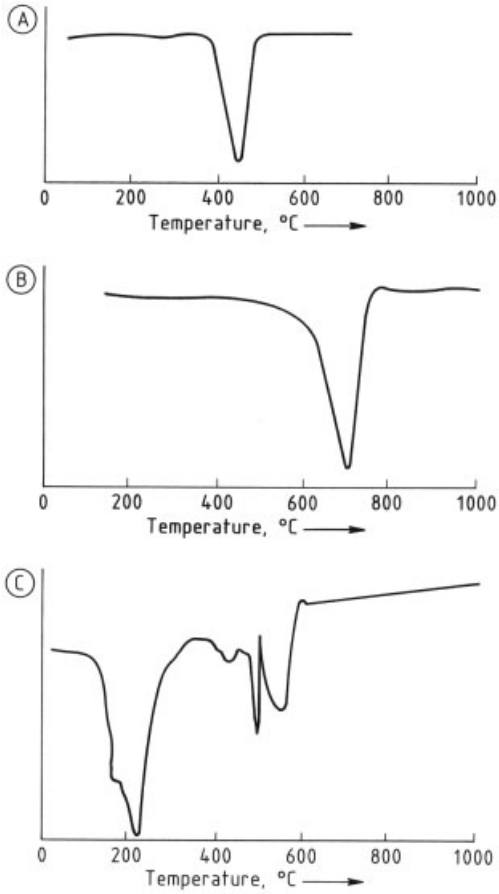

Figure 4. Differential thermal curves for $Mg(OH)_2$ (A), $MgCO_3$ (B), and $MgCO_3 \cdot 3\,H_2O$ (C) [85]

MgO and CaO Production by Sintering of Dolomite. Complete decarbonation of dolomite theoretically leaves a finely crystalline mixture of MgO (42 %) and CaO (58 %).

Since large supplies of natural, high-purity dolomitic rocks are available, removal of contaminants by beneficiation is not economical. The high-purity lump dolomite is calcined, and sintered in a rotary or shaft kiln. In a special procedure the dolomite is decarbonated, briquetted, and sintered at a lower temperature than the lump material to give a high-density product. The CaO and MgO crystals in sintered dolomite lie directly adjacent to one another (Fig. 3A); despite the intense burning they are much smaller than those of sintered magnesia because they mutually hinder one another's crystal growth.

4.2.5. Burning of Caustic and Sintered Magnesia

Caustic magnesia is a very reactive, finely crystalline material that is produced by burning $MgCO_3$ or $Mg(OH)_2$ slightly above the decomposition temperature. The decomposition curves for $Mg(OH)_2$, $MgCO_3$, and $MgCO_3 \cdot 3\,H_2O$ are shown in Figure 4. Caustic magnesia may be the end product of magnesia production (for uses, see Section 4.3.3) or may be further burned to give sintered magnesia.

Caustic magnesia is produced industrially by calcining lump $MgCO_3$ (up to 50 mm) or finer material at 600 – 1000 °C in shaft, rotary, or multiple-hearth (Herreshoff) kilns. Heat-exchange kilns are also employed. In the case of $Mg(OH)_2$, dewatered filter cakes are usually calcined in lump form at ca. 950 °C in multiple-hearth kilns. These kilns generally contain ten shelves (hearths), one above the other; each hearth is provided with four burners in the vertical cylindrical kiln wall. The material is fed continuously into the top of the kiln and its residence time can be adjusted via the rake that rotates above each hearth. The calcining conditions must be carefully adapted to the contaminants in the feed, otherwise overburning results in excessive growth of the reactive MgO crystallites which lowers their activity. In modern gas suspension furnaces, $Mg(OH)_2$ or flotation concentrate can be converted into homogeneous caustic magnesia; energy consumption is low [70].

Caustic magnesia was formerly produced exclusively from cryptocrystalline magnesite with a low iron content but is now also obtained from all types of magnesite and $Mg(OH)_2$. Its MgO content ranges from ca. 65 to 99 wt %, and may

even reach 99.9 %. The magnesia is often ground prior to use. Extremely reactive caustic magnesia may have a surface area of up to 160 m^2/g.

Depending on the burning temperature, the product is termed *light burned* (870 – 1000 °C) or *hard burned* (1550 – 1650 °C). Light-burned caustic magnesia becomes hydrated in cold water and is soluble in dilute acid. It has a loose bulk density of 0.3 – 0.5 g/cm^3 and a specific surface area (BET) of 10 – 65 m^2/g. Hard-burned caustic magnesia has a loose bulk density of 1.2 g/cm^3 (bulk density 2 g/cm^3).

Sintered Magnesia. Most (ca. 85 %) $MgCO_3$ and $Mg(OH)_2$ is converted into sintered magnesia (also known as magnesia clinker or dead-burned magnesia, Fig. 3B – G). Typical analyses and densities of sintered magnesia are given in Table 3). Uses are described in Sections 4.3.1 and 4.3.2. Sinter quality depends not only on chemical composition but also on the bulk density or porosity; a high density gives a better product. Sintering capability decreases considerably with increasing purity, making a high density difficult to achieve. In the early 1960s lump magnesite with a low iron content could not even be sintered to a density of 3.40 g/cm^3 (porosity ca. 5 vol %). Although some briquetted, low-iron, C_2S-containing sinters have densities of up to 3.46 g/cm^3 (porosity ca. 3.5 vol %), the density is usually in the range 3.35 – 3.40 g/cm^3. The final bulk density of the sinter depends on chemical composition, sintering temperature and time, and the degree of compression of the material prior to sintering (briquetting is therefore used and not pelletizing or granulation). The material is usually sintered in lump form.

Caustic magnesia (preferably hot, 200 – 500 °C) is compressed into briquettes on high-pressure roller presses (pressure 4 – 15 t/cm). The briquettes are then screened and dried or hardened. The screened fines are mixed with fresh caustic magnesia and briquetted.

Binding and adhesive agents are added to fine *magnesite flotation concentrates* prior to briquetting; magnesium sulfate solution, sulfuric acid, lignin sulfonates, and reactive MgO flue dust are usually used. Magnesite concentrates are also used to produce double-burned sinter. Here the concentrates are first calcined to reduce the mass and the resulting gas pressure; they are then briquetted and sintered.

In some cases dewatering of *magnesium hydroxide slurries* is so efficient (high-performance filter presses) that pieces of the hard filter cake can be fed directly into a rotary kiln. This saves a considerable amount of energy and the sintered magnesia product usually has a sufficiently high bulk density. In most modern plants, however, the hydroxide slurry is first calcined and then briquetted.

Shaft kilns are heated with coke, gas, or oil (with or without oxygen) and must be fed with large pieces of material due to the pattern of gas circulation and movement of material within the shaft. Today, shaft furnaces can be better controlled than was the case a few years ago [71]. Smaller lumps (0.5 – 60 mm, product 0 – 30 mm) are used in rotary kilns to minimize abrasion and guarantee even movement of the material from the rotating kiln wall. Rotary kilns are generally oil- or gas-fired, sometimes with addition of coal dust or petroleum coke. Depending on the content of fluxing agents, sintering is performed at 1700 to > 2000 °C for 2 – 15 h (rotary kiln) or ca. 20 h (shaft kiln). The sinter from rotary kilns is generally homogeneous but has a smaller particle size and contains more dust than that from shaft kilns; this can lead to undesirable, round particles and an excess of fines during brick production. Cryptocrystalline magnesite with a low content of fluxing agents gives a sinter with a needle-like structure also after sintering in a rotary kiln.

Energy consumption for the burning of high-quality sintered magnesia from various starting materials depends on kiln design and the water content of the $Mg(OH)_2$ slurry; typical values are as follows (MJ/kg MgO):

Magnesite: single-stage burning (magnesite → sinter)	4.5 – 12
Magnesite: double-stage burning (magnesite → caustic magnesia → sinter)	8 – 16
Seawater magnesia [$Mg(OH)_2$ → caustic magnesia → sinter, value includes energy required for production of precipitation agent]	22 – 32
Sulmag II process	23
Magnesia produced by pyrohydrolysis	40

In an effort to improve sinter quality, manufacturers first turned their attention to increasing the size of the periclase crystallites by adding sinter-promoting agents, such as ZrO_2. Burning conditions were subsequently modified. In larger crystallites the MgO is less easily reduced by carbon at high temperatures (e.g., in magnesia – graphite bricks; see also Fig. 3G).

Modification of Sinter Composition. In the case of concentrates, slurries, or caustic magnesia, the final composition of the sinter can be modified by mixing in small amounts of additives such as CaO, SiO_2, and Fe_2O_3 prior to briquetting. Magnesite concentrates obtained by flotation and fine $Mg(OH)_2$ or MgO can also be mixed with large quantities of other fine substances (primarily chromium ores, calcined lime or dolomite), briquetted, and sintered to form coclinker. In refractory bricks made from pure sintered magnesia and low-silicate chromium ores, the chromium ore only becomes fully integrated with the periclase at high temperature ($\geq$ 1700 °C) and in a suitable atmosphere. In briquettes made from magnesia and fine chromium ore, however, sintering results in complete dissolution of the mixed spinels of the ore in the periclase. On cooling, fine crystalline spinels or skeleton crystals precipitate in the periclase; some idiomorphic spinels (Mg, Fe^{2+})O (Cr, Fe^{3+}, Al)$_2O_3$ are evenly deposited between the periclase crystallites.

4.3. Uses

The use of magnesium oxide and hydroxide as intermediates in the production of magnesium metal is described elsewhere (→ Magnesium).

4.3.1. Use of Sintered Magnesia as a Refractory

Oxides are best suited as "vessel linings" for most industrial high-temperature processes. Magnesium oxide has the highest melting point of the moderately priced oxides and is therefore an important raw material for refractory bricks and other materials (→ Refractory Ceramics). It is the only material apart from ZrO_2 that can withstand long-term heating above 2000 °C. Refractory construction materials based on magnesia (with chromium ore or CaO) are also resistant to neutral and basic slags and gases.

The boom in the production of low-iron sintered magnesia occurred after it was discovered that the use of refractory magnesia linings in steelmaking vessels (open-hearth furnace, electric arc furnace, and basic oxygen furnaces) allowed phosphorus and sulfur to be removed in the basic slag.

Sintered magnesia is converted into fired bricks by techniques used for traditional ceramics. It is first broken up, ground, and classified. Batches (1 to > 2 t) of individual fractions are mixed with a temporary aqueous binder (e.g., lignin sulfonates) and pressed into bricks of the desired shape weighing up to ca. 35 kg. After drying, the bricks are usually fired at 1500 – 1800 °C in a tunnel kiln. They have a final porosity of 12 – 20 vol %.

Bricks (especially those that are to be used in basic oxygen furnaces) may then be *impregnated with pitch* to prevent the slag from infiltrating them and dissolving their surface. Fired magnesia bricks have a high thermal expansion coefficient and a relatively low thermal shock resistance, especially during periodic operation. Fired bricks made from sintered magnesia and chromium ore (or alumina or spinel) are more resistant to thermal shock.

A large proportion of the magnesia bricks are not fired (i.e., do not form a ceramic bond) but are *bonded with pitch or resin*. The individual particle fractions are mixed with hot pitch, compressed into bricks, and tempered at 200 – 500 °C. When the bricks are used, a coke structure is formed which makes them more elastic than the ceramic-bonded bricks. Addition of 5 – 20 % flake graphite increases heat transfer, lowers thermal expansion and attack by the slag, and improves thermal shock resistance. This development is applied in electric arc furnaces, converters, and metal ladles for secondary steelmaking processes. The low-oxygen atmosphere formed in these bricks during use can cause significant reduction of the magnesia. To prevent this, manufacturers have developed high-quality magnesia with larger periclase crystals and a low internal sinter surface area [72].

Chemically bonded bricks with chrome ore contain chemical bonding agents (e.g., Epsom salt, $MgSO_4 \cdot 7\ H_2O$) and are produced in the same way as fired magnesia bricks up until the

drying stage. The bonding agents decompose and form a ceramic when the bricks are used at high temperature.

Pig iron mixers, transport vessels, and secondary refining ladles used in the steel industry are lined with magnesia bricks. Sliding gates and nozzles are also made from sintered magnesia.

On account of their high refractoriness (refractoriness under load and low creep in compression), resistance to alkali, and high heat storage capacity, magnesia bricks are used in the crowns, port linings, and the hot parts of the checkerwork in glass-melting tanks. Magnesia–chrome or magnesia–spinel bricks are more suitable for the sintering zones of rotary cement kilns due to their good stress deformation behavior. Magnesia bricks are preferred for lime shaft kilns. Large quantities of sintered magnesia are used for unshaped repair materials.

The increased demands made on refractory materials as a result of higher operating temperatures and shorter tap to tap times in metallurgical furnaces and reactors can only be met by pure, high-density magnesia sinters. Small quantities of "contaminants" are disadvantageous if they form low-melting eutectics with MgO (e.g., with CMS at 1485 °C or with C_2F at 1200 °C, Table 2) because this leads to deterioration of mechanical properties (e.g., strength and volume stability) at high temperatures. High-quality sinters therefore have a low content of high-melting silicate phases (such as C_2S), a low B_2O_3 content, and a high degree of direct periclase–periclase contact (without intermediate silicate phases).

4.3.2. Use of Sintered Magnesia as a Heat-Storage Medium

Magnesia bricks have a high heat storage capacity (about 160 % that of chamotte) and a high thermal conductivity [73]. They are used in efficient off-peak storage heaters. The heat generated by a heating element is transferred to the magnesia brick and increases its temperature.

The most suitable bricks are fired and have the following properties: a high bulk density (up to 3.10 g/cm^3) and thus a high degree of sintering, a high thermal conductivity (up to 9 W m^{-1} K^{-1} at 600 °C), and a high specific heat (1.1 kJ kg^{-1} K^{-1} at 600 °C). The influence of impurities on these factors is not as critical as in refractories.

The bulk density depends on the chemical composition, the sinter porosity and the method used to produce the bricks. Thermal conductivity is increased by a high periclase content and a low porosity. The specific heat is only slightly lowered by SiO_2 and Al_2O_3, but is significantly lowered by CaO, Cr_2O_3, and Fe_2O_3. The bricks should not contain free CaO (risk of hydration) or crystal phases with different modifications.

4.3.3. Uses of Caustic Magnesia

Agriculture. In the 1970 s and 1980 s the use of caustic magnesia increased the most in fertilizers and animal feeds [74–77]. Magnesium is a constituent of plant chlorophyll and also plays a role in enzyme activity. It is regarded as an essential plant nutrient and is used in *fertilizers* (→ Fertilizers). Magnesium deficiency (symptoms: chlorosis, necrosis) became especially common after fertilizers containing a high percentage of potassium replaced potassium–magnesium fertilizers based on kainite ($KCl \cdot MgSO_4 \cdot 3\ H_2O$). In countries with salt deposits (e.g., Germany), magnesium is supplied in the form of kieserite, magnesia lime (> 10 % MgO), and calcined dolomite. In all other countries (e.g., Netherlands, UK, USA), caustic magnesia (particle size 4 mm, ≥ 70 % MgO) and magnesite are used.

Ground caustic magnesia (particle size ≤ 2 mm, ≥ 85 % MgO) is used as a *supplement for animal feeds*. Magnesium is usually present in green fodder, but cattle and sheep feeding on fodder from slightly acid soils may suffer from "grass staggers" in spring which is caused by magnesium deficiency.

Building Industry. Caustic magnesia mixed with concentrated solutions of magnesium salts and a small amount of sodium phosphate becomes extremely hard and sets in air (e.g., with $MgCl_2$ solutions it forms $3\ MgO \cdot MgCl_2 \cdot 11\ H_2O$). However, the hardened material is not stable in water and is therefore not a true cement (which must set in air and water) but a *magnesia binder*. These binders were originally referred to as Sorel cement or as stonewood (xylolith) when mixed with sawdust as a filler. Since their

high chloride content can corrode metals and other construction materials, magnesia binders must never be reinforced. They can be used as mastics or protective coatings at room temperature. Large quantities of caustic magnesia derived from carbonate are used for stonewood or plaster flooring in the Eastern Bloc. The low-iron caustic magnesia generally used for this purpose contains 75 – 87 % MgO and is ground (40 –45 % < 10 μm).

An important use of caustic magnesia is in *lightweight construction boards* (Heraklith) for thermal and acoustic insulation (thickness 5 – 100 μm). The boards are produced on fully automatic lines at ca. 60 °C from finely ground caustic magnesia (MgO content 65 – 80 %, starting material magnesite containing silicates and dolomite), $MgSO_4$ solution as a binder, and impregnated pinewood chips. Mineral wool is also used as a filler. Subsequent hardening results in highly nonflammable boards which can be made fire-retardant and resistant to moisture and microorganisms by coating one side with plaster. The prefabricated boards are used to make formworks for concrete, and large quantities are used in combination with polystyrene in laminated boards.

Other Uses. Caustic magnesia containing ca. 70 – 99 % MgO is used in wastewater treatment, in prevention of corrosion, as a buffer, in preparative chemistry, and in the pharmaceutical industry.

Caustic magnesia can be used to remove heavy metals and silicate from wastewater. Ammonia can be precipitated and removed in the form of $Mg \cdot NH_4 \cdot PO_4 \cdot 6\,H_2O$ following addition of phosphoric acid and caustic magnesia [78]. The magnesium ammonium phosphate can be used as a fertilizer. This principle is also used to remove phosphate. Caustic magnesia is a mild base and neutralizes acids and acidic water, it also acts as a weak buffer.

Low-iron caustic magnesia is used as a filler in the plastics and rubber industries and allows adjustment of viscosity and stiffness. It also acts as a chlorine accepter in chlorinated hydrocarbon polymers and stabilizes them.

Large quantities of pure magnesium oxide are used in the chemical industry for the synthesis of magnesium compounds. Basic magnesium sulfonates and finely dispersed MgO are used as lubricant and motor fuel additives to neutralize combustion residues. Magnesium oxide in fuels binds sulfur and other compounds and prevents formation of acidic exhaust gases.

Large quantities of caustic magnesia are used in the paper and cellulose industries (magnesium bisulfite digestion). Caustic magnesia with graded particle size is also used as a polishing agent.

Caustic magnesia, magnesium carbonate (precipitated from caustic magnesia), and high-purity magnesia are utilized in preparative chemistry and cosmetics. In the pharmaceutical industry magnesium oxide ($\geq$ 98 % MgO) is used as an antacid, as a powder base, and for magnesium substitution therapy.

A potential use of caustic magnesia is as a slag additive to reduce the dissolution of MgO from refractory bricks in the steel industry. Treatment of the surface of grain-oriented steel transformer sheets decreases magnetic losses. A suspension of caustic magnesia is applied to the sheet and annealed.

4.3.4. Uses of Magnesium Hydroxide

The most important use of magnesium hydroxide is as an intermediate in the production of magnesia. In Japan it is used to replace polyhalogenated diphenyl ether as a flame retardant in thermoplasts processed at 200 – 350 °C. Hydrated alumina cannot be used at these temperatures because of its low decomposition temperature. This application is being tested in Germany [79]. Well-dispersed single $Mg(OH)_2$ crystals seem to be more suitable than agglomerates.

4.4. Economic Aspects

Developments in the production capacities, production, and consumption of magnesia can be briefly summarized as:

1974 – world magnesia shortage [80]
1977 – from shortage to surplus [81]
1981 – agricultural oversupply and industrial underdemand for caustic magnesia [75]
1982 – contraction of refractory magnesia and stagnation of caustic magnesia [82]
1984 – China, the magnesite giant [83]

1987 – the world magnesia industry, smaller but fitter and purer [84]
1989 – magnesia markets – fit for optimism [60]

These changes reflect the dependence of economics on the economic and technological developments in the steel industry. The discovery of highly pure magnesite deposits in Australia in 1985, which will allow magnesia to be obtained directly from the carbonate, has led not only to optimism for producers there, but also to concern among those using the energy-intensive process via magnesium hydroxide.

Recent developments are aimed at reducing energy consumption by improving the filterability of precipitated products, optimizing kiln design, and increasing the size of the periclase crystallites and the bulk density of sintered magnesia. The estimated production capacities for caustic and sintered magnesia are given in Table 6. The estimated production of crude magnesite ore increased from 16×10^6 t/a in 1981 to 20×10^6 t/a in 1988.

4.5. Fused Magnesium Oxide

4.5.1. Properties

Fused magnesia is crystalline magnesium oxide (*mp* 2800 °C) that has been melted in an electric arc. It has the same crystal structure as sodium chloride (i.e., face-centered cubic) and occurs naturally as periclase. No phase changes take place on heating up to the melting point.

Pure fused magnesia is white, although the presence of iron impurities impart a greenish color due to the presence of the mixed crystal compound magnesiowüstite (Mg,Fe)O. When the pale green fused magnesia is heated in air, the color changes to brown owing to the formation of magnesioferrite ($MgFe_2O_4$) [86].

Fused magnesium oxide has a much lower tendency to undergo hydration than sintered magnesium oxide or caustic calcined magnesite [87], and is essentially stable towards the atmosphere. Thus, polished crystal faces of periclase only begin to lose their luster after several weeks, depending on atmospheric moisture content and temperature. Reaction with atmospheric carbon dioxide to form basic carbonates is also noticeable only after several weeks or months.

Fused magnesia is stable in an oxidizing atmosphere up to ca. 2200 °C, above which vaporization and partial dissociation take place. Theoretical vapor pressures of magnesium oxide and partial pressures of its dissociation products in the temperature range 1227 – 2727 °C are given in [88].

In a reducing atmosphere (e.g., in the presence of carbon), fused magnesia is stable only up to ca. 1700 °C. This temperature is also the limit for vacuum use because of the resultant vapor pressure [89].

The most important physical properties of fused magnesia (e.g., thermal conductivity and electrical resistivity) depend on the density, purity, particle size distribution, type, and distribution of the foreign mineralogical phases present. Consequently, data obtained from measurements on single magnesium oxide crystals [90] or pure, polycrystalline magnesium oxide (> 99 %) cannot be regarded as characteristic of commercial fused magnesia.

The following data (see also Chap. 4) are therefore partly based on information from manufacturers of fused magnesia and include the grades most commonly available (MgO content > 92 %, particle size 0 – 400 μm):

Density of pure periclase (theoretical)	3.56 – 3.58 g/cm^3
Knoop hardness	370
Mohs hardness	6
Linear thermal expansion coefficient	
(1000 °C)	14×10^{-6} K^{-1}
Specific heat (c_p)	
100 °C	0.96 kJ kg^{-1} K^{-1}
1800 °C	1.21 kJ kg^{-1} K^{-1}
Thermal conductivity (1000 °C) for densities of	
3.0 – 3.58 g/cm^3	ca. 2.9 – 8.4 W m^{-1} K^{-1}
Dielectric constant, ε (30 Hz – 1 MHz)	11
Electrical resistivity (technical grades, manufacturers' data)	
600 °C	$2\times10^9 - 10^{11}$ Ω · cm
800 °C	$3\times10^8 - 2\times10^9$ Ω · cm
1000 °C	$10^7 - 9\times10^7$ Ω · cm
Electrical resistivity of MgO single crystals (for comparison)	
500 °C	2×10^{13} Ω · cm
1000 °C	6×10^8 Ω · cm

4.5.2. Production

Fused magnesia is mainly produced from naturally occurring magnesite ($MgCO_3$) that

Table 6. Production capacities for sintered and caustic magnesia in 1988/1989 (10^6 t/a) [60]

Country	Sintered magnesia		Caustic magnesia	
	Natural magnesite	Seawater and brines	Natural magnesite	Seawater and brines
EEC	0.6	0.5	0.3	0.1
Western Europe (not EEC) and Asia Minor	0.9	*	0.2	*
Eastern Bloc	2.8	0.14*		< 0.1*
North and South America	0.4	0.7	0.2	> 0.1
Asia	2.0	0.6	0.3	< 0.1
Australia	$\ll 0.1$		$\ll 0.1$	
Africa	$\ll 0.1$		$\ll 0.1$	
World total	6.8	1.9	1.0	0.4

* Including Western Europe (not EEC), Asia Minor and Eastern Bloc.

has been calcined at 1200 – 1400 °C ("dead burned"). The calcined product usually contains > 92 % MgO, 1 – 4 % SiO_2, 0.5 – 2 % CaO, 0.1 – 2 % Al_2O_3, up to 0.2 % Fe_2O_3, and traces of ZrO_2, NiO, Na_2O, K_2O, B_2O_3, and S. Magnesite from seawater (see Section 4.2.2), containing > 97 % MgO, is also used as a raw material.

When fused magnesia is used as an electrical insulator for heating applications (tubular elements), the amount of conductive impurities in the raw material must be minimal. Therefore magnesites that contain, for example, iron oxides or sulfur in appreciable amounts cannot be used. Contamination with sulfur is avoided by calcining the magnesite in wood-fired furnaces.

Fused magnesia is usually produced in a batch process by melting the raw material at 2800 – 3000 °C in electric arc furnaces [91].

The furnaces are heated by either single-phase or, more usually, three-phase a.c. current (Higgins furnace, block-making process). They have a moving base and a conical, water-cooled steel cover. The raw material is added batchwise and acts as a refractory lining for the furnace walls, and the furnace base is usually lined with magnesite bricks or a rammed graphitic material bonded with bitumen. The furnaces are heated by either two or three graphite electrodes. The arcs are struck onto a layer of coke and the electrodes can be vertically adjusted to suit the height of the molten bath.

The current in, for example, a 1000-kW a.c. electric arc furnace (Higgins furnace) is 6000 – 9000 A at 90 – 150 V. The electric power required to produce the fused blocks is ca. 2000 – 4000 kW · h/t [91].

In addition to the traditional block-making process (Higgins furnace), tilting furnace technology is also used for the industrial production of fused magnesia [92]. The melting process is continuous, and the molten product is poured intermittently into water-cooled or refractory-lined molds. Manufacture of fused magnesia by an electric arc trough melting process [93] and in a high-frequency plasma [94] has also been described.

On completion of melting, the electrodes are withdrawn and, after a cooling period, the furnace cover is removed. The fused magnesia blocks can weigh as much as 20 t. After cooling to room temperature, preliminary size reduction is carried out. This is followed by further size reduction of the large broken pieces. The blocks are not homogeneous with respect to structure or chemical composition, and their surface is covered with a coating or crust of agglomerated sintered magnesia.

The impurities in the raw product consist of compounds formed from the quaternary system $MgO-CaO-SiO_2-Fe_2O_3$, e.g., magnesium silicates such as forsterite ($2\,MgO \cdot SiO_2$, *mp* 1890 °C) and monticellite ($CaO \cdot MgO \cdot SiO_2$, *mp* 1500 °C, incongruent melting) whose melting points lie far below that of pure magnesium oxide. The foreign phases are mainly concentrated locally in the surface zone of the fused block and constitute the "melt residues" along with imperfectly melted magnesium oxide (the sintered crust). They are hand separated from the fused product which is a white to pale green

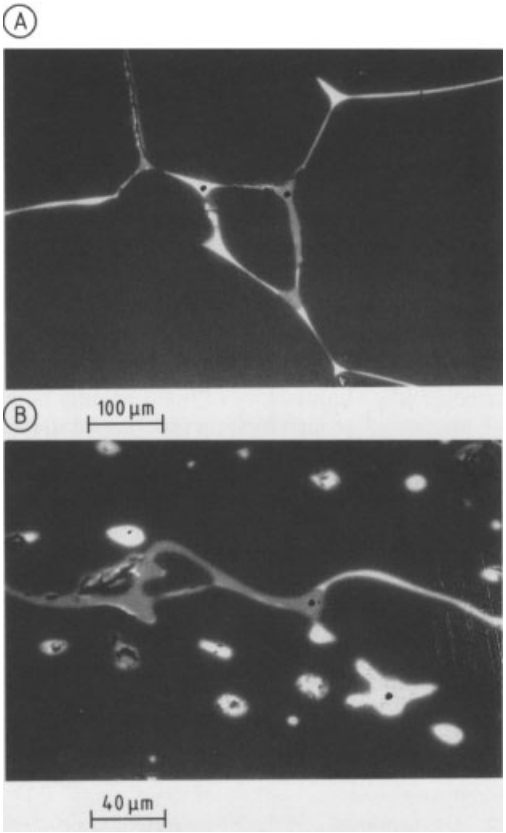

Figure 5. Photomicrograph of polished sections of fused magnesia (dark areas: MgO; light areas: foreign phases monticellite and forsterite)
A) Magnesia produced in a Higgins furnace: foreign phases deposited at the grain boundaries; B) Magnesia produced by casting in a tilting furnace: foreign phases occur mainly as intracrystalline material

crystalline periclase with crystallite sizes from several hundred micrometers up to several centimeters. Hand sorting is the normal industrial method of impurity removal, no better method has yet been found. However, some of the silicate impurities associated with the raw material form inclusions of foreign phases in the crystals of the pulverized magnesium oxide end product. This is because the rate of growth of the magnesium oxide crystals in certain regions of the solidifying melt is greater than the migration rate of the melt residues. Thus, magnesium oxide blocks from a Higgins furnace contain foreign phases (monticellite and forsterite) deposited at the grain boundaries, while magnesium oxide produced by casting from a tilting furnace usually exhibits an intracrystalline distribution of these phases (see Fig. 5). Further treatment of the hand-sorted, size-reduced, fused product includes several size reductions with jaw crushers, roller crushers, and gyratory crushers. Iron particles produced by abrasion and other magnetic impurities are removed by high power magnetic separators. Finally, the finely grained material is classified into particle size ranges using a series of sieves. Fused magnesia intended for electrical heating applications is heated in an oxidizing, neutral, or reducing (although not carbon-forming) atmosphere at 1100 – 1400 °C [95] to remove moisture and carbon (originating from breakdown of the electrodes) and to reduce surface and lattice defects of the magnesium oxide particles.

4.5.3. Uses

Fused magnesia is mainly used as an *insulating material* in the electrical heating industry (→ Insulation, Electric). The particulate material is packed into the space between the heating coil and the outer tube in a heating element for air or liquids [96]. Such elements are used in IR heaters, grill elements, tubular hot plates, and tubular heating elements for ovens, storage heaters, radiators, continuous-flow heaters, washing machines, coffee machines, electric irons, and industrial liquid heating equipment.

The most important factor in this application of fused magnesia is its combination of high electrical resistance with high thermal conductivity. This is superior to that of other refractory oxides or other materials.

The grade of magnesium oxide recommended depends on the application for which the heating element is used (surface temperature of the tube) and the electrical insulation requirements. The grades usually have a size range of 0 – 400 µm and a tap density of 2.2 – 2.7 g/cm^3. The product must have well-defined flow properties, which can be influenced by modifying the particle shape by treatment in a fluidized bed [97].

The electrical and thermal properties of fused magnesia can be improved by adding synthetic or natural minerals (e.g., pyrophyllites [98], fused zircon sand [99], enstatite [100], cordierite [101], or magnesium aluminum silicates [102]) to the prepared mixture of magnesia particles.

The mechanism of action of these additives has not yet been explained.

The chemical composition of some of the industrial grades of fused magnesia (electrical grade) are given in Table 7. The important properties of the mixture of fused magnesia particles (as used for electrical heating, i.e., thermal conductivity, long life time, sinter index, specific electrical impedance, leakage current, tap density, flow rate, grain size distribution) are tested by standard methods (ASTM D 2858, D 2900, D 3026, D 3215, D 3347; DIN 44 872).

Fused magnesia has a further application as a *refractory material*, especially as a lining for induction furnaces and crucibles. Induction furnace linings based on fused magnesia do not shrink (unlike those based on sintered magnesia) owing to the low porosity of the fused material. Magnesia–graphite bricks with an addition of fused magnesia [103] are used in ultrahigh-power furnaces for steel production. The low porosity of the fused magnesia and the size of the periclase crystals considerably improve the corrosion resistance of the refractory lining. The refractory – grade material has a lower SiO_2 content and a higher CaO : SiO_2 ratio than the electrical-grade fused magnesia (i.e., for electrical heating, see Table 7).

In *special applications*, fused magnesia is also used as a molding material for precision casting, as a raw material for magnesium oxide ceramics and as single crystals [104] for optical windows and lenses.

4.5.4. Economic Aspects

The most important countries and companies producing fused magnesia are:

France	Pechiney Electrometallurgie
Germany	Hüls AG
United Kingdom	Universal Abrasives, Thermal Syndicate
Austria	Österreichisch Amerikanische Magnesit AG (refractory grades)
United States	Muscle Shoals Minerals, Tateho America (formerly CE-Minerals), Insultherm
Canada	Baymag (refractory grades)
Japan	Tateho Chemical Ind.

Further information on these manufacturers (e.g., production capacities) is given in [105].

The world consumption of fused magnesia for electrical heating (electrical grade) is estimated to be ca. 35 000 – 40 000 t/a. There are no figures available for refractory grade material because the grades of fused magnesia are not listed separately in the statistics.

5. Magnesium Sulfate

Magnesium sulfate does not occur in nature in anhydrous form. It is found in the form of hydrates and double salts in salt and potash deposits (e.g., in the Unites States, Germany, Russia, and the Ukraine). Magnesium sulfate also occurs in dissolved form in salt lakes.

5.1. Properties

The transformation temperatures of the stable magnesium sulfate hydrates and their solubilities in water are given in Table 8. Table 9 lists the molecular masses, the specific heats, refractive indices, and crystal classes of the anhydrous salt and some of the hydrates [106], [107]. Anhydrous magnesium sulfate can be obtained without decomposition by dehydrating the hydrates at 400 – 500 °C. At 700 °C, however, the compound decomposes to give magnesium oxide, sulfur dioxide, and oxygen. This accounts for the fact that the melting point of magnesium sulfate cannot be accurately determined, it is between 1120 and 1150 °C.

Magnesium sulfate is hygroscopic and absorbs water to form hydrates, finally the heptahydrate. Various hydrates can be obtained by stepwise dehydration of higher hydrates or by crystallization from aqueous solution. Hydrates of analogous composition can be produced by these two methods, but X-ray crystallography shows that their structures differ due to destruction of the crystal lattice during dehydration.

Five stable hydrates can be crystallized from aqueous solutions: $MgSO_4 \cdot 12\,H_2O$, $MgSO_4 \cdot 7\,H_2O$ (Epsom salt, epsomite, Bittersalz), $MgSO_4 \cdot 6\,H_2O$ (hexahydrate), $MgSO_4 \cdot 5/4\,H_2O$ (5/4 hydrate or synthetic kieserite) and $MgSO_4 \cdot H_2O$ (kieserite). Three metastable forms also occur: a second form of the heptahydrate, $MgSO_4 \cdot 5\,H_2O$ (allenite), and $MgSO_4 \cdot 4\,H_2O$ (leonardite). Stepwise dehydration yields hydrates with 1,

Table 7. Chemical composition (typical) of some commercially available fused magnesia grades

Producer	Grade	MgO, wt %	SiO_2, wt %	CaO, wt %	Fe_2O_3, wt %	CaO : SiO_2 ratio	Density, g/cm^3
Pechiney, France	CA 250 [a]	95.5 (min.)	2.2	1.5	0.008	0.68	2.39 [c]
Hüls, FRG	Dynatherm [a] 1246 CS	96.3	2.0	1.5	0.10 (max.)	0.75	2.36 – 2.40 [c]
Thermal Syndicate, UK	Maglox [a] 1GN1	96.8	2.3	0.8	0.10	0.35	2.35 – 2.40 [c]
Universal Abrasives, UK	M 70 [a]	93.1	3.7	1.6	0.16	0.43	2.36 – 2.40 [c]
Muscle Shoals Minerals, USA	22 SR [a]	96.0	3.1	0.8	0.07	0.26	2.38 – 2.41 [c]
Pechiney, France	UR1 AT [b]	98.0	0.5	1.1	0.15	2.2	3.50 [d]
Baymag, Canada	Electromag [b]	96.7	0.2	2.2	0.5	11	3.50 [d]
Tateho, Japan	KMA [b]	98.5	0.35	0.9	0.15	2.6	3.57 [d]

[a] Electrical grades for high-duty tubular heating elements.
[b] Refractory grades.
[c] Tap density according to ASTM D 3347, grain size 40 – 325 mesh.
[d] Apparent density.

5/4, 2, 3, 4, and 6 molecules of water. The solubility of kieserite decreases with increasing temperature due to its negative heat of dissolution. Its dissolution rate increases with increasing temperature and decreasing pH.

Table 8. Stable magnesium sulfate hydrates

Hydrates	Transformation temperature, °C	Solubility, g $MgSO_4$/100 g H_2O
Ice – $MgSO_4 \cdot 12\,H_2O$	−3.9	22.0
$MgSO_4 \cdot 12\,H_2O$ – $MgSO_4 \cdot 7\,H_2O$	1.8	27.2
$MgSO_4 \cdot 7\,H_2O$ – $MgSO_4 \cdot 6\,H_2O$	48.1	49.5
$MgSO_4 \cdot 6\,H_2O$ – $MgSO_4 \cdot H_2O$	67.5	56.6

The acid sulfates $MgSO_4 \cdot H_2SO_4$ and $MgSO_4 \cdot 3\,H_2SO_4$ crystallize when $MgSO_4$ is dissolved in sulfuric acid. Magnesium sulfate ammine hydrates ($MgSO_4 \cdot 3\,NH_3 \cdot 3\,H_2O$, $MgSO_4 \cdot 2\,NH_3 \cdot 4\,H_2O$, and $MgSO_4 \cdot 2\,NH_3 \cdot 2\,H_2O$) are produced when Epsom salt reacts with gaseous ammonia. On reaction with alkali, magnesium hydroxide precipitates. A precipitate of nesquehonite ($MgCO_3 \cdot 3\,H_2O$) is formed with sodium carbonate. After addition of sulfite, phosphate, or stearate solutions, insoluble magnesium sulfite, magnesium phosphate, or magnesium stearate precipitates. Magnesium chloride reacts with magnesium oxide to form oxychlorides; magnesium sulfate reacts in the same way to form oxysulfates ($MgSO_4 \cdot 3\,MgO \cdot 11\,H_2O$ and $MgSO_4 \cdot 5\,MgO \cdot 8\,H_2O$): it can therefore be used as a slower type of sulfate Sorel cement binder (see Sections 3.7, 4.3.3). Magnesium sulfate forms double salts with other salts, most of which also occur in nature:

$KCl \cdot MgSO_4 \cdot 3\,H_2O$ [*67145-93-1*] kainite
$K_2SO_4 \cdot 2\,MgSO_4$ [*13826-56-7*] langbeinite
$K_2SO_4 \cdot MgSO_4 \cdot 4\,H_2O$ [*15226-80-9*] leonite

Table 9. Properties of magnesium sulfate and its hydrates

Compound [CAS No.]	M_r	Specific heat, kJ kg^{-1} K^{-1}	Density, g/cm^3	Crystal class	Refractive index
$MgSO_4$ [*7487-88-9*]	120.37	0.800	2.66	orthorhombic	
$MgSO_4 \cdot H_2O$ [*14168-73-1*]	138.38	1.047	2.57	monoclinic	1.523, 1.535, 1.586
$MgSO_4 \cdot 2\,H_2O$ [*17830-05-6*]	156.40	1.124			
$MgSO_4 \cdot 4\,H_2O$ [*24378-31-2*]	192.43	1.305	2.01	monoclinic	1.490, 1.491, 1.497
$MgSO_4 \cdot 6\,H_2O$ [*13778-97-7*]	228.46	1.525	1.75		1.456, 1.453, 1.426
$MgSO_4 \cdot 7\,H_2O$ [*10034-99-8*]	246.48	1.546	1.68	orthorhombic	1.432, 1.4554, 1.4609

$K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O$ [*15491-86-8*] schönite (picromerite)
$Na_2SO_4 \cdot MgSO_4 \cdot 4\,H_2O$ [*15083-77-9*] astrakanite (bloedite)
$6\,Na_2SO_4 \cdot 7\,MgSO_4 \cdot 15\,H_2O$ [*16633-52-6*] loeweite
$K_2SO_4 \cdot MgSO_4 \cdot 2\,CaSO_4 \cdot 2\,H_2O$ [*15278-29-2*] polyhalite

5.2. Occurrence and Raw Materials

The most important source of $MgSO_4$ in Germany is kieserite, which is a constituent of raw potash salt. In the United States, $MgSO_4$ is obtained from langbeinite and in Italy from kainite. The raw salts used by the potash industry in the Ukraine (Stebnik, Kalush) contain magnesium sulfate salts. Magnesium sulfate is also found in the brines of the Great Salt Lake and the Kara – Bogaz Gulf of the Caspian Sea and in other lakes, for example, the Dalagton Basin in China [108]. Seawater bitterns provide a further source (Epsom salt, see Section 5.3.2), the $MgSO_4$ is obtained by cooling. In many countries $MgSO_4$ is also produced from magnesite, dolomite, seawater magnesia, or other magnesium minerals.

5.3. Production

5.3.1. Production of Kieserite ($MgSO_4 \cdot H_2O$)

Production of Kieserite by Selective Dissolution of Residues from Potash Salt Production. The raw material for this process is a crude salt that contains kieserite (hard salt, carnallitite). The sodium chloride is selectively dissolved from residues of potash salt production. The filtered or centrifuged residues obtained during the production of potassium chloride are first washed with a partially saturated NaCl solution from a subsequent washing step. The NaCl in the residues is readily soluble and dissolves in the NaCl solution until it becomes saturated; the kieserite, however, remains largely undissolved due to its slow dissolution rate. Kieserite and residual NaCl are separated from the washing medium with a bucket conveyor or elevator. The kieserite – NaCl mixture is then washed with water in a second stage to dissolve any residual NaCl. Washing is performed in pipes lined with fused basalt that are several hundred meters long. The quantity of NaCl solution produced can be minimized by using the electrostatic production process (see below).

Ecological Aspects. The saturated NaCl solution contains 280 – 300 g NaCl per liter. Its disposal entails considerable ecological problems. Large quantities have been stored in jointed rock formations (platy dolomite) that are filled with slightly salted groundwater, but with only limited capacity [109].

Production of Kieserite by Flotation. Kieserite can be separated from rock salt by flotation. Prestaminol, oleic acid, fatty alcohol sulfates, or oxystearic acid are used as flotation agents for kieserite and anhydrite ($CaSO_4$) [110]. The kieserite can be further purified by separating it from the anhydrite and/or langbeinite. This is achieved by using oleic acid or a fatty acid amine as a flotation agent for the kieserite [111–113].

Electrostatic Separation of Kieserite. The electrostatic separation process (ESTA process) developed by Kali und Salz AG (Germany) [114] eliminates the formation of the large quantities of NaCl solution produced in the selective dissolution process (see above) [115], [116].

Kieserite can be electrostatically separated in one or two stages. In the one-stage process the raw salt is treated with aliphatic or a mixture of aliphatic and aromatic monocarboxylic acids that are combined with ammonium salts of aliphatic carboxylic acids, usually ammonium acetate. The salt is then heated at a defined relative humidity and electrostatically charged [117]. The kieserite becomes positively charged with respect to the potassium chloride and sodium chloride. Repeated separation of the kieserite deposited at the negative electrode yields $MgSO_4$ in a pure, marketable form.

In the two-stage process rock salt is first separated and then the highly pure kieserite. The raw salt is first treated with aromatic monocarboxylic acids, heated at a defined relative humidity, and then charged [118]. The rock salt becomes positively charged with respect to the

potassium chloride and the kieserite. The rock salt is removed and the KCl – kieserite mixture separated at the positive electrode. The mixture is treated with aliphatic monocarboxylic acids, heated once again at a defined relative humidity, and charged [117]. The kieserite now becomes positive with respect to the potassium chloride. Repeated separation of the kieserite deposited at the negative electrode gives a pure, marketable product.

The potassium chloride fraction is still contaminated with kieserite and sodium chloride; it is used for the production of potassium salts [119], [120].

Production of Granulated Kieserite. After addition of 3 parts of a solution containing preferably 18 wt % Na_2SO_4 or 20 wt % Na_3PO_4 to 47 parts of kieserite, the product can be pressed in a moist state and then ground [121]. After addition of ammonium sulfate and hot magnesium sulfate solution, kieserite can also be granulated with a dish granulator [122]. Mono-, di-, or polysaccharides (usually lactose) are highly suitable additives for preparing these rolled granulates. The kieserite should have a broad particle size range; at least 20 % of the particles must be less than 0.09 mm in diameter [123]. Addition of phosphate allows production of a rolled granulate of high strength even without drying [125], [126].

5.3.2. Production of Epsom Salt

Production of Epsom Salt from Kieserite. Kieserite is dissolved in hot water (90 °C) to give a clear solution. Dissolution is performed in single large vats with capacities of up to 500 m^3 that have a sieve base covered with coconut mats. The process has the disadvantage that the throughput is very low and the sieve plate gradually becomes encrusted with anhydrite and double salts. Kali und Salz AG therefore uses a multichamber cascade with slow stirring; the slurry density in the chambers is 70 wt %. Hot water is added from below and kieserite from above. The solution produced at the top of the cascade contains 400 g/L $MgSO_4$. At the bottom the residue containing anhydrite can be removed at intervals [127].

The hot concentrated solution is cooled in multistage vacuum coolers in which large crystals of Epsom salt are formed. After thickening, the Epsom salt is either filtered or centrifuged. The moist salt is dried in drum driers with air at ca. 50 °C.

Production of Epsom Salt from Magnesia, Magnesite, and Other Materials. Magnesium sulfate solution can be produced by treating seawater magnesia (MgO) or magnesite ($MgCO_3$) with sulfuric acid. Waste sulfuric acid contaminated with organic compounds may be used [128]. The resulting solution is clarified, evaporated to a density of 1.35 – 1.36 kg/L, and processed as described above. The product can be spray dried at 150 °C. Dolomite, olivine, serpentine, or peridotite can also be reacted with sulfuric acid. Magnesium sulfate can also be obtained by adding MgO to pyrite during roasting and subsequent extraction [129].

Waste serpentine from asbestos production can be reacted with ammonium sulfate above 250 °C and the reaction mixture extracted with water [130], [131]. The resulting solution is evaporated and the Epsom salt crystallizes.

Production of Epsom Salt from Boron- and Magnesium-Containing Raw Materials. The borate raw material containing SiO_2, Fe oxide, and Mg is decomposed with H_2SO_4. The solution is separated from the slurry. After crystallization, H_3BO_3 and $MgSO_4$ are separated by flotation [132].

Production of Epsom Salt from Seawater Bitterns or Salt Lake Brines. Pure Epsom salt can be obtained directly from seawater bitterns and from brines from salt lakes with a high $MgSO_4$ content (e.g., Kara – Bogaz Bay) by diluting with water and cooling to – 10 °C [133], [134].

Production of Epsom Salt from Langbeinite. Langbeinite ($K_2SO_4 \cdot 2\,MgSO_4$) is decomposed by treating with water for 6 h at 50 – 60 °C to produce $MgSO_4 \cdot 6\,H_2O$. On cooling the mother liquor to 20 – 35 °C the Epsom salt crystallizes [135].

Production of Epsom Salt from Magnesite or Dolomite and Gypsum. A mixture of ground dolomite and gypsum is treated with steam at

1.5 – 3.0 MPa and filtered to remove the $CaCO_3$ from the $MgSO_4$ solution [136]. Magnesite is calcined at 760 – 835 °C, hydrated in the presence of gypsum at 56 – 60 °C, and carbonized with carbon dioxide. The $MgSO_4$ solution is filtered to remove the $CaCO_3$ [137] and processed to obtain Epsom salt.

5.3.3. Production of Anhydrous Magnesium Sulfate

Anhydrous $MgSO_4$ does not occur in nature and can only be produced by dehydrating $MgSO_4$ hydrates. Kieserite is purified by removing anhydrite and langbeinite in a special electrostatic procedure. It is subsequently washed with water to remove residual chloride [138]; langbeinite and anhydrite may also be removed by flotation (Section 5.3.1). It is then dehydrated by heating at ca. 500 °C in a gas-heated calcination drum and finally cooled [139]. If Epsom salt is used as starting material instead of kieserite, six molecules of water of crystallization can be removed at < 70 °C and the remaining molecule at 500 °C [140].

5.4. Quality Specifications and Analysis

Magnesium sulfate is sold as kieserite (73 – 83 wt % $MgSO_4$), 99.9 % pure Epsom salt (48.8 % $MgSO_4$), and anhydrous $MgSO_4$ (> 98 % $MgSO_4$). Epsom salt is marketed as a chemically pure product, as a clinical product that complies with pharmacopeia specifications, and as an analytical grade reagent. Maximum contamination limits for clinical grades vary according to the pharmacopeia: chloride 100 – 300 mg/kg, iron 20 mg/kg, and heavy metals 10 mg/kg. Maximum limits for arsenic, selenium, and other substances may also be specified.

The analytical procedures used for quality control are the same as those employed for magnesium chloride (see Section 3.5). Sulfate, however, is determined by a gravimetric method [141].

5.5. Transportation and Storage

Depending on the atmospheric humidity, Epsom salt can lose or absorb water during storage and thus tends to harden. It is therefore preferably supplied in Clupack valve sacks. Kieserite is transported in railroad cars and trucks, both in bulk and in valve paper sacks. Anhydrous $MgSO_4$ is also supplied in bulk and in sacks.

5.6. Uses

Large quantities of kieserite and Epsom salt are used in Germany for the production of potassium sulfate (→ Potassium Compounds). Kieserite is used in fertilizers, either as an additive in multinutrient fertilizers or directly to combat magnesium deficiency. In multinutrient fertilizers kieserite aids granulation and improves granulate properties. In cases of acute magnesium deficiency, Epsom salt is an especially effective foliar fertilizer.

Magnesium sulfate is a component of magnesium binders (Sorel cement, see Section 3.7) and can also be used as a binding agent in the production of magnesia bricks. Addition of $MgSO_4$ and $Al_2(SO_4)_3$ leads to the formation of spinels during firing of refractory products. It is also used to produce abrasive wheels. Decomposition of $MgSO_4$ in refractory products at > 1100 °C results in the formation of sulfur dioxide and thus gives a light product. Magnesium sulfate is also used in the production of Portland cement from blast furnace slag and as an additive in gypsum-containing building materials.

Magnesium sulfate is employed in the sugar industry for refining the sugar liquor and for improving the sugar yield from molasses (Quentin procedure) [142]. It is utilized in fermentations for the production of amino acids, antibiotics, bakers' yeast, biomass, and citric acid. It increases the activity of glucose isomerase, the enzyme which converts glucose to fructose.

In the aluminum industry, a mixture of kieserite and calcium oxide or hydroxide is used to precipitate contaminants from aluminate solutions. Magnesium sulfate is also employed in fixative baths used to color aluminum surfaces by means of anodic oxidation. It is a component of the electrolytic baths used for chrome plating, zinc plating, copper plating, nickel plating, and

the electroplating of tungsten, manganese, vanadium, cobalt, and cobalt alloys. It is also used in the phosphating of iron.

In the glass industry, sinters with a defined porosity can be obtained by incorporating $MgSO_4$ of a defined particle size and then leaching it. Magnesium sulfate is used as an additive in enamels and pigments. In the detergent industry it prevents gel formation during spray drying; magnesium silicate synthesized from $MgSO_4$ stabilizes peroxides and thus improves bleaching. Magnesium sulfate is also used as a make-up chemical in the sulfite pulping process and as a stabilizer for cellulose during oxygen bleaching of pulp. In the paper industry it is used in the preparation of paper for coating and in the deinking of recycled printed paper.

Other uses include treatment of waste and wastewater and soil stabilization. The compound is also used as an additive in drilling fluids, lead – acid batteries, oil and coal fuels, ore preparation, "isotonic" beverages, low-NaCl salt mixtures, cosmetics, bath salts, infusions, photographic developers, dye baths, and animal feeds.

In the plastics industry $MgSO_4$ is employed as a coagulant, polymerization catalyst, stabilizer, and flame retardant. In the petrochemical industry it is employed together with manganese chloride for cracking petroleum. Use of $MgSO_4$ in the leather industry makes the leather more supple, increases its weight, and removes calcium carbonate. Magnesium sulfate is also used as a heat storage medium and as a raw material for the production of magnesium aluminate, magnesium alumosilicate, and other magnesium compounds. One of the important applications of Epsom salt is as a laxative.

5.7. Economic Aspects

Epsom salt is produced in many countries, the most important being Germany, Japan, and the United States. It is usually consumed in the country in which it is produced. Only Germany, and the Peoples Republic of China export the compound. Other commercially available products are kieserite, anhydrous $MgSO_4$, and fertilizers containing kieserite (Germany) and langbeinite (USA). Estimated production figures are given in Table 10.

5.8. Toxicology and Occupational Health

Magnesium sulfate and its hydrates are not dangerous to human health when handled in accordance with the appropriate regulations. The LDL_0' (mouse, oral) for $MgSO_4$ is 5000 mg/kg. Magnesium sulfate is not listed in the EEC directive of May 5th, 1976 concerning the discharge of harzardous substances into water; it is classified as a nonhazardous substance in the EEC directive of June 12th, 1987 (WGK:O).

6. Organomagnesium Compounds

Organomagnesium compounds have been of outstanding importance in preparative organic chemistry ever since Victor Grignard (1900) discovered that they are readily prepared from organic halides and elemental magnesium in ethers [143]. In 1912 Grignard received the Nobel Prize for their discovery and for the development of their synthetic potential. As a divalent element, magnesium can bond with one (RMgX, X is usually a halogen, Grignard reagent), or two organic groups (R_2Mg); this bonding does not, however, satisfy the coordination requirements of magnesium (Section 6.1).

The production, structure, and properties of organomagnesium compounds have been summarized in several reviews [144–149]. Industrial use is generally limited to the production of fine chemicals, in which the upper limit is of the order of several tons.

Occasionally, they are used for the synthesis of (polymerization) catalysts [150], [151]. Their use for the production of organic lead derivatives (e.g., tetramethyl lead) is declining.

6.1. Properties

Compounds of the type RMgX or R_2Mg (R = aryl or alkyl) are generally colorless solids unless R is a resonance-stabilized, colored anion. With few exceptions (e.g., *sec*-dibutylmagnesium, dineopentylmagnesium, dicyclopentadienylmagnesium) they are nonvolatile; only the dialkyl compounds where

Table 10. Estimated production and consumption of magnesium sulfate products in 1988 (10^3 t)

Product	Production	Consumption in fertilizers	Consumption in other areas
Anhydrous $MgSO_4$	110	40	70
Kieserite ($MgSO_4 \cdot H_2O$)	2 300	2 250	50
Epsom salt ($MgSO_4 \cdot 7\,H_2O$)	280	180	100
Sulfate of potash magnesia	1 000	1 000	
Total	3 690	3 470	220

$R > C_4H_9$ are soluble in hydrocarbons, especially aromatic hydrocarbons. Solubility, however, increases considerably in the presence of one or two equivalents of a coordinating ether, especially if intramolecular coordination occurs (i.e., by an ether function incorporated in the substituent R).

The structures of organomagnesium compounds in the solid state are being increasingly documented by crystal structure analysis. In the great majority of compounds, magnesium has a coordination number of four, which is generally attained by coordination with two Lewis bases, usually ethers (e.g., $RMgX \cdot 2\,Et_2O$); in the absence of a base, association takes place. The coordination sphere around the tetracoordinated magnesium is a distorted tetrahedron [145]. Association takes place by means of three-center, two-electron bonds, leading to the formation of dimeric or polymeric structures [152], [153]:

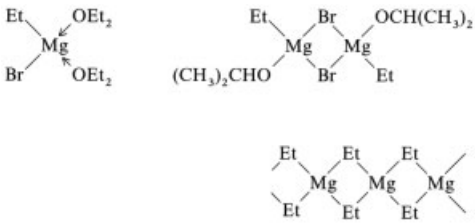

Organomagnesium compounds are thermally stable up to 100 – 170 °C, but at higher temperatures they can decompose explosively. They are extremely sensitive in air (i.e., in the presence of oxygen) and particularly toward water; they can sometimes ignite spontaneously. In ethereal solution the relatively heavy ether vapor protects them to some extent from exposure to air and thereby against decomposition. For small amounts or very pure preparations, an inert atmosphere of nitrogen or, preferably, argon, is recommended; in extreme cases an evacuated system should be employed.

For preparative purposes organomagnesium compounds are not isolated as such, but are used directly in solution following their preparation. Their structure in solution is basically the same as that in the crystal; nevertheless, a fundamental difference is that an equilibrium can exist between the various possible species, the position of which depends on the organic group R, the halogen X, the solvent Y, the concentration, and the temperature:

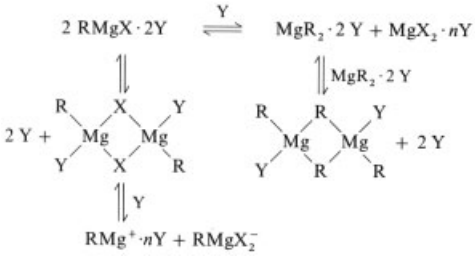

The most important of these equilibria is that between RMgX and R_2Mg, the so-called Schlenk equilibrium. It is strongly dependent on the solvent and lies, for example in diethyl ether, well to the left-hand side, whereas in tetrahydrofuran an almost statistical distribution exists (i.e., $R_2Mg : MgX_2 : 2\,RMgX = 1 : 1 : 2$).

Organomagnesium solutions can be quantitatively analyzed by various methods [154]. In addition to qualitative color tests [155], titrations can be used in which the active carbon–magnesium bond is either titrated directly with *sec*-butanol in the presence of a color indicator [156] or indirectly with hydrochloric acid after hydrolysis to give the free base, if necessary together with complexometric titration of the magnesium. Less convenient, but nevertheless reliable, is the formation and quantitative

determination of derivatives, e.g., with carbon dioxide or chlorotrimethylstannane.

$$RCOOH \xleftarrow{CO_2, H^+} RMgX \xrightarrow{ClSn(CH_3)_3} RSn(CH_3)_3$$

Apart from their instability in air and their caustic basic properties, organomagnesium compounds and the precipitated inorganic magnesium salts produced during their use display no pronounced toxicity or side effects of environmental concern. The pyrophoric properties of the compounds and the high flammability of the ethereal solvents demand the usual protective measures.

6.2. Production

In practice the synthesis of Grignard reagents RMgX (X = halide) from organic halides RX and metallic magnesium is by far the most important preparative method. The Grignard reagents are used directly or indirectly as starting materials for the production of other (organometallic) compounds and also for the diorganylmagnesium compounds R_2Mg.

6.2.1. Grignard Reagents

The production of solutions of Grignard reagents (RMgX) according to the equation

$$R-X + Mg \xrightarrow{2\,Y} R-Mg-X \cdot 2\,Y$$

R = alkyl or aryl group
X = halogen
Y = solvent (usually diethyl ether or tetrahydrofuran)

presents few problems if suitable protective measures are observed [145], [157]. The fluorides react very sluggishly; the iodides are very reactive and side reactions, particularly Wurtz coupling, occur frequently. Side reactions also occur to a small extent with the chlorides and the bromides:

$$2\,R-X + Mg \longrightarrow R-R + MgX_2$$

Organic bromides are most suitable for the production of Grignard reagents, but are frequently replaced in industrial applications by the cheaper chlorides.

The initiation of the reaction leading to formation of the Grignard reagent presents a considerable problem; it can be delayed or even fail as a result of reagent impurity or the low reactivity of the organic halide. Various strategies have been developed to overcome these difficulties [157], of which only a few will be mentioned here. Impurities causing delay in initiation can often be eliminated by the addition of a small amount of RMgX, possibly from an earlier batch. Most methods depend, however, on the activation of the magnesium, which is normally used in the form of turnings. Mechanical activation can be achieved by vigorous stirring or with ultrasound [158], chemical activation with numerous reagents (e.g., iodine, bromoethane, 1,2-dibromoethane, mercuric chloride, amalgam). A highly reactive, finely divided magnesium suspension can be obtained by (co-)sublimation of the metal (with the substrate), by reduction of magnesium chloride with potassium [159], or by heating anthracene – magnesium in tetrahydrofuran [160].

For reasons of cost and safety (inflammability, peroxide formation), the use of hydrocarbon solvents instead of the usual ethers is of interest. The synthesis of Grignard reagents in hydrocarbons (e.g., methylcyclohexane [150]) or an excess of the organic halide [161] unfortunately requires more severe conditions and leads to halogen-deficient products, whose reactivity often differs from that of the Grignard reagents synthesized in an ether solvent. Furthermore, solubility is also poor, so that suspensions must be used; however, the higher aliphatic halides (especially the iodides) give soluble organomagnesium compounds [162].

A relatively recent development is the synthesis of di-Grignard reagents, particularly those with only a few (1 – 3) carbon atoms between the two functional groups [163], [164]; they are of interest mainly in organometallic chemistry. Examples are the 1,1- or 1,3-di-Grignard reagents:

$$Br(CH_2)_nBr + Mg \xrightarrow[Mg]{n = 1, 3} BrMg(CH_2)_nMgBr$$
$$\downarrow n = 2$$
$$MgBr_2 + CH_2{=}CH_2$$

Even a tri-Grignard reagent has been produced (in 44 % yield) [165]:

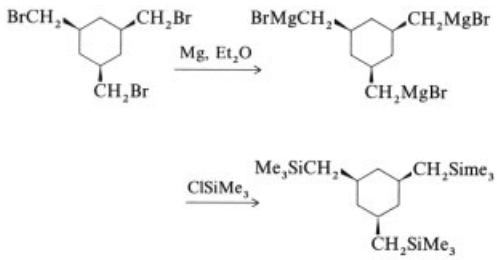

Finally, mention should be made of the Barbier reaction, i.e., the reaction of an organic halide with magnesium in the presence of the substrate (e.g., a ketone that is to be alkylated). This reaction was discovered by GRIGNARD's professor BARBIER and is the predecessor of the Grignard reaction. The Barbier reaction gives better yields when the Grignard reagents can only be isolated in poor yields, for example, with many allyl, benzyl-, or polyhalides [157], [158], [166], [167]: The method has been adapted for performance on a semi-industrial scale [168].

Other preparative methods have limited application and are only of special interest, e.g., the metallation of acidic hydrocarbons, metathesis with organolithium or organosodium compounds, metal exchange reaction with organomercury compounds, addition to unsaturated systems, or comproportionation of MgR_2 and $MgBr_2$ by the Schlenk equilibrium (in diethyl ether):

$$RMgX + R'C{\equiv}CH \longrightarrow R'C{\equiv}CMgX + RH$$

$$2\ \text{C}_5\text{H}_5 + Mg \longrightarrow (C_5H_5)_2Mg + H_2$$

$$RLi + MgBr_2 \longrightarrow RMgBr + LiBr$$

$$RHgBr + Mg \longrightarrow RMgBr + Hg$$

$$(CH_3)_3CMgCl + CH_2{=}CH_2 \longrightarrow (CH_3)_3CCH_2CH_2MgCl$$

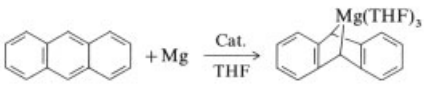

$$2\ CH_2{=}CH_2 + MgH_2 \xrightarrow[\text{THF}]{\text{Cat.}} Mg(CH_2CH_3)_2$$

$$MgR_2 + MgBr_2 \longrightarrow 2\ RMgBr$$

6.2.2. Diorganylmagnesium Compounds

Dialkyl- and diarylmagnesium compounds (R_2Mg) are generally of only limited practical importance. They are most frequently obtained from Grignard reagents by displacing the Schlenk equilibrium (see Section 6.1). In hydrocarbon solvents, the equilibrium already lies to the right. In diethyl ether, the equilibrium can be shifted by the addition of dioxane and precipitation of poorly soluble magnesium bromide–dioxanate [144], [145].

$$2\ RMgBr \underset{\times}{\overset{\text{Hydrocarbons}}{\rightleftharpoons}} R_2Mg + MgBr_2 \downarrow$$

$$2\ RMgBr \underset{\times}{\overset{\text{Dioxane}}{\rightleftharpoons}} R_2Mg + MgBr_2 \cdot (C_4H_8O_2)_2 \downarrow$$

Very pure R_2Mg can be obtained by shaking diorganylmercury compounds for several days at room temperature in ether with an excess of magnesium.

6.3. Uses

Organomagnesium compounds, especially Grignard reagents, have a wide range of application on account of their easy accessibility and their high reactivity. In organic synthesis they serve as nucleophiles for formation of carbon–carbon bonds. In organometallic chemistry they are used for the nucleophilic alkylation or arylation of metal salts; this reaction is suitable for the majority of main group and transition metals with a greater electronegativity than magnesium. In both areas the more reactive organolithium compounds are competing with the Grignard reagents. However, since the reactivity of Grignard reagents is usually high enough, they have lost little in popularity. This is also due to their ease of manipulation and their lower price, especially in industrial quantities.

The broad spectrum of reactions of Grignard reagents toward C=O, C=N, C≡N, C−X, C−H, and M−X groups can be illustrated by the following examples [144], [145], [147].

$$
\begin{aligned}
RMgX + R'CH{=}O &\rightarrow RR'CHOH \\
+ R'_2C{=}O &\rightarrow RR'_2CHOH \\
+ R'COOCH_2CH_3 &\rightarrow R_2R'COH \\
+ CO_2 &\rightarrow RCOOH \\
+ R'_2C{=}NR'' &\rightarrow RR'_2C{-}NHR'' \\
+ R'C{\equiv}N &\xrightarrow{\text{Cat.}} RR'C{=}O \\
+ R'X &\rightarrow R{-}R' \\
+ R'H &\rightarrow R{-}H + R'MgX \\
+ MCl_n &\rightarrow MR_n
\end{aligned}
$$

(M = metal from group 3–15)

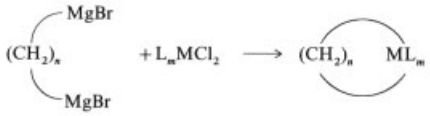

L = substituent or ligand ($n \geq 3$)

7. References

1. H. L. Pattinson, GB 9102, 1841.
2. W. Wetzenstein: "Limnische Huntit-Hydromagnesit-Magnesitlagerstätten in Macedonien/Nordgriechenland," *Miner. Deposita* **10** (1975) 129 – 140.
3. B. M. Coope: "Magnesia Markets. Fit for Optimism," *Ind. Miner. (London),* Sept. 1989, 45 – 57.

General References

4. *Gmelin* **27** A, 57 ff.
5. H. Ginsberg, K. Wefers, *Die Metallischen Rohstoffe,* vol. 15: Aluminium und Magnesium, F. Enke Verlag, Stuttgart 1971.

Specific References

6. Dead Sea Works, DE-OS 2 532 022, 1975 (J. A. Epstein et al.).
7. B. Breiter, J. Steußloff, *Neue Bergbautech.* **16** (1986) no. 1, 23 – 27. Standard-Messo Verfahrenstechnik GmbH, DD 229 678, 1984 (R. Schmitz, K. Wamser, H. Zahalka).
8. C. Dähne, B. Maurer, *Neue Bergbautech.* **15** (1985) no. 8, 291 – 294.
9. Bechtel Internat. Corp., US 230 986, 1972; US 347 222, 1973. Kaliforschungsinstitut Sondershausen, DE-OS 1 467 293, 1964 (H. Hoppe); DE-AS 1 198 802, 1964 (H. Hoppe).
10. J. Kenat, *Symp. Salt (Proc.)* **2nd, vol. 2,** 1965, 195 – 203.
11. Israel Chemicals Ltd., DE-OS 2 746 503, 1977 (A. Sadan, H. Pitnach).
12. E. E. Frolovskii, N. A. Alekseeva, L. B. Anikina, L. E. Bosiyak, *Khim. Promst. Ser. Kaliinaya Promst.* 1979, no. 3, 21 – 24; *Chem. Abstr.* **92** (1980) 78 938 f.
13. S. F. Estefan, *Hydrometallurgy* **10** (1983) 39 – 45.
14. Great Salt Lake Minerals and Chem. Corp., US 3 690 844, 1972 (U. Neitzel, H. G. Flint); P. Behrens, *Bull. Utah Geol. Miner. Surv.* **116** (1980) 223 –228.
15. Salzdetfurth AG, US 3 642 455, 1972 (B. J. Hahn, R. M. E. Reise, W. R. Raschka).
16. L. van den Assen, D. Buyze, H. P. Rogaar, *Rev. Metall. CIT,* April 1985, 295 – 305.
17. Kali und Salz, DE 2 513 947, 1975 (G. Budan, D. Kunze, W. Otto, I. Steingart).
18. H. Schnabel et al., DD 71 758, 1968.
19. Mines de Potasse d'Alsace, FR 8 104 391, 1981 (M. Bichara, M. Bodu, J. P. Koensgen, M. Meriaux, J.-P. Zimmermann).
20. E. S. Pereira, H. G. Riella: "Production of Pure Magnesium Chloride and Magnesium Oxide from Magnesite", *An. Assoc. Bras. Quim.* **44** (1995) no. 2, 15 – 18.
21. O. V. Lebedev: "Production of Magnesium Chloride", *Khim. Prom-st (Moscow) 1991* no. 2, 126; ISSN 0023-110X.
22. Aluminum Company of America, US 4 943 426, 1990 (L. E. Dastolfo, Jr., V. Cedro, III).
23. Kali und Salz, BASF, DE 3 119 968, 1981 (K. Wintermantel, D. Stockburger, A. Hollstein, D. Kunze, F. Werdelmann).
24. VEB Kombinat Kali, DD 211 105, 1982 (H. Scherzberg, W. Ulrich, R. Tober, W. Staufenbiel, M. Matern).
25. Mines de Potasse d'Alsace, DE 2 751 420, 1977 (A. Aubry, M. Bichara).
26. Nalco Chemical Company, DE-OS 3 010 755, 1980 (R. Allain, D. Braithwaite, J. P. Maniscalco). R. J. Allain, D. Braithwaite, *Proc. Annu. World Conf., Magnesium,* June 1979, 24 – 28.
27. Dow Chemical, US 3 282 642, 1966 (R. D. Goodenough, R. A. Gaska).
28. Kaiser Aluminium & Chem. Corp., US 3 181 930, 1931 (O. C. Olsen).
29. ESSO Res. & Eng. Co., US 3 347 625, 1965 (C. N. Kimberlin, Jr., W. F. Arey. Jr., F. J. Buchmann).
30. M. Nakayasu, Y. Suzukawa, W. Kobayashi: "Thermal Decomposition of $MgCl_2 \cdot NH_4Cl \cdot nNH_3$," *Denki Kagaku oyobi Kogyo Butsuri Kagaku* **51** (1983) no. 4, 396 – 402. *Chem. Abstr.* **99** (1983) 81 676 m.

31. J. T. May, V. E. Edlund, D. C. Seidel: "Dehydrating Magnesium Chloride by Double-Salt Decomposition". *Rep. Invest. U.S. Bur. Mines,* **RI 3277** (1978); *Chem. Abstr.* **89** (1978), 45 748 d.
32. Ruthner Engineering GmbH, DE-OS 2 638 123, 1976 (J. Jeney).
33. Ustav pro vyzkum rud, DE-OS 3 415 784, 1984 (V. Horák, V. Bumbálek, K. Matouš, F. Sehnálek).
34. Veitscher Magnesitwerke-AG, DE-OS 3 045 796, 1980 (M. Grill, H. Grohmann, E. Klein, A. Kullig).
35. *Kirk-Othmer,* 2nd ed., **vol. 12,** p. 672.
36. EEC Directive L 213/81 – 84, Method 5.1, Association of Official Analytical Chemists-Methods, 1984, no. 2.168, p. 30.
37. Association of Official Analytical Chemists-Methods (AOAC), 14th ed., 1984 no. 2.126 – 2.130 (AAS) and 2.141.
38. Association of Official Analytical Chemists-Methods (AOAC), 1984, Method no. 2.108 – 2.113.
39. Association of Official Analytical Chemists-Methods (AOAC), 1984, Method no. 2.173 – 2.176; Verband Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (Lufa): *Handbuch der landwirtschaftlichen Versuchs- und Untersuchungsmethodik (Methodenbuch)*, vol. 2: *Die Untersuchung von Düngemitteln,* Verlag J. Neumann-Neudamm, Melsungen 1973, 9_{62}– 9_{64}, Method 9.18.2.
40. J. Fries, H. Getrost: *Organische Reagenzien für die Spurenanalyse,* 3rd. ed., E. Merck, Darmstadt, Verlag Chemie, Weinheim 1966, p. 331 ff.
41. Lufa-Methodenbuch 923, Method 9.8.1.
42. Lufa-Methodenbuch 991 – 993, Method 9.28.2.
43. DIN 38405, part 17; H. Baron, *Fresenius Z. Anal. Chem.* **143** (1954) 339.
44. M. Zimmermann: *Fotometrische Metall- und Wasseranalysen,* Wiss. Verlagsanstalt, Stuttgart 1954.
45. E. Merck: *Organische Reagenzien für die Spurenanalyse,* 3rd ed., Verlag Chemie, Weinheim 1966, pp. 69 – 72.
46. Association of Official Analytical Chemists-Methods, Method no. 2.157 – 2.160, pp. 29/30.
47. R. Mathur, A. K. Misra, M. P. S. Chandrawa, *Res. Ind.* **31** (1986) June, 181 – 184. Kali und Salz, DE 3 230 962, 1982 (A. Singewald, K.-R. Löblich, H. Duyster). Kali und Salz, DE-OS 3 031 086, 1980 (H. Duyster, W. Prinzler, G. Nürnberger, K. Springer).
48. Saarbergwerke, DE-OS 2 204 281, 1972 (W. Schuhknecht, H. Kunz, H. G. Klinkner, G. Culmann); DE-OS 2 405 223, 1974 (H. G. Klinkner, G. Culmann).
49. Norsk Proco A/S, EP-A 0 241 103, 1987 (J. G. Berg, R. E. Smith-Johannsen).
50. M. C. Bell, *Feedstuffs* **50** (1978) no. 6, 24 – 25.
51. W. Pannekeet, *Chem. Ind. (London)* **21** (1983) 821 –824.
52. G. Gascho, *J. Fert. Issues* **3** (1986) no. 2, 62 – 65. Kali und Salz, DE-OS 3 432 327, 1984 (R. Wotschke, G. Budan, G. Kemmler).

General References

53. L. R. Duncan: "Today's Magnesia Industry," paper given at International Symposium on Refractories, Hangzhou, 1988.
54. E. Naujokat, D. v. Mallinckrodt, C. Zografou, *Keram. Zeitschr.* **35** (1983) 440 – 442.

Specific References

55. W. Zednicek: "Periklas. Ein bemerkenswertes und wesentliches Mineral in der Feuerfest-Keramik," *Radex Rundsch.* 1985, no. 4, 651 – 695.
56. V. Weiß: "Die Aufbereitung von Magnesit" in: *Handbuch der Keramik,* Verlag Schmid, Freiburg 1968.
57. H. Schmid: "Turkey's Salda Lake. A Genetic Model for Australias's Newly Discovered Magnesite Deposits," *Ind. Miner.* (*London)* August 1987, 19 – 31.
58. W. Wetzenstein: "Limnische Huntit-Hydromagnesit-Magnesitlagerstätten in Macedonien/Nordgriechenland," *Miner. Deposita* **10** (1975) 129 – 140.
59. W. Wetzenstein, D. Zachmann: "Sedimentäre magnesium-karbonatische Bildungen im Servia-Becken/Nordgriechenland," *Radex Rundsch.* 1977, no. 1, 29 – 49.
60. M. Coope: "Magnesia Markets. Fit for Optimism," *Ind. Miner. (London)* Sept. 1989, 45 – 57.
61. I. H. Howard-Smith: "Kunwarara Magnesite Deposit, Australia," Information Memorandum, Queensland Metals Corp., Toowong (Brisbane), Queensland, Australia, June 1989.
62. W. C. Gilpin, N. Heasman, *Chem. Ind. (London),* 1977, 567 – 572.
63. J. C. Drum, S. Tangney, *Trans. J. Br. Ceramic Soc.* **77** (1978) no. 4, 10 – 14.

64. T. F. W. Barth, C. W. Correns, P. Eskola: *Die Entstehung der Gesteine,* Springer Verlag, Berlin 1960.
65. L. van den Assem: "Planning for New Industrial Minerals Projects – Magnesium Oxide from Brine in the Netherlands," *Ind. Miner. (London),* January 1982, 35 – 42.
66. I. A. Epstein: "Utilization of the Dead Sea Minerals (a review)," *Hydrometallurgy* **2** (1976) 1 – 10.
67. H. Jedlicka: "Neue Anwendungen des Sprühröstverfahrens in der Chlorid-Naßmetallurgie," Schriftenreihe über Entwicklungen und Verfahren der Ruthner Industrieanlagen AG, Wien 1977, lecture Leningrad 1976.
68. H. K. Krivanec, W. F. Kladnig: "Herstellung von Oxidrohstoffen nach dem Andritz-Ruthner-Sprühröstverfahren," *TIZ* **112** (1988) no. 11, 762 – 769.
69. W. S. Ainscow: "Aufbereitung von Magnesit zu hochwertiger Sintermagnesia," *TIZ* **110** (1986) no. 6, 363 – 368. *Sulmag II the Sinter Magnesite Process,* Sulzer Brothers Ltd., Winterthur, Switzerland.
70. H. G. Schulte: "Polcal – An Economical Process for Calcining Magnesium Hydrate," *Ind. Miner. (London),* November 1982, 37 – 47.
71. H. Priemer, G. Mörtl: "High Temperature Shaft Kiln for Large Crystal Periclase Materials," *Proceedings of International Symposium on Refractories,* Hangzhou 1988, 167 – 178.
72. Z. Foroglou et al.: "Refractory Raw Materials of Natural Origin and High Performance Refractories Based on such Products," *Proceedings of International Symposium on Refractories,* Hangzhou 1988, 204 – 213.
73. A. Eschner: "Die wärmetechnischen Eigenschaften von Magnesiasteinen für Niedertarifspeicherungen," Dissertation Fakultät für Bergbau, Hüttenwesen und Maschinenwesen der Techn. Universität Clausthal 1978.
74. "Magnesium Oxide Makes a Comeback," *Chem. Week,* January 1981, 43.
75. B. M. Coope: "Caustic Magnesia Markets – Agricultural Oversupply and Industrial Underdemand," *Ind. Miner. (London)* February 1981, 43 – 51.
76. I. Watson: "Minerals in Animal Feedstuffs – Plenty of Food for Thought," *Ind. Miner. (London)* April 1982, 71 – 91.
77. "Why Apply Magnesium with Potash?" *Phosphorus Potassium* **161** (1989) May – June, 20 – 28.
78. R. Schulze-Rettmer, T. Yawari: "Versuche mit dem Verfahren der Fällung von Magnesium-Ammonium-Phosphat (MAP) aus verschiedenen Abwässern," *Vom Wasser* **71** (1988) 41 – 54.
79. "Flammschutzmittel und Füllstoffe auf Basis hochreiner Magnesiumverbindungen," *Kunststoffe,* June 1988. *Japan Chemical Week,* 26th March 1987.
80. "World Magnesia Shortage, Particularly in High Purity Grades," *Ind. Miner. (London)* May 1974, 23 –32.
81. "Magnesia – from Shortage to Surplus," *Ind. Miner. (London)* Sept. 1977, 31 – 55.
82. B. M. Coope: "Magnesia Markets – Refractory Contraction and Caustic Stagnation," *Ind. Miner. (London)* August 1983, 57 – 87.
83. H. Schmid: "China – the Magnesite Giant," *Ind. Miner. (London)* August 1984, 27 – 45.
84. B. M. Coope: "The World Magnesia Industry, Smaller but Fitter and Purer," *Ind. Miner. (London)* February 1987, 21 – 48.
85. R. C. Mackenzie: *The Differential Thermal Investigation of Clays,* Mineralogical Society (Clay Minerals Group), London 1957.
86. B. Brezny, *Ber. Dtsch. Keram. Ges.* **42** (1965) 308 –310.
87. I. S. Raeva, *Ogneupory* **24** (1983) no. 4, 31 – 34.
88. G. V. Samsonov (ed.): *The Oxide Handbook,* IFI/Plenum, Data Corporation, New York–Washington–London 1973, p. 185.
89. P. D. Johnson, *J. Am. Ceram. Soc.* **33** (1950) no. 5, 168 – 171.
90. S. P. Mitoff, *J. Chem. Phys.* **31** (1959) 1261 – 1269.
91. A. Cermak, *Silikattechnik* **18** (1967) no. 7, 219 – 223.
92. G. Schönfelder, *Elektrowärme Int. Ed. B* **41** (1983) B 5 214 – 221; Norton Co., DE-OS 2 333 601, 1973 (J. J. Scott).d
93. Norton Co., US 3 410 666, 1968 (J. J. Scott, N. C. Turnbull).
94. General Electric, DE-OS 1 592 116, 1966 (J. W. Rutter, E. M. Clausen).
95. Norton Co., US 2 280 515, 1939 (R. R. Ridgeway); Dynamit Nobel, DE 1 249 145, 1965 (H. R. Müller).
96. R. Czepek, *Elektrowärme Int.* **27** (1969) no. 12, 473 –481.
97. Dynamit Nobel, DE-OS 1 592 089, 1967 (P. Hack, M. Neidhardt, H. Warmann).
98. General Electric, US 3 592 771, 1968 (W. Vedder, J. Schultz).

99. Norton Co., US 3 457 092, 1966 (R. O. Tervo).
100. Dynamit Nobel, DE 1 921 789, 1969 (M. Neidhardt, J. Schneider).
101. Dynamit Nobel, DE 2 363 790, 1973 (H. Clasen, K. Deneke).
102. Dynamit Nobel, DE 2 525 441, 1975 (G. E. Bockstiegel, M. Neidhardt, G. Rehfeld).
103. P. Bartha, F. Metz, K. H. Nitsch, *Stahl Eisen* **107** (1987) no. 13, 639 – 643.
104. A. Rabenau, *Chem. Ing. Techn.* **36** (1964) 542 – 545.
105. T. Power in E. M. Dickson (ed.): *Raw Materials for the Refractories Industry:* 2nd ed., Industrial Minerals, London 1986, pp. 163 – 177.
106. D'ANS-LAX: *Taschenbuch für Chemiker und Physiker,* **vol. 1,** Springer Verlag, Berlin – Heidelberg –New York 1967, pp. 417 – 418.
107. *Kirk-Othmer,* **14,** pp. 636 – 639.
108. "Features and Genesis of the Dalangtan Salt Deposit in the Qaidam Basin," *Z. Geol. Wiss.* **18** (1990) no. 9, 813 – 824. *Chem. Abstr.* 114 : 126079y.
109. Kali und Salz, DE 3 223 366, 1982 (U. Neitzel, H. Schroth, A. Hollstein, A. Singewald).
110. DD 25 546, 1958 (S. Mildner, L. Döhler).
111. VEB Kombinat Kali, DD 220 237, 1983 (L. Herrmann, H. Baldauf, W. Kramer, H. Schubert, F. Aurich).
112. Wintershall, DE 1 064 891, 1957 (G. Budan).
113. Duval Corp., US 4 045 335, 1977 (B. F. Adams, J. Jr. Edward).
114. Kali und Steinsalz, Bd. 12, Heft 2, Sept. 1996.
115. Kali und Salz, DE 3 637 226, 1986 (A. Singewald, U. Neitzel, G. Fricke).
116. Kali und Salz, DE 3 637 227, 1986 (A. Singewald, G. Fricke, H. Schroth).
117. Kali und Salz, DE 1 667 814, 1968 (G. Fricke, A. Singewald).
118. Kali und Salz, DE 1 792 120 (A. Singewald, G. Fricke).
119. G. Fricke, *Kali Steinsalz* **7** (1979) no. 12, 492.
120. A. Singewald, U. Neitzel, *Potash Technol. Min. Process. Maint. Transp. Occup. Health Saf. Environ.* [Int. Potash Technol. Conf. 1st] (1983) 589 – 595.
121. Kali und Salz, DE 3 148 404, 1981 (H. Zentgraf, H. Fetzer, A. Hollstein, G. Fricke).
122. Kali und Salz, DE 2 748 152, 1977 (G. Bruns).
123. Kali und Salz, DE 3 618 058, 1986 (K.-R. Löblich, G. Bruns, G. Peuschel).
124. Kali und Salz, DE 3 707 785, 1987 (K.-R. Löblich, G. Bruns, H. Zentgraf, E. Czaplinsky).
125. Kali und Salz, DE-PS 4 303 984 Cl, 1994 (K. R. Löblich, M. Schroeder, F. Schrumpf, R. Werthmann).
126. Kali und Salz, DE-PS 4 303 985 Cl, 1994 (N. Knoepfel, F. Schrumpf, T. Fleischer, R. Werthmann, R. Schad).
127. Kali und Salz, DE 2 512 246, 1975 (E. Höfling, H. Eberle, A. Hollstein, B. Hahn).
128. Wigton-Abbot Corp., US 2 754 175, 1956 (F. J. Hendel); *Chem. Abstr.* **77** (1972) 22 391 v; JP 7 116 245, 1971. *Chem. Abstr.* **89** (1978) 77 069 z; B. Mikhailov et al., *Khim. Ind. (Sofia)* 1978, 50(3) 105 – 6.
129. Gypsum Industries Ltd., ZA 6 700 450, 1968 (N. Rudolf).
130. Société Nationale de l'Amiante, Thetford Mines, DE-OS 3 006 050, 1980 (J.-M. Lalancette); US 4 277 449, 1979.
131. *Chem. Abstr.* **95** (1989) 222 201 s; P. Bozadzhiev et al., *God. Vissh. Khim-Tekhnol. Inst., Sofia,* **24** (1978) no. 2, 19 – 24.
132. SU 1 629 245 Al, 1991 (L. F. Kardashina et al.).
133. J. A. Fernandez-Lozano: "Recovery of Epsomite and Sylvite from Seawater Bittern by Crystallisation," *Fourth International Symposium on Salt,* Northern Ohio Geological Society,vol. 2, 1974, pp. 501 – 510.
134. *Chem. Abstr.* **85** (1976) 162 612 h.
135. IMCC, US 2 733 132, 1956. Sincat, FR 1 363 066, 1964.
136. *Chem. Abstr.* **89** (1978) 26 924 s; RO 59 695, 1971 (S. Kiss et al.).
137. SU 1 127 846, 1983 (N. P. Balandin); *Chem. Abstr.* **102** (1985) 134 380 g.
138. Kali und Salz, DE 3 334 665, 1983 (G. Fricke, I. Geisler).
139. R. Burmeister, *Bergakademie* **16** (1964) no. 9, 558 – 559.
140. "Verfahrenstechnische Alternativen bei der Kalzinierung von Kieserit unter Berücksichtigung der Optimierung des Energieeinsatzes," Techn. Universität Clausthal 1993, 101 S.
141. Association of Official Analytical Chemists-Methods (AOAC), Method no. 2.182 – 2.183, 1984, p. 32.
142. J. F. T. Oldfield et al., *Int. Sugar J.* **82** (1979) 138 –143.
143. V. Grignard, *C. R. Hebd. Seances Acad. Sci.* **130** (1900) 1322.
144. K. Nützel in *Houben-Weyl,* **vol. 13/2 a,** pp. 47, 521.
145. W. E. Lindsell in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. 1,** Pergamon Press, Oxford 1982, p. 155.

146. W. E. Lindsell in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* **vol. II,** Pergamon, Oxford 1995, p. 57.
147. Ch. Elschenbroich, A. Salzer: *Organometallics,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1992, p. 40.
148. G. S. Silverman, P. E. Rakita (eds.): *Handbook of Grignard Reagents,* Marcel Dekker, New York 1996.
149. H. G. Richey (ed.): *Grignard Reagents,* John Wiley & Sons, Chichester 1999.
150. D. Bryce-Smith, E. T. Blues, *Org. Synth.* **47** (1967) 113.
151. Monsanto, US 4 438 034, 1984 (A. Y. Garner); *Chem. Abstr.* **101** (1984) 24 119 a.
152. A. L. Spek, P. Voorbergen, G. Schat, C. Blomberg, F. Bickelhaupt, *J. Organomet. Chem.* **77** (1974) 147.
153. E. Weiss, *J. Organomet. Chem.* **4** (1965) 101.
154. C. Blomberg: "Magnesium" in F. Korte (ed.): *Methodicum Chimicum,* vol. 7, Thieme Verlag, Stuttgart 1976, p. 79.
155. H. Gilman, F. Schulze, *J. Am. Chem. Soc.* **47** (1925) 2002.
156. S. C. Watson, J. F. Eastham, *J. Organomet. Chem.* **9** (1967) 165.
157. F. Bickelhaupt in A. P. Hagen (ed.): *Inorganic Reactions and Methods,* **vol. 10,** VCH Publishers, New York 1989, p. 237.
158. C. Einhorn, J. Einhorn, J. L. Luche, *Synthesis* 1989, 787.
159. R. D. Rieke, S. E. Bales, P. M. Hudnall, G. S. Poindexter, *Org. Synth.* **59** (1980) 85.
160. B. Bogdanović, *Angew. Chem. Int. Ed. Engl.* **25** (1985) 262.
161. Oxon Italia S. p. A., DE-OS 2 625 216, 1976; *Chem. Abstr.* **86** (1977) 140 234.
162. W. N. Smith, Jr., *J. Organomet. Chem.* **64** (1974) 25.
163. F. Bickelhaupt, *Angew. Chem.* **99** (1987) 1020.
164. F. Bickelhaupt, *Chem. Soc. Rev.* **28** (1999) 17, and references cited therein.
165. Ph. Boudjouk, R. Sooriyakumaran, C. A. Kapfer, *J. Organomet. Chem.* **281** (1985) C 21.
166. C. Blomberg, F. A. Hartog, *Synthesis* 1977, 18.
167. C. Blomberg: "The Barbier Reaction and Related One-Step Processes," in K. Hafner, J. M. Lehn, C. W. Rees, P. von Ragué Schleyer, B. M. Trost, R. Zahradnik (eds.): *Reactivity and Structure Concepts in Organic Chemistry,* **vol. 31,** Springer-Verlag, Berlin Heidelberg 1993.
168. B. A. Klokov, *Main Group Metal Chemistry* **22** (1999) 1.

Magnetic Information Storage → Information Storage Materials

Magnetic Materials

Wilhelm Ervens, Krupp Widia, Essen, Federal Republic of Germany

Heinrich Wilmesmeier, Krupp Widia, Essen, Federal Republic of Germany

1. Introduction

Magnetism is a fundamental property of matter. All substances exhibit some form of magnetic behavior, which has its ultimate source in the magnetism of atomic nuclei and electrons.

The magnetism of materials varies in nature and strength depending on their composition and state of order. The principal state of order in industrial magnetic materials is that of ferromagnetism, which (along with some variations) is discussed in Section 2.3.

Industrial magnetic materials are classified as soft or hard in accordance with how they behave in the presence of external fields. Hard magnetic materials are also called permanent magnets.

Soft magnetic materials find use where the magnetic effects associated with either permanent magnets or electromagnetic fields must be concentrated or shielded, or where low-loss transformation, exact transmission, modulation, or regulation of electrical signals is needed.

The practical importance of hard magnetic materials stems from their ability, once magnetized, to act on other ferromagnetic materials without contact, attracting or, in the case of other permanent magnets, repelling them. With no energy input and therefore at no operating cost, permanent magnets provide a magnetic flux or a magnetic field.

Magnetic materials are found in nearly every area of daily life. They perform coupling, modulating, or regulating functions in equipment and devices based on electromagnetic principles (e.g., motors, generators, loudspeakers, transformers) and also in magnetomechanical systems (e.g., clutches, separators). Magnetic materials are indispensable in office equipment, computer hardware, automobiles, entertainment electronics, communications, household appliances, and medical instruments.

With the rapid development of high-performance semiconductor components and integrated electronic circuits, the magnetic components and inductive devices used in electrical equipment must also satisfy more and more stringent requirements, chiefly that of smaller size with equal or enhanced efficiency. As a consequence, hard magnetic materials must have higher energy contents (see Section 3.2), while soft magnetic materials require improved magnetic conductances (see Section 3.3). Complicated shapes and close tolerances must also be realized, and stringent standards of thermal stability and corrosion resistance have to be satisfied.

Recently, major advances have been made in both hard and soft magnetic materials, due to newly developed materials and further development of existing ones. Important hard magnetic materials include primarily alloys containing rare earth metals; these are described in Section 4.2. Rapidly quenched amorphous and nanocrystalline materials and new plastic-bonded powder composites represent significant innovations in soft magnetic materials (see Section 5.2).

Thus, together with the classical hard and soft magnetic materials, a wide range of magnetic materials is available for various applications whose importance, both technical and economic, is steadily increasing.

2. Forms of Magnetism in Matter

The magnetic phenomena exhibited by matter depend on the electronic configuration of the atom, the crystalline and molecular structure, and the resultant coupling between electrons.

Magnetism is an inherent property of materials, arising from the orbital and spin magnetic moments of the nucleus and the electrons. The dominant contribution to macroscopic magnetic behavior is due to the electrons.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_001

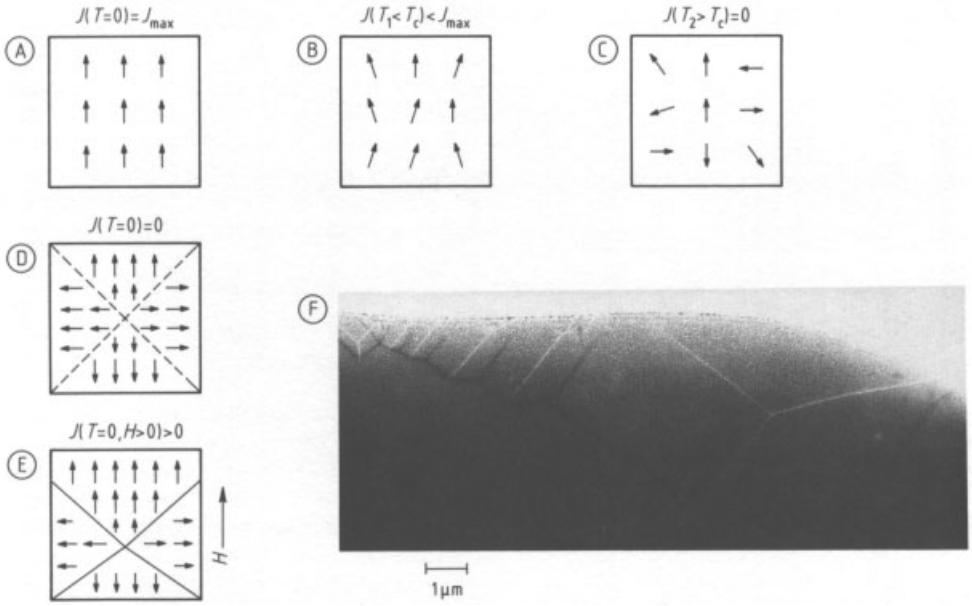

Figure 1. Schematic representation of ferromagnetic behavior (arrows represent atomic magnetic moments)
A) At absolute zero; B) Below the Curie temperature; C) Above the Curie temperature; D) Ferromagnetic domain structure with 90° domain walls; E) Ferromagnetic domain structure under the influence of an external magnetic field H; F) Electron micrograph (Lorentz microscopy) of the ferromagnetic domain structure in a thin FeCu sample near the edge of the foil [1]

2.1. Diamagnetism

Diamagnetism, a form of magnetism originating directly from the magnetic moments of the atom's constituents, is not associated with any particular state of aggregation. All substances are diamagnetic.

Diamagnetic materials are displaced out of inhomogeneous magnetic fields. Usually, however, diamagnetism is masked by other, stronger forms of magnetic interaction (see Sections 2.2 and 2.3).

Because of mutual cancellation of the individual moments, diamagnetic substances do not exhibit a resultant external magnetic moment. In the presence of an external magnetic field H, the diamagnetic moment or polarization J results from an additional precessional motion (Larmor precession) about the direction of the external field. The polarization J points in the opposite direction to H. These two quantities are linked by the negative susceptibility χ:

$$J = \mu_0 \chi H \tag{1}$$

where μ_0 is the magnetic field constant, equal to $4\pi \times 10^{-7}$ Vs/Am. Examples of diamagnetic substances are the noble gases, metals such as silver, gold and copper, water, and organic compounds such as ethanol and benzene.

2.2. Paramagnetism

If the magnetic moments of an atom do not completely cancel because of the electronic configuration, the atom exhibits a resultant moment. This is the case in paramagnetic substances.

In an external magnetic field, these moments, which were initially in a disordered state, align parallel to the field; this process competes with thermal disordering of the moments. The paramagnetic susceptibility χ is positive and depends strongly on temperature. At room temperature, it is of the order of $10^{-3} - 10^{-6}$.

Examples of paramagnetic substances are aluminum, potassium, magnesium, sodium, many gases, salts of the iron-group metals, and also all ferromagnetic and ferrimagnetic substances above the Curie temperature.

2.3. Ferromagnetism

Ferromagnetism is confined to the solid state of matter. In addition to the uncompensated magnetic moments of the individual atoms, in ferromagnetic materials there is also an exchange interaction between neighboring atoms that leads to adjacent moments becoming parallel in competition with the disordering effect of thermal motion. This is regarded as a spontaneous polarization that is inherent to the material and is present in macroscopic domains (Weiss domains) of the ferromagnetic material – even in the absence of an external magnetic field – and has its maximum value at absolute zero (see Fig. 1 A). If thermal energy is supplied, the order decreases (Fig. 1 B) until the paramagnetic state comes into being (Fig. 1 C) at the Curie temperature T_C, which is characteristic of the material. Ferromagnetism is thus a collective phenomenon based on an energetic interplay of many atoms in a definite geometrical association.

Despite its spontaneous magnetization in macroscopic domains, a ferromagnetic material can exhibit zero external net magnetism. The reason is that the entire volume of the material does not consist of a single domain spontaneously magnetized in one direction, but is subdivided into many domains in such a way as to minimize the magnetostatic energy (see Fig. 1 D). All these domains are spontaneously magnetized, but in various directions, so that the resultant moment can theoretically take any value between zero and the saturation polarization J_s(T), depending on the prehistory.

Changes in magnetization in the presence of an external magnetic field take place in one of two ways: favorably oriented domains grow by displacement of the domain walls (Bloch wall movement, Fig. 1 E), or entire domains collectively rotate their magnetic moments. Figure 1 F shows the ferromagnetic domain structure in FeCu.

The complexity of these processes means that the ferromagnetic susceptibility is not only markedly temperature dependent but also dependent on field strength, so that no single-valued relationship between the polarization J and the field strength H exists:

$$J = \mu_0 \chi(H) \cdot H \tag{2}$$

The relationship is illustrated by the hysteresis loop described in Section 3 (see Fig. 2).

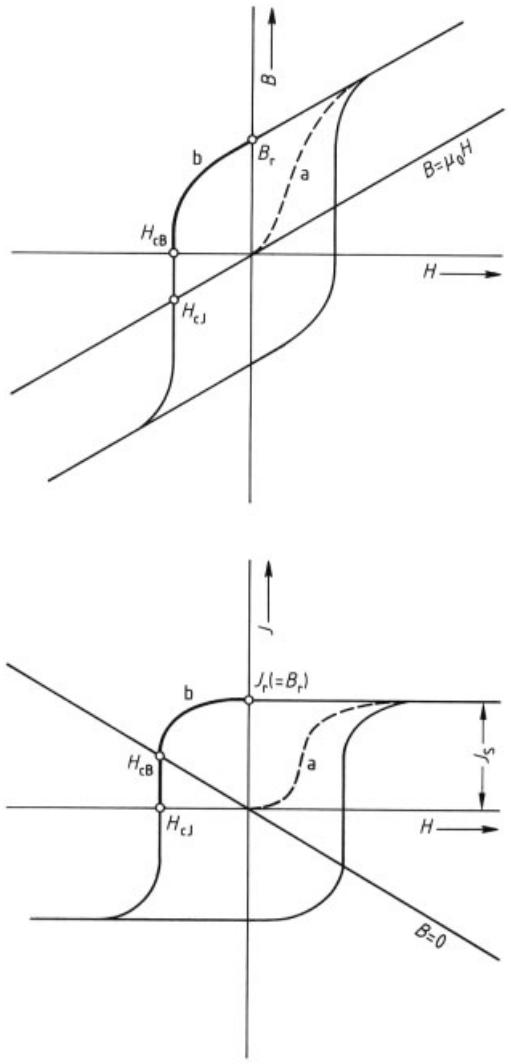

Figure 2. $J(H)$ and $B(H)$ hysteresis curves of ferromagnetic materials, showing the virgin curve (a) measured on magnetization of a previously demagnetized sample and the demagnetization curve (b)

The ferromagnetic susceptibility can have values between 0 and 10^6, depending on the material. Typical ferromagnetic materials are the elements Fe, Ni, and Co, as well as some of their alloys. Amorphous alloys can also exhibit ferromagnetic behavior.

2.4. Ferrimagnetism and Antiferromagnetism

Ferrimagnetic materials exhibit a structure consisting of two interpenetrating lattices (spinel structure), whose magnetic moments are unequal and opposite in direction. The effective moment is thus a differential moment, so that only small polarization values are obtained. This behavior, the much lower electrical conductivity, and the mechanical properties are the essential differences between ferrimagnetic and ferromagnetic materials.

The case in which the magnetic moments of the two sublattices cancel each other, so that the spontaneous magnetization with no external field is zero, is termed antiferromagnetism.

3. Magnetic Quantities and Their Measurement

3.1. General

Important terms that describe the magnetic properties of matter are the magnetic flux density or induction B, the magnetic polarization J, the magnetic field strength H, and other quantities associated with specific material behavior, such as the relative permeability μ or the magnetic susceptibility χ ($\mu = 1+\chi$).

The magnetic field and flux are related to electric field quantities by the fundamental vector equations of electric field theory, the most important of which are Maxwell's first and second laws:

$$\mathrm{curl}\ \vec{H} = \vec{j} \tag{3}$$

$$\mathrm{curl}\ \vec{E} = \partial \vec{B} / \partial t \tag{4}$$

The first equation relates the magnetic field strength H to a steady current distribution with current density $\vec{j}$. The law states that any conductor through which a current flows is surrounded by closed lines of magnetic field. For the magnetic field inside a long cylindrical or toroidal coil having length l and n turns, and carrying a current I, the magnetic field derived from (3) is $H = n\,I/l$. Accordingly, magnetic field strength has dimensions of current over length and units of A/m.

The second equation formulates the law of induction, according to which the time variation of a magnetic flux Φ through a surface A ($\Phi = BA$) induces an electrical voltage U. The unit of magnetic flux density is Vs/m^2 (1 Vs/m^2 = 1 Tesla).

The physical relationships expressed by the two formulas, modified in a variety of ways, explain how magnetic quantities can be measured by determining electrical quantities.

Other methods for measuring magnetic quantities are based on the force exerted by magnetic fields on electrically charged particles, such as the charge carriers in semiconductors. Probes employing this principle are called magnetoresistance cells or Hall probes. Another technique used for the precise determination of magnetic field strengths is nuclear resonance.

In order to identify and evaluate magnetic materials, the dependences of induction B and polarization J on magnetic field strength H must be known. Plots of these relationships for both soft magnetic and hard magnetic materials give hysteresis curves, though the crucial parameters characterizing the material are different in the two cases. Figure 2 shows a complete hysteresis loop in $J-H$ and $B-H$ plots, along with the associated "virgin" or initial magnetization curve obtained on the first magnetization of a magnetic material. The quantities B and J are related as follows:

$$B = \mu_0 H + J \tag{5}$$

The induction B thus consists of the polarization of the material J plus the polarization of a vacuum $\mu_0 H$.

The area enclosed by the hysteresis loop represents the work transformed to heat in a complete magnetization and demagnetization cycle. The quantities indicated in Figure 2 are explained in Sections 3.2 and 3.3.

3.2. Parameters of Hard Magnetic Materials

The quantities that describe a permanent magnet are the remanent induction, or remanence B_r measured in Tesla (T), the coercive forces of polarization H_cJ and induction H_cB in kA/m, and the energy product $(BH)_\mathrm{max}$ in kJ/m^3. These

quantities are labeled in the hysteresis curve of Figure 2. The branch of the loop in the second quadrant is called the demagnetization curve. The demagnetization curve is generally sufficient to characterize a hard magnetic material. It shows how the polarization (J – H plot) or the induction (B – H plot) behaves in the presence of an opposing magnetic field.

The coercive force H_{cJ} or H_{cB} is the external field strength at which J or B becomes zero. This is a measure of the stability of the magnetic state; H_{cJ} is always greater than or equal to H_{cB}.

The remanence B_r or the polarization J_r ($B_r = J_r$) is the induction in a field $H = 0$. Since a magnet is always subject to its own field, this quantity can be determined only in a closed magnetic circuit.

The energy product $(BH)_{max}$, the largest possible product of the B and H values along the demagnetization curve, is a measure of the magnetic quality of a permanent magnet. The greater this maximum for a material, the smaller the volume of magnet is needed to produce a given flux in the air gap of a magnetic system (the working region of the magnet). This fact accounts for the importance of high-quality permanent magnets for miniaturized components in many fields of modern technology.

At a remanence B_r, the theoretical $(BH)_{max}$ is given by

$$(BH)_{max} = \frac{B_r^2}{4\mu_0} \tag{6}$$

This value is reached when the coercive force H_{cB} is at least

$$H_{cB} = \frac{B_r}{2\mu_0} \tag{7}$$

Whether and to what extent this is really the case depends on the mechanism responsible for generating the coercive force. A marked magnetic anisotropy along one crystal axis, as is observed in hard ferrites and rare earth (RE) magnetic materials, is a prerequisite for a strong coercive force.

Thus, whereas hard ferrites and RE – Co and RE – Fe magnetic materials attain energy products within a few percent of $(BH)_{max}$ as calculated by Equation 6, in the AlNiCo and Cr – Fe – Co materials, in which precipitates having form anisotropy give rise to the coercive force, the $(BH)_{max}$ actually attained is usually less than 20 % of the theoretical maximum value.

Various measuring systems are available for determining the parameters of permanent magnets. Examples are highly sensitive fluxmeters, with which the integrated flux of the permanent magnet can be determined by the use of a search coil, and magnetic-field probes that use the Hall effect or a change in resistance to detect magnetic influences [2].

The measurement of the demagnetization curve or of the entire hysteresis loop is carried out in closed magnetic circuits consisting of a measuring yoke in which the magnetic specimens is clamped. The specimen, magnetized externally or in the yoke, is demagnetized by applying an opposing field. The flux variation of the magnet is determined by means of a search coil; the instantaneous induction B or polarization J of the magnet as a function of the opposing field strength is determined by measurement with a search coil or a Hall probe [3], [4].

3.3. Parameters of Soft Magnetic Materials

In contrast to the materials used for permanent magnets, magnetically soft materials must be easily magnetized or remagnetized. Even weak driving fields (a typical value for material characterization is $H = 0.4$ A/m) must produce the greatest possible flux density in the material.

The relation between B and H is described in terms of the relative permeability μ, with a subscript to identify the strength of the measuring field (e.g., $\mu_4 = \mu$ measured with $H = 0.4$ A/m). Starting with the unmagnetized state ($H = B = 0$), brought about either by thermal demagnetization or by an alternating field decaying to zero, the permeability increases with field strength up to the maximum permeability μ_{max}, then declines to unity as the flux density enters the magnetic saturation region.

By virtue of the relation $\mu = B/\mu_0 H$, the curve of μ versus H is derived from the "tip curve" of B versus H, which is made up of the end points of all smaller loops lying inside the largest hysteresis loop. This curve is obtained by reversing the field strength (from $+H$ to $-H$) and measuring the associated flux changes $\Delta\Phi$. It is very similar to the virgin magnetization curve (see Fig. 3).

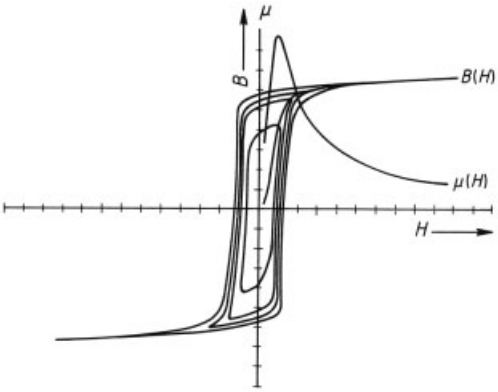

Figure 3. Hysteresis curves $B(H)$ for various maximum driving fields H, curve of normal magnetization $B(H)$, and permeability $\mu(H)$ for a soft magnetic material

The coercive force H_c and the magnetic loss P_{Fe} are used along with the permeability μ to characterize magnetically soft materials. Because of the low field strengths, H_{cB} and H_{cJ} are not usually distinguished.

The shape of the hysteresis curve is also important. It is described as rectangular, normal, or flat (see Fig. 4) and characterized by the remanence/saturation ratio B_r/B_s.

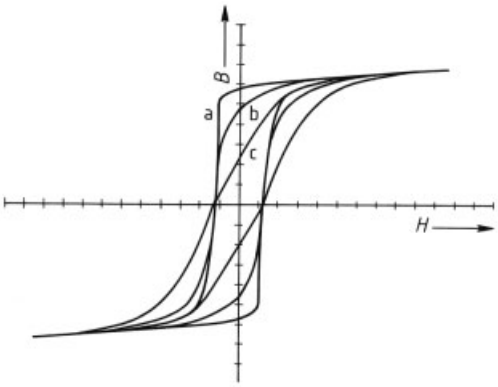

Figure 4. Typical types of hysteresis curves
a) Rectangular; b) Normal or round; c) Flat

All the parameters of soft magnetic materials must be classified as to whether they are measured statically or dynamically. Only static measurement yields information about the pure behavior of the material.

Dynamic driving involves alternating (a.c.) fields of given frequency, amplitude and form. Induction in the soft magnetic metal gives rise to eddy currents, which degrade the magnetic parameters (see Fig. 5). Classical eddy-current theory defines a cutoff frequency f_w at which the depth of penetration of the induced field in the specimen is equal to half the specimen thickness. The effect of the associated eddy currents becomes relatively strong above this frequency.

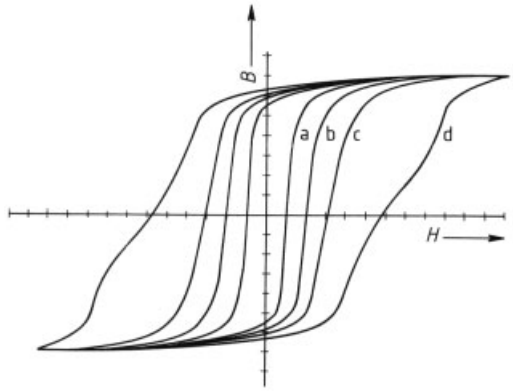

Figure 5. Hysteresis curves of a soft magnetic material measured at four different frequencies, with otherwise identical conditions.
a) 50 Hz; b) 200 Hz; c) 400 Hz; d) 1000 Hz

The variables influencing f_w include the initial permeability μ_i and resistivity ϱ of the material and the specimen thickness D. The cutoff frequency is given by

$$f_w \approx \frac{8\varrho}{\mu_0 \mu_i D^2}$$

In order to attain a high f_w at a high initial permeability level, the material must have a high resistivity and the core shape must be such that the specimen thickness is as small as possible in the direction perpendicular to the magnetic flux. Special core geometries devised to meet this condition include wound strip cores and laminated cores with sheet thicknesses of 0.05 – 0.35 mm having electrical insulation between sheets (see Fig. 6).

The parameters of magnetically soft cores are usually measured in closed magnetic circuits, often on toroidal specimens. In this way, external effects due to the geomagnetic field or parasitic stray fields in the environment, as well as measurement errors due to shearing effects such as occur in open circuits, are eliminated. Since the coercive force does not depend on shearing, it can be quickly determined in any geometry, provided specimen saturation can be established as a reproducible initial state.

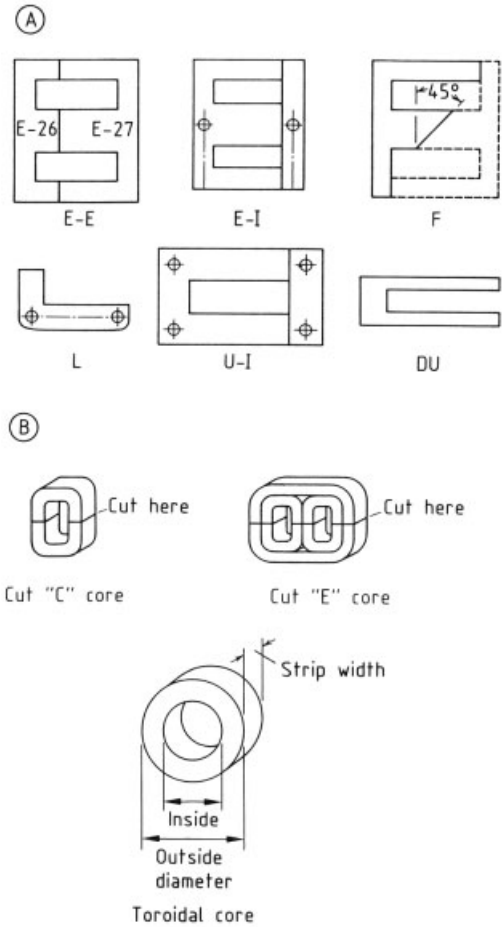

Figure 6. Typical core geometries for soft magnetic materials in dynamic applications
A) Core sheets; B) Core shapes

In dynamic measurements, the time-dependent current $I(t)$ and voltage $U(t)$ waveforms and therefore the field strength and flux density waveforms must always be taken into account, since they always have to be considered when the magnetic quantities are determined from electrical ones. In a closed magnetic circuit with a average magnetic path length l, the relation between the maximum current $\hat{I}$ and field strength $\hat{H}$ is given by

$$\hat{H} = \frac{n\hat{I}}{l} \tag{8}$$

The voltage and the flux density are related by

$$\hat{B} = \frac{1}{nA_e} \int u(t)\,\mathrm{d}t \tag{9}$$

For a sinusoidal voltage with effective value Ueff,

$$U_{\mathrm{eff}} = 4.44\hat{B}fnA_e \tag{10}$$

Efficient transformation of electrical voltages with small-cross-section A_e magnetic cores and few turns in the windings (i.e., with small inductive components) is possible if the flux density B is high at the operating point. The same result can be achieved by raising the operating frequency.

The purpose of developing improved magnetically soft materials is thus to lower the coercive force and losses and to insure a high flux density at the operating field strength.

4. Hard Magnetic Materials

Permanent magnets in industrial use can be classified as "classical" and "modern"; the classical ones are those developed up to about 1960, while the modern ones are those discovered and adopted in industry between the mid-1960s and the present. The most important of the classical materials are AlNiCo and hard ferrites. Modern materials include special rare-earth-containing alloys based on Sm−Co and Nd−Fe.

Table 1 lists the most important hard magnetic materials that have come into use since 1900. The years listed cover the time from the discovery of the material to the conclusion of the important development stages. The last two columns indicate the present importance of the materials.

From the 1800s up to the 1930s, the only materials for permanent magnets were special steels, which were "magnetically hardened" (i.e., given coercive force) by sometimes critical and costly heat-treatment processes. Tungsten, chrome and cobalt steels should be mentioned in this connection; in comparison with present-day hard magnetic materials, all of these exhibit lower stability of the permanent magnetic state. Mechanical effects, vibration, or the action of weak opposing magnetic fields irreversibly weakens the magnetic state.

An important advance in the field of permanent magnets took place in Japan in 1932, when precipitation-hardenable Al−Ni−Fe alloys were discovered [6], [7]. In the years following, these formed the basis for the AlNiCo

Table 1. Overview of hard magnetic materials in use from the early 1900s.

Material	Discovery/ development	Present-day importance	Trend
Hard magnetic steels	before 1930	important	
Cu – Ni – Fe	1935 – 1940	unimportant	
Cu – Ni – Co			
Pt – Co	1935 – 1940	minor	decreasing
V – Co – Fe	1940 – 1955	minor	decreasing
Fe – Co (ESD)	1950 – 1960	unimportant	
Cr – Fe – Co	1936 – 1960	minor	unchanged
Mn – Al – C	1958 – 1980	unimportant	
Al – Ni/AlNiCo	1932 – 1960	medium	decreasing
Hard ferrite	1950 – 1970	major	increasing
SmCo 1 : 5	1965 – 1975	major	increasing
SmCo 2 : 17	1975 – 1985	major	increasing
Nd – Fe – B	1983	medium	increasing considerably

materials, which are still of commercial importance.

Related to the AlNiCo alloys in metallography and magnetic qualities are the Cr – Fe – Co materials [8–10], which were also discovered in the mid-1930s but not industrially adopted until the beginning of the 1970s (see 4.1.1.2).

The second important group of classical hard magnetic materials are the hard ferrites. They were discovered around 1950, and developed and brought on the market between 1950 and 1960 [11], [12]. Because the raw materials are cheap and readily available and the manufacturing processes are economical, hard ferrites account for most permanent-magnet production despite their relatively low magnetic efficiency.

In addition to the AlNiCo and hard ferrite groups, a number of other materials were developed between 1930 and 1960, but these gained little commercial or industrial significance. They are listed in Table 1 and discussed in Section 4.1.3.

4.1. Classical Hard Magnetic Materials

Hard magnetic materials developed or known before 1960 include the AlNiCo group and hard ferrites, both of which are still commercially important, and also Cr – Fe – Co, Co – V – Fe, Pt – Co, Cu – Ni – Fe, Cu – Ni – Co, Mn – Al – C, and Fe – Co (ESD) materials. A few of those last named materials are still used for special applications, but their market share is shrinking.

4.1.1. Aluminum – Nickel – Cobalt Materials

Permanent magnets based on aluminum – nickel – cobalt – iron alloys, referred to as AlNiCo, form a broad spectrum of component-rich alloys with the following composition range: Al 6 – 13, Co 0 – 32, Cu 2 – 6, Ni 13 – 18, Ti 0 – 9, Nb 0 – 3 wt %, remainder Fe. Grades without cobalt, called AlNi magnets, are also available commercially; all of these have lower magnetic values. As the cobalt content increases, better values are obtained.

AlNiCo magnets are made by casting or sintering. In the casting process, the starting materials are melted in an open medium-frequency furnace and cast in sand or precision molds. In the powder-metallurgical processes, iron, cobalt, and other elements or appropriate master alloys are thoroughly mixed in powder form, shaped by compression at room temperature under a pressure of several kilobar, and sintered in an atmosphere free of oxygen and moisture either in vacuum or under an inert gas. Sinter temperatures range from 1250 to 1400 °C, depending on the alloy composition. The linear dimensions shrink by about 10 %, corresponding to an increase in density from 5 to 7 g/cm^3 during the sintering process.

Each of these processes has special features that influence the magnetic and mechanical properties of the products. Sintered magnets are finer grained than cast ones and have a uniform, cavity-free microstructure. They accordingly exhibit greater mechanical strength

than cast magnets. Furthermore, sintered magnets can be economically fabricated down to a volume of a few cubic millimeters. This possibility of making small parts is important given the present trend toward miniaturization. From the process engineering standpoint, powder-metallurgy techniques offer several further advantages over casting. Tolerances can be kept narrower, small holes and penetrations are easily made, and in general more intricate shapes can be realized.

In casting, however, controlled removal of heat during solidification makes it possible to create an oriented columnar structure in which the long direction of the grains lies preferentially close to a [100] direction of the cubic crystal lattice [13], [14]. Starting from this ordered state, excellent magnetic values in the grain axis direction can be achieved in subsequent magnetic refining.

For materials of the AlNiCo group, therefore, the selection of fabrication method is dictated by mechanical, magnetic, and economic considerations. Generally, heavy magnets are cast while lightweight ones are sintered; the boundary between these classes lies in the range 10 – 50 g. In special cases, the rule is violated; for example, rotors for high-speed motors require high strength, and although they weigh several hundred grams, they are produced by sintering. Regardless of the fabrication method, AlNiCo magnets are hard and brittle. The only way to machine them, at any stage of production, is by grinding.

After sintering or casting, the blanks have no permanent magnetic properties until they are given such properties in a multistage heat-treatment process. The cycle consists essentially of three stages, with parameters governed by the alloy composition and the type of magnet desired. A distinction is made between the production of isotropic and anisotropic magnets [15]. In anisotropic magnets, good magnetic properties are produced preferentially in one direction in space, whereas in isotropic magnets, all spatial directions are magnetically equivalent. Table 2 describes the most important steps in heat treatment.

Table 2. Heat treatment steps for AlNiCo materials

Treatment step	Isotropic magnets	Anisotropic magnets
1	Homogenization (1200 – 1300 °C)	Homogenization (1200 – 1300 °C)
2	Cooling at controlled rate to 600 °C (spinodal decomposition)	a) Cooling at controlled rate in a magnetic field (800 – 700 °C) or b) Isothermal treatment in a magnetic field near the Curie temperature (ca. 800 °C) (spinodal decomposition)
3	Tempering $T \approx 500 - 650$ °C	Tempering $T \approx 550 - 650$ °C

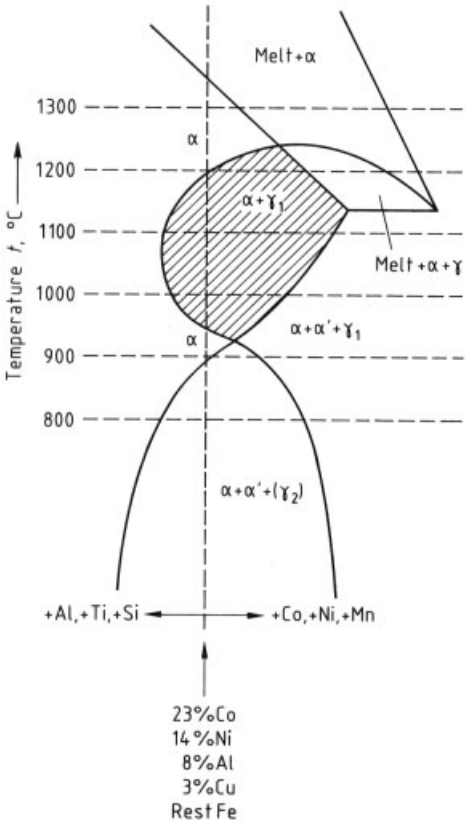

Figure 7. Pseudobinary Al – Ni – Co phase diagram (valid for the given composition) [20]

Understanding the processes that take place during heat treatment, requires knowledge of the phase diagram [16], [17]. Figure 7 shows the pseudobinary diagram of the Al – Ni – Co system. The dashed lines separate temperaturedependent phase regions for the compositions shown. The addition of the components shown at right and left shifts the line and thus the corresponding phase boundaries in the way indicated.

Homogenization, the first heat-treatment operation in Table 2, establishes the α-phase, which has a body-centered cubic lattice. Rapid cooling suppresses the γ-phase and freezes in theα-phase. Then, at temperatures below the $\alpha - \gamma_1$ mixedphase region, the α'-phase, which is also body-centered cubic, precipitates. This precipitation occurs by the spinodal decomposition mechanism. A highly uniform microstructure comes into being spontaneously and without the input of activation energy; the strongly magnetic, Fe – Co-rich α'-phase is present as inclusions in the weakly magnetic or nonmagnetic, Al – Ni-rich α-matrix. The shape and size of the precipitate can be controlled through temperature and time so that the finely dispersed α' inclusions are the size of magnetic single-domain particles, with transverse dimensions of ca. 10 nm and lengths of 80 – 150 nm.

It can be assumed that the single-domain particles obtained in the ideal case are remagnetized by domain rotation. For AlNiCo 700, in particular, it has been shown [18] that the experimental coercive force roughly corresponds with the value calculated from the particle composition and geometry. The hard magnetic properties of AlNiCo magnets thus bear a causal relationship to the shape anisotropy of the microstructure. The Curie temperature T_C for Al – Ni alloys is 760 °C; in AlNiCo materials, the value rises to ca. 850 °C with increasing Co content. At over 20 wt % Co (e.g., ca. 24 % Co in AlNiCo 500 or ca. 34 % Co in AlNiCo 450), the Curie temperature of the strongly magnetic α'-phase is in the region of the onset of spinodal decomposition. If phase decomposition takes place in a magnetic field, the columnar precipitates can be caused to align themselves with the field [15], [19] (each close to the nearest [100] direction of the crystallite). Depending on the alloy composition, the magnetic treatment is carried out by continuous heating in a magnetic field or by isothermal modifications of the heat-treatment procedure. In particular, high-coercivity AlNiCo materials containing titanium require finely controlled isothermal treatment in a magnetic field.

Phase decomposition in a magnetic field thus brings about an orientation with a microscopic preferential direction and leads to anisotropic magnets. The establishment of a preferential direction is especially efficient in the case of blanks with columnar grain structure when the magnetic field is parallel to the grain orientation. The best remanence and energy values are obtained in this way (see Table 3).

In tempering, the third heat-treatment operation in Table 2, the composition differences between the α- and α'-phases are heightened by diffusion processes [21], [22]. The phase-decomposition pattern, on the other hand, does not change. The coercive force is strengthened in this operation.

Table 3. Compositions and magnetic values for typical AlNiCo materials

Material designation (DIN)	Composition, wt %						Magnetic values				
	Al	Ni	Co	Cu	Ti	Fe	$(BH)_{max}$, kJ/m^3	Remanence B_r, mT	Coercive force H_{cB}, kA/m	Coercive force H_{cJ}, kA/m	Density ϱ, g/cm^3
Isotropic magnets											
AlNiCo 9/5	13	24	5	7		51	10	550	55	58	6.9
AlNiCo 12/6	11	20	15	4		50	12	600	54	58	7.1
Anisotropic magnets											
AlNiCo 35/5	9	14	24	3		50	40	1250	50	51	7.3
AlNiCo 40/12	8	14	36	3	5	34	44	880	115	120	7.3
AlNiCo 40/15	8	14	40	3	8	27	43	740	150	160	7.3
Columnar crystallized											
AlNiCo 52/6	9	14	24	3		50	62	1240	58	59	7.3
AlNiCo 80/10	8	14	40	3	8	27	80	930	150	155	7.3
Hysteresis materials											
AlNiCo 3/0.6	7	17				76	3	1000	6	6	7.1
AlNiCo 6/1.6	9	17	5			69	6	820	16	16	7.0

The many compositional variations in multicomponent AlNiCo alloys, coupled with the range of manufacturing processes, such as casting, crystallization in columnar grains, and sintering, as well as the treatment modifications leading to isotropic or anisotropic magnets, mean that there are many types of AlNiCo. About ten hard magnetic types are marketed along with several moderately hard types having intentionally low coercive force and intended for use as hysteresis materials in motors and clutches.

The isotropic, anisotropic, and columnar crystallized materials vary widely in B_r and $(BH)_{max}$ but have similar coercive forces. This comparison reflects the way in which the coercive force depends on the shape anisotropy of the precipitates, which are the same in isotropic and anisotropic qualities.

AlNiCo magnets also exhibit good temperature stability, with the remanence and coercive force being only weakly temperature dependent. The maximum service temperature of 550 °C is comparable to that of the last tempering step.

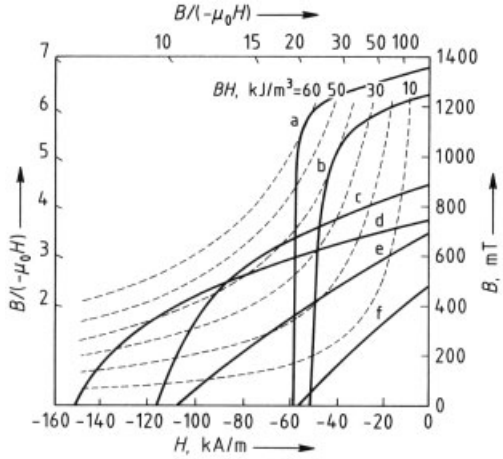

Figure 8. Typical demagnetization curves for isotropic (8/5 and 20/11) and anisotropic AlNiCo magnets
a) AlNiCo 52/6; b) AlNiCo 40/5; c) AlNiCo 40/12; d) AlNiCo 40/15; e) AlNiCo 20/11; f) AlNiCo 8/5

Figure 8 shows demagnetization curves for some isotropic and anisotropic AlNiCo magnets. Characteristic temperature data for AlNiCo magnets are listed in Table 4 along with those for other magnetic materials. The Curie temperatures for AlNiCo magnets range from 760 to 850 °C, depending on composition. The last two columns give the temperature coefficients for reversible changes in J_s or B_r and H_{cJ}. A comparison of the values with those for other materials shows that AlNiCo magnets have superior temperature stability. These properties dictate the use of AlNiCo magnets wherever high flux con-

Table 4. Characteristic temperatures for permanent magnets

Material	Curie temperature T_c, °C	Max. operating temperature, °C	$Tk(B_r)$, %/K	$Tk(H_{cJ})$, %/K
$SmCo_5$	720	250	−0.04	−0.25
SmCo 2 : 17 (high H_{cJ})	820	350	−0.03	−0.25
SmErCo 2 : 17	820	250	−0.002	−0.25
NdFeB	310	100	−0.12	−0.6
NdDyFeB	310	150	−0.10	−0.55
AlNiCo	800	500	−0.02	+0.03
Hard ferrite	450	250	−0.2	+0.40

stancy is needed, for example, in measuring instruments.

4.1.2. Chromium – Iron – Cobalt Materials

Hard magnetic Cr – Fe – Co alloys have many similarities to the AlNiCo group from metallographic and magnetophysical standpoints [23]. Chromium – iron – cobalt magnets can also be manufactured by casting or sintering. Whereas AlNiCo magnets are hard and brittle at all stages of production, the Cr – Fe – Co materials can be hot- or cold-worked at certain stages. This is the case after homogenization as well as after brief spinodal treatment. Before the final heat-treatment operations, in which the magnetic values are established, the material can be rolled, bent, punched, extruded or drawn, so that even complicated geometric forms can be fabricated. Once heat treatment is finished, Cr – Fe – Co magnets are hard and brittle. The magnetic properties are set after shaping, in a multistep heat treatment that differs in temperature range and duration from the treatment of AlNiCo magnets. A spinodal decomposition also takes place in this system, in the temperature range 600 – 700 °C, depending on alloy composition, and again an α'-phase precipitates from a high-temperature α-phase [24]. The Fe – Co-rich α'-phase forms inclusions in the chromium-rich matrix. As in the AlNiCo materials, the precipitates are anisotropic in form, and this property again forms the basis for the coercive force [25], [26]. Table 5 lists the important treatment steps for establishing the magnetic values of Cr – Fe – Co materials.

Table 5. Heat treatment steps for Cr – Fe – Co magnets

Treatment step	Isotropic magnets	Anisotropic magnets
1	Homogenization 950 – 1250 °C	Homogenization 950 – 1250 °C
2	Heat treatment 600 – 700 °C (spinodal decomposition)	Heat treatment in magnetic field at 600 – 700 °C or brief treatment at 600 – 700 °C followed by 60 – 70 % cold working (spinodal decomposition)
3	Multistep heat treatments between 600 and 700 °C	Multistep heat treatments between 600 and 700 °C

The application of a magnetic field during spinodal decomposition gives anisotropic magnets. Not only can anisotropy be introduced by magnetothermal treatment, but magnetic anisotropy in Cr – Fe – Co can also be achieved through a certain cycle of heat treatment and working. Globular precipitates are formed after brief spinodal treatment. Since the material is still ductile in this condition, the precipitates can be converted to particles with shape anisotropy by cold working (60 – 70 %) [27]. These extended phase regions are comparable in shape and effect to the columnar precipitates produced by thermomagnetic treatment.

In the final heat treatment, the concentration differences between the two phases are reinforced by diffusion processes. This leads to the generation of the coercive force. The treatment time required to set the magnetic values increases with decreasing Co content [28]. At about 23 % Co, the time for spinodal treatment and tempering is 3 – 4 h; at 3 % Co, it is ca. 150 h. Figure 9 shows schematically an entire heat-treatment cycle for Cr – Fe – Co alloys [29].

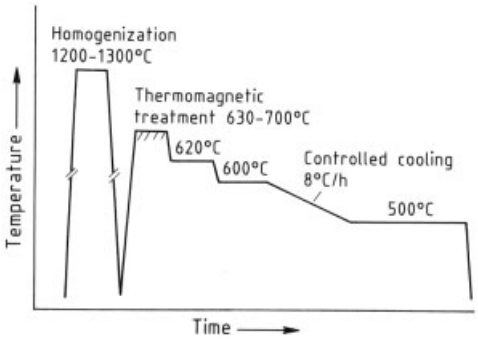

Figure 9. Progression of a Cr – Fe – Co heat-treatment cycle [29]

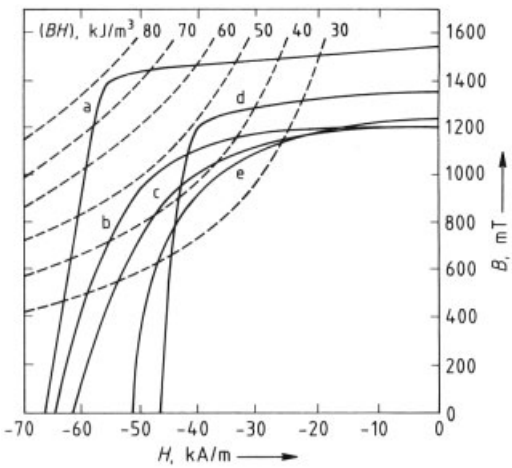

Figure 10. Comparison of the demagnetization curves for the Cr – Fe – Co and Al – Ni – Co alloys (columnar crystallized; magnetically tempered)
a) 24 % Cr, 15 % Co, 3 % Mo, remainder Fe (anisotropic through deformation) [29]; b) 33 % Cr, 11.5 % Co, remainder Fe [28]; c) 33 % Cr, 11.5 % Co, 2 % Cu, remainder Fe (magnetic field treatment) [28]; d) 27 % Cr, 9 % Co, remainder Fe [28]; e) AlNiCo 24 % Co, 8 % Al, 14 % Ni, 3 % Cu, remainder Fe

The composition range (in wt %) of the Cr – Fe – Co hard magnetic materials is as follows: Cr 20 – 33, Co 3 – 25, Si 0 – 3, Mo 0 – 3, Cu 0 – 3, remainder Fe.

The Cr – Fe – Co materials, both isotropic and anisotropic, have values corresponding to the middle range of AlNiCo materials. Comparable values are, however, attained in Cr – Fe – Co with significantly less Co. Table 6 summarizes the compositions and magnetic values of Cr – Fe – Co materials. For comparison, the last line describes an AlNiCo quality.

Although good magnetic values are obtained in permanent magnets with only 3 % Co, the processing times for such magnets are so long that these materials are of no commercial importance.

Figure 10 shows demagnetization curves for Cr – Fe – Co alloys and a magnetically comparable AlNiCo quality.

4.1.3. Other Metallic Hard Magnetic Materials

In addition to the AlNiCo and Cr – Fe – Co materials, a series of other metallic magnetic materials was developed in the same time period. Although these are now of little industrial importance, they are described here briefly for the sake of completeness.

Platinum – Cobalt. Hard magnetic materials based on the two-component system Pt – Co, containing 75 wt % Pt and 25 wt % Co, were developed in the mid-1930s [30]. Maximum *BH* values of ca. 80 kJ/m^3 at a coercive field strength of 50 kA/m (Table 7) were obtained after heat treatment at 600 – 1000 °C with controlled cooling, during which a two-phase structure consisting of a face-centered cubic and a tetragonal phase is formed. Due to the high cost of Pt – Co, it was used for small or very small magnets such as those in hearing aids and watches. Today, Sm – Co magnets are used for these applications. The high corrosion resistance of Pt – Co led to its use in medical technology.

Cobalt – Vanadium – Iron. Magnetic materials containing 50 wt % Co, 7 – 15 wt % V, and the remainder Fe were developed in the 1940s [31]; 3 – 5 % of the vanadium can be replaced by chromium [32]. These alloys were known as vicalloy. The material can be worked by cold rolling or drawing before the magnetic values have been established. Since a cold-working grade of ca. 90 % is required for optimum hard magnetic properties to be attained, only wires with a maximum diameter of 3 mm and ribbons with a maximum thickness of 0.6 mm can be produced. After cold rolling or drawing, the material can be formed by stamping, cutting, and bending. However, the high tensile strength limits the extent to which it can be bent and the

Table 6. Overview of the composition and magnetic values of anisotropic Cr – Fe – Co magnets

Composition, wt %						
Co	Cr	Fe	Additives	B_r, mT	H_{cB}, kA/m	$(BH)_{max}$, kJ/m^3
25	30	remainder	3 Mo	1150	62	40
15	30	remainder	5 V	1320	43	40
15	24	remainder	3 Mo	1540	67	76
9	27	remainder		1300	46	49
3	32	remainder		1250	40	39
24		remainder	Al, Ni, etc.	1220	50	40

Table 7. Comparison of hard values for various materials

Material	Isotropic or anisotropic	B_r, mT	H_{cB}, mA/m	H_{cJ}, mA/m	$(BH)_{max}$, kJ/m^3
Pt – Co	isotropic	600	380	500	70
Co – Fe – V	anisotropic	1200	30	32	20
Cu – Ni – Fe	isotropic	350	30	32	4
Cu – Ni – Fe	anisotropic	550	42	44	12
Co – Fe (ESD)	anisotropic	1050	75	77	50
Mn – Al – C	anisotropic	580	220	250	52

size of the holes that can be stamped. After forming, the magnetic values are established in a final heat-treatment process at 530 – 620 °C, during which part of the ferromagnetic phase transforms to a nonferromagnetic phase, and the magnetically easy direction rotates through 90° towards the rolling or drawing direction. This treatment also greatly increases the hardness of the alloy, so that it can then only be worked by cutting and grinding. The magnetic properties of these alloys are listed in Table 7.

Copper – Nickel – Iron and Copper – Nickel – Cobalt. The hard magnetic properties of the three-component system Cu – Ni – Fe, with the composition 60 % Cu, 20 % Ni, and 20 % Fe, were discovered in the 1930s [33]. Due to the presence of a strongly temperature dependent miscibility gap in the Cu – Ni – Fe system, the precipitation of an Fe – Ni-rich phase from the copper-rich matrix can be controlled by heat treatment, after the material has been homogenized at 1040 °C, quenched, and heat treated at 650 °C. This procedure gives anisotropic magnets [34]. Short tempering of the alloy at 650 °C, followed by 90 – 95 % cold forming and heat treatment at 650 °C establishes a magnetically easy direction in the direction of working [35]. Magnetically isotropic and anisotropic Cu – Ni – Fe materials are listed in Table 7. The Cu – Ni – Co magnets did not gain industrial importance.

ESD Magnets. The idea behind elongated-single- domain (ESD) materials was to prepare highly shape anisotropic powder particles from alloys with high saturation magnetization. Elongated-single- domain particles can be prepared from iron or iron – cobalt salts at room temperature by means of a mercury electrode [36], [37]. The resulting fine particles are enlarged by brief heat treatment at 100 – 200 °C, and traces of mercury are removed. The powder is converted to anisotropic magnets by compression in an orienting magnetic field, with addition of tin, antimony, or lead as a binding agent. Further heat treatment or mechanical working of the pressed magnets is not necessary. The ESD magnets, which were used for many years in the United States under the trade name Lodex, are now only of historical importance. Magnetic values are listed in Table 7.

Manganese – Aluminum – Carbon. The nonferromagnetic elements manganese and aluminum form a ferromagnetic $\bar{l}$-phase (69 – 75 wt % Mn, remainder Al) [38]. The tetragonal $\bar{l}$-phase is formed from a hexagonal high-temperature phase by cooling from 1000 °C at a rate of 350 – 780 °C/min. This $\bar{l}$-phase exhibits

a uniaxial magnetic anisotropy and has a Curie temperature of 300 °C. The addition of 0.4 – 1 wt % carbon increases the stability of the ferromagnetic phase from 400 °C to 700 °C. In the 1970s anisotropic Mn – Al – C magnets with good mechanical and magnetic properties were obtained by heat treatment at 650 – 750 °C with 60 – 70 % reduction in cross-sectional area [39], [40]. Despite the cheap raw materials and the good magnetic values, these materials failed to gain industrial importance because of the complex production process [41]. The magnetic values are listed in Table 7.

4.1.4. Hard Ferrites

Hard ferrites form a group of ceramic permanent-magnet materials based on phases $MFe_{12}O_{19}$ (where M = Ba, Sr, Pb) and solid solutions of these such as $Ba_{1-x}Sr_xFe_{12}O_{19}$. The barium and strontium ferrites have found industrial use, while lead ferrites are not manufactured for reasons of toxicity.

The name hard ferrite results from the ability of these materials to generate high coercive forces that can be reproducibly set under optimized production conditions. As ceramics, hard ferrites exhibit high mechanical hardness and are brittle and very susceptible to fracturing.

Hard ferrites crystallize in the hexagonal magnetoplumbite structure. The crystal lattice exhibits hexagonal close sphere packing for the oxygen ions as well as the barium and strontium ions, while the smaller iron ions occupy interstitial positions. The unit cell of this ferrite lattice contains two barium or strontium ions, 24 iron, and 38 oxygen, and thus includes two $Ba/Sr \cdot Fe_{12}O_{19}$ formula units. The theoretical densities of barium and strontium ferrites, calculated from the unit cell, are $\varrho = 5.28$ and 5.10 g/cm^3, respectively.

Hard ferrites exhibit uniaxial magnetic crystal anisotropy. Magnetic polarization spontaneously takes up the direction of the crystallographic *c* axis. The resulting large crystal anisotropy constant is the basis for the high coercive forces realized with these magnetic materials (see Section 4.2.1, Eq. 11). The crystal anisotropy constants, saturation polarizations, anisotropy field strengths H_A, and the coercive forces achieved are summarized in Table 8. As in the rare earth permanent magnets (see Section 4.2.1) defects and inhomogeneities mean that hard ferrites can reach only ca. 20 % of the theoretical coercive force H_A.

Table 8. Anisotropy constant K_1, H_A, H_{cJ} and J_s values for hard ferrites [44]

	K_1, kJ/m^3	H_A, MA/m	J_s, mT	H_{cJ}, kA/m
$Ba\,0 \cdot 6 \cdot Fe_2O_3$	300	1.34	460	150 – 350
$Sr\,0 \cdot 6 \cdot Fe_2O_3$	346	1.45	472	150 – 350
$Pb\,0 \cdot 6 \cdot Fe_2O_3$	220	1.13	464	

The carrier of the magnetic polarization of hard ferrites is iron. Trivalent iron ions located at five nonequivalent lattice sites couple antiferromagnetically via oxygen bridges. Because of partial compensation of the magnetic moments of the iron ions, hard ferrites exhibit a magnetic saturation polarization that is low in comparison with other hard magnetic materials [42] (see Table 9). On average, this ferrimagnetic coupling leaves only one of three magnetic moments with any external effect. High-quality anisotropic hard ferrites with remanences of 440 – 450 mT achieve 93 – 95 % of the theoretical saturation polarization.

Hard ferrites are made by powder sintering or with resin binders [43]. The essential steps in the manufacturing process are reaction sintering, grinding, molding of the powder into compacts, sintering, and mechanical finishing of the sintered articles if necessary.

Barium and strontium ferrites are produced by heating a mixture of the corresponding carbonate with iron(III) oxide at 1200 °C.

$$MCO_3 + 6\,Fe_2O_3 \rightarrow MO \cdot 6\,Fe_2O_3 + CO_2$$

Details of this multistage reaction process can be found in the literature [44].

Low contents of SiO_2 and Al_2O_3 do not cause difficulties, since small quantities of these substances are frequently added on purpose to control sintering behavior (SiO_2) and increase the coercive force (Al_2O_3). Optimal magnetic values are obtained when the Ba or Sr content is increased to a molar ratio of ca. 1 : 5.6.

The compounds formed in reaction sintering are ground in ball or attritor mills, with water as grinding medium, to yield powders with particle sizes ≤ 1 µm. In the manufacture of high-quality hard ferrites, it is very important to achieve a

Table 9. Magnetic values for hard ferrite sinter and injection-molded magnets

Material	Isotropic or anisotropic	B_r, mT	H_{cB}, kA/m	H_{cJ}, kA/m	$(BH)_{max}$, kA/m^3	Density, g/cm^3
Sinter magnets						
Hard ferrite 8/27	isotropic	230	155	290	9	~4.9
Hard ferrite 24/35	anisotropic	360	260	350	24	~4.9
Hard ferrite 32/25	anisotropic	410	240	250	32	~4.9
Hard ferrite 32/17	anisotropic	410	160	165	32	~4.9
Injection-molded magnets						
Hard ferrite 1/18 P	isotropic	80	60	170	1.0	~2.9
Hard ferrite 8/19 P	anisotropic	120	80	200	8.0	~3.5
Hard ferrite 12/12 P	anisotropic	250	180	220	12.0	~3.5

narrow, reproducible particle-size distribution in which the single- crystal particles are the size of single magnetic domains. Both the coercive force and the remanence depend on this point.

In the next step, the ferrite powder is molded, wet or dry, at pressures of 100 – 300 MPa, into compacts with densities of 3.3 – 3.5 g/cm^3. If an orienting magnetic field of 200 – 400 kA/m is imposed during compaction, the single-crystal particles line up parallel to the field [12]. This orientation is fixed by compaction and is not disrupted in sintering. In this way, anisotropic magnets with a distinct easy direction are obtained. Wet compaction, with 28 – 35 % water added to the mixture, gives the best alignment of powder particles in the compacts. Much of this water is expelled on compression. Both isotropic and anisotropic ferrite magnets of medium quality are made by the compaction of dry powders.

The compacts are sintered in air at 1150 – 1250 °C in continuous tube furnaces; the coercive force and remanence can be controlled within certain limits by varying the sintering parameters. The density increases in sintering, ultimately reaching 95 – 98 % of the theoretical material density. The linear contraction on sintering is ca. 16 % for isotropic magnets, while for anisotropic magnets, it is ca. 22 % parallel to the easy direction and ca. 13 % perpendicular to the easy direction. Hard ferrites do not require further heat treatment to set or optimize the magnetic values after sintering. The last step in production is grinding if the dimensional tolerances make this necessary.

Along with sintered magnets, resin-bonded ferrite magnets (plastoferrites) have also become economically and industrially important [44], [45]. These magnets are produced by mixing hard ferrite powders with, for example, rubber, poly(vinyl chloride), polyamide, or resins. This operation takes place in a mixer, mixer-extruder, or kneader. Rolling and injection molding or compression molding follow. Anisotropic magnets can be produced by injection molding and compression molding if orienting magnetic fields are imposed.

Depending on the production process, plastoferrites with various filler contents can be made. Injection molding makes it possible to obtain isotropic and anisotropic magnets with ferrite powder contents of 50 – 65 vol %; isotropic magnets made by injection molding or rolling contain up to 70 vol% filler, and compression-molded articles up to 80 vol %.

The reduction in magnetic values due to the lower volume fraction of the magnetic phase in resin-bonded magnets is offset by a number of technical and economic advantages. For example, plastoferrites have much better mechanical properties than brittle sintered magnets because of their elasticity and toughness. These properties, as well as the other physical and chemical properties, depend on the magnetic material as well as the filler content.

Improved mechanical properties offer substantial advantages for subsequent processing. The higher stability of the edges permits automatic handling of the magnets during assembly. The toughness of composite materials in conjunction with the close tolerances that are possible with injection molding makes it possible to use interference fits for attaching magnets to other components. By virtue of high edge strength and easy, high-precision fabrication of complicated shapes without finishing treatment, product costs can be held down and, in many cases, new products can be realized. The mechanical properties and manufacturing ad-

vantages of resin-bonded hard ferrites set them above all competing hard magnetic materials in applications where their relatively low magnetic level can be accepted.

Table 9 summarizes typical magnetic values for isotropic and anisotropic sintered and resin-bonded ferrites. The values in the table are for room temperature. Figure 11 shows demagnetization curves for isotropic and anisotropic hard ferrites.

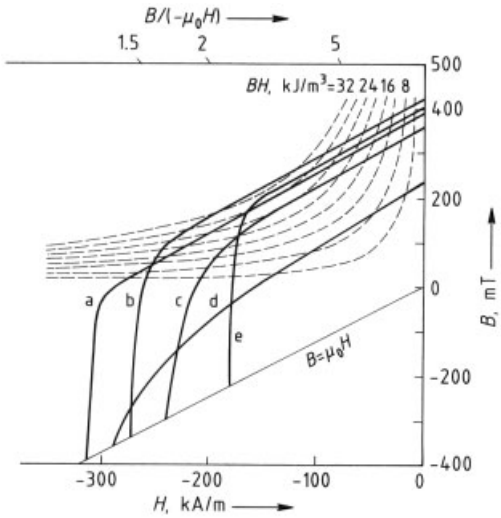

Figure 11. Demagnetization curves for hard ferrites; 8/27 isotropic, remainder anisotropic
a) Hard ferrite 26/29; b) Hard ferrite 32/26; c) Hard ferrite 26/29; d) Hard ferrite 8/27; e) Hard ferrite 28/15

The magnetic values of hard ferrites depend strongly on the temperature. In addition, the temperature variations of remanence and coercive force have opposite signs. Typical temperature data and temperature coefficients for hard ferrites are listed in Table 4.

4.2. Modern Hard Magnetic Materials

The development of hard magnetic materials with rare earth elements (RE) as essential components began in the mid-1960s and has yielded three groups of materials with magnetic values several times greater than those of classical AlNiCo and hard-ferrite magnets.

The $RECo_5$ magnets were the first generation of these materials [46], [47]. Of these, $SmCo_5$ is the magnetic material of greatest commercial and industrial significance today.

At the beginning of the 1970s, development started on a second generation of magnetic alloys based on rare earth elements and cobalt. These are multicomponent alloys containing not only rare earth elements and Co but also Cu, Fe, and alternatively Ti, Zr, or Hf [48–50]. The samarium-based alloys are most important in this group as well. As Table 11 shows, these materials contain ca. 10 % less Sm and ca. 15 % less Co than does $SmCo_5$.

The most recent group of rare earth hard magnetic materials are based on neodymium – iron – boron; these magnets were developed in 1983 in Japan [51] and the United States [52]. Their magnetic values surpass those of all other hard magnetic materials.

Figure 12 presents figures characterizing the quality of classical and modern hard magnetic materials: the energy product as a measure of magnetic performance and the coercive force as a measure of magnetic stability. The comparison makes clear the superiority of modern materials. The years listed mark the beginning of industrial production in each case. The rising trend in magnetic values has led to both smaller magnet systems and also new applications for permanent magnets.

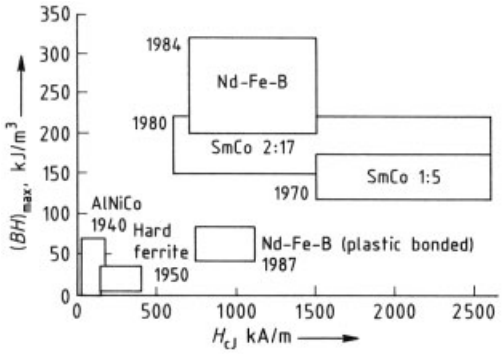

Figure 12. Hard magnetic material values (the year denotes the beginning of the industrial application)

The assessment of a hard magnetic material from the economic standpoint depends on not only the magnetic properties but also the availability and cost of raw materials and the reliability of supplies. In view of these points, Nd – Fe – B materials are much more promising than the Sm – Co group [53].

The most important rare earth elements for the manufacture of high-quality permanent

magnets, samarium and neodymium, occur in markedly different concentrations in ores. The abundance of Nd is 7 – 16 times greater than that of Sm. While ca. 350 t of Sm_2O_3 is available at present, enough for ca. 1000 t of Sm – Co magnets, current forecasts are that capacity increases will make 4000 – 5000 t/a of Nd available in the near future, enough to make more than 10 000 t of Nd – Fe – B magnets.

Rare earth permanent magnets based on Sm – Co and Nd – Fe – B are manufactured by powder sintering [54]. Other process variants exist for the production of Nd – Fe – B magnets; these are described in Section 4.2.2.

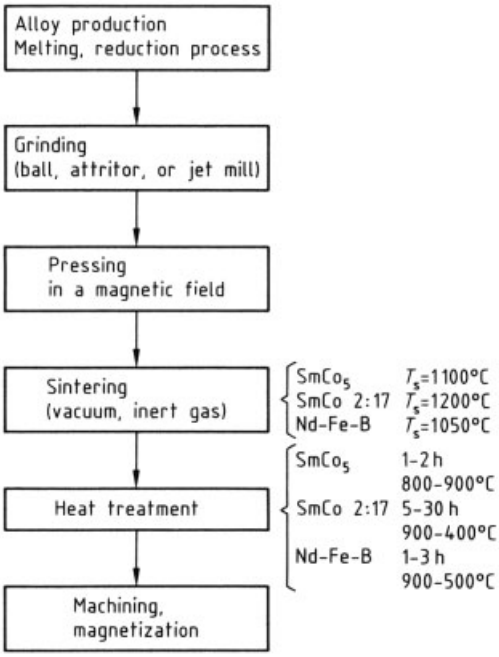

Figure 13. Flow sheet for the production of rare earth permanent magnets

The essential steps in the powder-metallurgical production process are illustrated in Figure 13. First the alloys are prepared. Two processes are used: melt-metallurgical production from the elements or master alloys, and preparation of the alloys by a reduction and diffusion process (co-reduction). The co-reduction process starts with oxides of the rare earth metals, such as Sm_2O_3 or Nd_2O_3, which are thoroughly mixed with the other alloy constituents and calcium in powder form and are then reduced by the calcium in reactors at ca. 1100 °C [55–57]. The desired alloys are formed directly in this process. The resulting powders, with particle sizes of $\leq 200\,\mu m$, are then fine ground, after the calcium has been washed out. Materials prepared by the melt-metallurgical technique, on the other hand, must first be precrushed and preground. In the case of Nd – Fe – B alloys produced by melt metallurgy (Section 4.2.2), the ingot can be disintegrated by hydrogenation [58]. To initiate this process it is sufficient to expose the material at room temperature to a hydrogen atmosphere of a few hundred millibar. The alloy absorbs hydrogen in an exothermic reaction, and the ingot disintegrates into particles smaller than 100 µm. In this way a powder suitable as fine-grinding feed is obtained, and a high grinding throughput is insured.

Fine grinding of the alloys takes place in ball, attritor, or opposed-jet mills. The mean particle sizes after grinding are 4 – 5 µm for Sm – Co materials and 3 – 4 µm for Nd – Fe – B; in either case, the goal is to obtain single-crystal particles.

Single-crystal powders are charged into the cavity of a press die and pressed into compacts in the presence of a magnetic field, or else they are isostatically compacted, also with a field. The magnetically anisotropic powder particles align with the direction of the applied magnetic field. This orientation is fixed by the compacting action.

Sintering is carried out under vacuum or an inert atmosphere at 1050 – 1200 °C, depending on the material. Between 16 and 20 % shrinkage occurs during sintering. Heat treatments as appropriate to the alloys are carried out afterward in order to produce a microstructure favorable for the coercive force (see Sections 4.2.1 and 4.2.2).

Because the fine-grained alloy powders have a strong affinity for oxygen, grinding, pressing, and sintering, and the handling of the powder and compacts between these operations must be performed under an inert atmosphere.

After heat treatment, the magnets are ground if necessary to meet tolerances. Larger blocks of materials can be cut into smaller parts. Magnets weighing as little as 0.1 g are not a rarity.

4.2.1. Rare Earth Metal – Cobalt Magnets

A number of binary and multicomponent RE – Co materials, with distinct uniaxial magnetic anisotropy, high saturation anisotropy, and a high Curie temperature, satisfy the essential requirements for the fabrication of high-quality permanent magnets.

Figure 14 shows the binary phase diagram for the Co – Sm system [59]. The striking point is the great number of compounds, some of which occur in very small ranges. The 1 : 5 compounds decompose, below certain system-dependent temperatures, into the equilibrium phases existing at those temperatures. Below 800 °C, $SmCo_5$ decomposes into the Sm_2Co_{17} and Sm_2Co_7 phases. However, rapid cooling from 800 to ca. 400 °C insures the existence of $SmCo_5$ as a metastable phase below 400 °C.

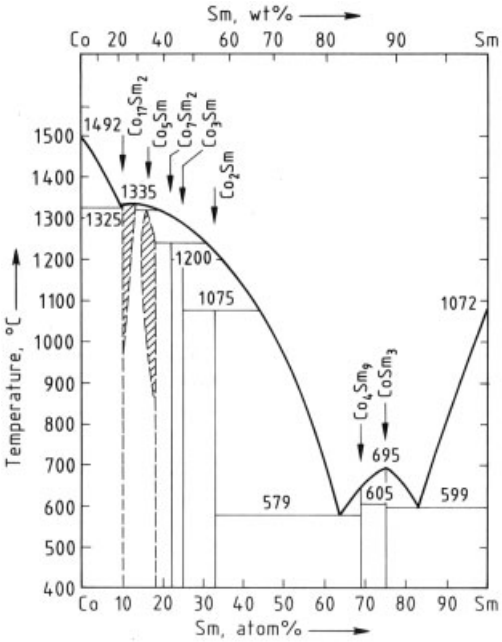

Figure 14. Phase diagram of the Sm – Co system

The phase diagrams for the other RE – transition-metal systems are very similar to the Co – Sm diagram, though not all the systems display identical compositions; for example, the compound $SmFe_5$ does not exist.

Figure 15 shows the saturation polarization for the $RECo_5$, $(RE)_2Co_{17}$, $(RE)_2Fe_{17}$ and $(RE)_2Fe_{14}B$ groups [60], the last of which is the crucial phase in Nd – Fe – B magnets (Section 4.2.2). The light metals from Ce to Sm become ferromagnetically coupled with the transition metals Co and Fe; hence these compounds exhibit a high saturation polarization. In contrast, the heavy rare earth metals from Eu to Lu couple antiparallel to the transition metal moments, so that some of these compounds have very low saturation polarizations.

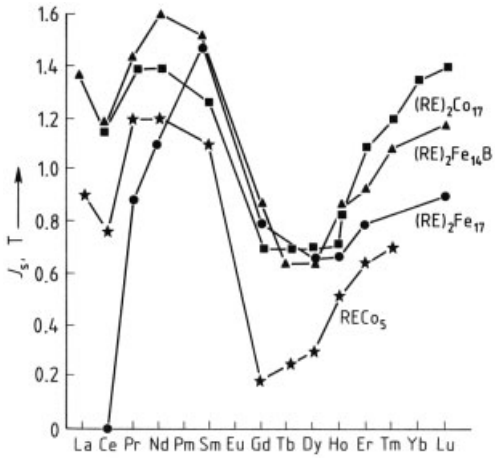

Figure 15. Saturation polarization J_s of $RECo_5$, $(RE)_2Co_{17}$, $(RE)_2Fe_{17}$, and $(RE)_2Fe_{14}B$ at room temperature

Figure 16 gives plots of Curie temperature for the compounds $RECo_5$, $(RE)_2Co_{17}$, $(RE)_2Fe_{17}$ and $(RE)_2Fe_{14}B$ [60]. The rare earth elements exhibit only very weak magnetic ordering and so have low Curie temperatures. Gadolinium has the highest Curie temperature of the whole group (T_C= 16 °C).

When the transition metals Ni, Fe, and Co – with respective Curie temperatures of 358, 768 and 1121 °C – combine with rare earth metals, the strong magnetic coupling of the 3 *d* electrons in the former is imposed on the latter through a pronounced 4 *f* – 3 *d* interaction.

The 1 : 5 compounds crystallize in a hexagonal structure of the $CaCu_5$ type. The 2 : 17 compounds occur in two uniaxial structural modifications, one hexagonal of Th_2Ni_{17} type, and one rhombohedral of Th_2Zn_{17} type (see Fig. 17) [61], [62]. The relationship of these to the 1 : 5 structure can be seen in the fact that both modifications can be built up by combining structural elements of the 1 : 5 lattice.

With the uniaxial magnetocrystalline anisotropy in the hexagonal or rhombohedral

lattice, the polarization vector points in the direction of the crystallographic c axis. This is the case in nearly all 1 : 5 compounds, at least at room temperature. In the binary 2 : 17 alloys, however, uniaxial anisotropy is the exception rather than the rule; the polarization is usually perpendicular to the hexagonal or rhombohedral axis. The polarization is free to rotate in the plane, meaning that high coercive forces cannot be achieved [63].

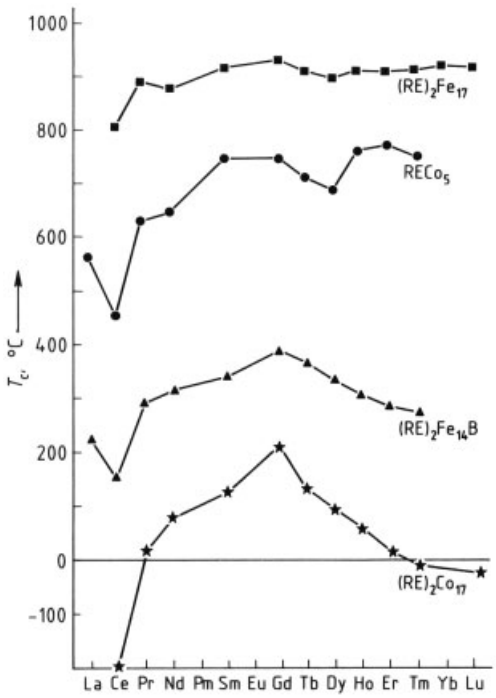

Figure 16. Curie temperature T_c for $(RE)_2Co_{17}$, $(RE)_2Co_5$, $(RE)_2Fe_{17}$, and $(RE)_2Fe_{14}B$

In many 2 : 17 compounds, the inclusion of various elements makes it possible to convert the planar anisotropy to uniaxial. Other potential hard magnetic alloys can be obtained in this way. Figure 18 shows how the direction of anisotropy depends on composition for several ternary 2 : 17 systems [60].

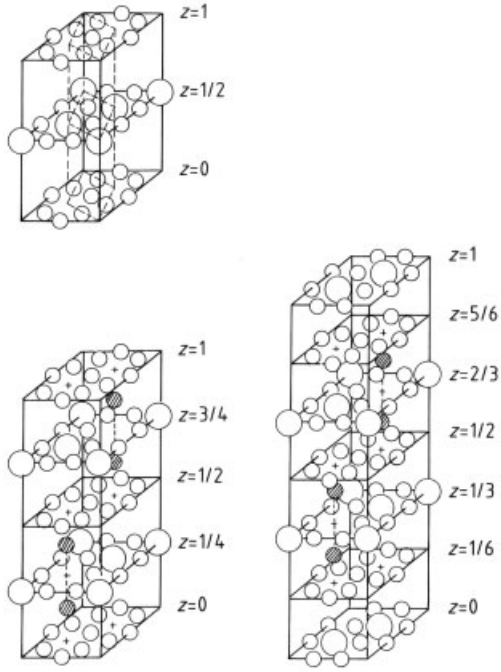

Figure 17. Structures of the 1 : 5 and 2 : 17 compounds
Big circle: = Rare earth atom; Hatched circle, Small circle = Cobalt or a metal atom

The magnitude of the magnetocrystalline anisotropy is expressed in the anisotropy constant K_1, which is positive for uniaxial and negative for planar anisotropy. The anisotropy constant is a measure of the work per unit volume that must be done to rotate the magnetic polarization from the preferential direction it assumes spontaneously (easy direction) into the direction perpendicular to it (difficult direction); for uniaxial anisotropy in the hexagonal lattice, this means from the c direction into the basal plane.

In the case of uniaxial anisotropy, the maximum possible coercive force or the anisotropy field strength H_A can be calculated from K_1 and the saturation polarization J_s [86]:

$$H_A = \frac{2K_1}{J_s} \tag{11}$$

Table 10 lists values of J_s and H_A for several 1 : 5 and 2 : 17 combinations. The large H_A of 31 MA/m for $SmCo_5$ and the comparatively small H_A of 6.8 MA/m for Sm_2Co_{17} are noteworthy [64–68].

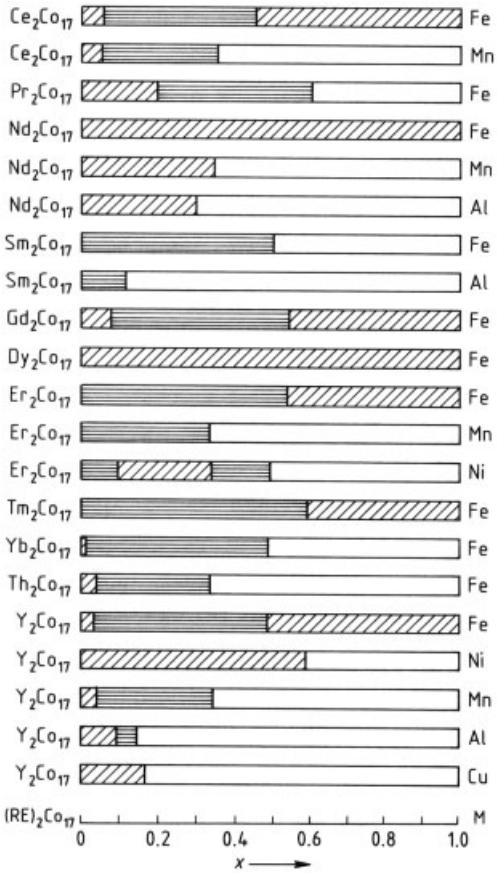

Figure 18. Dependence of preferred magnetic direction on composition for ternary $(RE)_2(Co_{1-x}M_x)_{17}$ alloys
Hatched = Easy plane; Horizontally striped = Easy axis; Plane = Mixed phases or unknown

Table 10. J_s and H_A values for some 1 : 5 and 2 : 17 material combinations [54]

Composition	J_s, T	H_A, MA/m
YCo_5	1.06	10.4
$LaCo_5$	0.91	14.0
$CeCo_5$	0.85	12.5
$PrCo_5$	1.20	13.9
$SmCo_5$	1.12	31.0
Sm_2Co_{17}	1.28	6.8
$Sm(Co, Cu)_{7.8}$	1.09	6.1
$Sm(Co, Fe)_{8.5}$	1.46	4.7

The coercive force H_c corresponding to H_A is found for the ideal case of a reversal of the magnetization by means of a coherent rotation of the polarization vectors. For real crystals with their various imperfections and inhomogeneities, however, this mechanism is not applicable. Remagnetization under the action of external fields generally occurs by nucleation of domain walls and their movement through the crystal, or else when walls already present break free from their positions and move. In the first case, the resulting coercive force is the field that can generate remagnetization nuclei and cause domain walls to move; in the second case, it is the field that can release domain walls from their pinning centers and shift them [54].

In real materials, the coercive force is governed by the resistance to domain-wall formation and/or the force pinning the existing walls and thus preventing their movement. The coercive forces of magnetic polarization produced by these mechanisms have been found to be of the order of $H_{cJ} < 0.3\,H_A$.

Both mechanisms – nucleation and pinning – occur in the RECo group of hard magnetic alloys, depending on composition. The nucleation mechanism governs H_{cJ} in the $RECo_5$ and the single-phase $RECo_{8-8.5}$ materials. These are built up from single-domain particles, and the grain walls hinder the movement of domain walls from grain to grain.

The virgin magnetization curves for materials with the nucleation mechanism rise steeply as a function of the magnetizing field strength, reaching the saturation polarization at relatively low fields. Obviously, domain walls are being completely displaced from the grain volume, against no great resistance, or else pinned fast in surface zones. Curve (a) in Figure 19 is the virgin curve for an $SmCo_5$ magnet [69].

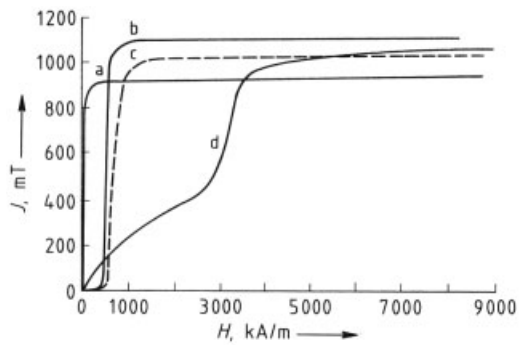

Figure 19. Virgin magnetization curves for 1 : 5 and various 2 : 17 magnets [54]
a) $SmCo_5$; b) $Sm(Co, Fe, Cu, Zr, Ti)_7$ low coercivity; c) $Sm(Co, Fe, Cu, Hf)_7$; d) $Sm(Co, Fe, Cu, Zr)_{7.5}$ high coercivity

In 1 : 5 and 2 : 17 alloys with copper, having compositions of $Sm(Co, Cu)_5$ and $Sm(Co, Cu, Fe, M)_{6.8-7.6}$ (M = Zr, Hf, Ti), the pinning mechanism determines H_{cJ}. Since the $Sm(Co, Cu)_5$ alloys are of no industrial significance because of their relative low magnetic values (see Table 11), the discussion that follows will center on 2 : 17 alloy types containing copper.

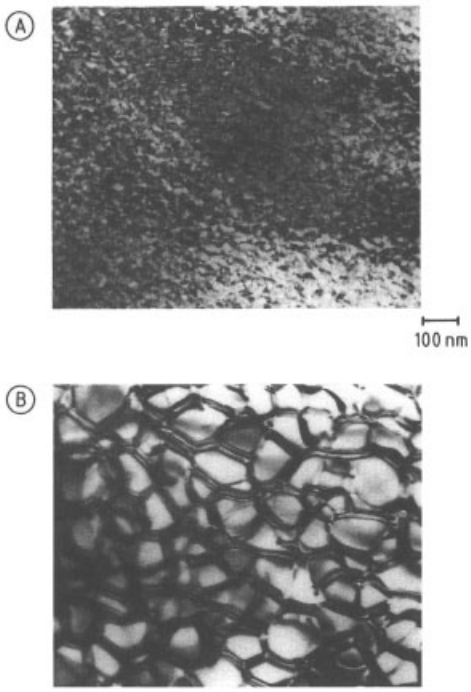

Figure 20. Precipitation microstructure of low- and high-coercivity Sm(Co, Cu, Fe, Zr) 2 : 17 alloys
A) Low coercivity; B) High coercivity

In the course of a composition-dependent heat treatment, the $Sm(Co, Cu, Fe, M)_{6.8-7.6}$ alloys (M = Zr, Hf, Ti, etc.) form a two-phase microstructure with a cobalt- and copper-rich matrix having the 2 : 17 structure and coherent precipitates of a copper-rich phase with the 1 : 5 structure. This two-phase structure, in which the phases differ with respect to saturation polarization and anisotropy, is responsible for the pinning of the domain walls.

Figure 20 shows transmission electron micrographs of the two-phase structure for alloys with markedly different coercive forces: Figure 20 A is for an $Sm(Co, Cu, Fe, Zr)_{6.8}$ alloy with $H_{cJ} \approx 560$ kA/m, and Figure 20 B is for an Sm-(Co, Cu, Fe, Zr)$_{7.5}$ alloy with $H_{cJ} > 1800$ kA/m. The cell diameters of the 2 : 17 phases are 40 – 60 nm in the first case and 100 – 120 nm in the second. For a quantitative understanding of the relationship between precipitate particle size and precipitate structure, see [70], [71].

The submicroscopic picture of the magnetically hardenable SmCo two-phase structure is very similar to the microstructure of AlNiCo and Cr – Fe – Co alloys in the magnetically optimized condition. The principal difference is the type of anisotropy. In AlNiCo alloys, for example, strongly magnetic Fe – Co-rich particles with shape anisotropy are created in a weakly magnetic Ni – Al-rich matrix; in the RE – Co alloys, strongly magnetic, crystallographically anisotropic zones are surrounded by weakly magnetic or nonmagnetic copper-rich phases.

Figure 19 shows the virgin curves of 2 : 17 alloys having low and high coercive force. The shape of the virgin curves for the precipitation-hardened 2 : 17 materials is quite different from that for $SmCo_5$ [54]. The low-coercivity 2 : 17 compounds show a rapid increase of polarization in magnetizing fields of the order of their H_{cJ} values. The polarization of the high-coercivity 2 : 17 alloy increases relatively slowly up to fields of about 2500 kA/m and does not begin to rise sharply until the range of 3000 – 4000 kA/m [54].

The various mechanisms responsible for generating the coercive force are manifest not only in the virgin curves but also in the demagnetization curve (plotted versus magnetizing field strength).

At magnetizing fields as low as 800 – 1000 kA/m, $SmCo_5$ magnets reach ca. 99 % of their remanence. The formation of the coercive force, however, requires fields of 2400 – 3000 kA/m. Only the action of these fields leads to the removal of domain walls out of the grains or their pinning in surface zones.

Low- coercivity 2 : 17 magnets with H_{cJ} values of 500 – 800 kA/m require field strengths of ca. 1000 – 1500 kA/m for complete magnetization, whereas high-coercivity 2 : 17 materials with coercive forces > 2000 kA/m require fields of ≥ 4000 kA/m for magnetization. The effect of magnetizing field strength on the shape of the demagnetization curve is shown in Figure 21 for $SmCo_5$ and high-coercivity 2 : 17 magnets. The

poor magnetizability of the latter leads to problems where, for reasons of manufacturing technology, the parts are not magnetized until they have been installed in the system. In most cases, the required magnetizing devices are not available. Magnetization at elevated temperature (see Fig. 22) makes possible a significant reduction in the field strength required for complete magnetization. This dependence is utilized in some instances.

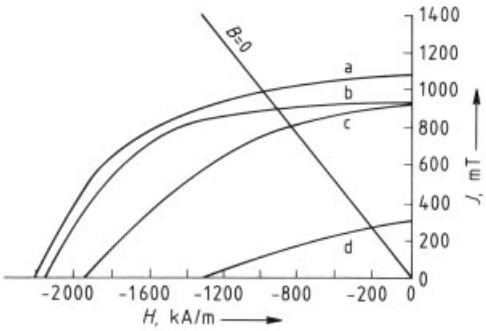

Figure 21. Effect of the magnetizing field strength on the demagnetization curves of 1 : 5 magnets and high-coercivity 2 : 17 magnets
a) 2 : 17, $H = 4800$ kA/m; b) 2 : 17, $H = 1000$ kA/m; c) 1 : 5, $H = 2800$ kA/m; d) 1 : 5, $H = 1000$ kA/m

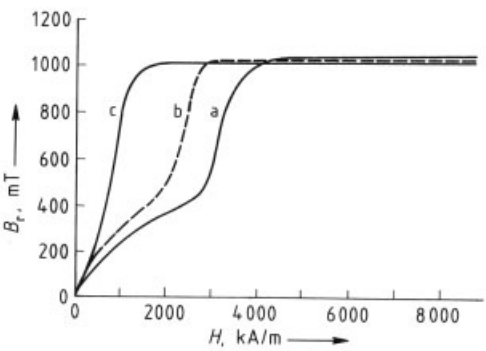

Figure 22. Dependence of remanence B_r on magnetic field strength and temperature [54]
a) Magnetized at room temperature; b) Magnetized at 200 °C; c) Magnetized at 400 °C

Table 11 gives a representative summary of the compositions and magnetic values for RE–Co hard magnetic materials. The 2 : 17 magnets contain ca. 10 % less samarium and 15 % less cobalt than $SmCo_5$, while the B_r and $(BH)_{max}$ values are higher.

Characteristic temperature data for $SmCo_5$ and 2 : 17 SmCo magnets appear in Table 4. As the temperature rises, 2 : 17 magnets display smaller magnetization losses than $SmCo_5$, and their maximum service temperature is also higher. The long-term thermal stability, measured by the relative flux loss at room temperature after holding at 200 and 300 °C, is illustrated by Fig. 23 [41], [54].

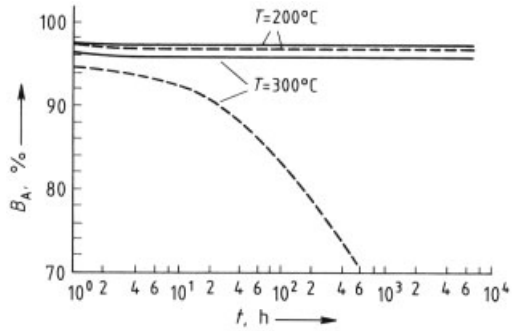

Figure 23. Comparison of the long-term thermal stability of 1 : 5 and 2 : 17 magnets
– – – = 1 : 5 magnets; —·— = 2 : 17 magnets; B_A = relative residual flux density; t = ageing time $B/\mu_0 H = -1$

In limited temperature ranges, the temperature coefficients of remanence Tk (J_s) for SmCo magnets can be reduced to ca. −0.002 to +0.002 %/K by the addition of small amounts of heavy rare earth metals [72], [73].

Besides sintered magnets, SmCo materials are also available in resin-bonded modifications (as are hard ferrites). These can be fabricated by injection or compression molding. The materials have maximum service temperatures of 80–100 °C (resin-bonded $SmCo_5$) and 120–140 °C (resin-bonded 2 : 17 SmCo).

4.2.2. Rare Earth Metal – Iron – Boron Magnets

Essential in permanent magnets based on RE–Fe–B is a tetragonal phase of composition $(RE)_2Fe_{14}B$, which exhibits a marked uniaxial magnetic crystal anisotropy for a number of rare earth elements (Table 12) [74], [75]. As in the case of samarium–cobalt materials, this anisotropy is responsible for the coercive force. The relationships between the crystal anisotropy, the saturation polarization, and the maximum possible coercive force (anisotropy field strength H_A) are analogous to those discussed in Section 4.2.1.

Table 11. Composition and magnetic values for rare earth (RE) – Co permanent magnets

Material	Typical composition, wt %								Typical magnetic values			
	Sm	Pr	MM∗	Co	Fe	Cu	Zr	Resin	$(BH)_{max}$, kJ/m^3	B_r, mT	H_{cB}, kA/m	H_{cJ}, kA/m
$(Sm_{0.2}MM_8)Co_5$	7		27	66					120	800	580	1000
$(Sm_{0.4}Pr_{0.6})Co_5$	16	18		66					200	1050	720	1200
$Sm(CoCuFe)_5$	34			49	5	12			110	770	290	300
$SmCo_5$	34			66					160	920	680	2500
$Sm(Co, Cu, Fe, Zr)_{6.9}$	24 – 26			48 – 50	10 – 16	5 – 12	1.5 – 3		240	1130	540	580
$Sm(Co, Cu, Fe, Zr)_{7.5}$	24 – 26			48 – 50	10 – 16	5 – 12	1.5 – 3		210	1080	780	800
$SmCo_5$ (resin bonded)	33		62					2.5	110	760	580	1000

∗ Misch metal (cerium-rich rare earth mixture)

Table 12. Anisotropy constant K_1, saturation polarization J_s, anisotropy field strength H_A, and Curie temperature for several $(RE)_2Fe_{14}B$ compounds [75]

$(RE)_2Fe_{14}B$	K_1, MJ/m^3	J_s, mT	H_A, kA/m	T_c, °C
Ce	1.7	1.17	2069	262
Pr	5.6	1.56	5970	294
Nd	5.0	1.60	5970	317
Dy	4.5	0.71		322
Ho	2.5	0.81	7320	303

Figures 15 and 16 show the saturation polarization J_s and the Curie temperature T_C for $(RE)_2Fe_{14}B$ compounds as functions of the RE component. As in the case of RECo, light RE elements couple ferromagnetically with transition metals and so exhibit higher saturation polarizations than the ferrimagnetically coupling heavy RE elements. Neodymium – iron – boron, the material in this series that has gained the greatest technical and economic importance, has J_s= 1.6 T, substantially higher than the saturation polarizations of $SmCo_5$ (1.07 T) and Sm_2Co_{17} (1.2 T).

The Nd – Fe – B permanent magnets became known in 1983 after parallel but independent development in Japan and the United States [51], [52]. A variety of processes are used for production of magnets.

The process developed in Japan involves powder sintering. The sequence of operations is similar to that for Sm – Co magnets (cf. Fig. 13). Nd – Fe – B alloys are prepared by melt metallurgy from the elements or prealloys (e.g., Nd – Fe), or by co-reduction with oxides (Nd_2O_3) and Ca as reducing agent. The alloys are then ground in ball, attritor, or opposed-jet mills; the product is a single-crystal powder with a mean particle size of 3 – 4 µm [56], [57]. The powder is compacted under an imposed magnetic field, and the compacts are sintered in vacuum or under an inert atmosphere. The optimal magnetic values are set in a subsequent heat treatment. The grinding process can be made much more efficient if the materials are embrittled by hydrogen absorption; the hydrogen is liberated again at temperatures of 200 – 700 °C in the sintering process [58].

The Nd content of Nd – Fe – B magnets is 31 – 34 wt %, higher than the 26.7 wt % Nd in the compound $Nd_2Fe_{14}B$. The magnets thus contain not only the dominant 2 : 14 : 1 compound but also other phases, such as $NdFe_4B_4$ and Nd-rich phases, which are important for sintering (sintering aids) and also for imposing the coercive force. The Nd – Fe – B powders have a higher oxygen affinity than the Sm – Co materials, so that even greater care and caution are necessary in the handling of the powders during the production process.

In the Nd – Fe – B magnet process developed in the United States, the alloys are produced by rapid quenching [52]. When molten alloys are quenched at cooling rates of 10^5 – 10^6 K/s, an amorphous or microcrystalline material is obtained; tempering converts this to a homogeneous microstructure. The polycrystalline, tinsel-like ribbon obtained in this process (30 – 50 µm thick and ca. 2 mm wide) is briefly ground to 200 µm. The powder particles consist of an agglomerate of isotropically distributed single crystals 15 – 150 nm in size.

Isotropic or anisotropic Nd – Fe – B magnets can be made from these materials by a variety of methods. Isotropic plastic-bonded Nd – Fe – B magnets are fabricated by injection molding

with a thermoplastic binder or by compression molding with an epoxy binder. Although these magnets have markedly lower magnetic values than anisotropic sintered magnets (see Fig. 12) because of their isotropy and the dilution of the magnetic material, they exhibit a number of advantages similar to those of plastoferrites. The simple fabrication technique, offering close tolerances and requiring no subsequent mechanical working, and the significantly lower brittleness, in comparison with sintered magnets, mean that resin-bonded Nd – Fe – B magnets are a technically and economically interesting approach for many applications.

The melt-spun material can be processed into compact isotropic magnets by hot-press molding at temperatures between 650 and 750 °C [76]. These have higher magnetic values than the resin-bonded composites discussed above.

Finally, plastic deformation of the isotropic compact magnets at ca. 700 °C produces a texture associated with a magnetic anisotropy [76], [77]. The magnetic values so achieved are comparable with those of anisotropic sintered magnets.

Other processes for making Nd – Fe – B alloys and magnets are now in the laboratory and pilot-plant stage. A mechanical alloying technique [78] begins with the starting materials neodymium, iron, and boron, in elemental form or as prealloys. These are intensively ground in ball or attritor mills, then subjected to a brief reaction anneal. Finely divided crystalline $Nd_2Fe_{14}B$ powders are obtained. Isotropic magnets can be produced by hot-press molding of these powders; anisotropic ones, by subsequent cold extrusion. The properties are comparable with those of magnets produced from melt-spun powders.

In melt-spraying techniques, such as the Ospray process [79] or liquid dynamic compaction (LDC) [80], melts are atomized in a gas stream and deposited on cooled substrates. Isotropic magnetic Nd – Fe – B layers with grain sizes of 1 – 20 µm are obtained. The remanence values are ca. 650 mT and the coercive forces up to 800 kA/m.

The high cost and delicacy of processing oxidation-sensitive Nd – Fe – B alloys in powder-metallurgical methods of magnet production has led to a search for alternatives. One such process involves subjecting ingots to special cold extrusion and cold upsetting treatments [81]: textures with a preferred magnetic direction result. This method is particularly successful with praseodymium – iron – boron – copper alloys [82], [83]; remanences of 1.25 T and coercive forces of around 800 kA/m are achieved.

Alloy compositions and magnetic values for various Nd – Fe – B alloys are listed in Table 13. The magnetic values are for room temperature after complete magnetization. The influence of the magnetizing field strength on the magnetic values of an anisotropic magnet is shown in Figure 24 [85].

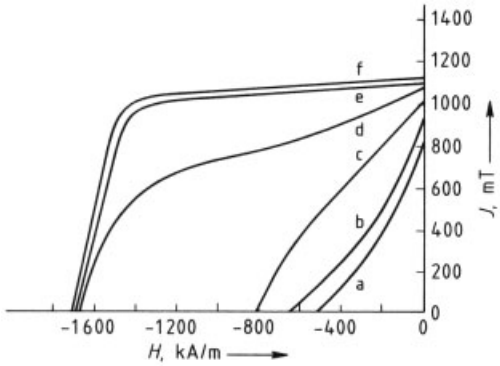

Figure 24. Effect of magnetizing field strength on the shape of the demagnetization curve for Nd – Fe – B alloys
The applied magnetic field strength in kA/m for the curves a – f is as follows:a) 400; b) 500; c) 600; d) 800; e) 1200; f) 2500

The Nd – Fe – B magnets have strongly temperature dependent magnetic properties, as is shown by the temperature coefficients for the reversible change of their J_s and H_{cJ} values (see Table 4). The strong temperature dependence of the coercive field strength means that the maximum permissible operating temperature depends on the coercive field strength at room temperature. Figure 25 shows the demagnetization curve $J(H)$ as a function of temperature for high- and low-coercive magnets [85]; the intercept of the curves with the line $B = 0$ gives the H_{cB} values. It can be seen that the 60 °C demagnetization curve of the low- coercive magnet exhibits a kink before it reaches the H_{cB} value, whereas the curve for the high- coercive magnet is linear on reaching H_{cB}, even at 100 °C. It is possible to produce Nd – Fe – B magnets with $H_{cJ} > 2000$ kA/m that have a linear demagnetization curve up to H_{cB} at 150 °C.

Table 13. Composition and magnetic values for Nd – Fe – B magnets

Production method	Isotropic or anisotropic	Composition, wt %					Magnetic values			
		Nd	Dy	Fe	B	Possible additives (0.5 – 3 %)	B_r, mT	H_{cB}, kA/m	H_{cJ}, kA/m	$(BH)_{max}$, kJ/m^3
Powder metallurgy/sintering	anisotropic	33.0		remainder	1.2		1200 – 1300	600 – 800	800 – 1000	280 – 330
	anisotropic	30.5	2.5	remainder	1.2	Al, G, V, Nb	1150 – 1200	900 – 950	1100 – 1300	260 – 280
	anisotropic	29.0	4	remainder	1.2		1050 – 1150	830 – 900	1800 – 2100	200 – 250
Melt spinning (extrusion)∗	isotropic	29		remainder	0.9		450 – 600	300 – 400	1000 – 1200	35 – 60
	isotropic	29		remainder	0.9		780 – 820	480 – 550	1100 – 1200	90 – 110
	isotropic	29		remainder	0.9		1170 – 1200	850 – 900	1000 – 1100	240 – 280

∗ Commercial product name of the powder = Magnequench.

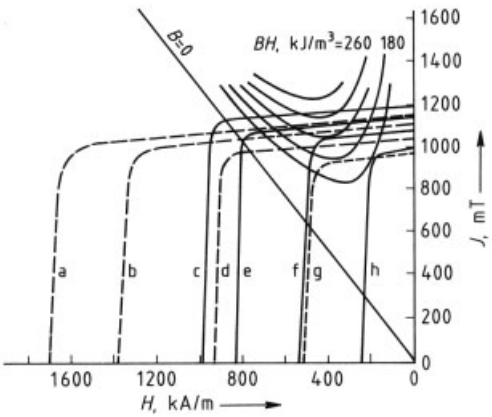

Figure 25. Demagnetization curves of a high- and low-coercitive Nd – Fe – B magnet at varying temperature
a) 25 °C; b) 60 °C; c) 25 °C; d) 100 °C; e) 60 °C; f) 100 °C; g) 150 °C; h) 150 °C – – – high coercivity; —- low coercivity

An increase in the coercive field strength therefore increases the maximum permissible operating temperature and widens the range of applications for Nd – Fe – B magnetic materials. Despite the greater temperature dependence of their magnetic properties, as compared to Sm – Co magnets, the Nd – Fe – B magnets have $(BH)_{max}$ values superior to those of SmCo 1 : 5 and even SmCo 1 : 17 magnets over a wide temperature range (see Fig. 26).

Apart from shearing effects and thermal influences, the chemical resistance of the Nd – Fe – B alloys is important for their long-term stability. Although corrosion is negligible at room temperature and normal humidity, Nd – Fe – B magnets rapidly form rust spots at elevated temperature and humidity. The magnets can be protected by coating the surface with metals, paints, or plastics.

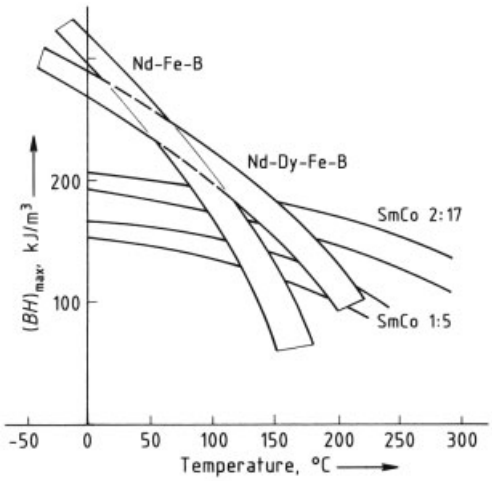

Figure 26. Temperature dependence of the $(BH)_{max}$ values for various magnetic materials

5. Soft Magnetic Materials

The ferromagnetic elements iron, nickel, and cobalt are the simplest representatives of the soft magnetic materials. Of these, however, only iron is industrially important in elemental form, because of its availability and low price.

The essential requirements for achieving the highest possible magnetic values in soft mag-

Table 14. Physical and magnetic properties of soft magnetic alloys

Material	Alloy class according to IEC 404-1	Composition in addition to iron, mass % or chemical formula, indices give atom %	Density, g/cm^3	Curie temperature, °C	Saturation flux density, T	Coercive force, A/m	Specific electrical resitivity, μ Ω cm
Classical soft magnetic materials							
Pure iron	A		7.80	770	2.15	60 – 240	15
Iron – silicon							
nonoriented	C 1, C 21	2.0 – 4.5 (Si)				12 – 24	
grain-oriented	C 22, C 23	2.5 – 3.5 (Si)	7.65	750	2.00	10	45
Iron – nickel	E 5	ca. 30 (Ni)	8.15	40 – 110	0.15 – 0.70	20 – 40	80
	E 4	35 – 40 (Ni)	8.15	250	1.30	16 – 40	75
	E 3	45 – 50 (Ni)	8.25	470	1.55	4 – 12	45
	E 2	54 – 68 (Ni)	8.25	500	1.50	1.2 – 8	40
	E 1	72 – 83 (Ni)	8.65	400	0.80	0.4 – 2.4	55 – 60
Iron – cobalt	F 1	47 – 50 (Co)	8.15	950	2.35	80	35
Soft ferrite	N, N 1		5.00	150 – 500	0.30 – 0.50	10 – 1000	$10^7 - 10^8$
Modern soft magnetic materials							
Powder-bound composite		1 (plastic)	7.00	irrelevant	1.80	300 – 400	
Amorphous	(I)						
Fe-based	(I 1)	$Fe_{81}B_{13}S_4C_2$	7.30	400 – 500	1.50 – 1.60	2.0 – 4.0	100 – 150
Fe – Ni-based		$Fe_{39}Ni_{39}Mo_4Si_6B_{12}$	270	0.70 – 1.00	0.6 – 4.0	100 – 150	
Co-based	(I 2)	$Co_{66}Fe_4(MoSiB)_{30}$	7.80	200 – 500	0.55 – 1.00	0.2 – 1.0	100 – 150
Nanocrystalline		$Fe_{74}Cu_1Nb_3Si_{14}B_9$		ca. 600		< 1.0	

Table 15. Dynamic magnetic properties (50 Hz) of soft magnetic alloys (for composition see Table)

Material	Band width, mm	Permeability		Losses, W/kg			
		μ_4	μ_{max}	$P*_{0.2}$	$P_{0.5}$	$P_{1.0}$	$P_{1.5}$
Classical soft magnetic materials							
Iron – silicon							
nonoriented	0.5	1000	30 000			1 – 3	2.0 – 4.0
grain-oriented	0.35	2000	35 000			0.3 – 0.6	1.0
Iron – nickel							
E 4	0.3	3000	16 000			0.55	
E 3	0.1	12 000	80 000			0.15	
E 2	0.1	80 000	150 000			0.10	
E 1	0.1	250 000	400 000		0.005		
Iron – cobalt	0.3	1000	11 000				2.4
Soft ferrite	massive	10 – 10 000					
Modern soft magnetic materials							
Powder-bound							
composite	massive	10 – 100					
Amorphous							
Fe-based	0.025		150 000	10 – 30			0.23
Fe – Ni-based	0.025		250 000	9		0.10	
Co-based	0.025	150 000	400 000	9 – 10			
Nanocrystalline	0.025	$> 100\,000$					

$*$ 20 kHz

netic systems are maximal purity and optimal microstructure. Purity, grain size, and texture are the main factors controlling the mobility of domain boundaries and, therefore, the magnetizability of the material. In elemental systems, these are also the only quantities that can be controlled in material fabrication and treatment. Impurities and grain structure are often related in such a way that impurity particles prevent the desired formation of large grains or allow the formation of only an inadequate texture.

In the case of iron, in particular, the elements carbon, nitrogen, sulfur, and oxygen are critical for the secondary magnetic parameters (coercive force and permeability). As a result, the magnetic values of even low-carbon steels are markedly lower than those of the pure iron (Armco iron) used in soft magnetic applications, which must exhibit carbon contents well below 0.01 wt %. The content of impurities also influences the degree of cold work needed to produce optimal magnetic properties.

In order to minimize the level of impurities in the material, special attention must be directed toward the initial purity and the amounts of substances added during processing to effect deoxidation and improve the hot or cold workability; the atmospheric conditions during working and casting must also be considered. The final microstructure can be influenced by a well-defined amount of hot and cold working.

Final annealing of the finished part under suitable thermal and atmospheric conditions leads to further reductions in impurities and the formation of the optimal grain structure. The same processing steps outlined for pure iron also apply to the melt-metallurgical production of all soft magnetic materials, although they are overlain there by many other controlling variables.

The main uses for soft magnetic pure iron are in formed or massive parts for relays, electromagnet pole shoes, and shielding enclosures. Because of its low resistivity, pure iron finds only minor use as a flux-guiding stator or rotor material in electric motors and generators. With small amounts of elements such as silicon added to increase the resistivity, it is marketed as low-cost unalloyed electric sheet for applications that impose no special loss requirements.

Binary alloys based on at least one of the three ferromagnetic elements, such as Fe – Si, Fe – Ni, and Fe – Co, as well as multicomponent systems based on Ni – Fe, have long been established in industry and so can be regarded as classical soft magnetic materials. They are manufactured by melt metallurgy with subsequent mechanical working. In special cases, they can also be produced by powder methods involving molding and sintering. The second category also includes the soft ferrites, whose resistivities, orders of magnitude higher than those of metallic materials, make them suitable for high-frequency applications.

The designation "classical" is not meant to detract from the current importance of such materials; it simply indicates that they form an industrial tradition going back decades in some cases. Continuing development (chiefly in the manufacturing area) on the original basic alloys, whose compositions were selected in order to optimize specific physical properties, has brought about a steady improvement in magnetic values. The use of these materials for inductive components in electrical and electronic circuits is firmly established, and they are found everywhere in the market, in a range of application-specific variations differing in magnetic properties and core geometry.

In contrast, modern soft magnetic materials include innovative developments of recent years. First place among these belongs to amorphous materials produced by rapid solidification. Amorphous cobalt-based alloys, with their low magnetostriction, represent a further distinct improvement in magnetic values over the highest-permeability crystalline Ni – Fe-based alloys.

In the area of small transformers, for example, where operating frequencies are high, these materials make it possible to produce smaller components so as to meet user demand for the miniaturization of inductive elements. Furthermore, special new types of magnetic sensors for control applications and electronic surveillance could not be created until these materials became available.

It is not possible to foresee at present how much acceptance these materials, which are metastable, in contrast to the crystalline materials, will gain for classical applications. Their production costs are still high; their thermal stability, relatively poor; and there is some uncertainty about the long-term behavior of their magnetic values.

In addition, powder composites, which had their origins (in iron-powder cores) around the turn of the century, are described in a modern resin-bonded variant used for dynamic flux-guiding components in electric motors and for choke cores.

Table 14 lists physical and magnetic properties of soft magnetic alloys, Table 15 gives their dynamic magnetic properties, and the magnetic characteristics can be found in Table 16.

Table 16. Magnetic characterization of soft magnetic materials

Material	Characteristic features
Classical soft magnetic materials	
Pure iron	high saturation flux density, cheap material for static applications
Iron – silicon	high saturation flux density, cheap material for dynamic applications
nonoriented	isotropy of the magnetic properties
grain-oriented	anisotropy of the magnetic properties (easy axis); low losses
Iron – nickel	
E 5	strong temperature-dependance of the flux density under 100 °C
E 4	small increase in permeability at low and medium field strengths
E 3	saturation maximum in the Fe – Ni system; texture material with rectangular curve
E 2	high permeabilities at low driving fields after magnetic field annealing
E 1	very high permeabilities at low driving fields; variation of the curve shape
Iron – cobalt	saturation maximum less than for the relevant soft magnetic alloys
Soft ferrite	low eddy current losses due to very high specific electrical resistivity
Modern soft magnetic materials	
Powder-bound composite	low eddy current losses due to interparticular isolation; mechanical treatment possible
Amorphous	useful soft magnetic behavior due to lack of crystal anisotropy
Fe-based	cheap alloys with low losses
Fe – Ni-based	magnetic values between Fe- and Co-based alloys
Co-based	high magnetic values due to zero magnetostriction
Nanocrystalline	new types of Fe-based alloys with good magnetic values

5.1. Classical Soft Magnetic Materials

Classical metallic soft magnetic materials are single-phase homogeneous mixed crystals, sometimes with (orientational) superlattice structures. The addition of one or more alloying elements to a ferromagnetic base element makes it possible to control the physical properties of the system in general and, in particular, the magnetic properties. The latter are influenced by magnetic anisotropies inherent in the structure, which lead to a preferred alignment of magnetic moments in the individual domains. There are five types of anisotropy:

1) Magnetic crystal anisotropy (K_1)
2) Magnetoelastic anisotropy ($\lambda_s\sigma$)
3) Induced uniaxial anisotropy (K_u)
4) Form anisotropy
5) Exchange anisotropy

The first three are especially important for soft magnetic materials. They are described in terms of the constants K_1 (aligning effect of the crystal lattice in J/m^3), λ_s (aligning effect of tensile or compressive stress), and K_u (effect of directional ordering in J/m^3). They are influenced by imperfections in the lattice, such as foreign bodies or internal stresses.

The anisotropies have energies associated with them, and an external magnetic field H must overcome these energies in order to rotate the magnetic moment vectors away from their preferred (easy) direction. The stronger the energetic coupling to the easy direction, the more difficult it is to remagnetize a magnetic region (such as a single domain) and the higher the coercive force.

In the case of the anisotropy associated with crystal lattice structure (constant K_1), the coercive force for a favorably oriented particle is [86]

$$H_A = 2K_1 J_s$$

The magnetoelastic anisotropy describes how much the magnetization is affected by mechanical stresses; if the crystal anisotropy is neglected, then [87]

$$H_{cJ} = s\lambda_s\sigma / J_s$$

where λ_s is the saturation magnetostriction, σ is the mean internal stress, and s is a constant ($0 < s < 1.5$). The saturation magnetostriction gives the relative elastic strain $\Delta l/l$ of a polycrystalline solid in a saturating magnetic field.

Soft magnetic materials, which are characterized by low coercive forces, must accordingly have small values of both the crystal anisotropy and the product $\lambda_s\sigma$ in the direction of the driving field (working direction). This is achieved by a suitable choice of alloy composition in conjunction with final annealing of the finished part in order to establish an optimal microstructure. Depending on the magnetic level of the soft magnetic part (e.g., shielding or coil core), the annealed part may be subjected to only a limited amount of mechanical working or loading, or to none; otherwise, plastic deformation may result

in new magnetic hardening. Dislocations generated in this way, like impurity particles, precipitates, and grain boundaries, lead to undesirable interactions with the domain boundaries.

Because the classical soft magnetic materials are all polycrystalline, minimizing the anisotropy effects in the working direction always involves either finding compromises or creating the most isotropic structure possible.

The crystal anisotropy can be used to control the shape of the hysteresis curve. If, for example, recrystallization annealing is used to produce a distinct texture in the strip direction in a cold-worked strip material, then the hysteresis loop measured in this direction will be extremely rectangular ($B_r/B_s \lesssim 1$).

Induced uniaxial anisotropy, which can be generated by tempering in a magnetic field, involves one-dimensional orientational superlattice structures of like atoms. This kind of anisotropy can also be used to control the shape of the hysteresis curve.

Magnetic anisotropies thus provide an important way of manipulating the secondary magnetic parameters of soft magnetic materials. But the effectiveness of these measures depends on the fabricated part having a geometry adapted to the driving field direction; this requirement leads to standardized core shapes.

5.1.1. Iron – Silicon Materials

Electrical power technology uses generators, motors, chokes, transducers, and transformers to produce, transform, and distribute electrical energy. The windings carry relatively large electrical currents at a constant voltage in the industrial frequency range (usually 50 or 60 Hz line frequency), and magnetic fields or magnetic fluxes result from these currents. Guiding the a.c. magnetic fluxes as efficiently as possible means using easily magnetizable, low-loss soft magnetic materials with high usable flux densities. This application requires large amounts of material, so that the first criterion for material selection is low cost. The term electric sheet covers all the iron-based core materials used in this field.

The technically and economically effective dynamic use of soft magnetic materials in electrical machinery and transformers requires low magnetic losses and high efficiencies. Pure iron satisfies these requirements to only a limited degree because of its low resistivity, and so it has been necessary to find low-cost alternatives.

The magnetic loss P_{Fe} of a magnetic material is the sum of hysteresis losses P_H (the area enclosed by the static hysteresis loop) and eddy-current losses P_E:

$$P_{Fe} = P_H + P_E$$

According to classical eddy-current theory, derived from Maxwell's equations and neglecting the domain structure, the eddy-current losses P_E (in W/kg) for a sinusoidal flux uniformly distributed in a sheet with density d is

$$P_E = \frac{\left(\pi D \hat{B}\right)^2}{6\varrho d} \cdot f^2$$

Small sheet thicknesses D and high resistivities ϱ are therefore necessary to reduce eddy-current losses.

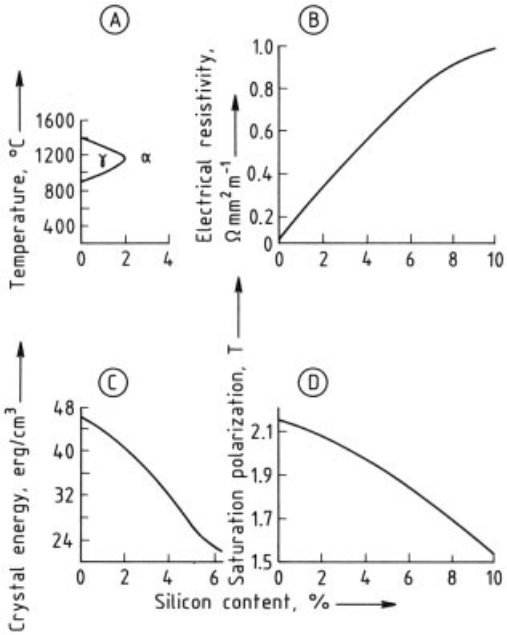

Figure 27. Effect of silicon in iron on the transition temperature γ-α (A); on the electrical resistivity (B); on the crystal anisotropy (C); and on the saturation polarization (D) [91]

The addition of silicon to iron gives a marked increase in resistivity (see Fig. 27), but at the cost of the saturation flux density. Since silicon forms a homogeneous solution in iron up

to a few wt % (see Fig. 28), no magnetic hardening occurs. Furthermore, silicon has the advantage of reducing crystal anisotropy. As the silicon content increases, however, the alloy loses workability, especially at room temperature, so that high-silicon electric sheet can be fabricated only by hot rolling. The commercial use of over 4.5 wt % Si is thus no longer attractive.

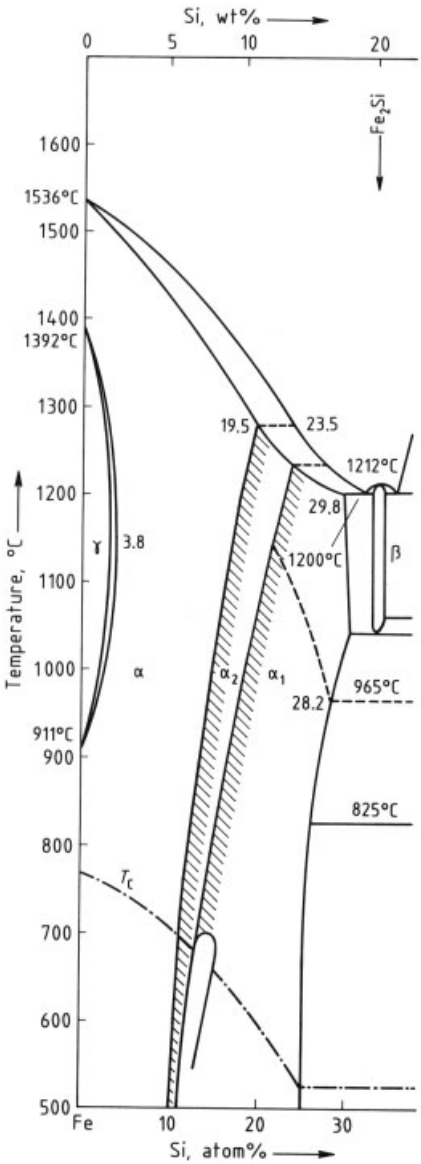

Figure 28. Iron-rich side of the Fe – Si phase diagram [92]

Siliconized electric sheet is available in thicknesses down to 0.05 mm. Such extremely thin sheets, however, are reserved for higher-frequency applications, where they are used in tape-wound toroidal or C (cut) cores for repeater coils, chokes, and other devices. A secondary use is in stampings for the power area, which involves sheets of standardized thicknesses down to 0.23 mm. Standard series of material qualities are differentiated by sheet thickness and loss values [88], [89].

Iron – silicon electric sheet materials are classified in two groups according to the crystallographic orientation of the strip material in the final annealed condition:

1) Non-oriented (NO) material
2) Grain-oriented (GO) material

Because of their isotropic magnetic values in the sheet plane, NO materials find use chiefly in rotating electrical machinery, while the anisotropic GO materials are employed for cores in power and distribution transformers [90]. Cores made of laminated sheets with insulation between layers come in the usual geometries.

Important gains in magnetic properties have been achieved for both classical electric-sheet groups, not least because of potential competition from novel amorphous iron-based alloys starting in the 1970s and, later, from high-silicon nanocrystalline alloys produced by rapid solidification.

5.1.1.1. Non-Oriented Electric Sheet

The fabrication of electric sheet from ordinary steel began in the late 1800s with the advent of electrification. Silicon was first used as an alloying element at the beginning of the 1900s. Only hot-rolled low-silicon sheet with a relatively high carbon content was available at first.

However, because hot rolling is a costly process for strip production, and because such strip is not as high in quality (purity, dimensional stability, surface finish) as cold-rolled strip, attempts to make cold-rolled electric sheet began soon afterwards. Today, electric sheet containing up to 3.5 wt % Si is generally made by cold rolling [90].

The addition of 0.3 – 0.6 % Al further reduces the magnetic loss [93], [94] by enhancing the resistivity, diminishing the crystal anisotropy, in-

creasing the grain size, and producing a beneficial change in the structure and size of inclusions [95]. Commercial NO electric sheet available today is generally an Fe – Si – Al alloy. Its production involves the following steps:

1) Melting the alloy by special methods to increase the purity [96]
2) Casting into ingots
3) Hot rolling to intermediate thicknesses of 1.8 – 2.5 mm
4) Pickling
5) Cold rolling to final thicknesses of 0.35 – 0.65 mm
6) Final annealing in two stages: a) decarburizing annealing in a humid atmosphere between 800 and 900 °C and b) coarse-grain recrystallization annealing above 1000 °C

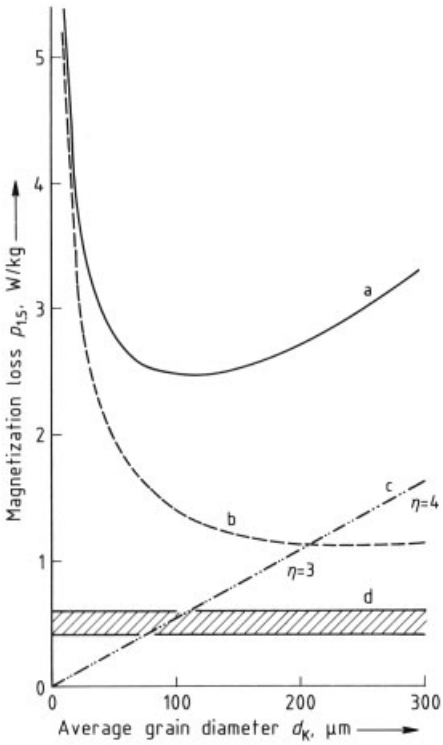

Figure 29. Effect of the average grain diameter on the magnetic loss [98]
a) Magnetic loss $P = P_H + P_E + P_A$; b) Hysteresis loss P_H; c) Abnormal loss P_A; d) Classical eddy current loss P_E (0.41 – 0.60 W/kg depending on the alloy grade)

Good NO electric-sheet qualities with low losses at high flux densities ($\hat{B}$ = 1.0 and 1.5 T) must meet stringent requirements for texture, grain size, purity and surface finish of the strip, along with the secondary condition that the silicon content be as high as possible. In particular this means [97], [98]:

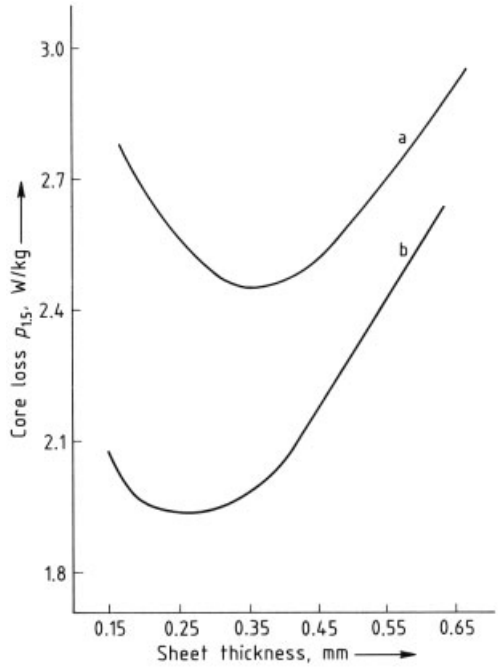

Figure 30. Magnetic loss as a function of sheet thickness for two surface states
a) Dry annealing; b) Two-stage (wet + dry) annealing

1) Establishment of a magnetically isotropic recrystallization texture in the strip plane– the cube-face texture (100) [0kl], also called a "wagonwheel" texture.
2) Recrystallization annealing (see Fig. 29) to produce a grain size (0.10 – 0.15 mm) optimal for the later service conditions. This step results in either optimization of the loss for a given flux density or optimization of the permeability.
3) Attainment of high purity, in particular with extremely low sulfur, oxygen, nitrogen, and carbon contents (in the ppm range). In this way, the formation of coarse grains and texture is improved, and hysteresis losses are reduced.
4) Prevention of surface oxidation and the resulting formation of fine grains at the surface. This is accomplished by optimizing the annealing atmosphere and eliminating

decarburizing annealing through the use of extremely low-carbon alloying techniques. Figure 30 shows how the magnetic loss $P_{1.5}$ depends on the sheet thickness for two surface conditions.

Commercially available NO electric-sheet qualities in thicknesses of 0.5 and 0.35 mm exhibit 50 Hz losses $P_{1.5}$ of between 2 and 4 W/kg (2 W/kg for the 0.5 mm sheet is the peak value).

5.1.1.2. Grain-Oriented Electric Sheet

Substantial improvements in the magnetic quality of electric sheet, in the sense of lower losses and higher permeabilities, are possible with the anisotropic GO materials. In principle, the requirements of high purity and low internal stress are the same as for the NO materials, but there is the constraint that lattice imperfections must be introduced in order to control texture formation and domain configuration.

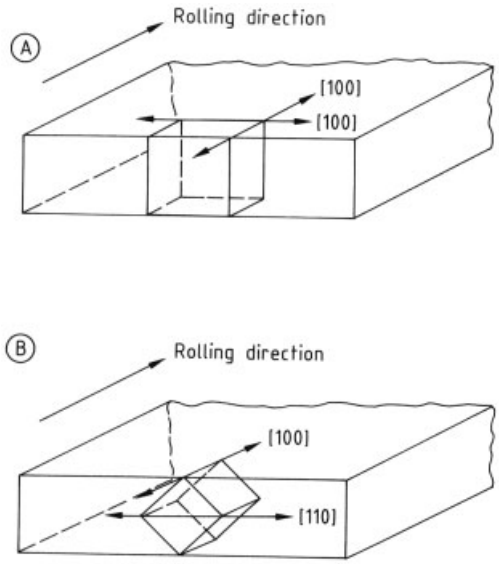

Figure 31. Textures of soft magnetic materials
A) Cube texture; B) Goss texture

Specially matched cold-rolling and annealing practices combined with the addition of "inhibitors" in melting produce a texture such that the grains have the [100] easy magnetization direction of the body- centered cubic mixed FeSi crystal preferentially oriented in the rolling direction. In the (100) [100] cube texture, the longitudinal and transverse directions of the strip are magnetically equivalent. A (110) [100] texture is called a " Goss texture" after N. P. Goss, who developed the process in the United States in 1934 [99]. All commercial GO electric sheet produced today has this orientation (see Fig. 31).

Figure 32 illustrates the history of high-quality GO electric sheet in a time plot of magnetic loss $P_{1.7}$ for RGO (regular grain-oriented) and HGO (high-permeability grain-oriented) qualities. The advantages of the HGO material [100], which appeared in the 1970s, result from improved sharpness of orientation and grain-size control with the simultaneous imposition of a mechanical tensile stress in the sheet surface by the insulation (stress coating). The magnetic loss can be further reduced by decreasing the strip thickness–provided that a good Goss texture is preserved (which becomes harder to do as the strip becomes thinner) [101–103] and the strip surface is kept smooth–in conjunction with domain size reduction by laser scribing (LS) or mechanical or chemical methods.

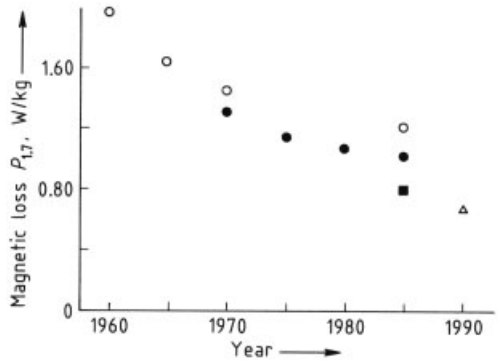

Figure 32. History of high-quality GO electric sheet in terms of the magnetic loss $P_{1.7}$ [98]
○ RGO, $d = 0.30$ mm; ● HGO, $d = 0.30$ mm; ■ HGO (LS), $d = 0.23$ mm; △ HGO (LS), $d = 0.20$ mm

Markedly lower losses $P_{1.7}$ of 0.21 W/kg at 50 Hz were measured on 0.031 mm thick laboratory fabricated sheets after tertiary recrystallization and domain size reduction by chemical etching [104]. Today, strip 0.23 mm thick is by far the most commonly used for commercial applications at line frequencies (50 Hz in Europe, 60 Hz in the United States) [105], where the lowest possible losses must be achieved along with safe mechanical handling.

Improvements in the process for making thinner strips (0.2 mm and less) are also being strongly advised in Europe [106]. By way of example, Figure 33 illustrates how magnetic loss depends on strip thickness in grain-oriented electric sheet. In contrast to the NO materials discussed in Section 5.1.1.1, as shown in Figure 32, the measured losses are distinctly lower despite the higher measuring flux density. Since 1970, as other sources [107] confirm, losses have been cut by an average of 50 % in GO electric sheet, and this development is clearly reflected in energy costs. According to [108], a 10 % reduction in magnetic losses yields an energy cost saving of € 36×10^6 in the Federal Republic of Germany alone.

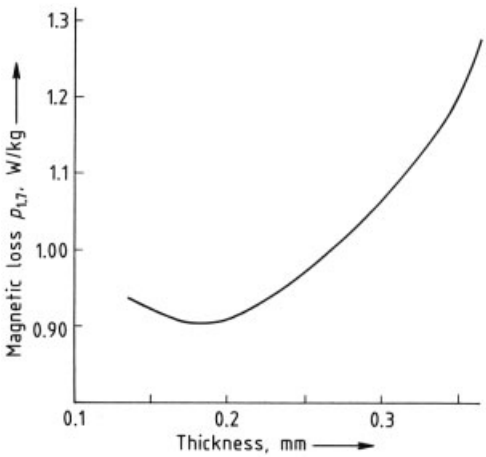

Figure 33. Magnetic loss as a function of sheet thickness in grain-oriented electric sheet [106]

The production of high-quality GO electric-sheet materials is complicated, and it is necessary to monitor a large number of process parameters that control the microstructure and bear crucially on the magnetic properties of the sheet in the final condition. A vitally important step is secondary recrystallization to generate the Goss texture, which should have the sharpest possible orientation. The mechanisms by which grain growth, texture, and control by particles of a second precipitated phase (such as manganese sulfide or copper sulfide, the "grain growth inhibitors") interact have been quantified only quite recently; according to [109], undisturbed secondary recrystallization with selective grain growth is possible only if certain preconditions–described in terms of an "effective grain radius" and an "inhibitor factor"–have been established in the primary structure. These quantities express the driving and inhibiting forces for optimal secondary recrystallization.

The strip surface should be as clean and smooth as possible; combined with the morphology of the electrical insulation on the strip, the surface finish represents an important factor determining the magnetic loss. The insulation layer is produced during the three-step final heat treatment of the cold-rolled strip. Table 17 lists the processes that take place in the several steps.

Domain size-reduction operations serve to cut losses by diminishing the anomalous eddy-current losses $P_A = P_W - P_E$ (excess losses), which result from the dynamics of the domain boundaries and cannot be derived directly from classical eddy- current theory. Excess losses result from the wide spacing of domain walls in very coarse structures (grain diameter 1 – 20 mm), which lead to movement of the domain walls over long paths upon remagnetization. Because no consistent theory of losses has yet been devised, the following simple treatment by Pry and Bean [110] is used to quantify this phenomenon.

If the spacing between domain walls is L and the sheet thickness is D, then for the case $2L > 0$ an anomaly factor

$$\eta = 1.63 \cdot 2L/D$$

is used to correct the classical eddy- current losses such that

$$P_{Fe} = P_H + \eta P_E$$

A domain structure with close wall spacing accordingly leads to lower total eddy-current losses, $P_W = \eta P_E$. If $L \ll D$, $\eta \to 1$, and the anomalous losses become negligible.

This analysis is being extended on the basis of statistical theories of loss [111–113]. Such theories also attempt to describe the situation in the NO materials, which is more difficult because the complicated microstructure implies more complex domain structures.

If a core (laminated or tape-wound toroidal) is made from strip already treated for domain size reduction, then for the treatment to remain effective in the final condition it must be stable against annealing, since mechanical working of electric sheet must generally be followed by a further stress-relief annealing step.

Table 17. Heat treatment in production of GO electric sheet

Heat treatment step	Conditions	Effect on	
		Fe – Si matrix	Insulation coating
1	fast continuous strip annealing at ca. 850 °C under moist H_2/N_2 atmosphere	decarburization	formation of fayalite (Fe_2SiO_4) and silicon dioxide (SiO_2) base layer
2	coarse-grain annealing in tight coil below 1200 °C for several days	recrystallization	reaction with MgO used as adhesive protection, yielding forsterite (Mg_2SiO_4) glass film
3	final annealing at 850 °C	stress relief	application and firing-in of a phosphate coating

Targets for development at present are a magnetic loss of $P_{1.7} = 0.4$ W/kg and economically produced strip 0.18 mm thick. Substantial increases in Si content (up to 6.5 %) through new strip-casting processes, in combination with smaller strip thicknesses, are promising trends in classical electric-sheet materials [106].

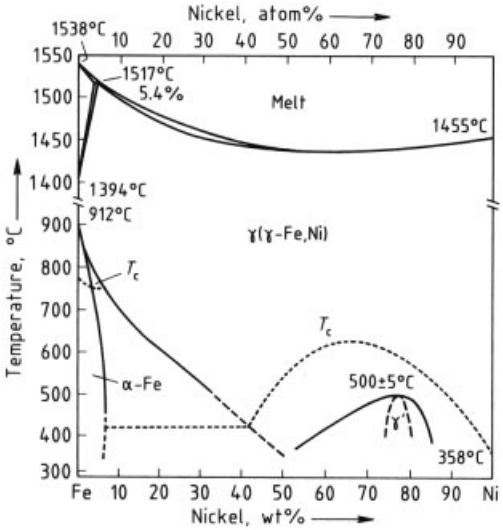

Figure 34. Iron – nickel phase diagram

5.1.2. Nickel – Iron Materials

The classical soft magnetic Ni – Fe materials cover a wide range of nickel contents, from 30 to 85 wt %; some contain small amounts of other substances. Figure 34 shows the phase diagram of the binary system Fe – Ni; Figure 35 presents curves of saturation magnetization and Curie temperature for this system. On the iron-rich side, the lattice is body-centered cubic below the Curie temperature; the industrially important nickel-rich alloys, on the other hand, crystallize in a face-centered cubic lattice. In the vicinity of Ni_3Fe, there is also an isotropic superlattice structure formed below 500 °C; this consists of short-range or long-range ordering and can be influenced by special cooling or tempering treatments. The superlattice structure is particularly important for the magnetic properties of alloys lying in this region.

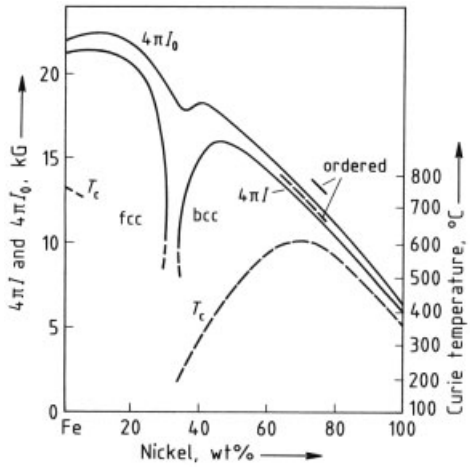

Figure 35. Absolute saturation $4\pi I_0$, saturation magnetization $4\pi I$ at 20 °C, and Curie temperature T_c in the Fe – Ni system [114]

The various crystallographic structures of the iron- and nickel-rich alloys result in a variety of easy directions for magnetization, because of the sign alternation of the crystal anisotropy constant K_1. Figure 36 shows how K_1 behaves in the Fe – Ni system.

When $K_1 \gg 0$, that is, on the iron-rich end, <100> is the easy direction. Above 36 % Ni, K_1 is small and positive; at ca. 75 % Ni, after rapid cooling from temperatures above 500 °C, it passes through zero; thereafter it falls to more negative values. In this region, the cube diag-

onals <111> of the face-centered cubic lattice are magnetically easy.

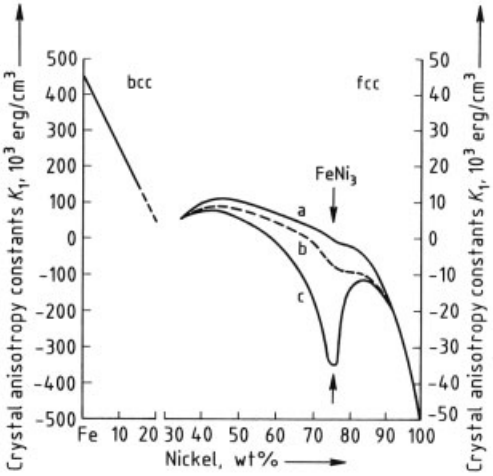

Figure 36. The crystal anisotropy constant K_1 in the Fe – Ni system [117]
a) Quenched; b) Oven cooled (ca. 50 K/h); c) Slowly cooled (ca. 2.5 K/h)

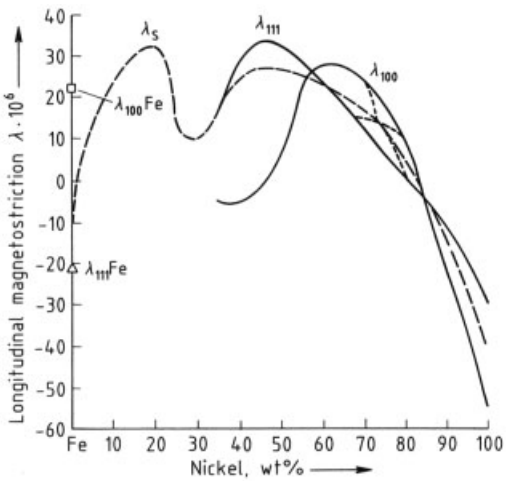

Figure 37. Magnetostriction constants λ_{100}, λ_{111}, and λ_s for the Fe – Ni system [118]

By altering the cooling conditions and the associated ordering, it is possible to shift the zero of K_1 toward lower Ni contents. In conjunction with further additions of Mo, Cu and Cr, this effect is particularly important for a markedly high level of the soft magnetic values in high-nickel materials (see Section 5.1.2.1).

The magnetostriction constants λ_{100}, λ_{111} and λ_s in the Ni – Fe system start out positive and pass through zero above 80 % Ni. Ordering processes do not play an important role here, however, as can be seen in the slight differences between the values after fast and slow cooling. Furthermore, λ_{100} crosses through zero at about 45 % Ni, while λ_s, the saturation magnetostriction in the field direction for polycrystalline structures, has a zero corresponding nearly to pure iron (see Fig. 37).

These points show which alloy compositions in the Ni – Fe system promise good soft magnetic values and how it may be possible to optimize the values by the addition of other elements. Because the smallest possible (or even vanishing) magnetostriction and/or crystal anisotropy are requisites for high permeabilities, while high saturation flux densities and high resistivities are desirable, essentially three alloy regions are of interest:

1) High-nickel alloys (> 70 % Ni)
2) Alloys with moderate nickel contents (45 – 65 %)
3) Low-nickel alloys (ca. 36 % Ni)

These alloys are prepared by melting at atmospheric pressure or under vacuum. Vacuum degassing is indispensable if maximum purity is to be achieved; small amounts of Al, Mg, or Si are added for deoxidation. The addition of Mn fixes the sulfur and thus improves the workability. Another possibility is to use the powder-metallurgical techniques of pressing and sintering, which can produce high-purity, highly homogeneous materials.

Whether produced by melting or sintering, the blocks are next thinned to a few millimeters by hot rolling, then cold-rolled, with well-defined intermediate annealing in the range 800 – 1000 °C, to thicknesses of 0.003 – 1 mm, depending on the desired final thickness. Surface treatments such as planing, pickling, and milling are needed especially after the hot-working operations.

Sheet rolled to the final thickness is made into tape-wound cores with insulation between layers, laminated cores, or special forms. Special heat treatment is required to produce the optimal magnetic properties for the particular composition. Initially there is a high-temperature final anneal above 1000 °C under a reducing atmosphere. In the course of this operation, the structure is purified by the removal of C, N,

S, and O, and grain growth takes place by secondary recrystallization. Thus an essential condition for good soft magnetic behavior is satisfied: the elimination of internal imperfections that hinder domain-boundary movement.

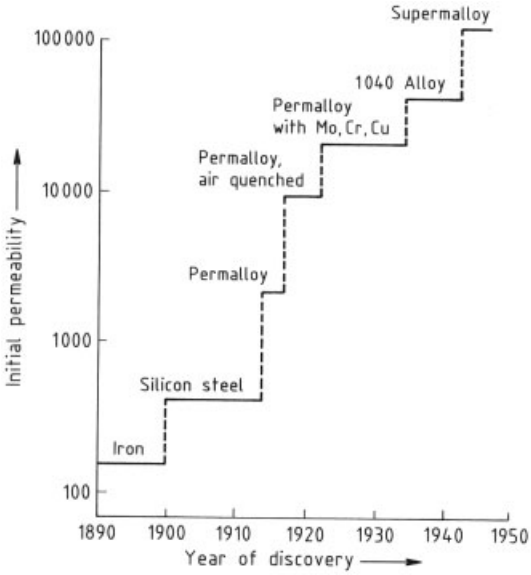

Figure 38. Initial permeability of important soft magnetic alloys as a function of the year of discovery [122]

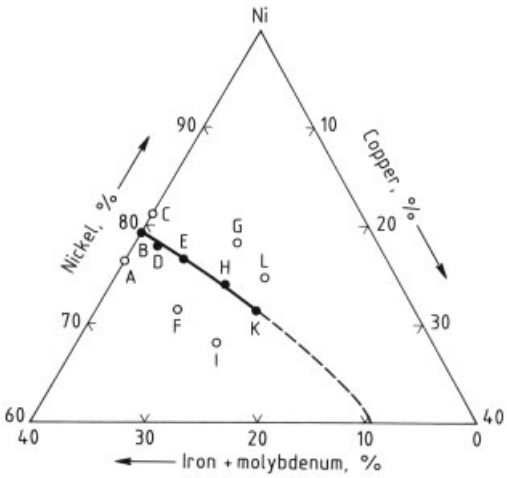

Figure 39. Favorable alloy compositions with Supermalloy behavior (B, D, E, H, K) in the Ni – Cu – (Fe + Mo) system [121]

This anneal is followed by tempering under conditions that depend on the alloy composition. This operation is carried out in the ordering region or in a magnetic field below the Curie temperature. Comprehensive accounts of soft magnetic iron – nickel alloys appear in [114–116].

5.1.2.1. High-Nickel Alloys

The development of high-nickel Fe – Ni alloys, also called Permalloys, began in the 1920s when the binary Permalloy (75 wt % Ni) was discovered by Elmen [119]. Later advances were mu-metal (76 % Ni, 5 % Cu, 2 % Cr), M 1040 (72 % Ni, 14 % Cu, 3 % Mo), and Supermalloy (79 % Ni, 5 % Mo, and others) [120]. Figure 38 shows the progressive improvement of the initial permeabilities in these alloys versus the year of discovery. For the sake of comparison, the earlier state of the art–soft magnetic iron and iron – silicon materials–is also shown; Permalloy represents a substantial gain over these. In Supermalloy, the composition for the maximum magnetic values is still restricted to a very narrow range. By varying the nickel content and adding copper and molybdenum, Assmus and Pfeiffer [121] showed that peak magnetic values are not limited to a certain composition but occur over a broader series of alloys. The copper-free Supermalloy thus stands at the beginning of an alloy series that extends to the ternary system Ni – Cu – (Fe + Mo) (see Fig. 39).

Replacing part of the Fe with Mo moves the zero point of K_1 from 75 % Ni in binary Permalloy (Fig. 36) toward higher nickel contents, while adding copper shifts the zero of λ_s (about 80 – 81 % Ni in Figure 37) toward lower nickel contents without affecting K_1. In comparison with current Permalloy materials, however, the maximum initial permeability of ca. 100 000 for the original Supermalloy appears relatively low.

The clear improvement now attained in the magnetic values of high-nickel alloys (μ_4 values $> 250\,000$; see Table 15) is due to optimized chemical compositions, based on the findings cited above, and the development of heat-treatment programmes specially tailored to them. In this way, the detrimental effects of crystal anisotropy and magnetostriction are further reduced. The alloy containing 77 % Ni, 5 % Cu, 4 % Mo, and 14 % Fe is an example of the high-nickel materials common today (77 % Ni Permalloy).

Furthermore, advances in technique and improvements in melting, processing, and annealing have led to continuing and significant reductions of internal perturbations in industrial-scale melt metallurgy. Impurities such as oxygen, sulfur, and carbon have a very harmful effect on soft magnetic behavior and must be eliminated as far as possible in the final high-temperature reducing anneal.

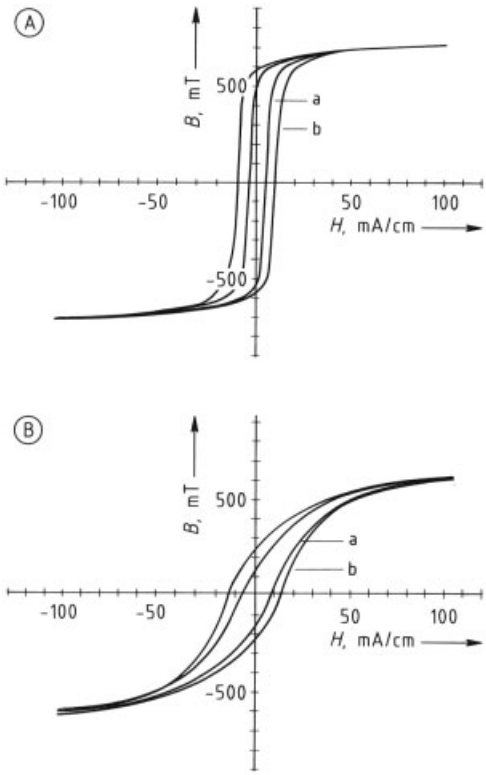

Figure 40. Static and dynamic hysteresis curves of a rich nickel alloy
A) Optimally tempered: H_c (static) = 0.4 A/m, B_r (static) = 0.52 T, μ_4 (dynamic) = 342 000, μ_{max} (dynamic) = 455 000; B) Magnetically tempered: H_c (static) = 0.7 A/m, B_r (static) = 0.20 T, μ_4 (dynamic) = 78 000, μ_{max} (dynamic) = 110 000
a) Static; b) Dynamic (50 Hz)

Heat treatments in very narrow temperature ranges below 550 °C, or under well-defined cooling conditions, are essential for establishing the charge-specific magnetically optimal tempering condition in order to realize the necessary critical degree of close-range order. In this way, departures from the optimal alloy composition can be subsequently corrected. Figure 40 shows the static and dynamic (50 Hz, sinusoidal flux density) hysteresis curves for an optimally treated alloy from this group, measured with a tape-wound toroidal core having a strip thickness of 0.1 mm.

In addition, the selection of tempering temperature makes it possible to control the temperature dependence of the permeability at typical service temperatures. In this way, either maximum permeability at room temperature or nearly constant permeability between − 25 and + 80 °C (temperature compensation) can be achieved. Figure 41 gives a plot of the permeability μ_7 versus temperature for the tape-wound toroidal core of Fig. 40 A and also for a temperature-compensated material.

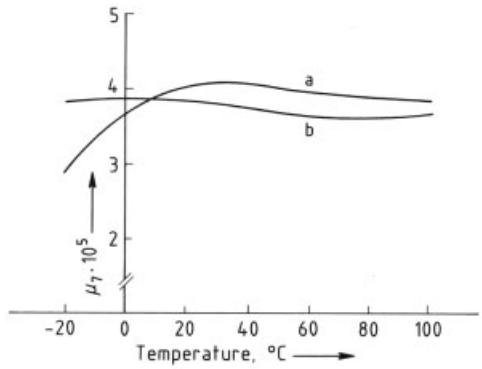

Figure 41. Permeability μ_7 of an optimally tempered (a) and temperature compensated (b) toroidal tape-wound core (made from a high-nickel alloy) as a function of temperature

Another possibility for controlling the shape of the hysteresis loop is tempering in a magnetic field below the Curie temperature, which leads to magnetically-induced anisotropic close-range ordering (directional ordering of Fe – Ni pairs). Figure 40 B shows a flat hysteresis loop realized in this way.

The microscopic causes of these effects lie in energetic stabilization of certain domain structures, which lead to round or rectangular hysteresis loops depending on the direction of the driving field. Practical application of these processes, however, presupposes certain core geometries such as the tape-wound toroid.

The very high magnetic values in industrial Permalloys, coupled with low coercive force, low magnetic loss, and high permeabilities at low driving field strengths, are important chiefly in inductive components subject to the most

stringent requirements. High-quality soft magnetic materials permit small core dimensions, so that the demand for miniaturized components can be satisfied.

5.1.2.2. Medium-Nickel Alloys

The group of nickel – iron alloys with moderate nickel contents (45 – 65 % Ni) exhibits the highest saturation flux densities of all the industrially important Fe – Ni alloys. The magnetic values are not as high as those of the modern Permalloy materials, but peak values are up to the lower limit of the high-nickel materials. Because the saturation flux densities are nearly twice as great, the hysteresis curve displays high $\hat{B}$ values or wide induction changes ΔB. This material group is thus also attractive for special applications in power transmission, primarily when very low losses are required.

Alloys with over 53 % nickel have only a weakly positive crystal anisotropy, so that they attain markedly higher magnetic values than do the lower-nickel materials. At high nickel contents (65 %) with optimized tempering below 550 °C, K_1 can even vanish. Furthermore, it is possible to carry out tempering in a magnetic field so as to bring about a large K_u value, since the high Curie temperature and the high saturation flux density create favorable conditions for magnetic-field-induced diffusion processes. In this way, as in the high-nickel alloys, the hysteresis loop can be varied from flat to rectangular (magnetic-field-induced rectangular loop). The addition of about 2 % Mo increases the resistivity and thus improves the dynamic behavior. In alloys with lower nickel contents (below 60 %), the effect of crystal anisotropy can be greatly reduced through mutual compensation of K_1 and K_u; this gain is also reflected in high magnetic values.

At nickel contents below 53 %, the stronger influence of crystal anisotropy ($K_1 > 0$, easy direction <100>) clearly lowers the level of magnetic values. In these alloys, however, it is possible to generate a cube texture in the strip plane through cold working (> 95 %) and coarsening of the primary recrystallized structure by annealing between 1000 and 1100 °C. As a result, with magnetization in the rolling direction and perpendicular to it, an extremely rectangular hysteresis loop is obtained; this behavior is especially favorable for the magnetization conditions in core laminations.

Annealing above 1100 °C leads to a large increase in grain size through secondary recrystallization and disrupts the cube texture if any secondary texture was still present. The initial permeability does increase, but the hysteresis loop loses its rectangular shape.

Typical magnetic values are listed in Table 15.

5.1.2.3. Low-Nickel Alloys

At the iron-rich end of the binary Fe – Ni system the group of alloys with about 36 % nickel forms the end of the series of industrially important high-permeability soft magnetic alloys. Alloys containing 30 % Ni are assigned to a separate group because of their strongly temperature-dependent magnetic properties, among which the saturation flux density is particularly important (see Section 5.1.5).

In comparison with the 45 – 49 % nickel materials, this group has lower permeabilities and higher coercive forces. The values are similar to those of siliconized electric sheet, but here the resistivities are far higher and the cold-working properties are significantly better. By virtue of the last point, thin strip can be produced and then worked further by deep drawing. The 36 % NiFe materials thus form, so to speak, a bridge between cheap electric sheet, with the mechanical constraints on its use, and more costly high-quality Ni – Fe alloys with more than 45 % nickel.

The saturation flux density of 1.3 T for the 36 % NiFe alloys is still relatively high. In combination with the high resistivity, this feature makes the material suitable for power applications in the medium-frequency range. If the nickel content is further reduced, the saturation drops sharply. Like the medium-nickel alloys, the 36 % Ni alloys are manufactured with the smallest possible amount of additives so as not to lower the saturation value further.

Special manufacturing processes make it possible to control the relationship between permeability and field strength, making the curve flatter or steeper (see Fig. 42).

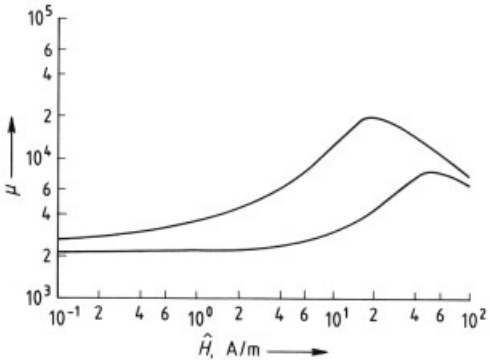

Figure 42. Steep and flat permeability curves of a 36 NiFe alloy

5.1.3. Iron – Cobalt Alloys

A ternary Fe – Co – V alloy (Permendur) represents the only material from this alloy group that is of any industrial value today. Binary Fe – Co alloys containing 27 and 35 % cobalt are of only marginal importance (special high-temperature applications). Figure 43 shows the phase diagram of the Fe – Co system, which can also be used in characterizing phase transitions in the Fe – Co – V system.

The alloy Fe49Co2V is characterized by a high saturation flux density of 2.35 T, which represents the largest value of all the soft magnetic alloys. This value is attained, however, only at relatively high driving field strengths, roughly 10 kA/m. Furthermore, the Curie temperature is high (T_C= 950 °C), and large maximum permeabilities can be found at high temperatures. The room-temperature magnetic values are only moderate. The iron – cobalt alloys are expensive because of their high cobalt content and are used chiefly for inductive elements in aerospace applications, where size and weight savings are important.

The mechanical working of iron – cobalt alloys causes some problems. In a purely binary 50 FeCo system, cold working would be impossible because of the very rapid superlattice structure formation in the α-phase, the resulting embrittlement of the material, and the tendency for intergranular cracking.

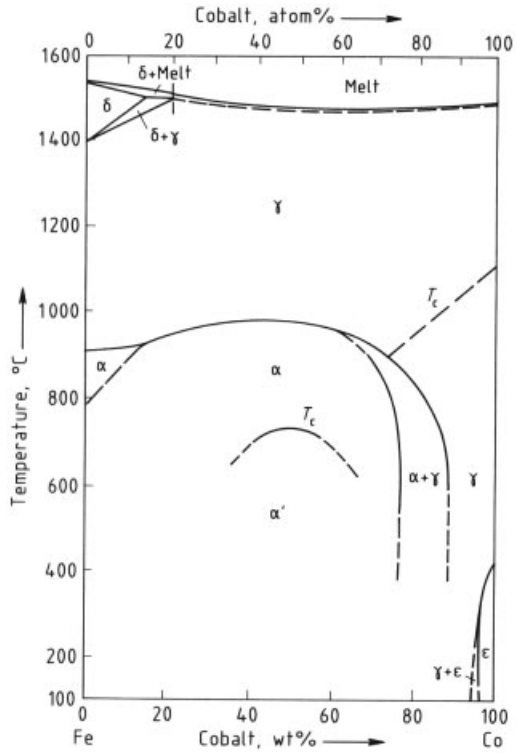

Figure 43. Iron – cobalt phase diagram [123]

The addition of vanadium slows down the ordering processes commencing below 750 °C, so that a disordered α state, generated by annealing just below the $\alpha - \gamma$ phase-transition boundary, can be quenched at room temperature. This creates better conditions for cold working by setting up a ductile mixed structure consisting of the α-phase and martensite and containing no superlattice structure.

For magnetic optimization, it is useful to reintroduce some degree of ordering in the final heat treatment (after the conclusion of mechanical working) in order to control the zero crossing point of K_1.

Experimental results [124] indicate that the addition of 0.2 % niobium or tantalum in place of the 2 % vanadium gives the same gain in ductility and further improves the magnetic values (B_s= 2.45 T).

5.1.4. Soft Ferrites

Ferritic materials have been known for over 50 years and have been in large-scale industrial use

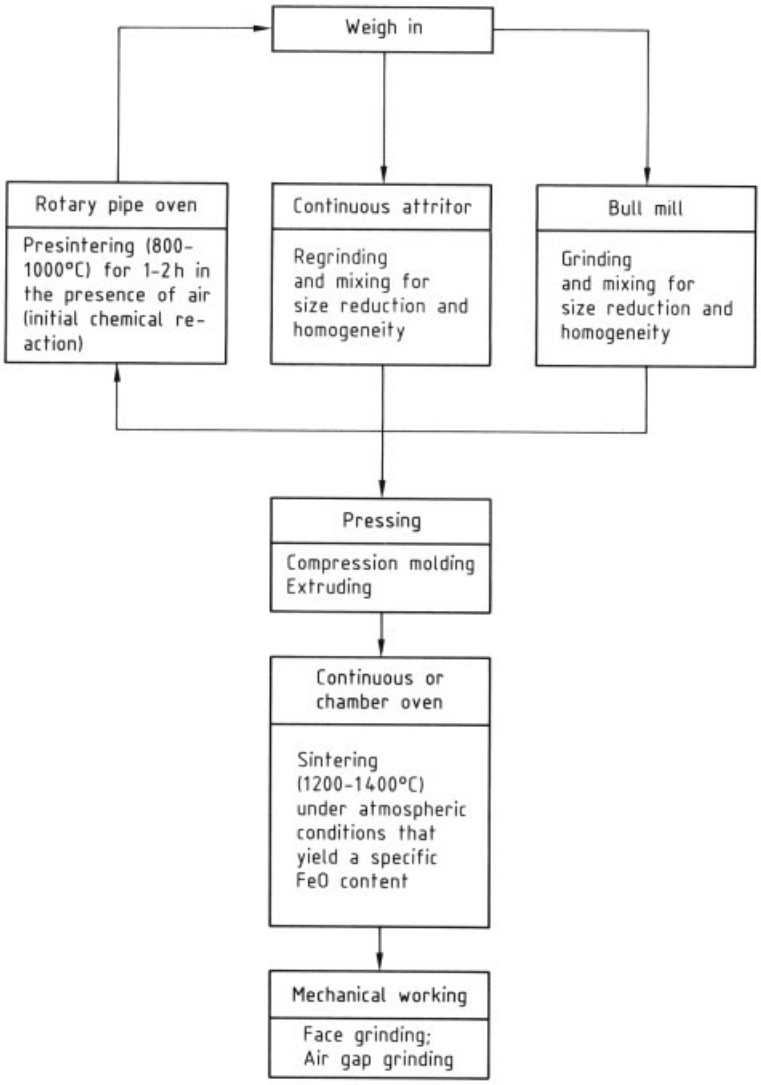

Figure 44. Flow-sheet for the production of soft magnetic ferrites

since the end of the 1940s. Among soft magnetic ferrites, the following classes are distinguished on the basis of the underlying crystal structure:

1) Cubic ferrites with spinel structure
2) Hexagonal ferrites with magnetoplumbite structure
3) Hexagonal ferrites of higher order
4) Ferrites with garnet structure
5) Orthoferrites

All these groups have found widespread application in industry [125].

Cubic ferrites have the chemical formula $MO \cdot Fe_2O_3$, where M is a divalent metal ion such as Fe, Ni, Co, Mn, Mg, Cu, Ti, Cd, or Zn, or a combination of such ions. They crystallize in the spinel structure, in which the oxygen ions form a face-centered cubic (fcc) lattice. The metal ions occupy the remaining tetrahedral (8 per unit cell) and octahedral (16 per unit cell) sites.

Other cubic soft ferrites of industrial importance, in addition to the nickel type, are the mixed ferrites, in which zinc figures as a sec-

ond metal ion. The MnZn and NiZn ferrites are the most common materials in this category:

Ni ferrites: $(Ni, Fe)O \cdot Fe_2O_3$
Ni – Zn ferrites: $(Ni, Zn, Fe)O \cdot Fe_2O_3$
Mn – Zn ferrites: $(Mn, Zn, Fe)O \cdot Fe_2O_3$

Hexagonal soft ferrites represent a mixture of cubic spinel ferrites and hexagonal barium ferrites with the magnetoplumbite structure (see Section 4.1.4).

Because of their ceramic nature, ferrites exhibit much higher room-temperature resistivities than metallic materials (see Table 14). They are thus especially suitable as cores in inductive components for super-high-frequency dynamic service. Cubic ferrites can be operated into the MHz range, and hexagonal ones into the GHz range, without any drastic loss of magnetic properties due to eddy- current losses. In comparison with metallic soft magnetic materials, however, the much lower saturation flux density (a result of the structure) must be considered a drawback. The brittleness of ferritic materials also restricts the possibilities of machining such magnets.

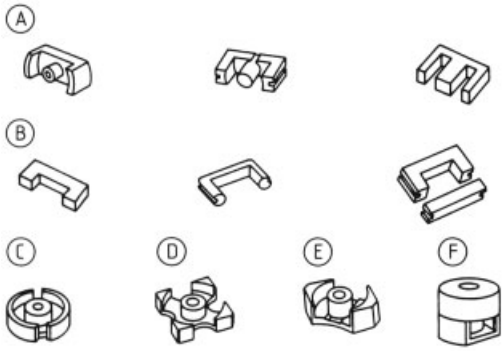

Figure 45. Shapes of soft ferrite cores [150]
A) E cores; B) U cores; C) Pot cores; D) Cross cores; E) RM cores; F) TT cores

Soft magnetic ferrites are produced by molding and sintering. As is common with soft magnetic materials, the starting materials must be as pure as possible. The magnetic properties are governed equally by the chemical reaction yielding the oxides and by the kinetics of sintering. The degree of oxidation, oxygen content, and grain size are among the important parameters that must be precisely controlled but which also offer possibilities for variation. In order to avoid lowering the saturation flux density, the residual porosity after sintering should be made as small as possible. Figure 44 shows a flowsheet for the production of soft magnetic ferrites [126]. A number of largely standard core shapes (e.g., U, E, EC, ETD, and RM cores; see Fig. 45) and matching accessories such as bobbins are available for applications in electronics.

The coercive forces of soft magnetic ferrites range from 10 to 1000 A/m; the maximum saturation flux density is ca. 0.5 T. Initial permeabilities of up to 40 000 can be realized with Mn – Zn ferrites, but only at operating frequencies of a few kHz and in extremely narrow ranges of working temperature (see Fig. 46). Nickel and Ni – Zn ferrites offer initial permeabilities of about 1000, but these figures hold well into the MHz range.

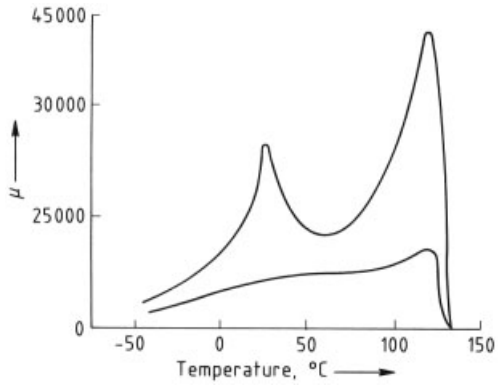

Figure 46. Permeability as a function of temperature for high permeable ferrites [127]

It must be acknowledged that only a portion of the magnetic properties can be optimized in any one ferritic material. Maximum initial permeabilities and low losses cannot be combined in one material, so that there is always an application-specific compromise. Typical compositions for high initial permeabilities (zero crossing of the crystal anisotropy) are close to 26 mol% MnO, 52.5 mol% Fe_2O_3, balance Zn; while high saturations and Curie temperatures for power applications can be attained at 35 mol% MnO, 54 mol% Fe_2O_3, balance Zn.

While ferrites were formerly used chiefly as RFI suppression choke core materials, more recently they have attracted interest in wide-band repeating coils and power applications (switched mode power supplies). Table 18 shows the relevant magnetic properties for various applications. The increasing demand for wide-band

Table 18. Relevant magnetic properties of soft magnetic ferrites for various applications

Application	Magnetic property*										
	a/μ_i	D/μ_i	$\tan/\mu_i$	μ_i	R_P	P_V	μ_a	T_c	$\Delta L/L_o$	η_B	A_L
Filters	x	x	x								
Wide-band repeater coils				x	x						
Repeater coils, chokes (Power applications)						x	x	x	x		
Inductance coils									x	x	x

* a/μ_i: temperature coefficient; D/μ_i: disacomodation factor; $\tan/\mu_i$: loss factor; μ_i: initial permeability; R_P: parallel loss resistance; P_V: power loss; μ_a: amplitude permeability; T_c: Curie temperature; $\Delta L/L_0$: inductivity rise; η_B: hysteresis correction value; A_L: inductivity factor

repeating coils results from new digital transmission techniques in communications (e.g., ISDN), where ferrites compete with metallic wound strip cores, often amorphous in nature, which offer higher permeabilities.

5.1.5. Special Soft Magnetic Materials

Materials whose magnetic polarization is strongly temperature-dependent under service conditions of −30 to 120 °C are used chiefly to compensate the temperature dependences of permanent-magnet systems. Such temperature-compensating alloys must satisfy the following requirements:

1) well-defined slope of polarization curve in service temperature range
2) exact fixing of Curie temperature above room temperature or of its apparent equivalent at a certain working field strength below saturation
3) reversible variation of magnetic properties on heating and cooling

Industrially important for this application area are Fe–Ni–C alloys with closely controlled nickel contents between 29 and 31 wt %. Variations of as little as 0.1 % in the nickel content or 0.01 % in the carbon content lead to changes of 10–20 mT in polarization at 20 °C. The addition of carbon up to about 0.3 wt % affords thermal stabilization of the ferromagnetic γ-phase. Other materials with similar thermomagnetic properties, but having lower slopes of the polarization versus temperature curve, are 30NiFe alloys with added chromium, and Ni40Cu.

The material is delivered as cold-rolled strip or stampings made from such strip, already heat-treated, and with a definite temperature dependence of polarization. Figure 47 shows three typical curves that can be produced by varying the nickel content slightly.

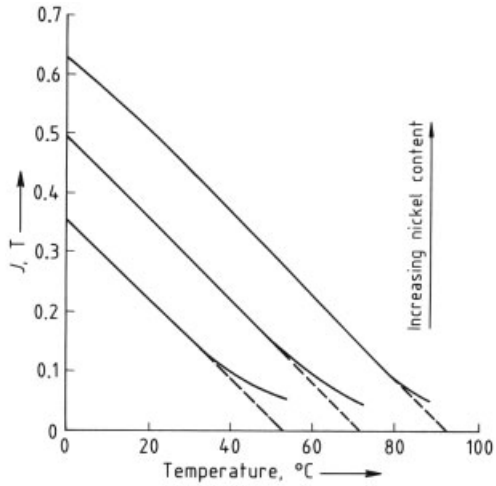

Figure 47. Polarization as a function of temperature for three temperature-compensation Fe–Ni–C alloys

Important applications are in magnetic shunts for the permanent-magnet flux in magnetic circuits of tachometers and measuring instruments, in thermomagnetic switches, in control components, and as temperature-dependent inductances.

If stringent requirements are imposed on the thermal or corrosion resistance of soft magnetic materials, soft magnetic chrome steels are candidates for use. These are ferritic stainless steels containing 14–24 wt % chromium and less than 0.08 wt % carbon. In comparison with the non-oriented electric sheet used under nor-

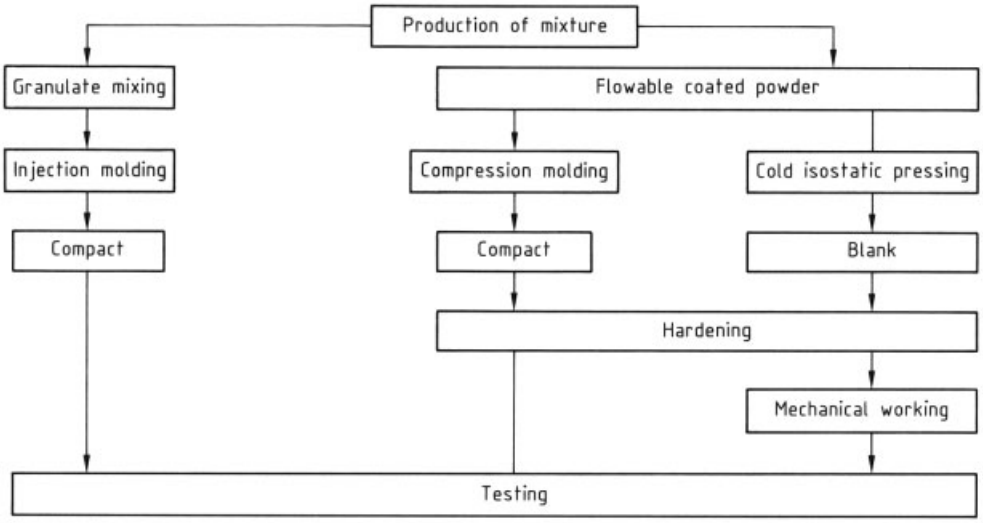

Figure 48. Production process for powder- composite magnetic materials

mal conditions, here the level of magnetic values is markedly lower while losses are higher and saturation flux densities and permeabilities are lower.

5.2. Modern Soft Magnetic Materials

5.2.1. Powder Composite Materials

Modern soft magnetic powder composites are a direct development of the classical powder core. The idea of a composite made up of particles electrically isolated from one another goes back to HEAVISIDE [128]. Iron powder cores were developed as early as 1932 by VOGT [129].

The production of soft magnetic sintered components starts with metal powders, which may consist of the elemental metal (Fe), a mixture of several elemental powders (Fe, Si, Ni, Co), or alloy powders (Fe – P, Fe – Si, Ni – Fe, Co – Fe, Ni – Fe – Mo). These powders are molded, then subjected to thermal compaction. Such powder composites are multicomponent structures. In addition to the basic magnetic powder, electrically insulating additives and organic or inorganic binders are included. In a preparatory step, the metal powder grains are surrounded as completely as possible by electrically insulating skins. The mixture is then converted to a compact under high pressure. Parts can also be fabricated by injection molding (see Fig. 48). Since these parts cannot be subsequently compacted by sintering, a binder must provide the mechanical stability of the formed part. Depending on the nonmetallic components used and their thermal stabilities, heat treatment serves to relieve mechanical stresses and harden the composite. The final density must accordingly be attained directly through the compaction process.

Modern powder composites are made with organic insulating and binding agents. Such resin composites can then be shaped by mechanical methods such as turning, milling, drilling and boring, so that large sizes and complicated geometries not directly obtainable by molding can still be realized. In contrast to laminated core shapes, a powder composite is isotropic in its magnetic properties.

From the magnetic standpoint, a powder composite represents a circuit strongly sheared by the air gaps inherent in the material. The permeability level in a static measurement is therefore very low. Iron-based powder composites offer a maximum initial permeability of 100. Because the magnetic particles are electrically isolated from one another, the electrical volume resistance is high; at the same time, however, eddy-current losses are drastically reduced, so that the permeability remains approximately

constant under dynamic driving up to the MHz range.

The strong shearing also leads to very good d.c. preloadability. Because the metallic base materials have higher saturation flux densities, this represents a clear advantage over ferrites. Given an appropriate choice of mean grain size and grain-size distribution of the magnetic particles in the range between 10 and 100 µm, hysteresis losses–which are the dominant loss component over a wide range of frequencies– can be minimized. If the grains are too large, the eddy-current losses increase to an undesirable degree; if they are too small, the background hysteresis loss becomes too great. The shape of the magnetization curve and the maximum saturation flux density can be adjusted through the metallic filling factor of the composite and the type of powder used (Fe, Fe – Si, Fe – Ni).

Figure 49 shows the static magnetization curves for three iron-based powder components having different filling factors, with the curve for a soft magnetic sintered iron for comparison.

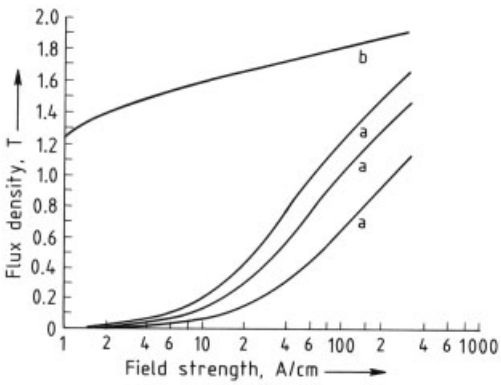

Figure 49. *B* (*H*) curves for three iron-based powder components of varying density (a) and a sintered iron material (b)

Given optimal isolation of the particles, the magnetic properties of soft magnetic powder composites can be improved only by reducing the hysteresis losses. Since the required annealing temperature is high, mechanical stress relief in the molded part is possible only with inorganic insulation. It is certainly advantageous to use amorphous metal powders with vanishingly low magnetostriction. At present, however, there is no low-cost compaction process for such future products.

5.2.2. Rapidly Quenched Materials

The magnetic materials dealt with in the previous chapters all exhibit a crystalline lattice structure. Associated with this is an inherent magnetocrystalline anisotropy, which must be minimized if optimal soft magnetic properties are to be obtained. In addition, all industrially used alloys have a polycrystalline grain structure, which must either be made as coarse grained as possible by secondary crystallization, or whose texture must be adapted to the shape of the intended magnetic core.

By means of rapid quenching with cooling rates of ca. 10^6 K/s, a technique developed in the 1960s [130], it is possible to suppress crystallization during solidification and to retain the amorphous structure of the melt (see also → Metallic Glasses). In this way, wires or strips of amorphous soft magnetic materials can be prepared directly from the melt [131–133]. The alloys contain iron, nickel, and cobalt as ferromagnetic elements (T), or mixtures of these, and crystallization-suppressing metalloids (X) such as silicon, boron, carbon, and phosphorus. Typical compositions lie in the range $T_{70-85}X_{15-30}$ [134].

The most conspicuous property of these amorphous alloys is that they are the first soft magnetic materials to exhibit high mechanical hardness ($HV_{0.2} = 700 - 1100$).

5.2.2.1. Amorphous Magnetic Materials

The amorphous soft magnetic materials, without crystal anisotropy and structural lattice defects, and the possibility to produce alloys free of magnetostriction were milestones in the development of soft magnetic materials. The magnetic values attained by these materials, in particular for dynamic applications, are excellent (see Table 15). Particularly favorable is the electrical resistance, which is about three times greater than that of crystalline materials, and the thinness of the ribbons resulting from the production process (15 – 50 µm). The high yield strength of metallic glasses, about 5 – 10 times greater than those of comparable crystalline materials, means that they are insensitive to plastic deformation,

which has advantages for use in magnetic heads or as magnetoelastic sensors [135].

Toroidal cores can be very easily wound from metallic glass ribbons after insulation of the surfaces, without the numerous forming and annealing steps that are necessary for crystalline materials. However, in the production of stampings the greater hardness of the amorphous alloys leads to 1000-fold faster wear of tools.

The magnetization process in amorphous soft magnetic alloys is determined by the magnetoelastic energy and the induced anisotropy. Internal mechanical stresses, resulting from inhomogeneous solidification [136], and surface roughness also play a role. The stresses can be relieved by thermal annealing below the crystallization temperature, whereas the state of the surface and the uniformity of the ribbon thickness depend mainly on the production process.

Magnetic tempering below the crystallization and Curie temperatures ($< 400-550\,°C$) allows the establishment of any desired hysteresis loop [137], as with crystalline Ni – Fe alloys. Noteworthy is that extremely rectangular or flat hysteresis loops are possible, with retention of the low coercive field strength.

Compared to soft ferrites, the classical materials for cores in high-frequency dynamic applications, the metallic glasses exhibit a better temperature dependence of the magnetic values and considerably higher saturation flux densities, which is important for the miniaturization of inductive components.

Amorphous Iron-Based Alloys. The amorphous iron-based alloys are composed of iron, silicon, boron, and/or carbon; examples are $Fe_{78}B_{19}Si_3$ and $Fe_{81}B_{13}Si_4C_2$. Since they are produced from cheap raw materials, they could be used for power applications in electrical technology.

In low-frequency applications, the amorphous iron-based alloys compete with grain-oriented electric sheet. Their losses are 60 – 70 % lower (0.15 – 0.2 W/kg at 1.4 T and 50 Hz) than those of conventional materials and their permeabilities are higher (see Fig. 50). However, the saturation flux density of 1.6 T is lower than that of crystalline Fe – 3 % Si, meaning that transformers can only be driven at up to ca. 1.4 T (see Fig. 51).

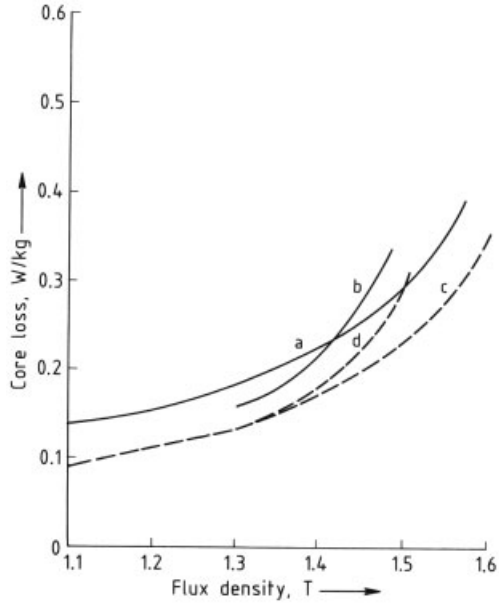

Figure 50. Core loss of two amorphous iron-based alloys at varying temperature
a) $Fe_{81}B_{13}Si_4C_2$ at 25 °C; b) $Fe_{81}B_{13}Si_4C_2$ at 100 °C; c) $Fe_{78}B_{19}Si_3$ at 50 °C; d) $Fe_{78}B_{19}Si_3$ at 100 °C

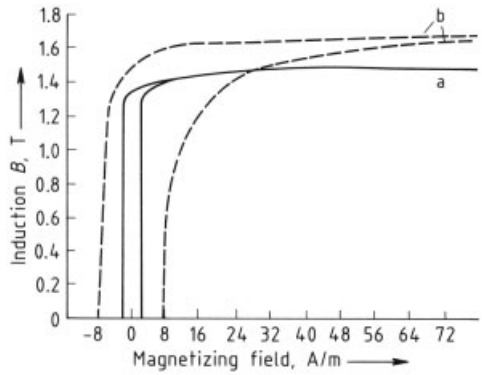

Figure 51. Comparison of the static $B(H)$ curves of an amorphous $Fe_{78}B_{13}Si_9$ alloy with that of grain-oriented electric sheet [138]
a) Amorphous $Fe_{78}B_{13}Si_9$, 30 µm thick; b) Grain-oriented silicon steel, 0.28 mm thick

An interesting potential use is in distribution transformers in the United States and Japan [139], [140]. Use of amorphous materials in all 35×10^6 distribution transformers in the United States could save 60×10^9 kW of energy per year [141]. However, the following properties make mass production of amorphous transformer cores unattractive: the stress sensitivity

of magnetostrictive iron-based alloys; their brittleness after annealing; the difficulty of punching the hard thin ribbons; and the resulting low filling factors of the cores. Nevertheless, amorphous distribution transformers with cross-shaped wound cores, covering the range up to 500 kVA, are now commercially available in the United States. The commercial viability has been demonstrated in product development and pilot series [140]. This application is, however, not of interest for Europe.

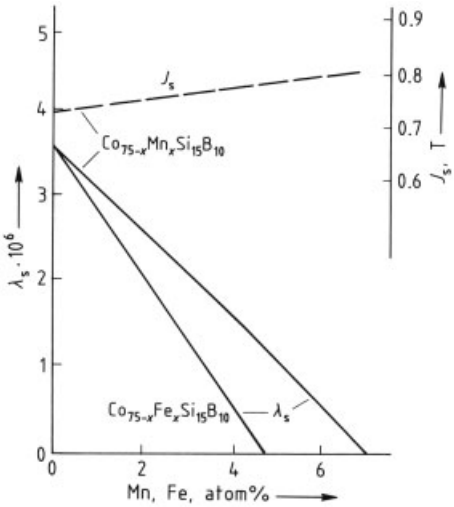

Figure 52. Variation of the magnetostriction constant λ_s and the saturation polarization J_s with the manganese and iron content of two amorphous cobalt-based alloys

Amorphous Cobalt-Based Alloys. The amorphous cobalt-based alloys have lower saturation flux densities than the iron-based alloys. Depending on the alloy composition, they are characterized by a small or almost vanishingly small magnetostriction constant λ_s ($|\lambda_s| < 10^{-7}$ possible) (Fig. 52). This leads to maximum values for the magnetic properties H_c, μ, and P_v and low stress sensitivity. At low working frequencies, the magnetic values are comparable to or better than those of Permalloys. At high driving frequencies (kHz range), the amorphous cobalt-based alloys are also superior to ferrites. These materials are therefore highly suited for highfrequency, high-tech applications [142]. The main disadvantage is the relatively high cost of these alloys resulting from their high cobalt contents. However, the ability to manufacture small, light cores, which is possible due to the high magnetic values, could contribute positively to the total cost of a complete inductive element.

Amorphous Nickel – Iron Based Alloys. These alloys contain approximately equal amounts of iron and nickel (ca. 40 at %), the rest being silicon and boron. Addition of molybdenum improves the thermal stability. The magnetic values lie between those of the amorphous iron- and cobalt-based alloys. The magnetostriction constant is ca. 10×10^{-6}. Figure 53 shows typical hysteresis curves, measured on toroidal cores in the tempered, as-produced condition and after two different magnetic temperings, which demonstrates the possibility of influencing the form of the hysteresis curve.

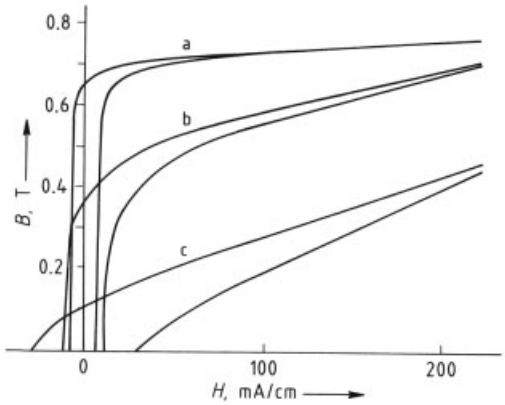

Figure 53. Static hysteresis curves of an amorphous $Fe_{39}Ni_{39}Mo_4Si_6B_{12}$ alloy subjected to various heat treatments [143]
a) Magnetic field annealing (longitudinal); b) Annealed (rapid cooling); c) Magnetic field annealing (transverse)

5.2.2.2. Nanocrystalline Magnetic Materials

In the 1980s a new group of soft magnetic materials with extremely fine grained structures was developed in Japan [144], [145]. They are produced by rapidly quenching an amorphous iron-based alloy (e.g., $Fe_{73.5}Cu_1Nb_3Si_{13.5}B_9$), followed by controlled heat treatment above the crystallization temperature (ca. 510 °C). This leads to a homogeneous, ultrafine structure of α-FeSi with a grain diameter of 10 – 20 nm, whose formation is attributed to the addition of Cu and Nb [144]. These materials exhibit excellent soft

magnetic properties ($H_c < 1$ A/m, $\mu_i > 10^5$) because K_1 averages zero in the nanocrystalline lattice.

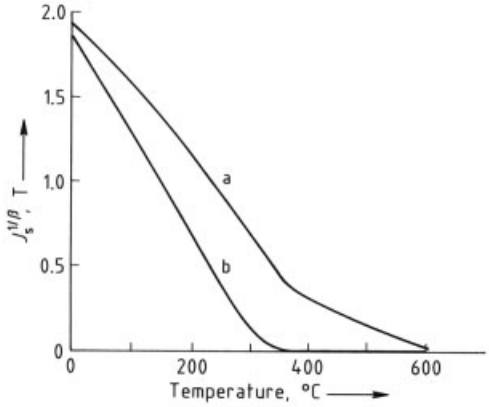

Figure 54. Temperature dependence of the saturation polarization J_s of amorphous and nanocrystalline $Fe_{73.5}Cu_1Nb_3Si_{13.5}B_9$
Ordinate shows $J_s^{1/\beta}$ with $\beta = 0.36$.a) Nanocrystalline (1 h, 520 °C) b) Amorphous, as quenched

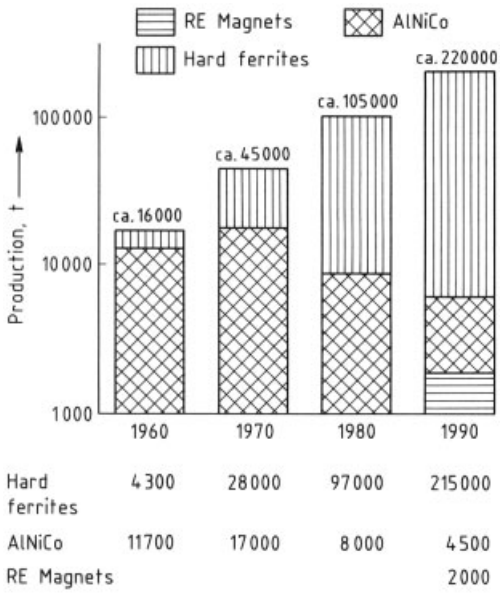

Figure 55. Annual world production of permanent magnets during the period 1960 – 1990

Figure 54 shows the saturation polarization $J_s^{1/\beta}$, with $\beta = 0.36$, as a function of temperature for a quenched probe and for a subsequently annealed probe [146]. After annealing, the Curie temperature rises from 320 to ca. 600 °C, while the saturation polarization at room temperature remains essentially unchanged. The kink in the $J(T)$ curve at ca. 320 °C indicates the presence of a second magnetic phase in the structure of the annealed probe. This is a grain-boundary phase, only a few atomic diameters wide, whose magnetic properties resemble those of the original amorphous phase (Table 19). This phase magnetically couples the nanocrystalline grains to one another. The form of the hysteresis curve of nanocrystalline materials can be influenced by magnetic annealing [147].

The nanocrystalline soft magnetic materials are innovative materials that combine the advantages of the amorphous iron- and cobalt-based alloys and should therefore have many applications, also in competition with classical soft magnetic materials.

6. Applications

6.1. Applications of Hard Magnetic Materials

From 1960 to 1990, the quantity of permanent magnets produced in the Western world increased from 16×10^3 t to 220×10^3 t (Fig. 55). The market share of the hard ferrites increased from ca. 27 % in 1960 to ca. 97 % in 1990. In 1990, ca. 4500 t AlNiCo magnets and ca. 2000 t of rare earth magnets (Sm – Co, Nd – Fe – B) were produced [84], [85], [148]. Figure 56 shows the division of the market for permanent magnets according to end use.

Apart from the magnetic properties and their temperature dependence, one of the most important factors in the choice of a magnetic material for a technical application is the cost. Therefore, because of their low cost, hard ferrites are used wherever small size and low weight are unimportant, whereas the rare-earth magnets are used when miniaturization is important. In the latter applications, the higher cost of the magnets is balanced to some extent by the smaller amounts of material required. AlNiCo magnetics are used in applications where high temperature stability of the magnetic properties is required (e.g., measuring instruments).

Table 19. Characterization of a nanocrystalline $Fe_{73.5}Cu_1Nb_3Si_{13.5}B_9$ alloy after tempering at 500 – 600 °C [146]

Phase	Composition	Grain size	Proportion of volume, %	T_c, °C	K_1
1	α-$FeSi_{20-23}$	10 – 20 nm	75 – 80	600	~ 0
2		intergranular phase $\sim$ 1 nm	20 – 25	320	

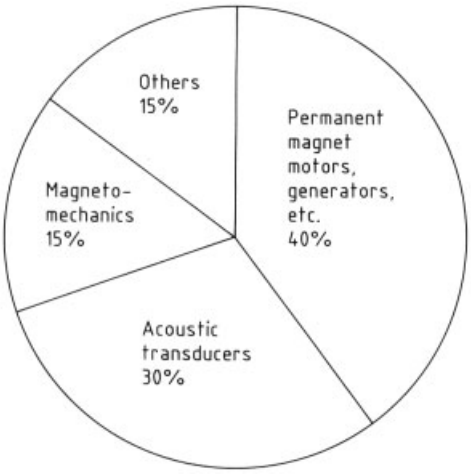

Figure 56. Distribution of the turnover from permanent magnets (ca. $\$ 2.0 \times 10^9$)

A comparison of commercial permanent magnets in terms of $(BH)_{max}$ values, relative costs, temperature behavior, and availability of raw materials is given in Table 20. Table 21 gives an overview of the most important applications of permanent magnets.

Electroacoustic transducers are among the best known applications of permanent magnets. About 30 % of the production of hard magnetic materials is used in loudspeakers, headphones, telephones, microphones, and phono cartridges. This section of the market is dominated by loudspeakers and headphones.

In magnetically driven machines, permanent magnetics provide the magnetic field in which an electrical conductor moves; this results in a force acting on the conductor and induces a potential difference across it. The performance of magnetically driven motors ranges from a few microwatts for quartz-controlled stepping motors in watches, through several kilowatts for large servo motors, and up to 100 kW for large electrical machinery. Magnetically driven motors are more efficient and safer than electrically driven motors.

A major use of permanent magnets in the field of magneto-mechanics are magnetic clutches, which allow contact-free transfer of force. For example, drives for chemical pumps can be manufactured in which the driving force is transferred from the driving half of the clutch through a wall to the other half of the clutch inside the pump. Advantages are that the driving force is transferred contact-free with virtually no losses; high revolution rates can be attained readily, and there are no problems with leakage.

Interesting uses of high-quality rare-earth magnets are found in plasma and particle physics. Permanent-magnet focussing and beam-guiding systems can be made which can act as replacements for energy consuming electrical or superconducting devices. Rare-earth magnets are used in high-intensity synchrotron radiation sources, which offer new and improved applications in medical science, material analysis, and solid-state physics. These devices contain a linear array of magnetic field amplitudes, produced by permanent magnets, which induces an electron beam to emit intensive, almost monochromatic synchrotron radiation [149].

6.2. Applications of Soft Magnetic Materials

Soft magnetic materials have uses in all areas of electrical engineering and electronics. The geometrical form of the magnetic core is of particular importance. These are used as tailor-made parts in magnetic systems or as inductive components. The main function of soft magnetic cores is to guide the magnetic flux as effectively as possible in magnet circuits, or to act as inducers in coil systems.

Table 20. Comparison of industrial permanent magnetic materials

Material	$(BH)_{max}$ value, kJ/m^3	Relative costs	Temperature behavior	Raw material base
Hard ferrite	30	1	satisfactory	very good
AlNiCo	50	4	very good	critical (Co)
SmCo 1 : 5	150	10	satisfactory	critical (Co), limited (Sm)
SmCo 2 : 17	200	7	good	critical (Co), limited (Sm)
1987 NdFeB	280	5	adequate	good
1992	280	3	satisfactory	good

Table 21. Overview of the most important applications of permanent magnets

Range of application	Examples	System function proportional to
Acoustic converters	loud speakers, head phones, telephone capsules, microphones, pickups	B
Motors and generators excited by permanent magnets, electromechanic converters	DC motors, EC motors, synchronous motors, stepping motors, generators, linear motors, torque transmitters	B
Mechanical systems using magnets	clutches, separators, magnetic conveying systems, bearings, transport systems	B^2
Magnetic field and focusing systems	dipoles, quadrupoles, hexapoles, undulators, wigglers, circulators, mass spectrometers, NMR systems	B

Solid magnetic components, produced by grinding or die pressing, are used for guiding magnetic flux, e.g., as pole shoes in electromagnets. Hot- and cold-rolled sheets, formed by stamping and bending, are used as pole pieces in electromechanical systems such as relays. The coercive field strength is the most important criterion for these applications. Casings formed by deep-drawing or by welding of sheets are used to protect sensitive measurement systems, such as automobile tachometers and computer monitors in industrial use, from stray magnetic fields.

Dynamic applications require laminated cores to minimize eddy current losses. High-permeability laminated cores are used in electrical safety applications, e.g., false current safety switches. Core sheets and compacted core sheets are used as in signal engineering. Furthermore, numerous inductive components, such as transformers, repeater coils, transducers, and chokes, require magnetic cores whose dimensions and electrical and magnetic properties are designed for the various operating conditions.

Table 22 gives an overview of the soft magnetic materials and their uses.

7. References

1. H. Wilmesmeier, Dissertation, Universität Göttingen 1985.
2. J. Koch, K. Ruschmeyer: *Permanentmagnete I, Valvo,* Philips GmbH, Hamburg 1983, pp. 164 – 203.
3. P. Erdmann, *Tech. Mitt. Krupp Forschungsber.* **26** (1968) 11 – 16.
4. E. Steingroever, *J. Appl. Phys.* **37** (1966) 1116 – 1117.
5. R. Lee, L. Wilson, C. E. Carter: *Electronic Transformers and Circuits,* 3rd ed., J. Wiley & Sons, Inc., New York 1988.
6. T. Mishima, *Iron Age* **130** (1932) 346.
7. W. Zumbusch, *Elektrotech. Maschinenbau* **60** (1942) 533 – 547.
8. Deutsche Edelstahlwerke, DE 638 652, 1936 (W. Köster).
9. H. Kaneko, M. Homma, K. Nakamura, *AIP Conf. Proc.* **5** (1971) 1088 – 1092.
10. H. Kaneko, M. Homma, K. Nakamura, M. Miura, *IEEE Trans. Magn.* **MAG-8** (1972) 347 – 348.
11. J. J. Went, G. W. Rathenau, E. W. Gorter, G. W. van Oosterhout, *Philips Tech. Rev.* **13** (1952) 194 – 208.
12. G. W. Rathenau, J. Smit, A. L. Stuijts, *Z. Phys.* **133** (1952) 250 – 260.

Table 22. Uses of soft magnetic materials

Material	Lamellated magnetic core		Mouldings	
	Uses	Delivery form	Uses	Delivery form
Classical soft magnetic materials				
Pure iron	radio interference suppression coils	combination cores with NiFe tapes	relays, motors, yokes for electromagnets	punched bent workpieces, finished annealed hardware, semifinished products
Iron – silicon				
nonoriented	rotating machines	core sheets (punched)	relays, motors	punched bent workpieces, finished annealed hardware, semifinished products
grain-oriented	transformers	tape (wound) cores		
Iron – nickel				
E 5			temperature compensation of permanent magnet systems	punched workpieces, finished annealed semifinished products
E 4	transformers for communication engineering, pulse transformers, throttles	core sheets (punched), tape cores/tape-wound cores	screenings, relay and magnet systems, pole screen and other flux-guiding workpieces	deep-drawing workpieces, connecting workpieces, finished annealed hardware, semifinished products (punched, sheets)
E 3	current transformers, throttles, transformers for communication engineering	toroidal tape cores, tape cores with air gap, tape cores, core sheets, laminated cores	relays, yokes/rotors, measuring systems	finished annealed hardware, semifinished products
E 2	current transformers, thyristor protection throttles	toroidal tape cores		
E 1	current transformers, transformers for communication engineering	toroidal tape cores, core sheets, laminated cores	screenings, highly sensitive relays	deep-drawing workpieces, connecting workpieces, semifinished products (plates, sheets), punched bent workpieces
Iron – cobalt	rotating machines, transformers	core sheets (punched), tape cores	flux-guiding workpieces	finished annealed hardware, semifinished products
Soft ferrite			inductive structural components	standardized mouldings
Modern soft magnetic materials				
Powder-bound composite			rotating machines, throttles, transformers	pressed and mechanically-worked stator workpieces, toroidal core (pressed), coated cores
Amorphous				
Fe-based	transformers	(toroidal) tape cores		
Co-based	transformers, throttles, sensors	toroidal tape cores		
Nanocrystalline	transformers, throttles	toroidal tape cores		

13. H. Zijlstra, *J. Appl. Phys.* **27** (1956) 1249 – 1250.
14. N. Makino, Y. Kimura, J. Yamahi, *J. Jpn. Inst. Met.* 1963, 582 – 587.
15. K. J. de Vos, *Z. Angew. Phys.* **14** (1962) 253 – 254.
16. A. J. Koch, M. G. van der Steeg, K. J. de Vos, *Proc. Conf. Magnetism Magnetic Materials,* Boston 1956, pp. 63 – 65.
17. A. J. Koch, M. G. van der Steeg, K. J. de Vos, *Ber. Arbeitsgem. Ferromagn.* 1959, 130 – 139.
18. W. Baran, *Tech. Mitt. Krupp Forschungsber.* **17** (1959) 150 – 152.
19. R. Wittig, *Z. Angew. Phys.* **14** (1962) 253 – 254.
20. E. Steinort, *Ber. Arbeitsgem. Ferromagn.,* Tagung Karlsruhe 1962, Bericht III.
21. K. J. de Vos, *Z. Angew. Phys.* **17** (1964) 168 – 174.
22. M. McCaig, *Z. Angew. Phys.* **21** (1966) 66 – 68.
23. W. Ervens, *Tech. Mitt. Krupp Forschungsber.* **40** (1982) 109 – 116.
24. H. Dombre et al., *J. Less Common Met.* **66** (1979) 1 – 11.

25. H. Kaneko et al., *IEEE Trans. Magn.* **MAG-13** (1977) 1325 – 1327.
26. T. Minowa, M. Okada, M. Homma, *IEEE Trans. Magn.* **MAG-16** (1980) 529 – 533.
27. S. Jin, *IEEE Trans. Magn.* **MAG-15** (1979) 1748 – 1750.
28. G. Y. Chin et al., *J. Appl. Phys.* **52** (1981) 2536 – 2541.
29. M. Homma, E. Horikoshi, E. Minowa, T. Okada, *Appl. Phys. Lett.* **37** (1980) 92 – 93.
30. W. Jellinghaus, *Z. Tech. Phys.* **17** (1936) 33 – 36.
31. E. A. Nesbitt, G. A. Kelsall, *Phys. Rev.* **58** (1940) 203.
32. H. Fahlenbrach, *Metall (Berlin)* **7** (1953) 413 – 421.
33. W. Köhler, W. Dannöhl, *Z. Metallkd.* **27** (1935) 220 – 226.
34. O. Dahl, J. Pfaffenberger, W. Schwarz, *Metallwirtsch. Metallwiss. Metalltech.* **14** (1935) 665 – 670.
35. H. Neumann, A. Büchner, H. Reinboth, *Z. Metallkd.* **29** (1937) 173 – 185.
36. T. O. Paine, L. J. Mendelsohn, F. E. Luborsky, *Phys. Rev.* **100** (1955) 1055 – 1059.
37. F. E. Luborsky, *J. Appl. Phys.* **33** (1962) 2385 – 2390.
38. A. J. J. Koch, P. Hokkeling, M. G. v. d. Steeg, K. J. de Vos, *J. Appl. Phys.* **31** (1960) Suppl. 75 – 77.
39. T. Ohtani et al., *IEEE Trans. Magn.* **MAG-13** (1977) 1328 – 1330.
40. Y. Sakamoto, A. Ibata, S. Kojima, T. Ohtani, *IEEE Trans. Magn.* **MAG-16** (1980) 1056 – 1058.
41. W. Ervens, *Tech. Mitt. Krupp Forschungsber.* **40** (1982) 117 – 122.
42. E. W. Gorter, *Philips Res. Rep.* **9** (1954) 295 – 365, 403 – 443.
43. H. Stäblein, *Tech. Mitt. Krupp Forschungsber.* **40** (1982) 85 – 92.
44. H. Stäblein in E. P. Wohlfarth (ed.): *Ferromagnetic Materials,* **vol. 3,** North Holland Publ., Amsterdam –New York – Oxford 1982, pp. 442 – 602.
45. R. Ebeling, H. Krause, *Tech. Mitt. Krupp Forschungsber.* **40** (1982) 93 – 98.
46. K. J. Strnat, G. Hoffer, J. Olson, W. Ostertag, *J. Appl. Phys.* **38** (1967) 1001 – 1002.
47. K. J. Strnat, *Kobalt* **36** (1967) 119 – 128.
48. Y. Tawara, H. Senno, *Jpn. J. Phys.* **12** (1973) 761 –763.
49. K. T. Inomata, T. Oshima, T. Ido, M. Yamada, *Appl. Phys. Lett.* **30** (1977) 669 – 670.
50. T. S. Ojima, S. Tomizawa, T. Yoneyama, T. Hori, *IEEE Trans. Magn.* **MAG-13** (1977) 1317 – 1390.
51. M. Sagawa et al., *J. Appl. Phys.* **55** (1984) 2083 –2087.
52. J. J. Croat, J. F. Herbst, R. W. Lee, F. E. Pinkerton, *J. Appl. Phys.* **55** (1984) 2076 – 2082.
53. P. Falconnet, *Workshop NdFe Permanent Magn.,* Brussels Oct. 25, 1984, CEC, 15 – 23.
54. W. Ervens, *Tech. Mitt. Krupp Forschungsber.* **40** (1982) 99 – 107; *Workshop NdFe Permanent Magn.,* Brussels Oct. 25, 1984, CEC, 111 – 122.
55. C. Herget, H. G. Domazer, *Goldschmidt Informiert* **35** (1975) no. 4, 3 – 33.
56. C. Herget, *Proc. Workshop Rare Earth Magnets Their Appl.* **8th** (1985) 407 – 422.
57. H. Edeling, C. Herget, *Proc. Workshop Rare Earth Magnets Their Appl.* **10th** (1989) 439 – 447.
58. I. R. Harris, *Proc. Workshop Rare Earth Magnets Their Appl.* **9th** (1987) 249 – 265.
59. K. H. Buschow, A. S. van der Groot, *J. Less Common Met.* **14** (1968) 323 – 328.
60. K. J. Strnat in E. P. Wohlfarth, K. H. J. Buschow (eds.): *Ferromagnetic Materials,* **vol. 4,** North Holland Publ., Amsterdam 1988, pp. 131 – 209.
61. C. W. Allen, D. L. Kuruzar, A. E. Miller, *IEEE Trans. Magn.* **MAG-10** (1974) 716 – 719.
62. C. W. Allen, K. C. Liar, A. E. Miller, *J. Less Common Met.* **52** (1977) 109 – 112.
63. H. R. Kirchmayr, C. A. Poldy, *J. Magn. Magn. Mater.* **8** (1978) 1 – 42.
64. M. G. Wells, K. S. V. L. Narasimhan, *Goldschmidt Informiert* **48** (1979) 15 – 22.
65. H. A. Leupold et al., *Proc. Int. Symp. Magn. Anisotropy Coercivity Rare Earth Transition Met. Alloys 2nd* 1978, 87 – 101.
66. A. Menth, H. Nagel, *Proc. Int. Symp. Magn. Anisotropy Coercivity Rare Earth Transition Met. Alloys 2nd* 1978, 65 – 86.
67. W. E. Wallace, V. U. S. Rao, S. G. Sankar, M. Merches, *Proc. Int. Symp. Magn. Anisotropy Coercivity Rare Earth Transition Met. Alloys 2nd* 1978, 118 –136.
68. H. R. Kirchmayr, *Appl. Phys.* **1** (1973) 187 – 201.
69. W. Ervens, *Proc. Int. Workshop Rare Earth Cobalt Perm. Magnets Their Appl.* **6th** (1982) 319 – 327.
70. H. Kronmüller, K. H. Durst, W. Ervens, W. Fernengel, *IEEE Trans. Magn.* **MAG-20** (1984) 1569 – 1571.

71. K. H. Durst, A. Kronmüller, W. Ervens, *Phys. Status Solidi A* **108** (1988) 403 – 416, 705 – 719.
72. K. J. Strnat, A. Tauber, *J. Less Common Met.* **93** (1983) 269 – 273.
73. H. F. Mildrum, J. B. Krupar, A. E. Ray, *J. Less Common Met.* **93** (1983) 261 – 273.
74. J. F. Herbst, J. J. Croat, F. E. Pinkerton, W. B. Yelon, *Phys. Rev. B: Condens. Matter* **29** (1984) 4176 – 4178.
75. K. H. J. Buschow: "Permanent Magnet Materials Based On 3 d-Rich Ternary Compounds," in E. P. Wohlfarth, K. H. J. Buschow (eds.): *Ferromagnetic Materials,* vol. 4, North-Holland, Amsterdam – New York – Oxford 1988, pp. 1 – 129.
76. R. W. Lee, *Appl. Phys. Lett.* **46** (1985) 790 – 791.
77. R. W. Lee, E. G. Brewer, N. A. Schaffel, *IEEE Trans. Magn.* **MAG-21** (1985) 1958 – 1961.
78. L. Schulz, J. Wecker, *Proc. Int. Workshop On Rare Earth Magnets Their Appl.* **9th** (1987) 301 – 309.
79. R. W. Evans, A. G. Leatham, R. G. Brooks, *Powder Metall.* **28** (1985) 13 – 20.
80. T. S. Chin et al., *J. Appl. Phys.* **59** (1986) 1297 – 1300.
81. W. Ervens in I. V. Mitchell et al.: *Concerted European Action On Magnets (CEAM),* Elsevier Appl. Sci., London – New York 1989, pp. 655 – 658.
82. T. Shimoda, K. Akioka, O. Kobayashi, T. Yamagami, *Proc. Int. Workshop On Rare Earth Magnets Their Appl.* **10th, vol. I** (1989) 389 – 398.
83. T. Ohki et al., *Proc. Int. Workshop On Rare Earth Magnets Their Appl.* **10th, vol. I** (1989) 399 – 408.
84. H. Stäblein, W. Baran, *Tech. Mitt. Krupp Forschungsber.* **43** (1985) 43 – 58.
85. R. Ebeling, W. Ervens, *Tech. Mitt. Krupp Forschungsber.* 1988, 15 – 30.
86. C. Kittel, *Phys. Rev.* **73** (1948) 810.
87. M. Kersten: *Grundlagen einer Theorie der ferromagnetischen Hysterese und der Koerzitivkraft,* Hirzel, Leipzig 1944, p. 99.
88. DIN 46 400, Parts 1 – 4.
89. IEC-Publications 404-8-2, 404-8-3, 404-8-4, 404-8-5, 404-8-7.
90. K. Honma et al., *IEEE Trans. Magn.* **MAG-21** (1985) 1903.
91. H. C. Heck: *Magnetische Werkstoffe und ihre technische Anwendung,* 2nd ed., Hüthig Verlag, Heidelberg 1975.
92. O. Kubaschewski: *Iron-Binary Phase Diagrams,* Springer Verlag, Berlin – Heidelberg – New York 1982.
93. H. Huneus, K. H. Schmidt, K. Hartig, *IEEE Trans. Magn.* **MAG-18** (1982) 246.
94. A. L. Geiger, *J. Appl. Phys.* **50** (1979) 2366.
95. I. Bóc et al., *J. Appl. Phys.* **64** (1988) 5350.
96. E. Höffken, H. D. Pflipsen, R. A. Weber, *Stahl Eisen* **103** (1983) 155.
97. K. Günther, F. Bölling, H. Huneus, *J. Appl. Phys.* **64** (1988) 5347.
98. F. Bölling et al.: *ETG Fachbericht 29: "* Magnetische Werkstoffe, Varistoren und Supraleiter: Anwendungen und Entwicklungstendenzen," VDE-Verlag, Berlin – Offenbach 1989, p. 5 ff.
99. N. P. Goss, US 1 965 559, 1934.
100. S. Taguchi, T. Yamamoto, A. Sakakura, *IEEE Trans. Magn.* **MAG-10** (1974) 123.
101. T. Yamaguchi, K. Takeda, *IEEE Trans. Magn.* **MAG-21** (1985) 41.
102. J. G. Benford, *IEEE Trans. Magn.* **MAG-20** (1984) 1545.
103. I. Ichijima, H. Nakamura, T. Nozawa, T. Nakata, *IEEE Trans. Magn.* **MAG-20** (1984) 1557.
104. K. I. Arai, K. Ishiyama, H. Mogi, *IEEE Trans. Magn.* **MAG-25** (1989) 3949.
105. G. Bertotti, A. Ferro Milone, F. Fiorillo, G. P. Soardo, *J. Phys. C* **46** (1985) C 6 – 149.
106. M. Espenhan, G. Giermann, W. Bald, H. Römmen, *Stahl Eisen* **108** (1988) no. 18, 853.
107. R. Baehr, *ETZ Elektrotech. Z.* **109** (1988) 1047.
108. Stahl Informationszentrum Düsseldorf, *Stahl Eisen* **109** (1989) 441.
109. G. Abbruzzese, *IEEE Trans. Magn.* **MAG-25** (1989) 3955.
110. R. H. Pry, C. P. Bean, *J. Appl. Phys.* **29** (1958) 532.
111. G. Bertotti, *J. Magn. Magn. Mat.* **54 – 57** (1986) 1556.
112. G. Bertotti, *IEEE Trans. Magn.* **MAG-24** (1988) 621.
113. G. Bertotti, *J. Appl. Phys.* **57** (1985) 2110.
114. F. Pfeifer in K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970, p. 56.
115. F. Assmus, *Arch. Tech. Messen Ind. Messtech.* **393/394** (1968) 227 – 232, 253 – 256.
116. G. Y. Chin, *IEEE Trans. Magn.* **MAG-7** (1971) 102.
117. F. Pfeifer in K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970, p. 58.

118. F. Pfeifer in K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970, p. 61.
119. G. W. Elmen, *J. Franklin Inst.* **207** (1929) 602.
120. F. Pfeifer in K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970, p. 73.
121. F. Assmus, F. Pfeifer, *Metall (Berlin)* **7** (1953) 189.
122. O. I. Boothby, R. M. Bozorth, J. Appl. Phys. **18** (1947) 173.
123. E. Josso, *IEEE Trans. Magn.* **MAG-10** (1974) 161.
124. R. V. Major, C. M. Orrock, *IEEE Trans. Magn.* **MAG-24** (1988) 1856.
125. T. Postupolski, *J. Phys. C* **46** (1985) C 6 – 159.
126. R. Dreyer: "Ferrite und induktive Bauelemente für die Leistungselektronik," in G. Taubitz (ed.): *ZVEI Symposium '87,* Frankfurt 1987, Fachverband Bauelemente der Elektronik im ZVEI, Pforzheim 1987, p. 5.
127. E. Roess, *Elektronik-Technol. Elektronik-Anwend. Elektronik-Mark.* **10** (1989) 48.
128. O. Heaviside, *Electrician* **397** (1887) 302.
129. H. Vogt, AT 140 508, 1935.
130. W. Klement, R. H. Willens, P. Duwez, *Nature (London)* **187** (1960) 869.
131. P. Duwez, S. C. H. Lin, *J. Appl. Phys.* **38** (1967) 4096.
132. H. H. Liebermann, *IEEE Trans. Magn.* **MAG-15** (1979) 1393.
133. H. R. Hilzinger, S. Hock, *Proc. Conf. Met. Glasses: Sci. Technol.* 1980.
134. R. Boll, H. R. Hilzinger, *ETZ Elektrotech. Z.* **102** (1981) 1096.
135. H. R. Hilzinger, *IEEE Trans. Magn.* **MAG-21** (1985) 2020.
136. H. Kronmüller, *J. Magn. Magn. Mat.* **24** (1981) 159.
137. H. R. Hilzinger, W. Kunz, *IEEE Trans. Magn.* **MAG-20** (1983) 1323.
138. H. H. Liebermann, N. DeCristofaro, *CHEMTECH* 1987, 367.
139. F. E. Luborsky, P. G. Frischmann, L. A. Johnson, *J. Magn. Magn. Mat.* **8** (1978) 318.
140. F. E. Luborsky, *IEEE Trans. Magn.* **MAG-14** (1978) 1008.
141. M. P. Sampat, L. A. Lowdermilk, *International Conference on Transformers,* New Delhi 1988.
142. G. Herzer, H. R. Hilzinger, *Phys. Scr.* **T 24** (1988) 22.
143. R. Boll, H. Warlimont, *IEEE Trans. Magn.* **MAG-17** (1981) 3053.
144. Y. Yoshizawa, S. Oguma, K. Yamauchi, *J. Appl. Phys.* **64** (1988) 6044.
145. Y. Yoshizawa, K. Yamauchi, S. Oguma, EP-A 0 271 657, 1988.
146. G. Herzer, *IEEE Trans. Magn.* **MAG-25** (1989) 3327.
147. Y. Yoshizawa, K. Yamauchi, *IEEE Trans. Magn.* **MAG-25** (1989) 3325.
148. W. Baran, R. Ebeling, D. Fernow: *ETG Fachbericht 29: Magnetische Werkstoffe, Varistoren und Werkstoffe,* VDE-Verlag, Berlin – Offenbach 1989, pp. 77 –88.
149. K. Halbach, *Nucl. Instrum. Methods Phys. Res.* **187** (1981) 109 – 117.
150. R. Lee, L. Wilson, C. E. Carter: *Electronic Transformers and Circuits,* 3rd ed., John Wiley & Sons, New York 1988.

Maleic and Fumaric Acids

Kurt Lohbeck, Hüls Aktiengesellschaft, Bottrop, Federal Republic of Germany (Chaps. 1–4)

Herbert Haferkorn, Hüls Aktiengesellschaft , Bottrop, Federal Republic of Germany (Chaps. 1–4)

Werner Fuhrmann, Hüls Aktiengesellschaft, Bottrop, Federal Republic of Germany (Chaps. 1–4)

Norbert Fedtke, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Chap. 5)

1. Maleic Acid 1
2. Maleic Anhydride (MA) 2
2.1. Properties 2
2.2. Production 2
2.2.1. Oxidation of Benzene 3
2.2.2. Dehydration of Aqueous Maleic Acid Solutions 3
2.2.3. Oxidation of C_4 Hydrocarbons 4
2.2.4. Purification 5
2.2.5. Byproducts, Construction Materials, Effluent, and Discharge Air 6
2.3. Quality, Storage, and Transportation 6
2.4. Uses 7
3. Citraconic and Mesaconic Acids . . 7
4. Fumaric Acid 8
5. Toxicology 9
6. References 10

1. Maleic Acid

Maleic acid [*110-16-7*], $C_4H_4O_4$, *cis*-butenedioic acid, M_r 116.07, does not occur naturally. It was first produced by Pelouze in 1834 by heating malic acid. Maleic acid became industrially important after the Barrett Co. in the United States successfully obtained the acid in 1919 by catalytic gas-phase oxidation of benzene. Maleic acid has been available commercially since 1928; maleic anhydride, since 1933.

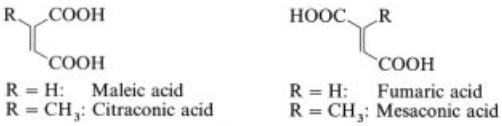

R = H: Maleic acid
R = CH_3: Citraconic acid

R = H: Fumaric acid
R = CH_3: Mesaconic acid

Maleic acid itself is not very important economically. Small amounts are used for maleinate resins and, as maleic acid esters, for copolymers.

Physical Properties. The acid crystallizes in the form of monoclinic prisms. Other physical properties are listed below:

mp	130.5 °C
d_{20}^{20} (solid)	1.590
Heat of formation	− 788.3 kJ/mol
Heat of combustion	− 1358.9 kJ/mol
Solubility in 100 g of water	
at 25 °C	78.9 g
at 40 °C	112.3 g
at 60 °C	148.8 g
at 97.5 °C	392.6 g
Dissociation constants in water	
at 25 °C	K_1 1.14×10^{-2}
	K_2 5.95×10^{-7}

Chemical Properties. Maleic acid is very reactive at both its carboxyl groups and its double bond. Thus, on heating above 100 °C, water is eliminated with the formation of maleic anhydride. On further heating–preferably with the addition of catalysts–decarboxylation occurs with the formation of acrylic acid. Maleic acid also undergoes all normal reactions of *carboxylic acids,* such as esterification and amidation, but it does not form an acid chloride.

Examples of reactions involving the *double bond* are addition of water to form malic acid at elevated temperature and pressure; reaction with ozone to form glyoxylic acid; addition of halogens to form compounds such as dichlorosuccinic acid; and catalytic hydrogenation to succinic acid. Isomerization to fumaric acid occurs slowly already at temperatures as low as 100 °C; in the presence of a catalyst it proceeds almost quantitatively even at 60 °C.

Production. Maleic acid is generally obtained by heating maleic anhydride in a small

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_053

amount of water. After the reaction mixture is cooled, the supernatant mother liquor is separated and the solid product is dried in vacuo. The mother liquor is diluted with water and returned to the hydration stage. It can, however, also be passed to the dehydration stage of maleic anhydride production (see Section 2.2.2), which avoids the problem of effluent purification. The apparatus must be made of corrosion-resistant steel.

Occasionally, maleic acid is also obtained directly from the wash water of maleic anhydride production; energy required for dehydration to the anhydride is thus saved. However, the wash water must first be purified with active charcoal; the charcoal must then be removed and disposed of. Furthermore, the purified solution must be concentrated in vacuo because fumaric acid is formed at elevated temperature and the chrome – nickel steel apparatus corrodes above 80 °C. Enriched wash water from *o*-xylene and naphthalene oxidation plants (phthalic anhydride production) is also used as a starting material for this process, as well as for the production of fumaric acid and maleic anhydride.

Quality Requirements. Maleic acid must meet stringent quality requirements. White crystals with a maximum fumaric acid content of 0.1 – 0.2 % are required, and the water content should be < 0.5 %.

2. Maleic Anhydride (MA)

Maleic anhydride [*108-31-6*], $C_4H_2O_3$, 2,5-furandione, M_r 98.06, has considerable industrial importance, in contrast to maleic acid.

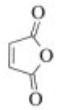

Maleic anhydride

2.1. Properties

Physical Properties. Maleic anhydride crystallizes as orthorhombic needles. The most important physical properties may be summarized as follows:

mp	52.85 °C
bp (101.3 kPa)	202.0 °C
Heat of combustion	− 1391.2 kJ/mol
Specific heat (liquid)	− 1.67 kJ mol^{-1} K^{-1}
Heat of evaporation	54.8 kJ/mol
Heat of fusion	13.66 kJ/mol
Solubility in xylene (29.7 °C)	163.2 g/L

Vapor pressure of maleic anhydride at various temperatures:

t, °C	44.0	78.7	111.8	135.8	179.5	202.0
p, kPa	0.13	1.33	5.33	13.33	53.33	101.33

Chemical Properties. Maleic anhydride reacts readily with water to form maleic acid. At low temperatures, addition of alcohols leads to the formation of semiesters, whereas at elevated temperature in the presence of esterification catalysts, water is eliminated with the formation of diesters. Addition of ammonia or amines to maleic anhydride results in formation of the corresponding semiamides, which can be converted by elimination of water into cyclic imides.

The double bond in maleic anhydride is also extremely reactive. Addition of halogens leads to the formation of dihalogenated succinic anhydrides or mono- and dihalogenated maleic anhydrides, depending on reaction conditions. Hydrogenation yields succinic anhydride, 1,4-butanediol, tetrahydrofuran, or butyrolactone, depending on conditions.

Alkenylsuccinic anhydrides are formed by addition of olefins. Maleic anhydride undergoes Diels – Alder reactions with conjugated dienes. The anhydride is also suitable for homopolymerization and copolymerization. A summary of the reactions of maleic anhydride is given in [1], [2].

2.2. Production

Maleic anhydride is produced industrially by catalytic oxidation of suitable hydrocarbons in the gas phase. Benzene used to be the predominant starting material, but in recent years oxidation of C_4 hydrocarbons has become increasingly important; the C_4-based capacities are continually increasing.

2.2.1. Oxidation of Benzene

To produce maleic anhydride (see Fig. 1), benzene is added to a preheated airstream so as to form a largely homogeneous mixture. If a large excess of air is used, the lower ignition limit of the mixture is not reached in the majority of methods.

Tubular reactors with a multiplicity of parallel vertical tubes of ca. 25-mm diameter are used for conventional fixed-bed methods. Reaction gas passes through the tubes over the catalyst, the process pressure being adjusted to an optimum value of 0.15 – 0.25 MPa. Supported catalysts are used, where an "active mass" of V_2O_5 and MoO_3 is applied on an inert carrier. The lifetime of the catalysts is up to four years and depends mainly on the operating temperature in the reactor; a uniform, interference-free operation; and the purity of starting materials. Both the desired main reaction

$$C_6H_6 + 4\,1/2\,O_2 \rightarrow C_4H_2O_3 + 2\,CO_2 + 2\,H_2O$$

$$\Delta H = -1875\ \text{kJ/mol}$$

and the undesirable secondary reaction of combustion to carbon monoxide, carbon dioxide and water are strongly exothermic, and "hot spots" of 340 – 500 °C occur in the catalyst. A considerable amount of heat is produced during the reaction. Approximately 27 MJ of heat must be dissipated per tonne of reacted benzene, which corresponds to almost 10 t of saturated steam; in practice, ca. 80 % of the steam can be utilized.

Heat is dissipated primarily by eutectic salt melts circulating around the reactor tubes. The hot salt melts are cooled with water in a secondary heat exchanger. Only part of the generated steam can be used in MA production. In many cases, excess steam–unless it can be utilized in other operations–is used to drive air compressors. The reaction can be optimized with regard to MA yield by means of the accurately adjustable (± 0.5 °C) salt bath temperature.

In a typical conversion of 100 mol benzene, 73 mol is oxidized to maleic anhydride and 23 mol is "burned." Unreacted benzene (4 mol) can largely be recovered by a subsequent adsorption step (e.g., with activated charcoal) and reused. Processes that involve recycling the major portion of reaction gas after MA separation are described in some patents; up to the present, they have not been used on an industrial scale [3], [4].

Separation of Crude Maleic Anhydride. After leaving the reactor, the gas mixture is initially cooled to 150 – 160 °C so that neither MA nor water condenses. The waste heat is utilized further. Different routes can then be used for further processing.

Partial Condensation. In liquid partial condensation the melting point of MA is the lower temperature limit. The reaction gas is cooled to ca. 55 °C in "partial condensers" [5] (e.g., bundles of vertical tubes 20 – 50 mm in diameter); MA condenses and can be removed as liquid from the collecting vessel. Discharge should occur as quickly as possible [6] because prolonged contact of the liquid product with the water-containing reaction gas leads to the formation of fairly large amounts of maleic acid. In this way, 40 – 60 % of the MA contained in the reaction gas can be separated; the remainder is normally washed out with water as maleic acid and processed further to MA (see Section 2.2.2).

Separation by Water Scrubbing. In principle, all the MA contained in the reaction gas can also be recovered as maleic acid solution by water scrubbing. To obtain MA from this solution, the acid must be dehydrated. Because of the associated high energy requirement, this method is justified only to obtain MA from reaction gases with a high water content, such as occur in the C_4 oxidation.

For separation by washing with organic solvents, see Section 2.2.3.

2.2.2. Dehydration of Aqueous Maleic Acid Solutions

Maleic anhydride can be obtained by dehydration from the maleic acid solutions that are formed by washing the reaction gas with water (see Fig. 1). Above 150 °C even small amounts of alkali in the wash solution can decompose the resultant crude MA to decarboxylation products [7]; thus, alkali-free wash water must be used.

Temperatures above 100 °C are required for dehydration of maleic acid ($\Delta H = +34.88$ kJ/mol). A reaction rate sufficient for industrial processes is, however,

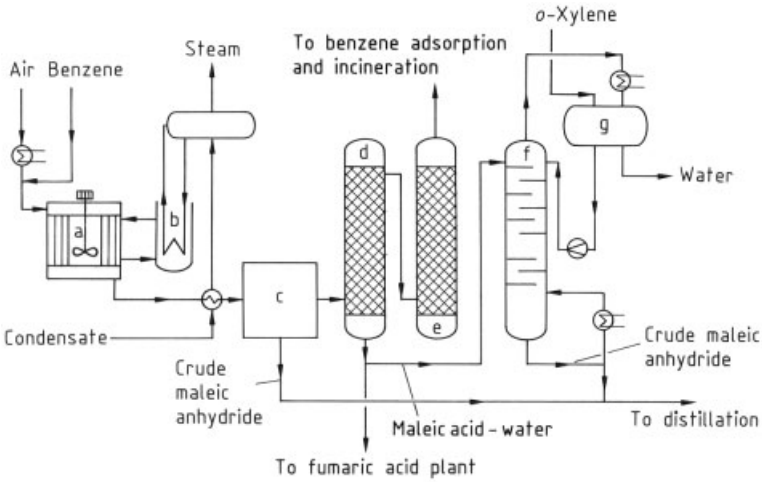

Figure 1. Plant for production of maleic anhydride by benzene oxidation
a) Reactor; b) Salt bath cooler; c) Partial condenser; d) Acid scrubber; e) Alkali scrubber; f) Dehydration column; g) Phase separator

achieved only above 130 °C. Isomerization of maleic acid to fumaric acid occurs concurrently ($\Delta H = -23.2$ kJ/mol).

Fumaric acid decomposes only when heated above 230 °C, without forming significant amounts of MA. The isomerization reaction must therefore be suppressed in the dehydration of maleic acid; this is achieved by having as short a residence time as possible at elevated temperature.

Entraining Agent Method. Aqueous maleic acid solution is added to the upper half of a distillation column (Fig. 1, f) whose bottom contains a boiling MA – entraining agent mixture. Aromatic hydrocarbons (e.g., xylene) are generally used as entrainers.

The entraining agent together with the water leaves the head of the column as an azeotropic mixture. Then it is separated as the organic phase in separator (g) and refluxed to the head of the column. The aqueous phase is returned to the reaction gas scrubber (d) of the oxidation plant [8].

The bottom product (MA, 10 – 40 % entraining agent, 1 – 5 % maleic acid, 1 – 3 % fumaric acid) is distilled to yield pure MA (see Section 2.2.4). The process is performed continuously as well as batchwise.

Thermal Process Without Entraining Agent. The thin-layer evaporator technique has been used for some time for thermal dehydration of aqueous maleic acid solutions. By this technique the water can be separated as steam, and the residence time of maleic acid at 150 – 200 °C is kept very short. Conditions for isomerization thus exist for only a short time, and only a small amount of fumaric acid is formed (1 – 3 %). Several thin-layer units are arranged in sequence [9], [10]. The MA separated as liquid in the partial condenser (see Fig. 1) is also fed to the second stage of this plant to dehydrate residual maleic acid (1 – 5 %). The 99 % MA obtained is distilled to yield pure product (see Section 2.2.4).

Fumaric acid and high-boiling residues are discharged from the bottom of the last thin-layer evaporator.

These continuously operating plants must be shut down relatively frequently for cleaning.

2.2.3. Oxidation of C_4 Hydrocarbons

Processes for the oxidation of C_4 hydrocarbons involve mainly the use of *n*-butane or *n*-butane – *n*-butene mixtures with a high paraffin content. The following types of process have been developed or are in the developmental stage:

fixed-bed process
fluidized-bed process
transport-bed process

In the United States the fixed-bed process with *n*-butane has been the only MA route used commercially since 1985; it is employed to a small but increasing extent in other parts of the world. With improved methods and sufficient raw material availability, C_4 oxidation has become increasingly important; the low raw material price has had a decisive influence on the economy of this process.

With C_4 oxidation lower conversion rates (80 %) and selectivities (70 %) contrast theoretically higher weight yields (+ 34 wt % compared to benzene oxidation) so that finally, weight yields for benzene and C_4 oxidation are comparable (ca. 95 wt % MA). Since the recovery of unreacted C_4 hydrocarbons is substantially more difficult than with benzene, reaction waste gases from C_4 oxidation must be passed to an incineration unit.

Fixed-Bed Process. In the fixed-bed process the catalyst is embedded in tubular reactors, similar to benzene oxidation. Since the publication of the first U.S. patents catalysts based on vanadium and phosphorus oxides (V – P – O catalysts) have mainly been used [11–15].

The concentration of hydrocarbons in the feed gas to the reactor is kept just below the explosion limit of 1.8 mol % [corresponding to 45 g/m^3 (STP) air] in the commercial fixed-bed process. Conversion at the catalyst in the selective reaction proceeds as follows:

$$C_4H_{10} + 3\,1/2\,O_2 \rightarrow C_4H_2O_3 + 4\,H_2O$$

$$\Delta H = -1260\ \text{kJ/mol}$$

Because of the increased formation of water per formula conversion compared to benzene oxidation, with C_4 oxidation only a fairly small amount of MA can be directly liquefied from the reaction gas by partial condensation. The larger proportion (65 – 70 %) is washed out as maleic acid in a subsequent water scrubber and must be dehydrated, with a high energy input.

In a modification of the water wash, MA contained in the reaction gas is alternatively absorbed by organic solvents [16–19]. More than 98 % of the anhydride can be absorbed in this way, without noticeable amounts of maleic acid being formed. The solvent – anhydride mixture is then subjected to fractional distillation, in which MA is distilled from the high-boiling solvent and the latter is returned to the absorption column.

Fluidized-Bed Process. The fluidized-bed process has the advantage of a particularly uniform temperature profile (without "hot spots"). A uniform temperature profile is the necessary precondition for high reaction selectivity, it is, however, offset to some extent by intensive remixing of the products as a result of fluidization. A particular problem of this process is mechanical stress on the catalyst (V – P – O), its abrasion, and erosion at the heat-dissipating surfaces. On the other hand, the fluidized bed constitutes an extremely effective flame barrier, with the result that the process can operate at higher C_4 concentrations (i.e., within the explosion range) than the fixed-bed process. This results in process engineering advantages plus a considerably smaller investment for fluidized-bed technology [20].

The *ALMA process* (Alusuisse, Lummus) is currently the most technically advanced process; it combines a fluidized-bed reactor with a nonaqueous processing unit [18], [19].

In October 1988 a 10 000-t/a MA plant [21] combining fluidized-bed technology with an aqueous processing unit has come on stream at MTC Sohio, Japan.

Transport-Bed Process. The transport-bed process developed jointly by Monsanto and Du Pont is at the trial stage [22], [23]. The process involves two reactors: in the first, spent catalyst (V – P – O) removed from C_4 oxidation is regenerated by atmospheric oxygen. Regenerated (oxidized) catalyst is separated from the oxygen-containing gas; then it passes, with the addition of butane, to the second reactor. Here, catalytic conversion to MA takes place, the reaction being carried out with virtual exclusion of atmospheric oxygen and simply with the amount of oxygen transferred by the catalyst itself. A high selectivity of MA formation (75 mol %) is said to be achieved under the quasi-stoichiometric conversion conditions of this process, which also largely prevents remixing of reaction gases.

2.2.4. Purification

Batchwise Distillation. Batchwise distillation is the conventional method for processing

crude MA. Directly separated MA, as well as mixtures from entraining agent dehydration, are added to the batch distillation column. The mixture is distilled under total reflux conditions until the remaining maleic acid is dehydrated. The small amounts of water eliminated are separated from the recirculating entraining agent (xylene). The entrainer is then distilled, and MA is recovered by distillation in vacuo. Batch stills of 50–150-m^3 capacity and distillation columns with 10–20 trays are used.

Continuous Distillation (see Fig. 2). For large-capacity MA plants, processing by continuous distillation is both possible and economical [24], [25]. The mixture of directly separated and crude MA obtained by dehydration is continuously "postdehydrated". Residual xylene is separated in a column (f), and pure MA is finally distilled from the top of a second column (g).

2.2.5. Byproducts, Construction Materials, Effluent, and Discharge Air

Benzene Oxidation.

Byproducts. Benzoquinone, formaldehyde, and formic acid occur in small amounts as byproducts of benzene oxidation.

Construction Materials. Directly condensed crude MA containing 1–5 % maleic acid can be stored at ca. 60 °C in normal steel vessels. Special stainless steel is essential in the "wet" region of the plants (i.e., for scrubbers, vessels containing maleic acid solution, dehydration column, and apparatus used for postdehydration). The partial condensers should preferably also be made of special stainless steel. For reasons of quality, stainless steel is used for the pure MA distillation columns.

C_4 Oxidation. Because acetic, acrylic, and crotonic acids, etc., are formed as byproducts of C_4 oxidation, a suitable special stainless steel must be chosen with particular care for the relevant equipment.

Environmental Protection. The MA plants can be operated almost without any acid effluent if aqueous maleic acid solutions from the dehydration stage are recycled. Byproducts are contained in the distillation residue, which is normally combusted.

Environmental legislation differs very much in various countries. In the Federal Republic of Germany the TA Luft (Clean Air Regulations) was revised in 1986. The emission of benzene, which is classified as a carcinogen, must not exceed 5 mg/m^3 in waste gas from benzene oxidation [26]. This necessitates installation of a combustion unit downstream of the benzene adsorption stage, as in C_4 oxidation plants [27].

2.3. Quality, Storage, and Transportation

Quality Specifications. A typical specification for pure MA includes the following:

mp	52.5 °C (min.)
Acid number	1140 mg KOH/g (min.)
Purity	99.7 wt % (min.)
Melt color index	20 APHA (max.)
Ash content	10 ppm (max.)
Iron content	5 ppm (max.)

Storage of Pure MA. Pure MA in the molten state at 55–65 °C is stored in stainless steel or pure aluminum vessels. It can be stored for several weeks in an inert-gas atmosphere without any change in quality.

Commercial Forms and Transportation.

Liquid MA is transported over large distances at 60–80 °C in well-insulated road tankers or tank containers. The vessels are provided with heating devices.

Solid MA is dispatched in the form of pastilles and briquettes weighing 0.5–20 g. Because of their high dust content, flakes are rarely produced. Multiply paper bags with welded polyethylene liners or bags of strong polyethylene film are normally used as packaging. Solid MA can be stored for several months in this packaging in a cool, dry place.

Maleic anhydride is included in the list of dangerous substances and preparations according to the EEC Directive 67/548/EEC and GefStoffV and must be specially packed and identified. However, on the basis of the most recent toxicological investigations, the identification according to GefStoffV [28] should be amended and amplified as follows [29]:

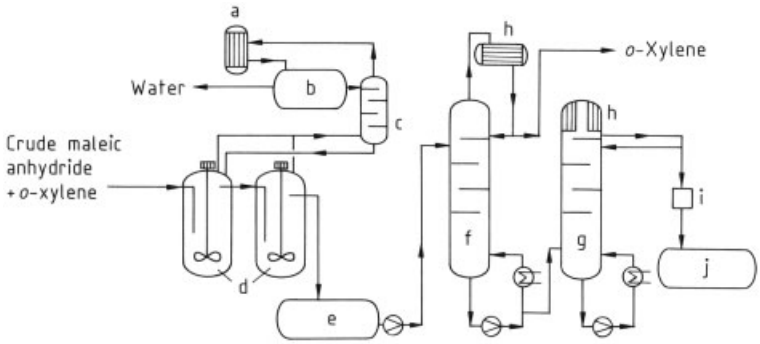

Figure 2. Continuously operating plant for production of pure maleic anhydride
a) Condenser; b) Separating vessel; c) Column; d) Postdehydration; e) Intermediate tank; f) *o*-Xylene column; g) Pure MA column; h) Cooler; i) Colorimeter; j) Pure MA vessel

Danger symbol	C (corrosive)
R	22 – 34 – 42/43
S	22 – 28 water – 39

The hazard classification for the transport of MA are as follows:

IMDG Code (GGVSee)	Class 8
GGVS/GGVE, ADR/RID	Class 8, no. 31 c
UN no.	2215

2.4. Uses

Because of its properties as a dicarboxylic anhydride with a double bond, MA can be used for both polycondensation and polyaddition. Polyester and alkyd resins, lacquers, plasticizers, copolymers, and lubricants are the most important technical end products. Examples of industrially important copolymers are: MA – styrene (for engineering plastics) and MA – acrylic acid (used in the detergent industry). Polyester and alkyd resins, in particular, are used in the production of fiberglass reinforced plastics, in the construction and electrical industries, and in pipeline and marine construction.

Tetrahydrophthalic anhydride is formed by a Diels – Alder reaction with butadiene; hydrogenation of the anhydride yields hexahydrophthalic anhydride. Methylhexahydrophthalic anhydride is obtained in a similar manner from isoprene and MA. These products are used as anhydride curing agents for epoxy resins.

Smaller amounts of MA are used in the production of pesticides (captan, malathion) and growth inhibitors (maleic acid hydrazide). Surfactants are produced from maleic acid esters by reaction with sodium hydrogen sulfite. When added to drying oils, MA reduces the drying time and improves the coating quality of lacquers.

3. Citraconic and Mesaconic Acids

In 1881 Anschütz obtained mesaconic [*498-24-8*], 2-methyl-*trans*-2-butenedioic acid and citraconic acid [*498-23-7*] (2-methyl-*cis*-2-butenedioic acid) for the first time by distilling citric acid. Citraconic acid forms an inner anhydride (citraconic anhydride [*616-02-4*], 3-methyl-2,5-furandione, *mp* 7 °C, *bp* (4 kPa) 109 °C, $d_4^{15.7}$ 1.2469), analogously to maleic acid. Citraconic anhydride is interesting particularly because of its low melting point (easier handling at lower temperature).

Many attempts have been made to develop industrial processes for obtaining *citraconic anhydride*, including its preparation from citric acid (1905), and gas-phase oxidation of alkyl-substituted olefins [30], methylbutanols [31], or mesityl oxide [32]. In the gas-phase oxidation of *o* -xylene, ca. 2 – 4 t of citraconic anhydride is formed per 1000 t of phthalic anhydride. Citraconic anhydride is present as citraconic acid after water washing. Three processes have been developed to isolate citraconic anhydride [33–35]. Its synthesis from succinic anhydride and formaldehyde has been proposed [36].

With regard to its industrial use, citraconic anhydride is frequently mentioned with MA in patents. Substitution of citraconic acid for MA, however, does not impart any special properties to the products. Citraconic anhydride is useful simply in producing liquid curing agents for epoxy resins: Diels – Alder synthesis with isoprene yields a nonhygroscopic curing agent that is liquid at room temperature [37].

4. Fumaric Acid

Fumaric acid [*110-17-8*], *trans*-butenedioic acid, M_r 116.07, occurs widely in nature. It received its name from WINCKLER, who isolated it from fumitory (*Fumaria officinalis*) in 1832. Subsequently, it was isolated many times under the names boletic acid, glaucinium acid, and lichen acid, and in 1834 it was described by PELOUZE who called it paramaleic acid. Fumaric acid is currently produced on an industrial scale, although it is less important than maleic anhydride.

Physical Properties. Fumaric acid crystallizes in the form of monoclinic prisms. Important physical properties are summarized below:

mp	286 – 287 °C (closed tube)
d^{25} (solid)	1.635
Heat of sublimation (92 °C)	136.1 kJ/mol
Dissociation constants in water	
	K_1 9.57×10^{-4}
	K_2 4.13×10^{-5}
Solubility in 100 g of water	
at 15.5 °C	0.428 g
at 100 °C	9.97 g

Chemical Properties. Fumaric acid sublimes without decomposition at 200 °C or slightly higher. Above 230 °C, however, it decomposes with the formation of maleic anhydride, water, and a considerable amount of residue. Addition of phosphorus pentoxide promotes the formation of maleic anhydride.

Its other chemical properties resemble those of maleic acid. Reactions proceed considerably more slowly, however, due partly to the lower water solubility of fumaric acid. In contrast to maleic acid, fumaric acid forms an acid chloride.

Production (see Fig. 3). Maleic acid or maleic anhydride, especially the maleic acid-containing wash water from the production of maleic anhydride or phthalic anhydride, serves as starting material for the manufacture of fumaric acid. The maleic acid concentration should be at least 30 %. Maleic acid is converted almost quantitatively by thermal [38] or catalytic isomerization into the sparingly soluble fumaric acid, which is recovered by filtration. Various substances have been proposed as catalysts: mineral acids (e.g., hydrochloric acid) [39]; sulfur compounds such as thiocyanates, thiazoles, thiosemicarbazides, thioureas; or bromine compounds in combination with peroxides (e.g., persulfate) [40]. Thiourea is most commonly used in practice [41].

The maleic acid-containing wash water contains impurities that can affect quality and yield. This problem can be largely avoided (1) by thermal pretreatment of the wash water [42], (2) by adding urea [43] if thiourea is used as catalyst, and (3) by addition of sulfites or passaged of sulfur dioxide [44] and addition of mineral acids [45]. The crude fumaric acid obtained is purified by recrystallization from water, combined with purification by active charcoal. Losses during purification are ca. 10 %.

Fumaric acid production is generally carried out batchwise, although continuous processes are also known. If hydrogen chloride does not have to be added, an apparatus made of corrosion-resistant austenitic steel is satisfactory. Disposal of the mother liquor from the isomerization stage, which in addition to soluble impurities also contains the isomerization catalyst, is especially difficult. One possibility is to combust the effluent.

In the United States, fumaric acid is also produced by fermentation of monosaccharides or starch with selected fungal cultures (*Mucor, Aspergillus, Rhizopus*) [46].

Uses. Fumaric acid is often used instead of maleic acid to produce polyesters and copolymers. A nearly identical mixture of *trans*- and *cis*-dicarboxylic acids is formed by isomerization during the preparation of these products [47], regardless of whether maleic or fumaric acid is used as starting material. Fumaric acid is often preferred because of its innocuous nature.

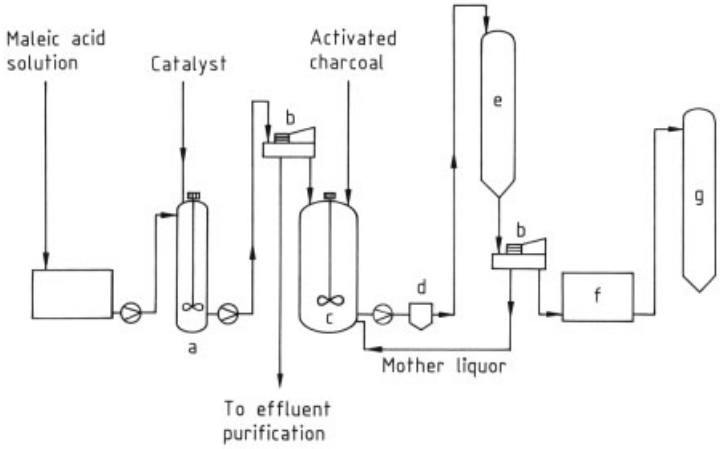

Figure 3. Plant for production of fumaric acid from aqueous maleic acid solutions
a) Isomerization vessel; b) Centrifuge; c) Dissolving tank; d) Filter; e) Crystallizer; f) Dryer; g) Silo

On the other hand, special effects are observed when fumaric acid is used, e.g., greater hardness of certain resins [48].

Because it is nontoxic, fumaric acid is used in the food industry, e.g., as an acidulant in baking powders and beverages, for which purpose particularly stringent requirements are placed on purity (U.S. Food Chemicals Codex). Fumaric acid (Fumaril) and iron(II) fumarate are used as special additives for animal feeds [49].

5. Toxicology

Maleic Anhydride. In aqueous solution, maleic anhydride slowly hydrolyzes to maleic acid; accordingly, the toxicological properties of the anhydride are in part determined by the toxicological activity profile of maleic acid. The *acute toxicity* is characterized by LD_{50} values of 481 mg (rat, oral), 465 mg (mice, oral), 2620 mg (rabbit, percutaneous), and 390 mg (guinea pig, oral) per kilogram of body weight [50]. The main acute toxic effects of this compound are its powerful local irritant and corrosive action on the skin, mucous membrane, and eyes. At a concentration of 1.5 – 2 ppm, nasal irritation was observed in workers after 1 min, and irritation of the eyes after 15 – 20 min [51]. Concentrations > 2.5 ppm are extremely irritant. Asthmatic symptoms may occur after long-term exposure to maleic anhydride concentrations > 1.2 ppm [52]. In addition to its acute irritant action, maleic anhydride is also characterized by a strongly sensitizing action, which has been observed experimentally in guinea pigs as well as in humans [52], [54].

No *systemic* toxic effects were observed in a 6-month inhalation exposure of rats, hamsters, and monkeys to up to 2.4 ppm of maleic anhydride [55]. Animal experimental studies did not reveal any evidence of a carcinogenic, teratogenic, or reproductive – toxicological action of maleic anhydride [55–57].

The MAK and TLV – TWA values were based on the acute irritant and sensitizing action of maleic anhydride and are 0.2 and 0.25 ppm, respectively [52], [53].

Maleic Acid. As an acid, the toxicological activity of maleic acid is characterized by marked irritant and corrosive action on the skin, mucous membrane, and eyes. A 20 % aqueous solution produced a mild and reversible skin irritation in humans [58]. Lower concentrations (< 5 %) are sufficient to produce serious ocular irritation [59]. Animal experiments confirm that the local irritant action is the most prominent toxicological feature of maleic acid. The *acute* oral and dermal toxicity is characterized by LD_{50} values of 708 mg (rat, oral), 2400 mg (mouse, oral), 1560 mg (rabbit, dermal), and > 1000 mg (guinea-pig, dermal) per kilogram of body weight [60].

Kidney damage has been observed as a *systemic* toxic action of maleic acid. Morpholog-

ical – functional changes in the kidneys were produced by intraperitoneal injection (100 – 350 mg/kg; rats, dogs) and also by 1-h inhalation exposure to ca. 720 mg/m^3 (rats) [61]. The damage to distal and proximal tubular resorption produced by maleic acid is reminiscent of the known (congenital or acquired) Fanconi syndrome in humans [62], [63]. This damage seems to be caused by the interaction of maleic acid with glutathione in tubular renal cells, leading to intolerable concentrations of free radicals and peroxides [64]. In addition to the resultant cellular damage, maleic acid appears to affect Na^+ and H^+ ion transport in the proximal tubes [65]. After chronic exposure of male rats (250, 500, 750 mg kg^{-1} d^{-1} for up to two years), an increased mortality as well as kidney damage and delayed growth was observed in all dosage groups [66]. Liver and testicular damage has also been found in the highest dosage group. No signs of *carcinogenic activity* of maleic acid were discovered with this long-term dosage, which however was not designed as a carcinogenicity study. No *genotoxic effects* were observed in the Ames test, with or without metabolizing fractions [67].

Fumaric Acid. Fumaric acid is a naturally occurring intermediate of the tricarboxylic acid cycle and is used as a food additive [68]. Accordingly the compound is found to be practically nontoxic in acute toxicity tests. An oral LD_{50} value of 10 g/kg of body weight was measured in rats, and in rabbits 20 g/kg of body weight applied dermally was not lethal [69]. In humans, no changes in the blood and urine parameters or in liver function were found after administration of 8 mg of fumaric acid per kilogram of body weight per day for one year [70]. Experimental investigations did not reveal any evidence of genotoxic [67], [71], carcinogenic [70], [72], or reproductive – toxicological [70] activity. The toxicological evaluation of fumaric acid is thus based primarily on the local irritant action on contact with the skin, mucous membrane, and eyes. In this respect, fumaric acid is classified as slightly irritant to the skin and moderately irritant to the eyes and mucous membranes.

6. References

1. L. H. Flett, W. H. Gardner: *Maleic Anhydride Derivatives,* J. Wiley & Sons, New York 1952.
2. B. C. Trivedi, B. M. Culbertson: *Maleic Anhydride,* Plenum Press, New York 1982.
3. R. Klar, DE-AS 1 069 150, 1957.
4. Chevron Res. Co., DE-OS 2 658 191, 1976 (K. R. Bakshi, D. M. Marquis, S. G. Paradis).
5. United Eng. & Constr., US 2 762 449, 1953 (M. P. Sweeney).
6. Halcon, DE-AS 1 468 586, 1962 (G. F. Vinci, M. Gans).
7. C. E. Vogler et al., *J. Chem. Eng. Data* **8** (1963) no. 4, 620 – 623.
8. Reichhold Chem. Inc., DE 833 039, 1950 (C. H. B. Jarl).
9. UCB, DE-OS 2 003 684, 1970 (E. Weyens).
10. UCB S.A., *Hydrocarbon Processing* **64** (1985) no. II, 143.
11. Princeton Chem. Res., US 3 293 268, 1966 (R. I. Bergman, N. W. Frisch).
12. Chevron Res. Co., US 3 864 280, 1975 (R. A. Schneider).
13. Monsanto Co., US 4 456 764, 1982 (J. T. Wrobleski).
14. H. Bosch et al., *Appl. Catal.* **31** (1987) 323 – 337.
15. G. Centi, J. R. Ebner et al., *Chem. Rev.* **88** (1988) 55 –80.
16. Ftalital S.p.A., US 4 314 946, 1982 (A. Neri, S. Sanchioni).
17. F. Budi, A. Neri, G. Stefani, *Hydrocarbon Processing* **61** (1982) no. 1, 159 – 161.
18. S. C. Arnold, L. Verde et al., *Hydrocarbon Processing* **64** (1985) no. 9, 123 – 126.
19. *Hydrocarbon Processing* **64** (1985) no. 11, 142.
20. G. Emig et al., *Catal. Today* **1** (1987) 477 – 498.
21. *Eur. Chem. News* **50** (1988) Oct. 31, 8.
22. Du Pont, EP-A 0 189 261 A1, 1986 (R. M. Contractor).
23. *Eur. Chem. News* **50** (1988) Dec. 12, 36.
24. VEBA-Chemie AG u. Metallges. AG, DE 2 061 336, 1970 (O. L. Garkisch, G. Ibing, G. Kammholz, H. Schirrmacher, K. Lohbeck).
25. K. Lohbeck, *Erdöl Kohle Erdgas Petrochem. Suppl.* **1** (1976) 320 – 327.
26. Technische Anleitung zur Reinhaltung der Luft, 27. 02. 1986, Ziff. 2.3 Krebserzeugende Stoffe.
27. Ziff. 3.1.7 Organische Stoffe/Anhang E. [26]
28. Verordnung über gefährliche Stoffe (GefStoffV), 26. 08. 1986, Anhang VI.
29. Hüls AG, Tox. Bericht Nr. 1390, 1989, unpublished result.

30. H. Pichler, DE-OS 2 206 713, 1972.
31. Pfizer Inc., DE-OS 2 522 568, 1974 (R. G. Berg, B. E. Tate).
32. Pfizer Inc., DE-OS 2 526 464, 1974 (R. G. Berg, D. S. Hetzel).
33. Produits Chimiques, FR 1 489 787, 1966 (M. P. Juston).
34. VEBA-Chemie, DE-AS 1 593 546, 1966 (G. Ibing, H. Haferkorn).
35. BP Chemicals Int., GB 1 398 586, 1972 (L. E. Cooper).
36. Pfizer Inc., DE-OS 2 352 468, 1972 (R. G. Berg, B. E. Tate).
37. VEBA-Chemie, DE 1 593 548, 1966 (G. Ibing, K. Neubold, H. Haferkorn).
38. California Res. Corp., US 2 816 922/3, 1954 (R. W. Stephenson).
39. Allied Chem. Corp., US 2 393 352, 1942 (L. Winstrom).
40. Halcon, US 3 262 972, 1963 (R. S. Barker).
41. United States Rubber Comp., US 2 454 385/7, 1946 (L. H. Howland, W. F. Brucksch).
42. Deutsche Texaco, DE 1 768 629, 1968 (P. Ackermann, H. Heinrichs).
43. Allied Chem. & Dye Corp., US 2 483 576, 1947 (G. E. Neuman de Végvár).
44. Montclair Res. Corp. + Ellis Foster Comp., US 2 843 628/9, 1953 (J. B. Rust).
45. Halcon, US 3 389 173, 1966 (R. L. Russel, H. Olenberg).
46. Chas. Pfizer & Co., Inc. + Merck & Co. Inc., US 2 327 191, 1939 (J. H. Kane, A. Finlay, P. F. Amann).
47. S. S. Feuer et al., *Ind. Eng. Chem.* **46** (1954) 1643.
48. R. C. Peter, *Ind. Chem.* **31** (1955) 564.
49. H. Brune et al.: "Fumarsäure in der Tierernährung," *Landwirtschaftl. Schriftenreihe Boden Pflanze Tier,* Heft 8, Parey, Hamburg 1979.
50. RTECS (Jan. 1988).
51. ACGIH, TLVs, 4th ed. and Supplement (1980).
52. D. Henschler, *MAK-Werte, Toxikologisch-arbeitsmedizinische Begründungen,* Verlag Chemie, Weinheim 1977.
53. ACGIH *Treshold Limit Values and Biological Exposure Indices for 1989 – 1990,* Cincinnatti 1989.
54. Hüls AG, unpublished results, Marl (1989).
55. R. D. Short et al., *Fundam. Appl. Toxicol.* **10** (1988) 517.
56. R. D. Short et al., *Fundam. Appl. Toxicol.* **7** (1986) 359.
57. Chemical Industry Institute of Toxicology (1983), cited in R. D. Short et al., *Fundam. Appl. Toxicol.* **10** (1988)517.
58. M. B. Britz, H. I. Maibach, *Contact Dermatitis* **5** (1979) 375.
59. L. Parmeggiani (ed.): *Encyclopedia of Occupational Health and Safety, 3rd ed.,* International Labour Organisation, Geneve 1983.
60. RTECS (Jan. 1988).
61. R. R. Verani et al., *Lab. Invest.* **46** (1982) 79.
62. E. D. Brewer et al., *Am. J. Physiol.* **245** (1983) F 339.
63. E. C. Foulkes, *Toxicol. Appl. Pharmacol.* **71** (1983) 445.
64. G. Gstraunthaler et al., *Biochem. Pharmacol.* **32** (1983) 2969.
65. N. A. Rebouchas et al., *Pflugers Arch.* **401** (1984) 266.
66. O. G. Fitzhugh, A. A. Nelson, *J. Am. Pharm. Assoc.* **36** (1947) 217.
67. W. H. Rapson et al., *Bull. Environ. Contam. Toxicol.* **24** (1980) 590.
68. Zusatzstoffverkehrs-VO, 7, Anlage 2, Liste 10, (20. Dezember 1977).
69. E. H. Vernot et al., *Toxicol. Appl. Pharmacol.* **42** (1977) 417.
70. S. Levey et al., *J. Am. Pharm. Assoc.* **35** (1946) 298.
71. M. Ishidate Jr. et al., *Food Chem. Toxicol.* **22** (1984) 623.
72. V. Saffioti, P. Shubik, *Natl. Cancer Inst. Monogr.* (1983) no. 10, 489.

Maleic Anhydride → Maleic and Fumaric Acids
Malic Acid → Hydrocarboxylic Acids, Aliphatic

Malonic Acid and Derivatives

Harald Strittmatter, Lonza Ltd., Visp, Switzerland
Stefan Hildbrand, Lonza Ltd., Basel, Switzerland
Peter Pollak, Lonza Ltd., Basel, Switzerland

1. **Introduction** 1
2. **Malonic Acid** 1
3. **Malonates** 3
4. **Cyanoacetic Acid and Cyanoacetates** . . 7
5. **Malononitrile** 10
6. **Quality Specifications and Analysis** . . . 13
7. **Economic Aspects** 13
8. **Toxicology and Occupational Health** . . 15
9. **References** 15

1. Introduction

Three-carbon 1,3-dicarboxylic acid derivatives (malonic acid, malonates, cyanoacetic acid, cyanoacetates, and malononitrile) are widely used in industry for the manufacture of pharmaceuticals, agrochemicals, vitamins, dyes, adhesives, and fragrances. The common feature of malonic acid and its derivatives is the high reactivity of the central methylene group. Due to the increasingly electron-withdrawing character of the substituents, the acidity of the hydrogen atoms in the 2-position increases in the order malonates < cyanoacetates < malononitrile. Therefore, all these compounds undergo reactions typical of 1,3-dicarbonyl compounds. For example they are easily alkylated or arylated, undergo aldol and Knoevenagel condensations, and they can be used for the synthesis of pyrimidines and other nitrogen heterocycles.

Malonic acid and derivatives are basically all produced by reaction of a two-carbon building block with a one-carbon unit (Fig. 1). Thus, cyanoacetates are prepared from sodium chloroacetate and sodium cyanide, followed by acidification and subsequent esterification. Malonates are produced either by a cobalt-catalyzed alkoxycarbonylation of chloroacetates with carbon monoxide in the presence of alcohol, or by hydrolysis of cyanoacetic acid followed by esterification. For the production of malononitrile, the continuous high-temperature reaction of cyanogen chloride and acetonitrile has almost totally replaced the former synthesis, based on dehydration of cyanoacetamide (**1**).

2. Malonic Acid

Physical Properties. Important physical properties of malonic acid (propanedioic acid, methanedicarboxylic acid) are listed in Table 1. Its pK_a values are 2.83 and 5.70. Malonic acid forms a colorless hygroscopic solid which sublimes in vacuum with some decomposition. Its solubility is 139 g in 100 g of water (at 22 °C) and 15 g in 100 g of pyridine (at 15 °C). Malonic acid is only slightly soluble in ethanol and diethyl ether, and is insoluble in benzene.

Chemical Properties. Malonic acid is found in small amounts in sugar beet and green wheat, being formed by oxidative degradation of malic acid. Reaction with sulfuryl chloride or bromine gives mono- and dihalogenated malonic acid, whereas treatment with thionyl chloride or phosphorus pentachloride leads to mono- or diacyl chloride. When heated with phosphorus pentoxide, malonic acid does not form an anhydride but rather carbon suboxide, a toxic gas that reacts violently with water to reform malonic acid. On heating the free acid above 130 °C, or an aqueous solution above 70 °C, decomposition to acetic acid and carbon dioxide takes place. The mono- and dianion of malonic acid are more stable. In aqueous solution the monosodium salt decomposes above 90 °C and the disodium

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_063.pub2

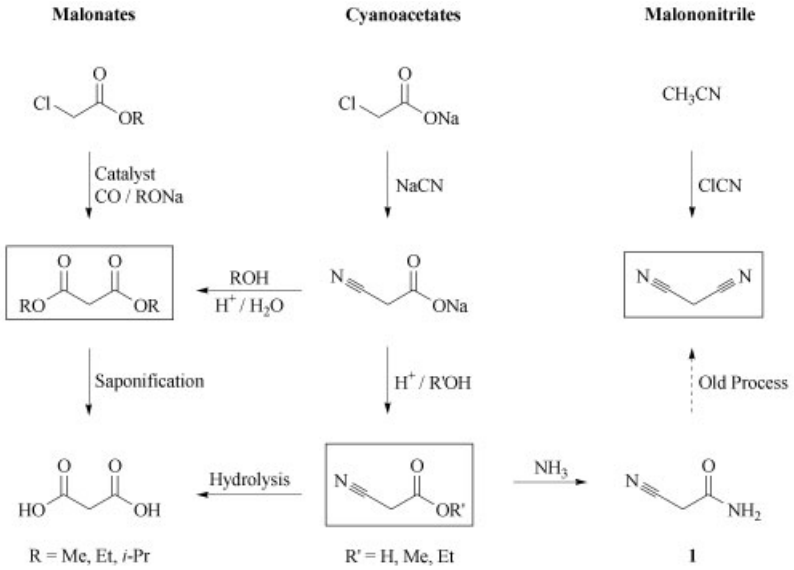

Figure 1. Synthetic routes for malonates, cyanoacetates, and malononitrile

Table 1. Properties of malonic acid, malonates, and derivatives

Name	CAS registry no.	Molecular formula	Molecular mass	*bp*, °C	Density (20 °C), g/mol	Refractive index (n_D^{20})	Flash point, °C*
Malonic acid	[*141-82-2*]	$C_3H_4O_4$	104.06	134 – 138**			157
Meldrum's acid	[*2033-24-1*]	$C_6H_8O_4$	144.13	92 – 95**			
Dimethyl malonate	[*108-59-8*]	$C_5H_8O_4$	132.12	181	1.151 – 1.156	1.413 – 1.414	88
Diethyl malonate	[*105-53-3*]	$C_7H_{12}O_4$	160.17	199	1.052 – 1.057	1.413 – 1.414	90
Diisopropyl malonate	[*13195-64-7*]	$C_9H_{16}O_4$	188.23	201 – 203	0.991	1.412	95
Diethyl methylmalonate	[*609-08-5*]	$C_8H_{14}O_4$	174.20	198.5	1.020 – 1.025	1.412 – 1.414	85

* Pensky–Martens, closed cup DIN 51 758.
** *mp* (decomp.).

salt above 130 °C [1]. These mild decarboxylation conditions combined with its ability to undergo Knoevenagel condensations makes malonic acid a useful reagent for the synthesis of α,β-unsaturated carboxylic acids. Thus cinnamic acids are formed by reaction of malonic acid with benzaldehyde derivatives, and acrylic acids from malonic acid and aliphatic aldehydes [2]. Finally, platinum complexes of malonic acid and its derivatives have been investigated as antitumor agents [3, 4].

Production. The industrial production of malonic acid is less important than that of the malonates. Malonic acid is usually produced by hydrolysis of cyanoacetic acid [5, 6] or by acid saponification of malonates [7–9].

Alternative more recently reported methods for the preparation of malonic acid are the ozonolysis of cyclopentadiene [*542-92-7*] [10], the palladium-catalyzed air oxidation of 1,3-propanediol [11], the platinum group metal-catalyzed oxidation of 3-hydroxypropionaldehyde or 3-hydroxypropionic acid [12] and the use of malononitrile as a substrate for a nitrilase [13].

Uses. Malonic acid is used for the introduction of an acetic acid moiety under mild conditions by Knoevenagel condensation and subsequent decarboxylation. Examples include the preparation of cinnamic acid, used for the production of the anti-inflammatory cinmetacin [*20168-99-4*] (**2**) [14, 15] and 3,4,5-trimeth-

oxycinnamic acid, the key intermediate of the vasodilators cinepazet [*23887-41-4*] (**3**, R = OEt) [16] and cinepazide [*23887-46-9*] (**3**, R = 1-pyrrolidinyl) [17]. The fragrance γ-nonanoic lactone [*104-61-0*] (**4**) is available by Knoevenagel condensation of malonic acid with heptaldehyde [*111-71-7*] followed by ring closure [18].

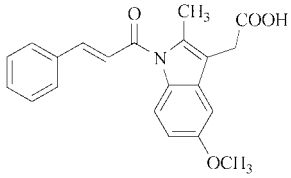

2

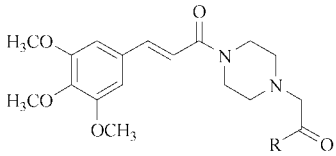

3

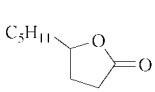

4

Derivatives. *Meldrum's acid* (2,2-dimethyl-1,3-dioxane-4,6-dione) (**5**) is a colorless crystalline solid which is produced from malonic acid and acetone in the presence of acetic anhydride and traces of sulfuric acid [19–22]. Some physical properties are listed in Table 1. Its chemistry was reviewed in several publications [23, 24]. Meldrum's acid is frequently used for the preparation of malonic acid monoesters **6** [25–27] and β-keto esters **7** by acylation and subsequent alcoholysis [28].

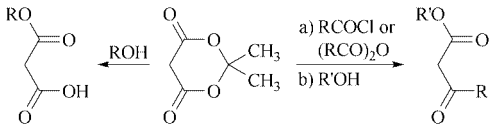

6 **5** **7**

3. Malonates

Physical Properties. The most important malonates are the dimethyl and the diethyl ester. Both are colorless liquids that are sparingly soluble in water (1 g/50 mL for diethyl malonate), but miscible in all portions with diethyl ether and ethanol. At room temperature, malonates ($pK_a \approx 13$) and alcoholic solutions thereof exist almost exclusively in the keto form.

Of the malonates derived from higher alcohols, only the diisopropylester is of industrial use. The physical properties of the most important malonates are shown in Table 1.

Chemical Properties [29]. Mono- and dialkylation of malonates with alkyl halides [30–33] or sulfonates [34, 35] are usually performed in the presence of alkoxides or with sodium hydride. Subsequent hydrolysis of one ester group and decarboxylation allows the introduction of the CH(H)–CO_2R (R = Me, Et) moiety into an organic molecule (*Malonic ester synthesis*). Similarly, acylation of malonates with acid chlorides [36] yields β-keto esters after hydrolysis and decarboxylation [37–39]. Reaction with aldehydes or ketones (Knoevenagel condensation) affords ylidene-malonates [40]. Reduction of dialkyl malonates with LiAl(O-*tert*-Bu)$_3$H proceeds with selective reduction of one of the ester groups to provide β-hydroxy esters in good yields [41]. Michael addition of malonates to mesityl oxide [*141-79-7*] gives 5,5-dimethyl-1,3-cyclohexanedione-4-carboxylic acid after ring closure. Subsequent decarboxylation leads to dimedone [*126-81-8*] (**8**) [42, 43]. Reaction of diethyl malonate with triethyl orthoformate affords diethyl (ethoxymethylene)malonate [*87-13-8*] (**9**) [44]. Mesoxalic acid diesters **10** can be obtained by metal-catalyzed air oxidation of malonates [45]. 2-(Acylamino)-malonates **11** are available by nitrosation of malonates followed by simultaneous acylation and hydrogenation [46–49]. Alkylation of **11**, followed by hydrolysis, and subsequent decarboxylation affords α-amino acids [50–52]. Dialkyl malonates can be used for the selective synthesis of monoalkyl malonates [53–55] and for the preparation of mixed malonates by transesterification [56, 57]. A two-step process for the preparation of Meldrum's acid (**5**) from diethyl malonate has been reported [58].

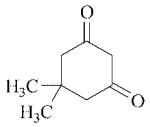

8

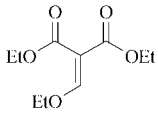

9

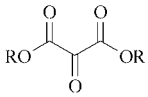

10

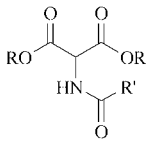

11

Finally, malonates and their 2-substituted derivatives can be converted into nitrogen heterocycles. Pyrazolones are obtained by reaction of malonates with hydrazines [59]. By condensation with urea or thiourea, 2,4,6-trihydroxypyrimidines or 2-thio-4,6-dihydroxypyrimidines are formed, and with guanidine 2-amino-4,6-dihydroxypyrimidines are obtained [60]. Similarly, 4,6-dihydroxypyrimidines are available from malonates and formamide [61–63] and 2-alkoxy-substituted-4,6-dihydroxypyrimidines by condensation of malonates with *O*-alkylisoureas [64, 65].

Production. Two processes have become established for the industrial production of malonates [66].

Hydrogen Cyanide Process. This process is based on hydrogen cyanide and chloroacetic acid; the intermediate cyanoacetic acid is simultaneously saponified and esterified in the presence of a large excess of mineral acid and alcohol (Figure 1).

A solution of sodium cyanide (ca. 25 %) in water is heated to 65 – 70 °C in a stainless steel reaction vessel. An aqueous solution of sodium chloroacetate is then added slowly with stirring; the temperature must not exceed 90 °C. Stirring is maintained at this temperature for an hour. Particular care must be taken to ensure that the hydrogen cyanide, which is formed continuously in small amounts, is trapped and neutralized. The solution of sodium cyanoacetate is concentrated by evaporation under vacuum and then transferred to a glass-lined reaction vessel for hydrolysis of the cyano group and esterification.

The alcohol and mineral acid (weight ratio 1:2 to 1:3) are introduced in such a manner that the temperature does not rise above 60 – 80 °C. For each mol of ester, ca. 1.2 mol of alcohol is added.

Hydrochloric acid, which is formed as byproduct from unreacted chloroacetic acid, is fed into an absorption column. After the addition of acid and alcohol is complete, the mixture is heated at reflux for 6 – 8 h, whereby the intermediate malonic acid ester monoamide is hydrolyzed to a dialkyl malonate. The pure ester is obtained from the mixture of crude esters by extraction with benzene, toluene, or xylene. The organic phase is washed with dilute sodium hydroxide to remove small amounts of the monoester. The diester is then separated from solvent by distillation at atmospheric pressure, and the malonic ester obtained by distillation under vacuum as a colorless liquid with a minimum assay of 99 % (gas chromatography). The aqueous phase contains considerable amounts of mineral acid and salts and must be treated before being fed to the waste treatment plant. The process is suitable for both the dimethyl and diethyl esters.

The yield based on sodium chloroacetate is 75 – 85 %. Various low molecular mass hydrocarbons, some of them partially chlorinated, are formed as byproducts. Although a relatively simple plant is sufficient for the reaction itself, a sizeable investment is required for treatment of the wastewater and exhaust gas.

Carbon Monoxide Process. According to this process, dimethyl, diethyl, and diisopropyl malonates are produced by a dicobalt octacarbonyl-catalyzed reaction of chloroacetates (methyl, ethyl, or isopropyl) with carbon monoxide in the presence of the appropriate alcohol [67–70]. In the case of diethyl malonate the reaction is conducted at around 100 °C and 18 bar at pH 5.7 (Fig. 2) [71]. Ethyl acetate is formed as a major byproduct. After completion of the reaction, sodium chloride and catalyst are separated. The alcohol and the low-boiling components are distilled off and the nonconverted chloroacetate recovered by distillation. The crude ethyl malonate obtained is purified by redistillation.

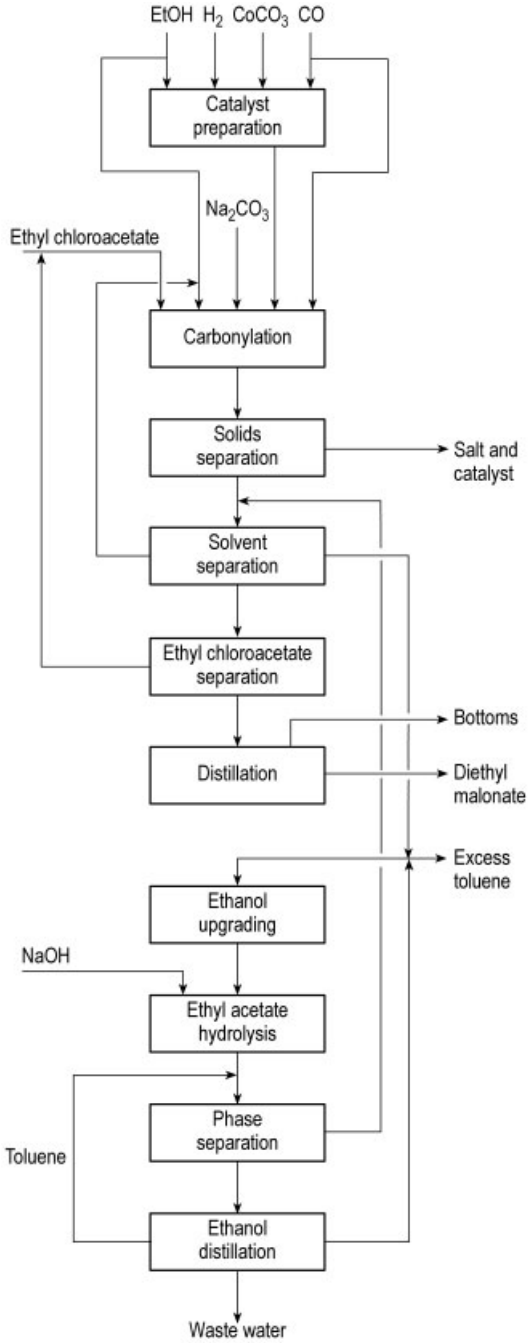

Figure 2. Flow-sheet for diethyl malonate production by the carbonylation process

Alternative methods for the production of malonates, which have not been used industrially are the palladium-catalyzed reaction of ketene [*463-51-4*] with carbon monoxide in the presence of an alkyl nitrite [72], the ozonization of diketene [73], and the alkoxycarbonylation of dichloromethane [74, 75].

Uses. Dimethyl and diethyl malonates are widely used for the production of vitamins, pharmaceuticals, agrochemicals, fragrances, and dyes (Fig. 3). The use of barbituric acid [*67-52-7*] for the synthesis of vitamin B_2 (riboflavin) [76] is not important any more, because the prevailing process is now based on fermentation. Whereas barbituric acid itself has no physiological effects, barbiturates (5,5-disubstituted barbituric acids) and thiobarbiturates are widely used as hypnotics (→ Hypnotics), antiepileptics (→ Antiepileptics, Chap. 2.5), or anesthetics (→ Anesthetics, General, Chap. 2.1).

The vasodilator naftidrofuryl [*31329-57-4*] is prepared by alkylation of diethyl malonate first with 2-tetrahydrofurfuryl chloride [*3003-84-7*] and then with 1-(chloromethyl)naphthalene [*86-52-2*] followed by hydrolysis, decarboxylation and introduction of the diethylaminoethyl moiety [77]. 2-Substituted malonates are used for the synthesis of the antiepileptic vigabatrin [*60643-86-9*], derived from diethyl malonate and 1,4-dichloro-2-butene [78], and the anti-inflammatory drugs phenylbutazone [*50-33-9*], obtained from ethyl 2-(*n*-butyl)malonate [*133-08-4*] and 1,2-diphenyl-hydrazine [*122-66-7*] [79], and carprofen [*53716-49-7*], synthesized by Michael addition of dimethyl methylmalonate [*609-02-9*] to 2-cyclohexene-1-one [*930-68-7*] followed by hydrolysis, subsequent decarboxylation and formation of the carbazole unit [80]. Reaction of (ethoxy-methylene)malonate (**9**) with arylamino derivatives gives access to (aminomethylene)malonates, which are further upgraded into a wide range of quinolone antibiotics such as nalidixic acid [*389-08-2*] [81, 82], oxolinic acid [*14698-29-4*] [83], ofloxacin [*82419-36-1*] (*X* = O, R = CH_3) [84], or rufloxacin [*101363-10-4*] (X = S, R = H) [85]. Finally, 2-(acetylamino)malonate **11** (R′ = CH_3) is used for the synthesis of the anti-inflammatory and antiulcer drug rebamipide [*90098-04-7*] [86].

In the agrochemical field, malonates are mainly used for the production of two classes of herbicides: the cyclohexanedione-type and the sulfonylurea herbicides. The synthesis of the

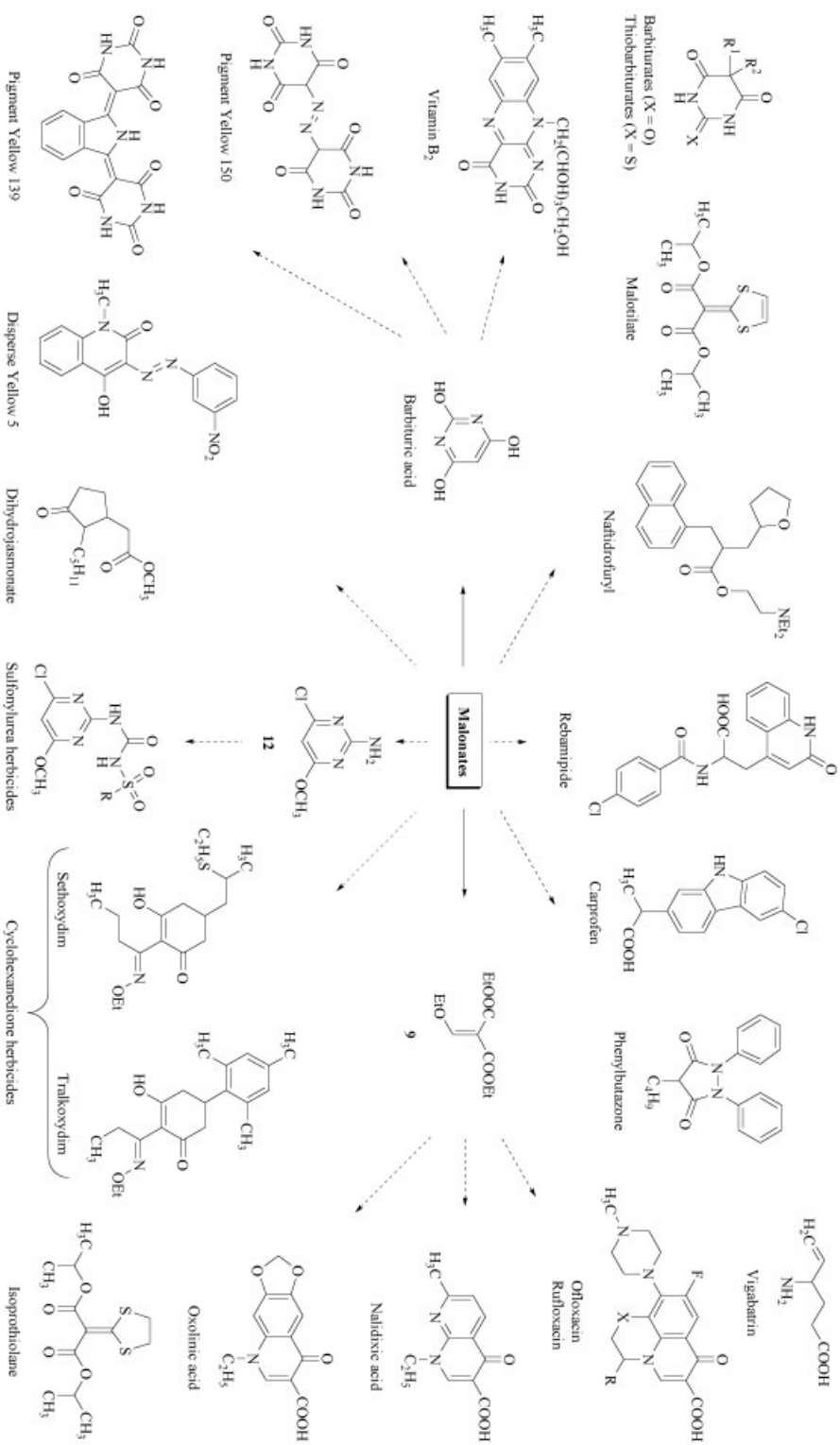

Figure 3. Important products derived from malonates

1,3-cyclohexanedione herbicides commences with a Michael addition of malonates to 1-substituted but-1-en-3-ones followed by ring closure and subsequent decarboxylation. Examples of such herbicides are sethoxydim [*74051-80-2*] [87] and tralkoxydim [*87820-88-0*] [88]. The other prominent class are the sulfonylurea herbicides, derived from 2-amino-4-chloro-6-methoxypyrimidine [*5734-64-5*] (**12**), for example chlorimuron-ethyl [*90982-32-4*] [89]. 2-Amino-4-chloro-6-methoxypyrimidine is prepared by condensation of malonates with guanidine followed by chlorination and displacement of one chlorine atom with methoxide [90–92].

Diisopropyl malonate is used for the synthesis of the fungicide isoprothiolane [*50512-35-1*] by reaction with carbon disulfide and ethylene dichloride [*107-06-2*] [93] and for the hepatic protectant malotilate [*59937-28-9*] [94, 95].

Malonates and barbituric acid are also used for the synthesis of dyes such as C. I. Pigment Yellow 139 [*36888-99-0*], C. I. Pigment Yellow 150 [*25157-64-6*], and C. I. Disperse Yellow 5 [*6439-53-8*]. Dihydrojasmonate [*24851-98-7*], which is synthesized by Michael addition of a malonate to 2-pentyl-2-cyclopenten-1-one [*25564-22-1*] [96], is used in perfumery for blossom fragrances, in particular jasmine.

4. Cyanoacetic Acid and Cyanoacetates

Physical Properties. The physical properties of cyano-acetic acid and the most important cyanoacetates are listed in Table 2. Cyanoacetic acid forms colorless, hygroscopic crystals (*mp* 69 – 70 °C), that undergo decarboxylation above 160 °C to give acetonitrile. The compound has a dissociation constant of 3.65×10^{-3} in water (25 °C) and dissolves readily in water, but is only moderately soluble in ethanol, and sparingly in benzene. The industrially most important cyanoacetate is ethyl cyanoacetate, followed by the methyl ester. Both esters are colorless liquids, which are miscible with ethanol and diethyl ether but insoluble in water.

Chemical Properties. The chemical properties of cyanoacetates are quite similar to those of the malonates [97].

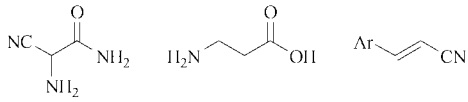

13 **14** **15**

Reaction with ammonia gives cyanoacetamide [*107-91-5*] (**1**) [98, 99]. Similarly, reaction of cyanoacetates with urea or its *N*-alkylated derivatives affords cyanocetylureas, which form pyrimidines upon ring closure [100, 101]. 2-Aminocyanoacetamide [*6719-21-7*] (**13**) is prepared by nitrosation of cyanoacetamide [102] or alkyl cyanoacetates [103, 104], followed by catalytic hydrogenation. Hydrogenation of cyanoacetic acid affords β-alanine (**14**) [105]. Cyanoacetates are also used as chain extenders. Thus, Knoevenagel condensation with aromatic aldehydes affords 3-substituted acrylonitriles after decarboxylation [106, 107], such as 3-thiophen-2-yl-acrylonitrile from thiophene-2-carbaldehyde [*98-03-3*] (**15**, Ar = thiophenyl) [108], 3-furan-2-yl-acrylonitrile from furfural [*98-01-1*] (**15**, Ar = furanyl) [109], or cinnamonitriles from benzaldehyde derivatives (**15**, Ar = phenyl) [110, 111]. With nonaromatic carbonyl compounds a shift of the double bond can occur to give β,γ-unsaturated nitriles [112, 113]. Finally, cyanoacetates can be converted into malonic acid monoesters or malonic acid monoamide monoesters [114] using microorganisms [115, 116].

Production. Cyanoacetic acid and the cyanoacetates are produced by the same route (and often in the same plant) as the malonates by cyanidation of sodium chloroacetate followed by acidification of the sodium cyanoacetate solution. Cyanoacetic acid is isolated by extraction with an organic solvent and evaporation of the solvent, whereas the cyanoacetates are obtained by mineral acid-catalyzed esterification with subsequent azeotropic removal of water [117–119]. Major byproducts are malonates and succinates. A process variation involves interchange of the cyanidation and esterification steps [120–122]. In this case, however, the formation of the side product dimethyl [*6283-71-2*] or diethyl [*10359-15-6*] 2-cyanosuccinate is difficult to suppress. In order to avoid the isolation, most of the cyanoacetic acid outlets are produced directly from solution.

Table 2. Properties of cyanoacetates and derivatives

Name	CAS registry no.	Molecular formula	Molecular mass	*bp*, °C	Density (20 °C), g/mol	Refractive index $(n)_D^{20}$	Flash point, °C*
Cyanoacetic acid	[*372-09-8*]	$C_3H_3NO_2$	85.06	108/0.02 bar			160
Methyl cyanoacetate	[*105-34-0*]	$C_4H_5NO_2$	99.09	203	1.125 – 1.130	1.418 – 1.419	110
Ethyl cyanoacetate	[*105-56-6*]	$C_5H_7NO_2$	113.12	208	1.058 – 1.063	1.417 – 1.419	109
n-Propyl cyanoacetate	[*14447-15-5*]	$C_6H_9NO_2$	127.14	220	1.023 – 1.028	1.4222	108
Isopropyl cyanoacetate	[*13361-30-3*]	$C_6H_9NO_2$	127.14	202	1.012 – 1.017	1.4171	91
Allyl cyanoacetate	[*13361-32-5*]	$C_6H_7NO_2$	125.13	89 – 90/0.005 bar	1.070 – 1.075	1.4222	118
n-Butyl cyanoacetate	[*5459-58-5*]	$C_7H_{11}NO_2$	141.17	109/0.013 bar	0.998 – 1.002	1.426 – 1.427	91
2-Ethylhexyl cyanoacetate	[*13361-34-7*]	$C_{11}H_{19}NO_2$	197.28	135/0.01 bar			
2-Methoxyethyl cyanoacetate	[*10258-54-5*]	$C_6H_9NO_3$	143.14	257/100 kPa	1.123 – 1.128	1.434 – 1.435	> 110
2-Ethoxyethyl cyanoacetate	[*32804-77-6*]	$C_7H_{11}NO_3$	157.17	141/0.02 bar	1.080 – 1.083	1.434 – 1.435	110

* Pensky–Martens, closed cup DIN 51 758.

The world's largest cyanoacetate plant was brought on stream in 1987 in Mobile (Alabama) by Hüls. The annual production capacity of this plant was later expanded to 8000 t.

Alternative methods for the preparation of cyanoacetic acid involve the carboxylation of acetonitrile [123–125] and the oxidation of partially oxidized propionitriles [126–132]. Other methods for the preparation of cyanoacetates include the cleavage of cyanoacetaldehyde dimethyl acetal [133], formed by a palladium-catalyzed oxidation of acrylonitrile [134–137], the alkoxycarbonylation of chloroacetonitrile [138, 139] and pyrolysis of 3,3′-bisisoxazole in the presence of an alcohol [140]. However, none of these processes have been used industrially.

Higher Cyanoacetates. Large quantities of cyanoacetates of higher alcohols are usually obtained by esterification of cyanoacetic acid as described above [141], whereas smaller amounts are prepared by trans-esterification of methyl or ethyl cyanoacetate in the presence of aluminum [142] (e.g., aluminum isopropoxide), titanium [143] or tin catalysts [144, 145].

Uses. The most important application of cyanoacetates is in the field of adhesives. Knoevenagel reaction of alkyl cyanoacetates with formaldehyde affords α-cyanoacrylates **16**, which are stable under exclusion of air but polymerize rapidly to polycyanoacrylates **17** after application on a surface even at room temperature. This polymerization is reversible on heating and the monomer can be formed back and subsequently polymerized again by traces of moisture. Esters of higher alcohols (see Table 2), such as 1-butanol, are also employed.

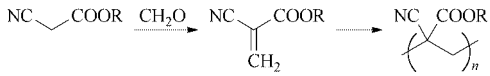

16 **17**

Otherwise, cyanoacetates are used for the synthesis of pharmaceuticals, agrochemicals, dyes, and UV absorbers (Fig. 4). For many applications cyanoacetic acid, cyanoacetates, or cyanoacetamide (**1**) can be used alternatively. Caffeine (R = CH_3) [*58-08-2*] is synthesized from theophylline (R = H) [*58-55-9*], which is accessible in several steps from cyanoacetic acid and 1,3-dimethylurea [*96-31-1*] [146, 147]. The traded cyanoacetic acid is mainly used for the production of the fungicide cymoxanil [*57966-95-7*] [148, 149] and the cough remedy dextromethorphan [*125-71-3*] [150].

Ring closure of cyanoacetylurea [*1448-98-2*] affords 6-aminouracil [*873-83-6*] which is further upgraded into lumazine [*487-21-8*] (**18**) and the antihypertensive amiloride [*2609-46-3*] [151, 152] or into the antiinfective sulfadimethoxine [*122-11-2*] [153–155]. Reaction of cyanoacetates with guanidine affords 4-hydroxy-2,6-diaminopyrimidine [*56-06-4*] (**19**), which can be transformed into guanine [*73-40-5*] (**20**, R = OH) and 2-amino-6-chloropurine [*10310-21-1*] (**20**, R = Cl) [156–159]. Antivirals such as acyclovir [*59277-89-3*] (**21**, R^1 = OH, R^2 = $CH_2OC_2H_4OH$) [160, 161], ganciclovir [*82410-32-0*] (**21**, R^1

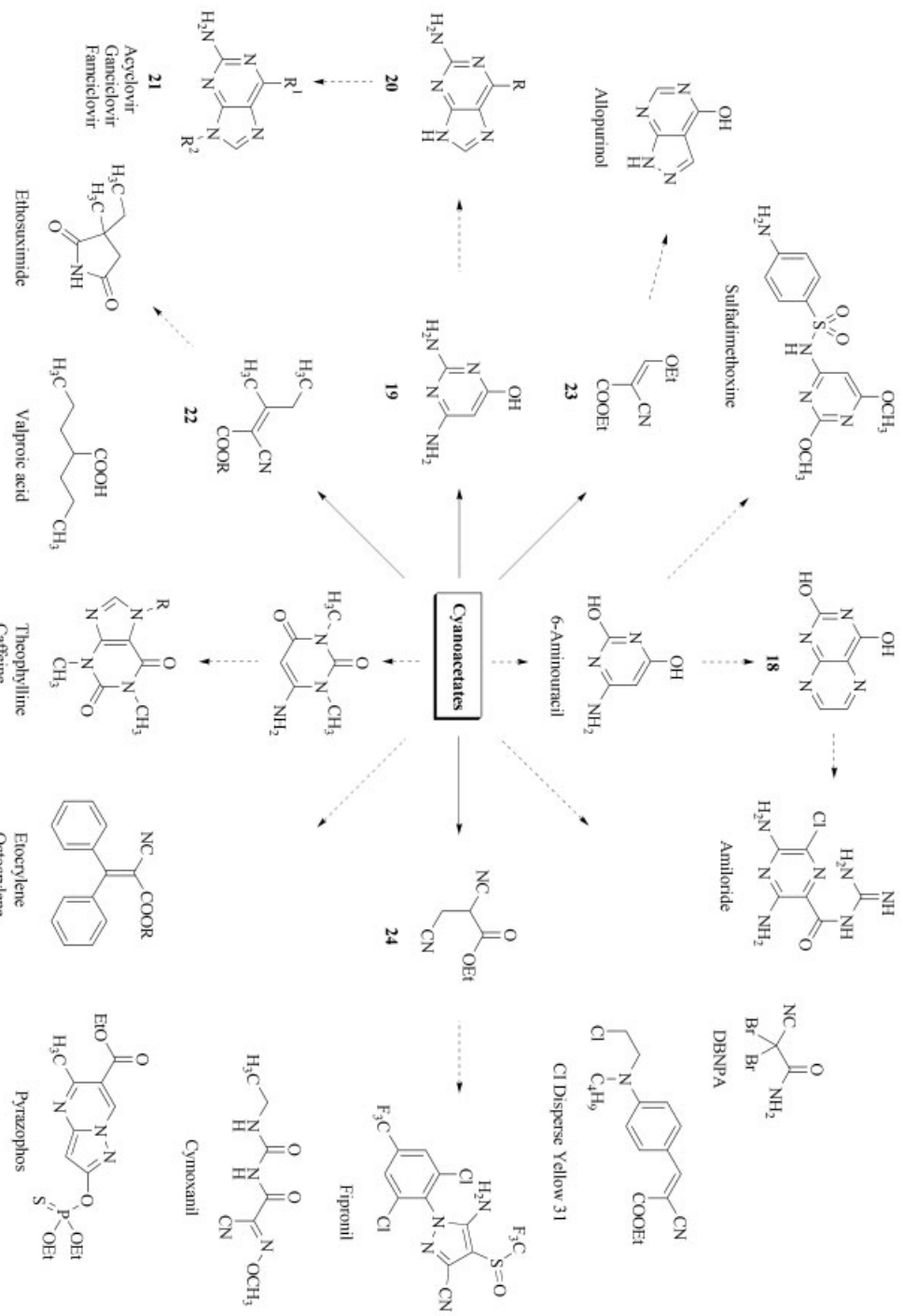

Figure 4. Important products derived from cyanoacetates

= OH, $R^2 = CH_2OCH(CH_2OH)_2$) [162], or famciclovir [*104227-87-4*] (**21**, R^1 = H, R^2 = $CH_2CH_2CH(CH_2OAc)_2$), [163] are significant outlets thereof. The synthesis of the antiepileptic ethosuximide [*77-67-8*] commences with a Knoevenagel condensation of a cyanoacetate with 2-butanone to give **22**. Cyanide addition, followed by hydrolysis, subsequent decarboxylation, and imide formation completes the synthesis [164, 165]. Other examples include the antiepileptic valproic acid [*99-66-1*] [166] and the uricosuric allopurinol [*315-30-0*], prepared from ethyl (2-ethoxymethylene)cyanoacetate [*94-05-3*] (**23**) and hydrazine [167]. Compound **23** is also a possible starting material for the preparation of the rice herbicide pyrazosulfuron-ethyl [*93697-74-6*] [168] and the fungicide fludioxonil [*131341-86-1*] [169]. Pyrazophos [*13457-18-6*] is another fungicide manufactured from cyanoacetate [170]. Reaction of ethyl cyanoacetate with sodium cyanide and paraformaldehyde affords ethyl 2,3-dicyanopropionate [*40497-11-8*] (**24**), which is further upgraded into the insecticide fipronil [*120068-37-3*] [171, 172]. The same building block might also be used for the preparation of the pesticide vaniliprole [*145767-97-1*] [173].

Reaction of cyanoacetamide (**1**) with bromine affords the biocide 2,2-dibromo-2-cyano-acetamide (DBNPA) [*10222-01-2*] [174]. Cyanoacetates are also used for the synthesis of dyes such as C.I. Disperse Yellow 31 (Celliton Fast Yellow 7G) [*4361-84-6*] [175]. Finally, the UV absorbers etocrylene (R = ethyl) [*5232-99-5*] and octocrylene (R = 2-ethylhexyl) [*6197-30-4*] are prepared by condensation of ethyl or 2-ethylhexyl cyanoacetate with benzophenone [*119-61-9*] [176] or benzophenone imine [*1013-88-3*] [177].

5. Malononitrile

Physical Properties. The physical properties of malononitrile are listed in Table 3. Malononitrile forms colorless crystals that are soluble in water, lower alcohols, diethyl ether, and acetonitrile, but insoluble in tetrachloromethane, petroleum ether, and xylene. After prolonged storage the crystals turn yellow to dark brown without significant degradation of quality.

Table 3. Physical properties of malononitrile

CAS registry no.	[*109-77-3*]
Other names	propanedinitrile
	malonic acid dinitrile
	dicyanomethane
Molecular formula	$C_3H_2N_2$
Molecular mass	66.06
Melting point, °C	31.6
Boiling point, °C	218 – 219
	105 – 107/0.01 bar
Density (35 °C) g/mL	1.0494
Refractive index (34 °C)	1.4146
Dipole moment (25 °C), D	3.57
pK_a	11.2
Solubility (20 °C), g/L:	
water	133
diethyl ether	200
ethanol	400
Flash point* , °C	> 86

* Abel-Pensky, closed cup DIN 51 755.

Chemical Properties. The chemical properties of malononitrile are determined by the nucleophilicity of the malononitrile anion, formed by deprotonation with relatively week bases and by the two electrophilic cyano groups [178–182]. Base-catalyzed dimerization of malononitrile affords 2-amino-1,1,3-tricyano-2-propene [*868-54-2*](**25**) [183, 184]. Reaction of malononitrile with trialkyl orthoformates gives (alkoxymethylene)-malononitriles (**26**, R^1 = OR, R^2 = H) which react with nucleophiles by displacement of the alcohol group. This reaction can be used for the synthesis of styryl dyes [185]. Dialkylation of malononitrile with alkyl halides [186] or acylation with acid chlorides and anhydrides is possible [187].

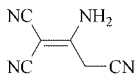

25

26

27

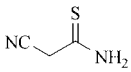

28

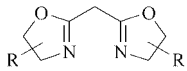

29

Knoevenagel condensation of malononitrile with aldehydes and ketones gives ylidene-malononitriles, which can undergo further reactions [188, 189]. Examples are the reaction with elemental sulfur (Gewald reaction) to give 3-cyano-2-amino-thiophenes [190, 191] or cy-

cloaddition reactions with azides to afford tetrazoles [192].

Bromomalononitrile [*1885-22-9*] (**27**, R^1 = Br, R^2 = H) is obtained in moderate yields by reaction of malononitrile with one equivalent of bromine [193]. With two equivalents of bromine dibromomalononitrile [*1885-23-0*] (**27**, $R^1 = R^2$ = Br) is obtained, which is usually isolated as a complex with potassium bromide. Thermal decomposition of this complex affords tetracyanoethylene [*670-54-2*] (**26**, $R^1 = R^2$ = CN) [194–198]. Reaction of dibromomalononitrile with potassium cyanide yields the potassium salt of tricyanomethane (**27**, R^1 = CN, R^2 = H) [199]. Alternatively, tricyanomethane is available from malononitrile and cyanogen chloride [200]. Tricyanomethane can be used for the synthesis of ionic liquids [201].

Reaction of malononitrile with carbon disulfide in the presence of base affords the disodium [*4885-93-2*] or dipotassium salt [*4777-48-4*] of (dimercaptomethylene) malononitrile (**26**, $R^1 = R^2 = S^-$) [202], which is used for the synthesis of antimicrobials [203, 204]. Dimethylation affords [bis(methylthio)methylene]malononitrile [*5147-80-8*] (**26**, $R^1 = R^2 = SCH_3$), a versatile building block for the preparation of heterocycles [205–208]. Cyanothioacetamide [*7357-70-2*] (**28**) is best prepared by reaction of malononitrile with hydrogen sulfide [209–211].

Among the most important building blocks derived from malononitrile are the imidates, which are prepared by acid-catalyzed reaction of one or both nitrile groups with alcohols [212, 213]. Diimidates are convenient starting materials for the synthesis of nitrogen heterocycles, such as 4,6-dimethoxypyrimidines or bis(oxazolines) **29** [214]. For example, 2-amino-4,6-dimethoxypyrimidine [*36315-01-2*] (**30**) is produced from dimethyl malonoimidate dihydrochloride [*71160-05-9*] (**31**) by ring closure with cyanamide [*420-04-2*] (see Fig. 7) [215]. The bis(oxazolines) **29** can alternatively be prepared directly from malononitrile by a Ritter-type reaction (→ Amines, Aliphatic) [216] or by using a catalyst [217]. Bis(oxazolines) derived from malononitrile (or derivatives thereof) and chiral aminoalcohols are efficient ligands in asymmetric transition metal catalyzed reactions [218].

2-Aminomalononitrile [*5181-05-5*] (**27**, $R^1 = NH_2$, R^2 = H) is obtained either by nitrosation of malononitrile followed by catalytic hydrogenation or by reduction of (phenylazo)malononitrile [*6017-21-6*] (**32**) (see Fig. 7) [219–221]. Reaction of aminomalononitrile with guanidine yields 2,4,5,6-tetraaminopyrimidine [*1004-74-6*] (**33**) [222]. 2,4,6-Triaminopyrimidine [*1004-38-2*] (**34**) is available by condensation of malononitrile with guanidine [223, 224], and 4-amino-5-cyano-pyrimidine [*16357-69-0*] by reaction of malononitrile with formamidine [225].

Production. Until the mid-1970s malononitrile was obtained by dehydration of cyanoacetamide [*107-91-5*] (**1**), with phosphorus pentachloride [226] or phosphorus oxychloride [227, 228] in a batch process (see Fig. 1). Nowadays, only Chinese producers use this process for the production of small amounts. But there are still some new patents being issued concerning this reaction [229–233]. By far the major part of malononitrile is manufactured by a continuous high-temperature process in which acetonitrile and cyanogen chloride are fed into a tube reactor above 700 °C (Fig. 5) [234, 235]. Almost complete conversion is attained in one pass. The reaction products are malononitrile, excess acetonitrile, hydrochloric acid, and traces of maleic, succinic, and fumaric acids. The products leaving the reactor are immediately quenched to 40 – 80 °C, and gaseous hydrogen chloride is simultaneously separated, fed into a scrubber, and recovered as dilute hydrochloric acid. Excess acetonitrile is removed by a combination of vacuum distillation and thin-film evaporation. The recovered acetonitrile contains very little hydrogen chloride and can be recycled.

Removal of maleic and fumaric acid from the crude malononitrile by fractional distillation is impractical because the boiling points of these acids and malononitrile differ only slightly. Therefore, the impurities are converted into high-boiling compounds in a conventional reactor by means of a Diels – Alder reaction. The byproducts are finally removed by two vacuum distillations. The yield of malononitrile is 66 % based on both cyanogen chloride and acetonitrile.

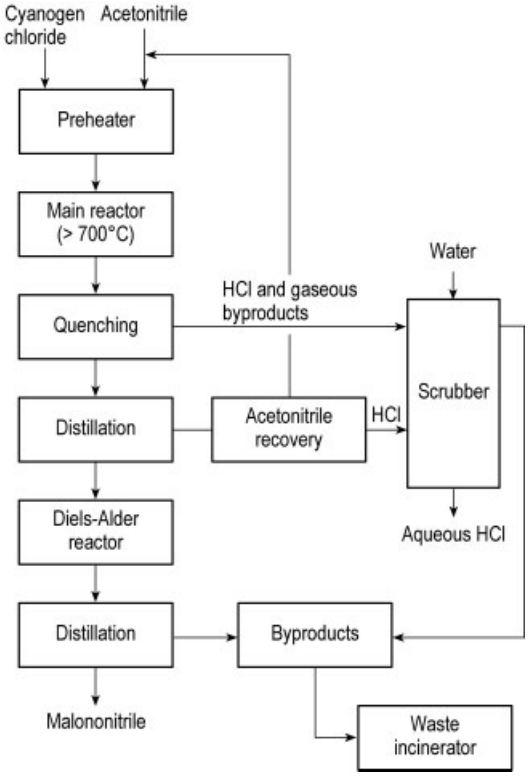

Figure 5. Process flow-sheet for the production of malononitrile

Lonza in Visp (Switzerland) brought the first industrial plant using this process on stream in 1973 (Fig. 6).

Alternative ways reported to prepare malononitrile include its direct formation from cyanoacetates [236] or malonates [237] by reaction with ammonia in the presence of a dehydrating catalyst; the molybdenum-catalyzed reaction of 3-hydroxypropionitrile [238] with ammonia and oxygen, vapor phase oxidation of 3-aminopropionitrile in the presence of ammonia [239], vapor phase reaction of allene with ammonia in the presence of oxygen [240], a high temperature reaction of dicyanogen with acetonitrile [241], and the thermal reaction of isonitriles with nitriles [242]. However none of these routes are used on an industrial scale.

Transportation. Malononitrile is shipped as solidified melt packed into 200-L lined drums or as a solution in methanol. Care should be exercised when remelting the solidified product, because spontaneous decomposition can occur when it is exposed for an extended time at elevated temperatures, particularly above 100 °C. A water bath is recommended for this purpose.

Uses. Malononitrile is extensively used as a building block for the synthesis of vitamins, agrochemicals, pharmaceuticals, and dyes (Fig. 7). The two most important outlets for malononitrile are thiamine (vitamin B_1) [*59-43-8*] [243, 244] and the sulfonylurea herbicides based on 2-amino-4,6-dimethoxypyrimidine [*36315-01-2*] (**30**), such as benzsulfuron-methyl [*83055-99-6*] [245], azimsulfuron [*120162-55-2*] [246], pyrazosulfuron-ethyl [*93697-74-6*] [168], or halosulfuron-methyl [*100784-20-1*] [247, 248]. Other outlets from malononitrile are the diuretic triamterene [*396-01-0*] [249, 250], the folic acid antagonist methotrexate [*59-05-2*] [251, 252], both prepared from 2,4,6-triaminopyrimidine [*1004-38-2*] (**34**), adenine [*73-24-5*] [253–256], the antihypertensive minoxidil [*38304-91-5*] [257, 258], and the pesticide dicyclanil [259, 260].

Furthermore, malononitrile is used for the synthesis of several yellow methine dyes with a *N*,*N*-dialkylaminoaryl group (aminoarylneutrocyanines) such as C. I. Disperse Yellow 90 [*6684-20-4*] (**35**, $R = CH_3$, $R^1 = C_2H_5$, $R^2 = C_2H_4OC(O)C_6H_5$) [261]. Blue aminoarylneutrocyanine dyes such as C. I. Disperse Blue 354 [*74239-96-6*] (**36**, $R=CH_3$, $R^1 = R^2 = C_6H_{13}$) are available by Knoevenagel condensation of 3-(dicyanomethylene)-2,3-dihydrobenzo[b]thiophene-1,1-dioxide [*74228-25-4*] with substituted benzaldehydes [262]. 1,4,5,8-Naphthalenetetracarboxylic acid [*128-97-2*] (**37**), used for the synthesis of perinone pigments, can be made by acylation of acenaphthene [*83-32-9*] with malononitrile followed by oxidation [263].

Charge-transfer salts of the malononitrile derivative 7,7,8,8-tetracyanoquinodimethane (TCNQ) [*1518-16-7*] are powerful electrical conductors [264]. TCNQ is made by condensation of malononitrile with 1,4-cyclohexanedione [*637-88-7*] followed by dehydrogenation [265]. Finally, condensation of *o*-chlorobenzaldehyde with malononitrile affords the tear gas (*o*-chlorobenzylidene)malononitrile [*2698-41-1*] (CS-gas) [266, 267].

Figure 6. Malononitrile plant of Lonza in Visp (Switzerland)

6. Quality Specifications and Analysis

Standard commercial quality of *malonic acid* has a minimum purity of 99 %. The analytical method usually applied for determination of purity is potentiometric titration with sodium hydroxide. Both equivalent points are measured, and the content is determined using the following equation:

$$\text{Content [\%]} = \frac{(E_2 - E_1) \times M \times 10.406}{W}$$

where E_1, E_2 are the volumes of standard sodium hydroxide solution at the first and second equivalence point, M is the molarity of the standard NaOH solution, and W the weight of the sample (in grams).

Malonate of standard commercial quality has a minimum purity of 99 % (gas chromatography).

Cyanoacetates are analyzed by gas chromatography, whereas potentiometric titration is used for cyanacetic acid. Standard commercial quality has a minimum assay of 99 %.

Malononitrile has a minimum purity of 99 wt % (measured by gas chromatography) and a solidification point of at least 30 °C.

7. Economic Aspects

The worldwide annual production capacity is believed to be more than 20 000 t for malonates, and more than 15 000 t for cyanoacetates. The production capacity has increased substantially over the past years, especially in the Eastern Hemisphere. A considerable overcapacity exists which has led to a price erosion. The most important products are dimethyl and diethyl malonate, methyl and ethyl cyanoacetate, cyanoacetic acid, and malononitrile. The ethyl esters are more important products than the methyl esters in both series. A large part of the cyanoacetic acid is used captively. The single most important producer for both cyanoacetates and malonates is Degussa–Hüls with plants in the USA and Europe (Table 4). The largest producer of malononitrile is Lonza in Switzerland. Worldwide production of malonic acid does not exceed a few hundred tons per year.

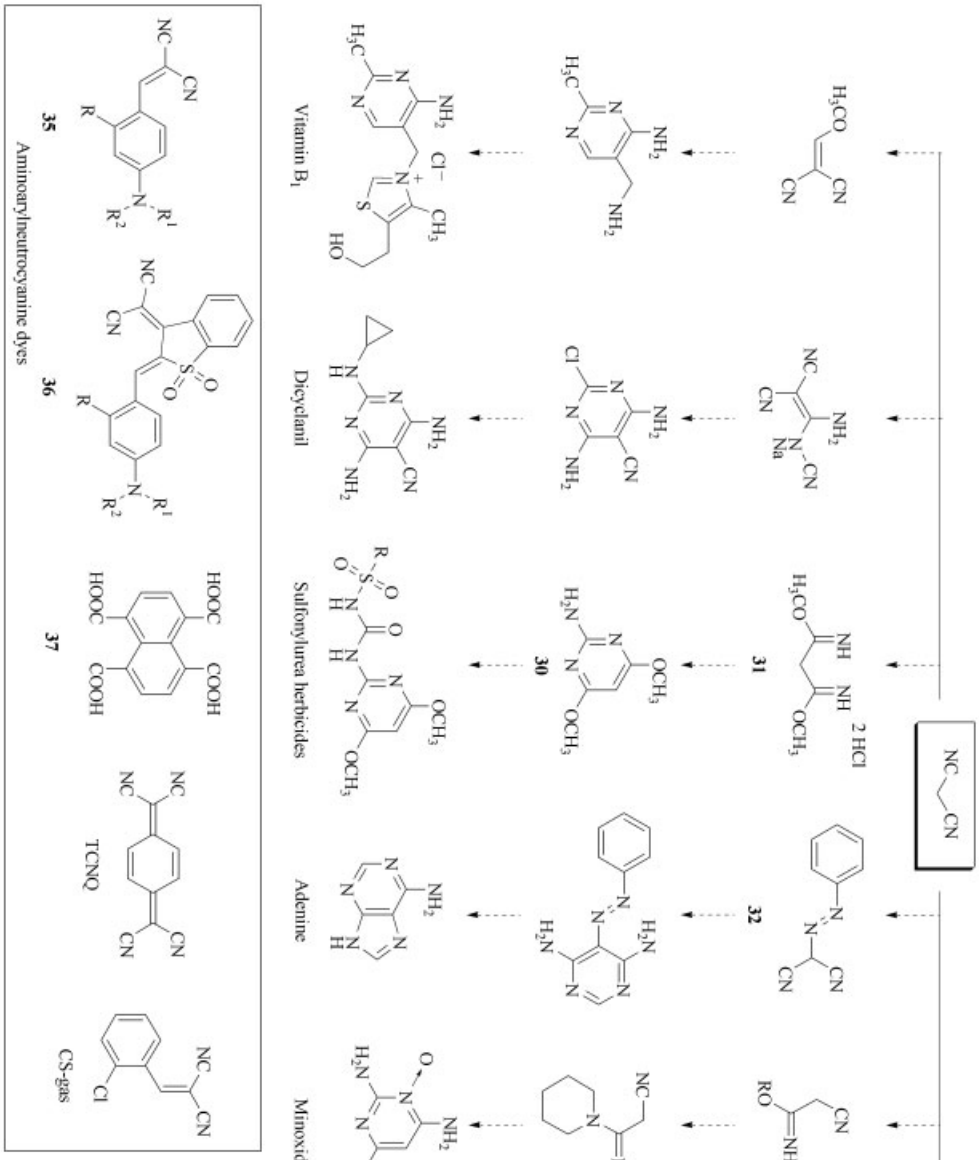

Figure 7. Reactions of malononitrile

Table 4. Estimated annual production capacity of cyanoacetates and malonates (t)

Country	Producer	Cyanoacetates	Malonates
United States	Degussa–Hüls	7000*	
Europe	Degussa–Hüls		7000*
Japan	Juzen, Tateyama	3000	4000
China		5000*	12 000
Korea	Samsung		2000
India		1000*	600

* Partly used for internal demand.

8. Toxicology and Occupational Health

Toxicology. Toxicological data of malonic acid and its derivatives are listed in Table 5.

Table 5. Toxicological data of malonic acid and important derivatives

Name	Skin irritation*	LD_{50}**, mg/kg
Malonic acid	mild	2750
Meldrum's acid	mild	> 5000
Dimethyl malonate	non-irritant	4520
Diethyl malonate	non-irritant	> 5000
Diisopropyl malonate	mild	> 5000
Diethyl methylmalonate	mild	> 5000
Cyanoacetic acid	mild	1500
Methyl cyanoacetate	mild	3062
Ethyl cyanoacetate	mild	2820
n-Propyl cyanoacetate	mild	> 5000
Isopropyl cyanoacetate	non-irritant	> 2000
Allyl cyanoacetate	non-irritant	160
n-Butyl cyanoacetate	mild	> 5000
2-Methoxyethyl cyanoacetate	non-irritant	4300
2-Ethoxyethyl cyanoacetate	mild	1250 – 5000
Malononitrile	mild irritant	14

* Rabbit,
** Oral, rat.

Occupational Health. Beyond normal safe handling measures no special precautions are necessary for the handling of *malonic acid*.

Dimethyl and diethyl malonate do not present any specific health hazards if handled with the usual precautions. Nevertheless, inhalation and skin contact should be avoided.

Because solid *cyanoacetic acid* and its solutions attack the skin, the acid should be handled carefully.

Because of the low LD_{50} (Table 5) and its chemical reactivity particular care should taken in handling *malononitrile*.

9. References

1. E. A. Halonen, *Ann. Acad. Sci. Fenn.* **55** (1954) (Ser. A2), 1.
2. G. L. Bolton in L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis*, vol. **5**, John Wiley & Sons, New York 1995, pp. 3213 – 3214.
3. Ferrer Internacional SA, ES 2004446, 1989 (R. F. Ambros et al.).
4. Korea Institute of Science and Technology, US 5426203, 1995 (S. Y. Sohn et al.).
5. N. Weiner, *Org. Synth., Coll. Vol. II*, (1943) 376.
6. Trikem. S. A., WO 9908988, 1999 (O. De Carvalho et al.).
7. Dow Chemical Co., US 2373011, 1945 (E. C. Britton, E. Monroe).
8. Huels AG, DE 4120704, 1992 (F. Englaender, M. Feld).
9. Huels AG, EP 0857712, 1998 (K.-D. Steffen).
10. CS 249237, 1988 (M. Matas, K. Fancovic).
11. Henkel KGaA, DE 4107986, 1991 (A. Behr et al.).
12. Degussa AG, DE 19629372, 1998 (T. Haas et al.).
13. Mitsubishi Rayon Co. Ltd., Nitto Chemical Industry Co. Ltd., EP 0187680, 1986 (K. Enomoto et al.).
14. Sumitomo Chemical Co., Ltd., ZA 6702683, 1968 (H. Yamamoto, M. Nakao).
15. H. Yamamoto, M. Nakao, *J. Med. Chem.* **12** (1969) no. 1, 176.
16. Delalande SA, GB 1168108, 1969 (C. P. Fauran et al.).
17. Delalande SA, BE 730345, 1969 (C. P. Fauran et al.).
18. H. Pyyasolo et al., *Finn. Chem. Lett.* **5** (1975) 129.
19. D. Davidson, S. A. Bernhardt, *J. Am. Chem. Soc.* **70** (1948) 3426.
20. Dow Chemical Co., US 4613671, 1986 (A. G. Relenyi, D. E. Wallick, J. D. Streit).
21. Leningrad State University, SU 1237666, 1986 (N. N. Khimich, V. A. Nikolaev).

22. Degussa–Huels AG, DE 19962346, 2001 (M. Bongen et al.)
23. H. McNab, *Chem. Soc. Rev.* **7** (1978) no. 3, 345.
24. B. C. Chen, *Heterocycles* **32** (1991) no. 3, 529.
25. H. Junek, E. Ziegler, U. Herzog, H. Kroboth, *Synthesis* **5** (1976) 332.
26. S. L. Schreiber et al., *J. Am. Chem. Soc.* **108** (1986) 2106.
27. D. Craig et al., *Org. Lett.* **7** (2005) 463.
28. Y. Oikawa, T. Yoshioka, K. Sugano, O. Yonemitsu, *Org. Synth.* **63** (1985) 198.
29. G. Romeder in L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis*, vol. **3**, John Wiley & Sons, New York 1995, pp. 1822 – 1825.
30. Hüls AG, DE 19752041, 1999 (J. Metz, C. Osterholt).
31. Wako Pure Chemicals Industries, Ltd., JP 62289574, 1987 (M. Tanaka et al.).
32. Lonza AG, EP 0378073, 1990 (R. Fuchs, J. McGarrity).
33. S. Nose et al., US 20050124825.
34. Nippon Zeon Co., JP 08245508, 1996 (K. Ogu, Y. Kondo).
35. B. Ashfeld et al., *Org. Lett.* **7** (2005) 4535.
36. Bayer AG, DE 4138616, 1993 (M. Preiss).
37. H. Henecka, *Houben-Weyl* **8** (1952) 610 – 616.
38. Bayer AG, DE 3142854, 1983 (K. Grohe, J. H. Zeiter, K. Metzger).
39. R. J. Clay et al., *Synthesis* **3** (1993) 290.
40. Hüls AG, DE 4237897, 1994 (M. Bongen et al.).
41. T. A. Ayers, *Tetrahedron Lett.* **40** (1999) 5467.
42. W. E. Parham, L. J. Reed, *Org. Synth.* **28** (1948) 60.
43. J. J. Wang et al., *Youji Huaxue* **23** (2003) no. 8,827.
44. Rhone-Poulenc, FR 2589466, 1987 (S. Ratton).
45. Montedison SpA, EP 0167053, 1986 (R. Santi et al.).
46. K. Shaw, C. Nolan, *J. Org. Chem.* **22** (1957) 1668.
47. Hüls AG, EP 0811607, 1997 (F. Bauer, M. Feld).
48. B. Arct et al., *J. Environ. Sci. Health, Part B* **B18** (1983) (4 – 5), 559.
49. Degussa AG, DE 10026108, 2001 (W. Kleemiss, G. Koehler).
50. L. Zhou et al., *Huadong Ligong Daxue Xuebao* **25** (1999) 424.
51. N. Albertson, S. Archer, *J. Am. Chem. Soc.* **67** (1945) 2043.
52. J. Fillman, N. Albertson, *J. Am. Chem. Soc.* **70** (1948) 171.
53. Mitsubishi Rayon Co, JP 08173174, 1996 (E. Ozaki et al.).
54. Creanova Spezialchemie GmbH, EP 0950653, 1999 (J. Muhr, M. Feld).
55. Degussa–Huels AG, EP 1070701, 2001 (J. Muhr).
56. A. L. Gutman et al., *J. Org. Chem.* **57** (1992) 1063.
57. Denki Kagaku K. K., Kogyo K. K., JP 62039544, 1987 (N. Tachikawa et al.).
58. I. N. Nesterova et al., *Khim.-Farm. Zh.* **8** (1994) 41.
59. Ciba-Geigy AG, EP 0115469, 1984 (N. Yokoyama).
60. D. J. Brown in E. C. Taylor (ed.): *The Chemistry of Heterocyclic compounds*, vol. **52**, John Wiley & Sons, New York 1994.
61. Hüls AG, DE 19640756, 1998 (A. Hunds).
62. Zeneca Limited, WO 9744327, 1997 (R. V. H. Jones et al.).
63. Degussa AG, EP 1284261, 2003 (A. Kuever, A. Hunds).
64. SKW Trostberg AG, EP 0603893, 1994 (H. Hintermaier, U. Maier, S. Weiss).
65. Lonza AG, CH 685497, 1995 (D. Macquarrie, R. Imwinkelried).
66. A. K. A. Rathi, S. A. Puranik, *Chem. Eng. World* **31** (1996) no. 10, 95.
67. Dynamit-Nobel, DE 2359963, 1975 (M. El-Chahawi et al.).
68. Dynamit-Nobel, DE 2524389, 1976 (U. Prange et al.).
69. Degussa AG, DE 10008899, 2001 (F. Bauer).
70. R. Tuba et al., *J. Mol. Catal. A: Chem.* **203** (2003) no. 1–2, 59.
71. M. El-Chahawi, U. Prange, *Chemiker Zeitung* **102** (1978) 1.
72. Ube Industries Ltd., EP 0006611, 1980 (K. N. Kenji et al.).
73. BASF AG, DE 4120721, 1992 (L. Schuster, K. Halbritter).
74. Mitsui Toatsu Chemicals, Inc., JP 63170338, 1988 (K. Watabe, H. Ono).
75. Nippon Kokan K. K., JP 03077848, 1991 (K. Fujimoto, T. Shikada).
76. M. Tishler et al., *J. Am. Chem. Soc.* **69** (1947) 1487.
77. Lyonnaise Industrielle Pharmaceutique, FR 1363948, 1964 (E. Szarvasi, M. Bayssat).
78. Merrell Toraude S.A., BE 873766, 1979 (M. W. Gittos, G. J. Letertre).
79. J. R. Geigy AG, US 2562830, 1949 (H. Stenzl).

80. Hoffmann-La Roche, DE 2337340, 1974 (L. Berger, A. J. Corraz).
81. Stirling Drug Inc., BE 612258, 1962 (G. Y. Lesher, M. D. Gruett).
82. G. Y. Lesher et al., *J. Med. Pharm. Chem.* **5** (1962) 1063.
83. Warner-Lambert, US 3287458, 1966 (D. Kaminsky et al.)
84. Daiichi Seiyaku Co., Ltd., EP 0047005, 1982 (I. Hayakawa et al.).
85. Mediolanum Farmaceutici S.r.l., JP 60208987, 1985;
EP 0252352, 1988 (P. Terni et al.).
86. Otsuka Pharmaceutical Co., Ltd., DE 3324034, 1984 (M. Uchida et al.).
87. Nippon Soda Co., Ltd., JP 56075408, 1981.
88. ICI Australia Ltd., EP 0080301, 1983 (R. B. Warner et al.).
89. E. I. du Pont de Nemours and Co., AT 8316181, 1984; BR 8303322, 1984; US 4547215, 1985 (A. D. Wolf).
90. Imperial Chemical Industries Ltd., GB 571014, 1945 (G. A. P. Tuey).
91. E. I. du Pont de Nemours and Co., WO 9507265, 1995 (W. C. Appleton, P. A. Parziale).
92. Chemie Linz GmbH., EP 0682018, 1995 (F.-T. Schwarz, J. Altreiter).
93. Nihon Nohyaku Co., Ltd., DE 2316921, 1973 (K. Taninaka et al.).
94. Nihon Nohyaku Co., Ltd., DE 2545569, 1976 (K. Taninaka et al.).
95. Nihon Nohyaku Co., Ltd., US 4327223, 1982 (H. Matsui et al.).
96. Firmenich et Cie, GB 907431, 1960 (A. Firmenich et al.).
97. F. Freeman in L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis*, vol. 2, 1410 – 1411 and vol. 4, 2416 – 2418, John Wiley & Sons, New York 1995.
98. R. Malinowski, J. Legocki, *Pol. Organika* (1977) 46.
99. Degussa–Huels AG, US 6288264, 2001 (M. Immig, M. Kaufhold).
100. C. K. Cain, M. F. Mallette, E. C. Taylor, *J. Am. Chem. Soc.* **68** (1946) 1996.
101. R. A. Hartz et al., *J. Med. Chem.* **47** (2004) 4741.
102. Eprova AG, EP 0924195, 1999 (H. R. Müller et al.).
103. Lonza AG, DE 2700733, 1977 (E. Catalucci).
104. Lonza AG EP 0342616, 1989 (H. P. Mettler, F. Previdoli).
105. Nation. Oil. Prod. Co., US 2365295, 1941 (K. H. Schaaf, F. D. Pickel)
106. G. Jones, *Org. React.* **15** (1967) 204.
107. Medichem S.A., ES 2149687, 2001 (A. Stampa Diez del Corral et al.).
108. J. A. Montgomery et al., *J. Med. Chem.* **36** (1993) 55.
109. J. Guyot, A. Kergomard, *Tetrahedron* **39** (1983) 1167.
110. Hoechst, DE 4408083, 1995 (M. Beller, K. Forstinger).
111. Asahi Glass Co., JP 57035552, 1982.
112. Haarmann und Reimer, EP 0419980, 1991 (R. Hopp et al.).
113. BASF, EP 0395982, 1990 (G. Lauterbach et al.).
114. Lonza AG, WO 9708152, 1997 (B. Schmidt et al.).
115. Mitsubishi Rayon Co., Ltd., JP 10229891, 1998 (E. Ozaki, K. Enomoto).
116. Mitsubishi Rayon Co., Ltd., Nitto Chemical Industry Co., Ltd., JP 10229890, 1998 (E. Ozaki, R. Endo).
117. J. K. H. Inglis, *Org. Synth. Coll. Vol. I* (1941) 254.
118. Lonza AG, GB 824640, 1959.
119. Monsanto, US 2539238, 1951 (C. M. Eaker).
120. I.G. Farbenindustrie, DE 640509, 1934.
121. Phillips, DE 1951032, 1969 (P. G. J. Wesselmann, W. Houtenlaan).
122. Knapsack, US 3384654, 1968 (K. Sennewald et al.).
123. Mitsui Toatsu Chemicals, JP 50071627, 1975 (M. Kawamata et al.).
124. Monsanto, US 3945896, 1974 (D. A. Tysee).
125. K. Chiba et al., *Chem. Lett.* **1** (1990) 39.
126. Nitto Chemical Industry Co., JP 53029995, 1978 (I. Watanabe, M. Kumakura).
127. Mitsui Toatsu Chemicals, JP 62032891, 1987 (T. Hagiwara et al.).
128. B. Wermeckes, F. Beck, *Chem. Ber.* **118** (1985) 3771.
129. Nitto Chemical Industry Co., JP 56021595, 1981.
130. Catalytica, WO 9212962, 1992 (A. R. Bulls, J. D. Fellmann, R. A. Periana).
131. Denki Kagaku Kogyo, JP 07233135, 1995 (H. Kaneko, T. Masuda).
132. Lonza AG, WO 0116092, 2001 (P. Hanselmann et al.).
133. Nippon Chemical Kogyo Co., Ltd, JP 2006016365, 2006 (K. Midorikawa et al.).
134. Ube Industries Ltd., EP 0055158, 1982 (K. Matsui et al.).
135. N. Manada et al., *Nippon Kagaku Kaishi* **7** (1994) 667.

136. Sagami Chem. Res., JP 06092931, 1994 (T. Fuchigami, N. Ko, H. Urata).
137. Denki Kagaku Kogyo, JP 08081396, 1996 (T. Ootsuka et al.).
138. Kawatetsu Chemical Industry Co., Ltd., JP 59148751, 1984.
139. T. Suzuki et al., *Nippon Kagaku Kaishi* **10** (1983) 1482.
140. Lonza AG, WO 0128983, 2001 (P. Chen et al.)
141. BASF AG, DE 19539293, 1997 (C. Grund et al.).
142. A. Said, *Chimia* **28** (1974) 234.
143. D.-W. Zhu, *Synthesis* **10** (1993) 953.
144. Nitto Chemical Industry, JP 10095760, 1998 (A. Koizumi, T. Narita).
145. Nitto Chemical Industry, WO 9517377, 1995 (K. Nakamura, Y. Takayanagi, M. Seya).
146. H. Bredereck, A. Edenhofer, *Chem. Ber.* **88** (1955) 1306.
147. S. Ganapathy, *Chem. Eng. World* **33** (1998) 110.
148. E. I. du Pont de Nemours and Co., US 3954992, 1973 (S. H. Davidson).
149. L. Greci et al., *J. Chem. Res., Synop.* **10** (2003) 655.
150. O. Schnider, A. Grüssner, *Helv. Chim. Acta* **34** (1951) 2211.
151. Merck and Co., Inc., US 3313813, 1967 (E. J. Cragoe).
152. Merck and Co., Inc., US 3625944, 1971 (E. Grabowski et al.).
153. W. Klötzer, H. Bretschneider, *Monatsh. Chem.* **87** (1956) 136.
154. H. Bretschneider et al., *Monatsh. Chem.* **92** (1961) 75.
155. H. Bretschneider et al., *Monatsh. Chem.* **92** (1961) 128.
156. Lonza AG, WO 9621664, 1996 (G. Stucky, G. Griffiths).
157. Boehringer Ingelheim, DE 4342131, 1995 (R. Ramert, R. Sobotta).
158. Kojin KK, JP 05213952, 1993 (T. Sato).
159. Smithkline Beecham PLC, WO 9213859, 1992 (J. C. Hanson).
160. Wellcome Foundation Ltd., DE 2539963, 1976, (H. J. Schaeffer).
161. H. Matsumoto et al., *Chem. Pharm. Bull.* **36** (1988) 1153.
162. Syntex, Inc., US 4355032, 1982; Syntex, Inc., US 4423050, 1983 (J. P. Verheyden, J. C. Martin).
163. Beecham Group PLC EP 355986, 1990 (P. M. Kincey).
164. S. S. G. Sircar, *J. Chem. Soc.* (1927) 600, 1252.
165. C. A. Miller, L. M. Long, *J. Am. Chem. Soc.* **73** (1951) 4895.
166. Labaz, US 4155929, 1979 (M. Chignac et al.).
167. Ciba Pharmaceutical Products Inc., US 2868803, 1959 (J. Druey, P. Schmidt).
168. S. Yamamoto et al., *Nippon Noyaku Gakkaishi* **21** (1996) 259.
169. Ciba-Geigy, EP 0333661, 1989 (B. Schaub, H. Kaenel, P. Ackermann).
170. Farbwerke Hoechst AG, DE 1545790, 1965 (O. Scherer, H. Mildenberger).
171. May and Baker Ltd., EP 0295117, 1988 (J. G. Buntain et al.).
172. Rhone-Poulenc, WO 9732843, 1997 (D. W. Hawkins et al.).
173. Rhone-Poulenc, EP 0511845, 1992 (J. Huang et al.).
174. Dow Chemical Co., US 3751444, 1973; DE 2326186, 1974 (R. H. Solem, L. D. Berg).
175. S. Hafenrichter, *Textil-Praxis* **16** (1961) 273.
176. BASF Corp., EP 0430023, 1991 (S. I. Goldstein, E. F. Labrie).
177. BASF AG, DE 19519894, 1996 (M. Holderbaum et al.).
178. F. Freeman, *Chem. Rev.* **69** (1969) 591.
179. A. J. Fatiadi, *Synthesis* (1978) 165, 241.
180. F. Freeman, *Synthesis* (1981) 925.
181. F. Freeman in L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis*, vol. **5**, John Wiley & Sons, New York 1995, pp. 3214 – 3217.
182. V. P. Litvinov, *Russ. Chem. Rev.* **68** (1999) 39.
183. R. A. Carboni, D. D. Coffmann, E. G. Howard, *J. Am. Chem. Soc.* **80** (1958) 2838.
184. E. C. Taylor, K. S. Hartke, *J. Am. Chem. Soc.* **81** (1959) 2452.
185. Bayer AG, EP 1329482, 2003 (J.-W. Stawitz, U. Feldhues).
186. Lonza AG, DE 2329251, 1973 (H. Marketz).
187. J. P. Fleury et al., *Bull. Soc. Chim. France* **2** (1964) 413.
188. F. Freeman, *Chem. Rev.* **80** (1980) 329.
189. E. Campaigne, S. W. Schneller, *Synthesis* (1976) 705.
190. S. M. Sherif, A. M. Hussein, *Monatsh. Chem.* **128** (1997) 687.
191. T. H. Affi, *Adv. Colour Sci. Technol.* **6** (2003) no. 3, 63.
192. D. Konwar, R. C. Boruah *Indian J. Chem., Sect. B: Org. Chem. Incl. Med. Chem.* **36B** (1997) 918.
193. P. Boldt et al., *Chem. Ber.* **100** (1967) 1281.
194. R. A. Carboni, *Org. Synth.* **39** (1959) 64.
195. E. I. du Pont de Nemours, US 2794824, 1955 (R. E. Heckert, E. L. Little).

196. E. I. du Pont de Nemours, US 3076836, 1963 (E. L. Martin).
197. A. J. Fatiadi, *Synthesis* (1986) 249.
198. A. J. Fatiadi, *Synthesis* (1987) 749, 959.
199. FMC, US 4059433, 1976 (L. K. Gibbons).
200. H. Schmidtmann, *Chem. Ber.* **29** (1896) 1168.
201. Lonza AG, WO 2006021390 (C. Täschler et al.).
202. Seitetsu Chemical Industry Co., Ltd., JP 62270556, 1987 (H. Itsuda et al.).
203. Dow Chemical Industry Co., US 5171743, 1992 (Y. C. Tong).
204. Dow Chemical Co., US 5180740, 1993 (D. R. Romer, R. G. Pews).
205. Novartis AG, WO 9814449, WO 9814450, WO 9814451, 1998 (G. Bold et al.).
206. Hoechst Schering AgrEvo G.m.b.H., DE 19751943, 1999 (A. Angermann et al.).
207. Lonza AG, EP 0508353, 1992 (G. Ryan, H.-P. Mettler, F. Previdoli).
208. VEB Fahlberg-List, DD 266101, 1989 (J. Spindler et al.).
209. F. M. Abdel-Galil, S. M. Sherif, M. H. Elnagdi, *Heterocycles* **24** (1986) 2023.
210. B. Y. Riad et al., *Heterocycles* **26** (1987) 205.
211. R. Liboska et al., *Synthesis* **12** (2002) 1649.
212. E. I. du Pont de Nemours and Co., US 4412957, 1983 (J. S. Gramm).
213. Lonza AG, DE 19926257 A1, 1999 (Y. Bessard).
214. A. K. Ghosh et al., *Tetrahedron Lett.* **37** (1996) 3815.
215. E. I. du Pont de Nemours and Co., EP 0024200, 1981 (J. J. Fuchs, E. K. W. Wat).
216. I. W. Davies et al., *Tetrahedron Lett.* **37** (1996) 813.
217. G. K. Jnaneshwara et al., *Tetrahedron Lett.* **39** (1998) 459.
218. A. K. Ghosh, P. Mathivanan, J. Cappiello, *Tetrahedron Asymmetry* **9** (1998) 1.
219. Salk Institue for Biological Studies, US 3670007, 1972 (J. P. Ferris).
220. Lonza, EP 0003335, 1979 (H. Junek, M. Mittelbach).
221. Lonza, EP 0298261, 1989 (H.-P. Mettler).
222. C. F. Shey et al., *Shih Ta Hsueh Pao (Taipei)* **29** (1984) 631.
223. E. Merck (Darmstadt), DE 165692, 1905.
224. A. A. Joshi et al., *Bioorg. Med. Chem. Lett.* **15** (2005) no. 1, 73.
225. T. W. Bell et al., *Org. Prep. Proced. Int.* **34** (2002) no. 3, 321.
226. B. B. Corson, R. W. Scott, C. E. Vose, *Org. Synth.* **10** (1930) 66.
227. A. R. Surrey, *J. Am. Chem. Soc.* **65** (1943) 2471.
228. Lonza Ltd., FR 1365202, 1964.
229. BASF AG, DE 19506993, 1996 (D. Hermeling, H. Siegel).
230. Zhennan Chemical Plant, CN 1264701, 2000 (J. Cao).
231. Tongyuan Fine Chemical Plant, CN 1250773, 2000 (Z. Wu et al.).
232. Creanova Inc., EP 1065189, 2001 (C. S. Subramaniam et al.).
233. Council of Scientific and Industrial Research, India, EP 1184369, 2002 (A. K. Bandyopadhyay et al.).
234. Lonza, CH 493473, 1970 (U. Arni, A. Faucci).
235. Lonza, US 3683003, 1972 (K. Aufdereggen et al.).
236. Ube Industries Ltd., JP 57131750, 1982 (S. Furusaki et al.).
237. Ube Industries Ltd., JP 56113752, 1981 (S. Tawara et al.).
238. Asahi Chemical Industry Co., Ltd., JP 58004755, 1983 (T. Katsumata et al.).
239. Takeda Chemical Industries Ltd., JP 47028968, 1972 (K. Morita et al.).
240. US 3578696, 1971 (J. H. Blanck, M. Prince).
241. Little, Arthur D., Inc., US 3497546, 1970 (P. L. Levins et al.).
242. Lonza AG, US 5959136, 1999 (P. Chen et al.).
243. R. Grewe, *Z. Physiol. Chem.* **242** (1936) 89.
244. Hoffmann-La Roche, FR 831110, 1938.
245. E. I. du Pont de Nemours and Co., EP 0051466, 1982 (R. F. Sauers).
246. E. I. du Pont de Nemours and Co., US 4746353, 1988 (G. Levitt).
247. Nissan Chemical Industries, Ltd., JP 60208977, 1985 (S. Yamamoto et al.).
248. K. Morimoto et al., *J. Heterocycl. Chem.* **34** (1997) no. 2, 537.
249. Henkel KGaA, DE 2651794, 1978 (D. Rose).
250. Smith-Kline and French Laboratories, US 3081230, 1963 (J. Weinstock, V. D. Wiebelhaus).
251. Lonza AG, DE 2741383, 1979 (E. Catalucci).
252. D. R. Seeger et al., *J. Am. Chem. Soc.* **71** (1949) 1753.
253. M. Sekiya, J. Suzuki, *Chem. Pharm. Bull.* **20** (1972) 209.
254. M. Sekiya et al., *Chem. Pharm. Bull.* **24** (1976) 1331.
255. Kohjin Co., Ltd., EP 0045503, 1982, (J. Suzuki).
256. Merck and Co., CH 630617, 1977, (M. C. Van der Zwan, D. F. Reinhold).
257. Upjohn Co., US 4032559, 1977 (J. M. McCall, J. J. Ursprung).

258. J. M. McCall et al., *Synthesis* (1978) 673.
259. Ciba-Geigy, WO 9503282, 1995 (J.-M. Assercq, H. P. Schwemlein, J. W. Perine).
260. Ciba-Geigy, BR 9500315, 1997 (J.-M. Assercq et al.).
261. American Aniline Products, Inc., US 3909198, 1975 (E. E. Renfrew, H. W. Pons).
262. Sandoz Ltd., GB 2026528, 1980 (W. Baumann).
263. Hoechst, DE 557665, 1932; DE 566154, 1932 (H. Greune, W. Eckert).
264. T. Nakamura, K. Kikuchi, *Kino Zairyo* **9** (1989) 43.
265. D. S. Acker, W. R. Hertler, *J. Am. Chem. Soc.* **84** (1962) 3370.
266. Federal Laboratories Inc., US 3963770, 1976 (J. S. Knapp).
267. H. Staats, *Schriftenr. Inst. Toxicol. Univ. Kiel* **18** (1991) 145.

Manganese and Manganese Alloys

David B. Wellbeloved, Samancor Limited, Johannesburg, Republic of South Africa

Peter M. Craven, Samancor Limited, Johannesburg, Republic of South Africa

John W. Waudby, Samancor Limited, Johannesburg, Republic of South Africa

1. History 2
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 3
3. Occurrence 3
4. Mining and Beneficiation 5
5. Reduction of Manganese Oxides . 6
5.1. Manganese – Oxygen System . . . 6
5.2. Reduction of Manganese Oxides with Carbon Monoxide, Hydrogen, and Carbon 7
5.3. Reduction of Mixed Oxides and Minerals Containing Manganese Oxides 10
5.4. Reduction of Manganese Oxides by Silicon 10
5.5. Reduction of Manganese Oxide by Aluminum 10
6. Production of Manganese-Containing Ferroalloys 11
6.1. Production of High-Carbon Ferromanganese 11
6.1.1. Production of Ferromanganese in Blast Furnaces 11
6.1.2. Production of Ferromanganese in Electric Arc Furnaces 15
6.1.2.1. Design and Operation of Electric Furnaces 19
6.1.2.2. Raw Materials Required for the Manufacture of High-Carbon Ferromanganese 20
6.1.2.3. Chemistry of the Process 22
6.1.2.4. Discard-Slag Practice 22
6.1.2.5. High-Slag Practice 23
6.2. Production of Silicomanganese . . 24
6.3. Production of Medium-Carbon Ferromanganese 26
6.3.1. Production of Medium-Carbon Ferromanganese by Oxygen Refining of High-Carbon Ferromanganese . . . 26
6.3.2. Silicothermic Production of Medium-Carbon Ferromanganese . 27
6.4. Production of Low-Carbon Ferromanganese 28
6.5. Gas Cleaning 28
6.6. Developments and Future Trends 29
7. Production of Manganese Metal . 29
7.1. Electrolysis of Aqueous Manganese Salts 30
7.2. Electrothermal Process 33
7.3. Aluminothermic Process 33
7.4. Silicothermic Process 34
7.5. Distillation of Ferromanganese . . 34
7.6. Electrolysis of Fused Salts 34
8. Manganese as an Alloying Component 35
8.1. Manganese in the Iron and Steel Industry 35
8.2. Manganese in Nonferrous Alloys 36
8.3. Nitrogen-Containing Manganese Alloys 36
9. Processing of Manganese Nodules and Encrustations 37
9.1. Occurrence, Origin, and Composition 37
9.2. Legal Considerations and Nodule Recovery 39
9.3. Processing 40
9.3.1. Gas Reduction and Ammoniacal Leach Process 40
9.3.2. Cuprion Ammoniacal Leach Process 40
9.3.3. Elevated Temperature and Pressure H_2SO_4 Leach Process 40
9.3.4. Reduction and HCl Leach Process . 41
9.3.5. Smelting and H_2SO_4 Leach Process 43
9.4. Economic Viability and Future Prospects 43
10. Economic Aspects 45
11. References 45

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_077

1. History [1], [2]

There is some dispute over the origin of the name "manganese". Deposits of manganese ore (probably pyrolusite) were apparently found near Magnesia in Asia Minor. This was referred to by PLINI around 50 B.C. as *magnesia negra* to distinguish it from *lapis magnes* (magnetite), which has a similar appearance and density. The name pyrolusite was derived only in 1826.

The useful properties of manganese were probably first identified in the mineral pyrolusite, which was used by the ancient Egyptians and Romans to control the color of glass. Small additions of the ore removed the greenish yellow discoloration caused by iron impurities, while further additions imparted pink, purple, or black colorations.

During the 1400s the name braunstein was used for all the manganese oxide ores. SCHEELE recognized manganese as an element in 1774, and GAHN succeeded in isolating the metal for the first time during the same year. The name manganese was first used by BUTTMAN in 1808.

Manganese ore may have been used in ironmaking for centuries, but found formal usage in iron- and steelmaking only at the beginning of the 19th century. Ferromanganese was first produced on an industrial scale in 1841 by POURCEL in France for use in crucible steels, and by 1850 spiegeleisen (an alloy of 20 % manganese and iron) had been produced commercially in Prussia. The Bessemer process was made practical by MUSHET, who advised the addition of spiegeleisen in 1856/1857. The use of ferromanganese as an additive to steel to counteract the harmful effects of sulfur and phosphorus was patented by SIEMENS in 1866, and by 1875, POURCEL was making 65 % ferromanganese in a blast furnace. HADFIELD developed a wear-resistant steel containing manganese as an alloying agent in 1888. Ferromanganese was first produced in an electric arc furnace during 1890.

A two-stage process for the production of refined ferromanganese with 1 – 2 % carbon was proposed in 1908, and manganese metal was first produced industrially by an aluminothermic process in 1898. In 1866 LECLANCHE filed a patent for the MnO_2-NH_4Cl-Zn dry-cell battery.

2. Properties [1–5]

2.1. Physical Properties

Manganese is a silver-grey metal, resembling iron. It is hard and very brittle, and its primary uses in a metallic form are as an alloying, desulfurizing, and deoxidizing agent for steel, cast iron, and nonferrous metals. Due to its high tendency to oxidize, annealing colors are often evident.

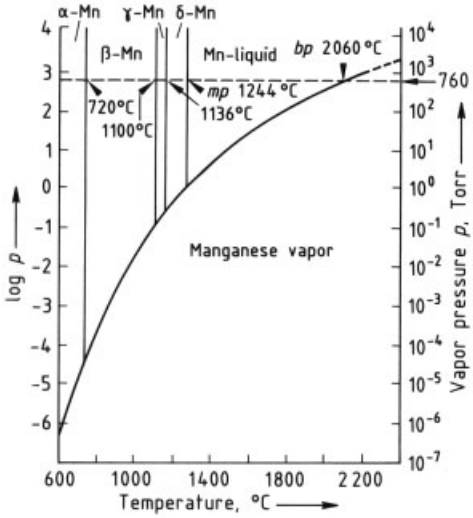

Figure 1. The vapor pressure of manganese [3]

Manganese belongs to group 7 and period 4 of the periodic table with an atomic number of 25 and atomic mass of 54.94. Six isotopes have been produced, of which only ^{55}Mn is stable. The other isotopes, ^{51}Mn, ^{52}Mn, ^{54}Mn, ^{56}Mn, and ^{57}Mn, have half-lifes ranging from 46 min to 310 d. Manganese exists in four allotropic forms: α, β, γ, and δ. Some important physical and thermodynamic properties are listed in Table 1. The vapor pressure of manganese in millibars is given by:

$$\log p = -13625/T + 8.8837 \ (770 - 885\,^\circ\mathrm{C})$$
$$\log p = -13994/T + 9.3360 \ (957 - 1097\,^\circ\mathrm{C})$$
$$\log p = -12275/T + 8.3730 \ (1250 - 1550\,^\circ\mathrm{C})$$

and the molar heat capacity in J mol^{-1} K^{-1} as a function of temperature by:

α-Mn (0.3 – 4.2 K):

$$C_p = [16\,T + 2.39\,T^{-2} + 0.0266\,T^{-3}]\cdot 10^{-3}$$

α-Mn (298 – 1000 K):

$C_p = 23.9 + 14.2 \cdot 10^{-3}\,T - 1.55 \cdot 10^5\,T^{-2}$
β-Mn (1000 – 1374 K):
$C_p = 34.9 + 2.765 \cdot 10^{-3}\,T$
γ-Mn (1374 – 1400 K): $C_p = 44.80$
δ-Mn (1410 K – *mp*): $C_p = 47.31$
Liquid Mn (*mp* – *bp*): $C_p = 46.06$

The vapor pressure diagram of manganese is shown in Figure 1.

2.2. Chemical Properties

Manganese exhibits certain similarities to its neighbors in the periodic table, iron and chromium. The electronic configuration is $1\,s^2\ 2\,s^2p^6\ 3\,s^2p^6d^5\ 4\,s^2$.

Manganese occurs in all oxidation states between Mn(−III) and Mn(VII), with Mn(II) being the most stable. Trivalent manganese is less stable and readily disproportionates into Mn(II) and Mn(IV). The remaining oxidation states occur mainly in complexes.

The basicity of the oxides decreases with increasing valency of the manganese. Thus, manganese(II) oxide is the anhydride of a base

$$Mn(OH)_2 \rightarrow MnO + H_2O$$

while manganese(VII) oxide Mn_2O_7 is an acid anhydride:

$$2\,HMnO_4 \rightarrow Mn_2O_7 + H_2O$$

Manganese(IV) oxide (MnO_2) is amphoteric. The oxidation states Mn(II) and Mn(VII) are of industrial importance, forming a large number of useful salts. In metallurgy, oxides of di-, tri-, and tetravalent manganese are important, especially Mn(II) oxide.

At normal temperatures, pure manganese is not attacked by oxygen, nitrogen, or hydrogen. At high temperatures it reacts violently with oxygen, sulfur, and phosphorus. It is therefore a powerful agent for deoxidizing and desulfurizing metals, whereby it is converted to a divalent oxide or sulfide. Reaction with chlorine forms manganese(II) chloride; with SO_2 the products are MnS and MnO.

Manganese dissolves in acids with liberation of hydrogen and formation of Mn(II) salts. Hot concentrated sulfuric acid dissolves manganese with evolution of SO_2, and nitric acid with evolution of hydrogen, nitrogen, and dinitrogen monoxide.

3. Occurrence [5–12]

Manganese is the twelfth most abundant element in the earth's crust, at 950 ppm [5].

Mineralogy. Manganese is contained in many minerals, of which approximately 250 can be regarded as true manganese minerals. The major ores, generally with manganese contents above 35 %, are hydrated or anhydrous oxides. The silicates and carbonates occur to a lesser extent. The major manganese minerals are listed in Table 2.

Economically significant manganese deposits are all of sedimentary origin, having been dissolved from crystalline rocks and deposited as the oxide, hydroxide, or carbonate.

Wad is the term given to the soft, earthy, hydrous mixture of manganese and iron oxides of variable manganese content.

Ore Quality. Typically, only ores containing a minimum of 35 % Mn are regarded as manganese ores. Ores containing 10 – 35 % Mn are categorized as ferruginous manganese ores, and iron ores with 5 – 10 % Mn are referred to as manganiferous ores.

The three most important areas of application of manganese ores and the specific requirements, are as follows:

1) Metallurgical grade ore for the iron and steel industry, which generally contains 38 – 55 % Mn. The phosphorus content should preferably be below 0.1 %, and the concentrations of Al_2O_3, SiO_2, CaO, MgO, and S are important. The manganese/iron ratio is critical; a 7.5 : 1 ratio, for example, is required for a standard ferromanganese alloy with 78 % Mn.
2) Battery grade ore, containing 70 – 85 % MnO_2 (44 – 54 % Mn). The ore should generally contain less than 0.05 % of metals more electronegative than zinc, such as copper, nickel, cobalt, and arsenic. The suitability of manganese dioxide for use in batteries depends on a number of factors, including the crystal structure, surface area, pore size distribution, particle shape and size, electrical conductivity, surface conditions, chemical composition, and structure defects [2].

Table 1. Properties of manganese

Property	α-Mn	β-Mn	γ-Mn	δ-Mn	Molten Mn
Atomic number	25				
Atomic mass	54.938				
Crystal structure	complex cubic	complex cubic	fcc	bcc	
Lattice parameter, nm	0.8894	0.6290	0.3774	0.3080	
Transition temperature, °C	$\alpha \leftrightarrow \beta$ 727	$\beta \leftrightarrow \gamma$ 1095	$\gamma \leftrightarrow \delta$ 1104	$\delta \leftrightarrow$ liquid 1244	liquid $\leftrightarrow$ gas 2032
Standard electrode potential, V	1.134				
Magnetic susceptibility, m^3/kg	1.21×10^{-7}				
Density, g/cm^3	7.44	7.29	7.21		$7.55 - 1.2 \times 10^{-3}\,t$ (1244 – 1500 °C)
Brinell hardness, MPa	3923 – 4119				
Hardness, Mohs scale	5				
Volume increase on transition, %	$\alpha \leftrightarrow \beta$ 3.4	$\beta \leftrightarrow \gamma$ 0.8	$\gamma \leftrightarrow \delta$ 0.8	$\delta \leftrightarrow$ liquid 2.9	
Resistivity, Ω cm	$150 - 260 \times 10^{-6}$	90×10^{-6}	40×10^{-6}		
Compressibility, m^2/N	8.4×10^{-7}				
Coefficient of linear expansion, K^{-1}	22.3×10^{-6}	24.9×10^{-6}	14.8×10^{-6}	41.6×10^{-6}	
Latent heat of fusion, J/g	244				
Latent heat of vaporization, J/g	4020				
Heat of transition, J/mol	2240	2282	1800	14 655	219 901
Specific heat, $J\,g^{-1}\,K^{-1}$	0.477	0.482	0.502		
Enthalpy, J/mol	4999		5112		
Entropy, $J\,mol^{-1}\,K^{-1}$	32.0		32.3		173.8

Table 2. Manganese minerals of economic importance

Mineral	Formula	Theoretical Mn content, %	Relative density
Pyrolusite	β-MnO_2	63.2	5
Braunite	$Mn^{2+}Mn_6^{3+}SiO_{12}$	66.6	4.7 – 4.9
Manganite	γ-MnOOH	62	4.3
Psilomelane	$(Ba, Mn^{2+})_3$ $(O, OH)_6Mn_8O_{16}$	45 – 60	4.4 – 4.7
Cryptomelane	KMn_8O_{16}	62	4.3
Hausmannite	Mn_2MnO_4	73	4.7 – 5.0
Jacobsite	Fe_2MnO_4	23.8	4.8
Bixbyite	$(Mn, Fe)_2O_3$	30 – 40	5
Rhodonite	$(Mn, Fe, Ca)SiO_3$	42	3.4 – 3.6
Rhodochrosite	$MnCO_3$	47.6	3.3 – 3.6
Bementite	$Mn_3Si_2O_5(OH)_4$	43.2	3.5

3) Chemical grade ore whose specifications vary considerably depending on the end use. Included in this category are feed stocks for electrolytic manganese and manganese dioxide, manganese chemicals, colorants and, in the Republic of South Africa, an oxidant in uranium extraction.

Ore Deposits. The most important manganese ore deposits are located in the Republic of South Africa, the Soviet Union, Australia, Gabon in West Africa, and Brazil. The deposits in India and Mexico are of declining importance, and the majority of the production is consumed domestically.

The largest known land-based manganese deposit is the Kalahari Field, located in the Northern Cape Province of the Republic of South Africa, which contains ca. 78 % ($13\,600 \times 10^6$ t) of the world's potential resources. The surface area underlain by this deposit is approximately 320 km^2. The deposit is sedimentary and interlayered with banded ironstone.

A breakdown of the world's total estimated manganese ore reserves is presented in Table 3.

Table 3. Manganese reserves and production (1987)

	Ore reserves, 10^6 t [13]	Ore production, 10^3 t [14]	High-carbon ferromanganese production, 10^3 t [14]	Silicomanganese production, 10^3 t [14]
Ore-producing countries				
Republic of South Africa	13 600	2892	300 [a, b]	275
former Soviet Union	2300	9400 [a]	[c]	[c]
Gabon	400	2400		
Australia	400	1680	51	42
Brazil	160	1970	136	188
China	100	1300	[c]	[c]
India	60	1200	170	
Others	90	1028		
Non-ore-producing countries				
Federal Republic of Germany			170	
Japan			303	135
Mexico			124	80
Norway			90	237
United Kingdom			92	
France			293	5
Others			256	312
Total	16 210	21 870	1985	1274

[a] Estimated.
[b] Includes medium-carbon production.
[c] Statistics not available.

Table 4. Typical composition of some important metallurgical manganese ores (chemical content on % dry basis)

Producer	Product	Mn	Fe	P	SiO_2	Al_2O_3	CaO	MgO
Samancor, RSA	Mamatwan Lumpy	37–38	4–6	0.08	4–6	0.5	14–16	2.4
Samancor, RSA	Mamatwan Ferruginous	32–34	4–6	0.08	4–6	0.5	14–16	3.5
Samancor, RSA	Wessels High Grade	50–51	9–11	0.08	5–7	0.5	4–6	0.6
Samancor, RSA	Wessels Low Grade	38–40	18–20	0.08	5–7	0.5	4–6	1.2
Groote Eylandt, Australia	Metallurgical Lump	49.0	2.5	0.06	6.0	4.0	0.05	0.10
Groote Eylandt, Australia	Siliceous Lump	44.0	2.5	0.06	15.0	4.0	0.05	0.10
Amapa, Brazil	Bitolado 48	48–49	5.5–6.9	0.08–0.11	2–3	3.7–5.5		
Amapa, Brazil	Carbonate Grosso	33–37	2.5–4	0.04–0.10	10–14	2.0–3.8	3.3–5.2	2.2–4.5
CVRD, Brazil	Lumpy	45.0	3.7	0.08	3.6	9.8	0.25	0.30
Comilog, Gabon	Gabon Lumpy	48–51	3–5	0.11–0.14	2.5–4	6	0.11	0.07
former Soviet Union	Tchiaturi Fines	48.2	1.1	0.142	9.2	1.6	2.26	0.50

Each of the major manganese producers offers a range of ore products, the detailed chemical and physical characteristics of which are given in [8]. Analyses of some typical commercial grades are presented in Tables 4 and 5.

Another potentially important source of manganese are the deep-sea nodules located over certain areas of the ocean floor.

4. Mining and Beneficiation

The methods employed for the mining and beneficiation of manganese ore vary widely depending on the size and nature of the deposit, ore type, end uses, and size of the operation.

Mining. The major producers of manganese ore utilize both underground and open-cast mining techniques. In the Republic of South Africa, Samancor extracts ore from the open-cast Mamatwan Mine and the underground Wessels Mine. At Mamatwan the ore body attains a maximum thickness of 45.2 m and is overlain by, on average, 45 m of overburden. The open pit is approximately 1900 m by 300 m, and the ore is extracted from three benches by drilling, blasting, loading, and hauling to an in-pit crusher. At Wessels the ore body is located at a depth of ca. 300 m and a room and pillar mining system is employed. The average stope height is ca. 5 m,

Table 5. Typical compositions of some important battery grade manganese ores (chemical content on % dry basis)

Producer	MnO_2	Mn	Fe	Pb	Cu	Ni	As	Co
CVRD, Brazil	76.5	51.0	3.2	0.035	0.050	0.080	0.004	0.035
Comilog, Gabon	83.6		1.8		0.03	0.05		0.095

and the ore is drilled, blasted, loaded, and hauled to underground primary crushing facilities.

At Groote Eylandt, BHP operates Australia's only manganese mine by open-cut methods from various quarries, each limited to an area ca. 250 m long by 60 m wide and normally having a single bench [15].

The former Soviet Union's primary manganese source is the Nikopol field, where the thickness of the ore seam is 1.5 – 2.5 m on average occurring at a depth of 100 – 120 m. A combination of underground and open-cut mining methods are employed [16].

Beneficiation. Ore treatment varies from simple crushing, washing, and sizing plants to fairly sophisticated beneficiation techniques. Gravity separation techniques are frequently employed, such as the dense-medium separation plants at Groote Eylandt and Mamatwan. Extensive investigations on a pilot-plant scale have been undertaken by Samancor on ore sorting techniques, based on neutron activation or X-ray fluorescence; this has not as yet led to a commercial operation. Froth flotation with fatty acids, amines, sulfates, sulfonates, and hydroxamates has been shown to be feasible on a laboratory scale [17], as have high-gradient magnetic separation techniques [18].

Beneficiation techniques are becoming progressively more complex as reserves of high-grade ores diminish and requirements for smelting and steelmaking become more stringent.

Intermediate treatment processes. Various intermediate treatment processes can be applied to manganese natural ores or concentrates prior to smelting to a ferroalloy or direct addition to a steelmaking process. These processes are employed for one or both of the following two reasons: 1) to convert a fine ore to a more suitable size, and 2) to calcine and/or reduce the ore to a more suitable final product.

Agglomeration processes include nodulizing, pelletizing, and briquetting. Reduction can be carried out in a rotary kiln or sinter plant.

Sintering has become the preferred process for intermediate treatment, and sintering grate areas varying from 3 m^2 to 80 m^2 and production capacities of 2 – 300 t/h.

Sinter is produced on a moving grate from a mixture of fine ore, fine reductant/fuel, and recycled sinter fines. The fuel in the bed is ignited, and ignition is propagated by applying suction to the grate. A sufficiently high temperature is generated to fuse the ore particles together. During sintering, the ore is calcined and partially reduced. Some of the benefits derived from sintering ore are:

1) Fine ore, which has a limited application and value in conventional smelting, is agglomerated and converted to a superior product.
2) Reduced gas volumes, and hence fewer furnace eruptions, result when smelting sinter.
3) Furnace availability and operating loads are increased.

5. Reduction of Manganese Oxides [19]

Manganese is only produced commercially from blends of ore in which manganese exists as an oxide. A knowledge of the thermodynamics of the reduction of these oxides is therefore essential to the understanding of the production of alloys from manganese ores.

5.1. Manganese – Oxygen System

Of the known oxides of manganese (Mn_2O_7, MnO_2, Mn_5O_8, Mn_2O_3, Mn_3O_4, and MnO), the only oxides found naturally in manganese ores are MnO_2, Mn_2O_3, Mn_3O_4, and MnO.

Under normal conditions, manganese dioxide occurs as stable β-MnO_2; the other modifications, α, γ, δ, and ε, are not pure varieties of MnO_2. Dimanganese trioxide (Mn_2O_3) exists in the α-modification, which is cubic above 20 °C and rhombic below this temperature; the γ-modification is metastable. Trimanganese tetroxide (Mn_3O_4) exists at atmospheric pressure in the tetragonal α-modification and changes to the cubic β-form at 1170 °C. Manganese oxide (MnO) in normal conditions exists in only one modification. This oxide is, however, nonstoichiometric and is more accurately described as $Mn_{1-x}O$, with $x = 0-0.25$ [20]. The complete phase diagram of the manganese – oxygen system above 1000 °C is given in Figure 2.

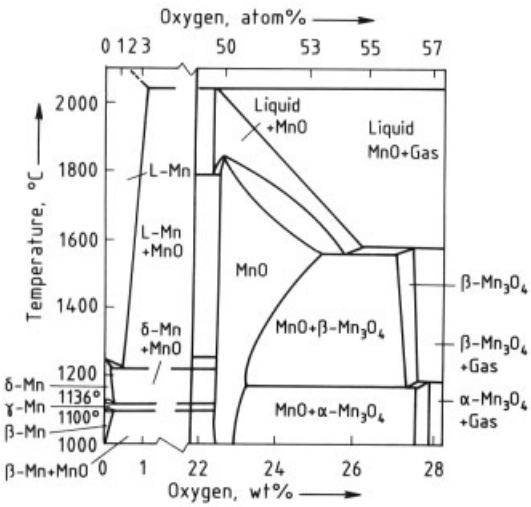

Figure 2. Mn – O phase diagram above 1000 °C [3]

The equations for the thermal dissociation of the important manganese oxides are as follows:

$$4\,MnO_2 \rightarrow 2\,Mn_2O_3 + O_2 \quad (1)$$
$$\Delta H_{298} = 166.0\,kJ$$
$$6\,Mn_2O_3 \rightarrow 4\,Mn_3O_4 + O_2 \quad (2)$$
$$\Delta H_{298} = 64.8\,kJ$$
$$2\,Mn_3O_4 \rightarrow 6\,MnO + O_2 \quad (3)$$
$$2\,MnO \rightarrow 2\,Mn + O_2 \quad (4)$$

The oxygen partial pressures exerted by these reactions are given as a function of temperature in Figure 3 [21]. According to the thermodynamic data in [22], the decomposition temperature at 101.3 kPa for reaction (1) is 500 °C and for reaction (2), 980 °C. For reaction (3) to occur at atmospheric pressure a reductant is necessary. Reaction (4) requires a high reducing potential. It is important, therefore, to consider the reduction of the oxides of manganese in the presence of carbon monoxide, hydrogen, and carbon. These are the reactions that usually take place in the blast furnaces and electric furnaces used to produce manganese alloys.

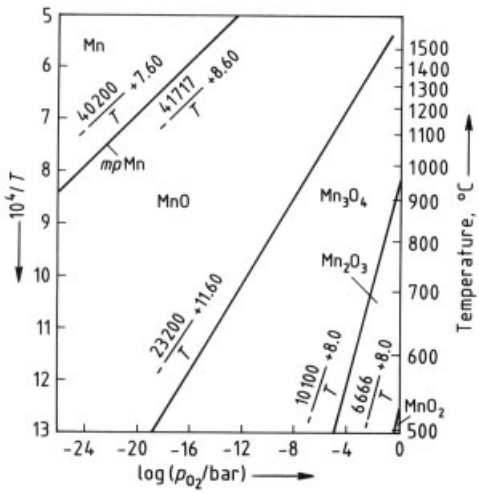

Figure 3. Variation of oxygen partial pressure in the Mn – O system with temperature

5.2. Reduction of Manganese Oxides with Carbon Monoxide, Hydrogen, and Carbon

The reduction of MnO_2 to Mn occurs in four main steps (see Eqs. 1 – 4). This reduction is strongly exothermic in the presence of carbon, methane, and carbon monoxide; the following equations apply:

$$MnO_2 + 1/4\,C \rightarrow 1/2\,Mn_2O_3 + 1/4\,CO_2 \quad (5)$$
$$\Delta H_{298} = -56.8\,kJ$$

$$MnO_2 + 1/8\,CH_4 \rightarrow 1/2\,Mn_2O_3 + 1/4\,H_2O + 1/8\,CO_2 \quad (6)$$
$$\Delta H_{298} = -58.6\,kJ$$

$$MnO_2 + 1/2\,H_2 \rightarrow 1/2\,Mn_2O_3 + 1/2\,H_2O \quad (7)$$
$$\Delta H_{298} = -79.5\,kJ$$

$$MnO_2 + 1/2\,CO \rightarrow 1/2\,Mn_2O_3 + 1/2\,CO_2 \quad (8)$$

$$\Delta H_{298} = -99.6\,\text{kJ}$$

The reduction of Mn_2O_3 in the presence of the same reducing agents is also exothermic. The following equations apply:

$$1/2\,Mn_2O_3 + 1/12\,C \rightarrow 1/3\,Mn_3O_4 + 1/12\,CO_2 \quad (9)$$

$$\Delta H_{298} = -16.3\,\text{kJ}$$

$$1/2\,Mn_2O_3 + 1/24\,CH_4 \rightarrow 1/3\,Mn_3O_4 + 1/24\,CO_2 + 1/12\,H_2O \quad (10)$$

$$\Delta H_{298} = -17.2\,\text{kJ}$$

$$1/2\,Mn_2O_3 + 1/6\,H_2 \rightarrow 1/3\,Mn_3O_4 + 1/6\,H_2O \quad (11)$$

$$\Delta H_{298} = -24.3\,\text{kJ}$$

$$1/2\,Mn_2O_3 + 1/6\,CO \rightarrow 1/3\,Mn_3O_4 + 1/6\,CO_2 \quad (12)$$

$$\Delta H_{298} = -31.0\,\text{kJ}$$

The third step is the reduction of Mn_3O_4, which is endothermic with carbon or methane and exothermic with hydrogen or carbon monoxide:

$$1/3\,Mn_3O_4 + 1/6\,C \rightarrow MnO + 1/6\,CO_2 \quad (13)$$

$$\Delta H_{298} = +11.8\,\text{kJ}$$

$$1/3\,Mn_3O_4 + 1/12\,CH_4 \rightarrow MnO + 1/12\,CO_2 + 1/6\,H_2O \quad (14)$$

$$\Delta H_{298} = +10.5\,\text{kJ}$$

$$1/3\,Mn_3O_4 + 1/3\,H_2 \rightarrow MnO + 1/3\,H_2O \quad (15)$$

$$\Delta H_{298} = -3.3\,\text{kJ}$$

$$1/3\,Mn_3O_4 + 1/3\,CO \rightarrow MnO + 1/3\,CO_2 \quad (16)$$

$$\Delta H_{298} = -17.2\,\text{kJ}$$

The last step, the reduction of MnO, is more difficult and requires a high reducing potential. The presence of carbon is therefore necessary, and at a carbon monoxide partial pressure of 101.3 kPa a temperature of 1420 °C is necessary for the reaction to proceed. However, the formation of manganese carbides in the presence of carbon lowers the reaction temperature to 1300 °C (see Section 6.3.1). The following, highly endothermic overall reaction applies:

$$MnO + C \rightarrow Mn + CO \quad (17)$$

$$\Delta H_{298} = 274.5\,\text{kJ}$$

This reaction can be broken down into

$$MnO + CO \rightarrow Mn + CO_2 \quad (18)$$

and the reaction of CO_2 with C to produce CO

$$CO_2 + C \rightarrow 2\,CO \quad (19)$$

which ensures that a high reducing potential is maintained.

Due to the importance of both carbon monoxide and carbon as reducing agents in the commercial production of manganese-containing alloys, these are discussed in more detail. It is possible to predict the ratio CO/CO_2 that will reduce the individual manganese oxides by considering the relationship between the oxygen partial pressures and the equilibrium constants for the reactions (1) – (4)

$$K_{p(1)-(4)} = p_{O_2} \quad (20)$$

(assuming that the activities of the solid phases are unity) and the equilibrium constant of the reaction

$$2\,CO + O_2 \rightarrow 2\,CO_2 \quad (21)$$

which is

$$\frac{1}{K_p} = \frac{p_{CO}^2 \cdot p_{O_2}}{p_{CO_2}^2} \quad (22)$$

and the relationships

$$G^0 = -RT \ln K_p \quad (23)$$

and

$$G^0 = -RT \ln p_{O_2} \quad (24)$$

the oxygen potentials of the individual reactions at various temperatures can be calculated. This is shown graphically in the Richardson diagram [21] (see Fig. 4). The same applies to the use of hydrogen as a reductant. Although the reduction of MnO_2, Mn_2O_3, and Mn_3O_4 takes place at low partial pressures of CO, the presence of carbon is necessary for the reduction of MnO. In this case the following reaction is important:

$$2\,C + O_2 \rightarrow 2\,CO \quad (25)$$

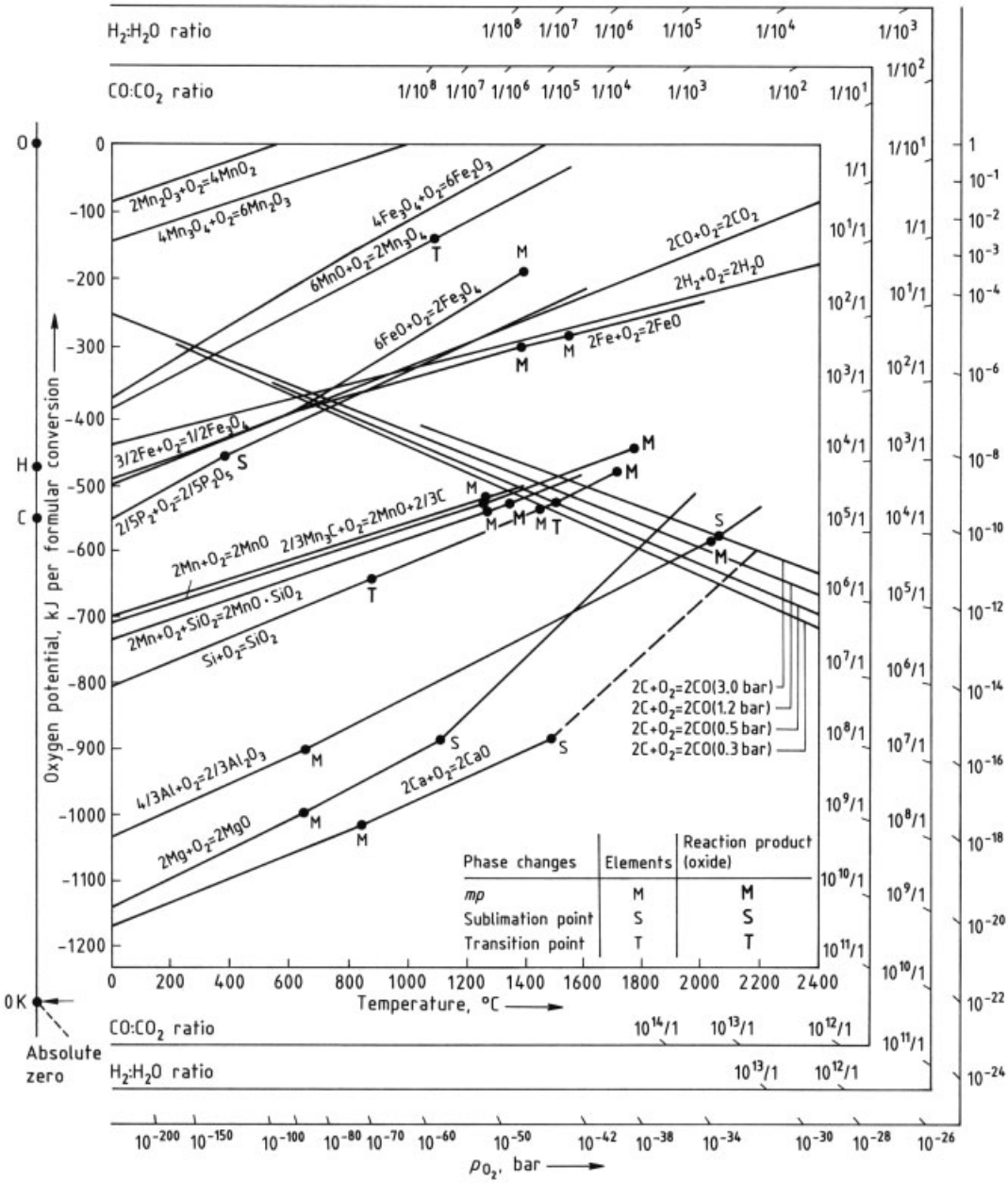

Figure 4. Temperature dependence and oxygen potentials of oxide systems [21]

for which

$$\log p_{O_2} = -\frac{11672}{T} - 9.16 \tag{26}$$

This reaction is pressure dependent due to the change in the number of moles of gas, and again the Richardson diagram can be used to predict under which conditions the reduction of MnO will occur (see Eqs. 17, 18, and 19). However, manganese carbides are always formed during the reduction of MnO with carbon, and this lowers the temperature at which the reaction occurs.

The products produced in this process always contain carbon. The phase diagram for the manganese – carbon system is shown in Figure 5. The melting point increases with increasing carbon content [23]. Commercial ferromanganese (76 % Mn, 16 % Fe) contains ca. 7 % carbon. The solubility of carbon decreases with increasing silicon content.

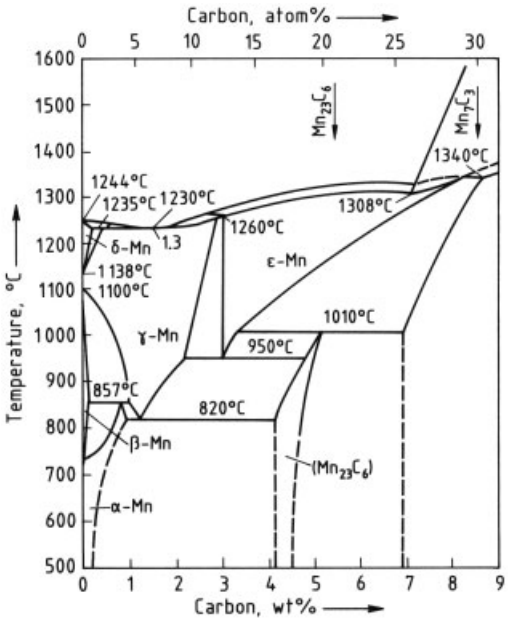

Figure 5. Phase diagram for the Mn – C system [23]

5.3. Reduction of Mixed Oxides and Minerals Containing Manganese Oxides

In the discussion of the reduction of manganese oxides it has been assumed that the activity of the solid phases is unity. However, this is usually not true in practice because in most manganese ores the manganese oxides are combined with iron oxides, silicates, aluminum oxides, calcium oxides, and phosphorus oxides.

The oxides FeO and P_2O_5 are reduced at lower temperature than MnO, and SiO_2 at only slightly higher temperature (see Fig. 4). Any iron oxides and oxides of phosphorus occurring in the ore, fluxes, and reductants used in commercial processes are reduced together with the manganese oxides and report to the product. A judicious choice of ores and raw materials is therefore necessary to limit the amount of these elements which report to the metal. The silica contained in the ores can also be reduced, and in the production of silicomanganese this is desirable. Quartz or quartzite is therefore an important raw material for the production of silicomanganese.

The presence of other oxides in the ores reduces the activity of the manganese oxide. This must be compensated for by increasing the reducing potential by increasing the CO/CO_2 ratio of the gas mixture. Alternatively, the reduction temperature must be increased if the reaction is carried out with carbon.

5.4. Reduction of Manganese Oxides by Silicon

Figure 4 shows that the reduction of manganese(II) oxide by silicon is also possible:

$$2\,MnO + Si \rightarrow SiO_2 + 2\,Mn \tag{27}$$

for which

$$K = \frac{a_{SiO_2} \cdot a_{Mn}^2}{a_{MnO}^2 \cdot a_{Si}} \tag{28}$$

This reaction is the basis of the Frisch process for the production of low-carbon ferromanganese and manganese metal from silicomanganese. The silicon content of the metal is restricted by using a basic slag that reduces the activity of SiO_2 by reaction with CaO, displacing reaction (27) as far to the right as possible.

5.5. Reduction of Manganese Oxide by Aluminum

It is easier to reduce manganese(II) oxide with aluminum than with silicon because the differences in the oxygen potentials of the dissociation equilibria of MnO and Al_2O_3 are greater than for MnO and SiO_2 (see Fig. 4). As in the case of carbon and silicon, the higher oxides of manganese are also reduced if present in the ore. The aluminothermic process is used in the production of manganese metal (see Section 7.3).

Table 6. Types of ferromanganese and their general compositions [26]

Alloy	Composition, %		
	Manganese	Carbon	Silicon
High-carbon ferromanganese (carburé)	72 – 80	7.5	< 1.25
Medium-carbon ferromanganese (affiné)	75 – 85	< 2.0	
Low-carbon ferromanganese (suraffiné)	76 – 92	0.5 – 0.75	
Silicomanganese	65 – 75	< 2.5	15 – 25
Ferromanganese silicon	58 – 72	0.08	23 – 35
Spiegeleisen	16 – 28	< 6.5	11 – 45

6. Production of Manganese-Containing Ferroalloys

A number of manganese-containing ferroalloys are manufactured which are used largely in the mild steel, foundry, and stainless steel industries [24–26]. The names and typical compositions of these alloys are given in Table 6, and the international standards for the most commonly used alloys, namely high-carbon ferromanganese HC FeMn and silicomanganese FeSiMn, are given in Table 7 [24]. These are generally classified as intermediate products and the range of their end uses is shown in Figure 6.

Generally, high-carbon ferromanganese and silicomanganese are produced from a blend of manganese-containing ores, and in the case of silicomanganese, slags and silica are added. Ferromanganese can be produced in either electric submerged furnaces or blast furnaces, although only four blast furnace producers exist in the Western world [25], whereas silicomanganese is largely produced in submerged arc furnaces. Producers of high-carbon ferromanganese and silicomanganese are listed in Table 8. High-carbon ferromanganese can be converted to medium-carbon manganese by an oxygen blowing process, and silicomanganese can be further refined into medium- or low-carbon ferromanganese as well as manganese metal (see Fig. 7).

6.1. Production of High-Carbon Ferromanganese

6.1.1. Production of Ferromanganese in Blast Furnaces [25–29]

Ferromanganese can be produced in blast furnaces in a manner similar to pig iron and spiegeleisen; however, in the Western world only four producers employ this method. These are Thyssen Stahl (Federal Republic of Germany), BSC Cleveland (United Kingdom), SFPO (France), and Mizushima (Japan) (see Table 10), [25]. In the former Soviet Union, the majority of the high-carbon ferromanganese is produced in blast furnaces [26]. The choice of the use of blast furnaces over electric furnaces is based on the relative price of coke and electricity. Blast furnaces are usually used where the cost of power is high in relation to coke. In blast furnaces, coke is used both as a reductant and as the energy source. The coke rate in blast furnaces is higher than in submerged arc furnaces, which use electricity as the power source [27]. An exception to this is SFPO, where off-gases from the furnace are used to produce electricity, which is sold back to the local power supplier.

The product produced from blast furnaces generally contains 76 % Mn and 16 % Fe; the ferromanganese produced in the former Soviet Union is generally of a lower grade (see Table 9) [26].

Raw Material Selection and Pretreatment. The raw materials required for the production of high-carbon ferromanganese are manganese ores, fluxes such as limestone, dolomite, or silica, and solid fuels and reductants such as coke.

In order to produce ferromanganese of the required grade a single ore is seldom suitable

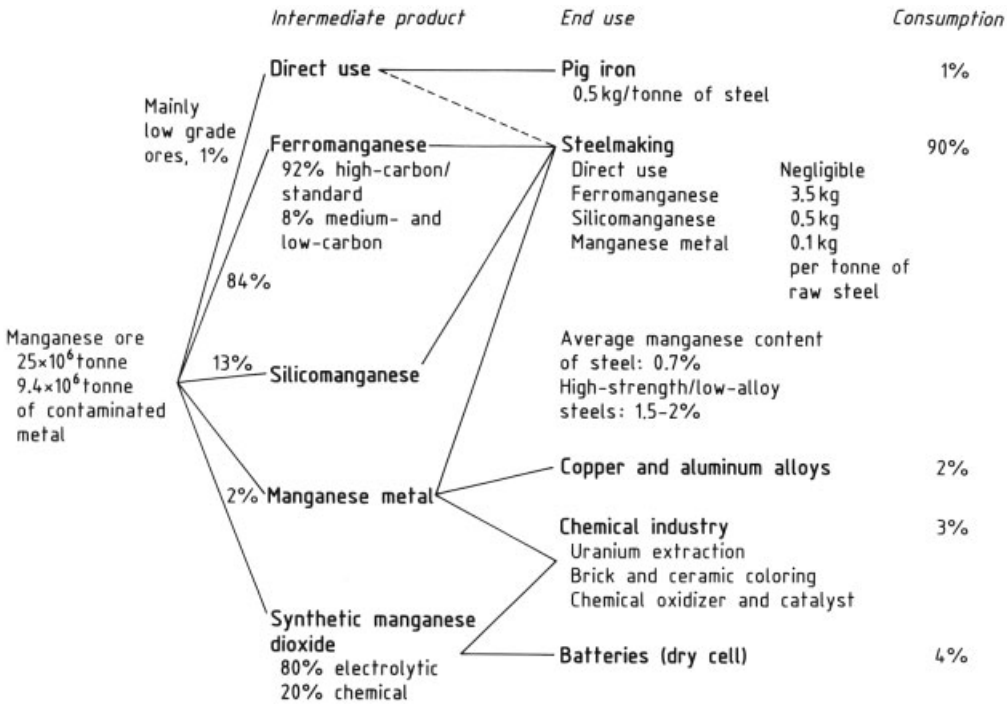

Figure 6. Manganese intermediate products and end uses

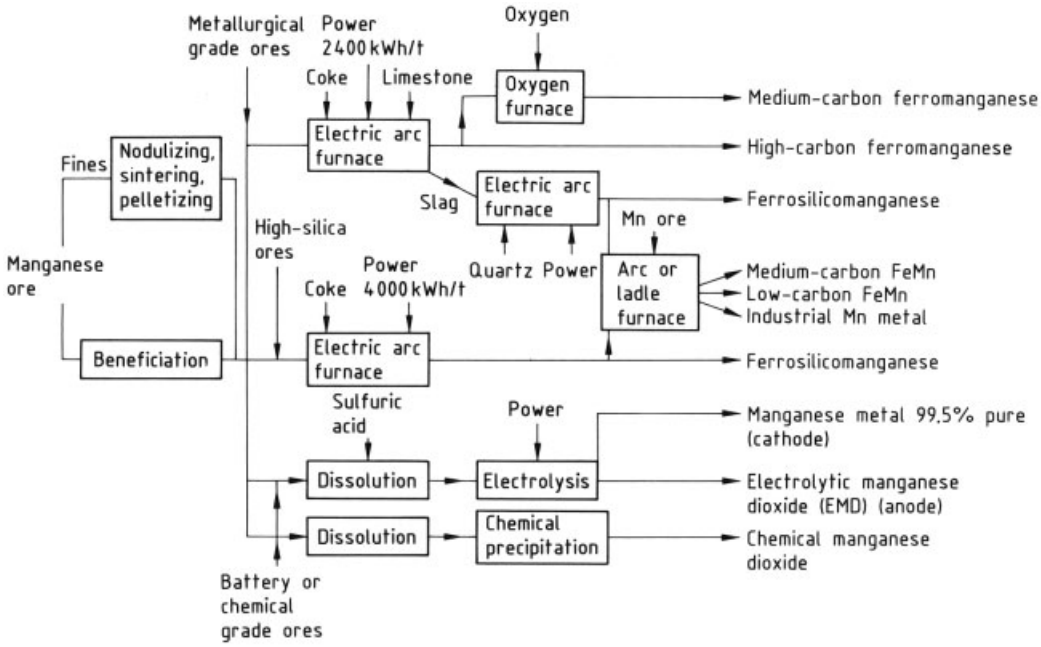

Figure 7. Summary of process routes

Table 7. Ferromanganese standards [24]

Alloy	Country or organization	Standard	Reference
FeMn	International Standards Organization	DIS	5446
	France	AFNOR NF	A-15-020
	Japan	JIS	G 2301
	United States	ASTM	A 99-66
	former Soviet Union	GOST	4755-70
	Federal Republic of Germany	DIN	17 564
FeSiMn	International Standards Organization	DIS	5447
	France	AFNOR NF	A-13-030
	Japan	JIS	G 2304
	United States	ASTM	A 701
	former Soviet Union	GOST	4756-70
	Federal Republic of Germany	DIN	17 564

because the desired Mn/Fe ratio of the charge determines the Mn content of the final product [27]. Ores from various sources are therefore blended to achieve the ideal ratio and to limit the contents of the deleterious components silica, alumina, and phosphorus in the raw material mix.

The raw material is crushed and screened to ca. 5 – 30 mm. Alternatively, sintered or pelletized fine ore can be used (see Chap. 4). Some deleterious components can be partially removed from the ore prior to melting by dense-medium separation or flotation. Slagging components (dolomite or limestone) can be added to the sintered or pelletized ore, which results in cost savings in the blast furnace. Partial reduction of the higher manganese oxides may also occur during sintering.

Blast Furnace Operation. In comparison to iron making, high gas temperatures are required in ferromanganese production because the reduction of manganese(II) oxide takes place at a higher temperature than is required for the reduction of wustite [28]. This is achieved by oxygen enrichment of the hot blast or, in the case of SFPO, by heating the blast with nontransferred arc plasma torches [29]. The plasma arc increases the flame temperature from 2200 to 2800 °C and considerably reduces the coke consumption, which usually ranges from 1270 to 2000 kg/t.

The recovery of manganese in the alloy is usually 75 – 85 %. This is influenced by the MnO content of the slag, the slag-to-metal ratio, and losses in the flue gases. The MnO content of the slag is highly dependent on the basicity ratio $(CaO + MgO)/SiO_2$ (see Fig. 8) [30], which can be controlled by the choice of ore and addition of flux. Losses to the flue gas can generally be recovered in the gas cleaning section (see Section 6.5). These materials can then be agglomerated and returned to the furnace.

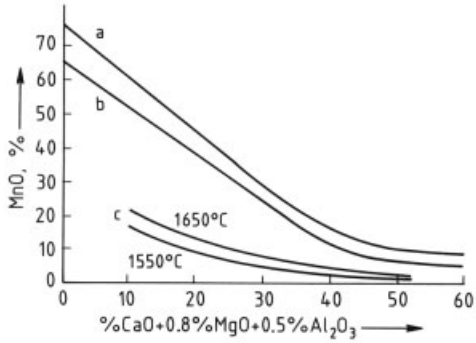

Figure 8. Variation of equilibrium MnO content of slag for high-carbon ferromanganese and silicomanganese production [30]
a) High-carbon ferromanganese (78 % Mn, 0.1 % Si, 7 % C, 14.9 % Fe); b) High-carbon ferromanganese (77 % Mn, 1.0 % Si, 7 % C, 15 % Fe); c) Silicomanganese (71 % Mn, 16 % Si, 1.8 % C, 10.2 % Fe)

At the Mizushima works, the double bell valve of the conventional blast furnace has been replaced with an arrangement incorporating a distribution chute (Fig. 9) [31]. This results in a better distribution of the burden in the shaft and therefore a more even flow of gas through the burden (see Fig. 10). The incorporation of a distribution chute lowers the coke consumption of the furnace.

In spite of the innovations mentioned above the raw material costs of blast furnaces remain higher than those of submerged arc furnaces [25]

Table 8. Producers of high-carbon ferromanganese and silicomanganese (producers of 10 000 t/a of either ferroalloy in 1988) [25]

Country	Company	Site	Capacity [a], $\times 10^3$ t (1988)		Production, $\times 10^3$ t (1988)	
			FeMn	SiMn	FeMn	SiMn
RSA	Samancor	Meyerton	256	169	255	165
RSA	Transalloys	Witbank	0	125	0	125
RSA	Ferroalloys	Cato Ridge	80	0	75	0
India	Tisco	Keonjhar	31	0	25	0
India	Facor	Shreeramnagar	10	0	10	0
India	Khandelwal	Khandelwal Nagar	20	11	18	10
India	Sandur Manganese	Hospet	30	16	26	10
India	UF & A	Maneck Nagar	19	30	10	30
India	Maharashtra Electrosmelt	Chandrapur	68	11	67	11
Japan	Nippon Denko	Hidaka	0	13	0	10
Japan	Kobe Steel	Kakogawa	61	32	30	35
Japan	Mizushima [b]	Kurashiki	94	0	95	0
Japan	Chuo Denki Kogyo	Kashima	85	57	60	30
Japan	NKK	Niigata	37	16	35	15
Japan	Nippon Denko KK	Tokushima	57	0	40	0
Japan	Pacific Metals	Hachinohe	43	0	13	0
Japan	JMC	Takaoka	85	31	60	30
Korea	Korea Ferroalloy	Pohang	36	30	28	26
Korea	Dongil Chungong	Pohang	16	28	13	24
Korea	Dongbu Industry	Tonghae	43	32	34	27
Taiwan	Chen Hsing Industrial	Taipei	20	0	18	0
Belgium	Sadacem	Gent	36	10	35	10
France	SFPO [b]	Boulogne	400	0	372	0
France	Pechiney	Dunkerque	0	60	0	25
Federal Republic of Germany	Thyssen Stahl [b]	Duisburg-Hamborn	240	0	246	0
Italy	Carlo Tassara	Breno	25	80	10	75
Italy	Italghisa	Bagnolo Meila	12	0	15	0
Norway	Elkem	Porsgrunn	65	60	65	50
Norway	Tinfos Jernverk	Kvinesdal	0	120	0	120
Norway	Elkem	Sauda	165	60	155	60
Spain	Hidro Nitro	Monzon	0	15	0	10
Spain	Fyesa	Boo de Guarnizo	38	35	9	25
Spain	Carburos Metalicos	Cee	15	21	10	14
United Kingdom	BSC Cleveland [b]	Middlesborough	120	0	100	0
Argentina	Grassi	Fray Luis Beltran	27	0	25	0
Argentina	Grassi	Malargue	0	20	0	12
Brazil	Paulista	Barbacena	15	27	15	27
Brazil	Sibra	Simoes Filho	106	53	100	50
Brazil	Paulista	Rancharia	0	40	0	35
Brazil	Prometal [c]	Aruja	18	22	8	18
Brazil	Paulista	Corumba	0	24	0	24
Brazil	Maringa	Itapeva	12	13	9	11
Brazil	Paulista	Caxambu	1	15	0	15
Mexico	Minera Autlan	Teziutlan	4	17	4	17
Mexico	Minera Autlan	Tamos	92	43	91	42
Venezuela	Hevensa	Puerto Ordaz	0	30	0	29
Canada	Elkem Canada	Beauharnois	50	45	50	25
USA	Elkem Metals	Marietta	70	60	65	60
Australia	Temco	Bell Bay	72	59	55	45
Total, Africa			336	294	330	290
Total, Asia			755	307	582	258
Total, Europe			1116	461	1017	389

Table 8. (Continued)

Country	Company	Site	Capacity [a], $\times 10^3$ t (1988)		Production, $\times 10^3$ t (1988)	
			FeMn	SiMn	FeMn	SiMn
Total, Latin America			275	304	251	280
Total, North America			120	105	115	85
Total, Australia			72	59	55	45
World Total			2673	1531	2350	1347

[a] Capacity is adjusted to reflect current product mix at furnaces that produce both products.
[b] Blast furnace.
[c] Silicomanganese figures for Prometal include production for conversion to medium-carbon and low-carbon ferromanganese.

Table 9. Composition of former Soviet blast-furnace high-carbon ferromanganese [26]

Grade and official grade code	Mn content, %	Si, %	Deleterious elements, not to exceed		
			P, % (group A∗)	P, % (group B∗)	S, %
Mn-5	75.0	2.0	0.35	0.45	0.03
Mn-6	70.0 – 75.0	2.0	0.35	0.45	0.03
Mn-7	70.0 – 75.0	1.0	0.35	0.45	0.03

∗ Furnace.

due to the high cost of coke. With the exception of SFPO, blast furnace production costs are higher than the average production cost of ferromanganese in electric furnaces.

The Reduction Process in the Blast Furnace. The reduction of the higher manganese oxides to manganese(II) oxide takes place in the upper zone of the shaft according to the reactions given in Section 5.2. These generally occur below 900 °C and are indirect. The reactions are exothermic, and the heat generated causes high top temperatures and necessitates water cooling of the furnace top.

The reduction of manganese(II) oxide

$$MnO + C \rightarrow Mn + CO$$

is highly endothermic, in contrast to the weakly endothermic reduction of wustite. This requires higher temperatures and, consequently, higher coke rates are required for the smelting of ferromanganese in blast furnaces.

6.1.2. Production of Ferromanganese in Electric Arc Furnaces

The majority of producers of ferromanganese in the Western world use submerged arc electric furnaces (see Table 10). Although electric furnaces have lower capacities than blast furnaces they have the advantage that the heat requirement is provided by electricity, and coke and coal are added to the feed only as reductants. Consequently the coke consumption is lower in electric furnaces than in blast furnaces, which is a considerable advantage in the light of dramatically increasing coke prices. An additional advantage is that the process is not entirely dependent on high-strength coke, unlike blast furnaces, and a portion of the carbon requirement can be supplied in the form of coal. In South Africa, where coking coal is in short supply, up to 70 % of the carbon for the production of ferro- and silicomanganese is supplied in the form of bituminous coal.

Originally electric furnaces were small (3 – 8 MVA); however, furnaces have grown progressively larger with time [32]. More recently built electric furnaces have capacities of 75 – 90 MVA. Smaller furnaces are still popular with producers because they offer flexibility in that they can be switched more easily between different products than their larger counterparts [33]. The larger size and more stable operation of modern electric furnaces, due largely to computer control, have resulted in lower electricity consumption. Electric furnaces used in the production of manganese alloys are generally cir-

Table 10. Profiles of manganese alloy producers [25]

Country	Company (location)	Furnaces utilized on ferromanganese	Furnaces utilized on silicomanganese	Comments
Argentina	Industrias Siderurgicas Grassi (Mendosai)	4×2.2 MVA	1×8 MVA, 1×5.4 MVA	low ore cost, no export sales
Australia	Temco (Bell Bay)	1×36 MVA, 1×29 MVA, 1×27 MVA	furnaces interchangeable to silicomanganese	captive ore supply, low power cost
Belgium	Sadacem (Gent)	1×24 MVA	convertible to silicomanganese	imported ore, locally available reductants
Brazil	Companhia Cimento Portland Maringa (Sao Paulo)	1×6.5 MVA	1×6.5 MVA, 1×2.5 MVA	high reductant costs
Brazil	Paulista (Barbacena)	6×4 MVA	interchangeable to silicomanganese	
Brazil	Paulista (Caxambu)	4×4 MVA	interchangeable to silicomanganese	
Brazil	Paulista (Corumba)		3×4 MVA	no export sales
Brazil	Paulista (Rancharia Ouro Preto)		1×24 MVA	high reductant costs
Brazil	Prometal (Aruja)	1×5 MVA	1×10 MVA, 1×2.5 MVA	high reductant costs
Brazil	Sibra (Simoes)	2×24 MVA, 6×5.4 MVA	interchangeable to silicomanganese	low overall costs
Brazil	Paulista (Barbacena)	6×4 MVA	interchangeable to silicomanganese	
Canada	Elkem (Beauharnois)	1×72 MVA	interchangeable to silicomanganese	low power costs
France	SFPO (Boulogne)	3 blast furnaces		imported ore, plasma torches, surplus power
France	Pechiney Electrometallurgie (Dunkerque)		1×45 MVA	imported ore
Germany	Thyssen Stahl (Duisburg-Hamborn)	3 blast furnaces		imported ore, high reductant cost
India	Facor (Sheeramnagar)	1×7.5 MVA		high power costs
India	Khandelwal Ferro Alloys (Khandelwal Nagar)	2×9 MVA	interchangeable to silicomanganese	high power costs
India	Maharashtra Electrosmelt (Chandrapur)	1×33 MVA	interchangeable to silicomanganese	high power costs
India	Sandur Manganese & Iron ores (Hospet)	1×15 MVA	interchangeable to silicomanganese	high power costs
India	Tisco (Keonijhar)	2×9 MVA		high power costs
India	Universal Ferro & Allied Chemicals (Maneck Nagar)	2×9.5 MVA, 2×16.5 MVA	interchangeable to silicomanganese	high power costs
Italy	Italghisa (Bagnolo)	1×4.5 MVA		imported ore, high labor costs
Japan	Chuo Denki Kogyo (Kashima)	1×40 MVA	1×50 MVA	imported ore, high power costs, prereduction
Japan	Japan Metals and Chemicals Co Ltd. (Takaoka)	1×38.5 MVA	1×18 MVA	imported ore, power supply partially captive
Japan	Kobe Steel (Kakogawa)	1×20 MVA	1×20 MVA	imported ore, high power costs
Japan	Mizushima (Kurashiki)	1 blast furnace		cheapest producer in Japan
Japan	Nippon Denko (Samani)		1×7.5 MVA	imported ore, power supply partially captive
Japan	Nippon Denko (Tokushima)	1×40.5 MVA		imported ores, high power costs
Japan	NKK (Nilgata)	1×40 MVA	interchangeable to silicomanganese	imported ore, high power costs, now closed
Japan	Pacific Metals (Hachinohe)	1×15 MVA		imported ore, high power costs
Korea	Dongbu Industry (Tonghae)	2×7.5 MVA	1×15 MVA	imported ore, high power costs
Korea	Dongil Chungong (Pohang)	1×5 MVA	1×7.5 MVA, 1×4.5 MVA	imported ore, high power costs
Korea	Korea Ferroalloy (Pohang)	1×12 MVA	1×15 MVA	imported ores, high power costs

Table 10. (Continued)

Country	Company (location)	Furnaces utilized on ferromanganese	Furnaces utilized on silicomanganese	Comments
Mexico	Minera Autlan (Tamos)	1×33 MVA		ore local and
		1×33 MVA, 2×15 MVA	interchangeable to silicomanganese	imported, high power costs
Mexico	Minera Autlan (Teziutlan)	interchangeable to ferromanganese	1×3.6 MVA, 1×2.5 MVA 1×2 MVA	ore local and imported,
			1×5 MVA	high power costs
Norway	Elkem (Porsgrunn)	1×45 MVA, 1×39 MVA, 1×18 MVA	interchangeable to silicomanganese	imported ore, low power costs
Norway	Elkem (Sauda)	1×5 MVA, 1×36 MVA 1×30 MVA,	all interchangeable to silicomanganese	imported ore, low power costs
		1×24 MVA 1×10 MVA, 2×6 MVA		
Norway	Tinfos Jernverk (Kvinesdal)		2×45 MVA	imported ores, low power costs
RSA	Ferroalloys (Cato Ridge)	2×9.6 MVA, 2×9 MVA		integrated producer
RSA	Samancor (Meyerton)	1×10 MVA, 2×75 MVA	1×81 MVA, 2×18 MVA	most units
			1×10 MVA, 1×7.8 MVA	interchangeable
			2×4.6 MVA	integrated producer
RSA	Transalloys (Witbank)		1×48 MVA, 1×22 MVA, 2×21 MVA	nonintegrated producer
Spain	Carburos(Cee)	1×24 MVA	interchangeable to silicomanganese	imported ore, high power costs
Spain	Fyesa (Boo De Guarnizo)	1×20 MVA	1×30 MVA	imported ore
Spain	Hidro Nitro (Monzon)		1×45 MVA	imported ore, high power costs
Taiwan	Chen Hsing Industrial (Taipei)	2×3 MVA, 1×1.5 MVA		imported ore, high power costs, low labour costs
UK	BSC Cleveland (Middlesborough)	1 blast furnace		imported ore, high coke rate
USA	Elkem Metals	unknown (70 000 t/a)	unknown (60 000 t/a)	imported ore, low power costs
Venezuela	Havensa (Puerto Ordaz)		2×7.7 MVA, 2×6 MVA	imported ore, very low power costs

cular and have three electrodes, each coupled to a separate electrical phase (see Fig. 11) [34]. The diameters of these furnaces range from 2 to 20 m.

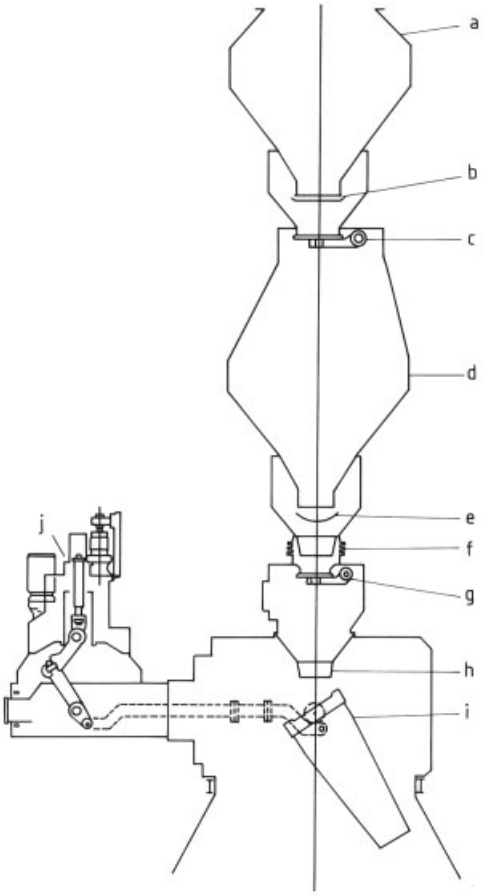

Figure 9. The cardan distributor on the blast furnace [31]
a) Upper bunker; b) Gate valve; c) Upper seal valve; d) Lower bunker; e) Material flow control gate; f) Expansion joint; g) Lower seal valve; h) Vertical chute; i) Distribution chute; j) Driving apparatus

In modern electric arc furnaces the raw material is usually fed by gravity from bunkers above the furnace. Fresh burden therefore automatically enters the furnace as the raw materials are melted and slag and metal are removed from the system. In older furnaces use is still made of charging cars to introduce raw materials to the top of the units.

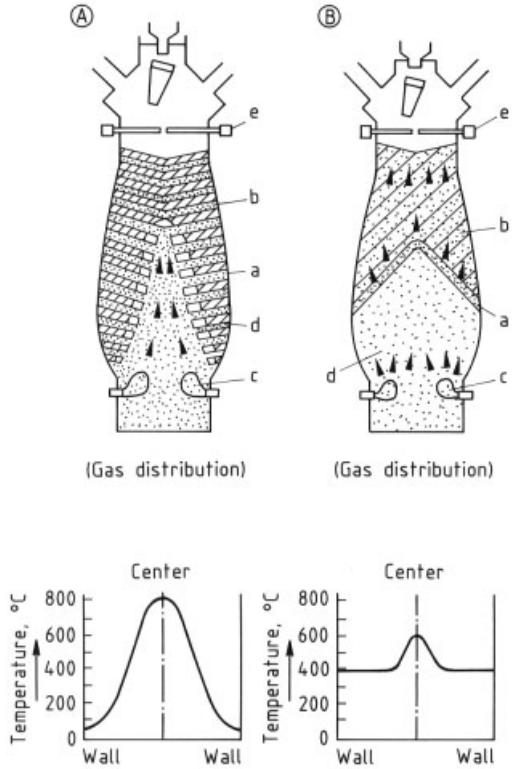

Figure 10. Effect of mixing on the distribution of gas in a blast furnace [31]
A) For layer by layer charging; B) For perfect-mixed charginga) Cohesive zone; b) Lumpy zone; c) Raceway; d) Dead man; e) Throat sonde

As the raw materials move down the furnace, the higher oxides of manganese are reduced to MnO by the gas leaving the furnace. The reduction of manganese(II) oxide occurs by the contact of carbon with the molten oxide in the slag phase. The overall reaction is:

$$MnO + 10/7\,C \rightarrow 1/7\,Mn_7C_3 + CO$$

for which [35]

$$G^0 = 265.7 - 0.18T \text{ kJ}$$

The heat required for this endothermic reaction, for heating the burden, and to compensate for heat losses is supplied by the electrical input to the furnace. Heating takes place by the flow of electricity from the tips of the electrodes, which are submerged in the burden, through the burden and slag to the metal, as well as through the flow of electricity between the electrodes [35].

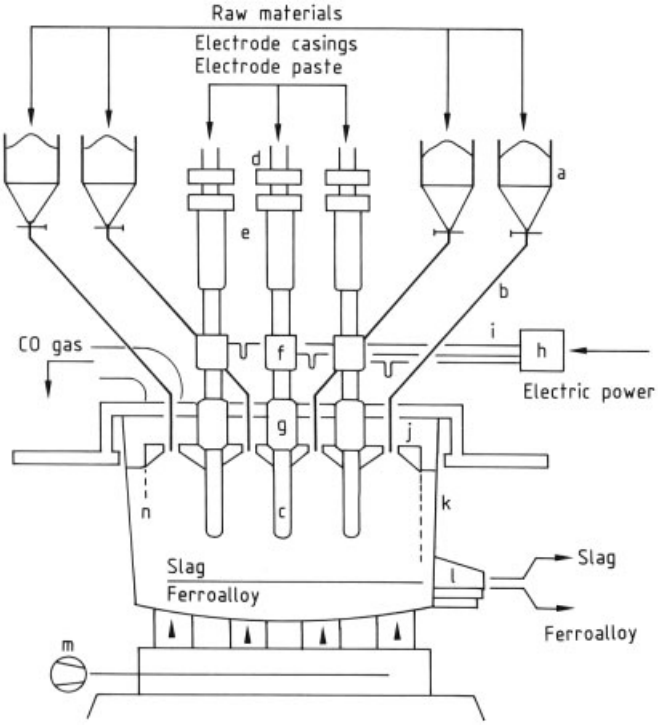

Figure 11. Layout of an electric arc furnace [34]
a) Charging bins; b) Charging tubes; c) Electrodes; d) Electrode slipping device; e) Electrode positioning devices; f) Current transmission to electrodes; g) Electrode sealing; h) Furnace transformer; i) Current bus bar system; j) Furnace cover; k) Furnace shell; l) Tap hole; m) Furnace bottom cooling; n) Refractory material

Table 11. Furnace design parameters

	Elkem Beauharnois	Elkem Sauda	Temco Bellbay	Samancor Meyerton M 10	Samancor Meyerton M 4	former Soviet Union PRO 2.5	former Soviet Union PKZ 33
Inside shell diameter, m	15.1	12.5	10.0	16.0	9.8	2.7	8.7
Shell height, m	8.8	6.0	5.2	8.0	5.7	1.2	3.0
Electrode diameter, m	1.90	1.90	1.4	1.90	1.20	0.30	1.5
Tapholes	2 Metal, 2 Slag	2 Metal, 1 Slag	2 Metal, 1 Slag	2 Metal, 1 Slag	2 Metal, 1 Slag		
Megawatt rating	4.5	30	13	46	20	2.5	33

6.1.2.1. Design and Operation of Electric Furnaces

The degree of heating depends on the electrical current flow as well as the resistance provided by the burden and the slag to the flow of electricity. In the production of ferromanganese the resistivity of the burden is low, hence low voltages between the electrodes are necessary to maintain satisfactory penetration of the electrodes in the charge. The vapor pressure of manganese is high; therefore overheating of the charge must be avoided. The current densities on the electrodes should accordingly be lower for ferromanganese production than for other ferroalloys [36]. The diameters of the electrodes are therefore larger in ferromanganese furnaces than in other ferroalloy furnaces to facilitate the high currents required for low voltage operation. The distance between electrode centers is usually larger than in other furnaces, hence the furnace diameters tend to be greater. The values of these design

parameters for a number of operating furnaces are given in Table 11.

Most electric furnaces have two tapholes offset by 60 °C which are used alternately to tap both slag and metal. The slag and metal are then separated either in the ladle or by means of a skimmer plate in the runner between the furnace and the ladle [37]. In larger furnaces, separate tapholes are included for metal and slag [33]. Tapholes are usually opened by taphole drills and closed with automatic mud guns.

An important feature in the design of submerged arc furnaces is the Soederberg electrodes. These are used because the large electrode diameters required for the production of manganese alloys make the use of prebaked electrodes uneconomic. The Soederberg electrode consists of a mild steel or stainless steel casing which is stiffened with internal fins and is filled with a carbonaceous paste, consisting of a solid aggregate, usually calcined anthracite, and a binder of coal-tar pitch [33]. The paste becomes plastic when hot and fills the entire volume of the casing. On further heating of the electrode by the electric current and furnace heat, the paste is baked and becomes solid. As the electrode is consumed, additional casings are welded onto the top. The current carrying capacity and strength of an electrode is a function of the quality of the paste, the electrode baking rate, and the cross-sectional area. Breakages of Soederberg electrodes are a major cause of downtime in electric furnaces, and proper management of the electrodes is therefore essential for efficient production [33].

A number of devices are commercially available to control the electrodes' movement and slipping rate (rate at which the electrode is allowed to move through the rings to compensate for its consumption in the furnace). One of these is designed and manufactured by Elkem [37]. The electrode is clamped by hydraulically operated rings and is moved up and down on hydraulic cylinders. Current is fed to the electrode through brass contact shoes clamped around its diameter.

Large electric furnaces are usually completely closed at the top, and the CO-rich gas leaves the furnace at approximately 290 °C and is cleaned in cyclones and venturi scrubbers. The gas is then either flamed off to the atmosphere or, more recently, is used to generate electricity. This is the case of SFPO [25] and at Tinfos in Norway [38]. At Elkem's plant in Canada the off-gases are used to fuel auxiliary equipment in the plant [39]. Smaller furnaces are either open or closed. In the case of open furnaces the gas is usually withdrawn by fans and cleaned in a bag-filter plant. In this case the gases are completely burnt in the furnace and have no commercial value.

Raw materials are usually batch weighed and blended according to a predetermined recipe and are then fed to bunkers above the furnace. The mix then gravitates into the furnace through feed chutes. To ensure an even distribution of material over the furnace, up to ten feed chutes are radially distributed around each electrode and one is positioned in the center of the furnace.

After the metal is tapped from the furnace it is cast into molds formed from ferromanganese fines or cast iron and allowed to solidify. The alloy is then removed, crushed, and screened into various size fractions, depending on the requirements of the user. An alternative to this practice is the use of a casting machine. In this case the metal is tapped directly onto a moving train consisting of small molds. The metal then solidifies and is ejected from the mold at the end of the strand. The advantage of this process is that the product is more even in size and cubical in shape. The generation of fines (−6 mm), which are generally unsalable, is also minimized. To date Elkem, Samancor, Transalloys, Chuo Denki, and Mizushima use casting machines for a portion of their products.

6.1.2.2. Raw Materials Required for the Manufacture of High-Carbon Ferromanganese

Manganese ores from different sources vary widely in their contents of manganese, iron, silica, alumina, lime, magnesia, and phosphorus. To produce standard ferromanganese (78 % Mn) and a slag containing 30 % MnO, the manganese to iron ratio in the charge must be 6 : 5. Since a single manganese ore of this ratio is seldom available, blending of ores from different sources is common practice to reach the desired manganese to iron ratio and to control the level of deletereous elements, particularly phosphorus. The ore mixes used in some operations

Table 12. Operating parameters of some ferromanganese electric arc furnaces

			Temco Bellbay	Elkem Porsgrunn	Elkem Beauharnois	Facor Sheermnagar	Samancor M 10	Samancor M 2
Raw material additions, tonnes/tonne alloy								
Manganese ore:	Groote Eylandt		1.171	(18 %)				
	South Africa:	high grade	0.780	(31 %)			0.927	0.926
		medium						
		Mamatwan ore		(51 %)			0.345	
		Mamatwan sinter					0.884	1.176
	Gabon:	Moanda ore			(78 %)			
		Moanda sinter			(22 %)			
	India:	average 44 % Mn				2.18 – 2.23		
Reductants:	coke		0.523		16 %	0.58 – 0.65	0.131	
	coal							
Fluxes:	limestone		0.121			0.55 – 0.60		
	dolomite							
	quartz						0.087	0.053
Operating results								
Average operating load, MW			13	25	34.4	5	45	16
Operating time, %			98	99		99	98	
Specific energy consumption, kWh/t			2430	2399	2050	3100	2560	2710
Manganese recovery, %			77.3	75	79	77	74	
Slag practice			high slag	high slag	high slag	discard slag	discard slag	discard slag
Slag composition, %								
MnO			40.4	43.1	44.7	20.0	20.0	20.0
MgO			1.8	4.4		2.5	7.3	7.4
Al_2O_3			13.8	11.1	28.8	9.00	4.3	3.8
CaO			15.2	11.9	1.0	32.5	35.6	34.8
SiO_2			28.2	24.2		30.5	32.4	31.5
Metal analysis, %								
Mn			78	78	78	74.5	76.5	76.5
Fe				14.6		17.6	15.4	15.4
Si			0.25	0.08		0.50	0.20	0.20
P			0.14	0.16		0.35	0.09	0.09
C			6.7	6.89		6.7	6.9	6.9
Ratios								
Slag/metal ratio			0.514				0.68	0.80
Basicity ratio (mass)			0.68			1.35	1.35	
Basicity ratio (molar)			1.9					

are shown in Table 12, [36]. The use of sinter as a source of manganese is becoming increasingly popular. In the sintering process a degree of prereduction is achieved, reducing the energy requirement in the furnace. The additional advantage of sinter is that fine ores, which are otherwise unusable, are agglomerated in the sintering process. Bag-house dust and sludge from gas cleaning plants can also be recycled to the furnace in the form of sinter. The maximum amount of sinter that can be fed to the furnace is a moot point and depends on its mineralogy and state of prereduction. When sinter replaces ore of high MnO_2 content, the energy required in the furnace increases because the highly exothermic reduction of MnO_2, Mn_2O_3, and Mn_3O_4, (see Section 5.2) no longer takes place in the furnace [36]. The use of Mamatwan sinter, in place of Mamatwan ore, results in power savings because the calcination of the carbonates in Mamatwan ore is energy consuming [40]. At the Zestafori ferralloy works in the former Soviet Union, up to 100 % sinter is used in the furnaces [41]. The bottom size of ore is also important because close packing of the ore in the furnace must be avoided [36]. This can result in the formation of calcined bridges in the furnace, which disrupt the distribution of gas and can cause eruptions when the

gas entrained under a bridge is suddenly released on its collapse [33]. Generally ores larger than 6 mm are used in large furnaces. The addition of small amounts of − 6 mm material to small furnaces is possible. Generally, ore is screened prior to batch weighing of the furnace mix.

The carbon required in the furnace is generally added in the form of coke. The size of the reductant is important. Coke and coal of too small a size can also cause close packing as well as affecting the resistivity of the burden. To maintain electrode penetration at Elkem Sauda [42] coke nuts (10 – 20 mm) are used. This is a compromise between the resistivity and the necessity of maintaining a coke bed at the bottom of the furnace.

Limestone, dolomite, or silica are added to the process as fluxes to adjust the basicity of the slag. The amount and type of flux added depends on the blend of ores and whether a discard- or high-slag practice is used. At Elkem's operation in Beauharnois the use of alumina as a flux makes the operation of the furnace possible with 100 % Moanda ore from Comilog [43].

6.1.2.3. Chemistry of the Process

A dig out of a 75 MVA ferromanganese furnace at the Metalloys operation in the Republic of South Africa showed that nine distinct zones exist around each electrode (see Fig. 12) [44]. This study showed that material descends rapidly down the side of the electrode (a) into the semi-active zone (b), where prereduction of higher manganese oxides to MnO takes place. Thereafter, the material moves into the active zones of the furnace (e, f), where reactions take place between the manganese(II) oxide in the melt and the coke particles in the coke bed:

$$7\,MnO + 10\,C \rightarrow Mn_7C_3 + 7\,CO$$

Equilibrium between the slag and metal was thought to exist under each electrode, and further from the electrode, layers of unreacted ore and coke were found to be present (h). This suggests that heat is concentrated under each electrode. The path of electrical transfer was deduced to be from the electrode tip through the coke bed and into the alloy layer (g).

The efficient production of high-carbon ferromanganese therefore depends on the degree of reduction of MnO by carbon as well as the prereduction that occurs in the upper region of the furnace. The ratio of CO and CO_2 in the off-gas is important and can be used to monitor the condition of the furnace. The higher the CO_2 content of the off-gas, the higher is the energy efficiency of the process, because the reducing potential of the gas is being more fully utilized (see Fig. 13) [42], [44]. Good operation of the furnace is indicated by a $CO_2/(CO_2 + CO)$ ratio of 0.55. This ratio, as well as the MnO content of the slag, can be used to control the coke rate of the furnace. Undercoking of the furnace is indicated by high MnO content of the slag and a low CO_2 content in the off-gas [42].

A further influence on the MnO content of the slag is the basicity ratio $(CaO + MgO) : SiO_2$. Generally the addition of basic oxides increases the melting point of the slag. The hotter slag improves the reaction between the slag and the coke and, consequently, more MnO is reduced. Increasing the basicity of the slag thus decreases the residual MnO content (see Fig. 8). The MnO content of the slag is also reduced by increasing the penetration of the electrodes, which also increases the slag temperature. The target MnO content of the slag depends on whether a discard- or high-slag process is used.

6.1.2.4. Discard-Slag Practice

Discard slags generally have MnO contents of 8 – 12 %. Slag is produced by the silica and other basic oxides entrained in the ore. Manganese(II) oxide is entrained in the silicate network and is released by addition of CaO or MgO. By increasing the basicity of the slag the recovery of manganese as metal is increased; however, consumption of carbon and electricity also increases. The discard-slag process is therefore only used where power is relatively cheap and the delivered cost of manganese ore is high [36]. The recovery of manganese to the metal is between 70 and 75 % when this practice is used. The practice usually involves the addition of limestone or dolomite to the furnace. However, in the Republic of South Africa where high proportions of Mamatwan ore are used, a basic slag is produced without the addition of limestone, due to the high CaO content of the ore. In India

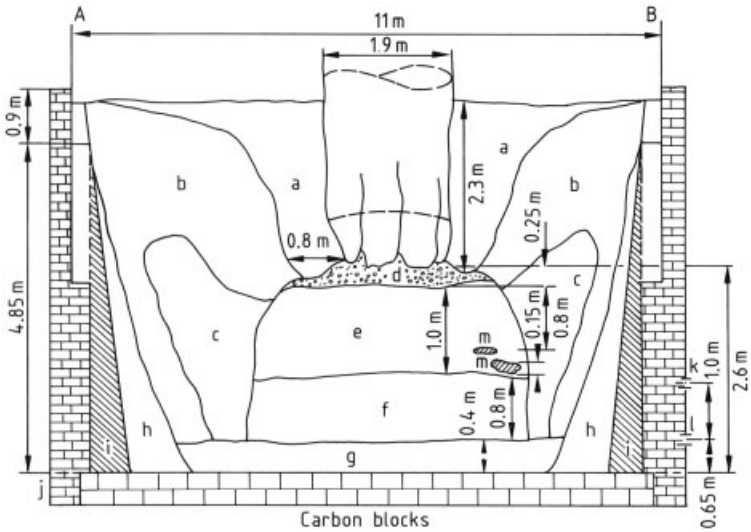

Figure 12. Zones in a ferromanganese furnace [35]
a) Loosely sintered burden; b) Loosely sintered material enriched in carbonaceous reducing agents; c) Coke and slag region, showing the active zone away from the electrode; d) Coke bed; e) Coke-enriched layer, with $CaO-MnO-SiO_2$ slag; f) MnO melt layer with some slag, coke, and additional carbonaceous reducing agent; g) Ferromanganese alloy layer intermixed with MnO melt; h) Graphitized and carbon-rich material; i) Carbon lining; j) Brick lining; k) Slag taphole level; l) Metal taphole level; m) Pieces of electrode

the discard-slag practice is used because high recoveries of manganese are necessary to produce 74 % Mn from lower grade ores.

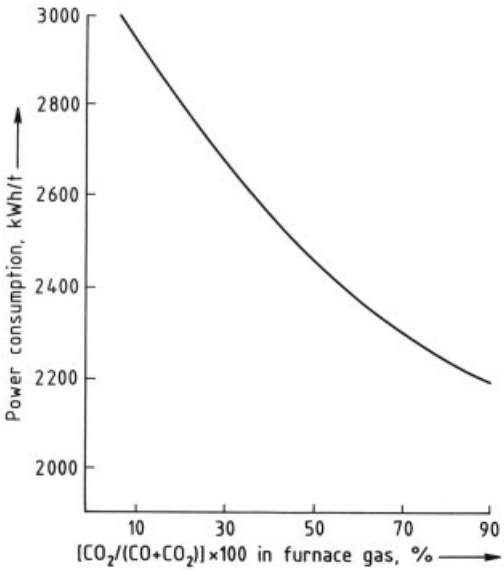

Figure 13. Relationship between off-gas composition and power consumption for ferromanganese production [42]

6.1.2.5. High-Slag Practice

In the high-slag practice less coke is required for reduction and little or no basic fluxes are used because the MnO content of the slag satisfies the requirement for basic oxides. Slags of this nature tend to contain more than 25 % MnO. In the high MnO practice the power consumption is reduced because a higher proportion of the reduction occurs by the gases and less MnO is reduced by solid carbon. Maximum use is therefore made of the exothermic nature of the prereduction reactions. The recovery of manganese is low in the ferromanganese furnace, but the overall recovery of manganese is high because the slag is usually used for the production of silicomanganese [36].

An additional attraction of the high-slag process is that an artificial ore, with a manganese to iron ratio of up to 100:1, can be made from relatively low-grade ores. This artificial ore can then be used to produce the highest grade of ferromanganese without the need to purchase costly high-grade ores. An additional benefit is the extremely low phosphorus content of the slag, which hence lowers the phosphorus content of any mix in which it is used. In countries

having only ores with high phosphorus contents, the first stage of the process is the production of a slag high in MnO and a low-manganese alloy [45].

Since the energy input in the production of ferromanganese by the high-slag process is lower than that of the discard-slag process, it is used by most producers. Slag compositions for various producers are shown in Table 12.

6.2. Production of Silicomanganese

Unlike ferromanganese, silicomanganese is only produced in electric arc furnaces, most of which can be used interchangeably to produce either of the manganese-containing alloys. Silicomanganese is used either as a substitute for ferromanganese and ferrosilicon in steelmaking or as a raw material for the production of medium- and low-carbon ferromanganese and industrial manganese metal. The composition of silicomanganese produced in the Western world is given in Table 6 and of that produced in the former Soviet Union in Table 13. Although silicomanganese generally contains 14 – 19 % Si, grades containing up to 35 % are produced for the production of extremely low-carbon alloys.

The solubility of carbon in the alloy decreases with increasing silicon content (Fig. 14) [46]. On cooling, sparingly soluble SiC comes out of solution.

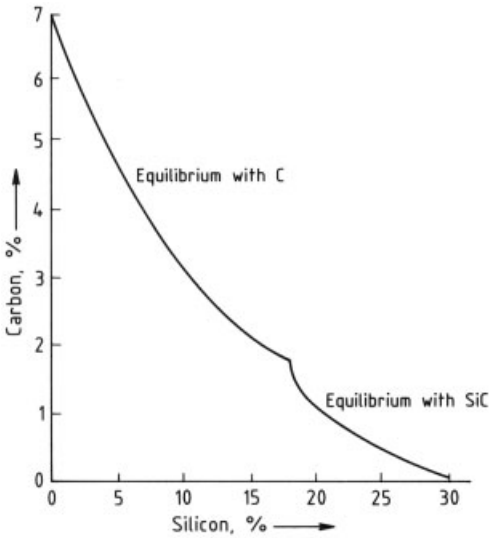

Figure 14. Carbon solubility in the Fe – Mn – Si – C system (50 – 80 % Mn) at 1420 °C [46]

There are three general routes for the production of silicomanganese:

1) Reduction of manganese ores and silica with coke and coal
2) Reduction of MnO-rich slags from ferromanganese production and quartzite with coke and coal
3) Reaction of standard ferromanganese and quartzite with coke

The first two processes are used for the production of alloys containing 15 – 25 % Si and can be carried out in the same furnaces used for ferromanganese production. The third method is used to produce alloys containing 30 – 35 % Si and is generally performed in smaller furnaces.

Raw Materials. The raw materials used in silicomanganese production are similar to those used in making ferromanganese. Silica is added to the furnace as quartz or quartzite, and ferromanganese slag can be used as an alternative additional source of manganese instead of manganese ore or sinter. As in the case of ferromanganese, production advantages can be gained by using sintered ore (particularly from Mamatwan). At the Kashima works of Chuo Denki, 40 % of the ore feed is in the form of sinter. The raw materials used to produce silicomanganese are given in Table 14.

Chemistry of the Process. In a dig out of a small electric arc furnace [47], four zones were distinguished: the burden zone, the zone of the coke bed, the melting zone, and the metal layer. In the burden zone, the higher oxides of manganese are reduced to MnO, while the higher oxides of iron are reduced to FeO and partially to metallic iron. Manganese(II) oxide is converted to complex silicates, which begin to melt at the bottom of the burden zone. Fine metallic particles exist in the coke bed zone, which are possibly caused by condensation of silicon and manganese in the hot areas under the electrodes. In the upper and lower parts of the melt zone, the reduction of manganese and silicon oxides occurs. The equilibrium is determined by the following reaction [30]:

$$SiO_2 + 2\,Mn \rightarrow Si + 2\,MnO$$

This equation is important in determining the silicon and manganese contents of the metal that

Table 13. Composition of former Soviet silicomanganese [26]

Grade and official grade code	Si,∗ %	Mn minimum, %	C, %	Deleterious elements, not to exceed		
				P, % (group A∗∗)	P, % (group B∗∗)	S, %
CMn26	26.0	60.0	0.2	0	0.05	0.03
CMn20	20.0 – 25.9	65.0	1.0	0.1	0.25	0.03
CMn17	17.0 – 19.9	65.0	1.7	0.1	0.35	0.03
CMn14	14.0 – 16.9	65.0	2.5	0.2	0.35	0.03
CMn10	10.0 – 13.9	65.0	3.5	0.2	0.35	0.03

∗ As reported.
∗∗ Furnaces.

Table 14. Operating parameters for silicomanganese production

	Elkem Marietta	Samancor M 12	Samancor M 13
Raw material additions, tonnes/tonne alloy			
Manganese ore: Groote Eylandt	0.227		
South Africa: medium grade		0.536	0.942
Mamatwan ore		0.647	
Mamatwan sinter			0.754
Ferromanganese slag	0.227		
Remelt metal	0.091		
Reductants: coke	0.141	0.198	
coal		0.566	0.836
Fluxes: limestone			
magnesite			
quartz	0.127	0.610	0.814
Operating results			
Average operating load, MW	28	46	11
Operating time, %	94	97	98
Specific energy consumption, kWh/t	4400	3870	3900
Manganese recovery, %	85	74	67
Slag composition, %			
MnO		78.0	15.0
MgO		3.9	4.5
Al_2O_3	29.7	7.7	6.1
CaO		20.9	21.1
SiO_2	33.8	48.5	47.0
Metal analysis, %			
Mn	65.9	67.7	65.5
Fe		16.0	15.5
Si	19.1	17.5	17.8
P		0.09	0.09
C	1.38	1.18	1.15
Ratios			
Slag/metal ratio		0.72	1.17
Basicity ratio	1.0	0.54	0.55

collects in the lowest part of the furnace and is influenced by the slag chemistry and the temperature of the process. Increasing the CaO content of the slag reduces the silicon content of the metal. The basicity requirement of the slag is therefore better supplied in the form of MgO. Higher temperatures tend to drive Si into the metal at the expense of Mn. Higher temperatures are therefore required in the production of silicomanganese than in the production of ferromanganese.

Silicomanganese is produced by most manufacturers of manganese alloys, and the slag and metal compositions and operating parameters of some producers are given in Table 14.

To produce manganese metal of 97 % Mn by the silicothermic method (see Section 7.4) a silicomanganese containing 28 % Si is required that is particularly low in phosphorus and iron. This

is made from a manganese slag produced from the partial reduction of manganese ore (see Section 6.1.2.5).

Operation of the Furnace. The operation of the furnace for silicomanganese production is similar to that of ferromanganese production. However, deeper penetration of the electrodes is necessary to provide the higher temperature required to drive the reduction of silicon [30]. The resistivity of the burden is therefore important, and the size and the activity of the reductant is critical for stable operation of the furnace.

6.3. Production of Medium-Carbon Ferromanganese

Medium-carbon ferromanganese contains 1 – 1.5 % carbon and has a manganese content of 75 – 85 % (see Table 6). Medium-carbon ferromanganese can be produced either by refining high-carbon ferromanganese with oxygen or by the silicothermic route, whereby the silicon in silicomanganese is used to reduce additional MnO added as ore or slag. The former process has considerable advantages [46] and is used by most producers. Elkem Metals in Norway and the United States use their patented manganese oxygen refining (MOR) process to refine high-carbon ferromanganese in plants having capacities of 5000 t/month and have closed their silicothermic plant in the United States. A similar facility was sold under license to the Torros plant of Minera Autlan in Mexico. Thyssen Stahl (Federal Republic of Germany) uses a similar process to produce medium-carbon ferromanganese from the high-carbon ferromanganese produced in their blast furnaces. Samancor (Republic of South Africa) and Usinor Sacilor (France) have patented a similar process, but have not yet built a plant [48].

Transalloys (Republic of South Africa) produces 20 000 t/a of medium-carbon ferromanganese by the silicothermic route and appears to be the only producer in the Western world to do so. In the former Soviet Union, medium- and low-carbon ferromanganese are produced silicothermically with manganese recoveries of 59 – 63 % [26].

The silicothermic and MOR processes are shown schematically in Figure 15.

6.3.1. Production of Medium-Carbon Ferromanganese by Oxygen Refining of High-Carbon Ferromanganese

In the MOR process patented by Union Carbide, high-carbon ferromanganese is decarburized in a similar manner to the steelmaking process in the basic oxygen furnace [46]. However, several distinctive differences are encountered in the case of ferromanganese:

1) A final temperature of 1750 °C compared to 1550 °C
2) Refractory attack is more severe
3) Difficult casting of the final alloy
4) The higher vapor pressure of manganese
5) The higher volume and temperature of the off-gas

In the MOR process, oxygen is blown into the molten high-carbon ferromanganese and the temperature is increased from its tapping value of 1300 to 1750 °C. The heat required is supplied by the oxidation of manganese to manganese(II) oxide and carbon to carbon monoxide. The need to increase the temperature is shown by the carbon – temperature relationship in Figure 16. In the early part of the blowing process, most of the oxygen is consumed by oxidation of manganese, and the temperature of the melt increases from 1300 to 1550 °C. Hereafter, carbon is rapidly oxidized and the temperature rises to 1650 °C. Above this temperature, the rate of carbon removal decreases and manganese is once again oxidized. The process is stopped at 1750 °C, which corresponds to a carbon content of 1.3 %. Further reductions in carbon content result in unacceptably high losses of manganese. In the MOR process, the recovery of manganese is ca. 80 % and the distribution of manganese can be broken down as follows [46]:

Alloy MC FeMn	80 %
Fume formed by vaporization	13 %
Slag formed by oxidation of Mn	5 %
Other losses, splashing	2 %

The manganese lost in the fume is recovered in the gas cleaning plant and is then pelletized and returned to the high-carbon ferromanganese furnace. The slag, which contains about 65 %

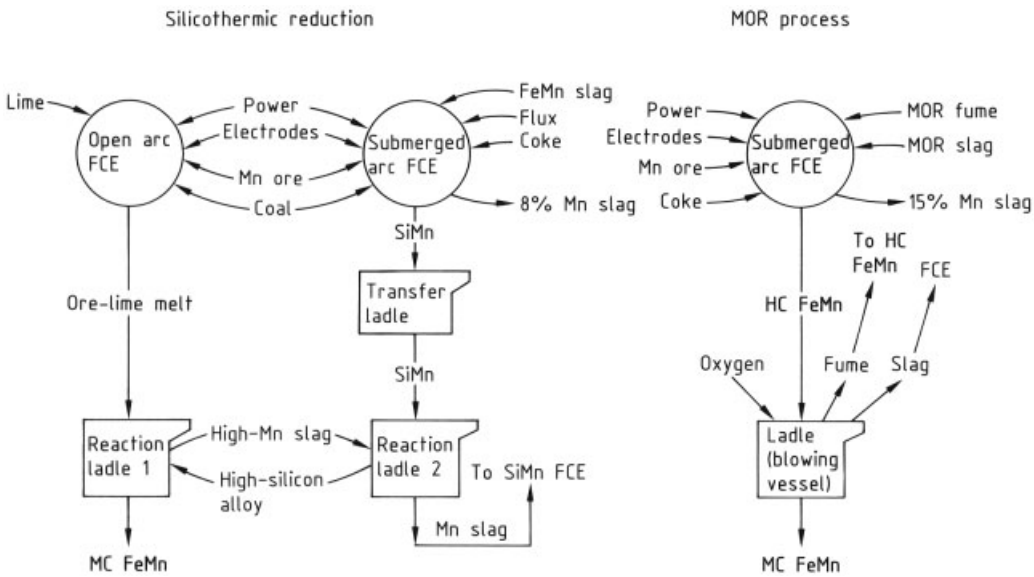

Figure 15. Process flow sheet comparison for silicothermic reduction and the MOR process [46]

MnO, is also returned to the high-carbon ferromanganese furnace. The successful operation of this process depends on the design of the blowing vessel and oxygen lance as well as giving careful attention to operational procedures. In the joint patent of Samancor and Usinor Sacilor, a bottom blown converter is used and steam is injected at the end of the blow as a coolant [48].

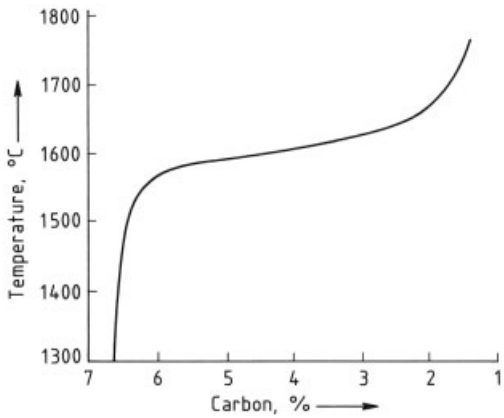

Figure 16. Dependence of carbon content on temperature for a ferromanganese alloy (Mn : Fe, 6 : 1) at 101.3 kPa [46]

The MOR process has many advantages over the silicothermic process: lower energy consumption, lower capital investment, lower production cost, and greater flexibility.

The main disadvantage of the process is that its use is limited to production of medium-carbon ferromanganese because the carbon content cannot be reduced to below 1.3 %.

6.3.2. Silicothermic Production of Medium-Carbon Ferromanganese

In the silicothermic production of medium-carbon ferromanganese, a high-grade slag or a melt containing manganese ore and lime is contacted with silicomanganese containing 16 – 30.1 % silicon [49]. The silicon in the alloy acts as the reducing agent in the process, which reduces the manganese(II) oxide in the melt. Similarly to silicomanganese production, the equilibrium is determined by the reaction:

$$Si + 2\,MnO \rightleftharpoons SiO_2 + 2\,Mn$$

The purpose of the lime is to reduce the activity of the SiO_2 in the melt, thus forcing the above reaction as far to the right as possible. The ratio of CaO to SiO_2 in the slag should be greater than 1.4 to ensure a sufficient reduction in the activity of SiO_2. The carbon entering the process in the silicomanganese remains entirely in the metal phase and is therefore found in the product. Thus, to produce a medium-carbon ferro-

manganese containing 1 % C, a silicomanganese containing 20 % Si is necessary (see Fig. 14).

The heat produced by the silicothermic reduction is not sufficient to sustain the process; hence it is usually carried out in an electric arc furnace. These furnaces are usually small and, unlike ferromanganese furnaces, are lined with magnesite bricks, which are fairly resistant to the highly basic slag. The power consumption is between 1000 and 3000 kW. These furnaces can be tilted so that the slag can be separated from the metal. At Transalloys (Republic of South Africa) the process is carried out in a mixing ladle.

Although the silicothermic reduction process is more energy intensive than the decarburization of high-carbon ferromanganese, it has the advantage that the final carbon content is limited only by the carbon content of the initial silicomanganese. The silicothermic process can therefore be used to produce low-carbon ferromanganese and industrial manganese metal (see Section 7.4).

6.4. Production of Low-Carbon Ferromanganese

Low-carbon ferromanganese contains 76 – 92 % Mn and 0.5 – 0.75 % C (see Table 6). The production of low-carbon silicomanganese is not possible by the decarburization of high-carbon ferromanganese without incurring extremely high losses of manganese [49]. Use must accordingly be made of a silicothermic reduction process.

The process is similar to that used in the silicothermic production of medium-carbon ferromanganese. High-purity ores are used and in particular ores containing iron and phosphorus should be avoided. An artificial manganese ore, produced as a high-grade slag (see Section 6.1.2.5) is particularly suitable because of its low impurity level and because all the manganese is present as MnO. The reduction of the higher oxides of manganese is therefore unnecessary.

The operating figures for 1 t of ferromanganese containing 85 – 92 % Mn, 0.1 % C, and 1 % Si with a manganese recovery of 75 % are:

Calcined manganese ore	1250 – 1350 kg
Silicomanganese (32 – 33 % Si)	800 – 850 kg
Quicklime	1000 – 1100 kg
Electrodes	10 – 12 kg
Electricity	1800 – 2500 kWh

Since the required silicon content of the metal is low, a slag high in MnO is necessary. The MnO content of the slag can, however, be reduced by the use of a two-stage refining operation. In the first stage, an excess of silicomanganese is maintained and a slag containing 6 – 8 % Mn is teemed and discarded. The second refining stage with manganese ore and lime results in a slag containing 10 – 14 % Mn, which is returned to the silicomanganese furnace.

6.5. Gas Cleaning

In all the processes described in this chapter, large volumes of gas are generated. These gases consist generally of CO, CO_2, and N_2, and contain large quantities of dust from the raw materials and condensed manganese droplets. These gases require cleaning prior to their venting to the atmosphere. In open furnaces, in which the gas is totally combusted, hoods are incorporated in the design through which the gas is removed. This gas is cooled in trombone coolers to ca. 200 °C and the dust is removed in bag filters. Ideally, after recovery the dust should be agglomerated and returned to the furnace. In closed furnaces, the gas is usually cleaned in cyclones and venturi scrubbers prior to combustion to the atmosphere or used in downstream processes. At Elkem Sauda in Norway, gas from the 36 MVA furnace is cleaned by two wet venturi scrubbers. The gas has an initial dust loading of 150 g/m^3 which is reduced to a maximum of 50 mg/m^3 [37]. The dust recovered from the cleaning plant is filtered, sintered and returned to the process. At Samancor's Meyerton works in South Africa the gas from the 81 MVA and two 75 MVA closed furnaces is cleaned by a cyclone followed by two venturi scrubbers. A mist eliminator is included in the gas plant of the 81 MVA furnace to reduce the moisture content because the gas is used for heating in other areas of the plant. The emissions from the gas plants are less than 50 mg/m^3, which is the statutory maximum. The gas from the remaining open furnaces is cleaned in a bag filter plant. Investigations are under way to briquette the dust from this plant

due to its high manganese content and high manganese to iron ratio.

It is expected that dust limits will be reduced to 25 mg/m^3. In addition to this the exposure limits allowed in the working environment in manganese producing facilities have been reduced due to the toxicity of manganese dust; the following limits apply:

United States, United Kingdom, Australia, Belgium, Brazil, Federal Republic of Germany	5 mg/m^3
Yugoslavia	2 mg/m^3
former Soviet Union, Poland	0.3 mg/m^3
Bulgaria	0.02 mg/m^3

6.6. Developments and Future Trends

Energy Saving Measures. In regions with high electricity prices developments have concentrated on the saving of energy. This is of particular importance to Japanese producers of manganese alloys that use electric furnaces. Developments include the use of the off-gas from the furnace to preheat and mildly prereduce the ore, either in a rotary kiln or in a shaft kiln above the furnace.

The former process is used at the Kashima Works of Chuo Denki, where the feed to the 40 MVA ferromanganese furnace is heated to 950 °C in a rotary kiln and the higher oxides of the manganese ore are partially reduced. The ore loses 23 % of its mass during this process [50]. The use of this system has resulted in savings in power, and an additional advantage is that coal can be used instead of coke to satisfy the carbon requirement of the furnace. Care must be taken to avoid rapid heating and consequent decrepitation of the ore in the kiln.

The process in which a shaft kiln is positioned above the furnace was invented by Nasu [51]. In this process the gas from the furnace is burnt under the furnace roof and the hot burnt gas leaves the furnace through a vertical shaft. The raw ore mix is introduced to the top of the shaft and is heated in a countercurrent fashion by the exhaust gas. The ore is then fed to the furnace, where further heating takes place as the ore is exposed to the radiant heat produced by the burning gas in the roof. Two furnaces of this type are in operation at the Mizushima plant and their operation has resulted in power savings.

Computer Control of Electric Furnaces. Computer control of electric furnaces is practiced by Elkem at Suada in Norway, at Marietta in Ohio, and at Beauharnois in Canada. Temco in Tasmania and Samancor in South Africa also use computer control [52]. The following improvements in operating parameters have been realized by computer control at Elkem's plant at Beauharnois:

Operating power level (MV)	5 % improvement
Power cost reduction from increased load	3 % improvement
Manpower cost reduction	9 % improvement
Statistical process control	lower variation in product

Elkem is investigating the potential of enhancing their process control system by the incorporation of an expert system. This system, based on artificial intelligence, will incorporate the experience of the operator in the process control system and will be used to diagnose reasons for poor furnace performance. Possible changes to improve the operation will be suggested by the expert system [53].

Use of Plasma Furnaces. Non-transferred arc plasma furnaces have been used successfully in the production of charge chrome. Attempts to produce ferromanganese in plasma furnaces have not been economically successful due to the high losses of manganese caused by volatilization in the arc attachment zone. In addition, no preheating of the ore takes place in a plasma furnace, unlike in the burden zone of the submerged arc furnace. In an effort to solve these problems and to utilize the higher specific throughputs that are obtainable in a plasma unit, a combined plasma/shaft furnace is being jointly developed by Voest Alpina and Samancor [54].

7. Production of Manganese Metal

[55], [56]

Pure manganese metal has only become an important commodity since ca. 1940. Today the world market is in excess of 140×10^3 t/a. Most

of the high-purity manganese produced is consumed by the steel and aluminum industries, although the use of manganese for the production of manganese – zinc soft ferrites has led to a demand by electronic component producers [57]. Whilst ferromanganese remains an economically sound additive for manganese steel producers, greater volumes of pure manganese metal are used in the formulation of specialty steels.

In aluminum production, manganese powder is added to the melt to make the product harder. Manganese powder has long been used in welding-rod fluxes, where it imparts superior flow properties to the melt forming the weld.

Producers of manganese chemicals enjoy the convenience of a pure, reactive starting material from which to manufacture manganese salts and compounds.

A number of methods for the production of relatively pure manganese metal are known:

1) Electrolysis of aqueous manganese salts
2) Electrothermal decomposition of manganese ores
3) Silicothermic reduction of manganese ores or slags
4) Aluminothermic reduction of manganese ores or slags
5) Distillation of ferromanganese
6) Electrolysis of fused salts

Techniques (1) and (2) are the most important; methods (3) – (6) are not of economic significance.

7.1. Electrolysis of Aqueous Manganese Salts

The electrolytic route to manganese metal was first investigated in 1930 by Davis. However, the process did not become significant until 1939 when demand for manganese by the steel producers (for armaments manufacture) prompted the U.S. Bureau of Mines to build a 1 t/a pilot plant at Knoxville, Tennessee. The plant was redesigned in 1940, and by 1944 it was producing ca. 1500 t/a. Chuo Denki Kogio of Japan commenced production in 1941. The U.S. Bureau of Mines built a second pilot plant at Boulder City in 1942 [58], and The Electrolytic Manganese Corporation of Krugersdorp, Republic of South Africa began production in 1955.

The Boulder City process forms the basis of modern electrolytic manganese plant operation and is encountered in many parts of the world. Production capacities according to country and company are listed in Table 15.

Table 15. Estimated production capacity for electrolytic manganese metal, 1988 [90]

Company	Country	Capacity, t
Fermavi Industria e Comercio	Brazil	80
Chuo Denki Kogio	Japan	3600
Tosoh	Japan	6000
Manganese Metal Company	RSA	43300
Elkem Metals	United States	10000
Kerr-McGee Chemicals	United States	11250
State organization	former Soviet Union	3000
Total		77230

The Manganese Metal Company of South Africa is currently the largest supplier but high prices in the late 1980s have encouraged investment by other countries.

The process may be conveniently divided into 3 stages:

1) Milling and chemical reduction of a suitable ore
2) Leaching of the soluble manganese and removal of deleterious solutes
3) Electrowinning of manganese metal

Ore Milling and Reduction. To ensure reactivity in both the reduction and leaching steps, the ore is milled to ca. 80 – 90 % $< 75\,\mu m$ [59]. The surface area (BET) is 2 – 5 m^2/g. It is usual to dry mill the ore since the hydrometallurgical part of the process is closed and unnecessary dilution of the liquor must be avoided. Ball or vertical spindle mills are suitable for this purpose. Transport of the ore may be effected pneumatically at this stage.

Manganese ores usually occur in an insoluble form comprising minerals largely in the 4 + oxidation state. Carbonate ores, which are readily soluble, occur in some parts of the world. To render refractory ores soluble in acid, the manganese is usually converted to the 2 + oxidation state by calcining under a reducing atmosphere at 850 – 1000 °C. Reductants are carbon sources

and may include various combinations of anthracite, coal, charcoal, hydrocarbon oils, natural gas, coal gas, or liquified petroleum gas. The reduction reactions can be simplified to:

$$CO + MnO_2 \rightarrow CO_2 + MnO$$

$$CO_2 + C \rightarrow 2\,CO$$

Other species present are also reduced. Iron is reduced to the iron(II) state and in some cases to iron metal. The Fe(II) content of the calcined ore may be used as an indicator of the degree of reduction achieved; a calcine containing 1 – 3 % Fe(II) can be assumed to have its manganese fully reduced to Mn(II).

The discharging calcined ore must be cooled to below 100 °C to exclude the possibility of reoxidation. The reactive nature of the calcine at this stage precludes pneumatic transport.

Reductive leaching has been investigated with bacteria [60] and with reducing chemicals such as sulfur dioxide, hydroquinone, and other organic compounds [61].

Leaching and Purification. The reduced ore is agitated with spent electrolyte which normally contains 2 – 5 % sulfuric acid (generated at the diaphragm in the electrolysis stage). Other soluble species, notably iron(II), are also dissolved and must be removed prior to electrolysis. The pH of the leachate should be 5 – 6, and air oxidation of Fe^{2+} to Fe^{3+} ensures the precipitation of iron(III) hydroxide to remove soluble iron [62]. Other deleterious transition metals are removed by precipitation of their sulfides through pH manipulation and addition of a soluble sulfide [63]. Some manganese sulfide is necessarily coprecipitated during this process and the relative amount of this present in the sulfide precipitate is a measure of the efficiency of the reaction [64–67]. The manganese concentration at this point is ca. 10 % expressed as the sulfate. If a substantial amount of soluble silica is present, it may be removed by coprecipitation with aluminum hydroxide. A solid – liquid separation is required to provide a perfectly clean solution for the electrolytic stage. Other methods of purification include ion exchange and solvent extraction [68], [69].

Electrowinning of Manganese Metal. Although manganese has a strongly electronegative electrode potential (– 1.13 V), its high hydrogen overvoltage allows electrowinning from aqueous solutions. Good current efficiencies are only obtained when other metallic solutes are absent from the catholyte.

The manganese is recovered in a diaphragm cell. The cathode may be made of titanium, stainless steel, or one of a number of proprietary alloys. The anode is usually a lead – calcium or lead – silver alloy, but success has recently been achieved with anodes coated with platinum-group metals. Diaphragms are usually made of synthetic polymer canvas.

Prior to introduction to the cell, the solution is conditioned by addition of sulfur dioxide or, more recently, selenium dioxide [70]. These additives improve current efficiencies and ensure the deposition of the hard and brittle α allotrope of manganese. It has been reported that the current efficiencies have been increased to as much as 90 % with selenium dioxide. However, the latter has the disadvantage that the metal obtained is often contaminated with selenium. This reduces the total manganese content of the metal and causes toxicological and environmental problems for the consumer. Addition of sulfur dioxide produces a metal with slightly higher sulfur content. Ammonium sulfate is used in the cell solution as a supporting electrolyte.

Typical cell operating conditions at the cathode are as follows:

$MnSO_4$ concentration	4.0 %
$(NH_4)_2SO_4$ concentration	13 %
pH	7.0
Temperature	35 – 45 °C
Current densities:	
Anode	1000 A/m^2
Cathode	500 A/m^2
Potential	5 – 7 V

These values are approximate and vary considerably between installations. Current efficiencies are variable, depending on additives, the purity of the catholyte, and optimization of solute concentrations. Values between 42 and 62 % are typical.

The following reactions occur [71]:
At the cathode

$$Mn^{2+} + 2\,e^- \rightarrow Mn^0$$
$$2\,H_2O + 2\,e^- \rightarrow H_2 + 2\,OH^-$$

At the diaphragm

$$(SO_4)^{2-} + 2\,H^+ \rightarrow H_2SO_4$$

At the anode

$$2\,H_2O \rightarrow 4\,H^+ + 4\,e^- + O_2$$
$$2\,e^- + Mn^{2+} + O_2 \rightarrow MnO_2$$

The electrolysis of water is largely responsible for the inefficiency of the electrolytic procedure. Power consumption is typically 9000–12 000 kWh/t. Polymeric smoothing agents may be added to the electrolyte to prevent the formation of dendrites and nodules on the cathodes, which usually cause the metal to be of inferior quality. Deposition time is dependent on current densities and varies between 12 and 30 h. Current efficiencies tail off rapidly as the thickness of the metal increases, and a compromise must be made between high efficiencies and the production of the thick metal flake required by the market. A thickness of 1.5–2.0 mm is considered optimum.

Metal Treatment. The cathodes are removed from the electrolysis bath after the appropriate deposition period. Since the α-manganese deposited is hard and brittle, it is readily stripped from the cathode starter-sheet by mechanical shock. The metal is then washed and dried. Absorbed hydrogen is removed by heating in air or under an inert atmosphere. The latter is necessary when manganese with a particularly low oxygen content is required. Metal destined to be milled must have hydrogen content $< 0.0005\,\%$ to minimize the explosion risk. A primary hydrogen explosion can initiate a secondary manganese dust explosion.

The metal is sold in flake, powder, or granular forms, depending on the end use. A generalization is that the steel industry requires thick flake, while the aluminum industry, welding-rod manufacturers, and manganese chemical producers require powders of various particle size distributions.

Other forms of manganese are also encountered; for example, powder and flake manganese which has had its surface passivated by mild oxidation. A market exists for briquettes of manganese and iron, or manganese and aluminum, which are conveniently added to melts of the appropriate metal to produce a particular alloy.

Quality Control and Analysis. Electrolytic manganese metal has a purity of $> 99.5\,\%$; no reliable method exists to determine manganese content directly. The sum of the impurities is therefore deducted from 100 % to obtain the assay of the metal.

A typical analysis (%) is:

Manganese	99.8
Oxygen	0.2
Nitrogen	0.005
Sulfur	0.02
Carbon	0.001
Iron	0.001
Copper	0.001
Cobalt	0.001
Nickel	0.001
Selenium	0.0005
Hydrogen	0.0006

When selenium dioxide is used as a cell solution additive, typical values (%) are:

Manganese	99.4
Sulfur	0.01
Selenium	0.3

Analytical methods described in this section are applicable to all types of manganese metal.

In large amounts, manganese is best determined by titration and many methods exist. The most reliable is the titration of the divalent ion with permanganate in a neutral pyrophosphate buffer that stabilizes the trivalent manganese species. A platinum redox indicator electrode is used with a suitable reference electrode. A microprocessor controlled automatic titrator facilitates precision and accuracy. The reaction can be summarized as:

$$4Mn^{2+} + MnO_4^- + 8\,H^+ + 15\,H_2P_2O_7$$
$$\rightarrow 5\,Mn(H_2P_2O_7)_3^{3+} + 4\,H_2O$$

The equivalence potential is ca. 0.47 V versus the saturated calomel electrode.

An alternative is to oxidize the manganese to permanganate with excess sodium bismuthate in the presence of nitric acid, add an excess of standard iron(II) solution and back-titrate with permanganate. This method has the advantage of not requiring sophisticated end-point detection equipment but suffers from a number of interferences. Variations on this oxidation/titration method include silver-catalyzed oxidation by persulfate followed by an arsenate titration [72].

Metallic impurities in both metal and process liquors are usually determined by absorption or emission spectroscopy, though various voltammetric techniques have been used to determine traces in pure metals. The half wave potential of manganese is − 1.57 V versus the saturated calomel electrode, making anodic stripping voltammetry ideally suitable for trace metal determination .

Nonmetallic impurities are universally determined by the use of commercially available equipment. The metal is usually heated above the melting point, in a gas stream, releasing the sought element in the form required. This is then determined in the gas phase by infrared absorption or thermal conductivity detectors. Nitrogen in greater than trace amounts has been determined by acid digestion, distillation, and colorimetric determination of ammonia (cf. Kjeldahl method for ammonia).

Summary. The electrolysis of aqueous solutions is the most important manufacturing route at present to high-purity manganese. Economic efficiencies are dependent on scrupulously clean solutions and strict control of operating parameters. The metal produced may be > 99.7 % pure and has many varied uses.

Manganese has also been produced by the electrolysis of aqueous solutions at a flowing mercury cathode. This process, which operates at high current efficiencies, can deliver the manganese in a variety of forms, and is less dependent on solution purities. High electrolysis temperatures, and the inconvenience of coping with toxic mercury are possible reasons why there are now no full-scale plants using this process.

7.2. Electrothermal Process

This practice is a refinement of the silicothermic reduction of manganese (see below). However, the purity of the product obtained from the electrothermal process is superior.

The production of bulk manganese by this method is a well protected secret and little information appears in the literature. It is essentially a multistage hot refining process based on the following principles. Manganese ore is smelted in an arc furnace with a limited amount of reductant. This preferentially reduces the iron to metal with a small amount of manganese to form a low-grade ferromanganese with a high phosphorus content. The slag is thus refined to a purified manganese starting material.

The slag is used to produce a high-silicon silicomanganese (> 30 % Si) to ensure a low carbon content. Other ingredients such as quartz, limestone, and coke must be selected carefully to ensure a low contamination level. Coke, in particular, may have to be produced from special low-phosphorus coals.

The liquid silicomanganese is treated in two steps, with a liquid basic slag made from the purified manganese starting material. The operation is carried out in a ladle or a shaking ladle.

Typical analyses of electrothermal manganese are given in Table 16. From these analyses it can be seen that electrothermic manganese is a relatively pure product that competes in some areas with electrolytic manganese. The sulfur content is particularly low. The aluminum industry is experimenting the use of these products as alternatives to electrolytic manganese.

7.3. Aluminothermic Process

This production route is based on the well known thermite reaction in which a reactive metal liberates a less reactive metal from its oxide with the evolution of heat. The reactions possible for manganese oxides are:

$$3/4\,MnO_2 + Al \rightarrow 3/4\,Mn + 1/2\,Al_2O_3$$

$$\Delta H = -447.2\ \text{kJ/mol Al}$$

$$1/2\,Mn_2O_3 + Al \rightarrow Mn + 1/2\,Al_2O_3$$

$$\Delta H = -357.1\ \text{kJ/mol Al}$$

$$3/8\,Mn_3O_4 + Al \rightarrow 9/8\,Mn + 1/2\,Al_2O_3$$

$$\Delta H = -316.5\ \text{kJ/mol Al}$$

$$3/2\,MnO + Al \rightarrow 3/2\,Mn + 1/2\,Al_2O_3$$

$$\Delta H = -259.6\ \text{kJ/mol Al}$$

A heat of reaction of between 272 and 314 kJ/mol aluminum is necessary to sustain the reaction and ensure that it is not explosive. Therefore, some ores may require modification of the relative proportions of the different manganese oxides prior to reaction. This is achieved

Table 16. Typical analyses of electrothermal manganese ∗

	Gimel A	Gimel F	Gimel BF	Gimel LP
Mn (% min)	96	97	98	97
C (% max)	0.050	0.030	0.030	0.030
Si (% max)	0.50	0.30	0.30	0.30
P (% max)	0.050	0.030	0.030	0.020
S (% max)	0.008	0.008	0.008	0.008
Fe (% max)	3.0	2.0	1.5	2.0
H (ppm)	5	5	5	5
N (max)	0.035	0.035	0.035	0.035

∗ Gimel is a registered trade mark of Pechiney Electrometallurgie

by calcination at ca. 1000 °C. Certain circumstances require reduction of the ore by calcining in a reducing atmosphere (see page 30).

Calcium oxide is added to the mix to improve the flow properties of the slag, ensuring good separation of slag and metal. Lime also decreases the amount of silica reporting to the metal phase.

The aluminothermic route to manganese production is attractive in that it requires simple technology which is well documented, and low capital investment. Its disadvantages are that aluminum is expensive, and the exhaust gases from the process carry manganese and aluminum dust into the atmosphere. Efficiencies of up to 87 % have been reported.

7.4. Silicothermic Process

The production of manganese by silicothermic reduction resembles that of low-carbon ferromanganese in electric arc furnaces of 1 – 3 MW power consumption [73], [74]. The raw material used is either low-iron manganese ore or a manganese slag concentrate. The silicon is in the form of a special silicomanganese with ca. 33 % Si and < 3 % Fe. It is made from a manganese slag concentrate or from pure ores. Reduction as complete as possible is ensured by the addition of lime, which forms a basic slag.

Many slag changes are carried out during the process, as in the silicothermic manufacture of low-carbon ferromanganese. The refined metal has the following typical percentage composition:

Manganese	93 – 97
Carbon	< 0.10
Silicon	1.5
Phosphorus	< 0.06

The system is moderately efficient and the manganese yield is 63 – 64 %, with slag losses accounting for the remainder.

7.5. Distillation of Ferromanganese

Pechiney has filed patents dealing with the vacuum distillation of ferromanganese at 1200 – 1350 °C. The process has not been used on a large-scale since the manganese produced is inferior to electrolytic manganese and the production costs are prohibitive.

7.6. Electrolysis of Fused Salts

Manganese has been produced on a laboratory scale by the electrolysis of fused salts. A system in which an untreated manganese ore is dissolved in a fused mixture of calcium and manganese fluorides and electrolyzed at 1150 – 1200 °C has been described [75]. The cell potential was 5.5 – 6.5 V and current density was 9 – 11 A/cm^2 at the anode, and 2 – 2.5 A/cm^2 at the cathode. Manganese purity in the range 80 – 99.3 % was obtained at current efficiencies of 53 – 65 %.

8. Manganese as an Alloying Component [24]

8.1. Manganese in the Iron and Steel Industry

More than 90 % of the manganese mined is used in the steel industry (see Fig. 6). In the production of normal mild steel, high-carbon ferromanganese and silicomanganese are used; however, in special low-carbon steels, medium-carbon ferromanganese, low-carbon ferromanganese or electrolytic manganese is used. A steel producer has to weigh up the additional costs of the refined manganese alloys against the extra time in the argon oxygen decarburization (AOD) process that would be necessitated by the use of carbon-containing alloys [50]. Manganese occupies a special place among all alloying elements as it is present in almost all types of steel and cast iron.

The addition of manganese to iron – carbon systems has the following effects:

1) It retards the rate of transformation (i.e., it increases the hardenability).
2) It lowers the martensitic start temperature. This effect is 40 °C at 1 % addition of manganese and increases linearly to 200 °C with 5 % addition.

Manganese has extensive solid solubility in austenitic iron; therefore, it has no solution hardening effect in austenite. The solubility of manganese in ferritic iron is limited to 3 %. Hence, manganese exhibits a solution hardening effect in ferrite (30 – 40 MPa per wt % Mn).

Hadfield's austenitic manganese steels (which were developed from 1882) contain 1 – 1.4 % carbon and 10 – 14 % manganese. When they are quenched from 1000 °C the normal hardening transformation temperature is suppressed by the manganese and the steel remains austenitic at room temperature. These steels are exceptionally tough, have high tensile strength, and are wear resistant due to work hardening. As a consequence, they are used extensively in the mining and earth moving industries.

Manganese is a milder deoxidant than silicon and aluminum. Manganese enhances the effectiveness of aluminum and silicon as deoxidizers of steel due to the formation of stable manganese silicates and aluminates.

The solubility of nitrogen in ferrite is increased by the presence of manganese. The solution hardening effect of manganese is considerable enhanced by nitrogen, which is valuable in high-temperature creep-resistant alloys. Both the yield and ultimate stress at elevated temperatures are increased by the presence of nitrogen in steel, and this effect is enhanced by the presence of manganese.

Nitrogen is a powerful austenizing agent in steels and can be used as a partial substitute for nickel in austenitic stainless steels. The solubility of nitrogen in stainless steels is increased by the addition of manganese, and this is the basis of the series 200 stainless steels.

Manganese has had a long association with the efforts of steel makers to remove sulfur and to minimize the harmful effects of residual sulfur in steel. In the absence of sufficient manganese, sulfur can concentrate in the interstitial liquid in the form of ferrous sulfide. This compound has a low melting point and can persist as a liquid down to low working temperatures. This can cause cracking (hot shortness) during steel rolling. Steel requires a manganese to sulfur ratio of 4 for adequate ductility in the hot working range (950 – 1150 °C). In practice, ratios of up to 20 are used.

One of the most significant properties of steel is its ability to harden with heat treatment. A number of alloying elements can influence the hardenability of steels but manganese has become virtually synonymous with hardenability due to its effectiveness, availability, and relatively low cost. Increasing the manganese content permits the use of lower heat treating temperatures for steel parts where the use of higher temperatures would result in deformation.

Shortages, notably during the Korean War, originally prompted the use of manganese as a replacement for some of nickel in austenitic stainless steels.

In the AISI 200 series, about half, and sometimes more, of the nickel in a more conventional 18/8 steel is replaced by a combination of manganese and nitrogen. AISI 201 is at present the flagship of the series. Type 205 should be receiving increasing attention.

During the depressed price of nickel at below $ 6/kg until late in 1987, the price of type 201 in the United States remained lower than that of

type 301 or 304, and thus type 201 continued to be produced and used in the face of the oversupply of nickel. There is a price advantage inherent in series 200, which could probably benefit from further research and development. For example, a type 20 X, which contains some copper in addition to manganese, has been proposed as a direct challenge to type 304.

Technical problems previously associated with some manganese-bearing stainless steels (refractory wear, nitrogen introduction, continuous casting, rolling) have been largely overcome.

Numerous proprietary steels, some covered by patents and trade names, have been developed. The feature of these steels is not so much the saving in nickel, but their high strength and austenite stability under stress at low temperatures. Also, their corrosion, wear, galling, and welding properties are enhanced to various degrees.

The new version of type 201, which has a maximum carbon content of 0.03 %, may have an advantage over standard type 301 in applications involving welding and subsequent exposure to intergranular or stress-corrosion environments. The composition of the various series 200 stainless steels is given in Table 17.

Table 17. Nominal compositions of manganese steels, %

	201 a	201 b	205	20 X
C	0.05	0.03	0.15	0.1
Mn	6.75	6.50	15.0	7.0
Ni	4.0	3.75	1.0	3.0
Cr	16.75	16.75	17.25	14.75
Si	0.5	0.3	0.3	0.3
N	0.15	0.25	0.35	0.08
Cu				2.5

High-manganese stainless steels (e.g., 22 % manganese, 13 % Cr, 5 % nickel) have been developed by Kobe Steel for use as structural material at 4 K. The magnetic permeability of these steels at that temperature is 1.001, with no deterioration even after cold-working. These developments will find increasingly wide applications in fusion reactors, magneto-hydrodynamic generators, electromagnetic thrusters, and magnetically levitated (Maglev) trains.

8.2. Manganese in Nonferrous Alloys

Manganese alloys with nonferrous metals are made with low-carbon manganese. The addition of manganese improves hot and cold strengths, resistance to corrosion, and formability. A number of manganese alloys with nonferrous metals have therefore been developed, but of these only the alloys with copper and aluminum have achieved importance.

In combination with magnesium and other alloying metals, manganese produces the desired modification and hardening of aluminum alloys, e.g., Duralumin.

The Ni – Cu – Mn alloys have high electrical resistivity. The manganese contents of the most important copper alloys are given in Table 18. The alloy Manganin is used on a large scale as an electrical resistance material as its resistivity varies only slightly at high operating temperatures.

8.3. Nitrogen-Containing Manganese Alloys

The following types of manganese are suitable for nitriding: manganese metal, electrolytic manganese, and refined ferromanganese. High-carbon manganese can take up only small amounts of nitrogen because of the manganese already combined as carbide.

The nitriding process is carried out in annealing furnaces (e.g., hood type) in an atmosphere of pure nitrogen. The manganese metal or ferromanganese is in the form of powder or small pieces and the temperature is 700 – 1100 °C. Ferromanganese absorbs between 4 and 6 % nitrogen; increased manganese content gives increased uptake of nitrogen.

According to a patent of the Gesellschaft für Elektrometallurgie mbH, Düsseldorf [76] the nitrogen uptake of ferromanganese is improved by mixing with other ferroalloys such as ferrovanadium or ferrochrome. The nitriding is carried out by heating ferromanganese in an annealing furnace for 8 h at 950 °C in a nitrogen atmosphere at 101 kPa. Results obtained for the nitriding of low-carbon ferromanganese and ferrovanadium both separately and as a mixture are given in Table 19.

Table 18. Composition of some manganese-containing copper alloys

Alloy	Mn, %	Cu, %	Ni, %	Zn, %	Pb, %	Al, %	Fe, %
Copper manganese	10 – 40	65 – 70					
Manganese bronze	5 – 15						
Resistin	15						
Hard brass	1.5	56		42			
Special brass 58 Al_2	2	58	3	37	0.5	1.6	
Special brass 59	2.5	58	3			1.5	
Special brass 64	4	65				7	4
Manganin	12	83	3 – 5	0.2			

Table 19. Nitriding of ferroalloys

			Nitrided alloys (sintered)		
	Low-carbon Fe – Mn	Fe – V	Low-carbon Fe – Mn	Fe – V Fe – V	Low-carbon Fe – Mn + Fe – V 1 : 1
Mn, %	90.9		83.2		40.3
V, %		80.2		71.3	35.6
N, %	0.03	0.013	7.7	11.0	11.1

Table 20. Nitrogen-containing manganese alloys

Alloy type	Composition, %				
	Mn	N	C, max	Si, max	P, max
Sintered	75 – 85	4 – 6	1	1.5	0.20
Melted	80 – 85	1.5 – 2.5	0.10 – 0.30	1.5	0.15

Ferromanganese may also be nitrided in a medium-frequency induction furnace [77]. The temperature of the fine grained ferromanganese (1 – 5 mm) is raised to ca. 1200 °C, i.e., a little below the melting point. On cooling the furnace under nitrogen pressure, a sintered block is obtained containing up to 7 % N. If the nitrided ferromanganese is heated above its melting point, decomposition of the manganese nitride takes place and a molten ferromanganese is obtained with 1.5 – 2.5 % nitrogen. This nitrogen is distributed very uniformly in the alloy so that its use as a steel additive gives very reproducible nitrogen contents (see Table 20).

Ammonia is also used for nitriding of manganese and ferromanganese [77]. In this case, low temperatures are used (600 – 700 °C) and the nitrogen contents achieved are 12 – 13 %. Ammonia also leads to decarburization; in the nitriding of a ferromanganese powder, the carbon content is reduced from 1 % to 0.1 %.

9. Processing of Manganese Nodules and Encrustations

The primary economic interest in manganese deposits on the ocean floor has never been for the manganese content, but rather for the associated valuable metals, including nickel, copper, and cobalt. There are, however, some indications that the manganese content is becoming more attractive in its own right.

The deep ocean resources of manganese nodules were first discovered during the British HMS Challenger expedition of 1872 – 1876. The commercial potential of these resources was recognized in 1958 as a result of work undertaken at the University of California.

9.1. Occurrence, Origin, and Composition [78–81]

Occurrence. Manganese nodules occur over most of the deep ocean floor, but are currently considered to have economic potential in one

primary and three subsidiary areas (see Fig. 17). The primary area is located in the northwestern equatorial Pacific (the Clarion – Clipperton zone), and the subsidiary areas are in the South Pacific, western equatorial Pacific, and the central Indian Ocean. All of the main areas of potentially economically viable manganese nodules occur on the abyssal seafloor.

Limited quantities of ore-grade nodules have been identified within the Exclusive Economic Zones (EEZ) of some of the western Pacific islands.

Manganese encrustations occur on oceanic seamounts, primarily in the Pacific ocean.

Origin. Manganese on the ocean floor originates from two primary sources: runoff from the continents and submarine volcanic activity. It is believed that plankton extract the manganese and associated metals from seawater and carry them to the ocean floor after death. The metals are liberated by organic decay processes into the pore water of the deep-sea muds, where, by a series of diagenetic reactions, they migrate to centers of nucleation within the uppermost sediments and precipitate to form manganese nodules. Encrustations are formed on exposed rocks.

Composition. The chemical composition of manganese nodules varies widely. Some 74 elements have been found to occur in manganese nodules and encrustations. The average chemical composition varies from ocean to ocean and with latitude, water depth, and mineralogy.

World averages of nodule composition are influenced by the disproportionately high number of stations located in the Pacific Clarion – Clipperton zone, in which the nodules average 25.4 % Mn, 6.66 % Fe, 1.27 % Ni, 1.02 % Co. Excluding these values from the world averages decreases the average nickel, copper, and manganese contents, increases the iron content, and has no effect on the cobalt content.

Principle metal contents, together with nodule concentration and depth, are listed in Table 21.

The top of a nodule is generally higher in Fe, Co, and Pb and lower in Cu, Ni, Mo, Zn, and Mn compared to the bottom. The nodule core is usually deficient in Fe, Co, and Pb.

Physical Characteristics. Manganese nodules generally exhibit five external shapes: spheroidal, ellipsoidal, discoidal, botryoidal, and flat faced. The surface textures, which may be smooth or granular, may vary on the opposite sides of the nodule. The size varies from 0.5 – 20 cm, with concretions larger than 20 cm generally assuming the form of slabs. The color is usually dark reddish brown to black. Some zonal pattern in nodule cross section, produced by variation in mineral content of the growth layers, is generally present. The porosity is relatively high and relative density low (2.1 – 3.1).

Water comprises about 45 – 50 wt % of nodules when removed from the ocean. Air drying removes about half of this water. Approximately 5 – 10 % water is contained within the crystal structure.

Table 21. Chemical composition of manganese nodules

Element	Mean content, %	Range, %
Mn	18.60	0.04 – 50.3
Fe	12.47	0.30 – 50.0
Ni	0.66	0.01 – 1.95
Cu	0.45	0.01 – 1.90
Co	0.27	0.01 – 2.23

Mineralogy. Nodules comprise a complex mixture of organic and colloidal matter, nucleus fragments, and crystallites of various minerals. The phases are fine grained, often metastable, poorly crystallized, and intimately intergrown.

Todorokite, birnessite, and vernadite are the predominant manganese oxide minerals, while various other MnO_2 modifications in the form of α-MnO_2, β-MnO_2, and γ-MnO_2 have also been reported.

Iron is present as oxide, oxyhydroxide, and hydrated oxide phases, the most commonly observed being ferroxyhite, goethite, and lepidocrocite.

The accessory minerals can be divided into the following three categories: sheet silicates and zeolites, clastic silicates, volcanic minerals, and biogenic minerals.

Known minerals of nickel, copper, or cobalt have not been identified; Cu, Mo, Ni, and Zn are associated almost entirely with the manganese mineral phase. Ba, Cd, Ca, Mg, Sb, Sr, and V are usually associated with manganese but are

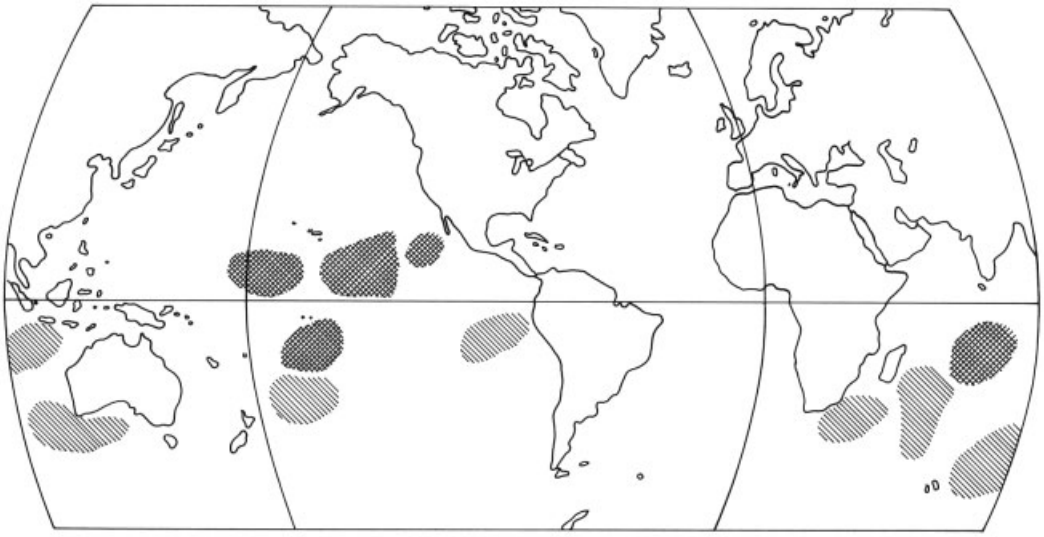

Figure 17. Manganese nodule occurrence
Hatched: areas of high nodule concentration; crosshatched: areas of potential economic viability

also found in other phases. Pb and Ti are associated almost entirely with the Fe and Mn phases.

9.2. Legal Considerations and Nodule Recovery [78], [80], [82–85]

Legal Considerations. The Law of the Sea as pertaining to the exploitation of mineral resources is of major importance to the future viability and prospects of commercial venture.

During the early 1960s interest in deep-ocean mining increased and the Third UN Law of the Sea conference began in 1974. The current Convention was eventually adopted by this conference in 1982.

The United States, the Federal Republic of Germany, and United Kingdom voted against the convention because of objections to certain provisions related to deep-sea mining.

The most important of the above regimes as pertaining to manganese nodule exploitation are the EEZ, extending to 200 nautical miles from the coastal state shoreline, and the so-called Area, extending beyond this limit. The coastal state has sovereignty over the EEZ, but the sea floor outside these zones falls under the jurisdiction of an International Seabed Authority which will be responsible for all mineral deposits and even mine these deposits through its own organization. The Authority would also license independent mining organizations to exploit these deposits, but in doing so would impose a number of conditions which some national governments and mining consortia find unacceptable.

The 200 mile EEZ would put under coastal state jurisdiction most of the minerals of the continental margins and, in some areas, those of the deep-ocean floor.

The convention also limits the amount of nickel that can be produced outside the EEZ over the first 20 years of commercial production in order to protect the markets of land-based producers.

The United States, Federal Republic of Germany, France, the United Kingdom, the former Soviet Union, and Italy have passed national deep-seabed mining legislation which, in many instances, is in contradiction of the UN Convention.

Nodule Recovery. Research into nodule recovery has been conducted since the 1960s and the technical feasibility, to a large extent, has been established.

Three main systems have been investigated: the continuous line bucket, the autonomous shuttle, and hydraulic pumping with a suspended conduit or dredge pipe.

Techniques based on continuous hydraulic pumping systems are the most technically and economically viable.

9.3. Processing [79], [80], [86], [87]

The following processes for recovering the metals of value (Mn, Co, Ni, Cu) from nodules have been identified as being potentially feasible:

1) Gas reduction and ammoniacal leach
2) Cuprion ammoniacal leach
3) Elevated temperature and pressure H_2SO_4 leach
4) Reduction and HCl leach
5) Smelting and H_2SO_4 leach

The first three of the above options recover Co, Ni, and Cu, while Mn can be optionally recovered from the tailings. The last two are more suited to the recovery of all four metals.

Ammoniacal processes are similar to those used for the treatment of land-based nickeliferous laterites which, in some respects, are similar to manganese nodules. Cobalt, nickel, and copper are soluble in $NH_3-(NH_4)_2CO_3$ (the Caron process) and $NH_3-(NH_4)_2SO_4$ solutions.

These processes involve selective reduction of the metals and disruption of the nodule matrix to permit rapid, complete dissolution of the valuable metals. Acidic chloride solutions are also capable of dissolving the valuable metals including manganese. Acidic sulfate solutions can solubilize cobalt, nickel and copper. Nickel is recovered from laterites by sulfuric acid leaching at elevated temperature.

The diverse chemical, physical and mineralogical composition of the nodules exclude physical enrichment prior to the complex hydrometallurgical processing. Separation on the basis of density, magnetic properties, or flotation has not been successful.

The recovery of manganese from the tailings of the ammoniacal processes (routes 1 and 2 above) involves flotation of the $MnCO_3$, which could then be converted to a manganese oxide product for further processing or sale. Manganese dioxide could be recovered from the H_2SO_4 process tailings (route 3 above) by dissolution and chemical the tailings from all the processes above for ferromanganese production is probably precluded due to the unacceptably high levels of trace metals, sulfur, and phosphorus.

9.3.1. Gas Reduction and Ammoniacal Leach Process

This process (Fig. 18) is an adaptation of the Caron process for nickeliferous laterites (see → Nickel).

The MnO_2 is reduced to MnO at 625 °C by a producer gas rich in CO. The mineral structure is disrupted by this reduction and the contained metals can be dissolved by an aqueous solution of 10 % ammonia and 5 % CO_2 at 40 °C and atmospheric pressure. The pregnant solution is separated from the solids, and copper and nickel are isolated by solvent extraction and aqueous stripping. The products are recovered as cathode copper and nickel by electrowinning.

The aqueous NH_3-CO_2 solution is contacted with hydrogen sulfide, and insoluble cobalt sulfides and residual copper, nickel, zinc, and other metals are precipitated.

This precipitate is removed and contacted with air and H_2SO_4 to selectively redissolve the cobalt and the small amount of nickel still present. The cobalt and nickel are recovered from the liquid in powder form by selective reduction with hydrogen at 3.4 MPa and 185 °C.

The residual ammonia and carbon dioxide are removed from the nodule residue by steam at 120 °C and 0.2 MPa. The resulting mixture is condensed and, together with the mixture of ammonia carbon dioxide and steam arising from the cobalt extraction step, is recycled.

9.3.2. Cuprion Ammoniacal Leach Process

Manganese dioxide is reduced in an aqueous ammoniacal solution containing an excess of copper(I) ions at 50 °C. The cobalt, nickel, and copper are solubilized in an aqueous solution of ammonia and carbon dioxide at 50 °C and atmospheric pressure. The separation and purification of the metals is identical to the Caron gas reduction process (see Section 9.3.1 and Fig. 18).

9.3.3. Elevated Temperature and Pressure H_2SO_4 Leach Process

In this process (Fig. 19) the cobalt, nickel, and copper content of the milled nodules is dissolved

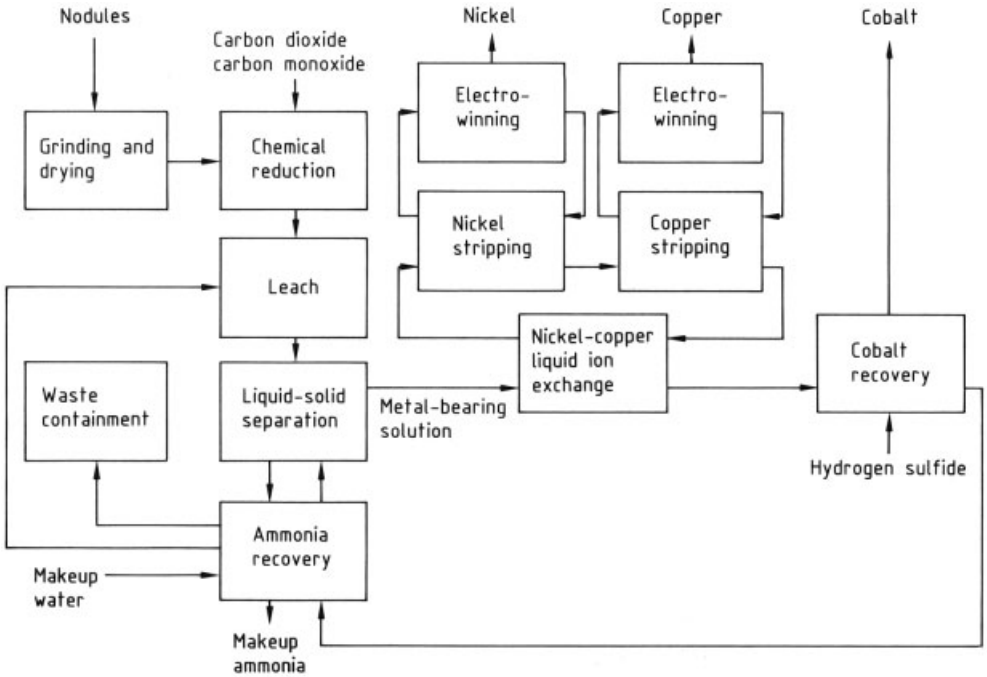

Figure 18. Gas reduction and ammoniacal leach process

in a 30 % sulfuric acid solution at 245 °C and 3.54 MPa. Manganese and iron are not solubilized to any appreciable extent. The resultant slurry is cooled and the acid solution separated from the nodule residue. The residue is washed, neutralized, and disposed of. The pregnant solution is pH adjusted and the copper and nickel are selectively removed by extraction with an organic solvent. These metals are then stripped as copper and nickel sulfate in acidic aqueous solutions and subsequently recovered by electrowinning.

The cobalt is precipitated with hydrogen sulfide together with residual nickel, copper, and zinc. The precipitate is separated from the solution and contacted with air and H_2SO_4 at 100 °C to selectively redissolve the cobalt and residual nickel, which are recovered as powders from the solution by reduction with hydrogen gas at 3.4 MPa and 185 °C.

The ammonia is recovered from the barren solution and recycled. The barren, ammonia-free solution is recycled for washing of the leached nodule material.

9.3.4. Reduction and HCl Leach Process

In this process (Fig. 20) the manganese dioxide is reduced to soluble manganese chloride with hydrogen chloride at 500 °C and 0.1 MPa. A portion of the hydrogen chloride is oxidized to chlorine. The excess hydrogen chloride is separated from this chlorine and water vapor and recycled. Dissolution of the iron in the nodules is minimized by injection of steam, which results in the hydrolysis of iron chloride to insoluble ferric hydroxide.

The soluble Co, Ni, Cu, and Mn chlorides are dissolved by leaching with HCl. Copper is selectively extracted by organic ion exchange and subsequently stripped into a strong H_2SO_4 solution. The copper is recovered by electrowinning.

The copper-free chloride solution is neutralized and cobalt is removed by a second selective solvent extraction stage. The cobalt is recovered as a powder by a sequence of stripping, precipitation as cobalt sulfide by hydrogen sulfide, selective leaching, and hydrogen reduction.

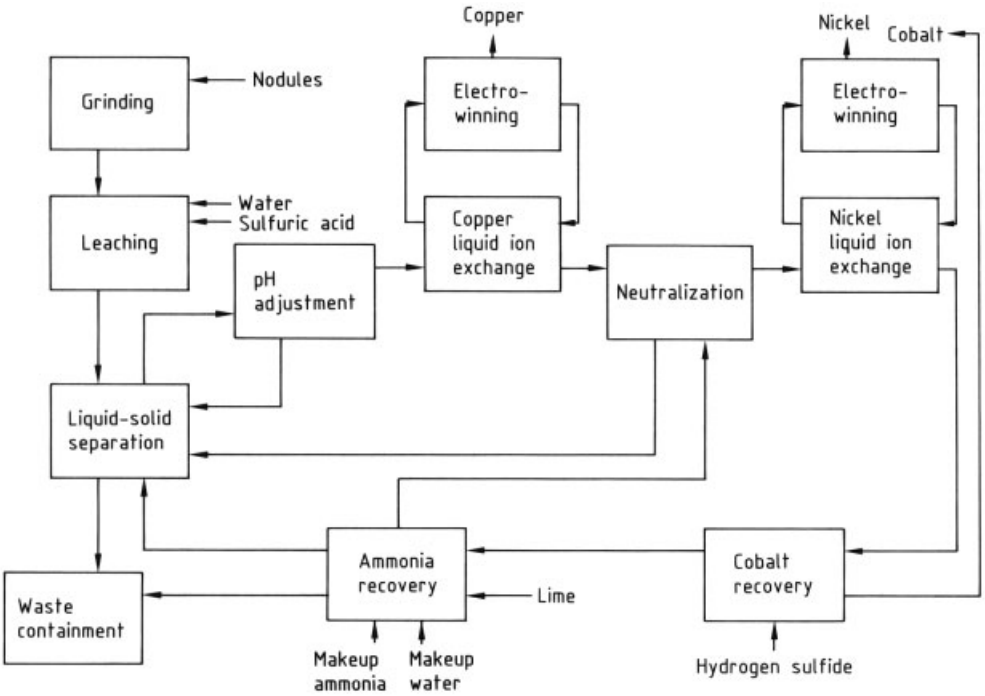

Figure 19. High-temperature, high-pressure sulfuric acid leach process

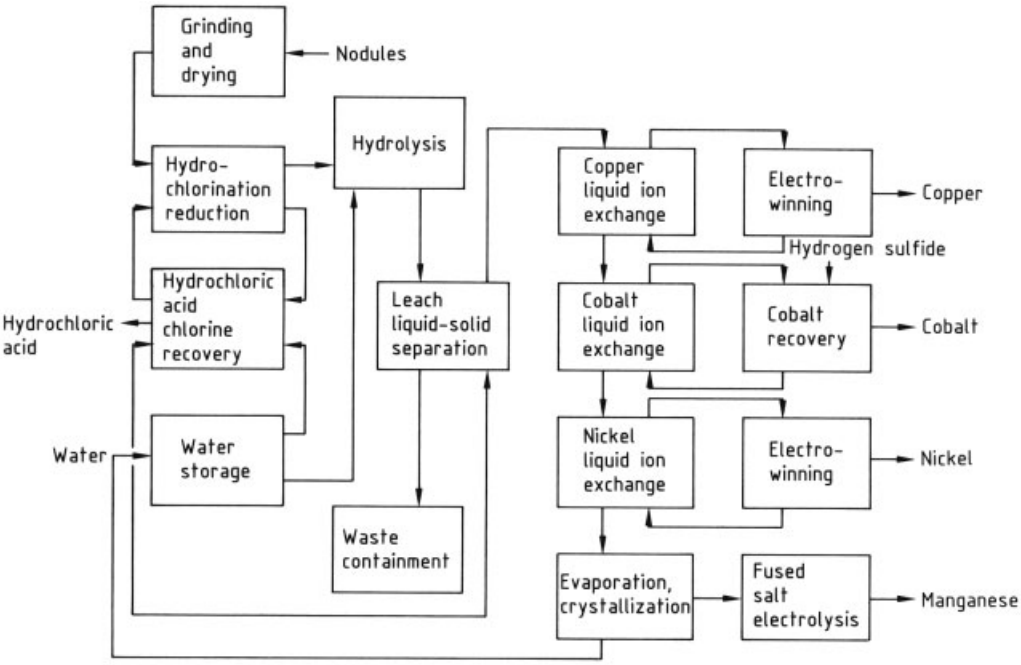

Figure 20. Reduction and hydrochloric acid leach process

The copper- and cobalt-free solution is subjected to a third solvent extraction stage to remove nickel. The nickel is stripped into an acidic sulfate solution for subsequent electrowinning.

The final aqueous chloride solution can be dried to produce impure manganese chloride, which can be charged to an electrolysis furnace where it dissolves in a molten alkali chloride bath. Molten manganese is tapped from the furnace and cast. Chlorine gas is also produced from the electrolysis furnace and, together with the chlorine resulting from the initial hydrochlorination stage, can be converted to hydrogen chloride for recycling.

9.3.5. Smelting and H_2SO_4 Leach Process

In the sulfuric acid leach process (Fig. 21), the nodules are dried and the manganese dioxide and ferric oxide reduced to manganous and ferrous oxides at 625 – 1000 °C with coal and carbon-monoxide-rich producer gas. The reduced nodules, together with coke and silica, are hot-charged to an electric furnace.

In this step most of the Cu, Ni, Co, and Fe, and some of the Mn is reduced at 1425 °C to form a molten alloy phase, which separates by gravity from the unreduced management slag. The alloy is transferred to converters where, after the addition of further silica, the Mn and most of the Fe is reoxidized with air, separated as a slag, and recycled to the furnace. Gypsum, coke, and possibly sulfur are added to the alloy, producing a metal sulfide matte phase containing the Cu, Ni, and Co. These are separated in the converter and the slag is recycled to the electric furnace, while the matte is water quenched and granulated.

A ferromanganese alloy can be produced from the furnace slag by smelting, together with iron-rich slags and coke, in an electric furnace.

The metals are leached from the granulated matte with a 5 % H_2SO_4 solution in the presence of oxygen at 1 MPa and 110 °C. Copper and nickel are recovered separately from this solution by selective solvent extraction. The pH of the solution is controlled by addition of ammonia. The metals are stripped into acidic aqueous solutions and the products recovered by electrowinning.

Cobalt and the residual other metals are recovered from the aqueous ammonium sulfate solution by precipitation with hydrogen sulfide. The separated precipitate is contacted with air and H_2SO_4 at 100 °C to selectively redissolve the cobalt and residual nickel, which are recovered in powder form by selective reduction with hydrogen at 3.4 MPa and 185 °C.

9.4. Economic Viability and Future Prospects [78], [80–82], [88], [89]

The constraints inhibiting the exploitation of the nodule reserves are complex–a combination of economic, technological, sociological, and political. Any operation would have to satisfy several prerequisites: technical feasibility of mining and processing; economic or strategic viability; legal authorization and security of tenure; and environmental acceptability.

Although most of the results of the extensive research and development work undertaken by several nodule recovery consortia has not been published, sufficient successful announcements have been made to indicate that the recovery and processing of nodules is technically feasible.

Numerous environmental impact assessments indicate that nodule dredging totally destroys the area of operation. Adjacent undredged areas, however, appear to be unharmed and provide the basis for rehabilitation. Adverse environmental effects are therefore short-lived.

The estimated economic viability of mining and processing nodules has been well reported, covering a wide range of scenarios. Estimates of the internal rate of return range from 3 to 18 %, with the more studies showing lower rates. The high degree of risk associated with nodule exploitation probably necessitates a rate of return of ca. 30 % to warrant investment. The studies also exhibit very high sensitivity to variations in commodity prices, metal recoveries, and deposit grades. From an economic viewpoint only, it is most likely that nodules will not be mined and processed in the foreseeable future without significant financial incentives in the form of price supports, market protection, and tax breaks.

The considerable uncertainty currently existing as to the legal authorization for mining beyond the limits of national jurisdiction (that is, outside the EEZ) casts more doubt on the feasibility of commercial nodule mining for some

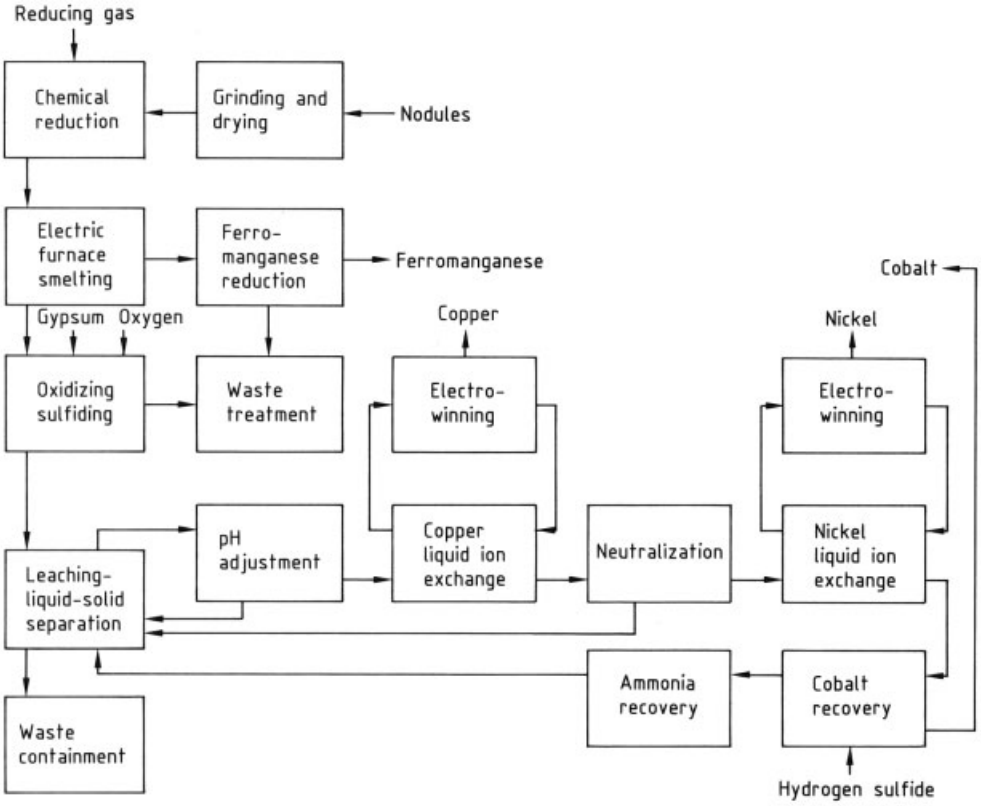

Figure 21. Smelting and sulfuric acid leach process

years to come. The prospects of commercial exploitation of the extensive deposits of cobalt-rich manganese oxide crusts lying within the EEZ are better than those for the nodules. The technical feasibility of recovering these crusts has not yet, however, been developed and tested.

The supply and demand situation, coupled to strategic considerations, for the four metals of interest will probably be the ultimate determining factor for future exploitation. Most mining consortia have only considered the recovery of cobalt, nickel, and copper. Recent reappraisals of manganese and its perceived value as a strategic metal (particularly in the United States) has led to an increasing interest in four-metal extraction. A modest increase in manganese usage is predicted in future, directly coupled to steel production. The perceptions of the instability of cobalt supply from African sources increase the strategic sensitivity of this metal.

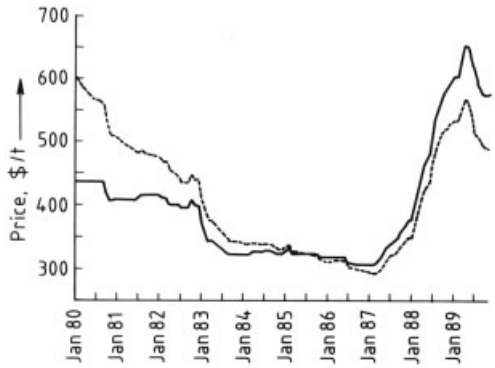

Figure 22. Price of ferromanganese
—- Actual price; – – – Price in 1985 $

Another important use for nodules is as a stack gas absorption mechanism. Elimination or reduction of sulfur and nitrogen oxides is possible by catalytic oxidation by compositions involving transition metals. Nodules possess the high specific surface area and transition metal

oxide content required for an oxidation catalyst. This use of nodules also provides a primary stage for the recovery of the metals of value. The complex metal oxides in the nodules are readily reduced to sulfides as a contaminated gas stream is passed through the filter pack of ground nodules. Well defined processing techniques, as described before, can then be applied to recover the four metals of interest.

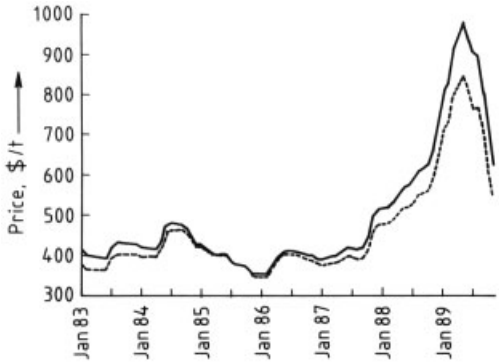

Figure 23. Price of silicomanganese
—- Actual price; – – – Price in 1985 $

10. Economic Aspects [90]

Ninety percent of manganese produced is consumed in the steel industry, and the growth of the manganese industry is therefore directly related to that of the steel industry. Steel production in the Western world dropped sharply in the early 1980s and bottomed out in 1982–1983. Economic recovery in the Western industrialized countries in the late 1980s resulted in strong growth in the steel sector. Between 1986 and 1988 the world production of crude steel increased by 9 percent to 780×10^6 t and is likely to continue at this level into 1990, when a drop off is expected. As a consequence the demand for ferromanganese alloys has increased considerably. The increase in demand placed upward pressure on the price of ferromanganese and silicomanganese in 1988 and 1989 and in terms of constant 1985 U.S. $ similar price levels are being achieved to those in the early 1980s (see Fig. 22). A sharp drop in the price of silicomanganese was experienced at the end of 1989 due to an excess of Chinese material in the market and the low ferrosilicon prices (see Fig. 23).

As steel demand is expected to flatten off in the 1990s high growth rates in the demand for manganese are not expected to continue. Improvements in the technology used in the steel industry will result in lower consumption of manganese alloys per tonne of steel. This trend will be further influenced by the development of the Optimal Refining Process (ORP) in Japan, in which manganese ore instead of alloys is added to the blast furnace runner prior to the basic oxygen furnace. In the short to medium term ferromanganese prices are likely to remain high due to the expected level of high-grade manganese ore prices. Silicomanganese prices should remain weak as the production of this material is not as dependent on high-grade ores and will continue to be influenced by the ferrosilicon price and availability.

The increased demand for ferromanganese alloys has highlighted the shortage in the production capacity of high-grade ores, and as a result large increases in the ore prices were seen in 1988 and 1989 (see Fig. 24). Demand for high-grade ores is likely to increase due to the additional demand materializing in the former Soviet Union, China and the Eastern-bloc countries.

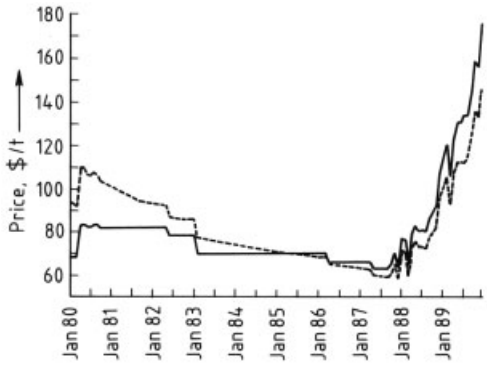

Figure 24. Manganese ore prices
—-Actual price; – – – Price in 1985 $

11. References

1. J. S. Stanko: *Manganese for Steelmakers,* Council for Mineral Technology, Johannesburg, Republic of South Africa 1989.
2. S. A. Weiss: *Manganese – the Other Uses,* The Garden City Press Ltd., Letchworth 1977.
3. E. A. Brandes, R. F. Flint: *Manganese Phase Diagrams,* The Manganese Centre, Paris 1980.

4. O. Kubaschewski, E. L. Evans, C. B. Alcock: *Metallurgical Thermochemistry,* Pergamon Press, Oxford 1967.
5. L. Habraken: "Manganese – an Alloying Agent,"*Proc. Semin. Non Ferrous Applications of Manganese,* The Manganese Centre, Paris 1977.
6. B. McMichael: "Manganese, EMD – the Power in the Market," *Ind. Miner.* (1989) May, 42 – 59.
7. J. F. Machamer: "A Working Classification of Manganese Deposits," *Min. Mag.* **157** (1987) no. 10, 348 – 351.
8. *Manganese Ores Analysis,* The Manganese Centre, Paris 1986.
9. A. S. E. Kleyenstuber: "The Correlation of the Mineralogy with the Sedimentary Cycles of the Proterozoic Manganese-Bearing Hotazel Formation in South Africa," *IAS Int. Symp. Sedimentology Related to Mineral Deposits,* 30th July – 4 August 1988, Beijing.
10. C. J. Nel, N. J. Beukes, J. P. R. de Villiers: "The Mamatwan Manganese Mine of the Kalahari Manganese Field," in C. R. Anhaeusser, C. R. Maske (eds.): *Mineral Deposits of Southern Africa,* vol. I, Geol. Soc. S. Afr., Johannesburg 1986, pp. 963 – 978.
11. S. Roy: *Manganese Deposits,* Academic Press, London 1981.
12. I. M. Varentson, G. Y. Graselly: *2nd International Symposium on the Geology and Geochemistry of Manganese,* Schweitzerbart'sche Verlagsbuchhandlung, Stuttgart 1980.
13. D. B. Wellbeloved: "Manganese in South Africa – an Overview," *Conference on Base Metals,* Johannesburg, May 17, 1988.
14. *Manganese Statistical Bulletin 1987,* The Manganese Centre, Paris 1988.
15. M. P. Sassos: "Manganese – GEMCO Mines Huge Resources at Groote Eylandt," *Eng. Min. J.* **185** (1984) 57 – 58.
16. A. R. Chernenko, A. S. Sidorov: "Updating the Technology and Machinery for Underground Manganese Ore Mining," 12th World Mining Congress, New Delhi 1984.
17. M. C. Fuerstenau, K. N. Han, J. D. Miller: "Flotation Behaviour of Chromium and Manganese Minerals," in P. Somasundaran (ed.): *Advances in Mineral Processing,* Society of Mining Engineers, Littleton 1986, pp. 289 – 307.
18. J. Chen, D. Chen, T. Chen, S. Liu: "Technological Innovations and Theoretical Approach to Shear Washing for a Refractory Manganese Ore," *Zhongnan Kuangye Xueyuan Xuebao* **19** (1988) no. 4, 371 – 378.
19. O. Holta, S. E. Olsen: "Composite Briquettes of Manganese Ore and Coal," *Electr. Furn. Conf. Proc.* **43** (1986) 273 – 277.
20. K. P. Abraham, M. W. Davies, F. D. Richardson: "Determination of Manganese Oxide in Silicate Melts," *J. Iron Steel Inst. London* **196** (1960) 82 – 86.
21. F. D. Richardson: *Physical Chemistry of Melts in Metallurgy,* **vol. 2,** Academic Press, London 1974, p. 333.
22. I. Barin, O. Knacke: *Thermodynamic Properties of Inorganic Substances,* Springer Verlag, Berlin 1974.
23. J. H. Downing: "Fundamental Reactions in Submerged Arc Furnaces," *Electr. Furn. Conf. Proc.* **19** (1963) 288 – 296.
24. J. S. Stanko: *Steel Makers Guide to Manganese,* Mintek, Randburg 1989.
25. Commodities Research Unit: *The Costs of Producing Ferroalloys,* vol. 2, London 1989.
26. V. V. Strishkov, R. M. Levine: *The Manganese Industry in the U.S.S.R.,* U.S. Bureau of Mines, Washington 1986.
27. H. Bromet: "Technical and Economic Comparison of the Various Furnaces used for Ferro and Silico Manganese," *INFACON 80, Proc. Int. Ferro-Alloys Congr.* **2nd,** 1980.
28. O. Kubaschewski, C. D. Alcock: *Metallurgical Thermochemistry,* 5th ed., Pergamon Press, Oxford 1979.
29. *Plasma Torches in Blast Furnaces,* Société du Ferromanganese de Paris, Outreau 1986.
30. W. J. Rankin: "Si–Mn Equilibrium in Ferromanganese Production," *J. S. Afr. Inst. Min. Metall.* **79** (1979) Sept., c 167 – c 174.
31. S. Sakurai: *Recent Operational Results of the Shaft Type Ferro-Manganese Smelting Furnace,* Mizushima Ferro-Alloy Company Japan, Mizushima 1986.
32. A. G. Arnesen, B. Asphaug: "Computer Control of Electric Smelting Furnaces," *Proc. Int. Ferro Alloy Conf.,* Johannesburg 1974.
33. Z. Van der Walt, W. A. Gericke: "The Effect of the Magnitude of Scale on Various Operating Characteristics in the Production of High-Carbon Ferromanganese," *Proc. Lat. Am. Inst. Iron and Steel,* Seminar Ferroalloys, 1975.
34. G. Glöckler, H. G. Müller: "A Comparison of the Electric Submerged Arc Furnace and the Plasma Smelting Furnace," *BHM Berg-Hüttenmänn. Monatsh.* **134** (1989) no. 5.

35. A. Kousaris, J. B. See: "Reactions in the Production of High Carbon Ferromanganese from Mamatwan Ore," *J. S. Afr. Inst. Min. Metall.* **19** (1979) 149 – 158.
36. R. T. Hooper, "The Production of Ferromanganese," *Electr. Furn. Conf. Proc.* **24** (1967) 141 – 145.
37. L. Rossenmyr: "Manganese Alloy Production in Large Submerged Arc Furnaces," *Electr. Furn. Conf. Proc.* **27** (1970) 121 – 124.
38. Tinfos Company Brochure, Tinfos Jernwerk AS, Oyestronda, Norway 1989.
39. "Electric Furnaces Beats Pollution," *Mod. Power Eng.,* May, 1974.
40. W. Gericke: "The Establishment of a 500 000 t/a. Sinter Plant at Samancor's Mamatwan Manganese Ore Mine," *Proc. Int. Ferroalloys Congr.* **5th** (1987) April.
41. A. A. Tskilishvili: "Production of High-Carbon Ferromanganese from Manganese Iron Sinter," *Stal'* (1975) no. 9, 816 – 817.
42. E. Svana: "Ferro Manganese Smelting," *Elkem Seminar on Smelting,* New Delhi 1974.
43. R. G. Rotzlaff: "High Carbon Ferromanganese Production on the High Alumina Practice," *Electr. Furn. Conf. Proc.* **44** (1987) 169 – 173.
44. H. W. Wise: "High Carbon Ferromanganese Metallurgical Variables," *Electr. Furn. Conf. Proc.* **24** (1968) 88 – 92.
45. J. F. Dery: "Metallurgical Calculations for Ore Evaluation and Operational Planning in the Electric Furnace Smelting of High Carbon Ferromanganese," unpublished paper (1964).
46. D. S. Kozak, L. R. Matricardi: "Production of Refined Ferromanganese Alloy by Oxygen Refining of High Carbon Ferromanganese," *Electr. Furn. Conf. Proc.* **38** (1980) 123 – 127.
47. R. Ando, T. Yamagishi, T. F. Fukusima, K. Kawasaki: "A Study on the Silicomanganese Process," *Electr. Furn. Conf. Proc.* **34** (1974) 107 – 115.
48. Creusot-Loire and S. A. Manganese Amcor Ltd., FR 79 0169, 1979.
49. N. A. Barcza, D. P. O'Shaughnessy: "Optimum Slag –Alloy Relationships for the Production of Medium to Low Carbon Ferromanganese," *Can. Metall. Q.* **20** (1981) no. 3, 285 – 294.
50. T. Chisaki, K. Takervchi: "Electric Smelting of High Carbon Ferromanganese with Preheated Prereduced Materials at Kashima Works," Kashima 1971.
51. I. Tanabe: "Preheating of Ore for a Ferromanganese Furnace. A Recent Trend in Japan," *Electr. Furn. Conf. Proc.* **23** (1967) 131 – 138.
52. C. T. Ray, A. H. Olsen: "Computer Control of Four Furnaces in Tasmania," *Electr. Furn. Conf. Proc.* **44** (1987) 217 – 223.
53. T. K. Leonard: "Expert Systems Can Control Ferroalloy Processes," *Proc. Int. Ferroalloy Congr.* **5th** (1989) 267 – 279.
54. H. Muller, D. B. Wellbeloved: "Examples of Plasma Potential for Industrial Applications," *Symp. Proc. Int. Symp. Plasma Chem.* 1989.
55. S. A. Weiss: *Manganese the Other Uses. A Study of the Non-steelmaking Applications of Manganese,* Metal Bulletin Books Ltd., London 1977.
56. J. W. Mellor: *A Comprehensive Treatise on Inorganic and Theoretical Chemistry,* vol. XII, Longmans, London 1972.
57. E. C. Snelling: *Soft Ferrites, Properties and Applications,* 2nd ed., Butterworths, London 1988.
58. J. H. Jacobs et al.: "Operation of Electrolytic Manganese Pilot Plant, Boulder City, Nevada," *U.S. Dep. Inter. Bull.* **463** (1948).
59. G. L. Hubred, R. T. Lewis, US 4 317 730, 1982.
60. T. I. Mercz, J. C. Madgwick: "Enhancement of Bacterial Leaching by Microalgal Growth Products," *Proc. Australas. Inst. Min. Metall.* **283** (1982) 43 – 46.
61. F. N. Bender, C. Rampacek: "Percolation Leaching of Manganese Ores With Sulphur Dioxide," *U.S. Dep. Inter. Rep. of Investigation* **5233** (1957).
62. P. K. Sahoo, S. C. Das, S. K. Bose, S. C. Sircar: "Separation of Iron From Manganese Ore Roast-leach Liquor," *J. Chem. Tech. Biotechnol.* **29** (1979) 307 – 310.
63. S. Joris: "La Cinetique de Precipitation des Sulphures de Cobalt et de Nickel par l'Hydrogene Sulphure," *Bull. Soc. Chim. Belg.* **78** (1969) 607.
64. P. M. Ambrose, US 2 316 750, 1943.
65. J. Koster, US 2 483 287, 1949.
66. E. M. Wanamaker, W. D. Morgan, US 2 325 723, 1943.
67. P. M. Ambrose, US 2 347 451, 1944.
68. J. J. Harper, US 4 311 521, 1982.
69. J. E. Reynolds, N. J. Lombardo, US 4 423 012, 1984.
70. The Consolidated Mining and Smelting Company of Canada Ltd., GB 540 228, 1939.
71. M. Harris, D. M. Meyer, K. Auerswald: "The Production of Electrolytic Manganese in South Africa," *J. S. Afr. Inst. Min. Metall.* **77** (1977) Feb. 137.

72. A. I. Vogel: *A Text Book of Quantitative Inorganic Analysis, Including Elementary Instrumental Analysis,* 3rd ed., Longmans, London, p. 953.
73. J. K. Tuset: "Silicothermic Reduction of Manganese," *Steelmaking Conf. Proc.* **69** (1986) 59 – 64.
74. M. M. Bourgeois: "Les Alliages de Manganese a Bas Carbonne: Silico-manganese et Ferro-manganese," *Rev. Metall.* (Paris) (1979) June.
75. K. Venugopalan, M. Kamaludeen, N. G. Renganathan, V. Aravamuthan: "Massive Manganese Metal By Fused Salt Electrolysis," *Trans. Indian Inst. Met.* **33** (1980) no. 1, 79 – 81.
76. Gesellschaft für Elektrometallurgie GmbH, Düsseldorf, DE 1 558 500, 1967.
77. R. Durrer, G. Volkert: *Metallurgy of Ferroalloys,* vol. 2, Springer Verlag, Berlin 1972.
78. D. S. Cronan: "Marine Mineral Resources," *Geol. Today* **1** (1985) no. 1, 15 – 19.
79. B. W. Haynes et al.: "Pacific Manganese Nodules, Characterization and Processing," *U.S. Bur. Mines Bull.* **679** (1985) 2 – 32.
80. V. E. McKelvey: "Subsea Mineral Resources," *U.S. Geol. Surv. Bull.* **1689-A** (1986) 72 – 95.
81. D. S. Cronan: "Manganese on the Ocean Floor," *Trans. Inst. Min. Metall. Sect. A* **94** (1985) July, 166 –167.
82. W. D. Siapno: "Manganese Nodules: Overcoming the Constraints," *Mar. Min.* **5** (1986) no. 4, 457 – 465.
83. J. S. Chung: "Advances in Manganese Nodule Mining Technology," *Mar. Technol. Soc. J.* **19** (1985) no. 4, 39 – 44.
84. S. Wakabayashi, E. Kuboki, T. Saito, A. Takata, T. Linuma: "Research and Development Project of Manganese Nodule Mining System in Japan," *Proc. Fifth Int. Offshore Mechanics and Arctic Eng. Symp.,* Tokyo 1986.
85. D. C. Tolefson, J. P. Latimer, R. Kaufman: "Design Considerations for a Commercial Deep Ocean Mining System," *Proc. Fifth Int. Offshore Mechanics and Arctic Eng. Symp.,* Tokyo 1986.
86. S. B. Kanungo, R. P. Das: "Extraction of Metals from Manganese Nodules of the Indian Ocean by Leaching in Aqueous Solution of Sulphur Dioxide," *Hydrometallurgy* **20** (1988) no. 2, 135 – 146.
87. S. Anand, S. C. Das, R. P. Das, P. K. Jena: "Leaching of Manganese Nodules at Elevated Temperature and Pressure in the Presence of Oxygen," *Hydrometallurgy* **2** (1988) no. 2, 155 – 168.
88. C. T. Hillman, B. B. Gosling: "Mining Deep Ocean Manganese Nodules. Inf. Circ. No. 9015," *U.S. Dept. Inter. Bur. Mines,* Washington 1985, p. 1 – 15.
89. H. Bozon: "La Future Exploitation des Nodules Polymetalliques Sous-Marins," *JFE J. Four Electr.* (1982) no. 9, 13 – 21.
90. Roskill Information Services: *The Economics of Manganese* 1989, 5th ed., 1989.

Manganese Compounds

ARNO H. REIDIES, Carus Chemical Company, La Salle, Illinois 61301, United States

1. Introduction 1
2. Manganese Oxides 2
2.1. Manganese(IV) Oxide 2
2.1.1. Properties 2
2.1.2. Natural Manganese Dioxide 2
2.1.3. Synthetic Manganese Dioxides . . . 3
2.1.3.1. Commercial Forms 3
2.1.3.2. Synthetic Manganese Dioxides for Dry Cells 4
2.2. Other Manganese Oxides 8
3. Manganese(II) Salts 9
4. Higher Oxidation-State Manganates 11
4.1. Potassium Manganates(V) and (VI) 11
4.2. Potassium Permanganate 12
4.2.1. Properties 12
4.2.2. Production 12
4.2.2.1. Fusion Processes 13
4.2.2.2. Anodic Oxidation of Manganate(VI) 14
4.2.2.3. Production Plants 15
4.2.3. Commercial Grades, Packaging, and Transportation 17
4.2.4. Uses 18
4.2.5. Economic Aspects 18
4.3. Sodium Permanganate 18
5. Miscellaneous Manganese Compounds 19
6. Occupational Health and Environmental Aspects 19
7. References 20

1. Introduction [1]

Manganese compounds occur in all oxidation states from − 3 to + 7, with + 2, + 4, and + 7 being the most important. Most of the species in which Mn displays a negative valence state are anionic; for example, the ion $Mn(CO)^{3-}$ [obtained by vigorous reduction of $Mn_2(CO)_{10}$] contains Mn in the − 3 state. The − 2 oxidation state can be found in a complex of manganese with phthalocyanine, whereas manganese in the − 1 state is represented by the manganese pentacarbonyl ion $Mn(CO)_5^-$. Ions representing the various positive valences display characteristic colors, as indicated in Table 1. The color of a particular ion in solution is not always the same as that of the corresponding solid compound. Thus, solid MnO is light green, $Mn(OH)_2$ is white, and Mn_3O_4, Mn_2O_3, and MnO_2 are black. Liquid Mn_2O_7 is green under reflected light but deep red under transmitted light.

Given the multitude of valence states for manganese compounds, redox reactions are of great importance. Under suitable pH conditions, compounds containing manganese with a valence of + 3 or more are effective oxidants. Furthermore, tri-, penta-, and hexavalent manganese compounds tend to undergo disproportionation reactions under the influence of H^+ and OH^- :

$$2\,Mn^{3+} \rightarrow Mn^{2+} + Mn^{4+}$$
$$2\,Mn^{5+} \rightarrow Mn^{4+} + Mn^{6+}$$
$$3\,Mn^{6+} \rightarrow Mn^{4+} + 2\,Mn^{7+}$$

The divalent state is generally regarded as the most stable, at least in acid to neutral media. Tetravalent manganese is also quite stable as MnO_2 under moderately acid and moderately alkaline conditions, while the heptavalent state displays maximum stability around pH 7. Precipitated manganese(II) hydroxide is oxidized at room temperature by oxygen to tri- and tetravalent manganese oxides in the presence of traces of alkali, while conversion to manganates (i.e., the penta- and hexavalent states) requires much higher temperatures and high alkali concentrations ($\geq$ 180 °C for 60 – 70 % KOH). The basicity of the manganese oxides decreases (and the acidity increases) with increasing valency. Thus, MnO behaves as a basic anhydride, MnO_2 is amphoteric, and Mn_2O_7 is an acidic anhydride. Even though manganese in its highest valence state (+ 7) resembles chlorine (i.e., perchlorate), its general reactivity corresponds more closely to that of iron.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_123

Table 1. Ionic species of manganese

Ion	Valence	Color
Mn^{2+}	+2	pink
Mn^{3+}	+3	red
Mn^{4+}	+4	brownish black
MnO_4^{3-}	+5	blue
MnO_4^{2-}	+6	green
MnO_4^{-}	+7	purple

2. Manganese Oxides

2.1. Manganese(IV) Oxide

2.1.1. Properties

Manganese(IV) oxide [*1313-13-9*], MnO_2, M_r 86.93, is deep black to dark brown and practically insoluble in water. It is more commonly referred to as manganese dioxide, occasionally as pyrolusite. Pyrolusite (German: Braunstein) is also the name of a specific MnO_2-containing mineral. The purest manganese dioxide – corresponding rather closely to the formula MnO_2 – occurs in the form known as the β-modification. The compositions of other natural or synthetic manganese dioxides range from $MnO_{1.7}$ to $MnO_{<2.0}$ with varying contents of lower-valent manganese, foreign cations (e.g., K^+, Na^+, Ba^{2+}), hydroxyl ions, and water molecules. At least six distinct modifications of manganese dioxide have been characterized (α, β, γ, δ, ε, and ramsdellite) [2], [3], which differ according to their degree of crystallization and their content of foreign ions. The β-modification (as in the mineral pyrolusite) is not only the least reactive form of MnO_2 but also the most highly crystalline, and it comes the closest to having a stoichiometric composition. By contrast, γ-MnO_2 is nearly amorphous and much more reactive both chemically and electrochemically. When heated up to 500 °C, manganese dioxides frequently release water and/or undergo phase transitions; above 500 °C they liberate oxygen. Between 500 °C and 600 °C MnO_2 is converted into Mn_2O_3, and above 890 °C into Mn_3O_4. Manganese dioxide acts as an oxidant toward readily oxidizable materials, its own valency changing from +4 to +3 or +2. For example, under acidic conditions manganese dioxide oxidizes chloride ions to chlorine; hydrazine and hydroxylamine to nitrogen; iron(II) to iron(III); carbon monoxide to carbon dioxide; alkylbenzenes to aromatic carboxylic acids, ketones, and, in some cases, aldehydes; and oxalic acid to carbon dioxide.

In the presence of strong alkalis and at elevated temperature, however, manganese dioxide is itself readily oxidized by oxygen to manganese(V) and (VI) compounds. Other noteworthy properties of the various MnO_2 modifications include their sorptive and ion-exchange capabilities and their catalytic and electrochemical activities. The latter is of great importance to the dry-cell battery industry, which consumes large amounts of "battery-active" MnO_2 (especially the γ- and ε-modifications) for use as a depolarizer (→ Batteries, Chap. 6.1.3.1.

2.1.2. Natural Manganese Dioxide

Only about 5 % of the world production of manganese ore (totalling about 27×10^6 t/a) is consumed in nonmetallurgical applications. In 1976, the following approximate tonnages of manganese ore were associated with applications other than steel making [4, p. 331]:

For dry cell batteries (battery-active ore plus ore consumed in the production of battery-grade electrolytic manganese dioxide and chemical manganese dioxide)	500 000 t/a
In brick and ceramic coloring, including in glass manufacture	200 000 t/a
In welding rod manufacture	200 000 t/a
For the production of Mn chemicals and diverse chemical and metallurgical products	450 000 t/a
Total	1 350 000 t/a

Within limits, these consumption figures are probably still applicable, because world ore consumption in 1976 (24×10^6 t/a) was only slightly less than that in 1986 (25×10^6 t/a), and about 5 % is still consigned to metallurgical uses. Organic MnO_2 oxidations are generally carried out in the presence of sulfuric acid, and they lead to manganese sulfate as a byproduct (see Chap. 3). Oxidation of aniline by MnO_2 in the presence of sulfuric acid was once the main industrial route to hydroquinone, but since the late 1960 s this technique has been largely replaced by other methods, at least in the Western world. For more information on the synthetic applications of native manganese dioxide, see [5, pp. 36 – 37].

Several processes are known for the recovery of manganese from low-grade MnO_2-containing ores (Mn content < 20 wt %) [6, pp. 727 – 731], [7], but they are currently of little practical significance since high-grade manganese ores (Mn > 40 wt %) are widely available at moderate prices.

2.1.3. Synthetic Manganese Dioxides

2.1.3.1. Commercial Forms

Several types of synthetic manganese dioxide are produced commercially for specific end uses. In principle, synthetic manganese dioxide can be prepared either by a strictly chemical route (*chemical manganese dioxide, CMD*) or by electrochemical methods (*electrolytic manganese dioxide, EMD*).

Procedures leading to CMD include the oxidation of Mn(II) salts or lower manganese oxides, reduction of permanganates, thermal decomposition of manganese(II) nitrates, thermal decomposition/oxidation of manganese carbonate, and disproportionation of Mn(III) compounds.

Manganese dioxides obtained by reduction of permanganate are commonly known as *manganites*. They correspond to the δ-modification of MnO_2 and should actually be regarded as salt-like combinations of hydrous manganese dioxide and cations of base-forming metals (e.g., $K_2O \cdot 4\,MnO_2$). The metal portion is exchangeable against other cations. Such manganites [which are actually alkali manganates(IV)] are often obtained as byproducts from technical permanganate oxidations, but they may also be synthesized intentionally by direct reduction of permanganate (e.g., with manganese(II) salts). Until the late 1960 s, the manganites were important in the manufacture of dry cells because they are highly active depolarizers. Manganites for this purpose were produced and sold in Europe under such trade names as Manganit and Permanox. However, in recent years they have been largely displaced by electrolytic manganese dioxide (EMD) and the other types of chemically prepared manganese dioxide (CMD) described in Section 2.1.3.2.

Also of interest is the *hydrated manganite,* better known as *active manganese dioxide,* used in organic syntheses under nonaqueous, neutral conditions. Manganese dioxide of this type is generally prepared from $MnSO_4$ and $KMnO_4$ under strictly controlled conditions. Numerous methods are available for the preparation of "active" manganese dioxide. Many involve oxidation of a Mn^{2+} salt with $KMnO_4$, but $NaClO_3$ or O_3 are also used as oxidants. In a typical procedure, a solution of manganese sulfate (151 g/2.87 L) is added with stirring to a solution of potassium permanganate (105 g/2 L), and the resulting suspension of hydrous MnO_2 is stirred at 60 °C for 1 h. After filtration and washing, the precipitate is dried to a constant weight at 60 °C [8, p. 122]. These conditions lead to precipitated MnO_2 in the poorly crystallized but very reactive γ-form. It is a specific oxidant whose uses include dehydrogenation (e.g., the preparation of unsaturated aldehydes and ketones from unsaturated alcohols) and coupling reactions.

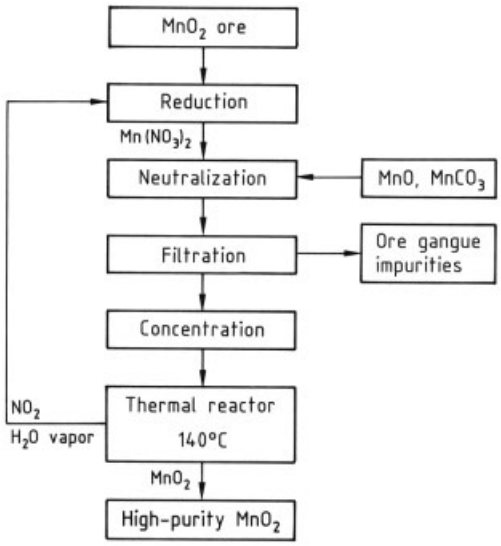

Figure 1. Manufacture of high-purity manganese dioxide MnO_2 (Chemetals process)

Mention should also be made of the increasing use of manganese-dioxide-based *oxidation catalysts*, particularly for air pollution abatement (removal of volatile organics, destruction of ozone) [9], [10].

Another commercial CMD is a *high-purity manganese dioxide* used in the manufacture of high-purity lower manganese oxides as well as ferrites (ceramic magnets) and thermistors for the electronics industry. This product (99.5 wt %

MnO_2) is obtained by thermal decomposition of manganese nitrate:

$$Mn(NO_3)_2 \overset{H_2O}{\rightleftharpoons} MnO_2 + 2\,NO_2$$

For a description of the original manganese nitrate process developed by IG Farbenindustrie (Bitterfeld) in the 1920 s, see [11]. A later modification of this process by Chemetals has been utilized at a plant in Baltimore since the early 1970 s [12], [13] (Fig. 1).

A concentrated solution of manganese nitrate (made from MnO_2 ore by reaction with NO_2) is first purified and then thermally decomposed at about 140 °C in a well-agitated, externally heated reactor. The resulting β-manganese dioxide precipitates as a fine, free-flowing powder; the nitrogen dioxide coproduct is recycled to generate new manganese nitrate. Chemetals' production capacity for high-purity manganese dioxide is about 6000 t/a.

2.1.3.2. Synthetic Manganese Dioxides for Dry Cells

From a commercial standpoint the most important synthetic manganese dioxides are those that are electrochemically active and are therefore useful as depolarizers in dry-cell batteries. Electrochemical or battery activity of MnO_2 results from a favorable combination of such factors as crystal structure, surface area, porosity, and chemical purity. Battery activity is not readily predictable; the only reliable way of establishing this property is by an actual performance test in a battery.

One important type is the so-called *activated manganese ore*, actually a semisynthetic product. It is made by roasting a high-grade oxidic manganese ore containing at least 80 wt % MnO_2 but with low or no battery activity at 600 °C. The product is then treated with hot sulfuric acid.

$$2\,MnO_2 \xrightarrow{600\,°C} Mn_2O_3 + 1/2\,O_2$$

$$Mn_2O_3 + H_2SO_4 \longrightarrow MnO_2 + MnSO_4 + H_2O$$

The second reaction is a disproportionation of trivalent manganese leading directly to battery-active γ-MnO_2. The coproduct $MnSO_4$ must be separated by leaching with water. Activated manganese ore is somewhat less effective as a depolarizer than EMD or fully synthetic CMD (see below), but it is still produced in France at a rate of several thousand tonnes per year under the names Ergogene and Philodyne.

Chemical Manganese Dioxide (CMD). Much more important than activated ores are the CMD products manufactured by the Sedema Division of Sadacem S.A. in Belgium under the trade names Faradiser M (for Leclanché and magnesium cells) and Faradiser WSZ (for zinc chloride batteries). The chemical compositions of the two products are practically identical: ca. 90 wt % MnO_2 (primarily the γ-modification), 2 wt % water, plus minor amounts of lower manganese oxides. Critical trace impurities (Co, Ni, Cu, Mo) are kept below 0.001 wt %. Types M and WSZ differ in such physical properties as surface area and density. Typical particle sizes are 80 % $<$ 44 µm for type M and 85 % $<$ 44 µm for type WSZ.

The first step in the production process [13], [14] (Fig. 2) is the reduction of MnO_2 ore to MnO using heavy fuel oil and a temperature of ca. 900 °C. The MnO is then treated with sulfuric acid to form manganese(II) sulfate. After careful neutralization with MnO to precipitate heavy metal impurities, solids are removed by thickening. The solution of $MnSO_4$ is then treated with ammonium carbonate (by simultaneous addition of NH_3 and CO_2) to precipitate manganese(II) carbonate, ammonium sulfate being generated as a byproduct. The $MnCO_3$ is separated and dry-roasted in the presence of air at about 320 °C to form a higher manganese oxide with the approximate composition $MnO_{1.80-1.85}$. Complete oxidation to MnO_2 is achieved by treatment with $NaClO_3$ in the presence of sulfuric acid. The key oxidation reactions may be represented as

$$MnCO_3 \xrightarrow{320\,°C} MnO_{1.85} + CO_2$$

$$MnO_{1.85} + NaClO_3 \xrightarrow{H_2SO_4} MnO_2 + Cl_2$$

Sedema introduced a process improvement in the course of expanding their production in which the MnO_2 is transformed directly to $Mn(NO_3)_2$ by reaction with nitric oxides. The procedure is similar to the first step in the

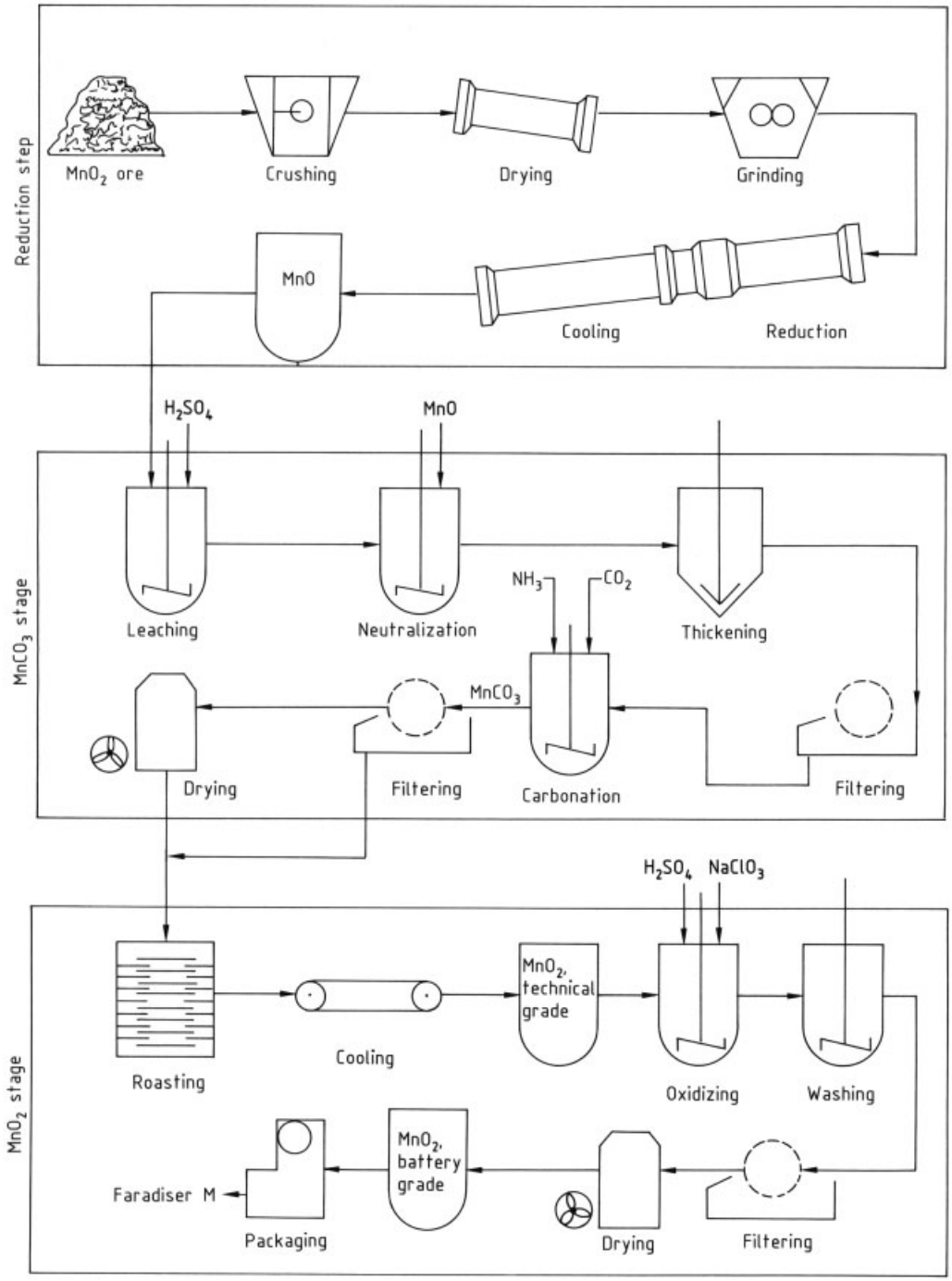

Figure 2. Production of chemical manganese dioxide (CMD) according to the Sedema process [13]

Chemetals synthesis of high-purity manganese dioxide (see Section 2.1.3.1). The $Mn(NO_3)_2$ solution thus obtained is treated with ammonium carbonate to give manganese carbonate (processed as before) and ammonium nitrate. The latter byproduct is more readily marketed as a fertilizer than the previously obtained manganese sulfate [15].

Electrolytic manganese dioxide (EMD) is the most important of the synthetic manganese dioxides even though it is a relatively new commodity. Practically all the EMD produced is used in the manufacture of dry cells and electronic materials such as ferrites.

The electrolytic preparation of MnO_2 was discovered during the first half of the 19th century, but the outstanding suitability of the product as a depolarizer for dry cell batteries was not recognized until 1918 [16]. Commercial production started in Japan and the United States during the 1930 s. By about 1952 the Japanese industry had perfected their technology to such an extent that the performance of dry cells made with EMD was about two to three times that of cells made with native battery ore [17]. Ever since, the Japanese EMD industry has maintained its world leadership in this field.

Electrolytic manganese dioxide is a black powder and typically contains 91 wt % MnO_2 (mostly the hexagonal ε-modification [18]), 3 – 5 wt % moisture, up to 1.3 wt % sulfate, < 0.02 wt % Fe, and very low residual concentrations (< 0.001 wt %) of such metals as Pb, Cu, and Co. The balance consists largely of lower manganese oxides. The true density of EMD ranges from 4.0 to 4.3 g/cm^3, and its tap density from 2.2 to 2.3 g/cm^3. The BET surface area is 40 – 50 m^2/g [17] with a particle size of < 74 µm (< 200 mesh).

The outstanding performance of EMD as a battery depolarizer (especially in high-performance alkaline cells) is a result of its unique solid state properties (which permit free transport of protons through the lattice) and the virtual absence of elements that promote the corrosion of zinc metal [19].

Production of EMD usually starts from manganese dioxide ore [13], [20, p. 385], [21, p. 637], [22] (Fig. 3), which is first reduced to MnO with either coal, heavy fuel oil, hydrogen, or natural gas. The MnO is then leached with sulfuric acid to form a manganese sulfate solution. Alternatively, if the rhodochrosite (manganese carbonate) is the starting material (as in Japan), only a leaching step with sulfuric acid is required. The acidic manganese sulfate solution is then purified by oxidation with MnO_2 and neutralized with lime to a pH of 4 – 6. This treatment precipitates any heavy-metal ions present (Fe, Pb, Ni, Co). Hydrogen sulfide may be added for even more complete purification. After filtration and adjustment of concentrations to 75 – 160 g/L $MnSO_4$ and 50 – 100 g/L H_2SO_4, the manganese sulfate solution is subjected to electrolysis, which can be represented in simplified form by the following reactions:

$$\text{Anode:}\quad Mn^{2+} + 2\,H_2O \rightarrow MnO_2 + 4\,H^+ + 2\,e^-$$
$$\text{Cathode:}\quad 2\,H^+ + 2\,e^- \rightarrow H_2$$
$$\text{Overall:}\quad Mn^{2+} + 2\,H_2O \rightarrow MnO_2 + 2\,H^+ + H_2$$

The electrolytic cells generally consist of open steel troughs equipped with an acid-resistant, electrically nonconductive lining (e.g., Hypalon, rubber, or ceramic). The anode is made from graphite, hard lead, or titanium. Anode current density is 70 – 120 A/m^2 at a cell voltage of 2.2 – 3.0 V. During the electrolysis, which is carried out at 90 – 98 °C (a paraffin layer is used to minimize evaporation), the concentration of $MnSO_4$ decreases and that of H_2SO_4 increases. Current yields range from 70 % in older installations to > 90 % in modern plants. In practically all commercial processes the EMD is deposited as a solid coating on the anode.

Electrolysis is terminated when the EMD deposit is 20 – 30 mm thick (requiring 14 – 20 d). The product is then mechanically removed from the anode, crushed, repeatedly washed with hot water, dried, and ground to the desired particle size (which depends upon the grade). The ground product is resuspended in water, adjusted to pH 6.5 – 7.0 with alkali, dried at 85 °C to a defined residual water content, and packaged.

Much of the recent progress toward economical production of high-quality EMD is related to technical advances in electrolytic cell (especially electrode) technology. Nevertheless, many EMD producers still use graphite and lead anodes (Table 2). The main disadvantage of graphite anodes lies in their relatively short life; lead anodes lead to undesirable product contamination. The modern trend is toward titanium an-

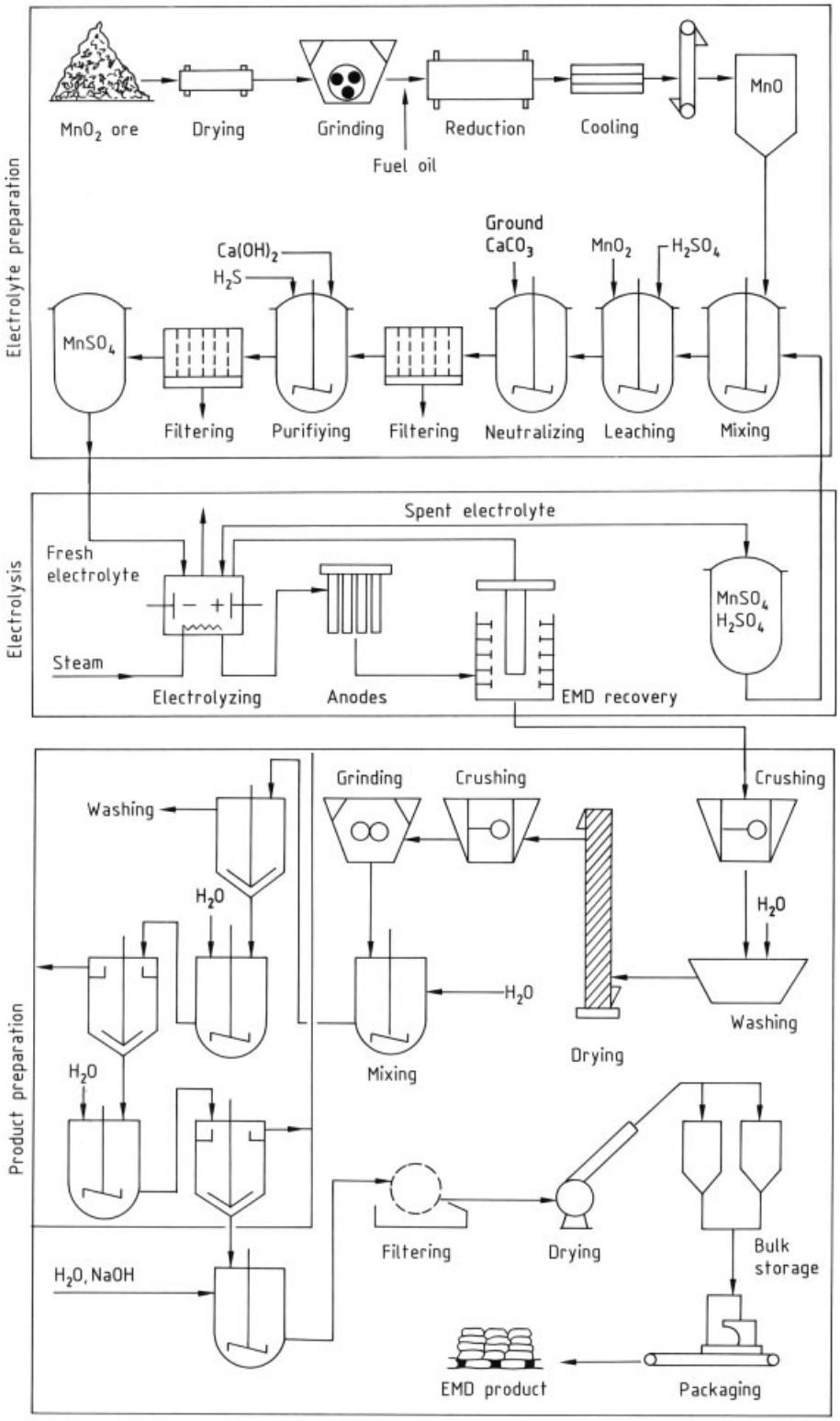

Figure 3. Production of electrolytic manganese dioxide (EMD) according to the process used by Tekkosha, Greece [13]

Table 2. Major producers of electrolytic manganese dioxide (EMD)

Country	Producer	Location	Production capacity, 10^3 t/a∗	Anode material [23]
Japan	Toyo Soda	Hyuga	24	Ti
Japan	Mitsui Mining and Smelting	Takehara	24	Ti
Japan	Dai-Ichi Carbon Company	Yokohama	6	C
Japan	Japan Metals and Chemicals Company	Takaoka	18	Ti
Greece	Tekkosha Hellas	Thessaloniki	15	C
Ireland	Mitsui Denman	Cork	12	Ti
United States	Kerr McGee	Henderson, NV	14.5	C
United States	Eveready Battery Co.	Marietta, OH	10	C
United States	ESB Materials Company	Covington, TN	3	C
United States	Chemetals, Inc.	New Johnsonville, TN	16	Ti
Spain	Cegasa	Onate	5	Pb
Soviet Union	state	Rustavi	5	Pb
People's Republic of China	state		8	C
India	Union Carbide	Thaha	5	C
India	T.K. Chemical	Trivandrum	1	C
Brazil	Union Carbide	Itapercerica	4	C
South Africa	Delta	Nelspruit	12	Pb
Total world capacity			194.5	

∗ Compiled by Hoechst AG, 1987.[86]

odes, which are mechanically and chemically stable and do not cause product contamination. One of the apparent limitations of titanium anodes is the formation of a passivated layer on the electrode surface at current densities $> 80-90\ A/m^2$, but this problem has been overcome by the addition of a mixture of finely ground manganese oxides (MnO_2, Mn_2O_3, Mn_3O_4) to the electrolyte [24]. Cells are now operated at current densities $\geq 150\ A/m^2$, resulting in a substantial increase in productivity.

Economic Aspects. World capacity for CMD is estimated to be ca. 40 000 t/a. This includes the Sedema Division of Sadacem in Belgium, Chemetals Corporation in the United States, and several plants in the former Soviet Union. Actual consumption is about 25 000 t/a [25]. Even though CMD has its own established niches in the battery industry, it also competes increasingly with EMD as a depolarizer in high-performance cells.

Electrochemical manganese dioxide is now the most important fastest growing synthetic manganese product due to the high growth rate of the dry-cell battery and electronic component markets. World production capacity for EMD is approximately 183 000 t/a (Table 2) and growing. The EMD capacities in Japan and the United States are 73 000 and 32 000 t/a, respectively. Additionally, two Japanese companies operate EMD plants in Ireland and Greece, with a combined capacity of ca. 25 000 t/a.

In view of the rapid growth of the EMD market in the 1980 s (averaging about 10 %/a), considerable interest has been stimulated in establishing new EMD production plants. Thus, announcements regarding the construction of new facilities were recently made by BHP in Australia (15 000 t/a), Metalman in Brazil (12 000 t/a), and CIA Minera Autlan in Mexico (13 200 t/a).

According to [26], the 1986 world demand for EMD was about 128 000 t/a, and current production is said to be 165 000 t/a [25]. Serious EMD oversupply and fierce competition among producers is already evident, and is expected to intensify in the future.

2.2. Other Manganese Oxides

Manganese(II) Oxide. Manganese(II) oxide [*1344-43-0*], manganese monoxide, MnO, M_r 70.94, green, ϱ 5.37 g/cm^3, *mp* 1945 °C, is practically insoluble in water and occurs in nature as manganosite. It dissolves readily in most acids. Synthetic manganese monoxide is obtained by reduction of manganese dioxide or by thermal decomposition of Mn(II) carbonate

under the exclusion of air. Depending on particle size (a function of the method of preparation), MnO reacts at varying rates with atmospheric oxygen even at room temperature, thereby forming Mn_3O_4.

Manganese monoxide has major significance as an intermediate in the manufacture of Mn(II) compounds (especially manganese sulfate and sequential products) from MnO_2 ores. Reducing agents for MnO_2 in the large-scale manufacture of manganese monoxide include finely powdered coal, hydrogen, carbon monoxide, natural gas, and heavy fuel oils. The reaction is carried out at 400 – 800 °C in a rotary kiln or shaft furnace. The freshly formed manganese oxide must be allowed to cool in a reducing atmosphere in order to ensure room-temperature stability with respect to atmospheric oxygen.

Manganese monoxide (in contrast to manganese dioxide) is readily assimilated by most plants. This fact accounts for its widespread use as a fertilizer for manganese-deficient soils (→ Fertilizers, → Fertilizers, Chap. 4.3.1.).

The consumption of manganese monoxide in the United States for fertilizer applications (alone or in combination with $MnSO_4$) is about 20 000 t/a. A similar amount is used as an animal feed additive (→ Foods 1. Survey, Chap. 3.5., → Foods 1. Survey, Chap. 6.3.). Fertilizer/feed grade manganese monoxide (containing 77 wt % MnO) is made by direct reduction of manganese ore.

High-purity manganese monoxide (obtained, for example, by reduction of 99.5 % MnO_2) is used in the production of specialty ceramics and glasses for electronic applications, in ferrites and thermistors, and for making welding rod fluxes and high-purity manganese chemicals.

Manganese(III) Oxide. Manganese(III) oxide [*1317-34-6*], Mn_2O_3, M_r 157.86, decomp. > 900 °C, ϱ 4.89 g/cm^3, exists in α- (rhombic or cubic) and γ- (tetragonal) modifications. The hydrate $Mn_2O_3 \cdot H_2O$ [*1332-63-4*] occurs as the mineral manganite and forms steel-gray, shiny, rhombic crystals. Manganese(III) oxide is made industrially from manganese dioxide by calcination at 600 – 800 °C, or by thermal decomposition and controlled air oxidation of manganese carbonate. Primary uses are the preparation of such electronic materials as ceramic magnets and semiconductors. It is a small-volume product, a few hundred tonnes are consumed annually.

Manganese(II, III) Oxide. Manganese(II, III) oxide [*1317-35-7*], Mn_3O_4, M_r 228.79, *mp* 1562 °C, occurs naturally as hausmannite and crystallizes tetragonally (ϱ 4.84 g/cm^3). It may be obtained from the other manganese oxides by heating above 950 °C in the presence of air. According to [27], high-purity Mn_3O_4 can also be made from an aqueous suspension of finely divided manganese metal by air oxidation at 30 – 100 °C in the presence of ammonium salts. The compound forms black crystals, but finely dispersed it appears as a red powder. Highly purified Mn_3O_4 is used in the manufacture of semiconductors and ceramic magnets. World production is estimated at ca. 2000 t/a.

Other Manganese Oxides. Oxides with penta-, hexa-, or heptavalent manganese have no industrial significance. Indeed, the hypothetical oxides Mn_2O_5 and MnO_3 have never been prepared. The anhydride of permanganic acid, Mn_2O_7, forms readily from $KMnO_4$ and concentrated sulfuric acid. It is an unstable, highly explosive, green-black liquid.

3. Manganese(II) Salts

Manganese acetate [*15243-27-3*], $Mn(CH_3COO)_2 \cdot 4\,H_2O$, M_r 245.08, ϱ 1.589 g/cm^3, forms pink crystals that are soluble in water, methanol, and ethanol. The compound dehydrates between 80 °C and 130 °C, and decomposes at 350 °C to Mn_2O_3. It is made industrially from acetic acid and either manganese carbonate, manganese(II) oxide, or electrolytic manganese metal. Its main use is as a catalyst (either alone or in combination with cobalt) in the liquid-phase air oxidation of hydrocarbons to carboxylic acids.

Manganese borate [*12228-91-0*], $MnB_4O_7 \cdot 8\,H_2O$, M_r 354.34, is a white to pale reddish-white solid. It is insoluble in water and ethanol, but soluble in dilute acids. The compound is precipitated from aqueous solutions of manganese chloride and sodium borate, and is used as a siccative.

Manganese(II) carbonate [*598-62-9*], $MnCO_3$, M_r 114.94, light pink, decomp. > 200 °C, ϱ 3.125 g/cm^3, occurs naturally as rhodochrosite. It is sparingly soluble in water (solubility product at 25 °C: 8.8×10^{-11}) but readily soluble in acids. It is produced commercially from manganese sulfate by precipitation with alkali-metal carbonates or hydrogen carbonates. If the presence of alkali-metal ions must be avoided in the product (as in the manufacture of ferrites), ammonium hydrogen carbonate may be used as the precipitating agent. According to a GDR patent [28], the use of ferromanganese as a starting material results in manganese carbonate of especially high purity. The precipitated $MnCO_3$ is filtered, washed, and dried at 110–120 °C. Heating above 200 °C causes decomposition to MnO and CO_2.

Manganese carbonate is used in the synthesis of other manganese(II) salts (by reaction with the corresponding acids) and also in the production of battery-active, chemical manganese dioxide (CMD) (see page 4). Manganese carbonate is also important in the preparation of high-quality manganese – zinc ferrites for the television and computer industries [4, p. 180]. World production of $MnCO_3$ is estimated to be about 9000–10 000 t/a.

Manganese(II) chloride [*7773-01-5*], $MnCl_2$, M_r 125.84, ϱ 2.977 g/cm^3, *mp* 690 °C, *bp* 1190 °C, is readily soluble in water and exists in the anhydrous form and as di-, tetra-, and hexahydrates. Unless precautions are taken (see below), thermal dehydration of the hydrates to the anhydrous form may result in some hydrolytic decomposition, with formation of gaseous HCl and manganese oxychloride.

Industrial preparation of manganese chloride is based either on the reaction of aqueous hydrochloric acid with MnO_2 ore, MnO, $MnCO_3$ (rhodochrosite), or on direct chlorination of manganese metal or ferromanganese. The process formerly employed in Bitterfeld (GDR) started from manganese ore (MnO_2). Chlorine gas is generated

$$MnO_2 + 4\,HCl \rightarrow MnCl_2 + Cl_2 + 2\,H_2O$$

and is absorbed in a suspension of hydrated lime [6, p. 739]. An alternative process used by Chemetals avoids the formation of chlorine by prereduction of the MnO_2 to MnO.

The $MnCl_2$ solution thus obtained is purified by neutralization with $MnCO_3$, causing iron and aluminum impurities to precipitate. After filtration and evaporation, the tetrahydrate crystallizes from the cooled solution in plate-like crystals. This material, which is itself a commercial product, requires very careful drying because it melts at 58 °C in its own water of crystallization. Dehydration requires a temperature above 200 °C. Fusion in the oven is avoided by addition of previously dehydrated product. The drying process may be carried out in the presence of excess dry HCl gas if anhydrous manganese chloride free of oxychlorides is required.

Uses of manganese(II) chloride include dry cell manufacture, preparation of hard and corrosion-resistant magnesium alloys, synthesis of methylcyclopentadienylmanganese tricarbonyl (Chap. 5) and brick coloring [4, p. 182].

Manganese(II) nitrate [*10377-66-9*], $Mn(NO_3)_2$, M_r 169.02, decomp. > 140 °C, exists in the anhydrous form and as mono-, tri-, tetra-, and hexahydrates. The commercial product is the tetrahydrate [*20694-39-7*], which is readily soluble in water. Manganese nitrate is prepared by nitric acid treatment of manganese carbonate or MnO, or by reaction of MnO_2 with nitric oxides. The compound is an important intermediate in the production of high-purity manganese oxides and is also used in the preparation of colorants for the ceramics industry.

Manganese(II) Phosphates. The most important manganese phosphate is the water-soluble $Mn(H_2PO_4)_2\cdot 2\,H_2O$ [*18718-07-5*], which is prepared by dissolving manganese carbonate, manganese(II) oxide, or manganese metal in phosphoric acid. This phosphate finds wide application in the phosphatizing of ferrous metals and is also used as a light stabilizer for polyamide fibers.

Manganese(II) sulfate [*7785-87-7*], $MnSO_4$, M_r 151.01, ϱ 3.25 g/cm^3, *mp* 700 °C, *bp* 850 °C (decomp.), is almost pure white in the anhydrous state; the corresponding mono-, tetra-, penta-, and heptahydrates are pink. Manganese(II) sulfate is readily soluble in water, but has a strongly negative solubility coefficient at temperatures > 24 °C. The solids content of a

manganese sulfate solution saturated at 24 °C is 39.3 wt %; at 100 °C the solids content drops to 26.2 wt %.

The customary commercial product (technical grade) is the monohydrate, containing 99 wt % $MnSO_4 \cdot H_2O$. Manganese sulfate is produced either directly by reaction of manganese(II) oxide (or manganese carbonate) with sulfuric acid or as a byproduct from organic oxidations with manganese dioxide.

Until recently the manufacture of hydroquinone (by oxidation of aniline with MnO_2 and H_2SO_4) was a major source of fertilizer-grade manganese sulfate, containing 78 wt % $MnSO_4$ [29], but modern hydroquinone processes are based on either diisopropylbenzene/O_2 or phenol and H_2O_2 (→ Hydroquinone).

Before 1977 Tennessee Eastman, a major U.S. hydroquinone producer, operated with a byproduct manganese sulfate capacity of about 35 000 t/a. Eastman's manufacturing technology has since been changed, and manganese sulfate is no longer produced. Other organic oxidations are still carried out with MnO_2 as the oxidant (e.g., the manufacture of 4-*tert*-butylbenzaldehyde and *p*-anisaldehyde), but the output of manganese sulfate from these processes is no more than a few thousand tonnes per year, so the majority of U.S. manganese sulfate is now derived directly from MnO or $MnCO_3$.

Manganese sulfate is commercially one of the most important manganese products. It is the starting material for electrolytic manganese metal, electrolytic manganese dioxide, fungicides such as Maneb (→ Fungicides), and other manganese compounds including manganese carbonate, manganese soaps (naphthenate, linolate, and resinate, all used as siccatives), and certain inorganic pigments. Manganese sulfate is also used in textile printing and glass making. Very significant in terms of volume is the use of manganese sulfate as a fertilizer for manganese-deficient soils, for example, in the vegetable- and citrus-growing areas of Florida. It is also valuable as a micronutrient additive for animal feeds [4, p. 187].

World production of commercial manganese sulfate is estimated to be about 120 000 – 130 000 t/a. Production in the United States is ca. 8000 t/a, with a demand of approximately 14 000 – 14 500 t/a. The deficit is covered by imports from Europe and the People's Republic of China.

Major producers of manganese sulfate include: Eagle Picher (USA), production capacity 6000 t/a; Sulfamex (Mexico) 21 000 t/a (start-up delayed); and Sedema (Belgium) 22 000 t/a. For more economic information on manganese sulfate, see [30].

4. Higher Oxidation-State Manganates

4.1. Potassium Manganates(V) and (VI)

Potassium manganate(V) [*12142-41-5*], K_3MnO_4, M_r 236.24, ϱ 2.78 g/cm^3, decomp. > 1000 °C, occurs as fine blue-green to turquoise crystals. In the presence of water it is readily hydrolyzed, disproportionating to K_2MnO_4, $KMnO_4$, and MnO_2. Solutions of K_3MnO_4 in 40 % potassium hydroxide have limited stability even below − 10 °C; however, in the presence of 75 % KOH and under nitrogen, potassium manganate(V) is stable up to 240 °C. Pure potassium manganate(V) can be heated to over 900 °C without decomposition. It is an important intermediate in the production of potassium permanganate.

Potassium manganate(VI) [*10294-64-1*], K_2MnO_4, M_r 197.14, ϱ 2.703 g/cm^3, decomp. 640 – 680 °C, exists in the pure state as brilliantly green, rhombic crystals. Upon aging, the color changes to black-green due to the formation of a surface film of $KMnO_4$. The compound is less prone to decomposition than K_3MnO_4; the deep green solutions in 15 – 20 % potassium hydroxide (for solubilities in KOH see [31]) are quite stable even at 50 – 60 °C. In more dilute solutions, and especially in the presence of acid, K_2MnO_4 disproportionates to $KMnO_4$ and MnO_2. Pure potassium manganate(VI) is stable up to ca. 600 °C; at higher temperatures it loses oxygen to form K_3MnO_4 and Mn_2O_3. By far the largest part of the K_2MnO_4 produced industrially is converted to potassium permanganate; smaller quantities are used in the surface treatment of magnesium metal [32] and as a chemical oxidant [5, p. 175].

4.2. Potassium Permanganate

The best known of the manganates, potassium manganate(VII) (potassium permanganate) was introduced as a commercial product in 1862 [33]. It soon became a very important chemical [34] not only industrially but also in the eyes of many generations of academic and industrial researchers. The former concerned themselves with the chemistry of $KMnO_4$ and its precursors, while the latter devised new and improved methods for its production. Unraveling the chemistry of the higher valence states of manganese (generally referred to as manganates) proved to be a major challenge. Since 1954 the roles of the various manganates [especially manganate(V)] in the industrial production of $KMnO_4$ have been fully clarified [35]. For more detailed historical information, see [36].

4.2.1. Properties

Potassium manganate(VII) [*7722-64-7*], $KMnO_4$, M_r 158.03, ϱ 2.703 g/cm^3, decomp. $>200\,°C$, gives dark purple to bronze-colored rhombic crystals. Its solutions are faint pink to deep violet, depending on concentration. The solids content of a saturated aqueous solution at 20 °C is ca. 6 wt % $KMnO_4$ (11 wt % at 40 °C and 20 wt % at 65 °C). Permanganate solutions are most stable in the neutral or near-neutral pH region; acidic or alkaline solutions decompose with loss of oxygen. For solubilities in aqueous KOH see [31]. Potassium permanganate is also soluble in several organic solvents, including acetone, glacial acetic acid, methanol, and sulfolane (tetrahydrothiophene-1,1-dioxide). Such solutions usually have limited stability because the permanganate ion slowly attacks the solvent.

When dry potassium permanganate is heated to 200–300 °C, an exothermic, autocatalytic decomposition takes place with the evolution of oxygen. Contact with combustibles, especially water-soluble organic substances (such as polyhydroxy compounds), can lead to spontaneous ignition and potentially violent combustion.

The usefulness of permanganate as an oxidant is a result of several unique properties. One is the compound's ability to function as an oxidant at all pHs. Also interesting is the fact that permanganate oxidations can be carried out in both aqueous and several nonaqueous solvents. Phase-transfer catalysis greatly extends the range of usable solvents [37].

Permanganate oxidizes a wide variety of inorganic and organic compounds (→ Oxidation). It is often regarded as the oxidant of choice for olefins and aromatic side chains even though it can also interact with other functional groups [5, pp. 183–288], [38]. Permanganate oxidations can be used not only for specific modification of organic molecules but also to destroy such materials as in pollution abatement applications.

In moderately alkaline, neutral, or slightly acidic media, $KMnO_4$ donates active oxygen according to the following equation:

$$2\,MnO_4^- + H_2O \rightarrow 2\,OH^- + 2\,MnO_2 + 3\,[O]$$

Only under substantially more acidic conditions, and with particular substrates, does permanganate oxidation follow an alternative course:

$$2\,MnO_4^- + 5\,H_2O + 6\,H^+ \rightarrow 2\,Mn^{2+} + 8\,H_2O + 5\,[O]$$

4.2.2. Production [6, pp. 741–762]

Most permanganate is made from MnO_2-containing ore by fusion followed by electrolysis. However, all-electrolytic methods also exist. One such method, practiced in the Soviet Union since the late 1950 s, involves direct single-step anodic oxidation of ferromanganese to permanganate [39]:

$$2\,Mn + 2\,OH^- + 6\,H_2O \rightarrow 2\,MnO_4^- + 7\,H_2$$

This process is not generally considered economical because of high energy consumption (about 15 kWh/kg), the high cost of preparing cast ferromanganese anodes, and the cost of cooling the electrolyte to the required operating temperature (20 °C).

In another patented all-electrolytic process [40] a 10–25 wt % suspension of precipitated manganese dioxide in aqueous KOH is electrolyzed at $\geq 60\,°C$:

$$MnO_2 + KOH + H_2O \rightarrow KMnO_4 + 3/2\,H_2$$

It is not clear whether this process is, or has been, used commercially.

4.2.2.1. Fusion Processes [41]

The industrial preparation of potassium permanganate and its precursors (K_3MnO_4 and K_2MnO_4) begins with a finely ground MnO_2 ore which is subjected to alkaline oxidative fusion:

$$2\,MnO_2 + 6\,KOH + 1/2\,O_2 \rightarrow 2\,K_3MnO_4 + 3\,H_2O \quad (1)$$
$$2\,K_3MnO_4 + H_2O + 1/2\,O_2 \rightleftharpoons 2\,K_2MnO_4 + 2\,KOH \quad (2)$$
$$MnO_2 + 2\,KOH + 1/2\,O_2 \rightarrow K_2MnO_4 + H_2O \quad (3)$$

In a separate step, potassium manganate(VI) is subsequently converted to potassium permanganate, usually by anodic oxidation (see Section 4.2.2.2).

Several continuous or batch-type fusion processes are in current use, but they may be categorized in two groups:

1) *Roasting processes* that employ a MnO_2: KOH molar ratio of 1 : 2 to 1 : 3 (the reaction mixture is a solid).
2) *Liquid-phase processes* with a MnO_2: KOH molar ratio $\geq$ 1 : 5.

Roasting processes usually involve two steps, with reactions (1) and (2) carried out sequentially. In the primary step, finely ground manganese dioxide is mixed with 50 % KOH (frequently in a highly dispersed form) and treated with air or oxygen at 300 – 400 °C. The KOH/MnO_2 reaction mixture is usually introduced as a fine spray into internally or externally heated rotary kilns (see Fig. 4). Stationary spray drier-like reactors are also in use. Most of the water quickly evaporates, and the highly concentrated KOH reacts within a few minutes with MnO_2 and oxygen to form K_3MnO_4 (Eq. 1). The dry, blue-green reaction product is ground to increase its surface area and then subjected to secondary oxidation (Eq. 2), in which it is exposed to a moisture-laden air or oxygen stream at 190 – 210 °C. Several hours are required to effect nearly complete conversion to the black-green K_2MnO_4.

Rotary kilns are commonly employed for this step, but other types of solid – gas contactor (e.g., specially designed "plate" reactors) are also utilized. The removal of hard crusts that form on the inner walls of the reactor can cause substantial downtime.

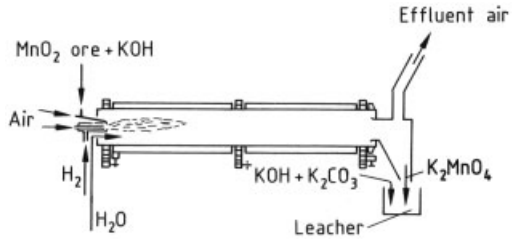

Figure 4. Rotary kiln with internal heating for the production of potassium manganate(VI)

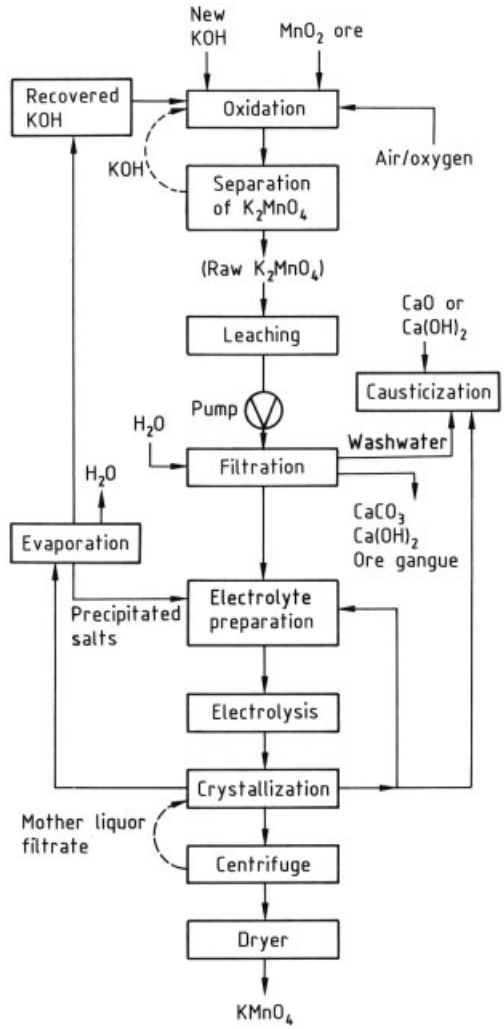

Figure 5. Production of potassium permanganate: liquid-phase oxidation process (overall flow schematic)

Liquid-Phase Processes. In liquid-phase processes (see Fig. 5) finely ground MnO_2 ore is added to an excess of concentrated, molten

potassium hydroxide (70 – 90 wt %) at 200 – 350 °C. The reaction mixture is agitated vigorously and a stream of finely dispersed air or oxygen is passed through the molten mass. The MnO_2 is converted to K_2MnO_4 in a single step (Eq. 3), although such an oxidation can also be effected in two stages [42–44]. Potassium manganate(VI) crystallizes from the melt, from which it must then be separated by suitable means (e.g., thickening, decantation, filtration, and centrifugation) [45, p. 870].

4.2.2.2. Anodic Oxidation of Manganate(VI)

Oxidation of the manganate(VI) to permanganate is always accomplished by electrolysis:

$$K_2MnO_4 + H_2O \rightarrow KMnO_4 + KOH + 1/2\,H_2 \quad (4)$$

Crude manganate from the fusion process is first leached in dilute potassium hydroxide (90 – 250 g KOH per liter, depending upon the nature of the subsequent electrolysis). The resulting leach solution is usually filtered to separate insolubles, originating from the gangue portion of the ore (see Fig. 5).

Effective control of key parameters is decisive for the overall success of the electrolysis, including the concentrations of K_2MnO_4, $KMnO_4$, and KOH, as well as the temperature (usually ca. 60 °C).

Cell Design. Typical electrolysis cells employ voltages of 2.3 – 3.8 V and anodic current densities of 50 – 1500 A/m^2. Current yield depends on cell design and mode of operation, and ranges from 40 – 90 %. Nickel or Monel Metal are the preferred materials for the anode, whereas iron is usually the metal of choice for the cathode.

Cells may be either monopolar or bipolar, rectangular or circular, and may be designed for either flow-through or batch operation.

Cells used by the Chemiekombinat Bitterfeld in the German Democratic Republic [6, p. 749] (see Fig. 6), for example, have a volume of about 4 m^3, are circular in shape, and are equipped with a built-in mechanical agitator. The electrode system is monopolar. These batch-operated cells use unfiltered electrolyte; the resulting $KMnO_4$ is allowed to crystallize within the cells and is drawn off periodically through a bottom valve. The crude product must be recrystallized to separate it from residual ore gangue.

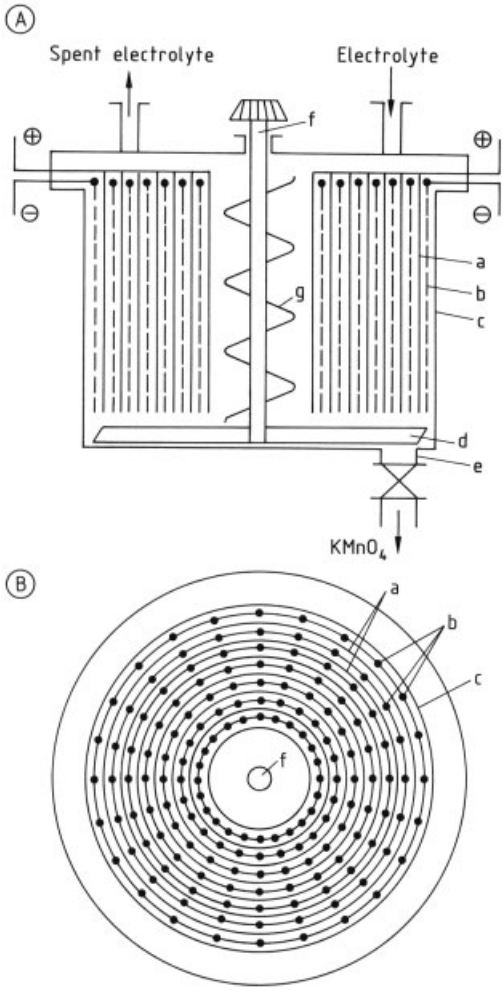

Figure 6. Electrolytic cell used in the Bitterfeld (GDR) process for the production of potassium permanganate
A) Vertical section; B) Cross sectiona) Anode; b) Cathode; c) Rubber-lined cell trough; d) Agitator; e) Bottom discharge valve; f) Agitator shaft; g) Spiral agitator

In the cell developed by the Carus Chemical Company, numerous bipolar electrodes are combined in an arrangement resembling a filter press to form a closed, diaphragmless, flow-through electrolyzer (Fig. 7) [46–48]. The anodic side of each cell sheet consists of Monel Metal screens, while the effective cathode area consists of a multitude of small steel protrusions. The rest of the cathode is covered with a corrosion-resistant

insulating material. A lower surface area for the cathode than for the anode maximizes the anodic oxidation of MnO_4^{2-} and minimizes undesirable cathodic reduction of both manganate and permanganate. The resulting potassium permanganate is crystallized outside the Carus cell in a continuous crystallizer system.

Crystallization. Crystallization of potassium permanganate is effected either directly in a specially designed electrolytic cell or in separate crystallizers. Modern installations use single- or multistage vacuum crystallizers, which produce a crystalline product directly from the mother liquor in such purity that only centrifugation and drying are required before the $KMnO_4$ is packaged and sold (see Fig. 8).

Removal of Impurities and Recycling. When the crude manganate is leached with dilute KOH a solid residue remains. This is composed of the insoluble portion of the ore gangue (e.g., iron oxide hydrates, precipitated aluminum silicate), unreacted MnO_2 ore, and some precipitated MnO_2 (from disproportionation of K_2MnO_4). These insoluble constituents are removed by the use of thickeners, vacuum filters, or filter presses. The Bitterfeld process operates with unfiltered electrolyte, but here the electrolyzers have been specially designed to cope with suspended matter.

In order to recycle potassium hydroxide generated in the electrolysis step (Eq. 4 in Section 4.2.2.2), the spent mother liquor is first concentrated by evaporation to a relative density of ca. 1.575 – 1.585. This "recovered" caustic potash is then recycled to the front end of the process and added to either the roaster or the liquid-phase oxidizer. Alternatively, after removal of dissolved impurities, the recovered KOH can be treated with carbon dioxide to produce potassium carbonate. This method has the drawback that there is a limited market for the substantial quantities of potassium carbonate coproduct.

Dissolved impurities (in particular silicate and aluminate from the gangue portion of the ore) must be removed from the process liquors in order to avoid formation of rock-like deposits on pipeline and equipment walls. Equally undesirable is the buildup of carbonate in the process liquors due to reaction of atmospheric CO_2 with KOH. Excessive carbonate concentrations in the recycle system reduce the rate of conversion of ore into manganates and can also cause low current efficiencies in the electrolytic cell.

All three major impurities (Si, Al, and carbonate) can be removed together with such minor impurities as Cu, Pb, Ni, and Co by treatment with calcium hydroxide [49]. Effective causticization of the potassium carbonate occurs according to

$$K_2CO_3 + Ca(OH)_2 \rightleftharpoons 2\,KOH + CaCO_3$$

and requires that the potassium hydroxide concentration be kept ≤ 80 g/L. The process used by Nippon Chemical Industrial Company [50] entails addition of $Ca(OH)_2$ as early as the roasting step to prevent formation of soluble carbonate and the solubilization of other impurities.

Waste Disposal. The treatment and disposal of solid wastes from permanganate production is an ecological as well as an economic problem. The waste product contains not only unreacted ore but also some precipitated MnO_2. The main constituents are derived from the insoluble portion of the ore gangue (Fe, Si, Al compounds) but these are accompanied by significant quantities of potassium and calcium salts. One permanganate producer has proposed recovering most of the adsorbed and chemically-bound potassium by treatment with calcium hydroxide. After washing, the treated waste could then be deposited in landfills [51]. The patent literature describes a number of potential uses for the waste from permanganate production, including extraction of residual manganese, or use of the waste as an absorbent for air pollutants [52], as a micronutrient additive for fertilizers [53], or in building materials [54].

4.2.2.3. Production Plants

This section gives a brief review of production plants for potassium permanganate in eight countries. Most of the reported production capacities are estimates.

India. Permanganate production began in India after World War II. Only one plant remains, that of Curti Chemicals (Goa) built in 1970 with a capacity of less than 1000 t/a. It uses a roasting step followed by electrolytic oxidation.

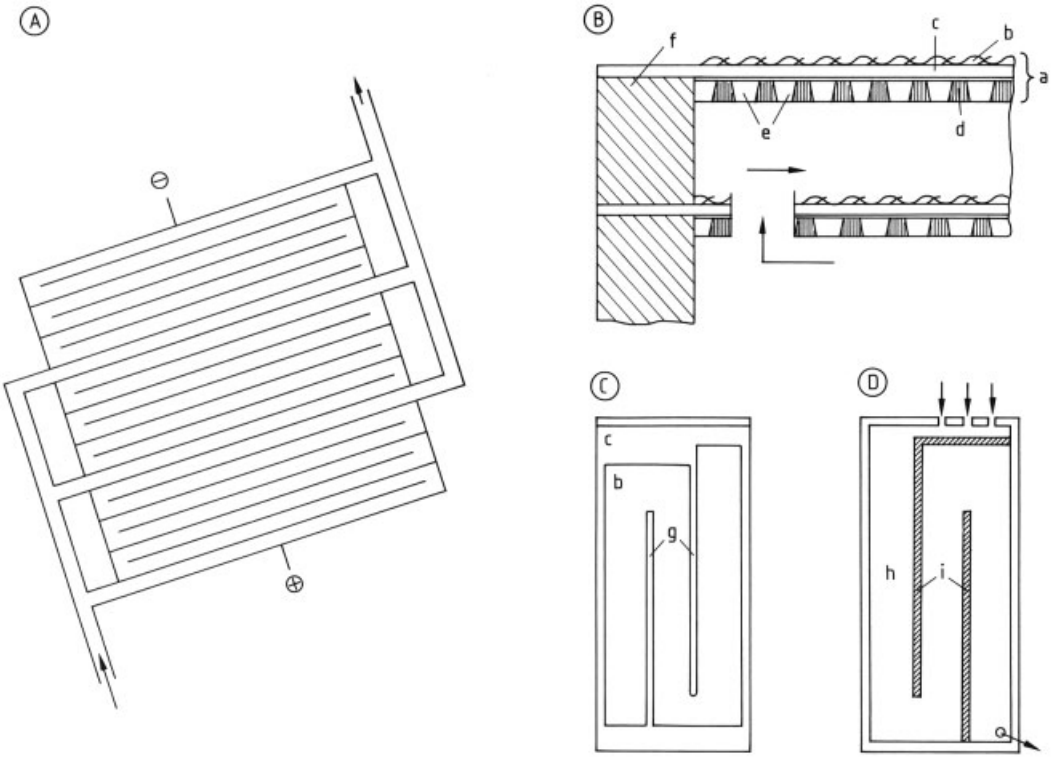

Figure 7. Carus Chemical electrolytic cell for the production of potassium permanganate from potassium manganate (VI) [47] A) Schematic overview of the cell bank; B) Side elevation of an individual cell section; C) Cross section of the anode side of a bipolar electrode; D) Cross section of the cathode side of a bipolar electrode a) Individual bipolar electrode; b) Anode (wire screen of Monel Metal); c) Base steel sheet; d) Steel cathode with perpendicular projections; e) Insulating plastic; f) Insulating spacer; g) Cutout for flow dividers; h) Cathode projections embedded in plastic; i) Flow dividers

Spain. The Spanish permanganate plant in Trubia was established in 1940 and modernized in 1977. The plant is operated by Asturquimica and capacity is 4000 t/a. A modified spray drier is used as the reactor in the first roasting step and a "plate" reactor for the second roasting step [55]. Electrolytic oxidation and vacuum crystallization occur continuously.

Japan. Nippon Chemical constructed a new, continuously operating plant in Aichi Prefectur in 1971 (capacity ca. 3000 t/a). An improved roasting process is employed for making the manganate intermediate, whereby the MnO_2–KOH reaction mixture containing hydrated lime is dewatered in a separate step before it is introduced into the two-stage roaster system [50]. The electrolysis cells are arranged in cascade fashion and overflow from the final stage is passed to a vacuum crystallizer.

Former Soviet Union. The Soviet Union operates three plants, built in the late 1950 s and early 1960 s. The plant in Rustavi has a capacity of 500 t/a and employs single-step anodic oxidation of ferromanganese [45, p. 875]. The second plant, in Saki, has a capacity of about 2500 t/a and uses technology similar to that of the Bitterfeld plant in the GDR. The third plant is also in Saki and has a capacity of about 3000 t/a. It uses discontinuous liquid-phase oxidation in the manganate step [45, p. 870], followed by batch electrolysis and crystallization.

Former Czechoslovakia. The Czech plant (Spolek pro Chimickou a Hutni Vyrobu n.p., Usti n.L.) was built in the 19 th century. It uses a two-stage roasting process with rotary hearth furnaces for the manganate step. Conversion of each batch of MnO_2 to K_2MnO_4 requires ca.

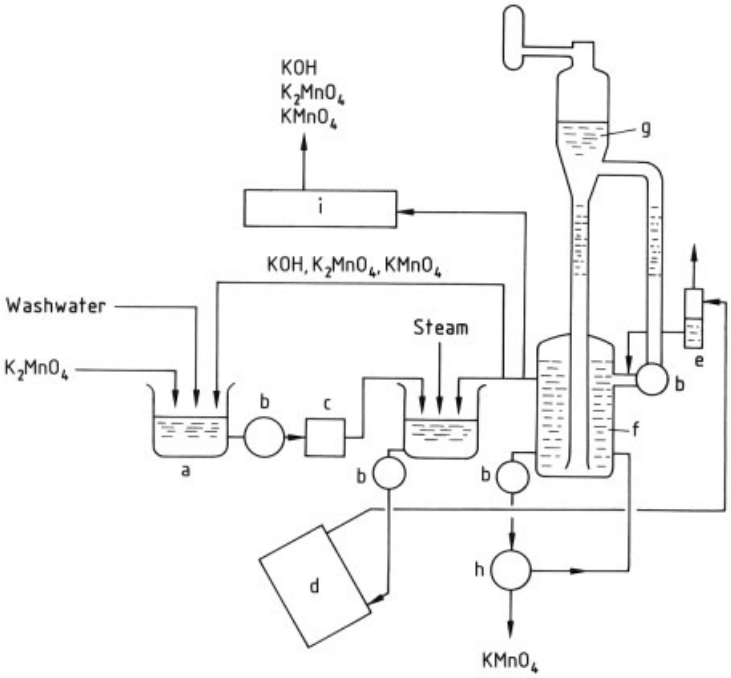

Figure 8. Production of potassium permanganate: continuous electrolyzing – crystallizing system
a) Leach tank; b) Pump; c) Filter; d) Electrolysis cell; e) Gas separator; f) Crystallizer; g) Vacuum evaporator; h) Centrifuge; i) Evaporator

48 h. The monopolar, continuously operating electrolytic cells contain asbestos diaphragms and electrodes. Plant capacity in 1981 was stated to be 2000 t/a [56].

People's Republic of China. China, with ten state-owned plants, has become a major factor in the world permanganate market, especially as an exporter. Estimates of production capacity range from 7000 – 15 000 t/a. The largest most modern installation (with production of ca. 3000 t/a), established in 1954 in Guangzhou, Canton, presumably uses a liquid-phase oxidation process in the manganate step. The other (smaller) plants are said to employ roasting methods.

United States. The Carus Chemical Company of La Salle, Illinois, has been a producer of potassium permanganate since 1915. New production technology was developed and installed during the late 1950 s and early 1960 s when patented processes included liquid-phase oxidation in the manganate step [42–44] and closed bipolar cells with continuous crystallization in the permanganate step [46–48]. Numerous improvements have since been introduced, and the current process is fully continuous, highly automated, and in full compliance with environmental regulations. Production capacity is ca. 15 000 t/a.

4.2.3. Commercial Grades, Packaging, and Transportation

Potassium permanganate is offered in several standard grades, including technical, technical free-flowing, U.S.P., and reagent. Guaranteed assays for the various products range from 97 % $KMnO_4$ for the free-flowing grade (which contains an anticaking additive) to 98 % for the technical grade and 99 % for both U.S.P. and reagent grades (typical assays are usually approximately 0.5 % above the guaranteed levels). In addition to assay requirements, reagent grade and U.S.P. grade potassium permanganate conform to specifications set forth in the U.S.P. [57] and the 7 th Edition of Reagent Chemicals [58], respectively. Additional specialty grades are also available.

Material intended for drinking water treatment must conform, for example, to NSF Standard 60 [59] in the United States and to DIN 19 619 [60] in the Federal Republic of Germany. Packaging is in steel drums or, for large users, bulk containers, bulk cars, and trucks.

The following transportation regulations are applicable: IMDG code D 5187, class 5; RID/ADR class 5.1, number 9 c; UN no. 1490; CFR 49: 172.101 oxidizing material.

4.2.4. Uses

The United States is the leading consumer (followed by Western Europe and Japan) of permanganate intended for environmental applications, such as drinking water treatment (removal of tastes, odors, iron, and manganese [61]; control of trihalomethanes [62]); wastewater treatment (destruction of hydrogen sulfide [63] and other toxic and/or corrosive compounds, including phenols [64] and plating wastes [65]); and air purification (degradation of malodorous or toxic constituents in industrial off-gases) [66].

In other parts of the world (particularly Eastern Europe and India), the principal use of potassium permanganate is in chemical processing, especially the manufacture of synthetic organics (e.g., saccharine, chloroamphenicol, ascorbic acid, isonicotinic acid, pyrazinoic acid). Even in the Western world, chemical processing applications are significant. Thus, numerous intermediates for pharmaceuticals and pesticides are made via permanganate oxidation, as are many fine organics, including flavoring. Significant tonnages are also consumed for the oxidative destruction or precipitation of undesirable impurities, often ones that impart unwanted colors or odors. $KMnO_4$ is used in the purification of such chemicals as methanol, ethanol, acetic acid, caprolactam, adiponitrile, plasticizers, carbon dioxide for dry ice manufacture, zinc chloride, and hydrofluoric acid. In recent years, $KMnO_4$ has played a major role in the bleaching of indigo and other dyes for cotton twill fashion garments. This involves processes known as "frosting", "icing", or "acid washing". Permanganate is also used for bleaching beeswax, natural sponges, jute fibers, and certain clays.

Other important uses include surface treatment of carbon steels and stainless steels (descaling and desmutting, especially in wire manufacture), etching of rubber and plastics, and decontamination of nuclear reactors. Fish farmers utilize $KMnO_4$ to prevent oxygen depletion and to control fish parasites.

4.2.5. Economic Aspects

Even though potassium permanganate is not a large-volume chemical it is industrially very important. Especially in the manufacture of specific organic intermediates (Section 4.2.4) $KMnO_4$ is an indispensable oxidant.

World production capacity for potassium permanganate is estimated to be 43 000 – 51 000 t/a, although actual demand is less than 30 000 t/a. Most of the idle capacity is to be found in countries with free market economies: the United States, Western Europe, and Japan. Two of the three West European permanganate plants have been shut down since 1975: Rhône Poulenc's installation in France and the Boots Company plant in the United Kingdom.

Much of the permanganate output from countries with controlled economies enters Western markets at very low prices as a result of extremely low labor costs and virtually no environmental overhead. Government-imposed antidumping duties provide some relief, but not enough to ensure competition on a comparable-cost basis between the imported and domestic product.

4.3. Sodium Permanganate

Many permanganates other than $KMnO_4$ are known [44, pp. 859 – 860], but only sodium permanganate is currently of industrial significance. $NaMnO_4 \cdot H_2O$ [*10101-50-5*], M_r 195.93, ϱ 1.972 g/cm^3, *mp* 36.0 °C, can be made in several ways, including anodic oxidation of ferromanganese in Na_2CO_3 solution [6, p. 759], via aluminum permanganate [67], or from $KMnO_4$ by the hexafluorosilicate method [68]. The price of sodium permanganate is about 5 to 8 times that of $KMnO_4$. This is mainly due to the fact that $NaMnO_4$ cannot be made in the same way as $KMnO_4$, because the oxidation of MnO_2 in a NaOH melt does not lead to the required Na_2MnO_4 (with hexavalent Mn) but only to

Na_3MnO_4 with pentavalent Mn. The latter is very unstable in dilute NaOH solution (and therefore cannot be converted electrolytically to the desired $NaMnO_4$). Even if electrolytic oxidation were possible, there would still be the difficult problem of isolating the extremely soluble $NaMnO_4$ from the alkaline mother liquor.

Sodium permanganate is readily soluble in water (900 g/L at ambient temperature) and is preferred for applications requiring high concentrations of the MnO_4^- ion. Thus, $NaMnO_4$ is used in the etching of plastic parts such as printed circuit boards [69], as well as in some special organic oxidation reactions, particularly if the presence of potassium ions is undesirable.

5. Miscellaneous Manganese Compounds

Manganese(III) fluoride [*7783-53-1*], MnF_3, M_r 111.93, ϱ 3.54 g/cm^3, is a red crystalline material, thermally stable to 600 °C, that is obtained by reacting Mn(II) halides with elemental fluorine at 200 °C. It decomposes in water and is used as a fluorinating agent for organic compounds.

Manganese hypophosphite [*10043-84-2*], $Mn(H_2PO_2)_2 \cdot H_2O$, M_r 202.91, is soluble in water, forms pink, odorless crystals, and decomposes spontaneously upon heating to flammable phosphine. It is prepared from calcium hypophosphite and manganese sulfate. The compound is used as a stabilizer to improve the heat and light resistance of certain man-made fibers and as a food additive and dietary supplement.

Manganese(III) acetylacetonate [*14284-89-0*], $Mn(C_5H_7O_2)_3$, *mp* 172 °C, is a brown to black crystalline solid that is insoluble in water but soluble in organic solvents. In the dry state it is relatively stable. It is synthesized by reacting manganese(III) sulfate, acetylacetone, and ammonia in an aqueous medium, and is used industrially as a catalyst for organic reactions (e.g., in the polymerization of 2-hydroxyethyl methacrylate to a soluble polymer [70]).

Manganese ethylenebis(dithiocarbamate) [*12427-38-2*], Maneb, $(CH_2NHCS_2)_2Mn$, M_r 265.3, is a yellow powder. It is prepared by the addition of an aqueous solution of ethylene diamine and ammonia to carbon disulfide, followed by neutralization with acetic acid and precipitation with $MnSO_4$ or $MnCl_2$. The compound is an important fungicide (→ Fungicides, Agricultural).

Methylcyclopentadienylmanganese tricarbonyl [*12108-13-3*] (MMT), M_r 218.08, ϱ 1.39 g/cm^3, *bp* 233 °C, is a yellow liquid that is insoluble in water but soluble in organic solvents. Several synthetic routes exist for this compound. For example, manganese(II) chloride may be allowed to react with cyclopentadienylmagnesium bromide, C_5H_4MgBr, to form biscyclopentadienylmanganese, an intermediate that reacts with carbon monoxide to give the tricarbonyl. This is then methylated in the presence of Friedel–Crafts catalysts. The product finds limited use as an antiknock additive in motor fuels and as a combustion aid in heating oils.

6. Occupational Health and Environmental Aspects

Manganese plays an important biochemical role in animal and plant life, and is regarded as an essential trace element [71]. It is added as a fertilizer (in the form of MnO or $MnSO_4$) to manganese-deficient soils and is also a constituent in dietary supplements for animals and humans.

The United States Food and Drug Administration has given GRAS (**G**enerally **r**ecognized **a**s **s**afe) clearance to manganese chloride, manganese citrate, manganese gluconate, and manganese sulfate as direct human food ingredients [72]. Nevertheless, excessive intake and long-term exposure, especially to airborne manganese, can lead to toxic effects [73–76]. Acute poisoning with manganese compounds in humans is very rare, but a few cases involving potassium permanganate are mentioned in the literature, either in connection with its attempted use as an aborticide (abortifacient) or in cases of attempted suicide. The human LDLo for $KMnO_4$ is reported to be 143 mg/kg, with a TDLo for women of 2400 mg kg^{-1} d^{-1} [77].

Chronic manganese poisoning is of much greater concern [78]. Inhalation of manganese in the form of oxidic dust or as vaporized metal over extended periods of time (3 months to several years) increases susceptibility of the res-

piratory organs to infection (manganese pneumonia). Manganese can also cause permanent damage to the central nervous system (manganism). The hazard of chronic manganese poisoning affects not only workers in manganese mines and manganese ore crushing/grinding facilities but also those employed in various manganese-consuming industries (e.g., the production and use of ferro- and silicomanganese in steelmaking, welding with manganese-containing welding rods, dry cell manufacture, and production of manganese chemicals) [79], [80]. The early symptoms of manganism are classed as psychological disturbances: fatigue, loss of appetite, incoordination, speech disorders, and psychotic behavior. Manganism at this stage is reversible provided exposure ceases. Continued exposure causes the disease to enter a second stage, in which permanent injury is inflicted on the central nervous system. Symptoms include severe rigidity of the limbs with jerky movements, gait problems, tremors, and excessive salivation and perspiration. There is some similarity between the symptoms of manganism and those of Parkinson's disease.

In the United States, the Federal Republic of Germany, Switzerland, and Japan a general ceiling value (MAK) has been set for manganese in air at 0.5 mg/m^3/BAT-Wert: 20 mg/l[87]. The ceiling value for Mn_3O_4 is 1 mg/m^3 [81]. OSHA has proposed lowering the "Permissible Exposure Limit" to manganese at the workplace to 1 mg/m^3 [82]. On the other hand, the United States Environmental Protection Agency decided not to designate manganese as a hazardous air pollutant under the Clean Air Act, mainly because public exposure to manganese is far below levels associated with serious noncarcinogenic health effects, and available data do not indicate that manganese is carcinogenic [83].

Detailed information about the location of major manganese air emission sources in the United States can be found in [84]. At current exposure levels, manganese does not constitute a general environmental risk [81]. The limit for manganese in drinking water in the United States and most West European countries is 0.05 ppm. This low value is based largely on aesthetic rather than toxicological considerations [85].

7. References

1. R. D. W. Kemmit: "Manganese" in J. C. Bailar, H. J. Emeleus, R. Nyholm, A. F. Trotman-Dickenson (eds.): Comprehensive Inorganic Chemistry, vol. 3, Pergamon Press Ltd., Oxford 1973, pp. 771 – 876. *Gmelin,* System-Number 56, Mangan.
2. O. Glemser, G. Gattow, H. Meisiek, *Z. Anorg. Allg. Chem.* **309** (1961) 1 – 19.
3. R. Giovanoli, E. Stähli, *Chimia* [Aarau] 24 (1970) 49 –61.
4. S. A. Weiss: *Manganese – The Other Uses, A Study of the Non-Steelmaking Manganese Trade and Industry,* Metal Bulletin Books, London 1977.
5. D. Arndt: *Manganese Compounds as Oxidizing Agents in Organic Chemistry,* Open Court Publishing Company, La Salle, Illinois, 1980.
6. H. Marcy: "Anorganische Manganverbindungen," in F. Matthes, G. Wehner (eds.): *Anorganisch-Technische Verfahren,* VEB Deutscher Vlg. für Grundstoffindustrie, Leipzig 1964.
7. A. D. Little, Inc.: *Recovery of Manganese from Low-Grade Domestic Sources, Bureau of Mines Open File Report 201-84 (1985),* National Technical Information Service, U.S. Department of Commerce, Springfield, VA, under PB85-119378.
8. A. J. Fatiadi: "The Oxidation of Organic Compounds by Active Manganese Dioxide," in W. J. Mijs, C. R. H. I. de Jonge (eds.): *Organic Syntheses by Oxidation with Metal Compounds,* Plenum Press, New York–London 1986, pp. 119 – 260.
9. Lambeg Industrial Research Association, US 3 900 425, 1975 (E. Robinson).
10. Carus Corporation, US 4 299 375, 1981; US 4 304 760, 1981 (P. G. Mein, A. H. Reidies).
11. *Ullmann,* 4th ed., **16,** 469.
12. Diamond Shamrock Corp., US 3 780 158, 1973 (J. Y. Welsh).
13. E. Preisler, *Chem. Unserer Zeit,* **14** (1980) no. 5, 137 –148.
14. Manganese Chemicals Corp., US 2 956 860, 1960 (J. Y. Welsh).
15. E. Preisler, Hoechst AG, personal communication.
16. G. D. Van Arsdale, C. G. Meier, *Trans. Electrochem. Soc.* **33** (1918) 109.
17. A. Kozawa, I. Tanabe, *Proc. Electrochem. Soc.* **87** (1987) no. 14, *Proc. Symp. Hist. Battery Technol.,* (1987), 39 – 46.

18. P. M. DeWolff, J. W. Visser, R. Giovanoli, R. Brütsch, *Chimia* **32** (1978) 257 – 259.
19. E. Preisler, *Dechema Monographien,* **102** (1986) 543 –559; **109** (1987) 123 – 137.
20. K. v. Kordesch: *Batteries, Manganese Dioxide,* vol. 1, M. Dekker Inc., New York 1974.
21. E. Preisler, A. Reidies: "Manganverbindungen," in *Winnacker-Küchler,* vol. 2, 1982, 633 – 650.
22. G. Kano, *Mem. Fac. Eng., Fukui Univ.* **34** (1986) no. 2, 163 – 172.
23. A. Kozawa: "New Materials and New Processes," in *New Materials and New Processes,* vol. 3, JEC Press Inc., Cleveland, Ohio, 1985, p. 36.
24. Japan Metals and Chemicals, DE 30 46 913 C2, 1987 (M. Misawa et al.).
25. K. E. Anthony, Broken Hill Proprietary Co., Melbourne, Australia, personal communication.
26. H. Takahashi, *Jpn. Chem. Week* **27** (1986) 8.
27. Kerr-McGee Chemical Corp., US 4 812 302, 1989 (W. C. Laughlin, D. A. Schulke).
28. H. J. Koch, F. Henneberger, DD 45 071, 1966.
29. *Ullmann,* 4th ed., **16,** 472.
30. *Chem. Mark. Rep.,* **Feb 15, 1988,** 33; **May 4, 1987,** 50; **Apr 23, 1984,** 58.
31. R. Landsberg, P. Örgel, *Chem. Technol.* **13** (1961) 665 – 667.
32. H. A. Evangelides, US 2 723 952, 1952.
33. H. Cassebaum, *Pharmazie* **34** (1979) 740 – 744.
34. A. H. Reidies, *Water Conditioning and Purification,* 1987, 32 – 33.
35. R. Scholder, W. Klemm, *Angew. Chem.* **66** (1954) 463 – 465.
36. *Ullmann,* 4th ed., **16,** 473; **3rd ed., 12,** 229.
37. D. G. Lee: "Phase Transfer Assisted Permanganate Oxidations," in W. S. Trahanowski (ed.): *Oxidation in Organic Chemistry,* Academic Press Inc., New York 1980, Part D, pp. 147 – 206.
38. A. J. Fatiadi, *Synthesis* 1987, no. 2, 85 – 127.
39. G. R. Popov, *Khim. Ind. (Sofia)* **4** (1968) 166 – 169; *Chem. Abstr.* **69** (1968) 1 023 326.
40. Nippon Chemical Industrial Company Ltd., US 3 986 941, 1976 (T. Okabe, E. Narita, Y. Kobayashi, M. Mita).
41. W. Baronius, J. Marcy, *Chem. Techn.* **18** (1966) 723 –727.
42. Carus Chemical Company, US 2 940 821, 1960 (M. B. Carus, A. H. Reidies).
43. Carus Chemical Company, US 2 940 822, 1960 (M. B. Carus, A. H. Reidies).
44. Carus Chemical Company, US 2 940 823, 1960 (A. H. Reidies, M. B. Carus).
45. A. H. Reidies in *Kirk-Othmer,* **14,** 844 – 895.
46. Carus Chemical Company, US 2 843 537, 1958 (M. B. Carus).
47. Carus Chemical Company, US 2 908 620, 1959 (M. B. Carus).
48. Carus Chemical Company, US 3 062 734, 1962 (M. B. Carus).
49. Carus Chemical Company, US 3 172 830, 1965 (M. B. Carus).
50. Japan Chemical Industrial Company, JP 72 30520, 1972 (S. Takasawa, Y. Kobayashi); *Chem. Abstr.* **78** (1973) 86 706 e.
51. Carus Chemical Company, US 4 085 191, 1978 (P. G. Mein, H. Adolf).
52. Academy of Sciences, Kasakh SSR, SU 874135, 1981 (A. V. Kotova, N. S. Buyanova, S. T. Deikalo, Y. N. Kolesnikov); *Chem. Abstr.* **96** (1982) 37 689 q.
53. Nippon Chemical Industrial Co., JP Kokai 74 11 664, 1974 (T. Takeda, K. Aihara); *Chem. Abstr.* **80** (1974) 144 813 s.
54. E. T. Berezhenko, B. M. Kuninets, E. V. Martynets, L. V. Pisarev, *Khim. Tekhnol. Kiev* **4** (1981) 29 – 30; *Chem. Abstr.* **95** (1981) 208 465 z.
55. Industrial Quimica del Nalón S.A., BE 905 130, 1986 (M. O. Gonzalez Garcia).
56. Spolek Pro Chemickov a Hutni Vyrobu n.p., Usti n. L., personal communication.
57. *The United States Pharmacopeia,* XIX revision (1975) p. 394; XX revision, 4th suppl. (1983) p. 711, United States Pharmacopeial Convention Inc., Rockville, MD.
58. *Reagent Chemicals,* 7th ed., American Chemical Society, Washington, DC, 1986, pp. 533 – 535.
59. NSF: *Standard 60 for Drinking Water Treatment Chemicals – Health Effects,* December 1987, revised June 1988, NSF, Ann Arbor, MI.
60. DIN 19 619, Kaliumpermanganat zur Wasseraufbereitung, Technische Lieferbedingungen, Feb. 1975.
61. K. Ficek in R. L. Sanks (ed.): *Water Treatment Plant Design for the Praciticing Engineer,* Ann Arbor Science Publications, Inc., Ann Arbor, MI, 1978, 461 – 479.
62. Ph. C. Singer, J. H. Borchardt, J. M. Colthurst, *J. Am. Water Works Assoc.* **72** (1980) 573 – 578.

63. K. S. Pisarczyk, L. A. Rossi, *Sludge Odor Control and Improved Dewatering with Potassium Permanganate,* paper presented at the 55th Annual Conference of the Water Pollution Control Federation in St. Louis, 5 Oct 1982; J. H. Jackson, *Pulp & Pap.* **58** (1984) 147 – 149.
64. Hiyuga Seirensho K.K., JP-Kokai 78 30 163, 1978 (K. Atsuchi, E. Hagiwara, K. Fujita).
65. C. Freudenberg K.-G., Langbein-Pfanhauser Werke AG, DE-OS 3 335 746, 1985 (L. Feikes, D. Schroeder, P. Reyer); *Chem. Abstr.* **103** (1985) 26 768 b.
66. K. J. Ficek, *Paint Varn. Prod.* **63** (1973) no. 5, 45.
67. Carus Chemical Company, US 2 504 131, 1950 (T. Jaskowiak).
68. M. M. Kushnir, *Khim. Prom. Nauk. Tekhn. Zb.* **79** (1963); *Chem. Abstr.* **59** (1963) 4813.
69. Morton Thiokol Inc., US 4 601 783, 1986; US 4 601 784, 1986 (G. Krulik).
70. Leningrad Technological Institute, SU 1 353 780, 1987 (N. A. Lavrov, A. F. Nikolaev, N. A. Astafeva); *Chem. Abstr.* **108** (1988) 205 263 h.
71. Committee on Medical and Biologic Effects of Environmental Pollutants, National Research Council, Manganese, Washington, DC, 1973.
72. FDA: *Fed. Regist.* **50** (1985) 19 164 – 19 167.
73. EPA: *Health Assessment Document for Manganese, Final Report,* EPA-60018-83-013F, NTIS PB 84-229954, Environmental Criteria Assessment Office, Cincinnati 1984.
74. EPA: *Health Effects Assessment for Manganese (and Compounds)*, EPA/54011-861057. PB 86-134681, Office of Health and Environmental Assessment, Cincinnati 1984.
75. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, NIOSH: *Registry of Toxic Effects of Chemical Substances,* 13th ed.,1983 – 1984 cumulative supplement to the 1981 – 1982 ed.
76. WHO: *Task Group on Environmental Health Criteria for Manganese, Environmental Health Criteria 17: Manganese,* WHO, Geneva 1981.
77. *Registry of Toxic Effects of Chemical Substances,* ed. 1985/86, item 55171, U.S. Department of Health and Human Services DHHS (NIOSH), publication no. 87 – 114.
78. M. Tolonen: "Industrial Toxicology of Manganese," *Work, Environ. Health* **9** (1972) 53 – 60.
79. H. Roels et al., *Am. J. Ind. Med.* **11** (1987) 297 – 305.
80. H. Roels et al., *Am. J. Ind. Med.* **11** (1987) 307 – 327.
81. R. Schiele, *Met. Umwelt* 1984, 471 – 477.
82. NIOSH: *Fed. Regist.* **53** (1988) no. 109, 21 000 –21 001.
83. EPA: *Fed. Regist.* **50** (1985) 36 627 – 36 628.
84. EPA: *Locating and Estimating Air Emission from Sources of Manganese,* EPA-45014-85-007h, NTIS PB 86 117 587, Sep 1985.
85. R. J. Bull, G. F. Craun: "Health Effects Associated with Manganese in Drinking Water," *J. Am. Water Works Assoc.* **69** (1977) 662 – 663.
86. *Kirk-Othmer,* **15,** 1036.
87. DFG *MAK- und BAT-Werte-Liste* Wiley-VCH, Weinheim (2001)

Mannitol → Sugar Alcohols

Margarines and Shortenings

Ian P. Freeman, Unilever, 8 Vicarage Close, Ravensden, MK44 2RW, United Kingdom

1.	**History**	1
2.	**Properties of Margarine**	2
2.1.	**Subjective Properties**	3
2.2.	**Objective Properties**	4
2.2.1.	Color	4
2.2.2.	Gloss	4
2.2.3.	Melting Properties	4
2.2.4.	Penetration	5
2.2.5.	Drop Size	5
2.2.6.	Salt Release	5
2.2.7.	Keeping Properties	6
3.	**Composition**	7
3.1.	**The Fat Phase**	7
3.1.1.	Choice of Fat Blend Components	7
3.1.2.	Preparation of Fat Blends	9
3.1.2.1.	Refining	9
3.1.2.2.	Hardening	9
3.1.2.3.	Interesterification	10
3.1.2.4.	Fractionation	10
3.2.	**The Aqueous Phase**	11
3.2.1.	The Aqueous Phase in Conventional Margarines	11
3.2.2.	The Aqueous Phase in Low-Fat Spreads	11
3.3.	**Emulsifiers**	12
3.4.	**Minor Components**	12
4.	**Processing**	13
4.1.	**Churn Process**	14
4.2.	**Votator Process**	14
5.	**Specialty Margarines and Shortenings**	15
5.1.	**Pastry Margarines**	15
5.2.	**Frying**	16
6.	**Packaging**	17
7.	**Economic Aspects**	18
8.	**References**	18

1. History

The term "margarine" was coined by the French chemist Michel Chevreul in 1814. He used it to describe one of what he thought to be the three primary constituents of natural fats, the others being *oleine* (from the Latin *oleum* for oil) and *stearine* (solid fat, from the Greek *stear* for tallow). "Margarine" was supposedly a third, semisolid constituent, so-named (from the Greek *margaron*, pearl) because of its pearl-like, crystalline appearance. Chevreul believed that these three substances were glycerides of the fatty acids oleic acid, stearic acid, and margaric acid, respectively. Subsequent investigation has shown that natural fats contain many more fatty acids than these three (→ Fats and Fatty Oils, Chap. 2.). Oleic acid and stearic acid have since acquired unique chemical identities, and some authors still use Chevreul's "margaric acid" as a trivial name for heptadecanoic acid even though his original concept has been shown to be erroneous. The terms olein and stearin have been retained to describe the lower and higher melting fractions into which natural fats can be separated, but the term margarine is no longer employed in that context. It is now used instead to describe a commodity resembling butter but manufactured from fats derived from sources other than milk (→ Dairy Products, Imitation).

Such a commodity was first produced by Hippolyte Mège – Mouries in 1869 in response to a prize offered by Napoleon III for a substitute for butter. Mège – Mouries removed the stearin from beef tallow by fractional crystallization, leaving a mixture of olein and margarine or *oleomargarine*. This he emulsified with milk, water, and a little alkali to produce a plastic, fat-continuous emulsion resembling butter in texture. The term oleomargarine (or "oleo") is still used for this product in the United States. In Europe the name was shortened to margarine, a name that was established by Act of Parliament as the legal term in Britain in 1887 in preference to the suggested alternative "butterine".

Early production of margarine was dependent upon the availability of the oleomargarine from beef tallow to provide the desired consistency. However, two inventions soon opened the way to a wide range of natural fats as potential raw

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_145

materials for margarine manufacture: *hardening* and *interesterification*.

Hardening, or hydrogenation, consists of reactions of unsaturated fats with hydrogen, thereby raising their melting points and in some cases increasing their oxidative stability. This procedure was introduced by NORMANN in 1903 [4].

Interesterification involves the exchange of fatty acid moieties between or within glycerides to create new glyceride components. The exchange of acid moieties between esters had been demonstrated by FRIEDEL and CRAFTS [5] as early as 1865, and its potential application to natural fats was demonstrated by NORMANN in the 1920s [6]. Because of the required high temperatures (> 200 °C) and long reaction times (several hours), however, the process was of no interest to margarine manufacturers until ECKEY introduced alkali-metal catalysts in 1948 [7]. The original conditions are not only economically prohibitive, they are also deleterious with respect to the fat. Using alkali-metal catalysts the reaction can be carried to completion below 60 °C within minutes.

A third major development in the manufacture of margarine was the introduction of the votator system in 1936, permitting the ingredients (i.e., fat, the aqueous phase, and minor components) to be premixed and then subjected to simultaneous rapid shearing and cooling. This not only causes the aqueous phase to be finely and uniformly dispersed throughout the fat phase but also gives the processor a greater degree of control over the final texture of the product.

Margarine was first conceived as a substitute for butter, and from the start it was seen as a commercial threat by the dairy industry. Margarine manufacture was soon controlled by legislation designed to protect the dairy interest. Suppression of the name "butterine" was considered to be an early victory for the dairy farmers in this long standing "war". Other legislation soon decreed that margarine must be sold as a product distinct from butter, and it prohibited the extending of butter products by incorporation of small proportions of margarine. In the United States, the addition of color to margarine to give it a more butter-like appearance was long prohibited, and manufacturers had to resort to including capsules of color with the packs of margarine, which consumers could blend in to their own satisfaction.

Despite such legislation, the relative cheapness of margarine soon made it a major component of the diet, especially for the poorer members of society, and enlightened governments quickly foresaw the dangers of malnutrition if nutritional standards were not laid down and maintained. The proportion of water was thus controlled at no more than 15 % in most countries, and the addition of vitamins A and D was made compulsory in order to maintain margarine at a nutritional standard comparable with that of butter.

Recent technical advances and new insights into nutrition and health have enabled margarine manufacturers to exploit more fully the flexibility inherent in the formulation of margarine. As a result, margarine is now a product in its own right, not simply a cheap substitute for butter. For example, wider usage of domestic refrigerators has produced a demand for products plastic enough to spread at low temperatures, a demand that has been met by the so-called *soft blend margarines*. These contain fats chosen in such a way that the product retains a spreadable plasticity over a wide range of temperatures. A high level of fat in the diet, especially saturated fat, is now considered undesirable, and this has led to the development of *low-fat spreads* and to the use of *fat blends* with a high proportion of polyunsaturated fatty acids (PUFA). The early protective legislation has also come under pressure from responsible nutritional authorities, who have urged that margarine manufacturers be permitted more flexibility. Consumers now have a wide choice of margarine-type products, covering the whole gamut of fat contents from conventional margarine at 85 % down to very low fat spreads at ca. 20 %, and including various melanges with dairy products such as *low calorie and high PUFA "butters"*.

Brand names for the various types of margarine are listed in Table 1.

2. Properties of Margarine

The properties of margarine can be considered under two headings:

Table 1. Some typical margarines and other spreads with brand names in use at the time of writing (1989)

Fat content, wt%	Brand name	Manufacturer	Country
Conventional Margarines			
General Purpose			
ca. 80	Stork	Unilever	UK
ca. 80	Imperial	Unilever	USA
ca. 80	Rama	Unilever	FRG
ca. 80	Lesieur	Lesieur Alim.	France
Soft Blend (i.e., spread at refrigerator temperature)			
ca. 80	Stork S.B.	Unilever	UK
ca. 80	Blue Band	Unilever	UK
PUFA (i.e., high in polyunsaturates)			
ca. 80	Flora	Unilever	UK
ca. 80	Becel	Unilever	Netherlands, FRG
ca. 80	Vitelma	Vandemoortele	Belgium
ca. 80	Sense	Van Dijk	Netherlands
ca. 80	Vitalite	Kraft	UK
Reduced Fat Spreads			
72	Promise	Unilever	USA
70	Krona	Unilever	UK
60	Stork Light	Unilever	UK
60	Light Soft	Fleischmann	USA
52	Country Crock	Unilever	USA
40	Delight	Unilever	UK
40	Gold	St Ivel	UK
40	Blue Band	Unilever	Netherlands
27.5	Outline	Unilever	UK
25	Gold Lowest	St Ivel	UK
Melanges (i.e., mixtures of butter fat and vegetable fats)			
78	Willow	Dairy Crest	UK
75	Golden Churn	Kraft	UK
75	Clover	Dairy Crest	UK
40	Clover Light	Dairy Crest	UK
40	St Hubert 4 I	St Hubert Dairies	France
40	St Florigny	Unilever	France
40	Tartine	Lesieur Alim.	France
Reduced Fat Butters			
40	Half Fat Anchor	Anchor Foods	UK
40	Die Leichte	Unilever	FRG
40	Sylphide	Fromagerie Bel	France
40	Vive la Vie	Unilever	France
40	Balade	Corman-France	France

1) *Subjective properties* that are assessed by the reaction of a consumer when presented with a sample of the margarine
2) *Objective properties* measured with laboratory instruments

2.1. Subjective Properties

The most obvious subjective property is *appearance,* which comes into play as soon as the consumer opens the package. First, the product must have an attractive color. This has traditionally meant a yellow color similar to that of butter. Margarine must also have a *smooth, creamy-looking texture*; i.e., it should not appear grainy or greasy. When the consumer applies the margarine to a piece of bread or biscuit, it should have good "spreadability"; i.e., it should be soft enough to spread smoothly and easily without breaking into lumps, exuding water, or dragging the bread surface. Furthermore, the margarine must have an acceptable flavor and "mouthfeel".

For Europe and many other parts of the world the standard for good *flavor* is fresh butter, and much research has been devoted to creating flavor cocktails imparting a butter-like taste. In some places, however, other flavors are preferred. Thus, in parts of South America fruit flavors such as pineapple are appreciated, while in some African countries fish flavors are popular.

Mouthfeel is a reflection of the physical changes that take place as margarine melts and breaks down under the action of teeth and tongue and mixes with saliva in the warm environs of the

mouth. This melt-down can take place smoothly or lumpily. The melt-down may be complete, or some unmelted fat may remain and be perceived as a waxy coating around the tongue and palate. The melted product can feel creamy or thin and watery. In some circumstances, a cool sensation is perceived, perhaps due to a rapidly melting fat blend. The flavor release pattern is also affected by the way in which margarine breaks down, releasing flavor components to the palate at different rates, which may in turn influence the consumer's overall perception of flavor.

All these effects can be related, in theory at least, to such physical properties of the margarine as melting characteristics or emulsion stability, but it is not possible to measure all of them directly with laboratory instruments.

2.2. Objective Properties

2.2.1. Color

The color of margarine is defined in terms of the tristimulus concept, which postulates that any color can be specified in terms of the three primary colors red, yellow, and blue. In practice, the color of light reflected from the surface of a margarine sample is matched against that of light which has passed through a series of filters based on the primary colors. The color is then defined in terms of the density of each of the filter colors required to give a match. A suitable commercial instrument for this purpose is the Lovibond Tintometer [8].

Color is primarily dependent on the amount of coloring material added to the margarine mixture, but it is also affected by the physical nature of the margarine (e.g., by the ratio of solid fat to liquid oil in the fat phase, the amount of the water phase, and the size and distribution of water droplets).

2.2.2. Gloss

Gloss is a measure of the efficiency with which light is reflected from the surface of the margarine. A mirror surface reflects a beam of light at the same angle at which the beam strikes the surface, with little or no scattering. A matt surface scatters the beam so that the intensity of the reflected light is uniform over a wide angle of reflection. A glossy surface falls between these two extremes: a portion of the light is reflected along the angle of incidence, but some is scattered. Glossiness can therefore be measured by comparing the intensity of incident light with that of reflected light at predetermined angles.

Like color, gloss is dependent upon the physical nature of the margarine; i.e., upon the proportion and distribution of solid fat crystals and water droplets.

2.2.3. Melting Properties

The melting characteristics of a margarine are essentially those of the fat blend used in its composition, and they are responsible for many of the product's subjective characteristics (see Section 2.1). A fat blend, and indeed a natural fat, consists of a complex mixture of glycerides, each with its own characteristic melting point and each influencing and influenced by the melting of other glycerides in the mixture through the formation of solutions and eutectics. As a result, a solid fat does not show a single sharp melting point, but rather a gradual softening over a wide range of temperature. The melting character of a fat must therefore be defined in terms of the solids content of the fat at a given temperature.

Dilatation. Traditionally, solids content was measured by dilatation [9]. A completely solid fat, like any other solid, shows a small increase of volume with increasing temperature. A fully liquid oil also expands when heated, slightly more than does a solid fat. However, when the solid melts to a liquid, a considerably greater volume change is observed. Therefore, the increase in volume in the course of heating between two temperatures is essentially a measure of the proportion of the sample that melts over that temperature range. In practice the change is measured with a dilatometer, which consists of a glass bulb containing the fat sample connected to a graduated capillary tube containing a marker liquid. As the volume of fat in the bulb increases the marker liquid is pushed up the capillary, permitting the volume change to be quantified. Such a bulb is immersed sequentially in a series of water baths at different temperatures and for each bath the level of liquid in the capillary is noted.

After correcting for expansion of the glass and the liquid oil it is then possible to calculate the proportion of fat melted at each temperature.

Nuclear Magnetic Resonance. Today it is generally more convenient to use a nuclear magnetic resonance (NMR) technique to measure solid fat content [10]. This is possible because a proton within a molecule responds differently to a strong magnetic field depending on whether it is held rigidly in a solid crystal matrix or only restrained in a liquid substrate. NMR instruments especially designed for this application are commercially available.

Polymorphism. One factor that can confuse the measurement of solid fat content by either dilatation or NMR is the phenomenon of polymorphism [11]. Thus, many compounds are capable of assuming any of a number of different physical forms, each with a different melting point. This phenomenon is especially evident in the behavior of fats. When a melted fat is cooled rapidly below its melting range, it usually crystallizes in an unstable form known as the α modification. On standing, this rearranges into a more stable β or β' form (cf. Section 3.1.1). Stable modifications can also be achieved by slow cooling or by crystallization from a solvent. The solid content of a particular sample depends on the relative proportions of these various polymorphs and therefore on the way in which the sample has been prepared. For quantitative comparisons, samples must therefore be subjected to a standard regime before their solids content is measured.

2.2.4. Penetration

Penetration is a measure of the resistance shown to a plunger pressed through the margarine under standard conditions. It depends upon both the viscosity of the liquid components of the margarine mix and the rigidity and extension of the network of solid fat crystals (see Section 3.1.1). Penetration is measured with a penetrometer [12]. A plunger with a standard conical tip and loaded with a standard weight is allowed to sink for a standard period of time and by gravitational force alone into the surface of the margarine. The depth to which the plunger penetrates is recorded on a dial. Such an instrument must normally be housed in a temperature-controlled environment, and the sample should be stored under controlled temperature conditions for at least a week prior to measurement.

The penetration value can be indirectly related to such subjective properties as spreadability and how the margarine feels when cut with a knife. It also gives an indication of how well the margarine will retain its shape during packing, storage, and transport.

2.2.5. Drop Size

The aqueous phase of a margarine is dispersed in the fat phase in the form of small droplets with a maximum diameter of about 100 µm. Average droplet size is 5 – 10 µm. In conventional margarines, drop size is an important factor in controlling the biological stability of the product. Most microorganisms can only develop in an aqueous environment and their growth throughout the product is therefore inhibited if water is present only in very small drops. In the case of low-fat spreads (where water constitutes 60 % or more of the total composition) the droplet size can also affect such subjective properties as hardness, mouthfeel and flavor release, but these considerations are less critical in conventional margarines (with only ca. 15 % water content).

Drop size can be measured directly with an optical microscope or indirectly by NMR techniques [13]. Both methods are subject to difficulties of interpretation because the droplets in a given sample have a range of sizes and are not necessarily spherical.

2.2.6. Salt Release

The salt release test provides an indication of how margarine is likely to break down in the mouth. Thus, a sample of margarine is transferred to a vessel containing water and two electrodes with which the conductivity of the water can be measured. The sample is then stirred and the temperature is slowly raised. When the sample begins to break down, the aqueous phase of the margarine is released from its surrounding fat phase and any salt present is liberated. As a result, the conductivity of the water increases.

For a good product this increase should occur at around mouth temperature, i.e., 33 – 37 °C. If the salt release occurs at a considerably higher temperature (e.g., 40 °C) then the product will have a dead, waxy feel in the mouth.

2.2.7. Keeping Properties

If stored for too long or under unfavorable conditions, desirable margarine properties may be lost or undesirable ones may develop. Physical, chemical, or biological changes may be responsible.

Physical Changes. Physical changes lead to a breakdown in the structure of the margarine.

1) *Emulsion instability* results in the coalescence of dispersed water droplets and visible separation of water from the product. Provided a suitable emulsifier has been chosen and the product has been correctly processed, this only becomes a problem if the product is subjected to extreme changes of temperature.
2) *Oily exudation* indicates a breakdown of the crystal network in the fat phase of the product, and there is visible separation of liquid oil at the surface (i.e., between the paper and the product in a wrapped margarine, or between the walls of the tub and the product in a tub-filled margarine). Exudation may arise as a result of inappropriate choice of fat composition, but in an established product it is more often an effect of exposure to extreme temperatures or to mechanical mishandling.
3) *Sandiness or graininess* is due either to continued growth of individual crystals or to aggregation of crystals after processing. The effect can be detected visually (by an excessively matt or granular surface appearance) or by mouthfeel [14]. Certain fats or fat blends (notably those containing palm oil and lard) are particularly prone to this fault due to the presence of large quantities of single species of glycerides, oleodipalmitins, or palmitodioleins, which readily form compound crystals. Critical proportions of these fats should be avoided. Interesterification can also be used to reduce the concentration of individual triglycerides responsible for the effect.

Chemical Changes. Chemical changes affect the flavor of a product, causing rancidity through either autoxidation or hydrolysis.

Autoxidation refers to reaction between fatty acid chains and atmospheric oxygen. The first step is removal of a hydrogen atom from the hydrocarbon chain of the fatty acid and its replacement by a hydroperoxide group. The latter then reacts further by addition to double bonds or by chain cleavage to give a complex mixture of aldehydes, ketones, and short-chain fatty acids. Many of these products are highly flavor-intensive, only a few parts per million being sufficient to impart a detectable flavor. Hydrogen atoms on carbon atoms adjacent to double bonds, and to a lesser extent those adjacent to ester groups, are most labile and therefore most readily attacked. Polyunsaturated fatty acids are more susceptible than monounsaturated acids, which in turn are more susceptible than saturated acids. Fish oils are highly unsaturated, with up to six double bonds per fatty acid chain, and these are particularly prone to autoxidation. They have the added disadvantage that their autoxidation products have characteristic rancid, fishy flavors unacceptable in a margarine product. For this reason fish oils are only used in margarine formulations after the number of double bonds has been reduced by hardening.

In general, autoxidation can be controlled by efficient refining (see Section 3.1.1). Refining not only removes indigenous flavors from natural oils or fats but also reduces the level of any pro-oxidants present. If a blend contains predominantly animal fats, further protection by antioxidants may be required. This precaution is not normally necessary for vegetable fats and oils, as these contain adequate quantities of natural antioxidants in the form of tocopherols.

Hydrolytic rancidity occurs when the ester linkage joining the fatty acid moiety to glycerol is split, leading to the release of free fatty acid. This can be a result of autoxidation, but it is more likely to be caused by hydrolysis. Long-chain fatty acids ($\geq C_{16}$) are virtually flavorless, so their release does not produce a flavor problem. On the other hand, short-chain fatty acids ($\leq C_{12}$) have quite intense intrinsic flavors, and their release causes the off-flavors variously described as rancid, soapy, or cheesey. Since butter contains a relatively high proportion of such

short-chain acids it is more prone than most margarines to this form of off-flavor development.

Biological Changes. Biological changes are the result of microbial contamination of a margarine. If allowed to develop, microorganisms can render a product unsightly and unwholesome as well as subject to physical and chemical changes. The primary means of controlling microorganisms is strict attention to hygiene during storage of ingredients, processing, and packaging, but preservatives are sometimes used to protect the stored product after packaging. Because of their high water content, low-fat spreads are more susceptible than conventional margarines to this type of deterioration.

3. Composition

All margarines consist of both a fat phase and an aqueous phase. In conventional margarines the fat phase constitutes ca. 85 % of the total mixture. Low-fat spreads can contain as little as 20 % fat phase, with a correspondingly larger amount of aqueous phase.

3.1. The Fat Phase

The fat phase consists predominantly of blends of natural or modified oils or fats. These comprise not only triglycerides but also various minor natural ingredients such as tocopherols and sterols. Small amounts of other substances must be added by the manufacturer, including emulsifiers, coloring agents, and antioxidants.

3.1.1. Choice of Fat Blend Components

The most essential characteristic of the fat blend for a margarine is that it contains a certain proportion (at least 15 – 20 wt%) of a fat that is solid at normal ambient temperature. This solid fat is known as the *hardstock*, and its function is to form a network of fat crystals through the otherwise liquid oil and so give it a certain rigidity [15]. Ideally, the hardstock should all have melted or dissolved at mouth temperature, otherwise the product has a heavy, waxy mouthfeel. However, a small proportion (≤ 2 %) of higher-melting fats can be tolerated without the effect being too noticeable. Indeed, a little high-melting material can be beneficial, in that it promotes crystallization of the main fraction of the hardstock and also stabilizes the plasticity of the blend over a wide range of temperature. In margarines designed for sale in tropical countries, ideal mouthfeel may have to be sacrificed to some extent and the proportion of high-melting fat increased to ensure that the margarine withstands high ambient temperatures.

Crystallization characteristics of the hardstock can also be critical, but this is not usually a concern for general-purpose margarines (i.e., those other than soft blend or high PUFA margarines). Typical formulations are based on a mix of fats with a wide range of melting points. The fats are typically hardened or partially hardened to ensure a substantial proportion of solid fat over a wide temperature range. An excessively high level of high-melting components must be avoided, as must components with undesirable crystallization characteristics. An example of such a characteristic is the tendency to crystallize slowly, causing too much growth of solid crystals subsequent to mechanical processing. This phenomenon, known as *posthardening*, gives rise to products with a hard, brittle texture and is especially associated with palm oil. The somewhat similar problem of sandiness or graininess is commonly associated with lard or blends of lard and palm oil.

Apart from these restrictions, the main criterion for choice of fat components is the cost, and the blend formulator will adapt his formulations to whatever is economically attractive at the time. He is aided in this by the concept of linear programming, which postulates that the melting characteristics of a blend of fats are additive functions of the individual components of the blend [16]. Thus, the solid fat content of a blend S can be calculated from the formula:

$$S = xA + yB + zC + \ldots$$

where $x, y, z, \ldots$ are the proportions of the individual components and $A, B, C, \ldots$ are the solid fat contents of those components. This formula is reasonably valid over a wide range of blends and conditions, and enables the blender to substitute components at will depending on availability and cost. For soft blend margarines, and

especially for high PUFA margarines, the crystallization characteristics of the hardstocks are more critical. In order to comply with physical or nutritional requirements, the proportion of hardstock must be kept to a minimum and the ability of the hardstock to form an intermeshing crystal network must be maximized. This means that fats with a tendency to crystallize in the β form, which exhibits small, compact, spherulitic crystals with little intercrystal interaction, are rejected in favor of those with a tendency to form β' crystals, with extended open dendritic growth (→ Fats and Fatty Oils, Chap. 3.).

Typical Oils and Fats Used in Margarine. The oil blender can select his components from a wide range of natural oils and fats but only about a dozen are of commercial importance at any one time (Table 2). The list varies depending on agricultural practice and availability, but the following applies at the time of writing (1989). Of the liquid oils, soybean oil [*8001-22-7*] is the most widely used, with rapeseed oil [*8002-13-9*] a close equivalent in Europe. Sunflower oil [*8001-21-6*] is favored in high PUFA blends because of its high content of linoleic acid (ca. 50 – 60 %) and the absence of the linolenic acid present in soybean oil, which reduces flavor stability. Safflower oil [*8001-23-8*] is more expensive than sunflower oil, but it is favored by some manufacturers because it has an even higher linoleic acid content and can therefore be used to maintain a high PUFA content with a higher level of hardstock. Peanut oil [*8002-03-7*] was at one time widely favored because of its high oxidative stability, but it is now more expensive than other liquid oils. Palm oil [*8002-75-3*] and lard [*61789-99-9*] are semiliquid oils whose usage in unmodified form is restricted by the crystallization characteristics discussed above. Apart from lard, the only animal fat to find wide application is beef tallow [*61789-97-7*] , but, as Mège – Mouries found (cf. Chap. 1), the high-melting stearin must be removed before beef tallow can be used. Butter fat is of course incorporated into melange products. The lauric fats (palm kernel oil [*8023-79-8*] and coconut oil [*8001-31-8*]), so called because they contain a high level of lauric acid, are also useful as hardstock components since they have relatively sharp melting points near mouth temperature. Marine oils are among the cheapest oils available, but because

Table 2. Typical oils used in margarine manufacture

Type of oil	Region of origin	Use*
Soybean	USA and South America	widely used as a cheap source of PUFA (linoleic acid) also, after hardening, as a hardstock
Cottonseed	USA and Eastern Mediterranean	in margarine mainly used after hardening as a hardstock especially in USA
Rapeseed	North America and Northern Europe	since the development of low-erucic varieties, this has been increasingly used as a replacement for soya bean oil in Europe and Canada
Sunflower	widely distributed in warmer temperature zones	main source of PUFA (linoleic acid)
Safflower	USA, Central America, North Africa, and India	richer in PUFA than sunflower but more expensive
Groundnut	widely distributed in semitropical areas	formerly one of the basic commodity oils for margarine, both as a liquid oil and, after hardening, as a hardstock but now relatively expensive
Palm	Malaysia and Central Africa	widely used both as the whole oil and after fractionation or interesterification when melting characteristics are important
Lard	widely distributed where pigs are raised, mainly USA and Europe	used in similar contexts to palm oil but only in circumstances where the animal source is not important
Beef tallow	USA, Europe, and Australasia	limited use of the olein in margarines, especially those products where texture is more important than health appeal
Palm kernel	as for palm oil	widely used as a hardstock because of its relatively sharp melting range at about mouth temperature
Coconut	Pacific Islands, West Indies, and Southern Indian Continent	as for palm kernel
Fish		
Herring	North Atlantic South	very cheap source of oil but because of flavor problems can only be
Menhaden Pilchard Anchovy	Atlantic North Pacific South Pacific	used in food products after hardening; however, increased exploitation as source of essential fatty acids and sale as such through pharmacies and health food stores may alter this pattern of use

* PUFA = polyunsaturated fatty acids

of flavor problems they can only be used (after hardening) in cheaper, lower quality margarines. At one time this source was dominated by whale oil, but embargoes placed on whaling for conservation reasons have caused whale oil to be replaced by oils derived from such oily fish as herring, menhaden, and anchovy. For further information on the oils and fats described in this section, see → Fats and Fatty Oils.

3.1.2. Preparation of Fat Blends

3.1.2.1. Refining

All crude oils are darkly colored and strongly flavored, and they contain such undesirable components as free fatty acids and phosphatides. They must therefore be subjected to a refining process before use to produce a bland, lightly colored, neutral oil with increased resistance to autoxidation. Conventional refining involves three main stages: neutralization, bleaching, and deodorization (→ Fats and Fatty Oils, Chap. 6.).

Neutralization removes free fatty acids and it is carried out in the early stages of refining because acids would interfere with such later procedures as bleaching. In the alkali neutralization process, the crude oil is stirred with an aqueous solution of sodium hydroxide, which reacts with free fatty acids to form soaps. After settling, the aqueous phase containing the soaps is drained off. In the alternative physical refining process the oil is neutralized by distillation, free fatty acids being more volatile than the parent triglycerides.

Oils with a high proportion of phosphatides (e.g., soybean oil) are subjected to *degumming*, in which boiling water is used to hydrate and precipitate the phosphatides. Phosphoric or citric acid is often included in the boiling water to facilitate hydrolysis of the more resistant phosphatides.

The dried, neutralized oil is next subjected to *bleaching*. This process decolorizes the oil and removes various minor components, including compounds of heavy metals (especially iron and copper), which promote autoxidation, and of alkali metals, which promote hydrolysis. An absorbent material (usually an activated bleaching earth like fuller's earth, but sometimes charcoal) is stirred into the oil and stirring is continued under an inert atmosphere for 20 – 30 min at 80 – 120 °C, after which the absorbent is removed by filtration.

The final stage of refining is *deodorization.* The oil is heated to 180 – 250 °C for several hours while steam is blown through it. A vacuum is also applied to remove volatiles, which include flavor components (especially short-chain autoxidation products and residual free fatty acids).

In addition to simple refining, natural oils are sometimes subjected to procedures designed to alter their physical or chemical characteristics. Examples include hardening (hydrogenation), interesterification, and fractionation, which are discussed in the following sections.

3.1.2.2. Hardening

Hardening or hydrogenation is the process by which unsaturated fatty acids in an oil are reacted with hydrogen in the presence of a catalyst (→ Fats and Fatty Oils, Chap. 8.). Hydrogen adds to carbon – carbon double bonds in the fatty acid chain, converting them to saturated bonds.

The most common catalyst in oil and fat processing is activated metallic nickel, although other activated metals are sometimes used. Catalysts are most active when freshly prepared, and they lose activity with use, especially when subjected to oils containing a significant amount of sulfur. Sulfur is the most common poison for this type of catalyst. Some catalysts are deliberately poisoned with high-sulfur oil to achieve a specific degree of activity, because less active catalysts show greater selectivity.

Selective catalysts favor the hydrogenation of polyunsaturated fatty acids relative to monounsaturated fatty acids. Thus, conversion of linoleic acid, with two double bonds, to a form of oleic acid with only one double bond, is favored over the conversion of oleic acid to stearic acid with no double bonds. With a *nonselective catalyst* on the other hand, double bonds are attacked randomly; saturation of linoleic acid thus proceeds in parallel with that of oleic acid.

Another potential reaction under hardening conditions is isomerization without the addition of hydrogen. That is, the positions of double bonds along the chain may change. Moreover,

cis double bonds, which prevail in natural oils and fats, may be converted into the trans form, a geometry that occurs only rarely in natural oils. Isomerization reactions are favored by less active catalysts and also under conditions of hydrogen starvation.

The main purpose of hardening is to raise the solids content of an oil. In general, the more unsaturated the oil or fat, the lower is its melting point. Melting points of trans fatty acids are also higher than those of the corresponding cis isomers. Nonselective hardening tends to produce a wider mix of solid components, giving a fat that is plastic over a wide range of temperature. Selective hardening and trans hardening both favor a less complex mixture, resulting in a narrower melting range. Hardening also helps to stabilize a polyunsaturated oil against autoxidation. Thus, the linolenic acid in soybean oil (with three double bonds) is reduced in the course of hardening, as are the highly unsaturated fatty acids (with five or six double bonds) in fish oils.

3.1.2.3. Interesterification

Interesterification alters the ester links in oils and fats joining fatty acid moieties and glycerol; thus, one fatty acid residue from one of the three positions on glycerol is interchanged with another , either on the same molecule or on another molecule (→ Fats and Fatty Oils, Chap. 4.2., → Fats and Fatty Oils, Chap. 9.). This process commonly results in a random fatty acid distribution over all the glyceride molecules present (*random interesterification*). In the case of a two-component blend, such as a liquid oil and a solid fat, interesterification gives a more complex spectrum of glycerides than would be obtained by simple physical mixing. The result is a plastic fat that melts gradually over a wide temperature range rather than a semisolid oil with high-melting components. Interesterification can also be used to eliminate some triglyceride components that tend to dominate the crystallization behavior of a mixture (e.g., to remove troublesome triglyceride species from lard or palm oil, see Section 2.2.7).

Interesterification is catalyzed by heavy-metal salts such as tin(II) chloride or by fatty-acid soaps of the alkali metals. These are generally introduced at a level of about 2 %. For such catalysts to be effective, the reaction must be carried out at about 200 °C. More efficient catalysis is provided by such compounds as the sodium salts of the short-chain alcohols (sodium methoxide or ethoxide), sodium glyceroxide, or anhydrous sodium hydroxide, all at a level of about 0.5 % or less. Reaction then proceeds at a much lower temperature (≤ 60 °C), reaching completion within a few minutes. This is a much more efficient technique for random interesterification. Moreover, it permits *directed interesterification*, in which the reaction is conducted at such a low temperature that the high-melting glycerides tend to crystallize during the reaction. The highest melting fatty acids are thus effectively removed, and equilibrium is displaced toward the formation and removal of even more of the high-melting glycerides. This continues until no more high-melting acids remain in the reaction mix. The result is segregation of the higher melting acids into a stearin fraction, which can be removed by fractionation. Application of this technique to sunflower oil gives an oil with a lower content of saturated fatty acids and a corresponding increase in polyunsaturated acids.

The use of sodium compounds as catalysts requires that the oil be dry, neutral, and protected by an inert atmosphere, since water, fatty acids, and oxygen all deactivate the catalyst. A common procedure calls for carrying out the reaction under vacuum so that water is removed rapidly. A sodium-based catalyst cannot be recovered from the reaction mixture because it is ultimately converted to soap. Any active catalyst remaining after interesterification is deactivated by the addition of a dilute acid (e.g., acetic acid). Sodium salts and soaps are then removed by washing, which is followed by a light earth bleaching and final deodorization.

3.1.2.4. Fractionation

Fractionation entails separating a fat into fractions with different melting ranges by crystallization and subsequent removal of the crystals (→ Fats and Fatty Oils, Chap. 7.). Such a process might be carried out for various reasons: to remove very high melting components (e.g., the stearin from beef tallow), to concentrate PUFA acids (as in removal of stearin from a directively interesterified sunflower oil), or to select a frac-

tion with a relatively sharp melting range for use as a hardstock component. The latter objective requires two-stage fractionation involving removal of both a stearin and an olein fraction to leave a mid-fraction. The process has been applied to palm oil to obtain a fraction rich in monooleodipalmitin and also to interesterified blends of lauric-rich fats with fats rich in stearic and palmitic residues, thereby producing a fraction rich in monolaurodipalmitin or -distearin glycerides.

Fractionation can be carried out either in a solvent or "dry". Acetone, light hydrocarbons, and methanol are used for *solvent fractionation.* The fat is first dissolved by heating it in the solvent, after which the solution is cooled to a temperature at which crystallization takes place. The temperature required depends on the oil being treated. In a typical palm-oil fractionation, for instance, an acetone solution of the oil is cooled to ca. 20 °C to precipitate the stearin, then to 0 °C to precipitate the mid-fraction. The resulting fat crystals are removed by filtration and solvent is recovered by distillation. Solvent fractionation gives a clean separation of the various glycerides, but it is an expensive operation because of the specialized equipment required for the handling and recovery of highly flammable solvents. Normally it is applied only to fats for products carrying a premium relative to ordinary margarine (e.g., confectionery fats).

In *dry fractionation,* liquid components of the oil blend (i.e., the olein) act in effect as solvents for the solid components or stearin. This means that "solvent" occluded in the separated stearin cannot be removed by distillation as in solvent fractionation. Instead, it remains as part of the stearin composition and the resulting separations are less clear-cut than in solvent fractionation. In order to achieve maximum separation efficiency special care must be taken in the choice of crystallization conditions. The rates of cooling and stirring are particularly critical for producing crystals with the best form for easy separation.

Separation techniques other than filtration may also be used to maximize separation of the olein. One such technique is pressing, in which slabs of stearin are squeezed between the plates of a hydraulic press. This is particularly effective for removing a small amount of olein from a large amount of stearin, as in the preparation of palm kernel stearin. Another valuable technique is the Lanza separation. Here the crystal slurry is stirred with an aqueous detergent solution, typically sodium lauryl sulfate. The detergent solution preferentially wets the solid fat crystals and draws them into the aqueous phase. Olein is then separated from the aqueous phase by centrifugation, leaving the stearin in the aqueous phase. The stearin is recovered by heating this mixed phase above the melting range of the stearin, followed by further centrifugation.

3.2. The Aqueous Phase

The aqueous phase comprises ca. 15 % of a conventional margarine but it can account for 60 % or more in a low-fat spread.

3.2.1. The Aqueous Phase in Conventional Margarines

In conventional margarine the aqueous phase consists mainly of water, and its main function is as a carrier of flavoring components. In some parts of the world, special flavors are common (cf. Section 2.1), but in the Western World margarine is expected to taste as much like butter as possible. Considerable research has been devoted to achieving this end. Most of the effort has focussed on identifying the natural flavor components in butter. This resulted initially in the introduction of diacetyl and short-chain fatty acids. Later work led to the use of the γ- and δ-lactones of C_8–C_{14} hydroxy fatty acids. Most manufacturers now employ their own closely guarded "flavor cocktails". Butter-like flavor is enhanced by use in the aqueous phase of skimmed milk that has been ripened with lactic acid bacteria.

Another flavor-contributing component of the aqueous phase is salt. The salt level must be carefully adjusted to correspond to the preferences of a particular national market.

3.2.2. The Aqueous Phase in Low-Fat Spreads

In low-fat spreads the aqueous phase not only carries the flavor components but also contributes to the structure of the product. Thus,

with a fat content ≤ 40 % it is necessary to invoke special techniques to ensure adequate rigidity. Rigidity can be achieved in two ways. A spread in which the aqueous phase is essentially water, can be stiffened by the use of powerful emulsifiers which disperse the water more finely and stabilize the resultant droplets. Such stable, highly dispersed droplets augment the limited number of fat crystals present in the system (cf. Chap. 4), thereby increasing the effective size of the elastic crystal network. The extent to which this principle can be applied is limited, however; an emulsion that is too stable adversely affects the mouthfeel of the product, producing an initial thick, heavy sensation that is followed by a rapid watery breakdown. The alternative is to thicken the aqueous phase by adding components that either gel the water or increase its viscosity to such an extent that the dispersed water droplets behave like semisolid particles. Components that have been used for this purpose include gelatin, caseinates derived from milk proteins, water-soluble carbohydrate gums (e.g., carrageen and locust bean gum), and modified starches. Additives are introduced at levels of 5 – 15 % depending on type. Together with the processing conditions such additives can significantly affect the mouthfeel of spreads, and the creamy breakdown associated with conventional margarines is more easily achieved in low-fat spreads in which thickeners have been used than in one with an unthickened aqueous phase.

3.3. Emulsifiers

Emulsifiers are essential to both conventional margarines and low-fat spreads. They facilitate dispersion of the aqueous phase and also stabilize the resulting dispersion. A typical emulsifier contains both polar and nonpolar groups. Since the affinities of these groups for the two phases differ, the emulsifier tends to concentrate at the phase interface. This alters both the interfacial tension and the interfacial viscosity, thus aiding the formation and stabilization of droplets of one phase dispersed within the other. Whether the dispersion is one of water in oil or oil in water depends on the relative number and nature of polar and nonpolar groups, the so-called hydrophilic – lypophilic balance (HLB). Emulsions are discussed in detail elsewhere (→ Emulsions).

The emulsifiers most widely employed in margarine and other spreads are monoglycerides and lecithins. In *monoglycerides*, polar character is provided by two free hydroxyl groups, and nonpolar character by the long-chain fatty acid moiety. Monoglycerides are prepared by interesterification of natural triglyceride oils with glycerol and concentrated by distillation.

In *lecithin,* nonpolar character is again conferred by long-chain fatty acid residues, but the polar component is the phosphatidyl moiety attached to the third hydroxyl group of glycerol (→ Lecithin). The lecithin most widely used in margarine occurs as a natural component of soybean oil and is separated from the oil at the degumming stage of refining.

Margarines designed for special applications (e.g., pastry) may incorporate specially selected emulsifiers to enhance particular end-use properties (see Section 5.1).

3.4. Minor Components

Minor components added to margarines include antioxidants, coloring agents, preservatives, pH regulators, and vitamins (→ Foods, 3. Food Additives).

The preferred *antioxidants* are tocopherols, which may be either synthetic or derived from natural sources, but synthetic compounds such as butylated hydroxytoluene (BHT) or butylated hydroxyanisole (BHA) are sometimes used instead. The antioxidants pick up or release free radicals in the course of assuming stable quinone or hydroquinone configurations and thereby interrupt the chain of free radical reactions responsible for autoxidation (→ Antioxidants). Natural tocopherol antioxidants occur in most vegetable oils, so that the use of antioxidants is generally limited to fat blends consisting primarily of animal fats.

Coloring agents are usually carotenoids. Examples include β-carotene, either from palm oil or synthetic, and annatto, a mixture of bixins derived from the Central American fruit of that name.

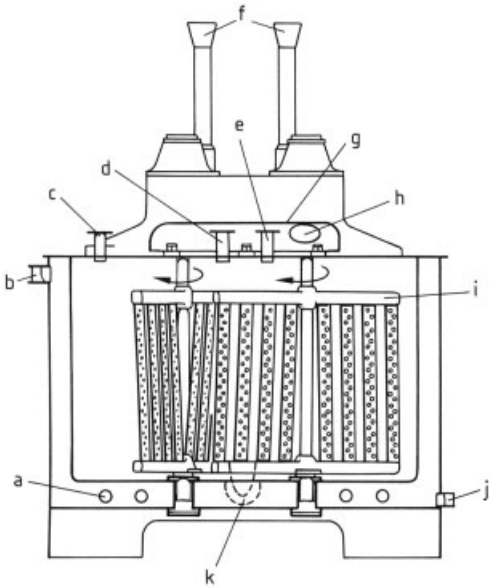

Figure 1. Emulsion churn for margarine production (reproduced with permission from [1])
a) Perforated steam coils; b) Heating jacket outlet; c) Salt solution inlet; d) Milk inlet; e) Oil inlet; f) Vents; g) Hinged flap; h) Inspection glass; i) Stirrer with pierced plates; j) Heating jacket outlet; k) Product outlet

The most widely used *preservative* for the control of biological activity is common salt, but in some circumstances potassium sorbate is added, particularly in low-fat spreads where increased water content favors microbial growth. Control of the pH may also be used to inhibit biological activity. Control of pH is also necessary in spreads in which the aqueous phase has been thickened with proteins (e.g., caseinates) because the resulting viscosity or elasticity is a function of pH. Salts of weak acids (e.g., sodium citrate or sodium phosphate) are used for this purpose.

Most food control authorities require the addition of vitamins A (27 – 33 IU/g) and D (2.8 – 3.5 IU/g) to margarine as a means of assuring adequate nutritional value.

4. Processing

Margarine processing procedures are designed to achieve three objectives:

1) dispersion of the aqueous phase as a water-in-oil emulsion throughout the fat phase,
2) formation of a fat crystal network to stabilize the resulting emulsion and give the product some degree of firmness, and
3) modification of the crystal network to produce the desired firmness and confer plasticity.

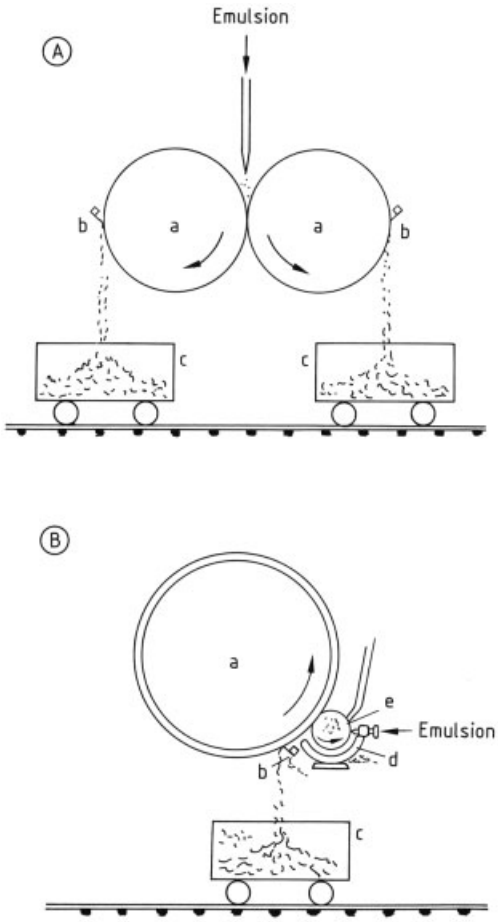

Figure 2. Cooling drum systems for margarine production (reproduced with permission from [1])
A) Double-drum system; B) Single-drum system with feeding rollera) Cooling drum; b) Scraper; c) Collecting wagon; d) Trough; e) Feed roller

Common procedures can be broadly classified according to whether these objectives are achieved independently (churn process) or simultaneously (votator process).

4.1. Churn Process

The churn process was the only one applied to margarine production prior to the mid-1930s.

In the first stage, *emulsion formation,* the aqueous phase and the fat phase, together with minor ingredients, are introduced into a stirred jacketed vessel at a temperature just above the melting range. The stirrer consists of a pierced plate that occupies the whole cross-section of the vessel (Fig. 1). The two phases are then mixed vigorously to produce a fine water-in-oil emulsion.

In the second stage, *crystal formation,* the emulsion is rapidly chilled by passage through heat exchangers or by use of a drum cooler (Fig. 2). In the drum cooler a thin layer of emulsion is spread on the surface of a cooled rotating drum and scraped off with a blade. As a result of rapid cooling, crystal nuclei develop and grow to form the network required for product rigidity. This initial product is too brittle for commercial use, however, and it must therefore be subjected to *plasticization*, in which it undergoes mechanical working (plodding) to break down part of the crystal network. Appropriate mechanical processes include extrusion through a screw press, squeezing through a gap between rollers, or agitation by rotating knife blades.

4.2. Votator Process

The term "Votator" was first coined by the Girdler Company in the United States to describe the continuous cooler – stirrer unit that they introduced in 1936. A flow scheme for the votator process is shown in Figure 3. Two votator units are involved in margarine production, the A unit and the C unit.

The *A unit* consists of a jacketed tube fitted with a rotating shaft that fills most of the cross section of the tube, leaving an annulus of 5 – 10 mm. This annulus is occupied by blades attached to the rotating shaft and touching the inner wall of the tube. The jacket is cooled by a refrigerant such as liquid ammonia. The shaft is rotated at high speed ($\geq$ 500 rpm) and the product mixture is forced through the annulus. It is thus subjected to rapid cooling by contact with the refrigerated walls of the jacket as well as to high shear.

The *C unit* also consists of a jacketed tube with a rotating shaft, but the shaft is smaller in diameter relative to the cross section of the tube, so the annulus is much wider. Moreover, the shaft is fitted not with wall-scraping blades but with a series of pegs. As the shaft rotates, the pegs move between similar pegs attached to the inner wall of the tube, subjecting the contents of the C unit to mixing at controlled temperature but not to the high shear conditions prevailing in the A unit.

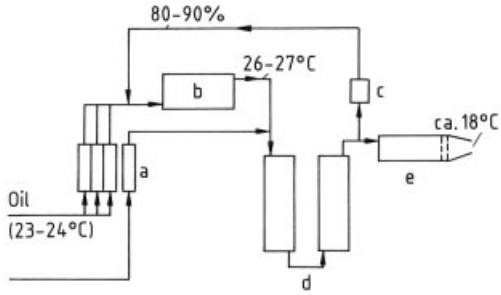

Figure 3. Schematic diagram of a votator plant with precrystallization and recycling (Unilever)
a) High-pressure metering pump; b) C unit; c) Pump; d) A unit; e) B unit

The required proportions of fat and aqueous phases may be continuously stirred in a premix tank to create a loose emulsion suitable for pumping into the votator line. Alternatively, the components may be pumped from independent holding vessels into the votator line through a proportionating pump adjusted to produce the required volume ratio.

The combination of high shear and rapid cooling in the A unit results in a finely dispersed emulsion with an even distribution of fine fat crystals. Low shear agitation in the C unit permits further growth of these crystals without the formation of an interacting crystal network. Only under subsequent quiescent conditions is further crystal growth able to produce such a network. Thus, it is the C unit that controls the ultimate plasticity of the product. If all the crystal nuclei emerging from the A unit were allowed to grow into crystals, all contributing to the crystal network, then the product would be too brittle.

A typical votator installation utilizes combinations of A units and C units coupled in series. The manufacturer selects a set of temperatures, residence times, and agitation conditions specifi-

cally tailored to the mixture composition and the required product type.

A further unit, the *B unit*, is necessary for margarine that is to be packaged in a paper wrapper. This unit consists of an open tube corresponding in cross-section to the final wrapped slab. The tube is without a rotating shaft, so product passes through with minimum agitation. This not only molds the product into the required shape for packing but also provides the degree of product firmness necessary for handling in a wrapping machine.

Product that is to be packaged in tubs or cartons is extruded directly into the container from the last votator unit. Firming proceeds to completion in the container during subsequent storage. The success of this technique depends on precise control of the degree of crystal network formation in the processing line, i.e., on an optimal balance of A- and C-unit working. "Overworking" (premature formation of a crystal network, which is subsequently destroyed by further mechanical working) produces a product that is too liquid, because most of the crystallization takes place in the absence of an interactive network. Such a product only becomes firm in the tub after prolonged storage, if at all. At the other extreme, too much potential for crystal growth in product leaving the line leads to excessive posthardening and a birttle product. Nevertheless, some posthardening must occur in the tub, since attempting to extrude stiffened product would result in high back pressure along the line, which would in turn interfere with the mechanical working of the system.

The votator system is intended for continuous operation, but it is nonetheless subject to occasional stoppages, with breakdown of the packaging equipment being the most common cause. Interruption of throughput typically results in product setting up in the votator line, with the potential for considerable loss of product, time, and energy. Consequently, the whole system (the mixture composition and the lay-out of the processing units) is designed to facilitate recycling without altering the nature of the final product or the settings of the processing parameters.

A satisfactory production line must also include provision for regular and rigorous cleaning and sanitization to eliminate the possibility of microbial contamination. This entails pipelines for detergent solutions as well as steam lines for equipment scalding. It is especially important to avoid dead pockets that might trap partially processed material and serve as isolated breeding grounds for microorganisms.

5. Specialty Margarines and Shortenings

Most margarines are intended for spreading on bread, but some are also used in cooked foods, including cakes and pastries, as well as in the preparation of fried foods. Specialty margarines exist for commercial use in the latter applications.

5.1. Pastry Margarines

Pastry margarines are mainly used for short crust pastry, cake mixes, and flaky pastry; i.e., flour confections in which fat is a major ingredient. Such baked goods differ considerably from bread. In *bread making*, the object is to work the dough so that it develops a gluten structure; i.e., protein molecules within the flour grains are encouraged to link up to form long chains and extended sheets, providing the characteristic cellular structure of bread. Fat is kept to a minimum in bread mixes because fat levels > 3 % interfere with the development of gluten structure. In *short crust pastry* the aim is reversed. Long protein chains are not allowed to develop, resulting in a product that is crumbly and, unlike bread, breaks easily into small fragments when bitten. This is achieved by giving the pastry mixture a relatively high fat content– ca. 30 % or more. Individual grains of flour thereby become coated and the gluten structure cannot extend from grain to grain. Baked pastry consists essentially of an aggregate of partially gelatinized starch granules cemented together with fat. The fat must therefore be easily and intimately mixed into the flour, and it must also have sufficient solid content to contribute to the rigid structure of the finished pastry. This calls for a fat blend containing relatively high-melting components together with a reasonable proportion of liquid fat, thereby providing plasticity over a wide temperature range. Animal fats such as lard and beef suet have traditionally been used for this purpose, but more

precise control of favorable properties is possible through the use of blended oils, including palm oil and hydrogenated vegetable oils. Water is essential only as a flavor carrier. Indeed, too much water renders the pastry sticky due to overgelatinization of the starch. The final water content of a mix must usually be adjusted by the baker to suit a particular flour, so margarine especially designed for short crust pastry frequently contains less water than conventional margarines.

The products referred to as *shortenings* contain no water. Soft, pumpable shortenings have been developed to meet the needs of large-scale bakers and confectioners, who must handle ingredients in bulk. Such shortenings are collected from the votator line in bulk containers. They are delivered in tankers and stored in bulk storage tanks. Storage and transport containers are held at 20 – 25 °C to maintain a pumpable consistency.

Cake mixes resemble short crust pastry mixtures in that they yield a product with a rigid, yet easily fragmented texture. However, a cake mix must rise (i.e., expand) during baking to produce a foam of air bubbles entrapped in baked dough. This effect depends upon the creaming power of the dough, i.e., its ability to entrap air during the mixing process. Here the water phase is an asset: not only does water dissolve air during the mixing stage but it is also converted to steam during the baking stage, thus increasing the number and volume of entrapped bubbles. Emulsifiers also enhance the creaming effect, so that cake margarines and shortenings often have a high monoglyceride content (2 – 5 % compared to 0.5 % in conventional margarines). This is the basis for the term *high ratio shortenings* [17]. Such a blend permits incorporation at a higher level of water into the mix as well as a higher ratio of sugar to flour, resulting not only in a greater volume increase but also a tenderer crumb and one that is less liable to go stale on keeping.

The creaming effect is inhibited by fat blends that contain a preponderance of a single glyceride species. Natural lard, for instance, is not ideal for cake making because of its high oleodipalmitin content, but its performance is greatly improved if this glyceride is broken down by interesterification [18].

The characteristic texture of *flaky pastry* in which thin layers of pastry are separated by layers of air, results from a special cake structure in which air is present not in the form of spherical bubbles but as laminae separating continuous layers of pastry. To achieve this the margarine or shortening is not intimately mixed into the flour as in a normal cake mix but remains in discrete, evenly distributed layers. In one process the fat is added to the dough mix in small lumps. The mixture is then rolled out, flattening the lumps of fat into layers. In an alternative process the dough is rolled out first and the margarine or shortening is subsequently spread over its surface. A further layer of dough is then rolled on top, followed by another layer of fat, and so on. During baking, air and water in the margarine layers expand, forcing the dough layers apart and producing a flaky texture. Margarine designed for this application must be plastic enough to be spread or rolled into thin layers, but not so soft as to diffuse into the surrounding dough layers before starch gelatinization has set these into flakes. Thus, a fat blend is required that offers a wide range of plasticity but a limited proportion of liquid oil [19].

5.2. Frying

Butter has traditionally been a popular medium for frying, so it is not surprising that attempts have been made to use margarine for the same purpose. However, the performance of the two substances in this application is not the same. Whereas butter can be heated without difficulty to the high temperatures required for frying (140 – 200 °C), margarine tends to spatter. Not only does this cause anything in the vicinity of the frying pan to become covered with fat droplets, but spattering is also a potential fire hazard, since the aerosol of fine fat droplets is highly flammable.

Spattering occurs because the fat crystals that help to stabilize the emulsion disappear as the margarine is heated above its melting range. The aqueous-phase droplets thus coalesce into relatively large drops. During frying, water is heated well above its boiling point. Since the water drops are suspended in liquid oil, there is little chance for the nucleation that would initiate evaporation. Instead the drops become superheated. When nucleation eventually does occur (e.g., by contact with the walls of the frying

pan), the drops vaporize very rapidly. The result is a minor explosion as a small volume of liquid water is converted to a significantly larger volume of steam over a very short period of time. Spattering is a consequence of numerous drops undergoing this violent transition over a short period of time.

In butter the water droplets are stabilized by encapsulation in casein micelles. Coalescence is thus inhibited and the droplets can diffuse to the surface, where they evaporate as foam. This desirable state can be approached in margarine by the use of suitable emulsifiers (e.g., the citrate esters of mono- and diglycerides, fatty acid esters of polyglycerol, and lecithin) [20].

Low-fat spreads are unsuitable for frying not only because of their high water content, but also because the gums or proteins used as thickeners are denatured at frying temperature and form unpleasant charred residues on the bottom of the pan.

Shortenings and frying oils do not contain an aqueous phase, so spattering problems do not arise. Spattering occurs here only when water is introduced along with the food that is to be fried (e.g., as lumps of ice on frozen food). In these cases the physical and chemical properties of the fat blend are thus the only factors to be considered in the choice of a frying medium.

As noted in Section 2.2.7, liquid oils are unsaturated, and therefore more susceptible to oxidation than saturated solid fats. Choice of a suitable blend for frying is therefore reduced to achieving a balance between the handling convenience of a liquid oil and the oxidative stability of a solid fat.

The domestic user tends to favor liquid oil, especially for deep frying. It is easier to heat a pan of liquid oil on a domestic stove than to melt a block of solid fat. Furthermore, the domestic user is not overly concerned with oxidative stability, since a given batch of oil is only used for a limited number of fryings. The favorable nutritional properties of liquid oils also have a strong influence on domestic users.

For the industrial fryer, on the other hand, oxidative stability is of major importance because the useful lifetime of an oil is an important economic factor. Industrial frying is a continuous process and is thus dependent upon keeping the oxidative state of the oil at an acceptable level. Ideally, an equilibrium is attained between removal of used oil from the fryer by incorporation into the fried product and addition of fresh oil; the level of oxidation products in the oil in the fryer thus remains constant. The rate of oxidation must be equal to the rate of turnover of the oil, a requirement that can to some extent be met by proper design of the fryer and adjusting the rate of throughput of product. Nevertheless, the oxidative stability of the oil is also a major factor.

Thus, for the industrial fryer, the preferred frying medium is a solid or semisolid fat. Solid fats present a minor inconvenience in that they are liable to congeal in pipelines and holding tanks, but this problem is readily overcome by the installation of suitable heating equipment.

Economic considerations are also important for small-scale commercial fryers, such as restaurants, canteens, and fast-food shops. However, a chef is in more direct contact with the frying process than is the quality control manager of a large frying operation, permitting a quicker reaction to changes in product quality. The choice of fat is therefore influenced heavily by personal preference and local tradition.

6. Packaging

Packing units for margarine vary in size from the very small individual packs used in restaurants (containing only a few grams of product), through domestic packs (250, 500 g), larger packs (25, 50 kg) for commercial bakers and caterers, up to tanker loads of several tonnes for industrial users.

Wrapping materials for retail packs should provide a barrier to both light and air to inhibit autoxidative deterioration, and they must be impermeable to fat and water to prevent leakage and consequent spoilage. For hard-blend margarines, greaseproof parchment paper adequately serves the latter purpose, but aluminum foil is also widely used because of its impermeability to both light and air. Soft blend margarines and low-fat spreads require rigid containers such as tubs made from plastic or waxed cardboard. In all cases the packaging material must be devoid of potentially harmful components that might migrate into the contents. Thus, plastic containers or wrappers must not contain residual monomers, fillers, or other harmful ma-

Table 3. Annual consumption of butter and margarine (10^3 t) ∗

	United States			Western Europe		
Year	Butter	Margarine	Total	Butter	Margarine	Total
1985	224 (20)	908 (80)	1132	1226 (44)	1569 (56)	2795
1986	207 (18)	919 (82)	1126	1189 (43)	1587 (57)	2776
1987	165 (15)	925 (85)	1090	1141 (42)	1599 (58)	2740
1988	155 (15)	907 (85)	1062	1081 (41)	1601 (60)	2682
1989 (projected)	147 (14)	900 (86)	1047	1030 (39)	1610 (61)	2640

∗ Based on information from the Office of Agricultural Affairs, U.S. Embassy, London, and from Unilever Department of Statistics, Unilever House, Rotterdam. Values in parentheses indicate percentages of total.

Table 4. Weekly consumption of spreadable fats (grams per person) in the United Kingdom and the Netherlands ∗

	United Kingdom				Netherlands			
Year	Butter	Margarine	Low-fat Spreads	Total	Butter	Margarine	Low-fat Spreads	Total
1975	159	74	6	239				
1976	147	88	6	241				
1977	133	99	6	238				
1978	130	99	6	235				
1979	128	102	9	239	49	177	36	262
1980	116	108	11	235	51	179	35	265
1981	105	116	14	235	50	178	35	263
1982	91	122	17	230	51	174	36	261
1983	94	116	14	224	50	174	36	260
1984	82	116	20	218	56	171	37	264
1985	79	108	26	213	58	169	38	265
1986	65	116	37	218	59	166	37	262
1987	60	113	37	210	59	162	36	257

∗ Based on information from the Office of the Agricultural Counsellor, The Royal Netherlands Embassy, London, and from The Food Research Association, Leatherhead, United Kingdom.

terials. Appropriate specifications and tests have been drawn up by food regulation authorities (→ Foods, 4. Food Packaging, Chap. 8.).

7. Economic Aspects

Tables 3 and 4 show the changes in pattern of consumption of butter, margarine and low fat spreads over recent years in the geographical areas for which figures are available.

The decrease in total fat consumption is evident in all regions. This reduction is mainly at the expense of butter relative to the other spreads. This is dramatically apparent in the United Kingdom figures. However, the United Kingdom was traditionally an area of high butter consumption. The effect is not apparent in the Netherlands figures where butter consumption has been low compared to margarine for some years. The increase in consumption of low-fat spreads compared to conventional margarine is, however, apparent in both those countries.

8. References

General References

1. A. J. C. Andersen, P. N. Williams: *Margarine,* Pergamon Press, London 1965.
2. S. F. Riepma: *The Story of Margarine,* Public Affairs Press, Washington 1970.
3. R. J. Hamilton, A. Bhati (eds.): *Fats and Oils, Chemistry and Technology,* Applied Science, London 1980.

Specific References

4. W. Normann, GB 1 515, 1903.
5. C. Friedel, J. R. Crafts, *Annalen* **133** (1865) 207 – 211.
6. W. Normann, *Chem. Umsch.* **30** (1923) 250 – 251.
7. E. W. Eckey, *Ind. Eng. Chem.* **40** (1948) 1183 – 1190.

8. *Official Methods of American Oil Chem. Soc. Cc13b45,* American Oil Chem. Soc., Champagne, Illinois 1987.
9. *Official Methods of American Oil Chem. Soc. Cd10-57,* American Oil Chem. Soc., Champagne, Illinois 1987.
10. A. J. Haighton, L. F. Vermass, C. den Hollender, *J. Am. Oil. Chem. Soc.* **48** (1971) 7 – 15.
11. D. Chapman, *Chem. Revs.* **62** (1962) 433 – 456.
12. A. J. Haighton, *J. Am. Oil Chem. Soc.* **36** (1959) 345 – 348.
13. D. Waddington in R. J. Hamilton, A. Bhati (eds.): *Fats and Oils, Chemistry and Technology,* Applied Science, London 1980, pp. 25 – 45.
14. D. R. Marker, L. C. Brown, L. C. Wiederman, *J. Am. Oil Chem. Soc.* **35** (1958) 130 – 133.
15. J. M. Denman, A. M. Beers, *J. Texture Stud.* **18** (1988) 303 – 318.
16. P. N. Williams: *Linear and Non-linear Programming in Industry,* Pitman, London 1967.
17. J. Moncrieff, *Baker's Dig.* **44** (1970) 60.
18. Swift and Co., US 2 625 478, 1953 (K. F. Mattel, F. A. Norris).
19. E. A. McGill, *Baker's Dig.* **49** (1975), 28.
20. Refining Unincorp., US 2 640 780, 1953 (M. Matlikow).

Marzipan → Confectionery

Matches

MICHAEL COX, Swedish Match, High Wycombe, United Kindom

1. **Introduction** 1
2. **History** 1
3. **The Matchmaking Process** 3
3.1. **Splint Manufacture** 3
3.2. **Outer-Box Making** 4
3.3. **Inner-Box Making** 5
3.4. **Composition Making** 5
3.5. **Match Assembly Line** 6
4. **Raw Materials** 7
5. **Head Composition** 9
6. **Friction Surfaces** 9
7. **The Chemistry of Matches** 10
8. **Match Testing** 12
9. **Match Standards** 13
10. **Quality Assurance** 13
11. **Safety and Handling** 13
12. **References** 14

1. Introduction

Matches are readily available in every country of the world. The match industry grew rapidly during the 1800s, but it is now declining in most developed markets due to reduced smoking, competition from mechanical lighters, and the growth in automatic ignition on domestic appliances. However, match usage is still growing in some developing countries such as India.

Worldwide production is estimated at around 6×10^{12} matches per year. To put this into context, if they were all placed end to end they would measure about 260×10^6 km, enough to reach to the sun and most of the way back. During every second of every day about 190 000 matches are being struck.

The match industry uses about 6×10^6 trees each year for making match sticks (splints), and a similar number of trees are felled to make cardboard for the outer boxes. To make the head composition, some 120 000 t of chemicals are used worldwide. The match industry is the most significant user of potassium chlorate, it is the only consumer of phosphorus sesquisulfide, and it has a significant demand for red phosphorus.

2. History

The development of matches depended on the availability of the right materials and an understanding of chemistry. In the early 1800s both of these factors had advanced sufficiently to make the invention of matches possible. Of particular significance was the manufacture of phosphorus and potassium chlorate on an industrial scale. From ca. 1820 to 1840, there was a rush of inventions relating to different types of matches.

The history of matches is documented in [4 – 6]. Here only a few key highlights of the history that caused significant changes to matches or the structure of the industry are mentioned.

The first matches, or friction lights as they were then known, appeared in 1826 and they were produced and sold by JOHN WALKER, a pharmacist from Stockton on Tees in the United Kingdom. These matches consisted of wooden splints tipped with a mixture of antimony sulfide and potassium chlorate together with a binder. They could be ignited by striking the tip on a rough surface. These matches were sold in bundles.

The Walker matches were not very sensitive and the phosphorus match, invented by SAURIA in the 1830s soon superseded them. These matches contained white (or yellow) phosphorus and were much easier to light. Like the Walker matches they were ignited by striking against a rough surface. These matches were so sensitive that sometimes they would ignite spontaneously if left in the sun. Furthermore, white phosphorus is very toxic, and this would have a significant effect on the development of the match industry.

Up until this point in the development of matches, all ingredients necessary for easy igni-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_163.pub2

tion were contained in the match head. Typically the active components were a powerful reducing agent (e.g., white phosphorus) together with a powerful oxidant (e.g., potassium chlorate). In 1844 Pasch invented a new type of match in which the powerful reducing agent (he used red phosphorus, the nontoxic and less reactive form of phosphorus) was painted on the side of the box, and the match head contained the powerful oxidant (potassium chlorate). These matches could only be struck on the painted friction surface on the side of the box. The Pasch match was developed and produced by the Lundstom brothers in Jonkoping, Sweden. The matches were known as safety matches because they were nontoxic, in contrast with the Sauria phosphorus matches containing white phosphorus.

At this stage in the development of matches the methods of production were simple and labor-intensive. The splints were cut by hand and manually placed in perforated plates or frames for dipping, boxes were made manually by outworkers, and the matches were packed by hand. From 1862 onwards the process was revolutionized by the work of Alexander Lagerman who single handedly developed machines for mass producing splints and for dipping and packing the product. The techniques developed by Lagerman for carrying out these processes are broadly similar to the techniques that are still in use today.

For a long time the Sauria phosphorus matches and the Pasch safety matches were produced and sold generally. However, the toxicity of white phosphorus led to a severe occupational disease, one of the first to be identified, amongst match workers called phosphorus necrosis (phossy jaw). Phosphorus matches were also responsible, accidentally or intentionally, for some deaths amongst the general public. To protect the public and the match workers, the sale of phosphorus matches was prohibited in Finland in 1872. Over the next 20 years more and more countries outlawed the sale of phosphorus matches, and in these countries the Pasch safety match then dominated the market.

In 1898 Sevene and Cahill invented a new type of match, in which toxic white phosphorus was replaced by the much less toxic phosphorus sesquisulfide. These matches could be struck on any rough surface, like the Sauria phosphorus matches. The use of white phosphorus in matches declined rapidly after this as all manufacturers changed over to the new compound.

The consequence of this history is that those countries that banned the use of white phosphorus between the years 1870 and 1900 even today only consume safety matches whereas those that banned the use of white phosphorus after 1900 use both safety matches and phosphorus sesquisulphide matches (usually called strike-anywhere matches).

The use of white phosphorus in matches was eventually banned throughout the world by international treaty (The Berne Convention, 1902). This is one of the few treaties banning the use of a substance on a global scale.

Until 1990 all matches contained some toxic heavy metals and sulfur. Increasing environmental pressures have caused a few manufacturers in Western Europe to develop new formulations that do not contain these substances.

Throughout the development of matches wood has been the preferred stem material by most manufacturers. However, in some countries other materials are used for the match stem. In Spain, Italy, and some Latin American countries, stems are made from waxed paper that is crimped and shaped into rods by pulling it through a forming die. For advertising matches, cardboard is often used for the stem. The cardboard is punched and cut into strips with teeth, and each tooth is tipped with match head composition. These combs are then fastened inside a cardboard cover that carries the advertising message.

Aside from the traditional matchbox (for retail matches) and matchbooks (for advertising matches) various other types of containers or packaging have been developed and used but they have never seriously rivaled the traditional forms of packaging.

No account of the history of matches can omit mention of the Swede, Ivar Kreugar. After the end of World War I, Kreugar acquired match manufacturing operations around the world until he eventually controlled 60 % of the world production of matches. This lead to the establishment of the Swedish Match company (Svensk Tandstickor AB) that dominated world trade in matches. Kreugar committed suicide in 1932, and today the Swedish Match company is still a global company with interests in matches, match manufacturing machinery, lighters, and tobacco

products, albeit with a much smaller share of the world match trade.

3. The Matchmaking Process

Match making is typically carried out in integrated plants where wood, cardboard and chemicals are the raw materials and the fully packed matches are the output (see Fig. 1). However, not all match factories make there own splints. Many factories buy in ready-made splints. These are usually supplied in boxes, although in some cases they are strapped together in big bundles known as cheeses.

3.1. Splint Manufacture

Wood Handling and Pretreatment. Wood arrives at the factory in the form of logs, typically 3 m long and 200 – 400 mm in diameter. These are stacked in log piles until they are needed in the process. During dry weather these piles are sprayed with water to prevent the wood from drying out.

The logs are brought into the factory when needed. If the logs are frozen then they are first thawed in steam-heated chambers. Some species, notably the resinous pines, also require steam cooking to soften the timber before further processing.

Crosscutting and Debarking. After pretreatment the logs are cut into billets about 600 mm long. The bark is removed from the billets. Sometimes the logs are debarked before cross-cutting into billets.

Peeling, Chopping, and Sieving. The debarked billets pass to a peeling lathe where they are rotated against a sharp knife, and the log is converted into a continuous sheet of veneer. To produce veneer with the best strength, the timber must be saturated with water at this stage. If it is too dry then the veneer will be weak and full of cracks.

The veneer is cut into sheets about 3 m long and stacked into a pile that is fed continuously to a chopper that guillotines through the pile of veneers. As the guillotine knife passes through the stack it is preceded by some knives at right angles to the guillotine blade and pitched the length of the desired match splint apart. Thus at each stroke of the guillotine blade the veneer is cut into individual match splints. Splints as cut contain about 50 % water.

The splints leaving the chopper contain some short splints and some long splinters. These arise from the edges of the billet and from where knots are cut. It is usual to remove as much of this waste as possible at this stage by butt sieving.

Impregnation. The splints are impregnated with a chemical to prevent after glow. Usually this is ammonium dihydrogen phosphate or diammonium hydrogen phosphate. Impregnation can be carried out by dipping the splints into a bath containing a dilute solution (ca. 2 %) of the afterglow inhibitor or by spraying the splints with a more concentrated solution (ca. 12 %) of the afterglow inhibitor in a rotating drum. The latter procedure is preferred since less water is added to the splints and consequently less water has to be evaporated during the drying process.

Drying. The impregnated splints are then dried. Usually the splints fall onto a flat perforated conveyor and pass through a heated drying chamber maintained at 80 – 100 °C. The residence time in the dryer is typically about 30 min. In the dryer a strong airflow is maintained through the bed of splints, and water-saturated air is exhausted. The drying conditions are critical; too high a temperature leads to scorching of the splints, which discolors and weakens them, while too low a temperature or air flow leads to poor drying and problems in the following polishing process. The typical water content of the splints is about 3 % after drying.

Polishing, Cleaning, and Sieving. At this stage the splints are dry but they are rough, i.e., the surface of each splint has many protruding fibers. The splints are passed to a rotating polishing drum where they are tumbled together. In the tumbling process the small protruding surface fibers are rubbed off, and the splints become much smoother. If the splints are dry then the surface fibers are brittle and the splints polish well. If the splints have a water content greater than 4 %, then the surface fibers are much tougher, and polishing is much less effective. Sometimes blocks or flakes of wax are added to the polishing drum. This is said to increase the effectiveness of the polishing process.

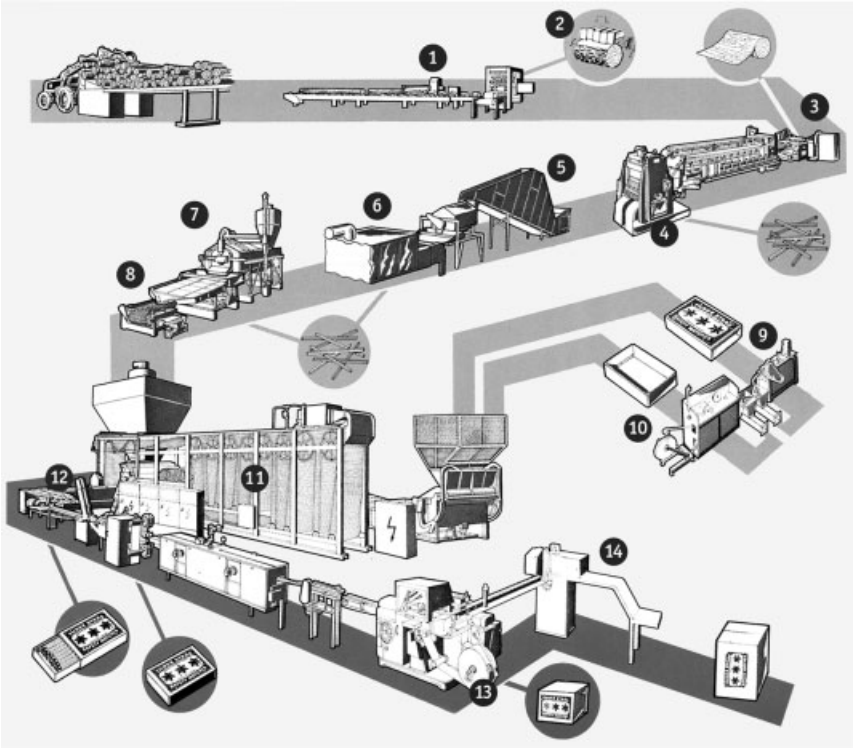

Figure 1. Schematic diagram showing match production
1. Crosscut saw, 2. Debarking machine, 3. Peeler, 4. Chopper, 5. Impregnating equipment, 6. Dryer, 7. Polishing drum, 8. Sorting and sieving machines, 9. Bar machine for outer boxes, 10. Inner box machine, 11. Continuous machine, 12. Filling into boxes, 13. Wrapping machine, 14. Final packing

The butt sieving carried out immediately after chopping does not completely remove short splints. Furthermore, many intrinsically weak splints are broken during the subsequent impregnation, drying, and polishing operations. Hence, after polishing, the splints must be cleaned and sieved to remove undersized pieces. In the first stage of this process the splints pass down gently sloping vibrating tables with holes in. The diameter of the holes is just under half the splint length. Short splints pass through the holes, whereas those of correct length pass over the screen. Many strongly curved splints are also removed in this sieving operation. After this sizing to the correct length the splints then fall onto a second sieving table. On this vibrating table the splints fall into a series of small cups or pockets which maintain them vertical. The bottom of each pocket is perforated with holes that permit only splints with the required cross section to pass. All the oversize splints remain in the pockets. The finished sieved splints then pass to bunkers for storage.

Waste. At each stage in the process some wastage occurs. Bark is rejected, as are the first cuts into the billet on the peeling machine. On the peeling machine it is not possible to peel right through to the center of the log. Usually a center core of about 75 mm in diameter is rejected at the end of the peeling operation. More waste material is filtered out during the butt sieving and final sieving and cleaning operations. Typically only 40 – 50 % of the logs entering the process is recovered as good splints. The remaining 50 – 60 % is waste. This is typically converted to chips and used as fuel for a boiler to generate process steam, which supplies heat for the drying operation and other processes in the factory.

3.2. Outer-Box Making

Printing. Various processes are used to print the outer box. These can be reel fed (for large

printing runs) or sheet fed (for smaller printing runs).

One variant uses a dedicated printing machine (Chambon machine) to take reels of cardboard up to 500 mm wide and produce cut printed skillets complete with friction paint. The printing process can be letterset or, more commonly these days, offset litho, enabling high-quality, high-definition, full-color images to be reproduced. UV inks (ultraviolet curing inks) and varnishes are commonly used, and these facilitate high-speed printing (web speeds of up to 140 m/min can be achieved). Friction paint is applied at the same time by using a gravure printing process. The friction paint is water-based, and high-speed radio-frequency drying is performed immediately after friction-paint application. After printing, the web passes through a rotary cutter, creaser, and stacker mechanism. The stacked bundles of skillets are supplied to the box-forming machine.

In another process, sheets of cardboard are fed to a conventional printing machine and printed. The printed sheets then pass to a second machine where the friction surface is applied by a gravure printing process. The sheets then pass to a third machine where they are then diecut and creased to form the skillets in a press. The skillets are then stacked into bundles.

In a third variant sheets are printed and cut into strips. Each strip consists of many skillets with the friction surfaces aligned together. The strips pass over a creasing roll and then a coating roll that applies a continuous strip of friction paint along the length of the strip. The strips are dried with hot air or IR lamps. The dried strips are slit into separate skillets and stacked in bundles.

There are other variants for producing printed skillets with a friction surface, but the three outlined above are the most common. Many match factories purchase ready made printed and creased skillets complete with the friction surface from specialist suppliers.

Folding. To produce the outer boxes, bundles of skillets are loaded onto a bar machine. In this machine skillets are fed from the bottom of the bundle by a conveyor system under a bar with the internal dimensions of the box outer. A fast setting glue (usually animal glue) is applied to one edge of the skillet, and the skillet is folded around the bar to form the outer box. A typical bar machine can produce outer boxes at speeds of around 1100 boxes per minute. It is common practice, nowadays, to feed the boxes directly to the match assembly line so that the orientation of the outer box with respect to the inner is always maintained.

Other types of machines for making outer boxes are available, but these are usually much slower than a bar machine. Consequently several machines are necessary to supply boxes at a sufficient rate to feed the match-assembly machine. This leads to complex conveyor systems on the infeed of the match line.

3.3. Inner-Box Making

Inner boxes are produced from reels of cardboard. The board is fed to the inner-box machine, which stamps out the correct shape, applies adhesive, and folds the box into its final form. Modern inner-box machines do the forming and gluing on a rotating drum and work at similar speeds to the outer box bar machine. Older machines punch and press each box individually, and these work at slower speeds. Several of these are then necessary to supply one match assembly line.

3.4. Composition Making

The making of match head composition used to be carried out in relatively small batches of around 50 kg. Each batch would be passed through a cone mill to ensure good dispersion of all the ingredients. This practice is very labor intensive and results in considerable batch-to-batch variation.

Match head compositions typically contain water, animal glue (as binder), potassium chlorate (as oxidizer), inert materials, sulfur or red phosphorus, and various minor ingredients to control foaming, viscosity, and microorganism growth. In addition to these ingredients, it is also important that the match composition contains a controlled amount of finely dispersed air.

Modern practice is to prepare composition in much larger quantities (ca. 500 kg) in a high-speed dispersing mixer to ensure uniform dispersion of all the ingredients. Normally, hot water is added to the mixer at the start, and then animal

glue, and the mixture is stirred at low speed until the glue is dissolved. The rest of the ingredients are then added, and the stirrer speed is increased. Once the ingredients are well dispersed, air is injected into the bottom of the mixer. Air injection is stopped once the composition reaches a certain volume. In this way it is possible to produce large quantities of head composition with a consistent amount of entrained air.

3.5. Match Assembly Line

The match assembly line, also known as the continuous machine, is fed with splints, paraffin wax, head composition, inner boxes, and outer boxes. The line assembles these components to form the finished boxes of matches.

Splints are fed from a hopper into a system that aligns them and punches them in rows into perforated steel bars about 1.5 m wide. At each stroke of the machine, one row of splints is inserted into a row of holes. The bars are linked together to form a continuous mat that conveys the splints through the whole process.

The splints are preheated by passing them over a heated plate (ca. 100 °C). This ensures satisfactory absorption of paraffin wax in the next stage.

The tips of the warm splints pass continuously through a bath of paraffin wax, typically maintained at 125 – 155 °C. The tips are immersed into the wax to a depth of ca. 5 mm and for from 2 – 5 s. The paraffin wax is the main fuel that sustains the flame once the match is alight.

After dipping in paraffin, the splints again pass over a heated plate that allows the paraffin wax to remain liquid so that it can be absorbed by the wooden splint. If paraffin wax remains on the surface of the splint, adhesion of the match head is compromised.

After this post-heating the splints are allowed to cool for ca. 3 min before they reach the composition-dipping station.

All the previous stages take place with the bar mat moving continuously. At the composition dipping station the bar mat must stop moving. A number of bars are pressed down so that the tips of the matches are dipped in the head composition. The splits are withdrawn slowly from the head composition to form a well-shaped match head.

The surface of the head composition must be continuously scraped (except during dipping) to prevent the formation of a skin. Various mechanisms have been designed to do this but the band-dip mechanism is generally recognized as being the best.

After dipping, the bar mat again moves continuously. A "lost-motion" mechanism is incorporated around the dipping station to facilitate this transition from continuous movement to intermittent movement and back to continuous movement.

After dipping, the matches remain on the continuous machine conveyor for ca. 40 – 60 min to allow the match heads to dry. To facilitate this, the conveyor zigzags vertically or horizontally along the length of the machine.

The conveyor finally brings the matches to the punch out. Here the matches are pushed out of the bars with pins and they fall into cups that then transfer them to the boxes. The inner and outer boxes are supplied on separate conveyors to the punch out mechanism.

After filling, the boxes are closed, usually in two stages. First the inner box is pushed about half way over the outer box. The surface of the outer box is then brushed to push any protruding matches into the box. Then the box is completely closed.

The boxes finally pass to an automatic wrapping station where they are wrapped in packets of 6, 10, or 12. Polypropylene film is generally used for wrapping, but paper wraps are also used. There are many packing options for the wrapped packets. The most common is to pack them into stout cardboard cases containing about 1000 boxes, although trays, caddies, and other packs are also common. Packing can be done by hand or automatic machines, depending on labor costs and the sophistication of the plant.

The continuous machine typically has 2×10^6 matches on the conveyor at any one time, and the production rate is around 2.4×10^6 matches per hour (up to 75 000 boxes per hour depending on the contents of the box).

Production of strike-anywhere matches is in general similar to the production of safety matches. The most important difference is that the outer box skillets are not printed with friction

paint. The skillets are formed into outer boxes. The rough friction surface is applied to the box either before or after filling the boxes.

4. Raw Materials

The main raw materials for making matches are wood, cardboard, paraffin wax and chemicals.

Wood. Wood for making match splints must satisfy several criteria:

- It must be straight grained and easy to peel on a peeling lathe.
- It must be stress-free while it dries; otherwise, the splints become too curved during the drying process.
- It must be strong. The higher the strength of the timber the thinner the cross section of the splints can be and the greater the yield of splints per cubic meter of timber.
- It must be tough and fibrous. Some timbers are strong but very brittle. If matches made from such timber broke during striking, then the end with the burning head is liable to detach from the part that is being held, leading to a potentially hazardous situation. Under the same conditions, splints made from tough fibrous timber break but the two halves remain attached to one another.
- The timber must be porous so that it will absorb paraffin wax. Wax is the main fuel for the burning of the match once the head has finished burning. It is important for the wood to absorb sufficient paraffin wax and that little remains on the surface of the wood after dipping.
- From the viewpoint of appearance the color of the wood should be as light as possible.

The preferred timber for match splints is Baltic aspen (*Populus tremaloides*). This is widely distributed in Northern Europe, and the principal producers are Sweden, Finland, the Baltic States, and Russia. Aspen is rarely, if ever, planted as a commercial timber crop. Aspen occurs adventitiously amongst other timber stands. It propagates readily by suckering, especially after felling. It grows fast and tends to establish itself rapidly and in so doing gains a head start on other species in the competition for light. Similar species exist in North America, and Canadian Aspen has also been extensively used for making match splints.

In Southern and central Europe, various poplar species are used for matchmaking. These timbers are softer, weaker, and more absorbent than aspen. Poplar is also used in the Americas and in Asia.

In other parts of the world many species are used for matchmaking, but in general these lack strength, and frequently are darker than Aspen. Notable are the various pine species (such as *Pinus ellioti*), araucaria, white mutti (India), and morototo.

Ideally timber should be converted into splints immediately after felling. In this state it is saturated with water and is hence easy to peel. However this ideal situation rarely applies and it is necessary to store the timber before use. It is important that stored logs are kept wet during dry weather and that they are properly thawed out before processing if they are frozen.

Wood is a natural product whose properties vary considerably from tree to tree, with the season, with the climate and with the soil. In addition, storage conditions affect the properties and processing of timber. There is thus always a spectrum of properties and a considerable variation from one splint to another.

Cardboard. Cardboard for box making must also satisfy certain criteria:

- It must be stiff enough to form a rigid box that will adequately protect the matches.
- It must print well. This is especially important in the fully developed European and North American markets.
- It must cut cleanly and crease easily.
- The friction paint must adhere well to the board. Certain coatings that are applied to the board may interfere with the good adhesion of the friction paint to the box.
- The board must glue well. Generally the water absorption of the reverse side of the board must be high to promote good adhesion and fast setting of the box-making adhesive.
- The board must not soften too much on exposure to high humidity.

Cardboard for box making varies considerably. Generally it has a basis weight of around 275 – 350 g/m^2 and a thickness between 0.5 and 1 mm. In developed markets, light coated board that provides good printability is used for the

outer boxes. Recycled boards are commonly used for both inner and outer boxes.

Paraffin waxes. Paraffin waxes for matchmaking are mainly derived from petroleum. Satisfactory waxes usually have melting points in the range 49 – 55 °C. The other important parameter is the oil content of the wax. If this is too high then head adhesion may be adversely affected, and if really excessive then the heads may be desensitized and the box stained because the oil migrates into the head and the cardboard. Generally waxes with oil contents up to 7 % are satisfactory, although new waxes should always be checked before implementation into production to ensure that the appearance and performance of the product is unaffected. Synthetic (Fischer – Tropsch) waxes have also been used, as has stearin.

Chemicals. *Ammonium dihydrogen phosphate* (monoammonium phosphate, MAP) is used for splint impregnation as an afterglow inhibitor. It is an acid salt and the materials of construction of the impregnation unit and dryers must be compatible with the substance. The specification of the material is not critical, but it should be free from gross contamination. Insoluble matter can cause problems with spray impregnation systems.

Diammonium hydrogen phosphate (diammonium phosphate, DAP) is an alternative to ammonium dihydrogen phosphate. It is an alkaline salt and it causes much less corrosion of mild steel than ammonium dihydrogen phosphate. However, a greater concentration of the diammonium salt is needed to achieve the same degree of afterglow inhibition. During the drying process, considerable amounts of ammonia are released, and this can make the working environment unpleasant.

Potassium chlorate is the major ingredient of most match head formulations, and it accounts for up to 60 % of the dry head weight. It is the principle oxidizing agent in matches. The most important properties of potassium chlorate for matchmaking are the particle size and the sodium content. The particle size has a great influence on the burn rate of the head, and it must be closely controlled. The sodium content is important, since a high sodium content makes the match heads hygroscopic, and this can severely limit the sensitivity and performance in high-humidity environments. Potassium chlorate tends to form lumps on standing, and often a small amount (ca. 0.5 %) of anticaking agent (such as tricalcium phosphate) is added to the salt before it is packed.

Animal glue (technical gelatin) is the most widely used binder for match head composition. It is extracted by partial hydrolysis of animal hides or bones. During the hydrolysis the collagen present in the hides or bones is broken down and becomes soluble in hot water. In addition to its binding properties glue has important influences on other properties of match head composition:

- It controls the viscosity of the head composition.
- It is the main fuel present in the match heads.
- It is an air-entraining agent and foam stabilizer. It thus has an influence on the amount and stability of the air entrained in the composition.
- The animal glue is only partly burned during the burning of the head composition and the carbon remaining is a component of the ash left by the head. This carbon has an influence on the propensity of the ash to drop off.
- It is a good dispersing agent and it helps to keep all the head composition components well dispersed.

Solutions of animal glue gel on cooling. The gel point is around 28 °C for good-quality glue. The quality of glue is measured by the Bloom jelly strength test and by the viscosity. Generally glues with a Bloom value of around 300 g or more are most suitable for matchmaking. High-Bloom glues usually have a high gelling temperature and permit matchmaking at high ambient temperatures. The ash content should be low (< 2 %).

Animal glue is also frequently used as an *adhesive* for outer boxes formed on the bar machine. For this purpose, glues of very high Bloom strength (ca. 400 or more) are required.

Casein is sometimes used instead of animal glue in tropical countries. It provides better resistance to humidity, but the heads are more brittle.

Many head compositions contain a *starch*, a *starch derivative*, or a *cellulose derivative* (e.g., hydroxyethyl cellulose or ethyl hydroxyethyl cellulose). These contribute to the binding of the

head, but their main function is to modify the viscosity and dispersing properties of the mix.

Sulfur is a common component of match heads. Usually powdered sulfur is used.

Red phosphorus is added to some head compositions as well as being the main ingredient in the friction paint. For head composition and for printed friction paint the maximum particle size should be ca. 50 µm. No large particles should be present. For friction paint to be applied by roller, a coarser grade of red phosphorus can be used. Phosphorus tend to become acidic in storage. This tendency can be greatly reduced if the phosphorus particles are coated with a stabilizer, usually magnesium oxide. These stabilized grades are normally used for matchmaking.

Phosphorus sesquisulfide is the active ingredient in strike-anywhere matches. It is produced by the reaction of white phosphorus with sulfur. It smells strongly of hydrogen sulfide due to slow hydrolysis in normal damp air.

Potassium dichromate is frequently added to head composition as a sensitizer. It is usually dissolved in the water used for making the mix.

Manganese dioxide is usually finely ground natural pyrolusite. It is sometimes added to head composition and to friction paint. It may act as a sensitizer and/or pigment.

Various *inert materials* are added to match head composition to moderate the reaction. These are mainly siliceous materials like silica, feldspar, powdered glass, diatomaceous earth, and kaolin. Basic materials like magnesium carbonate, calcium carbonate or zinc oxide are usually added to prevent the composition from becoming acidic. In general, these materials play little part in the chemical reactions that take place, but they do have a profound influence on the stability of the ash and the speed of burning.

Colorants include various dyes such as eosine or rhodamine, or pigments such as red iron oxide (Fe_2O_3), black iron oxide (Fe_3O_4), or manganese dioxide. The prime function of these is to impart color, but some dyes do seem to affect the glue properties, and the pigments may affect the sensitivity and/or ash stability.

Various *antifoams* may be added to control foam during composition making. Usually these are polyoxyethylene derivatives. After the head composition has been prepared, a *foam stabilizer* may be added. Usually these are alkyl aryl sulfonate salts.

5. Head Composition

Match compositions are always made up as an aqueous slurry with a water content of about 40 % by weight of the dry solids in the mix.

For best sensitivity, burning characteristics, and ash stability the head composition, safety or strike-anywhere, must contain about 25 vol % of entrained air dispersed as very small bubbles in the mix. This corresponds to a mixture density of about 1.25 g/cm^3.

Safety Matches. A typical head composition for safety matches contains the following ingredients:

Potassium chlorate	50 – 60
Glue	9 – 12 %
Starch or cellulose derivative	0 – 3 %
Inert materials	25 – 35 %
Sulfur	0 – 5 %
Red phosphorus	0 – 1 %
Potassium dichromate	0 – 1 %
Colorants	0 – 3 %
Foam stabilizer	0.1 %

For the common brown safety match heads, iron oxide is usually added to the mix as a colorant. For colored matches, the inert materials are carefully selected to have a light color, and a dye is added. Good white matches are difficult to make. For best results only white inert materials are used, no potassium dichromate is added, and the composition is heavily loaded with titanium dioxide pigment.

Strike-Anywhere Matches. A typical strike-anywhere head composition contains the following ingredients:

Potassium chlorate	35 – 45 %
Glue	10 – 20 %
Phosphorus sesquisulfide	5 – 10 %
Inert materials	35 – 45 %
Colorants	0 – 1 %
Antifoam	trace
Foam stabilizer	0.1 %

6. Friction Surfaces

For safety matches the friction paint is an important functional part of the match box. In principle, strike-anywhere matches can be struck on any rough surface. In practice they are generally struck on the rough surface provided on the box.

Safety Matches. Friction paint for printed friction:

Red phosphorus	60 – 80 %
Poly(vinyl/acetate)	20 – 40 % (as dry solids)
Antifoam	1 – 2 %

The water content of a typical paint for a Chambon printing machine is ca. 60 % by weight of the dry solid. No fillers are added, as these rapidly wear the doctor blade that scrapes the gravure cylinder.

Friction paint for roller applicator:

Red phosphorus	50 – 70 %
Poly(vinyl/acetate)	10 – 20 % (as dry solids)
Inert fillers	10 – 40 %

The water content of friction paint for roller application varies widely and must be adjusted to suit the applicator conditions.

Strike-Anywhere Matches. Friction surfaces for strike anywhere matches can be produced in many ways:

- A strip of abrasive paper can be stuck to the side of the box as the box is formed.
- The boxes on the bar machine are coated along one edge with an aqueous binder (usually urea – formaldehyde resin) as they are formed and then sprayed with sand that adheres to the wet binder. The boxes leave the bar and the frictioned surface is heated to cure the resin before the box is fed to the continuous machine.
- The boxes on the bar machine are coated along one edge with hot-melt adhesive as they are formed on the bar machine. The coated boxes are allowed to cool to solidify the hot melt adhesive strip and then fed to the continuous machine. After filling, the boxes are sprayed with hot sand that then adheres to the hot-melt adhesive. The advantages of this latter method are that the sand is applied after the continuous machine, so that the abrasive wear on the machine parts is reduced and the line can be made more compact.
- The boxes can be coated with a wet, gritty slurry by means of a roller applicator. Such slurries generally contain poly(vinyl acetate) binders, siliceous grit, and a thickening agent. UV-curing resins have also been used as binders.

7. The Chemistry of Matches

The chemistry of matches is not well understood, even though the product is 150 years old. The reactions taking place can be conveniently divided into three parts: the initiation phase, the head-burning phase and the splint-burning phase.

The Initiation Phase. The initiation phase begins when the match head is placed in contact with the friction. This puts the potassium chlorate in the head into close proximity with the red phosphorus in the friction surface. Any movement of the head causes the potassium chlorate and phosphorus to mix together. At the same time the interface between the head and the friction is heated by the work done in rubbing them together. The reaction of red phosphorus with potassium chlorate has a very low activation energy, and it can be postulated that this is the first reaction to take place:

$$5\,KClO_3 + 6\,P \rightarrow 5\,KCl + 3\,P_2O_5$$

This reaction is strongly exothermic and quickly causes further heating of the interfacial area until it is hot enough to start the second phase of the reaction.

The amount of air entrained in the mix has a considerable influence on the force required to strike matches. For normal head composition, this is about 25 vol %. In the past it was argued that this air somehow takes part in the initiation phase. However, this seems improbable, since 1 cm^3 of head composition contains 0.25 cm^3 of air, 0.05 cm^3 of which is oxygen. This is equivalent to 0.07 mg of oxygen, which compares with the 200 mg of oxygen contained in the potassium chlorate present in the same volume of head composition. It seems unlikely that such a small amount of oxygen can have a serious influence on the initiation phase. Furthermore it has been demonstrated that carbon dioxide bubbles are just as effective as air bubbles in sensitizing matches. A more likely explanation of the sensitizing action of the entrained air is that it reduces the thermal conductivity of the head composition. Hence, less of the frictional heat leaks into the head, and a higher temperature is achieved at the interface with the friction surface for the same mechanical work input. Conversely, the same interfacial temperature can be achieved with less input of mechanical work, i.e., the matches are more sensitive.

Another illustration of the importance of thermal conductivity of the system is provided by the fact that it is possible to strike safety matches, albeit with great difficulty, on a smooth surface like polished glass. It is in fact even easier to strike matches on a smooth polished silica surface that has a lower conductivity than glass, but it is impossible to strike matches on smooth polished stainless steel surface, which has a much higher conductivity than glass.

The Head-Burning Stage. The active ingredients in the head are potassium chlorate, animal glue and either sulfur or red phosphorus. These latter two ingredients are present in much smaller amounts than the animal glue. Furthermore, it is known that match heads without sulfur or phosphorus can be readily struck, so it can be argued that sulfur and phosphorus are not essential to the main reaction. Thus the principle reaction taking place is that between potassium chlorate and animal glue.

This strongly exothermic reaction can be expressed as follows:

$$\mathrm{C}_v\mathrm{H}_w\mathrm{O}_x\mathrm{N}_y\mathrm{S}_z + (2/3v+1/6w-x/3+2/3z)\,\mathrm{KClO_3} \rightarrow$$

$$(2/3v+1/6w-x/3+2/3z)\,\mathrm{KCl}$$

$$+v\mathrm{CO_2}+w/2\mathrm{H_2O}+y/2\mathrm{N_2}+z\,\mathrm{SO_2}$$

Even without sulfur or phosphorus it is easy to initiate the above reaction, but some times it spreads as a dull red glow throughout the material without the appearance of a flame. It can be argued that the reaction temperature is insufficient to ignite the gases formed by pyrolysis of the material ahead of the glowing zone. The inclusion of sulfur or phosphorus in the composition gives rise to some phosphorus or sulfur vapor in the pyrolysis products, and that these lower the spontaneous ignition temperature of the pyrolysis gases in the surrounding atmospheric oxygen, causing instantaneous ignition with the appearance of flame.

The underlying wood impregnated with wax also plays some role in the head reaction, since the pyrolysis products from the wood and wax also mix with the pyrolysis products formed ahead of the reaction zone.

The inert materials, mainly silicates, present in the mix serve to moderate the reaction of the active chemicals. These slow down the reaction by absorbing heat. The residue left after combustion of the head must also fuse and flux together to leave a stable ash that does not fall from the splint. The silicates present provide this fluxing action in combination with the molten potassium chloride that remains from the decomposition of the potassium chlorate. It is the correct choice of the inert materials that renders a match safe to use (i.e., smooth burning, nonexplosive, and not spitting or dropping hot ash).

The other important ingredient in the match head is a fine dispersion of small bubbles. The effect of these on sensitivity was noted above. However, the presence of these dispersed voids also controls the rate of burning of the head. Generally the greater the volume fraction of bubbles, the faster the head burns. The porosity also has an influence on the stability of the ash residue.

The reactions taking place are complex and involve interactions between the solid phases present at the start of the reaction, between the volatile pyrolysis products and the surrounding air, as well as various reactions involving the intermediate products. Often quite small changes to the formulation, such as the removal of potassium dichromate from the composition, can have a profound effect on the sensitivity and burning properties of the match head, as well as completely changing the structure and stability of the residual ash left behind.

The Splint-Burning Phase. By the time the head finishes burning the underlying paraffin-wax-impregnated wood has been heated quite strongly and is producing copious amounts of paraffin vapor, which is already alight and burning in the surrounding air. This in turn heats the splint ahead of the flame front, generating more vapor, and thus the flame transfers from the head to the splint and then travels along the splint consuming the paraffin as it progresses.

The main reaction taking place is the combustion of the paraffin vapor:

$$\mathrm{C}_n\mathrm{H}_{2n+2} + (3/2n+1/2)\,\mathrm{O_2} \rightarrow n\,\mathrm{CO_2} + (n+1)\,\mathrm{H_2O}$$

The wood takes relatively little part in this reaction. Indeed, on splints without paraffin wax the flame always extinguishes once the head has finished burning.

Wood consists mainly of the carbohydrate cellulose together with some lignin. Dry wood is readily combustible since the pyrolysis products are inflammable. Typically dry, unimpregnated wood yields combustible pyrolysis prod-

ucts on heating to 400 °C. The residue is charcoal that continues to glow in air with short blue flames. However, the splints for matches are impregnated with ammonium phosphate to prevent afterglow. This modifies the pyrolysis reaction considerably. On heating, the ammonium phosphate is pyrolysed to phosphoric acid. The phosphoric acid catalyzes the dehydration of the carbohydrate to carbon and steam [7]:

$$(C_6H_{10}O_5)_n \rightarrow 6n\,C + 5n\,H_2O$$

The consequence is that few combustible pyrolysis products are formed on wood treated with ammonium phosphate. The steam evolved merely serves as a diluent for the paraffin vapor. This dilution of the wax vapor with steam results in larger flame heights. The char left on pyrolysis of wood treated with ammonium phosphate is much more inert than the char from unimpregnated wood and consequently it does not readily react with the oxygen in the surrounding air. Thus, any tendency for the wood to continue glowing (afterglow) once the flame is extinguished is suppressed.

Strike-Anywhere Matches. The chemistry of strike-anywhere matches is basically similar to that of safety matches. The only difference is that the active reducing agent, phosphorus sesquisulfide, and the potassium chlorate are both in the match head. During striking, these two ingredients are ground together and heated by friction. The reaction has a low activation energy and gives phosphorus and sulfur oxides:

$$16\,KClO_3 + 3\,P_4S_3 \rightarrow 16\,KCl + 6\,P_2O_5 + 9\,SO_2$$

The relation between the oxygen supply present in the oxidizer and the oxygen demand of the reducing agents present in the match head can vary widely. Most safety matches contain a slight excess of oxygen, whereas most strike-anywhere matches are highly oxygen deficient. However, there are many examples of perfectly acceptable oxygen-deficient safety matches that have been produced.

Other Match-Head Systems. Relatively few combinations of active chemicals produce satisfactory friction matches. The combinations of potassium chlorate with red phosphorus, antimony sulfide or phosphorus sesquisulfide are the only systems that are currently used for making matches. A few other systems have been investigated, such as those based on potassium chlorate and calcium hypophosphite or lead thiosulfate, and shown to have some friction-sensitive properties. However, these are not as good as the more traditional systems, and they are not cost effective. Attempts to replace the potassium chlorate with other oxidants such as potassium nitrate or potassium perchlorate result in matches that are difficult or impossible to strike.

Matches have been made, in the past, from combinations of white phosphorus with other oxidants such as potassium nitrate, lead dioxide or even calcium carbonate that also function well. However, these would be unacceptable today because of the toxicity of white phosphorus.

8. Match Testing

Match tests are time-consuming and subjective. Typical tests that are carried out on matches follow:

- Count: the content of the box is counted.
- Appearance: the quality of the print, the friction surface, the splints, and the match heads are judged visually, often by comparison with reference samples. Match heads must have the correct shape and be without cracks or other damage, since these could lead to detachment of pieces of head material during striking.
- Splint strength can be assessed by using a pendulum gauge to measure the energy absorbed in breaking the splint or by applying a static load to the center of a splint supported at the two ends.
- Head adhesion can be measured by applying a load to the head to try to pull it off.
- Head impact strength is assessed by measuring the load, applied with a small drop hammer, that cracks the head.
- Strike test: matches are struck manually in a defined way using the minimum force to achieve ignition. Evidence of explosion or fragmentation, of dropping ash, of breaking splints, of afterglow; ignition of the friction surface; flame height; and flame duration are recorded.
- Durability of the friction surface: the usual requirement is that all the matches can be struck on half the available friction surface.

– Sensitivity. The minimum force required to ignite the match can be assessed with a strain-gauge force sensor and a peak measuring meter.

9. Match Standards

There are many national standards for matches. The only international standard is the CEN Standard EN 1783 (Matches — Performance Requirements, Safety and Classification), published by CEN in 1997.

10. Quality Assurance

Matches pose some unique challenges to a quality-control manager:

– Matches are produced at very high production speeds (ca. 2.4×10^6/h).
– Match testing is slow and laborious. In the time it takes to make a complete test to confirm conformance to the performance tests of EN 1783, a continuous machine will have produced about 10×10^6 matches.
– The costs of testing are very much higher than the cost of production. In the time it takes for one person to test the strike performance of 125 matches, another can produce 330 000 matches.
– The important tests are destructive, so high rates of testing would be uneconomic even if they were feasible.
– Production is not serial, which would allow automatic tests on every match, but parallel, with many matches produced at each stroke of the machine.
– Consumer expectation of defects is very low (complaints are measured in parts per billion), but sampling methods for testing can only detect defects at the level of 0.1 %.
– Matches are made from natural materials and derived products (wood and cardboard) that do not always behave in a predictable way.
– Some match defects lead to potentially hazardous, even life-threatening, situations. For example, detached burning heads or accidental ignition of matches can cause fires.

In practice it is impossible to measure the product quality to provide total quality assurance, since sampling methods can only detect defects at the 0.1 % level, whereas acceptable consumer quality requires some defects, such as head fragmentation during striking, to be at very much lower levels. The only way is to control the process. If the process has been stable and the resultant product quality has been acceptable in the market then there is no reason why, if the process is controlled to have the same process conditions, that it should not continue to produce satisfactory product. This means setting tight limits on all controllable variables and raw materials and constantly checking them to ensure that the process is within the control range. All changes to the process or raw materials must be closely monitored to ensure that there are no adverse effects on quality or consumer satisfaction. Product testing is used to supplement process controls to ensure that matches meet the performance and safety requirements of internal or external standards.

11. Safety and Handling

Match manufacturing is potentially a hazardous process since it involves handling highly reactive materials such as potassium chlorate and red phosphorus. Good practice demands that the highly reactive oxidants and reductants are kept in separate storage areas and that they only come together when they are processed to make the match-head composition. This mixing is always carried out in the presence of water and each component is well wetted and dispersed before adding the next. Once the head composition is prepared it is relatively safe to handle provided it is never allowed to dry out (except on the splints of course) and that good housekeeping is practiced.

On the continuous machine the head composition is applied to the splints and allowed to dry. The greatest potential for fire occurs when the matches are punched out into the match boxes. The matches at this stage are almost dry. A fire at the punch out will spread rapidly to all the matches on the continuous machine. Such incidents do occasionally happen in production (usually if the machine has stopped for some time and the matches have become very dry before boxing). However, continuous machines

are robustly built to withstand such fires and the operators are trained in fire-fighting techniques. Fires on match machines can be extinguished with carbon dioxide or water. Automatically triggered fire extinguishers are frequently fitted at the machine punch out.

Once the matches are boxed and packed into cases they represent little danger. Even if a case falls from a considerable height, and some boxes ignite inside the case, the fire usually self extinguishes because there is insufficient oxygen in the packaging to sustain burning. However, cases of matches should be stored in a cool, dry place and separated from highly flammable materials. The ignition temperature of matches is greater than 125 °C so there is no danger of self-ignition in normal storage environments.

In normal use there are relatively few serious accidents involving matches. It should be noted that many accidents involving matches result from children playing with fire and due care should be taken to ensure that matches are stored out of their reach

12. References

General References

1. C. A. Finch, S. Ramachandran: *Matchmaking: Science, Technology and Manufacture*, Ellis Horwood, Chichester 1983.
2. H. Ellern in *Kirk-Othmer*, 4th ed., vol. **13**, 160 – 166.
3. H. Ellern: *Military and Civilian Pyrotechnics*, Chemical Publishing Co., New York 1968, pp. 64 – 83.

Specific References

4. S. R. Johansson, *Ullmann*, 4th ed., **24**, pp. 801 – 810.
5. *Proc. Pyroteknikdagen* (1983) no. 7, 1 – 104.
6. *Proc. Int. Pyrotechnics Seminar* **11** (1986) 373 – 393.
7. G. Drake in *Kirk-Othmer*, 4th ed., vol. **9**, 302 – 303.

Melamine and Guanamines

George M. Crews, Melamine Chemicals, Inc., Donaldsonville, Louisiana, United States (Chap. 1, 2, 3, 4, 5, 6, 7, 8 and 9)

Willi Ripperger, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8 and 9)

Dietrich Burkhard Kersebohm, BASF Aktiengesellschaft, Ludwigshafen, Germany (Chap. 10)

Thomas Güthner, Degussa AG, Trostberg, Germany (Chap. 11)

Bernd Mertschenk, Degussa AG, Trostberg, Germany (Chap. 11)

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 2
3.1. Thermal Behavior 2
3.2. Hydrolysis 3
3.3. Salt Formation 3
3.4. Reaction with Aldehydes 3
4. Production 4
4.1. Low-Pressure Processes 4
4.1.1. BASF Process 5
4.1.2. Chemie Linz Process 5
4.1.3. Stamicarbon Process 6
4.2. High-Pressure Processes 7
4.2.1. Melamine Chemicals Process 8
4.2.2. Montedison (Ausind) Process 8
4.2.3. Nissan Process 9
5. Quality Specifications 10
6. Chemical Analysis 10
7. Storage and Transportation 10
8. Uses 11
9. Economic Aspects 11
10. Toxicology 11
11. Guanamines 12
11.1. Production 13
11.2. Uses 14
11.3. Toxicology 15
12. References 15

Melamines (2,4,6-triamino-1,3,5-triazines) are produced from urea. They are used in the fabrication of melamine – formaldehyde resins for laminating and adhesive applications. Melanine is used as cross-linker in heat-cured and high-solids paint systems.Guanamines (2,4-diamino-1,3,5-triazines) are produced from dicyandiamide and the corresponding nitrile by base-catalyzed cyclocondensation. They are used in similar polymer applications as for melamine. They give a lower cross-link density and higher flexibility because only two amino functionalities are present.

1. Introduction

Melamine was first prepared and described in 1834 by Liebig, who obtained it from fusion of potassium thiocyanate with ammonium chloride. In 1885, A. W. von Hoffmann published its molecular structure. Melamine [*108-78-1*] (2,4,6-triamino-1,3,5-triazine), $C_3N_6H_6$, M_r 126.13, exists mainly in the amino form:

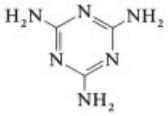

Not until 100 years later did melamine find industrial application in the production of melamine – formaldehyde resins (→ Amino Resins). The first commercial plants came on stream in the late 1930s. Since that time melamine has become an increasingly important chemical commodity. In 1970, world capacity was estimated at 200 000 t. Production in 1994 is 610 000 t/a. Most of the melamine produced is still used in the fabrication of melamine – formaldehyde resins.

Until about 1960, melamine was prepared exclusively from dicyandiamide [*461-58-5*] (→ Cyanamides, Chap. 3). This conversion was car-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_171.pub2

ried out in autoclaves at 10 MPa and 400 °C in the presence of ammonia, according to the equation

$$3\ H_2NC\,(NH)\,NHCN \rightarrow 2\ C_3N_6H_6$$

In the early 1940s, MACKAY discovered that melamine could also be synthesized from urea [*57-13-6*] at 400 °C with or without catalyst [6]. Today, melamine is produced industrially almost exclusively from urea. Most processes using dicyandiamide as raw material were discontinued or replaced at the end of the 1960s.

2. Physical Properties

Melamine is manufactured and sold as fine, white, powdered crystals. The most important physical data for melamine are summarized in the following list [7, 8]:

mp	350 °C (subl.)
Density ϱ	1.573 g/cm^3
Dissociation constants	
K_{b1} (25 °C)	1.1×10^{-9}
K_{b2} (25 °C)	1.0×10^{-14}
Heat of formation ΔH_f^0 (25 °C)	− 71.72 kJ/mol
Heat of combustion (25 °C)	− 1967 kJ/mol
Heat of sublimation (25 °C)	− 121 kJ/mol
Molar heat capacity (25 °C)	155 J K^{-1} mol^{-1}
Specific heat capacity c_p, J kg^{-1} K^{-1}	
at 273 – 353 K	1470
at 300 – 450 K	1630
at 300 – 550 K	1720
Entropy (25 °C)	149 J K^{-1} mol^{-1}
Entropy of formation ΔS_f^0 (25 °C)	−835 J K^{-1} mol^{-1}
Free energy of formation ΔG_f^0 (25 °C)	177 kJ/mol
Solubility (30 °C), g/100 mL, in	
Ethanol	0.06
Acetone	0.03
Dimethylformamide	0.01
Ethyl cellosolve	1.12
Water	0.5

The temperature dependence of the vapor pressure (in 10^5 Pa) in the range 417 – 615 K is described by the following equation [9]:

$$\log p = 9.7334 - 6484.9/T$$

Melamine solubility in water (in grams per 100 g of H_2O) over the range 20 – 100 °C coincides closely with the relationship [10]:

$$\log L = 5.101 - 1642/T$$

Reasonably reliable values can be obtained by means of this equation down to 0 °C.

Crystal Data. Melamine forms monoclinic crystals, space group $P2_1/a$ with a = 1.0537, b = 0.7477, c = 0.7275 nm, β = 112°9′, and Z = 4 [11].

3. Chemical Properties

The chemical properties of melamine are summarized in detail in [1] and [2]. The *s*-triazine ring is very stable and cleaves only under drastic conditions (e.g., heating above 600 °C or fusion with alkali compounds).

By X-ray diffraction studies, crystalline melamine has been shown to exist only in the symmetrical triamino structure; the same is true for the vapor phase and for both neutral and alkaline solutions. Although reactions are in some cases observed at the ring nitrogen atoms (the products being substituted isomelamines), the most important commercial reactions involve only the $-NH_2$ groups, which behave chemically as amido rather than amino functions.

3.1. Thermal Behavior

When melamine is heated above 300 °C in the absence of ammonia or at low ammonia partial pressure, deammoniation and condensation lead to compounds with higher molecular mass. Degradation starts with the release of ammonia and the formation of melem [*1502-47-2*] (2,5,8-triamino-1,3,4,6,7,9,9b-heptaazaphenalene):

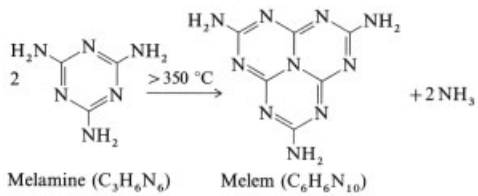

Further heating to ca. 600 °C yields more ammonia and melone [*32518-77-7*] [12]:

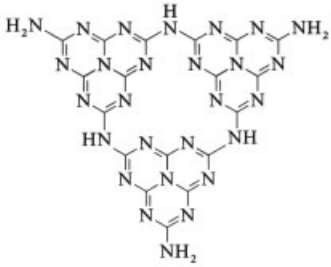

1

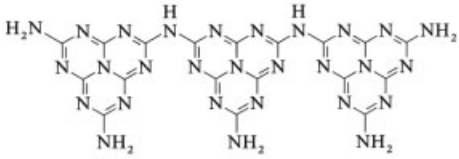

2

Various sources disagree on hydrogen analysis and other analytical data for melone. The material may represent a mixture of substances such as **1** and **2**.

Melam [(*N*-4,6-diamino-1,3,5-triazin-2-yl)-1,3,5-triazine-2,4,6-triamine], [*3576-88-3*] seems not to be an intermediate in the thermal degradation of melamine although the evidence is not entirely clear [12 – 14]. This substance can be prepared by heating melamine salts below 315 °C, and it arises as a byproduct of melamine synthesis.

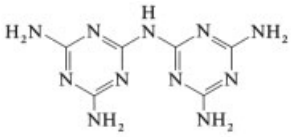

Melam ($C_6H_9N_{11}$)

The three deammoniation products—melam, melem and melone—are formed reversibly; addition of ammonia at high pressure and temperature regenerates melamine. Indeed, processes for melamine production invariably rely upon excess ammonia to suppress formation of these byproducts. If melamine is heated to 600 °C or higher it is partially cracked, leading to cyanamide along with other products. Hydrogen cyanide may also appear in the crack products, especially in the absence of oxygen [15].

3.2. Hydrolysis

Melamine is hydrolyzed by mineral acid or inorganic alkali. Hydrolysis proceeds stepwise, with loss of one, two, or all three amino groups:

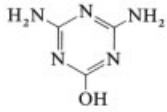

Ammeline

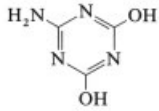

Ammelide

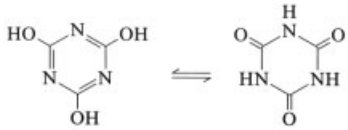

Cyanuric acid

The product spectrum varies with temperature, pH, and concentration; the end product is cyanuric acid [*108-80-5*]. Even small amounts of the oxotriazines (especially cyanuric acid and ammelide) markedly effect the condensation of melamine with formaldehyde by increasing the rate of condensation [16].

3.3. Salt Formation

Melamine is a weak base, forming well-defined salts with both organic and inorganic acids. The melamine ion is assigned the following structure:

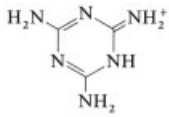

The water solubility of organic and inorganic salts of melamine is no higher than that of free melamine (see Table 1). Melamine cyanurate, melamine picrate, and melamine perchlorate are very insoluble in water, and are useful in the quantitative determination of melamine.

3.4. Reaction with Aldehydes

Melamine reacts with aliphatic and aromatic aldehydes to give a variety of products. Most

Table 1. Industrially important melamine salts

Molecular formula	Solubility (20 °C), g per 100 g H_2O	M_r	*mp*, °C
$C_3N_6H_6 \cdot H_3PO_4$	0.43	224.12	223
$C_3N_6H_6 \cdot H_2SO_4$	0.19	350.33	380 *
$C_3N_6H_6 \cdot HNO_3$	0.68	189.14	298 **
$C_3N_6H_6 \cdot C_{17}H_{35}COOH$	0.17	410.42	154
$C_3N_6H_6 \cdot HCOOH$	1.56	172.16	250 (decomp.)

* Sublimes at 316 °C.
** Sublimes at 186 °C.

important is the resinous material obtained from the reaction of formaldehyde with melamine:

$$Me(NH_2)_3 + 6\,CH_2O \rightarrow Me[N(CH_2OH)_2]_3$$

where Me represents that part of the melamine molecule that is not involved in the reaction.

All hydrogen atoms on the melamine molecule can be replaced by methylol groups, and products ranging from the monomethylol to the hexamethylol derivatives have been observed. The methylolmelamines are sparingly soluble in most solvents and are very unstable due to further condensation or resinification, e.g.,

$$MeNHCH_2OH + H_2N{-}Me \rightarrow MeNHCH_2NHMe + H_2O$$

$$2\,MeNHCH_2OH \rightarrow MeNHCH_2OCH_2NHMe + H_2O$$

Melamine – formaldehyde condensation products are characterized by good heat resistance and superior water-resisting properties. They are used, usually in combination with urea – formaldehyde resins, as glues in the woodworking industry, as impregnating resins for decorative laminates, and as a binder in molding materials containing a filler (e.g., cellulose or sawdust). Methylolmelamines can be etherified by heating with alcohol in the presence of an acid catalyst. Industrially most important are the products formed with methanol, *n*-butanol, and isobutanol. They are used as curing agents for surface coatings and as auxiliaries in the paper and textile industries. For additional information see → Amino Resins [17].

4. Production

Melamine can be synthesized from urea at 390 – 410 °C:

$$6\,H_2N - CO - NH_2 \rightarrow C_3N_3(NH_2)_3 + 6\,NH_3 + 3\,CO_2$$

The overall reaction is endothermic, requiring 649 kJ per mole of melamine starting with molten urea at 135 °C.

The processes themselves may be subdivided into two categories:

1) noncatalytic, high-pressure ($\geq$ 8 MPa) processes, and
2) catalytic, low-pressure processes (ca. 1 MPa).

Each type includes three stages:

synthesis
melamine recovery and purification
off-gas treatment

4.1. Low-Pressure Processes

Typical low-pressure processes utilize a fluidized catalyst bed at pressures from atmospheric to ca. 1 MPa and temperatures of 390 – 410 °C. The fluidizing gas is either pure ammonia or the ammonia – carbon dioxide mixture formed during the course of the reaction. Catalysts include alumina and materials of the silica – alumina type. Melamine leaves the reactor in gaseous form together with the fluidizing gas; it is separated from ammonia and carbon dioxide by quenching the gas stream either with water (followed by crystallization) or with cold reaction gas (desublimation).

In the catalytic processes the first reaction step is decomposition of urea to isocyanic acid

and ammonia, after which the isocyanic acid is transformed into melamine:

$$6\,(NH_2)_2CO \longrightarrow 6\,HN{=}C{=}O + 6\,NH_3 \qquad \Delta H = 984\ \text{kJ/mol}$$

$$6\,HN{=}C{=}O \longrightarrow C_3N_3(NH_2)_3 + 3\,CO_2 \qquad \Delta H = -355\ \text{kJ/mol}$$

$$6\,(NH_2)_2CO \longrightarrow C_3N_3(NH_2)_3 + 6\,NH_3 + 3\,CO_2 \qquad \Delta H = 629\ \text{kJ/mol}$$

The overall reaction mechanism is not yet fully understood, but isocyanic acid from the decomposition of urea is believed to be catalytically disproportionated into carbon dioxide and cyanamide or carbodiimide, which then trimerizes to melamine [18 – 20]:

$$6\,HOCN \longrightarrow 3\,[\underset{\text{Carbodiimide}}{HNCNH} \rightleftharpoons \underset{\text{Cyanamide}}{H_2NCN}] + 3\,CO_2$$

$$3\,HNCNH \longrightarrow C_3N_3(NH_2)_3$$

The melamine yield is ca. 90 – 95 % based on urea. Byproducts include melam, melem, and melone, as well as oxotriazines such as ammeline, ammelide, and cyanuric acid. Ureidotriazine is also observed as a product of reaction between melamine and isocyanic acid.

Some byproducts are formed in the reactor during synthesis; others are not generated until the melamine recovery section, where deammoniation or hydrolysis occurs [20 – 29].

Worldwide, three low-pressure processes are in commercial operation: the BASF process, the Chemie Linz process, and the Stamicarbon process.

4.1.1. BASF Process

The BASF process (see Fig. 1) is a *one-stage,* low-pressure, catalytic vapor-phase process. Molten urea is fed to the fluidized catalytic bed reactor (a) at 395 – 400 °C and atmospheric pressure. Alumina is used as a catalyst, and fluidization is accomplished with an NH_3– CO_2 mixture (the process off-gas).

The reactor temperature is held at ca. 395 °C by molten salt circulated through internal heating coils (b). The fluidizing gas is also preheated to 400 °C. To secure an ammonia-rich atmosphere in the reaction zone, make-up ammonia is added to both the fluidizing gas and the urea nozzles.

Gas leaving the reactor is a mixture of gaseous melamine, traces of melem, and unreacted urea (in the form of its decomposition products isocyanic acid and ammonia), as well as ammonia and carbon dioxide (part newly formed, part fluidizing gas). In addition, the gas mixture contains entrained catalyst fines; coarser catalyst particles are retained by cyclone separators inside the reactor.

The gas mixture leaving the reactor is cooled in the gas cooler (d) to a temperature at which only the byproduct melem crystallizes. Precipitated melem, in the form of a fine powder, is removed together with the entrained catalyst fines in adjacent gas filters (e).

The filtered gas mixture enters the top of the crystallizer (f) where it is blended countercurrently with recycled off-gas (140 °C). The temperature in the crystallizer is thereby reduced to 190 – 200 °C, and more than 98 % of the melamine crystallizes as fine crystals. Melamine is recovered from the gas in a cyclone (g), after which it is cooled and stored. It can be used without further treatment and has a minimum purity of 99.9 %.

The nearly melamine-free gas stream from the cyclone is fed to the urea washing tower (i) where it is scrubbed with molten urea (135 °C), which provides both cooling and washing. Clean gas leaving the urea scrubber (after passing through droplet separators) is partially recycled to the reactor as fluidizing gas and partially recycled to the crystallizer as quenching gas. The surplus is fed to an off-gas treatment unit.

A single-stage reactor has the advantage of converting the corrosive intermediate isocyanic acid immediately to melamine; also, the heat of this exothermic reaction is used directly for the endothermic decomposition of urea, the first step in melamine synthesis.

4.1.2. Chemie Linz Process

The Chemie Linz process (see Fig. 2) is a *two-stage* process. In the first step, molten urea is decomposed in a fluidized sand-bed reactor (b) to ammonia and isocyanic acid at ca. 350 °C and 0.35 MPa. Ammonia is used as the fluidizing gas. Heat required for the decomposition is supplied to the reactor by hot molten salt circulated through internal heating coils. The gas stream

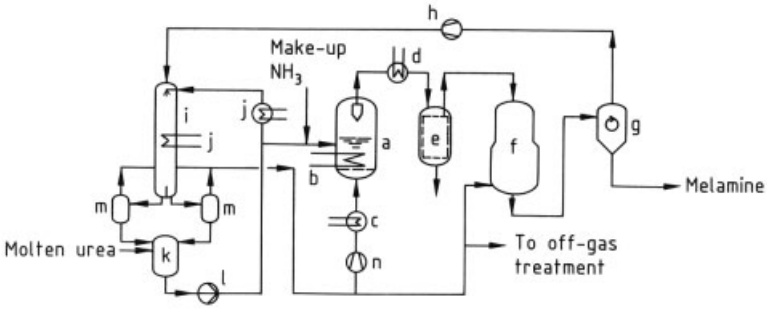

Figure 1. BASF process
a) Reactor; b) Heating coils; c) Fluidizing gas preheater; d) Gas cooler; e) Gas filter; f) Crystallizer; g) Cyclone; h) Blower; i) Urea washing tower; j) Heat exchanger; k) Urea tank; l) Pump; m) Droplet separator; n) Compressor

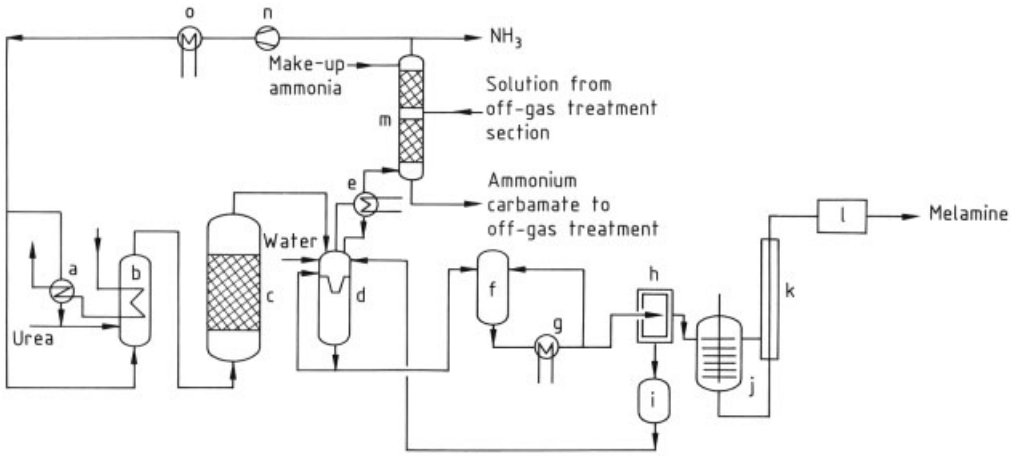

Figure 2. Chemie Linz process
a) Heat exchanger; b) Urea decomposer; c) Converter; d) Quencher; e) Heat exchanger; f) Suspension tank; g) Heat exchanger; h) Centrifuge; i) Mother-liquor vessel; j) Disk dryer; k) Elevator; l) Delumper; m) CO_2 absorption column; n) Compressor; o) Heat exchanger

is then fed to the fixed-bed catalytic reactor (c) where isocyanic acid is converted to melamine at ca. 450 °C and near-atmospheric pressure.

Melamine is recovered from the reaction gas by quenching with water and mother liquor from the centrifuges (h). The quencher (d) is specially designed to work quickly, thereby preventing significant hydrolysis of melamine to ammelide and ammeline. The melamine suspension from the quencher is cooled further to complete the melamine crystallization process. After being centrifuged, the crystals are dried, milled, and stored. A separate recrystallization step is not required.

Exhaust gas from the quencher is fed to an absorber (m) where carbon dioxide is removed as ammonium carbamate by washing with a lean carbamate solution from the off-gas treatment section. The wet ammonia gas is dried with make-up ammonia. Part of it is compressed and recycled to the urea decomposer, and part is exported. Remaining ammonia and carbon dioxide in the liquid effluent are then recovered in the off-gas treatment section.

4.1.3. Stamicarbon Process

Like the BASF process, the DSM Stamicarbon process (see Fig. 3) involves only a *single catalytic stage*. However, it differs from the former in that it is operated at 0.7 MPa, the fluidizing gas is pure ammonia, the catalyst is of the silica – alumina type, and melamine is recovered from the reactor outlet gas by water quench and recrystallization.

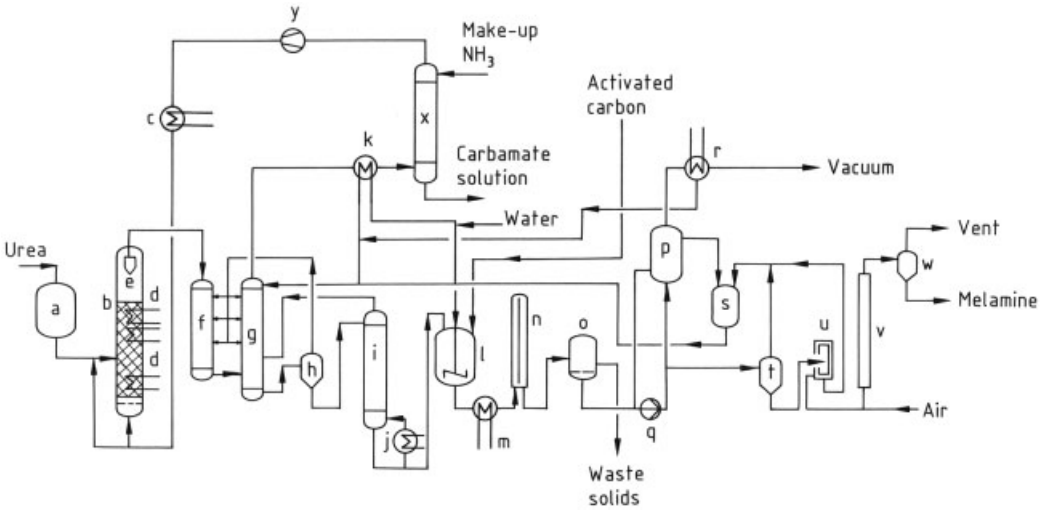

Figure 3. Stamicarbon process
a) Urea tank; b) Reactor; c) Preheater; d) Heating coils; e) Internal cyclone; f) Quench cooler; g) Scrubber; h) Hydrocyclone; i) Desorption column; j) Heat exchanger; k) Heat exchanger; l) Mixing vessel; m) Heat exchanger; n) Dissolving vessel; o) Precoat filter; p) Vacuum crystallizer; q) Pump; r) Heat exchanger; s) Mother-liquor vessel; t) Hydrocyclone; u) Centrifuge; v) Pneumatic dryer; w) Hydrocyclone; x) Absorption column; y) Compressor

Urea melt is fed into the lower part of the reactor (b). The silica – alumina catalyst is fluidized by preheated (150 °C) ammonia, which enters the reactor at two points: at the bottom of the reactor to fluidize the catalyst bed, and at the urea nozzles to atomize the urea feed. The reaction is maintained at 400 °C by circulating molten salt through heating coils within the catalyst bed.

The melamine-containing reaction mixture from the reactor is quenched first in a quench cooler (f) and then in a scrubber (g) with recycled mother liquor from the crystallization section. The resulting melamine suspension is concentrated to ca. 35 wt % melamine in a hydrocyclone (h), after which it is fed to a desorption column (i) where part of the ammonia and carbon dioxide dissolved in the suspension is stripped off and returned to the scrubber. The preceding steps are all carried out at reaction pressure; for the following stages, the pressure is reduced.

The suspension leaving the bottom of the desorber is diluted with recycled and preheated mother liquor and water. Activated carbon and filter aids may also be added. The melamine dissolves completely, although separate dissolving vessels (n) are necessary to allow sufficient time for dissolution. The resulting solution is filtered using precoat-type filters (o). Crystallization of melamine is carried out in a vacuum crystallizer (p), and crystals are separated from the mother liquor by hydrocyclone (t) and centrifuge (u). The crystals are dried in a pneumatic dryer and then conveyed to product bins.

Surplus ammonia must be recovered as fluidizing gas from the wet ammonia – carbon dioxide mixture leaving the desorption column and the scrubber. The hot gas mixture is partly condensed by heat-exchange (k) with the mother liquor from melamine dissolution. The condensate and uncondensed gas are then passed at 0.7 MPa to an absorption column (x). Liquid make-up ammonia is fed to the top of this column to condense any carbon dioxide remaining in the ammonia gas. The ammonia is then compressed and recycled as fluidizing and urea-atomization gas for the reactor.

4.2. High-Pressure Processes

High-pressure melamine synthesis systems differ from low-pressure processes by producing melamine in the liquid instead of the vapor phase. They have the advantage of providing high-pressure off-gas more suitable for use in the urea synthesis facility. Liquid phase operation also lends itself to smaller reaction vessels, but the highly corrosive nature of the system dic-

tates use of expensive, corrosion-resistant construction materials such as titanium.

High-pressure reactions occur without catalyst at > 7 MPa and > 370 °C. In general, molten urea is injected at high pressure into a molten melamine – urea mixture in the reactor, where it undergoes conversion to melamine. Sufficient residence time is provided in the reactor to ensure complete reaction, leading to melamine with a purity > 94 %. Heat is supplied to the reactor either by electric heater elements or by a molten salt heat-transfer system. Various types of off-gas separation and melamine purification follow.

High-pressure synthesis of melamine from urea proceeds via the intermediate cyanuric acid, which is subsequently converted to melamine under high pressure in an ammonia environment [30, 31]:

$$\underset{\text{Urea}}{3\,(NH_2)_2CO} \longrightarrow \underset{\text{Cyanic acid}}{3\,HOCN} + \underset{\text{Ammonia}}{3\,NH_3}$$

$$\underset{\text{Cyanic acid}}{3\,HOCN} \longrightarrow \underset{\text{Cyanuric acid}}{(NCOH)_3}$$

$$(NCOH)_3 + 3\,NH_3 \longrightarrow \underset{\text{Melamine}}{C_3N_3(NH_2)_3} + 3\,H_2O$$

$$3\,(NH_2)_2CO + 3\,H_2O \longrightarrow 6\,NH_3 + 3\,CO_2$$

$$6\,(NH_2)_2CO \longrightarrow C_3N_3(NH_2)_3 + 6\,NH_3 + 3\,CO_2$$

The net reaction is the same as in the low-pressure process.

4.2.1. Melamine Chemicals Process

Melamine Chemicals uses a continuous high-pressure *single-stage* process that produces melamine with a purity of ca. 96 – 99.5 %. Molten urea is converted into melamine in a liquid-phase reactor. The off-gases (ammonia and carbon dioxide) are separated in a gas-separating vessel. Liquid melamine is then quenched in a cooling unit, where liquid ammonia is used to solidify the crystals.

Process Description. Incoming urea is preheated by using it to scrub the reactor off-gas stream. This scrubber performs various functions, including (1) driving off any water that may be present in the urea feed, (2) preheating the molten urea, (3) removing melamine from the off-gases, and (4) recovering excess heat energy for subsequent use. The reactor is heated to ca. 370 – 425 °C with a heating-coil system and pressurized to about 11 – 15 MPa. Mixing is provided by heat convection and generation of gaseous reaction products.

Liquid melamine is separated from the off-gas in a gas separator, the product being collected at the bottom. The separator is held at about the same temperature and pressure as the reactor. The gaseous phase (ammonia and carbon dioxide, saturated with melamine vapor) is removed overhead to a urea scrubber. The melamine stream leaving this separator is then injected into the product cooling unit.

The product cooling unit employs liquid ammonia to both cool and solidify melamine. This is accomplished at a controlled temperature and pressure to minimize formation of such impurities as melam and melem. Product is removed from the pressurized cooling unit through a series of pressure-reducing hoppers. Depending on the required degree of purity, it may then be recrystallized [32].

The off-gas stream represents high-pressure (> 10 MPa) ammonia and carbon dioxide, which can be used directly as feed to the urea facility. Alternatively, this gas stream can be treated in a monoethanolamine (MEA) scrubber to recover ammonia and remove carbon dioxide.

4.2.2. Montedison (Ausind) Process

The Montedison process (see Fig. 4) operates at 370 °C and 7 MPa. The required temperature is maintained by a molten-salt heating system consisting of concentric bayonet-type tubes.

Molten urea at 150 °C is fed to the reactor (a) together with preheated ammonia. Average residence time of the mixture in the reactor is about 20 min. As the reaction mixture leaves the reactor, the pressure is lowered to 2.5 MPa, and the mixture is treated at 160 °C in a quencher (b) with an aqueous solution of ammonia and carbon dioxide to precipitate melamine. The water-saturated mixture of ammonia and carbon dioxide leaving the top of the quencher can be recycled to a plant for urea or fertilizer production.

The aqueous melamine slurry remains in the quencher at 160 °C for some time to decompose unconverted urea and such byproducts as biuret and triuret to ammonia and carbon dioxide. It is then fed to a steam stripper (c), where any

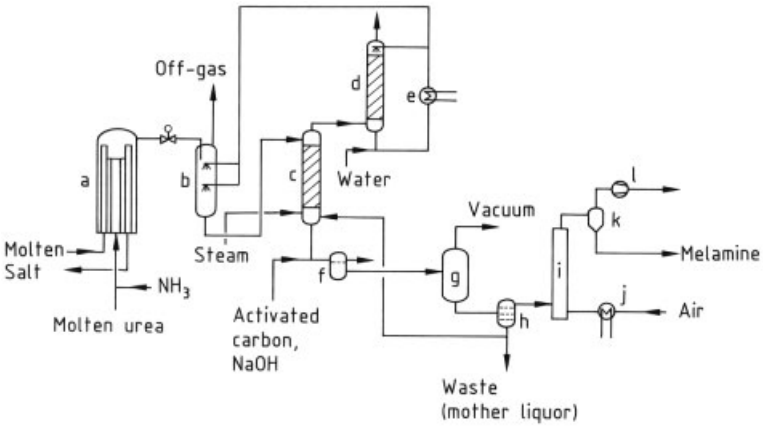

Figure 4. Montedison process
a) Reactor; b) Quencher; c) Stripper; d) Absorption column; e) Heat exchanger; f) Filter; g) Vacuum crystallizer; h) Filter; i) Pneumatic dryer; j) Heat exchanger; k) Cyclone; l) Blower

remaining ammonia and carbon dioxide are removed. Off-gas from the stripper is dissolved in water in an absorption column (d) and this solution is recycled to the quencher.

The ammonia- and carbon dioxide-free melamine slurry leaving the bottom of the stripper is diluted with mother liquor to dissolve melamine. Sodium hydroxide is also added, and the solution is then clarified with activated carbon (f). Melamine is crystallized from the clarified solution in a crystallizer (g) operated adiabatically under vacuum. Melamine crystals are separated from the mother liquor in a rotary filter (h), dried in a pneumatic conveyor – dryer (i), and stored.

4.2.3. Nissan Process

The Nissan melamine process (see Fig. 5) operates at 10 MPa and 400 °C. One characteristic feature is urea washing of the reactor off-gas. For this purpose, molten urea is also pressurized to 10 MPa and passed through a high-pressure washing tower (c) where it absorbs any melamine and unreacted urea present in off-gas leaving the reactor. The urea then flows into the reactor (a) by gravity. Ammonia is also fed to the reactor.

In a so-called level tank (b), effluent from the reactor is separated into gaseous and liquid phases. The gaseous phase passes through the previously described urea washing tower to an off-gas treatment facility. The liquid phase consists mainly of molten melamine. This melt is mixed with hot gaseous ammonia and fed to a "cushion" vessel (e) for aging (i.e., to allow by-products to be reconverted to melamine).

After aging, the melamine melt is quenched (f) under pressure with aqueous ammonia, in which it dissolves. The resulting 20 – 30 wt % melamine solution is retained in the quencher at 180 °C until any remaining impurities have decomposed.

Most of the added ammonia is next removed from the solution in an ammonia stripper (g) (operated at 1.5 MPa) and the solution is filtered. Recovered ammonia is recycled. Crystallization takes places in two crystallizers (k) operated in series. Mother liquor and melamine crystals are separated in centrifuges (l), after which the crystals are dried and crushed before storage.

Further treatment of the mother liquor starts with an ammonia stripper (n), in which oxoaminotriazines precipitate. The slurry from this ammonia recovery tower is therefore alkalinized before being fed to a third crystallizer (o) operating at reduced temperature and pressure. Additional melamine crystallizes here, and after separation from the liquid (p) it is returned to the second crystallizer.

Lowering the pH of the mother liquor causes oxoaminotriazines to precipitate; these are removed by decantation (q). The clear mother liquor is used to absorb ammonia leaving the crystallizers and is subsequently recycled to the

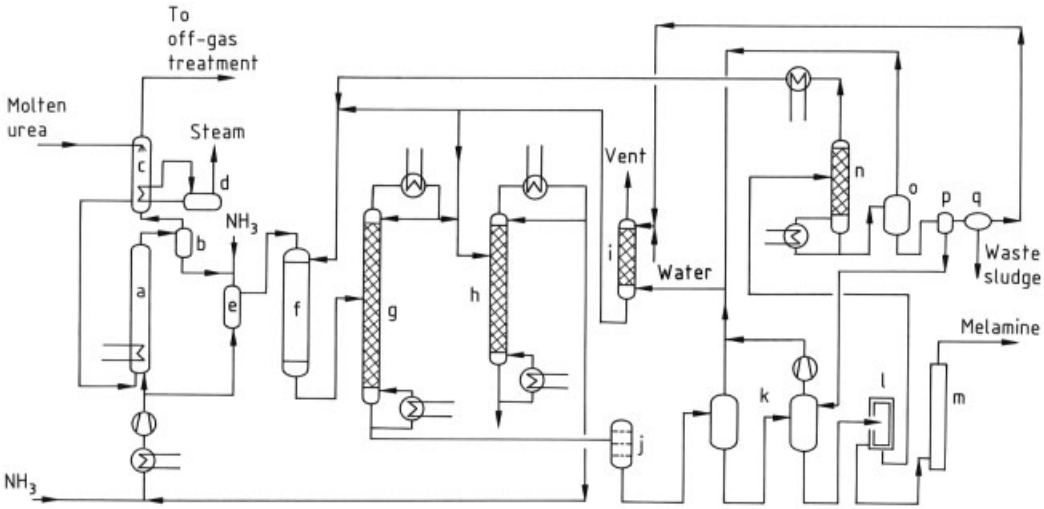

Figure 5. Nissan process
a) Reactor; b) Level tank; c) Off-gas washing tower; d) Steam drum; e) Cushion vessel; f) Quencher; g) NH_3-stripper; h) NH_3-distillation column; i) Absorber; j) Filter; k) Crystallizers; l) Centrifuge; m) Pneumatic dryer; n) Ammonia recovery tower; o) Crystallizer; p) Separator; q) Decanter

quencher together with ammonia released in the ammonia recovery tower.

5. Quality Specifications

Many consumers of melamine are satisfied with a purity of 99.9 %. Some, however, specify additional measurable criteria such as: content of inorganic ash, moisture, and ammeline-related compounds (alkali solubles); particle-size distribution; pH; resin reaction time; and resin color. Average particle sizes ranging from 15 to 100 µm are available for various applications. The rate of dissolution of melamine in formaldehyde solutions depends on particle size, an important parameter that is generally also reported by producers.

6. Chemical Analysis

Melamine is difficult to characterize by traditional chemical methods. Purity specifications are usually based on differences obtained after subtracting determined impurity levels for moisture, ash, and alkali solubles. Instrumental analysis is possible using liquid chromatography [33] and spectroscopic methods, although melamine tends to form complexes with the pH-control buffers used in liquid chromatography, leading to variability in the UV absorption observed at different pHs. Melamine can be precipitated for quantitative determination as the perchlorate or picrate salt (see Section 3.3).

7. Storage and Transportation

Melamine is stable when stored under normal warehouse conditions. Although not particularly hygroscopic, powdered melamine must still be protected from wetting because, like most powders, it will pack and lump over extended storage periods.

In the VDI guideline 2263 melamine is classified as having a burning index of 2, i.e., during a fire it ignites quickly but the flame is rapidly extinguished. Flammability tests performed in accordance with the EEC guideline 84/449 A 10 showed that a glowing platinum wire (>1000 °C) was not able to produce continous burning of melamine.

Shipping considerations are typical of those for other nonhazardous powders. Melamine is available in standard-weight paper bags and semibulk bags. Bulk shipment is by truck and railroad car.

Table 2. Melamine applications, in percent, by region

Application	Europe	United States	Japan
Laminates	47	35	6
Glue, adhesives	25	4	62
Molding compounds	9	9	16
Coatings	8	39	12
Paper, textiles	11	5	3
Other		8	1
Total	100	100	100

8. Uses

Most melamine is reacted with formaldehyde to produce resins for laminating and adhesive applications [16, 17, 34]. One of the major uses of melamine is in the upper sheet of laminated counter- and tabletops.

Another important use of melamine is as the amino cross-linker in heat-cured paint systems. In this case the methylated methylolmelamine is used, with varying molar ratios of melamine, formaldehyde, and alcohol for different paint system applications. High-solids paint systems for automotive applications also constitute a major market for melamine.

Other uses include the preparation of wet-strength resins for paper, water clarifying resins, ion-exchange resins, plastic molding compounds, adhesives, fire retardants in polyurethane foams, and intumescent paints. Important new applications are under development in the field of fire retardants for polymeric materials, especially polyurethane foams. Applications and uses of melamine differ widely among the main consumer countries or regions. Estimates are provided in Table 2.

9. Economic Aspects

About 18 melamine producers exist worldwide. Rated annual production capacity is ca. 550 000 t (data for 1990/1991; see Table 3). In the last decade the average annual growth rate for melamine consumption was ca. 2 %.

10. Toxicology

Acute Toxicity. From the standpoint of acute toxicity, melamine is not classified as a health risk. The oral LD_{50} for rats is > 5000 mg per kilogram of body weight [35]. Melamine applied to the skin and eyes of rabbits is a non-irritant [35]. Skin sensitization could not be provoked by patch tests on humans [36] or guinea pigs [37].

Table 3. Melamine production capacity worldwide

Country	Company	Capacity, t/a
Fed. Rep. Germany	BASF	42 000
Austria	Chemie Linz	55 000
Netherlands	DSM	90 000
Italy	Ausind	28 000
France	Norsolor	15 000
Western Europe		*230 000*
Poland	Polimex Cekop	28 000
Rumania	Romchim	12 000
Soviet Union	Techmashimport	10 000
Eastern Europe		*50 000*
United States	American Melamine Ind.	50 000
	Melamine Chem.	47 000
America		*97 000*
Japan	Mitsubishi Petrochemical	32 000
	Mitsui Toatsu Chemical	38 000
	Nissan Chemical	42 000
Korea	Korea Fertilizer	16 000
Taiwan	Taiwan Fertilizer	10 000
Saudi Arabia	Safco	20 000
China	Sichuan Chemical Works	12 000
India	Gujarat State Fertilizer	5 000
Middle and Far East		*175 000*
Total		**552 000**

Genetic Toxicity. Investigations into the potential genetic effects of melamine have included the following tests:

In vitro methods

1) Bacterial tests
2) Ames Test [38]
   Escherichia coli plate test [39]
3) Test with eukaryotes
4) Test with eukaryotes

Saccharomyces cerevisiae (gene conversion) [39]
Rat hepatocytes (DNA repair) [40]
Mouse lymphoma test (point mutation) [41]
CHO cells (chromosomal aberrations) [42]

In vivo methods

Drosophila melanogaster (sex-related lethal test) [43]
Mouse micronucleus test (oral administration of 1000 mg per kilogram of body weight) [44]

In none of these studies could melamine-induced mutagenicity or damage to genetic material be demonstrated.

Metabolism. Investigations into the metabolism and toxicokinetics of melamine showed that a single oral dose of 0.38 mg of ^{14}C-labeled melamine administered to a rat was eliminated unchanged in the urine to the extent of 90 % [45]. The plasma half-life was found to be 2.7 h, with highest concentrations in the bladder and kidneys.

Chronic Toxicity and Carcinogenicity. The target organ system for melamine toxicity after prolonged administration to mice and rats is the urinary tract. Both species were administered, in some cases extremely high doses in the diet (750 – 30 000 ppm; ca. 62 – 2490 mg per kilogram of body weight per day) over periods of 14 d, 90 d, and 2 a. All studies led to changes in the kidney (inflammation, calcified concretions in the proximal tubuli) and bladder (inflammation, ulceration, epithelial hyperplasia, bladder stones). Daily dosages up to 15 000 ppm for mice and 5000 ppm for rats administered in the diet over 14 d produced no changes. Female rats tolerated doses of 9000 ppm/d for 13 weeks without symptoms. The incidence of bladder carcinoma increased only for male rats receiving melamine in the diet over a two-year period (dosage: 4500 ppm). Of eight rats displaying tumors, seven also had bladder stones, suggesting that chronic mechanical irritation of the mucous membranes of the bladder may have been responsible for the tumors. Dosages that do not result in bladder stones are not expected to be carcinogenic. Female rats in this study displayed neither bladder stones nor carcinoma [39, 46].

An initiation-promotion study on female mouse skin failed to reveal any tumor-inducing effect [47]. A single dermal application of 1 μmol of melamine followed by a 31-week treatment (twice a week) with 10 nmol of a promoter (12-*O*-tetradecanoylphorbol-13-acetate) resulted in no increased incidence of tumor-bearing animals.

Reproduction Toxicology. A study reported by Thiersch on the reproduction toxicology of melamine was published in 1957. Intraperitoneal administration of a dose of 70 mg/kg to pregnant rats on days 4 and 5, 7 and 8, or 11 and 12 of pregnancy had no influence on the maternal or fetal development, nor was there any teratogenic effect [48].

11. Guanamines

In the course of heating guanidine acetate, Nencki (1874) obtained a new compound to which he assigned the name "guanamine"; later he changed it to acetoguanamine after the discovery of other homologues. Weith recognized in 1876 that this substance was in fact 2,4-diamino-6-methyl-1,3,5-triazine. The term guanamine has since been applied generally to 2,4-diamino-1,3,5-triazines substituted in the 6-position with alkyl, aryl, or alkaryl residues. Such substances are named on the basis of the carboxylic acid that contains one more carbon atom than is present in the substituent on the triazine ring:

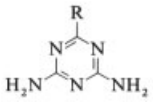

R = CH_3: acetoguanamine
R = C_3H_7: butyroguanamine
R = C_9H_{19}: caprinoguanamine
R = C_6H_5: benzoguanamine

Compounds of this type that have major commercial significance are benzoguanamine [*91-76-9*], $C_9H_9N_5$, M_r 187.20, and acetoguanamine [*542-02-9*], $C_4H_7N_5$, M_r 125.14 [49].

The physical properties of aceto-, benzo-, and caprinoguanamine [*5921-65-3*], $C_{12}H_{23}N_5$, M_r 237.35, are summarized in Table 4. Figure 6 illustrates the most important reactions of aceto- and benzoguanamine [50 – 53]. For information regarding other guanamines and their reactions, see [54 – 56].

Table 4. Physical properties of the most important guanamines

	Acetoguanamine	Benzoguanamine	Caprinoguanamine
mp, °C	277	228	105 – 120[a]
Solubility (at 20 °C), g/L, in			
Water	11.2	0.3	insoluble
Acetone	1.04	18.0	25.2
Benzene	0.07	0.3	11.2
Dimethylformamide	0.88	120.0	67.0

[a] Liquid crystalline intermediate phase

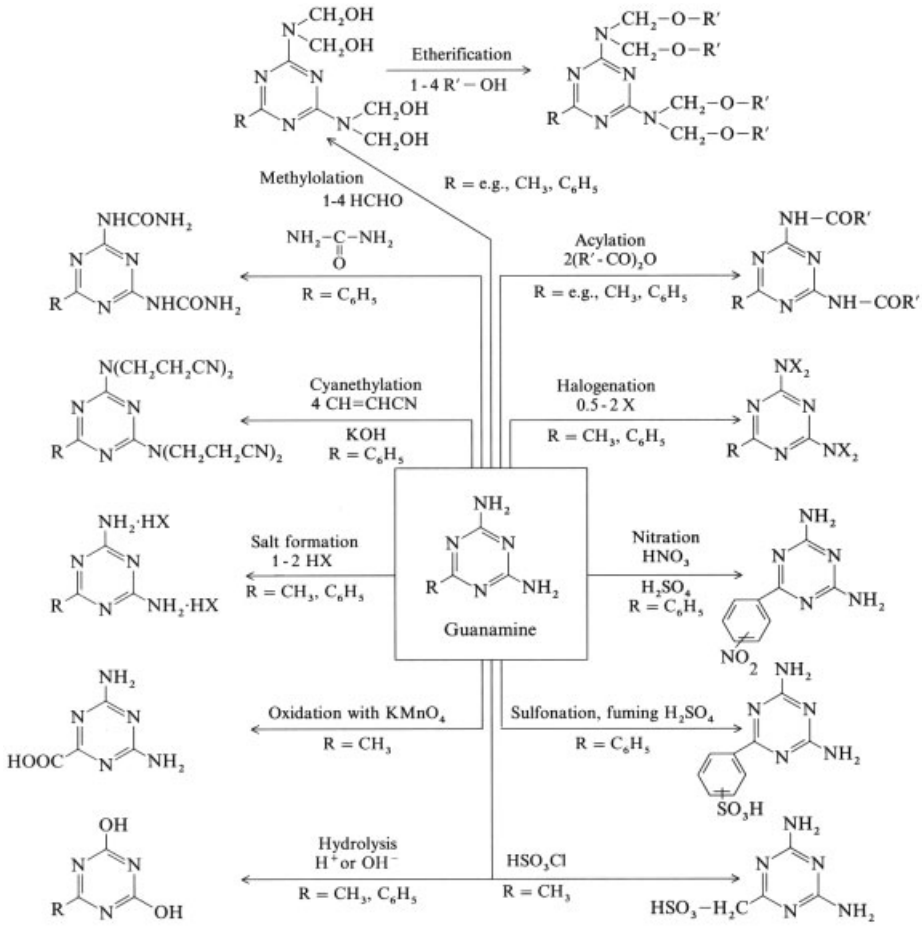

Figure 6. Reactions of guanamines

11.1. Production

2,4-Diamino-1,3,5-triazines can be prepared from several aliphatic C-N-compounds [5] including the historic synthesis from biguanides with esters [57, 58]. Aromatic guanamines can be obtained by Pd-catalyzed Suzuki coupling from 1,3,5-triazine precursors [59]. The only method of industrial relevance, however, involves the reaction of dicyandiamide (cyanoguanidine) with nitriles [60]:

Use of dinitriles leads to bisguanamines. Reaction occurs at 105 – 120 °C in the presence of alkaline catalysts (e.g., KOH) in polar solvents, usually alcohols, and high yields are obtained. The rate of reaction depends on the structure of the nitrile, the nature of the solvent, and the concentration of alkali. Aliphatic nitriles generally react more slowly than aromatic nitriles. The reaction can also be performed under microwave irradiation [61] or in ionic liquids [62].

11.2. Uses

Acetoguanamine is used as a condensation component in amino resins. Occasional use is made of the *pure* resins that result from reaction of acetoguanamine with formaldehyde. Such resins display a high degree of water tolerance and can be cured thermally. The rate of the condensation reaction is considerably less dependent on pH than in the case of melamine. Condensation may be carried out in weakly basic or weakly acidic media. Compared to melamine – formaldehyde resins, the curing of acetoguanamine resins occurs more slowly and at a more acidic pH.

Acetoguanamine is normally used to *modify* melamine – formaldehyde resins, conferring improved elasticity, higher gloss, and reduced resin shrinkage. Main application is in the decorative layer of high-pressure laminates (HPL) used for flooring and, e.g., kitchen plates. Acetoguanamine-modified melamine resins provide high gloss, low stain receptivity, and postforming properties. Even low amounts of acetoguanamine (5-10% based on melamine) are sufficient to allow small bending radii without affecting surface hardness or thermal stability [63]. Newer developments aim at reducing the formaldehyde content of melamine – acetoguanamine resins in order to avoid formaldehyde emissions after production [64]. Due to the relatively high cost of acetoguanamine, mixtures with other modifying agents, e.g., dicyandiamide have been applied, especially for short cycle laminates on wood-based materials [65](see also → Wood). Partially alkoxylated acetoguanamine – formaldehyde resins have been proposed for high-temperature fibers with good flexibility [66]. Acetoguanamine cyanurate, either alone or with melamine cyanurate, serves as an effective flame retardant in polyamides [67, 97].

Benzoguanamine is also used primarily in amino resins. Compared to acetoguanamine, benzoguanamine produces significantly more hydrophobic resins that have lower lightfastness due to the aromatic ring.

The most important application of benzoguanamine is in resins for industrial paints. Methylolated benzoguanamine resins that have been etherified with butanols are compatible with hydrocarbon solvents, oils, and various synthetic resins (e.g., alkyd, polyester, epoxy [68] and acrylic resins [69]). Upon curing, these OH functional resins react with the butoxymethyl groups of the benzoguanamine resin to form new cross-links, by splitting off butanol. In contrast to similar melamine resins, benzoguanamine resins give a lower cross-link density, and thus a higher flexibility and better surface quality. Typical applications are can coatings, coil coatings, and automotive basecoats. Benzoguanamine-carboxylic acids have been proposed as cross-linkers for water-borne coatings [70].

Due to its favorable physicochemical properties, *tetra(methoxymethyl)benzoguanamine* [*4588-69-6*] has been proposed as a cross-linker for powder coatings based on hydroxy polyesters [71, 72]. However, bubbles formed during the elimination of methanol have limited its use to date.

Fully condensed formaldehyde – benzoguanamine resins that mainly consist of methylene-bridged structures are amorphous, insoluble solids with high thermal stability. Due to their high refractive index, these resins are used as white pigments or colorants in plastics [73] or, in conjunction with fluorescent dyes, as pigments for daylight fluorescent inks and paints [74, 75]. Benzoguanamine is used as a flexibilizing agent in phenolic resins to improve the punchability, water resistance, and flame-retardance of FR2-type printed circuit boards [76]. In the same way, it can be used as additive to urea – formaldehyde foams [77] or bismaleimide copolymer resins [78]. Both, acetoguanamine and benzoguanamine have been proposed as epoxy cur-

Table 5. Typical storage lifetime of formaldehyde solutions stabilized with guanamines

Formaldehyde concentration	Guanamine added	Storage temperature	Storage lifetime
40 %	0.1 % benzoguanamine	10 °C	2 weeks
40 %	0.05 % caprinoguanamine	10 °C	4 weeks
50 %	0.05 % caprinoguanamine	40 °C	4 weeks

ing agents with improved properties for microelectronic encapsulation [79, 80]. By reaction with 4,4'-diaminodiphenyl ether, benzoguanamine forms a bridged bisguanamine which was proposed for resin applications [81].

Benzoguanamine can be used instead of dichlorophenyltriazine to produce Pigment Yellow 184, an anthraquinone dye with a triazine subunit [82].

Caprinoguanamine is one of the best known stabilizers for aqueous formaldehyde solutions. It prevents or retards the precipitation of polymeric paraformaldehyde at low storage temperatures and/or high formaldehyde concentrations [83]. Though less effective, benzoguanamine is sometimes used for the same purpose. Typical stability data of aqueous formaldehyde solutions is given in Table 5.

Caprinoguanamine has been applied as additive for melamine – formaldehyde impregnation resins with improved wetting characteristics [84]. It is also used in styrenic foams to improve the foam structure and provide flame retardancy [85] (see → Flame Retardants).

Other Guanamines. Guanamines with unsaturated groups in the substituent R can also be converted to polymers and copolymers [86]. The compound "CTU" (bis-3,9-cyanoethyl-2,4,8,10-tetraoxaspiro[5.5]-undecan)-guanamine [*22535-90-6*] (formula shown below) is employed (like benzoguanamine) primarily in paints [87].

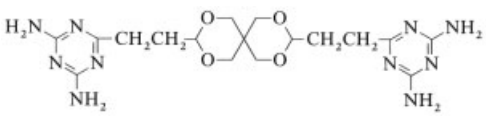

Guanamines with imidazole side chains, e.g., 2,4-diamino-6-(2′-methylimidazol-1′-yl)-ethyl-1,3,5-triazine are used as epoxy curing agents and accelerators as such or in form of adducts with, e.g., cyanuric acid [88].

Guanamine structures have also found application in pharmacology as c-AMP specific phosphodiesterase inhibitors [89], the most prominent being Irsogladine maleate (2′,5′-Dichlorobenzoguanamine, [*84504-69-8*]).

11.3. Toxicology

The acute oral LD_{50} for *acetoguanamine* in rats is 2740 mg/kg; for *benzoguanamine,* 1470 mg/kg. The corresponding value for *caprinoguanamine* is > 10 000 mg/kg, and its dermal LD_{50} (rabbit) is > 2800 mg/kg [90]. All three compounds are not irritating to skin and eyes [91, 92]. Acetoguanamine and benzoguanamine did not reveal mutagenic activity in the Ames test [93]. Long term feeding studies with benzoguanamine in male rats and mice failed to reveal carcinogenic potential [94]. Benzoguanamine was adopted by the Scientific Committee on Food as a monomer for food contact materials in List 3 with a restriction of 5 mg/kg of benzoguanamine in food [95]. The evaluation of this substance under the OECD/High Production Volume Program resulted in the conclusion, that there is low priority for further work (no further work recommended) [96].

12. References

General References

1. E. M. Smolin, L. Rapoport in A. Weissberger (ed.): *The Chemistry of Heterocyclic Compounds*, vol. **13**, Interscience, New York 1959, pp. 309 – 379.
2. J. M. E. Quirke in A. R. Katritzky, C. W. Rees (eds.): *Comprehensive Heterocyclic Chemistry*, vol. **3 part 2 B**, Pergamon Press, Oxford 1984, pp. 457 – 530.
3. SRI International: *Melamine*, Report No 122, 1978.
4. "The Manufacture of Non-Fertilizer Nitrogen Products," *Nitrogen* **139** (1982) 32 – 39.
5. Houben-Weyl, *Science of Synthesis*, vol. 17 (2004), 449 ff.

Specific References

6. Amer. Cyanamid, US 2 566 231, 1943 (J. H. Paden, J. S. Mackay).
7. Melamine Chemicals, Donaldsonville, Louisiana, 70346 (USA): *Product Brochure,* 1980, pp. 4 – 5.
8. C. C. Stephenson, D. J. Berets, *J. Am. Chem. Soc.* **74** (1952) 882 – 883.
9. R. C. Hirt, J. E. Steger, G. L. Simard, *Polymer Sci.* **XLIII** (1960) 319 – 323.
10. R. P. Chapmann, P. R. Averell, R. R. Harries, *Ind. Eng. Chem.* **35** (1943) 137.
11. A. C. Larson, D. T. Cromer, *J. Chem. Phys.* **60** (1974) no. 1, 185 – 192.
12. G. van der Plaats, H. Soons, R. Snellings, *Proc. Eur. Symp. Therm. Anal.* 2nd 1981, 215 – 218. L. Costa, G. Camino: "Thermal Behaviour of Melamine," *J. Therm. Anal.* **34** (1988) 423 – 429.
13. H. May, *J. Appl. Chem.* **9** (1959) 340 – 344.
14. A. I. Finkel'shtein, N. V. Spiridonova, *Russ. Chem. Rev. (Engl. Transl.)* **33** (1964) no. 7, 400 – 405.
15. T. Morikawa: "Evolution of Hydrogen Cyanide During Combustion and Pyrolysis," *J. Combust. Toxicol.* **5** (1978) 315 – 330.
16. R. I. Spasskaya, A. I. Finkel'shtein, E. N. Zil'bermann, Ts. N. Roginskaya, *Plastich. Massy*, (1980) no. 3, 7; *Int. Polym. Sci. Technol.* **7** (1980) no. 6, 69 – 71.
17. W. Woebcken (ed.): *Kunststoff-Handbuch*, vol. 10, "Duroplaste," 2nd ed., Hanser Verlag, München 1988.
18. M. Schwarzmann: "Make Melamine at Atmospheric Pressure," *Hydrocarbon Process.* **47** (1968) no. 10, 96 – 100.
19. A. Schmidt: "über den Reaktionsmechanismus der Bildung von Melamin aus Isocyansäure," *Monatsh. Chem.* **99** (1968) 664 – 671.
20. A. Schmidt: "Verfahrenstechnische Probleme bei der Herstellung von Melamin aus Harnstoff bei Atmosphärendruck", *Oesterr. Chem. Ztg.* **68** (1967) 175 – 179.
21. K. Abe et al.: "Synthetic Production of Melamine by a High-Pressure Process," *Kagaku Kogaku* **40** (1976) no. 6, 298 – 302.
22. Österreichische Stickstoffwerke, GB 1 054 502, 1967 (Stevens, Langner, Perry, Rollinson).
23. M. Schwarzmann: "Make Melamine at Atmospheric Pressure," *Hydrocarbon Process.* **48** (1969) no. 9, 184 –186.
24. Allied Chemical, US 3 250 773, 1966 (I. Christoffel et al.).
25. Montecatini, US 3 172 887, 1965 (E. Bondi).
26. Amer. Cyanamid, US 2 615 019, 1952 (W. J. Klapproth Jr.,)
27. *Kirk-Othmer*, 2nd ed., **2**, 366 – 367, **6**, 566.
28. "Melamine OSW Process," *Br. Chem. Eng.* **14** (1969) no. 10, 1336.
29. Nissan Chemical Industries, GB 1 032 507, 1966 (A. Murata et al.).
30. Y. T. Nakajima: "Development of a High-Pressure Process for Melamine Manufacture from Urea," *Nikkakyo Geppo* **28** (1975) no. 7, 175 – 179.
31. Office National Industriel de l'Azote, US 3 262 759, 1966 (J. L. Pomot et al.).
32. Melamine Chemicals, US 4 565 867, 1984 (R. E. Thomas, D. E. Best).
33. P. G. Stoks, A. W. Schwartz: "Determination of s-Triazine Derivatives at the nanogram level by Gas-Liquid Chromatography," *J. Chromatogr.* **168** (1979) 455 –460.
34. T. Götze, P. Dorries: "Neuentwicklungen auf dem Gebiet der technischen Melamin-Formaldehydharze," *Plastverarbeiter* **33** (1982) 1118 – 1122.
35. BASF Aktiengesellschaft, unpublished results.
36. C. B. Shaffer: *Melamine: Acute and Chronic Toxicity*, Amer. Cyanamid, Central Medical Department, Wayne, N.J., June 15,1955.
37. C. F. Reinhardt, M. R. Briselli in G. D. Clayton, F. E. Clayton (eds.): *Pattys Industrial Hygiene and Toxicology*, 3rd ed., John Wiley and Sons, New York 1981, p. 2770.
38. S. Haworth et al., *Environ*. Mutagen. Suppl. 1 (1983) 3.
39. ECETOC, Joint Assessement of Commodity Chemicals (JACC) no. 1, Melamine, CAS: 108–78–1, February (1983).
40. J. Mirsalis et al., *Environ. Mutagen.* **5** (1983) 482.
41. D. B. McGregor et al., *Environ. Mol. Mutagen.* **12** (1988) 85.
42. S. M. Galloway et al., *Environ. Mol. Mutagen.* **10** (1987) 1.
43. G. Röhrborn, *Vererbungs.* **93** (1962) 1.
44. R. W. Mast et al., *Environ. Mutagen.* **4** (1982) 340.
45. R. W. Mast et al., *Food Chem. Toxicol.* **21** (1983) 807.
46. R. L. Melnick et al., *Toxicol. Appl. Pharmacol.* **72** (1984) 292.
47. *IARC Monogr. Eval. Carcinog. Risk Chem. Man.* **39**, 11 – 18 June (1985) 333.
48. J. B. Thiersch, *Proc. Soc. Exp. Biol. Med.* **94** (1957) 36.

49. SKW Trostberg AG, Produktstudie *Guanamine,* 1979.
50. Henkel, DE-OS 2 019 675, 1970 (J. Brüning, H.-W. Eckert, P. Krings).
51. Henkel, DE-OS 1 964 793, 1969 (J. Brüning).
52. BASF, DE-OS 2 018 719, 1970 (M. Schwarzmann).
Henkel, DE-OS 1 445 907, 1964 (J. Schiefer).
53. E. Riesz, C. Heutrel, *Rev. Gen. Caoutch. Plast.* **46** (1969) 237 – 241.
54. Olin Mathieson Chem. Corp., DE-AS 1 245 385, 1963 (D. W. Kaiser, J. K. Zane).
55. E. Smolin, L. Rapoport: *s-Triazines and Derivatives*, Interscience, New York 1959, pp. 222 – 258.
56. *Houben-Weyl*, 4th ed., **E 9c**, 667ff.
57. Kaken Yakukako, DE-OS 2 121 694, 1971 (S. Murai, K. Yoshida, C. Tomioka).
58. American Cyanamid Co, US 3162633, 1964 (J. T. Shaw).
59. G. Cooke, H. Augier de Cremiers, V. M. Rotello, B. Tarbit, P. E. Vanderstraeten, *Tetrahedron* **57** (2001) 2787.
60. Olin Mathieson Chem. Corp., US 3 330 830, 1964 (D. W. Kaiser).
Armour Pharmaceutical Comp., US 3 655 892, 1969 (C. D. Bossinger, T. Enkoji).
Y. Ogata et al., *Tetrahedron* **20** (1964) 2755 – 2761,
22 (1966) 157 – 165.
Y. Iwakura et al., *Bull. Chem. Soc. Jpn.* **38** (1965) 1820 – 1824.
61. A. Diaz-Ortiz et al., *New Journal of Chemistry* **28** (2004) 952.
62. Y. Peng, G. Song, *Tetrahedron Lett.* **45** (2004) 5313.
63. J. Seeholzer, *Kunststoffe* **69** (1979) 263 – 265.
64. DSM N.V., NL 9 200 158, 1993 (H. M. C. Stijnen).
65. SKW Trostberg AG, US 5346 937, 1994 (J. Kuchler, H-G. Erben, J. Seeholzer).
66. BASF AG, WO 97/31958 (E. Guenther, W. Reuther).
67. Toray, US 4 383 064, 1983 (H. Jida).
68. General Mills Inc., US 3 271 350, 1963 (L. R. Vertnik).
69. Rohm & Haas, US 3 082 184, 1959 (D. R. Falgiatore, A. M. Levantin).
70. Dainippon Ink and Chemicals, EP 0860473, 1997 (R. Obayashi, T. Yamashita, S. Koshizawa, M. Iwahashi, T. Sato, K. Onishi, M. Takahashi).
71. S. Sander, *J. Appl. Polym. Sci.* **13** (1969) 555.
72. Hoechst, DE-OS 1 926 035, 1969 (H. Fischer, H.-J. Schubert).
73. Nippon Shokubai Kagaku Kogyo, DE-AS 2 349 964, 1973 (T. Tsubakimoto, J. Fuzikawa, O. Toyonaka).
DE-OS 2 634 415, 1976 (T. Tsubakimoto, J. Fuzikawa, O. Toyonaka).
74. Lawter Chemicals, US 3 116 256, 1961 (G. F. D'Alelio, R. W. Voedisch).
75. Nippon Shokubai K.K. Co., DE 3 628 244, 1987 (I. Fujikawa, M. Kushino, Y. Yamamoto).
76. Toshiba Chemical Co., JP 02 135 258, 1990, (T. Suzuki).
77. EP 319 751, 1989 (W. D. Schoellhorn).
78. Rhône-Poulenc Chimie, FR 2 645 539, 1990 (Y. Camberlin, R. Arpin).
79. Toyo Ink, JP-Kokai 62 218 454, 1962 (J. Katsumasa, T. Kojiro, K. Keiji).
Matsushita Electric, JP-Kokai 61 246 227, 1961 (H. Hirohisa, F. Taro).
F. Ueda, S. Aratani, K. Mimura, K. Kimura et al., *Arzneim. Forsch./Drug Res.* **34** (1984) 478 – 484.
80. Matsushita Electric Works Ltd., JP 61 246 227, 1986 (H. Hino, T. Fukui, S. Shinji, M. Tsujimoto).
81. Sumitomo Bakelite Co., JP 62 226 969, 1987 (R. Nakatsuka, K. Ishii, T. Yoshioka).
82. Ciba-Geigy A.G., US 3 839 332, 1974 (E. Moergeli).
83. Süddeutsche Kalkstickstoff-Werke, DE 1 205 073, 1964 (P. Bornmann, H. Michaud).
84. Hoechst AG, EP 710 682, 1996 (J. Ott et al.).
85. BASF AG, DE 3914292, 1990 (K. Hahn, U. Guhr, H. Hintz, K. H. Roehrig).
86. Cassella, GB 1 174 523, 1968 (H. v. Brachel, H. Kindler, H. Gattner).
Rohm & Haas, US 3 554 684, 1969 (W. D. Emmous, J. G. Brodmyan).
87. Toka Shikiso and Ajinomoto, US 3 557 059, 1971 (H. Yoko, O. Tamiki, T. Kijchiro, M. Takao).
88. Shikoku Chemicals Corp., EP 194 895, 1986 (K. Kamagata, T. Yamada, N. Sawa).
89. Nippon Shinyaku Co., WO 2005/026132 (K. Takashi).
90. Central Institute for Nutrition and Food Research TNO, unpublished results.
91. Huntington Research Center Ltd., unpublished results.
92. RCC NOTOX B.V., unpublished results.
93. Microtest Research Ltd., unpublished results.
94. Bio-Research Consultants, Inc., Contrast no. NIH-NCI-E68-1311, Carcinogenicity of Chemicals Present in Man's Environment, Final report (July 1973).

95. Opinion of the Scientific Committee on Food on the 16th additional list of monomers and additives for food contact materials, PM/Ref. No. 15310, SCF/CS/PM/GEN/M88 Final, December 21, 2001.
96. UNEP Publications, OECD SIDS 2,4-Diamino-6-phenyl-1,3,5-triazine, SIAM 13, Bern, Switzerland, November 6 – 9, 2001.
97. Shikoku Chemicals Corp., JP 2000/063365, (T. Masuda, S. Morishita, H. Kawata, N. Kano).

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Membranes and Membrane Separation Processes

Standard Article

Heinrich Strathmann[1]

[1]Institut für chemische Verfahrenstechnik, Universität Stuttgart, Federal Republic of Germany

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a16_187.pub2
Article Online Posting Date: April 30, 2005

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Historical Development of Membranes and Membrane Processes
1.2. Advantages and Limitations of Membrane Processes
1.3. The Membrane-Based Industry
1.4. The Future of Membrane Science and Technology
2. Fundamentals
2.1. Definition of Terms
2.1.1. Definition of a Membrane and its Function
2.1.2. Definition of the Various Membrane Processes
2.1.3. The Membrane Transport Mechanism
2.2. Materials and Structures of Synthetic Membranes
2.2.1. Symmetric and Asymmetric Membranes
2.2.2. Porous Membranes
2.2.3. Homogeneous Membranes
2.2.4. Ion-Exchange Membranes
2.2.5. Liquid Membranes
2.2.6. Fixed-Carrier Membranes
2.3. Fluxes and Driving Forces in Membrane Separation Processes
2.4. Mathematical Description of Mass Transport in Membrane Processes
2.5. Membrane Separation Capability
2.6. Chemical and Electrochemical Equilibrium in Membrane Processes
2.6.1. Osmotic Equilibrium, Osmotic Pressure, Osmosis, and Reverse Osmosis
2.6.2. Donnan Equilibrium and Donnan Potential
3. Principles of Membrane Separation Processes
3.1. Principle of Microfiltration
3.2. Principle of Ultrafiltration
3.3. Principle of Reverse Osmosis
3.3.1. Mass Transport in Reverse Osmosis Described by Phenomenological Equations
3.3.2. Description of Mass Transport in Reverse Osmosis by the Solution-Diffusion Model
3.4. Principle of Gas Separation
3.4.1. Gas Separation by Knudsen Diffusion
3.4.2. Gas Transport by the Solution-Diffusion Mechanism in a Polymer Matrix
3.5. Principle of Pervaporation
3.6. Principle of Dialysis
3.6.1. Dialytic Mass Transport of Neutral Components
3.6.2. Dialytic Mass Transport of Electrolytes in a Membrane without Fixed Ions
3.6.3. Dialysis of Electrolytes in Ion-Exchange Membranes

3.7. Principle of Electrodialysis
3.7.1. Electrical Current and Ion Fluxes
3.7.2. Transport Number and Membrane Permselectivity
3.7.3. Membrane Counterion Permselectivity
3.7.4. Water Transport in Electrodialysis
3.8. Electrodialysis With Bipolar Membranes
3.9. Energy Requirements of Membrane Processes
3.9.1. Minimum Energy Required for Separation
3.9.2. Practical Energy Requirements of Membrane Processes
4. Membrane Separation Process Design and Operation
4.1. Ultrafiltration, Microfiltration, and Reverse Osmosis
4.1.1. Recovery Rate, Membrane Rejection, Retentate and Filtrate Concentrations
4.1.2. Solute Loss in Filtration
4.1.3. Effect of Osmotic Pressure on the Filtration Rate
4.1.4. Process Operating Mode
4.1.5. Diafiltration Operating Mode
4.2. Gas Separation
4.3. Pervaporation
4.4. Dialysis
4.5. Electrodialysis
4.5.1. Mass Transport in an Electrodialysis Stack
4.5.2. Energy Requirements in an Electrodialysis Desalination Process
4.5.3. Electrodialysis Process Costs
4.6. Bipolar Membrane Electrodialysis Process Design
4.6.1. Operating Modes in Bipolar Membrane Electrodialysis
4.6.2. Energy Requirements in Electrodialytic Water Dissociation
4.6.3. Total Energy Required for Production of an Acid and a Base
4.6.4. Total Process Costs
4.7. Continuous Electrodeionization Process Design
4.8. Other Membrane Processes
5. Membrane Preparation and Membrane Module Constructions
5.1. Membrane Preparation
5.1.1. Preparation of Porous Membranes
5.1.2. Preparation of Asymmetric Membranes
5.1.3. Rationalization of the Phase-Inversion Membrane Preparation Process
5.1.4. Preparation of Composite Membranes
5.1.5. Preparation of Inorganic Composite Membranes
5.1.6. Preparation of Homogeneous Solid Membranes
5.1.7. Preparation of Liquid Membranes
5.1.8. Preparation of Ion-Exchange Membranes
5.2. Membrane Characterization
5.2.1. Characterization of Porous Membranes
5.2.1.1. Membrane Separation Properties Determined by Filtration Tests
5.2.2. Characterization of Homogeneous Membranes
5.3. Membrane Modules
5.3.1. Cartridge Membrane Module
5.3.2. Plate-and-Frame Membrane Module
5.3.3. Spiral-Wound Module
5.3.4. Tubular Membrane Module
5.3.5. Capillary and Hollow-Fiber Membrane Modules
5.3.6. Other Membrane Modules
6. Concentration Polarization and Membrane Fouling
6.1. Concentration Polarization in Filtration Processes
6.1.1. Concentration Polarization without Solute Precipitation
6.1.2. Concentration Polarization with Solute Precipitation at the Membrane Surface
6.2. Membrane Fouling and its Causes
6.3. Concentration Polarization in Other Membrane Separation Processes
6.3.1. Concentration Polarization in Dialysis
6.3.2. Concentration Polarization in Electrodialysis

6.3.3. Concentration Polarization in Pervaporation
6.3.4. Concentration Polarization in Gas Separation

[Top of Page]

List of Symbols

A	m^2	surface area
a	mol m^{-3}	activity
a	–	constant
C	mol m^{-3}, mol L^{-1}	concentration
D	m	batch cell diameter
D	m s^{-1}	dialysance
D	m^2 s^{-1}	diffusion coefficient
$Đ$	m^2 s^{-1}	Stefan-Maxwell diffusion coefficient
d	m	distance
d_H	m	hydraulic diameter
d	m	stirrer length
E	A V s	energy
F	96485 A s eq^{-1}	Faraday constant
f	N s m^{-1} mol^{-1}	friction coefficient
ΔG^m	J	Gibbs free energy of mixing
g	kg m s^{-2}	acceleration
g	–	osmotic coefficient
h	m	height
I	A	current
i	A m^{-2}	current density
i_{lim}	A m^{-2}	limiting current density
J	mole m^{-2} s^{-1}	flux
K_w	mol^2 L^{-2}	water dissociation constant
k	various	coefficient
k	J mol^{-1} K^{-1}	Boltzmann constant
L	mol^2 N^{-1} m^{-2} s^{-1}	phenomenological coefficient
L_{pp}	m	process path length
M	kg mol^{-1}	molecular weight
m	m s^{-1}	mobility
N	mol s^{-1}	flux
N	–	number
n	–	number of moles
P	mol m^{-1} s^{-1} J^{-1}	permeability
p	Pa	pressure
Q	m^3 s^{-1}	volume flux
R	Ω	electrical resistance
R	J mol^{-1} K^{-1}	gas constant
R	–	rejection
R_m	N m s^{-1}	membrane resistance
r	Ω m^2	area resistance
r	N s^{-1}	specific resistance
r	m	radius
Re	–	Reynolds number
S	–	permselectivity
Sc	–	Schmidt number
Sh	–	Sherwood number

T	K	temperature
T	–	transport number
t	s	time
U	V	electrical potential
u	$m^2\ s^{-1}\ V^{-1}$	ion mobility
u	$m\ s^{-1}$	velocity
V	m^3	volume
$\bar{V}$	$m^3\ mol^{-1}$	partial molar volume
v	$m\ s^{-1}$	velocity
w	m	width
X	m	cell length
X	$Nm\ mol^{-1}$	driving force
X	–	mole fraction
x	m	directional co-ordinate
Y	m	cell width
Z_b	μ	boundary layer thickness
z	m	directional coordinate
z	$eq\ mol^{-1}$	valence

Greek letters

α	–	separation factor
α	–	dissociation degree
β	–	enrichment factor
γ	$m^3\ mol^{-1}$	activity coefficient
Δ	m	cell thickness
Δ	–	difference
Δ	–	recovery rate
δ	–	fractional product loss
ε	–	porosity
η	$kg\ m^{-1}\ s^{-1}$	dynamic viscosity
η	$A\ V\ s\ mol^{-1}$, $J\ mol^{-1}$	electrochemical potential
θ	–	stage cut
κ	$\Omega^{-1}\ m^{-1}$	specific conductivity
Λ_{eq}	$m^2\ \Omega^{-1}\ eq^{-1}$	electrolyte equivalent conductivity
λ	m	distance
λ	$m^2\ \Omega^{-1}\ eq^{-1}$	ion equivalent conductivity
μ	$J\ mol^{-1}$	chemical potential
ν	$m^2\ s^{-1}$	kinematic viscosity
ν	–	stoichiometric coefficient
ξ	–	current utilization
π	Pa, $N\ m^{-2}$	osmotic pressure
ρ	$kg\ m^{-3}$	density
ρ	Ω m	specific resistance
σ	$mol\ N^{-1}\ s^{-1}$	reflection coefficient
σ	$N\ m^{-1}$	surface tension
τ	–	tortuosity factor
φ	°	contact angle
φ	V	electrical potential
φ	–	fugacity coefficient
ϕ	–	pressure ratio
Ψ	$N\ m\ s^{-1}$	dissipation function
Ψ	–	membrane permselectivity
ω	s^{-1}	circular velocity
ω		

	mol m^{-1}	permeability
ω	–	porosity

Subscripts

A	component
a	anion
B	component
c	cation
cell	dialysis cell
co	co-ion
con	convection
D	diffusion
Don	Donnan potential
des	desalination
diff	diffusion
e	electric charge
fix	fixed ion
g	gas
gas	gas
i	component
j	component
k	component
l	feed
m	membrane
p	pore
p	pressure
pro	product
s	salt
s	solute
s	solution
spec	specific
st	stack
tr	transition layer
v	volume
w	water, solvent
o	standard state

Superscripts

am	anion-exchange membrane
b	bulk solution
b	constant
bml	bipolar membrane
c	concentrate
c	constant
cm	cation-exchange membrane
D	diffusion
d	constant
d	dialysate
d	diluate
f	feed
fc	feed concentrate
fd	feed diluate
g	gas
in	inlet
k	Knudsen diffusion
m	membrane

out	outlet
p	product
p	permeate
r	retentate
s	solution
t	time
tr	transition region
w	membrane wall
_	average number
o	standard state

[Top of Page]

1. Introduction

The separation, concentration, and purification of molecular mixtures are major problems in the chemical industries. Efficient separation processes are also needed to obtain high-grade products in the food and pharmaceutical industries, to supply communities and industry with high-quality water, and to remove or recover toxic or valuable components from industrial effluents. For this task a multitude of separation techniques such as distillation, precipitation, crystallization, extraction, adsorption, and ion exchange are used today. More recently, these conventional separation methods have been supplemented by a family of processes that utilize semipermeable membranes as separation barriers. Membranes and membrane processes were first introduced as an analytical tool in chemical and biomedical laboratories, and they developed very rapidly into industrial products and methods with significant technical and commercial impact [1, 2]. Today, membranes are used on a large scale to produce potable water from seawater, to clean industrial effluents and recover valuable constituents, to concentrate, purify, or fractionate macromolecular mixtures in the food and drug industries, and to separate gases and vapors. They are also key components in energy conversion and storage systems, artificial organs, and drug delivery devices.

The membranes used in the various applications differ widely in their structure, their function, and the way they are operated in a separation process. However, all membranes and membrane processes share several features that make them particularly attractive tools for the separation of molecular mixtures. The separation is performed by physical means at ambient temperature without chemically altering the constituents. This is important for applications in artificial organs and in many drug delivery systems, as well as in the food and drug industries or in downstream processing of bioproducts where temperature-sensitive substances must often be handled. Furthermore, membrane properties can be tailored and adjusted to specific separation tasks, and membrane processes are often technically simpler and more energy efficient than conventional separation techniques and are equally well suited for large-scale continuous operations as for batchwise treatment of very small quantities.

A key property of a membrane is its permselectivity, which is determined by differences in the transport rate of various components through the membrane. The transport rate of a component through a membrane is determined by the structure of the membrane, the size of the permeating component, by the chemical nature and the electrical charge of the membrane material and permeating component, and by the driving force due to the chemical or electrochemical potential gradients, that is, concentration pressure and electrical potential differences. Some driving forces such as concentration, pressure, and temperature gradients act equally on all components, in contrast to an electrical potential as driving force, which is only effective with charged components. The use of different membrane structures and driving forces has resulted in a number of rather different membrane processes such as reverse osmosis, micro- and ultrafiltration, dialysis, electrodialysis, Donnan dialysis, pervaporation, gas separation, and so on. The versatility of membrane structures and functions makes a precise and complete definition of a membrane rather difficult. In the most general sense a membrane can be described as a barrier that separates two different regions and controls the exchange of matter, energy, and information between the regions in a very specific way. The term “membrane” describes two very different items, that is, biological membranes, which are part of a living organism, and synthetic membranes, which are man-made. Biological membranes carry out very complex and specific transport tasks in organisms. They accomplish them quickly, efficiently, and with minimal energy expenditure, often utilizing carrier-facilitated and active transport coupled to a chemical reaction. Synthetic membranes are not nearly as complex in their structure or function as biological membranes. They are usually less selective and energy-efficient but have significantly higher chemical, thermal, and mechanical stability. In practical separation processes biological membranes have no significance today and will therefore not be discussed in this article. Because membranes are capable of separating molecular mixtures very selectively and efficiently at room temperature by physical means their potential applications are extremely wide compared to conventional separation processes. Today membranes are used on a large technical scale in three distinct areas: The first area includes applications in which the use of membranes is technically feasible, but where they must compete with conventional separation processes on the basis of overall economy. This, for instance, is the case in seawater desalination and the treatment of certain wastewater streams. Here, membranes must compete with distillation and biological treatment, respectively. The second area includes applications for which alternative techniques are available, but membranes offer a clear technical and commercial advantage. This is the case in the production of ultrapure water and in the separation of certain food products. Finally, there are applications where there is no reasonable alternative to membrane processes. This is the case in certain drug delivery systems and artificial organs.

1.1. Historical Development of Membranes and Membrane Processes

Synthetic membranes are a rather recent development and the technical utilization of membrane processes on a large scale began just 40 years ago. The first recorded study of membrane phenomena and the discovery of osmosis dates back to NOLLET, who discovered in 1748 [3] that a pig's bladder passed preferentially ethanol when it was brought in contact on one

side with a water – ethanol mixture and on the other side with pure water. NOLLET was probably the first to recognize the relation between a semipermeable membrane, osmotic pressure, and volume flux. More systematic studies on mass transport in semipermeable membranes were carried out by GRAHAM [4] in 1833, who studied the diffusion of gases through different media and discovered that rubber exhibits different permeabilities to different gases. Most of the early studies on membrane permeation were carried out with natural materials such as animal intestines, bladders, and gum elastics. In 1867 TRAUBE [5] was the first to introduce an artificially prepared semipermeable membrane by precipitating cupric ferrocyanide in a thin layer of porous porcelain. This type of membrane was used by PFEFFER [6] in his fundamental studies on osmosis. However, the theoretical treatment and much of the interpretation of osmotic phenomena and mass transport through membranes is based on the studies of FICK [7] in 1877, who interpreted diffusion in liquids as a function of concentration gradients, and those of VAN T' HOFF [8] in 1887, who gave a thermodynamic explanation for the osmotic pressure of dilute solutions. Later NERNST [9] in 1889 and PLANCK [10] in 1890 introduced the flux equation for electrolytes under the driving force of a concentration or electrical potential gradient. With the classical publications of DONNAN [11] in 1911, describing the theory of membrane equilibria and membrane potentials in the presence of electrolytes, the early history of membrane science ends with most of the basic phenomena satisfactorily described and theoretically interpreted. With the beginning of the 1900s membrane science and technology entered a new phase. First, BECHHOLD [12] developed in 1908 a method of making the first synthetic membranes by impregnating a filter paper with a solution of nitrocellulose in glacial acetic acid. These membranes could be prepared and accurately reproduced with different permeabilities by varying the ratio of acetic acid to nitrocellulose.

Nitrocellulose membranes were also used in the studies of ZSIGMONDY and BACHMANN [13] in 1918 as ultrafilters to separate macromolecules and fine particles from an aqueous solution. These studies were later continued by many others, including ELFORD [14] in 1930, and MCBAIN and KISTLER [15] in 1931. The relation between the streaming potential, electroosmosis, and electrodialysis was treated in 1931 in a monograph of PRAUSNITZ and REITSTÖTTER [16]. Based on a 1922 patent of ZSIGMONDY [17], Sartorius GmbH in Germany began in 1937 the production of a series of nitrocellulose membranes with various pore sizes. These membranes were used in microbiological laboratories in analytical applications. The development of the first successfully functioning hemodialyzer by KOLFF and BERK [18] in 1944 was the key to the large-scale application of membranes in the biomedical area. All synthetic membranes available up to that time were microporous structures separating mixtures exclusively according to their size. In the early days of membrane science and technology membranes had been mainly a subject of scientific interest with only a very few practical applications. This changed drastically from 1950 on, when the practical use of membranes in technically relevant applications became the main focus of interest and a significant membrane-based industry developed rapidly. Progress in polymer chemistry resulted in a large number of synthetic polymers which ultimately became available for the preparation of new membranes with specific transport properties plus excellent mechanical and thermal stability. Membrane transport properties were described by a comprehensive theory based on the thermodynamics of irreversible processes by STAVERMAN [19] in 1952, KEDEM and KATCHALSKY [20], SCHLÖGL [21], and SPIEGLER [22] in 1961. A second route for describing membrane processes was based on postulating certain membrane transport models, such as the model of a diffusion-solution membrane by MERTEN [23] in 1966. The properties of ion-exchange membranes were studied extensively by HELFFERICH [24] in 1962 and many others.

A milestone in membrane science and technology was the development of a reverse osmosis membrane based on cellulose acetate which provided high salt rejection and high fluxes at moderate hydrostatic pressures by REID and BRETON [25] in 1959 and LOEB and SOURIRAJAN [26] in 1962, which were major advances toward the application of reverse osmosis membranes as an effective tool for the production of potable water from seawater.

The membrane developed by LOEB and SOURIRAJAN had an asymmetric structure with a dense skin at the surface, which determined the membrane selectivity and flux, and highly porous substructure, which provided the mechanical strength. It was shown by KESTING [27] in 1971 that the preparation of asymmetric cellulose acetate membrane was based on a phase inversion process in which a homogeneous polymer solution is converted into a two-phase system, that is, a solid polymer-rich phase providing the solid polymer structure and a polymer-lean phase forming the liquid-filled membrane pores. STRATHMANN et al. [28] rationalized in 1971 the formation mechanism of asymmetric membranes and correlated different structures with the various preparation parameters. Soon, other synthetic polymers such as polyamides, polyacrylonitrile, polysulfone, and polyethylene were used as basic materials for the preparation of synthetic membranes. These polymers often showed better mechanical strength, chemical stability, and thermal stability than the cellulose esters. However, cellulose acetate remained the dominant material for the preparation of reverse osmosis membranes until the development of the interfacial-polymerized composite membrane by CADOTTE [29] and independently by RILEY [30] in 1980. These membranes showed significantly higher fluxes, higher rejection, and better chemical and mechanical stability than the cellulose acetate membranes.

The first membranes developed for reverse-osmosis desalination and other applications were manufactured as flat sheets and then installed in a so-called spiral-wound module, based on the work of BRAY [31] and WESTMORELAND [32] in 1982. A different approach to membrane geometry was taken in the early 1960s by MAHON [33], who made self-supporting hollow-fiber membranes with a wall thickness of only 6 – 7 µm. However, the permeate flux was rather low, presumably because the fibers were isotropic in structure. This problem was overcome later by Du Pont Corporation, which made an asymmetric hollow-fiber membrane with main application in desalination of brackish and seawater.

Soon after the development of efficient membranes, appropriate membrane housing assemblies, called modules, were devised. The criteria for the design of such modules included high membrane packing density, reliability, ease of membrane or module replacement, control of concentration polarization, and low cost. Membranes were produced in three different configurations: as flat sheets, as hollow fine fibers or capillaries, and as tubes. In modern reverse-osmosis desalination plants mainly spiral-wound modules are used, while hollow-fiber membrane modules are utilized in gas separation and pervaporation. In medical applications such as artificial kidneys and blood oxygenators, capillary membranes play a dominant role today. Tubular membranes are mainly used in micro- and ultrafiltration.

Even earlier than the large-scale use of reverse osmosis for desalination of sea- and brackish water was the industrial-scale application of electrodialysis. The history of electrodialysis goes back to the work of MEYER and STRAUSS [34], who developed the first multicell stack just before World War II. However, modern electrodialysis became a practical reality with

the work of JUDA and MCRAE [35] in 1953, who developed the first reliable ion-exchange membranes having both good electrolyte conductivity and ion permselectivity. Electrodialysis was first commercially exploited for the desalination of brackish water by Ionics Inc. The commercial success of Ionics was due to their membranes, their compact stacking, and the mode of operation referred to as electrodialysis reversal, which provided a periodic self-cleaning mechanism for the membrane stack and thus allowed long-term continuous operation at high concentrations of scaling materials without mechanical cleaning of the stack [36].

While Ionics pioneered the development of electrodialysis for water desalination a completely different use of electrodialysis was envisioned in Japan. Here, electrodialysis was used for concentrating sodium chloride solutions from seawater to produce table salt, as described by NISHIWAKI [37]. In the early 1980s a completely new area for the application of electrodialysis was opened up. At this time LIU et al. [38] introduced bipolar membranes for the recovery of acids and bases from the corresponding salts by water dissociation induced by an electrical potential.

The large-scale separation of gases and vapors is also a relevant industrial area for membrane applications. Gas separation was pioneered by HENIS and TRIPODI [39] in 1981. Originally, the aim was to recover hydrogen from off-gases and to produce oxygen- or nitrogen-enriched air. Today, however, a large number of other applications such as the removal of CO_2 from natural gas and the recovery of organic vapors from off-gases are typical applications for gas and vapor separation. Pervaporation, which in its basic principle is closely related to vapor separation, was studied extensively by APTEL and NEEL [40] in 1968, who pointed out a large number of interesting potential applications, but so far very few large commercial plants have been built.

Other applications of membranes which were developed in more recent years and have reached major technical and commercial significance include the controlled release of drugs in therapeutic devices and the storage and conversion of energy in fuel cells and batteries.

1.2. Advantages and Limitations of Membrane Processes

In many applications membrane processes compete directly with the more conventional techniques. However, compared to these conventional procedures membrane processes are often more energy efficient, simpler to operate, and yield higher quality products. Furthermore, the environmental impact of all membrane processes is relatively low. There are no hazardous chemicals used in the processes that have to be discharged and there is no heat generation. However, there are certain limitations to the application of membrane processes. A major disadvantage of membranes, especially in water- and wastewater-treatment processes, is that until today the long-term reliability has not completely been proven. Membrane processes sometimes require excessive pretreatment due to their sensitivity to concentration polarization, chemical interaction with water constituents, and fouling. Membranes are mechanically not very robust and can easily be destroyed by a malfunction in the operating procedure. Another critical issue are the process costs. In general, membrane processes are quite energy efficient. However, the energy consumption is only part of the total process costs. Other factors determining the overall economics of a process include the investment-related cost, which is determined by the cost of the membranes and other process equipment and their useful life under operating conditions, and various pre- and post-treatment procedures of the feed solutions and the products. Plant capacity may also play a role in total cost. While, for example, in distillation processes usually a substantial cost reduction can be achieved with an increase in the plant capacity, the scale-up factor has only a relatively small effect on the process cost in desalination of water by reverse osmosis or electrodialysis. Depending on the composition of the feed solution and the required quality of the product water, a combination of processes may be appropriate. For example, if ultrapure water for certain industrial applications is required, a sequence of processes may be applied, such as reverse osmosis, ion exchange, UV sterilization, and microfiltration as a point-of-use-filter to remove traces of particles.

The environmental impact of all membrane processes is relatively low. There are no hazardous chemicals used in membrane processes which have to be discharged. In certain applications such as the purification of industrial effluents and wastewaters or the desalination of brackish water there may be a problem with the disposal of the concentrate. In these applications brine post-treatment procedures may be necessary.

1.3. The Membrane-Based Industry

Parallel to the development of membranes and membrane processes was the development of a membrane-based industry. The structure of this industry is quite heterogeneous as far as the size of the companies and their basic concept towards the market is concerned. Several companies have concentrated on the production of membranes only. Membrane producers are frequently divisions of major chemical companies. They offer a range of membrane products as flat sheets, hollow fibers, or capillaries with different properties and for different applications ranging from seawater desalination and hemodialysis to purification of wastewater and fuel cells. Other companies buy membranes or modules as key components from one or several membrane manufacturers, design and build the actual plant, and often also operate it, guaranteeing the customer a certain amount of product of a given quality. Sometimes, these companies provide different processes, such as reverse osmosis, nanofiltration, and electrodialysis according to the requirements of the separation task. These companies provide a solution to a customer's separation problem in general, which might be a combination of separation processes such as ion exchange, carbon adsorption, flocculation and precipitation, or various chemical and biological treatment procedures in addition to membrane processes. Although the sales of membranes and membrane modules to any one of these companies are often not very large they are of importance in the membrane industry because of their specific application know-how in different markets. Finally, there are companies that provide the membranes, the system or product design, and the operation of the process. The companies concentrate very often on a single, usually very large application such as production of potable water from sea- or brackish water or hemodialysis. They often not only provide the tools for producing potable water in the case of seawater desalination, but also operate the plant and distribute the water. Companies producing artificial kidneys also operate dialysis stations. The membrane market is characterized by a few rather large market segments, such as sea- and brackish-water desalination, production of ultrapure water, and hemodialysis and a large number of small market segments in the food, chemical, and pharmaceutical industries, analytical laboratories, and especially in the treatment and recycling of industrial wastewater streams. The larger markets for water desalination and hemodialysis are dominated by a relatively small number of large companies. A multitude of small companies are active in market niches such as treating

certain wastewater streams or providing services to the chemical or food and drug industries.

1.4. The Future of Membrane Science and Technology

In many applications today's membranes and processes are quite satisfactory, while in other applications there is a definite demand for further improvements of both membranes and processes. For desalination of sea- and brackish water by reverse osmosis, for example, there are membranes available today that are quite satisfactory as far as flux and salt rejection are concerned, and the processes have been proven by many years of operating experience. The same is true for hemodialyzers and hemofiltration. In these applications only marginal improvements can be expected in the near future. In micro- and ultrafiltration and electrodialysis the situation is similar. The properties of current membranes are satisfactory. However, there are other components such as the process design, application know-how, and long-term operating experience that are of importance in the use of micro- and ultrafiltration in the chemical and food industries and in wastewater treatment. Here, concentration polarization and membrane fouling play a dominant role, and new membrane modules and process design concepts which provide a better control of membrane fouling and thus result in a longer useful life of the membranes are highly desirable. In other membrane processes such as gas separation, pervaporation, fuel cell separators, and membrane reactors, the situation is quite different. Here, better membranes, improved process design and extensive application know-how and long-term experience are mandatory to establish membrane processes as a proven and reliable technology.

[Top of Page]

2. Fundamentals

Membrane processes such as dialysis, electrodialysis, reverse osmosis, and micro- and ultrafiltration differ widely as far as the applied membrane structures and the driving forces for mass transport are concerned. There are, however, some common fundamentals, such as certain thermodynamic and kinetic relations describing mass transport in membranes and at the membrane/bulk solution interface. Phenomena such as concentration polarization and membrane fouling occur in all membrane processes and have common causes and consequences. Thus, before treating the different membrane processes and their application in more detail, some common fundamentals are discussed.

2.1. Definition of Terms

The performance of membrane processes in a certain application is determined by a number of components and process parameters such as the membranes, the membrane modules, the driving forces applied to achieve the desired mass transport and certain chemical engineering aspects relevant for designing processes. These components and process parameters and their function must be defined.

2.1.1. *Definition of a Membrane and its Function*

A precise and complete definition of a membrane that covers all its aspects is rather difficult, even when the discussion is limited to synthetic structures as in this article. In the most general sense, a synthetic membrane is an interphase that separates two phases and restricts the transport of various components in a specific manner [1]. A membrane can be homogeneous or heterogeneous, symmetric or asymmetric in structure. It may be solid or liquid and may consist of organic or inorganic materials. It can be neutral or can carry positive or negative charges, or functional groups with specific binding or complexing abilities. Its thickness can be less than 100 nm or more than one millimeter. The electrical resistance may vary from more than $10^6\ \Omega\ cm^2$ to less than $1\ \Omega\ cm^2$. The term “membrane” therefore includes a great variety of materials and structures, and a membrane can often be better described by its function rather than by its structure. Some materials, such as protective coatings, or packaging materials, though not meant to be membranes, show typical membrane properties, and are in fact membranes. All materials functioning as membranes have one characteristic property in common: They restrict the passage of different components in a very specific manner. Separation of a mixture in a membrane process is the result of different transport rates of different components through the membrane. The transport rate of a component through a membrane is determined by driving forces such as concentration, pressure, temperature, and electrical potential gradients, and the concentration and mobility of the component in the membrane matrix.

2.1.2. *Definition of the Various Membrane Processes*

Membrane processes can be grouped according to the applied driving forces into pressure-driven processes such as reverse osmosis or ultra- and microfiltration, concentration gradient driven processes such as dialysis and Donnan dialysis, and electrical potential driven processes such as electrodialysis. The molecular mixture to be separated is referred to as feed, the mixture containing the components retained by the membrane is called the retentate, and the mixture composed of the components that have permeated the membrane is referred to as permeate or filtrate in micro- and ultrafiltration.

2.1.3. *The Membrane Transport Mechanism*

The mechanism by which certain components are transported through a membrane can also be very different. In some membranes, for example, the transport is based on viscous flow of a mixture through individual pores in the membrane caused by a hydrostatic pressure difference between the two phases separated by the membrane. This type of transport is referred to as viscous flow. It is the dominant form of mass transport in micro- and ultrafiltration but also occurs in other membrane processes. If the transport through a membrane is based on the dissolution and diffusion of individual molecules in the membrane matrix due to a concentration or chemical potential gradient the transport is referred to as diffusion. This form of mass transport is dominant in reverse osmosis, gas separation, pervaporation, and dialysis but may occur in other processes, too. If an electrical potential gradient across the membrane is applied to achieve the desired transport of certain components through the membrane the transport is referred to as migration. Migration occurs in electrodialysis and related processes and is limited to the transport of components carrying electrical charges, such as ions.

2.2. Materials and Structures of Synthetic Membranes

Synthetic membranes show a large variety in their physical structure and the materials they are made from. Based on their structure they can be classified into four groups:

- Porous membranes
- Homogeneous solid membranes
- Solid membranes carrying electrical charges
- Liquid or solid films containing selective carriers

Furthermore, the structure of membranes may be symmetric, that is, the structure is identical over the entire cross section of the membrane, or it may be asymmetric, that is, the structure varies over the cross section of the membrane.

The materials used for the preparation of membranes can be polymers, ceramics, glass, metals, or liquids. The materials may be neutral or carry electrical charges (i.e., fixed ions). The schematic in Figure 1 illustrates the morphology and materials of some technically relevant synthetic membranes.

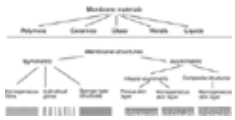

Figure 1. Schematic illustrating the various materials and structures of technically relevant synthetic membranes

2.2.1. *Symmetric and Asymmetric Membranes*

As indicated earlier synthetic membranes may have a symmetric or an asymmetric structure. In a symmetric membrane the structure and the transport properties are identical over the entire cross section, and the thickness of the entire membrane determines the flux. Symmetric membranes are used today mainly in dialysis and electrodialysis. In asymmetric membranes structural and transport properties vary over the membrane cross section. An asymmetric membrane consists of a 0.1 – 1 µm thick “skin” layer on a highly porous 100 – 200 µm thick substructure. The skin represents the actual selective barrier of the asymmetric membrane. Its separation characteristics are determined by the nature of the material or the size of its pores. The mass flux is determined mainly by the skin thickness. The porous sublayer serves only as a support for the mostly thin and fragile skin and has little effect on the separation characteristics or the mass transfer rate of the membrane. Asymmetric membranes are used primarily in pressure-driven membrane processes such as reverse osmosis, ultrafiltration, and gas and vapor separation, since here the unique properties of asymmetric membranes, that is, high fluxes and good mechanical stability, can best be utilized. Two techniques are used to prepare asymmetric membranes: one utilizes the phase-inversion process, which leads to an integral structure with the skin and the support structure made from the same material in a single process [27], and the other resembles a composite structure where a thin barrier layer is deposited on a porous substructure in a two-step process [29]. In this case barrier and support structures are generally made from different materials.

2.2.2. *Porous Membranes*

A porous structure represents a very simple form of a membrane, which closely resembles the conventional fiber filter as far as the mode of separation is concerned. These membranes consist of a solid matrix with defined holes or pores which have diameters ranging from less than 1 nm to more than 10 µm. Separation of the various components is achieved strictly by a sieving mechanism with the pore diameters and the particle sizes being the determining parameters. Porous membranes can be made from various materials such as ceramics, graphite, metals, metal oxides, and various polymers. Their structure can be symmetric, that is, the pore diameters do not vary over the membrane cross section, or asymmetric, that is, the pore diameters increase from one side of the membrane to the other, typically by a factor of 10 to 1000. The techniques for the preparation of porous membranes can be rather different and include simple pressing and sintering of polymer or ceramic powders, irradiation and leaching of templates, as well as phase-inversion and polymer-precipitation procedures or sol – gel conversion techniques. Porous membranes are used to separate components that differ markedly in size or molecular weight in processes such as micro- and ultrafiltration or dialysis [41].

2.2.3. *Homogeneous Membranes*

A homogeneous membrane is merely a dense film through which a mixture of molecules is transported by a pressure, concentration, or electrical potential gradient. The separation of the various components of a mixture is directly related to their transport rates within the membrane phase, which is determined by their diffusivities and concentrations in the membrane matrix. Therefore, homogeneous membranes are referred to as solution – diffusion-type membranes [23]. They can be prepared from polymers, metals, metal alloys, and, in some cases, ceramics, which may also carry positive or negative electrical charges. Since the mass transport in homogeneous membranes is based on diffusion, their permeabilities are rather low. Homogeneous membranes are used mainly to separate components which are similar in size but have different chemical nature in processes such as reverse osmosis, gas and vapor separation, and pervaporation [42]. In these processes asymmetric membrane structures are used which consist of a thin homogeneous skin supported by a porous substructure.

2.2.4. *Ion-Exchange Membranes*

Films carrying charged groups are referred to as ion-exchange membranes. They consist of highly swollen gels carrying fixed positive or negative charges. The properties and preparation procedures of ion-exchange membranes are closely related to those of ion-exchange resins [24]. There are two different types of ion-exchange membranes: 1) cation-exchange membranes, which contain negatively charged groups fixed to the polymer matrix, and 2) anion-exchange membranes, which contain positively charged groups fixed to the polymer matrix. In a cation-exchange membrane, the fixed anions are in electrical equilibrium with mobile cations in the interstices of the polymer. In contrast, the mobile anions are more or less completely excluded from the cation-exchange membrane because of their electrical charge, which is identical to that of the fixed ions. Due to this exclusion of the anions, a cation-exchange membrane permits transfer of cations only. Anion-exchange membranes carry positive charges fixed on the polymer matrix. Therefore, they exclude all cations and are permeable only to anions. Although there are a number of inorganic ion-exchange materials, mostly based on zeolites and bentonites, these materials are rather unimportant in ion-exchange membranes compared to polymer materials [24]. The main applications of ion-exchange membranes are in electrodialysis and electrolysis. They are also used as ion-conducting separators in batteries and fuel cells.

2.2.5. *Liquid Membranes*

Liquid membranes are mainly used in combination with the so-called facilitated transport which is based on “carriers” which transport certain components such as metal ions selectively across the liquid membrane interphase. Generally, it is no problem to form a thin fluid film. It is difficult, however, to maintain and control this film and its properties during a mass separation process. To avoid breakup of the film, some type of reinforcement is necessary to support such a weak membrane structure. Two different techniques are used today for the preparation of liquid membranes. In the first, the selective liquid barrier material is stabilized as a thin film by a surfactant in an emulsion-type mixture [43]. In the second, a porous structure is filled with the liquid membrane phase. Both types of membranes are used today on a pilot-plant scale for the selective removal of heavy metal ions or certain organic solvents from industrial waste streams [44]. They have also been used rather effectively for the separation of oxygen and nitrogen.

2.2.6. *Fixed-Carrier Membranes*

Fixed-carrier membranes consist of a homogeneous or porous structure with functional groups which selectively transport certain chemical compounds. Fixed-carrier membranes can have a symmetric or asymmetric structure depending on their application. They are used today, for example, in co- and countercurrent transport and in the separation of alkane – alkene mixtures.

2.3. Fluxes and Driving Forces in Membrane Separation Processes

Separation in membrane processes is the result of differences in the transport rates of different chemical species through the membrane. The transport rate is determined by the driving force or forces acting on the individual components and their mobility and concentration in the membrane. The mobility of a component in the membrane is primarily determined by its size and the physical structure of the membrane material, while the concentration of the solute in the membrane is primarily determined by the chemical compatibility of the permeating component and the membrane material. In membrane separation processes there are different forms of mass transport as indicated in Figure 2, which show schematically the mass transport through membranes separating two homogeneous phases.

In the most general form mass transport through a membrane is shown in Figure 2a, in which the membrane acts as barrier that restricts the flux of various components due to their chemical nature, size, or electrical charge. The transport of a component A is caused by a driving force across the membrane acting on this component. The driving force is a difference in energy of the component A in the two phases separated by the membrane. Thus, the driving force for the transport of a component through a membrane can be expressed in a very general form by the energy gradient, that is, the gradient in its electrochemical potential across the membrane. Gradients in the electrochemical potential of a component are the result of hydrostatic pressure, concentration, electrical potential, or temperature differences between the two phases separated by the membrane [46].

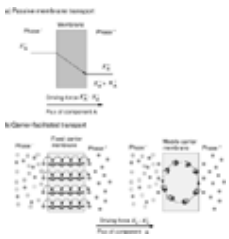

Figure 2. Schematic illustrating various modes of mass transport in membranes

In the example illustrated in Figure 2a it is assumed that the membrane is just a passive physical barrier without specific interactions between the permeating components and the membrane matrix. If the transport of certain components is facilitated due to specific interactions with a carrier component in the membrane matrix the process is referred to as facilitated transport. There are two types of membranes with carrier-facilitated transport properties. The first type is a membrane with the carriers fixed to the membrane matrix where the transported component moves through the membrane by “hopping” from one carrier site in the membrane to the next [45]. The second type is a liquid membrane with the mobile carrier dissolved in a liquid film. The process i illustrated in Figure 2b. Here, the transported components are coupled to the carrier on one side of the membrane, that is, the side facing the feed solution. Then they are transported through the membrane coupled to the carrier by diffusion. This facilitated transport is generally more selective than transport through a passive membrane. However, to achieve a flux of a component an electrochemical potential gradient is also required in a membrane with facilitated transport properties.

There may also be a coupling of the fluxes of individual components. One example of the coupling of fluxes is the transport of bound water with an ion which is transported across a membrane by an electrical potential gradient. More important in membrane processes is the coupling of ion fluxes which results when electrically charged components have different permeability in a membrane. Here, the coupling of fluxes is the result of the electroneutrality condition, which requires that on a macroscopic scale there is no excess of electrical charges. If there is no electrical current, positively or negatively charged components can only be transported across a membrane when the same number of opposite charges carried by counterions are transported in the same direction, or if components with the same charges carried by co-ions are transported in the opposite direction. Thus two forms of flux coupling can be distinguished: cocurrent coupled transport when co- and counterion are transported in the same direction, and countercurrent coupled transport when no counterions permeate the membrane and co-ion fluxes are in opposite direction. The transport of various ions through a membrane by coupling with co- or counterions is illustrated in Figure 3.

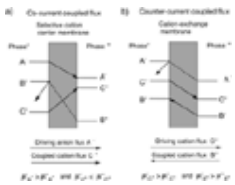

Figure 3. Schematic illustrating a) cocurrent and b) countercurrent coupled transport

In the cocurrent transport shown schematically in Figure 3a component A^- is an anion, and B^+ and C^+ are cations. The concentrations of the components A^- and B^+ in phase′ is higher than in phase″ and the concentration of the component C^+ in phase′ is lower than that in phase″. It is assumed that the membrane is an anion-exchange membrane and thus permeable for the anion A^-. Cation C^+ can permeate the membrane only by means of a specific carrier, and the cation B^+ can not permeate the membrane at all. Due to the concentration differences of the components in the two phases there is a strong driving force for anion A^- to permeate through the membrane from phase′ into phase″. However, because of the electroneutrality requirement it must be accompanied by a cation. The only cation capable of permeating the membrane due to a specific carrier is C^+, which will be transported against its chemical potential gradient from a lower concentration in phase′ to a higher concentration in phase″.

In the countercurrent coupled transport illustrated in Figure 3b a negatively charged cation-exchange membrane separates phase′ and phase″. The concentration of all components in phase′ is assumed to be higher than that in phase″. Thus, there is a driving force for all components to permeate the membrane from phase′ into phase″. Since anion A^- can not permeate the cation-exchange membrane, cation C^+ can only move from phase′ into phase″ if the same number of electrical charges of component B^+ are transported in the opposite direction against its concentration gradient. Both cocurrent coupled transport, which is usually referred to as diffusion dialysis, and countercurrent coupled transport, which is also called Donnan dialysis, are of practical relevance in wastewater treatment.

2.4. Mathematical Description of Mass Transport in Membrane Processes

Mass transfer in membrane processes is determined by the driving forces acting on the individual components in the membrane and by the friction which the components must overcome when moving through the membrane matrix. The driving forces acting on the components of a system can be expressed by gradients in their electrochemical potential and by a hydrostatic pressure gradient causing a bulk volume flow. The friction that must be overcome by the driving force or forces to achieve the transport of the component is generally expressed by a hydrodynamic permeability coefficient, a diffusion coefficient, or, in electrolyte solutions, also by the electrical resistance. To describe the mass transport through a membrane the thermodynamic (i.e., driving forces) and kinetic parameters (i.e., the friction coefficients) must be mathematically related.

In the most general form the mass transport through a membrane can be described by:

$$J_i = -P_i \frac{\mathrm{d}\,X_i}{\mathrm{d}\,z} \tag{2.1}$$

where J is the flux, P a general permeability coefficient, gradient $\mathrm{d}X/\mathrm{d}z$ represents the driving force, and subscript i refers to a component.

The driving force for the transport of a component through the membrane can be expressed by the gradient of its electrochemical potential in the membrane matrix, which is given by:

$$\mathrm{d}\,X_i = \mathrm{d}\,\eta_i = \mathrm{d}\,\mu_i + z_i F \mathrm{d}\,\varphi = \overline{V}_i \mathrm{d}\,p + RT\mathrm{d}\,\ln\,a_i + z_i F \mathrm{d}\,\varphi \tag{2.2}$$

where $\mathrm{d}\,\eta$ and $\mathrm{d}\,\mu$ are the gradients in the electrochemical and the chemical potential; $\mathrm{d}p$, $\mathrm{d}a$, and $\mathrm{d}\,\varphi$ the pressure, the activity, and electrical potential gradients across the membrane; $\overline{V}$ is the partial molar volume; F the Faraday constant; R the gas constant; z the charge number, which is also referred to as the valence of a charged component; and subscript i refers to component i.

In the most general form the flux of a component through a membrane can be described by inserting Equation (2.2) into Equation (2.1):

$$J_i = -P_i \frac{\mathrm{d}}{\mathrm{d}\,z}\left(\overline{V}_i \mathrm{d}\,p + RT\,\ln\,a_i + z_i F \mathrm{d}\,\varphi\right) \tag{2.3}$$

Equation (2.3) describes the flux of a component i through a membrane in reference to any other component in the membrane. If the membrane is used as frame of reference and there is convective flow, for example, through defined membrane pores this convective flow must be taken into account and an additional term must be added in Equation 2.3:

$$J_i = -P_i \frac{\mathrm{d}}{\mathrm{d}\,z}\left(\overline{V}_i \mathrm{d}\,p + RT\ln a_i + z_i F \mathrm{d}\,\varphi\right) + J_v C_i \tag{2.4}$$

where J_v is the convective flux in a membrane pore, and C_i the concentration of component i in the solution or mixture permeating the membrane by convective flow.

Note that the gradient of the chemical or electrochemical potential always refers to the membrane phase. It is usually assumed that the membrane phase is in equilibrium with the adjacent outside phases, that is, the electrochemical potentials

in the membrane η^m and in the adjacent phase η' are identical. This, however, is not true for the concentration of a component, which in the membrane is often very different from that in the adjacent phase. The concentration in the membrane is related to that in the adjacent phase by:

$$C^m = k\,C' \tag{2.5}$$

where *k* is a distribution coefficient of a component between an outside phase and the membrane matrix, which in homogeneous polymer membranes is identical with the solubility coefficient. Equations (2.3) and (2.4) are complete but rather complex descriptions of the transport of a component through a membrane in which three different additive forms of transport can be distinguished, that is, convection, diffusion, and migration.

Convection is a movement of mass due to a mechanical force, generally a hydrostatic pressure difference. Convective transport is generally less important in homogeneous solid materials such as ion-exchange membranes. It is, however, dominant in porous structures and stirred solutions.

Diffusion is a flux of molecular components due to a local gradient in the chemical potential. Gradients in the chemical potential of a component under isothermal conditions can arise from gradients in concentration or pressure. The dominant term for molecules or ions transported through an ion-exchange membrane or in a solution by diffusion is the concentration gradient. For the diffusion of ions it is mandatory that the electroneutrality requirement is fulfilled, that is, cations and anions have to diffuse in the same direction.

Migration is a movement of ions due to an electrical potential gradient. An electrical potential gradient is usually established by applying a voltage difference between two electrodes dipping into an electrolyte solution. The positively charged cations migrate towards the negatively charged cathode and the negatively charged anions migrate towards the positively charged anode, that is, cations and anions are transported in opposite directions.

Convection, diffusion, and migration are traditionally described by phenomenological equations that relate the fluxes to the corresponding driving forces such as the Hagen – Poiseuille law, Fick's law, and Ohm's law.

For instance, a flux of mass caused by a hydrostatic pressure gradient is described by the *Hagen – Poiseuille law*:

$$J_v = -L_p \frac{\mathrm{d}\,p}{\mathrm{d}\,z}[\mathrm{m\,s^{-1}}] \tag{2.6}$$

A flux of individual components caused by a concentration gradient is described by *Fick's law*:

$$J_i = -D_i \frac{\mathrm{d}\,C}{\mathrm{d}\,z}\left[\mathrm{mol\,m^{-2}\,s^{-1}}\right] \tag{2.7}$$

A flux of electrical charges caused by an electrical potential gradient is described by *Ohm's law*:

$$J_e = -\kappa \frac{\mathrm{d}\,\varphi}{\mathrm{d}\,z}[\mathrm{A\,m^{-2}}] \tag{2.8}$$

where *C* is the concentration, *p* the pressure, φ the electrical potential, *D* the diffusion coefficient, L_p the hydrodynamic permeability, κ the electrical conductivity, and the subscripts v, *i*, and e refer to volume, individual components, and electrical charges.

Unfortunately, the simple linear relations expressed in Equations (2.6) to (2.8) are not suited to describe the mass transport in many of the technically relevant membrane processes. First, the conditions in the membrane are generally not ideal, and the concentration of individual components must be replaced by their activities. Furthermore, in many membrane processes mass transport is the result of a combination of driving forces; for example, in reverse osmosis and ultrafiltration the mass flux is caused by pressure and concentration gradients, and in electrodialysis the flux is due to an electrical potential and a concentration gradient. In these processes transport must be described by more complex relations based on Equation (2.3). Only in membrane processes such as microfiltration, where the mass transport is the result of only a single driving force (i.e., a hydrostatic pressure gradient), or in dialysis, where the driving force is a concentration gradient, can Equation (2.3) be replaced by a simpler linear relation between one driving force and the corresponding flux, such as the Hagen – Poiseuille law or Fick's law.

Furthermore, in membrane processes driving forces and fluxes may be interdependent, giving rise to new effects. For instance, a concentration or activity gradient across a membrane may not only result in diffusion of certain components through the membrane. Under certain conditions it can also cause the buildup of a hydrostatic pressure difference across the membrane. This phenomenon is referred to as osmosis and the generated pressure is referred to as osmotic pressure. Similarly, a gradient in hydrostatic pressure may not only lead to volume flux, but may also result in the buildup of a concentration difference across the membrane. This phenomenon is called reverse osmosis. An electrical potential gradient may not only result in an electric current but also in the transport of individual ions and buildup of a hydrostatic pressure. These phenomena are referred to as electrodialysis and electroosmosis.

In practice mass transport in membranes can be described by postulating certain membrane models, such as the pore model and the solution-diffusion model. In these models any kinetic coupling between individual fluxes is neglected. A more comprehensive treatment of the mass transport in multicomponent mixtures is based on the thermodynamics of irreversible processes. Here, the fluxes of individual components are described in a set of phenomenological equations which relate the fluxes of electrical charges, momentum, and individual components to the driving forces and which also take a coupling of different fluxes into account. Another approach to describe mass transport in membranes, suggested by Stefan – Maxwell, also describes the transport of the individual components in a set of phenomenological equations.

Phenomenological description of mass transport. A flux of individual components, heat, and electrical charges can be expressed in the following general equation [47]:

$$J_i = \sum_k L_{ik} X_k (i, k = 1, 2, 3, \ldots, n) \tag{2.9}$$

Here J_i refers to fluxes of individual components, heat and electric charges, X_k refers to the driving forces related to the fluxes, and L_{ik} is a phenomenological coefficient relating flux and driving force.

For multicomponent systems with fluxes of heat, electric charges, and neutral components Equation (2.9) can be written as a matrix. The diagonal coefficients relate the fluxes to the corresponding driving forces, and the cross-coefficients express the coupling of fluxes with indirectly conjugated driving forces [47].

Membrane processes are irreversible. Therefore, the entropy increases during the process. Based on the thermodynamics of irreversible processes it can be shown that the diagonal coefficients are always positive, while the cross-coefficients may be positive or negative and their number is reduced by the so-called Onsager relation [48].

To describe the transport in a multicomponent system such as an electrolyte solution with gradients in hydrostatic pressure, chemical potential, and electrical potential as driving forces, a very large number of phenomenological coefficients must be known which may be a function of temperature, pressure, or composition but which are by definition independent of the driving forces. This limits the practical use of Equation (2.9) substantially. However, under practical conditions many of the phenomenological coupling coefficients can be neglected and the rather complex set of equations can be reduced to more simple relations such as Fick's law or Ohm's law.

Another approach to describe mass transport in membrane processes is based on a relation developed by Maxwell and Stefan. This approach is based on the assumption that at a constant flux of a component *i* the driving forces acting on this component are equal to the sum of the friction forces between the component *i* and other components in the system [49]. If viscous flow (i.e., convection) is excluded the Stefan – Maxwell equation is given by [50]:

$$X_i = \sum_k \frac{f_{ik}}{C_i}\left(v_i - v_k\right) = \sum_k \frac{RT}{Đ_{ik} C_i}\left(v_i - v_k\right) \tag{2.10}$$

where *X* is the driving force, *C* the concentration, *v* the linear velocity, *f* the friction coefficient, Đ the Stefan – Maxwell diffusion coefficient, and subscripts *i* and *k* refer to individual components. The flux of an individual component is always proportional to its velocity relative to the velocity of another component which is used as a frame of reference. Thus, the number of independent fluxes in a given system is equal to the total number of components in the system minus the one that is used as reference. A system containing a membrane with *n* components, including the membrane, has *n*–1 independent fluxes.

The flux of component *i* relative to that of component *k* which serves as a frame of reference is given by:

$$J_i = C_i(v_i - v_k) \tag{2.11}$$

In membrane processes the fluxes through the membranes are of interest. Therefore, the membrane is used as frame of reference and its velocity is assumed to be zero.

Combination of Equations (2.9) to (2.11) provides the relation between the various terms of the phenomenological equation and the Stefan – Maxwell equation:

$$\sum_i L_{ik} = \sum_i \frac{C_i}{f_{ik}} = \sum_i C_i \frac{Đ_{ik}}{RT} \tag{2.12}$$

Equation (2.11), that is, the Stefan–Maxwell relation, provides the same complete description of transport processes through a membrane separating two homogeneous mixtures as does the phenomenological Equation (2.9) based on the thermodynamics of irreversible processes. The only further boundary condition relevant for a system which contains electrically charged components is the electroneutrality requirement and the conservation of electrical charges. For all phenomena observed in membrane systems, such as osmosis, electro-osmosis, diffusion of individual components, viscous flow of a bulk solution, electric current, or the buildup of an osmotic pressure, a streaming and a diffusion potential can be described by applying Equation (2.9) or (2.10). However, as mentioned before, the practical value of these equations is rather limited since they are only applicable close to equilibrium because of the assumed linear relationships between fluxes and driving forces and the large number of different coefficients that must be known when multicomponent systems with viscous flow, diffusion, and migration are to be treated. But even if the phenomenological treatment of mass transport in membrane processes is of little practical use, it indicates how complex the processes in multicomponent systems and heterogeneous membranes really are. It shows that through the coupling of fluxes a transport of components can be obtained without a directly related driving force. The magnitude of the transport of certain components obtained through coupling with other fluxes depends on the coupling coefficient. In many cases the coupling between fluxes can be neglected. In some cases, however, it can be significant. For instance, in electrodialysis the coupling of water with an ion flux can lead to significant water transport through a membrane. A more simplified approach to describe mass transport in electrolyte solution is provided by the so-called Nernst – Planck flux equation which can be derived from Equation (2.9) or (2.10) by introducing the electrochemical potential as driving force for mass transport and by neglecting all other kinetic coupling between individual components.

The electrochemical potential η is given by the chemical potential μ and the electrical potential φ:

$$\eta_i = \mu_i + z_i F \varphi = \mu_i^o + \overline{V}_i p + RT \ln a_i + z_i F \varphi \tag{2.13}$$

where *a* is the activity, *p* the hydrostatic pressure, $\overline{V}$ the partial molar volume, *z* the valence, *F* the Faraday constant, *R* the gas constant, *T* the absolute temperature, subscript *i* refers to a component, and superscript o to a standard state.

Introducing the activity coefficient and the ion concentration for the ion activity into Equation (2.13) and differentiation in transport direction *z*, that is, perpendicular to the membrane surface, leads to:

$$\frac{\mathrm{d}\,\eta_i}{\mathrm{d}\,z} = \overline{V}_i \frac{\mathrm{d}\,p}{\mathrm{d}\,z} + RT\frac{\mathrm{d}\,\ln\,(C_i\,\gamma_i)}{\mathrm{d}\,z} + z_i F \frac{\mathrm{d}\,\varphi}{\mathrm{d}\,z} \tag{2.14}$$

where *C* is the concentration, and γ the activity coefficient.

In electrodialysis and other ion-exchange membrane separation processes the pressure difference across the membrane is kept as low as possible to minimize viscous flow. Therefore, the term $\overline{V}_i \frac{\mathrm{d}\,p}{\mathrm{d}\,z}$ in Equation 2.14 can generally be neglected and the term $RT\frac{\mathrm{d}\,\ln(C_i\gamma_i)}{\mathrm{d}\,z}$ can be expressed by $\frac{RT}{C_i}\left(1 + C_i\frac{\mathrm{d}\,\ln\gamma_i}{\mathrm{d}\,C_i}\right)\frac{\mathrm{d}\,C_i}{\mathrm{d}\,z}$.

Thus:

$$\frac{\mathrm{d}\,\eta_i}{\mathrm{d}\,z} = \frac{RT}{C_i}\left(1 + C_i\frac{\mathrm{d}\,\ln\,\gamma_i}{\mathrm{d}\,C_i}\right)\frac{\mathrm{d}\,C_i}{\mathrm{d}\,z} + Fz_i\frac{\mathrm{d}\,\varphi}{\mathrm{d}\,z} \tag{2.15}$$

Introducing the gradient of the electrochemical potential of a component in the membrane expressed in Equation (2.15) as driving force X_i into Equations (2.10) and (2.11) gives the anion and cation fluxes of a completely dissociated salt:

$$J_i = -\text{Đ}_i\left[\left(1 + C_i\frac{d\,\ln\,\gamma_i}{dC_i}\right)\frac{dC_i}{dz} + \frac{z_i C_i F}{RT}\frac{d\varphi}{dz}\right] + v_k C_i \tag{2.16}$$

If the solution is assumed to be ideal, that is, the activity of a component is identical to its concentration, the Stefan – Maxwell diffusion coefficient can be replaced by Fick's diffusion coefficient, and Equation (2.16) is simplified to:

$$J_i = -D_i\frac{\mathrm{d}\,C_i}{\mathrm{d}\,z} - D_i\frac{z_i C_i F}{RT}\frac{\mathrm{d}\,\varphi}{\mathrm{d}\,z} + v_k C_i \tag{2.17}$$

Equation (2.17) represents the extended Nernst – Planck flux equation where D_i is Fick's diffusion coefficient of component *i* relative to the water in the membrane.The first term $D_i\frac{\mathrm{d}\,C_i}{\mathrm{d}\,z}$ in Equation (2.17) represents diffusion, the second term $D_i\frac{z_i C_i F}{RT}\frac{\mathrm{d}\,\varphi}{\mathrm{d}\,z}$ migration, and the third term $v_k C_i$ convection.

Fick's diffusion coefficient is related to the Stefan – Maxwell diffusion coefficient and the phenomenological coefficient by:

$$D_i = L_i\frac{RT}{C_i}\left(1 + \frac{\mathrm{d}\,\ln\,\gamma_i}{\mathrm{d}\,\ln\,C_i}\right) \tag{2.18}$$

and

$$D_i = \text{Đ}_i\left(1 + \frac{\mathrm{d}\,\ln\,\gamma_i}{\mathrm{d}\,\ln\,C_i}\right) \tag{2.19}$$

The Nernst – Planck equation becomes identical with the phenomenological equation of the thermodynamics of irreversible processes and the Stefan – Maxwell equation if kinetic coupling between individual ions is neglected and if the activity coefficient is assumed to be unity. The Nernst – Planck flux equation is a very convenient approach to describe the transport of ions in ion-exchange membranes. It must, however, be realized that a series of assumptions has been made which in practical ion-exchange membrane separation processes are not always applicable.

2.5. Membrane Separation Capability

In most technically relevant applications the capability of a membrane to separate different components from each other, that is, the membrane permselectivity, is the most important membrane property. This separation capability can be expressed by the selectivity or by the rejection [51].

The membrane selectivity is given by:

$$S^{\mathrm{P}}_{\mathrm{A,B}} = \frac{P_{\mathrm{A}}}{P_{\mathrm{B}}} \tag{2.20}$$

and the membrane rejection by:

$$R_{\mathrm{A}} = \left(1 - \frac{C'_{\mathrm{A}}}{C''_{\mathrm{A}}}\right) \tag{2.21}$$

where $S^{\mathrm{P}}_{\mathrm{A,B}}$ is the permselectivity of a membrane for the components A and B, and P_A and P_B are their permeability; R_A is the rejection of component A by the membrane; C_A' and C_A'' are the concentrations of component A in the two phases separated by the membrane, for instance, a feed solution and a filtrate when a filtration process is considered.

2.6. Chemical and Electrochemical Equilibrium in Membrane Processes

Equilibrium between two systems such as a membrane and an adjacent solution or two solutions separated by a membrane is obtained when the systems are in the same state, that is, when the Gibbs free energies in both systems are identical. The Gibbs free energy of a system is a function of various state variables such as pressure, electrical potential, and the activity of individual components. Consider a system separated by a membrane which is impermeable to some components and permeable to others. These two systems may well be in equilibrium even if they have differences in pressures, electrical potentials, or the activities of individual components as long as these differences compensate each other and the sum of the state variables (i.e., the Gibbs free energy) in both systems is identical. Examples for two such equilibria are the osmotic equilibrium, in which pressure and activity differences compensate each other, and the Donnan equilibrium which is attained when an electrical potential difference is compensated by an activity difference. For membrane processes the osmotic and Donnan equilibrium are both of importance and will be discussed in more detail.

2.6.1. *Osmotic Equilibrium, Osmotic Pressure, Osmosis, and Reverse Osmosis*

If two aqueous salt solutions of different concentrations are separated by a membrane which is permeable for the solvent (e.g., water) but impermeable for the solute (e.g., salt), transport of water from the more dilute solution to the more concentrated solution is observed. This phenomenon (Fig. 4) is referred to as osmosis.

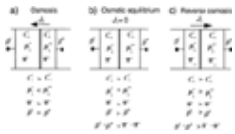

Figure 4. Schematic illustrating the osmotic phenomenon. It shows two phases (′ and ″) separated by a semipermeable membrane. The phases consist of a solvent and a solute, indicated by the subscripts l and s. *C* and μ are concentration and chemical potential, *p* is hydrostatic and π osmotic pressure, and *J* the membrane flux

The figure shows a schematic diagram of a membrane separating two solutions consisting of a solvent and a solute, a salt. The two solutions are indicated by ′ and ″. The membrane is assumed to be permeable to the solvent, but impermeable to the solute. Depending on the concentrations and hydrostatic pressures in the two phases separated by the membrane four different situations can be distinguished:

1. The hydrostatic pressures in the two phases separated by the membrane are equal but the solute concentration in solution′ is higher than that in solution″. In this case the osmotic pressure in solution′ is higher than that in solution″ and flow of solvent occurs from the more dilute solution into the more concentrated solution . This situation is referred to as *osmosis*.
2. The two phases separated by the membrane have different hydrostatic pressures, but the difference in hydrostatic pressure is equal to the difference in the osmotic pressures between the two solutions acting in opposite direction. This situation is referred to as *osmotic equilibrium* and there will be no flow of solvent through the membrane although the concentrations in the two solutions are different.
3. The two phases separated by the membrane have different hydrostatic pressures, but the difference in hydrostatic pressure across the membrane is larger than the osmotic pressure and acts in the opposite direction. Thus, solvent will flow through the membrane from solution′ with the higher solute concentration into solution″ with the lower solute concentration. This phenomenon is referred to as *reverse osmosis*.

The flux of solvent between two homogeneous solutions of different concentrations separated by a semipermeable membrane which is only permeable for the solvent to the more concentrated solution as a function of applied hydrostatic pressure is illustrated in Figure 5. Here, the solvent flux between two solutions of different concentrations through a strictly semipermeable membrane is shown as a function of the hydrostatic pressure applied to the more concentrated solution.

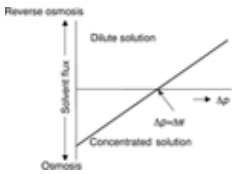

Figure 5. Solvent flux between two solutions of different concentrations through a strictly semipermeable membrane as function of hydrostatic pressure applied to the more concentrated solution

As long as the applied hydrostatic pressure is lower than the osmotic pressure difference between the two solutions, solvent will flow from the more dilute solution into the more concentrated solution by osmosis. When the hydrostatic pressure exceeds the osmotic pressure difference the flow is reversed and solvent will flow from the more concentrated solution to the dilute solution. This process is referred to as reverse osmosis.

The osmotic pressure π of a solution is proportional to the solute concentration C_s, but it can be calculated from the chemical potential of the solvent in a solution and is given by:

$$\pi = \frac{RT}{V_l} \ln a_l \tag{2.22}$$

For dilute solutions the osmotic pressure can be expressed to a first approximation by the concentration of the individual components in the solution [46]:

$$\pi = RT \sum_i g_i C_i \tag{2.23}$$

If the solute is a salt, as is the case in many practical applications, the concentration of the individual ions must be considered when determining the osmotic pressure, that is, the degree of dissociation and the stoichiometric coefficients of

the salt must be considered for determining the osmotic pressure.

The osmotic pressure of an aqueous solution with salts is given by:

$$\pi = RT\sum_i g_i(v_{ic} + v_{ia})C_{is} \tag{2.24}$$

where *g* is the osmotic coefficient and the subscripts s, c, and a respectively refer to a stoichiometric coefficient, cation, and anion.

2.6.2. *Donnan Equilibrium and Donnan Potential*

In discussing the osmotic pressure it was assumed that two solutions separated by a membrane are in chemical equilibrium. If the solution contains charged components (i.e., ions) and the membrane is permeable for at least one ionic component, the membrane will be in equilibrium with the adjacent solution if the electrochemical potential of all ions in the membrane and the solution are equal. Thus, for each ion in equilibrium is:

$$\eta_i^m = \eta_i^s = \mu_i^m + z_i F\varphi^m = \mu_i^s + z_i F\varphi^s \tag{2.25}$$

where μ_i and η_i are the chemical and the electrochemical potential, respectively of component *i*, φ is the electrical potential, *F* the Faraday constant, and superscripts m and s refer to the membrane and to the solution, respectively.

Thus, the electrochemical potential of an ion is composed of two additive terms: the chemical potential, and the electrical potential multiplied by the Faraday constant and the valence of the ion. Introducing the chemical potential μ_i as defined in Equation (2.13) into Equation (2.25) and rearranging gives the electrical potential difference between the membrane and the adjacent solution as:

$$\varphi^m - \varphi^s = \frac{1}{z_i F}\left[RT \ln \frac{a_i^s}{a_i^m} + \overline{V}_i \quad (p^s - p^m)\right] = \varphi_{Don} \tag{2.26}$$

where φ is the electrical potential, *a* the ion activity, $\overline{V}$ the partial molar volume, *z* the valence, *F* the Faraday constant, *p* the pressure, *T* the absolute temperature, *R* the gas constant, subscript *i* refers to individual components, and superscripts m and s refer to the membrane and the electrolyte solution, respectively. The potential difference between the membrane and the solution is referred to as Donnan potential φ_{Don}. [11].

The Donnan potential between an electrolyte solution and an ion-exchange membrane cannot be measured directly. It can, however, be calculated from the ion activities in the solution and the membrane, and by the pressure difference between the membrane phase and the adjacent solution, p^m–p^s, which is referred to as swelling pressure and is identical to the osmotic pressure difference between the solution and the membrane.

Introducing the osmotic pressure into Equation (2.26) gives the Donnan potential as a function of the ion and the water activities in the membrane and the solution:

$$\varphi_{Don} = \frac{1}{z_i F}\left[RT \ln \frac{a_i^s}{a_i^m} - \overline{V}_i \Delta\pi\right] \tag{2.27}$$

The numerical value of the Donnan potential φ_{Don} can be calculated either from the cation or anion activities. For a single salt it is:

$$\varphi_{Don} = \frac{1}{z_c F}\left[RT \ln \frac{a_c^s}{a_c^m} - \overline{V}_c \Delta\pi\right] = \frac{1}{z_a F}\left[RT \ln \frac{a_a^s}{a_a^m} - \overline{V}_a \Delta\pi\right] \tag{2.28}$$

Equation (2.28) gives a general relation for the cation and anion distribution at the interface between a solution and an ion-exchange membrane.

Note that the value of the Donnan potential is negative for an cation-exchange membrane and positive for an anion-exchange membrane in equilibrium with a dilute electrolyte solution.

The Donnan equilibrium describes the electrochemical equilibrium of an ion in an ion-exchange membrane and an adjacent solution and can be calculated for a single electrolyte by rearranging Equation (2.27):

$$\left(\frac{a_a^s}{a_a^m}\right)^{\frac{1}{z_a}}\left(\frac{a_c^m}{a_c^s}\right)^{\frac{1}{z_c}} = e^{\frac{\Delta\pi \overline{V}_s}{RT z_c \nu_c}} \tag{2.29}$$

where *z* is the valence, ν the stoichiometric coefficient of the electrolyte, *F* the Faraday constant, *R* the gas constant, *T* the absolute temperature, $\overline{V}$ the partial molar volume, and *a* the activity; the subscripts a and c refer to anion and cation, and the superscripts s and m to the solution and the membrane, respectively.

The Donnan equilibrium is an important relation in electro-membrane processes since it determines the distribution of co- and counterions between the membrane and the adjacent solution.

[Top of Page]

3. Principles of Membrane Separation Processes

Membrane separation processes can differ greatly with regard to membranes, driving forces, areas of application, and industrial or economic relevance. Table 1 summarizes industrially important membrane separation processes. The driving force in these processes can be a hydrostatic pressure gradient as in micro- and ultrafiltration, reverse osmosis, or gas separation; a concentration gradient as in dialysis; or an electrical potential gradient as in electrodialysis. In some membrane separation processes, for example, pervaporation and membrane distillation, partial pressure differences of individual components can be applied as driving force [51].

Table 1. Technically relevant membrane separation processes, their operating principles, and their applications.

Separation Process	Membrane type used	Applied driving force	Mode of separation	Applications
Microfiltration	symmetric porous membrane, pore radius 0.1–10 mm	hydrostatic pressure, 0.05 – 0.2 MPa	filtration (size exclusion)	water purification, sterilization
Ultrafiltration	asymmetric porous membranes, pore radius 2 – 10 nm	hydrostatic pressure, 0.1 – 0.5 MPa	filtration (size exclusion)	separation of molecular mixtures
Diafiltration	asymmetric porous structure,pore radius 2 – 10 nm	hydrostatic pressure, 0.1 – 0.5 MPa	filtration and dialysis (size exclusion)	purification of molecular mixtures, artificial kidney
Reverse osmosis	asymmetric skin-type solution-diffusion membrane	hydrostatic pressure, 1 – 10 MPa	solution and diffusion	sea- and brackish water desalination
Gas and vapor separation	homogeneous or porous membrane	hydrostatic and vapor pressure	diffusion, Knudsen diffusion	separation of gases, vapors, and isotopes
Pervaporation	homogeneous asymmetric membrane	vapor pressure	solution and diffusion	separation of azeotropic mixtures
Dialysis	symmetric porous membrane	concentration gradient	diffusion	artificial kidney
Electrodialysis	symmetric ion-exchange membrane	electrical potential	Donnan exclusion	water desalination
Donnan dialysis	symmetric ion-exchange membrane	concentration of ions	Donnan exclusion	water softening
Electrodialytic water splitting	bipolar membrane	electrical potential	Donnan exclusion	acid and base production
Membrane distillation	symmetric porous hydrophobic membrane	vapor pressure	diffusion	liquid – solid separation
Membrane contactors	symmetric porous or liquid membrane	chemical potential	diffusion and solution	solvent extraction

Membrane separation processes can conveniently be classified according to the driving forces applied to achieve the transport of certain components through the membrane into pressure-driven, concentration-driven, and electrical-potential-driven processes. Typical pressure-driven membrane processes are microfiltration, ultrafiltration, and reverse osmosis. In all three processes a feed solution is separated into a filtrate that is depleted of particles or molecules and a retentate in which these components are concentrated.

The term *microfiltration* is used when particles with a diameter of 0.1 – 10 µm are separated from a solvent or other low-molecular components. The separation mechanism is based on a sieving effect, and particles are separated according to their dimensions. The membranes used for microfiltration have pores of 0.1 – 10 µm in diameter. The hydrostatic pressure differences used are in the range of 0.05 – 0.2 MPa.

In *ultrafiltration* the components to be retained by the membrane are macromolecules or submicrometer particles. Generally, the osmotic pressure of the feed solution is negligibly small and hydrostatic pressures of 0.1 – 5 MPa are used. Ultrafiltration membranes are mostly asymmetrically structured, with pores in the skin layer having diameters of 2 – 10 nm.

Diafiltration is in principle identical to ultrafiltration. It is used to rinse out low molecular weight components such as salts from a solution with macromolecular components.

In *reverse osmosis* particles, macromolecules, and low molecular mass compounds such as salts and sugars are separated from a solvent, usually water. Feed solutions, therefore, often have a significant osmotic pressure, which must be overcome by the applied hydrostatic pressure. The osmotic pressure of a solution containing low molecular mass solutes can be quite high even at low solute concentrations. Since the membrane flux is a function of the effective pressure, which is given by the hydrostatic pressure difference between feed and filtrate solutions minus the difference in osmotic pressure between these solutions, and the osmotic pressure is relatively high, hydrostatic pressures on the order of 1 – 10 MPa are used in practical applications.

In *dialysis* certain solutes are transferred from a the feed solution to a receiving solution, which is also referred to as stripping solution or dialysate, through a membrane due to a concentration gradient across the membrane. The separation of various compounds is the result of differences in their diffusivity in the membrane matrix.

In *electrodialysis* an electrical potential driving force and anion- and cation-exchange membranes installed in alternating series between electrodes are used to separate charged components such as ions from a solution. The separation is the result of the permselectivity of the ion-exchange membranes, which are permeable only to counterions, that is, cation-exchange membranes are permeable to cations and anion-exchange membrane are permeable to anions.

Electrodialytic water dissociation is a process similar to electrodialysis in which monopolar ion-exchange membranes are used in combination with bipolar membranes which split water into H^+ and OH^- ions to produce acids and bases from the corresponding salt.

In *gas and vapor separation* two different processes and membranes are used. One is similar to reverse osmosis using asymmetric solution-diffusion membranes and a hydrostatic pressure as driving force. The membrane selectivity is determined by the solubility and diffusion of the components in the membrane matrix. In the second process, porous membranes are used through which gases permeate at elevated temperature under a hydrostatic pressure as driving force by Knudsen diffusion. The selectivity is very low and a large number of stages are required to obtain significant separation of the permeating components. Knudsen diffusion separation is used mainly for the separation of radioactive isotopes.

In *pervaporation* volatile organic compounds are removed from a liquid feed solution through a solution-diffusion membrane into the gas phase as a result of a difference in their partial vapor pressure. Both hydrostatic pressure and temperature differences are used as driving force for the transport of the different components through the membrane.

Membrane distillation is similar to conventional distillation. The main difference is that the feed and the product are separated by a porous membrane.

Membrane contactors are used in solvent extraction. Here, the membrane acts as barrier between the feed and the stripping solution.

3.1. Principle of Microfiltration

The principle of microfiltration is depicted in Figure 6. The membrane separating a feed solution from a permeate or filtrate has a symmetric porous structure with an average pore size between 0.1 and 10 µm. The driving force for the mass transport across the membrane is a pressure gradient. Only large particles with diameters in excess of about 0.1 µm are separated by the membrane. The diffusion of particles and the osmotic pressure difference between the feed and filtrate solution are negligibly low and the mass flux through a microfiltration membrane is given to a first approximation by:

$$J_v = \sum_i J_i \overline{V}_i \cong L_v \frac{\Delta p}{\Delta z} \tag{3.1}$$

where J_v is the volume flux across the membrane, the subscripts *v* and *i* refer to volume and components in the solution, $\overline{V}$ is the partial molar volume, *p* the pressure, and L_v a phenomenological coefficient referring to hydrodynamic permeability of the membrane, Δp the pressure difference between the feed and filtrate solution, and Δz the thickness of the membrane.

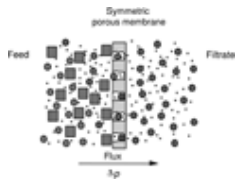

Figure 6. Schematic illustrating the principle of microfiltration

The mass transport in microfiltration membranes takes place by viscous flow through the pores. The solid membrane matrix can be considered as completely impermeable. Therefore, the hydrodynamic permeability can be expressed in terms of the membrane pore size and the porosity and the solution viscosity according to the Hagen – Poiseuille law, which is given by:

$$J_v = \frac{\varepsilon r^2}{8\eta\tau} \frac{\Delta p}{\Delta z} \tag{3.2}$$

where J_v is the flux, ε the membrane porosity, *r* the pore radius, η the viscosity, τ the tortuosity factor, Δp the pressure difference across the membrane, and Δz the membrane thickness.

The tortuosity factor is defined as the ratio of the actual pore length to the thickness of the membrane taking into account that the pore length in general is somewhat longer than the cross section of the membrane. Thus, τ is always greater than 1, and the porosity ε is always less than 1.

The flux through a microfiltration membrane, which is often referred to as filtration rate, is adequately described by Equation (3.2). Another parameter which is of interest for the practical application of microfiltration is the separation characteristic of a membrane. This is generally expressed by the retention or rejection of a particle of a certain size, which is given by:

$$R_i = \left(1 - \frac{C_i^p}{C_i^f}\right) \tag{3.3}$$

where *R* is the rejection coefficient, *C* is the concentration, subscript *i* refers to a given component in the feed and the permeate, and superscripts f and p refer to the feed and permeate or filtrate solutions. *R* is always less than or equal to 1 and a function of the particle and the pore size and pore size distribution.

3.2. Principle of Ultrafiltration

In ultrafiltration the driving force again is a pressure gradient and the mass transport is also dominated by the convective flux through pores. The principle of the process is depicted in Figure 7.

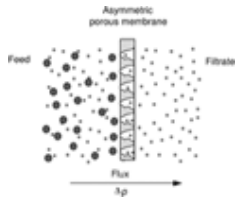

Figure 7. Schematic illustrating the principle of ultrafiltration

As in microfiltration a porous membrane and a hydrostatic pressure difference are used to separate certain components from a feed solution. However, the structure of an ultrafiltration membrane is asymmetric, with the smallest pores on the surface facing the feed solution, and its pores are significantly smaller than those of a microfiltration membrane. Their diameters at the feed side surface of the membrane are between 2 and 10 nm and the components retained by an ultrafiltration membrane have a molecular weight between 5000 and several million Dalton. Since ultrafiltration membranes also retain some relatively low molecular weight solutes, osmotic pressure differences between the feed and the filtrate can be significant and diffusive fluxes of the solutes across the membrane are no longer negligibly low. Therefore, the flux of individual components in ultrafiltration can be expressed as a function of a hydrostatic pressure and a difference in the chemical potential of the different components between the feed and filtrate solutions. The volume flux (i.e., the filtration rate) is given by the sum of the fluxes of the individual components:

$$J_v = \sum_i J_i \overline{V}_i = \sum_i \overline{V}_i L_i \frac{d\mu_i}{dz} + L_v \frac{dp}{dz} = \sum_i \overline{V}_i L_i \frac{d}{dz}(\overline{V}_i p + RT \ln a_i) + L_v \frac{dp}{dz} \tag{3.4}$$

where *J* is the flux, *L* a phenomenological coefficient referring to interactions of the permeating components with the membrane matrix, $\overline{V}_i$ the partial molar volume, μ the chemical potential, *p* the hydrostatic pressure, *z* a directional coordinate, *a* the activity, and the subscripts v and *i* refer to volume flow and individual components. The first term in Equation (3.4) describes the diffusive fluxes of all components in the pores of the membrane, and the second term the volume flow.

In ultrafiltration the total volume flux in a dilute solution can be expressed to a first approximation by the flux of the solvent, that is, $J_v \approx J_w$, and the activity of the solvent in the solution a_w can be expressed by an osmotic pressure. Assuming a linear relation for the pressure and activity gradients across the membrane, integration of Equation (3.4) gives the flux through an ultrafiltration membrane as a function of pressure difference between feed and permeate solution, the hydrodynamic permeability for the viscous flow, the osmotic pressure difference between feed and permeate solution, and the phenomenological coefficient determining the diffusive flow of water through the membrane pores:

$$J_v \cong J_w = \overline{V}_w^2 L_w \frac{\Delta p - \Delta \pi}{\Delta z} + L_v \frac{\Delta p}{\Delta z} \tag{3.5}$$

where J_v and J_w are total volume and solvent fluxes, respectively, Δp and $\Delta \pi$ are the hydrostatic and the osmotic pressure gradients across the membrane, L_v is the hydrodynamic permeability and L_w is the diffusive permeability of the solvent, and Δz is the thickness of the selective barrier of the membrane.

In most practical applications of ultrafiltration the first term of Equation (3.5) can be neglected since $L_w \ll L_v$. The osmotic water transport will effect the membrane volume flux only when the difference in osmotic pressure between the feed and permeate solution is significantly higher than the hydrostatic pressure difference. This is only the case for ultrafiltration of solutions containing a high concentration of retained components of relatively low molecular weight. Since generally only macromolecular components are retained the volume flux (i.e., the filtration rate in ultrafiltration) can be described to a first approximation by:

$$J_v = L_v \frac{\Delta p}{\Delta z} \tag{3.6}$$

For the transport of solutes through an ultrafiltration membrane the diffusive flux can generally not be neglected, and the flux of the solute through the membrane can be expressed as the sum of a diffusive flux due to a chemical potential driving force and a convective flux due to the concentration in the pore solution:

$$J_i = \overline{V}_i L_i \frac{d}{dz}(\overline{V}_i p + RT \ln a_i) + L_v {}^{m}C_i \frac{dp}{dz} \tag{3.7}$$

where ${}^{m}C_i$ is the concentration of the solute in the pore solution.

It can be shown that for hydrostatic pressure differences usually applied in ultrafiltration, $\overline{V}_i p << RT \ln a_i$. If it is assumed that the activity coefficient of the solutes in the membrane is 1, the phenomenological coefficient can be expressed by the Fick's

diffusion coefficient according to Equation (2.19). Introducing these approximations into Equation (3.7) leads to:

$$J_i = J_v {}^{m}C_i - {}^{m}D_i \frac{d\,{}^{m}C_i}{d\,z} \tag{3.8}$$

Integration of Equation (3.8) over the membrane thickness gives:

$$J_i = J_v \frac{{}^{m}C_i^f \exp\frac{J_v \tau \Delta z}{{}^{m}D_i} - {}^{m}C_i^p}{\exp\frac{J_v \tau \Delta z}{{}^{m}D_i} - 1} \tag{3.9}$$

where *J* is the flux, *C* the concentration, *D* the diffusion coefficient, τ the tortuosity factor, and Δz the thickness of the membrane; subscripts *i* and v refer to solute and the volume; and superscripts m, f,and p refer to the membrane, the solution in the membrane at the feed side, and the solution in the membrane at the permeate side.

Equation (3.9) shows that the flux of a component *i* which is partly retained by the membrane in ultrafiltration consists of two terms. The first term represents the viscous flow and is a linear function of the hydrostatic pressure. The second term represents the diffusion in the pore liquid and is an exponential function of the hydrostatic pressure.

In ultrafiltration two limiting cases exist for the transport mechanism. In the first case it is assumed that the hydrostatic pressure and thus the viscous flow is zero, so that the transport of a solute is determined by diffusion only:

$$\lim_{J_v \to 0} J_i = -{}^{m}D_i \frac{{}^{m}C_i^p - {}^{m}C_i^f}{\tau \Delta z} \tag{3.10}$$

The second limiting case is obtained when the pressure and thus the viscous flow is infinitely high. Then the transport of the solute is determined by its concentration in the membrane:

$$\lim_{J_v \to \infty} J_i = J_v {}^{m}C_i^p \tag{3.11}$$

Since the flux of the solute in ultrafiltration is a function of the viscous flux the retention of an ultrafiltration membrane is obviously a function of the applied hydrostatic pressure. It can be described by combining Equations (3.3) and (3.9) and by introducing partition coefficients to relate the concentrations of the solute in the feed and the permeate to that in the membrane:

$$R_i = 1 - \frac{k^f \exp\frac{J_v \tau \Delta z}{{}^{m}D_i}}{k^p - 1 + \exp\frac{J_v \tau \Delta z}{{}^{m}D_i}} \tag{3.12}$$

where

$$k^f = \frac{C_i^f}{{}^{m}C_i^f} \text{ and } k^p = \frac{C_i^p}{{}^{m}C_i^p} \tag{3.13}$$

where k^f and k^p are the partition coefficients which relate the concentrations in the feed and the permeate to that in the membrane.

The effect of the hydrostatic pressure on retention in ultrafiltration becomes more clear when two extreme cases are considered, that is, a very large and a very small hydrostatic pressure difference between feed and permeate solutions. At large pressure differences the viscous flux is high and the retention approaches a maximum value which is determined by the distribution coefficient of the solute between the membrane and the solution.

$$\lim_{J_v \to \infty} R_i = 1 - k^f \tag{3.14}$$

At very low hydrostatic pressure the retention goes to zero.

$$\lim_{J_v \to \infty} R = 0 \tag{3.15}$$

3.3. Principle of Reverse Osmosis

In reverse osmosis the driving force is also a hydrostatic pressure difference. The principle of the process is illustrated in Figure 8, which shows a solution containing a low molecular weight component separated by a membrane from the solvent and a hydrostatic pressure difference applied across the membrane resulting in a flux of solvent from the solution into the pure solvent.

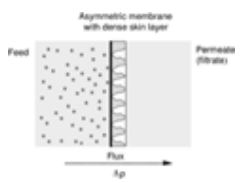

Figure 8. Schematic illustrating the principle of reverse osmosis

The membrane used in reverse osmosis has an asymmetric structure with a dense barrier layer at the side facing the feed solution. It is assumed that in this layer the individual components are transported by diffusion and that viscous flow through

pores or pinholes can be neglected. For describing the transport in reverse osmosis membranes two approaches are described in the literature [52]. One is the so-called solution-diffusion model, which neglects kinetic coupling of different fluxes [23]. The second is based on the phenomenological equation of the thermodynamics of irreversible processes [47]. This model does consider kinetic coupling between different fluxes. Both models describe the transport in reverse osmosis membranes quite accurately, and this indicates that kinetic coupling is not of great importance in reverse osmosis.

3.3.1. *Mass Transport in Reverse Osmosis Described by Phenomenological Equations*

In a reverse osmosis system which consists of a membrane and a two-component feed solution, i.e. a solvent and a solute, two independent fluxes will be obtained when the membrane is used as frame of reference. These fluxes can be described by:

$$J_i = \sum_k L_{ik} X_k \quad (i,\ k = 1, 2, 3,\ \ldots,\ n) \tag{3.16}$$

where J is the flux of the different components, L is the phenomenological coefficient, X the driving force, and subscripts i and k refer to different components.

Assuming a system without viscous flow composed of a solute, a solvent, and a membrane as frame of reference there are two independent fluxes:

$$J_w = L_w \, d\mu_w + L_{ws} \, d\mu_s \tag{3.17}$$

and

$$J_s = L_{sw} \, d\mu_w + L_s \, d\mu_s \tag{3.18}$$

where J is the flux, μ the chemical potential, L the phenomenological coefficient, subscripts w and s refer to the solvent, which is assumed to be water, and the solute, which is assumed to be salt, and subscripts ws and sw to the coupling between the fluxes of water and salt and salt and water.

The coupling coefficients L_{sw} and L_{ws} are identical according to the Onsager relation [48]. They are very difficult to measure in an independent experiment. However, the coupling coefficient can be converted to more easily measurable constants when the phenomenological equations are slightly modified.

Usually in reverse osmosis the total volume flux is measured. It is therefore possible to consider a volume flux J_v and a diffusion related flux J_D which are given by [20]:

$$J_v = J_w \overline{V}_w + J_s \overline{V}_s \tag{3.19}$$

$$J_D = \frac{J_s}{C_s} - \overline{V}_w J_w \tag{3.20}$$

The dissipation function for the above described set of equations is:

$$\Psi = J_w \, d\mu_w + J_s \, d\mu_s = J_w \, d(\overline{V}_w p + RT \ln a_w) + J_s \, d(\overline{V}_s p + RT \ln a_s) \tag{3.21}$$

where Ψ is the dissipation function, J the flux, C, a, μ, and $\overline{V}$ are the concentration, the activity, the chemical potential, and the partial molar volume in the membrane, p is the hydrostatic pressure, and subscripts v, w, s, and D refer to volume, solvent, solute, and diffusion.

Assuming a linear relation for the chemical potential gradients, introducing the osmotic pressure for the activities of water and solute as described in Equations (2.22) and (2.23), and inserting Equations (3.19) and (3.20) into Equation (3.21) leads to a new set of phenomenological equations:

$$\Psi = (J_w \overline{V}_w + J_s \overline{V}_s) \Delta p + \left(\frac{J_s}{C_s} - J_w \overline{V}_w \right) \Delta \pi = J_v \Delta p + J_D \Delta \pi \tag{3.22}$$

The first term in Equation (3.22) describes the total volume flux J_v as function of the applied pressure and the second term describes the diffusion flux of the salt J_D.

The new set of phenomenological equations is:

$$J_v = L_p \Delta p + L_{pD} \Delta \pi \tag{3.23}$$

$$J_D = L_{Dp} \Delta p + L_D \Delta \pi \tag{3.24}$$

By applying different experimental conditions and Equations (3.23) and (3.24) the phenomenological coefficients can be determined from experimental measurements.

Assuming no pressure difference is applied ($\Delta p = 0$), then the diffusive flux is given by:

$$(J_D)_{\Delta p = 0} = L_D \Delta \pi \tag{3.25}$$

and the volume flux is given by:

$$(J_v)_{\Delta p=0} = L_{pD}\,\Delta\pi \tag{3.26}$$

Assuming that there is no osmotic pressure difference ($\Delta\pi$ = 0), then the diffusive flux is given by:

$$(J_D)_{\Delta\pi=0} = L_{Dp}\,\Delta p \tag{3.27}$$

and the volume flux by:

$$(J_v)_{\Delta\pi=0} = L_p\,\Delta p \tag{3.28}$$

The hydrostatic pressure difference at zero volume flux, that is, the osmotic equilibrium, is given by:

$$(\Delta p)_{J_v=0} = -\frac{L_{pD}}{L_p}\Delta\pi = \sigma\,\Delta\pi \tag{3.29}$$

where σ is the so-called Staverman reflection coefficient [53], which is a measure for the permselectivity of a membrane.

Equation (3.29) shows that in osmotic equilibrium the hydrostatic pressure difference between two solutions separated by a membrane is equal to the osmotic pressure difference if $L_p = L_{pD}$ and therefore σ = 1. This is only the case for a strictly semipermeable membrane. Usually, the Staverman reflection coefficient has values between 0 and 1.

If the volume flux is 0 the ratio of salt flux to osmotic pressure difference is given by:

$$\left(\frac{J_s}{\Delta\pi}\right)_{J_v=0} = \frac{C_s L_p L_D - L_{pD}^2}{L_p} = \omega \tag{3.30}$$

where ω is the permeability of the salt.

Combining Equations (3.20), (3.23), (3.24), (3.29), and (3.30) leads to:

$$J_v = L_p(\Delta p - \sigma\,\Delta\pi) \tag{3.31}$$

$$J_s = C_s(1-\sigma)J_v + \omega\,\Delta\pi \tag{3.32}$$

Based on the phenomenological equations the mass transport in reverse osmosis of a solution of a single salt solution can be described by three coefficients:the hydrodynamic permeability of the membrane L_p, the salt permeability, ω and the reflection coefficient σ. All three coefficients can be experimentally determined.

3.3.2. *Description of Mass Transport in Reverse Osmosis by the Solution-Diffusion Model*

In the solution-diffusion model the flux of different components through a membrane is assumed to be caused by diffusion and can be described by the product of its concentration and mobility in the membrane and the chemical potential gradient of the component in the membrane. Furthermore, it is assumed that there is no coupling between fluxes, that the volume flux is approximately identical to the flux of the solvent, and that the solute flux due to the pressure gradient is negligibly small.

Thus, the flux of a component in the membrane is given by:

$$J_i = {}^{m}C_i m_i = L_i\,\mathrm{d}\,\mu_i = L_i(\overline{V}_i\,\mathrm{d}\,p + RT\,\mathrm{d}\ln a_i) \tag{3.33}$$

where J is the flux, ${}^{m}C$ the concentration, m the mobility, L a phenomenological coefficient composed of a concentration and a mobility term, μ the chemical potential, p the hydrostatic pressure, a the activity, $\overline{V}$ the partial molar volume, and subscript i refers to a component.

For i = w it is assumed that $J_v \cong J_w$ and thus $C_w\overline{V}_w = 1$. For i = s it is assumed that d $\mu_s \cong RT\mathrm{d}\ln a_s$ and $\pi \cong \frac{RT\ln a_w}{\overline{V}_w}$. Here the subscripts v, w, and s refer to volume, solvent, and solute.

Integrating over the cross section of the membrane and expressing the activity of the solvent by the osmotic pressure and that of the solute by its concentration gives the volume flux as

$$J_v = L_w\overline{V}_w(\Delta p - \Delta\pi) \tag{3.34}$$

and the solute flux as:

$$J_s = L_s\frac{RT}{C_s}\Delta^{m}C_s \tag{3.35}$$

The concentration of the different components in the membrane is proportional to the concentration in the outside solutions and given by:

$${}^{m}C_i = k_i C_i \tag{3.36}$$

where k_i is the partition coefficient of component i between the membrane and the adjacent solution, and C_i and ${}^{m}C_i$ are the concentrations in the outside solution and in the membrane.

Introducing Equation (3.36) and Fick's diffusion coefficient according to Equation (2.18) into Equations (3.35) and (3.34) gives the volume flux (i.e. the filtration rate) and the flux of the solute as function of the applied hydrostatic pressure, the osmotic pressure, the diffusion coefficient of the components in the membrane, their concentration in the feed and permeate solution, and their partition coefficient between the membrane and the solutions:

$$J_v = \frac{k_w D_w C_w \bar{V}_w}{RT}(\Delta p - \Delta \pi) = \frac{k_w D_w}{RT}(\Delta p - \Delta \pi) \tag{3.37}$$

and

$$J_s = k_s D_s \Delta C_s \tag{3.38}$$

where J_v is the volume flux, J_s the solute flux, D the diffusion coefficient, k the partition coefficient of the components between the membrane and the adjacent solutions, C the concentration, p the hydrostatic pressure and π the osmotic pressure of the solutions, subscripts s and w refer to solute and water, respectively, and the superscript m refers to the membrane. Furthermore it is assumed that $V_w C_w \cong 1$.

In Equations (3.37) and (3.38) all kinetic coupling between fluxes is neglected and activity coefficients in the membrane are assumed to be 1. Nevertheless, Equations (3.37) and (3.38) describe the mass transport in reverse osmosis for a membrane with a high solute rejection capability quite well, as a comparison with Equations (3.31) and (3.32) shows, since the reflection coefficient is very close to 1 for a membrane with a high salt rejection.

Besides the volume flux (filtration rate) and the solute flux, the rejection of the membrane is of interest. The rejection is given by Equation (2.22). In reverse osmosis the concentration in the permeate can be expressed by the solute flux divided by the volume flux, that is, $C_s^p = \frac{J_s}{J_v}$. Membrane rejection in reverse osmosis is given by:

$$R_s = 1 - \frac{C_s^p}{C_s^f} = 1 - \frac{J_s}{J_v C_s^f} = 1 - \frac{k_s{}^m D_s RT \Delta C_s}{k_w C_s^{fm} D_w (\Delta p - \Delta \pi)} \tag{3.39}$$

Equation (3.39) shows that rejection of a component in reverse osmosis not only depends on the diffusion coefficients of water and solute in the membrane and on their partition coefficients between the membrane and the adjacent solutions, but it is also a function of the applied hydrostatic pressure and approaches asymptotically a maximum value. The rejection disappears when the hydrostatic pressure is equal to the osmotic pressure.

3.4. Principle of Gas Separation

The principle of gas separation is illustrated in Figure 9, which shows two different gas mixtures separated by a membrane. The driving force for the gas to permeate the membrane is a pressure gradient. In gas separation both porous and dense membranes are used as selective barriers. In porous membranes the transport of gases is based on Knudsen diffusion, and in dense solid materials the gas transport is based on a solution-diffusion mechanism. Both Knudsen diffusion and solution-diffusion transport can result in selective transport of gases and thus result in separation of gases. However, the extent of the separation, that is, the separation factor, is much higher in solution-diffusion transport than in Knudsen diffusion.

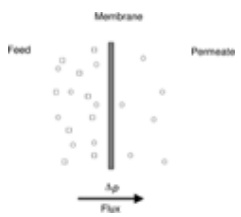

Figure 9. Schematic illustrating the principle of gas separation

3.4.1. *Gas Separation by Knudsen Diffusion*

Knudsen diffusion can be considered as viscous flow in narrow pores, that is, pores with a diameter that is smaller than the mean free path of the diffusing gas molecules. The mean free path is defined as the average distance a gas molecule travels before it collides with another gas molecule. The mean free path of a gas molecule depends on the nature of the gas, the temperature, and the pressure. It is given by [54]:

$$\lambda = \frac{kT}{\pi d_{gas} p \sqrt{2}} \tag{3.40}$$

where λ is the mean free path of a gas molecule, k the Boltzmann constant, d_{gas} the diameter of the gas molecule, and p the hydrostatic pressure.

The difference between viscous flow and Knudsen diffusion is illustrated in Figure 10.

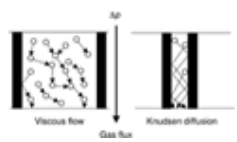

Figure 10. Schematic illustrating viscous flow and Knudsen diffusion

The figure shows two membrane pores of different diameter with a number of gas molecules. In the pore with the larger diameter the gas molecules have more interaction with each other than with the pore wall since the mean free path of the molecules is much shorter than the pore diameter. In the pore with the smaller diameter the gas molecules have more interactions with the pore wall than with each other. In the case of the larger pores the energy loss during transportation is mainly due to the interaction of the molecules with each other. It is expressed in the viscosity of the solution. Therefore, the flux is referred to as viscous flow in which all molecules are transported with the same rate. In the pores with the smaller diameter the energy loss during the transport of the gas molecule is due to their interaction with the pore wall. The process is similar to diffusion in a homogeneous solid phase and is therefore referred to as Knudsen diffusion. Since molecules of different molar mass have different interaction energies when colliding with the pore wall they are transported with different rates. This can be utilized to separate gases which differ in molecular weight. The mean free path of gases at room temperature and atmospheric pressure is on the order of several nanometers. This means that only membranes with a mean pore size below several nanometers can separate gases by a Knudsen diffusion mechanism. Since the free path of a gas molecule is pressure- and temperature-dependent, as Equation (3.40) indicates, membranes with larger pore diameters can be used when the process is carried out at higher temperature and low pressure.

The flux in Knudsen diffusion can be described by the following relation [55]:

$$J_i = \frac{\pi n r^2 D_i^k \Delta p}{R T \tau \Delta z} \tag{3.41}$$

where J is the flux through the membrane, n the number of pores in the membrane, r the pore radius, Δp the pressure difference across the membrane, Δz the thickness of the membrane, τ the tortuosity factor, and D^k the Knudsen diffusion coefficient.

The Knudsen diffusion coefficient is given by:

$$D_i^k = 0.66 r \sqrt{\frac{8RT}{\pi M_w}} \tag{3.42}$$

Equation (3.42) shows that the Knudsen diffusion coefficient of a gas molecule is inversely proportional to the square root of its molecular weight. Thus, the separation of two gases based on Knudsen diffusion is given by the ratio of the square root of the molecular weights.

$$\alpha_{j,k} \propto \sqrt{\frac{M_k}{M_j}} \tag{3.43}$$

Because of the Knudsen diffusion transport mechanism the selectivity that can be achieved with porous membranes is in general rather low when gases differ only marginally in their molecular mass. Nevertheless, porous membranes are used to separate $^{235}UF_6$ from $^{238}UF_6$. Since the selectivity is only about 1.0043, hundreds of stages are required to obtain significant separation.

3.4.2. *Gas Transport by the Solution-Diffusion Mechanism in a Polymer Matrix*

Significantly higher selectivity can be obtained in homogeneous polymer membranes, where the transport mechanism is based on the solution and diffusion of the various components within the membrane phase. The mass transport in a so-called solution-diffusion membrane is illustrated in Figure 11. It consists of three steps:

1. Sorption of the various components from a feed mixture according to their partition coefficients between the gas and polymer phase.
2. Diffusion of the individual components within the membrane phase according to their activity gradients.
3. Desorption of the components from the membrane in the permeate gas phase.

If viscous flow is excluded the driving force for the mass transport of gases in solution-diffusion membranes is the activity gradient of the permeating components within the membrane phase, which can be related to the difference between the chemical potentials of the component in the feed and the permeate.

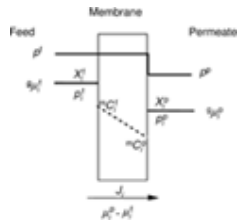

Figure 11. Schematic illustrating the three-step transport mechanism in solution-diffusion membranes (μ is the chemical potential, X the mole fraction, C the concentration, p the partial pressure; the subscript i refers to a component, and the superscripts f, p, g, and m to feed, permeate, gas phase, and membrane)

The flux across the membrane can be described again by the phenomenological equation as a function of the gradient of the chemical potential of the permeating component:

$$J_i = -L_i \frac{\mathrm{d}^m \mu_i}{\mathrm{d}z} \tag{3.44}$$

where J is the flux, L a phenomenological coefficient relating the flux to the corresponding driving force, μ the chemical potential, z the directional coordinate perpendicular to the membrane surface, subscript i refers to a component, and

superscript m to the membrane.

The chemical potential in the membrane at the membrane surface is equal to the potential in the adjacent gas phase. The change of chemical potential of a component as function of pressure and composition is given by:

$$\mathrm{d}\,\mu_i = \mu_i^o + \mathrm{d}\,RT\,\ln\,a_i + \overline{V}_i\,\mathrm{d}\,p \tag{3.45}$$

where *a* is the activity, *p* is the hydrostatic pressure, and $\overline{V}$ is the partial molar volume.

Introducing Equation (3.45) into Equation (3.44) gives the flux of component i in the membrane:

$$J_i = -L_i\left(\mathrm{d}\,RT\,\ln\,{}^{\mathrm{m}}a_i + {}^{\mathrm{m}}\overline{V}_i\,\mathrm{d}\,p\right) \tag{3.46}$$

Since in the membrane the pressure is constant (i.e., d*p* = 0) and $\mathrm{d}\,\ln\,a = \frac{\mathrm{d}\,a}{a}$ Equation (3.46) can be written as:

$$J_i = -L_i RT\frac{\mathrm{d}\,{}^{\mathrm{m}}a_i}{{}^{\mathrm{m}}a_i} \tag{3.47}$$

Integration of Equation (3.47) across the membrane gives the flux as function of the activity of component *i* in the membrane at the interface between the feed gas and the permeate gas:

$$J_i = -L_i\frac{RT}{{}^{\mathrm{m}}\overline{a}_i}\frac{{}^{\mathrm{m}}a_i^{\mathrm{p}} - {}^{\mathrm{m}}a_i^{\mathrm{f}}}{\Delta\,z} \tag{3.48}$$

where ${}^{m}\overline{a}_i$ is the average activity of component *i* in the membrane, and superscripts f and p refer to feed side and product side of the membrane.

Since at the interface between the membrane and the gas the chemical potentials of component *i* in the membrane and the gas are equal (i.e., ${}^{m}\mu_i = {}^{g}\mu_i$), the activity of component *i* in the membrane at the interface can be expressed by its activity in the gas. The chemical potential of component *i* in the membrane and in the gas is obtained by integration of Equation (3.45).

In the membrane the partial molar volume of the component can be considered as constant and equal to the molar volume of the condensed gas:

$${}^{\mathrm{m}}\mu_i = \mu_i^o + RT\,\ln\,{}^{\mathrm{m}}a_{\mathrm{g}} + {}^{\mathrm{m}}\overline{V}_i({}^{\mathrm{m}}p - p^o) \tag{3.49}$$

where the superscript o refers to a standard state.

In the gas the partial molar volume is a function of the pressure and must, in accordance with the ideal gas law, be replaced by $\frac{RT}{p}$.

The chemical potential in the gas is then given by:

$${}^{\mathrm{g}}\mu_i = \mu_i^o + RT\,\ln\,{}^{\mathrm{g}}a_{\mathrm{g}} + RT\,\ln\,\frac{p}{p_i^o} = \mu_i^o + RT\,\ln\,\frac{{}^{\mathrm{g}}a_i p}{p_i^o} \tag{3.50}$$

Since the chemical potentials in the membrane and the gas at the gas membrane interface are equal:

$${}^{\mathrm{g}}\mu_i = {}^{\mathrm{m}}\mu_i = RT\,\ln\,\frac{{}^{\mathrm{g}}a_i p}{p_i^o} = RT\,\ln\,{}^{\mathrm{m}}a_i + {}^{\mathrm{m}}\overline{V}_i({}^{\mathrm{m}}p - p^o) = \frac{{}^{\mathrm{g}}a_i p}{p_i^o} = {}^{\mathrm{m}}a_i\,\exp\,\frac{{}^{\mathrm{m}}\overline{V}_i({}^{\mathrm{m}}p - p^o)}{RT} \tag{3.51}$$

where p_i° is the saturation pressure of the gas.

It can be shown that the exponential term in Equation (3.51) even at high pressure is always close to 1. Therefore, the activity of component *i* at the interface with the gas can be expressed by:

$${}^{\mathrm{m}}a_i^{\mathrm{f}} = \frac{{}^{\mathrm{g}}a_i^{\mathrm{f}}p^{\mathrm{f}}}{p_i^o}\ \text{and}\ {}^{\mathrm{m}}a_i^{\mathrm{p}} = \frac{{}^{\mathrm{g}}a_i^{\mathrm{p}}p^{\mathrm{p}}}{p_i^o} \tag{3.52}$$

where the superscripts *f and p* refer to feed and permeate side of the membrane.

The activity of component *i* in the membrane and in the gas can be related to its concentration in the membrane and its molar fraction in the gas by:

$${}^{\mathrm{m}}a_i = {}^{\mathrm{m}}X_i\,{}^{\mathrm{m}}\gamma_i\ \text{and}\ {}^{\mathrm{g}}a_i = \frac{X_i\varphi_i}{p_i^0} \tag{3.53}$$

where *X* is the molar fraction, γ the activity coefficient, φ the fugacity coefficient, and superscripts m, g, and o refer to the membrane, the gas phase, and the saturation pressure.

Introducing Equations (3.52) and (3.53) into Equation (3.48) leads to:

$$J_i = -L_i \frac{RT}{{}^{m}X_i\,{}^{m}\gamma_i p_i^0} \frac{X_i^p p^p \varphi_i^p - X_i^f p^f \varphi_i^f}{\Delta z} \tag{3.54}$$

where L is a phenomenological coefficient, X the molar fraction in the gas, p the pressure, φ the fugacity coefficient, ${}^{m}X_i$ and ${}^{m}\gamma_i$ are the average molar fraction and the average activity coefficient of component i in the membrane, Δz is the thickness of the membrane, superscripts p, f, and o refer to permeate, feed, and saturation pressure, and subscript i refers to a component.

Introducing the relation between the phenomenological coefficient and the diffusion coefficient as expressed in Equation (2.18), that is, $L_i = \frac{D_i C_i}{RT}$, the relation between the molar fraction and the concentration given by ${}^{m}C_i = X_i \frac{{}^{m}\rho}{M_i}$, and assuming that to a first approximation $\varphi_i^p = \varphi_i^f$ Equation (3.54) becomes:

$$J_i = -D_i \frac{\varphi_i\,{}^{m}\rho}{{}^{m}\gamma_i\, p_i^0 M_i} \frac{X_i^p p^p - X_i^f p^f}{\Delta z} \tag{3.55}$$

Introducing a sorption coefficient given by $k_i = \frac{\varphi_i\,{}^{m}\rho}{{}^{m}\gamma_i\, p_i^0 M_i}$ into Equation (3.55) leads to:

$$J_i = -D_i k_i \frac{X_i^p p^p - X_i^f p^f}{\Delta z} \tag{3.56}$$

where D is the diffusion coefficient in the membrane, k_i the gas sorption coefficient of the membrane, ρ the density, M the molecular weight, and subscript i refers to a component.

The product of diffusion and sorption coefficient is the membrane permeability given by $P_i = D_i k_i$. Introducing the permeability P_i into Equation (3.56) gives the flux J_i of a component through the membrane as a function of its molar fractions in the feed and permeate X_i^f and X_i^p, its fugacity coefficient φ_i, its permeability in the membrane P_i, the membrane thickness Δz, and the pressures in the feed and permeate p^f and p^p.

$$J_i = -P_i \frac{X_i^p p^p - X_i^f p^f}{\Delta z} \tag{3.57}$$

In deriving Equation (3.57) several assumptions have been made which are generally valid for permanent gases. However, the assumption that the diffusion and the solubility coefficient are independent of concentration and pressure is not valid for easily condensable vapors. In permeation of organic vapors, the diffusion coefficient may vary by several orders of magnitude with the concentration of the permeating component in the membrane.

For the relationship between the diffusion coefficient and the concentration of the diffusing component different equations accounting for dual sorption or free volume and structural changes of the polymer are discussed in the literature. For gas and especially vapor permeation it is generally not possible to predict the mass transport behavior of various components in a mixture from single-component measurements.

For a practical application the separation efficiency of the membrane is a crucial parameter. In gas permeation the separation efficiency of a membrane is expressed by its selectivity and/or its separation factor. The selectivity of a membrane for various components of a mixture is defined by the ratio of the permeabilities. For a binary mixture of components k and j, the selectivity is:

$$S_{j,k}^P = \frac{P_j}{P_k} \tag{3.58}$$

where $S^P_{j,k}$ is the permeation selectivity of a membrane for the components j and k, and P_j and P_k are their permeabilities.

The permeation selectivity can be split into two terms, namely, the selectivity due to different diffusion coefficients of the components of the gas mixture and that due to their solubility in the membrane:

$$S_{j,k}^P = S_{j,k}^D S_{j,k}^k \tag{3.59}$$

where $S_{j,k}^D$ and $S_{j,k}^k$ are the diffusivity and solubility selectivity of a membrane for components j and k.

The selectivity is a useful parameter to characterize a membrane. For the design of a membrane plant, however, the separation factor is more useful. For a binary mixture the separation factor is defined by:

$$\alpha_{j,k} = \frac{X_j^p}{X_j^f} \frac{X_k^f}{X_k^p} \tag{3.60}$$

where α is the separation factor, X the mole fraction, subscripts k and j refer to the two components, and superscripts f and p refer to feed and the permeate.

The separation factor is defined to be always greater than 1. It is related to the membrane selectivity by:

$$\alpha_{j,k} = S_{j,k} \frac{\frac{X_j^{\mathrm{p}} p^{\mathrm{p}}}{p^{\mathrm{f}}} - X_j^{\mathrm{f}}}{\frac{X_k^{\mathrm{p}} p^{\mathrm{p}}}{p^{\mathrm{f}}} - X_k^{\mathrm{f}}} \frac{X_k^{\mathrm{f}}}{X_j^{\mathrm{f}}} \tag{3.61}$$

where *S* is the selectivity for a binary mixture, *p* the pressure, *X* the molar fraction, superscripts f and p refer to feed and the permeate, and subscripts *k* and *j* to the components in the mixture. The separation factor is always smaller than the selectivity and when the pressure ratio of feed to permeate becomes infinitely high:

$$\lim_{\frac{p^{\mathrm{p}}}{p^{\mathrm{f}}} \to 0} a_{j,k} = S_{j,k} \tag{3.62}$$

In vapor separation, the factor α is normally a function of the feed concentration.

3.5. Principle of Pervaporation

Pervaporation is a membrane process in which the permeation of certain components through a membrane from a liquid feed mixture into a vapor phase is combined with the evaporation of these components. The principle of the process is illustrated in Figure 12.

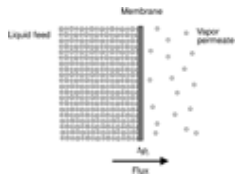

Figure 12. Schematic illustrating the principle of pervaporation showing an asymmetric membrane with a dense solution-diffusion barrier layer

The membranes and transport mechanism of a component through these membranes in pervaporation are the same as in gas separation. The driving force for transport is the chemical potential gradient of the permeating components in the membrane. The chemical potential gradient in the membrane can be related to the partial vapor pressures in the liquid and vapor phases. To establish a partial pressure difference of a component in the liquid and the vapor phase, generally, a vacuum is applied at the permeate side of the membrane. But differences in the partial pressure of a component can also be established by using a sweeping gas on the permeate side or by a temperature difference between the liquid feed and the vapor.

Mass transport in a pervaporation membrane can be described by the same mathematical relations as gas transport with the exception that the chemical potential in the membrane on the feed side of the membrane is not expressed by the molar fraction, the hydrostatic pressure, and the fugacity coefficient, but by the molar fraction and the activity coefficient in the liquid phase, and the saturation pressure of the component.

Thus, mass transport in pervaporation is given by:

$$J_i = -D_i k_i \frac{X_i^{\mathrm{p}} p^{\mathrm{p}} - X_i^{\mathrm{l}} \gamma_i^{\mathrm{l}} p_i^{\mathrm{o}}}{\Delta z} \tag{3.63}$$

where D_i is the diffusion coefficient of component *i* in the membrane, k_i the sorption coefficient of component *i* into the membrane matrix, *X* the molar fraction, *p* the pressure, γ the activity coefficient, subscripts *j* and *k* refer to components of the mixture, and the superscripts o, l, and *p* to saturation pressure, feed, and permeate, respectively.

Equation (3.63) describes the permeation flux of a component *i* as a function of the feed and permeate mixture composition and its diffusivity and solubility in the membrane, which can easily be determined by independent measurements. The practical application of Equation (3.63), however, is rather limited since the diffusion coefficient and the distribution coefficient are not constant but depend strongly on the concentration of the permeating component in the membrane. This dependency must be determined experimentally and may be different for different membrane materials and vapors.

The product of diffusion and distribution coefficient is referred to as permeability coefficient, and the ratio of the permeability coefficients of two components determine the membrane selectivity for the two components.

$$D_i k_i = P_i \tag{3.64}$$

$$S_{j,k} = \frac{P_j}{P_k} \tag{3.65}$$

where *P* is the permeability of the membrane, *S* the membrane selectivity, and *j* and *k* refer to the components of the mixture.

The separation factor in pervaporation of a two-component mixture is given by:

$$\alpha_{j,k} = \frac{X_j^{\mathrm{p}} X_k^{\mathrm{p}}}{X_j^{\mathrm{l}} X_k^{\mathrm{l}}} = S_{j,k} \frac{X_j^{\mathrm{p}} \varphi_j^{\mathrm{p}} - X_j^{\mathrm{l}} \gamma_j^{\mathrm{l}} p_j^{\mathrm{o}}}{X_k^{\mathrm{p}} \varphi_k^{\mathrm{p}} - X_k^{\mathrm{l}} \gamma_k^{\mathrm{l}} p_k^{\mathrm{o}}} \frac{X_k^{\mathrm{l}}}{X_j^{\mathrm{l}}} \tag{3.66}$$

where α is the separation factor, *S* the membrane selectivity, *X* the molar fraction, *p* the pressure, and γ the activity coefficient, subscripts *j* and *k* refer to components of the mixture, and superscripts o, l, and p to saturation pressure, feed, and permeate, respectively.

In pervaporation the partial pressure of the components on the permeate side is kept as low as possible by either using a sweeping gas or more commonly by applying a vacuum, i.e., $X_j^{\mathrm{p}} p^{\mathrm{p}} << X_j^{\mathrm{l}} \gamma_j^{\mathrm{l}} X_j^{\mathrm{0}}$ and $X_k^{\mathrm{p}} p^{\mathrm{p}} << X_k^{\mathrm{l}} \gamma_k^{\mathrm{l}} X_k^{\mathrm{0}}$. For a pervaporation experiment using a vacuum the separation factor is to a first approximation:

$$\lim_{p^{\mathrm{p}} \to 0} \alpha_{j,k} = S_{j,k} \frac{\gamma_j^{\mathrm{l}} p_j^{\mathrm{0}}}{\gamma_k^{\mathrm{l}} p_k^{\mathrm{0}}} \tag{3.67}$$

Thus, the separation factor consists of two terms: $S_{j,k}$, which represents the membrane selectivity, and the thermodynamic liquid/vapor equilibrium $\frac{\gamma_j^{\mathrm{l}} p_j^{\mathrm{0}}}{\gamma_k^{\mathrm{l}} p_k^{\mathrm{0}}}$, which represents the separation that would be achieved by distillation. In pervaporation the membrane selectivity can increase or decrease the distillation selectivity and eventually push the overall separation factor in the opposite direction.

3.6. Principle of Dialysis

Dialysis is carried out under isobaric conditions, and there is no temperature or electrical potential gradient. The principle of the process is depicted in Figure 13.

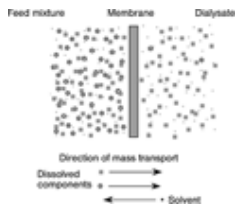

Figure 13. Schematic illustrating the principle of dialysis with two neutral components of different size

The schematic drawing shows a liquid solution containing a solvent and molecular components of different molecular weight or size, referred to as feed solution separated by a membrane from the solvent referred to as dialysate or “stripping” solution. Due to the driving force of a concentration difference the components will diffuse through the membrane from the feed solution to the dialysate. Their fluxes through the membrane and thus their separation is determined by their diffusion rate in the membrane matrix.

There are different forms of dialysis depending on the constituents of the feed solution and the structure of the membrane. The simplest form is the dialysis of neutral components.

3.6.1. *Dialytic Mass Transport of Neutral Components*

The mass flux in dialysis of a mixture, which contains neutral components only, can be expressed by the phenomenological equation, which can then be related to Fick's law by several approximations.

For a membrane and a feed solution with uncharged components the flux of a component at constant temperature and pressure is given by:

$$J_i = -L_i \frac{\mathrm{d}\,\mu_i}{\mathrm{d}\,z} = -L_i RT \frac{\mathrm{d}\ \ln\ a_i}{\mathrm{d}\,z} = -L_i RT \mathrm{d}\ \ln\ C_i \gamma_i \tag{3.68}$$

where J is the flux, L a phenomenological coefficient, μ the chemical potential, a the activity, C the concentration, γ the activity coefficient, z the directional coordinate perpendicular to the membrane, and the subscript i refers to a component.

Assuming that the activity coefficient is 1 and replacing the phenomenological coefficient by the diffusion coefficient according to Equation (2.18) leads to:

$$J_i = L_i \frac{RT}{C_i} \frac{\mathrm{d}\,C_i}{\mathrm{d}\,z} = D_i \frac{\mathrm{d}\,C_i}{\mathrm{d}\,z} \tag{3.69}$$

Equations (3.68) and (3.69) indicate that the mass transport in dialysis with feed solutions containing neutral components can be described adequately by Fick's law if there is no coupling of fluxes and if the activity coefficient of the components in the membrane is unity, which is the case in many practical applications.

3.6.2. *Dialytic Mass Transport of Electrolytes in a Membrane without Fixed Ions*

If feed solutions with charged components (i.e., electrolyte solutions) are treated by dialysis, the electroneutrality requirement must be satisfied. For a system consisting of a membrane and solutions with various ions the electroneutrality condition requires:

$$\sum_{\mathrm{a}} |\, z_{\mathrm{a}} \,|\, C_{\mathrm{a}} = \sum_{\mathrm{c}} z_{\mathrm{c}} C_{\mathrm{c}} \tag{3.70}$$

where z is the valance, C the concentration, and subscripts a and c refer to anion and cation, respectively.

If the membrane contains fixed ions they must be included in the electroneutrality requirement. The fixed charges are negative in a cation-exchange membrane and positive in an anion-exchange membrane. If there is no electrical current the conservation of electrical charge requires that cations and anions move in the same direction. Hence:

$$\sum_{\mathrm{a}} |\, z_{\mathrm{a}} \,|\, J_{\mathrm{a}} = \sum_{\mathrm{c}} z_{\mathrm{c}} J_{\mathrm{c}} \tag{3.71}$$

The concentration of the electrolyte (e.g., a single salt) is related to that of the ions by:

$$C_a = \alpha \nu_a C_s \text{ and } C_c = \alpha \nu_c C_s \tag{3.72}$$

where α is the degree of dissociation of the electrolyte, ν a stoichiometric coefficient, and subscript s refers to the electrolyte.

The stoichiometric coefficient relates the number of ions to that of the salt molecules. For a completely dissociated monovalent salt such as NaCl α and ν are 1. For a salt such as $CaCl_2$ ν_c is 1 and ν_a is 2.

In an electrolyte solution the fluxes of the ions are the result of an electrochemical potential gradient driving force. The electrochemical potential of ions is related to that of the electrolyte by [56]:

$$\eta_s = \nu_a \eta_a + \nu_c \eta_c \tag{3.73}$$

where η is the electrochemical potential.

Since in dialysis no external potential and no hydrostatic pressure difference is applied the flux of the ions can be described as a function of the diffusion coefficient and the concentration gradient by Equation (3.69) if the activity coefficients are assumed to be 1. However, the fluxes of the ions are related to each other by the electroneutrality requirement and the conservation of charges, with the consequence that in diffusion of an electrolyte which consists of a cation and an anion with significantly different diffusivities the faster moving ion will cause a minute charge imbalance and thus an electrical potential gradient. This potential gradient is referred to as diffusion potential, and although it may be very small it will slow down the faster diffusing ion and speed up the slower ion so that both ions diffuse with the same velocity.

The flux of the completely dissociated monovalent electrolyte is related to that of the ions by:

$$J_s = J_a = J_c = -D_s \frac{dC_s}{dz} = -D_a |z_a| \frac{dC_a}{dz} = -D_c z_c \frac{dC_c}{dz} \tag{3.74}$$

where *J* is the flux, *C* the concentration, *D* the diffusion coefficient, *z* the valence or the directional coordinate, and subscripts s, a, and c refer to salt, anion, and cation, respectively.

Rearranging Equation (3.74) under consideration of Equation (3.73) gives the diffusion coefficient of the salt as a function of the diffusion coefficients of the cations and anions:

$$D_s = \frac{D_a D_c (|z_a| + z_c)}{D_a |z_a| + D_c z_c} \tag{3.75}$$

Combination of Equations (3.74) and (3.75) gives the salt and ion fluxes for a completely dissociated monovalent salt at constant temperature and pressure in an ideal solution, that is, the activity coefficients are 1:

$$J_s = \frac{D_a D_c (|z_a| + z_c)}{D_a |z_a| + D_c z_c} \frac{dC_s}{dz} \tag{3.76}$$

where *J* is the flux, *C* the concentration, *D* the diffusion coefficient; *z* the valence or the directional coordinate, and subscripts s, a, and c refer to salt, anion, and cation, respectively.

Equation (3.76) describes the flux of an electrolyte as a function of the diffusion coefficients of the individual ions. This is the consequence of the electroneutrality requirement for the transport of ions which have different diffusivities. Because the faster moving ion will cause the buildup of a minute electrical potential gradient which will slow down the faster ion and speed up the slower ion, at steady state both ions move with the same velocity. Therefore, an average diffusion coefficient for each ion can be defined to describe the fluxes of the different ions, which are coupled by the conservation of charges and the electroneutrality condition. The average diffusion coefficient $\overline{D}$ is given by:

$$\overline{D}_c = \overline{D}_a = D_s = \left(\frac{D_a D_c |z_a| + z_c}{D_a |z_a| + z_c D_c} \right) \tag{3.77}$$

Equations (3.76) and (3.77) respectively describe the transport of a single electrolyte under the driving force of a concentration or activity gradient through a membrane that does not carry any fixed charges. In this case all ions of the electrolyte are transported in the same direction even if their diffusion coefficients are quite different. The situation, however, becomes more complex if a feed solution contains more than one electrolyte and if the membrane carries fixed charges. In this case certain ions may be transported against their concentration gradient.

3.6.3. *Dialysis of Electrolytes in Ion-Exchange Membranes*

The fluxes of ions through ion-exchange membranes in dialysis are even more complex and may result in transport of certain ions against their concentration gradient, as illustrated in Figure 14. Figure 14a shows countercurrent coupled transport of H^+ and Na^+ ions through a cation-exchange membrane. This process is referred to as Donnan dialysis. Figure 14b depicts cocurrent transport of H^+ ions and Cl^- ions through an anion-exchange membrane.

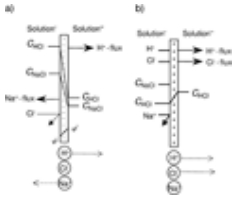

Figure 14. Schematic illustrating the fluxes across an ion-exchange membrane which separates two electrolyte solutions of different concentrations containing HCl and NaCl. a) Shows the fluxes through a cation-exchange membrane and represents countercurrent coupled flux, i.e., Donnan dialysis, and b) shows the fluxes through an anion-exchange membrane and represents cocurrent coupled transport, i.e., diffusion dialysis

In Donnan dialysis with a strictly permselective ion-exchange membrane the flux of the co-ion (i.e., in Figure 14a the Cl^- ions) is 0 and the fluxes of the counterions (i.e., the Na^+ and H^+ ions) are equal but in opposite directions. Thus, there is no net transport of electrical charges across the membrane, that is, there is no electrical current and the conservation of charges is fulfilled. For the system illustrated in Figure 14a the transport of a counterion (i.e., the Na^+ ion) is the result of a flux of another counterion (i.e., the H^+ ion) in the opposite direction under the driving force of a concentration gradient. Since no co-ions (i.e., Cl^- ions) can permeate the cation-exchange membrane, the flux of the H^+ ions must be identical to that of the Na^+ ions. The fluxes are given by:

$$J_{H^+} = -2\left(\frac{L_{H^+}L_{Na^+}}{L_{H^+}+L_{Na^+}}\right)\frac{RT\,\mathrm{d}\ln a_{H^+}}{\mathrm{d}z} = J_{Na^+} \tag{3.78}$$

The transport of Na^+ ions due to the H^+ ion flux results in a buildup of an activity gradient of the Na^+ ions, which causes a backtransport. When the Na^+ ion flux caused by the activity gradient is identical to that caused by the H^+ ion flux equilibrium is reached and there will be no further ion transport through the membrane. This equilibrium, which is referred to as Donnan equilibrium, is given by:

$$\mathrm{d}\ln a_{H^+} = \mathrm{d}\ln a_{Na^+} \tag{3.79}$$

The Donnan equilibrium between two solutions with different Na^+ ion and H^+ ion concentrations separated by a permselective cation-exchange membrane is:

$$\frac{a'_{Na^+}}{a''_{Na^+}} = \frac{a'_{H^+}}{a''_{H^+}} \tag{3.80}$$

where ′ and ″ refer to the two solutions separated by the membrane.

Donnan dialysis plays an important role in water softening, and Equation (3.80) determines the minimum and maximum concentrations of a counterion which can be achieved by a given flux of another counterion.

Cocurrent coupled transport is illustrated in Figure 14b, which shows a solution containing a mixture of a salt (NaCl) and an acid (HCl), separated by an anion-exchange membrane from a solution containing no salt. The Cl^- ion concentration in the solution containing the salt – acid mixture is assumed to be much higher than in the solution containing only the acid. Due to a concentration gradient the Cl^- ions permeate the anion-exchange membrane towards the solution containing only the acid. Because of the electroneutrality requirement an anion can permeate the membrane only when accompanied by a cation. The anion-exchange membrane, however, is impermeable to cations such as Na^+ ions but it is to some extent permeable to H^+ ions because protons are transported in aqueous solution and also in highly swollen ion-exchange membranes by a different transport mechanism. The Cl^- ion concentration gradient across the membrane acts as driving force for the transport of H^+ ions, which are transported against their concentration gradient. The minimum concentration of the acid in the salt solution and the maximum concentration of the acid that can be achieved is given when the activity gradients of the H^+ ion and the Cl^- ion across the membrane are identical. The cocurrent transport of an anion with a H^+ ion is of practical importance for the recovery of acids from spent pickling solutions containing the acid in a mixture with salts.

3.7. Principle of Electrodialysis

The principle of the electrodialysis is illustrated in Figure 15.

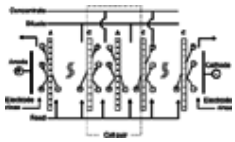

Figure 15. Schematic illustrating the principle of electrodialysis

In electrodialysis cation-exchange membranes carrying fixed negative electric charges and anion-exchange membranes carrying positive fixed charges are placed in alternating series between two electrodes. Cation- and anion-exchange membranes form individual cells filled with an electrolyte-containing feed solution. If an electric potential between the electrodes is established the cations migrate towards the negatively charged cathode while the anions migrate towards the positively charged anode. Cations can easily permeate a cation-exchange membrane but are retained by an anion-exchange membrane, while anions permeate the anion-exchange membrane but are retained by the cation-exchange membrane. Thus, ions are accumulated in alternating cells to form the concentrate solution while the other cells are depleted and form the diluate solution.

In electrodialysis mass transport is the result of an electrochemical potential gradient across the membrane, that is, a gradient in the chemical and electrical potential. An electrical potential driving force acts on charged components only and in such a way that negatively charged components (anions) migrate in the direction towards the positively charged anode and the positively charged cations migrate towards the negatively charged cathode, that is, cations and anions move in opposite directions.

To describe the mass transport of a salt in an aqueous solution through a membrane which is used as frame of reference three independent fluxes must be considered: the flux of the cations, the flux of anions, and the flux of the solvent. The transport of ions result from an electrochemical potential difference, and the transport of the solvent through the membrane from osmotic effects and coupling with the fluxes of the ions. Under certain conditions the solvent flux can be significant and must be taken into account.

Ion-exchange membranes carry positive or negative electrical charges fixed to a solid matrix and are therefore permselective as far as the transport of ions is concerned, that is, the cation-exchange membranes are preferentially permeable to cations and the anion-exchange membranes are preferentially permeable to anions.

Mass transport in electrodialysis can be described by a phenomenological equation giving the flux of the individual components by:

$$J_i = \sum_k L_{ik} \frac{\mathrm{d}\eta_k}{\mathrm{d}z} = \sum_k L_{ik} \left(RT \frac{\mathrm{d} \ln a_k}{\mathrm{d}z} + z_i F \frac{\mathrm{d}\varphi}{\mathrm{d}z} \right) \tag{3.81}$$

Assuming an ideal solution, that is, one in which the activity of a component is identical to its concentration and without kinetic coupling between individual components, and expressing the phenomenological coefficient by the Fick's diffusion coefficient, Equation (3.81) becomes identical to the Nernst – Planck flux equation, which is given by [57]:

$$\mathrm{J_i} = -D_i \left(\frac{\mathrm{d}C_i}{\mathrm{d}z} + \frac{\mathrm{z_i} F C_i}{RT} \frac{\mathrm{d}\varphi}{\mathrm{d}z} \right) \tag{3.82}$$

where D_i is the diffusion coefficient of component *i* with reference to the membrane.

The first term in Equation (3.82) $D_i \frac{\mathrm{d}C_i}{\mathrm{d}z}$ represents the diffusion, and the second term $D_i \frac{z_i C_i F}{RT} \frac{\mathrm{d}\varphi}{\mathrm{d}z}$ the migration.

Thus, the Nernst – Planck equation is an approximation of the more general phenomenological equation.

In ion-exchange membrane separation processes such as electrodialysis the flux of the cations through the cation-exchange membrane and the flux of the anions through the anion-exchange membrane are interdependent because of the electroneutrality requirement in the solution between the membranes, as can easily be seen when considering the electrodialysis of a solution of a monovalent single electrolyte such as NaCl. If strictly permselective membranes are assumed the fluxes of the cations through the cation-exchange membrane and of the anions through the anion-exchange membrane are equal. Thus, in a solution containing a monovalent, completely dissociated electrolyte such as NaCl and strictly permselective membranes it is [22, 56]:

$$^{\mathrm{cm}}J_{\mathrm{Na}^+} = {}^{\mathrm{am}}J_{\mathrm{Cl}^-} = J_{\mathrm{s}} = -2 \left(\frac{^{\mathrm{cm}}D_{\mathrm{Na}^+} \, {}^{\mathrm{am}}D_{\mathrm{Cl}^-}}{^{\mathrm{cm}}D_{\mathrm{Na}^+} + {}^{\mathrm{am}}D_{\mathrm{Cl}^-}} \right) \left(\frac{RT}{^{\mathrm{im}}C_i} \frac{\mathrm{d}\,^{\mathrm{im}}C_i}{\mathrm{d}z} + F \frac{\mathrm{d}\varphi}{\mathrm{d}z} \right) \tag{3.83}$$

where $^{\mathrm{cm}}J_{\mathrm{Na+}}$ and $^{\mathrm{am}}J_{\mathrm{Cl-}}$ are the fluxes of Na^+ ions through the cation-exchange membrane and of Cl^- ions through the anion-exchange membrane, $^{\mathrm{cm}}D_{\mathrm{Na+}}$ and $^{\mathrm{am}}J_{\mathrm{Cl-}}$ are the diffusion coefficients for the transport of Na^+ and Cl^- ions through the corresponding membrane, subscript *i* refers to cation or anion, and superscript *i*m to cation- or anion-exchange membrane.

3.7.1. *Electrical Current and Ion Fluxes*

In electrodialysis it is assumed that the total current through the membrane is carried by ions only:

$$i = \frac{I}{A} = F \sum_i |z_i| J_i \tag{3.84}$$

where *i* is the current density, *I* the current, *A* the membrane surface, *F* the Faraday constant, *J* the flux, *z* the valence, and subscript *i* refers to cations and anions.

The current density *i* can be related to the specific conductivity κ by:

$$i = \kappa \frac{\mathrm{d}\varphi}{\mathrm{d}z} \tag{3.85}$$

where κ is the specific conductivity, φ the electrical potential, and *z* the directional coordinate.

The specific conductivity κ can be expressed by the specific electrical resistance, the equivalent conductivity, the ion migration velocity, or the ion diffusivity:

$$\kappa = \frac{1}{\rho} = \sum_i |z_i| C_i \lambda_i = F \sum_i |z_i| C_i u_i = F^2 \sum_i |z_i| C_i \frac{D_i}{RT} \tag{3.86}$$

where ρ is the specific resistance, *C* the concentration, *F* the Faraday constant, λ the equivalent conductivity, *u* the ion migration velocity, *D* the ion diffusion coefficient, *R* the gas constant, *T* the absolute temperature, and subscript *i* refers to anions and cations.

Introducing Equations (3.85) and (3.86) into Equation (3.84) and rearranging leads to:

$$i = F \sum_i |z_i| J_i = F^2 \sum_i z_i^2 \frac{C_i D_i}{RT} \left(\frac{RT}{z_i C_i F} \frac{\mathrm{d}C_i}{\mathrm{d}z} + \frac{\mathrm{d}\varphi}{\mathrm{d}z} \right) \tag{3.87}$$

where *i* is the current density, *C* the concentration, *F* the Faraday constant, φ the electrical potential, *z* the valence, *D* the ion diffusivity, *R* the gas constant, *T* the absolute temperature, and subscript *i* refers to anions and cations.

The term $\frac{RT}{z_i C_i F} \frac{\mathrm{d}C_i}{\mathrm{d}z}$ has the dimensions of an electrical potential gradient and represents the concentration potential which is established between two electrolyte solutions of different concentrations. In electrodialysis this potential is established between the concentrate and the diluate solution. It represents an electromotive force which has to be overcome by the applied electrical potential.

3.7.2. *Transport Number and Membrane Permselectivity*

In an electrolyte solution the current is carried by both ions. However, cations and anions usually carry different portions of the overall current. In ion-exchange membranes the current is carried preferentially by the counterions. The fraction of the current that is carried by a certain ion is expressed by the ion transport number, which is given by:

$$T_i = \frac{|z_i| J_i}{\sum_j |z_j| J_j} \tag{3.88}$$

where T_i is the transport number of component *i*, J_i its flux, z_i its valence, and subscript *j* refers to all ions involved in the charge transport.

The transport number T_i indicates the fraction of the total current that is carried by ion *i*; the sum of the transport numbers of all ions in a solution is 1.

The membrane permselectivity is an important parameter for determining the performance of a membrane in a certain ion-exchange membrane separation process. It describes the degree to which a membrane passes an ion of one charge and retains an ion of the opposite charge. The permselectivity of cation- and anion-exchange membranes can be defined [58] by the following relations:

$$\Psi^{\mathrm{cm}} = \frac{T_{\mathrm{c}}^{\mathrm{cm}} - T_{\mathrm{c}}}{T_{\mathrm{a}}} \tag{3.89}$$

$$\Psi^{\mathrm{am}} = \frac{T_{\mathrm{a}}^{\mathrm{am}} - T_{\mathrm{a}}}{T_{\mathrm{c}}} \tag{3.90}$$

where Ψ is the permselectivity of a membrane, *T* the transport number, superscripts cm and am refer to cation- and anion-exchange membranes, and subscripts c and a to cation and anion, respectively.

The permselectivity of an ion-exchange membrane relates the transport of electric charges by a specific counterion to the total transport of electric charges through the membrane and the transport number of the ion in the solution. An ideal permselective cation-exchange membrane would transmit positively charged ions only, that is, for a transport number of a counterion in a cation-exchange membrane of $T_c^{cm} = 1$ the permselectivity $\Psi^{cm} = 1$. The permselectivity approaches zero when the transport number within the membrane is identical to that in the electrolyte solution, that is, for $T_c^{cm} = T_c$, $\Psi^{cm} = 0$. For the anion-exchange membrane, the corresponding relation holds.

The transport number of a certain ion in the membrane is proportional to its concentration in the membrane, which again is a function of its concentration in the solutions in equilibrium with the membrane phase, due to Donnan exclusion, as discussed above. The concentration of a co-ion in an ion-exchange membrane can be calculated from the Donnan equilibrium, which is given in Equation (2.29). For a monovalent salt and a dilute salt solution and assuming the activity coefficients of the salt in the membrane and the solution to be 1, the co-ion concentration is given to a first approximation by:

$${}^{m}C_{\mathrm{co}} = \frac{{}^{s}C_{\mathrm{s}}^2}{C_{\mathrm{fix}}} \tag{3.91}$$

where *C* is the concentration, subscripts co, s, and fix refer to co-ion, salt, and fixed ion of the membrane, and superscripts s and m to membrane and solution.

Equation (3.91) indicates that the co-ion concentration in the membrane increases with increasing solution salt concentration, and the membrane permselectivity decreases correspondingly.

3.7.3. *Membrane Counterion Permselectivity*

The transport number of different counterions in the membrane can be quite different. The transport rates of ions are determined by their concentration and their mobility in the membrane. The concentration of the counterions is always close to the concentration of the fixed charges of the membrane. The mobility of the ions in the membrane depends mainly on the radius of the hydrated ions and the membrane structure. The mobilities of different ions in an aqueous solution do not differ

very much from each other. An exception are the H^+ and OH^- ions, the mobility of which is higher than that of other ions by a factor of about 5 to 6. The exceptionally high mobility of the H^+ ion can be explained by the transport mechanism of protons. Salt ions move with their hydration shell through the solution. Protons interact with water dipoles and form hydronium ions, which are transported in the membrane mostly by a so-called tunnel mechanism. The transport mechanism of the proton is also the reason for the high permeability of anion-exchange membranes to protons, while these membranes generally have a very low permeability for salt cations. The same mechanism as that described here for the transport of protons also holds for the transport of hydroxide ions, and thus the permeability of hydroxide ions in a cation-exchange membrane is much higher than that of other salt anions.

The permselectivity of an ion-exchange membrane for different salt counterions is determined by the concentration and the mobility of the different ions in the membrane. The concentrations of the different counterions in the ion-exchange membrane are generally also different. A typical counterion-exchange sequence of a cation-exchange membrane containing SO_3^- groups as fixed charges is:

$$Ba^{2+} > Pb^{2+} > Sr^{2+} > Ca^{2+} > Mg^{2+} > Ag^+ > K^+ > NH_4^+ > Na^+ > Li^+$$

A similar counterion-exchange sequence is obtained for anions in an anion-exchange membrane containing quaternary ammonium groups as fixed charges [22, 56]:

$$I^- > NO_3^- > Br^- > Cl^- > SO_4^{2-} > SO_4^{2-} > F^-$$

3.7.4. *Water Transport in Electrodialysis*

Water transport in electrodialysis from the diluate to the concentrate process stream can effect the process efficiency significantly. If a convective flux as a result of pressure differences between flow streams can be excluded there are still two sources for the transport of water from the diluate to the concentrate solution. The first is the result of osmotic pressure differences between the two solutions, and the second is due to electroosmosis, which results from the coupling of water to the ions being transported through the membrane due to the driving force of an electrical potential.

Each of the two fluxes may be dominant, depending on the permselectivity of the ion-exchange membrane, the concentration gradient, and the current density. In a highly permselective membrane and for moderate differences in salt concentration in the two solutions separated by the membrane, the electroosmotic flux dominates and is generally much higher than the osmotic solvent flux. In electrodialysis the water flux due to electroosmosis can be expressed by a solvent transport number, which gives the number of water molecules transported by one ion:

$$J_w = {}^mT_w \sum_i J_i \tag{3.92}$$

where ${}^mT_w$ is the water transport number, J_w the water flux, and J_i the flux of ions through a given membrane.

The water transport number is thus:

$${}^mT_w = \frac{J_w}{\sum_i J_i} \tag{3.93}$$

The water transport number is the number of water molecules transferred by one ion through a given membrane. It depends on the membrane and on the electrolyte, that is, on the size of the ions, their valence, and their concentration in the solution. In aqueous salt solutions and commercial ion-exchange membranes the water transport number is on the order of 4 to 8, that is, 1 mol of ions transports ca. 4 – 8 mol of water through a typical commercial ion-exchange membrane.

3.8. Electrodialysis With Bipolar Membranes

Electrodialysis with bipolar membranes has gained increasing attention as an efficient tool for the production of acids and bases from the corresponding salts [59]. This process is economically attractive and has many potential applications. A typical arrangement of an electrodialysis stack with bipolar membranes is illustrated in Figure 16, which shows the production of an acid and a base from the corresponding salt in a repeating cell unit which consists of three individual cells containing the salt solution, the acid, and the base, and three membranes (a cation-exchange, an anion-exchange, and a bipolar membrane); 50–100 repeating cell units may be placed between two electrodes in industrial-scale stacks.

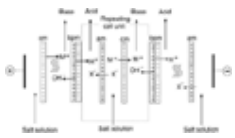

Figure 16. Schematic illustrating acid and base production from the corresponding salt by electrodialysis with bipolar membranes

The key element in electrodialysis with bipolar membranes is the bipolar membrane itself. Its function is illustrated in Figure 17, which shows a bipolar membrane consisting of an anion- and a cation-exchange layer arranged in parallel between two electrodes. If a potential difference is established between the electrodes, all charged components will be removed from the interface between the two ion-exchange layers. If only water is left between the membranes, further transport of electrical charges can be accomplished only by protons and hydroxyl ions, which have a concentrations of about 10^{-7} mol L^{-1} in completely deionized water. Protons and hydroxyl ions removed from the interface between the cation- and anion-exchange membranes are replenished because of the water dissociation equilibrium:

$$2\,H_2O \rightleftharpoons H_3O^+ + OH^-$$

The rate at which protons and hydroxide ions are removed from the bipolar membrane, that is, the maximum electrical current, is limited by the rate of water diffusion into the bipolar membrane and the water dissociation rate. The water dissociation is given by:

$$J_{H^+} = J_{OH^-} = k_d\, C_{H_2O} 2\lambda \tag{3.94}$$

where J is the maximum ion flux from the bipolar membrane into the outer phases, k_d the water dissociation rate constant, C the concentration in the transition layer, and 2λ is the thickness of the transition layer.

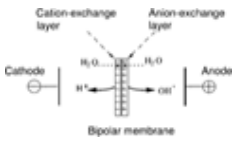

Figure 17. Schematic illustrating the function of a bipolar membrane

The water dissociation rate constant k_d at 25°C is 2.5×10^{-5} s^{-1}. The water concentration in pure water is 55.5 mol L^{-1}. Assuming a thickness of the transition layer of 1 nm, the maximum ion fluxes J_{H^+} and J_{OH^-}, respectively, calculated by Equation (3.95), would be 1.4×10^{-9} mol m^{-2} s^{-1}. The electrical current density i is given by:

$$i = F\sum z_i J_i. \tag{3.95}$$

Thus, the ion fluxes J_{H^-} and J_{OH^-} would correspond to current density of about 1.4 × 10×4 A m^{-2}. In practice, however, bipolar membranes can be operated at current densities in excess of 1000 A m^{-2} [60]. This means that the simple model of a bipolar membrane is incorrect and that water dissociation is at least 10^6 times faster in bipolar membranes than in solution. The actual mechanism of water dissociation, which to a large extent determines the overall efficiency of the process, is still the subject of controversial discussion, and various models are suggested in the literature to explain the accelerated water dissociation which is observed in the practical application of bipolar membranes [61-63].

The energy required for the production of acids and bases in a bipolar membrane at constant temperature and pressure can be calculated from the concentration chain of solutions with different H^+ ion activities. The Gibbs free energy of water dissociation in a bipolar membrane can be calculated from the difference in the pH values of the solution in the phases between the cation- and anion-exchange layers of the bipolar membrane and the solutions outside the bipolar membrane and is given by:

$$\Delta G = RT \ln \frac{a^i_{H^+} a^i_{OH^-}}{a^o_{H^+} a^o_{OH^-}} = RT \ln \frac{K_w}{a^o_{H^+} a^o_{OH^-}} = 2.3RT\Delta pH = F\Delta\varphi \tag{3.96}$$

where ΔG is the Gibbs free energy, R the gas constant, T the absolute temperature, F the Faraday constant, a the activity of H^+ and OH^- ions, $\Delta\varphi$ the electrical potential difference between the solution in the bipolar membrane and the solutions outside the bipolar membrane, ΔpH the difference in pH value of the two solutions separated by the bipolar membrane, K_w the dissociation constant of water, and superscripts i and o refer to the interface between the cation- and anion-exchange layers of the bipolar membrane and the outside solutions, respectively.

At 25°C –log K_w is 13.997. The free energy for the dissociation of 1 mol of water, and thus for the production of 1 mol L^{-1} acid and base at 25 °C, is 79.887 kJ or 0.0222 kW·h. The voltage drop between the two outside solutions is 0.828 V. The actual potential drop across the bipolar membrane is higher due to the electrical resistance of the membranes and the solutions.

3.9. Energy Requirements of Membrane Processes

The energy required in membrane processes is the sum of two terms: 1) the energy to transfer the components from the feed mixture through a membrane into the permeate mixture, and 2) the energy required to pump the solutions through the membrane apparatus and to operate the process monitoring and control devices. The second term depends on the plant capacity and process design. The energy required for separating different components permeating through a membrane consists of two parts. The first part is the reversible energy given by the Gibbs free energy obtained when the retentate and permeate solutions are mixed. It is the minimum energy required for the separation of components in a mixture. It is identical for any separation process. The second part is the energy required to overcome the friction between the transported components and the membrane matrix. It represents the irreversible energy loss.

3.9.1. *Minimum Energy Required for Separation*

The minimum work that is required for the separation of various components from a mixture in a membrane process is given by the Gibbs free energy of mixing. which is given at constant pressure and temperature by:

$$(d\Delta G^m)_{p,T} = \Delta G' - \Delta G'' = \sum_i (\mu'_i \Delta n'_i - \mu''_i \Delta n''_i) \cong RT\sum_i \Delta n_i \ln \frac{a'_i}{a''_i} \tag{3.97}$$

where ΔG^m is the Gibbs free energy of mixing two solutions indicated by ′ and ″, μ the chemical potential, a the activity, n the number of moles, and subscript i refers to individual components.

For filtration processes such as reverse osmosis the energy requirement for removing the solvent from the solution can be expressed by the difference in the osmotic pressure of the two solutions, which for a strictly semipermeable membrane is given by:

$$\Delta\pi = \frac{RT}{V_w} \frac{\ln a'_w}{\ln a''_w} \tag{3.98}$$

where the subscript w refers to the solvent, for example, water.

Introducing Equation (3.98) into Equation (3.97) leads to:

$$\Delta G^m = \Delta n_w V_w \Delta\pi \cong V_f \Delta\pi \tag{3.99}$$

where $\Delta\pi$ is the osmotic pressure difference between the two solutions ′ and ″ separated by the membrane, and V_w and V_f are the partial molar volumes of the solvent and the filtrate, which for a dilute solution is to a first approximation given by $\Delta n V_w \cong V_f$.

In electrodialysis the change of the Gibbs free energy for converting a feed solution into a concentrate and a diluate is given by [56]:

$$\Delta G_{total} = \nu_d RT\left(n_s^c \ln \frac{C_s^c}{C_s^f} + n_s^d \ln \frac{C_s^d}{C_s^f}\right) \tag{3.100}$$

where ΔG_{total} refers to the Gibbs free energy change required for the production of the diluate, C the salt concentration, ν_d the stoichiometric dissociation coefficient, and superscripts f, c, and d refer to the feed, the concentrate, and the diluate.

3.9.2. *Practical Energy Requirements of Membrane Processes*

The minimum energy requirement for the separation of molecular mixtures such as salt solutions, discussed in Section Minimum Energy Required for Separation refers to an idealized, completely reversible process. In practice, such conditions never prevail and the energy requirements are usually significantly higher and rather different in different membrane separation processes. They are also a function of the membrane module construction and the overall process design. The practical energy requirements in the different processes are discussed in Section Practical Energy Requirements of Membrane Processes.

[Top of Page]

4. Membrane Separation Process Design and Operation

In the previous chapters the fundamentals of mass transport in membranes and the operating principle of various membrane separation processes are discussed. In this chapter the membrane separation processes are treated as unit operations and the various process designs and operation concepts are discussed.

4.1. Ultrafiltration, Microfiltration, and Reverse Osmosis

Microfiltration, ultrafiltration, and reverse osmosis have many chemical engineering aspects and process design features in common. In all three processes a hydrostatic pressure is utilized as driving force to transport certain components through a membrane which shows different permeability for different components. As a result a feed solution is converted into a filtrate and a retentate. The filtrate contains all components that have permeated the membrane, and the retentate contains the components that are retained or rejected by the membrane. The design of a microfiltration, ultrafiltration, or reverse osmosis process is illustrated in Figure 18.

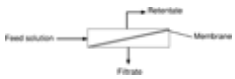

Figure 18. Process scheme of ultrafiltration, microfiltration, and reverse osmosis

The performance of a membrane in a pressure-driven separation process is determined by its filtration rate, the membrane flux, and by the membrane separation properties. Membrane flux and separation properties are a function of membrane permeability for the different components of the solution and the applied hydrostatic pressure. But they are also determined by process and systems design.

4.1.1. *Recovery Rate, Membrane Rejection, Retentate and Filtrate Concentrations*

The separation capability of a membrane in reverse osmosis or ultra- and microfiltration can be expressed in terms of *membrane rejection* which describes the “true” rejection of a membrane. The true rejection of the membrane is given by:

$$R = \left(1 - \frac{C_i^f}{C_i^r}\right) \tag{4.1}$$

where R is the rejection of the membrane for a given component i at a defined hydrostatic pressure and feed solution concentration, and C_i^f and C_i^r are the concentrations of rejected components at a given point at the surface of the membrane facing the filtrate and at the surface facing the retentate, respectively. At a given temperature and pressure the rejection depends only on the membrane properties.

In many practical applications it is more convenient to express the rejection in reference to the feed solution:

$$R_0 = \left(1 - \frac{C_i^f}{C_i^o}\right) \tag{4.2}$$

where R_0 is the rejection in reference to the feed entering a filtration device and to the filtrate leaving the device after a certain amount of the feed has been recovered as filtrate.

Thus, the rejection in reference to the feed is a function of the recovery rate, which is given by the ratio of the filtrate to the feed volume.

$$\Delta = \frac{V^f}{V^o} \tag{4.3}$$

where Δ is the recovery rate and V^f and V^o are volumes of the filtrate and the feed solution.

Since in a practical application the concentration in the retentate is limited by factors such as osmotic pressure and viscosity, the solvent of the feed solution can not be completely recovered as filtrate. The recovery rate has a value between 0 and 1.

The concentrations in the retentate and the filtrate are not only determined by true membrane rejection *R*, but they are also a function of the recovery rate and of the operational parameters such as cocurrent or countercurrent flow of the concentrate and permeate solutions in the membrane device. For a simple batch-type filtration with complete mixing of the retentate and filtrate, as illustrated in Figure 19, the membrane flux is identical over the entire membrane area, and the relation between the concentration of retentate, filtrate, and feed solutions is given by a mass balance:

$$V^o C_i^o = V^r C_i^r + V^f C_i^f \tag{4.4}$$

where C_i^o, C_i^r, and C_i^f are the concentrations in the feed solution, in the retentate, and in the filtrate.

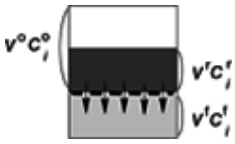

Figure 19. Schematic illustrating a batch filtration with complete mixing of filtrate and retentate

If concentration polarization effects are neglected the true membrane rejection can be calculated according to Equation (4.1) by assuming infinitesimal volume change, that is, the recovery rate approaches 0.

The concentration of a component in the retentate and filtrate at a given recovery rate is obtained from the mass balance by integration over the recovery rate:

$$C_i^r = C_i^o (1 - \Delta)^{-R} \tag{4.5}$$

The concentration in the filtrate is directly proportional to that in the retentate. Thus, for a given recovery rate, the filtrate concentration at any given time is found by introducing Equation (4.1) into Equation (4.5):

$$C_i^f = C_i^o (1 - R)(1 - \Delta)^{-R} \tag{4.6}$$

The filtrate concentration expressed in Equation (4.6) is the concentration corresponding to a given recovery rate, that is, during an infinitely small time interval. For most practical applications, however, the filtrate concentration obtained during an infinitely small time interval and at a certain recovery rate is of less interest than the *mixing cup concentration* [52], which is obtained when the entire filtrate up to a given recovery rate is collected and mixed. The mixing cup filtrate concentration $\overline{C}_i^f$ can be calculated by integration of Equation (4.5):

$$\overline{C}_i^f = \frac{C_i^o}{\Delta}\left[1 - (1 - \Delta)^{1-R}\right] \tag{4.7}$$

The mixing cup concentration is always lower than the concentration obtained during an infinitely short time period at a certain recovery rate. Figure 20 shows the retentate, filtrate, and mixing cup concentrations as a function of recovery rate for three different membrane rejection coefficients (*R* = 0.99, 0.9, 0.8, and 0.5).

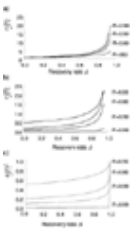

Figure 20. Ratios of a) the retentate to feed solution concentration, b) the filtrate to feed solution concentration, and c) the mixing cup filtrate to feed solution concentration as a function of the recovery rate calculated for different membrane retentions

Figure 20 shows that the retentate concentration strongly depends on the recovery rate and reaches high values for membranes with high retention properties. The concentration in the filtrate also depends on the recovery rate and reaches high values for membranes with low retention properties. The mixing cup filtrate concentration increases less strongly with recovery rate, especially for membranes with high retention properties, and approaches its maximum value of 1 at a recovery rate of 1.

In membrane evaluation or characterization tests, the rejection *R* of a membrane can be calculated from two easily measurable values: the mixing cup filtrate concentration $\overline{C}_i^f$ and feed concentration C_i^o.

$$R = 1 - \frac{\ln\left(1 - \frac{\overline{C}_i^f \Delta}{C_i^o}\right)}{\ln(1 - \Delta)} \tag{4.8}$$

The significance of the mixing cup and maximum retentate concentrations can be illustrated in practical examples. For instance, in the production of potable water from seawater or brackish water, the concentration of total dissolved salts in the filtrates should not exceed 500 mg/L. This concentration is the mixing cup filtrate concentration. For given membrane and feed solution properties it is a function of the recovery rate and can be calculated by Equation (4.7).

If the feed solution contains a solute with limited solubility such as $CaSO_4$ or $CaCO_3$, the maximum recovery rate is determined by the maximum retentate concentration achievable without solute precipitation. The maximum recovery rate for given membrane and feed solution properties can be calculated by Equation (4.5).

4.1.2. *Solute Loss in Filtration*

In a membrane filtration process, the “product” may be either the filtrate, for example, as in the production of ultrapure water, or the retentate, for example, as in the concentration of proteins from whey. For membranes that are not strictly semipermeable some solute will permeate the membrane with the filtrate. This may affect the quality of the filtrate or lead to product losses.

The *fractional solute loss* δ of a component *i* is usually expressed by the amount of solute lost with the filtrate divided by the total amount of solute in the feed solution:

$$\delta = \frac{V^f \overline{C}_i^f}{V^o C_i^o} = 1 - (1 - \Delta)^{1-R} \tag{4.9}$$

where V^f and V^o are the volumes of the filtrate and the initial feed solution, $\overline{C}_i^f$ is the mixing cup filtrate concentration, C_i^o the initial feed solution concentration, Δ the recovery rate, and *R* the membrane rejection.

Product loss may be significant even with membranes that have a relatively high solute rejection when high retentate concentrations are desired, as illustrated in Figure 21.

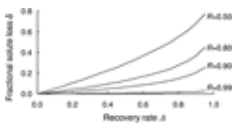

Figure 21. Fractional solute loss in membrane filtration calculated as a function of the recovery rate for different membrane retentions

4.1.3. *Effect of Osmotic Pressure on the Filtration Rate*

The main difference between reverse osmosis and ultra- or microfiltration is the size or molecular mass of the solutes separated from a solution. In reverse osmosis, low molecular mass components are separated from a solvent. A solution of low molecular mass components can have a significant osmotic pressure that must be overcome by the hydrostatic pressure applied as driving force for the separation. The osmotic pressure is always negligibly low in microfiltration. In ultrafiltration of macromolecular solutions it may have an effect on the filtration rate. In reverse osmosis of low molecular components such as salts the osmotic pressure plays a very important role. It often is the limiting parameter for the maximum concentration that can be achieved in the retentate.

The osmotic phenomenon and the osmotic pressure of a solution is discussed in Section Osmotic Equilibrium, Osmotic Pressure, Osmosis, and Reverse Osmosis. The osmotic pressure of a solution is proportional to its solute concentration and given by:

$$\pi = RT \sum_i g_i C_i \tag{4.10}$$

where *R* is the gas constant, *T* the absolute temperature, and *g* the osmotic coefficient, which is a correction factor for the nonideal behavior of a solution, C_i the concentration of a component *i*, which in a dissociated salt refers to individual particles (i.e. cations and anions).

The membrane flux rate is determined by the effective pressure, which is given by the difference between applied pressure and osmotic pressure. Since the osmotic pressure increases with increasing solution concentration and thus with increasing recovery rate, the membrane flux decreases with increasing recovery rate. In microfiltration and to a large extent also in ultrafiltration the osmotic pressure is negligibly low compared to the applied pressure because of the high molecular weight of the retained components. Therefore, the filtration flux will be independent of the feed or retentate concentration. This, however, is not the case in reverse osmosis. Here, the osmotic pressure is significant for even modestly concentrated

solutions and increases with increasing recovery rate. In many applications of reverse osmosis the recovery rate is limited by the osmotic pressure.

4.1.4. *Process Operating Mode*

In the practical application of filtration processes three modes of operation are used (Figure 22).

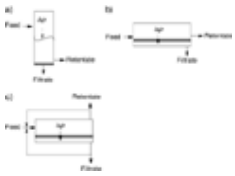

Figure 22. Schematic illustrating different operation modes of filtration processes: a) batch process, b) continuous process, and c) feed-and-bleed process

In the batch process depicted in Figure 22a a given feed volume is placed in a pressure vessel. Under a hydrostatic pressure certain components, mainly solvents, permeate the membrane and are collected as filtrate. The components retained by the membrane are concentrated. When a certain concentration in the retentate is achieved the process is terminated.

In the continuous process depicted in Figure 22b the feed solution is continuously fed into the filtration device, which generally consists of a membrane-lined channel or tube. Under a hydrostatic pressure certain components permeate the membrane and are collected as filtrate, which leaves the device through an outlet port. The retained components are concentrated during passage through the device and leave at the end of the process as the retentate. Depending on the flow velocity of the feed solution, the geometry of the device, and the filtration rate a certain recovery rate is achieved. If the recovery rate achieved in one process pass is not satisfactory a higher recovery rate can be obtained by applying a so-called feed-and-bleed operating mode. In this case part of the retentate is recycled to the device inlet and mixed with the feed solution, as shown in Figure 22c. Depending on the retentate recycling ratio, different recovery rates can be achieved. If a component is not completely retained by the membrane its concentration is proportional to the retentate concentration and thus increases with increasing recovery rate. If the concentration of partly retained components in the filtrate exceeds the desired maximum value in the given process path length part of the filtrate may be recycled to the feed inlet. By adjusting the ratio of the filtrate recycled to the feed to the overall filtrate the concentration in the filtrate can be kept to a desired value.

4.1.5. *Diafiltration Operating Mode*

A variation of the ultrafiltration process, generally referred to as diafiltration [64], is often used when more complete separation of micro- and macrosolutes is required. Diafiltration is used, for instance, to remove salts from a mixture with proteins or other macromolecular components. In this application it competes directly with dialysis. The principle of diafiltration is illustrated in Figure 23.

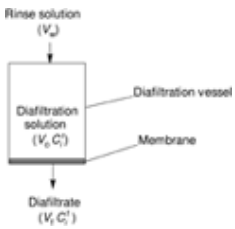

Figure 23. Schematic illustrating the diafiltration operating mode

The diafiltration membrane should retain macromolecular components, but should be permeable to microsolutes. If filtrate removed from the cell is replaced by pure solvent, the low molecular mass constituents are gradually washed out of the feed mixture. If the volume of the original solution is kept constant, the concentration in the diafiltration vessel can be calculated as a function of the membrane retention and the volume of diafiltration rinse solution (i.e., the pure solvent replacing the filtrate volume) by a mass balance which is:

$$\frac{V_f C_i^f}{dt} = -V_o \frac{d\,C_i^r}{d\,t} \tag{4.11}$$

By introducing $V_f = V_w$ and $C_i^f = (1–R)C_i^r$ into Equation (4.11) and integrating over a certain time period the concentration in the diafiltration vessel can be expressed by:

$$C_i^{rt} = C_i^{ro} \exp -(1-R)\frac{V_w}{V_o} \tag{4.12}$$

where C_i^f and C_i^r are the concentrations of component *i* in the filtrate and in the diafiltration vessel, C_i^{rt} and C_i^{ro} the concentrations in the diafiltration vessel at time *t* and at the beginning of filtration, V_o is the volume of the diafiltration vessel, V_f and V_w are the volumes of the filtrate and the rinse solution added during time *t*, and *R* is the rejection of the membrane for the component under consideration.

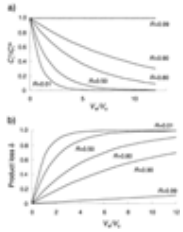

Figure 24. a) Reduced retentate concentration in diafiltration as a function of the ratio of rinse water volume to diafiltration vessel volume calculated for various membrane retentions, and b) fractional product loss in diafiltration as a function of the rinse water volume calculated for different membrane retentions

If the component is completely retained by the membrane (i.e., R = 1), the concentration in the batch cell remains constant. If the membrane is completely permeable to the component (i.e., R = 0), the concentration in the reactor decreases according to the exponential function of Equation (4.12). The “rinse-out” effect in diafiltration is illustrated in Figure 24a, which shows the reduced concentration in the diafiltration process as a function of the ratio of the rinse water volume to diafiltration vessel volume calculated for different membrane rejections.

If the macromolecular constituents are only partially rejected by the membrane, a substantial loss of product occurs. If complete removal of the low molecular mass components without substantial loss of macromolecular material is desired, rejection of the membrane should be close to zero for the component low in molecular mass and close to one for the macromolecular material.

As in ultrafiltration, product losses may be significant, for example, if the membrane does not reject the macromolecules completely in desalting a macromolecular solution. The fractional product loss δ is then given by

$$\delta = \frac{V^{\mathrm{f}}\overline{C}_i^{\mathrm{f}}}{V^{\mathrm{o}}C_i^{\mathrm{o}}} = 1 - \exp - (1 - R)\frac{V_{\mathrm{w}}}{V_{\mathrm{o}}} \tag{4.13}$$

where V is the volume, R the membrane rejection, $\overline{C}_i^{\mathrm{f}}$ the mixing cup concentration, C the concentration, superscripts o and f refer to the initial feed solution and the filtrate, and subscripts o and w to the original volume of the feed and the volume of the wash solution.

In Figure 24b the fractional product loss is shown as a function of the ratio of wash solution volume V_w to feed solution volume V_o for membranes with different rejection characteristics. The concentration of a low molecular mass solute is reduced to less than 1 % of its original value by a wash solution volume five times that of the feed solution volume when the membrane passes the solute unhindered (i.e., R approaching 0), as shown in Figure 24. If the solute is partially rejected by the membrane, correspondingly more wash solution is required to reduce the concentration of the low molecular mass solute below a certain value. On the other hand, when the membrane rejects the macrosolute completely (i.e., R approaching 1) no product is lost, but partial rejection leads to significant product loss, as shown in Figure 24b.

4.2. Gas Separation

In gas separation, a feed stream enters a membrane device as indicated in Figure 25. Under the driving force of a hydrostatic pressure difference, some components permeate the membrane while others are retained, and a feed stream is split into a permeate and a retentate.

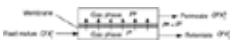

Figure 25. Schematic illustrating gas separation as a unit operation

The different flow streams illustrated in Figure 25 are related with each other by the mass balance:

$$Q^{\mathrm{f}}X_j^{\mathrm{f}} = Q^{\mathrm{p}}X_j^{\mathrm{p}} + Q^{\mathrm{r}}X_j^{\mathrm{r}} \tag{4.14}$$

where Q is the flow stream, X the molar fraction of the component j, superscripts f, p, and r refer to feed, permeate, and retentate, respectively, and subscript j refers to a component in the gas mixture.

If the membrane has a higher permeability for one of the components in the feed mixture than for others, this component is enriched in the permeate and depleted in the retentate. Gases can be separated in microporous and in homogeneous membranes. The selectivity of microporous membranes is generally rather low because of the Knudsen diffusion transport mechanism.

For a practical application of gas separation it is of interest:

- To what extent can a given gas mixture be depleted of a certain component?
- To what extent can a given component be enriched in the permeate?
- How much of the original feed stream can be recovered as permeate or retentate?
- Which hydrostatic pressure difference is needed to achieve the desired separation?
- How much membrane area is required for this operation?

There are three crucial parameters that determine the efficiency of a membrane separation process. These are the membrane selectivity $S_{i,k}$, the ratio of the feed to permeate pressure, and the stage cut. The selectivity is given by the membrane properties and can be expressed by:

$$S_{j,k} = \frac{P_j}{P_k} \tag{4.15}$$

The pressure ratio and the stage cut are operating-related parameter given by:

$$\Phi = \frac{p^{\mathrm{p}}}{p^{\mathrm{f}}} \tag{4.16}$$

and

$$\theta = \frac{Q^{\mathrm{p}}}{Q^{\mathrm{f}}} \tag{4.17}$$

where *S* and *P* are the membrane selectivity and the permeability, Φ and *p* the pressure ratio and the hydrostatic pressure, θ and *Q* the stage cut and the gas flow rate, and subscripts *j*, *k*, p, and f refer to the components *j* and *k*, the permeate, and the feed.

The membrane separation factor $\alpha_{j,k}$ and the membrane enrichment factor $\beta_{j,k}$ are related to the membrane selectivity and the pressure ratio by:

$$\alpha_{jk} = \frac{X_j^{\mathrm{p}} X_k^{\mathrm{f}}}{X_j^{\mathrm{f}} X_k^{\mathrm{p}}} = S_{j,k} \frac{X_j^{\mathrm{p}}\Phi - X_j^{\mathrm{f}} X_k^{\mathrm{f}}}{X_k^{\mathrm{p}}\Phi - X_k^{\mathrm{f}} X_j^{\mathrm{f}}} \tag{4.18}$$

and

$$\beta_j = \frac{X_j^{\mathrm{p}}}{X_j^{\mathrm{f}}} = \frac{\alpha_{j,k}}{1 + (\alpha_{j,k} - 1) X_j^{\mathrm{f}}} \tag{4.19}$$

where α is the separation factor, β the enrichment factor, *S* the membrane selectivity for a binary mixture, Φ the pressure ratio, *X* the molar fraction, subscripts *k* and *j* refer to two components, and superscripts f and p to the feed and the permeate.

The pressure ratio Φ has a value between 0 and 1. Thus, the maximum separation and enrichment is achieved when the pressure ratio approaches zero. The effect of the pressure ratio on the enrichment or the separation that can be achieved in a separation process depends also on the construction of the membrane module and the mode of operation, such as cocurrent or countercurrent or complete mixing of feed and permeate. This makes the calculation of the enrichment as a function of the pressure ratio rather complex and requires numerical solutions which are described in detail in the literature or provided by companies selling gas separation equipment [2, 65-67]. Only for very simple cases such as complete mixing of permeate and retentate or for a very low stage cut can the composition of the permeate be calculated to a first approximation by [68]:

$$X_j^{\mathrm{p}} = \frac{X_j^{\mathrm{f}}\Phi}{2} + \frac{1}{2} + \frac{\Phi}{2(S_{j,k} - 1)} - \left[\left(\frac{X_j^{\mathrm{f}}\Phi}{2} + \frac{1}{2} + \frac{\Phi}{2(S_{j,k} - 1)}\right)^2 - \frac{S_{j,k}\Phi X_j^{\mathrm{f}}}{(S_{j,k} - 1)}\right]^{0.5} \tag{4.20}$$

A practical application of gas separation is the enrichment of oxygen in air using a silicone rubber membrane which has a selectivity of 2. The molar fraction of oxygen in air is assumed to be 0.2 and the stage-cut is assumed to be 0. In Figure 26 the calculated enrichment factor β and the separation factor α are shown as a function of the pressure ratio.

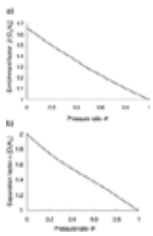

Figure 26. a) The oxygen enrichment factor α and b) the separation factor β as a function of the pressure ratio for a membrane selectivity $S_{O2/N2}$ of 2 and a feed composition of X_{O2}/X_{N2} of 0.2/0.8

The example of oxygen enrichment in air illustrated in Figure 26 shows that the maximum separation and maximum enrichment can only be achieved when the ratio of permeate over feed approaches 0, that is, when either the feed pressure is infinitely high or the permeate pressure approaches 0 and when the stage cut is 0. The above example also shows that the maximum molar fraction in air that can be achieved in gas separation using a silicone rubber membrane is about 0.33 if the pressure ratio and the stage cut approach 0.

The flux through a membrane is proportional to the pressure difference between the feed and the permeate side, as indicated in Equation (3.56). This means that to obtain high fluxes the pressure difference between feed and permeate should be as high as possible, and to obtain high separation the pressure ratio between the permeate and the feed should be as low as possible.

In gas separation the achievable separation or enrichment of a component in the permeate or retentate depends not only on the membrane properties and the applied hydrostatic pressure ratio but also on the so-called stage cut, which corresponds to the recovery rate used in filtration processes and which is given by:

$$\theta = \frac{V^{\mathrm{p}}}{V^{\mathrm{f}}} \text{ or } \theta = \frac{Q^{\mathrm{p}}}{Q^{\mathrm{f}}} \tag{4.21}$$

if a continuous process as shown in Figure 22 is considered.

Here $\hat{\theta}$ is the stage cut, V and Q are the volume and the volume flux, respectively, and superscripts p and f refer to permeate and feed. The stage cut has a value between 0 and 1.

The molar fraction of a component in the feed, the permeate, and the retentate is related to the stage cut by the mass balance expressed in Equation (4.14). Introducing the stage cut in this equation leads to:

$$X_j^{\mathrm{f}} = \theta X_j^{\mathrm{p}} + (1-\theta) X_j^{\mathrm{r}} \tag{4.22}$$

where X_j is the mole fraction of component j and superscripts f, p, and r refer to feed, permeate, and retentate.

The permeate and retentate concentration as a function of the stage cut is illustrated in Figure 27 for the separation of oxygen and nitrogen. The calculation is based on a membrane selectivity $S_{O2/N2}$ of 2 and a feed composition X_{O2}/X_{N2} of 0.2/0.8. Furthermore, it is assumed that the pressure ratio of permeate and retentate pressure approaches 0 and the retentate and permeate are completely mixed. Under the above conditions the highest oxygen concentration that can be reached in the permeate is 0.33 when the stage cut approaches 0. The lowest oxygen concentration in the retentate of 0.11 is reached when the stage cut approaches 1.

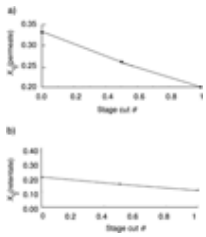

Figure 27. a) The oxygen permeate and b) the oxygen retentate concentration as a function of the stage cut assuming complete mixing of permeate and retentate, for a membrane selectivity $S_{O2/N2}$ of 2, and a feed composition of X_{O2}/X_{N2} of 0.2/0.8

In practical applications gas separation systems are not operated with complete mixing of the permeate and retentate, and the calculation of the permeate and retentate concentrations becomes significantly more complicated and depends on the membrane module design and flow conditions of permeate and retentate. In gas separating modules three idealized flow patterns can be assumed (Fig. 28). The effect of the various flow patterns on the performance of a unit is rather significant. The determination of the required membrane area and separation characteristics for the different flow patterns for binary and multicomponent mixtures are described in the literature, and computer programs for parametric studies are available for all flow patterns [65-67].

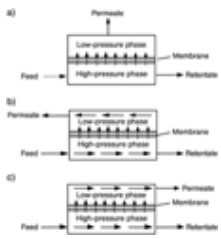

Figure 28. Schematic illustrating three idealized feed and permeate flow patterns used in gas separation systems. a) Complete mixing of feed and permeate; b) Countercurrent plug flow of feed and permeate; c) Cocurrent plug flow of feed and permeate

The separation obtained in a single permeation stage can be multiplied many times, if necessary, by connecting an appropriate number of stages in series to form a countercurrent *permeation cascade*. There are two possible arrangements. In the first arrangement there is no reflux of the retentate. A typical section of a permeation cascade without reflux of the retentate is shown in Figure 29a. In this simple arrangement the permeate from stage n becomes the feed for the next higher stage $n + 1$, and the retentate is disposed of. In the second case the retentate is refluxed, that is, the retentate of stage n is mixed with the next lower stage n–1 and so on. All permeate streams must be recompressed before entering a higher stage. The simple cascade without reflux of the retentate is only of use when the retentate is virtually of no value and large enrichment factors or product concentrations in the permeate are required. If a cascade with reflux of the retentate is used there are two sections depending on the position where the original feed solution is introduced into the cascade. One is the so-called enrichment section where the product is enriched in the permeate, and the other is the stripping section where the product is enriched in the retentate. The principle of a permeation cascade with reflux of the retentate is illustrated in Figure 29b.

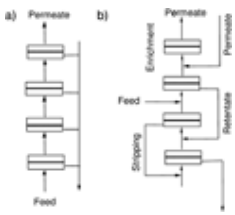

Figure 29. Flow diagram of permeation cascades a) without reflux of the retentate and b) with reflux of the retentate

The number of stages needed for a required separation factor can be determined by the McCabe – Thiele method, a graphical procedure commonly used in the design of distillation columns. The subject of cascade operation is of rather fundamental importance for all separation processes and therefore treated in detail in the corresponding literature [69].

The graphical procedure of determining the number of stages is based on two basic relations. One is the equilibrium curve expressing the relation between the composition of the permeate and the retentate in one stage; this is a function of the membrane properties, operating pressures, stage cut, and flow pattern. The second curve is the operating line, which describes the relation between the composition of the permeate leaving stage *n* and that of the retentate of the next higher stage. The operating line depends on the operating scheme used in a cascade (e.g., variable or constant stage cut) and is the material balance between two stages. The application of the countercurrent flow principle in a reflux cascade has led to the development of the membrane column [70]. In analogy to distillation a membrane column can be regarded as a reflux cascade with an infinite number of stages, and like a reflux cascade the membrane column will produce an enriched retentate in the stripping section and an enriched permeate in the enrichment section. Compared to the normal cascade the membrane column has the advantage of being able to produce highly enriched products both in the permeate and retentate without pressurizing the permeate before it enters the next stage of the cascade, and it thus requires less energy for compression of the gas. On the other hand, the membrane column requires a significantly higher membrane surface area. Whether a column or a cascade is optimal for a given separation problem must be decided by an economic analysis. It is highly dependent on membrane properties and module design. Although the membrane-column concept can in principle be applied to other membrane separation processes such as reverse osmosis, it is most important for gas separation.

4.3. Pervaporation

In pervaporation, volatile organic components are removed from a liquid feed mixture through a semipermeable membrane into a gas phase. The separation of components from a liquid mixture is determined not only by differences in their vapor pressures, as is the case in distillation, but also by their permeation rates through the membrane. As in gas separation the driving force for the transport is the chemical potential gradient of the permeating components in the membrane, which can be related to the partial vapor pressures in the liquid and vapor phase. In gas separation, the chemical potential difference is induced by a hydrostatic pressure difference. In pervaporation, however, the chemical potential gradient is usually induced by applying a vacuum on the permeate side of a membrane, or by applying a temperature difference between the liquid feed mixture and the permeate gas phase. A sweeping gas can also be used to remove the permeating component. The mass transport in a pervaporation membrane can be described by the same mathematical relations as the gas transport with the exception that the chemical potential in the membrane on the feed side of the membrane is expressed by the molar fraction and the activity coefficient in the liquid phase as discussed in Section Gas Separation.

The different modes of operation in pervaporation are illustrated in Figure 30.

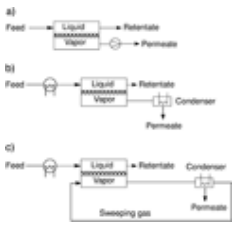

Figure 30. Schematic illustrating different operation modes used in pervaporation: a) vacuum-driven pervaporation, b) temperature-difference-driven pervaporation, and c) sweeping-gas-driven pervaporation

4.4. Dialysis

Dialysis was one of the first technically used membrane separation processes and today is still among the most important commercially. A large fraction of today's artificial kidneys are dialyzers, and hemodialysis is the largest single application of a membrane separation process.

In a dialyzer, one or more solutes are transferred from the feed solution to the dialysate or stripping solution through a membrane under the driving force of a concentration gradient. The principle of the process has been discussed earlier. The mass transfer in a dialyzer is illustrated in Figure 31. A dialyzer cell contains two chambers separated by a membrane. A feed solution to be depleted of a solute is pumped through one chamber while the receiving fluid (dialysate) is passed through the other chamber. The three most common modes of operating a dialyzer are parallel flow and countercurrent flow of feed solution and dialysate and feed flow with the dialysate completely mixed.

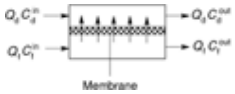

Figure 31. Operating scheme of a dialyzer (*Q* and *C* refer to volume flow rate and concentration of the component to be removed, subscripts f and d to feed solution and dialysate, and superscripts in and out to inlet and outlet)

The overall efficiency of a dialyzer is governed by two interdependent factors: the ratio of the flow rates of the two fluids and the rate constant for solute transport between the fluids, which is determined by membrane properties, membrane area, fluid channel geometry, and local fluid velocities. A material balance can be expressed by:

$$N_i = Q_f\left(C_i^{f\,-\mathrm{out}}\,C_i^{f}\right) = Q_d\left(C_i^{d\,-\mathrm{in}}\,C_i^{d}\right) \tag{4.23}$$

where *N* is the overall transport rate of the solute, *Q* and *C* are the volume flow rate and the concentration, subscripts *i*, f, and d refer to a component, the feed solution, and the dialysate, and superscripts f and d, and in and out refer to feed and dialysate, and inlet and outlet. The overall rate of solute transport is also expressed by:

$$N_i = J_i A = k^* A \Delta \overline{C}_i \tag{4.24}$$

where J_i is flux rate and $\Delta \overline{C}_i$ the average concentration difference of component *i* between the fluids, *A* the membrane area, and *k** the overall rate constant. The efficiency of a dialyzer is expressed in terms of its *dialysance* D_f, which is defined as [71]:

$$D_f = \frac{N}{C_i^f - {}^{in}C_i^d} = \frac{k^* A \Delta \overline{C}_i}{C_i^f - {}^{in}C_i^d} \tag{4.25}$$

The dimensionless ratio D_f/Q_f is a convenient efficiency parameter, because it represents the fraction of maximum attainable solute depletion actually achieved in the feed.

The proper average solute concentration difference is the logarithmic mean of the inlet and outlet differences. The mean concentration difference between the feed and dialysate solution depends on the membrane module design and the mode of operation, that is, complete mixing of feed and dialysate or co- or countercurrent flow. The calculation of the dialysance for the different operation modes by simultaneous solution of Equations (4.23) to (4.25) using the logarithmic mean concentration difference as driving force is described in the literature [71].

In dialysis with cation- or anion-exchange membranes, which is used to recover acids or bases from a mixture with salts, the recovery rate of the acid or base and the maximum concentration that can be achieved is of interest. The recovery rate is given by:

$$\Delta = \frac{{}^{out}C_i^d Q^d}{{}^{in}C_i^f \; Q^f} \tag{4.26}$$

where Δ is the recovery rate of the desired product (acid or base), *C* the concentration, *Q* the volume flow through the dialysis cell, and superscripts f and d, and in and out refer to feed and dialysate, and inlet and outlet.

The recovery rate Δ determines how much of the acid or base which is in a mixture with a salt can be recovered. Δ has a value between 0 and 1. It is related to the concentration of the recovered product and the concentration in the depleted solution at the cell exit by:

$${}^{out}C_i^d = \frac{Q_f}{Q_d} \Delta \, {}^{in}C_i^f \tag{4.27}$$

and

$${}^{out}C_i^f = (1 - \Delta) {}^{in}C_i^f \tag{4.28}$$

According to Equations (4.23) to (4.28) the concentration of a component such as an acid or base in the final dialysate or the depleted feed solution can be determined as a function of the overall mass transfer coefficient, the total membrane area, and the volume flow rates of the feed and the product for different dialysis operations modes [72].

The modeling of diffusion dialysis only on the basis of concentration differences, as suggested in the literature [72], does not consider electrochemical interactions of the different ions (i.e. a diffusion potential or osmotic and electroosmotic effects) and can therefore only be considered as a first approximation for the design of a diffusion dialysis process.

4.5. Electrodialysis

The basic concept of electrodialysis is discussed in Section Principle of Electrodialysis. Although there are many different components necessary for the proper operation of an electrodialysis plant such as the electrical power supply, pumps, and control and monitoring devices, the stack is a key component in an electrodialysis unit. For designing and operating an electrodialysis stack certain process parameters must be taken into account and be controlled such as the feed solution concentration and the desired product and brine concentrations. Various modes of operation are possible such as batch or continuous. In many cases the desired desalination or concentration of a feed solution cannot be obtained in a single pass through the electrodialysis cells. To achieve a higher degree of desalination or concentration two or more stacks are placed in series. When high product recovery rates are required part of the brine may be recycled. A cell arrangement of an electrodialysis stack is shown schematically in Figure 32.

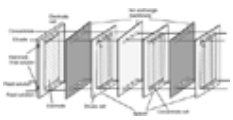

Figure 32. Schematic illustrating the construction of a sheet-flow electrodialysis stack

An electrodialysis stack is composed of a multitude of cells placed in parallel between two electrodes. The different cells are separated by ion-exchange membranes. In alternating cells a feed solution is concentrated and desalinated, respectively. A spacer-gasket arrangement separates the membranes and contains the manifolds to distribute the process fluids in the different compartments. In designing and operating an electrodialysis stack, several criteria concerning the hydrodynamic and electrical properties must be considered. A proper electrodialysis stack design should provide a maximum effective membrane area per unit stack volume and ensure equal and uniform flow distribution in each compartment. Any leakage between the diluate, concentrate, and the electrode cells should be prevented. The spacer screen should provide maximum mixing of the solutions at the membrane surfaces and should cause a minimum in pressure loss [73, 74]. In practice, two

different stack types are used on a large scale: sheet-flow and the tortuous path flow stacks. In a sheet-flow electrodialysis stack the solution flows in a straight path from the entrance to the exit ports, which are located on opposite sides of the gasket, as indicated in Figure 33a.

In the tortuous path flow stack, the membrane spacer gaskets have a long serpentine cutout which defines a long narrow channel for the fluid path and provides higher residence times of the solution in the cells even at high flow velocities. A tortuous path flow spacer gasket is shown schematically in Figure 33b.

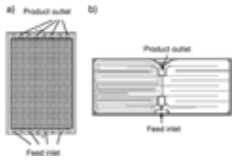

Figure 33. Concept of a) a sheet-flow and b) a tortuous path flow electrodialysis spacer gasket

The distance between the membrane sheets (cell thickness) should be as small as possible to minimize energy consumption due to electrical resistance of the solutions. In industrial-size electrodialysis stacks, membrane distances are typically between 0.3 and 2 mm. A spacer is introduced between the individual membrane sheets to support the membrane and to control the solution flow distribution.

Solution flow velocities in sheet-flow stacks are typically 3 – 10 cm s^{-1}, whereas in tortuous path flow stacks the solution flow velocities are 15 to 50 cm s^{-1}. Because of higher flow velocities and longer flow paths, higher pressure drops on the order of 2 – 3 bar are obtained in stacks with tortuous path flow spacers compared to sheet-flow systems, which have pressure drops between 0.5 and 2 bar. In a practical electrodialysis system, 200 to 1000 cation- and anion-exchange membranes are installed in parallel to form an electrodialysis stack providing 100 to 500 cell pairs. At each end of a stack is an end-plate that contains the electrodes. Depending on the feed solution composition and the product requirements electrodialysis units can be operated in batch-type, continuous, or feed-and-bleed mode with partial recycle of the diluate and concentrate streams.

In practical applications two basic concepts of electrodialysis are used. One is referred to as unidirectional electrodialysis [75] and the other as electrodialysis reversal [76]. In a unidirectionally operated electrodialysis stack the electrical potential gradient across the stack is permanently applied in one direction, and the diluate and concentrate cells are also permanently fixed over the period of operation. Unidirectionally operated electrodialysis plants are rather sensitive to membrane fouling and scaling and often require frequent stack-cleaning procedures. In electrodialysis reversal the change of the electric potential across the stack is accompanied with a reversal of the flow streams, so that the diluate becomes the concentrate and vice versa. In practical applications the polarity is changed at time intervals ranging from a few minutes to several hours. The advantage of electrodialysis reversal is that precipitates in the brine cells are redissolved when the polarity is reversed and the brine cell becomes the diluate cell. During the reversal of the polarity and the flow streams there is a brief period when the concentration of the desalted product exceeds the product quality specification and must be discharged. Thus, in electrodialysis reversal a certain amount of the product is always lost to the waste stream. This is generally no problem in desalination of brackish water, but it may be not acceptable in certain applications in the food and drug industry when feed solutions with high-value products are processed.

The main advantage of electrodialysis reversal is that suspended matter, polyelectrolytes, and multivalent salts that are precipitated on the membrane surfaces are removed when the polarity and the flow streams are reversed. This is illustrated in Figure 34, which shows a typical electrodialysis cell formed by a cation- and anion-exchange membrane between two electrodes. If an electric field is applied to a feed solution containing negatively charged particles or large organic anions these components will migrate to the anion-exchange membrane and be deposited on its surface. If the polarity is reversed the negatively charged components will migrate away from the anion-exchange membrane back into the feed stream and the membrane properties are restored. This procedure is very effective and is used today in almost all electrodialysis water desalination systems.

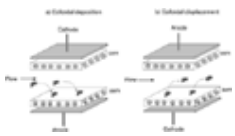

Figure 34. Schematic illustrating the removal of deposited negatively charged colloidal components from the surface of an anion-exchange membrane by reversing the electric field

The degree of desalination that can be achieved in passing the feed solution through a stack is a function of the solution concentration, the applied current density, and the residence time of the solution in the stack, which is determined by the process path length. If the required product concentrations are not achieved in a single process path, electrodialysis can be operated as a multistage process or as a process with feed and bleed in which the diluate or the concentrate or both are partially recycled. Staging can be done by complete physical separation of the stacks, with each having its own set of electrodes, pumps, and power supply, or by internal staging where one set of electrodes and pumps is needed. A disadvantage of simple staging is a constant recovery rate of the feed solution. If variable recovery rates are desired the system must be operated in a feed-and-bleed mode. In this operating mode both the brine and the product concentration can be controlled independently.

4.5.1. *Mass Transport in an Electrodialysis Stack*

The degree of desalination in electrodialysis can be expressed by a material balance between the feed, the concentrate, and the diluate solutions. It is a function of the residence time of the different solutions in the cells, that is, the flow velocities of the solutions, the membrane area in a cell pair, and the applied current density. Calculation of the complete mass balance between the diluate and concentrate flow stream can be rather complex for multicomponent solutions and high ion

concentrations, which means high osmotic pressures and activity coefficients that deviate largely from 1 and when the two flow streams have different flow velocities and directions. However, in many electrodialysis applications the concentrate and diluate cells have identical geometries and flow conditions, that is, the flow in both cells is cocurrent and of identical velocity. The transfer of ions in such a cell arrangement is illustrated in Figure 35.

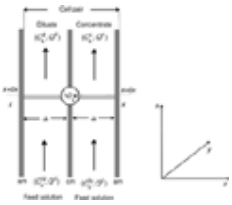

Figure 35. Schematic illustrating the flow of diluate and concentrate streams parallel to the membrane surface in a cell pair of an electrodialysis stack and the concentration change between diluate and concentrate solution ΔC_s due to the ion flux through the membrane

The figure shows a cell pair of an electrodialysis stack with diluate and concentrate flow streams parallel to the membrane surface in the *x* direction. The geometry of both cells are identical and the two flow streams have the same flow velocity. It is assumed that the concentration potential between the two flow streams due to different concentrations in the diluate and concentrate solutions is negligibly low compared with voltage drops due to the resistance of the solutions, and that changes in the resistance of the solutions due to the boundary layer effects are also negligible. Finally, it is assumed that all diffusion processes between the two flow streams can be neglected and the only driving force for the transport is the electrical potential gradient, which is constant over the entire cell length.

Under the above-stated assumptions the degree of desalination of a given feed solution, that is, the changes in the ion concentration in the diluate and concentrate cells in an area element expressed in equivalents, can be calculated as a function of the solution properties, the applied voltage, and the stack design by:

$$\mathrm{d}\,C_s^{\mathrm{d}} = -\mathrm{d}\,C_s^{\mathrm{c}} = \mathrm{d}^{\Delta}C_s = J_s \tag{4.29}$$

Furthermore:

$$J_s = \frac{i\xi}{FQ_{\mathrm{cell}}}\mathrm{d}A_{\mathrm{cell}}, \mathrm{d}A_{\mathrm{cell}} = Y\mathrm{d}x, \quad \text{and} \quad i = \frac{U_{\mathrm{cell}}}{\frac{\Delta\left(C_s^{\mathrm{c}}+C_s^{\mathrm{d}}\right)}{\Lambda_s C_s^{\mathrm{c}} C_s^{\mathrm{d}}} + r^{\mathrm{am}} + r^{\mathrm{cm}}} \tag{4.30}$$

where *i* is the current density, *F* the Faraday constant, ξ the current utilization, U_{cell} the applied voltage across the cell pair, $\mathrm{d}A_{\mathrm{cell}}$ a cell pair area element, Δ the thickness of the cell, Λ_s the equivalent conductivity of the solution (average value for the concentration range considered), r^{am} and r^{cm} are the resistances of the anion- and cation-exchange membranes, and $\mathrm{d}C_s^{\mathrm{d}}$ and $\mathrm{d}C_s^{\mathrm{c}}$ the concentration changes in a volume element of the diluate and concentrate cell, $\mathrm{d}^{\Delta}C_s$ is the concentration change in a volume element of the diluate and concentrate cell due to the ion flux through the membrane, Q_{cell} the flow stream in the diluate or concentrate cell, *Y* the cell width, and d*x* a distance change in direction of the flow stream parallel to the membrane surface.

Introducing Equation (4.29) into Equation (4.30) and integrating over the cell length leads to:

$$\ln\frac{C_s^{\mathrm{c}}C_s^{\mathrm{fd}}}{C_s^{\mathrm{d}}C_s^{\mathrm{fc}}} + \frac{\Lambda_s(r^{\mathrm{am}}+r^{\mathrm{cm}})\left(C_s^{\mathrm{fd}}-C_s^{\mathrm{d}}\right)}{\Delta} = \frac{\Lambda_s\xi U_{\mathrm{cell}}YX}{FQ_{\mathrm{cell}}\Delta} \tag{4.31}$$

where $C_s^{\mathrm{fd}}$and C_s^{fc} are the concentrations of the diluate and the concentrate at the cell inlet, and C_s^{d} and C_s^{c} the concentrations of the diluate and the concentrate at the cell outlet.

The total flow of the diluate or concentrate solution through a stack is given by:

$$Q_{\mathrm{st}} = N\,Q_{\mathrm{cell}} = N\Delta Yu \tag{4.32}$$

where Q_{st} is the total solution flow of the diluate or the concentrate through the entire stack, Q_{cell} the flow through one cell, *N* the number of cells in the stack, Δ the thickness and *Y* the width of a cell, and *u* is the linear flow velocity in the cell.

The total voltage across the stack U_{st} is given by:

$$U_{\mathrm{st}} = NU_{\mathrm{cell}} \tag{4.33}$$

Equations (4.31) to (4.33) give the relations between the various stack construction data such as cell length, cell width, and cell thickness, the solution properties such as the concentrations and conductivities of the feed, the concentrate, and the product, the flow velocity in the cells, and the applied voltage. However, to design an electrodialysis stack for a certain plant capacity on the basis of Equation (4.33) it must be taken into consideration that the current density cannot exceed a certain value, which is referred to as limiting current density. Since the applied voltage is related to the current density it should not exceed a certain value.

The current density and thus the voltage that can be applied in a practical electrodialysis, however, is limited due to concentration polarization and ion depletion at the membrane surfaces in the diluate cell. The causes and consequences of the limiting current density are discussed in more detail below.

Introducing the limiting current density into Equation (4.30) and multiplying by the number of cell pairs in the stack gives the limiting applied voltage across the stack $U_{st,lim}$:

$$U_{st,lim} = i_{lim}\frac{N\Delta}{\Lambda_s}\left[\frac{C_s^d}{C_s^c} + 1 + \frac{\Lambda_s C_s^d}{\Delta}(r^{am} + r^{cm})\right] \tag{4.34}$$

where i_{lim} is the limiting current density, C_s^d and C_s^c are the concentrations in the diluate and concentrate solutions, Λ_s is the equivalent conductivity of the salt solution, r^{am} and r^{cm} are the area resistances of the anion- and cation-exchange membranes, Δ is the cell thickness, and $\Re$ the number of cell pairs in the stack.

Combination of Equations (4.31), (4.33), and (4.34) and rearranging gives the required membrane area of the complete stack as function of the cell geometry, the applied voltage, the concentrations of the feed, concentrate, and diluate solutions, and the solution flow velocities:

$$A_{st} = N_{cell}\frac{\left[\ln\frac{C_s^c C_s^{fd}}{C_s^d C_s^{fc}} + \frac{\Lambda_s(r^{am}+r^{cm})\left(C_s^{fd}-C_s^d\right)}{\Delta}\right]}{\left[\frac{C_s^d}{C_s^c} + 1 + \frac{\Lambda_s C_s^d}{\Delta}(r^{am} + r^{cm})\right]}\frac{Yu\Delta u\Delta_s^d}{\xi\, i_{lim}} \tag{4.35}$$

where A_{st} is the total cell pair area installed in the stack, N_{cell} the number of cell pairs in the stack, C_s^{fd} and C_s^{fc} are the concentrations of the diluate and the concentrate at the cell inlet, C_s^d and C_s^c are the concentrations of the diluate and the concentrate at the cell outlet, Λ_s is the equivalent conductivity of the salt solution, r^{am} and r^{cm} are the area resistances of the anion- and cation-exchange membranes, Δ is the cell thickness, i.e., the distance between the cation- and anion-exchange membrane, Y the cell width, u the linear flow velocity, F the Faraday constant, i_{lim} the limiting current density, and ξ the current utilization.

From the cell pair area required for a certain plant capacity and the cell dimensions the process path length L_{pp} can be determined by:

$$L_{pp} = \frac{A_{cell}}{Y} \tag{4.36}$$

It becomes equal to the cell length X when desalination is achieved in one pass through the cell.

4.5.2. *Energy Requirements in an Electrodialysis Desalination Process*

The total energy requirements in a practical electrodialysis process is the sum of four terms:

- The energy required for the transport of ions from a feed to a concentrate solution.
- The energy necessary to pump the solutions through the stack from a feed solution reservoir to the product tank and brine disposal.
- The energy consumed by the electrode reactions.
- The energy required to operate various process control and measuring instruments.

The last two terms can generally be neglected in larger capacity plants. However, they can be significant in specific applications of small-capacity plants used in the chemical industry or in the treatment of certain wastewater streams.

The required electrical power for the transfer of ions in a desalination process is given by the total current passing through the stack multiplied by the applied voltage. The total current passing through a cell pair is given by:

$$I = \frac{N_{cell}Q_{cell}F\left(C_s^{fd} - C_s^d\right)}{\xi} \tag{4.37}$$

where I is the total electric current passing through the cell pair, Q_{cell} the flow rate of the solution in the diluate cell, C_s^{fd} and C_s^d are the diluate feed and product equivalent concentrations, and N_{cell} is the number of cell pairs in the stack.

The electric power P_{des} required for the desalination process in a stack is given by:

$$P_{des} = N_{cell}U_{cell}I = \frac{FQ_{cell}\left(C_s^{fd} - C_s^d\right)}{\xi}N_{cell}U_{cell} \tag{4.38}$$

Introducing Equations (4.30) and (4.31) into (4.38) and rearranging gives the power requirement for the desalination process in a cell pair:

$$P_{des} = N_{cell}U_{cell}I = \frac{N_{cell}}{A_{cell}}\left[\frac{\Delta\ln\frac{C_s^{fd}C_s^c}{C_s^{fc}C_s^d}}{\Lambda_s\left(C_s^{fd} - C_s^d\right)} + r^{am} + r^{cm}\right]\left[\frac{Q_{cell}^d F\left(C_s^{fd} - C_s^d\right)}{\xi}\right]^2 \tag{4.39}$$

Multiplying the power requirements for one cell pair with the time of operation and the number of cell pairs in the stack gives the energy required for the amount of product obtained in this time:

$$E_{\mathrm{des}} = N_{\mathrm{cell}} U_{\mathrm{cell}} I t = \frac{N_{\mathrm{cell}} t}{A_{\mathrm{cell}}} \left[\frac{\Delta \ln \frac{C_{\mathrm{s}}^{\mathrm{fd}} C_{\mathrm{s}}^{\mathrm{c}}}{C_{\mathrm{s}}^{\mathrm{fc}} C_{\mathrm{s}}^{\mathrm{d}}}}{\Lambda_{\mathrm{s}} \left(C_{\mathrm{s}}^{\mathrm{fd}} - C_{\mathrm{s}}^{\mathrm{d}} \right)} + r^{\mathrm{am}} + r^{\mathrm{cm}} \right] \left[\frac{Q_{cell}^{d} F \left(C_{\mathrm{s}}^{\mathrm{fd}} - C_{\mathrm{s}}^{\mathrm{d}} \right)}{\xi} \right]^2 \tag{4.40}$$

Thus, the specific energy required for a unit volume of product is:

$$E_{\mathrm{des/spce}} = \frac{N_{\mathrm{cell}} U_{\mathrm{cell}} I}{Q_{\mathrm{st}}} = \frac{N_{\mathrm{cell}}}{A_{\mathrm{cell}} Q_{\mathrm{st}}} \left[\frac{\Delta \ln \frac{C_{\mathrm{s}}^{\mathrm{fd}} C_{\mathrm{s}}^{\mathrm{c}}}{C_{\mathrm{s}}^{\mathrm{fc}} C_{\mathrm{s}}^{\mathrm{d}}}}{\Lambda_{\mathrm{s}} \left(C_{\mathrm{s}}^{\mathrm{fd}} - C_{\mathrm{s}}^{\mathrm{d}} \right)} + r^{\mathrm{am}} + r^{\mathrm{cm}} \right] \left[\frac{Q_{\mathrm{cell}}^{\mathrm{d}} F \left(C_{\mathrm{s}}^{\mathrm{fd}} - C_{\mathrm{s}}^{\mathrm{d}} \right)}{\xi} \right]^2 \tag{4.41}$$

where E_{des} and $E_{des/spec}$ are the desalination costs for electrodialysis and the cost per unit volume of product, I is the total current and U_{cell} the voltage per cell, N_{cell} the number of cell pairs in a stack, t the time of operation, C_s^{fd} and C_s^{fc} are the equivalent concentrations of the diluate and the concentrate at the cell inlet, C_s^{d} and C_s^{c} the concentrations of the diluate and the concentrate at the cell outlet, Λ_s is the equivalent conductivity of the salt solution, r^{am} and r^{cm} are the area resistances of the anion- and cation-exchange membranes, Δ is the cell thickness, ζ the current utilization, Q_{cell} and Q_{st} are the diluate flow velocities in a cell and in the entire stack, respectively, and A_{cell} is the cell pair area.

The energy needed to pump the various flow streams through the stack is given by:

$$E_{\mathrm{p/spce}} = \frac{E_{\mathrm{p}}}{Q^{\mathrm{d}} t} = k_{\mathrm{eff}} \frac{\left(Q^{\mathrm{d}} \Delta p^{\mathrm{d}} + Q^{\mathrm{c}} \Delta p^{\mathrm{c}} + Q^{\mathrm{e}} \Delta p^{\mathrm{e}} \right)}{Q^{\mathrm{d}}} \tag{4.42}$$

where $E_{p/spec}$ is the total energy per unit diluate water for pumping the diluate, the concentrate, and the electrode rinse solution through the stack, k_{eff} an efficiency term for the pumps, and Q^d, Q^c, and Q^e are the volume flow rates of the diluate, the concentrate, and the electrode rinse solution through the stack.

The energy consumed by the electrode reaction and required for the operation of the control and measuring instruments is generally neglected or taken as a fixed item which is between 1 and 3 % of the energy used for ion transfer and pumping of the solutions.

4.5.3. *Electrodialysis Process Costs*

The total production costs in electrodialysis are the sum of fixed charges associated with amortization of the plant investment costs and of operating costs such as energy and maintenance requirements. The investment costs include nondepreciable items such as land and depreciable items such as the electrodialysis stacks, pumps, electrical equipment, monitoring and control devices, and membranes. The required membrane area for a given plant capacity is inversely proportional to the current density and directly proportional to the quantity and quality of the product from a given feed solution.

The total operating costs of the electrodialysis plant include energy and maintenance costs and all pre- and post-treatment procedures. They are also a function of the membrane properties, feed and product composition, and several process and equipment design parameters such as stack construction. A dominant role in electrodialysis process costs is played by the applied current density, since it directly affects the investment and energy costs. The required membrane area decreases with increasing current density, while the consumed energy increases with increasing current density, as shown by Equations (4.35) and (4.41). Thus, the total product costs, which are the sum of energy costs, amortization, and maintenance costs, are a function of the current density and will reach a minimum at a certain current density. This is schematically shown in Figure 36.

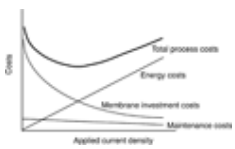

Figure 36. Schematic illustrating the process costs in electrodialysis as a function of the applied current density

Here the total process costs, the energy costs, the amortization on membrane and hardware investment, and the maintenance costs are shown as function of the applied current density. The optimum operating current density in electrodialysis depends to a large extent on the equipment, especially the membrane cost and life, and on the cost of energy. The current density to be applied in electrodialysis is determined by the limiting current density, which should not be exceeded because of possible pH changes in the diluate and concentrate. However, in certain cases it is possible to operate well above the limiting current density [77].

In many applications, electrodialysis competes with other separation processes. Although the theoretically required minimum energy is identical in all processes, the irreversible energy dissipation differs considerably. The desalination of a saline solution by, for example, reverse osmosis involves the passage of water through a membrane under the driving force of a hydrostatic pressure difference, whereas in electrodialysis the salt passes through a membrane under the driving force of an electrical potential difference. The irreversible energy loss in reverse osmosis is caused by hydraulic resistance of the membrane to the water flux and is thus independent of the feed salt concentration. In electrodialysis the irreversible energy loss is proportional to the number of ions transferred from a feed to a concentrate solution, and thus for a given product water concentration it is directly proportional to the feed water concentration. For feed solutions with low salt concentrations the

energy requirements are therefore generally lower for electrodialysis than for reverse osmosis; with feed solutions having a high salt concentration the situation is reversed.

Comparison of energy consumption in mass separation processes must consider the fact that in electrodialysis energy is required in the form of electricity, a relatively expensive form, whereas processes such as distillation use heat, a relatively inexpensive form.

4.6. Bipolar Membrane Electrodialysis Process Design

The basic concept of electrodialytic water dissociation with bipolar membranes is discussed in Section Electrodialysis With Bipolar Membranes. The main application of this process is the production of an acid and a base from the corresponding salt in a cell arrangement consisting of cation- and/or anion-exchange membranes and bipolar membranes in alternating series between two electrodes forming an array of individual cells. A repeating unit may be composed of a three-cell arrangement (Fig. 37a) or a two-cell arrangement (Fig. 37b and c) [78]. In the three-compartment cell design one cation-exchange, one anion-exchange, and one bipolar membrane with three flow streams in between them (i.e., the salt solution, the acid, and the base) represent a cell unit. In the two-compartment cell design a cation-exchange and a bipolar membrane, or an anion-exchange membrane and a bipolar membrane with two flow streams in between are arranged in parallel. One of the flow streams contains the salt solution in a mixture with the produced acid or base. In a three-cell arrangement relatively pure acids and bases are obtained. However, an additional monopolar membrane is needed which will increase the investment cost. The two-cell arrangement results in an acid or a base in a mixture with the salt. The two-cell arrangement is often applied when the product has a relatively low conductivity. This is, for instance, the case when certain organic acids which have a low degree of dissociation such as lactic acid are recovered from a fermentation broth. Without additional salt the poor conductivity in the acid compartment would result in excessive power consumption.

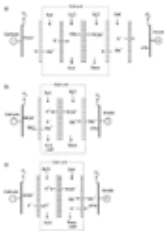

Figure 37. Schematic illustrating the cell arrangement in an electrodialytic water dissociation stack

a) A three-cell unit composed of an anion-exchange, a cation-exchange, and a bipolar membrane; b) A two-cell unit composed of a cation-exchange and a bipolar membrane; c) A two-cell unit composed of an anion-exchange and a bipolar membrane

As in conventional electrodialysis various modes of operation are possible such as batch, feed-and-bleed, and continuous processes [79].

The stack design used in electrodialysis with bipolar membranes is basically the same as that of the sheet-flow stack used in conventional electrodialysis, with the exception that the basic unit of an electrodialysis stack with bipolar membranes consists generally of three cells. Therefore, three separate flow streams must be fed through the manifold in the individual cells that form the repeating unit of the stack. The current density used in electrodialysis with bipolar membranes for the production of acids and bases is generally an order of magnitude higher than in most conventional electrodialysis applications. The voltage drop across a cell unit is also higher than in conventional electrodialysis because of the additional resistance of the bipolar membrane and the concentration potential between the reaction layer of the membrane and the outer phases. As a consequence of the high current density the cell unit area is usually significantly lower in bipolar membrane electrodialysis than that of conventional electrodialysis. Because of the higher voltage drop over a cell unit in electrodialysis with bipolar membranes the number of repeating cell units is also much lower than in conventional electrodialysis.

To design a bipolar membrane electrodialysis plant of a certain capacity and given feed solution and product requirements the necessary membrane area and the applied voltage must be determined. For a stack of three-cell units the required membrane area can be expressed as the cell unit area, which consists of a cation-exchange membrane, an anion-exchange membrane, and a bipolar membrane. The cell unit area is related to the total current required for the production of a certain amount of an acid and a base by the current density.

$$A_{\mathrm{cell}} = \frac{I}{i} \tag{4.43}$$

where A_{cell} is the cell unit area, I is the total current passing through the stack or stacks in series, and i is the average current density.

The total current for a certain production capacity is given by:

$$I = \frac{Q_{\mathrm{p}} F \Delta C_{\mathrm{p}}}{N_{\mathrm{cell}} \xi} \tag{4.44}$$

where Q_{p} is the total product flow, ΔC_{p} the concentration difference between the product solution in the feed at the entrance of the first stack and at the exit of the final stack, N the number of cell units in a stack, F the Faraday constant, and ξ the current utilization.

The average current density can be determined from the applied voltage and the resistance of the cell unit.

When concentration potentials between the three solutions are neglected the total voltage drop consists of three major contributions:

- The potential drop due to the water dissociation equilibrium
- The potential drop due to the resistance of the solutions
- The potential drop due to the resistance of the membranes

A cell unit contains the salt, the acid and base solutions, and the three membranes between these solutions. The total voltage drop is the sum of the water dissociation potential and the voltage drop due to the resistance of the solutions and membranes.

The voltage drop across a cell unit of a stack which consists of geometrically identical cells operated in cocurrent flow is given by:

$$U_{\mathrm{cell}} = \bar{i}\left[\Delta\left(\frac{1}{\Lambda_{\mathrm{s}}\overline{C}_{\mathrm{s}}} + \frac{1}{\Lambda_{\mathrm{b}}\overline{C}_{\mathrm{a}}}\right) + r^{\mathrm{tr}} + r^{\mathrm{am}} + r^{\mathrm{cm}} + r^{\mathrm{bml}}\right] + \frac{RT}{F}\ln\frac{C^{\mathrm{tr}}_{\mathrm{H}^+}C^{\mathrm{tr}}_{\mathrm{OH}^-}}{C^{\mathrm{b}}_{\mathrm{H}^+}C^{\mathrm{b}}_{\mathrm{OH}^-}} \tag{4.45}$$

where U_{cell} is the voltage drop across a cell unit, $\bar{i}$ the average current density along the cell in direction of the flow velocity parallel to the membrane surface, Δ the thickness of the individual cells, $\overline{C}$ the average concentration, Λ the equivalent conductivity, r the area resistance, R the gas constant, T the absolute temperature, F the Faraday constant, superscripts tr, am, cm, bml, and b refer to transition region, cation-exchange membrane, anion-exchange membrane, the two layers of the bipolar membrane, and the bulk solution, respectively, and subscripts s, a, and b to salt, acid, and base, respectively.

The average concentration is given by the integral average over the process path length. The average concentration for each ion is:

$$\overline{C}_i = \frac{C_i^{\mathrm{f}i} - C_i^{\mathrm{p}i}}{\ln\left(\frac{C_i^{\mathrm{f}i}}{C_i^{\mathrm{p}i}}\right)} \tag{4.46}$$

where $\overline{C}_i$ is the average concentration of ion i in a cell, $C_i^{\mathrm{f}i}$ and $C_i^{\mathrm{p}i}$ are the concentrations in the bulk solution at the cell entrance, i.e., the beginning of the process path length and the cell exit, i.e., the exit of the process path length.

Introducing Equation (4.46) into Equation (4.45) leads to:

$$U_{\mathrm{cell}} = \bar{i}\left[\Delta\left(\sum_i \frac{\ln\left(\frac{C_i^{\mathrm{f}i}}{C_i^{\mathrm{p}i}}\right)}{\Lambda_i\left(C_i^{\mathrm{f}i} - C_i^{\mathrm{p}i}\right)}\right) + r^{\mathrm{tr}} + r^{\mathrm{am}} + r^{\mathrm{cm}} + r^{\mathrm{bml}} + \frac{RT}{iF}\ln\frac{C^{\mathrm{tr}}_{\mathrm{H}^+}C^{\mathrm{tr}}_{\mathrm{OH}^-}}{\dfrac{C^{\mathrm{f}\mathrm{H}^+}_{\mathrm{H}^+} - C^{\mathrm{p}\mathrm{H}^+}_{\mathrm{H}^+}}{\ln\dfrac{C^{\mathrm{f}\mathrm{H}^+}_{\mathrm{H}^+}}{C^{\mathrm{p}\mathrm{H}^+}_{\mathrm{H}^+}}} \dfrac{C^{\mathrm{f}\mathrm{OH}^+}_{\mathrm{OH}^+} - C^{\mathrm{p}\mathrm{OH}^-}_{\mathrm{OH}^-}}{\ln\dfrac{C^{\mathrm{f}\mathrm{OH}^-}_{\mathrm{OH}^-}}{C^{\mathrm{p}\mathrm{OH}^-}_{\mathrm{OH}^-}}}}\right] \tag{4.47}$$

where U_{cell} is the voltage drop across a cell unit, $\bar{i}$ is the average current density in the cell unit, and subscript i refers to salt, acid, and base. All other symbols are as defined in Equations (4.44) and (4.45).

The required membrane area and the process path length for a given capacity plant and given feed solution and product concentrations as a function of the current density are derived from Equations (4.43) and (4.44):

$$A = \frac{Q_{\mathrm{p}}F\left(C_{\mathrm{p}}^{\mathrm{fp}} - C_{\mathrm{p}}^{\mathrm{p}}\right)}{\bar{i}\,\xi} \tag{4.48}$$

where A is the cell unit area, Q_{p} is the product volume flow, F is the Faraday constant, $C_{\mathrm{p}}^{\mathrm{fp}}$ and $C_{\mathrm{p}}^{\mathrm{p}}$ are the concentrations of the product at the cell entrance and the product at the cell exit, $\bar{i}$ is the average current density, ξ the current utilization, and N_{cell} the number of cell units in a stack.

For a given stack design and number of cell units in the stack the required process path length is given by:

$$L_{\mathrm{pp}} = \frac{A}{YN_{\mathrm{cell}}} \tag{4.49}$$

where L_{pp} is the process path length, Y the cell width, Q_{p} the product volume flow, F the Faraday constant, $C_{\mathrm{p}}^{\mathrm{fp}}$ and $C_{\mathrm{p}}^{\mathrm{p}}$ are the concentrations of the salt at the beginning and the end of the process path length; and N_{st} is the number of stacks in series.

4.6.1. *Operating Modes in Bipolar Membrane Electrodialysis*

There are several operating modes for the bipolar membrane stack. Depending on the feed solution concentration and the product requirements the process may be operated in a batch or continuous mode. In practical applications it is often operated in a feed-and-bleed mode as illustrated in Figure 38.

Figure 38. Flow scheme of a bipolar membrane electrodialysis stack operated in a feed-and-bleed mode, i.e., with partial recycling of the acid, the base, and the salt solution

With partial recycling of the acid, the base, and the salt solution a constant product concentration can be achieved by adjusting the feed and bleed streams and accordingly changing the recirculation rates of the acid, the base, and the salt solution flow streams.

4.6.2. *Energy Requirements in Electrodialytic Water Dissociation*

The energy required in electrodialytic water dissociation with bipolar membranes is the sum of three terms:

- The electrical energy required to transfer salt ions from the feed solution and protons and hydroxide ions from the bipolar membrane into the acid and base solutions.
- The energy required for the water dissociation in the bipolar membrane.
- The energy consumption for pumping the solutions through the stack.

Energy consumption due to electrode reactions can generally be neglected when more than 50 – 100 cell units are stacked between the two electrodes [72]. The energy required for the dissociation of water at constant temperature and pressure can be calculated for solutions with different H^+ ion activities (i.e., pH values). The Gibbs free energy required for the production of acids and bases in a bipolar membrane is:

$$\Delta G = -RT \ln \frac{C^{\mathrm{tr}}_{\mathrm{H}^+} C^{\mathrm{tr}}_{\mathrm{OH}^-}}{C^{\mathrm{b}}_{\mathrm{H}^+} C^{\mathrm{b}}_{\mathrm{OH}^-}} = -2.3\, RT \Delta\,\mathrm{pH} \tag{4.50}$$

where ΔG is the Gibbs free energy, R the gas constant, T the absolute temperature, F the Faraday constant, $C^{\mathrm{tr}}_{\mathrm{H}^+}$, $C^{\mathrm{tr}}_{\mathrm{OH}^-}$, $C^{\mathrm{b}}_{\mathrm{H}^+}$, and $C^{\mathrm{b}}_{\mathrm{OH}^-}$, are the proton and hydroxide ion concentrations in the transition region of the bipolar membrane and the bulk solutions in contact with the bipolar membrane, and ΔpH is the difference between the pH values of the two solutions separated by the bipolar membrane.

The Gibbs free energy for water dissociation varies slightly, depending on the salt being processed, the concentration of acid and base generated, and the temperature. For the production of 1 N acids and bases at 25 °C the theoretical free energy varies between 0.056 and 0.058 kW·h mol^{-1} and the electromotive force between 2.1 and 2.2 V.

4.6.3. *Total Energy Required for Production of an Acid and a Base*

The total energy required for acid and base production includes energy losses due to the resistances of the solutions and membranes and is given by the current passing through the stack multiplied by the total voltage drop encountered between the electrodes:

$$E_{\mathrm{pro}} = I\,U\,t \tag{4.51}$$

where E_{pro} is the energy consumed in a stack for the production of an acid and a base, I the current passing through a stack or a series of stacks, U the voltage applied across the stack (i.e., between the electrodes), and t the operating time.

The voltage drop in a stack arrangement is caused by the electrical resistances of the solutions and the membranes in the stack and by the potentials which are established between solutions of different ion concentrations. The voltage drop per cell unit under the assumption that the stacks consist of geometrically identical cells which are operated in cocurrent flow with equal velocity is given by Equation (4.47). The required current for the production of a given amount of acid and base is given by Equation (4.44). Introducing Equations (4.44) and (4.47) into Equation (4.51) and considering the voltage drop across the entire stack, the total energy for the production of a given amount of acid and base from the corresponding salt solution is given by:

$$E_{\mathrm{pro}} = I\,N_{\mathrm{cell}}\,U_{\mathrm{cell}}\,t = N_{\mathrm{st}}\bar{i}\left(\frac{\Delta}{\sum_i \Lambda \overline{C}_i} + r^{\mathrm{tr}} + r^{\mathrm{am}} + r^{\mathrm{cm}} + r^{\mathrm{bml}} + \frac{2.3\,RT\overline{\Delta}\mathrm{pH}}{\bar{i}\,F}\right)\left(\frac{Q_p F\left(C^{\mathrm{fp}}_{\mathrm{p}} - C^{\mathrm{p}}_{\mathrm{p}}\right)}{N_{\mathrm{cell}}\,\xi}\right) t \tag{4.52}$$

where E_{pro} is the energy for the production of a certain amount of acid and base, I the current passing through the stack, N_{cell} the number of cell units in a stack, U_{cell} the voltage drop across a cell unit, $\bar{i}$ the average current density, Δ the thickness of the individual cells, $\overline{C}_i$ the average concentration, Λ_i the equivalent conductivity, r the area resistance, ξ the current utilization, R the gas constant, T the absolute temperature, F the Faraday constant, $\overline{\Delta}$ pH the average difference in the pH value between the solution in the transition region of the bipolar membrane and the two adjacent bulk solutions, subscript i refers to salt, acid, and base, superscripts tr, am, cm, and bml to the transition region, the cation-exchange membrane, the anion-exchange membrane, and the two layers of the bipolar membrane, C_p^{fp} and C_p^{p} are the concentrations of the acid or base at the cell inlet and the cell outlet, Q_p is the total flow of the acid or base through the stack, and t the time.

The total current I passing through the stack, which is identical to the current through a cell unit, is related to the average current density $\bar{i}$ and the area of a cell A_{cell} unit by:

$$I = \frac{Q_p F\left(C_p^{fp} - C_p^{p}\right)}{N_{st}\,\xi} = \bar{i}\,A_{cell} \tag{4.53}$$

Introducing Equation (4.53) into (4.52) gives the energy for the production of a certain amount of acid and base:

$$E_{pro} = N_{st}\,A_{cell}\left(\frac{\Delta}{\sum_i \Lambda \overline{C}_i} + r^{tr} + r^{am} + r^{cm} + r^{bml} + \frac{2.3\,RT\overline{\Delta}\mathrm{pH}}{\bar{i}\,F}\right)\left(\frac{Q_p F\left(C_p^{fp} - C_p^{p}\right)}{A_{cell}\,N_{st}\,\xi}\right)^2 t \tag{4.54}$$

where the symbols are as defined in Equation (4.52).

The term $\frac{Q_p F\left(C_p^{fp} - C_p^{p}\right)}{A_{cell}\,N_{st}\,\xi}$ is identical to the current density. This means that for a given stack design the acid and base production energy E_{pro} is proportional to i^2.

The average concentrations of the acid, the base, and the salt in the bulk solutions $\overline{C}_i$ can be expressed by an integral average of the solutions at the cell inlet and the cell outlet C_i^{fi} and C_i^{pi}, respectively, as indicated by Equation (4.47). Thus, the electrical energy consumption for the production of an acid and a base in electrodialytic water dissociation with bipolar membranes can by calculated as a function of the applied current density, the produced acid and base concentrations, and the volume of acid and base, and by the current utilization, which is affected by hydraulic leaks between the cells, current flow through the manifold system, and transport of co-ions, especially protons and hydroxide ions, through membranes that are not strictly permselective. All these effects are expressed in the current utilization ξ.

4.6.4. *Total Process Costs*

The estimation of the total production costs in bipolar membrane electrodialysis follows the general procedure applied in conventional electrodialysis. The main difference is that in addition to the monopolar membranes bipolar membranes are needed and that the energy costs are increased by the additional energy required for water dissociation in the bipolar membrane.

Furthermore, the applied current density is only limited by water diffusion into the bipolar membrane and is generally significantly higher than in conventional electrodialysis. The optimum operating current density in bipolar membrane electrodialysis in practical applications is on the order of 1000 A m^{-2} [80].

4.7. Continuous Electrodeionization Process Design

Continuous electrodeionization is widely used today to produce high-quality deionized water for the preparation of ultrapure water in the electronics industry and in analytical laboratories [81-83].

The process design and the different hardware components needed in electrodeionization are very similar to those used in conventional electrodialysis. The main difference is the stack construction. The diluate cell, and sometimes also the concentrate cell, is filled with an ion-exchange resin, which affects the flow distribution and the pressure drop of the solution in the cell drastically. Therefore, the dimensions of the diluate and concentrate cells are generally quite different. The different concepts used for the distribution of the cations and anions in the cell are illustrated in Figure 39. In the conventional electrodeionization process the diluate cell is filled with a bed of mixed ion-exchange resins with a ratio of cation- to anion-exchange resin close to 1, as shown in Figure 39a. The cation- and anion-exchange resins can also be placed in separate beds in series in a stack as is illustrated in Figure 39b. The main difference between the electrodeionization system with the mixed ion-exchange resin bed and the system with separate beds is that in mixed-bed electrodeionization systems anions and cations are simultaneously removed from the feed while the solution leaving the diluate cell is neutral. In the electrodeionization system with separate anion- and cation-exchange beds the anions are first exchanged by the hydroxide ions generated at the cathode, and the solution leaving the cation-exchange bed is basic. This solution is then passed through the cell with the cation-exchange resin, where the cations are exchanged by the protons generated at the anode. At the exit of the cation-exchange cell the solution is again neutral. The production of the H^+ and OH^- ions can also be accomplished by bipolar membranes instead of electrodes [84].

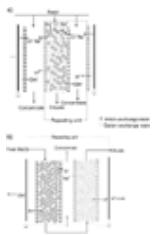

Figure 39. Schematic illustrating two different stack concepts used in continuous electrodeionization. a) A stack in which a repeating unit consists of a diluate cell filled with a mixed bed of ion-exchange resins and a concentrate cell; b) A repeating unit of a stack with a concentrate cell between two diluate cells, one filled with a cation-exchange resin and the other with an anion-exchange resin, between two electrodes

4.8. Other Membrane Processes

Ion-exchange membranes are used as fuel cell separators and in combination with electrochemical reactions in electrochemical synthesis such as the production of chlorine and alkali [85].

Facilitated-transport and membrane reactors are still in an early state of development and so far used mainly on a laboratory scale. The same is true for membrane distillation and membrane contactors.

[Top of Page]

5. Membrane Preparation and Membrane Module Constructions

The most important part in any membrane separation process is the membrane itself. As described in Section Materials and Structures of Synthetic Membranes membranes vary widely in structure, function, transport properties, transport mechanism, and the materials from which they are made. Just as diverse as the different membranes are the methods of making them. Some membranes are manufactured by simple sinter techniques from fine powders, others are prepared by irradiation and track etching of thin films or by inversion of homogeneous liquid mixtures or melts into heterogeneous solid phases. Composite membranes are prepared by dip-coating techniques, interfacial polymerization, and plasma polymerization. Inorganic membranes are prepared as composite structures by using powder sintering techniques and the sol – gel method. Liquid membranes with mobile selective carriers are prepared as emulsions or supported in porous structures. Fixed-carrier and ion-exchange membranes are prepared by introducing positively or negatively charged groups into suitable polymer structures.

Closely related to preparation of membranes is their characterization. In addition to the determination of their transport properties, membranes are also characterized by various methods in terms of their morphology and their mechanical, chemical, and electrochemical properties. Which of the different membrane characterization techniques are used depends on the membrane type and its possible application.

5.1. Membrane Preparation

The selection of a suitable base material and preparation technique depends on the application the membrane is to be used in. In some applications such as gas separation and pervaporation the membrane material used as the barrier layer is of prime importance for the performance of the membrane. In other applications such as micro- and ultrafiltration the membrane material is not quite as important as the membrane structure. Membrane preparation procedures are often developed empirically and described in the literature as detailed recipes [17, 86-89].

5.1.1. *Preparation of Porous Membranes*

Porous membranes consist of a solid matrix with defined pores which have diameters ranging from less than 2 nm to more than 20 µm. They can be made from various materials, such as ceramics, graphite, metal, metal oxides, and polymers. Their structure may be symmetric (i.e., pore diameters do not vary over the membrane cross section) or asymmetric (i.e., pore diameters increase from one side of the membrane to the other by a factor of 10 – 1000. Porous membranes are used mainly in microfiltration and ultrafiltration. They can be classified according to:

- The material they are made from
- Their structure (symmetric or asymmetric)
- Their pore size and pore size distribution

The membrane material determines the mechanical properties and the chemical stability of the membrane. The chemical properties of the material such as its hydrophilicity or hydrophobicity affect its performance in a practical application due to specific adsorption of constituents of the feed mixture.

The methods used most widely today for making porous membranes are sintering of powders, stretching of films, irradiation and etching of films, and phase inversion or sol – gel processes. The different structures, the material from which they are made, their preparation, and their typical applications are summarized in Table 2.

Table 2. Porous membranes and their preparation and applications

Membrane type	Membrane material	Pore size, µm	Preparation process	Applications
Symmetric	ceramic, metal, polymer, graphite	0.1 – 20	powder pressing and sintering	microfiltration, gas separation
Symmetric	polymer with partial crystallinity	0.2 – 10	extruding and stretching of films	microfiltration, battery separator
Symmetric	polymer, mica	0.05 – 5	irradiation and etching of films	microfiltration, point-of-use filters
Symmetric	polymer, metal, ceramic	0.5 – 20	template leaching of films	microfiltration
Symmetric	polymer	0.5 – 10	temperature-induced phase inversion	microfiltration
Asymmetric	polymer	< 0.01	diffusion-induced phase inversion	ultrafiltration
Asymmetric	ceramic	< 0.01	composite-membrane gel – sol process	ultrafiltration

Sintering is a rather simple technique to obtain porous structures from organic and inorganic materials. A powder consisting of particles of a certain size is pressed into a film or plate and sintered just below the melting point of the material. Figure 40a

shows the surface of a porous membrane made by pressing and sintering a plate of fine polytetrafluoroethylene powder. The process yields a porous structure of relatively low porosity in the range of 10 – 40 %, and a rather irregular pore structure with a very wide pore size distribution. Material selection for the preparation of sintered membranes is determined mainly by the required mechanical properties and chemical and thermal stability of the material in the application of the final membrane. Sintered membranes are made on a large scale from ceramic materials such as aluminum oxides, graphite, and metal powders such as stainless steel and tungsten. The particle size of the powder is the main parameter determining the pore sizes of the final membrane, which typically range from 0.2 to 20 µm. The lower limit of the pore diameter is determined by the particle size of the powder. Sintered membranes can be made in the form of disks, cartridges, or fine-bore tubes. They are used for the filtration of colloidal solutions and suspensions. They are also suitable for gas separation and widely used for the separation of radioactive isotopes.

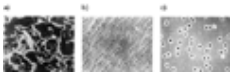

Figure 40. Scanning electron micrographs of a) a sintered membrane prepared from a polymer powder, b) a membrane prepared by stretching an extruded polytetrafluoroethylene film perpendicular to the direction of extrusion, and c) a capillary-pore polycarbonate membrane made by track etching

Porous carbon membranes are also prepared by *pyrolysis* of preformed polyacrylonitrile membranes in an inert atmosphere at 600 – 800 °C. The membranes are generally coated on a porous ceramic support or prepared as hollow fibers. Their pore size is 1 – 4 nm. The membranes are used for both ultrafiltration and gas separation. However, they are very brittle and thus difficult to handle in large-scale practical applications.

Another relatively simple procedure for preparing porous membranes is the *stretching* of a homogeneous polymer film with partial crystallinity [90]. This technique is mainly employed with films of polyethylene or polytetrafluoroethylene which have been extruded from a polymer powder at temperatures close to its melting point and rapidly drawn down. The crystallites in the semicrystalline polymer are aligned in the direction of drawing. After annealing and cooling the extruded film is stretched perpendicular to the drawing direction. This leads to partial fracture of the film, and relatively uniform pores with diameters of 0.2 – 20 µm are obtained. A typical stretched membrane prepared from polytetrafluoroethylene is shown in Figure 40b. The membranes can be produced as flat sheets, tubes, and capillaries. Because of the hydrophobic nature of the polymer these membranes are highly permeable for gases and vapors but impermeable for aqueous solutions. They are used in sterile filtration, blood oxygenation, and membrane distillation, and as water-repellent textiles. Because of its high porosity this membrane has a high gas and vapor permeability, but up to a certain hydrostatic pressure it is completely impermeable to aqueous solutions.

Porous membranes with very uniform, nearly perfectly round cylindrical pores are obtained by *track etching* [91]. In the first step, a homogeneous 6 – 15 µm thick polymer film is exposed to collimated charged particles in a nuclear reactor. As particles pass through the film, they leave sensitized tracks where the chemical bonds in the polymer backbone are damaged. In the second step, the irradiated film is placed in an etching bath, in which the damaged material along the tracks is preferentially etched to form uniform cylindrical pores. The pore density of a track-etched membrane is determined by the residence time in the irradiator, while the pore diameter is controlled by the residence time in the etching bath. The minimum pore diameter of these membranes is approximately 0.01 µm. The maximum pore size that can be achieved in track-etched membranes is determined by the etching procedure and is about 5 µm. Figure 40c shows a typical track-etched polycarbonate membrane. Capillary-pore membranes are prepared today mainly from polycarbonate and polyester films. The advantage of these polymers is that they are commercially available in very uniform films of 10 – 15 µm thickness, which corresponds to the maximum penetration depth of collimated particles from a nuclear reactor. Because of their narrow pore size distribution and low tendency to plug, capillary-pore membranes made from polycarbonate and polyester have found applications on a large-scale in analytical chemistry and microbiological laboratories, and in medical diagnostic procedures. Capillary-pore membranes are used on an industrial scale for the production of ultrapure water for the electronics industry. Here, they show certain advantages over other membrane products because of their short “rinse down” time and good long-term flux stability. Because of their surface filter characteristics, particles retained by the membranes can be further monitored by optical or scanning electron microscopy.

Other techniques to produce porous microfiltration membranes are based on *microlithography* and *template leaching*. These membranes often have a narrow pore size distribution and high fluxes. The template-leaching technique is also applied to prepare membranes from glass, metal alloys, and ceramics [92]. The preparation procedure for porous glass and metal membranes is relatively simple: Two different types of glass or metals are homogeneously mixed; then, one type is dissolved and a network with well-defined pore sizes of the undissolved material is obtained.

Most commercially available symmetric microporous membranes are prepared by the *phase-inversion process* in which a polymer is dissolved in an appropriate solvent and spread as a 20 – 200 µm thick film on a plate, a belt, or a fabric support. A precipitant such as water is added to this liquid film and causes separation of the homogeneous polymer solution into a solid polymer-rich and a liquid solvent-rich phase. The precipitated polymer forms a porous structure containing a network of more or less uniform pores. A microporous cellulosic membrane made by phase inversion is shown in Figure 41a.

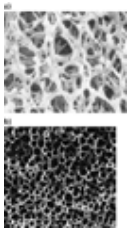

Figure 41. Scanning electron micrograph of the surface of a) a porous cellulose nitrate membrane prepared from a homogeneous polymer solution by precipitation with water vapor, and b) a porous polypropylene membrane precipitated by thermal gelation from a hot polymer solution

Porous phase-inversion membranes can be made from almost any polymer which is soluble in an appropriate solvent and

can be precipitated in a nonsolvent. By varying the polymer, the polymer concentration, the precipitation medium, and the precipitation temperature, porous phase-inversion membranes can be made with a very large variety of pore sizes, from less than 0.1 to more than 20 µm, and with varying chemical and mechanical properties. These membranes were originally prepared from cellulosic polymers by precipitation at room temperature in an atmosphere of approximately 100 % R.H. More recently, symmetric microporous membranes have also been prepared from various polyamides, polysulfone, and poly (vinylidene difluoride) by precipitation of a cast polymer solution in water. The phase-inversion process is today the most important technique for obtaining porous structures. Polypropylene or polyethylene can also be used for the preparation of porous membranes. However, since these polymers are not readily dissolved at room temperature, the preparation technique must be slightly varied. Polypropylene, for example, is dissolved in an appropriate amine at elevated temperatures. A solution of 20 – 30 % polymer is spread at elevated temperature into a film. The precipitation of the polymer, however, is not induced by the addition of a nonsolvent but merely by cooling the solution to a point at which a two-phase system forms. The resulting open foam structure is shown in Figure 41b. The pore size of membranes depends on the polymer concentration, the solvent system, the solution temperature, and the cooling rate. This membrane preparation technique is usually referred to as *thermal gelation*.

The symmetric, porous polymer membranes made by phase inversion are widely used for separations on laboratory and industrial scale. Typical applications range from the clarification of turbid solutions to the removal of bacteria or viruses, the detection of pathological components, and the detoxification of blood in artificial kidneys. The separation mechanism is that of a typical depth filter which traps the particles somewhere within the structure. In addition to the simple “sieving” effect, porous phase-inversion membranes often show a pronounced tendency for adsorption because of their extremely large internal surface area. They are therefore particularly well suited when complete removal of components such as viruses or bacteria is desired. They are suited for immobilization of enzymes for use in biotechnology. They are also widely used for culturing microorganisms in water quality control tests.

5.1.2. *Preparation of Asymmetric Membranes*

Most membranes used today in large-scale separation processes have an asymmetric structure which consists of a very thin (0.1 – 1 µm) selective skin layer on a highly porous, ca. 100 – 200 µm thick substructure (Fig. 42). The very thin skin represents the actual membrane. It can consist of homogeneous polymer or can contain pores. Its separation characteristics are determined by the nature of the polymer or the pore size in the skin, while the mass transport rate is determined by the thickness of the skin. The porous substructure serves only as a support for the thin and sometimes fragile skin and has little effect on separation characteristics or the mass transfer rate of the membrane.

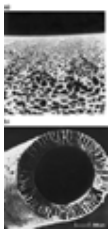

Figure 42. Scanning electron micrographs showing cross sections of a) an asymmetric reverse osmosis membrane with a graded pore structure and b) an asymmetric ultrafiltration capillary membrane with a finger-type structure

Asymmetric membranes are used primarily in pressure-driven membrane processes such as reverse osmosis, ultrafiltration, and gas separation, as it is here that the unique properties of high mass transfer rates combined with good mechanical stability can be best utilized. In addition to high filtration rates, asymmetric membranes are very resistant to fouling. Conventional symmetric structures act as depth filters and retain particles within their internal structure. These trapped particles plug the membrane, and so the flux declines during use. Asymmetric membranes are surface filters and retain all rejected materials at the surface where they can be removed by shear forces applied by the feed solution moving parallel to the membrane surface. Two techniques are used to prepare asymmetric membranes: One technique utilizes the phase-inversion process and leads to a membrane in which the skin and the substructure consist of the same polymer (integral asymmetric membrane). In the second technique an extremely thin polymer film is deposited on a preformed porous substructure to give composite membranes. The development of the first integral asymmetric membranes by phase inversion was a major breakthrough in the development of ultrafiltration and reverse osmosis. These membranes were made from cellulose acetate and yielded fluxes 10 – 100 times higher than symmetric structures with comparable separation characteristics. Today, however, most reverse osmosis membranes are composite structures.

Asymmetric phase-inversion membranes can be prepared from virtually all polymers that are soluble at a certain temperature in an appropriate solvent or solvent mixture and can be precipitated as a continuous solid phase by either changing the temperature or the composition of the system by the following general procedures [93-97]:

- Cooling of a homogeneous polymer solution which separates at a certain temperature into two phases.
- Evaporation of a volatile solvent from a homogeneous polymer solution with two or more solvents of different dissolution capacity.
- Addition of a nonsolvent or nonsolvent mixture to a homogeneous solution.

All three procedures result in the formation of two phases, that is, a liquid phase that forms the membrane pores and a solid phase that forms the membrane structure, which may be either symmetric and porous or asymmetric with a dense skin on one or both surfaces of a porous bulk phase. The only thermodynamic prerequisite for all three preparation procedures is that the system has a miscibility gap over a defined concentration and temperature range.

A phase separation that is the result of a temperature change is called temperature-induced phase separation, and a phase separation caused by the addition of a nonsolvent or nonsolvent mixture to a homogeneous solution is referred to as diffusion-induced phase separation. In practical membrane preparation combinations and certain variations of the basic

processes are used to produce membranes with tailor-made structures and properties.

The *diffusion-induced phase-separation process* consists of the following consecutive steps:

- A polymer is dissolved in an appropriate solvent to form a solution.
- The solution is cast into a film on the order of 100 – 500 µm thickness.
- The film is quenched in a nonsolvent, typically water or an aqueous solution.

During the quenching process the polymer solution separates into two phases: a polymer-rich solid phase which forms the membrane structure, and a solvent-rich liquid phase which forms the liquid-filled membrane pores. Generally, the pores at the film surface, where precipitation occurs first and most rapidly, are much smaller than in the interior or on the bottom side of the film. This leads to the asymmetric membrane structure.

Different variations on this general preparation procedure are described in the literature. For example, sometimes an evaporation step prior to the precipitation is used to change the composition in the surface of the cast film, and an annealing step is applied to change the structure of the precipitated membrane [98], [99].

The original recipes and subsequent modifications for preparing asymmetric membranes are deeply rooted in empiricism. Detailed descriptions of membrane preparation techniques are given in the literature [93, 100-103]. Only after extensive use of the scanning electron microscope, which provided the necessary structural information, was it possible to rationalize the various parameters responsible for the processes forming the membrane structure. It then became evident that both symmetric and asymmetric porous structures can be prepared by the diffusion-induced phase-separation process from a large number of polymers when the appropriate preparation parameters are chosen. The actual phase separation or phase inversion can be induced not only by addition of nonsolvent but also by controlled evaporation of a volatile solvent from a three-component mixture of solvent – precipitant – polymer, causing precipitation as the system becomes enriched in precipitant [104]. Alternatively, precipitation of a simple two-component polymer – solvent casting solution can be brought about by imbibing precipitant from the vapor phase [17]. This technique was the basis of the original porous membranes and is still used commercially today by several companies.

In the *temperature-induced phase-separation process* the precipitation of a casting solution is achieved by cooling a polymer solution which forms a homogeneous solution only at elevated temperature, for example polypropylene dissolved in *N,N*-bis-(2-hydroxyethyl)tallow amine [95]. The temperature-induced phase-separation process generally results in a symmetric porous structure. Under certain experimental conditions it can also result in asymmetric structures [97]. Temperature-induced phase separation is not only applicable to polymers. It is also used for the preparation of porous membranes from glass mixtures and metal alloys in combination with a leaching procedure.

5.1.3. *Rationalization of the Phase-Inversion Membrane Preparation Process*

Although the recipes given in the literature for the preparation of porous structures from polymers, metals, or glasses by phase inversion are very different they are all based on similar thermodynamic and kinetic parameters, such as the chemical potentials and diffusivities of the individual components and Gibbs free energy of mixing of the entire system. Identification of the various process parameters is the key to understanding the mechanism of membrane formation, which is a necessity for optimizing membrane properties and structures.

A quantitative treatment considering all thermodynamic and kinetic parameters involved in the phase-inversion process is difficult. But a phenomenological description of the process with the aid of the phase diagram of a mixture consisting of a polymer and one or more solvents and nonsolvents at constant or different temperatures is very useful for a better understanding of the relation between membrane structure and the different preparation parameters. However, the phase diagram is the thermodynamic description of an equilibrium state. In the membrane formation process phase separation is also determined by kinetic parameters, and thermodynamic equilibrium is generally not attained on a macroscopic scale. The quantitative description of the kinetics is difficult. However, a general thermodynamic description of the phase separation based on the phase diagram of the polymer – solvent – nonsolvent system provides valuable information concerning the membrane structures obtained by the phase-inversion process.

Temperature-induced phase separation of a two-component polymer mixture is illustrated in Figure 43a, which shows a phase diagram of a two-component mixture of a polymer and a solvent as a function of temperature. The diagram indicates a miscibility gap over a wide range of composition at low temperature. Above a certain temperature the polymer and solvent form a homogeneous solution at all compositions. At other temperatures the system is not stable at certain compositions and will separate into two phases. This region is referred to as the miscibility gap, which is surrounded by the binodal curve. If a homogeneous polymer – solvent mixture of a certain composition at temperature T_1, as indicated by point A in Figure 43a, is cooled to temperature T_2, as indicated by point B, it will separate into two different phases, the composition of which is indicated by points B′ and B″. Point B″ represents the polymer-rich phase and point B′ the solvent-rich polymer-lean liquid phase. The lines B′ – B and B″ – B represent the ratio of the amounts of the two phases in the mixture, that is, the overall porosity of the obtained porous system. When the polymer concentration in the polymer rich phase has reached a certain value its viscosity has increased to such an extent that it can be considered as solid. The polymer-rich phase forms the solid membrane structure and the polymer-lean phase the liquid-filled pores.

Figure 43. Schematic showing the formation of a microporous system by a) thermal precipitation of a two-component mixture and b) by addition of a nonsolvent to a homogeneous polymer solution in a three-component mixture exhibiting a miscibility gap under certain conditions of temperature and composition

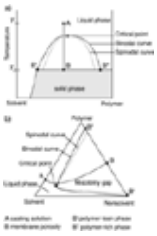

Phase separation induced by the addition of a nonsolvent to a homogeneous polymer solution is illustrated in Figure 43b, which shows a three-component isothermal phase diagram. The three-component mixture exhibits a miscibility gap over a wide range of compositions. If a nonsolvent is added to a solution consisting of polymer and solvent with composition indicated by point A on the solvent – polymer line and the solvent is removed from the mixture at about the same rate as the nonsolvent enters, the composition of the mixture will change following the line A – B. When the composition of the system reaches the miscibility gap it will separate into two phases: a polymer-rich phase represented by the upper boundary of the miscibility gap which is determined by the binodal curve, and a polymer-lean phase represented by the lower boundary of the miscibility gap. When the solvent is replaced completely by the nonsolvent the mixture reaches the composition represented by point B. Point B represents a mixture of the solid polymer-rich phase and a liquid polymer lean phase of compositions B′ and B″, respectively. Instead of adding a nonsolvent to a polymer – solvent mixture, phase separation can also be achieved by evaporation of a solvent from a solvent – polymer – nonsolvent mixture which is used as casting solution for the membrane [104].

The description of the formation of porous systems by means of the phase diagrams, as illustrated in Figure 43, is based on the assumption of thermodynamic equilibrium. It predicts under which conditions of temperature and composition a system will separate into two phases. But it does not provide any information on the size of the domains of the two phases and their shapes, that is, the pore size, the pore shape (i.e., “finger”- or “sponge”-like), and the pore size distribution of the membrane. These parameters are determined by kinetic parameters such as the diffusivities of the various components in the mixture, the viscosity of the solution, and the chemical potential gradients which act as driving forces for the diffusion of the various components in the mixture. Because these parameters change continuously during phase separation, which constitutes the actual membrane formation process, no transient states of equilibrium will be achieved. Therefore, the kinetic parameters are difficult to determine by independent experiments and thus are not readily available. This makes a quantitative description of the membrane formation mechanism very difficult.

By the phase-inversion process both symmetric and asymmetric membranes with rather different structures are produced commercially from a variety of polymers (Table 3).

Table 3. Polymers used for commercial membrane production by phase inversion and their applications

Membrane material	Membrane processes[a]
Cellulose acetate	EP, MF, UF, RO
Cellulose esters (mixed)	MF, D
Polyacrylonitrile (PAN)	UF
Polyamide (aromatic, aliphatic)	MF, UF, RO
Polyimide	UF, RO, GS
Polypropylene	MF
Polyethersulfone	UF, MF, GS, D
Polysulfone	UF, MF, GS, D
Sulfonated polysulfone	UF, RO, NF
Poly(vinylidene fluoride)	UF

[a]D = dialysis, EP = electrophoresis, GS = gas separation, MF = microfiltration, NF = nanofiltration, RO = reverse osmosis, UF = ultrafiltration.

5.1.4. *Preparation of Composite Membranes*

In processes such as reverse osmosis, gas separation, and pervaporation, the actual mass separation is achieved by a solution-diffusion mechanism in a homogeneous polymer layer. Since the diffusion process in a homogeneous polymer matrix is relatively slow, these membranes should be as thin as possible. Therefore, an asymmetric membrane structure is mandatory for these processes. Unfortunately, many polymers with satisfactory selectivities and permeabilities for the various components in gas mixtures or liquid solutions are not well suited for the phase-inversion membrane preparation process. This has led to the development of the composite membranes. A typical composite membrane is shown schematically and as a scanning electron micrograph in Figure 44. It is composed of a 20 – 1000 – nm thick dense polymer barrier layer formed over an approximately 50 – 100 µm thick porous film. Composite membranes are prepared in a two step process: 1) manufacture of the porous support, and 2) deposition of the barrier layer on the surface of this porous support layer.

Figure 44. a) Schematic of a composite membrane showing the porous support structure and the selective skin layer, and b) scanning electron micrograph of a composite membrane with polydimethylsiloxane as the selective layer on a polysulfone support structure

The advantages of composite membranes over integral asymmetric structures is that different polymers can be used for the porous support and the selective barrier layer. This means that polymers which show the desired selectivity for a certain separation problem but are not suited for fabrication into integral asymmetric membranes because of poor mechanical strength or poor film-forming properties, may be usable as the selective barrier in composite membranes. And polymers suited for the preparation of porous structures but without the required properties for a certain separation task can be used as support structure. This expands the variety of available materials for the preparation of membranes with properties required for a given separation task considerably.

Making composite membranes involves two steps. One is the preparation of a suitable porous support and the second is the preparation of the actual barrier layer on the surface of the support structure. The performance of a composite membrane is not only determined by the properties of the selective barrier layer, but it is also significantly affected by properties of the porous support. Although the selectivity of a composite membrane is determined almost exclusively by the actual barrier layer, its flux rate is also to some extent determined by the pore size and overall porosity of the substructure. The effect of pore size and overall porosity is illustrated in Figure 45, which shows an idealized homogeneous barrier layer on a porous support.

The actual diffusion pathway of the component through the barrier layer into a pore of the support structure is always longer than the thickness of the barrier layer z_o but shorter than the longest path z_{max}.

Figure 45. Schematic illustrating the diffusion path of a component through a selective barrier layer of a composite membrane (z_o represents the shortest and z_{max} the longest path)

If it is assumed that the transport in the support structure is exclusively through the pores the effective path length can be expressed as a function of the overall membrane porosity, the thickness of the barrier layer, and the radius of the pores in the support structure:

$$z_{\mathrm{eff}} = \omega z_o + (1-\omega)\frac{1}{2}\left(\sqrt{r^2\left(\frac{1-\omega}{\omega}\right)} + z_o^2 + z_o\right) \tag{5.1}$$

where z_{eff} and z_o are the effective diffusion path length and the thickness of the barrier layer, respectively, r is the radius of the pores in the support structure, and ω the surface porosity of the support structure.

This relationship, which can be obtained by simple geometric considerations [68], indicates that the effective diffusion path length z_{eff} is strongly affected by the surface porosity of the substructure. According to Equation 5.1 the effective diffusion length is approximately equal to the thickness of the barrier layer at relatively high surface porosity (> 50 %). However, it increases by an order of magnitude when the porosity is less than 1 % if the barrier layer thickness is assumed to be approximately the same as the pore radius. Accordingly, the flux decreases by the same magnitude. The geometrical consideration expressed in Equation 5.1 indicates that the surface porosity of the support layer significantly affects the membrane flux. The porosity should always be as high as possible to achieve optimal flux rates with thin-film composite membranes. To ensure adequate mechanical strength, the pore diameter should not be significantly greater than the film thickness.

The techniques used for the preparation of composite membranes can be grouped into four general procedures:

- Casting of the barrier layer, for example, on the surface of a water bath and then laminating it on the porous support film.
- Coating of the porous support film by a polymer, a reactive monomer, or prepolymer solution followed by drying or curing with heat or radiation.
- Gas-phase deposition of the barrier layer on the porous support film from a glow-discharge plasma.
- Interfacial polymerization of reactive monomers on the surface of the porous support film.

Casting an ultrathin film of cellulose acetate on a water surface and transferring the film onto a porous support was one of the earliest techniques used for preparing composite reverse osmosis membranes. However, this technique is no longer used today [30].

Coating a porous support structure by dipping it into a polymer or prepolymer solution is today applied mainly for the preparation of composite membranes to be used in gas separation and pervaporation. In particular, polymers such as polydimethylsiloxane which are available as soluble prepolymers and which can easily be cross-linked by heat treatment and thus become insoluble in most solvents are suited for the preparation of this type of composite membrane. If the pore dimension in the support membrane is selected properly, the prepolymer is unable to penetrate the support, and a rather thin uniform barrier layer of 0.05 – 1 µm thickness can easily be prepared. A typical composite membrane prepared by dip coating can be seen in Figure 44b, which shows a scanning electron micrograph of an asymmetric polysulfone ultrafiltration membrane which was coated with a 1 wt % solution of polydimethylsiloxane in an appropriate solvent followed by thermal cross-linking [68].

Gas-phase deposition of a barrier layer on a dry porous support membrane by plasma polymerization was also successfully used for the preparation of reverse osmosis membranes. Although reverse osmosis membranes with excellent desalination properties showing salt rejection in excess of 99 % and fluxes of 1.2 $m^3\ m^{-1}\ d^{-1}$ when tested with seawater have been prepared on a laboratory scale by plasma polarization, it is not used in large-scale industrial production of composite membranes [105].

Today, by far the most important technique for preparing composite membranes is *interfacial polymerization of reactive monomers* on the surface of a porous support film. The first membrane produced on a large scale with excellent reverse osmosis desalination properties was developed in the early seventies in the North Star Research Institute under the code name NS 100 [106, 107]. The preparation procedure of this membrane, which exhibited water fluxes of about 1 $m^3\ m^{-1}\ d^{-1}$ and salt rejections in excess of 99 % when tested with seawater at 6 MPa pressure, was rather simple [108]. A polysulfone support membrane was soaked in an aqueous solution of 0.5 – 1 % polyethyleneimine, which was treated interfacially at the membrane surface with a 0.2 – 1 % solution of toluene diisocyanate in hexane. A heat curing step at 110 °C leads to further cross-linking of the polyethyleneimine. The preparation of composite membranes based on piperazine and trimesoyl chloride is shown schematically in Figure 46.

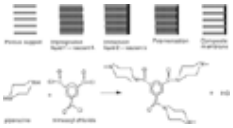

Figure 46. Schematic showing the formation of a composite membrane by interfacial polymerization of piperazine with trimesoyl chloride

The process seems to involve two types of reactions. In a first step, the piperazine reacts rapidly at the interface with the trimesoyl chloride to form a polyamide surface skin, while amino groups below this surface remain unconverted. In the second, heat-treatment step, internal cross-linking of these amino groups takes place. Thus, the final membrane has three distinct layers of increasing porosity: 1) the dense polyamide surface skin, 2) a thin cross-linked piperazine layer which extends into the pores of the support film, and 3) the actual polysulfone support membrane.

5.1.5. *Preparation of Inorganic Composite Membranes*

Ceramic membranes for use in micro-, ultra-, and nanofiltration as well as in pervaporation and gas separation are also prepared as multilayer composite structures [109]. A typical three-layer structure of an ultra- or nanofiltration membrane is shown schematically in Figure 47a and as a scanning electron micrograph in Figure 47b. The first support layer is generally prepared from a suspension of Al_2O_3 powder by using an organic polymer such as cellulose or poly(vinyl alcohol) as a binder and to increase the viscosity of the suspension. The suspension can be processed by slip casting or extrusion into flat sheets, hollow fibers, or tubes to form the so-called green structure, which is dried and converted into the final membrane by sintering at about 1500 °C to give a porous structure of α-Al_2O_3 with an average pore diameter of 5 µm or more and an overall porosity of about 35 %. The process is shown in Figure 48. A second layer, prepared by a sol – gel process, is deposited on the support layer by dip coating or filtration. The sol used in this process consists of colloidal particles, as indicated in the schematic of Figure 49. The material used for the preparation of the second layer is usually γ-Al_2O_3 or TiO_2. The pore size of this layer is in the range of 5 – 10 nm. Thus, the two-layer membrane can be used for micro- and ultrafiltration applications. For the separation of lower molecular weight components in nanofiltration and pervaporation a third layer is deposited. This layer is also prepared by the sol – gel process. However, the sol contains an inorganic polymer, as indicated in Figure 49, which becomes a polymer gel consisting of a three dimensional network. As material for the inorganic polymer structure TiO_2, ZrO_2, or SiO_2 is often used.

Figure 47. a) Schematic illustrating the structure of a three-layer inorganic composite membrane, and b) scanning electron micrograph of a three-layer ceramic membrane

Figure 48. Porous inorganic membrane preparation scheme based on slip casting of a ceramic powder and sintering at 1500 – 1800 °C

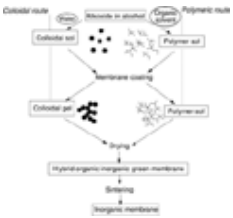

Figure 49. Schematic illustrating the colloidal and polymeric sol – gel process used in the preparation of composite inorganic membranes

A suspension is a dispersion of solid particles in a liquid; a sol is a suspension of colloidal particles or polymers. The main difference between suspension and sol is the size of the dispersed particles. Suspensions are prepared from a fine powder of aluminum oxide with an average particle diameter of 5 – 10 µm. The dispersion is mixed with an organic polymer such as poly(vinyl alcohol) or a cellulose derivative and poured in a mold or extruded in the desired shape as a flat sheet, tube, or capillary, dried, and sintered at 1500 – 1800°C. The process scheme is shown in Figure 48.

One surface of the support structure is then coated again with a suspension of finer particles, dried, and sintered again. Sometimes, more than one coating is applied to obtain the desired pore size at the surface of the membrane. The process of preparing a multilayer membrane by using suspensions of different particle sizes is referred to as slip coating. With the slip-coating process membranes with pore sizes down to 20 nm for use in microfiltration can be made. To obtain membranes with smaller pores a soll – gel process is applied and a colloidal or polymeric gel is used, which is prepared by controlled hydrolysis of a metal alkoxide to a hydroxide. The process can follow two different process paths. One path is referred to as the colloidal and the other the polymeric process path. Both pathways are illustrated in Figure 49.

In the colloidal process path a metal oxide such as aluminum oxide is dissolved in alcohol and then hydrolyzed and precipitated by addition of excess water at elevated temperature to form a stable colloidal solution. The solution is cooled to form a gel, which is then coated onto an appropriate support and sintered at about 800 °C. The average pore diameter obtained in the colloidal sol – gel process is in the range of 2 – 5 nm. The overall process consists of three steps:

Metal alkoxide formation: $Al_2O_3 + \text{alcohol} \longrightarrow Al(OR)_3$

Precipitation: $Al(OR)_3 + H_2O \longrightarrow Al(OH)_3 \longrightarrow \gamma\text{-}Al_2O_3 \cdot H_2O$

(Böhmite)

Coating and sintering: $\gamma\text{-}Al_2O_3 \cdot H_2O \longrightarrow \gamma\text{-}Al_2O_3$

In the polymeric sol – gel process the metal alkoxide is only partially hydrolyzed in an alcohol solution by a controlled addition of a minimum of water. The hydroxyl groups of the alkoxide react with each other and form an inorganic/organic polymer which forms a clear gel that is coated on the support structure. It is dried and sintered at 500 – 800 °C and further cross-linked to form a porous structure with a pore diameter of less than 1 nm. For the polymer sol – gel process often a titanium alkoxide is used which is converted into a cross-linked inorganic polymer in three steps:

- Hydrolysis of the metal alkoxide
- Polymerization and coating
- Sintering and cross-linking

5.1.6. *Preparation of Homogeneous Solid Membranes*

The selective barriers of many composite membranes can be considered as homogeneous solid membranes. In homogeneous solid membranes the entire membrane consists of a dense, solid, and pore-free structure. They are made from polymers and inorganic materials such as glass or metal. Because of their high selectivity for different chemical components, homogeneous membranes are used in various applications, which generally involve the separation of low molecular mass components with identical or nearly identical molecular dimensions. The most important applications are in gas separation.

One of the more important homogeneous solid metal membranes is the palladium or palladium alloy membrane used for the separation and purification of hydrogen [110]. The permeability of hydrogen in palladium, palladium alloys, and several other metals such as platinum, silver, and nickel is several orders of magnitude higher than that of any other gas. The permeability of hydrogen in palladium alloy membranes is highly temperature dependent, and separation is carried out at about 300 °C. Membranes generally consist of metal foils 10 – 50 µm thick. Because of their high selectivity, these membranes are used to produce high-purity hydrogen (> 99.99 %).

Homogeneous silica glass membranes are homogeneous structures that show promise as selective barriers for the separation of helium. Like metal membranes, glass membranes are employed at elevated temperature. Homogeneous glass membranes also have a high selectivity for hydrogen ions and are used as the selective barrier in pH electrodes. Their preparation is described in the literature [111].

5.1.7. *Preparation of Liquid Membranes*

Liquid membranes are important in combination with facilitated and coupled transport, which utilize selective “carriers” to transport certain components such as metal ions selectively and at a relatively high rate across the liquid membrane. Liquid membranes generally consist of a thin film separating two phases which may be aqueous solutions or gas mixtures. The material used for the liquid membrane should be immiscible with water and should have a low vapor pressure to guarantee long-term stability of the membrane. It is relatively easy to form a thin oil film between two aqueous or gas phases. It is difficult, however, to maintain and control this film and its properties during a mass separation process. Two different techniques are used today for the preparation of liquid membranes. In the first, which leads to supported membranes, the pores of a porous membrane are filled with the selective liquid barrier material (Fig. 50a). In the second, the liquid membrane is stabilized as a thin oil film by a surfactant in an emulsion-type mixture (Fig. 50b). In the supported liquid membranes, the porous structure provides the mechanical strength and the liquid-filled pores the selective separation barrier. Both types of membranes can be used for the selective removal of heavy-metal ions or certain organic components from industrial effluents [112, 114-116]. They have also been used rather effectively for the separation of oxygen and nitrogen [113].

Figure 50. Schematic of a) a supported hollow-fiber liquid membrane and b) the preparation of an emulsion-type liquid membrane

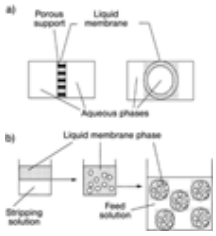

The preparation of supported liquid membranes is extremely simple. To maximize the useful lifetime of a membrane, the liquid membrane material should have low viscosity, low vapor pressure (i.e., high boiling point), and, when used in aqueous solutions, low water solubility. The porous substructure should have a high porosity and a pore size small enough to support the liquid membrane phase under hydrostatic pressure; the polymer of the substructure should be hydrophobic because most liquid membranes are used with aqueous feed solutions. In practice, liquid membranes are prepared by soaking a hydrophobic porous membrane made from polytetrafluorethylene such as Gore-Tex or from polyethylene such as Cellgard [117] in a hydrophobic liquid. The liquid may be a selective carrier such as an oxime or a tertiary or quaternary amine dissolved in kerosene. The disadvantage of supported membranes is their thickness, which is determined by the thickness of the porous support structure. The latter is in the range of 10 – 50 µm, that is, about 100 times thicker than the selective barrier of an asymmetric polymer membrane. Thus, the fluxes of supported liquid membranes can be low even when their permeabilities are high.

The preparation of the unsupported liquid membrane is more complex [118]. Here, two immiscible phases (i.e., the hydrophobic oil-type membrane phase and a water phase, often referred to as stripping solution) are mixed to form an emulsion of water droplets in a continuous oil phase, which is stabilized by addition of a surfactant. This emulsion is then added to the second aqueous phase (the feed solution), and the first emulsion forms droplets in the second aqueous phase. The overall result is the separation of two aqueous phases by an oil phase forming the liquid membrane. Ideally, droplets form in this process in which the aqueous phase is surrounded by a relatively thin hydrophobic phase, the membrane, which is surrounded by a second continuous aqueous phase. In reality, several aqueous droplets are found in a single hydrophobic droplet, as shown in Figure 50b, and the diffusion pathways become longer. With another aqueous solution, the component to be eliminated is supplied to the original emulsion and passes through the membrane into the internal solution.

5.1.8. *Preparation of Ion-Exchange Membranes*

Ion-exchange membranes consist of highly swollen gels carrying fixed positive or negative charges. The properties and preparation of ion-exchange membranes are closely related to those of ion-exchange resins. As with resins, there are many possible types with different polymer matrices and different functional groups to confer ion-exchange properties on the product. Although there are a number of inorganic ion-exchange materials, most of them based on zeolites and bentonites, these materials are rather unimportant in ion-exchange membranes and are not discussed further.

Cation-exchange membranes have negatively charged groups fixed to the polymer matrix, and anion-exchange membranes have positively charged groups fixed to the polymer matrix. In a cation-exchange membrane, the fixed anions are in equilibrium with mobile cations in the interstices of the polymer (Fig. 51).

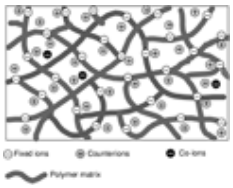

Figure 51. Schematic of a cation-exchange membrane showing the polymer matrix with the negative fixed charges, the positive counterions, and the negative co-ions

This figure shows schematically the matrix of a cation-exchange membrane with fixed anions and mobile cations (counterions). In contrast, the mobile anions, known as co-ions, are more or less completely excluded from the polymer matrix because their electrical charge is identical to that of the fixed ions. Due to the exclusion of the co-ions, a cation-exchange membrane permits transfer of cations only. Anion-exchange membranes carry positive charges fixed on the polymer matrix. Therefore, they exclude all cations and are permeable only to anions. The most desired properties for ion-exchange membranes are:

- High permselectivity: an ion-exchange membrane should be highly permeable for counterions and impermeable to co-ions.
- Low electrical resistance: the permeability of an ion-exchange membrane for the counterions should be as high as possible.
- Good mechanical and dimensional stability: the membrane should be mechanically strong and have a low degree of swelling on transition from dilute to concentrated ionic solutions.
- High chemical stability: the membrane should be stable over a large pH range and in the presence of oxidizing agents.

It is often difficult to optimize the properties of ion-exchange membranes because the parameters determining the different properties often act contrary to one another. For instance, a high degree of crosslinking improves the mechanical strength of the membrane but also increases its electrical resistance. A high concentration of fixed ionic charges in the membrane matrix leads to a low electrical resistance but generally causes a high degree of swelling and poor mechanical stability. The properties of ion-exchange membranes are determined by two parameters, namely, the polymer matrix and the type and

concentration of the fixed ionic moieties. The polymer matrix largely determines the mechanical, chemical, and thermal stability of the membrane. Often it consists of a hydrophobic polymer such as polystyrene, polyethylene, or polysulfone. Although these polymers show a low degree of swelling in water, they become hydrophilic by the introduction of the ionic moieties, and swelling in water is strongly increased. Therefore, the polymer matrix of ion-exchange membranes is very often cross-linked. The degree of cross-linking then largely determines the degree of swelling and the chemical and thermal stability, but it also has a large effect on the electrical resistance and the permselectivity of the membrane. The type and the concentration of the fixed ionic charges determine the permselectivity and the electrical resistance of the membrane, but they also have a significant effect on the mechanical properties. The degree of swelling especially is affected by the concentration of the fixed charges. The following moieties are used as fixed charges in cation-exchange membranes:

$$-SO_3^- \quad -COO^- \quad -PO_3^{2-} \quad -PHO_2^- \quad -AsO_3^{2-} \quad -SeO_3^-.$$

In anion-exchange membranes fixed charges may be:

$$-\overset{+}{N}H_2R, \quad -\overset{+}{N}HR_2, \quad -\overset{+}{N}R_3, \quad -\overset{+}{P}R_3, \quad -\overset{+}{S}R_2.$$

The sulfo group is completely dissociated over nearly the entire pH range, whereas the carboxyl group is virtually not dissociated below a pH value of 6. The quaternary ammonium group is completely dissociated at all pH values, whereas the primary ammonium group is only weakly dissociated. Accordingly, ion-exchange membranes are referred to as being weakly or strongly acidic or basic in character. Most commercial ion-exchange membranes are either homogeneous or heterogeneous. In homogeneous membranes the ion-exchange groups are homogeneously distributed throughout the polymer, whereas in heterogeneous membranes the ion-exchange particles are dispersed in a neutral polymer matrix. In general, heterogeneous membranes have a relatively high electrical resistance. Homogeneous ion-exchange membranes have a more even distribution of fixed ions and often a lower electrical resistance.

The preparation of ion-exchange membranes is described in a large number of detailed recipes in the literature [119-126]. For the practical preparation of ion-exchange membranes two rather different procedures are used. A quite simple technique is based on mixing an ion-exchange resin and a binder polymer such as poly(vinyl chloride) and extruding the mixture as a film at a temperature above the melting point of the polymer. The result is a heterogeneous membrane with relatively large domains of ion-exchange material and no conductive regions of the binder polymer. To obtain ion-exchange membranes with satisfactory conductivity the fraction of ion-exchange resin must be in excess of 50 – 70 wt %. This often leads to rather high swelling and to poor mechanical stability of the membrane. Furthermore, the size of the ion-exchange particles should be as small as possible (i.e., < 2 to 20 µm in diameter) to be able to make thin membranes with low electrical resistance and high permselectivity. Nevertheless, the method is still used today on a large scale for the preparation of ion-exchange membranes.

More recently homogeneous ion-exchange membranes have been produced by polymerization of monomers that carry anionic or cationic moieties or by introducing these moieties into a polymer which may be in an appropriate solution or a preformed film.

A typical example for the preparation of an ion-exchange membrane by polymerization and sulfonation of monomers is the polymerization of styrene with divinylbenzene and subsequent sulfonation according to the following reaction scheme:

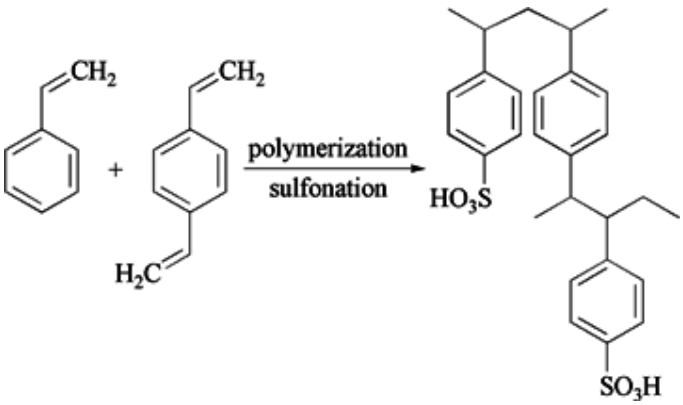

For the preparation of anion-exchange membranes positively charged groups are introduced into the polystyrene by chloromethylation and amination with triamine according to the following reaction scheme:

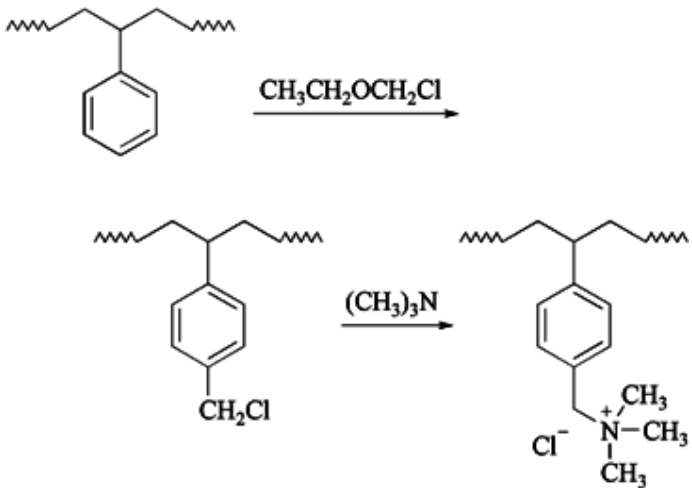

Cation-exchange membranes with good overall properties can be made by sulfonation of polyetherketone according to the following reaction scheme:

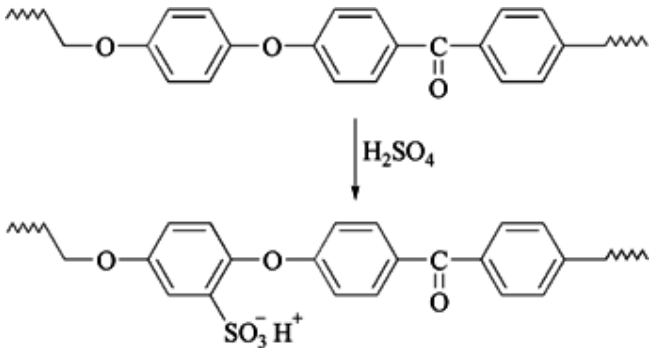

Anion-exchange membranes with excellent chemical stability can be prepared by incorporating the monoquaternary salt of 4,4′-diazabicyclo[2.2.2]octane into a chlormethylated polysulfone matrix giving an anion-exchange membrane of the following structure:

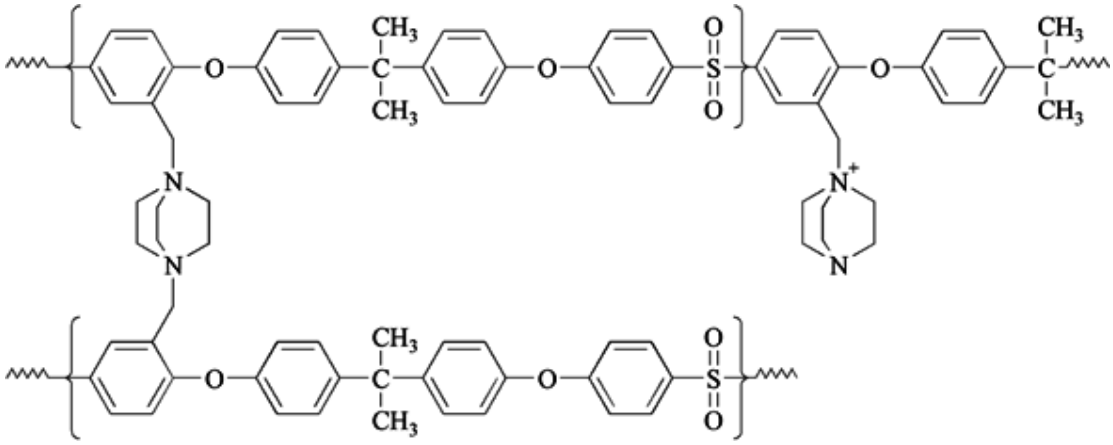

Both the cation- and the anion-exchange membrane have excellent chemical and mechanical stability and good electrochemical properties [72].

In the literature, there are numerous methods reported for the preparation of ion-exchange membranes with special properties, for instance, for use as battery separators, ion-selective electrodes, or in the chlor-alkali process. Especially membranes for the chlor-alkali industry are of commercial significance. These membranes are based on polytetrafluoroethylene and carry sulfone groups in the bulk of the membrane phase and carboxyl groups on the surface as the charged moiety. They combine excellent chemical stability with high selectivity and low electrical resistance. Today's commercially available perfluorocarbon membranes have the following basic structure:

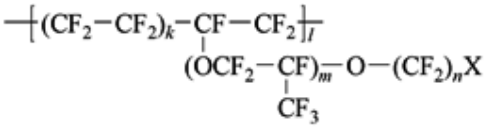

$k = 5\text{-}8$, $l = 600\text{-}1200$, $m = 1\text{-}2$, $n = 1\text{-}4$, $X = SO_3^-, COO^-$

The synthesis of the perfluorocarbon membranes is rather complex and requires a multistep process [124]. Perfluorinated membranes have excellent chemical and thermal stability. Their lifetimes under the aggressive conditions of the chlor-alkali

process are on the order of three years.

Other special-property membranes were developed for the production of table salt [127] and for use in diffusion [128] and Donnan dialysis. Significant effort has been also been concentrated on the development of anion-exchange membranes with low fouling tendencies [129] and low proton permeability [130].

A significant effort has been concentrated on the development of bipolar membranes, which have recently gained increasing attention as an efficient tool for the production of acids and bases from their corresponding salts by accelerated water dissociation. Bipolar membranes can be prepared by simply laminating conventional cation- and anion-exchange membranes back to back [131]. Laminated bipolar membranes often exhibit unsatisfactory water splitting capability. Special surface treatment of commercial ion-exchange membranes and subsequent laminating may yield bipolar membranes with satisfactory properties. Single-film bipolar membranes and multilayer bipolar membranes fulfill most of the practical needs [132].

5.2. Membrane Characterization

Membrane characterization is a very important part of membrane research and development because the design of membrane processes and systems depends on reliable data relating to membrane properties. To select a membrane to be used in a particular separation process its properties must be known. Since membranes are very different in their properties and applications a large number of different techniques are required for their characterization. Some of the more important membrane characterization procedures described in the literature or recommended by membrane manufacturers are briefly reviewed here.

5.2.1. *Characterization of Porous Membranes*

Porous micro- or ultrafiltration membranes are generally characterized in terms of their flux, their pore size, their pore size distribution, and their molecular mass cutoff. Other important characteristics are the nature of the membrane material (whether it is hydrophilic or hydrophobic or carrying positive or negative charges), the structure of the membrane (whether it is symmetric or asymmetric) and their mechanical, chemical, and thermal stability. The methods used to characterize porous membranes include simple filtration and retention measurements, microscopic techniques, surface tension measurements, and the determination of pore radii by liquid – gas or liquid – liquid replacement. The chemical and mechanical stability at various temperatures is usually measured in long-term tests. Which of the membrane characterization methods are applied depends on the anticipated use of the membrane.

Pure-Water Flux of Micro- and Ultrafiltration Membranes. A characteristic property of micro- and ultrafiltration membranes is their pure-water flux or pure-water permeability. The flux J through a porous medium is proportional to the applied hydrostatic pressure and inversely proportional to the solvent viscosity. It is expressed in Darcy's law:

$$J_v = \frac{L_p \Delta p}{\eta} \tag{5.2}$$

where J_v is the membrane flux, L_p the hydrodynamic solvent permeability, η the viscosity of the solution passing through the pores of the membrane, and Δp the applied pressure.

For the determination of the pure-water flux it is very important to use a certain water quality. Generally de-ionized or distilled water with a conductivity of less than 5 µS/cm^{-1} which has been treated by microfiltration to remove suspended materials is used as test solution. The flux is determined at different applied pressures and plotted as a function of the pressure. From the slope of the plot the hydrodynamic permeability is obtained. Prior to the flux measurements the membranes should be pressurized by filtration of pure water at a hydrostatic pressure that is at the upper limit of the recommended operating pressure for 1 – 2 h to compress the membrane to a stable state that will not further deform when used at lower pressures. Preservatives, which are often present in membranes, must be eliminated by a washing procedure.

Some inorganic membranes, like Al_2O_3-based types, show an apparent internal pore clogging when de-ionized water is used in pure-water flux measurements. In these cases an electrolyte solution is used for determining the membrane flux (e.g., an NaCl or KCl solution). Since there is no standard method for the determination of the pure-water flux, companies use their own standards, which makes a reliable comparison of pure-water fluxes rather difficult. Typical pure-water fluxes of microfiltration membranes vary from 500 to 50 000 L m^{-2} h^{-1} bar^{-1}, while those of ultrafiltration membranes vary from 50 to 800 L m^{-2} h^{-1} bar^{-1}.

In practical applications fluxes are normally much lower and a steep flux decline is observed during the first period of operation, mainly due to membrane fouling, which is generally more pronounced for microfiltration than for ultrafiltration. More details on membrane fouling are given in Section Membrane Fouling and its Causes.

Microscopic Techniques. A variety of microscopic techniques are used to study membrane structures, such as scanning electron microscopy, field emission electron microscopy, transmission electron microscopy, and atomic force microscopy. The advantage of these techniques is that direct visual information of the membrane morphology is obtained. The different microscopic techniques require different preparation methods and have different resolutions. Scanning electron microscopy, for example, has a resolution of up to 5 nm and provides good information on the structures of micro- and ultrafiltration membranes including pore sizes and pore shapes. To reach a high resolution high electron-beam energy has to be applied, which can damage the membrane samples. Another problem is that in electron microscopy the samples must be coated by a conductive layer. This is generally done in a high vacuum which can lead to a change in the structure of membranes made from hydrophilic polymers.

The pore sizes and pore shapes of ultrafiltration membranes can also be determined by transmission electron microscopy, which has a resolution of 0.4 – 0.5 nm, or by field emission electron microscopy, which has a resolution of 0.6 – 0.7 nm.

Transmission electron microscopy requires a significantly more complex preparation procedure and is mainly applied to determine the surface porosity of ultrafiltration membranes [133]. Because of the high resolution, pores in ultrafiltration membranes can be visualized by field emission electron microscopy [134].

Sample preparation for scanning electron microscopy and field emission electron microscopy is relatively easy compared to that for transmission electron microscopy. To visualize the cross sections of a membrane, for example, it is immersed in ethanol or a water – ethanol mixture prior to cooling in liquid nitrogen. The sample is fractured to give a rather clean fracture and glued with a conductive glue to a sample holder and sputtered with a thin layer of gold. The sputtered layer should be thinner than the structural features of the membrane to be studied. More recently, low-vacuum scanning electron microscopy (so-called environmental scanning electron microscopy) has become available, which makes it possible to scan samples without a conductive coating. However, the magnification of these microscopes is limited to 7500. Figures 40 to 42 show typical pictures of membrane structures that can be obtained by scanning electron microscopy.

Besides electron microscopy, atomic force microscopy is also applied to study membrane surface structures. It has a resolution of about 1 nm [135]. Sample preparation is fairly simple, since no conductive coating is required. Therefore, samples need not be dried and exposed to vacuum.

5.2.1.1. *Membrane Separation Properties Determined by Filtration Tests*

Both micro- and ultrafiltration are pressure-driven membrane processes using porous membranes. The separation of various components is based on a sieving mechanism in which particles or molecules which are larger than the pores of a membrane are retained by the membrane. The membrane retention as a function of the particle size and the pore size can be defined by a relation between the pore and the particle radius [136]. If the particles are spherical and have a radius r_s and the membrane pores are cylindrical with a radius r_p the following relation can be used to describe the membrane solute rejection:

$$R = \left[1 - \left(1 - \frac{r_s}{r_p}\right)^2\right]^2 \tag{5.3}$$

For the radius of the molecule either the radius of gyration or the hydrodynamic radius can be used. The so-called Ferry model [136] is rather simplified, but gives a good approximation of the expected retention when particle size and the membrane pore radius are known.

Retention and Molecular Weight Cutoff. To describe the retention of ultrafiltration membranes usually the molecular weight cutoff is used. Determination of the molecular weight cutoff is very sensitive to the experimental conditions, that is, the solutes and their concentration used in the test, the applied pressure, and the temperature. Therefore, a comparison of retention data given by different manufacturers is generally problematic. As test conditions for determining the ultrafiltration membrane cutoff a transmembrane pressure of 100 kPa, a feed solution concentration of 0.1 %, and a test temperature of 25 °C are recommended. Furthermore, maximum agitation of the feed solution at the membrane surface is required to minimize concentration polarization effects, and small amounts of solution should be filtered to assure a constant feed concentration.

The retention is generally expressed in per cent and defined by:

$$R = \left(1 - \frac{C_p}{C_f}\right)100 \tag{5.4}$$

where R is the retention or rejection of a membrane, and C_f and C_p are feed and permeate concentration.

The retention of molecules having identical molecular weights may be rather different when their shape is different. Globular proteins, for example, are rejected more easily than branched polysaccharides or flexible polymers [137]. Adsorption of the solute on the membrane surface, for example, may result in high “apparent” retention values when the experiment is performed for a very short time period. Accumulation of solutes at the membrane surface due to concentration polarization effects can also obscure the retention measurements and must be considered when the true (i.e., intrinsic) retention is to be determined. These effects can more or less be eliminated by using low pressures, low feed concentrations, and a high degree of agitation.

Manufactures of ultrafiltration membranes generally use marker molecules such as dextrans and proteins for the determination of the molecular weight cutoff of their membranes. In addition, the sharpness of the cutoff is important for practical applications. It is determined by measuring the retention of the membrane for components of different molecular weight and plotting the retention versus the molecular weight of the retained components (Fig. 52). The figure shows the characteristic curves of a membrane with sharp and diffuse cutoff. The sharpness of the cutoff of a membrane depends mainly on its pore size distribution. For practical applications the cutoff should be as sharp as possible.

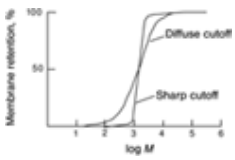

Figure 52. Schematic illustrating membrane cutoff characteristics

Bacterial Challenge Test. Micro- and ultrafiltration membranes are often used in the food and pharmaceutical industry for the sterilization of solutions. In this application it is important that membranes retain undesired microorganisms as completely as possible. Therefore, these membranes are subjected to a bacterial challenge test with a solution containing a minimum

concentration of organisms such as *Pseudomonas diminuta* per cubic centimeter of solution [138]. The data obtained in bacterial challenge tests are normally expressed in terms of the logarithmic reduction value, which is given by the logarithm of the ratio of feed to filtrate concentration. The bacterial challenge test is a destructive test and is therefore not suited as an on-line quality-control test. For this application the nondestructive diffusion and the bubble-point integrity tests are generally applied.

Membrane Pore Size Measurement. In addition to filtration tests micro- and ultrafiltration membranes are characterized by determining their pore size and pore size distribution by a number of techniques such as liquid – gas and liquid – liquid displacement, mercury intrusion, permporosimetry, thermoporosimetry, and the bubble point test.

The *bubble point test* is a structural integrity test in which the largest pore or hole in a membrane is determined. The bubble point method is based on the capillary effect of small pores due to surface tension forces. In equilibrium these forces are balanced with gravity of the liquid in the pores. The principle is shown in Figure 53.

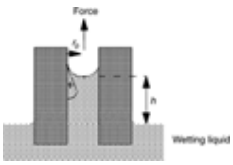

Figure 53. Schematic illustrating the capillary effect of a membrane pore

This equilibrium between surface tension forces and gravity of the liquid in the pore can be expressed by the following equation:

$$2\pi\, r_p \sigma \cos\, \varphi = r_p^2 \pi\, h \rho g \tag{5.5}$$

where r_p is the pore radius, σ the surface tension of the liquid in contact with air, cos φ the contact angle between the liquid and the wall of the pore, h the height of the liquid column, ρ the liquid density, and g the acceleration due to gravity.

The term $h\,\rho g$ in Equation 5.5 can be replaced by the pressure p to give the LaPlace equation:

$$r_p = \frac{2\sigma \cos \varphi}{p} \tag{5.6}$$

The bubble point method is a fairly simple technique where the pores of a membrane are filled with a liquid. From the bottom side of the membrane nitrogen gas or air is introduced with increasing pressure. At a certain pressure the gas will replace liquid in the largest pores permeating the membrane, and a bubble rising from the membrane surface can be observed. Equation 5.6 gives a relation for the surface tension between the liquid and the gas and the required pressure to open up pores of a certain size. In the bubble point test mostly water or isopropyl alcohol is used, since the surface tension of water – air is approximately 3.5 times higher than that of isopropyl alcohol – air. Other techniques to determine the pore size and pore size distribution of a porous membrane that are also based on the La Place equation, but cover different pore size ranges when the same pressure range is applied, are mercury porosimetry, gas – liquid displacement, and liquid – liquid displacement.

In *mercury porosimetry* a dry membrane is exposed to a certain volume of liquid mercury. The pressure applied to the mercury is slowly increased, while simultaneously the amount of mercury forced into the porous structure is measured. The largest pores are filled first and the required pressure for the mercury to penetrate the porous structure increases with decreasing pore size. The required pressure corresponds to a certain pore radius and the amount of mercury that enters the membrane corresponds to the total volume of these pores. The interfacial tension of mercury and air is very high. Therefore, relatively high pressures are required to fill small pores. In polymer membranes only pores having a radius in excess of 1 µm can be detected reliably by mercury porosimetry. Furthermore, this technique does not distinct between dead-end pores and interconnected pores.

The *gas-liquid displacement technique* is identical to the bubble point method. The pores of the membrane are generally filled with an organic solvent which has a lower surface tension than water. The solvent is forced out of the membrane pores by nitrogen gas, which is introduced with increasing pressure. As the pressure increases the liquid is replaced by nitrogen in the largest pores first and a convective gas flow through these pores occurs. The gas flow is measured as a function of the applied pressure. When all the pores are opened a linear relationship between pressure and gas flow is observed when the pressure is further increased. Figure 54 shows schematically typical gas fluxes obtained as a function of pressure with a dry and a liquid-filled membrane.

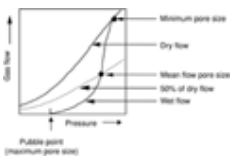

Figure 54. Schematic illustrating gas – liquid displacement measurements to determine the pore size and pore size distribution of microfiltration membranes

The mean flow pore size is given by the point where the 50 % dry-flow curve crosses the wet-flow curve. The maximum pore size is given by the point at which flow through the liquid filled membrane is first observed. The minimum pore size is given by the point where the fluxes of the dry and the liquid-filled membrane become identical. The LaPlace equation (Eq. 5.6) describes the relationship between the pore radius and the applied pressure.

Liquid – liquid displacement differs from gas – liquid displacement in that the liquid in the pores is displaced by a second liquid instead of a gas. The two liquids applied should be immiscible. A typical liquid pair used is water – isobutanol. Both liquids are first saturated with each other before one of them is used to fill the pores and the other liquid is applied as displacement liquid. The experimental setup for the determination of pore sizes by liquid – liquid displacement is shown in Figure 55. A slow increase in the displacement liquid pressure pushes the liquid out of the largest pores first. With increasing pressure the liquid in the smaller pores is also replaced. The pore size distribution can be calculated from a plot of the flux versus the applied pressure [139, 140].

Whether isobutanol or water is used as displacing liquid depends on the nature of the membrane material. Due to swelling phenomena different results might be obtained dependent on the liquid occupying the pores.

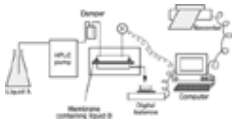

Figure 55. Schematic of a typical liquid – liquid displacement setup [139]

The relation between the pore radius and the pressure required to open pores of a certain size is again described by the LaPlace equation (Eq. 5.6) and the pore size distribution can now be calculated by measuring the flux as a function of the pressure. Like in gas – liquid displacement only active pores contributing to transport are taken into account. Pore diameters of approximately 5 – 100 nm can be analysed by liquid – liquid displacement. The advantage of liquid – liquid displacement over gas – liquid displacement is the lower interfacial tension between two liquids compared to a gas – liquid combination. Therefore, pores of the same size can be opened at much lower pressures in liquid – liquid displacement.

Other methods that can be used to determine pore sizes in porous membranes [141, 142] include permporometry and thermoporometry. *Permporometry* is a technique based on gas adsorption – desorption measurements in which pores are either filled by controlled condensation of a vapor phase or filled pores are emptied by controlled evaporation of the liquid inside the pores. The relation between the relative vapor pressure and the pore radius is given by the Kelvin equation. *Thermoporometry* is based on the microcalorimetric analysis of a solid – liquid transition by sensitive differential scanning calorimetry. Table 4 lists the pressures required to open pores of 10 nm for different displacement systems. An advantage is that membranes are characterized in a wetted state and therefore the test conditions are closer to the real operating conditions.

Table 4. Different porosimetry techniques and the influence of interfacial tension on the pressure required to open pores of 10 nm

Technique	System	Interfacial tension, mN m^{-1}	Pressure, bar
Mercury porosimetry	air/mercury	480	749
Bubble point	air/water	72.3	145
Coulter porometry	air/Porofil	16	32
Gas – liquid displacement	air/isobutanol	20.7	41
Liquid – liquid displacement	water/isobutanol	1.85	3.7
Liquid – liquid displacement	water/isobutanol/methanol	0.35	0.7

5.2.2. *Characterization of Homogeneous Membranes*

Homogeneous membranes are generally used for the separation of low molecular mass materials such as gases, salts, or solvents. The transport mechanism in these membranes is based on the solution and diffusion of components within the membrane matrix. The intrinsic properties of the membrane material are often more relevant for characterizing their transport behavior than is the macroscopic membrane structure. Characterization methods used with microporous structures have therefore been supplemented with other procedures, such as sorption and diffusion measurements or determination of the glass transition temperature and crystallinity of the membrane material. For ion-exchange membranes, the measurement of electrokinetic properties is required.

A detailed discussion of all methods used to characterize *homogeneous polymer membranes* is beyond the scope of this outline. In addition to the determination of diffusion and partition (solubility) coefficients, measurements providing information on the physical structure of the polymer, such as differential scanning calorimetry, X-ray studies, or determination of the free volume in a polymer, are important. These methods are described in [143, 145, 146]. Permeability and selectivity are the most interesting parameters; they are determined in permeation tests and expressed in terms of rejection coefficients (in reverse osmosis) or separation coefficients (in gas separation and pervaporation). These coefficients are usually not constant but a function of composition, temperature, and pressure. Furthermore, the determination of membrane fluxes and selectivities in practical measurements is often obscured by boundary layer effects at the membrane surface.

The methods used to determine the structural properties of *ion-exchange membranes* are similar to those used for characterizing other polymer membranes. Because ion-exchange membranes are used mainly in separation processes with electrical potential driving forces, their permeability and selectivity are determined under experimental conditions closely

related to their practical use, and their properties are expressed in terms commonly used in electrochemistry (e.g., electrical resistance, ion-transfer numbers, charge densities) [72].

5.3. Membrane Modules

To use a membrane in a separation process it must be installed in a proper device which is generally referred to as a membrane module. Membrane modules must meet certain requirements with regard to production costs, packing density, energy consumption, and control of concentration polarization. The modules used in the practical applications of membrane separation processes are quite different [147-149]. In small-scale laboratory applications often stirred batch cells or dead-end filter cells are used. In large-scale industrial applications six basic module types are used. These modules are shown schematically in Figures 56 to 61. They are quite different in their design, their mode of operation, their production costs, and the energy requirements for pumping a feed solution through the module. Commercially available modules, even if they are of the same type, can vary significantly in detail and are often tailor-made for a certain application. Here, only the basic concepts of the different module types are discussed.

In some modules, for example, tubular, plate-and-frame, and capillary type modules, concentration polarization and membrane fouling can effectively be controlled by proper feed flow. Other modules such as the spiral-wound and the hollow-fiber module are more sensitive to membrane fouling. There is no module type that serves all membrane processes and applications. The different modules which are commercially available today are designed for a certain application and process in which they provide the technically and commercially best solution. The same is true for the process design and mode of operation. In certain applications batch and feed-and-bleed operation is used, while in others continuous operation is more efficient.

5.3.1. *Cartridge Membrane Module*

The pleated cartridge filter membrane module (Fig. 56) is used mainly in dead-end microfiltration. It consists of a pleated membrane cartridge installed in a pressurized housing. The feed solution enters the filter from the housing side and the product is collected in a center tube which is sealed against the housing by an O-ring. Cartridge filters are operated at relatively low hydrostatic pressures of 1 – 2 bar. Their useful life is limited due to plugging of the membrane pores by the retained solutes.

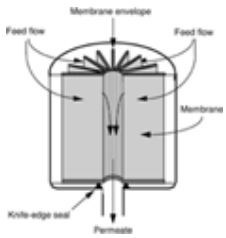

Figure 56. Schematic of a cartridge filter unit

The useful life of a cartridge filter depends very much on the feed solution constituents and their concentration. It can vary between a couple of days and several months. Cartridge filters are disposable items. If the membrane, which has the typical characteristics of a depth filter, is blocked by the retained particles it can generally not be cleaned and the original flux cannot be restored. The main application of cartridge filters is in sterile filtration of water, beverages such as wine, beer, and fruit juices, and pharmaceutical solutions. They are also used for the filtration of surface water for industrial or household use to remove particles and suspended matter and to serve as prefilters in reverse-osmosis water-desalination plants. An important application of cartridge filters is in the production of ultrapure water for the electronic industry, where they are used as point-of-use filters to remove all traces of particles from the ultrapure water used to rinse electronic components. The actual cartridge is made by pleating a membrane sheet and potting the ends by an appropriate resin or hot-melt adhesive. Since the process of pleating the membrane places a considerable mechanical stress on the membrane, only materials that are rather flexible and have a certain mechanical strength are suited for the fabrication of cartridge filters. The cartridge filter module provides a relatively high surface area per unit volume and the production costs are relatively low.

5.3.2. *Plate-and-Frame Membrane Module*

Another module type used on an industrial scale for various membrane separation processes including ultrafiltration, reverse osmosis, and gas separation is the plate-and-frame module. Its design has its origin in the conventional filter press concept. The membranes, porous membrane support plates, and spacers forming the feed flow channel are clamped together and stacked between two endplates and placed in a housing (Fig. 57).

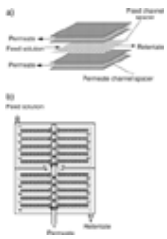

Figure 57. Schematic illustrating a) the concept of a plate-and-frame membrane module and b) a circular plate-and-frame filter device with one baffle to extend the feed flow path length

The feed solution is pressurized in the housing and forced across the surface of the membrane. The permeate leaves the module through the permeate channel to a permeate collection manifold, which in circular devices is a central tube, as shown in Figure 57b. Often the device contains one or more baffles to extend the path length of the feed solution in the device. There are various types of plate-and-frame modules on the market, but they offer only slight variations in their basic

configuration. In many plate-and-frame membrane modules the membranes can easily be exchanged. This makes the module suitable for batch-type operations and multipurpose applications using different membranes for different separation tasks. Plate-and-frame units are mainly used in small-scale applications such as the production of certain pharmaceuticals, bioproducts, or fine chemicals. The housings and other components of plate-and-frame modules to be used in the food and pharma industry are made from stainless steel so that they can easily be steam-sterilized. These units, however, are quite expensive and the exchange of the membranes is labor-intensive. Therefore, the plate-and-frame module is quite expensive. However, one exception is the electrodialysis stack, which resembles a plate-and-frame unit in which cation- and anion-exchange membranes are stacked in alternating series between two end plates. A typical sheet-flow electrodialysis stack is shown in Figure 58.

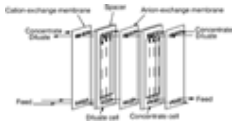

Figure 58. Arrangement of cation- and anion-exchange membranes in an electrodialysis stack

This device holds an array of membranes between electrodes in such a way that the streams being processed are kept separate. Gaskets between the membranes form the individual cells and contain manifolds to distribute the process fluids in the different compartments. The distance between the membrane sheets in industrial-size electrodialysis stacks is typically between 0.5 and 2 mm. In a practical electrodialysis system 200 – 1000 cation- and anion-exchange membranes are installed in parallel between two electrodes, and a stack contains 200 – 1000 m^2 of membrane area. Thus, the membrane area per unit volume is relatively high. The electrodialysis stack is discussed in Section Membrane Separation Process Design and Operation.

5.3.3. *Spiral-Wound Module*

A variation of the basic plate-and-frame concept is the spiral-wound module, which is widely used today in reverse osmosis, ultrafiltration, and gas separation. Its basic design is illustrated in Figure 59.

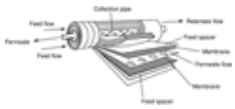

Figure 59. Schematic of a spiral-wound membrane module

The feed flow channel spacer, the membrane, and the porous membrane support form an envelope which is rolled around a perforated central collection tube and inserted into an outer tubular pressure shell. The feed solution passes in axial direction through the feed channel across the membrane surface. The filtrate is moved along the permeate channel and is collected in a perforated tube in the center of the roll. Small spiral wound units consist of just one envelope which limits the total membrane area that can be installed in one unit to about 1 – 2 m^2. The main reason for the limitation of the surface area which can be installed in a module containing one single envelope is the pressure drop encountered by the permeate moving down the permeate channel to the central collection tube. Because the channel in a practical unit is very narrow its length is limited to 2 – 5 m. A significantly longer path would result in an unacceptable pressure drop in the permeate channel. To install larger membrane surfaces in a spiral-wound module and to extend the path of the permeate a multileaf arrangement is used. Commercial spiral wound modules are about 1 m long and have a diameter of 10 – 60 cm. The membrane area in a spiral-wound element is 3 – 60 m^2. Generally, 2 – 6 elements are placed in series in a pressure vessel.

The spiral-wound module provides a relatively large membrane area per unit volume. Large-scale production is quite cost effective and module costs per membrane area are quite low. The major application of the spiral-wound module is in desalination of sea- and brackish water by reverse osmosis, but it is also extensively used in ultrafiltration and gas separation. However, the spiral-wound module is quite sensitive to fouling, and the feed channels can easily be blocked, so particles or fibers should be removed from the feed solution by a proper pretreatment procedure.

5.3.4. *Tubular Membrane Module*

While the previously described three membrane module types require flat sheet membrane material for their preparation, special membrane configurations are needed for the preparation of tubular, capillary, and hollow-fiber modules.

The tubular membrane module consists of membrane tubes placed into porous stainless steel or fiber glass reinforced plastic pipes. The pressurized feed solution flows down the tube bore and the permeate is collected on the outer side of the porous support pipe, as indicated in Figure 60a. The diameters of tubular membranes are typically between 1 and 2.5 cm. In some modules, the membranes are cast directly on the porous pipes, and in others they are prepared separately as tubes and then installed into the support pipes. Today, tubular modules are used in ultrafiltration at low hydrostatic pressures. This allows the membrane tubes to be made by welding or gluing of flat sheet membranes that are cast on a relatively thick and mechanically strong porous polyester support material. These tubes, which have a diameter of 0.5 – 1 cm, do not need additional support when operated at hydrostatic pressures of less than 2 – 4 bar.

Usually, 10 – 30 individual tubes are installed in a larger tube and potted at the end of the tube. The feed solution is fed in parallel through the tubular bundle while the permeate of the individual tubes is collected in the outer shell tube, as shown in Figure 60b. The main advantage of the tubular module is that concentration polarization effects and membrane fouling can be easily controlled, and plugging of the membrane module is avoided even with feed solutions that have very high concentrations of solid matter and thus high viscosity. The disadvantage of the tubular module is the low surface area that can be installed in a given unit volume and the very high cost. Therefore, tubular membrane modules are generally only applied in applications where feed solutions with high solids content and high viscosity must be treated and other module

concepts fail due to membrane fouling and module plugging. This is the case in certain applications in the food and pharmaceuticals industries and in the treatment of certain industrial effluents.

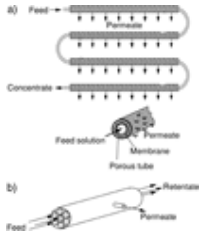

Figure 60. Schematic illustrating the tubular membrane module a) with individual tubes in series and b) with seven individual tubes in parallel bundled in a shell tube

5.3.5. *Capillary and Hollow-Fiber Membrane Modules*

The capillary membrane module (Fig. 61a) consists of a large number of membrane capillaries with an inner diameter of 0.2 – 3 mm arranged in parallel as a bundle in a shell tube. The feed solution is passed down the center of the membrane capillary, and the filtrate, which permeates the capillary wall, is collected in the shell tube. The capillary membrane module requires membranes in a self-supporting capillary configuration, which when asymmetrically structured, carry the selective barrier on the inner side of the capillary, as indicated in Figure 42b, which shows a typical capillary, ultrafiltration membrane prepared in a wet-spinning process.

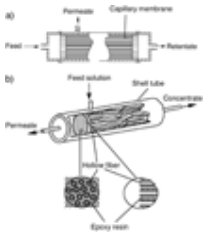

Figure 61. Schematic illustrating the construction a) of a capillary module and b) of a hollow-fiber module

The same basic spinning process is used for the preparation of hollow-fiber membranes, which have an outer diameter of 50 – 100 µm. In hollow-fiber membranes, the selective layer is on the outside of the fibers, which are installed as a bundle of several thousand fibers in a half loop with the free ends potted with an epoxy resin in a pressure tube, as shown in Figure 61b. The filtrate passes through the fiber walls and flows up the bore to the open end of the fibers at the epoxy head.

The hollow-fiber membrane module has the highest packing density of all module types available on the market today and its production is very cost effective. Hollow-fiber membrane modules can be operated at pressures in excess of 100 bar. The main disadvantage of the hollow-fiber membrane module is the difficult control of concentration polarization and membrane fouling. When operated with liquid solutions the modules do not tolerate any particles, macromolecules, or other materials that can easily precipitate at the membrane surface. Therefore, an extensive pretreatment is required when hollow-fiber membranes are used for the treatment of liquid mixtures. The main application of the hollow-fiber module is in reverse osmosis desalination of seawater and in gas separation. Both applications require high operating pressures and low-cost membranes which have a long useful life. In reverse osmosis of seawater extensive pretreatment of the feed water is required.

5.3.6. *Other Membrane Modules*

Other membrane module concepts include the rotating-disk module, in which the membranes are arranged as in the circular plate-and-frame module. However, the feed solution is not pumped through the module but the membrane plates rotate and thus provide some turbulence in the feed solution at the membrane surface. This module is rather costly per unit installed membrane area and is used only in very specific applications in the pharmaceutical industry when solutions with very shear sensitive constituents must be processed.

Table 5 lists the different membrane modules, their costs per unit membrane area, and main applications.

Table 5. Commercially available membrane modules, their costs, and their applications

Membrane module	Membrane area per unit volume, $m^2\ m^{-3}$	Membrane costs	Control of concentration polarization	Applications[a]
Filter cartridge	800 – 1000	low	very poor	dead-end MF
Plate-and-frame	400 – 800	medium	good	MF, UF, RO, D, ED
Spiral-wound	800 – 1200	low	good	UF, RO, GS
Tubular	20 – 100	very high	very good	MF, UF, RO
Capillary	600 – 1200	low	very good	UF, MF, D, SLM
Hollow-fiber	2000 – 5000	very low	very poor	RO, GS

[a]D = dialysis, ED = electrodialysis, GS = gas separation, MF = microfiltration, RO = reverse osmosis, SLM = supported liquid membrane, UF = ultrafiltration.

[Top of Page]

6. Concentration Polarization and Membrane Fouling

Even if the proper membrane for a certain mass separation task is available, various engineering aspects must be considered that can affect the economics of the process considerably, or even make it technically impossible. A major problem in all membrane separation processes is the decline of the membrane flux due to concentration polarization effects and the formation of cake or gel layers by feed solution constituents retained by the membrane. Equally devastating for the performance of a process is membrane fouling. Membrane fouling is a more general term. It may be the result of concentration polarization, or the consequence of adsorption of feed solution constituents at the membrane surface and, especially in microfiltration, also within the membrane structure. The control of concentration polarization and membrane fouling is a major problem in the design of membrane separation processes and equipment [150-152].

Concentration polarization occurs in all separation processes, but its consequences are especially severe in membrane processes. When in a mass separation procedure a molecular mixture is brought to a membrane surface, some components permeate the membrane under a given driving force while others are retained. This leads to accumulation of retained material and depletion of the permeating components in the boundary layers adjacent to the membrane surface. This phenomenon is referred to as concentration polarization. The causes and consequences of concentration polarization may differ in different membrane processes. Often the adverse effects of concentration polarization are intensified by adsorption of certain feed mixture constituents at the membrane surface. This phenomenon is referred to as membrane fouling. It is often observed when solutions containing biological materials such as proteins are processed, but inorganic materials, especially polyelectrolytes, can also cause severe fouling. In biofouling microorganisms accumulate at and attach to the membrane surface. The actual mechanism of biofouling is rather complex and involves the formation of a so-called conditioning film, which consists of organic polymeric material that then leaves the membrane surface within the biocompatible range for attachment of film-forming bacteria. Biofouling is especially severe in reverse osmosis of seawater and its control is crucial for the reliable operation of reverse osmosis desalination plants. Concentration polarization can be minimized by hydrodynamic means such as the feed flow velocity and the membrane module design. The control of membrane fouling, however, is more difficult. Membrane fouling is influenced by the chemical nature of the membrane material and the feed solution constituents. To avoid or control membrane fouling extensive feed water pretreatment and dosage of special chemicals may be required. The causes and consequences of concentration polarization and membrane fouling as well as necessary means to control them depend on the feed water composition and on the membrane process.

6.1. Concentration Polarization in Filtration Processes

The causes and consequences of concentration polarization in pressure-driven membrane processes such as reverse osmosis and ultra- and microfiltration have been studied extensively [153-155]. While the causes for concentration polarization are identical in all three processes the consequences are rather different. When water selectively permeates a reverse osmosis or ultra- and microfiltration membrane, the retained solutes are accumulated at the membrane – solution interface. Thus, a concentration gradient between the solution at the membrane surface and the bulk is established which leads to backtransport of the material accumulated at the membrane by diffusion and eventually other means. This phenomenon is referred to as concentration polarization. In reverse osmosis mainly low molecular weight materials are separated from a solvent such as water. The feed solutions often have a considerable osmotic pressure. For example seawater has an osmotic pressure of about 24 bar. In reverse osmosis concentration polarization leads to an increase in the osmotic pressure, which is directly proportional to the solute concentration at the membrane surface, and thus a decrease in the membrane flux at constant applied hydrostatic pressure. Furthermore, the quality of the filtrate is impaired since solute leakage through the membrane is also directly proportional to the solute concentration at membrane feed-side surface [156, 157].

In ultra- and microfiltration only macromolecules and particles are retained by the membrane. The osmotic pressure of the feed solution is generally quite low. However, the applied hydrostatic pressure is also low and in certain applications of ultrafiltration the osmotic pressure can not be neglected. Due to the rather high molecular weight of the components separated in ultra- and microfiltration their diffusion from the membrane surface back into the bulk solution is relatively slow. Therefore, the retained components often are precipitated and form a solid layer at the membrane surface [158]. This layer, which often exhibits membrane properties itself, can effect the membrane separation characteristics significantly by reducing the membrane flux and by changing the rejection of lower molecular weight components. This is especially problematic in the fractionation of materials of different molecular weight.

To describe concentration polarization in membrane filtration processes mathematically a relationship is needed which relates the solute concentration at the membrane surface to that in the bulk and permeate solution, the membrane flux, and the fluid flow conditions in the boundary layer between membrane surface and bulk solution. There are two basic approaches to this problem [159]:

- Application of the so-called film model with the appropriate mass transfer coefficients to describe solute transport into and out of the boundary layer.
- Solution of the transport equations for the fluid velocity field to obtain an expression for the local concentration at any point in the feed solution.

Concentration polarization without precipitation of solutes at the membrane surface can be described reasonably well as a function of the various hydrodynamic parameters and membrane and feed solution properties by the film model. The more rigorous treatment of concentration polarization is based on a material and momentum balance. The numerical solution of the transport equations for the fluid and the solutes is rather complex even after introducing certain simplifications such as complete solute rejection, concentration-independent transport coefficients, and no diffusive transport parallel to the

membrane surface. Mathematical modeling is even more difficult when concentration polarization leads to precipitation of the solutes at the membrane surface. Here time-dependent effects such as densification of a formed gel or cake layer often influence flux and separation characteristics of the membrane.

6.1.1. *Concentration Polarization without Solute Precipitation*

The treatment of concentration polarization without solute precipitation by the film model assumes turbulent flow with fully developed velocity and concentration profiles in the boundary layer at the membrane surface. The description of concentration polarization in laminar flow is based on considerations analogous to heat transfer in laminar feed flow conditions and must consider entrance and exit effects in the development of the velocity and concentration profiles in the feed solution.

The film model assumes that, even with *turbulent feed flow*, at a certain distance from the entrance a constant laminar boundary layer is established adjacent to the membrane surface in which longitudinal mass transport is negligibly low, so that mass transport within the film can be regarded as one-dimensional in the direction perpendicular to the membrane surface. During the filtration process, a steady state is reached in which the convective transport of solutes to the membrane surface is counterbalanced by a diffusive flux of the retained material back into the bulk solution. Therefore, a constant concentration profile of the retained material in the laminar boundary layer is obtained. This is illustrated in the schematic diagram of Figure 62.

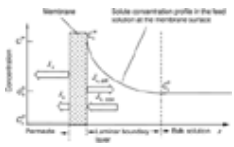

Figure 62. Schematic diagram illustrating the concentration profiles of a component retained by the membrane and the fluxes in the laminar boundary layer at the feed side surface of a membrane under steady-state conditions as assumed in the film model

The material balance for the solute in the differential element of the film is given by:

$$J_s = J_{s,con} - J_{s,diff} \tag{6.1}$$

where J_s is the solute flux through the membrane, $J_{s,con}$ the solute flux toward the membrane by convection, and $J_{s,diff}$ the solute flux from the membrane surface into the bulk solution by diffusion.

Furthermore:

$$J_s = J_v C_s^p, \; J_{s,diff} = -D_s \frac{d\,C_s}{d\,z}, \text{ and } J_{s,con} = J_v C_s \tag{6.2}$$

where J_v is the membrane volume flux, D_s is the diffusion coefficient, and dC_s/dz is the concentration gradient of the solute in the boundary layer solution.

Combination of Equations (6.1) and (6.2) and integrating with the boundary conditions $C_s = C_s^w$ at $z = 0$ and $C_s = C_s^b$ at $z = Z_b$ leads to:

$$\frac{J_v Z_b}{D_s} = \ln \frac{C_s^w - C_s^p}{C_s^b - C_s^p} \tag{6.3}$$

where Z_b is the boundary layer thickness, C_s^p the permeate solute concentration, and C_s^w and C_s^b are the solute concentrations in the feed solution at the membrane wall and in the bulk.

The membrane rejection R relates the permeate concentration to the concentration at the membrane wall:

$$R = 1 - \frac{C_s^p}{C_s^w} \tag{6.4}$$

Combination of Equations (6.3) and (6.4) leads to:

$$\frac{C_s^w}{C_s^b} = \frac{\exp \frac{J_v Z_b}{D_s}}{R + (1 - R) \exp \frac{J_v Z_b}{D_s}} \tag{6.5}$$

In Equation (6.5) the concentration polarization is expressed as the ratio of the solute concentration at the membrane surface C_s^w and that in the bulk solution C_s^b. It is a function of the membrane flux J_v, the thickness of the boundary layer Z_b, the diffusion coefficient of the solute in the boundary layer solution D_s, and the membrane solute rejection R.

To describe the concentration polarization under various feed flow conditions a solute mass transfer coefficient k_s is introduced. Assuming that the membrane flux is small compared to the feed flow parallel to the membrane surface the mass transfer coefficient can be expressed by:

$$k_s = \frac{D_s}{Z_b} \tag{6.6}$$

where k_s is the solute mass transfer coefficient, D_s is the solute diffusion coefficient in the boundary solution, and Z_b is the boundary layer thickness.

Introducing Equation (6.6) into Equation (6.5) leads to:

$$\frac{C_s^w}{C_s^b} = \frac{\exp \frac{J_v}{k}}{R + (1 - R) \exp \frac{J_v}{k}} \tag{6.7}$$

Equation (6.7) describes the concentration polarization as a function of the filtration rate, the membrane rejection, and the mass transfer coefficient. The filtration rate and the membrane rejection can easily be measured. The mass transfer coefficient is more difficult to determine. However, empirical correlations developed for various feed solution flow regimes under the assumption of steady-state conditions can be used to estimate the mass transfer coefficient as a function of the feed flow velocity and the filtration device geometry. These correlations may be expressed in a very general form by the Sherwood, Reynolds, and the Schmidt numbers [160]:

$$Sh = a\, Re^b Sc^c \left(\frac{d_H}{L}\right)^d \tag{6.8}$$

where *Sh, Re*, and *Sc* are respectively the Sherwood, Reynolds, and Schmidt numbers, d_H is the hydraulic diameter or other characteristic dimension, *L* the length of the feed flow duct, which may be a channel or a tube, and *a, b, c*, and *d* are constants which are characteristic for different filtration system geometries and which must be experimentally determined.

The Sherwood number is given by:

$$Sh = k_s \frac{d_H}{D_s} \tag{6.9}$$

where D_s is the diffusion coefficient.

The Schmidt number is given by:

$$Sc = \frac{v}{D_s} \tag{6.10}$$

where v is the viscosity.

The Reynolds number for turbulent flow in a tube or channel is given by:

$$Re = \frac{d_H u}{v} \tag{6.11}$$

where *u* is the linear velocity.

In a stirred batch cell the Sherwood and the Reynolds numbers can be expressed by the cell diameter D_c, the stirrer speed ω, and the stirrer length *d*. Thus, in a well-stirred batch cell the Sherwood number is given by:

$$Sh = k_s \frac{D_c}{D_s} \tag{6.12}$$

and the Reynolds number by:

$$Re = \frac{\omega\, d^2}{v} \tag{6.13}$$

When the film model is used to describe concentration polarization in turbulent flow regime, it is generally assumed that the flow is fully developed and entrance effects can be neglected. This can be done in most filtration systems that are used in large-scale industrial applications; therefore, exponent *d* in Equation (6.8) is 0. In laminar flow the situation is different, especially in modules consisting of relatively short thin channels. Here, the entrance length often covers the entire channel length.

Combination of Equations (6.5), (6.8), (6.10), (6.12), and (6.13) gives a relation for the mass transfer coefficient in a well-stirred batch cell:

$$\frac{C_s^w}{C_s^b} = \frac{\exp \frac{J_v D_c}{a\left(D_s \omega d^2\right)^{0.66} v^{-0.33}}}{R + (1 - R) \exp \frac{J_v D_c}{a\left(D_s \omega d^2\right)^{0.66} v^{-0.33}}} \tag{6.14}$$

where C_s^w and C_s^b are the solute concentrations at the membrane surface and in the bulk solution, D_s is the solute diffusion coefficient, D_c the diameter of the cell, *d* the stirrer length, v the viscosity of the solution, ω the stirrer speed, J_v the membrane flux, *R* the membrane rejection, and *a* a constant which has a value between 0.1and 0.6 depending on the cell design.

Equation (6.14) shows that concentration polarization for a given feed solution and batch-cell geometry increases with

increasing transmembrane flux and membrane rejection and decreases with increasing stirrer speed.

By introducing the appropriate relation for the hydraulic diameter and the length of the flow duct the concentration polarization can also be calculated for a rectangular channel and a circular tube. The hydraulic diameter is expressed in terms of the channel width *w* and the channel height *h* by:

$$d_{\mathrm{H}} = \frac{2wh}{w+h} \tag{6.15}$$

where *w* is the channel width and *h* the channel height.

Thus, the concentration polarization in a rectangular channel is given by:

$$\frac{C_{\mathrm{s}}^{\mathrm{w}}}{C_{\mathrm{s}}^{\mathrm{b}}} = \frac{\exp \frac{J_{\mathrm{v}}\left(\frac{wh}{w+h}\right)^{0.2} {}_{v}{}^{0.47}}{a\, D_{\mathrm{s}}^{0.66} {}_{u}{}^{0.8}}}{R + (1-R) \exp \frac{J_{v}\left(\frac{wh}{w+h}\right)^{0.2} {}_{v}{}^{0.47}}{a\, D_{\mathrm{s}}^{0.66} {}_{u}{}^{0.8}}} \tag{6.16}$$

where *w* is the channel width, *h* the channel height, and *u* the linear feed flow velocity.

The concentration polarization in a tubular membrane device can be calculated accordingly to:

$$\frac{C_{\mathrm{s}}^{\mathrm{w}}}{C_{\mathrm{s}}^{\mathrm{b}}} = \frac{\exp \frac{J_{\mathrm{v}} r_{\mathrm{t}}^{0.2} {}_{v}{}^{0.47}}{a\, D_{\mathrm{s}}^{0.66} {}_{u}{}^{0.8}}}{R + (1-R) \exp \frac{J_{\mathrm{v}} r_{\mathrm{t}}^{0.2} {}_{v}{}^{0.47}}{a\, D_{\mathrm{s}}^{0.66} {}_{u}{}^{0.8}}} \tag{6.17}$$

where r_{t} is the radius of the tubular membrane device.

The film model which pictures a laminar boundary layer at the membrane surface and a turbulent bulk solution can obviously not be applied to membrane devices with *laminar feed flow*. In laminar flow in thin-channel, tube, or capillary membrane modules the flow velocity is no longer constant over the cross section of the flow channel or capillary, and the velocity and concentration profiles develop over a much longer distance from the channel entrance than in turbulent flow. Therefore, entrance effects cannot be neglected in most commercially available membrane modules operated with laminar feed flow when the concentration polarization effects are described mathematically. Concentration polarization in laminar flow can be estimated with reasonable accuracy by describing the mass transfer in analogy to heat transfer in laminar flow by the Reynolds, the Schmidt, and the Sherwood numbers, taking the entrance length into account. The Sherwood number for laminar flow devices can be expressed by [161]:

$$Sh = a\left(ScRe\frac{d_{\mathrm{H}}}{L}\right)^{0.33} \tag{6.18}$$

In filtration processes with laminar flow the concentration polarization can be calculated as a function of the flow velocity, the channel height, and the channel length by combination of Equations (6.5), (6.6), and (6.18).

Although the film model assumes several simplifications such as steady-state conditions, that is, no variations in the concentration profile with time, no mass transport parallel to the membrane surface due to concentration or density gradients, and fully developed velocity profile in the turbulent flow regime, it has proved to be quite useful for describing concentration polarization in reverse osmosis and ultrafiltration under various feed flow conditions as long as no solutes precipitate at the membrane surface. The same is true for the treatment of concentration polarization in membrane modules with laminar flow.

The characteristic constants for the calculation the concentration polarization in practically used membrane filtration modules with turbulent and laminar flow are summarized in Table 6.

Table 6. Mass transfer correlation for the constants *a, b, c*, and *d* under various hydrodynamic feed flow conditions

Flow regime	Module geometry	Hydraulic diameter d_{H}	Characteristic constants			
			a	***b***	***c***	***d***
Turbulent	channel	$\frac{2hw}{h+w}$	0.023	0.8	0.33	0
Laminar	channel	$\frac{2hw}{h+w}$	1.62	0.33	0.33	0.33
Turbulent	tube	r_{t}	0.023	0.8	0.25	0
Laminar	tube	r_{t}	1.86	0.33	0.33	0.33

r_{t} = tube radius, *h* = channel height, *w* = channel width.

The *rigorous analysis* of concentration polarization involves the solution of the general transport equations for fluid and solute (i.e., the mass balance of solute and solvent and the momentum balance). The solution of the coupled continuity and

momentum equations is mathematically rather complex and it becomes even more complicated in the case of significant viscosity changes or precipitation of solutes at the membrane surface due to concentration polarization [159].

6.1.2. *Concentration Polarization with Solute Precipitation at the Membrane Surface*

In ultrafiltration and even more so in microfiltration the simple film model is no longer applicable because the feed solutions contain macromolecular components or suspended materials and the membrane fluxes are generally significantly higher than in reverse osmosis [162], [163]. Due to the relatively low diffusion coefficient of macromolecules or suspended particles, diffusive mass transport of components retained by the membrane from the membrane surface back into the bulk solution is slow and can not counterbalance the convective mass transport towards the membrane. Thus, the solute concentration at the membrane surface often exceeds the solubility of the feed solution constituents. This results in the precipitation of the feed solution components at the membrane surface and the formation of a dense solid layer on the membrane surface, which adds an additional hydrodynamic resistance to the membrane flux. When the membrane surface concentration reaches the solubility limit of the feed water constituents, the concentration difference between the membrane surface and the bulk solution and thus the diffusive transport from the membrane surface into the bulk solution reaches its maximum value. If the membrane flux and thus the convective solute transport towards the membrane is increased by increasing the hydrostatic pressure, more solutes are transported towards the membrane than can diffuse back. The additional solutes also precipitate and increase the thickness of the layer at the membrane surface and thus increase its resistance correspondingly, and in spite of an increase in the hydrostatic pressure driving force the membrane flux is not further increased. That means that in ultrafiltration and microfiltration of solutions containing high molecular components with limited solubility the membrane flux becomes independent of the hydrostatic pressure driving force as soon as these components precipitate and form a solid layer at the membrane surface. Typical results obtained in ultrafiltration of macromolecular solutions that precipitate at the membrane surface when a certain hydrostatic pressure or feed solution concentration is exceeded is shown in Figure 63. Here, membrane fluxes obtained when filtering hemoglobin solutions of different concentrations through an ultrafiltration membrane are shown as a function of the applied hydrostatic pressure and the hemoglobin concentration. The filtration apparatus is a stirred batch cell operated at constant stirrer speed. With pure water the membrane flux is directly proportional to the applied hydrostatic pressure. With the hemoglobin solution the membrane flux depends on the applied pressure only at very low pressures and approaches a constant, pressure-independent value at higher pressures. This value is a function of the hemoglobin concentration and is lowest at the highest concentration. Similar membrane flux – pressure relations are obtained in all filtration tests with feed solutions that contain macromolecular or particulate components which have low diffusivities and thus tend to precipitate at the membrane surface. Under conditions of precipitation-layer formation at the membrane surface the membrane flux can only be increased by a decrease in the boundary layer thickness or the bulk solution concentration. The boundary layer in membrane separation processes depends on hydrodynamic conditions of the system and depends on the degree of mixing, that is, on the Reynolds number achieved in the feed solution in turbulent flow or the channel height or capillary diameter under laminar flow conditions. The flux decline due to the formation of a precipitate layer also depends on the feed water constituents. Some components, such as proteins, dextrans, and metal oxides, form rather dense layers that affect the membrane flux significantly, while other materials, such as rigid particles, have a lesser effect on the flux.

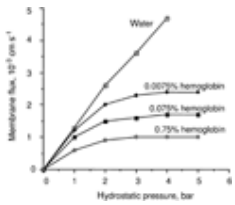

Figure 63. Ultrafiltration fluxes of water and hemoglobin solutions of different concentrations determined as a function of the hydrostatic pressure in a stirred batch cell at constant stirrer speed of 1800 rpm

The effect of the boundary layer thickness on the membrane flux has been demonstrated in several studies [164], [165]. Figure 64 shows the ultrafiltration flux of 1 % bovine albumin (M = 69 000) and 1 % dextran 110 (M = 110 000) as a function of the stirrer speed determined in a batch cell at 5 bar hydrostatic pressure [166]. With increasing stirrer speed, that is, decreasing boundary layer thickness, the membrane flux increases.

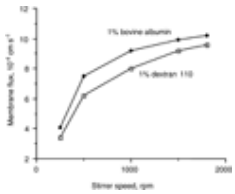

Figure 64. Membrane flux determined in a stirred batch cell with 1 % solutions of dextran 110 and bovine albumin at different stirrer speeds and constant hydrostatic pressure of 5 bar

The experimental results show that in the case of precipitation of solute at the membrane surface, as in ultrafiltration and more so in cross-flow microfiltration, the hydrodynamic resistance to membrane flux is not only a function of the membrane properties but is also strongly affected by the gel or cake layer formed by the retained solutes at the membrane surface. A simpler approach to describe the membrane flux in case of formation of a gel or a cake layer is to assume that the osmotic pressure of the feed solution in micro- and ultrafiltration can be neglected and express the flux in terms of the resistances of the membrane and the layer in series [167], [168]:

$$J_v = A\frac{1}{R_m + r_1\Delta z_1}\Delta p \qquad (6.20)$$

where J_v is the membrane flux, R_m the hydrodynamic resistance of the membrane, r_l the specific resistance of the layer, Δz_1 the thickness of the layer, Δp the hydrostatic pressure driving force, and A the membrane area.

Formation of a gel or cake layer in micro- and ultrafiltration is illustrated in Figure 65. This figure shows schematically a membrane, the gel layer at its surface, and the concentration and hydrostatic pressure profiles in the bulk solution, the laminar boundary layer, the gel layer, and the membrane. When the concentration of the retained solute at the membrane surface C_s^w reaches the gel layer concentration C_s^g, maximum diffusive transport of retained components from the membrane surface back into the bulk solution is attained. In the description of concentration polarization in terms of the film model it is assumed that diffusion is the only mechanism in the laminar boundary layer for the transport of components from the membrane surface to the bulk solution, and that in the steady state diffusive transport is just counterbalanced by the convective transport towards the membrane. This means that an increase in the hydrostatic pressure will lead only to short-term increase of the membrane flux and convective transport of solutes towards the membrane surface. Since diffusive transport from the membrane into the bulk solution is independent of the applied hydrostatic pressure, and additional components transported towards the membrane surface will precipitate and lead to an increase in gel layer thickness, the overall cake layer resistance will increase until the membrane flux has dropped back to its original value. This means that the membrane flux in ultra- or microfiltration is independent of the applied hydrostatic pressure when a gel layer is formed due to precipitation of the retained components at the membrane surface and it depends only on the transport of these components back into the bulk solution.

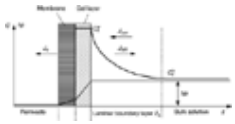

Figure 65. Schematic illustrating concentration polarization with the solutes forming a gel or cake layer at the membrane surface (C_s^b and C_s^w are the solute concentrations in the bulk solution and at the membrane wall, Δp is the pressure difference between the permeate and the feed, and J_v the filtrate rate)

Under conditions of precipitation of retained components at the membrane surface the membrane flux can only be increased by increasing the backtransport of the retained components into the bulk solution. This can be done, for example, by decreasing the boundary layer thickness or decreasing the bulk solution concentration, as can be seen from Equations (6.21) and (6.22), which describe the mass transport in the laminar boundary layer in a filtration device with turbulent bulk flow. For simplicity it is assumed that the dissolved components are completely retained by the membrane (i.e., $R = 1$):

$$\frac{C_s^w}{C_s^b} = \exp \frac{J_v Z_b}{D_s} \tag{6.21}$$

where C_s^b and C_s^w are the solute concentrations in the bulk solution and at the membrane surface, which is identical to the gel layer concentration C_s^g and constant for a given temperature and pressure. With $C_s^w = C_s^g$ = const. the flux is determined according to Equation (6.21) only by the bulk solution concentration and the boundary layer thickness. For certain mass transfer conditions and bulk solution concentration it is given by:

$$J_v = \frac{D_s}{Z_b} \ln C_s^b = k_s \ln C_s^b \tag{6.22}$$

Figure 66 shows the membrane flux obtained for three different feed solution concentrations in the filtration of solutions with constituents that precipitate at the membrane surface as a function of the applied hydrostatic pressures, as calculated by Equation (6.22). The flux versus pressure curves in Figure 66 show the same general pressure – flux relation as the experimental data obtained in ultrafiltration experiments with hemoglobin solutions of different concentrations (Fig. 63). The flux of pure water increases linearly with the applied pressure. The flux of a solution also increases linearly with pressure until a certain flux is attained. A further increase in pressure does not lead to a permanent increase in flux since the concentration of the retained solutes at the membrane surface has reached the solubility limit and precipitation occurs. The precipitated solute forms a dense layer which grows in thickness and in hydrodynamic resistance when, due to an increase in the applied pressure, the flux through the membrane is increased for a short time period and then reaches its former value again. The gel or cake layer at the surface of a micro- or ultrafiltration membrane not only impairs the membrane flux, but it can also influence separation properties of the membrane by acting as a permeable membrane which retains lower molecular weight materials not rejected by the membrane [158]. Especially in fractionation of biological fluid mixtures such as blood the formation of a gel layer impairs the entire process and can make a separation of different molecular weight components impossible.

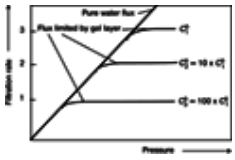

Figure 66. Schematic showing the permeate flux as a function of the applied pressure for the filtration of gel-layer-forming feed solutions of different concentrations

With the formation of a gel or cake layer during filtration of macromolecular or particulate solutions or suspensions with high-flux membranes, mathematical modeling of the process becomes significantly more complex, because the specific resistance of the gel layer depends on the layer-forming components and can be very different for different components. While some rigid materials such as latex balls form layers that resemble the packing of spheres, other materials such as certain proteins form gel-type layers. Many gel layers change their structure with time and their resistance often increases with time.

Furthermore, because of the low diffusivities of macromolecules and particles, diffusion can no longer be the only means of transportation of the components retained by the membrane in micro- and ultrafiltration. Grossly simplified, the process can be described as a three-step procedure (Fig. 67). In the gel layer the material retained by the membrane is more or less solidly integrated in a structure. The strength of this structure will depend on the intrinsic properties of the material but also on the hydrostatic pressure and eventually on the “age” of the gel layer. The first step in the process of bringing the material

back into the bulk solution therefore is to release it from the structure into the solution at the boundary layer. If the material is in solution it can diffuse or be transported by other means such as convection into the bulk solution. Steps 2 and 3 of the transport process are in parallel, while steps 1 and 2 or 3 are in series.

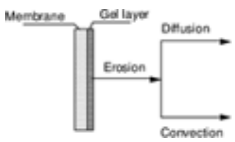

Figure 67. Schematic illustrating the mechanism of the transport of components retained by the membrane and accumulated in a gel layer from the membrane surface back into the bulk solution

In parallel processes, the fastest is the overall rate-determining step, and in consecutive processes the slowest is rate-determining. In ultra- and microfiltration with gel layer formation it is quite likely that the first step, that is, the release of the material from the gel layer structure, is the rate-determining step, and of the two parallel steps convective transport is fastest.

In many ultrafiltration tests a flux decline with time is observed even when all other operating parameters are kept constant. The original membrane flux can usually be restored when the membrane has been cleaned thoroughly with a proper solution. This is demonstrated in Figure 68, which shows schematically the results of a typical filtration test.

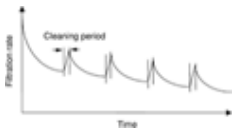

Figure 68. Membrane flux obtained during ultrafiltration of cheese whey at constant operating conditions but with daily cleaning periods as function of time (schematic)

The flux decline with time when all operating parameters are kept constant, as shown in Figure 68 and observed in many practical ultrafiltration tests, is difficult to explain with flow condition arguments only. It can, however, be rationalized by the assumption that the rate-limiting step for the transport of the material retained at the membrane surface into the bulk solution is the release rate of the material from the gel structure. This rate may well decrease with time when the gel structure ages and becomes more compact under pressure [169].

Modeling the mass transport at the membrane surface in micro- and ultrafiltration mathematically in terms of the hydrodynamic feed flow conditions, with inclusion of all phenomena observed in practical separation problems, is rather difficult when a gel layer has formed and the mass transfer at the membrane surface is no longer just a function of the diffusion coefficient and boundary layer thickness. Many mathematical models use fitting parameters and can be applied only to very specific conditions of membranes, feed solutions, and process parameters. A general model that fits all feed solutions and process parameters seems to be too complex to be of practical value.

6.2. Membrane Fouling and its Causes

In the literature the term “membrane fouling” is often used to describe a long-term flux decline caused by accumulation of certain materials at the membrane surface. The consequences of membrane fouling are obvious, but its causes are often not very clear [169], [170], [172]. Formation of a gel or cake layer is one cause of membrane fouling. Gel or cake layers can be formed by a variety of materials including inorganic precipitates such as $CaSO_4$, $Fe(OH)_2$, and other metal hydroxides; organic materials such as proteins, humic acids, and other macromolecular materials; and biological components such as microorganisms and their metabolic products. Membrane fouling can also occur without concentration polarization (i.e., direct transport to the membrane surface as in any mass separation process). The attachment of the substances to the membrane surface may be caused by adsorption due to hydrophobic interactions, van der Waals forces, or electrostatic forces. The fouling layer itself can be porous and thus permeable to aqueous solutions, as in the case of some inorganic precipitates, or highly impermeable when the layer consists of films of mineral oils or hydrophobic surfactants. The fouling mechanism also depends on the membrane process. In electrodialysis fouling is caused mainly by precipitation of polyelectrolytes or sparingly soluble salts such as $CaSO_4$ or $CaCO_3$. Membrane fouling in electrodialysis affects mainly anion-exchange membranes because most of the colloidal and macromolecular polyelectrolytes present in natural waters (e.g., humic acids and proteins) are negatively charged. In ultra- and microfiltration of biological solutions, but also in reverse-osmosis seawater desalination, biological fouling is a severe problem affecting the economics of the processes. In biomedical applications protein adsorption and protein denaturation at the membrane surface often impair the performance of the membranes.

The difference between concentration polarization and membrane fouling or scaling is illustrated schematically in Figure 69. Concentration polarization is a reversible process based on diffusion and takes place over a few seconds; it can be described adequately by a simple mathematical model and easily controlled by proper process design. Fouling is generally irreversible and the flux decline takes place over many minutes, hours, or even days. A constant flux is generally not reached. Membrane fouling is more difficult to describe and to control by experimental means. Membrane fouling is determined by a variety of different parameters including the feed solution constituents and their concentration, membrane material, and the fluid-dynamic system design. Membrane fouling can be caused by simple precipitation of insoluble materials or by reversible or irreversible adsorption of components at the membrane surface and within the membrane pores.

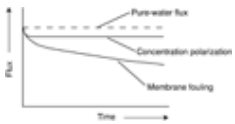

Figure 69. Schematic illustrating the difference between the flux decline due to concentration polarization and due to membrane fouling

The means of preventing or at least controlling membrane fouling are as heterogeneous as the different materials and mechanisms causing the fouling [171], [172]. The main procedures to avoid or control fouling involve:

- Pretreatment of the feed solution
- Membrane surface modifications
- Hydrodynamic optimization of the membrane module
- Membrane cleaning with appropriate chemical agents

Pretreatment of the feed solution can include chemical precipitation, prefiltration, pH adjustment, chlorination, and carbon adsorption. In some membrane module designs, for instance, hollow-fiber modules, the elimination of all particulate materials is of great importance for the proper function of the membrane.

Membrane surface modifications include the introduction of hydrophilic moieties or charged groups in the membrane surface by chemical means or plasma deposition.

Increasing the shear rate imposed by the feed solution on the membrane surface will in many cases reduce membrane fouling. High feed flow velocities and proper module design are efficient tools in controlling membrane fouling.

When, in spite of an adequate membrane and module design, the membrane flux decreases with operating time to an unacceptably low value it is necessary to clean the membrane to restore the flux in part or completely. Typical cleaning agents are acids and bases such as HNO_3 and NaOH, complexing agents, enzymes, and detergents. Another very effective method to minimize the effects of membrane fouling in microfiltration is backflushing, in which the applied pressure is reversed and the permeate pushed through the membrane to lift off fouling material precipitated on the feed side membrane surface and wash it out of the filtration device. Backflushing is done at certain time intervals for a few seconds.

6.3. Concentration Polarization in Other Membrane Separation Processes

In other membrane separation processes of technical relevance concentration polarization is of lesser significance than in reverse osmosis or ultrafiltration. It can generally be controlled and its detrimental effects be minimized by proper process design.

6.3.1. *Concentration Polarization in Dialysis*

In dialysis, concentration polarization is an accepted phenomenon. There is always depletion of the components in the boundary layer at the membrane surface facing the feed solution and accumulation at the membrane surface facing the dialysate (Fig. 70). However, the differences between the concentrations at the membrane surfaces and in the bulk solutions is low, because dialysis membranes are rather thick compared to the laminar boundary layers at their surfaces, and the transport in the membrane is rate-limiting for the overall mass transfer from the feed to the dialysate bulk solution. Thus, in dialysis concentration polarization effects are small.

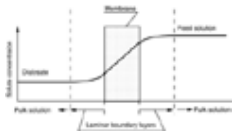

Figure 70. Schematic illustrating the effects of concentration polarization in dialysis

6.3.2. *Concentration Polarization in Electrodialysis*

In electrodialysis, concentration polarization decreases the ion concentration at the membrane surface facing the diluate cell and increases the concentration at the membrane surface facing the concentrate cell of the electrodialysis stack. The concentration increase may lead to a precipitation of salts. The concentration decrease lowers the limiting current density. Since the required membrane area is inversely proportional to the limiting current density, concentration polarization in electrodialysis has direct consequences for the overall process economics. It is controlled by using cell design and the flow velocities in the system to minimize the thickness of the laminar boundary layers at the membrane surfaces.

6.3.3. *Concentration Polarization in Pervaporation*

In pervaporation, concentration polarization can occur on both sides of the membrane. As in dialysis the boundary layer at the membrane surface facing the feed solution is depleted of the preferentially permeating components, which are then enriched at the membrane surface facing the permeate [173]. In pervaporation, concentration polarization is generally controlled by decreasing the laminar boundary layer thickness through hydrodynamic measures. This can lead to problems when high-flux thin-film composite membranes are used. At the permeate side the laminar boundary layer is always as thick as the porous substructure, and severe concentration polarization and capillary condensation may occur on the permeate side of the membrane. If the solubility of the permeating component in the feed solution is low the boundary layer at the feed side of the membrane may be depleted of the permeating component and the membrane flux and separation efficiency may be severely impaired. This, for example, is the case when organic solvent – water mixtures of limited solubility are treated.

6.3.4. *Concentration Polarization in Gas Separation*

In the separation of gases by solution-diffusion membranes, concentration polarization is of lesser importance and has generally little effect on the overall efficiency of the process. The reason is that the diffusivity of gases in the gas phase is much higher than in a solid polymer phase. Therefore, the transport rates in the boundary layer are comparatively high and there is less accumulation or depletion of components at the membrane surface. Nevertheless, concentration polarization in gas separation can not in all cases be neglected and must also be controlled by proper system design and flow conditions of the gases on the feed and permeate side of the membrane.

[Top of Page]

References

General References

1. H. K. Lonsdale: “The Growth of Membrane Technology”, *J. Membrane. Sci.* **10** (1982) 81. Links
2. R. W. Baker: *Membrane Technology and Applications*, J. Wiley & Sons, Chichester, U.K., 2004.
3. J. A. Nollet, (Abbe) Recherches sur les causes du Bouillonnement des Liquides. Histoir de l’Academie Royale des Sciences, Paris 1748, p. 57.
4. T. Graham: “On Osmotic Force”, *Philos. Trans. R. Soc. London* **114A** (1854) 177.
5. M. Traube: “Physiologie und wissenschaftliche Medizin”, *Arch. Anat. Physiol.* **87** (1867).
6. W. Pfeffer: *Osmotische Untersuchungen*, Leipzig 1877.
7. A. Fick: “Über Diffusion”, *Pogg. Ann.* **94** (1855) 59.
8. J. H. van’t Hoff: “Die Rolle des osmotischen Druckes in der Analogie zwischen Lösungen und Gasen”, *Z. Phys. Chem.* **1**, (1887) 481.
9. W. Nernst, *Z. Phys.Chem.* **2** (1888) 613.
10. M. Planck, *Ann. Phys. Chem. N.F.* **39** (1890) 161.
11. F. G. Donnan: “Theorie der Membrangleichgewichte und Membranpotentiale bei Vorhandensein von nicht dialysierenden Elektrolyten”, *Z. Elektrochem. Angew. Phys. Chem.* **17** (1911) 572.
12. H. Bechhold: “Durchlässigkeit von Ultrafilter”, *Z. Phys. Chem.* **64** (1908) 328.
13. R. Zsigmondy, W. Bachmann: “Über neue Filter”, *Z. Anorg. Allg. Chem.* **103** (1918) 119. Links
14. W. J. Elford: “A New Series of Graded Collodion Membranes Suited for General Bacteriological Use, Especially in Filterable Virus Studies”, *J. Pathol. Bacteriol.* **34** (1931) 505.
15. J. W. McBain, S. S. Kistler: “Ultrafiltration as a Test for Colloidal Constituents in Aqueous and Nonaqueous Systems”, *J. Phys. Chem.* **35** (1931) 130.
16. P. H. Prausnitz, J. Reitstotter: *Elektrophorese, Elektroosmose, Elektrodialyse*, Steinkopff Verlag, Dresden 1931.
17. US-Patent 1421341, 1922 (R. Zsigmondy).
18. W. J. Kolff, H. T. Berk: “The Artificial Kidney: A Dialyzer with Great Area”, *Acta Med. Scand.* **117** (1944) 121.
19. A. J. Staverman: “Non-Equilibrium Thermodynamics of Membrane Processes”, *Trans. Faraday Soc.* **48** (1952) 176.
20. O. Kedem, A. Katchalsky: “A Physical Interpretation of the Phenomenological Coefficients of Membrane Permeability”, *J. Gen. Physiol.* **45** (1961) 143. Links
21. R. Schlögl: *Stofftransport durch Membranen*, Steinkopff Verlag, Darmstadt 1964.
22. K. S. Spiegler, “Transport Processes in Ionic Membranes”, *Trans. Faraday Soc.* **54** (1958) 1408.
23. U. Merten: “Transport Properties of Osmotic Membranes”, in U. Merten (ed.): *Desalination by Reverse Osmosis*, MIT Press Cambridge, MA, 1966, pp. 15–54.
24. F. Helfferich, *Ion-Exchange*, McGraw-Hill., New York, NY 1962.
25. C. E. Reid, E. J. Breton: “Water and Ion Flow Across Cellulose Membranes”. *J. Appl. Polym. Sci.* **1** (1959) no. 2, 133. Links
26. S. Loeb, S. Sourirajan: “Seawater Demineralization by Means of a Semipermeable Membrane”, in R. Gould (ed.): *Advances in Chemistry*, ACS Series No. 38, American Chemical Society, Washington, DC 1962, pp. 117–132. US 3133132, 1964 (S. Loeb, S. Sourirajan).
27. R. E. Kesting: *Synthetic Polymeric Membranes*, McGraw-Hill, New York, NY, 1971.
28. H. Strathmann, K. Kock, P. Amar, R. W. Baker: “The Formation Mechanism of Asymmetric Membranes”, *Desalination* **16** (1975) 179–203. Links
29. J. E. Cadotte, R. I. Petersen: “Thin Film Reverse Osmosis Membranes: Origin, Development, and Recent Advances”, in A. F. Turbak (ed.): “*Synthetic Membranes*”, **Vol. I**, Desalination, ACS Symp. Ser. **153**, American Chemical Society, Washington, DC 1981, pp. 305–325.
30. R. L. Riley, H. K. Lonsdale, C. R. Lyons, U. Merten: “Preparation of Ultrathin Reverse Osmosis Membranes and the Attainment of Theoretical Salt rejection”, *J. Appl. Polym. Sci.* **11** (1967) 2143. Links
31. US 3 417 870, 1968 (T. D. Bray).
32. US 3 367 504, 1968 (J. C. Westmoreland).
33. US 3 228 876, 1966 (H. I. Mahon).
34. K. H. Meyer, W. Strauss, *Helv. Chim. Acta* **23** (1940) 795. Links
35. W. Juda, W. A. McRae: “Coherent Ion-exchange Gels and Membranes”, *J. Am. Chem. Soc.* **72** (1950) 1044. Links US 2 636 851, US 2 636 852, 1953 (W. Juda, W. A. McRae).
36. W. E. Katz: “The Electrodialysis Reversal (EDR) Process”, *Desalination* **28** (1979) 31. Links
37. T. Nishiwaki: “Concentration of Electrolytes Prior to Evaporation with an Electromembrane Process”, in R. E. Lacey, S. Loeb (eds.): *Industrial Processing with Membranes*, Wiley Interscience, New York, NY, 1972, pp. 83–106.
38. K. J. Lui, F. P. Chlanda, K. J. Nagasubramanian: “Use of Bipolar Membranes for Generation of Acid and Base: an Engineering and Economic Analysis”, *J. Membrane Sci.* **2** (1977) 108.
39. J. M. S. Henis, M. K. Tripodi: “A Novel Approach to Gas Separation Using Composite Hollow Fiber Membranes”, *Sep. Sci. Technol.* **15** (1980) 1059.
40. P. Aptel, J. Neel,. “Pervaporation”, in P. M. Bungay, H. K. Lonsdale, M. N. de Pinho (eds.): *Synthetic Membranes: Science, Engineering and Applications*, D. Reidel Publishing Company, Boston 1968, pp. 403–436.

41. W. Eykamp: “Microfiltration and Ultrafiltration”, in R. D. Nobel, S. A. Stern (eds.): *Membrane Separation Technology: Principles and Applications*, Elsevier, Amsterdam 1995.
42. Z. R. Raymond, G. K. Flemming, “Gas Permeation”, in W. S. Ho, K. K. Sirka (eds.): *Membrane Handbook*, Van Nostrand Reinhold, New York, NY 1992.
43. N. N. Li: “Permeation Through Liquid Surfactant Membranes”, *AIChE J.* **17** (1971) 459. Links
44. A. J. B. Kemperman, H. H. M. Rolevink, D. Bargeman, T. van den Boomgaard, H. Strathmann: “Stabilization of Supported Liquid Membranes by Interfacial Polymerization Top Layers”, *J. Membrane Sci.* **138** (1998) 43 Links
45. E. L. Cussler: “Membranes which Pump”, *AIChE J.* **17** (1971) 1300. Links
46. R. Haase: *Thermodynamik der Mischphasen*, Springer Verlag, Berlin 1956.
47. A. Katchalsky, P. F. Curran: *Nonequilibrium Thermodynamics in Biophysics*, Harvard University Press, Cambridge, MA 1967.
48. L. Onsager: “Reciprocal Relations in Irreversible Processes”, *Phys. Rev.* **37** (1931) 405; L. Onsager, *Phys. Rev.* **38** (1931) 2265.
49. J. A. Wesselingh, R. Krishna: *Mass Transfer*, Ellis Horwood, London, 1990.
50. G. Kraaijeveld: “The Maxwell-Stefan Description of Mass Transfer in Ion-Exchange and Electrodialysis”, PhD Thesis, University of Groningen, The Netherlands, 1994.
51. R. D. Noble, S. A. Stern (eds.): *Membrane Separations Technology, Principles and Applications*, Elsevier, Amsterdam 1995.
52. H. Strathmann: *Trennung von molekularen Mischungen mit Hilfe von Membranen*, Steinkopff Verlag, Darmstadt, Germany 1979.
53. A. J. Staverman, C. A. Kruising, D. T. F. Pals, *Trans. Faraday Soc.* **61** (1965) 2805.
54. P. W. Atkins: *Physical Chemistry*, Oxford University Press, Oxford, UK 1990.
55. E. A. Mason, A. P. Malinauskas: *Gas Transport in Porous Media*, Elsevier, Amsterdam 1983.
56. S. B. Tuwiner: *Diffusion and Membrane Technology*, Reinhold Publishing Corporation, New York 1962.
57. M. Planck, *Ann. Phys Chem. N.F.* **40** (1890) 161.
58. K. S. Spiegler: “Electrochemical Operations”, in F. C. Nachod, J. Schubert (eds.): *Ion-Exchange Technology*, Academic Press, New York, NY 1956.
59. K. J. Liu, F. P. Chlanda, K. J. Nagasubramanian: “Use of Bipolar Membranes for Generation of Acid and Base – an Engineering and Economic Analysis”, *J. Membrane Sci.* **2** (1977) 109.
60. H. Strathmann, J. J. Krol, H. J. Rapp, G. Eigenberger: “Limiting Current Density and Water Dissociation in Bipolar Membranes”, *J. Membrane Sci.* **125** (1997) 123 Links
61. S. Mafé, P. Ramirez, A. Alcaraz, V. Aguilella: “Ion Transport and Water Splitting in Bipolar Membranes”, in A. J. B. Kemperman (ed.): *Handbook Bipolar Membrane Technology*, Twente University Press, Enschede 2000.
62. R. Simon: “Water Splitting in Ion-Exchange Membranes”, *Electrochim. Acta* **30** (1993) 17
63. V. I. Zabolotsky, N. V. Sheldeshov, N. P. Gnusin: “Dissociation of Water Molecules in Systems with Ion-Exchange Membranes”, *Russ. Chem. Rev.* **57** (1988) 801.
64. M. C. Porter, A. S. Michaels: “Membrane Ultrafiltration”, *Chem. Technol.* **1** (1971) 56, 248, 440, 633.
65. C. T. Blaisdell, K. Kammermeyer: “Counter-Current and Co-Current Gas Separation”, *Chem. Eng. Sci.* **28** (1973) 1249. Links
66. S. T. Hwang, K. Ksammermeyer: “Membranes in Separations”, *Techniques of Chemistry*, Vol. VII, Wiley-Interscience, New York 1975.
67. S. Weller, W. A. Steiner: “Separation of Gases by Fractional Permeation Through Membranes”, *J. Appl. Phys.* **21** (1950) 279. Links
68. H. Strathmann, W. Gudernatsch, C.-M. Bell, K. Kimmerle: “Die Entwicklung von lösungsmittelselektiven Membranen und ihre Anwendung in der Gastrennung und Pervaporation”, *Chem. Ing. Tech.* **60** (1988) 590.
69. J. H. Perry: *Chemical Engineers Handbook*, McGraw-Hill, New York 1950.
70. S. T. Hwang, J. M. Thormann: “The Continuous Membrane Column”, *AIChE J.* **26** (1980) 558. Links
71. A. S. Michaels, *Trans. Am. Soc. Artif. Intern. Organs* **12** (1966) 387. Links
72. H. Strathmann: “Ion-Exchange Membrane Separation Processes”, *Membrane Science and Technology Series* **9**, Elsevier, Amsterdam 2004.
73. E. L. Huffmann, R. E., Lacey: “Engineering and Economic Considerations in Electromembrane Processing”, in R. E. Lacey, S. Loeb (eds.): *Industrial Processing with Membranes*, John Wiley & Sons, New York 1972, pp. 39–55.
74. L. H. Schaffer, M. S. Mintz: “Electrodialysis”, in K. S. Spiegler (ed.): *Principles of Desalination*, Academic Press, New York 1966, pp. 3–20.
75. J. R. Wilson: “Design and Operation of Electrodialysis Plants”, in J. R. Wilson, (ed.): *Demineralization by Electrodialysis*, Butterworth Publishers, London 1960.
76. H. Strathmann: “Electrodialysis and Related Processes”, in R. D. Nobel, S. A. Stern (eds.): *Membrane Separation Technology: Principles and Applications*, Elsevier, Amsterdam 1995.
77. T. Aritomi, T. van den Boomgaard, H. Strathmann: “Current-Voltage Curve of Bipolar Membrane at High Current Density”, *Desalination* **104** (1996) 13. Links
78. B. Bauer, H. Holdik, A. Velin: “Cell Equipment and Plant Design in Bipolar Membrane Technology”, in A. J. B. Kemperman (ed.): *Handbook Bipolar Membrane Technology*, Twente University Press, Enschede 2000
79. K. N. Mani: “Electrodialysis Watersplitting Technology”, *J. Membrane Sci.* **58** (1991) 117 Links
80. J. J. Krol, M. Jansink, M. Wessling, H. Strathmann: “Behavior of Bipolar Membranes at High Current Density, Water

Diffusion Limitation", *Sep. Purif. Technol.* **14** (1998) 41. Links

81. US 4 632 745, 1986 (A. J. Giuffrida., A. D. Jha, G. C. Ganzi).
82. US 4 747 929, 1988 (K. K. Siu, A. D. Dracut, G. C. Ganzi).
83. US 3 291 713, 1966 (E. J. Parsi).
84. S. Thate: "Untersuchung der electrochemischen Deionisation zur Reinstwasserherstellung", Ph.D.-Thesis, University of Stuttgart, Germany 2002.
85. T. Sata: "Ion-Exchange Membranes and Separation Processes with Chemical Reactions", *J. Appl. Electrochem.* **21** (1991) 283. 91 Links
86. US 3 615 024, 1971 (A. S. Michaels).
87. US 4 340 479, 1982 (D. B. Pall).
88. US 3 133 132, 1964 (S. Loeb, S. Sourirajan).
89. US 2106744, 1938 (H. P. Hood, M. E. Nordberg).
90. US 3962153, 1976 (R. W. Gore). K. Schneider, T. J. van Gassel, *Chem. Ing. Tech.* **7** (1985) 514.
91. R. L. Fleischer, P. B. Price, R. M. Walker, *Sci. Am.* **220** (1969) 30. Links
92. C. J. M. van Rijn: *Nano and Micro Engineered Membrane Technology*, Elsevier, Amsterdam 2004.
93. S. Manjikian: "Desalination Membranes from Organic Casting Solutions", *Ind. Eng. Chem. Prod. Res. Dev.* **6** (1967) 23.
94. R. E. Kesting: "Phase Inversion Membranes", in D. R. Lloyd (ed.): "Material Science of Synthetic Membranes", ACS Symp. Ser. **269** (1985) 131.
95. W. C. Hiatt, G. H. Vitzthum, K. Wagner, K. Gerlach, C. Josefiak: "Microporous Membranes Via Upper Critical Temperature Phase Separation", in D. R. Lloyd (ed.): "Material Science of Synthetic Membranes", ACS Symp. Ser. **269** (1985) 229–244.
96. H. Strathmann: "Production of Microporous Media by Phase Inversion Processes", in D. R. Lloyd (ed.): *Material Science of Synthetic Membranes*, ACS Symp. Ser. **269** (1985) 165.
97. K. Kamide, S. Manabe: "Role of Microphase Separation Phenomena in the Formation of Porous Polymeric Membranes", in D. R. Lloyd (ed.): "Material Science of Synthetic Membranes", ACS Symp. Ser. **269** (1985) 197.
98. H. Strathmann, P. Scheible, R. W. Baker, *J. Appl. Polym. Sci.* **15** (1971) 811. Links
99. J. G. Wijmans, J. P. B. Baaij, C. A. Smolders, *J. Membrane Sci.* **14** (1983) 263. Links
100. M. A. Frommer, I. Feiner, O. Kedem, R. Block, *Desalination* **7** (1970) 393. Links
101. S. Sourirajan, B. Kunst: "Cellulose Acetate and Other Cellulose Ester Membranes", in S. Sourirajan (ed.): *Reverse Osmosis and Synthetic Membranes*, National Research Council Canada, Ottawa, Canada 1977.
102. L. Broens, F. W. Altena, C. A. Smolders, *Desalination* **32** (1980) 33.
103. R. E. Kesting, *J. Appl. Polym. Sci.* **17** (1973) 1771. Links
104. R. Kesting, *Synthetic Polymeric Membranes, a Structural Perspective*, Wiley-Interscience, New York, NY 1985.
105. H. Yasuda: "Composite Reverse Osmosis Membranes Prepared by Plasma Polymerization", in S. Sourirajan (ed.): *Reverse Osmosis Synthetic Membranes*, National Research Council Canada, Ottawa 1977.
106. L. T. Rozelle, J. E. Cadotte, K. E. Cobian, C. V. Kopp: "Nonpolysaccharide Membranes for Reverse Osmosis: NS 100 Membranes for Reverse Osmosis and Synthetic Membranes", in S. Sourirajan (ed.): *Reverse Osmosis Synthetic Membranes*, National Research Council Canada, Ottawa 1977.
107. US 4 277 344, 1981 (J. E. Cadotte).
108. J. E. Cadotte: "Evolution of Composite Reverse Osmosis Membranes", in D. R. Lloyd (ed): "Material Science of Synthetic Membranes", ACS Symp. Ser. **269** (1985) 229.
109. A. J. Burggraaf, L. Cot: "Fundamentals of Inorganic Membrane Science and Technology", *Membrane Science and Technology Series* **4**, Elsevier, Amsterdam 1996.
110. K. Kammermeyer: "Gas and Vapor Separation by Means of Membranes", in S. Perry (ed.): *Progress in Separation and Purification*, Wiley-Interscience, New York 1972.
111. R. P. Buck: "Electroanalytical Chemistry of Membranes", *CRC Crit. Rev. Anal. Chem.* **5** (1976) 323.
112. S. G. Kimura, S. L. Matson, W. J. Ward III: "Industrial Applications of Facilitated Transport", in N. N. Li (ed.): *Recent Developments in Separation Science*, Vol. V, CRC Press, Boca Raton, FL 1979, pp. 11–25.
113. R. W. Baker, J. C. Roman, H. K. Lonsdale: "Liquid Membranes for the Production of Oxygen-Enriched Air", *J. Membrane Sci.* **31** (1987) 15. Links
114. E. L. Cussler: *Diffusion*, Academic Press, Cambridge, MA 1984.
115. W. J. Ward III: "Analytical and Experimental Studies of Facilitated Transport", *AIChE J.* **16** (1970) 405. Links
116. J. S. Schultz, J. D. Goddard, S. R. Suchdeo: "Facilitated Transport via Carriermediated Diffusion in Membranes, Part I: Mechanistic Aspects, Experimental Systems and Characteristic Regimes", *AIChE J.* **20** (1974) 417. Links
117. W. C. Babcock, R. W. Baker, E. D. LaChapelle, K. L. Smith, *J. Membrane Sci.* **7** (1980) 71. Links
118. N. N. Li: "Permeation Through Liquid Surfactant Membranes", *AIChE J.* **17** (1971) 459. Links
119. D. S. Flett: "*Ion Exchange Membranes*", Ellis Horwood, Chichester, UK 1983.
120. R. Kunin: *Ion-Exchange Resins*, Wiley-Interscience, New York 1958.
121. T. Sata, K. Motani, Y. Ohaski: "Perfluorinated Exchange Membrane "Neosepta-F" and its Properties", in. D. S. Flett (ed.): *Ion Exchange Membranes*, Ellis Horwood, Chichester, UK 1983.
122. US 891562, 1963 (F. deKörösy, Y. Schorr).
123. P. Zschocke, D. Quellmalz: "Novel Ion Exchange Membranes Based on an Aromatic Polyethersulfone", *J. Membrane Sci.* **22** (1985) 325. Links

124. US 3 784 399, 1974 (W. G. Grot).
125. W. Grot: “Perfluorinated Cation Exchange Polymers”, *Chem. Ing. Tech.* **47** (1975) 617.
126. A. Eisenberg, H. L. Yeager: “Perfluorinated Ionomer Membranes”, ACS Symp. Ser. **180** (1982).
127. T. Sata: “Monovalent Cation Permselective Ion-Exchange Membrane”, *Kolloid Z. Z. Polym.* **250** (1972) 980.
128. T. Sata: “Recent Trends in Ion-Exchange Research”, *Pure Appl. Chem.* **58** (1986) 1613. Links
129. E. Korngold, F. De Körösy, R. Rahav, M. F. Taboch: “Fouling of Anion Selective Membranes in Electrodialysis”, *Desalination* **8** (1970) 195. Links
130. Y., Kobuchi, Y. Matsunaga, Y. Noma: “Application of Ion-Exchange Membranes to the Recovery of Acids by Diffusion Dialysis and Electrodialysis”, in B. Sedlaceek, J. Kahovec (eds.): *Synthetic Polymer Membranes*, W. de Gruyter & Co., Berlin 1987.
131. R. Simons: “Preparation of a High Performance Bipolar Membrane”, *J. Membrane Sci.* **78** (1993) 13. Links
132. F. W. Wilhelm: “Bipolar Membrane Preparation”, in A. J. B. Kemperman (ed.): *Bipolar Membrane Technology*, Twente University Press, Enschede, The Netherlands 2000.
133. A. G. Fane, C. J. D. Fell, A. G. Waters: “The Relationship Between Membrane Pore Characteristics and Flux for Ultrafiltration Membranes”, *J. Membrane Sci.* **9** (1981) 245. Links
134. K. J. Kim, A. G. Fane, C. J. D. Fell, T. Suzuki, M. R. Dickson: “Quantitative Microscopic Study of Surface Characteristics of Ultrafiltration Membranes”, *J. Membrane Sci.* **54** (1990), 89. Links
135. A. K. Fritsche, A. R. Arevalo, M. D. Moore, C. O’Hara: “Surface Structure and Morphology of Polyacrylonitrile Membranes by Atomic Force Microscopy”, *J. Membrane Sci.* **46** (1992) 167.
136. J. D. Ferry, *Chem. Rev.* **18**, (1936) 373. Links
137. M. C. Porter: “Membrane Filtration”, in P. A. Schweizer (ed.): *Handbook of Separation Techniques for Chemical Engineers*, McGraw-Hill, New York 1979.
138. M. Cheryan: *Ultrafiltration and Microfiltration Handbook*, Technomic Publishing Co., Lancaster, PA 1986.
139. G. Capanelli, F. Vigo, S. Munary: “UF Characterization Methods”, *J. Membrane Sci.* **15** (1983) 289.
140. I. M. Wienk, B. Folkers, T. van den Boomgaard, C. A. Schmolders: “Critical Factors in the Determination of the Poresize Distribution of UF Membranes Using a Liquid Displacement Method”, *Sep. Sci. Technol.* **29** (1994) 1433.
141. F. P. Cuperus, D. Bargeman, C. A. Smolders: “Permporometry. The Determination of Active Pores in Ultrafiltration Membranes”, *J. Membrane Sci.* **71** (1992) 75.
142. F. P. Cuperus, D. Bargeman, C. A. Smolders: “Critical Points in the Analysis of Membrane Pore Structure by Thermoporometry”, *J. Membrane Sci.* **66** (1992) 45. Links
143. W. J. Koros, R. T. Chern: “Separation of Gaseous Mixtures Using Polymer Membranes”, in R. W. Rousseau (ed.): *Handbook of Separation Process Technology*, John Wiley & Sons, New York 1987.
144. D. R. Paul: “The Solution–Diffusion Model for Swollen Membranes”, *Sep. Purif. Meth.* **5** (1976) 33.
145. G. Morel, D. R. Paul: “CO_2 Sorption and Transport in Miscible Polyphenylene Oxide/Polystyrene Blends”, *J. Membrane Sci.* **10** (1982) 273. Links
146. A. R. Berens, H. B. Hopfenberg: “Diffusion of Organic Vapors at Low Concentrations in Glassy PVC, Polystyrene and PMMA”, *J. Membrane Sci.* **10** (1982) 283. Links
147. R. Günther, B. Perschall, D. Reese, J. Hapke: “Engineering of High Pressure Reverse Osmosis”, *J. Membrane Sci.* **121** (1996) 953.
148. US 3 367 504, 1968 (J. C. Westmoreland).
149. US 5 169 530, 1992 (T. J. Eckman).
150. P. T. L. Brian: “Mass Transport in Reverse Osmosis”, in U. Merten (ed.): *Desalination by Reverse Osmosis*, MIT Press, Cambridge, MA 1966, pp. 161.
151. W. F. Blatt, A. Dravid, A. S. Michaels, L. M. Nelsen: “Solute Polarization and Cake Formation in Membrane Ultrafiltration: Causes, Consequences and Control Techniques”, in J. E. Flinn (ed.): *Membrane Science and Technology*, Plenum Press, New York 1970, pp. 47–97.
152. G. Jonsson, C. E. Boesen: “Polarization Phenomena in Membrane Processes”, in G. Belfort (ed.): *Synthetic Membrane Processes*, Academic Press, New York 1984.
153. M. C. Porter: “Concentration Polarization with Membrane Ultrafiltration”, *Ind. Eng. Chem. Prod. Res. Dev.* **11** (1972) 234.
154. A. G. Fane: “Ultrafiltration of Suspensions”, *J. Membrane Sci.* **20** (1984) 249. Links
155. R. Rautenbach, R. Albrecht: *Membrane Processes*, John Wiley, New York 1989.
156. G. Belfort: “Fluid Mechanics in Membrane Filtration”, *J. Membrane Sci.* **40** (1989) 123. Links
157. H. Strathmann: “Untersuchungen zur Konzentrationsüberhöhung bei der Membranfiltration”, *Chem. Ing. Tech.* **44** (1972) 1160.
158. R. W. Baker, H. Strathmann: “Ultrafiltration of Macromolecular Solutions with High-Flux Membranes”, *J. Appl. Polym. Sci.* **14** (1970) 1197. Links
159. A. G. Fane, C. J. D. Fell, D. Wiley, R. M. McDonogh: “Concentration Polarization, Mass Transfer and Fluid Dynamics in Membrane Systems”, Proceedings of the ESMST Summer School on Engineering Aspects of Membrane Processes, Aarhus, Denmark, June 2–6, 1986.
160. T. K. Sherwood, P. T. L. Brian, R. E. Fischer, L. Dresner: “Salt Concentration of Phase Boundaries in Desalination by Reverse Osmosis”, *Ind. Eng. Chem. Fundam.* **4** (1965) 113.
161. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena*, John Wiley & Sons, New York 1965.
162. J. G. Wijmans, S. Nakao, C. A. Smolders: “Flux Limitation in Ultrafiltration: Osmotic Pressure Model and Gel Layer

Model", *J. Membrane Sci.* **20** (1984) 115. Links

163. R. M. McDonogh, C. J. D. Fell, A. G. Fane: "Predicted Flux Time Behavior of Flocculating Colloids in Unstirred Ultrafiltration", in E. Drioli, M. Nakagaki (eds.): *Membrane and Membrane Processes*, Plenum Press, New York 1986, pp. 199–208.
164. C. K. Colton: "Analysis of Membrane Processes for Blood Purification", *Blood Purif.* **5** (1987) 202. Links
165. M. J. Lysaght "Mass Transfer in Clinical Blood Ultrafiltration Devices – A Review", in T. H. Frost (ed.): *Technical Aspects of Renal Dialysis*, Pittmann Medical Publishers, London 1978, pp. 81–95.
166. H. Strathmann: "Konzentrationsüberhöhung bei der Filtration von makromolekularen Lösungen", *Chem. Ing. Techn.* **45** (1973) 825.
167. A. G. Fane: "Factors Affecting Flux and Rejection in Ultrafiltration", *J. Sep. Process Technol.* **4** (1983) no. 1, 15.
168. A. L. Zydney, C. K. Colton: "A Concentration Polarization Model for the Filtrate Flux in Cross-Flow Microfiltration of Particulate Suspensions", *Chem. Eng. Commun.* **47** (1986) 1. Links
169. H. Reihanian, C. R. Robertson, A. S. Michaels: "Mechanisms of Polarization and Fouling of Ultrafiltration Membranes by Proteins", *J. Membrane Sci.* **16** (1983) 237. Links
170. H.C. van der Horst, J. H. Haanemeijer: "Cross Flow Microfiltration in the Food Industry", *Desalination* **77** (1990) 235.
171. H. F. Ridgway, C. Justice, A. Kelly, B. H. Olson: "Microbial Fouling of Reverse Osmosis Membranes Used in Advanced Wastewater Treatment Technology: Chemical, Bacteriological and Ultrastructural Analyses", *Appl. Environ. Microbiol.* **45** (1983) 1066. Links
172. J. L. Nilsson: "Fouling of an Ultrafiltration Membrane by a Dissolved Whey Protein Concentrate", *J. Membrane Sci.* **36** (1988) 147. Links
173. R. Rautenbach, R. Albrecht: "Investigation of Mass Transport in Asymmetric Pervaporation Membranes", *J. Membrane Sci.* **36** (1988) 445. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Mercaptoacetic Acid and Derivatives

ROBERT RIPPEL, Hoechst AG, Frankfurt, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 2
5. **Analysis** 2
6. **Transportation and Storage** 2
7. **Toxicology** 2
8. **Uses** 3
9. **References** 3

1. Introduction

Mercaptoacetic acid [*68-11-1*], thioglycolic acid, $HSCH_2COOH$, is the simplest and industrially most important mercaptocarboxylic acid. The compound was first prepared in 1862 by CARIUS from chloroacetic acid and potassium hydrogen sulfide. There are numerous patents dealing with the applications of mercaptoacetic acid in the cosmetic and plastic industries.

2. Physical Properties

Mercaptoacetic acid is a clear, colorless liquid with a characteristic odor. Some physical properties are as follows:

M_r	92.11
mp	−16.5 °C
bp (2.3 kPa)	110–112 °C
bp (0.1 kPa)	79–80 °C
n_D^{20}	1.5027
d_4^{20}	1.325
Heat of combustion	1450 kJ/mol
Vapor pressure (30 °C)	0.02 kPa
Vapor density	3.18
Flash point	126 °C
Dissociation constants (in $NaClO_4$ solution at 20 °C)	
K_1	$(3.82 \pm 0.1)\,pK_a$
K_2	$(9.30 \pm 0.1)\,pK_a$

Mercaptoacetic acid is miscible with water, mono- and polyalcohols, ethers, ketones, esters, chlorinated hydrocarbons, and aromatic hydrocarbons, but not with aliphatic hydrocarbons. The esters of mercaptoacetic acid are mostly colorless liquids with characteristic odors. The cetyl and stearyl esters are pale yellow waxes [1].

3. Chemical Properties

Mercaptoacetic acid has two reactive centers: the mercapto group and the carboxyl group. The compound forms salts of the carboxyl and mercapto groups, esters, amides, anilides, thioethers, etc. Oxidation of mercaptoacetic acid, which occurs even on standing in air, gives dithiodiglycolic acid [*505-73-7*], $HOOCCH_2SSCH_2COOH$. The reaction is accelerated by traces of metals such as copper, iron, and manganese [2]. In concentrated solutions (80 %), linear and cyclic polycondensation products are formed, including a tetracarboxylic acid with a 1,4-dithiane ring structure [3]. Thioethers are formed by reactions of the sodium or potassium salts of the mercapto group with alkyl halides or by addition of mercaptoacetic acid or its derivatives to double or triple bonds. Depending on the catalyst used, either the Markovnikov or anti-Markovnikov product is obtained [4]. Mercaptoacetic acid and its derivatives react with aldehydes and ketones to form mercaptals, mercaptols, or α,β-unsaturated thioethers. The addition reactions are catalyzed by acids (e.g., mineral acids and toluenesulfonic acids) [5]. In the formation of oxathiolanes, the mercapto and carboxyl groups react simultaneously with a carbonyl group. For example, the reaction of mercaptoacetic acid with cyclohexanone affords 2,2-pentamethylene-1,3-oxathiolane-5-one [6]. The strongly reducing nature of mercaptoacetic acid is of importance both in investigations of biochemical redox and enzyme systems and in industrial applications such as cold-waving preparations and modification of wool. The use of mercaptoacetic acid for the permanent waving

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_265

of hair is based on the reductive cleavage of the cystine disulfide bridges of keratin [7].

4. Production

Industrially, mercaptoacetic acid is produced from chloroacetic acid or its salts and sodium or potassium hydrogen sulfide. Thiodiglycolic acid, dithiodiglycolic acid, and glycolic acid may be formed as side products. The mercaptoacetic acid is isolated from the acidified reaction mixture by extraction with organic solvents (ethers, alcohols, or chlorinated hydrocarbons) and purified by distillation. Both batch and continuous processes have been described in the patent literature as well as processes that operate under a partial pressure of hydrogen sulfide or carbon dioxide [8]. In one process, sulfur wastes, which arise from the reduction of aromatic nitro compounds, are used instead of alkali-metal hydrogen sulfides [9], while in another process mercaptoacetic acid is obtained by cleavage of imidazolyl mercaptoacetic acid with sodium sulfide [10]. It can also be produced through cleavage of the Bunte salt NaO_3SSCH_2COOH, prepared from chloroacetic acid and sodium thiosulfate, with dilute sulfuric acid; by reduction of dithioglycolic acid (prepared from chloroacetic acid and alkali-metal polysulfides); and by decomposition of the xanthate ester obtained from chloroacetic acid and potassium ethylxanthate [11]. Anhydrous mercaptoacetic acid is obtained by distillation of technical mercaptoacetic acid with toluene. During production, it is important to avoid contact of the mercaptoacetic acid with metals that catalyze its dimerization.

Esters of mercaptoacetic acid can be produced either by esterification with alcohols, whereby the alcohol can simply be used as the extractant after the synthesis of the acid, or by reaction of the corresponding alkyl chloroacetate with sodium thiosulfate [12].

5. Analysis

Mercaptoacetic acetic acid gives color reactions with numerous metal salts. Iron salts and sodium pentacyanonitrosylferrate(III) give a violet coloration [13]. Mercaptoacetic acid is analyzed quantitatively by titration with iodine [14]. Spectroscopic and chromatographic methods in industrial use include determination by HPTLC [15], analysis of cold-waving preparations by HPLC [16], and detection by gas chromatography [17]. Reference spectra for spectroscopic investigations can be found in [18], [19].

6. Transportation and Storage

Commercial forms of mercaptoacetic acid are the 99 % acid (d_4^{20} 1.325) and the 80 % acid (d_4^{20} 1.27). Standardized mercaptoacetic acid should contain not more than 2 % dithiodiglycolic acid and a maximum of 1 ppm iron, lead, copper, and arsenic [20]. Isooctyl mercaptoacetate [*25103-09-7*] and glycerol monothioglycolate [*30618-84-9*] are sold water-free. Mercaptoacetic acid must be transported in stainless steel (V_4A, V_2A) or polyethylene-lined polyester containers. The following transportation codes apply [21]:

IMDG (GGVSee)	Class 8 D E – F 8154 UN no. 1940
RID (GGVE)	Class 8 Rn 801 Section 32 b
ADR (GGVS)	Class 8 Rn 2801 Section 32 b
ADNR	Class 8 Rn 6801 Section 32 b
GefStoffV	607-090-00-6
GB Blue Book	Corrosives and IMDG code E 8154
United States	CFR 49: § 172.101 Corrosive M
ICAO/IATA-DGR	Class 8 UN no. 1940

7. Toxicology

Contact with mercaptoacetic acid leads to irritation of the skin and eye. The vapor causes irritation of the mucous membranes, coughing and shortness of breath after inhalation, badly healing wounds, and blisters [22]. Protective gloves and eyewear are therefore necessary when working with mercaptoacetic acid [22]. The TLV value has been set at 1 ppm (5 mg/m^3) [23]. The oral LD_{50} in the rat is 261 mg/kg [24]; more recent tests on mice are reported in [25]. Most derivatives of mercaptoacetic acid and commercial products containing them, for which legal maximum concentrations apply (see → Hair Preparations, Chap. 10.), are considerably less toxic and therefore do not present a health risk in normal use [26].

8. Uses

Since ca. 1950 mercaptoacetic acid and its derivatives, in particular the ammonium and ethanolammonium salts and to an increasing extent glycerol monothioglycolate, have been used for the production of cold-waving preparations [27] (→ Hair Preparations, Chap. 6.). The calcium salt is used in depilatories. Mercaptoacetic acid and derivatives are also used for industrial unhairing (e.g., leather production) and for permanent forming of wool (e.g., in hat making). Sulfur-containing organotin compounds of the type dialkyltin bis(isooctyl thioglycolate) and alkyltin tris(isooctyl thioglycolate) are the most important thermo-stabilizers for PVC [28]. The low toxicity and extremely low migration of, for example, dioctyltin bis(isooctyl thioglycolate) with various amounts of octyltin tris(isooctyl thioglycolate), has led to their use worldwide for the production of food packaging materials and pipes for drinking water [29]. Esters, particularly of higher alcohols such as isooctanol and decanol, and amides of mercaptoacetic acid are used as polymerization catalysts and antioxidants for the production of plastics and rubbers [30]. The sodium salt of mercaptoacetic acid is a component of the Brewer nutritive medium for anaerobic bacteria [31].

Mercaptoacetic acid is used in analytical chemistry for the separation of iron and aluminum and as a reagent for iron [13], nitrite, uranium, vanadium, chromium, copper, molybdenum, and palladium [32]. It is used in fast atom bombardment mass spectrometry as a matrix compound [33]. A mild process for the isolation of lignin involves the use of mercaptoacetic acid with boron trifluoride [34]. Other economically important uses include the production of derivatives of thiophene and dihydrothiophene, which are used as precursors for pharmaceuticals [35], and the synthesis of intermediates for dyes (e.g., *S*-aryl mercaptoacetic acids). A review of chelate complexes of mercaptoacetic acid is given in [36]. World production of mercaptoacetic is estimated to be $15-20\times10^3$ t/a.

9. References

1. G. T. Walker, *Seifen Öle Fette Wachse* **88** (1962) 11.
2. M. Kharasch, R. R. Legaut, A. B. Wilder, R. W. Gerard, *J. Biol. Chem.* **113** (1936) 537.
3. A. Schöberl, G. Wiehler, *Justus Liebigs Ann. Chem.* **595** (1955) 101.
4. *Houben-Weyl,* **9,** 199.
5. *Houben-Weyl,* **7/2 b,** 1938.
6. Carlisle Chemical Works Inc., US 3 209 012, 1965 (C. H. Miller et al.). P. D. Klemmensen, O. Z. Mortensen, S. O. Lawesson, *Tetrahedron* **26** (1970) 4641.
7. T. Bersin, J. Steudel, *Ber. Dtsch. Chem. Ges.* **B 71** (1938) 1015.
8. G. T. Walker, *Seifen öle Fette Wachse* **88** (1962) 402, 431. Denki Kagaku Kogyo K. K., JP 8 034 794, 1977. Hoechst, DE-OS 2 711 867, 1977 (H. Klug).
9. Martin-Luther-Universität, Halle-Wittenberg, DD 238 229, 1986 (H. Matschner, H. Stange, E. Mendow).
10. C. F. Spiess u. Sohn, DE-OS 2 832 977, 1978 (R. Himmelreich).
11. *Houben-Weyl,* **9,** 3 ff.
12. VEB Arzneimittelwerk Dresden, DD 270 063, 1989 (H. Stange, E. Mendow, E. Loewe, G. Nauwald).
13. D. Geffken, K. H. Surborg, *Dtsch. Apoth. Ztg.* **128** (1988) 1235.
14. *DAB 7* (1968) p. 317.
15. R. Klaus, *Chromatographia* **20** (1985) 235.
16. Y. Fukuda, F. Nakamura, Y. Morikawa, *Eisei Kagaku* **31** (1985) 235. J. Rooselar, D. M. Liem, *Int. J. Cosmet. Sci.* **3** (1981) 37.
17. N. Goetz, *Cosmet. Sci. Technol. Ser.* **4** (1985) 65. N. Goetz, P. Gataud, P. Bore, *Analyst (London)* **104** (1979) 1062. U. Hannestad, B. Soerbo, *J. Chromatogr.* **200** (1980) 171.
18. *The Aldrich Library of Infrared Spectra,* 3rd ed., Aldrich Chemical Company, Milwaukee 1981, p. 3096.
19. *The Aldrich Library of NMR-Spectra,* **vol. II,** Aldrich Chemical Company, Milwaukee 1974, p. 176 A.
20. *Kirk-Othmer* 2nd ed., **20,** 200.
21. GefStoffV, 26th Aug. 1986, *Product Safety Data, Mercaptoacetic Acid,* E. Merck, Darmstadt.
22. Kühn-Birett: *Merkblätter Gefährliche Arbeitsstoffe,* 13th Supplement, Ecomed Verlag, Landsberg/Lech 1980.
23. *Hazards in the Chemical Laboratory,* 3rd ed., The Royal Society of Chemistry, London 1981.
24. E. Merck, *Product Safety Data, Mercaptoacetic Acid,* Darmstadt 1988.
25. E. W. Schafer, Jr., W. A. Bowles, Jr., *Arch. Environ. Contam. Toxicol.* **14** (1985) 111.

26. B. Bach, *Seifen öle Fette Wachse* **105** (1979) 405, 448.
27. M. E. Marti, *Seifen öle Fette Wachse* **116** (1990) 23.
28. A. Holger in G. Becker, D. Braun (eds.): *Kunststoffhandbuch,* 2nd ed., **vol. 2/1,** Hanser Verlag, München 1985, p. 527.
29. R. Franck, H. Mühlschlegel: *Kunststoffe im Lebensmittelverkehr,* Carl Heymanns Verlag, Köln 1977. United States Food and Drug Administration, *Fed. Regist.* **48** (1983) no. 35, 7169.
30. E. Ceausescu et al., *J. Macromol. Sci. Chem.* **A 22** (1985) 525. *Comprehensive Polymer Science,* **vols. 1 – 7,** Pergamon Press, Oxford – New York – Beijing – Frankfurt – Sao Paulo – Sydney – Tokyo – Toronto. *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vols. 1 – 17 (incl. Suppl. Vol.),** J. Wiley and Sons, New York – Brisbane – Chichester – Toronto 1985 – 1989.
31. *DAB 9,* (1986) p. 423.
32. E. Merck: *Organische Reagenzien für die Spurenanalyse,* Darmstadt 1975.
33. J. L. Gower, *Biomed. Mass. Spectrom.* **12** (1985) 191.
34. I. Mogharab, W. G. Glasser, *Tappi* **59** (1976) 110.
35. DE 1 055 007, 1959 (H. Fießelmann). DE 1 083 830, 1960 (H. Fießelmann).
36. J. Malowska, J. Szmich, *Zesz. Nauk. Politech. Lodz. Technol. Chem. Spozyw.* **38** (1984) 5.

My Profile Log

WILEY InterScience®

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Mercury, Mercury Alloys, and Mercury Compounds

Standard Article

Matthias Simon[1], Peter Jönk[2], Gabriele Wühl-Couturier[3], Stefan Halbach[4]

[1]Rheinzink, Datteln, Germany
[2](E. Merck), Darmstadt, Germany
[3]Thor Chemie, Speyer, Germany
[4]GSF-Forschungszentrum für Umwelt und Gesundheit GmbH, München, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a16_269.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Properties
2.1. Physical Properties
2.2. Chemical Properties
3. Resources and Raw Materials
3.1. Deposits
3.2. Secondary Sources
4. Production
4.1. Extraction from Primary Sources
4.1.1. Dressing
4.1.2. Processing to Metallic Mercury
4.1.3. Furnace Systems
4.2. Extraction from Secondary Sources
4.3. Condensation of Mercury from Furnace Off-Gas
4.4. Treatment of the Stupp
5. Environmental Protection
5.1. Natural Distribution of Mercury
5.1.1. Mercury in Soil, Plants, and Animals
5.1.2. Mercury in Food
5.2. Mercury Emissions
5.2.1. Gas Purification
5.2.2. Water Purification
6. Quality Specifications
7. Chemical Analysis
8. Storage and Transportation
9. Uses
10. Mercury Alloys
10.1. General
10.2. Production of Amalgams
11. Mercury Compounds
11.1. Inorganic Compounds
11.1.1. Mercury Chalconides
11.1.2. Mercury Halides
11.1.3. Mercury Pseudohalides

11.1.4. Acetates, Nitrates, Sulfates
11.1.5. Mercury - Nitrogen Compounds
11.1.6. Analysis, Storage, and Transportation; Protective Measures
11.2. Organic Compounds
11.2.1. Production
11.2.2. Compounds of the Type R^1–Hg–R^2
11.2.3. Compounds of the Type R–Hg–X
11.2.4. Analysis
12. Economic Aspects
13. Toxicology and Occupational Health
13.1. Chemical Forms and Mechanism of Action
13.2. Exposure
13.3. Metabolism
13.3.1. Mercury Vapor and Inorganic Mercury
13.3.2. Organic Mercury Compounds
13.4. Toxic Effects
13.4.1. Mercury Vapor and Inorganic Mercury
13.4.1.1. Acute Toxicity
13.4.1.2. Chronic Toxicity
13.4.2. Organic Mercury Compounds
13.4.3. Dental Amalgam Fillings
13.4.4. Carcinogenicity
13.5. Indicator Media and Reference Values
13.5.1. Mercury Vapor and Inorganic Mercury
13.5.2. Methylmercury
13.6. Critical Levels, Risk Assessment

[Top of Page]

1. Introduction

Matthias Simon

Mercury [7439-97-6] (also called quicksilver), symbol Hg from the Greek *hydor* = water and *argyros* = silver, atomic number 80, A_r 200.59, was known as early as 1000 B.C. because of its liquid state at room temperature (melting point – 38.89 °C).

The discovery in 1938 of 1 kg of the metal in 2500-year-old sand layers on the eastern coast of Greece indicates that mercury was used in the extraction of gold at an early date [1]. Mercury was mentioned about 200 B.C. in India as well as in China (Han dynasty).

As early as 1556, GEORGES BAUER, known as AGRICOLA, reported five different methods for extracting mercury from its ores [2]. He also realized that mercury vapors are heavier than air and that they could therefore conveniently be trapped in condensers beneath the reaction vessel.

Because of the considerable vapor pressure of mercury even at room temperature and the toxicity of its vapors (see Chap. Toxicology and Occupational Health), particularly safe and reliable methods must be used in the extraction of mercury to avoid releasing even the slightest trace of the metal into the environment. The problem of gas and water purification is therefore particularly important (see Chap. Environmental Protection).

[Top of Page]

2. Properties

Matthias Simon and Peter Jönk

2.1. Physical Properties

Mercury is a silvery-white, shiny metal, which is liquid at room temperature. The most important physical properties are listed below:

Isotope masses (ordered according to decreasing abundance)	202, 200, 199, 201, 198, 204, 196
mp	– 38.89 °C
bp (101.3 kPa)	357.3 °C
Density (0 °C)	13.5956 g/cm^3
Specific heat capacity	

c_p (0 °C)	0.1397 J g^{-1} K^{-1}
Heat of fusion	11.807 J/g
Heat of evaporation (357.3 °C)	59.453 kJ/mol
Thermal conductivity (17 °C)	0.082 W cm^{-1} K^{-1}
Thermal expansion coefficient β (0 – 100 °C)	1.826×10^{-4} K^{-1}
Electrical conductivity (0 °C)	1.063×10^{-4} m Ω^{-1} mm^{-2}
Crystal structure	rhombohedral
Viscosity (0 °C)	1.685 mPa · s
Surface tension	480.3×10^{-5} N/cm
t_{crit}	1450 °C
p_{crit}	105.5 MPa
Critical density	5 g/cm^3
Evaporation number (25 °C)	0.085 mg K^{-1} cm^{-2}

Mercury has a relatively high vapor pressure, even at room temperature. Saturation vapor pressures at 0 – 100 °C are listed in Table 1 (corresponding to a specified mercury content in air). The temperature dependence of the density of mercury is given in Table 2.

Table 1. Saturation vapor pressure of mercury at different temperatures

t, °C	*p*, Pa	Hg content in air, g/m^3
0	0.026	0.00238
10	0.070	0.00604
20	0.170	0.01406
30	0.391	0.03144
100	36.841	2.40400

Table 2. Density of mercury as a function of temperature

t, °C	Density, g/cm^3	*t*, °C	Density, g/cm^3
– 38.89	13.6902	200	13.1139
0	13.5956	250	12.9957
50	13.4733	300	12.8778
100	13.3524	350	12.7640
150	13.2327		

Mercury vapor is excited to a state of luminescence by electrical discharge (mercury vapor lamps). Ultraviolet radiation is released primarily, it can be used to start and to promote chemical reactions.

The surface tension of mercury is ca. six times greater than that of water, which is the reason for its poor wettability by water.

The dynamic viscosity η of mercury (see Table 3) is of the same order of magnitude as that of water [3].

Table 3. Temperature dependence of the dynamic viscosity of mercury

t, °C	η, mPa · s	*t*, °C	η, mPa · s
– 20	1.855	60	1.367
– 10	1.764	80	1.298
0	1.685	100	1.240
10	1.615	200	1.052

20	1.554	300	0.950
30	1.499	340	0.921
40	1.450		

Some other metals, such as gold, silver, and zinc, dissolve readily in mercury to form amalgams. The solubility of mercury in water is strongly temperature dependent [4]. At room temperature, the solubility is ca. 60 µg/L; at 50 °C, ca. 250 µg/L; and at 90 °C, ca. 1100 µg/L. At low temperature, the addition of sodium chloride reduces the water solubility.

2.2. Chemical Properties

Mercury exists in the oxidation states 0, + 1, and + 2; monovalent mercury is found only in Hg–Hg bonds. The redox potentials at 298.15 K and 101.325 kPa relative to the standard hydrogen electrode are as follows [3]:

		E^0, V
$Hg^{2+} + 2\,e^-$	$\rightleftharpoons$ Hg	+ 0.851
$2\,Hg^{2+} + 2\,e^-$	$\rightleftharpoons$ Hg_2^{2+}	+ 0.920
$Hg_2^{2+} + 2\,e^-$	$\rightleftharpoons$ 2 Hg	+ 0.797
$Hg_2(CH_3COO)_2 + 2\,e^-$	$\rightleftharpoons$ $2\,Hg + 2\,(CH_3COO)^-$	+ 0.511

The standard potentials show that mercury is a relatively noble metal. Metallic mercury dissolves in nitric acid, aqua regia, warm concentrated hydrochloric acid and sulfuric acid. It is sparingly soluble in dilute HCl, HBr, and HI as well as in cold sulfuric acid. Most of its chemical compounds have densities of 5 – 9 g/cm^3. The oxide of mercury (HgO) decomposes at 400 – 500 °C. This effect is utilized in the extraction of mercury from oxidic sources. Mercury forms monovalent and divalent compounds with the halogens fluorine, chlorine, bromine, and iodine. It also forms monovalent and divalent compounds with sulfur.

[Top of Page]

3. Resources and Raw Materials

Matthias Simon

3.1. Deposits

All known mercury ores are relatively lowgrade ores, the average mercury content being ca. 1 %. Because mercury ores lie close to the earth's surface, the mining depth is ca. 800 m at most. The overwhelming proportion of mercury has always been produced in Europe; some of the most important deposits are listed below.Since the 1980s the former Soviet Union, Spain, China, Algeria, Turkey, and the United States have accounted for ca. 90 % of world mining production.

Spain. The deposits at Almadén have been mined for more than 2000 years. Almadén is situated in the southwest of the province of Ciudad Real in New Castile, about 200 km from Madrid, on the northern edge of the Sierra Morena. The ore-bearing deposits are porous sedimentary rocks (sandstone, bituminous shale, Silurian quartzite) that contain mercury sulfide (cinnabar). Needles of coarsely crystalline cinnabar as well as metallic mercury in lenticular ore-bearing bodies are also found over a roughly 20-km stretch in the valley of the Valdeazogues river. The ore is extracted in several mines (San Pedro y San Diego, San Francisco, San Nicolas). The richest strata contained 12 – 14 % mercury at a depth of 170 – 200 m. Mining is carried out today at a depth of 500 m; 1 t of ore yields about one flask of mercury (= 34.473 kg of mercury, corresponding to a 3.5 % mercury content in the ore).

This deposit was probably known to the Celts and Phoenicians, and was mined by the Romans from 150 B.C. These mines were subsequently worked by the Moors, then by the orders of knights, by the Spanish royal house, and from 1525 – 1645 by the Fuggers. Since then the mines have been under state management. From 1499 (when production records start) to the present, ca. 300 000 t of mercury has been mined.

In 1988 Almadén produced ca. 1380 t of mercury. The ore-bearing body has an average mercury content of 5 %. Because of the low price of mercury the plant operates for only a few months of the year. The operating company, Minas de Almadén y Arrayanes SA, is state-owned [5], [6].

Spain, which until 1989 was the world's largest producer of mercury, ceased nine production after 1990, although 100 t of mercury was produced in 1991 from stockpiled materials.China, Mexico, adn Russia were the largest producers in 1992.

Italy. A roughly 25-km-wide, 50-km-long belt running from Monte Amiata to the coast in southern Tuscany contains many closed mines as well as some that are still operating. Although these deposits had been mined by the Etruscans, they were not intensively worked by the Romans, to the benefit of the Spanish mines at Almadén. The Siele mine recommenced operation only in 1846, and the Abbadia San Salvatore mine, at present still the largest, in 1898.

Mercury occurs as cinnabar and metacinnabarite in pyrites, marcasite, antimonite, and realgar, mainly in Eocene sediments (shale, sandstone, marl, limestone) under a covering layer of trachyte. The gangue consists largely of clay or dolomite. The ores contain on average 0.2 – 0.8 % mercury. Mercury extraction is organized by the Monte Amiata works, its latent production capacity being up to 2000 t/a. No mercury has been mined since 1983.

Former Yugoslavia. The third largest mercury deposits in Europe are situated in Idrija in Slovenia, about 40 km west of Ljubljana. These deposits have been worked since 1497. The mines were under Austrian ownership from 1580 to 1918, then under Italian ownership, and finally reverted to former Yugoslavia after World War II.

The tectonics of this region are complicated. Shell marl and shale are impregnated with cinnabar and native mercury, and typically contain admixtures of pyrites, gypsum, and bitumen. The ore contains ca. 0.5 % mercury. Idrija has a capacity of about 600 t/a. About 30 t mercury was mined in 1992. Additional deposits in former Yugoslavia are located at Maškara in Croatia, Berg Avala to the south of Beograd in Serbia, Neumarkt in the Karavanken, and Montenegro.

Algeria. Algeria has become the second largest mercury producer in the western hemisphere after Spain, and in 1986 ca. 690 t of mercury was produced. In 1992 ca 425 t were produced.

Germany. In 1936 a modern mercury works was built at Landsberg near Obermoschel in the northern Pfalz (Palatinate), which had to be closed in 1942 because working the very low-grade ore, containing only 0.1 % Hg, was uneconomical.

Austria. Fahl ore containing 1.8 % mercury is mined in small amounts at Schwaz in the Tyrol. Other deposits at Dollach on the Drava are no longer mined.

Finland. Outokumpu obtains about 100 t of mercury per year from the processing of zinc concentrates.

Former Czechoslovakia. Three important deposits exist: at Kotrbachy, mercury is obtained as a byproduct of a fahl ore. Cinnabar is found southwest of Gelnica in a workable ore containing 0.25 % mercury. Finally, mercury has been obtained in the region between Mernik and Vranov since the end of the 17th century. This production ceased in 1937.

Rumania. Rumania has mercury ores at Zlatna and Baboja in Transylvania.

Turkey. In Turkey about 5 t of mercury is obtained annually from mining operations. At present, because of its low price, mercury is extracted from only two mines, which belong to the state-owned Etibank. The deposits are at Izmir-ödemi ş-Haliköy (1×10^6 t of ore containing 0.25 % mercury), Konya-Ladik, and Konya-Sizma (1.15×10^6 t of ore containing 0.23 % mercury). These deposits alone account for ca. 40 % of the Turkish reserves of mercury [7].

Former Soviet Union. The most important deposits are at Nikotovka in the Ukraine. Per year about 400 t of mercury is produced there from an antimonite ore containing 0.4 % mercury. Cinnabar deposits have recently been discovered in the Crimea, although no further details are available. Additional ore deposits are mined in the Northern Caucasus, Urals, Altai Mountains, and in Turkestan and Dagestan.

China. China has important deposits in the provinces of Yunnan, Hunan, and Kweitschan. The mines at Wanschantschang and Patschai are well known. Mined production was ca. 760 t of mercury in 1987 and 950 t in 1992.

Japan. In Japan, mercury is extracted on the island of Hokkaido.

Canada. Mercury ore deposits are situated in the west of Canada in British Columbia at Pinchi Lake and in the vicinity of Vancouver (Kamloops, Bridge River).

United States. The number of mines in the United States has decreased considerably in the last 20 years. Whereas 149 mines were operating in 1965, only 3 mines were still operating by 1981. The principal mining regions are California and Nevada. Placer, Inc., a subsidiary of the Canadian company Placer Development, accounts for ca. 99 % of production. The average mercury content in the ore is 0.3 – 0.5 %. Mining production in 1992 was ca. 64 t and falling.

Mexico. The 200 known mercury ore deposits are distributed over 20 states. The most important mines are in the provinces of Zacatecas, Guerero, Durango, Chihuahua, Guanajuato, San Luis Potosi Aguascalientes, and Queretaro; total mercury production is some 700 t/a.

South America. South America is of only minor importance as a mercury producer. Chile occupies first place among the mercury-producing countries. The Santa Barbara mine at Huancavelica in Peru is now exhausted, although it supplied substantial amounts of mercury in the 17th and 18th centuries : some 50 000 t was used for precious-metal extraction. Small amounts of mercury are mined in Venezuela and Bolivia.

A U.S. estimate made in 1983 gave the following amounts of mercury extracted by mining operations [8]:

Spain	50 000 t
People's Republic of China	17 200 t
Former Soviet Union	17 200 t
Former Yugoslavia	17 200 t
Italy	12 100 t
United States	12 100 t
Other countries	29 000 t
World total	154 800 t

Mining activities and exploitation of deposits are, as with most base metals, greatly dependent on current world market

prices. Too low returns have led in recent years to a number of mines being closed, which has resulted in a concentration of plants (less plants with increased production).

3.2. Secondary Sources

Like primary ores, industrial waste containing mercury also contributes to its production. The majority of plants used in chlor – alkali electrolysis employ liquid mercury cathodes, resulting in residues containing 10 % mercury or more. In addition to this major secondary source, mercury batteries, mercury fluorescent tubes, electrical switches, thermometer breakage, and obsolete rectifiers should be regarded as mercury sources.

Although the overall production of mercury has decreased since the 1970s, sufficient potential uses and thus secondary sources remain for the foreseeable future, thanks to the unique properties of this metal.

[Top of Page]

4. Production

Matthias Simon

4.1. Extraction from Primary Sources

4.1.1. *Dressing*

Preliminary concentration is desirable, especially for working low-grade ores. This is often performed in the mine by classification. Because of the brittleness of cinnabar, pieces of ore break easily at the mercury sulfide veins. Crystalline mercury sulfide in brittle ore can be concentrated by using settling tables.

Another method of concentration is the use of a flotation stage (⟶ Flotation) to increase the mercury content. Separation of the antimony and arsenic fractions is particularly important, and the ores are comminuted to grain sizes of ≤ 0.075 mm (200 mesh) in the case of antimony and ≤ 0.5 mm (25 mesh) in the case of arsenic. After comminution, lead(II) nitrate [$Pb(NO_3)_2$], butyl xanthate, and pine oil are added to the first flotation stage. Flotation is performed at pH 7.1. After purification, mercury is selectively floated, and a concentrate of ca. 50 % antimony is obtained. Potassium dichromate ($K_2Cr_2O_7$) is added as reactant. The end concentrate has a mercury content of ca. 70 % [9]. Another way of concentrating mercury by flotation is to add potassium oleate as collector [10].

Only a coarse preliminary grinding to ca. 50-mm grain size is necessary for high-concentration mercury sulfide ores. During subsequent thermal treatment, the ore particles burst because of the high vapor pressure, resulting in further comminution.

4.1.2. *Processing to Metallic Mercury [11], [12]*

The most important starting material for mercury extraction is mercury sulfide (cinnabar, cinnabarite) (see Section Mercury Chalconides); it is nonpoisonous and can be stored and transported without any problem.

The coarsely ground ore is processed in directly or indirectly heated furnaces, retorts, or muffles. Reaction with oxygen begins at 300 °C, according to

$$HgS + O_2 \rightleftharpoons SO_2 + Hg$$

Quicklime can be added as flux to bind the sulfur in solid form. The overall reaction in the *absence of oxygen* is as follows:

$$4\,HgS + 4\,CaO \rightleftharpoons 3\,CaS + CaSO_4 + 4\,Hg$$

In the *presence of oxygen,* the reaction is

$$HgS + CaO + 3/2\,O_2 \rightleftharpoons Hg + CaSO_4$$

Addition of metallic iron in the form of iron filings enables the following reaction to occur:

$$HgS + Fe \rightleftharpoons FeS + Hg$$

Instead of metallic iron, iron oxides can also be used:

$$7\,HgS + 2\,Fe_2O_3 \rightleftharpoons 4\,FeS + 3\,SO_2 + 7\,Hg$$

For this reason, sulfur-resistant steel should be used as construction material. Processing is normally performed with the addition of air, and the sulfur dioxide formed can be converted in a downstream wet scrubber.

4.1.3. *Furnace Systems*

Mercury sulfide-containing ores are processed pyrometallurgically. Distinction is made between *directly* and *indirectly* heated furnaces. This is important because the formation of *stupp* cannot be avoided in pyrometallurgical processing (the word stupp is an Austrian expression that simply means powder; in the case of mercury, a mercury-containing powder). The greater the formation of dust in the furnace, the greater is the occurrence of stupp. Because of the presence of combustion air in the furnace, directly heated furnaces generally create more dust than indirectly heated ones. Mercury in vapor form condenses to some extent on the small dust particles which act as condensation nuclei.

Another difference between directly and indirectly heated furnaces is that mercury vapor in indirectly heated systems is more

highly concentrated because it is not diluted by combustion gases.

Directly Heated Furnaces. Cylindrical Bustamente furnaces, multiple-hearth furnaces, and rotary kilns are used.

The *Bustamente furnace* is still used in Spain (Almadén) [5]. The cylindrical vessel is heated outside the shaft, only the combustion gases come into contact with the ore. A historical feature is that mercury vapors are condensed in air-cooled terracotta pipes.

The *multiple-hearth furnace* is advantageous for working up rich concentrates, mercury waste of a similar type, or stupp. This kind of furnace has been used for many decades in various mercury mines in North America, and in Almadén most of the ore extracted since 1961 has been processed in multiple-hearth furnaces. The hearths and outer walls of the furnaces are built of shaped fireclay bricks. The furnaces are lined with sheet steel. Cast steels with additions of titanium, chromium, and nickel have proved suitable for internal structural elements (stirrer shaft, raking arm). A modern furnace is up to 10 m high, has a diameter of ca. 5 m, and contains up to 16 hearths that are heated with wood, gas, or oil (mainly the upper and lower hearths). An attempt is made to keep the flow velocities for the reaction gas low (maximally 3 m/s) by skillful feeding of the material (if possible by special sliding surfaces at the throats in the case of small hearth interspacings) and by separately installed gas passageways. The amount of flue dust is thereby lowered (0.2 – 3 % of the feedstock), and the formation of stupp is minimized. Addition of combustion air through the insulated hollow shaft and raking arm and, if necessary, preheating of the fuel oil both reduce fuel consumption. By accurate metering of the inlet air, a lower amount of gas can be passed through the furnace, and the capacity of the subsequent condensation unit can be reduced. Table 4 shows some operational examples of multiple-hearth furnaces.

Table 4. Operational data for multiple-hearth furnaces for mercury extraction

	New Almadén	Pershing	Quicksilver Mines
Furnace diameter, m	4.9	4.9	5.5
Number of hearths	6	8	6
Worked material	ore	ore	ore
Throughput, t/d	38	77	50
Fuel consumption per tonne of material	47 L oil	21 L oil	0.15 m^3 wood

With regard to environmental protection, multiple-hearth furnaces are relatively difficult to make gastight and, in some cases, have had to be shut down.

Rotary kilns have been used since 1913 to extract mercury. Rotary kilns and rotary drum furnaces provide a high mercury yield with a good throughput rate of the ore because the reaction temperature can be controlled accurately; the ore does not have to be specially preheated; and material throughput can be regulated.

The ore is ground to about 65-mm particle size before it is charged. Lime, charcoal, or low-temperature coke may be added as fluxes, and the furnaces are heated directly by oil or gas. The furnaces are operated at 320 – 400 °C at the charging head and 700 – 800 °C at the discharge end. For normal furnace sizes and ores, heat consumption is 1×10^6 to 1.25×10^6 kJ per tonne of throughput. Without the addition of coal as flux, this corresponds to up to 30 L of oil or 250 m^3 (STP) of fuel gas per tonne of ore throughput. A disadvantage of rotary kilns is the dust formed by the intense material movement, which leads to increased stupp formation in subsequent condensation and can cause stoppage in the waste-gas line. Flue dust amounting to 0.75 – 6 % of the ore feedstock by weight is observed. In many cases, dust formation can be reduced by fluxes having a sintering effect. Some operating examples of older rotary kilns are listed in Table 5.

Table 5. Operational data for rotary kilns for mercury extraction

	Gelnica	Abbadia San Salvatore	New Indria	Nevada, Quicksilver Mines	Pershing, Quicksilver Mines	Mercury Mining Company
Furnace length, m	14	16	17	12.2	18.3	21
Internal diameter, m	1.5	1.25	1.22	0.92	1.22	1.22
Thickness for lining, mm	250	200	150			
Gradient, %	4	5	4			
Time for one revolution, s	72		42	45	50	
Moisture content of ore, %	5 – 10					

Hg content of ore, %	0.2	0.6 – 0.8				
Grain size of ore, mm	50	5		64	40	50
Ore throughput, t/d	40	100	125	40	45	45
Fuel consumption per tonne of ore	100 m^3 (STP) gas + 0.3 m^3 wood	100 kg wood	19 L oil	26.5 L oil	30 L oil	29 L oil

Indirectly Heated Furnaces. The basic advantage of indirect heating is the lower gas velocity in the reactor and the higher mercury concentration of the process gas. The furnaces are heated by gas, coal, or by electricity and consist in some cases of clay muffles, vessels lined with fireclay or ceramics, or iron vessels lined with silicon carbide.

F. Krupp – Grusonwerk AG has arranged several tubular individual muffles in a rotary kiln around a thermally insulated, centrally aligned shaft. A rotary kiln constructed from two concentric tubes was developed several years later. The ore passes continuously through the inner tube; the outer tube is heated. A highly thermally conducting, gastight silicon carbide is used as cladding.

Several proposals for processing mercury-containing material in indirectly heated furnaces have been made, and patents have been issued [13-15].

The subsequent condensation of mercury from the furnace off-gas is treated in Section Condensation of Mercury from Furnace Off-Gas.

4.2. Extraction from Secondary Sources

Because of the relatively high toxicity of metallic mercury and some of its compounds, spent products must be reprocessed to a large extent. Storage of slurry-like residues is complicated and expensive, and is governed by the limited storage capacity of the closed mines. Active carbon slurries from the effluent concentration resulting from chlor – alkali electrolysis constitute the major proportion of mercury residues that must be reprocessed.

In addition to these residues, the processing of fluorescent bulbs, which contain about 15 – 50 mg of metallic mercury per lamp, is becoming increasingly important. Several plants dealing specifically with the disposal of these lamps already exist in the Federal Republic of Germany [16].

The mercury content of normal zinc – manganese dioxide batteries has been reduced considerably over the last few years (addition of mercury inhibits the formation of gas by zinc). In the Federal Republic of Germany, the mercury content of batteries will be reduced to 0.15 % by 1990 (Fig. 1), starting from a level of 1 % [17].

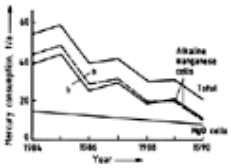

Figure 1. Reduction of mercury in primary batteries in the Federal Republic of Germany [17]

a) Mercury in sold primary batteries; b) Mercury in domestic refuse

In addition to zinc – manganese dioxide batteries, zinc – mercury(II) oxide batteries are also widely used as round cells. These contain ca. 30 % mercury, which must be recovered. Mercury in concentrations of ca. 1 % is also found in zinc – silver (II) oxide and zinc – oxygen batteries. A particular problem in reprocessing batteries is grading and sorting the various types, which are inadequately identified and, in some cases, have the same external dimensions for different electrochemical systems. Treatment processes have been developed in Japan [18], the Federal Republic of Germany [19], [20], and other countries [21], which can remove mercury by treatment in rotary kilns or in a distillation chamber. The individual types of batteries do not have to be sorted beforehand, and a mixture of batteries can be processed.

Because batteries also include a large number of plastics (e.g., as sealants), waste gases must be treated and burned if necessary.

Because a proportion of spent batteries end up in domestic refuse, high mercury concentrations can be found in waste gas from domestic refuse combustion. Mercury-containing older printing inks are another source. Because mercury can be precipitated from waste gas most efficiently by adsorption on activated carbon or similar substances, mercury-containing charcoal from waste gas purification will also require reprocessing in the future.

The mercury – selenium residue formed in the primary smelting of mercury-containing sulfidic ores requires special reprocessing. This substance, which occurs as a slurry, has been reprocessed in a multiple-hearth furnace in which the mercury fraction is evaporated [22]. Another company converts the residue to metallic mercury in a rotary kiln by adding lime fluxes, with a relatively inert residue remaining behind [23]. A hydrometallurgical reprocessing treatment has been investigated on a pilot scale in a research project [24]. Mercury is extracted in the form of mercury(II) oxide and sulfide, and can be added to the conventional extraction processes. The method does not appear to be economically practical at present

because of the relatively expensive reactants involved.

Additional secondary sources, such as thermometer breakage, electric switches, and amalgams, can generally be worked up by conventional distillation methods [25], [26]. A comprehensive monograph on the topic of evaporation and thermal dissociation of mercury sulfide is available [27]. Electrolysis has been proposed for the removal of mercury from gold-containing solutions: 90 – 95 % mercury and < 10 % gold are deposited at 1.0 – 1.5 V [28].

4.3. Condensation of Mercury from Furnace Off-Gas

The vapor pressure *p* (mbar of mercury) can be calculated according to a relationship given by BARIN and KNACKE [29]:

$$\log\ p = -\frac{3305}{T} - 0.795 \cdot \log\ T + 10.47983$$

where *T* is measured in kelvin.

The degree of saturation *s* (kg of mercury per cubic meter) of the gas at *t* (°C) or *T* (K) and *p* = 101.3 kPa can be determined from the pressure and density of mercury vapor (3.9091 kg/m^3 at 357 °C) (α = 1/273):

$$s = 3.9091 \frac{1 + 357 \cdot \alpha}{1 + \alpha t} \cdot \frac{p}{101.3} = 3.9091 \cdot \frac{630}{T} \cdot \frac{p}{101.3}$$

The degree of saturation *s* of mercury vapor calculated according to this formula for different temperatures is summarized in Table 6. Figure 2 shows *p* and *s* plotted as a function of 1/*T* on a semilogarithmic scale.

Table 6. Saturation content of mercury vapor at various temperatures

p, kPa	*T*, K	Saturation content *s*,	
		g/m^3	g/m^3(STP)
7.03×10^{-5}	283	0.00604	0.00626
1.71×10^{-4}	293	0.01406	0.01508
3.919×10^{-4}	303	0.03144	0.03489
8.514×10^{-4}	313	0.06612	0.07588
1.761×10^{-3}	323	0.1325	0.1568
3.486×10^{-3}	333	0.2545	0.3104
6.625×10^{-3}	343	0.4695	0.5893
0.012138	353	0.8359	1.081
0.02150	363	1.440	1.915
0.03690	373	2.404	3.284
0.06154	383	3.906	5.478
0.100	393	6.187	8.906
0.1585	403	9.558	14.110
0.2457	413	14.464	21.385
0.3727	423	21.417	33.175
2.2894	473	117.65	203.65
9.8743	523	458.97	879.25
32.808	573	1391.5	2921.0
101.3	630	3909.1	9021.1

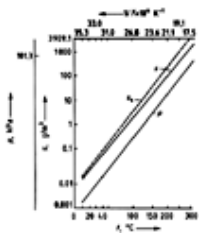

Figure 2. Vapor pressure *p* and saturation content *s* (g/m^3) as a function of temperature; s_s saturation content under standard conditions [g/m^3 (STP)]

Mercury losses during condensation can be calculated as follows: 1 m^3 of saturated vapor at 140 °C (413 K) enters the condensation unit laden with 14.464 g of mercury. There it is cooled to 20 °C (293 K) (*s* = 0.0141 g/m^3), and a volume contraction to 293/413 = 0.71 m^3 occurs. The saturated vapor leaving the unit therefore contains 0.71×0.0141 = 0.01 g of mercury. The loss of uncondensed mercury removed with the waste gas is thus 0.07 % of the initial amount. With a

condensation end temperature of 40 °C and otherwise identical conditions, the loss increases to 0.35 % Hg.

In practical operation the mercury content in the hot reaction waste gases is always far below the saturation limit because the theoretical vapor – liquid equilibrium is not reached. Directly heated mechanical roasting furnaces require a heat input of ca. 1.2×10^6 kJ per tonne of ore. The following amounts of waste gas [in m^3 (STP)] are produced per 1000 kJ for the different fuels with 20 % excess of air:

Wood	0.387
Generator gas	0.42
Oil	0.32

Per tonne of ore, this corresponds to about 470 m^3 (STP) with wood or generator gas firing and 375 m^3 (STP) with oil firing. With 5 % water content in the ore, 70 m^3 (STP) of steam is produced in addition. If an ore containing 0.3 % Hg is worked up, the mercury content with wood or gas firing is 5.56 g/m^3 (STP), and with oil firing 6.75 g/m^3 (STP), at 140 °C.

With a condensation end temperature of 20 °C—at which the gas still contains 0.0141 g of mercury per cubic meter corresponding to the saturation limit—a loss of 0.21 – 0.25 % mercury occurs, and at a condensation end temperature of 40 °C, a loss of 1.1 – 1.3 % mercury from the waste gas occurs.

In general, the mercury residual content in waste gas, compared to the amount at 20 °C, increases with a temperature increase of

10 °C by a factor of 2.2
20 °C by a factor of 4.7
30 °C by a factor of 9.4
40 °C by a factor of 18.1
50 °C by a factor of 33.4

Thus efficient cooling of the waste gas is necessary. In a plant that processes an ore with 5 % moisture content, as much as 50 % of the cooler capacity is used for condensation of evaporated moisture alone, without the steam originating from fuel combustion being taken into account.

Several practical requirements must be met for satisfactory operation: (1) The ore should be as dry as possible. (2) The amount of waste gas produced by the fuel should be low. (3) When the gas enters the condenser, its temperature should be only 10 – 20 °C above the mercury dew point. (4) The outlet temperature should be as low as possible. (5) Before entering the stack, the reaction gas should pass through an afterpurification section, if possible. Stupp formation can be kept low by (1) indirect heating, (2) separate processing of fine-grain classes of ore, (3) predrying of ore that is too wet, or (4) wetting of ore that is too dry. The flue dust should be separated as completely as possible from the still hot reaction gas before it enters the mercury condensation unit.

A conveniently arranged condensation system for a *directly heated* mercury smelting plant operating with mechanical roasting furnaces consists of the following parts:

1. Dust separator in the form of a cyclone, an electrostatic precipitator, or a combination of both
2. Fan (e.g., of Monel Metal)
3. Cooling unit for mercury condensation
4. Afterpurification chamber for separating mercury residues
5. Injection chamber for separating traces of mercury
6. Fan (e.g., of Monel Metal)

The cooling unit is itself in many cases also constructed of monel metal, and consists of tubular condensers (CERMAK type) formed from six to ten trains arranged in parallel, each with eight inverted U-tubes of elliptical cross section whose arms terminate in acutely tapering, water-filled deflection boxes. The tubes are sprinkled with water. The mutually corresponding tubes of all the individual trains in each case end in a common box, and in this way the stupp can be classified. Gate valves in each train enable the draft to be regulated or even permit the train to be separately disengaged.

The ratio of tube cross section to fan output must be matched so that the gas flow velocity is maximally 0.75 m/s in narrow parts of the tubular condenser. The overall unit must be designed so that half of the tube trains can be decommissioned for cleaning and repair, whereas the remaining tube trains ensure satisfactory cooling of the reaction gas while maintaining draft conditions in the furnace.

Gases from *indirectly heated furnaces* that contain only a small amount of flue dust can be condensed by spraying water directly into the mercury-containing waste-gas stream. To ensure sufficient mercury condensation, the gas must pass through a succession of water curtains. Fewer parallel tubes of larger diameter are used in the cooling unit for directly heated furnaces. The last stage operates with fresh water; for the other cooling stages, water is circulated and recycled through suitably dimensioned settling tanks.

All settling tanks within the condensation unit have sloping floors and connecting pipes through which mercury metal flows to the lowest part of the system. The tanks stand on feet on a smooth cement floor, which slopes so that any leakage can be detected and escaping mercury collected. The material for these tanks is wood or cast iron; stupp-collecting tanks are also

constructed of concrete. When the plant is operating continuously, all places prone to thick deposits must be mechanically cleaned once a week; the remaining areas can be cleaned at longer intervals.

4.4. Treatment of the Stupp

The amount of mercury contained in the stupp may represent a substantial proportion of production. In Idrija, when processing high-bituminous ore, for example, 78 – 91 % of the mercury is extracted from the stupp and only 9 –22 % is extracted directly. To extract mercury from the stupp, individual mercury droplets must be coagulated, which can generally be achieved by kneading and pressing the mass. Stirring the stupp with sievelike perforated rakes in iron vessels also gives good results. Addition of quicklime to the stupp neutralizes any acids contained in the condensate, saponifies any fats present, absorbs a large proportion of the water, and chemically reduces mercury sulfate.

Mercury and dust can also be mechanically separated by simple wash treatment with water. A tenfold mercury enrichment from the stupp by *flotation* was described in 1929 [30]. This very effective method of separating mercury enables the waste material to be dumped directly onto the waste tip in certain cases. Generally, however, extraction of mercury from the stupp is not so effective, with the result that the low-mercury fraction must be returned to the smelting process. For regularly occurring large amounts of stupp, a special furnace unit is worthwhile (retorts or multiple-hearth furnace).

[Top of Page]

5. Environmental Protection

Matthias Simon

5.1. Natural Distribution of Mercury

Because of its high vapor pressure, metallic mercury disperses relatively quickly into the atmosphere and, with suitable air movement, is taken up by plants and animals. The average concentration in the earth's crust is 0.08 ppm and in seawater 3×10^{-5} ppm. Mercury is thus one of earth's relatively rare elements. The natural mercury content of the atmosphere is 0.005 – 0.06 ng/m^3 [31] ; in plants, 0.001 – 0.3 µg/g (generally < 0.01 µg/g) ; and in meat, 0.001 – 0.05 µg/g [32]. A comprehensive list of mercury in the environment appears in [33].

A quantitative summary of the occurrence, distribution, and utilization of mercury is given in the following material [34]:

Mercury content of the oceans	70 000 000 t
Mercury content of the earth's crust (1-m-thick layer)	100 000 000 t
Natural atmospheric mercury emission (volcanoes, wind erosion, degassing)	25 – 100 000 t/a
Natural mercury emission in water (weathering in rivers)	5000 t/a
Use of fossil fuels (minerals, deposits)	8 – 10 000 t/a
Use of mercury and its compounds	6000 t/a

Natural air emissions from mechanical activity, wind erosion, and degassing constitute the largest proportion of emitted mercury. By contrast, the utilization of mercury is relatively small, although this should not minimize or obscure the danger of mercury at high concentration.

5.1.1. *Mercury in Soil, Plants, and Animals*

Relatively large amounts of mercury are circulated due to the constant exchange of mercury among water, soil, and the atmosphere. Some of this migrating mercury is retained in the soil in the form of humus compounds and, in certain cases, is also concentrated [35]. The mercury content of various soils in Austria unless otherwise stated (mg per kilogram of soil) is as follows [36], [37]:

Meadowland in Niederösterreich	0.039
Arable land in Niederösterreich	0.180
Vineyards in the Burgenland	0.080
Reed belt on the Neusiedlersee	0.340
Arable land (3 % humus) in Kremsmünster	0.065
Meadowland (4.8 % humus) in Kremsmünster	0.070
Woodland (10.2 % humus) in Kremsmünster	0.170
Garden soil in England	0.25 – 15.0

Rice field slurry in Japan	0.40 – 1.8
Fields for agricultural use in Niedersachsen, FRG	0.055 – 0.104
Woodland in Niedersachsen, FRG	0.249 – 1.672
Mean mercury value for Europe (10-cm-thick humus layer)	120 g/ha

Organically bound mercury is also added to the soil with seeds as a seed treatment agent (fungicide). Amounts of mercury used for seed treatment or soil disinfection are as follows:

Grain (wheat) added as seed treatment agent to the soil per 150 kg of seed material/ha	ca. 7 g/ha
Sugar beet (seed material in pill form) added as seed treatment agent to the soil per 200 000 pills/ha	ca. 1 g/ha
Cotton (seed treatment and disinfection)	ca. 20 g/ha
Lawns (disinfection)	ca. 80 g/ha

However, neither soils that have received mercury-treated seeds for many years, nor plants grown on such soils, exhibit a high mercury content [35]. As pot tests have shown, the mercury is partly retained in the humus and partly evaporates from the soil [38], [39]. The amount of residue passing into groundwater is insignificant [40] (see Fig. 3).

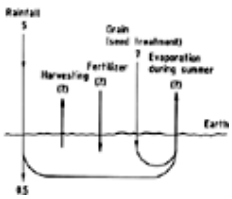

Figure 3. Microcirculation of mercury (in g ha^{-1} a^{-1}) in the soil [36]

According to investigations carried out in Norway, the mercury content of soil varies between 0.022 and 0.55 ppm (mean value 0.19), without any “recognizable” effect due to human activity [41]. The mercury content in soil is higher in geothermal or ore-containing regions and can be as high as several parts per million. Thus, mercury is often used as a tracer metal in geological exploration [42].

Mercury content has been investigated in soil and lichen in the mercury mining region around Monte Amiata (Italy). The mercury content in the soil and in the lichen decreases sharply with increasing distance from the mine. Mine waste gases and worked ore are the main sources of emission [43].

Mercury in Fungi and Mushrooms. The fleshy parts of fungi and mushrooms can contain unexpectedly high concentrations of mercury [44], [45]. Because of the affinity of mercury for the sulfur-containing constitutents of fungi, the latter can absorb and concentrate mercury from the local soil. Mercury-rich fungi and mushrooms are found everywhere, regardless of whether the soil has been anthropogenically contaminated [46]. With more serious contamination, especially atmospheric, the mercury content of fungi and mushrooms increases considerably [47].

Mercury in Fish and Seals. The mercury content in *fish* varies depending on type and habitat [48]. The assumption that fish concentrate mercury to 2000 times its concentration in the ambient medium is neither valid nor statistically relevant and has no significance as a rule of thumb [48]. Mercury uptake in warm tropical waters is higher than in cold northern waters [49]. Because of the concentration of mercury in sediment, dabbling fish, which collect their food from the bottom of the water, absorb more mercury than nondabbling fish. Predatory fish (e.g., pike) contain higher than normal levels of mercury only if they preferentially feed on dabbling fish (see also Section Mercury in Food).

The controversy surrounding canned tuna in the United States has revealed that tuna preparations in the Smithonian Institute from ca. 1880 have the same mercury concentration as present-day fish. Under the assumption that the values found in museum preparations constitute a lower limit—some of the mercury might well have evaporated over the course of time—the fact that present-day values are the same as previous values means that no concentration has occurred.

Data from the Danube upstream of large cities and industrial regions, where trout have been found with a mercury content of > 1 mg of mercury per kilogram of fish, point to factors other than industrialization. The high mercury concentration in the liver of seals in the northern section of the east coast of North America (Labrador), where no large cities or industrial regions are to be found, is attributed to underwater volcanic emissions.

5.1.2. *Mercury in Food*

Many investigations were carried out on foods from the 1920s and 1930s onward [50]. In the Federal Republic of Germany, the Zentrale Erfassungs- und Bewertungsstelle für Umweltchemikalien (central office for collecting and evaluating data on environmental chemicals; ZEBS) of the Federal Ministry of Health in Berlin in particular, has collected data from 1975, 1979, and 1982. According to these data, if average consumption patterns are assumed, mercury uptake in 1979 was about 0.052 mg per week for the “average” inhabitant of the Federal Republic of Germany (i.e., only about one-sixth of the upper limit recommended by WHO). The major proportion of mercury is absorbed from animal foods [51]. The legal limits are given in Table 7.

Table 7. Legal limiting values for total uptake of mercury (1980)

	Fish, mg/kg	Total uptake, mg/week
Switzerland	0.5	
Japan	0.4 (Hg)	0.17 mg of Hg per 50 kg of body weight
	0.3 (Hg methyl)	
United States	1.0	0.005 mg/kg of body weight (Hg)
		0.0033 mg/kg of body weight (Hg methyl)
Sweden	no limit (lakes that contain fish with more than 1 ppm Hg are placed on a blacklist) ; fish with a high Hg content may be legally rejected.	
Federal Republic of Germany	1.0	0.3 (max.)
WHO recommendation		

The ZEBS investigations [52] covering 1978 – 1982 are presented in Tables 8 and 9. The 1982 values essentially confirm the analytical data of 1979. The values for milk, condensed milk, cheese, and eggs; veal, beef, pork, calves liver, calves kidneys, poultry, sausages, and meat products; vegetable oils and fats; rice, rye, and potatoes; leafy, sprouting, fruit, or root vegetables and canned vegetables; pomaceous, stone, berry, citrus, shell, or canned fruit; fruit and vegetable juices; and beverages such as wine, beer, and chocolate are all < 0.020 mg of mercury per kilogram of fresh substance and, in most cases, < 0.010 mg/kg.

Table 8. Mercury in food

Food	Content, mg/kg of fresh substance	
	1982	1979
Cow's liver	0.021	0.015
Pig's liver	0.047	0.058
Cow's kidneys	0.077	0.066
Pig's kidneys	0.246	0.260
Freshwater fish	0.271	0.257
Saltwater fish, except Hg-susceptible fish	0.196	0.127
Hg-susceptible fish*	1.070	0.859
Fish products	0.208	
Canned fish	0.206	0.188
Wheat	0.026	0.003
Coffee	0.041	

* Hg-susceptible fish include: mackerel shark, dogfish, blue ling, halibut, black halibut, turbot, and Greenland shark.

Table 9. Uptake amounts of mercury in absolute figures and as a proportion of WHO values

	Hg
WHO value	
mg/week (70 kg)	0.35
mg/week (58 kg)	(0.290)
Men (70 kg)	
mg/week*	0.1229
Proportion of WHO figure, %	35.11
Women (58 kg)	
mg/week*	0.0933
Proportion of WHO figure, %	32.17

* The median values of toxic substance levels were taken into account.

The average amounts of mercury actually absorbed per week are compared in Table 9 with the limiting values specified by WHO. Foods contain the major proportion as follows: drinking water (14.8 %), wheat (12.9 %), milk (10.2 %), mineral water (8.1 %), canned fish (8.0 %), coffee (7.2 %), potatoes (6.0 %), and fish (5.6 %). Another detailed investigation carried out by ZEBS is concerned with the mercury, magnesium, and zinc content of breast milk, blood serum, and the fatty tissue of nursing mothers [53].

5.2. Mercury Emissions

5.2.1. *Gas Purification*

In a large number of combustion processes, mercury is transferred to the gaseous phase and must be removed again by subsequent treatment. The largest amounts of secondary mercury occur in the smelting of sulfidic ores. Mercury which is present as mercury sulfide is released during roasting. At present, two methods are primarily used for waste-gas purification.

In the first method, the gas is treated with sulfuric acid (90 %) at 200 °C. The mercury(I) sulfate formed is deposited in wash towers. Fine purification is performed by afterpurification with sodium sulfide [54]. In the second method, the cooled and dusted roast gas is treated with a mercury(II) chloride solution, and mercury precipitates as mercury(I) chloride (Hg_2Cl_2). Mercury is removed from the wash liquid by treatment with sodium sulfide. Part of the mercury(I) chloride produced is oxidized to mercury(II) chloride with gaseous chlorine and returned to the process. The final level of mercury in the waste gas is 0.05 – 0.1 mg/m^3 [55], [56].

In a third method, waste gas containing mercury and sulfur dioxide is treated with a wash solution containing Cu^{2+} and Hg^{2+} ions in addition to H_2SO_4 (200 – 300 g/L) and HCl (5 g/L). Mercury in the vapor is thereby oxidized to Hg^+, and the Hg^0 values of the waste gas are 0.02 –0.05 mg/m^3 [57].

Another source of mercury emission is refuse combustion. Mercury occurs in the form of thermometer breakage, fluorescent tubes, switches, and batteries in domestic refuse. Investigations have shown that mercury in flue gases is present as mercury(II) chloride [58]. Below ca. 150 °C, the flue ash contained in flue gases adsorbs mercury(II) chloride and reduces it to mercury(I) chloride. Active carbon or active carbon impregnated with iodine compounds can also be used as a mercury adsorbent [59-62]. The mercury content of domestic refuse is 3 – 5 g/t [63]. The percentage distribution of mercury in a typical refuse combustion plant is shown schematically in Figure 4. Most of the mercury passes into the residue to be dumped.

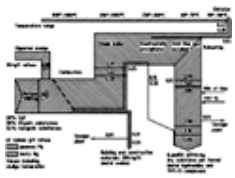

Figure 4. Specific mercury load distribution in the Bamberg refuse heating and power station with TMT-15 precipitation, with and without sludge combustion [63]

5.2.2. *Water Purification*

The solubility of mercury in water depends strongly on the temperature (see Section Physical Properties). It decreases with decreasing temperature and can be reduced further by addition of salts (e.g., sodium chloride).

A whole range of equipment and devices can be used in purification methods to remove mercury from water. Examples are activated carbon filters, ion exchangers, and electrolysis systems. Mercury can also be removed from solution by addition of suitable reagents.

The level of 10 – 30 mg of mercury per liter in wastewater from chlor – alkali electrolysis can be reduced to 0.1 mg/L by using *activated carbon* and graphite powder. The carbon has a grain size of 5 – 100 µm [64], [65]. In principle, all types of activated carbon can be used for the fine purification of mercury-containing water. Carbon is dispersed in the water and then filtered.

For *ion exchangers,* a distinction can be made between reusable and disposable ones [65-67]. In some cases, sulfur-containing substances are used as active groups, which remove mercury from the solution in the form of a mercury sulfide compound.

Copper fluidized-bed electrolysis provides another method for mercury separation. The copper powder cathode is held in suspension by the flowing solution, and mercury is deposited as an amalgam. The copper amalgam is purified by distillation [68].

By adding *small amounts of oil* which are dispersed in the aqueous solution, mercury can be concentrated in the oil phase. The mercury content of the purified aqueous solution is in some cases < 10 µg/L. The oil is separated by centrifugation, and fine purification is carried out by using a conventional oil – water emulsion splitting unit. The initial mercury concentration should be between 0.5 and 2.0 mg/L [4]. Tin(II) chloride dissolved in dilute hydrochloric acid has been proposed for treating mercury ion-containing wash water from combustion units. The reducing agent tin(II) chloride is added in above-stoichiometric quantities. Reduced mercury is added together with the wash water to an evaporation device and is expelled by passage of stripping gas at elevated temperature. Condensation is effected by cooling the gas stream. The mercury content is thus reduced from 5 mg/L to < 0.1 mg/L, depending on the concentration of tin(II) chloride [69].

To purify flue gas wash water from *refuse combined heating and power stations,* the additive TMT-15 developed by Degussa is used in some cases on an industrial scale (see Fig. 4). The active substance of the additive consists of trimercapto-*s*-triazine in the form of a sodium salt. Mercury is bound as $C_6N_6S_6Hg_3$. The compound is stable up to 210 °C and is only sparingly soluble in the elution test (a test to determine the soluble components of a solid material). The mercury content is reduced from ca. 4 mg/L to < 0.05 mg/L [70], [71]. The additional costs involved in TMT-15 precipitation amount to ca. 0.25 DM per tonne of refuse.

Mercury can be removed from *concentrated sodium hydroxide solution* by the use of ultrasound when the solution is filtered. The initial mercury content of ca. 20 mg/kg of aqueous sodium hydroxide is reduced to ca. 0.4 mg/kg after filtration [72].

The fine purification of *crude phosphoric acid* can be performed at pH 0.5 – 1.5 by using a diorganodithiophosphorus compound in conjunction with activated carbon as adsorbent. The final mercury concentration is < 0.02 µg/g of solution [73]. At present, no standard methods exist for removing mercury from sulfuric acid [74].

[Top of Page]

6. Quality Specifications

Matthias Simon

Fine Purification of Metallic Mercury [75]. Most of the mercury on the market is 4N material (99.99 % mercury). Higher purity is seldom required. Impurities in the form of gold (> 10 ppm) are manifested as dark particles after dissolution in nitric acid. Only a few purification methods exist.

1. *Dry Oxidation.* In this method, readily oxidizable constituents such as magnesium, zinc, copper, aluminum, calcium, silicon, and sodium can be removed by passing air or oxygen through the liquid metal. The oxides formed have a lower density than mercury and float on its surface. They can be removed by filtration, scooping, or withdrawing the mercury through an opening in the bottom.
2. *Wet Oxidation.* In an aqueous medium, mercury is dissolved by adding nitric, hydrochloric, or sulfuric acid with dichromate, permanganate, or peroxide, to oxidize impurities. Good dispersion of the mercury is extremely important in this method. The aqueous solution can be separated from the mercury by decanting, and traces of water can be removed with calcium oxide. A plant for the wet purification of mercury has been described [76].
3. *Electrolytic Refining.* Perchloric acid containing mercury oxide serves as the electrolyte.
4. *Distillation.* Mercury can be evaporated under atmospheric pressure or in vacuo. Distillation can be carried out in normal steel vessels or in a glass apparatus. Elements with a lower vapor pressure than mercury can be separated in this way. In many cases, mercury must be distilled repeatedly to achieve the desired purity, particularly if it is to be used to produce cadmium – mercury telluride.

Additional methods, adapted to the relevant processes, are available for purifying and working up larger amounts of mercury [77], [78].

[Top of Page]

7. Chemical Analysis

Matthias Simon

The oldest and simplest method for determining mercury in minerals is described by ESCHKA ; it involves a gravimetric method in which mercury is precipitated as an amalgam [79]. The ore sample is weighed in a porcelain crucible, intimately mixed with iron filings, and then covered with a layer of zinc oxide, magnesium oxide, or calcium oxide. The crucible is closed with a tightly fitting cover of gold foil having a cup-shaped depression in the middle, which is sprayed from above with cooling water. On careful heating of the sample, mercury is distilled and deposits on the underside of the cover. When a constant weight is reached at the gold foil, all the mercury has been collected. Measurements can be affected by cadmium and arsenic, which also condense on the gold foil. This method of mercury determination requires a high level of experimental skill and care on the part of the analyst, whose technique greatly affects the accuracy and reproducibility of the results. The reliably detectable minimum mercury content is between 10 and 50 ppm. Theoretically, the accuracy could be improved by increasing the amount of sample, although this would give rise to difficulties with regard to the apparatus.

A method for precipitating *monovalent* mercury as an iodate from a neutral or weakly acidic solution (nitric acid) is described in [80]. Mercury can be determined gravimetrically after the mercury(I) iodate precipitate is washed with ethanol and diethyl ether. Parallel gravimetric determination of the excess of potassium iodate precipitation agent is possible with thiosulfate after the mother liquor has been acidified with sulfuric acid and potassium iodide has been added. Both methods give good results.

A quick method for determining *mercury*(*II*) is based on the fact that a complex is formed when an excess of potassium iodide is added to a neutral or weakly ammoniacal mercury solution. This complexes with copper diethylenediamine sulfate to form violet crystals of the complex double salt copper diethylenediamine mercury iodide, which is practically insoluble in ethyl alcohol or diethyl ether [81].

In a rapid classical qualitative method for detecting mercury, ca. 1 g of the substance to be tested is digested with acid; the resultant solution is oxidized with a drop of bromine (the solution must not turn yellow) and then boiled with a few milliliters of the reagent (10 g of KI and 100 g of NaOH in 100 mL of H_2O) and filtered. In the filtrate, mercury is determined by the black precipitate formed on dropwise addition of a Sn^{2+} solution [82].

Modern operational monitoring employs physical analytical methods and test tube methods suitable for quick detection. Both *X-ray fluorescence spectroscopy* (XRF) and *atomic absorption spectroscopy* (AAS) have proved suitable for quantitative and qualitative mercury determination. The detection limit for these methods is so low that the maximum workplace concentration values can be monitored precisely [83].

Portable atomic absorption spectrometers, for example, are available, which indicate the atmospheric mercury concentration as an analog or digital display after a short warm-up time of the spectrometer. The result of the analysis is available immediately and simplifies the monitoring of a production plant. The measurement value can also be recorded continuously at stationary measurement sites with a recorder. The influence of interfering factors in AAS determination is discussed in [84].

Differential pulse anodic stripping voltammetry (DPASV) in conjunction with a gold electrode can be used to detect copper and mercury in natural water and wine [85], [86]. A mercury concentration of 0.02 µg/L can be measured.

Mercury in the air can be detected down to 0.1 ng/m^3 with the *Coleman mercury analyzer* system (based on a very selective cold vapor atomic absorption), to an accuracy of 10 %. One analysis takes about 3 min [83].

[Top of Page]

8. Storage and Transportation

Matthias Simon

A classification for the transportation of mercury, mercury oxide, mercury(I) chloride, and mercury(II) chloride is given in [87] (specification sheets 831, 863, 865, 868). DORIAS presents a detailed description of the properties, handling, storage, and transportation of dangerous substances [88], as well as a list of addresses of the relevant authorities (throughout Europe) for information purposes. The provisions differ from country to country and must be ascertained from the relevant authorities.

In general, containers of stainless steel, normal-quality steel, iron, glass, ceramics, and a range of plastics are suitable for storing mercury. When storing liquids that contain extremely low levels of mercury (µg/g – ng/g range) in plastic containers, mercury losses occur with a large number of plastics. When water containing 5 ng/g of mercury was stored in polyethylene bottles, only 5 % of the mercury was present after 21 d. Approximately 77 % of the mercury had been adsorbed on the side walls, and 18 % had evaporated. The Hg^{2+} ions are assumed to be reduced to Hg^{+} ions, which in turn disproportionate to Hg^0 and Hg^{2+} [89]. Addition of Au^{3+} ions in trace amounts is sometimes recommended [90]. Further investigations on the storage of mercury in dilute solution are discussed in [91-95].

[Top of Page]

9. Uses

Matthias Simon

Because of its special properties, mercury has had a number of uses for a long time. The conventional application is the thermometer. Mercury is frequently used in pressure gauges and for thermal content measurements.

In mechanical engineering, mercury is used in mercury vapor diffusion pumps for producing a high vacuum.

One important area of use is the lighting industry, where mercury is added to various types of bulbs. Among electrical components, mercury switches, rectifiers, oscillators, and primary batteries can be mentioned, which contain up to 30 % mercury.

Mercury is used as a liquid cathode in the production of chlorine and sodium hydroxide by chlor – alkali electrolysis. Because mercury forms alloys with a large number of metals, mercury alloys have a wide range of applications. Table 10 summarizes the areas of use of mercury and its compounds.

Table 10. Uses of mercury and its compounds [96]

Area of use	Form	Approximate	Emissions	Other	Remarks

		amount, t/a		sectors, waste elimination	
Chemicals, reagents	compound	40	laboratories		largely impossible to find suitable replacement; environmental damage can be reduced by other measures
Alkali – manganese batteries	chloride/metal	37		refuse combustion	reduced emission by separate collection possible
Mercury oxide batteries	oxide	20		refuse combustion	alternatives still too expensive; separate collection declining
Pesticides		27	distribution		now only for seed treatment; substitutes are available
Medicinal sector		26		refuse combustion	separate collection possible
Dental amalgams	metal				
Disinfection	compound				
Fungicides for paints	compound	10	weathering	exploitation; removal of old paint coats	declining; in some cases restrictions on use
Catalysts	compound	8		chemical industry	declining
Thermometers, barometers, manometers	metal	8	instrument and apparatus breakage	refuse combustion	replacement by other methods and processes possible (e.g., electronics)
Electrical engineering components	metal	6		refuse combustion	declining; substitution possible largely through electronic components
Interval switches				exploitation of scrap	
Fluorescent tubes	metal	3	lamp breakage	refuse combustion	demand for fluorescent tubes increasing, despite reduction in mercury use per tube; total use increasing
Pigments	sulfide	1		refuse combustion	declining

[Top of Page]

10. Mercury Alloys

Matthias Simon

10.1. General

The alloys of mercury (amalgams) occupy a special position among metal alloys because they can be solid, plastic, or liquid at room temperature.

Liquid amalgams are true solutions of the alloying elements in mercury, whereas *plastic* amalgams are suspensions of solid particles of the alloying partners in mercury or a saturated mercury solution. *Solid* amalgams are intermediary phases, often mixed with alloying partners or their primary mixed crystals. Solid amalgams may contain liquid-phase inclusions. An amalgam is sometimes difficult to identify experimentally. JANGG has suggested a suitable apparatus for synthesizing and

analyzing amalgams [97].

The solubility of many metals in mercury depends strongly on temperature (see Fig. 5). The solubility of some metals in mercury at room temperature is given in Table 11. Amalgam formation may be exothermic (e.g., sodium) or endothermic (e.g., gold).

Table 11. Solubility of some metals in mercury at 20 °C [97]

Metal	Solubility, %	Metal	Solubility, %
Tl	42.5	Mg	ca. 0.3
Cd	5.0	Au	0.131
Zn	1.99	Ag	0.035
Pb	1.47	Cu	ca. 0.002
Sn	ca. 0.9	Al	0.002
Na	0.62	Fe	ca. 10^{-5}
K	ca. 0.4	Si	virtually insoluble

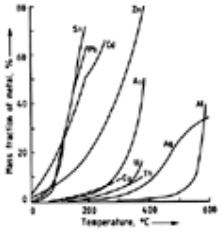

Figure 5. Solubility of some metals in mercury as a function of temperature [99]

Ammonium amalgam, which has been known for a long time, is an interesting case of the NH_4 group acting as a metal-like alloying constituent. The synthesis of tetramethylammonium amalgam is described in [98].

Technically important amalgams are those of tin – copper – precious metals used in conservative dental treatment.

Gold, silver, or tin amalgams are still used for much gold- and silver-plating work, as well as for production of certain types of mirrors. The measurement range of mercury thermometers can be extended to – 58 °C by addition of thallium to mercury.

Alkali amalgam is an important intermediate in chlor – alkali electrolysis by the amalgam method. Amalgams also play a role in the extraction of cadmium or aluminum [99].

A comprehensive list of thermodynamic data on amalgam formation has been collected by GUMINSKI [100]. After surveying the literature, he is of the opinion that liquid mercury greatly influences intermetallic compound formation. The reactions are comparable to the solid-state formation of ionic and nonionic substances in liquids.

10.2. Production of Amalgams

Three methods are used for the industrial production of amalgams:

1. *Powder Metallurgy Method.* The powdered alloying components are mixed with mercury. The reaction rate of spontaneously occurring amalgam formation is determined by the degree of dispersion of the powder, the rate at which reactants diffuse into one another, and the wettability of the powder by liquid mercury. If other components besides mercury are added simultaneously, they are preferably pulverized as master alloy in the desired weight ratio. In many cases, the powder reacts with mercury more quickly in the presence of a salt of the element to be amalgamated.
2. *Galvanic Amalgam Formation.* Many metals can be deposited from their aqueous solutions or from salt melts on mercury cathodes, with simultaneous formation of an amalgam. Because of the high overvoltage of hydrogen on mercury or amalgam cathodes, even nonprecious metals (Zn, Fe, Mn) can be deposited from acid solutions. Nonaqueous solutions may also be used (with Mg, Ti).
3. *Reaction with Sodium Amalgam.* Metal exchange can occur in a fast and stoichiometric reaction through the phase boundary by reaction of sodium amalgam with the salt solution of a precious metal (or, generally, by reacting a nonprecious amalgam with a noble-metal salt solution).

[Top of Page]

11. Mercury Compounds

Gabriele Wühl-Couturier

11.1. Inorganic Compounds

Mercury occupies a special position in group 12 of the periodic table. In contrast to the two other members of this group (zinc

and cadmium), it forms compounds in two valence states.

Compounds of *monovalent* mercury contain ions in the unusual form Hg_2^{2+}. These compounds are not very stable and disproportionate easily to form elemental mercury and the corresponding divalent mercury derivative. Most of the monovalent compounds are sparingly soluble in water. The more soluble salts, e.g., the nitrate, are partially hydrolyzed in aqueous solution: after acidification of these solutions, the poorly soluble compounds can be obtained by precipitation. In addition, compounds of monovalent mercury can be prepared from those of the divalent element by reduction with metallic mercury.

The compounds of *divalent* mercury can be divided into those that are strongly dissociated and those that are weakly dissociated. The former, such as the sulfate and the nitrate, undergo considerable hydrolytic cleavage in water. The weakly dissociated compounds, e.g., the chloride and the cyanide, are less prone to hydrolysis by water. With excess anions they form complexes that are more soluble than the salts themselves.

The starting material for all of these compounds is elemental mercury: the metal is treated initially with a suitable oxidizing agent, e.g., chlorine or nitric acid. The other compounds can be obtained from the resulting oxidation product by further reaction. Multistep processes are often necessary.

11.1.1. *Mercury Chalconides*

Chalconides of monovalent mercury are unknown: precipitation from a solution of mercury(I) nitrate with sodium hydroxide does not lead to the expected mercury(I) oxide, but rather to a mixture of finely divided elemental mercury and mercury(II) oxide. The chalconides of divalent mercury exist in nature as minerals: the oxide HgO as montroydite, the sulfide HgS as cinnabar and metacinnabar, the selenide HgSe as tiemannite, and the telluride HgTe as coloradoite. They can also be produced synthetically. Only the oxide and the sulfide are of practical importance.

Mercury(II) oxide [21908-53-2], HgO, M_r 216.59, ϱ 11.1 g/cm^3, is a red or yellow powder. The color depends on the size of the crystals: the yellow oxide consists of crystals < 2 µm; the red one of crystals > 8 µm in diameter. Samples with particle sizes between these two values appear yellow- to red-orange. Increasing the temperature leads to an intensification of the color: the yellow oxide becomes yellow-orange; the red one, dark red. The crystal lattice is rhombic and identical for both forms. Under certain preparative conditions a hexagonal form exists. This modification has no practical importance and can be converted to the more stable rhombic form by heating above 200 °C.

Heating above 450 °C causes the oxide to decompose into elemental mercury and oxygen. Mercury(II) oxide is sparingly soluble in water and in ethanol. With dilute mineral acids, solutions of the corresponding salts are formed, a method that can be used to prepare these salts.

Production. Mercury(II) oxide can be prepared via the *anhydrous route* by reaction of the elements at 350 – 420 °C under oxygen pressure or by thermal decomposition of mercury nitrates at ca. 320 °C. Production via the *wet route* by precipitation is more important: the oxide is precipitated from solutions of mercury(II) salts by addition of caustic alkali [usually mercury(II) chloride solutions with sodium hydroxide]. Whether the yellow or the red form is obtained depends on reaction conditions: slow crystal growth during heating of mercury with oxygen or during thermal decomposition of mercury(I) nitrate leads to relatively large crystals (i.e., the red form). Rapid precipitation from solution gives finer particles (i.e., the yellow form). Nevertheless, depending on the conditions during precipitation such as stirring speed, pH, temperature, and method of mixing the components, large crystals can be obtained by the wet route and, therefore, the red form is produced [101].

Uses. Red mercury(II) oxide in particular has become increasingly important for the production of galvanic cells with mercury oxide anodes in combination with zinc or cadmium cathodes. These cells are distinguished from other systems in that their voltage remains very constant during discharge; they are used mainly as small button-shaped batteries, e.g., for hearing devices, digital watches, exposure meters, pocket calculators, and security installations. Additional uses of mercury(II) oxide include the following: for the production of mercury(II) salts by treatment with the corresponding acids, and as a reagent in analytical chemistry. Its importance as an additive to antifouling paint for ships and in medicine (e.g., for eye ointment) has decreased.

Mercury(II) sulfide [1344-48-5], HgS, M_r 232.65, is the most important starting material for mercury extraction; it can exist in two forms: α-HgS (*cinnabar,* cinnabarite) has a density of 8.1 g/cm^3, and β-HgS (metacinnabar) has a density of 7.7 g/cm^3. The β-form slowly changes to α-HgS on heating. The latter sublimes at 583 °C. Of the two sulfide minerals, cinnabar is the most important ore for the production of mercury (see Production). When pure, the compound is bright red and forms hexagonal crystals. *Metacinnabar* is black and forms cubic crystals (zinc blende lattice). Both have extremely low water solubility; they are also insoluble in mineral acids and in caustic alkali. They dissolve only in aqua regia, to release sulfur, and in alkali sulfide solutions, to form thio complex-salt ions, such as $[HgS_2]^{2-}$.

Production by either the dry or the wet route is possible. In the former, a mixture of mercury and sulfur is heated. The elements react slowly together even on mixing. Production from aqueous solutions is more important. The sulfide is precipitated from solutions of mercury(II) salts by treatment with hydrogen sulfide, alkali, or ammonium sulfide solutions. Initially, the black sulfide is formed. It can be converted to the more stable red form by heating in the presence of the mother liquor or with ammonium polysulfide solutions. The reaction of mercury and sulfur by heating with a solution of sodium polysulfide has been described [102].

11.1.2. *Mercury Halides*

Halides of both mono- and divalent mercury are known. Of these, only the fluorides are ionic compounds: they undergo hydrolysis with water; mercury(I) fluoride simultaneously undergoes disproportionation. The other halides either are already composed of molecules in the crystal lattice or form these by dissolution or evaporation. Accordingly, their melting and boiling points are low. The halides of monovalent mercury are sparingly soluble in water: the solubility of divalent mercury halides

decreases with increasing molecular mass. Mercury halides form numerous basic compounds.

Mercury(II) fluoride [7783-39-3], HgF_2, M_r 238.59, *mp* 645 °C, ϱ 9.0 g/cm^3, forms colorless octahedral crystals with a cubic ionic lattice (fluorite type). It is unstable in humid air; hydrolysis yields a yellow color. Mercury(II) fluoride is insoluble in organic solvents.

The compound is *produced* from mercury and fluorine at elevated temperature or from mercury(II) oxide and hydrogen fluoride under oxygen pressure at 450 °C [103]. Synthesis from mercury(II) oxide and sulfur tetrafluoride has been suggested [104]. The compound is used in organic synthesis as a fluorinating agent.

Mercury(I) chloride [10112-91-1], calomel, Hg_2Cl_2, M_r 472.09, ϱ 7.15 g/cm^3, is rarely found as a mineral in nature. When pure, it exists as a heavy white powder or as colorless crystals with a silvery luster, having a tetragonal molecular lattice. It sublimes at 385 °C ; above 400 °C the molecules decompose into a vapor composed of mercury and mercury(II) chloride. The substance is sparingly soluble in water, ethanol, diethyl ether, and acetone. A black color occurs in ammonia solution, whereby a mixture of finely divided elemental mercury and mercury(II) ammonium chloride is formed by disproportionation—hence, the name *calomel*, from the Greek word meaning beautiful black.

Production. An intimate mixture of mercury and mercury(II) chloride is heated at 525 °C in closed iron or fused silica tubes, attached to cooled receivers in which calomel vapor condenses [105], [106]. Synthesis from the elements is also possible [101]. Very finely divided mercury(I) chloride can be obtained by precipitation from a dilute nitric acid solution of mercury(I) nitrate and sodium chloride [105].

Mercury(I) chloride finds *application* in calomel electrodes, which serve as standard electrodes for the measurement of electrochemical potential; it is also employed as a fungicide and insecticide in agriculture, and as a catalyst in organic synthesis. Mercury(I) chloride is mixed with gold for painting on porcelain.

Mercury(II) chloride [7487-94-7], corrosive sublimate, $HgCl_2$, M_r 271.50, *mp* 280 °C, *bp* 303 °C, ϱ 5.43 g/cm^3, is a white, heavy, crystalline powder with a rhombic crystal lattice. At the melting point the vapor pressure is 560 kPa; the substance can, therefore, sublime under reduced pressure. The sublimate is moderately soluble in cold water. Its solubility increases sharply with increasing temperature: the saturation limit is 6.2 % at 20 °C and 36.0 % at 100 °C ; the compound may therefore be purified by recrystallization from water. The sublimate is readily soluble in organic solvents, in contrast to mercury(I) chloride, so that a clear solution, for example, in ether, is an indication of the absence of calomel. Molecules of mercury(II) chloride exist as such in all solvents; mixing Hg^{2+} ions with Cl^- in aqueous solution leads immediately to undissociated mercury(II) chloride molecules; this process is used analytically to bind chloride ions in determining the COD of effluents. Aqueous solutions of sublimate are weakly acidic, resulting from hydrolysis of a small amount of the chloride. The sublimate is much more soluble in alkali chloride solution than in pure water, because of the formation of chloro complex ions, e.g., $[HgCl_4]^{2-}$.

Production. For the formation of mercury(II) chloride from the elements, mercury is oxidized with chlorine in heated retorts; the reaction is carried out with the appearance of flame at > 300 °C. The escaping sublimate vapor is condensed in cooled receivers, where it settles as fine crystals. Formation of mercury(I) chloride is avoided by the use of excess chlorine [101]. Mercury and chlorine also react in the presence of water; in this case, intensive stirring is necessary. The chloride formed precipitates as crystals after the solubility limit has been exceeded. If an alkali chloride solution is used in place of water, solutions of chloro complex salts are formed, which are used mainly for the production of other compounds of divalent mercury [107].

Mercury(II) chloride can also be prepared from other mercury compounds. Mercury(II) sulfate, for example, is heated in the dry state with sodium chloride, and the evolving mercury(II) chloride vapor is condensed to a solid in receivers. A warm sublimate solution is obtained from the reaction of mercury(II) oxide and a stoichiometric amount of hydrochloric acid; the chloride separates as crystals on cooling.

Uses. Mercury(II) chloride is an important intermediate in the production of other mercury compounds, e.g., mercury(I) chloride, mercury(II) oxide, mercury(II) iodide, mercury(II) ammonium chloride, and organic mercury compounds. The compound is used as a catalyst in the synthesis of vinyl chloride (⟶ Chlorinated Hydrocarbons – Vinyl Chloride from Acetylene), as a depolarizer in dry batteries, and as a reagent in analytical chemistry. It has minor importance as a wood preservative and retains some importance as a fungicide. Other uses (e.g., as a pesticide or in seed treatment) have declined considerably.

Mercury(II) bromide [7789-47-1], $HgBr_2$, M_r 360.41, *mp* 236 °C, *bp* 320 °C, ϱ 6.05 g/cm^3, forms colorless crystals with a rhombic layered lattice. Its water solubility is highly temperature dependent: the concentration limit is 0.6 % at 25 °C and 18 % at 100 °C. Mercury(II) bromide is readily soluble in diethyl ether and ethanol. The compound is *produced* from mercury and bromine in the presence of water; by dissolution of mercury(II) oxide in hydrobromic acid; or by precipitation from a nitric acid solution of mercury(II) nitrate with addition of sodium bromide [108]. It is *used* as a reagent for arsenic and antimony, as an intermediate in the production of bromine-containing organomercury compounds (see Organic Compounds), and as a catalyst in organic synthesis. The melt is used as a nonaqueous solvent.

Mercury(II) iodide [7774-29-0], HgI_2, M_r 454.45, exists in a red and a yellow form. *Red mercury(II) iodide* is soluble in diethyl ether, chloroform, and methanol. The compound is produced by adding an aqueous solution of potassium iodide to an aqueous solution of mercury(II) chloride with stirring; the precipitate is then filtered off, washed, and dried at 70 °C. A yellow product is formed by sublimation; on cooling, it turns red.

Yellow mercury(II) iodide, *mp* 259 °C, *bp* 354 °C, is unstable at room temperature and is converted to the red form at the slightest touch or upon heating. The transition temperature is 127 °C. Yellow mercury(II) iodide is prepared by pouring an alcoholic solution of the red form into cold water, whereby a pale yellow emulsion is formed, from which the mercury(II) iodide crystallizes after a few hours.

Potassium mercury iodide [7783-33-7], K_2HgI_4, M_r 786.45, readily soluble in water, is prepared by dissolving mercury(II) iodide in a concentrated potassium iodide solution.

11.1.3. *Mercury Pseudohalides*

The *cyanides* and *thiocyanates* of divalent mercury resemble the halides, in that they also exist in solution as undissociated molecules and form highly soluble complexes with an excess of the anion; moreover, numerous basic compounds are derived from them. Mercury(I) cyanide is unknown; the poorly soluble mercury(I) thiocyanate, which can be obtained by precipitation from mercury(I) nitrate solution, has no practical importance.

Mercury(II) Cyanide (⟶ Cyano Compounds, Inorganic – Cyanides of Mercury, Lead, Cobalt, and Nickel***)***

Mercury(II) thiocyanate [592-85-8], $Hg(SCN)_2$, M_r 316.75, ϱ 3.71 g/cm^3, is a white powder which is thermally unstable. Decomposition begins at 110 °C and becomes spontaneous at 165 °C, with the compound swelling to many times its normal volume. In air a blue flame appears, and a dark-colored, voluminous residue is left. The thiocyanate is sparingly soluble in cold water. It is produced by precipitation from mercury(II) nitrate solution with a stoichiometric amount of potassium thiocyanate solution and is used as an analytical reagent and as an intensifier in photography.

11.1.4. *Acetates, Nitrates, Sulfates*

The acetates, nitrates, and sulfates of mercury are composed of ions and undergo hydrolysis with water. The acetate and the sulfate of monovalent mercury are sparingly soluble and the nitrate is quite soluble in dilute acid. Compounds of divalent mercury are readily soluble in dilute acid.

Mercury(II) acetate [1600-27-7], $(CH_3COO)_2Hg$, M_r 316.68, ϱ 3.27 g/cm^3, *mp* 178 °C, exists as a fine white powder or colorless shiny crystal flakes. It is soluble in diethyl ether and ethanol. It is produced by dissolution of mercury(II) oxide in dilute acetic acid and concentration of the resulting solution. Mercury(II) acetate is used for the synthesis of organomercury compounds, as a catalyst in organic polymerization reactions, and as a reagent in analytical chemistry.

Mercury(I) nitrate [10415-75-5], $Hg_2(NO_3)_2 \cdot 2\,H_2O$, M_r 561.22, ϱ 4.68 g/cm^3, *mp* 70 °C, which forms colorless crystals (monoclinic ionic lattice), is produced by dissolving mercury in cold dilute nitric acid and crystallizing the compound from the resulting solution.

Uses. Mercury(I) nitrate is the most readily available soluble salt of monovalent mercury and is, therefore, an important intermediate for other mercury(I) derivatives; sparingly soluble compounds can easily be prepared from it by precipitation from aqueous solution. Thermal decomposition leads to red mercury(II) oxide. “Millons reagent,” a solution of mercury nitrate and nitrous acid in dilute nitric acid, is used as an analytical reagent for the indication of tyrosine-containing proteins.

Mercury(II) nitrate [10045-94-0], $Hg(NO_3)_2$, M_r 324.60 ; $Hg(NO_3)_2 \cdot H_2O$, M_r 342.62, forms colorless, hygroscopic crystals. Apart from the anhydrous salt and the monohydrate, several other hydrates and basic compounds are known. Mercury(II) nitrate is produced by dissolving mercury in hot concentrated nitric acid; the resulting solution is concentrated and the nitrate crystallized by cooling. *Uses* include the production of other divalent mercury derivatives, as a nitrating agent in organic synthesis, and as an analytical reagent.

Mercury(I) sulfate [7783-36-0], Hg_2SO_4, M_r 497.24, ϱ 7.56 g/cm^3, is a colorless microcrystalline powder (monoclinic crystal system) that is very sensitive to light. The compound is prepared by precipitation from a solution of mercury(I) nitrate with sulfuric acid or sodium sulfate solution, or by electrochemical oxidation of mercury in dilute sulfuric acid. It is used as a depolarizer in standard cells after Clark and Weston.

Mercury(II) sulfate [7783-35-9], $HgSO_4$, M_r 296.65, ϱ 6.49 g/cm^3, is a white powder (rhombic crystal type). It is prepared by fuming mercury with concentrated sulfuric acid or by dissolving mercury(II) oxide in dilute sulfuric acid and evaporating the resulting solution until the compound crystallizes. Mercury(II) sulfate is used in analytical chemistry to bind chloride ions in the determination of the COD of wastewater [109] ; as a catalyst in the production of acetaldehyde from acetylene (⟶ Acetaldehyde – Production from Acetylene) and of anthraquinonesulfonic acids (⟶ Anthraquinone Dyes and Intermediates – Anthraquinonesulfonic Acids) ; and as a depolarizer in galvanic elements.

11.1.5. *Mercury – Nitrogen Compounds*

Reaction of mercury(II) compounds with ammonia solution leads, depending on reaction conditions, to amine complexes or mercury(II) nitrogen compounds. Of the numerous compounds known, only mercury(II) amidochloride has any practical importance.

Mercury(II) amidochloride [10124-48-8], $HgNH_2Cl$, M_r 252.07, ϱ 5.38 g/cm^3, is a fine, white crystalline powder with a rhombic crystal lattice. It is insoluble in water and in ethanol, and soluble in warm acid, ammonium carbonate solution, and sodium thiosulfate solution. On heating it decomposes without melting. The compound can be precipitated from mercury(II) chloride solution with an ammonia solution. Mercury(II) amidochloride is used in the treatment of severe skin disease, as an eye ointment, and as a veterinary preparation; its importance is declining because of the development of mercury-free products.

11.1.6. *Analysis, Storage, and Transportation; Protective Measures*

Analysis. Determination of the purity of mercury compounds, which consists of the determination of trace amounts of foreign cations and anions, involves specific reactions of these ions [110] ; conventional methods for the trace analysis of cations have been supplemented or replaced by methods involving the simultaneous determination of several elements by means of plasma emission spectroscopy [111].

Storage and Transportation. Plastic-lined steel drums are normally employed for packing. Small amounts (e.g., for chemical laboratories) are usually placed in plastic or glass bottles. Many compounds are light-sensitive and must, therefore, be adequately protected from light sources. Proper consideration must be given to the chemical and toxic properties (see Chap. Toxicology and Occupational Health) of mercury compounds and the necessary protective measures; this is particularly true in container labeling.

Protective Measures. Most mercury compounds, because of their toxicity (see Chap. Toxicology and Occupational Health), require the same protective measures during production and processing as metallic mercury. Furthermore, environmental protection necessitates appropriate precautions (see Chap. Environmental Protection).

11.2. Organic Compounds [112], [113]

Organic mercury derivatives are among the oldest known organometallic compounds. Organometallic compounds of *divalent mercury* are stable toward air, oxygen, and water. In contrast, organic derivatives of *monovalent mercury* are unstable and can be prepared only at low temperature. After 1990, use of organic mercury derivatives in the Federal Republic of Germany is difficult because mercury is not allowed in industrial effluent or wastewater.

Organomercury compounds can be divided into two major groups: compounds of the type R^1–Hg–R^2, where R^1 and R^2 are aliphatic or aromatic groups, and compounds of the type R–Hg–X, where R is aliphatic or aromatic and X is a halogen or an acid group.

11.2.1. *Production [114], [115]*

Synthesis of organic mercury compounds can be carried out by reaction of Grignard reagents (⟶ Magnesium Compounds – Organomagnesium Compounds) with mercury halides. To obtain pure products, the mercury salt and the Grignard reagent must contain the same anion

$$RMgX + HgX_2 \rightarrow RHgX + MgX_2$$

Furthermore, organic mercury compounds can be produced by the reaction of sulfinic acids or their sodium salts with mercury(II) halides [116]:

$$R-SO_2H + HgX_2 \rightarrow R-HgX + SO_2 + HX$$

Hydroxy- or alkoxymercury derivatives can be obtained via the solvomercuration reaction (Hofmann – Sand reaction) by addition of mercury(II) compounds to alkenes in aqueous, alkaline, or alcoholic solutions [114], [117]:

$$R^1-CH=CH-R^2 + Hg(OAc)_2 + R^3OH \rightarrow (R^3O)CHR^1-CHR^2HgOAc + AcOH$$

Organic mercury compounds can be converted to other organometallic derivatives by electrolysis or transmetallization [118].

11.2.2. *Compounds of the Type R^1–Hg–R^2*

Dialkyl- or Alkoxyalkylmercury Compounds. *Dimethylmercury* [593-74-8], CH_3–Hg–CH_3, M_r 230.60, a colorless, sweet-smelling liquid, is very toxic; it has a density of 3.069 g/cm^3, *bp* 96 °C, and is soluble in ethanol and diethyl ether. In acidic aqueous solution, dialkylmercury compounds are hydrolyzed to monoalkylmercury derivatives. Dimethylmercury is formed in organisms by enzymatic methylation of elementary mercury or a mercury compound. In an organism it is easily converted to methylmercury (CH_3–Hg^+) or methylmercury(II) chloride (CH_3–Hg–Cl). These species can react with free S–H groups of biologically important molecules, leading to the diseases that have been caused by mercury and its derivatives (Minamata, Niigata, Iraq; see Chap. Toxicology and Occupational Health).

Mixed dialkylmercury compounds (e.g., methylpropylmercury compounds) are less volatile than symmetrical dialkylmercury compounds (e.g., diethylmercury). Dialkylmercury derivatives are generally very reactive and can also undergo transalkylation with simple alkyl halides (e.g., ethyl iodide).

Diarylmercury Derivatives. The best known diarylmercury derivative is diphenylmercury, $(C_6H_5)_2Hg$, which is produced by direct mercuration.

11.2.3. *Compounds of the Type R–Hg–X*

Monoalkyl- and Monoarylmercury Carboxylates. The best known representatives of the monoalkyl- or monoarylmercury carboxylate series are phenylmercury(II) acetate and phenylmercury(II) oleate.

Phenylmercury(II) acetate [62-38-4], C_6H_5–Hg–O–$COCH_3$, M_r 336.75, forms colorless, lustrous crystals, *mp* 149 °C; it is soluble in ethanol, benzene, and acetic acid and sparingly soluble in water. This compound was used as a preservative for paints in an acetic acid –ethanol solution.

Phenylmercury(II) oleate [104-60-9], C_6H_5–Hg–O–CO–$C_{17}H_{33}$, M_r 559.17, is a white, crystalline powder, *mp* 45 °C; it is

insoluble in water and soluble in organic solvents or oils. It is used as a preservative against mold and as a catalyst in the production of adhesives and plastics.

Organic mercury salts play an important role as intermediates in catalyzed reactions, making these products of interest in carbene chemistry [114], [115], [119-121]. Addition of an ethereal solution of diazomethane to a solution of benzylmercury(II) iodide in tetrahydrofuran at 0 °C leads to immediate and quantitative formation of the carbenoid.

Monoalkyl- and Monarylmercury(II) Halides, Borates, and Nitrates. *Phenylmercury*(*II*) *chloride* [100-56-1], C_6H_5–Hg–Cl, M_r 313.18, and *basic phenylmercury*(*II*) *nitrate* [8003-05-2], C_6H_5–Hg–OH · C_6H_5–Hg–NO_3, M_r 634.45, are used as antimycotics and antiseptics.

The halides have higher melting points and are less irritating to the skin than other ionogenic salts. These organic mercury compounds were once used as antisyphilitics and diuretics. Their use for this purpose has been restricted or stopped because of their numerous side effects and toxicity.

Phenylmercury(*II*) *borate,* C_6H_5–Hg–O– $B(OH)_2$, M_r 338.50, is soluble in water, ethanol, and glycerol. It is used as a skin disinfectant and for the disinfection of medical instruments.

Monoalkyl- and Monoarylmercury Hydroxides. *Phenylmercury*(*II*) *hydroxide* [100-57-2], C_6H_5–Hg–OH, M_r 294.72, is a white, crystalline powder, *mp* 197 – 205 °C, which is soluble in glacial acetic acid and ethanol, and sparingly soluble in water. Phenylmercury(II) hydroxide is used as a starting material for numerous organic mercury compounds [121], [122]. Esters of the type C_6H_5–Hg–O–CO–R, produced from the hydroxide, are used as fungicides (seed protectants) and disinfectants. Regulations for cosmetics in the Federal Republic of Germany permit the use of this compound as a preservative in eye makeup and makeup remover, up to a concentration of 0.007 % (based on Hg).

11.2.4. *Analysis*

Organic mercury compounds may be analyzed in the following ways: they can be determined *qualitatively* (1) by digestion with concentrated sulfuric acid and 30 % hydrogen peroxide or (2) by digestion with 10 % sulfuric acid and subsequent addition of dithizone (diphenylthiocarbazone) in carbon tetrachloride solution. In the presence of mercury the green solution turns orange. Organic mercury compounds can be determined *quantitatively* by atomic absorption spectroscopy.

[Top of Page]

12. Economic Aspects

Matthias Simon

A detailed description of the development of production is given in the metal statistics of Metallgesellschaft AG, Frankfurt [123] (Table 12). Over the last ten years, production figures have changed only slightly. Because of reduced demand, many mines and smelting plants are no longer operating or have greatly cut back production. Intervention by the former Soviet Union has seriously depressed the price of mercury to dumping price levels, although this has been resisted by Spain and Algeria. The change in the price of mercury since 1979 is shown in Table 13 [123]. A survey of previous mercury prices is included in [124].

Table 12. Production of mercury in tonnes [123]

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
EEC	113	87	91	59	328	212				1471	1553
Fed. Rep. of Germany	99	84	91	56	76	53					
Italy	14	3		3	252	159					
Spain	926	1020	1116	1721	1560	1540	1416	1520	1539	1471	1553
Finland	22	39	46	75	67	71	65	80	130	147	147
Former Yugoslavia	108						52	72	88	75	67
Europe [*]	1169	1146	1253	1855	1955	1823	1533	1672	1757	1693	1767
Turkey	162	173	163	154	204	246	162	182	226	262	202
Asia [*]	162	173	163	154	204	246	162	182	226	262	202
Algeria	1049	1055	506	841	877	386	828	587	795	690	773
Africa	1049	1055	506	841	877	386	828	587	795	690	773
Dominican Republic	18	17	21	6	3	2	4	2	1		
Mexico	333	76	68	145	240	295	221	384	264	345	350 [a]
United States	974	834	1018	1058	962	888	864	657	570	470	100 [a]
America	1325	927	1107	1209	1205	1185	1089	1043	835	815	450
Western countries	3705	3301	3029	4059	4241	3640	3612	3484	3613	3460	3192
Former Czechoslovakia	183	196	171	159	153	151	144	152	158	168	164
Former states of USSR [a]	2200	2000	2000	1800	1700	1700	1700	1600	1600	1500	1650

China [a]	700	600	700	800	800	800	850	800	800	850	900
Eastern countries	3083	2796	2871	2759	2653	2651	2694	2552	2558	2518	2714
Total world	6788	6097	5900	6818	6894	6291	6306	6036	6171	5978	5906

[*] Excluding Eastern-bloc countries.
[a] Estimates.

Table 13. Quoted prices for mercury in London and New York

Year	European price		U.S. price	
	£ per flask (34.473 kg)	€/kg	$ per flask (34.473 kg)	€/kg
1979	283.959	7.72	281.882	7.66
1980	392.631	10.60	380.447	10.52
1981	413.018	13.90	415.385	12.79
1982	371.133	13.37	370.934	13.37
1983	307.275	11.64	322.443	12.20
1984	302.280	12.78	314.381	13.30
1985	284.130	12.46	310.957	13.60
1986	187.500	6.13	232.785	7.56
1987	245.640	6.54	295.503	7.87
1988	297.17	7.77	335.52	8.76

The main producers of mercury extracted by mining are as follows [125] (figures refer to flasks):

Algeria	23 000
China	20 000
Finland	2 300
Former Yugoslavia	2 000
Mexico	10 000
Spain	42 000
Former Czechoslovakia	4 400
Turkey	6 000
Former Soviet Union	66 000
United States	14 000[*]

[*] The sole producer in the United States discontinued mining oprations from the middle of 1987 to February 1988. The production for 1988 is 14 000 flasks.

The Algerian mercury producer, Entreprise Nationale des Nonferreux et Substances Utiles (ENOF), quotes production prices of $ 300 per flask. Most of the Chinese production is exported to the United States. China claims to have the largest mercury resources in the world. Guizhou Province contains five mines, accounting for ca. 90 % of Chinese production. Italy, once a large mercury producer, now imports mercury from Algeria, the Netherlands, and former Yugoslavia. In Mexico, ca. 350 t of mercury (ca. 10 000 flasks) was extracted from mines in 1986, a large part of the production being exported to Brazil and Argentina. Spanish mercury production is ca. 40 000 fl. The main importers are the United States, Belgium, Luxemburg, and France. The former Soviet Union produces ca. 67 000 fl of mercury; in contrast to the high export level in 1986, almost the entire production is now reserved for domestic use. The capacity of production plants at the Anzob antimony – mercury complex in Tadzhikistan has been doubled. The major proportion of Yugoslav production of ca. 2000 fl of mercury is exported.

The development of prices and production figures will certainly be influenced substantially by acceptance of the metal mercury, and in large sectors no foreseeable substitute exists. The utilization of mercury in a highly industrialized country such as the Federal Republic of Germany is shown in Table 14 [126]. The purchase of mercury by individual sectors has decreased sharply since 1980. The main user of mercury is the electrical engineering industry, followed by the alkali chloride industry. The use of mercury in paints, pigments, and pesticides has fallen sharply.

Table 14. Breakdown of mercury use (in tonnes) in the Federal Republic of Germany according to sector [126]

	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985

Alkali chloride industry (without changes in stock levels)	128.0	127.0	99.0	103.0	78.0	87.5	87.5	72.0	50.4	46.3	40.7	31.6	41.8
Catalysis	45.0	29.0	9.0	10.0	11.6	4.8	27.5	18.1	14.1	17.9	6.0	4.5	3.9
Paints, dyes	19.5	18.9	5.6	12.4	12.6	12.6	9.1	3.8	3.5	3.8	1.0	0.6	0.3
Pesticides	30.5	33.7	26.6	27.6	28.7	28.5	33.0	31.8	26.7	4.1	19.0	14.0	9.0
Electrical engineering	26.2	42.1	40.6	43.4	40.4	43.5	45.4	51.5	55.0	54.2	53.1	56.2	65.8
Control instruments and apparatus construction	14.8	17.3	8.7	10.3	14.6	13.4	13.3	16.5	15.7	17.1	7.0	6.1	7.2
Chemicals and reagents	25.3	7.0	33.1	45.8	22.0	31.4	43.2	35.5	27.1	53.9	36.2	24.7	
Medicine	24.2	25.2	25.0	26.0	23.9	24.3	24.2	24.2	22.9	23.6	23.5	24.3	24.1
Miscellaneous	32.8	31.1	44.2	46.7	30.0	30.0	30.0	42.1	32.5	35.7	30.1	30.2	30.2
Total	346.3	331.3	291.8	325.2	261.8	276.0	313.2	295.5	247.9	256.6	216.6	192.2	182.3
Total, % (1980 = 100)	117	112	99	110	89	94	106	100	84	86	73	65	62

[Top of Page]

13. Toxicology and Occupational Health

Stefan Halbach

13.1. Chemical Forms and Mechanism of Action

The unique physical properties of mercury, especially its high vapor pressure and therefore partial occurrence in the vapor phase at room temperature, have been referred to above. In addition to the ingestion of mercury compounds (mercurials), intoxications can thus occur by the inhalation of the vapor. Among the numerous organic compounds, the short-chain alkylmercury compounds, mainly methylmercury, are most important from the toxicological point of view because of their chemical stability. In contrast, phenyl Hg compounds are rapidly degraded to inorganic mercury compounds in mammalian tissues and therefore are similar to inorganic mercury regarding their toxic effects.

The interaction of mercury with biological ligands is based on the very high affinity of the mercury cation (Hg^{2+} or $R\text{-}Hg^{+}$) to sulfur [127-129]. If the binding occurs to sulfhydryl groups relevant for the structure of a protein or for the cellular function, each will be seriously disturbed. Since sulfhydryl groups are ubiquitous, the high affinity of mercurials leads to their potent but unspecific inhibition as seen with many enzymes [130]. Nevertheless, cases of poisoning are characterized by defined clinical manifestations from the target organs. It is assumed, therefore, that the selective toxicity of the different forms of mercury results from a selective distribution to the various compartments. Distribution and the permeation of biological barriers (cell membranes, blood-brain barrier) depend on lipophilicity which is high for Hg^0 and for alkyl mercurials but low for mercury(II) salts and for organic compounds with polar groups [131], [132]. A further modulator of toxicity is the air-liquid partition coefficient of Hg^0 which favors the accumulation of mercury vapor in plasma and whole blood [133]. All these factors acting together with various routes of uptake create a distinct etiology and symptomatology in intoxications with different Hg compounds.

13.2. Exposure

Occupational exposure to Hg vapor occurs in specific industries and professions such as chlor-alkali electrolysis (mercury cell process), production and recycling of fluorescent lamps and batteries, and in dentistry offices. Background exposure originates from the release of minute amounts of vaporous and inorganic mercury from dental amalgam (see Section Dental Amalgam Fillings). Exposure to organic mercury compounds may occur in the chemical industry, whereas background exposure results from intake with food, mainly as methylmercury from seafood. In summary, the intake of all forms of mercury amounts to 11 – 28 µg per day for the general population in the USA and Europe [134] but is largely dependent on dietary habits.

13.3. Metabolism

13.3.1. *Mercury Vapor and Inorganic Mercury*

Both chemical forms have in many respects a common metabolism, except for the routes of uptake.

Absorption. Approximately 80 % of inhaled *Hg vapor* is absorbed from the lungs into the bloodstream. The dissolved vapor is rapidly oxidized to Hg^{2+}, partly in the erythrocytes and partly after diffusion into other tissues. This oxidation is catalyzed by the enzyme catalase in presence of traces of hydrogen peroxide [135], [136]. For humans, the absorption of inorganic mercury from ingested food was estimated to range between 5 and 10 % of the dose [137]. The gastrointestinal absorption of mercury(I) is sufficient to cause serious illness (pink disease, see Pink Disease.). The absorption from ingested *liquid mercury metal* is probably below 0.01 % [138]. The cutaneous absorption can be substantial as shown in patients suffering from nephrotic syndrome after the use of ointments containing inorganic mercury [139].

Distribution. In human blood, the red cell: plasma ratio of mercury after exposure to Hg vapor is between 1 and 2. Studies on animals and humans showed that mercury has an affinity for epithelial cells and glands [140]. Based on animal data, 50 – 90 % of the body burden is located in the kidneys, where a large portion is bound to metallothionein, a low molecular mass, cysteine-rich protein [141]. Inorganic mercury penetrates the placental barrier as shown by the good correlation between concentrations in fetal and maternal blood of mothers with amalgam fillings [142].

Excretion. As demonstrated by short-term exposure of volunteers to Hg vapor, half of the mercury excretion is through the feces and much less through the urine. After long-term exposure, the urinary route (60 %) predominates over the fecal route (40 %) which reflects the renal accumulation [143]. A small portion of absorbed ionic mercury is exhaled as vapor after reduction in tissues [144]. Tracer studies on volunteers revealed a fecal excretion of 75 – 92 % of the dose after ingestion of inorganic mercury [145].

Toxicokinetics. After exposure to Hg vapor or to mercury(II), the elimination half-life for the whole body is 58 d which is close to that of 64 d for the dominating Hg pool in the kidneys [138]. Upon short exposure to Hg vapor, about 90 % of the mercury in blood is cleared with a half-life of 2 – 4 d, followed by a second phase with a half-life of 15 – 30 d [146], [147]. Tracer studies on volunteers showed a half-life of 19 d for elimination of most of the mercury in brain [143], [147], [148]. Autopsies on mercury miners many years after termination of exposure revealed high amounts in the pituitary, probably bound as biologically inactive mercury selenide [149], which means that the half-life in brain could be in the range of years [138].

13.3.2. *Organic Mercury Compounds*

Reliable data are available for monomethylmercury (CH_3HgCl; Me-Hg) only which, therefore, shall be described here.

Exposure and Absorption. Nonoccupational exposure to Me-Hg occurs almost exclusively through the diet, mainly fish, and remains without effect at background levels. Subacute intoxication may occur after ingestion of contaminated food. Prenatal exposure leading to neurotoxic effects has been found in the epidemics (see Prenatal Exposure.). More than 90% of the methylmercury in the diet is absorbed into the bloodstream [150]. Exposure to dimethylmercury appears to result in a very high absorption through lung and skin (see Effects on the Nervous System.).

Distribution. About 90 % of the absorbed Me-Hg in whole blood is bound to the red cells; further distribution to the organs is completed within four days [151]. It appears from studies on volunteers that the brain-blood concentration ratio is about 6 [152]. Major depositions in other organs are found in liver and kidney. Of particular interest is the binding in hair during hair formation which facilitates the retrospective assessment of the exposure (see Effects on the Nervous System.).

Conversion and Excretion. In the mammalian organism, Me-Hg is converted to inorganic Hg. Macrophages in the spleen are possibly involved in this breakdown [153]; enteral Me-Hg may be demethylated by the intestinal microflora [154]. Binding of Me-Hg and of Hg^{2+} to glutathione in the liver and subsequent secretion into the bile seems to be the essential step for fecal excretion [155]. Inorganic Hg, being poorly absorbed in the intestine, is mostly excreted, whereas unmetabolized Me-Hg can be reabsorbed to a great extent (enterohepatic circulation).

Toxicokinetics. Me-Hg is so mobile in the body that the excretion is the rate-limiting step for elimination. Therefore, the elimination rate is proportional to the actual body burden and can be described by a single biological half-life [150]. The Iraq epidemic as well as studies on volunteers have shown that the whole-body half-life is 70 d. Half-life in blood appears to be about 50 d. For practical purposes, the amount of Me-Hg in 1 L of blood can thus be assumed to equal 1 % of the body burden or the daily intake for a 70-kg person in steady state [134].

13.4. Toxic Effects

13.4.1. *Mercury Vapor and Inorganic Mercury*

Acute exposure to high doses occurs accidentally by incautious handling of the metal or by suicidal ingestion of mercury salts. On the whole, such incidents are less frequent than long-time exposure to intermediate doses of Hg vapor in the occupational environment (at or below the threshold limit value) or to low doses released from dental fillings of amalgam (see Section Dental Amalgam Fillings). According to its importance for humans, the focus of this section will be on chronic exposure at low levels.

13.4.1.1. *Acute Toxicity*

Severe affections of the respiratory system (chest pains, dyspnoea, cough, hemoptysis, impairment of pulmonary function, and evidence of interstitial pneumonitis) resulted from the acute exposure of workers to mercury vapor at estimated levels up to 44 mg/m^3 for 4 – 8 h [156]. Mercurial pneumonitis was also observed in an episode of unsafe manipulations with metallic mercury at home [157].

The symptoms of acute toxicity of $HgCl_2$ usually consist of severe gastrointestinal damage, cardiovascular collapse, and acute renal failure. Given in acute massive doses, inorganic and organic mercury will denature proteins, inactivate enzymes, and cause disruption of any tissue with which it comes into contact in sufficient concentrations. Gastrointestinal lesions, ranging from mild gastritis to severe ulceration of the mucosa, and renal lesions resulting in renal failure were found in cases of human poisoning following oral ingestion of mercuric chloride. The lethal doses ranged from 29 to at least 50 mg/kg body weight [158].

13.4.1.2. *Chronic Toxicity*

Upon long-term exposure to toxic levels of Hg vapor, the central and peripheral nervous system is the critical organ. Chronic exposure to mercury salts, particularly in the working environment, often involves exposure to a mixture of salt and vapor and mainly affects the kidney.

Effects on the Nervous System. In early stages, occupational exposure to inorganic mercury may induce unspecific symptoms such as irritability, excitability, shyness, and insomnia. With longer lasting exposure, a fine tremor develops, initially affecting the hands and later spreading to the facial muscles, causing violent muscular spasms in the most severe cases. Parallel to the tremor, mercurial erethism develops which is characterized by severe behavioral and personality changes, increased excitability, loss of memory, and insomnia, which may develop into depression. In milder cases, erethism and tremor subside slowly over a period of years after cessation of exposure. Peripheral nerves may also be affected; pathological changes in conduction velocity have been observed in mercury-vapor exposed workers [159], [160].

Renal Effects. The nephrotoxic effect of inorganic mercury consists of two types of renal injury. The first is a glomerular lesion caused by an autoimmune reaction induced by mercury and resulting in antibody formation against the glomerular membrane, deposition of immune complex, glomerular nephritis, proteinuria, and nephrotic syndrome. The immunotoxic response has been extensively investigated in the rat model and appears to be controlled genetically [161]. The second lesion affects the proximal tubules in parallel with the accumulation of mercury in the tubular cells. This damage results in a loss of tubular enzymes (gamma-glutamyl transferase, beta-galactosidase and beta-glucuronidase) [162]. Reabsorption is also decreased which leads to an elevated excretion of trace elements such as zinc and copper [163].

Dose-Response Relationships of Neurotoxic and Nephrotoxic Effects. An extensive study on workers in chlor-alkali plants revealed that the persons exposed to mercury vapor at or below a level of 100 µg/m^3 showed no significant neurotoxic signs or symptoms [164]; in addition, there was no evidence of poisoning at exposure concentrations below 10 µg/m^3. Studies on groups with high occupational exposure have pointed to the importance of urine Hg peaks in excess of 500 µg/L for the development of neurological signs and symptoms [165], [166]. Urine values above 100 µg/L have been associated with decreased performance in psychomotor ability tests [167]. Evidence for effects at urinary concentrations < 50 µg/L remains inconclusive [168-170].

Studies on chlor-alkali workers showed the onset of *renal effects* at urine concentrations between 30 and 50 µg/L, corresponding to an exposure of 0.3 mg Hg/m^3 air and to a blood Hg level of ≥ 30 µg/L [162], [171]. On the other hand, no signs for renal dysfunction were found in occupationally exposed persons with a mean urine Hg concentration of 56 µg/g creatinine [172].

Thus, early effects on the two target organs for inorganic mercury occur at similar Hg concentrations in air and urine. Preclinical neurological and renal effects may be detected with subtle diagnostic methods between 50 and 100 µg Hg/L urine, corresponding to more than 20 µg/L blood, whereas clinical manifestations can be expected to increase at levels above 200 µg/L urine. Hence, these concentrations can be viewed as the lowest observed effect level (LOEL).

Skin Reactions. Metallic mercury has rarely been found to cause hypersensitivity. However, mercury compounds (thiomersal and ammoniated mercury) give rise to a type IV cell-mediated delayed hypersensitivity reaction [173]. Few cases of allergic dermatitis have been reported among dental personnel. A very rare condition observed in individuals with dental amalgam fillings are lesions of the oral mucosa such as lichen planus [174].

Pink Disease. A particular syndrome occurring in children exposed to inorganic mercury is pink disease (acrodynia, Feer's disease). It is characterized by irritability, sleeping disturbances, copious sweating, photophobia, and generalized rash. The extremities become cold, painful, red, and swollen, and the skin desquamates. Calomel-containing teething powder and ointments as well as sources of Hg vapor in the children's environment are the cause of this disease of unknown etiology [175-177].

13.4.2. *Organic Mercury Compounds*

Phenyl-Hg compounds are rapidly metabolized to inorganic Hg [143] which makes their toxicity similar to that of inorganic Hg. Alkyl compounds are more persistent and produce specific signs of disease.

Effects on the Nervous System. The syndrome occurring upon subacute and chronic exposure to *methylmercury* was first described for a group of occupationally exposed workers [178]; it has reappeared in the epidemics in Minamata [179], [180] and Iraq [181]. The typical sequence of effects is paresthesia in the extremities, ataxia, dysarthria, constriction of the visual field, and deafness. The beginning of symptoms is preceded by a latent period from one to four months.

The mechanism for the selective damage of methylmercury to the nervous system is still unclear and is debated in various experimental approaches. Muscular weakness observed in severe cases of poisoning is possibly related to the inhibition of acetylcholine transmission at the neuromuscular junction [182]. Protein synthesis in nerve cells appears to be particularly sensitive to methylmercury [183-185]. Activated oxygen species and an increased lipid peroxidation may also significantly contribute to the alkylmercury-induced neurotoxicity [186].

Upon exposure to *dimethylmercury*, similar but more intensive symptoms develop after a longer latency period. The few cases of poisoning reported in the literature all ended fatally, which means that this compound should be considered as extremely toxic [187].

Dose-Response Relationships. The relationships between concentrations of Me-Hg in blood or hair and the magnitude or frequency of an effect had been studied in the epidemics (see Effects on the Nervous System.) as well as in populations with an unusually high consumption of fish. Among the last group, no signs of poisoning were found below blood Hg levels of 200 – 360 µg/L [150]. This is in accordance with the observations on adult Minamata and Iraq victims. The most sensitive groups of these developed paresthesia at blood Hg levels above 200 – 400 µg/L, corresponding to hair mercury levels of 50 – 125 µg/g [150]. Considering 300 µg/l as the lowest observed effect level in blood means that the daily uptake is about 300 µg for a person of this group in steady state (see Dose-Response Relationships of Neurotoxic and Nephrotoxic Effects.).

Prenatal Exposure. Estimates between 1.2 and 1.9 for the infant : mother blood-mercury concentration ratio reveal the

importance of the transplacental transfer of methylmercury [188]. Follow-up studies on the Iraq epidemic indicated that maternal hair levels of 10 – 20 µg/g raised the incidence of psychomotor retardation of the infant above the background frequency [189]. According to [134] this range of values in maternal hair must be considered as critical for the development of the fetal nervous system. Epidemiological studies on fish-eating populations of the Faeroe islands in the 1990s reported the possibility of slight neuropsychological effects in children of mothers with hair Hg levels of 4.3 µg/g during pregnancy [190], while a similar study in the Seychelles found no indication of such effects at 6 µg/g [191]. The mechanism underlying the prenatal neurotoxicity of Me-Hg may be a derangement of the migration of neurons to their proper location in the cortical layers at the time of the fetal brain development [192].

13.4.3. *Dental Amalgam Fillings*

A special case of exposure to inorganic mercury is dental amalgam for this is a long-term exposure to low levels on an epidemiological scale. Urine and blood Hg levels are clearly elevated in presence of amalgam if compared to amalgam-free persons, but remain within the normal range [193], [194]. These levels are influenced, however, by dietary organic Hg, rendering the estimation of the amalgam-related portion of blood Hg more difficult. A parameter more representative for amalgam-related mercury may be the Hg level in blood plasma which responds with high sensitivity to alterations of the amalgam status [195]. A toxicokinetic evaluation of the available data has shown that the daily Hg uptake from amalgam varies between 3 and 17 µg [196] which is in the same range as the dietary intake and within the limits of the acceptable daily intake ([197]; see Section Critical Levels, Risk Assessment). More recent estimations of the absorbed daily dose consistently revealed average levels below 10 µg/d [193], [198], [199].

13.4.4. *Carcinogenicity*

Me-Hg and, to a lesser extent, inorganic Hg have been shown to be potentially genotoxic in animal and in-vitro studies [134], [200]. However, the relevance for human exposure seems to be limited as clinical cases of cancer have not been linked to mercury so far.

13.5. Indicator Media and Reference Values

13.5.1. *Mercury Vapor and Inorganic Mercury*

Ongoing occupational exposure is equally well reflected by mercury levels in blood or urine (24-h urine or reference to creatinine) [168]. Low-level exposure to Hg vapor or inorganic Hg is easier to assess through total Hg concentrations in urine than in blood because methylmercury from food, which is poorly excreted by the kidney, does not interfere with the measurement. People without occupational exposure have Hg concentrations in blood of 3 – 5 µg/L and in urine up to 5 µg/L [143], [196], [201]. Under steady-state conditions in long-term exposure, the following proportionality can be assumed: a mercury concentration in air of 40 µg/m^3 equals about 15 – 20 µg/L blood and about 50 µg/L urine [138].

13.5.2. *Methylmercury*

It follows from the toxicokinetics that whole blood or red cells are the most reliable media for the assessment of body burden of Me-Hg. Urine is not satisfactory as it is a minor route of excretion (see Section Mercury Vapor and Inorganic Mercury). The concentration in scalp hair during its growth parallels closely that in blood and is about 250 times higher than in blood, hence hair Hg levels are useful in estimating actual and previous exposure [143]. The average concentration in background exposed populations is up to 5 µg/L in whole blood and 2 µg/g in hair [134].

13.6. Critical Levels, Risk Assessment

Factors applied to the LOELs (see Dose-Response Relationships of Neurotoxic and Nephrotoxic Effects.) scale them down to levels of acceptable health risks or to those assumed to have no effect if they are not exceeded (critical levels, regulatory levels). Depending by which institution they have been set, these levels differ to some extent, and therefore more than one are given where possible. In the workplace, the Hg concentration in air shall not exceed 100 [202] or 50 µg/m^3 [203], whereas for permanent inhalation, levels between 0.2 (minimum risk level) [204] and 1 µg/m^3 [205] are considered safe. For an adult, the provisional tolerable weekly intake of mercury with food shall not be more than 300 µg of total mercury, containing not more than 200 µg of methylmercury, which is about 1/10 of the LOEL [134], [197], [206]. The upper limit in drinking water is set at 1 µg/L (Germany) and is based on the analytical detection limit rather than on toxicological evidence.

Regular intake of small doses of methylmercury over extended time periods leads to steady-state concentrations in organs and biological fluids [150], which means proportionality between these concentrations. This is likely to hold also for the exposure to Hg vapor. Therefore, Hg determinations in blood, urine and scalp hair can be used as indicators of the mercury burden of target organs or of the whole body. Blood levels, indicating exposure to all forms of mercury, range normally below 5 µg/L with a critical value of 50 in occupationally exposed persons. Urine levels, being representative of exposure to inorganic and vaporous Hg, are normally below 10 µg/L with a critical limit of 100 in exposed workers [138], [202]. In cases of dietary overexposure, symptoms appear at hair levels of 50 – 125 µg/g (LOEL) [134]. Because of the neurotoxic susceptibility of the fetal brain the critical level in maternal hair during pregnancy is estimated at 10 µg/g [134].

[Top of Page]

References

1. P. T. Craddock, *Bull. Metals Museum* **10** (1985) no. 10, 3 – 25.
2. G. Agricola: *12 Bücher des Bergbaus und des Hüttenwesens*, pp. 370 – 375.
3. R. C. Weast, M. J. Astle (eds.): *Handbook of Chemistry and Physics*, 64th ed., CRC-Press, Boca Raton, Florida 1983.
4. Hüls, DE 3 709 570, 1987 (S. Sridhar).
5. K. R. Suttill, *Eng. Min. J.* (1989) 24 – 27.
6. W. Peters, *Erzmetall* **35** (1982) no. 7/8, 389 – 395.
7. H. Köse, M. Kemal, F. Şim Şir, *Erzmetall* **42** (1989) no. 6, 276 – 278.

8. SRI International, *Chemical Economics Handbook*, Menlo Park 1984.
9. X. Wang, Z. Wu: *Mineral Processing and Extractive Metallurgy*, The Institution of Mining and Metallurgy, Portland Place, London 1984, p. 44.
10. L. D. Skrylio, L. M. Lopatenko, L. A. Sin'kova, *Sov. Non-Ferrous Met. Res. (Engl. Transl.)* **13** (1985) no. 3, 186 – 187.
11. V. Tafel: *Lehrbuch der Metallhüttenkunde*, 2nd ed., vol. **1**, Hirzel, Leipzig 1951, p. 590 ff.
12. F. Pawlek: *Metallhüttenkunde*, Berlin, Walter de Gruyter, 1983.
13. Lumalampan, DE 3 243 813, 1982 (A. Sikander).
14. Linden Chemical & Plastics, EP 0 042 509, 1982 (R. J. Burkett).
15. Elektro-Ofenbau Matthias Marcus, DE 3 609 517, 1986.
16. H. Kulander, *Sprechsaal* **119** (1986) no. 11, 1016 –1018.
17. F. Hiller: *Entsorgung von Gerätebatterien*, lecture held at the Technische Akademie Esslingen, März 1988.
18. N. Hirayama: *Behandlung von Sonderabfall 1*, EF-Verlag, Berlin 1987.
19. H. Pietsch, Bundesministerium für Forschung und Technologie, Forschungsprojekt FKZ: 143 03 553, Lurgi, Frankfurt 1986.
20. H. Laig-Hoerstebrock, Bundesministerium für Forschung und Technologie, Forschungsprojekt FKZ: 143 0273/0, Varta, Kelkheim 1986.
21. H. Kulander: *Behandlung von Sonderabfall 1*, EF-Verlag, Berlin 1987.
22. Preussag AG Metall, internal report, 1989, Goslar, Federal Republic of Germany.
23. Outokumpu O.Y., DE 2 406 119, 1973.
24. Preussag AG Metall, DE 2 558 115, 1975 (G. Heyer).
25. K. Hanusch, BMFT-Forschungsbericht NTS 0103/0, Nov. 1973, Harlingerode.
26. J. L. Gallant, *World Conf. Int. Nucl. Target Development Soc.*, Boston, Mass., 1 – 3 Oct. 1979, Plenum Press 8 (181) pp. 213 – 216. F. Desmet, L. Lemaître, A. P. van Petegham, *Mater. Chem. Phys.* **11** (1984) 305 – 309.
27. A. D. Pogorelyi, G. M. Tysh, *Tsvetn. Met.* 1979, no. 5, 10 – 18.
28. S. A. N. Sheya, J. H. Maysilles, R. G. Sandberg, *Rep. Invest. U.S. Bur. Mines* **9191** (1988) 1 – 12.
29. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances*, Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1973, p. 339.
30. Ch. G. Maier, *Am. Inst. Min. Metall. Eng. Techn. Publ.* no. 264 (1929).
31. R. Heindryckx et al., “Mercury and Cadmium in Belgian Aerosols,” in: *CEC 1974 Problems of the Contamination of Man and his Environment by Mercury and Cadmium,* Commission of the European Communities, Luxemburg, pp. 135 – 148.
32. NAS, *An Assessment of Mercury in the Environment*, National Academy of Sciences, Washington, D.C. 1978.
33. J. K. Piotrowski, D. O. Coleman: *MARC-Report 20, Environmental Hazards of Heavy Metals: Summary Evaluation of Lead, Cadmium and Mercury*, MARC-Monitoring and Assessment Research Center, GEMS Global Environmental Monitoring System, Genève, London 1980.
34. H. J. Rösler, H. Lange: *Geochemische Tabellen*, 2nd ed., Enke Verlag, Stuttgart 1976.
35. *Schwermetalle in Lebewesen und Böden*, ANS-Mitteilungen, Sonderheft 2 p. 18 ff.
36. P. Koronowski: *Nebenwirkung von Quecksilberverbindungen auf Mensch und Tier*, Berlin 1973.
37. S. Bombach, L. Peters, *Naturwissenschaften* **62** (1975) 575 – 576. Links
38. E. Heinrich, H. Paucke, H.-D. Nagel, D. Hansen: *Agrochemikalien in der Umwelt*, VEB Fischer Verlag, Jena 1976, p. 57.
39. R. D. Rogers, J. C. McFarlane, *J. Environ. Qual.* **8** (1979) no. 2, 255 – 260. Links
40. R. Taylor, T. Bogacka, M. Balcerska, *Environ. Prot. Eng.* **4** (1978) no. 2, 179 – 182.
41. J. Lag, E. Steiners, *Acta Agric. Scand.* **28** (1978) 393 –396.
42. R. W. Klusman, R. A. Landrers, *J. Volcanol. Geotherm. Res.* **5** (1979) 49 – 65.
43. R. Bargagli, F. P. Iosco, C. Barghigiani, *Water Air Soil Pollut.* **36** (1987) 219 – 225. Links
44. T. Stijve, R. Roschnik, *Mitt. Geb. Lebensmittelunters. Hyg.* **65** (1974) 208 – 220.
45. R. Seegers, *Z. Lebensm. Unters. Forsch.* **160** (1976) 303 – 312. Links
46. K. Aichberger, O. Horak, *Bodenkultur* **26** (1975) no. 1, 8 – 14. Links
47. W. Rauter, *Z. Lebensm. Unters. Forsch.* **159** (1975) 149 – 151. Links
48. E. Hauser, M. Mohadjerani, *Mittl. Geb. Lebensmittelunters. Hyg.* **67** (1976) 389 – 401.
49. M. Cember, M. Curtis, B. G. Blaylock, *Environ. Pollut.* **17** (1978) no. 4, 311 – 319.
50. A. Stock, *Z. Angew. Chem.* **41** (1928) 663.
51. *Blei-, Kadmium- und Quecksilbergehalte von Lebensmitteln in der Bundesrepublik Deutschland*, Bericht der Zentralen Erfassungs- und Bewertungsstelle für Umweltchemikalien im Bundesgesundheitsamt Berlin (ZEBS), Stand 1975, und ZEBS-Berichte 1/79.
52. *Arsen, Blei, Cadmium und Quecksilber in und auf Lebensmitteln*, ZEBS-Bericht, Berlin 1/1984.
53. *Quecksilber-, Magnesium- und Zinkgehalte in der Frauenmilch, im Blutserum und Fettgewebe der Mütter*, ZEBS-Hefte, Berlin 1/ 1986.
54. A. Kuivala, J. Poijärvi, *Erzmetall* **30** (1977) 556.
55. Boliden AB, EP 179 040, 1985 (F. Dyvik).
56. F. Dyvik: “Extraction Metallurgy 85,” *Proc. Conf.*, London 1985, The Institution of Mining and Metallurgy, London,

pp. 189 – 198.
57. Norddeutsche Affinerie AG, EP 278 537, 1988 (H. Winkler, C. Reppenhagen).
58. H. Braun, M. Metzger, H. Vogg, *Müll. Abfall* **18** (1986) no. 2, 62 – 71, 89 – 95.
59. Kernforschungszentrum Karlsruhe, DE 3 715 046, 1987 (H. Vogg, H. Braun, M. Metzger, A. Merz).
60. Kernforschungszentrum Karlsruhe, EP 0 289 810, 1988 (H. Vogg, H. Braun, M. Metzger, A. Merz).
61. Kernforschungszentrum Karlsruhe, EP 0 289 809, 1988 (H. Vogg, H. Braun).
62. Bergwerksverband, DE 3 715 526, 1987 (K. Knoblauch, K. Wybrands, K.-D. Henning, J. Degel, H. Ruppert).
63. D. Reimann, *Umweltmagazin*, Oct. 1984, 48 – 54.
64. VEB Chemieanlagenbau- und Montagekombinat Leipzig, DD 139 068, 1976 (H. Tischendorf, R. Boege, G. Kreutzberger).
65. M. D. Rosenzweig, *Chem. Eng. (London)* **82** (1975, Jan. 20) 60 – 61; *Chem. Eng. (London)* **82** (1975, Feb. 3) 36 – 37.
66. Chlorine Inst.: *Water Pollution Aspects*, 5. 3. 1971.
67. K.-H. Bergk, F. Wolf, S. Eckert, *Z. Chem.* **17** (1977) 85 – 89.
68. G. van der Heiden, C. M. S. Raats, M. F. Boon, *Chem. Ing. Tech.* **51** (1979) 631 – 653.
69. Kernforschungszentrum Karlsruhe, DE 3 721 141, 1987 (M. Metzger, H. Braun).
70. D. O. Reimann, *VGB Kraftwerkstechnik* **64** (1984) no. 3, 230 – 235.
71. "Umwelt und Degussa, TMT-15 für die Abtrennung von Schwermetallen aus Abwässern," *Degussa Broschüre*, 1982.
72. Wacker-Chemie, DE 3 335 127, 1983 (J.-H. Janssen, B. Bangler, K.-H. Fahrmeier).
73. Hoechst AG, EP 0 091 043, 1983 (H. von Plessen, R. Gradl, G. Schimmel).
74. R. Kola, *Erzmetall* **30** (1977) no. 12, 559 – 561.
75. R. K. Williardson, A. C. Beer, *Semicond. Semimetals* **18** (1981) 21 – 45.
76. Montedison S.p.A., EP 0 148 023, 1984 (M. Gramondo, G. Donati, G. Faita, G. L. Marziano).
77. *Chem. Eng. (London)* **82** (1975) no. 2, 60.
78. *Chem. Eng. (London)* **82** (1975) no. 3, 36.
79. Chemikerausschuß der Ges. Dtsch. Metallhütten- und Bergleute: *Analyse der Metalle*, vol. **2**, Springer Verlag, Berlin – Göttingen – Heidelberg 1953, p. 590.
80. *Z. Anal. Chem.* **96** (1934) 30.
81. *Z. Anal. Chem.* **89** (1932) 187.
82. *Z. Anal. Chem.* **98** (1934) 331.
83. R. Dumarey, R. Heindryckx, R. Dams, J. Hoste, *Anal. Chim. Acta* **107** (1979) 159 – 167. Links
84. G. R. Carnrick, W. Barnett, W. Slavin, *Spectrochim. Acta Part B*, **41 B** (1986) no. 9, 991 – 997.
85. L. Sipos, J. Golimowski, P. Valenta, H. W. Nürnberg, *Fresenius Z. Anal. Chem.* **298** (1979) no. 1, 1 – 8.
86. H. Bloom, B. Noller: *Trends in Electrochemistry*, Badford, Australia 1976, 241 – 252.
87. H. Dorias: *Gefährliche Güter, Eigenschaften, Handhabung, Lagerung und Beförderung*, Springer Verlag, Berlin 1984.
88. G. Hommel: *Handbuch der gefährlichen Güter*, Springer Verlag, Berlin 1985.
89. J. M. Lo, C. M. Wai, *Anal. Chem.* **47** (1975) no. 11, 1869 – 1870. Links
90. S. Dogan, W. Herdi, *Anal. Chim. Acta* **101** (1978) no. 2, 433 – 436. Links
91. T. Suzuki, M. Fujita, K. Iwashima, *Eisei Kagaku* **26** (1980) no. 5, 229 – 235.
92. J. C. Meranger, B. R. Hollebone, G. A. Blanchette, *J. Anal. Toxicol.* **5** (1981) no. 1, 33 – 41. Links
93. J. R. Kechtel, *Analyst (London)* **105** (1980) 826 – 829.
94. R. W. Heiden, D. A. Aikens, *Anal. Chem.* **51** (1979) no. 1, 151 – 156. Links
95. P. R. Ludlam, J. G. King, *Analyst (London)* **106** (1981) 488 – 489.
96. R. Schaaf, *Müll Abfall* **11** (1983) 277 – 283.
97. G. Jangg, H. Palman, *Z. Metallk.* **54** (1963) 364.
98. G. Brauer, G. Düsing, *Z. Anorg. Allg. Chemie* **328** (1964) 154.
99. G. Jangg, *Metall (Berlin)* **13** (1959) 407.
100. C. Guminski, *Z. Metallk.* **77** (1986) no. 2, 87 – 96.
101. N. P. Chopey, *Chem. Eng. (N.Y.)* **68** (1961) no. 25, 120.
102. A. Giordano, US 3 061 412, 1960.
103. Du Pont, US 2 757 070, 1956.
104. Du Pont, US 2 904 398, 1959.
105. R. A. Feldhoff, *Pharm. Ztg.* **75** (1930) 11.
106. R. Hirayama, JP 174 101, 1946.
107. Wood-Ridge Chem. Co., US 3 424 552, 1967.
108. Y. V. Karyankin, J. J. Angelov: Chisty Klimicheskie Reativy (pure chemical reagents), state-owned scientific-technical publisher for chemical literature, Moskva 1955, p. 457.
109. W. Leithe, *Vom Wasser* **37** (1970) 106.
110. Merck-Standards, pp. 807 – 830, 1971, Darmstadt, Federal Republic of Germany.
111. J. Dahmen, *Proc. 6th Indo-Ger. Semin. on Trace Element Analysis – Methods and Selected Applications*, Oct. 1978 (Maria Laach).

112. L. G. Makarova, A. N. Nesemeyanow: *The Organic Compounds of Mercury*, North-Holland Publ. Co., Amsterdam 1967.
113. *Beilstein*, **E IV (4)**, 4426 – 4463; **E IV (16)**, 1701 –1718.
114. C. Larock, *Angew. Chem.* **90** (1978) 28 – 38. Links
115. C. Larock: *Organo Mercury Compounds in Organic Synthesis*, Springer Verlag, Berlin 1985.
116. W. Peters, *Ber. Dtsch. Chem. Ges.* **38** (1905) 2567. Bayer, DE 1 003 733, 1955.
117. J. B. Johnson, J. P. Fletcher, *Anal. Chem.* **31** (1959) 1563. Links
118. D. Steinborn, U. Sedlak, *Z. Chem.* **25** (1985) 376 ff.
119. D. Seyferth, *Acc. Chem. Res.* **5** (1972) 65. Links
120. R. Scheffold, U. Michel, *Angew. Chem.* **84** (1972) 160. Links
121. B. Giese, *Angew. Chem.* **97** (1985) 555 – 567. Links
122. H. C. Brown, P. Geoghean, *J. Am. Chem. Soc.* **89** (1967) 1522; *J. Org. Chem.* **35** (1970) 1844. Links
123. Metallstatistik 1987 der Metallgesellschaft AG, Frankfurt/Main 1987.
124. *Ullmann*, 4th ed., **19**, 654.
125. United States Department of the Interior: *Minerals Yearbook 1986*, vol. **1**, Bureau of Mines, US Government Printing Office, Washington 1988, p. 659 ff.
126. A. Rauhut, *Metall (Berlin)* **42** (1988) no. 11, 1137 –1141.
127. G. Schwarzenbach, M. Schellenberg, *Helv. Chim. Acta* **48** (1965) 28.
128. D. L. Rabenstein, *Acc. Chem. Res.* **11** (1978) 100. Links
129. W. L. Hughes, *Cold Spring Harbour symposia on quantitative biology* **14** (1950) 79 – 84.
130. J. L. Webb: *Enzyme and Metabolic Inhibitors*, vol. **2**, Academic Press, New York 1966.
131. W. L. Hughes, *Ann. N. Y. Acad. Sci.* **65** (1957) 454. Links
132. S. Halbach, *Arch. Toxicol.* **64** (1990) 315. Links
133. J. B. Hursh, *J. Appl. Toxicol.* **5** (1985) 327. Links
134. WHO (1990). Methylmercury, Environmental Health Criteria 101, WHO, Geneva.
135. S. Halbach, T. W. Clarkson, *Biochim. Biophys. Acta* **523** (1978) 522. Links
136. L. Magos, S. Halbach, T. W. Clarkson, *Biochem. Pharmacol.* **27** (1978) 1373. Links
137. G.-G. Elinder, L. Gerhardsson, G. Oberdoerster, in T. W. Clarkson, L. Friberg, G. F. Nordberg, P. R. Sager (eds.): *Biological Monitoring of Toxic Metals*, Plenum Press, New York 1988, p. 1.
138. WHO (1991). Inorganic Mercury. Environmental Health Criteria 118, WHO, Geneva.
139. R. D. Barr *et al.*, *Brit. Med. J.* **2** (1972) 131. Links
140. A. Khayat, L. Dencker, *Acta Pharmacol. Toxicol.* **55** (1984) 145.
141. M. G. Cherian, T. W. Clarkson, *Chem. Biol. Interact.* **12** (1976) 109. Links
142. F. Stoz *et al.*, *Z. Geburtsh. u. Neonat.* **199** (1995) 35 – 41.
143. T. W. Clarkson, J. B. Hursh, P. R. Sager, T. L. M. Syversen, in T. W. Clarkson, L. Friberg, G. F. Nordberg, P. R. Sager (eds.): *Biological Monitoring of Toxic Metals*, Plenum Press, New York 1988, p. 199.
144. J. Dunn, T. W. Clarkson, L. Magos, *Brit. J. Ind. Med.* **35** (1978) 214.
145. T. Rahola, T. Hattula, A. Korolainen, J. K. Miettinen, *Ann. Clin. Res.* **5** (1973) 214. Links
146. M. G. Cherian, J. B. Hursh, T. W. Clarkson, J. Allen, *Arch. Environ. Health* **33** (1978) 109. Links
147. J. B. Hursh *et al.*, *J. Pharmacol. Exp. Ther.* **214** (1980) no. 3, 520. Links
148. J. B. Hursh *et al.*, *Arch. Environ. Health* **31** (1976) 302. Links
149. L. Kosta, A. R. Byrne, V. Zelenko, *Nature, London* **254** (1975) 238.
150. WHO (1976). Mercury, Environmental Health Criteria 1, WHO, Geneva.
151. T. G. Kershaw, T. W. Clarkson, P. H. Dhahir, *Arch. Environ. Health* **35** (1980) no. 1, 28. Links
152. M. Berlin, in L. Fridberg, G. F. Nordberg, V. Voulk (eds.): *Handbook on the Toxicology of Metals*, Elsevier, Amsterdam 1986, p. 387.
153. I. Suda, H. Takahashi, *Toxicol. Appl. Pharmacol.* **82** (1986) 45. Links
154. I. R. Rowland, M. J. Davies, J. G. Evans, *Arch. Environ. Health* **35** (1980) 155. Links
155. T. Refsvik, T. Norseth, *Acta Pharmacol. Toxicol.* **35** (1975) 67.
156. R. B. McFarland, H. Reigel, *J. Occup. Med.* **20** (1978) 532. Links
157. M. Levin, J. Jacobs, P. G. Polos, *Chest* **94** (1988) 554. Links
158. P. Troen, S. A. Kaufman, K. H. Katz, *New Eng. J. Med.* **244** (1951) 459. Links
159. P. L. Bidstrup, J. A. Bonnel, D. G. Harvey, S. Locket, *Lancet* **2** (1951) 856. Links
160. W. Von Glantzel *et al.*, *Z. Gesamte Inn. Med. Ihre Grenzgeb.* **35** (1980) 267.
161. P. Druet, in T. Suzuki, N. Imura, T. W. Clarkson (eds.): *Advances in Mercury Toxicology*, Plenum Press, New York 1991, p. 395.
162. V. Foa *et al.*, *Int. Arch. Occup. Environ. Health* **37** (1976) 115. Links
163. J. Chmielnicka, E. Brzeznicka, A. Sniady, *Arch. Toxicol.* **59** (1986) 16. Links
164. R. G. Smith, A. J. Vorwald, L. S. Patil, T. F. Mooney, *Am. Ind. Hyg. Assoc. J.* **31** (1970) 687. Links
165. G. D. Langolf, D. B. Chaffin, R. Henderson, H. P. Whittle, *Am. Ind. Hyg. Assoc. J.* **39** (1978) 976. Links

166. J. W. Albers *et al.*, *Ann. Neurol.* **24** (1988) 651 – 659. Links
167. M. Forzi, M. G. Cassitto, C. Bulgheroni, V. Foa: *Adverse Effects of Environmental Chemicals and Psychotoxic Drugs*, vol. **2**, Elsevier, Amsterdam 1976, p. 165.
168. H. Roels, S. Abdeladim, E. Ceulemans, R. Lauwerys, *Ann. Occup. Hyg.* **31** (1987) 135. Links
169. R. F. Fawer *et al.*, *Br. J. Ind. Med.* **40** (1983) 204. Links
170. F. Schuckmann, *Int. Arch. Occup. Environ. Health* **44** (1979) 193. Links
171. J. P. Buchet, H. Roels, A. Bernard, Jr. R. Lauwerys, *J. Occup. Med.* **22** (1980) no. 11, 741. Links
172. R. Lauwerys *et al.*, *Toxicol. Lett.* **17** (1983) 113. Links
173. North American Contact Dermatitis Group, *Arch. Dermatol.* **108** (1973) 537. Links
174. D. Herrmann, in L. T. Fridberg, G. N. Schrauzer (eds.): *Status quo and perspectives of amalgam and other dental materials*, Thieme Vlg., Stuttgart 1995, p. 51.
175. P. D. Logan, *Lancet* **1** (1949) 608. Links
176. W. W. Tunnessen, K. J. McMahon, M. Baser, *Pediatrics* **79** (1987) no. 5, 786. Links
177. K. E. Mühlendahl, *Lancet* **336** (1990) 1578. Links
178. D. Hunter, R. R. Bomford, D. S. Russell, *Quarterly Journal of Medicine* **9** (1940) 193.
179. T. Tsubaki, K. Irukayama, (1977). Minamata Disease Elsevier Science Publishers, Amsterdam.
180. A. Igata, in T. Suzuki, N. Imura, T. W. Clarkson (eds.): *Advances in Mercury Toxicology*, Plenum Press, New York 1991, p. 439.
181. F. Bakir *et al.*, *Science* **181** (1973) 230. Links
182. W. D. Atchison, *J. Pharmacol. Exp. Ther.* **237** (1986) no. 2, 672. Links
183. T. L. M. Syversen, *Toxicol. Lett.* **10** (1982) 31. Links
184. P. H. Fair, J. E. Balthrop, J. L. Wade, S. Braddon-Galloway, *Toxicol. Lett.* **36** (1987) 213. Links
185. S. Omata *et al.*, in T. Suzuki, N. Imura, T. W. Clarkson (eds.): *Advances in Mercury Toxicology*, Plenum Press, New York 1991, p. 223.
186. M. A. Verity, T. Sarafian, in T. Suzuki, N. Imura, T. W. Clarkson (eds.): *Advances in Mercury Toxicology*, Plenum Press, New York 1991, p. 209.
187. D. W. Nierenberg *et al.*, *New England J. Med.* **338** (1998) no. 23, 1672 – 1676. Links
188. L. Amin-Zaki *et al.*, *Am. J. Dis. Child.* **130** (1976) 1070. Links
189. D. O. Marsh *et al.*, *Arch. Neurol.* **44** (1987) 1017. Links
190. P. Grandjean *et al.*, *Neurotoxicol. Teratol.* **19** (1997) 417 – 428. Links
191. P. W. Davidson *et al.*, *Neuro Toxicology* **16** (1995) no. 4, 677 – 688.
192. B. H. Choi, in T. Suzuki, N. Imura, T. W. Clarkson (eds.): *Advances in Mercury Toxicology*, Plenum Press, New York 1991, p. 315.
193. K. R. Snapp, D. B. Boyer, L. C. Peterson, C. W. Svare, *J. Dent. Res.* **68** (1989) 780. Links
194. D. Zander *et al.*, *Zbl. Hyg.* **190** (1990) 325.
195. S. Halbach, *Envir. Res.* **77** (1998) 115 – 123.
196. T. W. Clarkson, L. Fridberg, J. B. Hursh, M. Nylander, in T. W. Clarkson, L. Fridberg, G. F. Nordberg, P. R. Sager (eds.): *Biological Monitoring of Toxic Metals*, Plenum Press, New York 1988, p. 247.
197. WHO (1972). Evaluation of mercury, lead, cadmium and the food additives, Food additives series No.4, WHO, Geneva.
198. A. Berglund, *J. Dent. Res.* **69** (1990) 1646 – 1651. Links
199. S. Halbach, *J. Dent. Res.* **74** (1995) no. 4, 1103 – 1109. Links
200. IARC (1993). Beryllium, cadmium, mercury, and exposures in the glass manufacturing indistry. IARC monographs on the evaluation of carcinogenic risks to humans. International Agency for Research on Cancer, Lyon.
201. C. Grasmick, G. Huel, *Sci. Tot. Environ.* **44** (1985) 101.
202. DFG, Deutsche Forschungsgemeinschaft: *MAK- und BAT-Werte-Liste 1998*, Wiley-VCH, Weinheim, Germany 1998.
203. ACGIH (1997). 1997 threshold limit values for chemical substances and physical agents and biological exposure indices. American Conference of Governmental Industrial Hygienists, Cincinnati, OH, USA.
204. ATSDR, (1999). Toxicological profile for mercury. Agency of toxic substances and disease registry. Atlanta, GA, USA.
205. WHO (1987) Air quality guidelines for Europe, p. 272–284.
206. WHO (1989). Evaluation of certain food additives and contaminants. Tech. Rep. Series, nr. 776, WHO, Geneva.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Metal-Complex Dyes

Klaus Grychtol, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 5)

Winfried Mennicke, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 6)

1. Introduction and History 1
2. Structure of Metal-Complex Dyes 1
2.1. Chelates 1
2.2. Stereochemistry and Isomerism . . 3
3. Production 4
4. Ecology and Toxicology 7
5. Uses of Metal-Complex Dyes 7
5.1. Textile Dyeing 7
5.1.1. Wool and Polyamides 7
5.1.2. Cotton 14
5.1.3. Polypropylene 15
5.1.4. Dust-Free Preparations 16
5.2. Nontextile Uses 16
5.2.1. Leather 16
5.2.2. Paper 17
5.2.3. Dyes for Anodized Aluminum 19
5.2.4. Solvent and Ink-Jet Dyes 19
5.2.5. Pigments 20
5.2.6. Redox Catalysts 23
5.2.7. Analysis 23
5.2.8. Photo Dyes 24
5.2.9. Charge Stabilizers 24
5.2.10. Optical Applications 25
6. Formazan Dyes 25
6.1. Physical Properties and Structure 26
6.1.1. Metal-Free Formazans 26
6.1.2. Metal-Containing Formazans 27
6.2. Production 30
6.2.1. Coupling of Diazonium Compounds with Hydrazones 30
6.2.2. Metallization 30
6.2.3. Double Coupling of Diazonium Compounds with Compounds Containing an Acidic CH Group 31
6.2.4. Oxidative Coupling of Azole Hydrazines with Hydrazones 33
6.3. Uses 34
7. References 36

1. Introduction and History

The use of metal complexes as dyes has its origins in *mordant dyeing*, a process that can be traced back to the Middle Ages. In mordant dyeing, textile dyes or the dyed fabrics themselves are treated with specific metal salts, which leads to both a bathochromic shift in color and a distinct dullness. Resistance to the effect of light or washing is also greatly improved. Metal compounds commonly employed earlier include aluminum, chromium, and iron salts.

Important events in the historical development of metal-complex dyes can be summarized as follows [5], [12]:

1887 Synthesis of the first metallizable azo dye, alizarin yellow, from diazotized 4-nitroaniline and salicylic acid (R. Nietzki)
1891 Formulation of the theory of coordination (A. Werner)
1893 Synthesis of the first metallizable 2,2′-dihydroxyazo dye (E. Bergmann and O. Borgmann)
1908 Elucidation of the stereochemistry of metal complexes (A. Werner)
1912 Synthesis of the first premetallized 1 : 1 chromium complex (R. Bohn, BASF)
1915 Patenting of a strong-acid dyeing process for 1 : 1 chromium complexes (Ciba)
1920 Marketing of 1 : 1 chromium complexes: the Neolan series (Ciba) and the Palatin Fast series (BASF)
1927 First synthesis of copper phthalocyanine (H. de Diesbach and E. von der Weid)
1949 Development of the first 1 : 2 chromium complexes lacking sulfo groups: Irgalan dyes (G. Schetty, Geigy)
1954 Marketing of Neopalatin dyes (BASF)
1962 Introduction of Lanacron dyes as homogeneous 1 : 2 chromium complexes containing a sulfo group (Ciba)
1970 Acidol M dyes (BASF) sold as 1 : 2 chromium complexes with two sulfo groups

2. Structure of Metal-Complex Dyes

2.1. Chelates

Complexes or *coordination compounds* are systems in which a metal ion (the central atom) is chemically linked with one or more electron donors (ligands). A ligand with several donor groups capable of coordination leads to cyclic coordination compounds, which are designated as *internal complexes* or *chelates*. Such polydentate ligands greatly increase the stability of a complex relative to similar monofunctional

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_299

nucleophiles. The resulting increase in the tendency to form complexes, known as the *chelate effect*, was discovered by SCHWARZENBACH [14] and is discussed in [13, pp. 121 – 122].

The stability of metal chelates is described extensively in [9]. Generally speaking, stability is related to the formation of strain-free five- and six-membered rings, to the number of chelate rings per ligand, to the basicity of the ligands, and to the nature of the metal ion.

A characteristic feature of metal complexes is that the coordination number of the metal ion is greater than its valence. Typical coordination numbers observed in metal-complex dyes are 4 (most divalent ions) and 6 (in most trivalent metal ions). Chelation leads to changes in many physical and chemical properties, including absorption spectra, stability, charge, and solubility. These properties are exploited in dye chelates to modify color tones, the affinity of a dye for particular substrates, and colorfastness.

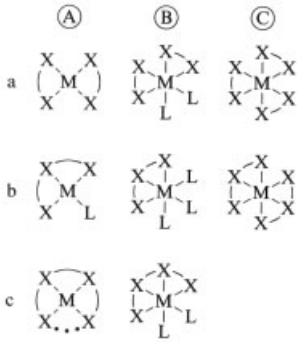

Figure 1. Types of chelate complexes
M = central metal atom; L = monofunctional ligand; X–X–X = chelate-forming ligand

Figure 1 illustrates the most important chelate types. Compounds in column A display a metal ion coordination number of 4, whereas those in columns B and C have the coordination number 6. Rows a, b, and c are assigned to bi-, tri-, and tetrafunctional ligands, respectively. Complex type Ac represents a tricyclic species capable of undergoing an additional ring closure to form a tetracyclic annulene. Metal ions with a coordination number of 4, such as copper(II), form 1 : 1 complexes with tridentate ligands, the fourth coordination site being occupied by another nucleophile (e.g., water, ammonia, or pyridine). Metal ions with a coordination number of 6, including chromium(III), combine with tridentate ligands to give 1 : 1 and 1 : 2 complexes (one metal ion joined to one or two dye molecules). In a 1 : 1 chromium complex, the three vacant coordination sites are usually occupied by water molecules because synthesis is performed in an aqueous medium.

Tridentate ligands are the most significant from the standpoint of applications. Metals having a coordination number of 4 usually give square planar complexes, whereas with a coordination number of 6 the ligands are arranged octahedrally around the central atom. Figure 2 shows the most important types of tridentate ligands from an industrial standpoint.

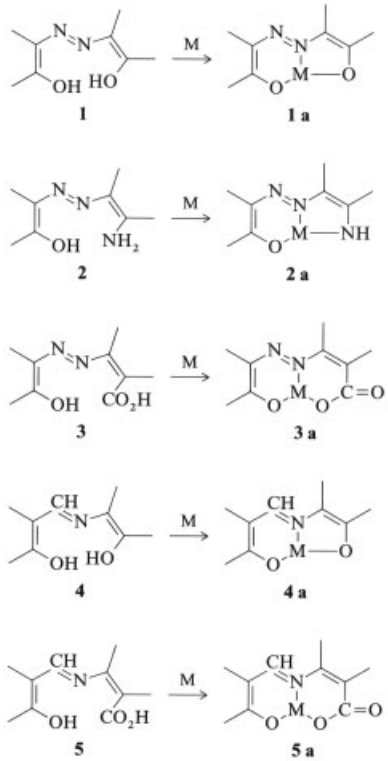

Figure 2. Tridentate ligands and their metal complexes
M = Cu^{2+}, Cr^{3+}, etc.

After metallization, dyes derived from 2,2′-dihydroxyazo (**1**), 2,2′-hydroxyaminoazo (**2**), and 2,2′-dihydroxyazomethine (**4**) ligands give the condensed five- and six-membered cyclic systems **1a**, **2a**, and **4a**, respectively, whereas 2,2′-hydroxycarboxyazo (**3**) and 2,2′-hydroxycarboxyazomethine (**5**) moieties lead to **3a** and **5a**, with two six-membered rings.

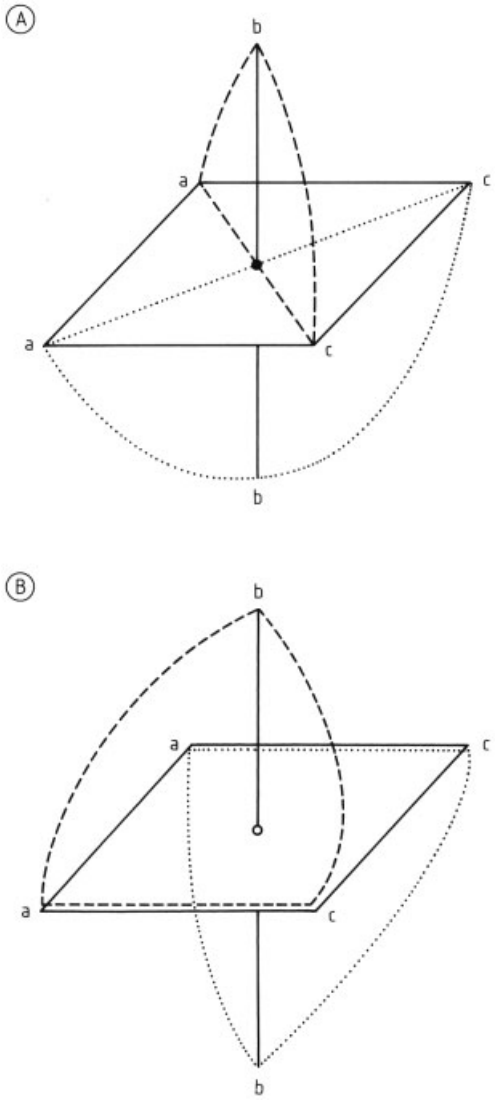

Figure 3. Geometric arrangements exhibited by octahedral complexes
A) Drew – Pfitzner structure, meridian arrangement with two tridentate ligands perpendicular to one another, one pair of enantiomers
B) Pfeiffer – Schetty structure, facial arrangement with two tridentate ligands parallel to one another, four pairs of enantiomers, one centrally symmetric isomer

Systems with larger numbers of ring atoms are extremely rare. One example described in the patent literature [15] is the 1 : 2 chromium complex of the dye formed by coupling 1-amino-2-phenylacetic acid with phenylmethylpyrazolone, resulting in one seven-membered and one six-membered ring. They are of no technical interest.

The "ligands" shown in Figure 2 are only fragments; they form part of benzene and naphthalene rings, pyrazolones, and open-chain keto – enol systems. The bonds joining metal atoms to nitrogen from the azo or azomethine groups are sometimes called "secondary valences", and may be characterized by arrows or dotted lines [13].

2.2. Stereochemistry and Isomerism

Three types of isomerism are associated with metal complexes bearing trifunctional ligands. The first arises from the fact that the ligands may occupy geometrically distinct positions around the central atom; the second is due to the possibility that either of the two nitrogen atoms in a given azo group can bind to the metal; finally, the complexing nitrogen in the hydrazone form of an azo dye is formally sp^3 hybridized, which produces two conformational possibilities.

The first type of isomerism is illustrated by octahedral 1 : 2 chromium and cobalt complexes, in which tridentate ligands may be arranged either perpendicular (Fig. 3 A) or parallel (Fig. 3 B) to each other. This results in eleven possible structural isomers. The meridial (*mer*) arrangement gives a pair of enantiomers [16], [17]. The resolution of these enantiomers into optically active forms was demonstrated by P. Pfeiffer and S. Saure [18]. Such "Drew – Pfitzner structures" are present in 2,2′-dihydroxyazo and azomethine dyes (5/6-membered rings). The perpendicular arrangement of the ligands has been verified by X-ray structural analysis [19].

The facial (*fac*) or sandwich orientation (Fig. 3 B) of 6/6-ring chelates is a consequence of steric factors [20]. Here, four enantiomeric pairs are possible, together with one centrally symmetric isomer. In the case of the 1 : 2 chromium complex of the azo dye derived from 1-(2-carboxyphenylazo)-2-naphthol, four isomers have been isolated by chromatography [21]. All four have the same molecular formula, but their spectra and solubilities differ. In solution, each of the isomers undergoes equilibration to regenerate the original mixture. Once again, the pro-

posed structure of one of these isomers has been confirmed by X-ray structural analysis [22].

With asymmetric 1 : 1 and 1 : 2 metal complexes a second type of isomerism becomes possible because different nitrogen atoms from the tricyclic ligands can be bound to the central atom in different ways. Two isomers may be distinguished in a 1 : 1 complex [23] and three in a 1 : 2 complex (Fig. 4) [24]. These isomers have been detected by NMR spectroscopy [25].

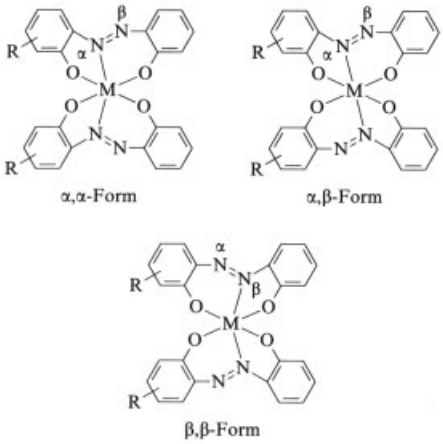

Figure 4. N-α,β-Isomers of 1 : 2 metal complexes $M = Co^{3+}$, Cr^{3+}

Such N-α,β-isomerism is precluded in complexes of the 1 : 1 and 1 : 2 types if the third position ortho to the azo function is occupied by methyl or methylene groups [26]. Thin layer chromatographic separation of N-α,β-isomers has been described [27].

The third type of isomerism is found in 1 : 2 complexes with 5/6-membered rings and is attributed to the tautomeric hydrazone structures of the azo ligands, in which an sp^3-hybridized nitrogen atom serves as the electron donor. Three conformationally distinct complexes arise [28]. A thorough study of this situation utilizing modern spectroscopic methods is reported in [29]. Figure 5 illustrates the tautomers containing only one 1-hydroxyl group, which can be used to synthesize cobalt complexes.

The fact that numerous isomeric complexes coexist in equilibrium means that the resulting colors tend to be rather dull. In contrast, planar dyes such as the copper phthalocyanines and copper formazans are brilliant in color.

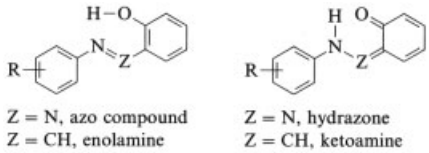

Figure 5. Tautomerism in azo and azomethine compounds

3. Production

The chief synthetic methods for metal complexes are summarized in Table 1. Most 1 : 1 nickel and copper complexes are prepared from the corresponding salts at pH 4 – 7 in the presence of bases such as sodium acetate and ammonia between 25 and 70 °C. *Dealkylative coppering* is the method of choice for water-soluble polyazo compounds. This may involve initial formation of a 1 : 1 copper complex in which the ether group is already coordinated to copper. Reaction is carried out in the presence of buffers such as sodium acetate, alkanolamines, or pyridine at elevated temperature (> 80 °C). Nucleophilic substitution is facilitated by the neighboring azo group, and if X = H (Table 1), simultaneous oxidation is required (*oxidative coppering* [30]). Oxidative degradation of the original dye is a danger in this reaction, so very mild conditions are required. The 2,2′-dihydroxy copper complexes are formed from the corresponding 2-halogen compounds in strongly alkaline medium [31]. 2-Sulfonic acids are also susceptible to nucleophilic exchange.

Dealkylative and oxidative coppering represent technical processes, because few direct methods are known for preparing azo metal complexes with 2-hydroxy substituents on both sides of the azo group. This is a consequence of the limited availability of 2-aminophenols and the difficulty of coupling the corresponding diazonium salts (2-quinone diazides). Copper complexes prepared in this way can subsequently be converted to metal-free dyes and then to the corresponding chromium complexes [32].

Iron complexes are prepared under weakly acidic conditions at slightly elevated temperature (40 – 80 °C).

The 1 : 1 cobalt(III) complexes cannot be prepared in acidic solution. Nevertheless, the greater stability of cobalt(III) amine complexes [9], [33] can be exploited to obtain 1 : 1 cobalt complexes with three NH_3 ligands [34] or

Table 1. Synthetic methods for metal complexes

Z = N,CH

X = H, Cl, SO_3H, OH, OCH_3 Y = O

X = CO_2H Y = CO_2

X = NH_2 Y = NH

Reagents	X	pH	Temperature, °C
Cu–Ni 1:1 complexes			
$CuSO_4$, NH_3, CH_3COONa	OH, CO_2H	4–6.5	25–70
$CuSO_4$, $NH_3/H_2NCH_2CH_2OH$	OCH_3	> 7	> 80
$CuSO_4$, CH_3COONa/H_2O_2	H	4.5–8	25–50
Fe 1:2 complexes			
$FeSO_4$, CH_3COONa	OH	5	40–80
Co 1:1, 1:2 complexes			
$CoCl_2$, NaOH, air	OH	5–8	40–80
$CoCl_2$, NaOH, H_2O_2	OH	5–6	40–70
$CoCl_2$, NH_3, glycols	OCH_3	> 7	80–120
Cr 1:1 complexes			
Cr^{3+} salt, acid, H_2O, or glycols	OH, CO_2H NH_2, OCH_3	< 3	100–140
Cr 1:2 complexes			
Cr^{3+} salt, NaOH	OH, CO_2H, NH_2	4–8	80–100
Cr^{3+} salt, tartaric acid, salicylic acid, NaOH		7–9	80–100
$Na_2Cr_2O_7$, glucose, NaOH		7–9	80–100
Cr^{3+} salt, amides, glycols		7–9	100–140
$Na_2Cr_2O_7$, glycols, NaOH	OCH_3	> 7	100–130

the even more stable complexes with ethylenediamine, triethylenediamine, or phenanthroline [35]. Nitrite as a strong nucleophile also produces stable complexes [36].

The metal ion in 1 : 2 cobalt complexes is also trivalent, but such dyes are manufactured from the more stable cobalt(II) salts. Atmospheric oxygen serves as the oxidant, and reaction conditions are very mild: 40 – 80 °C and near-neutral pH. If cobaltization proceeds slowly, the metal-free dye may suffer degradation because it can act as an oxidant. The result is a decrease in yield, but important characteristics of the dye may also be adversely affected. The usual remedy is dropwise addition of dilute hydrogen peroxide along with the cobalt(II) chloride solution. Demethylative cobaltization requires temperatures > 100 °C, achieved by carrying out the reaction in high-boiling glycols. Base (e.g., ammonia, alkanolamines, or sodium hydroxide) must also be added.

The 1 : 1 chromium complexes containing sulfonic acid groups are prepared below pH 3 at 100 °C (disulfo derivatives) and up to 140 °C (monosulfonated compounds). A reaction temperature > 100 °C is achieved either in water under pressure or in glycols. The 1 : 1 chromium dyes are isolated from acidic solutions as their slightly soluble betaines (**7**) [5].

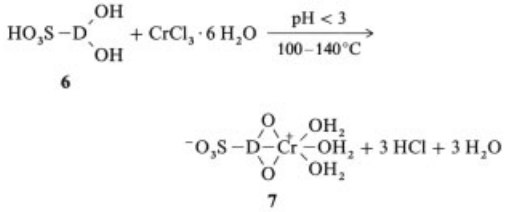

D = azo or azomethine residue

Irrespective of the number of sulfo groups, temperatures > 100 °C are always required in the demethylative incorporation of chromium. The usual source of chromium is chromium(III)

oxide (obtained as a byproduct of dichromate oxidation), which is first dissolved in hydrochloric, sulfuric, or formic acid.

Preparation of 1 : 2 chromium complexes generally requires higher reaction temperatures than the corresponding 1 : 2 cobalt compounds because the hexa-aquo chromium ion present in solution is very stable. This subject has been discussed in [9]. Chromium incorporation is usually conducted in near-neutral solution but can be accelerated by the addition of base. An anionic precomplex with tartaric, oxalic, or salicylic acid is used to prevent the precipitation of chromium hydroxide [37]. The 1 : 2 chromium complex with salicylic acid (**8**) is particularly effective.

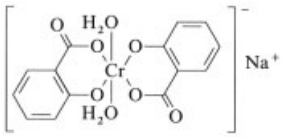

8 [*68214-27-7*]

An alternative method of preparing 1 : 2 chromium complexes under milder aqueous conditions is based on the use of an alkaline dichromate – glucose solution, in which a very reactive chromium(III) accessory complex forms as an intermediate [38]. The process is most important in the synthesis of 1 : 2 chromium complexes containing sulfonamide groups. Special precautions are necessary because of the toxicological risk of handling solid dichromate.

The introduction of alcohols [39] and carboxamides [40] as solvents for azo dyes that are only slightly soluble in water represented a technical advance. Formamide is particularly important because it promotes formation of 1 : 2 chromium complexes by buffering the acid that is released. The slightly soluble complexes are isolated by dilution with water.

Introduction of chromium by demethylation requires the use of sodium dichromate in glycol, which serves as both a high-boiling solvent and a reductant.

Two routes are used for preparing 1 : 2 metal complexes: the addition method and mixed metallization. The addition method leads first to a 1 : 1 chromium complex (**7**), which is isolated and subsequently reacted with the metal-free dye **6** to give the symmetric 1 : 2 complex **9**. Use in the second step of a dye with a different structure (**10**) but still capable of forming complexes, provides a route to uniform asymmetric 1 : 2 complexes (**11**).

For economic reasons, large-scale synthesis of symmetric 1 : 2 complexes is usually carried

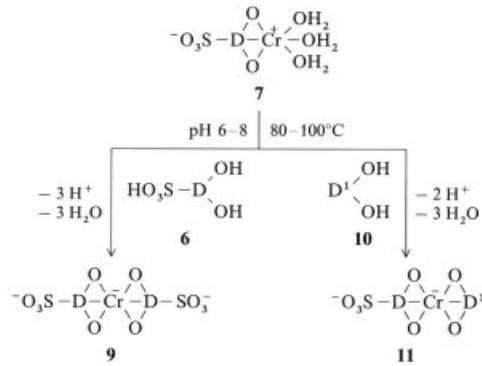

out without isolation of the intermediate, but synthesis of asymmetric 1 : 2 complexes (e.g., those with only a single sulfonic acid group) is necessarily a stepwise process.

Mixed metallization is particularly important in the synthesis of 1 : 2 cobalt complexes because of the inaccessibility of the corresponding 1 : 1 materials [41].

Two equivalents of an azo or azomethine dye are reacted with one equivalent each of a Cr(III) or a Co(II) salt to give a single symmetric 1 : 2 metal complex. If the reaction is conducted with two different dyes, both capable of forming complexes, a statistical mixture of 1 : 2 metal complexes arises, i.e., two symmetric complexes and one asymmetric complex:

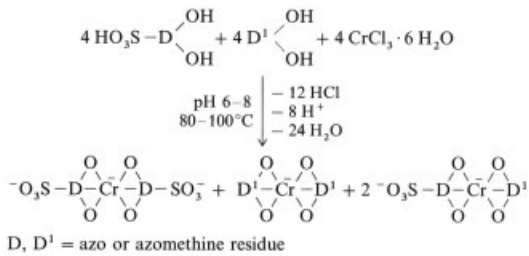

All three of the resulting dyes can be detected chromatographically. The solubility of such a mixture is often higher than that of a mechani-

cally prepared mixture of the same dyes. Mixed metallization opens the way to a large number of potential combinations, permitting the manufacture of brown, olive, marine blue, and black color tones that are otherwise difficult to achieve.

Use of three diazo and three coupling components can result in up to nine metal-free azo dyes, and these in turn can lead to 45 different 1 : 2 metal complexes. If one begins with ten components the number of possibilities increases to more than 5000.

4. Ecology and Toxicology

Dyes may enter the environment at several points: during production, in the dye works, and during use of the finished goods [42]. Losses during production and use depend on molecular structure, but average values from the textile, leather, and paper industries appear to be about 10 % of the total annual production of 450 000 t [43].

The effectiveness of a dye is a function of its stability with respect to a wide range of chemical and physical influences. Thus, dyes are degraded only slightly during normal wastewater treatment due to short residence times and aerobic conditions. Nevertheless, between 40 and 80 % of the material is adsorbed onto the activated sludge. Further degradation in the sediment by reductive anaerobic processes is rapid.

Toxicity studies in fish, involving more than 3000 commercial dyes, revealed that seven metal complexes had an $LC_{50} \leq 1$ mg/L. Such concentrations are not present in colorless effluents leaving a wastewater treatment plant (dye concentrations ≥ 1 mg/L are clearly visible). Typical distribution coefficients P of the complexes in n-octanol – water allow their bioaccumulation in fish to be estimated. Values indicate that the dyes are not bioaccumulated because solubilities in water are either too high or too low [44]. Studies also indicate that toxicity to algae and inhibitory effects on clarified sludge bacteria are negligible [45]. Metal-complex dyes possess only slight acute toxicity [46]. A published list of dyes classified as toxic ($LD_{50} \leq 200$ mg/kg) does not include metal complexes; their LD_{50} values all exceed 2000 mg/kg. Proposals for the safer handling of dyes are described in [47]. Relevant international regulations are reviewed in [48].

Economic and technological factors, together with ecologically suspect dyeing practices (e.g., large amounts of heavy metals in mordant baths), preclude the use of natural dyes as an alternative to synthetic substances [49]. Heavy-metal residues (including chromium) in effluents from dye production pose a risk of water pollution, but they can be eliminated by precipitation with calcium salts [50]. Processes for removing residual heavy metals from the reaction mixture in the form of water-soluble complex dyes are described in [51].

Different problems are posed by the after chroming process in which a metal-free complexable dye is used to stain wool in a process similar to application of an acid leveling dye. The dye is then converted into a metal-complex dye directly on the fibers with the aid of sodium dichromate. Improved procedures have drastically reduced the concentration of chromium salts in the corresponding effluents [52]. A report on the cleaning of effluents from textile mills is presented in [53].

5. Uses of Metal-Complex Dyes

Metal-complex dyes are unsurpassed in terms of range of application. Apart from certain restrictions affecting a few synthetic fibers, virtually all substrates can be dyed or printed with these substances but with dull shades. No other class of dyes provides such a generally high level of fastness, particularly lightfastness. Countless shades from greenish yellow to uniform black can be obtained by using different metals and combining different metal-complex dyes.

5.1. Textile Dyeing

5.1.1. Wool and Polyamides (→ Azo Dyes, Chap. 3.2.1.4.→ Azo Dyes, Chap. 3.2.2.)

Worldwide wool production is about 1.6×10^6 t, with a very low growth rate. Wool constitutes only about 4 % of all textile fibers, but because deep shades such as blue, brown, and black dominate, the corresponding fraction of total

dye consumption is twice as great, namely 8 %. Technical and economic criteria for selecting wool dyes are discussed in [54].

Premetallized complex dyes constitute 30 % of all wool dyes, a proportion slightly less than that assigned to secondary chromium incorporation dyes [52]. These mordant dyes are still important, especially for dyeing wool in dark shades with very high fastness. Premetallized dyes are 1 : 1 chromium, 1 : 2 chromium, or 1 : 2 cobalt complexes (no other metals are used). Their anionic character is responsible for their great colorfastness with respect to the highly polar wool fibers, and it also has a decisive influence on the corresponding application technology.

On account of their pH stability, the 1 : 1-chromium complexes are synthesized at pH ≤ 3, 1 : 2 chromium and cobalt complexes are prepared at ca. pH 7. The same pH values are required for the dyeing of wool. On prolonged heating at 100 °C and pH > 9 the 1 : 1 chromium complex decomposes into the 1 : 2 complex and chromium hydroxide. At 100 °C and pH < 4, however, the 1 : 2 chromium complexes undergo disproportionation to form the 1 : 1 complex and the metal-free dye.

Although a multitude of shades can be found in wool dyed with metal complexes, only a few combinations of diazo and coupling components are widely used, and all the applicable dyes are based on chromium and cobalt because of fastness considerations and problems related to synthesis. The most important 1 : 1 and 1 : 2 chromium and 1 : 2 cobalt wool dyes are listed in Table 2.

The chief starting materials in the yellow, orange, and red ranges are anthranilic acids and 1-aminophenols, together with acetoacetic acid anilides and pyrazolones substituted by electron-withdrawing groups. More bathochromic colors are obtained with naphthols as the coupling component, including the important blue and black shades resulting from incorporation of diazotized 1-amino-2-naphthol-4-sulfonic acid and the analogous 6-nitro compound.

Major suppliers of wool dyes together with relevant trade names are summarized in Table 3.

The 1 : 1 chromium complexes containing sulfonic acid groups precipitate as betaines (**7**) during synthesis and must be converted with base into soluble salts (**12**) for dyeing purposes:

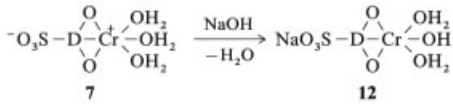

7 **12**

D = azo or azomethine residue

The coloristic and economic significance of 1 : 1 chromium complexes remains high even though they have been marketed since 1920. This can be attributed to high migration and high leveling capacity. Wetfastness in the light- and middle - color ranges is also excellent, remaining adequate even with deeper shades. However, the fact that strongly acidic conditions (pH 2) are required in the dyeing process is a disadvantage because wool is then subjected to hydrolytic damage. Use of more gentle conditions (i.e., pH 4) produces uneven dyeing due to compound formation between the chromium(III) ion and nucleophilic groups in the wool keratin, which in turn inhibits dye migration. This problem does not arise at pH 2 because of protonation [8]. Colorless bi- and trifunctional chelators can be used to displace H_2O ligands on 1 : 1 chromium complexes to give, without a significant shift in color tone, 1 : 2 chromium complexes suitable for gentle wool dyeing in weakly acidic baths. An example is the group of Neopalatin Fast dyes (**13**, BASF), with sulfosalicylic acid anilides as colorless ligands, which came on the market in 1954 [55]:

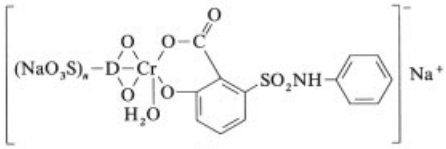

13

$n = 1, 2$
D = azo or azomethine residue

Other metal complexes suitable for dyeing wool are 8-hydroxyquinoline derivatives [56].

The 1 : 1 chromium complex modified in this way does not display any deficiencies from the standpoint of applications, but such dyes are not widely used for reasons of cost. An alternative gentle method proposed more recently for dyeing wool fibers involves 1 : 1 chromium com-

Table 2. Colors of 1 : 1 and 1 : 2 chromium complexes and 1 : 2 cobalt complexes

Diazo component \ Coupling component	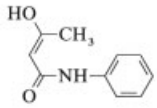	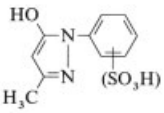	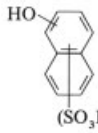	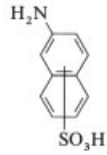
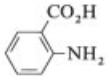		yellow		
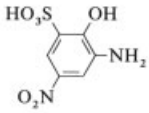	yellow	orange	bordeaux violet brown gray	green olive
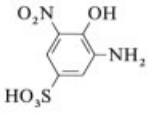		scarlet	blue black	
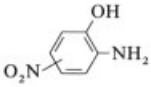 (4- or 5-nitro only)		red	blue black	green gray
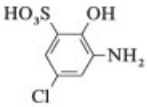	yellow	red	violet gray	
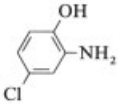			violet	
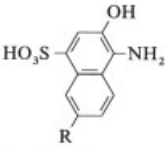 R = H, NO_2		bordeaux	blue black	

plexes in conjunction with anionic auxiliaries and alkali, ammonium, or silicofluorides (Neolan P dyes, Ciba-Geigy) [57].

The first nonsulfonated 1 : 2 chromium complex was marketed by Geigy in 1949. This dye, Irgalan Grey BL [*12218-94-9*] **14**, developed by G. SCHETTY, represented a milestone in complex chemistry [58]:

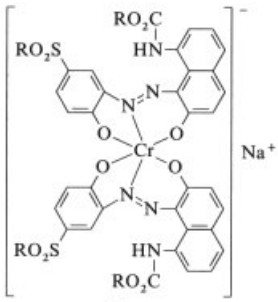

14

At about the same time, BASF was conducting research on the Vialone dyes. Unlike Irgalan Grey BL (which bears methylsulfone groups), these compounds require auxiliary water-solubilizing substituents. Vialone dyes were first marketed in 1951 [59] and are the least hydrophilic of the 1 : 2 metal complexes. For this reason they are used in dispersed form to dye nylon (e.g., C.I. Acid Yellow 59, 18 690, **15**):

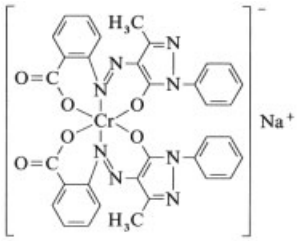

15

The commercial success of the Irgalan dyes marked the beginning of intensive research worldwide, which in turn led to better understanding of the relationship between the chemical structure of 1 : 2 metal complexes and their color and fastness properties. By 1964, several product lines were available commercially, representing more than 800 individual products [12]. All were free of sulfonic and carboxylic acid groups; their water solubility was enhanced by incorporation of nonionic alkylsulfonyl [60], alkylsulfonylamino [61], cyclic sulfone [62], and especially sulfonamide groups, sometimes in the form of mono- or dialkylated derivatives [63]. Unlike 1 : 1 chromium complexes, these dyes can be applied to wool under neutral to weakly acidic conditions, so that the wool fibers remain largely intact. The resulting colors are even and key properties such as wet- and lightfastness are outstanding.

The mechanism of the associated dyeing process has been the subject of numerous investigations [64]. The 1 : 2 metal complex dyes containing a single sulfonic acid group possess two negative charges. They are readily soluble in water, are applied in weakly acidic dye baths, and provide even dyeings with good fastness. Most are prepared in stages; that is, the 1 : 1 metal complex with sulfonic acid groups is isolated first, and this is added in the presence of base to a tridentate azo or azomethine dye to produce a homogeneous, asymmetric 1 : 2 metal complex [65] (e.g., **16** [66]):

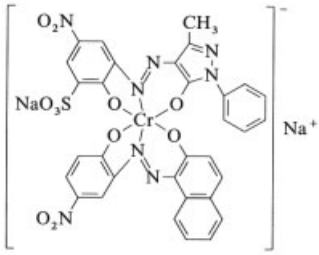

16

Proper selection of diazo and coupling components permits optimization not only of the color but also of such characteristics as leveling capacity and fiber affinity. Compared with 1 : 1 chromium complexes, monosulfonated 1 : 2 chromium and cobalt complexes result in more intense color and less damage to the wool; furthermore, they are generally easier to synthesize than unsulfonated complexes (reaction temperature $< 100\,°C$). Manufacturing processes have been described that avoid isolation of the intermediate 1 : 1 chromium complex [67].

The monosulfonated dyes marketed under the trade name Lanacron (Ciba) in 1960 were followed ten years later by a series of disulfonated metal-complex dyes (Acidol-M, BASF). These are the most hydrophilic of all metal-complex dyes, and possess three anionic charges (e.g., C.I. Acid Blue 193, 15 707 [*75214-58-3*], Acidol Dark Blue M-TR, **17**):

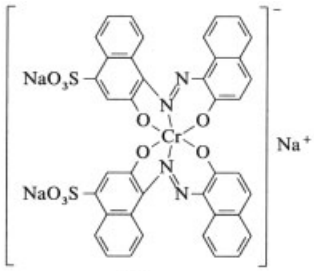

17

Dyes of this type have been known since 1932 [68] but were studied in detail only after 1956 [69]. The hydrophilic character of these dyes require the addition of suitable leveling agents for deep and even dyeing [70].

Economical, relatively brilliant colors with high lightfastness are obtained under weakly

acidic dyeing conditions. The wetfastness of these materials is superior to that of all other metal - complex dyes, which has enabled them to capture a relatively large market share within a very short time. Wetfastness can be increased further by the addition of polyquaternary ammonium compounds, particularly in the case of chlorinated (shrink-resistant) wool. The result is a machine-washable wool product [71].

The principal applications of wool-reactive dyes also occur in this context. Such dyes were first marketed in 1960 under the trade name Procilan (ICI). The reactive moiety is the acrylamide residue ($-NH-CO-CH{=}CH_2$) [72].

Compared with the corresponding chromium complexes, 1 : 2 cobalt complexes are generally distinguished by their much greater lightfastness, with color shades shifted in a hypsochromic sense.

Dyed synthetic polyamide materials undergo photochemical degradation when exposed to light and heat. This is particularly disturbing with such items as automobile seat covers. Stabilization can be achieved by the use of copper complexes [73], including complex **18** with a bisazomethine ligand:

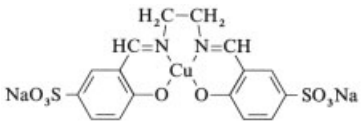

18

Synthesis of a 1 : 1 Cobalt Complex Stabilized with Nitrite [36]: A mixture of 28 g of $CoSO_4 \cdot 7\,H_2O$ and 42 g of $NaNO_2$ is dissolved in 400 mL of water. The solution is cooled to 10 °C, and 51 g of the dye is then added and partially dissolved. The dye is obtained by acid coupling of diazotized 2-amino-4,6-dinitrophenol with 2-phenyl-aminonaphthalene-4′-sulfonic acid. To ensure complete dissolution, 50 g of dimethylformamide is then added dropwise, the pH being kept at 5.0 – 5.5 by addition of HCl. Metallization commences immediately and is complete within 2h. The 1 : 1 cobalt complex is precipitated by addition of NaCl, isolated by filtration, washed with water, and dried.

Demethylative Incorporation of Chromium to Give the 1 : 1 Chromium Complex 19 [74]**:**

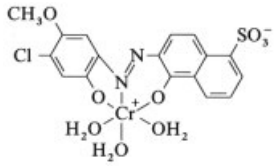

19 [*80004-31-5*]

To obtain **19**, 400 mL of 17 % HCl is added to 376 g of 5-amino-2-chlorohydroquinone dimethyl ether in 2800 g of ice water. Diazotization is accomplished by addition of 620 mL of 23 % $NaNO_2$ solution at 0 – 5 °C. Excess nitrite is reacted with amidosulfonic acid. The filtered diazo solution is introduced into a solution of 450 g of 1-hydroxynaphthalene-5-sulfonic acid in 3200 mL of ice water containing 70 g of NaOH solution and 100 g of Na_2CO_3. Coupling is complete within 2 h. The pH is then adjusted to 1 with HCl. Next, a solution of 300 g of Cr_2O_3 in 300 g of formic acid and 2000 mL of water is added, and the mixture is heated in an autoclave at 130 °C for 3 h. After cooling, the product is separated by filtration and dried. A black powder is obtained (yield: 750 g). The product dyes wool and leather grayish blue.

Synthesis of the 1 : 1 Chromium Complex 20 with a Colorless Ligand [55]:

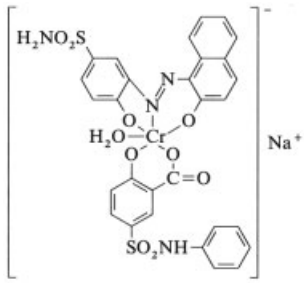

20

A mixture of 20 g of the 1 : 1 chromium-containing monoazo dye obtained from diazotized 3-amino-4-hydroxybenzenesulfonamide and 2-hydroxynaphthalene, 400 mL of ethanol, 14 g of crystallized sodium acetate, and 15 g of 2-hydroxy-5-sulfobenzoic acid phenylamide is boiled for 2 h with stirring. After cooling the

Table 3. Important dye series based on chromium and cobalt complexes

1 : 1 Metal complexes	1 : 2 Metal complexes without a sulfo group	1 : 2 Metal complexes with one sulfo group	1 : 2 Metal complexes with two sulfo groups
Neolan, Ciba-Geigy Palatin Fast, BASF Inochrom, ICI	Irgalan, Ciba-Geigy Isolan K, Bayer Ortolan, BASF	Lanacron S, Ciba-Geigy Neutrichrom S, ICI Lanasyn S, Sandoz	Acidol M, BASF Neutrichrom M, ICI Formalan M, Althouse-Tertre
Vitralon, Sandoz Tetroxan, Althouse-Tertre	Vialone, BASF Lanasyn, Sandoz	Isolan S, Bayer Formalan P, Althouse-Tertre	
	Formalan S, Althouse-Tertre		

solution is evaporated. The residue is a water-soluble dye for producing bordeaux red shades in wool from weakly acidic baths.

Synthesis of the Symmetric 1 : 2 Chromazo Dye 21 [75]:

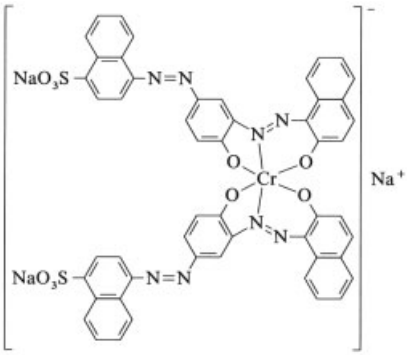

21

Diazotized 1-aminonaphthalene-4-sulfonic acid is coupled with 1-acetylamino-2-hydroxybenzene, after which the acetylamino group of the monoazo dye is saponified. The compound is then diazotized a second time and coupled with 2-naphthol. A 52-g portion of the resulting disazo dye is dissolved in 2000 mL of water containing 6 g of NaOH at 100 °C. To this solution is added dropwise, over the course of 30 min, 200 mL of an aqueous solution containing 8.1 g of $K_2Cr_2O_7$ and 13.5 g of glucose. Once chromium incorporation is complete, the mixture is cooled to 60 °C, mixed with 250 g of NaCl, and filtered. After drying, a brown-black powder is obtained which dyes wool in fast brown shades.

Stepwise Synthesis of the Asymmetric 1 : 2 Chromazo Dye 22 [74]**:**

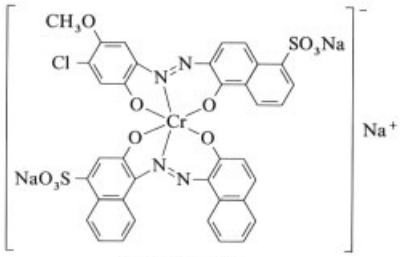

22 [*80004-32-6*]

A mixture of 40 g of the azo dye obtained from 1-diazo-2-naphthol - 4 - sulfonic acid and 2-naphthol is heated with 40 g of the 1 : 1 chromium complex **19** in 500 mL of water at 80 °C and stirred with aqueous NaOH at pH 7.5 for 5 h. The mixture is then cooled and the product is salted out with NaCl. Drying leads to a black powder that dyes wool and leather reddish blue. The resulting colors are fast, particularly on shrink-resistant wool.

Synthesis of the Statistical 1 : 2 Chromazo Dye Mixture 23 [76]:

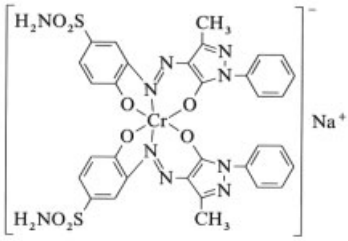

[30112-70-0]

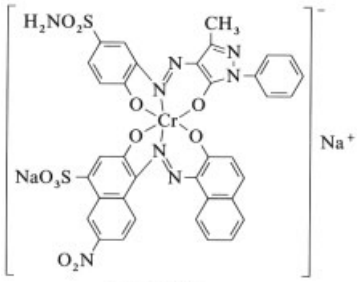

[81161-61-7]

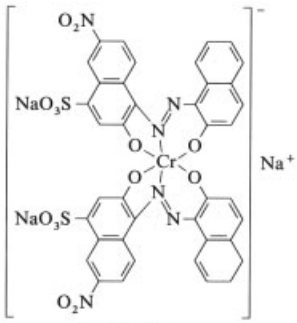

[53036-47-8]
23

A suspension of the azo dye obtained from 18.8 g of 2-aminophenol-4-sulfonamide and 17.4 g of 1-phenyl-3-methylpyrazolon-5-one is mixed with a suspension of the azo dye obtained from 29.5 g of 6-nitro-1-diazo-2-naphthol-4-sulfonic acid and 17.5 g of 2-naphthol, and the mixture is adjusted to pH 5.5 with glacial acetic acid. Then, 28 g of chromium(III) acetate is added and the mixture is refluxed for 4 h. Once chromium incorporation is complete the dye solution is filtered, cooled to room temperature, and treated with NaCl to precipitate the chromium mixed-complex dye, which is then filtered and dried. This dye imparts an intense dark brown color to wool and polyamide fibers; it also has good leveling and fastness properties.

Synthesis of the Asymmetric 1 : 2 Chromazo – Azomethine Dye 24 [77]:

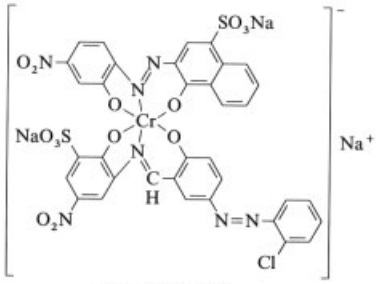

24 [110305-03-8]

A 1 : 1 chromium complex is prepared from diazotized 2-amino-5-nitrophenol and 1-hydroxynaphthalene-4-sulfonic acid. Then, 38.9 g of this dye (which contains ca. 5.2 g of chromium) is suspended in 500 mL of water and mixed with 23.4 g of 2-amino-1-hydroxy-4-nitrobenzene-6-sulfonic acid and 26 g of the monoazo dye obtained from diazotized 1-amino-2 - chlorobenzene and salicylaldehyde. The reaction mixture is heated to 80 – 85 °C, adjusted to pH 7 with NaOH, and maintained at this temperature and pH until starting materials are no longer detectable. The product is then precipitated by addition of NaCl and separated by filtration. After drying, a dark powder is obtained which dissolves in water to give a green solution that imparts olive-green shades with good fastness on polyamides.

Synthesis of the Symmetric 1 : 2 Chromium Bisazomethine Dye 25 [78]:

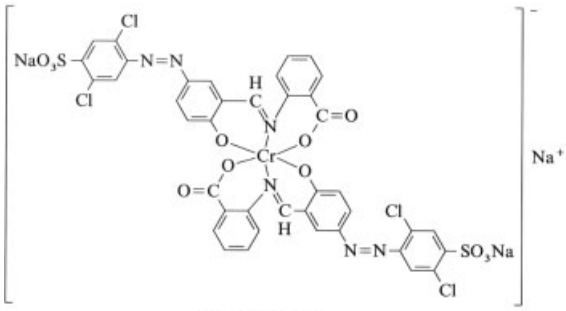

25 [82269-28-1]

A mixture of 112.5 g of the azo compound obtained from diazotized 2-amino-1,4-dichlorobenzene-5-sulfonic acid and salicylaldehyde is stirred with 42 g of anthranilic acid in 1200 mL of water and heated to 50 °C for 1 h with 12 g of NaOH and 50 g of sodium acetate. A chromium(III) formate solution prepared from 12 g of Cr_2O_3, 14 mL of formic acid, and 250 mL of water is then added, and the mixture is heated under reflux until thin layer chromatography reveals complete conversion to the 1 : 2 chromium complex. The pH is then adjusted to 7.8 with NaOH, and the complex dye is isolated by filtration and evaporation or spray drying. The brown dye powder (246 g, 33 % of which consists of sodium salts of acetic and formic acids), gives light- and wetfast reddish yellow colors with wool and synthetic polyamides.

Synthesis of the Symmetric 1 : 2 Cobalt Bisazomethine Dye 26 [79]:

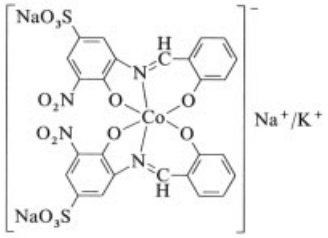

26 [*41043-60-1*]

A mixture of 44.9 g of 75.7 % 2-amino-6-nitrophenol-4-sulfonic acid and 18.5 g of salicylaldehyde is heated to boiling in 300 mL of 20 % sodium acetate solution. The suspension is then mixed at 60 °C with 19 g of $CoCl_2 \cdot 6\,H_2O$, and 17 mL of 30 % H_2O_2 solution is added within 15 min. This causes the suspension to pass almost entirely into solution. The product is precipitated with NaCl and KCl, filtered, and dried. It yields exceptionally lightfast colors with wool and polyamides.

Synthesis of the Asymmetric 1 : 2 Chromium Azomethine Complex Dye 27 [80]:

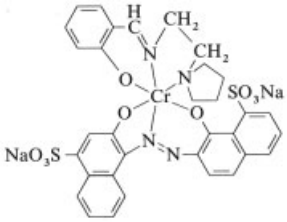

27 [*85140-75-6*]

2-Hydroxybenzaldehyde (12.2 g) is added dropwise with stirring to a solution of 11.0 g of 1-(*β*-aminoethyl)pyrrolidine and 12.0 mL of acetic acid in 100 mL of water. The mixture is stirred at 50 °C until the aldehyde has dissolved and a yellow solution of the azomethine is obtained. This mixture is further diluted with 100 mL of water at 70 °C, and 61 g of a 1 : 1 chromium complex is then added with stirring. This 1 : 1 complex is obtained as the coupling product of diazotized 1-amino-2-hydroxynaphthalene-4-sulfonic acid and 1-hydroxynaphthalene-8-sulfonic acid (C.I. Acid Blue 158). Stirring is maintained, and the pH is adjusted with dilute NaOH to 7.2 – 7.5 at 70 °C. About 13 g of NaOH is consumed before formation of the 1 : 2 complex is complete, a process that can be monitored by thin layer chromatography. The resulting complex is precipitated with 120 g of NaCl. The precipitate is filtered at room temperature and dried at 80 °C to give 82 g of a water-soluble dye that imparts a deep blue light- and wetfast color to wool and synthetic polyamide fibers.

5.1.2. Cotton

The application of copper-containing substantive (direct) dyes to cotton has long been known. One example of such a dye is Sirius Light Blue 3 RL [*13217-74-8*] (**28**), C.I. Direct Blue 93, 22 810 [81].

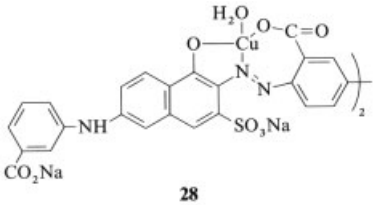

28

Dimeric copper-containing direct dyes based on 4,4′-diaminodiphenyl-3,3′-diglycolic acid

were marketed by I. G. Farbenindustrie as Benzo Fast dyes (e.g., Benzo Fast Red CGL, C.I. Direct Red 180). The incorporation of copper in complex-forming di- and polyazo direct dyes considerably improves their light- and wetfastness. The Indosol direct dyes introduced by Sandoz (mostly copper-containing direct dyes) show increased wetfastness [13, p. 135]. They are reactant-fixable dyes that are cross-linked on and with the fiber after dyeing in the presence of a special reactive agent.

The 1 : 2 metal complex dyes with low affinity for cellulose can be made to bind to cotton by introduction of an appropriate reactive group [72]. Thus, black print is obtained on cotton by employing reactive dye **29** [82]. The corresponding pure 1 : 2 chromium complex [*79828-44-7*] results in a cloudy blue color, whereas the pure 1 : 2 cobalt complex [*79817-89-3*] gives a glossy red.

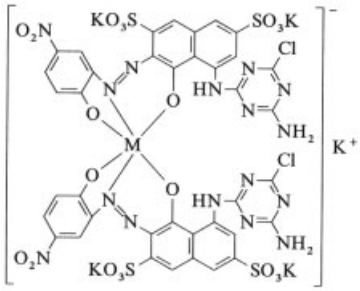

29
M = Cr:Co ≈ 3:1

By incorporating specific substituents metal-complex dyes can be used to print and dye cotton – polyester blends in various shades [83]. Previous commercial lines include Dybln (Du Pont) and Cellestren (BASF), containing tetracyclic nickel complexes such as **30**.

Synthesis of the Metal-Complex Dye 30 for Cotton [83]:

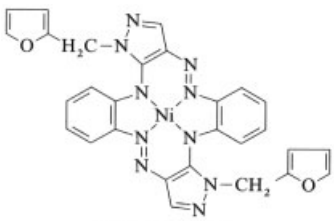

30 [*103850-03-9*]

The coupling product of 2-nitroaniline with furfuryl-5-aminopyrazole (31 g) in 300 mL of dimethylformamide is treated with 12 g of anhydrous nickel acetate and 56 g of K_2CO_3, and the mixture is stirred for 2 h at 130 °C. The olive-green reaction product is precipitated at room temperature by addition of 250 mL of methanol. The precipitate is separated by vacuum filtration, washed with a little methanol and then with water, and dried at 70 °C under vacuum. The resulting powder is crushed to permit preparation of a finely divided aqueous dispersion. Application to polyesters and mixed polyester – cottons according to patent DE-PS 1 811 796 results in fast olive colors.

Special auxiliaries are required when printing or dyeing with such substances because they act as disperse dyes.

5.1.3. Polypropylene

Polypropylene is a very hydrophobic fiber. It is used mainly in carpets and cannot be dyed with anionic dyes because it does not have any relevant functional groups. Since a high degree of lightfastness is required, the material is polymerized in the presence of nickel salts. Subsequent treatment with dyes capable of complexation results in the formation of nickel complexes during the dyeing process.

Appropriate dyes include nitrogen heterocycles with azo groups α to the nitrogen atoms and a hydroxyl group ortho to the azo group. 2-Aminothiazoles and 2-aminobenzothiazoles [84], 2-aminopyridines, pyrazoles, and imidazoles are used as the diazo components. Alternatively, the heterocycle may be incorporated into the coupling component [85].

Synthesis of the Dye 31 for Dyeing Polypropylene [85]: 2-Amino-4-chlorophenol (14.4 g) is slurried with 25 mL of 10 % NaOH, dissolved by the addition of 150 mL of water, and treated with 70 mL of 10 % $NaNO_2$ solution. The filtered solution is added dropwise with stirring (over ca. 30 min) to a mixture of 30 mL of concentrated HCl and 150 g of ice. Diazotization occurs within 1 h at 0 – 5 °C, and excess nitrite is subsequently decomposed by the addition of amidosulfonic acid. A solution of 13.2 g of 1-methylimidazo[1,5–*a*]pyridine in 70 mL of

10 % HCl is then added dropwise to the diazotization mixture. The solution is brought to pH 4 – 5 by the addition of ca. 100 mL of saturated aqueous sodium acetate, and the resulting dye is separated by vacuum filtration, washed, and dried. The product consists of 21.7 g of material corresponding to formula **31**, which is pure by thin layer chromatographic standards. It is a brown powder that dyes nickel-modified polypropylene fibers a greenish blue.

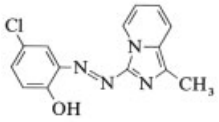

31 [*83156-84-7*]

5.1.4. Dust-Free Preparations

Toxicological, ecological, and process technology considerations have led to a demand for dye preparations that do not produce dust. Such formulations greatly simplify handling on the part of the customer [86]. The problem can be approached in several ways, including the addition of high-boiling mineral oils, use of liquid preparations, or dye granulation. The latter two variants have been intensively studied by major dye manufacturers in recent years.

The chief requirements with respect to liquid preparations are a high storage stability, a high concentration, and good miscibility with water. These are met by adding water-soluble solvents such as polyglycols, poly(glycol ethers) [87], or cyclic lactams [88].

Synthesis of a Liquid Dye [88]: β-Naphthol (4320 g) is dissolved in a mixture of 6277 mL (5712 g) of 25 % ammonia, 7740 g of pyrrolidinone, and 35 000 g of water (pH 12). The solution is then cooled to about 15 °C. During the course of 2 h, 14 334 g of 61.8 % 1-diazo-6-nitro-2-naphthol-4-sulfonic acid is added. If necessary, ammonia is used to adjust the pH to 9.5, and the solution is heated to reflux within 20 min. A mixture of 2232 g of $Na_2Cr_2O_7 \cdot 2\,H_2O$, 1500 g of glucose, and 6000 mL of water is then added within 15 min, and refluxing is maintained for approximately 1 h. A clear, stable dye solution results (C.I. Acid Black 52, 15 711 [*5610-64-0*]).

An alternative synthesis involving preparation of liquid dyes by phase separation is described in [89]. Another route for the production of highly concentrated, stable liquid preparations entails exchanging the usual countercations for ions that impart better solubility. For example, sodium ions required during synthesis can be replaced by lithium or substituted hydroxyalkyl-ammonium species [90].

Finally, salts produced during synthesis that interfere with the preparation of high-concentration liquid formulations can be removed by membrane separation techniques [91].

Dye granulates are solid dye preparations, and in this case stability is not a problem. Granulates are much more soluble in cold water than dye powders that have been treated with dust-reducing agents because the latter tend to clump when mixed with water [92].

5.2. Nontextile Uses

5.2.1. Leather

Leather can be dyed in two ways: by drum dyeing, usually with a dye powder, or by spray dyeing with liquid dyes. The most important leather powder dyes containing metal complexes are

Derma, Dermalicht (Sandoz)
Sellaecht, Sellacron (Ciba-Geigy)
Luganil (BASF)
Inoderm (ICI)
Baygenal (Bayer)

Many of these dyes can also be used for dyeing wool, furs, and hair. The most important liquid dyes are

Eukesolar (BASF)
Levaderm (Bayer)
Irgaderm (Ciba-Geigy)

Leather is often treated with di- and polyazo dyes of the direct type. Nickel salts sometimes serve as the metal ion donor, but copper salts and, especially in the important brown range, iron salts are more common. The use of chromium and cobalt compounds is relatively rare. In the case of dual coupling, resorcinol [93], 3-aminophenol [94], and *m*-phenylenediamine [95] are the preferred coupling components. The diazo

components are selected so that the greatest possible color intensity is obtained with chromium-tanned leather or leather that has been retanned with vegetable or synthetic products.

Various 1 : 2 chromium and cobalt complexes have been introduced for the *trichromic dyeing* of leather. In the trichromic process, mixtures of three dyes with different colors (e.g., yellow, red, blue, or gray) are used. An important prerequisite is that the dyes penetrate evenly into the leather and have the same affinity for different types of leather. This ensures that no color difference exists between the leather surface and the interior, a major consideration if polishing is contemplated [96].

Procedure for Dyeing Leather [97]: Chromium-tanned calfskin (100 g) is washed in 300 mL of water for 10 min at 30 °C and 24 rpm. The wash water is then removed and replaced by a mixture of 400 mL of water at 30 °C, 1 g of sodium formate, and 1 g of $NaHCO_3$. After neutralization for 45 min, the leather is again washed with 300 mL of water at 30 °C. The neutralized leather is then dyed in 300 mL of water at 50 °C with a dye mixture consisting of 9.072 g of the yellow dye **32**, 0.123 g of the red dye **33**, and 0.21 g of the gray dye **34**.

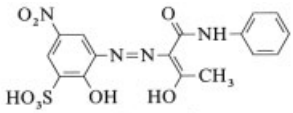

1:2 Co complex
32 [*10343-56-3*]

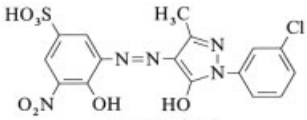

1:2 Cr complex
33 [*83914-74-3*]

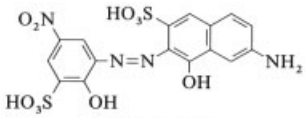

1:2 Cr complex
34 [*111308-46-4*]

After 30 min, 3 g of a lubricant based on sulfonated chloroparaffins is added to the dye bath, followed after another 30 min by 0.5 g of 85 % formic acid. After an additional 20 min, the leather is rinsed with cold water and processed as usual.

Polyazo 1 : 2 chromium complexes give leather colors that show high resistance to poly(vinyl chloride) migration, a particularly important characteristic for sports shoes [98]. The black dye **35** is an example of such a substance; the iron complex of **36** gives yellowish brown shades:

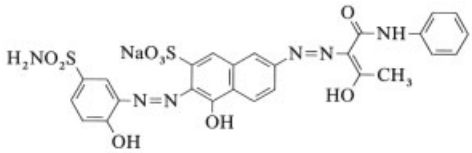

1:2 Cr complex
35 [*119654-54-5*]

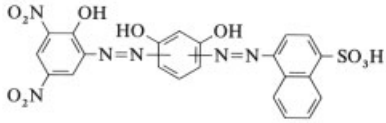

36 [*88332-66-5*]

Double 1 : 2 metal complexes, with up to six negative charges, show especially good incorporation into leather and outstanding fastness [99], [100]. For example, **37** produces fast brown shades.

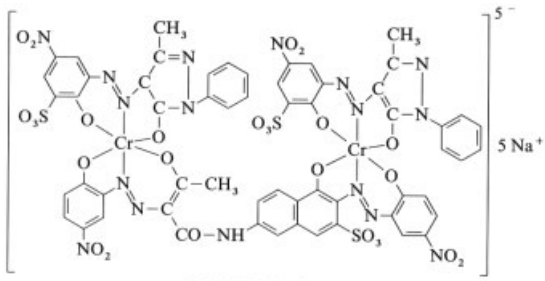

37 [*103761-82-6*]

5.2.2. Paper

Octahedral metal complexes are seldom used for coloring paper because of their low substantivity and associated heavy discoloration of effluent water. The 1 : 1 copper complexes predominate because they tend to be more lightfast than metal-free dyes. Such dyes are often prepared by demethylative or oxidative coppering [30]. Cationized 1 : 1 and 1 : 2 metal complexes have

been intensely studied by Sandoz. They provide an almost colorless effluent, and the resulting paper shows good lightfastness and does not "bleed" upon contact with milk, fruit juice, alcohol, or mineral water [101].

Synthesis of the 1 : 1 Copper Complex 38 by Demethylative Coppering [102]:

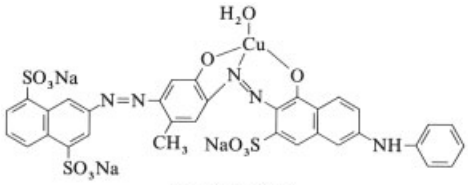

38 [*92341-30-5*]

The aminoazo dye obtained by the weak-acid coupling of diazotized 2-aminonaphthalene-4,8-disulfonic acid with 2-methoxy-5-methylaniline (45.1 g) is dissolved in 500 mL of water and mixed with as much NaOH as required to give a neutral solution. This solution is then treated with $NaNO_2$ (6.9 g) and allowed to flow into a mixture of 25 mL of 30 % HCl and 50 g of ice. More ice is added as required to keep the temperature below 5 °C. As soon as diazotization is complete, excess nitrous acid is destroyed with a little amidosulfonic acid. The resulting suspension of diazonium salt is introduced into a neutral solution of 31.5 g of 5-hydroxy-2-phenylaminonaphthalene-7-sulfonic acid in 60 mL of water. The pH is kept between 8 and 8.5 by simultaneous addition of 100 mL of 10 % Na_2CO_3 solution. The coupling mixture is heated to 90 °C, mixed with 50 g of diethanolamine and a solution of 25 g of $CuSO_4$ (crystalline), 45 mL of 25 % aqueous ammonia, and 60 mL of water. It is then stirred for 8 h at 90 °C. The copper-complex dye obtained is precipitated with NaCl and separated by filtration. Paper treated with this dye acquires a pure blue color (maximum absorption at 595 nm) with particularly good light- and wetfastness.

Synthesis of the Bis 1 : 1 Copper Complex 39 by Oxidative Coppering [103]:

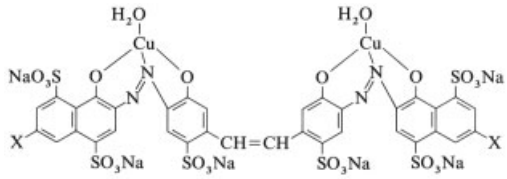

X = CH_3CONH–

39 [*119103-25-2*] (Cu-free)

The coupling dye obtained from diazotized 4,4′-diamino-stilbene-2,2′-disulfonic acid and 6-acetylamino-1-hydroxynaphthalene-4,5-disulfonic acid (88.5 g) is dissolved in 2000 mL of water at 60 °C. The solution is adjusted to pH 6 with a little glacial acetic acid, after which a solution of 25 g of $CuSO_4 \cdot 5\,H_2O$ in 75 mL of water is added. Then, 200 mL of 3 % H_2O_2 solution is added dropwise, the pH being maintained at 5 – 6 by addition of aqueous NaOH. After completion of oxidative coppering the mixture is heated to 80 °C and mixed with 450 g of NaCl. The precipitated dye is then isolated and dried. It produces blue shades in paper and cotton.

Synthesis of the Cationic 1 : 1 Copper Complex 40 by Demethylation [104]:

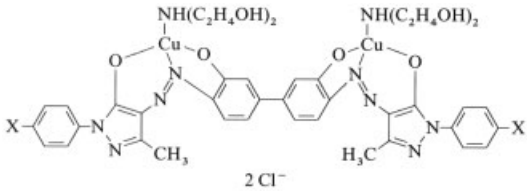

2 Cl^-

40

X = $(CH_3)_3\overset{+}{N}CH_2CONH$–

Dianisidine base (24.4 g) is coupled in the usual way with 65 g of the pyrazolone derivative **40** by using an aqueous medium acidified with HCl to pH 4. A red dye is obtained. A portion of this dye (34.4 g) is dissolved or suspended in a mixture of 400 mL of dimethylformamide, 42 mL of diethanolamine, and 13.3 g of copper acetate, and the reaction mixture is stirred for 4 h at 90 °C. The volume is then reduced by half with a rotary evaporator, and the red coppered dye is precipitated by the addition of 3000 mL of acetone.

5.2.3. Dyes for Anodized Aluminum

The 1 : 2 chromium and cobalt complexes of water-soluble azo dyes are suitable for coloring anodized aluminum [105]. For example, the symmetric 1 : 2 cobalt complex **41** imparts a black color to this substrate [106].

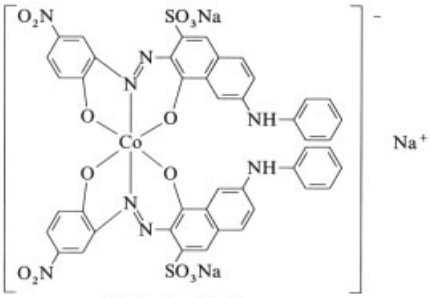

41 [*112144-04-4*]

5.2.4. Solvent and Ink-Jet Dyes

Solvent dyes are substances that are sparingly soluble in water but readily soluble in organic solvents such as alcohols, esters, and ketones. They have many applications, including foil printing and thread dyeing. Solvent dyes are also found in transparent lacquers, wood stains, and numerous office products (e.g., colors in felt-tip pen, ballpoint pastes, and ink for overhead projection transparencies). A few commercial product lines follow:

Orasol	(Ciba-Geigy)
Neozapon	(BASF)
Zapon	(BASF)
Savinyl	(Sandoz)
Lampronol	(ICI)
Valifast	(Orient Chemicals)
Aizen Spilon	(Hodogaya)

Anionic 1 : 2 chromium and cobalt complexes are particularly lightfast. To increase their solubility they are sometimes laked with organic bases such as long-chain aliphatic amines, guanidines, or cationic dyes. Laking converts the sodium salts produced during synthesis into salts containing organic cations [107]. Due to their insolubility in water, solvent dyes can be completely freed from salt by washing with water. They are marketed as 100 % dyes without additives. Solubilities up to 1000 g/L can be achieved in organic solvents, as in the case of the red dye **42** [108].

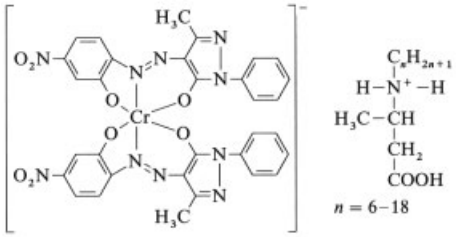

42

One new use of metal-complex dyes is ink-jet printing. In the ink-jet process, ink droplets are transferred to the substrate by spraying (→ Imaging Technology, Chap. 2.1.3.) [109]. The dye solutions are normally based on water, glycols, or cyclic aliphatic amides. The most important requirement is a total absence of salt because, otherwise, the capillaries of expensive print heads might become blocked or degraded by corrosion. The dye must not crystallize. Additional requirements for ink-jet colors are discussed in [110].

To obtain the desired properties, standard chromophores are replaced by residues such as **43** [109], **44** a formazan dye [111], or **45** [112] (see top of the next page).

Other examples include the laked (statistical) 1 : 2 chromium complex **46** used in printing inks (absorption maximum 580 nm) [113] and the 1 : 2 copper complex **47**, a lithium salt used as an ink-jet dye [114].

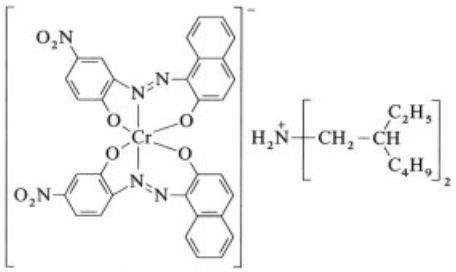

46

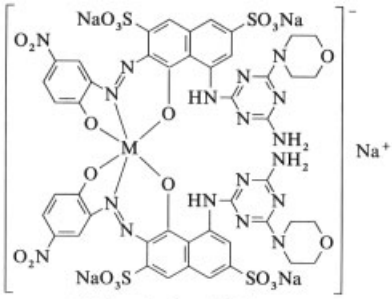

43 M = Cr:Co = 70:30
M = Co [*104815-64-7*]
M = Cr [*104815-63-6*]

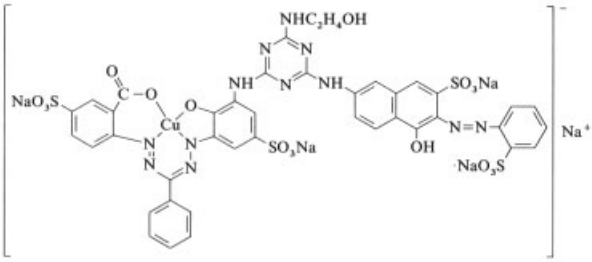

44 [*109973-79-7*]

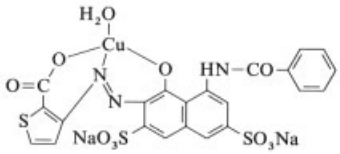

45 [*116932-38-8*]

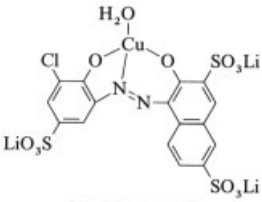

47 [*113989-79-0*]

5.2.5. Pigments (→ Pigments, Organic)

Azo and Azomethine Pigments. Metal-complex dyes for use as pigments contain bi-, tri-, and tetracyclic planar systems and are devoid of solubilizing groups. Complexes of azomethine dyes play a much greater role than azo pigments with complex-bound metals. Nevertheless, a few azo dyes with astonishingly low molecular mass display pigmenting properties as well. For example, the bicyclic 1 : 1 nickel complex **48** [115] is suitable for mass dyeing polyolefin, polyamide, or polyester fibers. Numerous additional pigment applications are described in [115].

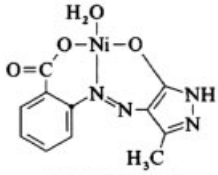

48 [*106335-41-5*]

Examples of yellow metal-complex pigments include 1 : 1 nickel compounds containing sym-

metric [116] or unsymmetric [117] heterocyclic azo moieties (e.g., **49**).

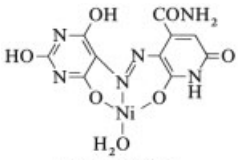

49 [*88717-24-2*]

Such systems are prepared from diazobarbituric acids (in turn obtained from benzenesulfonic acid azide by diazo group transfer) through coupling with 2,6-dihydroxypyridine-4-carboxamide followed by metallization with nickel chloride.

Reaction of salicylaldehyde with 2-amino-4-phenylphenol followed by treatment with copper acetate yields the bicyclic 1 : 1 copper azomethine dye **50** [118], obtained as a greenish yellow pigment.

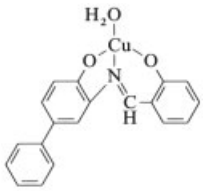

50 [*21405-81-2*]

2-Hydroxyazo compounds form neutral bicyclic 1 : 2 complexes with divalent metal ions. The yellow pigment **51** [119] is particularly useful for coloring paint.

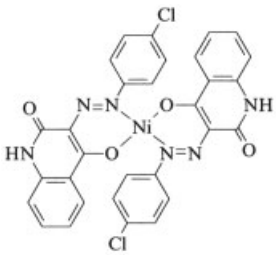

51 [*51931-46-5*]

Useful yellow pigments are also obtained upon condensation of iminoketoisoindolenine with 2-cyanomethylquinazolone. The yellow 1 : 2 nickel complex **52** [120] can be rolled into poly(vinyl chloride) films to give a yellow color with good fastness to light and migration. A paint pigmented with the nickel complex also exhibits very good resistance to light and weather.

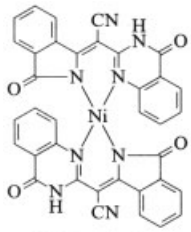

52 [*106646-72-4*]

Synthesis of the Bicyclic 1 : 2 Nickel Pigment 52 [120]. The condensation product (32 g) of 2-cyanomethylquinazolone and iminoketoisoindolenine is stirred with 26 g of $Ni(COOCH_3)_2 \cdot 4\,H_2O$ in 300 mL of dimethylformamide for 16 h at 150 °C. The precipitate is then removed by hot vacuum filtration and washed with methanol and water. The product is a greenish yellow powder (yield 26 g).

Bicyclic yellow to red pigments are obtained by nitrosation of *N*-arylacetoacetamides, followed by reaction with ammonia (**53**) [121] or hydroxylamine (**54**) [122] and subsequent treatment with divalent metal ions. The products are actually tetracyclic due to the formation of additional rings by hydrogen bonding.

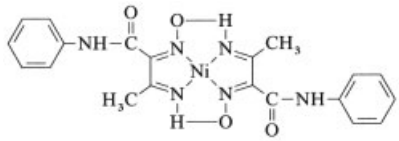

53 [*107861-02-9*]

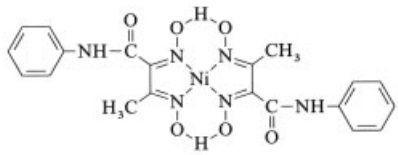

54 [*91602-80-1*]

The same applies to the nickel diacetyldioxime dyes that have long been utilized in analytical chemistry (see Section 5.2.7). The structures of the corresponding platinum and palladium complexes have been studied and calculated [123].

Yellow pigments are derived from the oximes of arylazosalicylaldehydes by coupling with nickel salts (**55**) [124]; many applications are described in [124, pp. 13, 16 – 18].

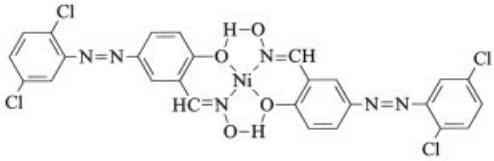

55 [*106223-22-7*]

In the case of tricyclic pigments, azomethines are the compounds of greatest industrial significance; an example is the bisazomethine compound (**56**) [125].

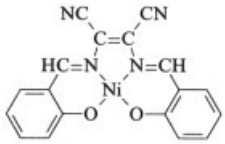

56 [*64696-98-6*]

Yellow and orange 1 : 2 nickel complexes arise from formylated quinazolo[3.2.1]quinolones and *o*-phenylenediamines (**57**) [126].

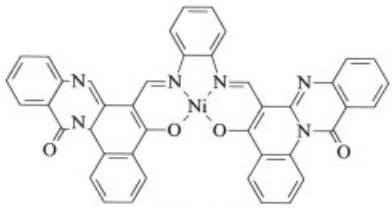

57 [*97253-30-0*]

Synthesis of the Tricyclic 1 : 2 Nickel Complex 57 [127]. 4-Formyl-3-methyl-1-phenyl-5-pyrazolone (34.38 g) is stirred with 51.51 g of 1-(cyanophenylcarbamoylmethylene)-3-hydrazonoisoindoline, prepared from 1-(cyanophenylcarbamoylmethylene)-3-iminoisoindoline and hydrazine hydrate, in 1000 mL of ethyl cellosolve at 110 – 115 °C for 3 h. The resulting brown suspension is cooled to 90 °C and filtered. The product is washed with ethyl cellosolve and ethanol, and dried overnight at 80 °C under vacuum. The yield is 57 g (i.e., 69 % of theoretical) of the ligand in the form of a brown powder. Reactions of 46.8 g of the above ligand with 26.3 g of nickel acetate tetrahydrate in 800 mL of dimethylformamide for 4 h at 120 – 125 °C gives 20.3 g (39 % of theoretical) of an orange-red nickel complex.

Pigments obtained from aminoiminoisoindolenines treated with CH-active compounds and substituted hydrazones have been the subject of extensive research. Thus, the 1 : 2 nickel complex **58** is used for scarlet-red shades with outstanding fastness in both plastics and paints [127].

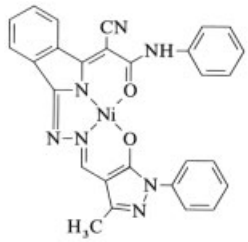

58 [*77135-16-1*]

Alternatives to the formylpyrazolone moiety in **58** include 3-acetyl-2,4-dioxyquinoline [128], 3-acetyl-4-hydroxy-6-methyl-2-pyridone [129], *N*-methyl-5-cyano-2-hydroxy-4-methyl-6-pyridone-3-aldehyde [130], and 1-formyl-2-hydroxy-3-naphthalenecarboxylanilide [131].

Phthalocyanine Pigments. See also → Phthalocyanines. The pigments produced in largest quantity are the copper phthalocyanines (tetraaza[18]annulenes). Because of their cyclically conjugated 18 π-electron systems, these tetracyclic compounds (e.g., **59**) are especially stable.

Phthalocyanine – metal complexes are synthetic aza analogues of the natural porphyrin

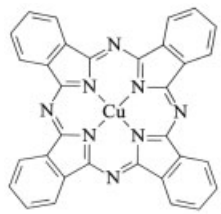

59

dyes hemin and chlorophyll (→ Phthalocyanines). Phthalocyanine is a tetrafunctional ligand that forms neutral complexes with divalent metal ions. Its copper complexes are impor-

tant as highly lightfast blue pigments; related halogen derivatives make excellent green pigments. Metal complexes of phthalocyanine are prepared directly from four identical subunits (e.g., phthalodinitrile) in the presence of the corresponding metal ions. The process occurs according to a "template reaction", in which four ligand molecules are coordinated to the metal ion in a spatial arrangement conductive to cyclization. Reaction of 2-aminobenzaldehyde with oxalyl chloride, followed by condensation with *o*-phenylenediamine, also produces tetracyclic pigments through a template reaction with divalent metals [132].

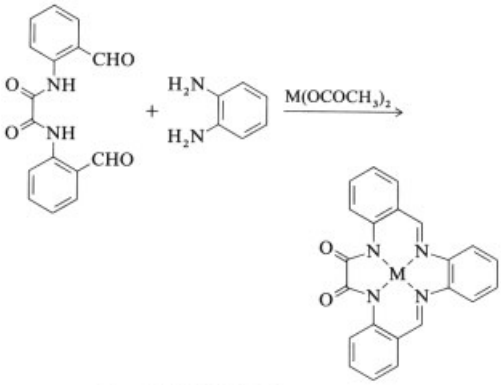

M = Ni [*59459-98-2*] (purple)
Cu [*76683-16-4*] (brilliant red)

5.2.6. Redox Catalysts

Metal-complex dyes used as redox catalysts are, like the pigments, polycyclic compounds. The well-known cobalt complex " salcomin" (**60**) binds reversibly with oxygen, thus serving, like the corresponding iron complex, as a model for biological redox catalysis.

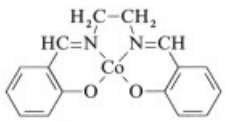

60 [*14167-18-1*]

The catalysts are prepared from salicylaldehydes and aliphatic or aromatic diamines. Tetracyclic tetraaza[14]annulenes are used commercially as redox catalysts in the production of triarylmethane dyes. For example, the iron complex **61** can be synthesized from malonic dialdehyde and *o*-phenylenediamine [133].

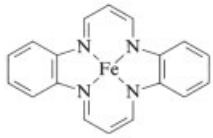

61 [*50792-65-9*]

Aza[14]annulenes can also be prepared via tricyclic 1 : 1 metal complexes, which then undergo ring closure [134]:

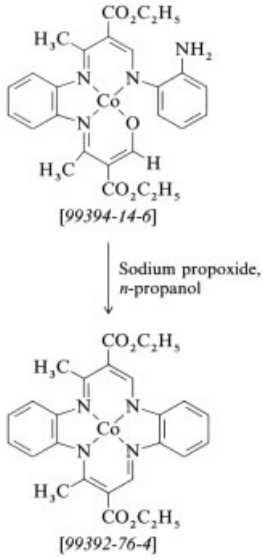

The resulting tetracyclic complex precipitates as dark red crystals. It can be used to catalyze redox processes with atmospheric oxygen, as an electrocatalyst, or as the catalyst in electrochemical sensors for the determination of oxygen and nitrogen. The 1 : 2 cobalt complex of 1-[2-hydroxynaphthyl]azo-2-hydroxybenzene-5-sulfonic acid acts as a catalyst in the hydrogen peroxide oxidation of leucocrystal violet to crystal violet lactone, an important color-forming agent in carbon-free copy paper [135].

5.2.7. Analysis

Colored chelating agents have many applications in analysis. For example, complex-forming

dyes serve as analytical reagents for the gravimetric determination of metal ions (e.g., nickel diacetyldioxime). Chelators such as Eriochrome Black T or dithizone are useful indicators in complexometric titrations [136]. The coupling compounds prepared from 3,5-halogen-substituted 2-aminopyridines and 1,3-dialkylaminobenzoic acids give specific color reactions with vanadium, iron, cobalt, nickel, copper, and palladium ions. The resulting purple to blue chelates are sparingly soluble in water, but readily soluble in organic solvents, and have high molar extinction coefficients [137].

5.2.8. Photo Dyes

Color photography makes use of two types of metal-complex dyes: 1 : 1 chromium complexes bearing colorless ligands and 1 : 2 chromium complexes that are both characterized by the presence of a developer group such as the hydroquinone residue [138]. An example is dye **62** [139]:

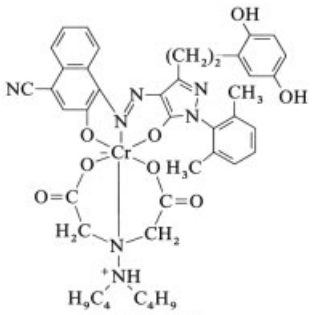

62 [*79230-32-3*]

Numerous patents describe the photographic applications of nickel complexes of 2-amino-6-arylazo-3-pyridinol dyes [140].

5.2.9. Charge Stabilizers

Metal-complex dyes are rapidly gaining importance in the electronics industry, especially in photocopying. The basis of photocopying and requirements for the corresponding dyes are described in detail in [141]; see also → Imaging Technology, Chap. 2.1.1.

Toner for photocopying is prepared by molten extrusion of a resin with both a pigment (often carbon) and a charge stabilizer (charge control agent), followed by grinding to a particle size of 5 – 15 µm. The toner is then mixed with a carrier composed of fine iron particles to provide an electrophotographic developer.

Charge stabilizers (charge control agents) are negatively [142] or positively [143] charged 1 : 2 metal-complex dyes containing chromium, cobalt, or iron. Dye **63** [144] is a negative charge control agent, whereas dye **64** [145] is a positive charge control agent:

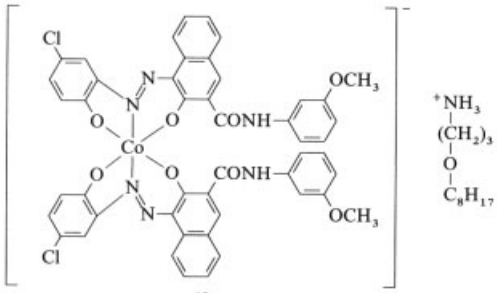

63

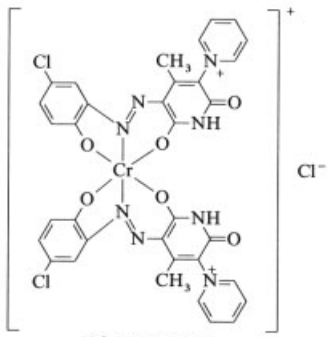

64 [*100012-90-6*]

Bicyclic 1 : 2 chromium salicylate complexes such as **65** are nearly colorless charge control agents [146].

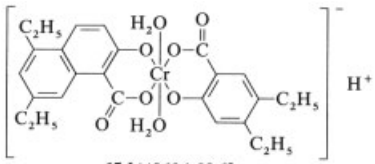

65 [*119604-58-9*]

5.2.10. Optical Applications

Binary nickel, palladium, and platinum complexes of dithiol compounds show intense absorption between about 750 and 950 nm. They are used as IR dyes for optical drawing media or laser components, and also as IR absorbers for sunglasses and tinted windshields. In addition, such compounds are used as antioxidants for polymers and other organic materials, as corrosion inhibitors, and as protective layers for color photographic films or textile colors [147]. They are produced by reacting α-hydroxyketones with phosphorus pentasulfide, followed by introduction of the appropriate metal salts:

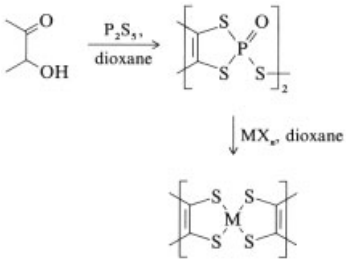

Bicyclic dithiolene complexes can be either neutral (**66**) [148] or anionic (**67**) [149]:

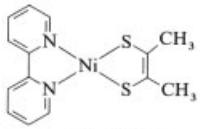

66 [*118422-25-6*]

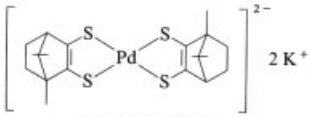

67 [*105892-91-9*]

Bicyclic copper or nickel complexes of indoanilines (**68**) [150] are also suitable for optical drawing media.

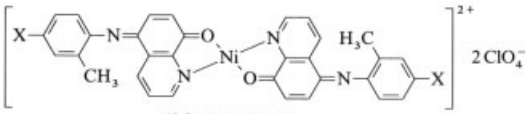

68 [*111792-88-2*]
X = $-N(C_2H_5)_2$

Certain bicyclic (**69**) [151] and tetracyclic (**70**) [152] metal complexes exhibit discotic mesophases

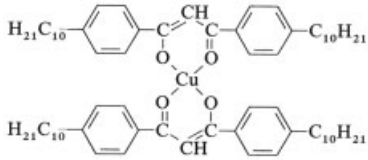

69 [*78452-82-1*]

and are thus attractive as liquid crystals for electro-optical and thermo-optical display elements.

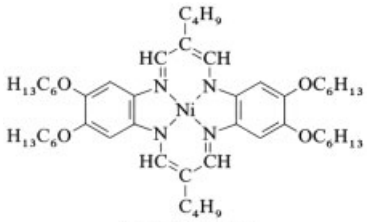

70 [*105358-51-8*]

Synthesis of 69 [151]. A solution of 160 mg of $CuCl_2$ in 3 mL of ethanol is added dropwise to a solution of 300 mg of bis(4-decylbenzoyl)methane in 1.5 mL of tetrahydrofuran at 40 °C. The mixture is stirred at 40 °C for 0.5 h, slowly mixed with 1 mL of 20 % ammonia solution, and stirred for 2 h. The resulting precipitate is separated by filtration and recrystallized from isopropanol to yield 250 mg of greenish bronze crystals. These are transformed into liquid crystals at 87 °C and pass into the isotropic fluid state at 130 °C.

6. Formazan Dyes

Formazan dyes are derived from the following basic structure:

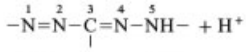

(N^1 meso N^2)

The position of the substituents is given according to IUPAC nomenclature using the prefixes 1-, 3-, and 5-; the prefixes N^1, meso, and N^2 are

sometimes used. Formazans unsubstituted in the 1- and 5-positions and 1,5-dialkyl-substituted formazans are unknown. Aryl or heteroaryl groups are the most common 1,5-substituents. The 3- or meso position can be occupied by a variety of substituents (e.g., aryl, heteroaryl, H, OH, S-alkyl, halogen, NO_2, CN, and alkyl).

The 1,5-diphenyl and 1,3,5-triphenyl derivatives were discovered almost simultaneously by H. v. PECHMANN [153] and E. BAMBERGER [154] in 1892. Interest in these compounds was renewed in 1939 when G. LAKON [155] and R. KUHN and D. JERCHEL [156] found that colorless tetrazolium salts were very sensitive to biological reduction processes and were converted to deeply colored formazans:

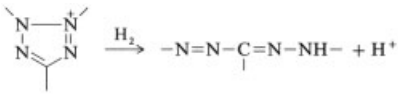

In 1941 L. HUNTER and L. B. ROBERTS produced metal complexes from formazan compounds [157]. R. WITZINGER and coworkers introduced other complex-forming groups into formazans [158] which ultimately led to the discovery that metallized formazans could be used for textile dying. These chromophores have become important reactive dyes for cotton. Formazan dyes are also used in analytical chemistry because of the high color intensity of many of the metal complexes. The chemistry of formazans was first exploited in 1962 by MACDONALD to produce color photographs [159]. Early reviews of formazans and their coordination compounds can be found in [160–166].

6.1. Physical Properties and Structure

6.1.1. Metal-Free Formazans

With different aryl substituents (Ar, Ar′) at the 1- and 5-positions, the tautomers **71** and **72** are imaginable but have not yet been detected:

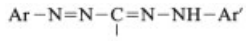

71

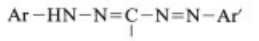

72

Each of these tautomers has four possible stereoisomers **73–76**:

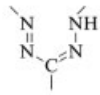

syn-trans
73

syn-cis
74

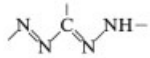

anti-trans
75

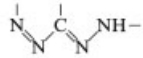

anti-cis
76

Irradiation with visible light leads to rearrangement of many formazans from the stable red syn-trans form **73** into the yellow anti forms **75** and **76**. The anti forms exist only in solution; in the dark they revert to the closed syn-trans form **73** [167].

No stable isomers of triphenylformazan [*49624-20-6*] have been found. The compound has no vibrational bands in the IR spectrum between 2500 and 3200 cm^{-1} [168] and does not form salts in an alkaline medium. Its structure can be described as a quasi-aromatic chelate or as a resonance hybrid with the limiting structures **77** and **78** [167].

Alkyl substitution at the 3-position weakens the intramolecular hydrogen bond, hence 3-methyl-1,5-diphenylformazan exists in the open, syn-trans configuration **79** [169]:

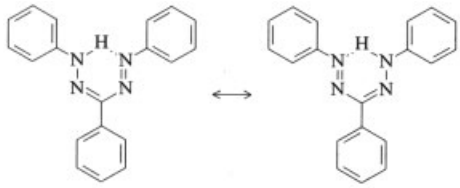

77 **78**

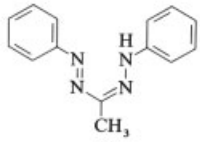

79

This is recognized by the NH-vibrational band at 3360 cm^{-1} and the ease of salt formation in aqueous alkaline solution.

H. ZIEGLER investigated the influence of structure on the absorption maximum. He discovered that a bathochromic shift occurred when (1) electron-donating substituents were introduced at the 1-, 3-, and 5-positions, (2) on reaction with strong bases, and (3) above all on formation of chelates with metal ions. He attributed this effect to the formation of an alternately polarized six-membered ring,

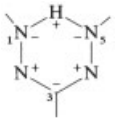

whereby reduction of polarization leads to a hypsochromic shift and an increase to a bathochromic shift [170].

Polarization can also be modified by means of heat. Positive thermochromism is observed with benzothiazolylformazans (**80**): in cold ethanolic solution they are red and at 60 °C violet – blue [171]. The UV – visible spectrum of the blue form corresponds to that of the sodium salt

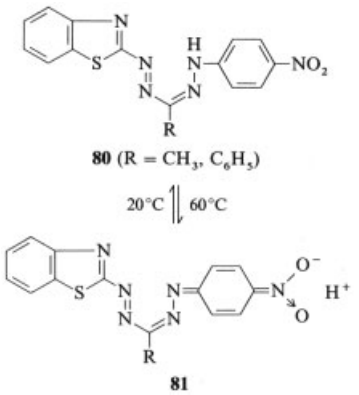

and is therefore assumed to have the anionic structure **81** [172].

In contrast, benzimidazolylformazans **82** exhibit negative thermochromism in oxygen-free solvents. On cooling to $-10\,°C$ in chloroform, for example, the color of the formazan with $R = CH_3$ and $R^1 = C_2H_5$ changes from violet ($\lambda_{max} = 545$ nm) to blue ($\lambda_{max} = 650$ nm). Deepening of color on cooling is due to hydrogen bond formation [173]:

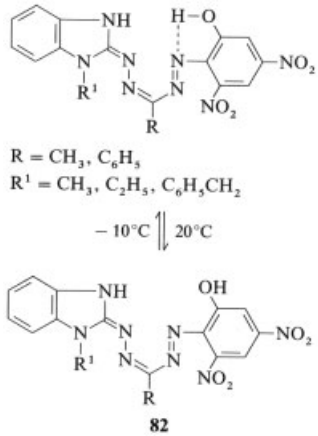

6.1.2. Metal-Containing Formazans

Bidentate Formazans. The most outstanding property of formazans is their ability to form coordination compounds with metals. They behave as monovalent, bidentate ligands and form neutral 1 : 2 metal chelates with divalent metals:

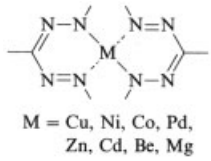

Opinions about the geometry of these compounds vary. A planar structure has been proposed on account of their magnetic properties [174]. However, with 1 : 2 Ni(II) complexes their partial resolution into mirror-image isomers with (+)-quartz and their diamagnetism suggest a highly nonplanar structure [175].

Reaction with Cu(II) salts can produce various compounds depending on the reaction conditions: 1 : 2 Cu(II) chelates [157], Cu(I) complexes with 1 : 1 [176] and 1 : 2 stoichiometry [177], and the corresponding tetrazolium salts. The very low magnetic moment (1.4 BM, Bohr magneton) for the Cu(II) complex produced from triphenylformazan [157] suggests that it is a mixture of **83** (1.8 BM) and **84**, which does not have a magnetic moment.

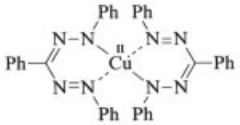

83 [*53708-91-1*]

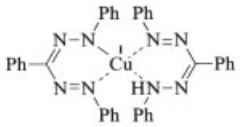

84

Tridentate Formazans. Formazans substituted with OH or COOH in the 2-position of the N_1- or N_5-aryl group have the same complexing properties as 2-hydroxy- (or carboxy)-2′-amino diaryl azo dyes. Similarly they also form 1 : 1 complexes with four-coordinate metals and 1 : 2 complexes with six-coordinate metals. As an amino system these formazans react more readily with cobalt salts than with chromium salts. The mostly blue cobalt(III) 1 : 2 complexes of type **85** [178] and the mostly gray-blue complexes of type **86** [179] are known:

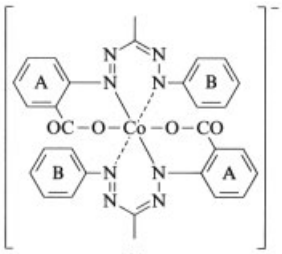

85

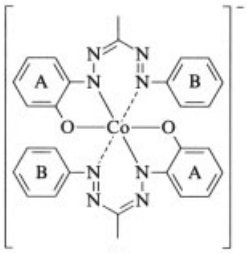

86

Rings A and B may contain other substituents.

As with the tridentate azo compounds, ligands coordinated in 6/6-rings of type **85** are arranged facially and those coordinated in 5/6-rings of type **86** meridionally. In the formazan system, unlike the azo system, the plane of the anellated 6/6-rings in the octahedron stays almost planar. Twisting of the aromatic groups does not occur during metallization and there is therefore no loss of color intensity. Compounds of type **85** have an even larger absorption area in the UV – visible spectrum than their metal-free precursors [178].

Tridentate formazans form 1 : 1 complexes with divalent metal salts:

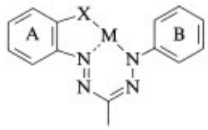

X = COO, O

These coordinatively unsaturated complexes can undergo coordination reactions with compounds having a free electron pair (particularly pyridine, amines, ammonia, and water). Coordination is associated with a marked hypsochromic color change. For example, the ligand-free nickel complex **87** is green and its pyridine adduct **88** is violet [180].

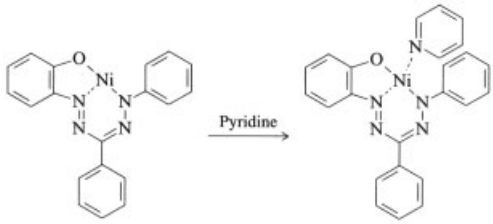

87 **88**

The coordinatively unsaturated complexes are even less soluble in organic solvents (e.g., diethyl ether, benzene, chloroform) than their saturated equivalents. The magnetic moment of copper complexes of formazans with a 2-hydroxyaryl substituent at N_1 is very low (1.44 BM) compared to the expected value for copper(II) complexes with an unpaired electron and strong covalent bonding. The low solubility and magnetic moment suggest that dimerization occurs with the formation of a saturated dinuclear complex **89**:

The saturated pyridine adduct **90**, which does not dimerize, has a magnetic moment of 1.7 BM. The copper complex **91** has an even higher magnetic moment of 1.81 BM [181].

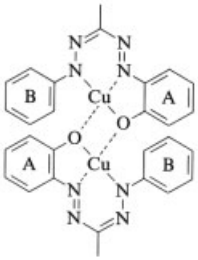

89

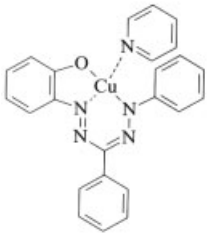

90

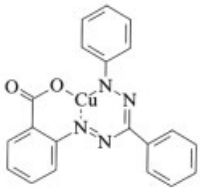

91
[*69239-84-5*]

Quadridentate Formazans. Formazans with two complex-forming groups in the 2- and 2′-positions of the N_1- or N_5-aryl substituents are potential quadridentate ligands. They form 1 : 1 coordination complexes with four- and six-coordinate metals:

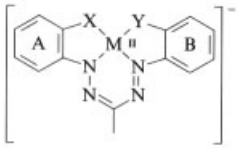

X, Y = O, COO

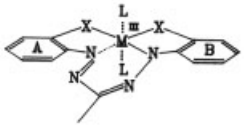

L = H_2O, NH_3, pyridine

The complexes assume a tetragonal planar structure so that two ligands can occupy free coordination sites at the apices of an octahedron.

WIZINGER and coworkers, who first produced the nickel and copper complexes [160], were not sure whether the complexes were of a bicyclic or tricyclic nature. The 1,5-bis(2-hydroxyphenyl)-3-cyanoformazan derivative of the copper complex **92** probably exists as a tricyclic structure because of the increased acidity of the phenolic hydroxyl group.

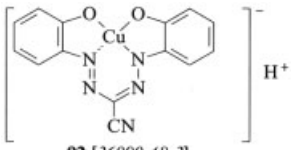

92 [*36090-18-3*]

Complexes with one or two carboxylic groups in the 2- or 2,2′-positions of the aryl substituents, however, appear to be bicyclic. Thus, a hydrate was obtained on reaction of 1,5-bis(2-carboxyphenyl)-3-cyanoformazan with copper acetate [182]. The coordinated water could not be removed by drying and produced a broad band in the IR spectrum at 3450 cm^{-1}. The copper complexes of formazans containing carboxyl groups **93** were then assigned the following structure:

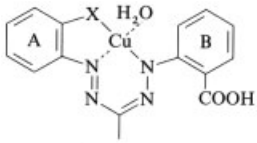

X = O, COO
93

In the patent literature there are indications that the structure of copper complexes of the 2-hydroxy-2′-carboxylic acid type can be modified. After coupling and metallization, lightfastness, color intensity, and color quality can be improved by stirring the reaction mixture for 2 h at low pH [183]. The copper complex can be stabilized by refluxing for 15 h [184]. The coordinatively bound water may possibly be displaced by the carboxyl group in the 2-position during this aftertreatment, converting the dye to the tricyclic form **94** with a 5/6/6-ring system. Maybe a rearrangement proceeds simultaneously, in the course of which the coordinative bound migrates from N^1 to N^2 of the formazan structure to give the less strained form with a 6/5/6-membered ring configuration (**95**):

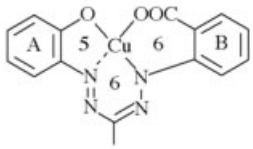

94

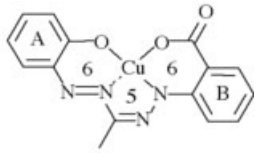

95

Another example of a tricyclic formazan structure with two six-membered rings and one five-membered ring is shown below (**96**), written in the customarily used 5/6/6-mode [226]:

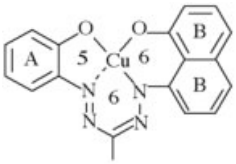

96

Formazan-like structures with a 6/5/6-system are exemplified by **97** [227] and **98** [228]:

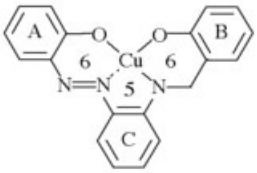

97

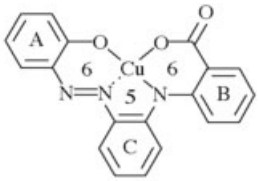

98

6.2. Production

6.2.1. Coupling of Diazonium Compounds with Hydrazones

The most common method for the production of formazans is the coupling of diazonium salts with aryl hydrazones in an alkaline medium, possibly an organic solvent:

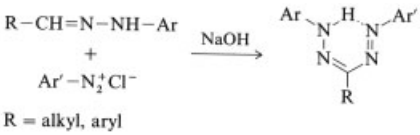

For preparation, see [185, example 1]. The industrially important 1-(2-hydroxyaryl)-3-aryl-5-(2-carboxyaryl)formazans are obtained by this process. The corresponding bis(2-hydroxy) compounds cannot be produced by this method because 2-hydroxyaryl hydrazines are very unstable compared to 2-carboxyaryl hydrazines. For the production of bis(2-hydroxy)formazans, see Section 6.2.3.

6.2.2. Metallization

Bidentate formazans that are insoluble in water can be warmed with cobalt, nickel, and copper salts (preferably acetates) to form the appropriate metal chelates in solvents such as methanol, ethanol, acetone, and dimethylformamide.

Metal complexes of tri- and quadridentate formazans are much more stable. Metallization with divalent metal salts occurs rapidly at room temperature. On reaction with diazotized 2-aminophenols or 2-aminonaphthols, coupling and metallization with divalent metal salts can take place concurrently under the same conditions.

When coupling is complete, the dye is usually fully metallized: warming for a short time at 80 °C is only necessary in rare cases (see also Section 6.1.2). For production of cobalt(III) 1 : 2 complexes, a cobalt salt is generally added after coupling and warmed to 60 – 90 °C in a weakly alkaline medium. Reaction with chromium requires more severe reaction conditions; for example, the formazan and chromium formate are refluxed for 20 h to produce **99** [186]:

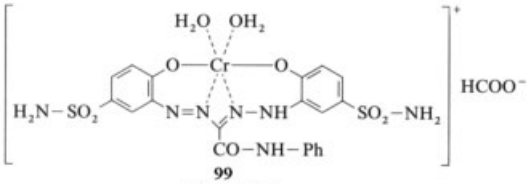

99
[100031-36-5]

Synthesis of the Copper Complex 100 [187, example 1]:

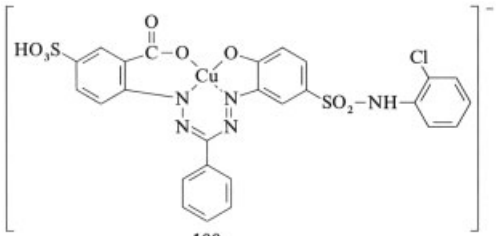

100

2-(3-Amino-4-hydroxyphenylsulfonamido) chlorobenzene (0.1 mol) is dissolved in 100 mL hot water, to which 21 mL of 10 mol/L NaOH solution is then added. After addition of 0.1 mol sodium nitrite, the solution is cooled to − 5 °C and then added to a well-stirred mixture of 42 mL concentrated hydrochloric acid, 100 mL water, and 50 g ice. After diazotization, excess nitrous acid can be destroyed with amidosulfonic acid. The diazonium salt suspension is neutralized with sodium hydrogen carbonate and combined with the weakly acidic solution of 0.1 mol of the hydrazone formed from 2-hydrazino-5-sulfobenzoic acid and benzaldehyde in 100 mL water. The reaction mixture is kept at 0 – 10 °C, and at pH 10 – 12 by dropwise addition of 10 mol/L NaOH. When coupling is complete, the pH of the brown formazan solution is reduced to 8 with hydrochloric acid, a dilute aqueous 0.1 mol/L solution of copper tetraaminosulfate is added and the solution is stirred for several hours. The copper complex formed is precipitated at 85 – 90 °C by the addition of sodium chloride, then filtered and dried. The powder produced is dark and colors wool in clear blue, fast tones.

6.2.3. Double Coupling of Diazonium Compounds with Compounds Containing an Acidic CH Group

Coupling with Compounds Containing an Activated Methyl Group. Some compounds with an activated methyl group can form symmetric formazans with strong diazo compounds by double coupling:

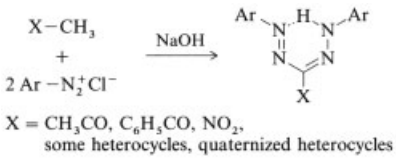

Examples of heterocycles with an activated methyl group are (I [188], II [189], III [190], IV [191]):

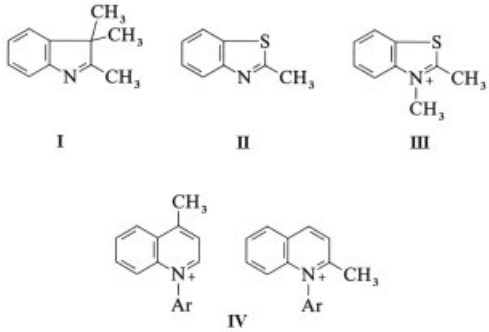

The formazans **101** produced from 2-methylbenzothiazolium salts exist as cations only in strongly acidic solution. When the pH is increased, they are converted into the uncharged olefinic disazo compounds **102**, and for the reaction with metal salts into the cationic formazan complexes **103** [190]:

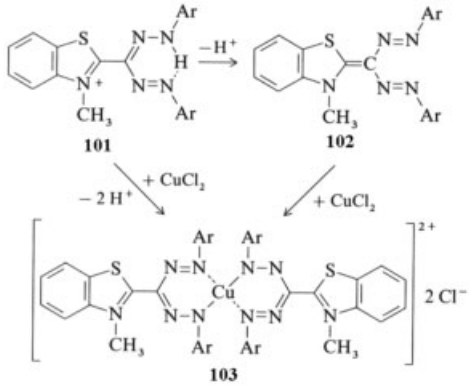

Coupling with Compounds Containing an Active Methylene Group. During coupling of many methylene groups linked to two activating groups X and Y, the activating group X is exchanged for an azo group in the second coupling step. If the first coupling step is easier than the removal of the activating group X, then unsymmetric formazans can be synthesized by coupling the primary hydrazone formed with another diazonium compound:

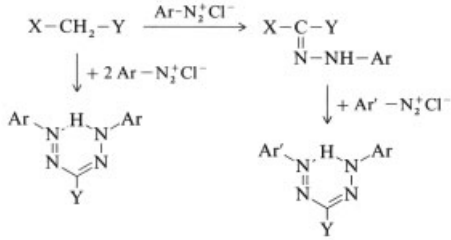

X = COOH, $CONH_2$, CHO, CO-alkyl, CO-aryl, COCOOH, SO-CH_3
Y = X, COO-alkyl, CN, C = N-NH-aryl, SO_2-CH_3, SO_2-aryl

Only symmetrical formazans can be obtained, for example, from (I [192], II [193], III [194]):

I II

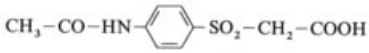

III

and from 4-oxopentanoic acid (levulinic acid), which couples to form the bisformazan **104** [195], [229]:

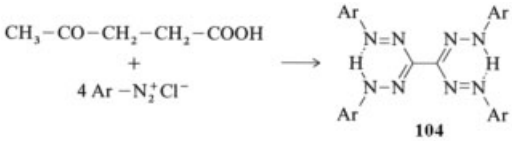

104

Unsymmetrical formazans can be prepared from the following compounds by stepwise coupling, preferably at differing pH values (I [193], II [196], III [197], IV [230]):

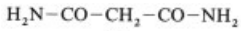

I

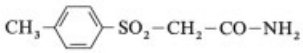

II

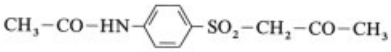

III

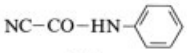

IV

Coupling with Compounds Containing an Active Trisubstituted Methane Group. Doubly activated methane compounds with aryl substituents can also be converted to formazans by elimination of both of the activating groups:

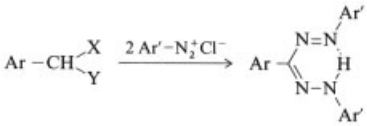

X = COOH, $CONH_2$, CHO, CO-alkyl, CO-aryl
Y = X, CN, COO-alkyl

H. ZIEGLER used coupling with diazotized 2-aminophenols to synthesize 1,5-di-2-hydroxyaryl-3-phenylformazans which are difficult to synthesize by the hydrazone method (see Section 6.2.1).

Ester and cyano groups generally resist substitution by arylazo groups and must first be hydrolyzed to produce carboxyl and carbonamide groups respectively, which can be more easily eliminated. Unsymmetrical formazans can be produced according to the following scheme :

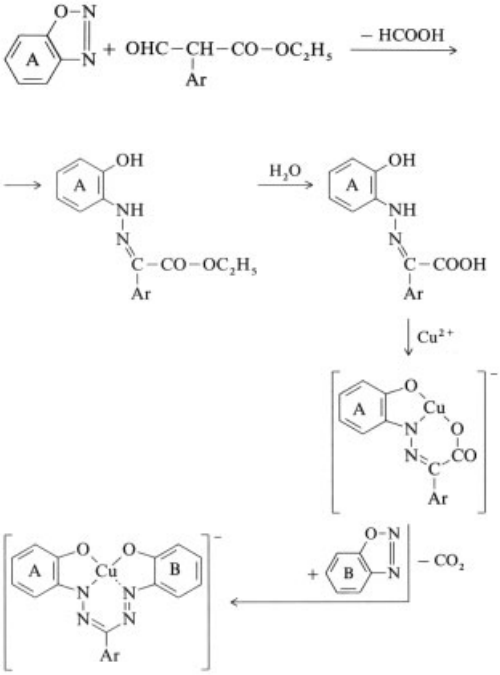

Synthesis of the Copper Complex 105 [198, example 1]:

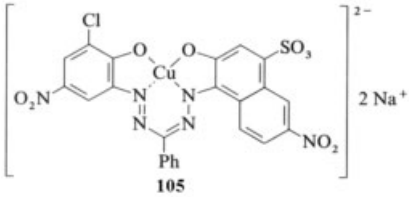

105

Ethyl phenylacetate (0.1 mol) and 14 mL 10 mol/L NaOH are stirred together at 0 °C in 300 mL water in the presence of an emulsifier. 14 g sodium carbonate and the diazonium salt from 0.1 mol 2-amino-6-chloro-4-nitrophenol are added and are coupled at < 25 °C. When the reaction is complete, the solution is neutralized with 10 mol/L hydrochloric acid, diluted with an equal volume of water, and brought to the boil; 80 mL 10 mol/L NaOH are added. The mixture is saponified for 10 min at 95 – 100 °C and neutralized again with 10 mol/L HCl. After cooling to 20 °C, the diazonium salt of 0.1 mol 1-amino-6-nitro-4-sulfo-2-naphthol is coupled in the presence of 0.1 mol copper sulfate and 26 g sodium carbonate. The coupling reaction is completed by heating to 60 °C. The resulting formazan is precipitated at 80 °C, filtered and dried. The dark powder colors wool, natural silk, and polyamide fibers an intense green tone with a blue cast.

A special process for the production of copper-containing 1,5-bis(2-hydroxyphenyl)formazans is based on the demethylating copperization of 1,5-bis(2-methoxyphenyl)formazans [199]. The method involves heating for a short time in pyridine, formamide, or dimethylformamide (see below).

In another process, the copper complexes are produced with alkylsulfonylamine as the complex-forming group [200].

Syntheses of formazans under phase-transfer conditions starting from hydrazones or from carbonyl compounds with active methylene groups are also possible. Azo coupling with aryldiazonium salts in two liquid phases occurs faster and in a better yield when catalyzed either by tetrabutylammonium bromide or dicyclohexano-18-crown-6 [231].

6.2.4. Oxidative Coupling of Azole Hydrazines with Hydrazones

Oxidative coupling of azole hydrazines with hydrazones takes place spontaneously in the presence of pyridine in air and is confined to the synthesis of heteroarylformazans, for example [201]:

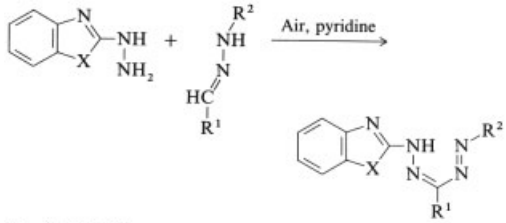

X = S, N-alkyl
R^1 = alkyl, aryl, heteroaryl
R^2 = aryl, heteroaryl

Alkylation of 1,5-diheteroaryl-substituted formazans produces dehydroformazans with a cationic structure [202]:

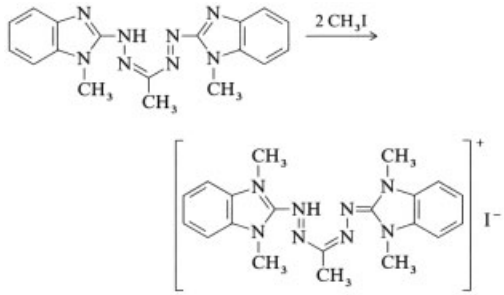

This type of dye can also be prepared by oxidative coupling of azines with hydrazones [203].

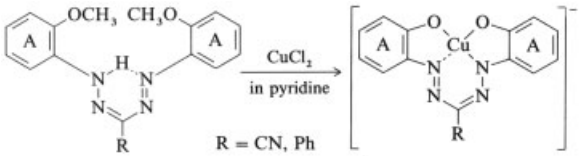

6.3. Uses

Textile Dyes. Many metallized formazans are well suited as dyes for textile materials on account of their coloring strength and good fastness. The blue to blue – green copper complexes of quadridentate formazans are particularly important because of their color clarity, which is similar to that of anthraquinone dyes. They can also be combined easily with yellow, orange, and red azo dyes and therefore present a low-priced alternative to blue anthraquinone dyes for trichromic dying [204]. Copper complexes from quadridentate formazans are preferred on account of their high stability. Some dicyclic copper complexes of tridentate formazans also show good fastness and good application properties [205].

Wool and Polyamides. Water-soluble formazan complexes containing sulfonamide, alkylsulfonyl, or sulfonic acid groups possess a high affinity for nitrogen-containing fibers. As with acid dyes, they can be used to dye wool and polyamide in a neutral to weakly acid bath. Commercial dyes of this type are C.I. Acid Blue 267 and 297, and the 1 : 2 cobalt complex C.I. Acid Black 180, 13 710 **106**.

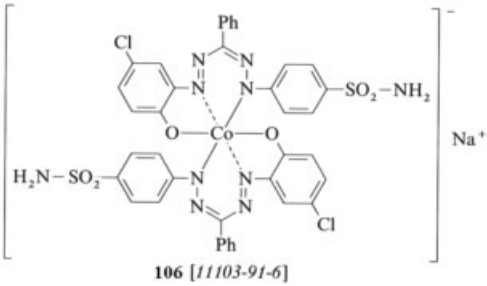

106 [*11103-91-6*]

As environmentally friendly alternatives to the Cr(III) and Co(III) metal complex acid dyes a range of 1:2 iron complex formazan dyes **107** was synthesized [232], which produce violet, blue, black and brown shades, furnishing dyeings on wool and polyamide with good wet and light fastness.

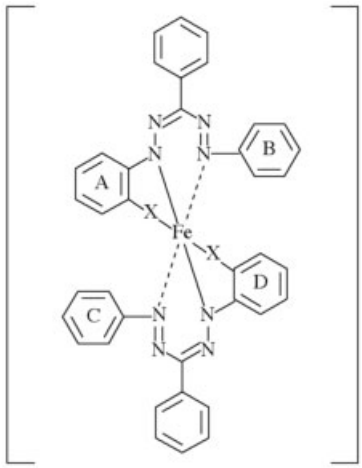

X = O, COO

107

Cotton. The most important formazan dyes are those containing reactive groups for dyeing and printing cellulose fibers. They are usually blue, occasionally green [206] copper complexes derived from 1-(2-hydroxyphenyl)-3-phenyl-5-(2-carboxyphenyl)formazans containing sulfonic acid and reactive groups. The following commercial products belong to this group: C.I. Reactive Green 15, Reactive Blue 70, 83, 84, 104, 157, 160, 182, 202, 209, 212, 216, 218, 220, 221, 226, 228, 235. The increasing importance of blue formazan dyes is mostly due to their suitability as a cheap constituent of triple dyes. The number of patents issued involving formazan dyes containing reactive groups was nearly 100 between 1985 and 1989 and nearly 65 between 1990 and 1994.

Analytical Reagents. As a result of their fast and intense color reaction, formazans can be used for the spectroscopic determination of a variety of metal ions. In a sensitive blue color reaction, Zn and Cu can be determined with the commercial product Zinkon [*135-52-4*] **108**. The detection limit for Zn ($\lambda_{max} = 625$ nm) is 0.02 ppm.

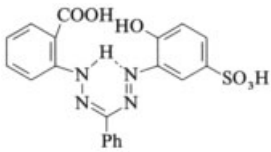

108 [*135-52-4*]

Some heterocyclic-substituted formazans are described as being very specific reagents. For example, **110** can be used for the detection of Co or Ni depending on the pH [207]; with **111**, depending on the pH, Sc or Sc and Y can be detected [208]. The formazan **112** is a reagent for Cu, Co, and Cd [209]; **113** a reagent for Hg [210]; **109** a selective reagent for Co [211]; and **114** a reagent for Cu and Ag [212]. Cu, Co, Ni, Cd, and Zn can be detected separately by colorimetry using **115** [213]. Microamounts of nickel in steel samples can be determined by spectrophotometry using **116** [233]. The anthryl formazans **117** give red Zn chelates and bluish green Cd chelates having a remarkably large difference of 110 nm in the wavelengths of their absorption maxima [234]:

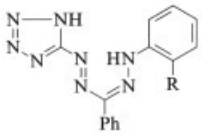

109 [*51475-08-2*] R = H
110 [*78733-47-8*] R = SO_3H

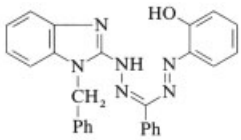

111 [*94816-77-0*]

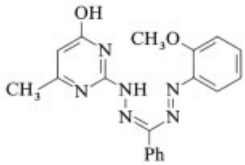

112 [*92829-23-7*]

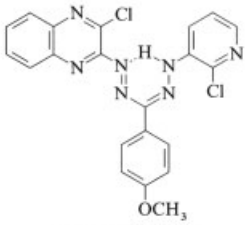

113 [*114546-65-5*]

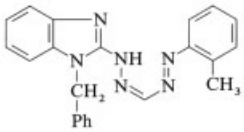

114 [*37704-00-0*]

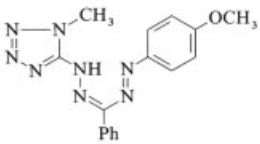

115 [*72607-20-6*]

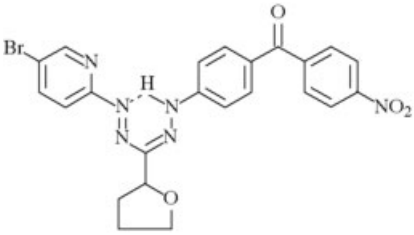

116 [157078-30-3]

Bioindicators. Water-soluble, colorless tetrazolium salts can be reduced to water-insoluble, deeply colored formazans. The reduction of tetrazolium salts in plant tissue at pH 7.2 was first demonstrated in 1941 [156]. Tetrazolium salts have since been used in biochemistry, cytochemistry, and histochemistry because of the irreversibility and great sensitivity of this reaction. They can be used to detect biological redox systems in blood serum, in living cells, tissues, tumors, and bacteria. Tetrazolium Blue [*1871-22-3*] **118** is a particularly sensitive reagent.

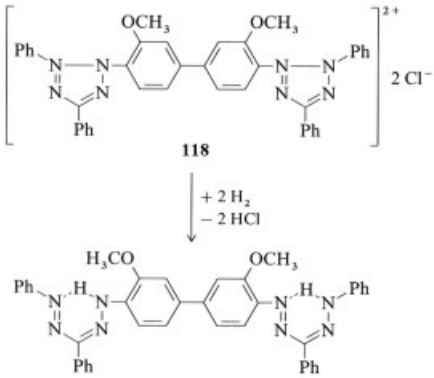

118

Indicator reactions are reviewed in [214], [215]. A more recent patent describes how cyanotetrazolium salts can be converted into fluorescent formazans for cytological diagnosis [216]. Another area of application is the fast determination of bacterial capacity in industrial wastewaters [217] and the determination of dehydrogenase activity in sludges [218].

Photoreagents. The conversion of tetrazolium salts into metallizable formazan dyes is used in color photography. Exposure of a photographic film to light results in the formation

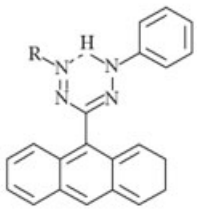

R = 2-pyridyl, 2-quinolyl

117

of silver from a silver salt. When the film is developed with sodium thiosulfate in the presence of a tetrazolium salt, for example, a formazan is formed with the silver metal. The formazan is stabilized with transition metals and optically amplified [219]. In a silver-free method, the formazan is developed physically in a high-gain amplification process. The color of the photograph can be modified by varying the tetrazolium salt [220]. In another physical development process, the silver produced on the film is also amplified with formazan dyes [221]. Formazan dyes are also effective in photographic emulsions as antifogging agents [222]. Structure **119** represents discoloration-resistant photochromic formazan dyes, whose color change from red to yellow by irradiation wih a xenon lamp [235]:

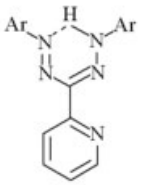

119

Other Uses. Formazan dyes can also be used in many other areas. For example, porous membranes suitable for ultrafiltration can be modified with the reactive dye **120** [223]. Formazans of the formula **121** are used as dyes in inks for ink jet printing [224].

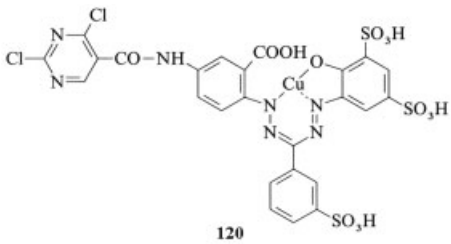

120

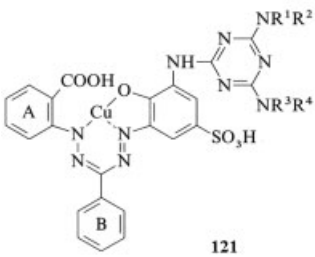

121

The cyclic formazans **122** are lithium-specific indicator dyes, useful for the quantitative colorimetric determination of lithium in biological fluids, e. g., blood [236]:

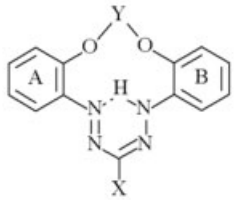

X = CN, NO_2
Y = CH_2-CH_2, CH_2, -CH_2-CH_2, CH_2-CH_2-O-CH_2-CH_2

122

Formazans are also used for transfer printing whereby the formazan derivative is on the transfer coating and the receiver coating contains zinc oxide as a chelating agent [225].

7. References

General References

1. A. E. Martin, M. Calvin: *Die Chemie der Metallchelatverbindungen,* Verlag Chemie, Weinheim 1958.
2. H. Zollinger: *Chemie der Azoverbindungen,* Birkhäuser Verlag, Basel – Stuttgart 1958.
3. P. Rys, H. Zollinger: *Leitfaden der Farbstoffchemie,* Verlag Chemie, Weinheim 1970.
4. K. Venkatavamann: *The Chemistry of Synthetic Dyes,* **Academic Press, New York; (a) vol. I (1952)** pp. 569 – 934; **(b) vol. III (1970)** pp. 303 – 383; **(c) vol. V (1971)** pp. 241 – 282; **(d) vol. VI (1972)** pp. 264 – 323; **(e) vol. VIII (1978)** pp. 44 – 73, 93 – 109, 371 – 379.
5. H. Baumann, H. R. Hensel: "Neue Metallkomplexfarbstoffe, Struktur und färberische Eigenschaften," *Fortschr. Chem. Forsch.* **7** (1967) 643 – 783.
6. *Houben-Weyl,* **10/3,** pp. 434 – 463.
7. *Houben-Weyl,* **6/2,** pp. 41 – 53.
8. H. Gerstner: *Die Chemie der Applikation von Komplexfarbstoffen,* Academie Verlag, Berlin 1959.
9. R. Price: "Some Aspects of the Chemistry of Metal Complex Dyestuffs," *Chimia* **28** (1974) 221.

10. F. Beffa, E. Steiner, *Rev. Prog. Color. Relat. Top.* **4** (1973) 61.
11. P. F. Gordon, P. Gregory: *Organic Chemistry in Colour,* Springer Verlag, Berlin – Heidelberg – New York 1983.
12. F. Beffa, G. Back, *Rev. Progr. Color. Relat. Top.* **14** (1984) 33.
13. H. Zollinger: *Color Chemistry,* VCH Verlagsgesellschaft, Weinheim 1987.

Specific References

14. G. Schwarzenbach, *Helv. Chim. Acta* **35** (1952) 2344.
15. Geigy, DE-AS 2 019 827, 1970 (G. Schetty, E. Steiner).
16. H. D. K. Drew, R. E. Fairbairn, *J. Chem. Soc.* (1939) 823.
17. H. Pfitzner, *Angew. Chem.* **62** (1950) 244.
18. P. Pfeiffer, S. Saure, *Ber. Dtsch. Chem. Ges.* **74** (1941) 935.
19. R. Grieb, A. Niggli, *Helv. Chim. Acta* **48** (1965) 317.
20. G. Schetty, *Helv. Chim. Acta* **45** (1962) 809; **46** (1963) 1132.
21. G. Schetty, W. Kuster, *Helv. Chim. Acta* **44** (1961) 2193. G. Schetty, *Helv. Chim. Acta* **45** (1962) 1095, 1473; **47** (1964) 921; **49** (1966) 49; **50** (1967) 15; *Chimia* **18** (1964) 244.
22. H. Jaggi, *Helv. Chim. Acta* **51** (1968) 580.
23. H. Pfitzner, *Angew. Chem.* **84** (1972) 351; *Angew. Chem. Int. Ed. Engl.* **11** (1972) 312.
24. G. Schetty, E. Steiner, *Helv. Chim. Acta* **57** (1974) 2149.
25. G. Schetty, E. Steiner, *Helv. Chim. Acta* **58** (1975) 1309.
26. F. Feichtmayr, H. Pfitzner, *Liebigs Ann. Chem.* 1980, 2055.
27. H. Pfitzner, H. Schweppe, *Z. Anal. Chem.* **268** (1974) 337. H. Pfitzner, *Melliand Textilber.* **58** (1977) 314.
28. G. Schetty, *Helv. Chim. Acta* **53** (1970) 1437. G. Schetty, E. Steiner, *Helv. Chim. Acta* **55** (1972) 1509.
29. J. A. Connor, D. J. Fine, R. Price, *J. Chem. Soc., Dalton Trans.* **2** (1981) 559. P. Ball, C. H. Nicholls, *Dyes Pigm.* **3** (1982) 5.
30. H. Pfitzner, H. Baumann, *Angew. Chem.* **70** (1958) 232. BASF, DE 807 289, 1948 (H. Pfitzner, H. Merkel).
31. I.G. Farbenindustrie, DE 571 859, 1931 (D. Delfs, R. Knoche); DE 658 841, 1934 (D. Delfs); DE 738 900, 1936 (D. Delfs).
32. BASF, DE 843 275, 1949 (H. Pfitzner, H. Merkel).
33. C. V. Stead, *Rev. Progr. Color. Relat. Top.* **1** (1970) 23.
34. ICI, FR 1 431 264, 1965.
35. Geigy, DE-AS 1 544 583, 1965 (P. Lienhard, F. Beffa, C. Soiron).
36. Sandoz, DE-AS 2 153 548, 1971 (J. Dore).
37. *Friedländer,* **21,** 919; **22,** 897.
38. Bayer, DE 929 567, 1952 (G. Dittmar, R. Pütter). *Chem. Abstr.* **50** (1956) 13, 459 e.
39. BASF, DE 479 373, 1926 (H. Krzikalla, W. Müller). *Friedländer,* **16,** 975.
40. BASF, DE 715 082, 1938 (K. Holzach, H. Pfitzner); DE 715 948, 1938 (K. Holzach, H. Pfitzner).
41. BASF, DE 742 939, 1938 (H. Pfitzner).
42. J. P. M. Ros, *Textilveredelung* **21** (1986) 352. B. L. Bangham, T. A. Perenich *Am. Dyest. Rep.* (1988) 19.
43. R. Anliker, E. A. Clarke, *J. Soc. Dyers Colour.* **98** (1982) 42. R. Anliker, *Textilveredelung* **16** (1981) 431.
44. R. Anliker, E. A. Clarke, P. Moser, *Chemosphere* **18** (1981) 263.
45. D. Brown, H. R. Hitz, L. Schäfer, *Chemosphere* **10** (1981) 245.
46. E. A. Clarke, R. Anliker, *Rev. Prog. Color. Relat. Top.* **14** (1984) 84. R. Anliker, *J. Soc. Dyers Colour.* **95** (1979) 317. R. Anliker, G. Düring, D. Steinle, E. I. Moricon, *J. Soc. Dyers Colour.* **104** (1988) 223.
47. R. Anliker, D. Steinle, *J. Soc. Dyers Colour.* **104** (1988) 377.
48. A. J. Hinton, *Rev. Prog. Color. Relat. Top.* **18** (1988) 1.
49. U. Sewekow, *Melliand Textilber.* **69** (1988) 271.
50. Z. Gorzka, J. Kraska, H. Lawniczak, *Dyes Pigm.* **5** (1984) 263.
51. Hoechst, DE-OS 3 829 265, 1988 (U. Mrotzek et al.).
52. P. A. Duffield, R. D. D. Holt, *Textilveredelung* **24** (1989) 40. P. A. Duffield, K. H. Hoppen, *Melliand Textilber.* **68** (1987) 195. A. C. Welham, *J. Soc. Dyers Colour.* **102** (1986) 126. Sandoz, DE-OS 3 532 386, 1985 (A. C. Welham).
53. K. Roennefahrt, *Melliand Textilber.* **70** (1989) 203. H. Dumser, *Melliand Textilber.* **70** (1989) 206. K. König, *Melliand Textilber.* **70** (1989) 209.
54. J. A. Bone, J. Shore, J. Park, *J. Soc. Dyers Colour.* **104** (1988) 12.
55. BASF, DE 846 142, 1950 (H. Pfitzner, O. Kaufmann); DE 925 725, 1953 (H. Pfitzner, O. Kaufmann).

56. Geigy, DE-OS 1 061 922, 1955 (G. Schetty, H. Ackermann); DE-OS 2 529 798, 1975 (G. Back, F. Beffa, A. Bühler).
57. Ciba-Geigy, EP-A 163 608, 1985 (G. Back, W. Mosimann); EP-A 203 890, 1986 (G. Back, W. Mosimann). SDC Report, *J. Soc. Dyers Colour.* **105** (1989) 103.
58. G. Schetty, *J. Soc. Dyers Colour.* **71** (1955) 705.
59. BASF, DE 734 990, 1940 (J. Müller, H. Pfitzner); DE 748 970, 1942 (G. v. Rosenberg, K. Anacker, H. Pfitzner).
60. Geigy, GB 730 036, 1951; GB 736 034, 1953.
61. Bayer, GB 755 853, 1953; DE 925 725, 1953; DE 758 016, 1957.
62. BASF, DE 846 142, 1950. Bayer, GB 796 759, 1955; GB 823 742, 1957; DE 940 483, 1952 (G. Dittmar, R. Pütter, F. Suckfüll).
63. Ciba, FR 836 297, 1938. Allied Chemicals, US 2 499 133, 1945 (K. F. Conrad). Ciba, DE 850 038, 1950 (W. Widmer, A. Bühler); DE 932 979, 1952 (A. Bühler, C. Zickendraht). Geigy, US 2 623 871, 1951 (G. Schetty, F. Beffa). Sandoz, US 2 753 336, 1953 (P. Maderni). Bayer, FR 1 111 970, 1954. General Anilin, US 2 809 962, 1953 (H. B. Freyermuth, A. F. Strobel). ICI, GB 748 421, 1953 (C. H. Reece). BASF, DE-AS 1 032 865, 1955 (H. Pfitzner, O. Kaufmann). Bayer, GB 796 723, 1956.
64. G. Back, H. Zollinger, *Melliand Textilber.* **37** (1956) 1316; *Chimia* **11** (1957) 103; *Helv. Chim. Acta* **41** (1958) 2242; **42** (1959) 1539.
65. Ciba, DE-AS 1 012 007, 1954 (J. Brassel, A. Bühler). Geigy, DE 1 079 247, 1957 (G. Schetty, W. Biedermann, F. Beffa); DE-AS 1 111 318, 1957 (W. Biedermann, F. Beffa, G. Schetty); DE 1 544 580, 1965 (F. Beffa, E. Steiner); US 3 459 727, 1966 (E. Steiner, W. Biedermann). Bayer, DE-AS 2 210 260, 1972 (W. Mennicke). BASF, DE-AS 2 461 481, 1974 (H. Baumann). Ciba-Geigy, DE 2 620 981, 1976 (F. Beffa, G. Back, E. Steiner). BASF, DE-OS 2 735 287, 1977 (H. Baumann, K. Grychtol).
66. Ciba, DE-AS 1 012 007, 1954 (J. Bvassel, A. Bühler).
67. Ciba-Geigy, EP-A 156 768, 1985; EP-A 157 733, 1985 (W. Iff, D. Parisot, H. Schaller); EP-A 241 414, 1987 (B. Chambon, H. Jäger, R. Oxenius, A. Püntener).
68. Ciba, DE 600 545, 1932.
69. BASF, DE 1 008 253, 1956 (H. Baumann, W. Federkiel). Ciba, DE 1 133 846, 1959 (A. Bühler, A. Fasciati, C. Zickendraht).
70. K. R. Schneider, *Textilveredelungspraxis* **2** (1972) 151.
71. K. Reincke, *Melliand Textilber.* **67** (1986) 191. Hoechst, DE-AS 2 845 536, 1978 (K. H. Röstermundt).
72. D. M. Lewis, *J. Soc. Dyers Colour.* **98** (1982) 165. Ciba-Geigy, EP 159 962, 1985 (G. Back, H. V. Schütz, K. Schlesinger). K. Venkataraman: "Reactive Dyes" in *The Chemistry of Synthetic Dyes,* vol. VI Academic Press, New York – London 1972.
73. Bayer, EP-A 51 188, 1981 (R. Preuss, J. Grütze, U. Beck, H. Pantke, K. Vogel). Sandoz, DE-OS 3 201 381, 1982 (L. Avar, E. Kalt, H. Reinshagen). Ciba-Geigy, EP-A 185 611, 1985 (G. Reinert, H. Schütz, G. Back).
74. BASF, DE-OS 3 010 872, 1980 (H. Baumann, K. Grychtol).
75. Bayer, DE 929 567, 1952 (G. Dittmar, R. Pütter).
76. Hoechst, DE-OS 3 023 664, 1980 (H. Fuchs, K. Filzinger).
77. Ciba-Geigy, DE-OS 3 642 927, 1986 (F. Beffa).
78. BASF, EP 47 427, 1981 (H. Baumann, K. Grychtol).
79. Bayer, DE-OS 2 123 453, 1971 (W. Mennicke).
80. BASF, EP-A 68 242, 1982 (H. Baumann, K. Grychtol).
81. I.G. Farbenindustrie, DE 553 045, 1929 (J. Hilger, K. Wiedemann). *Friedländer,* **18,** 1030.
82. ICI, GB 985 481, 1960 (A. H. Berrie, A. Johnson, C. E. Vellins).
83. BASF, DE-OS 3 007 966, 1980 (J. Dehnert); DE-OS 3 337 591, 1983 (J. Dehnert, W. Jünemann); DE-OS 3 417 237, 1984 (J. Dehnert).
84. Eastman Kodak, US 2 835 661, 1956 (J. M. Straley, J. G. Fisher); US 2 822 359, 1954 (J. M. Straley, J. Sagal); US 2 871 231, 1955 (J. M. Straley, D. J. Wallace, J. G. Fisher).
85. Agfa-Gevaert, EP-A 54 230, 1981 (V. Runzheimer, G. Wolfrum, H. Heidenreich, P. Bergthaller, G. Schenk).
86. W. Beckmann, F. Hoffmann, J. Wolff, *Textilveredlung* **24** (1989) 81. D. Link, J. Moreau, *Textilveredlung* **24** (1989) 87. W. Zysset, *Textilveredlung* **24** (1989) 91.
87. Sandoz, DE-OS 2 816 983, 1978 (J. v. Diest).
88. Bayer, DE-OS 3 208 141, 1982 (K. Linhart, H. Gleinig).
89. Bayer, EP-A 24 609, 1980 (W. Scholl, K. H. Schündehütte, W. Mennicke).

90. Bayer, DE-AS 2 443 483, 1974 (W. Mennicke, P. Suchanek, P. Scotch, P. Vogt). Ciba-Geigy, EP-A 44 805, 1981 (R. Lacroix).
91. Ciba-Geigy, EP-A 88 727, 1983 (P. Erzinger).
92. Ciba-Geigy, EP-A 24 655, 1980 (P. Hugelshofer, B. Bruttel). Ciba-Geigy, EP-A 95 740, 1983 (P. Balliello).
93. Ciba-Geigy, EP-A 82 117, 1982 (G. Back, F. Beffa). BASF, DE-OS 3 212 240, 1982 (U. Bergmann, G. Hansen, G. Zeidler). Ciba-Geigy, EP-A 99 857, 1983 (G. Back). BASF, EP-A 140 252, 1984 (U. Bergmann, G. Hansen, D. Lach, G. Zeidler).
94. BASF, DE-OS 3 503 588, 1985 (G. Hansen, U. Bergmann, G. Zeidler); EP-A 193 893, 1986 (G. Hansen, U. Bergmann, G. Zeidler).
95. BASF, DE-OS 3 428 836, 1984 (U. Bergmann, G. Hansen, G. Zeidler); DE-OS 3 520 393, 1985 (U. Bergmann, G. Hansen, G. Zeidler).
96. Ciba-Geigy, EP-A 61 997, 1982; EP-A 61 998, 1982; EP-A 61 999, 1982 (A. Püntener). Bayer, EP-A 260 561, 1987 (W. Mennicke). Sandoz, DE-OS 3 825 755, 1988 (M. Dien, W. Egli).
97. Ciba-Geigy, EP-A 231 147, 1987 (F. Beffa, J. Koller).
98. Sandoz, DE-OS 3 805 746, 1988 (H. Hollinger).
99. Ciba-Geigy, EP-A 9 466, 1979 (F. Beffa); EP-A 19 851, 1980 (F. Beffa); EP-A 48 693, 1981 (F. Beffa); EP-A 100 297, 1983 (G. Back).
100. Ciba-Geigy, DE-OS 3 516 421, 1985 (U. Schlesinger, F. Beffa).
101. Sandoz, DE-OS 3 313 695, 1983 (J. Doré, R. Pedrazzi); EP-A 92 520, 1983 (M. Greve, H. Moser, R. Pedrazzi, R. Wald); DE-OS 3 609 590, 1986 (H. Moser, R. Wald); DE-OS 3 715 066, 1987 (H. Moser, R. Wald); DE-OS 3 440 777, 1984 (J. Doré, R. Pedrazzi).
102. Sandoz, DE-OS 3 236 238, 1982.
103. Bayer, DE-OS 3 710 077, 1987 (K. Kunde).
104. Sandoz, DE-OS 3 117 127, 1981 (P. Doswald, E. Miriconi, H. Moser, H. Schmid).
105. C. H. Giles, *Rev. Prog. Color. Rel. Top.* **5** (1974) 49. Hodogaya, JP 57 073 052, 1980.
106. VEB Chemiekombinat Bitterfeld, DD 246 549, 1986 (G. Knöchel, G. Bley, W. Niehoff, H. Riedel, G. Stubenrauch).
107. IG Farbenindustrie, BASF, DE 743 848, 1939 (H. Pfitzner, B. Eistert); DE 746 839, 1939 (H. Pfitzner); DE-AS 1 260 652, 1960 (H. Baumann et al.).
108. Williams (Hounslow) Ltd., US 4 204 879, 1980 (K. N. Paskins, D. C. Redman, I. D. Turner).
109. ICI, EP-A 177 138, 1985 (A. G. W. Baxter, S. B. Bostock, D. Greenwood).
110. Canon, EP-A 233 769, 1987 (T. Eida, Y. Suga, K. Shirota).
111. ICI, EP-A 219 232, 1986 (A. G. W. Baxter, S. B. Bostock, D. Greenwood).
112. Canon, JP 63 117 080, 1986 (N. Ota et al.); Derwent Report 88-179 505/26.
113. BASF, EP-A 163 113, 1985 (H. Löffler, R. Dyllick-Brenzinger).
114. ICI, EP-A 246 763, 1987 (A. Quayle, C. V. Stead).
115. Ciba-Geigy, EP-A 183 651, 1985 (P. Lienhard).
116. Bayer, DE-AS 2 064 093, 1970 (K.-H. Schündehütte); EP-A 73 463, 1982 (M. Lorenz, K. H. Schündehütte).
117. Bayer, DE-OS 3 307 508, 1983 (M. Lorenz, K. H. Schündehütte, W. Bornatsch).
118. BASF, DE 1 544 404, 1966 (P. Dimroth, A. Haag).
119. Du Pont, US 2 396 327, 1941 (D. F. Kvalnes, H. F. Woodward).
120. BASF, DE-OS 3 443 465, 1984 (W. Lotsch).
121. Sandoz, DE-OS 3 625 575, 1986 (P. Bitterli, J. Goldmann, L. Bansi).
122. BASF, DE-OS 1 569 666, 1967 (P. Dimroth, H. Hiller). Politechnika Lodzka, PL 123 437, 1981 (J. Kraska, W. Czaijkowski).
123. S. Di Bella, M. Casarin, I. Fragala, G. Granozzi, J. Marks, *Inorg. Chem.* **27** (1988) 3 993.
124. Ciba-Geigy, EP-A 169 167, 1985 (G. Cseh, P. Lienhard).
125. Hoechst, DE-OS 2 850 900, 1978 (T. Papenfuhs, W. Teige).
126. Sun Chem. Corp., US 4 508 899, 1985 (T. R. Chamberlain).
127. Ciba-Geigy, EP-A 20 299, 1980 (A. Iqbal, P. Lienhard, A. Pugin).
128. Ciba-Geigy, EP-A 36 834, 1981 (A. Iqbal, P. Lienhard).
129. Ciba-Geigy, EP-A 74 925, 1982 (F. Felix, P. Lienhard).
130. Bayer, EP-A 172 512, 1985 (R. Neeff, M. Rolf, W. Müller).
131. Bayer, DE-OS 3 537 625, 1985 (R. Neeff, M. Rolf, W. Müller).
132. D. St. C. Black et al., *Aust. J. Chem.* **35** (1982) 2435.
133. BASF, DE 2 138 931, 1971 (H. Kast, H. Baumann, U. Mayer, A. Oberlinner); DE 2 152 703, 1971 (H. Kast, H. Baumann, A. Oberlinner). F. Beck, *Ber. Bunsen-Ges. Phys. Chem.* **77** (1973) 353.

134. Friedrich-Schiller-Universität Jena, DD 217 519, 1983 (D. Seidel, E. Jäger).
135. Hodogaya, DE 3 005 397, 1980 (M. Nagata, K. Wakasugi).
136. G. Schwarzenbach, W. Biedermann, *Helv. Chim. Acta* **31** (1948) 678. B. Bruttel, E. Ruhlmann, C. Wittwer, *Chimia* **28** (1974) 56.
137. Agency of Ind. Sci. Tech., JP 57 136 570, 1981; Derwent Report 82 659.
138. Polaroid Corp., US 3 970 616, 1973 (H. Bader, M. H. Feingold); US 4 206 115, 1977 (E. M. Idelson); US 4 231 950, 1979 (E. M. Idelson); US 4 656 117, 1982 (E. Chinoporos, E. M. Idelson, P. F. King); EP-A 149 158, 1984 (P. F. King, S. G. Stroud).
139. Polaroid Corp., US 4 267 252, 1980 (E. M. Idelson).
140. Agfa-Gevaert, DE-OS 3 322 058, 1983 (R. Stolzenburg, P. Bergthaller, G. Wolfrum, J. Strauss); DE-OS 3 329 774, 1983 (P. Bergthaller, G. Wolfrum, H. Heidenreich); DE-OS 3 337 118, 1983 (G. Wolfrum et al.). Eastman Kodak, EP-A 95 127, 1983 (J. A. Reczek, J. K. Elwood).
141. K. L. Birkett, P. Gregory, *Dyes Pigm.* **7** (1986) 341.
142. Kao Corp., JP 63 199 365, 1987 (Y. Sato, S. Ootani, K. Takemusa, R. Tsushima). Hodogaya, JP 59 228 259, 1983. Xerox Corp., US 4 623 606, 1986 (R. N. Ciccavelli). ICI, EP-A 162 632, 1985 (P. Bamfield, P. Gregory, K. L. Birkett).
143. Hodogaya, EP-A 141 377, 1984 (T. Kurahashi, N. Susuki, K. Motohashi, G. Sugiyama).
144. Hodogaya, EP-A 180 655, 1984 (I. Niimura, K. Motohashi, H. Yamaga, N. Akuzawa).
145. Hodogaya, JP 60 106 859, 1983.
146. Konishiroku Photo, JP 63 217 362, 1987 (T. Okuyama, S. Ikeuchi, A. Matsubava, M. Avai); Derwent Report 88-296 629/42.
147. Fuji Foto, JP 61 011 704, 1984; Derwent Report 86-059 321/09; JP 61 032 003, 1984 (Y. Suzuki, K. Hayashi); Derwent Report, 86-085 065/13; JP 61 057 674, 1984 (Y. Suzuki, K. Hayashi); Derwent Report, 86-116 734/18. Secretary of Navy, US 4 593 113, 1985 (M. H. Kauffman). BASF, DE-OS 3 505 751, 1985 (W. Schrott et al.).
148. Fuji Foto, JP 63 202 492, 1987 (Y. Suzuki, M. Tsuboi); Derwent Report, 88-275 389/39.
149. BASF, DE-OS 3 505 750, 1985 (W. Schrott et al.).
150. Mitsubishi, JP 63 227 569, 1987 (K. Yoshida, Y. Kubo); Derwent Report, 88-310 376/44.
151. Cent. Nat. Rech. Scientifiq, FR 2 486 946, 1980 (A.-M. Giroud-Godquin).
152. Ciba-Geigy, EP-A 162 804, 1985 (M. Hunziker).
153. H. v. Pechmann, *Ber. Dtsch. Chem. Ges.* **25** (1892) 3175.
154. E. Bamberger, E. Wheelwright, *Ber. Dtsch. Chem. Ges.* **25** (1892) 3201.
155. G. Lakon, *Ber. Dtsch. Bot. Ges.* **57** (1939) 191; **60** (1942) 299, 434.
156. R. Kuhn, D. Jerchel, *Ber. Dtsch. Chem. Ges.* **74** (1941) 949.
157. L. Hunter, L. B. Roberts, *J. Chem. Soc.* 1941, 820, 823.
158. R. Wizinger, V. Biro, *Helv. Chim. Acta* **32** (1949) 901.
159. Ilford, GB 908 299, 1962 (E. MacDonald).
160. R. Wizinger, Chimia Suppl. 1968, 82 – 94.
161. A. W. Nineham, *Chem. Rev.* **55** (1955) 355 – 483.
162. R. Pütter, *Houben Weyl* **10/3** (1965) 627 – 694.
163. H. Baumann, H. R. Hensel: "Neue Metallkomplexfarbstoffe, Struktur und färberische Eigenschaften," *Fortschr. Chem. Forsch.* **7** (1967) 714 – 727.
164. H. Iida, *Senryo to Yakuhin* **14** (1969) 39 – 45.
165. R. Price: *Comprehensive Coordination Chemistry,* **vol. 6,** Pergamon Press, Oxford, pp. 35 – 94.
166. N. Bednyagina, I. Postovskii, A. Garnovskii, O. Osipov: "Hetarylformazans," *Russ. Chem. Rev.* (Engl. Transl.) 44 (1975) 493 – 509.
167. R. Kuhn, H. Weitz, *Chem. Ber.* **86** (1953) 1199.
168. D. Schulte-Frohlinde, R. Kuhn, W. Münzig, W. Otting, *Justus Liebigs Ann. Chem.* **622** (1959) 43.
169. A. Zeif, G. Lipunova, N. Bednyagina, L. Shchegoleva, L. Chernyavskii, *Zh. Org. Khim.* **6** (1970) 2590.
170. H. Ziegler, *Ind. Chim. Belge* **22** (1957) 533.
171. Urals S. M. Kirov Polytechnic and Uralsk Wood Techn. Inst., SU 296 791, 1971 (Y. Rybakova, G. Lipunova, N. Bednyagina).
172. G. Lipunova, N. Gulemina, A. Zeif, N. Bednyagina, *Khim. Geterotsikl. Soedin.* 1974, 493.
173. Y. Rybakova, N. Bednyagina, *Khim. Geterotsikl. Soedin.* 1971, 970.
174. M. Ermakova, E. Krylov, I. Postovskii, *Zh. Obshch. Khim.* **30** (1960) 849.
175. H. Irving, J. Gill, W. Cross, *J. Chem. Soc.* 1960, 2087.
176. B. Hirsch, *Justus Liebigs Ann. Chem.* **637** (1960) 169. B. Hirsch, E. Jauer, *Justus Liebigs Ann. Chem.* **682** (1965) 99.

177. R. Price, *J. Chem. Soc. A* 1971, 3379.
178. F. Beffa, P. Lienhard, E. Staner, G. Schetty, *Helv. Chim. Acta* **46** (1963) 1369.
179. Instytut Przemyslu Organicznego, Warschau, PL 55 361, 55 362, 55 363, and 56 636, 1966 (E. Kubicky); PL 66 469, 1968 (K. Alchimowicz, E. Kubicky);Pabianickie Zaklady Farmaceutyczne "POLFA", Polytechnika Lodzka PL 66 469 (Z. Jankowski, R. Stolarski, W. Cieslak, R. Salagacki).Geigy, DE-AS 1 252 829 and DE-OS 1 444 688, 1962 (F. Beffa, P. Lienhard, E. Steiner).
180. R. Wizinger, *Angew. Chem.* **61** (1949) 33.
181. M. Kishita, Y. Muto, M. Kubo, *Aust. J. Chem.* **11** (1958) 309.
182. R. Price, *J. Chem. Soc. A* 1971, 3385.
183. Hoechst, DE-OS 2 945 493 and 2 945 537, 1979 (G. Schwaiger, E. Hoyer).
184. Bayer, EP-A 0 321 784, 1987 (E. Köcher, H. Jäger).
185. Geigy, DE-OS 1 444 688, 1962 (F. Beffa, P. Lienhard, E. Steiner).
186. J. Poskocil, Z. J. Allan, GB 927 128, 1963.
187. Geigy, DE-OS 1 419 841, 1960 (F. Beffa, G. Schetty).
188. H. Wahl, F. Gault, *C. R. Serie A* **251** (1960) 1523.
189. H. Wahl, M.-T. le Bris, F. Gault, F. Pierrot, *Teintex* **21** (1956) 697, 786.
190. H. Wahl, M.-T. le Bris, *Bull. Soc. Chim. Fr. ser. 5* **1954,** 248, 587. Francolor, FR 1 115 086, 1954 (H. Wahl, M.-T. le Bris).
191. O. Stashkevich, G. Pilyugin, V. Stashkevich, *Khim. Geterotsikl. Soedin.* 1970, 1104.
192. M. Lozinskij, S. Sanova, P. S. Pel'kis, *Z. Org. Chim.* **1** (1965) 314.
193. Sandoz, FR 1 338 068, 1962 (W. Steinemann).
194. Sandoz, FR 1 284 174, 1962 (W. Steinemann).
195. E. Bamberger, F. Kuhlemann, *Ber. Dtsch. Chem. Ges.* **26** (1893) 2978. Y. Gok, *Dyes Pigm.* **11** (1989) 101.
196. Sandoz, GB 984 451, 1960 (W. Steinemann).
197. Sandoz, GB 984 452, 1960 (W. Steinemann).
198. Soc. Carbochimique, DE-AS 1 062 851, 1952, 1953 (H. Ziegler).
199. Du Pont, US 3 663 525 (V. Witterholt, J. Dalgarn).
200. Geigy, DE-OS 1 419 838, 1959 (B. Beffa, G. Schetty).
201. N. Bednyagina, Y. Sedov, G. Petrova, I. Postovskii, *Khim. Geterotsikl. Soedin.* 1972, 390.
202. J. Rybakova, N. Bednyagina, *Khim. Geterotsikl. Soedin.* 1966, 287.
203. Francolor, BE 609 146, 1960 (R. Surreau, G. Kremer, V. Dupre).
204. Bayer, DE-OS 3 329 193, 1983 (R. Weitz, R. Neeff, H. Hugl). Ciba-Geigy, EP-A 0 360 735, 1988 (U. Lehmann, K. Hoegerle).
205. Hoechst, DE-OS 3 323 638, 1983 (G. Schwaiger); DE-OS 3 440 265, 1984 (G. Schwaiger).
206. Geigy, DE-OS 1 644 254, 1966 (P. Dussy, H. Meindl, H. Ackermann).
207. Urals Kirov Polytechnic, SU 825 527, 1979.
208. Urals Kirov Polytechnic, SU 1 056 052, 1982.
209. Work Safety Res. Inst., SU 1 105 814, 1983.
210. T. V. Chernyshova, SU 1 016 287, 1982.
211. Ural Polytechnic Inst., SU 689 953, 1977.
212. Work Safety Res. Inst., SU 874 628, 1979.
213. Urals Kirov Polytechnic, SU 825 526, 1979.
214. W. D. Hooper, *Rev. Pure Appl. Chem.* **19** (1969) 231.
215. F. Wohlrab, E. Seidler, K. H. Kunze: *Histo- und Zytochemie dehydrierender Enzyme,* Verlag J. A. Barth, Leipzig 1979.
216. DE-OS 3 418 852, 1984 (J. Stellmach, E. Severin).
217. V. Zotov, *Vodosnabzh. Sanit. Tekh.* 1972, 16.
218. J. Gajdusek, J. Kupec, J. Svancer, *Acta Hydrochim. Hydrobiol.* **2** (1974) 521.
219. Eastman Kodak, US 3 642 478, 1972 (A. T. Brault, V. L. Bissonette).
220. M. Lelental, H. Gysling, *J. Photogr. Sci.* **26** (1978) 135.
221. *Res. Discl.* **217** (1982) 139.
222. *Res. Discl.* **183** (1979) 393.
223. Aligena, EP-A 0 025 973, 1979 (C. Linder, M. Perry).
224. ICI, EP-A 0 219 232, 1985 (A. Baxter, S. Bostock, D. Greenwood).
225. Ricoh, JP 59 048 189, 1982; JP 59 133 260, 1983.
226. Bayer , EP-A 0 338 310, 1988(H. Henk, K. J. Herd, F. Stöhr).
227. Bayer, EP-A 0 240 839, 1986(H. Henk).
228. Bayer, EP-A 0 201 026, 1985H. Henk).
229. Y. Gok, B. Senturk, *Org. Prep. Proced. Int.* **27**(1995)87.
230. F. M. Abdel Galil, F. A. Khalifa, T. S. Abdin, *Dyes Pigm.* **12**(1990)49.
231. A. R. Katritzky, S. A. Belyakov, D. Cheng, H. D. Durst, *Synthesis*(1995)577.
232. J. Sokolowska-Gajda, H. S. Freeman, A. Reife, *Dyes Pigm.* **30**(1996)1.
233. Y. Wang, Y. Zhou, M. Wang, Z. Li, *Fenxi Ceshi Xuebao* **12**(1993)60.
234. A. Uchiumi, A. Takatsu, H. Tanaka, *Anal. Sci.* **7**(1991)459.
235. NOK Sogo Gijutsu Kenkyusho, JP 01 168 669, 1987(O. Kobayashi, K. Seki).
236. Eastman Kodak, EP-A 0 325 389, 1988(M. H. Delton, S. L. Eiff).

Metaldehyde → Acetaldehyde
Metal Effect Pigments → Pigments, Inorganic
Metal Fibers → Fibers, 5. Synthetic Inorganic

Metallic Glasses

Gary J. Shiflet, Department of Materials Science, University of Virginia, Charlottesville, VA 22903, United States

Yang Leng, Department of Materials Science, University of Virginia, Charlottesville, VA 22903, United States

Jeffery W. Hawk, Department of Metallurgy and Materials Science, University of Alabama, Tuscaloosa, AL 35487, United States

1. Introduction 1
2. Structural Characteristics of Metallic Glasses 2
2.1. Glass Formation and Crystallization 3
2.2. Structure 5
2.3. Relaxation 7
3. Processing 8
3.1. Sputtering and Evaporation 8
3.2. Chemical Deposition and Electrodeposition 8
3.3. Ion Implantation 8
3.4. Splat Quenching 8
3.5. Spray Deposition or Plasma Spraying 9
3.6. Melt-Spinning 9
3.7. Electron-Beam and Laser Surface Melting 10
3.8. Atomization and Comminution . . . 10
4. Physical Properties of Metallic Glass Alloys 10
4.1. Mechanical Properties 10
4.1.1. Elastic Properties 11
4.1.2. Anelastic Behavior 11
4.1.3. Magnetoelastic Interactions 11
4.1.4. Plastic Properties 12
4.1.5. Thermomechanical Interaction 13
4.1.6. Fracture 14
4.2. Magnetic Properties 15
4.2.1. The Curie Temperature 16
4.2.2. Saturation Magnetization 16
4.2.3. Anisotropy 18
4.2.4. Low-Field Properties 18
4.3. Electrical Properties 19
4.4. Chemical Properties 20
5. Uses of Metallic Glasses 21
5.1. Mechanical Uses 21
5.2. Magnetic Uses 22
6. References 23

1. Introduction [1], [2]

Previously, glass formation was restricted to melts, usually of oxides, in which crystallization is very sluggish. However, with the advent of new solidification techniques with which cooling rates $> 10^6$ K/s can be achieved, a new class of material–the metallic glass–has emerged. Metallic glasses have been the focus of major research and development over the last fifteen years, including a continuing search for commercial applications.

As a starting point, it is convenient to separate solid materials into two rather large classes based on structural order: crystalline solids and amorphous solids. Crystalline materials are characterized by a periodic repetition in space of atoms grouped in a clearly defined structural element–the unit cell. The long-range repetition of the unit cell in three dimensions produces the crystallinity associated with metals. Crystalline materials thus give rise to X-ray diffraction patterns with sharp intensity maxima which are related to the periodic nature of the specific crystal lattice of the solid. In amorphous materials, there is no long-range periodicity, and the atoms are arranged more or less randomly. In amorphous solids there may be, however, regions of local order due to the topological and chemical constraints imposed by the close packing of atoms as the melt solidifies. This order is not repeated over large distances in the solid, and in general these regions are very localized and widely separated. Thus, X-ray diffraction patterns from amorphous materials have no sharp intensity maxima.

It is also apparent that the characteristics of cooling from the liquid state to the solid state differentiates crystalline from amorphous materials. In crystalline materials, as the melt is

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_335

cooled, the transition from the liquid state to the solid state is sharp and, under equilibrium conditions, takes place at a well-defined temperature with an appreciable release of heat. In amorphous materials, the transition from the liquid to the solid is very broad and can be quantified by the progressive increase in the melt viscosity towards a value so high that the material can be considered a solid. During this transition the atoms remain in approximately the same configuration as found in the liquid, and the glass can be considered as a supercooled liquid.

Amorphous solids can additionally be subdivided into two groups: glasses (inorganically based) and polymers (organically based). Inorganic glasses, such as fused silica or mixtures of silica with oxides of Al, Na, Mg, Ca, etc., comprise the oldest group of amorphous solids. Other glasses based on elements of groups 15, 16, and 17 also form glassy solids (e.g., chalcogenide glasses). When cooled from the melt, the silicates and chalcogenides do not crystallize. However, liquid metals and alloys do crystallize upon freezing. In the last twenty-five years, however, novel processing techniques have led to the development of amorphous metallic alloys. Unlike silicates and chalcogenides, it is very difficult to prevent metallic alloys from crystallizing unless the melts can be cooled at very high rates (i.e., $\geq 10^6$ K/s). At these high cooling rates, the liquid becomes supercooled and crystallization is suppressed. A metallic glass is thus formed. However, it must be remembered that these supercooled glassy metals only persist in metastable equilibrium; hence metallic glass alloys should be regarded as thermodynamically metastable. These alloys may remain in a particular metastable state indefinitely for kinetic reasons; however, they tend to evolve with time towards other amorphous configurations having lower free energies until an equilibrium state characteristic of the particular temperature is reached or until the crystalline phase is formed.

Even though metallic glasses exist in metastable equilibrium, they have generated considerable interest for two reasons. First, the fact that amorphous metallic alloys could be produced at all generated scientific curiosity. However, as experimental research proceeded, interest grew in understanding the solid state structure, the unusual mechanical properties, the liquidlike electrical properties, and the ferromagnetic properties of these amorphous metallic alloys [1]. In fact, metallic glasses have become an important commercial material (e.g., Allied-Signal Metglas) precisely because of their magnetic properties.

Table 1. Metal – metalloid and metal – metal metallic glass formers [1]

Groups ∗	Representative systems	Composition, atom %
Major Systems		
T^2 metal	Pd – Si, Co – P,	15 – 25 (X)
+ metalloid (X)	Ni – P – B	
T^1 metal +	Zr – Cu	30 – 65 Cu
T^2 metal (or Cu)	Ti – Ni	30 – 40 T^2
	Nb – Ni	40 – 65 Ni
Miscellaneous Systems		
A metal + B metal	Mg – Zn	25 – 35 Zn
Actinide + T^1 metal	U – V, U – Cr	20 – 40 T^1
T^1 metal + A metal	(Ti, Zr) – Be	20 – 60 Be

∗ A metal = Li, Mg groups; B metal = Cu, Zn, Al groups; T^1 metal = early transition metal (Sc, Ti, V groups); T^2 = late transition metal (Mn, Fe, Co, Ni groups) or noble metal; metalloid = B, C, Si, Ge, P

2. Structural Characteristics of Metallic Glasses

It is generally acknowledged that pure metals cannot be quenched from the melt into a metallic glass with the presently achievable cooling rates of $10^6 - 10^8$ K/s. This means that all metallic glasses are in fact alloys containing two or more elements. No theory at present is capable of predicting the composition of a metallic alloy that is likely to be amorphous upon solidification from the melt. However, experimental results suggest some general guidelines for the development of amorphous metallic alloys [2].

Many glassy alloys contain both a strongly metallic element, such as Cu, Ag, Au, Fe, Co, Ni, Pd, or Pt, and a nonmetallic element (i.e., metalloid) of valence 3 (B), 4 (C, Si, or Ge), or 5 (usually P) [2]. There are exceptions to this general guideline, most notably the following binary alloys: Cu – Zr, Nb – Ni, Ta – Ni, Y – Cu, and Ti –Ni [1]. Table 1 contains a classification scheme for metal – metalloid and metal – metal metallic glass alloy systems as well as some miscellaneous systems.

Ternary metallic glass alloys contain either two metals and one metalloid (e.g.,

Pd – Ni – P) or one metal and two metalloids (e.g., Fe – P – B). In both cases the sum of the metallic elements amounts to 70 – 85 atom %, leaving 15 –30 atom % metalloid [2].

Metallic glasses containing up to 90 atom % aluminum were produced independently in Japan and the United States. These glassy aluminum alloys, unlike previous metallic glass alloys, have a low density, resulting in high specific strengths [3], [4].

2.1. Glass Formation and Crystallization

If an appropriate liquid is quenched rapidly enough to bypass normal solidification then a metallic glass is formed. The resulting metallic glass is inherently metastable and, at temperatures well below the melting point of the alloy, the glass can transform into one or more crystalline phases. It is therefore instructive to understand how metallic glasses form and then transform to their crystalline counterparts.

Upon melting, a metal or alloy loses its three-dimensional lattice arrangement of atoms. All long-range order is destroyed and the atoms subsequently vibrate about positions that constantly interdiffuse in the liquid state. During melting, the crystal and the liquid are in equilibrium. For pure metals, the enthalpy, entropy, and volume undergo a discontinuous change (i.e., enthalpy, entropy, and volume increase). As a liquid well above the melting temperature, the melt is in a state of internal equilibrium, and its structure and properties are independent of its thermal history.

During solidification, a liquid must be undercooled below its equilibrium crystallization temperature in order to surmount an energy barrier to nuclei formation. The degree of undercooling needed to form nuclei depends on the initial viscosity of the liquid, the rate at which the viscosity increases with decreasing temperature, the temperature dependence of the free energy difference between melt and crystal, the volume density and efficacy of heterogeneous nucleating particles and the imposed cooling rate [5]. The growth rates of crystals in metallic melts are quite rapid once nuclei have formed, and rapid recalescence occurs in those regions where the rate of heat removal to the surroundings is small.

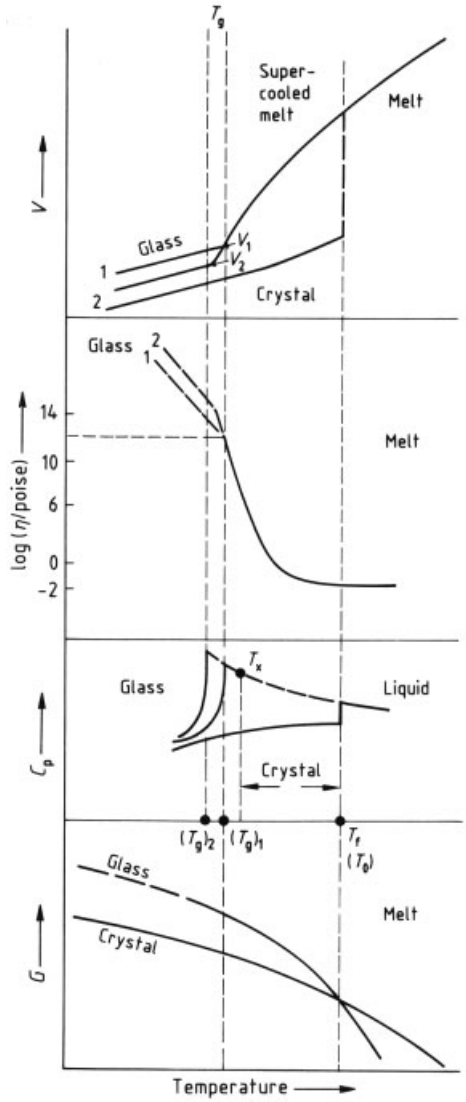

Figure 1. Variation of volume V, viscosity η, specific heat C_P, and free energy G, with temperature for crystalline and amorphous phases [1]

When a liquid is cooled rapidly by heat removal to a sink, the influence of heterogeneous nucleants is delayed for kinetic reasons. As the cooling rate is further increased, undercooling is enhanced and recalescence decreases. Thus, the temperature range over which crystallization proceeds becomes increasingly narrow. This leads to a refinement of the microstructure, an extension of solute solubility, and to the formation of metastable crystalline phases. When

the cooling rate is sufficiently fast, crystallization is completely suppressed because of insufficient time for the development and growth of nuclei. In this case the shear viscosity of the melt η increases continuously, as shown in Figure 1.

Although the driving force for nucleation continually increases, the rapid decrease in atomic mobility of the melt opposes nuclei formation and dominates at high undercooling. Eventually, the atomic configurations of the liquid departs from equilibrium and becomes homogeneously "frozen" at the glass transition temperature T_g. The structural freezing to the glass state is, by convention, considered to occur when the viscosity of the melt reaches 10^{13} poise. No unique glass structure exists in reality because the temperature and structure at which departure from configurational equilibrium occurs depends on the cooling rate. Thus, T_g and the glass structure are dependent on the cooling rate.

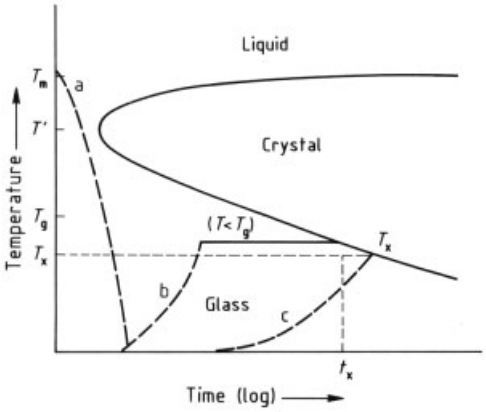

Figure 2. TTT diagram of glass formation and crystal growth in an undercooled melt [6]
(a) Fast cooling to form a glass; (b) Isothermal heating of glass (i.e., $T < T_g$) giving crystallization at time t_x; (c) Slow heating of glass producing crystallization at T_x

The glass-forming ability of a material is determined by the kinetics of the nucleation process and of the early stages of crystal growth [6]. This is illustrated by a TTT (time – temperature – transformation) diagram (Fig. 2). If the liquid is quenched from above T_m (i.e., the absolute melting temperature) to well below T' (i.e., the temperature at which the crystallization rate is a maximum) in a time t_{min} (i.e., the minimum time for crystallization), the undercooled liquid is retained, and at T_g, the configuration is frozen to form a glass. At room temperature, the time for crystallization of a glass is very long. If the glass is heated to a temperature $T < T'$ and held isothermally for a time t_x, it will begin to crystallize. Alternatively, if the glass is heated slowly from room temperature, crystallization begins at T_x and increases with increasing heating rate. In actuality, there are a number of crystallization reactions. Most metallic glasses initially crystallize into one or more metastable phases before crystallizing into the equilibrium phase.

Crystallization of metallic glasses does not occur at a unique transformation temperature. There are a range of temperatures at which crystallization can occur, although it is standard practice to define the crystallization temperature of a metallic glass as the temperature at which spontaneous crystallization begins as the glass is heated at a constant rate from room temperature. Usually the crystallization temperature T_x is defined as the onset temperature or the peak temperature of the first exotherm of a differential scanning calorimeter (DSC) scan. Upon crystallization, the density of a metallic glass increases by ca. 1 %. The enthalpy of crystallization is ca. 5 kJ/mol and represents about half the latent heat of melting [7]. The crystallization temperature for most metallic glasses lies between $0.4\,T_m$ and $0.65\,T_m$. Absolute values of T_x vary from ca. 400 K for magnesium-based glasses to 1200 K for glasses based on refractory metals. Ferromagnetic iron-based glasses possess T_x values near 700 K [7].

A number of factors increase the crystallization temperature in metal – metalloid glasses. These factors include low overall electron/atom ratios, the large atomic size difference between constituent transition metals, and the different structures of the stable metal – metalloid crystalline phases that occur in an alloy system [7].

The crystallization mechanism which occurs in a metallic glass can depend upon its composition. The following reactions have been observed:

1) polymorphous crystallization,
2) eutectic crystallization,
3) primary crystallization, and
4) separation into two distinct glassy phases.

Polymorphous crystallization occurs when the glass transforms into the crystalline phase

with no change in composition. As an example, the crystal phase in an iron – boron glass is usually Fe_3B although the equilibrium phase is Fe_2B [8].

Eutectic Crystallization. In this transformation reaction the glass forms two phases which grow in a coupled fashion. There is no overall change in composition from glass to eutectic colony. In an $Fe_{40}Ni_{40}P_{14}B_6$ glass, the crystal contains fine lamallae of iron – nickel austenite and the metastable $(Fe,Ni)_3PB$ phase. Growth of the dominant eutectic colonies is initiated by homogeneous nucleation of the boride phase, after which the iron-rich phase nucleates without difficulty [9].

Primary Crystallization. In this reaction, the first crystals to form have a composition different from the glass (e.g., nearly pure Al crystallizes from an Al – Fe – Gd glass; see Fig. 3) [1]. The rate of growth of the crystals is determined by the kinetics of the diffusion of rejected solute (i.e., Fe and Gd) away from the aluminum crystallite into the remaining glass. The solute-enriched glass then undergoes a polymorphous transition to another crystalline species.

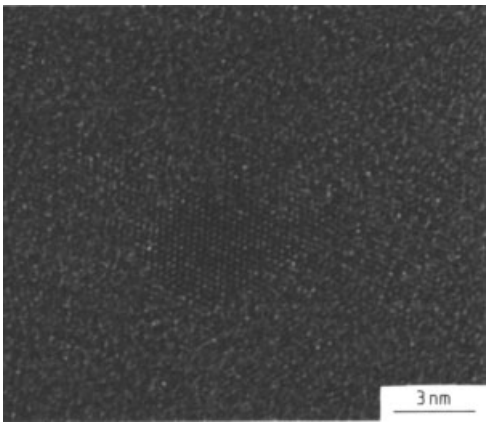

Figure 3. High resolution electron micrograph of $Al_{90}Fe_5Gd_5$ metallic glass after annealing at 175 °C for 150 min
The periodic region is an aluminum crystallite along the [110] direction; the remaining area is glassy [3]

Separation into Glassy Phases. For example, a $Zr_{36}Ti_{24}B_{40}$ glass separates into very distinct but glassy phases below T_x, with each glassy phase crystallizing separately [10].

Crystallization kinetics usually follow the Avrami equation for solid state nucleation and growth processes [11]

$$x(t) = 1 - \exp\left[-b\,(t-t_0)^n\right] \tag{1}$$

where $x(t)$ is the fraction crystallized in time t, t_0 is the incubation time, b is a temperature dependent rate constant [equal to $b_0 \exp(E_x/kT)$], and n is an exponent which varies from 1.5 to 4.0 [6]. The activation energy E_x in the crystallization process, which incorporates the activation energy for both nucleation and growth of the crystallite, varies from 40 to 400 kJ/mol [7].

Once crystals are nucleated in the metallic glass, growth proceeds in the conventional manner. For example, during polymorphous crystallization, in which diffusion is not necessary, the growth rate is independent of time and increases exponentially with increasing temperature. In primary crystallization, the process is diffusion controlled. The crystal dimensions are thus proportional to the square root of time. Eutectic crystallization is also diffusion controlled. Depending on the glassy alloy, the diffusion necessary to partition the solute between the crystalline phases may occur at the crystal – glass interface or in the glass ahead of the growing crystal. The spacing between phases is independent of growth rate, unlike conventional eutectic and eutectoid transformations [7].

2.2. Structure

In metallic crystals, determining the structure consists of identifying the coordinates of all atoms in the unit cell. For a glass, however, the structure can only be described on a statistical basis. Glasses do not possess a unit cell, and the local environments of different, chemically identical atoms may vary.

The determination of glassy structures resolves itself into two complementary tasks: determining a radial (or pair) distribution function (RDF) and modelling [1], [12]. In principle, the RDF does not provide a complete, unambiguous description of the glass structure. It constitutes an important test of the validity of a proposed structural model, which in turn must be augmented by tests of other properties (e.g., determination of macroscopic density, and the re-

sults of Mössbauer spectroscopy, nuclear magnetic resonance, and positron annihilation studies) [13].

Several models for metallic glasses deserve mention [14]. The first model, the Dense Random Packing (DRP) model, was developed by BERNAL, initally for liquid structures [15]. GASKELL [16] has proposed an alternative model based on prism packing which mimics the smallest identifiable units of a stable crystalline phase. A third model starts with the DRP model but incorporates the effects of chemical ordering [1], [12].

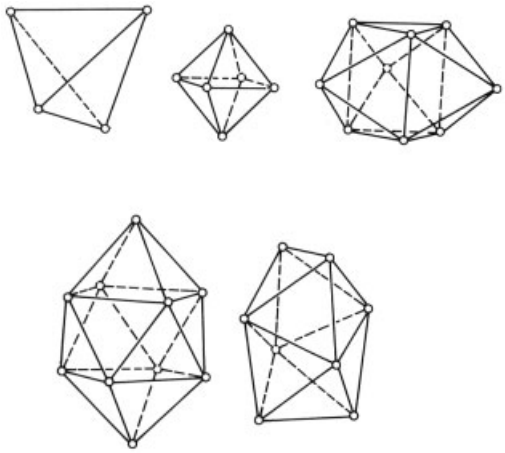

Figure 4. Polyhedra formed from dense random packing of hard spheres [15]

Bernal's empirical model [15] originally referred to the DRP of hard spheres of equal diameter. The packing was based on physical experiments with steel balls kneaded in a rubber bag. The packing of the spheres was statistically reproducible and consisted of a number of different polyhedra (Fig. 4). The small polyhedra (i.e., the tetrahedra and octahedra) make up the majority of structural units, although the frequency ratio of polyhedra is subject to dispute. For some metallic glasses, there is respectable agreement between the computed RDF and that deduced from diffraction experiments. However, in the case of an Ni – P glass, the computed RDF ignores the P atoms. A second problem with Bernal's approach is the use of "hard" spheres in the model. This model has been modified to account for these difficulties, and the replacement of "hard" spheres with "soft" spheres has led to better agreement between the modelled and "observed" RDF [17]. Even with these refinements, Bernal's approach suffers from numerous approximations and other difficulties.

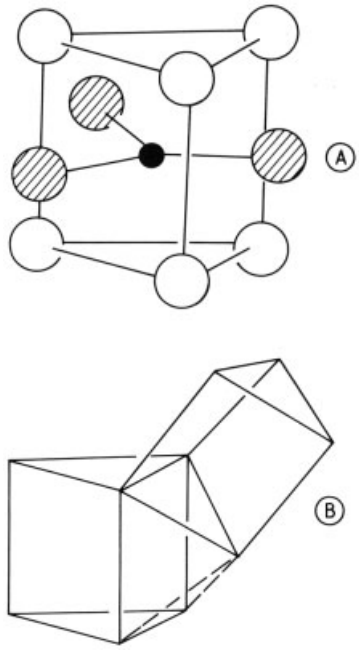

Figure 5. (A) Trigonal prismatic packing of transition metal atoms (hatched) and metalloid atom (solid), as found in Fe_3C or Fe_3P; (B) In a glass the prisms can be connected in various configurations after relaxation has occurred, of which this is the most important example [16]

In attempting to reconcile a deficiency in Bernal's model of the glass structure, GASKELL [16] proposed a prism packing model (Fig. 5) analogous to the network model used for oxide glasses. In metal – metalloid glasses with about 20 % metalloid atoms, GASKELL calculated that the metalloid diameter must be ≤ 0.48 of the metal atom diameter. It had been assumed that metal – metalloid glasses are stabilized by the insertion of small atoms into the larger structural voids in the assembly of large atoms. However, only carbon and boron can physically occupy these voids without lattice distortion, even though phosphorus, silicon, and arsenic are effective glass stabilizers. In the Pd_3Si glass, the linkage of prisms among themselves is refined by computer approximations in order to minimize atomic overlap. The resulting densities and RDF's agree well, and GASKELL proposed that the modelling procedure can apply to many metal – metalloid glasses. Further confir-

mation of the network model has been obtained for $(Ru_{84}Zr_{16})_{1-x}B_x$ glasses ($0.40 < x < 0.53$) [18]. With this high metalloid content, no DRP model gives even moderate agreement with the diffraction data. However, a form of irregular packing based on the Ru_2B crystal structure provides good, though not perfect, agreement to observed RDF's. This structural model accommodates more easily the high metalloid concentrations than either the hard or soft RDP models.

The third approach models metal – metal glass alloys [1], [12]. This concept is based on chemical short-range order (CSRO) where the nearest neighbor identities are not randomly determined. The philosophy behind this approach contends that a preexisting single component model incorporates the fundamental packing constraints in a dense, noncrystalline structure. With the topology precisely determined, complete data are available which allow the labelling of A and B atoms in a manner consistent with the desired local order. Different combinations of A and B labelling can be analyzed by allowing statistical variations in the topological and metrical perfection of the local ordering scheme. Thus relaxation calculations can be performed with periodic boundary conditions. A $Mg_{70}Zn_{30}$ metallic glass [19] has been studied by means of CSRO with potentials developed from pseudopotential theory [20].

2.3. Relaxation

A metallic glass is a metastable structure in which large-scale rearrangements are difficult due to low atomic mobility. The exact structural state developed in the glass is directly related to its cooling rate. If a glass is subsequently annealed below T_g, the structure will evolve toward a more stable state (e.g., toward the structure of the thermodynamically "ideal glass"). The relaxation process involves the reduction in free volume and changes in the chemical identity of nearest neighbor atom pairs. Associated with relaxation of the glassy structure is a change in its physical properties.

When a metallic glass relaxes, its density increases by ca. 0.25 % and its enthalpy decreases by ca. 1 kJ/mol [21]. The physical properties of the glass also change, with elastic modulus, Curie temperature, and viscosity increasing, while the thermal resistivity, diffusivity, and fracture toughness decrease [1], [21].

Although metallic glasses have no long-range atomic periodicity, they possess a degree of short-range order (i.e., local atomic coordination). Relaxation thus causes changes in the short-range order (SRO) and the defect distribution within the glass. When rapidly cooled from the melt, the solid glass is highly disordered. Upon annealing below T_g, various structural defects are removed and the SRO increases. This leaves a new structural configuration in metastable equilibrium at the specific annealing temperature. This process can be repeated at different annealing temperatures, and at each temperature the SRO changes to reflect the new metastable equilibrium condition. Two accepted types of microstructural changes are associated with changes in SRO [1], [21]:

1) topological short-range order, and
2) chemical short-range order.

Topological Short-Range Order (TSRO). Changes in TSRO constitute a more efficient packing of the constituent atoms irrespective of their chemical identity. This process effectively eliminates and/or redistributes the free volume within the structure. As a result, the density of the glass increases. Since TSRO involves the cooperative rearrangement of atoms over relatively large distances, this process is expected to occur slowly.

Chemical Short-Range Order (CSRO). Changes in CSRO occur upon rearrangement of individual species on the atom sites. These changes occur rapidly since long-range movement of atoms is not necessary.

Although the two ordering effects have been treated separately, in practice they are interrelated because changes in SRO are likely to involve both topological rearrangement and chemical ordering.

Relaxation processes affect the physical properties of metallic glasses to a great degree. For example, the elastic modulus of a metallic glass may change by 10 % upon relaxation. Since the quenching rate can change the structure and properties of the glass, the thermal history of a metallic glass is very important. Additionally, it is generally not possible to standardize the structural state of metallic glasses,

following different quenching rates, by a relaxation annealing below T_g. However, the measured properties will be more consistent if the metallic glass is relaxed as fully as possible prior to property measurements.

3. Processing

The critical processing step in producing metallic glasses is the quench from the liquid state. The quench rate must be very fast ($> 10^6$ K/s) to freeze in the desired metastable structure; therefore, rapid heat removal from the bulk is needed. This requires that any cooling process has a sufficiently high heat transfer coefficient at the interface of the liquid metal and the cooling medium and that the material possess a thin cross section in order that heat be rapidly conducted out of the metal. Another important factor in the preparation of rapidly quenched materials is undercooling the melt so that subsequent recalescence does not cause morphological destabilization of the material.

In removing heat from the bulk liquid metal alloy, it is important to maximize the contact area between the liquid and the cooling medium by rapidly increasing the surface area of the liquid alloy. This basic principle has led to a number of procedures for producing metastable alloys [1], [22].

3.1. Sputtering and Evaporation [23]

Sputtering and evaporation methods build metallic glass alloys atom-by-atom. Sputtering and evaporation techniques depend on the deposition of a vapor on cold substrates (often cooled to liquid nitrogen or liquid helium temperatures). In essence, these methods require that the kinetic energy of the condensing atom exceeds the binding energy of the atoms in the substrate surface. Sample preparation by these means is very slow (i.e., 20 µm/h) and is very sensitive to deposition conditions. For example, the presence of impurity atoms such as oxygen facilitates the formation of an amorphous atomic structure [24]. However, because this method maximizes the surface-to-volume ratio, very high quench rates may be achieved (10^8 K/s), allowing the formation of metallic glasses that cannot be produced by other melt quenching techniques.

3.2. Chemical Deposition and Electrodeposition

Various metal – metalloid amorphous alloy systems have been prepared by chemical deposition and electrodeposition. The final alloy composition depends on the deposition conditions and bath composition during sample formation. Deposition rates of ca. 0.1 mm/h are typical [25]. Most alloys produced by these methods to date have consisted of Fe, Ni, and/or Co with P, B, or S.

3.3. Ion Implantation

Ion implantation, which has an effective quench rate of 10^{14} K/s, allows a wide variety of atomic species to be introduced into the surface of many materials. This technique has produced amorphous materials that are virtually identical to alloys of the same nominal composition made by more conventional rapid solidification techniques. For example, nickel films can be amorphized by phosphorus ion implantation and are identical to films produced by evaporation or electrodeposition [26]. However, self-ion bombardment (e.g., Ni^+ into Ni) never generates a stable glassy structure. Ion implantation is attractive because it allows a degree of freedom from the compositional restrictions imposed by constitutional phase diagrams, although similar composition rules for metallic glass formation seem to apply for both implanted and conventionally quenched alloys. In essence, ion implantation is a surface modification technique because the distance of ion penetration is linked to the energy loss sustained by the penetrating ion. Single-stage ion implantation modification remains viable only for thin samples.

3.4. Splat Quenching

The gun technique developed by Duwez [27] was probably the earliest rapid solidification method used to produce metastable crystalline and amorphous alloys. In the gun technique,

a molten alloy is atomized by expulsion from a shock tube. The molten alloy droplets are then propelled against a substrate having a high thermal conductivity. The gun technique probably achieves the highest cooling rates of all splat quenching techniques for the very thin regions of the foil. However, the resultant samples are irregular in thickness (i.e., 0.1 – 50 µm) and are very small. As a result, other techniques have evolved from the gun method. For example, the piston and anvil [28] and twin-piston (Fig. 6) [29] methods rapidly flatten and quench molten alloy droplets between impacting substrate surfaces, leading to a more uniform splat. The arc-hammer method uses a water-cooled hearth to quench an arc-melted button of material struck by a hammer [30]. Modifications of these techniques include the use of better substrate materials (e.g., diamond [31]) and quenching techniques (e.g., magnetically driven substrate platens [32]). All splat quenching methods produce to a certain extent foils of irregular shape and variable thickness. These techniques are not suitable for producing large quantities of material.

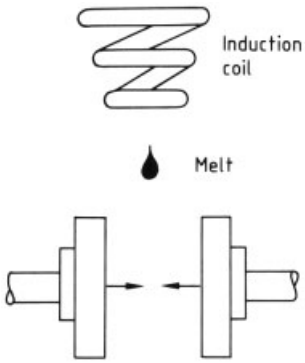

Figure 6. Twin-piston splat quenching with levitation induction melting

3.5. Spray Deposition or Plasma Spraying

Variations of melt droplet spraying and deposition have been used to fabricate rapidly solidified materials. One variation involves the injection of a molten alloy into a hot plasma flame which forms splats upon impact with a high-conductivity substrate [33]. Another modification involves the use of a rotating wheel substrate which substantially increases sample yield [34]. Spray techniques combine melting, quenching, and compaction into a single processing step [35].

3.6. Melt-Spinning

Melt-spinning (Fig. 7), or the production of filaments, ribbons, or tapes of rapidly solidified material, has the advantage of producing products in continuous form. Making continuous rapidly solidified materials from the melt improves both the sample geometry and the quench rate geometry.

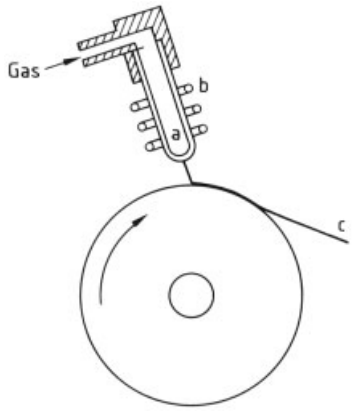

Figure 7. Melt-spinning with cooling on a metal drum
a) Melt; b) Induction coil; c) Ribbon

Filaments of circular cross section can be made by expelling the molten alloy through a circular orifice [36]. Melt-jet stabilization, solidification, and cooling are controlled by the quenching media [37], either gases [38] or liquids [39]. A solid substrate can also be used to quench the molten alloy, but this results in a quasi-circular cross section. Such processes include melt drag [40], crucible [41], and pendant-drop melt extraction [42].

Metallic ribbons have been made for over 80 years by allowing a stream of molten alloy to flow onto a rapidly moving substrate [43]. Chill block melt-spinning (CBMS) [44] utilizes the impingement of a jet of molten alloy from a circular orifice onto a rapidly moving substrate to produce a continuous ribbon. The puddle that results from the impingement of the melt jet serves

as a local reservoir from which the ribbon is continuously formed and chilled. Both the initial quench rate ($10^5 - 10^8$ K/s) and the final ribbon geometry depend upon the nature of the melt puddle. Ribbons 20 to 200 µm in thickness usually result from this technique. A smooth, unchanging puddle produces ribbons with smooth edges and surfaces [45]. Typical substrate surfaces are the inside of drums [46] or wheels [47], the outside of wheels [48], between twin-rollers [49], and on belts [50].

The difference between ribbons and tapes lies in the width of the resultant rapidly solidified material. The use of a slit several centimeters or more in width replaces the circular orifice in the CBMS method [51]. The use of a planar melt jet is difficult due to the destabilization of the stream that results from the high surface tension and low viscosity of the molten alloy [52]. As a result, planar flow, or strip, casting of molten metallic alloys has been developed in which the melt nozzle is held very close to the moving substrate [53]. This reduces perturbation of the melt on the substrate and thus results in the formation of rapidly quenched tapes of various widths.

3.7. Electron-Beam and Laser Surface Melting [54]

Rapid quench rates (i.e., $4\times10^4 - 5\times10^6$ K/s) have been generated by melting a thin layer on the surface of a suitable substrate material. Electron and laser beams are very efficient in melting small surface areas rapidly and continuously because of the high energy densities (100 kW/cm^2) they generate. For example, 400 µm thick amorphous surface layers have been created on a $Pd_{91.7}Cu_{4.2}Si_{5.1}$ substrate [55]. Aside from forming amorphous surface layers, these techniques can be used to form new surface compositions by melting thin films [56] or by continuously feeding alloy wire or powder onto the substrate surface [57].

3.8. Atomization and Comminution [58]

Atomization and comminution techniques are efficient in producing rapidly solidified powders having a refined microstructure and grain size as well as homogeneous, stable phases. Atomization is a process whereby powder particles are formed from the quenching and solidification of discrete droplets of liquid metal. Comminution is the chopping or comminuting of rapidly solidified ribbons or tapes of amorphous alloys.

Atomization techniques include ultrasonic gas atomization (10^5 K/s) [59], water atomization ($10^5 - 10^6$ K/s) [60], gas – liquid atomization (10^6 K/s) [61], centrifugal atomization ($10^5 - 10^6$ K/s) [62], roller atomization (10^5 K/s) [63], and spark erosion ($10^6 - 10^9$ K/s, depending on powder size; i.e., the smaller the powder particle the faster the cooling rate) [64].

The final product of all atomization and comminution processes is a powder which must ultimately be consolidated into a useful shape. For example, dynamic or explosive compaction has been used to consolidate powder while retaining the amorphous state [65]. Other consolidation techniques include warm die pressing and warm extrusion. However, whenever an amorphous material is heated, devitrification is always a problem. While much is known about processing crystalline powder alloys, creating amorphous powder materials is still a relatively underdeveloped area of research.

4. Physical Properties of Metallic Glass Alloys

4.1. Mechanical Properties

Metallic glasses exhibit mechanical properties quite different from those of crystals. The disordered atomic structure increases the flow resistance in metallic glasses so that the ratio of Young's modulus to yield strength E/σ_y is typically about 50, which approaches the theoretical strength of these materials. The fracture strength of iron-based glasses can be as high as 3.4 GPa. The average ratio of hardness to yield strength is about 3, which is also observed for crystalline metals, but the elastic stiffness is lower than that of its crystalline counterparts. The ductility of glassy metals depends strongly on the stress state. For example, metallic glasses behave in a ductile manner when bent (with strains as high as 100 %), but they are macroscopically brittle in tension. The fracture toughness in metallic glasses is also higher than in conventional

glasses; and it is higher in comparison with metals if based on strength.

4.1.1. Elastic Properties

A detailed review of metallic elasticity is given in [66]. In general, the elastic or shear modulus of a metallic glass is less than that of its crystalline counterpart. Typical experimental values for the ratio of the crystalline to amorphous moduli lie in the range 1.2 to 1.4. The lower modulus in the metallic glass is associated with its lower density and the presence of "loose" sites in the amorphous structure [67]. The moduli increase irreversibly after thermal treatment and correspond to structural relaxation during thermal treatment [68].

Poisson's ratio ν for metallic glasses is about 0.4. This is high when compared to metals and alloys ($\nu \approx 0.33$) and nonmetallic glasses ($\nu \approx 0.25$).

Generally, glassy metals exhibit isotropic elastic behavior by virtue of their random disordered atomic structure; however, anisotropy can be induced in certain ferromagnetic metallic glasses by magnetic annealing. Such anisotropic behavior is relatively minor and elastic isotropy is a reasonable assumption for most cases.

4.1.2. Anelastic Behavior

Metallic glasses exhibit time-dependent stress – strain behavior. This behavior results from the stress-induced adjustment of the glass as it passes from one state of metastable equilibrium to another, in response to the applied stress.

Relaxation processes include thermoelastic rearrangement, atomic diffusion, and magnetoelastic interaction. The thermoelastic effect involves the adiabatic temperature change in the glass due to applied stress; such a change causes a strain increment and results in a change in modulus.

Elastic relaxation in metallic glasses can be induced by atomic diffusion as in crystalline solids; however, in amorphous solids, the concept of a discrete lattice site is replaced by generalized concepts of free volume and "loose" sites. The reversible motion of "loose" sites can be envisioned as a micromechanism of anelasticity in glassy metals [69].

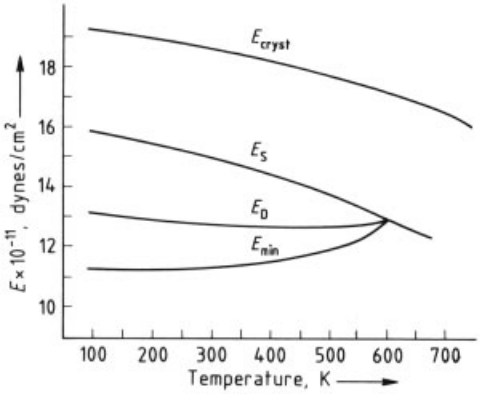

Figure 8. Young's modulus of $Fe_{75}P_{15}C_{10}$ metallic glass after various thermal annealing conditions [66]

4.1.3. Magnetoelastic Interactions

Certain ferromagnetic metallic glasses exhibit a ΔE-effect. The ΔE-effect refers to the change in the elastic modulus E with changing magnetization states. The ΔE-effect in metallic glasses can be large and is readily demonstrated at room temperature with low to moderate fields. Figure 8 shows how E varies with temperature for a Fe – P – C glass for different magnetic states. These curves represent the elastic modulus for the saturated magnetized condition (E_S), for the demagnetized condition (E_D), and for an intermediate level of magnetization where the modulus exhibits a minimum value (E_{min}). The uppermost curve depicts E after crystallization annealing at 450 °C. After crystallization, the modulus increases substantially and the ΔE-effect disappears. However, the loss of the ΔE-effect is not associated with a loss of ferromagnetism. The ΔE-effect depends in part on the alloy composition and prior treatment, as well as on the magnetization and temperature. Magnetoelastic effects do not arise from large magnetostriction, but rather reflect the absence of strong restraints to the motion of domains. It has been found that the change of elasticity at the intermediate state of magnetization is caused by a rotation process of the magnetic moment [70] and domain boundaries [71]. It is believed that domain movement

with applied stress produces the ΔE-effect responsible for anelasticity [66].

4.1.4. Plastic Properties

Basically, plastic flow in metallic glasses occurs in two ways: homogeneously and inhomogeneously. Homogeneous flow, in which the glass deforms uniformly, occurs at low stresses ($\tau < \mu/50$, where τ is the shear stress, and μ the shear modulus) and usually is observed in creep experiments. Inhomogeneous flow in which the strain is localized in a shear band, occurs at high stress ($\tau > \mu/50$) and low temperatures ($T < 0.7\,T_x$).

Homogeneous flow in metallic glasses is significant during deformation at elevated temperature. At sufficiently low stresses ($\tau < \mu/100$), the flow is Newtonian viscous ($\dot{\gamma}\,\alpha\,\tau$, where $\dot{\gamma}$ is the strain rate) and is characterized by a shear viscosity η ($\eta \equiv \tau/\dot{\gamma}$). At higher stress, the shear stress – strain rate relation is nonlinear and can be described as $\dot{\gamma}\,\alpha\,\sinh(A\,\tau)$, where A is a factor associated with the temperature and material properties.

The microscopic mechanisms for homogeneous flow have been reviewed in [72]. These mechanisms are based on models of free volume, configurational entropy, microscopic shear transformation, structure relaxation, and isoconfigurational flow.

The free volume and configurational entropy models have been used to interpret flow at high temperature ($T > T_g - 20$). The former model accounts for the temperature dependence of flow by examining changes in relative free volume in the glasses, while the latter explains the temperature dependence by considering changes in the size of the rearranging region.

For flow at low temperature ($T < T_g - 20$), structural relaxation is the most important micromechanism. The atomic structure deviates from the configuration required for equilibrium due to rapid cooling during glass formation. This "frozen" atomic configuration is allowed to approach its equilibrium state at a temperature slightly below T_g. The subsequent viscosity increase during isothermal annealing can be interpreted as an increase in structural order, which results in a decrease in the atomic defect density in the flow region, which is related to flow resistance. At high heating rates, ductility decreases less rapidly as a consequence of structural relaxation than it does for low heating rates [73]. Therefore, by controlling the composition and thermal processing, high homogeneous plastic deformation can be achieved.

Inhomogeneous Flow. At high stress levels, plastic flow in metallic glasses occurs in highly localized deformation shear bands. The strains sustained in individual shear bands are exceptionally large. There is evidence that these bands accommodate shear strains of ca. 10, with shear band width as narrow as 2 – 5 nm [74]. The number of shear bands and their orientation are determined by the kinematic constraints of the deformation process (e.g., they are different for uniaxial compression, bending, rolling, drawing, and localized indentation). Multiple shear bands are observed to form due to mechanical constraints which prevent catastrophic shear instability. Figure 9 shows an example of multiple shear band development in a Al – Fe – Gd glass upon shearing of the ribbon surface.

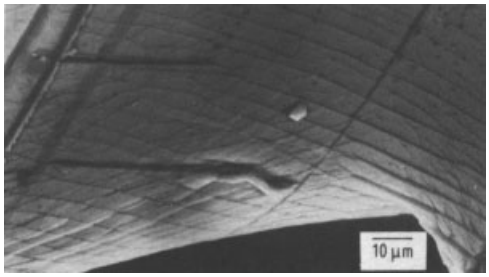

Figure 9. Primary and secondary shear bands produced by complex bending [3]

The appearance of intensely localized shear flow is sufficient to demonstrate the absence of work hardening in metallic glasses. As a consequence, the deformation of metallic glasses loaded in uniaxial tension is mechanically unstable, and tensile failure occurs almost immediately after yield. Therefore, the tensile fracture strength of a metallic glass is almost identical to its yield strength, and the tensile fracture strain is equivalent to that of brittle materials. In contrast, when a shear instability is avoided and a multiple shear configuration can be developed, the ductility of metallic glasses is remarkable (e.g., 5 % in compression and 100 % in bending).

There are two primary micromechanisms which explain the localized shear band deformation phenomena: the dislocation model and the fluid layer model.

Dislocation Model. Slip lines are observed to terminate inside an amorphous metal, and the boundary which separates the slipped from the unslipped regions can be defined as a "dislocation" in a metallic glass. Since the slip vector (both direction and magnitude) may not be constant along the dislocation line, it is called Somigliana dislocation [75].

The topology of forming localized dislocations in amorphous systems has been investigated by CHAUDHARI et al. [76], [77]. It was determined that stress fields, characteristic of both screw and edge dislocations, can be found in an amorphous solid.

The existence of dislocations is supported by the reversible nature of shear band formation. For example, a shear band can be sheared in a reverse direction [78]. This reversibility cannot easily be removed by annealing, and it persists until the material is embrittled by crystallization.

Fluid Layer Model. The viscosity of a metallic glass can decrease due to the dilatation at a stress concentrator, such as a microcrack [79]. The free volume, and hence the atomic mobility of a viscous fluid layer, can be generated in the shear band. After shear band initiation in the viscous layer, subsequent softening is sustained by strain-induced disordering. That is, the deformation process continually disorders the material in the shear band at a rate faster than structural relaxation processes can restore it. In this model, disordering is assumed to be a shear-induced, free-volume creation process. This results from atoms pushing their neighbors aside at high stress levels. This process is balanced by an ordering process during which free volume is annihilated [80]. Such creation and annihilation processes establish a steady state, in which the free volume, and hence the viscosity, is a function of the applied stress. This model accounts for the strain rate insensitivity during inhomogeneous flow.

Experimental support of the fluid layer model can be found in the vein pattern (Fig. 10) observed on the fracture surfaces of metallic glasses. It may be explained as characteristic of the breakup of a fluid layer between two solid surfaces and is similar to the vein pattern obtained by separating two solid surfaces containing a layer of fluid such as vaseline or grease. Questions about the applicability of the fluid layer model arise because the vein pattern is not unique to glassy metals (some crystalline materials also show this pattern, e.g., in shear rupture failure of 18 % Ni maraging steel) [81].

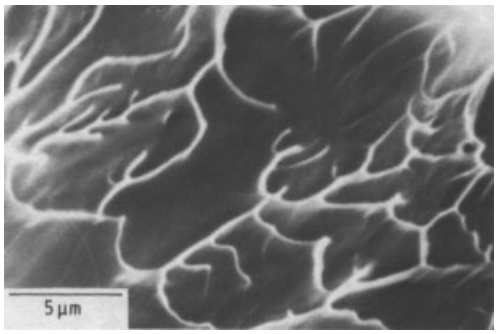

Figure 10. Fracture surface of $Al_{90}Fe_5Ce_5$ metallic glass [3]

4.1.5. Thermomechanical Interaction

The ductility of some amorphous alloys, (e.g., Fe, Co, Mn, and Cr based glasses), is sensitive to thermal treatment. These alloys embrittle upon annealing even though crystallization has not occurred. Figure 11 shows the surface bend strain at fracture in an Fe – Ni – P – B amorphous ribbon as a function of annealing temperature [82]. The annealing-induced embrittlement shows two well-defined stages: Stage I is characterized as fracture following plastic flow, while in Stage II no plastic flow occurs. DAVIS [81] has found that the loss of ductility begins at a temperature lower than that required for initial stress relief of the material.

Embrittlement varies with alloy composition. CHEN et al. [83] investigated the compositional dependence of the embrittlement for Fe – Co and Fe – Ni amorphous alloys. The embrittling temperature is defined as the annealing temperature at which a glass ribbon breaks into pieces when bent over a razor blade. Figure 12 shows the variation of crystallization temperature T_x, and the embrittlement temperature T_B for (Fe, Co)$_{75}$ PBAl and (Fe, Ni)$_{75}$ PBAl glasses with

the Co or Ni content. Mixing Fe, Co, or Ni with one another enhances embrittlement of the alloy glasses, as indicated by the lowering of T_B. This behavior contrasts with the increase in thermal stability of the glassy state, as evidenced by the increase in T_x upon alloying. Increasing the Ni content substantially raises the embrittlement temperature, and replacing Fe and Ni in these glasses with Cr and Mo lowers T_B.

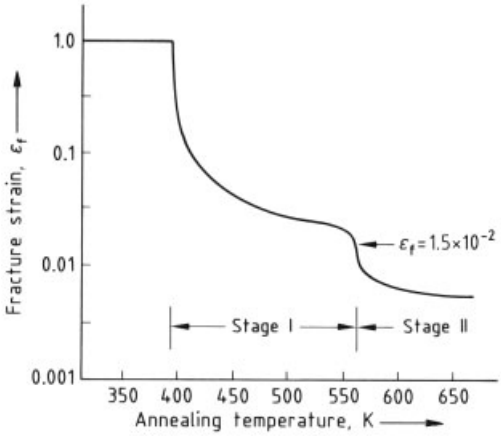

Figure 11. Relationship between surface bend strain at fracture ε_f and annealing temperature in a amorphous $Fe_{40}Ni_{40}P_{14}B_6$ ribbon [87]

So far the mechanisms of thermomechanical interaction are not clear. In some respects metallic glass embrittlement is phenomenologically similar to temper embrittlement of steel. In steels, temper embrittlement is attributed to a strong interaction between impurity and alloying elements which segregate to grain boundaries. In metallic glasses, phosphorus segregation has been observed to cause embrittlement [84]. It is suggested that structural relaxation that leads to local structural and compositional fluctuations may be responsible for the embrittlement of metallic glasses.

4.1.6. Fracture

Tensile specimens of metallic glass ribbon fail by fracture along shear bands oriented at $45 \pm 10°$ to the tensile axis. The fracture surface of the shear band is characterized by a vein pattern and a smooth shear offset region. The smooth region results from initial shear displacement along the shear band. When the shear becomes unstable, fracture occurs and is marked by a vein pattern.

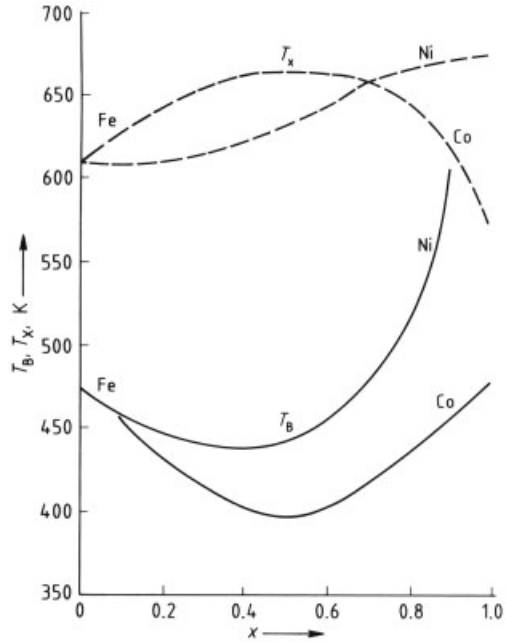

Figure 12. The dependence of embrittlement temperature T_B and crystallization temperature T_x on composition for $(Fe_{1-x}, Co_x)_{75}PBAl$ and $(Fe_{1-x}, Ni_x)_{75}PBAl$ glasses annealed for 20 h [83]

LEAMY et al. [85] first determined that the veins are hills on a flat background and suggested that shear disk cracks nucleate and spread through the shear band. At those places where the cracks run together, the material fails by internal necking, thereby producing smoothly rounded veins. The vein pattern results because of an instability that occurs during the motion of the interface between the fluid and the outside air. In a manner similar to that of dendrite formation during crystal growth, perturbations at the fluid – air interface grow as protrusions into the fluid layer. The protrusions then run into one another on an increasingly larger scale, resulting in the vein pattern. This growth is determined by the balance between the driving force for motion (the negative pressure gradient into the fluid) and an opposing force due to increased surface area (the fluid – air surface tension).

Shear tear tests have been used to determine the fracture toughness K_{IIIC} of metallic glass alloys [86]. In this case, the tearing energy G_t is given by

$$G_t = 2F/t$$

where F is the force required to extend the tear and t the specimen thickness. Therefore,

$$K_{IIIC} = \sqrt{G_t \mu}$$

where μ is the shear modulus.

The fracture toughness of single-edge-notch and center-cracked panels in tensile tests has been evaluated from the following relation [81]:

$$K_C \equiv \sigma_a \sqrt{\pi a}$$

where σ_a is applied tensile stress and a is the crack length.

Figure 13 compares the fracture toughness of metallic glasses with those of crystalline materials. This plot shows that the lower fracture toughness of the iron-based glasses is a natural consequence of their substantially increased yield strengths. On the other hand, the fracture toughness of metallic glasses is ca. 100 times higher than that of oxide glasses.

Considering that most as-quenched amorphous alloys do not form cracks before yielding in bend tests, Kimura and Masumoto [87] developed fracture parameters that describe the ductile behavior of amorphous alloys. These parameters include the plastic constraint factor and the microfactor stress.

The plastic constraint factor L of a double-notched tensile specimen is defined as follows:

$$L = \frac{P_{GY}}{\sigma_Y (W - 2a) t}$$

where P_{GY} is the general yield load, W the specimen width, a the notch depth, and t the specimen thickness. The plastic constraint factor was used to investigate amorphous $Pb_{78}Cu_6Si_{16}N$. A value of 2.4 was found for L, and this value is nearly equal to the upper limit of 2.96 for a deep crack predicted by slip-line field theory. This implies that this alloy is an almost perfectly ductile material.

The microfracture stress σ_f is used to express the fracture criterion for an amorphous, thin bend specimen:

$$\sigma_f = Q\sigma$$

where Q is the plastic concentration factor, and σ the yield stress. The loss of ductility after annealing can be explained as a decrease in the microfracture stress, and not as an inhibition of the slip process.

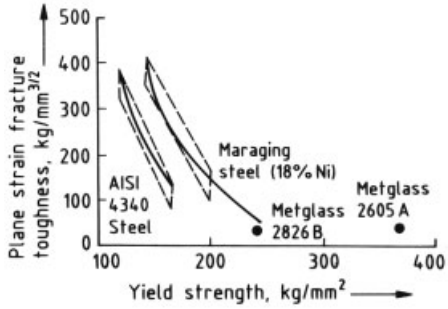

Figure 13. Fracture toughness of iron-based metallic glass (Metglas) compared to crystalline iron-based metals [81]

4.2. Magnetic Properties

The superior magnetic properties of metallic glasses provide one of the few commercial uses of this class of material [88], [89]. As a result of the random positions of atoms, the ferromagnetic magnetization vector should not have any special anisotropy. That this is not quite true is discussed later; however, this idea led to the strong motivation for the development of soft magnetic materials based on iron, cobalt, and nickel. In this section the soft magnetic properties of amorphous alloys are considered. The term soft refers to the fact that the response of the magnetization to an applied field is large; that is, the material is easily magnetized and demagnetized with low remanent magnetization. This large response of the magnetization coupled with low magnetic losses and low excitation power make amorphous magnetic alloys highly desirable in such applications as transformers and inductors [90]. Further, to be of commercial interest, the material should retain its ferromagnetic properties when heated to elevated temperature. Therefore, the exchange energy that keeps the spins aligned must be fairly large. Even though the coexistence of a long-range ordered ferromagnetic state with a structurally disordered material may, at first, seem surprising, ferromagnetism is usually taken to arise from the nearest neighbor interactions, which in dense random alloys are not very different from those

found in the equivalent crystalline material. This concept of the magnetic properties reflects the short-range order existing in the amorphous alloy, which, in general, mirrors the structure at the same composition in the corresponding crystalline alloys.

Almost all the compositions that produce useful magnetic amorphous structures stable at room temperature contain large fractions of transition-metal or rare-earth elements, many of which are ferromagnetic. More specifically, the class of alloys containing about 80 atom % of the transition metals Fe, Co, and Ni, and about 20 atom % of metalloid or glass-forming elements such as B, C, P, Ge, and Si, are strongly magnetic at room temperature. Since formation of the amorphous state is simpler (that is, lower cooling rates are sufficient) when more than one transition-metal element and metalloid element are employed, many of the available experimental data concern rather complex compositions. The transition metal content can be varied between ca. 75 and 84 atom %, and the alloys of interest magnetically are those containing mainly Fe, Ni, and Co. It is these alloys that offer potential important opportunities for engineering applications.

4.2.1. The Curie Temperature

In ferromagnetism the mutual interactive forces of atomic dipoles tend to align neighboring dipoles to one another, resulting in a permanent magnetic moment on a macroscopic scale. The transition temperature at which thermal agitation perturbs the parallel arrangement of neighboring spins such that the average net moment is zero is the Curie temperature T_C. The Curie temperature of binary iron-nickel, iron-cobalt, and nickel-cobalt alloys with a single metalloid or a combination of metalloid atoms such as B, P, Si, Ge, and C is given in Figure 14 [91]. The Curie temperature of these alloys varies with metalloid concentration in a systematic fashion. The figure can be summarized as follows [91], [92]:

1) T_C decreases with increasing Fe content, which suggests that the structure is behaving more like close-packed, rather than body-centered cubic iron.
2) Substituting Ni for Fe increases T_C only slightly or not at all up to an Fe/Ni ratio of 1 or slightly greater. Further replacement by Ni then decreases T_C in an approximately linear fashion to zero (for pure Ni).
3) The Fe – Co alloys generally show a broad maximum in T_C near a Fe/Co ratio of 1. The addition of Si significantly raises T_C.
4) Replacing Co by Ni results in a rapid drop in T_C in the Co – Ni alloys. Alloys richer than about 85 % Ni are nonmagnetic. Also, in Co – Ni alloys, the Curie point is not sharply defined and the scaling laws are not observed.

Other general observations include:

1) Alloys with one transition metal element have a more sharply defined T_C than alloys containing two or more transition metal elements. This implies that the range in T_C values is a result of chemical inhomogeneity, which may include clustering or phase separation, rather than an intrinsic property of homogeneous amorphous alloys.
2) Alloys rich in P have a tendency for lower values of T_C than alloys high in B, or B plus Si. Also, in the $Fe_x(PB)_y$ alloys, T_C increases as x decreases, which is to say that alloys of higher Fe content have lower Curie temperatures. However this is not always the case for iron only amorphous alloys. It may be that the T_C is sensitive to the Fe – Fe interatomic spacing, which is altered by both the metalloid species and concentration.
3) There is a suggestion [91] that in binary Fe amorphous alloys, the Curie temperature decreases in the order Fe – Si, Fe – B, Fe – C, Fe – P.
4) When other transition metals (i.e., Cr, Mn, Mo, V) are substituted for Fe the Curie temperature is dramatically lowered.

4.2.2. Saturation Magnetization

The saturation magnetization is the maximum magnetic moment (magnetization) per unit volume. The values for saturation magnetization or saturation induction have been reported for a large number of ferromagnetic amorphous alloys, with considerable scatter and disagreement in the results, which can be attributed in part to

the inherent difficulty of accurate magnetization measurements on small samples [91].

The saturation magnetization of binary iron – metalloid alloys varies almost linearly with metalloid content at low temperature [93]. As the temperature is increased above room temperature the saturation magnetization reaches a peak and then drops sharply because T_C is decreasing quickly with increasing metalloid content.

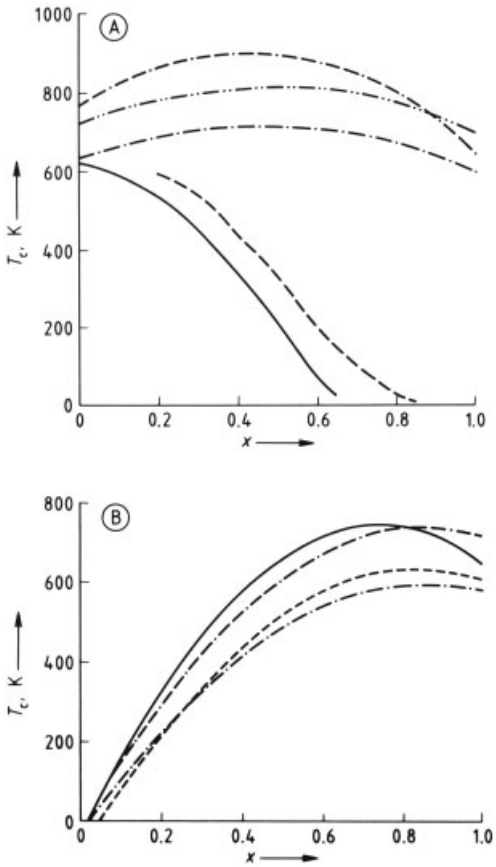

Figure 14. Dependence of Curie temperature on composition in glassy alloys
A) Iron – cobalt – metalloid and nickel – cobalt – metalloid alloys
– – · – $(Fe_xCo_{1-x})_{80}P_{10}B_{10}$; – · – · – $(Fe_xCo_{1-x})_{78}B_{12}Si_{10}$; – · – · $(Fe_xCo_{1-x})_{80}P_{13}C_7$; – – – – $(Ni_xCo_{1-x})_{78}P_{14}B_8$; — $(Ni_xCo_{1-x})_{75}P_{16}B_6Al_3$
B) Iron – nickel – metalloid alloys
– – · $(Fe_xNi_{1-x})_{78}B_{12}Si_{10}$; — $(Fe_xNi_{1-x})_{80}B_{20}$; - - - - $(Fe_xNi_{1-x})_{80}P_{14}B_6$; – · – · $(Fe_xNi_{1-x})P_{13}C_7$

KAZAMA et al., [94] found magnetization to be greatest in Fe_xB_{100-x}. Replacement of B by a different metalloid decreased the magnetization in the order P < C < Si < Ge, as would partially be expected on the basis of the charge transfer model. In similar studies LUBORSKY found P < Si < Ge, C, with the differences between Si, Ge, and C being small [92].

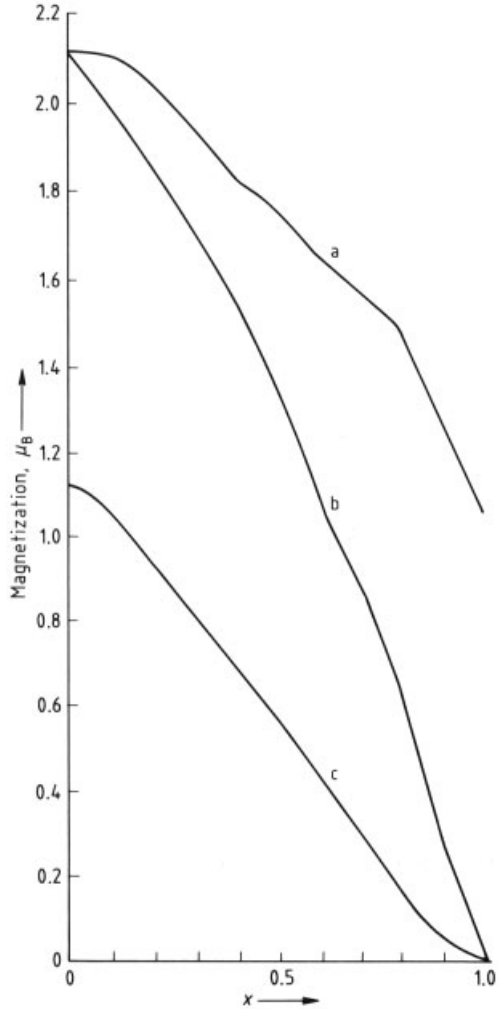

Figure 15. Magnetization (Bohr magnetons per metal atom) as a function of atomic fraction of transition metal
a) $(Fe_xCo_{1-x})P_{10}B_{10}$; b) $(Fe_xNi_{1-x})_{80}P_{10}B_{10}$; c) $(Co_xNi_{2-x})_{78}P_{14}B_8$

Figure 15 illustrates the magnetization in Bohr magnetons for Fe – Ni, Fe – Co, and Co – Ni amorphous alloys [93]. The magnetization is relatively insensitive to the metalloid content; this behavior is similar to crystalline Fe – Ni – Co alloys [95] except that the moments

are higher in the crystalline material (Fe: 2.2 to $2\,\mu_B$, Co: 1.6 to $1\,\mu_B$, Ni: 0.6 to $0\,\mu_B$). This change in the moment per atom in the amorphous state has been interpreted in a rigid-band model [91], in which the moment per metal atom is decreased because of electron contributions to the unfilled d-band by the metalloid elements. [96].

Employing a series of FePB amorphous alloys and altering only the P/B ratio led to the conclusion that boron-containing alloys have a slightly higher moment than those containing only phosphorus [97], [98]. Furthermore, a P/B ratio of 3 : 1 yields a maximum in the moment.

Finally, when transition metals to the left of iron in the periodic table (Mn, V, Cr) are added, a rapid drop in moment follows, with the drop being more rapid for Cr and V than for Mn [91]. The assumption is made that the moments of these atoms (Mn, V, Cr) align antiferromagnetically with the dominant iron moments.

The temperature dependence of magnetization has been measured for a variety of amorphous alloys. The decrease in magnetization with increasing temperature is observed to be more rapid than in crystalline metals. The distribution of atomic spacings which would, in turn, lead to a distribution of exchange interactions can explain this result [91].

4.2.3. Anisotropy

It might be thought that because metallic glasses are amorphous they should exhibit isotropic magnetic properties. They do not because they are not ideally homogeneous alloys, and real glasses have varying degrees of magnetic anisotropy. These result from chemical and structural vagaries which occur during electrodeposition or melt-quenching of the alloys. Stresses resulting from differences in cooling rates through the thickness of the ribbon can also lead to magnetostriction and, therefore, anisotropy. Processing or fabrication of the amorphous alloys contribute to further anisotropic behavior. Thermal annealing can help mitigate the anisotropy through removal of internal stresses.

Magnetostriction is the induced strain inside magnetic domains which results to minimize the elastic strain dependence of the anisotropy energy. Although generally small, it always leads to a contraction in the direction of the field. In spite of the size of the effect it is important in commercial practice because of the possibility of elastic energy being stored by the magnetostrictive force, and even small dimensional changes in laminated ferromagnetic transformer cores can be important. The saturation magnetostriction is expressed by λ_s. Values of magnetostriction in metallic amorphous materials are similar to those in crystalline alloys. Data for Fe – Ni and Fe – Co glasses exhibit nearly a linear decrease in λ_s (Fig. 16) [91]. Cobalt-rich Fe – Co alloys seem to show a negative λ_s. Due to the techniques used to obtain data (strain gages and cement employed) these values represent lower limits [92].

4.2.4. Low-Field Properties

In the application of metallic glasses the low-field properties are usually the controlling factors. Ferromagnetic materials subjected to alternating fields, as in transformers, must have a thin hysteresis loop, because of an energy loss per cycle that is determined by the area enclosed by the hysteresis loop. Because amorphous materials do not have grain boundaries or precipitates as pinning sites they have very low coercivity. However, pinning sites still remain and include defect structures in magnetostrictive alloys, structural rearrangements, and surface irregularities [99]. These can lead to varying degrees of coercive fields. To minimize the coercivity, defect structures must be reduced or elimiated. As the reversal of magnetization in amorphous alloys occurs by wall motion, wall pinning controls the coercive field. Herein surface roughness of the amorphous metal ribbons is a major factor contributing to the coercivity.

Amorphous alloys with high saturation magnetization have losses at 60 Hz that are much smaller than the best commercial grain-oriented Fe – 3.2 % Si alloys [100]. It is these low losses that make the application of amorphous metals in magnetic devices so attractive. Again these low losses are due to the lack of wall pinning centers as well as low anisotropy following stress-relief annealing, which reduces short-order and relaxation effects.

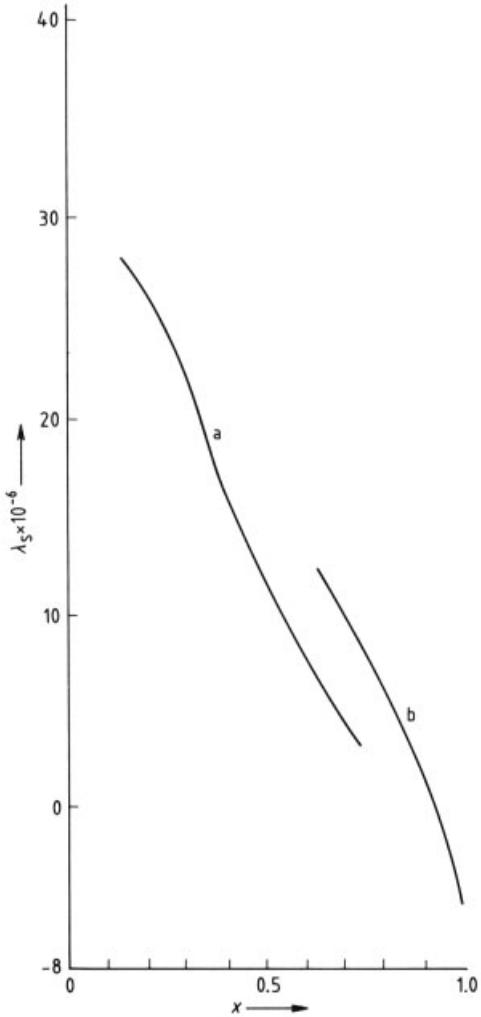

Figure 16. Saturation magnetostriction as a function of atomic fraction of transition metal at room temperature a) $(Fe_xNi_{1-x})_{80}B_{20}$; b) $(Fe_xCo_{1-x})_{80}B_{20}$

4.3. Electrical Properties [2], [101]

The electrical resistivity of metallic glasses is relatively high (i.e., $> 1\,\mu\Omega \cdot m$ at room temperature). In addition to the high electrical resistivity, the temperature coefficient is very small (in some cases, it can be zero or negative) and lies in the range of $\pm 10^{-4}\,K^{-1}$. As a result, glassy metals posses a high residual resistivity near 0 K. This in part stems from the random atomic arrangement of the amorphous structure, which effectively scatters conduction electrons. The randomness of the structure, however, limits the phonon contribution to scattering and leads to the small temperature coefficient of resistivity. Figure 17 is a resistivity – temperature curve for an $Fe_{75}P_{15}C_{10}$ glassy alloy [102]. The temperature dependence for the amorphous phase is small until 675 K when the alloy crystallizes. The residual resistivity near 0 K for the amorphous Fe – P – C alloy is ca. 4 times the values of the crystalline phase.

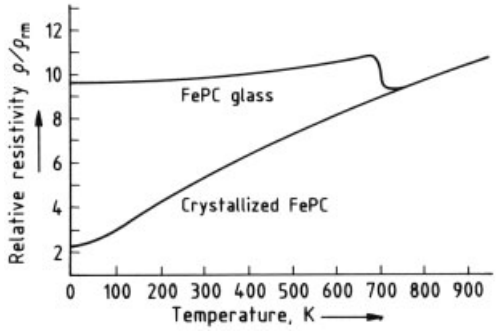

Figure 17. Variation in relative resistivity (ϱ/ϱ_{rm}) with temperature for the $Fe_{75}P_{15}C_{10}$ glass alloy [102]

The electrical resistivity – temperature curve for some amorphous alloys possesses a minimum below which the resistivity increases with decreasing temperature. Compared to crystalline materials, the minimum can occur at a relatively high temperature. In amorphous $Pd_{80}Si_{20}$ containing 7 atom % Cr, the minimum occurs at 500 K, compared to a temperature of 20 K for Au alloys containing small amounts of Fe [101]. Figure 18 shows this effect for a $Pd_{77.5}Si_{16.5}Cu_6$ alloy that contains 700 ppm Fe as an impurity [103]. The resistance minimum occurs at about 14 K. The curve represented by the dashed line shows what happens to the alloy when a magnetic field (0.7 T) is applied. This implies that the increase in resistance is due to magnetic scattering.

Superconducting amorphous metals and alloys have been obtained in thin film form since 1954 by vapor deposition of bismuth and gallium on cryogenic substrates [104]. However, these early films were unstable and they crystallized at temperatures as low as 50 K. New processing techniques have produced superconducting metallic glasses from the liquid state that are stable at room temperature [105]. The superconducting transition temperatures for these alloys lie between 1.5 and 8.7 K. Table 2 contains data

on superconducting amorphous metals and alloys [106]. These values are extremely low in comparison to conventional metal superconductors (ca. 23 K) and the newer high T_c superconducting oxides (> 90 K). One advantage that metallic glasses may have as superconductors is their amorphous structure. A potential problem with crystalline superconductors (e.g., Nb_3Sn) is the decrease in transition temperature (i.e., from 18 K to less than 4 K) when subjected to radiation. A nuclear fusion reactor is one such environment where superconducting materials could possibly be used and the effect of radiation on structure and properties is of critical importance. The progressive degradation in properties of the crystalline alloy stems from the disordering of the lattice due to neutron bombardment. The disordered structure of glassy metals is relatively unaffected by this type of environment [2].

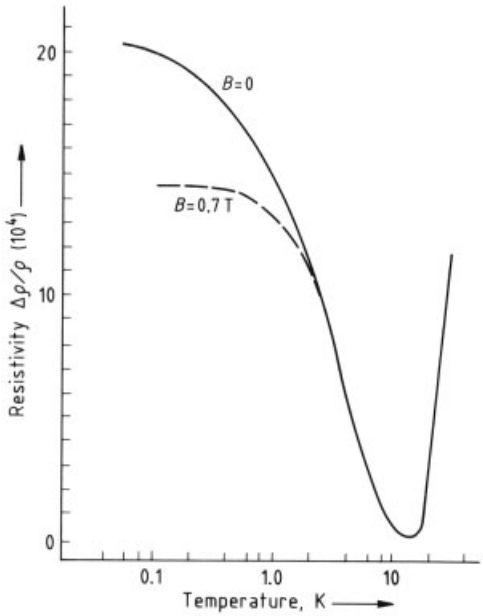

Figure 18. Low temperature resistivity of $Pd_{77.5}Si_{16.5}Cu_6$ wire in applied fields ($B = 0$ and 0.7 T) [103]

4.4. Chemical Properties [107]

Although the amorphous structure in glassy alloys is thermodynamically metastable, these materials are free of the defects usually associated with crystalline materials, such as grain boundaries, dislocations, and stacking faults; that is, they are free of regions of local electrochemical potential difference. Furthermore, solid-state diffusion is suppressed during rapid quenching. Therefore, metallic glasses are free of the second phases, precipitates, and segregates which usually occur during slow cooling or heattreatment associated with the processing of conventional alloys. The amorphous alloys can therefore be regarded as ideally homogeneous materials in a chemical sense with respect to the surface-related characteristics pertaining to corrosion.

Table 2. Superconductivity transition temperature for selected amorphous materials [106]

Material	T_c, K	Preparation method ∗
$Pb_{90}Cu_{10}$	6.5	VQ
$Sn_{90}Cu_{10}$	6.76	VQ
$La_{80}Au_{20}$	3.5	LQ
$Mo_{48}Ru_{32}B_{20}$	6.05	LQ
$Zr_{75}Rh_{25}$	4.55	LQ
$Nb_{60}Rh_{40}$	4.8	LQ
Nb_3Ge	3.6	S

∗ VQ = vapor quenched; LQ = liquid quenched; S = sputtered

Before the advent of glassy metals it was generally felt that the formation of amorphous passive films would enhance corrosion resistance [108]. Amorphous alloys, being noncrystalline, have an added advantage over crystalline materials in that matching of a substrate alloy with a passive film is more easily accomplished with an amorphous material than trying to find two crystalline alloys for which a quasi-perfect epitaxial relationship exists [109]. Thus, high corrosion resistance can be provided by a chemically homogeneous amorphous alloy. What is especially advantageous in glassy metals is the ability to increase the element beneficial in corrosion resistance while maintaining the chemical homogeneity of the alloy.

An amorphous metal – metalloid alloy containing only one metallic component has a lower corrosion resistance than the pure metal. As an example, the corrosion rate for an amorphous iron – metalloid system without a second metallic element is higher than that of crystalline iron [110]. The corrosion rate decreases, however, with the addition of almost all second elements (e.g., Ti, Zr, V, Cr, Mo, W, Co, Ni, etc.) [111]. In particular, the addition of Cr is very effective; passivation occurs spontaneously

in $Fe_{72}Cr_8P_{13}C_7$ in 2 M hydrochloric acid at room temperature [112]. Figure 19 compares current densities for amorphous $Fe_{70}Cr_{10}P_{13}C_7$ and 18 – 8 stainless steel as a function of time in a sulfuric acid solution [113]. The rate of corrosion for the amorphous Fe – Cr – P – C alloy is two orders of magnitude lower than for the stainless steel. These alloys also do not undergo pitting corrosion, even by anodic polarization, up to the transpassive region of chromium. Chromium also improves the corrosion resistance of amorphous cobalt – and nickel – metalloid alloys.

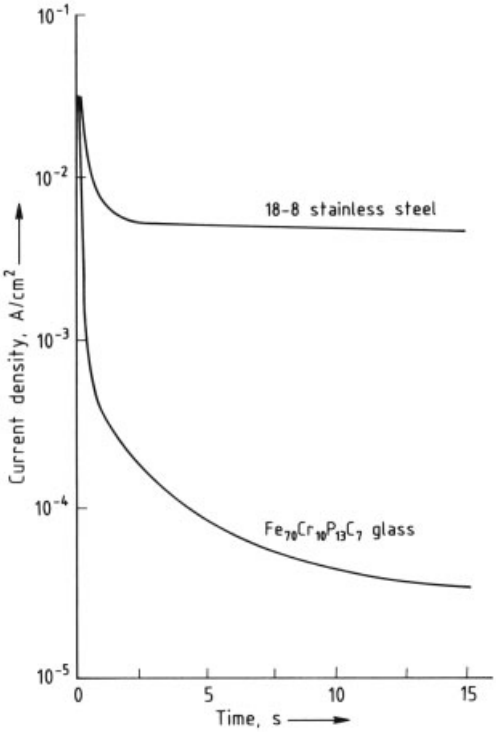

Figure 19. Current densities for amorphous $Fe_{70}Cr_{10}P_{13}C_7$ glass and crystalline 18 – 8 stainless steel as a function of time in 1 M H_2SO_4 + 0.5 M NaCl (300 mV) [113]

Amorphous metal – metal alloys are usually not corrosion resistant. Their corrosion resistance is always lower than that of the crystalline pure metal component with the highest corrosion resistance or the highest passivating ability [114]. Thus, the overall corrosion resistance is determined by the presence of a passivating element (for example, increasing Si content in an Fe – Si metallic glass improves the corrosion resistance due to the passivating ability of silicon) rather than the fact that the alloy is amorphous [115].

5. Uses of Metallic Glasses

5.1. Mechanical Uses

Although they are strong and inherently fracture tough, metallic glasses are rarely used as structural materials due to the shearing instability they manifest and size limitations resulting from their mode of processing, in which at least one dimension of the product must be small enough to obtain extremely high cooling rates. Plastic flow in metallic glasses occurs in the form of localized shear bands. When multiple shear bands are initiated, metallic glasses can exhibit high macroscopic, as well as microscopic, ductility. But in uniaxial tension, tensile failure is coincident with yielding, because of the highly localized shearing and the absence of strain hardening. The discovery of aluminum glasses [3], [4] has greatly increased interest in metallic glasses as structural materials [116]. Their great asset is their strength, which varies from 670 to 1140 MPa. These values are much higher than the best aluminum alloys (300 – 500 MPa) but suffer from the disadvantage of a lower Young's modulus (up to 71 GPa) [3]. These materials, with their low density and high strength, should be useful as reinforcing composites.

Composites containing metallic glasses may offer means to avoid the problems of low macroscopis ductility and toughness. Composite structures can provide an extrinsic mechanical constraint and also allow use of metallic glasses in bulk form. The combination of high strenght and isotropic two-dimensional form gives amorphous metals promise for planar composites. In fact, high-quality plates have been made of 1.5 cm Metglas [117] ribbons in 0 – 90 ° cross ply with an epoxy matrix [118]. These composites revealed a remarkable degree of in-plane isotropy and strength.

In cement reinforcement, amorphous metal ribbon fibers have been shown to attain superior performance [119]. A 2000-fold increase in specific fracture work was obtained with only 1 vol % of fiber. To improve cohesion to the mortar, the surfaces of the Metglas fibers were roughened by sandblasting.

Another commercial application for amorphous metal is brazing foils. Here, the eutectic alloys, based on nickel, nickel – palladium, and copper, are replacing many conventional brazing materials that have high silver and gold contents. High-strength, high-temperature metal joints are often produced by brazing with foils of eutectic compositions which easily flow into the joint at temperatures as low as possible, but still significantly above the service temperature of the joint. The strong and flexible amorphous metal brazing ribbon conforms well to curved surfaces and provides an optimal, controlled gap. An added advantage is that the glasses contain no binder, and are therefore noncontaminating and have a very long shelf life. In addition, the previously used noble-metal-based crystalline foils can be replaced by amorphous compositions of cheaper materials [89].

5.2. Magnetic Uses

Magnetic uses of metallic glasses already provide a commercial market and are the most promising uses of these materials. The primary commercial application of metallic glasses is in transformer cores. There are about 20 000 000 distribution transformers now in service in the United States alone. In 1984 new distribution transformers were sold at the rate of about 1 000 000 per year [120]. Electric power applications is the only major commercial venture for metallic glasses. The hysteresis of magnetic materials results in energy losses when they are used for cores in high-frequency voltage transformers. Metallic glasses are attractive because with amorphous-metal cores, core loss can be reduced by as much as 50 % compared to conventional transformers [120], depending on specific geometry and design as well as pattern of use [124]. In the United States about 2 % (energy costs greater than $ 10^8 per year) of the total electricity generated is lost in the last stage step-down elements in power distribution transformers before the consumers receive the electricity [122], [123]. Here, the use of amorphous alloy cores could significantly cut energy losses. If all distribution transformers in service in the United States were replaced by amorphous metal transformers, the power savings would amount to about 2000 MW per year [120]. The impact would be much less meaningful in Europe due to the difference in the distribution system. The energy loss in the distribution transformer includes both the magnetic loss in the transformer core and the winding loss in the copper windings. The magnetic loss in the amorphous alloy core is much less compared to silicon steel which is conventionally used for transformers. Also attractive is that the melt-quenched alloys are already in thin ribbon form when produced. This not only reduces the eddy current loss, but is less costly to produce than rolling thin crystalline ribbon tapes to similar dimensions. Although the loss is similar to the nonmetallic ferrite cores in the frequency range up to 10^5 – 10^6 Hz, because of their higher saturation induction the amorphous alloys are more attractive in terms of performance per volume [89], [121].

Some other uses include magnetomechanical transducers that utilize the high magnetoelastic coupling and high elastic limit of amorphous alloys. Based upon this, various applications such as displacement or force sensors have been proposed [125]. Magnetoacoustic applications are feasible because the amorphous alloys are atomistically homogeneous; not only do the domain walls easily move, but ultrasonic sound is scattered only slightly [124]. This contrasts with crystalline solids, in which sound is scattered by lattice defects such as grain boundaries and dislocations. Amorphous alloys are an excellent choice for magnetoacoustic applications, such as delay lines, due to strong magnetoelastic coupling and low acoustic attenuation. In recording and memory applications the sputter-deposited rare earth – transition metal (RE – TM) films can have a strong enough magnetic anisotropy with the easy axis perpendicular to the surface of the film. Magnetic bubble domains can therefore be sustained [126]. Recent interest in magnetooptical recording has revived interest in the use of RE – TM films as a mass storage medium.

A fabric woven from tapes has been used for magnetic shielding because it can be cut and bent into shape without degrading its magnetic performance. Also, the combination of high permeability and high yield strength can be utilized in applications requiring a shield that also serves as a spring (e.g., magnetic tape cassettes). Cobalt-based metallic glasses are being applied as magnetic tape recording head materials. These materials are processed (quenched) between twin

rollers to produce 50-µm thick and 15-mm wide amorphous tapes with acceptable smoothness on both sides. The high flux density coupled with excellent wear resistance results in recording heads produced from these tapes which are superior in overall performance to ferrite, permalloy, and Sendust heads. They have been suggested for use in audio, video, computer, and instrumental recording machinery [127].

The high magnetic permeability of the ferromagnetic amorphous metals makes them good generators of harmonics when placed in appropriate a.c. and d.c. fields [124]. This characteristic makes them useful as antitheft tags for protecting store merchandise or if planted in book spines in libraries and bookstores. The usual system consists of field-generating coils placed around a passageway and a detector of high order harmonics. Since amorphous strips have no crystals, bending does not change the magnetic distortion as it does with conventional crystalline materials.

6. References

1. R. W. Cahn in R. W. Cahn, P. Haasen (eds.): *Physical Metallurgy,* 3rd ed., Elsevier Science Publ., Amsterdam 1983, p. 1779.
2. P. Duwez in R. G. Lerner, G. L. Trigg (eds.): *Concise Encyclopedia of Solid State Physics,* Addison-Wesley, Reading 1983, p. 112.
3. Y. He, S. J. Poon, G. J. Shiflet, *Science (Washington, D.C.)* **241** (1988) 1640. G. J. Shiflet, Y. He, S. J. Poon, *Scr. Metall.* **22** (1988) 1661. Y. He, S. J. Poon, G. J. Shiflet, *Scr. Metall.* **22** (1988) 1813. G. J. Shiflet, Y. He, S. J. Poon, *J. Appl. Phys.* **64** (1988) 6863.
4. A. Inoue, K. Ohtera, A. P. Tsai, T. Masumoto, *Jpn. J. Appl. Phys. Part 2* **27** (1988) L 479. A. P. Tsai, A. Inoue, T. Masumoto, *Metall. Trans A* **19 A** (1988) 1369.
5. H. A. Davis in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 8.
6. M. G. Scott in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 144.
7. M. G. Scott in M. B. Bever (ed.): *Encyclopedia of Material Science and Engineering,* Pergamon Press, Oxford 1986, p. 2968.
8. A. L. Greer, *Acta Metall.* **30** (1982) 171.
9. D. G. Morris, *Scr. Metall.* **16** (1982) 585.
10. L. Tanner and R. Ray, *Scr. Metall.* **14** (1980) 1657.
11. J. Burke: *The Kinetics of Phase Transformations in Metals,* Pergamon Press, Oxford 1965, p. 192.
12. J. L. Finney in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 42.
13. F. Spaegen in M. B. Bever (ed.): *Encyclopedia of Material Science and Engineering,* Pergamon Press, Oxford 1986, p. 2976.
14. P. H. Gaskell, *J. Phys. C* **12** (1979) 4337.
15. J. D. Bernal, *Proc. R. Soc. London A* **284** (1964) 299.
16. P. H. Gaskell, *J. Non-Cryst. Solids* **32** (1979) 207.
17. D. S. Boudreaux, J. M. Gregor, *J. Appl. Phys.* **48** (1977) 152, 2057.
18. M. Mehra, A. R. Williams, W. L. Johnson, *Phys. Rev. B Condens. Matter* **28** (1983) 624.
19. L. von Mehimendahl, J. Hafner, *Phys. Rev. Lett.* **42** (1979) 386.
20. J. Hafner, *Phys. Rev. B Condens. Matter* **21** (1980) 406.
21. M. G. Scott in M. B. Bever (ed.): *Encyclopedia of Material Science and Engineering,* Pergamon Press, Oxford 1986, p. 2974.
22. H. H. Liebermann in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 26.
23. G. K. Wehner, G. S. Anderson In L. I. Maissel, R. Glang (eds.): *Handbook of Thin Film Technology,* McGraw-Hill, New York 1970, p. 1 – 3.
24. M. R. Bennett, J. G. Wright, *Phys. Status Solidi A* **13** (1972) 135. P. K. Leung, J. G. Wright, *Philos. Mag.* **30** (1974) 995. D. E. Polk, C. E. Dube, B. C. Giessen in B. Cantor (ed.): *Proc. 3rd Int. Conf. Rapidly Quenched Metals,* The Metals Society, London 1978, p. 220.
25. H. Jones, *J. Mater. Sci.* **19** (1984) 1043.
26. L. Mendoza-Felis et al., *Phys. Rev. B Condens. Matter* **26** (1982) 1306.
27. P. Duwez, R. H. Willens, W. Klement, *J. Appl. Phys.* **31** (1960) 1136.
28. P. Pietrokowsky, *Rev. Sci. Instrum.* **34** (1963) 455.
29. D. R. Harbur, J. W. Anderson, W. J. Maraman, *Trans. Metall. Soc. AIME* **245** (1969) 1055.
30. M. Ohring, A. Haldepier, *Rev. Sci. Instrum.* **42** (1971) 530.
31. P. Ramachandrarao, M. Laridjani, R. W. Cahn, *Z. Metallk.* **63** (1972) 43.
32. R. W. Cahn et al., *Mater. Sci. Eng.* **23** (1976) 83.

33. M. Moss, D. L. Smith, R. A. Lever, *Appl. Phys. Lett.* **5** (1964) 120. K. Yamauchi, Y. Nakagawa, *Jpn. J. Appl. Phys.* **10** (1971) 1730.
34. G. Thursfield, H. Jones, *J. Phys E.* **4** (1971) 675. P. H. Shengir, K. Shimomura, R. Ozaki, *Trans. Jpn. Inst. Met.* **20** (1979) 33.
35. K. D. Krishananand, R. W. Cahn in N. J. Grant, B. C. Giessen (eds.): *Rapidly Quenched Metals,* Sect. I, MIT Press, Cambridge 1976, p. 97. A. R. E. Singer in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Div., Baton Rouge 1978, p. 154. M. H. Kim, H. Jones in B. Cantor (ed.): *Proc. 3rd Int. Conf. Rapidly Quenched Metals,* **vol. 1,** The Metals Society, London 1978, p. 171.
36. J. W. Mottern, W. J. Privott (eds.): *Spinning Wire from Molten Metal,* American Institute of Chemical Engineers, New York 1978, p. 31.
37. I. G. Butler, W. Kurz, J. Gillot, B. Lux, *Fibre Sci. Technol.* **5** (1972) 243.
38. R. B. Pond, US 2 976 590, 1961. R. D. Schile, US 3 461 943, 1969.
39. H. Weisner, J. Schneider, *Phys. Status Solidi A* **26** (1974) 71. S. Kavesh, US 3 845 805, 1974. T. Masumoto, I. Ohnaka, A. Inoue, M. Hagiwara, *Scr. Metall.* **15** (1981) 293. I. Ohnaka et al., in T. Masumoto, K. Suzuki (eds.): *Proc. 4th Int. Conf. Rapidly Quenched Metals,* vol. I, Japan Institute of Metals, Sendai 1982, p. 3.
40. J. C. Hubert, F. Mollard, B. Lux, *Z. Metallk.* **64** (1973) 835.
41. C. E. Mobley, A. H. Clauer, B. A. Wilcox, *J. Inst. Met.* **100** (1972) 142. R. E. Maringer, C. E. Mobley, *J. Vac. Sci. Technol.* **11** (1974) 1067.
42. R. E. Marenger, C. E. Mobley, E. W. Collings in N. J. Grant, B. C. Giessen (eds.): *Rapidly Quenched Metals,* Sect. I, MIT Press, Cambridge 1976, p. 29.
43. E. A. Strange, C. H. Pim, US 905 758, 1908.
44. R. B. Pond, US 2 825 108, 1958; US 2 910 744, 1959. C. E. Mobley, US 3 710 842, 1973.
45. H. H. Liebermann in B. Cantor (ed.): *Proc. 3rd Int. Conf. Rapidly Quenched Metal,* **vol. I,** The Metals Society, London 1978, p. 34. M. Matsuura, M, Kikuchi, M. Yagi, K. Suzuki, *Jpn. J. Appl. Phys.* **19** (1980) 1781.
46. R. Pond, R. Maddin, *Trans. Metall. Soc. AIME* **245** (1969) 2475.
47. D. E. Polk, US 3 881 542, 1975. S. Kavesh, US 3 881 540, 1975. H. S. Chen, C. E. Miller, *Mater. Res. Bull.* **11** (1976) 49. J. R. Bedell, N. Hemmat, D. E. Polk, US 4 184 532, 1980.
48. H. H. Liebermann, C. D. Graham, *IEEE Trans. Magn.* **MAG-12** (1976) 921. S. Kavesh in J. J. Gilman, H. J. Leamy (eds.): *Metallic Glasses,* American Society of Metals, Metals Park 1978, p. 36 H. H. Liebermann, *Mater. Sci. Eng.* **43** (1980) 203.
49. H. S. Chen, C. E. Miller, *Rev. Sci. Instrum.* **41** (1970) 1237. T. Suzuki, M. Koda, Y. Hashimoto, R. Sato, *Rev. Sci. Instrum.* **51** (1980) 550. S., Uedaira, S. Ito, US 4 212 344, 1980.
50. M. C. Narasimhan, US 4 142 571, 1979.
51. H. H. Liebermann, *IEEE Trans. Magn.* **MAG-15** (1979) 1393.
52. H. H. Liebermann, *J. Appl. Phys.* **50** (1979) 6773.
53. S. Takayama, T. Oi, *J. Appl. Phys.* **50** (1979) 4962.
54. B. H. Kear, J. W. Mayer, J. M. Poate, P. R. Struti in J. K. Tien, J. F. Elliot (eds.): *Metallurgical Treatises,* AIME, Warrendale 1981, p. 321.
55. E. M. Breinan, B. H. Kear, C. M. Banas, *Phys. Today* **29** (1976) Nov., 44.
56. S. M. Coplex et al. in B. Cantor (ed.): *Proc. 3rd Int. Conf. Rapidly Quenched Metals,* **vol. I,** The Metals Society, London 1978, p. 147.
57. E. M. Breinan, D. B. Snow, C. O. Brown, B. H. Kear in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Dev., Baton Rouge 1978, p. 440.
58. S. A. Miller in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 506.
59. N. J. Grant in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Div., Baton Rouge 1978, p. 230. V. Anand, A. J. Kaufman, N. J. Grant in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* II, Claitor's Publ. Div., Baton Rouge 1980, p. 273.
60. R. J. Dain, US 4 191 516, 1980. R. J. Grandzol, J. A. Tallmadge, *Int. J. Powder Metall. Powder Technol.* **11** (1975) 103.
61. S. A. Miller, R. J. Murphy, *Scr. Metall.* **13** (1979) 673. S. A. Miller, R. J. Murphy in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* II, Claitor's Publ. Div., Baton Rouge 1980, p. 385.

62. M. R. Glickstein, R. J. Patterson, N. E. Schockey in R. Mehrabian, B. H. Kear, and M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Div., Baton Rouge 1978, p. 46.C. C. Thompson, US 2 284 394, 1980.P. R. Holiday, A. R. Cox, R. J. Patterson in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Div., Baton Rouge 1978.
63. A. R. E. Singer, A. D. Roche, L. Day, *Powder Metall.* **23** (1980) 81. A. R. E. Singer, A. D. Roche in E. N. Aqua, C. I. Whitman (eds.): *Modern Developments in Powder Metallurgy,* Metal Powder Industries Federation, Princeton 1977, p. 127.
64. T. Yamaguchi, K. Narita, *IEEE Trans. Magn.* **MAG-13** (1977) 1621. A. E. Berkowitz, J. L. Walter in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* II, Claitor's Publ. Div., Baton Rouge 1980, p. 294.
65. C. F. Cline, R. W. Hopper, *Scr. Metall.* **11** (1977) 1137. C. F. Cline, J. Mahler, F. Milton, W. Kuhl, R. Hopper in R. Mehrabian, B. H. Kear, M. Cohen (eds.): *Rapid Solidification Processing: Principles and Technologies,* Claitor's Publ. Div., Baton Rouge 1978, p. 380.
66. B. S. Berry in J. J. Gilman, H. J. Leamy (eds.): *Metallic Glasses,* American Society of Metals, Metals Park 1978, p. 161.
67. D. Weaire, M. F. Ashby, J. Logan, M. J. Weins, *Acta Metall.* **19** (1971) 779.
68. A. S. Aronin, N. P. Kobelev, A. V. Serebryakov, Y. M. Soifer, *Fiz. Tverd. Tela (Leningrad)* **30** (1988) 3160.
69. A. S. Argon, H. Y. Kuo, *J. Non-Cryst. Solids* **37** (1980) 241.
70. K. Kamagaki, S. Abe, H. Fujimori: "Elastic Properties of Amorphous Ferromagnets," *Conf. Magnetic Properties of Amorphous Metals,* Benalmadena, Spain,25 – 29 May 1987.
71. I. V. Zolotukhin, Y. E. Kalinin, I. B. Kelalo, B. G. Sukhodulov, *Metallofizika (Akad. Nauk Ukr. SSR Otd. Fiz.)* **6** (1984) 58.
72. F. Spaepen, A. J. Taub in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 231.
73. V. A. Khomik, V. A. Zelenskii, *Fiz. Met. Metalloved.* **61** (1986) 602.
74. T. Masumoto, R. Maddin, *Acta Metall.* **19** (1971) 725.
75. J. C. M. Li in: *Metallic Glasses,* American Society of Metals, Metals Park 1978, p. 224.
76. P. Chaudhari, A. Leve, P. Steinhardt, *Phys. Rev. Lett.* **43** (1979) 1517.
77. P. Chaudhari, F. Spaepen, P. Steinhardt in H. J. Guntherodt (ed.): *Glassy Metals,* II, Springer-Verlag, Berlin 1982, 163.
78. C. A. Pampillo, *J. Mater. Sci.* **10** (1975) 1194.
79. F. Spaepen, D. Turnbull, *Scr. Metall.* **8** (1974) 563.
80. F. Spaegen, *Acta Metall.* **25** (1977) 407.
81. L. A. Davis in: *Metallic Glasses,* American Society of Metals, Metals Park 1978, p. 190.
82. E. Kimura, D. G. Ast in T. Masumoto and K. Suzuki (eds.): *Proc. 4th Int. Conf. Rapidly Quenched Metals,* **vol. II,** Japan Institute of Metals, Sendai 1982, p. 457.
83. H. S. Chen, K. A. Jackson in: *Metallic Glasses,* American Society of Metals, Metals Park 1978,p. 74.
84. F. E. Luborsky, J. L. Walter, *J. Appl. Phys.* **47** (1976) 3638.
85. H. J. Leamy, H. S. Chen, T. J. Wang, *Metall. Trans.* **3** (1972) 699.
86. H. Kimura, T. Masumoto, *Scr. Metall.* **9** (1975) 211.
87. H. Kimura, T. Masumoto in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 187.
88. F. E. Luborsky, L. A. Johnson, *J. Physiol. (Paris)* **41** (1980) C 8, 820 – 26.
89. D. Raskin, C. H. Smith in F. E. Lubursky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, pp. 381 – 400.
90. F. E. Luborsky: *Amorphous Metallic Alloys,* Butterworths, London 1983, pp. 360 – 400.
91. C. D. Graham, Jr., T. Egami, *Ann. Rev. Mater. Sci.* **8** (1978) 423 – 57.
92. F. E. Luborsky, J. D. Livingston, G. Y. Chin: *Physical Metallurgy,* North Holland Publishers, Amsterdam 1983, pp. 1674 – 1734.
93. F. E. Luborsky: *J. Magn. Magn. Mater.* **7** (1978) 143.
94. N. S. Kazama, M. Mitera, T. Masumoto: *Proc. Third Int. Conf. on Rapidly Quenched Metals,* The Metals Society, London 1978, vol. 2, p. 164.
95. T. Mizoguchi, K. Yamauchi, H. Miyajima: *Amorphous Magnetism,* Plenum Press, New York 1973, pp. 325 – 330.
96. T. Mizoguchi, *AIP Conf. Proc.* **34** (1976) 286 – 291.
97. J. Durand: *IEEE Trans. Magn.* **MAG-12,** 945 – 947.
98. J. Durand, M. Yung: *Amorphous Magnetism II,* Plenum Press, New York 1977, pp. 275 – 288.

99. H. Kronmuller, *J. Appl. Phys.* **52** (1981) 1859.
100. F. E. Luborsky, J. J. Becker, *IEEE Trans. Magn.* **MAG-15** (1979) 1939.
101. K. V. Rao in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 401.
102. P. Duwez, *Annu. Rev. Mater. Sci.* **6** (1976) 83.
103. H. Dierker, H. Gudmundsson, A. C. Anderson, *Solid State Commun.* **29** (1979) 767.
104. W. Buckel, R. Hilsch, *Z. Phys.* **138** (1954) 109, 118.
105. W. L. Johnson, S. J. Poon, P. Duwez, *Phys. Rev. B Solid State* **11** (1975) 150. W. L. Johnson, S. J. Poon, *J. Appl. Phys.* **46** (1975) 1787.
106. W. L. Johnson in R. W. Cahn, P. Haasen (eds.): *Physical Metallurgy,* 3rd ed., Elsevier Science Publ., Amsterdam 1983, p. 1735.
107. K. Hashimoto in F. E. Luborsky (ed.): *Amorphous Metallic Alloys,* Butterworths, London 1983, p. 471.
108. T. P. Hoar, *J. Electrochem. Soc.* **117** (1970) 17 C.
109. A. G. Revesz, J. Kruger in F. P. Frankenthal, J. Kruger (eds.): *Passivity of Metals,* The Electrochemical Society, Princeton 1978, p. 137.
110. T. Masumoto, K. Hashimoto, M. Naka in B. Cantor (ed.): *Proc. 3rd Int. Conf. Rapidly Quenched Metals,* The Metals Society, London 1978, p. 435. M. Naka. K. Hashimoto, A. Inoue, T. Masumoto, *J. Non-Cryst. Solids* **31** (1979) 347. P. Cadet, M. Keddam, H. Takenouti in T. Masumoto, K. Suzuki (eds.): *Proc. 4th Int. Conf. Rapidly Quenched Metals,* **vol. II,** Japan Institute of Metals, Sendai 1982, p. 1447.
111. K. Hashimoto, M. Naka, T. Masumoto, *Sci. Rep. Res. Inst. Tohoku Univ. Ser. A* **26** (1978) 48. M. Naka, K. Hashimoto, T. Masumoto, *Sci. Rep. Res. Inst. Tohoku Univ. Ser. A* **26** (1978) 283. M. Naka, K. Hashimoto, T. Masumoto, *J. Non-Cryst. Solids* **29** (1978) 61. K. Hashimoto, K. Asami, M. Naka, T. Masumoto, *Sci. Rep. Res. Inst. Tohoku Univ. Ser. A* **27** (1979) 237. M. Naka, K. Hashimoto, T. Masumoto, *J. Non-Cryst. Solids* **31** (1979) 355. P. Kovacs et al. in T. Masumoto, K. Suzuki (eds.): *Proc. 4th Int. Conf. Rapidly Quenched Metals,* **vol. II,** Japan Institute of Metals, Sendai 1982, p. 1471.
112. K. Kobayaski, K. Hashimoto, T. Masumoto, *Sci. Rep. Res. Inst. Tohuku Univ. Ser. A* **29** (1981) 284.
113. K. Hashimoto, T. Masumoto, *Mater. Sci. Eng.* **23** (1976) 285.
114. M. Naka, K. Hashimoto, T. Masumoto, *J. Non-Cryst. Solids* **30** (1978) 29.
115. V. Brusic, R. D. MacInnes, J. Aboaf in R. P. Frankenthal, J. Drueger (eds.): *Passivity of Metals,* The Electrochemical Society, Princeton 1978, p. 170.
116. R. W. Cahn, *Nature (London)* **341** (1989) 183.
117. METGLAS is a trademark of Allied-Signal Corp., Parsippany, New Yersey, USA.
118. Y. T. Yeow, *J. Compos. Mater.* **Suppl. 14** (1980) 1.
119. A. S. Argon, G. W. Hawkins, H. Y. Kuo, *J. Mater. Sci.* **14** (1979) 1707.
120. R. T. Dawn, *Mach. Des.* **56** (1984) April 26, 50.
121. W. Kunz, D. Graetzer, *J. Magn. Magn. Mater.* **19** (1980) 183.
122. F. E. Luborsky, J. J. Becker, H. H. Liebermann in B. Cantor (ed.): *Rapidly Quenched Metals III,* **vol. 2,** The Metals Society, London 1978, p. 249.
123. F. E. Luborsky, J. J. Becker, P. G. Frischman, L. A. Johnson, *J. Appl. Phys.* **49** (1978) 1769.
124. T. Egami, *Rep. Prog. Phys.* **47** (1984) 160.
125. K. Mohri, E. Sudoh, *IEEE Trans. Magn.* **MAG-17** (1981) 1317.
126. P. Chaudhari, J. J. Cuomo, R. J. Gambino, *IBM J. Res. Dev.* **17** (1973) 66.
127. H. S. Chen, K. A. Jackson in H. Herman (ed.): *Ultrarapid Quenching of Liquid Alloys,* Academic Press, New York 1981, p. 215.

Metallic Soaps

Angelo Nora, FACI S.p.A., Carasco, Italy

Alfred Szczepanek, Faistenlohestr. 5, München, Federal Republic of Germany

Gunther Koenen, Peter Greven Fettchemie GmbH, Bad Münstereifel, Federal Republic of Germany

1. Introduction 1
2. Production 2
2.1. Raw Materials 2
2.2. Industrial Processes 3
2.2.1. Double Decomposition 3
2.2.2. Direct Reaction with Metal Compounds 3
2.2.3. Direct Reaction with Metals 4
2.3. Plant Equipment 4
2.4. Storage 5
3. Fire and Explosion Protection 5
4. Occupational Health and Environmental Protection 5
5. Individual Metallic Soaps 6
5.1. Sodium Soaps 6
5.2. Potassium Soaps 6
5.3. Lithium Soaps 6
5.4. Magnesium Soaps 7
5.5. Calcium Soaps 7
5.6. Barium Soaps 8
5.7. Aluminum Soaps 8
5.8. Lead Soaps 9
5.9. Copper Soaps 10
5.10. Zinc Soaps 11
5.11. Cadmium Soaps 12
5.12. Manganese Soaps 12
5.13. Iron Soaps 13
5.14. Cobalt Soaps 13
5.15. Nickel Soaps 14
6. Quality Specifications 14
7. References 15

1. Introduction

Originally the term metallic soaps was only used for metal salts of fatty acids of naturally occurring animal fats (mainly tallow and lard) and vegetable fats. Today, metallic soaps are understood to refer to the sparingly soluble or insoluble salts of saturated and unsaturated, straight-chain and branched, aliphatic carboxylic acids with 8 – 22 carbon atoms. Examples are saturated fatty acids like stearic acid (octadecanoic acid), lauric acid (dodecanoic acid), 12-hydroxystearic acid, and mixtures of acids with 8 – 22 carbon atoms; unsaturated fatty acids like oleic acid (*cis*-9-octadecenoic acid) and linoleic acid (9,12-octadecadienoic acid) (→ Fatty Acids); synthetic carboxylic acids like isostearic acid, 2-ethylhexanoic acid (→ Carboxylic Acids, Chap. 10.4.), dimethylhexanoic acids, trimethylhexanoic acids; and mixtures of synthetic aliphatic isocarboxylic acids.

The salts of the alicyclic naphthenic acids and resin acids (abietic acid) are also frequently regarded as metallic soaps due to properties related to their manufacture and applications. However, the metal salts of resin acids (resinates) no longer have great significance. Naphthenic acids are a byproduct of oil refining (→ Oil Refining). The byproduct is a mixture of linear and cyclic aliphatic hydrocarbons containing carboxylic acid groups.

The production and properties of soaps of almost all metals are described in the literature. However, only the metallic soaps of the following metals (arranged according to the periodic table) are industrially important: lithium, potassium, sodium (group 1); magnesium, calcium, barium (group 2); zirconium (group 4); manganese (group 7); iron (group 8); cobalt (group 9); nickel (group 10); copper (group 11); zinc, cadmium (group 12); aluminum (group 13); lead (group 14); as well as cerium and mixtures of rare earths.

Uses. Metallic soaps are used as thickeners for oils and other organic media, as lubricants (→ Lubricants and Related Products, Chap. 9.) and release agents, as well as drying agents (dryers). Dryers are metallic soaps that accelerate drying as, for example, in the film formation of paints based on drying oils and alkyd resins → Paints and Coatings). Metallic soaps are used

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_361

as acid scavenger and lubricants both internally and externally in polyolefins and PVC. Sodium and magnesium soaps are important as additives in the pharmaceutical and cosmetics industries, as well as in the building industry. Lithium soaps are especially important as components of lubricating greases.

Properties. The physicochemical properties of metallic soaps are determined by the nature of the metal ion and the organic acid. The metallic soaps of saturated straight-chain carboxylic acids are solids and either have a sharp melting point or melting range, or carbonize on heating without melting. Pastelike products contain water.

Unsaturated fatty acids and branched synthetic acids form metallic soaps with plastic properties. They are therefore usually produced and used in solution.

Types. Metals can form the following metallic soaps depending on their valence and properties:

Neutral soaps:	$RCOOM^{I}$
	$(RCOO)_2M^{II}$
	$(RCOO)_3M^{III}$
Acid soaps:	$(RCOO)_2M^{II} \cdot RCOOH$
	$(RCOO)_3M^{III} \cdot RCOOH$

Acid soaps contain bound free organic acid that can be removed by extraction.

In *basic soaps* $(RCOO)_2M^{II} \cdot (MO)_x$ ($x = 1$, 2), complexed metal oxides can also be present in the form of metal hydroxides or as mixtures of metal oxides and hydroxides.

Mixed soaps $(RCOO)(R'COO)M^{II}$ consist of mixed crystals or mechanically produced mixtures.

Complex soaps: Lead can form complex basic salts: $(PbO)_x \cdot PbSO_4 \cdot [(RCOO)_2Pb]_y$. Other inorganic anions can also be present instead of sulfate.

Overview. A selection of commercially available metallic soaps and those described in the literature is given in Table 1. The soaps are listed in alphabetical order according to the chemical symbol of the metal. The products selected are representative of other homologous soaps. At least one metallic soap of each type is described in the list of references.

2. Production

2.1. Raw Materials

The organic acids used for the production of metallic soaps are either solid or liquid. They are supplied in sacks, drums, containers, or tank cars. The use of products packed in paper sacks should be minimized because of high handling costs and danger of impurities. Acids purchased in drums and containers should be stored at the recommended temperature to avoid total or partial crystallization. For large-scale production, molten acids are generally supplied in tank cars. A tank farm is required, and the tanks should preferably be made of stainless steel to avoid attack on the tank material and darkening and contamination of the product.

Blanketing with an inert gas (e.g., nitrogen or carbon dioxide) may be necessary to prevent darkening caused by oxidation. Acids such as stearic acid, which are solid at room temperature and are supplied as liquids, must be kept above the melting point, preferably by circulation of warm water to avoid overheating at the wall. Stainless steel and aluminum are used as tank materials.

The acids are best dosed in the liquid state; the accuracy of pumps working by volume dosage is sufficient if temperature and density are taken into consideration. Nowadays, mass flow meters are used with better accuracy.

The melting of solid organic acids in the reaction vessel reduces throughput; a separate melting vessel should therefore be used.

The metal salts, which are usually solids, are supplied in sacks, drums, or containers; they are dosed by means of weighing systems. Because the salts may be very toxic, appropriate measures for personal and environmental protection must be taken.

The majority of metallic soaps marketed are stearates. These are salts of stearin, a mixture of mainly $C_{16}-C_{18}$ fatty acids that can be derived both from animal fats (tallow or lard) or from vegetable oils. Recently, the demand for stearates from vegetable sources has increased, mainly for applications in foods, animal feeds, pharmaceuticals, and cosmetics due to the bovine spongiform encephalopathy (BSE) problem, even though stearin from tallow is largely deactivated during its production.

Table 1. Metallic soaps

Element	Stearate	12-Hydroxy-stearate	Laurate	Oleate	Naphthenate	Resinate	2-Ethyl-hexanoate	Reference
Ag	+	+	+	+	+	+		[1], [2]
Al	+		+	+	+	+	+	[3–5]
Ba	+	+	+		+		+	[6]
Be	+		+					[7]
Bi	+			+		+	+	[8]
Ca	+		+		+	+	+	[9]
Cd	+	+	+	+	+		+	[10]
Ce	+				+			[11]
Co	+			+	+	+	+	[12]
Cr	+				+		+	[13]
Cu	+			+	+		+	[14]
Fe	+			+	+		+	[13]
Hg	+		+	+	+			[7]
K	+			+				
Li	+	+						[15]
Mg	+		+	+	+	+	+	[16]
Mn	+			+	+	+	+	[17]
Na	+			+				
Ni	+			+	+	+	+	[13]
Pb	+			+	+	+	+	[18]
Sb				+				[19]
Sn	+			+	+		+	[20]
Sr	+				+			[7]
Th	+							[7]
Ti	+					+		[21]
U	+			+				[7]
V					+			[22]
Zn	+	+	+	+	+	+	+	[23]
Zr	+				+	+	+	[21]

2.2. Industrial Processes

Because of the great number of possible combinations of metal and acid groups, only general production processes can be described here. Three types of reaction are possible:

1) Double decomposition (precipitation)
2) Direct reaction with metal oxides, metal hydroxides, or metal salts of volatile acids
3) Direct reaction with metals

2.2.1. Double Decomposition

In the first step of the double decomposition (precipitation) process the organic acid is saponified with alkali. In the second step, the resulting alkali soap is precipitated by addition of a solution of a water-soluble metal salt:

$$\mathrm{RCOOH} + 2\,\mathrm{NaOH} \longrightarrow 2\,\mathrm{RCOONa} + 2\,\mathrm{H_2O}$$

$$2\,\mathrm{RCOONa} + \mathrm{Ca^{2+}} \longrightarrow (\mathrm{RCOO})_2\mathrm{Ca} + 2\,\mathrm{Na^{+}}$$

The soap solution can also be added to the metal salt solution, or both solutions can be run simultaneously into a separate precipitation vessel. For sparingly soluble alkali soaps, alkaline solutions are used as the reaction medium. Plastic metallic soaps are precipitated in the presence of a water-immiscible solvent which absorbs the metallic soaps formed.

2.2.2. Direct Reaction with Metal Compounds

The reaction of organic acids with metal oxides, metal hydroxides, or salts of volatile acids, such as carbonates, acetates, or formates can be carried out by the following types of process:

1) Melt process
2) Reaction in the aqueous phase
3) Reaction in solvents or lubricants

Melt Process. In the melt process, metal oxides, metal hydroxides, or metal salts of volatile acids are fed into a stirred melt of the organic acid at elevated temperature. The rate of addition should be adapted to the reaction rate to prevent foaming caused by evaporation of water or volatile acid. With unreactive metal compounds, delays can occur in the initial phase. To initiate the reaction, small quantities of water or, for metal oxides, volatile acids such as acetic acid are often added.

The melt process is suitable for metallic soaps with melting points below 140 °C, which produce relatively nonviscous melts. These are generally metallic soaps with sharp melting ranges such as calcium, magnesium, lead, zinc, or cadmium stearate. Alkaline earth stearates are often used in combination with low-melting metallic soaps (e.g., as PVC stabilizers in the combinations barium – cadmium, barium – zinc, or calcium – lead), a low-melting acidic metallic soap is first produced in the melt; the high-melting soap is then formed in the melt of the low-melting one [24].

Reaction in the Aqueous Phase. In the aqueous-phase reaction, metal oxides or hydroxides react in the presence of catalysts such as amines, sparingly soluble alcohols, or other wetting agents. The processes are applicable only to solid metallic soaps (e.g., barium, magnesium, zinc, lead, and calcium stearates) [17].

Reaction in solvents is preferred if the direct melt process is impossible or gives low-quality products, or if the metallic soaps are to be used in solution. Dryers and liquid PVC stabilizers based on branched-chain aliphatic carboxylic acids and naphthenic acids are mainly produced by this process. The solvent (e.g., solvent naphtha or white spirits) also acts as an entrainer for removing reaction water by azeotropic distillation [25].

For solid metallic soaps the solvent can be replaced by organic compounds such as alkanes or fatty alcohols which are solid at room temperature and cannot be saponified by metal oxides. The reaction medium cannot be removed and remains in the final product, which may, for example, be used as a lubricant for plastics [26]. Lubricating greases provide a further example of a reaction product that remains in the reaction medium (→ Lubricants and Related Products, Chap. 9.).

Continuous Process. Metal oxides are dissolved in aqueous ammonia and converted to basic metal carbonates by passage of gaseous carbon dioxide. The molten fatty acids in aqueous ammonia are converted to saponified emulsions. The two components are continuously mixed in the blade region of a rotor – stator mixer with interlocking radial surfaces. The metallic soaps are produced in pastelike to semidry form; they are removed continuously, dried, and ground [27].

Dust-Free Granules. For many uses, metallic soaps do not have to be in powder form. To prevent explosion and promote better handling and dosage, dust-free granules are preferred. The granules are produced from straight-chain fatty acids (C_8–C_{22}) and metal oxides, hydroxides, or carbonates. Granule formation is controlled by the rate of addition of the metal compound, stirring rate, temperature, and water content of the reaction mixture, which is adjusted to ca. 10 wt %. The reactors are mixers equipped for vacuum drying. In the case of unreactive metal oxides (e.g., zinc oxide), chelating acids like citric, tartaric, or glycolic acid are added as catalysts. Because of the small quantity of water involved and the absence of filtration, drying, and pulverization stages, the process runs very economically by using the heat of reaction [28].

2.2.3. Direct Reaction with Metals

Organic acids are treated with metal granules, powders, or gauze at elevated temperature, usually with a solvent as the reaction medium. Direct reaction with metals is used when it offers economic advantages or allows the replacement of unreactive metal oxides by metals [29].

2.3. Plant Equipment

Direct Reaction in Aqueous Media. Reaction vessels, filtration, and drying equipment are necessary for reaction in aqueous media (the precipitation process).

Stainless steel *reaction vessels* are preferred for the production of alkali soaps. The acids

are saponified at 60 – 90 °C at a dilution of 1 : 5 – 1 : 10 depending on the type.

Metal compounds are dissolved at 40 – 50 °C, depending on their solubility, in stainless steel containers. Tanks for intermediate storage of reaction mixtures are equipped with slow stirrers or recirculation pumps.

In the *filtration* of metallic soaps, the filter cake should be as dry as possible for economic reasons. Nutsche (pan) filters, drum filter presses, and plate filter presses have now been largely replaced by membrane filter presses or continually functioning automatic filters. In the direct reaction in the aqueous phase, the filter water produced is used as circulation water and must be put into intermediate storage.

Belt, drum, and rotary dryers have proved successful as *drying units.* In drying units without integrated pulverization and classifying facilities, pulverization units are necessary.

Melt Process. For direct reaction by the melt process or in lubricants, stainless steel or glass-lined reactors fitted with stirrers, heaters (up to 160 °C), and vapor extractors are used. The melts are then processed into a solid, dust-free form by using flaking equipment, cooling belts, or spraying devices. If desired they can be ground.

Precipitation or Direct Reaction in Solvents. For precipitation in the presence of solvents or direct reaction in solvents, stainless steel or glass-lined steel reactors are used that are fitted with stirrers, heaters (up to 160 °C), vacuum connections, and reflux condensers with water separators. Suitable filter equipment is necessary for filtration of viscous solutions. Reaction with metals is also possible in this type of plant. Finally, the product is ground in various mills to obtain different particle size distributions.

2.4. Storage

The liquid to pastelike, solvent-containing metallic soaps are generally stored in packaging forms ready for sale (e.g., cans, drums, or containers). Bags or big bags are often used for powdered metallic soaps. For storage in large silos, protective measures against fire and explosion are necessary (see Chap. 3).

3. Fire and Explosion Protection

Fire Protection. Metallic soaps are flammable substances that burn with intense smoke production. The resulting oxides are carried into the air and may be toxic (heavy metals). This should be taken into consideration during fire fighting. Foam and carbon dioxide extinguishers are suitable fire-fighting agents.

For metallic soap solutions in organic media, local regulations for the handling and storage of flammable liquids must be observed. Plastic metallic soaps are not readily flammable because of their small surface area. Powdered metal soaps have a relatively large surface area and are flammable. With metallic soaps based on unsaturated fatty acids such as oleic and linoleic acids, self-ignition can occur if the packed products are not cooled sufficiently. Self-ignition is promoted by soaps of heavy metals, which are oxidation catalysts.

Explosion Protection. When organic solvents are used in the production of metallic soaps, plants must be protected against explosion by suitable equipment in compliance with prevailing regulations. Metallic soaps in powder form can, like other organic dusts, cause dust explosions, which are most likely to occur during drying, pulverizing, and storage. To prevent buildup of static charge (ignition source), good grounding (earthing) should be ensured. Formation of sparks by metal particles is difficult to avoid. Magnetic separators do not offer any safety improvements because the metals are either nonmagnetic or coated by the metallic soaps. To avoid serious damage, relevant parts of the plant must be equipped with pressure-release devices. If this precaution is not taken, a pressure wave can lead to severe secondary dust explosions and destruction of the plant.

4. Occupational Health and Environmental Protection

Occupational Health. Soaps of calcium, sodium, magnesium, potassium and zinc are nontoxic substances and do not require particular protective measures. Soaps of heavy metals are toxic substances that can lead to acute

or chronic poisoning. Suitable measures to protect the worker are dust extraction, protective clothing, dust masks (when dust sources cannot be avoided), and regular medical checkups. The MAK and TLV values are given for individual metals (Chap. 5).

Wastewater. The filter water produced in the direct reaction in the aqueous phase is collected and reused as reaction water; complete circulation eliminates wastewater production. In the double decomposition reaction, considerable quantities of neutral salts such as sodium sulfate, sodium chloride, and sodium nitrate are produced, which are difficult or impossible to remove by water treatment. This fact should be given particular consideration in the choice of plant location. Other sources of water pollution are bleeding of the filter presses and excess metal salts used for precipitation. Before wastewater is sent to the water treatment plant, the heavy metals are therefore precipitated by addition of lime and/or passing through CO_2 to give a pH of 8.5 – 9.

Air Pollution. Solvent emissions must be reduced in accordance with local regulations by means of a scrubber or absorption tower. Steam- or water-jet vacuum pumps are often used and entrain small quantities of mostly water-immiscible solvents; oil separators are therefore necessary.

The dryers for powdered metallic soaps may release residual dust-laden or water-vapor-containing air into the surroundings. To purify the air, filtration units with large surface areas are necessary. Automatic dust-measuring instruments built into the exhaust stack switch off the plant if the dust concentration becomes too high (e.g., because of a tear in the filter). The residual emission is discharged into the surroundings via a high exhaust stack. Permitted emission values are stipulated in national air purity regulations.

5. Individual Metallic Soaps

5.1. Sodium Soaps

Production. The sodium soaps of stearic and oleic acid have the most industrial importance. They are produced by direct reaction. A solution of sodium hydroxide is added to the fatty acid, and when the reaction is complete, the product is dried and grounded.

Properties. See Table 2.

Uses. Sodium stearate is used in various fields, but mainly in the pharmaceutical and cosmetics industries as an emulsifying and gelling agent. Sodium oleate is largely used in the building industry as an initial dispersing agent that becomes a waterproofing agent after application.

5.2. Potassium Soaps

Producton. The potassium soaps are produced by direct reaction. A solution of potassium hydroxide is added into the fatty acid, and after completion of the reaction the product is dried and ground.

Properties. See Table 2.

Uses. Potassium soaps, mainly stearate and oleate, are used for their swelling properties. They are used to thicken natural and synthetic oils and to raise the melting point and flexibility of waxes and paraffins.

5.3. Lithium Soaps

Production. The lithium soaps of stearic and 12-hydroxystearic acid have the most industrial importance. They are produced by direct reaction in aqueous medium. A very dilute solution of lithium hydroxide hydrate, $LiOH \cdot H_2O$ (concentration ca. 1 : 40, i.e., 2.5 wt %), is slowly added with intense stirring to the fatty acid dispersed in water in the ratio 1 : 20 at 90 °C. Lithium soaps are difficult to filter; flocculation is therefore improved by using an excess of lithium hydroxide or by adding neutral salts. The resulting lithium soap dispersion is, however, often not filtered but spray-dried, even though this involves evaporation of a large quantity of water. Under modified reaction conditions, lithium hydroxide hydrate can be replaced by lithium carbonate and ammonia.

Table 2. Technical data for some commercial soaps of sodium, potassium, lithium, and magnesium

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash or metal content, wt %	Moisture∗, wt %	Bulk density∗, g/L	Melting range, °C
Lithium stearate	[*4485-12-5*]	white powder	1	2.4 Li	0.5	230	190 – 210
Lithium 12-hydroxy-stearate	[*7620-77-1*]	white powder	1	2.3 Li	0.5	400	> 200
Magnesium stearate, technical	[*557-04-0*]	white powder	2	7 – 8	4 – 5	250	115 – 135
Magnesium behenate	[*43168-33-8*]	white powder	2	6 – 7	6.0	250	ca. 140
Sodium oleate	[*143-19-1*]	yellowish powder	0.1	22 – 24	4	350	200 – 210
Sodium stearate	[*822-16-2*]	white powder	1	22 – 24	2	250	210 – 220
Potassium stearate	[*593-29-3*]	white powder	2	12 K	1.5	300	215 min.

∗ Approximate values.

Properties. See Table 2.

Uses. Lithium soaps have particularly good swelling properties. Lithium stearate is used to thicken natural and synthetic oils; to raise the melting point and flexibility of microcrystalline waxes and paraffins; as an additive for coating waxes to increase water repellence; and as a lubricant in the production of injection-molded articles from light metal. Multipurpose lubricating greases have particularly good penetration properties, high oxidation stability, and dropping points up to ca. 200 °C. Lithium 12-hydroxystearate is used primarily for the production of lubricating greases based on synthetic oils (e.g., from silicon and ester oils). These lubricants cannot be produced by the usual saponification in mineral oil because the synthetic oils are attacked under the saponification conditions employed.

5.4. Magnesium Soaps

Magnesium soaps are produced as neutral salts by the double decomposition process, using water-soluble magnesium salts (primarily magnesium chloride) and alkali salts of organic acids in an aqueous medium, or by direct reaction with magnesium oxide. Magnesium soaps in powder form based on stearic acid are particularly important. The degree of purity required, especially with regard to heavy-metal impurities, determines the choice of raw materials in pharmaceutical products. Pharmaceutical quality must fulfill the requirements of the European, Japanese, and U.S. pharmacopoeias.

Properties. See Table 2. Magnesium soaps of long-chain fatty acids have a high specific surface area and can absorb considerable quantities of liquid, giving a creamy consistency. They are insoluble in alcohols, esters, ketones, ethers, and hydrocarbons.

Uses. Magnesium soaps are used in the plastics industry as lubricants and mold-release agents for thermosetting plastics and thermoplastics; in the pharmaceutical industry as processing aids in the production of dragées and as lubricants for tablet pressing; in the cosmetics industry as powder components and for consistency in water-free ointments; in the wax industry to increase the retention of creams based on semisolid wax products and in the production of lubricating waxes; and as a waterproofing agent for the building industry.

5.5. Calcium Soaps

Calcium soaps are generally neutral salts. They are obtained by double decomposition or direct reaction. The raw materials consist of saturated straight-chain aliphatic acids ($C_8 - C_{22}$), montan wax acids, unsaturated straight-chain carboxylic acids (e.g., oleic and linoleic acids), branched-chain aliphatic carboxylic acids (e.g., ethylhexanoic acids), naphthenic acids, and resin acids. Calcium chloride and calcium hydroxide are used as the metal compounds.

Very finely divided calcium soaps of the aliphatic straight-chain carboxylic acids are produced by double decomposition in aqueous

medium. Stearates, myristates, and laurates are commercial products, but the most important calcium soaps are stearates.

Direct reaction 2.2.2 (Melt process) is the most common production method for solid calcium soaps.

In both cases the products are free of water-soluble salts; their purity depends on the quality of calcium hydroxide used.

Calcium soaps are also produced as neutral soaps by the double decomposition process from water-soluble calcium salts and sodium soaps in an aqueous medium. This process can be used to obtain a very high specific surface area, but is mainly used to produce highly crystalline calcium soaps.

Aqueous calcium stearate pastes with a solid content of 40 – 50 % are also produced by the direct process and are used as additives for paper coating (slips).

Calcium soaps which are not in powder form and are therefore used mostly as solutions in organic media (e.g., 2-ethylhexanoates and naphthenates) are produced by direct reaction in an appropriate medium.

Properties. See Table 3.

Uses. Calcium stearate is the most important calcium soap. Calcium stearate is widely used in polyolefins (PE, PP) and in PVC as a lubricant (internal or external), release agent, and acid scavenger. Calcium soaps are used as physiologically inert lubricants and secondary stabilizers in the plastics industry to improve the flow properties and prevent the caking of hygroscopic substances; as lubricants and mold-release agents along with magnesium soaps in pharmaceutical tablet pressing; as waterproofing additives in agents used for building protection and for surface treatment of fillers; and in the paper industry in recording paper, diagram paper, and coating slips.

5.6. Barium Soaps

The most important organic acids in barium soaps are 12-hydroxystearic, lauric, 2-ethylhexanoic, and naphthenic acids. The metal compounds used are barium chloride and barium hydroxide (mono- or octahydrate). The TLV – TWA value for barium is 0.5 mg/m^3.

Properties. See Table 3.

Uses. Barium soaps are used in the paint industry for wetting pigments and as dispersion agents; in metal forming as a component of wire-drawing and other auxiliaries; and in the plastics industry as lubricants and costabilizers.

5.7. Aluminum Soaps

Commercial aluminum soaps are generally produced by treating alkali soaps with water-soluble aluminum salts. In precipitation in aqueous systems, agglomerates of aluminum hydroxide and carboxylic acids are first produced which then react with each other to form salts during the subsequent drying process.

Because aluminum is trivalent, three types of soap are possible:

1) Aluminum monosoaps $Al(OH)_2(OOC{-}R)$
2) Aluminum disoaps $Al(OH)\,(OOCR)_2$
3) Aluminum trisoaps $Al(OOCR)_3$

The polymeric structure of aluminum soaps is not conveyed by the formulas [30].

All types of organic acids are used for aluminum soaps, but particularly stearic, oleic, naphthenic, and 2-ethylhexanoic acids. Aluminum sulfate is preferred as the water-soluble aluminum salt, along with potassium aluminum sulfate and aluminum chloride. In industrial production of the three types of aluminum soap, the organic acids are saponified with 1, 2, or 3 mol of alkali, respectively. Because the three types differ greatly in their thickening properties, commercial products are available with intermediate metal contents. The metal, ash, and free fatty acid contents are given to characterize the soaps.

To achieve special thickening properties, above all for the production of lubricating greases based on mineral oil, fatty acids are combined with aromatic acids (e.g., benzoic or phthalic acid) or with dimeric fatty acids. Lubricating grease with high fastness to fulling resistance and a high dropping point is obtained in this way.

Aluminum soaps in powder form are obtained by double decomposition in aqueous medium.

Table 3. Technical data for some commercial soaps of calcium and barium

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash, wt %	Moisture∗, wt %	Bulk density, g/L	Melting range, °C
Calcium stearate, precipitated	[*1592-23-0*]	white powder	1 – 2	9 – 10	2 – 3	120 – 130	120 – 160
Calcium stearate, direct reaction	[*1592-23-0*]	white powder	0.5	9 – 10	2	200	120 – 160
Calcium arachidate	[*22302-43-8*]	white powder	1	8 – 9	1.5	200 (tapped)	ca. 150
Calcium 12-hydroxy-stearate	[*3159-62-4*]	white powder	1	8.5 – 9.5	2 – 3	250 (tapped)	ca. 140
Calcium laurate	[*4696-56-4*]	white powder	1	13 – 14	2 – 3	300 (tapped)	ca. 140
Calcium oleate	[*142-17-6*]	yellow mass	2	10 – 11	1		ca. 120
Calcium 2-ethyl-hexanoate	[*136-51-6*]	yellowish powder	2	16 – 17	1		ca. 160
Calcium ricinoleate	[*6865-33-4*]	yellowish powder	2	10 – 11	1.5		ca. 130
Calcium naphthenate	[*61789-36-4*]	solid brown mass		11 – 12	1		
Barium stearate	[*6865-35-6*]	white powder	1	27 – 28	0.5	250	200
Barium laurate	[*4696-57-5*]	white powder	1	35 – 36	0.5	275	200
Barium 2-ethyl-hexanoate	[*2457-01-4*]	yellowish paste		29 – 30	0.6		
Barium naphthenate	[*61789-67-1*]	dark paste		24 – 25	2		

∗ Approximate values.

Catalyzed reactions between fatty acids and freshly precipitated aluminum oxide hydrate are known but not used industrially. Aluminum oxide hydrate ages too quickly and is not suitable for melting or direct reaction processes in nonaqueous media because of its low reactivity. An exception is the in situ production of lubricating greases in which aluminum alkoxides and freshly precipitated aluminum hydroxide hydrate are used. The TLV – TWA value for aluminum is 10 mg/m^3 for metal dust and 5 mg/m^3 for pyro powders and welding fumes.

Properties. See Table 4. Aluminum soaps of long-chain fatty acids are insoluble in water, lower alcohols, esters, and ketones; they are soluble in hydrocarbons, toluene, xylene, ethylene glycol, chlorinated hydrocarbons, and vegetable and mineral oils. The most striking property of aluminum soaps is their capacity for thickening organic media and for forming solid gels with them. Gel consistency increases in the order trisoaps, disoaps, and monosoaps. The gels are destroyed by polar substances such as primary amines, ammonia, pyridine, and lower alcohols.

Uses. The gel-forming property of aluminum soaps is exploited in the mineral oil, cosmetics, and pharmaceutical industries. They are also employed in the paint industry as antisettling and matting agents and for wetting pigments. Aluminum soaps in powder form are used as lubricants in thermosetting plastics and thermoplastics, as waterproofing agents in building protection, and as flow promoters for caking powders.

5.8. Lead Soaps

Neutral lead soaps are produced by double decomposition, by direct reaction in the melt, or by direct reaction in aqueous media in the presence of catalysts such as triethanolamine and sparingly soluble alcohols [17].

Basic lead soaps are produced by direct reaction in the presence of catalysts in aqueous media. The addition salts contain a maximum of two molecules of lead(II) oxide to one molecule of neutral lead soap:

$$3\,PbO + 2\,HOOCR \rightarrow 2\,PbO \cdot Pb(OOCR)_2 + H_2O$$

Direct reaction in the melt does not go beyond the addition of one molecule of lead(II) oxide because of an increase in the melting point. In the presence of a fusible organic medium

Table 4. Technical data for some commercial aluminum soaps

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash, wt %	Moisture∗, wt %	Tapped bulk density, g/L	Melting range, °C
Aluminum tristearate	[*637-12-7*]	white powder	16 – 22	7 – 8	1.2	300	110 – 130
Aluminum distearate	[*300-92-5*]	fine white powder	3 – 7	10 – 11	1.2	82	150 – 160
Aluminum monostearate	[*7047-84-9*]	white powder	1 – 3	13 – 14	1.2	170	180 – 250
Aluminum di-12-oxystearate	[*7047-84-9*]	fine white powder	4 – 6	7.5 – 8.5	1.5	110	ca. 140
Aluminum dipalmitate	[*14236-50-1*]	fine white powder	5 – 7	10.5 – 11.5	1	300	ca. 160
Aluminum monopalmitate	[*555-35-1*]	very fine white powder	2 – 4	15.5 – 16.5	1.5	280	ca. 200
Aluminum dimyristate	[*56639-51-1*]	fine white powder	6 – 9	10.5 – 11.5	1	300	ca. 145
Aluminum dilaurate	[*817-83-4*]	white powder	4 – 6	10.5 – 11.5	1	280	ca. 150

∗ Approximate values.

(e.g., $C_{16}-C_{18}$ fatty alcohols, or a combination of fatty alcohols and paraffin or long-chain fatty acid esters), however, a further molecule of lead(II) oxide can be added to give dispersions of dibasic lead soaps [24]. The basicity of basic lead soaps in the direct reaction in solvents depends on their solubility. In general, only one molecule of lead(II) oxide can be added.

Basic Lead Complex Soaps. The starting material for basic lead complex soaps is an aqueous dispersion of a basic inorganic lead salt which reacts with fatty acids in the presence of catalysts in an aqueous medium [31]:

$$4\,PbO \cdot PbSO_4 + 2\,HOOCR \rightarrow$$
$$3\,PbO \cdot PbSO_4 \cdot Pb(OOCR)_2 + H_2O$$

Production. Technical-grade stearic, oleic, 2-ethylhexanoic, and naphthenic acids are used for the production of the most important lead soaps. The principal metal source is lead oxide. Lead acetate is used as the water-soluble lead salt in the double decomposition process. Direct reaction in aqueous medium is performed with reactive lead(II) oxide (litharge, commercial specification "canary yellow"). The red lead oxide content should be very low because this modification cannot form basic addition salts. The red modification is, however, suitable for the production of neutral salts in the melt or in an aqueous medium. In the melt process, the lead metal content of the lead oxide should be very low to avoid blockage of valves by agglomeration.

Regulations for personal and environmental protection must be observed, particularly with lead soaps in powder form. Even the absorption of small quantities of lead can lead to chronic illness (→ Lead, Chap. 9.; → Lead, Chap. 10.). The MAK value for lead is 0.1 mg/m^3 and the TLV – TWA value is 0.15 mg/m^3.

Properties. See Table 5.

Uses. Neutral and basic lead stearates are used as heat stabilizers and lubricants in PVC processing; as lubricants in the pencil industry; and as additives for specialty papers. Lead 2-ethylhexanoate, naphthenates, and resinates are used in the paint industry for wetting pigments and as dryers, and in the mineral oil industry as components of high-pressure lubricants.

5.9. Copper Soaps

Copper forms neutral soaps which are produced either by double decomposition or by the direct process in the melt or in organic media. Production from metallic copper is also known: copper powder is dispersed in solvent naphtha; copper soaps that are soluble in solvent naphtha are formed by passage of air in the presence of an organic acid.

Table 5. Technical data for some commercial soaps of lead and copper

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash or metal content, wt %	Moisture∗, wt %	Bulk density, g/L	Melting range, °C
Neutral lead stearate	[*1072-35-1*]	white powder	1	27 – 28 Pb	0.2	500	105 – 110
Monobasic lead stearate	[*90459-52-2*]	white powder	1	41 – 42 Pb	1	250 – 300	110 – 120
Dibasic lead stearate	[*56189-09-4*]	white powder	1	51 – 52 Pb	0.5	450	> 200 (decomp.)
Lead 2-ethylhexanoate	[*16996-40-0*]	white, plastic	1	ca. 45	0.5		
Lead oleate	[*1120-46-3*]	brown mass	1	30 – 31	1		ca. 65
Lead naphthenate	[*61790-14-5*]	dark, viscous mass	1	30 – 31	1		
Copper stearate	[*7617-31-4*]	blue-green powder	1	14 – 15	1		ca. 155
Copper laurate	[*19179-44-3*]	blue-green powder	2	17 – 18	1		ca. 125
Copper 2-ethylhexanoate	[*22221-10-9*]	blue mass		23 – 24	1		
Copper oleate	[*10402-16-1*]	blue-green mass		13 – 14	1		
Copper naphthenate	[*1338-02-9*]	gray-green paste		14 – 15	1		

∗ Approximate values.

Commercial copper soaps are produced from stearic, naphthenic, oleic, and 2-ethylhexanoic acids. Water-soluble copper salts, mainly copper sulfate, are used as copper source for the double decomposition; copper hydroxide or basic copper carbonate is used for direct reaction.

Production of copper soaps in powder form by double decomposition requires a separate plant, because cleaning the drying and filtration units is almost impossible. The MAK values for copper are 0.1 mg/m^3 (fume) and 1 mg/m^3 (dust); the TLV – TWA values are 0.2 mg/m^3 (fume) and 1 mg/m^3 (dusts and mists).

Properties. See Table 5.

Uses. Copper soaps are used mainly as fungicides in wood and ship paints, textiles (e.g., tents, awnings, sandbags, camouflage netting, ropes), paper, and cardboard. Compared with inorganic copper compounds, they have the advantage of being difficult to wash out. Naphthenates, oleates, and 2-ethylhexanoate are used as solutions in organic media.

5.10. Zinc Soaps

The most important zinc soaps are the zinc stearates.

Neutral zinc soaps are produced by double decomposition, direct reaction in the melt, direct reaction in aqueous or organic media, and reaction of zinc oxide with fatty acids below the melting point of the soaps.

Basic zinc soaps produced by double decomposition are obtained in powder form; the basic moiety can be in the form of a basic soap, zinc oxide, or zinc carbonate. In organic media, soluble zinc soaps (particularly naphthenates, 2-ethylhexanoate, and oleates) can be rendered basic during the direct reaction by addition of excess zinc hydroxide. Soaps with 1.7 equivalents of zinc hydroxide to 1 equivalent of acid can be obtained.

Commercial zinc soaps are produced from stearic, lauric, oleic, 2-ethylhexanoic, and naphthenic acids. Those derived from stearic acid are produced on the largest scale. Water-soluble salts, preferably zinc sulfate and chloride, are used as zinc source for double decomposition; zinc oxide, hydroxide, carbonate, and acetate are employed for direct reaction. When several types of product are prepared in a single plant, attention must be paid to the production order because zinc stearate acts as a gel breaker with gels from aluminum soaps. Normally calcium, magnesium, zinc, and aluminum soaps are produced in a single plant.

Properties. See Table 6.

Uses. Zinc soaps are used for the following purposes:

1) In the plastics industry as lubricants, release agents, and components of PVC stabilizers

Table 6. Technical data for some commercial soaps of cadmium and zinc

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash, wt %	Moisture∗, wt %	Bulk density, g/L	Melting range, °C
Cadmium stearate	[*2223-93-0*]	white powder	2	19 – 20	1	180	105 – 115
Cadmium 12-hydroxy-stearate	[*69121-20-6*]	white powder	2	18 – 19	1	220	110 – 118
Cadmium laurate	[*2605-44-9*]	white powder	2	26 – 27	1	300	
Cadmium oleate	[*10468-30-1*]	brown mass		19 – 20	2		
Cadmium 2-ethyl-hexanoate	[*2420-98-6*]	yellowish paste		32 – 33	2		
Zinc stearate, precipitated	[*557-05-1*]	white powder	0.5	13 – 14	1	140	118 – 122
Basic zinc stearate, direct reaction	[*557-05-1*]	white powder	0.5	15 – 17	1	300	118 – 122
Zinc stearate, direct reaction	[*557-05-1*]	white powder	0.5	13 – 14	1	300	118 – 122
Zinc 12-hydroxystearate	[*35674-68-1*]	white powder	1	12.5 – 13.5	1	220	145 – 155
Zinc laurate	[*2452-01-9*]	white powder	1	17 – 18	1	220	122 – 126
Zinc arachidate	[*22302-43-8*]	white powder	1.5	12 – 13	1	180	120 – 124
Zinc oleate	[*557-07-3*]	bright mass		13 – 14	0.3		
Zinc 2-ethylhexanoate	[*136-53-8*]	clear, highly viscous liquid		23 – 24	1		
Zinc naphthenate	[*12001-85-3*]	brown mass		14 – 15	0.4		

∗ Approximate values.

2) In the rubber industry, as release agents for unvulcanized products (wettable zinc stearate soaps and zinc stearate dispersions in water are also used)
3) In paints, as matting agents and abrasives
4) In building protection, as waterproofing agents
5) For textiles, also as waterproofing agents
6) In the cosmetics and pharmaceutical industries, as additives to body and face powders
7) As components of wire-drawing greases (special zinc arachidate is used in tube drawing because of its low ash content)
8) In paints, as components of dryers and as anticaking agents for powdered products

5.11. Cadmium Soaps

Neutral and basic cadmium soaps are produced by double decomposition, direct reaction in the melt, or direct reaction in organic media. Stearic, lauric, 12-hydroxystearic, oleic, and 2-ethylhexanoic acids are commonly used. Cadmium sulfate, chloride, and nitrate are used in double decomposition; cadmium oxide and hydroxide in the direct reaction. Cadmium soaps are highly toxic. Cadmium accumulates in organs such as the liver and adrenal gland and is released very slowly. Absorption of very small quantities of cadmium over a long period therefore leads to chronic illness. The TLV – TWA value is 0.05 mg/m^3.

Properties. See Table 6.

Uses. The use of cadmium soaps as components of PVC stabilizers is decreasing because of their toxic properties. Solid PVC stabilizers contain cadmium soaps of saturated fatty acids, whereas liquid stabilizers contain soaps of unsaturated or branched, short-chain fatty acids. The soaps act as heat and light stabilizers, and soaps of saturated fatty acids are used as lubricants in the processing of PVC. Unlike other metallic soaps used in the PVC sector, cadmium soaps do not affect transparency and are suitable for hard, transparent PVC products.

5.12. Manganese Soaps

Neutral manganese(II) soaps are produced by double decomposition from alkali soaps and water-soluble manganese salts in an aqueous medium or in the presence of organic solvents, and by direct reaction in the melt or in organic media with manganese oxides, hydroxides, or salts of volatile acids. The oxides and hydroxides are very unreactive and must be freshly prepared; conversion is generally incomplete. In

the reaction with metallic manganese in powder form or as divided gauze, lower carboxylic acids (e.g., acetic acid) are added to the solvent as catalysts.

Basic manganese(II) soaps are produced by double decomposition in an aqueous medium; oversaponified alkali soaps (i.e., more than 1 mol of alkali is used per mole of acid) react with water-soluble manganese salts. The excess alkali forms manganese hydroxide or manganese oxide hydrate, which adds to the neutral manganese soap. Whether a true addition salt or a simple mixture is formed is unclear. The basic manganese soaps are insoluble or sparingly soluble in organic solvents. Stearic, oleic, naphthenic, and 2-ethylhexanoic acids are commonly used. Sources of manganese for the double decomposition are water-soluble manganese salts such as manganese chloride and sulfate; for the direct reaction, manganese oxide, hydroxide, and acetate.

Manganese soaps are hazardous to health. A particular danger arises from dust formation by products in powder form. Because of their brown color, separate production plants are necessary for the powdered products. The TLV – TWA values for manganese are 5 mg/m^3 (dust) and 1 mg/m^3 (fume).

Properties. See Table 7.

Uses. Manganese soaps are used mainly as dryers in paints, printing inks, and varnishes based on drying oils or alkyd resins. Basic powdered products are also added to printing inks.

Manganese soaps can be used as catalysts (e.g., in paraffin oxidation or the reduction of fatty acids to alcohols).

5.13. Iron Soaps

Only neutral iron(III) soaps are produced industrially. Formation of basic soaps by addition of iron oxide or hydroxide is known but is not exploited commercially.

Iron soaps are obtained by double decomposition from alkali soaps and water-soluble iron salts in an aqueous medium with or without organic solvents. Direct reaction in the melt with iron oxides does not go to completion and is very slow; the unreacted free fatty acid is extracted with alcohol.

Raw materials used are stearic, oleic, naphthenic, and 2-ethylhexanoic acids. In double decomposition, iron(III) chloride is generally used as the iron source; in the direct reaction, iron oxide, hydroxide, or carbonate are used. Because of the intense color of iron soaps, separate precipitation units are necessary. Although iron soaps do not pose a danger to health, local regulations regarding personal and environmental protection for dust-producing products must be observed during handling.

Properties. See Table 7.

Uses. Iron soaps are components of dryer mixtures. Because of their intense color, they are used only with dark colors. The function of the iron as a dryer is based on polymerization at elevated temperature rather than catalysis of oxidation.

Further uses are as hydrogenation catalysts, as dry-film lubricants in the form of monomolecular adhesive layers, in the production of fast-copy paper, in waterproofing, and in pharmaceuticals as oil-soluble iron compounds for the treatment of anemia.

5.14. Cobalt Soaps

Neutral and weakly basic cobalt(II) soaps are known. They are produced by double decomposition using alkali soaps with water-soluble cobalt salts in an aqueous medium with or without organic solvents, or by direct reaction via the melt process, or in solvents. Reaction of metallic cobalt in powder or gauze form in organic media is also possible; air acts as an oxidizing agent.

Stearic, oleic, naphthenic, and 2-ethylhexanoic acids are used. For double decomposition, cobalt(II) sulfate is used as the water-soluble salt; for the direct reaction, cobalt acetate, cobalt(II) hydroxide, and cobalt(II) oxide are used.

Cobalt soaps have an intense blue-violet color. Soaps in powdered form must be produced in separate plants. Dust from cobalt soaps is hazardous to health. Local regulations concerning personal and environmental protection and the compulsory labeling of packaging must

Table 7. Technical data for some commercial soaps of iron, cobalt, nickel, and manganese

Compound	CAS registry no.	Appearance	Free fatty acid∗, wt %	Total ash or metal content, wt %	Moisture∗, wt %	Melting range, °C
Iron(III) stearate	[*555-36-2*]	red-yellow powder	5	14 – 15	1	ca. 90
Iron oleate	[*23335-74-2*]	red paste		9 – 10	2	
Iron 2-ethylhexanoate	[*6535-20-2*]	reddish paste		16 – 17	1	
Cobalt stearate	[*13586-84-0*]	blue – violet powder	1	13 – 14	1	ca. 150
Cobalt 2-ethylhexanoate	[*136-52-7*]	blue – violet mass		22 – 23	1	
Cobalt naphthenate	[*61789-51-3*]	dark blue mass		10 – 11 Co		
Nickel stearate	[*2223-95-2*]	bright green powder	1	14.5 – 15.5	1	80 – 86
Nickel oleate	[*13001-15-5*]	greenish paste	1	22 – 23	1	
Nickel 2-ethylhexanoate	[*4995-91-9*]	green paste		22 – 23	1	
Nickel naphthenate	[*61788-71-4*]	deep green mass	1	8.5 – 11.5 Ni	1	
Manganese stearate	[*3353-05-7*]	bright brown powder		13 – 14	1	ca. 80
Manganese(II) oleate	[*23250-73-9*]	brown mass		12 – 13	1	
Manganese 2-ethylhexanoate	[*15956-58-8*]	brown mass		ca. 16 Mn		
Manganese naphthenate	[*1336-93-2*]	dark brown mass		ca. 10 Mn	1	

∗ Approximate values.

be observed. The TLV – TWA value for cobalt is 0.05 mg/m^3.

Properties. See Table 7.

Uses. Cobalt soaps are very effective dryers for printing and writing inks because they catalyze oxidation. Cobalt stearate is used as an adhesive for rubber – metal bonding.

5.15. Nickel Soaps

Divalent nickel forms neutral to weakly basic soaps. They are produced by double decomposition from alkali soaps and water-soluble nickel salts in an aqueous medium in the presence of organic solvents or by direct reaction in the melt or in solvents with nickel oxide, hydroxide, or salts of volatile acids. In the melt process, decomposition occurs above 200 °C with formation of a black powder. The usual acids are stearic, oleic, 2-ethylhexanoic, and naphthenic acids. Nickel sulfate and chloride are used as water-soluble salts for double decomposition, and nickel(II) hydroxide, oxide, acetate, or carbonate for the melt process.

Nickel soaps in powder form must be produced in separate plants because of their intense green color. Dust from nickel soaps is hazardous to health. Local regulations concerning personal and environmental protection and compulsory labeling must be observed. In processing nickel oxide and carbonate, protective measures must be taken because of their carcinogenic properties. The TLV – TWA value for nickel is 1 mg/m^3.

Properties. See Table 7.

Uses. Nickel soaps are used as oxidation catalysts in dryers; as oil-soluble hydrogenation catalysts, which liberate finely divided nickel on heating or decomposition; as detergents in petroleum ether; and as additives in lubricating oils for preventing cyclization and resinification.

6. Quality Specifications

Metallic soaps can be characterized by a series of data. Certain application-orientated properties, such as solubility and swelling, may have

to be determined in special tests on test formulations.

Solid Metallic Soaps. The following parameters are determined:

Moisture Content. The sample is weighed before and after heating at 110 °C for 2 h in a drying oven.

Metal Content. Metal content is determined by one of the usual analytical methods after acid decomposition of the metallic soap in an aqueous extract.

Total Ash. The substance is preheated carefully in an incinerator and then heated slowly to 800 – 1000 °C in a muffle furnace as soon as all the organic substance has decomposed. After being cooled in a desiccator, the ash is determined by back-weighing. For lead soaps, attention must be paid to reduction to metallic lead.

Washed Ash. Residue from the determination of total ash is boiled in distilled water and filtered; the residue and the filter paper are incinerated. The final residue is expressed as a percentage of the weight of substance used for the total ash determination. This method is not applicable to water-soluble metal oxides.

Water-Soluble Salts. The difference between total ash and washed ash is described as soluble ash or water-soluble salt. For metallic soaps produced by double decomposition, this parameter gives a good indication of whether the salts have been washed sufficiently.

Free Fatty Acids. The substance is refluxed with acetone for 10 min and filtered; the fatty acid content is determined by titration with alcoholic potassium hydroxide using thymol blue and cresol red as a mixed indicator.

Particle Size. Particle size is determined as residue on a 71 µm air-stream sieve or by wet sieving of an aqueous dispersion with a surfactant.

Tapped Bulk Density. Tapped bulk density is determined in a commercial tamp volumeter by using a fixed number of impacts.

Bulk Density. The substance is passed through a 500-µm sieve into a measuring cylinder up to the 100-mL mark and weighed.

Melting Range. Melting range can be determined in a commercial melting point apparatus.

pH value. pH determination on the soap dispersion.

Fatty Acids Composition. Fatty acids composition is determined by gas chromatography with direct injection of a methyl esters solution obtained from the soap.

Metallic Soap Solutions. For metallic soap solutions (e.g., used as PVC stabilizers or dryers), metal content, density, viscosity, solid content, and Gardner color index must be determined. Industrial supply conditions are to be observed for metallic soap dryers [32].

7. References

1. I. W. H. Oldham, A. R. Ubbelohde, *J. Chem. Soc.* 1941, 368 – 375.
2. P. N. Cheremisinoff, *J. Am. Oil. Chem. Soc.* **28** (1951) 278 – 279.
3. Colgate-Palmolive-Peet Co., US 2 447 064, 1948.
4. G. H. Smith et al., *J. Amer. Chem. Soc.* **70** (1948) 1053 – 1054.
5. The Commonwealth Eng. Co., US 2 768 996, 1954.
6. C. I. Boner, *Ind. Eng. Chem.* **29** (1937) 58.
7. A. S. C. Laurence, *Trans. Faraday Soc.* **34** (1938) 665 –669.
8. E. A. Nikitina et al., *J. Gen. Chem.* **19** (1949) 1108 –1114.
9. G. H. Smith, S. Ross, *Oil Soap (Chicago)* **23** (1946) 77 – 80.
10. H. I. Braun, *Chem. Ztg.* **53** (1929) 913 – 914.
11. C. W. Hill, C. W. Stoddart, *Am. Soc.* **33** (1911) 1076.
12. I. Marwendel OHG, DE 845 800, 1950.
13. P. N. Cheremisinoff, *Am. Paint. J.* **37** (1953) 90 – 94.
14. A. Davidsohn, *Ind. Chem.* **26** (1950) 385 – 394.
15. Battenfeld Grease and Oil Corp., US 2 753 364, 1951.
16. Socony Vacuum Oil, US 2 389 873, 1945.
17. IG Farben, GB 335 863, 1929.
18. National Lead Co., US 2 650 932, 1953.
19. IG Farben, GB 478 587, 1938.
20. Socony Vacuum Oil, US 2 311 310, 1943.
21. L. W. Ryan, W. W. Pletscher, *Ind. Eng. Chem.* **26** (1934) 909 – 910.
22. IG Farben, US 2 095 508, 1937.
23. R. C. Pink, *Trans. Faraday Soc.* **37** (1941) 181.
24. Hoesch-Chemie GmbH, US 3 519 571, 1967.
25. Tenneco Chemicals Inc., DE-AS 2 159 347, 1971.

26. US 3 779 962, 1970, (G. Koenen, A. Szczepanek).
27. Supraton F. I. Zucker GmbH, US 4 376 079, 1981.
28. Mallinckrodt, Inc., EP 0 086 362, 1983.
29. Tenneco Chemicals Inc., EP-A 0 058 792, 1981.
30. Alexander Gray, *J. Phys. Colloid Chem.* **53** (1949) 9 – 39.
31. Chemische Fabrik Hoesch KG, DE 1 068 237, 1955.
32. ISO 4619: "Driers for Paints and Varnishes," Jan. 1982.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Metallurgy

Standard Article

Joachim Krüger[1], Jan Reisener[2], Markus Reuter[3], Klaus Richter[4]

[1]Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany

[2]Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany

[3]Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany

[4]Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule Aachen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a16_375
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Raw Materials and Products
2.1. Energy
2.2. Ores
2.3. Intermediate Materials
2.4. Scrap
2.5. Products
3. Material Concentration
3.1. Size Reduction
3.2. Particle Selection
3.3. Agglomeration
4. Pyrometallurgy
4.1. Feed Preparation
4.2. Reduction
5. Hydrometallurgy
5.1. Leaching
5.2. Purification and Enrichment
5.3. Metal Recovery
6. Electrometallurgy
6.1. Electrochemical Metal Production
6.2. Electric Furnace Operation
7. Metal Refining
7.1. Pyrometallurgical Methods
7.2. Hydrometallurgical Methods
8. Metal Recycling
9. Fabrication Technology

[Top of Page]

1. Introduction

Metallurgy, as discussed in this article, covers only the field of metal production and refining, starting with the handling of raw materials and ending with refined metals. This definition excludes the topics fabrication technology and properties of metals and alloys (physical metallurgy). The fabrication of metals is mentioned very briefly at the end of this article to place it into perspective with regard to extractive metallurgy as discussed here.

Concentration of minerals, scrap, and intermediate materials is the first major step in metallurgy and involves the size reduction of the aforementioned materials, and separation from the bulk of the gangue material, followed by agglomeration, if necessary. This renders the aforementioned materials into a suitable form for downstream processes.

The subsequent extractive metallurgical processes cover four main areas:

1. Pyrometallurgy: smelting operations at high temperature
2. Hydrometallurgy: aqueous solution operations and electrolysis at low temperature
3. Metal refining
4. Electrometallurgy: electrolysis, smelting, melting, and refining operations at high temperature, and aqueous electrolysis and refining at low temperature

A general flow diagram of metallurgy (Fig. 1) illustrates the interrelationship of these aspects.

Figure 1. General flow sheet of extractive metallurgy

[Top of Page]

2. Raw Materials and Products [10-13]

Without raw materials and energy no extractive metallurgy and hence no products could exist. This chapter presents an overview of all the raw materials (primary or secondary) and energy sources required to produce metallurgical products.

2.1. Energy

Most metallurgical processes require energy for the completion of the following operations: size reduction, heating, reactions, mixing, and phase separation. The required energy may be available in the form of chemical energy, heat from the combustion of fuels, electrical energy, sensible heat, mechanical energy, or buoyancy forces. In many cases an energy source is not used for only one purpose, but for several at the same time. On the other hand several energy sources may be used in one process. Examples of the application of these energy sources are:

1. Melting of metals and alloys in an induction furnace is performed with electrical energy which heats and agitates the melt.
2. During smelting in shaft furnaces, coke has a multitude of functions, e.g., heating, reduction, and mechanical stabilization in the shaft column (⟶ Iron).
3. Oxygen injection in a steel converter (⟶ Steel) is the only energy source for reactions (exothermic oxidation of impurities), mixing (by buoyancy forces of the gas), and heating (via energy released by exothermic oxidation reactions).
4. In Waelz kiln operations heating is done principally by the combustion of hydrocarbons, the intermediate reduction is brought about by the partial burning of added coke, and mixing is effected by the rotation of the furnace.

In some cases these different kinds of energy are interchangeable, in other cases, only one specific type of energy may be applied.

2.2. Ores

Most metals occur in nature in mineral form, although a small number of metals (e.g., Cu, Ag, Au, Hg, Pt) may occur in their metallic state. A mineral can be defined as a homogeneous substance of definite composition and crystal structure. These minerals are found in varying compositions in nature in close association with various gangue materials, this association is called an ore if it can be exploited economically. Another large source of primary minerals are the oceans and salt lakes. Table 1 depicts the usual classification of minerals (with examples) according to their chemical composition.

Table 1. Classification of minerals and examples

Substances	Examples
Elements	Au, Ag, Cu, As, Sb, Bi
Sulfides	CuS, Cu_2S, PbS, ZnS, NiS, Cu_5FeS_4,

	$CuFeS_2$, FeS, FeS_2, MoS_2
Oxides and hydroxides	Fe_2O_3, Fe_3O_4, $FeTiO_3$, $FeCr_2O_4$, TiO_2, MnO_2, SnO_2, Al_2O_3, $Al(OH)_3$, AlOOH, FeOOH
Halides	NaCl, KCl, CaF_2, Na_3AlF_6, $KMgCl_3 \cdot 6\ H_2O$, $Cu_2Cl(OH)_3$
Carbonates, nitrates, and borates	$MgCO_3$, $CaCO_3$, $FeCO_3$, $MnCO_3$, $ZnCO_3$, $CaMg(CO_3)_2$, $Cu_2CO_3(OH)_2$, $NaNO_3$, $Na_2B_4O_7 \cdot 10\ H_2O$
Sulfates, chromates, molybdates, and tungstenates	$BaSO_4$, $PbSO_4$, $CaSO_4$, $CaSO_4 \cdot 2\ H_2O$, $KFe_3(SO_4)_2(OH)_6$, $CaWO_4$, $FeWO_4$
Phosphates, arsenates, and vanadates	$(Ca, Ce, Nd)PO_4$, $(Ce, La, Th)PO_4$
Silicates	$ZrSiO_4$, $ZnSiO_4$, SiO_2, $LiAlSi_2O_6$

In most cases primary ores are further processed close to the mine site to decrease their volume and increase the concentration of valuable mineral(s) therein, before they are shipped or undergo further metallurgical treatment.

Additional materials are required to enhance the extraction of metals from the primary ores and for subsequent refining. These are substances which:

1. produce better phase separation during smelting,
2. serve as solvents for minor components during smelting, and
3. enable selective oxidation or reduction of certain components.

2.3. Intermediate Materials

Ores mostly contain more than one economically valuable mineral. However, the concentration of these valuable minerals may be too low in most cases for direct *pyrometallurgical* recovery, hence they are accumulated in certain *slags* (mixtures of oxides), *mattes* (mixtures of heavy-metal sulfides), *speisses* (mixtures of heavy-metal arsenides and antimonides), *drosses* (mixtures of metals and oxides), *skimmings* (mixtures of metals and oxides), *flue dusts* (mixtures of oxides, chlorides, and sulfates) etc., which are subsequently refined. In *hydrometallurgical* operations these elements accumulate in sludge, leach residue, anodic slime, and cementates. Processing of these materials not only prevents loss of valuable minerals but also minimizes the volume of the wastes and costs of disposal. Table 2 summarizes some intermediate materials.

Table 2. Intermediate materials in metallurgy

Material	Mixture of	Collector for
Pyrometallurgy		
Slags	stable oxides	Zn (Pb)
Mattes	heavy-metal sulfides	Cu, Ni, PM *
Speisses	heavy-metal	
arsenides and antimonides	Cu, Co, Ni, As, Sb, PM *	
Flue dusts	oxides, chlorides,	
and sulfates	Pb, Zn, Cd, Tl	
Drosses	metals and oxides	depends on process
Skimmings	metals, oxides, etc.	depends on process
Hydrometallurgy		
Sludges, leach		
residues	insolubles	Pb, PM *
Anodic slimes		PM *, Pb, As, Sb, Sn, Se, Te
Cementates	metal slimes	Cu, Ni, Co, Cd

* Precious metals.

2.4. Scrap

Scrap is metallic and generally requires only sorting, melting, and refining before products are manufactured. Processing of secondary metals usually requires much less energy than processing of primary materials and, in addition, results in less pollution. Three types of scrap may be distinguished:

1. New scrap (home scrap): generated during the production of metals
2. Prompt scrap: generated in the manufacturing of metal products
3. Old scrap (obsolete scrap): waste products such as used batteries, automobiles, cans

Because of its unknown composition, recycling of obsolete scrap is the only item that presents a problem during refining.

2.5. Products

The final products of hydro- and pyrometallurgical operations appear in various forms, the most common being metal alloys and, in some cases, inorganic compounds of the valuable metals. The characteristics of the product may be defined in terms of the following:

1. chemical state (metal, oxide, salt, etc.),
2. physical state (solid, liquid, bar, rod, powder, solution, etc.), and
3. chemical and physical standards which have to be met (purity, microstructure etc.). A variety of standards for metals and alloys has been established in the international metal industry (e.g., ASTM, SAE, DIN, EN and BS) to facilitate trade. Similar standards for inorganic metal compounds can be found in the different national pharmacopoeias (e.g., B.P., DAB, and U.S.P.). However, the chemical industry relies largely on producer specifications.

[Top of Page]

3. Material Concentration [14-17]

A series of relatively cheap processes, mostly physical, is applied to prepare ore and other raw feed materials for treatment in the various metallurgical processes. In general these processes can be discussed under the following headings:

1. *Size reduction,* during which the valuable minerals in the ore are released and exposed from the surrounding gangue material, or scrap and intermediate materials are rendered suitable as an input to subsequent processes
2. *Particle selection,* which concentrates and, to a large extent, separates valuable minerals or materials from the bulk of the gangue
3. *Agglomeration,* which agglomerates fine materials into larger particles, for use in blast furnaces, for example

The interrelations of these aspects are shown in Figure 2.

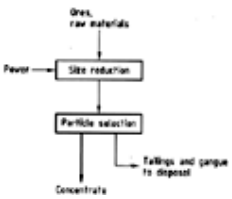

Figure 2. Flow sheet of material concentration

3.1. Size Reduction (⟶ Size Reduction)

Size reduction (comminution) is accomplished in a sequence of physical processes, which commence with *crushing* in jaw, gyratory, cone, roller, and fine crushers, and, if possible, in hammer mills. Oversize material is recycled to the appropriate crushing unit in a closed-loop configuration.

Subsequent *milling* is carried out in ball, rod, tube, and pebble mills in a closed-loop configuration. In many modern applications, however, crushing has been replaced by closed-loop autogenous and semiautogenous milling, in which hydrocyclones and other classification equipment are used to produce a suitable particle size distribution. The size reduction of scrap and intermediate material can also be effected by shredders.

3.2. Particle Selection

During sorting, various properties of the particles (e.g., size, density, luminosity, magnetic, electrostatic, and surface properties) may be utilized to separate the valuable mineral or material from the surrounding gangue.

Separation According to Particle Size. The separation of particles according to size is effected by the use of physical barriers and screens (⟶ Screening) or by the differential movement of solid particles through fluids. These operations are performed in, e.g., rake classifiers (⟶ Elutriation), sedimentation classifiers, and hydrocyclones (⟶ Centrifuges, Sedimenting, ⟶ Hydrocyclones).

Separation According to Density. The sorting of particles as a function of their density may be performed in unit operations and devices such as Humphreys Spiral, shaking tables, jigging and tabling operations, as well as dense-medium separations (⟶).

Separation due to Surface Properties. Under certain chemical conditions, air bubbles, which are blown through a pulp, attach selectively to the valuable particles (e.g., various sulfides) and float to a froth which is skimmed off and collected as concentrate. This flotation (⟶ Flotation) process is carried out in cells, columns, and air-sparged hydrocyclones. As

sufficient concentration cannot be achieved in one unit alone, they are arranged in banks which are interconnected so as to produce maximum recovery.

Magnetic and Electrostatic Sorting (⟶ ; ⟶). These techniques (wet or dry) utilize the magnetic (e.g., with magnetite), electrical conductive, and induced electrostatic properties of various minerals and materials to separate the valuable mineral or material from the surrounding gangue. This occurs mostly in rotating drumlike units arranged into suitable circuits. Electroosmosis, on the other hand, produces a separation between two electrodes.

X-Ray Sorting (⟶). When certain materials (e.g., diamonds) are irradiated by X-rays they become luminous and, hence, can be detected electronically and sorted from the gangue material.

Solid – Liquid Separation. The recovery of solids from liquid media, either to clarify solutions for hydrometallurgical treatment or to dewater pulps for pyrometallurgical operations, can be performed in drum, leaf, and sand filters (⟶ Filtration), thickeners, as well as decanting and centrifuging processes.

Gas Cleaning. Gas cyclones, bag filters, and electrostatic precipitators (⟶ Dust Separation) are used to clean gases or to recover valuable minerals from flue gases from pyrometallurgical operations, before they are vented to the atmosphere.

3.3. Agglomeration (⟶ Size Enlargement)

If the particle's size is too small for use in subsequent metallurgical unit operations, it is agglomerated into larger, more suitable lumps of appropriate size and strength. The unit operations used to achieve this objective are briquetting, pelletizing, sintering, and nodularizing.

Briquetting. Ore fines are mixed with a suitable binder and mechanically pressed into a suitable shape at room or higher temperature. The briquettes may subsequently be hardened in a kiln.

Pelletizing. A suitable binder and water are added to the ore fines, which are fed into a inclined rotating drum or disk. These green pellets may subsequently be fired in a furnace to increase their strength.

Sintering. Coal dust (5 – 6 %), ore fines, and water are mixed and fed onto the grate of a sinterstrand. The carbon in the sinter bed is ignited, and air sucked or blown through the bed produces an ignition front which facilitates drying, calcination, volatilization of the compounds in the bed, and fusion of the materials in the bed, imparting strength to the sinter. If sulfides are roasted and sintered, no carbon need be added.

Nodularizing. The same feed materials used for sintering are fed to an inclined rotating drum at an elevated temperature to produce nodules.

[Top of Page]

4. Pyrometallurgy [18]

Pyrometallurgical processes are used for the treatment of concentrates and ores to effect the following changes:

1. Conversion of certain compounds in ores and concentrates into other compounds
2. Separation of components and enrichment of valuable metals in certain phases
3. Reduction and refining of metals

All of these processes are characterized by high temperatures (200 – 3000 °C), large production rates (up to 10 000 t/d), and fast reaction rates.

Pyrometallurgy is discussed here under feed preparation and metal reduction. Refining is treated in Chapter Metal Refining, because it also includes hydrometallurgical processes. Each of these topics is discussed with reference to appropriate examples. The interrelation of these steps is depicted in Figure 3. An enrichment step is sometimes necessary before reduction can take place. Figure 3 shows that slag and flue dust are often recycled to recover their valuable metal content. The aforementioned steps, as they are applied in various pyrometallurgical operations in practice, are summarized in Tables 3 and 5).

Figure 3. Scheme for pyrometallurgical metal production

4.1. Feed Preparation

Ores and concentrates often do not have the optimum composition and physical state that is required for direct metal production. Thus, a certain amount of feed preparation is necessary. This includes drying (water removal), calcining (removal of hydroxide and/or carbonate), roasting (removal of sulfur as SO_2), and sintering to agglomerate fines to lumpy material. These processes concentrate the valuable compound only marginally, which makes separation of gangue and worthless

components from the metals worthwhile before reduction. This *enrichment* is done in several processes:

1. **Matte smelting** for copper or nickel enrichment in a sulfide phase (⟶ Copper – Pyrometallurgical Principles, ⟶ Nickel)
2. **Waelz process** for lead – zinc volatilization from different low-grade feeds into a high-grade Waelz oxide (⟶ Lead – Products of the Lead Blast Furnace)
3. **Slag reduction** of copper – nickel slags for final metal recovery
4. **Slag fuming** for lead – zinc – tin volatilization from slags into a fuming oxide (⟶ Lead – Products of the Lead Blast Furnace)

4.2. Reduction

Reduction of metal compounds takes place with oxides, chlorides, and sulfides. Reducing agents for oxides include coke for iron and lead oxide, hydrogen for tungsten trioxide, silicon for magnesium oxide, reactive metals such as sodium for $TiCl_4$ (Kroll process; ⟶), and sulfidic sulfur and oxygen (via intermediate formation of SO_2) for sulfides (copper converting, ⟶ Copper – Discontinuous Matte Conversion).

In these pyrometallurgical operations energy is necessary for heating and phase separation, and in the production of metal, energy is also required for the reduction steps. In many cases, the energy is available as chemical energy and the process may need no additional energy from external sources, as can be seen from Table 3.

Table 3. Process steps for pyrometallurgical operations

Reaction	Reactor	Metals	Energy source
Feed preparation			
Drying	rotary kiln	different	hydrocarbons
Calcining	fluidized bed, rotary kiln	Al	hydrocarbons
Roasting of sulfides	fluidized bed	Zn	autogenous with air
	multiple hearth	Mo	
Sinter roasting of sulfides	DL-sintering machine	Pb, PbZn	autogenous with air
Sintering of oxides	DL-sintering machine	Fe	coke
Enrichment			
Matte smelting	flash smelter	Cu, Ni	autogenous with air
Slag reduction	electric furnace	Cu, Pb	electric and coke
Slag fuming	fuming furnace	Pb, Zn	coal
Waelz process	rotary kiln	Pb, Zn	coal
Metal production			
Reduction with carbon	shaft furnace	Fe, Pb, PbZn	coke
Metallothermic reduction			
Oxides	various reactors	V, Mg, Cr	Al, FeSi
Chlorides	retort furnace	Ti, Zr	Na, Mg
Reduction with hydrogen	retort furnace	Mo, W	H_2– heating
Electrochemical reduction	electrolytic cells	Al, Mg	electrical
Conversion of sulfides	converters	Cu, Ni, Pb	autogenous
	flash smelter	Pb, Cu	autogenous

[Top of Page]

5. Hydrometallurgy [19-22]

Most hydrometallurgical operations commence with leaching during which the valuable exposed or liberated minerals go into solution. This initial step is then followed by various processes that purify the leach solution and finally precipitate the valuable metal(s) as compounds or elements in a series of controlled steps. These metallurgical processes are discussed under the headings leaching, solid – liquid separation, purification and precipitation, and metal production. The interrelation of each of these aspects is depicted in Figure 4.

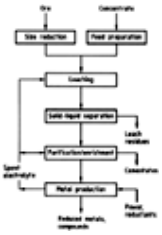

Figure 4. Scheme for hydrometallurgical treatment

5.1. Leaching (⟶ Liquid–Solid Extraction)

During leaching, the valuable mineral and other associated minerals are dissolved by an appropriate reagent under certain conditions (e.g., pH). Usually Pourbaix diagrams (⟶ Liquid–Solid Extraction) can give an indication of which leaching conditions are appropriate. Typical leaching reagents and reactions are:

1. water for water-soluble compounds (copper sulfate);
2. sulfuric, hydrochloric, and nitric acids or sodium hydroxide for metal oxides;
3. complexing agents e.g., cyanide (gold, silver) or ammonia (copper – nickel minerals);
4. reduction of a mineral by a suitable gas or oxidation, for example, manganese dioxide by sulfur dioxide, and
5. acid – base reactions, e.g., recovery of a tungsten complex from tungstates at high pH.

Bacterial leaching techniques involving microorganisms (e.g., *Thiobacillus ferrooxidans*) have been used to increase the leaching rates of various sulfides.

The physical method used to facilitate the leaching depends on the ore size, leaching rate, ore composition, and subsequent isolation and precipitation techniques.

Heap and Percolation Leaching. For particles in the size range of 6 – 200 mm, heap and percolation leaching techniques may be used. The leaching rates and recoveries, however, are not very high for these processes.

Agitation Leaching. Agitation leaching is more efficient than the two aforementioned techniques and is most widely used. In agitation leaching, the fine particles of a pulp – leaching reagent mixture are suspended in stirred tank reactors using either mechanical agitation (impellers, paddles, or turbines) or compressed air in Pachuca tanks. This process can be performed batchwise or continuously in a train of reactors.

Pressure Leaching. When the aforementioned techniques do not produce sufficient leaching rates, pressure leaching can be performed at elevated temperature [e.g., the Bayer process in which bauxite is leached (⟶ Aluminum Oxide – History and Procedure)].

In Situ Leaching. In situ leaching is performed at the ore body in the mine; the ore is flushed with an appropriate leaching reagent and the leach solution is collected for subsequent hydrometallurgical treatment.

Solid – Liquid Separation. After leaching and precipitation, various leach residues, gangue materials, or valuable compounds must be separated from the liquid phase. Various solid – liquid separation techniques are mentioned in Chapter Material Concentration .

5.2. Purification and Enrichment

During leaching a number of elements, in addition to the main element, go into solution which could affect the downstream processes or the final product quality. A variety of hydrometallurgical purification and enrichment techniques exist which could be used to produce suitable leach liquors, e.g., precipitation (crystallization), ion exchange (including carbon-inpulp/leach), solvent extraction, reverse osmosis, and ultrafiltration. Once the raw leach solutions have been purified and the appropriate metallic ions isolated or concentrated, they can be recovered by the methods discussed in Section Metal Recovery.

Purification.
Crystallization (⟶ Crystallization and Precipitation). Depending on the concentration of gangue ions in solution, appropriate selection of temperature and pressure can render the solution saturated; hence crystallization of these ions can occur, and they can be separated from the valuable leach liquor by solid – liquid separation techniques.

pH and Potential Control. By controlling the oxygen partial pressure, the potential, and the pH of a solution, specific ions may be selectively precipitated as sulfates, sulfides, or hydroxides, or partially reduced into more suitable ionic species. These precipitates are separated from the liquor by solid – liquid separation techniques.

Enrichment.
Ion Exchange. Various ion-exchange resins and fibers have been developed onto which cationic and anionic groups are loosely bonded; these can be replaced by appropriate metal ions, when the latter pass over these resins and fibers. These extracted metal ions can subsequently be reclaimed by stripping the resin or fibers with a suitable solvent. An example is the recovery of uranium from leach liquors. Activated carbon has a similar function and can be used to adsorb various metal ions and complexes from solution in continuous or batch stirred tank reactors. The carbon-in-pulp/leach process is applied, e.g., in the gold industry (⟶ Gold, Gold Alloys, and Gold Compounds – Recovery of Gold with Carbon).

Solvent Extraction (⟶ Liquid–Liquid Extraction). Some organic solvents that are immiscible with water can selectively extract certain metal ions and complexes from aqueous solutions. Examples of these extractants are

1. solvating extractants, which render certain metal ions neutral and hence soluble in organic solvents by solvation;
2. cation-exchange extractants (⟶ Liquid–Liquid Extraction – Introduction), i.e., organic acids in which hydrogen ions can be replaced by metal ions (⟶ Liquid–Liquid Extraction – Phosphoric, Sulfonic, and Carboxylic Acids.);
3. anion-exchange extractants (⟶ Liquid–Liquid Extraction – Introduction), in which metals present as complex metal anionic salts in solution are exchanged for chloride anions present as organic salts in the organic solvent (e.g., amines ⟶ Liquid–Liquid Extraction – Amine and Phosphine Salts.); and
4. chelating extractants (⟶ Liquid–Liquid Extraction – Introduction, ⟶ Liquid–Liquid Extraction), in which the organic compound forms an organometallic complex with the metal ion.

The valuable metal ions are subsequently recovered by various stripping techniques from the organic phase and the extractant is regenerated.

Reverse Osmosis (⟶ Membranes and Membrane Separation Processes – Principle of Reverse Osmosis). When a pressure in excess of the osmotic pressure is applied to a concentrated leach solution, water passes from the concentrated to the dilute solution (water), which is separated from the concentrated one by a semipermeable membrane. This leaves the metallic ions in a more concentrated form and suitable for subsequent precipitation.

Ultrafiltration (⟶ Membranes and Membrane Separation Processes – Principle of Ultrafiltration). Ultrafiltration involves selective sieving of molecules (M_r > 500) through the pores of a membrane, external pressure being the driving force.

5.3. Metal Recovery

A metal may be recovered from the leach solutions either as a compound or as an element. These aspects are discussed under precipitation and reduction, respectively.

Precipitation.
Crystallization and pH Control. The same discussion of enrichment and purification is true here also, but with reference to the valuable metal ion to be recovered. Examples are the production of aluminum hydroxide from bauxite ores via the Bayer process or the precipitation of ammonium paratungstate as discrete crystals under favorable conditions of temperature, pH (rate of addition of base), and agitation.

Reduction. Precipitation of the valuable metal in metallic form may be achieved by various methods, all of them involving an oxidation – reduction reaction.

Cementation. Cementation involves the oxidation of a less noble metal (more anodic) by a more noble metal (more cathodic) ion in solution, e.g., precipitation of copper ions in solution by iron or zinc which go into solution and zinc cementation in the older gold-processing technology or in zinc solution purification.

Gaseous Reduction. By bubbling appropriate gases (e.g., hydrogen) through solutions containing metal ions (e.g., reduction of nickel from ammine solutions), selective reduction can be achieved if the pH, the hydrogen partial pressure, and the activity of the metal ion are controlled.

Other gases that are used are hydrogen sulfide, sulfur dioxide, and carbon monoxide, e.g., the Mantos Blancos process utilizes sulfur dioxide to produce copper(I) chloride from atacamite leach solutions (⟶ Copper – QS Process).

Electrolysis. Metals can be recovered from their aqueous solutions by electrolysis. This involves deposition of the required metal on a suitable cathode constructed either from the metal being recovered or from a suitable alloy which provides a high hydrogen overpotential. The anode is usually chosen so that it does not dissolve during electrolysis. However, during electrolytic refining, anodes are made of the same metal that is being recovered; copper and zinc electrolysis are typical examples. Other applications of electrolysis are discussed in Chapters Electrometallurgy and Metal Refining.

[Top of Page]

6. Electrometallurgy [23, 24]

Electrometallurgy includes all processes that employ electrical energy for metallurgical purposes; hence, it is employed in both pyro- and hydrometallurgy. The two principal applications are electrochemical metal production and refining, and heating.

In this chapter only the electrolysis of fused salts is discussed (for aqueous electrolysis, see Chap. Hydrometallurgy).

6.1. Electrochemical Metal Production

Electrochemical reduction is used for metals with very low reduction potential, so low that they cannot be recovered from aqueous solutions. Hence, these metals are recovered via the electrolysis of fused salts; examples are aluminum, magnesium, titanium, and alkali metals. All these processes require pure salts as feed material (e.g., Al_2O_3, $MgCl_2$, $TiCl_4$, NaCl, and KCl), to avoid co-reduction of impurities. This type of operation is conducted at 600 – 950 °C, the heat originating mainly from the ohmic resistance of the melt. The electrolytes are fluorides and chlorides.

Electrochemical reduction produces metals of very high purity, but at a very high energy consumption and a low production rate, if compared to pyrometallurgical methods.

6.2. Electric Furnace Operation

Electric furnaces (⟶) are used in the pyrometallurgical process industry during metal production (see Table 4). Further applications are in the fields of metal and alloy melting (see Table 5).

Table 4. Metallurgical application of electric furnaces

Smelting	Metals
Smelting of ores	FeNi
Smelting of concentrates	CuNi
Reduction of ferroalloys	W, Mo, V, Cr, Si, Mn
Production of special materials	P, graphite, SiC, CaC_2, Ca_2Si

Table 5. Electric furnaces for metal melting

Metal melting operation	Open furnace	Vacuum furnace
Arc	steel	steel, Ti, superalloys
Plasma	steel	Ti
Induction	steel, Cu, Al, Ni, Zn	steel, Ni, superalloys
Resistance (electro slag refining)	steel	
Electron beam		steel, Nb, Ta, Ti, V, Zr, Hf, superalloys

The advantages of the application of electrical furnaces can be summarized as follows:

1. Very high temperatures can be reached.
2. Energy density is high.
3. High product purity and production rate can be achieved.
4. Metals of uniform composition can be produced.
5. Any desired atmosphere may be applied.
6. Waste gas volume is small; hence, environmental control is easier.

[Top of Page]

7. Metal Refining

Every metal produced by either pyro- or hydrometallurgical methods contains various amounts of impurities. Depending on the application of the metal, the impurity content must be lowered to levels determined by product specifications set by the user. These impurities may be divided into two groups i.e., those elements that are economically recoverable and secondly, those that have little economic value. Refining can be achieved by following a pyrometallurgical or a hydrometallurgical route. This, however, depends on the product specification or the economic value of the byproducts. A general refining scheme cannot be presented (Fig. 5 gives a simplification), but common refining methods are listed in Table 6.

Table 6. Common refining methods and examples

Pyrometallurgical	Hydrometallurgical
Physical methods	
Liquation (Pb, Zn)	Crystallization (Al, W)
Zone melting (Si, Ge)	
Distillation (Zn)	
Degassing (steel, Al)	
Vacuum melting (steel, Co – Ni alloys, Nb, Ta, V, Ti, Zr, Hf)	
Filtration (Al)	
Chemical methods	

Oxidation (steel, Cu, Pb)	Electrolytic refining (Cu, Pb, Au, Ni, Ag)
Chlorination (Al)	Precipitation (jarosite, Cu_2Cl_2)
Sulfurization (Pb)	Cementation (Cu, Co, Cd by Zn)
Formation of intermetallic compounds (Pb, Sn)	Solvent extraction (Cu, Zn, Co, Zr – Hf)
Deoxidation (steel, Cu)	
Electro slag refining (steel)	Ion exchange (U)

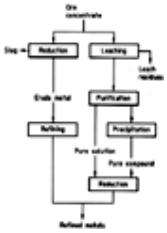

Figure 5. Scheme for metal refining

Some of the more common metals are refined *after* reduction (e.g., Cu, Ag, Au, Pb, Sn, Zn, and Fe), others must be purified *before* reduction (refractory metals, platinum group metals, rare earths, CoNi, Al, Zn) because their properties are similar to other metals associated with them (see Fig. 5).

Pyrometallurgical refining is generally fast, but not very selective, whereas hydrometallurgical refining has high selectivity, but usually lower production rates. Each of these routes is discussed separately and illustrated by appropriate examples.

7.1. Pyrometallurgical Methods

Pyrometallurgical methods may be subdivided into two paths: i.e., physical and chemical methods.

Physical Methods.
Liquation. Liquation is a rough separation process, mostly for nonferrous systems, where large concentration differences exist either between immiscible liquids or between solid and liquid phases, which are in equilibrium with one another. It is carried out by slow cooling of the liquid metal mixture in hearth furnaces or steel kettles. An example is copper liquation from liquid lead, which may be explained in terms of the low melting point eutectic mixture which exists in the copper – lead system at 99.94 wt % liquid lead: if the mixture is slowly cooled down to about 350 °C, most of the copper separates out on the surface of the lead and can thus be recovered (⟶ Lead – Pyrometallurgical Refining).

Zone Melting. Metals of very high purity are produced from solid metals by this method. A small molten zone is moved along a bar or rod of the feed metal. The ensuing concentration difference between the liquid and solid phases produces enrichment of the impurities in the moving liquid zone.

Distillation. Due to the difference in the vapor pressure of various components within a metal mixture, one or more of the more volatile components evaporates preferentially from the metal melt. The distillation and the bottom (distillation residue) products are recovered. The rate of distillation can be increased by the application of vacuum techniques. Liquid mixtures of metals with small differences in vapor pressures are separated by rectification.

Degassing. Product specifications require low levels of dissolved gasses in the final metal product, especially hydrogen, nitrogen, and oxygen, which must be removed. Carrier gas degassing and vacuum degassing are well-established techniques.

Vacuum Melting. Extremely low levels of impurities may be produced in metals and alloys if they are melted under vacuum. This achieves degassing and volatilization at the same time.

Filtration. Metal melts can be filtered to remove floating solid particles, which occur as oxides, carbides, and nitrides.

Chemical Methods. Compounds formed from dissolved impurities in the melt by chemical reactions can be separated from the melt by physical methods. This is the most widely used technique for metal purification. The compounds which are formed may be either in the solid, liquid, or gaseous state.

Oxidation, Chlorination, and Sulfurization. Under oxidizing (sulfurizing or chlorinating) conditions, the impurities in the melt may be removed as oxides (sulfides, chlorides), because their affinity for these substances is higher than that of the base metal. The oxidation product is usually dissolved by a slag or salt melt, which can easily be removed from the metal bath. Oxygen is fed to the melt as pure gas, as air, or in the form of an oxidizing agent; chlorine as gas or salt, and sulfur in its elemental solid state.

Deoxidation. Dissolved oxygen may be removed from melts by the addition of deoxidizing agents such as carbon, hydrogen,

or reactive metals and by formation of the corresponding oxides (e.g., CO, CO_2, H_2O, and Al_2O_3).

Formation of Intermetallic Compounds. Removal of an impurity from a liquid metal can be effected by precipitation of intermetallic compounds. This is performed by dissolving sufficient amounts of particular metals, which subsequently form such compounds with the impurities. After dissolution, agitation, and cooling, a crust of intermetallic phases is formed on the bath.

Electro Slag Refining. In this method a crude metal is refined by a molten slag that is heated electrically by resistance. The crude metal is usually a suspended electrode, the lower part being submerged in the molten slag bath. The molten metal forms droplets that fall through the molten slag, hence they are refined before collecting at the bottom of a water-cooled mold, from which the solid metal is recovered.

7.2. Hydrometallurgical Methods

Most of these methods are described in Chapter Hydrometallurgy , hence only electrolytic refining is discussed here.

Electrolytic refining is a common method for refining crude metals. The crude metal and impurities with a more negative potential (anode) are dissolved electrochemically and transferred to the cathode. Only the principal metal is deposited as a metal on the cathode. Metals with a more positive potential than the principal metal are not dissolved at all.

[Top of Page]

8. Metal Recycling

For economical and ecological reasons, recycling of various scrap materials (see Chap. Raw Materials and Products) is a very important aspect of extractive metallurgy. A general overview of various recycling routes is given in Figure 6.

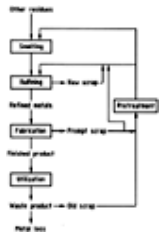

Figure 6. Scheme for metal recycling

[Top of Page]

9. Fabrication Technology [25-29]

Designing materials with the required characteristics for a particular application from the refined metals falls into the realm of physical metallurgy. These materials subsequently follow various fabrication steps during which they are transformed into usable finished products. Figure 7 gives a brief overview of fabrication methods and steps.

Figure 7. Brief overview of fabrication technology

For a description of the various methods and applications, see ⟶ Foundry Technology; ⟶ Metal-Matrix Composites; ⟶ Metals, Surface Treatment; ⟶ Powder Metallurgy and Sintered Materials; ⟶ Soldering and Brazing; ⟶ ; ⟶ Aluminum Alloys – Thermal and Mechanical Processes ⟶ Aluminum Alloys – Production Methods for Aluminum Alloys; ⟶ Fibers, 5. Synthetic Inorganic – Metal Fibers; ⟶ High-Temperature Materials – Metallic Materials.

[Top of Page]

References

General References

1. O. Buhtz: *Grundlagen der Metallurgie*, VEB Deutscher Verlag für Grundstoffindustrie, 1st ed., Leipzig 1975.
2. J. D. Gilchrist: *Extraction Metallurgy*, 2nd ed., Pergamon Press, New York 1980.
3. C. B. Gill: *Nonferrous Extractive Metallurgy*, Wiley & Sons, New York 1980.
4. *Winnacker Küchler*, 4th ed., Carl Hanser Verlag, München 1986.
5. P. Hayes: *Process Selection in Extractive Metallurgy*, Hayes Publishing Co, Brisbane 1985.
6. J. J. Moore: *Chemical Metallurgy*, Butterworths, London 1981.

7. F. Pawlek: *Metallhüttenkunde*, De Gruyter, Berlin 1983.
8. R. D. Phelke: *Unit Processes of Extractive Metallurgy*, 6th ed., Elsevier, Amsterdam 1973.
9. T. Rosenqvist: *Principles of Extractive Metallurgy*, McGraw Hill, New York 1974.

Specific References

10. *Minerals Yearbook*, U.S. Dept. of Interior, Bureau of Mines, published annually.
11. *Metals Handbook*, 9th ed., American Society of Metals, Metals Park, Ohio 1979.
12. W. Gocht: *Handbuch der Metallmärkte*, 2nd ed., Springer Verlag, Berlin 1985.
13. L. Müller-Ohlsen: *Die Weltmetallwirtschaft im industriellen Entwicklungsprozess*, J. C. B. Bohr, Tübingen 1981.
14. J. M. Coulson, J. F. Richardson, J. R. Backhurst, J. H. Harker: *Chemical Engineering,* 3rd ed., **vol. 2**, Pergamon Press, New York 1987.
15. A. L. Mular, R. B. Bhappu (eds.): *Mineral Processing Plant Design*, American Institute of Mining, Metallurgical and Petroleum Engineers, Inc., New York 1989.
16. R. H. Perry, C. H. Chilton: *Chemical Engineer's Handbook*, 6th ed., McGraw – Hill New York 1984.
17. H. Schubert: *Aufbereitung fester Mineralischer Rohstoffe*, VEB Deutscher Verlag für Grundstoffindustrie, 3rd ed., vol. **1**, Leipzig 1974, 2nd ed., vol. **2** 1977.
18. C. B. Alcock: *Principles of Pyrometallurgy*, Academic Press, London 1976.
19. R. G. Bautista, R. J. Wesely, G. W. Wesely: *Hydrometallurgical Reactor Design and Kinetics*, The Metallurgical Society, Inc., Pennsylvania 1986.
20. R. G. Bautista: *Hydrometallurgical Process Fundamentals*, Plenum Press, New York 1982.
21. J. O'M. Bockris, B. E. Conway, E. Yeager, R. E. White: *Comprehensive Treatise of Electrochemistry,* vol. 2, Plenum Press, New York 1981.
22. A. Schmidt: *Angewandte Elektrochemie*, Verlag Chemie, Weinheim, Germany 1976.
23. V. P. Elyutin, Y. A. Pavlov, B. E. Levin, E. M. Alekseev: *Production of Ferroalloys Electrometallurgy*, 2nd ed., The State Scientific Technical Publishing House, Moscow 1957.
24. G. Volkert, K.-D. Frank: *Metallurgie der Ferrolegierungen*, 2nd ed., Springer Verlag, Berlin 1972.
25. G. Spur, Th. Stöferle: *Handbuch der Fertigungstechnik,* vols. 1 – 5, Carl Hanser Verlag, München 1983.
26. Metals Handbook, *Forging and Casting*, vol. **5**, 8th ed., American Society of Metals, Metals Park, Ohio, 1977.
27. Metals Handbook, *Heat Treating*, vol. **4**, 9th ed., American Society of Metals, Metals Park, Ohio, 1981.
28. Metals Handbook, *Surface Cleaning, Finishing, and Coating*, vol. **5**, 9th ed., American Society of Metals, Metals Park, Ohio, 1981.
29. Metals Handbook, *Powder Metallurgy*, vol. **7**, 9th ed., American Society of Metals, Metals Park, Ohio 1981.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Metal-Matrix Composites

GERHARD IBE, Vereinigte Aluminiumwerke AG, Bonn, Federal Republic of Germany

1. **Introduction** 1
2. **Components** 2
3. **Mechanics of Reinforcement** 2
3.1. **Continuous Fibers** 2
3.2. **Short Fibers and Whiskers** 4
3.3. **Particles** 5
3.4. **Mechanical Properties** 6
4. **Interactions between the Metal Matrix and Reinforcing Material** 6
4.1. **Wetting and Chemical Reaction** 6
4.2. **Bonding and Composite Strength** . . . 7
4.3. **Chemical Compatibility** 8
5. **Production** 8
5.1. **Sheets and Tapes** 8
5.2. **Cast Parts** 9
5.3. **Semifinished Products** 10
5.4. **Hybrid Composites** 11
6. **Working and Machining** 12
7. **Uses** 13
8. **Occupational Health** 13
9. **References** 14

1. Introduction

In aerospace, automotive, and general engineering, new developments require new materials with a low density and improved properties. Thin, stiff, high-strength fibers embedded in a plastic matrix can be used for such applications (→ Composite Materials). In an increasing number of applications, however, the inherent properties of the plastic matrix (temperature stability, resistance to the environment, brittleness, strength, and stiffness) are not sufficient. Composite materials with a metal matrix, especially those with a low density, are becoming more important due to their significantly improved properties.

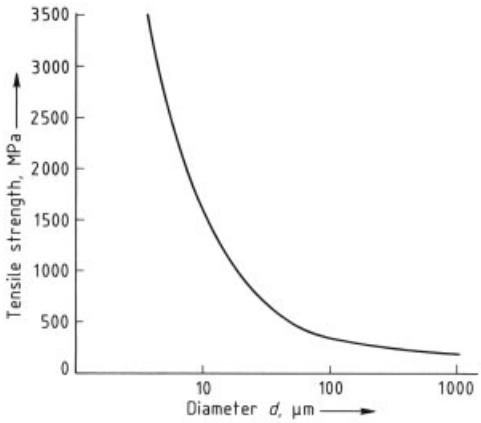

Figure 1. Tensile strength of glass fibers as a function of their diameter [2], [3]

The theoretical strength of crystalline or metallic bodies for all imaginable loading situations would be fully adequate if it could be achieved in practice. It is approximately 10 % of the modulus of elasticity (i.e., 5 – 50 GPa) and can be calculated from the energy required for splitting two atomic planes in the crystal lattice [1]. The theoretical strength cannot be achieved in practice, however, because even compact bodies made from specially produced materials have defects. These defects may be dislocations which cause plastic deformation and thus lead to ductile fracture, or the formation of crack nuclei in nonductile material followed by premature brittle fracture.

In 1920 A. A. GRIFFITH [2] observed that glass fibers increased in strength with decreasing diameter (Fig. 1), and concluded that this offered a way of attaining the theoretical strength. He proposed the so-called " fiber paradox" which states that "material in fiber form is stronger than in compact form and strength increases with decreasing thickness". In a compact brittle ceramic material with a defect density ϱ_d, the average defect distance is given by:

$$l = \varrho_d^{-3}$$

i.e., if $\varrho_d = 1000\,\mathrm{cm}^{-3}$ then $l = 1$ mm (Fig. 2 A) [3]. In fibers of diameter d with the same defect density, the defect distance is given by:

$$l = 4/\left(\pi d^2 \varrho_d\right)$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_389

i.e., when $d = 10\,\mu m$ then $l = 12.7\,m$ (Fig. 2 B). Such fibers can attain strengths of up to 5 GPa. Use of fibers thus allows production of a high-strength material. The fibers are spatially arranged and embedded in a suitable matrix which fixes the fibers, protects them, and transmits forces to the fibers.

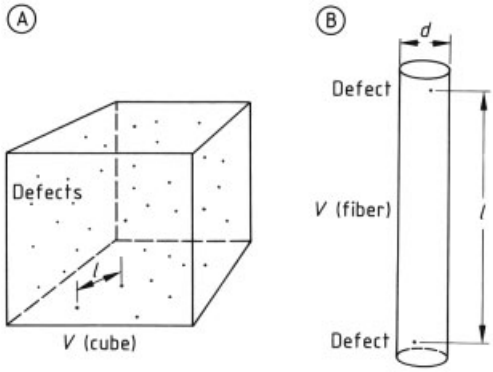

Figure 2. Distance between defects in bulk (A) and fibrous (B) ceramic materials [3]

More detailed discussions of the definition, the classification, the reinforcement mechanisms, and the resulting properties of fiber-reinforced composites as well as a detailed description of fiber reinforcement of plastics are given elsewhere (→ Composite Materials).

The present article concentrates on the reinforcement of metal matrices due to the increasing importance of these materials. More detailed information can be found in [4–6].

2. Components

Metal-Matrix Alloys. The properties of composite materials depend on the matrix properties especially when the fiber content is low or when reinforcement is achieved with short fibers or particles instead of continuous fibers: this is often the case for metallic composites. Table 1 contains important properties of the light metal alloys most commonly used in metal matrix composites [7], [8]; due to their low densities, high specific values for the properties of the composite are possible. Methods to increase the high-temperature strength of the nickel superalloy Nimonic 90 [9] by fiber reinforcement have been currently being investigated. Carbon steel and an epoxy resin are listed for purposes of comparison.

Continuous Inorganic Fibers. Due to the required high-temperature and chemical resistances of the fibers to the matrix alloy during production and also at the use temperatures, only ceramic fibers based on the thermodynamically stable oxides and carbides of light elements (high-silica, high-quartz, alumina, and silicon carbide fibers) can be considered. A summary of the composition and properties of the most important of these fibers is given elsewhere, → Composite Materials, Chap. 4.1.3., see also → Fibers, 5. Synthetic Inorganic, Chap. 3.4.1.

Short Fibers, Whiskers, and Particles. In addition to classical reinforcement with continuous fibers, reinforcement of metals with short fibers, whiskers, and ceramic particles is becoming more important. These composites can be further worked using techniques such as forging or extrusion. The short fibers can be cut from long fibers or produced by similar techniques to staple fibers. Aluminum hydroxide, for example, is spun in a sol – gel process to form bundles of short filaments. These gel-like filaments are dehydrated, dried, and calcined at high temperatures.

Whiskers are short, microscopically fine, single-crystal fibers produced by gas-phase reactions at high temperatures (→ Whiskers). As a result of their small diameter ($< 1\,\mu m$) and the fiber paradox (Chap. 1), they possess very high strength; Table 2 gives typical properties.

Fine ceramic particles (powder with a particle size $\leq 20\,\mu m$) are less expensive and are being introduced as a reinforcing material [10]. Their reinforcement effect is based on a different mechanism–fine grain hardening and strain hardening of the matrix (cf. Section 3.2).

3. Mechanics of Reinforcement

3.1. Continuous Fibers

The modulus of elasticity and the tensile strength of unidirectional, continuous fiber composites can be calculated, with a good approximation, using the linear "rule of mixtures" (→ Composite Materials, Chap. 3.2.). A linear relationship exists between the tensile strength

Table 1. Properties of commonly used matrix alloys and related materials

Material∗	Density (ϱ), g/cm^3	Modulus of elasticity (E), GPa	Tensile strength (R_m), MPa	Linear coefficient of thermal expansion (α), 10^{-6} K^{-1}	Reference
AlMgSiCu wrought alloy (6061)	2.70	70	290	24.0	[7]
AlZnMgCu wrought alloy (7075)	2.80	72	530	21.5	[7]
AlSi10Mg casting alloy (361)	2.65	73	280	20.0	[7]
TiAl6V4 wrought alloy (BZ65-58T-5)	4.43	119	1190	9.9	[8]
MgAl9Zn0.7Mn casting alloy (AZ91)	1.80	46	260	26.1	[9]
Nimonic 90	8.19	229	1250	11.5	[9]
Carbon steel (C 1050)	7.83	210	950	14.6	[9]
Epoxy resin (Epon 828/1031)	1.22	3.6	80	60	[9]

∗ Aluminum alloy designations are in accordance with the recommendations of the Aluminum Association, Washington, D.C.

Table 2. Properties of short fibers, whiskers, and particles [11]

Reinforcement material	Density (ϱ), g/cm^3	Average diameter, µm	Modulus of elasticity (E), GPa	Tensile strength (R_m), GPa	Linear coefficient of thermal expansion (α), 10^{-6} K^{-1}	Reference
Short fibers						
Al_2O_3 (Saffil, ICI)	3.30	3	300	2	5.0	[10, a]
Al_2O_3–SiO_2 (Fiberfrax, Sohio)	2.6	2–5	105	1.03	ca. 6	[10, b]
Al_2O_3 (FP, chopped, Du Pont)	3.95	20	377	1.4	6.8	[10, c]
Whiskers						
SiC (Tokamax, Nippon Carbon)	3.17	0.1–0.5	400–700	3–14	4.5	[10, d]
Al_2O_3	3.9		420	2–18	6–9	[11]
B_4C	2.5		450	2–6	4.5	[11]
Particles						
SiC (ESKH)	3.21	10–20	420	25	4.7	[10, e]

of the composite $\sigma*_c$ and the fiber volume fraction f:

$$\sigma_c^* = f\,\sigma_{F,\mathrm{eff}} + (1-f)\,\tilde{\sigma}_M \tag{1}$$

where $\sigma_{F,\mathrm{eff}}$ is the effective strength of the fibers and $\tilde{\sigma}_M$ is the matrix flow stress at fiber fracture (Fig. 3). However reinforcement above the matrix tensile strength $\sigma*_M$ only occurs when the critical fiber volume fraction (content) f'' is exceeded. For metal-matrix composites the effective fiber strength $\sigma_{F,\mathrm{eff}}$ is lower than the maximum fiber strength $\sigma*_F$ because of alignment problems of the fibers, orientation deviations, and insufficient bonding. The fiber efficiency factor η_F is defined by $\sigma_{F,\mathrm{eff}} = \eta_F \cdot \sigma*_F$ or after Equation (1) by

$$\eta_F = \frac{\sigma_{F,\mathrm{eff}}}{\sigma_F^*} = \frac{\sigma_c^* - (1-f)\,\tilde{\sigma}_M}{f\,\sigma_F^*} \tag{2}$$

When $\eta_F < 1$, the critical fiber volume fraction f'' must be increased to achieve the required strengthening (Fig. 3).

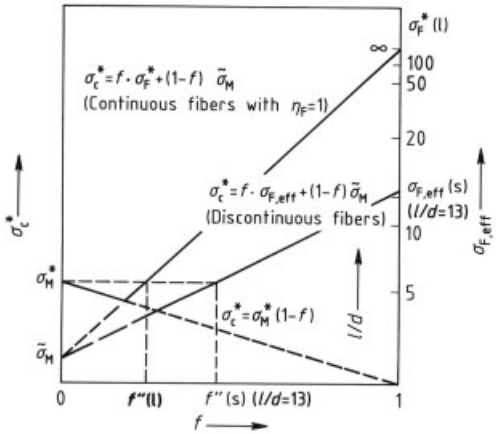

Figure 3. Composite strength $\sigma*_c$ as a function of fiber volume fraction f for continuous (l) and short discontinuous (s) fibers with different aspect ratios l/d (ROM rule of mixtures)
Symbols: $\sigma*_M$ = matrix strength; $\tilde{\sigma}_M$ = matrix flow stress; $\sigma*_F$ = maximum fiber strength; f'' = critical fiber volume fraction; η_F = fiber efficiency factor

3.2. Short Fibers and Whiskers

In short-fiber reinforcement the individual fibers do not carry the full tensile stress over their entire length. This is because this stress, which results from tensile and especially shear stresses in the fiber – matrix boundary surface, must be transmitted to the fibers. The stress from the fiber ends to the fiber centers can be described by the shear lag theory [1], [12]. A simplified representation can be seen for high boundary bonding strength and high tensile stresses when the matrix has exceeded its flow limit. The fiber ends transmit the tensile flow stress $\tilde{\sigma}_M$ and the fiber circumference the shear flow stress $\tilde{\tau}_M$ of the matrix (Fig. 4). Integration of this differential stress equivalency along the fiber

$$\frac{\pi}{4} d_F^2 \frac{d\sigma_F}{dx} dx = \pi d_F \tilde{\tau}_M dx \tag{3}$$

gives the average or the effective tensile strength of a fiber of length l (d_F is the fiber diameter, σ_F the tensile stress within the fiber, and x the length coordinate).

$$\text{For } l > l_c : \sigma_{F,eff} = \sigma_F^* \left(1 - \frac{l_c}{2l}\right) + \tilde{\sigma}_M \frac{l_c}{2l} \tag{4}$$

where l_c is the critical length and fracture is initiated by fiber breakage.

$$\text{For } l < l_c : \sigma_{F,eff} = \tau_M^* \left(\frac{l}{d_F} + 2\right) \tag{5}$$

where $\tau*_M$ is the shear strength of the matrix because fracture now occurs in the matrix.

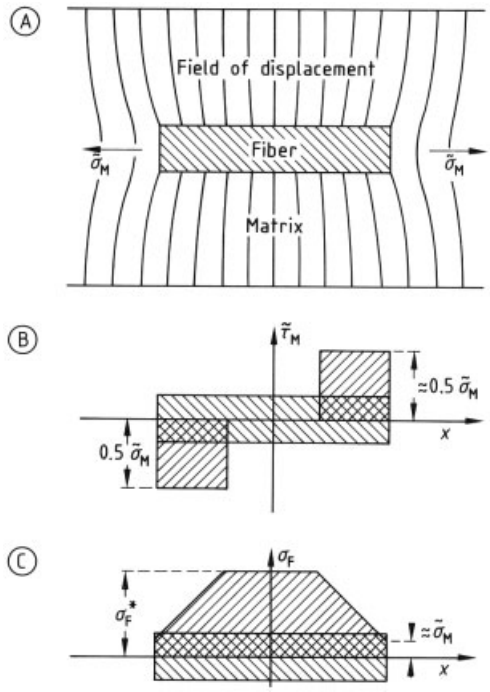

Figure 4. Stresses and strains of short fibers in a ductile matrix
A) Displacement field and tensile flow stress $\tilde{\sigma}_M$ within the matrix; B) Shear flow stress $\tilde{\tau}_M$ of the matrix at the fiber interface; C) Tensile stress σ_F within the fiberSymbols: $\sigma*_F$ = tensile strength of fiber; $\tilde{\sigma}_M$ = tensile matrix flow stress at fiber fracture; x = coordinate of fiber length

The parameter l_c is the critical fiber length below which the tensile stress in the fibers can never reach the true fiber strength $\sigma*_F$

$$l_c = \frac{\left(\sigma_F^* - \tilde{\sigma}_M\right) d_F}{2\tilde{\tau}_M} \tag{6}$$

Using the Schmid's shear stress relationship, the shear flow stress can be related to the more easily measurable tensile flow stress of the matrix:

$$\tilde{\tau}_{\mathrm{M}} \approx 0.5\tilde{\sigma}_{\mathrm{M}} \quad \tau^*_{\mathrm{M}} \approx 0.5\sigma^*_{\mathrm{M}} \tag{7}$$

According to Equation (2) the effective fiber strength and thus the fiber efficiency η_{F} depend only upon the length to thickness ratio (l/d) of the fiber and the matrix flow stress $\tilde{\sigma}_{\mathrm{M}}$ – especially when $l < l_{\mathrm{c}}$. The composite strength $\sigma*_{\mathrm{c}}$ with short fibers is thus more dependent upon the matrix properties than with continuous fibers. The influence of fiber length is especially significant in the case of the critical fiber content f'' where $\sigma*_{\mathrm{c}} = \sigma*_{\mathrm{M}}$.

$$\text{For } l > l_{\mathrm{c}} \quad f'' = \frac{\sigma^*_{\mathrm{M}} - \tilde{\sigma}_{\mathrm{M}}}{\sigma^*_{\mathrm{F}}\left(1 - \frac{\sigma^*_{\mathrm{F}} d_{\mathrm{F}}}{2l\tilde{\sigma}_{\mathrm{M}}}\right) - \tilde{\sigma}_{\mathrm{M}}} \tag{8}$$

$$\text{and for } l < l_{\mathrm{c}} \quad f'' = \frac{\sigma^*_{\mathrm{M}} - \tilde{\sigma}_{\mathrm{M}}}{\left(\frac{l}{2d}\right) - 1} \tag{9}$$

This strong dependency of the critical (the minimum allowable) fiber content on fiber length is shown in Figure 5 for alumina (Saffil and FP) fibers in an aluminum alloy (6061) as matrix material.

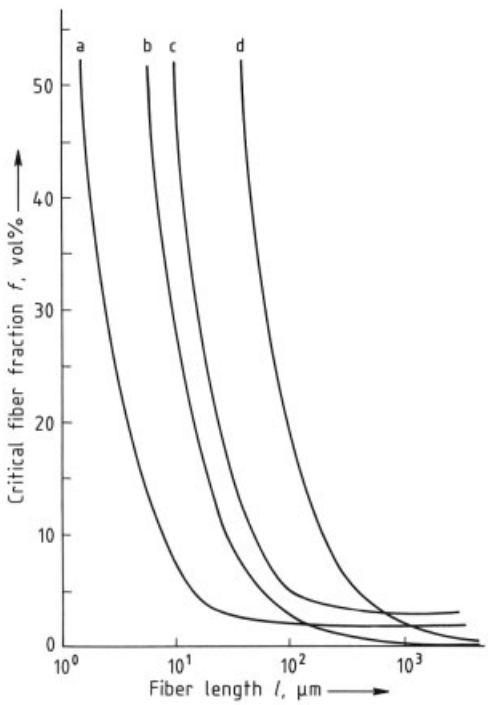

Figure 5. Dependence of the critical fiber content on the length of short alumina fibers in aluminum alloy 6061 [13]
a) Saffil fibers, room temperature; b) Saffil, 300 °C; c) FP, room temperature; d) FP, 300 °C

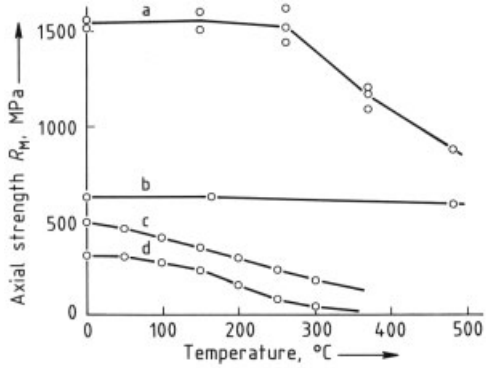

Figure 6. Axial strength of aluminum alloys reinforced with continuous silicon carbide fiber and whiskers at elevated temperature [31, a]
a) Al 6061 with 48 vol % SiC fibers aligned at 0 ° to the applied force; b) Al 6061 with 48 vol % SiC fibers arranged in four layers at 0 °/90 °/90 °/0 ° to the applied force; c) Al 6061-T6 with 16 vol % SiC whiskers; d) Al 6061-T6

3.3. Particles

In the case of particles (which can be regarded as very short fibers), fiber reinforcement does not occur even with a very high particle content. This is because the length of the particle is too short to transfer sufficient load by shear stresses from the matrix to the particle (cf. Eqs. 5 and 7). The observed high strength of such materials occurs because the particles increase the strength (i.e., flow stress) of the matrix. This increase in flow stress is due to several metallurgical mechanisms:

$$\Delta\tilde{\sigma}_{\mathrm{M}} = \Delta\tilde{\sigma}_{\mathrm{M}}(f) + \Delta\tilde{\sigma}_{\mathrm{M}}(V) + \Delta\tilde{\sigma}_{\mathrm{M}}(P) + \Delta\tilde{\sigma}_{\mathrm{M}}(HP) \tag{10}$$

where $\Delta\tilde{\sigma}_{\mathrm{M}}(f)$ is the increase of the flow stress that occurs because the total deformation of the material is restricted to the matrix volume share; $\Delta\tilde{\sigma}_{\mathrm{M}}(V)$ is the strengthening of the matrix caused by a reduction in the cooling stresses brought about by plastic deformation (these stresses are caused by the different thermal expansions that occur during production); $\Delta\tilde{\sigma}_{\mathrm{M}}(P)$ is the hardening of the matrix caused by particles hindering the movement of the dislocations (the Orowan mechanism) [14];

and finally $\Delta\tilde{\sigma}_{\mathrm{M}}(HP)$ is the hardening of the matrix where movement of dislocation groups is hindered by a fine grain-hardening effect (the Hall – Petsch effect) [14]. The last two of these four parameters are the most important. An increase in strength $\tilde{\sigma}_{\mathrm{M}}$ of only 0.5 MPa has been calculated for the Orowan mechanism in composites containing 16 vol % of short fibers with a diameter of 20 µm and a length of 100 µm. For very fine particles with a diameter of 0.1 µm, the corresponding increase is 70 MPa. This is due to the greater inhibition of the dislocation movement by finer dispersed particles with smaller interparticle distances. The Hall – Petsch effect increases the flow stress $\tilde{\sigma}_{\mathrm{M}}$ by 30 MPa both for the short fibers and the particles; this accounts for approximately half of the observed reinforcement.

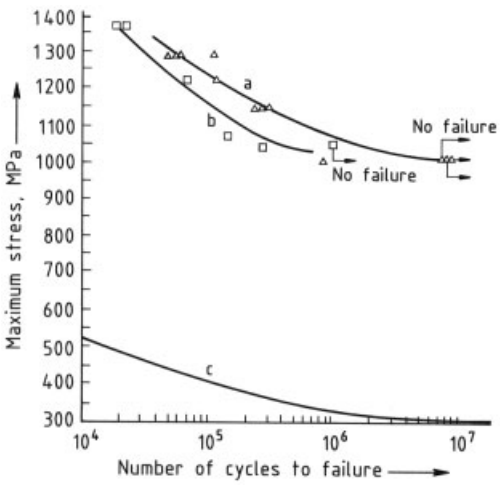

Figure 7. Fatigue strength of continuous-fiber-reinforced aluminum alloy (a, b) compared to a nonreinforced high-strength aluminum alloy (c) [31, a]
a) Al 6061 matrix reinforced with 48 vol % 12-ply SiC fibers, room temperature, minimum to maximum stress ratio = 0.1, maximum tensile strength 1420 MPa; b) Al 6061 matrix reinforced with 48 vol % 12-ply SiC fibers, 100 °C, minimum to maximum stress ratio = 0.1, maximum tensile strength 1517 MPa; c) AlZnMgCu1.5F53 (7075T73), minimum to maximum stress ratio = 0

3.4. Mechanical Properties

One of the most important advantages of metal-matrix composites compared to polymer-matrix composites is their higher thermal stability. Whereas the latter undergo severe softening at 120 – 250 °C, especially in humid atmospheres, metal-matrix composites retain their mechanical strength up to ca. 400 °C (Fig. 6). This is particularly marked for materials reinforced with continuous fibers (curves a and b). The late softening is due to reduction of the shear strength of the matrix which gradually eliminates transfer of load from broken to intact fibers, and transfer back to the intact parts of the broken fibers. The effective number of load-carrying fibers therefore decreases with increasing temperature.

Although enhanced thermal stability is less pronounced in metal-matrix composites reinforced with whiskers (Fig. 6, curve c), distinct reinforcement occurs.

Resistance to mechanical fatigue is also drastically enhanced by reinforcement with continuous fibers (Fig. 7). Due to their high strength the fibers act as crack stoppers and dividers (by delaminating the interface) so that the crack energy is prevented from concentrating into one single sharp crack. This also applies to a lesser degree for metal-matrix composites reinforced with short fibers.

4. Interactions between the Metal Matrix and Reinforcing Material

4.1. Wetting and Chemical Reaction

Wetting. The effective strength of short fibers obeys Equations (4) and (5) only when the tensile (bonding) strength and shear strength of the fiber – matrix boundary are higher than the flow stress of the matrix. If this is not the case, then these lower strengths must be taken into account and lead to significantly lower effective fiber strength and thus lower composite strength (Eq. 1). The necessary high bonding strength depends upon a low boundary surface energy, i.e., the fibers or particles must be wetted by the liquid matrix alloy. Wettability is measured by the sessile drop technique in which a droplet of alloy smelt is placed on a plate of the fiber material in a high vacuum and the resulting contact angle θ of the droplet is measured [15]. The angle θ decreases with increasing wettability and approaches 0° for complete wettability. The wettability increases with increasing temperature and θ decreases [16]. For example, aluminum alloys at 900 – 950 °C exhibit a rapid decrease of θ,

i.e., wetting occurs. Wetting can be improved by using alloying additives to lower the surface tension of the alloy (e.g., bismuth, lead, and titanium in aluminum) [17].

Wetting is time dependent and follows an S-shaped time curve; a slow start is followed by a rapid rise and a slow approach to saturation for 100 % surface coverage [18].

Chemical Reaction. The bonding strength can generally be increased when the matrix reacts slightly with the reinforcing components to give strong chemical bonds. Thick brittle intermediate layers must not, however, be formed. Growth of these layers is diffusion controlled and thus exceptionally rapid in the critical starting phase. This can be controlled by adding the reacting alloying elements in small amounts. The bonding strength of alumina fibers in aluminum alloys, for example, can be significantly increased by adding small amounts of lithium or magnesium. During heat treatment, the additional element diffuses to the boundary surface and reacts partially with the fiber surface to form $MgO \cdot Al_2O_3$ (spinel) [3].

Chemical reaction between the fibers and a main alloying element can produce a brittle compound which damages the fiber; this is the case with boron and carbon fibers in aluminum alloys [19]. Damage can be controlled by coating the fiber surface to prevent diffusion. Coating can be performed by methods such as chemical or electrochemical deposition, sputtering, or chemical vapor deposition. During production of loading no significant fiber damage occurs [20]. In titanium alloys that are reinforced with silicon carbide fibers and used for high-temperature constructions, the silicon in the fibers reacts with the titanium to form brittle titanium silicide [21]. A thin diffusion-hindering layer (e.g., of titanium carbide or nitride) delays the reaction successfully and prevents damage during the high-temperature production phase. Auger electron spectroscopy has confirmed the presence of titanium, silicon, and nitrogen in the boundary surfaces.

4.2. Bonding and Composite Strength

In composite materials with continuous fibers in a ductile metal matrix, the bonding of the fibers (expressed by the tensile and shear strengths of the interface) has only a small influence on the strength of the composite under static load. This is because the fibers are much longer than the critical length necessary to ensure full force transmission (Eq. 6) even for low boundary surface shear strength. However the toughness of the composite (the resistance to crack advancement during static or dynamic loading) depends largely upon the boundary surface shear strength [22]. If a crack in the matrix is perpendicular to a fiber of high bonding strength ($\tau*_{\mathrm{M}} > \tilde{\tau}_{\mathrm{M}}$), the critical length l_{c} (Eq. 6) is smaller than the stress field of the crack tip, i.e., the fracture strength of the fiber is exceeded and the crack advances in a straight line through the fiber and the entire composite.

If the fiber has a low bonding strength, then l_{c} is larger than the stress field and the fiber does not break but delaminates. As in a brittle matrix material, the crack branches leaving the fiber intact to bridge and relieve cracks. Toughness is significantly increased. The static strength of the composite depends directly upon the shear strength at the boundary surface when the fiber strength or diameter varies widely [23]. The strength increases with $\tau*_{\mathrm{M}}$ and the more so the wider the variation of $\sigma*_{\mathrm{F}}$ or d_{F}. A further influence on the composite strength is the reaction layer at the interface which is formed during production or heat treatment. These irregular and thus notch-forming layers grow approximately following a $\sqrt{t}$ law. Above a certain thickness (i.e., reaction time), they can lower the composite strength, as shown in Figure 8 for titanium alloys with silicon carbide fibers [24]. This behavior can be quantitatively related to the increasing notch effect of the layer which leads to delamination of fibers with a low bonding strength and crack advancement in fibers with a high bonding strength [25].

In metal-matrix composites with short fibers or particles, a sufficiently high bonding strength is a basic requirement. This ensures that stresses are transmitted to the short fibers (Section 3.2) and that effective dislocation barriers are formed with particles (Section 3.3). Very high bonding strengths have been achieved in a tensile test on an AlMgSiCu alloy reinforced with silicon carbide particles [26]. The lower limit for the bonding strength at the beginning of delamination was 1690 MPa. This bonding strength re-

quires a boundary surface reaction with chemical bonding forces.

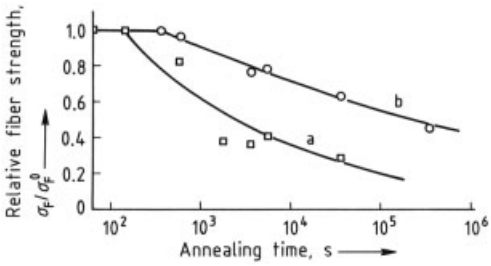

Figure 8. Fiber strength as a function of the annealing time (i.e., thickness of the reaction layer) for a titanium alloy matrix and fibers coated with boron (a) and boron – silicon carbide (b) [24]

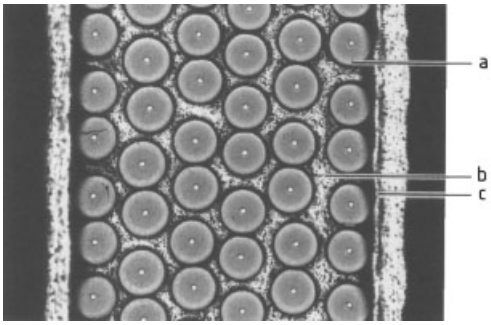

Figure 9. Hot-pressed, six-layer composite (aluminum alloy reinforced with silicon carbide fibers, ×60) [3]
a) 100 µm SiC fibers with a tungsten core; b) Intermediate AlMn1 layers clad on both sides with AlSi12 brazing alloy; c) Surface layers of AlMn1 clad only on the inner side with AlSi12

4.3. Chemical Compatibility

During long-term high-temperature treatment either during production or in use, the boundary surface reactions may proceed so far as to damage or even destroy the components. In order to calculate the thermodynamic stability combinations of different components and alloys, the enthalpies of formation of the used and possibly formed compounds should be known. This is, however, only the case for ionic bonds; data are not available for the many intermetallic compounds with covalent and metallic bonds. Empirical rules must be used to estimate the resistance and the reactivity of major combinations of components or elements, for example, as given in Table 3. In many cases the stability of the system can be increased by using thin boundary surface layers of a thermochemically stable compound (e.g., by coating fibers with tantalum, titanium carbide, or silicon nitride). A further possibility is to use reactive elements in the matrix that reduce the diffusion of the basic system. This diffusion controls the time dependency of the compatibility relationships. Thus, thermochemically unstable systems can be used for composite materials as long as the reactions occur slowly enough. The composite will then be sufficiently resistant to reactions at the temperatures and times encountered during production and use.

5. Production

5.1. Sheets and Tapes

Highly loaded spatial structures such as tubes or supports can be made from thin strips or tapes reinforced with continuous fibers. The strips or tapes may be brazed or diffusion bonded in autoclaves. They are produced by *plasma spraying* an aluminum alloy onto a single fiber layer on a thin aluminum support foil [28]. Alternatively *hot pressing* or *hot rolling* of fiber gauze between plated aluminum sheet can be used [3]; multilayered composites in unidirectional or other orientations can be produced by this technique (Fig. 9). A low-melting (eutectic) aluminum – 12 % silicon brazing alloy (*mp* 573 °C) is plated onto the higher-melting matrix material (an aluminum alloy such as AlMn1, *mp* 630 °C) by hot and cold rolling. This allows subsequent oxide-free bonding of the fibers and welding of the matrix sheets. Below the melting point of the braze, the silicon in excess of the eutectic diffuses from the brazing alloy into the matrix and ensures that localized melting does not occur and that the composite is fully heat stable [29].

Table 3. Chemical compatibility of metal-matrix materials and fiber materials* [27]

Matrix (T_m, °C; ϱ, g/cm^3)	Fiber (T_M, °C; ϱ, g/cm^3)*						
	B (2050; 2.3)	C (3800; 1.7)	Mo (2610; 10.2)	W (3410; 19.3)	SiC (2500; 3.2)	Si_3N_4 (1900; 3.2)	Al_2O_3 (2030; 4.0)	ZrO_2 (2580; 5.4)
Mg (650; 1.7)							R>900 °C	R
Al (660; 2.7)	R>700 °C	R>500 °C			R>700 °C			R
Ti (1670; 4.5)	R>900 °C				NR<700 °C		R>1400 °C	R>1400 °C
Co (1500; 8.9)	R	R>1000 °C					R	
Ni (1450; 8.8)	R>600 °C	R>1000 °C	R	R<1100 °C	R>1100 °C	R>1100 °C	NR	NR<1800 °C
W (3410; 19.3)	R	R						NR
Nb (2420; 8.4)	R	R					NR<1800 °C	NR

* R = reaction; NR = no reaction; T_m = melting or decomposition temperature.

In another continuous hot rolling technique, a layer of fiber gauze is preheated to braze temperature (590 °C) between plated aluminum tapes under an inert gas prior to entry in the rolling gap [30]. The excellent orientation and bonding of the fibers means that the fiber-reinforced tapes have a very high strength and modulus of elasticity (Table 4, row 7). For a fiber content of 50 vol % the tensile strength R_m is ca. 1350 MPa and the modulus of elasticity $E > 200$ GPa, i.e., they exceed the values of steel. Even higher values can be otained by smelting and casting techniques (Section 5.2, Table 4). The transverse strengths, R_m (cr.d.), are naturally much lower than the values in the longitudinal fiber direction, R_m (l.d.), but still considerably higher than those of fiber-reinforced plastics (20 – 40 MPa, see → Composite Materials).

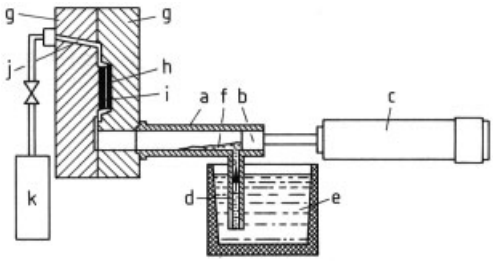

Figure 10. Production of metal-matrix composites by high-pressure melt infiltration with a vacuum die casting machine [33]
a) Filling chamber; b) Piston; c) Hydraulic cylinder; d) Suction pipe; e) Melt supply; f) Melt sucked into filling chamber; g) Die halves (closed); h) Form volume; i) Fiber preform; j) Vacuum channel; k) Vacuum system

5.2. Cast Parts

Starting material for fiber-reinforced semifinished goods and fiber-reinforced engineering components can be produced by high-pressure smelt infiltration methods, e.g., squeeze casting [32] or pressure die casting with fiber preforms. Injection with aluminum can, for example, occur very rapidly in pressure die casting carried out under vacuum in a cold chamber machine (Fig. 10). Wetting problems do not occur and the boundary surface reaction necessary for good bonding can be achieved during subsequent heat treatment. The fiber preforms for this technique are cut from porous sintered plates of short fibers (e.g., of Al_2O_3 fibers) or are made from fiber

Table 4. Mechanical properties[a] of aluminum and magnesium alloys reinforced with continuous fibers [31]

Metal matrix[b]	Fiber	Fiber content, vol %	Production route	E, GPa	A, %	R_m(l.d.), MPa	R_m(cr.d.), MPa	T_{max}, °C	Reference
AlMgSiCu (6061)	SiC – C (SCS)		plasma spraying and hot pressing (tapes)		0.9	1500	90	300	[31, a]
AlSi7Mg (A357)	SiC – C (SCS)	50	as cast	230	0.8	1700	105	300	[31, a]
AlLi2	Al_2O_3 (FP)	50	as cast	207	0.7	560	172	300	[31, b]
AlMgSiCu (6061)	SiC (Nicalon)	35	as cast	110		900	50	300	[31, c]
Al alloy	C	<50	as cast	210	0.5	700	50	300	[31, a]
AlCuMg (2024)	Al_2O_3 (Safimax)	50	as cast	185		1000		300	[31, d]
AlMn1 (braze plated)	SiC (SIGMA)	50[c]	hot rolled (tapes)	200	ca. 0.5	1350	ca. 120	300	Vereinigte Aluminium-werke
Mg	Al_2O_3	50	as cast	150		800		300	[31, d]
MgZn4RE 1.2Zr (ZE41)	SiC – C (SCS)	50	as cast	230	0.8	1350		300	[31, a]
AlZnMgCu (7075)	Aramid	ca. 10 (1 layer in 1.3 mm[d])	bonding with epoxy resin	68.3	1.0	775	420	160	[31, e]

[a] E = modulus of elasticity; A = plastic strain at fracture; R_m(l.d.) = tensile strength in longitudinal direction; R_m(cr.d.) = tensile strength in cross (lateral) direction; T_{max} = maximum use temperature.
[b] Aluminum alloy designations are in accordance with the recommendations of the Aluminum Association, Washington, D.C.
[c] Recalculated to 50 vol % fibers with equal efficiency of fiber properties.
[d] ARALL 1.

yarn or fabric that is bonded together with small additions of an inorganic binder and then sintered.

Fiber preforms (e.g., of FP fibers) for connecting rods are made from belts of continuous fibers for the shaft and the eyes, and short-fiber inserts for the stiffening of the shaft (Fig. 11 A) [3]. This and other techniques have been used to produce fiber-reinforced connecting rods from an ageable aluminum casting alloy AlSi10Mg (Fig. 11). Bonding of the fibers in the load-bearing section of the matrix is very good and they are evenly spaced, as has been confirmed by metallographic and microprobe investigations. Fiber reinforcement produces a significant increase in strength; the buckling strength for a connecting rod without reinforcement is 57 kN, reinforcing with 22 vol % alumina fibers raises the value to 149 kN [3]. The fatigue strength is, however, unsatisfactory probably due to poor fiber orientation. Weight saving as compared to steel is about 40 % and this factor also reduces vibration in the engine.

5.3. Semifinished Products

In the case of semifinished products which are to be subjected to hot deformation and are made from fiber-reinforced light metals (e.g., aluminum alloys), the starting material can be produced as described in Section 5.2 or by stirring or blowing the short fibers into the smelt [30]. Such material can be extruded to give simple shapes with a thin shell of fiber-free material. The fibers are orientated in the longitudinal direction so that virtually unidirectional reinforcement is achieved (Fig. 12) [30]. In conventional extrusion the fibers become fractured as a result of the high deformation forces. Fibers are thus shorter than the critical length (Eq. 6) and

full reinforcement is therefore impossible. This problem can only be solved using appropriate temperature control to extrude the alloy in a semiliquid state where the deformation forces are much lower [34]. Inspite of this handicap, the hot strength can be significantly improved. An aluminum – magnesium alloy with 16 vol % alumina has a strength of 120 MPa at 300 °C as compared to 60 MPa for a conventional piston alloy AlSi12CuMgNi [30].

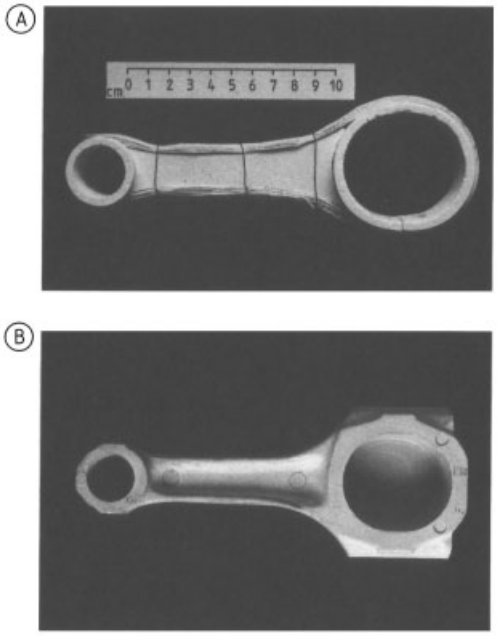

Figure 11. Fiber preform of FP (alumina) fibers (A) for a diecast, fiber-reinforced aluminum motor connection rod (B) [3]

The relatively low reinforcement with short fibers can be explained with the rule of mixing (Eqs. 1, 4, and 5). If the effective short fiber strength is used for $\sigma*_{\mathrm{F}}$ in Figure 3, then according to Equation (5) the increase in strength with increasing fiber content is much less than for longer fibers with $l > l_{\mathrm{c}}$. A correspondingly higher critical fiber content is required to produce reinforcement.

Since the short fibers are heavily fractured in conventional extrusion and other hardening mechanisms become effective in the heterogeneous structure (cf. Section 3.3), many experiments have been undertaken to reinforce aluminum alloys with ceramic particles (e.g., silicon carbide and alumina) [35]. The particles are either injected into the smelt or mixed with the matrix powder and subsequently processed by powder metallurgy. With a particle content of 10 – 40 vol %, strengths are increased by up to 50 % and the modulus of elasticity by up to 100 %; wear resistance is also improved (Table 5).

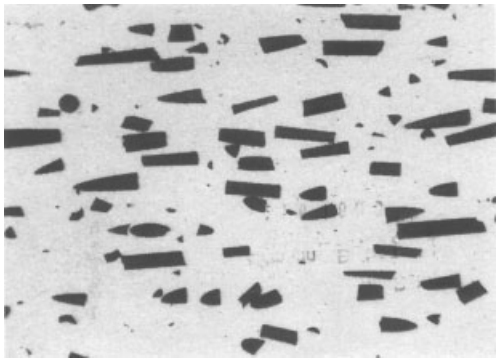

Figure 12. Longitudinal section of hot extrusion rod composed of aluminum reinforced with short FP (alumina) fibers (×120) [30]

If the theoretical portion of the short fiber or particle reinforcement is subtracted from the composite strength, a higher strength than that of the pure matrix is often obtained [13]. This induced hardening of the matrix according to Equation (10) has been determined to be ca. 60 MPa for an aluminum – magnesium alloy reinforced with short fibers. For material reinforced with extremely fine particles or whiskers this additional hardening effect is much higher because the contributions $\Delta\tilde{\sigma}_{\mathrm{M}}$ (P) and $\Delta\tilde{\sigma}_{\mathrm{M}}$ (HP) in Equation (10) are the more significant the smaller the particles and the shorter the distances between them at constant volume fraction.

5.4. Hybrid Composites

In hybrid composites laminates of epoxy resins reinforced with a monolayer of unidirectional aramid fibers are sandwiched between thin, high-strength aluminum alloy tapes (ARALL, Fig. 13) [37]. The composite can consist of one or several aramid-fiber-reinforced plastic layers. The aramid fibers have a higher modulus of elasticity than aluminum, a much higher strength, and a small but significant fracture elongation. The fibers do not greatly improve the strength,

Table 5. Mechanical properties[a] of aluminum – magnesium alloys[b] reinforced with particles [36]

Matrix	Particles	Particle content, vol %	Production route[c]	E, GPa	A, %	R_m(l.d.), MPa	R_m(cr.d.), MPa	K_{IC}, N/mm$^{1.5}$	α, 10^{-6} K^{-1}
AlCuMg (2024)		0	IM and extruded	72	13	490		780	24
AlCuMg (2024)	SiC	10	IM and extruded	83	2.5	490		560	16
AlCuMg (2024)	SiC	20	IM and extruded	108	2.5	500		600	12
AlCuMg (2024)	SiC	15	IM and extruded	96	3.2	502	474		14
AlZnMgCu (7075-T6)		0	IM and extruded	71	11	580		780	24
AlZnMgCu (7075-T6)	SiC	15	IM and extruded	100	2.5	610		630	14
AlMgSiCu (6061-T6)	SiC	15	IM and extruded	96	3.0	355	363		14
AlMgSiCu (6061-T6)	SiC	15	PM and extruded	98	7.5	460			15
AlMgSiCu (6061-T6)	SiC	25	PM and extruded	116	3.5	540			13
AlMgSiCu (6061-T6)	SiC	40	PM and extruded	147	1.75	560			11
AlZnMgCu (7090)	SiC	25	PM and extruded	117	2.0	805			13
AlZnMgCu (7090)	SiC	40	PM and extruded	147	0.9	720			11

[a] E = modulus of eleasticity; A = plastic strain at fracture; R_m(l.d.) = tensile strength in longitudinal direction; R_m(cr.d.) = tensile strength in cross direction; K_{IC} = fracture toughness; α = coefficient of thermal expansion; Densities range from 2.7 to 3.3 g/cm^3 depending on type and amount of powder (cf. Table 2).

[b] Aluminum alloy designations are in accordance with the recommendations of the Aluminum Association, Washington, D.C.

[c] IM = ingot metallurgy; PM = powder metallurgy.

but dramatically increase fatigue strength parallel to the fiber direction. This is due to bridging and relieving of the crack tip by the undamaged fibers [38]. This mechanism works optimally when the fibers are not too firmly bonded in the plastic. The critical length of the fiber is larger than the size of the crack stress field. For biaxial loading the fiber layers can be crossplied.

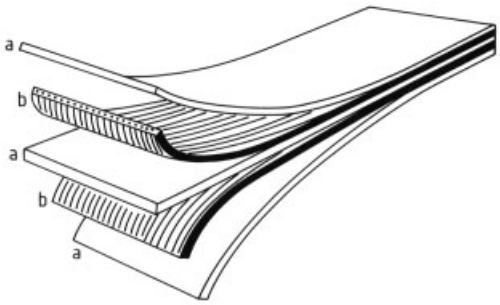

Figure 13. Schematic structure of a 3/2 ARALL-2 laminate composed of three aluminum alloy layers (a), and two epoxy layers reinforced with aramid fibers (b) [38]

6. Working and Machining

Cast billets or sections of thick, extruded billets are die forgeable (Fig. 14). Fiber fracture is much lower than in small extruded cross-sections because the degree of deformation and the resulting stresses are much lower.

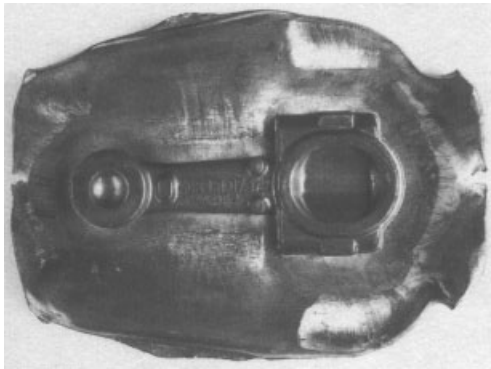

Figure 14. Forged compressor connecting rod made of aluminum alloy reinforced with short FP fibers [3]

Machining of metal-matrix composites (e.g., aluminum alloys reinforced with long and short alumina fibers) is possible with hard-metal cutting tools [39]. The life time of the tools can be raised by a factor of 100 when polycrystalline diamond is used as the cutting material. Surface roughness and dimensional stability are also improved due to the chip breaking effect of the fibers and the lack of pick-up on the cutting edge [39]. Other techniques such as grinding, turning, and milling with hard-metal tools produce similarly good results [3].

Table 6. Potential and realized technical applications of metal-matrix composites [11]

Application	Properties	Systems (reinforcement – metal matrix)	Production methods
Aerospace Structures			
Frames, stiffeners, antennas, fittings	high specific strength and stiffness, low coefficient of thermal expansion, high heat resistance	B – Al, SiC – Al, C – Al, C – Mg	melt infiltration (castings), hot pressing (tapes and sheets), diffusion and braze bonding (spatial structures)
Aircraft Structures			
Pylons, wing boxes, stiffeners, fan and compressor blades, transmission houses	high specific strength and stiffness, heat resistance, impact resistance	B – Al, SiC – Al, C – Al, SiC – Ti, Al_2O_3 – Al, Al_2O_3 – Mg	melt infiltration, hot pressing, diffusion and braze bonding
Turbine blades	high specific strength and stiffness, high heat and impact resistance	W – superalloy, directionally solidified eutectics, e.g., Ni_3Al – Ni_3Nb – Ni	melt infiltration, directional solidification of near net shape parts
Automotive Parts			
Frames, struts, connecting rods, pistons, piston pins, valve spring supports	high specific strength and stiffness, heat resistance, low coefficient of thermal expansion	SiC – Al, Al_2O_3 – Al, Al_2O_3 – Mg	melt infiltration, hot extrusion, forging
Battery plates	high stiffness	C – Pb, Al_2O_3 – Pb	melt infiltration
Electrical Equipment			
Motor brushes	high electrical conductivity, wear resistance	C – Cu	melt infiltration
Electrical contacts	high electrical conductivity, heat and corrosion resistance	C – Cu, Al_2O_3 – Al	melt infiltration, hot extrusion, forging
Technical Equipment			
Motorcycle and bicycle frames	high specific strength and stiffness, corrosion resistance	B – Al, SiC – Al, SiC – Mg	melt infiltration, hot pressing, braze bonding
Bearings	high specific strength and stiffness, wear resistance	C – Pb	melt infiltration, hot forming
Shuttles in textile machinery	high specific strength and stiffness, wear resistance	B – Al, SiC – Al, C – Al	melt infiltration, hot pressing, forging

7. Uses

Some uses of metal-matrix composites have already been mentioned in Chapter 5. Sheet and strip reinforced with continuous fibers have excellent properties and low densities (cf. Table 4). However, the production costs and fiber costs are presently so high that their use is limited to aerospace applications where the savings in weight and fuel consumption compensate the high production costs. An example here is the supporting structure of the U.S. Space Shuttle, made from aluminum tubes reinforced with silicon carbide fibers [40].

The aluminum sheets with a hybrid structure (ARALL) have lower production costs. They are being tested in aerospace applications; in cell constructions (e.g., in the lower side of an aircraft wing) where tensile loading is high [41]. Pressure die casting techniques using continuous and short fiber reinforcement of aluminum alloys are more economical methods of producing components. Various motor parts are presently being developed [3], [42] and some are already being produced and used (e.g., pistons, connecting rods, piston pins, and valve spring plates) [43]. Particle reinforcement of aluminum alloys is the cheapest method of production. Particle-reinforced components are being tested where high stiffness, hot strength, and especially good wear resistance are important (Table 6).

8. Occupational Health

During production and working with thin, inorganic (especially ceramic) fibers, fragments, and fine fibers, debris can escape into the atmosphere. As well as being a general inconvenience, this dust can be detrimental to health if inhaled. Investigations have shown that susceptibility to cancer depends upon the dimensions of the fibers. In cell culture tests very thin silicon carbide whiskers (diameter 0.1 – 1 μm) exhibit a similar toxicity as asbestos fibers (crocidolite) [44]. Tests using thicker alumina fibers (ca.

3 μm) show no significant shortening of cell life time. The same results also apply to equiaxial silicon carbide particles with similar thicknesses to the whiskers. They are only toxic when they have a large length to thickness ratio. The same protective measures must be used as for asbestos to prevent inhalation [45] (→ Asbestos, Chap. 9.

9. References

1. A. Kelly, N. H. Macmillan: *Strong Solids,* 3rd ed., Clarendon Press, Oxford 1986.
2. A. A. Griffith, *Phil. Trans. Soc. London A* **221** (1920) 163.
3. G. Ibe, J. Penkava, *Metall (Berlin)* **41** (1987) 590.
4. M. Taya, R. J. Arsenault: *Metal Matrix Composites – Thermomechanical Behavior,* Pergamon Press, Oxford – New York 1989.
5. M. R. Pigott: *Load Bearing Composites,* Pergamon Press, Oxford – New York 1987.
6. A. Kelly (ed.): *Concise Encyclopedia of Composite Materials,* Pergamon Press, Oxford – New York 1989.
7. *Aluminium-Taschenbuch,* 13th ed., Aluminium-Verlag, Düsseldorf 1987, p. 228.
8. U. Zwicker: *Titan und Titanlegierungen,* Springer Verlag, Berlin 1974, p. 283.
9. C. T. Lynch (ed.): *Metals, Composites, and Refractory Materials, CRC Handbook of Materials Science,* vol. II, CRC Press Inc., Boca Raton, FL, 1984.
10. Producer information: a) ICI, Chemicals & Polymers Group, Runcorn, Cheshire, England; b) Sohio Engineered Materials Co., Niagara Falls, USA; c) Du Pont, Textile Fibers Department, Wilmington, USA; d) Nippon Carbon Ltd., Tokyo, Japan; e) Elektroschmelzwerk Kempten, Kempten, FRG.
11. J. W. Weeton, D. M. Peters, K. L. Thomas: *Engineer's Guide to Composite Materials,* Sections 2 and 5, American Society for Metals, Metals Park, OH, 1987.
12. A. Kelly, W. R. Tyson, *J. Mech. Phys. Solids* **13** (1965) 329.
13. G. Ibe in G. Ondracek (ed.): *Verbundwerkstoffe und Stoffverbunde in Technik und Medizin,* **vol. 1,** DGM-Informationsgesellschaft, Oberursel 1989, p. 145.
14. P. M. Kelly, *Int. Met. Rev.* **18** (1973) 31.
15. L. J. Ebert, P. K. Wright in A. G. Metcalfe (ed.): *Interfaces in Metal Matrix Composites,* Academic Press, New York 1974, p. 31.
16. R. Warren, C. H. Anderson, *Composites (Guildford UK)* **15** (1984) 101.
17. G. L. Ranshofen, *Aluminium (Düsseldorf)* **49** (1973) 231.
18. T. Choh, T. Oki, *Mater. Sci. Technol.* **3** (1987) 378.
19. S. Ochiai, K. Osamura, Y. Murakami, *Z. Metallkd.* **74** (1983) 68.
20. T. Erturk, V. Gupta, A. S. Argon, J. A. Cornie in F. L. Matthews et al. (eds.): *Sixth International Conference on Composite Materials,* **vol. 2,** Elsevier Applied Science, London – New York 1987, p. 2.20.
21. R. Leucht, H. J. Dudek, G. Ziegler, *Z. Werkstofftech.* **18** (1987) 27.
22. M. Gouda, K. M. Prewo, A. J. McEvily: "Fatigue of Fibrous Composite Materials," *ASTM Spec. Tech. Publ.* **723** (1981) 101.
23. S. Ochiai, K. Osamura, *J. Mater. Sci.* **23** (1988) 886.
24. S. Ochiai, K. Osamura, Y. Murakami, *Z. Metallkd.* **75** (1984) 238.
25. S. Ochiai, K. Osamura, *Metall. Trans. A* **18 A** (1987) 673.
26. Y. Flom, R. J. Arsenault, *Mater. Sci. Eng.* **77** (1986) 191.
27. R. Taprogge, R. Scharwächter, P. Hahnel: *Faserverstärkte Hochleistungs-Verbundwerkstoffe,* Vogel-Verlag, Würzburg 1975, p. 76.
28. K. G. Kreider, K. M. Prewo in K. G. Kreider (ed.): *Metallic Matrix Composites,* Academic Press, New York 1974, p. 400.
29. G. Ibe, W. Gruhl, H. Gebhardt in W. Bunk, J. Hansen (eds.): *Faserverbundwerkstoffe,* **vol. 1,** Springer Verlag, Berlin 1980, p. 27.
30. G. Ibe, J. Penkava in J. Hansen (ed.): *Faserverbundwerkstoffe,* **vol. 2,** Springer Verlag, Berlin 1985, p. 77.
31. Producer information: a) AVCO Speciality Materials Division, Lowell, USA; b) Du Pont, Textile Fibers Dept., Wilmington, USA; c) Nippon Carbon Ltd., Tokyo, Japan; d) ICI, Chemical and Polymers Group, Runcorn, Cheshire, England; e) Alcoa Aluminum Corp., Pittsburgh, USA.
32. T. W. Clyne, J. F. Mason, *Metall. Trans. A* **18 A** (1987) 1519.
33. G. Ibe, J. Penkava, H. Woithe, DE-A 370 1218.5, 1984.
34. F. Girot et al. in F. L. Matthews et al. (eds.): *Sixth International Conference on Composite Materials,* **vol. 2,** Elsevier Applied Science, London – New York 1987, p. 2.330.

35. S. V. Nair, J. K. Tien, R. C. Bates, *Int. Met. Rev.* **30** (1985) 275.
36. Producer information: DWA Composite Specialities Inc., Chatsworth, CA; DACC Dural Aluminium Composites Corp., San Diego, CA.
37. J. Schijve, L. B. Vogelsang: *Development of ARALL, a New Material for Aircraft Structures,* Memorandum M-434, Delft University of Technology, Dept. of Aerospace Engineering, Delft 1982.
38. R. O. Ritchie, Y. Weikang, R. J. Bucci, *Eng. Fract. Mech.* **32** (1989) 361.
39. L. Cronjäger, P. Müller, *Z. Werkstofftech.* **18** (1987) 205.
40. R. B. Aronson, *Mach. Des.,* August 1985.
41. J. W. Gunnink, P. A. van der Schee, *Compos. Struct. (Proc. Int. Conf.)* **4, vol. 1,** (1987) 1.162.
42. G. Ibe et al. in B. Vierkorn – Rudolph, D. Lillack (eds.): *Symposium Materialforschung 1988,* **vol. 2,** Projektleitung Material- und Rohstofforschung der Kernforschungsanlage Jülich, 1988, p. 1099.
43. J. Donomoto et al., EP 0 080 562B1, (Priority JP 191918/81).
44. J. D. Birchall et al., *J. Mater. Sci. Lett.* **7** (1988) 350.
45. *Protection of Employees against Health Risks from Asbestos at the Working Place,* Directives of the Council of the European Community 83/477/EEC, Sept. 19th, 1983.

Metals, Cladding and Lining → Corrosion

Metals, Surface Treatment

Helmut Kunst, Consultant Engineer, Osterholz-Scharmbeck, Federal Republic of Germany (Chap. 1 and 4)

Brigitte Haase, Hochschule Bremerhaven, Bremerhaven, Federal Republic of Germany

James C. Malloy, Kolene Corporation, Detroit, Michigan, United States (Chap. 2.2)

Klaus Wittel, Chemetall GmbH, Frankfurt am Main, Federal Republic of Germany (Chap. 3)

Montia C. Nestler, Sulzer Metco (Deutschland) GmbH, Hattersheim, Federal Republic of Germany (Chap. 5)

Andrew R. Nicoll, Sulzer Metco (Deutschland) GmbH, Hattersheim, Federal Republic of Germany (Chap. 5)

Ulrich Erning, Sulzer Metco (Deutschland) GmbH, Hattersheim, Federal Republic of Germany (Chap. 5)

Gerhard Rauscher, Forschungsinstitut für Edelmetalle und Metallchemie, Schwäbisch Gmünd, Federal Republic of Germany (Chap. 6)

1. Introduction 2
2. Cleaning 2
2.1. Cleaning in Aqueous and Nonaqueous Solvents 3
2.1.1. Cleaning and Degreasing in Aqueous Media 3
2.1.2. Cleaning Agents Based on Organic Solvents 5
2.1.3. Cleaning Process Technology 6
2.2. Cleaning and Pickling with Salt Baths 7
2.2.1. Bath Composition and Properties . 7
2.2.2. Descaling of Metals 8
2.2.3. Casting Cleaning 10
2.2.4. Coatings Removal 12
3. Phosphating 13
3.1. Chemistry 13
3.2. Methods of Application 16
3.3. Uses of Phosphate Coatings 16
4. Case Hardening of Ferrous Metals 17
4.1. Case Hardening without Chemical Reaction via Phase Boundary . . . 18
4.1.1. Flame Hardening 18
4.1.2. Induction Hardening 19
4.1.3. Laser Hardening 20
4.1.4. Electron-Beam Hardening 20
4.2. Thermochemical Treatment 21
4.2.1. Gas Carburizing 21
4.2.1.1. Principles 21
4.2.1.2. Properties of Carburized Components 22
4.2.1.3. Carburizing Methods 22
4.2.2. Nitriding and Nitrocarburizing . . . 25
4.2.2.1. Principles 25
4.2.2.2. Properties of Nitrided and Nitrocarburized Components 26
4.2.2.3. Nitriding and Nitrocarburizing Methods 27
4.2.3. Boriding 32
4.2.3.1. Principles 32
4.2.3.2. Boriding Methods 33
4.2.4. Other Thermochemical Processes . 33
5. Thermal Spraying 34
5.1. Methods of Thermal Spraying . . 34
5.1.1. Flame Spraying 34
5.1.2. Electric Arc Spraying 35
5.1.3. Plasma Spraying 35
5.1.4. High-Velocity Oxy–Fuel Spraying (HVOF) 37
5.2. Properties of Thermally Sprayed Coatings 38
5.3. Materials 39
5.3.1. Spray Coating Materials 39
5.3.2. Substrate Materials 40
5.4. Applications 41
6. Coloring 42
7. Acknowledgement 42
8. References 42

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_403.pub2

1. Introduction

This article describes methods for the treatment of metal surfaces. It is only recently that metal surfaces have attracted the attention of theoreticians and those concerned with practical applications. Metal surfaces are subject to both corrosion and mechanical wear. Bending causes the formation of surface cracks. Surface treatment of metals has an aesthetic value; decorative aspects play an important role.

Chapter 2 describes procedures for cleaning metal surfaces. Cleaning is increasingly important, not only as a finishing process for parts and components with decorative functions, but surface cleanliness is more and more accepted as a property to be maintained and controlled throughout the manufacturing line. At the same time European and national law tends to restrict the use of chemicals, including components of cleaning agents, not only with respect to nonaqueous solvents (volatile organic solvents, VOC), inflammable or explosive compounds, halogenated hydrocarbons, biocides as present in preservatives, etc., but also by making the process of registration and permitting of chemical compounds more difficult and expensive.

Efficient cleaning can determine the success of subsequent treatments (e.g., electroplating, thermochemical and plasma-assisted treatments). In many areas of manufacturing, like automotive, aircraft, and medical implants, cleanliness is an essential quality of both product and production line, because the technological properties of the final product depend on it. The most recent development is the definition of technical cleanliness and instructions for its measurement and evaluation [1].

Chapter 3 deals with phosphating which is finding increasing application. Chapter 4 is concerned with the case hardening of ferrous materials. The procedures described improve surface mechanical properties such as hardness and wear resistance, as well as flexural strength, fatigue strength, endurance limits, and corrosion resistance of the component. New developments in case hardening include increasing automatization and the use of laser, electron-beam, and plasma techniques.

Chapter 5 is concerned with spraying methods and spraying materials that are used to solve special problems, especially those caused by wear. Chapter 6 discusses the coloring of metals, which is used almost exclusively for decorative purposes.

Surface treatment of metals is a rapidly developing field and involves an ever-increasing variety of methods and processes, which include high-temperature processes like chemical vapor deposition (CVD, at ca. 1000 °C) and physical vapor deposition methods, mostly plasma-assisted, at lower temperatures (PVD, 400 – 500 °C). Developments in methodology and coating materials are also being made in PVD. Titanium nitride, titanium carbide, and titanium aluminum nitride are supplemented by other materials that are still being tested in the laboratory, for example, diamond-like carbon. CVD and PVD are no longer exclusively applied to metal-working tools, but also to structural components and medical implants. These techniques are described under → Thin Films. Ion implantation is used, among other applications, for producing thin, wear-resistant and/or corrosion-resistant layers.

Prevention of corrosion by surface treatment, including cladding and lining, is discussed in → Corrosion. The technique of electroplating is treated in → Electrochemical and Chemical Deposition. One application is the formation of corrosion-resistant barrier layers, but, depending on the properties of the coating, electroplated layers can also reduce abrasive wear, improve running properties, and have a decorative effect. Finally, the treatment of metal surfaces in its broadest sense includes application of paints and coatings (→ Paints and Coatings).

2. Cleaning

Surface cleaning is the removal of unwanted components, particles, and films. During manufacturing, workpieces and parts come into contact with a number of chemicals (lubricants, greases, anticorrosives, coolants, drawing or cutting oils, emulsions, etchants, deburring fluids, quenching agents, etc.). Machining is responsible for chips and cuttings; storage and transport lead to general soiling. Surface contaminants are less and less tolerated and must be removed. Although the binding forces between the substrate and an adhering particle or film are much lower than those present in a true chemical bond, en-

ergy input is needed to overcome the substrate – contaminant bond.

This energy input can be subdivided into three contributions: chemical, thermal, and mechanical energy. Thermal and mechanical energy are provided by the cleaning installation. The cleaning agent makes the chemical contribution. Conventional cleaning agents used in metal surface cleaning are aqueous or nonaqueous solvents.

Nonconventional cleaning processes apply carbon dioxide in the liquid (supercritical) or solid state with promising results. As the cleaning effect of the conventional agents is limited to loosely adhering contaminants and particles, a higher energy input is sometimes needed to remove reaction films and layers or even the first few nano- or micrometers of substrate. Then processes such as blasting, chemical or electrochemical etching/pickling, deburring, electrochemical machining (ECM), and plasma etching are applied.

Cleaning processes and agents, including aspects of environmental hazards and occupational health and safety, are in treated detail in databases, e.g., CLEANTOOL (http://www.cleantool.org), SAGE (http://clean.rti.org/), BAUTEILREINIGUNG (http://193.158.124.195/bauteile/portal/portal.-php), TURI (http://www.cleanersolutions.org/), SolventCentral (http://www.solventcentral.com/), and EPA Solvents Database http://solvdb.ncms.-org/).

2.1. Cleaning in Aqueous and Nonaqueous Solvents

2.1.1. Cleaning and Degreasing in Aqueous Media

As most soil and surface contaminants are insoluble in water, pure water is not an appropriate cleaning agent in metal surface cleaning. Additives can be distinguished by function and chemical composition. Classes of additives or components in aqueous cleaners are:

- *Builders*: inorganic salts of weak acids, which increase the pH of the solution responsible for removal of solid insoluble particles such as cuttings, chips, and dust.
- *Surfactants* or *detergents*, subdivided into anionic, nonionic, and cationic surfactants, are organic molecules containing both hydrophobic and hydrophilic groups that enable the formation of micelles in aqueous solution and the formation of films, either on the surface to be cleaned or at the oil/water interface of an insoluble oil droplet, thus modifying surface tension and facilitating the removal of oil film and the formation of an emulsion, owing to the stabilization of small oil droplets in the aqueous phase. Whereas cationic surfactants are generally not used in metal cleaning, anionic surfactants are represented by alkyl aryl sulfonates, and nonionic surfactants by polyaddition products of fatty alcohols and ethylene oxide (polyethoxylates or polyglycols), with tailored properties (e.g., critical micelle formation temperature, turbidity point, and antifoaming behavior) depending on polyethoxylate chain length [2].
- *Complexing agents* increase the solubility of metal cations and reduce the problem of precipitation of sparingly soluble metal compounds, including metal soaps and water hardness.
- *Anticorrosives* are salts of organic acids and amino compounds that protect both cleaning installation and the cleaned metal parts from corrosion.
- *Buffers* maintain constant pH even if there is an input of acids to the cleaning bath.
- *Preservatives*: the increasing biodegradability of cleaning agent components leads to problems with microbial fouling, especially when the temperature of the cleaning bath is reduced. The use of preservatives, however, is being increasingly restricted by European laws and regulations.

The preferred components in aqueous cleaners are multifunctional: phosphates, for example, are known for their excellent capability to stabilize a dispersion of solid particles, have pH buffer properties, and form complexes with metal cations; amines prevent corrosion and can act as preservatives; nonionic surfactants help to remove an oil film from the metal surface and protect the cleaned surface from corrosion. Builders and surfactants which exhibit a significant synergistic effect are preferred (Fig. 1) [2].

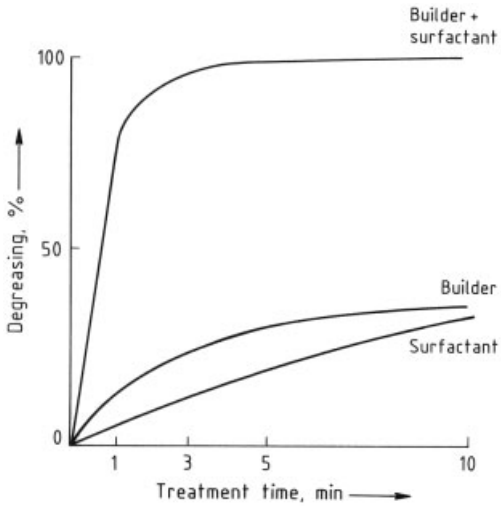

Figure 1. Synergistic action of builder (2 %) and surfactant (0.1 %) in degreasing [3]

Products normally contain mixtures of surfactants and builders, because no single representative can provide all of the desired properties in the complete formulation alone, i.e., foaming characteristics, emulsifying or de-emulsifying behavior, wetting properties, chemical stability, and cleaning action. Surfactants present in lubricant greases can pass into the cleaning bath, and this can have both beneficial and adverse consequences. Adverse effects include undesired foam formation in spraying operations or emulsification of the added lubricant. Beneficial effects include reduction of the treatment time or increased useful life of the bath. Surfactants must be biologically degradable in accordance with national regulations.

The so-called surfactant mechanism, with or without support of a builder, is based on the adsorption of surfactant and builder anions on the substrate, as well as on solid contaminants (builder anions) and hydrophobic films (surfactant molecules or anions). This leads to the modification of surface charge because of builder adsorption and to modification of surface energy or tension with respect to the oil film. Both effects weaken adhesive forces, so that the contaminants can be removed by mechanical energy, as in stirring, splashing, or ultrasound-assisted cleaning. Both builders and surfactants replace contaminants by adsorbed films of their own kind, by a mechanism significantly different from that of solvents. As all builders and many surfactants are nonvolatile and cannot evaporate (unlike solvents including water), they remain on the substrate surface after drying. The only method to avoid spots and films of cleaning agent residues is to rinse the parts diligently in deionized water (which may contain an anticorrosive).

Based on their behavior toward oils and lubricants, cleaners are classified as emulsifying or de-emulsifying, depending on whether a stable emulsion formed with the input oil or not. The de-emulsifying cleaners offer the opportunity to easily remove the input oil from the cleaning bath surface by simple phase separation.

Commercial products can be distinguished by their application pH as strongly alkaline (pH > 12), alkaline (9 < pH < 12), neutral (6 < pH < 9), or weakly acidic (pH < 6); see Table 1. From alkaline to neutral products, cleaning effectiveness decreases with decreasing pH, but neutral agents are increasingly popular owing to simpler storage and handling conditions with regard to environmental and health/safety regulations and less corrosion risk.

Table 1. Classification of aqueous cleaning agents for metals

Type of aqueous cleaning agent	Composition	Use
Strongly alkaline	alkali metal hydroxides, silicates, phosphates, complexing agents, surfactants	heavy-duty cleaning, removal of severe soiling, electrolytic degreasing, saponification of ester oils, particle removal
Alkaline	phosphates, borates, complexing agents, carbonates, surfactants	machined parts, removal of chips and cuttings, cutting oils and coolants/lubricants; precleaning
Neutral	surfactants, corrosion inhibitors, polyphosphates	removal of coolants/lubricants, anticorrosives etc., in rinsing baths for increased corrosion protection
Weakly acidic	acidic salts (hydrogenphosphates), surfactants	cleaning and phosphating (without coating)

Alkaline Cleaning Agents. Alkaline cleaning agents consist of a mixture of inorganic salts (called builders) and surfactants. The builders — alkali metal hydroxides, silicates, orthophosphates, condensed phosphates, borates, carbon-

ates — control the pH of the solution and intensify the action of the surfactants. They remove and help to disperse and emulsify contaminant particles, metal soaps, lubricants, oils, and greases.

What type of cleaning agent is selected depends on the substrate and on the type of surface contaminant. Alkali metal hydroxides and carbonates are used when fats and oils are to be saponified or if some form of chemical attack on the substrate (e.g., aluminum or zinc) is required. All electrolytic cleaners contain a high percentage of alkali to provide high electrolyte conductivity. Silicates have a high soil-removing capacity; they are suitable for removal of pigments and are preferred for difficult cleaning jobs. During cleaning, a thin layer of silicate is strongly adsorbed on steel, aluminum, and zinc; in the cases of aluminum and zinc this inhibits further attack by the alkaline solution [4]. Phosphates are typical multipurpose components; they have good soil-removing properties and reduce water hardness, aided by other complexing agents. Cleaning agents based on borates are used for aluminum and zinc-coated surfaces.

Neutral Cleaning Agents. Neutral cleaning agents usually do not contain inorganic components (builders). The formulations are based on a combination of surfactants (mainly nonionic), corrosion inhibitors for iron and steel based on fatty acids and amines, and inhibitors for nonferrous and light metals (triazide compounds). The pH of the aqueous solution is between 6 and 9.

Neutral cleaners have only mild cleaning properties and are not appropriate for heavy-duty cleaning, but have two advantages: 1) they are not prone to form low-volatility surface layers on the cleaned surfaces, and 2) they establish temporary corrosion protection if applied in the final bath of the cleaning installation, owing to a thin residual adsorbed film. Consequently, they can be used in simple single-chamber or single-step cleaning operations which do not provide a final rinsing stage; however, their cleaning effectiveness is limited.

Weakly Acidic (Passivating) Cleaning Agents. Acidic cleaning and passivating media are based on phosphoric acid or alkali metal hydrogenphosphates. Surfactants support the removal of soil, grease, oil, and metal fines, so that precleaning with a neutral or alkaline agent may not be necessary. The surfactants are nonionic and low-foaming, e.g., ethoxylated fatty alcohols. Liquid or powdered products are applied mainly by spraying, sometimes by dipping. In addition to grease removal, acidic alkali metal hydrogenphosphates create a bluish iridescent or gray iron phosphate layer (0.2–1.2 g/m^2) which adheres firmly to the conversion layer below. Because the coating thickness is of the same order of magnitude as the wavelength of visible light, very small variations give rise to color changes. The pH value of the cleaning solutions is between 3.5 and 5.5. They provide temporary corrosion protection as well as considerable adhesion, giving good bonding to organic coatings such as paints or resins. The final multilayer structure prevents corrosion and red rust. Weakly acidic cleaning media are used principally in the domestic sector, e.g., in the treatment of steel furniture, office furniture, and refrigerators.

2.1.2. Cleaning Agents Based on Organic Solvents

Industrial greases and lubricants used in metal machining are insoluble in water but readily soluble in organic solvents like hydrocarbons, halogenated hydrocarbons, alcohols, esters, and ketones, especially if the lubricants are based on mineral oils consisting of paraffinic, isoparaffinic, olefinic, or naphthenic hydrocarbons. Fatty ester based lubricants and emulsions have more limited solubility. Hence, organic solvents are used for surface cleaning and degreasing of metal parts. Most of the solvents are insoluble in water, but some systems are partly or completely miscible with water, for example, ester and alcohol derivatives.

One group of water-immiscible solvents are the halogenated hydrocarbons (HHCs), today comprising only chlorinated hydrocarbons. Chlorofluorohydrocarbons (CFCs), and especially 1,1,2-trichorotrifluoroethane (R113), are no longer used because of their ozone-depleting potential. Because of health and environmental risks, chlorinated hydrocarbons (CHCs) can only be applied in "zero-emission" plants [5]. To minimize the release of the volatile CHCs during transport, storage, and handling, special logistic concepts are available from suppliers. The advantage of CHCs is their flame resistance in

combination with excellent solubility for a broad variety of lubricants, oils, and greases, low boiling point, and low viscosity. Low surface tension enables them to penetrate narrow cleavages, capillaries, and bores. Low boiling points allow degreasing processes to be carried out in the vapor phase, and make the drying process simple.

Perchloroethylene and methylene chloride are nonflammable. The CHC most widely used in the past, because of its low price, trichloroethylene, must be replaced because of its carcinogenic potential [6]. Metal parts in a CHC-based degreasing installation should be dry, because CHCs tend to form hydrochloric acid in contact with water, oxygen, and light, which the causes corrosion of the cleaning plant and the cleaned metal parts. To prevent acid formation all CHCs contain stabilizers. As the stabilizer content decreases during operation, its concentration must be checked on a regular basis, so that stabilizer can be added if necessary.

Alternatively, hydrocarbon cleaners consisting of isoparaffins and naphthenes can be used. These solvents are inflammable and have flash points above 55 °C. Air and solvent vapor can form explosive mixtures, depending on composition. Their maximum working temperature is 15 °C below flash point [7]. However, even operating temperatures above the flash point are permitted if the danger of explosion is excluded owing to low-pressure or vacuum operation or the use of inert gases (e.g., during vapor degreasing). Drying installations are more complex, because of their low vapor pressures/high boiling temperatures (ca. 200 °C) and safety aspects. Although hydrocarbon cleaners have higher viscosities than CHCs, they have excellent degreasing characteristics, especially for mineral oil based contamination. Even insoluble contaminants such as chips and cuttings, pigments, metallic soaps, and inorganic solids (e.g., salt residues) can be removed from the surfaces if they are imbedded in soluble greases or lubricants. Hydrocarbon cleaners can be used as a dual-phase cleaner dispersed in water to increase the cleaning efficiency for hydrophilic contamination. After drying, the parts are temporarily protected from corrosion by a thin hydrophobic film on the surface.

Hydrocarbon cleaners are handled in closed installations. Solvent cleaners (provided they do not contain additives) have simple compositions compared to aqueous or surfactant cleaners. They contain no inorganic or organic salts and evaporate without residue at elevated temperatures.

2.1.3. Cleaning Process Technology

Cleaning efficiency is limited by process technology. Mechanical energy input plays an important role when parts with complex geometries (narrow clefts and bores) or bulky parts are treated. Mechanical energy input is realized by using ultrasound, spraying, or jet technologies. Especially when using aqueous cleaners, the appropriate cleaning agent must be selected with respect to foaming, ultrasonic stability, etc.

A multistep cleaning installation is illustrated in Figure 2. The cleaning of heavily contaminated parts requires a multistep cleaning installation with precleaning, cleaning, two or more rinsing baths, and a drying chamber. The precleaning stage removes heavy oil contamination and larger particles, so that the second cleaning bath is not overloaded with oils and lubricants. To save water, minimize waste, and ensure constant bath quality, all media within the cleaning installation should be treated and recycled. Efficiency is increased by managing all media throughputs in a cascade counterflow from lower to higher contamination, separately for the cleaning and the rinsing stages, so that contents of both contaminants and cleaning agents decrease from the initial to the final bath. Both cleaning baths should be treated by phase separation and micro- or ultrafiltration to separate insoluble and emulsified oils as well as particles in a bypass. Cleaning-agent concentration must be controlled in all baths. Diligent rinsing is crucial when aqueous, especially alkaline, cleaners are used, to remove nonvolatile cleaning-agent components which otherwise form milky-whitish layers or spots and can impede subsequent surface treatments [3].

Thus, at least one rinsing stage is necessary. Rinsing is performed with desalted or demineralized water. To prevent corrosion after drying, a corrosion inhibitor may be added to the second or the last rinsing bath. Attainable surface quality is limited by the composition (i.e., the degree of contamination) of the last rinsing bath.

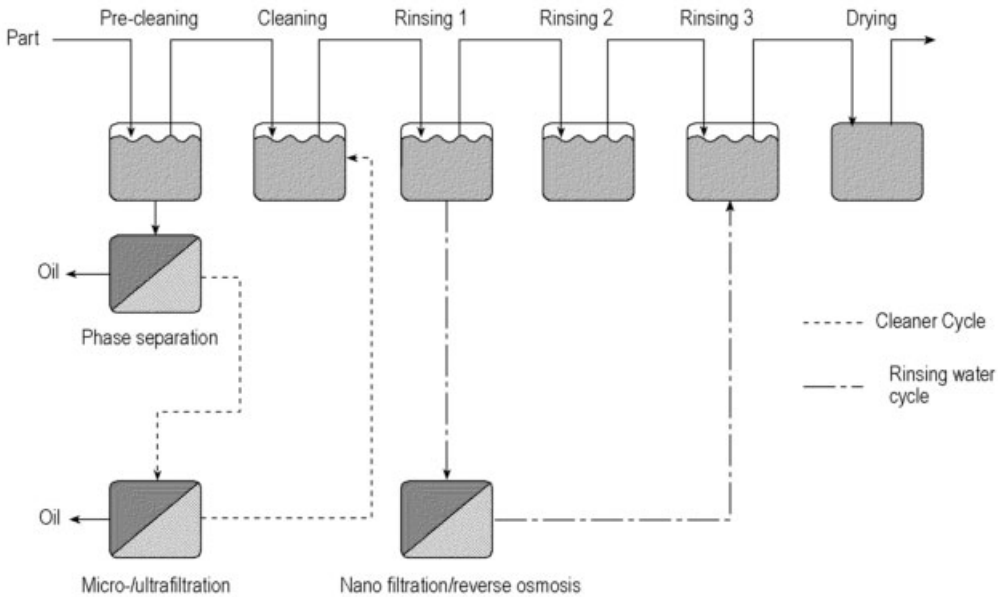

Figure 2. Cleaning installation

Organic solvents, especially chlorinated hydrocarbons, are applied in closed or even zero-emission installations, where solvent emissions are minimized by low-temperature condensation (e.g., hydrocarbons), activated-carbon filtration, or gas-phase exchange (zero-emission plant) so that they obey European and national law.

Alternative processes under development are high-pressure processes using, e.g., supercritical carbon dioxide or hot air jets with or without water vapor, whose effectiveness will depend strongly on the type of soil or contamination to be removed.

2.2. Cleaning and Pickling with Salt Baths

Although aqueous media are increasingly preferred, even for difficult scale removal and surface conditioning, molten salt baths are still in use. The latter are mixtures of inorganic compounds used in their anhydrous fused state at elevated temperature (200 – 540 °C depending on composition and application). Their commercial applications can be divided into three main categories: heat-transfer media, active heat-treating baths for carburizing and/or nitriding, and chemically reactive melts. The last category is of interest in the molten salt bath cleaning of metals.

The main function of a metal-cleaning salt bath is to react with an undesirable impurity, coating, or oxide on the surface of a metal without adverse attack on the base metal. Examples of the wide variety of inorganic and organic surface impurities which may be effectively removed with molten salt baths include metal oxides formed during thermal treatment or forming operations, lubricants, plasma wear-resistant coatings such as tungsten carbide which must be removed from a component for rework or repair, organic deposits such as paints or polymers, or a ceramic residue from a casting operation.

2.2.1. Bath Composition and Properties

Molten salt baths are excellent reactive media for metal cleaning. High chemical activity, in combination with good solvency and fluidity, allow reaction with compounds that are resistant to conventional aqueous or organic solvents and chemical solutions, because of their elevated operating temperature and the absence of oxygen.

The composition of a bath is governed by many variables including melting temperature, application or end use, and long-term thermal

stability. Depending on its composition, a bath may be reducing, neutral, or oxidizing. Some proprietary neutral compositions may, however, be either reducing or oxidizing when activated with low-voltage, high-amperage direct current [8].

When operated in properly designed and engineered furnaces, most salt bath processes have very good longevity of operation and, unlike most aqueous systems, do not require complete dumping or recharge with continued use.

Most cleaning operations with molten salt generate byproducts with limited solubility in the host molten salt. Even at and beyond saturation, the molten salts continue to react with and to remove surface contaminants. The byproducts simply precipitate and can be conveniently and safely removed from the working bath via solids collection systems (Fig. 3) [9]. "Soluble" impurities do not usually reach a detrimental concentration; drag-out and make-up tend to prevent buildup.

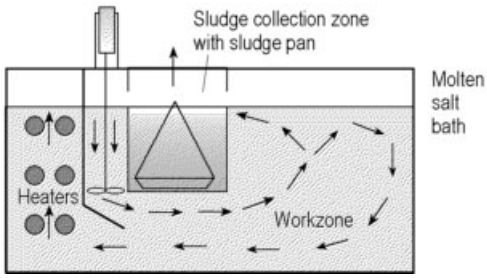

Figure 3. Removal of reaction byproducts from a molten salt bath with a dedicated sludge zone

Fresh salt is added to the bath after solids removal to restore the bath to its original operating level and to maintain proper bath chemistry. Fresh salt additions may be added manually in the form of granular dry salts. For high-volume production applications (such as continuous strip annealling and pickle lines for stainless steel), regenerating salt may be automatically fed to the bath as a concentrated aqueous feed stream [10]. The chemistry of the bath usually does not require routine chemical analysis or adjustment to its composition; proper chemical balance and reactivity is maintained by removal of the byproducts from the bath and additions of fresh process chemicals to the bath.

2.2.2. Descaling of Metals

One of the largest applications for molten salt bath cleaning is the descaling of metals, e.g., stainless steels, nickel and cobalt-based superalloys, and titanium alloys.

Mineral Acid Pickling. Scale and oxide are removed from metals by pickling in mineral acid solutions. Commonly used mineral acids include sulfuric, nitric, hydrochloric, and hydrofluoric acids, used either alone or in combination. Sulfuric and hydrochloric acids are used for oxide removal from low-carbon steels, while nitric or combinations of nitric and hydrofluoric acids are used for oxide removal from higher-alloy materials such as stainless steels and superalloys.

As newly developed alloys are increasingly acid- and corrosion-resistant, their respective oxides and scales are increasingly acid-resistant as well. Much of the scale removal from these alloys by mineral acids is based on "undermining" of the oxide layer at the oxide – metal interface. By dissolving some of the base metal at the interface, the oxide layers are loosened and are either carried away by gases generated during pickling and/or by mechanical scrubbing or high-pressure water rinsing.

Although mineral acid pickling is satisfactory for low-carbon materials, it has numerous drawbacks when used on high-alloy materials. In most cases, the oxides formed on stainless steels and superalloys are much less reactive to mineral acids than the metal substrate. High acid concentrations and operating temperatures, in conjunction with long immersion times usually needed for complete scale removal, can lead to irregular roughening, pitting, and intergranular attack on the base metal. Economic and environmental concerns include long process times, short acid life, excessive metal loss, fuming, and waste-acid disposal costs and liabilities. Although mineral acid pickling is still an important and integral part of scale and oxide removal from these alloys, it is now often used together with molten-salt conditioning.

Molten-Salt Conditioning. Preliminary scale conditioning processes such as molten salt bath descaling allow subsequent mineral acid pickling operations to be carried out at substantially lower acid concentrations/temperatures and shorter immersion times while pro-

ducing superior results. For example, the acid pickling of austenitic stainless steels may require 15 % nitric and 5 % hydrofluoric acid at 60 °C, while after salt conditioning the same alloys can be cleaned with 10 % nitric and 0.5 % hydrofluoric acid at 45 °C and with a shorter immersion time. The loss of base metal is significantly reduced and the useful life of the pickling solution is extended.

Reducing salt baths based on sodium hydroxide and sodium hydride have also been used for descaling stainless steels. Sodium hydride was formed in a chemical generator by treating molten sodium metal with dissociated ammonia. The hydride was then added to molten sodium hydroxide to form the working bath. The hydride reduced the metal oxides to form metallic deposits on the metal surfaces.

Most descaling baths are of an oxidizing nature, comprising alkali metal hydroxides, alkali metal nitrates, oxidizing agents, and in some instances catalysts such as alkali metal chromates, vanadates, and manganates to provide a consistent, controlled reaction potential [11, 12].

Molten salts are very efficient for chemically modifying the tenacious oxide films formed on metals during thermal processing operations (e.g., forging, annealing, and stress relieving).

Components to be descaled are immersed into an oxidizing molten salt which operates in the temperature range 315 – 510 °C depending on application and composition. Reaction times are short, ranging from a few seconds to several minutes. A molten salt bath facility may be used either as a batch operation in which fixtures or baskets containing the workpiece are immersed sequentially into the molten salt, water rinses, acid brighteners, etc., or may be used for processing of metal strip as in a continuous annealing and pickling line.

Chemical Reactions. The scale-modification reaction is a two-stage reaction between the original metal oxide film and the molten salt. The metal oxide film is first oxidized from its initial low valence state to a high valence state by means of the oxidizing potential of the molten salt. The second stage involves reactions between the newly formed, high-valent metal oxides and other constituents of the bath to form both bath-soluble and bath-insoluble byproducts.

Low-valent chromium oxides, for example, are rapidly oxidized to chromium(VI) oxide, which in turn dissolves and reacts with the bath to form an alkali metal chromate:

$$CrO + O_2 \rightarrow CrO_3$$

$$2\,Cr_2O_3 + 3\,O_2 \rightarrow 4\,CrO_3$$

$$CrO_3 + 2\,NaOH \rightarrow Na_2CrO_4 + H_2O$$

Analogous reactions take place with many of the metals used in important alloy systems including stainless steels, nickel and cobalt-based superalloys, titanium, and tungsten.

Alkali metal chromates are soluble both in the molten salt bath and in water. On completion of the molten salt bath conditioning cycle, components are then "quenched" in water at ambient temperature. Quenching cools the metal to a safe operating temperature, assists in thermally "shocking" any heavy scale deposits off the metal surface, and dissolves the thin salt film and any water-soluble reaction products from the surface of the metal. Up to 80 % of the original oxide thickness is removed by salt bath conditioning.

After water quenching, a thin oxide film composed of bath-insoluble species is usually still present on the metal surface. Figure 4 shows an Auger surface scan of annealed 304 stainless steel before (A) and after (B) salt bath conditioning. The film remaining after conditioning is composed of iron oxide and minor alloy oxides. The oxides are in high valence states and readily dissolve in acid pickling compositions. Prior oxidation of the surface scale by salt bath conditioning allows subsequent pickling operations to be carried out with more dilute acids and at lower temperatures than are possible without salt bath conditioning. Pickling times with salt bath conditioning are also usually shorter.

Refractive Metals and Alloys. The alloying elements responsible for producing high-temperature oxidation resistance and chemical resistance (e.g., nickel and chromium) make descaling and pickling of high-temperature alloys difficult. These alloys are most beneficially treated in molten oxidizing salt baths. In nickel-based superalloys, thermal treatments such as hot working, heat treating, and welding promote the formation of acid-resistant nickel and chromium oxides. Elements such as yttrium which

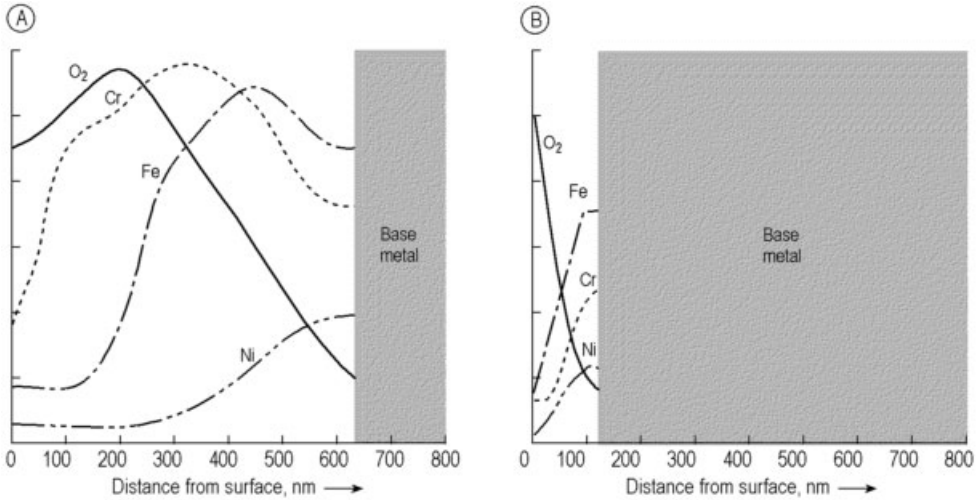

Figure 4. Auger surface analysis of annealed, 304 stainless steel
A) As annealed, total dioxide thickness 650 nm; B) After salt conditioning, total oxide thickness 130 nm

lend cyclic thermal stability to protective oxide films for high-temperature oxidation resistance also complicate the metal cleaning task.

Titanium and Titanium Alloys. Descaling of titanium and its alloys is also aided by molten salt bath conditioning. Surface oxides which are formed during annealing, stress relieving, or hot forming operations are converted to the respective alkali metal compounds in a manner analogous to the oxidation of chromium oxide. Lubricant residues from forging or forming operations, such as vitreous coatings, graphite, boron nitride, and molybdenum disulfide, react with the oxidizing molten salt to form alkali metal silicates, carbonates, borates, and sulfates, respectively.

After water quenching and rinsing, residual surface contamination is removed from the titanium by acid pickling. The surface has a uniformly reactive surface after salt bath conditioning. This is especially important if the titanium is to be subjected to chemical milling either for α-case removal or for gage reduction. (A brittle α-phase surface layer may form on β-titanium alloys due to surface alloying with oxygen or other interstitial elements. Chemical milling selectively reduces the thickness of titanium components or removes the α-case with a mixture of 35 % nitric and 3 % hydrofluoric acid.) Acid pickling is usually accomplished by first pickling the salt-conditioned titanium in hot sulfuric acid to expose the base metal, and then in a mixture of nitric and hydrofluoric acid for chemical milling and brightening.

Carbon Steel. Molten salt bath treatment is of little benefit in descaling carbon steel and low-alloy steel. Although oxidizing molten salts react with iron oxides, the resulting high-valent iron(III) oxides have very low solubility in the bath. The oxide film developed on carbon steel does not dissolve in the bath, and the metal surface becomes passive to any further chemical reaction. The formation of a passive iron oxide film on low-carbon steels allows salt bath furnace equipment to be manufactured from this cheap and readily available material.

2.2.3. Casting Cleaning

Sand and ceramic materials are commonly used for pattern and shell making in the metal casting industries (→ Foundry Technology). Residual core sand and/or shell may be difficult to remove by mechanical methods such as shot blasting because of casting geometries, or by chemical methods because of poor reactivity in conventional aqueous chemical systems.

Molten salt baths react with many of the commonly used shell and core compositions, which generally contain a substantial proportion of silicon dioxide [13]. Suitable bath compositions must include alkali metal hydroxides and may

contain proprietary components that cause the bath to be more aggressive and to reject the silicates from solution more readily.

Silica reacts readily with molten alkali metal salts to form alkali metal silicates:

$$SiO_2 + 2\,OH^- \rightarrow SiO_3^{2-} + H_2O$$

Other common shell or core constituents are zirconium silicate and aluminum silicate, which are either unreactive or only react very slowly with the molten salt. They are stripped off as the silicon dioxide is dissolved by the bath and the shell integrity weakens. The ceramic composition must include some silicon dioxide if the molten salt bath cleaning process is to be effective within a reasonable amount of time and at a reasonable operating temperature.

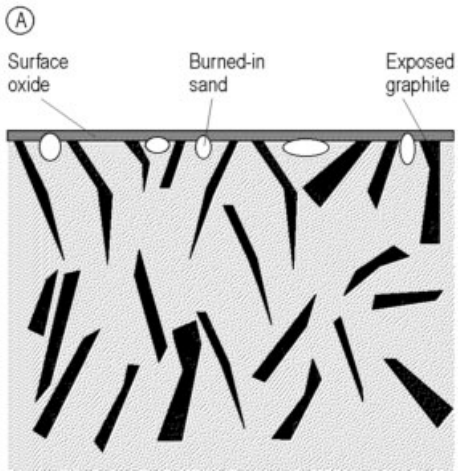

Figure 5. Electrolytic cleaning of cast iron with molten salt — A) Original cast iron surface

The cleaning of iron castings is another example of the versatility of molten salt bath processing (Figs. 5, 6, 7, 8, 9). Oxidizing or reducing conditions can be created within a single neutral salt bath by applying direct current potentials. Apart from surface oxides, iron castings usually contain some "burned-in" core sand which may come free during use and cause failure in critical components (Fig. 4 A). In addition, free surface graphite on as-cast or as-machined surfaces makes bonding operations such as babbitting or brazing difficult. (Babbitting is the application of tin- or lead-based alloys to journal areas of a component to provide a low-friction bearing surface.) By cleaning the castings in an electrochemically activated molten salt bath, surface scale, residual sand, and exposed surface graphite can all be removed without any detrimental attack on the casting surface. No significant dimensional changes occur.

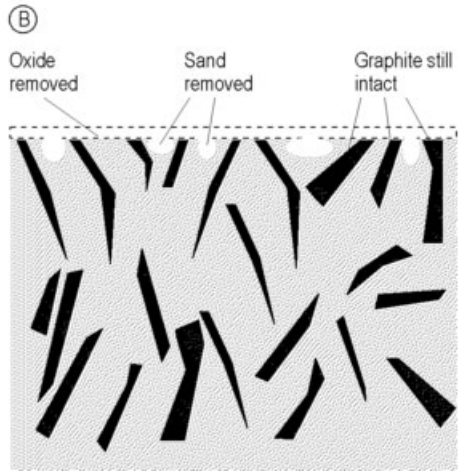

Figure 6. Electrolytic cleaning of cast iron with molten salt — B) After first reduction cycle

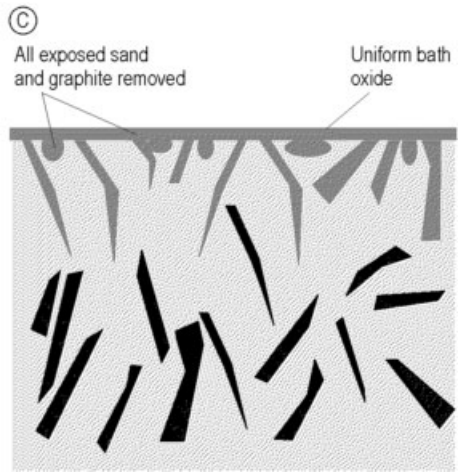

Figure 7. Electrolytic cleaning of cast iron with molten salt — C) After oxidation cycle

In the first step, the components to be treated are subjected to a cathodic potential at which oxides on the casting surface are electrochemically reduced, and any exposed sand particles are removed (Fig. 6 B). The polarity is then re-

versed to anodic, and a thin layer of iron oxide is formed on the metal surface, presumably accompanied by the oxidation of surface carbon to carbon dioxide (Fig. 7 C). A final reducing cycle yields an oxide-, sand-, and graphite-free surface (Fig. 8 D). When the cleaned casting is brazed or babbitted, a superior adhesive bond is formed because the coating spreads uniformly on the surface of the casting and penetrates into surface microvoids, e.g. formed by the removal of surface sand and graphite (Fig. 9 E).

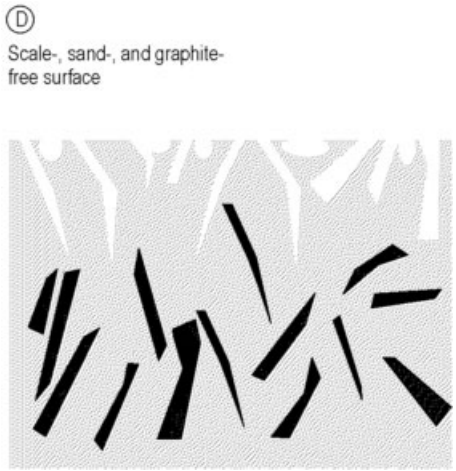

Figure 8. Electrolytic cleaning of cast iron with molten salt — D) After final reduction cycle

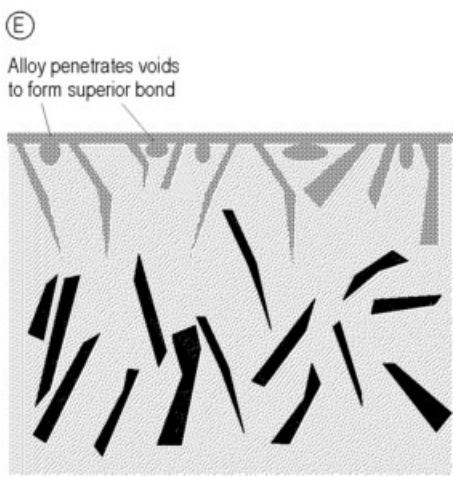

Figure 9. Electrolytic cleaning of cast iron with molten salt — E) After brazing and babbitting

2.2.4. Coatings Removal

Numerous coatings are applied to metal substrates for engineering and aesthetic purposes. Coatings (e.g., plasma- or flame-sprayed tungsten carbide and chromium carbide) are used to impart desired wear- and erosion-resistant characteristics to critical components. If these coatings must be removed, conventional chemical methods (see Section 2.1) are often inadequate because the coatings are resistant to aqueous solutions; mechanical methods are labor-intensive, and the coatings are difficult to remove because of their wear-resistant characteristics. Molten salt baths can react with the metal or the carbide fractions of these coatings, and remove them.

Tungsten carbide, for example, reacts with an oxidizing molten salt to form alkali metal tungstates and carbonates:

$$\mathrm{WC}_x + (\mathrm{Na,K})\,\mathrm{OH} + \mathrm{NO}_3^- \rightarrow (\mathrm{Na,K})_2\mathrm{WO}_3 + \mathrm{NO}_2^- + \mathrm{H}_2\mathrm{O} + (\mathrm{Na,K})_2\mathrm{CO}_3$$

Similar reactions occur with other commonly used coatings such as chromium carbides.

Removal of organic coatings such as paints and polymers relies on the oxidizing potential of molten salts to provide controlled chemical "combustion". The molten salt acts as a heat source to initiate the reaction, and as a heat sink that absorbs excess heat from the exothermic reaction of the organic materials being removed. These properties help to prevent localized overheating of the components being cleaned and assure complete removal without the formation of carbonaceous chars or soot. Typical uses for salt bath paint removal include both salvage stripping of rejected coated components and maintenance stripping of carriers, hooks, fixtures, and racks from production coating lines. The stripping operation may be carried out batchwise or continuously by integrating the stripping system into the production coating line.

Removal of solidified polymer residues is an important application for molten salt bath cleaning in the synthetic-fiber industry. The molten polymer is extruded through an orifice plate known as a spinneret. The dimensions and geometry of the spinneret holes are critical for the consistent production of a quality fiber, and represent a significant capital investment on the part of the fiber producer. Conventional thermal

cleaning methods such as incineration or burnoff can result in metallurgical damage from localized overheating. Molten salt bath cleaning prevents localized overheating during the polymer removal process and allows the spinneret to be cleaned without any adverse affects on performance [14].

3. Phosphating

Phosphating is the process of coating a metal surface with a layer of insoluble metal phosphates by treating it with an acidic phosphate-containing solution [15 – 20]. Zinc, iron, and manganese phosphate coatings are most common. The phosphate coating, specified as a conversion coating owing to the coating mechanism, provides an excellent base for paints, oils, or cold-forming lubricants. Iron and steel, steel coated with zinc or zinc alloys, and, to a lesser extent, aluminum alloys are the most common substrates.

Phosphating is used in the metalworking industry all over the world. Thus, virtually all passenger cars are zinc phosphated prior to painting; cold extrusion of steel would not be economically feasible without phosphating.

Phosphating chemicals are proprietary; major suppliers include Chemetall, Henkel, Nihon Parkerizing, Nippon Paint, and PPG.

3.1. Chemistry

Table 2 presents a collection of phosphating processes and phosphate coatings. The chemical basis for all of these processes is similar and will be exemplified in more detail for zinc phosphating.

Reaction Steps in Zinc Phosphating. Phosphating solutions are applied either by spraying or immersion (dipping).

Essential constituents of a zinc phosphating solution are zinc, phosphate ions, and an oxidizing agent (Ox), often sodium nitrite. The pH value is between 2 and 3.5. Concentrations vary considerably; additives such as nitrate, fluoride, silicofluoride, nickel ions, or manganese ions are common. As an example, the following approximate bath composition is suitable for the phosphating of steel sheets as a pretreatment prior to painting: Zn^{2+} 1.2 g/L; Ni^{2+} 0.1 g/L; H_3PO_4 + $H_2PO_4^-$ 15 g/L; and NO_2^- 0.1 g/L.

The pH of the solution is adjusted to 3.2 with sodium hydroxide. The process is usually carried out at up to 95 °C. For cold forming applications total concentrations may well be ten times higher.

In the ideal case, the phosphating reaction can be broken down into five steps:

Pickling reaction

$$Fe + 2\,H^+ \rightarrow Fe^{2+} + H_2 \text{ for Fe} \tag{1}$$

$$Zn + 2\,H^+ \rightarrow Zn^{2+} + H_2 \text{ for Zn} \tag{2}$$

$$AlOOH + 3\,HF \rightarrow Al^{3+} + 3F^- + 2\,H_2O \text{ for Al} \tag{3}$$

$$Al + 3\,H^+ \rightarrow Al^{3+} + 3/2\,H_2 \text{ for Al} \tag{4}$$

Acceleration

$$H_2 + 2\,Ox \rightarrow 2\,HOx \tag{5}$$

$$Fe^{2+} + H^+ + Ox \rightarrow Fe^{3+} + HOx \text{ for Fe} \tag{6}$$

Complex formation

$$Al^{3+} + 6\,F^- \rightarrow AlF_6^{3-} \text{ for Al} \tag{7}$$

Coating formation

$$3\,Zn^{2+} + 2\,H_2PO_4^- + 4\,H_2O \rightarrow Zn_3(PO_4)_2 \cdot 4\,H_2O + 4H^+ \tag{8}$$

$$2\,Zn^{2+} + Fe^{2+} + 2\,H_2PO_4^- + 4\,H_2O \rightarrow Zn_2Fe(PO_4)_2 \cdot 4\,H_2O + 4\,H^+ \tag{9}$$

Sludge formation

$$Fe^{3+} + H_2PO_4^- \rightarrow FePO_4 + 2\,H^+ \text{ for Fe} \tag{10}$$

$$AlF_6^{3-} + 3\,Na^+ \rightarrow Na_3AlF_6 \text{ for Al} \tag{11}$$

Pickling Reaction. Part of the metal surface is dissolved by the acid. The phosphating of aluminum requires fluoride ions to attack surface oxides which only slowly dissolve in phosphoric or nitric acid.

Acceleration. The pickling reaction is accelerated by oxidizing agents (Ox) called accelerators (Table 3), which also prevent the evolution of excessive amounts of molecular hydrogen (Eq. 5), thus minimizing the danger of hydrogen embrittlement.

Table 2. Phosphating processes and coatings

Phosphating process	Cations in the phosphating solution*	Phosphate coating		Coating weight, g/m ^{2}	Substrate
		Composition	Mineralogical name		
Zinc phosphating	Zn^{2+} (Fe^{2+})	$Zn_3(PO_4)_2 \cdot 4\,H_2O$	hopeite	1.3 – 40	Fe, Zn, Al
		$Zn_2Fe(PO_4)_2 \cdot 4\,H_2O$	phosphophyllite	1.3 – 40	Fe
	(Ni^{2+}, Mn^{2+})	$Zn(Ni, Mn, Zn)_2(PO_4)_2 \cdot 4\,H_2O$	isotypic hopeites	1.3 – 15	Fe, Zn, Al
Zinc calcium phosphating	Zn^{2+}, Ca^{2+}	$Zn_2Ca(PO_4)_2 \cdot 2\,H_2O$	scholzite	1.5 – 10	Fe, Zn
		$Zn_3(PO_4)_2 \cdot 4\,H_2O$	hopeite	1.5 – 10	Fe, Zn
		$Zn_2Fe(PO_4)_2 \cdot 4\,H_2O$	phosphophyllite	1.5 – 10	Fe
Manganese phosphating	Mn^{2+}, Fe^{2+}	$(Mn,Fe)_5H_2(PO_4)_4 \cdot 4\,H_2O$	hureaulite	3 – 30	Fe, Zn
Nickel phosphating	Ni^{2+}	$Ni_3(PO_4)_2 \cdot 8\,H_2O$		1 – 2	Fe
Iron phosphating	Na^{+}	$Fe_3(PO_4)_2 \cdot 8\,H_2O$	vivianite	0.2 – 2	Fe
Alkali phosphating		oxides and hydroxides			Fe, Zn, Al

* Ions in parentheses are optional.

Table 3. Accelerators for zinc phosphate processes

Accelerator	Method of application	Use			Coating weight, g/m^2
		Painting	Cold forming	Oiling	
Nitrite	spray, dip	+			1.5 – 4
	dip		+	+	5 – 15
Nitrate	dip		+	+	10 – 40
Chlorate	dip	+	+	+	3 – 8
Chlorate and nitrite	spray	+			1.5 – 3
Peroxide	spray	+			1.5 – 3
Organic nitro compounds	spray, dip	+			1.3 – 4
Hydroxylamine	dip	+			1.3 – 4
Nitroguanidine	spray, dip	+			1.5 – 5

Coating Formation. Metal dissolution in the pickling reaction results in a significant increase of the pH within the concentration boundary layer close to the metal surface. As a consequence, the equilibrium constant (Eq. 12) for the precipitation reaction is exceeded and zinc phosphate is precipitated. Since the metal surface offers many favorable nucleation sites (there are indications of epitaxial growth [21]), a phosphate coating forms. It ceases to grow further as soon as the metal surface is completely covered because metal oxidation ceases, and the pH within the boundary layer no longer increases. (Eqs. 1 – 4).

The mechanism demands that phosphating solutions be close to the equilibrium as given by the solubility constant of the precipitated phosphate. In spray applications slightly supersaturated solutions are common, whereas dip processes use solutions that are slightly (for low coating weights) to moderately more acidic (for heavy coating weights) than the equilibrium values.

Complex formation. When coating aluminum, fluoride ions must be available for complex formation (Eq. 7), since Al^{3+} ions in concentrations as low as 3 mg/L prevent the formation of zinc phosphate coatings. If different metals including aluminum are to be treated, use of fluoride-free phosphating solutions may allow phosphating of steel or zinc without coating the aluminum.

Sludge Formation. Dissolved Fe^{3+} and $AlF_6{}^{3-}$ ions from the pickling reaction are not enriched in the phosphating solution but are precipitated as iron(III) phosphate (Eq. 10) or trisodium hexafluoroaluminate (Eq. 11), respectively. When treating galvanized surfaces and in spraying processes, relatively small amounts of tertiary zinc phosphate sludge are formed as well.

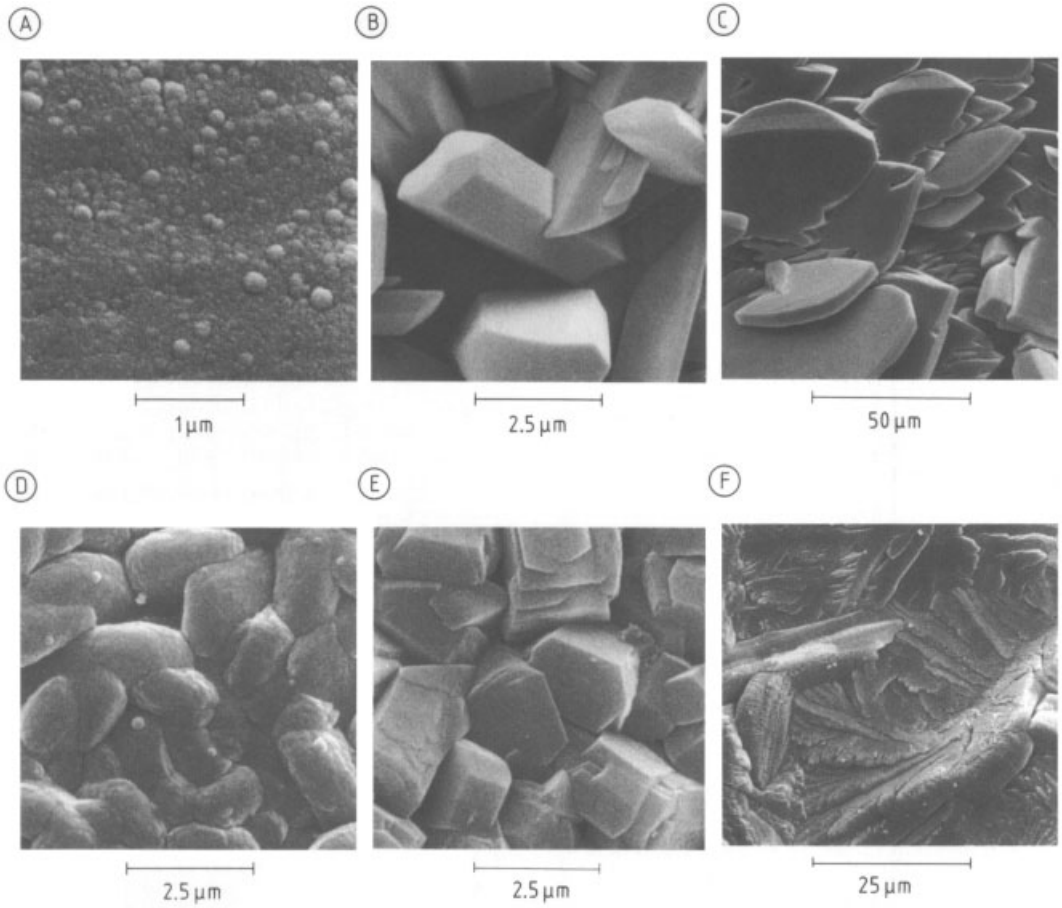

Figure 10. Scanning electron micrographs of phosphate coatings on steel A) Iron phosphate, spray; B) Manganese phosphate, immersion; C) Zinc phosphate, immersion; D) Zinc calcium phosphate, immersion; E) Low zinc phosphate, immersion; F) Low zinc phosphate, spray

Nitrate-accelerated processes often do not involve sludge formation. Iron(II) nitrate accumulates until an equilibrium is reached between iron dissolution and dragout. (Dragout denotes the solution that adheres to the treated parts and is carried over into the next treatment stage.)

Newer Processes for Zinc Phosphating. Zinc phosphating carried out prior to painting (prepaint process) can be classified as high-zinc or low-zinc processes. High-zinc processes operate at 3 – 4 g/L Zn^{2+}, and low-zinc processes at 0.7 – 1.5 g/L Zn^{2+}, the upper limit for dip applications. Low zinc concentration, i.e., high phosphate to zinc ratio, improves corrosion protection. Coatings on steel produced by low-zinc processes consist mainly of phosphophyllite and show superior paint-base performance to the high-zinc processes with their hopeite coatings [22]. Layer performance is further increased with the introduction of trication processes, which make use of the elements zinc, nickel, and manganese, both in solution and in the coating [23]. Consequently, these processes have become standard pretreatment in, for example, the automotive industry.

Low-temperature processes for cold forming are accelerated by nitrate, and the iron(III) concentration in the bath is limited to 5 – 8 g/L, owing to oxidation with air, preferably performed continuously in a separate reaction tank. Operating temperature has thus been lowered by ca. 30 to 50 – 60 °C [24], and the phosphating tank remains virtually free of sludge.

Health and environmental concerns have promoted the development of

- *Nitrite-free processes* with hydroxylamine [25], nitroguanidine, or hydrogen peroxide as accelerator [26, 27], the last-named offering the additional benefit that its reaction product, H_2O, is of no concern whatsoever.
- *Nickel-free processes* [28, 29].
- *Chromium-free* after-rinses based on organotitanium [30], inorganic zirconium [31], or polymeric compounds [32].
- *Effluent-free phosphating* lines using ultrafiltration for cleaning, hydrogen peroxide acceleration in phosphating, and precipitation plus ion exchange for chromium-free after-rinses, without compromising the performance of the painted parts [26]

Other phosphating processes are summarized in Table 2. Various phosphate coatings are shown in Figure 10.

Iron phosphating processes may not require acceleration. The coating weight varies with the bath composition. Coatings formed on ferrous surfaces contain iron oxides and phosphates. Iron phosphating solutions normally contain surfactants for cleaning, and soiled surfaces may thus be treated in one step (cleaner – coater).

Manganese phosphating processes are performed by dipping only, and iron(II) nitrate is often used for acceleration.

3.2. Methods of Application

The phosphating solutions are generally applied by spraying or by immersion depending on the number, size, and shape of the parts to be treated. The type of application may lead to differences in the composition and morphology of phosphate coatings (Fig. 10).

Strip is also coated by the roll-on, dry-in-place process, in which phosphating solution is applied to the strip and, without rinsing, is dried to form the phosphate coating.

Although phosphating is the crucial step in chemical pretreatment of metal surfaces, other steps may be necessary to complete the process cycle:

1) Degreasing and cleaning to remove oils, fats, and dirt
2) Rinsing with water
3) Pickling to remove rust or other oxides
4) Rinsing with water
5) Activation with hot water or with titanium or manganese phosphate dispersions to induce the formation of a fine-grain phosphate coating in the subsequent step
6) Phosphating
7) Rinsing with water
8) Final rinsing with deionized water (may contain corrosion inhibitors)

Rinsing with water is needed between the processing stages to prevent excessive contamination of treatment baths by preceding solutions.

3.3. Uses of Phosphate Coatings

Corrosion-Protection and Prevention of Red Rust. Heavy zinc and manganese phosphate coatings together with an anticorrosive oil provide substantial protection against corrosion (Fig. 11). Nuts, screws, bolts, and tubes are protected in this way.

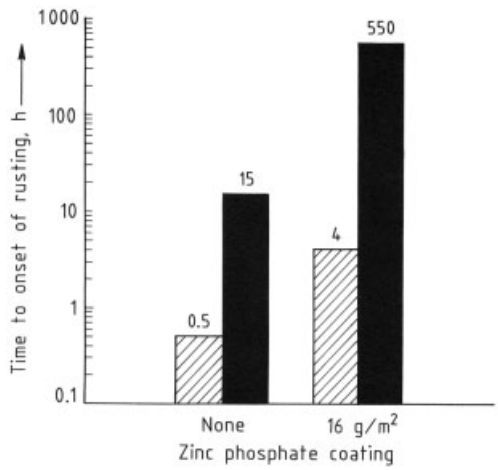

Figure 11. Zinc phosphating: corrosion protection in the salt spray test
Hatched: without oil; Full: with oil

Bearing Surface Lubrication. The presence of a crystalline manganese phosphate coating improves the formation of a uniform lubricant film on load-bearing surfaces. It provides some basic lubrication even if the supply of organic lubricant is interrupted. It also helps

to overcome surface irregularities and surface roughness, thus shortening the running-in period of gears and engines. Manganese phosphate coatings are therefore used for engine parts such as pinions, camshafts, pistons, gears, and valves.

Cold Forming. Cold forming of metals involves very high surface stresses. Under these conditions metal-to-metal contact may occur, damaging the workpiece and increasing tool wear. Phosphate coatings provide an effective barrier between metal and tool. Coating weight varies between 3 – 5 g/m^2 for light deformations, such as a calibration pass on a tube and 20 – 40 g/m^2 for heavy cold forging or cold extrusion. In addition, phosphate coatings retain more lubricant and may form highly lubricating zinc stearate coatings when treated with reactive soap lubricants. Phosphate coatings are used in all types of cold forming operations, i.e., drawing of wire, tube, or profile; deep drawing; cold heading; cold extrusion; and cold forging. Steel strip is zinc phosphated and treated with reactive soap. Electrogalvanized and hot dip galvanized steel strip is zinc phosphated in the galvanizing line to improve formability in subsequent drawing operations.

Paint Base. Phosphating enhances the adhesion and corrosion protection of paints. For automotive applications, zinc phosphate coatings of the low-zinc, manganese-modified type offer distinct advantages such as stone-chipping resistance and wet adhesion on galvanized substrates. For powder coatings, iron phosphate often gives optimum results. Figure 12 demonstrates results for a water-based paint system.

Special Uses. Special phosphating processes are used for a number of applications:

Electrical Insulation. Phosphate layers have been used to coat the silicon steel sheets forming the cores of electric motors, generators, or transformers. The insulation provided by a phosphate coating of thickness 1 – 6 μm is sufficient to prevent eddy currents.

Coil Coating. Chlorate-accelerated iron phosphate processes are commonly employed for the coil coating of steel strip; zinc phosphate processes are used for galvanized strip. The pretreatment considerably improves adhesion, which is essential during the forming operation, as well as corrosion protection of the paint film or plastic laminate.

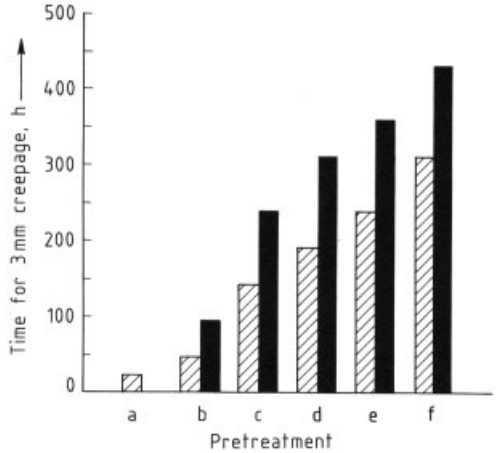

Figure 12. Performance of a water-based paint system (primer and top coat) in the salt spray test
Pretreatment: a) Vapor degreasing with tetrachloroethylene; b) Iron phosphate, spray; c) Zinc phosphate, spray, "normal zinc"; d) Zinc phosphate, spray, "low zinc"; e) Zinc phosphate, dip, "normal zinc"; f) Zinc phosphate, dip, "low zinc"
Final rinsing: hatched: deionized water; full: Cr^{VI}/Cr^{III}-containing passivation. Cr(VI) compounds must be substituted according to actual European and national law.

Nickel Phosphating Prior to Vitreous Enameling. Nickel phosphating can replace chemical nickel coating as a pretreatment for vitreous enameling. In this way the number of process steps and the amount of nickel-containing solid waste can be considerably reduced [33].

4. Case Hardening of Ferrous Metals

Parts and components are subjected to mechanical strain and stress, must resist shock and impact, and are affected by wear and corrosion in chemically, thermally or mechanically demanding environments. These stresses are at a maximum at a component's surface. To increase a component's resistance to fatigue, wear, and, in some cases, corrosion, case hardening is applied to obtain a surface layer several micrometers to several millimeters in thickness in which the substrate's hardness is increased significantly, whereas the substrate itself retains its toughness and ductility [34].

The term "case hardening" refers to quite different processes and treatments which have

in common that solely surface hardness is increased to provide increased fatigue strength and wear resistance and, depending on the chemical composition of the surface layer, corrosion resistance.

Surfaces with a high hardness can be produced by two methods:

1) If the component is composed of a "hardenable material (e.g., steel), the surface can be treated thermally (i.e., by short-term heating and subsequent quenching); no chemical reaction via the phase boundary is involved. (see Section 4.1).
2) Otherwise, additional elements can be embedded in the surface region by thermochemical treatment, i.e., thermal treatment in a chemically reactive environment, which can be gaseous, liquid, or solid and releases these elements by chemical reaction at the phase boundary and diffusion into the substrate.

Whereas both processes increase the substrate's surface hardness, owing to modified structure in terms of chemical or phase composition, alternatively hard coatings can be applied by means of chemical vapor deposition (CVD), physical vapor deposition (PVD), ion implantation processes (→ Thin Films), in various gas phases at ambient or reduced pressure; by spraying (Chap. Thermal Spraying), by electrochemical and electroless chemical processes, and electroplating in the liquid phase (→ Electrochemistry → Electrochemical and Chemical Deposition).

4.1. Case Hardening without Chemical Reaction via Phase Boundary

The steel base material must be hardenable, i.e., it should contain at least 0.35 wt % of carbon. Alloyed and unalloyed steels with a limited range of carbon content and good casting qualities are generally used, mostly in the as-tempered condition. Surface hardening may be followed by tempering, if necessary at low temperature.

Iron exists in a body-centered cubic modification (α-Fe, bcc) and a face-centered cubic modification (γ-Fe, fcc); ferrite is a solid solution of carbon in α-iron; austenite is the solid solution of carbon in γ-iron, with a significantly higher solubility. During hardening, steel is first heated to a temperature above A_{c3} (i.e., the temperature at which the $\alpha \rightarrow \gamma$ phase transition occurs) and then quenched (cooled rapidly). Above the transition temperature, carbon present as iron carbide Fe_3C, or cementite, dissolves because of the higher solubility of carbon in γ-iron. During rapid quenching, iron carbide precipitation cannot be completed because the precipitation rate is limited by diffusion. This leads to an extremely hard structure by the name of martensite, which is not favorable as substrate because of its brittleness. If, however, the transition temperature is reached only in the surface region of a steel substrate, owing to an extremely high heating rate, then only the surface will be heat-affected, i.e., hardened. As long as the hardened layer is thin, the internal (compressive) stresses remain bearable, and the desired result is a fatigue- and wear-resistant hard surface on the tough, ductile base material.

Thus, a precise heating technology providing high energy fluxes is crucial for successful case hardening. This is achieved by immersing the components in salt and metal baths, conductive or inductive heating, and heating by plasma, electron, and laser beams. The most important processes are described below.

Subsequent quenching must be performed with a high cooling rate, externally (immersion in or spraying with cooling media) or by the conduction of heat to the interior of the workpiece ("self-quenching").

Of the methods described here, flame hardening is the oldest, but is still widely used. The induction hardening process is the most widely used. Laser hardening and electron-beam hardening, in particular, are more recent developments. Table 4 lists some important process parameters for these techniques.

4.1.1. Flame Hardening

In flame hardening, heating is carried out with gas burners fuelled with a mixture of oxygen and natural gas, propane, or acetylene. Depending on the shape of the workpiece and on the fuel gas, various burner types can be used (e.g., nozzles, slots, and holes) that are arranged in a flat, circular, or contoured formation. Quenching is performed by water spraying or immersion in a vessel.

Table 4. Important process parameters

Parameter	Heating with gas burner	Induction	Laser beam	Electron beam
Energy flux, W/cm				
max.	6×10^3	10^4	10^{10}	10^8 – 10^{11}
used for hardening	$(1-6)\times10^3$	$10-10^4$	10^3 – 10^4 (10^5)	
Heating rate, K/s				
max.	10^2	10^4	10^{10}	10^9 (10^{10})
typical for hardening	10^2	$10-10^4$	10^3 – 10^4 (10^5)	10^3 – 10^4 (10^5)
Quenching	dipping or spraying	dipping or spraying	self-quenching	self-quenching
Thickness of hardened layer, mm	1.5 – 15	0.3 – 20*	0.01 – 1 (2)	0.1 – 1 (2)
Energy supply mode	continuous	continuous	continuous or pulsed	continuous or pulsed
Efficiency, %	50 – 70	50 – 70	70 – 90 (5 – 10)	70 – 80

* Dependent on frequency.

Two basic types of flame hardening process are used; a large number of variations exists. In *progressive hardening*, material is heated trackwise by a moving flame and is then immediately quenched; the burner and quenchant nozzle are designed as a unit [35, 36]. In *spot hardening* the entire surface area is first heated and then quenched. The problem of overlapping tracks in continuous hardening is discussed in Section 4.1.3.

Flame hardening provides case depths of 2 – 10 mm and is performed with a high degree of automation; it can still compete with more recent processes.

4.1.2. Induction Hardening

In induction hardening, heating is achieved by means of electromagnetic induction. An alternating magnetic field is produced when a current-carrying coil induces an alternating current within the workpiece. Direct conversion of electrical into thermal energy leads to heating of the workpiece. The highest temperature is attained in the surface layer owing to the skin effect, and it decreases exponentially towards the core. (The "skin effect" is a physical phenomenon — the current in an electrical conductor is not distributed uniformly throughout the material, but is concentrated in a thin surface layer.) Since the depth of penetration is inversely proportional to the frequency [35]:

Pulsed high frequency, 27 MHz 0.1 mm
High frequency, 400 – 2500 kHz 1 – 0.4 mm
Medium frequency, 3 – 10 kHz 8 – 5 mm

a broader range of case depths can be obtained by induction hardening than by flame hardening.

The shape of the inductor (the tool) matches the shape of the workpiece as closely as feasible; owing to the precise shape and design of the hardening tools, different case depths can be obtained in different areas of a workpiece. Because of the very specific and expensive tool, induction hardening is appropriate for bulk production of parts with complex topography. The inductor is made of pure copper or, for more demanding applications, of silver or sometimes silver-plated copper. Current densities in the inductor can amount to 6000 A/mm^2. Heat losses account for 10 – 25 % of the total power consumption, so internal cooling of the tool is necessary [36]. Quenching is performed with water jets or sprays, and, for small case depths, by self-quenching.

The adaptability of the process has led to the development of many variations, some of which are illustrated in Figure 13. Overlap problems always occur when the continuous-feed process is used (see Section 4.1.3).

During austenitizing, cementite particles of composition Fe_3C, i.e., 6.67 atom% carbon, must dissolve in a matrix with a carbon content of some orders of magnitude lower; diffusion in solids is slow, but a uniform carbon distribution must be achieved after a heating duration of some microseconds. Thus, the hardening result depends strongly on the as-received condition of the base material, that is, with respect to carbide precipitation, particle size, and spatial distributions. To overcome problems with less uniform substrate materials, operating temperatures can be much higher than the phase transition temperature A_{c3} [36]. Induction hardening is a fully developed, highly automated method that is perfectly integrated into production processes.

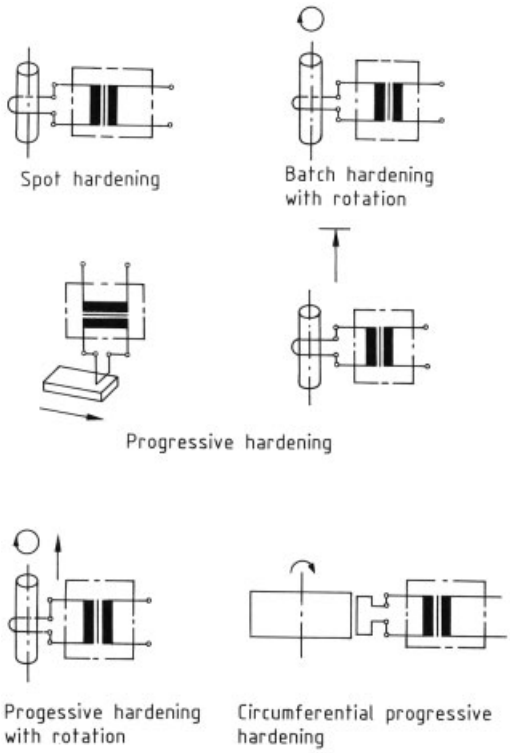

Figure 13. Methods of induction hardening [36]

4.1.3. Laser Hardening

A laser emits monochromatic, coherent electromagnetic radiation, and the laser beam exhibits nearly no dispersion. The energy density is high, so that an infrared laser is perfectly suited to heat a solid locally and rapidly to high temperature. For metallurgical purposes, carbon dioxide lasers (emission wavelength λ = 10.6 µm) are used. However, reflective metal surfaces absorb only less than 10 % of the emitted energy, which can be enhanced to 35 %. if the surface roughness is increased by, e.g., sand blasting. This is insufficient for practical application, so that black, absorbing coatings (e.g., sprayed graphite coatings) are applied which increase absorption to about 80 % of the emitted energy [37].

Methods such as spot hardening (Fig. 13) are not feasible by laser heating, because the laser beam must be permanently moved. With a continuously operated laser, any required geometrical pattern can be obtained with or without overlapping; pulsed lasers provide local hardening, as dots, small islands, lines, or circles.

However, even if the energy density of a laser beam is uniform along its cross section, and thus enables the corresponding area on the surface of a workpiece to be heated instantaneously to high temperatures, heat is transported within the metal along the temperature gradient by conduction, not only perpendicularly into the core, but also in parallel along the surface. In addition, the metallurgical reactions of steels are not only temperature-, but also time-dependent, especially with respect to rapid heating and cooling. Therefore, the individual laser tracks must not overlap, and the spaces between the tracks are crucial for the hardening result. Lack of process control leads to local overheating, heating durations or quenching rates that differ locally, and unwanted tempering and softening of previously hardened tracks. However, local heat transfer and phase transitions can now be modeled and simulated by finite element methods (FEM), so that individually adjusted motion patterns can be developed for the laser beam, to achieve the designed case depth (or case depth distribution) and microstructures. Consequently, laser hardening is now applied on industrial scale, and its use will presumably increase considerably. With accurate and precise process control complex hardening tasks can be fulfilled, and, in contrast to induction hardening, no expensive tool, individually manufactured for each type of workpiece, is needed.

4.1.4. Electron-Beam Hardening

In electron-beam hardening, impact of an electron beam on the workpiece causes the kinetic energy of the electrons to be converted to heat to a depth of 10 – 50 µm depending on the acceleration energy. Due to the nature of the electron beam, the process must be performed in a vacuum chamber, so that it is somewhat limited with respect to the size of the workpiece. An electron beam can be focused perfectly so that extremely fine or microscopic hardening patterns can be achieved, especially with the appropriate process control. Otherwise, problems of overlapping can occur (see Section 4.1.3). Hardening occurs by self-quenching, and tempering is usually avoided.

Electron-beam hardening is best carried out continuously and gives a case depth of 0.3 –1.5 mm at a track width of up to 50 mm [38]. Industrial application of this process has grown rapidly for mainly two reasons: 1) surface quality unlike in any other process is achieved because of the vacuum, so that the workpieces are not oxidized and remain bright and shiny; 2) high-frequency beam-deflection techniques open up possibilities never attained by other processes, for example, local hardening in workpieces of complex design, e.g. in deep cavities or bores. The treatment sites must not even be directly accessible from the outside.

4.2. Thermochemical Treatment

Thermochemical diffusion processes involve a heterogeneous chemical reaction at the interface of a metal substrate in a reactive environment that may be gaseous (including plasmas), liquid, or solid. The reaction products are in most cases of atomic scale, i.e., small nonmetal atoms, and thus able to diffuse into the substrate and thus modify the composition of the surface layer and the mechanical or chemical properties. Characteristically, the surface layer comprises both components from the reactive environment and from the substrate, and thus problems of defective layer adhesion are avoided.

The uptake of elements from the environment by diffusion can result in the formation of a solid solution with nonmetal atoms present at interstitial sites of the substrate's crystal lattice (diffusion layer), and/or in the formation of intermetallic compounds, in the form of precipitates (precipitation layer), e.g., at grain boundaries, or a closed and uniform surface layer (compound layer).

The thermochemical process follows a multistep mechanism including chemical reaction and transport of matter, mostly by diffusion. If chemical reaction is not kinetically hindered by the presence of passive surface films, diffusion, especially, diffusion in solids, is the slowest and thus rate-determining step. Consequently thickness and composition of the final surface layer depend on temperature and time, besides the concentration gradient and the diffusion coefficient. The residence time of the parts and workpieces in the furnace is generally used to control the resulting layer properties. Residence times can be in the range of minutes (e.g., 20 min) up to many hours (e.g., 100 h). Although higher temperature would enhance diffusion rates significantly, thermochemical treatments are performed at elevated temperatures between 400 and 1200 °C (mostly between 500 and 1000 °C), but within quite small temperature windows because modification of the substrate's microstructure (e.g., phase transition, grain growth, etc.) must be minimized.

The most commonly used thermochemical treatments — gas-phase nitriding and carburizing — are used for steels. Iron has high affinity for both carbon and nitrogen, and is reactive towards compounds like ammonia and hydrocarbons in gas phase. Provided the chromium content of a steel does not exceed 13 wt % (this limit depends on the content of additional alloying elements, especially C), it is not protected by a passive film which might significantly decrease the surface reactivity [39]. Other engineering metals and alloys, e.g., aluminum and titanium do not react readily in the gas phase because of their thermodynamically stable and impenetrable oxide layers. However, aluminum and titanium nitrides and carbides are known for their hardness and wear resistance. Both metals and their alloys can only be treated successfully by using plasma-assisted processes.

4.2.1. Gas Carburizing

4.2.1.1. Principles

Carburizing (carburization) increases the carbon content of the surface zone. Owing to high process temperature above austenitization temperature (A_{c2}) ferritic steels are transformed during the treatment. Diffusion of carbon into the surface zone of a steel with a low carbon content renders this zone hardenable, and it is then quench-hardened. The core of the workpiece remains unaffected, i.e., ductile. A hard, wear-resistant martensitic structure is produced in the surface zone. The process is used on "case-hardening steels", which are either unalloyed or only slightly alloyed, and contain $\leq 0.25\,\%$ carbon. Depending on the original material, the carbon content of the surface zone is 0.6 – 0.9

wt %. The usual layer thickness (the carburization depth, or, after hardening, the case depth) is usually 0.1 – 1.5 mm, less frequently up to 3 mm. As expected for a diffusion process, the carbon concentration decreases gradually from surface to core (Fig. 14, curve a). However, an initially constant carbon content and hence hardness are often required (curve b). This can be achieved by two-stage carburization (see Section 4.2.1.3). Small quantities of nitrogen are often allowed to diffuse along with the carbon (carbonitriding). The nitrogen decreases the temperature at which austenite formation occurs, and the diffusion process therefore takes place at lower temperatures. Workpieces that are susceptible to dimensional changes can then be more gently treated.

4.2.1.2. Properties of Carburized Components

Carburizing is one of the most widely used thermochemical processes. Treated parts have a tough, ductile interior and a hard, wear-resistant surface case with internal compressive stresses promoting layer adhesion and resistance to wear and corrosion cracking, thus improving fatigue strength and life, as well as load-bearing properties.

With carbon contents of > 0.5 wt % in the case after carburizing, some austenite remains after quenching because the end temperature of martensite formation is below room temperature. The amount of retained austenite depends on steel composition, carburizing conditions, and surface carbon content. Retained austenite reduces the internal compressive stresses in the case.

Furthermore, care must be taken to minimize oxidation when the carburizing atmosphere contains oxygen compounds such as water, CO_2, CO, or methanol. Intergranular oxidation occurs within a surface zone of 10 – 15 µm where oxides with aluminum, silicon, manganese, or chromium are formed, preferentially at grain boundaries. Thus, intergranular oxidation leads to microscopic "notches" which can initiate crack formation if the part is subjected to bending, so that fatigue resistance is affected. Furthermore, oxidation removes alloying elements, necessary for hardening, from the matrix. With high-load components, the oxidation zone must be removed by mechanical work.

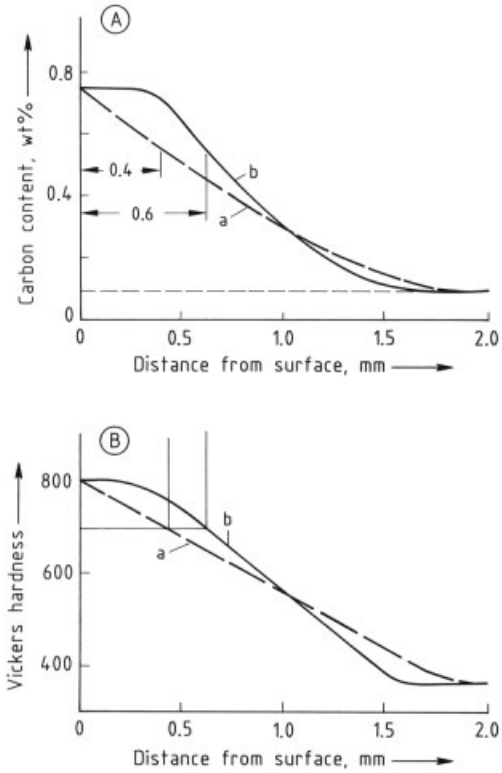

Figure 14. Carbon (A) and hardness (B) profiles after carburizing
a) One-stage carburizing; b) Two-stage carburizing

4.2.1.3. Carburizing Methods

The carburizing process is carried out mainly in the gas phase, to a lesser extent in a salt bath, and to a very small degree in powder. Developments are proceeding in two directions: use of plasma and use of low pressure, both combined with quenching in high-pressure gas.

Irrespective of the method, the carburizing mechanism is based on diffusion. The results depend not only on treatment temperature and time and the composition of the steel, but also on the characteristics of the carburizing media which affect carbon potential, carbon transfer factor, and carbon availability.

The *carbon potential* is defined as the percentage of carbon taken up by a test piece of pure iron in equilibrium with the carburizing medium

at a specified temperature. The *carbon transfer number* is the amount of carbon taken up from the carburizing medium by a defined surface area of the steel in unit time. The *carbon availability* is a measure of the amount of carbon in the carburizing medium per unit weight or volume.

Gas-Phase Carburizing. Three reactions are involved:

$2\ CO \rightleftharpoons (C) + CO_2$	Boudouard equilibrium
$CO + H_2 \rightleftharpoons (C) + H_2O$	water gas equilibrium
$CH_4 \rightleftharpoons (C) + 2\ H_2$	methane–hydrogen equilibrium

In these equations, (C) signifies the carbon dissolved in the austenite in equilibrium with the surrounding gas. When the carbon-releasing species, CO or CH_4, are not already present in the carburizing gas mixture, they are formed by decomposition in the carburizing furnace or in separate generators. The above equations are of fundamental importance for the measurement and control of the carbon level. Indirect methods of analysis are available such as IR absorption for the carbon dioxide and methane content and dew point determination for the water content. Increasing use has been made of oxygen-potential determination with an oxygen sensor. Direct methods include discontinuous determination of carbon using thin foils of iron, and continuous measurement of electrical conductivity changes in a wire owing to carbon absorption. The latter procedure permits continuous measurement of the most important parameter, the carbon potential [40]. Modern carburizing processes are computer-controlled and thus allow predefined carbon profiles with defined carbon contents, concentration gradients, retained austenite, and so on to be obtained and help to reduce carburizing times by at least 20 %. The profile illustrated in Figure 14 was obtained by using a very high carbon concentration to produce excess carburization of the surface. This was followed by a diffusion stage at a lower carbon concentration, during which the excess carbon leaves the highly carburized zone to give a greater depth of carburization (Fig. 15). This two-stage approach gave the desired carbon profile and shortened the time required for carburization [41].

Gaseous and liquid carburizing media provide such high carbon availability that the feed rate into the furnace must be controlled to avoid excessive carburization or soot formation. These problems can be avoided by using carrier gases that contain little or no carbon. The carbon concentration can be accurately adjusted by addition of carbon monoxide, methane, or propane. In industry the process gas used is "endo(thermic) gas", from the endothermic reaction of air and hydrocarbons, usually from natural gas. Endo gas has the following composition: 20 – 25 vol % CO, $\leq$ 0.5 vol % CO_2, 30 – 45 vol % H_2, $\leq$ 0.5 vol % CH_4, balance N_2. Processes based on nitrogen as carrier gas have been developed, mainly for economic reasons (i.e., shortage of natural gas or electric power) [42].

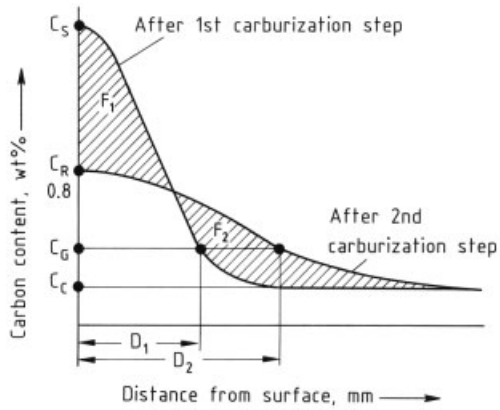

Figure 15. Carbon profile during two-stage carburizing [40] C_C= core carbon content; C_G= carbon content which determines case depth; C_R= desired case carbon content; C_S= saturation concentration of carbon in austenite; D_1= carburizing depth after first carburization; D_2= carburizing depth after second carburization; F_1= carbon distribution after first carburization; F_2= carbon distribution after second carburization

Liquids can also be used to produce carburizing gas. They can be added directly to the furnace chamber dropwise or by spraying. Alternatively, they may be vaporized in a reactor outside the furnace and led into the furnace as a gas. Lower alcohols (e.g., methanol, ethanol, or 2-propanol), ketones, and esters can also be used [35]. In the carbonitriding process, ammonia is also added as a source of nitrogen.

Past investigations into carburizing at reduced pressure employed methane as carburizing gas. Their aim was to reduce the carburizing time and hence to prevent intergranular oxidation of the case. These goals were not realized,

and the reproducibility and smoothness of the process were not satisfactory [43]. In new developments propane was used instead of methane, and the process was carried out at lower pressure. Acetylene is another possible carbon source. If the reaction gas is oxygen-free then quenching should also maintain oxygen-free conditions. Modern quenching media involve high-pressure gases, such as nitrogen, helium, or hydrogen [44].

As carbon diffusion in the metal controls the overall rate, and does not depend on the carbon source, the carburizing temperature is generally 930 – 950 °C. If the temperature is increased to 1000 °C, the diffusion coefficient of carbon into the steel is doubled, but the grain growth of the steel increases so much that the use of fine-grained steels provides only a partial remedy. (Coarse-grained steels are not as strong as fine-grained steels.) High-temperature carburizing has therefore only been used to a limited extent for large carburizing depths.

Carburizing is carried out in many types of furnaces (e.g., small shaft furnaces, multipurpose chamber furnaces with one or more chambers, and large continuously operating installations). A technical and economical solution is available for each kind of process, size of workpiece, and throughput.

Carburizing in Molten Salt Bath. In carburizing in molten salt baths, alkali metal cyanide is the most efficient carburizing agent. Much research has been aimed at substituting the cyanide with less toxic substances (e.g., carbides), but these salt baths are difficult to handle in practice, and give nonuniform results and poor reproducibility.

The decomposition reaction of cyanide in the salt bath is not yet fully understood, although it is assumed that carbon monoxide certainly acts as the carburizing agent. The reaction byproduct is carbonate.

In modern carburizing baths the cyanide is formed by decomposition, and the baths are initially set up from cyanide-free salts, with addition of a regenerator and an activator ($BaCl_2$) to maintain constant concentration of the active species. The regenerator (a nontoxic polymer) is added continuously to the bath and generates cyanide by reaction with the carbonate byproduct. The cyanide content can be held constant; waste salts are not produced. The achievable carbon content of the case depends on the activation (i.e., the alkaline earth metal chloride concentration). In practice the carbon content of the layer is adjusted by controlling the activator content of the strongly activated bath. Examples of trade names for commercial carburizing baths are Ceconstant and Durofer for regenerable baths, which both give carbon contents of 0.5, 0.8, and 1.1 wt % in the case. As in gas carburizing, a two-stage process aiming at specific carbon depth profiles is feasible, e.g., by performing two-thirds of the treatment in a highly activated bath, and one-third in a weakly activated bath [45].

When the carburized parts are subsequently treated in a warm salt bath, the adhering cyanide is destroyed by oxidation. When other cooling media (water, oil) are used, they must be detoxified afterwards. Modern automated installations include facilities for salt recovery, e.g., by treating the wastewater in an evaporation process.

Carburizing in Powder and Granular Material (Pack Carburizing). Only a small part of carburizing is performed with solid media (powder and granules). The process can be carried out without technical or economic difficulty, especially when a deep case is required. The solid medium consists of a carbon source (usually high-purity charcoal with low sulfur content) and an activator (usually carbonates of the alkali or alkaline earth metals). Activation probably occurs as follows:

$$BaCO_3 \rightarrow BaO + CO_2$$

$$C + CO_2 \rightarrow 2\,CO$$
$$CO + Fe \rightarrow FeC + \tfrac{1}{2}O_2$$
$$CO + \tfrac{1}{2}O_2 \rightarrow CO_2$$

Granules have two advantages over powder: powder tends to separate, and granules have higher gas permeability. Granules are usually formed by mixing the powder with a binder and compressing into shape. On heating the solid medium to 930 – 950 °C, carbon monoxide is formed as a carburizing agent; carburizing thus takes place via the gas phase.

Plasma Carburizing. The development of thermochemical processes involving reactive plasmas started with plasma nitriding, and is de-

scribed in Section 4.2.2.3. The industrial use of plasma carburizing was first reported quite recently [46]. Plasma carburizing is performed in a hot chamber, i.e., external heating is necessary, paraffinic hydrocarbons such as methane are used as reaction gases, and the quenching installation is integrated into the furnace. Controlled carbon distribution should be possible and treatment time is potentially shorter than that for conventional processes. As the reducing gas atmosphere does not contain oxygen compounds, it is possible to minimize intergranular oxidation.

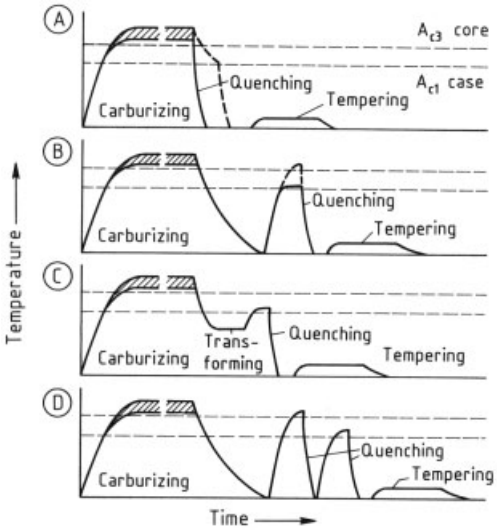

Figure 16. Hardening procedures used after carburization [47]
A) Direct hardening; B) Simple hardening; C) Hardening after isothermal modification; D) Double hardening A_{c1}, A_{c3} = transition points of iron

Hardening. The carburized steel is hardened by quenching in water, oil, or an oil – water emulsion according to one of the processes shown in Figure 16. The most economic method, whenever possible is direct quenching (A), in which the steel retains its original grain size. If, however, the material has a coarse-grained structure (e.g., due to long carburizing times and/or high carburizing temperature), it is first slowly cooled and then reheated (B) to convert it to fine-grained martensite; an isothermal transformation followed by quenching (C) may also be used. The double-hardening (D) process consists of initial hardening at austenite-formation temperature, and then at the lower temperature of the core; this method however is rarely used for economic reasons and because of large dimensional changes [47, 48]. After hardening, case-hardened workpieces are generally tempered at relatively low temperatures before use.

4.2.2. Nitriding and Nitrocarburizing

4.2.2.1. Principles

The term nitriding denotes the uptake of nitrogen into the surface zone of a ferrous metal workpiece. In general, this causes the formation of a case with an external *compound layer* which contains iron nitrides, namely, the ε-nitride, Fe_4N, and the γ'-nitride, $Fe_{2-3}N$. Below the external compound layer is a *diffusion zone* with an interstitial solution of nitrogen in the α-iron lattice; its thickness and concentration gradient depending on treatment temperature and residence time (Fig. 17). As nitriding is performed at slightly lower temperatures than nitrocarburizing the ε-nitride is usually only found after nitrocarburizing. Some alloying elements such as Al, V, Mo, Ti, and Cr form nitrides of their own in the compound layer or enable the formation of nitride precipitates in the diffusion (or precipitation) zone.

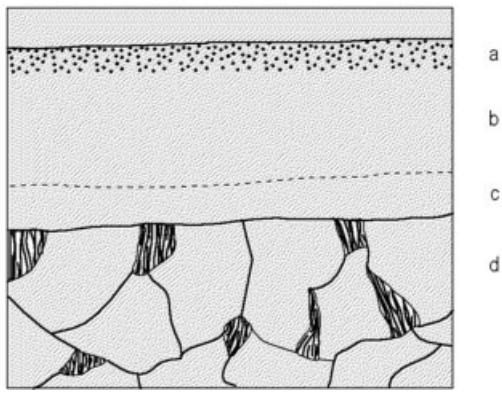

Figure 17. Nitrocarburized layer on an unalloyed steel (schematic)
a) Porous zone; b) ε-$Fe_{2-3}N$ layer; c) γ'-Fe_4N layer; d) Diffusion layer (solid solution of nitrogen in iron)

γ'-Iron nitride Fe_4N has a cubic structure and contains 5.7 – 6.1 wt % nitrogen; ε-iron nitride $Fe_{2-3}N$ has a hexagonal structure and contains 7.7 – 11.1 wt % nitrogen. In a dual-phase compound layer γ'-nitride often forms a narrow border between the ε-nitride and the base material or the diffusion zone, which is difficult to identify metallurgically or by X-ray techniques.

In nitrocarburizing both nitrogen and carbon are made to diffuse into the surface layer of a workpiece (see Section 4.2.2.3). The compound layer then additionally contains ca. 0.5 – 2 wt % carbon. Oxygen can be present in amounts up to several weight percent in nitrided or nitrocarburized cases.

The outer part of the compound layer can exhibit a porous structure, like a band of microscopic or submicroscopic pores, or cavities. Some of the cavities are joined to one another and to the surface, but most of them form isolated spheroidal voids with a diameter of < 1 µm. The mechanism of pore formation is not yet fully understood, but higher reaction rates, higher temperatures, and higher nitrogen availabilities must be involved, because pores are found preferably in the ε-nitride after nitrocarburizing, rarely in the γ'-nitride after nitriding, and never in the diffusion layer after prolonged nitriding. Nitriding and nitrocarburizing processes can be controlled by means of the nitrogen-source concentration, reaction temperature, and residence time.

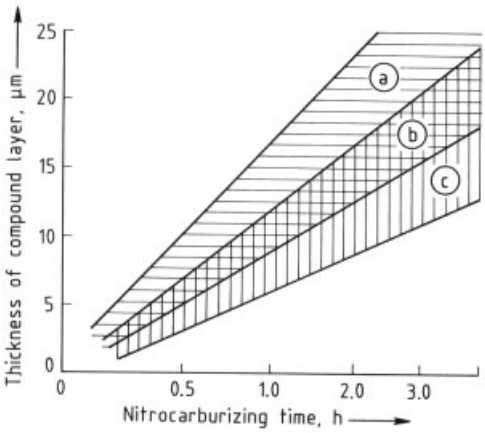

Figure 18. Thickness of compound layers in ferrous materials after nitrocarburizing in a molten salt bath at 580 °C
a) C 15 – C 45 steel; b) Quenched alloy steel; c) Hotworking 12 %-chromium steel, cast iron

The thicknesses of compound layer and diffusion zone are a function of the treatment process, the treatment parameters, and the base material. Figures 18 and 19 show the depths of these layers in various materials as a function of the nitrocarburizing time.

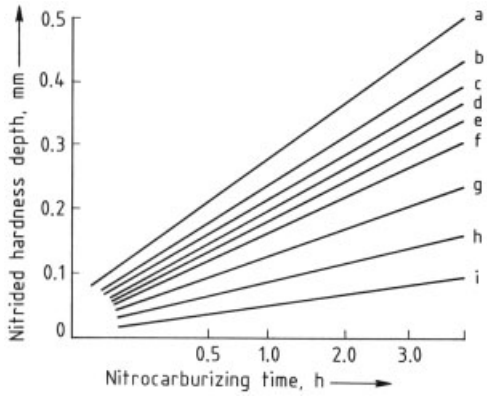

Figure 19. Case depth of steels after nitrocarburizing in a molten salt bath at 580 °C
a) C 15; b) C 45; c) 34 Cr 4; d) 42 CrMo 4; e) 50 CrV 4; f) 34 CrNiMo 6; g) 34 CrAlMo 5; h) X 210 Cr 12; i) Cast iron

4.2.2.2. Properties of Nitrided and Nitrocarburized Components

Nitriding and nitrocarburizing are quite simple processes. The solid solution of nitrogen in the diffusion layer is seen as responsible for increased fatigue strength, life, and endurance limit, whereas the compound layer provides excellent wear and some corrosion resistance. Consequently, the prominent goal is to produce surface layers perfectly adapted to the technological demands of the workpieces during operation, by improved process control. This can be achieved basically in one nitriding furnace by variation of the process parameters, so that the process result is solely a diffusion zone, or a diffusion zone in combination with a single- or dual-phase compound layer. In contrast to carburizing, in which the carbon content of the surface region is increased in the first step at a temperature above the transformation temperature of the steel, and the case is subsequently hardened by quenching, it is not necessary to quench nitrided parts, basically because no comparable

phase transformations are needed. In addition, distortion is less of a problem for the same reason, namely, the lower reaction temperature.

Fatigue stength and endurance life are improved mostly in the diffusion layer, where hardness increases gradually depending on the local nitrogen content, and with specific compressive stress distribution. It is plausible that the effects of roughness peaks, microcracks, and voids (and the accompanying local stress maxima) are smoothed out during nitrogen uptake by the slow expansion and growth of the diffusion zone. Consequently, fatigue strength and life, especially of notch-sensitive steels, are improved.

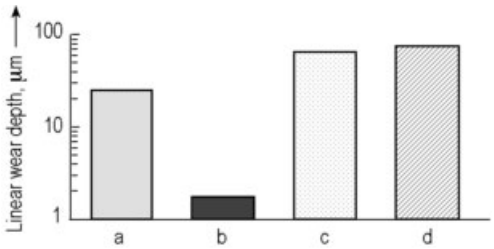

Figure 20. Results of laboratory wear tests (adhesive wear, pin – disk method) [49]
a) C45, hardened and tempered; b) 42CrMo4, nitrocarburized; c) 42CrMo4, borided; d) C100, vanadized

Adhesive wear is caused by metal-to-metal contact and cold welding, and is largely responsible for wear losses, e.g., in machines and motor vehicles. This type of wear can be very effectively combated by the nitride compound layers obtained by nitrocarburizing. Figure 20 shows the results of laboratory wear tests. In addition, nitrocarburizing also improves fatigue strength, the best results being obtained for components with unalloyed or low-alloy steels with high notch sensitivities. Finally, nitrocarburizing in combination with controlled cooling or quenching in specific atmospheres or media provides corrosion protection that can surpass that of electroplated surfaces or electroless coatings. Figure 21 shows the results of some corrosion tests [50].

In industrial application, nitriding and nitrocarburizing processes seem to de developing in different directions, for different demands and applications. Whereas nitriding is applied to form diffusion layers while avoiding nitride formation, for improved fatigue properties, nitrocarburizing is used to form compound layers providing wear and corrosion resistance. Thus, industrial nitriding processes are performed at lower temperatures and at residence times orders of magnitude greater than nitrocarburizing processes. The aim of nitriding is to obtain diffusion layers without compound formation, whereas the objectives of nitrocarburizing are highly specific single- or dual-phase compound layers without any pores.

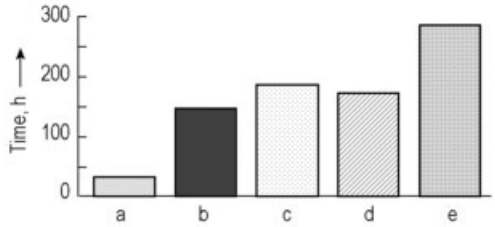

Figure 21. Results of laboratory corrosion tests (salt spray test, DIN 50 021, ASTM B117; time to appearence of first rust spot)
a) 40 µm hard Cr layer; b) 20 µm hard Cr + 20 µm soft Cr; c) 20 µm Ni + 20 µm hard Cr; d) Gas nitrocarburized, oxidized; e)Nitrocarburized in salt bath, oxidized.

4.2.2.3. Nitriding and Nitrocarburizing Methods

All nitriding and nitrocarburizing processes are carried out at fairly low temperatures (450 – 600 °C). Dimensional changes are consequently considerably smaller than with, for example, the carburizing process, which is carried out above the transformation temperature. Originally the maximum treatment temperature for nitriding or nitrocarburizing was the A_1 temperature in the iron – nitrogen system (590 °C) at which α-iron is transformed into γ-iron. Above this temperature an extra phase is formed below the compound layer which consists of nitrogen pearlite (braunite), nitrogen austenite, or nitrogen martensite depending on the base material and quenching conditions. More recently, this limit has been exceeded to reduce the treatment time and make better use of available plant capacity. Although the effect of the extra phase under the compound layer is not yet fully understood, the process is frequently operated at 600 – 650 °C. A precondition for this method is that the components must not be susceptible to distortion.

In all processes, nitrogen must be available in a diffusible form, i.e., the atomic state formed by dissociation, which can be thermal (gas-phase dissociation of ammonia), plasma-assisted (molecular nitrogen), or liquid-phase decomposition of alkali metal cyanate, catalyzed by the metal interface.

Gas nitriding is performed in mixtures of ammonia, nitrogen, and hydrogen, with ammonia being the reactant. The compound layer consists of nitrides of iron and alloying elements with no carbon; it is therefore brittle and unsuitable for many applications. The main aim of gas nitriding is to increase the bearing capacity of the surface zone of the workpiece while maintaining high core strength and small dimensional changes. Treatment is carried out batchwise in externally heated vessels at 480 – 520 °C for 30 – 100 h. Ammonia is catalytically decomposed into atomic nitrogen and hydrogen on the surface of the workpiece.

Appropriate process control enables the formation of a compound layer to be prevented by controlling the composition of the atmosphere. The control parameter is the nitriding number (or index) which is related to the equilibrium constant of ammonia formation (as discussed in the Section page 29). Gas nitriding without compound-layer formation is performed at low nitriding numbers. Thin compound layers formed nevertheless are removed by machining.

Process control is not simple because the composition of the nitriding atmosphere at a given temperature and total pressure depends on gas flow rate and on the catalytic surface area, which includes not only the steel parts but also the interior surface of the furnace; the latter contribution depends on the specific design and condition of the furnace.

Any steel can be nitrided in the gas process, but the achievable surface properties and the nitriding parameters depend strongly on its chemical composition. Thus construction steels with specifications without respect to composition will not exhibit uniform surface properties after nitriding unless they are from the same batch. Special nitriding steels (i.e., steels alloyed with nitride-forming elements such as aluminum, titanium, and chromium) are not widely used. [52]. Workpieces are nitrided in a hardened and tempered condition. Both heating and cooling are performed slowly, so that distortion and dimensional changes are minimized.

The objective to shorten nitriding time and to prevent the formation of compound layers has led to the development of multistage processes. For example, in a two-stage process [52], the first stage uses a nitriding medium with a high concentration of nitrogen, and the second a much lower concentration. Oxygen, water, or water vapor can be added to furnace atmosphere to activate or depassivate the surfaces.

Salt bath nitrocarburizing is performed in a molten salt mixture, typically containing an alkali metal cyanate (NaOCN, KOCN) as active component and 33 % of sodium or potassium carbonate. The cyanate decomposes thermally by the catalytic action of the surface of the workpiece, forming atomic nitrogen which can diffuse through the surface into the workpiece. Carbon in diffusible form is formed simultaneously. The byproduct of the reaction is alkali metal carbonate. Addition of a "regenerator" converts the byproduct carbonate back to cyanate, usually in a continuous process. The regenerator can be a melamine derivative or a urea polymer, though the latter is less active [53]. A side reaction between the cyanate and the workpiece generates a small quantity of cyanide, which must be removed during or after the process (see below).

Treatment is carried out at 570 – 580 °C for 45 – 180 min. The trend is towards higher treatment temperatures. Case depths produced under these conditions are shown in Figure 18.

Gas treatment processes have been highly automated for many years. Automation started later with salt bath processes but has now reached a similar level. A salt bath nitrocarburizing installation consists of precleaning, preheating, nitrocarburizing, quenching, and final cleaning steps. Precleaning is important for the success of the process. Although the salt bath is less sensitive than the gas process in this respect, residues from treatment liquids and washing media can lead to erratic results. In the next stage, the workpiece is preheated to ca. 300 – 350 °C. This avoids too great a reduction in the temperature of the bath when a charge is added to it, and it also ensures that the workpieces are dry when they are introduced into the salt bath. The next stage is the nitrocarburizing bath; several baths may be used to improve the flexibility of

the process. The salt baths are contained in crucibles made of pure titanium or nickel-based materials. Air is blown through the bath to ensure complete mixing and hence a constant composition and temperature. Quenching may be carried out in various ways. Unalloyed steels reach their maximum fatigue strength after quenching in water. Alloy steels are less sensitive in this respect. The quench water must be treated to eliminate its toxic cyanide content. Components which are extremely sensitive to distortion can be transferred from the nitriding bath to a vacuum chamber.

An oxidizing hydroxide bath containing nitrate has been developed for a commonly used quenching process carried out in the temperature range 330 – 400 °C. This bath immediately converts the cyanide and cyanate adhering to the workpiece to carbonate so that it is no longer necessary to detoxify the water used in the subsequent cooling and washing stages. A further advantage is that the stepwise and hence very mild quenching further reduces the tendency to dimensional change. Finally, treatment in the oxidation bath causes surface oxidation of the compound layer, forming a thin layer of Fe_3O_4 which gives good corrosion properties [56].

The corrosion-resistant property of the oxidized compound layer can be improved by a final surface treatment (e.g., vibratory grinding, polishing) and further final oxidation.

The salt bath nitrocarburizing processes described here have become known under the trade names Tenifer, Tufftride, and Melonite. The Sursulf process is also based on the use of alkali metal cyanate and carbonate with small amounts of cyanide. Sulfur is added to the bath to improve the running properties of the workpieces produced. However, the porosity is greatly increased by sulfur addition [55].

Gas Nitrocarburizing. Gas nitrocarburizing involves the diffusion of not only nitrogen but also of carbon, which contributes to the formation of the compound layer. The source of nitrogen is ammonia. To compare and evaluate different nitrocarburizing atmospheres, the nitriding index N has been introduced [57]:

$$N = \frac{p_{NH_3}}{p_{H_2}^{1.5}}$$

The equation for the decomposition of ammonia

$$2\,NH_3 \rightarrow N_2 + 3\,H_2$$

has the equilibrium constant

$$K = \frac{p_{NH_3}^2}{p_{N_2} \cdot p_{H_2}^3}$$

which is related to the nitriding index N by the equation

$$N = K^{0.5} \cdot p_{N_2}^{0.5}$$

An example of the use of the nitriding index N is given in Figure 22, which shows nitrocarburizing conditions in the ternary system Fe – N – C. Process control via the nitriding index N is finding increasing use in industrial practice.

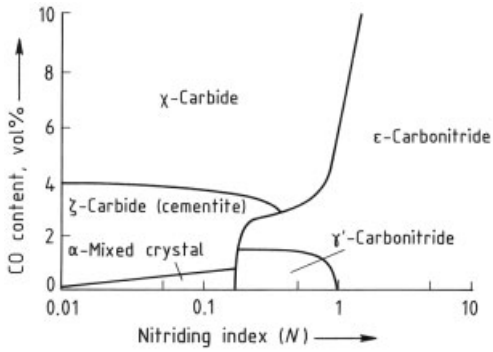

Figure 22. Ranges of formation of nitrides and carbides in the ternary system Fe–C–N at 575 °C [52]

The gases added to the ammonia as a source of carbon can be divided into three groups:

1) Carbon dioxide and carbon-dioxide-rich mixtures
2) Carbon monoxide and carbon-monoxide-rich mixtures
3) Carbon-rich gases

Multistage processes with atmospheres specific for each stage are used in industry.

Carbon dioxide is very suitable as the added gas for nitrocarburizing. For a given degree of dissociation of the ammonia, it increases the nitriding index; carbon is strongly taken up by the compound layer owing to the high carbon activity; and formation of the compound layer is much facilitated. Besides pure carbon dioxide, carbon dioxide – nitrogen mixtures or exo(thermic) gas are often used in practice, sometimes with additional carbon monoxide

and/or hydrogen. Exo gas denotes the reaction product of the near-stoichiometric combustion of hydrocarbons, e.g., natural gas. Approximate compositions are 0 – 3 vol % CO, 10 – 13 vol % CO_2, 0 – 4 vol % H_2, 2 – 3 vol % H_2O, balance N_2. Carbon activity increases with increasing carbon monoxide content.

In theory, carbon monoxide could be used as carbon source, and many references in the literature [57] refer to the use of carbon-rich carrier gases (e.g., methane, propane, and amines). However, neither method is used in practice. Carbon monoxide is added mainly in the form of endo gas, or is produced by pyrolysis of methanol. A commonly used endo gas composition is ca. 20 vol% CO, 35 vol% H_2, and 45 vol% N_2 as a 1:1 mixture with ammonia. The composition of the atmosphere under nitrocarburizing conditions in the furnace is far more complex and difficult to determine.

Gas nitrocarburizing is usually performed at 570 – 580 °C (like the salt bath process), but here also a trend towards higher temperatures exists. The treatment time is also similar to that of the salt bath process, i.e., 90 – 240 min. In general, the gas process takes longer on account of the slower heating.

Gas nitriding is mainly carried out in single-chamber furnaces and batch processes because of nitriding durations around 100 h. In gas nitrocarburizing more types of furnaces are in use, e.g., multiple-chamber installations including quenching chambers, which can be used for other treatments, e.g., carburizing. Installations are operated batchwise or continuously, the latter with residence times down to several minutes. These installations are single-purpose, mostly pusher-type furnaces. Throughput is then in the range of 30 t/d or more.

In all heat-treatment operations, aspects of occupational health and safety must be taken into account. Parts must be dry, and may under no circumstances contain water residues, e.g., enclosed in voids or bores. During gas nitriding or nitrocarburizing with ammonia, ignition may occur if the oxygen concentration in the furnace exceeds the explosion limit and an ignition source is present. However, this danger is difficult to assess because of the lack of information on the composition of the atmosphere. In practice, heating of the parts to process temperature in a batch furnace is performed in an inert atmosphere (i.e., nitrogen) so that the furnace is flushed; continuous processes apply a flame curtain to prevent any oxygen from entering. The exhaust gases are treated by combustion.

Process and furnace development for gas processes depend on developments in measuring and control. As the actual composition of the atmosphere depends on a multitude of parameters, including batch size as well as chamber design and condition, gas analysis of the multiple components is not feasible in practice. Thus various probes and sensors have been developed to assess control parameters such as the nitriding index. Gas probes based on various measuring principles, such as pressure measurement or electrochemical, are available in the market.

Gas-phase processes can be sensitive to passive metal surfaces. This refers to passive surface films, often present on steel parts containing 5 wt % or more chromium. Manufacturing residues (from lubricants, cleaning media, greases, and oils) can present problems, especially if they or their components are nonvolatile at process temperature. The removal of surface contaminants by diligent cleaning prior to heating is essential.

Gas nitrocarburizing processes are referred to by company names that have no relation to the process itself. Examples are Nitroc (ammonia and exo gas, Aichelin, Germany), Nitemp (ammonia and endo gas, Ipsen, USA), Triniding (ammonia and nitrogen, Midland-Ross, USA), and Deganit (two stages: ammonia and exo gas, then ammonia and endo gas, Degussa, Germany).

Plasma Nitriding and Nitrocarburizing. Plasma nitriding is based on the formation of the active, diffusible species from molecular nitrogen by plasma discharge. Unlike nitriding and similar to nitrocarburizing, relatively tough compound layers are formed which are said to be less prone to pore formation than in nitrocarburizing. Thus the plasma-assisted processes aim at the formation of a compound layer, independent of the presence of a carbon source in the reactive atmosphere.

Basically, the workpiece is the cathode inside a vessel which acts as the anode. The reaction gas in the previously evacuated container may simply consist of nitrogen at a pressure of 0.05 – 1 kPa. A potential difference of ca. 100 V is

applied between the vessel and the workpiece, producing a visible glow discharge seam surrounding the surface of the workpiece. Nitriding is assumed to proceed according to the mechanism presented in Fig. 23. The active species consists of reactive atomic and ionized nitrogen, present in the glow discharge seam or boundary layer near the metal surface. Owing to the directed movement of nitrogen ions in the electric field and depending on the potential difference in the boundary layer, iron atoms and ions become mobile and react to form nitrogen-rich iron nitrides, which are assumed to condense at the gas – solid interface. Since these intermediate nitrides are unstable, they decompose and release atomic nitrogen to diffuse into the substrate or to be desorbed into the atmosphere. The nitriding process thus is assumed to consist of cycles of removal and deposition. Direct diffusion of nitrogen from the plasma is said to occur only to a small degree, if at all.

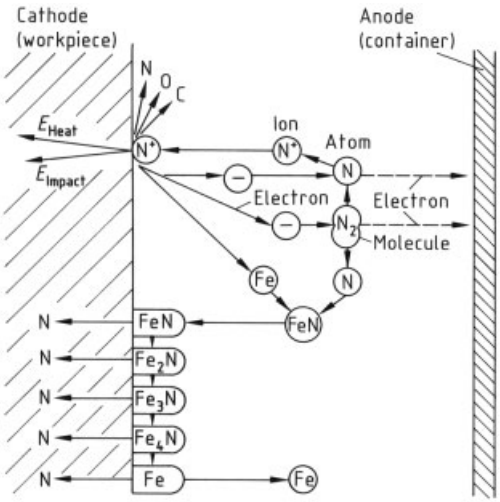

Figure 23. Plasma nitriding reactions [58]

In a d.c. plasma the substrate reaches nitriding temperature (i.e., the temperature needed for diffusion) without additional external heating (cold wall, cold chamber). However, today, because of multiple advantages with respect to uniform layer formation, pulsed plasmas (i.e., plasmas produced with pulsed direct current) are applied. The batch then requires external heating (e.g., by heating the walls of the vessel: hot wall, hot chamber) [59, 60]. Pulsed processes require computer control.

Compared to other nitriding processes plasma nitriding is exemplary for the great variety of modifications with respect to structure and properties of the hard layer. Thus, diffusion layers without any compound layer can be produced, or single-phase compound layers, or dual-phase mixed layers. Plasma nitriding is performed at 350 – 650 °C for 10 min to 48 h. At lower temperatures high core strengths are maintained; dimensional and shape changes are small. Temperatures above 590 °C can be applied, e.g., when thick single-phase γ'-layers are required. Reaction gas is nitrogen with or without methane, which must be supplied premixed from gas bottles or gas-mixing equipment. The gas composition can be adjusted during the process between 80 and 100 % nitrogen.

Similar to the gas processes without plasma assistance, parts and workpieces must be clean to obtain results without defects. Even if plasma activation in argon or oxygen plasmas is applied, passivation owing to thick layers of contaminants or passive oxide cannot be overcome in reasonable time, and there is the danger of contamination of the reaction vessel. Thus, the parts must be thoroughly cleaned prior to plasma treatment, preferably with aqueous cleaning agents. However, passive oxide layers, such as those on stainless steels, are easier to overcome than in the gas processes. Metals and alloys which cannot be treated in the gas phase because of very stable and dense oxide layers, like aluminum and titanium, can be nitrided in reactive plasmas.

Installations for plasma nitriding and nitrocarburizing are relatively simple and consist of a vacuum-tight cylindrical vessel, a system for supplying and mixing the gas, a vacuum system, and an electric power unit. Gas composition is usually controlled not in the reaction vessel but by careful mixing and metering of the gas components.

Up to now, a drawback of the plasma process was the limited size of the vessel and therefore, the workpiece, and the assembly of large batches of bulk parts. Nevertheless, plants are increasingly becoming available, and continuous setups are in development.

Future Developments. The use of nitriding and nitrocarburizing of ferrous components increased considerably during the 1980s, and this trend seems to continue. One reason is that in general more attention is paid to specification of the properties of the parts and the contribution of their surfaces. Especially the relatively low process temperature reduces costs for installation and operation. Recently, even complex processes have become manageable owing to automation, process simulation, and computer control.

The gas nitrocarburizing process accounts for ca. 40 % of the worldwide total, as does the molten salt process. The "classical" gas nitriding process continues to have a constant share of ca. 5 – 10 %. The plasma process makes up the remaining 10 – 12 % but still has some potential for growth.

Further development of the nitriding and nitrocarburizing processes will be in the direction of automatic operation, which will improve both economic and technological viability. The application of nitriding can be stretched further, e.g. to the conditioning of metal surfaces for plasma coating processes, or to obtain improved corrosion resistance. In the molten salt process compliance with environmental requirements is very important. Development is in the direction of closed systems in which spent salts, toxic wastewater, or other waste materials do not enter the environment. Similar environmental considerations apply to the gas processes.

4.2.3. Boriding

4.2.3.1. Principles

When boron diffuses into the surface zone of a ferrous workpiece, a compound layer is formed that consists of either one iron boride (Fe_2B) or two iron borides (Fe_2B in the inner, FeB in the outer layer). A diffusion zone containing boron as a solid solution is formed below the boride layer. The compound layer is strongly interlocked with the metal substrate, especially in the case of unalloyed or low-alloy steels (Fig. 24). With increasing content of alloying elements, however, this effect decreases; it is almost absent in high-alloy steels [61].

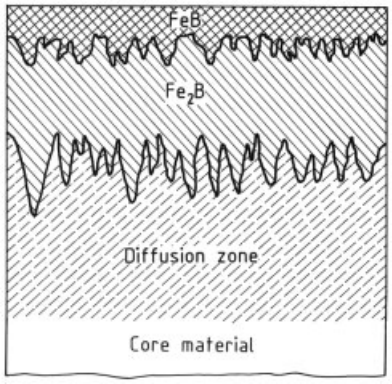

Figure 24. Schematic representation of boride layers

The two iron borides forming the compound layer ("boride layer") have different properties. The boron-rich rhombic FeB is harder and more brittle than the tetragonal Fe_2B with its lower boron content. Single-phase Fe_2B layers have much better properties than dual-phase layers. Different layer types are shown in Figure 25; those of types E and F are preferred in industrial practice. This cannot always be realized, especially with thick layers, and type D layers then must be accepted [62].

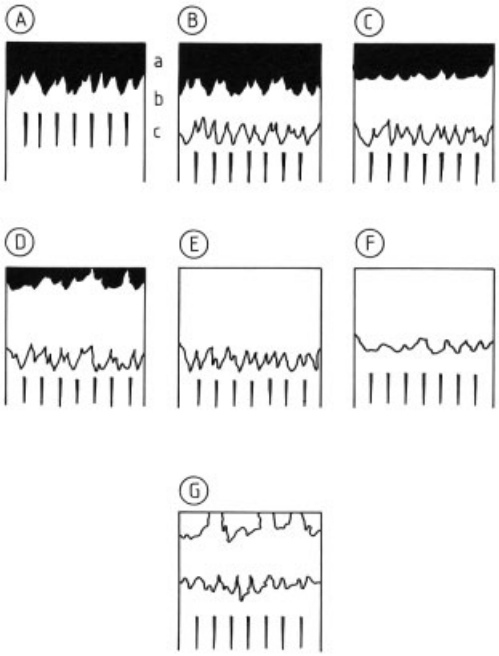

Figure 25. Types of iron boride layers (adapted from [62]) a) FeB; b) Fe_2B; c) Diffusion zone

The Vickers hardness values of the boride layer cited in literature vary between 1500 and 2300 [61]. Excellent wear properties are of higher importance, especially with respect to abrasive wear. Good results are also reported for high-speed wear and combined wear and corrosion [61].

Boron dissolved in the diffusion zone produces a considerable increase in hardenability. Since heat treatment can only be carried out after boriding, the higher hardenability results in improved load-bearing properties of the matrix material. Carbon is not soluble in the boride layer; it is substituted in the growing layer and enriched underneath. The increased carbon content contributes further to hardenability and load-bearing properties.

All types of ferrous materials are suitable for boriding, including sintered and cast products. Prolonged boriding of materials with high aluminum or silicon content ($> 0.5\,\%$) must be avoided because this leads to the formation of ferrite and and unsatisfactory cases (type G in Fig. 25).

The thickness of the borided compound layer can be varied between very wide limits: 20 – 30 µm layers are often used in machine manufacturing, while on molds for brick-making the layer can be up to 800 µm thick.

4.2.3.2. Boriding Methods

Boriding (boronizing) can be carried out in gaseous, liquid, and solid media. Although *boriding in gaseous media* has often been reported in the literature, no industrial application is known. Boron trihalides (e.g., BCl_3) and boron hydride (B_2H_6) have been proposed as gaseous media. The trihalides must be mixed with hydrogen or nitrogen. Diborane (B_2H_6) is extremely toxic and this has so far prevented its use [63].

Boriding in liquid media, (i.e., molten salts) has also been frequently reported [62] and is mainly employed in East European countries. The process is either electrolytical in molten borax or electroless in an inert melt of alkali metal chlorides containing boron carbide as boron source. Drawbacks including corrosion of the vessel, purity problems, and irreproducible results have prevented this method from becoming widely established.

Boriding with solid materials is performed with a mixture of boron carbide, an activator (frequently KBF_4), and a diluent (e.g., SiC, Al_2O_3, MgO) that allows adjustment of the boride potential. The parts to be treated are packed in this boriding medium, heated to 900 – 950 °C, and usually cooled in situ. If necessary, boriding can be followed by heat treatment. The powder can be replaced by an aqueous paste of similar composition; smaller amounts of boriding media are then required, and treatment of part of a workpiece is also possible. Boriding with powder can be carried out in air, but a protective gas (e.g., a mixture of 5 – 10 % hydrogen and 95 – 90 % nitrogen, or argon) is required for the paste.

Treatment time depends on the layer thickness required: for example, a boride thickness of 100 µm on a C45 steel requires 3 h at 900 °C (powder) or 2 h at 900 °C (paste). Single items and small series are treated in a chamber furnace; the pit-type furnace is also used for paste boriding. Large-scale boriding is carried out in a continuous furnace and can be almost fully automated.

Exploitation of the good properties of iron boride layers still fails in many cases due to the inadequate boriding methods. In the last few years much development work has been aimed at improving these processes, in particular the application of the plasma process to boriding. It remains to be seen whether this will be successful [64].

4.2.4. Other Thermochemical Processes

In addition to the processes described in Sections Case Hardening with Carburizing, Nitriding and Nitrocarburizing, Boriding a few other methods exist for special applications [36, 65]. They are rarely used.

Aluminizing. Diffusion of aluminum into the workpiece produces an aluminide phase which is resistant to oxidation up to 800 °C, especially if sulfur was present in the atmosphere.

Chromizing. Carbon steels are subjected to chromium diffusion to form chromium carbide cases with high wear resistance.

Siliconizing. Diffusion of silicon leads to solid solutions with high acid resistance.

Vanadizing. Diffusion of vanadium into carbon steels produces vanadium carbide cases with high abrasion resistance. Vanadizing is performed by a powder process or in molten borax with ferrovanadium (Japanese TD process) [66]. The TD process is still used to a small extent in Japan and several other countries for treating metal tools.

5. Thermal Spraying

All thermal spraying processes require a concentrated, high-energy heat source into which the coating material is fed in the form of wire, rod, or powder. The material melts and is impelled at high velocity onto the surface of the workpiece in the form of spray particles. Impact at the surface causes flattening of the liquid droplets and cooling as a result of heat transfer to the substrate. The particles then solidify and shrink. The temperature of the workpiece can be kept low during the coating process, so that even temperature-sensitive materials (e.g., low-melting light metals or plastics) can be deposited [67].

Many coating materials and methods can be used for thermally sprayed protective coatings. Research and development continually lead to new industrial applications. For the successful solution of technical problems, the complete system, including coating material, substrate material, and coating method must be considered [68]. Owing to the wide range of possibilities for combining substrate and coating material, thermal spraying processes find a large number of possible applications with increasing use in industry.

5.1. Methods of Thermal Spraying

5.1.1. Flame Spraying

The oldest thermal spraying processes employ a fuel gas – oxygen flame as heat source, which is used to spray metals and alloys and some ceramic materials. The material to be sprayed is fed into the flame where it melts, is broken down into fine particles (atomized), and sprayed onto the surface of the workpiece by the pressure of the flame and the carrier gases, sometimes using an additional atomizing gas such as compressed air (Fig. 26). Acetylene, propane, or hydrogen are used as fuel gases. The type of fuel gas determines the flame temperature, and the maximum temperature (3150 °C) is obtained with acetylene – oxygen. The kinetic energy of the combustion gases and the sprayed particles depends on the heat released during combustion and on the gas pressure.

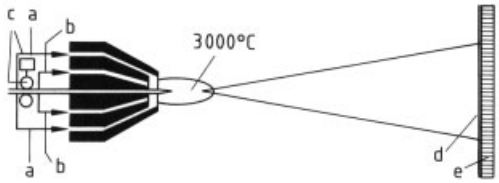

Figure 26. Principle of flame spraying
a) Atomizing air; b) Acetylene – oxygen mixture; c) Wire feeding device; d) Sprayed coating; e) Workpiece

Metallic materials are usually sprayed in the form of powder and wire. Ceramic materials are fed into the flame as powder or as rods formed by sintering the powder.

In a special type of flame spraying technique, nickel, cobalt, or iron alloys are used in powder form. The powder is sprayed onto the workpiece in the first step and then fused with the substrate in the second step. Owing to the very good wetting properties of these alloys, an enormous increase in adhesive strength is achieved; furthermore, the coatings become virtually nonporous after the fusing stage. Fusing can be either carried out manually with an autogenous flame, by inductive heating if the coated workpieces are rotationally symmetrical, or in a vacuum furnace if large numbers of parts are to be treated [69]. Figure 27 shows an abrasion- and corrosion-resistant Ni–Cr–B–Si alloy which has been flame sprayed and fused onto a mold core for glass products in glass industry.

A further development in flame-spraying technology, available on the market since the 1980s, is high-velocity flame spraying (HVOF: high-velocity oxygen–fuel spraying). The common feature of all HVOF processes is the increased pressure of the fuel gases, which results in an enormous increase in the velocity of the sprayed particles. In contrast to the conventional

flame-spraying procedure, combustion occurs in an extended, sometimes cooled combustion zone [70]. The higher kinetic energy of the spray gases leads to higher velocities of the spray particles, producing much denser coatings with much better adhesion to the substrate.

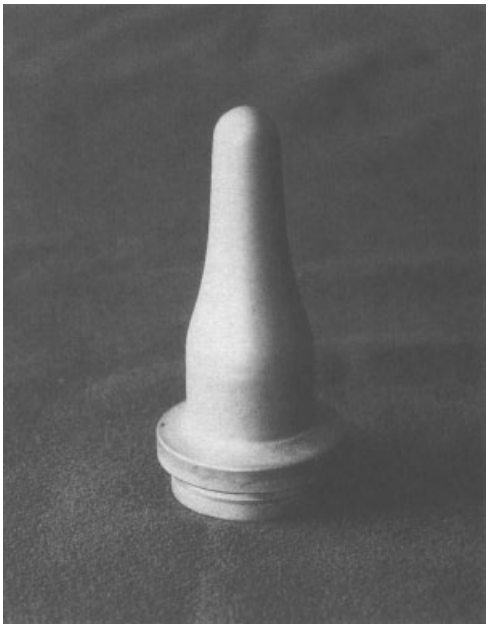

Figure 27. Mold core for raw glass preparation in the glass industry, coated with a self-fluxing Ni–Cr–B–Si alloy

5.1.2. Electric Arc Spraying

In electric arc spraying, two electrically conducting pieces of wire are introduced at constant, controlled rate and predetermined angle into an electric arc spray pistol (Fig. 28). A voltage is applied and the arc is ignited between the two electrodes, giving a temperature of ca. 4000 °C and causing both wires to melt. An auxiliary gas jet from a nozzle atomizes the molten material, accelerates the particles, and causes them to impinge on the substrate. The controlled feed rate ensures uniform melting of the wires. If wires of different metals or wires filled with hard phases are used, so-called pseudo-alloys are formed. The term pseudo-alloy refers to alloys that are not in thermodynamic equilibrium, and is often used to denote hard phases (e.g., carbides, borides) that are embedded in a metallic matrix made by spraying cored wires.

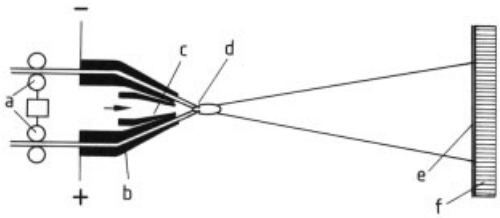

Figure 28. Principle of electric arc spraying
a) Wire feeding devices; b) Wire guides and electrical contacts; c) Atomizing air; d) Wire tips melting in the electric arc; e) Sprayed coating; f) Workpiece

More recent developments of the electric arc spraying process are aimed at spraying in a vacuum or in an inert gas atmosphere, which would enable application of otherwise reactive metals [71].

5.1.3. Plasma Spraying

Plasma spraying is one of the most commonly used spray coating processes for producing functional coatings. The method uses a heat source consisting of an electric arc inside a nozzle. This heats an inert gas stream (mixtures of argon, helium, nitrogen, or hydrogen). The ionization and recombination reactions of the gas produce very high temperatures — as high as 20 000 K in the center of the plasma flame. The coating material, usually in powder form, is introduced into this high-energy plasma jet by means of a carrier gas (usually argon). The powder particles are melted and impelled onto the prepared substrate material (Fig. 29).

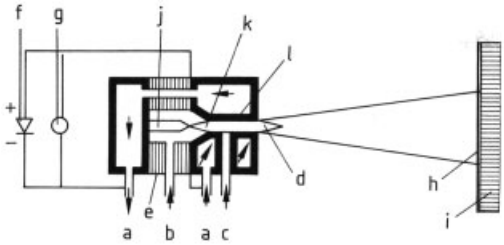

Figure 29. Principle of plasma spraying
a) Water cooling; b) Plasma gas; c) Powder coating material and carrier gas; d) Plasma jet; e) Insulation; f) Rectifier; g) High-frequency generator; h) Sprayed coating; i) Workpiece; j) Electrode (cathode); k) Electric arc; l) Nozzle (anode)

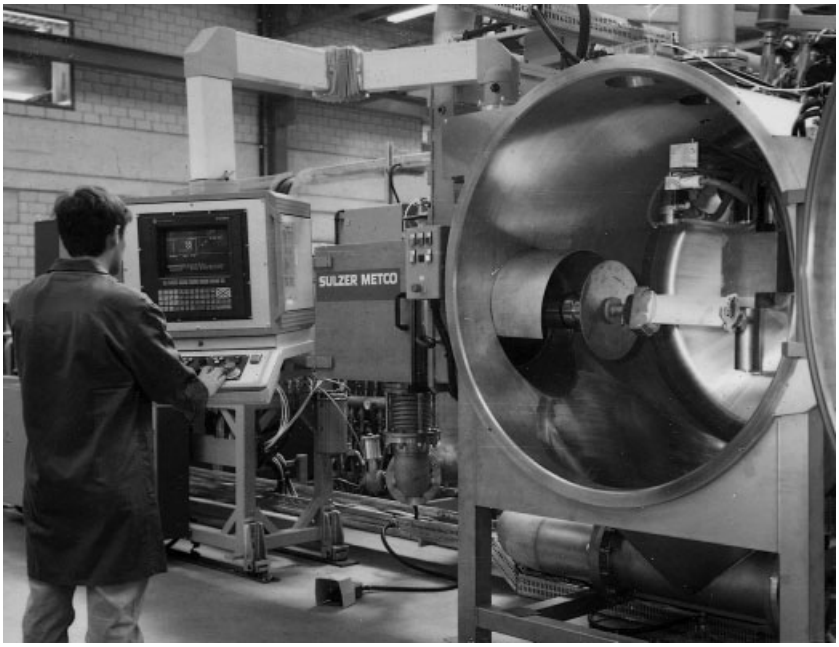

Figure 30. Modern vacuum plasma spraying system

Several process developments are described below. They are all basically carried out as described above, and differ mainly with respect to process parameters (air, inert gas, vacuum, water), which in turn determine plasma-gun design, and were developed for particular applications or spraying materials [72, 73]. Application of the various thermal-spraying methods may depend on the coating material and the quality or performance of the sprayed coating. Ceramics (mostly oxides) are, for example, sprayed by atmospheric plasma spraying. Oxygen-sensitive materials (e.g., refractory metals or aluminum-containing materials) are sprayed under inert conditions (inert gas plasma spraying, vacuum plasma spraying). Due to the reduced pressure in the vacuum chamber the spray particles attain higher velocities, resulting in denser coatings and better adhesion.

Atmospheric Plasma Spraying (APS). This process is performed in open air. The sprayed particles can react with oxygen during their flight from plasma gun to workpiece.

Inert Plasma Spraying (IPS). Unlike the atmospheric plasma spraying process, the IPS process is performed in a chamber containing an inert atmosphere (usually argon) either at or slightly above atmospheric pressure. Furthermore, an additional inert gas jacket can be used. The inert gas prevents reaction between the plasma jet, the workpiece, or the sprayed particles and the surroundings, but it is not a substitute for vacuum plasma spraying.

Vacuum Plasma Spraying (VPS). Vacuum plasma spraying is performed to meet higher specifications with respect to density, purity, and low oxide content of the coating. Spraying is carried out in a chamber containing an inert atmosphere at reduced pressure (5 – 30 kPa) to prevent chemical reaction of the plasma jet, sprayed material, and substrate with the surrounding atmosphere. Coatings are formed with higher energy since the particles undergo less cooling or deceleration in the low-density gas. Figure 30 shows a modern vacuum plasma installation, and Figure 31 a tantalum coating produced by the vacuum plasma spraying process [74, 75].

Underwater Plasma Spraying (UPS). Underwater plasma spraying features large reductions in fume, dust, and radiation emissions, with the option of using the process in offshore drilling operations in which coating can be done below the sea surface [76]. Figure 32 shows an underwater plasma spraying installation.

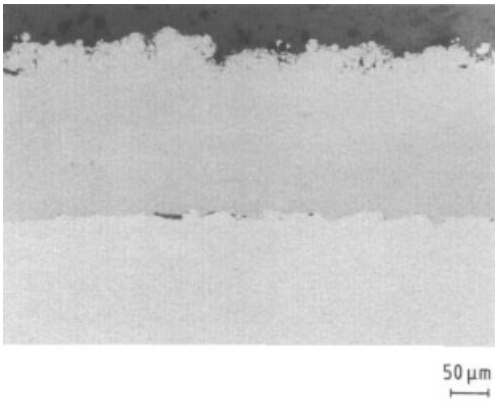

Figure 31. Vacuum plasma sprayed tantalum coating for prevention of corrosion

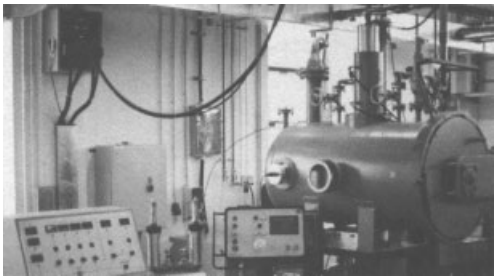

Figure 32. Underwater plasma spraying equipment for water pressures up to 1.5 MPa

High-Power Plasma Spraying (HPPS). High-power plasma spraying is still under development; high plasma energies are obtained using a special gun configuration. Increased deposition rates of refractive coating materials are obtained.

5.1.4. High-Velocity Oxy–Fuel Spraying (HVOF)

High-velocity oxy–fuel spraying (HVOF) involves a fuel gas – oxygen combustion process. Owing to strong gas expansion the resulting flame leaves the nozzle of the spray gun at supersonic speed (up to 2100 m/s). The powder with carrier gas is usually injected axially into the gun, melted, and accelerated to velocities which are much higher than those obtained with arc, flame, or atmospheric plasma spraying (Fig. 33) [77].

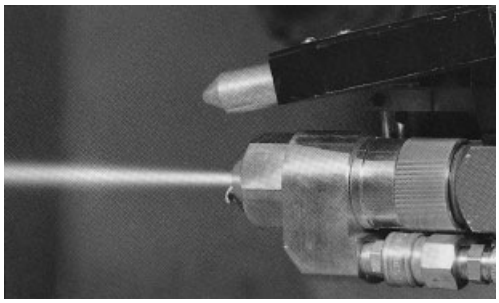

Figure 33. Spray stream of a Sulzer Metco Diamond Jet Hybrid HVOF spray gun

Compared the latter processes, HVOF coatings are characterized by high adhesion to

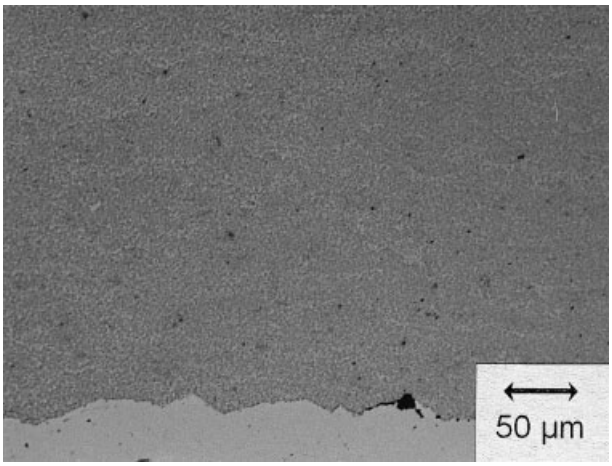

Figure 34. HVOF sprayed tungsten carbide/cobalt coating for wear protection

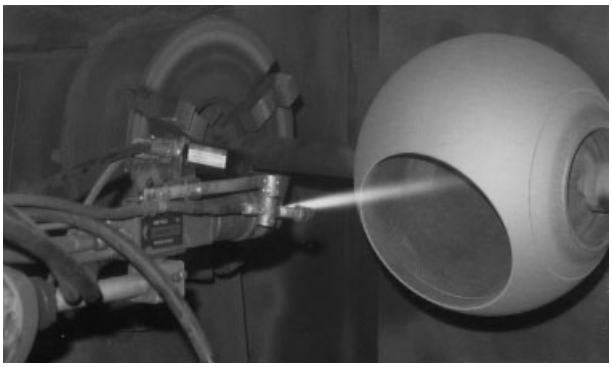

Figure 35. HVOF spraying of ball valves (Source: Bandak AS, Norway)

the substrate, good cohesion between the particles, low porosity, low surface roughness, low loss of alloying elements, and low oxide content. The HVOF process is most suitable for the production of carbide-containing wear-resistant coatings (Fig. 34) [80], corrosion resistant coatings, and high-quality protective coatings against wear and corrosion. The HVOF process has many interesting potential applications (Fig. 35).

In contrast to detonation gun spraying (Praxair), high-velocity oxy – fuel flame spraying is a continuous combustion process of oxygen and a gaseous or a liquid fuel. Typically propane, propylene, natural gas, hydrogen, acetylene, or kerosene is used as fuel; flame temperatures are between 2700 and 3150 °C. Flame velocities up to 2100 m/s can be obtained, implying particle velocities exceeding 650 m/s, depending on the fuel used, system hardware, and the material (grain size, density).

Characteristic of the HVOF process are visible shock diamonds in the high-velocity gas jet. These shock diamonds are speed nodes that result from compression and expansion waves. The spacing, size, and shape of the shock diamonds are related to flame velocity. The HVOF process combines very high particle velocity with controlled particle heating. Thus, thermally activated chemical reactions (e.g., oxidation, phase transformation, evaporation of elements, etc.) can be better controlled and limited.

The process features of HVOF are leading to wide industrial use, and steady growth in this spray technology for all kinds of coating applications is expected in the future.

5.2. Properties of Thermally Sprayed Coatings

Properties of thermally sprayed coatings depend strongly on spraying parameters, which must be optimized for each type of coating material, workpiece, and for each method of operation.

Coating Structure. The molten, high-velocity sprayed particles acquire a spherical shape as they move toward the substrate. High-energy impact on the substrate produces a lamellar, finely crystalline structure, owing to the energy of the spray and thus to process parameters such as the viscosity of the sprayed material. The lamellar structure is typical of sprayed coatings, and conversion to the compact, finely crystalline structure can only be achieved by thermal post-treatment. Depending on the sprayed material, annealing is sometimes feasible but not necessary in many cases.

Porosity. Depending on coating material, subsequent treatment, and geometry of the coated workpiece, thermally sprayed coatings have a degree of porosity, which can be 5 – 10 % for low-energy methods (flame spraying) down to < 1 % in the case of high-velocity oxy – fuel flame spraying or high-energy vacuum plasma spraying. Since the pores are mostly closed, thermally sprayed coatings have low permeability above a certain thickness, and protect the substrate from the surroundings. Thermally sprayed coatings are usually 200 – 500 μm thick and inhibit penetration of liquid media. Their thickness

can, however, range from ca. 15 μm up to several millimeters (depending on the material and the spraying process).

Adhesion. The adhesive strength of thermally sprayed coatings depends on spraying parameters and on the substrate. Adhesion generally decreases with increasing coating thickness because of the resulting increase in internal stresses. Special pretreatment of the surface is required for maximum adhesion of a spray coating to the substrate. This consists of mechanical roughening with corundum (Al_2O_3), carborundum (SiC), or boron carbide (B_4C), which also cleans and activates the surface. Thermal spraying must immediately follow blasting to avoid reduction of adhesive strength owing to contamination or oxidation. In the vacuum plasma spray process, additional pretreatment can be performed under vacuum. An electric arc is installed between the plasma gun and the substrate in addition to the plasma-generating arc in the plasma gun itself. This additional arc cleans and heats the workpiece.

Coatings consisting of thermally sprayed oxides (e.g., zirconia, alumina) or mixtures of oxides with metals are not applied directly to the substrate but to a thermally sprayed base coating (e.g., nickel – aluminum, nickel – chromium, or another metal alloy).

Surface Roughness. Thermally sprayed coatings have varying degrees of roughness. In the as sprayed condition, roughness R_z is 5 – 30 μm, and after polishing it can be < 0.2 μm (R_z is the mean difference between the five highest peaks and the five lowest valleys in the sampling length).

5.3. Materials

5.3.1. Spray Coating Materials

Coating materials used in thermal spraying can be classified according to their form and composition. They can be used in the form of solid wires, rods, filled wires, powders, and other shapes. Spray coating compositions cover a large number of natural and synthetic products both inorganic and organic. Materials should be capable of being liquefied without decomposition, sublimation, oxidation, or other adverse effects; otherwise these reactions must be controlled during spraying.

Materials for thermal spraying include basically all materials which exhibit a liquid phase on melting at ambient pressure. Materials without a liquid phase, such as SiC, can be processed in combination with matrix materials. Materials are usually available as powders or wires.

Powders. Spray materials are classified (e.g., DIN 32 529) according to their chemical composition:

- Pure materials (e.g., Cr, W, Ni, Cu, Al, Mo)
- Alloys (e.g., NiCr, NiAl, steels, CuAl, CoCrW, MCrAlY)
- Metal carbides (e.g., WC, TiC, Cr_2C_3 – NiCr, WC – Co, WC – CrCo)
- Self-fluxing alloys (e.g., NiCrBSi)
- Oxides (e.g., B. Al_2O_3, TiO_2, ZrO_2, Cr_2O_3, Al_2O_3 – TiO_2, ZrO_2 – Y_2O_3)
- Organic materials

In addition, blends and composite powders of the above groups are possible.

Powder spray coating materials can be classified according to particle size or composition. Particle diameters are between ca. 5 and 180 μm, and the particle size distribution is usually narrow. Typical powder grain size distributions are, for example, 22 – 45 μm or 22 – 90 μm, depending on the material (e.g., melting behavior) and the spraying process. Furthermore, the flow behavior in powder feeding is of importance for maintaining a stable coating process. The range of powders is almost unlimited, and their uses as thermal spray materials for spray processes must therefore be optimized [79].

Every spraying process requires characterization and specification of the appropriate powder. The gas temperatures are sometimes extremely high (e.g., 20 000 – 30 000 K in plasma spraying), as are the spraying velocities in some cases. There are many possible process gases and surrounding atmospheres, so that different processes require different powder materials with regard to chemical composition, grain size, shape, and particle structure (dense, porous, etc.) in order to obtain the best coating properties.

Production. Important basic steps in powder production include sintering and melting with subsequent mechanical size reduction, sometimes with an additional mixing stage, followed by classification.

Powders are manufactured by atomization, spray drying, reaction spray drying, fluidized-bed sintering, coating of the particles, agglomeration, plasma spheroidizing, and other processes. Virtually all combinations of powdered materials can be produced; most of them consist of a single component, but homogeneous and heterogeneous multicomponent powders (composite powders) are also available.

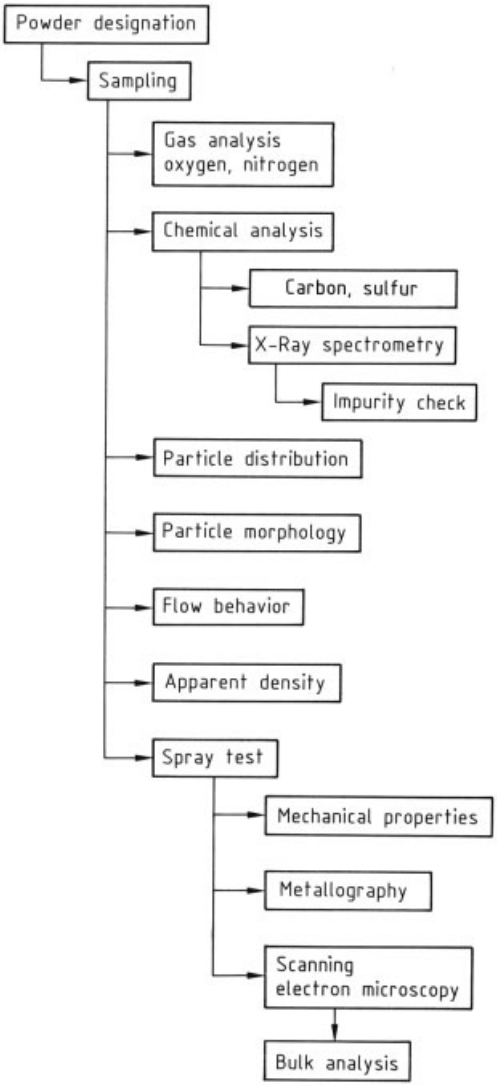

Figure 36. Testing procedures and quality control for powders and coatings

Characterization. The powders are characterized by means of their chemical composition and purity, physical properties, and morphology. Important morphological parameters of individual particles or collectives are shape, surface properties, porosity, texture, structure, homogeneity and phases. Their knowledge enables coating properties and manufacturing methods to be compared [78].

A practical procedure for quality testing and quality assurance of powders and coatings is shown in Figure 36 [79].

Wires. Wires can be classified, for example, according to DIN 8566 (Parts 1 – 3). Typical spray materials in wire form are solid wires of unalloyed and low-alloy steels, high-alloy steels, nonferrous metals such as Al and aluminum alloys, copper and copper alloys, tin alloys, nickel and nickel alloys, zinc, and molybdenum, as well as cored wires.

Solid wires are used mainly in flame and electric arc spraying. All steels and a wide variety of nonferrous metals and alloys can be used, including some refractory metals that can be manufactured in wire form. Several tens of thousands of tonnes of wire are used worldwide in these processes.

Cored wires can be produced in the form of overlapped cored wire, but can also be hollow tubes (generally of steel or nickel-base alloys) that are filled with a single- or multicomponent powder (often iron – chromium – carbon, tungsten carbide, or chromium carbide). Cored wire electrodes are already used for welding and are becoming of increasing importance for spraying, especially electric arc spraying.

5.3.2. Substrate Materials

Basically all engineering materials are suitable as substrate materials, provided appropriate pretreatment has been carried out, i.e., steels, cast iron, titanium and titanium alloys, aluminum materials, nickel and nickel alloys, copper and cobalt base materials, but also to some extent ceramics, plastics, glass, and wood. Restrictions exist because of temperature limitations of the substrate material and bonding limitations between coating and subtrate. Coating of thermochemically treated substrates (nitrided, case-hardened, borided) leads to irreversible modifi-

cations of the thermochemical layer, depending on coating temperature.

Satisfactory adhesion bonding requires careful, technically correct surface preparation and application of the coating within the shortest possible time. Surface contaminants such as oil, grease, lubricants, and particles must be avoided or removed. Degreasing can carried out with chemical agents, by heating, ultrasonically, or by steam cleaning. Surface roughening is most commonly performed by blasting. The blast parameters and the blasting material must be adapted to the area to be coated. Other methods include electrolytic or mechanical processes. Surface roughening should be carried out shortly before spraying.

5.4. Applications

Thermal sprays have many applications for the enhancement of surface properties such as wear resistance (i.e., abrasion, adhesion, erosion, and scratching, Fig. 37), and corrosion resistance (atmospheric, moisture, and hot gases). Further properties of coatings can be thermal insulation (Fig. 38), electrical insulation, electrical conduction including superconducting coatings (Fig. 39), self-lubricating coatings, abradable coatings, diffusion barriers, surface-active coatings, biocompatible coatings (Fig. 40), ionically conducting coatings, and decorative coatings.

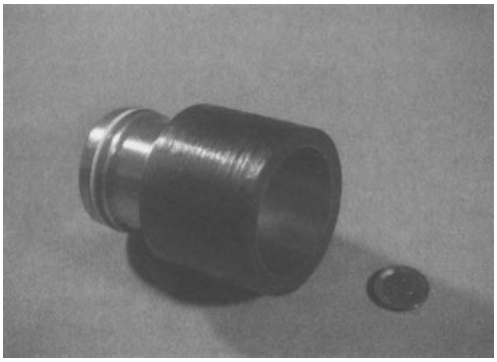

Figure 37. Carbon fiber reinforced resin with an inner coating of tungsten carbide and cobalt produced by thermal spraying

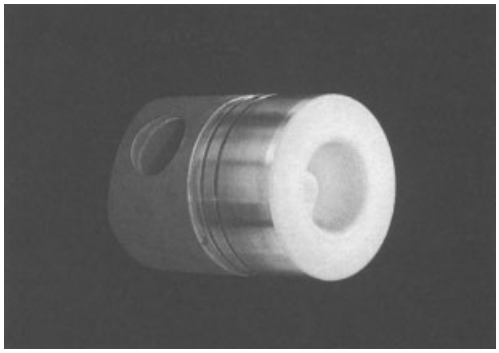

Figure 38. Sprayed yttria-stabilized zirconia coating used as a thermal barrier on a light-metal piston

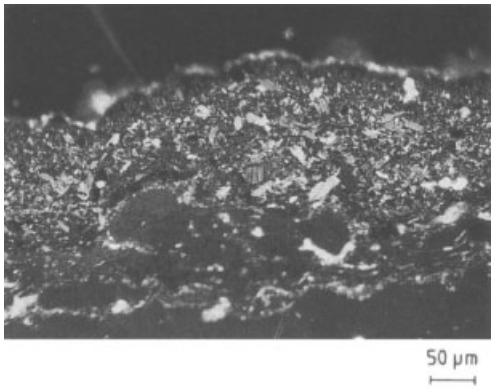

Figure 39. Section through high-temperature superconducting plasma-sprayed coating of the type Y–Ba–Cu–O Transition temperature 90.5 K; Critical current density 100 – 300 A/cm^2 (at 77 K).

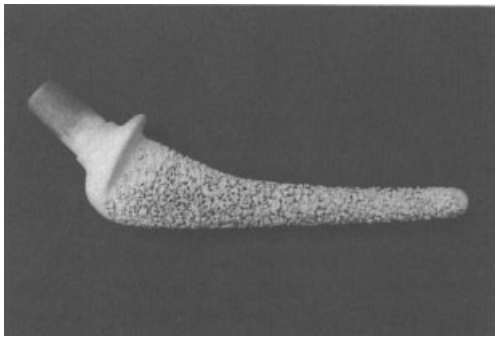

Figure 40. Plasma-sprayed hydroxyapatite on artificial hip joint implant for improvement of biocompatibility

Miscellaneous uses of thermal spraying include the manufacture of molds by spraying onto soluble cores, repair of worn parts (Fig. 41), recovery of products rejected because of faulty material or incorrect manufacture, and special powder treatments by means of melting and spheroidizing. Coating material and spraying process are selected according to the intended use of the coated article, its geometry and size, and the nature of the substrate and coating material.

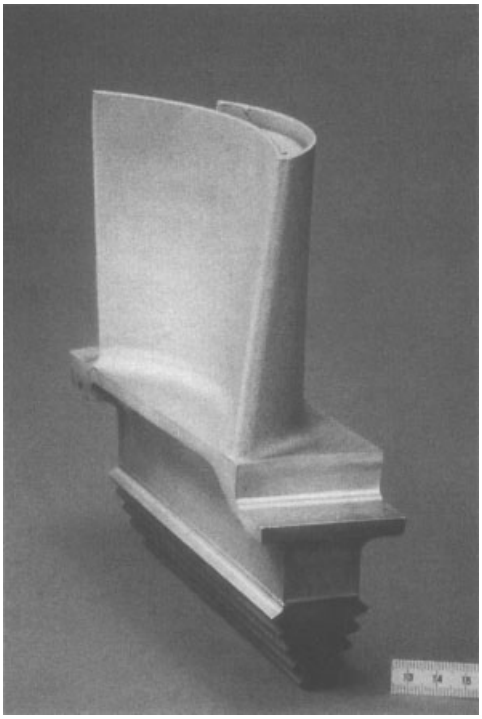

Figure 41. Turbine blade repaired by vacuum plasma spraying (material: Inconel 718)

6. Coloring

Decorative coloring of metals is no longer carried out on an industrial basis; anodized aluminum and aluminum alloys, as well as some special grades of stainless steel, are, however, exceptions. Only one important development has been made in the last few decades: the coloring of high-chromium stainless steel (mainly the DIN alloys 1.4301, X4CrNi18-10, and in special cases, 1.4512) to achieve a metallic luster even in outdoor applications. Metal coloring attained maximum popularity in the 1920s, but interest then declined, no further developments were made, and some of the old methods fell into oblivion. Another reason for the decline of this complicated technique was the fact that it could only be used by skilled craftsmen on individual items and was difficult to apply on an economic basis for mass production. Up to now chemical coloring has been used mainly in arts and crafts to achieve an antique decorative finishing on objects of copper (patina) and on brass and silver (brown to black coloration by immersion into polysulfide solutions). A special coloring method is the depletion of single alloy components from the surface by suitable chemical reactions, with simultaneous enrichment of the remaining components (also called surface enrichment); examples are the tumbaga gilding method (depletion gilding, mise en couleur) and depletion silvering [82]. For further details of coloring methods, see [83 – 97, 99] → Electrochemical and Chemical Deposition.

7. Acknowledgement

The entire topic was originally coordinated by Helmut Kunst, and the actual edition was completely revised by Brigitte Haase.

8. References

1. see: Richtlinie (Entwurf): Prüfung der Technischen Sauberkeit – Partikelverunreinigung funktionsrelevanter Automobilteile (VDA 19) Verband der Automobilindustrie, 1. Auflage 2004 – Gelbdruck).
2. H. G. Germscheid: *Galvanotechnik* **67** (1976) 215.
3. B. Haase, J. Dong: *Surface Engineering* **14** (1998) no. 1, 43 – 48.
4. J. Kresse: *Jahrb. Oberflächentech.* **42** (1986) 24 – 40.
5. The BImSchV 2 (1990) (Germany) permits the opening of the cleaning chamber and the removal of the cleaning parts only if the CHC concentration is below 1 g/cm^3.
6. GefStoffV, 2002 (Germany).

7. VbF (Germany): A3 solvents with flash points > 55 °C.
8. Kolene Corp., US 2 468 006, 1948 (H. G. Webster, J. H. Shoemaker).
9. Kolene Corp., US 2 863 465, 1955 (R. H. Shoemaker, J. A. Faler).
10. Kolene Corp., US 4 113 511, 1977 (R. H. Shoemaker, W. G. Wood).
11. Kolene Corp., US 3 260 619, 1965 (R. H. Shoemaker, H. Geschwender).
12. Kolene Corp., US 2 458 661, 1944 (H. G. Webster, C. L. Falter).
13. J. C. Malloy: *Sixth World Conference of Investment Casting*, Washington, D.C., October 1984.
14. J. C. Malloy: *Clemson Univ. Conference on In-Plant Extrusion*, June 1989.
15. I. I. Chain: *Theorie und Praxis der Phosphatierung von Metallen*, Verlag Chimija, Leningrad 1973.
16. G. Lorin: La phosphatation des métaux, Editions Eyrolles, Paris 1973. English translation: *Phosphating of Metals*, Finishing Publications Ltd., Hampton Hill, Middlesex, 1974.
17. Ph. Weiss, G. D. Cheever (eds.): *Interface Conversion for Polymer Coatings*, American Elsevier Publishing Company, New York 1968.
18. W. Rausch et al.: *Die Phosphatierung von Metallen*, 2nd ed., Eugen G. Leuze Verlag, Saulgau 1988.
19. D. B. Freeman: *Phosphating and Metal Pretreatment*, Woodhead-Faulkner, Cambridge.
20. DIN 50 942: Phosphatierung von Metallen, 1987.
21. M. Gebhardt, *Fachber. Oberflächentech.* **9** (1971) no. 3, 81 – 88.
22. K. Wittel, *Ind. Lackierbetr.* **51** (1983), 169 – 175, 210 – 215.
23. W.-A. Roland, K.-H. Gottwald, *Metalloberfläche* **42** (1988) 301 – 305.
24. R. Vey, *Wire* **45** (1995) 346 – 349.
25. W.-A. Roland, K.-H. Gottwald, *Metalloberfläche* **48** (1994) 790 – 796.
26. H. Schümichen, *JOT* **33** (1993) 42.
27. H. Gehmecker, *Handelsblatt*, April 13, 1994, p. B5.
28. W. Wimmer, J. Gottschalk, E. Busch, H. Gehmecker, *JOT* **37** (1997) no. 3, 39.
29. W. Wimmer, H. Gehmecker, Conference Proceedings of SURCAR 1997, in press.
30. Oakite Products Inc., US 5 053 081, 1990 (S. M. Jacob).
31. Th. Kolberg, *Metalloberfläche* **44** (1990) 6.
32. OCC, US 4 376 000, 1980 (A. Lindert).
33. Metallgesellschaft, EP-A 698 675, 1995 (M. Fleischhacker-Jeworrek, D. Jentsch, K. Wittel).
34. J. Grosch: "Werkstoffseitige Grundlagen des induktiven Randschichthärtens," *HTM Härterei-Tech. Mitt.* **44** (1989) no. 4, 205 – 209.
35. H. Benninghoff (ed.): *Wärmebehandlung der Bau- und Werkzeugstähle*, 3rd ed., BAZ Buchverlag, Basel 1978.
36. H.-J. Eckstein (ed.): *Technologie der Wärmebehandlung von Stahl*, 2nd ed., VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1987.
37. T. Rudlaff: "Arbeiten zur Optimierung des Umwandlungshärtens mit Laserstrahlen," Dissertation Universität Stuttgart, Teubner Verlag, Stuttgart 1993.
38. G. Stähli: "Kurzzeit-Wärmebehandlung," *HTM Härterei-Tech. Mitt.* **39** (1984) no. 3, 81 – 90.
39. Haase, B., M. Stiles, J. Dong: *Surf. Eng.* **16** (2000) no. 3, 263 – 268.
40. J. Wünning: "Prozeßrechner zur Steuerung des Diffusionsverlaufs während der Aufkohlung," *ZWF Z. Wirtsch. Fertigung* **77** (1982) no. 9, 424 – 426.
41. K. H. Weissohn, "Die Technik der C-Pegel-Regelung," *HTM Härterei-Techn. Mitt.* **51** (1996) no. 6, 393 – 270.
42. U. Wyss, "Verbrauch an Trägergas bei der Gasaufkohlung," *HTM Härterei-Tech. Mitt.* **38** (1983) no. 1, 4 – 9.
43. J. St. Pierre: "Recent Developments in Vacuum Carburizing" in: *Carburizing, Processing and Performance, Proceedings of an International Conference,* ASM, Lakewood, Colorado, July 1989. Publ. by ASM.
44. F. Schnatbaum, G. Schmitt, F. Preißer, "Puls-Plasmaaufkohlung mit Hochdruck-Gasabschreckung im industriellen Maßstab," *HTM Härterei-Tech. Mitt.* **51** (1996) no. 5, 265 – 270.
45. G. Wahl: "Development and Application of Salt Baths in the Heat Treatment of Case Hardening Steels" in [43]
46. G. Legge: "Plasma Carburizing — Facility Design and Operating Data," *Ind. Heat.* (1988) March, 26 – 30.
47. DIN, Deutsches Institut für Normung e.V. (ed.): *Wärmebehandlung metallischer Werkstoffe: Normen,* vol. 218, Beuth-Verlag, Berlin – Köln 1986.

48. H. Streng: *Zähigkeitsoptimierung einsatzgehärteter Gefüge*, Dissertation, TU Berlin 1986.
49. H. Czichos, K.-H. Habig: *Tribologie Handbuch*, Vieweg-Verlag, Braunschweig 1992.
50. G. Wahl: "Randschichtverfahren," in H. Kunst et al. (eds.): *Verschleißhemmende Schicheten*, expert-Verlag, Ehingen 1993.
51. D. Liedtke: "Grundlegende Betrachtung des Nitrierens und Nitrocarburierens von Eisenwerkstoffen," in R. Chatterjee-Fischer (ed.): *Wärmebehandlung von Eisenwerkstoffen: Nitrieren und Nitrocarburieren*, expert-Verlag, Sindelfingen 1986, pp. 1 – 60.
52. R. Hoffmann, H. Mallener: "Nitrieren und Nitrocarburieren in Gasen," in [51], pp. 63 – 124.
53. F.-W. Eysell: "Nitrocarburieren im Salzbad," in R. Chatterjee-Fischer (ed.): *Wärmebehandlung von Eisenwerkstoffen: Nitrieren und Nitrocarburieren*, expert-Verlag, Ehingen 1995.
54. F.-W. Eysell: "Nitrocarburieren im Salzbad," in [47], pp. 185 – 232.
55. H. W. Schläpfer: "Drei Salzbadnitrocarburierverfahren im Vergleich," *HTM Härterei-Tech. Mitt.* **39** (1984) no. 6, 267 – 272.
56. E. J. Mittemeijer, P. F. Colijn: "Oberflächenoxidation von Nitrierschichten, über die Ursachen der Verbesserung der Korrosionsfestigkeit," *HTM Härterei-Tech. Mitt.* **40** (1985) no. 2, 77 – 79.
57. D. Liedtke: *Beitrag zum technisch-wirtschaftlichen Optimieren des Nitrocarburierens von Bauteilen*, Dissertation, TU Berlin 1986.
58. W. Rembges: "Nitrieren und Nitrocarburieren im Plasma," in [51], pp. 125 – 183.
59. H. Wilhelmi, S. Strämke, H. C. Pohl: "Nitrieren mit gepulster Glimmentladung," *HTM Härterei-Tech. Mitt.* **37** (1982) no. 6, 263 – 269.
60. R. Grün, W. Exner, R. Zeller: "Puls-Plasma-Wärmebehandlung zum Oberflächen- und Randschichtbehandeln," *HTM Härterei-Tech. Mitt.* **42** (1987) no. 1, 42 – 49.
61. A. von Matuschka: *Borieren*, Hanser-Verlag, München – Wien 1977.
62. H. Kunst, O. Schaaber: "Borierverfahren," *HTM Härterei-Tech. Mitt.* **22** (1967) no. 4, 275 – 292.
63. F. Hegewaldt, L. Singheiser, M. Türk: "Gasborieren," *HTM Härterei-Tech. Mitt.* **39** (1984) no. 1, 7 – 15.
64. H.-J. Hunger, G. Trute, G. Löbig, D. Rathjen, "Plasmaaktiviertes Gasborieren mit Bortrifluorid," *HTM Härterei-Tech. Mitt.* **52** (1997) no. 1, 39 – 45.
65. H. Simon, M. Thoma: *Angewandte Oberflächentechnik*, Hanser-Verlag, München – Wien 1985.
66. Kabushiki Kaisha Toyota Kenkyusho, Nagakute, Aichi, JP 80 411–77, 1977 (Arai).
67. M. Villat: "Beschichtungen durch thermisches Spritzen," *SGT-Fachtagung*, Bern 1985.
68. E. Lugscheider, *Metalloberfläche* **41** (1987) 240.
69. O. Knotek, H. Reimann, E. Lugscheider: "Vakuumeingeschmolzene Oberflächenschichten aus Nickelhartlegierungen," *Metall (Berlin)* **34** (1980) no. 3, 228.
70. H. Kreye, P. Heinrich, G. Löwe: "Jet Kote Spraying with Different Fuel Gases," *Proc. ATTAC '88, Osaka* 1988, 73 – 78.
71. H. D. Steffens, M. Wewel: "Vacuum Arc Spraying — An Advanced Coating Technology," *Proc. 12th ITSC*, London 1989.
72. E. Lugscheider, Th. Weber: *Stand und Entwicklungstendenzen beim Plasmaspritzen*, Vulkan Verlag, Essen 1987. *Elektrowärme Int. Ed. B:* **45** (1987) no. 3/4, 190 – 195.
73. Plasmatechik AG, Work Report, *Plasma Spraying Technique — Basic Principles and Applications,* Wohlen, Switzerland 1975.
74. E. Lugscheider, H. Eschnauer, B. Häuser, D. Jäger: "Vacuum Plasma Spraying of Tantalum and Niobium," *Int. Conf. Metall. Coatings*, Los Angeles, USA 1985. *J. Vac. Sci. Technol. A* **3** (1985) no. 6, 1469.
75. H. Gruner: "Vacuum Plasma Spray Control," *Thin Solid Films* **118** (1984) 409 – 420.
76. E. Lugscheider, B. Häuser, B. Bugsel: *Int. Conf. Metall. Coatings*, San Diego, USA 1986.
77. M. Nestler, U. Erning, R. Henriksen: "Characteristics and Industrial Applications of the Diamond Jet Hybrid - The New Generation of HVOF Systems," *Conf. Proc. 4th Conf. HVOF Spraying, Erding, Germany* 1997.
78. E. Lugscheider, A. Krautwald, J. Wilden: "Laser Treatment of Thermal Spray Coatings," *Conf. Heat Treatment*, London 1987.
79. M. Villat: Lecture given at 1st Plasma-Technik Symposium, Lucerne 1988.

80. H. Eschnauer, E. Lugscheider, D. Jäger: *Proceedings 12th Int. Plansee Seminar*, Reutte 1989, vol. 1, pp. 773 – 791.
81. S. La Niece, P. Craddock (ed.): *Metal Plating and Patination*, Butterworth-Heinemann, Oxford 1993.
82. G. Buchner: *Die Metallfärbung, Handbuch für die chemische–elektrochemische und mechanische Metallfärbung*, 6th ed., Verlag M. Krayn, Berlin 1920.
83. H. Krause: *Metallfärbung*, 3rd ed., Carl Hanser Verlag, München 1951.
84. *Farbeffekte in der Oberflächentechnik — 11. Ulmer Gespräch* (Tagungsband, April 20 – 21, 1989, Ulm), DGO-VDI-TZ, Eugen-G. Leuze Verlag, Saulgau 1989.
85. R. Hughes, M. Rowe: *The Colouring, Bronzing and Patination of Metals*, 4th ed., Crafts Council London 1989.
86. Proceedings of the Conference on Coloring of Metals (British Museum), London 1990/91.
87. *Ullmann*, **16**, 548 – 551.
88. G. Oelsner in Dettner, Elze: *Handbuch der Galvanotechnik*, **vol. III**, Carl Hanser Verlag, Munich 1969, pp. 276 – 314.
89. A. M. Carpenter: *Techniques for the Coloring of Metals*, Special Report no. 36, The Worshipful Company of Goldsmiths, Technical Advisory Commitee, London, Nov. 1978.
90. W. Schwartz, *Plat. Surf. Fin.* **96** (1982) June, 26 – 29.
91. See "Färben von Eisen und Stählen," in [88], pp. 284 – 291, 313 – 214.
92. T. E. Evans, A. C. Hart, A. N. Skedgell, "The Nature of Film on Coloured Stainless Steel," *Trans. Inst. Met. Finish.* **51** (1973) 108 – 112.
93. V. Piddock, B. Mahmoudian, D. J. Arrowsmith, "Some Factors Affecting the Colouration of Stainless Steel by Immersion in a Mixture of Hot Chromic and Sulphuric Acid," *Trans. Inst. Met. Finish.* **61** (1983) 93 – 96.
94. "Blaugefärbte Stahlteile," *Galvanotechn.* **86** (1995) no. 12, 3980 – 3982.
95. "Cold Metal Colouring — a Viable Alternative," *Trans. Inst. Met. Finish.* **65** (1987) no. 3, 19 – 21.
96. T. Mingwei, "The Mechanism of Chemical Colouring of Stainless Steel — I. Calculation of the Colouring Initiation Time," *Corros. Sci.* **33** (1992) no. 6, 873 – 878.
"The Mechanism of Chemical Colouring of Stainless Steel — II. Calculation of the Diffusion Potential," *Corros. Sci.* **33** (1992) no. 6, 879 – 884.
97. Färben von Stahl, Mitteilung DFBO, 1950, no. 33, pp. 4 – 6.
98. A. Strauch et al. in: *Galvanotechnisches Fachwissen*, VEB Deutscher Verlag für Grundstoffchemie, Leipzig 1987, 361 – 363.
99. S. Wernick, R. Pinner, P. G. Sheasby: "Colouring Anodic Oxide Coatings with Dyes and Pigments." *The Surface Treatment of Aluminum and its Alloys*, 5th ed., **vol. II**, Chap. 10, ASM International, Metals Park, Ohio and Finishing Publications, Teddington 1987.
100. L. Vanino, E. Seitter: *Die Patina*, A. Hartleben's Verlag, Pest, Leipzig 1903.

Methacrylic Acid and Derivatives

WILLIAM BAUER, JR., Rohm and Haas Co., Philadelphia, Pennsylvania 19477, United States

1. **Introduction** 1
2. **Properties** 2
2.1. **Physical Properties** 2
2.2. **Chemical Properties** 2
3. **Production** 4
3.1. **Methacrylic Acid and Methyl Methacrylate from Acetone Cyanohydrin** . . 4
3.2. **Methacrylic Acid from Isobutene** . . . 5
3.3. **Methacrylic Acid from Isobutyric Acid** 7
3.4. **Methacrylic Acid From Ethylene** . . . 8
4. **Quality Specifications** 9
5. **Storage and Transportation** 9
6. **Uses** 10
7. **Economic Aspects** 10
8. **Toxicology and Occupational Health** . 10
9. **References** 12

1. Introduction

Methacrylic acid [*79-41-4*], α-methylacrylic acid, 2-methylpropenoic acid,CH_2=C(CH_3) COOH, M_r 86.09, is a colorless, moderately volatile, corrosive liquid with a strongly acrid odor. It was first prepared in 1865 from ethyl methacrylate, in turn obtained by dehydration of ethyl α-hydroxyisobutyrate [1].

Historical Aspects. OTTO RÖHM's doctoral thesis (1901) on polymerization products of acrylic acid derivatives represented the start of interest in acrylates and methacrylates as materials of potential commercial value [2]. RÖHM reported the preparation of colorless, clear, rubbery materials that he thought could be of practical utility. He and OTTO HAAS formed a company to manufacture leather treatment enzymes in 1907. Although laboratory studies of the preparation of acrylates and methacrylates continued, it was not until the 1920s that significant effort was spent on the development of commercial processes for the monomers and derived polymers.

After World War I, RÖHM developed a route for commercial production of acrylates from ethylene cyanohydrin, in turn derived from ethylene chlorohydrin, an intermediate used in the production of gases for chemical warfare. Later, he and WALTER BAUER developed an analogous synthesis of methyl methacrylate using acetone cyanohydrin. By 1931 BAUER succeeded in producing a methyl methacrylate polymer as a clear, transparent solid that softened above 100 °C. This first "organic glass" could be worked with saws, drills, and other common tools, which suggested significant practical applications. Research on forming cast sheet from methyl methacrylate was carried out in the 1930s by both Rohm and Haas and Du Pont in the United States, by Röhm and Haas AG in Germany, and by ICI in England.

Commercial production of methyl methacrylate monomer began in Darmstadt, Germany in 1933 (two years after the commercial production of acrylates) by the original acetone cyanohydrin route. The cyanohydrin was converted to ethyl α-hydroxyisobutyrate by treatment with ethanol and dilute sulfuric acid. This was then dehydrated using P_2O_5 [3–7]. In 1934 a patent was issued to ICI for the conversion of acetone cyanohydrin to methacrylamide sulfate using concentrated sulfuric acid [8]. The methacrylamide sulfate could be hydrolyzed to methacrylic acid and esterified to form methyl methacrylate [*80-62-6*]. RÖHM then cross-licensed his technology for manufacture of cast acrylic sheet with ICI so that the more efficient ICI process could be used for producing methyl methacrylate monomer. Both ICI and Röhm and Haas AG began commercial production of methyl methacrylate using the methacrylamide sulfate route in 1937. Rohm and Haas in the United States (a branch of the company that had become independent in 1917) was unable to obtain a similar license until the mid-1940s, so they continued to use the original process for several years [2].

The methacrylamide sulfate route provided the basis for commercial production of methac-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_441

rylic acid and methyl methacrylate until the 1983 introduction of isobutene oxidation processes, developed independently in Japan by Japan Catalytic Chemical and Mitsubishi Rayon. Although additional isobutene-based methacrylic acid capacity is anticipated in Japan, and a new ethylene-based facility is under construction by BASF in Ludwigshafen, most of the world demand for methyl methacrylate (1.4×10^6 t in 1988, 2.2×10^6 t in 1996 [75]) continues to be met via acetone cyanohydrin.

Table 1. Physical properties of methacrylic acid, methyl methacrylate, and butyl methacrylate

Property	Methacrylic acid	Methyl methacrylate	Butyl methacrylate
M_r	86.09	100.12	142.20
mp, °C	15.8	−48	−50
bp (101.3 kPa), °C	162	101	163
Density (d_4^{25}), g/cm^3	1.015	0.939	0.889
Refractive index, n_D^{25}	1.4288	1.4120	1.4220
Vapor pressure, kPa			
at 20 °C	0.09	3.87	0.20
at 60 °C	1.33	25.2	2.0
Viscosity (24 °C), mPa · s	1.38	0.53	0.92
Solubility (20 °C), g/kg			
in water	complete	15.9	1.0
water in	complete	11.5	ca. 1
T_{crit}, °C	370	291	
P_{crit}, Pa	4.70	3.68	
V_{crit}, cm^3/mol	270	323	
Heat of vaporization (101.3 kPa), kJ/mol	0.418	0.36	
Specific heat capacity, J g^{-1} K^{-1}	2.1	1.9	1.9
Flash point, °C			
Cleveland open cup	77		66
Tag closed cup	67		
Tag open cup		13	
Autoignition temperature, °C	400	435	294

Most acetone cyanohydrin-based operations lead directly to methyl methacrylate monomer, which is the most important methacrylate derivative. Methacrylic acid produced by other routes also serves as a key intermediate to methyl methacrylate.

Much smaller quantities of ethyl methacrylate [*97-62-3*], *n*-butyl methacrylate [*97-88-1*], and isobutyl methacrylate [*97-86-9*], as well as higher methacrylates (e.g., lauryl methacrylate [*142-90-5*] and stearyl methacrylate [*32360-05-7*]) are manufactured by direct esterification of methacrylic acid or transesterification of methyl methacrylate. In addition to these, 2-hydroxyethyl methacrylate [*868-77-9*], 2-hydroxypropyl methacrylate [*923-26-2*], and aminoalkyl methacrylates such as 2-dimethylaminoethyl methacrylate [*2867-47-2*] are also commercially available. Methacrylamide [*79-39-0*] is offered as a specialty monomer; it can be obtained from the intermediate methacrylamide sulfate.

2. Properties

2.1. Physical Properties

The most important physical properties of methacrylic acid and its derivatives are listed in Tables 1 and 2 [3], [9–13]. Liquid – liquid equilibrium data for the methanol – water – methacrylic acid – methyl methacrylate system are available [14–16], as are key vapor – liquid equilibrium data for methacrylic acid – water [17] and related methacrylic ester systems [18], [19]. Solid – liquid equilibria for methacrylic acid – water have also been published [20], as have the physical properties of aminoalkyl methacrylates [21].

2.2. Chemical Properties

Methacrylic acid and its esters are very reactive, displaying reactions typical of both the vinyl function and the carboxylic acid or ester group. The electron-withdrawing effect of the carboxylic acid or ester group polarizes the double bond and enhances its reactivity, but this reactivity is less pronounced than in the acrylates because of the electron-donating methyl group. Thus, nucleophilic Michael and Michael-like additions to the double bond take place only with electron-rich reagents.

Addition to the Carbon – Carbon Double Bond. Addition of hydrogen cyanide, hydrogen halides, hydrogen sulfide, mercaptans, alkyl amines, alcohols, phenols, or phosphines leads to β-substituted α-methyl propionates [3], [5], [22], [23].

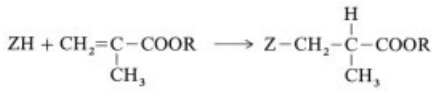

Table 2. Physical properties of methacrylic acid derivatives

Compound	*mp*, °C	*bp* (101.3 kPa), °C	Refractive index (n_D^{25})	Density (d_4^{20}), g/cm^3
Methacrolein	−81	68	1.4144 [b]	0.837
Methacrylonitrile	−36	90	1.3989	0.800
Methacrylamide	110			
Methacrylic anhydride		75 [a]	1.4520	
Methacryloyl chloride		96–98 [a]	1.4435	1.087 [c]

[a] At 0.67 kPa.
[b] At 20 °C.
[c] At 25 °C.

Diels–Alder Reactions. Diels–Alder reactions occur with dienes such as butadiene and cyclopentadiene [3], [23], [24].

Reactions of the Carboxylic Acid Function. Esterification of methacrylic acid with alcohols in the presence of catalytic quantities of sulfuric or sulfonic acids provides the corresponding esters. Acid-catalyzed addition of olefins also gives esters [25].

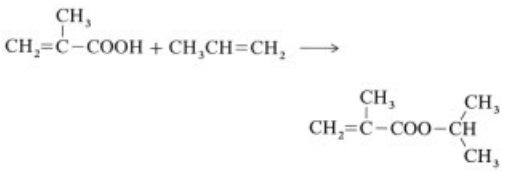

Transesterification of methyl methacrylate is another convenient technique for the manufacture of esters. Catalysts include dialkyl tin oxides [26], sodium phenoxide [27], and zinc chloride [28]. Methanol formed during the reaction is removed as a binary azeotrope with methyl methacrylate.

Reaction of the methacrylic acid group with epoxides permits the synthesis of hydroxyethyl and hydroxypropyl methacrylates. Appropriate catalysts include anion-exchange resins [29], ferric chloride [30], and lithium salts [31].

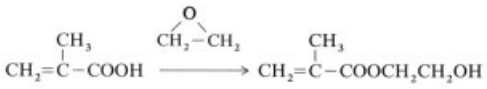

Methacryloyl chloride [*920-46-7*] is synthesized by reaction of the acid with thionyl chloride or phosphorus trichloride. Methacrylic anhydride [*760-93-0*] is prepared by reaction of the acid with acetic anhydride.

Methacrylamide is most conveniently prepared by reaction of acetone cyanohydrin with sulfuric acid [6], [32], [33]. Crude methacrylamide sulfate prepared as described in Section 3.1 is first neutralized with ammonia, after which water is added. The resulting methacrylamide can be separated from the aqueous ammonium sulfate by solvent extraction, filtration, or as a melt. It can be purified further by recrystallization. Alternatively, methacrylamide can be prepared by reaction of methyl methacrylate with ammonia.

Methacrylonitrile [*126-98-7*] is produced by ammoxidation of isobutene. It may be converted to methyl methacrylate by a process similar to the cyanohydrin route.

Oxidation. In the absence of free radical polymerization inhibitors, air oxidation of methyl methacrylate leads to a polymeric peroxide that decomposes to formaldehyde and methyl pyruvate.

Polymerization. Methacrylic acid, its esters, and other methacrylate derivatives polymerize readily upon heating in the presence of free radical initiators (e.g., peroxides, UV light, or ionizing radiation). Because the usual phenolic inhibitors function by reaction with peroxy radicals rather than with carbon radicals, these do not interfere with initiated polymerization. However, oxygen must be excluded during the pro-

cess because it converts alkyl radicals to hydroperoxy radicals.

Poly(methyl methacrylate) is unusual among polymers in that its pyrolysis gives high yields of the monomer. Heating scrap or off-grade polymer above 300 °C gives > 80 % recovery of monomer, which can be purified by fractional distillation. For further details of methacrylate polymers, see → Polyacrylamides and Poly(Acrylic Acids), → Polymethacrylates.

3. Production

This chapter is restricted to production methods for methacrylic acid and methyl methacrylate, since the preparation of methacrylic acid derivatives is summarized in Section 2.2. Disregarding the role of C_1 molecules (methanol, carbon monoxide, formaldehyde, and hydrogen cyanide), the commercially practicable routes to methacrylic acid and methyl methacrylate utilize ethylene (C-2 route), propene or acetone (C-3 routes), and isobutene (C-4 route) as carbon sources. Economic considerations for the various processes depend upon regional raw material supplies and costs; none of the newer processes is so superior to the acetone cyanohydrin route as to warrant scrapping existing cyanohydrin-based plants. Further, no single alternative process is clearly superior in an economic sense to the others.

3.1. Methacrylic Acid and Methyl Methacrylate from Acetone Cyanohydrin

The most common approach to methacrylic acid synthesis is the hydrolysis of methacrylamide sulfate, obtained from acetone cyanohydrin [6], [7], [23]. Methyl methacrylate may be prepared directly in a similar way by adding methanol in the final reaction step.

Dry acetone and hydrogen cyanide react in the presence of a basic catalyst to give the cyanohydrin, which is then reacted with excess concentrated sulfuric acid (1.4 – 1.8 mol per mole of cyanohydrin) to form methacrylamide acid sulfate:

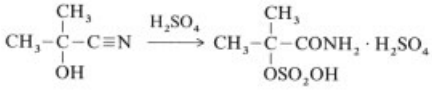

The sulfuric acid serves both as a specific reactant and as a solvent for the reaction, which appears to involve an α-sulfatoamide intermediate. If insufficient sulfuric acid is used, the reaction mass becomes a slurry or solid that is difficult or impossible to cool and pump. Both the sulfuric acid and the acetone cyanohydrin must be anhydrous in order to minimize hydrolysis of the sulfato derivative to α-hydroxyisobutyramide. The initial reaction is carried out continuously in a series of stirred tank reactors. Good heat transfer is required to assure removal of the heat of reaction. Thorough mixing is also necessary to avoid decomposition of the cyanohydrin to acetone and hydrogen cyanide, which can react to form byproduct acetone sulfonates and formamide sulfate. After initial reaction is completed at 80 –110 °C, the mixture is subjected to brief thermal cracking at ca. 125 – 160 °C to convert most of the α-hydroxyisobutyramide byproduct to methacrylamide sulfate, along with some acetone, carbon monoxide, and water. Total residence time at this stage of the process is about 1 h.

In a second stage, the methacrylamide sulfate stream is either hydrolyzed with excess water to give methacrylic acid and ammonium acid sulfate, or it is treated with aqueous methanol in a combined hydrolysis – esterification step to produce a mixture of methyl methacrylate and methacrylic acid:

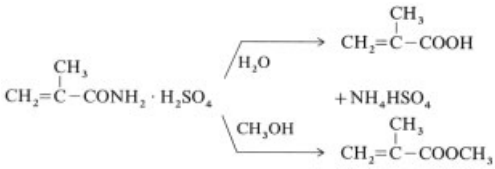

Several modifications are possible in the hydrolysis – esterification step. For example, the methacrylamide sulfate, water, methanol, and recycle streams can be led through a series of continuous reactors at 80 – 110 °C with a 2 – 4 h residence time. The reactor effluent then passes to a stripping column where crude methyl methacrylate, water, and excess methanol are removed overhead. The waste acid ammonium

sulfate residue can be either treated with ammonia for conversion to fertilizer or burned to regenerate sulfuric acid. Crude methyl methacrylate is extracted with water to recover methanol, which is concentrated and recycled to the esterification reactor. Washed ester is then purified by further distillation. The overall yield based on acetone cyanohydrin is in the range of 80 – 90 %.

In the manufacture of methacrylic acid, methacrylamide sulfate is reacted with water under conditions similar to those used for formation of the ester. The reactor effluent separates into two phases. The upper organic layer is distilled to provide pure methacrylic acid. The lower layer is steam stripped to recover dilute aqueous methacrylic acid, which is recycled to the hydrolysis reactor. The waste acid stream is treated as in the manufacture of the ester.

In an alternative process, shown in Figure 1, the methanolysis reaction is carried out at a pressure of ca. 800 kPa in one or more reactors (d) operated at 100 – 150 °C. The reactor effluent is separated into two layers (e) while still under pressure. The lower layer is steam stripped (g) to recover methacrylic acid for recycle to the esterification reactor (d), and acid waste is treated as described previously. The upper layer is passed to a distillation column (f) where light ends (primarily dimethyl ether) are removed. The bottoms from this column are washed with aqueous ammonia (h) to recover methanol and methacrylic acid for recycle to the esterification reactor. The crude, washed methyl methacrylate is then dehydrated in a downstream distillation column (i) and distilled in a product column (j) to provide pure methyl methacrylate.

A sulfuric acid regeneration plant is usually operated in conjunction with the methyl methacrylate plant, because approximately 1.6 kg of sulfuric acid is required to produce each kilogram of methyl methacrylate. The presence of a regeneration facility avoids the need to dispose of large quantities of ammonium sulfate contaminated with organic material.

One of the driving forces for development of the alternative routes discussed in Sections 3.2, 3.3, 3.4 has been the desire to eliminate the need for sulfuric acid regeneration. An additional concern is the hazard associated with transporting hydrogen cyanide, which is not always generated at the methacrylate plant site. On the other hand, both acetone (from phenol manufacture) and hydrogen cyanide (from acrylonitrile manufacture) have the economic advantage of being themselves industrial byproducts.

Asahi Chemical manufactures methyl methacrylate by the sulfuric acid hydrolysis of methacrylonitrile, using a plant originally designed for the acetone cyanohydrin process. The requisite methacrylonitrile is obtained by ammoxidation of isobutene, avoiding the need for hydrogen cyanide.

3.2. Methacrylic Acid from Isobutene

In recent years many companies have investigated the manufacture of methacrylic acid by two-stage catalytic oxidation of isobutene or *tert*-butanol. Nihon Methacryl Monomer (a joint venture of Sumitomo and Nippon Shokubai) and Mitsubishi Rayon have both constructed commercial plants using this technology.

In the first stage of the process [4], isobutene is oxidized to methacrolein (→Acrolein and Methacrolein, Chap. 3.2.), and in a second stage the methacrolein is oxidized to methacrylic acid:

$$CH_2{=}\underset{\underset{CH_3}{|}}{CH}{-}CH_3 + O_2 \longrightarrow CH_2{=}\underset{\underset{CH_3}{|}}{CH}{-}CHO + H_2O$$

$$CH_2{=}\underset{\underset{CH_3}{|}}{CH}{-}CHO + 1/2\,O_2 \rightarrow CH_2{=}\underset{\underset{CH_3}{|}}{CH}{-}COOH$$

A published account of process and catalyst developments [34], [35] contrasts methacrylic acid production from isobutene with a similar process for preparing acrylic acid from propene. Selectivity of the second-stage catalysts is best at modest conversions (65 – 85 %). In the Sumitomo–Nippon Shokubai process shown in Figure 2, the first-stage reactor is operated at high conversion, and its effluent passes directly to the second oxidation reactor [36]. Conversion in the second stage is kept low in order to ensure good catalyst selectivity and increase catalyst life. Unreacted methacrolein from the second-stage reactor effluent is separated and recycled. The overall yield of methacrylic acid from isobutene is about 65 – 70 %. Typical catalysts for the first stage are multicomponent metal oxides containing bismuth, molybdenum, and several other metals to promote activity and

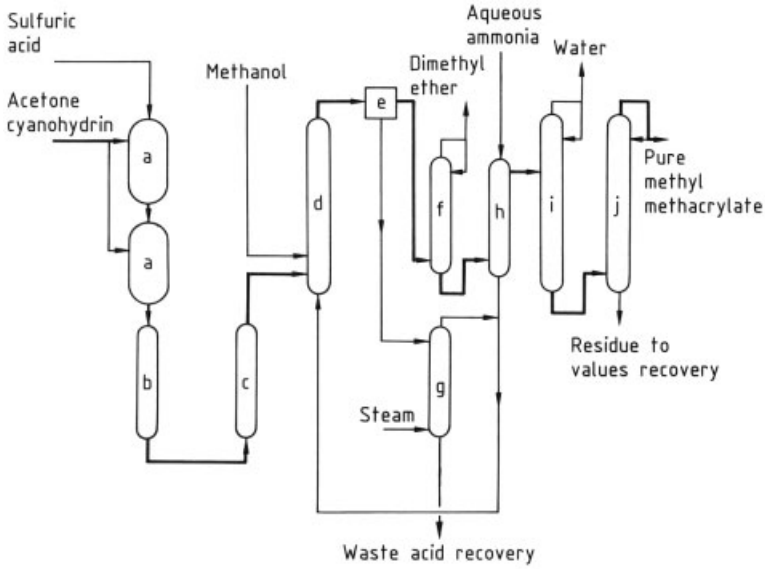

Figure 1. Production of methyl methacrylate from acetone cyanohydrin
a) Hydrolysis reactors; b) Cracker; c) Cooler; d) Esterification reactor; e) Separator; f) Flash column; g) Acid stripper; h) Wash column; i) Dehydration distillation column; j) Product distillation column.

modify selectivity [37–41]. Second-stage catalysts are based on phosphomolybdic acid, but they usually contain an alkali metal to control the acidity. Other elements such as copper and vanadium may also be present [42–46].

Reactor effluent from the second-stage oxidation passes to a quencher (b) where crude aqueous methacrylic acid is obtained. The gaseous effluent from the quencher is passed to an absorber (c) where the unreacted methacrolein is absorbed, usually in aqueous carboxylic acid. Absorber off-gases are sent to a combustion unit (d) before being discharged to the atmosphere. A portion of the incinerated gases may be recycled to the first-stage reactor to provide inert gas diluent for the feed. The methacrolein absorbate is transferred to a methacrolein recovery tower (e) from which methacrolein is recycled to the second-stage oxidation reactor; recovered absorbent solution is returned to the absorber.

The crude, aqueous methacrylic acid is sent to a solvent extraction unit (f) for methacrylic acid recovery. Next, a solvent recovery/dehydration tower (h) affords crude methacrylic acid as a bottoms product. The overhead organic solvent layer is recycled to the extraction step, whereas the overhead aqueous layer is combined with extractor raffinate and sent to a solvent stripping tower (g) before being subjected to wastewater treatment.

Dry, crude methacrylic acid from the extract stripper can be further purified if methacrylic acid is the desired end product, or it can be sent directly to an esterification reactor (i) where catalyst and methanol are added for conversion to methyl methacrylate. Crude ester is extracted with water (j) to recover excess methanol, which is removed by distillation (k) and recycled to the esterification reactor. The washed, crude ester is sent to a light ends stripper (l) for removal of light byproducts (e.g., methyl acetate) and then distilled (m) to provide pure methyl methacrylate. The bottoms from the final distillation column are recycled to the first extraction step with the exception of a small bleed for removal of inhibitor residues.

Plants based on the C-4 route were introduced in Japan by Nippon Shokubai in 1982 and by Mitsubishi Rayon in 1983 [35]. A joint venture plant operated by Sumitomo and Nippon Shokubai came on stream in 1984. Mitsui Toatsu and Kyowa Gas have also announced plans to construct a plant of this type. Several firms, including Nippon Kayaku, Mitsui Toatsu, Rohm and Haas, and Oxirane (a joint venture of Halcon SD and ARCO) have carried out extensive re-

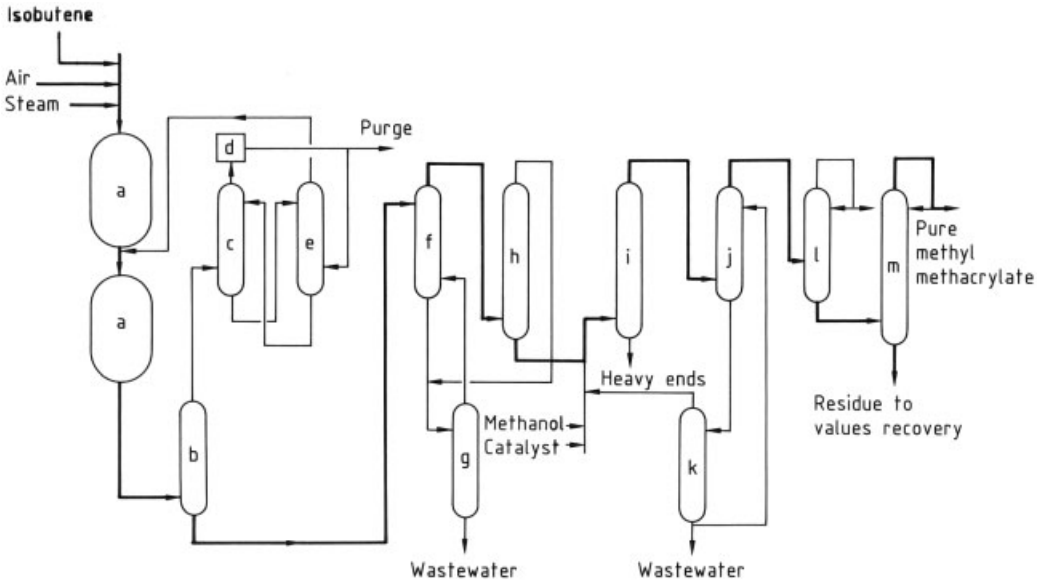

Figure 2. Production of methyl methacrylate from isobutene
a) Oxidation reactors; b) Quencher; c) Absorber; d) Combustion unit; e) Methacrolein recovery tower; f) Solvent extraction unit; g) Solvent stripping tower; h) Solvent recovery/dehydration tower; i) Esterification reactor; j) Extractor; k) Distillation unit; l) Light ends stripper; m) Product distillation column.

search on the isobutene process. In 1987 ARCO acquired exclusive worldwide rights to Halcon SD technology using C4-feedstocks [47].

Halcon has proposed a variation of this process that commences with dehydrogenation of isobutane [48], [49].

3.3. Methacrylic Acid from Isobutyric Acid

Acid-catalyzed carbonylation of propene to isobutyric acid, followed by oxidative dehydrogenation, presents still another route to methacrylic acid. In this case the starting material is propene itself rather than the oxydized derivative acetone, as in the acetone cyanohydrin route. Although the propene – isobutyric acid – methacrylic acid route is not currently in commercial use, several major methyl methacrylate manufacturers have research efforts aimed at its commercialization.

In the first stage of the process, propene, carbon monoxide, and water are reacted in the presence of a strong acid catalyst to produce isobutyric acid. Sulfuric acid, hydrogen fluoride, and boron fluoride have all been reported to be effective catalysts [50–53]. Patents to Röhm [54] and Ashland Oil [55] describe variations that involve preliminary preparation of isobutyroyl fluoride, which is then hydrolyzed to isobutyric acid. Alternatively, isobutyric acid may be synthesized directly by including carefully controlled amounts of water in the carbonylation step:

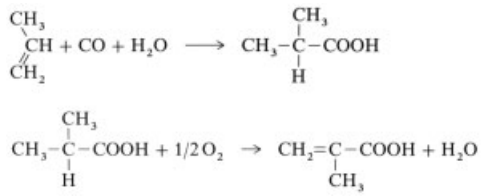

Hydrogen fluoride (which acts as both solvent and catalyst), carbon monoxide, and propene (14 – 40 : 1.5 : 1) are reacted in the presence of a slight stoichiometric deficiency of water relative to propene to generate isobutyric acid. Reaction conditions are included in the patent literature; these range from about 30 °C at 20 MPa to 120 °C at 14 MPa. Depending upon the temperature and pressure, residence time during the carbonylation step varies from about

5 to 30 min. Reactor effluent is passed to staged flash tanks, in the first of which excess carbon monoxide can be recovered for recycle to the carbonylation reactor. The second tank permits removal of inert gases, which can be passed to the atmosphere after scrubbing with caustic solution to remove any hydrogen fluoride or isopropyl fluoride. The bulk of the hydrogen fluoride is separated overhead in a distillation tower for recycle to the carbonylation reactor. The bottoms from this tower pass to a hydrolysis stripper where any remaining fluorine-containing materials are reacted with water; hydrogen fluoride is stripped off for recycle. A final distillation step provides isobutyric acid overhead for feed to the second part of the process. Bottoms from this distillation contain small amounts of C_7 and C_{10} carboxylic acids resulting from multiple condensations of propene prior to reaction with carbon monoxide. The overall selectivity with respect to propene is reported to be 95 – 97 %. The preceding steps must be carried out in a facility carefully designed to prevent fugitive fluoride emissions.

In the second stage of the process, isobutyric acid, steam, and air are passed over a fixed-bed catalyst [56–60] in a multitubular reactor, causing oxidative dehydrogenation to methacrylic acid. The reactor effluent is sent to a quench tower from which an aqueous methacrylic acid stream is obtained. This part of the process is similar to the methacrolein oxidation step in the C-4 process described in Section 3.2. In addition to carbon monoxide and carbon dioxide, which are incinerated with other noncondensable gases, the crude methacrylic acid stream contains acetone and acetic acid as byproducts. If desired, and with proper choice of quench tower conditions, the acetone can be directed to the incinerator along with the noncondensables.

If methacrylic acid is to be isolated, the crude product may be extracted into a solvent and dehydrated in a distillation tower. Isobutyric and acetic acids are then separated by distillation as light ends prior to final distillation of the methacrylic acid. The distillative separation of isobutyric acid from methacrylic acid is very difficult; normal boiling points of the two materials are 155 and 162 °C, respectively.

Catalysts for the oxidative dehydrogenation of isobutyric acid to methacrylic acid are of two general types. The first series, often referred to as Mo – P – V mixed oxide catalysts, were developed by Mitsubishi Chemical Industries [57], Röhm [58], [59], and others, and they are similar in composition to catalysts used in the oxidation of methacrolein. Most are phosphomolybdic acid derivatives, usually with some replacement of molybdenum by vanadium or tungsten. The better catalysts frequently contain at least some copper, and they are partially neutralized by cesium, rubidium, or potassium. Some catalysts are reported to achieve conversions of 99.8 % with selectivities above 74 % [59].

A second type of catalyst has been intensively studied by Ashland Oil [61]. These include iron phosphate materials and give selectivities of about 84 – 85 % at conversions in the 85 – 95 % range. Such catalysts are used at about 400 °C, in contrast to ca. 300 °C for the phosphomolybdate catalysts. Iron phosphate catalysts require high levels of steam in the reactor feed for optimum selectivity and life.

According to patents issued to Röhm [62], crude reactor product from the oxidative dehydrogenation step can be sent directly to the esterification reactor, which constitutes the starting point for the third section of the process. In this case, the product methyl methacrylate must be separated from methyl acetate and methyl isobutyrate. Separation of methyl isobutyrate from methyl methacrylate is difficult because its boiling point (92 °C) is close to that of methyl methacrylate (101 °C). The esterification step is similar to that described in Section 3.2 for the C-4 process.

Norsolor has an exclusive European license to the Ashland Oil technology, which Norsolor has further developed in a pilot plant at St. Avold (France) [63]. Röhm has developed its own process for this route in a pilot plant at Darmstadt [64].

An alternative process involves the hydroformylation of propene to give isobutyraldehyde, followed by oxidation of the aldehyde to isobutyric acid [65].

3.4. Methacrylic Acid From Ethylene

Other routes to methacrylic acid include the condensation of formaldehyde with propionic acid to generate methacrylic acid and the condensation of formaldehyde with propanal to give methacrolein [66–69].

BASF has developed such a process based on ethylene, synthesis gas, and formaldehyde. A plant with a production capacity of 40 000 t/a came on stream at Ludwigshafen in 1990 [70], [76].

Ethylene is first hydroformylated to give propanal, which is then condensed with formaldehyde to produce methacrolein. Catalytic air oxidation of methacrolein to methacrylic acid completes the synthesis, a step that is common to the C-2 and C-4 routes:

$$CH_2{=}CH_2 + CO + H_2 \longrightarrow CH_3CH_2CHO$$

$$CH_3CH_2CHO + HCHO \longrightarrow CH_2{=}\underset{\underset{CH_3}{|}}{C}HCHO + H_2O$$

An alternative process would entail oxidation of propanal to propionic acid, condensation of which with formaldehyde would give methacrylic acid directly.

4. Quality Specifications

Glacial methacrylic acid is typically available in a purity exceeding 99.4 % (analysis by GC and HPLC). Water (ca. 0.1 – 0.2 wt %) is the primary contaminant. Traces of α-hydroxyisobutyrate may also be present. The monomethyl ether of hydroquinone (MEHQ) at 100 – 250 ppm is used as an inhibitor during storage and transportation. Commercial methyl methacrylate is typically 99.9 % pure, with traces of acidity (< 0.003 %) and water (< 0.05 %). The storage and transportation inhibitor for the methyl ester is generally 10 – 50 ppm of MEHQ or 25 – 60 ppm of hydroquinine, though other phenolic inhibitors can also be used. Phenothiazine is an excellent inhibitor in this case because it has both anaerobic and aerobic activity. Nevertheless, it is not commonly used in products intended as polymer intermediates.

5. Storage and Transportation

The heat of polymerization of methacrylic acid is -56.5 kJ/mol; corresponding values for the methyl, ethyl, and *n*-butyl esters are -57.7, -57.7, and -56.5 kJ/mol, respectively. These highly exothermic reactions may occur with violence if the monomers are not properly inhibited against the formation of free radicals. Appropriate inhibitor levels have been established with great care by the manufacturers, resulting in excellent stability to polymerization under normal conditions of storage and transportation. Once stabilized, methacrylic acid and its derivatives can be handled as flammable liquids. The flammability limits of methyl methacrylate in air at 20 °C are 2.12 – 12.5 vol%.

The effectiveness of typical phenolic inhibitors depends upon the presence of oxygen. For this reason, methacrylic monomers should always be stored in contact with a vapor space whose oxygen concentration is 5 – 21 % (dry air) in order to maintain a sufficient level of dissolved oxygen [11].

Like many other unsaturated organic materials, methacrylic monomers react slowly with oxygen to give peroxides, hydroperoxides, and other compounds. Formation of such materials is greatly retarded by polymerization inhibitors and accelerated by increased temperature, but deterioration of product quality and stability must be anticipated upon extended storage (several months to a year). Storage temperatures should generally be kept below 25 °C.

Methacrylic acid has a relatively high freezing point (15.8 °C), and inhibitors tend to partition strongly into the liquid phase during freezing. Hence, thawing of frozen methacrylic acid can give initial melts that are significantly depleted in inhibitor. If such a liquid is heated, thermal initiation can give rise to violent polymerization. Care should thus be taken to avoid the possibility of freezing. If freezing takes place, thawing should preferably be conducted at 25 °C, and never above 40 °C. Furthermore, material should never be removed until melting is complete and the liquid has been thoroughly mixed to restore uniform inhibitor levels.

The preferred construction material for methacrylic acid storage vessels is type 316 stainless steel. Although less corrosion-resistant materials are suitable for methacrylate esters, stainless steel should be considered for these as well, because rust or iron salts can induce free radical decomposition of the peroxides that accumulate on long storage.

Transport Regulations. Methacrylic acid is classified as a flammable and corrosive liquid. Transport is governed by the IMDG-Code no.

D 8257, class B; UN-no. 2531; RID/ADR class 8, number 31 c; CFR 49 : 172.102 corrosive material.

6. Uses

Methacrylic acid and methacrylate esters are used to prepare a wide range of polymers [→ Polyacrylamides and Poly(Acrylic Acids), → Polymethacrylates]. Poly(methyl methacrylate) is the primary polymer in this category, and it provides water-clear, tough plastics that are used in sheet form in glazing, signs, displays, and lighting panels. Automotive lighting lenses and similar products can be prepared from molding pellets. Methyl methacrylate incorporated into copolymers forms the basis for durable coatings and inks. Higher methacrylate polymers are useful in the manufacture of oil additives, solventless inks and coatings, and binders for xerography. Salts of poly(methacrylic acid) can serve as the basis for water-soluble thickeners and detergent additives.

7. Economic Aspects

The cost of methyl methacrylate varied widely among the three major economic regions of the world in 1988. Japanese prices in early 1988 approached \$ 1.76/kg (80 cents/lb), the price in Europe was ca. \$ 1.43 – 1.48/kg (65 – 67 cents/lb), while the United States price was \$ 1.17 – 1.23/kg (53 – 56 cents/lb). World capacity is of the order of 1.4×10^6 t/a. The principal manufacturers and their existing capacities are listed in Table 3 [71], [72]. Projected capacities after completion of expansions and probable retirements of facilities are also given. The normal growth rate in demand is 3 – 3.5 %/a. Announced expansions by BASF in Germany, by Rohm and Haas and Du Pont in the United States, and by Sumitomo – NSKK and Mitsui Toatsu – Kyowa Gas Chemical in Japan are expected to satisfy demand into the 1990s.

8. Toxicology and Occupational Health

Acrylic monomers have been the subject of animal toxicity studies covering a variety of exposure routes. Larger esters are found to be less well-absorbed and less toxic than smaller esters. Generally speaking, methacrylates are less toxic than acrylates. Relevant data are provided in Table 4 [11], [73], [74].

With respect to acute toxicity, single exposures to acrylic monomers by the oral or dermal route appear to be slightly to moderately dangerous. This conclusion is based on lethality studies with rats and rabbits. The mucous membranes of the eyes, nose, and throat are particularly sensitive to irritation. Acrylic monomers can produce eye and skin irritation ranging from slight to corrosive depending upon the monomer.

Full eye protection should always be worn during handling of methacrylate monomers. Prolonged exposure to either liquid or vapor can result in permanent eye damage and possibly blindness. Overexposure to vapors causes nose and throat irritation, as well as dizziness or drowsiness (solvent narcosis) and possibly central nervous system depression. Thus, respiratory protection and proper ventilation are required in cases of exposure to high vapor concentrations.

Swallowing methacrylate monomers may lead to severe irritation of the mouth, throat, esophagus, and stomach, causing discomfort, vomiting, diarrhea, dizziness, and possibly collapse. Direct contact with the monomers causes skin redness and irritation ranging from slight to severe, perhaps including corrosion. Methacrylic acid is more corrosive than its esters. Because the monomers irritate and may be absorbed through the skin, gloves and protective clothing are required when handling these materials.

Repeated exposure to methacrylate monomers can produce allergic dermatitis (or skin sensitization), which results in rash, itching, or swelling. After one or more exposures to a given monomer, allergic responses sometimes are produced by single exposures to the same or other acrylic monomers.

Repeated exposure of animals to extremely high (near lethal) concentrations of monomer vapor has produced inflammation of the respiratory tract and degenerative changes in the liver, kidneys, and heart muscle. Liver and kidney changes have also been observed during prolonged oral exposure. These effects were the result of concentrations far above the threshold

Table 3. Worldwide methyl methacrylate capacity

Company	Location	Capacity, 10^3 t/a		Process ∗
		Current	Projected	
United States				
CyRo	Fortier (Louisiana)	80	92	ACN
ICI (formerly Du Pont)	Memphis (Tennessee)	134	191	ACN
Rohm and Haas	Deer Park (Texas)	300	360	ACN
Total		*514*	*643*	
Western Europe				
Norsolor	St. Avold (France)	60	60	ACN
Röhm	Worms (FRG)	115	115	ACN
Degussa	Wesseling (FRG)	45	45	ACN
BASF	Ludwigshafen (FRG)		36	C-2
Vedril (Montedison)	Rho (Italy)	50	50	ACN
ICI	Billingham (UK)	105	105	ACN
Paular	Tarragona (Spain)	30	30	ACN
Monacril (Röhm)	Palos-de-la-Fronters (Spain)	20	20	ACN
Total		*405*	*441*	
Japan				
Mitsubishi Rayon	Ohtake	75	75	ACN
		95	95	C-4
Japan Methacryl Monomer (Sumitomo/NSKK)	Niihama	40	80	C-4
Asahi Chemical	Kawasaki	60	60	MAN
Mitsui Toatsu	Mobara	14	0	ACN
Kyowa Gas Chemical	Nakajo	42	0	ACN
Mitsui Toatsu/ Kyowa Gas Chemical			40	C-4
Total		*326*	*350*	
Other countries		ca. 150	ca. 150	

∗ ACN = acetone cyanohydrin; MAN = methacrylonitrile.

Table 4. Toxicity data for methacrylic monomers

	Methyl methacrylate	Butyl methacrylates	Methyl acrylic acid
Oral LD_{50} (rat), g/kg	7.9 – 9.4	20.3	1 – 2
Dermal LD_{50} (rabbit), g/kg	> 9.4	10.2	> 2
Skin irritation (rabbit)	slight to moderate	moderate	corrosive
Eye irritation (rabbit)	slight to moderate	slight	corrosive
Inhalation LC_{50} (rat), ppm	7093	1720	> 1300
OSHA PEL, ppm [74]			20
TLV – TWA (1989 – 1990) [73]			
mg/m^3	410∗		70
ppm	100∗		20
Odor threshold, ppm	0.083	0.016	10

∗ These also represent the MAK values.

levels known to cause irritation. No such effects have been reported for humans, who presumably would leave a highly contaminated area rather than tolerate the irritation.

Current TLV – TWA and OSHA PEL values appear in Material Safety Data Sheets provided by methacrylate manufacturers upon request. Values for 1989 are also listed in Table 4 [73], [74].

Protection required for the safe handling of methacrylic acid and its esters includes chemical-resistant gloves and clothing, splash-proof goggles, and good ventilation of the workplace. Contact with these materials should be followed by flushing with copious amounts of water. Medical attention should be sought if symptoms appear.

9. References

1. E. Frankland, B. F. Duppa, *Ann. Chem. Pharm.* **136** (1865) 1 – 31.
2. S. Hochheiser: *Rohm and Haas,* University of Pennsylvania Press, Philadelphia 1986.
3. E. H. Riddle: *Monomeric Acrylic Esters,* Reinhold Publ. Corp., New York 1954.
4. J. W. Nemec, W. Bauer, Jr.: "Acrylic and Methacrylic Acid Polymers," in *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 1, J. Wiley and Sons, New York 1985, pp. 211 – 234.
5. *Ullmann,* 4th ed., **16,** 609 – 614.
6. H. Rauch-Puntigam, T. Volker: *Acryl- und Methacrylverbindungen,* Springer Verlag, Berlin 1967.
7. M. Salkind, E. H. Riddle, R. W. Keefer, *Ind. Eng. Chem.* **51** (1959) 1232 – 1238.
8. ICI, GB 405 699, 1934 (J. W. C. Crawford); *Chem. Abstr.* **28** (1934) 4745 – 9.
9. L. S. Luskin: "Acrylic Acid, Methacrylic Acid and Related Esters," in E. C. Leonard (ed.): *High Polymers, Vinyl and Diene Monomers,* vol. 24, part 1, Wiley Interscience, New York 1970.
10. L. S. Luskin: "Acrylic and Methacrylic Acids and Esters," in F. D. Snell, C. L. Hilton (eds.): *Encyclopedia of Industrial Chemical Analysis,* vol. 4, J. Wiley and Sons, New York 1967, pp. 181 – 218.
11. *Storage and Handling of Acrylic and Methacrylic Esters and Acids,* Bulletin 84C7, Rohm and Haas Co., Philadelphia 1987.
12. *Specifications and Typical Properties of Acrylic and Methacrylic Monomers,* Bulletin 77S2, Rohm and Haas Co., Philadelphia 1989.
13. *Glacial Methacrylic Acid and Glacial Acrylic Acid,* Publication CM41A, Rohm and Haas Co., Philadelphia 1975.
14. J. Kooi, *Rec. Trav. Chim. Pays Bas* **68** (1949) 34 – 42.
15. A. F. Frolov, M. A. Longinova, B. F. Ustavshchikov, *Zh. Obshch. Khim.* **36** (1966) 180 – 184.
16. A. F. Frolov, *Russ. J. Phys. Chem. (Engl. Transl.)* **39** (1965) 1538 – 1541.
17. A. F. Frolov, M. A. Longinova, A. P. Saprykina, A. B. Kundakova, *Zh. Fiz. Khim.* **36** (1962) 3282 – 3283.
18. E. A. Frolova, B. F. Ustavshchikov, S. Yu. Pavlov, *Zh. Fiz. Khim.* **48** (1974) 1865.
19. G. A. Chubarov, S. M. Danov, R. V. Efremov, *Zh. Fiz. Khim.* **48** (1974) 1047 – 1048.
20. G. A. Chubarov, S. M. Danov, G. V. Brovkina, *Zh. Prikl. Khim. (Leningrad)* **51** (1978) 1899 – 1900.
21. L. S. Luskin: "Basic Monomers: Vinyl Pyridines and Aminoalkyl Acrylates and Methacrylates," in R. H. Yocum, E. B. Nyquist (eds.): *Functional Monomers, their Preparation, Polymerization and Application,* vol. 2, Marcel Dekker, New York 1974, pp. 663 – 734.
22. I. N. Azerbaev, B. M. Butin, Y. G. Bosyakov, *Zh. Obshch. Khim.* **45** (1975) 1730 – 1734.
23. *Kirk-Othmer,* 3rd ed., **15,** 346 – 376.
24. Kohjin, DE-OS 2 217 623, 1972; US 3 813 438, 1974 (A. Oshima, K. Tsuboshima, N. Takahashi).
25. Rohm and Haas, US 3 087 962, 1963 (N. M. Bortnick).
26. Nitto Chemical Industry, DE-OS 2 752 109, 1978; US 4 301 297, 1981 (Y. Kametani, Y. Ino).
27. American Cyanamid, US 4 059 617, 1977 (T. Foster, T. S. Dawson).
28. Japan Catalytic Chemical Industry, DE-OS 2 311 007, 1977; US 3 872 161, 1975 (S. Fukuchi, N. Shimizu, T. Ohara).
29. Distillers, GB 1 120 301, 1968 (E. J. Percy, J. A. Wickings); Chem. Abstr. 69 (1968) 58 829 t.
30. Japan Gas-Chemical, DE-OS 2 027 444, 1970 (M. Murayama, K. Abe); *Chem. Abstr.* **74** (1971) 63 938 k.
31. Japan Oil and Fats, JP 72 51 328, 1972 (Y. Tanizaki, Y. Kubo); *Chem. Abstr.* **79** (1973) 19 377 f.

32. ICI, US 2 140 469, 1938 (J. W. C. Crawford, J. McGrath).
33. BASF, US 3 002 023, 1961; GB 873 603, 1961 (H. Fikentscher, H. Wilhelm).
34. S. Nakamura, H. Ichihashi in T. Seiyama, K. Tanabe (eds.): "Studies in Surface Science and Catalysis," *Proceedings 7th International Congress on Catalysis,* Tokyo, 30 June – 4 July 1980, Elsevier, New York 1981, pp. 755 – 767.
35. T. Nakamura, T. Kita, *CEER Chem. Econ. Eng. Rev.* **15** (1983) 23 – 27.
36. Nippon Shokubai Kagaku Kogyo Co. and Sumitomo Chemical Co., *Hydrocarbon Process.* **64** (1985) 143.
37. Nippon Kayaku KK, US 4 012 449, 1977 (Y. Shikakura, F. Sakai, H. Shimizu).
38. Ube Industries, US 4 171 328, 1979 (S. Umemura, K. Ohdan, K. Suzuki, T. Hisayuki).
39. Mitsubishi Rayon Co., DE-OS 2 427 670, 1975; US 3 972 920, 1976 (H. Ishii, H. Matsuzawa, M. Kobayaashi, K. Yamada).
40. Rohm and Haas, US 4 306 090, 1981; US 4 280 928, 1981 (L. S. Kirch, W. J. Kennelly).
41. Rohm and Haas, US 4 358 622, 1982 (J. W. Nemec, M. S. Cholod).
42. Sohio, US 4 136 110, 1979 (J. F. White, J. R. Rege).
43. Ube Industries, US 4 364 844, 1982 (S. Umemura et al.).
44. Sohio, US 4 301 031, 1981 (W. G. Shaw, P. L. Kuch, C. Paparizos).
45. The Halcon SD Group, US 4 374 757, 4 374 759, 1983 (S. Khoobiar).
46. Nippon Kayaku, US 4 172 051, 1979; US 4 273 676, 1981; US 4 467 113, 1984 (M. Matsumoto, A. Sudo, H. Sugi).
47. *Chem. Week* (1987) Aug. 26, 41.
48. R. V. Porcelli, B. Juran, *Hydrocarbon Process.* **65** (1986) 37 – 43.
49. The Halcon SD Group, US 4 413 147, 1983; US 4 532 365, 1985; US 4 535 188, 1985 (S. Khoobiar).
50. Sinclair Refining, US 2 975 199, 1961 (B. S. Friedman, S. M. Cotton).
51. Chem. Systems, US 4 256 914, 1981 (J. A. Jung, J. Peress).
52. Ashland Oil, US 4 303 594, 1981 (R. Norton et al.); US 4 495 110, 1985 (J. E. Corn).
53. Röhm, DE-OS 3 033 655, 1982; US 4 504 675, 1985 (S. Besecke, G. Schroeder, H. Siegert).
54. Röhm, DE-OS 3 213 395, 1983; US 4 452 999, 1984 (S. Besecke, G. Schroeder, H. Siegert).
55. Ashland Oil, US 4 270 983, 1981 (B. C. Trevedi, D. Grote, T. O. Mason).
56. Ashland Oil, US 4 299 980, 1981 (C. Daniel, P. Brusky).
57. Mitsubishi Chemical Industries, US 4 061 673, 1977 (T. Onoda, M. Otake).
58. Röhm, DE-OS 3 019 731, 1981 (H. Siegert).
59. Röhm, US 4 370 490, 1983 (W. Gruber, G. Schroeder).
60. M. Akimoto, Y. Tsuchida, K. Sato, E. Echigoya, *J. Catal.* **72** (1981) 83 – 94; M. Akimoto, K. Shima, H. Ikeda, E. Echigoya, *J. Catal.* **86** (1984) 173 – 186.
61. Ashland Oil, US 4 410 728, 1983; US 4 439 621, 1984 (C. Daniel).
62. Röhm, DE-OS 3 146 191, 1983 (W. Gaenzler, H. Siegert, H. J. Hohage, G. Schroeder); Chem. Abstr. 99 (1983) 38 912 f.
63. *Chem. Week* (1985) Oct. 23, 10 – 11.
64. *Chem. Ind. Int. (Engl. Transl.)* (1986) no. 1, 4.
65. M. Otake, T. Onoda in T. Seiyama, K. Tanabe (eds.): "Studies in Surface Science and Catalysis," *Proceedings 7th International Congress on Catalysis, Tokyo 1980,* Elsevier, New York 1981, pp. 780 – 791.
66. Röhm, US 4 147 718, 1979 (W. Gaenzler, K. Kabs, G. Schroeder).
67. Rohm and Haas, US 3 933 888, 1976 (F. Schlaefer).
68. Sohio, US 4 339 598, 1982 (A. T. Guttmann, R. K. Grasselli).
69. Rohm and Haas, US 4 490 476, 1984 (R. J. Piccolini, M. J. Smith).
70. S. J. Ainsworth, D. Hunter, S. Ushio, E. Johnson, *Chem. Week* **144** (1988) 64 – 65.
71. *Chem. Market. Rep.* **233** (1988) no. 19, 3.
72. A. S. Bakshi, *Oil Gas J.* **83** (1985) 99 – 101.
73. *Threshold Limit Values and Biological Exposure Indices for 1989 – 1990,* American Conference of Governmental Industrial Hygienists, Cincinnati, Ohio 1989.
74. *Fed. Regist.* **54** (1989) no. 12, 2631.
75. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998, p. 309
76. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998, p. 313

Methane

EDMOND-JACQUES NYNS, Unité de Génie Biologique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium

1. **Introduction** 1
2. **Physical and Chemical Properties of Biogas** 2
3. **Biological Methanogenesis** 2
4. **Technological Biomethanation** 3
5. **Feedstocks for Biomethanation** 5
6. **Process and Environmental Parameters for Stable, Efficient Biomethanation** 6
7. **Environmental Impact of Biomethanation** 8
8. **Storage, Pretreatment, Upgrading and Uses of Biogas** 9
9. **Analysis** 9
10. **Biogas in Developing Countries** 10
11. **Economic Aspects** 10
12. **References** 11

1. Introduction

Methane [*74-82-8*], CH_4, M_r 16.03, *mp* − 182 °C, *bp* − 162 °C, is a colorless gas that burns with a blue flame. Some physical properties of methane are as follows:

Specific heat capacity at − 100 °C	5.34 J g^{-1} K^{-1}
Density at − 170 °C	0.4362 g/cm^3
Viscosity at − 170 °C	0.142 mPa · s
Surface tension at − 170 °C	15.8 mN/m
Latent heat of fusion	0.942 kJ/mol
Latent heat of evaporation at	8.185 kJ/mol
$Kp_{101.3\,kPa}$	≈ 0.510 kJ/g
Minimum calorific value	50.0 kJ/g
Heat of formation	− 74.897 kJ/mol
Entropy	186.31 J K^{-1}mol^{-1}
Free energy of formation	50.83 kJ/mol
Critical values:	
Temperature	− 82.5 °C
Pressure	4.67 MPa
Density	0.162 g/cm^3
Minimum auto-ignition temperature (MAIT) (VDE 0173)	535 °C
MAIT varies according to method of determination	between 601 and 748 °C
Flame velocity	43.4 cm/s
Flame velocity varies according to method of determination	between 34 and 44 cm/s
Explosion limits in air at 20 °C and 0.1 MPa	
lower	5.0 vol %
upper	15.0 vol %
Compressibility factor	0.299
Solubility in water	
at 20 °C, 101.3 kPa CH_4	33.8 mL/L
at 35 °C, 101.3 kPa CH_4	25.4 mL/L
Henry's constant in water	
at 20 °C	1.45 × 10^{-3} mol/L atm
at 35 °C	1.005 × 10^{-3} mol/L atm

Methane is more soluble in organic solvents than in water. Ethanol dissolves 32 cm^3 methane per 100 cm^3 at 0 °C. Diethylether dissolves 106.6 cm^3 methane per 100 cm^3 at 0 °C. Methane forms a hydrate, $CH_4 \cdot 6\,H_2O$, which has a decomposition temperature at 0.1 MPa of − 29 °C, a critical decomposition temperature of 21.5 °C and a dissociation pressure at 0 °C of 2.6 MPa.

Naturally occurring methane is essentially of biological origin. Methane has been produced in nature by the dismutation of organic matter since the origin of life on earth. Fossil methane is described in → Natural Gas. Methane is formed in the production of gas from coal or hydrocarbons (see → Gas Production) and in refining of oil (→ Oil Refining).

All organic matter, left to itself, will ultimately be transformed (dismutated) by microorganisms under anaerobiosis into carbon dioxide and methane.

Indeed, nature obtains energy through redox reactions. The ultimate reducing step is the reduction of carbon to methane, and the ultimate oxidizing step, the oxidation of carbon to carbon dioxide [1].

The biological process by which organic matter is transformed into methane and carbon dioxide is called methanogenesis. When this biological process is housed in a reactor to create a technological process, it is known as biomethanation. The term anaerobic digestion is also widely used, but may lead to confusion with other anaerobic digestion processes that do not involve gen-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_453

eration of methane. The mixture of carbon dioxide and methane is called biogas, a registered trade name of the Institute of Gas Technology in Chicago, United States, but widely used by the public.

Biomethanation is of interest, firstly because the biogas produced is an energy vector of economic interest, and secondly because the raw material or feedstock is often waste, which can hence be treated by the process [2], [3]. Wastewaters can be treated by biomethanation to yield acceptable effluents for receiving surface waters, and sludges can be treated to yield stabilized end products with useful agricultural properties [4–6].

2. Physical and Chemical Properties of Biogas

Biogas contains variable proportions of carbon dioxide and methane, usually somewhat more of the latter than of the former. Biogas is usually saturated with water and also contains minor and trace compounds, the major of which is hydrogen sulfide (see Table 1). The composition of biogas depends firstly upon the composition of the feedstock from which it originates and secondly upon the physicochemical equilibrium reached or maintained in the digesting mixed liquor [7].

Table 1. Composition of biogas

Component	Content, vol %
Methane	52 – 95
Carbon dioxide	9 – 45
Hydrogen sulfide	0.001 – 2
Hydrogen	0.01 – 2
Nitrogen	0.1 – 4
Oxygen	0.02 – 6.5
Argon	0.001
Carbon monoxide	0.001 – 2
Ammonia	trace
Organics	trace

Biogas has many of the chemical and physical properties of its parent constituent methane. Of course, the presence of carbon dioxide changes the chemical and physical properties. As an example, Table 2 shows the dependence of the density of biogas upon composition.

Table 2. Densities of methane and biogas of various composition

Methane content, %	50	60	80	90	100
Gas density relative to air	1.040	0.942	0.745	0.652	0.555

Whereas methane has a minimum calorific value of 50.0 MJ/kg (35.9 MJ/m^3), upgraded biogas containing 90 vol % methane has a minimum calorific value of 45.0 MJ/kg (32.3 MJ/m^3). Mean biogas, with 60 vol % methane, has a minimum calorific value of 30.0 MJ/kg or 21.5 MJ/m^3.

The flammability of biogas is shown in Figure 1. The chemical and physical properties of biogas are collected in [8].

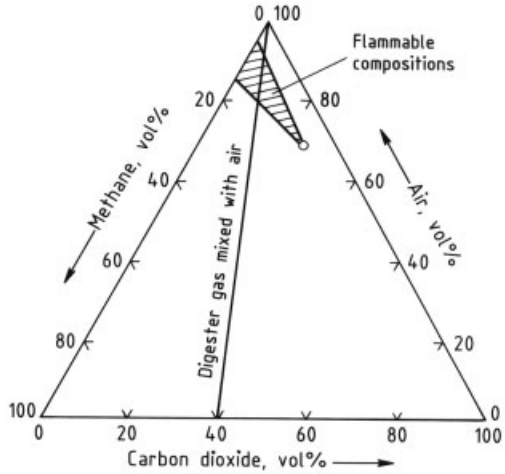

Figure 1. Flammability limits of biogas
The flammability limits of biogas are included in the dashed triangle. Point ○ refers to a gas mixture of 24 % carbon dioxide, 68 % air, and 8 % methane
From ref. [8], with permission from Elsevier Applied Science

3. Biological Methanogenesis

Microorganisms transform organic matter into methane by the process depicted in Figure 2. The process consists basically of two steps, linked by interspecies hydrogen transfer. In the first step, fermentative bacteria dissolve, depolymerize, and metabolize organic matter into acetate, inorganic carbon (e.g., CO_2, HCO_3^-, and CO_3^{2-}), and hydrogen. In the second step, methanogenic microbial species dismutate acetate into methane

and inorganic carbon, and reduce inorganic carbon with hydrogen to produce methane:

$$CH_3COO^- + H_2O \rightleftharpoons CH_4 + HCO_3^-$$
$$\Delta G^{0\prime} = -31.0\,\text{kJ/formular conversion}$$

$$4\,H_2 + HCO_3^- + H^+ \rightleftharpoons CH_4 + 3\,H_2O$$
$$\Delta G^{0\prime} = -135.6\,\text{kJ/formular conversion}$$

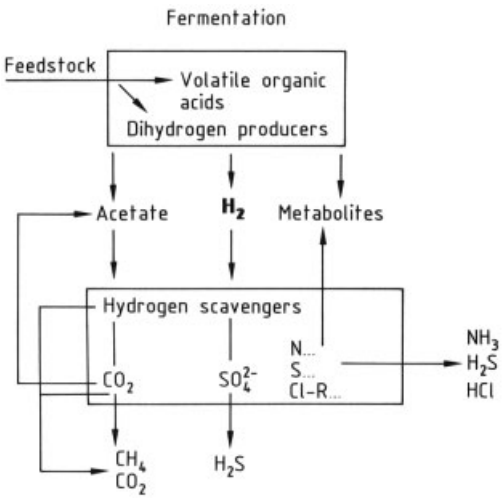

Figure 2. Flowsheet for biological methanogenesis

Whenever the first fermentative step produces organic metabolites other than acetate, for example, anions of short-chain carboxylic acids such as propionate and butyrate and alcohols such as ethanol or butanol, a link process, mediated by obligate hydrogen-producing acetogenic bacteria, transforms these metabolites into acetate and hydrogen. For thermodynamic reasons, the latter transformation can only take place at very low partial pressures of hydrogen. In stable, efficient methanogenic processes, the hydrogen partial pressure ranges between 1 and 10 Pa. An abrupt 2 – 4-fold increase in hydrogen partial pressure or an hydrogen partial pressure above 20 Pa is indicative of misfunctioning of the process. The increased partial pressure of hydrogen results in accumulation of intermediate organic metabolites, and in the vast majority of cases, the pH drops below 6 and methane generation stops.

The methanogenic microorganisms are archaebacteria [9], which are probably the oldest known living organisms. Other microorganisms interfere with hydrogen scavenging. Acetogenic bacteria can reduce inorganic carbon with hydrogen to produce acetate, and sulfate-reducing bacteria reduce sulfate with hydrogen to form hydrogen sulfide:

$$H^+ + SO_4^{2-} + 4\,H_2 \rightleftharpoons HS^- + 4\,H_2O$$
$$\Delta G^{0\prime} = -152\,\text{kJ/formular conversion}$$

More recently, it has been suspected that some hydrogen-scavenging microorganisms are also able to reduce organic molecules containing nitrogen, sulfur, or halogens. The methanogenic consortium appears in this way to play a fundamental role in detoxication and, hence, in the environmental protection of humans.

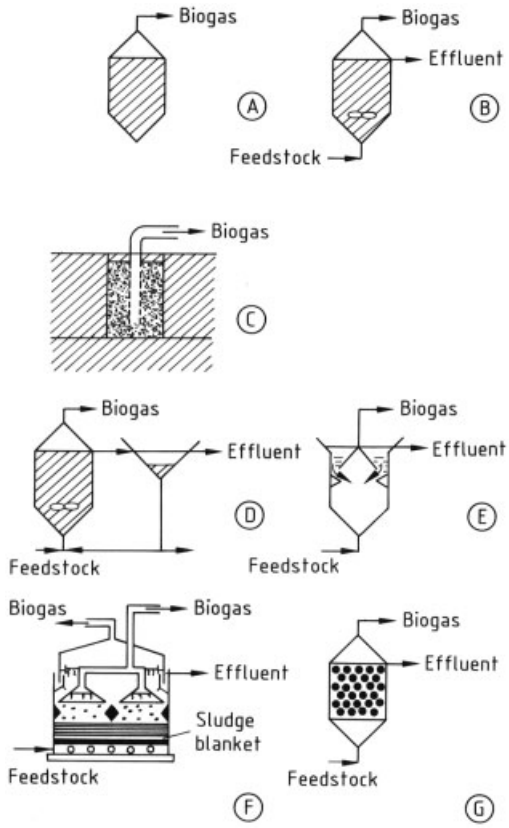

Figure 3. Schematics of the main designs of biomethanation reactors

4. Technological Biomethanation

The biological process of methanogenesis was discovered by VOLTA in 1776, but it was over

100 years later that this biological process was first exploited technologically. MOURAS in France (1881), SCOTT-MONCRIEFF (1890) and CAMERON (1892) in the United Kingdom, HOUSTON (1892) and TALBOT (1894) in the United States, and IMHOFF (1905) in the Federal Republic of Germany, were the first to apply methanogenesis to reduce the organic matter content of sewage sludge and wastewater [10].

There are three generations of reactors for biomethanation [11]. The first generation includes the completely mixed reactor, run either in batch or in continuous mode (Fig. 3 A and B). In the batch system, the feedstock is introduced at once and left in the reactor until the reaction is terminated. In the continuous mode, the feedstock is continuously loaded into the reactor and the effluent continuously removed (e.g., by overflow) at the same flow rate. Batch systems are mainly used for more solid feedstocks, whereas the continuous mode is better adapted to more fluid feedstocks.

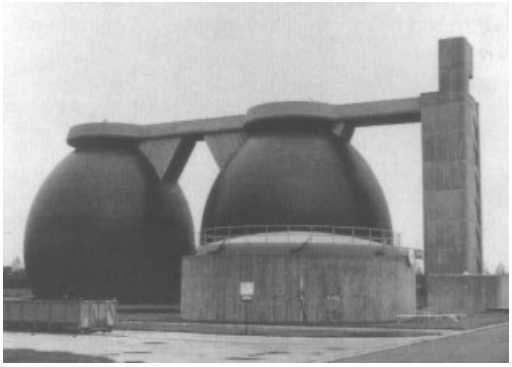

Figure 4. Egg-shaped reactor for the two-stage biomethanation of municipal sludge in the continuous mode at Schijnpoort, Antwerp, Belgium (courtesy of Biotim, Antwerp, Belgium)

As an example, in the European Community, there were in 1988 about 27 400 aerobic municipal wastewater treatment plants. Sewage sludge produced by these plants can be evaluated at 9×10^6 t/a of dry matter. The biomethanation of two-thirds of this represents a potential production of biogas equivalent to 1.5×10^6 t of oil per year [12]. Figure 4 shows a typical egg-shaped anaerobic digester for municipal sewage sludge.

As another example, animal manure has been frequently biomethanized in reactors run in the continuous mode to produce biogas as an energy source and stabilized manure as a fertilizer and soil conditioner. In the European Community in 1989, about 470 biogas plants on farms were in existence. In rural areas of developing countries, reactors of 5 – 20 m^3 capacity, run in the continuous mode on animal waste, mainly cattle dung, exist in very large numbers for producing energy and fertilizer at the family level. More than three million such digesters are in use in China, more than one million in India, and several thousand in other countries.

An interesting example of biomethanation in the batch mode is given by landfills (Fig. 3 C) [13]. Very large amounts of domestic solid waste are disposed of worldwide as huge landfills (several hundreds to several millions of tonnes each). Large amounts of biogas (several thousands of cubic meters per hour) are produced by natural fermentation in these landfills over long periods of time (10 – 20 years). The landfill gas escapes to the atmosphere, presenting a fire and explosion hazard, a source of damage to nearby vegetation, and a substantial contribution to the greenhouse effect. It can be made cost effective to extract and utilize the biogas produced in landfills. This environmental biotechnological industry started in the United States in the early 1970s, mainly in California, and spread in Europe, mainly in the Federal Republic of Germany and the United Kingdom. In the European Community, 107 landfills where biogas is exploited were identified in 1988, and this number is growing.

However, it quickly appeared that the first generation of anaerobic reactors, operating either in the batch mode or in the continuous mode without recycle (known in the United States as once-through mode), present severe limitations of performance for a number of reasons, the most important of which is the lowering in the concentration of active microorganisms (often referred to as active biomass) that arises because active biomass is lost when the effluent is removed.

A second generation of reactors for biomethanation was developed, in which the active biomass is either trapped in the anaerobic reactor or is recycled to it after sedimentation (Fig. 3 D and E). Sedimentation occurs either in a separate decanter or in a built-in decanter. The latter type was developed as early 1970 under the trade name Clarigester. In the Clarigester, the liquid waste enters at the bottom of

the reactor and flows upwards through a bed of active biomass. It then overflows into a built-in decanter, in which the active biomass sediments and returns to the reactor, while the effluent escapes from the top of the digester.

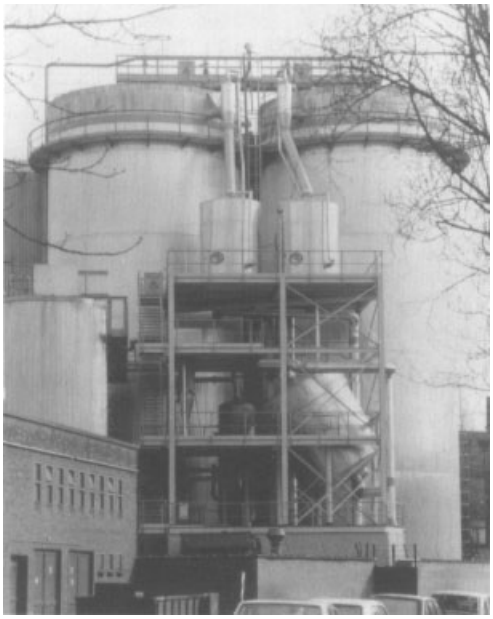

Figure 5. Tower anaerobic filter for the two-step anaerobic treatment of starch-containing wastewater
The towers are 22 m high and are filled with 16 layers of polyurethane foam matrices of 1.2 m thickness
By courtesy of Biotim, Antwerp, Belgium

It was found that the active biomass, under conditions that are still not well understood, organizes itself as heavy granules. This led to the upflow anaerobic sludge blanket (UASB) system (Fig. 3F), which was developed since 1970 by Lettinga in the Netherlands [14] and is a milestone in the application of biomethanation to the anaerobic digestion of agro-industrial and other wastewaters.

In a variant, the active biomass is either trapped within a polymeric foam, notably polyurethane, or adsorbed on an inert carrier (e.g., baked clay or plastic rings) (Fig. 3 G). The latter system carries the generic name anaerobic filter. The active biomass is referred to as a fixed film. When the inert carrier, usually large particles or matrices, is mainly immobile in the reactor, the system is known as a stationary bed. When the inert carrier, usually small particles such as sand or active carbon, expands and becomes mobile in the reactor, the system is termed a fluidized bed. Figure 5 shows a typical full-scale anaerobic filter for the anaerobic treatment of starch-containing wastewaters.

The third generation of biomethanation systems includes combinations of the former designs. For example, because the biological process consists of two steps, it has been found advantageous in some cases to optimize each biological step in a separate reactor. This is referred to as a two-step system and was first introduced in 1970 [15]. Other two-stage systems have an active biomethanation in the first stage and a finishing biomethanation stage with effluent clarification in the second stage. These are often applied to the biomethanation of sewage sludge and this explains, as can be seen in Figure 4, why egg-shaped digesters usually operate in pairs.

5. Feedstocks for Biomethanation [16]

Feedstocks for biomethanation include energy crops and organic wastes. Energy crops are plants grown mainly for producing energy. They include terrestrial and aquatic plants; the latter have been more intensively investigated. Typical aquatic energy crops are water hyacinths (*Eichhornia*) in soft waters and marine algae such as kelps (*Macrocystis*).

Organic wastes have become classical feedstocks for biomethanation [17]. They can be classified as domestic, industrial, or agricultural, according to their origin [18]. Feedstocks of domestic origin include municipal wastewaters, sewage sludges, and municipal solid wastes. Municipal wastewaters are progressively being investigated for anaerobic treatment. Sewage sludges have been stabilized by biomethanation for over 100 years. Domestic solid wastes are routinely biomethanized in landfills but are also being considered as feedstocks for biomethanation in engineered reactors.

Industrial wastes as feedstocks for biomethanation presently include mainly food wastes, originating from breweries, distilleries, fermentation industries, processing of coffee, fat, meat, milk, shellfish, fruit and vegetables, olive and palm oil mills, slaughterhouses, and starch and sugar refineries. They also include an increasing proportion of other agro-industrial wastes

such as those originating from pulp and paper mills and the textile and tanning industries. Even pharmaceutical, (petro)chemical and coal gasification wastes are presently being considered for biomethanation. Industrial wastes range in physical state from wastewaters through sludges to solid material. Agro-industrial wastewaters are now routinely treated anaerobically. More than one hundred mature, economic, and reliable full-scale plants have been built worldwide [19].

Agricultural wastes as feedstocks for biomethanation consist mainly of animal wastes [20]. Whereas all kinds of animal manure are suitable for biomethanation, cattle and pig manure are the main targets, not only for energy and methane production but also for pollution control. Very large scale biogas plants, have been developed in Denmark, Italy and The Netherlands, among others, that are intended for a group of farms appear more reliable and profitable than smaller biogas plants on single farms. Agricultural and animal wastes for biomethanation are of increased interest for developing countries.

In the European Community, the potential for biogas from domestic waste is equivalent to 10×10^6 t of oil; the potential for biogas from agricultural waste is equivalent to 23×10^6 t of oil and the potential for biogas from energy crops could be $9-57\times10^6$ t of oil [19]. In the European Community, there already exists a total biomethanation reactor volume of 6×10^5 m^3 producing 0.49×10^6 m^3 of methane per day (equivalent to 0.15×10^6 t of oil per year).

6. Process and Environmental Parameters for Stable, Efficient Biomethanation

The process parameters characteristic for running a biomethanation system are the space loading rate, the concentration of the feedstock, and the mean residence time. The space loading rate is the mass of feedstock (as total dry matter, organic matter, or COD) per unit digester working volume per day. The ratio of the concentration to the space loading rate gives the mean residence time (excluding any accumulation factor arising from trapping of solid feedstock by sedimentation in imperfectly mixed reactors), usually in days. The state parameters characteristic of the performance of a biomethanation system are the rate of gas production and the residual concentration of dry organic matter in the effluent or its residual COD. The yield is calculated either by the ratio of produced methane to added or consumed feedstock, or from the residual concentration in the effluent. Stability of the methanogenic process is an important aspect of performance. The process parameters related to stability are the concentration of volatile fatty acids (indicative of the equilibrium within the methanogenic microbial consortium), the hydrogen partial pressure (see Chap. 3), and the alkalinity or bicarbonate concentration (linked to the acidification potential of the feedstock).

Degradation of 1 kg of COD produces stoichiometrically 350 L (at STP) or 250 g of methane. One kilogram of average organic matter corresponds to 1 – 1.4 kg of COD, depending on its mean oxidation state.

In completely mixed systems conducted in the continuous mode without recycle, practical space loading rates remain below 3 kg per cubic meter per day. The mean retention time must remain above 14 d, and this severely limits the concentration of the feedstock. Gas production rates range between 0.5 and 1.5 (sometimes up to 3) volumes of gas per unit digester volume per day.

In systems with trapped active biomass, ca. 10 – 20 and, in exceptional cases, up to 40 kg of active biomass can be housed in 1 m^3 of digester volume. The food to microorganism ratio is the potential amount of feedstock transformed per day per kilogram of active biomass. Acceptable values of food to microorganism ratio range around 1, sometimes higher. Space loading rates in these systems with trapped biomass can thus be increased up to 10 – 40 kg per cubic meter per day. Mean residence times can be lowered to one day or less with favorable feedstocks, i.e., those for which the fermentative step is fast, such as sugar-containing wastewaters. Gas production rates can increase to 4 – 10 volumes of gas per unit reactor volume per day.

Yields vary from 30 % to over 95 %. Animal manure usually gives low yields: 30 % for cattle manure and up to 45 % for pig manure. Sugar-containing wastewaters can be 95 % digestible (at least in terms of BOD). Planning of biogas plants has been reviewed in [21], [22].

Several factors govern the process of methanogenesis in the technology of biomethanation, namely temperature, pH, anaerobiosis, carbon to nitrogen ratio, micronutrients, toxic elements in the feedstock, and inoculum.

There are two temperature ranges for optimum biomethanation, the mesophilic range at ca. 35 °C and the thermophilic range at ca. 60 °C. Thermophilic biomethanation results in faster gas production and higher yields, but the process is often less stable. The requirement for process temperatures above ambient temperature means that methane digesters need to be well insulated and that external process heat is often required. Part of the biogas produced (ca. 30 %) is usually used for this purpose. Between 20 and 30 °C, both methane production and methanogenic microbial growth are still possible, but at lower rates and yields. Below 20 °C, methane production remains possible, but microbial growth becomes progressively slower.

Biological methanogenesis occurs at pH 6.8 – 8.2. A pH below 6.5 inhibits the second, methanogenic step of the biological process but not the first, fermentative step. Under these conditions, the metabolization of organic carbon yields organic anions, for which the only possible pairing cation is the hydrogen ion. As a result of the accumulation of acid anions, the pH of the digesting mixed liquor drops. Bicarbonate alkalinity limits this pH drop by acting as a buffer and through formation of carbon dioxide.

$$H_3O^+ + HCO_3^- \overset{-H_2O}{\rightleftharpoons} H_2CO_3 \rightleftharpoons H_2O + CO_2\,(sol) \overset{-H_2O}{\rightleftharpoons} CO_2\,(g)$$

This explains in part why biogas contains carbon dioxide besides methane. The buffering action of nitrogenous compounds is discussed below.

Methanogenesis is a strictly anaerobic process. Some species of the methanogenic consortium are very sensitive to oxygen. However, the methanogenic consortium also includes facultative aerobic – anaerobic microbial species. Therefore, the methanogenic fermentation mixture can easily overcome an occasional limited exposure to air. The fermentative first step of the biological process of methanogenesis occurs at low redox potentials (E^0 well below 0 mV). The methanogenic second step occurs at even lower redox potentials (E^0 a couple of hundred mV below).

The carbon to nitrogen ratio is an important parameter for efficient biomethanation. First, nitrogen is required for cell growth. Average cell yield is 15 g of dry cells per 100 g of transformed feedstock in anaerobiosis. The nitrogen requirement is 1.7 g per 100 g feedstock. Second, the mineralization of organic nitrogen yields ammonia and ammonium ions. Ammonia produced in this way exerts a buffering capacity complementing that of bicarbonate and is therefore useful in pH control. Optimum mass ratios of carbon to nitrogen given in the literature range between 16 and 19. Shortage of nitrogen in a feedstock can best be overcome by mixing it with another feedstock richer in nitrogen.

Micronutrients are essential for efficient biomethanation. Metal ions seem to be necessary for methanogenic archaebacteria, notably up to 50 mM Na^+ and 2 – 100 mM Fe(II). Cobalt (50 nM) and nickel (100 nM) are essential trace elements for the reduction of carbon dioxide to methane by hydrogen.

Toxic compounds are often present in feedstocks at concentrations that inhibit biomethanation. For example, ammonia is inhibitory at concentrations of 1.7 g/L (as NH_4^+/NH_3). These concentrations are often encountered in pig manure. Heavy metal cations inhibit biological methanogenesis. However, most heavy metal cations form insoluble sulfides. Appropriate amounts of sulfate in the feedstock, a substrate for sulfate-reducing bacteria, normal hydrogen-scavenging members of the methanogenic consortium, help to overcome this potential inhibitory effect. This phenomenon has been conceived as a possible environmental biotechnology for the removal of heavy metals by concentration of insoluble compounds.

Many organic compounds are toxic to methanogenic microorganisms; an example is chloroform. Antibiotics such as monensin are used in animal feed to prevent methanogenesis in the rumen. Hence, animal manure may contain inhibitory concentrations of these antibiotics.

However, the methanogenic microbial consortium exhibits astonishing acclimation to toxic compounds [23]. Furthermore, the methanogenic microbial consortium includes detoxifying microbial species.

The methanogenic microbial consortium is more widely spread in nature than previously thought. However, due to the slow growth rate of several members of the consortium, and the even slower process of stabilization of the equilibrium between species in the consortium, the use of a large inoculum shortens the start-up period. This is especially true for solid feedstocks such as those in landfills. Methanogenic microorganisms do not seem to migrate rapidly over long distances (ca. 1 m). Appropriate and well mixed inocula may well prove to be essential for enhanced landfill gas generation.

When methanogenic microorganisms with adhesion properties are required for 2-d or 3-d generation methane reactors, it has proved more successful to use as inoculum the active flocculent sludge from aerobic wastewater treatment, as harvested at the bottom of the secondary decanter, rather than sludge obtained from the anaerobic stabilization reactor.

The use of specialized and acclimated inocula for given toxic or recalcitrant feedstocks can help reduce the start-up and acclimation periods. Enzymatic preparations have been claimed to be able to enhance biomethanation; however, no full proof for this exists. In the case of septic tanks, the enhancement was found to be due to inorganic contaminants which acted as inert carriers for microbial adhesion.

7. Environmental Impact of Biomethanation

Methanogenesis in nature and biomethanation by humans have a double impact on the environment. On the one hand, waste treatment has a positive impact, but biogas may have a negative impact.

Biomethanation is an environmental technology that can be applied at various scales from the family-size in rural tropical areas through farm-size in developed countries to industrial scale. In some cases, this environmental biotechnology can be made cost effective and give rise to an economic market.

As only the carbon of the organic matter is transformed to methane, all other elements remain in the digested effluent. The digested effluent of biomethanation is stabilized; that is, it undergoes further transformation only very slowly. Hence, it can be applied to soil without damage to plants.

Because it contains nitrogen, phosphorous and potassium, digested effluent has a fertilizer value. When the effluent is liquid, it can be used for "fertirrigation". When biomethanation proceeds from agricultural waste, it allows recycling of the valuable mineral elements. The digested effluent also contains lignocellulosic material, which is only poorly degraded because lignin is resistant to anaerobic degradation. This fibrous residue has excellent soil conditioning properties.

Biogas is flammable and explosive when appropriately mixed with air (see Fig. 1) and put in contact with an ignition source. Biogas generated in landfills migrates by diffusion and along pressure gradients. Biogas migrating in the root zones of plants creates asphyxia which can result ultimately in the death of vegetation. Biogas migrating in confined areas, such as the basements of buildings, creates explosion hazards. Several cases have been reported, some with fatal consequences for people. Nevertheless, landfill gas can easily be monitored and its migration is easily controlled. Furthermore, the gas can be profitably extracted, so that in the long term the concept of landfilling is not endangered.

It has been estimated that 1.2×10^9 t of methane are emitted annually into the atmosphere [24], resulting in a concentration of 1.4 ppm in air. Ten percent of this is fossil methane; 10 % originates from animal wastes and human sewage; 5 % from the anaerobic fermentation of human domestic wastes; and 5 % from biomass burning, notably from forest fires. Natural ecosystems, including notably termites, which account for 10 – 20 %, contribute to the remaining methane in the atmosphere. Methane has been calculated to have a mean residence time in the atmosphere of 3.3 years. Atmospheric methane concentration has doubled within a few centuries. Another doubling is expected 2050s. The atmospheric warming associated with the increase in methane concentration in the atmosphere during the last decade has been calculated to amount to about 38 % of the effect of carbon dioxide [25].

8. Storage, Pretreatment, Upgrading and Uses of Biogas

Biogas can be stored in low pressure gasholders, in medium-pressure tanks (1 – 100 m^3; 1 – 2 MPa), in high-pressure gas cylinders (30 – 50 L; 20 MPa), or as liquefied gas below the critical temperature of methane (− 82.5 °C). Low-pressure gasholders can be divided into wet and dry gasholders. Wet gasholders consist of floating rigid bells, usually of steel, with a capacity of 50 – 5000 m^3. Dry gasholders are made of rubber or plastics and store from 1 – 1000 m^3 of biogas.

Biogas must usually be pretreated before enduse. In any case, its 5 % water vapor content must be condensed. Condensed water is a major source of corrosion. Hydrogen sulfide whenever present in too large amounts must be removed. Chemical absorption of hydrogen sulfide onto iron oxides is one of the possible treatments. Whenever necessary, trace elements are removed by adsorption on active charcoal.

Biogas sometimes needs to be upgraded to methane, to be used as a substitute or a complement of natural gas, as town gas, or as a motor fuel. Carbon dioxide is removed by absorption in water or by molecular sieves. The pressure swing method, consisting of absorption under high pressure and release under low pressure periods, is often used. Selective membranes for the separation of methane and carbon dioxide are also widely used.

Biogas can be used to generate light using glowing mantles, a common practice in rural developing areas, and for heating and cooking; atmospheric and forced draft burners need to be adapted.

Biogas does not belong to any of the three families of combustible gases [8] and is thus not interchangeable with any gas belonging to these three families. The gas jet of burners must be enlarged and the pressure of the gas supply lowered. The air supply must be slightly increased. In some cases, modifications of the burner design is necessary.

Biogas can be used for mechanical power generation. Spark ignition, internal combustion, Otto engines require the addition of a gas carburettor to provide a stoichiometric air – biogas mixture to the combustion chambers. Compression ignition, internal combustion, diesel engines usually work with biogas on the dual-fuel engine basis. The latter engines use a homogeneous biogas – air mixture compressed below its autoignition conditions and ignited by the injection of pilot diesel fuel. Engine modifications concern air and biogas supply and diesel injection. The properties of biogas as a motor fuel are described in further detail in → Automotive Fuels.

Electrical power can be generated using a stationary biogas engine connected to a generator. Combined heat and power generation is very interesting in the case of biogas. The recovery of waste heat, as occurs for example in total energy modules (TOTEM), can provide the process heat necessary for controlling the temperature of the digester.

Whenever high rates of methane production are available, for example from landfills, gas turbines can be used.

9. Analysis

Biogas volumetric production rates are easily measured using conventional domestic gas flowmeters. Biogas composition is best measured by gas chromatography using Porapak Q as an inert support in a first column, a molecular sieve in a second column, and a catharometer for detection. Gases appear in the order hydrogen, oxygen, nitrogen, carbon dioxide, and methane. The ratio of carbon dioxide to methane in biogas can be estimated approximatively with a very simple device. A known volume of biogas is pumped into a small portable vessel, and the pressure is adjusted to an arbitrary 100 % value. Carbon dioxide is then absorbed in sodium hydroxide solution; the ratio is calculated from the pressure drop.

The COD or the content of organic dry matter are the most common parameters for assessing feedstock and effluent. Individual short-chain carboxylic acids are best determined by gas chromatography with Chromosorb W 100 as inert support, acidified Carbowax as stationary phase and a flame ionisation detector. The total concentration of acids can be determined by steam distillation followed by titration. Alkalinity is another important process parameter; it is usually determined by acid titration, removal of

carbon dioxide and back titration [26]. Low partial pressures of hydrogen (1 – 1000 ppm) can easily be determined by polarography.

10. Biogas in Developing Countries

[27], [28]

In rural areas of developing countries, biogas from domestic and agricultural wastes is a potential substitute for fuel wood. Most designs are based on the Chinese and Indian designs. A hybrid family digester designed by the German Technical Cooperation Agency (Gesellschaft für technische Zusammenarbeit, GTZ) is depicted in Figure 6. The optimum volume for a family digester is ca. 8 – 10 m^3. Between 10 and 15 kg of dry organic waste will produce about 3 m^3 of biogas containing 55 % of methane. This daily feedstock requirement corresponds to 60 – 80 kg fresh dung originating from six to eight head of cattle, half time in stabulation. The daily produced biogas suffices for 7.5 h use of a cooking burner, or alternatively 6 h of a cooking burner and 3 h of a biogas-fueled glowing-mantle lamp, or 3 h of a cooking burner, 3 h of lighting and 24 h of a 80 L biogas-fueled refrigerator, or 2 h of a 1 kW engine [29]. The replacement of fuel wood by biogas in developing countries will however require one or more generations.

11. Economic Aspects

Environmental biotechnology, for example wastewater treatment, is usually considered as a necessary cost, without any profit. Biomethanation should not follow this way of thinking.

A thousand methane digesters have been constructed in Europe for an average investment per cubic meter of working volume ranging from ca. € 300 for biogas plants on farms to ca. € 410 for industrial biogas plants. Methane digesters produce biogas which has an energy value and can hence be used as a substitute for conventional energy and can be valued for the substituted energy. In Denmark, the public authorities add a financial incentive equivalent to the energy taxes on top of that value. As an average, on the basis of the present energy prices, the value of biogas covers 20 % of the financial costs of biomethanation. The digested effluent, mainly when it is solid material has a potential value equivalent to that of compost, that is, ca. 40 € per ton. However, when the feedstock is municipal solid waste, marketing of the compost-like end-product is still proving difficult. The digested effluent can also be priced at the value of substituted energy. Preparing one ton of compost requires 6200 MJ. A digested effluent containing nitrogen, phosphorous and potassium can be valued for these elements. In a fertilizer, nitrogen values 18.5 MJ, potassium, 3.35 MJ, and phosphorous, 2.31 MJ.

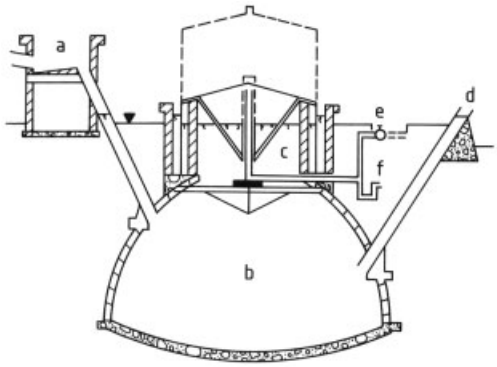

Figure 6. German – Chinese methane digester for a rural tropical family
a) Feedstock inlet; b) Digestion chamber; c) Floating dome gasholder; d) Overflow outlet; e) Gas outlet; f) Water trap

Whenever the feedstock is a waste, it has a cost to be paid by its producer as municipal taxes for domestic solid waste disposal or as pollution taxes for wastewater treatment. Pollution-tax abbatments for agro-industries that treat their wastewaters by biomethanation contribute 80 % of the financial costs of an anaerobic treatment plant. When the biogas produced can be used in the plant (e.g., to generate process heat), this waste treatment technology ends with a no-profit, no-cost balance in these industries. In developing countries, it now appears possible to build family-size digesters for a price acceptable to the buyer, for example the price of a head of cattle in Burundi or the price of a bicycle in China, with a minimum federal financial incentive, corresponding to the macro-economic savings (deforestation, hard currency).

12. References

1. E.-J. Nyns, *Trib. CEBEDEAU* **34** (1981) 351 – 356.
2. B. Lagrange: *Biomethane* vol. 1. *Une Alternative Crédible;* vol. 2. *Principe – Techniques – Utilisation,* Edisud, Aix-en-Provence 1979.
3. D. A. Stafford, D. L. Hawkes, R. Horton: *Methane Production from Waste Organic Matter,* CRC Press, Boca Raton 1980.
4. E. C. Price, P. N. Cheremisinoff: *Biogas. Production and Utilization,* Ann Arbor Sci., Ann Arbor 1981.
5. R. Braun: *Biogas. Methangärung Organischer Abfallstoffe,* Springer Verlag, Wien 1982.
6. D. L. Wise: *Fuel Gas Systems,* CRC Press, Boca Raton 1983.
7. A. Pauss, H. Naveau, E.-J. Nyns in D. O. Hall, R. P. Overend (eds.): *Biomass: Regenerable Energy,* Chichester 1987, pp. 273 – 291.
8. M. Constant, H. Naveau, G.-L. Ferrero, E.-J. Nyns: *Biogas End-Use in the European Community,* Elsevier Applied Science, London 1989.
9. W. E. Balch et al., *Microbiol. Rev.* **43** (1979) 260 – 296.
10. P. L. McCarty in D. E. Hughes et al. (eds.): *Anaerobic Digestion,* Elsevier, Amsterdam 1981, pp. 3 – 22.
11. E.-J. Nyns in H. J. Rehm, G. Reed (eds.): *Biotechnology,* VCH, Weinheim 1986, pp. 207 – 267.
12. R. Buvet in M. P. Ferranti, G.-L. Ferrero, P. L'Hermite (eds.): *Anaerobic Digestion. Results of Research and Demonstration Projects,* Elsevier, London 1986, pp. 210 – 217.
13. M. M. Schumacker: *Landfill Methane Recovery,* Noyes Data Corp., Park Ridge 1983.
14. G. Lettinga et al., *Biotechnol. Bioeng.* **22** (1980) 699 – 734.
15. F. G. Pohland, S. Gosh, *Environ. Lett.* **1** (1971) 255 – 266.
16. D. L. Wise: *Fuel Gas Production from Biomass,* **2 vols.,** CRC Press, Boca Raton 1981.
17. K. J. Thomé-Kozmiensky: *Biogas. Anaerobtechnik in der Abfallwirtschaft,* EF Verlag für Energie und Umwelttechnik, Berlin 1989.
18. E.-J. Nyns in O. Kitani, C. W. Hall (eds.): *Biomass Handbook,* Gordon and Breach, New York 1989.
19. M. Demuynck, E.-J. Nyns, W. Palz: *Biogas Plants in Europe,* Reidel, Dordrecht 1984.
20. P. N. Hobson, S. Bousfield, R. Summer: *Methane Production from Agricultural and Domestic Wastes,* Applied Science, London 1981.
21. R. Kloss: *Planung von Biogasanlagen,* Oldenbourg, München 1986.
22. P.-J. Meynell: *Methane. Planning a Digester,* Prism Press, Dorchester 1982.
23. R. E. Speece, *Adv. Solar Energy* **2** (1985) 51 – 123.
24. J. C. Sheppard et al., *J. Geophys. Res.* **87** (1982) 1305 – 1312.
25. H. Craig, C. C. Chou, *J. Geophys. Res.* **99** (1982) 1221 – 1224.
26. R. Buvet, E.-J. Nyns in R. Buvet, M. F. Fox, D. J. Picken (eds.): *Biomethane, Production and Uses,* Bowskell, Exeter 1984, pp. 109 – 119.
27. D. House: *Biogas Handbook,* Peace Press, Culver City 1948.
28. M. M. El-Halwagi (ed.): *Biogas Technology, Transfer and Diffusion,* Elsevier Appl. Sci. London 1984.
29. E.-J. Nyns, *Enzyme Microbial Technol.* **12** (1990) 151 –152.

Methanol

Eckhard Fiedler, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
Georg Grossmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
D. Burkhard Kersebohm, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
Günther Weiss, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
Claus Witte, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction . . . 1
2. Physical Properties . . . 2
3. Chemical Properties . . . 3
4. Production . . . 3
4.1. Principles . . . 3
4.1.1. Thermodynamics . . . 3
4.1.2. Kinetics and Mechanism . . . 5
4.1.3. Byproducts . . . 6
4.2. Catalysts . . . 6
4.2.1. Catalysts for High-Pressure Synthesis 6
4.2.2. Catalysts for Low-Pressure Synthesis 6
4.2.3. Production of Low-Pressure Catalysts 7
4.2.4. Catalyst Deactivation . . . 7
4.2.5. Other Catalyst Systems . . . 8
5. Process Technology . . . 9
5.1. Production of Synthesis Gas . . . 9
5.2. Synthesis . . . 10
5.3. Reactor Design . . . 11
5.4. Distillation of Crude Methanol . . . 13
5.5. Construction Materials . . . 13
6. Handling, Storage, and Transportation . . . 13
6.1. Explosion and Fire Control . . . 13
6.2. Storage and Transportation . . . 14
7. Quality Specifications and Analysis 15
8. Environmental Protection . . . 16
9. Uses . . . 16
9.1. Use as Feedstock for Chemical Syntheses . . . 16
9.2. Use as Energy Source . . . 17
9.3. Other Uses . . . 19
10. Economic Aspects . . . 20
11. Toxicology and Occupational Health 20
11.1. Toxicology . . . 20
11.2. Occupational Health . . . 22
12. References . . . 22

1. Introduction

Methanol [*67-56-1*], CH_3OH, M_r 32.042, also termed methyl alcohol or carbinol, is one of the most important chemical raw materials. Worldwide production capacity in 1989 was ca. 21×10^6 t/a. About 85 % of the methanol produced is used in the chemical industry as a starting material or solvent for synthesis. The remainder is used in the fuel and energy sector; this use is increasing. In 1993 world wide production capacity was 22.4×10^6t/a.

Historical Aspects. Methanol was first obtained in 1661 by Sir Robert Boyle through the rectification of crude wood vinegar over milk of lime. He named the new compound *adiaphorus spiritus lignorum*. Justus von Liebig (1803 – 1873) and J. B. A. Dumas (1800 – 1884) independently determined the composition of methanol. The term "methyl" was introduced into chemistry in 1835 on the basis of their work.

From ca. 1830 – 1923, "wood alcohol", obtained by the dry distillation of wood, remained the only important source of methanol. As early as 1913, A. Mittasch and coworkers at BASF successfully produced organic compounds containing oxygen, including methanol, from carbon monoxide and hydrogen in the presence of iron oxide catalysts during developmental work on the synthesis of ammonia. The decisive step in the large-scale industrial production of methanol was made by M. Pier and coworkers in the early 1920s with the development of a sulfur-resistant zinc oxide – chromium oxide catalyst. By the end of 1923 the process had been converted from the developmental to the production stage at the BASF Leuna Works.

Processes based on the above work were performed at high pressure (25 – 35 MPa) and 320 – 450 °C. They dictated the industrial pro-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_465

duction of methanol for more than 40 years. In the 1960s, however, ICI developed a route for methanol synthesis in which sulfur-free synthesis gas containing a high proportion of carbon dioxide was reacted on highly selective copper oxide catalysts. This and other related low-pressure processes are characterized by fairly mild reaction conditions (5 – 10 MPa, 200 – 300 °C). Methanol can now be produced much more economically worldwide by these low-pressure methods.

2. Physical Properties

Methanol is a colorless, neutral, polar liquid that is miscible with water, alcohols, esters, and most other organic solvents [1], [2]; it is only slightly soluble in fat and oil. Because of its polarity, methanol dissolves many inorganic substances, particularly salts.

The most important physical data for methanol follow [3], [4]:

Density (101.3 kPa), liquid	
at 0 °C	0.8100 g/cm^3
at 25 °C	0.78664 g/cm^3
at 50 °C	0.7637 g/cm^3
Critical pressure	8.097 MPa
Critical temperature	239.49 °C
Critical density	0.2715 g/cm^3
Critical volume	117.9 cm^3/mol
Critical compressibility	0.224
mp	− 97.68 °C
Heat of fusion (101.3 kPa)	100.3 kJ/kg
Triple-point temperature	− 97.56 °C
Triple-point pressure	0.10768 Pa
bp (101.3 kPa)	64.70 °C
Heat of vaporization (101.3 kPa)	1128.8 kJ/kg
Standard enthalpy of formation	
at 25 °C (101.3 kPa), gas	− 200.94 kJ/mol
at 25 °C (101.3 kPa), liquid	− 238.91 kJ/mol
Free enthalpy of formation	
at 25 °C (101.3 kPa), gas	− 162.24 kJ/mol
at 25 °C (101.3 kPa), liquid	− 166.64 kJ/mol
Standard entropy	
at 25 °C (101.3 kPa), gas	239.88 J mol^{-1} K^{-1}
at 25 °C (101.3 kPa), liquid	127.27 J mol^{-1} K^{-1}
Specific heat, c_P	
at 25 °C (101.3 kPa), gas	44.06 J mol^{-1} K^{-1}
at 25 °C (101.3 kPa), liquid	81.08 J mol^{-1} K^{-1}
Viscosity (25 °C)	
Liquid	0.5513 mPa · s
Vapor	9.68 × 10^{-3} mPa · s
Thermal conductivity (25 °C)	
Liquid	190.16 mW m^{-1} K^{-1}
Vapor	14.07 mW m^{-1} K^{-1}
Electrical conductivity (25 °C)	(2 – 7) × 10^{-9} Ω^{-1} cm^{-1}
Dielectric constant (25 °C)	32.65
Dipole moment	5.6706 × 10^{-30} C · m
Refractive index n_D^{20}	1.32840
n_D^{25}	1.32652
Surface tension in air (25 °C)	22.10 mN/m
Flash point (DIN 51 755)	6.5 °C
Open vessel	15.6 °C
Closed vessel	12.2 °C
Explosion limits in air	5.5 – 44 vol %
Ignition temperature (DIN 51 794)	470 °C

The temperature dependence of selected physical properties is given in [5]; thermodynamic data can be found in [6] and the heat capacity and enthalpy of the liquid in [7].

The vapor pressure of methanol is determined according to [8] by a Wagner equation of the form

$$\ln p = 8.999 + \frac{512.64}{T}\left(-8.63571q + 1.17982q^{3/2} - 2.4790q^{5/2} - 1.024q^5\right)$$

where $q = 1 - T/512.64$; T is the absolute temperature, and p the pressure in kilopascals. Further vapor pressure correlation data in the temperature range 206 – 512 K are given in [9], and critical data in [10]. A selection of binary azeotropes is shown in Table 1, and a comprehensive summary is given in [11].

Viscosity data of the pure components have been published in [5], [12], [13] for the liquid phase, and in [13] for the vapor. The viscosity and density of aqueous methanol solutions at 25 °C are shown in Table 2. Temperature-dependent densities of the binary mixture are given in [15] and [16]; viscosities are documented in [15] and [17]. The pressure dependence of viscosity has been measured [18], and isothermal compressibilities, coefficients of thermal expansion, partial molar volumes, and excess factors accounting for the difference between real and ideal behavior can be found in [19]. Information on the liquid – solid phase equilibrium in the methanol – water system is given in [20].

Data on the thermal conductivity of liquid methanol appear in [21]; the electrical conductivity of the pure liquid and dielectric properties are given in [22] and [23], respectively. Safety aspects have also been discussed [24], [25].

Table 1. Binary azeotropic mixtures of methanol [11]

Component	*bp* of component, °C	*bp* of azeotrope, °C	Methanol content of azeotrope, wt %
Acetonitrile	81.6	63.45	19
Acrylonitrile	77.3	61.4	61.3
Acetone	56.15	55.5	12
Ethyl formate	54.15	50.95	16
Methyl acetate	57.1	53.9	17.7
Furan	31.7	< 30.5	< 7
Thiophene	84	< 59.55	< 55
Methyl acrylate	80	62.5	54
2-Butanone	79.6	64.5	70
Tetrahydrofuran	66	60.7	31.0
Ethyl acetate	77.1	62.25	44
Methyl propionate	79.8	62.45	47.5
Methyl methacrylate	99.5	64.2	82
Cyclopentane	49.4	38.8	14
n-Pentane	36.15	30.85	7
Benzene	80.1	57.50	39.1
Cyclohexane	80	54	38
Cyclohexene	82.75	55.9	40
Toluene	110.6	63.5	72.5

Table 2. Viscosity and density of aqueous methanol solutions at 25 °C [14]

Mole fraction of methanol	Kinematic viscosity, mm^2/s	Density, kg/m^3	Absolute viscosity, mPa · s
0.0	0.893	997.1	0.890
0.0507	1.126	983.4	1.107
0.1125	1.385	966.9	1.339
0.1411	1.480	960.2	1.421
0.2276	1.657	941.1	1.559
0.2927	1.683	925.7	1.558
0.4198	1.593	898.4	1.431
0.4856	1.505	884.5	1.331
0.5542	1.396	869.9	1.214
0.7133	1.149	837.7	0.963
0.8040	0.992	821.0	0.814
0.8345	0.952	816.0	0.777
0.9140	0.825	800.1	0.660

3. Chemical Properties

Methanol is the simplest aliphatic alcohol. As a typical representative of this class of substances, its reactivity is determined by the functional hydroxyl group [26–28]. Reactions of methanol take place via cleavage of the C−O or O−H bond and are characterized by substitution of the −H or −OH group (→ Alcohols, Aliphatic, Chap. 2.2.) [29]. In contrast to higher aliphatic alcohols, however, β-elimination with the formation of a multiple bond cannot occur.

Important industrial reactions of methanol include the following (Fig. 1):

1) Dehydrogenation and oxidative dehydrogenation
2) Carbonylation
3) Esterification with organic or inorganic acids and acid derivatives
4) Etherification
5) Addition to unsaturated bonds
6) Replacement of hydroxyl groups

4. Production

4.1. Principles

4.1.1. Thermodynamics

The formation of methanol from synthesis gas can be described by the following equilibrium reactions:

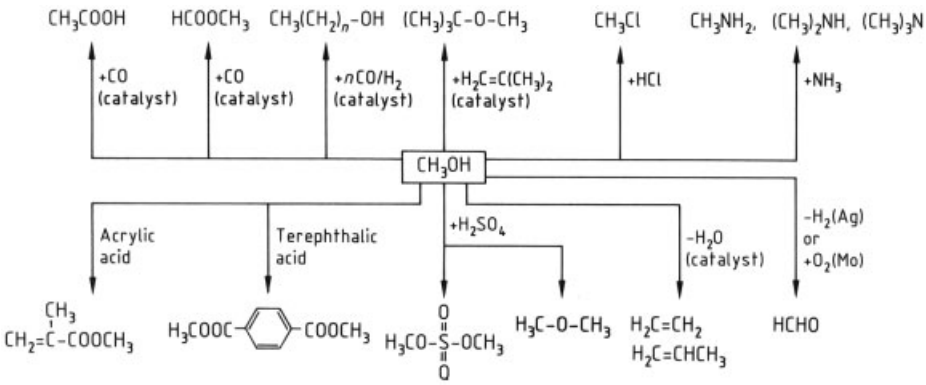

Figure 1. Industrially important reactions of methanol

$$CO + 2\,H_2 \rightleftharpoons CH_3OH \quad \Delta H_{300\,K} = -90.77\,\text{kJ/mol} \tag{1}$$

$$CO_2 + 3\,H_2 \rightleftharpoons CH_3OH + H_2O \quad \Delta H_{300\,K} = -49.16\,\text{kJ/mol} \tag{2}$$

Reaction enthalpies are determined from the standard enthalpies of the reactants and products [30]. Both reactions are exothermic and accompanied by a decrease in volume. Methanol formation is thus favored by increasing pressure and decreasing temperature, the maximum conversion being determined by the equilibrium composition.

In addition to the two methanol-forming reactions, the endothermic reaction of carbon dioxide and hydrogen (Eq. 3, the reverse water-gas shift reaction) must also be taken into account:

$$CO_2 + H_2 \rightleftharpoons CO + H_2O \quad \Delta H_{300\,K} = 41.21\,\text{kJ/mol} \tag{3}$$

For the sake of simplicity, Equations (1) and (3) can be discussed as independent reaction pathways. The conversion of carbon dioxide to methanol (2) is then the overall result of Equations (1) and (3), and the equilibrium constant K_2 can be described as $K_2 = K_1 \cdot K_3$. When the nonideal behavior of gases is taken into account, the equilibrium constants are determined as follows:

$$K_1 = \left[\frac{f_{CH_3OH}}{f_{CO}\, f_{H_2}^2}\right] = \left[\frac{\varphi_{CH_3OH}}{\varphi_{CO}\, \varphi_{H_2}^2}\right]\left[\frac{p_{CH_3OH}}{p_{CO}\, p_{H_2}^2}\right] = K_{\varphi 1} \cdot K_{p1} \tag{4}$$

$$K_3 = \left[\frac{f_{CO}\, f_{H_2O}}{f_{CO_2}\, f_{H_2}}\right] = \left[\frac{\varphi_{CO}\, \varphi_{H_2O}}{\varphi_{CO_2}\, \varphi_{H_2}}\right]\left[\frac{p_{CO}\, p_{H_2O}}{p_{CO_2}\, p_{H_2}}\right] = K_{\varphi 3} \cdot K_{p3} \tag{5}$$

where f_i is the fugacity, φ_i the fugacity coefficient, and p_i the partial pressure of the i-th component.

A number of numerical formulations exist for calculating the temperature-dependent equilibrium constants K_1 [31–38] and K_3 [36–39]; their results differ widely [40]. The binomial formulations of Cherednichenko (Eq. 6) [34] and Bisset (Eq. 7) [39] are examples:

$$K_1 = 9.740 \times 10^{-5} \exp\left[21.225 + \frac{9143.6}{T} - 7.492 \ln T + 4.076 \times 10^{-3} T - 7.161 \times 10^{-8} T^2\right] \tag{6}$$

$$K_3 = \exp\left[13.148 - \frac{5639.5}{T} - 1.077 \ln T - 5.44 \times 10^{-4} T + 1.125 \times 10^{-7} T^2 + \frac{49170}{T^2}\right] \tag{7}$$

The fugacity coefficients can be determined according to [41] by assuming ideal solubility for the individual pure components, or they can be calculated from suitable equations of state [42], [43].

The carbon monoxide and carbon dioxide conversions up to attainment of equilibrium are shown as a function of pressure and temperature in Table 3 [35]. A synthesis gas formed by

Table 3. Temperature and pressure dependence of the carbon monoxide and carbon dioxide equilibrium conversions [35]

Temperature,	CO conversion *			CO_2 conversion		
°C	5 MPa	10 MPa	30 MPa	5 MPa	10 MPa	30 MPa
200	96.3	99.0	99.9	28.6	83.0	99.5
250	73.0	90.6	99.0	14.4	45.1	92.4
300	25.4	60.7	92.8	14.1	22.3	71.0
350	− 2.3	16.7	71.91	9.8	23.1	50.0
400	− 12.8	− 7.3	34.1	27.7	29.3	41.0

* Negative sign denotes CO formation via Equation (3) [44]: $CO_2 + H_2 \rightleftharpoons CO + H_2O$.

steam reforming was chosen as the starting gas (15 vol % CO, 8 vol % CO_2, 74 vol % H_2, and 3 vol % CH_4). Equations (6) and (7) were used to establish temperature dependence, and the fugacity coefficients were determined according to the Soave – Redlich – Kwong equation. The negative sign for the carbon monoxide conversion denotes carbon monoxide formation by back-conversion [44].

4.1.2. Kinetics and Mechanism

The formation of methanol, as a typical heterogeneously catalyzed reaction, can be described by an absorption – desorption mechanism (Langmuir – Hinshelwood or Eley – Rideal). The nature of the active centers in the copper – zinc oxide – alumina catalysts used under industrial conditions is still a subject of discussion (see Chap. 5). The active species in low-pressure methanol synthesis may be a solution of copper(I) ions in the zinc oxide phase [45]. On the other hand, evidence can be found that copper(0) also catalyzes methanol formation. The feed gas composition (particularly the proportions of CO_2 and H_2O) also plays an important role in determining the activity and selectivity of catalysts in methanol production. Investigations have shown that various routes must exist for the formation of methanol via carbon monoxide or carbon dioxide, and that different active centers in the catalyst are involved [46–50].

According to [51], alumina exists in an X-ray amorphous form. The proposed functions of alumina in copper – zinc oxide – alumina catalysts include:

1) prevention of sintering of the fine copper particles by the formation of zinc spinel;
2) stabilization of the highly disperse copper – zinc oxide catalyst system; and
3) formation of surface defects by the incorporation of alumina clusters in the copper lattice [46], [51].

Which effect prevails in methanol synthesis is still not clear. However, alumina has an important function as a structural promoter in copper – zinc oxide catalysts by improving their mechanical stability and long-term activity.

Some kinetic investigations have concentrated on the role of carbon dioxide in methanol synthesis, which aroused a great deal of controversy during the 1980s [40], [46], [52–54]. Until the beginning of the 1980s, mechanistic considerations were based almost exclusively on the hydrogenation of carbon monoxide to methanol (Eq. 1, see 4.1.1) [55–59]. The increased yield achieved by adding carbon dioxide was ascribed to the displacement of the reverse water-gas shift equilibrium (Eq. 3). In addition, carbon dioxide was assumed to influence the oxidation state of the active centers in the catalyst [45].

In contrast, Kagan et al. [60] proposed that methanol was formed solely according to Equation (2) from carbon dioxide. Recent experiments with isotope-labeled reactants show that both reaction pathways (Eqs. 1 and 2) are possible [61], [62]. Similar results were obtained in other studies [63–65]. However, according to [62], formation via carbon dioxide predominates under conditions of large-scale industrial methanol synthesis.

4.1.3. Byproducts

Commercially available $Cu-ZnO-Al_2O_3$ catalysts for the low-pressure synthesis of methanol permit production of the desired product with high selectivity, typically above 99 % referred to the added CO_x.

The following impurities are important for the large-scale industrial process:

1) Higher alcohols formed by catalysis with traces of alkali [66–68]

$$n\,CO + 2n\,H_2 \rightleftharpoons C_nH_{2n+1}OH + (n-1)\,H_2O$$

2) Hydrocarbons and waxes formed by catalysis with traces of iron, cobalt, and nickel according to the Fischer–Tropsch process [67, 69, 70]

$$CO + 3\,H_2 \rightleftharpoons CH_4 + H_2O$$

$$CO_2 + 4\,H_2 \rightleftharpoons CH_4 + 2\,H_2O$$

$$n\,CO + (2n-1)\,H_2 \rightleftharpoons C_nH_{2n+2} + n\,H_2O$$

3) Esters [68], [70], [71]

$$(CH_2O)_{ads} + (RCHO)_{ads} \rightleftharpoons CH_3COOR$$

4) Dimethyl ether [70], [72]

$$2\,CO + 4\,H_2 \overset{Al_2O_3}{\rightleftharpoons} CH_3-O-CH_3 + H_2O$$

5) Ketones [73]

$$RCH_2CH_2OH \rightleftharpoons RCH_2CHO + H_2$$

$$2\,RCH_2CHO \rightleftharpoons RCH_2COCHRCH_3 + O_{ads}$$

The formation of most byproducts from synthesis gas, particularly C_2^+ species, is thermodynamically favored over methanol synthesis. Because methanol constitutes the main product, however, reactions yielding impurities are controlled kinetically rather than thermodynamically [40]. In addition to catalyst constituents and feed gas composition, the residence time at the catalyst [68], as well as the temperature [69], [70], mainly determine the extent of byproduct formation: an increase in these parameters raises the proportion of byproducts. A detailed discussion of individual byproduct classes is given in [40].

4.2. Catalysts

4.2.1. Catalysts for High-Pressure Synthesis

The first industrial production of methanol from synthesis gas by the high-pressure process employed a catalyst system consisting of zinc oxide and chromium oxide. This catalyst, which was used at 25–35 MPa and 300–450 °C, was highly stable to the sulfur and chlorine compounds present in synthesis gas [44], [54], [74], [75].

Production of methanol with zinc oxide–chromium oxide catalysts by the high-pressure process is no longer economical. A new generation of copper-containing catalysts with higher activity and better selectivity is now used. The last methanol plant based on the high-pressure process closed in the mid-1980s. For a detailed discussion of high-pressure methanol catalysts, see [74].

4.2.2. Catalysts for Low-Pressure Synthesis

Well before the industrial realization of low-pressure methanol synthesis by ICI in the 1960s, copper-containing catalysts were known to be substantially more active and selective than zinc oxide–chromium oxide catalysts. Copper oxide–zinc oxide catalysts and their use in the production of methanol were described by BASF in the early 1920s [76], [77]. These catalysts were employed at 15 MPa and 300 °C.

Their industrial use was prevented, however, by a serious disadvantage: impurities such as hydrogen sulfide and chlorine compounds in synthesis gas rapidly deactivated the catalysts. Nevertheless, the copper-containing catalyst systems proved to be the most promising candidates for producing methanol industrially at lower temperature and pressure. A series of publications on this topic appeared between 1925 and 1955 [74], [78], [79]. Investigations of copper catalysts continue to this day [53].

A low-pressure catalyst for methanol synthesis was first used industrially in the process developed by ICI in 1966. This copper oxide–zinc oxide catalyst was thermally stabilized with alumina. It was used to convert extremely pure (i.e., largely free of sulfur and chlorine compounds, $H_2S < 0.1$ ppm) synthesis gas to methanol [80]. Because this copper catalyst was extremely active, methanol synthesis could be carried out at

220 – 230 °C and 5 MPa. Premature aging due to sintering of copper was thereby avoided. The high selectivity of the new catalyst gave a methanol purity > 99.5 %. The formation of byproducts (e.g., dimethyl ether, higher alcohols, carbonyl compounds, and methane) associated with the old high-pressure catalyst, was drastically reduced or, in the case of methane, completely eliminated.

All currently used low-pressure catalysts contain copper oxide and zinc oxide with one or more stabilizing additives (Table 4). Alumina, chromium oxide, or mixed oxides of zinc and aluminum have proved suitable for this purpose [81], [82].

4.2.3. Production of Low-Pressure Catalysts

Catalysts now used in low-pressure methanol synthesis plants and based on copper – zinc – aluminum (or chromium) are obtained as metal hydroxycarbonates or nitrates by coprecipitation of aqueous metal salt solutions (e.g., nitrates) with sodium carbonate solution. Precipitation may occur in one or several stages. The quality of the subsequent catalyst is determined by the optimum composition of the metal components, the precipitation temperature, the pH used for precipitation, the sequence of metal salt additions, and the duration of precipitation. The stirring rate, stirring energy, and shape of stirrer also affect catalyst quality.

The precipitated catalyst precursors (largely metal hydroxycarbonates) are filtered off from the mother liquor, washed free of interfering ions (e.g., sodium), and dried at ca. 120 °C. Examples of such hydroxycarbonates are malachite rosasite $(Cu,Zn)_5(CO_3)(OH)_2$, hydrozincite $(Cu,Zn)_5(OH)_6(CO_3)_2$, and aurichalcite $(Cu_{0.3}Zn_{0.7})_5(OH)_6(CO_3)_2$ [40], [47], [82]. Aurichalcite derivatives with the composition $Cu_{2.2}Zn_{2.8}(OH)_6(CO_3)_2$, containing small amounts of alumina for stabilization, are obtained by coprecipitation of metal nitrates [87], [88]. The catalyst precursor is converted to finely divided metal oxide by subsequent calcination at ca. 300 – 500 °C [80]. The calcined product is then pelleted to commercial catalyst forms. Cylindrical tablets 4 – 6 mm in diameter and height are common [46], [47], [82], [89].

The catalysts still have a total BET surface area of 60 – 100 m^2/g and have to be activated [47]. They are activated by controlled reduction with 0.5 – 2 % hydrogen in nitrogen at 150 –230 °C. Particular care must be taken to avoid hot spots, which lead to premature catalyst aging. In their reduced (i.e., active) form, the synthesis-active copper surfaces of commercial catalysts have a surface area of 20 – 30 m^2/g [81].

Catalysts for the low-pressure synthesis of methanol can also be produced by other methods, e.g., impregnating a carrier with active components, kneading metal compounds together, and leaching Raney alloys [17], [82].

Catalysts must be devoid of interfering impurities. Alkali compounds reduce the useful life and adversely affect the selectivity of catalysts. Even iron or nickel impurities in the parts-per-million range promote the formation of hydrocarbons and waxes. Acidic compounds such as silicon dioxide increase the proportion of dimethyl ether in crude methanol [90].

4.2.4. Catalyst Deactivation

As mentioned in Section 4.1, efficient catalysts for low-pressure synthesis of methanol should have a highly disperse distribution of active centers stabilized by structural promoters. The longer a catalyst can retain these properties under industrial conditions, the more valuable it is for industrial operation: downtimes for catalyst replacement are reduced. Catalysts normally have useful lives of 2 – 5 years. Many factors can drastically reduce catalyst activity and, thus, useful life. Detailed review articles on catalyst deactivation and poisoning can be found in [40], [47], [90].

Even during catalyst production, manufacturing faults can seriously affect the complex structure of the active centers (see Section 4.2.3). Catalyst damage and, consequently, premature deactivation may also occur during reduction. The temperature conditions, hydrogen concentration of the reducing gas, and gas load must be strictly controlled. Deviations from specified reduction procedures may lead to hot spots in the pellets, resulting in sintering of the copper constituents; copper becomes mobile at 190 °C and can agglomerate from its finely divided form into fairly

Table 4. Summary of typical copper-containing catalysts for low-pressure methanol synthesis

Manufacturer	Component	Content, atom%	Reference
IFP	Cu	25 – 80	[83]
	Zn	10 – 50	
	Al	4 – 25	
Süd Chemie	Cu	65 – 75	[84]
	Zn	18 – 23	
	Al	8 – 12	
Shell	Cu	71	[85]
	Zn	24	
	rare-earth oxide	5	
ICI	Cu	61	[86]
	Zn	30	
	Al	9	
BASF	Cu	65 – 75	[87]
	Zn	20 – 30	
	Al	5 – 10	
Du Pont	Cu	50	[88]
	Zn	19	
	Al	31	
United Catalysts	Cu	62	[88]
	Zn	21	
	Al	17	
Haldor Topsoe	Cu	37	[88]
	Zn	15	
	Cr	48	

large crystallites. Reduction must be complete to obtain the entire active mass from the precursor compounds (see Section 4.2.3). Deviations from the specified reduction conditions may permanently decrease the active BET surface area and thus irrevocably damage the catalyst.

Another important point regarding the deactivation of copper catalysts is their high sensitivity to impurities in synthesis gas. Chlorine- and sulfur-containing contaminants long prevented the use of copper-containing catalyst systems in industrial methanol plants. These catalyst poisons must be removed from the feed gas prior to methanol synthesis. A certain degree of protection against deactivation by sulfur is afforded by catalysts containing zinc oxide because the sulfur is bound as zinc sulfide. After deactivation, the catalyst is still able to absorb large quantities of sulfur to protect subsequent catalyst layers against poisoning. Other synthesis gas impurities (e.g., silicon compounds, nickel carbonyls, or iron carbonyls) also cause catalyst damage [90].

The catalyst can also be deactivated by overheating during operation. Thermal damage to the catalyst can occur after use of nonoptimum recycled gas compositions, incorrect temperature control, or overloaded catalyst in the startup phase. The active surface area of the catalyst is decreased and phase transformations occur. The formation of copper spinels as well as malachite rosasite is observed. In effect, this removes active centers for methanol synthesis from the catalyst [40], [45], [91].

4.2.5. Other Catalyst Systems

A number of modified copper – zinc oxide – alumina catalysts have been prepared by doping with boron, manganese, cerium, chromium, vanadium, magnesium, or other elements [92–100]. Other basic types of catalyst systems have also been investigated: Raney copper catalysts, copper alloys with thorium or rare-earth oxides, and supported precious-metal catalysts [101–106]. Only copper alloy catalysts are reported to have a higher activity than conventional copper – zinc oxide – alumina catalysts [107]. Until now, however, exclusively copper-containing zinc oxide – alumina catalysts have been used in industrial methanol plants. These catalysts have high activity, very good selectivity, long-term stability, and favorable production

costs. They are still the most cost effective catalysts.

5. Process Technology

The oldest process for the industrial production of methanol is the dry distillation of wood, but this no longer has practical importance (→ Biomass Chemicals, Chap. 4.9.). Other processes, such as the oxidation of hydrocarbons and production as a byproduct of the Fischer – Tropsch synthesis according to the Synthol process, have no importance today.

Methanol is currently produced on an industrial scale exclusively by catalytic conversion of synthesis gas. Processes are classified according to the pressure used:

1) High-pressure process 25 – 30 MPa
2) Medium-pressure process 10 – 25 MPa
3) Low-pressure process 5 – 10 MPa

The main advantages of the low-pressure process are lower investment and production costs, improved operational reliability, and greater flexibility in the choice of plant size.

Industrial methanol production can be subdivided into three main steps:

1) Production of synthesis gas
2) Synthesis of methanol
3) Processing of crude methanol

5.1. Production of Synthesis Gas

All carbonaceous materials such as coal, coke, natural gas, petroleum, and fractions obtained from petroleum (asphalt, gasoline, gaseous compounds) can be used as starting materials for synthesis gas production. Economy is of primary importance with regard to choice of raw materials. Long-term availability, energy consumption, and environmental aspects must also be considered.

Natural gas is generally used in the large-scale production of synthesis gas for methanol synthesis. In a few processes (e.g., acetylene production), residual gases are formed which have roughly the composition of the synthesis gas required for methanol synthesis.

Synthesis gases are characterized by the stoichiometry number S:

$$S = \frac{[H_2] - [CO_2]}{[CO] + [CO_2]}$$

where the concentrations of relevant components are expressed in volume percent. The stoichiometry number should be at least 2.0 for the synthesis gas mixture. Values above 2.0 indicate an excess of hydrogen, whereas values below 2.0 mean a hydrogen deficiency relative to the stoichiometry of the methanol formation reaction.

Natural Gas. Most methanol produced worldwide is derived from natural gas. Natural gas can be cracked by steam reforming and by partial oxidation (Fig. 2, see also → Ammonia).

In *steam reforming* the feedstock is catalytically cracked in the absence of oxygen with the addition of water and possibly carbon dioxide (→ Gas Production, Chap. 2.→ Gas Production, Chap. 7.1.). The reaction heat required is supplied externally. In *partial oxidation,* cracking takes place without a catalyst (→ Gas Production, Chap. 3.2.). Reaction heat is generated by direct oxidation of part of the feedstock with oxygen. In a combination of the two processes, only part of the natural gas stream is subjected to steam reforming [108]. The remainder passes with the reformed gas to an autothermal reformer where the natural gas is partially oxidized by oxygen.

Only the production of synthesis gas by steam reforming is discussed here in some detail.

The catalysts used in steam reforming are extremely sulfur sensitive; sulfur concentrations < 0.5 ppm quickly poison the catalyst. A gas purification stage therefore precedes the reformer stage. If sulfur occurs primarily in the form of higher boiling compounds (e.g., mercaptans), batchwise adsorption on a regenerable activated charcoal bed is recommended. In the case of hydrogen sulfide, zinc oxide is used as adsorbent to remove sulfur as zinc sulfide at 340 – 370 °C. Hydrogenating desulfurization becomes necessary if organic sulfur compounds (e.g., COS) are present that cannot be removed with charcoal. Hydrogen (e.g., in the form of purge gas from methanol synthesis) is mixed with the gas stream to be desulfurized and passed over a cobalt or nickel – molybdenum catalyst at 290 – 370 °C.

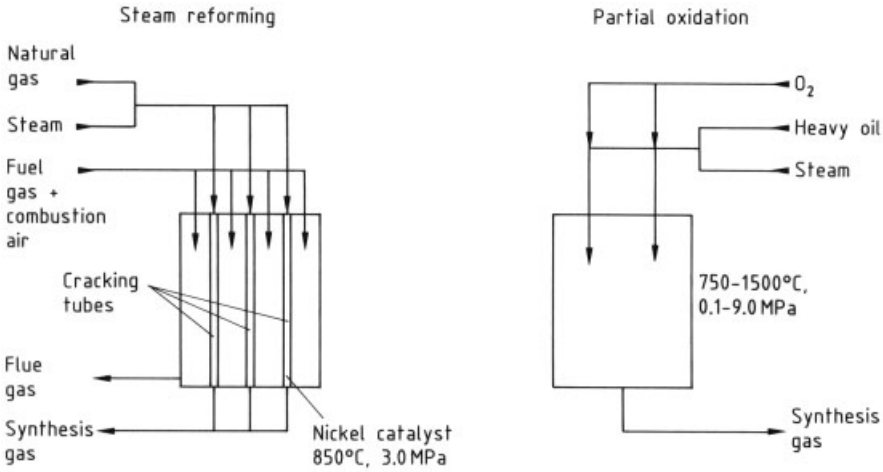

Figure 2. Processes for producing synthesis gases

The sulfur compounds are converted into hydrogen sulfide, which can be removed in a subsequent zinc oxide column.

In the reformer, natural gas is catalytically cracked in the presence of steam:

$$CH_4 + H_2O \rightleftharpoons CO + 3\,H_2 \quad \Delta H_{300\,K} = 206.3\,\text{kJ/mol}$$

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \quad \Delta H_{300\,K} = -41.2\,\text{kJ/mol}$$

The first of these reactions is endothermic and leads to an increase in volume, whereas the second is exothermic and proceeds without change in volume. The degree of conversion of methane increases with increasing temperature, increasing partial pressure of steam, and decreasing absolute pressure.

The interfering Boudouard equilibrium

$$2\,CO \rightleftharpoons CO_2 + C \quad \Delta H_{300\,K} = -172.6\,\text{kJ/mol}$$

which would lead to carbon deposits on the catalyst or on the walls of reformer tubes, can largely be prevented by using excess steam and avoiding long residence times in the critical temperature range above 700 °C.

To reach the stoichiometry necessary for methanol synthesis, carbon dioxide, if available, is mixed with exit gas from the steam reformer. If carbon dioxide is not available, the conversion must be performed with an excess of hydrogen. Hydrogen accumulates in the synthesis recycle gas and must be removed.

Other Raw Materials. Natural gas is not the only raw material for synthesis gas used in methanol production plants. Higher hydrocarbons (e.g., liquefied petroleum gas, refinery off-gases, and particularly naphtha) are also used (→ Gas Production, Chap. 2.1.); they are processed mainly by steam reforming. Crude oil, heavy oil, tar, and asphalt products (→ Gas Production, Chap. 3.1.) can also be converted into synthesis gas, but this is more difficult than with natural gas. Their sulfur content is considerably higher (0.7 – 1.5 % H_2S and COS) and must be removed. Synthesis gas also contains excess carbon monoxide and must, therefore, be subjected to shift conversion.

Coal can be converted into synthesis gas with steam and oxygen by a variety of processes at different pressures (0.5 – 8 MPa) and temperature (400 – 1500 °C); see also → Coal, Chap. 9.4.; → Gas Production, Chap. 4. Synthesis gas must be desulfurized and subjected to shift conversion to obtain the required stoichiometry for methanol synthesis.

5.2. Synthesis

Important reactions (Eqs. 1 – 3) for the formation of methanol from synthesis gas are discussed in Section 4.1. In one pass only about 50 % of the synthesis gas is converted because thermodynamic equilibrium is reached; therefore, after methanol and water are condensed out

and removed, the remaining synthesis gas must be recycled to the reactor. A simplified flow diagram for methanol synthesis is shown in Figure 3. The make-up synthesis gas is brought to the desired pressure (5 – 10 MPa) in a multistage compressor (f). The unreacted recycle is added before the recycle stage. A heat exchanger (b) transfers energy from the hot gas leaving the reactor to the gas entering the reactor. The exothermic formation of methanol takes place in the reactor (a) at 200 – 300 °C. The heat of reaction can be dissipated in one or more stages. The mixture is cooled further (c) after passing through the heat exchanger (b); the heat of condensation of methanol and water can be utilized at another point in the process.

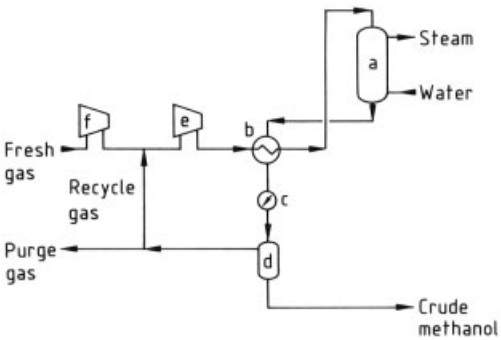

Figure 3. Methanol synthesis
a) Reactor; b) Heat exchanger; c) Cooler; d) Separator; e) Recycle compressor; f) Fresh gas compressor

Crude methanol is separated from the gas phase in a separator (d) and flashed before being distilled. Gas from the separator is recycled to the suction side of the recycle compressor (e). The quantity of purge gas from the loop is governed by the concentration and absolute amount of inert substances and the stoichiometry number. If hydrogen is used to adjust the composition of the fresh gas to give the required stoichiometry number it can be recovered from the purge gas by various methods (e.g., pressure swing absorption). The purge gas is normally used for reformer heating.

5.3. Reactor Design

Current industrial processes for producing methanol differ primarily in reactor design. Many different reactors are available [109]; they may be either adiabatic (e.g., ICI) or quasi-isothermal (e.g., Lurgi). The ICI process (Fig. 4) accounts for 60 %, and the Lurgi process (Fig. 5) for 30 % of worldwide methanol production.

Adiabatic Reactors. The ICI process (Fig. 4) uses an adiabatic reactor with a single catalyst bed [110]. The reaction is quenched by adding cold gas at several points. Thus, the temperature profile along the axis of the reactor has a sawtooth shape.

In the Kellogg process, synthesis gas flows through several reactor beds that are arranged axially in series [111]. In contrast to the ICI quench reactor, the heat of reaction is removed by intermediate coolers. The Haldor Topsoe reactor operates on a similar principle, but synthesis gas flows radially through the catalyst beds [112].

Ammonia – Casale S. A. has developed a reactor that employs a combination of axial and radial flow (mixed flow). This type of reactor initially developed for ammonia plants is offered by Davy McKee in ICI license [113].

Quasi-Isothermal Reactors. The Lurgi process (Fig. 5) employs a tubular reactor (f) with cooling by boiling water [114]. The catalyst is located in tubes over which water flows. The temperature of the cooling medium is adjusted by a preset pressure.

The Variobar reactor developed by Linde [115] consists of a shell-and-tube reactor coiled in several tiers, whose cooling tubes are embedded in the catalyst packing. The reactor temperature is adjusted by water cooling. As in other processes, the heat of reaction is utilized to produce steam, which can be used, for example, to drive a turbine for the compressor or as an energy source for subsequent methanol distillation.

Whereas synthesis gas flows axially through the two above-mentioned reactors, Toyo offers a reactor through which it flows radially [116]. The advantages, as in the Variobar reactor, lie in a high heat transfer rate with only slight pressure loss.

The Mitsubishi Gas Chemical (MGC) process uses a reactor with double-walled tubes that are filled in the annular space with catalyst [117]. The synthesis gas first flows through the inner tube to heat it up and then, in countercurrent, through the catalyst between the two tubes. The

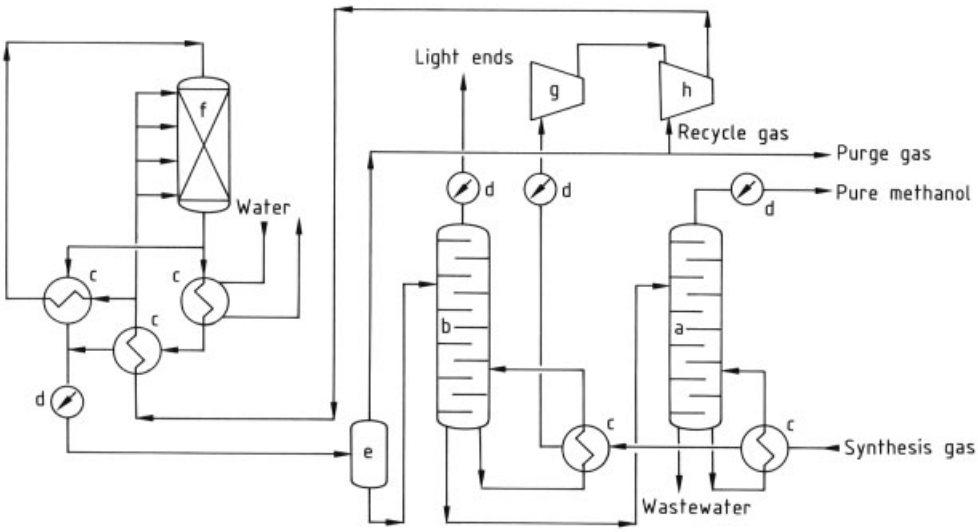

Figure 4. The ICI low-pressure methanol process
a) Pure methanol column; b) Light ends column; c) Heat exchanger; d) Cooler; e) Separator; f) Reactor; g) Compressor; h) Compressor recycle stage

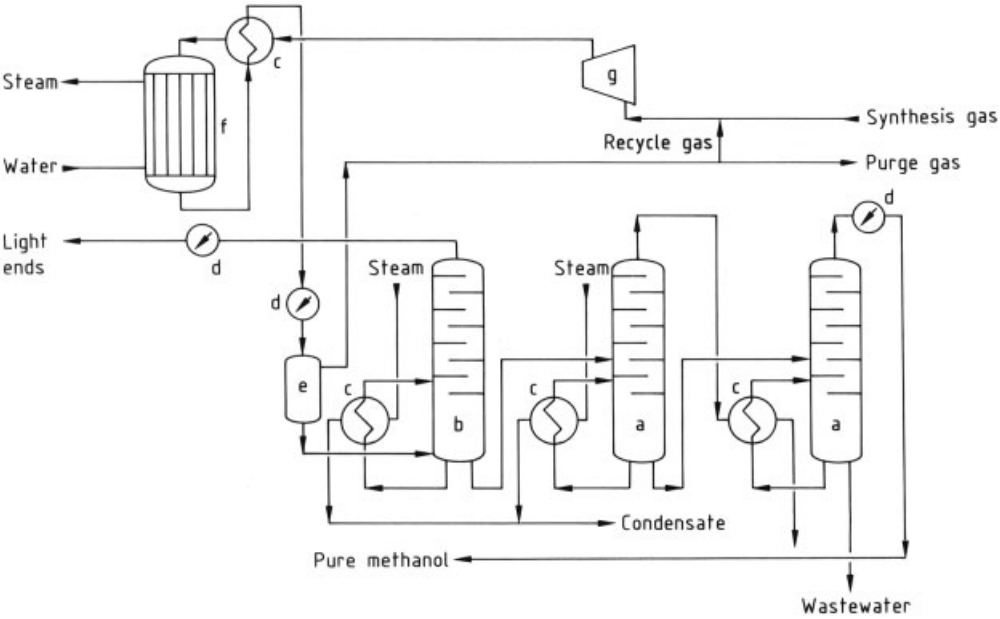

Figure 5. Lurgi low-pressure methanol process
a) Pure methanol columns; b) Light ends column; c) Heat exchanger; d) Cooler; e) Separator; f) Reactor; g) Compressor recycle stage

outer tubes are cooled by water, Mitsubishi considers the main advantage of this process to be the high conversion rate (ca. 14 % methanol in the reactor outlet).

5.4. Distillation of Crude Methanol

Crude methanol leaving the reactor contains water and other impurities (see Section 4.1). The amount and composition of these impurities depend on reaction conditions, feed gas, and type and lifetime of the catalyst. Crude methanol is made slightly alkaline by the addition of small amounts of aqueous caustic soda to neutralize lower carboxylic acids and partially hydrolyze esters.

The methanol contains low-boiling and high-boiling components (light and heavy ends). The light ends include dissolved gases, dimethyl ether, methyl formate, and acetone. The heavy ends include higher alcohols, long-chain hydrocarbons, higher ketones, and esters of lower alcohols with formic, acetic, and propionic acids. Higher waxy hydrocarbons consisting of a mixture of mostly straight-chain $> C_8 - C_{40}$ compounds are also formed in small amounts. They have low volatility and thus remain in the distillation bottoms, from which they can easily be removed because of their low solubility in water and low density.

The impurities in crude methanol are generally separated in two stages. First, all components boiling at a lower temperature than methanol are removed in a light ends column (Fig. 4 b, Fig. 5 b). Pure methanol is then distilled overhead in one or more distillation columns (Fig. 4 a, Fig. 5 a). If the columns operate at different pressures, the heat of condensation of the vapors of the column operating at higher pressure can be used to heat the column at lower pressure.

5.5. Construction Materials

Low-molybdenum steels are normally used as construction materials in methanol synthesis. Because organic acids are especially likely to be encountered in the methanol condensation stage, stainless steels are generally used then. Damage due to acids can also be prevented by the addition of small amounts of dilute caustic soda.

Stainless steels are normally employed in equipment operating at temperatures in which the formation of iron pentacarbonyl is likely (i.e., 100 – 150 °C). This applies, for example, to heat exchangers and compressors. Contamination with iron pentacarbonyl should be avoided because it decomposes at the temperatures used for methanol synthesis. Iron deposited on the catalyst poisons it and promotes the formation of higher hydrocarbons (waxy products).

6. Handling, Storage, and Transportation

6.1. Explosion and Fire Control

The flammability of methanol and its vapors represents a potential safety problem. The flash point is 12.2 °C (closed cup) and the ignition temperature 470 °C; in the Federal Republic of Germany methanol is thus included in ignition group B of the VbF [119].

Methanol vapor is flammable at concentrations of 5.5 – 44 vol %. The saturated vapor pressure at 20 °C is 128 kPa; a saturated methanol – air mixture is thus flammable over a wide temperature range. Methanol is included in ignition group G1, explosion class 1 (ExRL).

In premises and workshops in which the presence of methanol vapor is likely, electrical equipment must be designed in accordance with the relevant regulations:

Guidelines for explosion protection (ExRL)
Regulations governing electrical equipment in explosionhazard areas (ElE × V)
DIN VDE 0165
DIN EN 50 014 – 50 020

For international guidelines on the handling of methanol publications of the Manufacturing Chemists' Association should be consulted [118].

Pure, anhydrous methanol has a very low electrical conductivity. Measures to prevent electrostatic charging must therefore be adopted when transferring and handling methanol.

Fire Prevention. The VbF restrictions on the amount of methanol that can be stored in laboratory premises should be observed. When large amounts of methanol are stored in enclosed

spaces, monitoring by means of lower explosion limit monitors is desirable.

Permanently installed fire-extinguishing equipment should be provided in large storage facilities. Water cannons are generally installed in storage tank farms to cool steel constructions and neighboring tanks in the event of fire. Large tanks should have permanently installed piping systems for alcohol-resistant fire-extinguishing foams.

Fire Fighting. Conventional fire-extinguishing agents such as powder, carbon dioxide, or Halon can be used for small fires. Water is unsuitable as an extinguishing agent for fires involving large amounts of methanol because it is miscible with the compound; mixtures containing small amounts of methanol may also burn. Protein-based alcohol-resistant foams are suitable.

A methanol flame is practically invisible in daylight, which complicates fire fighting. The methanol flame does not produce soot, although formaldehyde and carbon monoxide form during combustion when oxygen is lacking. Respirators must therefore be worn when fighting fires in enclosed areas.

6.2. Storage and Transportation

Small-Scale Storage. Fairly small amounts (≤ 10 L) of methanol for laboratory and industrial use are stored in glass bottles or sheet-metal cans; amounts up to 200 L are stored and transported in steel drums. Some plastic bottles and containers cannot be used because of their permeability and the danger of dissolution of plasticizers. High-density polyethylene and polypropylene are suitable, whereas poly(vinyl chloride) and polyamides are unsuitable.

Large-Scale Storage. Large amounts of methanol are stored in tanks that correspond in design and construction to those used for petroleum products; cylindrical tanks with capacities from a few hundred cubic meters to more than 100 000 m^3 are normally used. With fixed-roof tanks, special measures (e.g., nitrogen blanketing) should be adopted to prevent the formation of an ignitable atmosphere in the space above the liquid surface. Emission of methanol may occur if the level fluctuates. To avoid these problems, large tanks are often equipped with floating roofs; attention should therefore be paid to guard against entry of rainwater.

For anhydrous and carbon dioxide-free methanol tanks, pipelines and pumps can be constructed from normal-grade steel; seals can be made from mineral fiber, graphite, and metal. Styrene – butadiene rubber, chlorine – butadiene rubber, and butyl – chlorobutyl rubber can be used for shaft seals.

Table 5. Federal specifications for pure methanol in the United States

Property	Grade A	Grade AA
Ethanol content, mg/kg		< 10
Acetone content, mg/kg		< 20
Total acetone and aldehyde content, mg/kg	< 30	< 30
Acid content (as acetic acid), mg/kg	< 30	< 30
Color index (APHA)	< 5	< 5
Sulfuric acid test (APHA)	< 30	< 30
Boiling point range (101.3 kPa), must include 64.6 ± 0.1 °C	< 1	< 1
Dry residue, mg/L	< 10	< 10
Density (20 °C), g/cm^3	0.7928	0.7928
Permanganate number	> 30	> 30
Methanol content, wt %	> 99.85	> 99.85
Water content, wt %	< 0.15	< 0.10
Odor	typical, non-persistent	

Large-Scale Transportation. Methanol is traded worldwide. The recent trend toward relocating production to sites that are remote from industrial centers where inexpensive natural gas is available, has meant that ca. 30 % of methanol produced worldwide must be transported by sea to consumer countries (Japan, Europe, United States). Specially built tankers with capacities up to 40 000 t are available for this purpose; ships built to transport petroleum products are also used.

The most important European transshipment point for methanol is Rotterdam. Methanol is distributed to inland industrial regions mainly by inland waterways on vessels with capacities up to 3000 t. Boats specialized for methanol transport are the exception; impurities can therefore be introduced into the methanol due to frequent change of cargo. Analysis prior to delivery is generally essential.

Methanol is also transported by road and rail tank cars. Permanently coupled trains consisting

of several large tank cars with common filling, discharge, and ventilation lines are used to supply large customers.

Transportation via pipeline is only of importance for supplying individual users within enclosed, self-contained chemical complexes.

Safety Regulations Governing Transportation. The transportation of methanol as less-than-carload freight in appropriate vessels, containers, and bulk, is governed by specific regulations that differ from country to country. An effort is being made, and is already well advanced, to coordinate these regulations within the EC. Relevant legal regulations governing less-than-carload and bulk transportation by sea, on inland waterways, and by rail, road, and air are as follows [120]:

IMDG Code (D-GGVSee)	D 3328/E-F 3087, Class 3.2, UN No. 1230
RID (D-GGVE)	Class 3, Rn 301, Item 5
ADR (D-GGVS)	Class 3, Rn 2301, Item 5
ADNR	Class 3, Rn 6301, Item 5, Category Kx
European Yellow Book	No. 603-001-00-X
EC Guideline/D VgAst	No. 603-001-00-X
FRG (Land, VbF)	B
Great Britain	Blue Book: flammable liquid and IMDG Code E 3087
United States	CRF 49, Paragraph 172.1.1, flammable liquid
IATA	RAR, Art. No. 1121/43, flammable liquid

In addition to water, typical impurities include acetone (which is very difficult to separate by distillation) and ethanol. When methanol is delivered by ships or tankers used to transport other substances, contamination by the previous cargo must be expected.

Comparative ultraviolet spectroscopy has proved a convenient, quick test method for deciding whether a batch can be accepted and loaded. Traces of all chemicals derived from aromatic parent substances, as well as a large number of other compounds, can be detected.

Further tests for establishing the quality of methanol include measurements of boiling point range, density, permanganate number, turbidity, color index, and acid number. More comprehensive tests include water determination according to the Karl Fischer method and gas chromatographic determination of byproducts. However, the latter is relatively expensive and time consuming because several injections using different columns and detectors must be made due to the variety of byproducts present.

The most important standardized test methods for methanol are

DIN 51 757	density
ASTM D 941	density
ASTM D 1078	boiling range
ASTM D 1209	color index
ASTM D 1353	dry residue
ASTM D 1363	permanganate number
ASTM D 1364	water content
ASTM D 1612	acetone content
ASTM D 1613	acid content

7. Quality Specifications and Analysis

Methanol for Laboratory Use. Methanol is available commercially in various purity grades for fine chemicals:

1) "Synthesis" quality (corresponding to normal commercial methanol)
2) Certified analytical quality
3) Extremely pure qualities for semiconductor manufacture

Commercial Methanol. In addition to laboratory grades, commercial methanol is generally classified according to ASTM purity grades A and AA (Table 5). Methanol for chemical use normally corresponds to Grade AA.

Apart from pure methanol, methanol obtained directly from synthesis without any purification, or with only partial purification, is sometimes used. This crude methanol can be used for energy (fuel methanol), for the manufacture of synthetic fuels, and for specific chemical and technical purposes; it is not normally available commercially. Composition varies according to synthesis conditions; principal impurities include, 5 – 20 vol % water, higher alcohols, methyl formate, and higher esters. The presence of water and esters can cause corrosion during storage due to the formation of organic acids (see Section 6.2); remedies include alkaline adjustment with sodium hydroxide and, if necessary, the use of corrosion-resistant materials.

8. Environmental Protection

Methanol is readily biodegraded; most microorganisms possess the enzyme alcohol dehydrogenase, which is necessary for methanol oxidation. Therefore, no danger exists of accumulation in the atmosphere, water, or ground; the biological stages of sewage treatment plants break down methanol almost completely. In the Federal Republic of Germany methanol has been classified as a weakly hazardous compound in water hazard Class 1 (WGK I, § 19 Wasserhaushaltsgesetz). In accidents involving transport, large amounts of methanol must be prevented from penetrating into the groundwater or surface waters to avoid contaminating drinking water. Little is known about the behavior of methanol in the atmosphere. Emissions occurring during industrial use are so small that harmful influences can be ignored. That situation could alter, however, if methanol were used on a large scale as an alternative to petroleum-based fuels.

In methanol production, residues that present serious environmental problems are not generally formed. All byproducts are used when possible; for example, the condensate can be processed into boiler feedwater, and residual gases or low-boiling byproducts can be used for energy production. The only regularly occurring waste product that presents some difficulties is the bottoms residue obtained after distillation of pure methanol; it contains water, methanol, ethanol, higher alcohols, other oxygen-containing organic compounds, and variable amounts of paraffins. The water-soluble organic substances readily undergo biological degradation; the insoluble substances can be incinerated safely in a normal waste incineration unit. In some cases this residual water is also subjected to further distillative purification; the resultant mixture of alcohols, esters, ketones, and aliphatics can be added in small amounts to carburetor fuel.

The spent catalysts contain auxiliary agents and supports, as well as copper (synthesis), nickel (gas generation), and cobalt and molybdenum (desulfurization) as active components. These metals are generally recovered or otherwise utilized.

Modern steam reformers can be fired so that emission of nitrogen oxides (NO_x) in the flue gas is maintained below 200 mg/m^3 without having to use secondary measures.

9. Uses

9.1. Use as Feedstock for Chemical Syntheses

Approximately 70 % of the methanol produced worldwide is used in chemical syntheses: in order of importance formaldehyde, methyl *tert*-butyl ether (MTBE), acetic acid, methyl methacrylate, and dimethyl terephthalate. Only a small proportion is utilized for energy production, although this use has great potential.

Formaldehyde. Formaldehyde is the most important product synthesized from methanol (→ Formaldehyde, Chap. 4.); in 1988, 40 %; in 1996 35 %, of the methanol produced worldwide was used to synthesize this product. The annual estimated increase in formaldehyde production from methanol is ca. 3 %, but because other bulk products have higher growth rates its share as a proportion of methanol use will decrease.

The processes employed are all based on the oxidation of methanol with atmospheric oxygen. They differ mainly with regard to temperature and nature of the catalyst used.

Methyl *tert*-butyl ether is produced by reacting methanol with isobutene on acid ion exchangers (→ Methyl tert-Butyl Ether). Increasing amounts of methanol are used in this form in the fuel sector. The ether is an ideal octane booster and has become extremely important due to the introduction of unleaded grades of gasoline and awareness of the possible harmfulness of aromatic high-octane components. In 1988, 20 %; in 1996 27 %, of worldwide methanol production was used for MTBE synthesis; annual increase rates of up to 12 % are expected. The availability of isobutene is becoming an increasing problem in MTBE synthesis, although the situation has recently been improved by the construction of plants for the isomerization of butane and subsequent dehydrogenation of isobutane.

Acetic Acid. Another 9 % of the methanol produced is used to synthesize acetic acid, and annual growth rates of 6 % are estimated. Acetic acid is produced by carbonylation of methanol with carbon monoxide in the liquid phase with cobalt – iodine, rhodium – iodine, or nickel – iodine homogeneous catalysts (→ Acetic Acid, Chap. 4.1.). The older BASF process operates at 65 MPa, whereas more modern processes (e.g., the Monsanto process) operate at 5 MPa. By varying operating conditions the synthesis can also be modified to produce acetic anhydride or methyl acetate.

Other Synthesis Products. In the intensive search after the oil crisis for routes to alternative fuels, processes were developed that allowed fuels to be produced from synthesis gas with methanol as an intermediate. Mobil in the United States has contributed decisively to the development of such processes, which involve mainly the reaction of methanol on zeolite catalysts. The most important and, up to now, the only industrially implemented process is methanol to gasoline (MTG) synthesis. A plant for producing and converting 4500 t/d of methanol from natural gas into 1700 t/d of gasoline has been built and operated as a joint venture between the New Zealand government and Mobil. Since the prices of petroleum products have not risen as expected, ways are now being sought to process the methanol from this plant into pure methanol and to market it as such.

Further synthesis routes that could become important in the event of a scarcity of petroleum products are the methanol to olefins (MTO) and methanol to aromatic compounds (MTA) processes [121].

A product that received great attention as a result of the discussion of environmental damage caused by chlorofluorocarbons is dimethyl ether (→ Dimethyl Ether). It can be used as an alternative propellant for sprays. Compared to propane – butane mixtures also used as propellants, its most important feature is its higher polarity and, thus, its better solubilizing power for the products used in sprays. Dimethyl ether is also used as a solvent, organic intermediate, and in adhesives.

Methanol is used to synthesize a large number of other organic compounds:

Formic acid	preservatives, pickling agents
Methyl esters of organic acids	solvents, monomers
Methyl esters of inorganic acids	methylation reagents, explosives, insecticides
Methylamines	pharmaceutical precursors, auxiliaries, absorption liquids for gas washing and scrubbing
Trimethylphosphine	pharmaceuticals, vitamins, fragrances, fine chemicals
Sodium methoxide	organic intermediates, catalyst
Methyl halides	organic intermediates, solvents, propellants
Ethylene	organic intermediates, polymers, auxiliaries (→ Ethylene)

9.2. Use as Energy Source

Methanol is a promising substitute for petroleum products if they become too expensive for use as fuels. As a result of the oil crisis in the early 1970s, a number of projects were started based on the assumption that the use of methanol produced from coal would be more economical in the medium term than the use of petroleum products. The estimates made at the beginning of the 1980s proved to be too optimistic, however, with regard to costs and to overcoming technical or environmental problems involved in producing synthesis gas from coal, and too pessimistic with regard to the price and availability of crude oil (Table 6). Nearly all the large-scale projects for coal utilization have been discontinued. Large-scale operational plants (e.g., Cool Water, United States and Rheinbraun, Wesseling, FRG) are being shut down or modified for use with other feedstocks [122].

Methanol as a Fuel for Otto Engines. The use of methanol as a motor fuel has been discussed repeatedly since the 1920s. Use has so far been restricted to high-performance engines for racing cars and aeroplanes. The combustion of methanol in four-stroke engines has been investigated for a long time. Methanol has been found to be an ideal fuel in many respects. Because of its high heat of vaporization and relatively low calorific value, a substantially lower combustion chamber temperature is achieved than with conventional motor fuels. Emissions of nitrogen oxides, hydrocarbons, and carbon monoxide are lower. This is offset, however, by increased emission of formaldehyde.

Table 6. Comparison of the efficiencies of natural gas conversion in liquid fuels

Energy carrier	Higher heating value, Gcal/t	Yield, t/t CH_4	Higher heating value, Gcal/t CH_4	Theoretical efficiency	Stoichiometric factor, [H] – [20]/[C]	Technical efficiency ∗ ∗
Methane	13.2	1	13.28	100	4	
Synthesis gas, partial oxidation	6.36	2	12.70	95.7	2	85 – 90
Synthesis gas, steam reforming	7.96	2.12 ∗	16.91 ∗		6	
Methanol	5.36	2	10.72	80.7	2	68 – 72
Ethanol	7.14	1.43	10.25	77.2	2	
Kerosene	11.00	0.87	9.57	72.1	ca. 2.05	
Diesel fuel	10.70	0.87	9.34	70.4	ca. 2	55 – 60 FT
Gasoline (average)	10.50	0.86	9.06	68.2	ca. 1.8	58 – 63 MTG 55 – 60 FT
Benzene	10.02	0.81	8.13	61.2	1	

∗ With extra heat energy for reformer.
∗ ∗ FT = Fischer – Tropsch; MTG = methanol to gasoline.

Table 7. Comparison of methanol and a typical fuel (gasoline) for use in Otto engines

Property	Gasoline	Methanol
Density, kg/L	0.739	0.793
Calorific value, kJ/kg	44 300	21 528
Air consumption, kg/kg	14.55	6.5
Research octane number	97.7	108.7
Motor octane number	89	88.6
Mixed research octane number		120 – 130
Mixed motor octane number		91 – 94
Reid vapor pressure, kPa	64	32
Boiling point range, °C	30 – 190	65
Heat of vaporization, kJ/kg	335	1174
Cooling under vaporization with stoichiometric amount of air, °C	20	122

The important properties of methanol for use as a fuel are compared with those of a conventional fuel (gasoline) in Table 7. Consumption is higher because of the lower calorific values.

Methanol can be used in various mixing ratios with conventional petroleum products:

M 3 Mixture of 3 % methanol with 2 – 3 % solubilizers (e.g., isopropyl alcohol) in commercially available motor fuel. This system is already widely used because modification of motor vehicles and fuel distribution systems is not required.

M 15 Mixture of 15 % methanol and a solubilizer with motor fuel; alterations to the motor vehicles are necessary in this case. The proposed use of M 15 to increase the octane number in unleaded gasoline has been supplanted by the large increase in the use of MTBE.

M 85 Methanol containing 15 % C_4–C_5 hydrocarbons to improve cold-start properties. Modified vehicles and fuel distribution systems are necessary.

M 100 Pure methanol–vehicles must be substantially modified and fully adapted to methanol operation.

The necessary modifications for methanol operation involve the replacement of plastics used in the fuel system (see Section 6.2). The ignition system and carburetor or fuel injection unit also have to be adapted. With M 85 and M 100 the fuel mixture must be preheated because vaporization of the stoichiometric amount of methanol in the carburetor results in a cooling of 120 K.

In mixtures with a low methanol content (M 3, M 15) phase separation in the presence of traces of water must be avoided. Absolutely dry storage, transportation, and distribution systems must be available for mixed fuels to prevent separation of water – methanol and hydrocarbon phases.

A further restriction on the use of methanol in gasoline is imposed by the increase in gasoline vapor pressure (Reid vapor pressure, RVP). In some warm regions of the United States, legal restrictions on the RVP have already been introduced to reduce hydrocarbon emissions, which are an important factor in the formation of photochemical smog and increased ozone concentration in the lower atmosphere. As a result, methanol can not longer be added to motor fuel because it increases the vapor pressure of the butane used as a cheap octane booster.

Widespread use of methanol as an exclusive fuel for cars is inhibited by its high cost and the lack of a suitable distribution system. Possible solutions to the latter problem include the construction of dual-purpose vehicles (flexible fuel vehicles), which can use either methanol or nor-

mal fuel. Another solution is to use methanol for company or government car fleets, which refill their tanks at a few specific filling stations. Trials based on this concept are underway in several countries; the largest is taking place in California [123].

Methanol as Diesel Fuel. Exclusive operation with methanol is not possible in diesel engines because methanol has a cetane number of 3 and will therefore not ignite reliably. To ensure ignition the engine must have an additional injector for normal diesel fuel; methanol is injected into the cylinder after ignition of the diesel fuel [124]. Additives are being developed to improve ignition performance.

Other Uses of Methanol in the Fuel Sector. In contrast to pure methanol, the use of MTBE in Otto engine fuels is not limited by considerations of miscibility or vapor pressure. The use of methanol for MTBE synthesis could soon quantitatively overtake its conventional uses. Arco, the world's largest producer of MTBE, is also promoting the use of oxinol, a mixture of methanol and *tert*-butanol.

An additional development in the use of methanol is the Lurgi Octamix process. Use of an alkali-doped catalyst and modified conditions (higher temperature, lower CO_2 concentration, higher CO concentration) in methanol synthesis yields a mixture of methanol, ethanol, and higher alcohols [125]. This mixture can be used directly in the engine fuel. The presence of higher alcohols is desirable not only because of the increase in octane number, but also because they act as solubilizers for methanol. However, this process is not yet used on an industrial scale.

Other Energy Uses of Methanol. A use that has been discussed particularly in the United States and implemented in pilot projects is the firing of peak-load gas turbines in power stations (peak shaving). Benefits include simple storage and environmentally friendly combustion in the gas turbine. The use of methanol as a fuel in conventionally fired boilers obviates the need for costly flue gas treatment plants but is not yet economically viable.

The gasification of methanol to obtain synthesis gas or fuel gas has often been proposed. Apart from exceptions such as the production of town gas in Berlin, here too, economic problems have prevented technical implementation.

9.3. Other Uses

Methanol's low freezing point and its miscibility with water allow it to be used in *refrigeration systems,* either in pure form (e.g., in ethylene plants) or mixed with water and glycols. It is also used as an *antifreeze* in heating and cooling circuits; compared to other commonly used antifreezes (ethylene glycol, propylene glycol, and glycerol), it has the advantage of lower viscosity at low temperature. It is, however, no longer used as an engine antifreeze; glycol-based products are employed instead.

Large amounts of methanol are used *to protect natural gas pipelines* against the formation of gas hydrates at low temperature. Methanol is added to natural gas at the pumping station, conveyed in liquid form in the pipeline, and recovered at the end of the pipeline. Methanol can be recycled after removal of water taken up from natural gas by distillation.

Methanol is also used as an *absorption agent* in gas scrubbers. The removal of carbon dioxide and hydrogen sulfide with methanol at low temperature (Rectisol process, Linde and Lurgi) has the advantage that traces of methanol in the purified gas do not generally interfere with further processing [126].

The use of pure methanol as a *solvent* is limited, although it is often included in solvent mixtures.

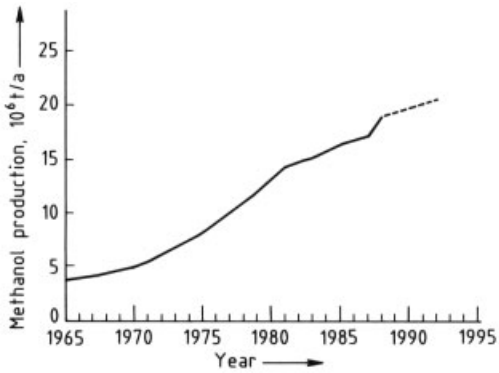

Figure 6. Worldwide methanol production
The estimate for 1989 – 1992 is based on a utilization of 80 % capacity [126], [127].

10. Economic Aspects

Economics of Methanol Production. The costs of methanol production depend on many factors, the most important being direct feedstock costs, investment costs, and costs involved in logistics and infrastructure.

Natural gas, naphtha, heavy heating oil, coal, and lignite are all used as feedstocks in methanol plants. In heavy oil-based plants and to an increasing extent in coal-based plants the principal cost burden is accounted for by capital costs. This means that, given the currently prevailing low energy prices, such plants have high fixed costs and are, therefore, uneconomical or economical only under special conditions. Under present conditions the balance between investment and operating costs clearly favors natural gas-based plants. All large plants currently being planned are designed for use with natural gas and some plants built for operation with naphtha have been converted. Less than 2×10^6 t of the currently installed worldwide capacity of ca. 21×10^6 t is based on raw materials other than natural gas.

Methanol on the World Market. After ammonia, methanol is quantitatively the largest product from synthesis gas. Worldwide capacity at the beginning of 1989 was 21×10^6 t. In 1988, 19×10^6 t of methanol was produced worldwide. The mean annual production growth rate is ca. 10 %. The production curve for methanol since 1965 is illustrated in Figure 6. Worldwide capacity in 1996 was 29.1×10^6 t/a, 24.3×10^6 t/a of methanol was produced.

The methanol industry underwent radical structural changes during the 1980s. Previously, companies that consumed large quantities of methanol produced the compound themselves from the most readily accessible raw materials at the site of use (i.e., highly industrialized countries with expensive energy sources). Since then the number of plants that produce methanol at remote sites exclusively for sale to processors has risen dramatically.

After the energy crisis of the 1970s, intensive oil prospecting led to the discovery of large natural gas fields in many remote regions. Because little demand for natural gas existed in these regions, the relevant countries in South America, Asia, and the Caribbean were interested in selling natural gas as such or in another form to industrialized countries.

Another, hitherto little-used energy source is the associated gas, which is often flared off. In addition to the transportation of liquefied methane and its use as a starting material for ammonia production, methanol production is often the most suitable alternative for marketing such gases. The technology of methanol production is relatively simple, and transport and storage involve inexpensive technology. On the basis of these considerations, 14 new natural gas-based plants producing methanol for export were built from 1974 to 1985 [128]. The largest single train plant based on this concept is located at Punta Arenas in southern Chile; it came on stream in 1988 and has an output of 750 000 t/a. As a consequence of this development, older methanol plants in industrialized countries such as the United States, Japan, and the Federal Republic of Germany have been shut down. The shift in capacities is illustrated in Figure 7.

Since a close relationship between supply and demand no longer exists, large price fluctuations occur, which are hardly justified by actual market conditions. This makes long-term price forecasts impossible and increases economic risks for new projects.

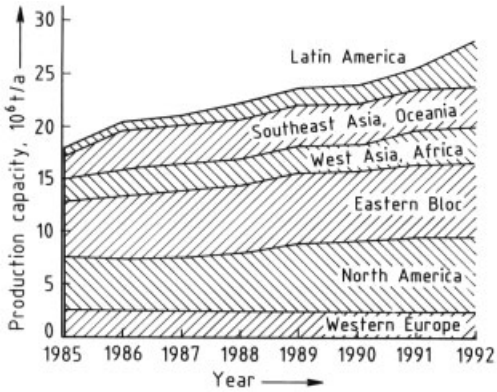

Figure 7. Distribution of existing and planned production capacity for methanol according to region [127]

11. Toxicology and Occupational Health

11.1. Toxicology

Human Toxicology. The first accounts of the poisonous action of "methylated spirits" were

published in 1855 [129]. However, the number of cases of poisoning increased only after the production of a low-odor methanol. In 1901, De Schweinitz reported the first cases of industrial poisoning [130].

Liquid methanol is fully absorbed via the gastrointestinal tract [131] and the skin [132] (absorption rate, 0.19 mg cm^{-2} min^{-1}). Methanol vapor is taken up in an amount of 70 – 80 % by the lungs [133]. The compound is distributed throughout body fluids and is largely oxidized to formaldehyde and then to formic acid [134]. It is eliminated unchanged through the lungs [132] and in the urine. Elimination half-life is ca. 2 – 3 h.

The metabolism of methanol to formic acid in humans and primates is catalyzed by the enzyme alcohol dehydrogenase in the liver. This enzyme can be inhibited competitively by ethanol. Formic acid is oxidized to carbon dioxide and water in the presence of folic acid. Because folic acid is not available in sufficient amount in primates, formic acid may accumulate in the body. This leads to hyperacidity of the blood (acidosis), which is ultimately responsible for methanol poisoning [134].

The symptoms of methanol poisoning do not depend on the uptake route (percutaneous, inhalational, oral) and develop in three stages. An initial narcotic effect is followed by a symptom-free interval lasting 10 – 48 h. The third stage begins with nonspecific symptoms such as abdominal pain, nausea, headache, vomiting, and lassitude, followed by characteristic symptoms such as blurred vision, ophthalmalgia, photophobia, and possibly xanthopsia. Depending on the amount of methanol, individual sensitivity, and the time when treatment is initiated, visual disturbances can either improve or progress within a few days to severe, often irreversible impairment of sight or even to blindness [136–139]. The symptoms are accompanied by increasing hyperacidity of the blood due to the accumulation of formic acid, with disturbances in consciousness, possibly deep coma, and in severe cases, death within a few days. The lethal dosage is between 30 and 100 mL per kilogram of body weight. Sensitivity to methanol varies widely, however. Cases have been reported in which no permanent damage occurred after drinking relatively large amounts of methanol (200 or 500 mL) [135], [140]; in another case, however, irreversible blindness resulted after consumption of 4 mL [141].

The treatment of acute oral methanol poisoning [137] should be initiated as quickly as possible with the following measures:

1) Administration of ethanol: In suspected cases of methanol poisoning, 30 – 40 mL of ethanol (e.g., 90 – 120 mL of whiskey) is administered immediately as a prophylactic before the patient is referred to a hospital. Because ethanol has a greater affinity for alcohol dehydrogenase than methanol, oxidation of methanol is inhibited; the production of formaldehyde and formic acid from methanol is thus suppressed.
2) Gastric lavage
3) Hemodialysis
4) Treatment with alkali: sodium bicarbonate is infused to control blood hyperacidity.
5) Administration of CNS stimulants (analeptics)
6) Drinking larger volumes of fluid
7) Eye bandage: the eyes should be protected against light
8) The patient should be kept warm

Methanol has a slight irritant action on the eyes, skin, and mucous membranes in humans. Concentrations between 1500 and 5900 ppm are regarded as the threshold value of detectable odor.

Chronic methanol poisoning is characterized by damage to the visual and central nervous systems. Case histories [142], [143] have not been sufficiently documented; whether poisoning is caused by chronic ingestion of low doses or ingestion of intermittently high (subtoxic) amounts is uncertain.

Animal Toxicology. Experiments on animals have shown that methanol does not cause acidosis or eye damage in nonprimates (e.g., rats, mice); it generally has a narcotic, possibly lethal, effect. Investigations on laboratory animals cannot, therefore, be extrapolated to humans, at least in the higher dosage range.

In a study on reproductive toxicology, methanol was administered to rats by inhalation during pregnancy. No embryotoxic effects were found after exposure to 5000 ppm [144]. The authors conclude that observance of the recommended

concentrations (MAK or TLV values) offers sufficient protection against fetal abnormalities in humans.

In the Ames test, the sex-linked lethal test on *Drosophila melanogaster* and the micronucleus test in mice, methanol was not mutagenic [145], [146].

11.2. Occupational Health

No special precautions need be taken when handling methanol because it is not caustic, corrosive, or particularly harmful environmentally. If methanol is released under normal conditions, no danger exists of buildup of acutely toxic concentrations in the atmosphere. (Chronic poisoning via the respiratory tract or oral ingestion is described in Section 1.) However, absorption through the skin does constitute a danger, and methanol should be prevented from coming in direct contact with skin.

Appropriate workplace hygiene measures should be adopted if methanol is handled constantly. Rooms in which methanol is stored or handled must be ventilated adequately. The TLV – TWA value (skin) is 200 ppm (262 mg/m^3), and the TLV – STEL value is 250 ppm (328 mg/m^3). The MAK value is 200 ppm (270 mg/m^3). Gas testing tubes can be used to measure the concentration in air. The peak limit should correspond to category II, 1: i.e., the MAK value may be exceeded by a maximum of 100 % for 30 min, four times per shift [146]. Respirators must be worn if substantially higher concentrations are present. Filter masks (filter A, identification color brown) can be used only for escape or life-saving purposes because they are exhausted very quickly. Respirators with a self-contained air supply and heavy-duty chemical protective clothing should be used for longer exposures to high methanol concentrations ($>$ 0.5 vol %).

12. References

1. R. C. Weast, D. R. Lide: *CRC Handbook of Chemistry and Physics,* 70th ed., CRC Press, Boca Raton 1989.
2. H. Stephen, T. Stephen (eds.): *Solubilities of Inorganic and Organic Compounds,* Pergamon Press, Oxford – London – Edinburgh – New York – Paris – Frankfurt 1964.
3. T. E. Daubert, R. P. Danner: *Data Compilation Tables of Properties of Pure Compounds,* Design Institute for Physical Property Data and American Institute of Chemical Engineers, New York 1985.
4. J. A. Riddick, W. B. Bunger: "Organic Solvents" in *Techniques of Chemistry,* 3rd ed., vol. 2, Wiley-Interscience, New York 1986.
5. K. N. Marsh et al. (eds.): "Non-Hydrocarbons," *TRC Thermodynamic Tables,* Suppl. 61, Thermodynamics Research Center 1989.
6. J. B. Pedley et al.: *Thermochemical Data of Organic Compounds,* 2nd ed., Chapman and Hall, London – New York 1986.
7. *Engineering Sciences Data,* vol. 79 028, ESDU International plc, London 1979.
8. *Engineering Sciences Data,* vol. 88 005, ESDU International plc, London 1988.
9. B. D. Smith, R. Srivastava: "Thermodynamic Data for Pure Compounds," in *Physical Sciences Data 25,* part B, Elsevier, Amsterdam – Oxford – New York – Tokyo 1986.
10. K. H. Simmrock et al.: "Critical Data of Pure Compounds," in D. Behrens, R. Eckermann (eds.): *Chemistry Data Series,* vol. 2, part 1, Dechema, Frankfurt 1986.
11. L. H. Horsley: "Azeotropic Data III," in *Advances in Chemistry Series,* vol. 116, American Chemical Society, Washington, D.C. 1973, pp. 74 – 83.
12. Engineering Sciences Data, vol. 83 016, ESDU International plc, London 1983.
13. C. L. Yaws: *Physical Properties, a Guide to the Physical, Thermodynamic and Transport Property Data of Industrially Important Chemical Compounds,* McGraw-Hill, New York 1977, pp. 197 – 226.
14. K. Noda et al., *J. Chem. Eng. Data* **27** (1982) 326.
15. Th. W. Yergovich et al., *J. Chem. Eng. Data* **16** (1971) no. 2, 222.
16. *Landolt-Börnstein,* **new series, group 4, vol. 1,** part b, 117 – 118.
17. *Landolt-Börnstein,* 6th ed., **vol. 2,** part 5, 366.
18. Y. Tanaka et al., *Int. J. Thermophys.* **8** (1987) no. 2, 147.
19. A. E. Easteal, L. A. Woolf, *J. Chem. Thermodyn.* **17** (1985), 49 – 62, 69 – 82.
20. J. Rex Goates et al., *J. Chem. Thermodyn.* **11** (1979) 739.
21. *Engineering Sciences Data,* vol. 75 024, ESDU International plc, London 1975.

22. *Landolt-Börnstein,* **7, Elektrische Eigenschaften II,** 18.
23. *Landolt-Börnstein,* **6, Elektrische Eigenschaften I,** 632, 711 – 712.
24. D. W. Nageli et al.: *Practical Ignition Limits for Low Molecular Weight Alcohols,* The Combustion Institute, March 1984, pp. 18 – 20.
25. N. J. Sax, R. J. Lewis: *Dangerous Properties of Industrial Materials,* 7th ed., **vol. 3,** Van Norstrand Reinhold, New York 1989, p. 2217.
26. J. A. Monick: *Alcohols, Their Chemistry, Properties and Manufacture,* Reinhold Publ. Co., New York 1968.
27. E. J. Wickson: "Monohydric Alcohols," *ACS Symp. Ser.* 1981, no. 159 (March 25 – 26, 1980).
28. C. Ferri: *Reaktionen der organischen Synthese,* Thieme Verlag, Stuttgart 1978.
29. P. Sykes: *A Guidebook to Mechanisms in Organic Chemistry,* 6th ed., Longman Group, London 1986.
30. D. R. Stull, H. Prophet et al. (eds.): *JANAF Thermochemical Tables,* 2nd ed., NSDRS, NBS-37, U.S. Government Printing Office, Washington, D.C. 1971.
31. W. J. Thomas, S. Portalski, *Ind. Eng. Chem.* **50** (1958) 967.
32. R. H. Newton, B. F. Dodge, *J. Am. Chem. Soc.* **56** (1934) 1287.
33. R. M. Ewell, *Ind. Eng. Chem.* **32** (1940) 149.
34. V. M. Cherednichenko, Ph. D. Thesis, Korpova, Physico-Chemical Institute, Moscow 1953.
35. T. Chang, R. W. Rousseau, P. K. Kilpatrick, *Ind. Eng. Chem. Process Des. Dev.* **25** (1986) 477.
36. G. H. Graaf, P. J. J. M. Sijtsema, E. J. Stamhuis, G. E. H. Joosten, *Chem. Eng. Sci.* **41** (1986) no. 11, 2883.
37. W. Kotowski, *Przem. Chem.* **44** (1965) 66.
38. *Kirk-Othmer,* **15,** 398 – 415.
39. L. Bisset, *Chem. Eng. (N.Y.)* **84** (1977) no. 21, 155.
40. G. C. Chinchen et al., *Appl. Catal.* **36** (1988) 1 –65.
41. O. A. Hougen, K. M. Watson, R. A. Ragatz: *Chemical Process Principles,* part 2, Wiley-Interscience, New York 1959.
42. G. Soave, *Chem. Eng. Sci.* **27** (1972) 1197.
43. D. Y. Peng, D. B. Robinson, *Ind. Eng. Chem. Fundam.* **15** (1976) 59.
44. F. Marschner, F. W. Moeller *Appl. Ind. Catal.* **2** (1983) 349 – 411.
45. K. Klier, V. Chatikavanij, R. G. Herman, G. W. Simmons, *J. Catal.* **74** (1982) 343.
46. K. Klier, *Adv. Catal.* **31** (1982) 243 – 313.
47. J. C. J. Bart, R. C. P. Sneeden, *Catal. Today* **2** (1987) 1 – 124.
48. G. C. Chinchen et al.: *ACS Symposium on Methanol and Synthetic Fuels,* Philadelphia, August 1984.
49. M. Bowker et al., *J. Catal.* **109** (1988) 263.
50. M. A. McNeil, C. J. Schack, R. G. Rinker, *Appl. Catal.* **50** (1989) 265.
51. A. Fischer et al., *Proc. Int. Congr. Catal. 7th, 1980,* 1981, 341.
52. P. J. Denny, D. A. Whan, *Catalysis (London)* **2** (1978) no. 3, 46 – 86.
53. H. H. Kung, *Catal. Rev. Sci. Eng.* **22** (1980) no. 2, 235.
54. R. J. Hawkins, R. J. Kane, W. E. Slinkard, J. L. Trumbley: "Methanol" in J. J. Ketta, W. A. Cunningham (eds.): *Encyclopedia of Chemical Processing and Design,* Marcel Dekker, New York 1988, pp. 418 – 483.
55. A. Deluzarche, R. Kieffer, A. Muth, *Tetrahedron Lett.* **38** (1977) 3357.
56. R. G. Herman et al., *J. Catal.* **56** (1979) 407.
57. G. Henrici-Olive, S. Olive, *J. Mol. Catal.* **17** (1982) 89.
58. W. Seyfert, G. Luft, *Chem. Ing. Tech.* **57** (1985) 482.
59. R. M. Agny, C. G. Takoudis, *Ind. Eng. Chem. Prod. Res. Dev.* **24** (1985) 50.
60. Y. B. Kagan et al., *Dokl. Akad. Nauk. SSSR* **224** (1975) 1081.
61. G. Liu, D. Willcox, M. Garland, H. H. Kung, *J. Catal.* **96** (1985) 251.
62. G. C. Chinchen et al., *Appl. Catal.* **30** (1987) 333.
63. M. Bowker, H. Houghton, K. C. Waugh, *J. Chem. Soc. Faraday Trans 1,* **77** (1981) 3023.
64. V. D. Kuznetsov, F. S. Shub, M. I. Temkin, *Kinet. Katal.* **25** (1984) no. 3, 606.
65. G. H. Graaf, E. J. Stamhuis, A. Beenackers, *Chem. Eng. Sci.* **43** (1988) no. 12, 3185.
66. K. J. Smith, R. B. Anderson, *J. Catal.* **85** (1984) 428.
67. K. Klier in S. Kaliaguine, A. Mahay (eds.): *Catalysis on the Energy Scene, Studies in Surface Science and Catalysis,* **vol. 19,** Elsevier, Amsterdam 1984, p. 439.
68. B. Denise, R. P. A. Sneeden, C. Hamon, *J. Mol. Catal.* **17** (1982) 359.
69. R. Bardet, J. Thivolle-Cazat, Y. Trambouze, *C_1 Mol. Chem.* **1** (1985) no. 3, 201.
70. E. Ramaroson, R. Kieffer, A. Kiennermann, *Appl. Catal.* **4** (1982) 281.
71. B. Denise, R. P. A. Sneeden, *C_1 Mol. Chem.* **1** (1985) 307.

72. E. R. A. Matulewicz, Ph. D. Thesis, University of Amsterdam, 1984.
73. D. J. Elliott, F. Pennella, *J. Catal.* **119** (1989) 359.
74. G. Natta: *Catalysis,* **vol. 3,** Reinhold Publ. Co., New York 1955, pp. 349 ff.
75. BASF, DE 441 433, 1923 (A. Mittasch, M. Pier, K. Winkler).
76. BASF, DE 109 495, 1923.
77. BASF, GB 237 030, 1925.
78. C. Lormand, *Ind. Eng. Chem.* **17** (1925) 430.
79. E. Blasiak, PL 340 000, 1947.
80. ICI, GB 1 159 035, 1965 (J. Gallagher, Y. H. Kiold).
81. *Crit. Report Appl. Chem.* **13** (1985) 102.
82. Ph. Courty, Ch. Marceilly: *Preparation of Catalysts III,* Elsevier, Amsterdam 1983, pp. 485 ff.
83. IFP, EP 0 152 314, 1987 (P. Courty et al.).
84. Süd Chemie, EP 0 125 689, 1987 (M. Schneider, K. Kochloefl, J. Ladebeck).
85. Shell, US 3 709 919, 1973 (E. F. Magoon).
86. ICI, GB 1 159 035, 1965 (J. Gallagher, Y. H. Kiold).
87. BASF, DE 2 846 614, 1978 (F. J. Bröcker et al.).
88. R. H. Höppener, E. B. M. Doesburg, J. J. F. Scholten, *Appl. Catal.* **25** (1986) 109 – 194.
89. P. Gherardi, O. Ruggeri, F. Trifiro, A. Vaccari: *Preparation of Catalysts III,* Elsevier, Amsterdam 1983, pp. 723 ff.
90. M. V. Twigg (ed.): *Catalyst Handbook,* 2nd ed., Wolfe Publishing Ltd., London 1989, pp. 441 – 468.
91. S. Lee, A. Sawant, K. Rodrigues, *Energy Fuels* **3** (1989) 2 – 7.
92. Mitsubishi Gas Chemical, GB 2 047 556, 1980 (K. Asakawa, Y. Yamamoto, S. Ebata, T. Nakamura).
93. Shell, GB 1 371 638, 1974 (E. F. Magoon).
94. Institut Ciezkiý Syntezy, GB 2 025 252, 1980 (W. Kotowski et al.).
95. R. G. Herman, G. W. Simmons, K. Klier: *New Horizons in Catalysis, Studies in Surface Science and Catalysis,* **vol. 7,** Elsevier, Amsterdam 1981, p. 475.
96. IFP, GB 2 037 176, 1980 (A. Sugier, P. Courty, E. Freund).
97. Shell, DP 2 154 074, 1972 (E. F. Magoon, L. H. Slaugh).
98. P. G. Bondar et al., US 4 107 089, 1978.
99. Metallgesellschaft, GB 1 287 327, 1971 (R. Herbert, H. Liebgott).
100. Mitsubishi Gas Chemical, GB 2 095 233, 1982 (K. Asakawa).
101. W. L. Marsden, M. S. Wainwright, J. B. Friedrich, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 551.
102. Unisearch, GB 2 066 856, 1981 (M. S. Wainwright, W. L. Marsden, J. B. Friedrich).
103. H. E. Curry-Hyde, D. J. Young, M. S. Wainwright, *Applied Catalysis* **29** (1987) 31 – 41.
104. H. E. Curry-Hyde, M. S. Wainwright, D. J. Young: *Methane Conversion,* Elsevier, Amsterdam 1988, p. 239.
105. W. G. Baglin, G. B. Atkinson, L. J. Nicks, *Ind. Eng. Chem. Prod. Res. Dev.* **20** (1981) 87.
106. J. E. France, W. E. Wallace, *Paper to the 12th Regional ACS Meeting,* Pittsburgh 1980.
107. J. B. Friedrich, M. S. Wainwright, D. J. Young, *J. Catal.* **81** (1983) 14.
108. E. Supp, *Energy Prog.* **5** (1985) no. 3, 127.
109. U. Zardi, *Hydrocarbon Process.* **61** (1982) no. 8, 129.
110. A. Pinto, P. L. Rogerson, *Chem. Eng. Prog.* **73** (1977) no. 7, 95.
111. L. R. LeBlanc, J. M. Rovner, H. Roos: *Advanced Methanol Plant Design,* 1989 World Methanol Conference, Houston, Crocco & Associates, 1989.
112. *Hydrocarbon Process.* **62** (1983) no. 11, 111.
113. R. E. Smith, G. C. Humphreys, G. W. Griffiths, *Hydrocarbon Process.* **63** (1984) no. 5, 95.
114. E. Supp, W. Hilsebein, *CEER Chem. Econ. Eng. Rev.* **17** (1985) no. 5, 18.
115. *CEER Chem. Econ. Eng. Rev.* **17** (1985) no. 5, 23.
116. K. Ohsaki et al., *CEER Chem. Econ. Eng. Rev.* **17** (1985) no. 5, 31.
117. I. Takase, K. Niva, *CEER Chem. Econ. Eng. Rev.* **17** (1985) no. 5, 24. *Hydrocarbon Process.* **64** (1985) no. 11, 146.
118. Manufacturing Chemists Association, Washington, D.C.: Safety Guide SG-3, Flammable Liquids Storage and Handling Drum Lots and Smaller Quantities.Chemical Safety Data Sheet SD-22, 1970. Manual TC-8, Recommended Practices for Bulk Loading and Unloading Flammable Liquid Chemicals to and from Tank Trucks. Manual TC-29, Loading and Unloading Flammable Liquid Chemical, Tank Cars.
119. Kühn/Birett: *Merkblätter gefährlicher Arbeitsstoffe,* 42. Erg.-Lfg. 2/89, Blatt M 10, Ecomed-Verlag, München.
120. Hommel: *Handbuch der gefährlichen Güter,* 4th ed., Merkblatt 123, Springer Verlag, Berlin – Heidelberg 1987.

121. P. J. Jackson, N. White: "Technologies for the Conversion of Natural Gas," *Austr. Inst. Energy Conference,* 1985.
122. F. Asinger: *Methanol – Chemie- und Energierohstoff,* Springer Verlag, Berlin – Heidelberg 1986.
123. R. Hagar, *Oil Gas J.,* December 19th, 1988, p. 23.
124. F. Pischinger, C. Havenith, G. Finsterwalder, *VDI-Ber.* **370** (1980) 331 – 338.
125. H. Hiller, E. Supp, *Erdöl Kohle Erdgas Petrochem.* **38** (1985) no. 1, 19 – 22.
126. *Linde Ber. Tech. Wiss.* **33** (1973) 6 – 12.
127. *Ullmann,* 4th ed., **16,** 621 – 633.
128. ChemFacts: *Methanol 1989,* Chemical Intelligence Service, London 1989.
129. J. F. McFarlan, *Pharm. J. Trans.* **15** (1855) 310.
130. G. E. De Schweinitz, *Ophthalmic. Rec.* **10** (1901) 289.
131. M. W. Haggard, L. A. Greenberg, *J. Pharmacol. Exp. Ther.* **66** (1939) 479.
132. B. Dutkiewicz, J. Konczalik, W. Karwacki: *Int. Arch. Arbeitsmed.* **47** (1980) 81.
133. G. Leaf, L. J. Zatmann, *Br. J. Ind. Med.* **9** (1952) 19.
134. J. Pohl, *Arch. Exp. Pathol. Pharmakol.* **31** (1893) 281.
135. R. Heinrich, J. Angerer, *Int. Arch. Arbeitsmed.* **50** (1982) 341.
136. US Department of Health Education and Welfare, NIOSH, *Criteria for a recommended Standard Occupational Exposure to Methyl Alcohol,* 1976.
137. W. Wirth, G. Hecht, C. Gloxhuber: *Toxikologie Fibel,* 2nd ed., Thieme Verlag, Stuttgart 1971, p. 200.
138. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 6th ed., Thieme Verlag, Stuttgart 1980, p. 259.
139. W. Forth, D. Henschler, W. Rummel: *Pharmakologie und Toxikologie,* 5th ed., Wissenschaftsverlag, Mannheim 1987.
140. I. L. Bennett, T. C. Nation, J. F. Olley, *J. Lab. Clin. Med.* **40** (1952) 405.
141. W. S. Duke-Elder: *Textbook of Ophthalmology,* **vol. 3,** The C. V. Mosby Co., St. Louis 1945, p. 3021.
142. M. Burk, *Mbl. Augenheilk.* **130** (1957) 845.
143. A. Dreyfus, *Z. Unfallmed. Berufskrankh.* **1** (1946) 84.
144. B. K. Nelson et al., *Fundam. Appl. Toxicol.* **5** (1985) 727.
145. V. F. Simmon, K. Kauhanen, R. G. Tardiff: "Mutagenic activity of chemicals identified in drinking water," in D. Scott, B. A. Bridges, F. H. Sobels, *Progress in Genetic Toxicology,* Elsevier, Amsterdam 1977.
146. E. Gocke et al., *Mutat. Res.* **90** (1981) 91.
147. *MAK-Werte 1986/Technische Regeln für Gefahrstoffe TRGS 900 (UVV Anlage 4),* Jedermann Verlag, Heidelberg 1986.

Methine Dyes and Pigments

Roderich Raue, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. History 3
3. Preparation 5
4. Properties 6
5. Uses 6
6. Economic Aspects 7
7. Cationic Methine Dyes 7
7.1. Streptocyanine Dyes 7
7.1.1. Trimethinestreptocyanine Dyes . . . 7
7.1.2. Pentamethinestreptocyanine Dyes and Higher Vinylogs 8
7.1.3. Aminoarylenamine Dyes 8
7.2. Hemicyanine Dyes 9
7.2.1. Cyclomethine Dyes 9
7.2.2. Dimethinehemicyanine Dyes 9
7.2.3. Higher Vinylogs of Hemicyanine Dyes 14
7.3. Phenylogous Hemicyanine Dyes . . 14
7.3.1. Zeromethine (Apocyanine) Dyes . . 14
7.3.2. Phenylogous Dimethinehemicyanine Dyes 15
7.3.3. Higher Vinylogs of Phenylogous Hemicyanine Dyes 25
7.3.4. Phenylogous Hemicyanine Dyes with Rings in the Methine Chain 25
7.4. Cyanine Dyes 26
7.4.1. Zeromethinecyanine Dyes 26
7.4.2. Monomethinecyanine Dyes 26
7.4.3. Dimethinecyanine Dyes 28
7.4.4. Trimethinecyanine Dyes 30
7.4.5. Pentamethinecyanine Dyes. 32
7.4.6. Higher Vinylogs of Cyanine Dyes . 33
7.4.7. Cyanine Dyes with Rings in the Methine Chain 33
7.4.8. Cyanine Dyes with Several Methine Chains 34
8. Anionic Methine Dyes (Oxonols) . 34
8.1. Monomethineoxonol Dyes 34
8.2. Trimethineoxonol Dyes and Higher Vinylogs 35
9. Methine Dyes with Betaine Structure 36
10. Neutral Methine Dyes 36
10.1. Enamineneutrocyanines 37
10.2. Aminoarylneutrocyanines 37
10.2.1. Aminoarylneutrocyanines with Open-Chain Terminal Groups 37
10.2.2. Aminoarylneutrocyanines with Cyclic Terminal Groups 41
10.3. Merocyanine Dyes 42
10.3.1. Merocyanine Dyes with Open-Chain Terminal Groups 42
10.3.2. Merocyanine Dyes with Cyclic Terminal Groups 43
11. References 46

1. Introduction

Methine dyes and polyene dyes are characterized by a chain of methine groups forming a system of conjugated double bonds. The methine chain in *polyene dyes* is even-numbered and the terminal groups do not affect the excitation of electrons in the dye. The most important group of polyene dyes, the carotenoids, bear aliphatic or alicyclic groups at the end of the methine chain, thus a very long conjugated chain is necessary to shift the excitation wavelength to the visible region spectrum. β-Carotene contains a chain of 22 methine groups. The dye absorbs at 450 and 478 nm.

Methine dyes are characterized by a chain of conjugated double bonds with an odd number of carbon atoms between the two terminal groups X and Y, which may be considered as being capable of charge exchange:

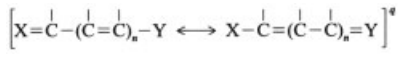

The substituents X and Y are groups of atoms, whereby one group acts as an electron acceptor and the other as an electron donor. Depending on the nature of the charge q, the dyes may be cationic, anionic, or neutral. Cationic methine dyes may be regarded as vinylogous amidinium salts:

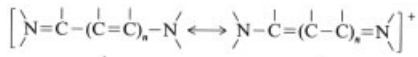

1 2

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_487

Similarly, anionic methine dyes may be classified as vinylogous carboxylate ions,

and neutral methine dyes as vinylogous carbonamides:

Methine dyes have an extended system of conjugated double bonds. A large number of resonance structures may therefore be formulated, of which (**1**) and (**2**) are the most important, i.e., the electron density is lowest on the nitrogen atoms. In all the following formulae only one resonance structure is shown. To indicate that a π-electron system with many possible resonance structures is involved, the total charge of the molecule is given outside a square bracket.

The electron density distribution in the ground state is generally different from that in the excited state, which means that light absorption by the dye molecule results in a redistribution of the electrons, but the total charge remains constant. Thus, the electronically excited state of a neutral methine dye molecule can be described by a betaine structure. The transition to the betaine structure may be associated with a large shift of the absorption wavelength if the chromophore system is formed by addition of energy with ring opening. Thermochromism refers to excitation by heat, and photochromism to excitation by light.

In the following classification of methine dyes only nitrogen and oxygen are considered as charge-carrying terminal atoms, although other heteroatoms are possible. The charge-carrying terminal atoms may be joined directly to the methine chain, may be linked via an aromatic group to the methine chain, or may form part of a heterocyclic ring.

There are several types of *cationic methine dyes.* Polymethine dyes that do not contain a heteroatom and in which the carbon atom is positively-charged are derived from azulenes.

Streptocyanine dyes contain both charge-carrying atoms at the ends of the methine chain:

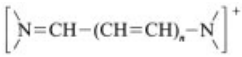

Aminoarylenamine dyes are obtained by insertion of an aryl ring:

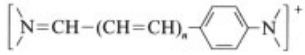

If the second charge-carrying terminal atom is separated from the methine chain by an aryl group, then di- and triarylmethane dyes and their vinylogs are obtained (→ Triarylmethane and Diarylmethane Dyes, Chap. 6.):

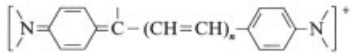

In hemicyanine dyes one of the charge-carrying terminal atoms is a member of a heterocyclic ring. Since the second nitrogen atom is directly coupled to the methine chain, these are also enamine dyes:

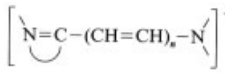

The phenylogous hemicyanine dyes are obtained by inserting an aryl group between the second nitrogen atom and the methine chain of the hemicyanine dyes. These dyes are often termed styryl dyes since they mostly contain an ethylene group linked to a phenyl group:

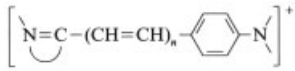

Finally, if the second charge-carrying terminal atom is also a member of a heterocyclic ring, then cyanine dyes are obtained:

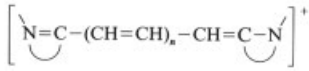

Carbinols are formed from cationic methine dyes by the action of alkalis. The hydroxyl group moves to the carbon atom with the lowest electron density. Carbinol ethers are similarly formed by reaction with alkoxides. If the dye contains a carboxyl group in a sterically favorable position, a lactone ring is formed under the action of alkalis with the elimination of water. All these carbinol derivatives are colorless compounds that are reconverted to the dyes by the action of protons.

Anionic methine dyes may be classified in a similar way. The simplest representative of the oxonols is the metal salt of glutaconaldehyde:

$$[O{=}CH{-}(CH{=}CH)_2{-}O]^-$$

The dioxytriphenylmethane dyes are phenylogous oxonols:

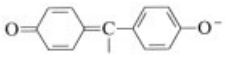

In the most important class of oxonol dyes both charge-carrying terminal atoms are bound to a ring system:

$$O{=}C{-}C{=}CH{-}(CH{=}CH)_n{-}C{=}C{-}O^-$$

Neutral methine dyes (neutrocyanine dyes) may be subdivided as follows. In enamineneutrocyanines the donor and acceptor groups are directly linked to the methine chain:

$$N{-}(CH{=}CH)_n{-}C{=}O \quad (\equiv N)$$

Aminoarylneutrocyanine dyes are obtained by inserting an aryl group between the amino group and the methine chain. Since the styrene structure is present, these dyes are also termed neutral styryl dyes:

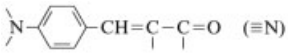

Finally, the electron-donating nitrogen can be a member of a heterocyclic ring, both in open-chain neutrocyanines

$$N{-}C{=}(CH{-}CH)_n{=}C{-}C{=}O \quad (\equiv N)$$

and in cyclic neutrocyanines (also called merocyanines):

$$N{-}C{=}(CH{-}CH)_n{=}C{-}C{=}O$$

The merocyanines also include the quinophthalones

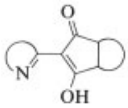

and the spiropyrans,

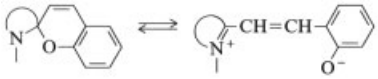

which undergo ring-opening on excitation by light to form hydroxyarylneutrocyanine dyes with the betaine structure.

The squaric acid dyes

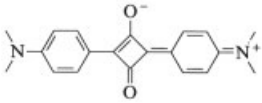

and the carbonylhomologous croconic acid (i.e., 4,5-dihydroxy-4-cyclopentene-1,2,3-trione) dyes always exist in the betaine form.

The methine dyes may be subdivided further according to the number of methine groups that are neither part of an aromatic ring nor part of a heterocyclic ring. A differentiation is made between mono-, di-, and trimethine dyes, etc. The lowest vinylogs of dimethine dyes do not contain a methine chain; the two terminal systems are directly coupled to one another. These dyes are termed zeromethine dyes or apocyanine dyes.

One or more members of the methine chain may be replaced by nitrogen, which is indicated by the term "aza". Depending on the number of nitrogen atoms introduced, monoaza- or diazadimethine dyes, etc., can be obtained. In 1977, 2 CH groups in the chain have also been replaced by phosphorus or arsenic to give the corresponding phosphamethine or arsamethine dyes. A comprehensive systematic classification of polymethine dyes can be found in [9].

2. History

The first methine dye was discovered in 1856 by G. Williams and was named cyanine on account of its brilliant blue color. In 1863 A. W. v. Hofmann established that the lepidine fraction in crude tar quinoline is essential for the formation of cyanine. S. Hoogewerf and W. A. Van Dorp prepared cyanine from pure quinoline and lepidine. In the same year W. Spalteholz discovered the violet isocyanines, which were formed by an analogous reaction from quinoline and quinaldine. In 1904 A. Miethe and G. Book correctly identified isocyanines as quinoline rings joined in the 2- and 4-positions by a

methine group. In 1906 W. König then proposed the correct structural formula. In 1912 A. Kaufmann and E. Vonderwahl discovered pseudocyanine, in which two quinoline rings are coupled in the 2,2-position by a methine bridge. The structure of this dye was verified by O. Fischer and G. Scheibe in 1920.

The chemical industry became interested in cyanines after their application as sensitizers in photography became known. In 1920 W. H. Mills and F. M. Hamer showed that pinacyanol had the structure of a vinylhomologous pseudocyanine, and it was therefore one of the first trimethinecyanine dyes to be synthesized. Trimethinecyanines were first prepared in 1922 by W. König using orthoformates and heterocycles with methyl or methylene groups capable of undergoing condensation. L. G. S. Brooker performed detailed investigations into the suitability of these dyes as sensitizers in silver-based photography. In 1912 W. König reported the preparation of styryl dyes; an important dye of this series is pinaflavol, which was discovered in 1920. Quinoline Yellow, discovered by E. Jacobson in 1882, was probably the first neutrocyanine; its structure, a zeromethineneutrocyanine, was established later. The first oxonol dye was Pyrazolone Yellow, discovered by E. Besthorn and L. Knorr in 1887.

With the development of cellulose acetate fibers, methine dyes became important for textile dyeing. Astraphloxin, the trimethine dye formed from 1,3,3-trimethyl-2-methyleneindoline and sodium formate, dyes cellulose acetate a brilliant pink. Progress was achieved in 1936 with the more lightfast styryl dyes synthesized from aminobenzaldehydes and 1,3,3-trimethyl-2-methyleneindoline. In 1938 dimethinecyanines were synthesized from the same base and indolaldehydes; and enamine dyes were synthesized from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde [systematic name: (1,3-dihydro-1,3,3-trimethyl-2 *H*-indol-2-ylidene)acetaldehyde] and aromatic amines. The Astrazon range for cellulose acetate printing, based on the above dyes, was introduced commerically in 1938 by IG Farben.

The importance of methine dyes for textile dyeing grew with the development of polyacrylonitrile fibers. It was found that a whole range of methine dyes could dye these fibers brilliantly and were extremely lightfast. Leading dye manufacturers throughout the world were subsequently engaged in intensive research and development programs. The diazadimethine-hemicyanine dyes (Geigy 1954), prepared by oxidative coupling of heterocyclic hydrazones with aromatic amines and described in the same year by S. Hünig became particularly important. Geigy produced these dyes by coupling of diazotized heterocyclic amines with aromatic amines and subsequent quaternization.

The apocyanine dyes synthesized from naphtholactam and aromatic amines (Bayer 1961) also became industrially important. Neutrocyanine dyes synthesized from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and alkyl cyanoacetates dye polyamide fibers in brilliant, lightfast shades (Bayer 1958). Neutral styryl dyes synthesized from aminobenzaldehydes and malononitrile were introduced commercially from 1960 onwards; they dye polyester fibers in brilliant yellow shades.

The range of dyes for dyeing paper (previously diphenylmethane, triphenylmethane, and rhodamine dyes) has been supplemented by methine dyes with yellow, orange, and red shades. Cationic neutrocyanine dyes synthesized from aminobenzaldehydes and cationic alkyl cyanoacetates (Bayer 1977); dimethine dyes synthesized from aminobenzaldehydes and quaternized γ-picoline (BASF, Bayer 1979); and enamine dyes synthesized from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and aromatic diamines and trimerized by cyanuric chloride (Ciba-Geigy 1980) have become important.

Brilliant blue neutrocyanine dyes are obtained by reacting aminobenzaldehydes with 3-dicyanomethylene-benzothiophene dioxide (Sandoz 1978). A brilliant red, classified as a cyclic neutrocyanine dye and having a bislactone structure, was introduced commercially by ICI in 1984.

Methine dyes have also proved valuable in research. They have been used to investigate the relationship between chemical structure and light absorption, the reactivity of organic compounds, the transmission of light energy, and the formation of associates.

Photochromic compounds are obtained by reacting 1,3,3-trimethyl-2-methyleneindoline with derivatives of 2-hydroxybenzaldehyde (Fischer and Hirshberg 1952). They have been used experimentally in optical data storage, but

the fulgides, derivatives of succinic anhydride discovered by HELLER and DARCY in 1981, proved to be more suitable. Polymethine dyes absorbing in the IR region and betaines formed from squaric acid and aromatic amines or thiopyrylium salts are used for nonerasable optical data storage (Philips 1983) (→ Information Storage Materials, Chap. 2.4.6.).

3. Preparation

Reactive methyl, methylene, and carbonyl groups and their derivatives are the most important groups for constructing the methine chains. Heterocyclic methylene compounds, in particular 1,3,3-trimethyl-2-methyleneindoline, as well as aromatic amines can be converted to the corresponding aldehydes by means of the Vilsmeier reaction. The Vilsmeier reagent is the product formed by reacting phosphorus oxychloride with dimethylformamide. Instead of phosphorus oxychloride, thionyl chloride or phosgene may also be used:

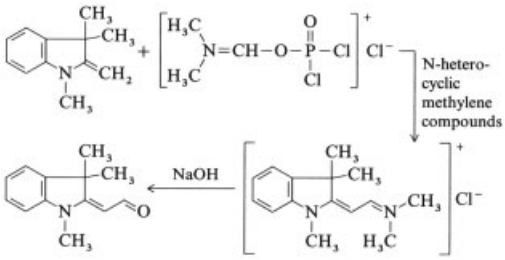

This heterocyclic α,β-unsaturated aldehyde reacts in glacial acetic acid with primary aromatic amines to form hemicyanine dyes, and with heterocyclic compounds containing activated methyl or methylene groups to form trimethinecyanines. Reaction with tertiary aromatic amines to form styryl dyes requires more aggressive conditions; phosphorus oxychloride is a suitable condensing agent. Styryl dyes can be prepared under milder conditions if aminoarylaldehydes are formed by the Vilsmeier reaction and then reacted with heterocyclic compounds containing a methyl or methylene group capable of undergoing condensation. Glacial acetic acid is a suitable condensing agent. In many cases it is not necessary to hydrolyze the intermediate formed in the Vilsmeier reaction, it can be condensed further to form the methine dyes. This one-pot reaction has economic advantages.

Preparation of styryl dyes by condensation of aminobenzaldehydes with 1,3,3-trimethyl-2-methyleneindoline in a paddle wheel dryer with addition of inorganic or organic acids has ecological advantages. Conversion is quantitative, with the result that a directly marketable dye is produced without any wastewater.

Suitable starting compounds for preparing symmetrical trimethinecyanines are formic acid derivatives (e.g., sodium formate, orthoformates, or chloroform). Methine dyes with longer chains are obtained by reaction with malonaldehyde, glutaconaldehyde, or their derivatives.

Monoazamethine dyes are prepared by using the corresponding nitroso compounds instead of the formyl compounds. Diazamethine dyes are prepared from heterocyclic azo dyes that are obtained by diazotizing heterocyclic amines and coupling with aromatic or heterocyclic coupling components. Diazadimethine dyes are obtained by quaternization at the heterocyclic nitrogen atom. Preparation of these dyes can be simplified by dissolving the heterocyclic amines with the aromatic or heterocyclic coupling components in an organic acid and reacting with sodium nitrite. The heterocyclic azo dyes are formed in a simultaneous diazotization and coupling reaction. If this solution is reacted with quaternization agents, solutions of diazadimethine dyes suitable for use in dyeing are obtained without the formation of wastewater.

Also in the case of diazadimethine dyes, total charge equalization (mesomerism) between the terminal groups is possible. This group of dyes is thus clearly distinguished from the cationic azo dyes (→Azo Dyes, Chap. 5.).

The Knoevenagel condensation of aromatic or heterocyclic aldehydes with enolate forming compounds (e.g., malononitrile, alkyl cyanoacetates, pyrazolones) is used industrially to prepare neutrocyanine dyes. If the aromatic aldehydes bear a hydroxy group in the 2-position, ring closure with the formation of coumarin takes place after condensation, and strongly fluorescent dyes are obtained.

The reaction products are generally sufficiently pure for dyeing. However, the dyes usually need to be further purified if they are to be used as sensitizers, chromatographic separation

on aluminum oxide is often necessary. The degree of purity must be checked by chromatographic analysis and absorption spectroscopy. As a more rigorous check, the dye is added to a silver halide emulsion which is then used to coat paper or a film. The sensitivity of the photographic layers is measured as a function of wavelength and light intensity. Very small amounts of impurities can be detected by comparison with standard photographs.

4. Properties

Polymethine dyes are generally stable to reducing agents but are decolorized irreversibly by oxidizing agents. Decolorization by acids is reversible; here a proton becomes attached to the free electron pair of the donor group. In alkaline media a carbinol is formed, OH^- being attached to the carbon atom of lowest electron density in the conjugated system, generally adjacent to the charge-carrying terminal atom. Styryl dyes with aromatic hydroxyl groups ortho to the methine chain form colorless spiro compounds with alkalis, accompanied by ring closure. In acidic media the methine dyes are reformed from the carbinol derivatives. Compounds of this type are important in the production of colorless copying papers.

In many cases the carbinols formed are unstable and decompose with cleavage of the methine chain. In streptocyanines the immonium group is particularly easily hydrolyzed to give corresponding aldehyde or ketone.

The stability of polymethine dyes decreases with an increase in the number of methine groups. Solutions are less stable than the pure substances. Mono- and trimethinecyanines are generally stable for several months in methanol, whereas solutions of undecamethine cyanines become unusable even after a few hours.

Lightfastness. Lightfast methine dyes for dyeing polyacrylonitrile fibers or plastics are the zero-, mono-, di-, and trimethine dyes. Lightfastness is frequently improved by introducing aza groups in the methine chain so that many diazadimethine dyes are very lightfast. Surprisingly, many tetraazapentamethine dyes are still lightfast when applied to polyacrylonitrile.

Light Absorption. Short-chain methine dyes absorb mainly in the shortwave region of the visible spectrum and are yellow, but some short-chain methine dyes with polynuclear heterocycles as terminal groups absorb in the longwave region of the spectrum. Zeromethine dyes synthesized from naphtholactam and aromatic amines and used to dye polyacrylonitrile are therefore blue. Extension of the methine chain by one vinyl group results in a 100 – 120 nm-bathochromic shift of the absorption maximum. Chain extension can lead to light absorption in the IR region (1500 nm).

All methine dyes have a high absorption intensity. The molar extinction coefficient reaches 30 000 even for the simplest dyes, and increases up to 250 000 for the longer-chain dyes. The absorption curve is generally unsymmetrical; the rise on the longwave side is steeper, while a second band appears on the shortwave side of the maximum. The extinction coefficient is greatest with symmetrical molecules. Steric hindrance interferes with the planar orientation of the conjugated systen. The half band width of the longwave band is thus increased and the extinction coefficient at the absorption maximum is generally reduced.

5. Uses

In the textile industry polymethine dyes are mostly used to dye polyacrylonitrile. The fibers are dyed in all processing stages such as loose stock, tow, yarn, and fabric. Dye mixtures are used to dye black. Mixtures of hydrazone, azo, diazacyanine, and naphtholactam dyes are of practical importance [10]. Styryl, triphenylmethane [11], and oxazine dyes [12] are also used. In order to reduce dust formation, liquid formulations are employed increasingly [13]. In the case of styryl dyes the preparation can be combined with the formation of concentrated dye solutions in a single stage [14]. Porous polyacrylonitrile fibers (Dunova, Bayer) are also dyed with polycationic dyes [15]. Level dyeing on polyacrylonitrile is obtained with low-molecular mass migrating dyes. Hydrazone, diazacyanine, and oxazine dyes are used [16]. The leveling ability of commercial polyacrylonitrile dyes can be improved by adding migrating dyes [17]. In the dry spinning process poly-

acrylonitrile is spun from an organic solvent, preferably dimethylformamide. Cationic dyes are added to these solutions to produce spun-dyed fibers. Bleeding into the drawing bath can be prevented by using special anions, e.g., nitrates [18], tetrafluoroborates [19], iodides [20], and rhodanides [21]. Styryl dye solutions suitable for spin dyeing can be prepared by reacting 4-aminobenzaldehydes with 1,3,3-trimethyl-2-methyleneindoline and toluene-4-sulfonic acid in dimethylformamide [22]. Concentrated solutions of cationic methine dyes are used on a large scale to dye polyacrylonitrile fibers during spinning in the gel state.

The dye formulations (all registered trademarks) of the most important producers of polyacrylonitrile dyes are as follows:

Aizencathilon	Hodogaya
Asahiacryl	Asahi
Astrazon	Bayer
Atacryl	Atlantic
Basacryl	BASF
Diacryl	Mitsubishi
Kayacryl	Kayaku Nippon
Maxilon	Ciba-Geigy
Nichilon	Nissei Kasei
Panacryl	Holliday
Sandocryl	Sandoz
Sumiacryl	Sumitomo
Synacril	ICI
Yoracryl	Yorkshire Chem.
Viocryl	Viochrom
Yation	Yamada

A whole range of methine dyes also dye acid-modified polyester fibers very lightfast. Black-dyeing can again be produced with dye mixtures. The same types of dye are used as for polyacrylonitrile [23]. Acid-modified polyamide is dyed using hemicyanine dyes to give lightfast yellow shades. Methine dyes containing triazine groups are used to dye paper [24].

Neutrocyanine dyes are suitable for dyeing polyamide and polyester fibers as well as for dyeing thermoplasts such as polystyrene and poly-(methyl methacrylate). On account of their low molecular mass, aminoarylneutrocyanines are used for sublimation thermotransfer printing on polyester fabrics, and especially as the yellow component for reproducing electronically recorded color photographs [25].

Some monomethinecyanines and their aza derivatives are used as optical brighteners, because they are colorless and fluoresce strongly. Some methine dyes are used for histological staining. Various cyanine dyes, e.g., benzothiazolepentamethinecyanine, act as antifilarial agents and anthelmintics, although colorless substances are preferred. In analytical chemistry anionic cyanines are also used to detect alkaline earth metals. Polymethine dyes are particularly important as spectral (optical) sensitizers in photography (→Photography).

6. Economic Aspects

In 1988, 26 000 t of cationic dyes with a value of 560×10^6 DM ($\$310\times10^6$) were required worldwide to dye 2.5×10^6 t of acrylic fibers. A polyacrylonitrile fiber consumption of 2.75×10^6 t is anticipated for 1993; the dye requirement will then amount to 27 000 t with a value of 610×10^6 DM ($\$400\times10^6$). The most important manufacturers of these dyes are Bayer and Ciba-Geigy. The requirement may be broken down according to dye shades: yellow 7300 t, orange 1300 t, red 5300 t, violet 1000 t, blue 7000 t, green 1000 t, black 3100 t. Cationic azo dyes represent a significant proportion of red dyes. Malachite Green continues to be an important dye for green and black shades. The remainder is covered by the methine dyes.

7. Cationic Methine Dyes

The most important group of methine dyes are the cationic compounds, especially those in which nitrogen is the cationic charge carrier.

7.1. Streptocyanine Dyes

In streptocyanine dyes both ends of the methine chain are joined directly to nitrogen atoms, a double enamine structure is thus present. The dyes are extremely susceptible to hydrolysis, particularly if they contain secondary nitrogen atoms. Stable dyes are obtained if the nitrogen is part of a heterocyclic ring system.

7.1.1. Trimethinestreptocyanine Dyes

Streptocyanine dyes with three methine groups have a coloring strength sufficient for practical

use if the nitrogen is part of a ring and if the aromatic ring bears an alkoxy group in the 4-position relative to the nitrogen atom.

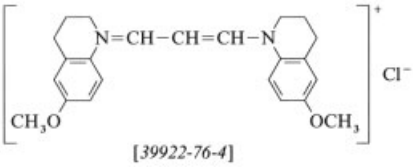

[39922-76-4]

Streptocyanine dyes are brilliant yellow dyes that dye polyacrylonitrile and acid-modified polyamide fibers with outstanding lightfastness [26]. Unsymmetrical trimethinestreptocyanine dyes are obtained by reaction of partially hydrogenated heterocycles with propyn-1-al and addition of the second heterocyclic component at the triple bond [27].

7.1.2. Pentamethinestreptocyanine Dyes and Higher Vinylogs

The pentamethinestreptocyanine dyes are formed by reaction of bromine cyanide, chlorine cyanide, or 2,4-dinitrochlorobenzene with pyridine and cleavage of the resultant compounds with primary or secondary amines. The tetraacetal of glutaconaldehyde is also used to form the methine chain [28]. The dyes can be used to dye paper. The greenish-yellow dyes prepared by reaction with *N*-ethyl- or *N*-methylaniline [29] were marketed as Astra Yellow G and R, respectively:

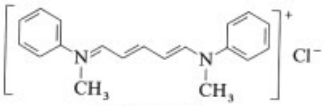

[13984-07-1]

Pentamethinestreptocyanine dyes synthesized from primary aromatic amines undergo ring closure on heating with the elimination of aniline, to form the colorless phenylpyridinium chloride. They are used as heat-bleachable antihalo dyes in photothermographic processes [30]. Ring closure involving both nitrogen atoms produces a considerable bathochromic shift. Rosol Scarlet is obtained with 2-methyl-2,3-dihydroindole as the amine component [31].

To prepare streptocyanine dyes containing an acyloxy group in the methine chain, vinylogs of furfural are used as starting material; the furan ring is cleaved with hydrochloric acid [32].

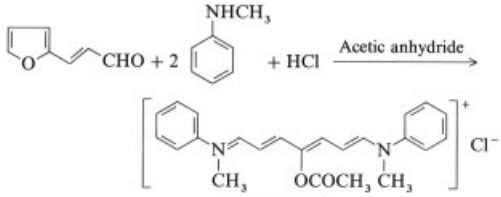

[126795-20-8]

Streptocyanine dyes are used to synthesize long-chain cyanine dyes.

7.1.3. Aminoarylenamine Dyes

Aminoarylenamine dyes are phenylogous streptocyanines, and are formed by condensing aromatic aminoaldehydes with aromatic amines. If the aromatic amine bears a substituent with electron-donor properties, the enamine dyes are sufficiently stable to be adsorbed on polyacrylonitrile without hydrolytic decomposition. The yellow dye

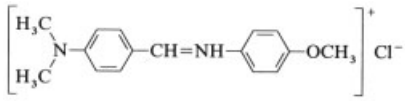

[85291-90-3]

is very lightfast [33]. The dye bases of these compounds are colorless or only slightly colored, and are used in carbonless copying paper. Solutions of the dye bases are encapsulated in microcapsules made of gelatin or polycondensation products and coated onto paper. If this is brought into contact with paper bearing a coating of a compound that acts as an electron acceptor (e.g., attapulgite) and the capsules are destroyed by pressure, the color bases are protonated to the enamine dyes, thereby producing yellow or orange dyeings of high intensity. The following compound

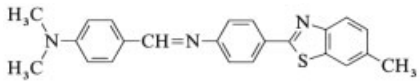

is particularly suitable in this respect and is also used for the preparation of black copying mixtures [34].

The color bases can also be sublimated image wise from a transfer ribbon onto an acid-developing receiver sheet. They then form the yellow component in color transfer prints [35].

7.2. Hemicyanine Dyes

In hemicyanine dyes one of the charge-carrying terminal atoms of the methine chain forms part of a heterocyclic ring. The lowest vinylogous hemicyanine dyes (cyclomethine dyes) do not contain a methine chain and the donor and acceptor groups are in the same ring system.

7.2.1. Cyclomethine Dyes

Cyclomethine dyes can be prepared by condensing N-heterocyclic carbonyl compounds with secondary amines. Intensely colored products are obtained by condensing 3-phenylpyrazoline with compounds such as *N*-ethylnaphtholactam:

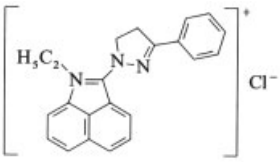

[34708-09-3]

This compound dyes polyacrylonitrile a brilliant, lightfast yellow. Orange-red dyes are obtained by condensing acridone with secondary amines [36].

7.2.2. Dimethinehemicyanine Dyes

Dimethinehemicyanine dyes and their aza analogues are important for dyeing polyacrylonitrile. Most of the dyes that color polyacrylonitrile in greenish to reddish-yellow shades belong to this class.

Enamine Dyes. Enamine dyes are obtained by condensation of heterocyclic methylene-ω-aldehydes with aromatic amines in an acid medium. Technically important dyes contain 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde as aldehyde component [37]. *C.I. Basic Yellow 11*, the condensation product formed with 2,4-dimethoxyaniline, is of particular importance:

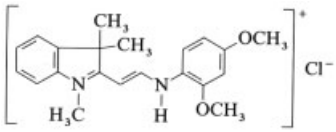

[4208-80-4]

This compound dyes polyacrylonitrile a lightfast, brilliant, greenish-yellow shade. In 1989, 900 t of Basic Yellow 11 were produced worldwide.

Preparation. A solution of 124 kg of 2,4-dimethoxyaniline in 1000 kg of 20 % sulfuric acid is added to a solution of 158 kg of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde in 820 kg of 20 % sulfuric acid. The mixture is stirred at room temperature until the formation of the dye is complete. The yellow crystals are pressed, dissolved, and allowed to crystallize from water, and the purified dye is dried. A yield of 217 kg is obtained, corresponding to 720 kg commercial grade.

Commercial Products. C.I. Basic Yellow 11 is included in practically all important dye ranges for dyeing polyacrylonitrile fibers. Commercial products include Astrazon Yellow 3GL, Maxilon Yellow 3GL, Basacryl Yellow X-3GL, Atacryl Yellow 4G, Yoracryl Yellow 4G, Aizencathilon Yellow 3GLH, Asahiacryl Yellow 3GL, Kayacryl Yellow 2GL, Sumiacryl Yellow 3G, Yation Yellow 3GS, Viocryl Yellow 3GS.

C.I. Basic Yellow 13 is obtained in a similar manner to Basic Yellow 11 by condensation of 4-anisidine with the ω-aldehyde. It dyes polyacrylonitrile a greener shade. In 1989, 150 t of this dye were produced worldwide.

Commercial Products. Astrazon Yellow 8GL, Basacryl Yellow X-4GL, Maxilon Brilliant Yellow 5GL, Panacryl Yellow 8GL, Sandocryl Brilliant Yellow B-6GL, Sumiacryl Brilliant Yellow 5G, Synacril Fast Yellow 8G, Nichilon Yellow 5GL.

The following amine components have also been proposed for condensation with methyleneindoline-ω-aldehydes: 4-aminophenylglycolether [38], 4-aminoacetanilide

[39], and 2-(4-aminophenyl)benzoxazole, -benzothiazole, -benzimidazole and -benzotriazole [40]. Diamines react with methyleneindoline-ω-aldehydes in a molar ratio of 2 : 1 to form yellow dyes suitable for dyeing paper. Suitable components are bis-4-aminophenylether, 4,4'-diaminodiphenylsulfone [41], or the 4-aminophthalic anhydride dimerized by aliphatic diamines [42]. Cyanuric chloride can be reacted twice or three times with 4-aminoacetanilide and the acetyl group can then be split off by acid hydrolysis. Reaction of the product with an equivalent amount of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde yields yellow dyes with a high affinity for paper [43]. Commercial products include Duasyn Direct Yellow KG liq. (Hoechst) and Pergasol Yellow FG liq. (Ciba-Geigy). The tetrafluoroborates of these dyes are suitable for spin dyeing polyacrylonitrile [44].

The low molecular mass bases of enamine dyes can be sublimed by means of a thermohead from a transfer ribbon onto a receiving material and serve as yellow components for producing color copies in sublimation transfer printing [45].

Vinylogous enamine dyes are obtained by reaction of aromatic amines with mucochloric acid (2,3-dichloro-4-oxo-2-butanoic acid), followed by condensation with quaternary quinaldine. After methylation of the enamine nitrogen with dimethylsulfate a dye is obtained that dyes bleached sulfite pulp red [46]:

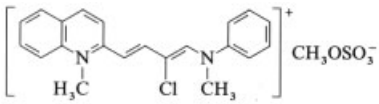

Enhydrazine dyes, synthesized from methyleneindoline-ω-aldehydes and the monoalkyl hydrazones of aromatic aldehydes, fluoresce in solution and on the fiber [47].

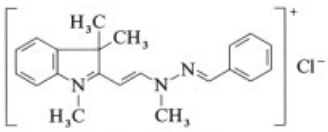

[56190-67-1]

Dyes prepared by condensation of methyleneindoline-ω-aldehydes with cyclic amino compounds have high lightfastness and resistance to hydrolysis [48].

C.I. Basic Yellow 21 is synthesized from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and 2-methyl-2,3-dihydroindoline. It dyes polyacrylonitrile in brilliant, greenish-yellow shades and is very lightfast:

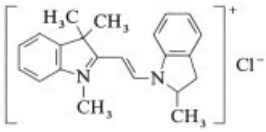

[6359-50-8]

Commercial Products. Astrazon Yellow 7GLL, Maxilon Yellow 6GL, Panacryl Yellow 7GLL, Diacryl Brilliant Yellow 7GL-N, Aizencathilon Yellow 7G-DP, Estrol Yellow 7GL-SD. In 1989 worldwide production was 400 t.

The shade is changed to golden-yellow by introducing a phenylazo group into C.I. Basic Yellow 21. Alkoxy substitution of the cyclic amine component in the 4-position to the nitrogen increases the coloring strength considerably [49]. Dyes of this type include Astrazon Yellow 8GSL and Sevron Yellow 8GMF:

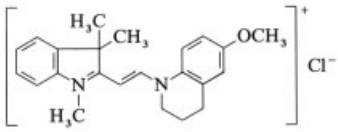

[27326-17-6]

Suitable N-heterocycles also include tetrahydrobenzoxazine [50] and tetrahydroquinoxaline. The latter produces a marked bathochromic shift. Sevron Yellow CP dyes polyacrylonitrile a neutral yellow shade [51]:

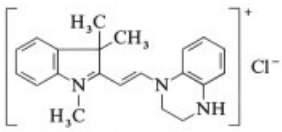

[28766-22-5]

A brilliant greenish-yellow dye is obtained by condensing methyleneindoline-ω-aldehydes with 3-phenylpyrazoline [52]:

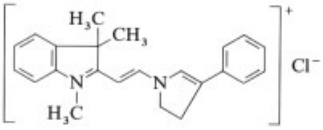

[31612-53-0]

If methyleneindoline-ω-aldehydes are condensed with cyclic amines substituted by an amino group in the aromatic ring, dyes are obtained that can be dimerized either by reaction with cyanuric chloride in a molar ratio of 2 : 1 or with phosgene. On account of their high substantivity these dyes are suitable for dyeing bleached sulfite pulp [53].

Instead of methyleneindoline-ω-aldehydes, the corresponding compounds derived from 2-methyl-benzothiazole have also been used. The resultant dyes are taken up by polyacrylonitrile quickly and are extremely lightfast [54].

Hydrazone Dyes. If the diazonium salts of aromatic amines are coupled with the methylene derivatives of N-heterocycles, hydrazone dyes (i.e., the monoaza derivatives of the enamine dyes) are obtained. Paper and leather can be dyed in yellow to red shades with these dyes [55]. They are moderately lightfast on polyacrylonitrile. If the diazo components contain substituents in the 2-position relative to the amino group that can form a hydrogen bond (e.g., 2-nitroaniline [56] or 1-aminoanthraquinone [57]), lightfastness is improved considerably:

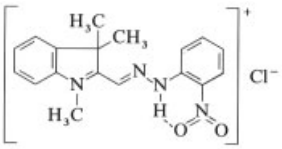

[27564-03-0]

Lightfastness can also be improved by a hydrogen-bond-forming substituent (e.g., a cyano group) in the coupling component [58]:

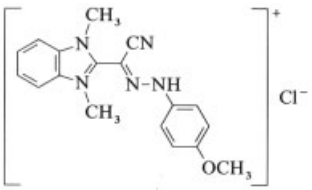

Such dye salts are converted to azo bases by alkali. Alkylating agents attack the nitrogen atom of the azo bond, forming alkylarylhydrazone dyes which dye polyacrylonitrile in lightfast, yellow to orange shades [59].

Diazotization, coupling, and methylation may be combined in a one-pot process by dissolving the aromatic amine and the heterocyclic methylene base in an organic acid and effecting the simultaneous diazotization and coupling by addition of sodium nitrite. The dye base is released by addition of alkali and is methylated with dimethyl sulfate to form the alkylarylhydrazone dye [60].

Preparation. 123 kg of *p*-anisidine is added to a mixture of 400 kg of water and 120 kg of glacial acetic acid under a nitrogen atmosphere. The reaction mixture is then cooled to 10 °C and 173 kg of 1,3,3-trimethyl-2-methyleneindoline added. A solution of 70 kg of sodium nitrite in 90 kg of water is added within 1 h at 10 °C. The temperature is then allowed to rise to 15 °C within 1 h and to 19 °C within the next 50 min. Cooling is then stopped. The suspension heats up as a result of the exothermic reaction, and after 20–30 min a temperature of 30–35 °C is reached. Beginning at 33 °C 50 kg of chlorobenzene is added, whereby the temperature rises to 61–62 °C. Conversion is complete after 30 min and the content of 1,3,3-trimethyl-2-methyleneindoline and *p*-anisidine is less than 0.1 %. Chlorobenzene (450 kg) is now added to the suspension and heated to 80 °C, the stirrer is switched off, and the aqueous phase is separated. The organic phase is cooled to 40 °C, 6 kg of magnesium oxide is added, 134 kg of dimethyl sulfate is added dropwise at 40 °C within 1 h followed by 18 kg of water within the next hour. The reaction mixture is then stirred for 6 h at 40 °C, a further 20 kg of dimethyl sulfate being added after 2 h. At this stage the content of unquaternized dye in the suspension is less than 0.03 %. After adding 200 L of water the pH value is adjusted to 2.5 with concentrated hydrochloric acid and the solvent is distilled off with steam. 1100 L of dye solution remain, which is salted out with 55 kg of sodium chloride, if necessary after clarification with activated charcoal at 50 °C. The dye suspension is then stirred for several hours at room temperature, and the dye is separated in a filter press, washed with 750 L of 5 % sodium

chloride solution, and dried. A yield of 440 kg is obtained, corresponding to 2000 kg commercial grade. Basic Yellow 28 dyes polyacrylonitrile golden yellow with very good lightfastness:

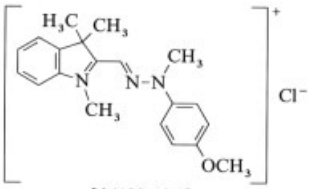

[*36528-10-6*]

Industrially important hydrazone dyes are:

1) Basic Yellow 28 (1988 production 3300 t worldwide): Astrazon Golden Yellow GL, Basacryl Golden Yellow X-GFL, Diacryl Golden Yellow GLN, Kayacryl Golden Yellow GL, Atacryl Yellow 2RL, Estrol Golden Yellow E-GL, Maxilon Golden Yellow GL, Panacryl Golden Yellow GL, Yation Yellow GLS.
2) Basic Yellow 29 [*38151-74-5*] (1989 production 200 t worldwide): Astrazon Yellow GRL, Atacryl Yellow GRL, Yation Yellow HS, Yracryl Yellow 2GL.
3) Basic Yellow 51 [*55850-01-6*] (1988 production 150 t worldwide): Astrazon Yellow 5 GL, Basacryl Yellow X-2GLN, Diacryl Yellow 3GN, Sumiacryl Yellow N-4G.
4) Basic Yellow 66: Diacryl Yellow 2GL-N.
5) Basic Yellow 70: Diacryl Yellow 3RL-F.
6) Basic Yellow 51, being a nitrate, is particularly suitable for preparing gravure and flexographic inks, and is marketed under the name Flexo Yellow 118 (BASF) [61].

Alkylation to form the cationic hydrazone dyes can also take place continuously in a two-phase system [62]. Alkylation in a purely aqueous medium is problematic on account of the hydrolysis of the dimethyl sulfate. Arylsulfonates are used in this case [63]. If special conditions are employed, the reaction can also be carried out with dimethyl sulfate using very small amounts of water [64] with the addition of urea [65] or sodium or magnesium sulfate [66].

Hydrazone dyes with an additional external cationic charge are suitable for use in writing inks, stamp pad inks and ball-point pen pastes [67]. Hydrazone dyes with acylamino groups can be hydrolyzed to compounds with a free amino group which dye polyacrylonitrile in red shades [68]. These compounds can be dimerized by reaction with phosgene, dicarboxylic acid chlorides, or cyanuric chloride in a molar ratio of 2 : 1. The dimerized compounds have a high affinity for cellulose and are suitable for dyeing bleached sulfite pulp [69]. The compounds with a free amino group can also be diazotized and reacted with aromatic or heterocyclic coupling components to form azohydrazone dyes [70].

In addition to the methyleneindolines, 4-methylene derivatives of 1,3-dialkylpyrimidin-2-one have also been used to prepare hydrazone dyes [59], [71]:

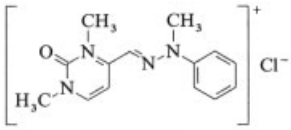

[*73297-08-2*]

A dye of this series is Basic Yellow 49, Basacryl Yellow X-RL.

Violet and blue hydrazone dyes are obtained with a methylene base derived from naphtholactam [72]:

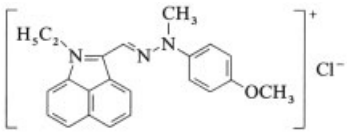

Heterocyclically substituted acetic acids (e.g., benzothiazolyl, benzimidazolyl [73], and pyrimidyl acetic acids [74]) are also suitable as coupling components for preparing cationic hydrazone dyes. In these compounds the methyl group in the 2-position is additionally activated. The carboxyl group is split off after coupling and the azo base is methylated.

Alkylarylhydrazone dyes can also be produced by reacting suitable heterocyclic aldehydes with arylhydrazines. Subsequent alkylation of the heterocyclic nitrogen atom produces dyes that dye polyacrylonitrile extremely lightfast [75]:

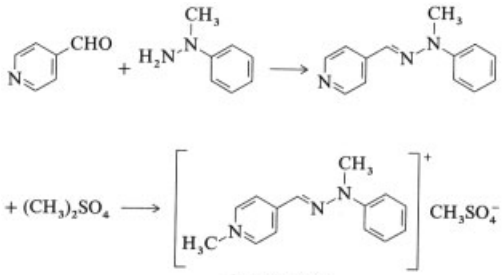

[42476-20-0]

The hydrazone dyes of the pyridine, quinoline, and acridine series can be prepared by this route. Hydrazone dyes that are derived from pyridine exhibit a particularly good leveling ability when dyeing polyacrylonitrile [76]. Dyes of this series include Maxilon Yellow M-3RL and Maxilon Yellow M-4GL.

Hydrazone dyes of the pyridine series can be alkylated in an aqueous medium [77]. Pyridinealdehydephenylhydrazones are also obtained by reaction of aminomethylpyridine with a ketone, rearrangement to the azomethine, and subsequent reaction with phenylhydrazine [78].

If N-amino nitrogen heterocycles are used as the hydrazine component, cyclic hydrazone dyes are obtained whose absorption maximum is shifted bathochromically with respect to the open-chain hydrazone dyes [79]. The dye shown below dyes polyacrylonitrile fibers a clear bluish-red shade:

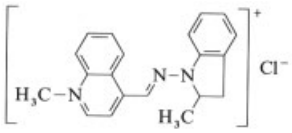

[72828-74-1]

Cyclic hydrazone dyes can also be prepared by intramolecular quaternization if diazo components are used that contain a leaving group in a sterically favorable position. The azo base formed from 2-(2-bromoethoxy)aniline closes the dihydrobenzoxazine ring with the elimination of a bromide ion. It dyes polyacrylonitrile a golden-yellow shade [80].

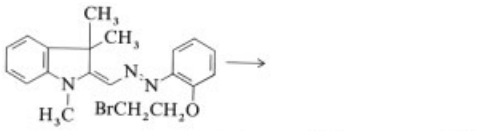

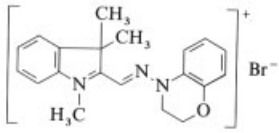

[69563-60-6]

Triazene Dyes (Diazadimethinehemicyanine dyes). If diazonium salts of N-heterocyclic amines are coupled to N-alkylanilines, e.g., diazotized 2-amino-6-methoxybenzothiazole to *N*-methyl-4-nitroaniline, a triazene is obtained. A cationic triazene dye is formed by subsequent N-alkylation, and dyes polyacrylonitrile a very lightfast yellow shade [81].

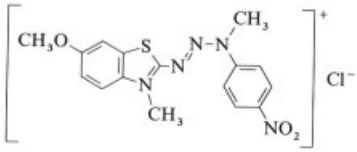

[14970-38-8]

Dyes from this range include Maxilon Yellow 4GL and 2RL. Worldwide production in Japan in 1989 was 200 t. Using 2-amino-5-dimethylamino-1,3,4-thiadiazole as diazo component and *N*-methylaniline as coupling component, a triazene dye is obtained that dyes polyacrylonitrile yellow and has a very good migration ability [82].

Of the enamine and hydrazone dyes, those with donor groups in the aryl system are particularly lightfast, whereas the lightfastness is reduced on the introduction of nitro groups. In contrast, triazene dyes with a nitro group attached to the aryl group are substantially more lightfast than those with a methoxy group attached to the aryl group.

Triazene dyes in which the terminal nitrogen atom is cyclically bound are obtained by coupling heterocyclic diazonium salts with partially hydrogenated N-heterocycles. In the following example alkylation was subsequently performed with ethylene oxide [83]:

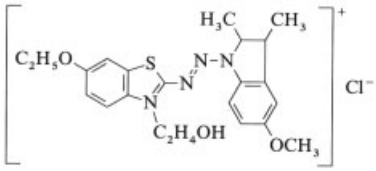

7.2.3. Higher Vinylogs of Hemicyanine Dyes

Streptocyanine dyes are semihydrolyzed under mild conditions to give ω-aminopolyenealdehydes. If these aldehydes are condensed with heterocyclic quaternary salts containing a reactive methyl group, the higher vinylogous hemicyanine dyes are obtained. Trimethinestreptocyanines yield tetramethinehemicyanine dyes, while pentamethinestreptocyanines yield hexamethinehemicyanine dyes [84], e.g.,:

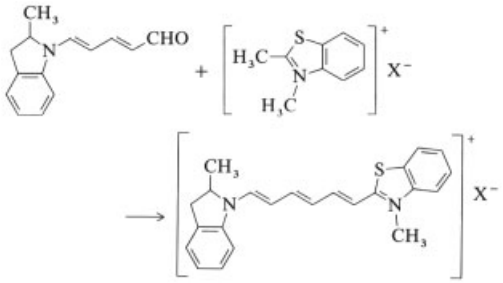

These dyes are used to sensitize photographic emulsions.

7.3. Phenylogous Hemicyanine Dyes

Phenylogous hemicyanine dyes are obtained by introducing a phenyl group between the charge-carrying nitrogen atom and the remaining dye system. The zeromethine (apocyanine) dyes are thus derived from the cyclomethine dyes. The styryl dyes are derived from the dimethinehemicyanine dyes by introducing a phenyl group between the charge-carrying nitrogen atom and the methine chain.

7.3.1. Zeromethine (Apocyanine) Dyes

Some representatives of the apocyanine dyes have been known for a long time, e.g., thioflavine T [*2390-54-7*]:

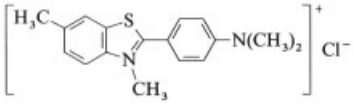

This dye is obtained by heating *p*-toluidine with sulfur and subsequent alkylation with methyl or ethyl sulfate. The dye colors silk and tanned cotton a greenish-yellow shade. Thioflavine S [*1326-12-1*] is obtained by sulfonation [85].

Apocyanine dyes of the acridine series are obtained by reacting acridinium iodide [86] or acridinium chloride [87] with tertiary aromatic amines. The dyes can also be obtained by ipso-substitution of acridinium iodide with 4-haloarylamines in dimethyl sulfoxide at room temperature [88].

The conversion of 2-methylbenzothiazole according to the Vilsmeier reaction leads to benzothiazolylmalonaldehyde. By condensation with 3-amino-*N*,*N*-dimethylaniline and subsequent methylation with dimethyl sulfate, a cationic dye is obtained that dyes paper a clear reddish-yellow shade:

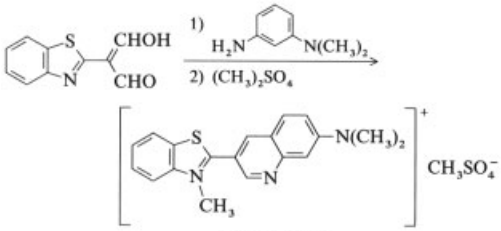

[*69157-68-2*]

This is a styryl dye whose dimethine chain has been incorporated into a quinoline ring system and is therefore an apocyanine dye. The reaction can also be applied to 2-methyl-benzoxazole, 2,3,3-trimethylindolenine, or quinaldine [89].

The most important method for preparing apocyanine dyes is the reaction of N-heterocycles containing 2- or 4-carbonyl groups with secondary or tertiary aromatic amines in the presence of dehydrating agents (e.g., phosphorus oxychloride or tin tetrachloride).

Naphtholactam is a particularly important condensation component for preparing apocyanine dyes. In 1989 ca. 400 t of naphtholactam dyes were produced worldwide. At first only the dyes unsubstituted at the lactam nitrogen were prepared and recommended for dyeing tanned

cotton and wool [90]. However, it was later found that the N-alkyl derivatives dye polyacrylonitrile in very lightfast violet and blue shades:

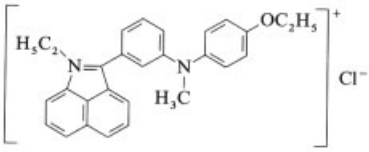

[35114-09-1]

The dyes are synthesized either from N-alkylnaphtholactams, obtained by intramolecular Friedel–Crafts reaction of the 1-alkylnaphthylamine carboxylic acid chlorides, or by subsequent alkylation of the dye bases obtained by condensation of naphtholactam with aromatic amines [91].

Secondary aromatic amines, in particular those that also bear a substituent in the 2-position to the amino group, condense with N-alkylnaphtholactams to form dyes that are particularly lightfast on polyacrylonitrile [92]. A marked bathochromic shift occurs if the aminoaryl nitrogen atom is part of a heterocyclic ring. By condensation of *N*-ethylnaphtholactam with 1-phenyl-3,5,5-trimethylpyrazoline a dye is obtained that dyes polyacrylonitrile a lightfast blue shade [93]:

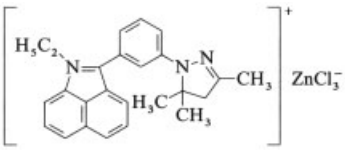

[40496-35-3]

Particularly lightfast dyes are obtained if 1-alkyl derivatives of 2,2,4-trimethyltetrahydroquinoline are used as amine components [94]:

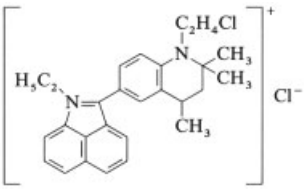

This dye colors polyacrylonitrile a clear blue shade.

Commercial Products. Astrazon Blue 5RL, FRR, F2RL, Diacryl Brilliant Blue H2RN, Diacryl Supra Brilliant Blue HBL, and Maxilon Blue TRL.

N-Alkylnaphtholactam can be condensed with formaldehyde to form a dinaphthylmethane derivative and then reacted with aromatic amines or indole derivatives. These dyes are suitable for spin dyeing polyacrylonitrile and for dyeing paper [95]. The dyes with an unsubstituted lactam nitrogen atom can be converted into dye bases and used in this form for sublimation transfer printing on polyacrylonitrile fabrics [96]. Use of the carbinols of naphtholactam dyes in heat- or pressure-sensitive copying or recording materials has been proposed [97]. If the lactam nitrogen bears a carbonamidoethyl group as substituent, intramolecular lactam ring closure occurs under the action of alkali. These compounds can also be used as color formers [98].

Dye bases of apocyanine dyes derived from pyridine which is unsubstituted on the nitrogen are color formers for heat- or pressure-sensitive copying paper and develop into yellow to orange shades [99]:

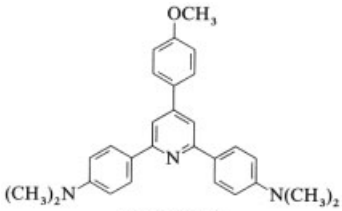

[84754-34-7]

A commercial product of this range is React Yellow (BASF).

Dye bases of apocyanine dyes are also used as organic photoconductors in electrophotography [100].

7.3.2. Phenylogous Dimethinehemicyanine Dyes

These dyes are used widely in photography and textile dyeing. The diazadimethinehemicyanine dyes are particularly important in the textile sector.

Styryl Dyes. Styryl dyes are produced by condensation of 4-aminobenzaldehydes with N-heterocycles bearing an activated methyl group

in the 2- or 4-position. The starting components are usually briefly heated in glacial acetic acid. The aminoaldehydes can also be condensed with quaternary heterocyclic salts in the presence of basic catalysts such as piperidine, diethylamine, or ammonia. The styryl dyes include pinaflavol [101], which is used as a sensitizer in silver-based photography:

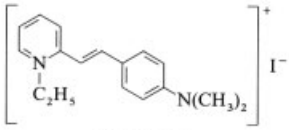

[3785-01-1]

Dyes containing nitro groups act as desensitizers (→Photography).

Styryl dyes for dyeing paper in brilliant shades are derived from quaternized γ-picoline. The dye obtained by reaction with 4-(*N*-ethyl-*N*-benzylamino)benzaldehyde has a brilliant orange color [102]:

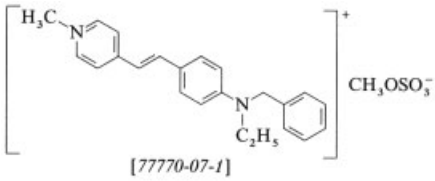

[77770-07-1]

The shade is shifted bathochromically by ring closure at the nitrogen atom of the amino group. Using 1-ethyl-2,2,4-trimethyl-6-formyltetrahydroquinoline, a dye is obtained that colors paper a brilliant bluish red [103]. The solubility and affinity of the dyes are increased if amino and hydroxyl groups are introduced into the quaternized pyridine substituents. For example, epichlorohydrin can be reacted with morpholine and then with 4-methylpyridine [104]. Primary amines can be reacted twice with epichlorohydrin and subsequently with 2 mol of 4-methylpyridine and two equivalents of aminobenzaldehyde to form dimeric styryl dyes [105]. Dimerization may also be effected via the aldehyde component if 1 mol of 1,4-bis(formylphenyl)piperazine is condensed with 2 mol of quaternized γ-picoline [106]. If dialdehydes are reacted with dimeric γ-picolinium salts, polymeric styryl dyes are obtained that are particularly suitable for dyeing paper [107]. Polymeric dyes are also obtained by intramolecular quaternization of the reaction product of 4-methylpyridine with 4-(*N*-ethyl-*N*-chloroethylaminobenzaldehyde in ethylene glycol with the addition of pyrrolidine at 140 °C [108]. Commercial products include Astra Brilliant Orange G and Basazol C-Orange 54L.

If quaternary 4-methylpyridine is condensed with 4-acetylaminobenzaldehyde and the acetyl group is then hydrolyzed, a styryl dye with a free amino group is obtained. Methineazo dyes can be prepared from the latter by diazotization and coupling. Resorcinol [109], acetoacetamides [110], and barbituric acid [111] have been used as coupling components. The dyes are all suitable for dyeing paper.

The most important styryl dyes for textile dyeing are obtained by condensation of 4-aminobenzaldehydes with 1,3,3-trimethyl-2-methyleneindoline and its derivatives. The dye synthesized from 4-dimethylaminobenzaldehyde and 1,3,3-trimethyl-2-methyleneindoline dyes polyacrylonitrile a brilliant red shade:

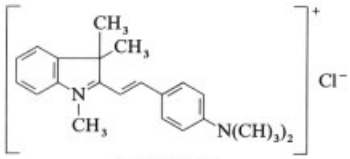

[54268-66-5]

Commercial products include Basacryl Brilliant Red G, Diacryl Supra Brilliant Pink R, and Tetradene Brilliant Pink FBB.

The corresponding dye (C.I. Basic Violet 16) obtained from 4-diethylaminobenzaldehyde dyes polyacrylonitrile a more intense bluish-red shade. In 1989, 700 t of the dye were produced worldwide.

Commercial products are Astrazon Red Violet 3R, Basacryl Brilliant Red BG, Diacryl Brilliant Red 5B-F, Panacryl Brilliant Red 2B, Tetradene Brilliant Red 2B, Viocryl Violet 3RSB, and Atacryl Red Violet 3RA.

In the Vilsmeier formylation of *N*-(β-hydroxyethyl)anilines the hydroxyl group is simultaneously replaced by chlorine [112]. The resulting aldehydes condense with 1,3,3-trimethyl-2-methyleneindoline to form very brilliant dyes [113]. With 4-(*N*-methyl-*N*-β-chloroethylamino)benzaldehyde a dye is obtained that is marketed as Astrazon Pink FG, Asahiacryl Pink FGL, Aizencathilon Pink FGH and Maxilon

Pink GA. In 1989, 100 t of this dye were produced worldwide. A methyl group in the 2-position to the methine chain produces a bluish-red color:

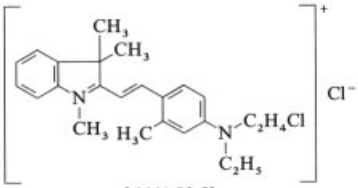

[6441-82-3]

Commercial products include Astrazon Red 6B, Sumiacryl Red 6B, and Aizencathilon Red 6BH.

Replacement of the chlorine substituent by a cyano substituent produces a marked hypsochromic shift [114]:

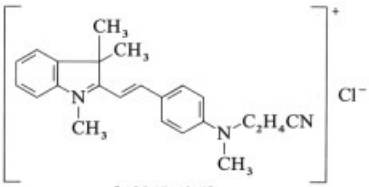

[12217-48-0]

This dye (C.I. Basic Red 14) is widely used to produce brilliant red dyeings on polyacrylonitrile; 500 t were produced worldwide in 1989. For this dye as well as the aforementioned styryl dyes from 1,3,3-trimethyl-2-methyleneindoline, a new and particularly environmentally friendly process has been developed in which no effluent occurs [115–117]. Since conversion exceeds 99 %, an immediately marketable dye is obtained.

Commercial Products. Astrazon Brilliant Red 4G, Basacryl Brilliant Red X-4G, Sevron Brilliant Red 4G, Sandocryl Brilliant Red B-4G, Sumiacryl Brilliant Red 4G, Aizencathilon Brilliant Red 4GH, Concordacryl Red 4G (Concorde Dyes, Taiwan), and Rifacationic Red 4G (Rifa, South Korea).

The cyanoethyl group can be hydrolyzed to the carbamoyl group by a strong acid. The resulting dyes are exhausted from the dyebath more slowly than compounds with an unsubstituted alkyl group [118]. The adsorption rate of the styryl dyes can also be reduced by introducing alkoxyalkyl groups at the aminoaryl nitrogen atom. Examples of such dyes include Aizencathilon Brilliant Pink BH and Aizencathilon Red 7 BNH. In 1984, 100 t of Aizencathilon Red 7 BNH were produced in Japan.

Lightfastness can be improved by introducing benzyl groups [119]:

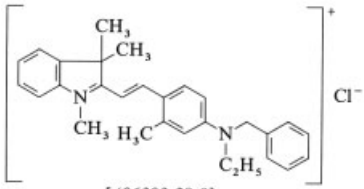

[106303-28-0]

Representatives of this series are Diacryl Supra Brilliant Red 2G and 3B as well as Aizencathilon Brilliant Scarlet RH.

The N-phenyl derivatives are even more lightfast [120]:

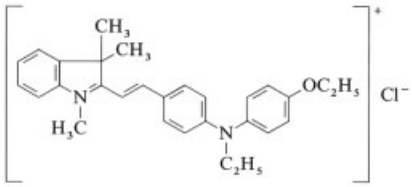

This range of dyes includes Astrazon Red Violet FRR and Astrazon Violet F3RL. They are used in electrophotography to make organic photoconductors [121]. Dyes derived from triphenylamine aldehyde are also suitable for the same application [122].

In addition to the unsubstituted 1,3,3-trimethyl-2-methyleneindoline, the 5-chloro derivative is often used as the heterocyclic condensation component [123]. Dyes of this series include Sandocryl Brilliant Red B-F and B-5B, and Sumiacryl Brilliant Pink F-B.

Styryl dyes with external amino groups (e.g., Kayacryl Brilliant Red 5G) are particularly suitable for dyeing polyacrylonitrile from a strongly acid medium [124]:

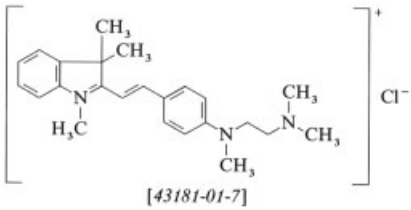

[43181-01-7]

The styryl dyes also include compounds in which the aminoaryl nitrogen atom is part of a carbazole (Astra Red 3G, see below) [125], phenoxazine [126], or tetrahydrobenzoxazine system [127].

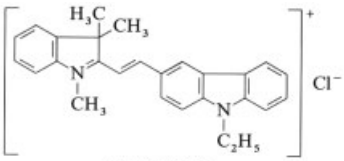

[72828-91-2]

Styryl dyes that contain the carbazole ring system are used in the production of photoconducting layers in electrophotography [128].

In addition to 1,3,3-trimethyl-2-methyleneindoline, a range of other heterocyclic compounds has also been used to prepare cationic styryl dyes. 1-Phenyl-2-methylene-3-methyl-1,2-dihydroquinoxaline [129] and 1,4-dimethyl-2-methylene-1,2,3,4-tetrahydroquinoxaline-3-one [130] condense with 4-aminobenzaldehydes to form styryl dyes that dye polyacrylonitrile a greenish-blue shade.

If a base obtained from *N*-alkylnaphtholactam by reaction with methyl magnesium iodide is used, the absorption range of the styryl dyes can be extended into the near IR region:

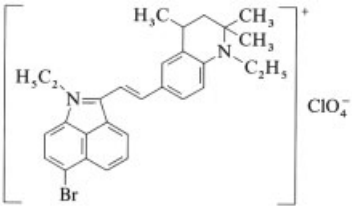

[111720-62-8]

This dye has an absorption maximum of 760 nm and is used in the production of discs suitable for information recording with the gallium arsenide semiconductor laser [131].

If 1,2-diaminobenzene is condensed with butyrolactone and the product quaternized, a benzimidazole derivative with a reactive methylene group is obtained. Condensation of this compound with 4-aminobenzaldehydes (which can also be dimerized via the nitrogen atom) produces styryl dyes which, on account of ring closure, can be classified as monomethinehemicyanine dyes. These dyes (e.g., Basazol C Yellow 50L) are used to dye paper [132]:

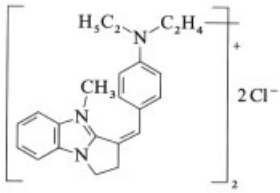

Nitrogen-free heterocycles (e.g., pyrylium salts) may also be condensed to form styryl dyes. Some cationic styryl dyes are suitable as active medium in dye lasers (→ Laser Dyes).

N-heterocycles with an activated methyl group can also react with aminobenzaldehydes to form dye bases without being quaternized. These compounds are widely used in electrophotography. Derivatives of quinoline [133], benzothiazole [134], and indolenine [135] have been described. A representative of these compounds is:

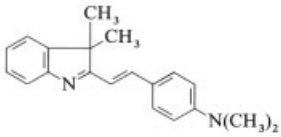

[34442-71-2]

This dye base can also be transferred onto polyacrylonitrile fabric by sublimation transfer printing [136]. The carbinols of styryl dyes are also suitable for transfer printing on polyacrylonitrile [137].

Carbinol Derivatives. If the alkyl group on the indolenine nitrogen atom bears a substituent (e.g., OH, COOH, or $CONH_2$) that is capable of forming ring-closed carbinol derivatives, then stability is substantially increased. The dye with the hydroxyethyl group on the indolenine nitrogen is stabilized in an alkaline medium to form an oxazolidine derivative:

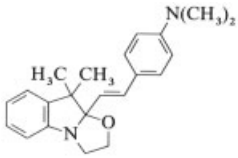

[59335-14-7]

When dissolved in propylene carbonate with the addition of tetrabutylammonium perchlorate as electrolyte, the compound is used in the production of electrochromic display elements [138]. A carboxymethyl group on the indolenine nitrogen atom closes the ring to form an oxazolone derivative. The compounds are used as color formers [139]. With a carbonamido group as substituent in the alkyl chain attached to the nitrogen, imidazolidinone derivatives [140] or hexahydropyrimidone derivatives [141] are obtained.

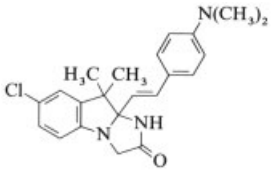

These compounds are more stable against ring opening by heat or protons and are therefore suitable as color formers for thermo- and pressure-copying processes.

Spiropyran Compounds. If 1,3,3-trimethyl-2-methyleneindoline is condensed with aminobenzaldehydes bearing a hydroxyl group in the 2-position to the aldehyde group, styryl dyes are obtained that are converted to colorless indolinospiropyran compounds on addition of alkali [142]. The compounds are suitable as color formers for producing colorless copying paper. In contact with acid inorganic materials (e.g., kaolin, bentonite, or attapulgite) they are cleaved into the open-chain colored form:

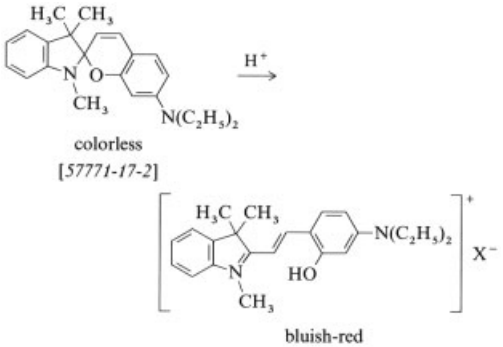

In spiropyran compounds derived from naphtholactam, the absorption of these color formers is shifted bathochromically to the limit of the IR spectrum [143]. Such color formers can also be prepared from aromatic 2-hydroxyaldehydes that do not contain an amino group. For example, indolinospironaphthopyran compounds that form red copies on acid clay are obtained by condensation of 1-carboxymethyl-2,3,3-trimethylindoleninium salts with 2-hydroxy-1-naphthaldehyde in the presence of triethylamine [144]:

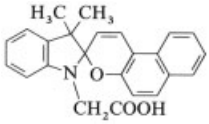

If the 2-hydroxyaldehydes also contain nitro groups, photochromic compounds are formed that are suitable for short-term data recording. Irradiation with UV light produces an intense color which disappears on heat treatment or during storage [145]:

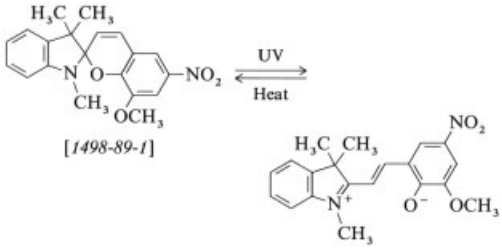

Storage stability of the colored form is essential for photochromic applications and can be improved either by introducing long-chain alkyl groups [146] or by incorporation into polymer systems [147].

A carboxyl group adjacent to the hydroxyl group stabilizes the open-chain form. Such compounds yield recording materials that exhibit inverse photochromism (colorless when irradiated, colored when not irradiated) [148].

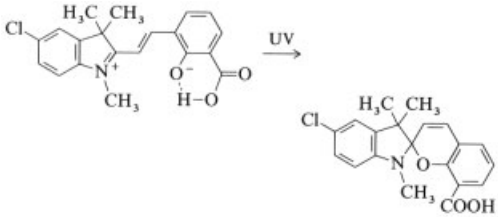

Instead of indolenine, derivatives of benzothiazole [149], quinoxaline [150], pyrazole [151],

benzothiopyran [152], benzoxazine [153], and pyrimidine [154], may also be used to prepare spiropyran compounds.

Condensation of suitably substituted benzopyrilium salts with 4-(diethylamino)salicylaldehyde, produces styryl dyes with a heterocyclic terminal system in which oxygen carries the cationic charge. The spirobipyrans are obtained from the latter by ring closure, for example:

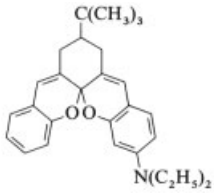

With protonating agents, a blue cation is formed by ring opening. The compounds are therefore suitable as color formers for pressure-sensitive recording materials [155]. A commercial product of this range is React Blue R (BASF).

Spirothiopyran compounds are obtained with derivatives of thiosalicylaldehyde as the starting compounds [156].

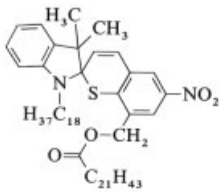

These compounds have outstanding photochromic properties and are used in the production of optical recording media. The indolenine ring system may also be substituted by amino [157] or alkoxy groups [158]. The absorption range of the open-ring compound is shifted to the near IR range ($\lambda_{max} = 780$ nm) by benzanellation. These compounds are suitable for information recording with the gallium arsenide semiconductor laser [159].

Coumarin Compounds. 7-Aminocoumarin compounds that bear a quaternized heterocycle in the 3-position may be regarded as ringclosed cationic styryl dyes. They are prepared by reacting 4-(dialkylamino)-2-hydroxybenzaldehydes with N-heterocyclic cyanomethyl compounds followed by alkylation [160]. They may also be prepared by condensation with the previously quaternized 2-cyanomethyl compounds [161].

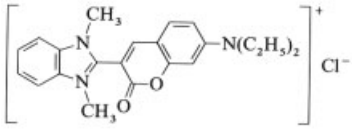

[29556-33-0]

The iminocoumarin compound formed as intermediate may be isolated as the chlorozincate [162]. If the latter is incorporated by condensation in a toluenesulfamide-paraformaldehyde-melamine melt, a fluorescent pigment is obtained [163]. Analogous fluorescent dyes can be prepared from the cyanomethyl compounds of 2-quinazolinone [164] and pyridine [165]. A dye of this series is Maxilon Brilliant Flavin 10GFF.

The analogous quinoline dyes also fluoresce and dye polyacrylonitrile in brilliant yellow and orange shades [166].

Phenylogous Monoazadimethinehemicyanine Dyes. By reacting 4-nitroso compounds of tertiary aromatic amines with the quaternary salts of N-heterocycles containing a reactive methyl group in the 2- or 4-position, dyes similiar to styryl dyes are obtained, the CH group being replaced by nitrogen [167].

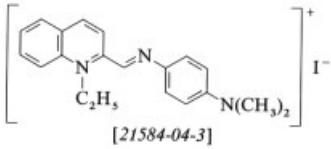

[21584-04-3]

These compounds are of no importance for dyeing because they are readily hydrolyzed at the CH = N group with the formation of the corresponding aldehydes, which are used as intermediates in the preparation of alkylarylhydrazone dyes [75], [79].

If the heterocyclic nitrogen bears a hydroxyethyl group, then an oxazolidine ring is formed in alkaline medium:

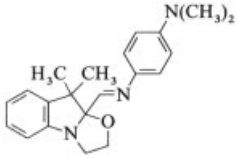

[59334-76-8]

Such compounds are suitable for multicolored electrochromic display elements [168].

Analogously to the 2-hydroxybenzaldehydes, condensation of the 2-hydroxynitroso compounds with 1,3,3-trimethyl-2-methyleneindoline yields monoazadimethine dyes that are converted to spiro compounds on the addition of alkali. They can be used as photochromic substances. Starting materials are nitrosophenols [169] and nitrosonaphthols [170].

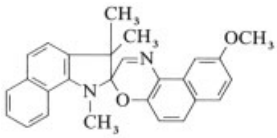

[119995-07-2]

Phenylogous Diazadimethinehemicyanine Dyes. If N-heterocyclic compounds with an amino group in the 2- or 4-position are diazotized and then coupled to aromatic amines, yellow to red azo dyes are obtained. If these azo dyes are reacted with alkylating agents, alkylation occurs at the heterocyclic nitrogen atom. The absorption maximum of the compounds thereby undergoes a considerable bathochromic shift and becomes more intense indicating delocalization of the cationic charge. The resultant dyes are dimethinehemicyanines, in which the two CH groups have been replaced by N.

Dyes with 2-amino-6-methoxybenzothiazole as diazo component are of industrial importance because their alkylated products dye polyacrylonitrile in clear very lightfast blue shades. With *N*,*N*-dimethylaniline as coupling component, C.I. Basic Blue 54 is obtained:

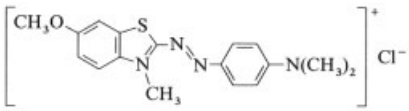

[29767-87-1]

Worldwide production in 1989 was 300 t. The dye with *N,N*-diethylaniline as coupling component is also produced industrially. Dimethyl and diethyl sulfates are used as alkylating agents [171].

Commercial Products. Maxilon Blue RBL, Basacryl Blue GL, Atacryl Blue GNA, Nabor Blue AF, Diacryl Supra Blue RBL, Aizencathilon Blue BRLH, Estrol Blue GSL-SD, Yation Blue M, and Yoracryl Blue BL.

With butylcyanoethylaniline a blue dye is obtained that produces very steam-fast dyeings on polyacrylonitrile [172]. The dyes with *N*-methyldiphenylamine as coupling component dye polyacrylonitrile in greenish-blue shades. Commercial products include Basacryl Blue GGL, Maxilon Blue GL, and Kayacryl Blue GL.

The dye shown below (C.I. Basic Blue 41)

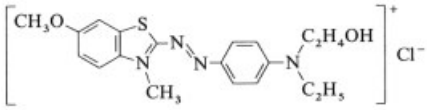

[26850-47-5]

is exhausted somewhat more slowly and can therefore be combined particularly well with other dyes. It is contained in most ranges for dyeing polyacrylonitrile. In 1988, 3000 t were produced worldwide.

Commercial Products. Maxilon Blue GRL, Sandocryl Blue B-RLE, Astrazon Blue FGGL, Basacryl Blue X-3GL, Synacril Blue G, Concordacryl Blue GRL, Nichilon Blue GRL, Panacryl Blue GRL, Viocryl Blue RLS, Yation Blue BS-S, Yoracryl Blue R-GL, Tetradene Fast Blue RL, Diacryl Supra Blue GRL.

The tetrachlorozincate of Basic Blue 41 can be prepared by adding zinc oxide or zinc carbonate as acid-binding agent in the methylation step and salting out the dye with sodium chloride [173]. Stable, highly concentrated solutions of the aforementioned dyes are produced by performing the methylation in mixtures of water and glycols, glycol ethers, or ethylenecyanohydrin [174].

Many diazadimethine dyes decompose during dyeing in a weakly acid to neutral medium, whereby accompanying fibers can be stained. This hydrolysis is suppressed by complexing anions such as benzoate, benzenesulfonate, or phenolate ions [175]. Dyes that are more stable to

hydrolysis are obtained if derivatives of indoline [176] or tetrahydroquinoline [177] are used as coupling components.

The exhaustion rate can be reduced by introducing ether [178] and polyether groups [179] into the aminoalkyl group of the coupling component. The same effect is achieved if the diazo component is substituted by a dioxane ring [180]. Dyes that show a lower affinity for the wool fiber in mixed fabrics are substituted on the aminoaryl nitrogen by acyloxy groups. Suitable coupling components are the benzoates or carbonates of *N*-ethyl-*N*-hydroxyethylaniline [181] or the carbamates obtained by reaction with phenyl isocyanate [182]. With diisocyanates dimeric dyes are obtained that are suitable for spin dyeing polyacrylonitrile [183]. If diazacyanine dyes bearing a free amino group in the benzothiazole group are dimerized via a urea bridge, dyes are obtained that are suitable for dyeing paper [184]. The above dyes, in particular those whose aminoaryl nitrogen is part of a ring, include compounds that are suitable as photoreceptors for electrophotographic applications [185].

The quaternization of heterocyclic azo dyes to form diazadimethine dyes can also be carried out in an acid medium if compounds containing vinyl groups are added to the protonated azo dyes. Acrylic acid [186], acrylates [187] and, in particular, acrylamide [188] are suitable for this purpose. The following dye

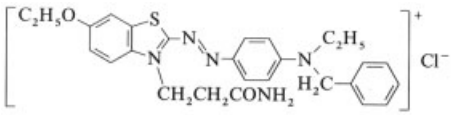

[85959-18-8]

colors polyacrylonitrile in greenish-blue shades. Commercial products are Aizencathilon Blue NBLH, NRLH, GLH, and Atacryl Blue LLR, LL2R.

In acid medium oxiranes, in particular ethylene oxide [189], propylene oxide (see below), butylene oxide and their substitution products [190], can also add on to heterocyclic azo dyes:

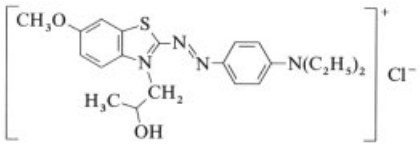

Commercial products include Yation Blue BS, BRS, GRS, Sumiacryl Blue F-GSL, Daitoacryl Blue GL, Acrylon Blue BLN, and Estrol Blue N-BRL.

With 2-aminothiazole as diazo component, violet to reddish-blue dyes are obtained [191]. If the coupling components have a low molecular mass, the dyes exhibit a high leveling ability [192]. One example of this is the migrating blue dye:

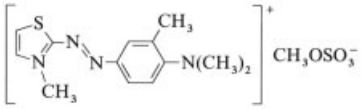

[71196-94-6]

Biscationic dyes, which are usually used for spin dyeing polyacrylonitrile, are also derived from thiazole [193]. The following compound

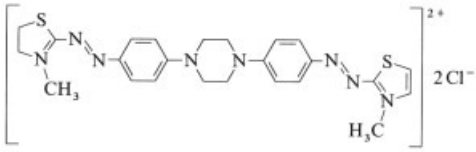

dyes acrylic fibers blue with good steam fastness [194]. Commercial products are Diacryl Supra Blue BL, Sumiacryl Blue G, and Atacryl Violet L6B.

With 3-methyl-5-aminoisothiazole as diazo component, lightfast and decatizing-fast diazacyanine dyes for dyeing polyacrylonitrile are obtained. The dye

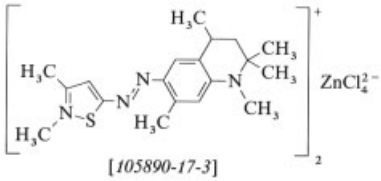

[105890-17-3]

colors polyacrylonitrile a reddish-blue shade [195]. Other suitable coupling components are tetrahydrobenzoxazine [196] and tetrahydroquinoxaline [197].

The coupling of diazotized 3-amino-1,2,4-triazole to aromatic amines affords azo dyes that require two equivalents of an alkylating agent to convert them to diazadimethine dyes. A mixture of two isomeric dyes is obtained, e.g.,:

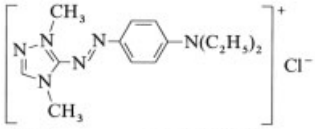

Main component [*74109-48-1*]

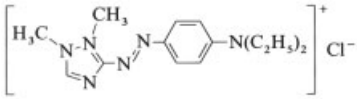

Byproduct [*47089-11-2*]

Dyes that are obtained with *N,N*-dimethyl- and *N,N*-diethylaniline as coupling component and with dimethyl or diethyl sulfate as alkylating agent dye polyacrylonitrile in lightfast bluish-red shades [171], [198]. They have a lower affinity for polyacrylonitrile, but on account of their low molecular mass they are exhausted with a high diffusion rate by the fibers. They are therefore particularly suitable for producing level dyeings in light shades. The same dyes can also be formed with 3-amino-1,2,4-triazole-5-carboxylic acid as diazo component since the azo dyes can easily be decarboxylated [199].

Commercial Products. Maxilon Red BL, Sandocryl Red B-BLE, Astrazon Red F3BL, Synacril Red 3B, Basacryl Red MX-BL, Bezacryl Red BL, and Tetradene Fast Red CBL. In 1989 300 t of Basic Red 22 were produced worldwide.

The affinity of the dyes for polyacrylonitrile and accordingly their exhaustion rate is increased by introducing aromatic groups. Basic Red 46 is of particular industrial importance:

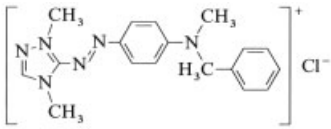

[*29508-47-2*]

In 1988, 1500 t of this dye were produced worldwide.

Commercial Products. Maxilon Red GRL, Astrazon Red FBL, Basacryl Red X-GRL, Aizencathilon Red FG-DP, Estrol Red GSL-SD, Panacryl Red GRL, Sandocryl Red B-RL, Sevron Fast Red GRL, Synacril Red G, and Yoracryl Red BGL.

N,N-Dibenzylaniline and *N*-methyldiphenylamine as well as the arylcarboxylates and aryl carbonates of *N*-(hydroxyethyl)aniline are also used to increase the exhaustion rate. Arylalkyl radicals in the 1-position of the triazole ring likewise increase the exhaustion rate [200]. The commercial products are Aizencathilon Red FRLH, Basacryl Red FL, and Maxilon Red 3BL.

Diazotization, coupling, and quaternization can also be combined in a one-pot process for this class of dyes [201]. Pure dyes are isolated if the separation is performed in the presence of nonionic surfactants [202]. To produce highly concentrated dye solutions, only part of the azo base is added at the start of the reaction, the remainder being added to the dye melt during the methylation process [203]. Many dyes of the triazolediazacyanine series tend to agglomerate during storage. Exceptions are the hydrobromic acid salts [204].

The introduction of a methyl group in the 2-position relative to the azo bond increases stability of these dyes in the dye bath [205]. Dyes that are dimerized by an alkylene bridge via the aminoaryl group are suitable for dyeing porous polyacrylonitrile fibers (Dunova) [206] and for spin dyeing polyacrylonitrile from the gel state [207]. (Commercial product is Diacryl Red 2BLE.)

Additive alkylation in an acid medium with alkylene oxides [208], acrylonitrile [209], acrylamide [210], and acrylic acid [211] has also been described for the triazole dye series.

Two series of dyes are derived from 3,5-diamino-1,2,4-triazole [*1455-77-2*], also known as guanazole. The monoazo dyes form red diazadimethine dyes after alkylation [212]:

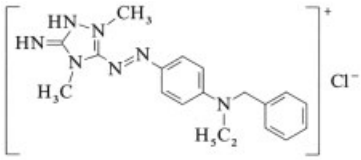

A commercial product of this series is Aizencathilon Red K-BLH. The bisazo dyes formed from guanazole produce reddish-blue diazadimethine dyes on alkylation [213], e.g.,:

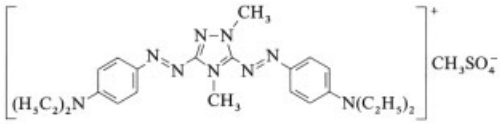

Commercial products include Aizencathilon Marine Blue RLH, GRLH, and Aizencathilon Royal Blue GH.

Blue diazadimethine dyes with a high dye bath stability are derived from 2-amino-5-dialkylamino-1,3,4-thiadiazole. The following dye

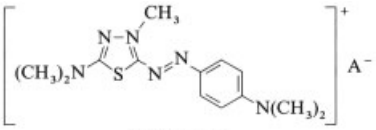

[83930-05-6]

also has a very good leveling ability [214]. The commercial products are Astrazon Blue FBL [*81921-76-8*] and FGRL.

Orange to red diazadimethine dyes are obtained in a similar manner with 3-aminopyrazole as diazo component [215]. The nitrogen atom in the 2-position of pyrazole can be substituted by a cyclohexyl, benzyl [216], or phenyl group [217]:

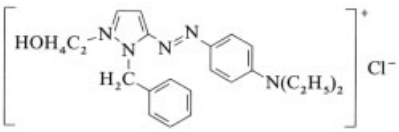

[68123-01-3]

Commercial products are Maxilon Orange 4RL and Basacryl Red X-GFL, X-3GL.

The nitrogen atom in the 1-position may also bear a phenyl group [218]. Partially hydrogenated N-heterocycles such as tetrahydroquinoline [219] and tetrahydrobenzoxazine [220] may also be used as coupling components.

Benzoanellation of pyrazole produces a considerable bathochromic shift; the corresponding indazole dyes are therefore violet to blue [221].

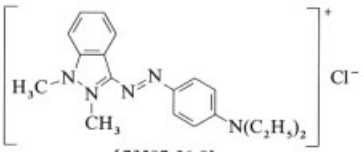

[73507-56-9]

Alternatively, indazole dyes can also be prepared by oxidative coupling of 1,2-dialkylindazol-3-one hydrazone with aromatic amines [222]:

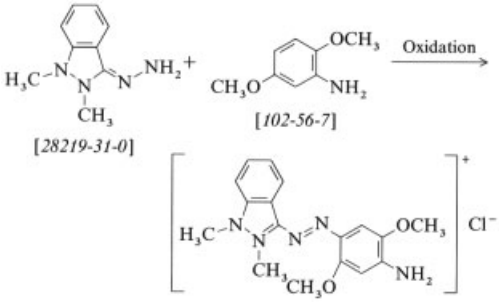

[28219-31-0] [102-56-7]

A further synthesis route is the condensation of a 1,2-dialkylindazolium-3-sulfonate with a 4-acylaminophenylhydrazine, followed by oxidation [223].

Commercial products are Basacryl Violet FL, Maxilon Violet RL.

In contrast to the aforementioned N-heterocyclic compounds, imidazole may be used as coupling component for diazotized aromatic amines. Dyes with primary amino groups are obtained by coupling diazotized 2- or 3-chloro-4-aminoacetanilide with imidazole, methylating the resultant azo dye to form the diazacyanine, and finally hydrolyzing the acetyl group in acid medium [224]. The dye of the formula

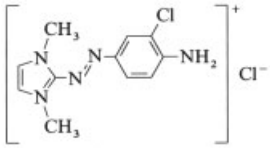

has an outstanding migration ability and accordingly dyes polyacrylonitrile in a lightfast orange shade with high levelness. Reaction of diazo components containing chlorine or a methoxy group in the 4-position to the amino group with imidazole, and subsequent quaternization yields compounds in which the halogen or alkoxy substituents (in the 4-position to the azo group) can be replaced by amino groups [225]. The dyes obtained with lower aliphatic amines (e.g., C.I. Basic Red 51)

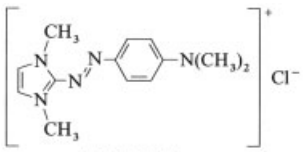

[12270-25-6]

color polyacrylonitrile in lightfast red shades

[226]. On account of their low molecular mass they have a high diffusion rate and are therefore particularly suitable for producing level combination dyeings.

Commercial products of Basic Red 51 include Basacryl Red X-BL, Basacryl Red Violet X-GL, and Maxilon Red M-4GL, M-RL.

Biscationic dyes of the imidazole series are particularly suitable for dyeing acid-modified polyamide fibers [227]. 2-Aminopyridine cannot be converted directly to its diazonium salt. The following violet diazadimethine dye is therefore prepared indirectly via the 2-aminopyridine-*N*-oxide [228]:

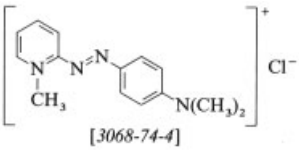

[3068-74-4]

Alternatively, 2-aminopyridine can be reacted with 4-nitroso-*N,N*-dialkylanilines in the presence of sodium amide, followed by quaternization [229]. These dyes are used in analytical chemistry for spectrophotometric determinations [230], [231]. They are also used as hair dyes [232].

A series of N-heterocycles with amino groups (e.g., 4-amino-2-methylquinoline) can also be diazotized as quaternary salts and coupled to form diazadimethine dyes [233]. The following dye gives polyacrylonitrile a blue color:

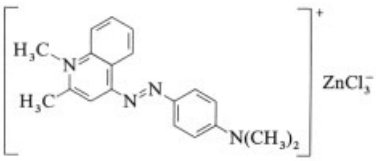

Red to blue-violet dyes for polyacrylonitrile are obtained by oxidative coupling of 2-alkylpyridazonehydrazone-3-derivatives with aromatic amino compounds [234], e.g.,:

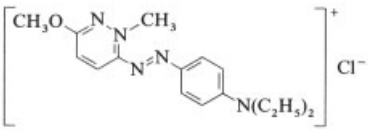

2-Aminothiophene derivatives can be condensed by reaction with potassium thiocyanate and oxidative ring closure to form 2-aminothiophenothiazoles. These heterocyclic amino compounds are starting material for the production of blue diazacyanine dyes [235]. The dye of the formula

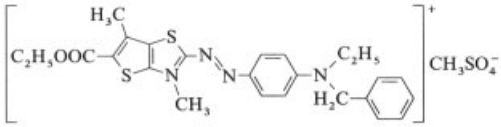

[68560-36-1]

colors polyacrylonitrile a clear greenish-blue shade.

7.3.3. Higher Vinylogs of Phenylogous Hemicyanine Dyes

Condensation of 4-(dialkylamino)cinnamaldehydes with N-heterocycles that bear a reactive methyl or methylene group in the 2- or 4-position to the heterocyclic nitrogen atom produces tetramethinehemicyanine dyes, which are vinylogs of styryl dyes. The dye synthesized from quaternary 2-methylbenzothiazole absorbs at 572 nm. The cyano-substituted dye of the formula

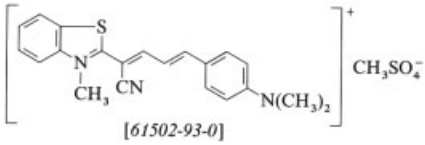

[61502-93-0]

obtained from 2-cyanomethylbenzothiazole exhibits a strong bathochromic shift with an absorption maximum of 647 nm [236].

7.3.4. Phenylogous Hemicyanine Dyes with Rings in the Methine Chain

In the higher vinylogs of phenylogous hemicyanine dyes the methine chain may also be part of a ring; this increases the stability of the dyes. The dye

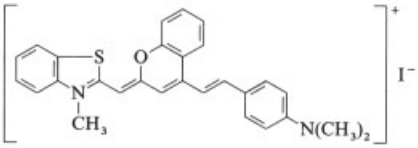

absorbs at 600 nm and is used for optical data storage. This dye is prepared by reacting quaternary 2-methylbenzothiazole with 4-methylbenzopyran-2-thione and condensing the resultant quaternary intermediate with 4-dimethylaminobenzaldehyde [237].

7.4. Cyanine Dyes

In cyanine dyes both charge-carrying terminal atoms of the methine chain are constituents of heterocyclic rings. Depending on the nature of the heterocycle, the methine chain may be coupled in the 2-, 3-, or 4-position to the charge-carrying terminal atom. Accordingly, either one, two, or three ring atoms participate directly in the mesomeric methine system.

According to the general definition of methine dyes, the number of atoms in the conjugated system must always be an odd number. In symmetrical dyes the sum of the charge-carrying atoms in both heterocyclic rings is always even. The methine chain must therefore contain an odd number of atoms. In cyanine dyes with different heterocyclic groups, however, the sum of the atoms involved in the charge-conducting system in both rings can be an odd number. In this case the methine chain must contain an even number of atoms. Cyanine dyes with a substituted methine chain or in which parts of the chain are constituents of a ring are particularly important in photography (→Photography).

7.4.1. Zeromethinecyanine Dyes

Naphtholactam and its N-alkyl derivatives can be condensed with indoles in the presence of acid condensation agents such as phosphorus oxychloride to form red zeromethinecyanine dyes that dye polyacrylonitrile very lightfast shades [238]. The indole may contain a methyl or phenyl group in the 2-position. If the 2-position is not substituted, the coloring strength increases because there is less steric hindrance [239], e.g.,:

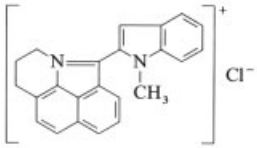

Condensation of indoles with N-alkylquinolones or N-alkylpyridones yields yellow and orange zeromethine dyes that dye polyacrylonitrile in lightfast shades [240]. If the indole nitrogen is unsubstituted, intensely colored anhydro bases are formed from the zeromethinecyanine dyes. Dealkylation of the heterocyclic quaternary nitrogen atom also occurs [241].

The second heterocyclic ring of the zeromethinecyanine dyes can also be formed in the last synthesis step. The methylene group of 1,3,3-trimethyl-2-methyleneindoline can react with alkylene oxides at elevated temperature. If the hydroxyl group is replaced by a halogen and the resultant compound is coupled with aromatic diazonium salts, intramolecular alkylation occurs simultaneously and a pyrazoline ring is formed. These dyes color polyacrylonitrile in brilliant, fluorescent shades that are extremely lightfast [242]:

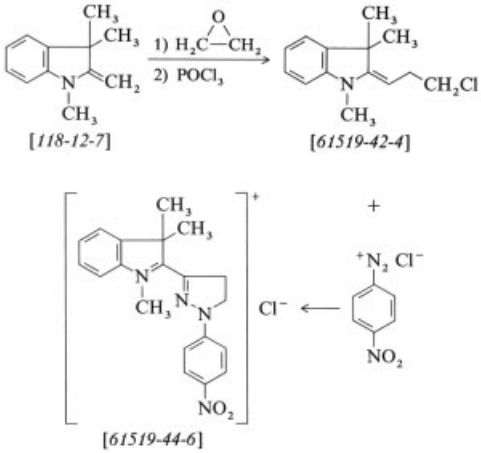

The solubility of these dyes is increased by introduction of polyether groups [243].

7.4.2. Monomethinecyanine Dyes

The oldest dye of this series is *cyanine*, a blue dye after which the whole class of dyes was named. Cyanine was synthesized by WILLIAMS

in 1856 from crude tar quinoline by quaternization with pentyl iodide, followed by treatment with alkali:

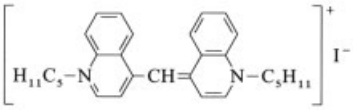

[862-57-7]

The unsymmetrical monomethine dye coupled in the 2- and 4-position of quinoline is termed *isocyanine*. Monomethine dyes containing two molecules of quinoline coupled to each other in the 2-position (also called pseudocyanines) are obtained by the alkaline-catalyzed reaction of *N*-alkylquinaldinium salts with quinolinium compounds [244]:

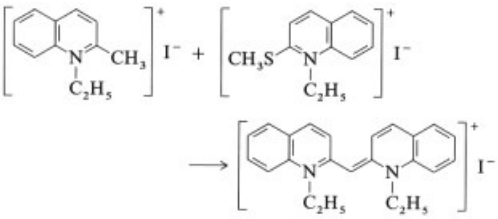

[977-96-8]

Monomethinecyanine dyes of indolenine are obtained by reaction of 1,3,3-trialkyl-2-methylindoleninium salts with esters or salts of nitrous acid in the presence of acetic anhydride. A nitroso compound is formed which reacts with a further mole of the indoleninium salt with the elimination of hydrocyanic acid to form a monomethine dye [245]:

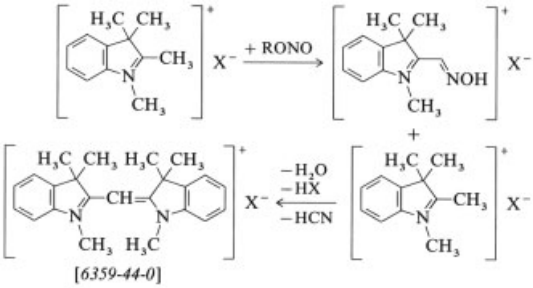

[6359-44-0]

This class of dyes colors tanned cotton, cellulose acetate, and polyacrylonitrile in greenish yellow shades with good lightfastness. The unsubstituted dye of this series was converted with phosphomolybdotungstic acid into a yellow pigment.

The allene compound is obtained by the action of sodium hydride on the monomethine dye [246].

The benzoxazolemonomethinecyanine dyes are obtained in a similar way to the monomethinecyanine indolenine dyes [247].

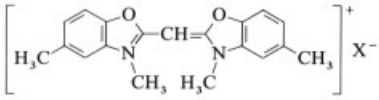

[47296-06-0]

These dyes are suitable for sensitizing silver chloride emulsions for the violet and blue regions of the spectrum. Bacteriostatically active dyes have also been discovered in this series [248]. Their intense blue-violet fluorescence makes them suitable for use as lightfast and washfast brighteners for natural and synthetic fibers [249]. Uvitex ALN is a commercial product of this range.

The hydrogen at the methine carbon atom in monomethine dyes can be reversibly substituted by halogen [250]. Monomethinecyanines can be coupled with aryldiazonium salts at the central (meso)-carbon atom to form mesoarylazomonomethinecyanines, which according to their absorption spectra have a phenylhydrazone structure [251].

If the methyl or methylene group of the N-heterocycle is activated by a substituent (e.g., a cyano group), it can be condensed directly with the relevant carbonyl compound in the presence of a dehydrating agent, especially phosphorus oxychloride. *N*-alkylnaphtholactam [252] and its derivatives [253] condense with 1,3,3-trimethyl-2-cyanomethyleneindoline to form red and violet dyes that dye polyacrylonitrile in lightfast shades:

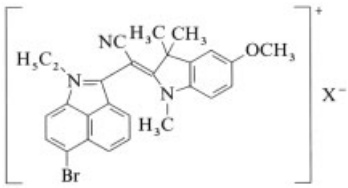

2,6-Diphenyl-[2*H*]-pyran-4-one condenses with acetophenone to form the following pyryliummonomethine dye:

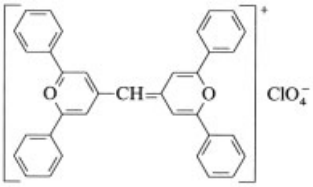

[47793-49-7]

This dye absorbs in the IR region [254] and is used to prepare nonerasable optical recording materials [255].

The monoazamonomethinecyanine dyes are obtained by reaction of imines of benzothiazol-2-ones with strong, preferably anhydrous acids [256]:

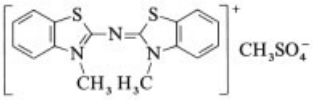

[66179-76-8]

The compounds display blue fluorescence and are used as optical brighteners (e.g., Ultraphor AL, BASF) for polyacrylonitrile [257].

7.4.3. Dimethinecyanine Dyes

In the most important dimethinecyanine dyes the methine chain links two indole moieties in the 2- and 3-positions.

The dyes were initially used for dyeing and printing cellulose acetate [258], but became increasingly important when it was discovered that they can dye polyacrylonitrile in brilliant and lightfast shades.

C.I. Basic Orange 21

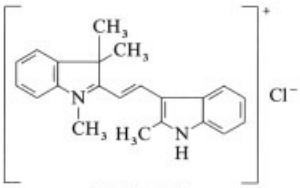

[47346-66-7]

is prepared either from 2-methylindoline-3-aldehyde and 1,3,3-trimethyl-2-methyleneindoline or from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and 2-methylindole. It is used in many ranges for dyeing polyacrylonitrile. In 1989, 250 t were produced worldwide.

Preparation. 1230 kg of 30 % hydrochloric acid are brought into a 1500-L enameled vessel, followed by 104 kg of 2-methylindole at room temperature, while stirring. When the 2-methylindole has completely dissolved after 1 – 2 h, 154 kg of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde is quickly added at room temperature and the reaction mixture is stirred for 24 h at room temperature; the dye starts to crystallize after ca. 15 min. The reaction mixture is diluted with 400 L of cold water and the dye suspension is suction filtered, washed twice with 200 L of cold water, and thoroughly dried under suction. After drying at 95 – 100 °C, a yield of 323 kg is obtained, corresponding to 1043 kg of commercial product.

Commercial Products. Astrazon Orange G, Maxilon Orange GA, Aizencathilon Orange GLH, Asahiacryl Orange GL, Atacryl Orange LR, Cathilon Orange GL, Concordacryl Orange GL, Hispacril Orange G, Kayacryl Orange 2G, Nabor Orange G, Sumiacryl Orange G, Viocryl Orange GLS, and Yation Yellow RM.

A red-orange dye is prepared in a similar manner from 1-methyl-2-phenylindole. The commercial products are Astrazon Orange R, Maxilon Orange RA, Aizencathilon Orange RLH, Nabor Orange R, and Sumiacryl Orange R.

If this dye is reacted with an aqueous solution of potassium cyanide, the cyanide group is introduced into the 2-position of the indoline ring system and the dye is decolorized [259].

Dimethinecyanine dyes in which the indole ring is unsubstituted in the 2-position are used for dyeing polyacrylonitrile and for histological staining [260].

Reaction of pyrrole-2- [261] and pyrazole-4-aldehydes [262] with 1,3,3-trimethyl-2-methyleneindoline has also been described. The resultant dimethinecyanine dyes color polyacrylonitrile yellow. Dimethinecyanine antipyrine dyes are obtained by condensation of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde with a pyrazol-5-one and subsequent alkylation. The following dye

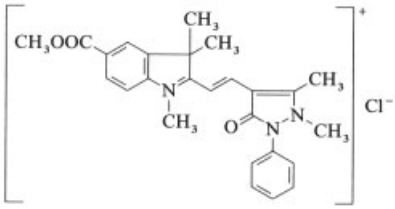

dyes polyacrylonitrile in reddish-yellow shades. Dimethinecyanine dyes prepared by condensation of indole-3-aldehydes with N-alkyl-γ-picolinium salts [264] color paper brilliant yellow shades:

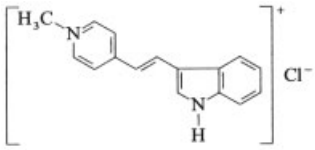

[64651-39-4]

Diazadimethinecyanine dyes are formed by coupling diazotized N-heterocyclic amines in the β-position to the nitrogen atom of other heterocycles followed by alkylation. The most important coupling agent is 1-methyl-2-phenylindole. Reaction of this compound with diazotized 2-aminothiazole and subsequent methylation with dimethyl sulfate, yields a dye that draws up very quickly on polyacrylonitrile and dyes it a lightfast red shade [265]:

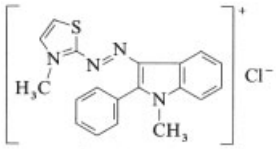

[42373-04-6]

Commercial products re Basacryl Red GL, Cat acryl Red GL, and Diacryl Red GL-N.

Additive alkylation with acrylamides yields dyes that have a medium sorption rate [266]:

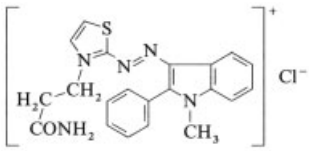

[50905-70-9]

Commercial products nclude Aizencathilon Red GLH and Atacryl Scarlet LL7B.

N-Alkylation with alkylene oxides and their derivatives also yields dyes of medium sorption rate [267]. Commercial products are as follows: Diacryl Red CS-N, Estrol Red N-GL, Sumiacryl Red FRL, N-RL, and Yation Red GS.

Coupling diazotized 3-amino-1,2,4-triazole with 1-methyl-2-phenylindole and double alkylation, yield dyes that dye polyacrylonitrile in lightfast, reddish-yellow shades [268]:

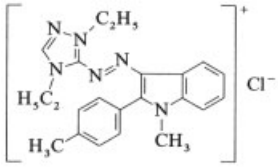

Commercial products nclude Basacryl Yellow 5RL, Aizencathilon Yellow DS-DP, Diacryl Yellow 2RN, Estrol Yellow N-2RL, Genacryl Yellow RRL, Kayacryl Yellow 3RL-ED, and Sumiacryl Yellow F-2RL.

Dyes that have dodecyl sulfate as an anion are suitable for spin dyeing polyacrylonitrile from acetone solution [269]. In this series of dyes additive alkylation can also be successfully performed with alkylene oxides [270] or cycloalkylene oxides [271].

Diazacyanine dyes with two different alkyl substituents on the triazole nitrogen atoms are obtained by coupling diazotized 3-amino-1-benzyl-1,2,4-triazole to 1-ethyl-2-phenylindole and quaternizing the azo dye with dimethyl sulfate or, in the acid pH range, with acrylamide to form the diazacyanine dye [272]. The commercial products are Aizencathilon Yellow 3R-DP, K-3RLH, and RPLH.

Mono- and bisdiazacyanine dyes are derived from guanazole. Whereas the monodiazacyanine dyes dye polyacrylonitrile reddish-yellow shades [273], the bisdiazacyanines [274] produce red dyeings, e.g.:

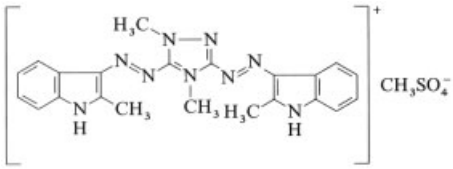

Dyes with 3-aminopyrazole or its derivatives [275] as diazo component dye polyacrylonitrile yellow, e.g.:

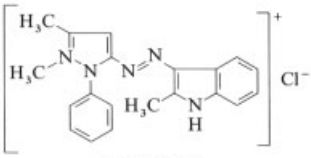

[33639-59-7]

The tetrafluoroborate of this dye is used for spin dyeing polyacrylonitrile [276]. Commercial products are Maxilon Yellow GL and Basacryl Yellow X-GL.

Benzoanellation of pyrazole produces a bathochromic shift (e.g., Basacryl Orange FL) [277]. Further suitable heterocyclic starting components are the amino derivatives of *N*-methyl-benzimidazole [278], 4,5-dicyanoimidazole [279], and pyridine [280].

7.4.4. Trimethinecyanine Dyes

Trimethinecyanine dyes are important sensitizers in photography (→Photography). The first sensitizer for the longwave red region was cryptocyanine [*4727-50-8*], obtained from quaternized lepidine:

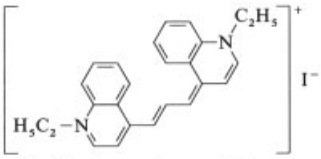

Sensitization maximum: 735 nm

Even more important was pinacyanol [*605-91-4*], which was obtained from quaternary quinaldine by reaction with formaldehyde to form the leuco compound and atmospheric oxidation in an alkaline medium [281]:

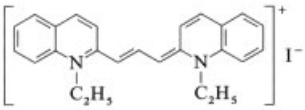

Sensitization maximum: 637 nm

Oxidation can also be performed in an acid medium. Suitable oxidizing agents are atmospheric oxygen, lead dioxide, or chloranil [282]. Oxidation is not necessary if condensation is carried out with chloroform, bromoform, or iodoform [283]. The Vilsmeier complex formed from phosphorus oxychloride and dimethylformamide is also suitable for introducing the middle methine carbon atom [284]. Pinacyanol is used as an electrophotographic photoreceptor [285]. This class of dyes also includes compounds having antimicrobial properties [286]. Fluorine-containing substituents (e.g., the trifluoromethylsulfonyl group) produce a bathochromic shift of the absorption maximum [287]. The trifluoromethylsulfonyl substituent has also been introduced into trimethinecyanine dyes of the benzimidazole series by condensation of the substituted quaternary salt with triethyl orthoformate [288].

Astraphloxin [*6320-14-5*] is obtained by reaction of 2 mol of 1,3,3-trimethyl-2-methyleneindoline with 1 mol of orthoformate [289] or other formic acid derivatives [290]:

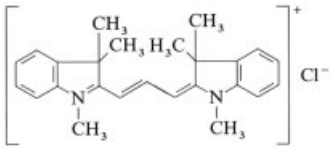

This compound dyes polyacrylonitrile, cellulose acetate, and paper in brilliant red shades. It is used in the form of its phosphomolybdotungstic salt [291] or as the tetraphenyl borate [292] to produce colored toners for electrophotographic applications. An improved process for preparing astraphloxin uses the mixed anhydride of acetic and formic acids (formed by reacting acetic anhydride with sodium formate) to introduce the methine carbon atom in the meso position [293].

Preparation. A mixture of 200 kg of glacial acetic acid and 200 kg of acetic anhydride is cooled to −6 °C. Sodium formate (93.2 kg) is added and produces a weakly exothermic reaction: the temperature rises to 2 – 6 °C. The resulting suspension thickens on prolonged stirring. When the temperature reaches 5 °C 200 kg of 1,3,3-trimethyl-2-methyleneindoline is added, the temperature increases to 10 – 15 °C. As soon as all the base has been added, the brine cooling is removed and the reaction mixture is stirred for 1 h, the temperature rises to 20 °C. The reaction mixture is heated for one more hour at 50 – 53 °C and stirred for a further 18 h at this temperature. If the 1,3,3-trimethyl-2-methyleneindoline con-

tent is still greater than 1 %, the melt is stirred for a further 5 – 10 h at 50 – 53 °C. The reaction mixture is then diluted with 150 L of water and stirred for 30 min at 50 °C until it is homogeneous. The condensation reaction mixture is then added to a 35 – 40 °C solution of 210 kg of 30 % hydrochloric acid in 2800 L of water. After cooling the mixture to room temperature, the dye is salted out with 500 kg of sodium chloride. The suspension is stirred for 5 h at room temperature, filtered, and washed with a solution of 15 kg of 30 % of hydrochloric acid and 90 kg of sodium chloride in 850 L of water. Finally, the press cake is washed with a solution of 95 kg of sodium chloride in 850 L of water. A yield of 229 kg of astraphloxin is obtained.

To prepare a concentrated liquid dye formulation, when the 1,3,3-trimethyl-2-methyleneindoline content has fallen to < 1 %, the following components are added to the condensation melt at 50 °C: 43 kg of glacial acetic acid, 75 kg of ethylene glycol, 11 kg of the addition product of 40 moles of ethylene oxide with oleyl alcohol, and 140 kg of water. This solution can be used directly to dye paper containing wood pulp.

The lightfastness of trimethinecyanine dyes is considerably improved by introducing cyano groups into the methine chain [294]. *N*-alkylnaphtholactam can be converted into the corresponding methylene compound by the Grignard reaction [295]. Condensation of the latter with 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and its substitution products yields violet to blue dyes for dyeing polyacrylonitrile. The cyano-substituted dyes are very lightfast [296].

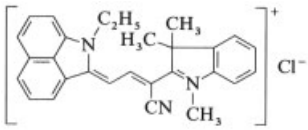

Blue and green trimethinecyanine dyes are obtained with 3-methyl-2-methylene-1-phenyl-1,2-dihydroquinoxaline as condensation component [297]. Introduction of pyrylium terminal groups extends the absorption range further into the IR region. The following dye is obtained by condensation of 4-methyl-2,6-diphenylpyrylium perchlorate with dimethoxyethylacetate [298].

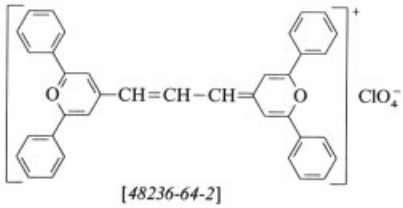

[48236-64-2]

In order to prepare unsymmetrical pyrylium dyes, 4-formyl-2,6-diphenyl-methylenepyran is reacted with pyrylium salts. Thiopyrylium dyes can also be obtained in the same way [299]. The dyes are used for nonerasable optical data storage [300]. Trimethinecyanine dyes with benzopyrylium [301], benzothiazolium, and benzimidazolium terminal groups [302] are also suitable for this purpose. Trimethinecyanine dyes with benzopyrylium terminal groups are used in electrophotography to produce organic photoconductors [303].

Monoaza-, Phospha-, and Arsatrimethinecyanine Dyes. If the methine groups in the α-position are replaced by *nitrogen*, a marked hypsochromic shift is produced. With the following dye

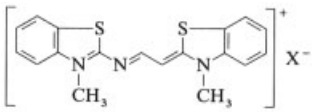

[31787-22-1]

this shift is 81 nm [304]. These dyes are used to produce organic photoconductors [305]. If, however, the β-methine group is replaced by nitrogen, a bathochromic shift of about 55 nm is produced.

To synthesize *β-phosphatrimethinecyanine* dyes, 1,3,3-trimethyl-2-methyleneindoline is reacted with phosphorus tribromide to form bromo-bis(1,3,3-trimethylindoline-2-ylidenemethyl) phosphane, which is then reacted with trimethyloxonium tetrafluoroborate. The replacement of the β-methine carbon atom by phosphorus leads to a bathochromic shift of 33 nm (λ_{max} = 586 nm) [306].

Similiarly *arsatrimethinecyanine* dyes are obtained with arsenic tribromide. The maximum

bathochromic shift in this case is considerable: 111 nm (λ_{max} = 664 nm) [307].

Diazatrimethinecyanine Dyes. The reaction of *N*-alkylbenzothiazol-2-onehydrazones with benzimidazole-2-aldehydes produces dye bases which can be alkylated to give greenish-yellow diazatrimethinecyanine dyes. These dyes fluoresce strongly and dye polyacrylonitrile brilliant shades [308]. They are also used as fungicides for aquatic animals [309].

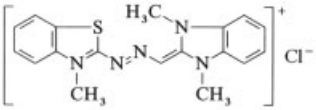

Commercial products are Basacryl Yellow 5GL and Kayacryl Brilliant Yellow 5GL.

Diazatrimethinecyanine dyes can also be prepared by coupling diazotized 2-aminobenzothiazole or 3-amino-1,2,4-triazole with benzimidazol-2-yl acetic acid, decarboxylation, and alkylation of the heterocyclic nitrogen atoms [310]. Oxidative coupling of heterocyclic hydrazones with *N*–heterocycles containing reactive methyl or methylene groups can also be employed [311].

Cyano-substituted diazatrimethinecyanine dyes are obtained from cyanomethyl or cyanomethylene derivatives. For example, diazotized 2-aminobenzothiazole can be coupled with 2-cyanomethylbenzimidazole or 1,3,3-trimethyl-2-cyanomethyleneindoline, the resultant dye bases are alkylated [312].

Triazatrimethinecyanine Dyes. Replacement of all three methine groups of trimethine dyes by nitrogen results in a hypsochromic shift. Such compounds are synthesized by diazotizing a heterocyclic amine and coupling with the same or another amine. The resultant triazene is alkylated to give the triazatrimethine dye [313]. The following compound

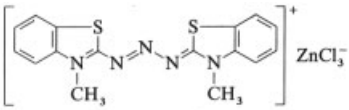

[3068-61-9]

dyes polyacrylonitrile a brilliant, very lightfast yellow-orange shade. Corresponding dyes containing benzimidazole terminal groups are obtained by reacting 2-chloro-1,3-diethylbenzimidazolium tetrafluoroborate with sodium azide to form the azidinium salts, followed by reaction with sodium azide in dimethylformamide [314].

7.4.5. Pentamethinecyanine Dyes.

Pentamethinecyanine dyes can be obtained analogously to trimethinecyanine dyes from the methyl-substituted heterocyclic quaternary salts by condensation with β-ethoxyacrolein acetal [315] or malonaldehyde tetraethyl acetal [316] in the presence of acetic anhydride or pyridine. The vinylogous dye of pinacyanol synthesized from quaternary quinaldine absorbs in the IR region (λ_{max} = 813 nm) and is used to produce nonerasable storage discs for optical information recording [317]:

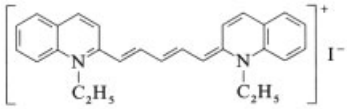

[14187-31-6]

Dyes with a substituted central methine carbon atom are obtained from the correspondingly substituted malonaldehydes. The introduction of electronegative substituents into the meso-position of the pentamethine chain produces, in contrast to the trimethinecyanine dyes, a hypsochromic shift of the absorption maximum. The dye prepared from sodium nitromalonaldehyde and quaternary quinaldine absorbs at 660 nm (i.e., at a 153 nm shorter wavelength than the unsubstituted dye). Bromine, chlorine, fluorine and cyano or carboethoxy groups can be introduced into the meso-position [318]. The influence of substituents has been investigated with ^{13}C-NMR-spectroscopy [319].

Pentamethinecyanine dyes that are substituted in the meso-position by an azo group are obtained with arylazomalonaldehydes [320]. Aminomalonaldehydebisdimethylacetal, which serves as a starting material for meso-acylamino-substituted pentamethinecyanine dyes, is obtained by catalytic hydrogenation of the arylazomalonaldehyde [321]. Dialkylaminomalonaldehydes, which are start-

ing compounds for the synthesis of dialkylaminopentamethinecyanines, can be prepared from bisdialkylaminoethylenes by the Vilsmeier reaction [322]. Pentamethinecyanine dyes with benzo[*c,d*]indolenine [323] and pyrylium [324] terminal groups absorb in the IR region and are used to produce nonerasable optical storage media.

Tetraazapentamethinecyanine Dyes. Aliphatic or aromatic aldehydes react with N-heterocyclic hydrazones to produce aldazines, which react with a further mole of the heterocyclic hydrazone in the presence of oxidizing agents to form tetraazapentamethinecyanine dyes. The dye

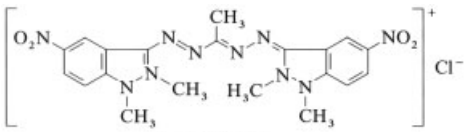

[*94108-98-2*]

colors polyacrylonitrile a lightfast green shade.

7.4.6. Higher Vinylogs of Cyanine Dyes

Heptamethinecyanine dyes are obtained by reacting the quaternary salts of 2- or 4-methyl-N-heterocycles or the corresponding methylene compounds with pentamethinestreptocyanine dyes (see Section 7.1.2). For example, condensation of glutaconaldehydedianil with 1,3,3-trimethyl-2-methyleneindoline yields

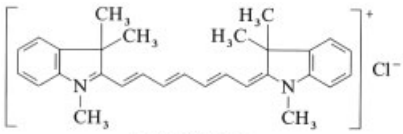

[*126609-59-4*]

which is used to produce nonerasable optical storage media [326].

Use of the complex salts of these dyes with phosphorus-containing heteropolyacids (e.g., phosphomolybdate, phosphotungstate, or phosphotungstomolybdate) has also been suggested for the same purpose [327]. Heptamethinecyanine dyes containing sulfonic acid groups and indolenine terminal groups are used as IR filter dyes [328].

Substitution of the dyes in the central (meso)-position produces a hypsochromic shift of the absorption bands in the case of electron donors and a bathochromic shift in the case of electron acceptors [329]. Nonamethine dyes and higher vinylogs are obtained from the corresponding streptocyanine dyes and heterocyclic quaternary salts [330].

7.4.7. Cyanine Dyes with Rings in the Methine Chain

The stability of polymethine dyes is improved by incorporation of ring systems in the methine chain [331]. Suitable starting materials are the dialdehydes obtained by the Vilsmeier reaction from cyclohexanone and cyclopentanone [332]. Reaction of the dialdehyde derived from cyclopentanone with 2-methyl-4,6-diphenylpyrylium perchlorate yields [333]:

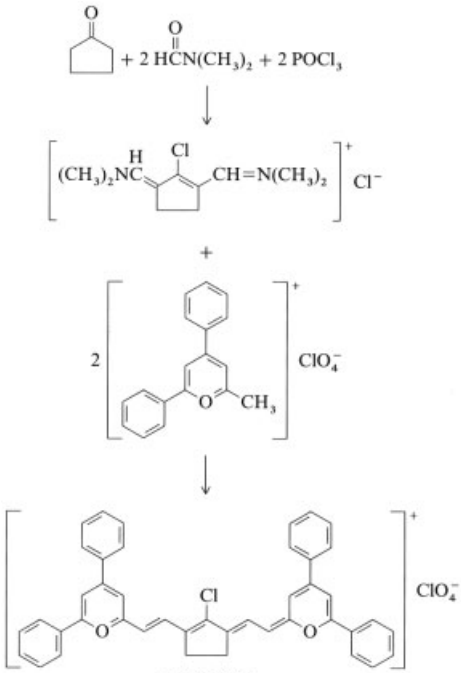

[*61010-22-8*]

λ_{max} = 1.095 nm, ε = 167000

The above aldehydes react with 1,3,3-trimethyl-2-methyleneindoline and its substitution products to yield IR-absorbing dyes. These dyes are brought together with sublimable dyes onto a transfer ribbon. The absorption of laser energy and the dye transfer in laser video printing is thus improved [334]. The halogen atom in the meso-position can be substituted by amines [335] to give dyes that serve as sensitizers in electrophotography. Replacement of the halogen by compounds with a reactive methylene group (e.g., malononitrile [336] or barbituric acid [337]) leads to the formation of IR-absorbing dyes that are used to produce nonerasable optical storage elements. The absorption range of this class of dyes can be increased to 1328 nm by vinylogous chain extension [338].

7.4.8. Cyanine Dyes with Several Methine Chains

Branched cyanine dyes are obtained by reacting triformylmethane with N-heterocyclic quaternary salts or the corresponding methylene compounds. With 1,3,3-trimethyl-2-methyleneindoline the following dye is obtained [339]:

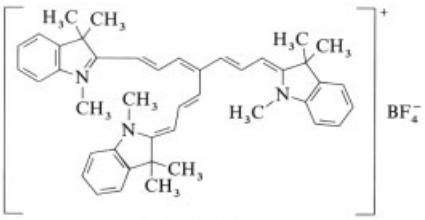

[80137-15-1]

A propeller-like configuration has been confirmed by ^{13}C-NMR spectroscopy, demonstrating its structural similarity to the triphenylmethane dyes [340].

Monomethinecyanine dyes with the 2,6-dimethylpyrylium terminal system can be reacted with aromatic or heterocyclic aldehydes to give dyes containing several methine chains [341]. The following dye

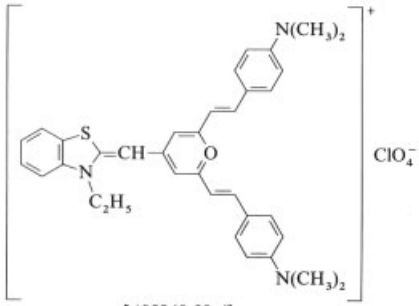

[105845-39-4]

is used as an organic photoconductor in electrophotography.

8. Anionic Methine Dyes (Oxonols)

Anionic methine dyes are formed by elimination of a proton from their conjugated acids:

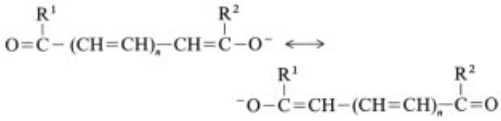

Since oxygen is the charge-carrying terminal atom in the majority of these dyes, they are also termed oxonols. The simplest member of this series is the sodium salt of malonaldehyde; the next higher vinylog is the sodium salt of glutaconaldehyde:

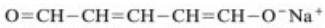

Only oxonol dyes derived from cyclic compounds containing reactive methylene groups are important in dyeing. They are used in the protonated form in acid media as disperse dyes.

8.1. Monomethineoxonol Dyes

The following symmetrical monomethineoxonol dye was first synthesized by E. Besthorn and L. Knorr in 1897 by the reaction of 1-phenyl-3-methylpyrazol-5-one with chloroform and alkali:

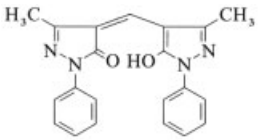

[*4174-09-8*]

Other reactants for synthesizing these dyes include orthoformates, dialkyl or alkylarylformamides and, in particular, formamide [342].

Unsymmetrical monomethineoxonol dyes are obtained by Vilsmeier formylation of pyrazolone derivatives and reaction with a further molecule of pyrazolone [343]. The compounds exist in the cis-form as 8-membered ring proton chelates [344]. The dyes are suitable for dyeing and printing textiles made from aromatic polyesters, polyamides, polyurethanes, and cellulose esters [345]. The dyes are also employed for extremely lightfast transparent varnishes [346]. The 4,4'-methinebis(pyrazolone) dyes are used for the transparent dyeing of thermoplasts such as polystyrene and poly(methyl methacrylate). The commercial products are Macrolex Yellow G (Bayer) and Thermoplast Yellow G (BASF).

Metal salts of 5,5'-methinebis(barbituric acid) derivatives are used as pigments for dyeing poly(vinyl chloride) [347].

The fulgides discovered by H. G. Heller may be classified as monomethineoxonol dyes. A typical representative is obtained as follows:

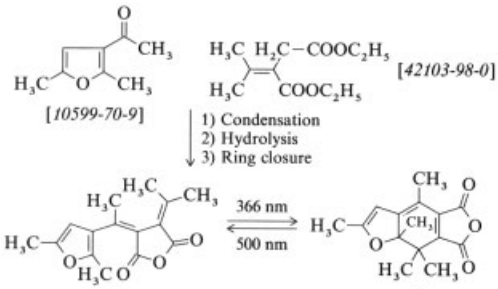

Irradiation at 366 nm converts the pale yellow monomethineoxonol dye (**3**) into the deep red cyclic oxonol (**4**). When irradiated with visible light (**4**) undergoes ring opening into the pale yellow monomethine oxonol (**3**). In contrast to other photochromic compounds, both forms are stable. Investigations aimed at using these compounds for the production of erasable optical memories are underway [348]. Long-chain groups on the methine carbon atom further stabilize the two forms [349]. Imides may also be used instead of anhydrides [350].

Dispersol Red CB-N (ICI) is also a cyclic oxonol dye. It contains a double oxonol structure as indicated by the dotted line [351]:

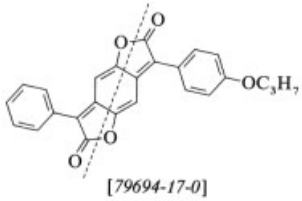

[*79694-17-0*]
λ_{max} = 506 nm

The ammonium salt of purpuric acid (murexide [*3051-09-0*]) can be classified as a monoazamonomethineoxonol dye [352]:

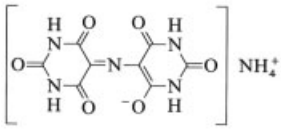

The corresponding symmetrical dye, rubazonic acid [*909-59-1*], synthesized from phenylmethylpyrazolone and the corresponding unsymmetrical dyes containing the pyridone group should also be classified as monoazamonomethineoxonol dyes [353].

8.2. Trimethineoxonol Dyes and Higher Vinylogs

The reaction of cyclic α-ketomethylene compounds with β-alkoxyacroleins yields trimethineoxonol dyes [354]. With 3-cyano-1-ethyl-4-methyl-6-hydroxypyrid-2-one the following dye is obtained:

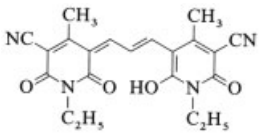

This dye is used in photography and in electrophotography for producing organic photo-

conductors [355]. The detection of lipoperoxides with thiobarbituric acid is based on the formation of a trimethineoxonol [356]. Pentamethinecyanineoxonol dyes are used in thermoreactive recording materials [357].

9. Methine Dyes with Betaine Structure

Condensation of squaric acid with two equivalents of a N,N-disubstituted aniline in the presence of a strong acid and a primary alcohol produces dyes with a betaine structure in the ground state. Their absorption maximum exceeds 600 nm [358].

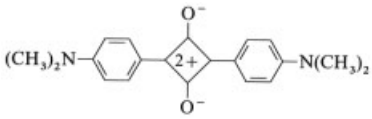

[13458-60-1]

Condensation may also be performed in the presence of an alkylamine [359] and by azeotropic distillation of the water formed in the reaction [360].

Dyes substituted by halogen in the 2-position relative to the squaric acid are used in organic photoconductors for electrophotography [361]. Dyes with hydroxyl groups in this position are used to produce IR-sensitive recording materials [362]. Dyes in which the aniline nitrogen has a benzyl substituent have proved particularly useful in electrophotography [363].

Instead of aromatic amines, compounds with a reactive methylene group can also be used as condensation components. Dyes formed from 2,6-dimethyl-4-phenyl- and 2-methyl-4,6-diphenylpyrylium salts absorb in the IR region [364]. The following dye

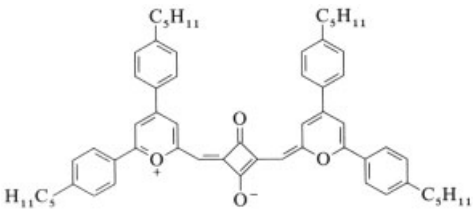

absorbs at 862 and 912 nm. Dyes with a thiopyrylium terminal absorb in the same region [365]. The dyes are used to produce nonerasable optical discs. Dyes synthesized from squaric acid [366] and croconic acid [367] with azulenium terminal systems can be used for the same purpose.

10. Neutral Methine Dyes

In cationic and anionic methine dyes, the two heteroatoms at the ends of the methine chain can partially take up charge. The terminal atoms in neutrocyanines have different properties. The methine system includes both a donor (D) and an acceptor group (A). If a betaine structure is formed in the transition to the first electronically excited state, then the donor group carries the positive charge and the acceptor group the negative charge. Thus light absorption results in charge transfer.

In the industrially important dyes the *donor* is always an amino group. However, the donor may also be oxygen or sulfur. Under photoexcitation the free electron pair on nitrogen is transferred to the conjugated system with the formation of a positive charge at the nitrogen. The nitrogen may be bonded within an open chain or a ring:

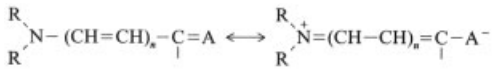

In the most important industrial dyes the electron *acceptors* are carbonyl, thiocarbonyl, and nitrile groups. All three groups take up the negative charge in the first excited state

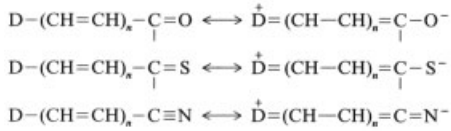

The acceptor group may also be a member of a ring.

The neutrocyanine dyes are insoluble in water and therefore must be dispersed if they are to be used in aqueous dye baths. They may however also contain hydrophilic groups (e.g., sulfonic acid or quaternary ammonium groups) and then acquire the character of cationic or anionic dyes with localized charge. In cationic and anionic methine dyes the charge is delocalized.

10.1. Enamineneutrocyanines

In the enamineneutrocyanines the amino group acting as the electron donor is coupled directly to the methine chain. *N*-vinylindole [368] and *N*-vinylcarbazole [369] can react with tetracyanoethylene to form enamineneutrocyanines. The compound

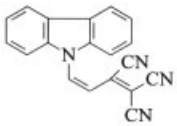

[31317-52-9]

is used as a photoconductor for electrophotographic applications.

10.2. Aminoarylneutrocyanines

The aminoarylneutrocyanine dyes are the phenylogs of the enamineneutrocyanines, an aromatic group being inserted between the donor amino group and the methine chain. Open-chain as well as cyclic compounds can be used as condensation components with reactive methylene groups.

10.2.1. Aminoarylneutrocyanines with Open-Chain Terminal Groups

Condensation of 4-dialkylaminobenzaldehydes with cyanoacetates or cyanoacetamides yields dyes that color cellulose esters, particularly cellulose acetate, brilliant greenish-yellow shades [370]. The following dye (Celliton Fast Yellow 7G, Telasol Yellow 4 GLS) is particularly suitable for dyeing cellulose acetate [371]:

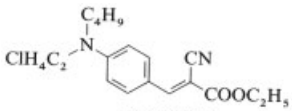

[4361-84-6]

The lightfastness of these dyes on poly(ethylene glycol)terephthalate is only moderate and can be improved considerably if malononitrile is used as condensation component.

Dyes with short-chain alkyl substituents on the amino group tend to sublime from polyesters. This property is used to advantage, however, in transfer printing, especially in the transfer from paper to textile fabrics and in the production of colored copies by sublimation transfer printing [372]. For dyeing polyester fibers, various functional groups are introduced to reduce the sublimation.

Dyes in which the hydroxyethyl group on the nitrogen atom is esterified by aromatic carboxylic acids have a particularly good fiber affinity capacity [373]. An example is

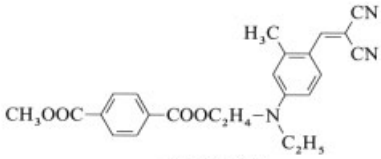

[41284-31-5]

Commercial products are Resolin Brilliant Yellow 7GL (Bayer) and Serisol Brilliant Yellow 6GL (Yorkshire Chemical).

Esterification of the hydroxyethylamino group with isocyanates yields dyes with carbamate groups [374]. Commercial products include Foron Brilliant Yellow 6GFL, S-7GL (Sandoz), Palanil Yellow 7G (BASF), and Kayalon Polyester Light Yellow 6GSF (Nippon Kayaku). The stability of these dyes to hydrolysis in a neutral dye bath is increased by introducing aromatic groups. Aryl ethers [375], aryl thioethers [376] and benzothiazolyl thioethers [377] have been used as substituents. The following dye is stable in a dye bath up to pH 7:

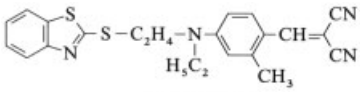

[34613-03-1]

Commercial products are Amacron Yellow 6GSP (Koppers) and Proxester Brilliant Yellow 8GL (Protex).

Dyes that are bonded via the alkyl group of the aminoaryl nitrogen to a sterically hindered piperidine (radical quenchers) exhibit high lightfastness [378].

Dyes derived from 2,2,4,7-tetramethyltetrahydroquinoline and whose ethyl group in the 1-position is substituted with a carbamate group are industrially important [379]:

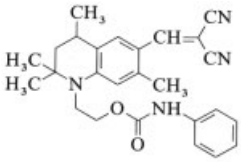

[63467-19-6]

Commercial products include Polycron Yellow 6GP (Holliday), Serilen Brilliant Yellow 6GLS (Yorkshire Chemical), and Sumikaron Yellow 7GTH (Sumitomo).

The above-mentioned dyes are used as the yellow component for producing colored copies in sublimation transfer printing (video printing) [380]. The more sublimation-fast representatives are especially useful; on account of the high energy density at the thermohead they are transferred from the colored ribbon onto the substrate but then do not diffuse out when the colored copy is stored.

Dyes with two separate chromophore systems are obtained from dialdehydes that are prepared by reacting *N*-hydroxyethyl- or *N*-chloroethyl-aminobenzaldehydes with bifunctional reactants. Coupling is preferably effected via ester groups [381]. The following dye

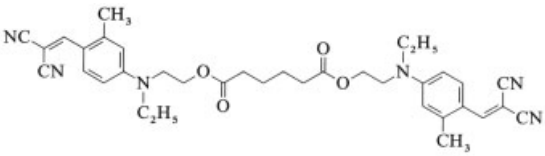

[25857-05-0]

colors polyester and cellulose acetate a brilliant greenish-yellow shade. Commercial products are Terasil Brilliant Yellow 6G and Cibacet Brilliant Yellow 6G (Ciba-Geigy).

Dyes obtained from *N*,*N*-bishydroxy-ethylaminobenzaldehydes and malononitrile are added before or during the polyaddition reaction used to produce polyurethanes. The dyes are bonded covalently to the polyurethanes and diffusion-fast colored foamed plastics are obtained [382]:

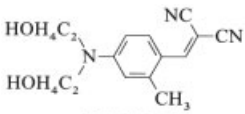

[25665-01-4]

The dyes can also be produced by hydrolyzing the corresponding bisacetoxy compounds [383].

Aminoarylneutrocyanines synthesized from cyanoacetates [384] or malonates [385] can be incorporated into bulk polyesters by condensation reactions.

A cyano substituent on the methine group produces an extremely strong bathochromic shift [386]. The following dye

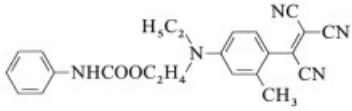

colors polyester fibers a red-violet shade. In order to improve sublimation fastness, dyes with aryl ether [387] and heteroaryl thioether groups [388] in the aminoaryl group have been proposed. Dyes that are derived from tetrahydroquinoline are also industrially important [389]; they can be applied to polyester textiles by transfer printing [390]. These dyes are also used as red components (magenta) in sublimation transfer printing [391].

Dyes containing a N-heteroaryl group instead of the aminoaryl group are also suitable for dyeing polyester a brilliant greenish-yellow shade. Suitable heterocycles include pyridine and thiazole. The following tricyanovinyl dye is used as the red component in video printing:

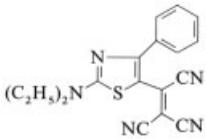

[117555-60-9]

Suitable open-chain methylene compounds include cyanomethyl-substituted N-heterocycles, in particular 2-cyanomethylbenzimidazole [394] and 2-cyanomethyl-4-quinazolin-4-one [395]. On reaction with 4-dialkylaminobenzaldehydes, compounds are formed that dye polyester yellow:

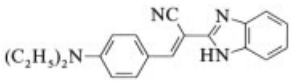

[51553-32-3]

Other suitable condensation components include the cyanomethyl derivatives of 1,3,4-

thiadiazole [396], benzoxazole [397], and benzothiazole [398].

Coumarin Dyes. If the dialkylaminobenzaldehyde bears a hydroxyl group in the 2-position relative to the aldehyde group, ring closure between the hydroxyl and the cyano group occurs, with formation of a coumarin [399]:

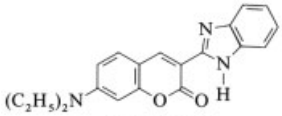

[27425-55-4]

The resultant 3-heteroaryl-7-dialkylaminocoumarins have an intense fluorescent greenish-yellow color and are suitable for dyeing polyester, polyamide and cellulose acetate fibers [400] (→Fluorescent Dyes, Chap. 3.

Other heterocyclic substituents have also been incorporated at the 3-position of coumarin to produce fluorescent dyes (e.g., thiazole [401], benzothiazole, benzoxazole [402], quinazolin-4-one [403], 2*H*-1,2,4-benzothiadiazine-*S,S*-dioxide [404], 1,3,4-oxadiazole [405], and 1,3,4-thiadiazole [406]. All of these dyes color synthetic fibers an intensely fluorescent greenish-yellow shade. Coumarin compounds with a 7-alkoxy or 7-heterocyclic substituent absorb in the shorter wavelength region and have a blue fluorescence. They are used as optical brighteners [407] (→Optical Brighteners). The coumarin dyes include many important laser dyes (→Laser Dyes, Chap. 2.3.).

Coumarinimines are primary products of coumarin ring closure. Red fluorescent dyes are obtained by reaction with ethoxymethylenemalononitrile [408]. Condensation with malononitrile yields a dicyanomethylenechromene derivative [409]. If the heterocyclic substituent includes a free NH group, further ring closure with one of the cyano groups can take place to produce a pentacyclic ring system [410].

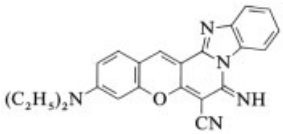

[52372-39-1]

This dye colors polyesters and thermoplasts red with an orange fluorescence.

Coumarin dyes that contain a cyclic methylene compound in the 2-position are obtained by a condensation reaction of β-dicarbonyl compounds, in which one of the carbonyl groups is part of a cyclic system, with dialkylaminosalicylaldehyde [411]. The following dye is obtained by reacting 3-methyl-1-phenyl-4-acetylpyrazol-5-one with 4-diethylaminosalicylaldehyde:

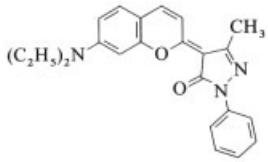

[71494-86-5]

This dye colors polyesters a brilliant yellowish-red shade.

Commercial Products. Setaron Brilliant Flavin 8GFF (Ciba-Geigy), Resolin Brilliant Yellow 10GN (Bayer), Foron Brilliant Flavin S8GF (Sandoz), Sumikaron Brilliant Flavin S-10GL (Sumitomo), Macrolex Fluorescence Yellow 10GN (Bayer), Fluorol Red BK, GK, and Palanil Luminous Red G (BASF).

Red 4-cyanocoumarin dyes with a strong orange to red fluorescence are obtained by addition of sodium cyanide to the 3,4-bond in coumarin, followed by oxidation with bromine or lead tetraacetate [412]. These dyes can be used in solar collectors [413]. 4-Cyanocoumarin compounds with a thiazole group in the 3-position have also been proposed for the same purpose [414]:

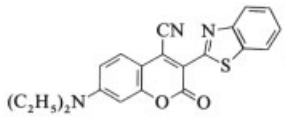

[70546-25-7]

The pentacyclic coumarin dyes described above can also be substituted by a cyano group at the 4-position of the coumarin system. These dyes color polyesters blue [415]. 4-Cyanocoumarin dyes with an alkoxy, sulfonamide, or heterocyclic group instead of the 7-dialkylamino group are yellow fluorescent dyes

that are used to dye plastics or in solar collectors [416].

Cationic Aminoarylneutrocyanines. Cationic aminoarylneutrocyanines contain a quaternary ammonium group bonded via an alkylene group. The positive charge is therefore not part of the mesomeric system. Dyes in which the quaternary ammonium group is bonded via an alkylene group to the aminoaryl nitrogen atom dye polyacrylonitrile in brilliant greenish-yellow shades. Coloring strength is substantially increased if the quaternary ammonium group is bonded to a cyanoacetate moiety [417]. The following dye is used to dye paper in greenish-yellow shades:

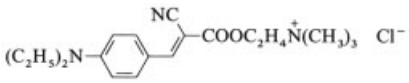

[64992-16-1]

Papers colored with such dyes can be fully bleached during recycling. Cationic charge may also be introduced by a cycloammonium group bonded via an ester group (e.g., the pyridinium [418] or imidazolium group) [419]. The commercial products are Astra Yellow 4G (Bayer), Basazol Yellow 46L, 47L, 51L (BASF), and Albright Yellow BR (American Color and Chem.).

Anionic Coumarin Dyes. Aminoarylneutrocyanine dyes that bear sulfonic acid groups are not sufficiently resistant to hydrolysis in the dye bath. This disadvantage does not apply to coumarin derivatives prepared from 2-hydroxy-4-dialkylaminobenzaldehydes. The following compound dyes polyamide in a fluorescent greenish-yellow shade [420]:

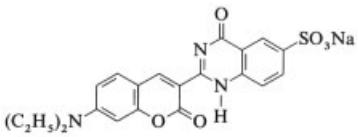

[56330-12-2]

The commercial product is Nylosan Brilliant Flavin E-8G (Sandoz).

Anionic coumarin dyes are also derived from 7-hydroxycoumarin. The anion fluoresces extremely strongly. 3-Benzoxazolyl-7-hydroxycoumarin and the corresponding sulfonic acid are used to produce counterfeit-proof papers [421]. The heteroaryl phosphates obtained from 3-benzimidazolyl-7-hydroxycoumarin on reaction with phosphorus oxychloride can be used for the fluorimetric determination of phosphatase activity in physiological systems [422].

Vinylogous Aminoarylneutrocyanines. The condensation of dialkylaminocinnamaldehydes with malononitrile, cyanoacetates, or benzenesulfonyl acetonitrile yields dyes suitable for dyeing polyester or cellulose acetate. Dyes with a substituted trimethine chain are synthesized by reacting benzonitrile with acetonitrile in the presence of sodium amide to form a ketimine. The latter is reacted with malononitrile to yield the dicyanomethylene compound which is in turn reacted with dialkylaminobenzaldehydes. The resulting dye is used to dye textile fibers and plastics, and also as a sensitizer in electrophotography [423]:

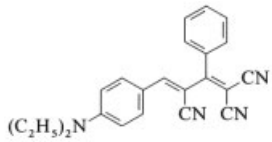

[106577-47-3]

7-Dialkylaminocoumarin-3-aldehyde, which can be obtained by the Vilsmeier reaction of 7-dialkylaminocoumarin, should be regarded as a ring closure derivative of the corresponding dialkylaminocinnamaldehyde. The coumarin aldehyde produces a brilliant greenish-yellow shade in transfer printing on polyester [424]. Condensation of the aldehyde with malononitrile yields a dye that colors polyester in brilliant fluorescent scarlet shades [425]:

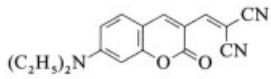

[57818-82-3]

A brilliant greenish-yellow dye is obtained with benzoyl hydrazide [426], and a red fluorescent dye with cyanomethylbenzoxazole [427]. The vinylogs of these compounds are obtained by the Vilsmeier reaction of 3-acetylcoumarins

and reaction with compounds containing a reactive methylene group [428].

10.2.2. Aminoarylneutrocyanines with Cyclic Terminal Groups

Compounds that dye polyesters yellow to orange are obtained by the condensing dialkylaminobenzaldehydes with 1,2-diarylpyrazolidin-3,5-dione [429]. The following dye

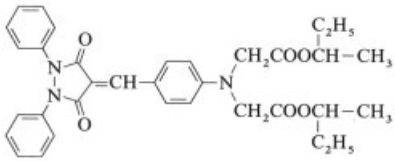

is used as the yellow component for producing colored copies in sublimation transfer printing (video printing) [430]. The products obtained from dialkylaminobenzaldehydes and 1-phenyl-3-dialkylaminopyrazol-5-one may also be used for the same purpose [431]. Replacement of the methine hydrogen atom by a cyano group yields a magenta dye for sublimation copying [432]. Yellow dyes that are used in the charge-generating layer of organic photoconductors are obtained by reaction of dialkylbarbituric acid with aminobenzaldehydes [433]. Here too the methine hydrogen atom can be replaced by the cyano group to produce dyes suitable for dyeing polyester [434]. If *N*-benzoylaminoacetic acid is condensed with terephthaldehyde, an azlactone dye is formed that dyes polyester a fluorescent yellow [435].

A new class of aminoarylneutrocyanine dyes absorbs at particularly long wavelengths due to the dicyanomethylene group in the cyclic terminal system. The compound synthesized from 2-methyl-4-dihexylaminobenzaldehyde and 3-dicyanomethylenebenzothiophene *S,S*-dioxide dyes polyester a brilliant blue shade and has good lightfastness [436]:

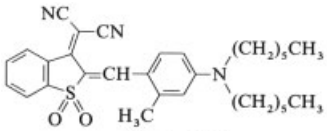

[*122533-66-8*]

Dimeric dyes formed from aminobenzaldehydes coupled via an alkylene bridge have high sublimation fastness [437]. 1-Methyl-2-phenylindole-3-aldehyde produces a red-violet shade in the same dye series [438]. Aminobenzaldehydes containing sulfonic acid groups condense with 3-d ic ya no me th yl en eb en zo th io ph en e-*S,S*-d io x-ide to give blue polyamide dyes [439]. The commercial products are Foron Brilliant Blue S-R (Sandoz) and Kayalon Polyester Brilliant Blue FR-S (Nippon Kayaku).

Pyrrole dyes with an exocyclic methylene group are formed by condensation of dialkylaminobenzaldehydes with dimeric malononitrile, replacement of the methine hydrogen atom by a cyano group, and ring closure [440]:

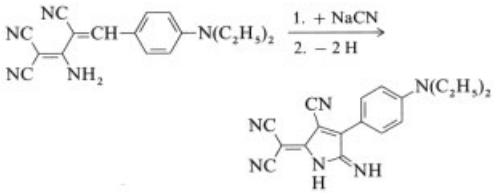

[*112304-76-4*]

These dyes color polyester a brilliant blue shade. They can also be synthesized by reacting 1-alkyl-3-c hl or o-4-c ya no-5-d ic ya no me th yl en ep yr ro li d-2-one with aromatic amines [441]. The dyes are suitable for sublimation thermotransfer printing [442]. Commercial products are Kayalon Polyester Brilliant Blue F2B-S and Kayalon Polyester Blue-Green FG-S (Nippon Kayaku).

Aminoarylazaneutrocyanines. These dyes can be synthesized from 4-nitroso-*N,N*-dialkylanilines and 1-phenyl-3-methylpyrazol-5-one:

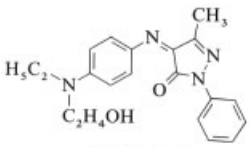

[*4719-52-2*]

They may be used as sublimable dyes for transfer printing [443] and for organic photoconductors used in electrophotography [444].

10.3. Merocyanine Dyes

In merocyanine dyes the nitrogen atom acting as electron donor is part of a heterocyclic ring. The terminal acceptor group may either be in an open chain or a member of a ring.

10.3.1. Merocyanine Dyes with Open-Chain Terminal Groups

If N-heterocyclic carbonyl compounds are condensed with derivatives of cyanoacetic acid, zeromethine dyes are formed. The vinylogous dimethineneutrocyanine dyes are obtained by condensation of the N-heterocyclic methylene-ω-aldehydes with cyanoacetic acid derivatives.

Zeromethine Dyes. If N-alkylnaphtholactam is condensed with cyanoacetic acid derivatives (e.g., malononitrile) in the presence of acid condensation agents, yellow disperse dyes are obtained that dye polyester, polyamide, and polyacrylonitrile in brilliant, lightfast shades [445]. The following dye

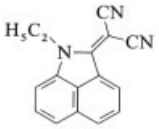

[57232-94-7]

is used in transfer printing [446]. If a thiophenyl group is introduced in the 4-position to the naphtholactam nitrogen atom, the color shifts to a brilliant orange shade. Additional ring closure results in a violet dye for dyeing polystyrene [447]. Five- and six-membered N-heterocycles with cyanomethyl substituents can be condensed with naphtholactam and N-alkylnaphtholactams to give orange and red dyes. With the thiophenyl group in the 4-position of the naphthalene ring, red dyes for polyester materials are obtained by reaction with 2-cyanomethylbenzothiazole. Additional ring closure yields a reddish-blue dye [448]:

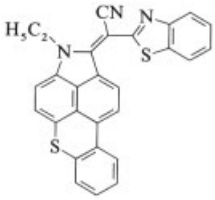

Other suitable starting materials are the cyanomethyl compounds of 1,3,4-oxadiazole [449] and quinoxalinone [450]. Dyes synthesized from naphtholactam and 2-cyanomethylbenzothiazole and containing sulfonic acid groups form zinc complexes that are suitable for spin dyeing polyamide in fluorescent red shades [451].

The zeromethinemerocyanines also include pigments obtained by reaction of aminoiminoisoindolenine with cyanoacetamides. These lightfast yellow pigments are used in paints, lacquers, and plastics [452]. Heat-stable red pigments for coloring polystyrene are obtained with cyanomethylbenzimidazole [453].

Dimethinemerocyanine Dyes. Dimethinemerocyanine dyes are obtained by condensation of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehydes with cyanoacetates or malononitrile in the presence of alkaline catalysts. The yellow dyes are suitable for dyeing synthetic fibers. Introduction of a carboxylate group at the 5-position of the indoline moiety produces dyes that are very lightfast on polyamide fibers [454]:

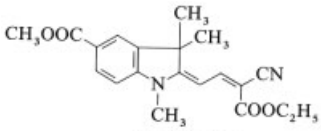

[99750-70-6]

These dyes are also suitable for transfer printing on polyester and polyamide textiles [455].

Neutrocyanine dyes whose methine chain contains a cyano substituent are obtained if 1,3,3-trimethyl-2-cyanomethyleneindoline derivatives are formylated according to the Vilsmeier reaction and then reacted with malononitrile [456]. The dyes can also be obtained by reacting 1,3,3-trimethyl-2-methyleneindoline with tetracyanoethylene [457]. A

commercial product is Resolin Brilliant Yellow P8GLN (Bayer).

10.3.2. Merocyanine Dyes with Cyclic Terminal Groups

These dyes include sensitizers (→Photography) and textile dyes.

Zeromethinemerocyanine Dyes. With acid condensation agents (e.g., phosphorus oxychloride) naphtholactam is converted into the 2-halogen derivative and can then be condensed with compounds containing a reactive methylene group to form zeromethinemerocyanine dyes. Dyes synthesized from naphtholactam and N-alkylated barbituric acids produce lightfast, sublimation-fast yellow dyeings on polyester [458]:

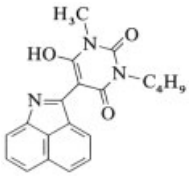

Thioaryl substitution in the 4-position to the naphtholactam nitrogen atom produces a brilliant scarlet shade [459]. If a carboxylate group is introduced into the 4-position relative to the naphtholactam carbonyl group, then condensation with pyrazolones or barbituric acids yields red disperse dyes [460]. A polynuclear heterocycle, in which the partial structure of barbituric acid is recognizable is obtained by reaction of 2-cyanomethylbenzimidazole with isocyanates followed by ring closure. Condensation with naphtholactam yields a dye which dyes polystyrene a brilliant, lightfast scarlet shade [461]:

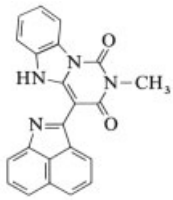

[77365-28-7]

The corresponding compounds with a fused benzothiazole structure dye polystyrene a brilliant red [462]. Red dyes for plastics are also obtained from naphtholactam and derivatives of 1-alkyl-2-hydroxypyrid-6-one [463]. Condensation of naphtholactam with 4-hydroxycoumarin cannot be performed with phosphorus oxychloride because the hydroxyl group is replaced by chlorine. The alternative route via the 4-morpholinoenamine, which condenses with naphtholactam to prepare dyes that color polyester in brilliant fluorescent yellow, orange, and red shades, is therefore employed [464]:

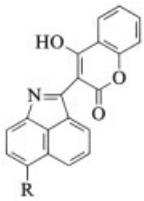

R = H: yellow [65626-20-2]
R = Br: orange [65625-85-6]
R = SC_6H_5: red [65626-05-3]

The zeromethinemerocyanines also include pigments that are obtained by condensation of aminoiminoisoindolenine with cyclic compounds containing a reactive methylene group. Derivatives of barbituric acid [465] and pyrazolone [466] are preferably used. The dye

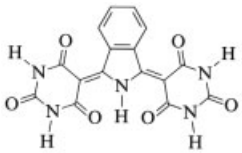

[36888-99-0]

is a brilliant yellow pigment used for coloring poly(vinyl chloride), polyethylene, and paints.

Quinophthalone Dyes. Quinophthalones are a particularly important class of zeromethinemerocyanine dyes [467]. They take their name from the oldest member of this series, which was prepared by E. Jakobsen in 1882 by melting phthalic anhydride with the quinaldine-containing quinoline bases from coal tar. The dye exists in the tautomeric form that is stabilized by hydrogen bonds:

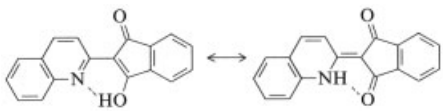

[83-03-9] [83-08-9]

The unsubstituted dye was marketed as alcohol-soluble Quinoline Yellow (C.I. 47 000) and was used for coloring wax, paraffin, and alcoholcontaining paints. The sulfonated dye (C.I. 47 005) was previously widely used for wool, silk, and paper dyeing as well as in the graphics industry. Quinoline Yellow KT (C.I. 47 035) is the sulfonated quinophthalone of 5,6-benzoquinaldine and phthalic anhydride [468].

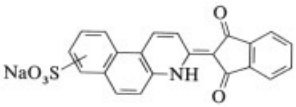

The benzothiazole-substituted quinaldine obtained from dehydrothiotoluidinesulfonic acid was condensed with phthalic anhydride to give a quinophthalone dye that is marketed by Bayer as Levacell Brilliant Yellow 5G for dyeing paper. Condensation in the presence of quinoline yields a purer dye in higher yield [469].

The lightfastness of the dyes is considerably improved by introducing a hydroxyl group into the 3-position of the quinoline system, because the system is stabilized by a further hydrogen bond [470].

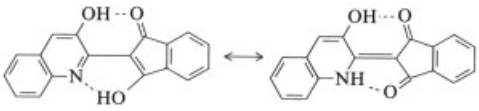

[*7576-65-0*]
C.I. Disperse Yellow 54

3-Hydroxyquinophthalone, produced by G. Kranzlein in 1934 as a precursor for Supra Light Yellow GGL, proved to be an excellent dye for dyeing poly(ethylene terephthalate) and plastics in bulk. The compound is obtained by condensation of 2-methyl-3-hydroxyquinoline-4-carboxylic acid with phthalic anhydride and decarboxylation. Particularly pure dyes are obtained if the conversion is carried out in N,N-dialkylanilines as solvent [471].

The *commercial products* of Disperse Yellow 54 (C.I. 47 020) are as follows: Palanil Yellow 3GE, Thermoplast Yellow 066 (BASF), Samaron Yellow 3GL (Hoechst), Terasil Yellow 2GW, Teraprint Yellow 2G (Ciba-Geigy), Resolin Yellow 4GL (Bayer), Dianix Yellow 3G-E (Mitsubishi), Kayaset Yellow A-G (Nippon Kayaku), Miketon Polyester Yellow F3G (Mitsui), Seriplas Yellow 3GL (Yorkshire Chemical), Sumiplast Yellow HLR (Sumitomo), Vondispers Yellow A-3G (Vondelingenplaat), and Waxoline Yellow 3GP-FW (ICI).

3-Hydroxyquinophthalone and its substitution products are used as yellow components in transfer printing on polyester textiles as well as for producing colored copies in sublimation transfer printing [472]. A wool dye of outstanding lightfastness is obtained by sulfonation of 3-hydroxyquinophthalone. Commercial products are Basovit Yellow 095E (BASF) and Quinoline Yellow 70 (Hoechst).

The wet fastness and sublimation fastness of 3-hydroxyquinophthalone can be improved by a halogen substituent. Pure 4-bromo-3-hydroxyquinophthalone is obtained by reaction of 3-hydroxy-quinophthalone with bromine in 80 – 100 % sulfuric acid at low temperature (C.I. Disperse Yellow 64) [*10319-14-9*] [473]. In dilute sulfuric acid, however, bromine reacts at 100 – 105 °C to form more highly brominated compounds [474].

The *commercial products* of C.I. Disperse Yellow 64 are as follows: Samaron Yellow H3GL (Hoechst), Palanil Yellow 3G, Thermoplast Yellow 103 (BASF), Resolin Yellow 3GL (Bayer), Foron Yellow SE-3GL (Sandoz), Dianix Yellow HG-SE (Mitsubishi), Kayacelon Yellow E-3GL (Nippon Kayaku), Miketon Polyester Yellow RSE (Mitsui), and Serilen Yellow V4GL (Yorkshire Chemical).

Ethers and thioethers of 3-hydroxyquinophthalone are obtained by reaction of 4-halogen-3-hydroxyquinophthalones with aliphatic, araliphatic, or aromatic hydroxy or thio compounds [475]. A commercial product is Thermoplast Yellow R (BASF).

An ether group can also be introduced by using 4-hydroxyphthalic anhydride as condensation component and subsequently etherifying the hydroxyl group. Etherification with halides of 1,3,5-triazine followed by replacement of one or both halogen atoms by aliphatic amines yields dyes that are suitable for dyeing textiles containing both cotton and synthetic fibers [476]. The triazine ether group may also be bonded to the quinophthalone dye via a carboxylic acid anilide group [477].

Alkanesulfonates and in particular benzenesulfonates of 4-hydroxyphthalic anhydride yield, when reacted with 3-hydroxyquinaldine-

4-carboxylic acid, sublimation-fast, yellow dyes with high affinity for polyesters [478]. Sulfonates and sulfonamides of quinophthalone are obtained from the corresponding sulfochloride by reaction with phenols [479] and aliphatic amines [480]. They are sublimation-fast dyes with high affinity for polyesters.

Quinophthalone dyes with a carboxyl substituent in the carbocyclic ring system are obtained by condensation of trimellitic anhydride with 3-hydroxyquinaldine-4-carboxylic acid. A particularly suitable solvent for this reaction is sulfolane; the dye can be esterified at the carboxyl group without intermediate separation [481]. The dyes are suitable for dyeing polyester and for producing colored copies in sublimation transfer printing [482]. A commercial product is Sumikaron Yellow SE-RPD (Sumitomo).

Quinophthalone dyes synthesized from benzenetetracarboxylic acid, whose free carboxyl groups are esterified by long-chain alkyl groups, are used in liquid crystal displays [483]. Quinophthalone dyes have also been prepared from naphthalene-2,3-dicarboxylic acid [484]. Commercial products are Dialight Yellow GRM and Diaresin Yellow 5110 (Mitsubishi).

Introduction of cationic substituents yields watersoluble dyes that dye paper in brilliant yellow shades with good lightfastness. The cationic groups are introduced by reacting quinophthalone dyes with formaldehyde and 1-alkylimidazole [485] or by condensing with chloroacetylmethylolamide (Einhorn reaction), the chlorine being replaced by pyridine [486].

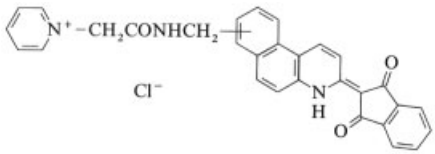

Quinophthalones synthesized from tetrachloro- or tetrabromophthalic anhydride and derivatives of quinaldine are suitable for dyeing plastics (e.g., polyolefins, polystyrene, polyacrylates) [487].

8-Aminoquinaldine is particularly suitable for producing pigments because it can react with 2 mol of tetrachlorophthalic anhydride. The amino group forms an imide, and the methyl group reacts to form the quinophthalone [488]. 8-Aminoquinaldine can also be converted into the imide with naphthalic acid prior to reaction with tetrachlorophthalic anhydride [489].

A particularly suitable starting material for producing pigments is 3-hydroxyquinophthalone carboxyloyl chloride. Reaction with aromatic mono- or diamines gives pigments that dye poly(vinyl chloride) yellow with good migration fastness and lightfastness [490].

Condensation of pyromellitic anhydride (1,2,4,5-benzene tetracarboxylic acid anhydride) with equimolar amounts of a quinaldine yields quinophthalone dicarboxylic acids which, after conversion into the anhydride, can react to form dyes containing imide groups. Monoamines yield pigments for paints and printing inks [491]:

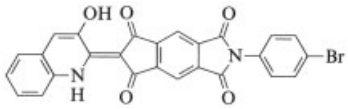

[58130-76-0]

Diamines such as 1,4-diaminobenzene [492], 2,6-diaminonaphthalene [493], 4,4'-diaminodiphenylsulfone [494], and dichlorobenzidine [495] yield migration-stable yellow pigments for dyeing poly(vinyl chloride).

Mixed quinophthaloneisoindolenine pigments are obtained by reaction of 8-aminoquinaldine with aminoisoindolenine derivatives, followed by condensation with tetrachlorophthalic anhydride. The following pigment dyes poly(vinyl chloride) a lightfast and migration-fast orange shade [496]:

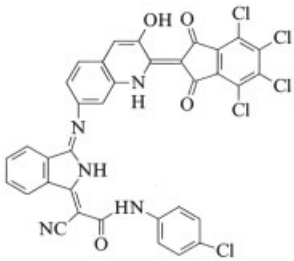

Dimethineneutrocyanine Dyes. Neutrocyanine dyes suitable as the yellow component for producing colored copies in sublimation transfer printing are obtained by reaction of 1,3,3-trimethyl-2-methyleneindoline-ω-

aldehyde with 1-phenyl-3-methylpyrazol-5-one [497] and 1-phenyl-3-dimethylaminopyrazol-5-one [498]. The lightfastness of these dyes on polyamide is significantly improved by introducing a carboxylate group at the 5-position of the indole moiety [499]:

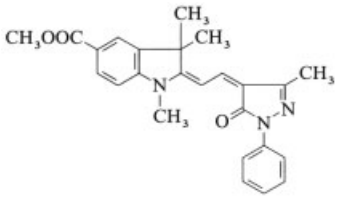

[5718-26-3]

The dye is also suitable for the transparent dyeing of polystyrene and poly(methyl methacrylate) in brilliant orange shades [500]. Reaction of 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde with 1-alkyl-3-cyano-4-methyltetrahydropyridine-2,6-dione produces brilliant red shades with poly(ethylene terephthalate) [501]. Dimethineneutrocyanines with benzothiazole and quinaldine as heterocyclic terminal system are used as organic photoconductors in electrophotography [502].

Dimethineneutrocyanine dyes from cyclic compounds with a reactive methylene group bearing a dicyanomethylene group absorb at long wavelengths. The compound synthesized from 1,3,3-trimethyl-2-methyleneindoline-ω-aldehyde and 3-dicyanomethylenebenzothiophene *S,S*-dioxide dyes polyester a reddish-blue shade [503]:

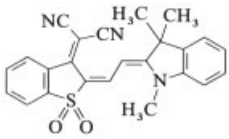

[115281-48-6]

Dyes of this series are also suitable for producing nonerasable discs for optical information recording [504].

11. References

General References

1. A. Weissberger: *The Chemistry of Heterocyclic Compounds,* **vol. 18,** Wiley Interscience Publishers, New York-London 1964.
2. F. M. Hamer: *The Cyanine Dyes and Related Compounds,* Wiley Interscience Publishers, New York-London 1964.
3. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 4,** Academic Press, New York-London 1971.
4. H. Meier: *Die Photochemie der organischen Farbstoffe,* Springer Verlag, Berlin-Göttingen-Heidelberg 1963.
5. W. Foerst: "Optische Anregung organischer Systeme," *2. Internationales Farbensymposium, Ellmau 1964,* Verlag Chemie, Weinheim 1966.
6. *Ullmann,* 4th ed., **16,** 635 – 669.
7. M. Okawara, T. Kitao, T. Hirashima, M. Matsuoka: "Organic Colorants," *A Handbook of Data of Selected Dyes for Electro-Optical Applications,* Kodansha Tokyo, Elsevier, Amsterdam-Oxford-New York-Tokyo 1988.
8. H. Zollinger: *Color Chemistry, Syntheses, Properties and Applications of Organic Dyes and Pigments,* Verlag Chemie, Weinheim-New York 1987.

Specific References

9. S. Dähne, *Z. Chem.* **5** (1965) no. 12, 441 – 451.
10. Ciba-Geigy, EP 112 282, 1982 (P. Erzinger, A. R. Läpple). Bayer, EP 56 947, 1981 (R. Foppe, R. Raue, H.-P. Kühlthau).
11. GAF Corp., US 3 985 500, 1974 (C. Streck).
12. Sumitomo Chem., JP 52 37 885, 1975.
13. Ciba-Geigy, EP 114 574, 1982 (A. R. Läpple, A. Nicopoulos).
14. Sumitomo Chem., JP 55 13 732, 1978
15. Hoechst, DE-OS 3 438 101, 1984 (M. Hähnke et al.).
16. Ciba-Geigy, DE-OS 2 548 009, 1974 (J. Koller et al.); DE-OS 2 943 750, 1978 (J. Haase, H. U. Berendt). Hoechst, DE-OS 2 716 246, 1977 (N. Kollodzeiski et al.).
17. Ciba-Geigy, EP 57 457, 1981 (H. Herrmann, R. Jenny, M. Metter).
18. Prod. Chem. Ugine Kuhlmann, DE-OS 2 504 997, 1974 (F. Manier, P. Caullet).
19. Hoechst, DE-OS 2 433 233, 1974 (M. Hähnke, E. Fleckenstein); DE-OS 2 822 913, 1978 (M. Hähnke, R. Mohr, K. Hohmann). Ciba-Geigy, EP 55 224, 1980 (P. Loew).

20. ICI, GB 2 033 401, 1978 (B. Parton).
21. Hoechst, DE-OS 3 546 294, 1985 (K. Hohmann, P. Mischke, W. Teige).
22. Bayer, DE-OS 3 301 024, 1983 (H. Bara, M. Wiesel).
23. Sumitomo Chem., US 3 942 946, 1974; US 3 958 287, 1974; JP 50 145 680, 1974.
24. BASF, DE-OS 3 103 110, 1981 (R. Dyllick-Brenzinger et al.).
25. Ciba-Geigy, DE-OS 2 945 686, 1978 (R. Bäuerle). Kodak, GB 2 039 541, 1979 (D. Mullen, A. W. McCann). Mitsubishi, JP 60 53 564, 1983. Matsushita Elec., JP 60 223 862, 1984. Sumitomo Chem., JP 60 253 596, 1984.
26. Bayer, DE-OS 2 135 156, 1971 (A. Brack, H. Psaar).
27. J. Vavrova, M. Svitak, H. Kralove, CS 187 293, 1977.
28. Mobil Finishes, SU 732 246, 1977.
29. Bayer, FR 1 490 350, 1965.
30. Minnesota Mining, US 4 594 312, 1984 (G. J. Sabongi et al.).
31. Bayer & Co, DE 218 904, 1908 (W. König).
32. IG Farben, DE 725 290, 1933 (W. König, W. Alt).
33. Bayer, DE-AS 1 108 659, 1956 (R. Raue, W. Müller).
34. Ciba-Geigy, DE-OS 2 718 225, 1976 (R. Garner, B. Greenmount).
35. Matsushita Electric, EP 97 493, 1982.
36. Bayer, DE-OS 2 036 505, 1970 (A. Brack).
37. IG Farben, DE 686 198, 1935 (K. Winter et al.)
38. BASF, DE-OS 2 735 263, 1977 (H. Kast).
39. Hoechst, FR 2 011 334, 1968.
40. Nihon Kasei Kogyo, JP 47 32 183, 1971.
41. ICI, DE-OS 2 658 638, 1975 (J. Hodgkinson et al.).
42. VEB Chemiekombinat Bitterfeld, DD 244 348, 1985 (H. Noack et al.).
43. Ciba-Geigy, EP 38 299, 1980 (W. Stingelin, P. Loew); EP 94 642, 1982 (W. Stingelin et al.); EP 145 656, 1983 (W. Stingelin et al.).
44. Ciba-Geigy, EP 74 926, 1981 (W. Stingelin, P. Loew). VEB Chemiekombinat Bitterfeld, DD 240 213, 1985 (H. Noack et al.).
45. BASF, EP 273 307, 1986 (J. P. Dix et al.).
46. BASF, DE-OS 2 839 379, 1978 (K. Grychtol).
47. Bayer, DE-OS 2 352 247, 1973 (A. Brack, R. Raue).
48. IG Farben, DE 710 750, 1933 (P. Wolff).
49. Du Pont, DE-OS 2 002 609, 1969 (F. R. Hunter). Bayer, DE-OS 2 130 790, 1971 (H. Psaar).
50. Sandoz, DE-OS 2 917 996, 1978 (B. Gertisser). Ciba-Geigy, DE-OS 2 926 183, 1978 (P. Loew, R. Zink).
51. Du Pont, DE-OS 1 956 158, 1968 (F. R. Hunter).
52. Bayer, DE-OS 1 945 053, 1969 (A. Brack).
53. Bayer, DE-OS 3 322 318, 1983 (K. Meisel, R. Raue).
54. Bayer, DE-OS 2 353 537, 1973 (H. Psaar, R. Raue).
55. Bayer, DE-AS 1 044 588, 1956 (R. Raue).
56. Ciba, DE 1 004 748, 1954 (R. Ruegg).
57. Ciba, DE 1 061 931, 1954 (R. Ruegg).
58. Bayer, DE-OS 2 013 791, 1970 (R. Raue, H.-L. Dorsch).
59. Bayer, DE 1 083 000, 1957 (R. Raue).
60. Bayer, EP 42 556, 1980 (R. Raue, H.-P. Kühlthau).
61. BASF, DE-OS 3 131 531, 1981 (U. Mayer, G. Henze, H. Schmidt).
62. Ciba-Geigy, EP 272 212, 1986 (R. Ronco, E. Ciapponi, P. Pessonneaux).
63. Bayer, DE-OS 2 931 772, 1979 (R. Raue, H. Giesecke).
64. Bayer, DE-OS 3 013 272, 1980 (H. Gleinig et al.); DE-OS 3 013 273, 1980 (H. Gleinig et al.).
65. VEB Chemiekombinat Bitterfeld, DD 145 635, 1979 (H. Noack et al.).
66. Bayer, DE-OS 3 545 605, 1985 (H.-L. Dorsch et al.).
67. Hoechst, DE-OS 2 945 028, 1979 (R. Mohr, M. Hähnke).
68. Bayer, DE 1 133 053, 1959 (R. Raue).
69. Bayer, EP 87 706, 1982 (R. Raue).
70. BASF, DE 1 150 474, 1960 (H. Baumann et al.).
71. BASF, GB 1 186 745, 1966.
72. Bayer, DE 1 233 520, 1963 (R. Raue, A. Brack, E.-H. Rohe).
73. Hodogaya Chem., JP 55 102 654, 1979.
74. Bayer, DE-OS 3 235 613, 1982 (H. Fürstenwerth).
75. Bayer, DE 1 150 475, 1959 (R. Raue); DE 1 133 054, 1959 (R. Raue).
76. Ciba-Geigy, DE-OS 2 548 009, 1974 (J. Koller et al.).
77. Ciba-Geigy, DE-OS 2 620 790, 1975 (C. Frey, P. Moser). Bayer, EP 65 717, 1981 (H. Giesecke).
78. Ciba-Geigy, US 4 237 275, 1978 (R. Bader, D. Reinehr).
79. Bayer, DE 1 209 679, 1961 (R. Raue, E.-H. Rohe).
80. Bayer, DE-OS 2 726 437, 1977 (K.-H. Hermann).
81. Geigy, BE 579 303, 1958.

82. Bayer, EP 53 751, 1980 (H. Fürstenwerth).
83. Hodogaya Chem., JP 71 35 630, 1968.
84. IG Farben, DE 714 882, 1939.
85. Cassella, DE 51 738, 1888 (J. Rosenheck).
86. O. N. Chupakhin et al., *Zh. Org. Khim.* **12** (1976) no. 7, 1553 – 1557.
87. V. N. Charushin et al., *Zh. Org. Khim.* **14** (1978) no. 1, 140 – 146.
88. O. N. Chupakhin et al., *Dokl. Akad. Nauk. SSSR* 239 (1978) no. 3, 614 – 616.
89. Bayer, DE-OS 2 721 190, 1977 (A. Brack).
90. IG Farben, DE 483 234, 1926 (G. Kränzlein, M. Heyse, P. Ochwat).
91. Bayer, DE 1 190 126, 1961 (A. Brack, D. Kutzschbach).
92. Bayer, DE-OS 2 138 029, 1971.
93. BASF, DE-AS 1 569 645, 1966 (H. Baumann, G. Hansen).
94. Bayer, DE-OS 2 351 296, 1973 (H.-P. Kühlthau). DE-OS 2 557 503, 1975 (H.-P. Kühlthau).
95. Bayer, DE-OS 3 148 104, 1981 (H.-P. Kühlthau, H. Harnisch).
96. Bayer, DE-OS 2 747 110, 1977 (A. Brack, H.-P. Kühlthau).
97. Ciba-Geigy, DE-OS 2 233 261, 1972.
98. Ciba-Geigy, DE-OS 3 321 130, 1982 (R. Zink).
99. Sterling Drug, EP 61 128, 1981 (P. J. Schmidt et al.). Matsushita Electric, JP 59 71 898, 1982.
100. Ricoh, JP 59 75 257, 1982. Canon, JP 60 195 550, 1984.
101. MLB Hoechst, DE 394 744, 1922 (R. Schuloff).
102. Bayer, DE-OS 2 932 092, 1979 (H. Beecken).
103. Bayer, DE-OS 3 012 599, 1980 (H. Beecken).
104. BASF, DE-OS 3 149 047, 1981 (R. Dyllick-Brenzinger).
105. BASF, DE-OS 3 136 259, 1981 (R. Dyllick-Brenzinger, K. Grychtol).
106. BASF, DE-OS 2 901 845, 1979 (H.-J. Degen, K. Grychtol).
107. BASF, DE-OS 2 908 921, 1979 (H.-J. Degen, K. Grychtol).
108. BASF, DE-OS 2 852 919, 1978 (H.-J. Degen).
109. Ciba-Geigy, EP 176 473, 1984 (P. Möckli).
110. Ciba-Geigy, EP 176 472, 1984 (P. Möckli).
111. Ciba-Geigy, EP 176 474, 1984 (P. Möckli).
112. IG Farben/Bayer, DE 711 665, 1935.
113. IG Farben/Bayer, DE 891 120, 1935 (W. Müller, C. Berres).
114. IG Farben, DE 721 020, 1935 (K. Winter, N. Roh).
115. Bayer, EP 74 569, 1981 (R. Raue, V. Hühne, H.-P. Kühlthau).
116. Bayer, DE-OS 3 213 966, 1982 (R. Raue, H.-P. Kühlthau).
117. Bayer, EP 89 566, 1982 (R. Raue, H.-P. Kühlthau, K.-F. Lehment).
118. Bayer, DE-OS 2 243 627, 1972 (H. Psaar).
119. IG Farben, DE 721 020, 1935 (K. Winter, N. Roh).
120. IG Farben, DE 744 019, 1939 (B. Eistert, G. Kochendörfer). Bayer, DE-OS 2 135 834, 1971 (A. Brack, W. Gomm).
121. Canon, JP 63 74 069, 1986.
122. Fuji Photo, US 4 708 922, 1985; DE-OS 3 503 480, 1984 (S. Kondo et al).
123. IG Farben, DE 742 039, 1938 (G. Kochendörfer et al.). Sandoz, DE-OS 2 101 223, 1970 (R. Entschel, W. Käppeli).
124. Nippon Kayaku, JP 74 04 530, 1970.
125. Bayer, DE 835 172, 1950.
126. Bayer, DE 1 569 606, 1967 (A. Brack).
127. Ciba-Geigy, EP 37 374, 1980 (R. Zink).
128. Vickers PLC, EP 112 169, 1982 (A. J. Chalk et al.).
129. IG Farben, DE 693 025, 1936 (W. Müller, O. Wahl, W. Teupel).
130. Bayer, BE 606 907, 1960.
131. Mitsubishi Chem., US 4 756 987, 1985 (S. Maeda et al.).
132. BASF, DE-OS 2 902 763, 1979 (H.-J. Degen, K. Grychtol); DE-OS 2 733 468, 1977 (H.-J. Degen, K. Grychtol).
133. Hitachi, EP 92 255, 1982 (A. Kakuta et al.).
134. Hitachi, JP 88 20 344, 1980.
135. Hitachi, JP 57 185 045, 1981; JP 58 120 257, 1982.
136. Hodogaya Chem., JP 52 114 788, 1976.
137. Sumitomo Chem., JP 50 142 876, 1974.
138. Matsushita Electric, JP 60 200 233, 1984; JP 61 147 235, 1984.
139. Nippon Kanko Shikis, JP 57 14 652, 1980; JP 52 113 999, 1976.
140. Bayer, DE-OS 3 622 009, 1986 (H. Psaar, R. Raue).
141. Bayer, DE-OS 2 510 238, 1975 (H. Psaar).
142. Hodogaya Chem., DE-OS 2 516 383, 1974 (M. Otsuzumi et al.); JP 50 137 212, 1974.
143. Nippon Photosensitive Paper, JP 61 258 867, 1985.
144. Fuji, DE-OS 1 961 921, 1968 (Ch. Osada et al.).
145. Fuji, JP 47 21 128, 1971. R. Guglielmetti, *J. Photogr. Sci.* **22** (1974) 77 – 83.
146. Sony Corp., JP 57 59 956, 1980. Matsushita Electric, EP 179 436, 1984 (J. Myazaki et al.); JP 62 147 454, 1985; US 4 737 427, 1985.
147. E. Goldburt et al., *Macromolecules* **17** (1984) no. 6, 1225 – 1230.

148. Nippon Kanko Shikiso, JP 74 45 229, 1969.
149. G. Didier et al., *Bull. Soc. Chim. Fr.* **1981** (1-2 Pt. 2) 14 – 18. Agency of Ind. Sciences and Techn., JP 63 54 377, 1986. Sony, JP 59 227 972, 1983.
150. BASF, DE-OS 2 230 225, 1972 (A. Oberlinner et al.). Kyodo Print., JP 53 53 675, 1976.
151. 3 M, GB 1 315 825, 1969.
152. Fuji, JP 52 46 079, 1975.
153. B. S. Lukyanov et al., *Khim. Geterotsikl. Soedin.* 1978 no. 2, 161 – 166.
154. V. A. Anisimova et al., *Khim. Geterotsikl. Soedin.* 1978 no. 1, 108 – 112.
155. BASF, DE-OS 2 430 568, 1974 (H. Baumann, A. Oberlinner); DE-OS 2 737 207, 1977 (H. Baumann, A. Oberlinner). Fuji Photo, JP 55 44 830, 1978. F. Feichtmayr, A. Oberlinner, *Liebigs Ann. Chem.* 1979 no. 9, 1337 – 1345.
156. Matsushita Electric, JP 62 146 978, 1985.
157. Sony, JP 62 299 386, 1986.
158. Sony, EP 184 808, 1984.
159. Sony, JP 60 177 089, 1984.
160. Geigy, DE 1 098 125, 1958 (H. Häusermann, J. Voltz).
161. BASF, DE-OS 2 005 968, 1970 (J. Dehnert, G. Grau). Hodogaya Chem., JP 54 145 735, 1978.
162. Yamamoto Kagaku, JP 50 39 733, 1973.
163. K. K. Shinroihi, JP 50 48 052, 1973.
164. Sumitomo Chem., JP 48 00 723, 1971.
165. Ciba-Geigy, DE-OS 2 142 411, 1970 (S. Koller et al.).
166. BASF, DE-OS, 2 363 459, 1973 (K. Grychtol).
167. A. Kaufmann, DE 243 078, 1911.
168. Matsushita Electric, JP 58 48 030, 1981; JP 61 147 236, 1984.
169. Fr. Schiller Univers. Jena, DD 238 611, 1985 (M. Reichenbächer, P. Czerney). Seiko Epson, JP 63 267 785, 1987.
170. Seiko Epson, JP 63 267 783-4, 1987. Pilkington, EP 245 020, 1986 (M. Rickwood, J. P. Hepworth).
171. Geigy, DE 1 044 023, 1954 (W. Bossard, J. Voltz, F. Favre); DE 1 050 940, 1954 (W. Bossard, J. Voltz, F. Favre).
172. BASF, DE 1 619 328, 1966 (H. R. Hensel, M. Rosenkranz).
173. Hoechst, DE-AS 2 733 178, 1977 (R. Neeb, R. Mohr).
174. VEB Chemiekombinat Bitterfeld, DD 139 663, 1978 (A. E. Jauer et al.). Bayer, DE-OS 2 616 412, 1976 (R. Raue, H.-P. Kühlthau).
175. A. P. D'Rozario et al., *J. Chem. Soc. Perkin Trans. 2* 1987 no. 12, 1785 – 1788.
176. Bayer, DE-OS 2 255 058, 1972 (H.-P. Kühlthau); DE-OS 2 344 901, 1973 (H. P. Kühlthau).
177. Bayer, DE-OS 2 255 059, 1972 (H. P. Kühlthau); DE-OS 2 344 672, 1973 (H. P. Kühlthau).
178. Bayer, DE-OS 2 409 543, 1974 (W. Kruckenberg). Ciba-Geigy, EP 105 031, 1982 (R. Zink).
179. Bayer, DE-OS 2 559 737, 1975 (W. Kruckenberg).
180. Sandoz, DE-OS 2 803 288, 1977 (B. Henzi, V. Käppeli). ICI, DE-OS 2 556 917, 1974 (D. M. Fawkes, J. L. Leng).
181. Bayer, DE-OS 2 440 021, 1974 (H.-P. Kühlthau).
182. Bayer, DE-OS 2 401 246, 1974 (H.-P. Kühlthau, H. Beecken).
183. VEB Chemiekombinat Bitterfeld, DD 268 964, 1985 (H. Noack et al.).
184. Ciba-Geigy, DE-OS 3 234 785, 1981 (V. Ramanathan).
185. Canon, JP 61 107 251, 1984.
186. Hodogaya, JP 74 21 285, 1970.
187. Daito Kagaku, JP 72 50 210, 1970.
188. Hodogaya Chem., US 3 132 132, 1960. Sumitomo Chem., DE-OS 2 136 974, 1970 (M. Ohkawa et al.); JP 73 06 931, 1970 (M. Ohkawa et al.); JP 73 01 997, 1970.
189. Sandoz, CH 512 565, 1969.
190. Asahi Kagaku, JP 73 30 990, 1967. Sumitomo Chem., DE-OS 2 129 271, 1970. Daito Kagaku, JP 73 00 165, 1970. Hodogaya Chem., JP 49 83 727, 1972.
191. Sumitomo Chem., BE 825 455, 1974. ICI, GB 2 070 050, 1980 (P. Gregory, D. Thorp).
192. Hoechst, DE-OS 2 840 121, 1978 (R. Neeb, K. Hohmann, R. Mohr). Bayer, DE-OS 2 758 107, 1977 (H.-L. Dorsch, K.-H. Hermann).
193. Hoechst, DE-OS 2 822 913, 1978 (M. Hähnke, R. Mohr, K. Hohmann).
194. Sumitomo Chem., GB 1 186 753, 1966.
195. Ciba-Geigy, CH 664 577, 1985 (P. Moser); DE-OS 3 540 090, 1984 (P. Moser).
196. Sandoz, DE-OS 3 029 746, 1979 (B. Gertisser).
197. Sandoz, CH 643 285, 1980 (B. Gertisser).
198. Geigy, FR 1 145 751, 1955.
199. BASF, DE-AS 1 117 233, 1955 (H. Baumann, S. Hünig).
200. Hodogaya Chem., JP 51 56 828, 1974. Sandoz, DE-OS 2 509 095, 1974 (B. Henzi).
201. Bayer, DE-OS 3 009 267, 1980 (H.-P. Kühlthau).
202. VEB Chemiekombinat Bitterfeld, DD 208 474, 1982 (H. Noack et al.).
203. Hoechst, DE-OS 3 602 587, 1986 (K. Hohmann, P. Mischke).

204. Ciba-Geigy, EP 45 719, 1980 (R. Zink, P. Loew).
205. Bayer, EP 286 971, 1987 (A. Brack).
206. Hoechst, DE-OS 3 438 101, 1984 (M. Hähnke et al.).
207. Hoechst, EP 44 059, 1980 (M. Hähnke, R. Mohr, K. Hohmann).
208. Yamada Kagaku, JP 73 09 094, 1968. Sumitomo Chem., JP 76 07 173, 1971. Hodogaya Chem., JP 51 18 738, 1974.
209. Hodogaya Chem., JP 51 23 516, 1974.
210. Hodogaya Chem., JP 51 56 828, 1974.
211. B. Hirsch et al., DD 216 039, 1982.
212. Hodogaya Chem., DE-OS 2 104 624, 1970 (M. Otsuzumi et al.).
213. Hodogaya Chem., DE-OS 2 022 624, 1969 (M. Otsuzumi et al.).
214. Bayer, DE-OS 2 811 258, 1978 (H. Fürstenwerth).
215. Geigy, BE 666 693, 1964. Prod. Chim. Ugine Kuhlmann, EP 1 731 1977 (M. Champenois).
216. BASF, DE-OS 2 234 348, 1972 (M. Eistert, K. Schmeidl). Yorkshire Chem., GB 1 591 532, 1977 (K. Bramham, M. Robert).
217. Ciba-Geigy, DE-OS 3 544 574, 1984 (W. Stingelin, P. Moser).
218. Montedison, DE-OS 2 750 952, 1976 (R. Battisti et al.). M. V. Gorelik et al., SU 897 813, 1980.
219. ICI, GB 1 541 826, 1975 (D. M. Fawkes). Ciba-Geigy, DE-OS 2 639 179, 1975 (S. Koller, R. Zink, P. Loew).
220. Ciba-Geigy, EP 15 872, 1979 (R. Zink).
221. BASF, DE-AS 1 222 185, 1961 (J. Dehnert, H. Baumann). ICI, GB 1 508 056, 1975 (D. M. Fawkes).
222. FMC, DE-AS 1 227 584, 1960 (R. F. M. Sureau et al.).
223. Ugine Kuhlmann, DE-AS 1 544 471, 1965 (G. V. H. Kremer, R. F. M. Sureau).
224. Sandoz, DE-OS 2 819 197, 1977 (B. Henzi). Bayer, DE-OS 2 837 908, 1978 (H.-P. Kühlthau). Sandoz, DE-OS 2 927 205, 1978 (B. Henzi, U. Blass).
225. BASF, DE-AS 1 137 816, 1958 (H. Baumann, J. Dehnert).
226. Bayer, DE-OS 2 837 953, 1978 (H.-P. Kühlthau); DE-OS 2 837 987, 1978 (H.-P. Kühlthau); DE-OS 2 908 135, 1979 (H.-P. Kühlthau).
227. Du Pont, US 3 649 162, 1970.
228. Am. Cyanamid, US 30 51 697, 1958; US 3 117 116, 1961.
229. Geigy, DE-AS 1 061 923, 1955 (W. Bossard, J. Voltz, F. Favre).
230. S. Motomizu, K. Toei, *Anal. Chim. Acta* **120** (1980) 267 – 277.
231. M. Oshima et al., *Anal. Chim. Acta* **134** (1982) 73 –80.
232. L'Oreal, DE-OS 2 538 363, 1974 (G. Lang, A. Bugaut).
233. Hoechst, FR 1 462 723, 1965.
234. Ugine Kuhlmann, FR 1 378 853, 1963; FR 1 403 670, 1964.
235. ICI, GB 1 508 500, 1975 (D. M. Fawkes); GB 1 533 260, 1975 (D. M. Fawkes).
236. IG Farben, DE 742 039, 1938 (G. Kochendörfer et al.). R. E. Koval'chuk et al., *Ukr. Khim. Zh. (Russ. Ed.)* **42** (1976) no. 11, 1174 – 1178.
237. Nippon Kanko Shikiso, JP 61 136 553, 1984. Fuji Photo, JP 62 186 205, 1986; JP 62 01 754, 1985.
238. Bayer, DE 1 184 882, 1962 (A. Brack).
239. Bayer, BE 671 139, 1964; BE 662 178, 1964.
240. Ciba, DE 1 569 690, 1965.
241. T. V. Stupnikowa et al., *Khim. Geterotsikl. Soedin.* 1980 no. 7, 959 – 964. T. V. Stupnikowa et al., *Khim. Geterotsikl. Soedin.* 1980 no. 10, 1365 – 1368.
242. Bayer, DE-OS 2 520 816, 1975 (H. Psaar).
243. Bayer, DE-OS 2 638 232, 1976 (H. Psaar, W. Kruckenberg).
244. Hoechst, DE 167 159, 1903; DE 167 770, 1903. IG Farben, DE 710 748, 1932 (R. Walter, H. Dürr). H. Meerwein et al., *Justus Liebigs Ann. Chem.* **641** (1961) 19.
245. IG Farben, DE 459 616, 1926 (O. Wahl). R. Kuhn et al., *Ber. Dt. Chem. Ges.* **63** (1930) 3176 – 3184.
246. W. Grahn, *Liebigs Ann. Chem.* 1981, 107 – 121.
247. Kodak, DE 666 041, 1934 (F. M. Hamer); DE 724 608, 1934 (F. M. Hamer).
248. Eastman Kodak, US 4 232 121, 1978 (P. B. Gilman jr. et al.).
249. Ilford, FR 1 014 189, 1950 (J. D. Kendall, D. J. Fry).
250. R. Steiger, *Photogr. Sci. Eng.* **25** (1981) no. 1, 10 – 20.
251. A. Klapproth, K. Dimroth, *Justus Liebigs Ann. Chem.* 1975, 1839 – 1849.
252. Bayer, DE 1 170 569, 1962 (A. Brack).
253. Bayer, DE-OS 2 202 316, 1972 (A. Brack, E. Schmitt).
254. G. A. Reynolds, J. A. van Allan, *J. Heterocycl. Chem.* **15** (1978) no. 7, 1225 – 1226.
255. Canon, JP 62 132 690, 1985.
256. BASF, DE 1 144 280, 1961 (M. Seefelder, H. G. Reppe).

257. BASF, DE 1 089 721, 1959.
258. IG Farben, DE 614 325, 1933 (P. Wolff). DE 615 130, 1933 (P. Wolff).
259. I. K. Khanna, V. Venkataraman, *Indian J. Chem. Sect. B* **16B** (1978) 755 – 759.
260. Comb. Chimic Giurgiu, RU 86 458, 1983.
261. Eastman Kodak, US 3 758 466, 1971 (J. G. Fisher, C. A. Coates).
262. BASF, DE-OS 2 008 677, 1970 (H. J. Sturm, H. Armbrust).
263. Bayer, DE 1 158 646, 1959. Yorkshire Dyeware, GB 1 181 951, 1968.
264. Ciba-Geigy, DE-OS 2 711 521, 1976 (P. Möckli).
265. BASF, DE 1 163 775, 1955 (H. Pfitzner et al.).
266. Hodogaya Chem., US 3 136 751, 1960. Sumitomo Chem., DE-OS 2 114 747, 1970 (Y. Tanaka et al.); DE-OS 2 136 974, 1970 (M. Ohkawa et al.).
267. Yamada Kagaku, JP 73 28 394, 1968. Sumitomo Chem., DE-OS 2 129 271, 1970 (M. Ohkawa et al); JP 73 17 727, 1970. Yamada Kagaku, JP 50 19 827, 1973. Hodogaya Chem., JP 52 39 729, 1975.
268. BASF, DE 1 228 586, 1956 (H. Pfitzner et al.); DE 1 205 637, 1960 (H. Pfitzner, H. Baumann).
269. Hodogaya Chem., JP 62 104 874, 1985.
270. Hodogaya Chem., JP 51 18 739, 1974. Sumitomo Chem., JP 48 18 316, 1971.
271. Hodogaya Chem., JP 50 156 540, 1974.
272. Hodogaya Chem., DE-OS 1 925 491, 1968 (M. Jizuka et al.); JP 73 00 851, 1970; JP 49 24 229, 1972.
273. Hodogaya Chem., JP 72 50 208, 1970.
274. Hodogaya Chem., US 3 822 247, 1969 (M. Ozutsumi, S. Maeda, Y. Kawada). Hodogaya Chem., JP 49 24 225, 1972.
275. BASF, BE 660 201, 1964. Geigy, BE 683 733, 1965. Montedison, IT 1 110 641, 1979 (F. Merlo et al.).
276. Ciba-Geigy, EP 55 224, 1980 (P. Loew).
277. BASF, DE 1 217 521, 1961 (H. Baumann, J. Dehnert).
278. Rostow State Univ., SU 682 521, 1977.
279. Hodogaya Chem., JP 51 124 122, 1975.
280. Ciba-Geigy, DE-OS 2 600 042, 1975 (S. Koller).
281. MLB Hoechst, DE 172 118, 1905; DE 175 034, 1906; DE 178 688, 1906; DE 189 942, 1906.
282. Ciba, DE 607 394, 1929.
283. MLB Hoechst, DE 200 207, 1907.
284. S. M. Makin, A. I. Pomogaev, *Zh. Org. Khim.* **18** (1982) no. 10, 2173 – 2176.
285. Fuji Xerox, JP 59 86 058, 1982.
286. V. K. Patratii et al., *Khim. Farm. Zh.* **13** (1979) no. 5, 72 – 76.
287. Z. Ya. Krainer et al., *Khim. Geterotsikl. Soedin.* 1978 no. 1, 76 – 78.
288. Inst. Organ. Khim., SU 775 114, 1978 (V. I. Troitskaya et al.). V. I. Troitskaya et al., *Khim. Geterotsikl. Soedin* 1981 no. 12, 1675 – 1679.
289. IG Farben, DE 410 487, 1922 (W. König); DE 415 534, 1922 (A. Blömer); DE 641 799, 1928 (K. Schmidt).
290. A. Kuhn, R. Winterstein, DE 599 560, 1929. Ciba, DE 539 784, 1929. Schleussner, DE 812 944, 1948.
291. Bayer, DE-OS 3 707 262, 1987 (R. Raue, H. Psaar, H. Berneth).
292. Ricoh, JP 63 201 665, 1987.
293. Bayer, DE-OS 3 105 778, 1981 (R. Raue); EP 58 863, 1981 (R. Raue).
294. Bayer, DE 1 070 316, 1958 (M. Coenen, O. Weißel); BE 576 230, 1958.
295. Bayer, FR 1 395 927, 1963.
296. Bayer, BE 652 123, 1963.
297. IG Farben, DE 693 025, 1936 (M. Müller, O. Wahl, E. Teupel).
298. J. A. van Allan et al., *J. Heterocycl. Chem.* **15** (1978) no. 3, 365 – 367.
299. A. I. Tolmachev et al., *Khim. Geterotsikl. Soedin.* 1982 no. 9, 1178 – 1183.
300. T. G. Deligeorgiev, N. I. Gadjev, *Dyes Pig.* **12** (1990) 157 – 162.
301. Canon, JP 63 168 392, 1986.
302. Victor Co. of Japan, JP 63 167 352, 1986.
303. Matsushita Electr., JP 62 272 263, 1986; JP 63 229 450, 1987.
304. L. G. S. Brooker, *J. Chem. Soc.* 1951, 1087.
305. Canon, JP 61 137 156, 1984.
306. N. Gamon, Ch. Reichardt, *Liebigs Ann. Chem.* 1980, 2072 – 2094. N. Gamon, Ch. Reichardt, *Angew. Chem.* **89** (1977) 418.
307. N Gamon, Ch. Reichardt, *Tetrahedron Lett.* 1979 no. 3, 225 – 226. N. Gamon et al., *Chem. Ber.* **114** (1981) no. 10, 3289 – 3305.
308. BASF, DE 1 065 544, 1956 (H. Baumann, E. Kern), DE 1 038 522, 1956 (H. Baumann et al.).
309. Kabushiki Kaisha, JP 59 20 204, 1982.
310. Sumitomo Chem., JP 49 126 728, 1973; JP 76 07 571, 1973.
311. S. Hünig, *Angew. Chem.* **71** (1959) 369.
312. BASF, BE 610 662, 1960.
313. Geigy, DE 1 069 563, 1957 (J. Voltz, W. Bossard).
314. H. Balli, R. Maul, *Helv. Chim. Acta* **59** (1976) no. 1, 148 – 155. R. Naef, H. Balli, *Helv. Chim. Acta* **61** (1978) no. 8, 2958 – 2973.
315. Bayer & Co., DE 410 487, 1922 (W. König).

316. GAF, DE 1 090 801, 1959.
317. Canon, JP 62 132 689, 1985.
318. Ch. Reichardt, *Justus Liebigs Ann. Chem.* **715** (1968) 74 – 89.
319. W. Grahn, Ch. Reichardt, *Tetrahedron* **32** (1976) 125 – 134.
320. Ch. Reichardt, *Angew. Chem.* **77** (1965) no. 11, 508. Ch. Reichardt, W. Grahn, *Chem. Ber.* **103** (1970) 1072 – 1087.
321. Ch. Reichardt, U. Rust, *Z. Naturforsch.* **37B** (1982) 236 – 245.
322. Ch. Reichardt, K. Schlagerer, *Liebigs Ann. Chem.* 1982, 530 – 535.
323. BASF, DE-OS 3 608 214, 1986 (M. Dust et al.)
324. Fuji Photo, DE-OS 3 630 389, 1985 (K. Kitatani, K. Sano). M. A. Kudinova et al., *Khim. Geterotsikl. Soedin.* 1980 no. 7, 898 – 902.
325. FMC, DE 1 222 608, 1960 (F. M. Sureau et al.). Ugine Kuhlmann, FR 1 353 497, 1962.
326. Ricoh, JP 59 202 892, 1983; JP 62 227 958, 1986; JP 62 232 461, 1986; JP 62 280 263, 1986; JP 63 189 291, 1987. Mitsubishi Chem., JP 62 67 068, 1985. Fuji Photo, US 4 832 992, 1986 (M. Yaba, Y. Inagaki).
327. Mitsubishi Chem., JP 62 124 986, 1985.
328. Eastman Kodak, EP 288 076, 1987 (R. L. Parton et al.).
329. A. F. Vompe et al., *Zh. Org. Khim.* **16** (1980) no. 8, 1767 – 1771. W. Grahn, W. Mrosek, Ch. Reichardt, *Chem. Ber.* **110** (1977) no. 5, 1674 – 1683.
330. S. M. Makin et al., *Zh. Org. Khim.* **20** (1984) no. 2, 439 – 444. N. V. Monich et al., *Zh. Org. Khim.* **21** (1985) no. 5, 1093 – 1097. M. S. Lyubich et al., *Zh. Org. Khim.* **19** (1983) no. 9, 1976 – 1979.
331. R. Raue, H. Harnisch, K. H. Drexhage, *Heterocycles* **21** (1984) 167 – 190.
332. J. Zemlicka, Z. Arnold, *Collect. Czech. Chem. Commun.* **26** (1961) 2852. Z. Arnold, *Collect. Czech. Chem. Commun.* **30** (1965) 2783.
333. G. A. Reynolds, K. H. Drexhage, *J. Org. Chem.* **42** (1977) no. 5, 885 – 888.
334. Eastman Kodak, EP 321 923, 1987 (C. D. de Boer).
335. Eastman Kodak, EP 251 282, 1986 (G. St. Proehl et al.).
336. Jap. Res. Inst. f. Photosens. Dyes, JP 59 217 761, 1983. Olympus Optical, JP 62 193 885, 1986.
337. BASF, EP 335 236, 1988 (M. Acker et al.).
338. Lambda Physik, DE-OS 3 316 666, 1983 (K. H. Drexhage et al.).
339. Ch. Reichardt, *Tetrahedron Lett.* **44** (1967) 4327 –4332. Ch. Reichardt, *Chem. Ber.* **116** (1983) no. 5, 1982 – 2000.
340. W. Grahn, *Tetrahedron* **32** (1976) 1931 – 1939.
341. Fuji Photo JP 63 136 054, 1986; JP 62 187 301, 1986; JP 61 171 766, 1985.
342. BASF, BE 598 870, 1960.
343. Bayer, BE 583 928, 1958.
344. E. D. Sych, L. T. Gorb, *Ukr. Khim. Zh. (Russ. Ed.)* **42** (1976) no. 9, 961 – 964.
345. GAF, US 2 840 443, 1954. Bayer, DE 1 164 973, 1958 (O. Weißel, R. Raue, W. Wunder); DE 1 157 582, 1958 (O. Weißel, R. Raue, W. Wunder).
346. BASF, DE 1 119 436, 1960 (D. Leuchs, H. Otterbach).
347. Ciba-Geigy, DE-OS 2 935 519, 1978 (A. Roueche). A. Roueche, *Dyes Pigm.* **1** (1980) no. 1, 59 – 76.
348. H. G. Heller, DE-OS 2 532 224, 1974; DE-OS 2 647 850, 1976. H. G. Heller, J. R. Langan, *J. Chem. Soc. Perkin Trans. 2* 1981, 341 – 343.
349. Ag. of Ind. Sc. and Techn., JP 63 273 688, 1987.
350. J. Hibino, E. Ando, JP 01 09 282, 1987.
351. ICI, EP 33 583, 1980 (D. F. Newton).
352. VEB Laborchemie, Apolda, DD 240 206, 1985.
353. Ciba-Geigy, EP 35 698, 1980 (W. E. Long).
354. Mitsubishi Paper Mill, JP 52 104 537, 1976.
355. Ciba-Geigy, GB 2 184 743, 1985 (D. J. Edwards et al.).
356. K. Nakashima et al., *Chem. Pharm. Bull.* **31** (1983) no. 7, 2523 – 2525. K. Nakashima et al., *Chem. Pharm. Bull.* **32** (1984) no. 4, 1654 – 1657.
357. Fuji Photo, JP 63 176 171, 1987.
358. Xerox, DE-OS 3 441 041, 1983 (K.-Y. Law, F. C. Bailey); EP 144 206, 1983 (J. F. Janus); US 4 524 219, 1984 (K.-Y. Law).
359. Xerox, US 4 524 220, 1984 (K.-Y. Law).
360. K.-Y. Law, F. C. Bailey, *Dyes Pigm.* **9** (1988) no. 2, 85 – 107.
361. Xerox, EP 145 401, 1983 (K.-Y. Law, F. C. Bailey); US 4 521 621, 1983 (J. F. Janus, W. W. Limburg).
362. Fuji Xerox, JP 60 260 554, 1984. IBM, EP 172 321, 1984 (R. B. Champ, R. L. Vollmer).
363. Fuji Xerox, JP 63 210 167, 1987; JP 61 34 546, 1984; EP 146 123, 1983 (H. Tanaka et al.); EP 161 005, 1984 (S. Kni et al.).
364. Mitsubishi Chem., JP 61 141 784, 1984; JP 61 167 680, 1985.
365. Mitsubishi Chem., JP 61 167 681, 1985.
366. Canon, JP 62 141 538, 1985.

367. Canon, JP 62 141 654, 1985; JP 62 142 688, 1985.
368. V. I. Smirnov, SU 1 043 582, 1982.
369. AS Sibe. Irkut. Org. Chem., SU 1 083 153, 1982.
370. IG Farben, DE 665 557, 1931 (O. Wahl).
371. IG Farben/Bayer, DE 891 120, 1935/1952 (W. Müller, C. Berres).
372. Ciba, DE-OS 1 936 461, 1968 (H. Angliker et al.). Mitsubishi Chem., JP 60 239 290, 1984. Sumitomo Chem., JP 60 253 594, 1984.
373. Bayer, BE 658 426, 1964. Ciba, BE 660 762, 1964. Yorkshire Dyeware, GB 1 027 026, 1964. Bayer, FR 1 469 253, 1965. Yorkshire Chem., GB 2 107 342, 1981.
374. Sandoz, BE 559 324, 1956; DE-OS 2 542 376, 1974 (U. Zirngibl). Ciba-Geigy, DE-OS 2 147 810, 1970 (H. R. Schweizer, W. Weiskat); DE-OS 2 147 811, 1970 (H. R. Schweizer, W. Weiskat). BASF, DE-OS 2 042 498, 1970 (H. Scheuermann). Yorkshire Chem., GB 1 516 713, 1975 (K. Bramham et al.). Aziende Colori Nationale, IT 1 092 257, 1977. T. A. Peters, M. S. Wild, *J. Soc. Dyers Colour* **94** (1978) no. 3, 106 – 113.
375. ICI, GB 1 110 714, 1965. Bayer, DE-OS 2 316 766, 1973 (H. Beecken); DE-OS 2 447 229, 1974 (H. Beecken); DE-OS 2 610 674, 1976 (H. Beecken). Ciba-Geigy, DE-OS 2 910 662, 1978 (S. Koller).
376. Ciba, BE 703 181, 1966. Bayer, DE-OS 2 445 583, 1974 (H. Beecken).
377. Am. Aniline, US 3 635 957, 1969; DE-OS 2 301 461, 1972 (G. R. Genta). VEB Chemiekombinat Bitterfeld, DD 222 321, 1983 (H.-W. Modrow et al.).
378. Sumitomo Chem., JP 59 147 054, 1983.
379. Eastman Kodak, FR 1 387 387, 1963; US 4 180 663, 1977 (M. D. Frishberg).
380. Eastman Kodak, EP 229 374, 1985 (M. A. Weaver, K. K. Lum, W. H. Moore). Sumitomo Chem., JP 61 162 396, 1985; US 4 833 123, 1987 (K. Hashimoto et al.).
381. Eastman Kodak, US 3 597 434, 1967.
382. Bayer, DE-OS 2 530 809, 1975 (H. Hugl, G. Wolfrum).
383. Bayer, EP 70 445, 1981 (H. Hugl, G. Wolfrum, K.-F. Lehment).
384. Eastman Kodak, WO 87 01 121, 1985 (W. P. Pruett et al.).
385. Eastman Kodak, WO 86 04 904, 1985.
386. Du Pont, DE 1 099 671, 1953 (R. E. Heckert); DE 1 094 387, 1953 (R. E. Heckert). Sandoz, BE 574780, 1958. Mitsubishi Chem., JP 60 38350, 1983.
387. Sumitomo Chem., JP 59 129 263, 1983.
388. Sumitomo Chem., JP 59 129 264, 1983.
389. Mitsubishi Chem., JP 59 230 060, 1983. Sumitomo Chem., JP 59 152 956, 1983.
390. Ciba-Geigy, DE-OS 2 741 392, 1976 (P. Möckli).
391. Mitsubishi Chem., JP 61 148 096, 1984; JP 61 241 191, 1985. Sumitomo Electric Ind., JP 61 163 895, 1985.
392. Ciba-Geigy, DE-OS 2 515 138, 1974 (N. B. Desai, V. Ramanathan); IN 144 354, 1974.
393. VEB Chemiekombinat Bitterfeld, DD 136 147, 1978 (H. Niegel et al.). BASF DE-OS 3 638 756, 1986 (K.-H. Etzbach, R. Sens).
394. Sumitomo Chem., JP 73 29 517, 1970; JP 74 23 806, 1970.
395. Sumitomo Chem., JP 73 24 433, 1970; JP 74 04 810, 1970
396. Kodak, GB 2 143 541, 1983 (K. Hunt).
397. Eastman Kodak, US 4 316 013, 1980 (K. Hunt).
398. Eastman Kodak, GB 2 077 751, 1980 (K. Hunt).
399. Geigy, DE 1 098 125, 1958 (H. Häusermann, J. Voltz).
400. Geigy, DE-OS 1 469 770, 1964 (J. Voltz, H. Häusermann); DE-OS 1 619 567, 1966 (J. Voltz, T. Somlo, H. Häusermann).
401. Techn. Univ. Dresden, DD 204 094, 1982. BASF, DE-OS 2 807 761, 1978 (C. Vamvakaris et al.).
402. Bayer, DE-OS 2 030 507, 1970 (H. Harnisch); DE-OS 2 065 076, 1970 (H. Harnisch); DE-AS 2 739 841, 1977 (H. Harnisch). Ciba-Geigy, DE-OS 2 710 285, 1976 (P. Loew).
403. Sandoz, US 4 007 188, 1973 (W. Koch).
404. Sandoz, CH 580 610, 1972 (W. Koch).
405. Ciba-Geigy, DE-OS 2 319 230, 1972 (H. Schwander).
406. BASF, DE-OS 2 529 434, 1975 (M. Patsch, C. Vamvakaris); DE-OS 2 553 294, 1975 (M. Patsch, C. Vamvakaris).
407. Bayer, DE-OS 3 533 075, 1985 (F. Seng).
408. Ciba-Geigy, DE-OS 2 717 599, 1976 (N. B. Desai). VEB Carl Zeiss, DD 221 191, 1983 (P. Czerney, H. Hartmann).
409. BASF, DE-OS 2 144 591, 1971 (W. Mach, H. Scheuermann).
410. BASF, DE-OS 2 253 538, 1972 (W. Mach, D. Augert, H. Scheuermann); DE-OS 2 415 661, 1974 (H. Scheuermann et al.). Nippon Kayaku, JP 57 03 860, 1980; JP 57 31 958, 1980.
411. Bayer, GB 2 013 703, 1978 (H. Harnisch).
412. Ciba-Geigy, US 4 547 579, 1978 (P. Möckli). P. Möckli, *Dyes Pigm.* **1** (1980) no. 1, 3 – 15.

413. Bayer, EP 25 136, 1979 (U. Claussen, H. Harnisch).
414. BASF, DE-OS, 2 937 422, 1979 (G. Seybold, C. Vamvakaris); DE-OS 3 609 804, 1986 (E. Hahn, G. Seybold, A. Stange).
415. Ciba-Geigy, DE-OS 2 844 299, 1977 (P. Möckli).
416. Bayer, DE-OS 2 925 546, 1979 (H. Harnisch).
417. Du Pont, US 3 742 012, 1972; DE-OS 2 716 765, 1976 (W. D. Steiner). Bayer, DE-OS 2 714 653 (H. Beecken). Am. Color and Chem., US 4 235 806, 1979 (K. A. Desai).
418. Am. Cyanamid, DE-OS 2 747 365, 1976 (F. F. Cesark et al.); US 4 017 486, 1976 (B. G. Buell, F. R. Loffelmann).
419. BASF, DE-OS 3 136 892, 1981 (H.-J. Degen et al.).
420. Sandoz, DE-OS 2 447 878, 1973 (W. Koch). Ciba-Geigy, EP 23 022, 1979 (H. Schwander, C. Zickendraht).
421. Bayer, DE-OS 3 044 128, 1980 (H. Harnisch); DE-OS 3 100 295, 1981 (H. Harnisch).
422. Bayer, DE-OS 3 248 043, 1982 (O. S. Wolfbeis, H. Harnisch); DE-OS 3 321 041, 1983 (O. S. Wolfbeis, H. Harnisch).
423. Nippon Kayaku, JP 60 226 555, 1984; JP 61 103 862, 1984.
424. Bayer, DE-OS 2 413 281, 1974 (H. Harnisch).
425. Bayer, DE-OS 2 413 371, 1974 (H. Harnisch).
426. Bayer, DE-OS 2 413 281, 1974 (H. Harnisch).
427. Bayer, EP 47 881, 1980 (H. Harnisch).
428. P. Czerney, H. Hartmann, J. Liebscher, DD 141 673, 1979. P. Czerney, H. Hartmann, *J. Prakt. Chem.* **324** (1982) no. 2, 255 – 266.
429. Sandoz, BE 626 369, 1961.
430. Eastman Kodak, EP 332 922, 1988 (S. Evans, H. Weber).
431. Eastman Kodak, EP 332 924, 1988 (S. Evans, H. Weber).
432. Eastman Kodak, EP 332 923, 1988 (S. Evans, H. Weber).
433. Fuji Photo, GB 2 094 493, 1981.
434. Nippon Kayaku, JP 61 268 760, 1985.
435. BASF, DE-OS 1 809 949, 1968 (G. Hansen, H. R. Müller); DE-OS 2 415 819, 1974 (G. Hansen, W.-D. Kermer); DE-OS 2 440 405, 1974 (N. Zimmermann et al.).
436. Sandoz, DE-OS 2 929 001, 1978 (W. Baumann).
437. Sumitomo Chem., JP 59 12 969, 1982.
438. Sumitomo Chem., JP 59 124 955, 1983.
439. Sumitomo Chem., JP 58 125 749, 1982.
440. Nippon Kayaku, JP 62 220 557, 1986.
441. Sumitomo Chem., JP 60 158 262, 1984.
442. Dai Nippon Insatsu Kabushiki, EP 327 077, 1988 (J. Kanto et al.); EP 327 063, 1988 (J. Kanto, H. Saito).
443. Dai Nippon Printing, JP 63 205 288, 1987.
444. Canon, JP 63 249 149, 1987.
445. Cassella, DE 1 569 692, 1966 (K. Mix).
446. Bayer, DE-OS 2 349 980, 1973 (A. Brack). Sublistatic Holding, CH-AS 515/75, 1975.
447. BASF, DE-OS 2 429 760, 1974 (E. Schefczik).
448. Bayer, DE-OS 2 608 020, 1976 (H. Harnisch).
449. Ciba-Geigy, DE-OS 2 724 444, 1976 (H. Schwander, C. Zickendraht); DE-OS 2 724 493, 1976 (H. Schwander, C. Zickendraht).
450. Ciba-Geigy, DE-OS 2 607 869, 1975 (H. Schwander). Ciba-Geigy, DE-OS 2 736 914, 1976 (H. Schwander, C. Zickendraht).
451. Ciba-Geigy, DE-OS 2 551 575, 1974 (H. Schwander et al.). Bayer, DE-OS 2 848 622, 1978 (H. Harnisch, P. Rasche).
452. Ciba-Geigy, DE-OS 2 814 526, 1977 (J. v.d.Crone). BASF, DE-OS 3 622 898, 1985 (W. Lotsch, R. Kemper).
453. BASF, DE-OS 2 830 501, 1978 (P. Neumann, W. Elser, G. Bock).
454. Bayer, US 3 255 204, 1958.
455. Bayer, DE-OS 2 300 487, 1973 (R. Raue, W. Kühnel).
456. Bayer, BE 581 913, 1958.
457. C. Hubschwerlen, J.-P. Fleury, F. Fleury, *Helv. Chim. Acta* **60** (1977) no. 4, 1312 – 1321.
458. Bayer, DE-OS 2 341 657, 1973 (A. Brack).
459. BASF, DE-OS 2 428 198, 1974 (E. Schefczik).
460. Ciba-Geigy, DE-OS 2 724 566, 1976 (H. Schwander, C. Zickendraht).
461. BASF, DE-OS 2 930 333, 1979 (E. Schefczik).
462. BASF, DE-OS 2 924 067, 1979 (E. Schefczik).
463. BASF, DE-OS 2 611 665, 1976 (E. Schefczik); DE-OS 2 705 108, 1977 (E. Schefczik).
464. Bayer, BE 850 841, 1976 (H. Harnisch).
465. BASF, DE-OS 2 628 409, 1976 (W. Lotsch); DE-OS 3 007 300, 1980 (W. Lotsch, R. Kemper).
466. BASF, DE-OS 3 311 375, 1983 (W. Lotsch); DE-OS 2 909 644, 1979 (W. Lotsch).
467. B. K. Manukian, A. Mangini, *Chimia* **24** (1970) 328 – 339.
468. BASF, EP 145 003, 1983 (R. Bermes).
469. Bayer, DE-OS 3 022 837, 1980 (A. Brack).
470. B. K. Manukian, P. Nicklaus, H. Ehrsam, *Helv. Chim. Acta* **52** (1969) no. 5, 1259 – 1273.
471. Mitsui Toatsu, JP 52 138 527, 1976.

472. Sublistatic Holding, DE-OS 1 771 812, 1968 (P. Sander); DE-OS 2 101 910, 1971 (N. de Plasse, P. Eckert); DE-OS 2 210 956, 1972 (N. de Plasse). Mitsubishi Kasei, JP 63 182 192, 1987; EP 318 032, 1987 (N. Taguchi et al.). Matsushita Electric, EP 323 259, 1987 (N. Taguchi et al.).
473. Hoechst, DE-OS 1 670 880, 1967 (E. Spietschka, F. Ische).
474. BASF, DE-OS 2 232 447, 1972 (E. Schefczik). Mitsubishi Chem., JP 53 27 285, 1970; JP 77 37 006, 1970.
475. Hoechst, BE 733 605, 1969. Teijin, JP 53 129 223, 1977.
476. Mitsubishi Chem., JP 61 155 466, 1984; JP 61 166 857, 1985; JP 61 166 858, 1985.
477. Mitsubishi Chem., JP 58 109 688, 1981; US 4 500 710, 1983; DE-OS 3 320 430, 1983 (T. Niwa, H. Hihira).
478. Bayer, DE-OS 2 225 960, 1972 (M. Groll, V. Hederich, H.-S. Bien).
479. Bayer, DE-OS 2 530 124, 1975 (M. Groll, H.-S. Bien).
480. Bayer, DE-OS 2 530 123, 1975 (M. Groll, H. S. Bien).
481. Du Pont, US 3 639 405, 1968. GAF, DE-OS 2 247 971, 1972 (W. E. Wallace). Mitsui Toatsu, JP 58 59 971, 1977; JP 59 108 769, 1977.
482. Mitsubishi Kasei, JP 63 191 861, 1987.
483. Mitsui Toatsu, JP 59 43 086, 1982.
484. Mitsubishi Chem., JP 78 16 009, 1970. Teijin, JP 51 34 927, 1974.
485. BASF, DE-OS 3 329 340, 1983 (H.-J. Degen).
486. Ciba-Geigy, EP 88 418, 1982 (J.-M. Adam).
487. Teijin, JP 52 87 443, 1976; JP 52 98 758, 1976.
488. BASF, GB 1 225 336, 1969. Teijin, DE-OS 2 638 528, 1975 (K. Shimada et al.); JP 53 50 231, 1976.
489. Teijin, JP 51 137 724, 1975; JP 52 150 438, 1976.
490. Ciba, CH 438 542, 1964 (P. Gangheux). Prod. Chim. Ugine Kuhlmann, DE-OS 2 901 096, 1979.
491. Teijin, JP 51 102 024, 1975; JP 52 10 330, 1975; JP 52 10 342, 1975.
492. Teijin, JP 52 10 329, 1975.
493. Teijin, JP 51 102 035, 1975.
494. Teijin, JP 51 102 023, 1975.
495. Teijin, JP 52 10 341, 1975.
496. BASF, DE-OS 3 625 007, 1985 (W. Lotsch, P. Neumann, H. Kowarsch).
497. Eastman Kodak, EP 257 580, 1986, G. W. Byers, D. D. Chapman). G. W. Byers, US 4 725 574, 1987.
498. Eastman Kodak, EP 257 577, 1986 (G. W. Byers et al.).
499. Bayer, BE 581 913, 1958.
500. Bayer, EP 87 044, 1982 (B. Wehling).
501. Mitsubishi Chem., JP 52 99 379, 1976; JP 53 14 734, 1976.
502. Fuji Photo, DE-OS 3 226 197, 1981 (S. Horie et al.). Canon, JP 63 249 855, 1987.
503. Ciba-Geigy, EP 258 180, 1986 (R. Naef).
504. Canon, JP 58 220 792, 1982.

Methylamines

August B. van Gysel, UCB, Brussels, Belgium

Willy Musin, UCB, Brussels, Belgium

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Production 2
4.1. Synthesis 2
4.2. Purification 3
5. Environmental Protection 4
6. Quality Specifications 4
7. Storage and Transportation 4
8. Uses 5
9. Economic Aspects 6
10. Toxicology and Occupational Health . 6
11. References 6

1. Introduction

Monomethylamine [*74-89-5*], methanamine, aminomethane, CH_3NH_2, M_r 31.06; dimethylamine [*124-40-3*], *N*-methylmethanamine, $(CH_3)_2NH$, M_r 45.08; and trimethylamine [*75-50-3*], *N,N*-dimethylmethanamine, $(CH_3)_3N$, M_r 59.11, are colorless gases at room temperature. All three are produced by the reaction of methanol and ammonia over solid acid catalysts.

Methylamines are the simplest members of the amine family and constitute an economical source of reactive amino nitrogen. Methylamines are valuable for the synthesis of a broad range of products applicable in such diverse fields as medicine, agriculture, rubber, plastics, and synthetic fibers.

Methylamines display chemical reactivity typical of small aliphatic amines. They bear a strong resemblance to ammonia from which they are derived, and their physical properties largely parallel those of ammonia.

Even at concentrations as low as 0.1 ppm in air, methylamines have a distinct, offensive, fishy odor. At higher concentrations, their odor is more ammoniacal.

2. Physical Properties

The most important physical properties of methylamines are listed in Table 1. Table 2 summarizes the physical properties of aqueous methylamine solutions [1], [2].

3. Chemical Properties (→ Amines, Aliphatic, Chap. 2.)

All three methylamines readily form salts with acids (e.g., mono-, di-, or trimethylammonium halides–also known as the corresponding amine hydrohalides–with the hydrohalic acids). Reaction with alkyl halides leads to further alkylation, up to and including the formation of quaternary ammonium salts, $R_{4-n}(CH_3N)_n{}^+X^-$ (e.g., methylethylamine, methyldiethylamine, and methyltriethylammonium halide in the reaction of ethyl halide with monomethylamine).

Reaction of mono- or dimethylamines with carboxylic acids and their derivatives results in *N*-methylamides. Similarly, substituted mono- and dimethylsulfonamides can be prepared by treatment with a sulfonyl chloride.

Monomethylamine (MMA) reacts with phosgene to produce methyl isocyanate, with carbon disulfide in the presence of sodium hydroxide to give sodium monomethyldithiocarbamate, and with urea to give *N*-methylurea via the intermediate ammonium cyanate. A mixture of chloroform and alkali converts monomethylamine to methyl isonitrile. Alkaline hypochlorites lead to a mixture of methylmono- and dichloroamines. Monomethylamine is decomposed by nitrous acid to methanol and nitrogen, and by nitrosyl chloride to methyl chloride and nitrogen. Ethylene oxide reacts with monomethylamine to produce methylethanolamine and methyldiethanolamine.

Dimethylamine (DMA) and phosgene react to produce tetramethylurea. Reaction with carbon disulfide and alkali in this case gives an

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_535

Table 1. Physical properties of methylamines

Property	Monomethylamine	Dimethylamine	Trimethylamine
bp (101.33 kPa), °C	−6.33	6.88	2.87
mp, °C	−93.46	−92.19	−117.3
Density (d^{20}), g/cm^3			
at 101.33 kPa (gas)	0.0014	0.0020	0.0027
at 20 °C (liquid)	0.6624	0.6556	0.6331
Surface tension (25 °C), 10^{-3} N/m	19.19	16.33	13.47
Heat of vaporization, kJ/mol			
at 25 °C	24.249	23.663	22.864
at *bp*	26.0	24.626	27.708
Heat of fusion, kJ/mol	6.054	5.945	6.548
Standard heat of formation, kJ/mol			
at 25 °C (liquid)	−47.31	43.96	−45.80
at 25 °C (gas)	−22.98	−18.46	−23.86
Heat of combustion, standard state at 25 °C, liquid, kJ/mol	−1061.35	−1744.63	−2422.60
Heat capacity, ideal gas (25 °C), J K^{-1} mol^{-1}	49.27	70.84	90.97
t_{crit}, °C	156.9	164.5	160.1
p_{crit}, MPa	7.46	5.31	4.07
Dielectric constant (25 °C), liquid	9.4	5.26	2.44
Dipole moment in benzene (25 °C), D	1.47	1.18	0.87
Flash point (closed cup), °C	0	− 18	− 7
Ignition temperature in air, °C	430	400	190
Lower explosion limit in air, vol %	4.95	2.8	2
Upper explosion limit in air, vol %	20.75	14.4	11.6

alkali dimethyldithiocarbamate, whereas urea leads to the expected *N,N*-dimethylurea. Both nitrous acid and nitrosyl chloride convert dimethylamine to the corresponding nitrosamine. Ethylene oxide reacts with dimethylamine to give dimethylethanolamine.

Oxidation of *trimethylamine* (TMA) with hydrogen peroxide or a peracid gives trimethylamine oxide. The reaction with ethylene oxide in this case yields choline. Trimethylamine can form quaternary ammonium compounds by *N*-alkylation.

4. Production

4.1. Synthesis

Methylamines are prepared commercially by the exothermic reaction of methanol with ammonia over an amorphous silica – alumina catalyst at 390 – 430 °C. All three possible methylamines are produced. The reaction proceeds to thermodynamic equilibrium, whose position is governed by the temperature and the nitrogen – carbon ratio as shown in Figure 1 [3], [4].

The equilibrium composition favors the formation of trimethylamine, whereas market demand is greater for mono- and especially dimethylamine. Excess mono- and trimethylamines are recycled over the same catalyst, leading to additional dimethylamine by dismutation or disproportionation.

The kinetics of the overall reaction have been studied by using several amorphous solid acid catalysts [5], [6]. Amine synthesis can be modeled by six bimolecular equilibrium reactions, three involving amine synthesis and three describing dismutation:

Amination reactions

$$NH_3 + CH_3OH \rightleftharpoons CH_3NH_2 + H_2O$$

$$(CH_3)_2NH + CH_3OH \rightleftharpoons (CH_3)_3N + H_2O$$

Dismutation reactions

$$CH_3NH_2 + CH_3NH_2 \rightleftharpoons (CH_3)_2NH + NH_3$$

$$CH_3NH_2 + (CH_3)_2NH \rightleftharpoons (CH_3)_3N + NH_3$$

$$(CH_3)_2NH + (CH_3)_2NH \rightleftharpoons (CH_3)_3N + CH_3NH_2$$

All six equilibrium constants are known [3], [7]. A parallel reaction leads to the formation of dimethylether, which may also be an intermediate in the synthesis of the amines:

Table 2. Physical properties of aqueous solutions of methylamines

Property	40 % Monomethylamine	40 % Dimethylamine	60 % Dimethylamine	40 % Trimethylamine
bp (101.33 kPa), °C	49.4	51.5	36.0	30.8
mp, °C	−38	ca. − 38	−60	1.7
Flash point, °C	−10	−10	−30	ca. − 20
Density (15.5 °C), g/cm^3	0.904	0.898	0.827	0.880
Vapor pressure (total), kPa				
20 °C	34	27	57	61
30 °C	53	42	80	90
40 °C	80	65	115	131
50 °C	120	97	158	185
Viscosity (15.5 °C), mPa · s	1.5	1.7		1.8

$$2\,CH_3OH \rightleftharpoons CH_3{-}O{-}CH_3 + H_2O$$

Research has been directed toward improving the selectivity for dimethylamine through the use of shape-selective zeolite catalysts [8–16].

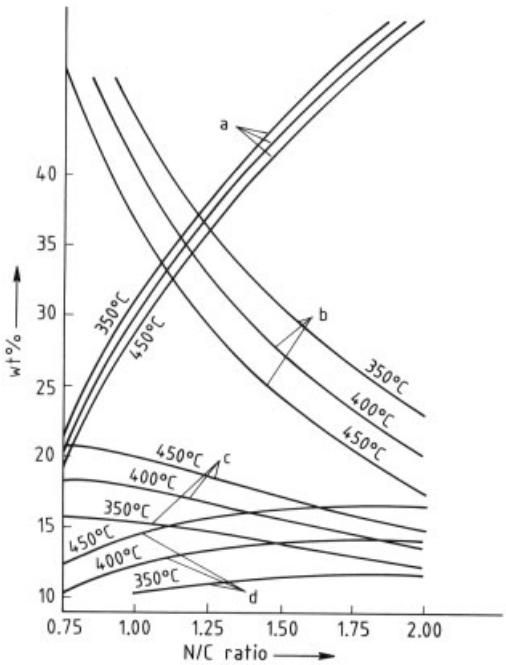

Figure 1. Composition of an equilibrium mixture of methylamines as a function of the nitrogen – carbon ratio at 350, 400, and 450 °C
a) Ammonia; b) Trimethylamine; c) Dimethylamine; d) Monomethylamine

Reactant selectivity is a consequence of one of the reactants being too large to pass through the zeolite channels, whereas *product selectivity* is observed when only certain products are of the proper size and shape to diffuse out of the channels [17]. The activity of zeolites with respect to the conversion of methanol is much higher than that of amorphous silica – alumina, but coke formation within the catalyst pores limits catalyst life even at 300 °C (instead of 400 – 430 °C used with amorphous silica – alumina catalysts) [18].

4.2. Purification

The crude reaction mixture consists essentially of excess ammonia; mono-, di-, and trimethylamines; reaction water; and unconverted methanol. Purification is generally effected in a train of four to five distillation columns.

In one technique (Fig. 2) the operating pressures of the columns are chosen in such a way that the top condensors can be cooled by standard 30 °C cooling water, and that the bottom reboilers are fed with saturated steam of about 195 °C. The first column (a) (known as the ammonia column), which operates at 1.7 MPa, removes excess ammonia. The latter is recycled to the synthesis section. In the second column (b), pure trimethylamine is recovered overhead by using extractive distillation with water and a pressure of 1.0 MPa. Extraction water, reaction water, and unconverted methanol are removed from the bottom of the third column (c); the mixture of mono- and dimethylamines is distilled overhead. This mixture is separated in the fourth distillation column (d) at 0.7 MPa, MMA being the overhead and DMA the bottom product.

In another technique, (Fig. 3), a supplementary column (b) operating at 0.7 – 0.9 MPa is inserted after the ammonia column (a) for bottoms removal of reaction water, certain impurities,

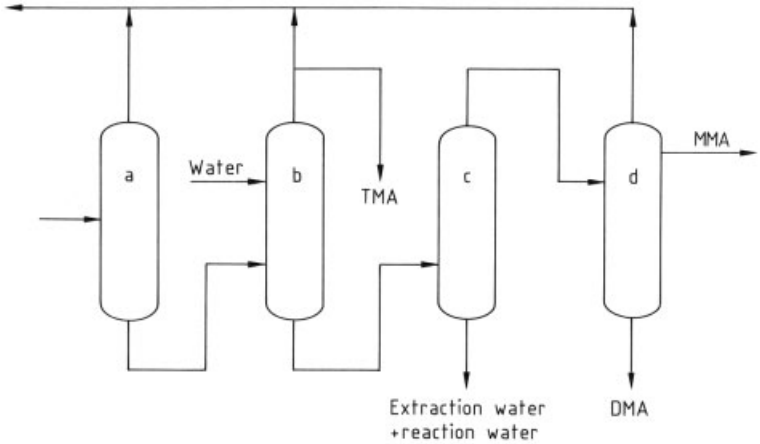

Figure 2. Schematic of methylamine purification
a) Ammonia column; b) Extractive distillation column; c) Column for water removal; d) MMA and DMA product column

and unconverted methanol. Most of the extraction water may be recycled; the small remaining quantity of bottom water, containing some impurities and methanol, is disposed of easily either by incineration or biological treatment.

5. Environmental Protection

Analysis of intermediate byproducts and final products occurs automatically in the operations control room with the aid of an on-line chromatographic analyzer coupled to an on-line computer. Centralizing the data handling in this way assures safe plant operation.

Two effluents are formed in the process:

1) gas containing amines and various byproducts (H_2, CO, N_2, O_2, ethylene), which is incinerated at high temperature to avoid formation of nitrosoamines; and
2) aqueous wastes, whose nature depends on the selectivity and activity of the catalysts. These aqueous solutions contain unconverted methanol and longer chain alkylamines, and can be treated by biological degradation with activated sludge.

6. Quality Specifications

Product purity as determined by gas chromatography is high ($>99\,\%$). Commercial methylamines must be free of higher chain alkylamines. Tables 3, 4, 5 provide typical commercial specifications.

Table 3. Specifications for monomethylamine

	Anhydrous	Aqueous solution
Monomethylamine, %	min. 99.2	min. 40
Di- and trimethylamine, %	max. 0.5	max. 0.2
Ammonia, %	max. 0.05	max. 0.02
Methanol, %	max. 0.1	max. 0.05
Water, %	max. 0.3	balance

7. Storage and Transportation

Methylamines are sold as anhydrous compressed liquefied gases and as aqueous solutions. They are extremely flammable products that should be stored in a well-ventilated area and protected from fire risk.

Monomethylamine. The hazard classifications for anhydrous monomethylamine are as follows:

ADR/RID	Class 2, no. 3 bt; labels no. 3 + 6.1, flammable; orange panel 236/1061
	Class 2, D 2200
UN no.	1061

whereas 40 % aqueous solutions of monomethylamine are classified as

ADR/RID	Class 3, no. 22 a/b; labels no. 3 + 8, flammable, caustic; orange panel 338/1235
IMDG Code	Class 3.1, D 3152
UN no.	1235

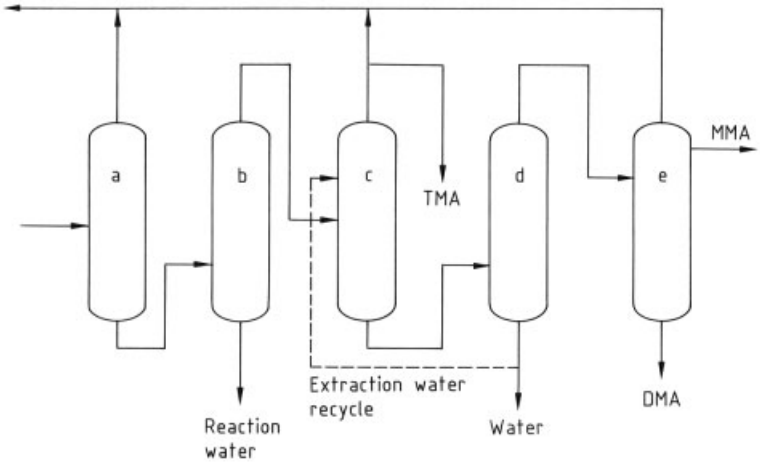

Figure 3. Methylamine purification with supplementary column
a) Ammonia column; b) Supplementary column; c) Extractive distillation column; d) Column for water removal; e) MMA and DMA product column

Table 4. Specifications for dimethylamine

	Anhydrous	Aqueous solutions	
		40 %	60 %
Dimethylamine, %	min. 99.5	min. 40.0	min. 60.0
Mono- and trimethylamine, %	max. 0.3	max. 0.15	max. 0.2
Ammonia, %	max. 0.02	max. 0.01	max. 0.01
Methanol, %	max. 0.1	max. 0.05	max. 0.05
Water, %	max. 0.2	balance	balance

Dimethylamine. Anhydrous dimethylamine must be classified according to

ADR/RID	Class 2, no. 3 bt; labels no. 3 + 6.1, flammable; orange panel 236/1032
IMDG Code	Class 3.1, D 3152
UN no.	1235

The hazard classification for aqueous solutions containing > 40 % DMA are

ADR/RID	Class 3, no. 22 a/b; labels no. 3 + 8, flammable, caustic; orange panel 338/1160
IMDG Code	Class 3.1, 3.2, D 3275
UN no.	1160

Table 5. Specifications for trimethylamine

	Anhydrous	Aqueous solution
Trimethylamine, %	min. 99.0	min. 45
Mono- and dimethylamine, %	max. 0.3	max. 0.15
Ammonia, %	max. 0.1	max. 0.05
Methanol, %	max. 0.1	max. 0.05
Water, %	max. 0.5	max. 0.25

Trimethylamine. Hazard classifications for anhydrous trimethylamine are as follows:

ADR/RID	Class 2, no. 3 bt; labels no. 3 + 6.1, flammable; orange panel 236/1083
IMDG Code	Class 2, D 2256
UN no.	1083

For an aqueous solution of ≤ 30 % trimethylamine, the following hazard classifications are used:

ADR/RID	Class 3, no. 22 a; flammable, caustic
IMDG Code	Class 3.2, D 3413
UN no.	1297

8. Uses

Monomethylamine is an intermediate in the synthesis of pharmaceuticals (e.g., theophylline), pesticides (carbaryl, sodium metham, carbofuran), surfactants, photographic develop-

ers, explosives, and solvents such as *N*-methyl-2-pyrrolidone.

Dimethylamine is also a synthetic intermediate. Examples of products based on dimethylamine include fungicides and rubber accelerators [zinc bis(dimethyldithiocarbamate) (Ziram), tetramethyl thioperoxydicarbonic diamide (TMTD), tetramethyl thiodicarbonic diamide (MTMT)], the propellant 1,1-dimethylhydrazine, various pharmaceuticals, monomers such as dimethylaminoethyl methacrylate, solvents (*N,N*-dimethylformamide, *N,N*-dimethyl- acetamide), catalysts [e.g., 2,4,6-tris[(dimethylamino)methyl]phenol (DMP 30)], the insecticide dimefox, surfactants, and ion-exchange resins.

Trimethylamine is used in the manufacture of choline salts, cationic starches, disinfectants, flotation agents, intense sweeteners, and ion-exchange resins.

Table 6. Methylamine production capacity

Location	Company	Capacity, 10^3 t
North America		170
United States	Du Pont	
	Air Products	
	Pitman-Moore	
	GAF	
Canada	Chinook	
Mexico	Celanese	
South America		10
Brazil		
Western Europe		210
Fed. Rep. of Germany	BASF	
Netherlands	Akzo	
Belgium	UCB	
United Kingdom	ICI	
Italy	ICT	
Spain	Ertisa	
Eastern Europe		30
former German Democratic Republic	VEB Leuna	
former Soviet Union		
former Rumania		
Asia		60
Japan	Nitto	
	Mitsubishi Gas	
India	Fertilizers Chem.	
Korea	Korea Fertilizer	
China	(various plants)	

9. Economic Aspects

Table 6 lists the principal manufacturers of methylamines by country, together with summary data on production capacity. The global market breakdown for methylamines is estimated to be 24 % MMA, 57 % DMA, and 19 % TMA [19].

10. Toxicology and Occupational Health

All gaseous amines are regarded as hazardous materials. At high airborne concentration, their odor resembles that of ammonia and the methylamines display an asphyxiating character. At lower concentration, they induce severe irritation and damage to the eyes and respiratory tract; skin contact results in burns. The slightest leak is easily recognized due to the characteristic fishy odor of the compounds, which is detectable above 0.1 ppm by volume. The TLV and the MAK value are 10 ppm (13 mg/m^3) by volume.

Monomethylamine (RTECS 1.1.78 – 31.1.87, NIOSH PF 6 300 000)

LD_{50} (mouse, inhalation): 2400 mg/m^3 over 2 h

Effects on skin (guinea pig): severe irritant at a dose of 100 mg (Draize test)

Dimethylamine (RTECS 1.1.78 – 31.1.87, NIOSH IP 8 750 000)

LD_{50} (rat, oral): 698 mg/kg

LC_{50} (rat, inhalation): 4540 ppm over 6 h

Toxic effects on eyes, respiratory system, and pulmonary system

Trimethylamine (RTECS 1.1.78 – 31.1.87, NIOSH PA 0 350 000)

LD_{50} (rat, inhalation): 3500 ppm over 4 h

11. References

1. J. A. Riddick, W. B. Bunger, T. K. Sakano: "Organic Solvents," in A. Weissberger (ed.): *Techniques of Chemistry,* vol. II, 4th ed., Wiley-Interscience, New York 1986.

2. N. I. Sax, R. J. Lewis, Sr.: *Dangerous Properties of Industrial Materials,* 7th ed., Van Nostrand Reinhold, New York 1989.
3. J. Issoire, C. Van Long, *Bull. Soc. Chim. Fr.* (1960) 2004; cited in D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds,* Wiley, New York 1969.
4. J. Ramioulle, A. David, *Hydrocarbon Process. Int. Ed.* **60** (1981) no. 7, 113 – 117.
5. F. J. Weigert, *J. Catal.* **103** (1987) 20 – 29.
6. M. Keane et al., *Appl. Catal.* **32** (1987) no. 1 – 2, 361 – 366.
7. G. Schmitz, *J. Chim. Phys.* **77** (1980) 393; **72** (1975) 579.
8. D. G. Parker, A. J. Tompsett, GB 2 019 394 B, 1982.
9. F. J. Weigert, US 4 313 003, 1982.
10. I. Mochida, JP 49 340, 1983.
11. M. Deeba, R. N. Cochran, EP-A 85 408, 1983.
12. R. N. Cochran, M. Deeba, US 4 398 041, 1983.
13. A. J. Tompsett, T. V. Whittam, EP-A 107 457, 1984.
14. A. J. Tompsett, US 4 436 938, 1984.
15. Y. Ashina, T. Fujita, M. Fukatsu, J. Yagi, EP-A 125 616, 1984;EP-A 130 407, 1985.
16. M. Deeba, W. J. Ambs, R. N. Cochran, US 4 434 300, 1984.
17. P. B. Weisz, *Pure Appl. Chem.* **52** (1980) 2091.
18. I. Mochida, A. Yasutake, H. Fujitsu, K. Takeshita, *J. Catal.* **82** (1983) no. 2, 313 – 321.
19. T. Gibson et al.: "Alkylamines," *Chemical Economics Handbook,* SRI International 611.5030, 1984.

Methylaniline (N-) → Aniline

Methyl Tert-Butyl Ether

Automotive Fuels and Octane Enhancers are separate keywords.

Udo Peters, Oxeno Olefinchemie GmbH, Marl, Federal Republic of Germany (Chaps. 1 – 10)

Franz Nierlich, Oxeno Olefinchemie GmbH, Marl, Federal Republic of Germany (Chaps. 1 – 10)

Ekkehard Schulte-Körne, Oxeno Olefinchemie GmbH, Marl, Federal Republic of Germany (Chaps. 11, 12)

Michael Sakuth, Hüls Infracor GmbH, Marl, Federal Republic of Germany (Chaps. 1 – 10)

Rula Deeb, Malcolm Pirnie, Inc., Emeryville, USA (Chap. 13)

Maryline Laugier, Malcolm Pirnie, Inc., Emeryville, USA

Martti Suominen, Fortum Oil and Gas Oy, Finland (Chap. 13)

Michael Kavanaugh, Malcolm Pirnie, Inc., Emeryville, USA (Chap. 13)

1. Introduction 1
2. Physical and Chemical Properties 2
3. Resources and Raw Materials . 2
4. Production 3
5. Environmental Protection 6
6. Quality Specifications 6
7. Chemical Analysis 7
8. Storage and Transportation . . . 7
9. Legal Aspects 7
10. Uses 7
10.1. As an Octane Enhancer 7
10.2. Other Uses 8
11. Toxicology and Occupational Health 9
11.1. Toxicological Profile 9
11.2. Human Experience 9
11.3. Health Risk Characterization, Occupational Health 9
12. Ecotoxicology 10
12.1. Ecotoxicity Data 10
12.2. Behavior in the Environment . . 10
13. Removal of MTBE from Impacted Groundwater 10
13.1. Overview of MTBE Remediation Technologies 10
13.2. Phase Transfer Technologies . . 11
13.2.1. Ex-Situ Phase Transfer Technologies 11
13.2.1.1. Air Stripping 11
13.2.1.2. Granular Activated Carbon (GAC) 11
13.2.1.3. Other Ex-Situ Phase Transfer Technologies 12
13.2.2. In-Situ Phase Transfer Technologies 12
13.2.2.1. Soil Vapor Extraction (SVE) and Multi-Phase Extraction (MPE) . . 12
13.2.2.2. In-Situ Air Sparging (IAS) 12
13.2.2.3. Phytoremediation 12
13.3. Transformation or Destruction Technologies 13
13.3.1. Ex-Situ Transformation Technologies 13
13.3.1.1. Advanced Oxidation Processes (AOPs) 13
13.3.1.2. Biological Treatment 13
13.3.2. In-Situ Transformation Technologies 14
13.3.2.1. In-Situ Chemical Oxidation 14
13.3.2.2. In-Situ Bioremediation 14
14. References 14

1. Introduction

Methyl *tert*-butyl ether, 2-methoxy-2-methylpropane (MTBE) [*1634-04-4*], M_r 88.15, was first synthesized (by the classical Williamson ether synthesis) and characterized in 1904 [1].

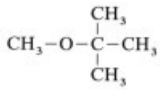

Methyl *tert*-butyl ether (MTBE)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_543

Extensive studies in the United States during World War II demonstrated the outstanding qualities of MTBE as a high-octane fuel component [2]. Even so, it was not until 1973 that the first commercial plant went on stream in Italy.

Reduction of the lead content in gasoline in the mid-1970s led to a drastic increase in demand for octane enhancers, with MTBE being used increasingly in this role. Political decisions about gasoline quality (e.g., lower aromatics content, lower vapor pressure and a defined oxygen content) especially in the United States had led to a significant increase in MTBE consumption in the 90th. In 1997 the world wide MTBE production reached some 19×10^6 t [44]. After years with two-digit growth rates (1990 – 1995) the increase in MTBE consumption is expected to be less than 2 % per year in the near future.

Table 1. Selected physical properties of MTBE

mp	− 108.6 ° C
bp	55.3 ° C
n_D^{20}	1.3692
Dielectric constant (20 °C)	4.5
Viscosity (20 °C)	0.36 mPa · s
Surface tension (20 °C)	20 mN m^{-1}
Specific heat (20 °C)	2.18 kJ kg^{-1} K^{-1}
Heat of vaporization (at *bp*)	337 kJ/ kg^{-1}
Heat of formation (25 °C)	− 314 kJ mol^{-1}
Heat of combustion	− 34.88 MJ kg^{-1}
Flash point (Abel – Pensky)	− 28 °C
Ignition temperature (DIN 51 794)	460 ° C
Explosion limits in air	1.65 – 8.4 vol %
Critical data	
t_{cr}	224.0 °C
p_{cr}	3.43 MPa

2. Physical and Chemical Properties

Physical Properties. Methyl *tert*-butyl ether is a colorless, readily mobile liquid with a characteristic terpene-like odor. Its most important physical properties are listed in Table 1:

Vapor pressure, density, and solubility in water, as well as the composition and boiling points of azeotropes with water and methanol, are given in Tables 2 and 3 [3]. Methyl *tert*-butyl ether has unlimited miscibility with all ordinary organic solvents and all hydrocarbons.

Chemical Properties. Methyl *tert*-butyl ether is very stable under alkaline, neutral, and weakly acidic conditions. In the presence of strong acids, it is cleaved to methanol and isobutene. Depending on reaction conditions the latter can form isobutene oligomers.

3. Resources and Raw Materials

MTBE is produced by the reaction of isobutene, contained in C_4-fractions, and Methanol. At present, *isobutene* from the following sources is used as feedstock for MTBE production [45]:

1) isobutene in raffinate 1, which is formed as a coproduct of butadiene production from steam cracker C_4 fractions (see Table 4 for typical composition).
2) pseudoraffinate 1, which is obtained by selective hydrogenation of butadiene in mixed C_4 fractions from steam crackers (composition, see Table 4)
3) isobutene contained in the C_4 fraction of fluid catalytic crackers (FCC – C_4's; for typical composition, see Table 4). FCC – C_4 is used as feedstock for ca. 29 % of MTBE production

Table 2. Vapor pressure, density, and water miscibility of MTBE

Temperature, °C	Vapor pressure, ∗ kPa	Density, kg/m^3	Miscibility	
			Water in MTBE, wt %	MTBE in water, wt %
0	10.89	761.3	1.19	7.3
10	17.59	751.0	1.22	5.0
20	27.21	740.7	1.28	3.3
30	40.88	730.4	1.36	2.2
40	59.69	720.1	1.47	1.5

∗ Antoine Parameters: A = 6.254909, B = 265.4, C = 242.517
$\log p = A - B / (C + \vartheta)$; $[\vartheta] = °C$

4) isobutene from dehydrogenation of isobutane, which is obtained both from refineries and from field butanes after isomerization (35 % of MTBE feedstocks); and
5) isobutene by dehydration of *tert*-butanol, a coproduct of propylene oxide synthesis (Halcon/Arco process, see → Propylene Oxide) (ca. 15 % of MTBE feedstocks)

Table 3. Binary azeotropes with MTBE

Azeotrope	*bp*, °C	MTBE content, wt %
MTBE – water	52.6	96∗
MTBE – methanol	51.6	86
MTBE – methanol (1.0 MPa)	130	68
MTBE – methanol (2.5 MPa)	175	54

∗ Condensate separates into two phases.

Raffinate 1 and pseudoraffinate 1 provide isobutene for approximately 21 % of world's total MTBE production [46].

Table 4. Typical composition of C_4 hydrocarbon streams from steam crackers (raffinate 1) and fluid catalytic crackers (FCC – C_4)

Compound	Raffinate 1, wt %	Pseudoraffinate 1, wt %	FCC – C_4, wt %
Isobutane	4	2	36
n-Butane	12	7.5	13
Isobutene	44	24	15
Butene-1	24	39	12
cis-Butene-2	6	8	9
trans-Butene-2	9	19	14
Butadiene-1,3	0.5	0.0	0.3
Balance	0.5	0.5	0.7

As demand for MTBE increases, the first source to be exploited will probably be free FCC – C_4 reserves. Any further expansion – and this also holds for the raffinate 1 route – is thus tied to new cracker construction. Field butanes will grow more than proportionally in importance. The relative share of *tert*-butanol as an isobutene source for MTBE production, on the other hand, is expected to decline because it is formed as a coproduct and thus linked to future propylene oxide demand. Raffinate 1 and FCC – C_4 can be utilized directly in MTBE synthesis. By contrast, isobutane must be dehydrogenated. The same holds true for field butanes after isomerization of the *n*-butane fraction to isobutane. A number of industrial processes have been established during the last decades [4], [47]. For the primary butane isomerization, the Butamer process is most commonly used. For the isobutane dehydrogenation, the commercially available processes Oleflex (UOP), Catofin (ABB Lummus Crest, Inc.), STAR (Phillips Petroleum Co.), and FBD – 4 (Snamprogetti SpA), are at present of industrial importance.

To produce MTBE from *tert*-butanol [5], isobutene must first be obtained by elimination of water from the alcohol before the olefin can be used for ether production.

Methanol (→ Methanol), the second reactant in MTBE synthesis, is produced at a typical purity of > 99.9 % and is used directly for ether synthesis without further purification. Comparison of today's methanol capacities, which amounts to 32×10^6 t/a, and the methanol demand of 26×10^6 t/a [48] would allow an additional increase of MTBE production of ca. 16×10^6 t/a. Some 25 % of the present methanol output is consumed by MTBE.

4. Production

Methyl *tert*-butyl ether can be obtained by the acid-catalyzed addition of methanol to isobutene [6]. Suitable catalysts are solid acids such as bentonites [7], zeolites [8], [49], [50] and – commonly used in industrial world scale MTBE-production units – macroporous acidic ion-exchange resins [9]. The reaction is weakly exothermic with a heat of reaction of − 37.7 kJ/mol.

The kinetics of MTBE formation have been investigated intensively [9], [51], [52]. It has been shown recently, that the observed reaction rates can be described sufficiently both by a kinetic model according to a Langmuir – Hinshelwood mechanism and by an Eley – Rideal (ER) approach [53]. However, the ER-model seems to be the most probable one. The commonly used kinetic model has been developed by Rehfinger et al. [54]. Because of equilibrium limitations only 92 % conversion can be achieved with equimolar amounts of isobutene and methanol at 333 K. An excess of methanol not only increases the conversion of isobutene but also suppresses its dimerization and oligomerization. Dimerization of isobutene is the most important side reaction of MTBE synthesis [55], [56]. At a molar methanol ex-

cess as low as 10 %, the selectivity for MTBE is practically 100 %.

In recent years, in addition to the Snamprogetti [57] and Hüls (now Oxeno) processes [10], [58], [59] processes developed by Arco [11], IFP [60], and CDTECH (ABB Lummus Crest and Chemical Research Licensing) have been established [12].

Other industrial processes have been developed by DEA (formerly Deutsche Texaco) [13], Shell (Netherlands) [14], Phillips Petroleum [60], and Sumitomo [15]. More than 140 MTBE plants with a total installed capacity of ca. 20×10^6 t/a are on stream [44], [46].

Process Description. Most commercially available processes are comparable and consist of a reaction and a refining section.

Reaction Section. All processes have in common the reaction of isobutene with a certain molar excess of methanol on a macroporous acidic ion exchanger at 50 – 90 °C. In the Snamprogetti, Hüls, and Arco processes, the pressure in the reaction section is chosen so that the reactants over the catalyst are entirely in the liquid phase over the catalyst bed, i.e., between 1.0 and 1.5 MPa. This enhances catalyst life time by reducing the polymerization of isobutene at the catalyst's surface and promotes the selectivity for the MTBE formation. For example, only ca. 100 – 200 ppm by weight of isobutene is converted to diisobutene (2.2.4-trimethylpentene-1 and -2) in the "Hüls-MTBE-Process".

Design of the reaction section is largely determined by the heat of reaction, which must be removed. The heat of reaction depends on the isobutene concentration in the feedstock. In the three processes mentioned (Snamprogetti, Hüls, Arco) adiabatic reactors are always employed with a *FCC – C_4 feedstock*. If *raffinate 1* is used as feedstock, Snamprogetti and Oxeno prefer tubular reactors for reasons of process engineering, whereas Arco employs recycle reactors. In the adiabatic reactors employed by IFP (Institute Français du Pétrol) or CDTECH, which also requires a recycle at higher isobutene concentration, an isobutene conversion of approximately 85 % is achieved. The heat of reaction is removed by partial evaporation of the C_4 hydrocarbons in the reactor as the reaction proceeds. This process design, however, gives rise to a certain loss in selectivity for MTBE. A final isobutene conversion of ca. 97 % is achieved by a subsequent *catalytic distillation.*

In most industrial plants, isobutene conversion of 95 – 97 % is sufficient. *Residual butenes* are mainly used for the manufacture of alkylate gasoline, recycled to the cracker, or simply burned. If they are to be utilized for other chemical purposes such as the production of polymer-grade butene-1, the degree of isobutene conversion must be significantly increased. To overcome the conversion limit of 95 to 97 % posed by the chemical equilibrium, a higher excess of methanol in the feed stream can be used. Unfortunately, this excess methanol remains in the MTBE product. It can, nevertheless, be recovered by distillation, e.g., as a methanol-poor minimum boiling azeotrope with MTBE [19]. Not only does this give MTBE with purities > 99.7 % (balance mainly *tert*-butanol produced by a reaction of isobutene with water dissolved in the feedstock), it also results in isobutene conversions of ca. 98 %. In the Oxeno-MTBE process isobutene conversions of > 99.9 % could be achieved in a two-stage process (see Fig. 1). This conversion corresponds to ca. 300 ppm per weight of residual isobutene in the spent C_4 fraction. Usually, if the feedstock is a raffinate 1 stream, the spent C_4 fraction is called raffinate 2.

This high conversion can also be obtained by using highly sulfonated acidic resins in the reaction section followed by an additional catalytic distillation column in the refining section.

Refining Section. Catalytic distillation (CD) or reactive distillation (RD) refers to a process, where both catalytic reaction and distillation are carried out simultaneously in the debutanizer column. From the viewpoint of reaction engineering, this column acts as a two-phase countercurrent flow, fixed-bed catalytic reactor. Usually CD is defined as a process, in which a heterogeneous catalyst is located in a distillation column. A more common term for this operation is RD. This term does not distinguish between a homogeneous or heterogeneous catalyzed reaction in distillation columns.

The most important advantage of using CD for MTBE synthesis lies in the elimination of equilibrium limitation of isobutene conversion as a result of continuous removal of the reaction product MTBE from the reaction mixture.

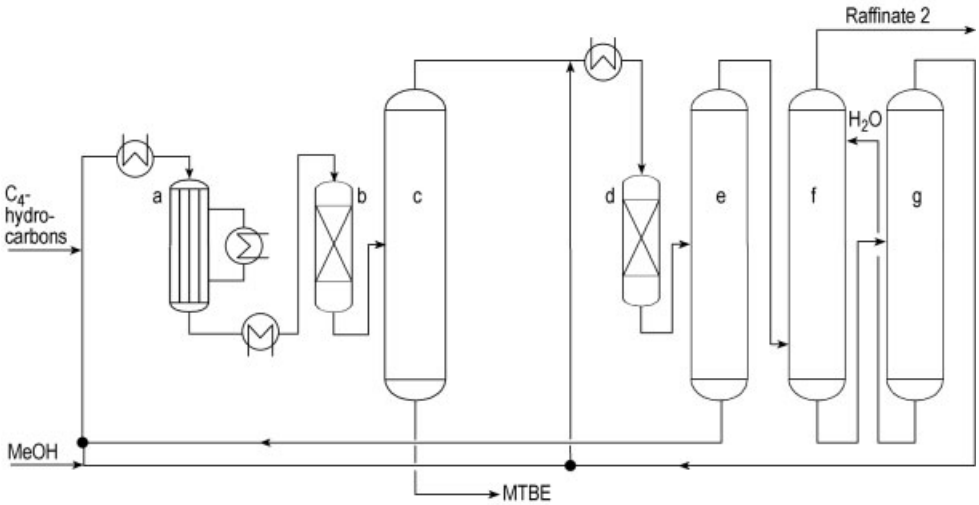

Figure 1. Two-stage Oxeno-MTBE-Process
First stage:
a) Multitubular reactor; b) Adiabatic reactor; c) First C_4 distillation tower (debutanizer)
Second stage:
d) Secondary adiabatic reactor; e) Second C_4 distillation tower
Methanol recovery:
f) Methanol extraction; g) Methanol tower

The partly reacted mixture from the reaction section – which is usually in chemical equilibrium – enters the CD column below the catalyst packing zone to ensure the separation of the high-boiling component MTBE from the feed stream. The catalyst packing is installed in the upper mid portion of the column with normal distillation sections above and below.

The CD-based MTBE process was first developed by CR & L using a patented catalyst support system (commercialized as CD MTBE or CD ETHEROL) [16], [61]. The catalyst used in the CD columns is essentially the same as that employed in the reaction section. It is an acidic ion-exchange resin, which is supported by a fiberglass cloth reinforced with stainless steel wire mesh. These fiberglass cloths are rolled to catalyst bales and stacked on sieve trays in the column.

A CD-process has been developed also by UOP and Hüls in 1992, utilizing Koch Engineering's KataMax-technology [62] (commercialized as ETHERMAX process). In this kind of process technology the catalytic section of the CD-column uses a conventionally structured, wire-cloth distillation packing, which contain pellets of strong acidic ion-exchange resin [63]. The benefit in using this type of packing is a very good distribution of liquid and vapor phase at a low pressure drop, an efficient contact of reactants with catalyst pellets, and instantaneous distillative removal of reactants. Therefore, the packing exhibits excellent separation performance combined with an efficient mass and heat transfer for chemical reaction. An analogous catalytic packing system have been developed by Sulzer (commercialized as KataPak) [64] and Montz (commercialized as MultiPak) [65].

CD-technology has the important advantage over conventional unit design (reaction section followed by debutanizing in a distillation column) of lower capital investment costs, if a very high conversion of isobutene has to be achieved. Because in the CD-process both chemical reaction and distillation are carried out in the same

device, at least one process step is eliminated. Beyond this, the exothermic heat of reaction can be used to vaporize the reaction mixture. Therefore, the maximum temperature in the reaction zone is limited to the boiling point of the mixture, so that an efficient and secure temperature control is possible.

One evident disadvantage of the CD-process technology is the fact, that since the debutanizer column must be taken out of operation during exchange of the catalytic packing, also the whole MTBE unit must be shut down. In contrast, in the conventional design, where the reaction section usually consists of two or more fixed bed reactors (e.g., in the two stage Hüls-MTBE-Process), a continued production can be performed at a lower feed rate, while the catalyst is exchanged. In the CD-technology it must also be ensured that a small excess of methanol exists along the reaction zone. In the absence of methanol the exothermic dimerization of butenes take place at high reaction rates. The ensuing sharp temperature rise ("hot spot") will cause irreversible catalyst deactivation and catalyst damage.

Nevertheless, since the mid-1980s most of the MTBE plants, which have been built, were equipped with a CD-column.

In the removal of nonconverted C_4 hydrocarbons from MTBE by distillation, excess *methanol* appears in the distillate, because of formation of an methanol-C_4 azeotrope. The methanol content in the azeotrope depends on pressure and C_4-composition, e.g., at 0.6 MPa the azeotrope contains ca. 2 to 4 wt % methanol. This methanol can completely be removed from the C_4-stream in an extraction column using a countercurrent water wash. The methanol can be recovered by distillation and is recycled to the reaction section.

If methanol losses of ca. 200 ppm by weight are acceptable, methanol recovery from the distillate with molecular sieves is also appropriate [17]. This 200 ppm residual methanol together with dimethyl ether, also formed in trace amounts, can be separated from the butene fraction in a subsequent molecular sieve adsorption unit [18].

Construction Materials. Because corrosive media are not used anywhere and do not result from catalyst disintegration, the entire plant is fabricated from standard carbon steel. Moreover under normal operating conditions, residues, process wastewater, and gaseous emissions are not formed.

Investment costs (inside battery limits, IBL) for a 200 000 t/a MTBE plant (Oxeno-MTBE-Process) in the Federal Republic of Germany, running on raffinate I feedstock, amount to 14×10^6 DM.

5. Environmental Protection

The small amounts of byproducts (see Chap. 6) in MTBE need not be removed if the product is to be used for gasoline. The catalyst is regenerable, and no environmental problems arise from its disposal.

Because MTBE is soluble in water, suitable precautions against groundwater contamination must be taken in loading and storage areas. Recently published complains about groundwater contamination and the reactions taken by public authorities in parts of the United States arose from careless storage of the product. For the environmental significance of MTBE, see Section 11. For emission control in the storage and handling of MTBE, see Chapter 8.

6. Quality Specifications

The usual purity of commercial MTBE is 98 – 99 wt %. The byproducts, *tert*-butanol and diisobutenes, as well as excess, residual methanol, do not have a detrimental effect on octane number improvement through the use of MTBE. Depending on the quality of the C_4 feedstock mixture, the MTBE product may also contain C_5 and C_6 hydrocarbons. A removal of these components by distillation is not necessary in most cases.

A typical MTBE composition adopted worldwide for use in the fuel sector is given below:

MTBE	98 – 99 wt %
Alcohols (methanol, *tert*-butanol)	0.5 – 1.5 wt %
Hydrocarbons (C_5, C_6-hydrocarbons, diisobutenes)	0.1 – 1.0 wt %
Water	50 – 1500 ppm by weight
Total sulfur	max. 10 ppm by weight
Residue on evaporation	max. 10 ppm by weight

For special applications (see Section 10.2), a high-purity MTBE (ether content > 99.98 % wt)

is marketed under the trade name Driveron S by Oxeno. High purity MTBE is also available from Shell and ARCO.

7. Chemical Analysis

Pure MTBE is analyzed by gas chromatography, preferably in capillary columns with a highly polar stationary phase, e.g., TCEP [1,2,3-tris-(2-cyanoethoxy)propane], Carbowax 20 M, or DX-1. For GC analysis of MTBE-containing fuels, an oxygen-specific detector (O-FID) [20] or a column combination technique [21], [22] can be employed. For other references on analysis of MTBE, see [23], [24], [25].

8. Storage and Transportation

Being noncorrosive, MTBE can be handled in the same way as fuel, and the existing distribution system can also be used for fuels containing MTBE. The ether has an indefinite storage life even in the presence of air. Because water is miscible with MTBE (though only slightly), dry storage is required. The product can be stored in nonpressurized containers. Carbon steel can be used as container material, as well as aluminum, brass, copper, polyethylene, or polypropylene. Teflon, Buna-N, and other fuel-resistant plastics and rubbers can be used for seals; the use of vinylidene fluoride-hexafluoropropene copolymer (Viton) is not recommended.

Methyl *tert*-butyl ether has a vapor pressure of 61 kPa at 40 °C. Emission from storage facilities can be controlled or prevented by ordinary measures. The usual safety precautions for highly inflammable liquids must be employed.

Extinguishing agents for fire fighting are powder, carbon dioxide, and alcohol-resistant foam such as Tutogen L (→ Fire Extinguishing Agents, Chap. 3.2.). With alcohol-resistant foams, a higher application rate is necessary for MTBE than for pure hydrocarbons.

Transport classifications are as follows:

ADR/RID II	class 3, no. 3 b; Name of substance: 33/2398 Methyl-*tert*-butylether
ADNR II	class 3, no. 3 b
GGVSee/IMDG	class 3.1
UN no.	2398
ICAO Code/IATA-DGR II	class 3
UN no.	2398

9. Legal Aspects

The principal reason for the current importance of MTBE is environmental legislation (especially in the United States, Japan, and Western Europe). Although laws do not require the use of MTBE, it is often the best way to satisfy their provisions.

In the *European Community*, the Council Directive 85/536/EEC, December 5, 1985 [26], does not require special labeling at filling stations for fuels containing up to 15 vol % MTBE.

In the *United States*, the use of oxygenates in gasoline is governed by EPA waivers from the Clear Air Act. Sun Refining received a provisional waiver in 1988, permitting the use of up to 15 vol % MTBE (2.7 wt % oxygen) [27].

Legislation has developed since oxygenates in fuel have become generally recognized. To reduce carbon monoxide emission and comply with provisions of the Clean Air Act, some U.S. cities and states mandate the use of ca. 2.7 wt % oxygen in fuel (mainly as MTBE) in winter months.

10. Uses

10.1. As an Octane Enhancer

(→ Automotive Fuels)

More than 95 % of MTBE produced are used in the gasoline pool. In the 1980s – 1990s huge production capacities have been erected in order to cover the demand of MTBE. This process has been driven especially by the CAA regulation in the United States. In addition to many other favorable characteristics, the current importance of MTBE is based primarily on its exceptionally good octane-enhancing properties when used as a gasoline blendstock [28], [31]. These antiknock properties are especially important because the use of cheap but toxic alkyllead compounds has been restricted or banned by law, both on environmental grounds and to permit the use of exhaust catalytic converters. Depending on the composition of the base gasoline, blend octane numbers of 115 – 125 (research octane

number, RON) and 92 – 120 (motor octane number, MON) can be achieved. The highest blend values are obtained with saturated, paraffin-rich gasolines; the lowest, with olefin-rich gasolines. The lead and aromatics content also influences the blend octane numbers of MTBE. Because of the relatively low boiling point of MTBE, the effect on the front-end octane number (FON or R-100 °C-ON, i.e., the research octane number of the fraction with a boiling range < 100 °C) is especially pronounced (see Fig. 2) [29].

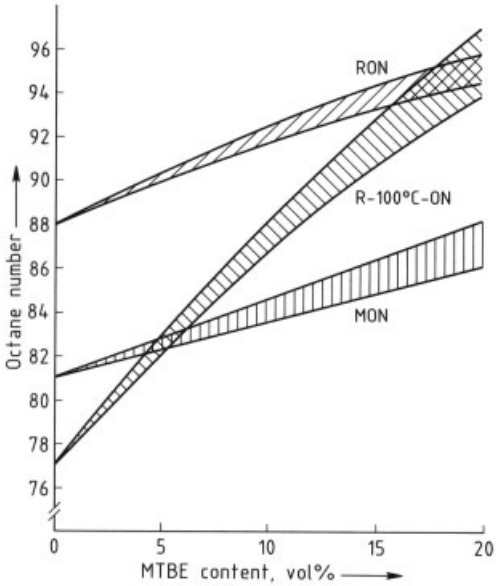

Figure 2. Ranges of octane improvement by addition of MTBE to an unleaded gasoline with RON = 88, MON = 81, and R-100 °C-ON = 77

Besides increasing the octane number, addition of MTBE to fuel has other positive effects. It does not necessitate any modifications to existing vehicles. The fuel vapor pressure (Reid vapor pressure, RVP) is decreased, so that vapor emissions during automobile fueling and operation are reduced. Addition of MTBE cuts down exhaust emissions, particularly carbon monoxide, unburned hydrocarbons, polycyclic aromatics, and particulate carbon. Although MTBE has a somewhat lower heat of combustion than gasoline, addition of up to 20 vol % neither impairs motor power nor increases fuel consumption. Easier cold starting and prevention of carburetor icing are other advantages. With regard to the hot-weather drivability of modern automobiles, no difference exists between fuels containing MTBE and those without it [30]. Because of the miscibility of water with MTBE, the fuel's cloud point is significantly lowered. The MTBE-blended fuels are compatible with all materials used in automobile manufacture (e.g., gaskets; lacquers; metals in the carburetor, injection pump, or elsewhere; and elastomers).

10.2. Other Uses

Methyl *tert*-butyl ether is also used in the petrochemical industry. Production of isobutene by splitting MTBE is the only application that has been used on an industrial scale (for applications of isobutene, → Butenes, Chap. 9.).

By reversal of its formation reaction, MTBE can be cracked to isobutene and methanol on acidic catalysts at > 100 °C [32]. The methanol obtained as a coproduct is recycled to MTBE synthesis. For economic and ecological reasons, the MTBE-cracking route and that involving the splitting of *tert*-butanol obtained by direct hydration of isobutene [33] will continue to displace the conventional sulfuric acid extraction process for the isolation (recovery) of pure isobutene from C_4 streams [34]. At present, isobutene production with MTBE as feedstock is employed by Exxon Chemical and Sumitomo.

Methyl *tert*-butyl ether itself can be used in a number of chemical reactions, for example, the production of methacrolein and methycrylic acid [35] and of isoprene [36]. For other possible applications, see [37–42] .

The lack of acidic hydrogen atoms makes MTBE a suitable *solvent* for chemical reactions such as Grignard reactions. The preferred grade for this purpose is high-purity MTBE with a residual alcohol content < 500 ppm by weight, such as the product marketed by Oxeno under the trade name Driveron S. Because of its negligible tendency to form peroxides, high ignition temperature, and narrow explosion limits, MTBE is a good solvent for analytical use [42]. It is also used as an *extractant*, for example, in solvent dewaxing of hydrocarbon oils [42], [43].

11. Toxicology and Occupational Health

11.1. Toxicological Profile

As regards its widespread application and the amounts used, MTBE ranges within the top 2 % of all tested chemical substances. According to all of the toxicity data reviewed, in respect of European Union regulations MTBE has been classified as a skin irritant.

Acute Toxicity. Tests with different animal species on oral, dermal, and inhalation application routes indicate a low acute toxicity for MTBE (oral LD_{50} well above 2000 mg per kilogram bodyweight, dermal $LD_{50} > 10\,000$ mg per kilogram bodyweight, inhalation LC_{50} values in the range from 85 – 142 mg/L for 4 h). At high concentrations anesthesia is the most immediate effect (CNS depression). At nonlethal concentrations local irritations, typical for organic solvents, have been observed besides transient behavioral effects.

MTBE was found to be only slightly irritant to the rabbit eye but causes reversible moderate to severe skin effects in rabbits. Animal studies showed no sensitizing potential for neat MTBE.

Chronic Toxicity. The main effects observed in repeated dose toxicity studies via the oral and inhalation routes were CNS depression, irritation of the respiratory and gastrointestinal tract and enlargement of the liver. For the oral route the no observed adverse effect level (NOAEL) in a 90 day test is 300 mg/kg bodyweight per day while it is 800 mL/m^3 or 2880 mg/m^3 in 90 days inhalation tests.

Genotoxicity, Mutagenicity, Carcinogenicity. Neither the structure of MTBE nor the comprehensive number of available in vitro and in vivo mutagenicity tests with MTBE and its metabolite *tert*-butanol provide any indication of possible genotoxic or mutagenic effects. The other metabolite, formaldehyde, has a proven potential of DNA damage, but will be rapidly detoxified when formed in the body. Therefore available test data strongly suggest that with regard to MTBE itself and its metabolites no genotoxic activity should be assumed by MTBE exposure. The absence of genotoxic potential is important when discussing possible carcinogenic effects. MTBE produces tumors in rats and mice at high concentrations. No observed adverse effect concentrations (NOAEC) of 400 mL/m^3 by inhalation and 250 mg/kg bodyweight at oral dose are derived from the tests results. The tumors found have been formed via nongenotoxic mechanisms, and are most probably not relevant to humans. Based on the available data several independent bodies, such as the International Association of Research on Cancer (IARC, WHO-agency), the European Union Working Group on Classification and Labelling of Dangerous Substances, and other bodies recently have not come to the conclusion that MTBE is a human carcinogen. Based on the available data MTBE is not classifiable with respect to its carcinogenic potential for humans.

Reproductive Toxicity. Tests on reproductive toxicity and developmental development shows that MTBE is not significantly toxic to reproduction nor to fetal development.

11.2. Human Experience

As the odor threshold of MTBE in air (0.053 mL/m^3 = 0.19 mg/m^3) as well as odor and taste thresholds in water (15 µg/L and 40 µg/L, respectively) are very low, at low exposure levels effects on humans may be transient and subjectively caused by taste and odor. At 180 mg/m^3 no objective symptoms on the CNS and no signs of irritation of the respiratory tract were reported. Although it is not reported, repeated exposure on skin may result in de-fatting of the skin.

As indicated by all reported human experiences there is no evidence of any adverse effect caused by low exposure levels.

11.3. Health Risk Characterization, Occupational Health

Since the exposure concentrations for workers involved in production of MTBE, blending, distribution and handling of gasoline containing MTBE, service-station attendants and garage workers, and also for consumers filling

MTBE containing gasoline in cars are significantly lower than the no effect levels in toxicity tests, MTBE production and use indicates no concern for human health.

However, compliance by workers with the short- and long-term occupational exposure limits and protection against skin contact is necessary to be protective against the described effects of MTBE. The lowest recommended occupational exposure limits are 270 mg/m^3 (75 mL/m^3) as the short-term and 90 mg/m^3 (25 mL/m^3) as the long-term value.

12. Ecotoxicology

12.1. Ecotoxicity Data

The result of a large amount of acute aquatic toxicity test data available for organisms at different trophic levels indicates low toxicity to aquatic organisms. All measured EC_{50} and LC_{50} values for microorganisms, algae, invertebrates and fish are well above 100 mg/L. Because of the fact that the measured concentrations of MTBE in the environment are significantly lower than the no effect levels in the mentioned tests and the calculated no effect levels for sediment organisms and terrestrial organisms like earth worms, it is verified that the risk for environmental organisms is low.

12.2. Behavior in the Environment

A calculation with McKay level 1 partitioning model estimates that in case of a spillage most of the MTBE (95 %) will distribute in the atmosphere. There MTBE is readily degraded by photochemically produced hydroxyl radicals. The half-life for photodegradation in the air is about 3 – 7 days.

Aerobic and anaerobic degradation in water and groundwater depends strongly on the conditions. With inoculum of predominantly domestic sewage nearly no biodegradation occurs within 28 days in water in a standard test. However, with adapted bacteria MTBE is readily biodegradable under aerobic conditions.

As MTBE is relatively easily soluble in water (about 50 g/L) and will not adsorb on soil, it is easily washed into the groundwater. Biodegradation in contaminated soil and groundwater (remediation) is discussed in Chapter 13.

Due to its physical-chemical properties and metabolism in organisms MTBE is not expected to bioaccumulate and this has been shown to be correct by a bioaccumulation test.

13. Removal of MTBE from Impacted Groundwater

As a component of gasoline, MTBE can enter subsurface environments through a number of release scenarios including leaking underground fuel tanks (LUFT), dispensing equipment, accidental spills during tank refills, leaking fuel pipelines, minor on-going spill incidents, and larger volume surface spills. In many cases, gasoline retail stations and production terminals are sources of MTBE in groundwater. The concentration of MTBE in groundwater and the extent of groundwater contamination depend largely on the volume and duration of a gasoline release and whether remediation has been initiated at the site. As with any remedial effort for most contaminants, source removal is critical for the success of remediation systems at MTBE-impacted sites. A quick response following an accidental release of an MTBE-blended gasoline is of great importance regardless of the selected remediation method since it decreases the size of the area requiring characterization and the volume of groundwater needing treatment thereby minimizing site characterization and remediation costs.

13.1. Overview of MTBE Remediation Technologies

The selection of a remedial strategy at an MTBE-impacted site depends on the volume of contaminated soil and groundwater, contaminant concentrations and mass, site-specific hydrogeological characteristics, and risk-based or regulatory-driven cleanup objectives. Of particular interest at MTBE-impacted sites is the behavior of MTBE relative to other gasoline components in subsurface environments.

The properties of MTBE greatly influence its fate in the environment, govern transport

mechanisms, phase partitioning and biological interactions, and impact the success of remedial technologies. Key physiochemical properties of MTBE were presented in Tables 2 and 3. As compared with toluene, a common gasoline component and groundwater contaminant (→ Toluene, Chap. 2.1.), MTBE has a lower density and boiling point, but a higher vapor pressure and much higher water solubility. The properties of MTBE suggest the following regarding its fate in subsurface environments: (1) MTBE does not readily sorb to soil particles in aquifers; (2) MTBE volatilizes readily from non-aqueous phase liquid (NAPL) sources; and (3) MTBE strongly partitions into the aqueous phase and preferentially remains in water. As a consequence, MTBE can be highly mobile in subsurface environments relative to other gasoline constituents.

Contarary to early reports, most of the technologies used historically at gasoline-contaminated sites are effective at removing MTBE from groundwater.

Successful technologies include groundwater extraction followed by air stripping, advanced oxidation or granular activated carbon systems. In-situ technology options include air sparging, multi-phase extraction, soil vapor extraction, bioremediation and phytoremediation. In addition to single technologies traditionally used at gasoline-contaminated sites, emerging technologies, modifications to existing technologies or the use of a treatment train involving more than one technology can greatly improve MTBE removal efficiency and/or reduce life cycle remediation costs.

13.2. Phase Transfer Technologies

Phase transfer technologies move MTBE from the water phase to either the gas phase (e.g., air stripping) or the solid phase (e.g., granular activated carbon) without chemically transforming it. Consequently, additional treatment steps to destroy MTBE (e.g., off-gas treatment from air strippers) or waste disposal (e.g., spent granular activated carbon) may be required.

13.2.1. Ex-Situ Phase Transfer Technologies

13.2.1.1. Air Stripping

Air stripping is a physical process for removing volatile organic compounds from water. In general, the effectiveness of air stripping increases with a compound's Henry's constant and the operating air-to-water ratio. MTBE's Henry's constant is much lower than that of common groundwater contaminants such as toluene (0.018 – 0.026 vs. 0.27, [66]). Thus, high air-to-water ratios (100 to 250) are needed to remove MTBE from water. Air stripping has been used to successfully remove MTBE at concentrations that are typically associated with groundwater contamination from LUFTs [67]. Evaluation of MTBE performance data from eight case studies of low profile and packed tower air strippers showed that air stripping has the potential to be more widely used in both drinking water and remediation applications [68]. For each of the studies evaluated, MTBE removal efficiencies exceeded 90 %. Because state and/or local air quality regulations sometimes required stripper off-gas treatment (granular activated carbon, thermal and catalytic oxidation, advanced oxidation), process costs can increase significantly.

13.2.1.2. Granular Activated Carbon (GAC)

Although MTBE has a low affinity for sorption to solids, several types of activated carbon including coconut shell or coal-based GAC can be used to remove low concentrations of MTBE in water (200 µg/L) cost-effectively [67], [69]. GAC can be used to treat low flows of water from private wells with low MTBE concentrations. For treatment of higher concentrations of MTBE, GAC can be used as a polishing step in a treatment train following air stripping or chemical oxidation [69]. Background water quality and co-contaminant concentrations influence MTBE removal efficiencies. High concentrations of natural organic matter (NOM) and other gasoline constituents in influent water streams compete with MTBE for GAC sorption sites thereby increasing GAC usage rates and process costs.

13.2.1.3. Other Ex-Situ Phase Transfer Technologies

Results from laboratory and pilot scale studies suggest that synthetic resins have the potential for successful remediation of MTBE-impacted sites [67]. One of the advantages of using synthetic resins is the ease of regeneration relative to GAC. One of the limitations of resins is the experimentally observed reduction in MTBE sorption due to the presence of other gasoline constituents, such as benzene, toluene, ethylbenzene, and xylene (BTEX) compounds and *tert*-butanol (TBA, *tert*-butylalcohol), in contaminated water. This is analogous to the effects of NOM and BTEX on GAC sorption as discussed above. The use of synthetic resins is costly relative to GAC but resins can be designed to achieve a higher degree of selectivity and trend to have a longer lifetime than GAC.

In addition to synthetic resins, other sorbents such as high silica zeolites have been shown to be effective in removing MTBE from contaminated water [70].

Other technologies, such as membranes and solvent extraction, can be used to remove MTBE from contaminated groundwater [71], [72], [73] but only a few bench- and pilot-scale studies have been conducted.

13.2.2. In-Situ Phase Transfer Technologies

Demonstrated in-situ phase transfer technologies include soil vapor extraction (SVE), multiphase extraction (MPE), in-situ air sparging (IAS) and phytroremediation. Emerging phase transfer technologies include steam injection/vacuum extraction and six-phase heating.

13.2.2.1. Soil Vapor Extraction (SVE) and Multi-Phase Extraction (MPE)

Soil vapor extraction (SVE) is commonly used for the remediation of vadose zones at gasoline-contaminated sites. In order to remove MTBE from both saturated and unsaturated zones, other remediation technologies can be used in combination with SVE. The success of SVE has been documented at MTBE-impacted sites when implemented either as the sole technology or in conjunction with other groundwater treatment technologies [74–76]. The combination of SVE and groundwater extraction, known as multiphase extraction, is effective in removing MTBE from contaminated aquifers [75] and is routinely used at many sites. Effectiveness with MTBE can be enhanced if the remediation system is operated under high vacuum.

13.2.2.2. In-Situ Air Sparging (IAS)

In-situ air sparging involves the injection of air below the groundwater table to transfer contaminants from the aqueous phase into the gaseous phase. The injection of air has the additional advantage of introducing oxygen to subsurface environments thereby potentially enhancing aerobic contaminant biodegradation rates. Laboratory air sparging column studies in different soils showed that the rate of MTBE volatilization is very rapid in sandy soils [77]. Results from another laboratory study showed that 85 % of MTBE in water equilibrated with gasoline was successfully removed using air sparging [78]. MTBE recovery rates were shown to increase with increased air injection rates and water saturation levels. Results from a number of field studies revealed that MTBE groundwater concentrations at gasoline-contaminated sites were reduced by 88 to 99 % after two years of air sparging [79]. As a general rule, if the site hydrogeology is conducive to successful use of IAS, e.g., soil porosity is satisfactory, this technology could be very successful at removing MTBE from groundwater.

13.2.2.3. Phytoremediation

Phytoremediation involves the use of plants or trees to either remove or degrade groundwater contaminants. Phytoremediation occurs by plant uptake of the contaminants followed by translocation and accumulation into plant tissues, or transpiration through leaves. In some cases, the breakdown of contaminants can take place either within plant tissues or in the rhizosphere [80]. Phytoremediation is most effective for contaminants with log K_{ow} values in the range of 0.5 to 3.0 [81]. Since MTBE has a log K_{ow} of 1.2, phytoremediation is expected to

be an effective technology at MTBE-impacted sites. Several plant and tree species have been recently reported to uptake MTBE including alfalfa plants [82], poplar trees [80], [83], [84] and Eucalyptus trees [84], [85]. In laboratory studies, poplar trees were shown to remove 30 % of the MTBE in test solutions at concentrations of 300 to 1600 µg/L in a week [81]. Results of several laboratory studies suggest that phytoremediation has the potential to be effective for the removal of MTBE from groundwater, especially at sites where the contaminated groundwater is shallow.

13.3. Transformation or Destruction Technologies

The underlying principle of transformation or destruction technologies is the transformation of MTBE in water to another chemical form either partially (e.g., acetone) or completely (carbon dioxide). Transformation or destruction technologies can either be implemented ex-situ following groundwater extraction (e.g., chemical oxidation, biological treatment) or in-situ (e.g., in-situ chemical oxidation, in-situ bioremediation).

13.3.1. Ex-Situ Transformation Technologies

13.3.1.1. Advanced Oxidation Processes (AOPs) (→ Water, Chap. 10.3.4.)

Advanced oxidation processes involve the use of ozone-peroxide, ozone – UV, peroxide – UV, Fenton's reaction (hydrogen peroxide in combination with iron(II)salts), high-energy electron beam irradiation (E-beam) or sonication-hydrodynamic cavitation. Most of these processes are characterized by the production of hydroxyl radicals. AOP technologies have been tested for MTBE destruction in both laboratory and pilot-scale studies. In addition, some full-scale applications have been reported [86–91]. Common problematic byproducts include bromate due to the reaction of ozone with bromide, and TBA, *tert*-butyl ether (TBE), acetone and formaldehyde from the use of UV – H_2O_2 [91]. Some of these compounds are either more toxic than MTBE or more resistant to oxidation [87], [88].

AOPs are believed to be most efficient for groundwater with MTBE concentrations ranging between 0.1 and 80 mg/L [87]. Although the use of AOPs is effective at treating high concentrations of MTBE, AOP systems can be expensive and can lead to the formation and accumulation of toxic or undesirable byproducts. More research to evaluate the cost-effectiveness and performance of AOPs at field-scale levels is needed.

13.3.1.2. Biological Treatment

Alkyl ethers, such as MTBE, are relatively difficult to degrade because of the high energy required by microorganisms to cleave the ether bond and the resistance of the branched carbon structure to microbial attack [92]. Several early studies reported that MTBE is biologically recalcitrant under most environmental conditions [93–95]. Recent laboratory and field studies, however, have shown that a number of bacterial and fungal cultures from various environmental sources are capable of degrading MTBE under aerobic [96], [97] and anaerobic [98–102] conditions. Microorganisms can either utilize MTBE as a sole source of carbon and energy, or can degrade it cometabolically following growth on short-chain alkanes such as propane.

The biodegradation of MTBE has been demonstrated using both suspended growth and fixed film bioreactors [103–109]. In a growth bioreactor microorganisms were able to reduce MTBE concentrations from 2400 mg/L to 1.6 mg/L with an average removal rate of 96.2 % over the course of the study.

Several types of ex-situ reactors can be used for MTBE treatment including fluidized bed, porous pot, membrane and trickling filter bioreactors. Membrane bioreactors sustain high densities of microorganisms and are therefore suited for treating high strength waste streams as well as waste streams containing slowly biodegraded compounds such as MTBE. In one study, the use of a membrane bioreactor reduced MTBE from high (2000 mg/L) to very low (40 g/L) concentrations during normal operation [108]. The potential for success of ex-situ bioreactors has increased over the last several years with the

isolation of pure cultures capable of degrading MTBE and TBA at relatively rapid rates [96]. For example, two pure cultures designates PM1 and ENV735 are being used commercially in fluidized bed reactors (FBRs). Such systems show great promise for biologically treating MTBE contaminated water streams.

13.3.2. In-Situ Transformation Technologies

13.3.2.1. In-Situ Chemical Oxidation

In-situ chemical oxidation (ISCO) is based on the same principles of advanced oxidation. Most ISCO technologies involve the injection of oxidants into groundwater to produce hydroxyl radicals. Common oxidants include peroxide, permanganate, ozone, and Fenton's reagent. In a successful field scale demonstration, hydrogen peroxide was injected at a site in New Jersey. MTBE concentrations dropped from 6000 µg/L to below detection limits in less than a year [111]. Injection of Fenton's reagent into groundwater at a site in Texas reduced MTBE concentrations by 83.4 % [112]. One of the advantages of ISCO is the potential to destroy MTBE in a relatively short time frame. Technology limitations include effective oxidant delivery and the potential of byproduct formation.

In addition to the injection of ozone as an oxidant, technologies involving ozone sparging are currently being tested. Fine bubbles with a high surface-to-volume ratio are injected into the saturated zone to extract dissolved MTBE from contaminated groundwater. Ozone contained within the bubble and thin film around the bubble is expected to react rapidly and destroy MTBE in the gas phase.

13.3.2.2. In-Situ Bioremediation (see also → Bioremediation)

In-situ bioremediation involves the use of indigenous or exogenous microorganisms to either destroy or immobilize contaminants. In-situ bioremediation strategies could involve direct metabolism, cometabolism, bioaugmentation or some combination thereof. Results from pilot studies showed that bioremediation approaches involving oxygen injection and/or bioaugmentation have a strong potential for success at MTBE-impacted sites [110], [113]. In addition, a recent study has demonstrated the success of propane injection and bioaugmentation at an operating service station in New Jersey where a 93 % reduction in MTBE concentration was achieved within two months of system operation [108].

Fortum Oil and Gas Oy and Oxeno Olefinchemie GmbH, two member companies of EFOA (The European Fuel Oxygenates Association – www.efoa.org) have been contributing to the Chapters 11, 12, and 13.

14. References

1. C. Henry, *Chem. Zentralbl.* (1904) 1065.
2. H. C. Barnett et al., *Natl. Advis. Comm. Aeronaut.,* Wartime Report ACR No. E 4 H 03 (1944).
3. *Ullmann,* 4th ed., **17,** 58.
4. K. C. Rock, R. O. Dunn: "Economics of Producing Butene Intermediate Feedstocks from Butanes", *Energ. Prog.* **8** (1988) no. 4, 191 – 195.
5. D. J. Miller: "Arco's Processes for Oxygenated Fuels – MTBE/TAME and ETBE", *Dewitt 1989 Petrochemical Review,* Houston, March 28 – 30, 1989.
6. F. Ancilotti et al., *J. Mol. Catal.* **4** (1978) 37 – 48.
7. A. Gicquel, B. Torck, *J. Catal.* **31** (1983) 27 – 31.
8. J. M. Adams et al., *Clays Clay Miner.* **34** (1986) no. 5, 597 – 609.
9. F. Ancilotti et al., *J. Catal.* **46** (1977) 49 – 57.
10. F. Obenaus: "Wege zur besseren Nutzung der C_4-Kohlenwasserstoffe", *DGMK-Meeting Karlsruhe,* February 8, 1982.
11. Th. A. Beck, D. J. Miller: *Arco MTBE – The Maximum in Flexibility,* Dewitt 1988, Petrochemical Review, Houston, TX., March 23 – 25, 1988.
12. P. H. O. Dixon et al., *Catalytic Destillation Technology and MTBE-Production,* 1989 NPRA Annual Meeting, San Francisco, March 19 – 21, 1989.
13. DEA (formerly Deutsche Texaco), DE 27 06 465, 1987 (H. Humberg, H. Wegner).
14. Shell NL, DE-OS 29 11 077, 1979 (H. Groeneveld, A. Kwantes).
15. Sumitomo Chemical Company, EP-A 0 206 594, 1986 (S. Arakawa et al.).

16. CR & L, EP-A 0 008 860, 1979 (L. A. Smith Jr.).
17. E. Trubac: "Oxygenate Removal from MTBE Raffinate", lecture presented at the *AIChE Meeting,* Minneapolis, Minn., Aug. 16 – 19, 1987.
18. Union Carbide Corp., EP-A 0 229 994, 1987 (M. Nagji).
19. Hüls AG, DE 2 629 769, 1976 (F. Obenaus et al.).
20. W. Schneider, J.-Ch. Frohne, H. Bruderreck, *J. Chromatogr.* **245** (1982) 71 – 83.
21. J. Winskowski, *Erdöl Kohle Erdgas Petrochem.* **41** (1988) no. 1, 30 – 36.
22. J. M. Levy, J. A. Yancey, *J. High Resolut. Chromatogr. Commun. HRC CC J. High Resolut. Chromatogr.* **9** (1986) no. 7, 383 – 387.
23. AFNOR (Association Française de Normalisation) M 07-084, Dec. 1984, Norme experimentale.
24. DIN 51 413, Parts 4 and 5.
25. BSI (British Standards Institution) proposed method JP-STG-6 C (The Institute of Petroleum).
26. EEC Council Directive 85/536/EEC, Dec. 5, 1985 (Column 3).
27. *Chem. Eng. News,* Oct. 24, (1988) 12.
28. G. H. Unzelmann, *Oil Gas J.* (1989, April 10) 33 – 37.
29. F. Obenaus, W. Droste, *Erdöl Kohle Erdgas Petrochem.* **33** (1980) no. 6, 271 – 275.
30. M. Denham, M. Mays, K. Schug, M. Sposini, lecture presented at the *8th International Symposium on Alcohol Fuels,* Nov. 13 – 16, Tokyo 1988.
31. W. J. Piel, *Energ. Prog.* **8** (1988) no. 4, 201 – 204.
32. Fattore et al., *Hydrocarbon Process.* **60** (1981) Aug., 101. August (1981) 101.
33. *Chem. Eng. N.Y.* **90** (1983) no. 25, 60.
34. H. Kröper, K. Schlömer, H. M. Weitz, *Hydrocarbon Process* **48** (1969) no. 9, 195 – 198.
35. Nippon Shokubai Kagaku Kogyo Co., EP-A 0 304 867, 1988 (K. Kinumi et al.).
36. Kuraray Co, JP-Kokai 60 193 932 <85 193 932>, 1985 (Y. Ninagawa et al.). Sumitomo Chem., US 3 574 780, 1971 (Y. Watanabe et al.).
37. Degussa AG, EP 0 050 869, 1981 (B. Lehmann).
38. Anic S.p.A., DE-OS 3 130 428, 1982 (F. Ancillotti et al.). BASF AG, DE-OS 27 40 590, 1979 (F. Merger et al.).
39. BASF AG, DE-OS 29 11 466, 1980 (H. Naarmann et al.).
40. BASF AG, DE-OS 29 28 098, 1981 (W. Hoffmann).
41. Bayer AG, DE-OS 30 02 203, 1981 (G. Bonse et al.).
42. H. Grangette, C. Bernasconi, S. Randriamahefa, R. Gallo, *Inf. Chim.* **246** (1984) Jan./Feb., 135 – 140.
43. Exxon Research and Engineering Co., EP-A 00 88 603, 1983 (T. H. West).
44. *Chem. Week* (1997) Dec. 3, 60.
45. M. L. Morgan, DGMK Conference "C_4 Chemistry – Manufacture and Use of C_4-Hydrocarbons", Oct 6 – 8 (1997) Aachen.
46. K. D. Miller, 23rd Annual Dewitt Petrochemical Review, March 31 – April 2 (1998) Houston.
47. P. R. Sarathy, G. S. Suffride, *Hydrocarbon Process.* (1993) Feb, 43 – 51.
48. *Chem. Week* (1997) Oct. 29, 44.
49. P. Chu, G. H. Kühl, *Ind. Eng. Chem. Res.* **26** (1987) 366.
50. A. Kogelbauer et al., *J. Catal.* **152** (1995) 122 – 129.
51. K. Sundmacher, U. Hoffmann, *Chem. Eng. Sci.* **49 (18)** (1994) 3077 – 3089.
52. T. Zhang, R. Datta, *Ind. Eng. Chem. Res.* **34** (1995) 730 – 740.
53. D. Parra et al., *Chem. Eng. Sci.* **49 (24A)** (1994) 4563 – 4578.
54. A. Rehfinger, U. Hoffmann, *Chem. Eng. Sci.* **45** (1990) no. 6, 1605 – 1617.
55. F. Obenaus, *Oil Gas J.* **77** (1979) 76 – 77.
56. A. Rehfinger, U. Hoffmann, *Chem. Eng. Technol.* **13** (1990) 150 – 156.
57. S. Rossini et al., DGMK-Conference "C_4 Chemistry – Manufacture and Use of C_4-Hydrocarbons", Oct. 6 – 8 (1997), Aachen.
58. *Hydrocarbon Process.* **72** (1996) Nov .,113.
59. W. Droste, *Tech. Mitt.* **80** (1987) 277 – 282.
60. *Hydrocarbon Process.* **75** (1996) Nov, 114.
61. L. A. Smith, M. N. Huddleston, *Hydrocarbon Process.* (1982) March,121 – 123.
62. J. L. DeGarmo et al., *Chem. Eng. Prog.* (1992), March, 43 – 50.
63. Koch Engineering Company, EP-A 0 428 265, 1991 (A. P. Gelbein, M. Buchholz).
64. Gebr. Sulzer AG, EP 0 396 650, 1992 (R. Shelden, J.-P. Stringaro).
65. DE-OS 19 701 045, 1998 (A. Górak, L. Kreul)

66. J. Zogorski, A. Morduchowitz, A. Baehr, B. Baumann, D. Conrad, R. Drew, N. Korte, W. Lapham, J. Pankow, E. Washington: Fuel Oxygenates and Water Quality. Coordinated by Interagency Oxygenation Fuel Assessment, Office of Science and Technology Policy, Executive Office of the President, Washington, DC, 1997.
67. California MTBE Research Partnership: Treatment technologies for removal of methyl *tert*-butyl ether from drinking water. National Water Research Institute, Fountain Valley, CA 2000.
68. R. Deeb, S. Sue, D. Spiers, M. Kavanaugh, Proceedings of the Air and Water Management Association's 94th Annual Conference and Exhibition, Orlando, FL 2001.
69. California MTBE Research Partnership: Treating MTBE-impacted drinking water using granular activated carbon. National Water Research Institute, Fountain Valley, CA (2002).
70. M. Anderson, *Environ. Sci. Technol.* **34** (2000) no. 4, 725 – 727.
71. L. Vane, F. Alvarez, B. Mullis, *Environ. Sci. Technol.* **35** (2001) no. 2, 391 – 397.
72. A. Keller, B. Bierwagen, *Environ. Sci. Technol.* **35** (2001) no. 9, 1875 – 1879.
73. J. Choi, I. Song, W. Lee, *J. Membr. Sci.* **166** (2000) 159 – 175.
74. D. Creek, J. Davidson, Proceedings of the 1998 Petroleum Hydrocarbons and Organic Chemicals in Ground Water: Prevention, Detection, and Remediation, National Ground Water Association, Westerville, OH 1998, pp. 560 – 568.
75. P. Johnson, R. Ettinger, in C. Ward, J. Cherry, M. Scalf (eds.): *Subsurface Restoration,* Ann Arbor Press, Inc., Chelsea, MI 1997.
76. United States Environmental Protection Agency: MTBE fact sheet #2: Remediation of MTBE contaminated soil and groundwater. EPA 510-F-98-002, 1998.
77. A. Mortensen, K. Jensen, T. Sonnenborg, E. Arvin, *Groundwater Monit. Rem.* (2000) 87 – 95.
78. C. Bruce, C. Gilbert, R. Johnson, P. Johnson, in G. Wickramanayake, R. Hinchee (eds.): *Physical, Chemical, and Thermal Technologies: Remediation of Chlorinated and Recalcitrant Compounds,* Battelle Press, Columbus, OH 1998, pp. 293 – 298.
79. California MTBE Research Partnership. Overview of MTBE remediation technologies. National Water Research Institute, Fountain Valley, CA 2002.
80. E. Rubin, A. Ramaswami, *Water Res.* **35** (2001) no. 5, 1348 – 1353.
81. J. Burken, J. Schnoor, *Environ. Sci. Technol.* **32** (1998) no. 21, 3379 – 3385.
82. Q. Zhang, L. Davis, L. Erickson, *Environ. Sci. Technol.* **35** (2001) no. 4, 725 – 731.
83. M. Hong et al., *Environ. Sci. Technol.* **35** (2001) no. 6, 1231 – 1239.
84. L. Newman et al., *Prepr. Extend. Abstr. Am. Chem. Soc.* **40** (2000) no. 1, 288 – 289.
85. L. Newman et al., *J. Soil. Contam.* **7** (1998) no. 4, 531 – 542.
86. J. Acreo et al., *Environ. Sci. Technol.* **35** (2001) no. 21, 4252 – 4259.
87. S. Cater, M. Stefan, J. Bolton, A. Safarzadeh-Amiri, *Environ. Sci. Technol.* **34** (2000) no. 4, 659 – 662.
88. P. Chang, T. Young, *Water Res.* **34** (2000) no. 8, 2233 – 2240.
89. J. Kang, H. Hung, A. Lin, M. Hoffman, *Environ. Sci. Technol.* **33** (1999) no. 18, 3199 – 3205.
90. A. Safarzadeh-Amiri, *Water Res.* **35** (2001) no. 15, 3706 – 3714.
91. M. Stefan, J. Mack, J. Bolton, *Environ. Sci. Technol.* **34** (2000) no. 4, 650 – 658.
92. G. White, N. Russell, E. Tidswell, *Microbiol. Rev.* **60** (1996) 216 – 232.
93. Y. Fujiwara, T. Kinoshita, H. Sato, I. Kojima, *Yugagaku* **33** (1984) 111 – 114.
94. H. Jensen, E. Arvin, in F. Arendt, M. Hinsenveld, W. van den Brink (eds.): *Contaminated Soil '90,* Kluwer Academic Publishers, Dordrecht 1990, 445 – 448.
95. J. Suflita, M. Mormile, *Environ. Sci. Technol.* **27** (1993) 976 – 978.
96. R. Deeb, K. Scow, L. Alvarez-Cohen, *Biodegradation* **11** (2000) no. 2 – 3, 171 – 186. C. Yeh, J. Novak, *Water Environ. Res.* **67** (1995) 828 – 834.
97. A. Stocking et al., *Biodegradation* **11** (2000) no. 2 – 3, 187 – 201.
98. P. Bradley, F. Chapelle, J. Landmeyer, *Appl. Environ. Microbiol.* **67** (2001) no. 4, 1975 – 1978.
99. P. Bradley, F. Chapelle, J. Landmeyer, *Environ. Sci. Technol.* **35** (2001) no. 23, 4643 – 4647.
100. K. Finneran, D. Lovley, *Environ. Sci. Technol.* **35** (2001) no. 9, 1785 – 1790.
101. J. Landmeyer, F. Chapelle, H. Herlong, P. Bradley, *Environ. Sci. Technol.* **35** (2001) no. 6, 1118 – 1126.

102. United States Environmental Protection Agency: Natural attenuation of MTBE in the subsurface under methanogenic conditions. EPA/600/R-00/006, 2000.
103. K. Acuna-Askar, A. Englande, C. Hu, G. Jin, *Water Sci. Technol.* **42** (2000) no. 5 – 6, 153 – 161.
104. D. Dupasquier, S. Revah, A. Auria, *Environ. Sci. Technol.* **36** (2002) no. 2, 247 – 253.
105. N. Fortin, M. Deshusses, *Environ. Sci. Technol.* **33** (1999) no. 17, 2980 – 2986.
106. N. Fortin, M. Deshusses, *Environ. Sci. Technol.* **33** (1999) no. 17, 2987 – 2991.
107. P. Hatzinger, R. Steffan, S. Drew, *Contaminated Soil, Sediment and Water Journal, AEHS,* 2001, 81 – 84.
108. R. Steffan, P. Hatzinger, Y. Farhan, S. Drew, *Proceedings of the Petroleum Hydrocarbons and Organic Chemicals in Ground Water.* November 14 – 16, Houston, Texas 2001, pp. 252 – 263.
109. S. Vainberg, A. Togna, P. Sutton, R. Steffan, Submitted to *J. Env. Eng.* 2001, in press.
110. J. Salanitro et al., *Environ. Sci. Technol.* **34** (2000) no. 19, 4152 – 4162.
111. United States Environmental Protection Agency. Field applications of *in*-situ remediation technologies: Chemical oxidation. EPA 542-R-98-008, 1998.
112. J. Leethem, *Contaminated Soil, Sediment and Water Journal, AEHS,* 2001, 54 – 58.
113. R. Wilson, D. Mackay, K. Scow, *Environ. Sci. Technol.* **36** (2002) no. 2, 190 – 199.

Methyl Cellulose → Cellulose Ethers
Methyl Chloride → Chlorinated Hydrocarbons
Methyl Ethyl Ketone → 2-Butanone
Methyl Iodide → Iodine and Iodine Compounds
Methyl Isobutyl Ketone → Ketones
Methyl-2-Pentanone(4-,3-) → Ketones
Methyl-1-Pentene(4-) Polymers → Polyolefins
Methyl-2-Pyrrolidone(N-) → Pyrrole
Methyl-2-Pyrrolidone(N-) → 2-Pyrrolidine
Methylene Chloride → Chlorinated Hydrocarbons

Mica

RICHARD J. BENBOW, Industrial Minerals Division, Metal Bulletin plc, London, United Kingdom (Chaps. 1 – 7)

BERNARD H. W. S. DE JONG, Institute of Earth Sciences and the Debye Institute, University of Utrecht, Utrecht, The Netherlands (Chap. 8, in part)

JANE W. ADAMS, Corning Research, Corning, New York 14 831, United States (Chap. 8, in part)

1. Introduction 1
2. Structure and Properties 1
3. Occurrence 4
4. Processing 5
4.1. Sheet Mica 5
4.2. Ground Mica 6
5. Uses 7
5.1. Sheet Mica 7
5.2. Ground Mica 8
6. Toxicology and Occupational Health . 9
7. Economic Aspects 10
8. Fluorinated Micas 10
9. References 13

1. Introduction

Mica consists of a group or family of closely related silica-based minerals which share a number of characteristic physical, chemical, and optical properties. It has had mystical associations for millenia. In classical Hindu writings from 2000 B.C., mica was regarded as petrified lightning flashes and was believed to possess potent medicinal powers. It was known as Abrak and Abhra, Hindu words for clouds and heaven. The actual name mica is thought to be derived from the Latin "micare", meaning to shine or glitter, a term used by AGRICOLA in 1546.

2. Structure and Properties

Structure. Micas are members of a class of silicates known as phyllo- or sheet silicates (→ Silicates), a term which reflects their crystal structure. The general chemical formula for mica is $W(X,Y)_{2-3}Z_4O_{10}(OH,F)_2$, where W corresponds to K, Na, and Ca, or more rarely Ba, Rb, and Cs. The X, Y site is occupied by any two ions of Al, Mg, Fe, or Li, and less frequently of Mn, Cr, and Ti. Normally Z is Si or Al, but may also be Fe or Ti. As with other silicates the primary building unit of mica is the SiO_4 tetrahedron. The tetrahedra are linked together via their three basal oxygen ions to form a network of hexagonal cells with the apical oxygen ions all pointing in the same direction (Fig. 1). The idealized basic structural unit of mica comprises two layers of tetrahedra in which the apical oxygen ions point toward each other. The voids between the bases of tetrahedra in adjacent layers and between adjacent apical oxygen ions provide the cation sites (Figs. 2 and 3). The interapical plane is occupied by the octahedrally coordinated X, Y ions which may be any two of Al, Mg, Fe, or Li. The W sites (filled by K, Na, and less frequently Ca cations) are in 12-fold coordination with the basal oxygen ions. Most commonly these stacking sequences lead to either one- or two-layered monoclinic cells (denoted as 1 M and 2 M_1), an alternative two-layered monoclinic cell (2 M_2), or a three-layered trigonal unit (3 T).

Micas can be subdivided into di-octahedral and tri-octahedral classes where the number of X and Y ions is 4 and 6, respectively.

The various types of mica are summarized in Table 1. The most commonly occurring forms are *muscovite* (potassium – aluminum) and *phlogopite* (magnesium – iron). Biotite and phlogopite are members of a continuous structural and chemical series. Differentiation between them is somewhat arbitrary; some authorities use the Mg : Fe ratio as a criterion, phlogopite having an Mg : Fe ratio $> 2:1$ and biotite a ratio $< 2:1$ [1], [2].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_551

Table 1. Most important types of mica

Name	CAS registry number	Formula	Unit cell dimensions $a : b : c$, nm	Cell designation
Muscovite	[*99401-63-5*]	$K_2Al_4[Si_6Al_2O_{20}](OH,F)_4$	0.519 : 0.904 : 2.008	$2M_1$
Biotite	[*112593-95-0*]	$K_2(Mg,Fe^{2+})_{6-4}(Fe^{3+},Al,Ti)_{0-2}[Si_{6-5}Al_{2-3}O_{20}](OH,F)_4$	0.53 : 0.92 : 1.02	1M
Phlogopite	[*110710-26-4*]	$K_2(Mg,Fe^{2+})_6[Si_6Al_2O_{20}](OH,F)_4$	0.5314 : 0.9204 : 1.0314	1M
Lepidolite	[*114705-28-1*]	$K_2(Li,Al)_{5-6}[Si_{6-7}Al_{2-1}O_{20}](OH,F)_4$	0.53 : 0.92 : 1.02	1M
			0.92 : 0.53 : 2.00	$2M_2$
		$3K(Li,Al)_{2.5-3.0}[Si_{3-3.5}Al_{1-0.5}O_{10}](OH,F)_2$	0.53 : – : 3.00	3T
Zinnwaldite	[*116813-68-4*]	$K_2(Fe^{2+}_{2-1},Li_{2-3}Al_2)[Si_{6-7}Al_{2-1}O_{20}](OH,F)_4$	0.527 : 0.909 : 1.007	1M
Paragonite	[*106495-33-4*]	$Na_2Al_4[Si_6Al_2O_{20}](OH)_4$	similar to muscovite polymorph	$2M_1$
Glauconite	[*102785-61-5*]	$(K,Na,Ca)_{1.2-2.0}(Fe^{3+},Al,Fe^{2+},Mg)_{4.0}[Si_{7-7.6}Al_{1-0.4}O_{20}](OH)_4 \cdot n\,(H_2O)$	similar to biotite 1M polymorph	
Margarite	[*1318-86-1*]	$Ca_2Al_4[Si_4Al_4O_{20}](OH)_4$	0.513 : 0.892 : 1.950	$2M_1$
Clintonite, xanthophyllite	[*12199-34-7*]	$Ca_2(Mg,Fe)_{4.6}Al_{1.4}[Si_{2.5}Al_{5.5}O_{20}](OH)_4$		$2M_1$
				1M

The less common micas include the lithia-rich varieties, *lepidolite* and *zinnwaldite* (→ Lithium→ Lithium), as well as the sodic variants, *paragonite* and *glauconite*. Substitution of Ca for K and Na in the W sites produces the so-called brittle micas– *margarite*, *clintonite,* and *xanthophyllite*.

The presence of cations such as Mg, Fe, Al, and Li in the crystal lattice introduces a stagger which manifests itself in the monoclinic symmetry of the mineral. A feature also shared by most micas is their water content which occurs as hydroxyl ions in the center of the hexagons [4 (O) 2 (OH) planes in Fig. 2 A, see also Fig. 3] and can attain 4 – 5 wt %.

Figure 1. Illustration of a tetrahedral $(Si,Al)_4O_{10}$ layer in mica with apical oxygen atoms pointing upward [2]
Open circles indicate oxygen atoms, filled circles indicate silicon (or aluminum) atoms.

Stains occur in sheet mica and can take the form either of mineral inclusions or discoloration due to the presence of inorganic and organic compounds. These stains are detrimental to the quality of the sheet and may be coincident with, or cut across, the cleavage plane. Examples include those formed during mica crystallization and include mottling, mineral inclusions, and intergrowths. Circulating groundwater may also cause staining.

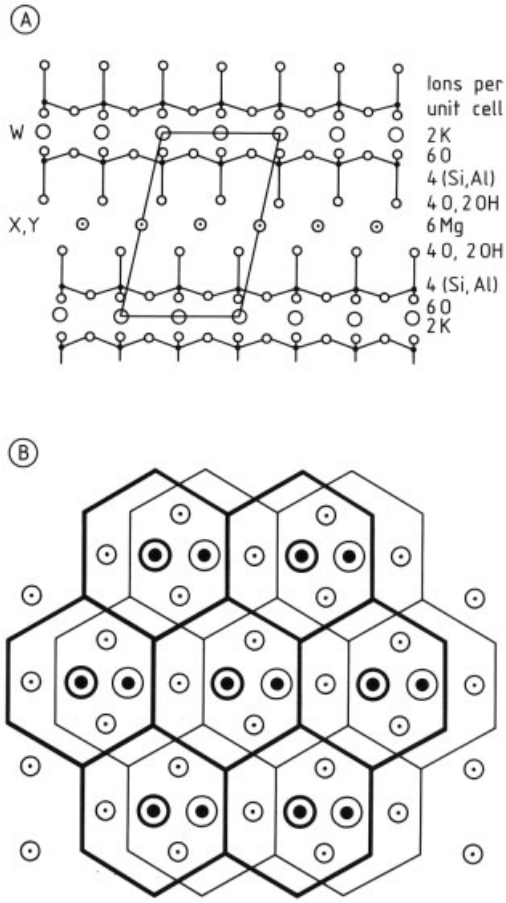

Figure 2. Illustration of the simplest unit cell of mica [2]
A) Vertical section through the tetrahedral $(Si, Al)_2O_5$ layers showing number of ions per unit cell
B) Lower (——) and upper (——) tetrahedral $(Si, Al)_2O_5$ layers viewed from above
⊙ Octahedrally coordinated X, Y cations (mainly Al, Mg, Fe, or Li); • Additional hydroxyl ions; ○ W ions below lower layer (K, Na, Ca); ○ W ions above lower layer (K, Na, Ca).

Properties. Physical properties of muscovite and phlogopite are listed in Table 2. Possibly the most characteristic physical property shared by mica minerals is a perfect basal cleavage which permits them to be split into thin sheets. This is an expression of the layered crystal structure of the mica lattice. A special physical characteristic of mica is its flexibility and elasticity which distinguishes it from minerals such as vermiculite, talc, and chlorite.

Micas are generally resistant to chemical attack with the exception of muscovite (which decomposes in hydrofluoric acid) and phlogopite (which is mildly attacked by hydrochloric acid and decomposes in hot sulfuric acid). Micas can withstand elevated temperatures, muscovite being stable up to 400 – 500 °C, whereas phlogopite decomposes at ca. 850 – 1000 °C. For this reason mica is used as a thermal insulator and optical window in a variety of furnace applications.

In addition to its inertness and thermal resilience, mica also possesses valuable electrical properties including high dielectric strength, uniform dielectric constant, capacitance stability, low power loss, and high electrical resistivity. Hardness typically varies in the range 2 – 3 on the Mohs scale although in brittle micas it can exceed 4.

Muscovite tends to be lighter in color (from colorless to various shades of green – red) than phlogopite and biotite. The color of phlogopite is determined principally by its iron and titanium content. High titanium levels produce a red –brown shade whereas low titanium and high iron impart a green – brown hue. Biotite is generally much darker, often occurring in shades of brown and black.

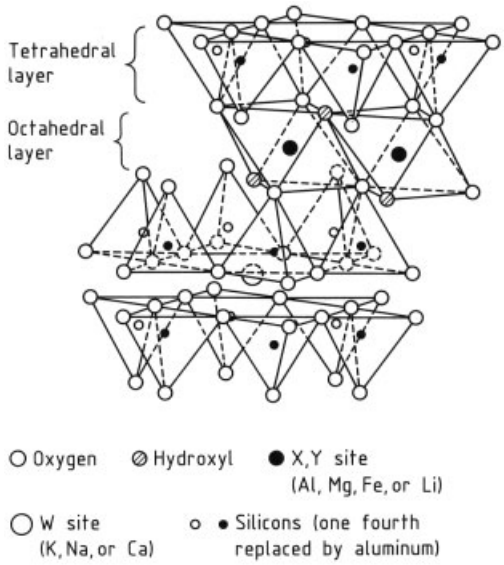

Figure 3. Three-dimensional structure of mica [2]
In fluoromicas the hydroxyl ions are partially or totally replaced by fluoride ions.

Table 2. Physical properties of mica

Property	Muscovite	Phlogopite
Density, g/cm^3	2.75 – 3.1	2.75 – 3.2
Modulus of elasticity, kN/mm^2	160 – 220	160 – 190
Tensile strength, N/mm^2	250 – 300	250 – 300
Mohs hardness	2.8 – 3.2	2.5 – 2.8
Electric strength, kV/mm for		
0.1 mm thickness	200	150
Volume resistivity, $\Omega \cdot$ cm		
23 °C	$10^{16} - 10^{17}$	10^{14}
500 °C	$10^{8} - 10^{10}$	10^{10}
Relative permittivity at 1 MHz	6 – 8	5 – 6
Dielectric loss at 1 MHz		
23 °C	10^{-4}	10^{-3}
100 °C	10^{-3}	5×10^{-2}
Thermal conductivity, W m^{-1} K^{-1}	0.3 – 0.8	0.4
Linear thermal expansion coefficient, 10^{-6} K^{-1}	10	13
Thermal resistance, °C	500	700
Specific heat capacity, J kg^{-1} K^{-1}	850	850

3. Occurrence

Muscovite occurs in various geological environments, including acid igneous rocks such as granites and pegmatites, and also in regionally metamorphosed sediments exemplified by phyllites, schists, and gneisses. It occurs less frequently in sedimentary rocks.

Biotite is the most geologically ubiquitous mica, being the most common mica in acid igneous rocks. It also occurs in gabbros, norites, and diorites and is a characteristic component of calcalkaline rocks. Less frequently it is found in lavas such as rhyolite, trachyte, andesite, and basalt. *Phlogopite* has two principal occurrences, as a component of ultrabasic rocks such as peridotite, and also in metamorphosed limestones.

The lithium micas, *lepidolite* and *zinnwaldite*, are found principally in pegmatites and high-temperature veins. *Glauconite* occurs solely in sedimentary environments.

Only two types of mica are commercially important–muscovite and phlogopite; lepidolite was formerly extracted as a source of lithium (→ Lithium and Lithium Compounds). Two main forms of commercial mica are produced–sheet and ground mica. India is the primary source of sheet mica while the United States is the major producer of ground mica (predominantly muscovite).

North America. In the United States the principal mica-bearing regions are the Blue Ridge and Piedmont provinces of North and South Carolina. North Carolina produces ca. 60 % of the total U.S. output, derived principally from alaskite in the Spruce Pine district which contains 60 % feldspar, 25 % quartz, and 15 % muscovite [3].

Kings Mountain in the Piedmont province of North Carolina hosts a belt of lithium-bearing pegmatite dikes which contain feldspar, quartz, and ca. 60 % muscovite.

Several mica-producing plants exist in South Carolina, one of which supplies sericite (a fine-grained muscovite) from a schist.

In the west of the United States, muscovite is produced in small tonnages from pegmatites and mica schists. In the Picuris range in New Mexico mica is derived from a Precambrian quartz – muscovite schist. The ore typically comprises 38 % muscovite and 60 % quartz [4]. In South Dakota one pegmatite source of by-product mica is currently being exploited. There is also some pegmatite production in the New England region.

In Canada mica production is confined to the large phlogopite deposit at Suzor in Quebec. This medium- to coarse-grained intrusive rock comprises 90 % phlogopite, 8 % pyroxene, and 2 % feldspar. Estimated mica reserves within 50 m of the surface total 27×10^6 t [5].

South America. Brazilian muscovite is derived from a belt of pegmatites some 450 km long running parallel to the Minas Gerais coastline. The pegmatites tend to be zoned, with mica occurring in cavities as well as disseminated throughout the body. They often exhibit preferential weathering and the bodies show considerable variation in form, size, and attitude to the country rocks. The restricted size and remoteness of these bodies severely limit their exploitation [6].

India. India is the world's principal source of sheet mica, accounting for 80 % of global supply. In addition it is an important producer of mica scrap. Bihar is the major producing state, accounting for more than 50 % of output, while Andhra Pradesh produces almost 30 %. The balance comes from Rajasthan, Tamil Nadu, Ker-

ala, and Orissa. Practically all output is muscovite although Kerala produces phlogopite [7].

The mica from the three principal state producers is derived from granitic pegmatites. These bodies vary in attitude from horizontal to subvertical while thickness is in the 1 – 30 m range.

Bihar is the country's principal source of premium-grade sheet muscovite and is renowned for its ruby mica. The production zone is 32 km wide and 145 km long. The mica-bearing pegmatites occur in mica schists and micaceous quartzite.

Rajasthan is the second most important Indian source with production confined to a belt 100 km wide and 320 km long. The rocks are similar to those in Bihar and the pegmatite dikes are steeply dipping bodies, some of which are zoned and possess a complex mineralogy [8].

Most Andhra Pradesh output comes from the Nellore mica district, an arc-shaped zone 100 km long and 15 km wide. Around 70 % of micaceous pegmatites occur in biotite schist and 10 % in hornblende schist [9]. Most output is of green muscovite.

Africa. There are three main sources of mica in Africa–Madagascar, the Republic of South Africa, and Zimbabwe. Tanzania was formerly a source of sheet mica.

Madagascar, along with Canada and Finland, is one of the world's principal sources of phlogopite. The mineral is mined in the southeastern part of the country where it occurs in 0.3 – 2 m thick pegmatite dikes and pockets in association with pyroxenite bodies. Underground mining is carried out in the Berakety and the Ambovombe districts. Nominal amounts of sheet mica are produced [3].

In the Republic of South Africa, muscovite is produced in the Letaba district of eastern Transvaal where the mica occurs in coarse-grained pegmatites. Mica also occurs in the Kenhardt and Namaqualand districts of the Cape Province where it is hosted by complex pegmatites. Mica production from pegmatite source rocks in Zimbabwe is concentrated in a region northwest of Harare.

Europe. In Europe phlogopite is produced as a coproduct of apatite mining at Siilinjarvi in central Finland from a subvertical tabular micaceous carbonatite ore body [10]. Some mica is also produced in Norway as a byproduct of feldspar mining. France (principally Brittany) and Spain produce sizeable quantities of mica as a byproduct of kaolin extraction.

former Soviet Union and China. The former Soviet Union and China are both major mica producers. In the former Soviet Union mica occurs in Karelia (where it is produced as a byproduct of feldspar extraction from zoned pegmatites), the Kola Peninsula, the East Sayan region, and the Kondakovsky region.

4. Processing

Mica can be used in two principal forms–sheet and ground. The former is categorized in terms of sheet purity, sheet thickness, and useable area. Ground mica is processed either dry or wet.

4.1. Sheet Mica

Sheet mica occurs naturally in pegmatites as hexagonal blocks which, owing to the mineral's perfect basal cleavage, are referred to as "books". Its mode of occurrence and the nature of the mineral require sheet mica to be mined and dressed by hand. Upon removal the "book" is cobbed or cleaned by hand to remove adherent quartz or feldspar, and then split with knives into successively thinner sheets. Block, thins, films, and splittings are the four categories of sheet mica classified according to lamina thickness and defined by the Indian Standards Institute, the British Standards, and the American Society for Testing and Materials.

Block mica is knife-dressed sheet mica of which 95 wt % has a minimum thickness of 0.2 mm, with the remainder having a minimum thickness of 0.18 mm. The minimum useable area is 625 mm^2. *Thins* are defined as knife-dressed sheets with a thickness of 0.05 – 0.18 mm. *Films* are split from superior quality block mica; they overlap with thins in that they have thicknesses of 0.02 – 0.18 mm. *Splittings* are laminae with a maximum thickness of 0.03 mm and a minimum useable area of 450 mm^2.

Table 3. Indian areal classification system for mica [11]

Designation		Useable rectangle,	Minimum dimension,
Old ∗	New	cm^2	cm
OOEE Sp.	630	>645.2	10.2
OEE Sp.	500	516.1 – 645.2	10.2
EE Sp.	400	387.1 – 516.1	10.2
E Sp.	315	309.7 – 387.1	10.2
Sp.	250	232.3 – 309.7	8.9
1	160	154.8 – 232.3	7.6
2	100	96.8 – 154.8	5.1
3	63	64.5 – 96.8	5.1
4	40	38.7 – 64.5	3.8
5	20	19.4 – 38.7	2.5
5.5	16	14.5 – 19.4	2.2
6	06	6.4 – 14.5	1.9
7	05	4.8 – 6.4	1.6

∗ O = over; E = extra; Sp. = special.

In addition to classifying sheet mica on the basis of lamina thickness, a grading system also exists for sheet purity and the maximum useable rectangular area that can be cut from a single lamina. Muscovite sheet can be also divided into two color groups–ruby and green. Minor differences exist between the Indian and American classification systems: the Indian system classifies according to 13 grades (Table 3) whereas the American system uses 12. Similarly, on the basis of sheet clarity and freedom from deleterious stains, the Indian system identifies 16 grades (Table 4).

Table 4. Indian visual classification system for mica

Designation	Appearance
V-1	ruby clear
V-2	ruby clear and slightly stained
V-3	ruby fair stained
V-4	ruby good stained
V-5	ruby stained “A”
V-6	ruby AQ
V-7	ruby stained “B”
V-8	ruby BQ
V-9	ruby heavy stained
V-10	ruby densely stained
V-11	black dotted
V-12	black spotted
V-13	black/red stained
V-14	green/brown 1st quality
V-15	green/brown 2nd quality
V-16	green/brown stained or BQ

Small offcuts of sheet mica may be bonded together artificially to form *worked* or *manufactured mica* which is generally classified into three categories [13]:

1) *Micanite or Built-up Mica.* This is produced from thin interleaved splittings arranged in layers and cemented by an artificial binder and compressed at 170 – 180 °C. Typical binders include shellac, epoxy resins, and silicone resins.
2) *Reconstituted Mica or Mica Paper.* This consists of fine flakes (40 – 2000 µm) of high-quality scrap mica pressed into sheets in a continuous film and generally impregnated with a binder.
3) *Glass-Bonded Mica.* This consists of fine particles bonded by borate or borosilicate glass with a low melting point.

4.2. Ground Mica

The raw material used as feedstock for ground mica is mica scrap. Crude mica ore is crushed with jaw and roll crushers. Originally mica scrap was discarded waste from the mining and production of sheet mica but nowadays scrap is predominantly derived from raw material unsuitable for sheet mica. More than 90 % of ground mica output relies on muscovite scrap obtained from pegmatites or as a byproduct from feldspar and kaolin extraction [14]. The origin, type, and quality of the raw material determines its end use.

Dry Ground Mica. Production of dry ground mica in the United States is influenced by its use in plasterboard jointing compounds. In this respect the United States differs from most other countries. Most production utilizes < 3.35 mm (6 mesh) scrap as feedstock which is subjected to air micronization; the oversize fraction > 150 – 250 μm (10 – 60 mesh) is removed on dry screens. Mica used in oil-well drilling is generally a byproduct of feldspar and kaolin extraction, having been segregated by flotation or water classification. The mica is dried and then sieved with multideck screens, the oversize particles (> 8 mm) are subjected to impact milling.

Outside the United States dry ground mica is obtained primarily from premium scrap (> 25 mm) which is introduced into high-speed hammer mills before being "topped" and "tailed" through vibrating screens. These products tend to be highly delaminated and are termed high aspect ratio flakes and powders. Some oil-well grade mica is derived from the production of mica scrap at the mine by "winnowing" the output from the primary crusher through air boxes and cyclones [14].

Wet Ground Mica. In wet grinding high-quality mica scrap is ground with up to 20 wt % water in edge-runner mills. It is then dried and classified to 150, 75, and 45 μm (100, 200, and 325 mesh sizes, respectively). Wet ground mica particles have a typical thickness of 0.1 – 0.8 μm and a high aspect ratio.

Due to the nature of the wet grinding process, the resultant flakes possess smooth edges and tend to be quite flat with good lubricity characteristics. In contrast dry ground mica particles tend to have a poor aspect ratio. The milling process also produces irregular edges which adversely affect lubricity. Consequently wet ground or micronized mica (see below) tends to be used in higher value markets.

Micronized mica is produced by grinding fine-grained byproduct scrap to give a particle size below 40 μm. This is achieved by propelling the particles at high speed into a confined enclosure which results in interparticle comminution. Superheated steam or compressed air can be used as the vehicle to effect micronization, particle size being controlled by residence time in the mill and removal of oversize particles by air classification [14].

5. Uses

5.1. Sheet Mica

The electrical industry has traditionally been the most important market for sheet mica, accounting for over 90 % of output. Within this industry sheet mica enjoys a number of applications by virtue of its thermal and electrical insulating properties, its inertness and resistance to chemical attack, and its translucency (→ Insulation, Electric, Chap. 3.1.1.1.).

Sheet mica possesses several unique electrical properties including high dielectric strength, uniform dielectric constant, capacitance stability, low power loss, and high electrical resistivity. Formerly sheet mica was used in the electrical heating elements of household appliances. Nowadays this market generally uses the cheaper reconstituted micanite product.

Vacuum tubes and capacitors for radios and televisions still incorporate sheet mica, although in color televisions mica has been largely replaced by solid state electronics.

Stacked and silvered mica are used in capacitors. Stacked mica consists of a multilayered "sandwich" of thin mica plates alternating with silver foil. The mica flakes have a typical thickness of 10 – 100 μm. Capacitance is determined by the number of repetitions in the stack and also by the mica thickness, the thinner the mica the higher the capacitance. In silvered mica, silver electrodes are printed and fired onto the mica plates which are then cut to the requisite size and the edges reinforced by silver paste. The plates are then stacked to give the required capacitance. Stacked mica is commonly used in transmitter circuits whereas the silvered variety (available in riveted, clamped, and sintered forms) finds applications in oscillators, filters, transmission circuits, and delay lines. Alternative capacitor materials (ceramics, glass, and plastic) although cheaper than mica, do not possess its range of properties which include dielectric stability at high frequencies up to 500 MHz. A tolerance of ± 1 % can be achieved with mica and mica capacitors exhibit negligible drift. High-grade ruby sheet is favored in capacitor applications

because of its chemical purity. The green variant is not suitable owing to the presence of deleterious elements such as iron which cause current leakage and are consequently intrinsically weak insulators.

The electrical, mechanical, and thermal properties of mica are utilized in the armaments industry. Phlogopite is incorporated in the battery units of heat-seeking missiles. By virtue of its softness and superior thermal insulation characteristics, Madagascar phlogopite is reportedly a favored medium for this application [13]. This mineral is capable of maintaining its thermal integrity at 900 – 1000 °C, giving it a 300 °C advantage over muscovite.

Mica is used for diaphragms in underground breathing apparatus for miners. Tanzanian mica is preferred because of its flatness which permits a good airtight seal. Gauge protectors in high-pressure boilers utilize the translucency, thermal resilience, and mechanical strength of mica. In addition resistance of mica to attack by many hazardous chemicals is a further asset. Superior-grade ruby or green mica is also used for circular compass cards in naval gyro compasses. Green muscovite is also commonly used in optical pyrometers; other optical applications include retardant plates which utilize the birefringent properties of the mineral.

5.2. Ground Mica

Particle sizes of ground mica products used in various applications are listed in Table 5.

Jointing Compounds and Fire Boards. Dry ground mica used in joint cements consisting of gypsum-based plaster imparts crack resistance and promotes smooth finishing. Mica is used at levels of 10 – 12 wt % in such cements to fill the gaps between plasterboard sheets. Its main function is to inhibit the formation of cracks that would allow passage of fire. In gypsum-based plaster cements its platy nature (i.e., occurrence in the form of flat flakes) aids smoothing because the flakes orientate themselves parallel to the substrate; its thermal insulating properties are also beneficial.

Mica is also important in the fire protection industry and is incorporated directly into cementitious fireboards. Mica's main virtues are its lack of shrinkage and its thermal resilience at high temperatures which minimize cracking and help partitions to maintain their structural integrity. For this application the mica should have a low magnesium oxide content, because high levels can cause shrinkage problems.

Plastics. Both muscovite and phlogopite can be incorporated as fillers in plastics (→ Plastics, Additives), phlogopite being chosen where color is unimportant. Dry ground, wet ground, and micronized mica can all be used depending on the type of plastic. Generally the principal properties mica confers on a plastic are:

1) Significant increase in tensile and flexural modulus and decrease in elongation
2) Reduction in thermal expansion and inhibition of nonuniform shrinkage likely to cause warping
3) Increased heat distortion temperature
4) Enhanced impermeability to fluids because of the planar orientation of mica platelets
5) Inertness to UV light, microwaves, and most chemical reagents
6) Improved dielectric properties

A further improvement has been the advent of silane and organosilane compounds for coating both wet and dry micas. Silane coating is thought to facilitate covalent bonding between the mica surface and the surrounding host polymer at elevated temperature and also improves dispersion of the mica within the polymer. Silane-treated wet ground mica is only marginally inferior to glass fibers as a polypropylene reinforcement additive [15].

One major area of utilization is in the automotive industry for "under the hood" applications of plastics such as headlight surrounds, dashboard components, molded rear seat backs, and car trims [16].

Oil-Well Drilling. In this application mica acts as a contingency commodity and its use in drilling is intermittent. It is used in drilling muds to inhibit water loss during drilling. The mica flakes block any cracks or fissures thereby reducing or eliminating loss of drilling fluid. It is used in a variety of grain sizes from coarse (5 – 12 mm) to fine (1.3 – 2 mm). In terms of product specifications this market places the

Table 5. Particle sizes of ground mica products ∗ [12]

Particle size,	Dry ground					Wet ground,	Micronized,
µm	Joint cement	Plastics	Oil-well		Welding rods	rubber grade	paint grade
6400			100				
2800			80	100	100		
1200			30	98	98		
500			10	50	60		
355		100		10	40		
250	100	99			10		
150	99	70				100	
105	95	50				99.9	
75	85	35				85	
53	75	20				80	100
44	70	5				75	99.9

∗ Values are expressed as percentage of total particles smaller than the specified size.

lowest demands on mica quality, only low-grade dry ground material is used.

Welding Rods. Dry ground mica is used as a component of rutile electrodes for welding mild and general steels. The mineral prevents cracking, serves as a flux control, and aids the coating of flux onto the electrode.

Rubber. Wet ground mica is also used as a mold release agent in the manufacture of rubber tyres.

Paints. Dry ground, wet ground, and micronized mica are used as fillers in paints. Micronized mica is widely used in emulsions with loading levels in the 10 – 20 % range. It improves the quality of the paint by promoting a matt finish, durability, and resistance to chemical attack and by acting as an anticracking agent. In addition mica helps to control paint flow and enhances scrub resistance and lightfastness. The plate-like mica particles align themselves parallel with the substrate. This not only improves adhesion of the paint to the substrate but also inhibits entry of water thereby improving durability. High aspect ratio wet ground mica is commonly favored for this application because the particles are flatter and, having been ground in an aqueous environment, are less likely to contain soluble salts that can reduce the shelf life of a paint [16]. Due to costs, micronized mica is, however, replacing wet ground mica in paints. Micronized mica is used almost entirely in paint formulations.

Nacreous Pigments. Nacreous (pearlescent) pigments rely upon light transmission, reflectance, and interference to reproduce a "mother of pearl" effect (→ Pigments, Inorganic). Mica is used as a substrate upon which a thin coating of titanium dioxide or other metal oxide is deposited. Both muscovite and phlogopite can be used. Mica is selected by virtue of its relative cheapness, platy morphology, and translucency. Mica particles used in pearl luster pigments are usually 3 – 150 µm in diameter with thicknesses of 200 – 500 nm. Particles below 20 µm produce a satin sheen, those in the 20 – 50 µm range give an optimal pearlescent effect, while those greater than 100 µm produce a sparkling finish. A major area of application is the automotive industry. Printing inks and undercoats for household paints are other markets.

Textured Coatings. Dry ground mica acts as a reinforcing filler in plaster for textured coatings where its main function is to control shrinkage and cracking, and maintain plasticity. The textured plaster coatings are used solely for decorative purposes.

6. Toxicology and Occupational Health

It has yet to be proved if mica dust is hazardous but several authorities have defined limits on permissible dust levels. The ACGIH has adopted threshold values for mica of 6 mg/m^3 for total

dust and 3 mg/m^3 for respirable dust. The Health and Safety Executive of Great Britain specifies a total airborne dust level for all mica grades of 10 mg/m^3 with an additional limit of 1 mg/m^3 for the respirable portion [14].

7. Economic Aspects

World mica output for 1987 totalled approximately 267 000 t of which more than half (147 000 t) was produced by the United States (Table 6). The equivalent figure for 1986 was 134 000 t, of which ground mica sales amounted to 116 000 t (96 000 t dry ground mica and 20 000 t wet ground material). Jointing compounds account for approximately 50 % of the U.S. ground mica market. India's output in 1987 amounted to 10 547 t, accounting for approximately 80 % of global sheet mica production.

Sheet mica has witnessed a gradual, probably terminal erosion of its traditional markets over the last three decades following the advent of solid state electronics. This is in sharp contrast to ground mica where demand for fireboard and jointing compounds is underpinning the market. The greatest potential for future growth in ground mica consumption lies in its expanding use in pearlescent pigments and as a filler in plastics.

8. Fluorinated Micas

In fluorinated micas (fluoromicas) hydroxyl ions in naturally occurring micas (see Fig. 3) are replaced in part or completely by fluorine. Substitution occurs with a minimum of strain as judged by the similarity of the Shannon and Prewitt atomic radii for fluorine (0.1285 nm) and oxygen (0.135 nm; interatomic distance (OH) = 0.097 nm). Fluorinated micas became of importance during World War II because of the limited access to natural micas. The mica of choice was *fluorophlogopite* [*12003-38-2*], $KMg_3AlSi_3O_{10}F_2$, which could be manufactured in sheets up to 7.5×12.5 cm^2 at atmospheric pressure by melting or solid state reaction of the raw materials (potassium fluorosilicate, feldspar, magnesia, alumina, and quartz sand). Care had to be taken that the fluorine levels were sufficiently high (ca. 12 mol %) to prevent the formation of fluoroamphiboles.

Fluorophlogopites have significantly different properties from their hydroxyl counterparts. They are harder, have a higher specific resistivity and dielectric constant, a smaller interlayer distance (0.9978 versus 1.0149 nm), and a higher maximum use temperature (1200 °C versus 400 – 500 °C for muscovite or 850 – 1000 °C for phlogopite). In spite of these improved properties, the industrial uses of fluorophlogopite as a dielectric or insulating material disappeared with the advent of microelectronics. The high use temperature of fluorophlogopite may make it desirable as a pillared clay molecular sieve – use of hydroxyl phlogopite as a cesium-selective ion sieve has been described [16]. Fluorophlogopites are also suited as a coating component for high-quality welding electrodes, which require hydroxyl-free minerals to minimize the concentration of diffusable hydrogen in the weldmetal.

The fluoromica field has been given new impetus with the invention of machinable and castable glass ceramics (→ Glass Ceramics, Chap. 4.3.) and synthetic micas. Two types of glass ceramics have been developed, one based on fluorophlogopite, $KMg_3AlSi_3O_{10}F_2$, with the trade name Macor [17], the other based on tetrasilicic mica, $KMg_2 \cdot 5\,Si_4O_{10}F_2$, with the trade name Dicor [18]. Compositions of these glass ceramics are given in Table 7.

Table 7. Commercial fluoromica glass-ceramic compositions (mol %) ∗

Oxide	Macor	Dicor
SiO_2	42.2	46 – 57
B_2O_3	6.6	
Al_2O_3	8.8	0 – 1
MgO	19.3	22 – 23
K_2O	5.4	8 – 9
F	17.8	13 – 21
ZrO_2		0 – 1
CeO_2		0.02

∗ Macor and Dicor are registered trade names, their composition is:

Macor $K_{1-x}Mg_3Al_{1-x}Si_{3+x}O_{10}F_2$

Dicor $K_{1-x}Mg_{2.5+x} \cdot 2\,Si_4O_{10}F_2$ $x < 0.2$

Macor glass ceramic is manufactured by pouring a fluorine-rich, phase-separated glass into rods, bars, or sheets, ceramming the cast material with a controlled heat cycle to cause for-

Table 6. World mica production (in tonnes)

Country	1985	1986	1987
United States [a, b]	125 000	134 000	147 000 [e]
former Soviet Union [e]	50 000	50 000	50 000
China [c, e]	16 800	19 700	18 000
Canada [e]	13 500	15 000	17 000
India	7 482	7 519	10 547
France [c]	10 084	10 834	9 300
Norway [d, e]	4 000	3 000	
Brazil [f]	931	1 700	
Morocco	1 440	1 500 [e]	–
Republic of South Africa	2 153	2 150	1 325
Zimbabwe	581	1 390	–
World total	239 000	254 000	267 000

[a] Sold or used by producers.
[b] Including mica recovered from mica schists and/or kaolin beneficiation.
[c] Conservative British Geological Survey estimates based on reconstructed exports.
[d] Crude only.
[e] Estimate.
[f] Exports.

mation of the fluorine-based mica crystals in the parent glass, and finishing the material to standard shapes and sizes; no firing is required after machining. Macor is a machinable glass ceramic that consists of ca. 55 vol % crystal and 45 vol % residual glass. The key to its machinability lies in its two-phase microstructure–randomly oriented mica crystals are formed in a glass matrix. Physical properties follow:

Density	2.52 g/cm^3
Porosity	0 %
Knoop hardness	250
Maximum use temperature	1000 °C
Linear thermal expansion coefficient	94×10^{-7} K^{-1}
Compressive strength	340 MPa
Flexural strength	100 MPa
Dielectric strength (a.c.)	1000 mV
Volume resistivity	$> 10^{14}$ Ω · cm

Properties are shown as a function of temperature in Figure 4. Macor has been employed in a wide range of applications such as vacuum-tight hermetic seals, headers and windows for microwave tube parts, retaining rings and hinge points for the space shuttle orbiter, and welding equipment nozzles.

Dicor glass ceramics are used in dental restoration. A potassium magnesium fluorosilicate melt is cast at 1400 °C by the lost wax technique. A centrifugal force generated by spinning pulls the melt tightly into the mold cavity. After removal of the cast, a glass crown is left which upon heat treatment turns into a glass ceramic. The modulus of rupture and compressive strength of this glass ceramic are three times those of conventional dental porcelain. The thermal conductivity and hardness are comparable to those of natural teeth.

Some forms of fluoromica sheet silicates may be processed into gels due to their propensity to hydrate, swell, and disintegrate into small particles in aqueous solutions. Hydration of the octahedral layer causes debonding of the tetrahedral sheets. Such gels can be processed into thin, continuous sheets of mica (e.g., for paper). Suspensions of these thin sheets can be formed into articles by casting, spraying, or doctor blading.

Processes for the manufacture of *inorganic paper* from water-swelling mica gels were developed in the 1960s by the U.S. Bureau of Mines and 3M [19], [20]. These generally rough, porous papers tend to have good strength and handleability. The manufacturing process involves fusion casting of a clear melt, followed by controlled cooling to create large crystalline books of mica. The mica is delaminated and screened, sedimented or decanted to remove the coarse fraction producing a gel from which paper is produced. As in the case of fluorophlogopite manufacture, lack of crystallite size control with concomitant additional processing such as dry milling makes the process expensive.

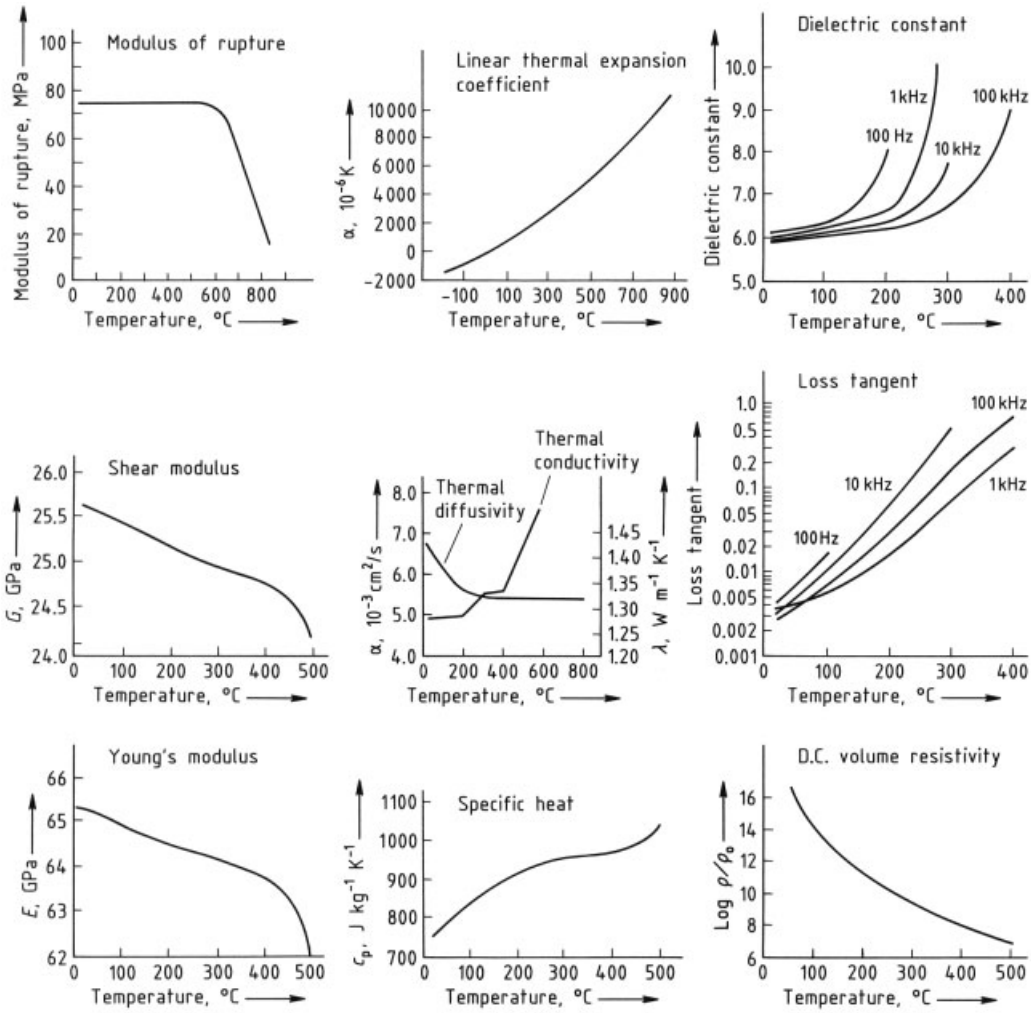

Figure 4. Physical properties of Macor as a function of temperature.

Lithium magnesium fluorosilicate (fluorohectorite) glass ceramics made by Corning Inc. alleviate a number of the tedious steps associated with manufacture of mica paper, and give a superior, fine-grained product from which smoother paper and film can be manufactured [21–23]. The lithium fluorohectorite crystallites in this glass ceramic consist of tetrahedral sheets of silica interspersed with layers of magnesium atoms octahedrally coordinated to oxygen. One of every four magnesium sites is occupied by lithium requiring an exchange of fluorine for oxygen to maintain charge balance.

The benefit of using a glass ceramic lies in the ability to control nucleation and crystal growth resulting in a uniformly fine-grained mica ceramic that disintegrates in water to give 10 µm particles. The critical issue in using glass ceramics lies in minimizing the amount of glassy, nonmicaceous phases, while promoting mica crystal growth and maintaining good dispersing properties in water. The prime benefit of using a glass ceramic is that it bypasses mechanical grinding, enabling the expedient processing of a stable gel of dissociated mica laminae. Electron micrographs indicate interlacing crystallites less than 10 nm thick and about 10 µm in average diameter.

The solid to liquid ratio of a fluorohectorite mica gelled suspension, its viscosity, and pH

may be controlled and allow a variety of forming techniques. Conventional fourdrinier process papermaking or film extrusion equipment can be used to form large continuous sheet ceramic paper. The gels can also be extruded as fiber. By addition of controlled concentrations of dissolved electrolytes the gels can also be flocculated, hence enabling the manufacture of foamed, spray dried, or cast thick-walled articles. Potential uses for fluorohectorite mica ceramics include dielectric and microwave transparent papers, boards, and cloths; floor tiles; and cheap, lightweight, disposable dinnerware [24], [25].

9. References

1. S. W. Bailey: *Micas, Reviews in Mineralogy,* **vol. 13,** Min. Soc. America, 1984, p. 584.
2. W. A. Deer, R. A. Howie, J. Zussman: *An Introduction to the Rock-Forming Minerals,* Longman, London 1966, p. 528.
3. P. W. Harben, R. L. Bates: *Industrial Minerals – Geology and World Deposits,* Industrial Minerals Div., Metal Bulletin plc, London 1989.
4. R. J. Beckman: "Mica Resources of the Western United States," Proc. 17th Annual Forum on the Geology of Industrial Minerals, *New Mexico Bureau of Mines and Min. Res. Circ.* **182** (1982) 35 – 37.
5. G. C. Hawley: "Expanding Markets and Technologies from Suzor Township, Quebec," *AIME Annual Meeting,* **preprint no. 81 – 121,** Chicago 1981, p. 10.
6. W. T. Pecora et al.: "Mica Deposits in Minas Gerais, Brazil," *U.S. Geol. Surv. Bull.* **964-C** (1950) 205 – 305.
7. M. L. Rajgarhia: *The Story of Mica,* Mica Manufacturing Co., Calcutta 1986, p. 99.
8. M. V. N. Murthy: "Mica Fields of India," *Proc. 22nd Int. Geol. Congress,* India 1964, p. 17.
9. G. Clarke: "Mica – A Review of World Developments," *Ind. Miner. (London)* **189** (1983) June, 27 – 50.
10. J. Griffiths: "Kemira Mica – Utilising Resources," *Ind. Min. (London)* **219** (1985) Dec., 77 – 83.
11. Indian Standard Institution Methods for Grading and Classification of Muscovite Mica Blocks, Thins and Films (IS: 1175), 1957.
12. ASTM, Standard Specification for Natural Muscovite Block Mica and Thins Based on Visual Quality, D 351-371 (1977) in *1984 Annual Book of ASTM Standards: Electrical Insulation – Specifications: Solids, Liquids, and Gases; Test Methods: Liquids and Gases, Electrical Protective Equipment for Workers,* Philadelphia, 1984, pp. 136 – 143.
13. J. Robbins: "Sheet Mica – and its Changing Face," *Ind. Min. (London)* **209** (1985) Feb., 33 – 47.
14. G. Chapman: "The World Mica-Grinding Industry and its Markets," *Proc. 4th Industrial Minerals International Congress,* Toronto 1984, pp. 197 – 203.
15. J. Rieger: "Silane Treated Wet Ground Mica as a Reinforcing Filler for Polypropylene," *Plastics 85, Proceedings of the SPE 43rd Annual Technical Conference and Exhibition,* Washington 1985, pp. 960 – 962.
16. J. Benbow: "Mica: Markets Built on Dry Ground," *Ind. Min. (London)* **245** (1988) Feb., 19 – 31. S. Komarneni, R. Roy, *Science* **239** (1988) 1286 – 1288.
17. G. H. Beall, DE-OS 2 133 652, 1978.
18. D. G. Grossman, US 3 839 055, 1974; US 3 839 056, 1974.
19. R. A. Hatch, US 3 001 571, 1961.
20. H. R. Shell, K. H. Ivey: "Fluorine Micas," *Bull. US Bur.* of Mines 647 (1969) 291.
21. G. H. Beall, D. G. Grossman, S. N. Hoda, K. R. Kubinski, US 4 239 519, 1980.
22. G. H. Beall, D. G. Grossman, S. N. Hoda, K. R. Kubinski, US 4 297 139, 1981.
23. F. P. Fehlner, W. J. Wein, US 4 432 852, 1984.
24. S. N. Hoda, A. R. Olszewski, US 4 559 264, 1985.
25. S. H. Wu, US 4 675 235, 1987.

Microbial Transformation → Biotechnology

Biocides

KEN R. PAYNE, Chesterfield, United Kingdom (Chaps. 1 – 2)

EDWARD HILL, ECHA Microbiology Limited, Cardiff, United Kingdom (Chaps. 1, 3)

1. Introduction 1
2. Chemical Biocides 2
2.1. Phenolics 2
2.2. Chlorine and Chlorine-Releasing Compounds . . . 3
2.3. Iodine and Iodophores 4
2.4. Quaternary Ammonium Compounds 4
2.5. Biguanides 5
2.6. Peroxygens 5
2.7. Alcohols 5
2.8. Aldehydes 6
2.9. Hydroxybenzoate Esters 6
2.10. Heavy Metals 6
3. Application of Biocides 6
3.1. Factors Affecting Biocide Selection . 7
3.2. Challenge Testing 7
3.3. Spoilage 9
3.4. Monitoring 9
3.5. Predetermined and Automatic Biocide Additions 11
4. References 11

1. Introduction

The term microbiocide (or more simply biocide) is used to describe all agents that kill microbial life and thus includes antibiotics (→ Antibiotics), antimicrobial agents used in medicine (→ Chemotherapeutics), disinfectants (→ Disinfectants), and agricultural fungicides (→ Fungicides, Agricultural). This article will, however, deal mainly with the compounds used to prevent microbial growth in industrial products, particularly in the food industry, in cosmetics, pharmaceuticals and toiletries, in paints and plastics, in water cooling systems, and in paper manufacture. The use of microbiocides in wood preservation is discussed elsewhere (→ Wood Preservation).

History. Chemicals have been used since the very earliest times in the prevention of disease and the preservation of food. Notable examples were the use of burning sulfur to control the spread of plague and the widespread use of spicing, salting, and smoking to prevent the deterioration of food. The use of chemicals in earlier times was quite empirical since nothing was known about the mechanisms by which they effected the preservation of food or the control of disease. However, these empirical observations were supplemented by a more rigorous scientific investigation following the discovery of the microscope by VAN LEEUWENOEK in the 18th century.

VAN LEEUWENOEK demonstrated the presence of major classes of bacteria, protozoa, algae, and yeasts. The causal relationship between microorganisms and disease was established by the research work of BASSI and PASTEUR on the infection of silk worms by fungi and protozoa, respectively. Major contributions were subsequently made by KOCH and LISTER in the isolation and identification of pure bacterial strains as pathogens responsible for disease. Recognition of the fact that bacteria are responsible for disease and food spoilage created much interest in the development of chemicals for the control of these bacteria. Methods developed by KOCH and others were used in evaluating the effectiveness of these chemicals in destroying bacteria and curing disease. In the latter half of the 19th century, LISTER demonstrated that the use of phenol markedly reduced the incidence of serious infection during surgery. LISTER's work demonstrated the potential of chemical disinfectants in hospitals; subsequently, the use of less toxic and more efficient chemicals was extended to other areas of hospital practice.

The preventative role of chemical biocides was also becoming recognized in the industrial field. The use of biocides in metal working fluids, for example, is an important factor in maintaining product quality and the health of the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_563

work force. In oil recovery, biocides are used to control sulfate-reducing bacteria and to improve yields. In water cooling systems and paper production, added biocides prevent the formation of bacterial or fungal slimes thus avoiding the necessity of costly shut-downs. Generally, microorganisms proliferate in the presence of an aqueous medium and biocides are therefore used in aqueous emulsion paints, polymer emulsions, pharmaceutical products, and toiletries. The microorganism *Cladosporium resinae* can occur in the traces of water found in aircraft fuel tanks and can cause corrosion if not controlled.

Important industrial applications exist for biocides in addition to those normally associated with human and animal health (e.g., hospitals, agriculture, public health and the home; → Disinfectants). Commercially used biocides consist of an active chemical formulated with a solvent or surfactant for direct application. A wide range of chemicals possess biocidal activity and include such diverse groups as phenolics, quaternary ammonium compounds, halogen compounds, aldehydes, alcohols, organometallic compounds, peroxygens, biguanides, and heterocycles.

2. Chemical Biocides

2.1. Phenolics

Phenol [*108-95-21*] itself has little direct use as a disinfectant, but many phenol derivatives, particularly the alkylated and chlorinated derivatives are still widely used.

The high-boiling *tar acids* consist of complex mixtures of alkylated phenols obtained from coal tar by distillation; they are used in the formulation of the so-called black and white fluids. In *black fluids* the tar acids are formulated with castor oil soap and either a neutral or aromatic oil or a cresylic creosote as a solvent or an extender. In *white fluids,* the tar acids are emulsified with either casein or a synthetic detergent. Both fluids are used in agriculture and also as crude domestic disinfectants. A mixture of isopropylphenols produced by the direct alkylation of phenol and marketed by Ciba-Geigy as Alkylated Phenols F 12 is used as a substitute for the tar acids and is marginally more effective.

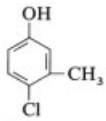

4-Chloro-*m*-cresol

The bactericidal efficiency of phenol is enhanced by introduction of chlorine atoms and substitution with alkyl or aryl groups (especially if they are ortho to the phenolic group). 4-Chloro-*m*-cresol [*59-50-7*] (PCMC) is used as an antiseptic and disinfectant in the domestic field and in hospitals (as a topical antiseptic). It is produced by direct chlorination of *m*-cresol.

Two powerful disinfectants are derived from 3,5-xylenol by direct chlorination: 4-chloro-3,5-xylenol [*88-04-0*] (PCMX) and 4,6-dichloro-3,5-xylenol [*133-53-9*] (DCMX). The former is used as a topical and urinary antiseptic and is the active disinfectant in Dettol (Reckitt and Colman, UK). The latter, as well as being bacteriostatic in soaps, is also used in domestic pine-type disinfectants.

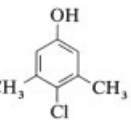

PCMX

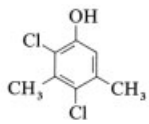

DCMX

2-Benzyl-4-chlorophenol [*120-32-1*] (chlorophene) is produced by reacting benzyl chloride with sodium or potassium phenolate. It is a powerful germicide and an excellent fungicide.

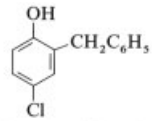

2-Benzyl-4-chlorophenol

2,2′-Methylenebis(4-chlorophenol) [*97-23-4*] (dichlorophen) is synthesized by the acid-catalyzed condensation of formaldehyde and 4-chlorophenol. It is an effective antimicrobial which is used as a germicide in toiletries. It is used in cooling water systems, in the food and oil industries, and in textiles and paper manufacture (trade name: Panacide, BDH, UK). Dichlorophen has a low mammalian toxicity (LD_{50}, rat, oral 2.69 g/kg) and skin irritancy.

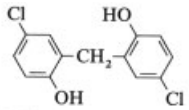

2,2′-Methylenebis(4-chlorophenol)

2-Phenylphenol [*90-43-7*] (trade name: Dowicide A, Dow, USA) is a fungicide which may either be synthesized directly or isolated as a byproduct in the production of phenol by the hydrolysis of monochlorobenzene. Direct chlorination of 2-phenylphenol gives a mixture of 4-chloro-2-phenylphenol [*1331-46-0*] and 6-chloro-2-phenylphenol, the former being a very effective biocide which is used as the active ingredient of pine type disinfectant fluids.

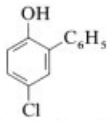

4-Chloro-2-phenylphenol

2.2. Chlorine and Chlorine-Releasing Compounds

Chlorine [*7782-50-5*] is widely used in water treatment, particularly in the treatment of crude water before it is supplied for either potable or industrial use. Many cooling circuits in steam electricity generating stations, in food processing plants, and in oil refineries are treated with chlorine to prevent the formation of microbial slimes and to maintain hygiene standards. The bactericidal effect is due to the reaction between chlorine and water:

$$Cl_2 + H_2O \rightarrow HOCl + HCl$$

The non-ionized hypochlorous acid is believed to be responsible for the bactericidal activity of chlorine, hypochlorites, and chlorine-releasing agents. Since hypochlorous acid ionizes to give H^+ and OCl^- ions, the concentration of hypochlorous acid, and hence the bactericidal efficiency will depend on the pH of the solution. The equilibrium diagram for chlorine – water – hypochlorite mixtures (Fig. 1) shows the variation of hypochlorous acid concentration with pH. Hypochlorous acid is believed to react with and inactivate enzymes that are essential for bacterial metabolism, particularly triose phosphate dehydrogenase.

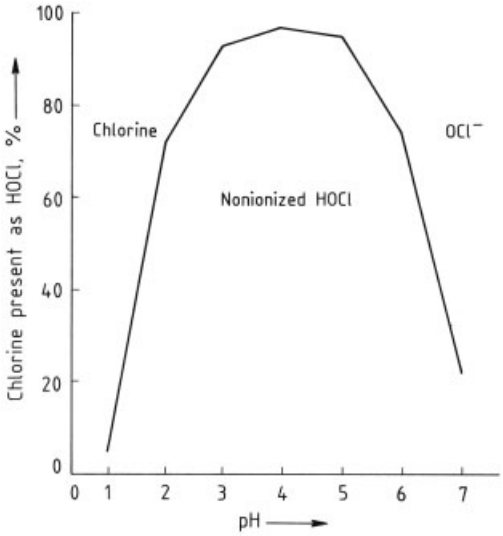

Figure 1. Equilibrium diagram of chlorine – water – hypochlorite mixtures.

Hypochlorites are principally used as the sodium salt [*7681-52-9*] and are cheap, effective biocides that are used widely throughout the EEC in hospitals, agriculture, dairying, food processing, and public health and hygiene. Sodium hypochlorite decomposes with increasing temperature and under the influence of light and certain metallic salts. For this reason the more stable calcium hypochlorite [*7778-54-3*] is used in certain applications, for example in swimming pool disinfection.

The chlorine-releasing compounds are essentially organic chlorine derivatives which are stable when dry, but generate hypochlorous acid in the presence of water. Hypochlorous acid is again the active biocide with the organic substituent controlling its release. Chloramine T [*127-65-1*] (the sodium salt of *N*-chloro-4-methylbenzenesulfonamide) and Dichloramine T [*473-34-7*] *N,N*-dichloro-4-methylbenzenesulfonamide are chlorine-releasing compounds that have been on the market for many years and were used as bactericides during World War I.

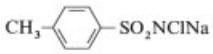

Chloramine T

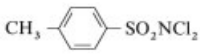

Dichloramine T

Trichloroisocyanuric acid [*87-90-1*] is an important chlorine-releasing agent that is produced by the chlorination of cyanuric acid in caustic soda solution. It finds an important use as a disinfectant/sanitizer in the food industry. A major manufacturer is Chlor-Chem of Hauxton, UK.

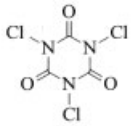

Trichloroisocyanuric acid

1,3-Dichloro-5,5-dimethylhydantoin [*118-52-5*] is sparingly soluble in water with which it reacts to give controlled release of hypochlorous acid. 1-Bromo-3-chloro-5,5-dimethylhydantoin [*126-06-7*] reacts with water to form both hypochlorous and hypobromous acids. The latter has a wider pH range of bactericidal activity than hypochlorous acid. When hypobromous acid comes into contact with nitrogenous waste, bromamines are formed which, unlike the chlorine analogues, retain a significant biocidal activity.

2.3. Iodine and Iodophores

Iodine [*7553-56-2*] has a wide range of antimicrobial activity and is effective against gram-positive and gram-negative bacteria, mycobacteria, fungi, and viruses. It also kills bacterial spores after an extended contact time. The active species (the iodine molecule) is only sparingly soluble in water; iodine is therefore used as a solution of iodine in alcoholic potassium iodide.

Iodine stains the skin and may lead to sensitization of the skin and mucous membranes. Iodophores (iodine carriers) were developed in the 1950s in order to overcome these disadvantages. Iodophores are essentially complexes of iodine with various organic compounds which release iodine on demand. The most important complexes used are those with poly(propylene oxide) or poly(ethylene oxide), cationic surfactants (quaternary ammonium salts), or polyvinylpyrrolidone. The iodophores are used in hospitals primarily for skin disinfection, and also in dairy farming where they are important as udder washes in the prevention of bovine mastitis as, for example, the Iosan range of iodophores (Ciba-Geigy, Switzerland) or Iodel FD (Diversey, UK).

$RNH(CH_2)_2NH \cdot (CH_2)_2NH \cdot CH_2 \cdot COOH \cdot HCl$

2.4. Quaternary Ammonium Compounds

Quaternary ammonium compounds are most effective against gram-positive bacteria at concentrations as low as 1 : 200 000. They are less active against gram-negative organisms: concentrations as high as 1 : 20 000 are required to control *Pseudomonas aeruginosa.* Further limitations of quaternary ammonium compounds are their inactivity at and below pH 3.5 and their reduced effectiveness under dirty conditions. They are nontoxic and are used quite widely in hospitals, in the food industry, and as household disinfectants. Commercially important quaternary ammonium disinfectants are benzalkonium chloride [*8001-54-5*], benzethonium chloride [*121-54-0*], and cetylpyridinium chloride [*123-03-5*] (see bottom of the page).

Amphoteric quaternary ammonium disinfectants, usually alkylated aminocarboxylic acid

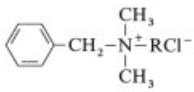

Benzalkonium chloride

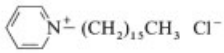

Cetylpyridinium chloride

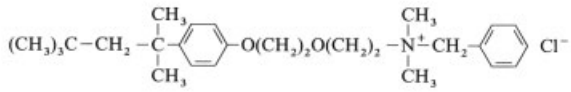

Benzethonium chloride

derivatives, are also used commercially. The amphoterics have greater caustic stability than the normal quaternaries and are more easily removed from surfaces. This is a great advantage in the food industry where amphoterics are widely used. The formula of Tego 103 S (Goldschmidt, FRG) is given below:

2.5. Biguanides

Although the biguanides had been recognized as biologically active by IG Farben in the 1930s, their potential as active bactericides was first discovered by ICI in 1946 [1], [2] in the form of the biguanide polymer, poly(hexamethylene biguanide)hydrochloride [*32289-58-0*] (PHMB). Subsequently, 1,6-bis(4′-chlorophenyl biguanide) hexane [*55-56-1*] (chlorhexidine) was also found to be a potent biocide. Chlorhexidine and PHMB, are the major biguanide compounds which are used today as disinfectants and sanitizers [3].

Chlorhexidine is bactericidal at concentrations of 100 µg/mL and bacteriostatic at concentrations of 1 µg/mL. It is often formulated with 2-propanol and quaternary ammonium compounds. It is absorbed by the skin and is widely used as a topical disinfectant in hospitals. It is sold under the trade name of Hibitane (ICI, UK) (see below).

PHMB is an industrially important biocide for the control of most gram-positive and gram-negative bacteria, yeasts, fungi, and aquatic algae. It is used in oil recovery, food processing, sugar extraction, and as a disinfectant in swimming pools.

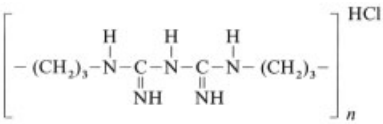

PHMB

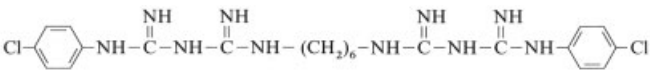

Chlorhexidine

2.6. Peroxygens

Hydrogen peroxide and peracids are powerful antimicrobial agents and effective sporicides. A 35 wt % solution of hydrogen peroxide [*7722-84-1*] can be stored for prolonged periods. It is easy to handle, noncorrosive, and mixes readily with water. Hydrogen peroxide is used in hospitals as a topical disinfectant and in swimming pools in conjunction with the ICI product Baquacil (PHMB). It is also important in the control of algae in agriculture and in industrial cooling systems [4].

Peracetic acid [*79-21-0*] and the peracids are nonfoaming, water-soluble liquids which are converted into environmentally acceptable decomposition products, namely, dilute aqueous solutions of the corresponding organic acids. Their bactericidal activity is not affected by hard water and only slightly reduced by the presence of organic matter. An important use for peracetic acid is in the brewing industry.

Trade Name. Proxitane (Interox, UK) [5].

2.7. Alcohols

The aliphatic alcohols, principally ethanol [*64-17-5*], 1-propanol [*71-23-8*], and 2-propanol [*67-63-01*] have long been used as disinfectants in hospitals and medicine. They are active against vegetative bacteria but are not sporicidal. The fact that they are also cleansing agents and are volatile makes them particularly useful in topical applications.

Ethanol is bactericidal at concentrations of 60 – 95 vol % in aqueous solution, but is commonly used at a concentration of 70 % for skin and instrument disinfection and for surface cleaning. It acts synergistically with such bactericides as iodine and chlorhexidine which are frequently used in aqueous ethanol solution. An aqueous 70 % ethanol solution is a powerful virucide and mycobactericide.

2-Propanol is a slightly more effective bactericide than ethanol, but is more toxic. It is also used in preoperative skin treatment.

Other alcohols used as disinfectants and preservatives are phenylethanol [*60-12-8*] and 2-phenoxyethanol [*122-99-6*]. 2-Bromo-2-nitropropane-1,3-diol [*52-51-7*] (Bronopol: Boots, UK) is used as a disinfectant and preservative in pharmaceuticals, toiletries, cosmetics, and household products.

2.8. Aldehydes

A number of aldehydes are antimicrobials but only formaldehyde and glutaraldehyde are widely used.

Formaldehyde [*50-00-0*] is used in the vapor-phase decontamination of infected buildings, particularly in agriculture. It is also used in hospitals as a 37 % aqueous solution with added methanol as a polymerization inhibitor. Formaldehyde is an irritant which exerts a powerful effect upon the mucous membranes, and there has been recent concern about its possible carcinogenicity.

Glutaraldehyde [*111-30-8*], $CHO(CH_2)_3CHO$, is a broad-spectrum bactericide which is also sporicidal. It is most effective at pH 8, but tends to polymerize at this pH. For this reason glutaraldehyde is supplied in acid solution and subsequently buffered to pH 8. It is used in hospitals, particularly in the sterilization of surgical equipment such as endoscopes which cannot be heat sterilized. It is also used as a biocide in the oil industry.

2.9. Hydroxybenzoate Esters

A number of 4-hydroxybenzoate esters are used as preservatives in pharmaceutical emulsions, creams, and lotions, particularly the methyl ester [*99-76-3*] which is more water soluble (0.25 %) than the higher esters.

Trade Names. Nipagin M (Nipa Laboratories, UK), Tegosept M (Thos. Goldschmidt, FRG).

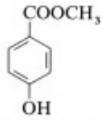

Methyl Paraben

2.10. Heavy Metals

Copper and silver vessels were used in ancient times to prevent the spread of disease through drinking water. Today the best known heavy-metal derivatives used as antiseptics are the organomercury compounds although doubts are raised from time to time concerning their toxicity. Thimerosal [*54-64-8*] is produced by reacting ethyl mercuric chloride with thiosalicylic acid. It is used as a preservative in contact lens solutions.

Trade Name. Merthiolate Sodium (Eli Lilly, USA).

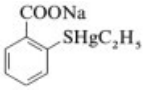

Thimerosal

3. Application of Biocides

A wide range of antimicrobial agents are available as industrial preservatives and biocides (Chap. 2). Their characteristics are sufficiently dissimilar to satisfy the great variety of market niches. Fitting the appropriate biocide into the specific niche should follow a logical selection process. When deficiencies exist, activity can usually be enhanced by appropriate formulation or by using several active agents.

The successful application of selected biocides/preservatives frequently falls short of expectations. Good selection of both agent and concentration is often nullified by poor on-site control. Monitoring strategies are available and deserve wider use. It may well become as important for health and safety reasons to keep in-use concentrations below a specified level as it is to keep them at or above target concentrations.

3.1. Factors Affecting Biocide Selection

The term (micro)biocide inherently implies killing life, preferably microbial life, but it is widely used to include preservatives. Preservatives are merely required to prevent spoilage development and are not expected to sanitize systems that are already infected. Preservatives can also be regarded as biocides – Pharmacopoeia protocols for preservative challenge testing usually demand not only "no increase" but a reduction in viable (culturable) numbers of microorganisms. Speed of action is not essential for preservatives but long-term stability and persistence is (see also → Foods, 3. Food Additives, Chap. 3.1.). Conversely in many practical applications, a fast "disinfectant" action is of prime importance and long-term persistence is of little consequence. The first consideration in the selection of a biocide should thus be its suitability for preservation and/or disinfection. This information should be available in the suppliers' technical data sheets. However the frequently cited minimum inhibitory concentrations are evidence of "no increase", not "kill" and certainly not the rate of kill.

Further selection is an "armchair" exercise using data sheets and published literature to draw up a short list of candidate biocides. This exercise should consider:

1) The biological spectrum of biocide activity – what are the likely short-term and long-term contaminating organisms? Is the biocide effective against them? Organisms occurring as a "biofilm" are much more resistant to biocides than organisms in suspension. Physiologically active organisms are generally more sensitive than "resting" organisms. Killed organisms may still be detrimental if the biocide does not inactivate their enzymes.
2) The influence of product pH on biocide solubility, stability, and activity.
3) The presence of two phases: the biocide will partition between the two according to its relative solubility. Biocide in an "oil" phase is not immediately available to exert its antimicrobial effect in an associated water phase.
4) Environmental impact and toxicology. There must be an acceptable method for disposing of the biocide. All biocides are toxic to humans and thus invoke a risk assessment for their application under the 1988 COSHH regulations [6].
5) Presence of known biocide inactivators or potentiators in the product to be protected.
6) Undesirable effect on the product, e.g., color, smell, pH shift, chemical reactivity, viscosity, flash point.
7) The influence of biocide concentration on activity (concentration exponent). Activity is not usually directly proportional to concentration. For example, reducing phenol concentration by a half would reduce its rate of action 64 fold (the concentration exponent of phenol is 6).
8) The influence of temperature on biocide activity and stability. Increase in activity is usually regulated by the formula

   $$t_1 = t_2 \theta^{(T_2 - T_1)}$$

   where t_1 and t_2 are times to kill, T_2 and T_1 are temperatures (in °C), and θ is a temperature coefficient. For example if θ is 3 (phenol), a 10 °C rise in temperature increases the kill rate 59 000 fold.
9) Ability to monitor active biocide concentration and hence determine "shelf life" or in-use "contact time".
10) Physical properties (powder, liquid, granules, volatility, etc.) which affect handling and measurement.

There is no simple, single, universally applicable biocide. No biocide is likely to satisfy the required specifications and the final choice will be a compromise. Deficiencies can to some extent be compensated by stabilizers, odor masks, dispersants, and combination with other biocides. Technical assessment will then be balanced against cost effectiveness and the final choice may be deferred until laboratory or field trials have been conducted.

Legal regulations vary nationally depending on the nature of the biocide, its formulation, and field of application. Much useful information on the practical application of biocides can be found in the Report of the Expert Advisory Committee on Biocides [7].

3.2. Challenge Testing

Following the armchair selection process, one or more biocides are listed as potentially cost-

effective candidates. Confirmation or final selection probably proceeds based on challenge testing. A great variety of test methods are available from national organizations (e.g., BS, ASTM, Féderation Industrielle Pharmaceutique, Pharmacopoeias). They may not provide the exact required protocol and are frequently modified.

Test methods are divided into two groups according to whether disinfection or preservation is the objective.

Disinfection Challenge Testing. Conventional disinfectant tests for public health microbiology are rarely adapted for industrial purposes but similar principles are involved:

1) A substantial challenge is used.
2) An effect is required in a short contact time.
3) Antimicrobial activity is terminated by the addition of a neutralizer.
4) Action takes place in an unfavorable environment.
5) A quantitative "kill" is determined.
6) The "challenge" is repeated.

Repetition of the challenge is not always necessary for industrial applications. The contact time for public health applications is usually a few minutes whereas several hours are normal in industry. The challenge usually consists of one or more well-standardized pure microbial cultures of public health interest; a mixed inoculum (possibly a previously spoiled sample) is more relevant in industry.

Public Health disinfectant tests are suspension tests. A suspension of test organisms is mixed with the test solution and the number of surviving organisms is determined after certain exposure times. This test forms the basis for the classic disinfectant assays such as the phenol coefficient test [8], the Rideal – Walker test [9], Chick – Martin test [10], the Kelsey – Sykes test [11], the AFNOR test [12], and the 5-5-5 test [13]. In industry it is often as important to kill biofilms, particularly if sulfate-reducing bacteria are to be eradicated. The protocol chosen to evaluate a biocide may include a sequence which confirms efficacy against microbes that are entrained in a biofilm.

Preservation Challenge Testing. Simple preservation tests merely establish "no increase" in the number of microorganisms. More elaborate tests appear in the Pharmacopoeias for Group 2 pharmaceuticals and demand a mixture of "no increase" results and specified logarithmic reductions of microbial numbers. The principles of the best protocols are:

1) A small challenge is used.
2) Surviving organisms are determined several times over a long contact time (days, weeks).
3) A neutralizer is used before testing for survivors.
4) Known species from a wide spectrum of microorganisms are inoculated separately.

For industrial purposes repetition of the challenge is often desirable to simulate potential contamination when a container is opened, partly used, and resealed. The inoculation range may also be extended by including "wild" strains of known significance or mixtures of organisms.

Simulation Tests. Simulation tests mimic real situations on a laboratory scale. The paint industry has possibly progressed furthest in simulation tests to evaluate paint film "fungicides". The deactivating mechanisms of water leaching and UV degradation have been intensively studied, the principle colonizing organisms are known, and thus reasonable simulation regimes can be standardized. For other biocide applications a myriad of simulation devices and procedures have been described. Those for metal working fluids have been the subject of much ingenuity because metal working fluid systems are complex and exhibit marked oxygen, temperature, and Eh (redox potential) gradients; procedures are reviewed in [14]. Biofilms develop slowly; various strategies have been adopted for growing them so that antimicrobial agents can be properly evaluated as regards their penetration of preformed films and their inhibition of film formation. Some procedures are simple and inexpensive [15] whilst others attempt to produce 'standard' biofilm replicates to improve comparison of results [16].

Legionella pneumophila is often a specific target organism in cooling tower water and simulators for this application have been described [17].

Whatever simulation regime is adopted, pass and fail criteria have to be imposed. These may relate to the expected shelf life or in-use life of

the product to be protected and to the anticipated degree of contamination. Although these tend to be arbitrary, the experienced tester is able to rank competitive biocides with a great deal of confidence.

3.3. Spoilage

Spoilage occurs in a wide range of physical and chemical conditions. For a detailed discussion of biodeterioration, see [18–25].

In approximate descending order of importance with regard to volume of biocide sales (not economic consequences) the industries involved are principally: surface coatings, paper, timber, textiles, adhesives and sealants, oil-field, water treatment, metal working, cosmetics and toiletries, fuel, and leather.

Within each industry, a multitude of applications exist for biocides. In some applications the antimicrobial agent (e.g., formaldehyde) is a one-shot treatment and is incorporated into the product during manufacture. When the agent fails, the product fails. In a large number of applications antimicrobial activity can be "topped up" by a tank-side addition when a product is in use. The agent used is frequently different from that originally incorporated in the product and should correct developing spoilage, fouling, or microbial corrosion. It should persist for some time, exerting preservative properties. Although "top-up" can be carried out at arbitrary intervals, this involves the risk of underdosing and loss of control, or overdosing and potential risks for health and environment. Regimes for semicontinuous addition of biocide utilize some form of monitoring.

3.4. Monitoring

Three principal strategies are used for monitoring the effectiveness of biocides: (1) subjective and functional criteria, (2) determination of microbial population, and (3) determination of biocide or preservative concentration.

Subjective and Functional Criteria. In this strategy the effectiveness of the biocide is assessed by means of the effect it produces: prevention of discoloration, visible slime formation, filter plugging, change in pH, corrosion, and system malfunctions. Although crude, this method is still widely practiced.

The procedure necessitates the addition of high biocide concentrations and results in the production of a large amount of dead microbial material and endotoxins. A partially spoiled system cannot be improved, further deterioration is merely delayed.

Monitoring Microbial Populations. In this method the number of microorganisms present in the system is determined; it attempts to arrest a developing spoilage situation instead of preventing it. Biocide depletion or failure is assumed if microbial regrowth is demonstrated.

The reappearance of a potential spoilage flora may, however, also be due to the arrival or generation of microbes resistant to that biocide; a new biocide must then be used. If the microbiological test procedure is based on culturing to yield visible growth, there is a time lag of 1 – 3 days before a result is available; during this time a situation may continue to deteriorate. A false low or negative microbiological result may be obtained due to carry-over of active residual biocide onto the culture medium. To avoid delays, on-site microbiological tests are widely used for semiquantitative results. Traditional laboratory techniques are justified if accurate numbers and identities are needed.

Limit values must be set which trigger an appropriate response. For example a total bacterial count of 10^5/mL and/or the detection of sulfate-reducing bacteria could invoke an addition of the original biocide. The detection of molds and/or yeasts in a system prone to bacterial spoilage could initiate the addition of a different, antifungal agent.

Prior to the introduction of the "Red Spot" test in 1967, there were virtually no practical onsite procedures for quantifying bacterial numbers [26]. This test utilizes the ability of microorganisms to reduce a colorless tetrazolium salt to red formazan. In 1971 the Methylene Blue Reduction Test for quantifying milk microorganisms was applied to metal working fluids [27]. This test exploits the ability of some microorganisms to reduce the blue dye to a colorless base. Din slides, developed in 1975, can distinguish between bacteria, yeasts, and molds semiquan-

titatively and are simple to use on-site [28], [29]. Individual invisible microbes reproduce on the agar surface of the slide to yield visible, countable colonies.

Interpretation of the results is sometimes suspect if the calibration data supplied for aqueous fluids are applied to fluids of differing viscosity and "wettability" such as diesel fuels and lubricants.

A new concept for on-site use, the Sig Test, was described in 1984 [30]. The most useful industrial versions rapidly and semiquantitatively detect nitrite-utilizing and sulfide-generating bacteria. A detecting gel in a tube is flooded with the test sample and undergoes characteristic color and consistency changes.

Rapid instrumental methods such as measuring ATP or impedance are popular in the food and water industries. The impedance test relies on the electrical changes that occur when microbes grow in a culture medium [31]. Problems occur, however; for example, residual biocides may quench the enzyme reactions in ATP-measuring devices; industrial samples, particularly emulsions, may mask impedance changes due to microorganisms.

Monitoring Biocide or Preservative Concentration. The concentration of biocide is not linearly related to its activity (see Section 3.1). A small reduction in the target concentration of a biocide may have a considerable impact on the ability to resist a microbial challenge. In many cases the correct initial concentration is not even achieved due to a lack of knowledge of the system size, of the active concentration of the agent, or of its partition between oil and water phases.

The advantages of monitoring biocide concentration are that biocide concentration can be frequently adjusted and spoilage is prevented not "corrected". If coupled to microbiological tests a realistic "target" biocide concentration can be established and maintained. Biocides can be measured in a number of ways.

Simple colorimetric assays for formaldehyde released from donor biocides have been available for some years (e.g., E. Merck, Darmstadt, FRG) but are not widely used although they are suitable for on-site use. Many simple colorimetric test kits are also available for measuring free and/or combined chlorine but they do not necessarily reflect activity.

Chemical and physical assays for other biocides usually necessitate inital extraction and/or concentration followed by analysis using procedures such as HPLC. The determination may not correlate with known antimicrobial activity, particularly if more than one active agent is present or if biocide activity is potentiated by a formulated component. The techniques are not suitable for on-site use.

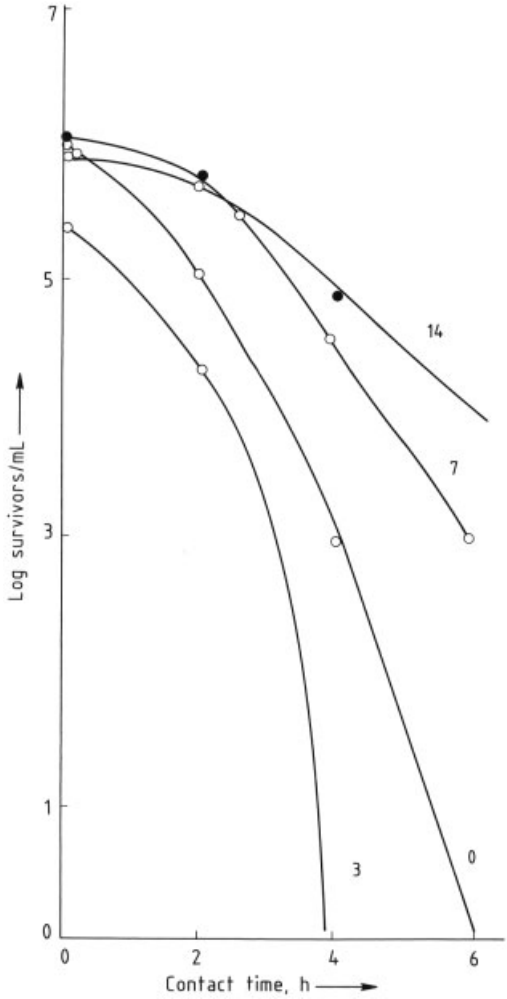

Figure 2. Survival curve of *Pseudomonas aeruginosa* for assaying biocides in metal-working fluids
A preserved 3 % cutting oil emulsion was held at 30 °C and 20 mL aliquots removed after 0, 3, 7, and 14 days; 0.1 mL of a suspension of an overnight culture of *P. aeruginosa* was added to the aliquot on removal and the numbers of survivors were determined after 0, 2, 4, and 6 h

A variety of biological techniques are available for assaying antimicrobial activity, they all

involve the inhibition of growth of a sensitive microorganism.

In 1982 a biological method of assaying biocides in metal-working fluids was proposed, which involved challenging aliquots of the preserved metal-working fluid at various storage or in-use time intervals with a standard inoculum of *Pseudomonas aeruginosa* and determining the survival curve over a 6-h period (Fig. 2) [32]. The procedure was relative rather than absolute; comparison of the survival curves revealed biocide depletion or inactivation.

A much simpler biological assay was proposed for metalworking fluids in 1986 [33] and for fuel water bottoms in 1987 [34] (Biocide Monitor, ECHA Microbiology, Cardiff). Spores of a *Bacillus* sp., dried nutrients, and a tetrazolium salt (as a growth indicator) were carried on a pad mounted on a 5×55 mm plastic strip. This was dipped briefly into the test fluid and incubated overnight at 37 °C. Reduction of the tetrazolium to red formazan accompanied spore germination and growth and thus indicated the presence or absence of a chemical inhibitor. The strip could be calibrated against known biocides and biocide mixtures. The minimum inhibitory concentration (MIC) so determined could then be utilized to assay unknown biocide concentrations by establishing which dilution of a sample just contained the MIC. A simple calculation then yields the concentration of biocide present.

Using this device it is relatively easy to determine whether target biocide concentrations have been achieved and are being maintained.

3.5. Predetermined and Automatic Biocide Additions

It is often argued that any form of monitoring is not financially justified for small systems. A compromise is to set up a top-up dosing regime based on an initial monitoring period, preferably based on monitoring the biocide (preservative) concentration (Section 3.4).

Automatic dosing via proportioning pumps has been very successful in systems where there are large volume losses. Cooling water systems are an example of such a system; losses are due to evaporation, misting, and bleed-off. The bleed-off may be triggered automatically by setting a limit value for conductivity which increases due to evaporation losses. The make-up flow is measured and linked to a proportioning pump which doses biocide to achieve a predetermined concentration. The whole system must be dosed with biocide before operation commences.

4. References

1. F. H. S. Curd, F. L. Rose: "Synthetic antimalarials Part X. Some aryl diguanide derivatives," *J. Chem. Soc.* 1946, 729 – 734.
2. F. L. Rose, G. Swain: "Bisbiguanides having antibacterial activity," *J. Chem. Soc.* **4** (1956) 4422 – 4425.
3. *ICI Antiseptics in Practice,* ICI Pharmaceutical Division, Wilmslow, Cheshire, UK, 1981.
4. M. S. Yeshpe-puter, E. Efan: "Disinfection of water by hydrogen peroxide," *Health Lab. Sci.* **5** (1968) no. 4, 233.
5. J. A. L. Fraser: *The use of Peracetic Acid as a Disinfectant,* Interox Chemicals Limited, Widnes, UK, 1980, p. 40.
6. *The Control of Substances Hazardous to Health Regulations,* HMSO, London 1988.
7. *Report of the Expert Advisory Committee on Biocides,* HMSO, London 1988.
8. Association of Official Agricultural Chemists (AOAC): *Official Methods of Analysis of the Association of Agricultural Chemists,* 13th ed., Washington D.C. 1980.
9. British Standard BS : 541: Rideal – Walker Coefficient of Disinfectants, British Standards Institute, London 1934.
10. British Standard BS : 808: *Modified Technique for the Chick – Martin Test for Disinfectants,* British Standards Institute, London 1938.
11. J. C. Kelsey, I. Maurer: "An improved Kelsey – Sykes test for disinfectants," *Pharm. J.* **213** (1974) 528 – 530.
12. Association Française de Normalisation (AFNOR): *Recueil de normes françaises des antiseptiques et desinfectants,* 1st ed., Paris 1981.
13. B. Van Klingeren, A. B. Leussink, L. J. Van-Wijnguarden: "A collaborative study on the repeatability of the Dutch Standard Suspension Test for the evaluation of disinfectants," *Zb. Bakt. Hyg., I Abt. Orig. B.* **164** (1977) 521 – 548.
14. J. L. Shennan, K. W. A. Chater in K. W. A. Chater, E. C. Hill (eds): *Monitoring and Maintenance of Aqueous Metal-Working*

Fluids, J. Wiley & Sons, Chichester 1984, pp. 113 – 129.
15. P. E. Cook, C. C. Gaylarde, *Int. Biodeterior. Bull.* **24** (1988) 265 – 270.
16. A. C. Peters, J. W. T. Wimpenny, *Biotechnol. Bioeng.* **32** (1988) 263 – 270.
17. R. Elsmore, *Isr. J. Med. Sci.* **22** (1986) 647 – 654.
18. A. H. Rose (ed.): *Microbial Biodeterioration,* Academic Press, London 1981.
19. A. H. Walters, J. E. Elphick (eds.): *Biodeterioration of Materials,* Elsevier, London 1968.
20. A. H. Walters, E. H. Hueck-Van der Plas (eds.): *Biodeterioration of Materials,* Applied Science Publishers Ltd., London 1972.
21. J. M. Sharpley, A. M. Kaplan (eds.): *Proceedings of the Third International Biodegradation Symposium,* Applied Science Publishers Ltd., London 1976.
22. T. A. Oxley, G. Becker, D. Allsopp (eds.): *Biodeterioration,* Pitman Publishing Ltd., London 1980.
23. T. A. Oxley, S. Barry (eds.): *Biodeterioration 5,* J. Wiley & Sons, London 1983.
24. S. Barry, D. R. Houghton (eds.): *Biodeterioration 6,* C. A. B. International, Slough, U.K., 1986.
25. D. R. Houghton, R. N. Smith, H. O. W. Eggins (eds.): *Biodeterioration 7,* Elsevier Applied Science, London 1988.
26. E. C. Hill, I. Davies, J. A. V. Pritchard, D. Byron, *J. Inst. Pet.* **53** (1967) 275 – 279.
27. H. W. Rossmore, *Int. Biodeterior. Bull.* **7** (1971) no. 4, 147 – 154.
28. E. C. Hill in R. J. Gilbert, D. W. Lovelock (eds.): *Microbial Aspects of the Deterioration of Materials,* Academic Press, London 1975, pp. 127 – 136.
29. C. Genner, *Process Biochem.,* July/August (1976), 39 – 41.
30. E. C. Hill in K. W. A. Chater, E. C. Hill (eds.): *Monitoring and Maintenance of Aqueous Metal-Working Fluids,* J. Wiley & Sons, Chichester 1984, pp. 97 – 112.
31. B. Jarvis in K. O. Habermehl (ed.): *Rapid Methods and Automation in Microbiology and Immunology,* Springer Verlag, Berlin 1984, pp. 593 – 602.
32. E. C. Hill in W. J. Bartz, J. Wolff (eds.): *Lubrication in Metal-Working,* Technische Akademie, Esslingen 1982, pp. 1 – 4.
33. E. C. Hill et al. in W. J. Bartz (ed.): *Additives for Lubricants and Operational Fluids,* **vol. 2,** Technische Akademie Esslingen 1986, pp. 11. 10-1 – 10-7.
34. E. C. Hill in E. C. Hill, J. L. Shennan, R. J. Watkinson (eds.): *Microbial Problems in the Offshore Oil Industry,* J. Wiley & Sons, Chichester 1987, pp. 219 – 229.

Microencapsulation

Christopher A. Finch, Pentafin Associates, Weston Turville, Aylesbury HP 22 5 TT, United Kingdom

Roland Bodmeier, College of Pharmacy, Freie Universität Berlin, Berlin, Germany

1. Introduction 1
2. Reasons for Microencapsulation . . . 2
3. Materials 3
4. Production 4
4.1. Phase Separation 5
4.2. Interfacial and in situ Polymerization 8
4.3. Spray Drying, Spray Congealing . . . 9
4.4. Solvent Evaporation Processes 10
4.5. Coating Processes 10
5. Other Forms of Microencapsulation . 12
6. Characterization of Microparticles . . 12
7. Uses 13
8. References 14

1. Introduction

Microparticles can be classified into microparticles in the literal sense and microcapsules. With microparticles, the active ingredient is dispersed/dissolved in the carrier matrix (matrix system), while with microcapsules, the active substance is surrounded by a membrane (reservoir system). Microparticles range in diameter from ca. 1 to 1000 µm, although at the upper end of this size distribution, the classification of the particles as microparticles becomes open to question. Nanoparticles are particles in the submicron range. Most gaseous, liquid, and solid materials – whether hydrophobic or hydrophilic – can be encapsulated. The wall/carrier material and also the process of microencapsulation are selected according to the physical properties of the core and the intended appliation. The carrier material generally consists of natural or synthetic polymers or lipids [1–3].

Microencapsulation can modify the color, shape, volume, apparent density, reactivity, durability, pressure sensitivity, heat sensitivity, and photosensitivity of the encapsulated substance. The microcapsule is able to preserve a substance in a finely divided state, and release the substance from the enclosed capsule as required. Encapsulation of a liquid effectively converts the liquid to a fine powder solid form, whilst preserving the reactivity of the interior phase. Microencapsulation therefore improves the "ease of handling" of the liquid reactants, and allows chemical reaction to commence when the microcapsule is ruptured. The density of a substance can be increased by encapsulation, and can be decreased by including air in the capsule. A dense solid can thus be converted by encapsulation into a product which will float on water.

Microparticles have many useful functions, and have been employed in many different areas, frequently connected with applications in which the contents of the microcapsule are released under controlled conditions into the surrounding environment. Sometimes, however, a core material to be enclosed is microencapsulated and then used directly in the powdered form [4–22].

The storage life of a volatile compound can be increased markedly by microencapsulation. Microcapsules can protect a core substance from the effects of UV rays, moisture, and oxygen. Chemical reaction between two active species can be prevented by physical separation by microencapsulation. Finely divided powders can be microencapsulated to reduce agglomeration problems.

The techniques and practice of microencapsulation cover several scientific disciplines, including colloid and physical chemistry, polymer chemistry and physics, materials science, suspension and drying technology, and, in controlled-release applications, pharmaceutical and agricultural technology. Many of these interact with each other in practical applications, and in the resulting commercial implications [4–8].

Most of the possible deficiencies in microcapsules depend on the properties of the wall materials. Premature failure of the microcapsule wall in service before the required conditions for rupture are achieved can cause diffi-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_575

culties, which must be considered in the design of the microcapsule wall system. Many of the more recent developments in microencapsulation technology have been directed at improving the physical properties of the wall polymers in service, so that the microcapsules do not release the core contents, without rupture or leakage, until required.

The design of a microencapsulation system must take into account the total system: the active and carrier materials, the mechanism of release, and the ultimate fate of all the ingredients. Each of these parameters must be optimized if a satisfactory product is to be obtained. Many methods exist for production of microparticles which allow many variations, depending on core and wall-polymer solubility, capsule size, wall thickness and permeability, type and rate of release of core contents, and physical properties.

In choosing processes for the production of microcapsules for particular applications, several physical properties must be considered, including:

1) *Core Wettability.* The key property is the ability of the core to be wetted by the wall material. In principle, this property can be predicted, but, in practice, it is usually determined during the microencapsulation process.
2) *Core Solubility.* The contents of the microcapsule core should not be soluble in the solvent for the wall polymer, and the polymer should not dissolve significantly in the liquid core. With spray coating, a water-soluble solid can be coated with a water-soluble polymer solution, because the water evaporates rapidly during microcapsule formation.
3) *Wall elasticity* is determined by the nature of the wall polymer, its thickness, and the size of the microcapsules formed.
4) *Wall permeability* determines whether the microcapsule core contents can be retained indefinitely (impermeable wall) until ruptured, or may be released at a predetermined rate (permeable wall) as in controlled-release applications.
5) *Wall polymer adhesive characteristics* are markedly affected by solution temperature and concentration, and depend on the physical properties of the wall polymer, notably the melting point and glass transition temperatures, degree of crystallinity, and degradation rate (under microcapsule formation conditions). Stickiness during capsule formation, and "stringiness" during spray drying can be major problems affecting microcapsule manufacture and storage behavior.

2. Reasons for Microencapsulation

The principal reasons for using microencapsulation are listed below.

Controlled Release. Release of the active component is controlled by the carrier material, various release profiles can be obtained. Many of the actual or potential uses of microparticles can be found within this category, usually with additional functional requirements.

Protection of Core Contents Against Atmospheric Effects. Vitamin A palmitate (widely used as an animal feed additive, especially for egg-laying hens), for example, is microencapsulated with gelatin to increase resistance to moisture and atmospheric oxygen, and therefore improve stability. Other fat-soluble vitamins are similarly treated. Microencapsulated feeds are used in industrial fish farming. These are basically high protein – fat mixtures which are microencapsulated to reduce oxidation of the product during storage, and to delay the rate of release into the aqueous environment of the fish farm. Such protection greatly increases the feeding efficiency and growth rate of the fish. Many flavors and fragrance compounds (e.g., methyl salicylate and peppermint oil) can be prevented from evaporating by microencapsulation.

Protection of Hygroscopic Core Contents. The flow properties of mixtures of the hygroscopic B group vitamins (e.g., thiamine hydrochloride, riboflavin, and niacin) with iron phosphate are improved by microencapsulating them before they are compressed into tablets. Choice of suitable coating materials can improve the performance (in compaction and disintegration) of tablets made from encapsulated drugs.

Reduction of Health Hazards. Pesticides are often encapsulated to reduce the hazard

to the user. The antibiotic ampicillin has been microencapsulated to reduce the possibility of sensitization of factory staff.

Reduction of Gastrointestinal Irritation. Encapsulation of drugs [e.g., iron(II) sulfate, acetylsalicylic acid] reduces gastrointestinal irritation by controlling the rate of release into the digestive system. Enteric polymers, for example, do not release the drug in the low pH environment of the stomach, but in the intestine. Gastric upset can therefore be avoided.

Masking of Taste and Odor. Microencapsulation can be used to disguise the unpleasant taste or odor of compounds such as pharmaceuticals. However, taste masking can also be achieved at lower cost by packaging pharmaceuticals in hard gelatin capsules, or by incorporating the unpleasant compound in sugar- or film-coated tablets.

Odor Release Control. In general, the same considerations apply as in masking. However, in addition to protection and reduction of core volatility, the principal use of microencapsulation for perfume-like materials has been in the preparation of "scratch-and-sniff" coatings which can be applied by printing techniques to advertising material. Upon rupture of the microcapsules by mechanical scratching, an odor related to the product being advertised is released.

Masking of Core Properties. The chemical properties of the microcapsule core (e.g., pH or catalytic properties) can be masked until release is required. The major application of this type is the control of the color change of leuco dyes upon pH change, which is the basis of carbonless copy paper.

Separation of Incompatible Components for Functional Reasons. An example of this application is the use of microencapsulated free-radical catalysts (e.g., benzoyl peroxide) mixed with curable polyester resins, which harden in situ at the point of use.

Other Reasons. Other reasons for microencapsulation include the protection of the active component from physical damage and conversion of a liquid active compound into a fluidized "solid powder" form. The latter is sometimes employed with food additives, so that the active component (e.g., a flavor oil, or a bakery additive) can be evenly distributed in a dry powder mix.

3. Materials

Core Materials. A wide range of materials have been encapsulated. These include:

Solvents: benzene, toluene, cyclohexane, chlorinated phenyls, paraffins, esters, ethers, alcohols, and water.
Plasticizers: citrate-, phthalate-, adipate-, and phosphate-type, silicones, and chlorinated hydrocarbons.
Acids and bases: boric acid, caustic alkali, and amines.
Catalysts: curing agents, oxidants, free radical initiators, and reducing agents.
Colorants: pigments and dyes, especially leuco dyes for carbonless copying papers.
Adhesives: polysulfides, cyanoacrylates, isocyanates, epoxy resins, and thermally-sensitive adhesive compositions.
Fragrances: menthol, essences, and specialty compositions.
Foods: oils, fats, spices, and flavors.
Pharmaceuticals: aspirin, vitamins, and peptides/proteins.
Agricultural chemicals: herbicides, insecticides, and fertilizers.
Recording materials: reprographic toners, coupling agents, developers, silver halides, fixing agents, photochromic compounds, and liquid crystals.
Rust inhibitors: zinc chromate and other compounds.
Others: detergents, bleaches, and fire retardants.

Wall Materials. Wall materials [22] may be natural, semisynthetic, or synthetic:

Natural Materials: gum acacia (gum arabic), agar, agarose, maltodextrins, sodium alginate, calcium alginate, dextran, fats and fatty acids, cetyl alcohol, milk solids, molasses, gelatin, gluten, albumin, shellac, starches, caseinates, stearins, sucrose, and waxes (e.g., beeswax, carnuaba, and spermaceti).

Semisynthetic Materials: cellulose acetate, cellulose acetate butyrate, cellulose acetate phthalate, cellulose nitrate (collodion), ethylcellulose, hydroxypropylcellulose, hydroxypropylmethylcellulose, hydroxypropylmethylcellulose phthalate, methyl cellulose, sodium carboxymethylcellulose, hydrogenated tallow, myristyl alcohol (1-tetradecanol), glyceryl mono- or dipalmitate, hydrogenated castor oil, glyceryl mono-, di-, or tristearate, and 12-hydroxystearyl alcohol.

Synthetic Materials: acrylic polymers and copolymers [e.g., polyacrylamide, poly(alkyl cyanoacrylate), and poly(ethylene – vinyl acetate)], aluminum monostearate, carboxyvinyl polymers (Carbopol), polyamides, poly(methyl vinyl ether – maleic anhydride), poly(adipyl-L-lysine), polycarbonates, polyterephthalamide, poly(vinyl acetate phthalate), poly(terephthaloyl-L-lysine), polyarylsulfones, poly(methyl methacrylate), poly(ε-caprolactone), polyvinylpyrrolidone, polydimethylsiloxane, polyoxyethylenes, polyesters, polyglycolic acid, polylactic acid and copolymers, polyglutamic acid, polylysine, polystyrene, poly(styrene – acrylonitrile), polyimides, and poly(vinyl alcohol).

4. Production

Microencapsulation Methods. Many different methods have been proposed for the production of microcapsules (several hundred methods/modifications have been identified in the patent literature) with many variations in detail, depending on core and wall-polymer solubility, particle size, wall thickness and wall permeability, type and rate of release of core contents required, physical properties, and detailed and overall economics of manufacture. The choice of a particular preparation method and a suitable polymer will depend on the physicochemical properties of the active substance, the desired release characteristics, the therapeutic goal for drug substances, the route of administration, the biodegradability/biocompatibility of the carrier material and regulatory considerations. From a technological point of view, the successful selection of a preparation method will be determined by the ability to achieve high loadings with the active substance, high encapsulation efficiencies, high product yields, and the potential for easy scale-up. For example, methods with high encapsulation efficiencies but with only low active loading capacity are limited to very potent active subtances.

The preparation of colloidal carrier systems, such as polymeric nanoparticles, generally also belongs to the area of microencapsulation (Chap. 5). Microencapsulation methods can be classified into the following groups (see also Table 1):

1) Phase separation (Section 4.1)
2) Interfacial and in situ polymerization (Section 4.2)
3) Spray drying, spray congealing (Section 4.3)
4) Solvent evaporation (Section 4.4)
5) Coating (Section 4.5)

Table 1. Microencapsulation processes

Mircoencapsulation process	Core material	Approximate particle size, µm
Phase separation (coacervation)	solids, liquids	2 – 1200
Interfacial polymerization	solids, liquids	2 – 2000
Spray drying and congealing	solids, liquids	6 – 600
Solvent evaporation	solids, liquids	5 - 500
Centrifugal extrusion	solids, liquids	1 – 5000
Air suspension	solids	35 – 5000

Alternatively, encapsulation methods may be classified according to the nature of the suspending medium (liquid or vapor):

Liquid Suspending Media. If a liquid is used as a suspending medium, microencapsulation is achieved by emulsification or dispersion of two or more immiscible phases. These methods involve:

Complex coacervation
Polymer – polymer incompatibility
Interfacial polymerization
In situ polymerization
Desolvation
Solvent evaporation from emulsion
Gelation
Pressure extrusion
Supercritical fluid technology

Vapor Suspending Media. If the vapor is the suspending medium, the microcapsules are formed by atomization of a liquid phase; methods include:

Spray drying and congealing
Fluidized-bed process (including Wurster process)
CoextrusionVvacuum coating
Gelation
Electrostatic deposition

Most microencapsulation systems employ a combination of several processes. In choosing processes for particular applications, the physical properties of the system must be considered.

Microparticle Size and Quality. Many factors affect the size and quality of microparticles. Some of these influence the performance of controlled release of drugs and active components in other applications (e.g., pesticides and aromas).

Factors affecting the size of microcapsules during production are:

Stirring rate
Solids content of organic phase
Viscosity of organic phase
Viscosity of aqueous phase
Concentration and type of surfactant (if any)
Configuration of vessel and stirrer
Quantity of organic and aqueous phases
Temperature profile during production
Nozzle design

Factors affecting the quality of microcapsules during production are:

Choice of solvent
Mixtures of solvents used
Aqueous solubility of active agent
Rate of solvent removal
Drying conditions
Type and molecular mass of carrier
Crystallinity of polymer

4.1. Phase Separation

Phase separation can be divided into aqueous or organic phase separation, depending on the solubility of the wall material. In phase separation, the core material is first suspended in a solution of the wall material. The wall polymer is induced to separate as a viscous liquid phase (not as a precipitate) by several different methods (e.g., by adding either a nonsolvent for the polymer or salts, by lowering the temperature, or by adding an incompatible polymer). This separation process is known as coacervation. Coacervation is recognized by the appearance of turbidity, droplet formation, or actual separation of liquid layers [23–25]. Coacervation may be simple or complex, or may occur as a result of a salting out effect.

Simple Coacervation. In simple coacervation, addition of a water-miscible nonsolvent (e.g., ethanol) to an aqueous polymer solution causes formation of a separate polymer-rich phase due to a partial miscibility effect. An example of a simple coacervation system consists of water, gelatin, and ethanol. The system is, however, difficult to control, and the core material must be insoluble in both ethanol and water.

Typical aqueous polymer solutions and the appropriate nonsolvents are:

Gelatin	ethanol, acetone
Agar	acetone
Pectin	dioxane, 2-propanol
Methyl cellulose	acetone
Poly(vinyl alcohol)	1-propanol
Fibrinogen	1-propanol

An example for organic phase separation is the microencapsulation of water-soluble organic compounds which can be finely dispersed in a solution of cellulose acetate butyrate [*9004-36-8*] (CAB) in methylene chloride. Toluene (a nonsolvent for CAB) is slowly added, causing coacervation of the polymer around the emulsified droplets of the aqueous solution of the organic compound as it separates from solution. The capsule walls can then be hardened by addition of an aliphatic hydrocarbon solvent with a low solvent power (e.g., hexane). Spherical microspheres with thin elastic walls are usually obtained.

Salt Coacervation. In salt coacervation, a polymer separates from an aqueous solution due to "salting out". An electrolyte is added to the aqueous polymer solution so that phase separation occurs. This method can be used to encapsulate water-insoluble oils or dispersed solid parti-

cles. Large amounts of salt are employed, which have to be washed out of the system. Salting out has the limitation that it is difficult to control capsule size and the resulting microcapsules tend to agglomerate. The microcapsules can be stabilized by adjusting the pH or temperature. The methods can be employed with polymers such as gelatin, poly(vinyl alcohol), or anionic polymers such as sulfonated polystyrene, hydrolyzed styrene – maleic anhydride copolymer, starch, or cellulose acetate phthalates. According to KONDO [9], the ranking of the coacervating power for gelatin for different salts is:

1) Addition of cations: $Na^+ > K^+ > NH_4^+ > Li^+$
2) Addition of anions: sulfate > citrate > tartrate > chloride. Sodium and ammonium sulfates are normally employed.
3) Change of pH causing insolubilization and precipitation of polymer.

Encapsulation may be carried out by using a combination of salt and heat with a system based on a mixture of methyl cellulose and carboxymethyl cellulose; the solubility of methyl cellulose decreases with increasing temperature.

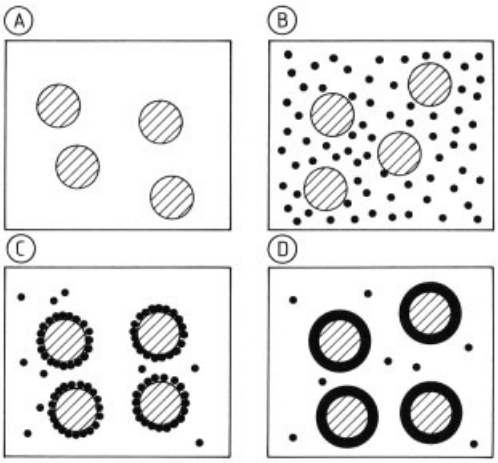

Figure 1. Microencapsulation by coacervation
A) Droplets of core material dispersed in gelatin – gum arabic solution; B) Onset of coacervation by separation of finely divided microcoacervates from the solution; C) Gradual separation of the microcoacervate on to the surface of the core material droplets; D) Coalescence of the microcoacervate into wall material of the droplets

Complex Coacervation. Complex coacervation results from the mutual neutralization of two or more oppositely charged colloids in aqueous solution. For example, the positively charged gelatin (pH < 8) forms a complex coacervate with negatively charged gum arabic.

Complex coacervation is closely related to the precipitation of colloidal material from solution: coacervation immediately precedes precipitation. Microencapsulation by complex coacervation consists of three stages (Fig. 1):

1) Dispersion of the active component to be encapsulated into an aqueous solution of a polyelectrolyte;
2) Deposition around the core material of the coacervate formed by addition of an aqueous solution of a second electrolyte of opposite charge; and
3) Gelation of the coacervate.

This principle of phase incompatibility has also been used for production of microcapsules based on the incompatibility of different polymers. Polymer – polymer incompatibility depends on the free energy of mixing of the polymers. An example of this system is the encapsulation of coffee flavors shown in Figure 2. The components are emulsified to form microcapsules, and then stabilized by addition of a hydrocarbon oil and a dehydrating liquid.

The complex coacervation process was originally developed in the 1950s for the coating process used in the manufacture of "no carbon required" carbonless copying paper by the NCR Company (Appleton, Wisconsin, USA) using the protein gelatin and the carbohydrate gum arabic as the electrolyte components [26].

In carbonless copying papers, the contents of the microcapsule core are released by mechanical action when the paper is struck with a typewriter key. These contents consist of a leuco dye (e.g., crystal violet lactone; → Cationic Dyes, Chap. 2.2.) which changes color due to the pH change occurring when it interacts with the acid clay coating of another face of the paper. In terms of volume, carbonless paper copying systems are by far the largest market for encapsulated products, although other processes, apart from gelatin microencapsulation, are now used for the production of these papers (see Section 4.2).

The dispersion of microcapsules contains an acid-sensitive leuco dye system and is coated on the underside of the top paper sheets. Impact from a writing instrument or type element ruptures the microcapsules, and releases the dye

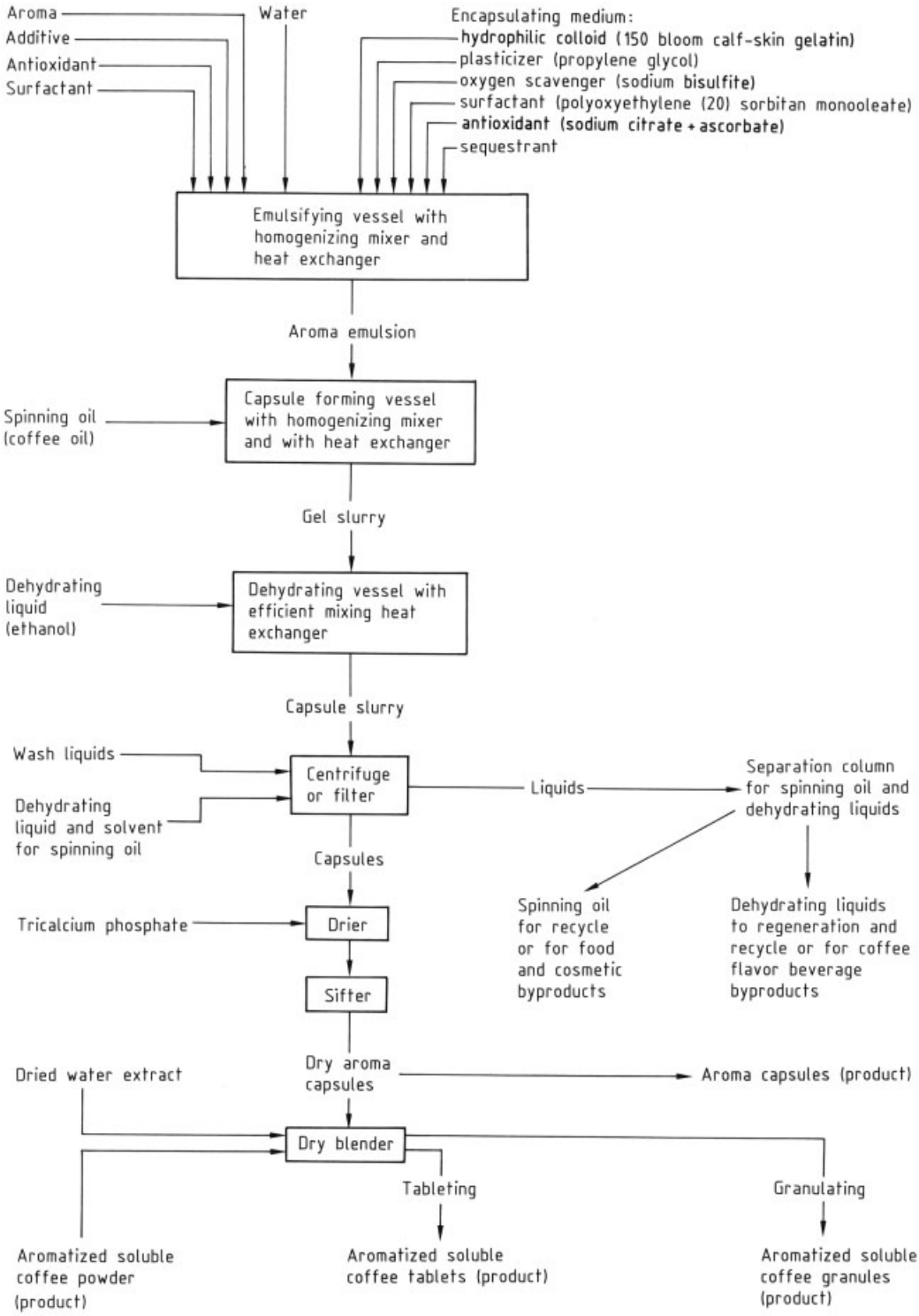

Figure 2. Flow diagram for encapsulation of coffee flavors [27]

solution. The dye thus comes into contact with the acid clay coating the upper surface of the copying paper and develops its color (Fig. 3).

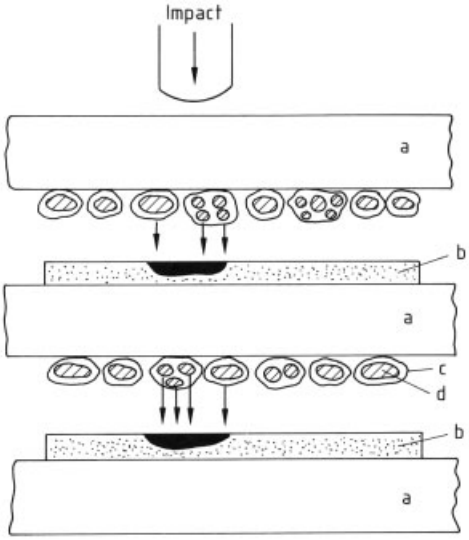

Figure 3. Three-sheet carbonless copying paper system with microcapsules
a) Paper; b) Acid clay; c) Microcapsule; d) Oil solution of leuco dye

Coacervation is reversible, and in general, the formation of microcapsules by phase separation methods, such as coacervation, is an equilibrium process. For gelatin –gum arabic microcapsules, the equilibrium is stabilized by cross-linking or "curing" the gelatin.

Gelatin with a high isoelectric point (pH 8, normally a pigskin-origin gelatin) and gum arabic are added to a suspension containing the active leuco dye core material at pH 4 – 5. The temperature is maintained above 35 °C to ensure that the coacervate is in the liquid phase. As the positively charged gelatin and the negatively charged gum arabic interact, viscous liquid microdroplets of the polymer coacervate separate. These particles are stabilized (i.e., cross-linked or cured) by reaction with formaldehyde or glutaraldehyde; reaction with $CrCl_3$ or $CuSO_4$; addition of alum; reaction with tannic or gallic acid or their iron salts, as activated phenolics; or heat treatment.

The mixture is then cooled to below 10 °C, adjusted to pH 9 and the microcapsules removed by filtration. In a variation of this process, two different gelatins, prepared by different manufacturing methods from different raw materials, are used as the two polymers.

An example of a simple and reliable method for the laboratory-scale production of microcapsules by the coacervation system is outlined as follows [27]:

1) Add 10 g gelatin to 225 g water, and allow to swell for 15 min.
2) Heat to 42 – 45 °C, and add 100 mL of internal phase (active compound in liquid form or in solution) and disperse to the desired droplet size.
3) Add 10 g gum arabic in 225 g water heated to 42 –45 °C.
4) Adjust pH to 4.5 – 5.5. This will form capsules with a size of 200 – 700 µm. For smaller capsules (30 –100 µm) add up to 700 mL water.
5) Cool to 28 °C with stirring.
6) Cool below 10 °C with ice.
7) For hardening, add 5 mL of 25 % formaldehyde or glutaraldehyde solution. (Use of formaldehyde gives an elastic capsule wall. Use of tannic acid gives a brittle wall.)
8) Wash 3 – 5 times to remove the free coacervate.
9) Add dispersing agent (colloidal silica) and wetting agent.
10) Filter microcapsules.
11) Dry with cool air in a fluidized bed.

4.2. Interfacial and in situ Polymerization

The principal chemical reaction methods for the production of microcapsules involve *interfacial polymerization,* typically between a diacyl chloride and an amine or alcohol. The resulting polymer films may be a polyester, polyurea, polyurethane, or polycarbonate. A range of pesticides have been encapsulated by this method, in which a solution of the pesticide and a diacyl chloride is emulsified in water and an aqueous solution of amine and a polyfunctional isocyanate is added. Upon addition of a base, the microcapsules are formed at the interface between the aqueous medium and the emulsion droplets. The isocyanate acts as a cross-linking agent to reduce the permeability of the capsule walls and to increase their resistance to damage.

Durable, impact-resistant microcapsules with high-tensile-strength walls have been developed for pesticide and pheromone formulations [28].

In the related process of *interfacial coacervation*, a polymer coacervate of an emulsified water-soluble natural extract is emulsified with a cold saturated solution of diethyl ether in water that contains a polyethylenesorbitan ester emulsifer (Tween 20); a collodion solution is then added. The capsule walls are hardened with *n*-butyl benzoate.

In a process developed by 3M, a water-insoluble core material is emulsified into an aqueous solution of a low molecular mass urea – formaldehyde (UF) resin, followed by reduction of the pH to about 1.5 – 3 with formic acid, causing the formation of a UF polymer shell. The microcapsules formed have a diameter of 10 – 50 µm, and have been used in the encapsulation of fragrances for advertising purposes [29].

Methods based on *in situ polymerization* have been employed to prepare microcapsules for carbonless copying paper systems. Condensation polymerization systems are used to form the microcapsule wall polymers that contain the color systems. The principles of carbonless copying paper are described in Section 4.1.

The production of microcapsules for carbonless copying paper using melamine – formaldehyde (MF) and UF resins condensates as wall polymers follows [30–33]:

1) Emulsify and mix aqueous solution of MF/UF precondensates and acrylamide – acrylic acid copolymers with capsule contents (solvent + dye)
2) Add dilution water and stir for 30 min
3) Adjust pH with acetic acid to pH 4.7
4) Maintain at 55 °C for 2 h
5) Cool and stir overnight
6) Raise pH to 10.0
7) Microcapsule mixture ready for coating

The more recently developed carbonless copying paper systems using MF resins and a controlled acid developing system are described in detail in [34], [35]. The polymers used for microencapsulation of the leuco dye comprise methylated MF precondensates with acrylamide – acrylic acid copolymers in aqueous solution. Water-soluble polymers (starch or cellulose ethers) are added as viscosity controllers. The microcapsule leuco dye system is coated on the back of the top paper sheet. The acid component is coated onto the front of the lower sheet of the self-copy system and is based on a dispersed hydrated silica coated with an aluminum salt [36].

The (acid) aluminum compound is distributed preferentially on the surface of the silica, and may be modified with other metal ions. There are further developments, especially with the introduction of improved cross-linking and encapsulating systems for the dyes.

4.3. Spray Drying, Spray Congealing

Spray drying is a very attractive and versatile microencapsulation process for a variety of reasons. Active materials with differing solubility properties can be encapsulated with various wall materials and partitioning of the active substance between two immiscible phases is avoided. In addition, when compared to other multi-step microencapsulation processes, spray drying is a single step operation, which has been scaled up in other industrial applications. In the spray-drying process, a microparticle powder is formed by spraying (atomizing) a solution/dispersion of active material and the film-forming wall material as fine droplets into hot air. The water then evaporates and the dried solid is separated, usually by air separation [9] (see also → Drying of Solid Materials, Chap. 2.1.5.). Several process variables can be important in achieving the desired properties, including the core : wall material ratio, and the concentration, viscosity, and temperature of the starting solution. An aqueous solution, an organic solution, or a slurry of the material to be coated may be used.

Drug-containing biodegradable poly(lactide) microparticles were prepared by spray drying an organic polymer solution [37]. Solvent-free polymeric microparticles were prepared by spray drying an aqueous polymer dispersion, whereby the colloidal polymer particles coalesced into larger microparticles [38].

The design of the spray dryer can be important (e.g., the structure of the drying drum, flow pattern and flow rate of the air current, drying temperature, and type of collector). Some exam-

Table 2. Microencapsulation by spray drying∗ [9]

Core	Polymer solution	Drying temperature (inlet – outlet), °C	Microcapsule size, µm
Water-insoluble polymers			
Liquid paraffin (100)	acetyl cellulose (50) acetone (600)	130	50
Dye solution (50)	zein (40) methanol (200)	130	1 – 60
Carbon black (1)	polystyrene – butyl methacrylate latex (15 %) (16)	95 – 80	7 – 11
Paraformaldehyde (1)	polycarbonate (2) chloroform (100)	110 (steam)	10
Water-soluble polymers			
Liquid paraffin (100)	hydroxyethyl cellulose (150) water (300)	130	5
TiO_2 (200) $CaCO_3$ (385)	polyacrylic acid (325)	50 – 35	2
Choline chloride (63)	clay (11) water (37)	160 – 190	100
Lemon oil (37)	carboxylated dextrin (150)	85 – 38	10 – 20
Dimethyl polysiloxane (50)	methyl cellulose (5) solid milk (45)	265 – 120	100

∗ The figures in parentheses indicate parts by weight.

ples of materials encapsulated by this technique are given in Table 2.

Similar types of spray-drying systems using different operating conditions can be used for microencapsulation of labile materials because of the brief contact time in the dryer. The method has been used to encapsulate fragrance oils, which are emulsified into a water-soluble polymer solution (e.g., with gum arabic) to make fine particles. These are used in products such as bath powders, when the formulation is controlled to release the aroma when the mixture is dissolved.

Spray Congealing. The related spray-congealing process is carried out by mixing the core material with a molten wax or fat. The molten formulation is then sprayed and chilled below its melting point to form particles. The system can be used for microencapsulation of such solid materials as citric acid, vitamin C, potassium chloride, and iron(II) sulfate.

As a drying process, spray drying is relatively expensive, but appropriate variations on the basis principle can be an economic form of production of microcapsules.

4.4. Solvent Evaporation Processes

The solvent evaporation method is a technique widely used for the preparation of biodegradable and nondegradable microspheres [39–41]. In this method, the active material and the polymer are dispersed or dissolved in a volatile, water-immiscible organic solvent (e.g., methylene chloride, chloroform, ethyl acetate). This solution or dispersion is then emulsified in an external aqueous phase containing an emulsifying agent by using conventional emulsification equipment to form an o/w-emulsion. Microspheres with a size range generally between 5 – 250 µm are obtained after solvent evaporation. Depending on the solubility of the active substance, various modifications of the solvent evaporation method exist. A multiple emulsion (w/o/w-method) or an external oil phase can be used for water-soluble active compounds.

4.5. Coating Processes

Small, particulate cores can also be coated by air suspension coating (fluidized-bed coating technologies). In air suspension techniques, particles are suspended in an upward-moving air stream and are coated by spraying polymer solutions or aqueous colloidal polymer dispersions onto the particles.

Equipment for this (and other) types of coating is made by several firms [42], [43]. The most widely used device is the *Wurster coating chamber*. "Combination" sets of equipment can be

modified for use with different coating systems. The equipment is automatically controlled and is designed for safe operation, taking account of the explosive hazard problems of electrostatic charges which can occur in coating systems employing flammable organic solvents with a low boiling point. For this reason, pressure-release devices are normally incorporated into production equipment. It is claimed that air control and powder coating systems provide more uniform film coatings on the active materials. This in turn gives reliable reproduction of predetermined release profiles for controlled-release microencapsulation systems. A typical small-scale fluidized-bed unit for pharmaceutical applications including granulating, drying, and film coating of solids (powders, tablets, and pellets) is shown in Figure 4.

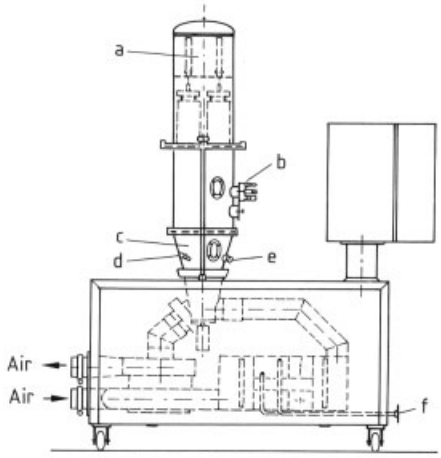

Figure 4. Small-scale general-purpose dryer/encapsulator for pharmaceutical and related products (Glatt AG)
a) Exhaust air filter; b) Spray nozzle; c) Product container; d) Temperature sensor; e) Sample port; f) Drain

Wurster Process. The Wurster process (Fig. 5), developed progressively since the mid 1950s, has been offered in the United States since 1976, and more recently in Europe [43]. It is particularly suited for uniformly coating particles in a single operation. It is used for pharmaceutical products (particularly for controlled release and taste masking), foods, and animal feeds (to improve shelf stability and reduce caking), chemicals (for controlled release and to change surface and adsorption properties), agricultural products (seed coating for easier handling and machine planting), pesticides, nutrients, and rodenticides (for taste masking and controlled release).

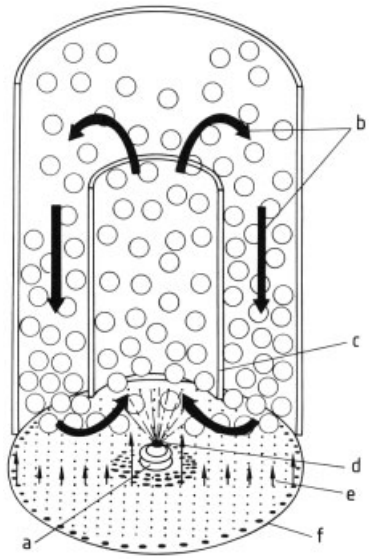

Figure 5. Diagram illustrating the Wurster process (Coating Place Inc.)
a) Hydraulic or pneumatic nozzle; b) Controller particle flow; c) Coating partition; d) Coating spray; e) Air flow; f) Air distribution plate

The Wurster system is reproducible and controllable, and can be run on a small as well as on a production-size scale.

The operating chamber is designed to induce a smooth cyclic flow of particles past a nozzle which atomizes the coating solution. The wall material could also be a molten wax (e.g., a fatty material), or dissolved in an aqueous or organic solvent. The coated particles are lifted on the air stream and dried as they rise on the airstream from the nozzle. They then settle out and descend to begin another cycle. The coating film thickness is controlled by the type of coating, the solvent concentration, the air volume and temperature, atomizing conditions, and the number of coating cycles; since many thin coats are applied, the systems can be used with irregularly shaped particles.

Because this process uses large volumes of air, the system has a large drying capacity and the particles are separated as they are carried on the air stream. Small particles can thus be coated with good control of agglomeration.

Suitable coating polymers are offered by Röhm, FMC, Colorcon, or BASF. The release of the active substance can be controlled by several process and formulation variables.

5. Other Forms of Microencapsulation

Nanoparticles are active substance-containing colloidal particles with a size $< 1\,\mu m$. The active compound can be present in various physical states; it can be dissolved in the polymeric matrix, it can be encapsulated, or it can be adsorbed or attached to the surface of the colloidal carrier. The techniques used to prepare nanoparticles are generally classified into two groups. In the first group, nanoparticles are formed from preformed polymers. Alternatively, nanoparticles are prepared by various polymerization reactions of lipophilic or hydrophilic monomers.

Nanoparticles Prepared from Preformed Polymers. Nanoparticles have been prepared from a variety of both water-soluble and water-insoluble polymers of synthetic, semisynthetic, or natural origin. The use of preformed polymers has several advantages as compared to nanoparticles prepared by polymerization of monomers. Advantages include the use of polymers with well-characterized physicochemical properties, established safety and approval standards, the absence of residual monomers or polymerization reagents (e.g., initiators or catalysts) and the lack of possible reactions between drugs and monomers. In addition, it is not possible to obtain nanoparticles by emulsion polymerization in the case of polymers where suitable monomers are not available (e.g., polylactic acid and lactic – glycolic acid copolymers, cellulose derivatives such as ethyl cellulose and cellulose acetate or natural polymers such as gelatin or sodium alginate).

Most techniques to prepare nanoparticles from *synthetic water-insoluble polymers* are derived from conventional emulsification technologies, which have been used to prepare microparticles, such as the solvent evaporation method. To prepare nanoparticles and not microspheres, the o/w-emulsion is homogenized under high shear with an appropriate homogenization equipment (e.g., microfluidizer, sonication) prior to the precipitation of the polymer to further reduce the particle size of the internal organic phase into the colloidal size range [44–47]. Drug – polymer particles in the nanometer size range are formed after solvent diffusion into the aqueous phase and evaporation at the water/air interface causing polymer and drug precipitation. Nanoparticles have been prepared with conventional laboratory homogenizers [44], by ultrasonication [45], [46], [48] or by microfluidization [47].

Hydrophilic polymers are primarily of natural origin. The nanoparticles are prepared either by an w/o-emulsification method or by aqueous phase separation techniques (e.g., coacervation). Preparation of nanoparticles of hydrophilic polymers (e.g., albumin, chitosan, gelatin or carbohydrates) by w/o-emulsification techniques has been developed for the preparation of albumin microspheres [49–52]. Basically, an aqueous polymer solution is emulsified into an external, water-immiscible phase, such as an oil or organic solvent followed by homogenization. Upon water removal, the polymer droplets solidify. Insoluble nanoparticles can be obtained by further hardening/insolubilizing the polymer through chemical cross-linking with aldehydes or other cross-linking agents or through denaturation at elevated temperatures.

Nanoparticles Prepared by Polymerization of Monomers. The methods used to prepare nanoparticles by polymerization are derived from methods developed to prepare latices. The monomer, which are primarily of acrylic origin, can be either dissolved or emulsified in the continuous phase, with the resulting polymer being insoluble in either case. The preparation of nanoparticles by polymerization reactions is generally classified according to the resulting polymer or the resulting particle structure.

6. Characterization of Microparticles

Formulation and process parameters of the described preparation techniques will strongly affect the performance of the microparticles. A precise characterization of the physicochemical

parameters is important in order to develop a product of high quality.

The parameters generally used to characterize microparticles include the particle size, morphology, and surface charge, the drug loading and encapsulation efficiency, the drug release profile, the physicochemical state of the drug and polymer, molecular weight of the polymeric carriers and in vitro biodegradation.

7. Uses

A great diversity of products have been microencapsulated for many different purposes. Only a limited number of these can be described in addition to those already mentioned.

Food Industry. Many foods and animal feed additives have been encapsulated in gelatin and vegetable gums.

Food additives are encapsulated to improve storage life by reducing volatilization and oxidative degradation. However, small amounts of free active material may be released during storage and contaminate the encapsulated product. Spray drying may overcome this problem at the cost of loss of low-boiling volatile components. Alternatively, the mixture may be microencapsulated by extrusion under pressure into alcohol at 100 – 120 °C to solidify the capsules. The resulting particles are larger than those obtained by spray drying, and the process may also adversely affect heat-sensitive components.

Examples of encapsulated food additives include flavors and aromas (such as citrus oils, spices, and coffee oil, see Section 4.1). Bakery flavors (e.g., ethyl vanillin) have been encapsulated in ethyl cellulose and gum arabic. Liquid additives are dispersed as droplets in a matrix of carrier material. Garlic oil has been encapsulated to reduce odor difficulties in handling [53–56].

In animal feeds, microencapsulation is used to protect atmosphere-sensitive additives: the largest application is the coating of synthetic vitamin A with gelatin for use as a poultry feed additive to enhance the color of egg yolk. In addition, microencapsulation is being increasingly used to produce edible microglobules for fish farming. Trout feed, for example, is obtained from gelatin microcapsules that are cross-linked with glutaraldehyde to control the rate of dissolution into water [56]. Uses of such products greatly increases the food conversion and growth rates of farmed fish and crustaceans, notably trout and prawns.

Pigments. Encapsulated pigments (e.g., in cosmetics) have been prepared using acrylic copolymers as the film-forming agents. Polyvinylpyrrolidone has been used as the wall-forming polymer for microencapsulating additives for specialty paints.

Laundry Products. Microcapsules are used in the formulation of laundry products, notably fabric softeners [57], optical brighteners [58], [59], and chlorine donors [59], [60].

Paper. The principal application of microcapsules in the coating of paper is the production of carbonless copying paper (see Section 4.1). Microencapsulated, air-containing opacifiers have also been used to increase the whiteness of the product.

Pharmaceuticals. Microcapsules are employed in many applications in pharmaceuticals (→ Pharmaceutical Dosage Forms). For example, aspirin is coated with ethyl cellulose and other cellulose esters to reduce stomach irritation [61]. Use of cellulose acetate phthalate as part of the capsule wall results in capsules which are insoluble in the acid environment of the stomach and dissolve only at higher pH in the intestines [62].

Of the many drugs based on microencapsulated controlled delivery systems, the "Contac"-type cold-relief remedies (and many similar competitive products) have been among the first products available on the market. The active components (usually phenylpropanolamine hydrochloride [*700-65-2*] and chlorpheniramine maleate [*113-92-8*]) are encapsulated in particles with a range of different sizes. The particles are pan-coated to different wall thicknesses, all contained in a hard-shell gelatin capsule. The particle sizes and coating thicknesses are designed so that the microcapsules dissolve in the digestive system at different rates, and release their active contents over a period of time.

In recent years, biodegradable microparticles based on poly(lactide – glycolide) copolymers

have been developed. These microparticles are used as encapsulation for some peptide drugs.

Cosmetics. Microencapsulation is used in the manufacture of beauty aids, including encapsulated perfumes (delayed release) and cosmetic aids (e.g., make-up removers).

Pest and Weed Control. There have been many applications of microcapsules in agriculture, to improve the delivery of active compounds (i.e., pesticides and herbicides) to the required target organisms. In many cases, microcapsules have been manufactured from specific products for particular uses. Capsule properties (i.e., storage life, and the time and rate of release of active component under field conditions) must be related to the identified biological behavior of the organism (such as a harmful pest or weed) which is to be controlled. Many pesticides have been microencapsulated, including methyl parathion [*298-00-0*], diazinon [*333-41-5*], thiocarbamates, and pyrethrins.

Durable microcapsules have been developed for the sustained and controlled release of biologically active materials, notably pesticides, to provide long-term effectiveness. These microcapsules have an outer wall of gelatin and an inner wall of a synthetic polymer, which may be polysulfone, polycarbonate, poly(methyl methacrylate), or a poly(styrene – acrylonitrile) copolymer. The active components are mixed in dichloromethane with the wall polymer, then emulsified with gelatin and nonionic surfactant at 45 °C until all the solvent has evaporated [63]. This microencapsulation system has been used not only for conventional insecticides (e.g., Naled), but also for the controlled release of pheromones such as " gossyplure" (pink bollworm pheromone) and " disparlure" (gypsy moth pheromone). The main advantage of the system is that the double-walled microcapsules have high impact resistance, and a long storage life compared to those produced by other systems.

Miscellaneous Uses. *Scented strips*, made by coating perfume-loaded microcapsules onto paper, have been widely used for advertising purposes in "scratch-and-sniff" sales promotions. Both gelatin-based and amino-resin-based capsule walls have been used.

Microencapsulated anaerobic *adhesives* are used to coat the threads of bolts and other fasteners. When a nut is screwed on to a coated thread, the microcapsules are fractured, and the adhesive seals the voids, so securing the fastening. Microencapsulated epoxy resin hardeners are also produced for control of adhesive resin setting rates.

Microencapsulation has also been used to encapsulate enzymes and cells (→ Immobilized Biocatalysts, → Immobilized Biocatalysts) and to make artificial cells [64].

Other microencapsulated products include dental amalgams containing encapsulated mercury; photographic materials and temperature-sensitive liquid crystals; tobacco flavors; weather modification systems (using microencapsulated hygroscopic seeding agents); microencapsulated reagents and catalysts; and flame retardants (as components of rigid and flexible foams).

8. References

General References

1. S. Benita: *Microencapsulation – Methods and Industrial Applications,* Marcel Dekker, New York 1996.
2. P. B. Deasy: *Microencapsulation and Related Processes,* Marcel Dekker, New York 1984.
3. J. Kreuter: *Colloidal Drug Delivery Systems,* Marcel Dekker, New York 1994.
4. *Journal of Microencapsulation,* Taylor & Francis, London.
5. *Journal of Controlled Release,* Elsevier, New York.
6. *Ullmann,* 4th ed., **16,** 675 – 682.
7. *Kirk-Othmer,* 3rd ed., **vol. 15,** 470 – 493.
8. C. Thies: "Microencapsulation" in *Encyclopedia of Polymer Science and Engineering,* vol. 9, Wiley-Interscience, New York 1987, pp. 724 – 745.

Specific References

9. A. Kondo in J. Wade van Valkenburg (ed.): *Microcapsule Processing and Technology,* Marcel Dekker, Basel 1979.
10. M. W. Ranney: *Microencapsulation Technology,* Noyes Data Corp., Park Ridge, New Jersey 1969.
11. M. H. Gurcho: *Capsule Technology and Microencapsulation,* Noyes Data Corp., Park Ridge, New Jersey 1972.

12. M. H. Gutcho: *Microencapsulation and Microencapsulation Techniques,* Noyes Data Corp., Park Ridge, New Jersey 1976.
13. M. H. Gutcho: *Microcapsules and other Capsules – Advances since 1975,* Noyes Data Corp., Park Ridge, New Jersey 1979.
14. J. R. Nixon (ed.): *Second International Symposium on Microencapsulation,* Marcel Dekker, New York 1976.
15. P. B. Deasy: *Microencapsulation and Related Drug Processes,* Marcel Dekker, New York 1987.
16. P. Guiot, P. Couvreur (eds.): *Polymeric Nanocapsules and Microspheres,* CRC Press, Boca Raton 1986.
17. L. Illum, S. S. Davis (eds.): *Polymers in Controlled Drug Delivery,* Wright, Bristol 1987.
18. D. S. T. Hseih (ed.): *Controlled Release Systems: Fabrication Technology,* vol. I, Chap. 5, vol. II, CRC Press, Boca Raton 1988.
19. P. I. Lee, W. R. Good (eds.): "Controlled Release Technology: Pharmaceutical Applications," *ACS Symp. Ser.* **348** (1987).
20. S. S. Davis, L. Illum, J. G. McVie, E. Tomlinson (eds.): *Microspheres and Drug Therapy: Pharmaceutical, Immunological and Medical Aspects,* Elsevier, Amsterdam 1984.
21. S. J. Risch, G. A. Reineccius (eds.): "Flavor Encapsulation," *ACS Symp. Ser.* **370** (1988).
22. C. A. Finch, *Chem. Ind. (London)* 1985, 752 – 756.
23. H. G. Bungenberg de Jong, H. R. Kruyt, *Kolloid Z.* **50** (1930) 39.
24. H. R. Kruyt: *Colloid Science,* **vol. 2,** Elsevier, New York 1949, p. 216.
25. T. N. Evreinova, W. N. Karaukov, T. W. Marmontova, G. R. Ivanski, *J. Colloid Interface Sci.* **36** (1971) 18.
26. NCR Co., US 2 800 457, 2 800 458, 1957; US Patent reissue no. 24 899, 1960 (B. K. Green).
27. Omni TECHNIK Mikroverkapselungs GmbH, Munich, FRG.
28. Bond Research Inc., US 4 670 250, 1987 (R. W. Baker). Rhône-Poulenc SA, EP-A 278 878, 1988 (G. Vivaur).
29. 3M, US 3 516 846, 1970; 3 516 941, 1970 (G. W. Matson).
30. Wiggins Teape Ltd., EP-A 138 448, 1983 (P. Pinot de Moira). Wiggins Teape Ltd., EP-A 145 225, 1983 (K. J. Shanton, F. Azizan).
31. Fuji Photo Film Co., US 3 944 502, 1975 (H. Matsukawa, K. Saeki, T. Shimada).
32. Polaroid Inc., US 3 396 026, 1968; 3 702 829, 1972 (L. D. Taylor).
33. Wiggins Teape Ltd., GB 1 507 739, 1978 (D. J. Hasler, T. A. McGee).
34. Wiggins Teape Ltd., GB 1 545 466, 1979 (P. R. Raine, L. Wescott).
35. Wiggins Teape Ltd., CA 1 156 725, 1981 (GB Priority, 1980) (= US 4 396 670, 1983).
36. Wiggins Teape Ltd., US 4 391 850, 1983 (GB Priority, 1980) (= CA 1 159 648, 1981).
37. R. Bodmeier, H. Chen: "Preparation of Biodegradable Poly(dl-lactide) Microparticles Using a Spray-Drying Technique," *J. Pharm. Pharmacol.* **40** (1988) no. 11, 754 – 757.
38. X. Guo, R. Bodmeier: "Polymeric Microparticles Prepared by Spray-drying a Drug-Containing Aqueous Colloidal Acrylic Polymer Dispersion, Eudragit RS 30D," *S.T.P. Pharma Sciences* **7** (1997) no. 6, 521 – 528.
39. L. R. Beck et al.: "A new long-acting injectable microcapsule system for the administration of progesterone," *Fertility and Sterility* **31** (1979) 545 – 551.
40. S. Benita, J. P. Benoit, F. Puisieux, C. Thies: "Characterization of drug-loaded poly(d,l-lactide) microspheres,"*J. Pharm. Sci.* **73** (1984) 1721 – 1724.
41. R. Bodmeier, H. Chen, P. Tyle, P. Jarosz: "Pseudoephedrine HCl microspheres formulated into an oral suspension dosage form," *J. Control. Rel.* **15** (1991) 65 – 77.
42. Glatt Maschinen und Apparatbau AG, Prattlen, Switzerland.
43. Coating Place Inc., Verona, Michigan, USA.
44. R. Gurny, N. A. Peppas, D. D. Harrington, G. S. Banker: "Development of biodegradable and injectable latices for controlled release of potent drugs," *Drug. Dev. Ind. Pharm.* **7** (1981) 1 – 25.
45. H. J. Krause, A. Schwarz, P. Rohdewald: "Polylactic acid nanoparticles, a colloidal drug delivery system for lipophilic drugs," *Int. J. Pharm.* **27** (1985) 145 – 155.
46. F. Koosha, R. H. Müller, S. S. Davis, M. C. Davies: "The surface chemical structure of poly(β-hydroxybutyrate) microparticles produced by solvent evaporation process," *J. Control. Rel.* **9** (1989) 149.
47. R. Bodmeier, H. Chen: "Indomethacin polymeric nanosuspensions prepared by microfluidization," *J. Control. Rel.* **12** (1990) 223 – 233.
48. F. Koosha, R. H. Müller, C. Washington: "Production of polyhydroxybutyrate (PHB) nanoparticles for drug targeting," *J. Pharm. Pharmacol.* **39** (1987) 136P.

49. B. Ekman, I. Sjöholm: "Incorporation of macromolecules in microparticles: Preparation and characteristics," *Biochem.* **15** (1976) 5115.
50. A. F. Yapel, Jr: "Albumin microspheres: Heat and chemical stabilization," in K. J. Widder, R. Green (eds.): *Methods in Enzymology: Part A, Drug and Enzyme Targeting,* Academic Press, Orlando 1985, pp. 3 – 18.
51. W. E. Longo, E. P. Goldberg: "Hydrophilic albumin microspheres," in K. J. Widder, R. Green (eds.): *Methods in Enzymology: Part A, Drug and Enzyme Targeting,* Academic Press, Orlando 1985, pp. 18 – 26.
52. E. Tomlinson, J. J. Burger: "Incorporation of water-soluble drugs in albumin microspheres," in K. J. Widder, R. Green (eds.): *Methods in Enzymology: Part A, Drug and Enzyme Targeting,* Academic Press, Orlando 1985, pp. 27 – 43.
53. L. L. Balassa, G. O. Fanger, *CRC Crit. Rev. Food Engineering,* **2** (1971) 245.
54. US 3 495 988, 1970 (L. L. Balassa).
55. International Flavours and Fragrances Inc., US 3 647 481, 1972 (M. V. Brodnitz, J. V. Pascale).
56. Capsulated Systems Inc., US 3 922 373, 1975 (R. G. Bayless).
57. Colgate-Palmolive Inc., US 3 896 033, 1975 (J. E. Grimm).
58. National Cash Register Co. Inc., US 3 401 123, 1968 (C. Brynko, J. A. Bakan). Purex Corporation, US 3 666 680, 1972 (B. R. Briggs).
59. Ciba-Geigy, US 3 846 324, 1974 (F. Lohman et al.).
60. Lever Bros., US 3 908 045, 1975 (D. S. Alterman, K. W. Chun).
61. Fuji Photo Film Co., US 3 703 576, 1972 (A. Kondo et al.).
62. Sterling Drug Co., US 3 488 418, 1970; 3 524 910, 1960 (W. M. Holliday, M. Berdick, S. A. Bell, G. C. Kiritis).
63. Bend Research Inc., US 4 670 250, 1987 (R. W. Baker).
64. D. Martin, *Chem. Britain* **25** (1989) 761.
65. *Controlled Release Systems – Technology and Markets,* CPS Research, Oxford 1989.

Microorganisms → Biotechnology

Milk and Dairy Products

ROLF JOST, Nestlé Product Technology Center, Konolfingen, Switzerland

1. Introduction 2
2. Historical Aspects 2
3. Composition of Milk 3
3.1. Milk Fat 4
3.1.1. Composition 4
3.1.2. Properties of Fat Globules 5
3.2. Milk Proteins 5
3.2.1. Major Nitrogen Fractions in Milk . 5
3.2.2. Nomenclature and Properties of Cow's Milk Proteins 5
3.2.3. Caseins 7
3.2.4. Whey Proteins 7
3.2.5. Nonprotein Nitrogen (NPN) 9
3.3. Lactose 9
3.4. Minerals 9
3.5. Vitamins 10
3.6. Contaminants 10
4. Physical and Physicochemical Properties of Milk 12
5. Microbiology of Milk 14
6. Nutritional Value of Milk 15
6.1. Milk as a Nutrient Source for Humans 15
6.2. Health Aspects of Milk Consumption 15
7. Dairy Technology 17
7.1. Milk Production and Transport . 17
7.2. Pretreatment in Dairies 18
7.3. Thermal Treatment 19
7.3.1. Thermal Processes 19
7.3.2. Effects of Heat on Milk 22
7.3.2.1. Effect on Microorganisms 22
7.3.2.2. Inactivation of Milk Enzymes . . . 23
7.3.2.3. Heat-Induced Destruction of Nutrients 24
7.4. Homogenization 26
7.5. Separation Processes 28
7.5.1. Microfiltration 28
7.5.2. Ultrafiltration 28
7.5.3. Nanofiltration and Reverse Osmosis 29
7.5.4. Ion Exchange Processes 29
7.5.5. Electrodialysis 30
7.6. Cleaning and Disinfection 30
7.7. Water Treatment 31
8. Fluid Milk 31
8.1. Consumer's Milk 31
8.1.1. Legal Classes of Consumer's Milk . 31
8.1.2. Characterization According to Heat Treatment 32
8.1.3. Modified Consumer's Milks 32
8.1.4. Packaging of Fluid Milks and Milk Products 33
8.2. Reconstituted and Recombined Milk 33
9. Milk Products 33
9.1. Flavored Milk and Milk Beverages 34
9.2. Evaporated (Condensed) Milk . . 34
9.2.1. Technology of Concentration by Evaporation 34
9.2.2. Unsweetened Condensed Milk . . . 35
9.2.3. Sweetened Condensed Milk 35
9.3. Dry Milk Products 36
9.3.1. Drying Processes 36
9.3.2. Whole and Skim Milk Powders . . . 39
9.3.3. Butter and Buttermilk Powders . . . 40
9.3.4. Whey Powders and Permeates . . . 41
9.3.5. Milk Protein Powders 41
9.3.6. Powdered Infant Formulas 43
9.4. Fermented Milk Products 44
9.4.1. Legal Definitions 45
9.4.2. Specific Fermented Products 45
9.5. Cream 48
9.5.1. Production 48
9.5.2. Types of Cream 48
9.6. Butter 49
9.6.1. Production 49
9.6.2. Steps in Buttermaking 50
9.7. Anhydrous Milk Fat and Butter Oil 52
9.8. Dairy Spreads 53
9.9. Dairy Desserts 53
10. Quality Control 54
11. Legislation 55
12. Economic Aspects 55
13. Acknowledgement 58
14. References 58

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_589.pub3

1. Introduction

Milk is the sole source of nutrients for all newborn mammals, including humans, in the first few months after birth. In dairy technology, the term milk refers exclusively to cow's milk that is collected as unaltered mixed milk from a large number of (usually several hundred) healthy cows. Milk from other sources must be labeled according to the species of the animal (e.g., goat's milk). Milk usually delivered at the dairy originates from small herds or from individual cows and shows a somewhat larger variation in properties and composition than mixed milk.

The milk is separated into two fractions (cream and skim milk) by centrifugation or spontaneous creaming. The *cream* is rich in fat, and the *skim milk* is practically fat-free. *Buttermilk* is the liquid remaining when butter is made from cream. The liquid that separates on acid curdling of milk is called *acid whey* and that obtained through rennet curdling is known as *rennet whey*.

A cow can be milked for approximately ten months after the birth of a calf. The composition and properties of the milk change during the course of this lactation period. The colostrum (colostral milk), obtained during the first few days after birth, contains high concentrations of antibodies that protect the newborn calf against infection. Colostrum is unstable when heated and consequently unsuitable for normal consumption.

The production of and trade in milk and dairy products are regulated by national laws and international agreements (Chap. 11).

2. Historical Aspects [1, 5]

The domestication of animals and the consumption of milk and simple milk products by humans probably began about 6000 to 8000 B.C. The Egyptians possessed dairy cattle even in 4000 B.C. and references have been made to milk production in ancient India (2000 B.C.). The milking of cows and the processing of milk, presumably a type of primitive buttermaking, are depicted in a frieze in a Sumerian temple located in the Euphrates valley and dating from 3100 B.C. The ancient Greek art of dairy farming was recorded by HOMER in the Iliad and Odyssey (ca. 900 B.C.) and by other authors such as HERODOTUS and ARISTOTLE. The Romans also practiced dairy farming, as mentioned in the writings of many authors (CATO, VERGIL, and STRABO). In addition to the consumption of milk and butter as food, these products were also used in ancient times for cosmetic and cultic purposes. Numerous references to the high esteem in which milk ("the land of milk and honey"), cattle, and cheese production were held can be found in the Old and the New Testaments. The first descriptions of the production of cheeses such as Glarner Schabzieger, Roquefort, Chester, and Emmentaler date from the 11th and 12th centuries. In the 13th century, MARCO POLO referred to a type of dried milk produced by the Tatars and the Mongols.

It was not until 1541 that K. GESSNER, a scientist from Zürich, wrote a book on dairy farming. Milk was examined scientifically for the first time a century later; the properties of lactose were described in 1633 (F. BARTOLETTI) and the presence of fat globules (A. VAN LEEUWENHOEK) and lactalbumin (TH. A. QUEVENNE) in milk were discovered in 1673 and 1841, respectively.

The Pilgrims of the "Mayflower" did not bring cattle with them when they settled in North America. Cattle were first introduced into the Plymouth Colony in 1624 leading to a great improvement in the settlers' nutrition [1].

The first methods for the conservation of milk were developed in the 19th century. Milk was first sterilized by N. F. APPERT in 1804; the evaporation of milk containing added sugar (NEWTON) was introduced in 1835; and the production of sweetened condensed milk by evaporation under vacuum conditions (G. BORDEN) was carried out in 1853. The first condensed milk factory was erected in the United States (Borden) in 1856, the milk separator was invented in 1859; the year 1866 was marked by the successful pasteurization of milk and by the erection of the first condensed milk factory in Europe (Anglo-Swiss Condensed Milk Co., Cham, Switzerland). The autoclave sterilization of condensed milk was reported in 1885 (J. B. MEYENBERG). The production of dried malt milk was achieved in 1887, the first synthetic resin (galalith from casein) was available in 1897, and the first homogenizer was invented in 1900. The use of a roll dryer for the production of milk powder was introduced in 1902 by J. A. JUST; spray-drying, invented in 1872, was also applied to milk at the beginning

Table 1. Composition of milk of various species * [11, 19, 21, 22]

Main Constituents, g/100 g whole milk	Cow's [11]	[19]	Human [11]	[19]	Goat [11]	[19]	Ewe [11]	[22]	Buffalo ** [11]	[21]
Protein (N × 6.38)	3.3	3.3	1.1	1.1	3.7	3.1	5.3	5.5	4.0	3.6
Fat	3.8	3.8	4.0	4.5	3.9	3.5	6.3	7.5	8.0	7.4
Carbohydrates	4.8	4.7	7.1	7.1	4.8	4.4	4.7	4.8	5.1	5.5
Organic acids	0.2	0.3								
Minerals (ash)	0.7	0.7	0.2	0.2	0.8		0.9	0.9	0.7	0.8
Sum of solids	12.8	12.7	12.5	12.9	13.2	11.0	17.1	18.7	17.8	17.2
Energy, kJ/100 g milk	280	290	290	310	290	250	400	400	450	440

* Energy content: 17 kJ/g for protein and for carbohydrate; 38 kJ/g for fat; 10 kJ/g for organic acids.
** *Bubalus bubalis* (water buffalo)

of the 20th century. The production of yogurt and processed cheese commenced in Germany in 1907 and in 1911, respectively (Gerber & Co., Thun, Switzerland).

The introduction of stainless steel gave a tremendous boost to the milk industry and resulted in an improvement in the quality of almost all milk products.

The production of sterile consumer's milk with a long shelf life (6 – 12 months) by continuous high-temperature short-time pasteurization followed by aseptic packing has a long history (Section 7.4):

- The first continuous process for sterile milk (USA, Germany) was introduced soon after 1890 [20]
- By 1914 the first patents were granted for the sterilization of milk by spraying with a hot jet of gas or steam [20]
- Direct heating of the milk by steam injection was introduced in 1927 (G. Grindrod, USA)
- Uperization (steam injection 150 °C, 2.4 s) was invented in 1952 [29]
- The Laguilharre process (injection of milk into steam) was introduced in 1954 [30]
- Indirect heating, ultrahigh temperature (UHT), was introduced in 1981 [13, 14, 24].

The technique of continuous heating required the development of methods for filling the product into commercial packs under sterile conditions. Examples of successful processes are:

- 1951, Martin Aseptic Canning System (USA) [31]
- 1961, packs made of coated cardboard (Tetra Pak, Sweden; Alpura Ltd., Switzerland) [32].

3. Composition of Milk [2, 5, 8, 9, 11, 18]

Whole milk is a colloidal system composed of a continuous aqueous phase in which lactose, whey proteins, and minerals are dissolved. The dispersed phase consists of casein micelles (see Section 3.2.3) and fat globules (see Section 3.1.2). Whole cow's milk contains typically 90 g/kg of solids non fat (SNF) and between 35 and 40 g/kg of fat. Both the fat globules and the casein micelles contribute to the whiteness of milk by light scattering.

Removal of the milk fat by centrifugation leads to skim milk on the one hand and to cream on the other.

Cream can be considered as a concentrated dispersion of fat globules.

Skimming of milk results in *skim milk* with $< 0.1\,\%$ of residual milk fat.

The term total solids (TS) refers to all dissolved or dispersed components of milk, it is typically 130 g/kg for pooled cow's milk.

The major constituents of bovine milk and milks from a few other species such as goat, ewe, or buffalo, which are economically significant, show some species dependent variation (Table 1).

Thus, milk from water buffalo is very rich in fat, while ewe's milk is both fat and protein rich, if compared with bovine milk. The composition of human milk is rather unique with respect to its low protein content. Its energy content is similar to that of cow's milk but more energy is derived from fat and lactose than in bovine milk.

There is also some variation in the composition of bulk milk of the same species throughout the milking season and also in dependence of the stage of lactation. The breed of the cow has a marked influence on the composition of its milk.

Table 2. Fatty acid composition of the milk fat of various species

			Bovine milk, % fatty acid in total milk fat			Milk of other species, % fatty acid in total milk fat			
Fatty acid	Symbol	*fp*, °C	Cow's [10]	Cow's [24]	Butter [25]	Human [26]	Goat [24]	Ewe [24]	Buffalo [24]
Butyric acid	C4:0	− 7.9	3.6	3.3	4	traces	2.6	4	3.6
Capronic acid	C6:0	− 3.9	2.3	1.6	2.4	traces	2.9	2.8	1.6
Caprylic acid	C8:0	16.3	1.3	1.3	1.3	traces	2.7	2.7	1.1
Capric acid	C10:0	31.3	2.7	3	2.9	2.2	8.4	9	1.9
Lauric acid	C12:0	44	3.3	3.6	3.6	4.4	3.3	5.4	2
Myristic acid	C14:0	54.4	10.7	9.5	11.2	6.3	10.3	11.8	8.7
Myristoleic acid	C14:1		1.4						
Pentadecanoic acid	C15:0	52.1	1.2						
Palmitic acid	C16:0	62.9	27.6	26.3	28.2	23	24.6	25.4	30.4
Pamitoleic acid	C16:1		2.6	2.3		3.8	2.2	3.4	3.4
Margaric acid	C17:0	61.3	0.9						
Stearic acid	C18:0	69.6	10.1	14.6	9.4	8.4	12.5	9	10.1
Oleic acid	C18:1	13.4	26	29.8	23.5	37	28.5	20	28.7
Linoleic acid	C18:2	− 5	2.5	2.4	2.1	12.9	2.2	2.1	2.5
Linolenic acid	C18:3		1.4	0.8	1.7	1.4		1.4	2.5
Total unsaturated fatty acids, %			33.9	35.3	27.3	55.1	32.9	26.9	37.1

The type of feeding (pasture or other) also influences the composition of bulk milk, especially the protein and fat contents.

3.1. Milk Fat

The milk fat in whole milk is present in form of globules, which are surrounded by a lipoprotein membrane. After its extraction from the fat globules, e.g., in butter production, essentially plain milk fat is obtained, containing some water in form of a water-in-oil emulsion.

3.1.1. Composition [23, 24]

Milk fat is composed to ca. 99 % of neutral lipids. Among these, triglycerides (triacyl glycerols) are the predominant group. Diglycerides (1 – 2 %) and very small amounts of monoglycerides (< 0.1 %), together with some sterols and free fatty acids, complement this fraction.

Complex milk lipids account for about 1 % of the total milk fat and are mainly composed of phospholipids in the following order of abundance:

phosphatidylcholine (lecithin) > phosphatidylethanolamine > other phospholipids

This fraction contains also trace amounts of ceramides and gangliosides.

The fatty acid composition of bovine milk fat is listed in Table 2, and compared with that of the milk fat of other species. Although the total number of different fatty acids in milk fat exceeds 200, only 15 fatty acids are present in important amounts. Bovine milk fat contains a relatively high proportion of saturated fatty acids, a characteristic shared by goat's, ewe's, and buffalo's milk, which contrasts with the relatively high level of unsaturated fatty acids in human milk fat.

The short-chain fatty acids ($C_4 - C_8$) of cow's milk are critical for its taste qualities. Release of short-chain fatty acids by lipase causes rancidity.

The fatty acid profile of milk fat varies depending on season and type of feeding. The degree of saturation, expressed by the iodine number (iodine number = grams of iodine bound per 100 g of fat), is 25 – 30 in European winter with hay feeding, while it is 35 – 40 in the summer season. Parallel with this goes a change in the melting properties of the milk fat. The more saturated winter fat melts at higher temperature than the less saturated summer fat. Above 35 °C, milk fat is completely liquid. At room temperature, the solid fat fraction amounts to about 20 % and this value increases at refrigerator temperature (4 – 6 °C) to about 50 % (see Section 9.6.2).

3.1.2. Properties of Fat Globules

Whole cow's milk with a fat content of 38 g/kg, contains approximately 2×10^9 fat globules per milliliter. The range of diameters of the globules (d_{VS}, volume to surface diameter) is about 2 – 5 µm, the mean value is near 3.4 µm.

Homogenization (see Section 7.4) reduces the mean fat globule size to diameters below 1 µm and increases the number of fat globules per unit volume by the third powder of the ratio of the two diameters, e.g.:

$$d_{\mathrm{initial}}/d_{\mathrm{homogenized}} = (3.4/1)^3 = 39.3$$

Thus, the increase in the number of fat globules per milliliter after homogenization will be 40-fold or more, for mean fat globule diameters below 1 µm. According to the temperature of the milk the fat inside the globules is solid, semisolid, or liquid. The state of crystallization of the milk fat inside the globules has considerable impact on cream- and buttermaking properties, it influences the stability of whipped creams as well as their stiffness.

The *milk fat globule membrane (MFGM)* is a true biological membrane of complex composition. Main constituents are polar lipids (phospholipids, cerebrosides, and gangliosides), lipoproteins, and glycoproteins [24]. Practically all of the milk cholesterol (ca. 130 mg per kilogram whole milk) is located in the MFGM. During skimming of milk, some membrane fragments remain in the defatted milk. The cholesterol content of skim milk is about 20 – 30 mg per kilogram skim milk.

Characteristic membrane proteins are, among others the enzyme xanthine oxidase (see Section 3.2.4) and the glycoprotein butyrophilin. The natural MFGM provides emulsion stability and protection against lipolysis of the fat. Homogenization leads to drastic changes in the membrane structure of fat globules. As the total fat area is enlarged, proteins from the bulk aqueous phase are adsorbed to the newly exposed fat surface and form a secondary membrane.

Fat globules can form clusters and produce aggregates, which lead to rapid creaming. A heat sensitive immunoglobulin fraction, IgM, is capable of agglutinating fat globules into clusters.

3.2. Milk Proteins

3.2.1. Major Nitrogen Fractions in Milk [27]

By convention, the total nitrogen content of milk (NT) is divided into two fractions, named according to Rowland [28]: i.e., casein nitrogen (CN) and noncasein nitrogen (NCN). The latter consists of two fractions, the serum protein nitrogen (SPN) and the nonprotein nitrogen (NPN).

The different nitrogen fractions of milk are established as follows:

NCN: = Nitrogen content determined according to Kjeldahl of the filtrate obtained by precipitation of the total casein at pH 4.6,

NPN: = Kjeldahl nitrogen of the filtrate obtained by precipitation of the total "true" proteins with trichloroacetic acid,

SPN: = Noncasein nitrogen minus nonprotein nitrogen (NCN − NPN)

A conversion factor of 6.38 is used to convert the nitrogen content measured into protein content.

The "true protein" content of milk is also defined as (NT − NPN) × 6.38. The value 6.38 assumes that the mean nitrogen content of (isolated) milk protein is 15.7 % which is close to reality, as shown by calculation of the nitrogen contents of milk proteins of known primary structure.

The casein fraction amounts to 75 – 80 % of the total milk protein with an average near 78 %. The serum (whey) protein fraction accounts for 15 – 18 % of the total protein and the NPN contributes to about 5 %.

3.2.2. Nomenclature and Properties of Cow's Milk Proteins [34]

Milk proteins are classified according to their primary structure, as presented in Table 3.

The major *caseins* are α-caseins, β-casein, and κ-casein. γ-Caseins are fragments of β-casein, resulting from a selective cleavage of this protein by plasmin (see Section 3.2.4). The *proteose peptones* 5, 8-slow, and 8-fast, correspond to the N-terminal fragments formed by plasmin catalyzed cleavage of β-casein. Proteose peptone 3, has a different origin, as it is derived from the MFGM. The most abundant whey protein of bovine milk, β-lactoglobulin, is absent

Table 3. Nomenclature, abundance, and properties of cow's milk proteins [34]

Protein species, common name [34]	Abbreviation [34]	Av. concentr. (g/kg milk)	Molecular mass	Genet.variant considered [a]	amino acid residue/mole	S – S bridges mole	perfree SH groups/mole	– PO_4 per mole	Isoionic pH	Glycoprotein yes/no
Caseins and Casein fragments										
αs1-Casein	αs1-CN	11	23614	B-8P	199	0	0	8	4.9	no
αs2-Casein [b]	αs2-CN	3	25230	A-11P	207	1	0	11	5.3	no
β-Casein	β-CN	9	23983	A2–5P	209	0	0	5	5.1	no
κ-Casein [c]	κ-CN	3	19007	B-1P	169	1	0	1	5.9	yes
γ1-Casein	β-CN-1P(f29–209)		20522		181	0	0	1		
γ2-Casein	β-CN-(f106–209)		11823		104	0	0	0		
γ3-Casein	β-CN-(f108–209)		11558		102	0	0	0		
Total γ-Caseins		0.8								
Proteose peptone component 5	β-CN-5P(f1–105/107)		12178	(f1–105)	105	0	0	5		
Proteose peptone comp. 8-slow	β-CN-1P(f29–105/107)		8717	(f29–105)	77	0	0	1		
Proteose peptone comp. 8-fast	β-CN-4P(f1–28)		3479		28	0	0	4		
Total proteose peptones 5 & 8		0.5								no
Proteose component 3	PP3	0.3		28000/18000 (2 components)					present	yes
Whey proteins										
α-Lactalbumin	α-LA	1.2	14175	B	123	4	0	0	4.5	minor variant
β-Lactoglobulin	β-LG	3.2	18277	B	162	2	1	0	5.2	minor variant
Serum albumin	SA	0.4	66267		582	17	1	0	5.1	no
Immunoglobulins of diff.classes [d]										
Immunoglobulin G1	IgG1	0.3 – 0.6	ca. 150000			4	0	0		yes
Immunoglobulin G2	IgG2		ca. 150000			4	0	0		yes
Immunoglobulin A	IgA									
Immunoglobulin M	IgM									
Total immunoglobulins		0.8								
Lactoferrin [e]	LF	< 0.2	80000		689	17	0	0		yes
MFG Membrane proteins [f]	MFGM	ca. 0.4								yes

[a] Calculated from the respective sequence data of the indicated genetic variant.
[b] This casein forms monomers and randomly disulfide bridged dimers.
[c] Disulfide bridges are also intermolecular and form S-S bonded oligomers of the proteins.
[d] Immunoglobulins are composed of H(heavy)-chains (M_r ca. 55000) and L(light)-chains (M_r ca. 25000). IgG1 and G2 are composed of 2 H-chains and 2 L-chains each. The 2 H-chains are linked via 2 disulfide bridges, 2 further disulfide bridges link each L-chain to a H-chain. IgA is a dimer of 2 H- and 2 L-chains, joined via one J-chain (16500). IgM is a pentamer of IgG and includes a central J-chain.
[e] Iron saturated lactoferrin carries 2 ferric ions per mole.
[f] MFGM Proteins are a heterogeneous group of glyco- and lipoproteins.

in human milk. Human milk contains as major whey proteins α-lactalbumin, lactoferrin, and immunoglobulin.

3.2.3. Caseins [35, 38]

The caseins are synthesized in the mammary gland, stressing their eminent role in the nutrition of the newborn mammal. α-Caseins (α-s_1- and α-s_2-casein) are the predominant casein species in bovine milk (see Table 3), while β-casein is the most abundant casein in bovine, goat's, mare's and human milk. A common feature of all caseins is the presence of one (in κ-casein) to several (α- and β-caseins) phosphoserine residues per mole.

Caseins in cow's milk are present in form of colloidal particles (*micelles*), with spherical diameters in the range of 20 – 400 nm. Micelles separated from milk by ultracentrifugation contain 7 – 8 % of minerals, in form of calcium phosphate and calcium-magnesium citrate complexes. The colloidal calcium phosphate complex of milk has a structure resembling that of hydroxyapatite. The micelles are strongly hydrated and bind about their four-fold weight of water.

This indicates that casein micelles must have a very open, sponge like structure.

Self-association of caseins via interaction of their hydrophobic regions is one of the key elements leading to micelles. Micelles are furthermore kept together by calcium phosphate bridges linking smaller casein aggregates (submicelles) to form larger aggregates (micelles).

Casein micelles lose coherence at low temperature (4 – 10 °C), due to the weakening of the hydrophobic protein-protein interactions. Under these conditions, a fraction of the casein is released into solution ("serum casein"), a process which is reversed if the milk is again warmed up. Micelles resist intense heat treatments such as UHT-heating or retort sterilization to a remarkable extent.

The lower pH limit for stability of the micelle is near pH 5.0. At a pH of 4.6, the total casein has precipitated and can be recovered as insoluble *acid casein*. The precipitated casein can be solubilized by addition of base; dissolution is achieved by raising the pH slowly to 6.8 or higher. This leads to the *caseinates* (see 3.2.3).

Rennet Clotting and Curd Formation [36]. A highly specific proteolytic process is the molecular basis of cheese making. Renneting milk with calf rennet or microbial "rennets", leads to the formation of a curd which is essentially composed of the milk casein (para-casein) and included fat and minerals (→ Cheese, Processed Cheese, and Whey).

3.2.4. Whey Proteins [35, 37, 38]

In contrast to the caseins the whey proteins are soluble proteins. The most abundant whey protein, *β-lactoglobulin* is present as a dimer in both, milk and in sweet whey. The dimer dissociates to monomers at pH < 3.5. In the pH range 3.7 – 6.5, the protein aggregates to form an octamer. This reaction is reversible. In contrast to the heat stable casein, whey proteins are readily denatured by heat treatments above 70 °C. As a result of such heat treatments, denatured whey proteins aggregate, coagulate or, depending on their concentration and environmental conditions such as pH and ionic strength, form protein gels. Although most of the whey proteins participate in such heat induced reactions, the role of β-lactoglobulin is dominant. Upon heating, the protein exposes its reactive thiol group (associated with cysteine [120]). The thiol group acts as a promoter of disulfide exchange and polymerization reactions, particularly at pH $>$ 6.7.

When whey proteins are denatured in milk, visible coagulation or gelling does not occur, as the denatured whey proteins interact with casein micelles to form a colloidal complex (see also Section 7.3.2). The formation of this complex has considerable technological importance in evaporated, shelf stable milks (see Section 9.2), in yogurt production (see Section 9.4), and is also the basis of coprecipitate production (see 3.2.4). While serum protein denaturation favorably influences the stability of evaporated milks and prevents syneresis in yogurt gels, it has a negative impact on the curd forming properties of cheese milks. Whey protein denaturation leads here to an increase in the rennet coagulation time, it lowers curd strength and reduces curd yield.

Table 4. Technologically relevant enzymes in milk [35, 39, 40]

Enzyme	EC classific	Concentration in raw milk, mg/L	Action in milk	Heat stability
Plasmin	3.4.21.7	0.1 – 1	proteolysis of caseins	survives direct UHT
Lipase	3.1.1.3	1 – 2	lipolysis of milk fat	*D* (72 °C) ca. 6 s
Alkaline phosphatase	3.1.3.1	300 – 500 mU/mL	hydrolysis of organic phosphoesters	*D* (72 °C) ca. 5 s
Lactoperoxidase	1.11.1.7	ca. 30	oxidation of thiocyanate	*D* (72 °C) ca. 100 s
Xanthine oxidase	1.2.3.2	ca. 35	conversion of xanthine into uric acid	partly inactivated in pasteurization
Catalase	1.11.1.6		decomposition of peroxide	survives pasteurization
Sulfhydryl oxidase	1.8.	ca. 10	oxidizes thiol groups to disulfides	partly inactivated in pasteurization

The whey protein quantitatively next important to β-lactoglobulin, α-*lactalbumin*, is devoid of a free thiol group while it contains four disulfide bonds which provide considerable structural stability.

α-Lactalbumin is produced in the endoplasmatic reticulum of the mammary gland and its function is to synthesize lactose in combination with the enzyme galactosyltransferase. α-Lactalbumin is found in the milk of all mammals. Native α-lactalbumin carries one calcium ion per mole. Although the protein has a somewhat lower denaturation temperature (60 – 61°C as compared with about 75 °C for β-lactoglobulin), it appears more heat stable as it resists heat-induced precipitation better than other whey proteins.

Milk *serum albumin* originates from blood and is also identical with blood serum albumin of the same species. Remarkable features of this protein are its primary structure stabilized by 17 disulfide bonds and folded into three major domains, as well as its capacity to act as a carrier of long-chain fatty acids.

Milk contains *immunoglobulins* of the *IgG, IgM* and *IgA* type; colostral milk being particularly rich in immunoglobulins. The biological function of immunoglobulins consists in a passive immunization of the calf via ingested colostrum.

From a milk processing point of view, the immunoglobulin fraction of early lactation milks reduces its heat stability; this is the reason why milks from early stages of lactation are excluded from production milks.

Milk Enzymes as Minor Proteins of the Whey Fraction [35]. The number of different endogenous enzymes in milk is very high ($>$ 50) but only some of them are technologically relevant. The most important of these enzymes and their catalytic activities are listed in Table 4. Most of these enzymes are found in whey, only a few of them remain micelle associated (plasmin and to some extent lipase), others are constituents of the MFGM (e.g., xanthine oxidase).

Milk plasmin is a protease derived from the fibrinolytic system of the blood and transferred from blood into the milk. Milk plasmin effects specific cleavage of β-casein to form γ-caseins and proteose peptones (components 5 and 8). Although β-casein is by far the most sensitive plasmin substrate, cleavage of other caseins, such as α-s_2-casein takes place as well. Plasmin induced proteolysis of milk casein may also play a role in cold stored milk. Further proteolysis of the caseins induces bitterness in the milk product via formation of bitter tasting, hydrophobic peptides. Plasmin occurs in freshly drawn milk to a large extent in form of its inactive precursor, *plasminogen*. Plasmin activity increases during storage of the milk by conversion of plasminogen into active plasmin.

Milk lipase (also called lipoprotein lipase), is important in milk processing. If milk is homogenized without prior pasteurization or severely sheared by pumping and other mechanical treatment, the milk fat undergoes rapid lipase induced lipolysis. Standard pasteurization will lead to a sufficient ($>$ 90 %) inactivation of the lipase to prevent lipolytic deterioration.

Lactoperoxidase (LPO), present in a concentration of ca. 30 mg/L in milk, has antibacterial activity through its ability to oxidize thiocyanate ions (SCN^-) into the antibacterial hypothiocyanite ion ($OSCN^-$). As lactoperoxidase is inactivated in the temperature range above 70 °C, its activity can be used to differentiate pasteur-

ized milk (LPO active) from "ultra-pasteurized" (LPO negative) milk (see Section 7.3.2).

Alkaline phosphatase is found after creaming of the raw milk in both, the skim milk phase, and in the cream. Alkaline phosphatase activity is the indicator for milk pasteurization see Section 7.3.2).

3.2.5. Nonprotein Nitrogen (NPN) [41]

About 5 % of the total nitrogen (NT) of bovine milk consist of NPN. In absolute terms NPN amounts to about 25 – 30 mg nitrogen per 100 g of milk. Most of the milk NPN derives from urea, while the minor part of it is composed of different amines, including small amounts of amino acids and oligopeptides.

3.3. Lactose [42]

The lactose content of bulk milk (ca. 48 g/kg) is rather invariable. The milk sugar represents roughly 50 % of the total dry solids of skimmed cow's milk.

Lactose (→ Lactose and Derivatives) is produced in the mammary gland from glucose by phosphorylation and epimerization to UDP-galactose. UDP-galactose is then condensed with another molecule of glucose to 4-*O*-β-D-galactopyranosyl-D-glucopyranose (lactose).

In dry milk, lactose occurs essentially in an amorphous state. As amorphous lactose is hygroscopic, milk powders, stored in a humid atmosphere, increase their water activity (a_w).

An increasing water activity induces partial crystallization of lactose and triggers the onset of Maillard type reactions leading to browning of the powder.

Hydrolysis and Fermentation of Milk Lactose [43]. Enzymatic hydrolysis of lactose by β-galactosidase (lactase) is one of the possibilities to produce lactose-reduced milk and milk products intended for lactose intolerant consumers.

Lactose is the substrate of lactic fermentation by lactic acid bacteria. These metabolize the sugar to lactate. The lactic acid concentration in milks fermented with typical yogurt culture is ca 1 %; corresponding to the fermentation of about 20 % of the initial lactose in milk (see Section 9.4).

3.4. Minerals [11, 42]

The mineral content of bovine milk is compared to that of human and buffalo milk (Table 5). Human milk has a relatively low mineral content in comparison with bovine, goat's, sheep, or buffalo milk. In all milks, potassium is the most abundant mineral, followed by calcium.

The variation in mineral content during the season is relatively moderate; higher than aver-

Table 5. Mineral content of bovine and human milk [11]

Minerals	Cow's	Human	Buffalo
Macrominerals, g/kg whole milk			
Calcium	1.20	0.31	1.95
Magnesium	0.12	0.04	
Potassium	1.57	0.53	1.00
Sodium	0.48	0.16	0.40
Phosphorus	0.92	0.15	
Inorganic phosphate			1.30
Chloride	1.02	0.40	0.62
Citric acid	2.10	0.85	1.50
Trace elements, mg/kg whole milk			
Manganese	0.02	0.01	0.05
Iron	0.46	0.29	1.50
Copper	0.10	0.35	0.25
Zinc	3.80	2.20	6.00
Aluminum	0.46		
Fluorine	0.17	0.17	
Iodine	0.11	0.06	
Selenium	0.09	0.03	

Table 6. Vitamin content of bovine and human milk [11, 52]

Vitamin	Cow's	Human	Buffalo	Recomm. intake for adults, mg/d [52]
Water soluble vitamins:				
in mg/100 g milk				
Vitamin C	1.7	4.4	2.5	55
Niacin	0.09	0.17	0.08	15
Panthothenic acid	0.35	0.21	0.37	10
Thiamine (B1)	0.04	0.02	0.05	2
Riboflavin (B2)	0.18	0.04	0.1	2
Pyridoxine (B6)	0.05	0.01	0.03	2
in µg/100 g milk				
Biotin	4	0.6	11	0.3
Folic acid	6	5		0.4
Cobalamin (Vit B 12)	0.4	0.05	0.3	0.004
Fat soluble vitamins:				
in µg/100 g milk				
Retinol (vitamin A)	30	54	64	1
Calciferol (vitamin D)	0.06	0.05		0.01
Tocopherol (vitamin E)	88	520		10

age chloride content may reveal udder disease (mastitis).

The high calcium content (31 mmol/L) is a biological requirement for the newborn mammal to enable build-up of bone mass. The rather limited solubility of the mineral is overcome by the existence of a colloidal complex of calcium with phosphate and citrate ions, in which negatively charged residues of the casein (phosphoserine residues and carboxyl groups) participate. Nearly 70 % of the milk calcium is present in this colloidal state, as well as about 45 % of the total milk phosphate.

Most of the *trace elements* in milk are bound to proteins or associated with MFGM components.

3.5. Vitamins [11, 42]

The role of milk as a source of vitamins is illustrated in Table 6. Riboflavin (vitamin B_2) and cobalamin (vitamin B_{12}) are particularly important, 1 L of milk covering 100 % of the recommended daily intake in these vitamins. The levels of other water-soluble vitamins such as thiamine (vitamin B_1), pyridoxine (vitamin B_6), niacin, and folic acid are much lower and the contribution of milk to the recommended daily intake is modest. The type of feed of the cows influences the milk content in fat-soluble vitamins such as retinol, calciferol, and α-tocopherol. Fat-soluble vitamins are virtually absent in skimmed milk, while the content of skimmed milk in water soluble vitamins is very similar to that of whole milk.

3.6. Contaminants [44]

Bacterial Enzymes and Metabolites. Bacterial proliferation in milk results in the contamination with (released or secreted) bacterial enzymes, of which the proteases and lipases are of critical importance. Both enzymes can seriously reduce shelf life of UHT milks or milk concentrates by development of bitterness and by causing so-called age gelling. The high resistance of bacterial proteases to UHT treatment leads to the gelling of long shelf life milk after a few weeks or months. Gelling is believed to result from partial proteolysis of κ-casein by the bacterial protease and subsequent destabilization of the casein micelle.

Cold storage of raw milk may end with very high viable counts of psychrotrophic bacteria such as *Pseudomonas* species. *Pseudomonas* species produce heat stable, extracellular proteases, even at refrigerated storage conditions. Inactivation of such proteases could be achieved at temperatures between 50 – 60 °C where an autodigestion process reduces the content in active protease [45].

Bacterial or Fungal Toxins. *Aflatoxins* find their way into the feed of dairy cows, e.g., via oilseed cakes. Ingested aflatoxin is metabolized to the likewise toxic hydroxy form *aflatoxin M1* (M stands for milk); aflatoxin M is being excreted in the milk. All aflatoxins are potent carcinogens and induce liver cancer, when ingested by test animals. Surveys carried out in milk and milk products revealed, in most cases, contamination levels in the order of < 0.5 µg/kg milk; the presently proposed tolerance for aflatoxin M1 in milk is 0.5 µg/kg.

Pesticides [44]. Pesticide residues detectable in milk originate from chemical treatment of weeds with herbicides, or by the treatment of skin parasites and insects on farms. Occasional sources of contamination may be oilseed cakes given as supplementary feed. Most countries restrict or forbid the use of *organochlorine pesticides* (e.g., aldrin, dieldrin, hexachlorocyclohexane, or DDT) in the direct neighborhood of dairy farms.

These compounds are not easily metabolized in vivo and, due to their high lipophilicity, accumulated in the fat cells of the adipose tissue. In the dairy cow, the ingested and stored organochlorine pesticides are secreted with the milk fat. Contamination levels of the order of 0.01 mg/kg of milk fat have been established in surveys in several countries.

These contamination levels do not represent any risk of toxic effects.

Organophosphates and carbamates (e.g., dichlorvos, parathion, or malathion) are less persistent and much more biodegradable. For this reason these compounds do not accumulate in the organism to a significant extent. Secretion into milk is therefore much less important than with organochlorine pesticides.

Veterinary Drugs and Hormones. Chemotherapeutics such as *antibiotics* (e.g., penicillins, tetracyclines, chloramphenicols, and streptomycins) and *sulfonamides*, as well as *parasiticides* are applied in the treatment of a wide spectrum of diseases of dairy cows. Udder infections are controlled by direct intra-mammary injection of antibiotics and sulfa drugs. Depending on the nature of the drug, withholding periods of up to 21 days are legally prescribed to protect production milks from contamination with such drugs. The presence of antibiotics creates a number of problems during further processing of the milk: it compromises the maturation of cheeses by inhibition of lactic and propionic acid bacteria, decreases acid and flavor production in fermented milks, and diminishes growth of starter cultures in yogurt, sour milk, or fermented buttermilk production. It cannot be excluded that consumption of some antibiotics via contaminated milk can lead to antibiotic resistance in the consumer. Antibiotics such as penicillins can elicit allergic reactions in sensitive persons.

Polypeptide Hormones. Treatment of dairy cows with recombinant *bovine somatotropin* (BST) is permitted in some countries with the objective to raise the milk output per cow. The hormone is mainly used in the USA, where its application in dairy cattle was approved in 1994. In the European dairy countries the use of this hormone remains banned, mainly for food safety considerations and animal health concerns [46]. The natural, endogenous polypeptide hormone occurs in bovine milk only in ppm concentrations. The natural hormone cannot be distinguished from the virtually identical recombinant hormone.

Heavy Metals. Heavy metal pollution due to industrial and traffic induced air pollution, has led to an increasing concentration of these toxic elements in animal feed and in food for humans.

In the past, most of the *lead* in milk originated from pasture pollution by automobile exhaust gases. The introduction of unleaded gasoline has been a decisive step to reduce this source of pollution. A risk of lead contamination comes from lead containing solder in cans used for evaporated milk.

Mercury contamination of milk is rare and usually the result of carry-over of mercury containing disinfection agents. Normal contents of lead, cadmium, or mercury in milk and milk products are extremely low and their reliable analytical determination needs specialized laboratory techniques.

Radioactive Elements. Fallout from nuclear weapon tests and atmospheric pollution by accidents in nuclear plants can lead to radioactive contamination of milk with radioisotopes, via the ingestion of polluted fodder by the

dairy cow. Radioisotopes resulting from such a nuclear pollution of the geosphere include ^{131}I, ^{134}Cs, ^{137}Cs, ^{90}Sr, and ^{239}Pb. Explosion of the Chernobyl reactor caused the emission of dangerous radioisotopes such as ^{131} I, ^{137}Cs, and ^{132}Te into the atmosphere. Besides extensive contamination of the local land, fallout of these isotopes in northern Europe and other regions caused contamination of milk and other foods with these isotopes.

Nitrate, Nitrite, and Nitrosamines. The endogenous nitrate content of milk varies between 20 µg/100 g and > 1 mg/100 g. Nitrate may be added to cheese milk with the purpose to inhibit gas formation in the maturating cheese, which is induced by the metabolic activity of *Clostridia* species. According to German and Dutch legislation, the addition of up to 15 g potassium nitrate per 100 kilogram of cheese milk is permitted.

In contrast to nitrate, *nitrite* is considered to be a contaminant, normally not present in milk. In raw milk, nitrite may be present in ppm amounts while for *nitrosamines* only traces are tolerable, if any.

4. Physical and Physicochemical Properties of Milk [7, 8, 47]

The physical properties of milk are of greatest importance in its analytical control and in its industrial processing. The most important of these properties are presented in Table 7.

The mean density of skim milk is 1035 kg/m^3, while that of whole milk is about 1030 kg/m^3.

The total solids content of any bulk milk can be approximated from the density and the fat content according to the Fleischmann formula [47]:

$$\mathrm{TS} = 1.2 \times \mathrm{fatcontent} + 266.5\,[(d-1)d]$$

Measuring the *freezing point* of milk it a routine procedure to detect dilution of milk with water or whey. An increase by 0.01 °C corresponds to an increase in the water content by 2 %. Lactic acid fermentation leads to a decrease in the freeing point, as the concentration of osmotically active species is increased (1 mol of lactose yields 4 mol of osmotically active lactate).

The viscosity of milk, at a given temperature, is determined by the viscosity of the outer phase (the milk serum with a viscosity very close to that of water) and the volume fraction occupied by both, casein micelles and fat globules. During concentration of skim milk the viscosity rises as a result of an increasing volume fraction ϕ occupied by the casein micelles and whey proteins. Such an increase in viscosity is higher after denaturation of the whey proteins (in the nonconcentrated milk).

The temperature dependence of milk viscosity at a fat content of 3.7 % is given by the empirical equation [47]:

$$\eta = 1.6 \times 10^{-7} \times \mathrm{e}^{(2741.5/t+273)} + 0.0002$$

The *specific heat* of skimmed milk is slightly higher than that of whole milk, due to the lower water content of a same volume of whole milk. This effect becomes important with creams of increasing fat content; 35 % fat cream has a much smaller heat capacity than whole milk (ca. 3000 Jkg^{-1} K^{-1} in comparison with 3900 Jkg^{-1} K^{-1}.

The *redox potential* of milk is mainly determined by its ascorbic acid (vitamin C) content, by the amount of dissolved oxygen, and by the concentration of reducing SH groups. The value for unheated milk is near 0.3 V. The value decreases for heated milk due to expulsion of oxygen and also due to the destruction of ascorbic acid.

Proliferation of bacteria leads to a strong decrease in the E_0 value, due to oxygen consumption and the activity of reducing enzymes.

The *specific conductivity* of milk is in the order of 4.5 – 5.0 × 10^{-3} S/cm. Milk with markedly higher values of conductivity indicate udder infection, as a result of increased chloride content of milks derived from cows with udder disease. The electrical conductivity of milk decreases with increasing fat content.

The *normal pH* of milk is 6.5 – 6.7. Raw milk with a higher pH than 6.8 can come from mastitic cows, or can be the result of contamination with detergents. A pH < 6.5 indicates beginning acidification; this is also indicated by higher °SH values (> 8 °SH). A direct relation between pH and titratable acidity expressed in °SH, is missing. Heat treatment of raw milk leads to a small

Table 7. Physical and physicochemical properties of bulk milk [7, 8, 47]

	Average	Range	Specifics
Density, kg/m^3			
Whole milk (3.5 % fat)	1030	1029 – 1031	measured at 20 °C
Skim milk	1035	1034 – 1037	
Freezing point, °C		− 0.51 to − 0.55	each increase by 0.01 °C indicates addition of 2 % water
Dynamic viscosity, mPa · s			
Whole milk (3.5 % fat)		1.4 – 1.7	measured at 20 °C
Skim milk		1.2 – 1.5	
Condensed milk (10 % fat)	ca. 25		
Cream (15 % fat)	ca. 10		
Electrical conductivity, mS/cm			measured at 25 °C
Whole milk	5	4.0 – 5.5	increased Na^+ and Cl^- concentrations (subclinical mastitis) lead to higher conductivities
Specific heat, J kg^{-1} K^{-1}			
Skim milk	4000	3980 – 4020	valid for temperature range 0 – 70 °C
Whole milk	3900		fat content 3 – 4.3 %
Thermal conductivity, W m^{-1} K^{-1} (at 20 °C)			
Skim milk		0.54 – 0.64	
Whole milk		0.52 – 0.6	
pH			
Normal whole milk		6.5 – 6.7	
Mastitic milk		$\geq$ 6.8	
Sour milks		4.6 – 4.9	
Yogurts		4.0 – 4.3	
Titratable acidity, °SH			
Fresh whole milk		6 – 8	titratable acidity mainly due to phosphate
Yogurt		38 – 52	44 SH correspond to a lactic acid content of 1 %
Redox potential, V			
Raw whole milk		0.25 – 0.35	main factors are ascorbic acid content, dissolved oxygen and reactive SH-groups
Past. whole milk			
Sterilized milk	− 0.05		intensive heat treatment destroys ascorbic acid, removes oxygen.
Yogurt	− 0.15		proliferation of bacteria reduces oxygen

pH increase (by release of CO_2) and, at higher heat loads, to a pH decrease (acid formation via Maillard reaction, formation of tricalcium phosphates and release of H^+ ions). Concentration by evaporation leads to a lowering of the pH.

Titratable acidity is measured by titration of milk to pH 8.3 with 0.25 mol/L sodium hydroxide solution. The Soxhlet-Henkel number (°SH) equals the milliliters of titrant needed to titrate milk to the end point of the indicator. 1 °SH corresponds to 0.0225 % lactic acid.

Color and Lightness. The optical appearance of milk, expressed in the CIE-LAB color system, is dominated by its lightness. The Hunter L-value measures the relation between reflected and absorbed light without consideration of a specific wavelength. Whole milk has a higher L-value (range 80 – 84, depending on fat content and heat treatment) than skim milk, due to the light scattering contribution of the fat globules. The value for skim milk depends on the heat treatment and ranges from ca. 75 to 78.

Heat denaturation of whey proteins leading to their aggregation, causes a small increase in lightness.

Severe heat treatments followed by advanced Maillard reaction reduce the lightness via formation of brown pigments.

5. Microbiology of Milk [49 – 51, 143]

Milk is a good growth medium for most microorganisms, desirable ones such as lactic bacteria used in the production of fermented milks, as well as undesirable ones (spoilage bacteria and pathogenic bacteria). Although milk contains usually < 1000 viable bacteria per milliliter on leaving the udder ducts, the temperature of the freshly drawn milk (37 °C), favors their propagation during the milking process. Contamination by exogenous microorganisms occurring during the milking procedure and during storage, adds to further increase in the total viable count of raw milk.

While *psychrotrophic bacteria* constitute an insignificant proportion of the milk flora in freshly drawn raw milk, these organisms may proliferate in milk during refrigerated storage. This is typical for bacteria of the genus *Pseudomonas*, *Flavobacterium*, and *Alcaligenes* species. Many of these bacteria produce extracellular enzymes such as proteases, lipases, and others, implicated in flavor defects or stability problems encountered with UHT milks.

Enterobacteriacea such as *Escherichia*, *Enterobacter*, *Citrobacter*, *Klebsiella*, *Proteus*, and *Serratia* species are major representatives of the spoilage flora of milk. High levels of *Enterobacteriacea* are occasionally associated with odor and flavor deterioration, slime formation or color development.

High viable counts (> 100 colony forming units, i.e., cfu/mL) of *coliform organisms* indicate serious contamination of the raw milk due to handling or unclean equipment.

Among the *spore forming bacteria*, *Bacillus* and *Clostridium* species produce heat resistant spores surviving pasteurizing treatments. *Bacillus cereus*, capable of growing at refrigeration temperatures, presents a particular problem in fresh or pasteurized milk products.

Clostridia such as *Clostridium tyrobutyricum* and others are a problem in hard cheeses, as the growth of this organism leads to production of gas and late cheese blowing.

Lactic bacteria of the genus *Streptococcus, Lactobacillus* and *Leuconostoc*, although suppressed in their growth at low temperature, are a problem in insufficiently refrigerated milk as they cause rapid souring.

The *total viable count,* expressed in colony forming units per milliliter of raw milk varies between several ten thousands and a million at the moment of the delivery to the dairy. Depending on the country, payment penalties due to insufficient bacteriological quality are applied at bacterial loads of $> 10^5$ cfu/mL [143].

Pathogenic bacteria [51] in milk originate either from the milk of infected cows or they are introduced by contamination of milking equipment and handling of milk or processing equipment. The most critical pathogens presenting a potential health hazard in raw milk are: *Mycobacterium tuberculosis*, *Brucella abortus*, *Coxiella burneti*, *Staphylococcus aureus*, *Streptococcus agalacticae*, *Salmonella* species, *Campylobacter*, and *Listeria monocytogenes.* The pathogenic flora is efficiently destroyed by pasteurization (see Section 7.3). Rigorous veterinary control of herds has limited the risk of infection of milk with *M. tuberculosis,* in developed countries.

Yeast and molds are of less importance in the spoilage of milk or cream, as they do not compete vigorously with the bacterial flora except in milk products with a low pH (such as fermented milks). They take also some importance in ripening of cheese where they may contaminate the surface of the cheese.

Milk *somatic cell count* (number of leukocytes) is used as a criterion of milk quality, particularly for the mastitis incidence in a dairy herd. The number of somatic cells in milk from healthy cows ranges between 10^5 and ca. 3×10^5 cells/mL. The rejection limit for milk with too high somatic cell counts is fixed at levels of $> 4 \times 10^5$ to $> 8 \times 10^5$ cells/mL, depending on the country [143].

Table 8. Milk as an essential nutrients source [52]

Nutrient	RDI [1] for woman >20 a	in 100 g milk (3.5 % fat)	500 g milk equiv. % RDI * [52]
Energy, kJ	9600	280	15
Protein, g	44	3.35	38
Linoleic acid, g	3	0.09	15
Minerals, mg			
Calcium	800	120	75
Magnesium	300	12	20
Phosphorus	800	92	58
Iron	18	0.05	1
Zinc	15	0.4	13
Vitamins			
Vitamin A (µg retinol)	800	30	19
Vitamin D (µg calciferol)	8	0.1	6
Vitamin E (mg tocopherol)	8	0.07	4
Niacin, mg	14	0.08	3
Riboflavin, mg	1.3	0.18	69
Thiamin, mg	1.1	0.04	18
Vitamin B 6, mg	2	0.05	13
Vitamin C, mg	60	1.5	13
Folate, mg	0.4	0.006	8
Vitamin B 12, µg	3	0.3	50

* RDI = recommended daily intake.

6. Nutritional Value of Milk [10, 11, 52, 53]

6.1. Milk as a Nutrient Source for Humans

Milk being the exclusive source of nutrients of all newborn mammals, has a sufficient and balanced content in essential nutrients required for the growing young mammal. After the weaning period, milk consumption makes up for a more or less big fraction of the total nutrient intake of a mammal, including humans. The adequacy of cow's milk to cover human nutritional needs is therefore a function of age. Cow's milk is not the adequate substitute of mother's milk for the human infant. An adaptation in several respects is necessary to produce a valid infant formula based on cow's milk (see Section 9.3.6).

For the growing child and for adults, milk and milk products remain a major source of essential nutrients (see Table 8).

Quality of Milk Protein [53 – 55]. Cow's milk protein is rich in essential amino acids (Table 9). Intensive heat processing of milk such as in the manufacture of evaporated milks or drying reduces the bioavailability of some essential amino acids of the milk protein. Lysine is particularly vulnerable as it undergoes a condensation reaction with lactose to form a nutritionally unavailable amino-1-desoxy derivative (Section 7.3.2.3). Other essential amino acids such as tryptophan, cysteine/cystine and methionine may also suffer damage by extensive heat processing or by drying of milk.

6.2. Health Aspects of Milk Consumption [10, 56]

Milk Fat [57, 58]. Milk fat is a good energy source due to its content in short- and middle-chain fatty acids, which are rapidly metabolized. Consumption of milk fat has, however been associated with development of arteriosclerosis in humans. The abundance of saturated fatty acids such as lauric acid (C_{12}), myristic acid (C_{14}), and palmitic acid (C_{16}) in milk fat is being considered the cause of its artherogenic effect. Whole milk, when added to a diet, was not found to raise total plasma cholesterol in humans, while the replacement of unsaturated vegetable fat by milk fat in test diets did so. This suggested that artherogenicity of milk fat may be little or not expressed when the intake is in form of milk, possibly due to "anti-cholesterogenic factors", of yet unknown nature.

Unsaturated fatty acid with trans configuration of the double bond physically resemble saturated fatty acids and saturated fat when in-

Table 9. Amino acid composition of the milk protein [53]

Amino acid, g/100 g protein	Total milk protein	Cow's milk whey protein	Total casein	Human mature milk
Aspartic acid	7.32	9.87	6.46	9.48
Threonine	4.23	6.47	4.22	4.99
Serine	5.51	5.05	5.47	5.19
Glutamic acid	21.17	17.19	19.97	16.96
Proline	8.7	6.03	10.48	7.95
Glycine	1.87	2.14	1.79	2.83
Alanine	3.3	4.61	2.78	4.38
Cystine	0.78	1.65	0.33	1.5
Valine	5.51	6.03	6.11	5.06
Methionine	2.39	1.91	2.53	1.8
Isoleucine	4.49	6.13	4.9	4.53
Leucine	9.08	9.53	8.62	9.77
Tyrosine	4.52	2.89	5.27	3.65
Phenylalanine	5.12	3.68	4.74	3.85
Lysine	7.42	8.87	7.36	7.63
Histidine	2.54	1.86	2.64	2.82
Arginine	3.12	2.42	3.39	4.29
Tryptophan	1.37	2.19	1.14	1.87
Ammonia	1.56	1.48	1.8	1.45
Sum all amino acids + NH_3	100	100	100	100
Tot N from amino acids + NH_3	15.72	15.72	15.72	16.34
calc. N-Protein conversion factor	6.36	6.36	6.36	6.12

corporated into triglycerides. Their assimilation leads to an increase in low density lipoprotein (LDL) cholesterol, whereas the high density lipoprotein (HDL) cholesterol is being decreased. Their consumption by hypercholesterolemic patients is therefore disadvised. Milk fat contains on average ca. 3 % of its total fatty acids in the trans form. The origin of the trans isomers is in the cow's rumen where unsaturated fatty acids are biohydrogenated, and subsequently desaturated, a reaction leading to *cis – trans* isomerization. In comparison to milk fat, margarine and other hydrogenated vegetable fats contain, following their hardening by catalytic hydrogenation, rather high concentrations of trans isomers [58].

Milk fat includes about 300 mg of cholesterol per 100 g fat. Although dietary cholesterol has little impact on the serum cholesterol levels of normal individuals, oxidized forms of cholesterol (oxysterols) generated from milk cholesterol during processing or storage, have been shown to be artherogenic in animal studies.

Lactose [43, 59]. Lactose intolerance results from intestinal malabsorption of the milk sugar and causes increased osmotic lumen pressure, abdominal pain, diarrhea and other symptoms. The intolerance is due to a deficiency in the mucosal enzyme β-galactosidase (lactase). Such a deficiency is very frequent in the black population of the USA. In Asian and Latin American more than 50 % of the population suffer from lactose intolerance. In European countries, the frequency of lactose intolerance is low but it a decrease in lactose tolerance occurs with progressing age.

However, many milk intolerant individuals tolerate fermented milks. This is probably due to the activity of ingested bacterial lactases coming from the starter organisms of fermented milk products. Aseptic addition of microbial lactase to pasteurized or UHT-sterilized milk, is a technological approach to reduce the lactose content of milk for lactose intolerant consumers. Such milks are on the European and American market since a few years.

Milk Protein. Milk protein consumption causes cow's milk allergy in 2 – 5 % of newborns [60 – 62]. Sensitization to cow's milk proteins occurs during early feeding the infant with milk formulas and other milk based products. It is likely to occur also when even minor concentrations of cow's milk proteins have found their way via blood circulation, into the breast milk of lactating mothers.

A hereditary disposition for cow's milk allergy has been clearly established. Cow's milk allergy has a transient character, as shown by

its much lower incidence in the adult population. The typical symptoms of cow's milk allergy include skin affections, intestinal affections such as colic, and respiratory symptoms like asthma. Formulas based on enzyme hydrolyzed milk protein have been found in many cases effective in dietary management of cow's milk allergic patients [62].

Milk protein, particularly the casein micelle, appears to play a positive role in *dental protection* as suggested by caries reducing effects of certain cheeses and lactose free milk protein.

Milk protein has a buffering effect and neutralizes, to some extent, acids generated by dental plaque bacteria. It also favors the mineralization of damaged enamel by providing high oral concentrations of soluble calcium. The protective effect of cheese may be related to both, its calcium providing or to the observed stimulation of saliva flow during cheese consumption [64].

Milk Calcium [67]. Milk is an ideal source of calcium and phosphorus for the build up of the skeletal mass of the young individual. Milk and dairy products are likewise a main calcium source in adult and in aging individuals.

7. Dairy Technology

Dairy technology is applied in the processing of raw milk to consumer milk and milk products. Technological processing of milk must guarantee its good hygienic quality and complete safety of the all milk products. It must provide dairy products which have attractive taste quality and texture properties and their full nutritional value.

7.1. Milk Production and Transport [1, 4, 14]

Milking. Today, *manual milking* is restricted to farms with a limited number of dairy cows. With larger herds, *mechanical milking* is performed with the aid of milking machines. The milk is extracted from the teat and udder by a pulsating vacuum in rubber-lined teat cups, fitted around the four teats.

The pulsator ensures physiologically adequate stimulation of milk secretion in the mammary gland by alternating application of vacuum (about 0.5 bar) and atmospheric pressure on the teat cups. The milk then flows through pipelines in a refrigerated storage tank. To avoid propagation of bacteria in the freshly drawn milk, cooling should be as rapid as possible, followed by cooled storage at temperatures ≤ 6 °C. For prolonged storage in cooled tanks, slow agitation is required to prevent the cream from rising.

Cows are usually milked twice a day. A typical daily milk yield per cow is about 20 kg, a mean annual milk production per cow ranges between 5 000 and 7 000 kg; the mean lactation period being about 300 days a year.

Milk Collection and Transport. Time laps between milking and milk collection and its transport to the dairy have become longer due to the growing size of dairy processing plants, collecting milk from a wide area. Mean transport times have also increased accordingly. Cooled storage at the dairy farm or in a collection center usually lasts 24 to 48 h before the milk is further processed in the dairy. In such a case a storage temperature below 6 °C is critical to reduce the growth of mesophilic and thermophilic bacteria. Transport of larger volumes of collected milk takes place in specially devised road tankers equipped with insulated milk tanks of 10 000 and more liters in size. The tankers may be equipped with cooling units to maintain low temperatures during longer transport in hot seasons.

Milk collected in alpine regions may be transported via milk pipelines made of polyethylene or poly(vinyl chloride) to local dairies located in the valleys.

Milk Reception in the Dairy. Control measures carried out upon milk reception in the dairy include the following:

Determination of specific gravity or the freezing point (see Chap. 4) to detect adulteration, e.g., by dilution with water; testing for the absence of antibiotics; rapid assessment of the bacteriological quality (methylene blue reduction test); and qualitative assessment of somatic cells by the so-called Schalm test. In addition to these parameters, the lactose, total protein, and fat contents are regularly monitored by automated infrared milk analysis.

Quality Control. Depending on the country, different criteria are included to determine the quality grade of milk of a particular supplier. These criteria include both hygienic and composition data. Among the hygienic parameters the total viable count and the somatic cell count, are the most important ones (see Chap. 5). Compositional criteria are the total protein content, the fat content, the lactose, and the total solids content.

7.2. Pretreatment in Dairies

Clarification. [4, 14]. Solid particles such as dirt particles, insects, or udder cells are removed from the milk by filtration or centrifugation, using continuous disk centrifuges rotating at 5000 – 6000 rpm with a flow of 10 – 30 m^3/h. The separated sludge accounts for less than 0.1 vol % of the treated milk.

Bactofugation and Bactofiltration. *Bactofugation* is a centrifugation process designed to remove microorganisms such as yeast, spores, and bacteria from milk. The process was introduced in the 1960s to improve the quality of cheese milk and the keeping properties of consumer milks. The most desired effect achieved with a bactofuge is the reduction in the spore content of production milks, as pasteurization does not inactivate bacterial spores and UHT-heating damages the curd forming properties of the milk. Optimum temperature for bacteria separation in a bactofuge is 60 – 70 °C; depending on the density difference between the milk and the microorganisms, a reduction of 1 to 2 orders of magnitude in their number can be achieved.

The *Bactotherm process* by Alfa-Laval, Sweden, uses a combination of bactofugation and UHT-heating (to treat the bacteria enriched bactofugate) and remix it, after heat sterilization, with the debacterized milk.

A more recent nonthermal process to reduce the bacteria and spores content of raw milk, is the *Bactocatch process* by Tetrapak, Sweden. The process is based on a cross-flow microfiltration of skim milk on inorganic membranes with a pore size of 1.4 µm. This microfiltration process can only process skim milk, thus fat separation has to precede and the cream must be separately heat treated. The aluminum oxide membrane is capable of reducing the bacterial population by three to six orders of magnitude, depending on the type of germ. Late cheese blowing does no more occur when cheese milks have been processed with this equipment [68].

Separation and Fat Standardization [4, 12 – 14]. The density difference between milk fat and skim milk is sufficient to allow efficient separation of milk fat in the form of cream. The reduced viscosity of the milk at 50 – 60 °C gives a sharper separation in the cream separator.

Continuously operating disk centrifuges as shown in Figure 1, are used in the industry.

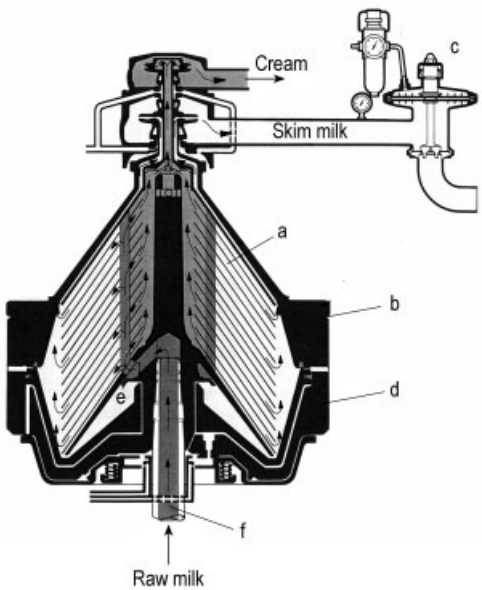

Figure 1. Continuous hermetic disk centrifuge for cream separation (reproduced from TetraPak Dairy Processing Handbook [4] with permission)
a) Disk stack; b) Lock ring; c) Counter-pressure valve; d) Bowl body; e) Distributor; f) Hollow bowl spindle

The whole milk is introduced into the separator via vertically aligned holes in the disks. Under the influence of the centrifugal force, the fat droplets settle inward the channels, close to the axis of rotation and the cream moves further along to an axial outlet. The skim milk moves outwards to the space outside the disk stack to a concentric skim milk outlet. The cream fraction usually represents about 10 % of the total throughput, accordingly a cream with 40 %

fat results from a milk having a fat content of 4 %. Depending on the desired fat content of the cream, the throttling valves in the cream and skim milk outlets are adjusted.

Fat standardization of production milks (for milk products of any kind) but also to some extent for consumer milks, is current practice in the dairy industry. To adjust the varying fat content of incoming milk to a specified fixed value, milk fat has to be added or removed. This adjustment can be done batchwise by mixing the appropriate volumes of skim milk and cream to achieve the desired fat content. In large dairies the standardization is often done automatically and in-line by means of continuous density monitoring of the fat content in the cream stream and computer controlled regulation of the flow of the streams of both skim milk and cream.

Degassing. Milk contains about 6 vol % air after milking. The amount of included air may increase to about 10 % due to handling and transportation. This leads to problems in measuring density and other physical parameters in the dairy. The same air will adversely affect the functioning of heat exchangers, centrifuges, and other equipment. Degassing of milk is achieved by either increasing its temperature and/or reducing the pressure. In most cases, deaeration is achieved in a vacuum tank in which the pressure is reduced to such an extent that the temperature of the milk drops by about 8 °C. This causes the evaporation of water, air and other volatile components are entrained.

Degassing is particularly important for good taste quality of butter. During UHT-heating, volatile substances are also removed together with water vapor in an expansion vessel (steam injection process). The vacuum expansion following the heat treatment in an UHT plant eliminates certain volatile components such as mercaptans from the milk.

7.3. Thermal Treatment

7.3.1. Thermal Processes [4, 12 – 14, 69]

(see also → Foods, 2. Food Technology, Chap. 4; → Heat Exchange)

Of all the steps used in milk processing, thermal treatment is by far the most important. The aims of heat treatment are the extension of shelf life, achievement of bacteriological and hygienic safety, the improvement of technological and functional properties of milk constituents and the development of desirable flavor properties. Thermal treatment destroys microorganisms and inactivates enzymes. It causes denaturation of whey proteins and accelerates chemical reactions such as the activation of SH groups; isomerization of lactose; Maillard reactions involving lactose and the ϵ-amino group of lysine; and others. The multiple benefits of heat treatment must be balanced with its negative impact on some heat sensitive nutrients in milk, such as certain vitamins. Heat treatments must therefore be carefully optimized and the best compromise found between safety and shelf life demands on the one hand, and minimum damage to nutritional quality on the other.

Thermization is a treatment of raw milk at temperatures below pasteurization temperature, aimed at reducing the number of bacteria during a cold storage of the milk prior to its final processing. Temperatures used for thermization range from 57 to 68 °C, and holding times are in the range of 10 – 15 s. Thermized milk must remain phosphatase positive.

Pasteurization [70] (→ Foods, 2. Food Technology, Section 4.1) has the aim to destroy all pathogenic microorganisms in milk, notably *Mycobacterium tuberculosis* and other pathogens such as *Brucella species, Salmonella* or *Escherichia coli* bacteria. While reducing the number of viable vegetative bacteria by several orders of magnitude, pasteurization does not reduce the content in viable spores. In view of the low intensity of a pasteurizing heat treatment, the nutritional quality of pasteurized milks is virtually unimpaired.

Depending on the equipment used in a dairy plant, different pasteurization conditions are applied.

Batch or *holding pasteurization*, the oldest method of pasteurization, is performed at 62 – 65 °C for 30 min. This process has been largely replaced by the continuous *high-temperature short-time (HTST)* procedure, where the milk is rapidly (within a few seconds) heated to temperatures ranging from 71 to 74 °C and held for

15 to 20 s at this temperature. Typical pasteurization conditions in European dairies are 72 °C and 15 – 20 s. The alkaline phosphatase activity of the heated milk is used to check pasteurization efficiency; adequate pasteurization results in a negative phosphatase test (see Section 7.3.2).

HTST pasteurization is performed with plate heat exchangers transferring the heat across the metal wall from the heating medium (hot water) to the product (milk) or, in the regeneration section, from the outflowing heated milk to the incoming, cool milk. Between the heating and the cooling section, a holding tube is inserted to provide the necessary holding time of the milk at the pasteurization temperature. The heated milk flows through the cooling section where it is cooled with ice water or brine.

For *cream* pasteurization a more intense heat treatment is required to achieve the same bacteriological safety as with pasteurized milk [71]. Pasteurization temperatures of 75 – 80 °C and holding times up to 30 s are adequate. To increase the shelf life of pasteurized cream, thermal treatments (85 – 90 °C and 10 – 20 s) are applied to reduce the number of more thermoresistant bacteria such as Streptococci, Micrococci, and Bacilli. This type of thermal treatment is also called high heating.

The expression *ultra pasteurization* is used to designate very short heat treatments (typically about < 2 s at 138 °C), mainly applied in the United States and Canada to produce consumer milk with an extended (refrigerated) shelf life (ESL milk), about twice that of standard pasteurized milk [33].

Ultrahigh Temperature (UHT) Heating (see also → Foods, 2. Food Technology).

The term UHT heating applies to a continuous process leading to a sterile product. The characteristic temperature range for UHT heating is 135 – 150 °C with holding times of a few seconds. The aim of an UHT process is to achieve a sterile product storable at ambient temperature. The heat treatment must destroy all microorganisms (spores of heat resistant bacilli included) and the product must be packaged aseptically into sterile, hermetically sealed packages. This is achieved in aseptic filling machines such as the Tetra Pak fillers.

UHT heating is performed according to a direct or an indirect process.

Direct UHT Process. In the direct heating process, the product is brought into contact with steam under pressure. The fast condensation of the injected steam heats up the product via transfer of the heat of vaporization of the steam. A consequence of this mixing of steam and milk is the dilution of the milk (by about 10 % or more) by the steam, which needs to be corrected at a later stage in the process.

Two main variants of direct heating plants exist: in the *injection system*, pressured steam is injected into the milk stream in a mixing chamber. An example of an injection plant ("uperization" plant) is shown in Figure 2. While preheating of the milk to the preheating temperature (typically 80 °C) takes a fraction of a minute, the heating time to the sterilization temperature (148 – 150 °C) is extremely short, in the order of a second. The milk is kept under adequate pressure in the holding tube by a restriction valve inserted between holding tube and the expansion vessel. This maintenance of pressure guarantees the appropriate holding temperature during the short holding time of 2 – 4 s. Once the heated milk has passed this valve and reached the expansion chamber, it boils and, at the same time rapidly cools down to near preheating temperature (actually 1 – 2 °C above the preheating temperature, see below). The cooled product is pumped through the homogenizer using an aseptically designed pump. A homogenization step after direct UHT heating by steam injection is required to disperse aggregates formed during the heat treatment. From the homogenizer the product passes through the regeneration section for further cooling and is pumped to an aseptic filling machine.

In the other version called *infusion system*, the milk is sprayed into a chamber filled with steam at the appropriate sterilization temperature. As the milk falls through the chamber, it condenses steam and equilibrates with the steam temperature.

In both variants of the direct UHT process, adequate removal of water during the vacuum expansion step is required, to correct for the dilution caused by the injected, condensed steam. This is achieved by evaporating water from the expansion vessel by adjustment of the temperature to ca. 2 °C above the preheating temperature at which the milk had entered the mixing chamber. The vacuum expansion allows, at the

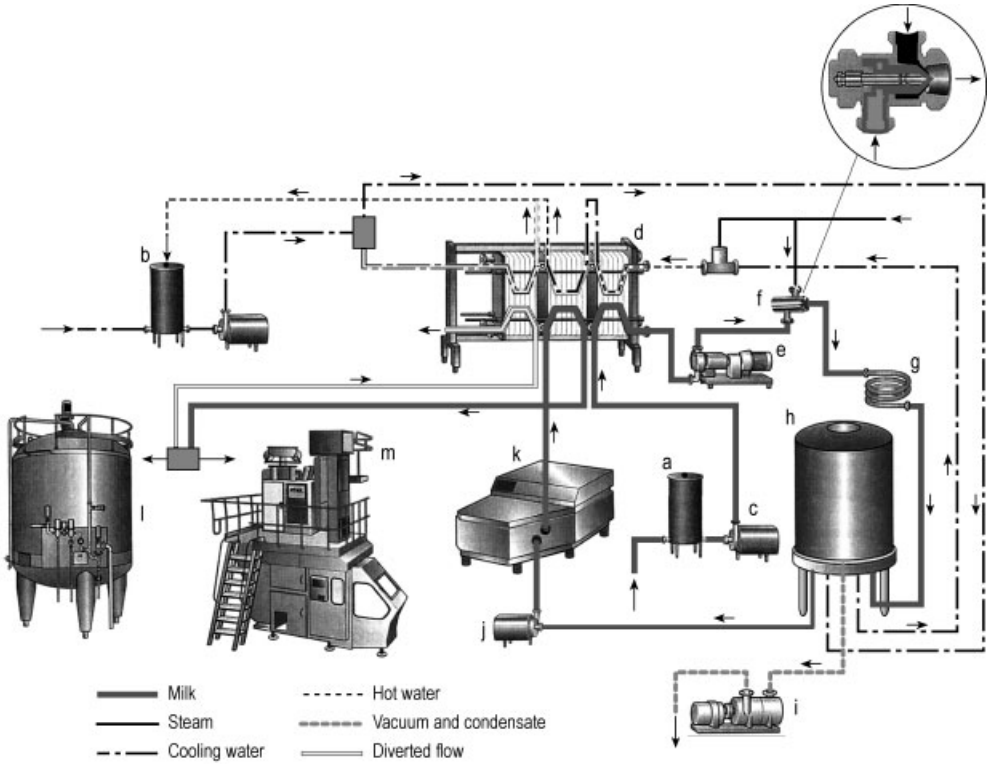

Figure 2. Direct UHT steam injection plant (reproduced from [4] with permission)
a) Balance tank milk; b) Balance tank water; c) Feed pump; d) Plate heat exchanger; e) Positive pump; f) Steam injection head; g) Holding coil; h) Expansion chamber; i) Vacuum pump; j) Centrifugal pump; k) Aseptic homogenizer; l) Aseptic tank; m) Aseptic filling

same time, to strip off undesirable volatile components from the milk.

Indirect UHT Process. Depending on the type of equipment, indirect UHT plants operate either with plate heat exchangers or with tubular heat exchangers. The common feature of all indirect plants is that the heated product does not get into contact with the heating medium (steam), but the two phases are separated by a thin (stainless steel) wall having a very high heat transfer capacity.

Plate heat exchangers present the highest possible heat exchange area on limited space. However for UHT processes, the plates must withstand higher temperatures and resist higher internal pressures. The plate exchanger systems are relatively prone to deposit formation and fouling.

Tubular heat exchangers are better suited to withstand high pressures than plate heat exchangers. Thus tubular exchangers are able to hold more easily the homogenization pressure (15 – 20 MPa).

Indirect UHT plants for UHT milk production often have a nonaseptic homogenization step in the septic part of the plant and end with an aseptic homogenization downstream of the UHT heater. In the case of milk, this second, aseptic homogenization may not be strictly required, as is the case with a direct heating process. Indirect UHT plants have the advantage of a better energy economy, their major disadvantage is fouling of the heat exchanger surfaces, resulting form deposits of heat denatured protein and minerals. Preheating the milk to 90 °C with a holding period of 1 – 2 min causes denaturation of the whey proteins and reduces subsequent deposit formation in the UHT section. A comparison of the time – temperature profiles of direct versus indirect UHT heating process reveals that the indirect process results in a higher

total heat load, due to the relatively long heating time [72].

Retort Sterilization in the Package. Sterilization in the package using an autoclave (or steam retort) is the oldest industrial sterilization technique. The product is being sterilized in its final package such as a metal can, a glass bottle, or a heat-stable plastic (e.g., polyethylene) container. The heat treatment is carried out in a sterilization chamber filled either with saturated steam or hot water under pressure. In batch autoclaves, the product in the package is accommodated in rotating baskets keeping the content of the cans, flasks, or other packages in a constant motion to promote uniform heat treatment of the whole product volume. In the case of a semisolid or solid product, the "cold spot" of the product defines the minimum heating conditions. Time – temperature profiles for retort sterilization of milk products are: 115 – 125 °C (product temperature) for about 10 min. In-package retorting is practiced for the production of evaporated milks, coffee creams, whipping creams, or milk based desserts.

For large industrial productions, continuously operating retorts are used, particularly for the manufacture of canned milk or milk in bottle sterilized milk.

Combined Sterilization Processes. Sterilization in the package is technically much simpler than UHT sterilization followed by aseptic filling. As UHT sterilization is much less product damaging than retorting, it can be an advantage to combine UHT sterilization with a short "post-sterilization" in a retort. Due to the high sterilizing effect of the UHT treatment, the post-sterilizing process operates with a much lower heat load than in normal in-can sterilization; its main task being limited to destroy microorganisms recontaminating the milk during nonaseptic filling.

In retort processes, the total lethal heat load is expressed by the F_0 *value*, a multiple of the sterilizing effect achieved by 1 min holding at a product temperature of 250 °F (121.1 °C) (→ Foods, 2. Food Technology, Section 4.2.1). For a complete retort sterilization process, the F_0 value must be extracted from the temperature – time profile of the operation (including heating and cooling sections). F_0 values between 5 and 7 are mostly applied for retorted milks and evaporated milks.

7.3.2. Effects of Heat on Milk [12, 13, 72, 73]

7.3.2.1. Effect on Microorganisms

The primary goal of heating milk and milk products is to guarantee their microbiological safety and to obtain a sufficient shelf life. This is achieved by a sufficient thermal reduction of viable bacteria and other microorganisms, or by their complete destruction. The inactivation of microorganisms follows a first order process:

$$-\ln(C_t/C_0) = k \cdot t$$

In this equation C_0 is the initial, C_t the final (after heat treatment) concentration of viable microorganisms, k is the temperature dependent death rate constant (expressed in s^{-1} or min^{-1}) and t the heating time, expressed in min. Transformation into

$$\log(C_0/C_t) = (k/2.3) \cdot t$$

leads to a definition of the *decimal reduction time* (D), which expresses the time (in s or min) at a particular temperature to reduce C_0 by a factor of ten. When the ratio of viable counts prior and after the heat treatment $C_0/C_t = 10$; as log 10 = 1, the time to achieve a tenfold (one log) reduction in the initial number of viable bacteria at the temperature ν, is D_ν.

The dependence of the D value on the temperature, for a given microorganism, is described by:

$$z = (v_1 - v_2)/(\log D_1 - \log D_2)$$

Accordingly, the z *value* indicates the temperature increase in °C required, to reduce the D_ν value by a factor of ten (→ Foods, 2. Food Technology, Section 4.2.1).

Thermal death data on bacteria and bacterial spores in milk are shown in Table 10. Time – temperature combinations of heat leading to comparable inactivation of mesophilic or thermophilic spores are apparent from Figure 3.

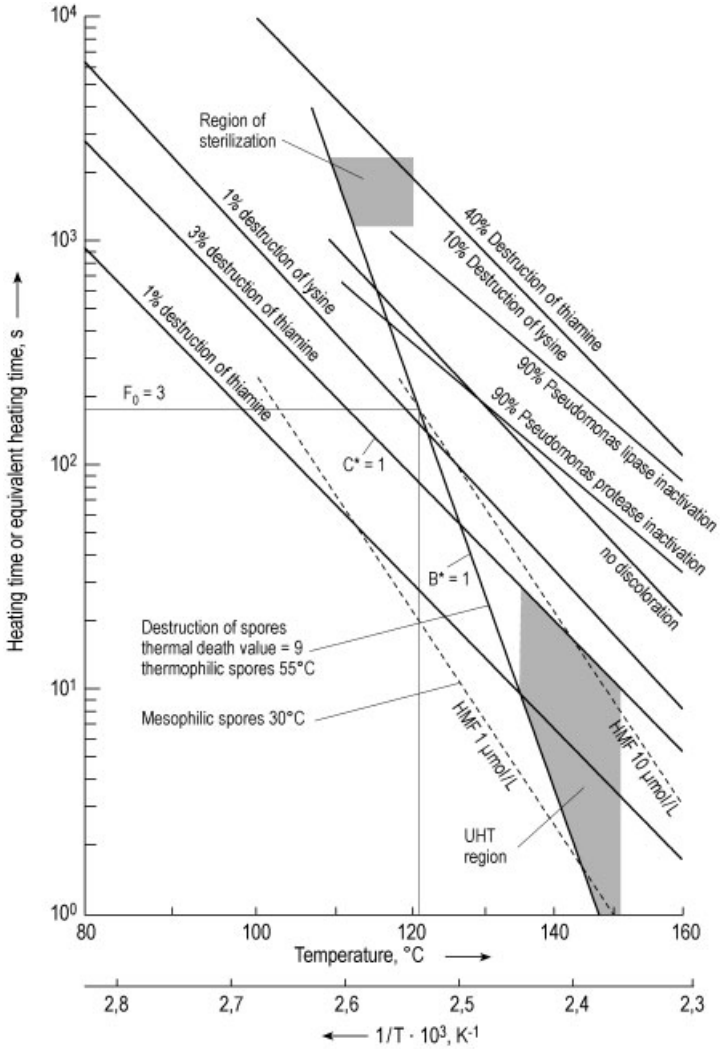

Figure 3. Effects of heat treatment on milk components (reproduced from [12] with permission)

7.3.2.2. Inactivation of Milk Enzymes [35, 39, 40, 70]

There are considerable differences in the heat sensitivity of important milk enzymes (see Section 3.2.4 and Table 4). Milk lipase and alkaline phosphatase are among the most heat labile enzymes and inactivated to a large extent during milk pasteurization. Other milk enzymes such as catalase, peroxidase, or acid phosphatase are much more resistant and survive pasteurization.

Milk *lipase* is sufficiently inactivated at an early stage of milk processing by standard milk pasteurization, to avoid lipolytic enzyme action during further processing (particularly homogenization). Homogenization greatly enhances the action of lipase on milk fat. Therefore, raw whole milk must never be homogenized without being pasteurized before. Pasteurization is not sufficient to destroy lipase activity from bacterial origins such as the lipases of *Pseudomonas* species and other, mainly psychrotrophic microorganisms (Fig. 3). High numbers of viable psychrotrophs may therefore lead to development of rancidity in UHT milks during their shelf life.

Alkaline phosphatase is inactivated to a high extent during standard milk pasteurization. The

p-nitrophenol method for the determination of residual phosphatase activity can easily detect inadequate pasteurization conditions or recontamination with small volumes of unpasteurized milk [76].

Table 10. Thermal death data on bacteria and spores in milk [69, 70, 73, 74]

Microorganism	*D*-value, s	*z*-value, °C
Endogenous milk flora		
Endog. mesophilic milk flora	9 – 12 (121 °C)	10 – 11
Endog. thermophilic milk flora	23 – 26 (121 °C)	10 – 11
Spores suspensions from:		
Bacillus stearothermophilus	180 (121 °C)	9.4
Bacillus subtilis	20 (121 °C)	6.7
Bacillus cereus	4 (121 °C)	36
Bacillus licheniformis	11 (121 °C)	8
Clostridium sporogenes	4 (121 °C)	11
Pathogenic vegetative germs:		
Mycobacterium tuberculosis	(60 – 70 °C) *	ca. 6
Salmonella enteriditis	12 – 31 (60 °C)	3.3
Listeria monocytogenes	35 – 72 (63 °C)	ca. 5

* Inactivation range of *M. tuberculosis*, no exact *D*-value available

While the *D* value of alkaline phosphatase at 70 °C is near 20 s, that of *lactoperoxidase* is >100 s; this explains that pasteurization has only a small effect on the activity of the latter enzyme. This provides a possibility to differentiate between pasteurized and UHT heated milk, as in the latter, the peroxidase is likewise inactivated. "Overpasteurization" can also lead to a negative peroxidase test.

Milk *plasmin* shows extreme heat resistance, active plasmin is found even in UHT heated milks. The residual plasmin activity in long shelf life milks is held responsible to some extent for taste deterioration (development of bitter taste) and gel formation.

Enzymes from the psychrotrophic microflora such as proteases from *Pseudomonas species* are equally highly heat resistant and limit the keeping quality of UHT milks by proteolysis and age gelling.

7.3.2.3. Heat-Induced Destruction of Nutrients [10, 12, 13, 69]

Vitamins. Losses in both water-soluble and fat-soluble vitamins are negligibly small in standard milk pasteurization. During UHT sterilization, small losses in thiamine, pyridoxine, folic acid, and vitamin B_{12} occur. Thus, losses in thiamine resulting from UHT sterilization range between 1 – 3 % of the initial content (Fig. 3). The oxidized form of vitamin C is destroyed during UHT heating of milk. Oxygen removal from milk to <1 ppm considerably reduces losses in vitamin C during UHT treatment. Losses in heat labile vitamins are much higher in retort sterilization of milk; e.g., for vitamin C loss may exceed 90 %, while for pyridoxine and folic acid, destruction amounts to about 50 % and for thiamine to about 30 %.

Damage to Proteins [77, 78]. As a result of intense heat treatment, several essential amino acids undergo irreversible chemical modifications which, if a significant fraction of the particular amino acid is involved, lead to a reduction in the nutritional value of the milk protein.

The ϵ-amino group of lysine undergoes a condensation reaction with lactose, when milk is heated, evaporated or dried. During the first and reversible stage of the *early Maillard reaction*, the ϵ-amino group condenses with a reducing aldehyde group of the lactose to form the protein bound Schiff base ϵ-lactosyllysine. The Schiff base then rearranges into ϵ-deoxylactulosyllysine. The latter represents an irreversibly blocked and nutritionally unavailable form of protein bound lysine.

In the further progress of *advanced Maillard reactions*, a variety of reactive decomposition products are formed. Among them, *5-hydroxymethylfurfural,* is a good indicator for progressing Maillard reaction in milk and milk products. Reactive intermediates of advanced Maillard reaction condense to form colored polymers (melanoidines). A great number of heterocyclic compounds with pronounced flavor characteristics are produced in the late stages of advanced Maillard reaction. These include *N*-, *S*-, and *O*-heterocyclic structures. The characteristic caramel aromas of severely heated milks are due to *O*-heterocyclic enolones [77].

Analytical determination of "blocked" lysine is based on the release of *furosine* (ϵ-*N*-(furoylmethyl)-L-lysine), and pyridosine from protein bound ϵ-deoxylactosyllysine, during total acid hydrolysis of the protein. Furosine, produced in about 40 % yield on lactosylated lysine, is determined by chromatographic procedures and allows calculation of the amount of blocked lysine in modified protein. UHT-sterilized milks show furosine contents in the range of 20 – 100

Table 11. Indicators for heat treatment in liquid milks [76, 79, 80]

	Lactulose, mg/L		Furosine, mg/L		Undenaturated β-lactoglobulin, mg/L	
Type of milk processing	Typical	Range	Typical	Range	Typical	Range
Raw Milk	0	0	0	0 – 5	3500	3000 – 3800
Pasteurization	5	0 – 10	6	2 – 10	2800	2600 – 3000
Draft EU Directive 92/46 [76]		non detectable				>2600
UHT						
Direct UHT-process	150	100 – 300	20	10 – 40	600	300 – 1000
Indirect UHT-process	300	200 – 1000	40	20 – 90	150	100 – 500
Draft EU Directive 92/46		>100				>50
Sterilization	1000	600 – 1200	100	75 – 150	0	0
Draft EU Directive 92/46 [76]		>600				< 50

mg/L. Retort sterilized milks give furosine contents of 100 mg/L and higher. A furosine content of 100 mg/L corresponds to the blockage of ca. 10 % of the total protein-bound lysine. Losses of >10 % of the total lysine significantly reduce the biological value of the protein.

The sulfur-containing amino acids *methionine* and *cysteine* can be partially desulfurated during intense heat treatments. Removal of sulfur from cysteine or methionine gives rise to volatile mercapto derivatives such as dimethylsulfide, dimethyldisulfide, methylmercaptan, or hydrogensulfide. These compounds are found in UHT milks produced in indirect heating plants in concentrations of up to 100 µg/L. The compounds are responsible for the characteristic "cooked" flavor of such milks. In the direct UHT-process, most of the volatile sulfur-derivatives are removed during the application of vacuum in the expansion step.

Formation of Intramolecular Cross-Links Involving Lysine [80, 81]. Phosphoserine residues in casein can undergo heat-induced elimination of the phosphate group in the β-position of the side chain, leaving a highly reactive dehydroalanine residue. The reactive double bond forms a cross-link with a proximate lysine ϵ-amino group. The corresponding lysine derivative, lysinoalanine, is the chemical trace of this reaction after total acid hydrolysis of the (cross-linked) proteins. The resulting lysinoalanine crosslink reduces the proteolytic digestibility of the protein. Significant amounts of lysinoalanine can be found in caseinates, probably as a consequence of temporary exposure of the casein to alkaline pH, during its production.

Lysine ϵ-amino groups can also undergo condensation with asparagine and glutamine residues to form isopeptide bonds with simultaneous release of ammonia from the amide group of asparagine or glutamine. This reaction is particularly favored by heating of milk proteins in the dry state or during the drying process in hot air.

Denaturation of Whey Proteins [12, 13, 83]. Heat treatment of milk above 65 °C causes the denaturation of whey proteins. Heat sensitivity of the major whey proteins decreases in the following order:

immunoglobulins > serum albumin > β-lactoglobulins > α-lactalbumin

Milk pasteurization leads to significant denaturation of immunoglobulin and serum albumin, while the effect on β-lactoglobulin and α-lactalbumin is negligibly small. As shown in Table 11, the content in undenatured β-lactoglobulin in pasteurized milks is ca. 3000 mg/L, or close to the initial content of the raw milk. In UHT milks, 60 – 90 % of the total β-lactoglobulin is denatured, while in retorted milk denaturation of all whey proteins is complete. Heat denaturation of β-lactoglobulin in whey leads to oligomerization and polymerization of the protein, denaturation in milk leads to complexation between β-lactoglobulin and casein micelles. In this complex, intermolecular disulfide bonds link denatured β-lactoglobulin to κ-casein of the micelle. Complex formation between the serum proteins and the casein micelle is pH dependent and reaches a maximum at pH 6.6 – 6.7. If the pH of the milk is raised prior to heat treatment, a major part of the κ-casein dissociates from the micelle and reacts with β-lactoglobulin in the serum phase. Such a reac-

tion reduces the heat stability of the milk, as a κ-casein depleted micelle is more sensitive to aggregation induced by heat treatment in presence of calcium ions.

Modification of Lactose. At temperatures >90 °C (e. g., during UHT or retort sterilization) lactose is epimerized) to lactulose (4-*O*-β-D-galactopyranosyl-D-fructose; → Lactose and Derivatives, Section 2.1), which is entirely absent in unheated milk. The measured content of lactulose represents a chemical indicator for the heat history of processed milk. Lactulose itself is not entirely stable but is converted, under the influence of heat, to galactose and formic acid.

The lactulose content of a sterilized milk allows to make an easy distinction between UHT sterilized and retorted milks, as well as between high-pasteurized and UHT milks (Table 11).

A lactulose concentration of 1 g/L milk corresponds to the isomerization of ca. 2 % of the total milk lactose. Lactulose can undergo Maillard reaction with lysine ϵ-amino groups; however, the reactivity of lactulose is lower than that of lactose.

Modification of Milk Fat [73]. Milk fat is relatively stable towards heating. There are no major chemical changes known to result from standard dairy heat processing of milk. On the other hand, some effects of heat treatment on the fat globules and their membrane are known. Heat coagulation of the "agglutinins" (fat globule aggregating immunoglobulins) interferes with cluster formation and decreases the velocity of cream separation. As a result of heat treatment, some of the membrane phospholipids may be transferred to the soluble phase of milk. Direct UHT heating of whole milk has a homogenizing effect, as the strong turbulence during steam injection followed by an exposition to cavitation forces during the expansion step, causes disruption of fat globules.

Changes in Mineral Equilibrium [73]. Milk is supersaturated in calcium and phosphate by virtue of the ability of the casein micelle to keep micelle-linked calcium phosphate in a colloidal state. In the milk serum calcium is mainly chelated by citrate ions, and, to a lesser extent, by phosphate ions.

When milk is heated, the pH is lowered due to lactose degradation and the formation of organic acids, such as formic acid. Tricalcium phosphate is formed and some of it may be precipitated from the milk. Calcium phosphate deposits on heat exchangers and in evaporators (fouling) are frequently occurring in dairy plants and are formed during the processing of milk as well as of whey. The deposits usually are not purely mineral but also contain significant amounts of protein. This reaction is responsible for the formation of calcium phosphate deposits during storage of UHT milks.

7.4. Homogenization [4, 12, 13, 85, 86]

Raw milk forms, upon standing, a cream layer within less than an hour. The speed by which the individual fat globules rise to the surface increases with their size and further depends on the density difference between the milk fat (0.93 g/mL) and the aqueous phase of the milk (1.034 g/mL), as well as on the viscosity of the milk itself. In quantitative terms, the creaming velocity v (cm/s), follows Stokes law:

$$v = [d^2(\varrho_d - \varrho_c)g]/18\eta$$

where ρ_d is the density of the dispersed phase (milk fat), ρ_c the density of the skim milk (continuous phase); η the dynamic viscosity of the milk and g the gravitational acceleration (9.81 m/s^2). As the diameter of the fat globules enter the equation in the second power, a reduction in fat globule size has a large effect on creaming velocity. The fat globule size in raw milk varies from about 1 to about 10 µm and application of Stokes law to such diameters leads to creaming times of several hours. Reason for the more rapid creaming of raw milk is the clustering of fat globules (see Section 3.1.2). To achieve sufficient creaming stability for pasteurized milks with a relatively short shelf life, reduction of fat globule size to a mean diameter of 1 – 2 µm is sufficient, while the long shelf life UHT milks require a much greater reduction in fat globule size (0.7 µm or less).

Another effect of homogenization is the large increase in exposed fat area, which greatly favors the action of lipases. For this reason, it is indispensable to first pasteurize the milk (in order to inactivate the lipase), prior to homogenization. The shear and cavitation forces leading to the

disruption of the milk fat globules also disrupt the native MFGM. The released membrane material will adsorb to newly exposed fat surface. However, the increase in uncovered fat surface is so large that the material derived from the MFGM is insufficient to cover it. Protein present in the continuous phase of the milk will adsorb to the newly exposed interface. This rapid adsorption of proteins from the continuous phase of the milk prevents a coalescence of the smaller fat globules after passage through the homogenizer's valve. In the homogenized milk, the fat globules are coated with a secondary membrane with a different protein composition than the original MFGM. This has practical importance in recombined milks in the production of which butterfat or vegetable fat (filled milks) is being homogenized with skim milk solids.

The relationship between the homogenization pressure and the fat globule size for fat contents of up to about 12 % is the following:

$$\log d_{\mathrm{v/s}} = \mathrm{const} - 0.6 \log P$$

$d_{\mathrm{v/s}}$ being the volume/surface related diameter and P the homogenization pressure. In the homogenization of milk, an increase of the homogenization temperature above 50 °C has little effect on homogenization efficiency, while at higher fat contents such as in creams, homogenization temperatures $\geq$ 65 °C lead to a better result.

Standard dairy homogenizers (Fig. 4) are operated at pressures of 5 – 30 MPa. Usually homogenization is carried out in two stages, a first stage at 15 – 20 MPa, followed by a second stage operated at lower pressure (typically 5 MPa). The principle of the homogenizer valve is shown in Figure 5. The dimension of the slit through which the milk is forced with pressure, is in the range 20 – 300 µm, depending on the type of homogenizer. A product velocity of 50 – 250 m/s is reached within the narrow slit of the homogenizer head, the average residence time in the slit is in the range of a few microseconds. In the aseptic design (see 7.4), the homogenizer's head can be exposed to pressurized steam to guarantee sterility of upstream UHT-heated milk.

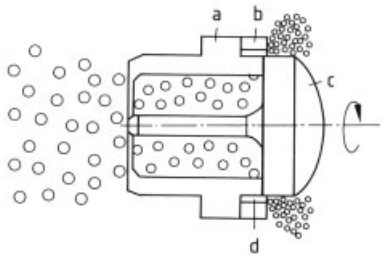

Figure 5. Homogenizer head (reproduced from [4] with permission)
a) Outer ring; b) Homogenizer ring; c) Core; d) Gap

In addition to prevent the creaming in milk and dairy products, homogenization has further desirable effects such as:

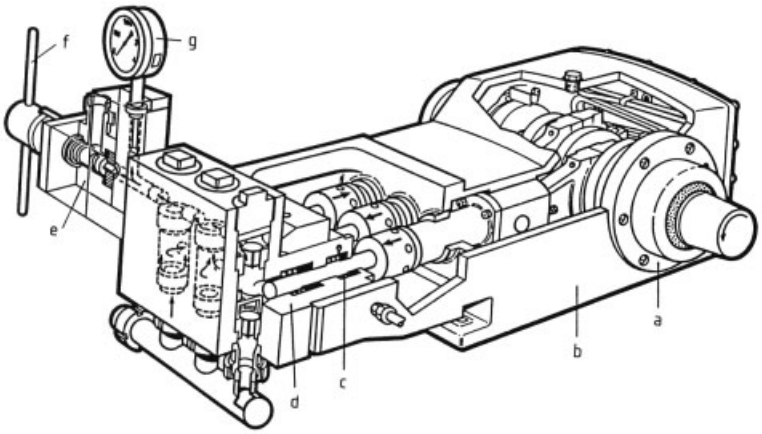

Figure 4. Milk homogenizer (reproduced from [4] with permission)
a) Crankshaft; b) High-pressure cylinder block; c) Piston rings; d) Pistons; e) Homogenizer head; f) Pressure setting lever; g) High-pressure manometer

- To increase the whiteness of milk and milk products by an increase in light scattering due to a higher number of fat globules;
- To improve the whitening effect of creams in coffee or tea;
- To improve the "mouthfeel" of milk products by increasing their smoothness and to reduce "grittiness" caused by heat-induced formation of aggregates.

Disadvantages are the increased sensitivity to lipolysis and an increased photosensitivity of the milk fat. Homogenization decreases to a certain extent the protein stability of milks, particularly evaporated milks, during heat treatment [87]. This is due to the formation of aggregates of protein-loaded fat globules during the heat treatment.

Homogenization of cream has the desirable effect to increase its viscosity, but the formation of clusters due to "sharing" of fat globule adsorbed protein, reduces the coffee stability of the homogenized cream. This feathering effect can be controlled by a second, aseptic homogenization of the UHT cream at a low pressure.

7.5. Separation Processes

Fractionation of milk components by either membrane- and/or ion exchange processes is used to produce "tailor-made" milk fractions for specific uses, or to replace conventional processes. Membrane technologies (→ Membranes and Membrane Separation Processes, Section 4.1) have a particularly high importance in the modern technology of milk fractionation.

7.5.1. Microfiltration [68]

Among membrane processes, microfiltration bridges between ultrafiltration and traditional dead-end filtration. The particle size range covered by microfiltration goes from about 0.1 µm to ca. 5 µm.

Thus, microfiltration, at its upper range, removes particles like fat globules, bacteria, bacterial spores and other, particulate milk components such as somatic cells or cell fragments. Such membranes produce ultra-clean milks with very low counts of viable germs or spores, with improved taste properties due to a minimum heat treatment, and with an increased shelf life when refrigerated (see Section 7.2).

Microfiltration membranes with lower pore sizes from 0.1 – 0.2 µm split skim milk into a casein-rich retentate and a "native" milk serum, devoid of casein. Casein micelles are sufficiently large to be rejected by the membrane, while the soluble serum proteins pass across it. The micelle fraction produced in this separation process has excellent cheese making properties, while the corresponding permeate fraction is a high quality source of native whey proteins. It can be further fractionated for isolation of specific whey proteins such as purified β-lactoglobulin, α-lactalbumin and others (see Section 9.3.4).

7.5.2. Ultrafiltration [88]

Ultrafiltration (UF) is applied to both skim milk and whey, with the objective to concentrate the respective protein components. Both polymer and mineral membranes are used for this purpose. Pore size of UF membranes are in the 10 – 100 nm range and provide cut-off values for spherical molecules between molecular masses of 10^3 to about 10^6.

Ultrafiltration of milk [89] is mainly applied to concentrate protein (and fat) to a level desirable for further processing into milk products such as fresh cheeses, milk powders, or yogurts. Lactose and soluble minerals and NPN are progressively removed with the permeate during this operation. UF is used to increase the protein content of cheese or yogurt milks for optimum product yield and texture improvement. Protein enrichment is useful to regulate the cheese yield throughout the seasons during which the protein content of the milk varies. In yogurt production, UF is an alternative to the addition of skim milk powders with the purpose to increase the protein and solids content in set style yogurts. As caseins and whey proteins are completely retained by the membrane during UF of milk, the mass ratio casein:whey protein is not being modified.

Ultrafiltration is applied to concentrate cheese milk to a level of solids (35 % or higher) at which curd formation proceeds without separation of serum. In this "pre-cheese process" retentate is directly transformed into a cheese

curd by addition of starter or rennet, or both. This process is advantageous for production of some fresh cheeses, as the draining of the serum can be avoided and the protein recovery is increased. The presence of whey proteins in the curd can however lead to texture defects or flavor problems in ripened cheeses. In a more recent version of the "pre-cheese process", the cheese milk is first heated and the casein coagulated, prior to its concentration on the membrane.

Ultrafiltration of Whey [111]. Since the 1960s UF has been used to recover the whey proteins in an essentially undenatured form from cheese whey or from casein whey (see Section 9.3.5). At a later stage, UF has become applicable to produce total milk protein concentrates (MPCs) from skimmed milk. In both processes, a protein-free permeate is produced which is mainly composed of lactose, minerals, and NPN. Such permeates can be used as a raw material for lactose production or, in most cases, are employed as animal feed.

7.5.3. Nanofiltration and Reverse Osmosis

Both processes use composite type polymer membranes such as composites of polysulfone or polyamide. The pore size of the nanofiltration membranes is in the range of a few nanometers, while reverse osmosis membranes have even lower pore sizes (0.1 – 1 nm).

Nanofiltration (NF) systems are operated at medium pressures in the range of 1 – 5 MPa. The membranes have selective permeability for minerals; i.e., high permeability for monovalent cations and anions and lower permeability for bivalent cations. Lactose is retained to a large extent, if not completely. These properties make NF interesting in pre-concentration of whey to about 20 % total solids. At the same time, a reduction of about 60 % of Na^+, K^+, and 30 % reduction in Ca^{2+} can be achieved. The process is often used to pre-concentrate the whey in a cheese factory to reduce transport volume for further dispatching to more or less distant whey processing plants.

Reverse osmosis (RO) membranes are only permeable to water but not to minerals, thus, RO is a pressure driven concentration process. The applied pressure has to compensate the osmotic pressure of the whey or milk solutes (mainly lactose and minerals) and to reverse the flux of water in the direction of the permeate fraction. Pressures of 5 – 10 MPa are required to achieve this (the osmotic pressure of a 30 % lactose solution is already near 4 MPa). In terms of energy, NF and RO are more economic than evaporation, as far as concentration to a certain level of solids (20 – 30 %) is concerned. In the concentration of whey, pre-concentration by RO is often combined with final concentration to >50 % solids on an evaporator.

7.5.4. Ion Exchange Processes (→ Ion Exchangers)

The application of ion exchangers in the food and dairy industry requires a high chemical stability of the resin, in order to avoid any migration of resin monomers or components into the food product.

Demineralization of Whey. Objective of the process is to reduce the mineral load of whey to the extent required for its use in human foods; incorporation of whey in various food products is, due to its high salt content problematic.

For the use of whey in the production of infant formulas, demineralization is mandatory in order to reach lower mineral levels compatible with an infant formula composition.

Whey demineralization often uses a combination of ion exchange and electrodialysis. The ion exchange process consists of two stages, a first cation exchange step, in which whey cations are exchanged against resin H^+ ions. In the second step the whey anions such as chloride, sulfate and phosphate, are exchanged for resin OH^-. The resin used is typically a sulfonated polystyrene resin:

Cation exchange step:

$$\mathrm{Resin{-}SO_3H{+}KCl} \rightarrow \mathrm{Resin{-}SO_3^-K^+{+}HCl}$$

Anion exchange step:

$$\mathrm{Resin-N(R_2)-OH{+}HCl} \rightarrow \mathrm{Resin{-}N^+(R_2)\,Cl^-{+}H_2O}$$

The demineralization efficiency of this process is high; it has to be accounted for some loss

of NPN and a very minor part of the true protein. Main drawback of the process is the need for large volumes of acids and bases for resin regeneration, connected with a high wastewater load.

Partial Demineralization of Milk. Ion exchange allows to obtain a partial replacement of calcium ions with sodium ions, with a simultaneous reduction of milk acidity. Calcium reduced milk has a higher heat stability. In order to split the calcium phosphate, a weakly acidic ion exchange resin is used which binds Ca^{2+} more strongly than Mg^{2+} or monovalent cations. Ion exchange processes can be applied to remove radioactive Cs or Sr ions in radioactively contaminated milk.

7.5.5. Electrodialysis (→ Membranes and Membrane Separation Processes, Chap. 3; → Membranes and Membrane Separation Processes, Section 3.7)

In electrodialysis (ED), low molecular mass ions migrate in an electrical field, across cationic and anionic membranes, the membranes being arranged in alternating manner in the ED stack, between the cathode and the anode. The membrane compartments are fed with brine and product to be demineralized, respectively, also in alternating manner. In the stack, anion exchange membranes repulse the migrating cations and let pass the anions; inversely, the cationic membranes are permeable for anions while they repulse the cations. The efficiency of demineralization of the feed is proportional to the intensity of the electrical field; the latter depends on a sufficiently high conductivity of the product stream. Main application of ED in the dairy industry is the demineralization of whey.

In order to increase the conductivity, whey is often preconcentrated to about 20 % total solids, prior to its demineralization by ED. The efficiency of demineralization in ED is a function of ion mobility and depends on the charge density of an ion. The efficiency is highest for K^+, Na^+, and Cl^-, while it is less efficient for Ca^{2+} and Mg^{2+} or phosphate ions. The ash content of skim milk (ca. 0.8 %) can be reduced by several passes of ED to less than 0.3 %.

7.6. Cleaning and Disinfection [4, 12 – 14] (→ Cleansing Agents, Chap. 2)

Cleanliness and perfect hygiene are required for consistent production of stable, high-quality products. Therefore, cleaning and disinfection are of fundamental importance for a dairy. After cleaning and disinfection, thorough rinsing with pure water is required to prevent contamination of the products and corrosion during holding.

Cleaning. The cleansing agent should swell proteins, emulsify fats, dissolve mineral residues, and make surfaces wettable for pure washing water without chemically or mechanically attacking them. Cleansing agents used in practice consist of a mixture of chemicals that dissolve proteins and milk stone which consists mainly of calcium phosphate (70 %); surfactants that wet and dissolve dirt particles and emulsify fat; corrosion inhibitors; and softeners that reduce water hardness. Large production plants are subjected to cyclic cleaning without disassembling (CIP = cleaning in place) which lends itself to automation.

Acidic cleansing agents are mainly effective against inorganic residues such as milk stone. Aqueous 0.5 – 5 % solutions of inorganic acids (nitric, phosphoric, and hydrochloric acids) and organic acids (citric, levulinic, gluconic, and tartaric acids) are used.

Alkaline cleansing agents are particularly effective in the swelling, dispersion, and dissolution of proteins; they also have a bactericidal effect. Silicates are added as corrosion inhibitors and polyphosphates function as emulsifiers and counteract water hardness. Alkaline cleansing agents contain sodium hydroxide, sodium carbonate, sodium ortho-, meta-, and polyphosphates, as well as sodium silicates in concentrations of 0.2 – 2 %.

Surface-active cleansing agents (alkyl sulfates, alkyl sulfonates, quaternary ammonium bases) improve permeative and wetting power, increase the dirt-suspending power, emulsify fat, and to some extent disinfect.

Chelating agents (ethylenediaminetetraacetic acid and nitrilotriacetic acid) prevent the precipitation of insoluble calcium and magnesium salts and reduce water hardness.

Table 12. Wastewater load from dairy processing [12]

Parameter	BOD_5, mgO_2/L	COD_5, mgO_2/L	TOC, mgC/L	Total solids
Milk (3.5 % fat)	114	183	70	12.5
Skim milk	90	147	62	9.2
Sweet whey	42	65	27	6.2
Acid whey	35	60	25	5.6
Cream (30 % fat)	400	750	186	38
Typical BOD load, kgO_2/t				
Milk processing (dairy)	0.5 – 3			
Milk drying	0.4 – 1			

Disinfection. Every cleaning procedure must be combined with a disinfection step to destroy all vegetative cells of bacteria, bacilli, yeasts, and molds. The most important disinfecting agents are:

1) Hypochlorite and chloramine: these solutions are strongly germicidal but also corrosive
2) Quaternary ammonium compounds and ampholyte salts
3) Hydrogen peroxide: primarily used for surface sterilization of packaging materials (15 – 30 % solution at temperatures up to 80 °C)
4) Hot water and superheated steam: temperatures of 120 – 130 °C are needed for complete sterilization

7.7. Water Treatment

Water Quality in Dairy Processing Plants. The water used in dairy plants must have drinking water quality; that means it is free of pathogenic bacteria, has < 300 total CFU/mL and no coliform germs in 100 mL.

Local water supply often provides hard water, unsuitable for dairy processing. For many membrane operations such as nanofiltration, ultrafiltration or microfiltration, partially or completely deionized water is being used.

"Milk water" is condensate formed during milk concentration on falling-film evaporators. The condensate from the first or eventually the second effect of evaporators is sufficiently pure, to be fed feed back into the steam generator's water reservoir.

Wastewater [14, 152 – 157]. A milk processing plant produces approximately twice as much wastewater as milk processed in the plant or even three to six times this quantity if the cooling water is also fed into the wastewater. Measures taken in the plant (e.g., the collection of washwater for animal feeding, repeated use of the butter washwater, and utilization of byproducts) can reduce the amount of wastewater by approximately 30 %. The BOD_5 and TOC values of some milk constituents and milk products are given in Table 12.

8. Fluid Milk

Fluid milks include plain milks from raw milk to heat treated consumer milk of different fat contents.

8.1. Consumer's Milk

8.1.1. Legal Classes of Consumer's Milk

The *European regulation* [90] has the following definitions for consumer milks. *Raw milk* is milk which was not been heated above 40 °C. *Whole milk* is milk that has been heat treated (pasteurized or sterilized). It can be adjusted in its fat content: *nonstandardized whole milk* has an unchanged fat content but not less than 3.5 %; *fat-adjusted whole milk* has a modified fat content but not below 3.5 % and the actual content is indicated on the package.

Partially skimmed milk is heated milk with an adjusted fat content which is ≤ 1.8 % and ≥ 1.5 %.

Skimmed milk is heated milk with a fat content of maximum 0.5 %.

In comparison with these European definitions, the minimum fat content of whole milk in the *USA* is 3.25 %; *reduced-fat milks* have fat contents of 2 %, *low-fat milks* of 1 %, while skim milk has ≤ 0.5 % fat and may be called "fat free".

8.1.2. Characterization According to Heat Treatment [91]

Pasteurized milks have received heat treatments sufficient to guarantee the inactivation of pathogenic microorganisms (see Section 7.3.1). Typically milk is continuously heated at 72 °C for 15 – 20 s.

The effect of pasteurization is controlled by the alkaline phosphatase activity test, which must be negative. Pasteurized milk must, on the other hand, give a positive peroxidase test.

The International Dairy Federation (IDF) [92] defined a routine method to determine alkaline phosphatase activity in milks. The colorimetric method can detect a contamination with as little as 0.5 % raw milks. A more sensitive fluorimetric method has been developed which can detect the contamination of pasteurized milk with as little as 0.01 % raw milk [93]. According to EC directive 92/46 [75], pasteurized milks, besides giving a negative phosphatase test do not contain measurable amounts of lactulose. It has also been suggested to use the content in undenatured β-lactoglobulin to fix the upper limit of adequate pasteurization, while the phosphatase test is targeted to define its lower limit. There is not yet agreement between different countries on the exact value for the lower limit. Pasteurized milk contains ca. 3000 mg of undenatured β-lactoglobulin per kilogram, the β-lactoglobulin content in unheated milk varies from 3300 – 3500 mg (Table 11).

UHT milks are sterile milks, which are packaged aseptically into sterilized, hermetically sealed containers. The UHT process is aimed to destroy all microorganisms in the product, which can be stored at room temperature.

The minimum lethal effect on heat resistant spores required is $F_0 = 3$ min. In most UHT processes, this is achieved with heating temperatures of 130 to 150 °C. The lower limit of this range is mainly exploited by indirect heating plants, while the higher end of the range is the area for direct heating by steam injection or infusion processes. The minimum lethal effect is reached at 135 °C in 7 – 8 s holding time, while at 148 °C, 0.4 s would be sufficient. In the practice, the applied heating conditions considerably exceed the minimum requirement of $F_0 = 3$ min; this has the objective to guarantee a sufficient shelf life under nonrefrigerated conditions.

The content in lactulose is the chemical indicator to differentiate between pasteurized and UHT milks. While pasteurized milks must not contain significant amounts of lactulose, directly sterilized UHT milks show lactulose contents of a few hundred milligrams per kilogram. The lactulose content of indirectly processed UHT milks is much higher (usually above 500 mg/kg). UHT heating results in extensive (>90 %) denaturation of the whey proteins; thus the content in undenatured β-lactoglobulin of UHT milks is usually below 100 mg/kg.

Despite the somewhat lower taste profile of UHT milks if compared with pasteurized milk, UHT milks have increased their market share at the costs of pasteurized milk during the last years (convenience driven preference of UHT milk).

In-Bottle Sterilized Milks. In several countries, retort sterilization of consumer's milk is current practice. This is mainly achieved in glass bottles, but alternatively sterilizable polyethylene packages can also be used.

Retorted milks differ from UHT milks by their even higher lactulose content ($\geq$ 1000 mg/kg). The retort process results in a somewhat caramelized taste, due to considerable Maillard reaction (see Section 7.3.2.3).

UHT sterilization may be combined with a poststerilization in the package, in this way the heat load in the final heat treatment can be significantly reduced compared with that of standard retort sterilization.

8.1.3. Modified Consumer's Milks

For the *European Community* EC directive 2597/97 [90] intends that the following modifications of milk composition are authorized for the production of special purpose consumer milks: milk may be enriched in milk protein, in minerals or vitamins; or milk may be reduced in their lactose content by (enzymatic) hydrolysis of the lactose.

The EC member states are, however, free to restrict or forbid such modifications of the milk composition in their proper countries.

In the *USA*, supplementation of consumer milks with vitamin D is optional practice, while

low fat consumer milks or fat free milks must be fortified with vitamin A.

In the last years, *lactose hydrolyzed* consumer milks have reached the market, intended for consumers that are lactose intolerant (see Section 6.2).

Since milk and milk products are being promoted as ideal dietary sources of calcium, dairies in all parts of the world have started to produce and market *calcium enriched* milks (pasteurized, UHT, or powdered) with total calcium contents of 1.8 – 2.0 g/kg, instead of the natural 1.2 – 1.3 g/kg.

8.1.4. Packaging of Fluid Milks and Milk Products [94, 95]

Packaging materials must be waterproof, have low oxygen permeability, provide sufficient light protection and be nonpenetrable by bacteria.

At present, laminated paperboard packages have the largest share among all packages (about 60 % in Western Europe). Plastic bottles, plastic cups, and glass are next important package materials. New trends in packaging of liquid dairy products are driven by cost factors and environmental considerations. Pouches could gain substantial market share for consumer milks due to their cost saving potential. For UHT milks and other shelf stable milk products, plastic based packages are especially suitable as these allow to create specific barrier properties and offer new possibilities for distinctive milk packaging concepts.

The requirement to reduce waste volume encourages the introduction of biodegradable packages, such as recently introduced for yogurts.

8.2. Reconstituted and Recombined Milk

Recombination is the production of consumer milks or milk products from milk powder (whole milk powder or skim milk powder) and a fat source, either anhydrous milk fat or vegetable fat.

If milk nonfat solids are recombined with vegetable fat, this is called *filled milk*.

Recombination has a high economic importance in nondairy countries with no or an insufficient supply in fresh milk (see Chap. 12).

The key raw materials for recombining are, besides whole milk powder, skim milk powder (SMP), and anhydrous milk fat (AMF). In the production of consumer milks by recombination, low heat or medium heat skim milk powders are the preferred raw material to achieve a good milk flavor of the recombined milk, resembling that of pasteurized fresh milk.

Two alternative procedures are used to produce recombined milk:

1) Anhydrous milk fat or butterfat is liquefied and added to an aqueous dispersion of skim milk solids, at about 45 °C. The mixture is homogenized at a temperature of 55 – 65 °C at 25 MPa and then subjected to a pasteurizing or an UHT heat treatment.
2) Anhydrous milk fat or butterfat is dispersed together with part of skim milk solids to give a cream with a fat content of about 30 %. The cream is homogenized at 20 MPa/5 MPa at 55 – 60 °C and then diluted to the desired SNF/fat ratio by mixing it with a skim milk dispersion.

Filled Milks. In order to save costs for expensive (imported) milk fat, and use local fat sources in its place, filled milks are produced and marketed in many nondairy countries. In order to provide cream flavor, buttermilk powder is often added during the recombination process.

Toned Milks. Toned milks are obtained by mixing whole fresh milk of local provenience with skim milk powder, to achieve a defined fat- and protein content in the final mix. An example is the production of mixed milks in India, from skim milk powder added to liquid high-fat buffalo's milk. This allows to obtain a more balanced protein : fat ratio.

9. Milk Products

According to EC directive 1898/87 [96], the designation "milk product" is, in addition to dried or evaporated milks, applicable to caseins, cheeses, cream, butter and buttermilk, butter oil, and anhydrous milk fat. Milk products are also fermented milks such as yogurt, kefir, or kumis.

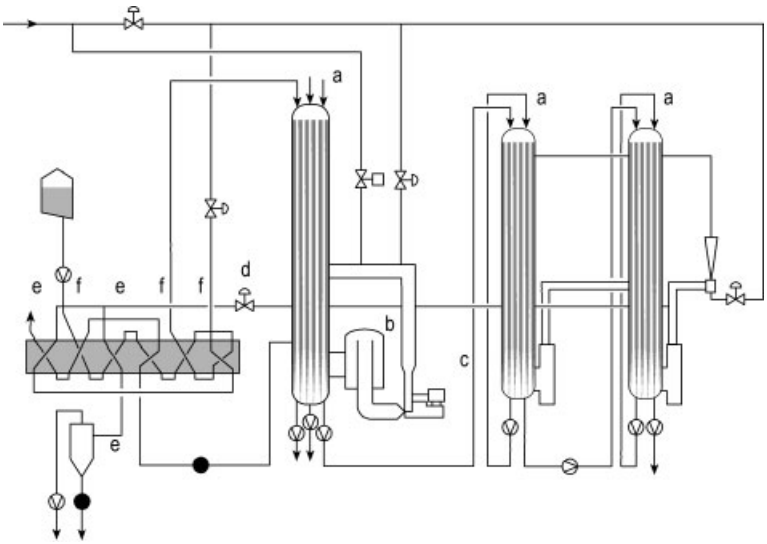

Figure 6. Multistage falling film evaporator (reproduced with permission of APV Nordic, Anhydro)
a) Calandria; b) Vapor separator; c) High-pressure fan; d) Plate heat exchanger; e) Condenser; f) Preheater

The delimitation between "milk products" and "milk containing other products", resulting from mixtures between milk, whey, and nonmilk food components is set at 30 % nonmilk solids.

9.1. Flavored Milk and Milk Beverages [14, 97]

Fluid milks may be combined with ingredients such as cocoa powder, chocolate, vanilla or fruit extracts and fruit juices. Milk beverages may contain thickeners and gelling agents (gums, carrageenans or chemically modified cellulose), in order to suspend insoluble or marginally soluble ingredients. They often also contain sucrose.

"*Chocolate milks*" are by far the most prominent among this category of products. A typical chocolate milk contains 1 – 3 % of cocoa powder, 3 – 6 % of sucrose and some hydrocolloid type stabilizer.

Cocoa particles will sediment in the milk beverage very rapidly, unless the suspension is stabilized by the addition of a thickening agent. In the case of pasteurized chocolate milks, the thickener is usually starch, gelatin or alginate, while for UHT chocolate milks, stabilization is based on the use of carrageenan (0.2 – 0.3 wt %). Retorted chocolate milks are also produced, in this case the preferred stabilizer is modified cellulose.

Fruit milks are combinations of milk with fruit extracts or fruit aromas. The content of all nonmilk ingredients is regulated by legislation. It must not exceed a certain value, e.g., 30 % according to German legislation.

9.2. Evaporated (Condensed) Milk [4, 98]

Concentration and drying of milk leads to products with a smaller volume and a longer shelf life. In the case of evaporated milks sold in cans or plastic cups, shelf stability is achieved by a sterilization step. In the case of sweetened condensed milk, product stability and shelf life is guaranteed by adding sugar to the extent where a sufficiently low a_w is reached. First attempts to extend the shelf life of milk by removal of water can be traced back to the early nineteenth century.

9.2.1. Technology of Concentration by Evaporation [4, 12 – 14]

Evaporation in the dairy industry is most often used in conjunction with spray drying. The most frequently used evaporator is a *falling film tubular evaporator* consisting of several stages ("ef-

fects"). The vapor consumption needed to evaporate a given volume of water from milk decreases with the number of effects, as the heat of the vapor from one effect can be reused in another effect, after compression of the steam in a thermal or mechanical steam compressor.

The scheme of a multiple-stage falling film evaporator is given in Figure 6. By applying vacuum to the individual effects, the boiling temperature can be kept low: between 70 and 40 °C; the highest temperature prevailing in the first effect. For the production of evaporated milks, the concentration factor is ≥ 2, as the milk is concentrated to $\geq 20\,\%$ solids non fat (evaporated skim milk) and to a minimum of 25 % solids in the case of evaporated whole milk.

9.2.2. Unsweetened Condensed Milk

The production of evaporated milk includes the stages shown in Figure 7.

The quality of the raw milk used for production of evaporated milk is critical, particularly with respect to the heat stability of the protein and the spore content which should be as low as possible.

The adjustment of the fat: SNF ratio involves the addition of cream or skim milk to attain the desired ratio in the final product. By preheating the standardized milk, bacterial growth during subsequent evaporation is reduced, and the required heat stability for later sterilization of the concentrate is achieved by denaturing the serum proteins at the level of the nonconcentrated milk. The optimum preheating conditions depend on milk quality and equipment factors. The milk is concentrated in a falling film evaporator. The concentrate is homogenized at pressures of 12 – 25 MPa, cooled and cold stored. At this point, a sterilization test with variable concentrations of stabilizer salt is carried out to evaluate the heat stability of the concentrate. Depending on the result of the test, an adequate amount of stabilizer salt (Na_2HPO_4 or another phosphate under E-no. 339–341) is added and the concentrate canned and retorted in a rotating autoclave or in a continuous retort. Typical retort conditions are 121 °C/10 min, corresponding to $F_0 \approx 10$ min. Test cans are stored for 2 to 3 weeks at 30 °C, others are incubated at temperatures up to 55 °C, in order to control sterility.

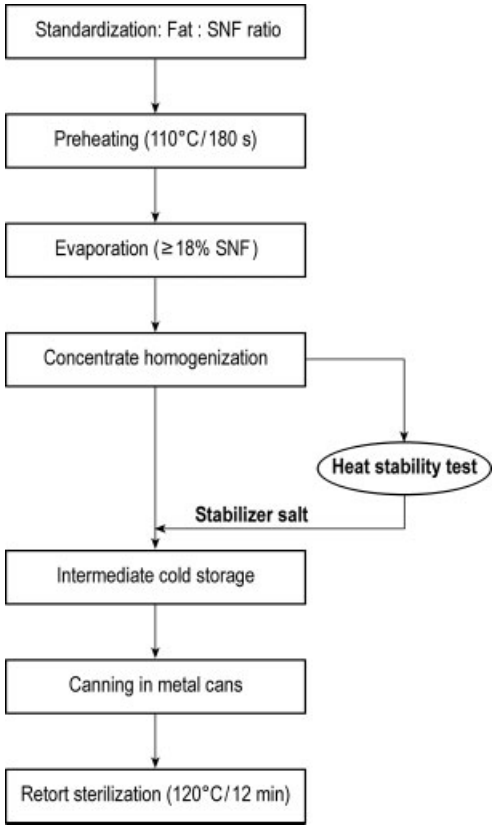

Figure 7. Flow sheet of production of evaporated milk

Evaporated milk is also UHT sterilized and aseptically filled in paperboard packages.

The legal definitions of unsweetened evaporated milks are shown in Table 13. EC and the Codex Alimentarius have harmonized their definitions of four categories of evaporated milks.

9.2.3. Sweetened Condensed Milk

Sweetened condensed milk (SCM) is essentially evaporated milk to which sugar has been added. It can be produced from whole milk, from skimmed, or from recombined milk.

In order to prevent bacterial growth in the product, the sugar concentration must be $\geq 62\,\%$ in the aqueous phase to assure an $a_w \leq 0.85$.

For a full cream product this corresponds to a sugar content of about 45 %, 11 – 12 % lactose derived from the milk concentrate adds to

Table 13. Legal regulations for evaporated milks and sweetened condensed milks

Product designation	EC [99]		Codex Alimentarius [104]		
	Fat content, wt %	Total solids from milk	Fat content, wt %	Total solids from milk	Milk solids non fat
Unsweetened evaporated milks					
Evaporated milk (whole)	min. 7.5	min. 25	min. 7.5	min. 25	
Evaporated skimmed milk	max. 1.0	min. 20	max. 1.0	min. 20	
Evaporated partly skimmed milk	$>1.0 < 7.5$	min. 20	$>1.0 < 7.5$	min. 20	
Evaporated high-fat milk	min. 15	min. 26.5	min. 15	min. 26.5	min. 11.5
Sweetened evaporated milks					
Sweetened condensed (whole) milk	min. 8	min. 28	min. 8	min. 28	
Sweetened condensed skimmed milk	max. 1.0	min. 24	max. 1.0	min. 24	
Sweetened condensed partly skimmed	$>1.0 < 8$	min. 24	$>1.0 < 8$	min. 24	min. 20
Sweetened condensed high-fat milk			min. 16		min. 14

In the Codex Alimentarius a minimum protein content of 34 %/TS is specific for both, unsweetened and sweetened condensed milks; the Codex Alimentarius also regulates eventual protein standardization.

this figure to give a total carbohydrate content of 56 – 57 %. The solids non fat content of a fat-reduced condensed milk has to be correspondingly increased, in order to reach a similarly high osmotic pressure in the aqueous phase as with the whole fat version.

The safe water : sucrose weight ratio is near 0.6 for all variants of SCM. Lower ratios, although being on the safe side, include an increased risk of lactose crystallization during shelf life.

Legal definitions of sweetened condensed milks are also given in Table 13 [104].

The required amount of sugar is added either in dry form to the standardized production milk, or as a concentrated syrup, during evaporation. The concentrate may be homogenized at a low homogenization pressure, if it appears desirable to regulate its viscosity properties. The final concentrate, after cooling, is supersaturated in lactose, at room temperature. Crystallization of lactose in form of large crystals (> about 10 µm), produces a sandy mouthfeel of the product. In order to avoid this, the concentrate is seeded at 25 – 30 °C with finely ground lactose and cooled under stirring to about 15 °C. After about 12 h at this temperature, crystallization of lactose in form of small crystals is usually completed and the product can be filled. Finished SCM has a yellowish color and a mayonnaise-like consistency. If the package is not opened and recontamination occurs, its shelf life is >1 a.

9.3. Dry Milk Products

Dry milk is a storable form of whole milk, skim milk, or milk fractions. From a world production of ca. 470×10^6 t/a of cow's milk (corresponding to 42×10^6 t/a of SNF), some 6×10^6 t of whole and skim milk powders are produced (corresponding to about 14 % of the total milk SNF world production). A major part of this milk powder is exported to nondairy countries and used for the production of milk products by recombination. Another important fraction is consumed as food ingredient by the food industry.

Further dried milk commodities include whey powders, buttermilk powders, and milk proteins.

9.3.1. Drying Processes [4, 12 – 14]

The objective of drying is to produce a shelf stable powder, with a remaining humidity near 3 %. This low water content provides a high chemical and microbiological stability.

The most commonly used processes to dry milk are spray drying (by far the most important one) and, for some special uses, roller drying. Freeze-drying is normally not used in dairy drying operations.

Pretreatment is quite different for roller and for spray drying. The kind of pretreatment also depends on the type of milk powder to be produced, i.e., whole milk powder, skim milk powder, or another milk product. The main

steps of pretreatment are preheating of the fat-standardized production milk and its concentration by evaporation. A third possible step is the homogenization of the liquid concentrate prior to its drying.

Roller Drying. In roller drying of milk, the concentrate is dried by contact with a steam heated metal drum, the temperature of which is 100 – 130 °C. The evaporated water is drawn off through a hood located above the drums (most versions of roller dryers have two drums rotating in opposite sense). Two main versions exist, the *trough-fed dryer*, depicted schematically in Figure 8 and the *spray-fed dryer* in which the milk is sprayed from a nozzle or a set of nozzles onto the drums. Factors influencing the speed of dehydration are the temperature of the roller surface, the thickness of the product film, and the rotation speed of the drums. Drying is usually complete within 2 – 3 s. Dried milk forms a layer on the drum which is continuously scraped off by blades.

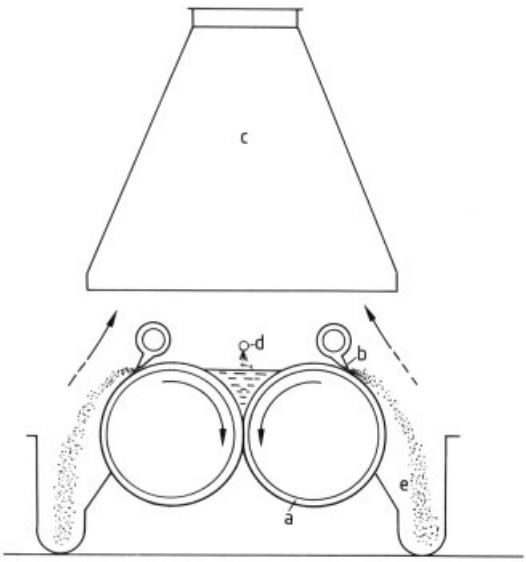

Figure 8. Principle of the trough-fed roll dryer (Just-Hatmaker, reproduced from [4] with permission)
a) Roll; b) Knife; c) Hood; d) Milk concentrate inlet; e) Dried milk

Characteristics of roller-dried milk are the high degree of protein denaturation and Maillard reaction taking place during drying. In the case of whole milk, roller drying also leads to a high proportion of free fat, due to damaging of the FGM. A high content in free fat (≥ 80 % of the total fat) is a desired property for milk powders used in milk chocolate manufacture. Although spray processes leading to high free fat contents have been recently developed, roller drying remains the most used drying procedure for chocolate powders.

Spray drying of milk or milk derivatives involves the following stages:

- Preconcentrated milk (45 – 55 % total solids) is pumped on the dryer and dispersed through a nozzle into a jet of fine droplets (atomization);
- The droplet spray is mixed with a hot air stream which leads to immediate evaporation of the droplets;
- The powder particles are separated from the air stream and recovered at the bottom of the dryer while the humid air is moved out of the drying chamber.
- In a two- or multistage dryer, the primary powder is conducted to an afterdryer in which final drying and eventually an agglomeration of powder particles is achieved on a vibrating fluidized bed.

Drying requires more energy than evaporation. In a multistage dryer, about 1.8 kg of vapor are needed to remove 1 kg of water from concentrate. Due to the less favorable energy balance of the drying process, concentration by evaporation is pushed as high as possible (actual limits for evaporation of whole milk powder are near 55 % solids, for skim milk powder near 50 % solids). Figure 9 shows the scheme of an industrial spray dryer with its main compartments.

A high-pressure pump transports the concentrate (ca. 50 % TS) on top of the drying chamber and forces it through the atomizer nozzle. Different types of nozzles exist: stationary ones and rotating disks or wheels. The hot air also enters the tower on the top. The atomization pressure determines the size of the droplets in the spray. High atomization pressures (>20 MPa) lead to smaller droplets which evaporate more easily but lead also to more powder fines (dust-like particles) which are not recovered in the powder but blown out with the outlet air.

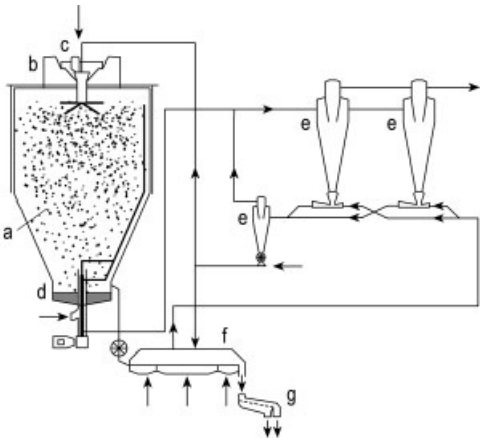

Figure 9. Spray dying tower for milk powder production (reproduced with permission of APV Nordic, Anhydro)
a) Drying chamber; b) Air inlet; c) Concentrate feed and central atomizing nozzle; d) Integrated fluidized bed; air broom for agglomeration; e) Cyclones; f) fluidized-bed afterdryer; g) Sifter (powder outlet)

Lower atomization pressures such as 5 – 15 MPa create larger droplets in the spray and reduce the drying efficiency but impose less shear stress to the concentrate.

The hot air temperature varies between 170 – 250 °C, depending on the type of dryer and the sensitivity of the product to be dried. The outlet air temperature is in most cases 60 – 80 °C. Residual moisture content of the primary powder is governed by the outlet air temperature, which influences also the maximum product temperature, of the powder during its travel through the drying chamber. When the powder arrives on the bottom of the drying chamber, its moisture content is usually between 5 – 7 %. In the APV/Anhydro type of dryer shown in Figure 10 an integrated fluidized bed of particles is formed on the bottom of the chamber on which the powder particles can be agglomerated. The agglomerated (instantized) particles are discharged at the end of the chamber.

Final drying takes place in the afterdryer at an air temperature markedly below that of the main drying chamber (typically 130 – 150 °C). This second stage of drying provides milder conditions for the removal of the remaining more firmly bound water. Here, the powder reaches its final moisture of about 3 %.

Fines removed by the outlet air flow are captured in air-solids separators or cyclones, collected and re-injected into the drying chamber at an appropriate location. In this type of dryer, the fines are moved with an air broom upwards to the wet zone of the drying chamber. By collision with droplets, the dust particles can be condensed with larger powder particles.

Major constructors of industrial dryers are APV/Anhydro AS (Denmark), Niro AS (Denmark) and Laguilharre *SA* (France). The Nestlé "Egron" dryer is operated by the company since the late 1930s in milk powder production plants.

Instantizing designates a drying process leading to "instantly" dispersible powder. Instant solubility is achieved by changing the particle structure (size of the particle, surface, pore structure, and fat binding). Two basic instantizing processes are used: in the *straight-through process* an agglomeration of the powder particles is achieved by collision of droplets and powder particles in different stages of drying. Such collision leads to a sticking together of individual particles to form agglomerates.

Thus, agglomeration is done in the drying chamber of the tower by means of controlled humidity and the collision of thermoplastic particles of different moisture content, followed by afterdrying.

In the *rewet process* powder agglomeration is done by rewetting powder particles with a water spray or with steam and agglomerate the wetted particles on a fluidized bed in an afterdryer (Fig. 10). Thus in the rewet process, agglomeration takes place outside the drying tower, in the afterdryer.

Lecithination improves the wettability of fat-rich powders such as whole milk powders or high fat powders in water. In lecithination a solution of lecithin in butterfat is sprayed against the spray of concentrate in the drying chamber. This leads to formation of a fine surface layer of lecithin on the particle surface and promotes rapid wetting of the powder particles in water.

EU-directive 76/118 [100] allows the use of lecithin (E 322) as a processing aid for milk powder production. The use of lecithin, limited by Good Manufacturing Practice (GMP) is also regulated by the Codex Alimentarius.

Characteristics of Milk Powder [101]. The *maximum moisture content* of milk powders is subject of regulation and a major quality criterion (see Section 9.3.2). Whole milk powder

Table 14. ADMI Specifications for quality of milk powders [106]

	Whole milk powder		Skim milk powder	
Parameter/grade	Standard	Extra	Standard	Extra
Maximum moisture content, %	4	4	5	4
Maximum milkfat content, %			1.5	1.25
Minimum milkfat content, %	26	26		
Maximum titrable acidity, %	0.17	0.15		0.15
Maximum total viable plate count, CFU/g	100000	50000	100000	50000
Scorched particles (disk) *	B (15 mg)	C (22.5 mg)	B	B or better
Maximum insolubility index, mL	1	0.5		1.25

* B, C are different levels (amounts) of scorched particles collecting on the filter disk, compared to an ADMI standard. The values in parenthesis indicate the corresponding mass of retained particles.

freshly filled after drying has usually $< 3.0\,\%$ of moisture. This low moisture content leads to a_w values < 0.2 and provides optimal stability under appropriate storage conditions.

The *bulk density* of the powder is an important property in economic terms. Increase in the powder density gives a more favorable weight:volume ratio and thus reduces packaging costs. Typical densities for whole milk powders range between 0.4 – 0.5 kg/L. Powder density is related to the amount of air included in the particles, but also to powder particle size. For whole milk powders, *particle sizes* vary from < 100 – 500 µm in diameter.

Powder *wettability* is a function of the particle volume and its size. Agglomerated powders have improved capillarity, thus better wetting properties. Wettability is defined as the time in seconds necessary for a given amount of powder to pass an unstirred surface of water [102]. The *dispersibility* of the powder is its ability to disperse into individual particles without leaving any lumps. Agglomeration and protein denaturation strongly influence dispersibility. *Solubility* of the powder describes as to what extent the powder dissolves well in water and forms a stable dispersion with a minimum of sediment [103].

For whole cream or high-fat milk powders, the *free fat content* is another important quality parameter. Free fat is defined as the amount of fat extractable from the powder by a nonpolar solvent such as hexane or carbon tetrachloride. The extractable fat fraction depends on factors such as fat globule coating in the liquid concentrate, fat globule membrane rupture during atomization, migration of free fat in the powder particles during final drying, and on the porosity of the powder particles. When good reconstitution properties are required (see page 38), the free fat content should be as low as possible.

9.3.2. Whole and Skim Milk Powders

The legal definitions of whole and skimmed milk powders according to Codex Alimentarius are documented in the Codex STAN A-5 of 1971.

In addition the American Dry Milk Institute has established quality standards (ADMI quality

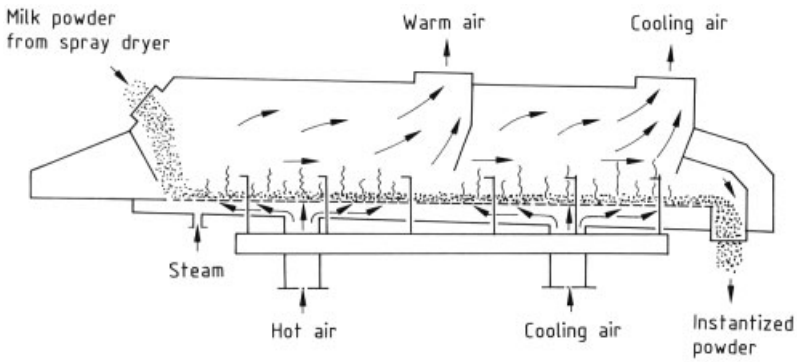

Figure 10. Fluidized bed for instantizing milk powder (reproduced from [4] with permission)

standards) for both whole and skim milk powders (Table 14).

Whole milk powder has a minimum milk fat content of 260 g/kg. It is mainly consumed after its reconstitution in water, like fresh milk. Its good reconstitution properties and pleasant milky taste are therefore the most important quality factors. Whole milk powders, unlike skim milk powders, are not otherwise categorized for trade.

Cream powders have a minimum fat content of 420 g/kg and there main use is in confectionery coatings and fillings, but also in the manufacture of dehydrated soups and sauces, in which the use of powdered cream can be advantageous.

Skim milk powder has a maximum fat content of 15 g/kg and is the key raw material for producing milk products by recombination (see Section 8.2). In addition to this, the food industry, particularly ice cream and chocolate manufacturers are major consumers of skim milk powders. The intensity of heat treatment during processing and the corresponding degree of whey protein denaturation, significantly influence the performance of skim milk powders in an industrial application. A classification of these powders according to the extent of whey protein denaturation has been established by the ADMI in 1971. This classification distinguishes three categories of skimmed milk powders: low heat, medium heat, and high heat powders.

The IDF has established a more refined classification with four instead of three heat classes: low heat, medium heat, medium to high heat, and high heat powders. In either classification, the amount of undenatured whey protein nitrogen (UDWPN) expressed in milligrams per gram of powder determines the heat class. On an average, 1 g skim milk powder has 56 mg of total nitrogen, of which ca. 9 mg is from whey protein. While a value of 9 mg/g of UDWPN is normally not found in industrial milk powders (minimum pasteurization requirements required for evaporation already reduce it), powders with 7 – 8 mg UDWPN/g can be produced and are suitable for cheese manufacturing. The largest number of applications will be best covered by medium heat powders; such as the production of consumer milks, yogurt production, ice cream manufacture, and others. For producing evaporated milks or sweetened condensed milks, for use in bakery products or in meat products, high heat powder may be the better choice [107].

9.3.3. Butter and Buttermilk Powders

Butter powders are a powdered form of butter, used mainly as an industrial food ingredient. Butter powder contains $\geq$ 82 % milk fat, some 5 – 7 % of SNF and caseinate which has been used to form an emulsion of sufficient stability to enable its further processing to powder. Processing aids used in the production of butter powders are small amounts of lecithin and citrate.

A functional advantage in industrial use of powdered butter are the good blending properties with flour which facilitate its incorporation in dough.

Butter production from butter powder by recombination, requires the addition of about 15 % water.

Buttermilk powder (BMP) is produced by spray-drying of buttermilk (see Section 9.6) and finds applications in the bakery, confectionery,

Table 15. Composition of dried milk, buttermilk and whey powders [11]

Constituent	Skim milk	Whole milk	Buttermilk	Rennet whey
Protein (N × 6.38), g/100 g powder	36.0	26.0	34.0	12.0
Fat, g/100 g powder	0.7	26.8	5.0	1.2
Carbohydrates, g/100 g powder	51.0	38.0	48.0	72.3
Minerals (ash), g/100 g powder	8.2	6.0	7.9	8.2
Humidity, g/100 g powder	3.0	2.3	3.0	7.1
Energy * , kJ/100 g powder	1505	2106	1584	1479
Calcium, g/100 g powder	1.3	1.0	0.9	0.9
Magnesium, g/100 g powder	0.1	0.1	0.1	0.2
Potassium, g/100 g powder	1.6	1.2	1.3	1.9
Sodium, g/100 g powder	0.6	0.4	0.4	1.3
Phosphorus, g/100 g powder	1.0	0.8	0.9	0.6

* Energy content 17 kJ/g for protein and for carbohydrate; 38 kJ/g for fat.

and ice cream production. BMP represents a dairy source of phospholipids acting as emulsifiers. Thus BMP with high phospholipid content can be used as a dairy ingredient to improve fat emulsification. The typical composition of milk, buttermilk, and whey powder is given in Table 15.

9.3.4. Whey Powders and Permeates [108, 109]

Whey powders have been obtained by spray drying whey of different origins (mainly rennet whey, lactic whey, casein whey). According to its origin, there are compositional differences in whey of different type (→ Cheese, Processed Cheese, and Whey, Chap. 7).

For a typical composition of rennet whey powder see Table 15. A key problem in spray drying of whey is the stickiness of the powder resulting from the strong hygroscopic properties of amorphous lactose and the lactic acid. Whey (total solids content ca. 6 %) is pre-concentrated by evaporation to ca. 60 % solids. At a certain level of concentration, lactose crystallization is induced and α-lactose allowed to crystallize (see → Cheese, Processed Cheese, and Whey, Chap. 7; → Lactose and Derivatives). Whey demineralization is another important processing route (see Section 7.5.4) leading to a key raw material in the manufacturing of infant formulas (see Section 9.3.6).

9.3.5. Milk Protein Powders [110, 111, 116]

Although a widely accepted legal definition of "milk proteins" does not exist at present, a tentative definition has been given by the International Dairy Federation (1988), in which the distinction is made between "milk proteins" and "milk protein products". The former have $\geq$ 85 % protein on dry matter, the latter $\geq$ 50 %. The proximate and mineral compositions of important commercial milk protein powders are given in Table 16.

Milk protein concentrate (MPC) [111] is produced by UF of skim milk and is a concentrate of the total true milk protein; thus the ratio of caseins to whey proteins remains essentially unchanged. During concentration of milk by UF caseins maintain their native, micellar structure, therefore micelle associated calcium and phosphate ions are retained to a large extent together with the protein. Protein contents of commercial MPC's are typically 65 – 85 % based on dry matter (Table 16).

Main end use of MPC is in the protein standardization of cheese milks; protein fortification with MPC leads to more regular curd formation and improved cheese yield throughout the milking season. Other applications are in protein fortification of yogurt milks with the objective to increase the gel strength and texture, and improve water retention properties. Milk protein concentrates are a valuable protein and calcium source for low lactose dietetic food formulations.

Caseins, Caseinates and Coprecipitates [35, 115].

Acid casein is produced through pH adjustment of skim milk to pH 4.6 with mineral acid (HCl or H_2SO_4) or lactic acid. Acidification and curd formation is also achieved by fermentation with starter culture producing lactic acid. The precipitated caseins are separated from the acid whey by a centrifuge process and the curd is washed free of lactose and other non-protein solutes. The washed curd (moisture content is ca 50 %) is then dried. Powdered acid casein has a protein content (N $\times$ 6.38) of $\geq$ 90 % based on dry matter and a rather low mineral content (Table 16). It is entirely insoluble in water, but dissolves if its aqueous dispersion is neutralized with a base. Quality criteria for edible caseins are laid down in Codex Alimentarius STAN A-18 of 1995 (Table 17).

Sodium, Potassium, and Calcium Caseinates [115]. Caseinates are produced by addition of base to acid casein curd up to a pH of about 6.8. Sodium and potassium caseinates are truly soluble proteins the solutions of which have considerable viscosity. In contrast to theses two, calcium caseinate forms a colloidal dispersion in which casein monomers are associated in micelle-like particles; calcium caseinate dispersions have relatively low viscosity.

Caseinates have wide application as protein ingredient in dairy and nondairy food formulations.

Sodium and potassium caseinates are key ingredients in emulsified food systems such as

Table 16. Composition of commercial milk proteins [110]

Milk protein fraction	Process involved	Proximate and mineral composition, g/100 g powder								
		Protein	Lipids	Lactose	Ash	Humidity	Calcium	Potassium	Sodium	Phosphorus
Acid casein	mineral acid precipitation of casein from skim milk	86.8	1.5	0.8	1.2	8.2	0.05	0.09	0.06	0.74
K Caseinate	acid curd neutralization with potassium hydroxide	87.4	1.2	0.3	5.3	3.9	0.1	2.3	0.07	0.65
Ca Caseinate	acid curd neutralization with calcium hydroxide	89.1	1.5	0.2	3.9	4	1.19	0.04	0.21	0.77
Micellar casein	microfiltration of skim milk	84.7	1.4	0.7	8.3	3.8	2.5	0.18	0.05	1.6
Rennet casein	rennet curd of skim milk	81	0.7	0.2	8.7	10	2.8	0.42	0.16	1.73
Milk protein concentrate	ultrafiltration of skim milk	79.1	1.5	7.5	7.8	3.4	2.8	0.3	0.24	1.2
Whey protein concentrate	ultrafiltration of sweet whey	77	7.2	5	2.7	3.6	0.37	0.61	0.2	0.33
Delipidated WPC	microfiltration and ultrafiltration of sweet whey	83	0.3	6.5	3	3.8	0.63	0.33	0.42	0.13
Whey protein concentrate	ultrafiltration of acid casein whey	78.6	4.7	4.7	1.3	3.7	0.1	0.12	0.02	0.5
Whey protein isolate	ion exchange on cation exchange resin	88.6	0.6	0.2	2	3.5	0.13	0.19	0.48	0.03
Whey protein isolate	ion exchange on Spherosil-S cation exchange resin	80.7	0.8	0.3	7.5	4.1	0.26	1.72	1.24	0.4

Table 17. Quality standard for edible caseins

Parameter/grade	Acid casein [112]		Caseinates [113]		Rennet casein [114]	
	Standard	Extra	Standard	Extra	Quality A	Quality B
Moisture max. % tot. solids	12	12	8	6	12	12
Protein min. % on dry solids	90	95	88	90	84	88
Milkfat max. % on dry solids	2	1.5	1.5	1.5	2	1.5
Lactose max. % on dry solids	1	0.2	1	0.5		0.5
Ash min. % on dry solids					7	7
Max. tot. viable plate count, CFU/g	100000	30000	100000	10000		

soups, dresses, sauces, or comminuted meat products. They find use in coffee creamers and coffee whiteners.

Calcium caseinate is applied as a nutritional protein and calcium base in dietetic formulations, performance nutrition and clinical nutrition products.

Coprecipitates [115] are produced by calcium addition and pH lowering of preheated skim milk, in a similar fashion as acid casein. In contrast to the former, milk is preheated at 90 – 95 °C for several minutes, in order to denature the serum proteins and by this, achieve their coprecipitation with the casein. Depending on the amount of $CaCl_2$ added to the milk and the pH of precipitation, coprecipitates with variable calcium content are obtained. The pH of precipitation for high calcium products is 5.8 – 5.5, for low calcium products the pH for precipitation is 4.6.

Such a process makes it possible to recover a major proportion of the whey proteins, together with the casein, in form of a protein concentrate. Coprecipitates are sparingly soluble in water and give colloid-disperse systems. Their application is restricted to food systems where good solubility of the protein is less important, such as in dough, bakery, and confectionery products and, to some extent also in meat products.

Rennet casein is produced by curd formation with calf rennet or a microbial rennet, followed by cutting and dispersing of the curd, washing the water insoluble casein, dewatering the dispersion and drying. Rennet casein remains insoluble in water unless the pH is raised above 9.0, and/or chelating agents are added. The major food uses of rennet casein do not required its solubilization but are more based on its heat-induced melting properties. Rennet casein is a raw material for producing imitation cheeses,

pizzas, and other fast food products. Rennet casein and coprecipitates are also processed to textured snack type foods by extrusion cooking in presence of alkali or alkaline earth metals.

Rennet casein has also nonfood industrial uses such as in paper coating.

Whey Protein Concentrates (WPCs) [110, 116, 117]. Ultrafiltration of sweet or acid whey leads to protein concentrates with, depending on the concentration factor and eventual diafiltration conditions used, protein contents ranging from >30 – 85 % on dry solids. Depending on whether the whey was delipidated or not prior to ultrafiltration, the fat content of the concentrates varies from < 1 % to 7 % on dry solids (see Table 16). Ultrafiltration of rennet whey is usually performed at milk pH or at a slightly lowered pH to improve the solubility of calcium phosphates. The concentrates obtained in this way have relatively high calcium and phosphate contents. Ultrafiltration at low pH (pH 3 – 3.5, acid whey) leads to retentates with low mineral content (Table 16). An important difference in protein composition of retentates from rennet whey and those from acid whey is the presence of the glyco-macropeptide fraction in the former. This fraction can amount to 15 – 20 % of the total protein in WPCs produced from rennet whey.

The food uses of WPCs depend mainly on their protein content. Low protein concentrates (30 – 40 % protein/TS) are mainly used as "skim milk replacers" in human nutrition, but also in calf feeding. A typical use of WPC is as a partial substitute of skim milk powder in ice cream production. High protein WPCs (>70 % protein/TS) find major application in dietetic food: infant formulas, clinical nutrition, sports nutrition, etc. It is also exploited for its functional performance (gelling, water binding, foaming) in culinary food as a substitute of egg white.

Whey Protein Isolates (WPIs) [117 – 119] have protein contents of up to 90 % based on dry solids, paired with low fat and moderately low mineral contents (Table 16). Isolates are produced by adsorption of whey proteins to ion exchange resins [118]. In the *Palmer process*, also called *Vistec process* (from Viscose Group, UK), whey at pH $\approx$ 3.5 is contacted with a modified cellulose based cation-exchanger. Following adsorption of the protein, the resin is washed free of lactose, fat and other nonprotein solutes. The protein is then desorbed from the resin by addition of alkali to the mobile phase, which causes the proteins to elute in an anionic form. This isolate is produced commercially in the United Kingdom (Bio Isolates) and in the United States (Davisco, at Le Sueur).

An alternative process, the *Spherosil process*, uses a silica bead substituted with quaternary ammonium groups [119]. Whey proteins are bound to the beads in anionic form. The Spherosil QMA resin was developed by Rhône-Poulenc (France) and is used for commercial production of WPI by Armor Proteins (France). Isolates usually have a bland taste and display high functionality in gelling, foaming, and water binding. Their end uses are mainly in dairy and nondairy culinary food in which they improve water retention, texture properties and mouthfeel.

9.3.6. Powdered Infant Formulas [120 – 122]

In production of powdered substitutes of mother's milk, the key raw materials are solids non fat from bovine milk or whey, together with vegetable fat. The task consists in creating a nutrient composition which approaches that of human milk (see Tables 1, 2, and 5). This requires an adaptation of milk SNF in several respects: The mineral content and the protein content have to be considerably reduced; the fat and the lactose content of the composition must be increased. The mean energy density of mature human milk is about 320 kJ/100 g with 55 % of the energy derived from fat, 38 % from lactose, and only 7 % from protein. The energy density of cow's milk is 275 kJ/100 g and 20 % of the energy comes from protein (protein density of breast milk: 1.6 g/100 kcal or ca. 1 g/dL; current infant formulas have protein densities $\geq$ 2 g/100 kcal ($\geq$ 1.3 g/dL).

The adaptation in terms of global composition is achieved by mixing milk SNF with demineralized whey (see Section 6.1) to achieve the desired protein content (12 – 14 % based on dry solids) and increase the lactose content to 54 – 58 % based on dry solids.

Adaptation in terms of protein quality involves a recombination of casein and whey pro-

Table 18. Protein quality in infant formulas

Protein quality and quantity	Codex [120]	EU [121]	USA	
			Old [122]	New [159]
Protein quality:				
Method of assessment	PER *	EAA ** pattern	PER *	EAA ** pattern
Value	min. 85 % of casein	min. 80 % of BM	min. 70 % of casein	reference pattern
Protein quantity:				
Minimum protein density formula	1.8 g/100 kcal	1.8 g/100 kcal	1.8 g/100 kcal	1.7 g true protein/100 kcal
Maximum protein density	4.0 g/100 kcal	3.0 g/100 kcal	4.5 g/100 kcal	3.4 g/100 kcal

* PER = Protein efficiency ratio (as determined in growing rats).
** EAA = Essential amino acids.

teins leading to a balance of the two categories of protein closer to the one prevailing in human milk. In the latter, noncasein proteins predominate over casein. By mixing adequate proportions of milk SNF and demineralized whey, casein : whey protein mass ratios of 40 : 60 – 50 : 50 can be established. This adaptation is beneficial with respect to the amino acid profile of the protein, coming closer to the one of mother's milk than an 80 : 20 ratio. The reduction of the casein proportion in the formula also leads to formation of weaker curds in the infant's stomach.

The appropriate fatty acid composition of an infant formula can be achieved by mixing vegetable fat such as corn or soybean oil with milk fat. The use of milk fat is not mandatory. The mix must have a linoleic acid content not lower than 0.3 g/100 kcal (418 kJ), or ca. 3.0 g linoleate/100 g total solids (total fat content 26 – 28 %).

The nutrient composition of an infant formula is subject of regulation by Codex Alimentarius, and by specific regulations valid within the EC, and the United States (Infant Formula Act of 1985). The Infant Formula Act regulates minimum and, in some cases, maximum contents of individual nutrients, protein quality, and protein density (see Table 18). While powdered infant formulas are the usual form of this product in most parts of the world, liquid in-can sterilized or aseptically filled, ready-to-feed products and liquid concentrates dominate the market in the United States.

In addition to standard infant formulas a number of *special formulas* have been developed as a response to specific needs in infant feeding, including:

1) Premature or low-birth weight formulas, covering particular nutritional needs of a growing number of premature born infants (higher needs for protein, calcium and polyunsaturated fatty acids).
2) Hypoallergenic infant formulas, recommended for infants with an acquired intolerance to cow's milk proteins, or with a genetic predisposition for allergy. In these formulas, the protein component has been subjected to enzymatic hydrolysis in order to attenuate the allergenic potential of bovine milk proteins [123].
3) Acidified infant formulas, with or without living lactic bacteria. Such formulas are particularly useful in countries where hygienic problems in bottle feeding (insufficient water quality and others) are frequent, because lactic acid efficiently lowers the growth rate of bacterial contaminants.
4) Semi-elemental diets, designed for infants with acute feeding problems of different origin. In such formulas the protein is usually provided in form of an oligopeptide fraction and lactose replaced by maltodextrins. To provide an easily absorbable form of fat, medium chain triglycerides are incorporated in combination with milk or vegetable fat.

9.4. Fermented Milk Products [97, 124, 125]

Fermented (cultured) milk products are among the oldest milk products, resulting in early periods from natural contamination of milk with lactic acid bacteria and yeast. Lactic acid fermentation leads to the development of texture, pleasant taste, and improved storage quality of milk.

Table 19. Properties of yogurt cultures [124]

Starter culture	Temperature optimum for growth, °C	Acidity, % lactate	Lactate isomer formed	Proteolysis	Other characteristics
Streptococcus					
S. cremoris	30	0.5 – 0.7	L(+)	low	creaminess by formation of slime
S. lactis	30	0.5 – 0.7	L(+)	high	
S. thermophilus	>40	0.6 – 0.8	L(+)	weak	diacetyl formation
Lactobacillus					
L. acidophilus	35 – 38	up to 1.9	L/D		
L. bulgaricus	>40	up to 1.8	D(−)	high	acetaldehyde formation
L. brevis	30	up to 1.5	L/D		
L. casei	30	up to 1.5	L(+)	high	
L. helveticus	>40		L/D		acetic acid taste
Bifidobacterium					
B. bifidum			L(+)	low	

Traditional and industrial production of fermented dairy products uses specific bacterial cultures, known as *starters*. Characteristic biochemical properties of the most important lactic acid bacteria used in starter cultures are shown in Table 19.

The health benefits of fermented milk products are long known. More recently, some of the beneficial properties of yogurt and other fermented dairy products have been attributed to the presence of living lactic bacteria, which are believed to act favorably on the intestinal microbial flora of the consumer. The term *probiotic* designates "live feed supplements which beneficially affect the hosts by improving their intestinal microbial balances" [126]. Fermented dairy products produced with probiotic cultures have been put on the market in recent years by: Danone (Actimel), MD Foods (Gaio yogurt), Nestlé (LC1 yogurt), Bauer (LA 7 yogurt), and others. Yakult, the Japanese fermented milk based beverage is being increasingly promoted in Western Europe as a "probiotic food". Anticipated and established benefits of fermented dairy products are reviewed in [127].

9.4.1. Legal Definitions

For yogurt, sour milks, and related fermented milk products, most countries regulate the minimum content in milk SNF and the maximum content of non-milk constituents such as sucrose, or fruit preparations. Thus, the maximum content of nondairy constituents in a fruit yogurt is limited to 30 wt % in the EC. The presence of a minimum viable count of lactic bacteria (*Lactobacillus bulgaricus* and *Streptococcus thermophilus*) in yogurt, is a likewise legal requirement. Milks that have been acidified chemically (by addition of lactic or citric acid) are not classified as fermented milk products. The addition of hydrocolloids such as pectins, carrageenans, and other gums, or gelatin, is furthermore subject of regulation on European [128] or on national level.

9.4.2. Specific Fermented Products

Sour milk is usually made from pasteurized milk by fermentation with mesophilic cultures of *Lactobacillus acidophilus* and other starters with moderate acid forming activity, such as *Streptococcus cremoris, S. lactis*, or *S. thermophilus*. The pH of the fermented, cooled product is 4.6 or lower, the consistency viscous. Sour milks are commercialized with different fat contents.

Yogurt, by definition in the Codex Alimentarius (FAO/WHO 1977), is milk fermented with *L. bulgaricus* and *S. thermophilus* and the microorganisms in the final product must be "viable and abundant". Addition of evaporated milk or milk powder into yogurt milk is current practice.

The characteristic flavor of yogurt is largely due to carbonyl compounds, particularly acetaldehyde (25 – 50 mg/kg), volatile fatty acids, and to some extent, due to acetone (ca. 1 mg/kg) and acetoin. Free amino acids resulting from casein proteolysis by proteases from the starter, amount to 0.3 – 0.5 g/kg and have some contribution to the flavor profile. The lactic acid content of yogurt varies in the range 0.8 – 1.2 %

(36 – 54 °SH), corresponding to the fermentation of about 20 – 30 % of the total lactose in milk. According to the fermentative properties of the starter cultures (Table 19), fresh yogurt contains $\geq$ 50 % of L(+) lactate based on total lactic acid. With increasing age, the proportion shifts in favor of the D(−) isomer due to the higher fermentative activity of *L. bulgaricus* at an elevated acidity. The typical pH of "ripened" yogurt varies from 4.0 – 5.0, according to type of yogurt, fat content, and specific starter cultures used. Low-acidity yogurts have a pH above 4.6, while normal acidity yogurt has a pH $< 4.6 \geq$ 4.0.

The two main types of yogurt are the *set-style yogurt* with a firm gelled texture, and the *stirred-style yogurt*, with a fluid texture of higher or lower viscosity. Fruit yogurts are always of the stirred type. While fermentation for the set-style yogurt must take place in the package (formation of the yogurt gel starts already at pH < 5.5); the stirred-type yogurt is fermented in fermentation tanks. After fermentation at 40 – 42 °C for 3 – 5 h (temperature and duration depend on the cultures used), the product is stirred at a low speed prior to its transport by pumping and filling. A moderate to low shear rate is critical to achieve the desired viscous properties of stirred-style yogurt. The process for production of standard stirred-type yogurt is shown in Figure 11.

Stirring of the yogurt milk during fermentation leads to products with low viscosity (*drinking style yogurt*). Yogurt can be frozen like ice cream. *Frozen-type yogurt* is common in the USA. Stirred yogurt can be dehydrated to give *yogurt powder*. Drying conditions have to be adjusted to guarantee the survival of an important fraction of the yogurt bacteria. Yogurt powder is a food ingredient used in bakeries and the confectionery industry, but also in the production of dairy desserts.

Bioghurt is the trade name of yogurt produced by fermentation with a selected population of starter organisms including *S. thermophilus, L. acidophilus and B. bifidum.* A main property of Bioghurt is that lactic acid is predominantly present as the L(+) isomer, the form considered to be the physiologically active form of lactic acid. As a result of a weak post-fermentative acidification during refrigerated storage, the yogurt has a mild taste.

Bifidus yogurt is the name for a mild tasting yogurt produced mainly in Germany, Japan, and the USA by using selected cultures of *B. bifidum*, *B. longum*, *S. thermophilus*, and *L. delbrueckii.*

Kefir. The name is derived from the Caucasian word *kefy,* meaning pleasant taste. Kefir originates from the Northern Caucasus and was originally produced from both cow's and goat's milk by using kefir grains. Kefir is sour and contains low concentrations of ethanol (0.02 – 0.04 %) and CO_2. Kefir grains contain lactic bacteria and yeast cells immobilized in a polysaccharide matrix. The microflora consists of heterofermentative lactic acid streptococci (*S. lactis*, *S. cremoris*), mesophilic lactobacilli such as *L. kefir*, and *Leuconostoc species*, acetic acid bacteria,and yeasts (*Candida kefir, Kluyveromyces marxianus* and others).

To produce kefir, preheated milk is inoculated with kefir grains and allowed to ferment at ambient temperature (20 – 23 °C) for 10 – 20 h; the kefir grain is then filtered off and the beverage cooled and cold stored. The flavor of kefir is very different from that of sour milk or yogurt; mainly due to the presence of yeasts, formation of ethanol, and CO_2. Formation of CO_2 presents an obstacle to the industrial production and distribution of kefir (blowing up of packages, unless glass bottles are used).

The most important industrial kefir production is in Russia. In this production, a kefir culture is used instead of the grain. The ethanol content of industrial kefir is about 0.02 %.

Kumys, also called Kumiss, Koumiss, or Kymys, originates from central Asia, where it was produced from mare's milk. In today's kumys production (mainly in Russia and the Mongolian plains of central Asia), the beverage is produced from pasteurized mare's milk by fermentation with starter culture. The starter consists of *L. delbrueckii* and lactose fermenting yeast (*Sacharomyces lactis*). In Russia, kumys is also produced from cow's milk [129].

Ymer is a Danish cultured milk product produced from cow's milk with an increased total solids content, to achieve a viscous consistency. The normal manufacturing procedure consists in fermenting skimmed milk with cream starter at 20 – 22 °C until a pH of 4.6 is reached. This

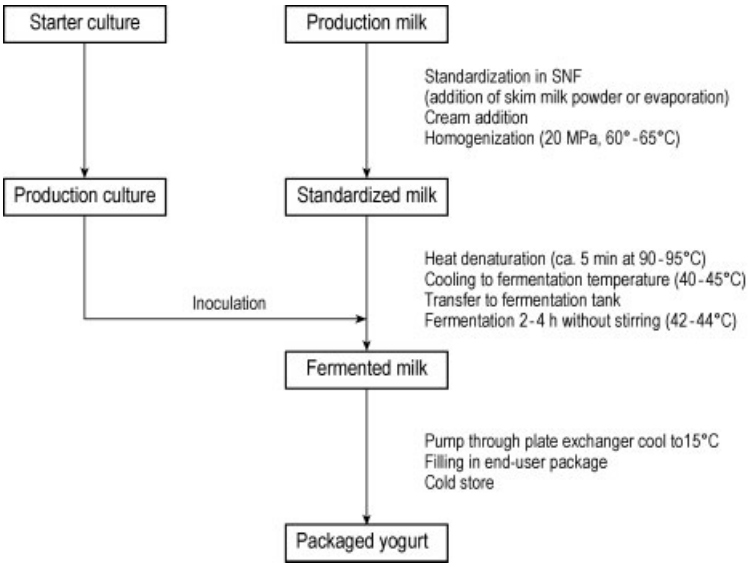

Figure 11. Flow sheet of production of stirred-type yogurt

is followed by cutting the resulting curd and whey removal. Following addition of cream and homogenization at low pressure, the product is cooled and packaged. Another process employs UF to concentrate the protein in the production milk about two-fold. The retentate is fat adjusted (typically 35 g milk fat/L), the mix homogenized and pasteurized. The resulting composition is fermented with cream starter for about 20 h; it has a gel-like consistency [130].

Yakult (registered trade name since 1935 in Japan) is a Japanese fermented milk beverage. After addition of glucose and glycine skim milk is heated at 90 – 95 °C for about 30 min. Following its cooling to 45 °C, the heated milk mixture is inoculated with "Shirota culture" (strain of *L. casei*, isolated by M. Shirota in the 1930s from human intestine).

Incubation lasts 6 – 7 d at 37 – 38 °C. Following fermentation, the product is diluted with water and sugar, gums and lactic acid are added prior to packaging in small cups or tins. Yakult enjoys great popularity as probiotic food in Japan, but it is conquering new European and Asian consumers.

Table 20. IDF Survey on cream definition in various countries [71]

Country	Minimum fat content	Other parameters	Classes according to fat content	Minimum fat content of cream class
Australia	350 g/kg	max plate counts and coliforms	rich cream	480 g/kg
		pH limits, limits for acidity	reduced cream	250 g/kg
			light cream	180 g/kg
			extra light cream	120 g/kg
Switzerland	150 g/kg	SNF minimum 85 g/kg	double cream	450 g/kg
			cream/whipping cream	350 g/kg
			half cream/coffee cream	150 g/kg
France	300 g/kg		light cream	$\geq 120 < 300$ g/kg
Germany	100 g/kg	milk solids can be increased	whipping cream	300 g/kg
UK	180 g/kg		clotted cream	550 g/kg
			double cream	480 g/kg
			whipping cream	350 g/kg
			cream (single c.)	180 g/kg
			half cream	120 g/kg
Netherlands	100 g/kg		whipping cream	350 g/kg
			coffee cream	200 g/kg
			half cream	100 g/kg

Filmjölk is a Swedish type of sourmilk produced from fat-standardized milk by fermentation with buttermilk starter (including *S. lactis*, *S. diacetilactis*, and *S. cremoris*). The resulting product resembles to some extent acidified buttermilk, it has smooth texture and a fresh, acid taste.

9.5. Cream [4, 12 – 14, 131, 132]

Cream is the fat-rich part obtained by skimming of whole milk in a separator. The legal definition of cream and cream types varies considerably from country to country. An overview of legal definitions in different countries has been compiled by the IDF (Table 20). Cream is an oil-in-water emulsion, similar to whole milk. In contrast to whole milk, the dispersed fat volume fraction (Φ) is much higher in cream ($0.1 < \Phi < 0.4$ for most creams, ca. 0.04 in whole milk).

9.5.1. Production

The traditional way of obtaining cream is to let the milk stand until spontaneous rising of the fat globules is completed. Industrial production of cream relies on milk skimming in continuously operated disk centrifuges (see Section 7.2). Depending on the type of cream to be produced, the fat content is adjusted by adding skim milk into cream for fat contents below that produced by the separator (normally $>35 < 40$ %). Cream has to be pasteurized. A negative phosphatase test is achieved with cream of ≥ 20 % fat by holding for 15 s at 80 °C. This treatment also ensures sufficient inactivation of the lipoprotein lipase, of which about one fourth of the total activity remained in the cream after skimming. While homogenization may be omitted for pasteurized cream with relatively short, refrigerated shelf life, it is indispensable for UHT creams with extended shelf life.

Homogenization conditions for creams with >20 % fat must be carefully adapted to provide sufficient emulsion stability (avoid cream plugs in the package), without causing unacceptably high viscosity and, in the case of coffee cream, enhanced flocculation tendency in coffee. Due to the unfavorable protein : fat ratio in cream, high homogenization pressure leads to a large increase of fat surface and causes cluster formation via protein sharing between different fat globules. The formation of such clusters leads to a sharp rise in the viscosity of the cream, which can be attenuated by a second stage of homogenization at a lower pressure [132].

During processing of cream, great attention must be paid not to damage MFG membranes by pumping, shearing, and air introduction. Such damage causes the release of free fat; the latter unfavorably affects several properties of cream. Pumping of cream should be done at temperatures >40 °C, at which the milk fat is entirely liquid. The risk of forming free fat by membrane rupture is highest at temperatures of 20 – 30 °C, where part of the fat has crystallized, while another part of it remains liquid and can coalesces after its release from the globule.

During production of *whipping cream*, attention must be given to the transport of the cream, preferably done at 60 – 62 °C using low shear pumps. Following its pasteurization, the cream is cooled in plate heat exchangers to about 8 °C and then transported (by means of special pumps) to the ripening tanks, where the final cooling without agitation takes place. "Ripening" means fat crystallization in a process requiring several hours. The ripened cream can be further transported to the filling machine once crystallization is almost completed.

9.5.2. Types of Cream

According to the desired functionality, different types of creams with different fat contents are produced and marketed.

Whipping Cream [133]. Whipping of cream becomes difficult if the fat content falls below about 30 %. Problems in whipping can also arise due to seasonal variations in the milk fat composition (summer fat has a lower melting range than winter fat). Good whipping properties include a sufficient volume increase ("overrun") and high stability of the froth without development of syneresis. To achieve a good froth, whipping cream should be kept at 6 °C or lower. In the whipped cream, fat globules associated with the air-water interface stabilize the air bubbles. Due to rupture of the membrane in part of the fat globules, crystalline free fat glues

different globules together to form a protective lamella around the air bubble.

Whipping cream quality is estimated by its overrun (which should be 90 – 100 % of the initial volume), the rate of serum separation and the mechanical strength after a defined whipping time in a standardized beater.

Coffee cream [134, 135], depending on the country of commercialization, have fat contents of 12 – 20 %. Coffee cream with 12 – 15 % fat is also called *crème légère* in France.

These creams are homogenized, to increase their viscosity to a desirable level, and to enhance their whitening effect in coffee or tea (stronger light scattering due to higher number of fat globules).

Homogenization has, on the other hand, the undesirable effect to increase the flocculation tendency of the cream in hot coffee (clustered fat globules form macroscopically visible flocs when aggregated by heat and acid). Adequate combination of distinct heat treatment and homogenization steps is able to minimize this problem.

Spray Cream (Aerosol Cream). Cream in containers under gas pressure is a relatively new development. Using a *bag-in-bottle* package permits to produce an UHT version of the aerosol cream. Bag-in-bottle systems are not pressurized and the outer container is often made of semi-rigid plastic. A so-called *bag-in-can* system is used for pasteurized cream. In the latter system, the propellant gas (e.g., N_2O) is injected in the space between the bag and the outer can, pressurizing the cream in the bag. The structure of the foam produced by the action of a propellant is quite different from that of a mechanically whipped cream.

9.6. Butter [4, 12 – 14]

Legal definitions of butter are not yet fully harmonized between different countries, but the term butter is reserved to a spread made from milk fat only; butter must be obtained either directly form milk or from cream. An EC draft regulation issued in 1991 proposes the following definitions of butters, according to their fat content: butter (80 – 82 % fat), three-quarter fat butter (60 – 62 % fat) and half-fat butter (39 – 41 % fat). Products with < 39 % fat are designated as low-fat dairy spreads.

Mixed fat spreads (milk fat with vegetable fat) require, according to Codex Standard A1 for butter, milk fat contents between >15 % and < 80 % of the total fat.

9.6.1. Production

To produce 1 kg of butter, about 24 L of whole milk need to be processed.

Butter is produced from whole milk or cream according to a process involving distinct stages (see Fig. 12):

1) Production of cream of a standardized fat content (usually near 40 %)
2) Cream pasteurization and ripening (including fermentation in case of sour cream butter)
3) Churning, buttermilk separation and kneading of butter corns into butter.

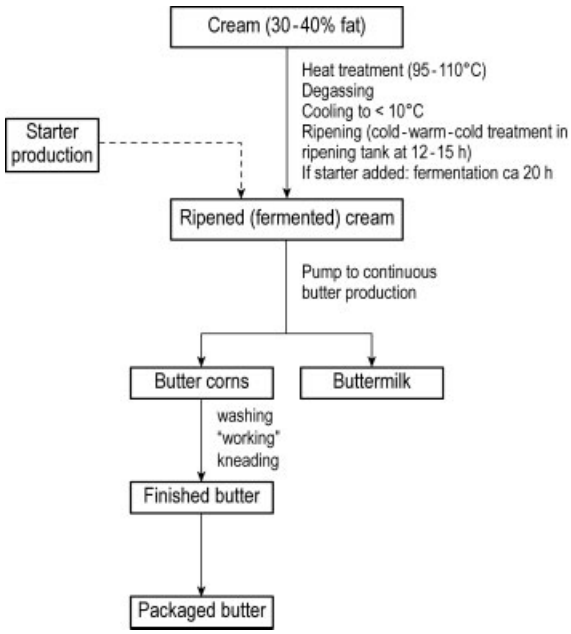

Figure 12. Flow sheet of butter production

9.6.2. Steps in Buttermaking

Cream Ripening. The adjusted fat content of the cream used for buttermaking depends on the type of process used. For traditional buttermaking, the fat content is 25 – 30 %, for continuous, industrial processes it varies from 30 – 50 % and, in the Alfa process (see page 50), a fat content >80 % is required.

The temperature range for pasteurization is 95 – 110 °C; pasteurization is done continuously without an added holding period. Prior to cooling of the cream vacuum flashing removes undesirable volatiles.

If sour cream butter is to be produced, 1 – 6 % of starter culture is added to the cooled cream in the ripening tank. The typical starter culture consists of *S. lactis*, *S. cremoris*, *S. diacetylactis*, and *Betacoccus citrovorus*. During the ripening and fermentation period of 18 – 24 h, lactic acid formation decreases the pH of the cream to 5.0 – 4.6. The pH of unfermented cream after ripening is 6.0 or higher. Fermentation leads to special flavor notes due to formation of diacetyl (2,3-butanedione) and acetoin. The production of diacetyl occurs mainly when the pH has already dropped below 5.2.

Physical ripening of the cream is a key step in butter production and has the objective to influence the hardness and consistency of the butter. Without this thermal process taking place in the ripening tank, no constant butter quality can be produced throughout the seasons. Due to a significantly higher content of saturated fatty acids in the winter milk, the butter obtained from winter milk is harder and less well spreadable while, due to a higher content in unsaturated triglycerides in summer milk, softer butter is obtained during the summer. The temperature program for cream tempering may be based on its iodine number (see Section 3.1.1). Typical winter cream with a low iodine value, requires a program which leads to less solid fat during the churning operation, by crystallizing out a low melting olein phase which will become the continuous oil phase in the churned butter. This can be achieved by minimizing the amount of mixed fat crystals in the "cold-warm-cold" process (rapid cooling to 6 – 8 °C, gentle warming to about 20 °C, cooling to ca. 15 °C).

When the task consists in making soft (iodine values >39) summer fat harder, an inverse sequence of temperature events is chosen.

Churning of the Ripened Cream into Butter. Butter grains are formed when cream is subjected to high mechanical stress, provided that the fat is present in the fat globules both in liquid and crystallized form. Liquid fat will, after rupture of the fat globule membranes, coalesce to form the continuous phase of the nascent butter.

When the cream is strongly agitated at the butter-making temperature (12 – 15 °C for winter cream and 10 – 12 °C for summer cream) a foam of protein stabilized air bubbles is produced, to which the membranes of the fat globules are drawn. As the air bubbles become smaller during agitation, the foam compacts and this presses liquid fat out of the fat globules. It also favors the rupture of the MFGM. The content of individual fat globules then condenses to form butter grains. The latter will steadily grow to reach a size of a few millimeters. At that stage, buttermilk is drained off and the butter grains are further kneaded to remove interstitial water. "Working" of the butter grains at the same time reduces the size of the remaining, interstitial water droplets, the average diameter of which should be < 10 μm, to prevent bacteria from growing in the remaining aqueous environment (residual water content should not exceed 16 %).

Specific Buttermaking Processes.

Batch type buttermaking in traditional wooden churning cylinders equipped with inside beaters and kneading rolls is presently only of historical interest. Today, metal churns of different technical design are operated in most dairies. Typical capacities of such churns are 8000 – 12 000 L of cream. The churns are filled to about half their volume (to allow for overrun) with stirred, temperature-adjusted cream and operated at selected rotating speed (typically 25 – 30 rpm). About 40 – 60 min are required until the butter grains appear. Buttermilk is drained off through a sieve and the butter grains are washed with cold water, in order to remove nonfat solids. More recent versions of such machines permit the reuse of the washing water several times to reduce the wastewater load. If desired, salt can be added to the butter in form

of solid table salt or a concentrated saline solution. Subsequent kneading of the nascent butter is conducted at a markedly lower speed (10 – 20 rpm) and results in compaction of the butter grains and expelling of surplus interstitial water.

Continuous Buttermaking Processes. The most widely spread process used in Western Europe was initiated by Fritz in 1940 and further developed by Eisenreich (1959). Key features of this process are shown in Figure 13. While the original Fritz process was only applicable to sweet cream, the further development by Eisenreich allows the production of butter from sour cream. The process is based on a double cooled churning cylinder fitted with beaters, in which the cream is rapidly churned. The churned cream then passes on to a separator (rotating screw) where buttermilk is drained. The butter phase is moved through a perforated plate in which the grains are further narrowed, water droplet size is reduced, and the water finely dispersed within the butter. The second "working section" consists of compartments, an injection compartment, a vacuum working section, and a final working stage including a moisture control unit. A nozzle located at the outlet of this section discharges the product in form of a continuous ribbon. The capacity of such machines goes from 200 to 10 000 L/h.

The *Alfa process* (named after Alfa Laval, today Tetra Pak, Sweden), also called the concentration process, uses a hermetically closed centrifugal separator to concentrate pasteurized cream with 30 – 40 % fat to a fat content of 82 % (close to the fat content of butter). The resulting concentrated oil-in-water emulsion is then cooled and sheared in a cooling screw to achieve phase inversion. As no buttermilk must be removed from the churned product, the MFGM material becomes part of the butter. Butters produced according to this concentration process therefore have an increased phospholipid content. Another characteristic feature of butter produced by the Alfa process is the virtual absence of air in the product, while butters produced according to modified Fritz type processes contain 3 – 6 % of air.

Another technical variant widely used in the United States, is the *Cherry Burrel process*. This procedure also uses concentrated cream and a homogenizing device to destabilize the fat globule and induce phase inversion.

Packaging and Storage. Automated packaging of the finished butter should immediately follow its production, in order to avoid contamination with environmental airborne microorganisms and texture damage.

Packaging material for butter must provide light protection, be fat impermeable and oxygen-tight, in order to provide sufficient refrigerated shelf life of the product. The packaging material must be devoid of migrating metal ions and it must not adsorb and transmit odors from the environment.

Common packaging materials are parchment paper and aluminum foil with coatings on the product side.

Refrigerated shelf life for butter at 4 – 6 °C is about 4 weeks. For longer storage periods, butter can be stored frozen at − 25 °C.

Butter Defects. *Odor and flavor defects* of butter may originate from deficient cream quality or may be the result of the buttermaking process itself. Oxidative defects can result from in-

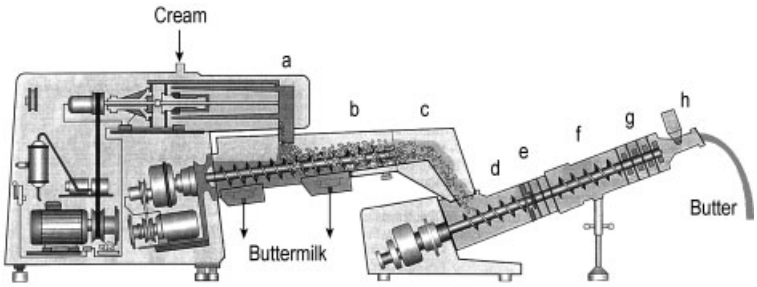

Figure 13. Continuous buttermaking plant (reproduced from [4] with permission)
a) Churning cylinder; b) Separation section; c) Squeeze drying section; d) Second working section; e) Injection section; f) Vacuum working stage; g) Final working stage; h) Moisture control unit.

creased copper or iron contents of the cream or from contaminated process water. To avoid lipolysis, adequate cream pasteurization is critical as it ensures inactivation of the lipases from microbial origin.

Texture defects in butter in finished butter with completed fat crystallization may result during its mechanical handling. If mechanical treatment leads to the coalescence of finely dispersed water droplets, water leakage can occur at a later moment. Coalesced interstitial water also favors bacterial growth in the stored product.

9.7. Anhydrous Milk Fat and Butter Oil [4]

Definitions and Properties. According to IDF Standard 68/1977 [136], *anhydrous milk fat* (AMF) must have 99.8 % fat at the minimum. Other quality criteria are the peroxide value, which has to be ≤ 0.2 meq/kg; and the free fat acidity which must not exceed 0.3 % (expressed as olein). The peroxide value is a measure for reactive lipid hydroperoxide intermediates; the free fat acidity is related to the extent of lipolytic deterioration. The latter can be determined by titrating the extracted fat with a weak alkali. Anhydrous milk fat must be produced either from butter or from fresh cream.

Butter oil is liquid milk fat with a minimum fat content of 99.3 %. While, according to the same IDF standard of identity, butter oil may contain up to 0.5 % moisture, AMF must have a maximum moisture content of 0.1 %. Anhydrous milk fat is solid at temperatures below 16 °C and has, if stored under N_2 gas at 4 – 6 °C, a shelf life of several months. It is a more suitable form of milk fat for transport and storage than butter. Major importers of AFM are nondairy countries that utilize AMF for recombining. Anhydrous milk fat and butter oil function also as raw materials for chocolate production, for the bakery, confectionery, and the ice cream industry.

Production [137]. Anhydrous milk fat is produced either from butter or from fresh cream. If *sweet cream butter* is the raw material, the butter is melted and warmed to 60 °C, and it is hold for some time at this temperature. From the holding tank the butter oil is concentrated in a centrifugal concentrator which separates the light butter oil phase from the heavier aqueous emulsion. The oil is then heated to >90 °C and residual moisture is stripped off in a vacuum chamber, up to a residual level of 0.1 % or less. When the raw material is *fresh cream*, the first step is concentration of the cream to at least 75 % fat in a centrifugal separator. The concentrate is then passed through a homogenizer at a high pressure, in order to induce phase inversion. Following this churning process, another concentration step is needed to bring the fat content to ca. 99.5 %. Heating at 95 – 98 °C and vacuum stripping of residual moisture will lead to the anhydrous product. After vacuum "drying", the oil is cooled to packaging temperature ($\approx$ 40 °C) and filled into containers or drums. The containers are first gassed with N_2, which sinks to the bottom because of its higher density as compared to oxygen. Upon filling of the liquid milk fat, the gas overlays the fat in the package. The packaging material has to be air tight and light proof.

Fractionation of Milk Fat [138 – 140]. At 28 °C melted milk fat can be separated by selective crystallization into a hard *stearin fraction* (melting range 40 – 45 °C) and a soft *olein fraction* (melting range 26 – 28 °C).

A semicontinuous process was developed in Belgium by A. Tirtiaux in 1976, which allows to separate sequentially crystallized triglycerides on a stainless steel vacuum belt filter. Milk fat fractions with different melting ranges can be produced by the Tirtiaux process. Such fractions are produced to satisfy specific requirements of the food industry. Main European producer of such tailor-made milk fat fractions is Corman in Belgium. Outside Europe, New Zealand Dairy Board has developed a continuous process for a spreadable butter, based on blending of milk fat fractions with different melting behavior.

Cholesterol reduction is another reason for refining butter oil or AMF. Several procedures lead to a reduction of the cholesterol content of the milk fat by about 90 % [139].

Mixing of the oil with powdered or granulated charcoal can remove up to 90 % of the cholesterol by adsorption. The adsorbent also removes color and flavor components from the

milk fat, these may have to be reintroduced after cholesterol removal.

In the *cyclodextrin* procedure inclusion complexes are formed between cholesterol and cyclodextrin, which can be precipitated from the oil phase. This procedure is less efficient than adsorption by carbon particles.

More recently, *supercritical CO_2 extraction* has become applicable to remove cholesterol. Up to 90 % of the total milk fat cholesterol can be removed using this procedure.

9.8. Dairy Spreads [141]

In the last two decades, numerous dairy spreads appeared on the markets in Western Europe and the United States. These spreads differ in their total fat level and in their proportions of butterfat to nondairy fat. Legislation regarding these products varies from country to country. Main raw materials for the manufacture of *pure dairy spreads* are butter oil, AMF and low-melting milk fat fractions. In the manufacture of *mixed dairy* spreads, vegetable oils, e.g., soybean oil or rapeseed oil, are mixed with milk fat in adequate proportions to obtain good spreadability at refrigeration temperature. In the production of *reduced-fat spreads*, often further components such as milk proteins, gelatin, modified starch, or carboxymethyl cellulose are needed to improve product stability and water immobilization.

Mixed spreads with a high fat content but improved spreadability have been on the market for some time.

Bregott is the name of a Swedish product with a total fat content of 80 % of which 15 – 30 % is soybean oil. The exact proportion of vegetable oil varies and is adapted to the (seasonal) hardness of the milk fat. Bregott has, at refrigeration temperature, better spreadability than butter, however it becomes too liquid at room temperature. Bregott has been on the market since the early 1970s.

Clover is a similar mixed spread developed in the United Kingdom and marketed since 1983. In this spread, hydrogenated vegetable oil is used, which makes the product less liquid at ambient temperature.

Reduced- or low-fat spreads have total fat contents $< 60\,\%$ (*reduced-fat spreads*) or $< 39\,\%$ for *low-fat spreads*. Spreads with fat contents of 39 – 41 % are also called *half-fat spreads*. Corresponding water-in-oil emulsions are produced, by dosing the aqueous phase (e.g., buttermilk) into the liquid fat phase. High shear devices such as scraped surface coolers or special homogenizers are needed to finely disperse the aqueous phase in the oil phase. The increased amount of dispersed water causes a sensation of empty mouthfeel and poor palatability. Addition of a small amount of milk proteins or gelatin improves the situation. The large amount of water also leads to microbial instability, thus the use of preservatives such as potassium sorbate or others is required.

9.9. Dairy Desserts [97]

This section only describes industry-made, ready-to-eat, or powdered desserts produced from milk, cream, and other milk constituents. Frozen desserts such as ice cream and frozen yogurt are dealt with in (→ Ice Cream and Frozen Desserts).

Ready-to-eat Desserts. Main categories of ready-to-eat dairy desserts include:

Puddings (Custards, Flans). These are mostly firmly textured, demoldable gels which are easily removable from their container. Typical examples are vanilla, chocolate, or caramel puddings with a caramel covering. Egg custards are mainly based on egg solids but most recipes contain significant proportions of milk SNF and cream. In dairy puddings, gelling is induced alternatively by heat setting (for whey protein and egg protein as well as for starch-based gels), by acid setting (for casein-based gels) or, as in the case of fresh curd products, by rennet-induced gelling of the dessert composition. It is also possible to create the gelled texture during a cooling step of the product, if gelatin or carrageenans are being used.

Creamy desserts include a wide variety of canned custards and liquid puddings. In contrast to custards or flans, these products are more or less viscous liquids with a creamy texture.

Aerated desserts can be based on fresh cheese curd, on cream, or on milk proteins (caseinates, whey proteins). Foam stabilizers such as starches, carrageenans, gelatin, and emulsifiers are used in such foamed dessert products.

Acid dairy desserts include fresh curd products (curd formation by lactic fermentation), yogurt-based desserts, and milk containing fruit preparations. Their creation often involves the use of highly methoxylated pectins, capable of producing gels at low pH.

Multi-layer desserts are made of several layers including chocolate layers, vanilla cream layers, fruits or fruit extracts. An upper layer of whipped cream is very common to these products, most of which have to be kept under refrigerated conditions.

Ready-to-eat desserts can also be classified according to their shelf life and storage conditions, and related with it, the mode of heat treatment used for their production. *Pasteurized products* require refrigerated storage and have shelf life of a few days up to several weeks. Hot filling after pasteurization allows 3 – 4 weeks of refrigerated shelf life.

Canned, retorted products are produced in rotary autoclaves by retorting in the temperature range of 115 – 125 °C for several minutes. This technology is particularly suited to produce long-shelf-life puddings and crèmes in metal cans; an example is *Crème Stalden,* a product from Nestlé in Switzerland.

UHT-sterilized desserts have long shelf life at ambient temperature, like the retorted ones but the heating process is less damaging. UHT-sterilization must be linked with aseptic filling. Packages mostly consist of paperboard, polypropylene, or polystyrene.

Powdered Desserts. Shelf-stable dairy dessert compositions can be produced in powdered form by spray-drying the liquid composition, as well as by dry mixing dry ingredients.

Powdered dessert compositions based on milk powders and containing pre-cooked starches and other ingredients are on the market. After their reconstitution in water or milk these powders are further processed into the ready-to-eat product by the consumer.

10. Quality Control

An inventory of international adopted (IDF, ISO, AOAC) methods of analysis and sampling of milk and milk products has been issued by IDF [142].

Quality control is a process taking place at different levels, from the raw material (raw milk) to the packaged product and its distribution and retailing.

Quality features concern in a first line food safety (in terms of microbiological and toxicological risks, contaminant and residue levels). Other quality features concern composition, such as maximum or minimum values for water and nutrients. Finally, adequate packaging and storage must ascertain that the required quality is maintained during the entire shelf life of the product.

Quality Control of the Raw Material. Quality control of raw milk includes both permanent and periodic control of specific quality factors. For virtually any kind of industrial transformation of raw milk, basic parameters such as acidity (°SH), pH, water addition (freezing point), or absence of antibiotics are established. Frequent periodical controls concern composition (total solids, total protein, fat content by automated infrared analysis), bacteriological quality and somatic cell content.

A somewhat longer periodicity is applied for specific controls such as testing for *Brucellosis* (every 2 – 3 months) and other udder diseases.

In a survey by the IDF in 1998 [143] the practice of quality payment was compared for >30 different countries. Besides hygienic standards, the total protein and the fat content are used everywhere as criteria. Some countries in addition use the total solids non fat (SNF) content.

Quality Control in Processing. In the industrial transformation of milk, preventive control measures along the production line are of greatest importance. The concept of Hazard Analysis & Critical Control Point (HACCP) aims at the identification of safety hazards within the production line (process) and to suppress such hazards by adequate control. It is considered that this procedure best guarantees the safety of the final product.

Finished Products Control. Analytical control of finished products has to ascertain that the product complies with legal requirements and with declared nutrient contents. Sterility testing is important for shelf stable, liquid milk products such as retort or UHT-sterilized, fluid products.

11. Legislation

Milk and dairy products are regulated in specific chapters of the Food Legislation on a world level by the Codex Alimentarius, on regional levels such as the European Union, and on a national basis. Today, national governments are pushed to implement Codex rules into national legislation.

Also, e.g., European countries are in the process to "harmonize" their national regulations with current EU legislation.

Codex Legislation Relative to Milk and Dairy Products [144]. Of major importance in the international trade with milk products is the "Code of Principles concerning Milk and Milk Products". This code is presently adopted by more than 70 countries throughout the world. Chapters relevant for the dairy industry are the volumes 1 a "General Requirements", and volume 1 b: "Food Hygiene" [145]. Volume No 4 contains requirements for infant foods and milk-based infant formulas [146].

Table 21. Codex Alimentarius standards for specific milk products

Type of product	Standard	Codex reference
Fresh milk products	fermented milks	STAN 243-2003
	creams and prepared creams	STAN A-9
Preserved milk products	evaporated milks	STAN A-3
	sweetened condensed	STAN A-4
	milk and cream powders	STAN 207-1999
	whey powders	STAN A-15
	edible casein products	STAN A-18
Milk fat products	butter	STAN A-1
	butter oil	STAN A-2

In addition to these general Codex rules, some 15 individual Codex Standards concerning specific milk products (not including cheeses) have been issued (Table 21). The standards, although not all yet widely adopted in national legislation, are becoming increasingly important in the international trade with milk products. Objective of the standards is to define the identity of a product, and to provide composition data and data relevant for quality.

European Union Legislation. The EU issued a number of directives concerning the identity, the denomination of milk and milk products, and on the use of processing aids and food additives in such products.

Of particular importance are the directives no. 1898/87 [96] on the denomination of consumer milks, as well as the complementary directive 2597/97 [90] concerning consumer milks.

Proposed Directive C 231/20 [99] deals with evaporated, sweetened condensed milks, and powdered milks. The use of additives (others than colors and sweeteners) in evaporated or powdered milks and other milk products is regulated in Directive 95/2/EC [128].

12. Economic Aspects

World Milk Production [147]. Total world production of milk of different species has surpassed 500×10^6 t/a. Since 1987 annual production of *cow's milk* amounted to nearly 470×10^6 t/a. In a ten years' period, total milk production has moderately increased, while the number of dairy cows has, in the same period, significantly decreased.

While production in Western Europe is decreasing, that of North America, Australia and New Zealand was increasing. Also some Latin American countries such as Argentina and Brazil have markedly increased their milk production.

Another feature is the increase in buffalo milk production, mainly due to India, which lead to a share of $>10\,\%$ of the total world production in 1997 (Table 22). Production of milk from goat and sheep has remained stable at a level of about 10×10^6 t/a for goat's milk and 8×10^6 t/a for ewe's milk.

Industrial Transformation of Milk. Out of 470×10^6 t/a of cow's milk produced worldwide in 1997, ca. 60 % was processed in dairies. (This percentage, however, depends on whether

Table 22. World milk production by species [147, 160]

Species	Milk production, 10^6 t						Increase since 1992, %
	1992	1994	1996	1998	2000	2002	
Cow's	463	465	466	475	487	506	9.3
Buffalo	46	48	52	59	66	74	60.9
Sheep	8	8	8	8	8	8	0
Goat	10	10	10	12	12	12	20
Total	527	531	536	554	573	600	13.8

Table 23. Production of major milk products 1987 – 2002 [147, 149, 150, 158, 160]

	Milk supply to dairies, 10^6 t/a			Butter & butterfat, 10^6 t/a			Cheese production, 10^6 t/a			Whole milk powders (incl. half skimmed), 10^6 t/a			Skim milk powders, 10^6 t/a			Condensed and evaporated milk, 10^6 t/a		
Country	1987	1996	2002	1987	1996	2002	1987	1996	2002	1987	1996	2002	1987	1996	2002	1987	1996	2002
European Community	101.88	113.38	115.6	1.91	1.87	1.7	4.63	5.95	6.6	0.79	0.91	0.79	1.63	1.25	1.08	1.29	1.24	1.20
Poland	11.03	6.50	7.2	0.26	0.13	0.15	0.46	0.34	0.45	0.05	0.04	0.03	0.16	0.12	0.15	0.02		
Russia and Ukraine	74.90	24.14	17.0	1.67	0.44	0.36	0.86	0.31	0.44	0.27		0.11	0.79		0.15	0.63		
USA	63.64	69.97	76.5	0.50	0.53	0.61	2.42	3.25	3.90	0.07	0.06	0.02	0.49	0.48	0.72	0.27	0.21	0.26
Canada	7.59	7.18	7.4	0.09	0.09	0.07	0.25	0.27	0.33	0.01			0.10	0.06	0.08	0.10	0.10	0.06
Australia	5.95	9.00	11.3	0.10	0.15	0.14	0.18	0.25	0.33	0.06	0.11	0.22	0.13	0.22	0.21	0.04		
New Zealand	6.59	10.14	12.3	0.25	0.29	0.36	0.11	0.26	0.27	0.18	0.34	0.53	0.15	0.17	0.30	0.00		
Japan	7.25	8.54	8.3	0.07	0.09	0.09	0.08	0.13	0.13	0.03	0.02	0.02	0.16	0.20	0.18	0.06	0.05	0.04
Total	278.83	248.85	257.0	4.85	3.58	3.5	8.98	10.73	12.45	1.46	1.48	1.72	3.60	2.50	2.87	2.41	1.60	1.58

Table 24. International trade of whole milk powder in 1993 and 2000 [151, 160]

Exporting/importing country	1993		2000	
	Exporter	Importer	Exporter	Importer
Total exports ($\times 10^6$ t)	1.13		1.52	1.52
New Zealand (% of global)	29.3		25.9	
Australia	8.4		11.1	
Netherlands	20.2			
France	11.5			
Denmark	8.7			
UK	3.8			
European Union (EU)			37.9	
Others	18.1		25.1	
Algeria		12.4		6.3
S. Arabia		6.3		2.8
Malaysia		5.8		5.0
Taiwan		5.6		
Venezuela		4.8		4.6
Mexico		4.7		2.2
Brazil				7.1
Others		60.4		

production occurs in an industrialized or in a developing country. In industrialized countries >90 % of the production goes to the dairy processing plant, while the percentage can be very low in developing countries.) The remaining ca. 40 % of the milk production, not processed in dairies, are consumed as loose milk, are used by the nondairy food industry or go to animal feed.

Table 23 shows data for the quantities of industrially produced milk products during the period 1987 – 1997. The largest fraction of the milk intake leaves the dairies in form of consumer milks (pasteurized milks or others). Among the industrially transformed milk, cheese production is clearly the largest segment, followed by butter and butterfat produc-

Table 25. International trade of skim milk powder in 1993 and 2000 [151, 160]

	1993		2000	
Exporting/importing country	Exporter	Importer	Exporter	Importer
Total exports ($\times 10^6$ t)	0.9		1.28	
Australia (% of global)	26.6		19.8	
New Zealand	20.9		13.5	
USA	10		8.8	
Canada	3.3		2.5	
Germany	8.4			
Netherlands	5.7			
European Union (EU)			27.9	6.1
Others	25.1		27.5	
Mexico		20.5		9.2
Brazil				4.2
Malaysia		10.3		
Philippines		10		7.5
Japan		6.4		4.1
Algeria		6.4		3.8
Thailand		5.4		
Others		41		65

Table 26. International trade of butter and butter fat in 1993 and 2000 [151, 160]

	1993		2000	
Exporting/importing country	Exporter	Importer	Exporter	Importer
Total exports ($\times 10^6$ t)	0.9		0.74	
New Zealand (% of global)	38.7		45.6	
Australia	13.2		14.7	
USA	15.3		1.2	2.6
Netherlands	12.3			
France	3.7			
Finland	3			
European Union (EU)			23.7	14.1
Others	13.8		14.8	
Russia		17.8		7.3
Eastern Europe		12		
Egypt		7.6		6.6
Morocco				3.7
Mexico		6.6		4.6
Japan		4.3		
Others		52		58

tion. Milk powder production accounts for some 4×10^6 t/a.

Some of the largest private companies transforming milk into consumer products include: Nestlé (Switzerland), Dairy Farmers (USA), Unigate (UK), Danone, Besnier, Bongrain (all France), Friesland Coberco, Campina Melkunie (both Netherlands), MD Foods (Denmark), Parmalat (Italy), Snowbrand, Meiji and Morinaga (all three Japan).

The milk price paid to milk producers in 1997 varied from \$ 0.2 to \$ 0.6 per kilogram of whole milk, depending on country, an EU average being ca. \$ 0.35 per kilogram.

International Trade [148]. Shelf-stable milk products such as skim milk powders, whole milk powders, butter powders and other powdered products are exported into nondairy countries or countries with insufficient own milk production for the production of dairy products by recombining, as well as for other uses in the food industry. Other very important export products are cheeses and milk fat in form of AMF or butter oil. Data about exporting and importing countries for milk powders and butterfat are shown in Tables 24–26. Besides the EU, New Zealand and Australia are the major exporting countries in the world.

Consumption of Milk Products. Consumption data for 1996 (Table 27) illustrate the per capita consumption of liquid milk and major milk products in various regions of the

Table 27. Consumption of dairy products in 1996 [149]

	1996 per capita annual consumption, kg							
Product	EC (15)	U.S.A.	Mexico	Brazil	Argentina	Australia	New Zealand	Japan
Liquid milks total	75.1	101.8	54.9	72.4	57.4	108.1	91.9	40.2
Fermented milks	21.6	7.2	1.6		6.3	5.2		14.2
Cream (as butterfat equiv.)	3.5	3.7			0.5			0.5
Butter	4.5	2	0.3	0.5	1.4	3.2	8.1	
Cheese	16.3	13.7	4.5	2.6	10.9	3.2	9.2	
Total listed milk products	121	128.4	61.3	75.5	76.5	119.7	109.2	54.9

world. Since 1987, the consumption of liquid milk in industrialized countries such as the EU is stagnating, while that of fresh dairy products (yogurts, fresh cheeses and milk desserts) is increasing. Pasteurized milks lost market shares in favor of UHT milks and milk beverages.

On a world scale, consumption of butter remains roughly stabile, while that of cream shows some increase. While butter consumption via retail trade may have decreased, this loss was compensated by increased industrial consumption. Positive developments are seen for fresh dairy products and for cheese. Milk products consumption shows highest growth rates mainly in East and South Asia (4 – 5 % growth per year). There are also growing markets in Latin America.

13. Acknowledgement

Based on the article by Jehuda L. Stein and Kurt Imhof from the Fifth Edition of Ullmann's Encyclopedia

14. References

General References

1. G. H. Schmidt, L. D. Van Vleck, M. F. Hutjens: *Principles of Dairy Science*, Prentice-Hall Inc., Englewood Cliffs, New Jersey 1988.
2. Ch. Alais: *Science du Lait*, 4th ed., Sepaic Edition, Paris 1984.
3. A. Nienhaus (ed.): *Behr's Handbuch Milch 2*, Verlag B. Behr's GmbH, Hamburg 1985.
4. *Dairy Processing Handbook*, Tetra Pak Processing Systems AB, Lund, Sweden 1995.
5. E. Renner: *Lexikon der Milch*, Volkswirtschaftlicher Verlag GmbH, München 1988.
6. P. Walstra, R. Jennes: *Dairy Chemistry and Physics*, J. Wiley & Sons, New York 1984.
7. A. Toepel: *Chemie und Physik der Milch*, 2nd ed., VEB-Fachbuchverlag, Leipzig 1981.
8. N. P. Wong, R. Jennes, M. Keeney, E. H. Marth (eds.): *Fundamentals of Dairy Chemistry*, 3rd ed., Van Nostrand Reinhold Co, New York 1988.
9. P. F. Fox (ed.): *Developments in Dairy Chemistry*, vols. **1 – 3**, Elsevier Applied Science Publishers, London, New York 1982 1985.
10. E. Renner: *Milch und Milchprodukte in der Ernährung des Menschen*, Volkswirtschaflicher Verlag GmbH, München 1982.
11. S. W. Souci, W. Fachmann, H. Kraut: *Die Zusammensetzung der Lebensmittel. Nährwerttabellen*, Wissenschaftliche Verlagsgesellschaft mbH, Stuttgart 1979.
12. H. G. Kessler: *Lebensmittel- und Bioverfahrenstechnik, Molkereitechnologie*, 4th ed., Verlag A. Kessler, Freising 1996.
13. H. G. Kessler: *Food Engineering and Dairy Technology*, Verlag A. Kessler, Freising 1981.
14. E. Spreer: *Technologie der Milchverarbeitung*, Behr's Verlag, Hamburg 1997.
15. R. K. Robinson (ed.): *Modern Dairy Technology*, vols. **I and II**, Elsevier Applied Science Publishers, Barking, UK 1986.
16. R. Early (ed.): *The Technology of Dairy Products*, Blackie Academic and Professional VCH Publishers, New York 1994.
17. W. Varro: *Technisches Handbuch für die Molkereipraxis*, vol. **71/72**, "Molkereitechnik," Th. Mann Verlag, Gelsenkirchen 1986.
18. R. G. Jensen (ed.): *Handbook of Milk Composition*, Academic Press, London 1995.
19. *Wissenschaftliche Tabellen Geigy*, 8th ed., Ciba-Geigy, Basel 1979.
20. *Ullmann*, 5th ed., **A16**, 589 – 639.

Specific References

21. IDF Bulletin No 137, "Buffalo as a candidate for milk production", International Dairy Federation (IDF), Brussels 1981.

22. IDF Bulletin No 140, "The composition of ewe's and goat's milk", International Dairy Federation (IDF), Brussels 1981.
23. E. Schlimme, W. Buchheim: *Milch und ihre Inhaltsstofe*, Verlag Th. Mann, Gelsenkirchen 1995.
24. P. Fox (ed.): *Advanced Dairy Chemistry-2 (Lipids)*, Applied Science Publishers, London – New York 1995.
25. E. Schlimme: *Kompendium zur milchwirtschaftlichen Chemie*, Volkswirtschaftlicher Verlag, München 1990.
26. R. G. Jensen, R. M. Clark, *J. Dairy Sci.* **74** (1991) 3228 – 3243.
27. IDF: *Protein Definition*, Proceedings of the first IDF Symposium held in Minneapolis, Oct 1993, Brussels 1994.
28. S. J. Rowland, *J. Dairy Sci.* **9** (1938) 47 – 57.
29. Alpura AG, CH 284 061, 1952.
30. FR 1 075 502, 1954 (M. J. Ciboit).
31. C. R. Harishorst, *Food Ind.* **7** (1951) 72 – 74.
32. H. Hostettler, *Schweiz. Milchztg.* (1961) nos. 42 and 76.
33. D. K. Henyon, *Int. J. Dairy Technol.* **52** (1999) 95 – 101.
34. W. N. Eigel (chaiman), "Nomenclature of Proteins of Cow's Milk: 5th Revision", *J. Dairy Sci.* **67** (1984) 1599 – 1631.
35. P. F. Fox (ed.): *Advanced Dairy Chemistry-1 (Proteins)*, Elsevier Applied Science, London – New York 1992.
36. IDF Bulletin No 238, "Heat-induced changes in milk", International Dairy Federation (IDF), Brussels 1989.
37. L. K. Creamer, A. K. H. McGibbon, *Int. Dairy J.* **6** (1996) 539 – 568.
38. B. Ribadeau-Dumas, *Rev. Lait. Franç.* **400** (1981) 17 – 32.
39. B. C. Richardson, *N. Z. J. Dairy Sci. Technol.* **18** (1983) 233 – 245.
40. L. Pellegrino, A. Tirelli, F. Mascotti, P. Resmini: "Heat treatments and alternative methods", in *IDF Proceedings of the IDF Symposium* Vienna, Sept.1995 International Dairy Federation (IDF), Brussels 1996, p. 373.
41. A. Wolfschoon-Pombo, H. Klostermeyer, *Milchwissenschaft* **36** (1981) 598 – 600; 721 – 724.
42. P. F. Fox (ed.): *Developments in Dairy Chemistry-3 (Lactose and minor constituents)*, Elsevier Applied Science Publishers, London – New York 1985.
43. C. A. Barth et al.: *Neue Entwicklungen bei den Sauermilchprodukten, SMK Schrift No 2*, Verlag Schweizerische Milchkommission, Liebefeld – Bern 1989.
44. IDF Monograph No 9101, "Residues and contaminants in milk and milk products", International Dairy Federation (IDF), Brussels 1991.
45. Research Triangle Institute, US 4 175 141, 1977 (D. M. Adams, J. T. Barch, M. L. Speck).
46. IDF Bulletin No 319, "Bovine Somatotropin", International Dairy Federation (IDF), Brussels 1997, p5.
47. M. Houska (ed.): *Milk, Milk Products and Semiproducts. Thermophysical and Rheological Properties of Foods*, Institute of Agriculture and Food Information, Prague 1994.
48. T. H. M. Snoeren, A. J. Damman, H. J. Klok, *Neth. Milk Dairy J.* **36** (1982) 305 – 316.
49. R. K. Robinson (ed.): *Dairy Microbiology*, vol **1 and 2**, Applied Science Publishers, London – New York 1983.
50. T. A. Roberts, J. I. Pitt, J. Farkas, F. H. Grau (eds.): *Microorganisms in Foods*, vol **6**, ICMSF, Blackie Academic & Professional, London 1998.
51. P. C. Vasavada: "Pathogenic Bacteria in Milk (Review)", *J. Dairy Sci.* **71** (1988) 2809 – 2816.
52. Food and Nutrition Board, National Academy of Sciences — National Research Council: "Recommended Daily Dietary Allowences", Revised 1980.
53. FAO/WHO Tables: "Amino acid content of Foods", FAO/WHO, Rome 1976.
54. Joint FAO/WHO Expert Consultation of Protein Quality Evaluation by amino acid scoring procedure, Bethesda, MD 1989, FAO/WHO, Rome 1991.
55. J. C. Dillon, *Cah. Nutr. Diét.* **27** (1992) 54 – 58.
56. IDF Bulletin No 222, "Milk Products and Health", International Dairy Federation (IDF), Brussels 1988.
57. R. E. Jensen in C. K. Chow (ed.): *Fatty Acids and Their Health Implications*, Marcel Dekker, Basel 1992.
58. D. Precht, J. Molketin, *Die Nahrung* **39** (1995) 343 – 374.
59. IDF Bulletin No 225, "Cultured Dairy products in Human Nutrition", International Dairy Federation (IDF), Brussels 1991.
60. D. Rheinhardt, E. Schmidt (eds.): *Food Allergy, Nestlé Nutrition Workshop Series*, vol. **17**, Raven Press, New York 1988.
61. A. Höst, *Pediatr. Allergy Immunol.* **5** (1994) Suppl. 5, 683 – 688.

62. T. J. David: *Food and Food Additive Intolerance in Childhood*, Blackwell Scientific Publications, Oxford 1993.
63. J. M. Martin, B. Trink, D. Danemann, H. M. Dosch, B. Robinson, *Annals of Medicine* **32** (1991) 447 – 452.
64. *Nutr. Rev.* **46** (1988) 215 – 217.
65. E. C. Reynolds, A. Del Rio, *Archs oral Biol.* **29** (1984) 927 – 933.
66. J. R. Neeser, A. Chambaz, S. Del Vedovo, M. J. Prigent, B. Guggenheim, *Infection and Immunity* **56** (1988) 3201 – 3208.
67. IDF Bulletin No 255, "Dietary Calcium and Health", International Dairy Federation (IDF), Brussels 1991.
68. IDF Bulletin No 320, "Implications of Microfiltration on Hygiene and Identiy of Dairy Products", International Dairy Federation (IDF), Brussels 1997.
69. H. Burton: *UHT Processing of Milk and Milk Products*, Elsevier Applied Science, London – New York 1988.
70. IDF Bulletin No 200, "Monograph on Pasteurized Milk", Brussels 1986.
71. IDF Bulletin No 271, "Monograph on the Pasteurization of Cream", Brussels 1992.
72. A. Nangpal, H. Reuter, *Kieler Milchwirtschaftl. Fortschrittsber.* **42** (1990) 31 – 41.
73. P. F. Fox (ed.): *Heat Induced Changes in Milk*, 2nd ed., IDF, Brussels 1995.
74. D. B. Shah, J. G. Bradshaw, J. T. Peeler, *J. Food. Sci.* **56** (1991) 391 – 393.
75. R. Behringer, H. G. Kessler, *Int. Dairy J.* **2** (1992) 83 – 93.
76. Council Directive 92/46/EEC of July 16 1992, Council of the European Communities, Brussels 1992.
77. J. Mauron, *Prog. Food Nutr. Sci.* **5** (1981) 5 – 35.
78. P. A. Finot, R. Deutsch, E. Bujard, *Prog. Food Nutr. Sci.* **5** (1981) 345 – 355.
79. R. Van Renterghem, J. De Block, *Int. Dairy J.* **6** (1996) 371 – 382.
80. L. Pellegrino, I. De Nomi, P. Resmini, *Int. Dairy J.* **5** (1995) 647 – 659.
81. P. J. de Koning, P. J. Van Rooijen, *J. Dairy Res.* **49** (1982) 725 – 736.
82. R. J. Fritsch, H. Klostermeyer, *Z. Lebensm. Unters. Forsch.* **172** (1981) 440 – 445.
83. F. Dannenberg, H. G. Kessler, *J. Food Sci.* **53** (1988) 258 – 263.
84. R. Brommage, M. A. Juillerat, R. Jost, *Le Lait* **71** (1991) 173 – 180.
85. P. Walstra, *Neth. Milk Dairy J.* **29** (1975) 279 – 294.
86. H. A. Kurzhals, *Milchwissenschaft* **28** (1973) 637 – 645.
87. A. W. M. Sweetsur, D. D. Muir, *J. Soc. Dairy Technol.* **35** (1982) 120 – 126.
88. M. Cheryan: *Ultrafiltration and Microfiltration Handbook*, Technomic Publishing Inc., Lancaster – Basel 1998.
89. IDF Bulletin No 311, "Advances in Membrane Technology for Better Dairy Products", International Dairy Federation (IDF), Brussels 1996.
90. Council Directive no. 2597/97 of December 18 1997, Council of the European Communities, Brussels 1997.
91. W. Buchheim, W. Heeschen, E. Schlimme, *Eur. Dairy Magazine*, (1994) March, 42 – 44.
92. IDF Standard 82A, "Detection of Phosphatase Activity", International Dairy Federation (IDF), Brussels 1987.
93. IDF Standard 155, "Determination of Alkaline Phosphatase Activity, Fluorimetric Method", International Dairy Federation (IDF), Brussels 1992.
94. P. Russel: *Developments in the Packaging of Milk, Liquid Dairy Products and Fruit Juices*, Pira International, Surrey 1998.
95. W. Sturm: *Verpackung Milchwirtschaftlicher Lebensmittel*, Edition IMQ, Kempten 1998.
96. Council Directive No 1898/87 of July 2 1987, Council of the European Communities, Brussels 1987.
97. H. J. Klupsch: *Saure Milcherzeugnisse, Milchmischgetränke und Desserts*, Verlag Th. Mann, Gelsenkirchen 1984.
98. O. F. Hunziker: *Condensed Milk and Milk Powder*, Published by the Author, La Grange, IL 1949.
99. Council Directive 96/C231/05, Proposition from May 30 1998, Council of the European Communities, Brussels 1996.
100. Council Directive 76/118, Brussels 1976.
101. J. Pisecki: *Handbook of Milk Powder Manufacture*, Niro Als, Copenhagen 1997.
102. IDF International Standard 87, "Determination of the Dispersibility and Wettability of Instant Dried Milk", International Dairy Federation (IDF), Brussels 1979.
103. Joint IDF/ISO Provisional Standard, "Determination of insolubility index", International Dairy Federation (IDF), Brussels 1985.
104. Codex Alimentarius, "Draft revised Standard for evaporated milks", Alinorm 97/11, Appendix IV. FAO/WHO Rome 1997.

105. IDF Provisional Standard 114, "Dried milk, assessment of heat class", International Dairy Federation (IDF), Brussels 1982.
106. American Dry Milk Institute "Standards for grades of dry milk", Chicago 1971.
107. A. Sjollema, *Neth. Milk Dairy J.* **42** (1988) 365 – 374.
108. IDF Bulletin No 212, "Trends in Whey Utilization", International Dairy Federation (IDF), Brussels 1987.
109. T. Siecievicz, C. L. Riedel: *Whey and Whey Utilization*, Th. Mann Verlag, Gelsenkirchen 1990.
110. R. Jost, *Trends Food Sci. Technol.* **4** (1993) 283 – 288.
111. IDF Monograph No 9201, "New applications of membrane processes", International Dairy Federation (IDF), Brussels 1992.
112. IDF Standard No 45, "Compositional standard for casein (edible and industrial)", International Dairy Federation (IDF), Brussels 1969.
113. IDF Standard No 72, " Edible caseinate: compositional standard", International Dairy Federation (IDF), Brussels 1974.
114. FAO/WHO Codex Norm A-14, FAO/WHO, Rome 1995.
115. C. R. Southward, *New Zealand J. Dairy Sci. Technol.* **20** (1985) 79 – 101.
116. C. V. Morr, E. Y. W. Ha, *Crit. Rev. Food. Sci. Nutr.* **33** (1993) 431 – 476.
117. J. N. de Wit, G. Klarenbeek, M. Adamse, *Neth. Milk Dairy J.* **40** (1986) 41 – 56.
118. D. E. Palmer, *Process. Biochem.* (1977) June, 24 – 28.
119. B. Mirabel, R. Goudal, *Revue Laitière Française* (1981) no. 400, 77 – 81.
120. Codex Alimentarius Commission CAC/RS 72/74, FAO/WHO 1976.
121. European Society for Paediatric Gastorenterology and Nutrition (ESPGAN), *Acta Paediatrica Scandinavica* **287** (1981) Suppl.
122. Food and Drug Administation: "US Infant formula Act", *Federal Register* **50** (1985) 45106 – 45108.
123. R. Jost, J. C. Monti, J. J. Pahud in M. Friedman (ed.): *Nutritional and Toxicological Consequences of Food Processing*, Plenum Press, New York – London 1991.
124. J. L. Rasic, J. A. Kurmann: *Yoghurt*, Published by the authors, Printed in Switzerland by Staempfli & Cie AG, Bern 1978.
125. IDF Bulletin No 227, "Fermented milks-Science & technology", International Dairy Federation (IDF), Brussels 1988.
126. R. Fuller (ed.): *Probiotics, the Sientific Base*, Chapman & Hall, London 1992.
127. C. Daly, G. F. Fitzgerald, L. O'Connor, R. Davis, *Int. Dairy J.* **8** (1998) 195 – 205.
128. European Council Directive No 95/2/EC of Feb. 20 1995.
129. H. J. Klupsch, *Dtsche. Molkerei Ztg.* **106** (1985) 292 – 296.
130. E. G. Samuelsson, P. Ulrich in IDF (ed.): *Proceedings of the 21st Int. Dairy Congress*, vol. **1**, Mir Publishers, Moscow 1982.
131. A. Fik, *Dtsche. Milchwirtschaft* **38** (1987) 1134 – 1139.
132. H. Eibel, H. G. Kessler, *Dtsche. Molkerei Ztg.* **8** (1986) 176 – 179.
133. C. M. Bruhn, J. C. Bruhn, *J. Dairy Sci.* **71** (1988) 857 – 862.
134. J. Koops, *Neth. Milk Dairy J.* **21** (1967) 29 – 49.
135. S. Geyer: "Zur Flockungsstabilität homogenisierter UHT Sahne in Kaffee und deren technologische Beeinflussung", Thesis. TU München-Weihenstephan 1987.
136. IDF Standard 68 A, "Anhydrous milk fat, anhydrous butteroil or anhydrous butterfat, butteroil or butterfat, ghee: standards of identity", International Dairy Federation (IDF), Brussels 1977.
137. H. R. Lehmann, E. Doller, A. Uphus, *Technical-Scientific documentation No 9*, Westfalia Separator Company, Oelde 1988.
138. C. Versteeg, *CSIRO Food Research Quarterly* **51** (1991) 32 – 54.
139. J. Arul, A. Bourdreau, J. Makhlouf, R. Tardif, B. Grenier, *J. Dairy Res.* **55** (1988) 361 – 371.
140. IDF Bulletin No 260, "Utilisations of milk fat", International Dairy Federation (IDF), Brussels 1991.
141. E. Frede, B. Buchheim, *J. Soc. Dairy Technol.* **47** (1994) 17 – 27.
142. IDF Bulletin No 312, "Inventory of IDF/ISO/AOAC International adopted methods of analysis and sampling for milk and milk products," International Dairy Federation (IDF), Brussels 1996.
143. IDF Bulletin No 331, "Payment Systems for Ex-Farm Milk", International Dairy Federation (IDF), Brussels 1998.
144. IDF Bulletin No 331, "Codex procedures and their importance", International Dairy Federation (IDF), Brussels 1998.
145. Codex Alimentarius, vol. 1, A General Requirements, 1B Hygiene, 2nd ed., FAO/WHO, Rome 1995.

146. Codex Alimentarius, vol. 4, *Foods for special dietary uses (including foods for infants and children) 2nd ed.,* FAO/WHO, Rome 1994.
147. IDF Bulletin No 333, "World Dairy Situation 1998", International Dairy Federation (IDF), Brussels 1998.
148. IDF Bulletin No 325, "Major changes in the international dairy trade", International Dairy Federation (IDF), Brussels 1997.
149. IDF Bulletin No 323, "World Dairy Situation 1997", International Dairy Federation (IDF), Brussels 1997.
150. IDF Bulletin No 243, "World Dairy Situation 1989", International Dairy Federation (IDF), Brussels 1989.
151. IDF Bulletin No 316, "The Dairy World-twenty-five years of change 1985 – 2010", International Dairy Federation (IDF), Brussels 1996.
152. E. J. Mann, *Dairy Ind. Int.* **44** (1979) no. 6, 21 – 22.
153. E. J. Mann, *Dairy Ind. Int.* **48** (1983) no. 4, 27 – 37.
154. E. C. Synnott, "Dairy Effluents," FIL Bull. no. 184, Brussels 1984.
155. A. D. Wheatley, *Biotechnol. Genet. Eng. Rev.* **1** (1984) 261 – 309.
156. F. Wagner, *Dtsch. Milchwirtsch. (Gelsenkirchen Ger.)* **36** (1985) 697 – 700.
157. J. Immerz, K. Guthy, *Dtsch. Molk. Ztg.* **106** (1985) 1110 – 1116, 1138 – 1144.
158. IDF Bulletin No 339, "World Dairy Situation 1999", International Dairy Federation (IDF), Brussels 1999.
159. D. J. Raiten, J. M. Talbot, J. M. Waters, *J. Nutrition* **128** (1998) no. 5, 2059 – 2293.
160. IDF Bulletin No 391, "World Dairy Situation 2004", International Dairy Federation (IDF), Brussels 2004.

Mirrors

Mark P. Taylor, Corning Inc., Corning, New York 14831, United States (Chaps. 1 – 4)

Rudolf Seitz, Flabeg GmbH, Fürth, Federal Republic of Germany (Chap. 5)

1. History and Definitions 1
2. Substrate Materials and Forming Techniques 2
2.1. Materials 2
2.2. Forming Techniques 4
3. Grinding and Polishing 5
4. Coatings 6
5. Production of General-Purpose Mirrors 6
6. References 7

1. History and Definitions

Mirrors provide humans with a unique ability to control and direct light by reflection from a surface. In this article, the development and function of mirrors will be described.

Cosmetic and General-Use Mirrors. Low-cost mirrors such as those used in home furnishings and automobiles are known as *second surface mirrors.* They are produced by application of a chemical silvering agent or evaporating a reflective coating (usually aluminum) on the surface of large sheets of flat glass (Chap. 5). The mirrors are then cut using diamond scribes. The glass is mounted with the coated surface to the back so that the coating comes from the back or second surface. These mirrors perform acceptably, although the clarity of the reflected image is reduced by reflections from the first glass surface. These reflections account for ca. 4 % of the incoming light compared to about 90 % from the coated surface.

A *first surface mirror* has the reflective coating on the outer or first surface of the glass substrate. Such mirrors are used where higher accuracy is demanded. Since all the reflection comes from this surface, only a single image is formed resulting in excellent clarity.

Flat mirrors simply reflect the rays of light arriving at their surface without changing their relative directions. *Curved mirrors* can be used to focus, spread, or collimate a bundle of rays. They are used in imaging systems ranging from flashlight reflectors to lasers to large telescopes. Substrates for curved mirrors may be glass, metal, or plastic. Whereas glass mirrors can be sagged to shape, metal and plastic mirrors are produced by depositing the substrate materials on a curved master, then coating with a highly reflective material. Designs often include a housing to protect the coated surface from contamination and damage.

Laser Mirrors. The development of lasers and laser optics has increased standards for high-quality finishing and coating. Ring laser gyros demand surface roughnesses of 0.1 – 0.2 nm and multilayer dielectric coatings producing reflectivities > 99 %. These mirrors measure only about 5 mm in diameter and are machined from a solid disk of β-quartz glass ceramic. Their supporting structure is optically contacted to the laser frame and the surface is activated by piezoelectric controls.

Scattering from the surface of laser mirrors is a particular problem, since the laser must maintain a very stable standing wave. This places special demands on the procedures used in grinding and polishing to eliminate subsurface fractures. Coatings must be very uniform and designed to withstand the high heat load generated by the laser.

Mirrors in Astronomy. The most spectacular advances in mirror design have come from the most demanding of mirror users, the astronomers. The art and science of producing mirrors was revolutionized with the invention of the reflecting telescope by Sir Isaac Newton in 1668. Previously, all telescopes were refracting, i.e., they used glass lenses and prisms to collect and focus light. These had serious limitations because of the mass of the lenses, the great length of the tube required, chromatic aberration (inability to focus all wavelengths at a single point)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_641

and the need for high-purity, low-distortion optical glass.

NEWTON's invention addressed two needs of astronomers that are still important today: (1) high light gathering power and (2) low distortion of the arriving light wave by the optical apparatus and the surrounding atmosphere. Great light gathering power is needed to collect the faint light from a distant object and is achieved by using large-diameter primary mirrors, high-reflectivity coatings, and sensitive detectors. Low distortion of the light wave is needed to achieve high image resolution. The full aperture of the primary mirror must maintain a precise curved shape, even under thermal variations. The compactness and practical execution of NEWTON's design allowed the apparatus to be located on remote mountain peaks, well above disturbances of the atmosphere.

The eighteenth century (1789) saw the production of the first large (48-inch) metal telescope mirror in England under the direction of WILLIAM HERSCHEL. This mirror weighed ca. 1000 kg and was contained in a tube 12 m long. It proved to be difficult to use because it contained no tracking mechanism.

Development of Glass Mirrors. The major disadvantage of metal mirrors was their tendency to distort with temperature due to a relatively large thermal expansion coefficient. Silicate glass has a much lower thermal expansion coefficient and can be polished to higher finishes than metals, but it does not have a high reflectivity. The invention of the silver coating of glass by reduction of silver nitrate revolutionized the design of mirrors. This process was credited to KARL STEINHEIL, a German physicist, in 1856. In 1857 the French physicist FOUCAULT independently announced his invention of a process for figuring (grinding to shape) and coating with a silver solution. Major advances in the technology of telescope mirrors are summarized in Table 1.

Parts of a Mirror. Figure 1 illustrates the parts of a lightweight mirror. The concave *faceplate* supports the reflective coating. The faceplate is bonded to a lightweight glass *core* to increase the stiffness. The core is mounted on a glass *backplate*. The faceplate is ground to a spherical or paraboloidal shape and focuses light from an infinite object at the focal length f, which is related to the radius of curvature R by $f = R/2$. A *spherical shape* is easier to obtain by standard grinding techniques, but forms an accurate image only at small angles where the spherical surface closely approximates a paraboloid (the paraxial region). This means that the mirror will be relatively flat and that long focal lengths are needed.

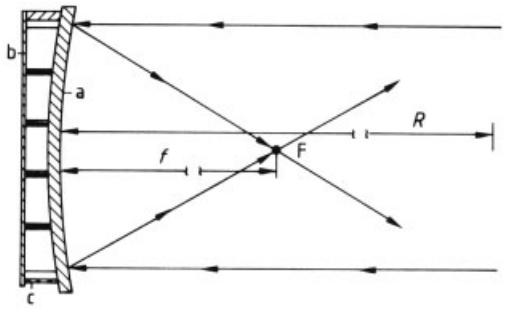

Figure 1. Lightweight mirror design
a) Concave faceplate; b) Lightweight fused-strut core; c) Planar backplateA spherical mirror focuses rays of incoming parallel light at point F which is at a distance one half the radius of curvature, R.

A *paraboloidal shape* forms much sharper images because rays from all points of the mirror focus sharply to the same point. Paraboloidal mirrors are, however, more difficult to produce because the radius of curvature is different for each point on the mirror.

The core is composed of cast or fused struts of material with the same thermal expansion as the faceplate. The backplate may be flat or curved, it provides the reaction surface to carry the load of the faceplate and core as well as the mounting points for joining the mirror to the mounting structure. In an *active mirror*, the core is replaced by a network of actuators which constantly monitor and adjust the surface of the mirror.

2. Substrate Materials and Forming Techniques

2.1. Materials

The mirror substrate material must meet two primary requirements. Firstly, its mechanical and

Table 1. Milestones in the development of telescope mirrors

Mirror	Material ∗	Size, m	Significant technology
NEWTON's mirror	metal	0.1	first reflector telescope, 1668
HERSCHEL's mirror	metal	1.2	first large mirror, 1789
Hale telescope	Pyrex glass	5.08	first large ribbed mirror, 1934
McDonald observatory	fused silica glass	2.7	large low expansion glass mirror produced by hex fusion, 1959
Hubble space telescope	ULE glass	2.4	first large fusion bonded lightweight, 1979
Monacle relay mirror	ULE glass	1.0	first ultralightweight frit bonded mirror, 1984
Keck Observatory	Zerodur glass ceramic	10	first large segmented aspheric mirror (in polishing)
Columbus telescope	borosilicate glass	8	spin-cast lightweight asphere (in planning)

∗ Pyrex, ULE, and Zerodur are registered trade marks; ULE = ultralow expansion.

thermal properties must ensure minimum distortion from external influences. These properties must be homogeneous throughout the mirror body and constant over the expected service life of the mirror. Glass and glass ceramics provide excellent stiffness and display elastic behavior over their entire stress range. Secondly, the materials must be capable of being polished to a surface roughness measured in tenths of nanometers and be capable of accepting and holding an evaporated reflective surface. Single-phase optical glass is ideal for this because it is homogeneous and can be heated to several hundred degrees Celsius for coating. Glass ceramics can also be polished to a high finish provided that the residual glass component has similar mechanical and chemical characteristics to the crystalline phase.

Thermal and mechanical properties of some common mirror substrate materials are shown in Table 2.

Borosilicate Glass (→ Glass, Chap. 2.2.4.). Borosilicate glass (Pyrex) was invented in the early 1900s for use in railroad lamps. Its low thermal expansion prevented breakage when rain splashed on the hot surface of the glass. It is a relatively difficult glass to form when compared to soda lime compositions (window glass), but proper compositions resist devitrification (crystallization) or phase separation. Borosilicate glass is relatively easy to finish and is inexpensive. Because of its high thermal expansion, it is used in instrument mirrors that do not require high precision.

Silica Glass (→ Glass, Chap. 2.2.1.). Silica glass is pure SiO_2 and is referred to as fused silica, vitreous silica, fused quartz, or just quartz. It may be produced by melting natural quartz crystals or beach sand, but the very best optical-quality glass is produced by vapor phase deposition in which $SiCl_4$ gas is oxidized in a high-temperature burner and deposited on a substrate. Mirror materials are cut from these large boules of very pure glass.

Silica glass is usually the material of choice for infrared mirrors because its thermal expansion is very low under cryogenic conditions. Because of its simple formula and amorphous structure, its thermal expansion is very consistent and predictable.

ULE Glass. Ultralow expansion glass (ULE) was invented in the early 1940s when researchers at Corning found that substituting about 7.5 wt % TiO_2 into SiO_2 reduced the thermal expansion from 0.5 to $< 0.03 \times 10^{-6}\,K^{-1}$. Due to its high viscosity, it must also be synthesized by vapor phase deposition. This process allows very precise control of the composition, and hence the thermal expansion coefficient.

The thermal expansion and the elastic modulus are both correlated to the titania content. The thermal expansion of the material can be measured nondestructively at any location in the boule by measuring the speed of sound. Boules may therefore be selected and matched for thermal expansion in very critical applications.

Glass Ceramics. Glass ceramics are materials which can be formed using conventional glass-forming technologies such as casting,

Table 2. Mechanical and thermal properties of some common mirror materials

Material	Thermal expansion coefficient (25 °C), 10^{-6} K	Elastic modulus, GPa	Thermal conductivity, $W\,m^{-1}\,K^{-1}$
Titanium silicate glass (ULE)	0.00 ± 0.03	67.6	1.3
LAS glass ceramic (Corning 9600)	− 0.01	95.3	1.5
LAS glass ceramic (Zerodur)	− 0.03	91.0	1.6
Silica glass (Corning 7940)	0.52	73.1	1.3
Borosilicate glass (Pyrex, Corning 7740)	3.2	62.7	1.1
Molybdenum	5.0	324	146
Beryllium	11.5	303	182
Nickel	15.3	204	74.9
Aluminum	23.4	68.9	222

blowing, or pressing, but are then reheated and crystallized to yield a low thermal expansion, highly crystalline body (→ Glass Ceramics).

Schott Glaswerke (Mainz, FRG) has extensively developed β-quartz glass ceramics (Zerodur) for telescope mirrors. This required solving the vexing problem of how to limit the stress induced by volume change during crystallization. Very long crystallization cycles and retention of a phosphorous-rich residual glass phase has permitted crystallization of spun-cast plates up to 4 m in diameter.

Glass ceramics are particularly well suited for ring laser gyro mirrors where low thermal expansion, transparency, and impermeability to helium are of paramount importance. Corning Code 9600 has been optimized for this application by reducing the thermal expansion to under $0.01 \times 10^{-6}\,K^{-1}$. Its very small crystal size results in low scattering and permits polishing to less than 0.1 nm root mean square surface roughness.

Metals and Composites. The development of glass mirrors practically eliminated the use of metal mirrors in precision telescope applications. Nevertheless, metal mirrors continue to be used in parabolic reflectors and in high-energy laser mirrors which require cooling. Nickel is used in many parabolic reflectors because it can be chemically deposited on a master form. Once deposition is complete, the thick layer of nickel is removed intact from the master resulting in a finished mirror which replicates the shape of the master form. Molybdenum has been used extensively for high-energy mirrors, but beryllium is becoming increasingly popular because of its low density and high thermal conductivity. A disadvantage of beryllium is its high toxicity which demands special machining precautions.

Aluminum reinforced with silicon carbide whiskers shows some promise for mirrors; the whiskers increase the modulus of elasticity and reduce thermal expansion. Carbon fibers have near-zero thermal expansion, excellent thermal conductivity, and very high mechanical stiffness. These fibers may be combined with epoxy, glass, or glass ceramic matrices to produce high-strength, high-toughness composite mirror substrates with very low thermal expansion. Although the composite material cannot be polished well enough to produce a high-quality surface, polishable surface materials can be applied as laminates. For example, ULE or silica glass faceplates can be bonded to a composite back structure.

2.2. Forming Techniques

Monolithic Casting and Hex Fusion. *Monolithic casting* is rarely used now for large mirrors. It involves directly filling a mold made from refractory bricks with molten glass by ladling or by tapping a glass tank. The glass must then be cooled quickly enough to avoid

devitrification, reheated to the annealing temperature, and very slowly cooled to eliminate thermal stress. If a glass ceramic is used, the annealed glass must be crystallized in a high-temperature thermal cycle to develop the desired low expansion phase. Schott has used centrifugal casting to produce 4-m meniscus blanks and plans a similar approach for the 8-m blanks for the European Southern Observatory.

Hex fusion was developed for large silica glass or ULE castings that cannot be poured or ladled into a mold. Small (ca. 1 m diameter) boules are cut into hexagonal tiles and carefully fitted together in a mold. The glass is heated and fuses together into a large mass. This approach has the advantage that each boule can be inspected before it is committed to the mirror blank.

Lightweight Centrifugal Casting. The Stewart Observatory (University of Arizona) has pioneered the technique of centrifugal casting of structured mirrors. A mold is constructed using low-density fibrous refractory blocks with the areas which will become ribs appearing as narrow gaps between the blocks. A gap between the mold back and the blocks also allows glass to flow under the blocks forming the backplate.

Borosilicate glass cullet (first-sized broken blocks) is loaded into the cold mold. Heat is applied and when the glass begins to soften, the entire apparatus is spun at about 8 rpm. Spinning causes the surface of the glass to assume a parabolic shape. Once the glass cools below the settling temperature, spinning is stopped. Residual stresses are annealed out of the glass blank by very slow cooling.

The Stewart group has also reported success in spinning 0.5 m disks of Corning Code 9600 glass ceramic; this approach could combine the advantages of a low thermal mass design and a low expansion material.

Ultralightweight Frit Bonding. This approach constructs the lightweight core by welding together thin slabs of precut ULE glass. The frit, a powdered glass ceramic designed to match the thermal expansion of ULE, is suspended in an organic vehicle and applied to the ends of the core struts. The faceplate and backplate are then placed on the core and the entire structure is heated to the sintering temperature of the frit. The frit forms a fillet weld between the core and front and back surfaces and is stronger than the ground glass itself.

Frit-bonded mirrors are attractive for space-based instruments where launch weight is critical. The mirror can be built up step by step and inspected at each critical juncture.

3. Grinding and Polishing

Grinding. After the blank has been produced and annealed, it is placed on a machine which grinds the surface into the desired radius of curvature. The grinding tool typically consists of fixed abrasives such as diamond or silicon carbide buttons on a metal surface which has been machined to the same radius of curvature as the mirror.

Grinding must be performed very slowly to avoid damaging the glass surface and may require months or years for a very large mirror. Since heat can distort the grinding tool, grinding is often carried out in temperature-stabilized environments.

Very high thermal "spikes" can occur at locations where the diamond cuts the substrate. This can lead to thin layers of stressed glass on the ground surface which must be removed by an acid etch treatment.

Polishing. The ground, shaped mirror blank is moved to a polishing machine. Starting with coarse grits, the blank is moved through progressively finer grinding materials. Typical polishing materials include diamond, silicon carbide, aluminum oxide, and cerium oxide for the final polish. They are either suspended in a liquid that is applied to the surface or contained on fabric tool surfaces.

During finishing, the optician must often check the surface curvature or *figure* of the mirror with an interferometer to ensure that it corresponds to the desired spherical or parabolic shape. The final quality of the mirror is usually determined by this deviation from the ideal figure. High-quality optics have deviations of less than a tenth of the wavelength of light across their aperture.

Parabolic shapes are particularly desirable for very large aperture telescope mirrors since they can reduce the focal length of the mirror. These

so-called "fast" mirrors require very special polishing tools and techniques because the lap (polishing tool) must fit the surface of the mirror to within about 3 µm, but the radius of curvature is different for each point on the mirror's radius. To solve this, the Stewart Observatory Mirror Laboratory is developing deformable laps which can change their radius of curvature on each stroke. Itek is stressing the mirror during polishing for the segments of the Keck parabola. The European Southern Observatory's 3.8 m New Technology Telescope has employed a pressure-controlled lap.

4. Coatings

The role of the massive glass faceplate and support structure, actuators, and mounting apparatus is to support the few grams of coating material where photons of light are reflected.

Coatings are usually applied by *vacuum deposition*. The entire mirror substrate is first scrupulously cleaned to guarantee adhesion of the coating. It is then placed in a vacuum chamber and the coating is applied by heating a metal or oxide target with an electron beam to deposit a thickness of a few hundred nanometers on the substrate surface. During use, coatings may tarnish or become scratched. Most metal coatings can be stripped using acids and a fresh surface can be applied.

Metal Coatings. Metals are the standard coatings for mirrors because of their high reflectivity and ease of application. The preferred metal coating for applications in the visible and near infrared is aluminum. Although it does not have the reflectivity of silver, aluminum is less prone to tarnishing. Fresh aluminum coatings provide reflectivities of about 90 % over most of the visible whereas silver may be as high as 98 %. Additional layers of dielectric materials (e.g., MgF_2) may be used to enhance reflectivity over a specific wavelength range and for protection against oxidation.

Silver provides the highest reflectivity, but is very prone to tarnishing. It is most often used on mirrors which provide internal reflection, e.g., on the faces of glass prisms. In this application they can be protected with overcoats of other metals such as copper to prevent oxidation.

Gold coatings are preferred for infrared mirrors because of their excellent tarnishing resistance and very high reflectance (96 to > 99 % for wavelengths of 0.7 to 5.0 µm). Their use is limited to controlled clean environments because they are prone to scratching and difficult to clean. Rhodium coatings are often used with electrodeposited nickel parabolic reflectors. They provide about 80 % reflectivity, but are very resistant to tarnishing, even in salt water.

Dielectric Coatings. Performance can be enhanced by using multiple layers of dielectric materials (e.g., SiO_2, TiO_2, MgF_2, and Al_2O_3) that are arranged so that as the incident beam moves from one layer to the next, part of it is reflected exactly in phase with the reflection from the previous surface. These coatings require precise control of the deposition chamber and typically operate over a limited wavelength and incident angle range, but have excellent reflectivity (> 99 %) and high durability and tarnish resistance. Coatings can also be designed to reflect certain wavelength ranges and transmit others. For example, the reflector in the lamp used in a dentist's office uses a "cold" mirror that reflects most of the visible but transmits most of the infrared. This is reversed for a "hot" mirror which reflects the infrared radiation and transmits the visible. A coating of this type is used on the windows of the Space Shuttle to reduce solar infrared heating.

5. Production of General-Purpose Mirrors

Production of general-purpose mirrors comprises four main stages

1) Cleaning of the glass surface
2) Sensitization of the glass surface
3) Formation of a silver layer
4) Application of a protective coating on the silver layer

Industrial mirror production is performed in units equipped with conveyor belts having a width of 1 – 3.4 m and a length of 50 – > 200 m (Fig. 2). The belt speed is 1 – 10 m/min. The reagents used to produce the silver and protective copper coatings are sprayed onto the glass.

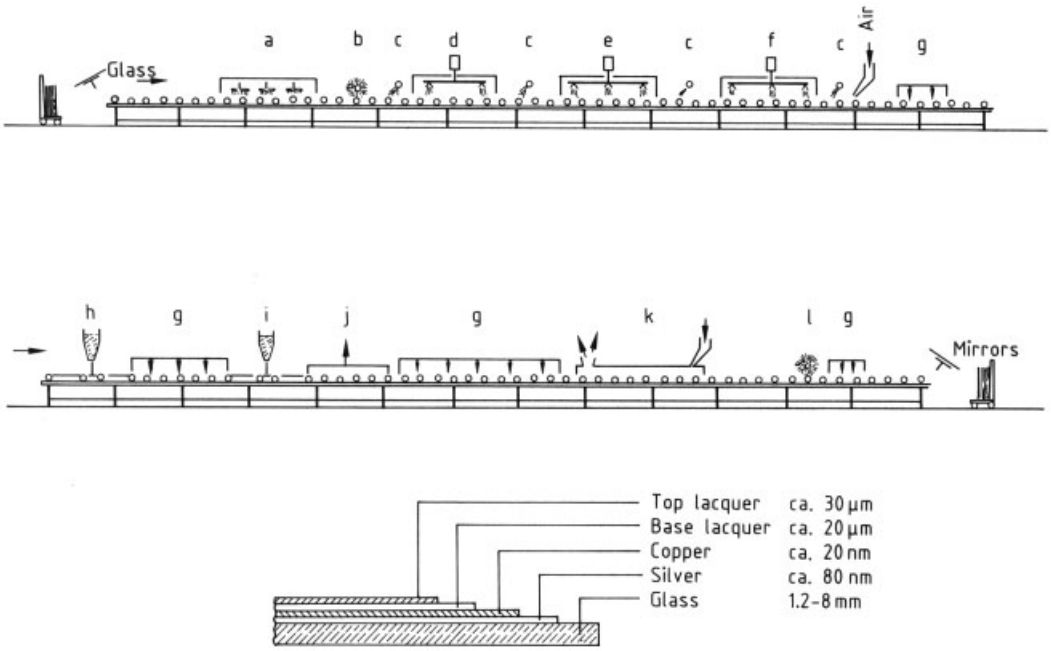

Figure 2. Large-scale production of mirrors
a) Polishing brushes; b) Washing brushes; c) Water wash; d) Stannous chloride; e) Silver nitrate – dextrose solution; f) Copper sulfate solution, zinc dust; g) Drying; h) Application of base lacquer coat; i) Application of top lacquer coat; j) Evaporation; k) Cooling; l) Cleaning

Cleaning is performed with rotating brushes and/or rotating or oscillating polishing felts with polishing agents. The glass is subsequently washed.

Sensitization. The glass surface is sensitized with a stannous chloride solution to ensure uniform adhesion of the silver precipitate. Excess stannous chloride is washed off because it would interfere with the silvering process.

Silvering. Silver is precipitated by reducing an ammoniacal silver nitrate solution with agents such as dextrose, hydrazine, formaldehyde, tartaric acid, or gluconic acid. The glass is then washed again.

Application of Protective Coating. The protective coating consists of a layer of copper and one or two coats of lacquer. The copper is applied on top of the silver either by chemical reduction of copper(II) ions to copper(I) ions or by a contact process. In contact coppering a nonnoble metal (zinc) powder is sprayed onto the mirror together with an acidic copper sulfate solution. The copper(II) ions are precipitated by the metal powder. After application of the lacquer by pouring, rolling, or spraying, the mirror is subsequently dried, cooled, cleaned, and stacked.

6. References

1. G. W. Flynn: *Cutting and Polishing Optical and Electronic Materials,* A. Hilger, Bristol, Philadelphia, 1988.
2. S. Musikant: *Optical Materials: An Introduction to the Selection and Application,* Marcel Dekker, New York 1985.
3. K. Krisciunas: *Astronomical Centers of the World,* Cambridge University Press, Cambridge 1988.
4. R. W. Smith: “Space Telescope” in: *Science, Technology and Politics,* Cambridge University Press, Cambridge 1989.
5. G. E. Pendray: *Men, Mirrors, and Stars,* Funk and Wagnalls Comp., New York–London 1935.
6. Product Catalogs, Melles Griot Inc., Irvine, California.

Mischmetal → Cerium Mischmetal, Cerium Alloys, and Cerium Compounds
Mixing → Continuous Mixing of Fluids
Mixing → Mixing, Introduction
Mixing → Mixing of Highly Viscous Media
Mixing → Mixing of Solids
Mixing → Stirring

Molluscicides

Hans-Jürgen Schnorbach, Bayer AG, Leverkusen, Germany(Chap. 2)

Hans-Detlef Matthaei, Bayer AG, Leverkusen, Germany(Chap. 3)

Franz Müller, (formerly Novartis Crop Protection, Basel), Allschwil, Switzerland (Chap. 1, 4)

1. Introduction 1

2. Agents for the Control of Slugs and Snails in Agriculture and Horticulture . 1

3. Combating Gastropods as Parasite Hosts 4

4. Toxicology 5

5. References 7

1. Introduction

Molluscicides are agents employed to control gastropods. Gastropods form the class containing the most species in the phylum Mollusca. The pulmonate gastropods (Pulmonata, Stylommatophora) feed on plants and are mainly responsible for damage in horticulture, agriculture, and crop storage; they also spread infectious plant diseases, spoilage bacteria, and cause contamination [1–6]. The importance of the aquatic pulmonates (Pulmonata, Basommatophora) lies in their role as intermediate hosts for pathological human and animal parasites [7], [8]. Control of the two groups of molluscs is based on their respective habits and often requires the use of different control agents or measures.

2. Agents for the Control of Slugs and Snails in Agriculture and Horticulture

Damage caused by gastropods has increased as a result of modern methods of cultivation. Reduced mechanical cultivation of the soil improves their chance of survival. Gastropods represent a serious problem in crops such as potatoes or transplanted vegetables which are grown at low plant densities. The loss of even a single plant per meter of row can result in significant stand and yield reductions. Other crops which are especially prone to attack (particularly in the seedling stage of growth) are cereals, rape, legumes, and ornamental plants. In plants intended for canning or deep freezing (e.g., vegetables and strawberries), soiling with slime and excreta is a particularly serious problem.

The damage done by slugs is particularly serious, especially that caused by the common slugs (*Deroceras agreste* L. and *D. reticulatum* Müll.) which are found worldwide in regions with a temperate humid climate. Under mild conditions they are active and capable of reproducing throughout the year; like all pulmonates they are hermaphrodites. The slug species *Arion hortensis* Fér., *A. ater* (L.), *A. circumscriptus* Johnst., and *Milax budapestensis* Hazay are all increasing in importance. The latter species lives underground and is penetrating more and more into southern Europe, causing damage primarily in green-houses.

Snails are better adapted than slugs for surviving dry periods and are therefore also very important in warmer regions. Important species in the Mediterranean region include the Mediterranean sand snail (*Theba pisana* Müll.) in citrus groves and the European edible snail (*Helix aspersa* Müll.), which has also been introduced into the United States. The giant African snail (*Achatina fulica* Bowdich) has colonized many countries and is also of great importance [3].

Since the introduction of direct sowing in aqueous rice cultivation, aquatic snails belonging to the operculate group (Prosobranchia) have caused problems in many countries. These snails possess a double breathing system consisting of gills and a breathing tube (see Chap. 3 for the control of these water snails). The "golden apple snail", *Pomacea* sp. (*Ampullaris*) was introduced into Taiwan from Central America at the start of the 1980s and has developed into a sig-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_649

nificant pest in rice fields [9] ; it is now spreading rapidly throughout the Philippines [10].

In contrast to the large number of insecticides, only a few chemical agents are available for the control of terrestrial gastropods.

Corrosive and Dehydrating Agents. Corrosive and dehydrating agents were employed initially for the control of terrestrial slugs and snails, slugs being particularly sensitive. High water loss as a result of extremely high slime excretion paralyzes the animal and leads to death. Quicklime CaO, freshly slaked lime $Ca(OH)_2$, and finely ground kainite $KMg(Cl/SO_4) \cdot 3\,H_2O$ are effective when applied in dry weather at a concentration of 300 – 600 kg/ha. The slugs or snails must, however, come in direct contact with these chemicals and be treated twice within a short period of time. The best results are obtained when the compounds are applied early in the morning or in the evening at the time of greatest gastropod activity. However, because of the potential crop damage from such treatments, it is recommended that these applications be made before the crop plants emerge above the soil. Rainfall rapidly inactivates these corrosive agents, as well as soluble agents such as copper sulfate (see below) and physical barriers such as sand, ash, soot, and cattle salt. Protective strips of fine sand, ash, and soot have an indirect drying effect due to the fact that the fine particles adhere to the moist foot and body of the slug or snail. The animals are then unable to crawl on such substances and thus dry out in warm weather. Cattle salt dehydrates the gastropods. Protective strips rapidly become ineffective in the presence of moisture and have to be reapplied at frequent intervals.

Copper(II) sulfate [*7758-99-8*], cupric sulfate, $CuH_{10}O_9S$, M_r 249.7, consists of blue crystals which are soluble in water, moderately soluble in methanol, and practically insoluble in most organic solvents (→ Copper Compounds, Chap. 4.4.1.[11, p. 270].

$CuSO_4 \cdot 5\,H_2O$

Copper sulfate is employed in considerable quantities against terrestrial gastropods, particularly in the developing countries. The compound is usually pulverized and applied at a concentration of 10 – 20 kg/ha; for ease of handling, it is often mixed with commercial fertilizer in a ratio of 1 – 2 : 40. Spraying with 12 kg of the agent dissolved in 400 L of water per hectare is also recommended. Because copper sulfate is a contact poison, snails and slugs must come into direct contact with the compound. Standing crops should not be treated in order to avoid damaging the plants.

Trade Names. Blue Viking (Griffin); Super Bouillie (La Cornubia); Triangle Brand (Phelps-Dodge).

Calcium cyanamide [*156-62-7*], $CaCN \cdot CN$ has been used as a repellent along the edges of fields to prevent the entry of slugs and snails.

Metaldehyde [*108-62-3*], 2,4,6,8-tetramethyl-1,3,5,7-tetroxocane, $C_8H_{16}O_4$, M_r 176.2, *mp* 246 °C, is the cyclic tetramer of acetaldehyde (→ Acetaldehyde, Chap. 8.2. It forms a crystalline powder which is moderately soluble in water, fairly soluble in toluene, and readily soluble in methanol [11, p. 796]. Metaldehyde is produced by polymerization of acetaldehyde in aqueous ethanol in the presence of HBr and alkaline-earth metal bromides [12].

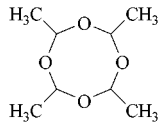

The molluscicidal activity of metaldehyde has been exploited in South Africa since 1934 and was discovered independently in France in 1936 [13]. Ingestion of metaldehyde causes irreversible permeability of the gut walls of gastropods and leads to their death. However, under favorable weather conditions (low temperature, high relative humidity), gastropods often survive the paralysis produced on contact with the compound and the increased slime production caused by irritation. Metaldehyde is also effective in the vapor phase and kills gastropod eggs. Different species are affected to differing degrees.

Metaldehyde is usually applied in the form of pelletized bait based on wheat bran or barley flour containing 4 – 6 % of the agent. Despite the addition of various attractants (proteins, dextrose, casein, etc.), these baits are not often eaten

by purely herbivorous gastropods and act mainly as contact poisons; the bait should therefore always be spread evenly rather than in small heaps. In spite of their similar composition, the amounts of individual baits employed differ greatly, depending on the size of the bait particles. Pellet lengths range from 1 to 10 mm, but 2 – 5 mm is most common. If the bait is of good quality (high attractant power and stability during rainfall), 25 particles per square meter can guarantee adequate control for the most destructive gastropod species. Small bait particles (2 – 4 mm long, diameter ca. 2 mm) allow application of smaller quantities (3 – 9 kg/ha) without loss of efficacy [14].

Repellents, such as nonanoic acid vanillylamide are often added to discourage the consumption of slug bait by mammals [15].

Slug pellets containing metaldehyde are manufactured by many companies throughout the world. The official plant protection agent registers of Germany, for instance, list a large number of commercial preparations [16].

Metaldehyde can also be sprayed or dusted as a powder, but larger quantities are required (ca. 5 kg of active ingredient per hectare) than when it is employed in bait. Gastropods must come into direct contact with the spray or dust. A viscous paste containing 4 % metaldehyde is available. Drops of the paste are applied around the plants.

Trade Names. Cekumeta (Sequisa); Halizan (Tamogan); Hardy (Chiltem); Meta (Lonza); Metason (Jewin-Joffé); MifaSlug (FCC).

Methiocarb [*2032-65-7*], mercaptodimethur, 3,5-dimethyl-4-(methylthio)phenyl methylcarbamate, $C_{11}H_{15}NO_2S$, M_r 225.3, *mp* 119 °C, consists of colorless crystals which are sparingly soluble in water and hexane, and fairly soluble in dichloromethane, isopropyl alcohol, and toluene [11, p. 813].

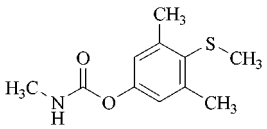

Methiocarb is produced by the reaction of methyl isocyanate with 3,5-dimethyl-4-methylthiophenol, which in turn is prepared from 3,5-dimethylphenol and methyl sulfenyl chloride [17].

Methiocarb was first employed as an insecticide in 1962 and has been used as a molluscicide since 1964; it also possesses acaricidal acticity. Methiocarb is a molluscicide with neurotoxic action (cholinesterase inhibition) and is used to control slugs and snails in a wide range of agricultural situations. Like metaldehyde, methiocarb is employed almost exclusively in the form of bait at a concentration of 1, 2, or 4 %. Slug bait formulated as small pellets is generally applied in amounts of 3 – 5 kg/ha, corresponding to 20 pellets per square meter. As with metaldehyde, the amount of agent required increases considerably if spray application is employed (e.g., 0.5 – 1.75 kg of active ingredient per hectare); slugs and snails must also come into direct contact with the spray. Seed treatment with 0.1 – 0.2 % methiocarb is employed when sowing wheat to avoid damage by slugs.

Gastropods are rapidly paralyzed on feeding or contact; those affected do usually not recover [18]. The effect is largely independent of temperature and is particularly good under moist conditions.

Trade Names. Mesurol, Draza (Bayer).

Thiodicarb [*59669-26-0*], 3,7,9,13-tetramethyl-5,11-dioxa-2,8,14-trithia-4,7,9,12-tetra-azapentadeca-3,12-diene-6,10-dione, $C_{10}H_{18}N_4O_4S_3$, M_r 354.5, *mp* 173 – 174 °C, consists of colorless crystals which are sparingly soluble in water but readily soluble in dichloromethane, acetone, methanol, and xylene [11, p. 1195].

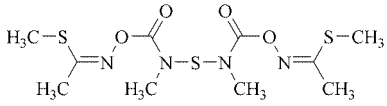

Thiodicarb is produced by reaction of *N*,*N*′-thiobis(methylcarbamic acid fluoride) with 2-methylthioacetaldoxime in the presence of a base [19].

Thiodicarb is a carbamate insecticide (cholinesterase inhibition) with molluscicidal properties. It was introduced in France in 1988. The recommended application rate for pelleted bait containing 4 % active ingredient is 5 kg/ha [20].

Trade Names. Skipper, Larvin (Rhône – Poulenc).

Fumigants are required for quarantine measures to prevent the introduction of gastropod pests from foreign countries. A mixture of 10 % ethylene oxide and 90 % carbon dioxide (ca. 320 g/m^3 for 72 h) is effective against snails even in their closed state (against *Achatina*), as are hydrogen cyanide (8 mg/L for 2 h under reduced pressure), and methyl bromide (ca. 130 g/m^3 for 72 h). Methyl bromide appears to be tolerated particularly well by plants when used in vacuum fumigation and allows the decontamination of imported seeds [21].

Insecticides and Fungicides. A range of carbamate insecticides also have molluscicidal activity. This is particularly true for Trimethacarb.

Trimethacarb [*12407-86-2*], mixture of 3,4,5- and 2,3,5-trimethylphenyl methylcarbamate (ratio: 3.5 : 1), $C_{11}H_{15}NO_2$, M_r 193.2, *mp* 105 – 114 °C, is a buff to brown crystalline product which is sparingly soluble in water and in organic solvents [11, p. 1254].

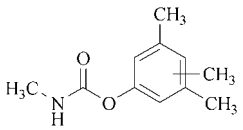

Trimethacarb is obtained by reaction of a mixture of 2,3,5- and 3,4,5-trimetylphenol with methyl isocyanate [22]. Trimethacarb is a cholinesterase inhibitor that is used to control a wide range of mollusc pests, and it also acts as a mammal and bird repellent.

Fentin Acetate [*900-95-8*], triphenyltin acetate, $C_{20}H_{18}O_2Sn$, M_r 409.0, *mp* 121 – 123 °C, consists of colorless crystals which are sparingly soluble in water, and fairly soluble in ethanol, ethyl acetate, hexane, dichloromethane, and toluene [11, p. 533].

Fentin acetate is produced by treatment of triphenyltin chloride with sodium acetate [23]. Fentin acetate is used for the control of algae and snails in rice fields, and also for the control of water snails in fish ponds.

Trade Names. Brestan (AgrEvo); Badar (Productos OSA); Suzu (Nihou Nohyaku).

Other Agents. Phosphate esters are active against gastropods only in exceptional cases. A good effect against slugs has been described for phorate (→ Insecticides). Azinphos-methyl (→ Insecticides) has been employed successfully for a considerable time against *Helix aspersa* in Californian citrus groves. Persistent molluscicidal activities have also been detected for nitrophenols (→ Insecticides, Chap. 4.) and for copper-containing fungicides (→ Fungicides, Chap. 6.1.) when employed for postharvest treatment in berry crops.

New Developments. The Rothamsted Experimental Station in the United Kingdom is working on the development of a new group of molluscicides. These are metal chelate complexes that contain iron or aluminum as the central atom [24].

3. Combating Gastropods as Parasite Hosts

Gastropods play an important role in human and veterinary medicine as hosts (vectors) in the life cycles of digenetic trematodes (flukes). In Europe, *Lymnaea truncatula* is the vector for the great liver fluke (*Fasciola hepatica*). In East Asia, *Fasciolopsis buski*, the "giant intestinal fluke" is spread by snails of the family Planorbidae (*Segmentina* species). *Dicrocoelium dendriticum*, the small liver fluke, requires pulmonate snails (e.g., *Helicella ericetorum, H. itala,* and *Zebrina detrita*) for its development. The life cycle of the Opisthorchidae trematode species (*Opisthorchis felineus, O. sinensis*) involves operculates belonging to the genus *Bithynia*.

According to the World Health Organization, about 200×10^6 people suffer from schistosomiasis (bilharziasis). In this context, several species of *Schistosoma* are of great importance

to humans in Africa, Central and South America, and East Asia [7]. *Schistosoma haematobium* (urinary schistosomiasis) and *S. mansoni* (intestinal schistosomiasis) require freshwater snails (*Bulinus* spp., *Biomphalaria* spp.) as hosts; *S. japonicum* (intestinal schistosomiasis) utilizes operculates (Prosobranchia *Oncomelania* spp.).

Successful control with molluscicides has been achieved only with schistosomiasis carriers. The agents and methods employed are, however, also applicable to the plant-damaging water snails discussed in Chapter 2. In the past, sodium pentachlorophenolate and copper sulfate were employed. Only one preparation is now available that can be judged as the agent of choice with regard to both its toxicology and its efficacy, this is niclosamide.

Niclosamide [*50-65-7*] 2',5-dichloro-4'-nitrosalicylanilide, $C_{13}H_8Cl_2N_2O_4$, M_r 327.1, *mp* 230 °C, consists of almost colorless crystals which are sparingly soluble in water but soluble in common organic solvents [11, p. 874].

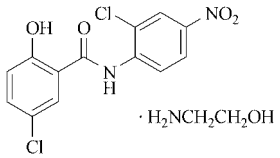

Niclsoamide is produced by reaction of 2-chloro-4-nitroaniline with 2-hydroxy-5-nitrobenzyl chloride in the presence of a base [25].

Niclosamide is a molluscicide with respiratory and stomach action used to control golden apple snails in rice. Niclosamide is very effective against snails, snail eggs, and the free-living larval stages (miracidia and cercariae) of the schistosomes [26], [27]. It is also used for the control of schistosomiasis and fascioliasis in man by killing fresh-water snails which act as intermediate hosts, and for veterinary control of tapeworm infestations.

Niclosamide is available as a powder for sprays (70 %) and as an emulsion concentrate (25 %). The concentration applied to flowing and stagnant waters is 0.5 – 1 ppm for 8 h or 4 – 8 ppm for 1 h. Its biological effectiveness in water is stable at pH 5 – 9 and is also largely independent of the concentration of calcium and magnesium ions [28].

Niclosamide shows no phytotoxic effects on irrigated crops (e.g., bananas), it does not damage zoo- or phytoplankton. The compound does not accumulate in the food chain; it is biodegradable and metabolized by certain microorganisms. Niclosamide shows, however, lethal effects on fish at the concentrations in which it is applied.

Trade Name. Bayluscid (Bayer).

Trifenmorph [*1420-06-0*], *N*-tritylmorpholine, M_r 329.42, was formerly used against gastropods.

Older molluscicides, such as *yurimin* [*3281-96-7*], 3,5-dibromo-4-hydroxy-4′-nitroazobenzene, were employed in Japan against *Oncomelania hupensis nosophora*. Yurimin has a rather unfavorable acute toxicity. In addition, it is more toxic to fish than to gastropods. The product is no longer on the market.

Organotin compounds such as bis(tri-*n*-butyltin) oxide exhibit a molluscicidal effect, but because too little is known about their environmental behavior, work being done on these compounds may be discontinued [7].

Endod, a herbal extract of *Phytolacca dodecandra*, has been extensively investigated in Ethiopia [29]. It exhibits an adequate effect against *Biomphalaris pfeifferi*. Endod has only local importance; it is not available commercially.

4. Toxicology

Copper(II) sulfate [*7758-98-7*] [11, p. 270]

LD_{50} (rat, oral)	ca. 300 mg/kg
LD_{50} (mouse, oral)	87 mg/kg
NOEL/NOAEL:	
Chronic.	
Rat (2-year feed)	25 mg kg^{-1} d^{-1}

Metaldehyde [*108-62-3*] [30]

LD_{50} (rat, oral)	227 – 690	mg/kg
LD_{50} (mouse, oral)	200	mg/kg
LD_{50} (guinea pig, oral)	175 – 500	mg/kg
LD_{50} (rabbit, oral)	290 – 1250	mg/kg
LD_{50} (dog, oral)	100 – 1000	mg/kg
Nonteratogenic		
NOEL/NOAEL:		
Reproduction.		
Rat (3-generation feed)	10	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)
Chronic.		
Rat (2-year feed)	10	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.025 $mg\,kg^{-1}\,d^{-1}$ [11, p. 796]

Methiocarb [*2032-65-7*] [31]

LD_{50} (rat, oral)	13 – 135	mg/kg
LD_{50} (guinea pig, oral)	14.1 – 100	mg/kg
LD_{50} (dog, oral)	10 – 25	mg/kg
LD_{50} (hen, oral)	175	mg/kg
LD_{50} (rat dermal)	> 300 to > 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.322	mg/L air
Nonmutagenic		
NOEL/NOAEL:		
Reproduction.		
Rat (3-generation feed)	15	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	3	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	3	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Chronic.		
Rat (2-year feed)	1.3	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	0.13	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$ [32]

Thiodicarb [*59669-26-0*] [33]

LD_{50} (rat, oral)	39 – 136	mg/kg
LD_{50} (mouse, oral)	226	mg/kg
LD_{50} (guinea pig, oral)	160	mg/kg
LD_{50} (rabbit, oral)	> 400	mg/kg
LD_{50} (chicken, oral)	582 – 660	mg/kg
LD_{50} (monkey, oral)	467	mg/kg
LD_{50} (rat, dermal)	2540	mg/kg
LD_{50} (rabbit, dermal)	> 6310	mg/g
LC_{50} (rat, 4-h inhalation)	0.1155 – 0.220	mg/L air
Rabbit: nonirritating to skin		
Rabbit: moderately irritating to eyes		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL:		
Reproduction.		
Rat (3-generation feed)	> 19	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	3	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 3	$mg\,kg^{-1}\,d^{-1}$ (developmental)

ADI: 0.03 $mg\,kg^{-1}\,d^{-1}$ [34]

Trimethacarb [*12407-86-2*] [35]

LD_{50} (rat, oral)	125 – 232	mg/kg
LD_{50} (dog, oral)	225	mg/kg
LD_{50} (rabbit, dermal)	> 2000 to > 2500	mg/kg
Rabbit: not a skin irritant		
Rabbit: mild eye irritant		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL:		
Subchronic.		
Rat (90-d feed)	19 – 21	$mg\,kg^{-1}\,d^{-1}$
Mouse (90-d feed)	119 – 138	$mg\,kg^{-1}\,d^{-1}$
Reproduction.		
Rat (3-generation feed)	< 40	$mg\,kg^{-1}\,d^{-1}$ (parental)
Teratogenicity.		
Rat (gavage)	< 30	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 30	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	< 50	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 50	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	40	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	4.6 – 4.9	$mg\,kg^{-1}\,d^{-1}$

ADI: not available

Fentin Acetate [*900-95-8*] [36]

LD_{50} (rat, oral)	140 – 298	mg/kg
LD_{50} (mouse, oral)	81 – 93	mg/kg
LD_{50} (guinea pig, oral)	25	mg/kg
LD_{50} (rabbit, oral)	80	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LD_{50} (mouse, dermal)	350	mg/kg
LC_{50} (rat, 4-h inhalation)	0.044	mg/L air
Rabbit: nonirritating to skin		
Rabbit: severe irritant to eyes		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL:		
Reproduction.		
Rat (2-generation feed)	0.4	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)
Teratogenicity.		
Rat (gavage)	3.2	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	0.1	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	0.32	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	< 0.3	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	1.0	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	0.2	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0005 $mg\,kg^{-1}\,d^{-1}$

Niclosamide [*50-65-7*] [37]

LD_{50} (rat, oral)	3710 to ≥ 5000	mg/kg	
LD_{50} (mouse, oral)	1500	mg/kg	
LD_{50} (rabbit, oral)	> 5000	mg/kg	
LD_{50} (cat, oral)	> 1000	mg/kg	
LD_{50} (dog, oral)	> 250	mg/kg	
LD_{50} (rat, dermal)	> 1000	mg/kg	
LC_{50} (rat, 4-h inhalation)	20	µg/L air	[11, p. 874]

Rabbit: severe eye irritant [11, p. 874]
Rabbit: skin irritant, after repeated exposure [11, p. 874]
Nonmutagenic [11, p. 874]
NOEL/NOAEL:
Chronic.

Rat (2-year feed)	100 – 400	$mg\,kg^{-1}\,d^{-1}$	[11, p. 874]
Mouse (2-year feed)	30	$mg\,kg^{-1}\,d^{-1}$	[11, p. 874]
Dog (one-year feed)	100	$mg\,kg^{-1}\,d^{-1}$	[11, p. 874]

ADI: $3\,mg\,kg^{-1}\,d^{-1}$ [11, p. 874]

5. References

General References

1. V. Fretter, J. Peake (eds.): *Pulmonates,* vol. 1, Academic Press, London 1975.
2. V. Fretter, J. Peake (eds.): *Pulmonates,* vol. 2 A, Academic Press, London 1978.
3. V. Fretter, J. Peake (eds.): *Pulmonates,* vol. 2 B, Academic Press, London 1979.
4. D. Godan: *Pest Slugs and Snails, Biology and Control,* Springer Verlag, Berlin-Heidelberg 1983.
5. I. Henderson (ed.): *Slugs and Snails in World Agriculture,* British Crop Protection Council Monograph No. 41, Thornton Heath 1989.
6. C. M. Port, G. R. Port: "The Biology and Behaviour of Slugs in Relation to Crop Damage and Control," *Agric. Zoology Rev.* **1** (1986) 255 – 299.
7. P. Jordan, G. Webbe (eds.): *Schistosomiasis, Epidemiology, Treatment and Control,* William Heinemann Medical Books Ltd., London 1982.
8. K. W. Frank: *Parasitologie,* E. Ulmer, Stuttgart 1976.

Specific References

9. E. Y. Cheng in [5]pp. 69 – 75.
10. C. B. Adalla, B. Morallo-Rejesus in [5], pp. 417 – 422.
11. C. D. S. Tomlin, *The Pesticide Manual,* 11th ed., British Crop Protection Council, Farnham, UK 1997.
12. Publicker Ind., US 2 426 961, 1944 (R. S. Wilder).
13. M. A. Wedgwood, S. E. R. Bailey, *Ann. Appl. Biol.* **112** (1988) 439 – 457.
14. G. Hermann, *Meded. Fac. Landbouwwet. Rijksuniv. Gent.* **36** (1971) 197 – 206.
15. Lonza, CH 9310-74, 9369-74 = DE-OS 2 512 163, 2 512 164, 1975.
16. Biol. Bundesanst. f. Land- u. Forstwirtsch. (Braunschweig) : *Pflanzenschutzmittel-Verz.,* Parts 1, 2, 36th ed., 1988, pp. 224, 229.
17. Bayer, EP 912 895, 1960 (E. Schegk, G. Schrader, K. Wedemeyer).
18. T. J. Martin, J. D. Forrest, *Pflanzenschutz Nachr. Bayer (Ger. Ed.)* **22** (1969) 212 – 253.
19. Rhône Poulenc, DE 2 654 331, 1977 (T. D. J. D′Silva).
20. R. Schwer, J. M. Gaulliard, *Int. Symp. o. Fytoformacie en Fytiatrie, Samenvattingen van de Mededelingen,* Gent 1988, p. 73.
21. H. Roth, J. W. Kennedy, *J. Econ. Entomol.* **66** (1973) 935 – 936.
22. Rhône Poulenc, DE 2 532 982, 1976 (J. K. F. Chan, E. Tobler, H. E. Johnson).
23. R. Bock, *Residue Rev.* **79** (1981) 1.
24. I. F. Henderson et al., in [5], pp. 289 – 294.
25. E. Schraufstätter, W. Meiser, R. Gönnert, *Z. Naturf. B* **16** (1961) 95.
26. R. Gönnert, *Pflanzenschutz Nachr. Bayer (Ger. Ed.)* **15** (1962) 4 – 25.
27. P. Andrews, J. Thyssen, D. Lorke, *Pharmacol. Ther.* **19** (1983) 245 – 295.
28. R. Strufe, R. Gönnert, *Pflanzenschutz Nachr. Bayer Ger. Ed.* **15** (1962) 50 – 70.
29. A. Lemma, *Ethiop. Med. J.* **3** (1965) 187. A. Lemma, *Bull. WHO* **4** (1970) 597.
30. W. J. Hayes, Jr.: *Pesticide Studies in Man,* Williams & Wilkins, Baltimore – London 1982, p. 627.
31. Pesticide Residues in Food-1981. *FAO Plant Prod. Protect. Paper* **42** (1982) 317.
32. Pesticide Residues in Food-1984. *FAO Plant Prod. Protect. Paper* **61** (1985) 239.
33. Pesticide Residues in Food-1985. *FAO Plant Prod. Protect. Paper* **72/2** (1986) 171.
34. Pesticide Residues in Food-1986. *FAO Plant Prod. Protect. Paper* **78/2** (1987) 185.
35. R. L. Baron in W. J. Hayes, Jr., E. R. Laws, Jr. (eds.): *Handbook of Pesticide Toxicology,* Academic Press, San Diego 1991, p. 1125.
36. Pesticide Residues in Food-1991. *WHO/PCS* **92.52** (1992) 173.
37. G. Hecht, C. Gloxhuber, *Pflanzenschutz Nachr. Bayer Ger. Ed.* **15** (1962) 34 – 41.

Molybdenum and Molybdenum Compounds

Roger F. Sebenik, Cyprus Minerals Company, Englewood, Colorado 80112, United States (Chap. 1; Sections 4.1 – 4.4 in part)

A. Richard Burkin, Royal School of Mines, Imperial College, London, United Kingdom (Chaps. 2, 3, and 9)

Robert R. Dorfler, Cyprus Minerals Company, Green Valley, Arizona 85622, United States (Sections 4.1 – 4.4 in part)

John M. Laferty, Climax Molybdenum Company, Wheatridge, Colorado 80033, United States (Sections 4.1 – 4.4 in part)

Gerhard Leichtfried, Metallwerk Plansee GmbH, Reutte, Austria (Sections 4.5 – 4.8, Chap. 5)

Hartmut Meyer-Grünow, Murex Rainham, Rainham, United Kingdom (Chap. 6, based on the corresponding chapter in Ullmann's, 4th. ed. written by Rudolf Fichte)

Philip C. H. Mitchell, The University of Reading, Reading, United Kingdom (Chap.7)

Mark S. Vukasovich, 1457 Woodland Drive, Ann Arbor, Michigan 48103, United States (Chap. 8)

Douglas A. Church, Climax Metals Company, AMAX Inc., AMAX Center, Greenwich, Connecticut 06836-1700, United States (Chap. 10)

Gary G. Van Riper, Amax Mineral Resources Company, Golden, Colorado 80401-3293, United States (Chap. 11)

James C. Gilliland, Amax Mineral Resources Company, Golden, Colorado 80401-3293, United States (Chap. 11 and Chap. 12 in part)

Stanley A. Thielke, Climax Metals Company, AMAX Inc., AMAX Center, Greenwich, Connecticut, 06836-1700 (Chap. 12 in part)

1. Introduction . . . 2
2. Properties . . . 2
3. Occurrence . . . 3
3.1. Minerals . . . 3
3.2. Deposits . . . 3
4. Production . . . 5
4.1. Concentration . . . 5
4.2. Processing of Concentrate . . . 7
4.3. Recovery from Spent Petroleum Catalysts . . . 9
4.4. Recovery during Production of Tungsten Ores . . . 10
4.5. Production of Molybdenum Metal Powder . . . 11
4.6. Production of Compact Molybdenum Metal . . . 11
4.7. Processing of Molybdenum . . . 12
4.8. Molybdenum-Base Alloys . . . 14
5. Uses . . . 14
6. Production of Ferromolybdenum . . . 16
6.1. Ferromolybdenum Grades . . . 16
6.2. Raw Materials . . . 16
6.3. Submerged Arc Furnace Carbothermic Reduction . . . 17
6.4. Metallothermic Reduction . . . 19
7. Molybdenum Compounds . . . 22
7.1. Overview of Molybdenum Chemistry 22
7.2. Molybdenum Oxides . . . 24
7.3. Molybdenum Chalcogenides . . . 25
7.4. Molybdenum Halides . . . 27
7.5. Molybdates, Isopolymolybdates, and Heteropolymolybdates . . . 28
7.6. Other Molybdenum Compounds . . . 30
8. Uses of Molybdenum Compounds . . 30
8.1. Catalysis . . . 30
8.2. Lubrication . . . 32
8.3. Corrosion Inhibition . . . 33
8.4. Flame Retardancy and Smoke Suppression . . . 34
8.5. Pigments . . . 34
8.6. Agriculture . . . 35
9. Analysis . . . 35
10. Economic Aspects . . . 37
11. Environmental Aspects . . . 38
12. Toxicology and Occupational Health . 40
13. References . . . 43

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_655

1. Introduction [1–4]

Molybdenum commonly occurs in nature as the mineral molybdenite, MoS_2, and is found as veins in quartz rock. It was discovered by SCHEELE in 1778, but for the next hundred years, molybdenite was merely a laboratory curiosity. The first major use came during World War I when additions of molybdenum produced steels with excellent toughness and strength at high temperatures for use as tank armor and in aircraft engines.

Historical Aspects. The main source of molybdenite throughout the 19th century was the Knaben mine in Norway in which the molybdenite was concentrated by hand. The World War I demand for molybdenum, however, spurred development of the Climax Mine (Colorado), which was put on stream in 1918 by the Climax Molybdenum Company. By that time the froth flotation process had replaced hand sorting for concentrating ores.

Although the Climax Mine was closed after the war because of the suddenly decreased demand, it was reopened in 1924 as new peacetime uses of molybdenum, largely in the automotive industry, were developed. Demand for molybdenum steadily grew as new chemical and metallurgical applications continued to be developed, largely stimulated by the research efforts of the Climax Molybdenum Company.

In 1933 a process for the selective flotation of molybdenite from copper porphyry ores was developed by the Anaconda Company at its Greene Cananea subsidiary in Mexico. The Kennecott Copper Company soon followed at its Utah mines. The technology was later extended to Chile at Kennecott's El Teniente Mine and Anaconda's Chuquicamata mine. Thus began the age of molybdenum production as a byproduct of copper.

During World War II, the United States supplied about 90 % of the world molybdenum demand, most coming from Climax with the balance from Kennecott's Utah mines and Molycorp's Questa mine. The largest of the other Western-world producing countries were Chile, Mexico, and Norway.

Climax Molybdenum Company continued to be the largest Western-world producer into the 1980s, doubling its capacity when it opened the massive Henderson mine (near Empire, Colorado) in 1976. Production from Mexico and Norway remained small, but less expensive byproduct molybdenum production from Chile and the United States continued to grow, and the Endako and other mines in Canada began molybdenum production. In 1977, total world production exceeded 200×10^6 pounds (ca. 90 000 t) for the first time.

During the late 1970s the high price of molybdenum stimulated the opening of several new primary mines and the expansion of many byproduct molybdenum facilities. When the price declined in the early 1980s, byproduct molybdenum became the dominant economic force in the marketplace. Climax Molybdenum Company closed the underground mine at Climax, but was still a major producer with the Henderson mine and the open pit at Climax. Cyprus Minerals, with four operating mines, became a major producer. Other important Western-world producers at the end of the 1980s were Codelco's Chuquicamata mine in Chile, Placer's Endako mine in Canada, the La Caridad mine in Mexico, and the Cuajone and Toquepala mines in Peru. Annual world production and consumption averaged $(180-200) \times 10^6$ pounds (80 000 – 90 000 t) of molybdenum during the late 1980s.

2. Properties

Physical Properties. Molybdenum [*7439-98-7*], Mo, A_r 95.94 ± 0.01, atomic number 42, electronic configuration $[Kr]\,4d^5\,5s^1$, is the second member of group 6 of the periodic table. As a transition element it may have a valency of 2, 3, 4, 5, or 6. In the massive state molybdenum is a lustrous silver-white solid with typically metallic properties. When produced as a powder it is dull grey. It has a body-centred cubic lattice with $a_0 = 0.31472$ nm. The physical properties of molybdenum depend to a large degree on the method used to produce it and on its subsequent treatment. For example, its relative density calculated from a_0 is 10.22 whereas an earlier measurement gave 9.01 and later work showed that the lower value increases with the amount of mechanical working to which the metal is subjected. The accepted values of the main physical properties of the metal are given below. A more

extensive compilation is available giving, in particular, data of importance to metallurgists [5].

mp	2617 °C
bp (101.3 kPa)	4612 °C
Latent heat of fusion at *mp*	35.6 kJ/mol
Mean specific heat (0 – 100 °C)	251 J kg^{-1} K^{-1}
Density (20 °C)	10.22 g/cm^3
Thermal conductivity (0 – 100 °C)	137 W m^{-1} K^{-1}
Electrical resistivity (20 °C)	5.7 μΩ · cm
Temperature coefficient (0 – 100 °C)	4.35×10^{-3} K^{-1}
Elastic constants of polycrystalline metal (20 °C)	
Young's modulus	324.8 GPa
Rigidity modulus	125.6 GPa
Bulk modulus	261.2 GPa
Poisson's ratio	0.293
Linear coefficient of thermal expansion (0 – 100 °C)	5.1×10^{-6} K^{-1}
Standard electrode potential $E^0_{Mo^{3+},Mo}$	− 0.200 V

Chemical Properties. Molybdenum retains its luster almost indefinitely in air, particularly when it has been drawn to fine wire. It can be passivated by oxidation, especially by electrolytic oxidation, becoming chemically unreactive. On prolonged heating in air below 600 °C, the metal becomes covered with its trioxide; at 600 °C the oxide sublimes and rapid oxidation occurs. Molybdenum burns in oxygen at 500 – 600 °C. It is slowly oxidized by steam, is attacked by fluorine when cold and by chlorine and bromine when hot. Dilute acids and concentrated hydrochloric acid have very little effect on the metal. Moderately concentrated nitric acid dissolves it but concentrated nitric acid soon passivates the surface and reaction ceases. Molybdenum is dissolved by a mixture of concentrated nitric and concentrated hydrofluoric acids. It is practically unaffected by alkaline solutions and very nearly so by fused alkali-metal hydroxides. However fused oxidizing salts such as sodium peroxide, sodium or potassium nitrate or perchlorate dissolve the metal rapidly. It reacts on heating with carbon, boron, nitrogen, and silicon and forms many alloys. It is used in a variety of catalysts, especially combined with cobalt in the desulfurization of petroleum. Molybdenum is a biologically active metal which is involved in the functioning of enzymes causing reduction of nitrogen to ammonia, and of nitrates.

3. Occurrence

3.1. Minerals

Molybdenum occurs in the earth's crust in an abundance of about 10^{-4} %, mainly as molybdenite [*1309-56-4*], MoS_2. Small quantities are also associated with other metals having similar chemical properties. Thus wulfenite [*14913-82-7*], ($PbMoO_4$) is found in oxidized parts of sulfidic lead deposits in many regions of the world; powellite [*14020-51-0*], [$Ca(Mo,W)O_4$] is a calcium molybdate usually formed from the modification of molybdenite in deposits containing tungsten as scheelite ($CaWO_4$) and tungsten substitutes for molybdenum up to about 10 %. Ferrimolybdite [*15079-04-6*] ($Fe_2Mo_3O_{12} \cdot 8\,H_2O$) is the product of oxidation of molybdenite in the presence of iron(III) compounds. Molybdenite is the primary source of molybdenum. Should demand for molybdenum rise so that additional sources become necessary, wulfenite, powellite, and ferrimolybdite may become commercially significant. Other molybdenum minerals include achrematite, belonesite, chillagite, eosite, ilsemannite, jordisite, koechlinite, lindgrenite, and paterite. These are of no commercial importance however.

3.2. Deposits

There are five genetic types of molybdenum deposits:

1) porphyry deposits in which metallic sulfides are disseminated throughout large volumes of altered and fractured rock,
2) contact-metamorphic zones and bodies in which silicated limestone is adjacent to intrusive granites,
3) quartz veins,
4) pegmatites, and
5) deposits bedded in sedimentary rocks.

Types (1) – (3) are of hydrothermal origin and provide almost all of the molybdenum ore currently mined. The last two are of no economic importance at present. Average molybdenite concentrations in primary porphyry deposits range from 0.05 to 0.25 %; in

secondary copper – molybdenum porphyry deposits molybdenite concentrations are much lower (0.01 – 0.05 %) so that the mineral can only be recovered as a byproduct.

In hydrothermal deposits, the cracks and interstices of the intrusive rock forming the matrix were originally penetrated by aqueous metal-containing solutions. Such solutions are often formed close to magma and so are at a high temperature and pressure and contain sulfide species. As the solution cools, minerals are deposited within the intrusive matrix and also commonly in the surrounding country rock. Thus the famous deposit at Climax, Colorado (USA) was formed by an intrusion of a quartz – monzonite magma into a granitic country rock. The hydrothermal solution deposited quartz and molybdenite in even the smallest fractures so that the host rock appears to consist of angular fragments separated by narrow bands of the minerals. In many porphyry deposits small amounts of molybdenite are disseminated with copper sulfide minerals throughout large volumes of granitic rock. The chief mineral values are chalcopyrite ($CuFeS_2$) and chalcocite (Cu_2S), associated with pyrite (FeS_2), and small amounts of other sulfides, including molybdenite. Where secondary enrichment of the copper ore has occurred, chalcocite has replaced pyrite and chalcopyrite. In limestone – silicate sedimentary rocks containing intrusions of granitic rocks, small amounts of molybdenite are occasionally widely distributed along the interfaces between the rocks and are often associated with scheelite ($CaWO_4$), bismuthinite (Bi_2S_3), or copper sulfides. These deposits may contain up to 0.6 % molybdenite.

Pegmatites are the result of igneous activity and subsequent solidification of the magma. The slow cooling of the mass results in the formation of coarse crystals, almost exclusively of quartz and feldspar. Molybdenite occurs in small amounts as an accessory mineral, usually as large crystals; pegmatites are not an important source of molybdenum. Nevertheless, a pegmatite deposit containing molybdenite together with bismuth was mined until recently at Val d'Or and Preissac (Quebec, Canada). Sedimentary rocks in which molybdenum minerals may occur include coals, shales, phosphorites, and sandstones. They are not used as sources of molybdenum and the technology to recover it from them has not been developed.

Molybdenum is obtained commercially almost exclusively from molybdenite, which is either mined and concentrated as the primary product of the mine or open pit, or recovered as a concentrate during the processing of ore from a copper mine. In the latter case the molybdenite can be either a by- or a coproduct, depending on its economic importance to the output of the mine. Molybdenum is also recovered from solutions obtained by leaching scheelite ores in the production of tungsten. Molybdenum supply in the Western world can be broken down into four production segments (see also Chap. 10) [6]:

1) primary mine production (ca. 40 %),
2) by- or coproduction output from copper and scheelite mines (ca. 55 %),
3) imports from the People's Republic of China (3 %), and
4) molybdenum recovered from the processing of spent petroleum catalyst (2 %).

Reserves of molybdenite in the market economy countries have been estimated by evaluating identified ore bodies [7], i.e., those which have been explored as well as those which have been exploited. The results indicate the following amounts of recoverable molybdenum (10^3 t): United States, 4100; Chile, 1770; Canada, 928; Mexico, 306; Peru, 288; other countries, 356. Ore bodies producing primarily molybdenum contain 55 % of the reserves identified; only 29 % of these ore bodies were being exploited at the time of the survey (January 1985). Properties producing molybdenum as a byproduct contained the remaining 45 % of reserves and 67 % of them were producing molybdenite. The total recoverable molybdenum from primary and byproduct reserves has been estimated as follows (10^3 t molybdenum in concentrate) [8]:

Primary producing properties	1265
Primary nonproducing properties	3036
Byproduct producing properties	2316
Byproduct nonproducing properties	1142

Molybdenum occurs widely in all continents but usually in small quantities. The large-scale mining, milling, and processing facilities now required for economic production of molybdenum compounds are only justified where large

reserves of ore exist. Where these are exploited they are the primary molybdenum producing mines. In Canada, Endako (Placer Dome) is the only primary molybdenum mine and is in British Columbia. The Cumobabi mine (Empresso Frisco SA de CV) is in Mexico; the Jinduicheng mine is the largest molybdenum producer in China and all of the others, except perhaps for some in the Soviet Union, are in the United States. In Alaska, Quartz Hill is under long-term development. The famous, very large Climax and Henderson mines (both Amax) are in Colorado; Thompson Creek (Cyprus Minerals) is in Idaho; Tonopah (Cyprus Minerals) is in Nevada; and Questa (Molycorp) is in New Mexico. A detailed list of major molybdenum-producing companies is given in [8] (see also Chap. 10).

Molybdenum is extracted as a byproduct or coproduct of copper production in Canada, Chile, Iran, Mexico, Papua New Guinea, Peru, Philippines and, in the United States (Arizona, Nevada, New Mexico, and Utah). It is extracted from the scheelite mine at Sang-dong, Gangwon-do, Korea (Korea Tungsten Mining) as a byproduct of tungsten recovery.

4. Production

4.1. Concentration [9]

Molybdenum ore is mined by underground and open-pit methods. A typical primary molybdenum ore body contains 0.05 – 0.25 % Mo, and secondary ore bodies (copper porphyry ores) average 0.3 – 1.6 % Cu and 0.01 – 0.05 % Mo. Flotation (→ Flotation) remains the preferred and almost exclusive method of upgrading the molybdenite mineral to an industrial-grade molybdenite concentrate containing 90 – 95 % MoS_2 [*1317-33-5*], the balance being gangue.

The flotation process [10], [11] starts by pulverizing the ore to liberate the molybdenite (either alone or in association with copper sulfide minerals) from the host rock (mainly quartz monzonite), and then agitating the pulverized ore with water, a collector oil, and other special chemicals to cause preferential wetting of the host rock particles. The unwetted copper and molybdenum mineral particles are carried to the surface by air bubbles as the wetted host rock particles settle or remain suspended. A frothing agent is used to stabilize the bubbles, which contain the sulfide minerals, for easy skimming from the surface.

Molybdenite flotation is so spontaneous that it is termed a natural floater. Other minerals, such as clay and covellite (CuS), that are commonly associated with molybdenite-containing ores are, however, also natural floaters. For this reason the flotation process for molybdenum production from primary ore bodies differs from that of secondary ore bodies. Factors such as the grinding method, particle size, alkalinity, slurry density, and reagent usage dictate the recoveries and process economics.

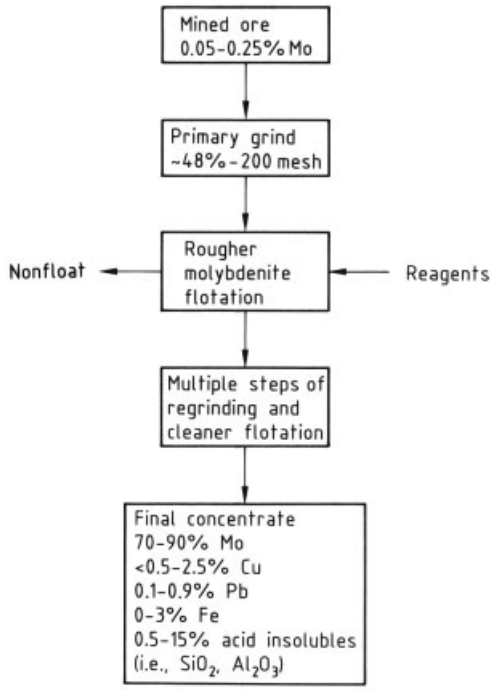

Figure 1. Primary molybdenite recovery process [9]

Recovery from Primary Ore Bodies. Figure 1 illustrates a basic recovery flowsheet from a primary molybdenum ore body. Although the methods of comminution vary, the objectives are to break down the ore to liberate the molybdenite. For example, Climax uses primary, secondary, and tertiary crushing with screening and rougher ball mills in closed circuit with

spiral and hydrocyclone classification to produce a flotation feed [12]. Henderson [13] and Thompson Creek [14] use primary crushing and semi-autogenous grinding in closed circuit with hydrocyclone classification. Detailed descriptions of comminution methods are given elsewhere, → Size Reduction.

An initial primary grind [ca. 48 % – 200 mesh (74 μm)] is usually acceptable to produce a 2.5 – 5.0 % Mo concentrate in the first "rougher" flotation stage. This concentrate is then upgraded by further grinding and "cleaner" floating. This concept of upgrading progressively smaller volumes of concentrate has proved to be the most cost effective means of handling large tonnages of low-grade molybdenum ores.

Sufficient regrind and cleaner steps follow to obtain the degree of liberation required for a final concentrate grade. For example, six cleaner flotation steps are used at Thompson Creek [14], five at Climax [12], and four at Henderson [13]. The nonfloat from the last stage of cleaner flotation is circulated countercurrently through the previous cleaner stages. Overall recovery is generally about 85 %.

At each flotation step molybdenite particles are floated and collected, and contaminants are depressed and rejected by adding reagents. Different reagent programs are used depending on the properties of the mined ore [15]. Simple hydrocarbon collectors can be used such as kerosene, stove oil, or light oil. Pine oils and Syntex VB (sulfated monoglyceride of coconut oil), which also have emulsifying and collecting properties, are used as frothers; lime or soda ash is used to control alkalinity, usually at pH 8.5. Contaminants are rejected by controlling the pH at an optimum level and by adding depressing reagents such as Noke's reagent (a product formed by reaction between stoichiometric quantities of sodium hydroxide and phosphorus pentasulfide) and sodium cyanide. Sodium silicate is used to disperse slimes.

Recovery from Secondary Ores. The generally accepted practice for secondary molybdenum recovery (Fig. 2) is to float a copper – molybdenum sulfide concentrate and then extract the molybdenum from the combined concentrate by methods governed by the type of copper mineralization present.

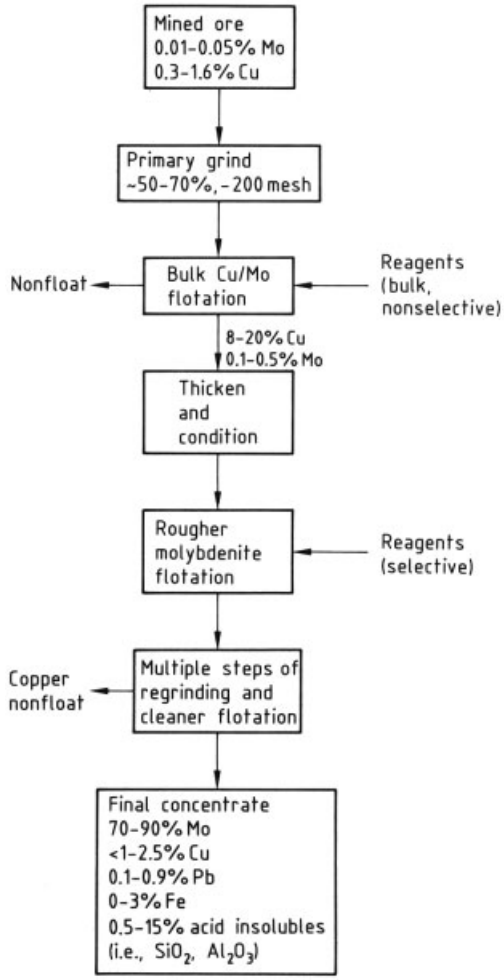

Figure 2. Byproduct copper – molybdenite recovery process [9]

In practice, fine grinding [50 – 70 % – 200 mesh (74 μm)] is usually performed to maximize copper recovery [16]. Then bulk flotation of the molybdenum – copper sulfide concentrate generally employs strong alkyl xanthates (e.g., potassium amyl xanthate or sodium isopropyl xanthate) and nonselective frothers (e.g., methylisobutylcarbinol) to float as much of the total copper – molybdenum product as possible.

The bulk copper – molybdenum flotation stream is thickened by settling the solids and decanting the clear process water to obtain a copper – molybdenum slurry with a higher solids content. The slurry is conditioned by

adding reagents during the thickening step as a method of premixing the reagents that will be used in the rougher flotation step.

On a worldwide basis, molybdenum losses in the copper – molybdenum bulk flotation process amount to about 48 % of the total molybdenum feed, but this level is gradually being reduced. Most of the molybdenum losses are incurred because the circuits are designed and operated for optimum copper recovery.

Selective collectors [17] are then used for separating molybdenum from the bulk copper – molybdenum concentrate beginning in the rougher flotation stages. For example, chalcopyrite ($CuFeS_2$), the most common type of copper mineralization, is particularly amenable to sodium hydrosulfide treatment. It is depressed with a 40 % solution of sodium sulfide or hydrosulfide. The hydrosulfide ion in the solution preferentially desorbs the bulk collector from the copper sulfide mineral, leaving it in the water-wetting, nonfloating condition.

Hypochlorides, peroxides, permanganates, and dichromates can also oxidize the adsorbed collectors giving an initial depressing effect to copper sulfides; cyanides will further depress copper to yield a high-grade molybdenite concentrate. When using these oxidants and depressants, chalcocite (Cu_2S) ores require more circuit design, i.e., additional flotation stages and more reagent addition points.

Covellite (CuS) ores float naturally and have been removed from flotation (depressed) with lignin sulfonate but now the simpler process of adding excess collector is used. Clays have been depressed with aluminum sulfate. Noke's reagent is a strong but short-lived depressant of copper sulfides. Lime is used for pH control at 9.5 – 11.5 to diminish pyrite flotation.

Sufficient grind and cleaner steps follow to obtain the desired degree of separation of the copper ore and molybdenite. Common reagents used for cleaner flotation are diesel oil and other light hydrocarbons. The nonfloat copper ore is filtered and dried before being converted to copper metal by smelting. The floated MoS_2 is filtered to remove bulk water. The wet, oily MoS_2 is not treated any further before roasting to MoO_3.

A careful balance of all reagent combinations must be determined in actual plant practice. The scarcity of collector oils has initiated intensive studies of collector substitutes, and environmental considerations are causing concentrators to move away from cyanides, arsenates, chromates, and other hazardous reagents.

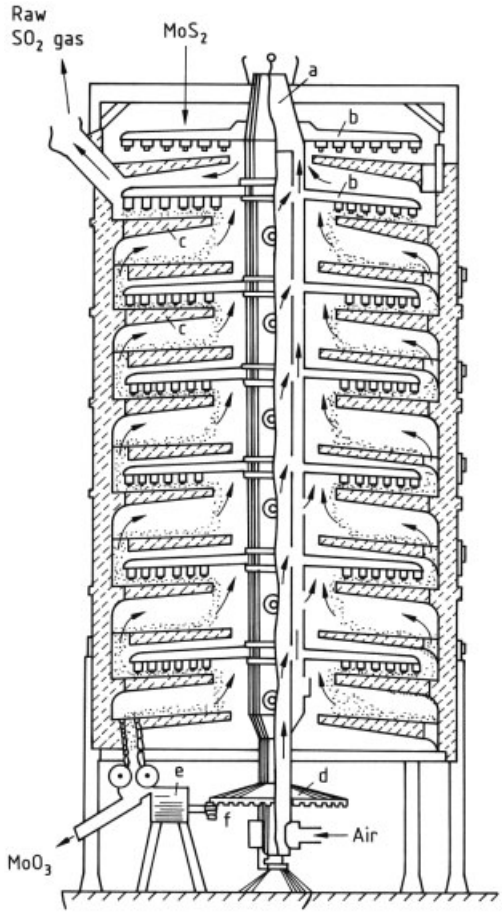

Figure 3. Multiple-hearth roasting furnace of the Nichols – Herreshoff type [20]
a) Rotating hollow shaft for passage of air to hearth; b) Rabble arms connected to shaft; c) Multilevel hearths; d) Beveled gear; e) Motor; f) Drive gear

4.2. Processing of Concentrate

Direct Use. Some commercial-grade molybdenite concentrate is subjected to additional grinding and flotation to produce lubricant-grade molybdenum disulfide (99 % MoS_2). Residual flotation oils and water are then volatilized by passing the upgraded concentrate through an inert-gas-swept kiln. This high-

purity product is added to grease and oils as a solid lubricant.

Production of Molybdenum Trioxide. The principal commercial molybdenum product is *technical-grade molybdenum trioxide* (MoO_3) [*1313-27-5*]. In addition to being the major compound for adding molybdenum to steel, it is also the starting compound for all other molybdenum products including ammonium dimolybdate, pure molybdenum trioxide, sodium molybdate, molybdenum metal (Sections 4.3 and 4.4), and ferromolybdenum (Chap. 6).

Molybdenite is converted to technical-grade MoO_3 by roasting it in air in a multiple-hearth furnace of the Nichols–Herreshoff or Lurgi design [18], [19], (Fig. 3). The roasted MoO_3 product typically contains $< 0.1\,\%$ sulfur. The multiple-hearth furnace is used because the hearths can be segmented to accomplish specific and different conversion operations within a single unit. Figure 4 shows the hearth-by-hearth composition of the molybdenum during roasting [18].

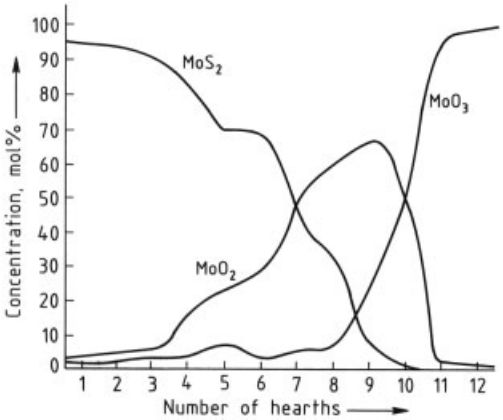

Figure 4. Composition of molybdenum sulfide and oxide in multiple-hearth roasting [18]

The upper hearths of the furnace are mainly used to burn off flotation oils and evaporate water. This part of the process is usually augmented by burning fossil fuels to ensure consistent temperature control and uniform roasting throughout the remainder of the unit. Upper hearth temperatures range between 600 and 700 °C.

The bulk of the sulfur is removed by oxidation in the intermediate hearths:

$$2\,MoS_2 + 7\,O_2 \rightarrow 2\,MoO_3 + 4\,SO_2$$

$$MoS_2 + 6\,MoO_3 \rightarrow 7\,MoO_2 + 2\,SO_2$$

$$2\,MoO_2 + O_2 \rightarrow 2\,MoO_3$$

Since the sulfur oxidation reactions are highly exothermic, additional heating is not needed in this section. The hearth temperatures are controlled between 600 and 650 °C by the addition of excess air and/or with water spray cooling.

As sulfur burning nears completion, the composition of the roaster hearth is mostly MoO_2 with $< 20\,\%$ MoO_3 and MoS_2. The reaction of MoO_3 with MoS_2 to form MoO_2 is driven by the strong oxidizing power of MoO_3, so that the production of MoO_3 is limited until most of the MoS_2 is gone.

Once the availability of MoS_2 is less than the availability of MoO_3, MoO_2 is rapidly converted to MoO_3. Although this reaction is exothermic, it is not sufficient to maintain the temperatures above 525 °C which are required to complete sulfur removal and conversion to MoO_3. Hence supplementary heating is required in these lower hearths. The oxide product is discharged from the roaster, cooled, and milled to form technical-grade MoO_3. Technical-grade MoO_3 typically contains 85–90 % MoO_3, the balance being silica with some Fe_2O_3 and Al_2O_3.

Sulfur dioxide and sulfur trioxide leave the roaster in the flue-gas stream. This gas stream is handled by methods which comply with emission regulations. The most common are contacting the off-gas with lime to produce disposable $CaSO_4 \cdot 2\,H_2O$ and the production of sulfuric acid. Sulfuric acid is produced by catalytically oxidizing the sulfur dioxide to sulfur trioxide which is absorbed by a H_2SO_4 solution.

In addition to sulfur removal, certain metal contaminants tend to vaporize during the roasting process. Any rhenium oxide [*1314-68-7*], Re_2O_7, in the concentrate can be scrubbed from the off-gas stream and recovered by solvent extraction or ion-exchange processes (→ Rhenium and Rhenium Compounds). Selenium oxide [*7446-08-4*], SeO_2, is also preferentially vaporized and can be collected to avoid creating an environmental hazard.

In many cases chemical leaching of the concentrate is required prior to roasting to preferentially remove nonferrous impurities detrimental to steel production. Three leaching processes are practiced:

1) sodium cyanide leachate is used for copper and gold removal;
2) iron(III) chloride for copper, lead, and calcium removal; and
3) hydrochloric acid for lead and bismuth removal.

Pure molybdenum trioxide can be produced from technical-grade MoO_3 via calcination of ammonium dimolybdate (ADM) or by sublimation.

Ammonium dimolybdate [*27546-07-2*] is obtained by first leaching the technical-grade oxide with water at ca. 75 °C to remove soluble alkali impurities, primarily potassium [20], [21]. The oxide is then filtered and dissolved in a 10 – 20 % ammonium hydroxide solution at 40 – 80 °C to form ammonium molybdate:

$$MoO_3 + 2\,NH_3 + H_2O \rightarrow (NH_4)_2MoO_4$$

The ammonium molybdate solution is separated from the insoluble gangue minerals (primarily silica and some iron) by filtration and then further purified by precipitation of the heavy metals (e.g., copper) with sulfide. The filtered solution is then fed to an evaporative crystallizer operating at atmospheric pressure to produce ADM crystals:

$$2\,(NH_4)_2MoO_4 \rightarrow (NH_4)_2Mo_2O_7 + 2\,NH_3 + H_2O$$

Varying degrees of rejection for sodium, calcium, magnesium, and aluminum occur during crystallization and these impurities build up in the mother liquor. They are eliminated by periodic purging of the crystallizer. The ADM crystals are recovered by centrifugation and then dried. Calcining the ammonium dimolybdate in an indirectly fired rotary kiln at temperatures in excess of 420 °C produces pure MoO_3.

Pure MoO_3 is also produced by *sublimation* of technical-grade MoO_3 [22]. The oxide is heated to 1100 – 1200 °C as a thin layer on a doughnut-shaped furnace hearth that rotates under electrical resistance heating elements. Air is drawn over the heated surface and sweeps away the vaporized MoO_3, leaving the nonvolatile gangue and other metallic impurity elements behind. The charge on the hearth makes one pass under the heating elements. The residue, still rich in molybdenum, is discharged for use as a steel additive. The sublimation yield is about 60 – 70 %. The sublimed vapor is then cooled, condensed, and collected in baghouses prior to densifying, screening, and packaging.

The molybdenum oxides produced by the ADM and sublimation processes are both of high purity (99.9 % MoO_3). Compared to calcined ADM, sublimed pure oxide has a significantly lower alkali content, but higher silica, iron, lead, and tin levels. Sublimed pure oxide also tends to be finer in size and acicular (needlelike) in shape.

Sodium molybdate [*7631-95-0*] (and other similar molybdates) may also be produced from technical-grade MoO_3 [23], although pure MoO_3 is a common, but more expensive, starting material. Technical-grade MoO_3 is leached as in the ADM process and then dissolved in sodium hydroxide at 50 – 70 °C to produce an almost saturated solution of sodium molybdate. A minimum excess of caustic (NaOH) is maintained to inhibit silica dissolution; a small quantity of sulfide ion is added to precipitate heavy metals. If pure oxide is used, the initial leaching step is eliminated.

After filtration, the solution is fed to an evaporative crystallizer to produce sodium molybdate dihydrate ($Na_2MoO_4 \cdot 2\,H_2O$). If crystallization is performed below 10 °C, the decahydrate is formed ($Na_2MoO_4 \cdot 10\,H_2O$). The crystals are recovered by centrifugation and dried at 70 °C. Anhydrous sodium molybdate (Na_2MoO_4) is produced by drying at > 100 °C.

4.3. Recovery from Spent Petroleum Catalysts

Molybdenum-containing catalysts are widely used in the petroleum refining industry for mild hydrogenation and removal of heteroatoms such as sulfur (hydrodesulfurization), nitrogen, and oxygen, as well as metals like nickel and vanadium. A typical catalyst uses molybdenum in combination with cobalt or nickel on a porous alumina (Al_2O_3) support. The composition range of such a catalyst when spent is 2 – 10 % Mo, 0 – 12 % V, 0.5 – 4 % Co, 0.5 – 10 % Ni, 10 % S, 10 % C, with the balance being Al_2O_3.

Numerous approaches have been explored to recover the metal values from spent catalysts. The fundamental strategy involves:

1) Initial heat treatment in air to remove the residual sulfur, carbon, and hydrocarbons,

and to oxidize the metals to soluble molybdate and vanadate.
2) A leaching step resulting in preferential solubilization of molybdate and vanadate, leaving the Ni – Co – alumina as a solid.
3) Separation of the Mo and V.
4) Treatment of the Ni – Co – alumina residue to recover Ni and Co.

Production of Molybdate and Vanadate. One basic process is to roast the ground (– 100 mesh, < 149 μm), spent catalyst in air at 600 °C and then mix the roasted material thoroughly with soda ash (Na_2CO_3) [24]. The Na_2CO_3 – spent catalyst mix is roasted again in air at 600 – 800 °C to convert the molybdenum and vanadium oxides to soluble sodium molybdate and sodium vanadate, respectively:

$$MoO_3 + Na_2CO_3 \rightarrow Na_2MoO_4 + CO_2$$

$$V_2O_5 + Na_2CO_3 \rightarrow 2\,NaVO_3 + CO_2$$

The roasted mix is leached with water at up to 100 °C to dissolve the soluble molybdenum and vanadium compounds. The molybdenum- and vanadium-containing solution is recovered by filtration for subsequent separation; the insoluble Ni – Co – alumina filter cake is treated separately to recover the respective metal values. A similar process uses sodium chloride instead of soda ash and results in the same sodium molybdate – sodium vanadate solution [25].

An alternative process uses wet-air oxidation to combine the first two oxidation steps [26]. The as-received and ground (– 100 mesh, < 149 μm) spent catalyst is slurried with water and soda ash at temperatures up to 300 °C and air pressures up to ca. 17 MPa (2500 psi). The molybdenum and vanadium sulfides are directly converted to soluble sodium molybdate and vanadate, respectively:

$$2\,MoS_2 + 6\,Na_2CO_3 + 9\,O_2 \rightarrow 2\,Na_2MoO_4 + 4\,Na_2SO_4 + 6\,CO_2$$

$$4\,VS + 6\,Na_2CO_3 + 11\,O_2 \rightarrow 4\,NaVO_3 + 4\,Na_2SO_4 + 6\,CO_2$$

In addition all the carbon and hydrocarbons are oxidized (burned off). The nickel and cobalt sulfides are also oxidized and the sulfur is converted to sodium sulfate. The Ni – Co – alumina remains insoluble as in the previous processes.

Recovery of Molybdenum and Vanadium. A number of technologies are available for recovering the metal values from the molybdenum- and vanadium-containing solution [26]. For example, addition of ammonium chloride or ammonium sulfate precipitates the vanadium as ammonium vanadate, $(NH)_4VO_3$, which is removed by filtration [24]. The molybdenum can then be precipitated as calcium molybdate by adding lime.

In another method, treating the solution with hydrogen sulfide under very acidic conditions (1 – 2 mol/L H_2SO_4) preferentially precipitates molybdenum trisulfide. The vanadium remaining in solution can be precipitated as the hydrated oxide, $V(OH)_4 \cdot 1.5\,H_2O$, by neutralization of the solution with sodium hydroxide, or as red cake ($Na_2H_2V_6O_{17}$) by first oxidizing the vanadium in solution with $NaClO_3$ and adjusting the pH to about 6 with NaOH or Na_2CO_3. In a final solvent extraction method the vanadium is preferentially extracted from the solution with a quaternary ammonium compound such as Aliquat 336 (tricapryl monomethyl ammonium chloride). The molybdenum can then be extracted with a secondary amine such as Adogen 283, a ditridecylamine.

Recovery of Nickel and Cobalt. A number of technologies are also available to recover the nickel and cobalt from the alumina. They can be leached from the alumina with 5 % HCl or H_2SO_4 solution at 75 °C [27]. They can also be recovered by caustic digestion of a slurry of the Ni – Co – alumina residue in a 25 % caustic (NaOH) solution at 250 °C for 2 h [26]. The dissolved nickel and cobalt are recovered and separated by other methods standard in the nickel industry.

4.4. Recovery during Production of Tungsten Ores

The occurrence of tungsten ores, such as scheelite ($CaWO_4$), is commonly associated with trace amounts of powellite ($CaMoO_4$). Most of the powellite stays with the scheelite during its processing to a sodium tungstate solution by soda ash digestion in an autoclave. The powellite

forms sodium molybdate and is solubilized. This is not a significant source of molybdenum and the recovered molybdenum is usually discarded because it is contaminated with tungsten. However the molybdate is routinely removed from the tungstate solution to purify the tungsten. For a full description of tungsten production, see → Tungsten and Tungsten Compounds.

The most common method of removing molybdenum from a sodium tungstate solution is to treat the solution with sodium sulfhydrate, $NaSH \cdot 2\,H_2O$ [28]. This compound reacts with the sodium molybdate and some of the sodium tungstate to precipitate the trisulfides, MoS_3 and WS_3, respectively. The precipitate generally contains two to three parts of molybdenum to one part of tungsten; it can be calcined to give the relevant oxides and then leached with strong hydrochloric acid to dissolve the molybdenum, leaving the tungsten.

4.5. Production of Molybdenum Metal Powder [29]

Molybdenum metal powder is produced industrially by reducing high-purity molybdenum compounds with hydrogen. The following compounds are used:

1) Molybdenum trioxide, MoO_3 (grey – green powder).
2) Ammonium hexamolybdate, $(NH_4)_2Mo_6O_{19}$ (yellow powder).
3) Ammonium dimolybdate, $(NH_4)_2Mo_2O_7$ (white powder).

To allow further processing by powder metallurgy, reduction conditions should result in a powder that can be pressed and sintered. Reduction is usually performed in two stages:

1) Reduction to MoO_2 (molybdenum red).
2) Reduction of MoO_2 to molybdenum metal powder.

This gives satisfactory yields of a completely reduced powder with the desired particle size distribution and a sufficiently high specific surface area in a reasonable length of time. Since reduction to MoO_2 is exothermic, this step is performed at 600 °C to prevent caking due to the melting of MoO_3 (*mp* 800 °C). The MoO_2 is reduced to molybdenum powder at ca. 1050 °C. The powder has a particle size of 2 – 10 µm, a specific surface area of 0.1 – 1 m^2/g, and an oxygen content of 100 – 500 mg/kg (partly adsorbed and partly as oxide).

Reduction is usually performed in an electrically heated continuous furnace (e.g., a pusher furnace, a walking beam furnace, or a rotary kiln) with a counterflowing stream of hydrogen. Reduction in fluidized-bed reactors is also known. Introduction of impurities should be avoided at all stages. Molybdenum powder 5N5 with an impurity content of ≤ 5 mg/kg can be produced; this figure does not, however, take into consideration gases and refractory metals. Special purification procedures such as multiple recrystallization, ion exchange, solvent extraction and adsorption are used to remove undesirable impurities such as alkali metals, ferrous metals, and radioactive elements from molybdenum to be used in the semiconductor industry. Special mixer linings, furnace linings, and sintering boats prevent contamination with iron and ceramic particles.

Good flow properties and a constant apparent density are critical for satisfactory processing of molybdenum powder in automatic presses. Molybdenum powder is typically mixed with an organic binder, e.g., poly(vinyl alcohol), and a volatile agent, e.g., water. The slurry is sprayed into a heated free-fall chamber where spherical agglomerates are formed due to surface tension. Heating causes vaporization of the volatile agent, giving a hard, densely packed agglomerate (spray-drying process).

4.6. Production of Compact Molybdenum Metal [30], [31]

More than 95 % of the total world production of molybdenum and molybdenum alloys high in molybdenum is produced by powder metallurgy, the remainder is obtained by vacuum arc remelting and electron beam melting.

Molybdenum produced initially by melting has to be worked by extrusion due to its coarse-grained, columnar microstructure; metal produced by powder metallurgy can, however, be worked by almost all standard hot metalworking techniques.

Production by Powder Metallurgy. More than 99 % of industrially produced molybdenum is prepared by the method shown on the left-hand side of Figure 5. In the case of large workpieces and round bars, the molybdenum powder is filled into rubber containers that are sealed and then compressed isostatically at 100 – 300 MPa. Due to frictional losses, a somewhat higher pressure (150 – 500 MPa) is used with die presses.

Sintering with resistance heating according to the Coolidge method is still used but indirect heating with tungsten heating elements is now the most common method. This procedure has a lower specific energy consumption, produces a homogeneous structure, and can be used to sinter large and complex components. Large items weighing up to 5 t (such as are used in isothermal forging dies) can only be produced by this method.

Depending on furnace design and the molybdenum alloy, sintering is performed at 1600 – 2200 °C for 3 – 30 h in a hydrogen atmosphere. Minimum sinter densities of 90 %, but preferably 95 %, are required to avoid subsequent processing difficulties. Neck formation between the particles starts at ca. 1200 °C and shrinkage at ca. 1600 °C. Addition of small quantities of platinum, palladium, or nickel allows the sintering temperature to be reduced to 1300 °C. These dopants have a low solubility in molybdenum and accumulate at the grain boundaries where they accelerate diffusion but have an adverse effect on ductility.

Production by Melting. Vacuum arc furnaces are used for large-scale melting of molybdenum. Electron beam devices are, however, used for special applications, particularly for the production of sputtering targets. Presintered molybdenum rods are fed into the electric arc furnace as electrodes and melted. In electron beam devices, sintered or hot isostatically compressed molybdenum rods are used. The resulting molybdenum billets are extruded by the Sèjournet process. In this process the heated billet is rolled over a bed of glass powder, which softens during contact, insulates the tooling from the hot billet, and lowers oxidation.

More recent efforts have been directed at producing a fine-grained as-cast structure by vacuum double-electrode remelting or cold hearth electron beam melting. This should give molten molybdenum which can be processed immediately without intermediate hot extrusion.

4.7. Processing of Molybdenum

Manufacture of Semifinished Products. Molybdenum produced by powder metallurgy (PM molybdenum) can be processed into semifinished articles by almost all classical forming techniques. High temperatures are, however, necessary and special equipment and tools are required. In molybdenum produced by melting the initial forming steps are important for obtaining a fine-grained structure.

The sintered starting material first has to be preheated to 1200 – 1400 °C to prevent any damage occurring during the first deformation steps. In this temperature range the tensile strengths of some molybdenum alloys correspond to those of hot-work tool steels at 700 °C. Preheating can be performed in electrical or gas-fired furnaces in a neutral or reducing atmosphere. Processing is carried out in air even though considerable oxidation occurs at this temperature. Embrittlement does not occur because oxygen and nitrogen are almost completely insoluble in molybdenum.

As forming proceeds, ductility increases and the temperature can gradually be reduced to room temperature. After hot and cold working, the molybdenum is subjected to stress-relief annealing at ca. 800 °C.

Large items are produced by press forging. Round products (diameter 10 µm to ca. 200 mm) are manufactured by extruding (Sèjournet method), forging, swagging, rolling, or drawing. Sheets are available in thicknesses of 10 µm to 50 mm.

Manufacture of Finished Products. Molybdenum can be worked by most of the customary techniques (e.g., deep drawing, flow turning). The working temperature should be higher than the ductile-to-brittle transition temperature but lower than the recrystallization temperature (1100 °C).

Molybdenum can readily be machined by chip removal and by electrical discharge. Tools used for milling, turning, and planing should comply with ISO hardmetal K specifications and have a highly positive cutting geometry. Since the cutting edges of the tools must be extremely

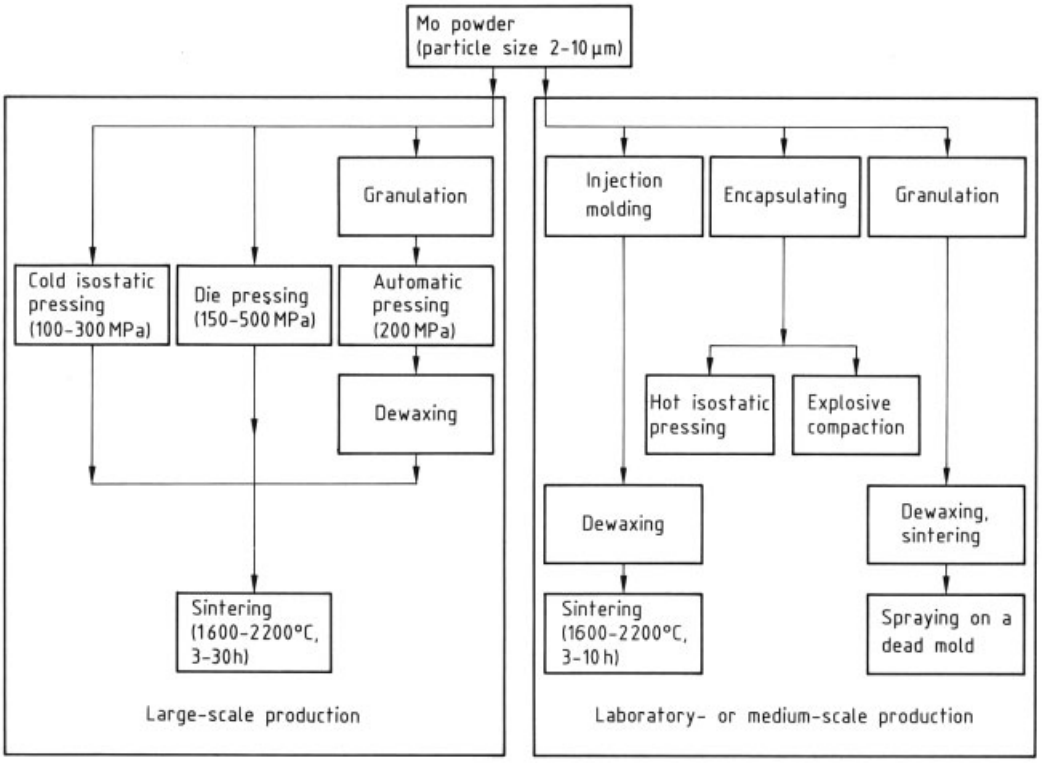

Figure 5. Production of molybdenum by powder metallurgy (PM)

sharp, coated indexable inserts are unsuitable. Commercially available emulsions or, in special cases, highly chlorinated hydrocarbons are used as coolants and lubricants.

When ceramic silicon carbide disks are used for grinding, low speeds must be used (≤ 33 m/s) to avoid buildup of heat. Molybdenum is machined by electrical discharge with zinc-coated brass wires and behaves similarly to hard metals. Graphite or tungsten – copper electrodes are recommended for cavity sinking by electrical discharge machining.

Joining [32]. Molybdenum and molybdenum alloys can be welded and brazed but mechanical joining methods are widely used to avoid embrittlement.

Welding. Tungsten inert-gas welding (in gastight chambers filled with argon) and electron-beam welding are usually employed. The oxygen partial pressure must be extremely low (< 0.5 Pa) because molten molybdenum rapidly takes up oxygen which leads to the precipitation of MoO_3 on cooling and thus to pore formation.

Diffusion welding at a welding temperature slightly below the recrystallization temperature (1100 °C) is particularly suitable for workpieces that are to be subjected to high mechanical loading. Resistance welding is primarily used in the electron tube and lamp industries for joining thin wires, ribbons, and foils.

Brazing is a suitable alternative to welding for operating temperatures below 1400 °C. Since the thermal expansion coefficient of molybdenum is about the same as that of graphite, molybdenum – graphite composites with advantageous properties can be produced, particularly for applications in nuclear fusion research and aerospace. Amongst others AgCu, PdCu, PdAg, Zr, and Ti are used as brazing filler materials.

Oxidation-Resistant Coatings [33]. Molybdenum alloys that can withstand oxidizing conditions at high temperature are still not available. Oxidation resistance decreases

drastically above 500 °C and above 600 °C significant loss of material occurs due to formation of volatile MoO_3. Owing to the increasing demand for high-temperature materials (e.g., in aerospace, turbine construction, and chemical reactors), oxidation resistant coatings are being developed to permit the use of molybdenum materials in these applications.

Coatings based on aluminides, silicides, ceramics (e.g., Al_2O_3, ZrO_2), and metals (chromium, noble metals) have been studied. Coating processes such as dip alloying, electroplating, but primarily pack cementation and slurry techniques, as well as plasma spraying, physical vapor deposition, and chemical vapor deposition have all been used.

Coatings based on silicides (e.g., $MoSi_2$, maximum use temperature 1600 °C) and complex silicides (e.g., SiCrFe, maximum use temperature 1500 °C) provide the best protection against oxidation but are only effective for a limited period.

Surface Hardening (→ Metals, Surface Treatment, Chap. 4.). Nitriding, carburizing, and boriding have been increasingly used to increase the surface hardness and abrasion resistance of molybdenum materials. These thermochemical processes harden the surface to a depth of ca. 0.2 mm. Plasma-assisted processes are also becoming more important.

4.8. Molybdenum-Base Alloys [34–36]

Solution hardening and particle hardening are responsible for the strengthening of molybdenum alloys. Commercially available molybdenum alloys are listed in Table 1. The most commonly used is TZM, an alloy that is particle-hardened with carbides. It has a high heat resistance, a high creep resistance, favorable welding properties, and a recrystallization temperature that is at least 300 °C higher than that of pure molybdenum (1100 °C). The MHC, ZHM, and TZC alloys are even more heat resistant due to their more stable carbides. Heat resistance can be increased even further by solution hardening with tungsten in alloys dispersion-hardened with hafnium carbide; ductility is, however, decreased.

Alloying with rhenium increases the strength and ductility, the latter effect is especially pronounced in alloys containing 40 – 50 % rhenium. These alloys also have excellent welding properties.

Doping with potassium silicate or potassium aluminum silicate and a high degree of deformation yield a material that forms large, oriented grains on recrystallization leading to excellent creep properties up to 1800 °C.

Molybdenum – tungsten alloys are extremely resistant to molten zinc and are therefore used in zinc metallurgy for tanks, pipes, pumps, and stirrers.

5. Uses

Molybdenum is mainly used as an alloying element in steel, cast iron, and superalloys (→ Steel) to increase hardenability, strength, toughness, and corrosion resistance. However, only the use of elemental molybdenum and molybdenum-base alloys is discussed in this chapter.

Initially, molybdenum was primarily used in the lamp industry. It is, however, now an increasingly important material in a wide range of applications (Table 2). In high-temperature applications, molybdenum competes with iron-, nickel- and cobalt-base superalloys, ceramics, and other high-melting metals (tungsten, tantalum, and niobium). Superalloys can be used up to 1200 °C; molybdenum materials, especially the carbide-hardened alloys and molybdenum doped with potassium silicate, show adequate heat resistance (see Fig. 6) and creep properties up to 1800 °C. The lack of resistance to oxidation should, however, always be borne in mind. Molybdenum materials have a higher failure tolerance and ductility than ceramics and are less expensive than tantalum and niobium. Heating elements, thermal radiation shields, and furnace ware made of molybdenum are used, for example, in hot isostatic presses, sintering furnaces for the ceramic industry, and heat treatment furnaces.

Table 1. Properties of commercial molybdenum-base materials (amount of deformation, 90 %)

Type of material	Composition, wt %	Hardening mechanism	Recrystallization temperature, ° C	Typical strength at 1000 ° C [a], MPa	ASTM B 386-85 material number [d]
Mo			1100	250	MM 360, 365 PM 361
TZM	Mo, 0.5 Ti, 0.08 Zr, 0.01 – 0.04 C	particle hardening	1400	600	MM 363 PM 364
MHC	Mo, 0.5 – 2.0 Hf, 0.04 – 0.2 C	particle hardening	1550	800	
ZHM	Mo, 0.4 Zr, 1.2 Hf, 0.1 C	particle hardening	1550	800	
TZC	Mo, 1.0 – 1.5 Ti, 0.2 – 0.35 Zr, 0.06 – 0.2 C	particle hardening	1550	800	
Mo – W – HfC	Mo, 25 – 45 W, 0.9 – 1 Hf, C	particle and solution hardening	1650	900	
Mo – Re	Mo, 3 – 50 Re	solution hardening	1200 (5 % Re) 1300 (41 % Re)	400 600	
Mo – W	Mo, 5 – 30 W	solution hardening	1100 (10 % W) 1200 (30 % W)	280 350	MM 366 (30 % W)
Mo doped with potassium silicate	Mo, 0.005 – 0.1 K, 0.005 – 0.1 Si, 0.01 – 0.2 O_2	particle hardening	1800 [b]	300 [c]	

[a] Subjected to stress relieving annealing.
[b] Amount of deformation 98 %.
[c] Recrystallized.
[d] Produced by powder metallurgy (PM) or melting (MM).

Table 2. Uses of molybdenum and molybdenum alloys

Area of application	Examples of uses
Lamp and lighting industries	support wires, sealing ribbons, dimming caps
Electronic and semiconductor industries	semiconductor base plates, heat sinks, contact pins, sputtering targets, control grids, klystron and travelling-wave tube components
High-temperature and vacuum furnace construction	heating elements, thermal radiation shields, furnace ware
Glass and ceramic industries	glass-melting electrodes, installations in glass production tanks, drawing dies, crucibles for producing sapphire single crystals
Casting technology and metalworking	forging dies for isothermal forging, extrusion dies, die-casting molds, hot-galvanizing equipment
Coating	spray wire, spray powder, evaporation boats, sputtering targets, components of chemical vapor deposition equipment
Nuclear technology	furnace parts and charging equipment for sintering UO_2; first wall, divertor, and limiter components for experimental fusion reactors
Medicine	rotating X-ray anodes, collimators

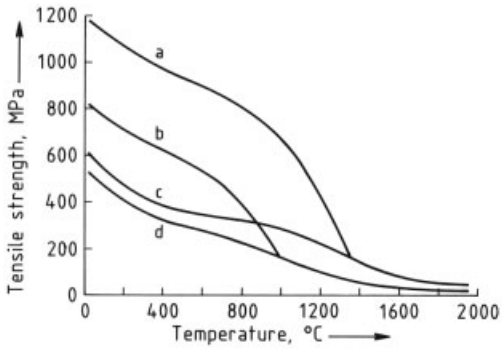

Figure 6. Tensile strengths of 1-mm thick molybdenum and TZM sheets as a function of temperature
a) TZM subjected to stress-relieving annealing; b) Molybdenum subjected to stress-relieving annealing; c) Recrystallized TZM; d) Recrystallized molybdenum

Molybdenum is resistant to corrosion by most glasses and is therefore used as an electrode material in electric glass-melting furnaces. Glass produced in such furnaces is of better quality than that produced in fuel-fired furnaces and the process is more acceptable environmentally.

Crucibles and mandrels used to draw quartz glass or crucibles used for making single sapphire crystals have to withstand temperatures up to ca. 2100 °C. Molybdenum ribbon is used for electrical lead-in (e.g., in halogen lamps). In this application a vacuum-tight seal is formed by squeezing the quartz tubing around the ribbon at high temperature (ca. 2100 °C) and pressure. The high heat resistance of molybdenum and its low coefficient of thermal expansion are important prerequisites for these applications.

Electronic applications have and will continue to be a high-growth market for molybdenum parts. Molybdenum disks are used as base plates and heat sinks for power transistors and silicon rectifiers.

Molybdenum is also an important sputtering target material for electronic applications, mainly for codeposition of $MoSi_x$ films from ultrahigh-purity molybdenum and silicon targets. Integrated circuits employing these films have low parasitic capacitances and also low gate and interconnection propagation delays.

Molybdenum layers produced by thermal spraying possess exceptionally good sliding properties and high abrasion resistance. They are deposited on machine components, especially car engine parts such as piston rings, synchrorings, and selector forks.

More than 80 % of the rotating X-ray anodes currently used in clinical diagnostics are made of molybdenum alloys. Loading can be increased by a thin tungsten – rhenium layer on the molybdenum body. Rhenium in tungsten has the same effect on ductility as rhenium in molybdenum and decreases the susceptibility to cracking initiated by thermal stress. Brazing a thick graphite disk onto the back of a target increases its heat and emission capacity.

6. Production of Ferromolybdenum

6.1. Ferromolybdenum Grades

In the steel and foundry industry molybdenum is added to melts, either as technical oxide or in the form of ferromolybdenum [*11121-95-2*]. Table 3 summarizes the internationally standardized ferromolybdenum alloys; the typical molybdenum content ranges from 55 to 75 %. Compared to pure molybdenum, ferromolybdenum dissolves much more easily in the steel melt and is cheaper to produce. Practical experience has shown that alloys with up to 75 % molybdenum do not cause any dissolution problems. Nevertheless there is still a demand for 65 % molybdenum alloys which are even less critical as regards dissolution characteristics. The Fe – Mo phase diagram is shown in Figure 7. Due to the limited information available, the melting temperatures of the alloy at high molybdenum concentrations can only be estimated. In practice they are likely to be considerably lower than indicated by the phase diagram because of significant concentrations of silicon and other impurities.

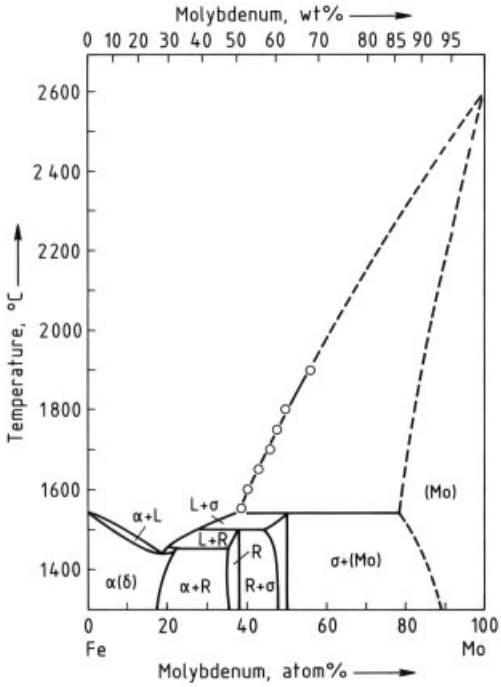

Figure 7. Fe – Mo phase diagram showing estimated liquidus (L) compositions at higher molybdenum concentrations [37]
α,δ, and R are intermetallic phases.

6.2. Raw Materials

The principal raw materials used for ferromolybdenum production are molybdenum concentrates (see Section 4.1) and technical-grade molybdenum trioxide (see Section 4.2). One method proposed for the production of

Table 3. International specifications for ferromolybdenum

Specification	Symbol/ designation	Composition, wt %							
		Mo	Si (max.)	C (max.)	S (max.)	P (max.)	Cu (max.)	Pb (max.)	Sn (max.)
ASTM A 132-74 (reapproved 1979)	Ferromolybdenum (formerly grade B)	60.0 min.	1.0	0.10	0.15	0.05	1.0	0.01	0.01
DIN 17 561	FeMo 70 (0.4270 [a])	60 – 75 [b]	1.0	0.10	0.10	0.10	0.5		
	FeMo 62 (0.4262 [a])	58 – 65 [b]	2.0	0.50	0.10	0.10	1.0		
ISO 5452	FeMo 60	55.0 – 65.0 [c]	1.0	0.10	0.10	0.05	0.5		
	FeMo 60 Cu 1	55.0 – 65.0	1.5	0.10	0.10	0.05	1.0		
	FeMo 60 Cu 1.5	55.0 – 65.0	2.0	0.50	0.15	0.05	1.5		
	FeMo 70	65.0 – 75.0 [c]	1.5	0.10	0.10	0.05	0.5		
	FeMo 70 Cu 1	65.0 – 75.0	2.0	0.10	0.10	0.05	1.0		
	FeMo 70 Cu 1.5	65.0 – 75.0	2.5	0.10	0.20	0.10	1.5		
Japanese Industrial Standard, JIS G 2307 1978									
Class:									
High-carbon ferromolybdenum	FMoH	55.0 – 65.0 [c]	3.0	6.0	0.20	0.10	0.5		
Low-carbon ferromolybdenum	FMoL	60.0 – 70.0 [c]	2.0	0.10	0.10	0.06	0.5		

[a] Material numbers in preparation, proposed numbers.
[b] The range of variation within a lot may not exceed ± 2 wt %.
[c] The range of variation within a graded lot should not exceed 3 wt % absolute.

ferromolybdenum used molybdenum concentrates which were reacted with ferromanganese to form ferromolybdenum and manganese sulfide [38]:

$$MoS_2 + 2\,Mn \rightarrow Mo + 2\,MnS$$

Due to the high Gibbs free energy of manganese(II) sulfide formation, this reaction is claimed to yield ferromolybdenum with a relatively low sulfur concentration, but a further sulfur removal step appears to be necessary. This process is not, however, used on a commercial basis.

Technical-grade molybdenum trioxide is commonly used as the molybdenum raw material. Because of its high value and the fact that MoO_3 starts to sublime at 700 °C, (see Section 7.2), electrostatic precipitators or bag houses must be used to collect the molybdenum dust and sublimates for recirculation. Good housekeeping is also essential to minimize molybdenum handling losses.

Two principal processes are used for molybdenum oxide reduction: *carbothermic reduction* in a submerged arc furnace and *metallothermic* reduction. The most commonly applied method is the silicothermic reduction, a type of metallothermic reduction.

6.3. Submerged Arc Furnace Carbothermic Reduction

The carbothermic reduction of MoO_3 can proceed as follows:

$$2\,MoO_3 + 2\,C \rightarrow 2\,MoO_2 + 2\,CO$$

$$MoO_2 + 2\,C \rightarrow Mo + 2\,CO$$

or with formation of molybdenum carbide:

$$MoO_2 + 2.5\,C \rightarrow 0.5\,Mo_2C + 2\,CO$$

The Gibbs free energies of oxide and carbide formation are shown in Figure 8 as a function of temperature. The reduction of MoO_3 starts at fairly low temperatures. If the process were run at these temperatures (i.e., < 600 °C) then Mo_2C would be formed because it is more stable than molybdenum. However, since the process is run at 1700 – 2000 °C, any intermediately formed molybdenum carbide reacts to form molybdenum metal:

$$2\,Mo_2C + MoO_2 \rightarrow 5\,Mo + 2\,CO$$

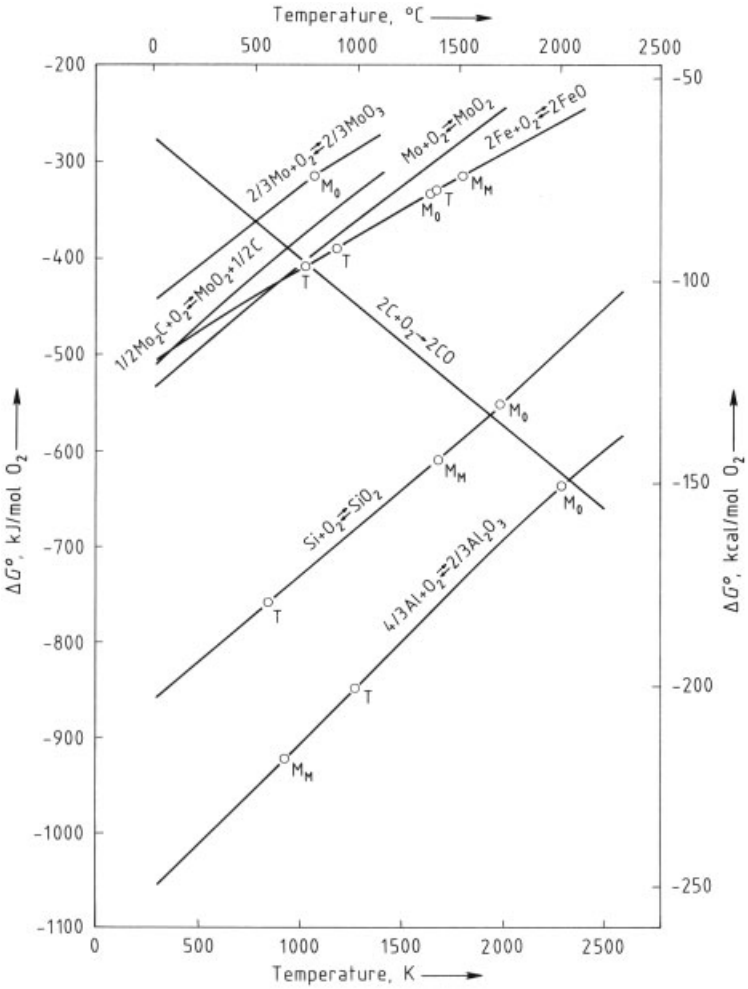

Figure 8. Gibbs free energy of oxide and carbide formation [39]
T = transformation point; M_M = melting point of metal; M_O = melting point of oxide

Under practical conditions the resulting metal alloy is low in carbon.

Because of the commercial disadvantages of the process, detailed literature describing the process is limited [40–43]. The process was previously run in single-phase 300 – 500 kW and three-phase 1500 – 2000 kW electric arc furnaces. The use of the single-phase arc furnace seems to have been more successful because of its higher energy density within the furnace. A single-phase system with a cylindrical magnesite or carbon-based rammed furnace body and equipped with a bottom carbon electrode is shown in Figure 9. The top electrode is adjustable and the whole furnace arrangement runs on tracks so that it can easily be replaced with a new unit. The process is run batchwise, producing a semimolten metal block and a slag. The slag contains up to 10 % molybdenum oxide and must be retreated by remelting. Melting of a 1.5 – 2 t metal block takes about 20 – 28 h. To separate metal and slag, the furnace assembly must be stripped and later rebuilt for reuse.

Raw materials are technical-grade molybdenum oxide; iron ore or millscale, steel punchings, nail nips or turnings; lime and fluorspar as fluxing agents; and charcoal as reducing agent. To minimize MoO_3 sublimation, the concentrate and carbon material are briquetted. The reaction is started by melting slag from a previous re-

action and adding the briquetted mixture to the slag to avoid molybdenum sublimation. Once the initial reduction to MoO_2 has taken place, sublimation is avoided. The resulting slag consists of some unreduced MoO_2 and FeO which has a higher Gibbs free energy than MoO_2 (see Fig. 8). To improve metal – slag separation, lime and fluorspar are added. The slag should be fluid and is tapped toward the end of the reduction cycle. The metal block is stripped while it is still red hot and quenched in water to aid subsequent crushing. The metal must be hand-sorted to recycle any parts of the block contaminated with slag. The resulting molybdenum-rich slag has to be remelted in a separate process into a low-molybdenum, high-carbon intermediate alloy which is recycled.

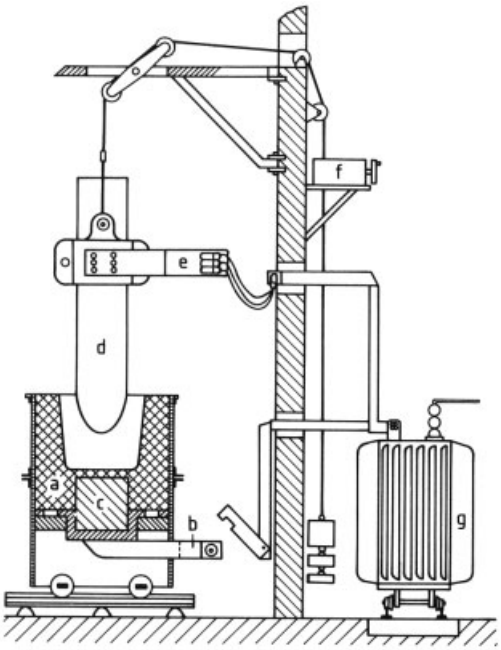

Figure 9. Single-phase submerged arc furnace [43]
a) Furnace with rammed lining; b) Bottom electrode power supply bar; c) Bottom electrode (carbon); d) Top electrode; e) Electrode clamp with power supply cables; f) Electrode control; g) Transformer

Table 4 shows the consumption of electric energy and ingredients required to produce ferromolybdenum using the carbothermic process.

Table 4. Consumables required to produce 1 t of ferromolybdenum in the carbothermic process [43]

Consumable	Single-phase furnace, 300 kW	Three-phase furnace, 1500 kW
Roasted Mo concentrate, kg	1160	1420
Remelt material, kg	280	330
Slag treatment metal, kg	165	168
Nail nips, kg	172	230
Platescale, kg	290	102
Lime, kg	150	178
Fluorspar, kg	150	178
Charcoal, kg	255	310
Electrodes, kg	95	135
Power, kW · h	4300	4450
Products		
Ferromolybdenum, kg	1000	1000
Remelt, kg	270	250
Dust, kg	66	

6.4. Metallothermic Reduction

Metallothermic reduction in which silicon (or aluminum) is used to reduce the MoO_3 is the commonly used production process for ferromolybdenum. For a detailed description of this technique, see → Aluminothermic Processes. Its advantages are that it is a single-step process which results in a saleable metal and a low-molybdenum slag and that it requires only basic process equipment. Due to the large difference in the Gibbs free energy for aluminum oxide and silicon dioxide formation compared to molybdenum oxide formation (Fig. 8), the reduction process approaches 100 % molybdenum reduction.

An aluminothermic or silicothermic reaction must be sufficiently exothermic to melt all the reaction constituents resulting in satisfactory separation of the liquid metal and liquid slag phases before cooling. According to SHEMTCHUSHNY, a minimum specific heat of reaction of − 2700 kJ/kg is required for a reaction to be selfsustained [44]. This assumes that no additional heat is added to the reaction (e.g., by preheating the charge or by adding booster oxides and further reducing metal). Consideration of the specific heats of reaction shows that both aluminum and silicon can be used to reduce MoO_3 in selfsustained reactions:

$$2/3\ MoO_3 + Si \rightarrow 2/3\ Mo + SiO_2 \quad -3241\ kJ/kg$$

$$1/2\ MoO_3 + Al \rightarrow 1/2\ Mo + 1/2\ Al_2O_3 \quad -4689\ kJ/kg$$

If aluminum is reacted with MoO_3, the resulting heat of reaction is very high, leading to excessive evaporation of MoO_3 and the danger of an explosion of the reaction mixture.

Consequently, silicon is the preferred reducing element. Furthermore, less silicon is necessary compared to aluminum (stoichiometric consumption is 0.439 kg silicon per kilogram of molybdenum, compared to 0.562 kg aluminum per kilogram of molybdenum) and silicon in the form of ferrosilicon is considerably cheaper than aluminum.

Practical experience, even when using charges much larger than a 200 kg, shows that silicon alone does not result in a selfsustained reaction. Some of the silicon must therefore be replaced with aluminum or other elements with a similar oxygen affinity. Some of the required iron is often added in the form of millscale or iron ore concentrates which react with silicon or aluminum to generate additional heat. Part of the iron remains as FeO in the slag and acts as a fluxing agent. Further fluxing agents are lime and fluorspar, which are necessary to ensure good separation of metal and slag and low molybdenum concentrations (typically $< 1\ \%$) in the slag. The dust generated in the process can contain up to 5 % of the molybdenum in the charge. Any dust and sublimed MoO_3 must be extracted, recovered, and recycled to achieve an overall molybdenum yield of 97 – 99 %.

To achieve reproducible reactions, a homogeneous mixture of the precisely weighed reaction ingredients is required. All ingredients must be dry, well-calcined, and ideally their particle size should be < 1 mm. A typical firing arrangement is shown in Figure 10. The crucible unit consists of a circular or square, sand- or brick-lined base section and a cylindrical refractory-lined, steel section that contains the reaction slag. The mix feeding arrangement allows part of the reaction mixture to be placed into the well-dried crucible unit. The reaction is started by adding a small amount of a mixture of barium peroxide and aluminum to a special starter mix which tends to be more exothermic than the main mix. Once the main mix has started to react, the bulk of the reaction mix is continuously fed into the crucible arrangement. Total reaction times of 30 – 60 min for metal block weights of 1 – 4 t can be expected. After completion of the reaction, metal and slag are allowed to segregate for another 30 – 60 min before the bulk of the slag is tapped off. The red-hot metal block is then stripped off the furnace base and quenched in water to aid crushing. Jaw crushers are commonly used for crushing the metal. Considerable care has to be taken when sampling the metal for molybdenum content because the molybdenum tends to segregate within the block. Any highly segregated metal and metal contaminated with slag inclusions is recycled for remelting. Good housekeeping is necessary to recycle all splashings, floor sweepings etc. and achieve a good total molybdenum yield.

In a process modification used by Duval Corporation, Duval Sierrita [46], the entire reaction mix is charged into the reaction crucible before ignition. The advantage is that controlled addition of the reaction mix is not required. The disadvantages are that the reaction speed cannot be controlled by varying the feed rate and that a larger crucible unit is required to contain the initial mixture.

More recent developments have centered on improving process economics. Automatic handling and weighing of the raw materials as well as finished metal and ferromolybdenum slag is one area of activity. Other activities center around the optimization of ingredient and consumable costs.55

Metallothermic mixtures used to produce ferromolybdenum are summarized in Table 5. Mixtures A – D represent different possibilities of increasing the reaction heat by using aluminum or calcium in the form of a calcium – silicon or a ferrosilicon aluminum alloy [43]. Reaction D can be regarded as a fairly typical ferromolybdenum reaction mix; it gives 87 kg ferromolybdenum with 67 – 73 % molybdenum, 0.05 – 0.08 % carbon, 0.03 – 0.05 % sulfur, 0.25 – 0.6 % silicon, and 5 kg remelt metal as well as a slag containing 0.34 % MoO_3, 13.85 % Fe_2O_3, 7.4 % CaO, 9.9 % Al_2O_3, and 64.5 % SiO_2. The reaction mix also contains recycled dust and metal remelt which is normal practice when continuously producing ferromolybdenum.

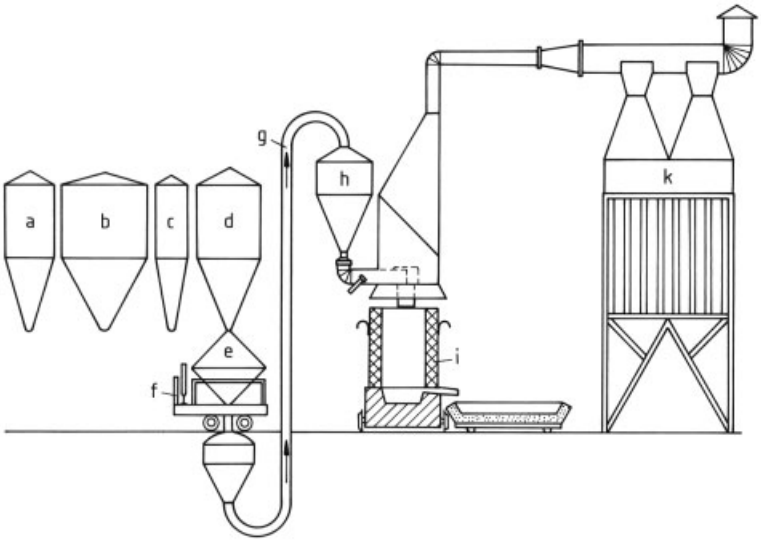

Figure 10. Metallothermic firing arrangement [45]
a) Storage hopper for iron ore or millscale; b) Storage hopper for roasted Mo concentrate; c) Storage hopper for fluorspar; d) Storage hopper for aluminum and ferrosilicon (75 % Si) powder; e) Mixing hopper; f) Balance; g) Air lift system; h) Firing hopper; i) Reaction vessel; k) Electrostatic precipitator

Table 5. Reaction mixtures used for ferromolybdenum production

Component	Durrer, Volkert [43]				Climax Molybdenum			Literature from the former Soviet Union		
					[46]	[47]	[47]	[48]	[49]	[50]
	A	B	C	D	E	F	G	H [a]	I [b]	J [c]
Roasted Mo concentrate, kg	100	100	100	100	100	100	100	100	100	100
MoO_3 concentration, wt %	ca. 90	ca. 90	ca. 90	ca. 90	85	85	85	81.7 – 84	80.3	85
Platescale, kg	29	16 – 18	35	51		24.2	22.5			
Fe ore, kg					26.9			26	22	29.6
Fe concentration, wt %					69				65.7	67.1
Fe turnings, kg	5 – 6	18						19.5		
Fe scrap, kg						17.3 [d]	5.2 [d]		19	19.2
Ferrosilicon, kg	31.5	21 – 23	65 [e]	38	48.9	20.8	31.2	35.1	31.6	36.8
Si concentration, wt %	75	75	45	75	50	75	75	75	77	73.6
Ca–Si alloy, kg		17 – 18								
Al powder, kg	9.0			5	5.1	12.1	6.9	3.7	4.3	3.8
Al concentration, wt %	99.5			99.5	93				93	99
Lime, kg				2.5	7	8.7		3.0		3
Fluorspar, kg			7	7.5	2.2		5.2	2.0	3	2
CaF_2 concentration, wt %					95					
Dust, kg				9.5			5.2			
Remelt, kg				5.5			6.9			

[a] Up to 36 mixes used in one reaction.
[b] Mix preheated to 150 °C.
[c] Calculated mix.
[d] Fe shot.
[e] FeSi 45 Al 10 alloy.

Mixtures E – G have been reported to have been used by Climax Molybdenum in their Langeloth (USA) plant [46]. The mixtures are based on 100 kg technical molybdenum oxide input with 85 % MoO_3 in the technical oxide. Reaction E can be regarded as a typical reaction mix for a ferromolybdenum with 58 – 64 % molybdenum. The maximum impurity levels quoted were 0.1 % carbon, 0.25 % sulfur, 0.5 % copper, and 1 % silicon. Reaction G can be regarded as a typical reaction mix yielding an approximately 70 % molybdenum alloy.

Mixtures H – J have been taken from literature from the former Soviet Union [48–50]. Mixture H gives ferromolybdenum containing 60 – 65 % molybdenum, 0.15 – 0.9 % silicon, 0.04 % phosphorus, 0.08 % sulfur, 0.05 % carbon, and ca. 0.5 % copper and slag containing 0.06 – 0.17 % molybdenum, 67 – 70 % SiO_2, 7 – 11 % FeO, 10 – 13 % Al_2O_3, 6 – 8 % CaO, and 1 – 3 % MgO. Mixture I uses a mix preheating to 150 °C, and should have resulted in a ferromolybdenum with ca. 56 % molybdenum (calculated). Mixture J is part of a calculated formulation according to Elyutin [50]. The calculation is based on 96.8 kg ferromolybdenum containing 58.14 % molybdenum, 0.53 % silicon, and 0.07 % carbon and 87 kg slag containing 65.65 % silicon oxide, 10.06 % aluminum oxide, and 24.28 % iron (II) oxide. A heat balance has been calculated for this formulation. The total heat generated is 422.45 MJ. This heat is assumed to be consumed to provide the latent and melting heat for the metal (25 %), the slag (50 %), and to cover the heat losses of 25 %.

7. Molybdenum Compounds

For molybdenum carbide, see → Carbides, Chap. 2.9.

7.1. Overview of Molybdenum Chemistry

As a transition element, molybdenum shows variable valency and forms many complexes and colored compounds. The outstanding feature of molybdenum is its chemical versatility: oxidation states from 2 − to 6 +; coordination numbers from 4 to 8 and, accordingly, a very varied stereochemistry; the ability to form compounds with most inorganic and organic ligands, with a particular preference for oxygen, sulfur, fluorine, and chlorine donor atoms; the formation of bi- and polynuclear compounds containing bridging oxide and chloride ligands and/or Mo – Mo bonds.

Molybdenum is very similar to tungsten, the third member of Group 6, but different from the first member, chromium. Molybdates(VI), unlike chromates(VI), are not strong oxidizing agents: much of the familiar chemistry of molybdenum is that of the 6 + and 5 + oxidation states.

In recent years research on molybdenum compounds has greatly increased because of their academic interest and their developing technical applications as lubricant additives (molybdenum disulfide and sulfur complexes), corrosion inhibitors, pigments, smoke suppressants, and heterogeneous catalysts (molybdenum trioxide and molybdates).

Recognition of the central role of molybdenum in a number of metalloenzymes (nitrogenase, nitrate reductase, xanthine oxidase) and the problem of molybdenum toxicity in ruminants have stimulated research in both molybdenum biochemistry and inorganic chemistry. The fact that molybdenum compounds are nontoxic to humans has lead to interest in the possibility of replacing compounds of toxic metals with safer molybdenum compounds, e.g., chromium in corrosion inhibitors, antimony in flame and smoke suppressants.

For more background information on molybdenum chemistry, see [51–56].

Compound Types. Typical compounds of molybdenum(VI) are the trioxide [*1313-27-5*], MoO_3; sodium molybdate [*7631-95-0*],Na_2MoO_4; and ammonium heptamolybdate, [*12054-85-2*], $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$ and [*10102-40-6*], $(NH_4)_6Mo_7O_{24} \cdot 2H_2O$. These compounds are the starting points for the preparation of many other molybdenum compounds. The chemistry of the higher oxidation states (4^+ to 6^+) is dominated by oxomolybdenum cations, e.g., MoO^{2+}, $Mo_2O_4^{2+}$, $Mo_2O_3^{4+}$, MoO^{3+}, MoO_2^{2+} (see Section 7.6). The oxide ligands may be bound to just one molybdenum (terminal oxide) or to two

or more molybdenum atoms (bridging) [57]. Replacement of one or more oxide ligands by sulfide has produced many molybdenum – sulfur complexes and clusters [58], [59].

The higher halides are covalent, e.g., molybdenum hexafluoride [*7783-77-9*], MoF_6, and molybdenum pentachloride [*10241-05-1*], $MoCl_5$.

Compounds of molybdenum(II) and lower oxidation states are stabilized by unsaturated ligands such as carbon monoxide, e.g., in molybdenum hexacarbonyl [*13939-06-5*], $Mo(CO)_6$, and by molybdenum – molybdenum bonds, and so are bi- or polynuclear, e.g., molybdenum(II) acetate [*14221-06-8*], $Mo_2(CH_3COO)_4$, and molybdenum(II) chloride [*11062-51-4*], Mo_6Cl_{12}.

Preparative Chemistry. The starting point in preparative molybdenum chemistry is molybdenum(VI) oxide, MoO_3 (Fig. 11). Molybdenum disulfide, the naturally occurring form of molybdenum, is unreactive.

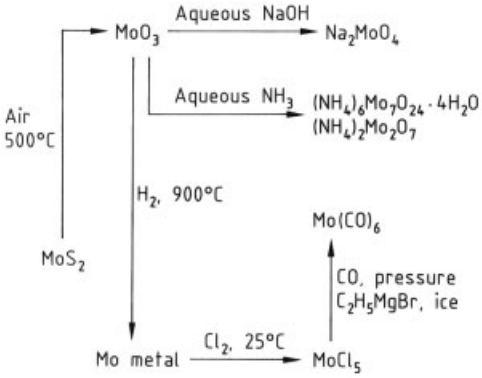

Figure 11. Relationship between some molybdenum compounds

Aqueous Solution Chemistry. The aqueous chemistry of molybdenum is complicated [52]. It is dominated by oxo species which are prone to dimerize or polymerize. Hydrated cations (i.e., cations having only water bound to molybdenum) are known only for oxidation states 2+ and 3+; they are powerful reducing agents which oxidize in air. The aqueous chemistry of molybdenum is summarized in Figure 12. The aqueous ion of molybdenum(IV) (not shown in Fig. 12) is the red trimer, $[Mo_3O_4(H_2O)_9]^{4+}$ [*74353-85-8*].

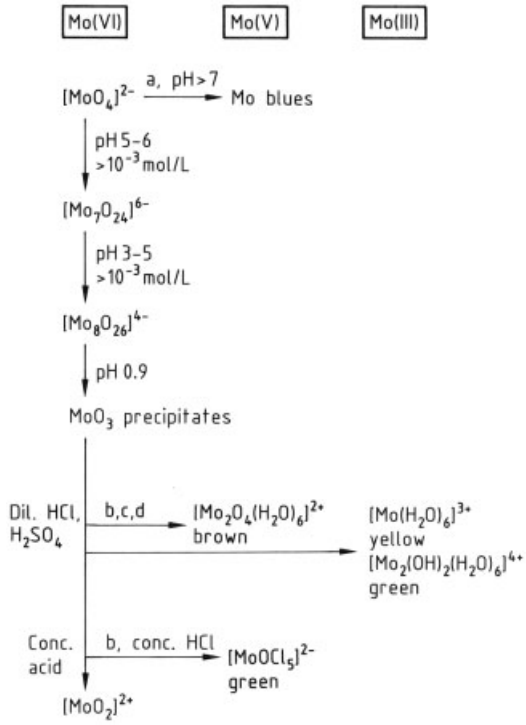

Figure 12. Aqueous chemistry of molybdenum showing species [52], reducing agents, and conditions
Reducing agents: a) Sodium dithionite, $Na_2S_2O_4$; b) Hydrazinium hydrate, $N_2H_4 \cdot H_2O$; c) Tin(II) chloride; d) Liquid mercury

Redox Chemistry. Molybdenum exemplifies the general trend of a transition metal group; the higher oxidation states of the heavier elements (molybdenum and tungsten) are more stable (resistant to reduction) and the lower oxidation states less stable (more prone to oxidation) than those of the first element (chromium). Reduction potentials are summarized in Figure 13 [60]. Note that reduction in water bypasses molybdenum(IV).

For the reduction of molybdate in neutral or alkaline solution, sodium dithionite is useful, e.g., in the preparation of the molybdenum(V) cysteine complex [*26998-42-5*], $Na_2[Mo_2O_4\{SCH_2CH(NH_2)COO\}_2] \cdot 5\,H_2O$ [61].

Molybdenum(III) solutions oxidize rapidly in air and must be handled under nitrogen or argon. With molybdenum(V), aerial oxidation is slow. Titration of reduced molybdenum solutions with cerium(IV) sulfate or permanganate

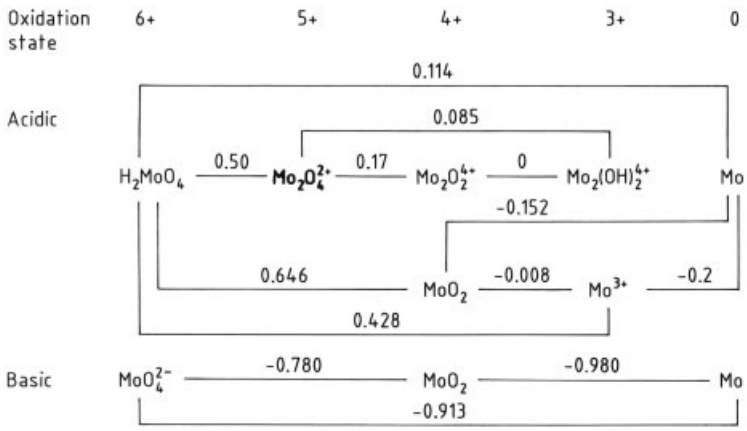

Figure 13. Reduction potentials (in volts) of molybdenum couples in acidic and alkaline aqueous solutions [60]
The more negative the potential, the more reducing is the couple, i.e., the more stable is the higher oxidation state.

is a volumetric method of determining molybdenum [62].

7.2. Molybdenum Oxides

Well characterized oxides are molybdenum trioxide, MoO_3, and molybdenum dioxide, MoO_2. Mixed-valency oxides and molybdenum blues are also important.

Molybdenum Trioxide. Molybdenum trioxide [*1313-27-5*], molybdenum(VI) oxide, MoO_3, M_r 143.95, is the ultimate product of the oxidation of molybdenum compounds. It forms white orthorhombic crystals that are photosensitive, i.e., turn blue in light. The compound is weakly paramagnetic and an *n*-type semiconductor. Other properties are listed in Table 6. It is produced by roasting molybdenum disulfide in air at 600 °C and may be purified by sublimation (see Section 4.2). It can also be prepared by reacting ammonium molybdate at 550 °C with oxygen or by precipitation from an aqueous ammonium molybdate solution treated with concentrated nitric acid [65].

The structure of molybdenum trioxide comprises layers built from linked, distorted MoO_6 octahedra. Crystal morphology depends on the method of preparation: needles occur after vapor condensation, platelets after crystal growth on silica or graphite [66]. Molybdenum trioxide has a characteristic IR spectrum showing Mo – O vibrations (996, 870, 821, and 570 cm^{-1}).

Table 6. Properties of molybdenum trioxide and dioxide

Property	MoO_3	MoO_2
Crystal structure	orthorhombic	monoclinic
Space group	*Pbnm*	$P2_1c$
Z	4	4
a, nm	0.39628	0.56109
b, nm	1.3855	0.48562
c, nm	0.36964	0.56285
α, β, γ	90°, 90°, 90°	90°, 120.95°, 90°
Mo coordination	octahedral, distorted	octahedral, distorted rutile
ΔH_f°, kJ/mol∗	− 745.2	− 588.7
ΔG_f°, kJ/mol∗	− 667.9	− 533.1
S° J/K∗	77.7	47.2
Density, g/cm^3 (21 °C)	4.692	6.29
mp, °C	801	
bp, °C	1155	
Sublimation, °C	ca. 700	ca. 1100
Solubility, g/L		
Water 28 °C	0.49	insoluble
100 °C	1.5	insoluble
HNO_3 (4 mol/L, 20 °C)	ca. 14	insoluble
Aqueous ammonia and alkali	soluble	insoluble
Organic solvents	insoluble	insoluble

∗ Standard enthalpy and free energy of formation from the elements and entropy at 298.15 K [63], [64].

Molybdenum trioxide is an acidic oxide soluble in aqueous alkali and ammonia. Freshly precipitated MoO_3, which is much more reactive than the aged oxide, dissolves in aqueous solutions of citric, malic, and other hydroxycarboxylic acids [67], and in solutions of

ethylene glycol, glycerol, and other polyols to form molybdenum(VI) complexes [68]. Molybdenum trioxide also reacts with β-diketones (e.g., acetylacetone) to form complexes such as $MoO_2(CH_3COCHCOCH_3)_2$ [*17524-05-9*] which is a useful starting compound for preparing other molybdenum(VI) complexes by ligand exchange reactions. It is reduced by hydrogen at 500 – 600 °C to MoO_2 and at 900 – 1000 °C to molybdenum metal.

Molybdenum trioxide in combination with more basic metal oxides finds application in heterogeneous catalysis (see Section 7.5). Studies of MoO_3 specimens with different crystal habits have demonstrated structure sensitivity of MoO_3 selective oxidation catalysts; the type of reaction catalyzed depends on the predominant crystal face. Platelike crystallites predominantly expose the (010) face which is responsible for insertion of oxygen into allylic compounds; activation of, for example, propene occurs at the (100) and (101) faces, and total oxidation at all faces [66].

Hydrates of MoO_3. The monohydrate, $MoO_3 \cdot H_2O$ [*39082-25-2*], precipitates when a concentrated solution of sodium molybdate is acidified with hydrochloric acid. The dihydrate, $MoO_3 \cdot 2\,H_2O$ [*25942-34-1*], precipitates when a solution of ammonium heptamolybdate is acidified with nitric acid [65]; it is commonly referred to as molybdic acid ($H_2MoO_4 \cdot H_2O$). Both hydrates are yellow.

Molybdenum Dioxide. Molybdenum dioxide [*18868-43-4*], molybdenum(IV) oxide, MoO_2, M_r 127.95, is a dark blue crystalline solid and consists of chains of distorted MoO_6 octahedra. The distortions arise from Mo – Mo bonding leading to alternate long (0.31118 nm) and short (0.25106 nm) Mo – Mo distances [69], [70]. Properties are listed in Table 6. Molybdenum dioxide is a metallic conductor and is weakly paramagnetic. It is produced by reducing MoO_3 with hydrogen, ammonia, and carbon monoxide at 600 °C.

Molybdenum dioxide is a neutral oxide that is insoluble in acid and alkali. It reacts with chlorine giving dichlorodioxomolybdenum(IV) [*13637-68-8*], MoO_2Cl_2, and with carbon tetrachloride giving molybdenum tetrachloride [*13320-71-3*], $MoCl_4$.

Molybdenum Blues. Molybdenum blues [*66771-43-5*], so called because of their blue color, are mixed oxide – hydroxides of molybdenum(VI) and molybdenum(V) (see Fig. 12) [71]. They are amorphous solids that are soluble in water and alcohols. A typical species is $(Mo_3^{6+}Mo_3^{5+}O_{18}H)^-$ which has an absorption maximum at 450 nm.

The formation of molybdenum blues is the basis of a colorimetric method for determining phosphorus [72]. Molybdate and orthophosphate ions condense in acidic solution to give molybdophosphoric acid which is reduced by hydrazinium sulfate to a molybdenum blue with maximum light absorbance at 820 – 830 nm. The color intensity is proportional to the amount of phosphate incorporated. The reduced molybdophosphate is an example of a heteropoly blue [71].

Mixed Valency Oxides. The mixed Mo(VI) – Mo(V) oxides (molybdenum oxide bronzes) have an intense color and metallic luster [73].

The binary bronzes MoO_x (x between 2 and 3), constitute a series of deep blue or purple compounds. They are prepared by heating MoO_3 in vacuo or with molybdenum metal. The structures are built from edge- and corner-sharing $[MoO_6]$ octahedra, or $[MoO_6]$ octahedra and $[MoO_4]$ tetrahedra. The Magneli phases, Mo_4O_{11} [*12033-38-4*], and Mo_8O_{23} [*12058-34-3*] have metallic properties at room temperature.

The ternary oxides include the red bronzes, $A_{0.33}MoO_3$ (A = Li, K, Rb, Cs, Tl), which are semiconductors, the blue bronzes $A_{0.3}\,MoO_3$ (A = K, Rb, Tl), purple bronzes $A_{0.9}Mo_8O_{17}$ (A = Li, Na, K, Tl), and a rare-earth bronze $La_2Mo_2O_7$; they behave like metals at room temperature. Interest in these compounds is focused on their electrical and optical properties.

7.3. Molybdenum Chalcogenides

The most important chalcogenide of molybdenum is molybdenum disulfide, which is the principal source of molybdenum, in the form of the mineral molybdenite. Molybdenum selenide [*12058-18-3*], $MoSe_2$, and telluride [*12058-20-7*], $MoTe_2$, are also known.

Molybdenum Disulfide. Molybdenum disulfide [*1317-33-5*], molybdenum(IV) sulfide, MoS_2, M_r 160.07, is a black crystalline solid that occurs in two forms: hexagonal, which is found in the mineral molybdenite [*1309-56-4*] and rhombohedral (Table 7). The solid is diamagnetic and a semiconductor. It is built of S−Mo−S layers whose sulfur atoms are in contact (Fig. 14). Within each layer molybdenum is at the centre of a trigonal prism of sulfur atoms. Since the forces operating between the layers are weak van der Waals attractions, the layers can slide over one another giving molybdenum disulfide a slippery texture and solid-state lubrication properties like graphite.

Table 7. Properties of molybdenum disulfide

Property	Hexagonal crystal structure	Rhombohedral crystal structure
Space group	$P6_3/mmc$, D_{6h}^4	$R3m$, C_{3v}^5
Z	2	3
a, b, c, nm	0.316, 0.316, 1.229	0.317, 0.317, 1.838
α, β, γ	90°, 90°, 120°	90°, 90°, 90°
Density, g/cm^3	5.05	5.02
Mo coordination	trigonal prismatic	
Interatomic distances, nm		
Mo – S	0.242	
Mo – Mo	0.316	
S – S	0.349	
ΔH_f°, kJ/mol∗	− 234.1	
ΔG_f°, kJ/mol∗	− 225.7	
S°, J/K∗	62.8	
mp, °C	> 1600	

∗ Standard enthalpy and free energy of formation from the elements and entropy at 298.15 K [63], [64], [74].

The hexagonal form is produced by heating molybdenum with sulfur, ammonium tetrathiomolybdate [$(NH_4)_2MoS_4$], or molybdenum trisulfide at 1100 °C. The compound is amorphous at 250 – 400 °C; crystallization requires prolonged heating at 1100 °C. The rhombic form is obtained by heating the hexagonal form in vacuo at 7.5 GPa and 1200 °C, or by reacting MoO_3 with sulfur and potassium carbonate at 1100 °C.

Molybdenum disulfide dissolves only in strongly oxidizing acids (e.g., aqua regia) to give molybdenum(VI) compounds. When heated in vacuo at 1200 °C it decomposes to Mo_2S_3 and metallic molybdenum. Heating in air at 500 °C produces MoO_3. Reduction with hydrogen at 1100 °C or > 1500 °C yields Mo_2S_3 or molybdenum metal, respectively.

Molybdenum Sesquisulfide. Molybdenum sesquisulfide [*12033-33-9*], dimolybdenum trisulfide, molybdenum(III) sulfide, Mo_2S_3, M_r 288.08, has a different structure from MoS_2; the molybdenum is at the center of a distorted octahedron of sulfur atoms. Molybdenum sesquisulfide is a grey solid and is prepared by heating molybdenum and sulfur (2 : 3) in vacuo at 1400 °C. It dissolves in, and is oxidized to molybdenum(VI), by aqua regia. In air at 350 °C it oxidizes to molybdenum trioxide. Hydrogen at 1500 °C reduces it to molybdenum metal.

Molybdenum Trisulfide. Molybdenum trisulfide [*12033-29-3*], molybdenum(VI) sulfide, MoS_3, M_r 192.14, is the sulfide precipitated in conventional inorganic analysis when hydrogen sulfide is passed through an acidified molybdate solution. Molybdenum trisulfide, like the sulfides of tin, arsenic and antimony, dissolves in yellow ammonium sulfide giving a tetrathio salt, MoS_4^{2-}.

According to X-ray absorption spectroscopy, the composition MoS_3 corresponds to $Mo(S^{2-})_2(S_2^{2-})_{0.5}$, one third of the sulfur is disulfide (S_2^{2-}); the formal oxidation state of the molybdenum is 5^+ [75].

When heated in nitrogen at ca. 350 °C molybdenum trisulfide decomposes to amorphous molybdenum disulfide. Heated in air, it begins to oxidize to molybdenum trioxide at 200 °C.

Chevrel Phases. The Chevrel phases are ternary molybdenum chalcogenides, $M_xMo_6X_8$ (where M is a metal and X a chalcogenide), for example, $PbMo_6S_8$ [*39432-49-0*] [76]. They are black powders prepared by heating the constituent elements in vacuo at 1000 °C. They consist of linked [Mo_6S_8] clusters having sulfur atoms at the corners of a cube and molybdenum atoms at the face centers (the same structure as the [Mo_6Cl_8] structure, see Section 7.4). These compounds have attracted great interest as superconductors with critical temperatures as high as 15 K and as battery materials.

Ammonium Tetrathiomolybdate. Ammonium tetrathiomolybdate [*15060-55-6*], $(NH_4)_2[MoS_4]$, M_r 260.28, crystallizes from

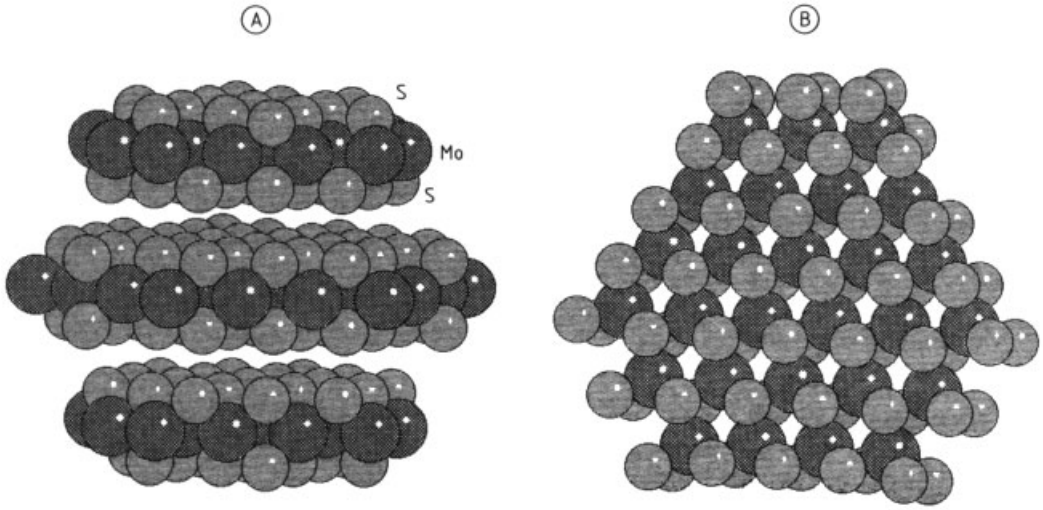

Figure 14. Computer-drawn structure of molybdenum disulfide
A) Side view showing the S – Mo – S layers; B) Top view of a hexagonal fragment of one layer

a solution of ammonium molybdate saturated with hydrogen sulfide to form deep red crystals. The $[MoS_4]^{2-}$ ion is tetrahedral and is the thio analogue of molybdate. Ammonium tetrathiomolybdate has an extensive chemistry and is an entry point to much molybdenum – sulfur chemistry. The sulfur atoms form bridges to metal ions so forming clusters, e.g., the species $[S_2MoS_2CoS_2MoS_2]^{2-}$ with Co^{2+} [52].

7.4. Molybdenum Halides

Binary halides are known for molybdenum in oxidation states (II) to (VI). The preparation of molybdenum fluorides and chlorides is summarized in Figure 15. The only molecular halide is molybdenum hexafluoride, MoF_6, which is a volatile liquid. The coordination number of molybdenum is six in the halides of Mo(V), (IV), and (III), the molybdenum atoms being linked by bridging halogens.

Molybdenum(II) chloride is an extended solid containing $[Mo_6Cl_8]^{4+}$ clusters linked by bridging chlorines so giving the formula Mo_6Cl_{12}. In the cluster eight chlorine atoms are at the corners of a cube. Molybdenum atoms, linked by Mo − Mo bonds, are at the face centers of the cube and so constitute an octahedral Mo_6 cluster. The cluster is very stable both to hydrolysis and oxidation and persists through many reactions.

A feature of the halides of Mo(IV), (III), and (II) is the formation of bonds between molybdenum atoms. Consequently, there is pairing of electron spins on adjacent molybdenum atoms and so the magnetic moments of the halides are less than the spin-only values.

Molybdenum Pentachloride. Molybdenum pentachloride [*10241-05-1*], molybdenum(V) chloride, $MoCl_5$, M_r 273.21, is the most important molybdenum halide. The dark green – black crystals (*mp* 194 °C) consist of Mo_2Cl_{10} dimers; the vapor and solutions contain $MoCl_5$ monomers.

Molybdenum pentachloride is soluble in carbon tetrachloride, benzene, and many other organic solvents. It is very reactive, undergoing hydrolysis in air, water, and damp organic solvents to form brown oxomolybdenum(V) chlorides. It is supplied and stored in sealed ampoules and must be handled under argon or nitrogen or in vacuo.

The outstanding property of molybdenum pentachloride is its ability to abstract an oxygen atom from water, sulfur dioxide, triphenylphosphine oxide (Ph_3PO), dimethyl sulfoxide (Me_2SO), and other oxygen-containing compounds to form oxomolybdenum complexes,

$MoOCl_3A_2$, (e.g., A = Ph_3PO, Me_2SO). In concentrated hydrochloric acid molybdenum pentachloride forms the deep green complex ion $[MoOCl_5]^{2-}$.

Molybdenum pentachloride is the starting point for much nonaqueous and organomolybdenum chemistry.

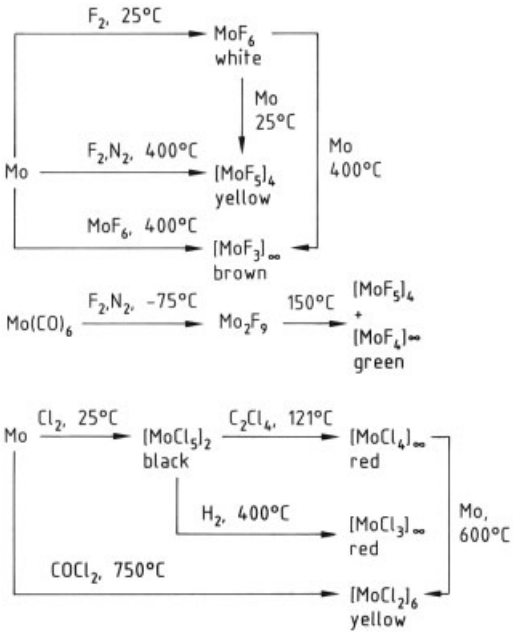

Figure 15. Preparation of molybdenum fluorides and chlorides

Oxomolybdenum Halides. In oxidation states 5+ and 6+ molybdenum has a strong tendency to form multiple bonds to oxygen. The oxomolybdenum halides are volatile solids which may be sublimed in vacuo. They are more stable than the binary halides and are formed when the binary halides are exposed to air and water.

Oxomolybdenum(VI) compounds are of two types: $MoOX_4$ (X = F, Cl), and MoO_2X_2 (X = F, Cl, Br). Oxomolybdenum(V) compounds are generally monoxo, e.g., $MoOCl_3$.

Trichlorooxomolybdenum(V). Trichlorooxomolybdenum(V) [*13814-74-9*], $MoOCl_3$, M_r 218.30, is a green solid consisting of octahedral $[MoOCl_5]$ units linked through cis-chlorine atoms. It forms adducts with many organic ligands, e.g., $MoOCl_3 \cdot L$ (L = ligand) with acetylacetone and 2,2′-bipyridyl, and $MoOCl_3 \cdot 2\,L$ with diethyl ether and pyridine.

7.5. Molybdates, Isopolymolybdates, and Heteropolymolybdates

Sodium molybdate and other simple molybdates contain the tetrahedral $[MoO_4]^{2-}$ ion. Ammonium heptamolybdate (an isopolymolybdate), $(NH_4)_6Mo_7O_{24} \cdot 4\,H_2O$, is composed of linked, distorted $[MoO_6]$ octahedra. In the heteropolymolybdates, molybdenum and another element are present in the same structure. The properties of some of the more important molybdates are summarized in Table 8.

Ammonium and Alkali-Metal Molybdates. Ammonium and alkali-metal molybdates are soluble in water (Table 8). They are made by crystallizing or evaporating solutions of molybdenum trioxide dissolved in aqueous alkali. Ammonium dimolybdate is used as an intermediate for purifying technical-grade MoO_3 (see 4.2). For the production of sodium molybdate, see 4.2.

Ammonium molybdates are made by dissolving MoO_3 in aqueous ammonia solution; $(NH_4)_2MoO_4$ then crystallizes from aqueous ammonia, $(NH_4)_6Mo_7O_{24} \cdot 4\,H_2O$ from a solution kept at the boiling point for 1 h, and $(NH_4)_2Mo_8O_{26} \cdot 5\,H_2O$ from an acidified aqueous solution of $(NH_4)_6Mo_7O_{24} \cdot 4\,H_2O$. The dimolybdate, $(NH_4)_2Mo_2O_7$, may be crystallized from hot ammoniacal molybdate solutions.

The structures of the isopolymolybdates are based on linked, distorted $[MoO_6]$ octahedra. The structure of ammonium dimolybdate consists of infinite chains of pairs of edge-shared $[MoO_6]$ octahedra with adjacent pairs linked by $[MoO_4]$ tetrahedra [79].

Ammonium di- and heptamolybdates are used as sources of water-soluble molybdenum, for example, in the preparation of catalysts. When the catalyst is calcined, the molybdate decomposes to molybdenum trioxide and ammonia.

Molybdates of Di- and Trivalent Cations. Molybdates of divalent cations (except Mg^{2+}) are insoluble in water (Table 8). They are prepared by precipitation from a sodium or ammo-

Table 8. Properties of some molybdates [53], [77]

Molybdate	M_r	Structure [a]	Color	Solubility in water, wt % [b]	ΔH_f°, kJ/mol [c]
$(NH_4)_2MoO_4$	196.01	4	white	39	
$(NH_4)_2Mo_2O_7$	339.95	4 or 6		30	
$(NH_4)_6Mo_7O_{24}$ $\cdot 4H_2O$	1235.8	6	white	30	
Na_2MoO_4	205.92	6+4	white	40	−1466
$MgMoO_4$	184.24	6+6	white	16	−1248
$CaMoO_4$	200.01	8+4	white	0.005	−1548
$SrMoO_4$	247.56	8+4	white	0.003	−1449
$BaMoO_4$	297.26	8+4	white	0.005	−562
$CoMo_4$ [d]	218.87				−1032
a		6+4	violet		
b		6+6	rose		
$NiMoO_4$ [d]	218.63				−947.7
a		6+4	green		
b		6+6	green		
$PbMoO_4$	367.14	8+4	white	10^{-5}	−1076
$ZnMoO_4$	225.33	5 or 6+4	white	0.5	−1140
$Fe_2(MoO_4)_3$	591.50	6+4	yellow		−2963

[a] Coordination of the countercation (except ammonium) by oxide and coordination of molybdenum (4 = tetrahedral; 6 = octahedral).
[b] Weight percent of anhydrous salt in 100 g saturated solution at 25 °C.
[c] Standard heat of formation from the elements at 298.15 K.
[d] Two polymorphs: low-temperature form, b, transforms to the high-temperature form, a, at 400 °C [78].

nium molybdate solution or by heating mixed solid oxides. The coordination environment of the molybdenum may be tetrahedral or octahedral.

Molybdates of trivalent cations have the formula $A_2(MoO_4)_3$ or $A_2Mo_3O_{12}$, with A = Al, Cr, Fe, Bi, lanthanide. Coordination of the molybdenum is tetrahedral; the other cation is usually octahedral.

A number of these molybdates have important uses. Cobalt, iron, and bismuth molybdates are catalysts for the selective oxidation of organic compounds (see Section 8.1): olefins are oxidized to unsaturated aldehydes and ketones, e.g., propene to acrolein (cobalt and bismuth), propene and ammonia to acrylonitrile (bismuth), methanol to formaldehyde (iron). Molybdenum is the source of selectivity in these reactions. The first step is generally activation of a C – H bond over the more basic (nonmolybdenum) oxide and the molybdenum then controls oxygen insertion. The well known Co–Mo/alumina hydrodesulfurization catalyst does not, however, strictly contain cobalt molybdate.

Other actual or proposed applications are: calcium and strontium molybdates doped with neodymium as lasers; lanthanide molybdates for optical and optoelectric applications; zinc molybdate as a white pigment and corrosion inhibitor.

Heteropolymolybdates. The heteropolymolybdates consist of $[MoO_6]$ octahedra incorporating atoms of a different element, the heteroatom [52], [80]. The heteroatoms are completely surrounded by the oxygen atoms of the $[MoO_6]$ octahedra. The resulting coordination of the heteroatom may be tetra- or octahedral. The 12-molybdo species, $[X^{+n}Mo_{12}O_{40}]^{(8-n)-}$, is an important group with tetrahedrally coordinated heteroatoms (X). An example is 12-molybdophosphoric acid [*12026-57-2*], $H_3[PMo_{12}O_{40}] \cdot 28\,H_2O$, prepared by dissolving molybdenum trioxide in phosphoric acid; it is yellow and readily soluble in water. Ammonium 12-molybdophosphate, which precipitates when ammonium molybdate is added to a solution of disodium hydrogen phosphate, is used for the gravimetric determination of phosphate [76]. The phosphomolybdates are selective oxidation catalysts, e.g., in the conversion of acrolein to acrylic acid [81].

7.6. Other Molybdenum Compounds

Molybdenum forms a large number of coordination compounds or *complexes* [52]. Complexes with organosulfur ligands attract a great deal of interest because of their significance in molybdenum biochemistry (see Section 7.1) and applications as oil-soluble antiwear and friction-reducing additives [82]. Examples of the latter are phosphorodithioates (dithiophosphates) and dithiocarbamates, $[Mo_2O_3L_4]$ and $[Mo_2O_2S_2L_2]$ where L is $(RO)_2PS_2^-$, $R_2NCS_2^-$.

Molybdenum hexacarbonyl [*13939-06-5*], $Mo(CO)_6$, M_r 264.002, *mp* 150 °C, *bp* 156.4 °C (sublimes), is a white, air-stable solid that is insoluble in water but soluble in organic solvents. It is made from molybdenum pentachloride (see Fig. 11) and is commercially available. It is a starting point for the synthesis of low-valent and organomolybdenum compounds [83]. The structure consists of CO groups octahedrally disposed around molybdenum and bound through carbon.

8. Uses of Molybdenum Compounds

8.1. Catalysis

Molybdenum-containing catalysts are used for a broad range of reactions [90] (see also → Heterogeneous Catalysis and Solid Catalysts). In some cases, the catalyst is a single molybdenum compound, and either soluble (homogeneous) or insoluble (heterogeneous) in the reaction medium. In other cases, the catalyst is a binary mixture of the oxides and/or sulfides of molybdenum and of another element (usually Co, Ni, Fe, W, or Bi). The catalytic properties of these binary compositions differ greatly from those of the individual compounds. The heterogeneous catalysts are often formed and supported on a substrate such as activated alumina, silica, carbon, or zeolite. The support can greatly improve catalyst activity, selectivity, and mechanical strength. In industrial practice other elements or compounds (e.g., potassium, phosphorus, tungsten, or silica) are often added to the binary compositions as additives to further enhance activity, selectivity, and strength.

The scope and versatility of molybdenum catalysts are illustrated in Table 9 [91]. The activity and selectivity of catalysts based on MoO_3 and MoS_2 are summarized for redox reactions, isomerization, polymerization, addition, and decomposition (usually classified as acid – base reactions).

One of the largest uses for molybdenum catalysts is in the desulfurization of petroleum, petrochemicals, and coal-derived liquids, in which organosulfur compounds react with hydrogen at the catalyst surface and sulfide ions are removed as hydrogen sulfide. Desulfurization is used to improve product color, smell, and stability, to eliminate sulfur dioxide emission on fuel combustion, and to make possible subsequent reforming processes which may use sulfur-sensitive catalysts. Today, environmental air quality is a major reason for desulfurization.

The most common desulfurization catalyst is a binary mixture of cobalt oxide and an excess of molybdenum trioxide on an activated alumina substrate. After being charged into the industrial reactor, the catalyst reacts with the sulfur-containing reducing atmosphere to form molybdenum and cobalt sulfides as the working catalyst.

A major development in the history of hydrocarbon oxidation was the discovery of bismuth molybdate catalysts for the partial oxidation and ammoxidation of propene to acrylonitrile in one step [92].

$$CH_2{=}CHCH_3 + NH_3 + 1\,1/2\,O_2 \rightarrow CH_2CHCN + 3\,H_2O$$

The original catalyst was reported to have a composition of 50 % $Bi_9PMo_{12}O_{52}$ and 50 % SiO_2, with the silica gel support preventing catalyst disintegration in the fluidized bed operation. This catalyst and its modifications account for virtually all of the acrylonitrile production in the world.

Another major catalyst based on ferric molybdate ($5\,MoO_3 \cdot Fe_2O_3$) is used to selectively oxidize methanol to formaldehyde (→ Formaldehyde, Chap. 4.2.) [93]:

$$CH_3OH + 1/2\,O_2 \rightarrow CH_2O + H_2O$$

Molybdenum catalysts are also important in commercial coal liquefaction processes where coal is converted into upgraded liquid fuels. Coal liquefaction processes commonly utilize either alumina-supported cobalt molybdate or molybdenum trioxide as catalyst to effect desulfurization, denitrogenation, deoxygenation, cracking, and hydrogenation.

Table 9. Reactions catalyzed by molybdenum oxides and sulfides [91]

Type of reaction	Example	Catalyst	Temperature range, °C	Conversion, %	Selectivity, %
Reactions with molecular hydrogen					
Isotopic exchange	$H_2 + D_2 \rightarrow 2\,HD$	$MoO_3/Al_2O_3-SiO_2$	80		very high
Hydrogenation	benzene → cyclohexane	MoO_3	420–450	60	
	acetylene → ethylene	MoO_3-CoCl_2/SiO_2	260–300		100
	naphthalene → decalene	MoS_2	340		yield 77
	sulfur → hydrogen sulfide	MoS_2	200	100	
	selenium → hydrogen selenide	MoS_2	330–350	100	
	maleic anhydride → succinic anhydride	$CuMoO_4/SiO_2$	120–160	100	
Hydrogenolysis	α- and β-methyl naphthalene → naphthalene + methane	$CoO-MoO_3/Al_2O_3$			yield 40
	ethylcyclohexane → products	MoO_3/Al_2O_3	510		
Reduction	phenol → benzene + H_2O	MoS_2	350		yield 88–98
	thiophenol → benzene + H_2S	MoS_2	200	100	
	thiophene → $C_4H_{10} + H_2S$ C_4H_8 C_4H_6	MoS_2	300	98	
	organic sulfur compounds → hydrocarbons	$CoO-MoO_3/Al_2O_3$	225–425		
Reactions with molecular oxygen					
Selective oxidation	propene → acrolein	$Bi_2O_3-MoO_3$	450	95	95
	butene → maleic anhydride	$CoO-MoO_3-P_2O_5/$			
SiO_2	450	76	50		
	benzene → maleic anhydride	$V_2O_5-MoO_3-P_2O_5/$			
Al_2O_3	425–450		85		
	o-xylene → phthalic anhydride	$V_2O_5-MoO_3$	400		48
Oxidation of functional groups	methanol → formaldehyde	$Fe_2(MoO_4)_3$	300–400	92	92
	acrolein → acrylic acid	$CoO-MoO_3-TeO_2$	383	85	65
Oxidative condensation	propene + NH_3 → acrylonitrile	$Bi_2O_3-MoO_3$	400–500		yield 80
	toluene + NH_3 → benzonitrile	$V_2O_5-MoO_3/Al_2O_3$	415		yield 94
Oxidative dehydrogenation	butene → butadiene	$Bi_2O_3-MoO_3$	370–550		90
	ethylbenzene → styrene	MoO_3-MgO	420–450		82–84
Oxyhydration	propene → acetone	SnO_2-MoO_3	135		85
Oxidative dehydrocondensation	*n*-butane + SO_2 → thiophene	MoO_3/Al_2O_3	595	70	44
	butene + H_2S → thiophene	$CoO-MoO_3/Al_2O_3$	570		yield 37
Oxidation with other oxidants					
Epoxidation	propene + H_2O_2 → propylene oxide	MoO_3 in H_2O_2 and pyridine	60	80	100
Isomerization					
Structural isomerization	*n*-pentane → isopentanes	MoO_3/Al_2O_3	425–450	50	95
Ring contraction	cyclohexane → methylcyclopentane	MoO_3/Al_2O_3	455–495		yield 37
Disproportionation					
Olefin metathesis	propene → butene + ethylene	MoO_3/Al_2O_3	66–288	43	94
Polymerization					
Polymerization	ethylene → polyethylene	MoO_3/Al_2O_3	200–260	50	
	acetylene → benzene	$CoO-MoO_3/Al_2O_3$	62	98	100
Etherification					
Formation of complex ethers	succinic acid + hexanol → ether	MoS_2/carbon	130–220	45–97	100
Addition					
Addition to C=C	allylic alcohol + H_2O_2 → glycerol	MoO_3 in acetic acid	70–100		
Decomposition					
Dehydration	2-propanol → propene	MoO_3	191–224		100
	ammonium benzoate → benzonitrile	MoO_3	225–245		
Dehydrogenation	ethylbenzene → styrene	MoS_2	414		yield 94
	cyclohexane → benzene	MoO_3/Al_2O_3	500		yield 63
Dehydrocyclization	*n*-heptane → toluene	MoO_3/Al_2O_3	450–500		yield 75
Dehydrocondensation	isobutene → xylenes	MoO_3/Al_2O_3	200		yield 90

Phosphomolybdate compounds are acidic oxidants and are used in both homogeneously and heterogeneously catalyzed processes such as the hydration of propene in liquid phase and the oxidation of isobutyric acid to methacrylic acid.

Molybdenum hexacarbonyl is a versatile homogeneous and heterogeneous catalyst. It is widely used in the production of methanol from carbon monoxide, to effect olefin metathesis, desulfurization, aqueous coal liquefaction, and epoxidation.

8.2. Lubrication

Molybdenum disulfide is an important solid lubricant, being used primarily to reduce wear, friction, and sustain lubrication under boundary sliding conditions (→ Lubricants and Related Products). It is a black powder that is insoluble in ordinary aqueous and organic solvents but dissolves by complex formation in potassium cyanide, and reacts with strong oxidizing agents such as aqua regia or hot concentrated sulfuric acid to give hexavalent molybdenum species. It is an excellent high-temperature lubricant which is stable in vacuo or an inert atmosphere to 1200 °C, and to about 350 °C in air.

The primary commercial source for MoS_2 is the mineral molybdenite, from which > 98 % pure powder is recovered and purified using oil flotation and other techniques (Chap. 4) with carbon being the major impurity. The lubricating powder is available in several grades based on its median particle size which may range from ca. 1 to 40 µm.

The crystal structure of natural MoS_2 is hexagonal (Fig. 14). Its excellent intrinsic lubricating properties are attributed to the wide spacing and weak van der Waals bonding between the S – Mo – S sandwich layers, and to the net positive charge on the surface which promotes easy shear by means of electrostatic repulsion [94]. This intrinsic lubricating quality contrasts with graphite, which requires the presence of condensable vapors (e.g., from water or hydrocarbons) to promote crystal cleavage and low friction [95].

The fundamental and practical aspects of MoS_2 lubrication are reviewed in [87], [96], [97]. Molybdenum disulfide is sometimes simply rubbed on surfaces as a thin film to provide short-term lubrication. As an additive, it enhances the performance of other lubricants like grease or oil. It can also be incorporated into paint-like coatings for longer service life. Finally, it can be incorporated into plastics, rubber, and powder metal matrices as an alternative to conventional lubricants to impart lifetime self-lubrication.

Summaries have been published describing the use of MoS_2 in greases [98], rubber [99], bonded coatings [100], and motor oils [101].

The largest consumption of MoS_2 is in *greases* used in mining, manufacturing, and transportation [98]. In greases, 1 – 20 wt % molybdenum disulfide gives added protection from wear and galling, and provides back-up lubrication should the grease deplete or thermally degrade. Pastes composed of grease and containing 20 –60 wt % MoS_2 are used with open gears, universal joints, spline drives, and in metal forming, press-fitting and various wear-in operations. They have outstanding boundary lubrication behavior which prevents fretting, galling, and seizing.

As an extra-fine powder, MoS_2 (0.5 – 1 wt %) is suspended in *industrial oils, motor oils, and synthetic fluids*, with the aid of chemical dispersants such as polyethylene where it behaves as an extreme-pressure and antiwear additive. In motor oils, it also improves vehicular fuel economy.

Paint-like coatings containing 30 – 85 vol% MoS_2 in either an organic or inorganic silicate binder are used on fasteners, machine tools, forklift chains, door locks, chutes, and sliding parts of office equipment to reduce friction and wear. Temporary coatings of MoS_2 in water-soluble binders (e.g., poly(vinyl alcohol) and polyacrylates) are used in die casting as a mold release, and in metal forming operations such as ironing, wire drawing, and extrusion.

In *plastic, rubber, and powdered metal compositions,* 0.5 – 4 wt % MoS_2 imparts self-lubricating behavior obviating the requirement for externally applied lubrication. These materials are used in articles such as cams, thrust washers, ball bearing retainers, compressor piston rings, sheaves, gears, conveyor belting, and dynamic seals.

Molybdenum disulfide is not the only molybdenum compound used in lubrication. *Oil-soluble molybdenum–sulfur compounds* have also long been used as antiwear and extreme-pressure additives in lubricating oils and greases [102]. Commercial oil-soluble additives are compositions based on tetrathiomolybdates and complexes of molybdenum with xanthates, dithiocarbamates, dithiophosphates, diketones, dithiols, and dimercaptothiadiazoles [102].

8.3. Corrosion Inhibition (→ Corrosion)

In 1939, two patents first described the use of readily soluble sodium, potassium, and ammonium molybdates as corrosion inhibitors for motor vehicle engine coolants [103], [104]. These and other inorganic molybdates are now among the most popular corrosion inhibitors because of their favorable properties and behavior. Molybdate is an anodic inhibitor, i.e., it inhibits by increasing the polarization of the anode component of the metal corrosion cell. It does this by precipitating escaping metal cations as molybdate species to block anodic sites and strengthen developing metal oxide films. Molybdate inhibits the corrosion of more ferrous and nonferrous metals over a wider pH range than any other inhibitor apart from chromate. Unlike chromate, however, the simple molybdate anion is not an oxidant and thus can be used with chemicals and systems that do not tolerate the strongly oxidizing chromate. Of even greater importance, molybdate has a very low toxicity. The toxicity, environmental aspects, behavior, mechanistic studies, and many applications of the molybdates are reviewed in [105], [106].

Molybdates are seldom used alone – as with other anodically active inhibitors, molybdate efficacy is improved and its concentration requirement significantly reduced when it is combined with other inhibitors. The most widely employed synergists of molybdate inhibition are the cathodically active compounds, especially zinc salts (e.g., zinc sulfate). Other synergists are the oxidizing inhibitors (e.g., nitrite) and the filming inhibitors (e.g., amine phosphonates and hydroxyfatty acids) [106]. Filming inhibitors strongly adsorb to the metal surface and protect it from attack.

Sodium molybdate is a component of many of the factory-fill and after-market engine coolants used in the United States, Europe, and Japan. A modern engine coolant concentrate formulation containing molybdate and other inhibitors is as follows [107] (concentrations in weight percent):

Ethylene glycol	95.56
Sodium nitrate	0.10
Sodium molybdate dihydrate	0.20
Sodium tetraborate pentahydrate	0.40
Sodium silicate solution N*	0.30
Phosphoric acid 85 %	0.15
Sodium mercaptobenzothiazole, 50 % solution	0.50
Sodium tolyltriazole, 50 % solution	0.20
Sodium hydroxide	0.235
Water	2.30
Polyglycol	0.05
Dye	0.005

∗ Registered trademark of PQ Corp., Valley Forge, Pennsylvania.

The largest consumption of molybdates for corrosion inhibition is in the treatment of cooling water in open and closed recirculating cooling systems. Open systems include cooling tower waters associated with power generation, manufacturing operations, and with metals production and chemicals processing. A formulation of a corrosion inhibition treatment in an open cooling tower system (milligrams per liter) follows:

1-hydroxyethylidene-1,1-diphosphonic acid, 60 % solution	10
Sodium molybdate dihydrate	10
Zinc sulfate monohydrate	2
Sodium tolyltriazole, 50 % solution	2

Closed loop waters are used in stationary power engines, in refrigeration and humidity control, and in chilled-water air conditioning; a formulation of a corrosion inhibition treatment for a closed chilled-water air conditioning system (milligrams per liter) follows:

Sodium molybdate dihydrate	150
Sodium nitrite	150
Sodium tolyltriazole, 50 % solution	4

Not only the highly soluble sodium, potassium, lithium, and ammonium molybdate salts are used in corrosion inhibition; the very slightly soluble zinc, strontium, and calcium molybdates are effective corrosion-inhibiting pigments in

primer paints for steel and aluminum [108]. The pure compounds are no longer used; they have been replaced by equally and, in some cases, more effective pigments made by precipitating the pure compounds on less expensive, synergistic core materials such as zinc oxide, calcium carbonate, talc, or zinc phosphate [109–111].

Other products and processes which utilize molybdates as corrosion inhibitors include water-based hydraulic and metalworking fluids, lithium halide refrigerating brines, boiler waters, hot forging lubricants, aluminum anodizing processes, coal – water slurries, temporary rust-preventive coatings, rinses for phosphate conversion coating, oil field drilling muds, brake linings, water used for quenching after heat treatment, paint-spray water curtains, wet ball milling of ores, salt brines for ice-making, passivation treatments for galvanized zinc and tin plate, packaging paper and board used for metal parts, temporary coating of aluminum to prevent staining by water during storage and transport, and as a pitting inhibitor for stainless steels in mineral acids. These and other applications are reviewed in [106].

8.4. Flame Retardancy and Smoke Suppression

Treatments based on molybdenum compounds exhibit both flame retardancy and smoke suppression; they are particularly effective smoke suppressants for synthetic polymers.

In the early 1930s, heteropolymolybdates were found to act as flame retardants for wood [112]. Some 15 years later, ammonium heptamolybdate and sodium molybdate were shown to retard the flammability of cotton fabric, although the treatment was not durable to laundering [113]. Other molybdenum compounds were subsequently used as additives to suppress the flammability and smoke output of plastics [nylon, poly(vinyl chloride), poly(vinylidene chloride), polyesters, and epoxy resins] [114].

The following molybdenum compounds are effective flame retardants and smoke suppressants for both rigid and plasticized poly(vinyl chloride) and halogenated polyester systems [115]:

- Molybdenum disulfide [*1317-33-5*], MoS_2
- Molybdenum(VI) oxide [*1313-27-5*], MoO_3
- Ammonium dimolybdate(VI)(2-) [*27546-07-2*], $(NH_4)_2Mo_2O_7$
- Ammonium heptamolybdate(VI)(6-)-4-water [*12027-67-7*], $(NH_4)_6Mo_7O_{24} \cdot 4\,H_2O$
- Ammonium octamolybdate(VI) (4^-) [*12411-64-2*], $(NH_4)_4Mo_8O_{26}$
- Sodium molybdate(VI) [*7631-95-0*], Na_2MoO_4
- Calcium molybdate(VI) [*7789-82-4*], $CaMoO_4$
- Zinc molybdate(VI) [*13767-32-3*], $ZnMoO_4$
- Copper molybdate(VI) [*13767-34-5*], $CuMoO_4$
- Nickel molybdate(VI) [*12673-58-4*], $NiMoO_4$
- Iron(III) molybdate(VI) [*13769-81-8*], $Fe_2(MoO_4)_3$

Flame retardancy in these compounds is linked to the accelerated loss of halogen; their smoke retardancy depends on increased char formation [116].

Molybdenum trioxide and extended proprietary zinc and calcium molybdate compositions are employed as flame retardants in the above-mentioned plastics that are used in the building, transportation, and wire and cable industries.

8.5. Pigments

Sodium molybdate [*7631-95-0*] is the usual starting material for the production of inorganic molybdate chrome pigments (→ Pigments, Inorganic) and a number of organic pigments (toners) based on cationic or basic dyes. Small amounts of sodium molybdate are also used as a condensation catalyst for phthalocyanine pigments as well as to fix and shade the coloring of furs and hair with azines and the amino- and hydroxy-substituted aromatic compounds used for red – brown shades.

Inorganic Pigments. Molybdate chrome pigments (molybdenum orange [*12656-85-8*]) are generally considered to be modified lead chrome pigments, i.e., compounds of lead chromate and mixed lead chromate and sulfate whose hues range from lemon yellow to primrose. Molybdate chrome pigments exhibit bright

red – orange to bright yellow – red hues with the redder hues containing more molybdenum. Molybdate orange pigments are mixed tetragonal crystals whereas the lead chrome pigments form monoclinic and rhombic crystals.

Molybdate chrome pigments are made by coprecipitating lead chromate, molybdate, and sulfate from lead nitrate solutions with sodium chromate, molybdate, and sulfate solutions. The stoichiometry and conditions of precipitation are adjusted to determine the hue. Lead molybdate is itself white, but in the coprecipitated product it distorts the monoclinic lead chromate crystals sufficiently to produce the tetragonal form.

The molybdate chrome pigments exhibit clean colors, good tinctorial and hiding properties; are nonbleeding and soft-grinding; have high heat resistance, and better film gloss and light fastness than lead chrome. They are often used with organic pigments to supply needed hiding power.

Molybdate chrome pigments are more expensive than the lead chromes, but are preferred in many applications for their reddish hue, cleanliness and brightness of color. They are employed in paints, plastics, rubber, and often in printing inks where their opacity, low cost, and heat stability are advantages over organic red pigments.

Organic Pigments. Sodium molybdate is also used in the production of organic pigments. Organic pigments have a higher tinctorial strength and brightness than inorganic pigments, but their thermo- and photochemical stabilities are lower and they are more transparent and expensive. Their biggest single use is in the coloring of plastics.

Molybdenum compounds are used in the production of toners derived from cationic or basic dyes that have free or substituted amino groups. The most important dyes are:

Diphenylmethane	Auramine	C.I. 655
Triarylmethane	Malachite Green	C.I. 657
	Brilliant Green	C.I. 662
	Rhoduline Blue 6G	C.I. 658
	Acronol Brilliant Blue	C.I. 664
	Methyl Violet B	C.I. 680
	Victoria Pure Blue BO	C.I. Pr198
Xanthene	Rhodamine B	C.I. 749
	Rhodamine 6G	

The dyes are made into pigments (i.e., insolubilized) by replacing the simple anions (chloride and sulfate) with a heteropolymolybdate acid salt, such as 12-molybdophosphate, $[PMo_{12}O_{40}]^{3-}$. This acid salt is formed from disodium phosphate and sodium molybdate in the presence of a strong mineral acid (e.g., HCl). The commercial phosphomolybdate pigments or toners are collectively known as PMA colors. The heteropolymolybdate complexes with the basic dye may contain 1 mol P_2O_5 with from 12 to 24 mol MoO_3 as either the free acids or acid sodium salts. Pigments made with combinations of both phosphotungstate (PTA) and phosphomolybdate are known as PTMA colors. The PTMA colors show better fastness than either the PMA or PTA colors, but are also generally costlier.

8.6. Agriculture

Molybdenum is an essential trace element for enzymes which fix nitrogen in leguminous crops. Sodium molybdate is used as a soil additive in areas where natural molybdenum is deficient and plant growth suffers. It is usually applied as part of fertilizer or seed treatments. About 0.25 kg molybdenum per acre (0.62 kg/ha) may be used.

9. Analysis

Molybdenum can be determined by gravimetric, volumetric, and colorimetric methods [117]. As long as great accuracy is not required, the most satisfactory method is, however, atomic absorption spectroscopy.

Atomic Absorption Spectroscopy. The absorption lines used and the achievable sensitivities in atomic absorption spectroscopy are shown in Table 10. Sensitivity may be increased by using a three-slot burner or by inert gas separation of the flame. Interference due to the presence of calcium, strontium, manganese, iron, and sulfate is reported to be eliminated in the presence of large amounts of aluminum [121] or 2 % ammonium chloride [118].

Table 10. Atomic absorption spectroscopic analysis for molybdenum

Absorption line, nm	Sensitivity, ppm/1 % abs**	Reference
313.26	0.8	[118], [119]
313.26*	0.4	[120]
317.03	1.1	[118]
379.83	1.3	[118]
319.40	1.4	[118]
386.41	1.7	[118]
390.30	2.4	[118]
315.82	2.8	[118]
320.88	5.9	[118]

* Nitrous oxide – acetylene flame, in all other cases air – acetylene flames were used.
** Defined as the concentration in parts per million (μg/mL) in aqueous solution which produces a 1 % absorption signal (0.0044 absorbance) under optimum conditions.

In the first report of the use of atomic absorption for molybdenum analysis [121], stainless steel samples were employed with an air – acetylene flame and ammonium chloride as a releasing agent. If the composition in the molybdenum standard solutions is matched with that of the sample solution by addition of appropriate quantities of the other metal salts present in the samples [122], a releasing agent is not necessary. An alternative is to separate the molybdenum from other metals in the sample solution by liquid – liquid (or solvent) extraction. This technique can also be used to concentrate the molybdenum solution before analyzing it. For example, quantities as low as 10^{-4} % have been determined in silicate rocks by extracting molybdenum from the aqueous sample with 8-hydroxyquinoline [123]; toluene-3,4-dithiol and ammonium pyrrolidine dithiocarbamate can also be used as extraction agents. Rock samples can be fused with sodium hydroxide and baked at ca. 650 °C for 30 min; after cooling the residue is treated with sulfuric acid and filtered. Alternatively, the sample can be treated directly with acid and filtered. Molybdenum in the filtrate has been extracted with α-benzoinoxime into chloroform [124].

Volumetric Methods. In the volumetric procedure, molybdenum in solution is reduced by zinc in a Jones reductor, the emerging solution is led under the surface of a ferric sulfate solution and titrated with potassium permanganate [117].

Gravimetric Methods. The older of the two main gravimetric methods is the precipitation of lead molybdate by slow addition of a lead acetate solution to a hot dilute acetic acid – ammonium acetate buffer solution containing the molybdenum. Free mineral acids prevent complete precipitation of the molybdenum. Cobalt, copper, magnesium, manganese, mercury, nickel and zinc do not interfere. Sulfate must be absent if alkaline-earth metals are present. If chloride is present, addition of a large excess of the lead acetate solution must be avoided. Since many metals interfere with the determination (antimony, arsenic, chromium, iron, phosphorus, silicon, tin, titanium, tungsten, and vanadium), it is not suitable for the determination of molybdenum in steels.

The other important gravimetric method is fairly satisfactory for steels and involves precipitation of molybdenum with α-benzoinoxime and ignition of the precipitated complex to molybdenum trioxide, which is weighed.

The sample is dissolved in dilute (1 : 6) sulfuric acid and the solution is treated with a minimum amount (ca. 2 mL) of concentrated nitric acid to decompose carbides and oxidize the iron and molybdenum. The solution is boiled to expel nitrogen oxides and filtered if necessary. To the cool solution sufficient iron(II) ammonium sulfate solution is added to reduce vanadium(V) and chromium(VI) acids. An excess of a 2 % solution of α-benzoinoxime in ethanol is added at 5 °C, followed by bromine water until the solution is pale yellow, and then more α-benzoinoxime. The precipitate on the filter paper is washed with a freshly prepared solution containing 10 mL of the 2 % reagent solution and 2 mL of concentrated sulfuric acid diluted to 200 mL. After charring, ignition at 500 – 525 °C, and weighing, the residue is treated with 5 mL of concentrated ammonia to remove the molybdenum and tungsten trioxides and the residual solid is ignited and weighed. If tungsten is present the ammoniacal extract is acidified with sulfuric acid and evaporated to dense white fumes. After diluting, the tungsten is precipitated using cinchonine and weighed as the trioxide after ignition at 750 – 850 °C.

Another gravimetric method often described involves precipitation of molybdenum as sulfide and ignition to oxide. This is not satisfac-

tory because loss due to volatilization is likely to occur. The complex of molybdenum with 8-hydroxyquinoline, $MoO_2(C_9H_6ON)_2$, is precipitated within the pH range 3.3 – 7.6. It differs from complexes of the reagent with other metals in being insoluble in organic solvents and in many concentrated inorganic acids. The freshly precipitated compound dissolves only in concentrated sulfuric acid and in hot solutions of alkali-metal hydroxides. This determination allows complete separation of molybdenum and rhenium.

Colorimetric Methods. Molybdenum(VI) in acidic solution can be reduced to give mainly molybdenum(V), which forms a red complex with thiocyanate. This complex can be extracted from water with organic solvents, particularly oxygenated compounds, providing a very sensitive spectrophotometric method. Tin(II) chloride is a suitable reducing agent, preferably used in the presence of a little iron(II). The method can be used to determine molybdenum in steels.

The steel sample is dissolved in mixed hydrochloric – perchloric acids, heated until dense white fumes are evolved and finally heated for a further 6 – 7 min. After cooling the solution is diluted with water and made up to a known volume. An aliquot is reduced with tin(II) chloride in the presence of iron(II) and potassium thiocyanate is added. The color depends on the acid concentration (optimum value 1 mol/L) and the thiocyanate concentration (should not be less than 1 %). The red complex is extracted into 3-methyl-1-butanol and the optical density measured at 465 nm. When the quantity of Mo in a 1-cm cell is 10 µg the following elements cause an error greater than 3 % when present in the amount specified: vanadium > 0.4 mg (interference is prevented by washing the organic extract with tin(II) chloride solution); chromium(VI) > 2 mg; tungsten(VI) > 0.15 mg; cobalt > 12 mg; copper > 5 mg; lead > 10 mg; titanium(III) in the presence of fluoride > 30 mg. The method is also recommended for the determination of molybdenum present as a trace element in fertilizers and animal feeds, the complex being extracted from the aqueous solution using a 1 : 1 mixture of carbon tetrachloride and 3-methyl-1-butanol [125]. The sample solution is prepared by dry ashing the feed at $\leq 450\,^{\circ}C$, treating the ash with hydrochloric acid, and evaporating to dryness. The residue is successively extracted with hot hydrochloric acid, the penultimate treatment being to take to dryness with a little nitric acid. The combined extracts are diluted to give an approximately 1 mol/L solution of acid. Inorganic fertilizers are extracted directly with hydrochloric acid. After evaporation to dryness the residue is treated in the same way as the ash from feeds.

10. Economic Aspects

Three factors form the basic economic character of the molybdenum industry:

1) Molybdenum consumption
2) Molybdenum supply
3) Excess production capacity

The iron and steel industry accounts for ca. 80 % of all *molybdenum consumption* in the Western world. The balance of consumption is split between nonferrous metallurgical uses and chemical applications. Estimated consumption of molybdenum in the Western world in 1989 according to end use is as follows (exports to Eastern-Bloc countries are not included):

Alloy steel	35 %
Stainless steel	27 %
Mo-base and superalloys	12 %
Chemicals	11 %
Tool steel	9 %
Cast irons	6 %

The metallurgical uses of molybdenum are primarily for the fabrication of plant and equipment; its consumption follows the capital goods cycle. The chemical uses for molybdenum involve chiefly catalyst, lubricant, aqueous corrosion inhibition, and color applications whose demand adheres to general industrial activity.

Sources of *molybdenum supply* in the Western world in 1989 were estimated to be as follows:

Primary mines	41 %
By- and coproducts of copper mining	54 %
Imports from China	3 %
Recovery from spent petroleum catalysts	2 %

In general, the cash cost of nonprimary production is significantly lower than the cost of primary mine production. This cost advantage

Table 11. Estimated Western world molybdenum demand (10^3 t contained Mo)

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990 [166]	1992 [166]
United States	27.2	26.3	15.4	14.1	17.7	16.8	17.2	18.6	23.1	24.5	24.0	22.0
Western Europe	29.5	26.8	25.0	24.0	29.5	29.5	29.5	30.4	37.2	37.2	38.0	33.0
Japan	12.7	12.7	11.8	10.9	11.8	12.2	11.8	11.8	15.4	15.0	17.0	14.0
Other Western countries	5.9	6.4	5.9	6.4	6.8	7.3	7.7	7.7	8.2	8.2	8.6	9.0
Exports to Eastern Bloc	11.3	11.3	11.3	10.4	11.3	11.8	10.5	10.4	10.0	10.8	–	–
Total demand	86.6	83.5	69.4	65.8	77.1	77.6	76.7	78.9	93.9	95.7	87.6	79.0

Table 12. Estimated Western world molybdenum production∗ (10^3 t contained Mo)

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
United States	68.3	63.3	38.5	15.2	46.3	49.6	42.5	35.0	43.8	64.0
Chile	13.7	15.4	20.0	15.3	16.9	18.4	16.6	16.9	15.5	17.0
Canada	15.4	17.5	14.4	9.3	8.9	7.4	11.9	12.4	12.5	14.4
Other∗∗	3.4	3.9	8.2	8.4	9.3	9.8	12.5	13.4	11.4	11.5
Total production	100.8	100.1	81.1	48.2	81.4	85.2	83.5	77.7	83.2	106.9
Metal loss	1.5	1.5	1.2	0.7	1.2	1.3	1.3	1.2	1.0	1.5
Net production	99.3	98.6	79.9	47.5	80.2	83.9	82.2	76.5	82.2	105.4

∗ Includes both mine production and recovery from spent petroleum catalyst.
∗∗ Includes Western world imports from the People's Republic of China.

enables nonprimary producers to sell at lower prices than their primary counterparts. Primary production is the first supply segment to react to changes in market conditions.

The molybdenum industry has significant *excess capacity*. Encouraged by unrealistic demand expectations, nearly 50 000 t of primary mine capacity was installed during the late 1970s and early 1980s. This, in combination with the continued growth of byproduct production and the emergence of China as a supplier, has resulted in a Western world supply capacity of over 125 000 t. Western world demand in 1989, including exports to the Eastern-Bloc countries, though at record levels, was only 95 700 t.

Tables 11 and 12 contain estimates for molybdenum supply and demand which show that production exceeded demand during most of the past decade. Tables 13 and 14 list the major molybdenum mines and conversion operations (molybdenum roasting and ferromolybdenum conversion).

11. Environmental Aspects

General Considerations. The regulation of chemical discharges into the environment is increasing dramatically and will continue to do so in the 1990s. Existing restrictions on molybdenum discharges will likely be enforced at the state or local level in the United States and at the national level in Europe.

General aspects of molybdenum in the environment are reviewed in [126]. Additional studies and reviews further explore the impact of molybdenum in the environment [127], [128]. Although molybdenum can cause problems in the environment when discharged in excessive quantities, the problems are well defined and are easily corrected in most instances.

Wastewater. Soluble molybdenum compounds exhibit medium to low levels of toxicity when discharged into either fresh water, salt water, or sewage treatment facilities. Table 15 summarizes relevant ecotoxicity data.

In the United States, effluent guidelines for molybdenum have been established for some industrial categories. These guidelines are technology-based and not toxicity-related. Calculation of individual discharge limitations is based on process production rates. The EPA has not established generic molybdenum limits for industrial discharges nor has it established

stream or drinking water standards for molybdenum.

Table 13. Major molybdenum mines[a]

Parent company	Mine[b]	Location
AMAX – Climax	Henderson (P)	Colorado, USA
AMAX – Climax	Climax (P)	Colorado, USA
Cyprus	Thompson Creek (P)	Idaho, USA
Cyprus	Tonopah (P)	Nevada, USA
UNOCAL – Molycorp	Questa (P)	New Mexico, USA
Placer	Endako (P)	B.C., Canada
BHP	Island Copper	B.C., Canada
Codelco	Andina	Chile
Codelco	Chuquicamata	Chile
Codelco	El Salvador	Chile
Codelco	El Teniente	Chile
Cyprus	Bagdad	Arizona, USA
Cyprus	Sierrita	Arizona, USA
Highland Valley Copper	Highland Valley	B.C., Canada
Magma Copper	San Manuel/ Pinto Valley	Arizona, USA
Mexicana de Cobre	La Caridad	Mexico
Montana Resources	Butte	Montana, USA
National Iranian Copper	Sar Chesmeh	Iran
Noranda[c]	Brenda	B.C., Canada
Phelps Dodge	Chino/Morenci	Arizona, USA
RTZ – Kennecott	Bingham Canyon	Arizona, USA
Southern Peru Copper	Cuajone	Peru
Southern Peru Copper	Toquepala	Peru

[a] Source: Climax Metals Company. [b] P = primary mine.
[c] Closed June 1990 with depletion of mining reserves.

Table 14. Major molybdenum converters∗

Parent company	Location
AMAX – Climax	USA, UK, Netherlands
Anglo Blackwells	UK
Codelco	Chile
Cyprus	USA
Eldorado Gold Mines	Canada
Empresas Frisco	Mexico
Ferro Alloys	UK
H.C. Starck	FRG
Japan Metals & Chemicals	Japan
Molycorp	USA
Molymet	Chile
M. P. Trollhattan	Sweden
Murex	UK
Nippon Denko	Japan
Pechiney	France
Placer	Canada
RTZ – Kennecott	USA
Sadacem	Belgium
Showa Denko	Japan
Treibacher C.W.	Austria

∗ Source: Climax Metals Company.

Exhaust Gas. Molybdenum compounds are not specifically regulated as a component of process exhaust gases. Control of molybdenum compounds in exhaust gases is accomplished by use of standard equipment such as bag houses, cyclones, and scrubbers. Efficiency of removal is typically monitored by noting total suspended particulate discharge rates. Work place exposure

Table 15. Ecotoxicity of molybdenum compounds

Species or medium	Molybdenum compound	Toxicity	Reference
Freshwater fish	sodium molybdate dihydrate	LC_{50} (96 h) fathead minnow, 7630 mg/L	[129]
		LC_{50} (96 h) bluegill, 6790 mg/L	
		LC_{50} (96 h) rainbow trout, 7340 mg/L	
		LC_{50} (96 h) channel catfish, $>$ 10 000 mg/L	
	ammonium dimolybdate	LC_{50} (48 h) bluegill, 157 mg/L	[130]
		LC_{50} (48 h) rainbow trout, 135 mg/L	
	molybdenum trioxide (pure)	LC_{50} (48 h) bluegill, $> 87 < 120$ mg/L	[130]
		LC_{50} (48 h) rainbow trout, $> 65 < 87$ mg/L	
Daphnia	sodium molybdate dihydrate	LC_{50} (96 h), 3940 mg/L	[129]
Saltwater species	sodium molybdate dihydrate	LC_{50} (96 h) mysid shrimp, 3997 mg/L	[131]
		LC_{50} (96 h) sheepshead minnow, 6590 mg/L	
		LC_{50} (96 h) pink shrimp, 3997 mg/L	
		EC_{50} American oyster, 3526 mg/L	
	ammonium molybdate	LC_{50} (48 h) marine shore crab, 1018 mg/L as Mo	[132]
Activated sludge	sodium molybdate dihydrate	no effect level $>$ 2522 mg/L (slug dose)	[133]
Anaerobic sludge	sodium molybdate dihydrate	no effect level $>$ 2522 mg/L (slug dose)	[133]
Mammals	molybdenum trioxide (pure)	LD_{50} oral-SD rat 2.73 g/kg	[134]
	molybdenum trioxide (technical)	LD_{50} oral-SD rat 6.66 g/kg	[135]
	sodium molybdate	LD_{50} ipr-mus 344 mg/kg	[136]
	molybdenum disulfide	no effect level, oral-SD rat $>$ 15 g/kg	[137]

for insoluble molybdenum compounds in the United States is limited to 10 mg/m^3 [138] which is the same as the TLV – TWA for nuisance dusts. Soluble molybdenum compounds in the work place are controlled at 5 mg/m^3 [138]. Exposure limits in other countries may differ slightly from the U.S. values and should be consulted accordingly. Under the United States Emergency Planning and Community Right-To-Know Act, molybdenum trioxide is considered a toxic chemical and is subject to annual chemical release reporting [139].

Solid Waste. Within the European Economic Community (EEC) and some other European countries (e.g., Switzerland), the molybdenum content of sewage sludge used for agricultural purposes is subject to regulation. In the United States, the EPA is developing guidelines for use of sewage sludge in agriculture, and will set acceptable limits for metals. Industrial waste in the United States and the EEC is not specifically regulated as to molybdenum content.

The Netherlands has adopted soil and groundwater contamination guidelines for molybdenum [140]. The pollutants regulated are classified in the "Black" and "Grey" lists, and are derived from a 1976 EEC Directive [141]. Chemicals on the Black list are considered very toxic and are persistent and/or bioaccumulate. Chemicals on the Grey list, which includes molybdenum, are of general concern in the environment.

The most widely studied environmental problem with molybdenum is its impact on ruminants that graze on forage with a high molybdenum content [142]. Environmental protection must therefore first identify the probable land use and then develop the control strategy necessary to mitigate the impact. For example, treatment of soil with sewage sludge has to be controlled so that the soil molybdenum content does not exceed 10 mg/kg [143]. Liming of soil lowers the accumulation of molybdenum in the surface plow layer because of the increased solubility at alkaline pH [143]. Due to the biological antagonism of copper toward molybdenum, dietary supplementation of ruminants with copper mitigates the effects of molybdenosis [143]. Sewage sludge with a high molybdenum content usually has a high copper content too which reduces the probability of molybdenosis in ruminants foraging on sludge-treated pasture. Sludge with less than 50 mg/kg molybdenum on a dry weight basis should therefore be acceptable for long-term land application at an annual addition rate of 10 000 kg/ha.

12. Toxicology and Occupational Health

" Molybdenum compounds in general are of a low order of toxicity, both from the point of view of observed clinical effects as well as from the histopathologic point of view" [144]. Toxicity is primarily associated with soluble molybdenum compounds. For example, both inhalation and ingestion of insoluble molybdenum disulfide and calcium molybdate failed to produce significant histopathological findings in either rat or guinea pigs [144].

Human Nutrition. Molybdenum is an essential trace element. Leguminous plants require molybdenum for bacterial fixation of nitrogen. In animals and humans it serves as a cofactor for the enzymes xanthine oxidase (which oxidizes xanthine to uric acid[145]) and aldehyde oxidase. Probably 25 – 75 % of molybdenum in the diet is absorbed through the gastrointestinal tract and is rapidly eliminated, primarily via the urine. The biological half-life is not well established for humans, but most of the absorbed molybdenum is eliminated within a few days or weeks. Animal data indicate that the metabolism of molybdenum is closely related to the metabolism of copper and sulfur compounds [146].

The normal daily human dietary intake of molybdenum in the United States is estimated to be 100 – 500 µg. High molybdenum concentrations are generally found in leafy vegetables (up to 5 µg/g wet weight); legumes with edible roots have a low molybdenum content (< 1 µg/g wet weight). There are no human or animal studies implying accumulation of molybdenum in the body during the life cycle [147].

The relationship of molybdenum deficiency and esophageal cancer in humans was first reported in 1966 [148]. The high incidence of this cancer was attributed to the consumption of food locally grown in soil low in molybdenum. As reported in 1981, a study in China demonstrated an

inverse relationship between the mortality rate of esophageal cancer and the content of Mo, Zn, Mn, Si, Ni, Fe, Bi, and I in cereal and drinking water samples [149]. Levels of molybdenum in the serum, hair, and urine of the inhabitants of the high-risk area were lower than those in the low-risk areas. Utilizing animal models, this inhibitory effect of molybdenum on esophageal, fore stomach, and mammary carcinogenosis has been demonstrated in laboratory studies [150], [151], [152].

Human Toxicity. Information on the toxic effects of molybdenum in humans is scarce. A 1961 study reported a high incidence of gout in an area of Armenia in which the soil contained 77 mg/kg of molybdenum and 39 mg/kg of copper [147]. Daily dietary intake of molybdenum was calculated to be 10 – 15 mg and that of copper 5 – 10 mg; corresponding values in the control area were 1 – 2 mg Mo and 10 – 15 mg Cu. The study showed an increased prevalence of goutlike joint disorders, elevated blood levels of uric acid, and an increase in the urinary excretion of uric acid [147]. The results are difficult to evaluate since no details are given concerning the study population and the selection of subjects. Furthermore the control group consisted of only five subjects.

A study in another high-molybdenum province in the formerSoviet Union failed to detect biological changes in the study population. A tentative explanation is that a higher intake of natural copper may have protected the population against the molybdenum exposure [147]. Animal studies have demonstrated a biological antagonism of copper toward molybdenum.

A study of four humans in India, detected a significant increase in copper metabolism with dietary intakes of molybdenum ranging from 160 to 2540 µg/d. However, the study failed to detect any effect on the urinary excretion of uric acid [153].

On the basis of the above studies, MERTZ suggests that a dietary intake of more than 10 mg/d of molybdenum is dangerous to human health, but that intakes of up to 1.5 mg/d do not appear to have an effect except for an increase in urinary copper excretion [154]. With a daily molybdenum intake of 540 µg, copper excretion was elevated but was still within normal range. In adult humans with a normal copper intake (1.5 – 2.5 mg/d), MERTZ proposes that a dietary molybdenum intake of 0.5 – 1.0 mg/d can be assumed to be safe. Based on balance studies conducted in areas where the diet is generally adequate in molybdenum and no molybdenum disease has been observed in humans, the U.S. National Academy of Sciences has estimated an adequate and safe intake of molybdenum for humans as being 0.15 – 0.5 mg/d [155].

Animal Toxicity. Most observations on the toxicity of molybdenum are based on animal data. Excessive exposure to molybdenum via food can cause severe disease in ruminants which involves diarrhea, anemia, and emaciation, and may progress to death. The disease was first identified in 1938 as the cause of a long-recognized problem in cattle grazing in areas called "teart" pastures in Somerset, England. By 1943 large doses of copper sulfate were shown to prevent or cure "teart" disease. The clinical signs of molybdenum toxicity in animals are primarily due to copper deficiency. After long exposure, the hair loses its color and lameness with a characteristic stiff gait may develop. The latter symptoms may be associated with disturbance of phosphorous metabolism resulting in osteoporosis and joint abnormalities [156].

Young lambs grazing on high-molybdenum pastures suffer from a bone or joint disorder known as swayback. Sheep also develop steely wool, but not the severe diarrhea seen in cattle [157]. Copper deficiency in lambs also lowers their resistance to infections, such as pneumonia [158].

Evidence indicates that cattle are the most susceptible to molybdenum toxicity of any species, followed by sheep. Horses grazed the "teart" pastures of Somerset without showing any clinical signs of toxicity. Rabbits [159], guinea pigs, rats, pigs, and chickens [158] are comparatively tolerant to molybdenum.

Due to the great variability in results, the data available for cattle and sheep provide little assurance about safe levels of molybdenum; a level of 5 – 10 ppm in pasture forage is usually considered a conservatively safe molybdenum level. The clinical symptoms of molybdenum toxicity or molybdenosis are primarily due to a conditioned or secondary copper deficiency. The most reliable indication of molybdenosis or copper deficiency is alteration of the symptoms by cop-

per supplementation. In some areas where animal forage is deficient in copper, pastures with molybdenum levels normally considered to be in the safe range could trigger signs of molybdenosis. At the other extreme, forage that is naturally high in copper produces copper toxicity and supplemental dietary molybdenum may be necessary.

The results of acute toxicity tests (LD_{50} and LC_{50} values) are given in Table 15. In literature citing acute testing, low LD_{50} values (i.e., high toxicity) are sometimes given for molybdenum compounds. Some of these low levels are the result of long-term feeding studies but are interpreted as single-dose acute studies. For example, the acute oral LD_{50} in rats for molybdenum trioxide is frequently listed in the literature as 125 mg/kg, indicating that it is highly toxic. This LD_{50} was, however, determined by a 44-day feeding study [144], not by a single dose. The acute oral LD_{50} of molybdenum trioxide, pure grade, when determined by a single dose in rats, ranges from 2960 mg/kg to 3520 mg/kg [160].

Occupational Health. Several studies on the industrial exposure of humans to molybdenum have been reported from the former Soviet Union. A 1963 study reported 3 cases of pneumoconiosis out of 19 workers exposed to metallic molybdenum and its oxide. Early signs of pneumoconiosis were identified by X-ray in a 44-year old woman exposed for 5 years to molybdenum concentrations ranging from 1 – 3 mg/m^3 and in a 44-year old man exposed for 4 years to concentrations between 6 and 19 mg/m^3. A 34-year old man exposed for 7 years to concentrations between 6 and 19 mg/m^3 showed fully developed pneumoconiosis [161].

In another former Soviet study 73 workers in a copper – molybdenum plant complex and 10 control subjects were studied [162]. Workers with the highest exposure had the highest blood level of uric acid, and increased blood uric acid was found in 34 out of the 37 workers who complained of arthralgia. The actual values of uric acid in blood were not reported in this study, nor was the exposure defined.

Hyperuricemia was also described in another study on 85 workers in a copper – molybdenum plant in Kadzaran, former Soviet Union [163]. An increase in blood concentrations of bilirubin, globulins, and cholesterol was observed but no values or exposure data were provided. In a study on 500 workers from a former Soviet molybdenum and copper mine, nonspecific symptoms and changes in the central nervous system were reported. Molybdenum dust levels in the mines were in some cases 10 – 100 times higher than the Soviet maximum permissible level of 6 mg Mo/m^3 [164].

The effects of molybdenum exposure was studied in 25 male workers [mean age 28.3 years (19 – 44 years), mean duration of employment 4.0 years (0.5 – 20 years)] from a roasting plant in Denver, Colorado, where molybdenum sulfide was converted to molybdenum oxides [165]. Total dust samples were collected in the roaster area and the 8-h time weighted average (TWA) molybdenum exposure was calculated to be 9.47 mg/m^3. The respirable ($< 10\,\mu$m) molybdenum concentration in the dust at the base of the roaster was 1.02 mg/m^3 and the minimum daily body burden of molybdenum was calculated to be 10.2 mg. Urinary and plasma molybdenum levels were higher for the worker group than for the control group. Except for some generalized medical complaints, the only adverse biochemical findings were elevated serum ceruloplasmin levels and a small increase in serum uric acid the mean value of which was within the normal range of the control group. No gout-like symptoms were reported. Urinary copper excretion was normal in 13 of the 14 workers.

In the United States, health and safety regulations have established permissible exposure limits (PELs) in the occupational environment at 5 mg/m^3 for soluble molybdenum compounds. The Threshold Limit Values (TLVs) have also been established by the ACGIH for soluble and insoluble molybdenum at 5 and 10 mg/m^3, respectively. Both the PELs and TLVs are time-weighted averages (TWAs) for 8 h/d. Although the study of the 25 workers [165] may be considered limited because of the small population and high employment turnover, the absence of any adverse health effects at an exposure level approximately twice the soluble molybdenum PEL is of considerable value in assessing the potential worker risk to molybdenum exposure.

13. References

1. A. Sutulov: *Molybdenum Extractive Metallurgy,* University of Conception, Chile 1965.
2. A. Sutulov: *International Molybdenum Encyclopedia 1778 – 1978,* vol. II, Intermet Publications, Santiago, Chile 1979.
3. J. W. Blossom: *Minerals Yearbook,* **vol. I,** Metals and Minerals, U.S. Bureau of Mines, U.S. Government Printing Office, Washington, D.C. 1987, pp. 633 – 640.
4. J. W. Blossom: *Mineral Facts and Problems,* U.S. Bureau of Mines, Bulletin 675, U.S. Government Printing Office, Washington, D.C. 1985, pp. 521 –534.
5. L. Northcott: *Molybdenum,* Butterworths Scientific Publications, London 1956.
6. S. M. Johnson in: *Mining Annual Review* 1989, Mining Journal, London 1989, p. C 72.
7. Bureau of Mines: *An Appraisal of Minerals Availability for 34 Commodities,* Bulletin 692, U.S. Department of the Interior, Washington, D.C. 1987, p. 191.
8. J. Banfield (ed.): *Financial Times International Year Book 1989,* Longman Group UK Ltd., Harlow, Essex, United Kingdom; St. James Press, Chicago, Illinois 1988.
9. R. R. Dorfler, J. M. Laferty: "Review of Molybdenum Recovery Processes," *J. Met.* **31** (1981) no. 7, 48 – 54.
10. "50 Years of Flotation . . . It's Men, Ideas, and Machines," *Eng. Min. J.* **162** (1961) 83 – 91, 116, 120, 123, 125, 127.
11. V. A. Glembotskii: *Flotation,* Primary Sources, New York 1963.
12. C. A. Born, F. N. Bender, O. A. Kiehn: "Molybdenite Flotation Reagent Development at Climax, Colorado," in D. W. Fuerstenau (ed.): *A. M. Gaudin Memorial Vol. II,* AIME, New York 1976, pp. 1147 – 1184.
13. R. E. Cuthbertson: "Unusual Reagent Combination Improves Flotation at Climax," *Min. Elec. Mech. Eng.* **24** (1944) no. 292, 1675 – 1680.
14. K. P. Neumann, W. A. Gibson: "Amoco Develops Thompson Creek Molybdenum Mine," *Min. Eng. Littleton. Color.* **36** (1984) 27 – 30.
15. D. Malhotra, R. M. Rowe, A. K. Bhasin: "Evaluation of Collectors for Molybdenite Flotation," *Minerals and Metallurgical Processing,* V3, m3, Aug. 1986, pp. 184 – 186.
16. J. F. Shirley: "By-product Molybdenite Plant Design," *Can. Min. J.* **102** (1981) no. 3, 27 – 28.
17. R. D. Crozier: "Flotation Reagent Practice in Primary and By-product Molybdenum Recovery," *Min. Mag.* **7** (1979) 174 – 178.
18. L. F. McHugh, J. Godshalk, M. Kuzlor: "Climax Conversion Practice III (for Molybdenite)," *Conference of the Metallurgical Society of CIM, 1977,* Canadian Institute of Mining and Metallurgy, Montreal 1978.
19. D. G. Lindsay: "Endako Roasting Practice," *Conference of the Metallurgical Society of CIM, 1977,* Canadian Institute of Mining and Metallurgy, Montreal 1978.
20. Amax, Inc., US 4 207 296, 1980 (H. H. K. Nauta, J. W. Kok, J. Harte).
21. Herman C. Starck, US 3 860 419, 1975 (T. A. Weber, R. F. Borrmann).
22. L. F. McHugh, P. L. Sallade: "Molybdenum Conversion Practice," Paper no. 86 – 154, presented at SME Annual Meeting, New Orleans, LA, March 2 – 6, 1986.
23. Engelhard Min. & Chem. Corp., US 4 296 077, 1981 (S. R. Heuer, Z. Zbranek).
24. Akzona Inc., US 4 087 510, 1978 (G. E. Steenken).
25. Union Carbide Corp., US 3 773 890, 1973 (J. S. Fox, J. E. Litz).
26. Amax, Inc., US 4 495 157, 1985 (R. F. Sebenik, P. P. LaValle, J. M. Laferty, W. A. May).
27. Marubeni Corp. and Fuji Fine Chemical Co, US 4 145 397, 1979 (S. Toida, A. Ohno, K. Higuchi).
28. Individual, US 3 256 058, 1966 (B. Burwell, Sr.).
29. *Gmelin,* System no. 53, Molybdän; suppl., part A 1.
30. G. Dowson, *Met. Powder Rep.* **44** (1989) no. 4, 272 – 275.
31. W. C. Hagel, J. A. Shields, S. M. Tuominen, CONF-8308130, US Department of Energy, Oak Ridge, USA 1984.
32. A. J. Bryhan, *Weld. Res. Counc. Bull.* **312** (1986) 1 – 20.
33. C. A. Krier, DMIC, Battelle Information Center, Rep. 263 (1961) Columbus, Ohio.
34. R. Eck, J. Tinzl, *12th International Plansee Seminar, Proc.* **1** (1989) 829 – 843.
35. J. Wadsworth, T. G. Nieh, J. J. Stephens, *Int. Mat. Rev.* **33** (1988) no. 3, 131 – 150.
36. T. J. Patrician, V. P. Sylvester: "Physical Metallurgy and Technology of Molybdenum and its Alloys," *Proc. of Symp.* AMAX Material Research Center, Ann Arbor, Michigan 1985, pp. 1 – 11.

37. J. K. Thorne, J. M. Dahle, L. H. Van Vlack, *Met. Trans.* **1** (1970) 2125 – 2132.
38. Ferrolegeringar Trollhaetteverken AB, DE-OS 2 716 591, 1977 (J. Wallen).
39. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin and Verlag Stahleisen mbH, Düsseldorf 1973.
40. L. Northcott: *Molybdenum,* Butterworth, London 1956.
41. A. G. E. Robiette: *Electric Smelting Processes,* Charles Griffin & Co. Ltd., London 1973.
42. British Intelligence Objectives Sub-Committee: *The German Ferro-Alloy Industry,* B.I.O.S. Final Report No. 798, Trip No. RAT. 71/72 Dec., 1945.
43. R. Durrer, G. Volkert in G. Volkert, K. D. Frank (eds.): *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin 1972.
44. *Ullmann,* 4th ed., **7,** 355.
45. H. Haag, *Erzmetall* **15** (1962) 229 – 234.
46. *Eng. Min. J.* **176** (1975) no. 6, 98 – 99. M. W. Murphy, E. S. Wheeler, A. Linz, *Min. Metall* **27** (1946) 350 – 352.
47. J. H. Young Jr.: *Comparison of Electric Smelting Process with Thermite Process;* Preprint von ILAFA – Latin American Seminar on Ferroalloys, Salvador de Bahia, June 22 – 27, 1975.
48. M. Riss, Y. Khodorovsky: *Production of Ferroalloys,* MIR Publishers, Moscow 1967.
49. V. I. Vasil'ev et al., *Sb. Tr. Chelyab. Elektrometall. Komb.* **4** (1975) 55 – 59; *Chem. Abstr.* **84** (1976) 138 836 t.
50. V. P. Elyutin, Y. A. Pavlov, B. E. Levin, E. M. Alekseev: *Production of Ferroalloys Electrometallurgy,* 2nd ed., Israel Program for Scientific Translations, Jerusalem 1957.
51. *Gmelin,* Molybdenum Suppl., vols. B 1, B 2.
52. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 5th ed., Wiley, New York 1988, pp. 804 – 847.
53. J. Aubry, D. Burnel, C. Gleitzer in P. Pascal (ed.): *Compléments au Nouveau Traité de Chimie Minérale,* vol. 5, Masson, Paris 1976.
54. C. L. Rollinson in A. F. Trotman-Dickenson (ed.): *Comprehensive Inorganic Chemistry,* **vol. 3,** Pergamon Press, Oxford 1973, pp. 632 – 770.
55. G. Brauer (ed.): *Handbook of Preparative Inorganic Chemistry,* 2nd ed., **vol. 2,** Academic Press, New York 1965, pp. 1401 – 1417.
56. *Chem. Uses Molybdenum, Proc. Int. Conf.* **1st,** 1973 *(J. Less-Common Met.* **36** (1974) 1 – 542); **2nd,** 1976 (*J. Less-Common Met.* **54** (1977) 1 – 564); **3rd,** 1979; **4th,** 1982; **5th,** 1985 (*Polyhedron* **5** (1986) 1 – 606).
57. P. C. H. Mitchell, *Q. Rev. Chem. Soc.* **20** (1966) 103 – 118; E. I. Stiefel, *Prog. Inorg. Chem.* **22** (1977) 1; Z. Dori, *Prog. Inorg. Chem.* **28** (1981) 239.
58. F. A. Cotton, *Polyhedron* **5** (1986) 3.
59. A. Muller, *Polyhedron* **5** (1986) 323.
60. A. J. Bard, R. Parson, J. Jordan: *Standard Potentials in Aqueous Solutions,* Marcel Dekker, New York 1985, p. 480; J. Emsley: *The Elements,* Clarendon Press, Oxford 1989, p. 116.
61. A. Kay, P. C. H. Mitchell, *Nature (London)* **219** (1968) 267; *J. Chem. Soc. A* 1970, 2421.
62. J. Bassett, R. C. Denney, G. H. Jeffery, J. Mendham: *Vogel's Textbook of Quantitative Inorganic Analysis,* 4th ed., Longman, London 1978, p. 367.
63. T. L. Woods, R. M. Garrels: *Thermodynamic Values at Low Temperatures for Natural Inorganic Materials (mean values),* Oxford University Press, Oxford 1987.
64. Climax Molybdenum Company: *Thermodynamic Properties of Molybdenum Compounds,* Bulletin Cdb-2, Climax Molybdenum Company, Ann Arbor, Michigan 1954 (some values have been superseded, see [63][74]
65. See [55]p. 1412.
66. J. E. Germain in M. Che and G. C. Bond (eds.): *Adsorption and Catalysis on Oxide Surfaces,* Elsevier, Amsterdam 1985, p. 355; K. Bruckman, J. Haber, T. Wiltowski, J. Catal. 106 (1987) 188; J. Haber, E. Serwicka, Polyhedron 5 (1986) 107; K. Bruckman et al., J. Catal. 104 (1987) 71.
67. C. B. Knobler et al., *J. Chem. Soc. Dalton Trans.* 1983, 1299.
68. F. A. Schroder, J. Scherle, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **28 B** (1973) 46; see also C. B. Knobler, B. R. Penfold, G. T. Wilkins, *J. Chem. Soc. Dalton Trans.* 1980, 248.
69. A. Magneli, *Acta Chem. Scand.* **9** (1955) 1378.
70. B. G. Brandt, A. C. Skapski, *Acta Chem. Scand.* **21** (1967) 661; *Acta Crystallogr. Sect. A Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **A 24** (1968) 699.
71. V. K. Rudenko, *Koord. Khim.* **5** (1979) 307; *(Sov. J. Coord. Chem. (Engl. Transl.)* **5** (1979) 231.

72. See [62]p. 756.
73. M. Greenblat, *Chem. Rev.* **88** (1988) 31.
74. K. C. Mills: *Thermodynamic Data for Inorganic Sulphides, Selenides, and Tellurides,* Butterworths, London 1974.
75. K. S. Liang et al., *J. Non-Cryst. Solids* **42** (1980) 345.
76. R. Chevrel, M. Hirrien, M. Sergent, *Polyhedron* **5** (1986) 87.
77. Climax Molybdenum Company: *Properties of the Simple Molybdates,* Bulletin Cdb-4, Climax Molybdenum Company, Ann Arbor, Michigan 1962.
78. F. Corbet, C. Eyraud, *Bull. Soc. Chim. Fr.* 1961, 571; J. Chojnacki, R. Kozlowski, J. Haber, *J. Solid State Chem.* **11** (1974) 106; J. Haber in P. C. H. Mitchell (ed.): *Climax First International Conference on the Chemistry and Uses of Molybdenum,* Climax Molybdenum Co. Ltd., London 1973, p. 146.
79. A. W. Armour, M. G. B. Drew, P. C. H. Mitchell, *J. Chem. Soc. Dalton Transactions* 1975, 1493.
80. M. T. Pope: *Heteropoly and Isopoly Oxometalates,* Springer Verlag, Berlin 1983.
81. J. B. Black et al., *Polyhedron* **5** (1986) 141.
82. P. C. H. Mitchell, *Wear* **100** (1984) 281.
83. See [52]p. 1046.

General References

84. D. H. Killefer, A. Linz: *Molybdenum Compounds, Their Chemistry and Technology,* Interscience, New York 1952.
85. O. Weisser, S. Landa: *Sulfide Catalysts, Their Properties and Applications,* Pergamon Press, New York 1973.
86. F. J. Clauss: *Solid Lubricants and Self-Lubricating Solids,* Academic Press, New York 1972.
87. T. J. Risdon: *Properties of Molybdenum Disulfide,* Bulletin C-5c, AMAX Mineral Sales, AMAX Inc., Greenwich, Connecticut, February 1987.
88. J. W. Lyons: *The Chemistry and Uses of Fire Retardants,* Wiley-Interscience, New York 1970.
89. R. J. Cole, P. C. H. Mitchell: *Molybdenum Compounds in the Paint and Allied Industries,* Bulletin of Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1971.
90. *Molybdenum Catalyst Bibliography,* Bulletins C-40, C-54, C-73, C-77, C-96, C-111, C119, C-130, C-135, C-139, Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1950 – 1983.

Specific References

91. J. Haber: *The Role of Molybdenum in Catalysis,* Bulletin of Climax Molybdenum Co. Ltd., London, AMAX Inc., Greenwich, Connecticut 1981, pp. 3 – 4.
92. F. Veatch, J. L. Callahan, J. D. Idol, Jr., E. C. Milberger, *Chem. Eng. Prog.* **56** (1960) no. 10, 65 – 67.
93. H. Adkins, W. Peterson, *J. Am. Chem. Soc.* **53** (1931) 1512 – 1516.
94. W. E. Jamison, *ASLE Trans.* **15** (1972) 296 – 305.
95. E. R. Braithwaite, *Sci. Lubr.* **18** (1966) 13 – 18.
96. W. O. Winer, *Wear* **10** (1967) 422 – 452.
97. J. P. G. Farr, *Wear* **35** (1975) 1 – 22.
98. R. S. Barnett: *Molybdenum Disulfide as an Additive for Lubricating Greases,* Bulletin of Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1981.
99. R. D. Loban, D. A. Gresty: *Molybdenum Disulfide in Rubber,* Bulletin of Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1976.
100. R. L. Johnson: *Bonded Lubricant Coatings – A Status Report,* Bulletin of Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1979.
101. T. J. Risdon, D. A. Gresty: *An Historical Review of Reductions in Fuel Consumption of United States and European Engines with* MoS_2, SAE Paper No. 750 674, Houston, Texas, June 1975.
102. P. C. H. Mitchell, *Wear* **100** (1984) 281 – 300.
103. Carbide and Carbon Chemicals Corp., US 2 147 395, 1939 (A. L. Bayes).
104. Carbide and Carbon Chemicals Corp., US 2 147 409, 1939 (H. Lamprey).
105. M. S. Vukasovich, J. P. G. Farr, *Mater. Perform.* **25** (1986) no. 5, 9 – 18.
106. M. S. Vukasovich: *Molybdate – The Versatile Inhibitor,* NACE Paper No. 444, NACE, New Orleans, Louisiana, April 1989.
107. L. C. Rowe, R. L. Chance, M. S. Walker, *Mater. Perform.* **22** (1983) no. 6, 17 – 23.
108. H. O. Schoen, B. G. Brand, *Off. Dig. Fed. Soc. Paint Technol.* **32** (1960) 1522 – 1543.
109. The Sherwin Williams Co., US 3 353 979, 1967 (J. V. Hunn).
110. The Sherwin Williams Co., US 3 677 783, 1972 (T. Kirkpatrick, J. J. Nilles).
111. American Metal Climax Inc., US 3 726 694, 1973 (F. W. Moore, D. R. Robitaille, H. F. Barry).

112. T. R. Truax, C. A. Harrison, R. H. Baechler, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **231** (1933) 107 – 123.
113. J. E. Ramsbottom: *The Fireproofing of Fabrics,* His Majesty's Stationery Office, London 1947.
114. F. W. Moore, G. A. Tsigdinos: "The Role of Molybdenum in Flame Retardancy and Smoke Retardation," in P. C. H. Mitchell, A. Seaman (eds.): *Proc. 2nd Intl. Conf. Chemistry and Uses of Molybdenum,* Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1976, pp. 145 – 149.
115. F. W. Moore, G. A. Tsigdinos: "Advances in the Use of Molybdenum Additives as Smoke Suppressants and Flame Retardants for Polyvinyl Chloride," *Proc. Intl. Symp. Flammability and Fire Retardants,* Technomic Press, Westport, Connecticut 1978, pp. 160 – 172.
116. M. Das, P. J. Haines, T. J. Lever, G. A. Skinner: "The Role of Molybdenum Trioxide as a Flame Retardant and Smoke Suppressant in Halogenated Polyester Thermosets," in H. F. Barry, P. C. H. Mitchell (eds.): *Proc. 4th Intl. Conf. Chemistry and Uses of Molybdenum,* Climax Molybdenum Co., AMAX Inc., Greenwich, Connecticut 1982, pp. 218 – 223.
117. A. I. Vogel: *Quantitative Inorganic Analysis,* 5th ed., Longman, London 1989.
118. R. A. Mostyn, A. F. Cunningham, *Anal. Chem.* **38** (1966) 121.
119. *Analytical Methods in Atomic Absorption Spectrophotometry,* Perkin Elmer Corp., Norwalk, Connecticut 1971.
120. M. D. Amos, J. B. Willis, *Spectrochim. Acta* **22** (1966) 1325.
121. D. J. David, *Analyst (London)* **86** (1961) 730.
122. D. M. Knight, M. K. Pyzyna, *Atomic Absorption Newsletter* **8** (1969) 129.
123. L. R. P. Butler, P. M. Mathews, *Anal. Chim. Acta* **36** (1966) 319.
124. D. Hutchinson, *Analyst (London)* **97** (1972) 118.
125. *Official Methods of Analysis,* 10th ed., Association of Official Agricultural Chemists, Washington, D.C. 1965.
126. W. R. Chappell, K. K. Peterson: *Molybdenum in the Environment,* vols. 1 and 2, Marcel Dekker, New York – Basel 1976.
127. C. F. Mills, I. Bremner, J. K. Chesters (eds.): "Trace Elements in Man and Animals," *Proc. Int. Symp. Trace Elem. Man Animals 5th,* Commonwealth Agricultural Bureaux, Farnham, United Kingdom 1985.
128. L. S. Hurley, C. L. Keen, B. Lonnerdal, R. B. Rucker (eds.): *Trace Elements in Man and Animals 6,* Plenum Press, New York – London 1988.
129. *Acute Toxicity of Sodium Molybdate to Bluegill (Lepomis macrochirus), Rainbow Trout (Salmo gairdneri), Fathead Minnow (Pimephales promeals), Channel Catfish (Ictalurus punctatus), Water Flea (Daphnia magna) and Scud (Gammarus fasciatus),* Bioassay Report Submitted to Climax Molybdenum Company of Michigan, Bionomics, Inc., Wareham, Massachusetts, December 1973.
130. *Acute Toxicity of Ammonium Molybdate and Molybdic Trioxide to Bluegill (Lepomis macrochirus) and Rainbow Trout (Salmo gairdneri),* Bioassay Report Submitted to AMAX, Inc., Bionomics, Inc., Wareham, Massachusetts, January 1975.
131. D. W. Knothe, Dr. G. G. Van Riper: "Acute Toxicity of Sodium Molybdate Dihydrate (Molyhibit 100) to Selected Saltwater Organisms," *Bull. of Environ. Contam. and Toxicol.* **40** (1988) 785.
132. O. J. Abbott: "The Toxicity of Ammonium Molybdate to Marine Invertebrates," *Marine Poll. Bull.* **8** (1977) no. 9, 204 – 205.
133. *Evaluation of Acute Effects of Sodium Molybdate on the Activated Sludge Process and on the Batch Anaerobic Sludge Digestion Process,* Laboratory Report for AMAX Inc., Stearns Catalytic Corp., January 1985.
134. *Acute Oral LD_{50} Assay in Rats,* FDRL ID: 81-0393 for AMAX Inc., Aug. 1981.
135. *Acute Oral LD_{50} Assay in Rats,* FDRL ID: 81-0394 for AMAX Inc., Aug. 1981.
136. N. Irving Sax: *Dangerous Properties of Industrial Materials,* 6th ed., Van Nostrand Reinhold Co., New York 1984, p. 1953.
137. *Acute Oral Toxicity in Rats of MoS_2,* FDRL ID: 9589 A for AMAX Inc., Nov. 29, 1987.
138. *Threshold Limit Values and Biological Exposure Indices for 1989 – 1990,* American Conference of Governmental Industrial Hygienists, Cincinatti, Ohio 1990.
139. *Emergency Planning and Community Righ-to-Know Act of 1986,* enacted by Public Law 99 – 499,October 17, 1986.
140. Wet Bodembescherming (Wbb), Soil Protection Act, The Netherlands 1988.
141. Council of the European Communities Directive, 76/464/EEC, May 4, 1976.
142. W. S. Ferguson, A. H. Lewis, S. J. Watson, *Nature (London)* **141** (1938) 553.

143. S. B. Hornick, D. E. Baker, S. B. Guss: "Crop Production and Animal Health Problems Associated with High Soil Molybdenum," *Molybdenum in the Environment,* **vol. 2,** Marcel Dekker, New York – Basel 1977.
144. L. Fairhall, R. Dunn, N. Sharpless, E. Pritchard: "The Toxicity of Molybdenum," *U.S. Public Health Bull.,* no. 293, 1945.
145. J. Johnson, K. Rajagopalan: "Molecular Basis of the Biological Function of Molybdenum," *J. Biol. Chem.* **249** (1974) 859 – 866.
146. L. Friburg, P. Boston, G. Nordberg, M. Piscator, K. Robert: *Molybdenum – A Toxicological Appraisal,* EPA 600/1-75-004, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 1975.
147. V. Kovalskii, G. Yarovaya, D. Schmavonyan: "Changes in Purine Metabolism in Humans and Animals Living in Biogeochemical Areas with High Molybdenum Concentrations," *Z. Obsc. Biol.* **22** (1961) 179.
148. R. Burrell, W. Roach, A. Shadwell: "Esophageal Cancer in the Bantu of the Transkei Associated with Mineral Deficiency in Garden Plants," *J. Natl. Cancer Inst.* **35** (1966) 201 – 209.
149. S. Yang et al.: "Molybdenum Deficiency and Esophageal Cancer in China," *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **40** (1981) 918.
150. S. Yang, H. Wei, X. Luo, H. Sproat: "Effect of Molybdenum on N-Nitrososarcosine Ethyl Ester-Induced Carcinogenesis in Rats," *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **41** (1982) 280.
151. S. Yang, X. Luo, H. Wei: "Inhibitory Effects of Molybdenum on Esophageal and Forestomach Carcinogenesis in Rats," *JNCI J. Natl. Cancer Inst.* **17** (1980) 75 – 80.
152. S. Yang, H. Wei, X. Luo: "Effects of Molybdenum and Tungsten on Mammary Carcinogenesis in S. D. Rats," *JNCI J. Natl. Cancer Inst.* **74** (1985) 469 – 473.
153. Y. Doesthale, C. Gopalan: "The Effect of Molybdenum Levels in Sorgham (Sorgham Vulgare Pers) on Uric Acid and Copper Excretion in Man," *Br. J. Nutr.* **31** (1974) 351 – 353.
154. W. Mertz: "Defining Trace Element Deficiencies and Toxicities in Man," *Molybdenum in the Environment,* vol. 1, Chap. 18, Marcel Dekker, New York 1976.
155. National Academy of Sciences: *Mineral Tolerance of Domestic Animals,* Washington, D.C. 1980.
156. G. Nordheim, G. Frostlie: "Copper Molybdenum Interactions in Relation to Chronic Copper Poisoning in Sheep in Norway," *Nutr. Rev. Suppl.* 1985, 535 – 538.
157. A. Lesperance, G. Cook, V. Bohman, E. Jensen: "Interrelationship of Molybdenum and Certain Factors to the Development of the Molybdenum Toxicity Syndrome," *J. Animal Sci.* **25** (1966) 96 – 101.
158. N. Suttle: "Copper Deficiency in Ruminants; Recent Developments," *Vet. Rec.* **119** (1986) 519 – 522.
159. L. Arlington, G. Davis: "Molybdenum Toxicity in the Rabbit," *J. Nutr.* **51** (1953) 295.
160. Food and Drug Research Laboratories, Inc: *Acute Oral LD_{50} Assay in Rats for Molybdenum Trioxide, Pure Grade,* Waverly, New York 1981.
161. O. Mogilevskaja: "Experimental Studies on the Effect on the Organism of Rare, Dispersed and Other Metals and Their Compounds Used in Industry," in Z. I. Izraelson (ed.): *Toxicology of the Rare Metals,* Israel Program for Scientific Translations Ltd., Jerusalem 1967.
162. O. Akopajan: "Some Biological Shifts in the Bodies of Workers in Contact with Molybdenum Dust," *Second Scientific Conference of The Institute of Labor of Hygiene and Occupational Diseases on Problems of Labor Hygiene and Occupational Pathology,* Erevan 1963, pp. 103 – 106.
163. M. Avakajan: "A Dynamic Study Of The Experimental Effect Of Molybdenum On Some Metabolic Processes," *Scientific Session On Problems of Labor Hygiene and Occupational Pathology In The Chemical and Mining Industries,* Alastan, 1966.
164. S. Eolajan: "The Effects of Molybdenum on the Nervous System," *Z. Exp. Klin. Med.* **5** (1965) 70 – 73.
165. P. S. Walravens et al.: "Biochemical Abnormalities in Workers Exposed to Molybdenum Dust," *Arch. Environ. Health* **34** (1979) 302 – 308.
166. *Kirk-Othmer* **16** (1996) 926.Wiley

Monochloroethane → Chlorinated Hydrocarbons
Monochloromethane → Chlorinated Hydrocarbons

Monoclonal Antibodies

Giovanni L. Galfre, Celltech Ltd., Slough, United Kingdom (Chap. 1-5, 7)

David S. Secher, Celltech Ltd., Slough, United Kingdom (Chap. 6)

Patrick E. Crawley, Celltech Ltd., Slough, United Kingdom (Chap. 1-5, 7)

1. Introduction 1
2. Derivation of Murine Monoclonal Antibodies by Somatic Cell Fusion 2
3. Human Monoclonal Antibodies 5
4. Production 6
5. Purification 8
6. Monoclonal Antibody Patents 9
7. Uses 9
8. References 11

1. Introduction

Antibodies are proteins produced in the blood of all vertebrate species and are one of the mechanisms for defending the organism against invasion by bacteria, viruses, or other foreign material. Each animal can produce millions of different antibodies, each antibody being able to bind specifically to a particular foreign substance known as an *antigen*.

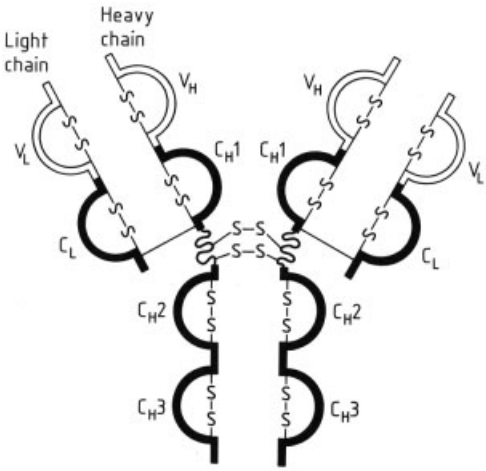

Figure 1. Schematic representation of an immunoglobulin (Ig) molecule (reproduced with permission [2])
The tertiary structure of Ig molecules is composed of interacting protein domains. The heavy chain domains (C_H2 and C_H3) interact with like domains; the light chain domains (C_L and V_L) interact with heavy chain domains (C_H1 and V_H). Inter- and intrachain disulfide bridges (–S–S–) stabilize the structure.

All antibodies have a similar structure based on a unit of four polypeptide chains (Fig. 1): two identical *heavy chains* (M_r ca. 50 000) and two identical *light chains* (M_r 25 000) [1]. The protein fraction containing antibodies (the immunoglobulin or Ig fraction) can be purified from blood by simple physicochemical procedures (e.g., salt precipitation, →Immunotherapy and Vaccines, Chap. 5.1.). This fraction contains a very large number of antibody molecules that are structurally similar but differ in their antigen-binding specificity. Further purification of a single antibody must, therefore, rely on specific antigen-binding properties. Purification methods based on this principle have been extensively used but the heterogeneity of the normal immune response to even a single, simple antigen has limited their usefulness.

In the 1960s it was shown that antibodies are made by *lymphocytes* (a type of white blood cell) and that each lymphocyte makes only a single antibody. Maintenance of clones derived from single lymphocytes in long-term culture, could therefore provide a source of single, pure antibodies. Occasionally, this happens naturally in a form of cancer (*myeloma*) that results from the uncontrolled proliferation of a clone of cells derived from a single lymphocyte. Pure antibody in large quantities can be derived from the blood of tumor-bearing animals but the rarity of the disease and the apparent randomness of the transformation event make it impossible to identify the specificity of the antibody.

In 1975 Georges Köhler and César Milstein discovered a general method for the immortalization of lymphocytes making antibodies of predefined specificity [3].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_699

They injected a mouse with a single antigen to boost the number of lymphocytes making specific antibodies. Lymphocytes removed from the mouse were then fused with cells derived from a previously isolated mouse myeloma. The resulting hybrid cell (*hybridoma*) was able to grow permanently like the parent myeloma and secreted the same specific antibody as the parent lymphocyte. Hybridomas derived from a single lymphocyte are monoclonal and the antibodies they produce are called *monoclonal antibodies*.

Since its introduction, monoclonal antibody technology has had a dramatic impact on life science research and clinical medicine. The extensive introduction of immunological methods based on monoclonal antibodies amounts to a revolution, the importance of which was recognized by the award of the Nobel Prize for medicine in 1984 to Köhler and Milstein.

Although other methods of lymphocyte immortalization have been used to produce monoclonal antibodies, the original method of Köhler and Milstein (Chap. 2) remains the basis for the majority of monoclonal antibody derivation today. It has been very successfully used to produce mouse and rat hybridomas, but the derivation of human monoclonal antibodies has proven particularly difficult (Chap. 3). The derivation of hamster hybridomas is becoming a well-established method using mouse myelomas as fusion partners to derive hamster – mouse *heterohybridomas* [4]. The production of heterohybridomas of either rabbit [5] or cow origin have been reported [6] but not developed into general methods. With the rapid progress of genetic engineering techniques, direct antibody gene cloning may become the method of choice for the future derivation of monoclonal antibodies (Chap. 3).

2. Derivation of Murine Monoclonal Antibodies by Somatic Cell Fusion

Somatic cell fusion is the process by which two somatic (body) cells are fused together to give a viable hybrid cell that carries some of the genotypic and phenotypic characteristics of the two parent cells. For the derivation of hybridomas, poly(ethylene glycol) (PEG) is used to fuse a large number of spleen cells from an immunized donor with cells of a suitable myeloma line.

Spleen cells from immunized animals can produce specific antibodies, but die after a short time in culture. Myeloma cells have been adapted to grow permanently in culture, and mutants have been isolated that lack either the enzyme hypoxanthine guanine ribosyltransferase (azaguanine-resistant) or thymidine kinase (bromodeoxyuridine-resistant). Unlike normal cells, they are unable to utilize hypoxanthine or thymidine for DNA replication. Such mutants therefore cannot grow in media containing aminopterin and supplemented with hypoxanthine and thymidine (HAT-selective media); aminopterin blocks the utilization of essential amino acids for the synthesis of these purine and pyrimidine bases. Hybrids between these mutant cells and spleen cells can therefore be selected from the parents because they are the only cells that actively multiply in HAT-selective media.

From the growing hybrids, individual clones can be selected that secrete the desired antibodies. Such antibodies are thus of monoclonal origin. The selected clones, like ordinary myeloma lines, can be maintained indefinitely.

A scheme for the general procedures involved in the derivation of monoclonal antibodies is illustrated in Figure 2. Overall the process takes between three months and one year depending on the antigen and on the monoclonal antibody characteristics required. A number of well-defined steps can be identified:

1) The animal is immunized with the antigen of choice
2) The myeloma cells are prepared in sufficient quantities and proper growth phase (spinner culture)
3) Spleen cells and myeloma cells are fused and hybrids selected by the use of HAT medium
4) Antibodies of particular interest are isolated and established in culture by repeated cycles of cloning and screening
5) Frozen stocks of hybridomas are prepared throughout the procedure so that if individual cultures are intentionally or accidentally stopped it is possible to go back to earlier stocks
6) Selected clones are propagated in vitro or in vivo for large-scale production of monoclonal antibody and for the preparation of consistent banks of vials that can be kept below −80 °C for several years

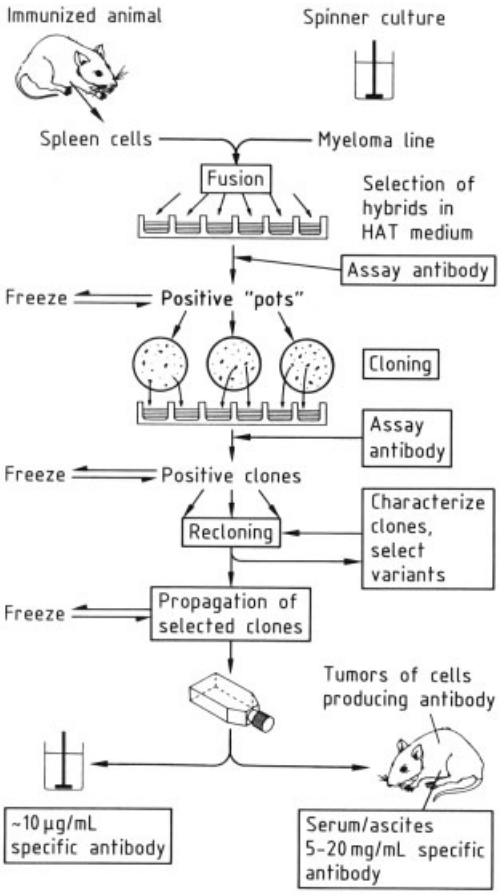

Figure 2. The process used for the derivation of monoclonal antibodies (reproduced with permission from [7])

Although this basic scheme has remained unchanged since 1976, a large number of variations have been introduced; for efficient derivation of specific monoclonal antibody the procedure must be carefully planned and prepared.

Cells and Fusion. The vast majority of monoclonal antibodies are made from immunized mice, but rats are also used. The choice of animal strains and the immunization procedure depend on the antigen to be used and on the requirements for the monoclonal antibody product.

A number of mouse and rat myeloma cell lines have been developed specifically for hybridoma derivation (Table 1). They differ in their specific drug resistance and their ability to secrete myeloma proteins. Since myelomas originate from lymphocytes, they generally secrete immunoglobulins of undefined specificity (hence the term myeloma proteins rather than antibody). Hybridomas derived from them secrete both the antibodies derived from immunized spleen cells and the myeloma protein. The two sets of immunoglobulin are difficult to separate, and it is generally a considerable advantage to use mutant myelomas that no longer make myeloma proteins but are still able to secrete antibody when fused with spleen cells. A number of protocols have been developed for cell fusion [10]. Most are based on mixing spleen cells and myeloma cells and treating the mixture for a few minutes with poly(ethylene glycol) solutions. An important alternative to the use of a chemical fusing agent is the use of a strong electric field [18].

Table 1. Myeloma cell lines used for hybridoma derivation

Cell line	Strain	Immunoglobulin expression ∗	References
Mouse Lines			
P3-X 63 Ag.8	BALB/c	mH/mL	[3]
X 63 Ag 8.653	BALB/c	none	[8]
NSI/1 Ag 4.1	BALB/c	mL (nonsecreted)	[9]
NS 0/U	BALB/c	none	[10]
SP 2/0-AG 14	BALB/c	none	[11]
MPC 11.45.6 Tg 1.7	BALB/c	mL	[12]
MPC 11.45.6 Tg 1.7.5	BALB/c	mL	[13]
S 194/5.XX 0.Bu	BALB/c	none	[14]
Rat Lines			
IR 983 F	LOU	none	[15]
Y 3 Ag 1.2.3	LOU	rL	[16]
YB 2/0	(LOU × AO)	none	[17]

∗ mH, mL, and rL refer to mouse heavy, mouse light, and rat light immunoglobulin chains of myeloma origin (myeloma proteins).

Screening. Fusion results in a viable mixture of residual nonfused spleen cells and a few hybridomas (generally 1 for every 10 000 spleen cells). This mixture is aliquoted in a number of cultures and incubated with selective media. In 5 – 20 days, the only remaining cells in the cultures will be hybridomas. Intensive screening is then necessary to identify the cultures that contain a hybridoma secreting the antibody of interest. Screening consists of harvesting spent culture media and subjecting these to immuno-

logical tests to detect the presence of specific antibody.

The choice of assay used for the screening stage is extremely important. Different screening assays detect antibodies with different characteristics, e.g., different antigen specifity, cross-reactivity with molecules similar to the specific antigen, different binding affinity, and different antibody effector functions. Immunologists have developed an enormous variety of methods for detecting the presence of antibodies ranging from precipitation reactions and radioimmunoassays to assays based on the biological activity of the recognized antigen [19]. These methods can very often be adapted.

Binding assays are probably most generally used on account of their specificity, accuracy, and versatility. Insoluble antigens (either naturally insoluble or absorbed on an appropriate solid support) are allowed to react with the antibody in the culture fluid. The free antibody is washed away. The amount of monoclonal antibody bound is measured directly (direct binding assay) or by binding a second, labeled antibody capable of recognizing the first (indirect binding assay). This second antibody can be labeled in several ways: radioactively labeled, fluoresceinated, and enzyme-linked derivatives are the most common.

Although less extensively used, *assays based on the biological activity* of the antigen are of particular importance because they allow the detection of antibody to antigens that are only identified by their biological activity but cannot be chemically purified. The first example of this was the production of monoclonal antibody to human interferon [19]. A biologically active antigen preparation can be added to the culture supernatants to be screened. After incubation a decrease in biological activity is assumed to be preliminary evidence for the presence of inhibitory antibody. Alternatively, the precipitation of antibody – antigen complexes can be effected (for instance by addition of anti-mouse or anti-rat immunoglobulin) thus removing part of the biological activity from the supernatant.

Cloning. Once cultures of hybridomas secreting specific antibody are identified, they can be cloned (i.e., individual cells physically separated and propagated in individual cultures). The best way is to culture the cells embedded in a semisolid agar-containing medium [10], [20]. Clones are generally rescreened, often using a battery of screening assays to select clones meeting more stringent criteria of antibody production and specificity.

Probably the most important feature of hybridoma technology is that immortalization by cell fusion and cell cloning allow the preparation of antibody specific for an individual antigenic determinant present in a mixture, without the need for purifying the antigen.

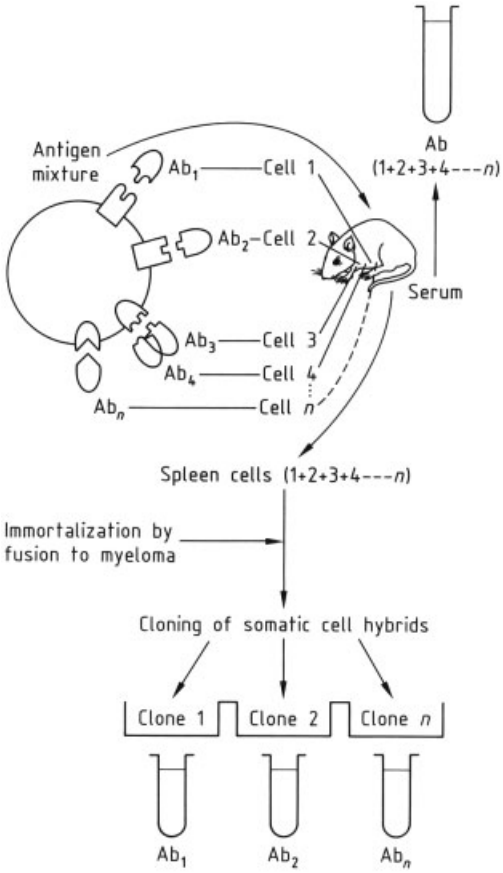

Figure 3. Antibody (Ab) production by animals and by hybridomas (reproduced with permission from [21])
Isolation of hybridoma clones allows the preparation of monospecific antibodies starting from a complex mixture of antigens.

When an animal is injected with an antigenic mixture, its usual response is the production of a highly heterogeneous population of antibody directed against the immunogen. Some of these antibodies are specific for various com-

ponents of the antigenic mixture and are mixed with many others that express other properties. When monoclonal antibodies are prepared, the collection of randomly derived clones represents a cross-section of such a heterogeneous population. The strength of the response to each individual antigenic determinant is reflected in the hybridoma population producing antibody to that determinant (Fig. 3).

Thus the derivation of monoclonal antibody to antigens present in a complex mixture has become a powerful tool for the definition and identification of previously unknown components of the mixture. The most spectacular example is the derivation of a very large number of monoclonal antibodies for mammalian cell surface antigens to define or "map" the structure of the cell surface [22].

3. Human Monoclonal Antibodies

For many potential medical applications of monoclonal antibodies the ultimate goal is a general method for the production of human monoclonal antibodies. Although several instances of human monoclonal antibody production have been reported, two main problems have been encountered in the development of a general methodology. The first is the problem of obtaining highly immunized human lymphocytes for use in the fusion. The second is the production of a monoclonal antibody in the absence of suitable myeloma partners.

A key feature of the standard technique for mouse monoclonal antibody production is the ability to repeatedly inject laboratory mice with the specific antigen. Although this is sometimes possible in humans (e.g., tetanus toxoid vaccination), it is clearly unethical to immunize humans with most antigens. The need for alternative sources of immunized human lymphocytes has led to the development of two new techniques that avoid the requirement of injection of the antigen into humans. The first consists of adding the antigen to a suspension of human blood lymphocytes in a culture medium that allows the necessary differentiation of the lymphocytes to antibody-producing cells [23]. Although not yet a routine operation, this in vitro immunization method has become more feasible with the availability of recombinant growth factors and lymphokines [23]. A more recent and promising advance is the discovery that human lymphocytes can be grown in a strain of mice suffering from Severe Combined Immune Deficiency disease (SCID mice). Human lymphocytes are used to repopulate the immune system of the mouse, and the animal responds to immunization by making human antibodies. Human cells growing in the spleen of such an animal have been fused with myeloma cells to produce a human monoclonal antibody [24].

This exciting development may open the door to the first routine method for the derivation of human monoclonal antibodies. A step beyond transfer of human lymphocytes to a mouse host would be the transfer of the genes coding for human antibodies. This may soon become reality with the development of transgenic mice [25].

However immunization is performed, a second major problem in the production of human monoclonal antibody is the absence of suitable myeloma partners. Mouse – human heterohybridomas are unstable and none of the available cell lines has become generally accepted. Despite considerable efforts, no reliable, consistent methodology based on somatic cell fusion alone has emerged. Alternative approaches to lymphocyte immortalization have therefore been explored.

Even before the advent of hybridoma technology the natural transformation of human lymphocytes by viruses had been used to derive stable cell lines. In most cases, however, these were unable to secrete useful amounts of antibody over a long period of time. The combination of viral transformation followed by somatic cell fusion has resulted in the most successful technique presently available for the derivation of human monoclonal antibodies [26].

Due to the above problems, the production of human monoclonal antibodies by hybridomas is likely to be based on genetic engineering. Antibodies of all species have relatively independent protein domains of about 110 amino acids folded in a characteristic shape (Fig. 1). Although the amino acid sequence is species-specific, the domain shape is common to all known antibodies and the domain carrying the antigen-binding specificity (V domains in Fig. 1) can be transplanted from the antibody of one species to another. Although this cannot be achieved directly on the protein, recombinant DNA technology

has allowed antibody genes to be cloned and antibody gene domains to be reshuffled. The engineered genes are then expressed in a suitable mammalian host cell line that can be grown like a hybridoma. An example of a "reshuffled" antibody is the "chimeric" antibody that consists of the antigen-binding domain of the mouse monoclonal antibody B 72.3 attached to the remaining domains from a human antibody. The original (mouse) B 72.3 recognizes an antigen present on human tumor cells; the chimeric antibody has identical binding properties to the mouse antibody and is now being tested in clinical trials as a potential therapy for colon cancer [27]. Chimerization of antibodies has become a subject of considerable industrial interest.

A recent refinement of this technology has been achieved by grafting onto a human antibody molecule only those parts of the antigen-binding domain that are actually in physical contact with the antigen [28]. Whether this will prove to be a practical improvement for chimeric antibodies remains an open question.

4. Production

Monoclonal antibodies accumulate in the spent media of growing hybridomas at an average concentration of 10 – 100 mg/L. The productivity is a characteristic of individual clones, but is strongly affected by culture conditions. Under optimal conditions productivity may reach 500 – 600 mg/L.

Small volumes of supernatant can be collected either from stationary cultures ("flasks" of up to 1 L) or from slowly rotating bottles ("rollers" of up to 3 L). These conditions are generally satisfactory only for laboratory-scale production (<1 g). For larger quantities two alternatives are available: the hybridomas are either grown as tumors in animals or in an industrial-scale cell-culture plant.

Whichever culture technique is used, a number of important considerations apply:

1) The system must be effectively sterilized before introduction of cells, and maintained free of microorganism contamination for prolonged periods (months for continuous cultures). Small vessels may be sterilized by autoclaving, but large production plant must be sterilized in situ, generally by steaming at high temperature and pressure. Materials and engineering standards have to be of high quality.
2) For effective cell growth and antibody production, the system must provide adequate mixing. Heat and mass transfer characteristics must be sufficient to maintain correct temperature, pH, dissolved gas and nutrient concentrations.
3) Maximum economy of scale cannot be achieved by simply increasing the number of production units.

Production in Animals. When hybridomas are grown in animals (either as solid tumor or as ascites), the monoclonal antibody accumulates in the blood or in the peritoneal cavity at very high concentration (5 – 20 mg/L) but the volume that can be harvested from each animal is rather small and therefore large numbers of animals are required (around 20 000 mice to produce 1 kg of antibody).

Production in Cell Cultures. In addition to the ethical considerations, production in cell cultures has other considerable advantages over production in animals:

1) Monoclonal antibody purified from blood is always contaminated with other antibodies of the same species and class, naturally produced by the animal, and these are difficult to eliminate.
2) The risk of contamination with rodent viruses is eliminated.
3) The process can be automated and made reproducible
4) Scaling up the process by increasing the number of animals does not provide the economy of scale offered by scaling up of industrial fermentation.

With product safety legislation becoming ever more stringent it may soon even be impossible to produce antibodies for therapeutic use in animals.

A wide variety of systems have been developed for in vitro production of monoclonal antibody and are treated in depth in [29]. The most relevant methods will only be summarized here. They depend on growth in free suspension (homogeneous suspension cultures) or

as immobilized cell cultures. More than 50 % of monoclonal antibody production is probably performed with homogeneous suspension cultures.

Homogeneous Suspension Culture. Two types of homogeneous suspension systems have been developed and are used extensively: continuously stirred deep-tank fermenters (originally developed for microorganisms) and airlift fermenters. Although not yet extensively used for monoclonal antibody, stirred deep-tank fermenters (up to 80 000-L capacity) have been used to grow lymphoblastoid cells for interferon production [30]. The airlift fermenter has an internal, concentric draught tube (Fig. 4). Gas mixture introduced from a sparge ring inside the base of the tube causes the culture medium to circulate. Control of gas pressure controls the circulation, and changes in gas mixture composition are used to adjust the pH and dissolved oxygen tension. Airlift fermenters of up to 2000-L capacity have been used for monoclonal antibody production [31].

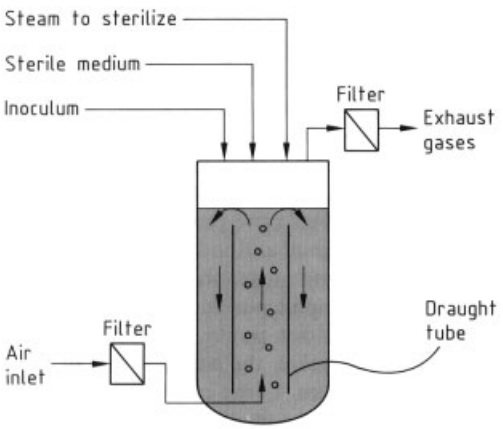

Figure 4. Schematic diagram of an airlift fermenter

Suspension cultures are usually operated in a batch mode but may also be used for continuous culture in which a feed of fresh medium is added to the culture over a prolonged period and the culture is at the same time continuously harvested at an appropriate scale to maintain a steady state [32].

Immobilized Cell Cultures. Various systems have been developed in which the cells are immobilized on a solid matrix (microcarrier), in microbeads, in hollow-fiber reactors, or on ceramic or glass matrices.

The use of *microcarriers* was first developed for bulk cultivation of anchorage-dependent cells [33] but microcarriers with surface charge properties suitable for the absorption of suspension cells are now used. Suitable microcarrier materials include cross-linked dextran, collagen, gelatin, glass, plastic, cellulose, and even droplets of a liquid that is immiscible with the culture medium (perfluorinated hydrocarbons, [34]). Cells immobilized on microcarriers can be grown in either perfusion culture systems or in fluidized-bed reactors.

In *microbead culture*, cells are encapsulated in semipermeable microbeads of agarose or alginate. The technique of alginate encapsulation, pioneered by Lim and Son [35] has been developed for industrial-scale cultivation by Damon Corporation [36]. As with microcarriers, encapsulation of cells lends itself to perfusion culture. Cells can grow to very high population densities and crude products can be retained at high concentration within the capsules.

Monoclonal antibodies have also been produced using *hollow-fiber reactors* which consist of loose bundles of semipermeable capillary fibers enclosed within a culture chamber. Cells grow within the culture chamber in the extracapillary space while the culture medium is circulated through the capillary. Gases, nutrients, and catabolites of low molecular mass diffuse between the culture chamber and the inside of the fibers along their respective concentration gradients; macromolecular products including the monoclonal antibodies are retained within the culture chamber from where they can be harvested at relatively high concentration [37].

The hollow-fiber principle is also used in the flat-bed membrane bioreactor [38] which contains three chambers: a culture chamber bounded on one side by a perfusion chamber and on the other by a harvest chamber. A semipermeable membrane separates the central culture chamber from the perfusion chamber through which a feeding medium is circulated. The other side of the culture chamber is separated from the harvest chamber by a microporous membrane, which allows the passage of macromolecules but not cells. Macromolecules, including the desired antibody product can thus be recovered from the harvest chamber.

Ceramic or glass matrices have also been developed for the immobilization of hybrido-

mas [39]. They are used in reactors in which the culture chamber is filled with porous rigid honeycomb units or beads.

5. Purification

Antibody can be recovered from cell cultures or from ascitic fluid by a wide variety of methods depending on the purity required. For preparation of diagnostic reagents, ascitic fluids or concentrated spent culture media can often be used directly, although some simple purification is generally used. High-purity material is, however, necessary when the monoclonal antibody is to be injected into patients for diagnostic (e.g., tumor imaging) or therapeutic purposes. Levels of contaminants such as proteins, nucleic acids, and pyrogens must also be reduced to acceptable levels. Furthermore the purification process must be validated to confirm that it removes contaminating substances irrespective of whether their presence can be demonstrated in the antibody source.

To achieve high purity, control of the environment in which the purification process is run is essential. Equipment used exclusively for monoclonal antibody production must be installed in a purpose-built facility.

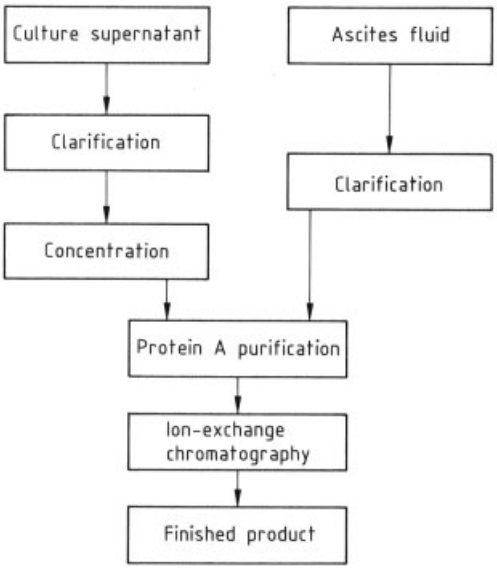

Figure 5. Flow sheet of monoclonal antibody production

The most important purification methods will only be summarized, more details can be found in specialized references [40]. A typical flow sheet is shown in Figure 5. In most cases the first step in antibody purification is some form of *clarification* to eliminate macroimpurities (lipids in ascites and cell debris from culture supernatants). This is generally achieved by either filtration and/or centrifugation (see also → 11. Biochemical Separations, Chap. 4.1.→ 11. Biochemical Separations, Chap. 4.2.). Culture supernatants are then usually concentrated by ultrafiltration, or (for smaller scale production) by protein precipitation.

Nearly all published antibody purification processes employ a chromatographic step. A more detailed description of industrial-scale chromatography is given elsewhere (→ Process-Scale Chromatography). One of the most common methods is *ion-exchange chromatography* because cellulose-based ion exchangers are economical and their use in protein purification is well documented. Each monoclonal antibody has an isoelectric point in the pH range 4.7 – 8.3 and this dictates the operating conditions and the particular matrix for optimal purification.

Since most serum proteins are acidic, ion exchange is particularly favorable for separating the basic antibodies. Ion exchange has been used both as a single purification step to achieve relatively pure antibodies used for diagnostic or immunopurification purposes, and as a final step to clean affinity-purified antibody for therapeutic purposes.

Purification of monoclonal antibody by *affinity chromatography* using, for example, Protein A Sepharose (Pharmacia) is another very popular method. This offers the advantages of giving a high degree of purity and high yield in a single step, even from very dilute supernatants. Not all antibody subclasses and species bind equally well to protein A. However, since the interaction of antibody with protein A is primarily hydrophobic, the introduction of a buffer with a high salt concentration enhances binding considerably and effective purification processes for most monoclonal antibodies can be designed [41]. Affinity chromatography is best used as the first chromatographic purification step due to its very high purification power. Although the matrix is very expensive, the columns can be used

for over 100 cycles. Large quantities of monoclonal antibody are commonly purified by using several purification cycles for each batch. Thus, with a 500-mL protein A column of 4 g capacity, up to 400 g of antibody can be purified in 100 cycles. Automation reduces costs and allows fine tuning of the purification process.

Affinity chromatography matrices are not easy to sterilize. Although Protein A Sepharose columns can be sterilized with a variety of agents (e.g., guanidine hydrochloride, hibitane and benzyl alcohol, propylene oxides) the purification process is usually designed to minimize the need for repeated sterilization. All solutions are sterilized before they come into contact with the column.

Although very high purity may be obtained in a single step by affinity chromatography, antibody for therapeutic use generally requires further purification to reduce the levels of contaminants to acceptable limits. This is often achieved by ion-exchange chromatography to bind the impurities. Because of high concentration and relatively small volumes involved, better antibody yields are obtained than if ion exchange is used at the outset.

6. Monoclonal Antibody Patents

The opportunity to obtain patent protection relating to hybridoma – monoclonal antibody technology was missed when the original publication was allowed to proceed without a patent application being filed [3]. With the benefit of hindsight, it is likely that a generic technology patent position could have been established with patent terms extending into the mid-1990s.

Subsequently, attempts were made to establish broad monoclonal antibody patents directed to classes of antigens, e.g., the Koprowski – Wistar patent applications relating to monoclonal antibodies to viral antigens and tumor-associated antigen [42]. Patents were granted on these filings in the United States, although the corresponding patents were rejected by the United Kingdom Patent Office. Broad patent applications were also filed by Hybritech directed to the use of monoclonal antibodies in immunoassays [43]. Again, patents have been granted in the United States and have now been upheld by the United States Courts in patent infringement actions. The corresponding European Hybritech patent would, however, probably be held invalid if challenged. The difference reflects the United States "First to Invent" and European "First to File" patent law.

To obtain patent protection, it is necessary to have identified a new, nonobvious invention over the "State of the Art" at the initial filing date. Furthermore once monoclonal antibody technology had been established, its application to a particular known antigen is unlikely to give rise to a patentable invention. Thus, following the early broad patents of Koprowski and Hybritech, the scope for other broad monoclonal patents was somewhat limited. Early exceptions, however, were the patents relating to T-cell markers (i.e., Ortho OKT series patents [44]) where the application of hybridoma – monoclonal antibody technology resulted in the identification and dissection of a new class of antigen, and the Secher and Burke patent where special problems had been encountered in obtaining monoclonal antibody to interferon [45].

Following on from these precedents, monoclonal patent opportunities are likely to arise in three main circumstances:

1) The identification of subsets of monoclonal antibodies with special benefits over other monoclonal antibodies for the antigen in question; for instance, as the result of identification of a critical antigenic epitope.
2) The identification of an individual hybridoma cell line. Although specific cell lines have been successfully patented in this way, it is generally thought that the protection afforded by such a patent can be more effectively achieved by controlling the distribution of the cell line.
3) The development of new, useful technologies for preparation of monoclonal antibodies. An exciting opportunity would be a satisfactory solution to the problem of deriving human monoclonal antibodies.

7. Uses

Explosive growth in the annual number of new scientific publications relating to monoclonal

antibodies (Fig. 6) reflects the enormous diversity of the uses of monoclonal antibodies. Extensive reviews can be found in [40], [46].

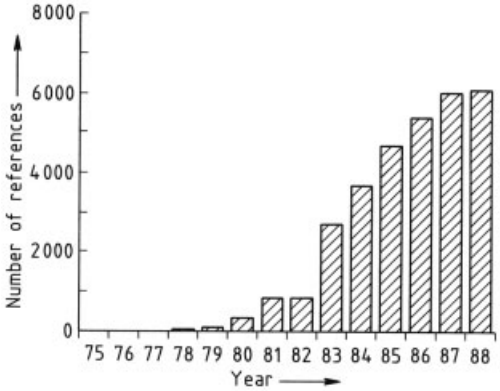

Figure 6. Number of scientific publications relating to monoclonal antibodies from 1975 to 1988 (Medline scientific references data bank)

In this short section a few general statements will be made, followed by a highly selected but illustrative list of applications.

Research Applications. The opportunity that monoclonal antibodies offers to dissect complex antigenic mixtures (Chap. 2) is the key to the dramatic impact that the technology has had on biological research. Immunology, neurobiology, and embryology for example have all witnessed great leaps forward in the past decade as a result of the application of hybridoma technology to describe the complex structure of cell surfaces. Disciplines less familiar with immunology (e.g., botany and chemistry) were slower to realize the full potential of monoclonal antibodies, but are now catching up rapidly.

In immunoassays, such as that used to measure thyroid hormones, monoclonal antibodies have almost completely replaced conventional antisera. Furthermore, because of their availability in large quantities and defined specificity and affinity, monoclonal antibodies have allowed the extension of immunoassay techniques into areas that were previously impractical. For example, assays using labeled antibodies (immunometric assay) are overtaking assays in which the antigen is labeled (radioimmunoassay).

The use of antibodies immobilized onto chromatographic matrices for immunopurification had previously been described using polyclonal antibodies but only with monoclonal antibodies has this technique become of general practical use, e.g., for the purification of interferon [19]. The exquisite specificity of monoclonal antibodies has also had an impact on the reliability and quality of immunohistochemistry.

Monoclonal antibodies are now also used as pharmacological tools to neutralize biologically active molecules or to interfere with specific ligand – receptor interactions.

Clinical Practice. The first important clinical use of monoclonal antibody was for the typing of human blood.

Diagnostics (→Diagnostic Reagents, Chap. 3.→Diagnostic Reagents, Chap. 4. Process-Scale Chromatography,). Monoclonal antibodies have allowed the development and commercial introduction of prepacked "kits" for the measurement of a variety of clinically relevant molecules. The introduction of diagnostic methods relying on enzyme rather than radioisotope labels has led to a rapid expansion of immunoassays in clinical practice.

Within clinical laboratories, the tendency towards full automation has been greatly accelerated by the introduction of monoclonal antibodies. Of equal or even greater importance is the development of simple kits that can be used in the doctor's office or even at home. The best example of this trend is the phenomenal growth in the market for home-based pregnancy detection and fertility testing. Another growth area is the development of tests for infectious disease agents. Here the specificity of monoclonal antibodies has been of paramount importance in developing tests of high diagnostic efficiency.

Rapid growth has also been seen in oncology, where tests for cancer markers, which are soluble tumor-related substances in the blood, can be used to monitor the progress of tumor therapy. Although not yet widely available, a different diagnostic approach for cancer is tumor imaging. This consists of injecting the patient with tumor-specific monoclonal antibody labeled with a γ-ray emitting radioisotope. Tumors can then be detected by γ-ray cameras after the labeled antibody has targeted to the cancer cells (→Diagnostic Reagents). This methodology may soon be extended to the detection of other abnormalities, such as atherosclerotic plaques, blood clots, and sites of infection.

Therapy. Compared to the success of diagnostic monoclonal antibodies the development of therapy based on monoclonals has been disappointing.

Proposed therapeutic uses of monoclonal antibodies range from the replacement of conventional antisera (e.g., in passive immunization against rabies) to more speculative novel applications (e.g., in the control of autoimmune disease; see also →Immunotherapy and Vaccines, Chap. 6.). In practice only a single monoclonal antibody for therapy has received a product licence in the United States (Orthoclon, Ortho Pharmaceutical Corporation). Orthoclon is an antibody to T-lymphocytes that is used to suppress patients' immune response following kidney transplantation, thus reducing the incidence of organ rejection. Antibodies are in clinical trials in three main areas:

1) The treatment of cancer using antibodies to target radioisotopes.
2) The use of monoclonal antibodies linked to cytotoxic drugs or biological toxins such as ricin for "purging" bone marrow before transplantation to remove leukemic cells or T-lymphocytes that are responsible for Graft versus Host Disease.
3) Neutralization of noxious substances such as rabies virus, bacterial endotoxins, or Tumor Necrosis Factor.

Four major development areas may bring about more widespread introduction of therapeutic monoclonal antibodies:

1) The problem of making human monoclonal antibodies will be solved by the approaches described in Chapter 3. This will allow repeated administration of monoclonal antibodies without the adverse reactions currently experienced.
2) The limited access of monoclonal antibodies to tumors and the prolonged residence time in the blood will be improved by genetically engineering antibody fragments.
3) These "cut-down" antibodies may be further modified by replacing the non-antigen-binding parts of the antibody molecule with other biologically active proteins or protein fragments.
4) Improvements in expression systems and large-scale production and purification of monoclonal antibodies will drastically reduce their manufacturing cost, particularly if the production of monoclonal antibody fragments in microorganisms can become generally applicable [47], [48].

8. References

1. D. S. Secher, "Structure of immunoglobulins," in E. S. Lennox (ed.): *Defense and Recognition,* 2nd ed., University Park Press, Baltimore 1979, pp. 1 – 48.
2. G. Galfre, G. W. Butcher in T. L. Wang (ed.): *Immunology in Plant Science,* University Press, Cambridge 1986, p. 5.
3. G. Kohler, C. Milstein, "Continuous cultures of fused cells secreting antibody of predefined specificity," *Nature (London)* **256** (1975) 495 – 497.
4. F. Sanchez-Madrid, P. Szklut, T. A. Springer, "Stable hamster – mouse hybridoma producing IgG and IgM hamster monoclonal antibodies of defined specificity," *J. Immunol.* **130** (1983) 309 – 312.
5. T. J. G. Raybould, M. Takamashi, "Production of stable rabbit – mouse hybridomas that secrete rabbit monoclonal antibody of defined specificity," *Science (Washington D.C.)* **240** (1988) 1788 – 1790.
6. D. M. Strickland, S. A. Saeed, M. L. Casey, M. D. Mitchell, "Bovine mouse hybridomas that secrete bovine immunoglobulin Gl.," *Science (Washington D.C.)* **220** (1982) 522 – 524.
7. G. Galfre, C. Milstein, "Preparation of monoclonal antibodies: strategies and procedures," *Methods Enzymol.* **73** (1981), part B, "Immunochemical techniques".
8. J. F. Kearney, A. Radbruch, B. Liesegang, K. Rajewsky, "A new mouse myeloma cell line that has lost immunoglobulin expression but permits the construction of antibody-secreting hybrid cell lines," *J. Immunol.* **123** (1979) 1548 – 1550.
9. G. Köhler, S. C. Howe, C. Milstein, "Fusion between Ig-secreting and non-secreting myeloma cell lines," *Eur. J. Immunol.* **6** (1976) 292 – 295.
10. G. Galfre, C. Milstein, "Preparation of monoclonal antibodies: strategies and procedures," *Methods Enzymol.* **73** (1981) part B, "Immunochemical techniques" 3 – 46.

11. M. Shulman, C. D. Nilole, G. Köhler, "A better cell line for making hybridomas secreting specific antibodies," *Nature (London)* **276** (1978) 269 – 270.
12. M. L. Gefter, D. H. Margulies, M. D. Scharff, "A simple method for polyethylene glycol-promoted hybridisation of mouse myeloma cells," *Somatic Cell Genet.* **3** (1977) 231 – 236.
13. J. Sharon, S. L. Harrison, E. A. Kabat, "Formation of hybridoma clones in soft agarose: effect of pH and of medium," *Somatic Cell Genet.* **6** (1980) 435 – 441.
14. I. S. Trowbridge, "Interspecies spleen – myeloma hybrid producing monoclonal antibodies against mouse lymphocyte surface glycoprotein, T 200," *J. Exp. Med.* **148** (1978) 313 – 323.
15. H. Bazin, "Production of rat monoclonal antibodies with the lou rat non secreting IR 983F myeloma cell line," C. I. Hybridoma Technology, 615 – 618, reprinted from: Protides of the Biological Fluids 29th Colloqium 1981-Ed-H Peeters-Pergamon Press, Oxford-New York 1982.
16. G. Galfre, C. Milstein, B. Wright, "Rat x rat hybrid myelomas and a monoclonal anti-Fd portion of mouse IgG," *Nature (London)* **277** (1979) 131 –133.
17. J. V. Kilmartin, B. Wright, C. Milstein, "Rat monoclonal antitubulin antibodies derived by using a new non-secreting rat cell line," *J. Cell. Biol.* **93** (1982) 576 – 582.
18. V. Zimmermann, "Electrofusion of cells: principles and industrial potential," *Trends Biotechnol.* **1** (1983) 149 – 155.
19. D. S. Secher, D. C. Burke, "A monoclonal antibody for large-scale purification of human leucocyte interferon," *Nature (London)* **285** (1980) 446 – 450.
20. D. M. Weir: *Handbook of Experimental Immunology,* 4th ed., Blackwell Scientific Publications, Oxford 1986.
21. C. Milstein, G. Galfre, D. S. Secher, T. Springer, "Monoclonal antibodies and cell surface antigens," *Hum. Genet.: Possibilities Realities,* Excerpta Medica, June 1979.
22. W. Knapp et al., "Towards a better definition of human leucocyte surface molecules," *Immunol. Today* **10** (1989) no. 8.
23. C. A. K. Borrebaeck, L. Danielsson, S. A. Moller, "Human monoclonal antibodies produced by primary in vitro immunisation of peripheral blood lymphocytes," *Proc. Nat. Acad. Sci. U.S.A.* **85** (1988) 3995 –3999.
24. J. M. McCune et al., "The SCID-hu mouse: murine model for the analysis of human hematolymphoid differentiation and function," *Science (Washington D.C.)* **241** (1988) 1632 – 1639. D. E. Mosier, R. J. Gulizia, S. M. Baird, D. B. Wilson, "Transfer of a functional human immune system to mice with severe combine immunodeficiency," *Nature (London)* **335** (1988) 256 – 259.
25. U. Storb et al., "Expression of immunoglobulin genes in transgenic mice and transfected cells," *Ann. N.Y. Acad. Sci.* **(5 NM) 546** (1988) 51 – 56.
26. K. James, G. T. Bell, "Human monoclonal antibody production. Current status and future prospects," *J. Immunol. Methods* **100** (1987) 5 – 40.
27. D. Colcher et al., "Characterization and biodistribution of recombinant and recombinant/chimeric constructs of monoclonal antibody B 72.3," *Cancer Res.* **49** (1989) 1738 – 1745.
28. L. Reichmann, M. Clark, H. Waldmann, G. Winter, "Reshaping human antibodies for therapy," *Nature (London)* **332** (1988) 323 – 327. M. Verhoeyen, L. Riechmann, "Engineering of antibodies," *BioEssays* **8** (1988) no. 2, 74 – 78.
29. J. Feder, W. R. Talbert (eds.): *Large-Scale Mammalian Cell Culture,* Academic Press, New York 1985.
30. K. Pullen et al., *Devel. Biol. Stand.* **70** (1988) 175 – 177.
31. M. Rhodes, J. R. Birch, "Large-scale production of proteins from mammalian cells," *Bio/Technology* **6** (1988) 518.
32. J. R. Birch et al., "The industrial production of monoclonal antibodies in cell culture," in J. M. Cardoso Duarte, L. J. Archeg, A. T. Bull, G. Holt (eds.): *Prospective in Biotechnology,* Plenum Publishing, New York 1987, 97 – 109.
33. A. L. Van Wezel, "Growth of cell strains and primary cells on micro-carriers in homogeneous culture," *Nature (London)* **216** (1967) 64 – 65.
34. C. R. Keese, I. Giaever, "Cell growth on liquid interfaces: Role of surface active compounds," *Proc. Nat. Acad. Sci. U.S.A.* **80** (1983) 5622 – 5626.
35. F. Lim, A. Sun, "Microencapsulated islets as bioartificial endocrine pancreas," *Science (Washington D.C.)* **210** (1980) 908 – 910.
36. R. G. Rupp, "Use of cellular microencapsulation in large-scale production

of monoclonal antibodies," in J. Feder, W. R. Talbert (eds.): *Large-Scale Mammalian Cell Culture,* Academic Press, New York 1985.
37. P. Calabresi et al., "Monoclonal antibody projection in artificial capillary cultures," *Proc. Am. Assoc. Cancer Res.* **22** (1981) 362 – 365.
38. G. Klement, N. Scheiver, H. N. D. Katinger, "Construction of a large-scale membrane reactor system with different compartments for cells, medium and product," *Devel. Biol. Stand.* **66** (1987) 221 – 226.
39. B. K. Lydersen et al., "The use of a ceramic matrix in a large scale cell culture system," in J. Feder, W. R. Talbert (eds.): *Large Scale Mammalian Cell Culture,* Academic Press, New York 1985, pp. 39 – 58.
40. A. C. Kenney, "Large-scale production of monoclonal antibodies," *Monoclonal Antibodies: Production and Application,* Alan R. Liss Inc., New York 1989, pp. 143 – 160.
41. H. Juarez-Salinas, G. S. Ott, US 4 704 366, 1987.
42. H. Koprowski, C. M. Croce, US 4 172 124, 1979.
43. G. S. Davis, H. E. Greene, UK 2 086 041 A, 1981;US 4 376 110, 1983.
44. Ortho Diagnostic, US 4 361 549, US 4 361 550, 4 364 932, 4 364 933, 4 364 934, 4 364 935, 4 364 936, 1982 (P. C. Kung, G. Goldstein).
45. D. S. Secher, US 4 423 147, 1983; US 4 514 507, 1985.
46. J. W. Golding: *Monoclonal Antibodies: Principles and Practice. Production and Application of Monoclonal Antibodies in Cell Biology, Biochemistry and Immunology,* Academic Press, New York 1987.
47. E. S. Ward et al., "Binding activities of a repertoire of single immunoglobulin variable domains secreted from Escherichia coli," *Nature (London)* **341** (1989) 544.
48. W. D. Huse et al., "Generation of a large combinatorial library of the immunoglobulin repertoire in phage," *Science (Washington D.C.),* in press.

Monosodium Glutamate

TETSUYA KAWAKITA, Central Research Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan
CHIAKI SANO, Central Research Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan
SHIGERU SHIOYA, Central Research Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan
MASAHIRO TAKEHARA, Central Research Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan
SHIZUKO YAMAGUCHI, Central Research Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan

1. Introduction 1
2. Properties 1
3. Production 2
4. Uses 3
5. Toxicology 6
6. Economic Aspects 7
7. References 7

1. Introduction

Monosodium glutamate (MSG) is a monosodium salt of L-glutamic acid; it is used in huge quantities as a flavor enhancer throughout the world. L-Glutamic acid is one of the naturally occurring amino acids and occurs both in bound form as a constituent of proteins together with other amino acids, and also in the free form.

L-Glutamic acid was first isolated in 1866 by RITTHAUSEN [6] from the acid hydrolysate of wheat gluten, and its chemical structure was identified synthetically by WOLFF in 1890 [7]. In 1908, KIKUNAE IKEDA discovered the ability of the monosodium salt of glutamic acid to enhance the flavor of foods [8].

A kelp-like seaweed harvested around Japan, known as Konbu (*Laminaria Japonica*), has been used for centuries by Japanese cooks to improve the flavor of soups and certain other foods. IKEDA was curious why the taste of soup was enriched when its stock was prepared from konbu. He isolated 30 g of L-glutamic acid by extracting 40 kg of the kelp with hot water and discovered that glutamate was the substance which made the difference. Immediately, he filed a patent for the production of MSG as a new seasoning [9]. IKEDA developed a process to extract MSG from wheat flour. In 1909, MSG was first produced commercially under the trade name Ajinomoto.

2. Properties

Monosodium L-glutamate monohydrate [*142-47-2*], $C_5H_8NO_4Na \cdot H_2O$, M_r 187.13, crystallizes from aqueous solution at room temperature as rhombic prisms. Its crystal structure, determined by X-ray analysis, indicates that the sodium ions are coordinated octahedrally by four (3 α and 1 γ) carboxyl oxygen atoms and two water molecules [10].

The solubility of MSG is given by the following equation:

$$S = 39.06 + 0.108t + 0.0013t^2$$

where S is the concentration of MSG in wt % in a saturated solution at t °C [11]. Below −0.8 °C, MSG crystallizes as the pentahydrate [11]. The pH of a 3 wt % aqueous solution of MSG at 25 °C is 7.0.

Crystalline MSG is less hygroscopic than sodium chloride. The critical humidities of MSG are 96.0, 94.8 and 90.0 % at 20, 30 and 50 °C, respectively [12].

MSG loses its water of crystallization at 120 °C; intramolecular dehydration occurs at 155 °C. The decomposition temperature is 225 °C.

The α-carbon of glutamic acid is chiral. The specific optical rotation $[\alpha]_D^{20}$ of a solution of 10 g MSG in 100 ml of 2 N HCl is +25.16°. Besides the naturally occurring dextrorotatory L-glutamic acid, D-glutamic acid [*6893-26-1*] and the racemic mixture, DL-glutamic acid [*617-65-2*], are also known. D-Glutamic acid is

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_711

only found as a constituent of the peptideglycans of cell walls and of bacitracins (polypeptide antibiotics) produced by some microorgansims (e.g., *Bacillus subtilis*). The unique flavor enhancing characteristics are possessed only by the L-form.

Monosodium glutamate does not racemize even at 100 °C in neutral solution. At 100 °C, a pH above 8.5 results in thermal racemization. Racemization is also brought about by a microbial racemase. The racemase that is specific for glutamic acid is found in *Lactobacillus fermenti*.

The α-amino and γ-carboxyl groups of L-glutamic acid form an intramolecular peptide bond upon dehydration.

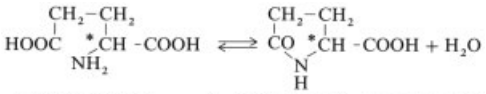

The equilibrium of this reaction favors the formation of 2-pyrrolidone-5-carboxylic acid [*98-79-3*] (PCA) when the pH is weakly acidic or weakly basic. At neutral pH, the rate of dehydration is at a minimum. Under strongly acidic or alkaline conditions, the equilibrium favors hydration of PCA [13]. Under normal cooking conditions, neither racemization nor dehydration occurs.

The threshold value (minimum perceptible concentration) of MSG is 0.03 %, which is smaller than that of sodium chloride (0.2 %) or sucrose (0.5 %). The flavor enhancing activity of MSG is synergistically increased in the presence of 5′-inosinate [*131-99-7*] and 5′-guanylate [*85-32-5*].

L-Glutamic acid is classified as a nonessential amino acid. It is a key substance in the nitrogen metabolism of amino acids. Inorganic nitrogen is incorporated by reductive amination of 2-oxoglutarate [*328-50-7*] to L-glutamate, catalyzed by glutamate dehydrogenase. There are many kinds of aminotransferases in various organisms that transfer an amino group from an amino acid to 2-oxo-acid, and many of them use the glutamate/ 2-oxoglutarate pair as the amino donor/acceptor. L-Glutamate is also the direct precursor of L-glutamine [*56-85-9*], the neuro transmitter γ-aminobutylic acid (GABA) [*56-12-2*], γ-carboxyglutamate [*53861-57-7*] in calcium binding proteins, and the tripeptide glutathion [*70-18-8*] [14].

3. Production

Classical Processes. From 1909, when industrial production of MSG started, until 1965, when production by the extraction process ceased, wheat gluten separated from wheat flour was the main raw material. It was used because of its high content of L-glutamic acid (ca. 25 wt %). After 1934, defatted soybean flake was also employed as raw material, since it did not produce starch as a byproduct.

The raw material is hydrolyzed by heating with hydrochloric acid. After concentration under reduced pressure and addition of concentrated hydrochloric acid, the hydrolysate is cooled to crystallize L-glutamic acid hydrochloride [*138-15-8*]. L-Glutamic acid can easily be separated from other amino acids in the form of its hydrochloride because of its very low solubility in concentrated hydrochloric acid. The hydrochloride is collected by filtration and then dissolved in warm water and the solution filtered to remove insoluble humic materials, which are mainly formed by the reaction of amino acids with carbohydrates. The pH of the filtrate is adjusted to 3.2, the isoelectric point of L-glutamic acid, with sodium hydroxide or ammonia to precipitate crystals of L-glutamic acid. The yield of L-glutamic acid is high because of its low solubility in water (0.864 g per 100 mL at 25 °C).

The crude L-glutamic acid crystals are suspended in water and neutralized with sodium hydroxide. The solution is decolorized with activated charcoal to give a clear solution and concentrated to crystallize MSG crystals. The crystals of MSG are then separated from the mother liquor by centrifugation and dried for packaging.

In the United States and some European countries, PCA, contained in byproducts of the sugar beet industry, e.g., Steffen's waste or vinasse, has been used as raw material for MSG. The PCA is hydrolyzed at weakly alkaline pH and moderate temperature (e.g., pH 10.5 – 11.5 at 85 °C for 2 h) [15] to avoid racemization. The pH of the hydrolysate is adjusted to 3.2 with a mineral acid to precipitate crystals of L-glutamic acid. The L-glutamic acid is transformed to MSG as described above.

Fermentation Processes. In the early 1950s, DAGLEY et al. reported the excretion of small amounts of amino acids in the culture medium of *Escherichia coli*. They found that addition of excess ammonium salts resulted in increased amino acid production [16]. In 1957, KINOSHITA and co-workers discovered bacteria that accumulate large quantities of L-glutamic acid [17]. The method of KINOSHITA et al. gave yields of L-glutamic acid of about 30 % using *Micrococcus glutamicus* (later renamed *Corynebacterium glutamicum*), which was immediately applied for the industrial production.

Bacteria that produce L-glutamic acid are widely distributed in nature, as shown in Table 1 [18]. Most of these bacteria are Gram-positive, non-spore-forming, non-motile, and require biotin for growth.

Table 1. Microbial strains producing l-glutamic acid

Genus	Species
Corynebacterium	*C. glutamicum, C. lilium, C. callunae, C. herculis*
Brevibacterium	*B. divaricatum, B. aminogenes, B. flavum, B. lactofermentum, B. saccharolyticum, B. roseum, B. immariophilum, B. alanicum, B. ammoniagenes, B. thiogenitalis*
Microbacterium	*M. salicinovolum, M. ammoniaphilum, M. flavum var. glutamicum*
Arthrobacter	*A. glubiformis, A. aminofaciens*

For industrial production of MSG, molasses and starch hydrolysate are generally used at present as raw materials. The accumulation of L-glutamic acid is governed not by its biosynthesis in the bacterial cells but by its secretion. Biotin is one of the vitamins essential for cell growth, but a biotin concentration in the culture medium sufficient for cell growth makes the cell membranes impermeable for L-glutamate and results in poor accumulation. The critical biotin content of the cells for L-glutamic acid production is 0.5 µg/g dry cells. Thus the accumulation of L-glutamic acid is at a maximum when the biotin concentration is suboptimal for maximum growth. Therefore, raw materials that are rich in biotin, such as beet molasses and cane molasses, could not be used until the discovery of the biotin-inhibiting effects of C_{16}–C_{18} saturated polyoxyethylene fatty acid esters. Limiting the amount of biotin is believed to cause incomplete biosynthesis of oleic acid, which results in a decrease of the phospholipids in the cell membrane and makes the membrane permeable [19].

Ammonium salts and urea are used as nitrogen sources for both microbial growth and product formation. The culture medium becomes acidic because of assimilation of ammonium ions and formation of L-glutamic acid. Gaseous ammonia is used advantageously to maintain neutral pH and to avoid dilution of the broth, because it contains neither hydroxyl ions nor water.

As the biosynthesis of L-glutamic acid is an aerobic process, the pressure of dissolved oxygen must strictly be kept above 1 kPa by aeration and agitation in the fermenter.

Genetically, many attempts have been made to improve the productivity of the process. L-Glutamic acid is synthesized biochemically through the glyoxylate cycle as an oxaloacetate generating system without carbon dioxide fixation, and through phosphoenolpyruvate [*138-08-9*] to form oxaloacetate [*328-42-7*] with carbon dioxide fixation (Fig. 1). The use of several mutants with high phosphoenolpyruvate carboxylase activity [20] or low isocitrate lyase activity [21] led to improved productivity of L-glutamic acid and increased the extent of carbon dioxide fixation. Application of genetic engineering techniques for further improvement is also in progress [22–25].

Figure 2 shows a typical large scale fermentation system equipped with automatic control which can be performed by computer aided systems [26].

Progress in fermentation technology has made it possible to raise the accumulation and the yield of L-glutamic acid above 100 g/L and 60 %, respectively. The pH of the fermentation broth is adjusted to 3.2 to recover L-glutamic acid crystals, which are then converted to MSG by the same method as in the extraction process.

4. Uses

Seasoning. MSG is widely used as a long-established seasoning or flavor enhancer to improve the palatability of foods. The effect of MSG is due to its characteristic taste, umami. Developments in food sciences, including electrophysiology, biochemistry, psychophysics, food chemistry, and nutrition, have provided evidence that umami should be re-

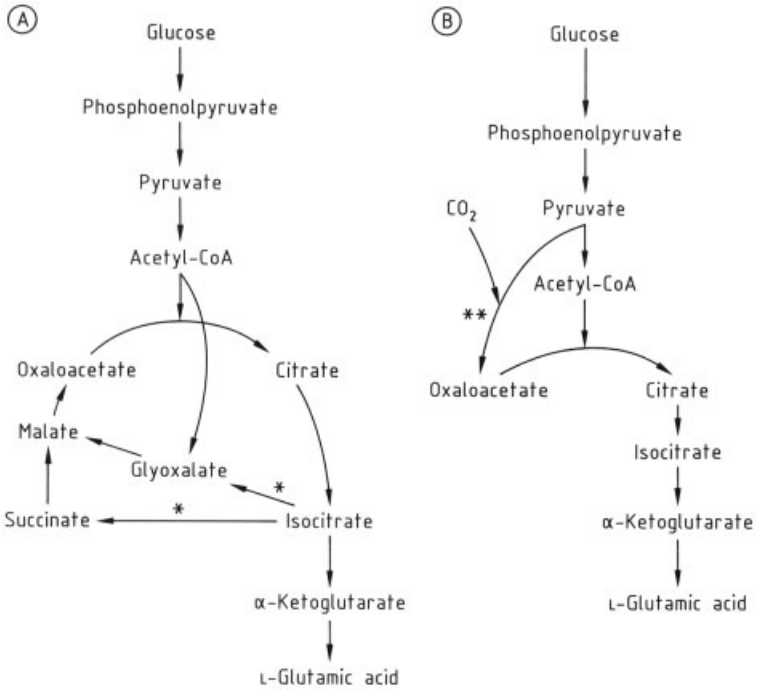

Figure 1. (A) Biosynthetic pathway of L-glutamic acid through the glyoxylate cycle as an oxaloacetate generating system without carbon dioxide fixation; (B) Biosynthetic pathway of L-glutamic acid through phosphoenolpyruvate as an oxaloacetate generating system with carbon dioxide fixation.
∗ Isocitrate lyase. ∗∗ Phosphoenolpyruvate carboxylase.

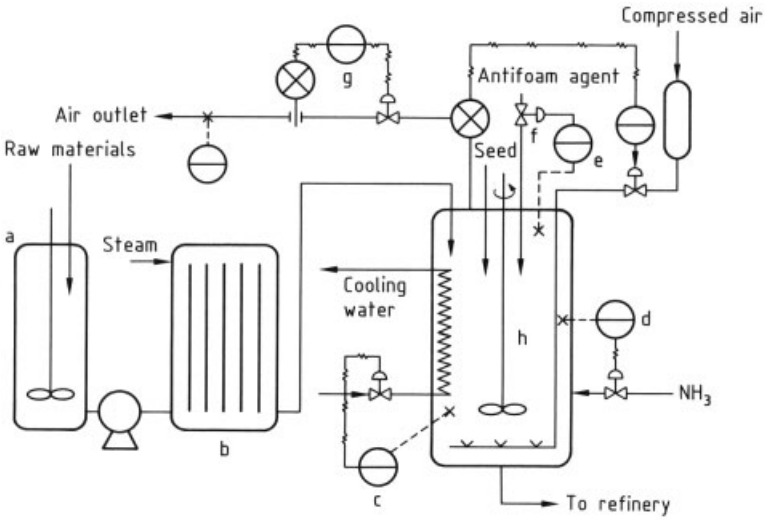

Figure 2. Automatically controlled fermentation process for the production of MSG
a) Medium preparation tank; b) Continuous sterilizer; c) Temperature recording and control; d) pH Recording and control; e) Level control; f) Pressure control; g) Flow rate recording and control; h) Fermentation tank

garded as a basic taste independent from the traditional four basic tastes: sweet, sour, salty, and bitter [27]. Umami cannot be composed from the classical four basic tastes and is essential to describe the human taste modality in general. Glutamates are ubiquitous in foods such as human milk, tomatoes, cheese, and seaweed. Other umami substances, i.e., 5′-ribonucleotides such as inosinate and guanylate are found in meats, fish, vegetables, and mushrooms. When glutamates and nucleotides coexist, a very strong synergistic effect occurs and umami is dramatically enhanced [28]. When foods contain umami substances under the taste threshold, just a small amount of additional umami substance makes a remarkable contribution to the flavor.

The taste threshold of MSG is around 0.03 % in aqueous solution. The umami intensity increases linearly according to the logarithm of the concentration of MSG. MSG does not enhance the taste of other basic taste substances nor is it enhanced by them. The synergistic effect of MSG with 5′-ribonucleotides is expressed as $y = u + \gamma uv$ where u and v are the respective concentrations of MSG and 5′-ribonucleotides, y is the equivalent concentration of MSG alone, and γ is a constant: 1200 for disodium 5′-inosinate (IMP) and 2800 for disodium 5′-guanylate (GMP) [29]. In some commercial umami seasonings, IMP or GMP or both are added to take advantage of the synergism with MSG.

Monosodium glutamate works well in a wide variety of foods, notably in meats, seafoods, vegetables, soups, stews, and chowders. Alcoholic beverages, fruits, and confectionery are not considered to be improved by addition of MSG. The effect of MSG on smell (or aroma) is negligible. When MSG or umami taste is added to an appropriate food at an appropriate level, certain flavor characteristics of food, such as body, mouthfulness, mildness, and continuity of flavor are increased. Thus the flavor appeal of foods is improved [30].

Although MSG increases the palatability of food, it is not always palatable by itself. When MSG is tasted in crystal form or in simple solutions, it does not cause a pleasant sensation. It gives a pleasant sensation only in flavored solutions or in actual foods [31]. Another important property of MSG is the concentration dependence of its pleasantness rating. An excess of MSG causes an unpleasant sensation, and thus the amount of MSG that is added to foods should be limited [31]. The desirable usage level of MSG depends on a number of properties of the food, including the contents of original umami substances, flavor intensity, flavor quality, density, and others. Results of extensive taste studies indicate that the optimum MSG level in food ranges from 0.1 to 0.8 wt %, varying from food to food.

In most cases, MSG is used in combination with salt. A relation between the levels of MSG and salt and palatability of foods has been clarified. The lower the salt concentration, the higher the MSG concentration required to maintain the palatability and vice versa. When umami substances are added to meals, the amount of salt can be reduced; thus the total sodium intake can be reduced without a decrease in the palatability or degree of satisfaction.

Umami is considered to be a universal taste sensation for mankind. An ethnic comparison of taste sensitivity has shown that there is no difference in the thresholds of MSG and IMP between Japanese and Euro-American Caucasians [32]. Even in the very first hours of extrauterine life, the human neonate responds to MSG with facial expressions. When a vegetable soup containing MSG was fed to neonates, they showed more pleasant and satisfied facial and oral expressions than for soup without MSG [33].

Industrial Uses. Depending on the pH, glutamic acid can be used to neutralize acidic or basic compounds. For example, arginine glutamate is used as a pharmaceutical and raw material for cosmetics. The reaction of glutamic acid with hydrochloric acid affords glutamic acid hydrochloride, which is used as a gastric acidifier. Glutamic acid's properties as an amphoteric electrolyte are exploited in chelating agents, buffers, and builders for detergents.

The sodium salt of pyroglutamic acid, which is obtained by dehydration of glutamic acid followed by neutralization with sodium hydroxide, is very hygroscopic and is used as a component of a natural moisturizing factor for human skin and as a numectant for cosmetics [34]. A polymer of methyl glutamate is used as a coating agent for synthetic leather [35], [36]. A condensate with formaldehyde is useful as a retarder for plaster [37]. The *N*-acylated compound pro-

duced by reaction with a long-chain fatty acid is an anionic surfactant and is popularly used for skin and hair cleansers due to its mildness and high safety [38]. The dibutylamide of this *N*-acylglutamic acid gelatinizes nonpolar oils and can be used as a recovering agent for spilt marine oil [39].

5. Toxicology

Glutamic acid is one of the most abundant amino acids in the diet and constitutes 10 – 35 % of the amino acids in proteins. Many foods also contain free glutamate, which contributes to their taste. MSG is metabolized in the same way as naturally occurring free glutamate and glutamate derived from the digestion of protein. After oral ingestion, glutamate is absorbed from the intestine by an active transport system. During this process, a large proportion is metabolized to alanine and 2-oxoglutarate, which enters the tricarboxylic acid cycle. Glutamate is further metabolized in the liver to give glucose, lactate, glutamine, and other amino acids. Consequently, blood glutamate levels do not rise significantly unless very large concentrated doses are administered. The blood glutamate level rises transiently when a large concentrated dose is ingested on its own, but the co-ingestion of foods that contain metabolizable carbohydrates increases its metabolism and eliminates or greatly attenuates this rise. Human infants, including premature infants, metabolize glutamate similarly to adults [40], [41].

The acute toxicity of MSG is low; the LD_{50} values, when administered in different ways to different animals, are shown in Table 2. The oral LD_{50} extrapolated to humans would represent a single dose of over 1 kg for a person weighing 70 kg; the oral LD_{50} for sodium chloride in rats is 3.75 g/kg [43].

The concentration of glutamate is higher in the brain than in the blood. In monkeys, guinea pigs, rats, and mice, glutamate levels in the brain remain unchanged after the administration of large oral doses of MSG. When large amounts of MSG are forcibly administered to some infant animals by injection, brain and retina lesions can occur. There are species- and age-dependent differences in sensitivity to these neuronal lesions. In mice, the most sensitive species, the newborn are sensitive but the adults are not. Hamsters and guinea pigs are less sensitive, dogs are not sensitive, and in nonhuman primates, even the injection of MSG into the foetus does not produce neuronal lesions. When MSG is administered orally in the diet or drinking water, neuronal lesions do not occur even in infant mice, except following dehydration by water deprivation.

The main use of MSG is as a food additive, and so its toxicity when used in the diet is the most important aspect of its safety in use. Both short term and chronic toxicity studies on MSG in the diet of several species at doses of up to 4 % (ca. 6 – 8 g per kilogram body weight per day) of daily food intake showed no specific toxic effects, and no evidence for carcinogenicity or mutagenicity. Reproduction and studies of up to three generations have not revealed any adverse effects of dietary MSG ingestion, with fertility, gestation, viability, and lactation indices and the pre- and post-weaning performance of offspring being unaffected. In primates, it has been found that the placenta serves as an effective metabolic barrier to the transfer of glutamate from the mother to the foetus, and in humans the glutamate level in mother's milk is hardly affected by oral ingestion of large doses of MSG [40], [41], [44].

Since 1968, much publicity has been given to a postprandial symptom complex, involving facial pressure, chest pain, and a burning sensation, called "Chinese restaurant syndrome" (CRS). Initial reports claimed that the symptoms occurred after eating at Chinese restaurants and suggested that MSG was the causative agent. However, a large scale epidemiological survey has found that only 0.2 % of the surveyed population could be classified as having experienced the syndrome after ingesting Chinese food, and controlled double-blind studies have failed to show an unequivocal relationship between CRS and the consumption of MSG. The symptoms are subjective and can be experienced after the ingestion of substances not containing MSG, such as orange juice, coffee, and spiced tomato juice. Although the administration of large concentrated doses of MSG in the laboratory can provoke CRS-like symptoms in some people, when self-proclaimed CRS sufferers were given MSG or a placebo in a double-blind test, they failed to respond selectively to MSG [40], [41], [45], [46].

Table 2. LD_{50} (g/kg body weight) of MSG in various species [41], [42]

Species	Route of administration			
	Oral∗	Subcutaneous	Intraperitoneal	Intravenous
Mouse	13.0, 16.2 17.0, 19.2	8.3	6.1	3.5
Rat	16.5, 16.6 19.9	6.0	5.2	3.3
Guinea pig			15.0	
Rabbit	> 2.3∗∗			

∗ Multiple values represent LD_5^0 data obtained by different researchers.
∗∗ Glutamic acid.

When used normally as a food additive, MSG does not show any harmful side effects, and in 1987 the World Health Organization (WHO) and the Food and Agricultural Organization (FAO) of the United Nations established the acceptable daily intake (ADI) for L-glutamic acid and its monosodium, potassium, ammonium, calcium and magnesium salts as being 'not specified'. This means that on the basis of the available data, the total daily intake of the substance, arising from its use at the level necessary to achieve the desired effect and from its acceptable background in food, in the opinion of the joint FAO/WHO expert committee on food additives (JECFA), does not represent a hazard to health, and so it was not deemed necessary to establish a numerical limit [41].

6. Economic Aspects

The annual production of MSG was estimated at 350 000 t worldwide in 1988 [47]. The world capacity for the production of MSG is more than 500 000 t/a [47] including the People's Republic of China and it is still increasing. The international price of MSG is $ 1500 – 1800 per tonne. The demand for MSG is expected to increase in developing countries.

7. References

General References

1. *Kirk Othmer,* 3rd ed., **2,** 410 – 421.
2. F. M. Clydedale: *Encyclopedia of Food Technology,* **vol. 2,** The AVI Publishing Company, Inc., Westport, Connecticut 1974, p. 480, pp. 702 – 704.
3. L. B. Sjostrom, *Hand Book of Food Additives,* CRC Press, Cleveland, Ohio 1973, pp. 513 – 521.
4. L. J. Filer, Jr., S. Garattini, M. R. Kare, W. A. Reynolds, R. J. Wurtman (eds.): *Glutamic Acid: Advances in Biochemistry and Physiology,* Raven Press, New York 1979.
5. G. C. Barrett (ed.): *Chemistry and Biochemistry of the Amino Acids,* Chapman and Hall, London-New York 1985.

Specific References

6. H. Ritthausen, *J. Prakt. Chem.* **99** (1866) no. 6, 454.
7. L. Wolff, *Justus Liebigs Ann. Chem.* **260** (1890) 79.
8. K. Ikeda, *J. Chem. Soc. Tokyo* **30** (1909) 820. *K. Ikeda, Original Communications, Eighth International Congress of Applied Chemistry (Washington and New York)* **18** (1912) 147. A. E. Marshall: "History of Glutamate Manufacture," *Monosodium Glutamate – A Symposium,* March 4, 1948, pp. 4 – 6.
9. K. Ikeda, JP 14 805, 1908. Suzuki & Co., Ltd., GB 9440, 1909 (K. Ikeda, S. Suzuki); *Chem. Abstr.* **5** (1910) 836. Suzuki & Co., Ltd., US 1 015 891, 1912 (K. Ikeda, S. Suzuki); *Chem. Abstr.* **6** (1912) 717.
10. C. Sano, N. Nagashima, T. Kawakita, Y. Iitaka, *Anal. Sci.* **5** (1989) 121 – 122.
11. T. Ogawa, *Kogyo Kagaku Zasshi* **52** (1949) 29.
12. Aijnomoto Co. Inc., *Amino Acids,* Tokyo 1965, pp. 96 – 97.
13. H. W. Wilson, R. K. Cannan, *J. Biol. Chem.* **119** (1937) 309.
14. D. A. Bender: *Amino Acid Metabolism,* 2nd ed., Wiley & Sons, New York 1985, pp. 119 – 143.

15. International Minerals and Chemicals Corp., US 2 373 342, 1945 (C. Royal); *Chem. Abstr.* **39** (1946) 4510.
16. S. Dagley, E. A. Dawes, G. A. Morrison, *Nature (London)* **165** (1950) 437.
17. S. Kinoshita, S. Udaka, M. Shimono, *J. Gen. Appl. Microbiol.* **3** (1957) 193 – 205. US 3 002 889, 1961 (S. Kinoshita et al.). Kyowa Ferment. Ind. Co., Ltd., US 3 003 925, 1961 (S. Kinoshita et al.)
18. K. Yamada, S. Kinoshita, T. Tsunoda, K. Aida: *The Microbial Production of Amino Acids,* Kodansha, Tokyo 1972, p. 5.
19. Y. Hirose, H. Enei, H. Shibai in M. Moo-Young (ed.): *Comprehensive Biotechnology,* **vol. 3,** Pergamon Press, Oxford 1985, pp. 593 – 600.
20. Ajinomoto Co., Inc., JP 5 332 193, 1978 (Y. Yamada, T. Takagi, K. Takinami).
21. Ajinomoto Co., Inc., JP 56-92 795, 1981 (O. Tosaka et al.).
22. H. Kaneko, T. Tanaka, K. Sakaguchi, *Agric. Biol. Chem.* **43** (1979) 867 – 868.
23. H. Momose, S. Miyashiro, M. Oba, *J. Gen. Appl. Microbiol.* **22** (1976) 119 – 129.
24. J. A. Shapiro, *Appl. Environ. Microbiol.* **32** (1976) 179 – 182.
25. Ajinomoto Co., Inc., JP 56-148 295, 1981 (T. Tsuchida, K. Miwa, S. Nakamori, H. Momose).
26. M. Hikuma, T. Yasuda, I. Karube, S. Suzuki, *Ann. N.Y. Acad. Sci.* **369** (1981) 307 – 319.
27. Y. Kawamura, M. R. Kare (eds.): *Umami: A Basic Taste,* Marcel Dekker, New York-Basel 1987.
28. A. Kuninaka, *Nippon Nogei Kagaku Kaishi* **34** (1960) 489 – 492.
29. S. Yamaguchi, *J. Food Sci.* **36** (1971) 846 – 849.
30. S. Yamaguchi, A. Kimizuka in L. J. Filer, Jr., S. Garattini, M. R. Kare, W. A. Reynolds, R. J. Wurtman (eds.): *Glutamic Acid: Advances in Biochemistry and Physiology,* Raven Press, New York 1979, p. 35 – 54.
31. S. Yamaguchi, C. Takahashi, *Agric. Biol. Chem.* **48** (1984) 1077 – 1081.
32. S. Yamaguchi, M. Kimura, R. Ishii, *Proceedings of the 22nd Japanese Symposium on Taste and Smell,* Japanese Association for the Study of Taste and Smell, Gifu 1988, pp. 73 – 76.
33. J. E. Steiner in Y. Kawamura, M. R. Kare (eds.): *Umami: A Basic Taste,* Marcel Dekker, New York-Basel 1987, p. 97 – 123.
34. K. Laden, R. Spitzer, *J. Soc. Cosmet. Chem.* **19** (1967) 351 – 360.
35. Asahi Chemical Industry, JP 68-23 360, 1963 (R. Wakasa et al.).
36. Ajinomoto Co., Inc., JP 74-29 400, 1968 (Y. Takeda, Y. Takagi, S. Mori).
37. Domtar Ltd., US 3 451 832, 1965 (R. A. Kuntze).
38. M. Takehara, R. Yoshida, M. Yoshikawa, *Cosmet. Toiletries* **94** (1979) 31 – 39.
39. M. Honma, T. Ikeda, *Kagaku no Ryouiki* **36** (1982) 697 – 701.
40. L. J. Filer, Jr., S. Garattini, M. R. Kare, W. A. Reynolds, R. J. Wurtman (eds.): *Glutamic Acid: Advances in Biochemistry and Physiology,* Raven Press, New York 1979.
41. Joint FAO/WHO Expert Committee on Food Additives: *Toxicological Evaluation of Certain Food Additives,* World Health Organization, Geneva 1988.
42. H. Moriyuki, M. Ichimura, *Oyo Yakuri* **15** (1978) 433 – 436.
43. Merck & Co., Inc., *Merck Index,* 11th ed., Rahway, N.J. 1989, p. 1359.
44. W. Oleny, *Science* **164** (1969) 719 – 721.
45. H. M. Kwok, *N. Engl. J. Med.* **278** (1968) 796.
46. A. Kenny, *Food Chem. Toxicicol.* **24** (1986) 351 – 354.
47. S. Ishimoda, *The Wines, Spirits & Provisions Monthly Statistics,* Nikkan Keizai Tsushiusha, Tokyo, March 1988, pp. 33 – 42.

Morpholine → Amines, Aliphatic
Mortar → Cement and Concrete

Automotive Fuels

Werner Dabelstein, Shell Global Solutions (Deutschland) GmbH (retired), Hamburg, Germany

Arno Reglitzky, Shell Global Solutions (Deutschland) GmbH (retired), Hamburg, Germany

Andrea Schütze, Shell Global Solutions (Deutsch Land) GmbH, Hamburg, Germany

Klaus Reders, Shell Global Solutions (Deutschland) GmbH (retired), Hamburg, Germany

Related Articles Automobile Exhaust Control, Aviation Turbine Fuels, Heating Oil, Liquefied Petroleum Gas, Methyl Tert-Butyl Ether, Octane Enhancers, and Oil Refining are separate keywords.

1. History . . . 2
1.1. The Spark Ignition (Otto) Engine and Its Fuel . . . 2
1.2. The Diesel Engine and Its Fuel . . . 2
2. Engine Technology . . . 3
2.1. Otto Engines . . . 3
2.2. Diesel Engines . . . 5
3. Fuel Composition and Engine Efficiency . . . 6
3.1. Quality Aspects of Gasoline . . . 6
3.1.1. Octane Quality . . . 6
3.1.2. Volatility . . . 7
3.1.3. Fuel Composition to Reduce Toxicity and Exhaust Emissions . . . 8
3.1.4. Stability, Cleanliness, etc. . . . 9
3.1.5. Performance Additives . . . 9
3.2. Quality Aspects of Diesel Fuels . . . 10
3.2.1. Ignition Quality . . . 10
3.2.2. Density . . . 10
3.2.3. Sulfur Content . . . 10
3.2.4. Cold Flow Properties . . . 10
3.2.5. Lubricity . . . 11
3.2.6. Viscosity . . . 11
3.2.7. Volatility . . . 11
3.2.8. Diesel Fuel Stability, Cleanliness, etc. . . . 11
3.2.9. Diesel Fuel Effects on Exhaust Emissions . . . 12
3.2.10. Performance Additives . . . 13
4. Fuel Components . . . 13
4.1. Gasoline Components . . . 13
4.1.1. Straight-Run Gasoline . . . 14
4.1.2. Thermally Cracked Gasoline . . . 14
4.1.3. Catalytically Cracked Gasoline . . . 15
4.1.4. Catalytic Reformate (Platformate) . . . 15
4.1.5. Isomerate . . . 15
4.1.6. Alkylate . . . 15
4.1.7. Polymer Gasoline . . . 15
4.1.8. Oxygenates . . . 15
4.2. Diesel Fuel Components . . . 16
4.2.1. Straight-Run Middle Distillate . . . 16
4.2.2. Thermally Cracked Gas Oil . . . 17
4.2.3. Catalytically Cracked Gas Oil . . . 17
4.2.4. Hydrocracked Gas Oil . . . 17
4.2.5. Kerosene . . . 17
4.2.6. Synthetic Diesel Fuel . . . 17
5. Fuel Additives . . . 18
5.1. Gasoline Additives . . . 18
5.1.1. Corrosion Inhibitors . . . 18
5.1.2. Detergents . . . 20
5.1.3. Antioxidants . . . 21
5.1.4. Metal Deactivators . . . 21
5.1.5. Anti-Icing Additives . . . 22
5.1.6. Additives for Combating Combustion Chamber Deposits . . . 22
5.1.7. Valve Seat Recession Protection Additives . . . 23
5.1.8. Antiknock Agents . . . 23
5.1.9. Dehazers and Antistatic Additives . . . 23
5.2. Additives for Diesel Fuel . . . 23
5.2.1. Ignition Improvers (Cetane Improvers) . . . 23
5.2.2. Detergent Additives . . . 24
5.2.3. Cold Flow Additives . . . 24
5.2.4. Lubricity Additives . . . 26
5.2.5. Antifoam Additives . . . 27
5.2.6. Additives for Increasing Storage Stability – Antioxidants . . . 27
5.2.7. Dehazers . . . 27
5.2.8. Biocides . . . 27
5.2.9. Antistatic Additives . . . 29
5.2.10. Reodorants . . . 29
6. Fuel Standardization and Testing 29
7. Storage and Transportation . . . 30
8. Alternative Fuels . . . 30
9. References . . . 34

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_719.pub2

1. History

1.1. The Spark Ignition (Otto) Engine and Its Fuel

In 1876 NIKOLAUS AUGUST OTTO developed a stationary, single-cylinder, four-stroke engine, that ran on coal gas. This invention, which was later named after him, set a development in motion that substantially shaped the industrial age and has not yet come to a halt [1, 2]. About ten years later the four-stroke engine was first used to drive vehicles (Daimler and Maybach automobiles, Benz patent car). The first vehicle engines ran on so-called *light gasoline* which had a low density (< 700 kg/m^3) and a very low final boiling point (ca. 85 °C). Light gasoline was available as a cleaning agent at pharmacies and could be evaporated in the first surface carburetors. The development of spray nozzle carburetors finally enabled the use of *straight-run gasoline* with a broader boiling range and higher final boiling point [3]. A higher proportion of crude petroleum could thus be used as a fuel, and the rising demand for gasoline could be satisfied.

After the development of a reliable mixture preparation system, attention focused on the further improvement of the engine power by increasing the compression ratio (from ca. 1918 on). The increase in compression ratio was shown to be limited by abnormal combustion, so-called engine knock, which can damage the engine. Since it was observed that knocking combustion can be related to gasoline formulation, systematic research began to improve the knock resistance of fuels.

Discovery of the strong antiknock effect of tetraethyllead by MIDGLEY and BOYD in 1922 was a milestone in the development of gasolines. A little later, refinery conversion processes were established that allowed the demand for gasoline to be met by improving the yield, and which also led to a further improvement in antiknock performance, so-called octane quality (→ Oil Refining). The further advances in the development of engine technology and in the quality of gasoline were mutually dependent. From 1950 on, increasing attention was given to the further improvement of gasoline by the addition of small quantities of highly active additives (see Section 5.1).

Since 1970, measures for improving the environmental compatibility of automotive use have become increasingly important worldwide. The first step was the successive reduction in the lead additive content of gasolines. Later, completely unleaded gasolines were developed, which were a prerequisite for the introduction of catalytic converters.

The reduction of sulfur content in gasoline has occurred more recently and it is still ongoing. This measure improves the efficiency and durability of modern exhaust gas aftertreatment devices and reduces the emission of sulfur compounds. In this context "sulfur free" gasolines (less than 10 ppm sulfur) have been made available on a wider scale. In countries with tax incentives for "sulfur free" fuels like Germany this fuel quality has a high market share. The limitation of benzene, other aromatics, and olefins concentration in gasoline is intended to reduce the resulting-post catalyst exhaust emissions. The worldwide rising demand for oil products has led to noticeable price increases and because of this development alternative fuels or fuel components (e.g., alcohols from biomass) have again achieved a certain significance.

1.2. The Diesel Engine and Its Fuel

The diesel engine, patented in 1892 by RUDOLF DIESEL, is characterized by self-ignition of the fuel [4]. Self-ignition is achieved by compressing the incoming air to such an extent that the resulting temperature increase is sufficient to ignite the fuel.

In the development of his engine, DIESEL's aim was an engine operation as close as possible to the Carnot cycle – which gives the highest theoretical efficiency – with the result that the first diesel engine built in 1893 at the Maschinenfabrik Augsburg Nürnberg AG (MAN) operated at a hitherto unachieved efficiency of 26.2 %. The power of this engine was 17.8 metric horsepower (13.1 kW) at 194 rpm, and it ran on a mixture of gasoline and lamp oil. After 1897, experiments with other fuels started, which had little success at first.

The injection of small quantities of fuel at high pressure caused the greatest problem in the initial phase of the development. Blowing the fuel in with compressed air, which was tried at

first, did not solve the problem satisfactorily. High-pressure liquid injection was developed in 1910; and in 1927, large-scale production of fuel injection systems started and the way was opened for broad expansion of the diesel engine. MAN and Daimler – Benz introduced the diesel engine to commercial vehicles in 1923 – 1924. About 12 years later, Daimler – Benz introduced the first passenger car diesel engine (type 260 D, 2.6 L, 45 metric horsepower at 3000 rpm).

Even the first diesel engines were often run on the so-called gas-oil fraction of crude petroleum (boiling range 200 – 350 °C). This type of fuel, which is provided in sufficient quantity mainly by crude oil distillation at moderate production cost is basically still used today, because of its suitable ignition quality (see Section 3.2.1).

In addition to high ignition quality (cetane number; see Section 3.2.1), cold flow properties and low tendency toward coking and sooting became important at an early stage.

During World War II, large quantities of diesel fuel were produced from coal in Germany. Mixtures of components of low ignition quality (tar oil) and very high ignition quality (from the Fischer – Tropsch synthesis) were used [3].

Because of the increasing demand for light oil products that followed World War II, gas-oil fractions from conversion processes have also been used in diesel fuel. However, these components are generally less suitable because of their lower ignition quality. For environmental protection and improved engine performance a number of measures were taken to raise the diesel fuel quality. For example, the maximum permitted sulfur content of diesel fuel was lowered stepwise from 1.0 (1955 in Germany) to 0.035 mass % (2000) and further to 0.005 mass % (2005) in Europe. Also a fuel quality with a sulfur content less than 0.001 mass % (sulfur free) must be available. The maximum permitted sulfur content in different areas is given below:

United States	0.001 (max average of refinery production) mass %
EU	0.005 and 0.001 mass %
Japan	0.0015 mass %

The reduction of sulfur in the diesel fuel helped also in the introduction of particulate traps in the exhaust systems.

Synthetic components with ultra-high ignition quality available from, e.g., natural gas synthesis (basicly Fischer-Tropsch synthesis with a newly developed catalyst) are used in some markets to formulate a diesel fuel with superior quality. Also high-quality components produced from biomass will shortly be available in limited quantities for commercial use.

Reliable engine operation and a reduction in exhaust gas emissions is also the primary aim in the treatment of diesel fuels with multifunctional additives, which has come into increased use from 1985 onwards. In addition, acceptability of diesel fuels could be improved (foaming, odor; see Section 5.2).

2. Engine Technology

2.1. Otto Engines

The basic principle of the spark ignition (Otto) engine–a combustion engine with timed ignition from an external source – can be realized in a number of variants with different types of fuel metering and mixture preparations [5].

The most frequently used type is the *four-stroke* reciprocating Otto engine in which gas exchange processes during these four strokes are controlled by valves. The sequence of strokes is

1) intake of the fuel – air mixture,
2) compression of the mixture and timed ignition,
3) combustion and expansion (working stroke), and
4) exhaust of combustion gases.

To carry out one working stroke the crankshaft turns twice. The valves are operated by a camshaft connected to the crankshaft.

The *rotary engine* (frequently called the *Wankel* engine after its inventor) is a variant of the four-stroke engine. It differs from the reciprocating piston engine in its rotary piston and in the use of ports for control of the inlet and exhaust gases [6, 7]. The rotary piston engine has not been very successful up to the present, compared to the reciprocating engine because of its relatively high fuel consumption and exhaust gas emissions.

In the standard *two-stroke* Otto engine the gas exchanges are also controlled by ports (inlet, transfer, and exhaust ports) in the cylinder.

A working cycle is achieved with each rotation of the crankshaft, and high power output at low swept volume is achieved. However, in the two-stroke process, scavenging losses must be accepted that lead to relatively high fuel consumption and hydrocarbon-exhaust emission. The two-stroke engine is used only in small motorcycles, outboard engines and light-weight machines [8]. Technically advanced models controlled by valves can overcome many of the disadvantages normally associated with these engines to a certain degree. However, more stringent exhaust gas requirements will reduce the use of two-stroke engines in the future.

The Otto engine runs with throttled intake air (quantity governing) and an almost homogeneous, approximately stoichiometric fuel – air mixture. Traditionally, the composition of the mixture depended on the running condition of the engine: i.e., at full load and directly after cold start the engines operated with an excess of fuel (rich mixture) whereas under part load nearer to the stoichiometric mixture. For the preparation of the fuel-air mixture previously carburetors [9] were used.

Modern electronic fuel injection systems, are superior to carburetors with respect to performance, fuel consumption, and accuracy of mixture strength control necessary for the efficient use of three-way catalytic converters [10]. Carburetors were therefore gradually replaced by fuel injection systems. Latest development are the so-called direct injected engines, which offer optimum fuel efficiency [11]. The main advantage of this technology is the avoidance of throttling losses at part load conditions. Whereas conventional Otto engines run at fixed air-fuel ratio and adjust power at part load by reducing the cylinder charge by the throttle valve, direct injected engines run in this respect rather like a diesel with almost unthrottled cylinder charge and varying amounts of fuel injected. To ensure proper ignitability of very lean mixtures at low load a locally gasoline enriched layer of air – fuel mixture ("stratified charge") must be created.

In the Otto engine the combustion process is started by an electric spark. Ignition occurs shortly before the piston has reached the top dead center and is advanced with increasing speed to allow a timed combustion beginning shortly before the top dead center [12].

To achieve the highest efficiency in the Otto engine, combustion induced by the ignition spark must develop in a controlled way under all operating conditions. If the resistance of the gasoline against autoignition, i.e., the octane quality of the fuel does not meet the octane number requirement of the engine, after compression by the piston either before or after ignition by the spark plug, autoignition of the yet unburned portion of the mixture occurs. This autoignition (engine knock) leads to a considerably higher combustion rate with a very sharp increase in pressure and temperature (Fig. 1). An anomalous combustion of this type manifests itself by a knocking or pinking noise. Whereas occasional knock that occurs on acceleration usually does not damage the engine, permanently knocking combustion at high speeds and loads over a longer period increases the risk of engine damage. Modern engine knock sensors allow safe engine operation even if gasolines of varying octane qualities are used, because they adjust the spark timing according to octane quality. However, gasolines with low octane quality require retarded ignition which leads to efficiency losses and, e.g., higher exhaust gas temperatures.

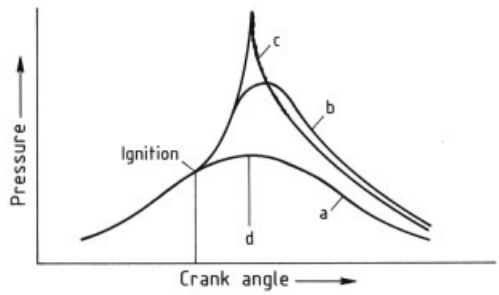

Figure 1. Combustion in the Otto engine
a) Without ignition; b) Normal combustion; c) Knocking combustion; d) Top dead center

Apart from design parameters the octane number requirement of an Otto engine varies with the engine speed and the load. Furthermore increasing deposit formation in the combustion chamber with increasing milage can raise the octane number requirement [13]. In designing an engine on the basis of a particular type of fuel, a safety margin must therefore be provided, to avoid knocking combustion even under unfavorable conditions.

2.2. Diesel Engines

The diesel engine differs from the Otto engine in the heterogeneous composition of its mixture and the self-ignition of the fuel. The temperature necessary for self-ignition is produced in the compression stroke, in which compression of air to ca. 3.0 – 5.5 MPa leads to temperatures of 700 – 900 °C. Fuel is injected into the heated air of the combustion chamber shortly before the end of the compression stroke, where it self-ignites. In principle the diesel engine can be either of the four- or two-stroke type. In practice, four-stroke engines are found in the vast majority of vehicles, whereas a valve-controlled two-stroke process with precompressed air is used in large ship engines.

In the diesel engine, air is generally taken in unthrottled and the fuel – air ratio varies with the quantity of fuel injected (quality governing) thus leading to a very lean fuel – air mixture ($\lambda > 1.3$, for definition of λ, see → Automobile Exhaust Control).

Two types of combustion systems have been developed for the diesel engine that differ in the preparation of the combustible gases [14]. *Direct injection* is characterized by a one-piece combustion chamber, into which fuel is injected without further preparation. Indirect injection is characterized by a two piece combustion chamber and was a mechanical solution to control the rate of combustion.

One of the main problems of the early diesel engines was the control of the combustion rate. Depending on the fuel and engine parameters ignition was often delayed and the subsequent rate of combustion more rapid than desired for best efficiency, low noise and emissions, etc. In those days before the development of modern electronic control units, various mechanical solutions were available, such as divided combustion chambers or specific mechanically controlled injection spray patterns to adjust the rate of combustion. Very common, especially for passenger car diesel engines, were concepts with divided combustion chambers, so called indirect injection engines (IDI engines). These *precombustion chamber* processes, such as the *swirl chamber* and *prechamber* processes, have a two-piece combustion chamber. Fuel is injected into the precombustion chamber, partially homogenized with air, and self-ignited. Partially oxidized combustible gases and evaporated fuel reach the main combustion chamber via connecting channels and burn there after further mixing with air. Indirect injection engines were used predominantly in passenger cars. Their engine noise is less harsh, but their fuel consumption is higher than that of the direct injection diesel engine. The split combustion chamber gives rise to pumping losses which impair the efficiency of precombustion chamber engines. This is also brought about by their higher compression ratio leading to higher friction losses.

In contrast to the divided combustion chamber engine concepts with a one-piece combustion chamber, i.e., direct injection engines (DI engines), were originally used mainly for heavy-duty transport and machinery. Their undivided combustion chamber allows higher engine efficiency. In modern direct injection engines sequentially electronically controlled high-pressure fuel injection systems allow optimum combustion and pressure rise thus avoiding previous problems, e.g., noise or limited speed range. The so-called *common rail technology* allowing high injection pressures is now well implemented for passenger car diesel engines.

The compression ratio of direct injection engines (ca. 15 – 17 : 1) lies near that optimum for most favorable total efficiency, which occurs at a compression ratio of ca. 15 : 1.

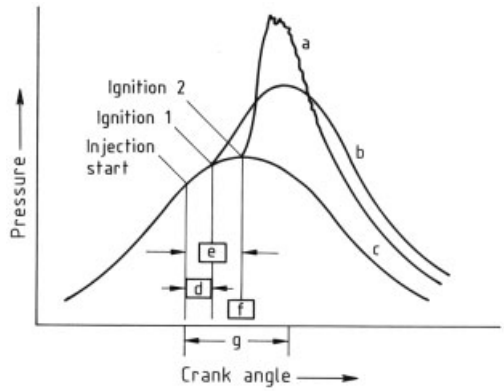

Figure 2. Combustion in the diesel engine
a) Noisy combustion; b) Normal combustion; c) Without combustion; d) Ignition delay 1; e) Ignition delay 2; f) Top dead center; g) Injection period

As mentioned before the diesel fuel is injected into the combustion chamber near the end

of the compression stroke shortly before the piston has reached the top dead center. The period between the beginning of the injection mode and the start of the combustion is called ignition delay. The length of ignition delay depends on ambient conditions, e.g., temperature, design and construction of the engine and the ignition quality of the diesel fuel. Optimum engine operation, necessary for high efficiency, low noise, and exhaust emissions, requires short ignition delay and therefore high diesel fuel ignition quality, i.e., a high cetane number. Prolonged ignition delay causes an unacceptable steep pressure rise causing, e.g., high noise and nitrogen oxide emissions (see Fig. 2) [15, 16].

3. Fuel Composition and Engine Efficiency

Because of the different modes of operation of Otto and diesel engines (see Chap. 2), completely different requirements exist for gasoline and diesel fuels and their composition. The effects of important fuel properties on the performance of engines are discussed below. The standardization of minimum requirements derived from them is described in Chapter 6.

3.1. Quality Aspects of Gasoline

The most important quality criteria for gasoline are its resistance against autoignition (engine knock), i.e., octane quality, the evaporation properties to provide an ignitable air – fuel vapor in the combustion chamber, environmental acceptability and low toxicity, and also cleanliness and stability. While all of these aspects are influenced by the main gasoline components and thus by refinery technologies, some can also be improved by additives, i.e., small proportions of organic substances, added to the gasoline mainly at the end of the production process and prior to delivery to the service stations.

3.1.1. Octane Quality

High octane quality is required to allow knock-free combustion in high compression engines to enable optimum engine efficiency. While the engine efficiency can be raised generally by increasing the compression ratio – leading to a higher octane requirement of the engine – the production of matching high octane fuels is connected with a higher energy demand in the refinery. The overall optimum efficiency (taking into account engine and refinery) is obtained with the most widely used premium and super gasolines with research octane numbers of 95 to 100 and engines designed for that octane quality.

The octane level of a gasoline depends on that of the individual refinery streams and on that of other, e.g., oxygen containing components such as ethers or alcohols. The octane quality of gasoline is described by the *octane number.* The octane number of a gasoline is determined by comparative measurements of its octane quality and that of binary mixtures having variable concentrations of *n*-heptane (low octane quality) and 2,2,4-trimethylpentane (isooctane; high octane quality). By definition the octane number of *n*-heptane is 0 and that of isooctane is 100. The octane numbers of mixtures are given by their percentage by volume of isooctane; octane numbers > 100 can be determined with lead-containing isooctane or toluene-containing mixtures.

Octane number determination (e.g., according to EN 25164 or EN 25163) is carried out in single-cylinder, four-stroke test engines specially developed for the purpose. The CFR (Cooperative Fuel Research) engine is most widely used worldwide. The test engines have a mechanically adjustable compression ratio. This ratio is raised during octane number measurement of a fuel until a certain standardized level of knocking, shown by a so-called knock meter, is obtained. The gasoline's octane number is determined by finding the *n*-heptane – isooctane mixture that gives the same knock level. Determination of the octane number is carried out under two different running conditions:

1) **Determination of the research octane number (RON)**: according to EN 25164, ASTM D 2699, in which comparatively mild conditions are used (600 rpm, no prewarming of the mixture, constant ignition timing).
2) **Determination of the motor octane number (MON)**: according to the EN 25163, ASTM D 2700, in which a higher mechanical and thermal load is used (900 rpm, prewarming of

the mixture to 150 °C, and variable ignition timing).

Because of the more severe test conditions the MON is usually lower than the RON; the difference between both is called *sensitivity* [17]. The RON describes the knocking performance of a gasoline at low and medium engine speeds ("acceleration knocking") whereas the MON defines the knock behavior under high speed and load ("high speed knocking"). The knock rating of a fuel is generally sufficiently characterized by its RON and MON. Historically it was necessary to use an additional characteristic, the *front octane number* (FON) describing the RON knock rating of the low boiling fuel components. An insufficient FON leads to knocking during acceleration, particularly in carburetor vehicles. This problem does no longer exist in fuel-injected vehicles.

For research and development purposes it is also possible to measure the octane quality in vehicle Otto engines under normal road use. The so-called *road octane number* gives information on the octane quality in vehicle Otto engines under normal road use. The road octane number is specifically related to the vehicle type and is also affected by driving conditions and weather. Usually the road octane number is given by an expression of the type a RON + b MON + c, where a, b, and c are constants that depend on the vehicle. The road octane number of a gasoline lies between the RON and MON, nearer the RON under acceleration conditions and approaching the MON under constant speed.

3.1.2. Volatility

Fuel volatility, expressed by distillation behavior (EN ISO 3405, ASTM D 86) and vapor pressure, controls the formation of an ignitable air – fuel mixture and the ability of the engine to start and run under cold and hot conditions. For safe operation the fuel must have a balanced distillation performance [13]. A certain proportion of volatile fuel components are needed for trouble-free performance at low temperature. An excess of these components leads, however, to hot fuel handling problems with a warm engine or high ambient temperature (see Fig. 3). To ensure trouble-free engine running, the boiling characteristics are usually matched to climatic conditions and change during the year. An excessively high final boiling point can cause deposit formation and lubricant dilution, and consequently engine wear.

Gasoline volatility also controls exhaust gas emissions and evaporative emissions. Vehicle design elements such as the activated carbon canister have of course a larger effect on these emissions. However, in a vehicle of given design the volatility of the fuel can contribute to guarantee low emissions. For example, a sufficiently high vapor pressure and high volatility ensures low exhaust gas emissions at low ambient and engine temperatures and sufficiently low vapor pressure reduces evaporative losses at high ambient and fuel system temperatures.

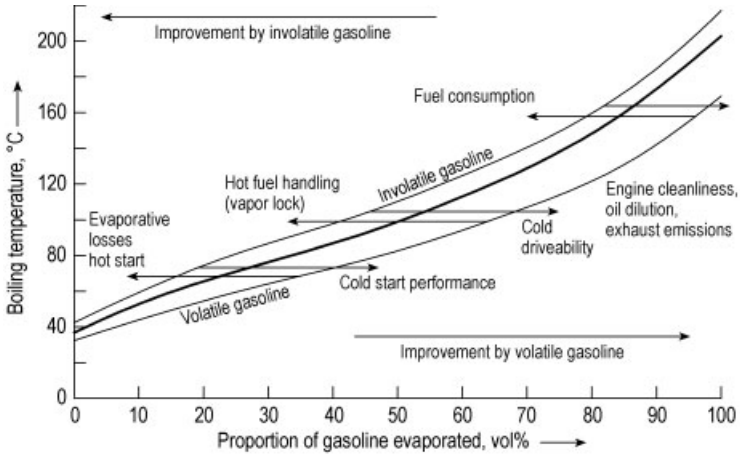

Figure 3. Significance of gasoline distillation properties for engine performance and emissions [19]

Table 1. Summary of effects of gasoline properties on emissions in noncatalyst cars (source: EPEFE program)

	Change	Lead	CO	HC exhaust	HC evaporation	NO_x	Benzene	Butadiene	Aldehydes
Add oxygenate	0.0 to 2.7 mass % O_2	0	▼▼▼	▼	0 – ▼	± 0	0	0	▲▲
Reduce aromatics	40 to 25 vol %	0	▼	▼	0	▼	▼▼	0	▲
Reduce benzene	3 to 2 vol %	0	0	0	– 0	0	▼▼	0	0
Reduce olefins	10 to 5 vol %	0	±0	▲	– 0	▼	0	▼▼	0
Reduce sulfur	300 to 100 ppm	0	0	0	0	0	0	0	0
Reduce Reid vapor pressure	70 to 60 kPa	0	0	± 0	▼▼▼	0	0	0	0
Increase E 100*	50 to 60 %	0	+ 0?	▼?	± 0	0	0	0	0
Increase E 150**	85 to 90 %	0	0	▼▼	0	▲?	0	▲?	▲?

0 = no effect. ± 0 = −2 to +2 %. ▼ or ▲ = 2 – 10 % effect. ▼▼ or ▲▲ = 10 – 20 % effect. ▼▼▼ or ▲▲▲ = > 20 %. ? = insufficient information.
* E 100 = % evaporated at 100 °C.
** E 150 = % evaporated at 150 °C.

3.1.3. Fuel Composition to Reduce Toxicity and Exhaust Emissions

Some of the gasoline constituents, especially benzene and previously used lead antiknock additives, are toxic and other components can form toxic materials during combustion. Both benzene and alkylleads were formerly used to improve the octane quality of gasolines. Their maximum allowed concentration was legally reduced stepwise. While lead has completely disappeared from gasoline in most countries also because of its poisoning effects on exhaust gas catalysts, benzene content has been restricted in Europe initially to maximum 5 vol % and since the year 2000 to 1 vol %. The major German oil companies produced their top quality Super Plus since 1995 already with benzene concentrations below 1 vol % on a voluntary basis. Benzene has not been used as an active blending component for a long time. It is extracted from the relevant gasoline components (e.g., reformate) mainly by distillation and also its formation in the catalytic reformer is minimized by extraction of "benzene precursors" in the reformer feed material.

To keep the octane quality of gasolines without lead and benzene above the standardized level, contents of other nontoxic materials were increased, namely of isomerates, higher aromatics, and to some extent ethers (→ Methyl Tert-Butyl Ether) and alcohols.

Ideally the gasoline is combusted completely to water and carbon dioxide, however, in practice other unwanted components are emitted because a small proportion of the gasoline (usually less than 1 %) leaves the combustion chamber unburned, e.g., because of quenching at the combustion chamber surfaces, or partly oxidized due to the high pressures and temperatures. Apart from carbon monoxide, unburned hydrocarbons, and nitrogen oxides, other toxic substances such as aldehydes, benzene, and 1,3-butadiene are emitted at very low concentrations. All of these are oxidized or reduced by the exhaust gas catalysts by more than 95 %. Their residual concentration can be influenced by gasoline composition and hence by the gasoline formulation in the refinery. Gasolines forming less toxic and environmentally harmful material have been called "reformulated gasolines".

The effects of gasoline properties and composition on exhaust gas and evaporative emissions have been studied in a wide range of modern vehicles, partly in joint motor – oil industry projects, (e.g., auto – oil programs) in the US and in Europe. The results of the European program are summarized in Tables 1 and 2 [20].

The reduction of sulfur and benzene has a clear-cut positive effect. The other measures, e.g., the reduction of total aromatics, the increase of mid-range volatility and the use of methyl *tert*-butyl ether have positive effects on some of the emission components but negative on others. This applies also to the reduction of vapor pressure (reduction of butane), which reduces the evaporative losses but increases emission of CO_2 and unburned hydrocarbons under cold start conditions.

A further objective is to reduce the formation of ozone and smog that occurs as a result of

Table 2. Summary of effects of gasoline properties on emissions in catalyst cars (source: EPEFE program)

	Change	Lead	CO	HC exhaust	HC evaporation	NO_x	Benzene	Butadiene	Aldehydes
Add oxygenate	0.0 to 2.7 mass % O_2	0	▼▼	▼	0 – ▲	+ 0	0	0	▲▲
Reduce aromatics	40 to 25 vol %	0	▼	▼?	0	▲?	▼▼	+ 0	▲?
Reduce benzene	3 to 2 vol %	0	0	0	– 0	0	▼▼	0	0
Reduce olefins	10 to 5 vol %	0	0	+ 0	– 0	– 0	0	▼▼	0
Reduce sulfur	300 to 100 ppm	0	▼?	▼?	0	▼?	▼?	▼?	▼?
Reduce Reid vapor pressure	70 to 60 kPa	0	0	– 0	▼▼	0	0	0	0
Increase E 100*	50 to 60 %	0	+ 0?	▼?	0	0	0	0	0
Increase E 150**	85 to 90 %	0	0	▼▼	0	▲?	0	▼?	▼?

0 = no effect. ± 0 = −2 to +2 %. ▼ or ▲ = 2 – 10 % effect. ▼▼ or ▲▲ = 10 – 20 % effect. ▼▼▼ or ▲▲▲ = > 20 %. ? = insufficient information.
* % evaporated at 100 °C.
** % evaporated at 150 °C.

photochemical reactions between hydrocarbons and nitrogen oxides. The ozone forming potential of hydrocarbons differs depending on their structure and size. Thus, the composition of the unburned hydrocarbons can be optimized by the composition of the gasoline to reduce the ozone forming potential of evaporative losses and exhaust emissions. As an example the reduction of the highly reactive olefins in the fuel is a measure to reduce the ozone forming potential, however olefins are desired fast burning gasoline components which are also readily converted in the exhaust gas catalysts.

The evaluation of the results of the above mentioned programs have led to the production of specific "reformulated gasolines" in some areas, e.g., California and to the stepwise introduction of more stringent gasoline specifications, e.g., in Europe, as listed below.

	2000	2005
Aromatics	42 vol %	35 vol %
Benzene	1.0 vol %	1.0 vol %
Olefins	18 vol % *	18 vol %
Sulfur	150 ppm	50 ppm and 10 ppm

* max. olefin content regular grade 21 vol %

In Germany some mineral oil companies had voluntarily introduced sulfur free gasoline qualities, e.g., Shell Optimax. Since 2005 at least one sulfur free gasoline must be available at the gasoline stations. Currently a tax incentive for sulfur free gasoline ensures that the majority of gasoline sold in Germany is sulfur free. Low and ultra-low sulfur levels can also be found in other countries, e.g., in Japan and in the US. Sulfur free gasolines (sulfur content less than 10 ppm) are required for new highly efficient exhaust gas aftertreatment systems especially for lean burn direct injection engines. The new specifications require changes to the refinery units including substantial investment.

3.1.4. Stability, Cleanliness, etc.

Problem-free gasoline storage, distribution, and vehicle/engine operation requires a practical freedom from water, acids, and foreign matter particles. The gasoline must have long term and high temperature stability in the presence of air, (oxygen) and must be neutral to metals and elastomers. A number of laboratory tests with specified maximum tolerable limits are used to define these properties (see Chap. 6). Some of these properties can be influenced by additives (see Chap. 5), e.g., water shedding, anticorrosion properties, oxidation stability. The long-term effect of fuel on elastomers in the fuel system is controlled by the selection of suitable materials by fuel system manufacturers. These materials have been tested in standardized and critical reference fluids.

3.1.5. Performance Additives

To ensure safe operation of the engine, high-quality gasolines contain special additives which ensure optimum inlet system cleanliness (see

Section 5.1.2) and prevent corrosion (corrosion inhibitors, see Section 5.1.1).

3.2. Quality Aspects of Diesel Fuels

The diesel combustion process requires a fuel, which in contrast to gasoline ignites easily. Ignition quality is therefore the main quality criterion of diesel fuel. Density is important as an indicator for fuel consumption. Cold flow performance, fuel lubricity, viscosity, stability, and cleanliness control mainly reliability of engine operation. All the above mentioned aspects together with sulfur content have effects on emissions. Fuel volatility is far less important than that of gasoline.

3.2.1. Ignition Quality

Since the fuel is not ignited by an external source in the diesel engine, good ignition performance of diesel fuel is an important criterion for its quality. Ignition quality is determined by characteristics of the components used in production of the fuel (see Section 4.2) and can be improved by additives (ignition improvers, cetane improvers, see Section 5.2.1).

The *cetane number* is used to describe the ignition quality of diesel fuels. Like the octane number, the cetane number is determined by comparative measurements in single-cylinder test engines (CFR swirl chamber engine (EN ISO 5165, ASTM D 613), or BASF prechamber engine) [21]. Mixtures of α-methylnaphthalene with very low ignition quality (cetane number 0) and cetane (*n*-hexadecane) with very high ignition quality (cetane number 100) are used as references. The cetane number of a reference mixture is given by the volume percentage of cetane in α-methylnaphthalene. For the determination of ignition quality the ignition delay (time interval between start of injection and start of self-ignition) is kept constant by varying the compression ratio of the engine depending on the test fuel quality.

The ignition quality of diesel fuels can also be calculated using the *cetane index* method (EN ISO 4264, ASTM D 4737). However the cetane index gives only an indication of the ignition quality of a diesel fuel without cetane improvers. This index is calculated by means of a complex formula based on density and distillation range. The formula is updated in regular intervals depending on changes in diesel fuel quality and production technology.

A high cetane number is advantageous for the ignition and starting behavior, the reduction of exhaust gas and noise emission ("clatter") [22, 23].

3.2.2. Density

The density of a diesel fuel (EN ISO 12185, ASTM D 4052) has a considerable effect on the engine performance. Because the quantity of fuel injected into an engine is metered by volume, the mass of fuel introduced into the engine increases with density. A higher fuel density leads to an enrichment of the fuel – air mixture which in principle yields a higher engine power output. On the other hand a high density increases also particulate emissions and black smoke [24]. The maximum density for diesel fuels was therefore reduced from 860 to 845 kg/m^3 in Europe.

3.2.3. Sulfur Content

Exhaust emissions, specifically particulates, are increased by the sulfur in the diesel fuel. In addition acidic combustion products arising from sulfur (sulfur oxides) can lead to engine corrosion. In Europe the maximum sulfur content had therefore been limited to 350 ppm from January 1, 2000 onward and was further reduced to 50 ppm by 2005. Furthermore, it was decided to make "sulfur free" diesel fuels (sulfur contents less than 10 ppm) generally available to allow the use of improved exhaust gas after treatment systems. A tax incentive in Germany for sulfur free diesel fuels led to a widespread availability of these fuels.

3.2.4. Cold Flow Properties

The composition of a diesel fuel affects its filterability at low temperature to a great degree. Particularly, *n*-paraffins with high ignition quality tend to form wax crystals at low temperature,

which can lead to clogging of the fuel filter. For this reason during winter periods the proportion of low boiling components such as kerosene and others with reduced concentration of *n*-paraffins is often increased in relation to higher *n*-paraffin-containing gas-oil fractions. Diesel fuels with filterability at very low temperatures, such as arctic fuels, consequently may have lower densities, viscosities, and also ignition performance than normal diesel fuels.

The low-temperature filterability of diesel fuel has been substantially improved since the 1970s by use of highly active flow-improving additives at reasonable costs [25]. The cold filter plugging point (CFPP), a laboratory test EN 116, gives an indication of the low-temperature behavior of diesel fuels (see Section 5.2.3). The cold flow performance of fuel components is also characterized by the cloud point, EN 23015, i.e. the temperature at which wax crystals begin to form. For fuels containing cold flow additives the cloud point has little relevance. However, it is used for the specification of eratic fuels as a safeguard with values 10 °C above the CFPP filterability.

3.2.5. Lubricity

The introduction of low and ultra-low sulfur fuels initially led to severe problems due to wear in fuel injection pumps, because the desulfurization process removes substantial amounts of molecules acting as natural lubricants. These have to be substituted by highly efficient lubricity additives (see Section 5.2). The lubricity performance of diesel fuels is defined by a laboratory rig test according to EN ISO 12156-1.

3.2.6. Viscosity

To ensure optimal engine performance the fuel's viscosity at 40 °C has to comply with narrow limits, e.g., in Europe it must be between 2.0 and 4.5 mm^2/s. (EN ISO 3104). Too low a viscosity can lead to wear in the injection pump; too high a viscosity deteriorates injection and mixture formation.

3.2.7. Volatility

The diesel engine is far less sensitive to changes in the fuel volatility than the Otto engine. However, the boiling range of a diesel fuel (EN ISO 3405) influences other properties such as viscosity, flash point, autoignition temperature, cetane number, density, and cold flow performance. Raising the back end temperature, for example, also raises the cloud point; lowering the temperature of the front end also lowers the flash point of the fuel. Generally speaking the variations in volatility of commercial diesel fuels are modest and usually far less significant than the influence of individual engine design features. To reduce exhaust emissions the amount of high boiling fuel components has been limited by lowering the specified 95 % point (temperature for 95 % evaporation) from previously 370 °C to 360 °C. Because of the intercorrelation with the other above mentioned physical properties it is not necessary to specify the light end of the diesel fuel by distillation properties.

3.2.8. Diesel Fuel Stability, Cleanliness, etc.

Safe storage and transportation of diesel fuels, their filterability, injector cleanliness, and compatibility with metals is achieved by meeting a number of specification criteria:

The minimum flash point (EN ISO 2719) of diesel fuel has been set to 55 °C in Europe by law. This limit has been adapted to the safety design elements of diesel fuel tanks which are less rigid than those for gasoline. Contamination of diesel fuel with small proportions of gasoline reduces the flash point below the legal limit and makes storage and transportation unsafe. However, it is legal to blend diesel fuel with gasoline in the vehicle tank to improve the cold flow performance at extremely low temperatures.

Total dirt (EN 12662) and water content (EN ISO 12937) are limited to guarantee acceptable filterability especially at low temperatures when filter elements may additionally be loaded with wax formed from paraffinic fuel components.

Carbon residue (EN ISO 10370) and oxidation stability (EN ISO 12205, ASTM D 2274) give an indication of possible deposit formation in the fuel system. Deposit formation is also controlled by addition of additives (see Section 5.2.2).

Table 3. Summary of effects of diesel fuel properties on emissions in light duty vehicles[b] (Source: EPEFE program)

	Change	CO	HC	NO_x	PM[d]
Reduce sulfur	2000 to 500 ppm	0	0	0	▼
Reduce density[c]	850 to 820 kg/m^3	▼	▼	0	▼▼▼
Reduce polyaromatics[c]	6 to 3 vol %				
Increase cetane number	50 to 55	▼▼	▼▼	± 0	▼
Reduce T 95[e]	370 to 330 °C	− 0	− 0	− 0	− 0

0 = no effect. ± 0 = −2 to +2 %. ▼ or ▲ = 2 – 10 % effect. ▼▼ or ▲▲ = 10 – 20 % effect. ▼▼▼ or ▲▲▲ = > 20 %. ? = insufficient information.
[b] Effects are not necessarily additive.
[c] Insufficient data is available to reliably separate the influence of these parameters.
[d] PM = particulate matter.
[e] T 95 = temperature for 95 % evaporation.

Table 4. Summary of effects of diesel fuel properties on emissions in heavy duty vehicles[b] (Source: EPEFE program)

	Change	CO	HC	NO_x	PM[d]
Reduce sulfur	2000 to 500 ppm	0	0	0	▼▼
Reduce density[c]	850 to 820 kg/m^3	0	0	▼	▼
Reduce polyaromatics[c]	6 to 3 vol %				
Increase cetane number	50 to 55	▼	▼	▼	▼
Reduce T 95[e]	370 to 330 °C	− 0	− 0	− 0	− 0

0 = no effect. ± 0 = −2 to +2 %. ▼ or ▲ = 2 – 10 % effect. ▼▼ or ▲▲ = 10 – 20 % effect. ▼▼▼ or ▲▲▲ = > 20 %. ? = insufficient information.
[b] Effects are not necessarily additive.
[c] Insufficient data is available to reliably separate the influence of these parameters.
[d] PM = particulate matter.
[e] T 95 = temperature for 95 % evaporation.

The anticorrosion properties are specified by means of the copper corrosion test (EN ISO 2160). Again they can be improved by specific additives (see Section 5.2.6).

3.2.9. Diesel Fuel Effects on Exhaust Emissions

The performance of a diesel engine is less sensitive to changes in fuel quality than that of the gasoline engine. However, in the diesel engine the fuel effects on emissions are more complex for several reasons: (1) The reduction of undesired exhaust gas constituents was so far achieved mainly during the combustion process and to a lesser extent by aftertreatment; (2) there is a wider range of engine technology in use (however there is now more conformity, i.e., direct injection engines with electronically controlled high pressure injection systems); (3) the diesel fuel composition is more complex than that of gasoline, i.e., because of the higher boiling range the individual molecules are less well defined (e.g., monoaromatics are mainly chemically bonded to long chain paraffins). Also properties and composition are often interlinked, such as aromatics concentration and density.

The effect of fuel properties and composition on exhaust emissions has been studied by oil companies and motor manufacturers – partly in joint programs, e.g., the European Program on Emissions, Fuels and Engine Technologies (EPEFE).

The effect of some fuel parameters, i.e., sulfur and monoaromatics content had been sufficiently documented previously. Sulfur in the fuel increases the emission of sulfur dioxide and of exhaust gas particulates, it has also negative effects on aftertreatment systems. Monoaromatics in the fuel have no measurable effect on exhaust emissions. The effect of other important fuel properties, i.e., density, polyaromatics content, cetane number, and final boiling point has been studied in the EPEFE program. The results of this investigation are summarized in Tables 3 and 4 [20].

For overall reduction of unwanted exhaust gas components it was decided to tighten the specification of diesel fuels in two steps in the years 2000 and 2005, (see also Chap. 6) by increasing the cetane number, reducing the maxi-

Table 5. Important properties of various gasoline components (approximate values) [3]

Component	ϱ_{15}, g/mL	MON	RON	E 70, [a] vol %	E 100, [b] vol %
Straight-run gasoline	0.680	62	64	70	100
Butane	0.595	87 – 94 [c]	92 – 99	100	100
Pyrolysis gasoline [d]	0.800	82	97	35	40
Light coker gasoline	0.670	69	81	70	100
Light catalytically cracked gasoline	0.685	80	92	60	90
Heavy catalytically cracked gasoline	0.800	77	86	0	5
Light hydrocracked gasoline	0.670	84	90	70	100
FR reformate (94) [e]	0.780	84	94	10	40
FR reformate (99) [f]	0.800	88	99	8	35
FR reformate (101) [g]	0.820	89	101	6	20
Isomerizate (isopentane)	0.625	87	92	100	100
Alkylate	0.700	90	92	15	45
Polymer gasoline	0.740	80	100	5	10
Methyl *tert*-butyl ether	0.745	98	114	100	100
Methanol – *tert*-butanol (1 : 1)	0.790	95	115	50	100

[a] E 70 = amount of fuel components with $bp < 70\,^{\circ}$C.
[b] E 100 = amount of fuel components with $bp < 100\,^{\circ}$C.
[c] Depending on the isobutane content.
[d] BTX aromatics extracted.
[e] Full range platformate with RON = 94.
[f] Full range platformate with RON = 99.
[g] Full range platformate with RON = 101.

mum sulfur content, the final boiling point and the density and by introducing a maximum limit for polyaromatics.

3.2.10. Performance Additives

To ensure optimum operation of the car or truck, high-quality diesel fuels contain a variety of special additives which ensure maximum cleanliness of the engine (detergents), minimize the foaming of the diesel during fueling processes (antifoaming additives), prevent corrosion (corrosion inhibitors), improve lubricity (lubricity improver) and improve the cold flow performance (flow improver, wax antisettling additives). Also ignition improvers may be used (see Section 5.2).

4. Fuel Components

The minimum requirements for fuel quality are defined in standards or specifications (see Chap. 6) to guarantee trouble-free engine operation and acceptable exhaust emissions. These requirements must be fulfilled in refineries by best use of the available product streams [26]. The blending of gasoline and diesel fuels therefore requests the careful selection of suitable refinery components, with sometimes very different physical and chemical properties in different refineries (see also → Oil Refining) or as a result of crude oil properties. More stringent requirements for the improvement of exhaust emissions post catalyst bring with them more and more intense refining processes such as, e.g., desulfurization.

4.1. Gasoline Components

The most important criteria that determine the suitability of gasoline components are the octane numbers (RON and MON), volatility (boiling properties and vapor pressure) and more recently chemical composition (aromatics, absence of benzene, olefins and sulfur) (Table 5). Gasoline fuels consist of hundreds of individual

Table 6. Approximate composition of various fuel components (in volume percent)

Component	Paraffins	Olefins	Aromatics
Straight-run gasoline	94	1	5
Butane	100		
Pyrolysis gasoline*	ca. 20	ca. 10	ca. 70
Light coker gasoline	ca. 57	ca. 40	ca. 3
Light catalytically cracked gasoline	61	26	13
Heavy catalytically cracked gasoline	29	19	52
Light hydrocracked gasoline	100	0	0
FR reformate (94)**	45		55
FR reformate (99)**	38		62
FR reformate (101)**	29	1	70
Isomerizate	98		2
Alkylate	100		
Polymer gasoline	5	90	5

* Reflux from petrochemical plants; after partial hydrogenation.
** For abbreviations, see Table 5.

chemical substances belonging mainly to one of the following three groups: aromatics, paraffins, and olefins (Table 6).

Oil refineries normally have extremely interdependent structures (→ Oil Refining) and, accordingly, many different fuel components are produced. The properties of fuel components, however, depend not only on the type of conversion process and the feed used, but also substantially on the mode of operation of the process.

4.1.1. Straight-Run Gasoline

This gasoline component is obtained by distillation and is therefore easiest to produce and also longest in use. The composition of straight-run gasoline depends on the type of crude oil used (frequently high paraffin content). The octane quality is usually low.

In the past, when highly leaded fuels were used, straight-run gasoline was a primary component of fuel. Currently (ca. 2000), the proportion of straight-run gasoline in motor fuel is generally very low (0 – 20 %).

Butane also belongs to the straight-run products but it has a high octane level. It is added to gasoline, besides being used as liquefied petroleum gas (→ Liquefied Petroleum Gas). Because of its high vapor pressure, butane is good for engine cold-start. On the other hand, too high a vapor pressure can lead to hot-fuel handling problems (warm start problems, stumbling engine, vapor lock) and increased evaporative hydrocarbon emissions [27]. Thus, addition of butane is limited by the vapor pressure specifications.

The yield of straight-run products supplied naturally by crude petroleum cannot meet the high demand for gasoline either qualitatively or quantitatively. Thus, high-boiling components of crude petroleum are cracked either thermally or catalytically into lower boiling products that can be used in gasoline production.

4.1.2. Thermally Cracked Gasoline

Thermally cracked gasoline is often used in the form of *pyrolysis gasoline.* It is obtained in particular as a byproduct from steam crackers (→ Ethylene) in petrochemical plants and has a high content of olefins and aromatics (with good octane quality). BTX (benzene, toluene, xylene) aromatics are frequently extracted from pyrolysis gasoline as valuable raw materials for chemical processes, so only the residues from petrochemical plants are sometimes used in gasoline production [28].

Variants of thermal cracking are *visbreaking* and *coking* (→ Oil Refining, Chap. 3.6.4.1). The relatively small quantities of gasoline obtained (visbreaker gasoline and coker gasoline) are olefin-rich and have a low octane quality.

4.1.3. Catalytically Cracked Gasoline

Catalytically cracked gasoline is more important for gasoline production. It has a high concentration of olefins and has to be desulfurized to be suitable for modern gasolines. Catalytically cracked gasoline is usually separated into high-boiling and low-boiling fractions, which have different applications. The octane quality (especially the MON) of catalytically cracked gasoline is usually sufficient only for regular gasoline. In order to meet the legally defined olefin specification of gasolines the amount which can be used in gasoline production is limited. The high olefin content of these fuel components also requires special precautions to be taken against polymerization and oxidation during their use (see Section 5.1.3).

Another variant of catalytic cracking is *hydrocracking*, a process in which hydrogen is added. The hydrocracked gasoline produced by this process contains practically no unsaturated components and consists principally of iso- and cycloparaffins. Hydrocracked gasoline has a high MON, a moderate RON, relatively low density, and high volatility [29].

4.1.4. Catalytic Reformate (Platformate) (→ Oil Refining, Chap. 3.4)

One of the most important motor gasoline component is catalytic reformate, sometimes referred to as platformate. Reformate has an exceptionally high aromatic content and a very high octane quality. It is obtained by reforming (isomerization and dehydrogenation reactions) from paraffin-rich straight-run gasoline with a low octane quality [30]. The properties of the reformate essentially depend on the mode of operation in the reforming process [31]. When very high octane numbers are achieved, the aromatics content and the density increase, but at the same time the yield and volatility decrease. Because of its high octane quality, reformate is the backbone for the production of lead-free premium fuels. The restrictions of benzene and aromatic content in gasoline fuels limit the use of these components and also requires specific treatment of the reformer feed and/or the reformed product to remove the benzene from this stream (mainly distillation and hydrogenation techniques).

4.1.5. Isomerate (→ Oil Refining, Chap. 3.7.3)

The demand for volatile gasoline components with sufficient octane quality cannot be met by, e.g., unrefined straight-run gasoline [32]. For this reason, from the 1970s onward the low-octane straight-run gasoline was increasingly converted into the higher octane isomerate (mainly isoparaffins). Because it is free from aromatics isomerate is becoming increasingly attractive for environmental reasons.

4.1.6. Alkylate (→ Oil Refining, Chap. 3.7.1)

Another attractive fuel component consisting mainly of isoparaffins is alkylate, which is moderately high in octane quality (especially with regard to the MON), however, less suitable to formulate higher octane premium fuels like Super Plus (minimum RON 98). Alkylate is obtained by the reaction of isobutane with olefins and can only be produced in limited capacity because of the low availability of isobutane.

Since the introduction of mandatory maximum aromatic levels in gasolines alkylate has become one of the most important components for the production of premium fuels. Alkylate is the main component in aviation gasoline (Avgas).

4.1.7. Polymer Gasoline (→ Oil Refining, Chap. 3.7.2)

A less common gasoline component is polymer gasoline, obtained by oligomerization of propene or butenes produced in cracking plants. It consists primarily of $C_6 - C_{12}$ olefins.

4.1.8. Oxygenates

High octane quality of unleaded gasoline can be obtained more easily by the addition of oxygen-containing fuel components [33, 34]. In particular, the addition of methyl *tert*-butyl ether (MTBE) (→ Methyl Tert-Butyl Ether) has proved to be very useful for maintaining the octane quality of unleaded high octane fuels, even

more since the introduction of maximum aromatic levels of 42 vol % (2000) and 35 vol % (2005). Besides its high octane quality MTBE is a very clean burning gasoline component.

Furthermore, methanol, ethanol produced from biomass, or higher alcohols can be used in low concentrations.

4.2. Diesel Fuel Components

The self-igniting diesel combustion process requires molecules with different structures than used in gasoline, i.e., in contrast to gasoline diesel fuel components must decompose easily at high temperature and pressure. Normal *n*-paraffins are for that reason ideal diesel fuel constituents, in contrast to olefins and aromatics. However, because of the higher boiling range of diesel fuel, the hydrocarbon molecules often consist of a combination of different structural moieties (e.g., an aromatic ring with a long paraffinic or olefinic side chain). The main characteristics determined by the composition of the diesel fuel are its ignition performance, cold flow properties, volumetric calorific value, and density. The latter can affect the air – fuel ratio and therefore influences particulate emissions (smoking tendency) under maximum torque conditions.

None of the classes of substances present in diesel fuel fulfills all the criteria equally well; for example, *n*-paraffins, which have a very good ignition performance and low smoking tendency, show poor low-temperature behavior and have a low calorific value (Table 7). However the combination of, e.g., paraffinic and aromatic substructures in the molecules can reduce cold flow problems without significantly impairing the combustion properties.

The diesel engine can also theoretically be run on lower and higher boiling hydrocarbons. Marine diesel engines for example use higher boiling residual fuels (→ Heating Oil, Chap. 2). The particular value of the automotive diesel engine lies, however, in its ability to utilize economically the middle distillates in the distillation range 150 – 360 °C. Accordingly, the diesel fuel components produced in refineries must correspond to the aforementioned requirements as far as possible.

4.2.1. Straight-Run Middle Distillate

Straight-run middle distillate (distillate gas oil) represents a valuable diesel fuel component because of its usually high cetane number (see Fig. 4). The quality of straight-run middle distillate nevertheless depends strongly on the quality of crude petroleum used. In many cases, distillate gas oil has a relatively high paraffin and low aromatic content.

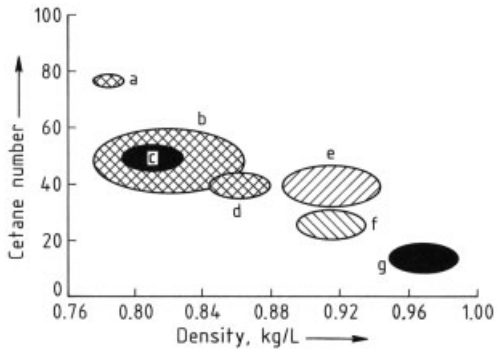

Figure 4. Typical properties of gas-oil components
a) Synthetic middle distillate; b) Straight-run gas oil; c) Hydrocracked gas oil; d) Thermally cracked gas oil; e) Syncrude gas oil; f) Coker gas oil; g) Light catalytically cracked cycle oil

Although middle distillates can be used for the diesel engine without further processing (apart from desulfurization), other diesel fuel components are also employed. The following

Table 7. Properties of various hydrocarbon groups with regard to their suitability as diesel fuels

	Ignition quality	Cold flow properties	Volumetric calorific value	Density	Smoking tendency
n-Paraffins	good	poor	low	low	low
Isoparaffins	low	good	low	low	low
Olefins	low	good	low	low	moderate
Naphthenes	moderate	good	moderate	moderate	moderate
Aromatics	poor	moderate	high	high	high

Table 8. Properties of diesel fuel components

Process	Component	Composition	
		Before treatment	After treatment
Distillation	distillate gas oil (straight run)		content of aromatics: medium; olefins: very low; paraffins: high
Thermal cracking	thermocracked gas oil	content of aromatics: low; olefins: medium; paraffins: high	content of aromatics: low; olefins: none; paraffins: high
Catalytic cracking	catalytically cracked gas oil	content of aromatics: high; olefins: low; paraffins: low	content of aromatics: high; olefins: none; paraffins: low
Hydrocracking	hydrocracked gas oil		content of aromatics: very low; olefins: none; paraffins: very high
SMDS synthesis	synthetic gas oil		content of aromatics: none; olefins: none; paraffins: very high
Distillation or hydrocracking	kerosene		content of aromatics: low; olefins: none; paraffins: very high

reasons are decisive: (1) the natural middle distillate fraction in crude petroleum is generally not sufficient to meet the demand for middle distillate; (2) unwanted heavy components in crudes are present in excess and are, therefore, converted into lighter products; (3) the refinery structures also provide a certain "compulsory fraction" of middle distillates from conversion processes; and (4) particular quality requirements can also make the production of special diesel fuel components necessary.

4.2.2. Thermally Cracked Gas Oil

Thermally cracked gas oil originates from either the visbreaking or the coking process and tends to have a lower cetane number and higher density than distillate gas oil (Table 8). Thermally cracked gas oil usually goes through a subsequent hydrogenation step because of the formation of unstable olefinic components during the cracking process. After hydrogenation, it represents a usable diesel fuel component.

4.2.3. Catalytically Cracked Gas Oil

Catalytically cracked gas oil is generally characterized by a very low cetane number and high density. Because of its instability, catalytically cracked gas oil must normally be hydrogenated [35]. Because of the poor ignition performance resulting from the high aromatic content and the high smoking tendency, it can be added to diesel fuel only in limited quantities.

4.2.4. Hydrocracked Gas Oil

Hydrocracked gas oil is a very valuable diesel fuel component. Because of its very high paraffins content, the ignition performance is good and the smoking tendency low.

4.2.5. Kerosene

Improved cold flow properties of diesel fuel can be achieved by the addition of kerosene. Because kerosene is predominantly used as jet fuel and, therefore, expensive and in short supply, its addition is limited mainly to meet the requirements of low temperature climatic conditions. The density, calorific value, and viscosity of kerosene are in most cases, however, much lower than those of other diesel fuel components and therefore its use is limited anyway. The low density and low calorific value would lead to decreased engine power and increased volumetric fuel consumption. The low viscosity would require additives to avoid wear in the injection pump because of poor lubricity.

4.2.6. Synthetic Diesel Fuel

Diesel fuel can also be obtained from fossil energy sources other than crude petroleum. Dur-

ing World War II diesel fuel was produced from coal by the Fischer –Tropsch process in Germany. This synthetic fuel (*kogasin*) had a range of outstanding properties: cetane number ca. 100, almost sulfur-free composition, low density, and high gravimetric calorific value. This synthetic diesel fuel component was used principally for enhancing low-value coal tar oil. Synthetic diesel fuel is produced in small quantities by the Shell Middle Distillate Synthesis (SMDS) process in Malaysia from natural gas [36]. The process includes the following steps:

1) Steam reforming of natural gas (conversion to synthesis gas),
2) Buildup of long-chain *n*-paraffins (Fischer – Tropsch synthesis),
3) Isomerization to isoparaffins
4) Distillative separation into kerosene and diesel fraction

The product was initially used mainly for improvement of the poor quality of the straight-run middle distillate otherwise available in Malaysia. The middle distillate produced by the SMDS process results in lower exhaust emission in diesel engines because it has a high cetane number (> 60). To meet increasing requirements for diesel fuel quality of modern low emission engines synthetic middle distillates are meanwhile used as components in top-quality diesel fuels.

A process is currently under development is to convert biomass to gas on a commercial scale with subsequent use of the SMDS process.

5. Fuel Additives

By the generally accepted definition only agents that are added to fuels in a concentration of $<$ 1 % should be described as "additives." For substances present in higher concentration, the term "fuel components" is more appropriate.

Fuels are treated with additives for various reasons: in the past the addition of octane enhancers to gasolines enabled the production of high octane grades. Nowadays the use of cetane improvers offers the possibility of upgrading a base fuel very economically compared with refinery measures.

Performance additives, on the other hand, which are added to modern high quality fuels, improve the behavior of a base fuel in operation, and offer technical advantages that often cannot be achieved by measures taken in the refinery [37]. In many cases, these types of additives offer the only possibility of guaranteeing trouble-free engine performance over a longer running period. Finally, treatment of fuels with additives is a major route to achieve product differentiation and trademarked quality.

Thus it is not surprising that in Western Europe further increase in the use of additives is generally anticipated [38] with the exception of the alkyllead compounds which due to their incompatibility with exhaust catalysts and their toxicity have been phased out as of January 1, 2000 in the EU. A derogation of two years for the lead phase out was granted to Spain, Italy, and Greece by the EU commission. Lead phase out is a general trend worldwide, except for developing countries with low car population.

5.1. Gasoline Additives

The treatment of gasoline with additives is almost as old as the fuel itself [39]. At first (around 1900), the search for so-called brisance-increasing additives was predominant (Fig. 5). In 1918, treatment with certain additives was recognized to allow an increase of compression ratio without dangerous knocking. An increase in the octane number became the major aim for decades in additive technology. Systematic research into additives that could positively influence many aspects of engine behavior began around 1950 on a wider scale. Since ca. 1970, conventional, ash-forming antiknock additives became increasingly less important for environmental reasons. Nowadays additive development is driven by environmental specifications as well as demanding new technologies like direct injected gasoline cars.

5.1.1. Corrosion Inhibitors

Water is almost always present in traces and can ingress into the fuel during processing, transfer, and storage by condensation. Moreover alcohols, when present in the fuel, can also lead to corrosion problems either by themselves or because of their water-solubility. Corrosion must

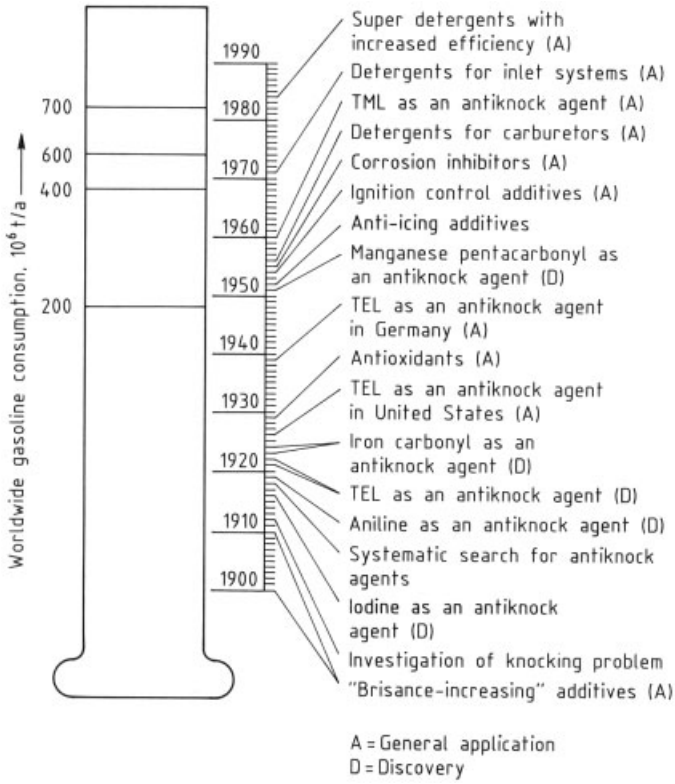

Figure 5. Historical development of gasoline additives

be inhibited not only to avoid damage of the tank or the fuel system but also because fine rust particulates can block fuel filters and injectors.

The effect of corrosion inhibitors is based on their ability to cover a metal surface with a monomolecular layer of inhibitor molecules and thus protect it from contact with corrosive components [40]. The structure of the corrosion inhibitors corresponds to this requirement: a polar end of the molecule (e.g., a carboxylate, ester, or amine group) is attached to the metal surface; while a nonpolar, high molecular mass alkyl chain at the other end of the molecule extends into the hydrocarbon phase (Fig. 6). An oily covering layer is thus formed that effectively repels water.

High alcohol contents in fuel (e.g., gasohol in Brazil) are a problem because alcohols, in addition to their own corrosiveness, can also dissolve the protective film. Apart from these cases, which require exceptionally high concentrations (e.g., 100 ppm) of corrosion inhibitors, 10 – 20 ppm, is generally sufficient for effective corrosion protection.

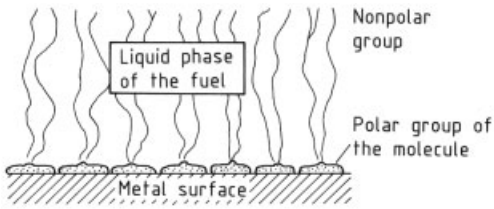

Figure 6. Mode of action of corrosion inhibitors

The efficiency of corrosion inhibitors can be tested in the corrosion test according to DIN 51585, ASTM D 665, in which gasoline – water mixtures are allowed to react with steel test objects.

Examples of suppliers are BASF (Kerokorr additives), Octel (DCI series), Lubrizol (LZ series), Bayer AG (RC additives), Infineum, Nalco/Exxon, Baker/Petrolite (Tolad) Betz-

Dearborn, Apollo Technologies (PRI), UOP (Unicor), Ethyl (Hitec), and Champion Technologies (RPS).

5.1.2. Detergents

In Otto engines, high molecular mass organic deposits can be formed in various sites in the fuelling and inlet and mixing systems. These deposits can lead to considerable problems in engine operation. Deposits on the *carburetor* and *injectors* prevent the formation of an optimal fuel – air mixture and cause increased fuel consumption and poor exhaust gas emissions [41 – 43]. In extreme cases the engine can even stall. Severe deposit formation on the *inlet manifolds* and *valves* causes, in addition to increased fuel consumption and deterioration in exhaust emissions, a decrease in power [44 – 47]. Sometimes even mechanical engine damage can occur.

The formation of deposits is caused by several factors [48 – 51]: Fuel components with high olefin or heteroatom content (which also have a marked tendency toward gum formation) are relatively prone to deposit formation. This is also true for fuels with very high final boiling points. Engine oil, which enters the inlet system via the valve stem guides, also contributes to deposit formation. Finally, combustion gases play a role in deposit buildup. These gases get into the inlet system either with the recirculated blowby gases from the crankcase or directly from the combustion chamber because of valve overlap. The multiplicity of factors contributing to deposit buildup shows that the problem cannot be solved by individual improvements in single areas but only by the addition of detergents with extensive effectiveness.

Detergents for Carburetor and Injector Cleanliness. Detergent additives were first used in the 1950s to maintain carburetor cleanliness ("keep clean effect"). Later, detergent additives were developed which removed deposits already present in the carburetor ("cleanup effect"). First-generation detergents were specially designed to work in the relatively cold region of the carburetor. These detergents include alkylamines, alkyl phosphates, alkyl-substituted succinimides, imidazolines, and oleylamides, which were added to the fuel in concentrations of 50 – 100 ppm. These compounds are also suitable, in principle, for keeping injectors clean, although concentrations of 150 – 300 ppm are then required [52].

To assess the efficiency of a carburetor detergent additive, the CEC developed a test cycle with a Renault 5 engine [53].

Detergents for Inlet Valve Cleanliness. From the mid-1970s, the search for decreasing fuel consumption and increased power led to leaner fuel – air mixtures and temperature increases in the inlet system. This created several problems that could not be overcome by the first-generation detergent additives, because these were effective only in the low-temperature range. In individual cases the carburetor detergents caused even higher deposit accumulation on the inlet valves due to decomposition.

The development of second-generation detergent additives made it also possible to keep the hot inlet valve clean [47, 54 – 56]. Long-chain, high molecular mass alkyl compounds (molecular mass around 500 – 1500 g/mol) with a polar end group, are very effective. The mechanism is analogous to that of corrosion inhibitors described in Section 5.1.1. Effective, modern detergent additives are based on polyisobuteneamine, polyisobuteneamide/succinimide, or polyetheramine technology. To meet most stringent environmental legislation these additives have to be produced halogen-free. These detergents, except for the polyetheramines, are normally used in combination with very thermally stable carrier oils that keep the surface of the valve lubricated and ensure continuous flowing off of deposit particles. To keep inlet valves clean relatively high concentrations of detergent (300 – 400 ppm) and carrier oil (300 – 400 ppm) are used. Depending on the polar end group modern detergents provide both keep-clean as well as cleanup performance.

A number of tests have been developed to assess the detergency performance of these gasoline additives: the Mercedes Benz M102E test was developed in Europe by CEC [57, 58].

Examples of suppliers are BASF (Keropur multipurpose additives), Infineum (Vectron), Ethyl (Hitec), Octel (OGP), Oronite (OGA series), and Lubrizol (ADX).

5.1.3. Antioxidants (→ Antioxidants)

As early as 1930 the aging of gasoline was known to be inhibited considerably by antioxidants (oxidation inhibitors). Nevertheless, antioxidants achieved importance only after World War II, when olefin-rich fuel components from cracking processes were increasingly used (→ Oil Refining). Olefins, in particular conjugated dienes, undergo liquid-phase free-radical chain reactions relatively easily at low temperature, which finally lead to the formation of highly polymeric residues, the so-called gum [59]. Other potentially gum-forming components are aromatics and the nitrogen and sulfur compounds present at trace levels. The efficiency of antioxidants is based on their capability to interfere with the mechanism of gum formation. In this mechanism, the initiating step is formation of a radical ($R^1-H \rightarrow R^1\cdot$); in the second step, hydroperoxides are formed by reaction with dissolved oxygen from the air and the chain reaction is started (Eqs. 1 and 2):

$$R^1\cdot + O_2 \rightarrow R^1-O-O\cdot \quad (1)$$

$$R^1-O-O\cdot + R^2H \rightarrow R-O-OH + R^2\cdot \quad (2)$$

Gum-forming reactions continue until the chain reaction is terminated by the coupling of two radicals (Eqs. 3 and 4). The chain termination reaction can also be taken over by an antioxidant (AH), which leads to the formation of an inactive radical A· that is not able to continue the chain reaction (Eq. 5):

$$R^1\cdot + R^2\cdot \rightarrow R^1-R^2 \quad (3)$$

$$R^1-O-O\cdot + R^2\cdot \rightarrow R^1-O-O-R^2 \quad (4)$$

$$R^1-O-O. + AH \rightarrow R^1-O-O-H + A \quad (5)$$

The aging of gasoline must be inhibited not only because of gum formation, which (1) blocks carburetors and injectors, filters in engines or dispenser pumps; and (2) can lead to dirt accumulation in the inlet system. An additional function of the antioxidant used to be the stabilization of tetraethyllead.

The gum content of gasoline fuels is measured according to EN ISO 6246, ASTM D 381, giving the washed gum evaporation residue. The residue has to be washed with *n*-hexane to remove detergent additives and carrier oils used in high performance fuels.

The induction period (EN ISO 7536) is a laboratory test to measure the storage stability of gasolines. Most antioxidants belong to one of the following two classes of substance.

para-Phenylenediamines have the general formula $R^1-NH-C_6H_4-NH-R^2$, in which R^1 and R^2 can be *sec*-butyl-, isopropyl-, 1,4-dimethylpentyl-, or 1-methylheptyl-. *para*-Phenylenediamines are very effective at concentrations of 5 – 20 ppm, particularly for gasolines with high olefin content. Their basicity, however, makes them susceptible to reaction with acidic impurities or fuel components. On the other hand, phenylenediamines offer the advantage of catalyzing the oxidation of mercaptans (which have foul odors even when present in trace amounts) to odorless disulfides.

Hindered Alkylphenols. Alkylphenols are optimally efficient if their hydroxyl group is sterically hindered by alkyl groups in positions 2 and 6, e.g., as in 2,6-di-*tert*-butylphenol.

Alkylphenols are particularly suitable for olefin contents $< 10\,\%$. Because of steric hindrance, the acidity of the hydroxyl group is lowered to such a degree that the danger of side effects is low. To achieve the same efficiency as phenylenediamines, higher concentrations (5 – 100 ppm) are necessary. Because of their positive synergism, mixtures of phenylenediamines and alkylphenols are also used sometimes [60].

Examples of suppliers are BASF (Kerobit additives), Octel (AO), and Ciba (Irganox).

5.1.4. Metal Deactivators

The aging (oxidation) of gasoline can be accelerated considerably by traces of metals that get into the fuel during transfer. Mercaptans, phenols, and alcohols increase the solubility of metals in fuel. Copper has the highest catalytic activity, accelerating the oxidation of fuel considerably at concentrations as low as 0.01 mg/L.

The function of metal deactivators is based on the fact that they form stable chelate complexes with dissolved metal atoms so that the metal loses its undesirable catalytic activity

[40]. The addition of *N,N′*-disalicylidene-1,2-propanediamine is particularly effective,

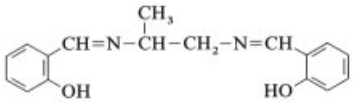

and copper is complexed as follows:

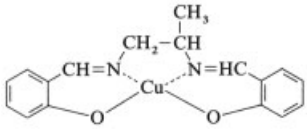

Examples of suppliers are BASF (Keromet additives) and Octel (DMD).

5.1.5. Anti-Icing Additives

The problem of carburetor icing has virtually disappeared because of the introduction of multipoint fuel injection systems which are practically immune to icing.

During the evaporation of fuel in the carburetor or in single point injection systems, heat is absorbed from the surroundings so that even at ambient temperatures > 0 °C, temperatures below 0 °C can occur and lead to freezing of the air moisture there. Besides its temperature, the relative humidity of the air is an important factor for the formation of icing (Fig. 7) [54].

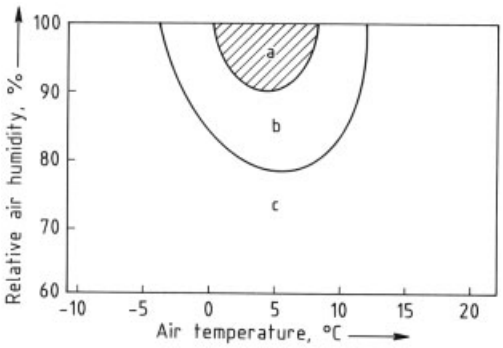

Figure 7. Effect of relative air humidity and air temperature on carburetor icing [54]
a) Severe icing; b) Icing; c) No icing

Carburetor icing was particularly noticeable during idling (engine stalling because of ice formation) or constant speed driving making itself noticeable by engine power loss, high fuel consumption, unfavorable exhaust gas emissions, and finally, engine stalling. To avoid carburetor icing, two different types of additives were used: (1) surfactants and (2) freezing point depressants (→ Antifreezes).

Surface-Active Substances. Surfactants produce protective films over the metal surface so that ice is prevented from sticking to it. Examples of surface-active compounds are amines, diamines, amides, or glycol esters of fatty acids, which are effective at 10 – 60 ppm. Other surface-active substances added to fuels, such as detergents, act as anti-icings in addition.

Freezing Point Depressants. Effective freezing point depressants include alcohols, glycols, dimethylformamide, and other water-soluble polar substances. Depending on the type used they were added in concentrations of 0.02 – 2 vol %.

5.1.6. Additives for Combating Combustion Chamber Deposits

Mainly when leaded gasolines were used considerable quantities of lead compounds, but also carbon, and engine oil components could deposit in the combustion chamber and on the spark plugs [61 – 64]. This could lead to knocking combustion and to spark plug fouling problems.

Organophosphorus and organoboron compounds were therefore added to fuel at an early stage. These additives do not remove the deposits but are at least able to convert them to materials with high glow temperature and low electrical conductivity (*deposit modifiers*). The commercial products contained, for example, tricresyl phosphate, triisopropyl phosphite, borazane, or boric acid esters. These additives are no longer on the market because (1) the phase out of lead compounds reduced the problem and (2) phosphorus compounds are incompatible with catalytic converters because of their tendency toward catalyst poisoning.

Research was devoted to the development of *ash-free* additives, which counteract the buildup of combustion chamber deposits and control any increase in the octane number requirement of the engine [*octane requirement increase control*

(ORIC) additives]. These additives have shown some positive effects in test engines. However their effect under actual road conditions seems to be limited. Reduced engine oil consumption, precise fuel metering to obtain stoichiometric air-fuel ratios, the shape of modern combustion chambers and of course the use of unleaded gasoline have reduced the problem of combustion chamber deposit formation. Specific additives to reduce these deposits are therefore not in general use.

5.1.7. Valve Seat Recession Protection Additives

Some older cars with "soft valve seats" (e.g., cast iron) relied on leaded gasoline, since lead compounds inhibited corrosion and wear of the valve seats. Unleaded gasoline for older cars can be provided with organic potassium or sodium compounds as lead replacement. These can be added to the finished fuel by the end consumer using a bottle solution. However, valve seat recession occurs predominantly under high-duty (motorway) operation. As vintage cars with soft valve seats are normally not operated under these conditions it is generally not necessary to use these additives.

5.1.8. Antiknock Agents (→ Octane Enhancers)

The most important antiknock additives were alkyllead compounds, i.e., tetraethyllead and tetramethyllead. As pointed out the use of these additives was discontinued because of their poisoning effects on the exhaust-gas aftertreatment systems and their toxicity. Other metal-based antiknock additives have either similar unwanted side effects or limited activity. Organic antiknock additives are less effective than ash forming products. Best known are nitrogen-containing aniline derivatives. None of these antiknock additives is used commercially on a larger scale. Instead oxygen-containing blending components, mainly ethers (→ Octane Enhancers; → Methyl Tert-Butyl Ether) are used to increase the octane quality of unleaded gasolines.

5.1.9. Dehazers and Antistatic Additives

Dehazers and antistatic additives are rarely used in gasoline. They are similar to the corresponding types of additives in diesel fuel (see Sections 5.2.7 and 5.2.9).

5.2. Additives for Diesel Fuel

Apart from a few exceptions treatment of diesel fuel with additives was introduced much later than that of gasoline. Several reasons existed for the introduction of additives: (1) a significant progress in diesel engine technology occurred from about the mid-1970s onward which led to a greater market share of diesel engines in passenger cars; (2) fleet operators and car owners required improved reliability and comfort; (3) more stringent emission standards required improved diesel quality.

5.2.1. Ignition Improvers (Cetane Improvers)

The cetane number (see Section 3.2) is a measure of the ignition quality of a diesel fuel. Because a range of aspects essential to the operation of diesel engines are also connected with the cetane number (cold-start properties, particulate, e.g., white and black smoke emission, noise emission, fuel consumption, engine durability), the increase in cetane number by additives is a valuable means of improving diesel fuel quality [65].

The effect that can be achieved by the use of ignition improvers depends on the base diesel fuel. Unfortunately, base fuels with high natural cetane numbers react better to additive treatment than those with lower cetane numbers, for which an increase in the cetane number is more essential (Fig. 8). In view of constantly increasing efforts to reduce exhaust gas emissions, the fact that even small quantities of ignition improvers can improve the exhaust gas quality of the diesel engine is important (Fig. 9) [66, 67].

The effect of ignition improvers is based on their rapid decomposition, with the formation of free radicals that accelerate the chain reaction of diesel fuel combustion. Suitable compounds

are alkyl nitrates, ether nitrates, alkyldiglycol nitrates, and organic peroxides. From a commercial point of view only alkyl nitrates such as 2-ethylhexyl nitrate (EHN) have been successful. With these compounds, the ready cleavage of the $RO-NO_2$ bond leads to radical formation [69].

Examples of suppliers are Octel (CI additives) and Ethyl (Hitec).

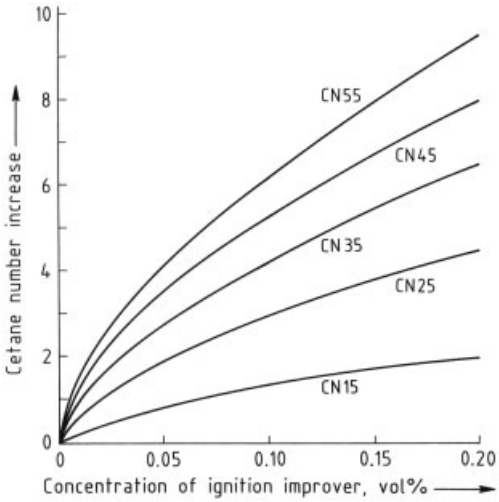

Figure 8. Response of different base fuel types to addition of ignition improvers

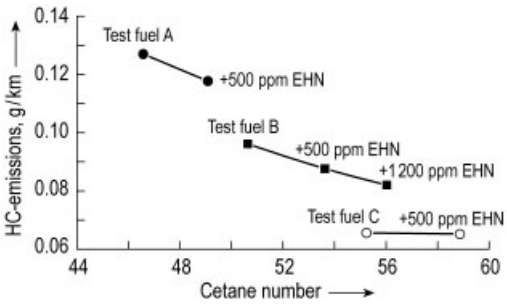

Figure 9. Improvement of exhaust gas emissions by ignition improvers; average emissions of three passenger cars with pre- and swirl chamber diesel engines [67]

5.2.2. Detergent Additives

Corresponding to the different construction of the diesel engine compared with the Otto engine, the problem of deposit formation occurs in a different form. In diesel engines the injection system is particularly affected, where gum- or carbonaceous deposits are formed on the injectors. This inhibits the ability of the valve needle to move and the formation of a finely dispersed spray of fuel. The mixing system is thus disturbed, and increased emissions (particulates), higher fuel consumption, and lower engine power, result.

A range of substances are suitable as detergents in diesel engines: amines, imidazolines, amides, succinimides, polyalkyl succinimides or amines, and polyetheramines [70, 71]. The following mechanism of action has been proposed: the surface activity of the polar groups ensures the formation of a protective film, whereas the dispersing effect of the polymeric chain prevents agglomeration and deposit formation by smaller particles. In this way a "keep clean effect" can be achieved; however, for already dirty nozzles, a solubilization of formed deposits material is also necessary. The addition of detergents offers other advantages for diesel fuel, such as better storage stability and a certain amount of protection against corrosion [72].

Examples of suppliers are Infineum (Vectron), Octel (OMA), Oronite (ODA series), and Lubrizol.

5.2.3. Cold Flow Additives

Diesel fuel has the unfortunate property that crystallization of high molecular mass *n*-paraffins (waxes) can occur at low ambient temperature. These wax crystals can clog the fuel filter and lead to engine stoppage after a short period of operation. The tendency toward wax separation depends strongly on the composition of the diesel fuel, which is determined mainly by the crude petroleum used and the refinery structure [73]. In principle, the low-temperature behavior of a diesel fuel could be improved by removing the particularly critical *n*-paraffins. However, this would be unacceptable not only because of the high costs, but also because *n*-paraffins are very valuable diesel fuel components due to their good ignition performance and low smoking tendency [74].

Before the development of highly effective additives for improving low-temperature properties, owners of diesel vehicles were frequently forced to add up to 30 % regular gasoline to

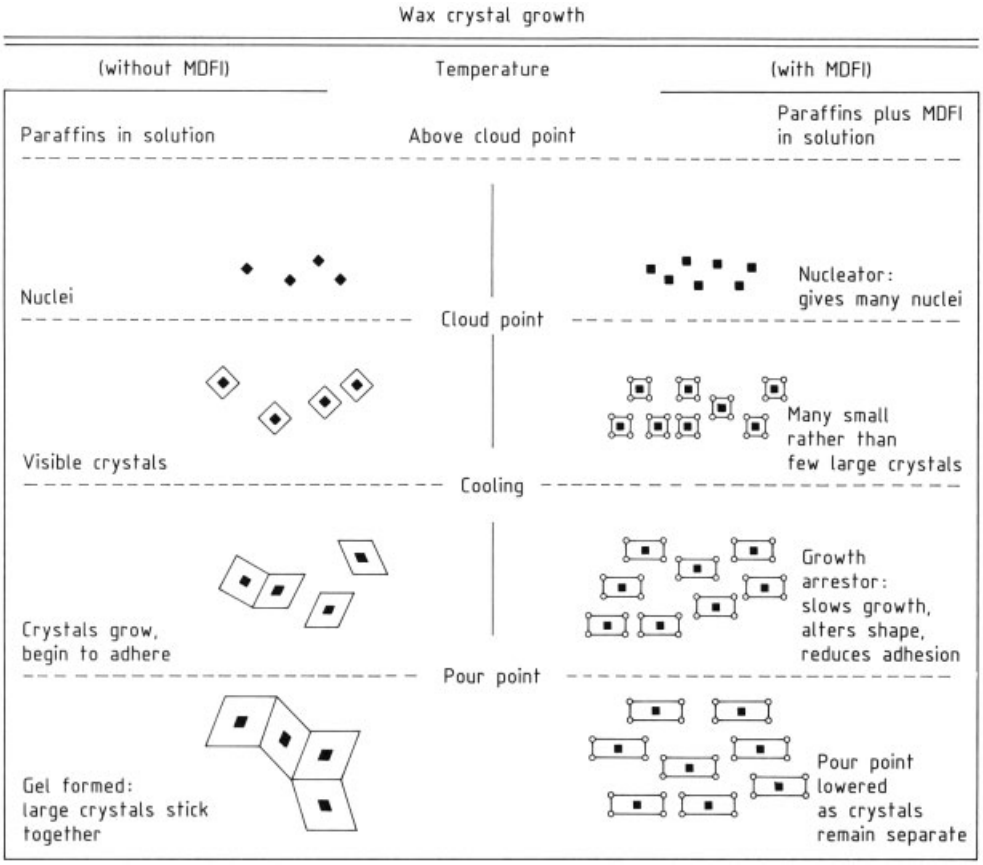

Figure 10. Mode of action of flow improvers [79]
MDFI = Middle distillate flow improvers

diesel fuel, risking engine damage and increased exhaust emissions.

To assess the low-temperature behavior of diesel fuel the following two laboratory test procedures are used:

1) **Determination of the Cloud Point** (EN 23015, ASTM D 2500). The cloud point is the temperature at which diesel fuel begins to become turbid at a defined cooling rate because of the separation of small wax crystals. The cloud point is especially important for the evaluation of refinery components in order to determine type and amount of cold flow additive that has to be added to the components. The cloud point does not define the operability of the final, additivated diesel fuel.
2) **Determination of the Cold Filter Plugging Point (CFPP)** (EN 116) [75, 76]. The CFPP is the limiting temperature at which, under defined conditions, a diesel fuel sample can no longer pass through a metal sieve within a specified length of time. Generally speaking the CFPP gives a reliable indication of the operability limit expected for a diesel fuel in modern vehicles. Since the development of wax antisettling additives, which drastically lower the CFPP, the operability of critical older vehicles (e.g., without heating of the fuel system) is not as low as suggested by the CFPP result.

Flow Improvers. Flow improver additives improve the cold flow properties of diesel fuels in a cost- effective way. Refineries are given the flexibility to use high cetane quality paraffinic product streams for diesel fuel, kerosene, otherwise necessary for achieving good low-temperature behavior of diesel fuel, for the production of jet fuel.

Flow improvers have been used increasingly since 1973. Their mechanism of action is based on two different effects [77]. First, the additive forms crystallization nuclei on reaching the cloud point, to which the wax crystals become adsorbed. Second, further crystal growth is slowed down by adsorption of the additive on the growing crystal surface. Thus, both the crystal size and the tendency to agglomerate are reduced (Fig. 10). The small crystals thus formed have a considerably lower tendency to block fuel filters because they (1) pass through the filters to some extent, (2) produce a still penetrable layer on the filter, and (3) dissolve again rapidly when the engine is warmed.

Flow improvers are effective only when they have been added to the fuel at temperatures well above the cloud point of the base fuel, i.e., at about + 40 °C, and preferably in the refinery. Addition after crystal formation is not effective [78]. Ash-free polymers are used as flow improvers, in the vast majority of cases ethylene – vinyl acetate (EVA) copolymers. The dosage with which a particular CFPP value can be achieved depends on the base fuel and is normally 50 – 500 ppm (Fig. 11).

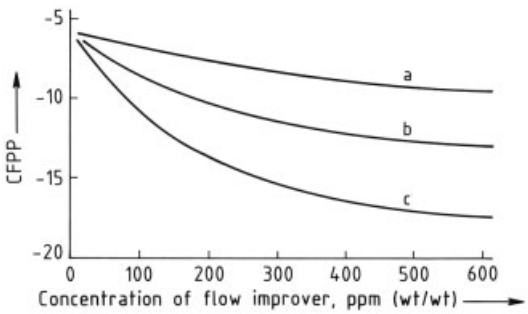

Figure 11. Effect of a flow improver on various base diesel fuels [54]
a) Poorest response behavior; b) Normal response behavior; c) Best response behavior

Examples of suppliers are BASF (Keroflux additives), Octel (OFI), Infineum, and Clariant (Dodiflow).

Wax Antisettling Additives (WASA). Wax antisettling additives keep small wax crystals in suspension and thus inhibit the formation of wax layers, which can rapidly lead to filter clogging [25]. This is achieved by their ability to slow the growth of wax crystals, which is even greater than that of flow improvers. According to Stokes' law, small particles have a considerably lower settling rate than large ones. By reducing the size of wax crystals by one-fifth, the sedimentation time can be extended from an hour to a day. Wax antisettling additives have been used since 1986 – generally in combination with flow improvers. Typical concentrations are 100 – 500 ppm. They are usually based on modified ethylene – vinyl acetate copolymers.

Examples of suppliers are BASF (Keroflux ES additives) and Infineum [79].

Cloud Point Depressants. The use of conventional flow improvers to lower the CFPP value generally has hardly any influence on the cloud point. However, polymer additives (based on ethylene – vinyl acetate, unsaturated esters, imides, or olefins) have been developed that permit additional lowering of the cloud point. Their mode of action is based on shifting the thermodynamic equilibrium by irreversibly decreasing the crystallization temperature. This particularly increases the solubility of higher molecular mass *n*-paraffins. With a dosage of 500 ppm the cloud point can be lowered by ca. 3 °C. The market importance of cloud point depressants is considerably smaller than that of flow improvers and wax antisettling additives [73].

5.2.4. Lubricity Additives

The intrinsic, natural lubricity of diesel fuel is reduced during the desulfurization process, because surface active components are destroyed during the hydrogenation. In the case of low and ultra-low sulfur diesel fuels the lubricity of the finished product can be too low, resulting in wear problems in the injection pumps. To ensure safe operation low and ultra-low sulfur diesel fuels have to be additivated with efficient lubricity additives. Examples of surface active substances are fatty acid derivatives which are added in the range of 50 – 200 mg/L.

The lubricity performance of fuels is evaluated in the high frequency reciprocating wear rig (HFRR) laboratory test (ISO 12156–1). A maximum wear (HFRR value) is defined and specified in EN 590.

Examples of suppliers are Octel (OLI), Ethyl (Hitec), Infineum, and Lubrizol (LZ).

5.2.5. Antifoam Additives

When a vehicle tank is filled with diesel fuel, severe foam formation can occur because of intensive mixing with air. This has unpleasant effects: either the time necessary to fill the tank completely is extended or the amount of fuel taken into the tank is considerably lower.

Silicon-containing antifoams, mostly polysiloxanes, have been developed and added to diesel fuel at a concentration of 10 – 50 ppm to avoid excessive foam formation. An effective antifoam additive not only depresses foam formation but also accelerates foam decomposition.

Examples of suppliers are Goldschmidt (Tegopren series), OSI Witco (MR additives), and Dow Corning.

5.2.6. Additives for Increasing Storage Stability – Antioxidants

During the storage of diesel fuel, aging and corrosion effects can occur in a way similar to gasoline (see Sections 5.1.3 and 5.1.4). The aging of diesel fuel also occurs via oxidation and radical polymerization reactions, which first cause darkening and later sediment formation [80]. The problem of unstable diesel fuel has increased particularly since ca. 1970 because higher proportions of diesel fuel components from cracking processes are in use [81]. These components have a higher content of unsaturated hydrocarbons than straight-run middle distillates. Middle distillates from cracking processes are generally subjected to hydrogenation ("hydrotreating"); however, additional treatment with antioxidants is often necessary.

In diesel fuel aging, sulfur- and nitrogen-containing substances can also be responsible for sediment formation to a considerable extent, although they sometimes have an oxidation-inhibiting effect. In addition, a high acidity (caused by carboxylic acids or phenols) accelerates aging catalytically. Sediment formation brought about by aging can clog the fuel filter; gumlike deposits can affect the flow of the fuel in the injection pump and the injectors.

The additives used to inhibit aging are similar to those used for gasoline (see Sections 5.1.3 and 5.1.4), namely sterically hindered phenols and phenylenediamines. Trialkylamines are also very efficient. The additive dosage is usually ca. 50 ppm; however, the efficiency depends on the particular diesel base fuel.

Since dissolved metal ions (especially copper) can also catalyze the aging of diesel fuel, ca. 5 ppm of a *metal deactivator* can be added.

Corrosion inhibitors (see Section 5.1.1), which generally consist of high molecular mass amines when used in diesel fuel, are also effective against aging because of their basicity.

The *detergents additives* added to diesel fuel principally to keep the injection nozzles clean are often effective in preventing diesel fuel aging. Additives for stabilizing diesel fuel are frequently used in the form of a multifunctional mixture at a concentration of ca. 100 ppm.

5.2.7. Dehazers

Through contact with water vapor in refinery processes and possible entry of water during transportation and storage, diesel fuel often becomes cloudy due to the presence of finely dispersed water droplets. Since such a cloudy, water-containing fuel would lead to considerable operational problems, the water droplets must have separated before the fuel is distributed. Especially in diesel fuels rich in aromatics, water separation may take a long time so that large tank capacities are necessary. The settling of water can be accelerated considerably by the addition of dehazers. These additives mainly lower the surface tension between water droplets, so that agglomeration to form larger, more rapidly sinking drops is achieved. Quaternary ammonium salts in a concentration of 5 – 50 ppm are effective dehazers.

Examples of suppliers are Rechem (ER and EXP additives), Petrolite (Tolad additives), and Nalco (Nalco).

5.2.8. Biocides (→ Biocides)

As explained in Section 5.2.7, the contact of diesel fuel with water cannot be ruled out totally.

Table 9. European standard for gasoline; extract from EN 228

	Regular	Super	Super Plus	Test method
ϱ(15 °C), kg/m^3	720 – 775	720 – 775	720 – 775	EN ISO 3675, EN ISO 12185
Knock rating				
RON	min. 91.0	min. 95.0	min. 98.0	Pr EN ISO 5164
MON	min. 82.5	min. 85.0	min. 88.0	Pr EN ISO 5163
Lead content mg/L	max. 5	max. 5	max. 5	Pr EN 237
Distillation range, total proportion distilled, vol % up to 70 °C:*				
Winter and intermediate	22 – 50	22 – 50	22 – 50	EN ISO 3405
Summer	20 – 48	20 – 48	20 – 48	
up to 100 °C:				
Winter and intermediate	46 – 71	46 – 71	46 – 71	
Summer	46 – 71	46 – 71	46 – 71	
Up to 150°C	min. 75	min. 75	min. 75	
Final boiling point, °C	210	210	210	
Vapor pressure, kPa *				EN 13016–1
Winter	60 – 90	60 – 90	60 – 90	
Intermediate and summer	45 – 60	45 – 60	45 – 60	
Benzene content, vol %	1.0	1.0	1.0	EN 238, EN 12177, Pr EN 14517
Aromatics content, vol %	max. 35.0	max. 35.0	max. 35.0	ASTM D 1319, Pr EN 14517
Olefin content, vol %	max. 21.0	max. 21.0	max. 21.0	ASTM D 1319, Pr EN 14517
Sulfur content, mg/kg	max. 50 or max. 10	max. 50 or max. 10	max. 50 or max. 10	EN ISO 20846, EN ISO 20847, EN ISO 20884
Oxygenates, vol %				
Methanol	max. 3.0	max. 3.0	max. 3.0	
Ethanol	max. 5.0	max. 5.0	max. 5.0	
2-Propanol (isopropyl alcohol, JPA)	max. 10.0	max. 10.0	max. 10.0	
2-Methyl-2-propanol (tert-butyl alcohol, TBA)	max. 7.0	max. 7.0	max. 7.0	
2-Methyl-1-propanol (isobutyl alcohol, JBA)	max. 10.0	max. 10.0	max. 10.0	
Ethers containing 5 or more carbon atoms	max. 15.0	max. 15.0	max. 15.0	
Other mono alcohols and ethers with a finalboiling point below 240 °C	max. 10.0	max. 10.0	max. 10.0	
Oxygen content, mass %	max. 2.7	max. 2.7	max. 2.7	EN 1601, EN 13132
Induction period, min	min. 360	min. 360	min. 360	EN ISO 7536
Copper corrosion visual rating	max. 1	max. 1	max. 1	EN ISO 2160

* mandatory limits vary depending on climatic conditions, figures shown are valid for central European climate.
* Regular grade and Super plus are standardized locally, figures shown are valid for Germany.

In some cases, water layers are formed at the bottom of the storage tank due to insufficient house keeping. At the interface between water and the diesel fuel, conditions are provided that allow the growth of microorganisms such as bacteria, yeast, and fungi. Infestation with microorganisms leads, on the one hand, to sediment formation which increases the danger of filter blocking. On the other hand, sulfate-reducing bacteria create a corrosive atmosphere that produces rust particles and can finally lead to tank destruction [82]. Through the addition of biocides, microbial infestation can be prevented or already affected tanks can be disinfected.

Microorganisms can develop resistance to biocides under certain circumstances, so biocides should be used only sparingly for specific purposes. Generally careful cleaning of the affected storage tank and consequent removal of water from the tank solve the problem.

A whole range of substances can act as biocides: formaldehyde derivatives, isothiazolones, triazines, *N,N*-methylenebis(5-methyloxazolidine), and boric acid derivatives.

Biocides are usually applied in a range of 50 – 400 ppm.

Biocides are supplied, e.g., by Schülke & Mayr, Norderstedt, Germany.

5.2.9. Antistatic Additives

When hydrocarbons are pumped rapidly, a buildup of static electricity can occur which, on discharge, may lead to fuel vapor ignition. In comparison to kerosene, this danger is relatively small for diesel fuel because of its lower volatility and its content of polar components that lead to increased conductivity. However, the use of small quantities of antistatic additives (up to 5 ppm) which, for example, contain calcium compounds is sometimes advisable where, very high pumping rates (e.g., at refineries, transfer, or truck refueling) can lead to the danger of static charge buildup. Electrical conductivity can be determined by means of a testing method ASTM D 2624, developed for jet fuel.

Antistatic additives are supplied, e.g., by Octel (Stadis).

5.2.10. Reodorants

The odor of diesel fuel is often regarded to be unpleasant, with the result that people can reject diesel engines as "dirty". If the skin or clothes are contaminated with diesel fuel, the unpleasant odor can remain for a long time because of its low volatility.

The unpleasant smell is caused particularly by certain sulfur-containing and unsaturated components. Desulfurization with hydrogen removes most of the unpleasant smelling ingredients. Modern low-sulfur fuels have generally a neutral smell so that specific reodorants are no longer required.

Reodorants are based mostly on nature-identical substances that are toxicologically harmless, such as vanillin or terpenes. They usually work by blocking the olfactory nerves and can be added in concentrations of 30 – 100 ppm.

Examples of suppliers are Haarmann & Reimer (Kompensols) and Bush Boake Allen (LG additives).

6. Fuel Standardization and Testing

The engine performance strongly depends on the physical and chemical properties of a fuel (see Chap. 3). To guarantee adequate fuel quality for a broad range of engine requirements, minimum specifications must be met, which are defined, for example, in *Europe* in EN 228 (for gasoline) and EN 590 (for diesel) (Tables 9 and 10).

Table 10. European standard for diesel fuel; extract from EN 590

		Test method
ϱ (15 °C), kg/m^3	820 – 845	EN ISO 3675, EN ISO 12185
Cetane number	min. 51	EN ISO 5165
Cetane index	min. 46	EN ISO 4264
Distillation range, total proportion recovered vol%		Pr EN ISO 3405
up to 250 °C	below 65	
up to 350 °C	min. 85	
95 % recovered	max. 360 °C	
Flash point °C	above 55	EN ISO 2719
Viscosity (40 °C), mm^2/s	2.00 – 4.50	EN ISO 3104
Filterability (CFPP), °C **		EN 116
Summer	max. 0	
Winter	max. − 20	
Intermediate	max. − 10	
Sulfur content, mg/kg***	max. 50 and max. 10	EN ISO 20846, EN ISO 20847, EN ISO 20884
Conradson carbon residue, mass %	max. 0.30	EN ISO 10370
Ash content, mass %	max. 0.01	EN ISO 6245
Copper corrosion visual rating	max. 1	ISO 2160
Oxidation stability, g/m^3	max. 25	EN 12205
Sediment, mg/kg	max. 24	EN 12662
Water content, mg/kg	max. 200	EN ISO 12937
Lubricity HFRR, µm *	max. 460	EN ISO 12156–1
Polycyclic aromatics, mass %	max. 11	EN 12916

* Protection against wear of injection pumps defined in the High Frequency Reciprocating wear Rig (HFRR).
** mandatory limits vary depending on climatic conditions, figures shown are valid for central European climate.
*** The specification of 10 mg/kg supports the introduction of "sulfur free" Diesel fuels.

In these specifications the great climatic differences within Europe (Sweden – Portugal, for example) are taken into account be defining appropriate limits for, e.g., volatility of gasoline and cold-flow performance of diesel fuels for each country.

The ASTM testing methods used in the *United States* are similar to the corresponding European (EN, ISO) methods. For all testing procedures a variation in measured values is unavoidable, and determination of a "true value" is

impossible. The terms "true value", repeatability, and reproducibility are precisely defined in EN ISO 4259.

The EN standards for gasoline and diesel fuel are mandatory. The quality of the fuels and their conformity with standards is, for example, tested on a random basis by the appropriate authorities in the Federal Republic of Germany and is followed up by court in the case of an offense. This is all the more important because not conforming to standards can cause damage to engines and provide economic advantages for disreputable suppliers.

7. Storage and Transportation

To transport fuels from the refinery to a vehicle tank a sophisticated infrastructure is necessary.

Storage. In the refinery or in gasoline depots, fuels that are ready for use are first placed in cylindrical, generally steel tanks for intermediate storage. To store volatile gasoline fixed roof tanks with *internal floating roofs* are generally used. The roof, which is sealed against the wall of the tank, can rise and fall with the liquid level. To avoid hydrocarbon emissions the tanks breathe into a sealed common system (→ Oil Refining, Chap. 4.2.1.1). For the storage of diesel fuel, *fixed roof tanks* are frequently used.

Small quantities of fuels can be stored in steel drums, steel cans or fuel-resistant plastic containers.

Transportation. Fuels can be transported *from* the refineries via pipelines, tank ships, rail tankers, or road tank trucks. In some countries mineral oil companies exchange base fuels from their refineries. The shortening of the transport routes allows cost saving and reduces safety risks.

In the refineries or depots, vapor recovery plants with a retention efficiency of ca. 99 % are sometimes employed for the filling of gasoline tankers [83]. When underground gasoline station tanks are being filled, displaced gasoline vapors are often returned to the tank trucks (vapor recovery stage 1). In Europe and the United States the emission of gasoline vapors is also controlled when car tanks are being filled by means of special dispensing nozzles (vapor recovery stage 2) [84].

Gasoline belongs to hazard class I, whereas diesel fuels are assigned to hazard class III because their flash point is > 55 °C. Hazard classifications for the transport of gasoline and diesel fuel are as follows:

Gasoline	
GGVS, ADR/RID	Class 3, number 3 b
IMDG Code	Class 3.1
UN no.	1203
ICAO/IATA-DGR	Class 3
Diesel fuel	
GGVS, ADR/RID	Class 3.3, number 31 c
IMDG Code	Class 3.3
UN no.	1202
ICAO/IATA-DGR	Class 3

8. Alternative Fuels

The search for alternative fuels is not new. During earlier efforts to introduce alternative fuels (e.g., during World War II or after the first oil crisis in 1973), concern about the extent of crude petroleum reserves or too great a dependence on oil-producing countries was predominant. Overcoming these problems was attempted by shifting fuel production from crude petroleum to other fossil energy sources. A true picture of the primary energy consumption can only be obtained by drawing up energy chains that take all conversion steps into account.

Liquefied Petroleum Gas (→ Liquefied Petroleum Gas). With liquefied petroleum gas (LPG) an additional fraction of crude petroleum (predominantly propane – butane) can be used as fuel for Otto engines. Because of the relatively expensive modification costs for vehicles and other disadvantages (higher volumetric consumption, lower engine power) use of LPG has become established in only a few places where taxation is favorable (e.g., the Netherlands, Italy, Japan, Australia).

Natural Gas (→ Natural Gas). Natural gas (predominantly methane) can also be used as an engine fuel either as compressed natural (CNG) or as liquefied natural gas (LNG). However storage in liquefied form requires considerable sophistication. Until now, natural gas is only used

Table 11. Important fuel properties for the performance of an Otto engine – comparison of methanol, ethanol, and regular gasoline [89]

	Methanol (CH_3OH)	Ethanol (C_2H_5OH)	Regular gasoline
Chemical composition, kg/kg fuel			
C	0.38	0.52	0.86
H	0.12	0.13	0.14
O	0.50	0.35	
Calorific value, MJ/kg	19.7	26.8	43.0
RON	ca. 114	ca. 111	ca. 92.6
MON	ca. 94	ca. 94	ca. 82.7
Stoichiometric air – fuel ratio, kg air/kg fuel	6.4	9.0	14.7

Table 12. Fuels from biomass

Raw material	Average yield of harvest, $t\ ha^{-1}\ a^{-1}$	Crude ethanol/ crude rape seed oil, $L\ ha^{-1}\ a^{-1}$	Acreage per volume unit of fuel, m^2/L	Energy per acreage unit $kW \cdot h\ m^{-2}\ a^{-1}$	Net energy profit [93] including production expense, %
Wood	10.0 [a]	1.500 [b]	6.7	0.88	0
Sugar beet	46.0	4.600 [b]	2.2	2.71	20
Potatoes	25.0	2.850 [b]	3.5	1.68	20
Wheat	4.5	1.760 [b]	5.7	1.04	10
Maize	5.0	2.000 [b]	5.0	1.18	20
Rape seed	2.5	1.250 [c]	8.0	1.26	35

[a] The amount harvested is related to energy farming.
[b] Crude ethanol.
[c] Crude rape seed oil.

as CNG on a broader scale in certain countries (Argentina, Italy, Russia). However, the relatively low carbon dioxide emission resulting from the high hydrogen content is favoring its use in the future. In Germany BMW has developed duel fuel vehicles for the use of CNG which are operated by fleet owners predominantly. Also a limited number of public service stations exist which offer CNG.

Fuels from Fischer – Tropsch Synthesis. *Synthetic gasoline* or *middle distillate* can be produced from coal or natural gas via synthesis gas from the Fischer – Tropsch synthesis [85]. This method, however, is often associated with a very poor overall energy efficiency and has been used in only a few special cases (Germany: World War II; South Africa: Sasol Process). However, new catalysts and different process conditions allow nowadays an improved energy efficiency, e.g., the Shell Middle Distillate synthesis process which produces a high cetane quality diesel fuel component from natural gas on small scale.

Methanol. Methanol is produced from synthesis gas [86], which can be obtained from crude oil, coal and even biomass. A number of experiments have been conducted on its use as a potential alternative fuel (Table 11). However, the use of almost pure methanol or of a 15 % mixture with gasoline has not yet found a practical application. On the technical side, problems in cold- and hotrunning performance and aldehyde emissions with cold engine require special attention whereas the specific power output and engine efficiency is advantageous in comparison to gasoline engines. From the economic point of view the high consumption and high engine adaption costs are a disadvantage [87 – 89].

Ethanol. The use of ethanol, which can be obtained form crude oil, coal and biomass, as an alternative fuel has been put into practice in Brazil. Both neat ethanol (hydrated at about 95 %) fuel and a 23 % blend in gasoline are marketed. Ethanol is produced mostly from biomass and could therefore potentially help reduce carbon dioxide emissions because of a closed carbon dioxide cycle. On the other hand catalytic exhaust gas treatment is also required even when ethanol is used because of other emissions (CO, HC, NO_x, and aldehydes).

Table 13. Properties of various vegetable oils in comparison with diesel fuel

Property	Fuels			
	Sunflower seed oil	Soybean oil	Rape seed oil	Diesel fuel*
Density	0.924	0.923	0.920	0.830
Viscosity (40 °C), mm^2/s	31.0	31.2	34.4	2.8
Sulfur, ppm	50	50	70	1500
Flash point, °C	196	220	202	60
Cetane number	33	27	40	53
Lower calorific value, MJ/kg	36.2	39.4	35.2	42.6
Appearance	slightly cloudy, yellow	clear, dark amber-colored	clear, amber-colored	clear, bright
Low-temperature behavior (limiting temperature for operability), °C	> 0	> 0	+ 15	− 22

* Typical values for Germany.

Table 14. Energy storage in motor vehicles – energy equivalent to 50 L of gasoline

	Total tank volume, L	Tank mass, kg	Tank filling time, min
Gasoline	55	47	2
Diesel fuel	50	47	2
Methanol	103	88	2
Ethanol	84	72	2
Hydrogen (liquid)	330	94	7 – 80
Metal hydride (FeTi)	385	1175	20
Electricity, NaS battery	3565	3800	600

From the viewpoint of technical applications, the use of ethanol is somewhat less problematic than that of methanol [90, 91]; however, the high production costs and the huge cultivation area for biomass have kept its use as a fuel within narrow limits (see Table 12). Ethanol was used in Germany in former times ("*Reichskraftstoff*": 1/3 ethanol, 1/3 benzene, 1/3 gasoline).

Fatty acid methyl esters (FAME) are produced from vegetable oils by esterification. Especially rape seed oil methyl ester (RME) is used as a fuel for diesel engines on a small scale in some countries (mainly France, also Austria [92] and Germany). In principle vegetable oil fuels offer the advantages of greenhouse gas emissions benefits because of the closed carbon cycle, new outlets for agricultural products, the possibility of usage in standard diesel engine with minimum modification, and good biodegradability. On the other hand they offer little improvement over conventional diesel fuels in view of local air quality. Although exhaust smoke is somewhat lower, regulated and nonregulated emissions are scarcely different form those of diesel fuel. The global impact of the potential reduction in carbon dioxide emissions is limited by production capacity of RME which is predicted to rise but will still be less than 10 % of the diesel market [93, 94].

Vehicles fueled with RME show reduced power due to the lower calorific value and the higher viscosity as compared with diesel fuel. The volumetric fuel consumption is higher because of the oxygen content of RME. Technically, the use of rape seed oil fuel is also associated particularly with an increased engine wear under certain circumstances (caused by deposit formation) and poor cold flow performance (Table 13).

Hydrogen can be used either as an engine fuel or to feed fuel cells for on board generation of electricity. Neither of these concepts has gained a commercial status so far, partly because of the difficulties connected with the handling and storage of hydrogen (see Tables 14 and 15). For detailed information see → Hydrogen, Chap. 9.3.3.3.

Table 15. Environmental compatibility and availability of various fuels

Fuel	Environmentally sensitive substances	Availability
Gasoline (with catalyst)	CO_2, ozone among others	worldwide
Diesel fuel	CO_2, particle, NO_x, ozone	worldwide
LPG	CO_2, among others	some countries
Methanol		
fossil	CO_2, formaldehyde	locally possible in short time
biomass	formaldehyde	?
Ethanol (biomass)	acetaldehyde	Brazil
Vegetable oils	particles, among others	possible locally to some extent
H_2		
fossil	CO_2, NO_x, among others	?
nonfossil		?
Electricity		
fossil	CO_2, NO_x, among others	worldwide
nonfossil		possible to some extent

Electricity. For the use of electricity as an alternative fuel, global environmental compatibility is again determined by the production method. Electricity production based on fossil energy sources also leads to carbon dioxide emission as well as to the formation of regulated harmful substances. In addition, the total energy efficiency (based on the energy chain from production to utilization) of electrically driven vehicles is clearly lower than that of conventional fuels. However, the use of electric vehicles for areas of higher emission sensitivity (indoor areas, city centers) could become increasingly important. The *storage* of electricity in the vehicle represents a great problem because of the weight and the size of the batteries [95]. The development of efficient batteries such as the sodium – sulfur battery led to an increase in the energy density by a factor of ca. 4 in comparison with the lead – acid battery. However, compared to liquid fuels currently in use, there is still a considerable disadvantage in energy storage (and therefore in vehicle range). A solution to this problem is the onboard production of electricity by fuel cells which are currently under development (→ Fuel Cells, Chap. 4.1 Zero Emission Vehicle). Ultimately these can be fueled with

Table 16. Evaluation of alternative systems

	Gasoline		Diesel	LPG [b]	Natural gas [c]	Hydrogen [c]	Bioethanol [c]	Methanol [c]	Fuel cell/hydrocarbon Fuel	Electricity	
	Without catalytic converter	With catalytic converter									
Suitability	0	0	0	−	−	−	0/−	−/0	0		− −
Availability	0	0	0	−	0	− − −	− −	− −	0		− − [d]
Economy	0	0	+	+/−	−	− − −	− −	−	+		− −
Infrastructure	0	0	0	− −	− −	− − −	− − − [e]	− − −	0		− −
										from fossil fuels	from nuclear and solar technology
CO	0	++	++	+	+	+++	+	+	+++	+	+++
HC	0	++	++	+	+	+++	+ [f]	+ [f]	+++	+	+++
NO_x	0	++	+	++	++	+	+	+	+++	0	+++
Particulates	0	+	− −	+	++	+++	+	+	+++	−	+++
CO_2	0	0	+	0	++	+++	++	0/− −	+	− −	+++
Nonregulated emissions	0	++	++	++	++	+++	−	− −	+++	0	radioactivity from nuclear production?

0 similar; + somewhat better; ++ clearly better; +++ very much better; − somewhat worse; − − clearly worse; − − − very much worse.

[b] Dual system without catalyst.

[c] With catalyst.

[d] For electricity: 0.

[e] Except Brazil.

[f] Aldehydes: −.

hydrogen. Because of the handling and availability problems use of intermediate fuels such as hydrocarbons and methanol are currently discussed. The fuel cell technology has also the advantage of a noticeably improved efficiency compared to the combustion engine.

Hybrid Concepts. The introduction of alternative fuels is associated with considerable changes in the infrastructure that cannot be carried out in short term. For this reason, dualfuel hybrid concepts can greatly ease the introduction of new fuels. Up to now, prototypes for the alternative use of gasoline – methanol ("flexifuel" concept) [87], gasoline – hydrogen [96], and diesel – electricity [2, 97] have been developed. Gasoline – LPG and gasoline – CNG vehicles are commercially available in some areas, however, their market share is small mainly because of the additional initial costs and the reduced load capacity because of the second tank. At present, considerable discrepancy still exists between environmental compatibility and other important criteria such as suitability, availability, economic factors, and infrastructure (Table 16) [98].

9. References

1. O. von Fersen: *Ein Jahrhundert Automobiltechnik*, VDI-Verlag, Düsseldorf 1985.
2. U. Seiffert, P. Walzer: *Automobiltechnik der Zukunft*, VDI-Verlag, Düsseldorf 1989.
3. W. Dabelstein, A. Reglitzky, K. Reders, N. Lucht: *100 Jahre Kraftstoffe für den Straßenverkehr*, Shell Technischer Dienst, Hamburg 1986.
4. J. S. Meurer, *VDI-Ber.* **714** (1988) 1 – 30.
5. F. Pischinger, *Verbrennungsmotoren*, RWTH Aachen, 1980.
6. F. Wankel, *Motortech. Z.* **25** (1964) no. 12, 489.
7. F. Huf, *Automobil Rev.* **49** (1961) 23.
8. K. Groth: *Grundzüge des Kolbenmaschinenbaues*, Groth, Hannover 1975.
9. A. Pierburg: *Vergaser für Kfz-Motoren*, 4th ed., VDI-Verlag, Düsseldorf 1970.
10. H. Scholl, *Motortech. Z.* **34** (1973) 4.
11. R. Krebs, B. Stiebels, L. Spiegel, E. Pott: "*FSI Gasoline direct injection Engine for the Volkswagen Lupo*", Lecture held at the 21. Wiener Motorensymposium, 2000.
12. Bosch: *Kraftfahrtechnisches Handbuch*, 20th ed., VDI-Verlag, Düsseldorf 1987.
13. K. Reders, W. Dabelstein, A. Reglitzky, W. Lange, *Mineralöltechnik* (1989) no. 12, ibid. no. 13, 38.
14. G. Sitkei: *Kraftstoffaufbereitung und Verbrennung im Dieselmotor*, Springer Verlag, Berlin – Heidelberg 1964.
15. M. Gairing, M. Fortnagel, F. Scherer, Lecture held at the Technische Arbeitstagung Hohenheim (1984), *Mineralöltechnik* (1984) no. 12 – 13.
16. P. Heinze, *Mineralöltechnik*, (1987) no. 4, 5.
17. H. Waldmann, G. H. Seidel: *Kraft- und Schmierstoffe*, Walter de Gruyter Verlag, Berlin 1979, pp. 986 – 993.
18. R. L. Allen, A. Reglitzky, *Qualität von Ottokraftstoffen*, Shell Kraftstoff-Symposium, Hamburg Sept. 1985, Shell Technischer Dienst, Hamburg.
19. W. W. Lange, A. Schütze, H. Krumm, K. Reders: *Ottokraftstoff Lexikon*, Shell Technischer Dienst, Deutsche Shell, Hamburg.
20. H. Krumm: *Fortschritte der Fahrzeugtechnik, Stadtauto/Mobilität Ökologie, Ökonomie, Sicherheit*, Vieweg Verlag, Braunschweig 1995.
21. K. Groth: *Brennstoffe für Dieselmotoren heute und morgen*, Expert-Verlag, Ehingen 1989, pp. 195 ff.
22. H. May, M. Regitz, "Studie Dieselkraftstoffe," Forschungsvereinigung Verbrennungskraftmaschinen Heft 430, Frankfurt 1989.
23. *CONCAWE Rep.* 1987, no. 10.
24. K. Weidmann, H. Menrad, K. Reders, R. Hutcheson, *SAE-Paper* **881 649**, 1988.
25. D. Rickeard, A. Ramsdem, N. Thompson, *SAE-Paper* **890 048**, 1989.
26. B. Riediger: *Die Verarbeitung des Erdöles*, Springer Verlag, Berlin – Heidelberg 1971.
27. W. Dabelstein, A. Reglitzky: *Umweltschutz neue Determinante für die Energiepolitik*, Verlag TüV, Rheinland, Köln 1988, pp. 167 – 221.
28. W. Keim, A. Behr, G. Schmitt: *Grundlagen der Industriellen Chemie*, Verlag Salle + Sauerländer, Frankfurt 1986, p. 133.
29. Deutsche BP: *Das Buch vom Erdöl*, 5th ed., Reuter und Klöckner, Hamburg 1989, pp. 161.

30. F. Kienert: "Ein Beitrag zur Untersuchung von Reformingkatalysatoren im integralen Rohrreaktor und einem neuen Kreisgasreaktor," Dissertation RWTH Aachen, 1989.
31. Refining Process Handbook, *Hydrocarbon Process.* **63** (1984) 67.
32. G. Zürn, K. Kohlhaase, K. Hedden, J. Weitkamp, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) 62 ff.
33. H. Bruderreck, *Erdöl Erdgas* **102** (1986) 463 ff.
34. D. Rohe, *Chem. Ind.* **7** (1986) 570 – 572.
35. W. Danaher, L. Palmer, *Fuel* **67** (1988) 1441 – 1445.
36. M. J. Van der Burgt, I. Van Klinken, S. T. Sie, "The Shell Middle Distillate Synthesis Process," Lecture Washington D.C., Nov. 11 – 13 (1985).
37. J. Falbe, W. Hasserodt: *Katalysatoren, Tenside u. Mineralöladditive*, Thieme Verlag, Stuttgart 1978, pp. 221.
38. Ingenieur conseil BEICIP, *Fuel Additives in Western Europe*, Nov. 1988.
39. E. Bamberg, K. Reders, "Verbesserung von Ottokraftstoffen durch Additive," Lecture held at the Technische Arbeitstagung Hohenheim, Shell Technischer Dienst, 1986.
40. J. Lodwick, *Chem. Rundsch.* **20** (1967) no. 1, 1 – 3; no. 2, 5 – 7; no. 3, 8 – 9.
41. R. Tupa, *SAE-Paper* 872 113, 1987.
42. B. Taniguchi, R. Peyla, G. Parsons, K. Hoekmann, D. Voss, *SAE-Paper* 861 534, 1986.
43. J. Beson, P. Yaccarino, *SAE-Paper* 861 533, 1986.
44. B. Bitting et al., *SAE-Paper* 872 117, 1987.
45. F. Carraciolo, R. Stebar, *SAE-Paper* 872 111, 1987.
46. K. Starke, "Ablagerungen in den Ansaugsystemen von Ottomotoren," Lecture held at the Technische Arbeitstagung, Hohenheim 1977, Shell Technischer Dienst, Hamburg 1977.
47. C. Jewitt, G. Bostick, V. Kersey, *SAE-Paper* 872 114, 1987.
48. J. Fabri, "Chemische Untersuchungen zum Mechanismus und den Einflußparametern der Ablagerungsbildung auf den Einlaßventilen von Ottomotoren," Dissertation RWTH Aachen, 1988.
49. G. Lepperhoff, J. Schommers, O. Weber, H. Leonhardt, *SAE-Paper* 872 115, 1987.
50. E. Dimitroff, A. Johnston, *SAE-Paper* 660 784, 1966.
51. E. Dimitroff, J. Moffit, R. Quillian, *Trans. ASME J. Lubric. Technol.* **91** (1969) no. 7, 406 – 416.
52. R. Tupa, D. Koehler, *SAE-Paper* 861 536, 1986.
53. Coordinating European Council, Tentative Test Method No. CEC F-03-T-81, (1981).
54. K. Owen (ed.): *Gasoline and Diesel Fuel Additives, Critical Reports on Applied Chemistry*, vol. **25**, John Wiley and Sons, New York 1989, pp. 19 ff.
55. H. Rath, K. Starke, H. Vogel, "Wirkung optimierter Kraftstoffadditive," Lecture held at the Technische Arbeitstagung, Hohenheim 1988, published.
56. R. Lewis, H. Newhall, *SAE-Paper* 830 938, 1983.
57. J. Schommers, "Ursachen der Ablagerungsbildung auf den Einlaßventilen von Verbrennungsmotoren," Dissertation RWTH Aachen, 1986.
58. M. Gairing, *Mineralölrundschau* **34** (1986) no. 11, 209 – 215.
59. P. Polss, *Hydrocarbon Process.* **52** (1973) 61 – 68.
60. J. Johnson: *Antioxidants*, Noyes Data Corporation, New York – London 1975.
61. Y. Yonekawa, Y. Nakamura, N. Okamoto, *Sekiyu Gakkaishi* **24** (1981) no. 2, 85 – 91.
62. P. Meyers, O. Uyehara, R. DeYoung, DOE/CS/50 020; NTIS-No. DE 82 012 576, 1981.
63. J. Mokita, B. Sturgis, Proc. World Pet. Congr. (1955) Sect. 4 F.
64. I. Spilners, J. Hedenburg, *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **26** (1981) no. 2, 632 – 638.
65. H. Giere, B. Nierhauve, *Mineralöltechnik* **6** (1988) no. 6, p. 5, pp. 14 – 19.
66. D. Sutton, *SAE-Paper* **860 222**, 1986.
67. R. Olree, D. Lenane, *SAE-Paper* **840 108**, 1984.
68. W. W. Lange, H. Krumm, K. Reders: *Dieselkraftstoff Lexikon*, Shell Technischer Dienst, Deutsche Shell, Hamburg.
69. Ting-Man Li, R. Simmons, Presentation at the 21st International Symposium on Combustion, München 1987.
70. X. Montagne, D. Herrier, J. Guibet, *SAE-Paper* 872 118, 1987.
71. D. Sutton, M. Rush, P. Richards, *SAE-Paper* 880 635, 1988.
72. T. Russel, Jugoma Professional Publication, Dubrovnik 1987.

73. T. Coley, F. Rossi, M. Taylor, J. Chandler, *SAE-Paper* 861 524, 1986.
74. C. v. Eberan-Eberhorst, L. Geldern, F. Pischinger, G. Seidel, *Erdöl Kohle Erdgas Petrochem.* **37** (1984) 69 ff.
75. T. Coley, L. Rutishauser, H. Ashton, *J. Inst. Pet.* **52** (1966) no. 510, 173 – 189.
76. Coordinating European Council Report P-01–74, 1974.
77. J. Zielinski, F. Rossi, A. Stevens, *SAE-Paper* 841 352, 1984.
78. British Technical Council (BTC), Publication BTC/F1/79, 1979.
79. Paramins Corp.: *Fuel Additives Manual-Wax Crystal Modifiers*, Abingdon 1988.
80. R. C. Tupa, C. J. Dorer, *SAE-Paper* 861 179, 1979.
81. H. Bechert, H. Hauschildt, G. Heilmann: "Künftige Dieselkraftstoffqualitäten und ihre Auswirkungen auf das motorische Betriebsverhalten," *Motortech. Z.* **43** (1982) no. 6, 237 – 241.
82. R. N. Smith (ed.): Papers presented at Microbiology Committee Conference Oct. 1986, Institute of Petroleum, 1986.
83. DGMK-Tagungsbericht 8902, Fachbereichstagung "Umwelt," Köln 1989.
84. R. Sander, *MOT-Technik* **20** (1989) 116 – 122.
85. D. Osterroth: *Von der Kohle zur Biomasse*, Springer Verlag, Berlin 1989, pp. 149 ff.
86. F. Asinger: Methanol-Chemie- und Energierohstoff, Springer Verlag, Berlin 1986, pp. 57 – 95.
87. Bundesministerium für Forschung und Technologie (ed.): *Alternative Energien für den Straßenverkehr – Methanol*, Verlag TüV Rheinland, Köln 1983.
88. F. Asinger: Methanol-Chemie und Energierohstoff, Springer Verlag, Berlin 1986, pp. 96 – 155.
89. H. Menrad, A. König: *Alkoholkraftstoffe*, Springer Verlag, Berlin 1982.
90. The Swedish Commission for Oil Substitution: *Strategy for Alternative Fuels*, DSI 1982 : 13, Stockholm 1983.
91. Europäische Zusammenarbeit auf dem Gebiet der Wissenschaftlichen und Technischen Forschung (COST): *Alternativkraftstoffe für Straßenfahrzeuge,* EUCO-COST 304/4/85, 1985.
92. Österreichischer Verein für Kfz-Technik: *Praxiserfahrungen mit Rapsölmethylester als Ersatz für Dieselkraftstoff*, Wien, Nov. 1988.
93. Bundesrat Drucksache 388/89, *Nachwachsende Rohstoffe*, Juli 1989.
94. E. Fiala, *Automobilindustrie* **3** (1989) 291 – 297.
95. K.-N. Regar, W. Strobl, R. Heuser, "PKW-Antriebe mit Elektro- und Wasserstoffspeicher," Tagungsband 2. Aachener Kolloquium Fahrzeug- und Motorentechnik (1989) 463 – 483.
96. W. Strobl, W. Peschka, *VDI-Ber.* **602** (1987).
97. P. Walzer, "Antriebskonzepte für künftige Energie- und Umweltprobleme," Tagungsband 2. Aachener Kolloquium Fahrzeug- und Motorentechnik (1989) 795 – 809.
98. J. Fabri, W. Dabelstein, A. Reglitzky, "Chancen alternativer Kraftstoffe unter besonderer Berücksichtigung der Umweltverträglichkeit," Lecture held at the Technische Arbeitstagung, Hohenheim April 1990, Shell Technischer Dienst, Hamburg 1990.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Mutagenic Agents

Standard Article

Wolfgang Pfau[1]

[1]Umweltmedizin Hamburg e.V. and Institute of Toxicology, Hamburg University Medical School, Hamburg, Germany

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a16_755.pub2
Article Online Posting Date: April 30, 2004

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Mutagenesis
2.1. Gene Mutations
2.1.1. Point Mutations
2.1.2. Frameshift Mutations
2.2. Chromosomal Aberrations
2.3. Genome Mutations
3. DNA Repair
4. Mutations and Cancer
5. Mutagenicity Tests
5.1. Microbial Tests
5.2. Eukaryotic In Vitro Tests
5.2.1. Mammalian Gene Mutagenicity Tests
5.2.2. In Vitro Cytogenetic Assays
5.3. In-Vivo Tests
5.3.1. Micronucleus Assay
5.3.2. Indicator Tests
5.3.3. Transgenic Animals
6. Testing Strategies
7. Metabolic Activation
8. Types of Mutagenic Agents
9. Conclusions

[Top of Page]

1. Introduction

Mutations are changes in the composition of the genetic information that are passed on to the daughter cells. Mutagenic agents are, thus, chemical or physical agents that damage the genome and are also referred to as *genotoxic agents*.

However, changes in the genome are also the basis of evolution and may, in rare occasions, confer a selective advantage. Since evolution is a slow process, this may be observed in our time scale only in rapidly proliferating organisms, such as bacteria or viruses, which may acquire resistance towards antimicrobial or antiviral drugs by mutations. The overall majority of mutations are, however, either silent or detrimental.

There are a number of hereditary diseases that have been assigned to specific mutations in relevant genes (such as sickle cell anemia, hereditary breast cancer). Also, hereditary enzyme deficiencies have been associated with a mutated genetic code for the respective enzymes.

As an instrument in molecular biology site directed mutagenesis is a means to elucidate proteins by systematically introducing changes in the amino acid sequence.

Because mutagenicity is linked to carcinogenicity it is import to determine if a substance has a mutagenic potential. Genotoxic carcinogens are generally considered to have no safe dose, no threshold, and therefore stringent regulations are imposed on mutagenic agents minimizing human exposure to these agents to the lowest (reasonably) achievable level.

[Top of Page]

2. Mutagenesis

Mutations in a *germ line cell* result in changes in the genetic code, that are manifest in all cells of the off-spring and will be hereditary. However, these are very rare events. More common are mutations in *somatic cells* which may, in the worst case, lead to cancer or, if introduced in the developing embryo, to teratogenic effects.

Mutagenic agents may induce premutagenic lesions. These are transformed into mutations during DNA replication or by incorrect DNA repair processes (see below).

There are different types of mutations resulting from different types of damage in the DNA: gene mutations (point and frameshift mutations), chromosomal aberrations, and genome mutations.

2.1. Gene Mutations

Gene mutations are small changes in the DNA at the level of the bases. Upon replication a wrong nucleotide may be incorporated into the daughter strand and as a consequence the mutation is then passed on to the following generations of cells.

2.1.1. *Point Mutations*

By substitution of nucleotides the sequence of the bases is changed. This may be of no consequence when this mutation occurs in a nontranscribed region of the genome or when a silent mutation is introduced (because the genetic code is degenerate, different codons coding for the same amino acid). However, changing a codon may result in changes of the amino acid sequence of the resulting protein or the introduction of a stop codon.

Endogenous or exogenous agents may induce chemical changes in the bases, for example oxidative damage may result in deamination of adenine by nitrous acid. This deamination results in the incorporation of cytosine during replication and within two cell divisions a A–T base pair is converted into a G–C pair.

The most common damage induced by mutagenic agents is the covalent binding of a reactive, electrophilic metabolite to the nucleophilic DNA bases. Small *alkylating agents* bind to the O^6- or N7-position of guanine, the O^4-position of thymidine, or the N6, N3 or N7 of adenine. These small DNA adducts may lead to a mispairing upon replication and, thus, to the introduction of a point mutation (see Fig. 1.).

Figure 1. Normally guanine pairs with cytosine. Methylation at the O^6-position freezes the enolic form of guanine and may result in introduction of thymine in the opposing strand. Upon replication one of the daughter cells will, thus, contain a TA base pair instead of CG.

Polycyclic aromatic amines and nitroarenes are metabolically transformed and the ultimate electrophilic agents, the nitrenium ions bind to the C8-position of guanine and sometimes to the N^2- or N^6-exocyclic amino groups of the purine bases (Figure 2) (⟶ Carcinogenic Agents – Metabolic Activation).

Polycyclic aromatic hydrocarbons bind generally, upon metabolic activation, to the exocyclic amino groups of purine bases (Figure 2) but may also, after a one-electron oxidation, bind to the N7-position. Binding to the N7-position results in a weakening of the *N*-glycosidic bond and may lead to a loss of the base, resulting in an apurinic site.

Bivalent electrophilic agents, such as unsaturated carbonyl compounds, dicarbonyl compounds, or dihalogenated aliphatic agents may form cyclic adducts, binding to both exocyclic and endocyclic nitrogen atoms in the purine bases or cytidine.

2.1.2. *Frameshift Mutations*

Addition or elimination of bases to the sequence may result in a change of the reading frame. Generally, this leads to more serious changes in the resulting protein. A classical example is the congenital phenylketonuria, where the enzyme responsible for the conversion of phenylalanine to tyrosin is inactive due to a (hereditary) frameshift mutation.

Chemicals with a planar aromatic structure may intercalate into the DNA helix. This, but also covalent DNA adducts may result in a slip or an addition of one base during replication or DNA repair.

2.2. Chromosomal Aberrations

When parts of a chromosome end up in a different location in the same or a different chromosome, are lost or amplified, this is called chromosomal aberration. These types of damage are caused by breakages, deletions, or rearrangements during the cell cycle. The places where breaks or exchanges occur are usually not randomly distributed over the chromosome.

Chromatid-type aberrations include micronuclei, deletions, and interchromatid exchanges. *Chromosome-type aberrations* include bicentric chromosomes, terminal deletions, and acentric rings.

Microscopically, chromosomal aberrations may be observed after fixation and staining of the chromosomes during the metaphase stage of the cell cycle.

Agents inducing chromosomal aberrations are called *clastogens*, these include ionizing radiation and numerous chemical agents.

An example of a genetic disease resulting from structural chromosome aberrations is the Philadelphia chromosome. Here, a portion of the long arm of chromosome 22 is translocated to chromosome 9, an aberration, that is observed in the bone marrow cells of 90% of patients with chronic myeloid leukemia.

2.3. Genome Mutations

Genome mutations are also referred to as numerical chromosomal aberrations and involve a change in the number of chromosomes. *Euploidy* refers to a multiple of the complete set of chromosomes (haploid, diploid, triploid). *Aneuploidy* describes the situation when the number of some chromosomes is different from the normal diploid set or exact multiples of it. This may result from non-disjunction of chromosomes at anaphase or chromosome loss during cell division with an uneven distribution of chromosomes between daughter cells.

Aneuploidy may be induced by agents affecting the spindle apparatus, such as the drugs colchicine, taxol and vinca alkaloids. Also aneugenic environmental chemicals have been identified, for example high levels of the pesticide trichlorfon in the diet of mothers may be responsible for a cluster of congenital abnormalities including Down's syndrome, observed in a Hungarian village in 1990.

The frequency of aneuploidy in humans appears to be high: At conception a rate of about 16–50% is estimated. Most of the affected embryos are not viable and are spontaneously aborted. About one in 200 newborns displays aneuploidy, and a strong increase in this frequency is observed with the mother's age.

Trisomy of chromosome 21 results in Down's syndrome characterized by physical abnormalities, physical and mental growth retardation. This most common autosomal abnormality occurs with a frequency of about 1 in 600 children.

[Top of Page]

3. DNA Repair

A large variety of DNA repair enzymes continuously scan the DNA and replace damaged nucleotides. Therefore, despite the frequent occurrence of spontaneous or induced lesions (estimated to amount to several thousand in a typical mammalian cell each day) the number of mutations occurring is low and the genetic information is largely maintained intact.

DNA repair depends on the presence of a separate copy of the genetic information in each strand of the DNA double helix. A lesion on one strand can therefore be cut out by a repair enzyme and a good strand resynthesized using the undamaged strand as a template. There are a number of repair systems that identify and amend specific lesions in the DNA strand prior to the replication. For example, alkyl residues at the guanine O^6-position are being removed by a dedicated methyl transferase. In *base excision repair* an altered base is removed by a DNA glycosylase enzyme, followed by excision of the resulting sugar phosphate. In *nucleotide excision repair* a small region of the strand surrounding the damage is removed from the DNA helix as an oligonucleotide. In both cases the small gap left in the DNA helix is filled in by the sequential action of DNA polymerase and DNA ligase.

While some of these repair enzymes are constitutively expressed, others are inducible, in an adaptive response these enzymes are expressed when a DNA damage occurs.

DNA double strand breaks may be repaired by homologous recombination using the intact homologous chromosome as a template (this may lead to sister-chromatid exchanges) or by rejoining the ends of two DNA strands thereby losing several nucleotides at the joining point and introducing mutations.

In order to transform a DNA lesion into a mutation it is necessary that the DNA polymerase system is by-passing this lesion and is introducing a (wrong) base opposite the lesion in the daughter strain. The molecular mechanisms of these translesion DNA synthesis with a family of error-prone polymerases have been identified most recently.

[Top of Page]

4. Mutations and Cancer

Hereditary predispositions towards certain types of cancer have been linked to mutations in one of the two copies (alleles) of specific genes that humans carry in every cell. For example, a tumor suppressor gene may be damaged, such as BRCA1 in hereditary breast cancer. Induction of a mutation that damages the intact copy of this tumor suppressor will lead to the development of a tumor at a specific organ site.

Similarly, it is believed that mutations in oncogenes may lead to the development of malignant tumors. Generally, several genetic changes are necessary for a malignant tumor to develop. This high incidence of specific mutagenic events is possible only because in tumor cells the systems ensuring the integrity of the genome are also lost or damaged.

Also, chromosomal mutations (aneuploidy) play a major role in the induction and development of malignant neoplasms.

It has been proposed that the number and pattern of mutations, the so-called mutation spectrum, may be an indicator of the damaging agent. For example the fingerprint of mutations detected in several thousand of cancer cases in a specific gene (p53) is indicative of polycyclic aromatic hydrocarbons in cases of lung cancer, whereas in cases of skin cancer the majority of mutations resembled the damage induced by UV-irradiation.

[Top of Page]

5. Mutagenicity Tests

A large number of test systems can be employed to determine the mutagenic potential of an agent. Basically, these systems include in vitro tests in bacteria, yeast, or mammalian cells and in vivo short-term tests in experimental animals. The extent of testing required for new substances depend on the intended use (pesticide, drug, food additive, etc.) and the production volume.

Several tests have been standardized and the testing procedures are laid down in OECD Guidelines, EU test guidelines, or by the Environmental Protection Agency of the USA.

The principles of these tests are described in detail in ⟶ Toxicology – *In Vitro* and *In Vivo* Short-term Tests for Genotoxicity.

5.1. Microbial Tests

Most commonly the so-called *Ames test* is performed. Here gene mutations (point mutations or frameshift mutations) are recorded in a battery of mutated *Salmonella typhimurium* strains that require histidine. With or without addition of external metabolic activation (added homogenate of rat liver) reverse mutations are detected by regaining histidine-prototrophy (OECD Guideline 471, EU test guideline B.13/14, ⟶ Toxicology – Microbial Tests for Mutagenicity).

5.2. Eukaryotic In Vitro Tests

Whereas microbial tests are easier to conduct, assays involving mammalian cells are generally considered more relevant for the prediction of animal or human in vivo effects and also allow detection of clastogenic or aneugenic potency.

5.2.1. *Mammalian Gene Mutagenicity Tests*

Most commonly the V79 cell line (Chinese Hamster) or the mouse lymphoma cell line L5178Y $TK^{+}/^{-}$ are employed. These tests may be conducted with or without external metabolic activation. In principle, mutations in the hypoxanthine-guanine phosphoribosyl transferase gene (HPRT) are detected in V79 cells. This gene, which is a available only in one copy because it is located on the X chromosome, confers susceptibility towards purine derivatives such as 6-thioguanine. Similarly, the mouse lymphoma cell line L5178Y $TK^{+}/^{-}$ possesses only one copy of the gene coding for thymidine kinase, an enzyme involved in nucleotide metabolism and leading to susceptibility towards the pyrimidine derivative trifluorothymidine. Inactivation of these enzymes (by mutation) leads to insensitivity towards these agents. Cells with a mutation survive and may be scored as growing clones.

In the mouse lymphoma test small mutant colonies are considered the result of chromosome damage while large mutant colonies are considered the result of gene mutations (OECD Guideline 476, EU test guideline B.17, ⟶ Toxicology – Eukaryotic Tests for Mutagenicity).

5.2.2. *In Vitro Cytogenetic Assays*

Chromosomal aberrations are analyzed microscopically in cell populations in metaphase. Mostly cells of the Chinese hamster such as V79 and CHO are utilized but also human lymphocytes and mouse lymphoma cells. Metaphase chromosomes are stained and inspected for rearrangements. In the *sister-chromatid exchange assay*, chromosomes are differentially stained and inspected for the presence of exchanges between sister chromatids.

In recent years the *micronucleus assay* has become more popular: Micronuclei (MN) may be formed during the cell division and may arise from a whole lagging chromosome (aneugenic event leading to chromosome loss) or an acentric chromosome fragment detaching from a chromosome after breakage (clastogenic event). The MN test may be performed with different cultured mammalian cells and involves the treatment of cells with the actin polymerization inhibitor cytochalasin B to identify cells that divide as binucleate cells. Micronuclei are detected microscopically. By means of kinetochore-antibody staining or fluorescence in situ hybridization they may be differentiated into micronuclei containing a whole chromosome or an acentric chromosome fragment, indicating aneugenic or clastogenic activity, respectively, of the test compound (OECD Guideline 473, EU test guideline B.10, ⟶ Toxicology – Tests for Chromosome Aberrations (Cytogenetic Assays)).

5.3. In-Vivo Tests

5.3.1. *Micronucleus Assay*

As the in vitro micronucleus test, this assay tests mutagenic effects on the chromosome and genome level, detecting clastogenic and aneugenic agents. The in vivo micronucleus test is generally performed in mice, and the tissues most often assessed for frequency of micronuclei are bone marrow and peripheral blood. Erythrocytes are the cells that are scored in the bone marrow or the blood for presence of micronuclei. Because they are constantly produced from precursor cells and the nucleus is expulsed from the cell immediately after full differentiation is accomplished, erythrocytes are an ideal cell type for a micronucleus test. If a stem cell is damaged by a chemical and a micronucleus is formed as a consequence of this damage, the micronucleus remains in the cell after the main nucleus has been expulsed and can be observed microscopically.

Mouse bone-marrow micronucleus tests typically employ one to three treatments with the chemical under study; with doses extending up to the maximum tolerated dose. The route of administration in these short-term tests is usually either intraperitoneal injection or oral gavage. Based on the cell cycle and maturation times of the erythrocytes, harvesting of the

bone marrow usually occurs 24 h after the final dosing. At that time, about 50% of the erythrocytes in the bone marrow are immature, newly formed erythrocytes, and these are the cell types that are checked for presence of micronuclei. The bone marrow is flushed from the femurs and spread onto slides. The slides are dried, fixed, and stained with a fluorescent DNA-specific stain that easily illuminates any micronuclei that may be present. 2000 polychromatic erythrocytes (PCEs, or immature erythrocytes) are scored per animal for frequency of micronucleated cells. In addition, the percentage of PCEs among the total erythrocyte population in the bone marrow is scored for each dose group as a measure of toxicity. Short-term studies that give positive results are repeated to confirm the response.

Critical parameters in analyzing micronucleus test data are the number of animals per dose group (a minimum of three is required); dose levels and number of doses administered; route of administration; tissue and cell type analyzed; sample time (interval between last dosing and harvesting of cells for analysis);, frequencies of micronucleated cells in the negative and positive controls; and the results of the statistical analyses. In the case of a negative result it is important to establish that the test chemical reached the target organ (bone marrow), which may be verified by signs of bone-marrow toxicity.

5.3.2. *Indicator Tests*

Assays that measure damage to DNA may be performed either in vitro or in experimental animals. These tests include detection of *unscheduled DNA synthesis* (UDS), indicating the repair of damaged DNA. The test is most commonly performed in primary hepatocytes in vitro or in hepatocytes following in vivo treatment (OECD test guideline 486, EU test guideline B39).

The formation of DNA adducts may be analyzed by ^{32}P-postlabeling analysis. Here covalent modifications are analyzed chemically with a sophisticated procedure involving enzymatic DNA hydrolysis, enzymatic labeling with ^{32}P-phosphate and chromatographic separation of labeled modified nucleotides (⟶ Toxicology – Test Systems Providing Indirect Evidence for DNA Damage.)

The analysis of DNA strand breaks may be accomplished by single cell electrophoresis (comet assay). Here nuclei are exposed to an electric field under DNA denaturing conditions in an agarose gel. Damage to DNA results in DNA fragments that move faster in the gel. This can be visualized upon fluorescence staining and microscopically scored.

The advantage of short-term in vivo assays such as ^{32}P-postlabeling analysis and comet assay is that a number of organs may be assayed in parallel and genotoxic activity testing is not confined to one organ such as the liver in UDS tests or bone marrow in the case of the micronucleus assay.

The more common application of these, liver and bone marrow specific, tests is based on the assumption that mutagens exhibit a general genotoxicity which would be detectable in all organs. However, this is contradicted by the organ specificity observed by some chemical carcinogens.

5.3.3. *Transgenic Animals*

Assays based on transgenic animals such as Muta Mouse or Big Blue Mouse (and rats) allow determination of genotoxic potency in any organ. These commercially available animals have been genetically modified so that each cell contains a shuttle vector containing several copies of the lactose operon of *E. coli*, lacZ or lacI, respectively. These serve as indicators of gene mutations. Briefly, these animals are treated with a test compound, animals are sacrificed, DNA is isolated after a certain delay period, and the shuttle vector is transferred to *E. coli*. Detection of mutations in the shuttle vector sequences is accomplished by white/blue selection of growing bacteria in plates containing appropriate indicators.

These assays show a high degree of correlation with conventional in vitro gene mutation assays and also mutagenic activity was detected in target organs of numerous carcinogens. While these assays are detecting mutations in nongenomic DNA and may not detect chromosomal or genome mutations, there are new developments under way such as the gpt delta mouse, for example, that may be capable of detecting large deletions induced by clastogens.

[Top of Page]

6. Testing Strategies

Depending on the type of substance there are guidelines and legislation, for example in the European Union, that regulate the requirements for genotoxicity testing.

Basic testing for genotoxicity includes one gene mutation assay (mostly a bacterial assay) and one in vitro mammalian assay (most commonly the in vitro chromosomal aberration test). If these tests are negative, there is no further testing required for existing chemicals and cosmetic ingredients.

A third in vitro test (mammalian cell gene mutation test) is required for plant protection products, biocides, food additives, food contact materials, and human pharmaceuticals.

For human pharmaceuticals and veterinary drugs for food producing animals, basic testing includes a bacterial gene mutation assay, mammalian cell assay (mouse lymphoma assay or in vitro chromosome aberration), and an in vivo assay (testing chromosomal aberration or micronucleus formation).

New chemical substances require testing depending on tonnage of production.

If these basic tests are consistently negative the test substance is considered nonmutagenic. However, further testing is required in the case of positive findings in one or all basic assays conducted.

Generally, in vitro tests are preferred to tests in animals in line with animal welfare considerations. However, in vivo test are more important in the final assessment of a test substance, for examples a positive result in a microbial assay is generally overruled by negative outcome of one or more in vivo assays.

In order to assess a test substance the judgement of an experienced toxicologist is required.

[Top of Page]

7. Metabolic Activation

Strictly speaking most so-called mutagenic chemicals are not mutagenic themselves but, in order to exhibit their genotoxic activity require metabolic activation. By enzymatic transformation reactive metabolites are formed, such as α-hydroxy derivatives of *N*-nitrosamines, hydroxylamines from aromatic amines and nitroarenes, epoxides from hydrocarbons, or radicals that are capable of binding to the DNA. Covalent binding to DNA results in DNA adduct formation (Fig. 2.). These adducts are mostly removed by repair mechanisms but may result in the introduction of mutations (⟶ Carcinogenic Agents)

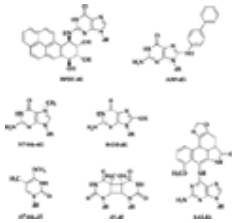

Figure 2. Prototypes of DNA adducts: Major deoxyguanosine-adducts formed by benzo[a]pyrene (via its diol-epoxide, BPDE-dG), 4-aminobiphenyl (C8-ABP-dG), hydroxyl radicals (8-OH-dG) and *N*,*N*-dimethylnitrosamine (N7-Me-dG). Alkylating agents may also modify thymidine such as O^4-Me-dT. The thymidine dimer formed is the major lesion formed by UV-light and the N^2-deoxyadenosine adduct is induced by the herbal carcinogen aristolochic acid I (AAI-dA).

UV- or visible light may also serve as an activating agent. In assays investigating photomutagenicity treated cells are exposed in vitro to a source of UV(and/or visible) light. A number of examples of agents inducing gene mutations upon photoactivation have been reported. These agents include polycyclic aromatic hydrocarbons, and fluoroquinolones. Also clastogenic activity may be induced by photoactivation for example in the cases of 8-methoxypsoralen or chlorpromazine. These effects are especially relevant for ingredients of cosmetic products.

[Top of Page]

8. Types of Mutagenic Agents

The most prominent mutagenic agents are also carcinogenic and are described in Carcinogenic Agents. Computer programs are available that, based on biological data, assess chemical structures for functional elements (“alerting structures”), which may act as mutagens. (⟶ Carcinogenic Agents – Modeling). These programs are useful to scan large libraries of chemicals being tested early in high throughput drug development, but even highly sophisticated calculation methods are presently not able to replace biological testing.

Members of the following classes of chemical compounds have been identified as mutagenic:

- *N*-Nitroso compounds
- Benzene
- Polycyclic aromatic hydrocarbons
- Nitroaromatic compounds
- Aromatic amines
- Halogenalkyls
- Aldehydes
- Alkylating agents
- Metals

In addition, a number of natural compounds and mixtures show mutagenic potential. Mutations can also be induced by oxygen damage. *Physical mutagenic agents* include nonionizing and ionizing radiation.

Aneugenic agents with no direct genotoxic activity are listed below. These may act by inhibition of repair enzymes, cell cycle control or apoptosis related proteins, tubulins of the mitotic or meiotic spindle or interference with detoxification or activating enzymes:

- Benomyl
- Cadmium chloride
- Carbendazim
- Chloralhydrate
- Colchicine
- Diethylstilbestrol
- Griseofulvin
- Hydroquinone
- Vinblastine

[Top of Page]

9. Conclusions

In addition to the application as short-term test to detect the mutagenic potency of test chemicals, assays such the micronucleus test, comet assay or ^{32}P-postlabeling are applied as biomarkers in molecular epidemiological studies. Increased levels of DNA adducts have been observed in cohorts occupationally exposed, e.g., to polycyclic aromatic hydrocarbons or styrene, to aflatoxins in the diet, to tobacco smoke, or to genotoxic drugs (including chemotherapeutic agents or herbal drugs).

In vivo animal experiments reflect the human situation only to a certain extent but generally the predictive value of animal assays is believed to be quite high. In contrast, in vitro assays are highly artificial and several aspects such as metabolic activation and toxicokinetics are not included. Therefore, data from in vivo assays are considered more relevant for the evaluation of a test chemical. In vitro assays are, however, indicative of the mechanism of action of a mutagen.

Antimutagenic agents inhibit or reduce the activity of mutagens. These may act by inducing detoxifying mechanisms, by antioxidative or receptor antagonism.

Co-mutagens are not mutagenic but may in combination with other agents increase mutagenic activity. Here, induction of activating enzymes, inhibition of detoxifying mechanisms, or formation of mutagenic metabolites is involved.

There are also compounds that are positive in several short-term tests for mutagenicity but failed to induce tumors in animal carcinogenicity studies. Examples are flavonoids such as quercetine or *p*-phenylendiamine. This may relate to high doses, that have been tested in vitro but are too toxic to be administered in vivo. However, these cases are highly disputed in the scientific community.

On the other hand there are carcinogenic agents that act via a nongenotoxic mechanism of action, such as mitogenic agents, tumor promoters, or hormones (⟶ Carcinogenic Agents).

Most genotoxic agents inducing gene mutations are also clastogenic. There are aneugenic agents that act solely by inhibiting the synthesis and assembly of the mitotic spindle, with no direct activity on the DNA, such as the methyl benzimidazole carbamate pesticides benomyl and carbendazim. Other indirect mechanism may involve inhibition of DNA repair (e.g., action of cadmium).

Based on theoretical considerations, *indirect genotoxins* are believed to be thresholded, i.e., there exists a safe exposure level. DNA damaging agents (*direct genotoxins*) are considered to act without a threshold, i.e., already a single molecule could induce mutation and cancer albeit with a exceedingly small but finite probability. Therefore, it is important to identify the mechanism by which a mutagenic agent is exerting its activity.

[Top of Page]

References

1. H. Marquardt, S. G. Schäfer, R. O. McClellan, F. Welsch (eds.): *Toxicology*, Academic Press, San Diego 1999.
2. Niesink, DeVries, Hollinger (eds.): *Toxicology*, CRC Press, Boca Raton 1996.
3. *BgVV: Strategies for genotoxicity testing of substances*, 2002.
4. Kirsch-Volders *et al.*, *Toxicol. Lett*. **140–141** (2003) 63 – 74.
5. Parry *et al. Mutagenesis* **17** (2002) 509 – 521. Links

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Molecular Modeling

DONALD B. BOYD, Indiana University-Purdue University at Indianapolis (IUPUI), Indianapolis, IN, United States

1. **Definitions** 1
2. **Purpose of Modeling** 1
3. **Brief History** 2
4. **Methodologies of Molecular Modeling** 3
4.1. **Computer Graphics** 3
4.2. **Databases** 4
4.3. **Conformational Modeling** 5
4.4. **Quantum Mechanical Modeling** 5
4.5. **Force Field Modeling** 6
4.6. **Statistical Modeling** 7
4.7. **Homology Modeling** 8
4.8. **Drug Design Modeling** 8
5. **Economic Aspects** 9
6. **References** 11

Molecular modeling is used to understand the characteristics of molecules and to design new molecules in industry and elsewhere. The computer-based techniques of molecular modeling include molecular mechanics, molecular graphics, and many other methodologies of computational chemistry. Most large research-based companies involved in molecular science, particularly pharmaceutical companies, use molecular modeling daily.

1. Definitions

In the narrowest sense used by some chemists, molecular modeling refers to conformational energy calculations and interactive manipulation of three-dimensional (3D) molecular structures on a computer screen. For instance, an early definition of molecular modeling was offered by TROST: "Molecular mechanics, [which is] the empirical computation of conformational energy, combined with computer graphics" [2]. This definition reflects the perspective of many organic chemists who want the computer to help them better understand the 3D shapes of molecules. HOPFINGER, a computational chemist, suggested a somewhat broader definition: "The generation, manipulation, and/or representation of molecular structures and associated physicochemical properties" [3]. From the perspective of a chemist in industry (DuPont), an even broader definition was set forth: "Virtually anything which is done to depict, describe, or evaluate any aspect of the properties or structure of a molecule that requires the use of a computer" [4].

The terms computational chemistry and molecular modeling are intertwined and are often used interchangeably in industry. Computational chemistry has been broadly defined as "those aspects of molecular research that are expedited or rendered practical by computers" [5]. Definitions for molecular modeling and computational chemistry have been discussed on the Internet (see, e.g., [6]). Deliberations of an IUPAC working party [7] resulted in a recommendation to define computational chemistry loosely as "a discipline using mathematical methods for the calculation of molecular properties or for the simulation of molecular behavior. It also includes, e.g., synthesis planning, database searching, and combinatorial library manipulation". The IUPAC group defined molecular modeling as "the investigation of molecular structures and properties using computational chemistry and graphical visualization techniques in order to provide a plausible three-dimensional representation under a given set of circumstances".

2. Purpose of Modeling

Modeling involves working with what is usually a simplified description of a system. Despite the simplifications, the model should capture the essence of a particular object, property, or process of interest. The models help scientists sort out complex relationships between variables. Being able to devise and use good

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a16_a35.pub2

models can help rationalize experimental data. Modeling can be used to estimate values of molecular properties in cases where experiment is difficult or impossible [8].

It is important to keep in mind that because of the mathematics involved, modeling requires attention to quantitation and detail. When molecules are being generated, manipulated, represented, depicted, etc., it is important to strive for realistic molecular structures and associated physicochemical properties. Although it is fairly easy to display a molecular structure on a computer screen, the model should closely represent an experimentally determined structure, so that any conclusions that are derived from the model are useful. State-of-the-art hardware and software make it easy for even the novice to create a structure on a computer screen, but does it mean anything? It may not be immediately apparent from just looking at a structural depiction that some bond lengths are too short or too long, for instance. A bench chemist may be used to writing chemical structures without the hydrogen atoms, but to minimize the quantum mechanical energy of a molecular structure leaving off hydrogen atoms will lead to useless results. Proper use of the models requires an appreciation of quantitative details and the underlying theory.

3. Brief History

Models of molecules have been around for a long time. Chemists have used hand-held molecular models to help them visualize conformation and understand stereochemistry. These models were built of cork, wood, metal, or plastic. Well-respected types, such as Dreiding or CPK, were once familiar in most chemistry laboratories. Because of the prevalence of computers today, some young students may not have ever heard of these models unless they happen to look in some old books and journals. Today computers perform many of the functions of the hand-held models and can do much more.

Theoretical chemistry had its origins in the quantum mechanics developed in the 1920s. However, it was not till the late 1950s that the early slow computers became available to research chemists. Two camps of quantum chemists developed [9]. One camp was interested in rigorous applications of the theory, which of necessity were confined to small molecules of only a few atoms, too small to be of interest to most industrial chemists. This ab initio (from first principles) research was mainly of interest to theorists, spectroscopists, and physicists. The other camp was willing to forego theoretical purity to render the methods applicable to molecules large enough to be of interest to a wide body of chemists. The latter, so-called semiempirical approach began to be applied to molecules of biological and commercial interest.

Simultaneously, organic chemists, starting with WESTHEIMER, ALLINGER [10], and others, recognized the importance of understanding the conformations of organic molecules and developed empirical methods for computing the relative energies of molecular structures. Likewise, protein chemists, such as SCHERAGA [11], became interested in empirical modeling because it allowed study of the conformations of polypeptides and proteins. The empirical or molecular mechanics methods were faster than either ab initio or semiempirical quantum mechanical methods and opened the door to computer simulations of large biomolecules, including enzymes, nucleic acids, and materials. The use of molecular mechanics spread in the 1970s and continues to grow [12, 13].

Other physical organic chemists were interested in the lipophilicity of molecules because of the relevance of this property to drug action. Starting in the 1960s, computational methods were developed for estimating the partitioning of solutes between aqueous and nonpolar solvents [14]. These methods have been used frequently to quantitate the relationship of biological activity and chemical structure.

Visionary scientists in the pharmaceutical industry recognized in the 1960s that theoretical modeling could add value to their company's research programmes and convinced their managers to hire one or two computational chemists in the 1960s and 1970s. However, it was not until the 1980s that industry broadly recognized the need to hire scientists to do molecular modeling. The field blossomed because of four factors:

1) Mini- and supermini-computers, particularly the VAX series from Digital Equipment Corporation.

2) High-resolution computer graphics terminals and workstations, such as the Evans and Sutherland Picture System.
3) User-friendly, multifunctional computational chemistry software packages.
4) The promising results of the initial modeling efforts.

Today the use of molecular modeling continues to grow due to:

1) Ever more powerful workstations (running UNIX) and personal computers (running Windows or Linux) at low relative cost.
2) Modern computational chemistry software with graphical user interfaces (GUI).

All major research-based pharmaceutical companies and many biotechnology companies employ computational chemists for molecular modeling. Large companies may have hundreds of scientists using molecular modeling. For a more detailed history of the use of computers and software in the pharmaceutical industry, see [15]. In addition, molecular modeling groups can be found in the agrochemical, chemical, and other industries.

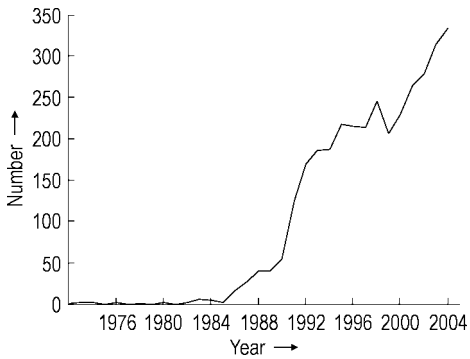

Figure 1. A search of the Chemical Abstract Service CAPLUS database indicates that as of late 2005 more than 3100 original research papers and nearly 400 book chapters and review articles are indexed as relevant to the two terms molecular modeling and computer. The number published annually is plotted.

Thousands of papers have been written about the methods and uses of molecular modeling. The growth in the annual number of publications can be seen in Figure 1. Use of the term molecular modeling with respect to computers increased in the 1980s, accelerated in the 1990s, and continues to grow. The most popular venues for these publications have been the Journal of Medicinal Chemistry, the Journal of Molecular Graphics and Modelling, and the Journal of Computer-Aided Molecular Design. Many books and reviews of molecular modeling continue to appear, see, e.g., [16 – 28].

4. Methodologies of Molecular Modeling

Molecular modeling entails a wide variety of complex, computer-based methods. These methods include ones for determining properties, accessing prior knowledge about them (databases), and analyzing data. Brief qualitative descriptions of some of these methods are presented.

Standard molecular modeling programs [29] are in use at many companies and universities, whereas other software packages, especially those specialized for molecular structure database management and drug design, are found mainly in the pharmaceutical industry because of their expense. Compilations [30, 31] have presented brief descriptions of programs, information about developers and vendors, and resources on the World Wide Web (WWW). More recent information about software can be found with the Google search engine on the WWW. A number of databases (e.g., PubChem) and software tools are in the public domain. Again, a search of the WWW will reveal both free and commercial resources.

4.1. Computer Graphics

The most important component of molecular modeling is computer graphics because this is the interface between the user and the computer. Molecular structure being the universal language of chemists, a user sitting at a computer can draw a molecule to be modeled on the screen using a “point and click” paradigm. Program options can be set up through a graphical user interface (GUI). After a calculation (such as described in the sections below) is complete, the results may be examined visually on the computer screen.

The complexity of molecular models makes high resolution, color, and stereo graphics highly

desirable, if not essential [32]. The simplest graphical depiction of molecules has lines representing bonds connecting points representing atoms. More complex graphics representing molecules as solid objects with colored or translucent surfaces help to convey molecular properties that pertain to each region of the molecule. Examples of computer graphics are shown in Figures 2 [33] and 3 [34 – 37].

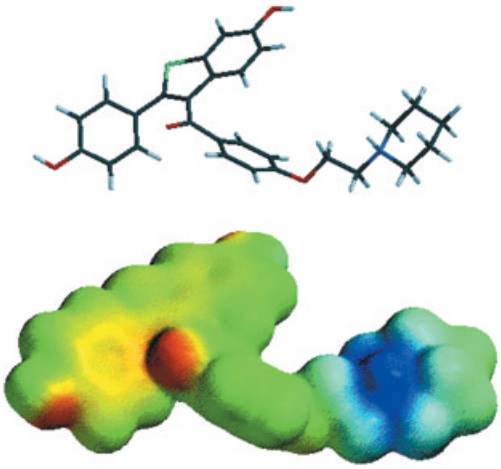

Figure 2. Molecular graphics of raloxifene (Evista), a drug for preventing osteoporosis and other health problems in women. Atoms are color-coded by element: carbons in black, oxygens in red, hydrogens in light blue, etc. The upper image is a stick figure; the lower image shows the electrostatic potential color-coded on an outer contour of the electron density distribution. Red denotes regions with high concentration of negative charge that would strongly attract a positive probe point charge; regions with a slight excess of electron density are in yellow; regions with a slight deficiency of electron density are in green; regions that would strongly repel a positive probe point charge are in blue.

Computer graphics is also used to visualize the results of statistical analyses. Complex response surfaces can be color-coded to show how several variables relate to each other [38]. The quantum chemical and molecular simulations methods (see below) are computationally intense and generate large quantities of data; computer visualization turns the data into something meaningful to chemists.

4.2. Databases

Almost all pharmaceutical and chemical companies have their own proprietary databases of compounds that have been synthesized or purchased for testing in various biological assays. In addition, there are commercial databases of compounds, such as the Chemical Abstracts Service's Registry database, and MDL Information Systems' Available Chemicals Directory (ACD). Industrial chemists use these databases to answer questions such as where can a compound be purchased, or has a chemical structure already been synthesized, or what compounds exist that are similar to a structure of interest?

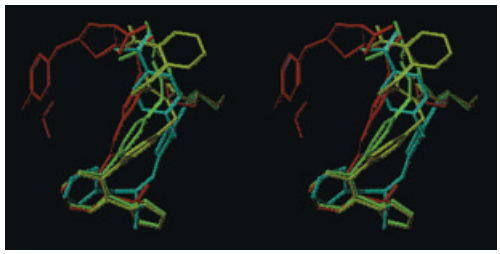

Figure 3. Stereo image of a possible superposition of four nonpeptide angiotensin II receptor antagonists.
a) losartan (green); b) eprosartan (light blue); c) Lilly's LY301875 (red); d) 3-*n*-butyl-5-[(2-carboxybenzyl)thio]-4-[[2′(1*H*-tetrazol-5-yl)biphenyl-4-yl]methyl]-4*H*-1,2,4-triazole (yellow). This overlay is based on alignments of the alkyl chains (right), aromatic rings (bottom), acidic groups (tetrazole or sulfonic acid) (bottom), and hydrogen bonding groups (top).

Database management tools specifically designed to handle chemical skeletons are used for rapid storage and retrieval. There are two types of molecular structure database: two-dimensional and three-dimensional. The 2D type stores atoms (chemical elements) and connectivity information (which atoms are bonded in a molecule). The 3D type stores, in addition, the x, y, z Cartesian coordinates of each atom in a molecule.

Structures determined by X-ray crystallography [39] or NMR spectroscopy [40] are stored in 3D databases. International repositories of the atomic coordinate data are available [41], such as the Cambridge Structural Database (CSD) for small organic and organometallic compounds, and the Protein Data Bank (PDB) and the Nucleic Acid Database for biomacromolecules. The structures can be used as starting points for molecular modeling experiments.

4.3. Conformational Modeling

Shape and energy are two molecular properties by which chemists understand how compounds interact and react. The properties are thus important in the pharmaceutical and agrochemical industries, where compounds must be found that will bind specifically to a receptor mediating a biological effect. To bind to a receptor, a molecule must have the shape — or the ability to adopt the shape — required for fitting into the receptor site. Most molecules are flexible, and at room temperature they are constantly undergoing conformational changes. The lowest energy structure is the global minimum one, but it may have no relevance to the conformation adopted upon receptor binding. As more ligand – receptor structures become solved by X-ray crystallography, it has become clear that is it not uncommon for the bioactive conformation of a ligand to be several kJ/mol higher in energy than the global minimum conformation. This situation arises because, when a ligand interacts with a receptor site, additional interatomic forces, not present in aqueous solution, occur that can induce the ligand to adapt its shape for better fit. The other forces arise from electrostatics, hydrogen bonding, van der Waals interactions, and hydrophobic effects. However, if too much energy is required to force a molecule to adopt a conformation required for receptor recognition, it is unlikely that the molecule will bind well.

The process of exploring all the conformations a molecule can adopt is conformational searching [42]. This involves molecular modeling techniques to map out the potential energy surface of a molecule, which describes the energy of the molecule as a function of the spatial location of its atoms. One type of conformational searching is systematic searching, where, as the name implies, the torsional angles at the rotatable bonds are varied systematically in increments. Another type is the Monte Carlo method, in which the geometrical variables are varied randomly, and the energy of each random conformation is calculated [43]. These and other methods for conformational searching tend to be computationally intensive and become prohibitively lengthy for molecules with more than about ten rotatable bonds. Energy evaluation during a search can be done with a force field or quantum chemical method (see below), although the force field methods are preferred because of their much greater speed and comparable accuracy.

Once the conformations have been determined, then molecular shape comparisons are possible. Molecular modeling offers several ways to compare and superpose molecular structures: alignments can be made by least-squares fitting of atoms or of projections to hypothetical hydrogen bonding points. Molecules can also be aligned based on their "surfaces" (e.g., the van der Waals surface) [44] or a computed field (measuring, e.g., steric, electrostatic, or lipophilic effects). Overlaying molecules is easy with computer models and is one of the most common applications of modeling.

4.4. Quantum Mechanical Modeling

Quantum mechanics (QM) is embodied in the Schrödinger equation, $H\Psi = E\Psi$, where H is a mathematical operator of kinetic and potential energy terms, Ψ is the wavefunction describing the positions of particles (electrons and nuclei) in a system, and E is the energy of the system [45, 46]. In molecular orbital (MO) theory, which has been the most common implementation of quantum mechanics used by chemists, electrons are distributed around the atomic nuclei until they reach a so-called self-consistent field (SCF), that is, until the attractive and repulsive forces between all the particles (electrons and nuclei) are in a steady state, and the energy is at a minimum. By computing the energy of different nuclear configurations, the lowest energy structure can be found. Such geometry optimization is useful because the bond lengths and bond angles that give the lowest energy are usually closest to what is observed experimentally. It has been pointed out that reliable reproductions of potential energy surfaces of flexible molecules sometimes need more stringent SCF and optimization convergence criteria during the calculations [47].

In ab initio MO calculations, very few approximations are made [48], and in principle they have the best ability to give agreement with experimental molecular properties. The electron density is represented by distributing it among certain standard sets of mathematical functions, called basis sets [49]. The different levels of ab

initio theory are referred to as "theoretical models" [50]. Unless the basis set is adequate, there is no great advantage of doing a calculation ab initio. Sometimes electron correlation (the phenomenon of electrons avoiding each other) and relativistic effects must be included in a calculation to get good results. Of the various quantum methods available, ab initio calculations are preferred, but unfortunately these calculations are sometimes too slow for molecular modeling applications of interest to industrial chemists. In certain situations, however, a molecule with several hundred atoms can be treated by an ab initio method [51, 52]. Large molecular systems are best left to one of the force field approaches (see Section 4.5).

In semiempirical MO calculations, additional approximations are made, such as treating only the outermost (i.e., valence) electrons while assuming the core (inner) electrons neutralize part of the nuclear charge [53 – 55]. Parameters in semiempirical theories are set to reproduce as closely as possible selected experimental molecular properties of well-studied compounds. However, parameters that work well at reproducing a small set of data are usually not robust enough to reproduce a broad range of properties and compounds. Semiempirical methods are sometimes used for "intermediate-sized" molecules (20 - 200 atoms). Semiempirical methods can sometimes predict molecular properties adequately, at least for similar molecules in which the effects of systematic errors are minimized.

A third implementation of QM is density functional theory (DFT) [56 – 59]. It is particularly useful in materials applications and is increasingly used for determining molecular properties. DFT is popular because it offers a compromise: it is faster than ab initio methods and more accurate than semiempirical methods. In DFT, the energy of a molecule is determined from the electron density, not from an MO wavefunction. Whereas ab initio theory includes the kinetic energy of the electrons and the potential energies of electron-electron repulsion and electron-nuclear attraction, DFT also includes an energy term called exchange-correlation that improves the result.

Electronic structure calculations are suited for predicting a variety of molecular systems (ground states, excited states, and transition states) and properties (optimum bond lengths and bond angles, conformational preferences, ionization potentials, electron affinities, charge distributions, hydrogen bonding, etc. [60, 61]). Molecular similarity can be calculated based on electronic structure [62]. Many types of reactions have been modeled (see, e.g., [63]). For modeling an enzymatic reaction, a chemically reasonable subset of atoms in the active site can be treated quantum mechanically and the rest of the system can be treated empirically [64]. This is the so-called quantum mechanics/molecular mechanics (QM/MM) approach.

Although the charge on an atom is not an experimental observable, charge distributions are of great interest to chemists [65, 66]. From a wavefunction, one can calculate the molecular electrostatic potential (MEP), which is an energy of attraction or repulsion experienced by a hypothetical unit charge as it moves in the vicinity of a molecule [67]. The MEP provides a model of how one molecule looks to another as they approach, such as when a ligand is approaching a receptor.

4.5. Force Field Modeling

In molecular mechanics (MM), as well as in molecular dynamics (MD) and Monte Carlo (MC) simulations, an empirical force field (FF) describes how the energy of a molecule changes as a function of bond lengths l, bond angles θ, and torsional angles ϕ. Atoms are treated as points in space except for the van der Waals term, which represents the atoms as spheres with soft surfaces. A set of energy terms and associated parameters is called a force field (FF) [68]. Simpler FFs assume harmonic force constants. The energy is a combination of bond stretching/compression, bond angle bending, torsional, van der Waals, and electrostatic forces that tend to hold the atoms at their equilibrium positions [69 – 73].

$$E_{\text{total}}=E_b+E_\theta+E_\phi+E_\tau+E_{el}+E_{vdw}$$

$$E_b=\sum\nolimits_{\text{bonds}} k_b(l-l_0)^2$$

$$E_\theta=\sum\nolimits_{\text{bond angles}} k_\theta(\theta-\theta_0)^2$$

$$E_\phi=\sum\nolimits_{\text{dihedral angles}} k_\phi[1+\cos(n\phi-\delta)]$$

$$E_r = \sum_{\text{improper angles}} k_\tau (\tau - \tau_0)^2$$

$$E_{el} = \sum_{i,j>i} \frac{q_i q_j}{\varepsilon r_{ij}}$$

$$E_{vdw} = \sum_{i,j>i} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

Here the force constants are the various parameters k, the "natural" bond lengths and angles are indicated by the subscript zero, ε is the dielectric constant, r_{ij} is the interatomic distance between atoms i and j, and "improper" angles τ refer to a mathematical trick to keep sp^2 atoms from deviating too far from planarity. Partial charges are denoted by q. n is 1, 2, or 3 depending on the periodicity of bond rotation, and δ determines where the minima occur. A_{ij} and B_{ij} are parameters determining the depth and distance of the energy minimum in a van der Waals interaction, which is attractive at long range and repulsive at short distances. The van der Waals and Coulombic interactions are summed between nonbonded atoms, whereas the other FF terms involve atoms bonded to each other. In better force fields, additional terms take into account anharmonicity and special electronic effects. The FF parameters are calibrated to reproduce experimental geometries and relative energies.

In molecular mechanics, the objective is to compute the energy of a molecular structure and/or to minimize the energy as a function of geometrical degrees of freedom, i.e., the potential energy surface. The computed energy reflects how far a geometrical arrangement around each atom deviates from an idealized bonding situation. A "better" geometry has a lower energy. Optimization methods [74] are used for energy minimization.

The MM model is conceptually simple, gives quite accurate molecular structures, and provides its information at least an order of magnitude faster than do QM methods. These strengths make MM well suited for treating both small and large molecules. Limitations of MM are that it requires empirical parameters to describe each "spring", and predictions depend on the care with which parameterization has been done, just as in quantum calculations where the results depend on the adequacy of the basis set and theoretical model.

Molecular dynamics (MD) simulations give information about the variation in structure and energy of a molecule over an interval of time [75, 76]. In MD, atoms move according to Newton's equations of motion (force = mass × acceleration, and force = negative gradient of potential energy). Thus, atoms under the most strain move fastest and farthest. To satisfactorily solve Newton's equations requires extremely short time steps, typically only 1 fs (10^{-15} s). If the time increment is too long, atoms can come too close in each step, and the energy of interaction from the FF become so high that the system becomes erratic in subsequent steps [77].

MD simulations are useful in computing accurate relative free energies. Free energy perturbation (FEP) or thermodynamic cycle calculations simulate the mathematical conversion of one molecule to another [78 – 80]. By changing the FF parameters in tiny increments from those of one chemical structure to those for another similar structure, the change in free energy can be accurately simulated. Assuming the incremental changes are small enough that the molecular system stays in equilibrium, FEP calculations can predict relative binding energies of two similar ligands associating with a receptor. Although in principle FEP calculations should be useful for testing molecular design ideas, in practice the calculations may be too slow for many industrial applications [81, 82].

Monte Carlo (MC) simulations are like MD calculations, but instead of proceeding along a time course, the system follows a random walk through configuration space [83]. Huge numbers of more or less randomly chosen Cartesian coordinates of the atoms of the system are considered. An ensemble average of the computed energies is used to calculate thermodynamic properties via statistical mechanics.

4.6. Statistical Modeling

By performing regression analysis and other statistical techniques, it is possible to quantitate the relation between molecular descriptors and an observed molecular property for a related set of compounds. The molecular property can be, for instance, the wavelength of an electronic transition, bioactivity, boiling point, chemical stability, etc. Descriptors can be almost

any quantity (— calculated or experimental —) that helps describe a molecule, such as molecular weight, atom counts, bond topological indices [84], molecular shape [85], or lipophilicity [14]. Descriptors related to substituents on an aromatic or aliphatic template have been assembled into large collections [86]. If a quantitative structure – property relationship (QSPR) can be found, then it can be used to predict (preferably by interpolation) the property of new structures. With the proliferation of descriptors, machine-learning techniques (see, e.g., [87]) have been developed to help choose the best ones [88].

4.7. Homology Modeling

Because of advances in the human genome project and in molecular biology in general, amino acid sequences of proteins are becoming known experimentally at a high rate, tens of thousands per year. Unfortunately, structural chemists cannot solve the 3D structures of these at the same rate, and it is very difficult to predict reliably how an amino acid sequence will fold.

One approach to obtaining a new 3D structure of a protein is to use homology modeling [89 – 92]. Here, the 3D structure has to be known for a sequence homologous to the new protein. The new sequence is threaded onto the backbone of the known structure, taking into account insertions and deletions in the sequence. The new model is then refined by energy minimization and an MD technique called simulated annealing. Thereby the theoretical temperature of the protein (determined by the kinetic energy of the atoms in the system) is gradually lowered in the hope that the structure will settle into a low minimum on the potential energy surface.

4.8. Drug Design Modeling

Some uses of the above methods have already been mentioned in regard to computer-aided drug design. In addition, a panoply of methods has been created specifically for applying molecular modeling to the design of chemical structures of interest to the pharmaceutical industry [93 – 98]. The vast majority of medicines perform their effect by binding to an enzyme or other protein. The rest target nucleic acids or other natural chemicals in the human body. Generally, synthetic molecules that bind to a biomacromolecule are called ligands.

Software tools [99, 100] have been developed to convert 2D chemical structures (as typically found in corporate and commercial databases) to 3D for further analysis by molecular modeling. Proprietary and commercial 3D databases are used by computational chemists in the pharmaceutical industry to search for structures matching a pharmacophore (the smallest set of atoms in a specific geometrical arrangement that will allow a molecule to be recognized by a biological receptor and bind to it). If appropriate hits from a 3D search are obtained, samples of the compounds can be submitted for testing in biological screens.

By combining quantitative structure – activity relationships (QSAR) and database tools, molecular modeling comes into play as a way to measure and maximize structural diversity in selecting compounds from databases for screening or to focus new compound libraries around a known lead structure. The ability of molecular modeling to provide molecular descriptors is important in database mining [101, 102] and combinatorial chemistry [103].

Comparative molecular field analysis (CoMFA) is a QSAR method that was developed in an attempt to model how ligands may be perceived in 3D by a receptor [104, 105]. The nonbonded van der Waals and electrostatic energies of a probe atom interacting at grid points surrounding a set of ligands are used as descriptors. The resulting data set has far more columns (descriptors) than rows (compounds). This situation is handled by the statistical method of partial least squares (PLS) of latent variables. If the ligands can be aligned in a strictly comparable way, then CoMFA can sometimes reveal which spatial regions surrounding a set of ligands enhance a property and which detract from it.

Some software tools combine conformational analysis and statistical modeling. These programs are designed to analyze the shapes of a series of compounds that are known to bind to the same receptor [106]. The algorithm tries to find conformations that let the more active compounds present a pharmacophore that the less active ones cannot. A statistical relationship is sought between activity and features in

the molecules, such as hydrophobic regions, hydrogen bonding sites, and ionized acidic or basic groups with formal positive or negative charge.

The 3D structure of many drug targets has not been determined. In such cases, methods for QSAR and modeling known ligands can be employed to elucidate the requirements for biological activity. In an increasing number of cases, a 3D structure of the drug target is known from experiment or can be built by molecular modeling. With a 3D structure in hand (structure-based drug design, SBDD), ligands can be docked in the receptor site [107 – 110]. Visual inspection of the docked molecules may reveal where adding or subtracting atoms/groups to/from the ligands would increase binding affinity. Usually, the better a drug binds to its intended target, the lower the required dose, and hence the lower the risk of side effects. The process of estimating the binding affinity is called scoring. A ligand may be able to orient in more than one way in a receptor; each orientation is called a pose. Computers may be used to automate the process in docking and scoring; this is called virtual screening [111].

A 3D model of a receptor can be further exploited with de novo design modeling programs [112, 113]. These programs generally have a library of molecular fragments, such as peptides or organic subunits, which they automatically assemble into possible 3D ligands. The hypothetical structures are evaluated in terms of the computed energy of interaction with the target receptor or other means. A weakness of early de novo design efforts was that they assumed that the target receptor structure is frozen. However, it is known both experimentally and computationally that some receptor sites change shape significantly when a ligand binds. Moreover, X-ray crystallography on ligand – receptor complexes shows that even similar ligands can occupy the receptor in different ways. Nevertheless, the de novo programs may suggest designs not previously considered, which can be followed up if they are synthetically feasible.

Pharmaceutical and biopharmaceutical companies currently use synthetic strategies, collectively called combinatorial chemistry, to produce thousands of compounds for biological screening. These sets of compounds are called libraries. Molecular modeling techniques are employed to increase the likelihood that bioactive compounds will be in the libraries [114, 115].

Genetic algorithms (GA) is a computational technique for solving optimization problems [116, 117]. The properties or constituents of a ligand can be modeled mathematically by a string of binary bits. Algorithms recombine and mutate these bit strings and then measure the resulting predicted property until the optimum combination of bits is found. GA can be used, for instance, to optimize the substituents on a molecular framework to fill a receptor cavity.

Computational drug design is extending beyond simply designing new ligands. Some of the new directions include studying the absorption, distribution, metabolism, and elimination (ADME) properties of compounds being considered for pharmaceutical development [118 – 121]. In addition, modeling is done to study potential toxicity of compounds. The preferred route of administration for most medicines is by mouth. Molecular modeling is used to study oral bioavailability [122]. One of the characteristics a compound must have to become a medicine is sufficient stability. Even this property is being scrutinized with molecular modeling [123].

The new field of cheminformatics is evolving mainly in regard to drug design as applied in industry [124, 125]. Cheminformatics arises from the cross-pollination of computer science, information science, and computational chemistry.

5. Economic Aspects

Most of the applications of molecular modeling alluded to in this article have dealt with compounds of interest to the pharmaceutical industry. This is because this area has to date been the largest area of applicability of the techniques. The total economic impact of molecular modeling is unknown, but the impact can be partially appreciated from the following data for the pharmaceutical industry. The major pharmaceutical companies operating in the United States, for instance, invested steadily greater amounts of their income in research and development: $\$ 2 \times 10^9$ in 1980, $\$ 15.2 \times 10^9$ in 1995, and $\$ 39.4 \times 10^9$ in 2005. Much of the money goes for clinical testing, and only a very tiny fraction is invested in modeling. Because the number of new drugs

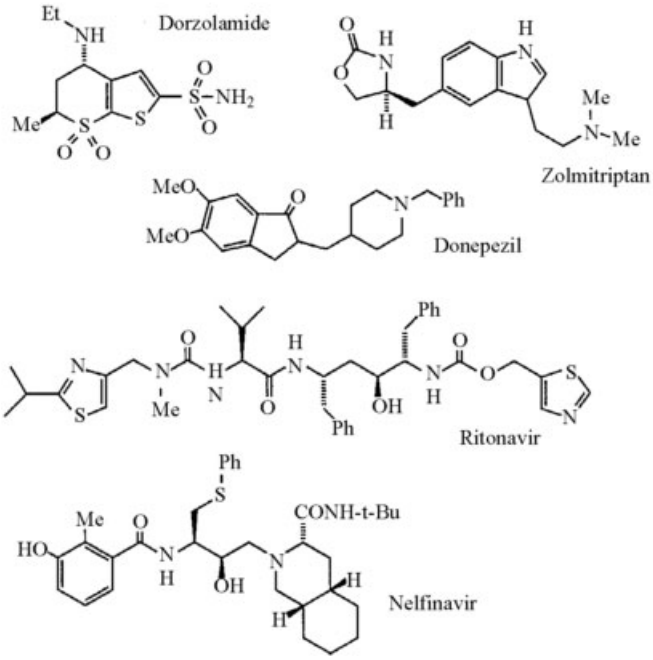

Figure 4. Examples of drugs discovered with some help of computer modeling: dorzolamide (Trusopt), an agent for treating glaucoma; zolmitriptan (Zomig), a serotonin-1D agonist; donepezil (Aricept), an acetylcholinesterase inhibitor; ritonavir (Norvir) and nelfinavir (Viracept), both HIV protease inhibitors.

reaching the market has been relatively flat (due to scientific difficulty and regulatory factors), it is not surprising that the average cost of discovering and developing a new pharmaceutical product is huge and has grown from \$ 597×10^6 in 1994 to about \$ 10^9 ten years later. Almost the entire burden of the R& D costs must be born by sales of existing pharmaceuticals. The incentive to innovate and discover new and better medicines is made possible by the high cost of prescription drugs. Innovative medicines save lives, improve the quality of lives, and in general help the economy of a nation by increasing productivity.

A large company may spend \$ 10×10^6 per year to support molecular modeling; this amount includes the investment in the computational chemists and computer support staff, the cost of new computers to stay current with technology, the purchase of new software and databases, and the licenses to maintain existing software and databases. However, a small company can launch an entry-level molecular modeling effort for under \$ 50 000.

Because of the extreme difficulty of finding viable new pharmaceutical products, researchers must use every effective technology to help them discover the few new compounds that meet medical and market needs. Molecular modeling has proven to be one such technology. Rarely, if ever, can it be claimed that a marketed product was designed purely by molecular modeling. Usually modeling plays a supportive role to the other scientific disciplines of a research organization. Sometimes modeling supplies information that answers specific research questions. Sometimes modeling can nudge a project in a direction that eventually leads to a new product. Cases of the discovery of marketed pharmaceuticals which were influenced or assisted by some type of computer modeling have been documented [126 – 135]. Structure-based drug design efforts have been particularly productive [136]. The chemical structures of some of these compounds are shown in Figure 4. In addition, modeling has proved useful for finding marketed agrochemicals, such as those depicted in Figure 5 [137 – 139].

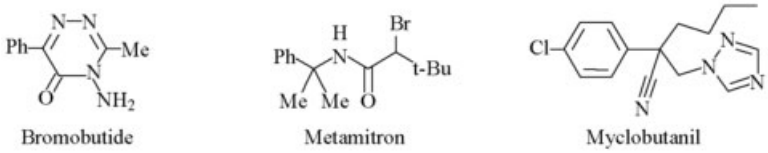

Figure 5. Examples of agrochemicals supported by computer modeling: bromobutide, a herbicide; metamitron, a photosystem II herbicide; myclobutanil (Systhane, Rally), a cytochrome P450 oxidase inhibitor.

Molecular modeling applications extend well beyond the pharmaceutical and agrochemical industries. Today, the methods are being used in industries commercializing adhesives, carbohydrates, catalysts, chromatographic stationary phases, clays [140], cleansing agents, corrosion inhibitors, crystals [141], explosives, flavorants, food components, fullerenes [142], glasses [143], hair colorings, lubricants and other petroleum products, materials [144], nanomaterials [145], nonlinear optical materials [146], nutrients, paint pigments, pesticides [147], photographic dyes, polymers, polysaccharides, semiconductors, soils [148], surfactants, water [149, 150], zeolites [151], and a host of other interesting materials (see, e.g., [152]). Uses of molecular modeling continue to grow because the computer methodologies help researchers gain a better appreciation of substances at a fundamental level.

6. References

General Reference

1. A. R. Leach: *Molecular Modeling: Principles and Applications*, 2nd ed., Prentice Hall, Harlow 2001.

Specific References

2. B. M. Trost, *Science* **227** (1985) 908–916.
3. A. J. Hopfinger, *J. Med. Chem.* **28** (1985) 1133–1139.
4. D. A. Pensak, *Pure Appl. Chem.* **61** (1989) 601–603.
5. K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990.
6. Archives of messages posted on the Computational Chemistry List (CCL) are available on the World Wide Web at http://www.ccl.net. See, e.g., the postings of A. Hocquet on 16 June 1997 and 4 July 1997, and A. Hocquet, L'actualite Chimique **7** (1997) 27–33.
7. H. Van de Waterbeemd et al., *Pure Appl. Chem.* **69** (1997) 1137–1152.
8. D. B. Boyd, J. D. Snoddy in D. Shugar, W. Rode, E. Borowski (eds.): *Molecular Aspects of Chemotherapy*, Springer-Verlag, New York 1992, p. 1–22.
9. J. D. Bolcer, R. B. Hermann in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 5, VCH Publ., New York 1994, p. 1–63.
10. U. Burkert, N. L. Allinger: *Molecular Mechanics*, Am. Chem. Soc., Washington DC 1990.
11. H. A. Scheraga in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 3, VCH Publ., New York 1992, p. 73–142.
12. D. B. Boyd, K. B. Lipkowitz in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 6, VCH Publ., New York 1995, p. 317–354.
13. D. B. Boyd, *J. Mol. Struct.: THEOCHEM* **369** (1997) 219–225.
14. P.-A. Carrupt, B. Testa, P. Gaillard in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 241–315.
15. D. B. Boyd, M. M. Marsh in S. Ekins (ed.): *Computer Applications in Pharmaceutical Research and Development*, Wiley, Hoboken 2006, p. 1–50.
16. P. Kollman, *Annu. Rev. Phys. Chem.* **38** (1987) 303–316.
17. A. D. French, J. W. Brady (eds.): *Computer Modeling of Carbohydrate Molecules*, American Chemical Society, Washington, DC, 1990.
18. G. Barnickel, *Pharm. Z.* **138** (1993) no. 45, 9–15.
19. G. Barnickel, *Chem. unserer Zeit* **29** (1995) no. 4, 176–185.
20. P. Comba, T. W. Hambley: *Molecular Modeling of Inorganic Compounds*, VCH Verlagsgesellschaft., Weinheim 1995.
21. B. J. Howlin, *Annu. Rep. Progr. Chem., Sect. C: Phys. Chem.* **92** (1996) 75–95.

22. W. B. Smith: *Introduction to Theoretical Organic Chemistry and Molecular Modeling*, VCH Publ., New York 1996.
23. H. D. Höltje, G. Folkers: *Molecular Modeling — Basic Principles and Applications*, VCH Verlagsgesellschaft., Weinheim 1996.
24. W. Gans, A. Amann, J. C. A. Boeyens (eds.): *Principles of Molecular Modeling*, Plenum Press, New York 1996.
25. N. B. Leontis, J. SantaLucia Jr. (eds.): *Molecular Modeling of Nucleic Acids*, Am. Chem. Soc., Washington DC 1998.
26. J. M. Goodman: *Chemical Applications of Molecular Modeling*, Royal Society of Chemistry, Cambridge 1998.
27. M. F. Schlecht: *Molecular Modeling on the PC*, Wiley-VCH, New York 1998.
28. H.-D. Höltje, W. Sippl, D. Rognan, G. Folkers: *Molecular Modeling — Basic Principles and Applications*, 2nd ed., Wiley-VCH, Weinheim 2003.
29. N. C. Cohen et al., *J. Med. Chem.* **33** (1990) 883–894.
30. D. B. Boyd in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, p. 303–380.
31. D. B. Boyd in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 373–399.
32. T. E. Ferrin et al., *J. Mol. Graphics* **9** (1991) 27–32, 37–38.
33. D. B. Boyd, R. D. Coner, *J. Mol. Struct.: THEOCHEM* **368** (1996) 7–15.
34. J. V. Duncia et al., *J. Med. Chem.* **33** (1990) 1312–1329.
35. W. T. Ashton et al., *J. Med. Chem.* **36** (1993) 591–609.
36. R. M. Keenan et al., *J. Med. Chem.* **36** (1993) 1880–1892.
37. D. B. Boyd et al. in C. H. Reynolds, M. K. Holloway, H. K. Cox (eds.): *Computer-Aided Molecular Design: Applications in Agrochemicals, Materials, and Pharmaceuticals*, Am. Chem. Soc. Symp. Series 589, Washington DC 1995, p. 14–35.
38. D. B. Boyd, *J. Med. Chem.* **36** (1993) 1443–1449.
39. K. Müller in B. Jensen, F. S. Jorgensen, H. Kofod (eds.): *Frontiers in Drug Research — Crystallographic and Computational Methods*, Munksgaard, Copenhagen 1990, p. 210–221.
40. A. E. Torda, W. F. van Gunsteren in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 3, VCH Publ., New York 1992, p. 143–172.
41. P. E. Bourne, H. Weissig (eds.): *Structural Bioinformatics*, Wiley-Liss, Hoboken 2003.
42. A. R. Leach in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 1–55.
43. M. Saunders et al., *J. Am. Chem. Soc.* **112** (1990) 1419–1427.
44. P. G. Mezey in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 265–294.
45. D. B. Boyd in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 321–354.
46. C. J. Cramer: *Essentials of Computational Chemistry: Theories and Models*, Wiley, Hoboken 2002.
47. D. B. Boyd et al., *J. Comput. Chem.* **9** (1988) 387–398.
48. R. J. Bartlett, J. F. Stanton in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 5, VCH Publ., New York 1994, p. 65–169.
49. D. Feller, E. R. Davidson in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 1–43.
50. W. J. Hehre et al.: *Ab Initio Molecular Orbital Methods*, Wiley, New York 1986.
51. E. R. Davidson in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 373–382.
52. J. Cioslowski in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 4, VCH Publ., New York 1993, p. 1–33.
53. J. J. P. Stewart in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 45–81.
54. M. C. Zerner in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 313–365.
55. C. J. Cramer, D. G. Truhlar in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 6, VCH Publ., New York 1995, p. 1–72.
56. L. J. Bartolotti, K. Flurchick in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, p. 187–216.

57. A. St.-Amant in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, p. 217–259.
58. E. Wimmer, *Analysis* **24** (1996) no. 1, M37–M42.
59. F. M. Bickelhaupt, E. J. Baerends in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 15, Wiley-VCH, New York 2000, p. 1–86.
60. S. Scheiner in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 165–218.
61. C. E. Dykstra et al. in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 83–118.
62. P. Bultinck, X. Gironés, R. Carbó-Dorca in K. B. Lipkowitz, R. Larter, T. R. Cundari (eds.): *Reviews in Computational Chemistry*, vol. 21, Wiley-VCH, Hoboken 2005, p. 127–207.
63. N. Matsunaga, S. Koseki in K. B. Lipkowitz, R. Larter, T. R. Cundari (eds.): *Reviews in Computational Chemistry*, vol. 20, Wiley-VCH, Hoboken 2004, p. 101–152.
64. J. Gao in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, p. 119–185.
65. D. E. Williams in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 219–271.
66. S. M. Bachrach in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 5, VCH Publ., New York 1994, p. 171–227.
67. P. Politzer, J. S. Murray in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 273–312.
68. D. B. Boyd, K. B. Lipkowitz, *J. Chem. Educ.*, **59** (1982) 269–274.
69. U. Dinur, A. T. Hagler in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 99–164.
70. R. L. DeKock et al. in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 4, VCH Publ., New York 1993, p. 149–228.
71. K. B. Lipkowitz, M. A. Peterson, *Chem. Rev.* **93** (1993) 2463–2486.
72. C. R. Landis, D. M. Root, T. Cleveland in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 6, VCH Publ., New York 1995, p. 73–148.
73. I. Pettersson, T. Liljefors in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 9, VCH Publ., New York 1996, p. 167–189.
74. T. Schlick in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 3, VCH Publ., New York 1992, p. 1–71.
75. W. F. van Gunsteren, H. J. C. Berendsen, *Angew. Chem. Int. Ed. Engl.* **29** (1990) 992–1023.
76. K. E. Gubbins, N. Quirke: *Molecular Simulation and Industrial Applications: Methods, Examples and Prospects*, Gordon and Breach, Amsterdam 1996.
77. D. B. Boyd in A. Kent, J. G. Williams (eds.): *Encyclopedia of Computer Science and Technology*, vol. 33, Marcel Dekker, New York 1995, p. 41–71.
78. T. P. Lybrand in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 295–320.
79. T. P. Straatsma in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, 9, VCH Publ., New York 1996, p. 81–127.
80. H. Meirovitch in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, 12, Wiley-VCH, New York 1998, p. 1–74.
81. L. M. Balbes, S. W. Mascarella, D. B. Boyd in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 5, VCH Publ., New York 1994, p. 337–379.
82. M. R. Reddy, M. D. Erion, A. Agarwal in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, 16, Wiley-VCH, New York 2000, p. 217–304.
83. R. J. Woods in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 9 VCH Publ., 1996, p. 129–165.
84. L. H. Hall, L. B. Kier in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 367–422.
85. G. A. Arteca in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, 9, VCH Publ., New York 1996, p. 191–253.
86. C. Hansch, A. Leo, D. Hoekman: *Exploring QSAR; Hydrophobic, Electronic, and Steric Constants*, Am. Chem. Soc., Washington, DC, 1995.
87. K. L. Peterson in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, 16, Wiley-VCH, New York 2000, p. 53–140.

88. D. J. Livingstone, D. W. Salt in K. B. Lipkowitz, R. Larter, T. R. Cundari (eds.): *Reviews in Computational Chemistry*, 21, Wiley-VCH, Hoboken 2005, p. 287–348.
89. J. M. Troyer, F. E. Cohen in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 2, VCH Publ., New York 1991, p. 57–80.
90. M. F. Hibert in C. G. Wermuth (ed.): *Practical Medicinal Chemistry*, Academic Press, London 1996, p. 523–546.
91. H. Cid, M. Bunster, *Biol. Res.* **29** (1996) 77–100.
92. N. Eswar et al., *Nucleic Acids Res.* **31** (2003) 3375–3380.
93. P. Gund et al., *Science* **208** (1980) 1425–1431.
94. A. S. V. Burgen, G. C. K. Roberts, M. S. Tute (eds.): *Molecular Graphics and Drug Design*, Elsevier, Amsterdam 1986.
95. J. G. Vinter, M. Gardner (eds.): *Molecular Modelling and Drug Design*, CRC, Boca Raton 1994.
96. N. C. Cohen (ed.): *Guidebook to Molecular Modeling and Drug Design*, Academic Press, San Diego 1997.
97. A. Tropsha, W. Zheng in O. M. Becker (ed.): *Computational Biochemistry and Biophysics*, Dekker, New York, 2001 p. 351–369.
98. P. Bultinck, H. De Winter, W. Langenaeker, J. P. Tollenaere (eds.): *Computational Medicinal Chemistry for Drug Discovery*, Dekker, New York 2004.
99. J. Sadowski, J. Gasteiger, *Chem. Rev.* **93** (1993) 2567–2581.
100. J. Sadowski, C. H. Schwab, J. Gasteiger in P. Bultinck, H. De Winter, W. Langenaeker, J. P. Tollenaere (eds.): *Computational Medicinal Chemistry for Drug Discovery*, Dekker, New York 2004, p. 151–212.
101. G. M. Downs, P. Willett in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, 1–66.
102. A. C. Good, J. S. Mason in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 7, VCH Publ., New York 1995, 67–117.
103. E. J. Martin et al. in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 10, VCH Publ., New York 1996, p. 75–100.
104. T. I. Oprea, C. L. Waller in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 127–182.
105. G. Greco, E. Novellino, Y. C. Martin in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 183–240.
106. A. Smellie, S. D. Kahn, S. L. Teig, *J. Chem. Inf. Comput. Sci.* **35** (1995) 285–304.
107. J. P. Tollenaere, *Pharm. World Sci.* **18** (1996) no. 2, 56–62.
108. J. M. Blaney, J. S. Dixon in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 5, VCH Publ., New York 1994, p. 299–335.
109. I. Muegge, M. Rarey in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 17, Wiley-VCH, New York 2001, p. 1–60.
110. H.-J. Böhm, M. Stahl in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 18, Wiley-VCH, New York 2002, p. 41–87.
111. P. Kolb et al. in J. Alvarez, B. Schoichet (eds.): *Virtual Screening in Drug Discovery*, CRC Press, Boca Raton 2005, p. 349–378.
112. M. A. Murcko in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 1–66.
113. D. E. Clark, C. W. Murray, J. Li in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 11, Wiley-VCH, New York 1997, p. 67–125.
114. R. A. Lewis, S. D. Pickett, D. E. Clark in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 16, Wiley-VCH, New York 2000, p. 1–51.
115. A. Dominik in W. Bannwarth, B. Hinzen (eds.): *Combinatorial Chemistry — From Theory to Application*, 2nd ed., Wiley-VCH, Weinheim 2006, p. 559–613.
116. R. Judson in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 10, Wiley-VCH, New York 1997, p. 1–73.
117. D. E. Clark (ed.): *Evolutionary Algorithms in Molecular Design*, Wiley-VCH, Weinheim 2000.
118. A. M. ter Laak, N. P. E. Vermeulen in B. Testa (ed.): *Pharmocokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*, Verlag Helvetica Chimica Acta, Zurich 2001, p. 551–588.
119. D. Butina, M. D. Segall, K. Frankcombe, *Drug Discovery Today* **7** (2002) S83–S88.
120. S. Ekins, P. W. Swaan in K. B. Lipkowitz, R. Larter, T. R. Cundari (eds.): *Reviews in Computational Chemistry*, vol. 20, Wiley-VCH, Hoboken 2004, p. 333–415.

121. M. P. Payne in M. T. Cronin, D. J. Livingstone (eds.): *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton 2004, p. 205–227.
122. H. van de Waterbeemd, H. Lennernäs, P. Artursson, *Drug Bioavailability: Estimation of Solubility, Permeability, Absorption and Bioavailability*, Wiley VCH, Weinheim 2003.
123. S. W. Baertschi (ed.), *Pharmaceutical Stress Testing: Predicting Drug Degradation*, Taylor and Francis, Boca Raton 2005.
124. A. R. Leach, V. J. Gillet, *An Introduction to Chemoinformatics*, Kluwer, Dordrecht 2003.
125. T. I. Oprea (ed.): *Chemoinformatics in Drug Discovery*, Wiley-VCH, Weinheim 2005.
126. D. B. Boyd in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 355–371.
127. J. Greer et al., *J. Med. Chem.* **37** (1994) 1035–1054.
128. D. J. Kempf et al., *Proc. Natl. Acad. Sci. USA* **92** (1995) 2484–2488.
129. R. C. Glen et al., *J. Med. Chem.* **38** (1995) 3566–3580.
130. A. K. Patick et al., *Antimicrob. Agents Chemother.* **40** (1996) 292–297.
131. Y. Kawakami et al., *Bioorg. Med. Chem.* **4** (1996) 1429–1446.
132. J. R. Damewood, Jr. in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 9, VCH Publ., New York 1996, p. 1–79.
133. D. B. Boyd in T. Liljefors, F. S. Jorgensen, P. Krogsgaard-Larsen (eds.): *Rational Molecular Design in Drug Research*, vol. 42, Munksgaard, Copenhagen 1998.
134. D. B. Boyd in A. L. Parrill and M. R. Reddy (eds.): *Rational Drug Design: Novel Methodology and Practical Applications*, American Chemical Society, Washington, DC, 1999, p. 346–356.
135. D. B. Boyd in P. v. R. Schleyer, N. L. Allinger, T. Clark, J. Gasteiger, P. Kollman, H. F. Schaefer III (eds.): *Encyclopedia of Computational Chemistry*, vol. 1, Wiley, Chichester 1998, p. 795–804.
136. M. Congreve, C. W. Murray, T. L. Blundell, *Drug Discovery Today* **10** (2005) 895–907.
137. O. Kirino, *Nippon Noyaku Gakkaishi* **9** (1984) 571–579.
138. W. Draber, *Z. Naturforsch. Teil C* **42** (1987) 713–717.
139. T. T. Fujimoto et al., *Pesticide Biochem. Physiol.* **30** (1988) 199–213.
140. S.-H. Park, G. Sposito in S. M. Auerbach, K. A. Carrado, P. K. Dutta (eds.): *Handbook of Layered Materials*, Marcel Dekker, New York 2004, p. 39–89.
141. P. Verwer, F. J. J. Leusen in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 12, Wiley-VCH, New York 1998, p. 327–365.
142. Z. Slanina, S.-L. Lee, C.-H. Yu in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 8, VCH Publ., New York 1996, p. 1–62.
143. A. N. Cormack, J. Du, T. Zeitler in *Proceedings of the 19th International Congress on Glass*, Society of Glass Technology, Sheffield 2001, p. 584–588.
144. J.-R. Hill, C. M. Freeman, L. Subramanian in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 16, Wiley-VCH, New York 2000, p. 141–216.
145. R. Q. Topper et al. in K. B. Lipkowitz, R. Larter, T. R. Cundari (eds.): *Reviews in Computational Chemistry*, vol. 19, Wiley-VCH, Hoboken, 2003, p. 1–41.
146. H. A. Kurtz, D. S. Dudis in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 12, Wiley-VCH, New York 1998, p. 241–279.
147. E. L. Plummer in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 1, VCH Publ., New York 1990, p. 119–168.
148. M. H. Gerzabek et al., *Bodenkultur* **52** (2001) 133–146.
149. J. C. Shelley, D. R. Bérard in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 12, Wiley-VCH, New York 1998, p. 137–205.
150. A. Wallqvist, R. D. Mountain in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 13, Wiley-VCH, New York 1999, p. 183–247.
151. B. van de Graff, S. L. Njo, K. S. Smirnov in K. B. Lipkowitz, D. B. Boyd (eds.): *Reviews in Computational Chemistry*, vol. 14, Wiley-VCH, New York 2000, p. 137–223.
152. C. H. Reynolds, M. K. Holloway, H. K. Cox (eds.): *Computer-Aided Molecular Design: Applications in Agrochemicals, Materials, and Pharmaceuticals*, Am. Chem. Soc. Symp. Series 589, Washington, DC, 1995.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Naphthalene and Hydronaphthalenes

Standard Article

Gerd Collin[1], Hartmut Höke[2], Helmut Greim[3]

[1]DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany
[2]Weinheim, Federal Republic of Germany
[3]Institut für Toxikologie und Umwelthygiene, TU München, Freising-Weihenstephan, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a17_001.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Physical and Chemical Properties
2. Production
3. Uses
4. Alkylnaphthalenes
5. Hydronaphthalenes
6. Economic Aspects
7. Toxicology

[Top of Page]

Naphthalene was discovered in coal tar by A. GARDEN in 1819.

[Top of Page]

1. Physical and Chemical Properties

Gerd Collin

Naphthalene [91-20-3], $C_{10}H_8$, M_r 128.18, *mp* 80.29 °C, *bp* 217.95 °C (101.3 kPa), forms colorless flakes or monoclinic crystals with a characteristic odor. It is readily soluble in benzene, diethyl ether, chloroform and carbon disulfide, soluble in ethanol, and insoluble in water. Naphthalene is volatile in steam and readily sublimable. It forms azeotropic mixtures with, for example, ethylene glycol, acetamide, *m*-cresol, and benzyl alcohol. Some important physical properties are as follows:

Density (20 °C)	1.1789 g/cm^3
Refractive index (99.5 °C)	1.5829
Heat capacity (25 °C)	1.294 kJ/kg
Heat of fusion	148 kJ/kg
Heat of vaporization	352 kJ/kg
Vapor pressure at 70 °C	0.525 kPa
at 88 °C	1.33 kPa
at 181 °C	40 kPa
at 270 °C	295 kPa
at 349 °C	1.01 MPa
at 479 °C	4.13 MPa
Critical temperature	475 °C
Flash point (closed cup)	80 °C

Ignition temperature	540 °C
Explosion limits in air	
upper	5.9 vol %
lower	0.88 vol %
Dielectric constant (20 °C)	2.47
Odor threshold	0.004 mg/m^3

Naphthalene can be catalytically hydrogenated in the liquid phase, preferably in the presence of nickel, with high selectivity to give tetra- and decahydronaphthalene, and in the gas phase in the presence of oxidic catalysts less selectively to give mixtures of tetra-, hexa-, octa-, and decahydronaphthalene. Liquid-phase oxidation with various oxidants or gas-phase oxidation with air in the presence of vanadium pentoxide catalysts gives (via 1,4-naphthoquinone) phthalic acid, which is readily dehydrated to phthalic anhydride (⟶ Phthalic Acid and Derivatives). The 1-position of naphthalene is the most nucleophilic. Consequently, electrophilic substitution reactions, such as nitration, halogenation, sulfonation, and alkylation at lower temperature, preferentially give the corresponding 1-substituted products; at higher temperature, isomerization occurs to give the thermodynamically more stable 2-substituted products. Nitration with nitric acid at low temperature produces 1-nitronaphthalene, and with H_2SO_4– HNO_3 at high temperature, 1,5- and 1,8-dinitronaphthalene (⟶). Halogenation forms 1-halo- and/or 1,4- and 1,5-dihalonaphthalenes, depending on the reaction conditions. Polychloronaphthalenes are obtained by catalytic chlorination at higher temperature (⟶ Chlorinated Hydrocarbons). Sulfonation with concentrated sulfuric acid gives 1-naphthalenesulfonic acid, and at higher temperature, 2-naphthalenesulfonic acid; oleum gives 1,5-, 1,6-, and 2,7-naphthalenedisulfonic acids and trisulfonic acids, depending on the SO_3 content (⟶). Naphthalene can be alkylated at higher temperature with alkyl halides, alcohols, and olefins in the presence of strong acids or Friedel-Crafts catalysts, with preferential formation of 2-alkyl derivatives (⟶ Acylation and Alkylation). Dehydrogenating condensation of naphthalene (Scholl reaction) gives perylene [198-55-0]. Acylation of naphthalene leads to mixtures of the 1- and 2-acyl derivatives, with the ratio depending on the solvent used (⟶ Acylation and Alkylation – Acylation of Polynuclear Aromatic Compounds). 1-Chloromethylation with paraformaldehyde and hydrochloric acid affords 1-chloromethylnaphthalene.

[Top of Page]

2. Production

Gerd Collin

The main feedstock for naphthalene production is coal tar. At an average content of around 10 %, naphthalene represents the most important compound in high-temperature coal tar (coke-oven tar) in terms of quantity. It can already be concentrated to over 90 % in primary tar distillation, in the naphthalene fraction boiling between 210 and 220 °C, with a yield in excess of 90 %. Further refining requires separation of the co-boiling compounds (phenols, nitrogen bases, benzo[*b*] thiophene, and other aromatic hydrocarbons). For production of technical naphthalene (e.g., *mp* 78.5 °C), redistillation of the naphthalene fraction is sufficient. Generally, phenols are extracted with sodium hydroxide solution before distillation (Fig. 1). Pure naphthalene (*mp* ≥79.6 °C) is produced by, e.g., crystallization or hydrorefining.

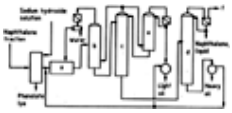

Figure 1. Naphthalene recovery from coal tar by extraction and distillation

a) Dephenolated naphthalene fraction; b) Dewatering column; c) Light oil column; d) Naphthalene column; e) Cooling; f) To the vacuum unit

Crystallization especially separates benzo[*b*]-thiophene, which boils at only 2 °C above naphthalene. Benzo[*b*]thiophene forms a eutectic mixture with naphthalene at a benzo[*b*]thiophene content of 93 %, and mixed crystals elsewhere in the phase diagram. Whereas until recently, naphthalene was predominantly recovered by crystallization from the melt, with formation of suspensions, newer processes of crystallization from the melt result in the formation of layers or blocks. Benzo [*b*]thiophene and other impurities can be separated from naphthalene by multistage countercurrent crystallization from the melt without adding selective solvents. Four to six crystallization stages are needed to produce pure naphthalene with a melting point of 80 °C from a coal tar primary fraction containing 90 % naphthalene.

Hydrorefining of naphthalene is carried out at 400 °C and ca. 1.4 MPa in the presence of cobalt – molybdenum catalysts, whereby benzo[*b*]-thiophene is converted to hydrogen sulfide and ethylbenzene. Other co-boiling compounds are broken down to lower-boiling hydrocarbons and can be separated by distillation. Naphthalene produced by this method has a sulfur content of ≤100 ppm and contains tetralin as the main by-product (ca. 1 %).

Other possible methods for separating benzo[*b*]thiophene include selective sulfonation, polycondensation with formaldehyde and acid, reaction with sodium metal, and azeotropic distillation with glycols.

Naphthalene can also be recovered from pyrolysis residue oils (from the pyrolysis of hydrocarbon fractions to olefins) by distillation and crystallization, in a manner similar to the recovery from coal tar. These pyrolysis residue oils normally contain 10 – 16 % naphthalene [6].

In addition to recovery from coal tar, naphthalene is also produced in the United States by hydrodealkylation of aromatized petroleum-derived fractions (Fig. 2) [7], [8]. Typical feedstocks are residue oils from catalytic naphtha reforming, from naphtha pyrolysis in olefin plants, or catcracker recycle oils. These aromatized oils are distilled to give a fraction containing naphthalene and alkylnaphthalenes and, in the case of catcracker recycle oils, subjected to extraction of the aromatics. The

bicyclic aromatic fraction, with a boiling range of, e.g., 220 – 270 °C, is dealkylated under hydrogen pressure, either thermally above 700 °C, or catalytically in the presence of a chromium oxide/aluminum oxide or cobalt oxide/molybdenum oxide catalyst at 550 – 650 °C. Methyl- and dimethylnaphthalene fractions from coal tar can also be hydrodealkylated. The crude dealkylation product is processed to a low-sulfur pure naphthalene (*mp* 80 °C) by distillation.

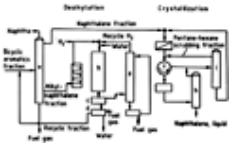

Figure 2. Naphthalene recovery by catalytic hydrodealkylation of alkylnaphthalene fractions (Unidak process)

a) Naphthalene column; b) Reactor; c) High-pressure gas separator; d) Low-pressure gas separator; e) Methane scrubber; f) Centrifuge; g) Melting vessel; h) Stripping column; i) Solvent recovery column

[Top of Page]

3. Uses

Gerd Collin

Worldwide, the main derivative of naphthalene is still phthalic anhydride (from catalytic gas-phase oxidation of naphthalene), an intermediate for PVC plasticizers (⟶ Phthalic Acid and Derivatives). Other important applications for naphthalene are dyes, particularly azo dyes produced from 2-naphthol [135-19-3] and naphthalene sulfonic acids (⟶ Azo Dyes). Naphthalene and alkylnaphthalene sulfonates are used as surfactants, and naphthalene sulfonate – formaldehyde condensates as tanning agents (syntans) and dispersants, e.g., as superplasticizers for concrete (⟶ Disperse Systems and Dispersants – Formaldehyde Condensation Products). Naphthalene is used in the production of moth repellents and fumigants and the insecticide carbaryl (1-naphthyl-*N*-methylcarbamate). It is also used for the synthesis of tetrahydroanthraquinone from naphthoquinone and butadiene (⟶ Anthraquinone – Production) and for alkylnaphthalene solvents for carbonless copy paper (Kureha Micro Capsule Oil, KMC, a mixture of diisopropylnaphthalene isomers), the solvents tetralin and decalin, as well as chlorinated naphthalenes (⟶ Chlorinated Hydrocarbons), of which only the monochloronaphthalenes retain commercial significance, e.g., as dye dispersants and as fungicides and insecticides for preservatives. Polychlorinated naphthalenes are rarely produced due to their toxicity and resistance to environmental degradation. 1-Methyl- and 1-chloromethylnaphthalene are used in the production of the plant growth regulators 1-naphthaleneacetamide [86-86-2] and 1-naphthaleneacetic acid [86-87-3]. The alkaline-earth salts of sulfonated dinonylnaphthalenes are important lube-oil additives. New applications are opening up for 2,6-naphthalenedicarboxylic acid and 6-hydroxy-2-naphthoic acid as intermediates for liquid crystal polymers and heat-resistant polyester fibers and films. For further details on the uses of naphthalene, see Figure 3.

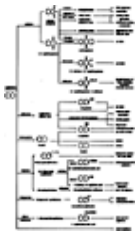

Figure 3. Uses of naphthalene

[Top of Page]

4. Alkylnaphthalenes

Gerd Collin

1-Methylnaphthalene [90-12-0], $C_{11}H_{10}$, M_r 142.20, *bp* 244.4 °C (101.3 kPa), *mp* –30.5 °C, d^{20} 1.0203, is a colorless, blue-fluorescing liquid which is readily soluble in benzene, ethanol, and diethyl ether, and insoluble in water. 1-Methylnaphthalene is present in high-temperature coal tar in a concentration of 0.5 % and is produced industrially from the methylnaphthalene fraction, which boils between 240 and 245 °C, by redistillation of the 2-methylnaphthalene filtrate following crystallization and separation of 2-methylnaphthalene. 1-Methylnaphthalene can be isomerized to 2-methylnaphthalene in the presence of, e.g., BF_3, phosphoric acid, and phase-transfer catalysts [9], or on zeolite catalysts [10]. It is used for the synthesis of 1-naphthaleneacetic acid and, in mixtures with 2-methylnaphthalene, as a solvent and heattransfer oil.

2-Methylnaphthalene [91-57-6], $C_{11}H_{10}$, M_r 142.20, *bp* 241.1 °C (101.3 kPa), *mp* 34.6 °C, d^{20} 1.029, forms colorless crystals. The compound is readily soluble in benzene, ethanol and carbon disulfide, but insoluble in water. It is recovered industrially by crystallization from the methylnaphthalene fraction of high-temperature coal tar, which contains 1.5 % 2-methylnaphthalene. 2-Methylnaphthalene can also be produced by isomerization of 1-methylnaphthalene (see above). 2-Methylnaphthalene is the feedstock for the production of 2-methyl-1,4-naphthoquinone [58-27-5] (menadione, vitamin K_3), and mixtures with 1-methylnaphthalene are used as solvents and heat-transfer oils. It is used in small quantities for the production of alkylmethylnaphthalene sulfonates as textile auxiliaries, surfactants, and emulsifiers. 2-Methylnaphthalene can be used via 2-methyl-6-acetylnaphthalene [11], [12], or in an older process (also as a mixture with 1-methylnaphthalene) by oxidation with subsequent Henkel rearrangement, to give 2,6-naphthalenedicarboxylic acid [1141-38-4] (⟶ Carboxylic Acids, Aromatic). This compound is used in the production of special heat-resistant polyester fibers and films, and of liquid crystal polymers. 2-Methylnaphthalene can be acylated in 1,2-dichlorbenzene to give 1-acyl-7-methylnaphthalene [13], which can be used as an intermediate for pharmaceuticals [14].

2,6-Dimethylnaphthalene [581-42-0], $C_{12}H_{12}$, M_r 156.23, *bp* 261 °C (101.3 kPa), *mp* 110 °C, forms colorless crystals which are slightly soluble in ethanol and insoluble in water. It can be recovered from the dimethylnaphthalene fraction of coal tar by redistillation and crystallization; from aromatized catcracker recycle oil by aromatics extraction, redistillation, hydrogenation of the isomer mixture to dimethyltetralins, isomerization, and dehydrogenation; or by disproportionation of methylnaphthalenes to give naphthalene and dimethylnaphthalenes, from which the 2,6-isomer can be obtained analogously to the production from aromatized catcracker recycle oil. 2,6-Dimethylnaphthalene is also obtained by alkylation of 2-methylnaphthalene or naphthalene on zeolite catalysts [15], [16].

2,6-Dimethylnaphthalene can be converted to 2,6-naphthalene dicarboxylic acid by liquid-phase oxidation with oxygen in acetic acid as solvent in the presence of a cobalt acetate/manganese bromide catalyst and a co-catalyst, e.g., ruthenium chloride [17].

2-Isopropylnaphthalene [2027-17-0], $C_{13}H_{14}$, M_r 170.25, *bp* 268.9 °C (101.3 kPa), *mp* 15.1 °C, d^{20} 0.9762, is a colorless liquid. It can be used to synthesize 1-naphthol [90-15-3] by oxidation via the hydroperoxide (Hock synthesis) (⟶). Catalytic gas-phase dehydrogenation produces polymerizable 2-isopropenylnaphthalene [3710-23-4] [18].

Diisopropylnaphthalenes. Selective Friedel-Crafts alkylation of naphthalene with propene in the presence of an aluminum silicate catalyst yields as the main product a mixture of diisopropylnaphthalene isomers that is liquid over a wide temperature range. This isomer mixture is used as a solvent for carbonless copy paper [19], [20]. The solvent, trade name KMC (Kureha Micro Capsule Oil), is odorless, colorless, and environmentally harmless [21]. KMC can also be used as a solvent for scintillation measurements and as a heat transfer oil. Some physical properties of KMC are as follows:

Boiling range (101.3 kPa)	290 – 299 °C
Pour point	<–40 °C
Flash point	>130 °C
Density (15 °C)	0.96 g/cm^3
Refractive index (25 °C)	1.565
Specific heat (20 °C)	1.71 kJ/kg K
Thermal conductivity (20 °C)	0.12 W/Km
Dielectric constant (90 °C)	2.46

Technically pure *2,6-diisopropylnaphthalene* [24157-81-1] (*mp* 70 °C) can be isolated from the isomer mixture by crystallization [22]. It can be converted to 2,6-naphthalene dicarboxylic acid [23-25] or to 6-hydroxy-2-naphthoic acid [16712-64-4] by liquid-phase oxidation [26]. These compounds are used as intermediates for heat-resistant liquid crystal polyesters [27].

[Top of Page]

5. Hydronaphthalenes

Gerd Collin

1,2,3,4-Tetrahydronaphthalene [119-64-2], tetralin, $C_{10}H_{12}$, M_r 132.21, *bp* 207.3 °C (101.3 kPa), *mp* –35 °C, d^{20} 0.9729, n_D^{20} 1.5461, is a colorless liquid which is infinitely soluble in chloroform, petroleum ether (ligroin), diethyl ether, and ethanol, but insoluble in water. Tetralin is produced industrially by selective hydrogenation of low-sulfur naphthalene in the presence of nickel catalysts at 180 – 260 °C and 1 – 1.5 MPa.

Tetralin is used as a solvent for fats, resins, and paints; for the production of detergents; and for the synthesis of carbaryl via 1-naphthol. It can be applied as a hydrogen-donor solvent in coal extraction (⟶ Coal Liquefaction – Solvents).

Decahydronaphthalene [91-17-8], decalin, $C_{10}H_{18}$, M_r 138.25, exists in two stereoisomeric forms: *trans*-decalin, *bp* 185.5 °C (101.3 kPa), *mp* –32.5 °C, d^{20} 0.8700 , n_D^{20} 1.4696; and *cis*-decalin, *bp* 195 °C, *mp* –45 °C (101.3 kPa), d^{20} 0.8967, n_D^{20} 1.4811.

The isomer mixture is a colorless liquid with a camphor-like odor. Technical decalin is infinitely soluble in butanol, soluble in chloroform, acetone, carbon disulfide and benzene, and slightly soluble in ethanol and methanol.

Decalin is produced by liquid-phase hydrogenation of naphthalene under pressure in the presence of nickel catalysts. The proportion of the two stereoisomers formed depends on the reaction conditions.

Like tetralin, decalin is used as a solvent for fats, resins, waxes, and paints. It can also be used to make cyclodecanone, an intermediate in the production of polyamide 10.

[Top of Page]

6. Economic Aspects

Gerd Collin

World production of naphthalene in 1987 was ca. 10^6 t. About one-fourth of this came from Western Europe, one-fifth each

from Japan and Eastern Europe, and one-eighth from the United States. In addition to coal-tar based naphthalene, petroleum-derived naphthalene has been recovered since 1961, predominantly in the United States, by dealkylation of aromatics-rich fractions from reforming and catalytic cracking. The proportion of petroleum-based naphthalene has decreased during the last decade from ca. 15 % to ca. 5 %, because coal-tar naphthalene is available in sufficient quantity. Table 1 summarizes the main uses of naphthalene as percentage of consumption in Western Europe, Japan, and the United States.

Table 1. Uses of naphthalene in % (1987)

Use	Western Europe	United States	Japan
Phthalic anhydride	40	65	70
Dyes	29		10
Surfactants, tanning agents, dispersants	15	14	5
Moth repellants, fumigants	9	5	2
Insecticides		14	
Solvents and other uses	7	2	12

An application for alkylnaphthalenes is the use of diisopropylnaphthalene isomer mixtures as solvents for carbonless copy paper. Production capacities for these mixtures, trade name KMC, total 10 000 t/a each in Japan and the Federal Republic of Germany. World production of 2-methylnaphthalene and 1-/2-methylnaphthalene mixtures is estimated at 1500 t/a each. For hydronaphthalenes, tetralin production is estimated to be 25 000 t/a worldwide. Naphthalenesulfonate – formaldehyde condensates are produced worldwide in quantities of ca. 40 000 t/a, predominantly for concrete plasticizers. Although polyester products derived from 2,6-naphthalenedicarboxylic acid or 6-hydroxy-2-naphthoic acid have been on the market for some time, they have not yet gained major commercial importance.

[Top of Page]

7. Toxicology

Hartmut Höke and Helmut Greim

Naphthalene

Human Health. The health consequences of human exposure to naphthalene have been repeatedly evaluated [48-51]. Most of the information given subsequently derives from these reports.

Effects assessment. Naphthalene is readily absorbed during oral, inhalation and dermal exposure. In rodents it is metabolized by epoxidation and subsequent hydroxylation. These reactive metabolites undergo glutathione conjugation or reduction forming the 1,2-dihydrodiol. In humans, 1-naphthol, 2-naphthol, and 1,2-, and 1,4-naphthoquinone are formed. More recent data demonstrate that cytochrome P450-2F2 in cellular systems of rats and mice metabolizes naphthalene to 1R,2S-naphthaleneoxide [52] which rearranges to 1-naphthol and forms the 1,2-dihydrodiol by hydrolase activity. Oxidation of 1-naphthol forms 1,2-, and 1,4-naphthoquinone [53]. In Clara cells isolated from the lungs of naphthalene treated mice ZHENG et al. [54] reported covalent binding of 1,2-naphthoquinone to protein.

In humans, single or repeated exposures to naphthalene caused severe hemolytic anemia esp. in infants exposed to textiles (clothing/bedding), which have been stored with naphthalene mothballs. It may cause slight irritation on contact with the skin and eyes. The absence of case reports indicates that naphthalene is not a skin or respiratory sensitizer in humans.

In rodents naphthalene is of low acute toxicity and did not show sensitizing activity. Repeated oral administration of naphthalene causes cataracts in rats and rabbits at doses of 700 mg/kg bw/d and above. Despite its widespread use there are no reliable reports of cataracts in humans. Repeated inhalation exposure for 28 days produces irritation of the nasal epithelium, with mild effects at levels as low as 5 mg/m^3 in rats (LOAEL). In mice, signs of chronic respiratory tract inflammation were noted at 50 mg/m^3. There was no indication of hemolytic anemia in rodent studies.

Naphthalene was not mutagenic in bacterial assays and did not induce UDS *in vitro*. In CHO cells clastogenic effects and sister chromatid exchanges (SCE) were found *in vitro* in the presence and in the absence of metabolic activation. In human peripheral lymphocytes no SCEs have been induced in the presence or absence of microsomal enzymes. Other *in vitro* tests (micronuclei in MCL-5 cells [55], DNA fragmentation in macrophages [56], chromosomal aberrations in mice embryo cultures [57, 58]) indicated genotoxic effects. *In vivo*, naphthalene did not induce micronuclei in bone marrow cells of mice or UDS in rat liver cells. The metabolic formation of an epoxide, which may also be formed in the nasal cavity of rats in the

carcinogenicity studies, implies that the involvement of a genotoxic mechanism in tumor formation at this target cannot be excluded.

In a carcinogenicity study [59] by inhalation, naphthalene produced epithelial adenomas and olfactory epithelial neuroblastomas in the nasal cavity of rats from the lowest exposure concentration (50 mg/m^3). This must be considered of high relevance because neuroblastomas are highly malignant and the P450 isozymes that metabolically activate naphthalene in the nasal cavity of rodents are also present in humans. In mice an increased incidence in pulmonary adenomas was seen. Since lung tumors were not seen in rats, and due to the higher capacity of the mouse lung in metabolic activation and inactivation, the pulmonary tumors seen in mice seem to be of little relevance to humans.

No studies to evaluate reproductive toxicity are available. Although no changes in the reproductive organs have been detected in repeated dose studies, appropriate studies are needed.

Exposure. Highest occupational exposures occur in mothball manufacture, from the manufacture of grinding wheels and the use of creosote.

Exposure via inhalation also occurs when organic material is incompletely combusted (e.g., in processing of coal, crude oil and natural gas, metal foundries, and power plants).

Consumer exposure to naphthalene may occur through the use of moth repellents, tar shampoos and soaps and through damp proofing operations.

Significant exposure in the mg/kg bw/d range was estimated from dermal and inhalation exposures for workers and consumers. Professional use of tar soap and shampoos exposure was considered to be low.

Indirect human exposure via the environment is low. Infants, in particular, may significantly be exposed to textiles (clothing/bedding), which have been in contact with naphthalene mothballs.

Human exposure via the environment is estimated to range from 65 ng/kg/d at the regional level to 0.25 mg/kg/d at the local level with releases from the manufacture of grinding wheels or mothballs contributing most. Environmental airborne levels are 0.14 µg/m^3.

Risk characterization. The available information indicates that there are no concerns regarding irritation, sensitization and mutagenicity.

The critical health effects are hemolytic anemia and carcinogenicity.

Since a NOAEL for hemolytic anemia cannot be established, there is concern due to exposure of neonates and infants to clothes that have been in contact with mothballs and following dump-proofing. The predicted exposures are high and can be close to concentrations at which local damage to the respiratory tract has occurred in rats.

In the carcinogenicity studies in rats, chronic inflammation may be a key mechanism in the development of the nasal tumors. However, it cannot be concluded whether a secondary genotoxicity due to chronic inflammation or a genotoxic effect is the relevant mechanisms of naphthalene to induce tumors. Formation of an epoxide and the clastogenic effects in vitro suggest involvement of a genotoxic mechanism.

1-Methylnaphthalene and 2-methylnaphthalene show only low toxicity; the oral LD_{50} values in rats were determined at 2800 and 3850 mg/kg, respectively [39]. At high oral doses in rats, acute symptoms are unspecific and indicate impairment of the nervous system and hemorrhagic edema of inner organs. After intraperitoneal injection into mice, epithelial necrosis of the lung was induced, morphologically similar to that found with naphthalene [40], [41]. Other specific phenomena that are characteristic of naphthalene exposure, i.e., hemolytic anemia and formation of cataracts, are unknown for methylnaphthalenes. In humans, no intoxications related to these aromatic compounds have been reported.

Diisopropylnaphthalenes (mixture of isomers, KMC) cause no toxic symptoms in rats up to high dosage levels: in the range from about 3.0 g/kg under acute conditions to 30 mg kg^{-1} d^{-1} in a long-term study for 24 months. There are generally involved unspecific, mostly reversible phenomena such as loss of body weight, digestive dysfunction, enlargement of the liver, and local disturbed circulation (hyperemia) [42]. In skin and eye irritation tests conducted on rabbits, diisopropylnapthtalenes exhibited slightly irritant, but reversible effects, which allows the product to be classified as nonirritant. Likewise, there is no evidence of sensitization [42]. No adverse effects were observed in a group of 80 workers who had had contact with diisopropylnapthalenes during production of carbonless copy paper [42]. Mutagenicity and carcinogenicity tests (24 months, oral, rat) gave no evidence of a carcinogenic potential, nor were teratogenic malformations found (mice, oral) [42]. Diisopropylnaphthalenes were rapidly released from the body, i.e., > 80 % within 15 h (mice) [42], and within 10 d (carp) [43], and they proved to be biodegradable [44].

Tetralin and decalin are irritant to the skin, eye, and to the mucous membranes on inhalation. They may cause nervous disturbances, headache, or numbness. The lowest concentration of decalin in air that exhibited an effect on humans was ca. 100 ppm; a 4 h exposure to 500 ppm was lethal to 4 from 6 rats [45]. In guinea pigs, morphological changes of the kidney, liver, and lung were observed after subacute inhalation of decalin [45]. Methemoglobinemia was induced by tetralin in cats; the same effect was found in infants after accidential exposure to tetralin [46]. The color of human urine may change to brownish green or green-gray after incorporation of decalin or tetralin [45]. Acute range finding studies revealed an oral LD_{50} (rat) of 4.17 g/kg for decalin and 2.68 g/kg for tetralin [45]. A TLV of 25 ppm has been suggested for decalin [47].

[Top of Page]

References

General References

1. *Beilstein,* naphthalene **5** 531, **5**(1), 257, **5**(3) 432, **5**(3) 1549, **5**(4) 1640; 1-methylnaphthalene **5** 566, **5**(1) 265, **5**(2) 460, **5**(3) 1617, **5**(4) 1687; 2-methylnaphthalene **5** 567, **5**(1) 266, **5**(2) 460, **5**(3) 1627, **5**(4) 1693; 2,6-dimethylnaphthalene **5** 570, **5**(1) 268, **5**(2) 468, **5**(3) 1651, **5**(4) 1714; 2-isopropylnaphthalene **5** 571, **5**(2) 470, **5**(3) 1657, **5**(4) 1723; 2,6-diisopropylnaphthalene **5**(4) 1768; tetralin **5** 491, **5**(1) 236, **5**(2) 382, **5**(3) 1219, **5**(4) 1388; decalin **5** 92, **5**(1) 46, **5**(2) 56, **5**(3) 242, **5**(4) 310.
2. H.-G. Franck, G. Collin: *Steinkohlenteer*, Springer Verlag, Heidelberg/Berlin/New York 1968, pp. 60 – 61, 180 – 181.
3. H.-G. Franck, J. W. Stadelhofer: *Industrial Aromatic Chemistry*, Springer Verlag, Berlin 1988, pp. 298 – 308, 328 – 333, 336 – 339.
4. G.-P. Blümer, G. Collin, *Erdöl, Kohle, Erdgas, Petrochem.* **36** (1983) 22 – 27.
5. G. Collin, M. Zander, *Nat. Resour. Dev.* **27**, 61 – 81 (1988).

Specific References

6. Rütgerswerke AG, DE 1 815 568, 1968 (O. Wegener, R. Oberkobusch, G. Collin, M. Zander, H. Buffleb); DE 3 442 275, 1984 (J. Talbiersky, N. Drescher, H. E. Carl).
7. F. Asinger: Die petrolchemische Industrie, vol. 1, Akademie-Verlag, Berlin 1971, pp. 603 – 609.
8. G. F. Asselin, R. A. Erickson, *Chem. Eng. Prog.* **58** (1962) 47 – 52. Links
9. Rütgerswerke AG, DE 3 028 199, 1980 (G.-P. Blümer).
10. Rütgerswerke AG, DE 3 723 104, 1987 (J. Weitkamp, M. Neuber, W. Höltmann, G. Collin).
11. Rütgerswerke AG, DE 3 519 009, 1985 (F. Kajetanczyk, R. Steinbach, I. Ruppert, K. Schlich); DE 3 519 009, 1985 (R. Steinbach, I. Ruppert, K. Schlich).
12. Mitsubishi Gas, JP 87 273 260, 1987 (T. Hirai, S. Kitaoka).
13. Rütgerswerke AG, DE 3 701 960, 1987 (R. Steinbach, F. Kajetanczyk).
14. Max-Planck-Gesellschaft, DE 3 120 099, 1981 (H. G. Schlossberger).
15. Hoechst, DE 3 334 084, 1983 (K. Eichler, E. I. Leupold).
16. Rütgerswerke AG, EP 280 055, 1988 (J. Weitkamp, M. Neuber, W. Höltmann, G. Collin, H. Spengler).
17. Rütgerswerke AG, DE 3 520 841, 1985 (G. Schmitt, K.-R. Kurtz).
18. K. Handrick, *Proc. Neue Verfahren Kohleveredlung,* Luxemburg, November 26 – 28, 1979, pp. 27 – 37. Bergwerksverband, DE 2 644 624, 1976 (R. C. Schulz, D. Engel).
19. Kureha, GB 1 359 512, 1974 (A. Konishi, M. Takahashi, F. Kimura, T. Toguchi).
20. Rütgerswerke AG, DE 3 735 976, 1987 (R. Zellerhoff, M. Gürtler).
21. J. W. Stadelhofer, R. B. Zellerhoff, *Chem. Ind. (London)*, 1989, April 3, 208 – 211.
22. Rütgerswerke AG, EP 216 009, 1986 (W. Höltmann, R. Zellerhoff, R. Oberkobusch, P. Stäglich, B. Charpey).
23. Teijin, JP 88 63 150, 1986 (I. Hirose).
24. Kureha, DE 3 531 982, 1986 (T. Yamauchi, S. Hayashi, A. Sasakawa).
25. Amoco, EP 329 273, 1989 (P. A. Sanchez, D. A. Young, G. E. Kuhlmann, W. Partenheimer, W. P. Schammel).
26. Kureha, DE 3 517 158, 1985 (A. Iizuka, Y. Konai, T. Yamauchi, S. Hayashi).
27. Celanese, EP 172 012, 1986 (D. E. Stuetz).
28. W. W. Zuelzer, L. Apt, *J. Am. Med. Assoc.* **141** (1949) 185.
29. H. W. Gerarde (ed.): *Toxicology and Biochemistry of Aromatic Hydrocarbons*, Elsevier, Amsterdam/London/New York/Princeton 1960, pp. 225 – 232.
30. R. van Heyningen, A. Pirie, *Biochem. J.* **102**, (1967) 842. Links
31. S. J. Fanbury, *Arch. Dermatol. Syphil.* **42** (1940) 53 – 55.
32. D. Mahvi *et al.*, *Am. J. Pathol.* **86** (1977) 559. Links
33. D. L. Warren *et al.*, *Chem. Biol. Interact.* **40** (1982) 287. Links
34. E. Knake, *Virchows Arch.* **329** (1956) 141. Links
35. B. Adkins *et al.*, *J. Toxicol. Environ. Health* **17** (1986) 311. Links
36. United States Environmental Protection Agency (EPA), Rep. No. 600/8–87/-055 F, Washington D.C., 1987.
37. R. B. Franklin in R. Snyder (ed.): *Ethel Browning's Toxicity and Metabolism of Industrial Solvents*, Elsevier, Amsterdam/New York/Oxford 1987.
38. Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker: *Naphthalin, BUA-Stoffbericht 39*, VCH Verlagsgesellschaft, Weinheim, Germany 1989.
39. Huntingdon Research Centre (by order of Rütgerswerke AG), Münster, Federal Republic of Germany, 1979 (unpublished results).
40. In [36] p. 176.
41. A. R. Buchpitt, R. B. Franklin, *Pharmacol. Ther.* **41** (1989) 393. Links
42. Rütgers Kureha Solvents: *Toxicological and Physiochemical Studies on KMC*, Duisburg 1985 (unpublished results).
43. T. Yoshida, H. Kojima, *Chemosphere* (1978) no. 6, 491 – 496. Links
44. T. Yoshida, H. Kojima, *Chemosphere* (1978) no. 6, 497 – 501. Links
45. *Patty*, vol. **2B**, pp. 3233, 3240 – 3242.
46. *Ullmann*, 3rd ed., **12**, p. 589.
47. D. J. de Renzo (ed.): *Solvents Safety Handbook*, Noyes Data, Park Ridge, N.J. 1986, pp. 174, 626.

48. BUA (Beratergremium umweltrelevanter Altstoffe der Gesellschaft Deutscher Chemiker): Naphthalin, BUA Stoffbericht 39, VCH Verlagsgesellschaft, Weinheim 1989.
49. IPCS (International Programme on Chemical Safety): Environmental Health Criteria 202: Selected Non-heterocyclic Polycyclic Aromatic Hydrocarbons, World Health Organization, Geneva 1998.
50. DFG: Deutsche Forschungsgemeinschaft. *Naphthalin*, Gesundheitsschädliche Arbeitsstoffe, H. Greim (ed.): Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten, Wiley-VCH, Weinheim, Germany 2001.
51. EU (European Union): Risk Assessment Report Naphthalene, October 2001 (R020_0110_env_hh).
52. MA Schulz, PV Choudary, AR Buckpitt: “Role of murine cytochrome P-450 2F2 in metabolic activation of naphthalene and metabolism of other xenobiotics,” *J. Pharmacol. Exp. Ther.* **290** (1999) 281 – 290. Links
53. AS Wilson, CD Davis, DP Williams, AR Buckpitt, MM Pirmohamed, BK Park: “Characterization of the toxic metabolite(s) of naphthalene,” *Toxicology* **114** (1996) 233 – 234. Links
54. J Zheng, M Cho, AD Jones, BD Hammock: “Evidence of quinone metabolites of naphthalene covalently bound to sulfur nucleophiles of proteins of murine clara cells after exposure to naphthalene,” *Chem. Res. Toxicol.* **10** (1997) 1008 – 1014. Links
55. JC Sasaki, J Arey, DA Eastmond, KK Parks, AJ Grosowsky: “Genotoxicity induced in human lymphoblasts by atmospheric reaction products of naphthalene and phenanthrene,” *Mutat. Res.* **393** (1997) 23 – 35. Links
56. M Bagchi et al.: “Naphthalene-induced oxidative stress and DNA damage in cultured macrophage J774A.1 cells,” *Free Radical. Biol. Med.* **25** (1998) 137 – 143.
57. LS Gollahon, P Iyer, JE Martin, TR Irvin: “Chromosomal damage to preimplantation embryos in vitro by naphthalene,” *Toxicologist* **10** (1990) 274.
58. JE Martin, P Iyer, TR Irvin: “Development of preimplantation rodent embryo culture systems for identification of developmentally-toxic chemical agents,” *In Vitro Cell Dev Biol* **26**, 20A, (1990).
59. NTP: National Toxicology Program, TR-500: Toxicology and Carcinogenesis Studies of Naphthalene (CAS No. 91-20-3) in F344/N Rats (Inhalation Studies), December 2000.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Naphthalene Derivatives

Alkylnaphthalenes →Naphthalene and Hydronaphthalenes ; Chloronaphthalenes →Chlorinated Hydrocarbons ; Cyanonaphthalenes Nitriles; Hydronaphthalenes →Naphthalene and Hydronaphthalenes ; Naphthalenedicarboxylic Acids →Carboxylic Acids, Aromatic.

GERALD BOOTH, Booth Consultancy Services, Thorpe House, Uppermill, Oldham OL3 6DP, United Kingdom

1. Introduction 1
2. Naphthalenesulfonic Acids 2
2.1. Production and Properties 2
2.2. Monosulfonic Acids 3
2.3. Disulfonic Acids 5
2.4. Tri- and Tetrasulfonic Acids 6
2.5. Alkylnaphthalenesulfonic Acids 7
3. Naphthols 7
3.1. 1-Naphthol 7
3.2. 2-Naphthol 9
3.3. Naphthalenediols 11
3.4. Hydroxynaphthoic Acids 12
4. Hydroxynaphthalenesulfonic Acids . 14
4.1. Production and Properties 15
4.2. 1-Hydroxynaphthalenesulfonic Acids 17
4.3. 2-Hydroxynaphthalenesulfonic Acids 19
4.4. 1-Hydroxynaphthalenedisulfonic Acids 20
4.5. 2-Hydroxynaphthalenedisulfonic Acids 21
4.6. Hydroxynaphthalenetrisulfonic Acids 23
4.7. Dihydroxynaphthalenesulfonic Acids 23
4.8. Dihydroxynaphthalenedisulfonic Acids 24
5. Aminonaphthalenes 25
5.1. Naphthylamines 25
5.2. Naphthalenediamines 27
6. Aminonaphthalenesulfonic Acids . . . 28
6.1. Production and Properties 28
6.2. 1-Aminonaphthalenesulfonic Acids . . 31
6.3. 2-Aminonaphthalenesulfonic Acids . . 34
6.4. 1-Aminonaphthalenedisulfonic Acids 35
6.5. 2-Aminonaphthalenedisulfonic Acids 36
6.6. Aminonaphthalenetrisulfonic Acids . 37
6.7. Diaminonaphthalenesulfonic Acids . . 39
6.8. Diaminonaphthalenedisulfonic Acids 40
6.9. Toxicity 40
7. Aminonaphthols 41
8. Aminohydroxynaphthalenesulfonic Acids 42
8.1. Production and Properties 42
8.2. Aminohydroxynaphthalenemonosulfonic Acids 43
8.3. Aminohydroxynaphthalenedisulfonic Acids 49
9. References 52

1. Introduction

Although the basis of naphthalene chemistry can be said to have started with ERLENMEYER in 1866, conclusive chemical evidence for the structure of naphthalene and its early derivatives (e.g., mono- and dinitroderivatives and α-naphthylamine) was published in 1888 by REVERDIN and NOELTING [5].

By 1900, hundreds of new naphthalene derivatives had been prepared as components for azo dyes. Not only was this early work, reported in German patents and journals, remarkably accurate in view of the comparatively primitive analytical techniques available, but the key compounds that emerged at that time remain dominant to this day. The trivial names given to these dye intermediates (Table 1) have also survived and are commonly used in the industry.

Table 1. Letter acids and other code-named naphthalene intermediates

A acid	3,5-dihydroxynaphthalene-2,7-disulfonic acid
Amino epsilon acid	1-aminonaphthalene-3,8-disulfonic acid
Amino F acid	2-aminonaphthalene-7-sulfonic acid
Amino G acid	2-aminonaphthalene-6,8-disulfonic acid
Amino J acid	2-aminonaphthalene-5,7-disulfonic acid
Amino R acid	2-aminonaphthalene-3,6-disulfonic acid
Armstrong acid	naphthalene-1,5-disulfonic acid
B acid	1-aminonaphthalene-4,6,8-trisulfonic acid
Badische acid	2-aminonaphthalene-8-sulfonic acid
BON acid	3-hydroxy-2-naphthoic acid
Böniger acid	1-amino-2-hydroxynaphthalene-4-sulfonic acid

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_009

Table 1. (Continued)

Bronner acid	2-aminonaphthalene-6-sulfonic acid
C (Cassella) acid	2-aminonaphthalene-4,8-disulfonic acid
Chicago (2 S) acid	1-amino-8-hydroxynaphthalene-2,4-disulfonic acid
Chromotropic acid	1,8-dihydroxynaphthalene-3,6-disulfonic acid
1,6-Cleve's acid	1-aminonaphthalene-6-sulfonic acid
1,7-Cleve's acid	1-aminonaphthalene-7-sulfonic acid
Crocein (Bayer) acid	2-hydroxynaphthalene-8-sulfonic acid
Cyanol	1-amino-7-naphthol
D (Dahl's) acid	2-aminonaphthalene-5-sulfonic acid
Dahl's acid II	1-aminonaphthalene-4,6-disulfonic acid
Dahl's acid III	1-aminonaphthalene-4,7-disulfonic acid
Delta (δ) acid	1-hydroxynaphthalene-4,8-disulfonic acid
Epsilon (ε) acid	1-hydroxynaphthalene-3,8-disulfonic acid
F acid	2-hydroxynaphthalene-7-sulfonic acid
Freund's acid (1,3,6)	1-aminonaphthalene-3,6-disulfonic acid
Freund's acid (1,3,7)	1-aminonaphthalene-3,7-disulfonic acid
G acid	2-hydroxynaphthalene-6,8-disulfonic acid
Gamma (γ) acid	2-amino-8-hydroxynaphthalene-6-sulfonic acid
H acid	1-amino-8-hydroxynaphthalene-3,6-disulfonic acid
J acid	2-amino-5-hydroxynaphthalene-7-sulfonic acid
K acid	1-amino-8-hydroxynaphthalene-4,6-disulfonic acid
Kalle's acid	1-aminonaphthalene-2,7-disulfonic acid
Koch acid	1-aminonaphthalene-3,6,8-trisulfonic acid
Laurent's (L) acid	1-aminonaphthalene-5-sulfonic acid
M acid	1-amino-5-hydroxynaphthalene-7-sulfonic acid
Naphthionic acid	1-aminonaphthalene-4-sulfonic acid
NW (Nevile and Winther) acid	1-hydroxynaphthalene-4-sulfonic acid
Oxy Chicago acid	1-hydroxynaphthalene-4,8-disulfonic acid
Oxy Koch acid	1-hydroxynaphthalene-3,6,8-trisulfonic acid
Oxy L acid	1-hydroxynaphthalene-5-sulfonic acid
Oxy Tobias acid	2-hydroxynaphthalene-1-sulfonic acid
Peri acid	1-aminonaphthalene-8-sulfonic acid
Purpurol	1-amino-5-naphthol
R acid	2-hydroxynaphthalene-3,6-disulfonic acid
2 R (Columbia) acid	2-amino-8-hydroxynaphthalene-3,6-di-sulfonic acid
RM acid	2-amino-3-hydroxynaphthalene-6-sulfonic acid
S acid	1-amino-8-hydroxynaphthalene-4-sulfonic acid
Schaeffer acid	2-hydroxynaphthalene-6-sulfonic acid
T acid	1-aminonaphthalene-3,6,8-trisulfonic acid
Tobias acid	2-aminonaphthalene-1-sulfonic acid
Violet (RG) acid	1-hydroxynaphthalene-3,6-disulfonic acid

The important Armstrong – Wynne rules for polysubstitution of naphthalene and its derivatives by nitration and sulfonation were also formulated in this early period. Although empirical, they have stood the test of time against increasingly sophisticated theories and calculations.

Improved processes based on the key unit processes of sulfonation, nitration, reduction, hydroxylation, and amination (Bucherer reaction) for a wide range of naphthalene derivatives were developed during the next 40 years [6]. Although these were not published [7] until after 1945 with the end of the I.G. Farbenindustrie era, all the important subsequent reviews [5], [8–10] were based largely on these comprehensive data.

The only significant production development work published in the last 40 years is related to the project for the new Schelde-Chemie plant (Bayer Ciba-Geigy joint venture) designed to manufacture 14 000 t/a of naphthalene intermediates at Brunsbüttel [16]. The process development on important letter acids (H, J, γ, C, Peri, and Laurent's) required a major change from the traditional processes, optimized over 80 years, to meet present-day energy and environmental requirements [17].

Azo dyes and pigments continue to be major outlets for naphthalene intermediates (see also →Azo Dyes). The Colour Index lists some 270 different naphthalene intermediates as precursors to many more colorants [18]. This represents about 20 % of the total list of intermediates. Supplementary updating volumes list only seven new naphthalene intermediates and relatively few new outlets for existing intermediates [19]. From this one may infer that very little new colorant research is being carried out based on novel naphthalene derivatives. In contrast, other areas such as agrochemicals and pharmaceuticals have been most active in exploiting new naphthalene derivatives over the last 20 years [11].

2. Naphthalenesulfonic Acids

2.1. Production and Properties

Controlled sulfonation using a range of sulfuric acid and oleum strengths under a variety of reaction conditions leads to formation of mono-, di-, tri-, and tetrasulfonic acids, whose separation is frequently complicated by desulfonation (i.e., reverse sulfonation) or isomerization.

Table 2. Salts of key naphthalenesulfonic acids

Salt	Hydrate	Solubility, g in 100 mL H_2O
1-Sulfonic acid		
Sodium	1/2 H_2O	9.1 (10 °C)
Potassium	1/2 H_2O	7.7 (10 °C)
Calcium	2 H_2O	6.1 (10 °C)
Barium	1 H_2O	1.2 (10 °C)
2-Sulfonic acid		
Sodium		5.9 (25 °C)
Potassium	1/2 H_2O	6.7 (10 °C)
Calcium	1 H_2O	1.3 (10 °C)
Barium	1 H_2O	0.35 (10 °C)
1,5-Disulfonic acid		
Sodium	2 H_2O	11.1 (18 °C)
Potassium	2 H_2O	6.7 (18 °C)
Calcium	2 H_2O	2.5 (18 °C)
Barium	1 H_2O	0.21 (18 °C)
1,6-Disulfonic acid		
Sodium	7 H_2O	33.3 (18 °C)
Potassium		20.0 (18 °C)
Calcium	4 H_2O	10.0 (18 °C)
Barium	31/2 H_2O	6.2 (100 °C)
2,6-Disulfonic acid		
Sodium	1 H_2O	11.9 (18 °C)
Potassium		5.2 (18 °C)
Calcium		6.2 (18 °C)
Barium	1 H_2O	<0.1
2,7-Disulfonic acid		
Sodium	6 H_2O	45 (18 °C)
Potassium	2 H_2O	71 (18 °C)
Calcium	6 H_2O	16 (18 °C)
Barium	2 H_2O	1.2 (19 °C)

A common method for isolating naphthalenesulfonic acids and substituted derivatives obtained by sulfonation is the liming-out process. This consists of neutralizing the quenched sulfonation mass (i.e., after addition to excess water) with lime and filtering off the precipitated calcium sulfate while still hot. The solution of the calcium salt of the product is then titrated with sodium carbonate to form the sodium salt. Precipitated calcium carbonate is removed by filtration, and the solution is either evaporated to give the solid sodium salt or used to precipitate a less soluble salt. This technique may be employed in the laboratory to give pure products by using barium instead of calcium, provided the barium salt of the product has sufficient solubility in hot water.

The twelve sulfonated products most readily obtained from naphthalene are shown in Figure 1.

Separation of naphthalenesulfonic acids and downstream derivatives for analysis may be accomplished on small-scale columns by ion-exchange [20], ion-pair [21], or high-performance liquid [17] chromatography.

The naphthalenesulfonic acids are all extremely water soluble and are strong acids in solution. Alkali-metal and alkaline-earth salts are also water soluble but with a progression toward sparingly soluble barium salts, as shown in Table 2. The moderately soluble salts are readily precipitated from the mineral acid solution, with variations in crystal form according to conditions.

Naphthalene-2-sulfonic acid is isolated for conversion to 2-naphthol, but the sulfonation mass is further sulfonated in situ for the production of di- and trisulfonic acids. Naphthalene-1-sulfonic acid is not usually isolated on a large scale because 1-naphthol is produced advantageously by other routes. The 1-isomer is, however, nitrated in situ for the manufacture of Peri and Laurent's acids (see Chap. 6).

Naphthalenedisulfonic acids are converted into naphthol sulfonic acids (Chap. 4) and dihydroxynaphthalenes (Section 3.3), or are nitrated to form intermediates for aminonaphthalenedisulfonic acids and aminonaphtholsulfonic acids (Chap. 8). Naphthalenetrisulfonic acids similarly afford aminohydroxynaphthalenedisulfonic acids (e.g., H acid). The one readily accessible naphthalenetetrasulfonic acid has no large-scale application.

2.2. Monosulfonic Acids

Naphthalene-1-sulfonic acid [*85-47-2*] (**1**), naphthalene-α-sulfonic acid, $C_{10}H_8O_3S$, M_r 208.23, crystallizes from aqueous solutions as the dihydrate, *mp* 90 °C; dehydration (P_2O_5, 1.33 kPa) gives the anhydrous acid, *mp* 139 – 140 °C. Aqueous solubilities of the alkali-metal salts are given in Table 2.

Naphthalene-1-sulfonic acid is desulfonated to naphthalene by heating with dilute mineral acid. Halogenation also removes the sulfonic acid group, mainly with formation of 1,5-dihalonaphthalenes. Naphthalene-1-sulfonyl chloride, *mp* 68 °C, is prepared by treatment of the anhydrous acid with phosphorus pentachloride at 100 °C.

Production. Naphthalene-1-sulfonic acid is produced by adding naphthalene to 96 % sul-

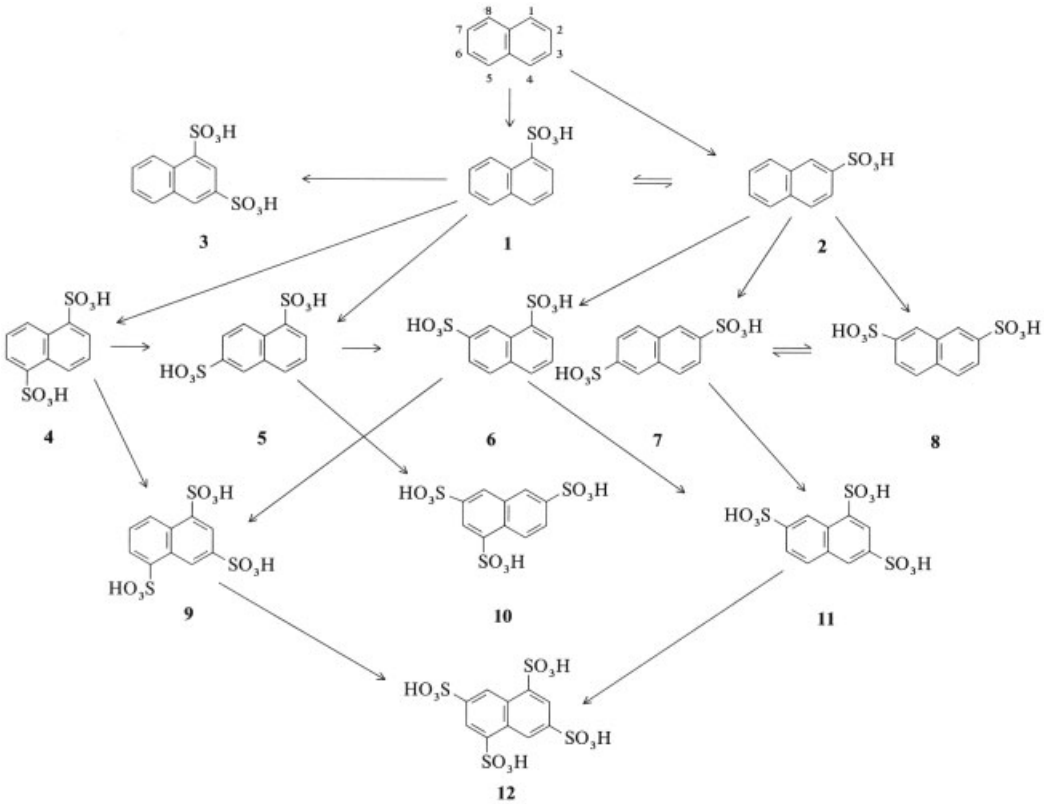

Figure 1. Sulfonation of naphthalene

furic acid at 20 °C and slowly raising the temperature to 70 °C [7]. After 3 h at 70 – 75 °C the reaction mass is poured into water and limed out. A technical-grade product is obtained by evaporation and a purer product by precipitating the aniline salt.

An alternative sulfonation process uses sulfur trioxide in a solvent such as tetrachloroethane.

Uses. Naphthalene-1-sulfonic acid is further sulfonated or nitrated without isolation. At lower temperature (35 °C), sulfuric acid with oleum yields the 1,5-disulfonic acid, whereas at higher temperature (100 °C) the 1,6-disulfonic acid is predominantly obtained. Nitration gives mainly the 5-nitro and 8-nitro derivatives (see Section 6.1). Production of 1-naphthol by caustic fusion (hydroxylation) has long since been superseded by alternatives for reasons of product quality (Section 3.1).

Naphthalene-2-sulfonic acid [*120-18-3*] (**2**), naphthalene-β-sulfonic acid, $C_{10}H_8O_3S$, M_r 208.23, crystallizes from aqueous solutions as the hydrate (*mp* 124 °C) or trihydrate (*mp* 83 °C). Aqueous solubilities of the free acid and its salts are somewhat lower than those of the corresponding α-isomers (Table 2). The isolated sodium salt is known as β salt.

Naphthalene-2-sulfonic acid is desulfonated by hot aqueous mineral acids with much more difficulty than naphthalene-1-sulfonic acid; the rate has been measured at 50 times slower [22]. Halogenation attacks either or both the 5- and the 8-positions. Bromination may be used to separate a mixture of naphthalene-1-sulfonic acid from naphthalene-2-sulfonic acid because the latter forms a soluble product whereas the former is desulfonated.

Production. Molten naphthalene is added to 96 % sulfuric acid in an iron vessel and the mixture is agitated at 163 °C for 2 h. A complex

work-up consisting of gradual dilution, heating, and neutralization with caustic soda and sodium sulfite ensures an 88 % yield of isolated β salt and desulfonation of the α-isomer, which is formed as a coproduct in 5 – 10 % yield [7].

Uses. The major outlet for β salt is the production of 2-naphthol (Section 3.2). The 2-sulfonic acid is a stage in the production of the 1,6-, 2,6-, and 2,7-disulfonic acids and the 1,3,6-trisulfonic acid. Nitration gives primarily the 5- and 8-nitro derivatives as intermediates for Cleve's acids (Section 6.1). The 2-sulfonic acid condenses with formaldehyde or alcohols to form surface-active agents (see Section 2.5).

The intermediate 2-thionaphthol can be obtained by catalytic hydrogenation of naphthalene-2-sulfonic acid, but is traditionally made by zinc reduction of naphthalene-2-sulfonyl chloride.

2.3. Disulfonic Acids

Naphthalene-1,3-disulfonic acid
[*6094-26-4*] (**3**), $C_{10}H_8O_6S_2$, M_r 288.28, is produced only in small proportion on disulfonation of naphthalene and cannot be separated. It is prepared in good yield in the laboratory by diazotization and copper-catalyzed deamination of 2-aminonaphthalene-6,8-disulfonic acid [23].

Naphthalene-1,5-disulfonic acid
[*81-04-9*] (**4**), Armstrong acid, $C_{10}H_8O_6S_2$, M_r 288.28, crystallizes as the tetrahydrate but dehydrates at 125 °C to give the anhydrous acid (*mp* 240 – 245 °C). Solubilities of the alkali-metal salts are given in Table 2.

The disulfonic acid forms an anhydride when treated with oleum below 50 °C. This anhydride reacts with ammonia to give a mixture of naphthalene-1,5-disulfonamide and naphthalene-1-sulfonamide-5-sulfonic acid. Treatment of the disodium salt of naphthalene-1,5-disulfonic acid with phosphorus pentachloride at 110 °C gives naphthalene-1,5-disulfonyl chloride (*mp* 184 °C), which may also be prepared by chlorosulfonation of naphthalene. Reaction of an aqueous solution of naphthalene-1,5-disulfonic acid with bromide – bromate gives 1,5-dibromonaphthalene so readily that the yellow precipitate may be used to identify very low concentrations of the disulfonic acid.

Production. Naphthalene is mixed with 20 % oleum at 20 – 35 °C followed by gradual addition of 65 % oleum and further naphthalene alternately. After heating for 6 h at 55 °C the reaction mixture is added to water, and the product is precipitated as the free acid by cooling or as the disodium salt by the addition of alkaline sodium sulfate. In each case, the isolated yield is ca. 53 %. The 1,6-disulfonic acid can be recovered from the filtrate.

Higher yields of **4** have been reported for alternative processes; for example, solvent sulfonation with sulfur trioxide and chlorosulfonic acid in a solvent such as tetrachloroethylene [24] or sulfonation with oleum through a double screw-feed apparatus [25].

Uses. The disodium salt of naphthalene-1,5-disulfonic acid gives 1-hydroxynaphthalene-5-sulfonic acid or 1,5-dihydroxynaphthalene (see Section 3.3) by fusion processes. Further sulfonation of the disulfonated reaction mixture gives naphthalene-1,3,5-trisulfonic acid (**9**), and nitration leads to 1- and 2-nitronaphthalene-4,8-disulfonic acids as intermediates to the corresponding aminonaphthalene disulfonic acids (Chap. 6).

Naphthalene-1,5-disulfonic acid is an important stabilizer for diazo compounds [26].

Naphthalene-1,6-disulfonic acid
[*525-37-1*] (**5**), $C_{10}H_8O_6S_2$, M_r 288.28, crystallizes as the tetrahydrate, *mp* 125 °C (decomp.). Solubilities of the alkali-metal salts are given in Table 2.

Production. Separation from the 1,5-disulfonic acid process (see above) usually meets requirements for the preparation of **5**. It can also be prepared by sulfonation of naphthalene-2-sulfonic acid with oleum, which gives naphthalene-1,7-disulfonic acid (**6**) as a byproduct [12].

Uses. Alkali fusion of **5** yields 1,6-dihydroxynaphthalene (Section 3.3), and sulfonation with oleum produces naphthalene-1,3,6-trisulfonic acid (**10**). However, the main outlet for naphthalene-1,6-disulfonic acid is via nitration to 1-nitronaphthalene-3,8-disulfonic acid, which is the intermediate for 1-aminonaphthalene-3,8-disulfonic acid (Section 6.4).

Naphthalene-1,7-disulfonic acid
[*5724-16-3*] (**6**), $C_{10}H_8O_6S_2$, M_r 288.28, is

produced by sulfonation of naphthalene-2-sulfonic acid (**2**) with cold oleum to give a mixture of 1,6- (80 %) and 1,7- (20 %) acids. Although these are difficult to separate, the corresponding disulfonyl chlorides are separable because of their differing solubilities in, for example, benzene.

Naphthalene-2,6-disulfonic acid [*581-75-9*] (**7**), $C_{10}H_8O_6S_2$, M_r 288.28, crystallizes as the dihydrate, *mp* 129 °C. The alkali-metal salts are much less soluble than those of the 2,7-isomer (Table 2).

Production. Naphthalene is added to sulfuric acid – monohydrate ($SO_3 \cdot H_2O$) mixture at 135 °C, and the reaction is completed by heating at 170 – 175 °C for 5 h. After quenching in water, the disodium salt of **7** is precipitated at 90 °C by adding sodium chloride and sodium sulfate. Filtration at 95 °C followed by washing with sodium sulfate solution gives the required product, free from the 2,7-isomer, in 21 % yield. The 2,7-isomer is then isolated from the filtrates.

Uses. The disodium salt of **7** gives 2-hydroxynaphthalene-6-sulfonic acid or 2,6-dihydroxynaphthalene (Section 3.3) by fusion processes. Nitration of the crude sulfonation product containing 2,6- and 2,7-disulfonic acids gives the 4-nitro derivatives, which, on reduction, give the mixed 1-aminonaphthalene-3,6(3,7)-disulfonic acid (Freund's acid).

Naphthalene-2,7-disulfonic acid [*92-41-1*] (**8**), $C_{10}H_8O_6S_2$, M_r 288.28; the sodium salt of **8** is much more soluble than that of the 2,6-isomer (Table 2) but can be isolated and purified by precipitation of the calcium salt at lower temperature followed by recrystallization.

Production. The 2,7-disulfonic acid is always the main product of high-temperature sulfonation of naphthalene, but at 160 °C it converts slowly to an equilibrium mixture containing ca. 30 % of the 2,6-isomer [12]. Crystallization of the pure product from conventional sulfonation is difficult, and alternative processes, (for example, reaction of naphthalene vapor with sulfur trioxide in the gas phase at 220 °C) are said to give a purer product.

Uses. The disodium salt of **8** gives 2-hydroxynaphthalene-7-sulfonic acid or 2,7-dihydroxynaphthalene (Section 3.3) by fusion processes. Sulfonation by oleum gives the 1,3,6-trisulfonic acid, and nitration gives 1-nitronaphthalene-3,6-disulfonic acid.

In sulfuric acid solution, the disodium salt of naphthalene-2,7-disulfonic acid undergoes an unusual reaction at the 4-position with 4,4′-bis-(dimethylamino)benzhydrol. The resulting condensation product is oxidized with lead dioxide to give the acid basic dye C.I. Acid Green 16.

2.4. Tri- and Tetrasulfonic Acids

Naphthalene-1,3,5-trisulfonic acid [*6654-64-4*] (**9**), $C_{10}H_8O_9S_3$, M_r 368.34, is extremely soluble in water; the sodium salt crystallizes as the tetrahydrate.

Production. Naphthalene and 65 % oleum are added simultaneously to sulfuric acid monohydrate over a period of 32 h at 30 – 35 °C. The mixture is then heated at 50 °C for 1 h, 70 °C for 1 h, and finally 90 °C for 7 h. This procedure minimizes formation of the 1,3,6-isomer, but a purer product may be obtained by using isolated naphthalene-1,5-disulfonic acid as starting material.

Uses. The total reaction mixture is nitrated and the resulting 8-nitro derivative is reduced to give a mixture of 1-aminonaphthalene-3,6,8- and 4,6,8-trisulfonic acids.

Naphthalene-1,3,6-trisulfonic acid [*86-66-8*] (**10**), $C_{10}H_8O_9S_3$, M_r 368.34, crystallizes as the hexahydrate, *mp* 170 °C.

Production. Naphthalene is initially sulfonated with sulfuric acid monohydrate under programmed temperature control between 80 and 145 °C. Then, 65 % oleum is added gradually at 40 °C and heating is continued for 2.5 h at 145 °C. Trisulfonation is complete after the further addition of 65 % oleum and heating for 3 h at 150 °C. This complex procedure [7] maximizes the formation of **10** with a conversion of up to 75 %.

Uses. The total reaction mixture is nitrated, and the resulting 8-nitro derivative is reduced to give 1-aminonaphthalene-3,6,8-trisulfonic acid (Koch acid), the key intermediate in the production of the important H acid. The crude trisodium salt may be precipitated from the quenched mixture after dilution to 75 % acid strength to give a product (Azoguard) used as a diazo stabilizer.

Naphthalene-1,3,7-trisulfonic acid [*85-49-4*] (**11**), $C_{10}H_8O_9S_3$, M_r 368.34, is much less important than the 1,3,5- and 1,3,6-isomers. It is produced by sulfonation of naphthalene-2,6-disulfonic acid (**7**) with 65 % oleum in sulfuric acid monohydrate [7].

Naphthalene-1,3,5,7-tetrasulfonic acid [*6654-67-7*] (**12**), $C_{10}H_8O_{12}S_4$, M_r 448.4, has no technical importance although it is easily obtained as the end product of the oleum sulfonation of naphthalene, along with the 1,3,7-trisulfonic acid (**11**). In the latter, unlike the 1,3,5- and 1,3,6-trisulfonic acids, direct entry of a fourth sulfonic acid group does not occur.

2.5. Alkylnaphthalenesulfonic Acids

A wide range of mixed alkylnaphthalenesulfonic acids of undefined or undisclosed constitution are produced as surface-active wetting agents and dispersants. For example, methylnaphthalenes are sulfonated with sulfuric acid at 160 °C, or naphthalenesulfonic acids are alkylated with alkenes, higher alcohols, or formaldehyde. Propene or butene produce isopropyl- or butylnaphthalene sulfonates, respectively. Formaldehyde gives the dimer **13** as a simple product from β salt, or polymeric condensation products (**14**), which are important as synthetic tanning agents or as plasticizers for concrete. Dinonylnaphthalene, prepared from naphthalene and nonene, is sulfonated to give derivatives that are important as lubricating oil additives [13].

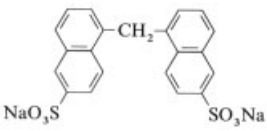

13

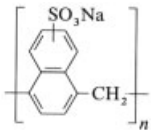

14

3. Naphthols

Naphthols are the most important naphthalene derivatives because they are key intermediates in the production of many chemicals other than dyes and pigments. World production is estimated at 40 000 t/a of 1-naphthol and 100 000 t/a of 2-naphthol.

The production methods for naphthols are as follows (the first four methods are analogous to those for phenols):

1) Alkali fusion of naphthalenesulfonic acids
2) Alkaline hydrolysis of chloronaphthalenes
3) Hydroperoxidation of 2-isopropylnaphthalene to form 2-naphthol and acetone (not applicable to 1-naphthol)
4) Dehydrogenation of tetralones
5) Direct, single-step replacement of the amino group of naphthylamine by a hydroxyl group

The last method is not applicable to phenols, for which the corresponding reaction requires two steps, diazotization of the aminobenzene followed by hydrolysis.

Naphthols resemble phenols in their chemical properties, but their hydroxyl groups are more reactive. For example, they are readily converted to ethers by reaction with alcohols, and to amines by heating with ammonia and bisulfite (Bucherer reaction).

3.1. 1-Naphthol

1-Naphthol [*90-15-3*] (**15**), α-naphthol, 1-naphthalenol, 1-hydroxynaphthalene, $C_{10}H_8O$, M_r 144.16, forms colorless prisms (from toluene) which darken on exposure to air or light. The compound is steam volatile and sublimable. Other physical properties are listed in Table 3.

Table 3. Physical properties of naphthols

Property	1-Naphthol	2-Naphthol
mp, °C	96	123
bp, °C	280	295
d_4	1.224	1.217
Solubility in water, %	0.03 (25 °C)	0.075 (25 °C)
Dissociation constant (H_2O)	1.4×10^{-10} (20 °C)	1.14×10^{-10} (20 °C)

1-Naphthol is reduced by sodium in liquid ammonia to give 5,6,7,8-tetrahydro-1-naphthol and is oxidized by $NaOCl-FeCl_3$ or I_2-KI to give a violet color. It is chlorinated by phosphorus pentachloride at 150 °C to give 1-chloronaphthalene, by sulfuryl chloride to give 4-chloro-1-naphthol,

or by Cl_2-CH_3COOH to give 2,4-dichloro-1-naphthol. Similarly bromine forms 2,4-dibromo-1-naphthol, and this reaction may be used quantitatively for titration.

Nitration of 1-naphthol gives complex mixtures. The crude 2,4-dinitro derivative was used in the mid-1800s as an acid yellow dye comparable to picric acid. Sulfonation at 50 °C readily yields 1-naphthol-2,4-disulfonic acid. Nitrous acid gives mainly the 2-nitroso derivative, contaminated with the 4-nitroso isomer.

Production. Historically 1-naphthol was produced via caustic fusion of naphthalene-1-sulfonic acid, but this was superseded by the I.G. Farbenindustrie process involving hydrolysis of 1-naphthylamine (Section 5.1) with aqueous 22 % sulfuric acid at 200 °C under pressure in a lead-lined autoclave [7]. To obtain a purer product, Union Carbide developed a process based on catalytic oxidation of tetralin to 1-tetralol and 1-tetralone followed by dehydrogenation (see Fig. 2) [27]. The two-stage catalytic process is claimed to give an overall yield of 72 % 1-naphthol, with an overall efficiency of 97 %.

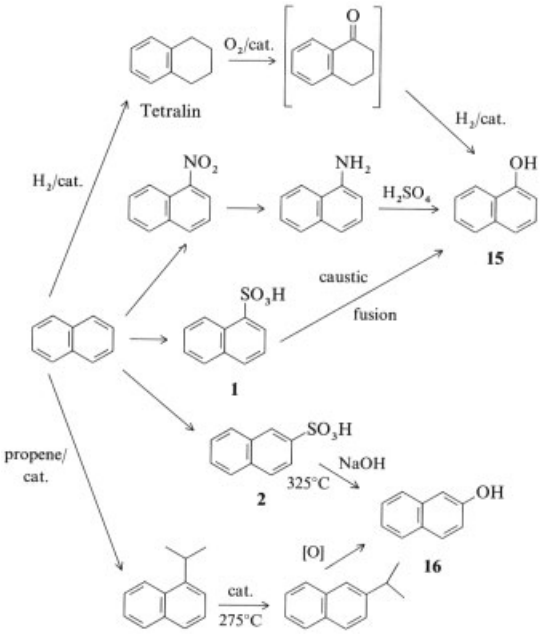

Figure 2. Industrial routes to 1-naphthol and 2-naphthol

Uses. Several important agrochemicals are derived from 1-naphthol. 1-Naphthoxyacetic acid (**17**, R = H) is prepared by aqueous alkylation with chloroacetic acid. Although 2-naphthoxyacetic acid (**31**) appears to be more widely used, ref. [28] suggests that the 1-isomers, especially (**17**, R = Me), have greater herbicidal activity. Devrinol [*15299-99-7*] (**18**), Stauffer's 2-(1-naphthoxy)-*N*,*N*-diethylpropionamide herbicide, is obtained by the reaction of 1-naphthol and *N,N*-diethyl-α-bromopropionamide in MeOH – NaOMe at reflux [29]. Sevin is Union Carbide's insecticide based on 1-naphthyl methylcarbamate [*63-25-2*], carbaryl (**19**), which is produced from 1-naphthyl chloroformate and methylamine [30] or by reaction of 1-naphthol with methyl isocyanate. Its main use lies in the control of earthworms and certain caterpillars. Many analogues and derivatives of **19** have been patented [11].

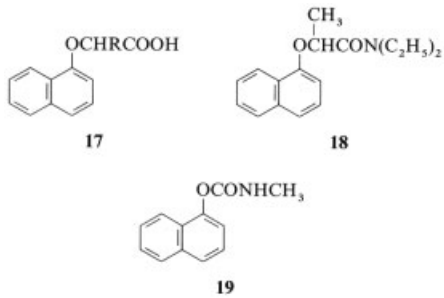

Important drugs are also derived from 1-naphthol. The ester 1-naphthyl salicylate [*550-97-0*], Alphol (**20**), has been used as an antiseptic and antirheumatic, and ICI's 1-isopropylamino-3-(1-naphthoxy)-2-propanol [*3506-09-0*], Inderal (**21**), is an important member of a new class of adrenergic β-blockers for the treatment of heart disease [31]. Production of **21** involves reaction of **15** with epichlorohydrin followed by isopropylamine.

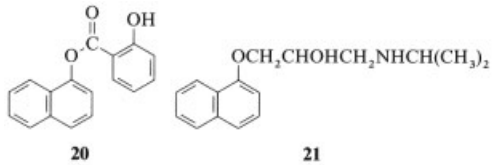

Related products used as β-blockers or antiarrhythmics include nadoxolol (**22**) and nadolol (**23**). Many analogues and alternative syntheses of this type of compound have been patented [11], [12].

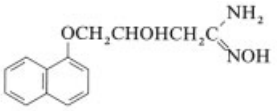

22

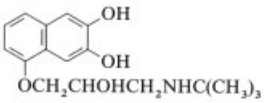

23

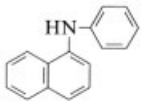

24

Condensation of 1-naphthol with aniline at 180 °C gives *N*-phenyl-1-naphthylamine (**24**) which is used as a rubber antioxidant.

1-Naphthol is much less important than 2-naphthol as a dye intermediate. Apart from the historical interest in simple azo derivatives such as Orange I (**25**) (GRIESS, 1876; C.I. Acid Orange 20), the equilibrium with the keto form (**26**) is an example of the prevalence of keto forms in many azo naphthol derivatives [32].

Certain diazo components, for example, 1-amino-2-hydroxynaphthalene - 4 - sulfonic acid, under strongly alkaline conditions, couple with 1-naphthol in the 2-position rather than the conventional 4-position. This leads to compounds such as C.I. Mordant Black 3 (**27**) which form stable chromium complexes. Excess diazo compound can lead to disazo acid dyes such as C.I. Acid Brown 43 (**28**).

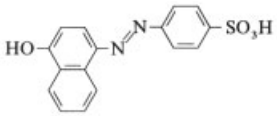

25

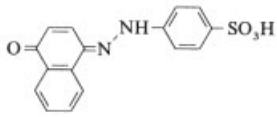

26

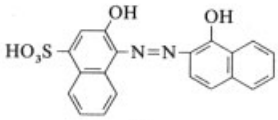

27

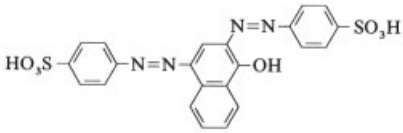

28

Nitroso compounds can condense with 1-naphthol for example, *N,N*-dimethyl-4-nitrosoaniline reacts to give indophenol (**29**).

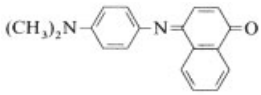

29

Halogenated derivatives of 1-naphthol have minor uses as shown by the reaction of 4-chloro-1-naphthol and 4-bromo-1-naphthol with substituted isatin-α-chlorides to form blue indigoid vat dyes [18]

Toxicity. 1-Naphthol is moderately toxic; the dust is liable to cause skin irritation and eye injury [14].

3.2. 2-Naphthol

2-Naphthol [*135-19-3*] (**16**), β-naphthol, 2-naphthalenol, 2-hydroxynaphthalene, $C_{10}H_8O$, M_r 144.16, forms colorless plates upon sublimation, which darken on exposure to air or light. Unlike 1-naphthol it is nonvolatile in steam. Other physical properties are presented in Table 3.

2-Naphthol is reduced with sodium or with hydrogen in the presence of a catalyst to give mainly 1,2,3,4-tetrahydro-2-naphthol (unlike 1-naphthol which gives the arylphenol under the same conditions). Oxidation with ferric chloride forms 2,2′-dihydroxy-1,1′-binaphthyl, with any blue coloration indicating the presence of 1-naphthol. Air oxidation at 300 °C with a vanadium pentoxide catalyst also yields 1,1′-bi-2-naphthol, which dehydrates at higher temperature to the oxide and finally decomposes to give benzoic acid. Treatment with sulfuryl chloride in carbon disulfide or with NaOCl – NaOH leads to 1-chloro-2-naphthol, whereas chlorine in sodium carbonate gives 8-chloro-2-naphthol (plus 1, 1′-bi-2-naphthol). Reaction with phosphorus pentachloride at 150 °C gives

2-chloronaphthalene or at lower temperature, tri-2-naphthyl phosphate (*mp* 111 °C), which can be isolated after treatment of the reaction mixture with alkali. Reaction with one equivalent of bromine in acetic acid gives 1-bromo-2-naphthol, whereas excess bromine yields 1,6-dibromo-2-naphthol. The latter is readily debrominated in dilute mineral acid to yield 6-bromo-2-naphthol. Excess bromine reacts with 2-naphthol at 100 °C to form 1,5,6-tribromo- and 1,3,5,6-tetrabromo-2-naphthol.

Careful nitration with strong nitric acid in acetic acid gives 1,6-dinitro-2-naphthol, whereas nitrous acid readily forms 1-nitroso-2-naphthol [*131-91-9*]. Sulfonation leads to a series of commercially important products (see Chap. 4).

Production. 2-Naphthol is produced by caustic fusion of naphthalene-2-sulfonic acid (Section 2.2). Typically, the sodium salt of the sulfonic acid is added gradually to 50 % sodium hydroxide liquor at 300 °C; the melt is then heated further at 320 °C in a gas-fired iron vessel with vigorous agitation. After completion of the reaction, the melt is run into excess water, possibly including filtrate from the previous batch at a proven tolerable level, and the naphtholate solution is neutralized to pH 8 with dilute sulfuric acid. If the temperature is maintained at >100 °C during neutralization, the crude product comes out of solution as an oil, which is separated, washed with hot water, and distilled under vacuum to give pure 2-naphthol. The molten material is processed through a flaker to give the final product for packaging. The fusion yield is about 80 % of the theoretical value, resulting in an overall yield of 70 % based on naphthalene. Typical specifications for 2-naphthol are (1) clear solution in dilute caustic soda, (2) *mp* 120.5 °C, and (3) 1-naphthol content <0.3 %.

The newer method of manufacture (see Fig. 2) is economically and environmentally favored in the United States, because despite requiring three stages, it is more amenable to continuous operation with recycle streams. The alkylation and isomerization are carried out up to 240 °C with a phosphoric acid catalyst [33]. Final catalytic oxidation at 90 – 110 °C gives the hydroperoxide, which is cleaved with dilute sulfuric acid to give 2-naphthol in high overall yield in spite of modest oxidation conversion.

Uses. The sulfonated and carboxylated derivatives are described in Chapter 4 and Section 3.4.

The simple ether derivatives, 2-methoxynaphthalene [*93-04-9*], nerolin, and 2-ethoxynaphthalene [*93-18-5*], nerolin "new", are readily prepared by conventional alcohol esterification processes for use in perfume formulation and, more recently as intermediates for drugs. The β-hydroxyethyl ether, anavenol (**30**), prepared by condensation of 2-naphthol with ethylene chlorohydrin or ethylene oxide, has been used as an animal sedative. Reaction of 2-naphthol with chloroacetic acid in aqueous alkaline medium gives 2-naphthoxyacetic acid (**31**), which is used as a growth promoter for fruit.

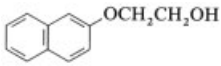

30

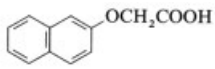

31

The lactate, benzoate, and salicylate esters have been used as antiseptics and are improvements of the naphthols themselves.

Reaction of 2-naphthol with aniline at 180 °C gives *N*-phenyl-2-naphthylamine (PBN, **32**) an antioxidant for rubber processing which is gradually being displaced by improved products.

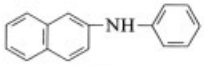

32

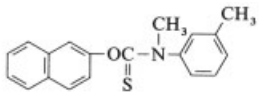

33

The major pharmaceutical products based on 2-naphthol are the antifungal tolnaftate (**33**), produced by reaction with thiophosgene and *N*-methyl-*m*-toluidine [34]; the semisynthetic penicillin nafcillin (**34**), produced via 2-ethoxynaphthalene; and the antirheumatic naproxen (**35**), produced via 2-methoxynaphthalene.

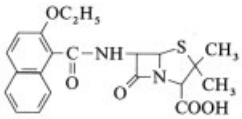

34

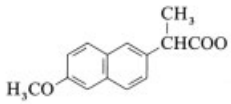

35

The C-alkylation of 2-naphthol by using 2-chlorocyclohexanone in boiling dioxane in the presence of potassium carbonate has been ex-

ploited in the preparation of the antitussive naphthanone (**36**).

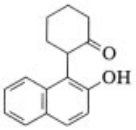

36

2-Naphthol is used in solvent dyes, e.g., C.I. Solvent Yellow 14 (**37**, R = H); in acid dyes, e.g., C.I. Acid Orange 7 (**37**, R = SO_3H); in metallizable dyes, e.g., C.I. Mordant Black 15 (**38**); and in pigments, e.g., C.I. Pigment Red 3 (**39**).

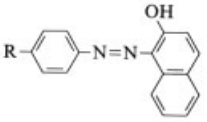

37

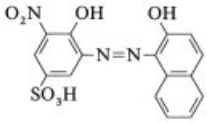

38

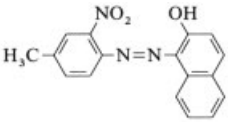

39

However, 3-hydroxy-2-naphthoic acid and derived arylamides (Section 3.4) are more important coupling components for pigments than is 2-naphthol.

Toxicity. 2-Naphthol is moderately toxic but less of a skin irritant than 1-naphthol [15].

3.3. Naphthalenediols

All ten isomeric naphthalenediols are known. They are usually prepared by alkali fusion of sulfonic acids or by acid hydrolysis of aminonaphthalenes, provided appropriately substituted compounds are available as precursors. With α,β-disubstituted naphthalenes, hydroxylation takes place more readily in the α-position; this is particularly important if hydroxynaphthalenesulfonic acids (Chap. 4) are required [12].

Conversely, with amination of dihydroxynaphthalenes by the Bucherer reaction, the β-position is more readily attacked. Thus, 1,7-dihydroxynaphthalene gives 7-amino-1-naphthol.

Of the ten dihydroxynaphthalenes, only 1,5-naphthalenediol has significant applications.

1,5-Naphthalenediol [*83-56-7*], 1,5-dihydroxynaphthalene, Azurol, $C_{10}H_8O_2$, M_r 160.16, forms colorless needles (from aqueous methyl cellosolve). The compound is sublimable, readily soluble in ether and acetone, but only sparingly soluble in water and hydrocarbons. It is oxidized to 5-hydroxy-1,4-naphthoquinone (juglone) by CrO_3. Amination gives 1,5-naphthalenediamine; coupling with diazotized anilines occurs in the 2- or 4-position.

Production [7]. The sodium salt of naphthalene-1,5-disulfonic acid (Section 2.3) is added gradually to excess 75 % sodium hydroxide. The mixture is heated under pressure to 276 °C over the course of 7 h and the reaction is completed by heating for a further 5 h at this temperature. After precipitation with sulfuric acid, the product is isolated in 90 % yield.

Uses. 1,5-Naphthalenediol is used as a coupling component in azo dyes as exemplified in the metallizable (Cr) wool dye, C.I. Mordant Black 9 (**40**).

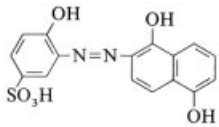

40

1,2-Naphthalenediol [*574-00-5*], *mp* 108 °C, is prepared by reduction of 1,2-naphthoquinone with hydrazine hydrate or sodium dithionite. It is oxidized under mild conditions to 1,2-naphthoquinone and couples with diazotized anilines in the 4-position.

1,3-Naphthalenediol [*132-86-5*], naphthoresorcinol, *mp* 124 °C, is synthesized by hydrolysis of 1,3-naphthalenediamine with sulfuric acid; by heating 1-amino-3-hydroxy-naphthalene-4-sulfonic acid with acid; or by cyclization of ethyl phenylacetoacetate. It is used as an analytical reagent for sugars and glucuronic acid in urine. It is oxidized in alkaline solution to 2-hydroxy-1,4-naphthoquinone and reacts with ammonia at 140 °C to give 3-amino-1-naphthol and 1,3-naphthalenediamine.

1,4-Naphthalenediol [*571-60-8*], *mp* 195 °C, is synthesized by reduction of 1,4-naphthoquinone with sodium dithionite, or by hydrolysis of 4-amino-1-naphthol with 20 % sulfuric acid. It undergoes oxidation to give 1,4-naphthoquinone; amination to give 4-amino-1-naphthol; and methylation (MeOH – HCl) to give 4-methoxy-1-naphthol.

1,6-Naphthalenediol [*575-44-0*], *mp* 138 °C, is synthesized by caustic fusion of naphthalene-1,6-disulfonic acid at 330 °C. It couples with diazotized anilines to form disazo and trisazo derivatives.

1,7-Naphthalenediol [*575-38-2*], *mp* 181 °C, is prepared by caustic fusion of 2-hydroxynaphthalene-8-sulfonic acid at 300 °C or by acid desulfonation of 2,8-dihydroxynaphthalene-6-sulfonic acid. It couples with diazotized anilines to form monoazo and disazo derivatives.

1,8-Naphthalenediol [*569-42-6*], *mp* 144 °C, is produced by caustic fusion of naphthosultone (Section 4.2) or naphthosultam (Section 6.1), and by acid desulfonation of 1,8-dihydroxynaphthalene-4-sulfonic acid. It is oxidized to juglone by chromium(VI) oxide; gives a dark green color with iron(III) chloride; undergoes amination at 300 °C to give 1,8-naphthalenediamine; and couples with diazotized anilines in the 4-position.

2,3-Naphthalenediol [*92-44-4*], *mp* 161 °C, is prepared by caustic fusion of 3-hydroxynaphthalene-2,6-disulfonic acid followed by treatment with dilute sulfuric acid under pressure, or by acid desulfonation of 2,3-dihydroxynaphthalene-6-sulfonic acid. It gives a dark blue color with iron(III) chloride; is aminated to 2-amino-3-naphthol at 140 °C and 2,3-naphthalenediamine at 240 °C; and couples with diazotized anilines in the 1- or 1,4-positions.

2,6-Naphthalenediol [*581-43-1*], 2,6-naphthohydroquinone, *mp* 222 °C, is synthesized by potassium hydroxide fusion of 2-hydroxynaphthalene-6-sulfonic acid at 295 °C. It undergoes amination at 250 °C to give 2,6-naphthalenediamine and couples with diazotized anilines in the 1-position under acid conditions or the 1,5-positions under alkaline conditions.

2,7-Naphthalenediol [*582-17-2*], *mp* 194 °C, is produced by caustic fusion of naphthalene-2,7-disulfonic acid or 2-hydroxynaphthalene-7-sulfonic acid at 300 °C. It undergoes amination at 250 °C to give 2,7-naphthalenediamine and couples with diazotized anilines in the 1-position or in the 1,8-positions. It forms a resin with formaldehyde.

3.4. Hydroxynaphthoic Acids

The industrially important hydroxynaphthoic acids are all produced by carboxylation (with CO_2) of the corresponding sodium or potassium naphtholate (Kolbe – Schmitt reaction). 3-Hydroxy-2-naphthoic acid (BON acid) is one of the most important naphthalene-based colorant intermediates. Arylamide derivatives were introduced in 1912 as improved azoic dyes, replacing 2-naphthol which had been in use since 1880 as a component in ice colors. Apart from their use in azoic dyes for cotton, these arylamides are intermediates for important pigments.

For isomers that are not accessible by carboxylation, the precursor cyano group is introduced by Sandmeyer reaction of the appropriate aminonaphthalenesulfonic acid, followed by hydrolysis and caustic fusion.

Although carboxylation of naphtholates may be carried out in a solvent whereby the polarity of the solvent has a marked effect on the reaction [35], large-scale operation is usually effected by using a dry melt in a rotating baker or specially designed agitated vessel. Whereas sodium α-naphtholate yields 1-hydroxy-2-naphthoic acid at 130 °C and sodium β-naphtholate yields 2-hydroxy-1-naphthoic acid at 120 °C; somewhat surprisingly, 3-hydroxy-2-naphthoic acid is produced from β-naphthol above 200 °C. The mechanism is not clear, and the course of the reaction is affected by temperature, pressure, alkali metal, and the presence of water [36]. However, the identified intermediate stages from sodium β-naphtholate are the formation of **41** at 120 °C and its conversion to equimolar quantities of **42** and 2-naphthol at 145 °C, which explains why the maximum yield is only 50 %. At 200 °C, **42** rearranges to **43**, which also has a maximum yield of 50 %.

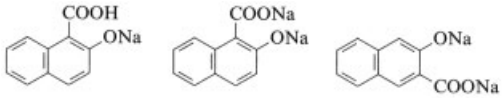

41 42 43

The Marasse modification (1893) of the Kolbe – Schmitt reaction uses potassium carbonate with carbon dioxide and in certain cases gives better yields [36]. This has been confirmed for hydroxynaphthoic acids at 300 °C [37]. However, with potassium β-naphtholate an alternative rearrangement can take place, in which 6-hydroxy-2-naphthoic acid is coproduced at 220 °C.

2-Hydroxy-1-naphthoic acid [*2283-08-1*], $C_{11}H_8O_3$, M_r 188.17, has a *mp* of 156 – 157 °C (decomp. to 2-naphthol and CO_2). The monosodium salt is unstable in hot aqueous solution. 2-Hydroxy-1-naphthoic acid is decarboxylated by nitrous acid, forming 1-nitroso-2-naphthol, and by diazo salts, forming a 1-azo-2-naphthol. An ethanolic solution of 2-hydroxy-1-naphthoic acid gives a dark blue color with iron(III) chloride. Chlorosulfonic acid at 30 °C gives the 6-sulfonyl chloride derivative.

Production. [7] Treatment of the dried sodium or potassium salt of 2-naphthol with carbon dioxide at 120 °C under pressure gives an 80 % yield of product after careful work-up. Solvents such as toluene, tetralin, or diphenyl oxide are patented alternatives [15].

Uses. 2-Hydroxy-1-naphthoic acid is an intermediate of minor importance in the synthesis of azo dyes.

3-Hydroxy-1-naphthoic acid [*19700-42-6*], *mp* 248 – 249 °C, is prepared by hydrolysis of diazotized 3-amino-1-naphthoic acid with boiling aqueous copper sulfate. It gives a reddish brown color with iron(III) chloride. Coupling with diazo compounds occurs in the 4-position.

4-Hydroxy-1-naphthoic acid [*7474-97-7*], *mp* 188 – 189 °C, is synthesized by potassium hydroxide fusion (200 °C) of 4-sulfo-1-naphthoic acid or 4-cyanonaphthalene-1-sulfonic acid (from 1-aminonaphthalene-4-sulfonic acid). It undergoes decarboxylation with bromine, nitrous acid, or diazo salts to form the corresponding 4-derivative.

5-Hydroxy-1-naphthoic acid [*2437-16-3*], *mp* 236 °C, is prepared by potassium hydroxide fusion (245 °C) of 1-cyanonaphthalene-5-sulfonic acid. It couples with diazo compounds in the 8-position.

6-Hydroxy-1-naphthoic acid [*2437-17-4*], *mp* 213 °C, is produced by hydrolysis (10 % KOH) and potassium hydroxide fusion (260 °C) of 1-cyanonaphthalene-6-sulfonic acid (from 1-aminonaphthalene-6-sulfonic acid and cyano-Sandmeyer reaction).

7-Hydroxy-1-naphthoic acid [*2623-37-2*], *mp* 256 °C, is synthesized by hydrolysis (50 % KOH) of 8-cyano-2-naphthol (from 8-amino-2-naphthol and cyano-Sandmeyer reaction). It couples in the 8-position with diazo compounds.

8-Hydroxy-1-naphthoic acid [*1769-88-6*], *mp* 169 °C, is prepared by hydrolysis (60 % H_2SO_4) of 8-cyanonaphthalene-1-sulfonic acid and caustic fusion of the resulting anhydride of 8-sulfo-1-naphthoic acid. It couples with diazo compounds in the 5-position.

1-Hydroxy-2-naphthoic acid [*86-48-6*], $C_{11}H_8O_3$, M_r 188.17, *mp* 200 °C, is only sparingly soluble in hot water but readily soluble in alkali and ethanol. It couples with diazo compounds in the 4-position and, with excess, in the 4,7-positions. It is sulfonated by oleum to give the 4-sulfonic and then the 4,7-disulfonic acids. Halogenation occurs in the 4-position, and nitrosation gives 2-nitroso-1-naphthol.

Production. [7] Treatment of the dried sodium salt of 1-naphthol with carbon dioxide at 135 °C under pressure gives an 80 % yield of product.

Uses. 1-Hydroxy-2-naphthoic acid is used as an intermediate for azo and triphenylmethane dyes and, more recently, for color-film dyes [38].

3-Hydroxy-2-naphthoic acid [*92-70-6*], BON acid, $C_{11}H_8O_3$, M_r 188.17, *mp* 222 °C, forms pale yellow crystals from aqueous ethanol. Alkaline solutions have a green fluorescence. Aqueous ferric chloride produces a blue color, and warming with $CHCl_3$– NaOH results

in a green color. Oxidation with alkaline permanganate gives 2-carboxyphenylglyoxylic acid and then phthalic acid. Dry distillation produces dibenzoxanthone. Bromination gives the 4,7-dibromo derivative; nitration, the 4-nitro derivative; and sulfonation with oleum, the 5,7-disulfonic acid derivative.

Production is invariably by the Kolbe – Schmitt reaction of carbon dioxide with the dried sodium salt of 2-naphthol at 220 – 260 °C [7]. Many improvements in both process and equipment have been developed. The main problems are removal of the 50 % unreacted 2-naphthol and avoidance of a tarry byproduct. Yields of 80 % based on the amount of 2-naphthol consumed are possible despite the low conversion.

Uses. The major use of 3-hydroxy-2-naphthoic acid as a dye intermediate is reflected by the quoted production figure of 21 000 t/a in Japan alone. With annual imports of 2500 t/a in the United States, BON acid is the largest imported naphthalene intermediate and one of the largest of all dye and pigment intermediates. Naphtol AS (Naphtol Anilid-Säure) (**44**) represents the most important arylamide derivative. It is produced by forming the acid chloride of 3-hydroxy-2-naphthoic acid with phosphorus trichloride in a solvent such as toluene and then adding aniline for condensation in situ. Many other arylamides are manufactured similarly by using a range of substituted anilines and other arylamines. These arylamides are used in azoic dyes or as intermediates for azo pigments.

Sulfonation of 3-hydroxy-2-naphthoic acid with 98 % sulfuric acid at 30 °C gives a mixture of the 5- and 7-sulfonic acid derivatives, which can be separated via their potassium salts. After isolation, the more soluble 5-isomer can be fused with NaOH – KOH at 260 °C to give 3,5-dihydroxy-2-naphthoic acid (**45**).

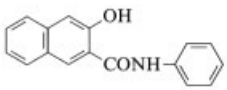

44

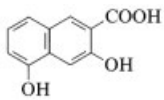

45

4-Hydroxy-2-naphthoic acid [*1573-91-7*], *mp* 225 – 226 °C, is produced by air oxidation of 1-tetralone-3-carboxylic acid.

5-Hydroxy-2-naphthoic acid [*2437-18-5*], *mp* 215 – 216 °C, is prepared by potassium hydroxide fusion (260 °C) of 5-sulfo-2-naphthoic acid.

6-Hydroxy-2-naphthoic acid [*16712-64-4*], *mp* 245 – 248 °C, is synthesized by carboxylation of the potassium salt of 2-naphthol at 170 – 230 °C. It couples with diazo compounds in the 5-position and undergoes sulfonation in the 4- and 7-positions.

7-Hydroxy-2-naphthoic acid [*613-17-2*], *mp* 274 °C, is prepared by sulfonation (160 °C) of 2-naphthoic acid to give the 7-sulfo derivative, followed by potassium hydroxide fusion at 260 °C. It couples with diazo compounds in the 8-position.

8-Hydroxy-2-naphthoic acid [*5776-28-3*], *mp* 229 °C, is produced by the cyano-Sandmeyer reaction of 2-aminonaphthalene-8-sulfonic acid, followed by hydrolysis and potassium hydroxide fusion.

Di- and trihydroxynaphthoic acids and hydroxy- and dihydroxynaphthalenedicarboxylic acids are also obtainable, but are of only academic interest at present [5].

4. Hydroxynaphthalenesulfonic Acids

The importance of the various mono- and dihydroxynaphthalene mono- and disulfonic acids is demonstrated by the fact that about 25 products in this group have been produced over the years with many of them still available and referred to by code names (see Table 1). Most of the products have been used as azo dye coupling components, but they are equally important as intermediates for aminonaphthalenesulfonic acids (Chap. 6) and aminohydroxynaphthalenesulfonic acids (Chap. 8). Only those isomers that have a history of commercial use are dealt with on an individual basis. Many of the other isomers are also obtainable [5] and may be of present or future interest.

Because in usage this group of products is considered to be derived from 1-naphthol or 2-naphthol they are dealt with on this basis rather

than as hydroxy derivatives of naphthalenesulfonic acids. The latter alternative is the IUPAC convention, and these more formal names are given in parentheses alongside alternative trivial names.

4.1. Production and Properties

Production depends on four main routes to achieve the required substitution patterns. All routes involve selectivity and separation at some stage, usually the sulfonation stage (cf. Section 2.1). The use of HPLC is essential for following separations and assessing the purity of products.

Direct Sulfonation of Naphthols. The reaction of 1-naphthol with sulfuric acid at 60 °C gives mainly 1-hydroxynaphthalene-2-sulfonic acid (**46**) with a small proportion of 1-hydroxynaphthalene-4-sulfonic acid (**47**), from which it can be separated. Increasing the sulfuric acid – naphthol ratio and raising the temperature produces first 1-hydroxynaphthalene-2,4-disulfonic acid (**48**) and then 1-hydroxynaphthalene-2,4,7-trisulfonic acid (**49**).

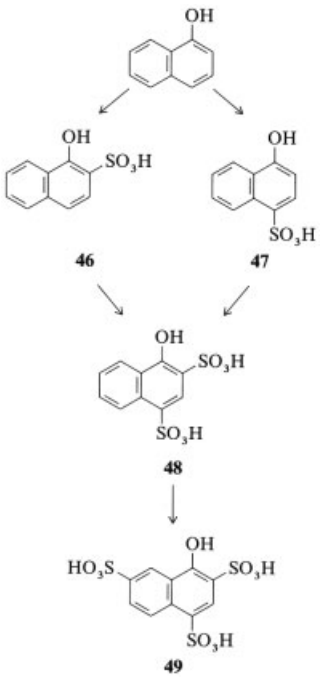

Sulfonation of 2-naphthol is more complex because concentration and amount of sulfonating agent, temperature, and reaction time all significantly influence the nature and yield of product. At 0 – 20 °C, sulfuric acid rapidly forms 2-hydroxynaphthalene-1-sulfonic acid (**50**); with longer reaction times a second sulfonic acid group is introduced to form 2-hydroxynaphthalene-1,6-disulfonic acid (**51**). At 50 – 55 °C, sulfonation occurs in the 6- and 8-positions giving **52** and **53**. More vigorous conditions lead to the disulfonic acids **54** and **55** with some 1- and, to a lesser extent, 8-desulfonation being involved to affect isomer yields. Finally, oleum at higher temperature sulfonates the disulfonic acids to give 2-hydroxynaphthalene-3,6,8-trisulfonic acid (**56**).

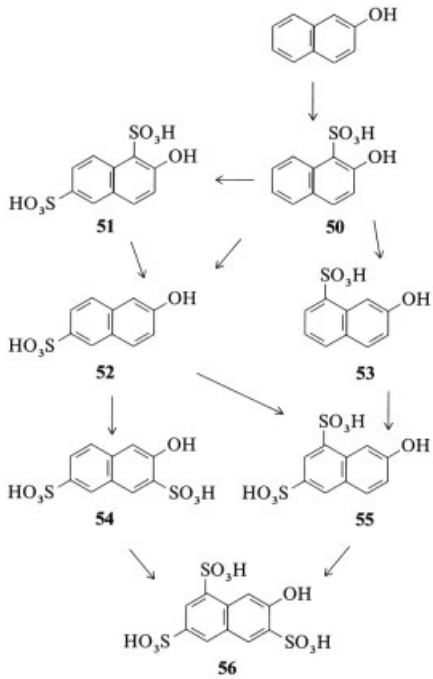

Alkali Fusion of Naphthalenedisulfonic Acids or Hydroxynaphthalenedi- and Trisulfonic Acids (cf. Chap. 3). Hydrolysis of sulfonic acid substituents to form substituted naphthols is a key process with selective monoreaction able to be controlled in disulfonic acids. This is easier with α,β-disulfonic acids, in which reaction at the α-group is much more facile.

The following ten derivatives (**57** – **66**) prepared by this route are all commercially important. The asterisks indicate hydroxyl groups introduced by the fusion process.

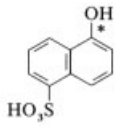

57

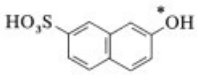

58

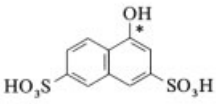

59

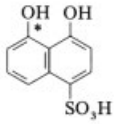

60

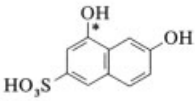

61

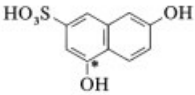

62

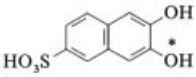

63

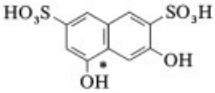

64

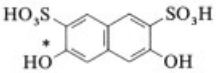

65

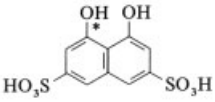

66

Hydrolysis of Aminonaphthalenesulfonic Acids. Certain naphthylaminesulfonic acids resemble 1-naphthylamine in hydrolyzing to the corresponding naphthol on prolonged heating with dilute sulfuric acid at 180 – 200 °C under pressure. This process is preferred for the synthesis of 1-hydroxynaphthalene-3-sulfonic acid (**67**), 2-hydroxynaphthalene-5,7-disulfonic acid (**68**), and 1-hydroxynaphthalene-3,6,8-trisulfonic acid (**69**).

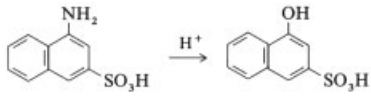

67

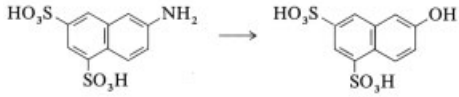

68

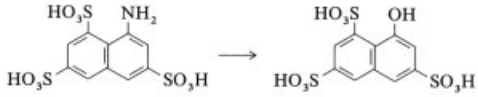

69

An alternative to sulfuric acid hydrolysis is the use of sodium bisulfite in a so-called reverse-Bucherer reaction. This is the preferred process for 1-hydroxynaphthalene-4-sulfonic acid (**47**).

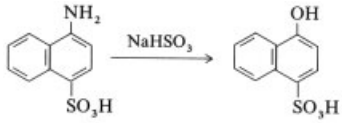

47

Finally, where neither sulfuric acid nor bisulfite hydrolysis is applicable, indirect hydrolysis via diazotization of the aminonaphthalenesulfonic acid and treatment with dilute sulfuric acid are employed. Although this method is generally applicable, it is more expensive to operate on a large scale, and its use is confined to derivatives of 1-hydroxynaphthalene-8-sulfonic acid.

Acid Desulfonation of Hydroxynaphthalenedi- or Trisulfonic Acids. With certain substitution patterns, sulfonic acid groups may be removed by heating with 40 % sulfuric acid. This reaction should be regarded as reverse sulfonation rather than hydrolysis, and includes the examples already given in Section 3.3 for the formation of naphthalenediols. An example is the selective desulfonation of 2-hydroxynaphthalene-4,8-disulfonic acid to give 2-hydroxynaphthalene-4-sulfonic acid (**70**).

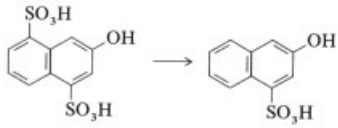

70

Although hydroxynaphthalenesulfonic acids are usually soluble in water and more soluble than their alkali-metal salts, they can be isolated by salting out from solution. The hydroxyl group is dissociated in aqueous solution as shown by the dissociation constants (Table 4) [15].

Table 4. Dissociation constants for hydroxyl groups of hydroxynaphthalenesulfonic acids in water at 20 °C

Hydroxynaphthalenesulfonic acid	pK_2	pK_3	pK_4
1-Hydroxynaphthalene-3-sulfonic acid	8.7		
1-Hydroxynaphthalene-4-sulfonic acid	8.2		
1-Hydroxynaphthalene-5-sulfonic acid	9.0		
2-Hydroxynaphthalene-1-sulfonic acid	10.96		
2-Hydroxynaphthalene-6-sulfonic acid	9.1		
2-Hydroxynaphthalene-7-sulfonic acid	9.2		
1,8-Hydroxynaphthalene-4-sulfonic acid	5.33	>12.7	
2,3-Hydroxynaphthalene-6-sulfonic acid	8.24	12.13	
2,8-Hydroxynaphthalene-6-sulfonic acid	8.61	10.21	
1,8-Hydroxynaphthalene-3,6-sulfonic acid		5.53	>12.75

Purification and characterization are difficult unless chromatographic standards have been obtained [20]. The simplest form of analysis is to determine strength by coupling via titration of an alkaline solution with a standard solution of a diazonium salt.

Many of the hydroxynaphthalenesulfonic acids are used as azo coupling components. Where they serve as precursors in the synthesis of downstream intermediates the most common reactions are amination of the hydroxyl group by the Bucherer reaction to form aminonaphthalenesulfonic acids (Chap. 6) and caustic fusion of the sulfonic acid group to form naphthalenediols (Section 3.3) or their sulfonic acid derivatives.

4.2. 1-Hydroxynaphthalenesulfonic Acids

1-Hydroxynaphthalene-2-sulfonic acid [*567-18-0*] (**46**), Baum's acid, $C_{10}H_8O_4S$, M_r 224.23, is sparingly soluble in water and is precipitated from aqueous solution by hydrochloric acid. Coupling with diazo compounds and nitrosation both take place in the 4-position. Sulfonation gives 1-naphthol-2,4-disulfonic acid and nitration gives 2,4-dinitro-1-naphthol.

Production. The reaction of 1-naphthol with oleum in the cold yields a mixture of the 2- and 4-sulfonic acids; the former may be separated from the aqueous quenched reaction mixture by salting out the potassium salt after removing unreacted 1-naphthol.

1-Hydroxynaphthalene-3-sulfonic acid [*3771-14-0*] (**67**) (4-hydroxynaphthalene-2-sulfonic acid), $C_{10}H_8O_4S$, M_r 224.2, and its salts are very soluble in water. Coupling with diazo compounds occurs in the 2-position with weak diazo compounds, but in the 4-position with strong diazo compounds such as diazotized nitroanilines. Heating with ammonia at 170 °C gives 1,3-naphthalenediamine.

Production. 1-Aminonaphthalene-3-sulfonic acid (Section 6.1) is heated with dilute sulfuric acid by using high-pressure steam to maintain a pressure of 1 MPa for 20 h [7]. After cooling, screening, and adding salt to the hot liquor, the sodium salt crystallizes on cooling and is isolated in 80 % yield.

Uses. 1-Hydroxynaphthalene-3-sulfonic acid is used as a coupling component for the production of direct dyes.

1-Hydroxynaphthalene-4-sulfonic acid [*84-87-7*] (**47**) (4-hydroxynaphthalene-1-sulfonic acid), Nevile and Winther's acid, NW acid, $C_{10}H_8O_4S$, M_r 224.23, and its salts are water soluble, but the sodium salt may be precipitated from solution by salting out. Coupling with diazo compounds takes place in the 2-position. Nitration yields 2,4-dinitro-1-naphthol, whereas nitrosation gives 2-nitroso-1-hydroxynaphthalene-4-sulfonic acid. Sulfonation leads to a mixture of 1-hydroxynaphthalene-2,4-di- and 2,4,7-trisulfonic acids.

Production. Sodium naphthionate (Section 6.1) in aqueous alkaline solution is heated with sodium bisulfite and liquid sulfur dioxide for 24 h at 95 °C [7]. The batch is heated with excess lime to hydrolyze the bisulfite compound and boil off ammonia. After filtration the solution of the calcium salt of the product is neutralized with sodium carbonate, and precipitated calcium carbonate is filtered off. The solution of the sodium salt can be used directly as a coupling component, or salt can be added to precipitate and isolate the product.

An alternative synthesis involves selective sulfonation of 1-naphthol in the 4-position by using liquid sulfur trioxide in solvents such as trichloroethane or tetrachloroethane.

Uses. 1-Hydroxynaphthalene-4-sulfonic acid is used as a coupling component for a wide range of acid azo dyes, e.g., C.I. Acid Orange 19, C.I. Mordant Brown 35, and C.I. Food Red 3 (**71**).

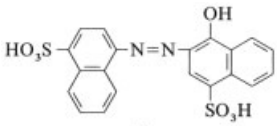

71

1-Hydroxynaphthalene-5-sulfonic acid [*117-59-9*] (**57**) (5-hydroxynaphthalene-1-sulfonic acid), oxy L acid, Azurin acid, $C_{10}H_8O_4S$, M_r 224.2, and its salts are very soluble in water, but the sodium salt may be precipitated from solution by salting out. Coupling with diazo compounds normally occurs in the 2-position but takes place in the 4-position with stronger diazo compounds such as diazotized nitroanilines. Sulfonation yields 1-hydroxynaphthalene-2,5-disulfonic acid; vigorous alkali fusion gives 1,5-dihydroxynaphthalene. Bromination in acetic acid yields 2,3-dibromo-1,4-naphthoquinone-5-sulfonic acid.

Production. The disodium salt of naphthalene-1,5-disulfonic acid (Section 2.3) is heated in an autoclave with 20 % sodium hydroxide at 230 °C for 24 h to selectively hydrolyze one sulfonic acid group [7]. Dilution, salting out, and acidification give a technical product in 90 % yield. A purer product, in lower yield, is obtained by acidification of the diluted melt with sulfuric acid, followed by liming and salting out.

Uses. 1-Hydroxynaphthalene-5-sulfonic acid is used as a coupling component for a range of azo colorants, for example, C.I. Pigment Red 54 and C.I. Mordant Black 29.

1-Hydroxynaphthalene-7-sulfonic acid [*20191-62-2*] (**72**) (8-hydroxynaphthalene-2-sulfonic acid), $C_{10}H_8O_4S$, M_r 224.2, is very soluble in water, and most of its metal salts are also readily soluble. It couples with diazotized 4-nitroaniline to give the 2,4-disazo compound. Nitration gives 2,4-dinitro-1-naphthol-7-sulfonic acid, which was known as Flavianic acid when used as a yellow dye for wool.

Production. 1-Naphthol is sulfonated with concentrated sulfuric acid at 130 °C to give a mixture of 2,7- and 4,7-disulfonic acids. This mixture is difficult to separate, but dilution with water and heating the resulting dilute sulfuric acid solution at 120 °C for 30 min result in selective desulfonation in the 2- and 4-positions [39].

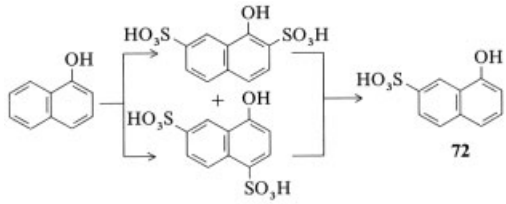

72

1-Hydroxynaphthalene-8-sulfonic acid [*117-22-6*] (**74**) (8-hydroxynaphthalene-1-sulfonic acid), $C_{10}H_8O_4S$, M_r 224.2, crystallizes as the monohydrate with *mp* 106 °C. Sulfonation gives 1-hydroxynaphthalene-4,8-disulfonic acid, and alkali fusion with potassium hydroxide at 230 °C yields 1,8-dihydroxynaphthalene. Coupling with diazo compounds occurs in the 2-position.

Production. 1-Aminonaphthalene-8-sulfonic acid (Section 6.1) is diazotized over 16 h at 50 °C and then heated to 80 °C [7]. After cooling, the precipitated naphthosultone (**73**) is filtered off and washed. The sultone is heated with dilute sodium hydroxide at 100 °C, which causes ring opening; the product is isolated after salting out.

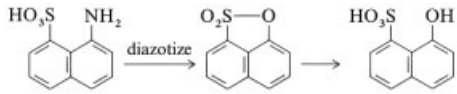

73 **74**

Uses. The naphthosultone (**73**) is usually satisfactory as an intermediate for further processing; however, for azo dye preparation the free naphthol (**74**) must be used, for example, in C.I. Acid Black 54 and C.I. Acid Blue 158 (the chromium complex of **75**).

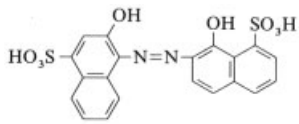

75

4.3. 2-Hydroxynaphthalenesulfonic Acids

2-Hydroxynaphthalene-1-sulfonic acid [*567-47-5*] (**50**), oxy Tobias acid, $C_{10}H_8O_4S$, M_r 224.2, and its salts are very soluble in water; neutral solutions give a strong blue color with iron(III) chloride. The sulfonic acid group is sufficiently labile for desulfonation to take place slowly in aqueous solution and rapidly with diazo compounds to form 1-azo-2-hydroxynaphthalene derivatives (i.e., the same product as from 2-naphthol), and with bromine to form 1-bromo-2-hydroxynaphthalene. Low-temperature sulfonation gives 2-hydroxynaphthalene-1,6-disulfonic acid, whereas higher temperatures result in 1-desulfonation and formation of 2-hydroxynaphthalene-6- and 8-sulfonic acids. Nitration affords 1,6-dinitro-2-hydroxynaphthalene.

Production. 2-Naphthol is dissolved in 1,2-dichloroethane and chlorosulfonic acid is added at 0 °C; the reaction is allowed to go to completion at low temperature. During work-up and isolation, desulfonation is avoided by careful extraction of the product into aqueous alkali and salting out of the sodium salt. Alternative solvents, such as dichloromethane, tetrachloroethane, or 1,2-dichlorobenzene, and other sulfonating agents, (e.g., liquid sulfur trioxide) can be used.

Uses. Virtually all the 2-hydroxynaphthalene-1-sulfonic acid produced is converted to 2-aminonaphthalene-1-sulfonic acid (Tobias acid) by amination (Section 6.1).

The calcium salt Asaprol is soluble in aqueous ethanol and has been used in place of gypsum in surgical plaster.

2-Hydroxynaphthalene-4-sulfonic acid [*6357-85-3*] (**70**) (3-hydroxynaphthalene-1-sulfonic acid), $C_{10}H_8O_4S$, M_r 224.2, and its salts are extremely soluble in water and difficult to isolate. It undergoes nitrosation and couples with diazo compounds in the 1-position. Sulfonation with oleum gives 2-hydroxynaphthalene-4,8-disulfonic acid. Amination with ammonia under pressure gives 2-aminonaphthalene-4-sulfonic acid.

Production. 2-Hydroxynaphthalene-4,8-disulfonic acid (Section 4.5) is heated with dilute sulfuric acid at 190 °C for 10 h in an autoclave under pressure [7]. After cooling, a small quantity of 2-naphthol is filtered off and sulfuric acid is removed by liming. The resulting solution of the calcium salt of the product is converted to the sodium salt, which is used directly as a solution or evaporated to isolate the product.

2-Hydroxynaphthalene-6-sulfonic acid [*93-01-6*] (**52**) (6-hydroxynaphthalene-2-sulfonic acid), Schaeffer acid, $C_{10}H_8O_4S$, M_r 224.23, is readily soluble in water, from which it can be recrystallized as the monohydrate (*mp* 129 °C). The sodium salt is sparingly soluble in cold water (1.7 %) but forms a 30 % solution at 80 °C. Coupling with diazo compounds results in substitution in the 1-position as does treatment with bromine, nitrous acid, or oleum. Fusion with potassium hydroxide gives 2,6-dihydroxynaphthalene (Section 3.3), amination gives 2-aminonaphthalene-6-sulfonic acid (Section 6.1), and sulfonation (20 % oleum, 25 °C) results in 2-hydroxynaphthalene-1,6-disulfonic acid.

Production. 2-Naphthol is added to a mixture of sulfuric acid and sodium sulfate, and the reaction is continued at 90 °C for 12 h [7]. After pouring the reaction mixture into water, the sodium salt of the product is precipitated by addition of salt and filtered at 60 °C to separate the isomers.

Uses. 2-Hydroxynaphthalene-6-sulfonic acid is used as a coupling component for a wide range of azo dyes, e.g., C.I. Acid Orange 12, C.I. Food Yellow 3, and C.I. Food Orange 2 (**76**). It is also used as an intermediate for more highly substituted dye components and synthetic tanning agents. The 1-nitroso derivative forms an iron complex (C.I. Acid Green 1) which was formerly used as a wool dye.

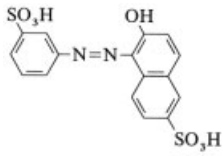

76

2-Hydroxynaphthalene-7-sulfonic acid [*92-40-0*] (**58**) (7-hydroxynaphthalene-2-sulfonic acid), F acid, Cassella acid, $C_{10}H_8O_4S$, M_r 224.23, is soluble in water, from which it

can be recrystallized as the monohydrate (*mp* 89 °C). The sodium and potassium salts are also readily soluble, but the barium salt is sparingly soluble. Iron(III) chloride gives a dark blue color with neutral solutions of **58**. Azo coupling and nitrosation occurs in the 1-position. Sulfonation gives a variety of products depending on the conditions.

Production. Naphthalene-2,7-disulfonic acid (Section 2.3) is heated with aqueous sodium hydroxide under pressure at 230 °C for 5 h and then at 260 °C for another 5 h. After neutralizing and screening, the liquor is acidified to precipitate the product, which is isolated after boiling off the sulfur dioxide and cooling to 30 °C.

Uses. 2-Hydroxynaphthalene-7-sulfonic acid is used as an intermediate in the production of 2-aminonaphthalene-7-sulfonic acid (Section 6.1), 2-hydroxynaphthalene-3,7-disulfonic acid, and 2,7-dihydroxynaphthalene (Section 3.3). It has minor uses as an azo coupling component.

2-Hydroxynaphthalene-8-sulfonic acid [*132-57-0*] (**53**) (7-hydroxynaphthalene-1-sulfonic acid), crocein acid, Bayer's acid, $C_{10}H_8O_4S$, M_r 224.23, is more soluble in water than 2-hydroxynaphthalene-6-sulfonic acid and cannot readily be isolated. Its salts are also more soluble than their 2,6-analogues and are difficult to isolate. On evaporation of its solutions, the free acid (**53**) decomposes to 2-naphthol. Coupling with diazo compounds takes place in the 1-position, but reaction is hindered by the peri-sulfonic acid group so that concentrated conditions and sometimes catalysts are required. Above 60 °C, sulfuric acid converts **53** into 2-hydroxynaphthalene-6-sulfonic acid; sulfonation then proceeds further to give the 6,8-disulfonic acid. Amination (Bucherer reaction) gives 2-aminonaphthalene-8-sulfonic acid (Section 6.1).

Production. 2-Naphthol reacts with excess 98 % sulfuric acid at 20 – 35 °C to form 2-hydroxynaphthalene-1-sulfonic acid. Further reaction at 55 – 60 °C produces a mixture of 6- and 8-sulfonic acids, from which the less soluble 6-isomer is separated before and after liming to give a crude crocein acid liquor containing 16 % Schaeffer acid. If pure **53** is required, the latter is separated by selective coupling with diazotized *o*-toluidine and removal of the precipitated azo dye.

Uses. Crocein acid is used as a coupling component for azo dyes e.g., C.I. Acid Red 25 (**77**).

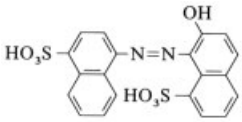

77

4.4. 1-Hydroxynaphthalenedisulfonic Acids

1-Hydroxynaphthalene-3,6-disulfonic acid [*578-85-8*] (**59**) (4-hydroxynaphthalene-2,7-disulfonic acid), RG acid, Violet acid, $C_{10}H_8O_7S_2$, M_r 304.3: the free acid (pK_a 8.56) and its alkalimetal salts are readily soluble in water, but the acid sodium salt can be precipitated from concentrated solution with salt. Sulfonation with oleum at 125 °C gives a tetrasulfonic acid, and amination with ammonia and ammonium chloride at 160 °C gives a mixture of 3-amino-1-hydroxynaphthalene-6-sulfonic acid and 6-amino-1-hydroxynaphthalene-3-sulfonic acid. Alkali fusion at 170 °C gives mainly 1,6-dihydroxynaphthalene-3-sulfonic acid, and at 250 °C yields 1,3,6-trihydroxynaphthalene. Diazo coupling takes place in the 2-position.

Production. Naphthalene-1,3,6-trisulfonic acid (Section 2.4) is heated with 50 % sodium hydroxide at 180 °C for 15 h in an autoclave [7]. After dilution with water, the product is salted out as the monosodium salt and purified by washing with brine.

Uses. 1-Hydroxynaphthalene-3,6-disulfonic acid is used as a coupling component for azo dyes, e.g., C.I. Acid Red 7 (**78**).

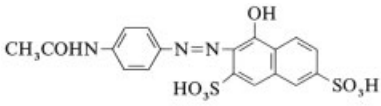

78

1-Hydroxynaphthalene-3,8-disulfonic acid [*117-43-1*] (**79**) (8-hydroxynaphthalene-1,6-disulfonic acid), epsilon acid, $C_{10}H_8O_7S_2$, M_r 304.3: the sodium salt crystallizes as the hexahydrate and is readily soluble in water.

Alkali fusion at 200 °C gives 1,8-dihydroxynaphthalene-3-sulfonic acid and amination with ammonia and ammonium chloride at 170 °C gives 6,8-diaminonaphthalene-1-sulfonic acid. Diazo coupling takes place in the 2-position and bromination in the 4-position. The 1,8-sultone is formed on dehydration with concentrated sulfuric acid.

Production. The acid sodium salt of 1-aminonaphthalene-3,8-disulfonic acid (Section 6.2) is hydrolyzed by heating with water under pressure at 180 °C.

Uses. Epsilon acid is used as a coupling component for azo dyes, e.g., C.I. Direct Red 47 (**80**).

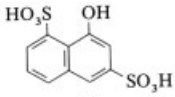

79

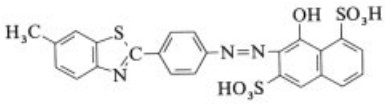

80

1-Hydroxynaphthalene-4,7-disulfonic acid [*6361-37-1*] (**81**) (4-hydroxynaphthalene-1,6-disulfonic acid), D(Dahl's)acid, $C_{10}H_8O_7S_2$, M_r 304.3: its alkali-metal salts are readily soluble in water. Boiling with 65 % sulfuric acid results in desulfonation and formation of 1-hydroxynaphthalene-7-sulfonic acid (**72**). Diazo coupling and nitrosation take place in the 2-position.

Production. 1-Aminonaphthalene-4,7-disulfonic acid (Section 6.2) is heated with sodium bisulfite under conditions similar to those described for the production of 1-hydroxynaphthalene-4-sulfonic acid (**47**) from sodium naphthionate (see Section 4.2). High yields of **81** have been claimed via sulfonation of 1-(dichlorophosphinyloxy)naphthalene (1-naphthol + $POCl_3$) [40].

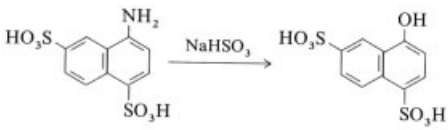

81

1-Hydroxynaphthalene-4,8-disulfonic acid [*117-56-6*] (**83**) (4-hydroxynaphthalene-1,5-disulfonic acid), Schollkopf's acid, delta acid, oxy Chicago acid, $C_{10}H_8O_7S_2$, M_r 304.3: the sodium salt is readily soluble in water, but the barium salt is only sparingly soluble. Cold 5 % oleum dehydrates **83** to the sultone **82**, whereas with hot 25 % oleum it undergoes sulfonation to give 1-hydroxynaphthalene-2,4,8-trisulfonic acid. Amination (Bucherer) gives 1-aminonaphthalene-4,8-disulfonic acid and caustic fusion gives 1,8-dihydroxynaphthalene-4-sulfonic acid. Diazo coupling and nitrosation take place in the 2-position.

Production. 1-Hydroxynaphthalene-8-sulfonic acid-1,8-sultone (**73**) (Section 4.2) is heated with sulfuric acid at 80 – 90 °C. The reaction mixture is then quenched by pouring into concentrated brine, whereby the sodium salt of **82** crystallizes. After isolation, this intermediate is hydrolyzed by heating with aqueous sodium carbonate and the product is isolated as its sodium salt after cooling.

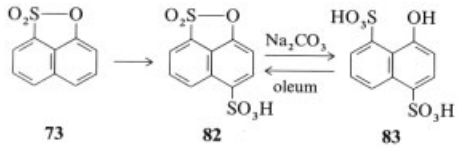

73 **82** **83**

Uses. Delta acid is used as a coupling component for azo dyes, e.g., C.I. Acid Blue 169.

4.5. 2-Hydroxynaphthalenedisulfonic Acids

2-Hydroxynaphthalene-3,6-disulfonic acid [*148-75-4*] (**54**) (3-hydroxynaphthalene-2,7-disulfonic acid), R acid, $C_{10}H_8O_7S_2$, M_r 304.3, is readily soluble in water and ethanol. Although the sodium salt (R salt) is soluble in water, it can be precipitated by sodium chloride. The barium salt is sparingly soluble in cold water but is 8 % soluble in boiling water. Whereas sulfonation with 20 % oleum yields 2-hydroxynaphthalene-3,6,8-trisulfonic acid (**56**), the reaction with Na_2SO_3–MnO_2 gives 2-hydroxy-1,3,6-trisulfonic acid. Alkali fusion with sodium hydroxide gives 2,3-dihydroxynaphthalene-6-sulfonic acid and some 2,3-dihydroxynaphthalene. Amination (Bucherer) produces

2-aminonaphthalene-3,6-disulfonic acid. Diazo coupling and nitrosation take place in the 1-position.

Production. 2-Naphthol is heated in excess 98 % sulfuric acid to 60 °C, and anhydrous sodium sulfate is added [7]. The mixture is then heated for 24 h at 105 °C, 12 h at 110 °C, and 12 h at 120 °C before being diluted with more sulfuric acid and quenched in water. Sodium chloride is added at 60 °C, and the mixture is cooled slowly to 30 °C with crystallization of the product as its sodium salt. The isolated yield is 68 %, and the product contains about 3 % of Schaeffer's salt as the major impurity. Alternatively, R salt may be recovered as a byproduct from the G acid process (see page 22). Further purification may be effected via the aniline salt [41].

Uses. The so-called R acid is used as a coupling component for a wide range of azo colorants, e.g., C.I. Mordant Red 9, C.I. Pigment Red 60, C.I. Acid Red 27, and C.I. Food Red 9 (**84**).

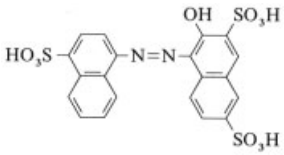

84

The aluminum salt of R acid has been used as an astringent and antiseptic (Alumnol).

This acid is an intermediate in the manufacture of 2-aminonaphthalene-3,6-disulfonic acid and 2,3-dihydroxynaphthalene-6-sulfonic acid.

2-Hydroxynaphthalene-3,7-disulfonic acid [*6361-38-2*] (**85**) (3-hydroxynaphthalene-2,6-disulfonic acid), $C_{10}H_8O_7S_2$, M_r 304.3: the alkali-metal salts of **85** are readily soluble in water, but its barium salt is only sparingly soluble even at 100 °C. Sulfonation with 20 % oleum at 85 °C gives mainly 2-hydroxynaphthalene-1,3,7-trisulfonic acid. Amination yields 2-aminonaphthalene-3,7-disulfonic acid, and caustic fusion yields 2,7-dihydroxynaphthalene-3-sulfonic acid. Diazo coupling takes place in the 1-position.

Production. 2-Hydroxynaphthalene-7-sulfonic acid (Section 4.3) is heated at 120 °C with concentrated sulfuric acid for 12 h, the mixture is poured into water, and the product is separated after liming out.

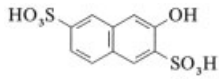

85

2-Hydroxynaphthalene-5,7-disulfonic acid [*575-05-3*] (**68**) (6-hydroxynaphthalene-1,3-disulfonic acid), $C_{10}H_8O_7S_2$, M_r 304.3, and its alkali-metal salts are readily soluble in water. Diazo coupling takes place in the 1-position.

Production. 2-Aminonaphthalene-5,7-disulfonic acid (Section 6.2) is heated with 5 % sulfuric acid for 15 h at 185 °C under pressure [7]. After liming out and removal of ammonia and calcium ions, the product is obtained as a solution of the sodium salt for further use.

Uses. The acid has some use as an azo coupling component and as an intermediate in the production of 2,5-dihydroxynaphthalene-7-sulfonic acid.

2-Hydroxynaphthalene-6,8-disulfonic acid [*118-32-1*] (**55**) (7-hydroxynaphthalene-1,3-disulfonic acid), G acid, $C_{10}H_8O_7S_2$, M_r 304.3: most salts of **55** are more readily soluble in water than those of R acid, with the notable exception of the potassium salt. Sulfonation with 20 % oleum at 100 °C yields 2-hydroxynaphthalene-3,6,8-trisulfonic acid (**56**). Diazo coupling takes place in the 1-position, although the reaction is sluggish, as in the case of 2-hydroxynaphthalene-8-sulfonic acid.

Production. 2-Naphthol is added to 98 % sulfuric acid at 40 °C. Then, 20 % oleum is added over 6 h and the temperature is allowed to rise to 60 °C. The reaction is completed by heating for a further 16 h at 60 °C followed by 15 h at 80 °C. After the reaction mixture has been poured into water, potassium chloride is added to the hot solution and the batch is cooled slowly to 35 °C before filtering off the crystalline dipotassium salt of G acid and carefully washing out any residual R salt; the yield is 60 %. In addition, R salt can be recovered from the filtrate in 12 % yield after conversion to its sodium salt.

Uses. The G acid is aminated (Bucherer reaction) to give 2-aminonaphthalene-6,8-disulfonic acid and subjected to caustic fusion at 200 °C to

yield 2,8-dihydroxynaphthalene-6-sulfonic acid (a precursor of γ acid). It is also used as a coupling component for a wide range of azo colorants, e.g., C.I. Acid Orange 10, C.I. Acid Red 187, and C.I. Food Red 7 (**86**).

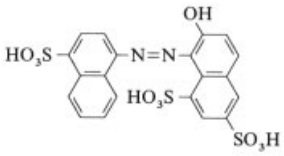

86

4.6. Hydroxynaphthalenetrisulfonic Acids

1-Hydroxynaphthalene-3,6,8-trisulfonic acid [*3316-02-7*] (**69**) (8-hydroxynaphthalene-1,3,6-trisulfonic acid), oxy Koch acid, $C_{10}H_8O_{10}S_3$, M_r 384.3, and its alkali-metal salts are readily soluble in water. Amination at 170 °C gives 1-aminonaphthalene-3,6,8-trisulfonic acid. Diazo coupling takes place in the 2-position.

Production. 1-Aminonaphthalene-3,6,8-trisulfonic acid (Section 6.3) is heated with dilute sulfuric acid at 180 °C under pressure for 20 h [7]. After the mixture is run into excess sodium hydroxide solution, ammonia is removed by heating and the total liquor is used directly for the production of chromotropic acid (Section 4.7) by caustic fusion.

2-Hydroxynaphthalene-3,6,8-trisulfonic acid [*6259-66-1*] (**56**) (7-hydroxynaphthalene-1,3,6-trisulfonic acid), $C_{10}H_8O_{10}S_3$, M_r 384.3, and its alkali-metal salts are readily soluble in water; nevertheless, the sodium salt is obtainable by salting out. Amination at 240 °C gives 2-aminonaphthalene-3,6,8-trisulfonic acid. Alkali fusion gives a mixture of dihydroxynaphthalenedisulfonic acids and finally 4,6,7-trihydroxynaphthalene-2-sulfonic acid. Diazo coupling occurs slowly in the 1-position.

Production. 2-Naphthol is heated with sulfuric acid monohydrate at 115 °C for 3 h. The reaction mixture is cooled and 65 % oleum is added, followed by heating at 115 °C for 12 h [7], after which it is poured into water. Salt is then added to the hot solution and the sodium salt of the product crystallizes. Further purification is effected by redissolving in water and salting out.

Uses. The acid is used as an azo coupling component in, e.g., C.I. Solvent Red 31 (the dicyclohexylamine salt of **87**).

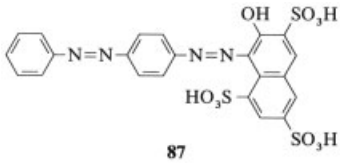

87

4.7. Dihydroxynaphthalenesulfonic Acids

Although many isomers are known [5], only the four compounds that have been used industrially (mainly as azo coupling components) are described. All are obtained in high yield by selective fusion processes.

4,5-Dihydroxynaphthalene-1-sulfonic acid [*83-65-8*] (**60**) (1,8-dihydroxynaphthalene-4-sulfonic acid), dioxy Chicago acid, dioxy S acid, $C_{10}H_8O_5S$, M_r 240.2, is sparingly soluble in water, but the sodium salt is readily soluble. Heating with dilute sulfuric acid at 150 °C gives 1,8-dihydroxynaphthalene (Section 3.3).

Production. 1-Hydroxynaphthalene-4,8-disulfonic acid (Section 4.4) is heated with 82 % caustic soda liquor for 3 h at 235 °C [7]. The melt is poured into excess hydrochloric acid and sulfur dioxide is expelled. After cooling, the product is filtered off and washed with brine; the yield is 83 %.

4,6-Dihydroxynaphthalene-2-sulfonic acid [*6357-93-3*] (**61**) (2,8-dihydroxynaphthalene-6-sulfonic acid), dioxy G acid, $C_{10}H_8O_5S$, M_r 240.2, is readily soluble in water, but the sodium salt is less so and can be salted out.

Production. The potassium salt of 2-hydroxynaphthalene-6,8-disulfonic acid (Section 4.4) is heated with 70 % caustic liquor for 3 h at 230 °C and then at 240 °C for completion of the reaction. The melt is diluted with water and poured into excess hydrochloric acid. After sulfur dioxide has been expelled, salt is added and

the batch is cooled to precipitate the product, which is filtered off and washed with brine; the yield is 80 %.

Uses. Dioxy G acid is used as an intermediate in the production of 2-amino-8-hydroxynaphthalene-6-sulfonic acid (γ acid) and its N-substituted derivatives by means of the Bucherer reaction with ammonia or primary amines. It is also used as a coupling component for azo dyes such as C.I. Mordant Black 56 (**88**).

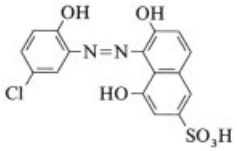

88

4,7-Dihydroxynaphthalene-2-sulfonic acid [*6357-94-4*] (**62**) (2,5-dihydroxynaphthalene-7-sulfonic acid), dioxy J acid, $C_{10}H_8O_5S$, M_r 240.2, is readily soluble in water; its sodium salt is much less soluble and is easily salted out. Amination with NH_3–$NaHSO_3$ yields J acid (Section 8.2).

Production. 2-Hydroxynaphthalene-5,7-disulfonic acid (Section 4.4) is heated with aqueous NaOH – KOH for 5 h at 175 °C [7]. The melt is poured into water and acidified with hydrochloric acid, and the precipitated solid is filtered off and washed with brine; the yield is 80 %.

Uses. Although not used as a precursor for J acid itself, **62** can be used advantageously for the production of some N-substituted J acids.

6,7-Dihydroxynaphthalene-2-sulfonic acid [*92-27-3*] (**63**) (2,3-dihydroxynaphthalene-6-sulfonic acid), dioxy R acid, $C_{10}H_8O_5S$, M_r 240.2, is readily soluble in water. The sodium salt is sparingly soluble in cold water but readily soluble in hot water. Heating with dilute sulfuric acid at 190 °C gives 2,3-dihydroxynaphthalene (Section 3.3).

Production. 2-Hydroxynaphthalene-3,6-disulfonic acid (Section 4.4) is heated with 70 % caustic soda liquor for 4 h at 270 °C, after which the reaction mixture is diluted with water and poured into excess hydrochloric acid [7]. The precipitated product is filtered off and washed with brine; the yield is 90 %.

4.8. Dihydroxynaphthalenedisulfonic Acids

Although many isomers are known [5], only the three significant ones are described here. All are obtained by selective fusion processes, although yields are lower than for the monosulfonic acids (Section 4.7).

3,5-Dihydroxynaphthalene-2,7-sulfonic acid [*81344-22-1*] (**64**), A acid, $C_{10}H_8O_8S_2$, M_r 320.3.

Production. 2-Hydroxynaphthalene-3,6,8-trisulfonic acid (Section 4.6) is heated with caustic soda liquor at 240 °C, and the melt is poured into water and neutralized. Careful addition of salt precipitates the product and leaves the more soluble 2,3-dihydroxy byproduct in solution.

3,6-Dihydroxynaphthalene-2,7-disulfonic acid [*23894-07-7*] (**65**), $C_{10}H_8O_8S_2$, M_r 320.3.

Production. 2-Hydroxynaphthalene-3,6,7-trisulfonic acid is heated with 66 % caustic soda liquor at 200 – 300 °C and the product worked up by dilution, neutralization, and salting out.

4,5-Dihydroxynaphthalene-2,7-disulfonic acid [*148-25-4*] (**66**) (1,8-dihydroxynaphthalene-3,6-disulfonic acid), chromotropic acid, $C_{10}H_8O_8S_2$, M_r 320.3: the disodium salt is readily soluble in water. Alkali fusion yields 1,3,8-trihydroxynaphthalene-6-sulfonic acid. Diazo coupling occurs in the 3- and 6-positions.

Production. 2-Hydroxynaphthalene-3,6,8-trisulfonic acid total liquor (Section 4.6) is concentrated and heated with caustic soda liquor to 158 °C for 20 – 30 h [7]. The melt is diluted with water and poured into excess hydrochloric acid. Salt is added, sulfur dioxide expelled, and the mixture cooled to precipitate the product as the disodium salt in 88 % yield.

An alternative process is based on direct fusion (caustic liquor at 228 °C under pressure) of 1-aminonaphthalene-3,6,8-trisulfonic acid (Section 6.3) to chromotropic acid without isolating the intermediate oxy Koch acid [7].

Uses. Chromotropic acid is used as an analytical reagent and as a coupling component for a wide range of azo dyes, e.g., C.I. Acid Violet 6, C.I. Mordant Blue 13, and C.I. Direct Blue 84 (the dicopper complex of demethylated **89**).

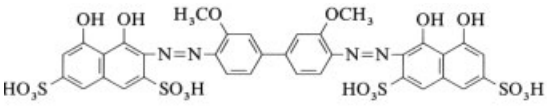

89

5. Aminonaphthalenes

The two general methods for the synthesis of monoamino- and diaminonaphthalenes are reduction or hydrogenation of the corresponding nitronaphthalene derivative, and amination (Bucherer reaction) of the corresponding hydroxynaphthalene derivative. Although the former method is analogous to the general method for substituted anilines, the Bucherer reaction is not applicable to phenols.

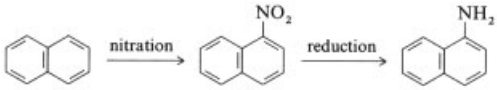

90

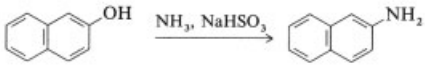

91

5.1. Naphthylamines

1-Naphthylamine [*134-32-7*] (**90**), 1-naphthaleneamine, 1-aminonaphthalene, α-naphthylamine, $C_{10}H_9N$, M_r 143.18, *mp* 50 °C, *bp* 301 °C, pK_a 10.0, d_4^{20} 1.131, sublimes and is steam volatile. It crystallizes from aqueous ethanol as colorless needles, which slowly turn red on exposure to air. Solubility in water is 0.16 % at 20 °C. It behaves as a typical primary aromatic amine in forming salts with strong acids (but not with acetic or benzoic acid) and readily forming *N*-acyl derivatives.

1-Naphthylamine couples with diazo compounds in the 4-position, with up to 10 % byproduct being formed by coupling in the 2-position.

Oxidation of 1-naphthylamine salts with iron(III) chloride in aqueous solution gives an insoluble blue-violet compound, whereas oxidation with chromic acid yields 1,4-naphthoquinone and phthalic acid. Reduction with Na – NaOEt leads to the 5,6,7,8-tetrahydro compound, which can also be formed together with the 1,2,3,4-tetrahydro derivative by catalytic hydrogenation. Sulfonation gives a complex mixture of mono-, di-, and trisulfonic acids.

Production. Naphthalene is nitrated with H_2SO_4–HNO_3 at 50 – 60 °C [7]. The crude product (*mp* 52 °C) is separated, washed with hot water, and further purified by crystallization or sweating (draining off the lower melting isomer) to remove 2-nitronaphthalene (3 %) and traces of 2,4-dinitronaphthalene. Purified 1-nitronaphthalene was traditionally reduced with iron in boiling dilute hydrochloric acid, but modern plants use hydrogenation with a nickel catalyst. The 1-naphthylamine produced is further purified by distillation under vacuum. The content of 2-naphthylamine in the commercial product is specified at <10 ppm.

Uses. A large proportion of the 1-naphthylamine produced is converted to 1-naphthol (Section 3.1). Other major uses are in the production of aminonaphthalenesulfonic acid dye intermediates (Chap. 6) and as a component of azo dyes. In the latter, use as a coupling component, followed by diazotization and further coupling of the resulting aminoazo compound, predominates, e.g., for C.I. Acid Blue 113 (**92**).

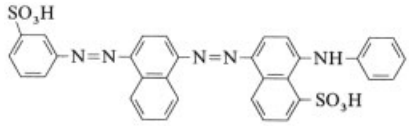

92

Other important derivatives are C.I. Solvent Brown 1, C.I. Acid Red 20, C.I. Direct Brown 208, and C.I. Mordant Black 68.

The rubber antioxidant 1-phenylaminonaphthalene (**24**) can be obtained from 1-naphthol (Section 3.1). However, because the latter is now manufactured from 1-naphthylamine, it is more convenient to manufacture **24** directly from 1-naphthylamine and aniline by a vapor-phase catalytic process (alumina gel, 300 °C). *N*-Alkyl-1-naphthylamines are usually manufactured from 1-naphthol; for example, *N*-ethyl-1-naphthylamine is of particular importance as an intermediate for Victoria Blues.

The selective preemergence herbicide naptalam [*132-66-1*] (**93**), *N*-1-naphthylphthalamic acid, is produced by acylation of **90** by heating with an equimolar quantity of phthalic anhydride in a solvent such as xylene [42].

The rodenticide ANTU [*86-88-4*] (**94**), 1-naphthalenethiourea, is manufactured from 1-naphthylamine by reaction with thiocyanate ion in aqueous acid solution. The elimination of ammonia from **94** by heating in chlorobenzene is a convenient method for producing 1-naphthyl isothiocyanate.

Acylation of *N*-methyl-1-naphthylamine with fluoroacetyl chloride yields the insecticide and acaricide Nissol [*5903-13-9*] (**95**). An improved process is based on reaction of the corresponding chloroacetamido derivative with potassium fluoride in a glycol solvent [43].

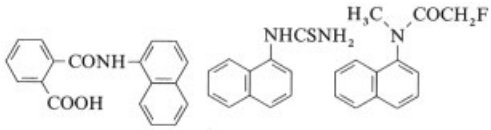

93 **94** **95**

Toxicology [14], [44]. 1-Naphthylamine is toxic to humans by inhalation or skin absorption and is classed as a potential carcinogen. Samples are often contaminated with highly carcinogenic 2-naphthylamine, which confused many earlier results. Although frequently quoted, data on toxicity to animals are inconclusive compared with those available for the 2-isomer.

The metabolism of 1-naphthylamine in animals and humans has been studied, with a variety of results. *N*-(1-naphthyl)hydroxylamine and *N*-hydroxy-1-naphthylacetamide appear to be key metabolic products, in contrast to the behavior of 2-naphthylamine.

Use in the United Kingdom is controlled by the Carcinogenic Substances Regulations of 1967 and in the United States by the Federal Register of Carcinogens (1973).

2-Naphthylamine [*91-59-8*] (**91**), 2-aminonaphthalene, 2-naphthaleneneamine, *β*-naphthylamine, $C_{10}H_9N$, M_r 143.18, *mp* 112 °C, *bp* 306 °C, d_4^{98} 1.061 is steam volatile and is soluble in hot water, ethanol, and ether. It crystallizes from aqueous ethanol as white crystals that turn red on exposure to light and air.

Oxidation with hypochlorite, lead dioxide, or air in UV light gives dibenzophenazine. Reduction with sodium in amyl alcohol or catalytic hydrogenation gives mainly 2-aminotetralin. Substition reactions (e.g., halogenation) occur first in the 1-position unless the amino group is protonated, in which case substituents enter the 5- and 8-positions.

Production. 2-Naphthol is heated with an $NH_4OH - NH_4HSO_3$ mixture at 180 °C for 20 h, and the molten amine is separated after basification and removal of excess ammonia. After being dried, the oil is distilled under vacuum. Production of 2-naphthylamine ceased worldwide in the 1950s.

Uses. Historically, 2-naphthylamine was an important colorant intermediate. Two examples are solvent sulfonation to Tobias acid (Section 6.1) and formation of Naphthol BN (**96**). Until 1950 the latter was the most important azoic coupling component after Naphthol AS (**44**). After production of 2-naphthylamine ceased, Naphthol BN was prepared for a time via Tobias acid and desulfonation.

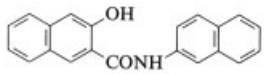

96

Although *N*-phenyl-2-aminonaphthalene (**32**) was produced originally from 2-naphthylamine, all production of this antioxidant is now based on 2-naphthol (Section 3.2).

Toxicology [14], [44]. 2-Naphthylamine is a human carcinogen that leads predominately to formation of tumors in the epithelium of the bladder. The conclusive epidemiological studies on humans are backed up by a wealth of animal data. However, although a proven carcinogen in mouse, hamster, dog, and monkey, 2-naphthylamine exhibits little, if any, carcinogenicity in rat or rabbit. Several metabolic pathways have been identified, but the main one involves N-hydroxylation followed by rearrangement to 2-amino-1-hydroxynaphthalene. Only the former is a proven animal carcinogen.

Measurement of 2-naphthylamine impurity levels requires sophisticated analysis. For example, 22 ng per cigarette was measured by GLC of the *N*-pentafluoropropionyl derivative. The levels found were considered to be biologically insignificant [45].

Use in the United Kingdom is prohibited under the Carcinogenic Substances Regulations of 1967; in the United States, by the Environmental Protection Agency; and in all industrialized nations, by comparable regulatory authorities.

Although *N*-phenyl-2-aminonaphthalene (PBN, **32**) has been classed as a potential carcinogen, its toxicity is more likely to be due to the presence of primary amines as impurities.

5.2. Naphthalenediamines

All ten of the isomeric diamines are known [5]. Table 5 lists their melting points, CAS registry numbers, and solubilities. Only the 1,5- and 1,8-isomers are readily available and of commercial significance.

Table 5. Physical properties of naphthalenediamines

Isomer	CAS re gistry no.	*mp*, °C	Solubility
1,2	[*938-25-0*]	98	soluble in hot water, alcohol, and ether
1,3	[*24824-28-0*]	96	
1,4	[*2243-61-0*]	120	slightly soluble in hot water
1,5	[*2243-62-1*]	189	soluble in hot water and alcohol
1,6	[*2243-63-2*]	78	soluble in hot water and alcohol
1,7	[*2243-64-3*]	117	soluble in alcohol
1,8	[*479-27-6*]	66	soluble in alcohol and ether
2,3	[*771-97-1*]	199	soluble in alcohol and ether
2,6	[*2243-67-6*]	220	slightly soluble in alcohol and ether
2,7	[*613-76-3*]	166	

1,5-Naphthalenediamine [*2243-62-1*] (**97**), 1,5-diaminonaphthalene, Alphamin, $C_{10}H_{10}N_2$, M_r 158.2: oxidation of **97** with iron(III) chloride in water produces a blue-violet color. Treatment with boiling aqueous sodium bisulfite followed by addition of alkali gives a mixture of 1-amino-5-hydroxynaphthalene and 1,5-dihydroxynaphthalene. Sulfonation (5 % oleum, 100 °C) gives 1,5-diaminonaphthalene-2-sulfonic acid, and nitration in acetic acid produces 2,4,6,8-tetranitro-1,5-diaminonaphthalene. Coupling with diazo compounds takes place in the 2-position; reduction of the resulting azo compound with $SnCl_2$–HCl produces 1,2,5-triaminonaphthalene.

Production. 1-Nitronaphthalene can be nitrated further to give a 40 : 60 mixture of 1,5- and 1,8-dinitronaphthalenes. Similar results are obtained by direct nitration of naphthalene with H_2SO_4–HNO_3 under careful control of temperature over the range 40 – 80 °C. Although separation of the isomers by fractional crystallization or solvent extraction [46] is usually carried out at this stage, the mixed isomers can also be reduced and the resulting diamines separated. Reduction of the dinitronaphthalenes is achieved by treatment of a nonaqueous solution with iron or hydrogen in the presence of a catalyst.

An alternative process for 1,5-naphthalenediamine involves amination of 1,5-dihydroxynaphthalene with ammonia and ammonium bisulfite. Although less efficient on a stage basis it offers an economical alternative to nitration and reduction if the 1,8-naphthalenediamine (**98**) is not also required.

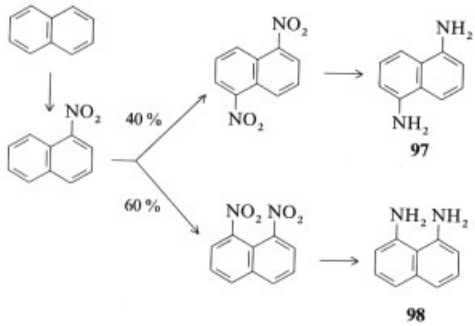

Uses. 1,5-Naphthalenediamine is an intermediate in the manufacture of naphthalene-1,5-diisocyanate [*3173-72-6*] (**99**), which is used for specialty polyurethane polymers. It has only limited use as a dye intermediate.

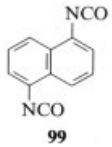

Toxicity. Like 1-naphthylamine, 1,5-naphthalenediamine is classed as having limited carcinogenicity in animals [14], [44].

1,8-Naphthalenediamine [*479-27-6*] (**98**), 1,8-diaminonaphthalene, Deltamin, $C_{10}H_{10}N_2$, M_r 158.2, turns brown in air even in the ab-

sence of light. Its solutions give a brown precipitate on addition of iron(III) chloride. Treatment with boiling aqueous sodium bisulfite followed by addition of alkali gives 8-amino-1-hydroxynaphthalene. Sulfonation takes place in the 4-position. Coupling with azo compounds also occurs in the 4-position, but in almost neutral solution 4,5-bis coupling can occur. The resulting bisazo compound is reduced with $SnCl_2$–HCl to give 1,4,5,8-tetraaminonaphthalene.

Production. Reduction of 1,8-dinitronaphthalene is carried out with iron and acetic acid in xylene or by catalytic hydrogenation.

Uses. 1,8-Naphthalenediamine has minor use as an antioxidant for lubricating oils and as an analytical reagent for the detection of selenium and nitrites.

Treatment of a solution of the hydrochloride of **98** with acetic anhydride yields methylperimidine (**100**), a colorant intermediate. Most important are the phthaloperinone pigments obtained by reaction of **98** with substituted phthalic anhydrides [47]. The parent product (**101**) is a lightfast, heat-stable orange dye for poly(methyl methacrylate) and is used to color the covers for automobile indicator lights.

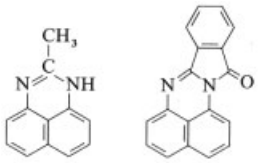

100 **101**

6. Aminonaphthalenesulfonic Acids

Many aminonaphthalenesulfonic acids are important intermediates either for the direct production of azo dyes or for conversion to more highly substituted intermediates for azo dyes. About 24 products in this group have been made industrially, and many of them are still commercially available and referred to by code names (Table 1). The most important are the C, Peri, and Laurent's acids [17]. Only those isomers that have a history of large-scale use are dealt with here on an individual basis. Many other isomers are also obtainable and may be of potential present or future interest [5].

Because in usage this group of products is considered to be derived from 1- and 2-naphthylamines they are dealt with on this basis rather than as amino derivatives of naphthalenesulfonic acids. The latter is the IUPAC convention, and these names are given in parentheses alongside alternative trivial names.

6.1. Production and Properties

Production depends on four main routes to achieve the required substitution patterns. As with most naphthalene substitution reactions, isomeric mixtures often arise so that isomer separation or purification to remove unwanted byproducts is required. High-performance liquid chromatography is an essential analytical tool for following separations and assessing the purity of products.

Sulfonation of Naphthylamines. Sulfonation of 1-naphthylamine gives a variety of products depending on conditions (Fig. 3). Further sulfonation of the separated and isolated primary products leads to di- and trisulfonic acids.

Although 2-naphthylamine can be sulfonated to give a range of isomeric mono- and disulfonic acids, this method is not discussed here because use of this starting material is now prohibited.

Nitration – Reduction of Naphthalenesulfonic Acids. Both naphthalene-1-sulfonic acid and 2-sulfonic acid undergo nitration to give separable mixtures of 5- and 8-nitro compounds, which are isolated and reduced to the corresponding aminonaphthalenesulfonic acids (Fig. 4).

Naphthalene-1,5-disulfonic acid is nitrated to give the 3-nitro derivative, which is reduced to 2-aminonaphthalene-4,8-disulfonic acid (**118**). Other available disulfonic acid isomers are similarly mononitrated and reduced to aminonaphthalenedisulfonic acids (see. Fig. 5).

Amination of Hydroxynaphthalenesulfonic Acids. The Bucherer reaction is most readily applicable to 2-hydroxynaphthalenemono- and disulfonic acids, as shown in the following examples.

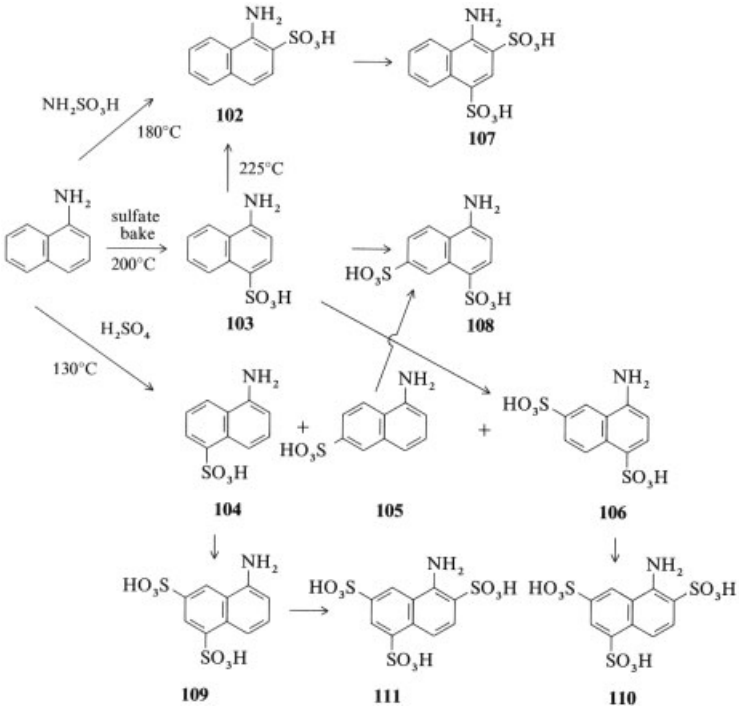

Figure 3. Sulfonation of 1-naphthylamine

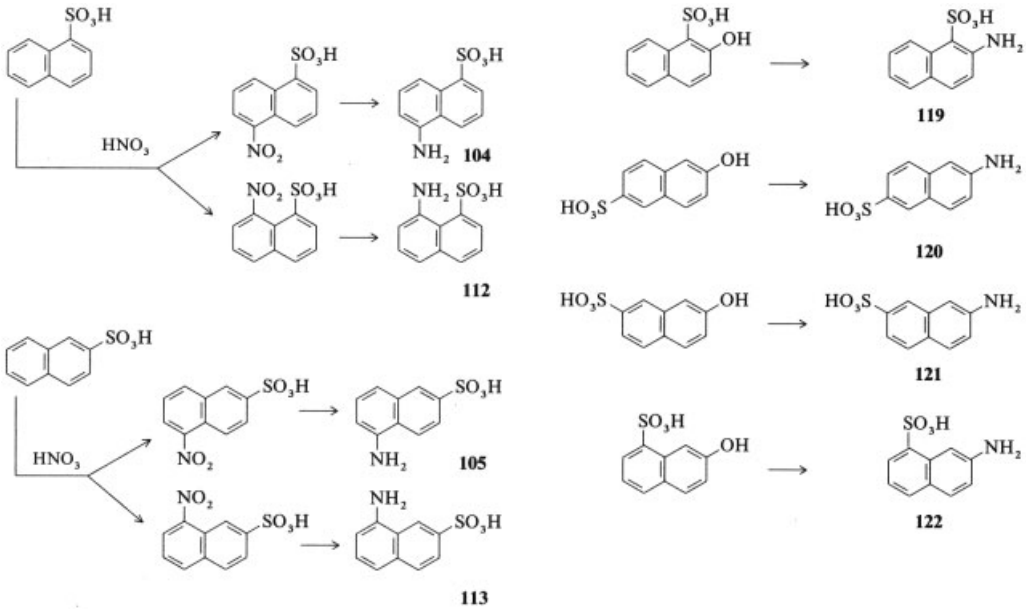

Figure 4. Nitration – reduction of naphthalenesulfonic acids

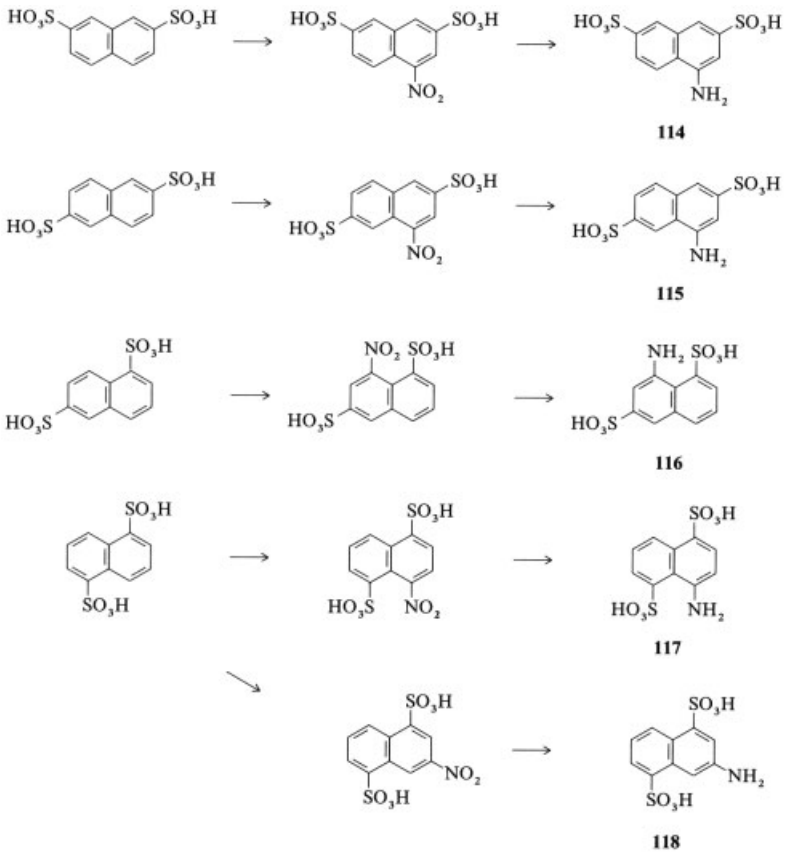

Figure 5. Nitration – reduction of naphthalenedisulfonic acids

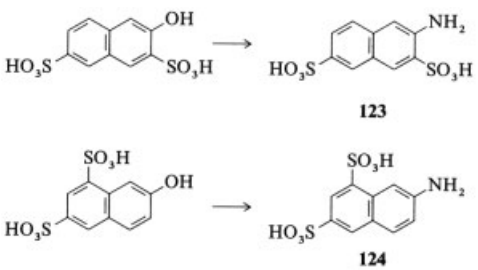

Desulfonation. As with certain hydroxynaphthalenedi- and trisulfonic acids (Section 4.1), selective desulfonation is exploited with aminonaphthalenedi- and trisulfonic acids. Heating with dilute sulfuric acid is the standard process; sulfonic acid groups adjacent to the amino group are usually labile. A different mechanism may apply when sodium amalgam is required in the case of peri-desulfonation, e.g., with **116** (see top of the next page).

Although many naphthalene derivatives exhibit UV fluorescence in solution, aminonaphthalenesulfonic acids are notable in commonly showing solution fluorescence in daylight. The fluorescent color is in the green – blue – violet area [15], with no obvious correlation between color and structure. Nevertheless, an attempt has been made to systematize the spectral identification of substituted naphthalenesulfonic acids as a contribution to the structural determination of azo dyes after reductive scission [48].

The fluorescence has been exploited with 5-dimethylamino-1-naphthalenesulfonyl chloride [*605-65-2*] (dansyl chloride), which is reacted with amino acids for labeling purposes in protein sequencing.

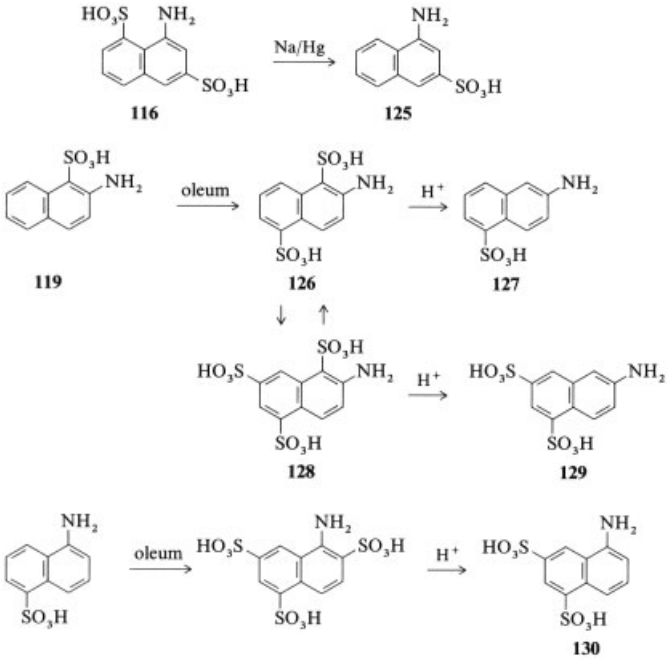

6.2. 1-Aminonaphthalenesulfonic Acids

1-Aminonaphthalene-2-sulfonic acid [*81-06-1*] (**102**), *ortho*-naphthionic acid, $C_{10}H_9NO_3S$, M_r 223.24, *mp* 262 – 265 °C (decomp.), is sparingly soluble in cold water but dissolves in hot water to give a 3 % solution. The sodium salt is more readily soluble in water (1.7 % at 20 °C and 10 % at 100 °C). The calcium salt is similarly soluble, but the barium salt is almost insoluble.

Production. 1-Naphthylamine (Section 5.1) can be sulfonated by heating with sulfuric acid at 180 – 185 °C, but the transient intermediate naphthalene-1-sulfamic acid is best prepared from 1-aminonaphthalene-4-sulfonic acid (naphthionic acid) [7]. When sodium naphthionate is heated in a solvent such as *o*-dichlorobenzene or xylene at 225 °C, thermal rearrangement occurs to give the 1,2-isomer in high yield.

1-Aminonaphthalene-3-sulfonic acid [*134-54-3*] (**125**) (4-aminonaphthalene-2-sulfonic acid), $C_{10}H_9NO_3S$, M_r 223.24, is only sparingly soluble in water, but its alkali-metal salts are readily soluble. Sulfonation with 20 % oleum yields 1-aminonaphthalene-3,5-disulfonic acid, and heating with water under pressure gives 1-hydroxynaphthalene-3-sulfonic acid, which can also be obtained by diazotization and acid hydrolysis of **125**.

Production. 1-Aminonaphthalene-3,8-disulfonic acid (Section 6.4) in dilute sodium hydroxide solution is pumped at 40 °C through a special reactor that effects contact with continuously produced sodium amalgam [7]. The resulting liquor is used directly for production of 1-hydroxynaphthalene-3-sulfonic acid (Section 4.2).

1-Aminonaphthalene-4-sulfonic acid [*84-86-6*] (**103**) (4-aminonaphthalene-1-sulfonic acid), naphthionic acid, Piria's acid, $C_{10}H_9NO_3S$, M_r 223.24, is sparingly soluble in water (0.25 % at 100 °C), but the sodium, potassium, calcium, and barium salts are all readily soluble.

Chlorination of **103** with copper(II) chloride produces 4-chloro-1-aminonaphthalene, whereas the action of chlorine on an aqueous sulfuric acid solution gives 2,3-dichloro-1,4-naphthoquinone. High-temperature bromination gives 2,4-dibromo-1-aminonaphthalene, whereas vigorous nitration yields 2,4-dinitro-1-hydroxynaphthalene. Sulfonation with 25 %

oleum gives 1-aminonaphthalene-4,6- and 4,7-disulfonic acids, whereas prolonged heating with sulfuric acid at 130 °C leads to 1-aminonaphthalene-5- and 6-sulfonic acids. Hydrolysis of the 1-amino group to give 1-hydroxynaphthalene-4-sulfonic acid is best achieved by a reverse-Bucherer reaction but also results from prolonged heating with 50 % sodium hydroxide.

Production. Naphthionic acid was first prepared by PIRIA by heating 1-nitronaphthalene with ammonium sulfite; however, the formation of 1-aminonaphthalene-2,4-disulfonic acid and other byproducts makes this route unattractive for large-scale production. Although it can also be prepared by direct sulfonation of 1-naphthylamine, it is more conveniently produced by a solvent-bake process [6]. 1-Naphthylamine is dissolved in *o*-dichlorobenzene and H_2SO_4 (molar ratio 1 : 1) added to form a suspension of the sulfate. After being heated gradually to 180 °C, the reaction is completed by maintaining this temperature until all the water of reaction has been distilled off with solvent recycle. The product is extracted from the final mixture by addition to aqueous sodium carbonate solution. After separation, the aqueous layer is steam distilled to remove residual solvent, and the product is salted out, filtered, and washed with brine to give sodium naphthionate as a paste in 90 % yield [7].

Uses. Naphthionic acid is used as an intermediate in the production of 1-hydroxynaphthalene-4-sulfonic acid, 1-aminonaphthalene-2-sulfonic acid, and 1-aminonaphthalene-4,6- and 4,7-disulfonic acids. Use as a diazo component has been previously described, e.g., C.I. Food Red 3 (**71**), C.I. Food Red 7 (**86**), C.I. Food Red 9 (**84**), and C.I. Acid Red 25 (**77**).

1-Aminonaphthalene-5-sulfonic acid [*84-89-9*] (**104**) (5-aminonaphthalene-1-sulfonic acid), Laurent's acid, L acid, Purpurin acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes from hot water as needles of the monohydrate. The alkali-metal salts are all readily soluble in water.

Bromination yields the 2,4-dibromo derivative, vigorous nitration gives 2,4-dinitro-1-hydroxynaphthalene, and sulfonation results in 1-aminonaphthalene-2,5,7-trisulfonic acid. 1-Hydroxynaphthalene-5-sulfonic acid is obtained by heating **104** with aqueous bisulfite (reverse-Bucherer reaction) or via the diazo compound. Caustic fusion produces 5-amino-1-hydroxynaphthalene (Chap. 7). Coupling with diazo compounds takes place in the 2-position.

Production. Laurent's acid is coproduced with Peri acid as the minor component of a 5:2 isolated product ratio.

Uses. Laurent's acid is used as an intermediate for 1-hydroxynaphthalene-5-sulfonic acid, 5-amino-1-naphthol, 1-aminonaphthalene-5,7-disulfonic acid, and 8-amino-4-hydroxynaphthalene-2-sulfonic acid.

1-Aminonaphthalene-6-sulfonic acid [*119-79-9*] (**105**) (5-aminonaphthalene-2-sulfonic acid), 1,6-Cleve's acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes from boiling water as plates of the dihydrate. Its sodium, potassium, and calcium salts are all readily soluble in water, but the barium salt is only sparingly soluble. Sulfonation with 10 % oleum yields 1-aminonaphthalene-4,6-disulfonic acid. Caustic fusion gives 5-amino-2-hydroxynaphthalene, and Bucherer-type hydrolysis gives 1-hydroxynaphthalene-6-sulfonic acid.

Production. Sulfonation of 1-naphthylamine with concentrated sulfuric acid at 130 °C gives mainly 1-aminonaphthalene-6-sulfonic acid. However, the latter is always made, together with the 1,7-isomer, by nitration of naphthalene-2-sulfonic acid to give the mixed 5- and 8-nitronaphthalene-2-sulfonic acids follwed by reduction to "mixed Cleve's acids." A β-sulfonation mass (Section 2.2) is diluted with sulfuric acid, then 67.5 % nitric acid is added slowly at 34 °C. After being quenched and neutralized with calcium carbonate, the filtered liquor is reduced by boiling with iron powder, slight acidity being maintained with sulfuric acid. After basification with magnesia the iron residues are filtered off and the filtrate is evaporated until the magnesium salt of the product crystallizes on cooling. Filtration followed by washing with magnesium sulfate gives a 34 % yield of the magnesium salt of 1,6-Cleve's acid. Purification by dissolution in hot water, acidification, and filtration of the free acid while hot reduce the yield to 25 % based on naphthalene.

1-Aminonaphthalene-7-sulfonic acid [*119-28-8*] (**113**) (8-aminonaphthalene-2-sulfonic acid), 1,7-Cleve's acid, $C_{10}H_9NO_3S$, M_r 223.24, is 0.5 % soluble in water at 25 °C and

3 % soluble at 100 °C. The sodium, potassium, calcium, and magnesium salts are readily soluble in water but the barium salt is only sparingly soluble. Sulfonation with oleum at 50 °C gives 1-aminonaphthalene-4,7-disulfonic acid. Alkali fusion gives 8-amino-2-naphthol, and Bucherer-type hydrolysis gives 1-hydroxynaphthalene-7-sulfonic acid.

Production. The filtrate from the production of 1,6-Cleve's acid is acidified to precipitate crude 1,7-Cleve's acid (34 % yield), which can be purified by recrystallization of the sodium salt in an overall yield of 29 % based on naphthalene.

Uses. For many azo dye outlets, isolated mixed Cleve's acid from acidification of the reduction liquor is adequate as an economical component in the production, for example, of C.I. Acid Black 36 (**131**).

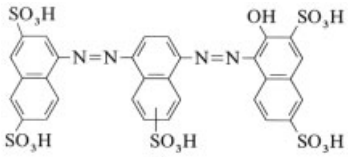

131

1-Aminonaphthalene-8-sulfonic acid [*82-75-7*] (**112**) (8-aminonaphthalene-1-sulfonic acid), Peri acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes as the monohydrate and is sparingly soluble even in boiling water. The sodium salt is also sparingly soluble in water.

Caustic fusion at 200 °C produces 8-amino-1-naphthol, whereas slow hydrolysis with 10 % sodium hydroxide at 220 – 260 °C yields 1,8-dihydroxynaphthalene. Sulfonation with cold 10 % oleum produces 1-aminonaphthalene-4,8-disulfonic acid, whereas hot concentrated sulfuric acid results in slow isomerization into naphthionic acid.

Production [7]. Naphthalene is converted to naphthalene-1-sulfonic acid (see Section 2.2). The crude sulfonation product is diluted with sulfuric acid, and 67.5 % nitric acid is added slowly at 30 °C to give the 5- and 8-nitro derivatives. The products are converted to the readily soluble magnesium salts by addition of a slurry of dolomite. The reaction mixture is neutralized with chalk and filtered hot. The filtrate is run into a suspension of iron powder in boiling water, which is kept slightly acidic with sulfuric acid. When reduction is complete, dissolved iron is precipitated by addition of magnesia, the iron residues are filtered off while hot, and the resulting liquor is acidified to pH 4.5 with sulfuric acid to give 99 % pure Peri acid in 48 % yield. The filtrates are further acidified to precipitate crude Laurent's acid. This may be purified via its magnesium salt to give a 17 % yield of pure 1,5-isomer.

Uses. Dehydration with phosphorus oxychloride yields naphthosultam (**132**), and diazotization followed by hydrolysis results in naphthosultone (**133**), the intermediate for 1-hydroxynaphthalene-8-sulfonic acid. Diazotization followed by a Sandmeyer cyanide reaction gives 8-cyanonaphthalene-1-sulfonic acid, one of the most industrially important Sandmeyer reaction products because the nitrile forms the lactam, naphthostyril [*130-00-7*] (**134**), upon treatment with concentrated alkali at 185 °C. Hydrolysis of naphthostyril with dilute alkali yields 8-aminonaphthalene-1-carboxylic acid (**135**), which is the key intermediate for production of anthanthrone vat dyes via 1,1′-binaphthyl-8,8′-dicarboxylic acid (**136**).

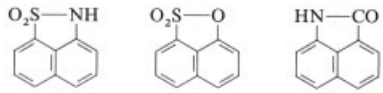

132 **133** **134**

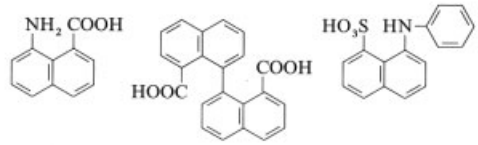

135 **136** **137**

Although Peri acid couples in the 4-position to form azo dyes, its N-aryl derivatives [e.g., *N*-phenyl (**137**) or *N*-*p*-tolyl] are generally used. These are produced by heating Peri acid with excess amine at 170 °C in the presence of amine hydrochloride. A typical and important acid dye for wool based on **137** is C.I. Acid Blue 113 (**92**). The monoazo compound derived from H acid (C.I. Acid Blue 92) is purified to anazolene sodium (**138**) for diagnostic use as a plasma volume indicator.

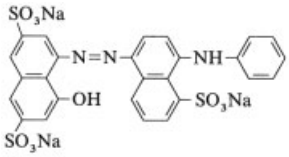

138

6.3. 2-Aminonaphthalenesulfonic Acids

2-Aminonaphthalene-1-sulfonic acid [*81-16-3*] (**119**), Tobias acid, $C_{10}H_9NO_3S$, M_r 223.24, is sparingly soluble in water but the sodium salt is readily soluble.

Coupling with diazo compounds takes place in the 1-position with elimination of the sulfonic acid group.

Production. 2-Hydroxynaphthalene-1-sulfonic acid (Section 4.3) is heated with a mixture of ammonia and ammonium sulfite in an autoclave at 150 °C and 1 MPa for 30 h [7]. The reaction mixture is added to a slurry of lime, and excess ammonia is driven off. Inorganic calcium salts are filtered off at 80 °C, and the filtrate is acidified with hydrochloric acid at 40 °C to precipitate Tobias acid in 95 % yield. Care must be taken throughout the process to avoid desulfonation and formation of 2-naphthylamine.

Uses. Sulfonation with oleum gives 2-aminonaphthalene-1,5-disulfonic acid, which can be desulfonated to produce 2-aminonaphthalene-5-sulfonic acid.

Tobias acid is used as a diazo component in acid dyes (e.g., C.I. Acid Yellow 19), red reactive dyes, and pigments such as the important Lithol Reds, [e.g., C.I. Pigment Red 49 (**139**)]. The free acid of **139** is also remarkably insoluble and can be used as a pigment.

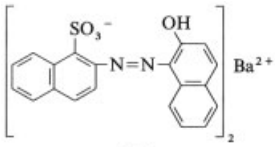

139

2-Aminonaphthalene-5-sulfonic acid [*81-05-0*] (**127**) (6-aminonaphthalene-1-sulfonic acid), Dahl's acid, Dressel acid, D acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes from water as needles although only 0.4 % soluble at 100 °C. The sodium, calcium, and barium salts are moderately soluble in water.

Bromination in acetic acid gives the 1-bromo derivative. Sulfonation with oleum gives a mixture of 2-aminonaphthalene-5,7-disulfonic acid together with some 1,5-disulfonic acid. Fusion with potassium hydroxide at 260 °C yields 6-amino-1-naphthol.

Production. Tobias acid is added to a mixture of sulfuric acid and oleum, and the reaction is completed by stirring for 12 h at 25 °C. The resulting sulfonation mass, containing mainly 2-aminonaphthalene-1,5-disulfonic acid (**126**), is poured into water and heated at 105 °C for 3 h to desulfonate and precipitate the product which, after cooling, is filtered off; the yield is 95 %.

2-Aminonaphthalene-6-sulfonic acid [*93-00-5*] (**120**) (6-aminonaphthalene-2-sulfonic acid), Bronner acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes from water as the monohydrate in plates although only 0.2 % soluble at 100 °C. Sulfonation with 20 % oleum at 20 °C gives a mixture of 1,6- and 6,8-disulfonic acids.

Production. 2-Hydroxynaphthalene-6-sulfonic acid (Section 4.3) is aminated by a Bucherer reaction under conditions similar to those used for Tobias acid.

2-Aminonaphthalene-7-sulfonic acid [*494-44-0*] (**121**) (7-aminonaphthalene-2-sulfonic acid), amido F acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes as the monohydrate in needles or prisms although only 0.3 % soluble at 100 °C.

Bromination gives the 1-bromo derivative, caustic fusion results in 7-amino-2-naphthol, and sulfonation with oleum gives a mixture of disulfonic acids.

Production. 2-Hydroxynaphthalene-7-sulfonic acid (Section 4.3) is aminated by a Bucherer reaction under conditions similar to those used for Tobias acid.

2-Aminonaphthalene-8-sulfonic acid [*86-60-2*] (**122**) (7-aminonaphthalene-1-sulfonic acid), Badische acid, $C_{10}H_9NO_3S$, M_r 223.24, crystallizes from water as needles or prisms although only 0.5 % soluble at 100 °C. Alkali-metal salts are quite soluble in water. Caustic fusion yields 7-amino-1-naphthol, sulfonation in the cold with 20 % oleum gives 2-aminonaphthalene-6,8-disulfonic acid, but nitration gives a mixture of products. 2-Aminonaphthalene-8-sulfonic acid reacts with diazo compounds to

form diazoamino compounds rather than coupling to give azo derivatives.

Production. 2-Hydroxynaphthalene-8-sulfonic acid (Section 4.3) is aminated by a Bucherer reaction under conditions similar to those used for Tobias acid.

6.4. 1-Aminonaphthalenedisulfonic Acids

Although preparation of fourteen of the possible isomers has been reported [5], only processes for those seven isomers that have been used as intermediates or dye components are described here.

1-Aminonaphthalene-3,6-sulfonic acid [*6251-07-6*] (**114**) (4-aminonaphthalene-2,7-disulfonic acid), Freund's acid(1,3,6), $C_{10}H_9NO_6S_2$, M_r 303.3, and its salts are readily soluble in water. Caustic fusion results in a mixture of aminonaphtholsulfonic acids, but amination occurs selectively, producing 1,3-diaminonaphthalene-6-sulfonic acid.

Production. Naphthalene is sulfonated with sulfuric acid at 160 °C for 14 h to give a mixture of mainly 2,6- and 2,7-disulfonic acids [7]. The reaction mass is nitrated with H_2SO_4–HNO_3 at 30 °C, poured into water, and limed out. The filtrate is then reduced with iron and hydrochloric acid, and the reduction liquor is worked up to selectively salt out the sodium salt of 1-aminonaphthalene-3,7-disulfonic acid in 17.5 % yield. The filtrates and brine wash are then acidified to precipitate crude 1-aminonaphthalene-3,6-disulfonic acid, which is purified by dissolution in aqueous potassium hydroxide and precipitation with acid; the yield is 35 %.

1-Aminonaphthalene-3,7-disulfonic acid [*6362-05-6*] (**115**) (4-aminonaphthalene-2,6-disulfonic acid), Freund's acid(1,3,7), $C_{10}H_9NO_6S_2$, M_r 303.3, is readily soluble in water, but its salts are less soluble than those of the 1,3,6-isomer. The amino group is hydrolyzed by water at 180 °C to give 1-hydroxynaphthalene-3,7-disulfonic acid. Caustic fusion results in formation of 1-amino-7-hydroxynaphthalene-3-sulfonic acid, and amination gives 1,3-diaminonaphthalene-7-sulfonic acid.

Production. The 1,3,7-acid is produced together with the 1,3,6-isomer [7].

1-Aminonaphthalene-3,8-disulfonic acid [*129-91-9*] (**116**) (8-aminonaphthalene-1,6-disulfonic acid), amino epsilon acid, $C_{10}H_9NO_6S_2$, M_r 303.3, and its sodium salt are very soluble in water, but the barium salt is only sparingly soluble. Reduction with zinc and alkali (or with sodium amalgam) desulfonates **116** to give 1-aminonaphthalene-3-sulfonic acid. The amino group is hydrolyzed by water at 180 °C to give 1-hydroxynaphthalene-3,8-disulfonic acid. Caustic fusion results in 1-amino-8-hydroxynaphthalene-3-sulfonic acid, whereas reaction with 10 % sodium hydroxide at 250 °C produces 1,8-dihydroxynaphthalene-3-sulfonic acid. Amination produces 1,3-diaminonaphthalene-8-sulfonic acid (**149**), but the production of 6,8-di(phenylamino)naphthalene-1-sulfonic acid (diphenyl epsilon acid; **140**) by reaction with aniline is more important.

Production. Naphthalene is sulfonated under conditions that give predominately the 1,6-disulfonic acid. The sulfonation mass is nitrated with H_2SO_4–HNO_3 at 30–40 °C [7]. After addition of dolomite and neutralization with lime, the liquor is reduced with iron and basified with magnesia, and the iron residues are removed. Careful adjustment of temperature and acidity precipitates the product as the acid magnesium salt, which is purified by dissolution in aqueous sodium carbonate and removal of magnesium carbonate; the diacid is precipitated with acid in 32 % overall yield.

Uses. Amino ε acid is used mainly as an intermediate rather than as a color component, but the derived diphenyl ε acid (**140**) is used for safranine dyes such as C.I. Acid Blue 61 (**141**).

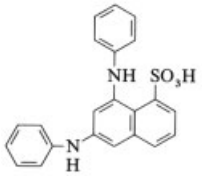

140

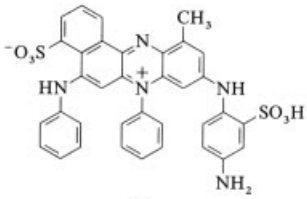

141

1-Aminonaphthalene-4,6-disulfonic acid [*85-74-5*] (**108**) (4-aminonaphthalene-1,7-disulfonic acid), Dahl's acid II, $C_{10}H_9NO_6S_2$, M_r 303.3, and its alkali-metal salts are soluble in water; the calcium salt is soluble in aqueous ethanol. Caustic fusion at 180 °C yields 1-amino-6-hydroxynaphthalene-4-sulfonic acid, but at higher temperature 4,7-dihydroxynaphthalene-1-sulfonic acid is formed. Sulfonation with 35 % oleum at 80 °C gives 1-aminonaphthalene-2,4,6-trisulfonic acid.

Production. Naphthionic acid is sulfonated with 25 % oleum at 20 °C to give a 70 : 30 mixture of 1,4,7- and 1,4,6-trisulfonic acids, which is used as mixed Dahl's acids.

1-Aminonaphthalene-4,7-disulfonic acid [*85-75-6*] (**106**) (4-aminonaphthalene-1,6-disulfonic acid), Dahl's acid III, $C_{10}H_9NO_6S_2$, M_r 303.3, and its sodium salt are readily soluble in water, but the calcium and barium salts are only sparingly soluble. Conversion to 1-hydroxynaphthalene-4,7-disulfonic acid is achieved by reverse-Bucherer reaction or diazotization and hydrolysis. Sulfonation (35 % oleum, 90 °C) yields 1-aminonaphthalene-2,4,7-trisulfonic acid.

Production. Although separation of **106** from the mixture obtained by sulfonation of naphthionic acid (see above) is not practicable on a large scale, the 1,4,7-isomer can be obtained directly in high yield by sulfonation of 1-aminonaphthalene-7-sulfonic acid with oleum.

1-Aminonaphthalene-4,8-disulfonic acid [*117-55-5*] (**117**) (4-aminonaphthalene-1,5-disulfonic acid), $C_{10}H_9NO_6S_2$, M_r 303.3, and its acid monosodium salt are sparingly soluble in cold water, but the disodium salt is readily soluble. Hydrolysis with sodium hydroxide at 200 °C yields 1-amino-8-hydroxynaphthalene-4-sulfonic acid; at higher temperature, 1,8-dihydroxynaphthalene-4-sulfonic acid is formed. Heating with bisulfite produces 1-hydroxynaphthalene-4,8-disulfonic acid. Diazo compounds couple with 1-aminonaphthalene-4,8-disulfonic acid to form azo compounds, in contrast to 2-aminonaphthalene-4,8-disulfonic acid, which reacts to form diazoamino compounds.

Production. Naphthalene is sulfonated under conditions that give predominantly the 1,5-disulfonic acid, and the sulfonation mass is nitrated at 20 °C [7]. 3-Nitronaphthalene-1,5-disulfonic acid is precipitated from the quenched reaction mixture as its iron(II) salt and removed. After dilution and liming out, the remainder of the mixed nitro solution is reduced with iron and dilute acid, and the 1,4,8-isomer is salted out preferentially in ca. 40 % yield. A small proportion (ca. 10 % of 1-aminonaphthalene-3,8-disulfonic acid) is then salted out, after which the more soluble 1-aminonaphthalene-5,7-disulfonic acid crystallizes on concentration.

1-Aminonaphthalene-5,7-disulfonic acid [*13306-42-8*] (**130**) (5-aminonaphthalene-1,3-disulfonic acid), $C_{10}H_9NO_6S_2$, M_r 303.3, and its sodium salt are very soluble in water. Caustic fusion produces 1-amino-5-hydroxynaphthalene-7-sulfonic acid, and sulfonation (40 % oleum, 130 °C) yields 1-aminonaphthalene-2,5,7-trisulfonic acid.

Production. 1-Aminonaphthalene-5-sulfonic acid (Section 6.2) in sulfuric acid solution is treated with 65 % oleum, and the mixture is heated at 120 °C for 4 h. The sulfonation mass, which contains mainly the 2,5,7-trisulfonic acid, is diluted with water, and the batch is heated at 130 °C for 10 h to cause desulfonation. The product crystallizes on cooling and, after isolation, is deacidified by treatment of the solution with lime. The final sodium salt solution is used directly for production of 1-amino-5-hydroxynaphthalene-7-sulfonic acid.

6.5. 2-Aminonaphthalenedisulfonic Acids

Although preparation of ten of the possible isomers has been reported [5], only processes for those five isomers that have been used as intermediates or dye components are described here.

2-Aminonaphthalene-1,5-disulfonic acid [*117-62-4*] (**126**), 5-sulfo-Tobias acid, $C_{10}H_9NO_6S_2$, M_r 303.32, is purified by crystallization of the dipotassium salt from water. Alkali fusion results in formation of 2-amino-5-hydroxynaphthalene-1-sulfonic acid. Sulfonation (40 % oleum, 100 °C) gives 2-aminonaphthalene-1,5,7-trisulfonic acid. Heating with aqueous sulfuric acid produces 2-aminonaphthalene-5-sulfonic acid. The diazo compound is

converted to 2-diazo-1-hydroxynaphthalene-5-sulfonic acid on treatment with alkaline hypochlorite.

Production. Sulfonation of Tobias acid to produce 2-aminonaphthalene-5-sulfonic acid (Section 6.3) proceeds via **126**, which is isolated from the quenched reaction mixture with careful avoidance of desulfonation.

Uses. As a diazo component in orange and red azo reactive dyes, **126** gives more bathochromic shades than Tobias acid.

2-Aminonaphthalene-3,6-disulfonic acid [*92-28-4*] (**123**) (3-aminonaphthalene-2,7-disulfonic acid), amino R acid, $C_{10}H_9NO_6S_2$, M_r 303.3.

Production. The sodium salt of 2-hydroxynaphthalene-3,6-disulfonic acid (Section 4.5) is aminated by heating with ammonia and ammonium bisulfite for 24 h at 160 °C. The charge is added to excess lime and heated to remove excess ammonia, and calcium sulfite is filtered off. The product is precipitated by acidification of the filtrate and purified by redissolving the paste in alkali and reprecipitating with acid. The yield is 90 %.

Uses. Amino R acid is used mainly as an intermediate for conversion to 2-amino-3-hydroxynaphthalene-6-sulfonic acid by caustic fusion at 250 °C.

2-Aminonaphthalene-4,8-disulfonic acid [*131-27-1*] (**118**) (3-aminonaphthalene-1,5-disulfonic acid), Cassella acid, C acid, $C_{10}H_9NO_6S_2$, M_r 303.32, and its salts are readily soluble in water.

Reduction with sodium amalgam desulfonates **118** to give 2-aminonaphthalene-8-sulfonic acid, whereas boiling with dilute sulfuric acid results in desulfonation and hydrolysis to give 2-hydroxynaphthalene-4-sulfonic acid. Alkali fusion (KOH, 215 °C) produces 2-amino-4-hydroxynaphthalene-8-sulfonic acid.

Production. The iron(II) salt of 3-nitronaphthalene-1,5-disulfonic acid (nitro-Armstrong acid) (Section 6.4) is reduced with iron in dilute acid [7].

A claimed improvement involves isolation of nitro-Armstrong acid from the nitration products of naphthalene-1,5-disulfonic acid as the crystalline magnesium salt prior to catalytic reduction to C acid [49].

Uses. Cassella acid is an important diazo component for yellow direct and reactive dyes.

2-Aminonaphthalene-5,7-disulfonic acid [*118-33-2*] (**129**) (6-aminonaphthalene-1,3-disulfonic acid), amino J acid, $C_{10}H_9NO_6S_2$, M_r 303.32, and the bis-alkali-metal salts are very soluble in water, but the mono-alkali-metal salts, especially the acid potassium salt, are sparingly soluble in cold water. Reduction with sodium amalgam produces 2-aminonaphthalene-7-sulfonic acid, and heating with ammonium bisulfite gives 2-hydroxynaphthalene-5,7-disulfonic acid.

Production. 2-Aminonaphthalene-1-sulfonic acid is sulfonated by heating with excess oleum for 12 h at 100 °C. This produces 2-aminonaphthalene-1,5,7-trisulfonic acid (**128**), which is desulfonated by heating the quenched acid mixture at 105 °C for 3 h. The product is filtered off after cooling and dissolved in hot aqueous alkali, and excess sulfate is removed by precipitation with lime. The resulting sodium salt liquor (80 % yield) is used directly for caustic fusion to give J acid, the major outlet.

2-Aminonaphthalene-6,8-disulfonic acid [*86-65-7*] (**124**) (7-aminonaphthalene-1,3-disulfonic acid), amino G acid, $C_{10}H_9NO_6S_2$, M_r 303.3, and its bis-alkali-metal salts are very soluble in water, but the acid salts, especially the acid potassium salt, are much less soluble.

Production. The disodium or dipotassium salt of 2-hydroxynaphthalene-6,8-disulfonic acid (Section 4.5) is heated with excess aqueous ammonia and ammonium bisulfite at 185 °C for 18 h. The reaction mixture is basified, heated to remove excess ammonia, neutralized, and evaporated while any residual sulfite is removed. The resulting amino G acid liquor (95 % yield) is used directly for caustic fusion to give γ acid, the major outlet.

6.6. Aminonaphthalenetrisulfonic Acids

Only those isomers that have been used as intermediates or dye components are described here. Many other isomers have been reported, as have aminonaphthalenetetrasulfonic acids [5], but they are of only academic interest. The major outlet for aminonaphthalenetrisulfonic acids

is in caustic fusion reactions to produce important aminohydroxynaphthalenedisulfonic acids.

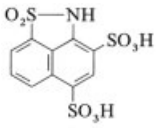

142

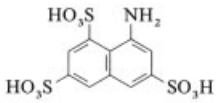

143

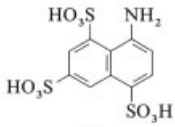

144

1-Aminonaphthalene-2,4,8-trisulfonic acid [*76530-15-9*] (4-aminonaphthalene-1,3,5-trisulfonic acid), $C_{10}H_9NO_9S_3$, M_r 383.36, is not isolable as the free acid and always reverts to the sultam (**142**) with elimination of water between the peri-substituents. The trisodium salt is also water soluble but can be crystallized in the form of yellow plates. Desulfonation occurs with 20 % hydrochloric acid at 150 °C to form 1-aminonaphthalene-8-sulfonic acid and with 40 % sulfuric acid at 110 °C to give 1-aminonaphthalene-2,8-disulfonic acid.

Production. A sulfuric acid solution of 1-aminonaphthalene-8-sulfonic acid (Section 6.2) is treated with 65 % oleum in the presence of anhydrous sodium sulfate as a moderator, and the temperature is slowly increased to 85 °C. The quenched solution is neutralized with chalk, calcium sulfate is removed, and the calcium salt solution is basified with sodium carbonate. After removal of calcium carbonate, the resulting solution of the disodium salt is evaporated for subsequent caustic fusion to produce Chicago acid.

1-Aminonaphthalene-3,6,8-trisulfonic acid [*117-42-0*] (**143**) (8-aminonaphthalene-1,3,6-trisulfonic acid), Koch acid, T acid, $C_{10}H_9NO_9S_3$, M_r 383.36, and its alkali-metal salts are readily soluble in water. The solutions are unusual for this class of compound in that they exhibit no fluorescence. Hydrolysis with dilute sulfuric acid produces 1-hydroxynaphthalene-3,6,8-trisulfonic acid (Section 4.6), whereas diazotization followed by acid hydrolysis results in formation of the corresponding sultone. Reduction with zinc and alkali desulfonates **143** to give 1-amino-naphthalene-3,6-disulfonic acid. Amination gives 1,3-diaminonaphthalene-6,8-disulfonic acid.

Production. Naphthalene in sulfuric acid solution is treated with gradual addition of 65 % oleum and temperature profiling (to 160 °C) to optimize formation of the 1,3,6-trisulfonic acid. A mixture of sulfuric acid and nitric acids is added slowly at 40 °C to form 1-nitronaphthalene-3,6,8-trisulfonic acid (nitro-Koch acid) which is converted to neutral liquor by liming out. The liquor is reduced with iron powder in the presence of acetic acid and, after removal of iron residues, the crude product is isolated, as the acid calcium – sodium salt. This is purified by dissolution with sodium carbonate and removal of calcium carbonate. The trisodium salt liquor is concentrated before direct use in H acid manufacture. To obtain pure Koch acid, the 1-nitro-3,6,8-trisulfonic acid can be isolated before reduction. The reduction step can alternatively be carried out by hydrogenation of the aqueous solution of nitro-Koch acid in the presence of a nickel catalyst [50].

Uses. Apart from caustic fusion to give H acid, the other main use is the production of chromotropic acid (**66**) by treatment with strong caustic under pressure.

1-Aminonaphthalene-4,6,8-trisulfonic acid [*17894-99-4*] (**144**) (8-aminonaphthalene-1,3,5-trisulfonic acid), B acid, $C_{10}H_9NO_9S_3$, M_r 383.36, is reduced with zinc and alkali to 1-aminonaphthalene-4,6-disulfonic acid. Heating with aqueous sulfuric acid causes desulfonation to form 1-aminonaphthalene-6,8-disulfonic acid. Sulfonation (25 % oleum, 90 °C) gives the sultam of 1-aminonaphthalene-2,4,6,8-tetrasulfonic acid.

Production. Naphthalene is sulfonated under conditions that maximize formation of the 1,3,5-trisulfonic acid (Section 2.4), followed by nitration with H_2SO_4–HNO_3 in situ at 35 °C. After quenching, neutralization with calcium carbonate, and reduction of the nitro-derivative with iron under acid conditions, the byproduct 1-aminonaphthalene-3,6,8-trisulfonic acid (18 % yield) is selectively precipitated by acidification. The desired isomer is then precipitated by further acidification and salting of the filtrates, and

finally purified by reprecipitation. Overall yield is 25 – 30 %.

Uses. Alkali fusion of B acid yields K acid. A special outlet for purified B acid is in the synthesis of the drug suramin sodium (Bayer 205) (**145**). 1-Aminonaphthalene-4,6,8-trisulfonic acid is acylated with 2-nitrotoluyl chloride, and this product is reduced and acylated with *m*-nitrobenzoyl chloride. This intermediate is reduced and finally phosgenated to give **145**, which has been used to treat trypanosomiasis and onchocerciasis.

145

2-Aminonaphthalene-1,5,7-trisulfonic acid [*55524-84-0*] (**128**) (6-aminonaphthalene-1,3,5-trisulfonic acid), $C_{10}H_9NO_9S_3$, M_r 383.36, and its salts are readily soluble in water. Alkali fusion (65 % NaOH, 220 °C) forms 2-amino-5-hydroxynaphthalene-1,7-disulfonic acid. Desulfonation to give 2-aminonaphthalene-5,7-disulfonic acid occurs on heating with dilute sulfuric acid. Diazotization and treatment with sodium carbonate yield the diazo oxide of 2-diazo-1-hydroxynaphthalene-5,7-disulfonic acid.

Production. Vigorous sulfonation of Tobias acid to produce 2-aminonaphthalene-5,7-disulfonic acid is described in Section 6.5. The intermediate **128** is obtained by quenching the sulfonation mass and liming out carefully to avoid desulfonation. The calcium salt solution is converted to the sodium salt with sodium carbonate and used as a total liquor for conversion to 2-amino-5-hydroxynaphthalene-1,7-disulfonic acid.

2-Aminonaphthalene-3,6,8-trisulfonic acid [*118-03-6*] (**146**) (7-aminonaphthalene-1,3,6-trisulfonic acid), $C_{10}H_9NO_9S_3$, M_r 383.36, and its disodium salts are soluble in water, but the dipotassium salt is much less soluble. Reduction with zinc dust and alkali results in 2-aminonaphthalene-3,6-disulfonic acid, and caustic fusion at 240 °C gives 2-amino-8-hydroxynaphthalene-3,6-disulfonic acid.

Production. 2-Aminonaphthalene-6,8-disulfonic acid (Section 6.5) is sulfonated with 65 % oleum at 125 °C for 12 h in the presence of anhydrous sodium sulfate. After quenching and liming out, the hot solution of calcium salt is basified with sodium carbonate and the resulting solution of the trisodium salt is used directly for caustic fusion to give 2-amino-8-hydroxynaphthalene-3,6-disulfonic acid.

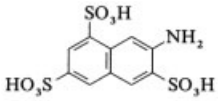

146

6.7. Diaminonaphthalenesulfonic Acids

4,5-Diaminonaphthalene-1-sulfonic acid [*6362-18-1*] (**147**) 1,8-diaminonaphthalene-4-sulfonic acid, $C_{10}H_{10}N_2O_3S$, M_r 238.27.

Production. 1-Nitronaphthalene (Section 5.1) is sulfonated with oleum at 20 °C, and the sulfonation mass is nitrated at 35 °C [7]. After quenching and liming out, the liquors are reduced with iron powder under slightly acidic conditions. After removal of iron residues, the product **147** is isolated in 40 % yield by acidifying and salting.

Uses. 4,5-Diaminonaphthalene-1-sulfonic acid is used as an intermediate for 4,5-dihydroxynaphthalene-1-sulfonic acid (Section 4.7) and perimidine derivatives.

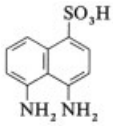

147

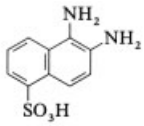

148

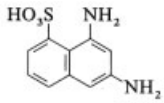

149

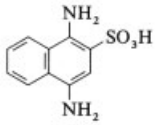

150

5,6-Diaminonaphthalene-1-sulfonic acid [*84-92-4*] (**148**) 1,2-diaminonaphthalene-5-sulfonic acid, $C_{10}H_{10}N_2O_3S$, M_r 238.27.

Production. 1-Aminonaphthalene-5-sulfonic acid (Section 6.2) is coupled with diazotized aniline, and the resulting isolated azo compound (purpurin base) is reduced by heating with zinc and ammonia at 100 °C and allowing ammonia and aniline to distill off [7]. The product (**148**) is isolated in 76 % yield by acidification of the screened liquors.

Uses. 5,6-Diaminonaphthalene-1-sulfonic acid is used as an intermediate for naphthotriazoles. A green pigment (C.I. Pigment Green 9) is obtained by treatment with iron(III) salts.

6,8-Diaminonaphthalene-1-sulfonic acid (**149**) 1,3-diaminonaphthalene-8-sulfonic acid, M_r 238.27, is prepared by amination of 1-aminonaphthalene-3,8-disulfonic acid (Section 6.4) or 1-hydroxynaphthalene-3,8-disulfonic acid but has no significant use.

6,8-Dianilinonaphthalene-1-sulfonic acid [*129-93-1*] (**140**), diphenyl epsilon acid, $C_{22}H_{18}N_2O_3S$, M_r 390.38.

Production. 1-Aminonaphthalene-3,8-disulfonic acid (Section 6.4) is heated with excess aniline in the presence of aniline sulfate at 130 °C.

Uses. A dye component as exemplified in **141**.

1,4-Diaminonaphthalene-2-sulfonic acid [*6357-89-7*] (**150**), $C_{10}H_{10}N_2O_3S$, M_r 238.27.

Production. 1-Aminonaphthalene-2-sulfonic acid (Section 6.2) is coupled with diazotized aniline, and the resulting azo compound is reduced in a process analogous to that used for **148**.

6.8. Diaminonaphthalenedisulfonic Acids

3,8-Diaminonaphthalene-1,5-disulfonic acid [*19659-81-5*] (**151**), amino Cacid, $C_{10}H_{10}N_2O_6S_2$, M_r 318.32.

Production. Naphthalene-1,5-disulfonic acid sulfonation mass is nitrated to give the 3,8-dinitro derivative [7]. The quenched and limed-out liquor is reduced with iron to give **151**.

4,8-Diaminonaphthalene-2,6-disulfonic acid [*6362-06-7*] (**152**), $C_{10}H_{10}N_2O_6S_2$, M_r 318.32.

Production. Naphthalene-2,6-disulfonic acid is dinitrated and reduced in a similar way to the 1,5-isomer [7].

4,5-Diaminonaphthalene-2,7-sulfonic acid [*6362-11-4*] (**153**), Alen acid, $C_{10}H_{10}N_2O_6S_2$, M_r 318.32.

Production. Naphthalene-2,7-disulfonic acid is dinitrated and reduced in a similar way to the 1,5-isomer [7].

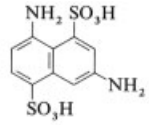

151

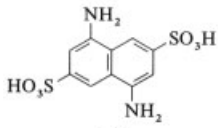

152

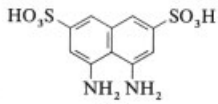

153

6.9. Toxicity

Aminonaphthalenesulfonic acids were once regarded as innocuous compared to the parent amines. Later, however, some risk was recognized to be involved [9]. Although several aminonaphthalene mono- and disulfonic acids are listed in the EPA inventory [14] as having mildly toxic and irritant properties, no evidence of carcinogenic properties exists.

The hypothesis that water-soluble agents are less likely to be absorbed by bodily organs is supported by the full and rapid excretion of ^{35}S-labeled Tobias acid in experiments with rats [51]. Neither 2-naphthylamine nor its derivatives were detected in these experiments, indicating that, given the short dwell time of water-soluble compounds in the body, desulfonation is unlikely.

7. Aminonaphthols

Although all 14 isomeric aminonaphthols have been described [5], they are less useful than their sulfonic acids (Chap. 8) and only those of commercial significance will be identified. As a class they are amphoteric and dissolve in both acid and alkali. One reason for their lack of use is that their alkaline solutions are very readily oxidized.

There are four main methods of preparation:

1) Caustic fusion of aminonaphthalenesulfonic acids as previously mentioned under properties of 1-aminonaphthalene-5-sulfonic acid, 1-aminonaphthalene-6-sulfonic acid, 1-aminonaphthalene-7-sulfonic acid, and 1-aminonaphthalene-8-sulfonic acid (Section 6.2), as well as 2-aminonaphthalene-5-sulfonic acid, 2-aminonaphthalene-7-sulfonic acid, and 2-aminonaphthalene-8-sulfonic acid (Section 6.3). With some isomers (e.g., 1,4- and 2,6-), hydroxynaphthalenesulfonic acids are formed by replacement of the amino group rather than the sulfonic acid group.
2) Bucherer amination of dihydroxynaphthalenes with ammonia and bisulfite. Alternatively, the reverse-Bucherer reaction can be applied to certain naphthalenediamines.
3) Desulfonation of certain aminohydroxynaphthalenesulfonic acids by heating with dilute acid or by treatment with sodium amalgam.
4) Reduction of arylazo, nitroso, or nitro derivatives of 1-naphthol or 2-naphthol by zinc or sodium. A more recent process uses aqueous hydrazine as the reducing agent [52].

4-Amino-1-naphthol [*2834-90-4*] (**154**), 4-hydroxy-1-aminonaphthalene, 4-amino-1-naphthalenol, $C_{10}H_9NO$, M_r 159.2.

Production. Orange I (**25**) in aqueous alkaline solution is treated with sodium dithionite at 40 – 50 °C, and the resulting suspension is heated to 70 °C to complete the reaction. The crude product is collected by filtration and dissolved in hot aqueous hydrochloric acid in the presence of a reducing agent. On cooling, needles of the hydrochloride salt crystallize and can be isolated in 70 % yield. Alternatively, the crude paste may be dissolved in dilute alkali in the presence of a reducing agent and acylated with acetic anhydride or acryloyl chloride, for example, to give 4-acetylamino-1-naphthol or 4-acryloylamino-1-naphthol, which have been used as azo coupling components.

Uses. 4-Amino-1-naphthol is used as a polymerization inhibitor.

5-Amino-1-naphthol [*83-55-6*] (**155**), 5-hydroxy-1-aminonaphthalene, 5-amino-1-naphthalenol, Purpurol, $C_{10}H_9NO$, M_r 159.2.

Production. 1-Aminonaphthalene-5-sulfonic acid paste (Section 6.2) is added to molten sodium hydroxide at 200 °C, and the reaction is completed by heating slowly to 250 °C and maintaining this temperature for 3 h [7]. After dilution and careful neutralization with hydrochloric acid, 5-amino-1-naphthol precipitates in 85 % yield (*mp* 192 °C).

Uses. 5-Amino-1-naphthol is used as a coupling component for azo dyes (coupling takes place ortho to the amino or hydroxyl group, or para to the hydroxyl group depending on the pH), and as an intermediate for sulfur dyes.

7-Amino-1-naphthol [*4384-92-3*] (**156**), 8-hydroxy-2-aminonaphthalene, 7-amino-1-naphthalenol, $C_{10}H_9NO$, M_r 159.2.

Production. 2-Aminonaphthalene-8-sulfonic acid (Section 6.3) is heated with molten potassium hydroxide for 2 – 3 h at 260 °C in an autoclave. The product is obtained in 74 % yield after dilution and acidification of the melt.

1-Amino-1-naphthol [*2834-92-6*] (**157**), 1-amino-2-naphthalenol, is prepared by reduction of Orange II (1-*p*-sulfobenzeneazo-2-naphthol) by a process analogous to that described for 4-amino-1-naphthol, but it has no significant use.

1-Amino-2-ethoxynaphthalene [*118-30-9*] (**158**), $C_{12}H_{13}NO$, M_r 188.

Production. A solution of 2-ethoxynaphthalene (Section 3.2) in 1,2-dichlorobenzene is reacted with nitric acid at 45 °C. The organic layer is separated and washed at 60 °C before cooling to crystallize 1-nitro-2-ethoxynaphthalene (*mp* 95 – 100 °C) in 80 % yield. Reduction with iron gives **158** in 74 % yield.

Uses. Sulfonation (H_2SO_4, 60 °C) of **158** gives 1-amino-2-ethoxynaphthalene-6-sulfonic acid for use principally as an intermediate component in azo direct green dyes.

8-Amino-2-naphthol [*118-46-7*] (**159**), 1-amino-7-naphthol, 8-amino-2-naphthalenol, Cyanol, $C_{10}H_9NO$, M_r 159.2.

Production. The sodium salt of 1-aminonaphthalene-7-sulfonic acid (Section 6.2) is added to 80 % potassium hydroxide at 210 °C, and the reaction is completed by heating for 6 h at 230 °C [7]. After dilution and careful acidification with hydrochloric acid, the acid solution is screened to remove impurities and neutralized to precipitate **159** in 80 % yield (*mp* 207 °C).

Uses. Sulfonation gives 1-amino-7-hydroxynaphthalene-4-sulfonic acid. The *N*-acetyl derivative is an important coupling component for after-chrome and metal-complex dyes, e.g., C.I. Mordant Black 38 and C.I. Acid Black 60.

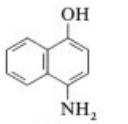

154

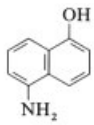

155

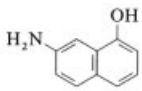

156

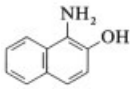

157

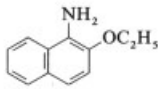

158

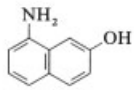

159

8. Aminohydroxynaphthalenesulfonic Acids

Many of the letter acids listed in Table 1 are aminohydroxynaphthalenesulfonic acids, (aminonaphtholsulfonic acids), the most important of which are probably the H, J, and γ acids [17]. Most of the large-scale use of this class involves the production of water-soluble azo dyes. Only those products of known use are dealt with on an individual basis. Registration of compounds in the European core inventory (ECOIN) in 1981 has been used as a guide to commercial applications. Thirteen aminonaphthol monosulfonic acids and seven aminonaphthol disulfonic acids are listed therein.

Nomenclature is used as in Table 1, based on tradition (and current common usage) with precedence given to the amino and then the hydroxyl group. The IUPAC convention, based on derivatives of naphthalenesulfonic acids, is added in parentheses for completion.

A number of aminohydroxynaphthalenetrisulfonic acids and diaminohydroxynaphthalenesulfonic acids have been prepared [5] by routes analogous to these described here but they have no technical advantage over the simpler derivatives.

8.1. Production and Properties

Four main routes are available depending on the substitution pattern required (route 1 predominates as the basis for all major aminonaphthol sulfonic acids).

1) Caustic fusion hydrolysis of aminonaphthalenedi- and trisulfonic acids.
   One sulfonic acid group of aminonaphthalenedi- and trisulfonic acids (see Chap. 6) can be hydrolyzed in a way similar to the selective hydrolysis described for naphthalenedi- and trisulfonic acids (Chap. 4).
   The following ten derivatives (**160–169**) manufactured by this route are all commercially important and are described individually later. With α, β-disulfonic acids, the α-group is the more mobile; with α, α-configurations, reaction occurs with the sulfonic acid group in the ring not containing the amino group. Hydrolysis of a sulfonic acid group in the peri-position to the amino group is particularly facile.

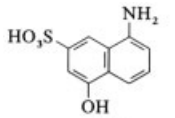

160

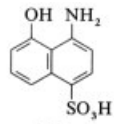

161

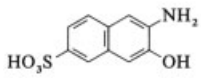

162

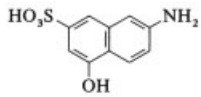

163

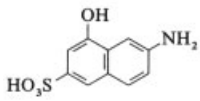

164

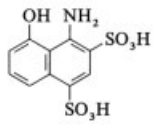

165

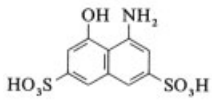

166

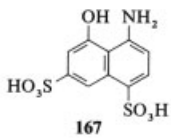

167

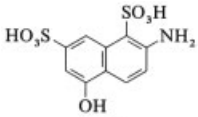

168

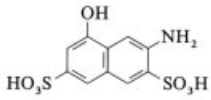

169

2) Reduction of nitroso or arylazo derivatives of hydroxynaphthalenesulfonic acids.
   Hydroxynaphthalenesulfonic acids react readily with nitrous acid or diazo compounds with the introduction of substituents ortho to the hydroxyl group. Reduction of the nitroso or azo substituent leads to an ortho arrangement of the amino and hydroxyl groups that is difficult to achieve by other routes.
   The nitrosation route can be extended to include 1-naphthol or 2-naphthol, because simultaneous reduction and sulfonation can be achieved to obtain, for example, 1-amino-2-hydroxynaphthalene-4-sulfonic acid (**170**) directly from 2-naphthol.
3) Sulfonation of aminonaphthols.
   When an aminonaphthol is readily accessible (Chap. 7), sulfonation may lead to useful aminohydroxynaphthalenesulfonic acids that are not otherwise readily obtainable. However, isomer formation and the need for separation may be a problem.
4) Selective desulfonation of aminohydroxynaphthalenedisulfonic acids.
   Facile removal of certain α-sulfonic acid groups may occasionally make an aminohydroxynaphthalenedisulfonic acid (e.g., K acid) a viable starting material.

Aminohydroxynaphthalenesulfonic acids resemble aminonaphthalenesulfonic acids in forming internal salts. The monosulfonic acids are therefore sparingly soluble in water and are isolated as the free acid. The disulfonic acids are much more soluble in water and are often isolated as the monosodium or monopotassium salts even under acidic conditions.

Coupling with diazo compounds is the most important property of aminohydroxynaphthalenesulfonic acids, and many of them couple under both acidic and alkaline conditions. The hues and properties of the resulting *o*-aminoazo or *o*-hydroxyazo dyes vary widely. The most useful products are derived from alkaline-coupled J acid (monoazo oranges) or H acid (monoazo reds), acid-coupled γ acid (monoazo reds), and acid- and alkali-twice-coupled H acid (disazo blues and blacks), even though all of these letter acids undergo both acid and alkaline coupling.

Although all aminohydroxynaphthalenesulfonic acids are diazotizable, only 1-amino-2-hydroxynaphthalene-4-sulfonic acid and its derivatives are industrially important as diazo components.

8.2. Aminohydroxynaphthalene-monosulfonic Acids

1-Amino-2-hydroxynaphthalene-4-sulfonic acid [*116-63-2*] (**170**) (4-amino-3-hydroxynaphthalene-1-sulfonic acid), Böniger acid, $C_{10}H_9NO_4S$, M_r 239.2, is obtained as white or gray needles of the hemihydrate, which is only sparingly soluble in hot water. The alkali-metal salts are soluble in water, but the solutions slowly oxidize in air. Oxidation with nitric acid gives 1,2-naphthoquinone-4-sulfonic acid.

Production. 2-Naphthol is dissolved in dilute aqueous sodium hydroxide and precipitated in finely divided form by addition of sulfuric acid at 5 °C [7]. Sodium nitrite solution is added to the slightly alkaline suspension, followed by sulfuric acid to effect the nitrosation while the temperature is held at 5 – 8 °C. The suspension of 1-nitroso-2-naphthol is neutralized with aqueous sodium hydroxide; a solution of sodium bisulfite is then added. The temperature is allowed to rise to 20 °C, and the soluble bisulfite addition complex (**171**) is formed. Excess sulfuric acid is then added and the temperature is raised to 40 °C, whereupon reductive rearrangement takes place and the product slowly crystallizes at 50 °C. Filtration and washing give the product as a 25 % paste in 80 % yield. All plant items and reactants must be free from iron to avoid complex formation.

Uses. The 1,2,4-acid paste is usually converted directly to the diazo oxide (**172**) by addition to excess sodium nitrite solution in the presence of copper(II) sulfate and in the absence of mineral acid [7]. After completion of the reaction, the product is precipitated by careful addition of excess sulfuric acid and isolated in 95 % yield as an orange-gray sandy paste. This can be dried if required for further nonaqueous processing at a temperature not exceeding 50 °C to avoid decomposition.

The so-called diazo-1,2,4-acid (**172**) is usually thought to exist as the naphthoquinone – diazide (**173**; 4-diazo-3,4-dihydro-3-oxonaphthalene-1-sulfonic acid) [26]. It is an important component in azo dyes for after-chroming, e.g., C.I. Mordant Black 3 (**27**), C.I. Mordant Black 17, and C.I. Mordant Red 7 (**174**), and for producing preformed chromium complex dyes such as C.I. Acid Violet 90 (the 2 : 1 chromium complex of **174**).

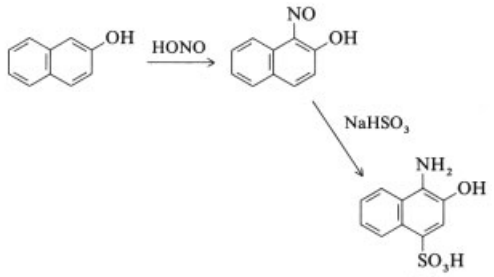

170

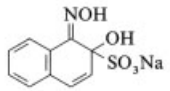

171

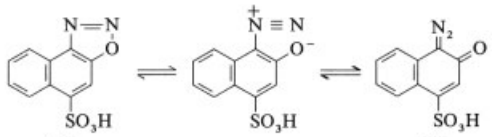

172 **173**

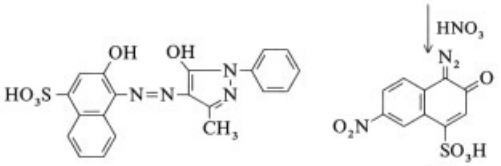

174 **175**

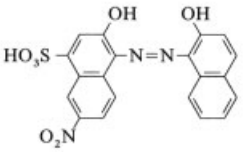

176

The quinone – diazide **173** [*4857-47-0*], is remarkably stable, which allows direct nitration in sulfuric acid solution by treatment with mixed acid over 8 h at 10 – 15 °C to give the so-called 6-nitrodiazo-1,2,4-acid [*5366-84-7*] (**175**) as a paste in 85 % yield after isolation and slurry washing. This nitroderivative is also considered to exist in the stable quinone – diazide form (4-diazo-3,4-dihydro-7-nitro-3-oxonaphthalene-1-sulfonic acid) and is an even more important component for metallized and metallizable dyes than is **173**; typical examples are C.I. Acid Black 52 (chromium complex of **176**), C.I. Direct Red 173, C.I. Reactive Black 1, C.I. Mordant Black 1 (**176**), and C.I. Mordant Black 11.

Chlorination of dry **173** in chlorosulfonic acid solution under a positive pressure of chlorine results in formation (50 % yield) of the diazo oxide of 6-chloro-1-amino-2-hydroxynaphthalene-4-sulfonic acid, which is, however, of lesser importance as a diazo component.

1-Amino-2-hydroxynaphthalene-6-sulfonic acid [*5639-34-9*] (**177**) (5-amino-6-hydroxynaphthalene-2-sulfonic acid), amino-Schaeffer acid, $C_{10}H_9NO_4S$, M_r 239.2, is obtained as needles or prisms that are only slightly soluble in boiling water. The sodium salt is soluble in water but can be isolated as the hemipentahydrate. Diazotization does not occur in mineral acid medium. The diazo oxide is formed by reaction with sodium nitrite in the presence of copper(II) sulfate.

Production. 2-Hydroxynaphthalene-6-sulfonic acid is nitrosated in aqueous solution, and the resulting nitroso derivative is reduced in situ by addition of zinc dust to the acidified batch, with the temperature allowed to rise to 40 °C [7]. The precipitated product is isolated in 70 % yield.

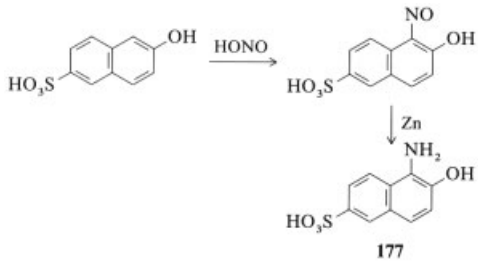

177

Uses. The sodium salt of **177** has found some use in the detection of potassium and as a photographic developer (Eikonogen). The *O*-ethyl derivative of **177** is an important middle component for dis- and trisazo green dyes.

1-Amino-2-ethoxynaphthalene-6-sulfonic acid [*118-28-5*] (**178**) can be obtained by ethylation of **177** with diethyl sulfate in methanol [53] or by sulfonation of 1-amino-2-ethoxynaphthalene (**158**; Chap. 7). The older, established route is based on Schaeffer acid and a three-step process involving ethylation, nitration, and reduction.

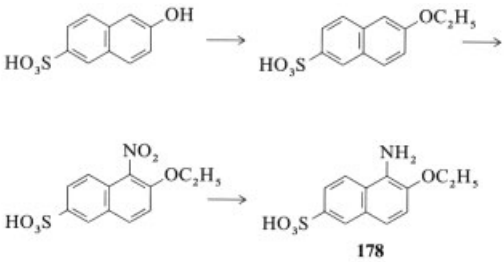

Important products derived from **178** are C.I. Direct Greens 23, 33 (**179**), 34, and 38.

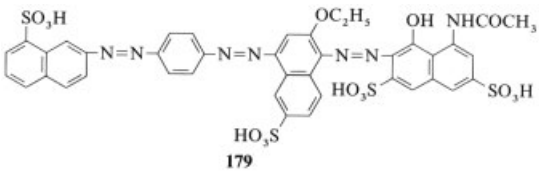

1-Amino-5-hydroxynaphthalene-6-sulfonic acid [*58596-07-9*] (**180**) (5-amino-1-hydroxynaphthalene-2-sulfonic acid), $C_{10}H_9NO_4S$, M_r 239.2.

Production. 5-amino-1-naphthol (Chap. 7) is sulfonated with 96 % sulfuric acid at 20 °C, and the sparingly soluble product is isolated after quenching.

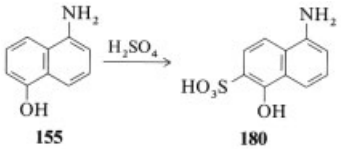

1-Amino-5-hydroxynaphthalene-7-sulfonic acid [*489-78-1*] (**160**) (8-amino-4-hydroxynaphthalene-2-sulfonic acid), M acid, $C_{10}H_9NO_4S$, M_r 239.2, and its sodium salt are sparingly soluble in water. Coupling with diazo compounds takes place under acidic and alkaline conditions, but a bisazo derivative cannot be obtained. Reaction with bisulfite gives 4,8-dihydroxynaphthalene-2-sulfonic acid.

Production. 1-Aminonaphthalene-5,7-disulfonic acid (Section 6.4) is heated with concentrated sodium hydroxide for 16 h at 172 – 176 °C. After the reaction mixture is quenched, acidified, and purged of sulfur dioxide, the product is isolated in 80 % yield.

1-Amino-8-hydroxynaphthalene-4-sulfonic acid [*83-64-7*] (**161**) (4-amino-5-hydroxynaphthalene-1-sulfonic acid), S acid, $C_{10}H_9NO_4S$, M_r 239.2, is only sparingly soluble in hot water. Reaction with bisulfite gives 4,5-dihydroxynaphthalene-1-sulfonic acid. Diazo compounds couple under acid conditions α to the amino group and under alkaline conditions α to the hydroxy group; bisazo derivatives can be obtained.

Production. 1-Aminonaphthalene-4,8-disulfonic acid (Section 6.4) is heated with concentrated potassium hydroxide at 200 – 230 °C.

Uses. The S acid and its acetyl derivative are used as azo coupling components in e.g., C.I. Mordant Green 14, C.I. Mordant Blue 7, and C.I. Direct Green 39.

1-Amino-8-hydroxynaphthalene-6-sulfonic acid [*35400-55-6*] (**181**) (5-amino-4-hydroxynaphthalene-2-sulfonic acid), $C_{10}H_9NO_4S$, M_r 239.2.

Production. 1-Aminonaphthalene-6,8-disulfonic acid gives **181** on fusion with 50 % potassium hydroxide at 200 °C, but because the former is not readily available, a more convenient process is the desulfonation of K acid (**167**) by boiling with zinc dust in aqueous sodium hydroxide solution.

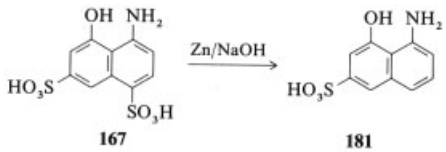

2-Amino-1-hydroxynaphthalene-4-sulfonic acid [*567-13-5*] (**182**) (3-amino-4-hydroxynaphthalene-1-sulfonic acid), $C_{10}H_9NO_4S$, M_r 239.2.

Production. Although **182** may be obtained from 2-nitroso-1-naphthol by heating with bisul-

fite, this is less efficient than the corresponding reaction with 1-nitroso-2-naphthol. The preferred method is reduction of the azo dye derived from aniline and 1-hydroxynaphthalene-4-sulfonic acid with sodium dithionite or hydrazine [52].

Uses. 2-Amino-1-hydroxynaphthalene-4-sulfonic acid forms the stable quinone – diazide, 3-diazo-3,4-dihydro-4-oxonaphthalene-1-sulfonic acid [*20680-48-2*] (**183**) when diazotized in the presence of copper(II) sulfate. The stability of **183** and **173**, together with their moderate photosensitivity, led to their use in the earliest diazotype reprographic processes.

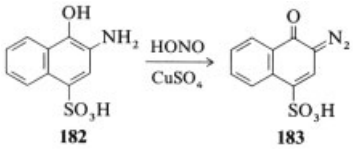

2-Amino-3-hydroxynaphthalene-6-sulfonic acid [*6399-72-0*] (**162**) (6-amino-7-hydroxynaphthalene-2-sulfonic acid), RM acid, $C_{10}H_9NO_4S$, M_r 239.2, is only sparingly soluble in hot water.

Production. 2-Aminonaphthalene-3,6-disulfonic acid (Section 6.5) is fused with 85 % potassium hydroxide at 180 °C for 20 h [7]. After quenching and acidification, the product is isolated in 75 % yield.

2-Amino-5-hydroxynaphthalene-7-sulfonic acid [*87-02-5*] (**163**) (7-amino-4-hydroxynaphthalene-2-sulfonic acid), J acid, $C_{10}H_9NO_4S$, M_r 239.2, is only slightly soluble (0.1 %) in cold water but readily soluble at 100 °C. The alkali-metal salts are readily soluble in water. Diazo compounds are normally coupled with J acid under alkaline conditions with reaction in the 6-position (ortho to hydroxyl); however, under acid conditions, coupling can occur in the 1-position. Reaction with bisulfite gives 4,7-dihydroxynaphthalene-2-sulfonic acid. The N-acylation, Nalkylation, and N-arylation reactions are important.

Production. A concentrated disodium salt solution of 2-aminonaphthalene-5,7-disulfonic acid (Section 6.5) is added to 72 % sodium hydroxide at 150 °C [7]. After heating to 187 °C over 10 h reaction is continued at this temperature for a further 8 h until completion of the reaction is indicated by giving equivalent diazotization and coupling titers for a worked-up sample. Quenching, acidification, purging of sulfur dioxide, and filtration at 55 °C followed by washing with water at 50 °C result in an 88 % yield of paste. Purification by salting out the sodium salt from a hot alkaline solution gives a 90 % recovery of the product as gray crystals.

An improved process with 90 % yield is claimed by feeding the concentrated 2-aminonaphthalene-5,7-disulfonic acid liquor and the caustic liquor simultaneously into the fusion reactor to maintain more stable conditions [54].

Uses. The J acid is an important coupling component in a wide range of direct dyes [e.g., C.I. Direct Blue 71 (**184**)] and reactive dyes [e.g., C.I. Reactive Orange 1 and C.I. Reactive Red 9 (**185**)].

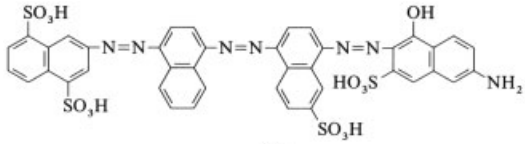

184

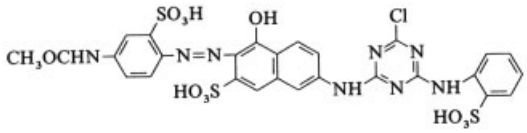

185

Many J acid derivatives, e.g., *N*-acetyl, *N*-methyl, *N*-phenyl, di-J acid, and the urea carbonyl J acid, are of comparable importance to J acid itself.

7-Acetamido-4-hydroxynaphthalene-2-sulfonic acid [*6334-97-0*] (**186**), *N*-acetyl J acid, is produced by reaction of the sodium salt of J acid with acetic anhydride in aqueous solution at 20 – 30 °C, followed by heating at 95 °C with sodium carbonate to hydrolyze any *O*-acetyl derivative present (yield 85 %). It is used to produce C.I. Acid Orange 27, C.I. Acid Red 137, and C.I. Direct Red 77.

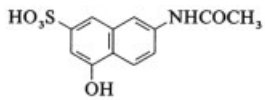

186

4-Hydroxy-7-methylaminonaphthalene-2-sulfonic acid [*22346-43-6*] (**187**), *N*-methyl J acid, is produced by reaction of 2,5-dihydroxynaphthalene-7-sulfonic acid (Section 4.7) with methylamine and aqueous sodium bisulfite at 96 °C for 36 h (yield 83 %). It is used to produce C.I. Reactive Orange 4, C.I. Reactive Orange 13, and C.I. Reactive Red 33.

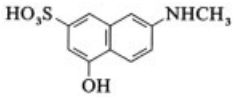

187

4-Hydroxy-7-phenylaminonaphthalene-2-sulfonic acid [*119-40-4*] (**188**), *N*-phenyl J acid, is produced by reaction of J acid with aniline in aqueous sodium bisulfite at reflux for 40 h (yield 85 %). It is used to produce C.I. Direct Violet 7, C.I. Direct Violet 9, C.I. Direct Violet 51, C.I. Direct Blue 78, and C.I. Direct Blue 168.

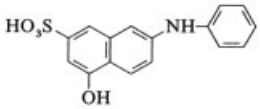

188

7,7′-Iminobis(4-hydroxynaphthalene-2-sulfonic acid) [*87-03-6*] (**189**), J acid imide, di-J acid, is produced by heating the sodium salt of J acid in aqueous sodium bisulfite at 106 °C for 36 h (yield 89 %). It is used to produce C.I. Direct Reds 31 and 149, as well as C.I. Direct Violets 48, 66, and 102.

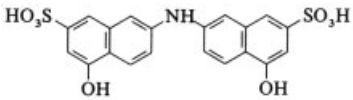

189

7,7′-Urylenebis(4-hydroxynaphthalene-2-sulfonic acid) [*134-47-4*] (**190**), J acid urea, carbonyl J acid, is produced by reacting the sodium salt of J acid with phosgene in aqueous solution at 30 °C and maintaining alkalinity with sodium carbonate. After being heated to 80 °C, the product is salted out with sodium chloride (yield ca. 100 %). It is used to produce C.I. Direct Orange 26; C.I. Direct Reds 4, 14, 23, and 80; and C.I. Direct Brown 112.

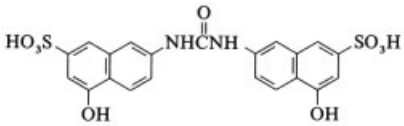

190

The related *N*-chloroacetyl, *N*-benzoyl, *N*-(4-aminobenzoyl), *N*-(3-aminobenzoyl), *N*-(3-methyl-4-aminobenzoyl), *N*-(4-methoxyphenyl), *N*-(3-carboxyphenyl), and *N*-(4-aminophenyl) derivatives are prepared by similar methods but have comparatively minor importance.

2-Amino-6-hydroxynaphthalene-8-sulfonic acid [*86-61-3*] (**191**), (7-amino-3-hydroxynaphthalene-1-sulfonic acid), $C_{10}H_9NO_4S$, M_r 239.2.

Production. 6-Nitrodiazo-1,2,4-acid (**175**) is heated with aqueous sodium sulfite solution in the presence of copper(II) sulfate. The solution is then acidified and heated further to cause desulfonation. The resulting nitrohydroxynaphthalenesulfonic acid is reduced with iron in the presence of acid [55].

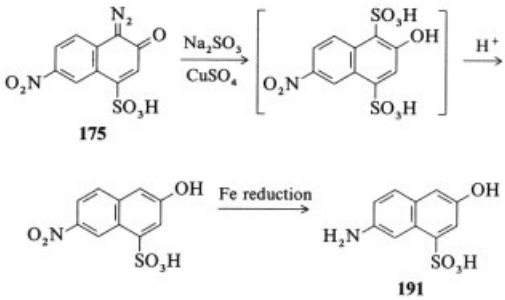

This process is preferred to the traditional deamination of **175** with ethanol in the presence of a copper catalyst.

2-Amino-8-hydroxynaphthalene-6-sulfonic acid [*90-51-7*] (**164**) (6-amino-4-hydroxynaphthalene-2-sulfonic acid), γ acid, $C_{10}H_9NO_4S$, M_r 239.2, is obtained as white or gray crystals which are only sparingly soluble in water even at 100 °C (0.4 %). The sodium salt dissolves readily in water to form solutions that exhibit a blue fluorescence. Coupling with

diazo compounds is usually carried out under acid conditions and occurs in the 1-position; coupling in the 7-position can be effected in alkaline solution.

Production. 2-Aminonaphthalene-6,8-disulfonic acid liquor (Section 6.5) is added over 16 h to 70 % sodium hydroxide at 190 °C, while the water is allowed to evaporate to maintain the temperature [7]. The reaction mixture is then heated for a further 6 h. Completion of the reaction is indicated by a worked-up sample giving identical titers for both diazotization and coupling. The batch is diluted with water, acidified, and heated at 75 °C to expel sulfur dioxide. The product is filtered off at 55 °C and washed with warm water; yield is 84 %.

An improved process with 90 % yield is claimed by feeding concentrated 2-aminonaphthalene-6,8-disulfonic acid dipotassium salt liquor and caustic soda liquor simultaneously into the fusion reactor [54].

Uses. The so-called γ acid is an important component in azo dyes; coupling can be carried out under acid or alkaline conditions. The *o*-amino azo configuration obtained under acid conditions leads to high lightfastness of special importance in monoazo acid red dyes for wool and polyamide, such as C.I. Acid Reds 37, 57, and 266 (**192**). Disazo direct dyes are also obtained by acid coupling (e.g., C.I. Direct Red 75) but trisazo blacks, exemplified by C.I. Direct Black 80 (**193**), which use γ acid as an alkali-coupled end and middle component, are of more significance as replacements for benzidine-based black direct dyes.

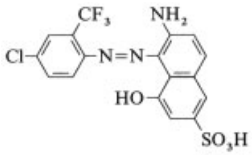

192

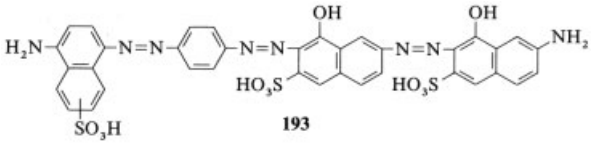

193

As with J acid, the *N*-acetyl and *N*-methyl derivatives are also important.

6-Acetamido-4-hydroxynaphthalene-2-sulfonic acid [*6361-41-7*] (**194**), *N*-acetyl γ acid, is produced by reaction of the sodium salt of γ acid with acetic anhydride in aqueous solution at 20 – 30 °C (yield 87 %). It is used to produce C.I. Acid Red 68, C.I. Reactive Orange 7, and C.I. Reactive Orange 16.

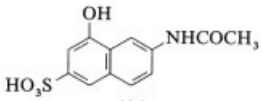

194

4-Hydroxy-6-methylaminonaphthalene-2-sulfonic acid [*6259-53-6*] (**195**), *N*-methyl γ acid, is produced by reaction of 2,8-dihydroxynaphthalene-6-sulfonic acid (Section 4.7) with methylamine in aqueous sodium bisulfite at 96 °C for 36 h (yield 95 %). It is used to produce C.I. Acid Brown 91.

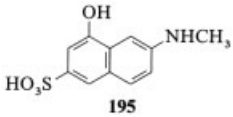

195

Other substituted γ acids that have been similarly used in specialized azo dyes include the *N*-β-hydroxymethyl, *N*-phenyl, *N*-(4-methoxyphenyl), *N*-(4-carboxyphenyl), *N*-(3-sulfophenyl), and *N*-(4-amino-3-sulfophenyl) derivatives.

2-Amino-8-hydroxynaphthalene-7-sulfonic acid (Na salt, [*74525-31-8*]) (**196**) (7-amino-1-hydroxynaphthalene-2-sulfonic acid), $C_{10}H_9NO_4S$, M_r 239.2.

Production. 7-Amino-1-naphthol (Chap. 7) is sulfonated at 0 °C with 96 % sulfuric acid

to yield 65 % of **196** together with 16 % of the isomeric 2-amino-8-hydroxynaphthalene-5-sulfonic acid [*21013-47-8*] (**197**). Separation by fractional precipitation of the sodium salt is not easy, but should the latter isomer be required the sulfonation temperature is increased to 30 °C, where-upon the product ratio changes from 65 : 16 to 50 : 35.

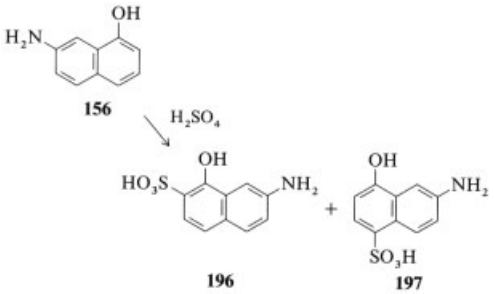

8.3. Aminohydroxynaphthalenedisulfonic Acids

1-Amino-2-hydroxynaphthalene-3,6-disulfonic acid [*2007-20-7*] (**198**) (4-amino-3-hydroxynaphthalene-2,7-disulfonic acid), $C_{10}H_9NO_7S_2$, M_r 319.3.

Production. R acid (Section 4.5) is coupled with diazotized aniline, and the isolated azo product is then reduced with sodium dithionite by a process similar to that employed for other *o*-aminonaphthols.

1-Amino-8-hydroxynaphthalene-2,4-disulfonic acid [*82-47-3*] (**165**) (4-amino-5-hydroxynaphthalene-1,3-disulfonic acid), Chicago acid, 2 S acid, $C_{10}H_9NO_7S_2$, M_r 319.3, and its acid sodium salt are readily soluble in water; the latter can be salted out with sodium chloride. In alkaline solution, coupling with diazo compounds takes place ortho to the hydroxyl group. Alkaline hydrolysis at 170 °C gives 1,8-dihydroxynaphthalene-2,4-disulfonic acid. Heating with dilute sulfuric acid results in desulfonation to give 1-amino-8-hydroxynaphthalene-2-sulfonic acid.

Production. A solution of the disodium salt of naphthosultam-2,4-disulfonic acid (**142**; Section 6.6) is concentrated and mixed with potassium hydroxide liquor at 130 °C [7]. The temperature is raised to 155 °C over 20 h to complete the reaction. After being cooled and diluted with water, the melt is run into excess hydrochloric acid and the monopotassium salt is collected by filtration at 30 °C and slurry washed with recycling of the wash liquors. A yield of 87 % is obtained.

Uses. Chicago acid is used as a coupling component for C.I. Direct Blue 1 (**199**), C.I. Direct Blue 22, and C.I. Direct Blue 76 (the dicopper complex of demethylated **199**).

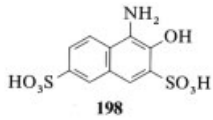

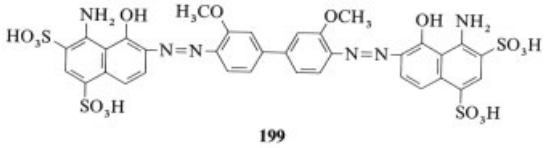

1-Amino-8-hydroxynaphthalene-3,6-disulfonic acid [*90-20-0*] (**166**) (4-amino-5-hydroxynaphthalene-2,7-disulfonic acid), H acid, $C_{10}H_9NO_7S_2$, M_r 319.3, is usually isolated as its monosodium salt [*5460-09-3*], a gray powder that is slightly soluble in cold water and readily soluble in hot water; the disodium salt is very soluble in water. With diazo compounds, H acid couples ortho to the amino group under acid conditions and ortho to the hydroxyl group under alkaline conditions. Bisazo derivatives are formed by stepwise acid and alkaline coupling, with the much slower acid coupling being performed first.

Production. 1-Aminonaphthalene-3,6,8-trisulfonic acid (Section 6.6) as the trisodium salt liquor is heated with aqueous sodium hydroxide in an autoclave [7]. The temperature is raised gradually from 150 to 182 °C, and heating is continued until the test for completion (strength by coupling equals strength by diazotization) is satisfied. The melt is diluted, added to excess dilute sulfuric acid, heated at 95 °C to expel sulfur dioxide, cooled, and filtered. The paste is washed and dried for ease of handling. The yield obtained is 72 %.

Given a yield from naphthalene of 62 % for Koch acid, the maximum overall yield of H acid from naphthalene is thus 45 %. Large-scale H

acid production uses naphthalene as starting material, and the through process to Koch acid involves the three process stages of sulfonation, nitration, and reduction. These all require optimization with a view to yield and purity of the ultimate H acid [56]. At the fusion stage, a higher temperature (200 °C) may be used to shorten the reaction period [57]. In a fully integrated process, many of the stage parameters are interrelated [58]. Equally important to the fusion reaction conditions are the work-up conditions to achieve good physical form, a major factor in final quality. The H acid is precipitated as the monosodium salt which has a strong tendency to be thixotropic and is hence difficult to wash.

Uses. Like certain other aminonaphthol derivatives, H acid has been claimed to be a useful corrosion inhibitor. The versatility of H acid as a coupling component is exemplified by alkaline coupling to give monoazo reds such as C.I. Acid Red 33 (**200**) and C.I. Reactive Red 12 (**201**). Even more important, on a volume basis, are disazo blues and blacks such as C.I. Direct Blue 15, C.I. Direct Blue 218, C.I. Acid Black 1, C.I. Direct Black 19, and C.I. Reactive Black 5 (**202**).

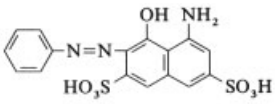

200

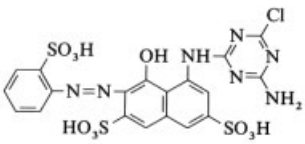

201

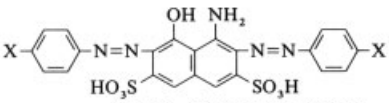

202, X = $SO_2CH_2CH_2OSO_3H$

Some N-substituted derivatives of H acid are isolated for more specialized uses. However, in situ N-acylation is more common with agents such as *p*-toluenesulfonyl chloride or fiber-reactive systems containing an acyl group or equivalent (e.g., as in **201**).

4-Acetylamino-5-hydroxynaphthalene-2,7-disulfonic acid [*134-34-9*] (**203**), *N*-acetyl H acid, is produced by reacting H acid in aqueous alkaline solution with acetic anhydride at 70 °C for 1 h, followed by heating at 95 °C for 2 h to hydrolyze the *O*-acetyl derivative (yield 88 %). It is used to produce C.I. Acid Red 1, C.I. Acid Red 35, and C.I. Acid Violet 7.

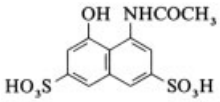

203

4-Benzamido-5-hydroxynaphthalene-2,7-disulfonic acid [*117-46-4*] (**204**), *N*-benzoyl H acid, is produced by reaction of H acid with benzoyl chloride in aqueous alkaline sodium carbonate solution at 35 °C for 5 h, followed by heating at 95 °C for 2 h to hydrolyze the *O*-benzoyl derivative (yield 80 %). It is used to produce C.I. Reactive Red 4 and C.I. Direct Green 13.

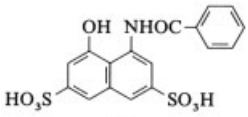

204

1-Amino-8-hydroxynaphthalene-4,6-disulfonic acid [*130-23-4*] (**167**) (4-amino-5-hydroxynaphthalene-1,7-disulfonic acid), K acid, $C_{10}H_9NO_7S_2$, M_r 319.3: the acid sodium salt of **167** is very soluble in water, but the acid potassium salt is less soluble and is more easily salted out. K acid couples under acid and alkaline conditions in a similar manner to H acid.

Production. 1-Aminonaphthalene-4,6,8-trisulfonic acid paste (Section 6.6) is heated with sodium hydroxide liquor in an autoclave at 170 °C for 12 h, after which the mixture is added to excess dilute hydrochloric acid. After heating to remove sulfur dioxide, the product (including ca. 8 % of the isomeric 5-amino-2-hydroxynaphthalene-4,8-disulfonic acid) is completely precipitated by cooling and is isolated in 86 % yield. Purification of crude K acid may be effected by dissolving it in aqueous alkali and filtering off the less soluble 5,2,4,8-isomer before reprecipitating.

Uses. The N-substituted derivatives of K acid, such as *N*-acetyl, *N*-benzoyl, *N*-(2,5-dichlorobenzoyl), and *N*-benzyl, are used as monoazo coupling components in special cases where an improvement in dye properties over the H acid analogue is demonstrable.

4-Benzamido-5-hydroxynaphthalene-1,7-disulfonic acid [*6361-49-5*] (**205**), *N*-benzoyl K acid, is produced by reacting purified K acid with benzoyl chloride in aqueous alkaline sodium carbonate solution at 35 – 40 °C for 18 h, followed by heating for 1 h at 90 °C to hydrolyze the *O*-benzoyl derivative (yield 80 %). It is used to produce C.I. Acid Red 133, C.I. Acid Red 157, and C.I. Mordant Red 21.

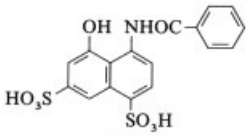

205

2-Amino-5-hydroxynaphthalene-1,7-disulfonic acid [*6535-70-2*] (**168**), sulfo J acid, $C_{10}H_9NO_7S_2$, M_r 319.3.

Production. 2-Aminonaphthalene-1,5,7-trisulfonic acid (Section 6.6) is heated with sodium hydroxide liquor at 170 °C and worked up in a similar way to the K acid. Alternatively, J acid may be sulfonated with 65 % oleum.

Uses. Sulfo J acid is used as a coupling component in azo reactive dyes, e.g., C.I. Reactive Red 6 (the copper complex of **206**).

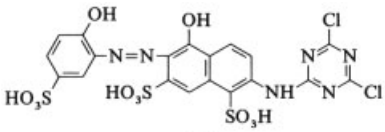

206

2-Amino-5-hydroxynaphthalene-4,8-disulfonic acid [*74832-35-2*] (**207**) (3-amino-8-hydroxynaphthalene-1,5-disulfonic acid), oxy C acid, $C_{10}H_9NO_7S_2$, M_r 319.3.

Production. 3,8-Diaminonaphthalene-1,5-disulfonic acid (Section 6.8) is monoacetylated in aqueous solution, and the free amino group is diazotized. The diazo compound is added to dilute sulfuric acid at 100 °C to hydrolyze both the diazo and the acetylamino groups [59].

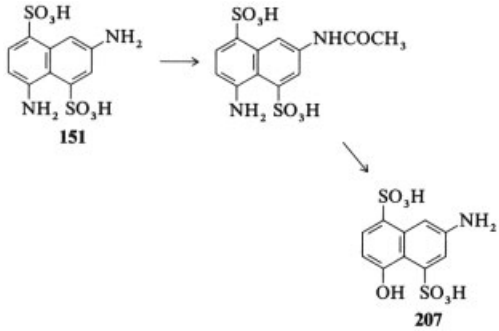

2-Amino-8-hydroxynaphthalene-3,6-disulfonic acid [*90-40-4*] (**169**) (3-amino-5-hydroxy-2,7-disulfonic acid), RR acid, 2 R acid, Columbia acid, $C_{10}H_9NO_7S_2$, M_r 319.3, and its alkali salts are readily soluble in water. Coupling with diazo compounds takes place under alkaline conditions in the position ortho to the hydroxyl group.

Production. 2-Aminonaphthalene-3,6,8-trisulfonic acid liquor (Section 6.6) is concentrated by evaporation and heated with sodium hydroxide liquor in an autoclave [7]. The temperature is raised to 195 °C over 10 h and held at 195 °C for another 8 – 10 h until completion of the reaction is determined by working up a test sample and obtaining a strength by alkaline coupling equal to the diazotization titer. After dilution, acidification, and stirring while hot until all sulfur dioxide is expelled, the product is filtered off as its monosodium salt at 55 °C and washed with brine. Most of the isomeric 2-amino-3-hydroxynaphthalene-6,8-disulfonic acid impurity is removed in the filtrate. Crude RR acid is obtained in 65 % yield and is suitable for certain outlets. Further purification is necessary for critical derived dyes, and this is achieved by dissolving in dilute alkali, filtering at 80 °C, and reprecipitating by addition of sodium chloride and hydrochloric acid. Filtration at 55 °C gives 90 % recovery of product.

Uses. The RR acid is used as a coupling component in azo dyes, e.g., C.I. Food Black 2 and C.I. Direct Brown 151.

9. References

General References

1. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin 1964.
2. *Houben-Weyl,* 4th ed., **9,** 474 – 492.
3. E. N. Abrahart: *Dyes and their Intermediates,* Arnold, London 1977.
4. E. H. Rodd: *Chemistry of Carbon Compounds,* 2nd ed., vol. 3 G, Elsevier, Amsterdam 1978.
5. N. Donaldson: *The Chemistry and Technology of Naphthalene Compounds,* Arnold, London 1958.
6. G. Booth: *The Manufacture of Organic Colorants and Intermediates,* Society of Dyers and Colourists, Bradford 1988.
7. BIOS, FIAT reports: German dyestuffs and dyestuffs intermediates, including manufacturing processes, plant design and research data 1945 – 1948.
8. H. E. Fierz-David, L. Blangey: *Fundamental Processes of Dye Chemistry,* Interscience, New York 1949.
9. K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. 1, Academic Press, New York 1952.
10. H. A. Lubs (ed.): *The Chemistry of Synthetic Dyes and Pigments,* Reinhold Publ. Co., New York 1954.
11. H. Hiyama: "Medicines and Agricultural Chemicals Derived from Naphthalene," *Kagaku to Kogyo (Osaka)* **54** (1980) 238 – 247.
12. M. Kodama, N. Sakota: "Recent Advances in Application and Synthesis of Naphthalene Compounds," *Yuki Gosei Kagaku Kyokaishi* **39** (1981) 960 – 972.
13. H-G. Franck, J. W. Stadelhofer: *Industrial Aromatic Chemistry,* Springer-Verlag, Berlin 1988.
14. N. I. Sax, R. J. Lewis: *Dangerous Properties of Industrial Materials,* 7th ed., **vol. 3,** Van Nostrand Reinhold, New York 1989.
15. *Ullmann,* 4th ed., **17,** 83. **3rd ed., 12,** 617. *Ullmann,* 3th ed., **12,**617.

Specific References

16. G. Dittmar, *Chem. Rundsch.* **30** (1977) no. 46, 23.
17. H. Bretscher, G. Eigenmann, E. Plattner, *Chimia* **32** (1978) 180.
18. Colour Index, vol. 4, 3rd ed., Society of Dyers and Colourists, Bradford 1971.
19. Colour Index, vol. 6 (1975); vol. 7 (1982), Society of Dyers and Colourists, Bradford.
20. J. J. Kirkland: *Modern Practice of Liquid Chromatography,* Wiley-Interscience, New York 1971.
21. L. R. Snyder, J. J. Kirkland: *Introduction to Modern Liquid Chromatography,* Wiley-Interscience, New York 1979.
22. A. A. Spryskov, N. A. Orsyankina, *J. Gen. Chem. USSR (Engl. Transl.)* **20** (1950) 1043.
23. H. E. Fierz-David, C. Richter, *Helv. Chim. Acta* **28** (1945) 257.
24. Sumitomo Chem. Co. Ltd., JP 7 3 38 699, 1970 (T. Ikeda et al.).
25. Bayer, DE 2 728 070, 1977 (H. U. Alles et al.).
26. K. H. Saunders, R. L. M. Allen: *Aromatic Diazo Compounds,* Arnold, London 1985.
27. Union Carbide, US 3 356 743, 1963; US 3 378 591, 1964 (B. T. Freure).
28. F. A. Yukhro et al., Deposited Doc. 1981, SPSTL 566 Khp-D 81, *Chem. Abstr.* **98** 160 364.
29. Stauffer, US 3 480 671, 1967 (H. Tilles et al.).
30. Union Carbide, US 2 903 478, 1959 (J. A. Lambrech).
31. ICI, GB 994 918; 995 800, 1962 (A. F. Crowther, L. H. Smith).
32. R. Kuhn, F. Bar, *Justus Liebigs Ann. Chem.* **516** (1935) 143.
33. Amer. Cyanamid, US 3 504 045; 3 504 046, 1968 (E. J. Scharf, H. R. Kemme).
34. Japan Soda Co., FR 1 337 797, 1961; *Chem. Abstr.* **60** 2871.
35. F. Seidel et al., *J. Prakt. Chem.* **2** (1954) 53. M. Scalera, H. Z. Lecher, *J. Prakt. Chem.* **3** (1956) 232.
36. A. S. Lindsey, H. Jeskey, *Chem. Rev.* **57** (1957) 583.
37. Henkel and Cie, DE 1 203 281, 1962.
38. J. Bailey, L. A. Williams in K. Venkataraman (ed.): *Chemistry of Synthetic Dyes,* vol. IV, Academic Press, New York 1971.
39. Sugai Chemical, JP 7 0 10 935, 1966 (H. Hiyama et al.).
40. Bayer, DE 1 943 543, 1969 (C. P. Mertig).
41. Amer. Cyanamid, DE 2 346 459, 1972 (M. L. Feldman et al.).
42. U.S. Rubber, US 2 556 665, 1951 (A. E. Smith, O. L. Hoffman).
43. Japan Soda, JP 6 7 17 016, 1964 (S. Kano); *Chem. Abstr.* **68** (1968) 29 479.
44. *IARC Monograph on the Evaluation of Carcinogenic Risk of Chemicals to Man,* **vol. 4,** IARC, Lyon 1974. "Some Aromatic Amines and Azo Dyes in the General and

Industrial Environment," *Environmental Carcinogens Selected Methods of Analysis,* **vol. 4,** IARC, Lyon 1981.
45. Y. Masuda, D. Hoffman, *Anal. Chem.* **41** (1969) 650.
46. Bayer, DE 2 517 437-9, 1975 (H. U. Blank et al.).
47. CFMC, FR 1 075 110, 1954 (J. Dassigny). CFMC, FR 1 108 109, 1956 (J. Robin).
48. H. Koepernik, R. Borsdorf, *J. Prakt. Chem.* **325** (1983) 1002.
49. Ciba-Geigy, DE 2 856 567, 1977 (B. Albrecht).
50. Bayer, DE 2 703 076; 2 747 714, 1977 (O. Barth et al.).
51. M. A. Marchisio et al., *Br. J. Ind. Med.* **33** (1976) 269.
52. Taoka Chem. Co., JP 76 146 447, 1975 (S. Kawasaki et al.).
53. Bayer, DE 3 010 372, 1980 (R. Puetter et al.).
54. Ciba-Geigy, DE 2 813 570, 1978 (H. Breitschmid).
55. Geigy, US 2 124 070, 1938 (A. Krebser, F. Vamotti).
56. T. Hayashi, *Yamaguchi Daigaku Rigakkai Shi* **2** (1951) 67; *Chem. Abstr.* **49** (1955) 2390.
57. Bayer, DE 2 716 030, 1977 (H. Behre et al.).
58. Sumitomo, DE 2 727 345, 1977 (N. Kotera et al.).
59. Bayer, DE 1 117 235, 1959 (K. H. Schuendehuette et al.).

Naphthalene Dicarboxylic Acid → Carboxylic Acids, Aromatic
Naphthalic Acid → Carboxylic Acids, Aromatic

Naphthalimide Dyes and Pigments

Sylke Haremsa, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **Alkoxynaphthalimides** 2
3. **4-Aminonaphthalimides** 2
4. **1′,8′-Naphthoylenebenzimidazoles** . . . 3
5. **Imides of Naphthalene-1,4,5,8-tetracarboxylic Dianhydride** 3
6. **1′,8′-Naphthoylenebenzimidazole Peridicarboximides** 4
7. **Bis-benzimidazole Derivatives of Naphthalene-1,4,5,8-tetracarboxylic acid** 4
8. **1′,8′-Naphthoylenepyrazoles** 4
9. **Benzo[*k,l*-]xanthene- and Benzo[*k,l*-]-thioxanthene-3,4-dicarboximides** 5
10. **Azo- and Azomethinenaphthalimides** . 5
11. **Perylene-3,4,9,10-tetracarboxydiimides** 6
12. **References** 7

1. Introduction

On account of their versatility in application and the excellent fastness of some of them, the naphthalimide dyes have now acquired great industrial importance. They are derivatives of naphthalene-1,8-dicarboximide (naphthalimide) (**1**)[*81-83-4*].

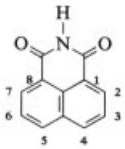

1

Naphthalimide itself is a colorless, crystalline compound, prepared by the oxidation of acenaphthene to naphthalic anhydride [*81-84-5*] [1], followed by conversion to the imide [2].

Naphthalimides only become of interest for color chemistry after substitution with electron-donating groups, preferably in the 4- and 5-positions. The electron-attracting effect of the anhydride or imide group allows the ready exchange of chlorine, nitro, or sulfonic acid groups for alkoxy, aryloxy, amino, and mercapto substituents.

4-Alkoxy- and 4,5-dialkoxynaphthalimides are almost colorless, but are of commercial interest as optical brighteners because of their fluorescence properties (see also → Fluorescent Dyes). The introduction of nitrogen or sulfur in the 4-position gives colored compounds. Together with the perylene-3,4,9,10-tetracarboxydiimides and the perinone dyes, which also belong to the naphthalimide dyes, it is possible to cover the whole color range. The range of applications is very diverse: alongside classical fields, such as dyeing of natural and synthetic textile fibers, plastics, coating and paint pigments, increasingly novel applications are coming to light, in which esthetic considerations are no longer of primary importance. These include, for example, applications in histology for cell staining [3], as laser dyes [4], as the charge-creating layer in electrophotography [5], and in liquid crystal displays [6].

The number of naphthalimide dyes described in the literature is now immense. For this reason it is only possible to describe a few selected types, which have been divided into the following groups on the basis of their structures:

- Alkoxynaphthalimides
- 4-Aminonaphthalimides
- 1′,8′-Naphthoylenebenzimidazoles
- Imides of naphthalene-1,4,5,8-tetracarboxylic dianhydride
- 1′,8′-Naphthoylenebenzimidazole peridicarboximides
- Bis-benzimidazole derivatives of naphthalene-1,4,5,8-tetracarboxylic acid
- 1′,8′-Naphthoylenepyrazoles
- Benzo[*k,l*]xanthene- and benzo[*k,l*]thioxanthene-3,4-dicarboximides
- Azo- and azomethinenaphthalimides
- Perylene-3,4,9,10-tetracarboxydiimides

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_059

2. Alkoxynaphthalimides

Naphthalimides with alkoxy substituents in the 4- and 5-positions find application as chlorite-stable optical brighteners. Their relatively low molar extinction coefficients are a disadvantage in comparison with other brighteners. These compounds are produced by reaction of 4-chloro- or 4,5-dichloronaphthalimides with alkoxides. Alternatively, 4-chloronaphthalic anhydride can be treated first with an alkoxide and then with an amine, hydrazine, or hydroxylamine to yield the corresponding imide. The nature of the imide residue determines the field of application. Thus, nonionic naphthalimides such as **2** [*56148-88-0*] are used for brightening polyesters, cellulose acetates, and cotton [7, 8] and also in the base coat of automotive lacquers [9]. Basic cationic naphthalimides such as **3** [*51286-95-8*] are employed for polyacrylonitrile and wool [10, 11].

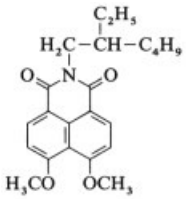

2

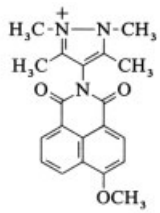

3

ω-Sulfoalkoxynaphthalimides are suitable for brightening synthetic, semisynthetic, and natural fibers. In contrast to most other optical brighteners, which have a red hue, they have a green hue [12].

Another field of application for alkoxynaphthalimides is as electrophotographic toners [13].

3. 4-Aminonaphthalimides

The introduction of an amino group into the 4-position yields brilliant greenish yellow compounds. The usual method of preparation is to treat 4-chloronaphthalic anhydride [*4053-08-1*] with a primary amine to yield the corresponding imide. The nucleophilic substitution with ammonia that follows is frequently carried out under copper catalysis. The arylimides, e.g., the 2′,4′-xylidide (**4**), are commercially important disperse dyes, which, however, can also be employed for coloring plastics or for the production of daylight fluorescent colors [14].

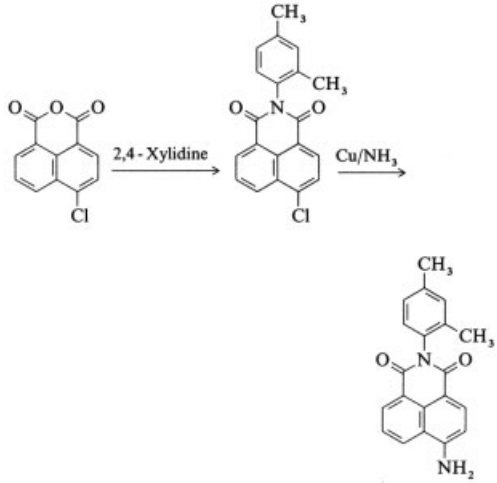

4

Moreover, the corresponding 4-sulfonaphthalimide can also be employed as starting material for the synthesis of 4-*N*-alkylaminonaphthalimides, as illustrated by the example of 4-cyclohexylaminonaphthalene-1,8-dicarboxylic acid *N*-methylimide [*33209-23-3*], which is suitable for dyeing polyesters [15]. Wool, silk, and synthetic polyamides can be dyed in brilliant greenish yellow hues with 4-aminonaphthalimides, but the lightfastness is low. The affinity of the dyes to the fibers is increased by introduction of a sulfonic acid group, preferably in the 3-position. An example of this is Brilliant Sulfoflavin FF (**5**) (C.I. 56 205) [*2391-30-2*] [16], whose further sulfonation in the 6-position leads to a strongly fluorescent, water-soluble dye that can be employed as a biological tracer [3].

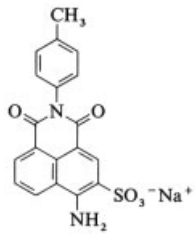

5

The usual anchors for reactive dyes, such as the 4,6-dichlorotriazinyl [17] or 2,4-difluoro-5-chloropyrimidyl groups [18], are also encountered in the 4-aminonaphthalimides. The cotton dyes are fast to washing, but as in the examples discussed above are less than satisfactory as far as lightfastness is concerned.

4. 1′,8′-Naphthoylenebenzimidazoles

The reaction of naphthalic anhydrides with *o*-phenylenediamines yields 1′,8′-naphthoylenebenzimidazoles [19]. The 7-*H*-benzimidazo-[2,1*a*]-benz[*d,e*]isoquinolone-(7) framework can also be constructed by first condensing naphthalic anhydride with an 2-nitroaniline and then carrying out ring closure after reduction of the nitro group [20]. Isomers are produced if the aniline or the naphthalic anhydride is substituted, and these can differ in their absorption maxima [21]. Various substituents can also be introduced into the naphthalene moiety by treating 4′-chloro-1′,8′-naphthoylenebenzimidazole with nucleophiles. The 1′,8′-naphthoylenebenzimidazoles are suitable for dyeing polyesters, polystyrene, and other polymers. As a result of their solubility in hydrocarbons, they are also suitable for nondestructive materials testing. The ionic compound **6** [*58888-73-6*], for example, is suitable for polyacrylonitrile fibers, which it dyes in brilliant, lightfast, green-shot yellow hues [22].

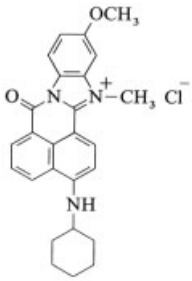

6

The polyester dye **7** [*5504-69-8*] is obtained by condensation of naphthalic anhydride with 1,2-diamino-4-methoxybenzene [23].

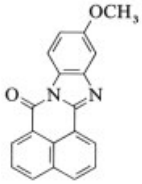

7

Polyester dyes with brilliant, highly fast, red tones are obtained by annellation of a benzothiophene group [21]. For this purpose, 2-(2-aminophenylthio)-1,8-naphthalic anhydride is diazotized, which results in ring closure to yield benzo[*b*]naphtho[3,2-*d*]thiophen-6,7-dicarboxylic anhydride. Condensation with *o*-phenylenediamine then leads to the dye **8** [*99634-44-3*].

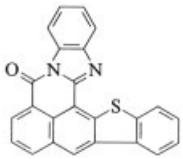

8

In the case of 4-alkylamino substituents, which are introduced by nucleophilic substitution of *N*-methyl-4′-bromo-1′,8′-naphthoylenebenzimidazolium chloride, itis possible to produce brilliant, light-fast polyacrylonitrile dyes, whose color scale ranges from yellow to violet [22].

5. Imides of Naphthalene-1,4,5,8-tetracarboxylic Dianhydride

The simple imides of naphthalene-1,4,5,8-tetracarboxylic dianhydride have no role as dyes because of the lack of a suitable donor – acceptor substitution. The acceptor behavior of the diimide allows electrochemical reduction to complex salts such as **9** [*119298-43-0*], which possesses the surprisingly high conductivity of 7×10^{-3} S/cm [24].

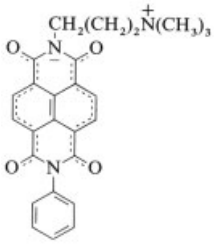

9

6. 1′,8′-Naphthoylenebenzimidazole Peridicarboximides

If naphthalene-1,4,5,8-tetracarboxylic dianhydride is treated first with a primary amine and then condensed with a 1,2-diaminobenzene, a dye of the 1′,8′-naphthoylenebenzimidazole peridicarboximide type is obtained [25]. Originally introduced as vat dyes, they have acquired industrial importance since the appearance of polyester fibers. The most important representatives, which dye in yellow shades, are [26]:

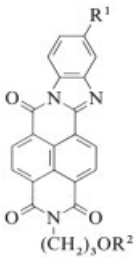

10 $R^1 = CH_3$; $R^2 = CH_3$
[*27767-97-1*]
11 $R^1 = Cl$ $R^2 = (CH_2)_3CH_3$
[*56532-67-3*] [21]
12 $R^1 = CN$ $R^2 = CH_3$
[*22176-47-2*]

7. Bis-benzimidazole Derivatives of Naphthalene-1,4,5,8-tetracarboxylic acid

Naphthalene-1,4,5,8-tetracarboxylic acid can also react on both sides with 1,2-diaminobenzenes. This yields *cis – trans* isomeric mixtures, which are of importance as vat dyes of indanthrene quality and as pigments in the paint industry. In the simplest case naphthalene-1,4,5,8-tetracarboxylic acid is condensed with 1,2-diaminobenzene to give an isomeric mixture of *cis*- and *trans*-**13**.

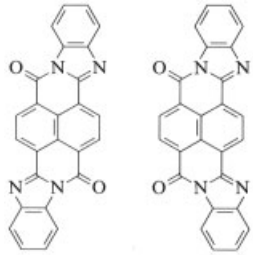

trans - **13** *cis* - **13**

The two isomers have different color tones and can be readily separated on the basis of their different solubilities, e.g., by recrystallization from sulfuric acid [27, 28], or by formation of an addition complex with potassium hydroxide [29] or 4-toluenesulfonic acid [30]. This yields the isomers Indanthrene Brilliant Orange GR (C.I. 71 105) [*4216-02-8*] and the cis form Indanthrene Bordo RR (C.I. 71 105) [*4216-02-8*].

If 4-chloro-1,2-diaminobenzene or 4-ethoxy-1,2-diaminobenzene is employed in place of 1,2-diaminobenzene, then the products obtained are Indanthrene Printing Brown 5R [*4601-61-0*] or Indanthrene Printing Brown B [*5492-43-2*] [31].

8. 1′,8′-Naphthoylenepyrazoles

Isomeric mixtures also occur in 1′,8′-naphthoylenepyrazoles. Starting from 4-chloronaphthalene-1,8-dicarboxylic hydrazide [*37877-34-2*] [32] a nucleophilic substitution is first carried out at the 4-position, preferably with mercapto or amino groups. Finally, the hydrazone is produced from an acetoacetic acid derivative, which is then cyclized under acidic or basic conditions [33, 34]. Dyes such as **14**, **15** are suitable for synthetic fibers, e.g., polyesters, acetylcellulose and polyamides [33] (see figure on next page).

Other representatives of this group are, for example, the bis-pyrazole derivatives of naphthalene-1,4,5,8-tetracarboxylic acid [35] and the products of the reactions of 4′,5′-dichloro-1′,8′-naphthaloylenepyrazoles with 2-mercaptobenzimidazoles [36].

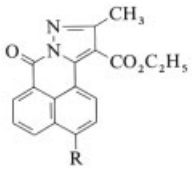

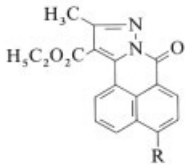

cis-**14** *trans*-**14** R = −S−

cis-**14** *trans*-**15** R = −N O

9. Benzo[*k,l*-]xanthene- and Benzo[*k,l*-]thioxanthene-3,4-dicarboximides

The starting compounds for this group of dyes are the benzo[*k,l*]xanthene- or benzo[*k,l*]thioxanthene-3,4-dicarboxylic acid anhydrides **16** [*36310-05-1*] and **17** [*14121-49-4*]. To synthesize

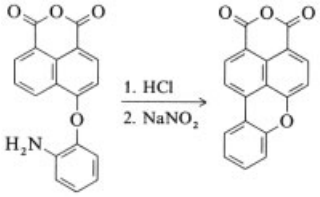

16

these compounds, 4-chloro- [37] or 4-nitronaphthalic anhydride [38] is first treated with a 2-aminophenol or 2-aminothiophenol. After diazotization, cyclization to give the heterocycle

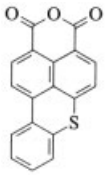

17

takes place. An alternative route is the reaction of 4-amino-5-bromonaphthalic anhydride with a phenol. Cyclization to the benzoxanthene also occurs via diazotization [37].

Anhydrides prepared in this manner can be treated with amines [39], diaminobenzenes [37], and hydrazines [40]. They subsequently yield pyrazoles [41] on condensation with acetoacetic acid derivatives. Benzo[*k,l*]thioxanthene-3,4-dicarboxylic acid imides can also be oxidized to the 7,7-dioxides [42]. Structural examples are the compounds **18** [*59373-45-5*] [43], **19** [41], and **20** [*82130-17-4*] [42].

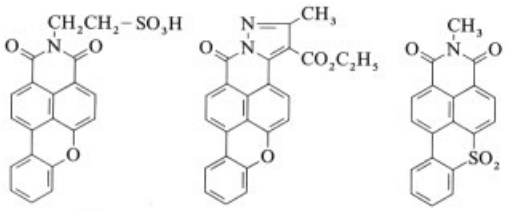

18 **19** **20**

The intense, lightfast, yellow to orange compounds are suitable for the dyeing of polyesters [37,44], mineral oils, thermoplastics [39], polyamides [43], for paint and printing pigments [40], and as liquid crystal dyes [45,46].

10. Azo- and Azomethinenaphthalimides

The azonaphthalimide dyes are generally characterized by good fastness and high thermal stability [47,48]. The absorption maxima of the 3- and 4-substituted compounds can differ by 60 nm. The bathochromic shift upon 4-substitution is attributed to the strong donor–acceptor interaction [49]. Alkaline hydrolysis of 4-halonaphthalimides leads to the corresponding 4-hydroxy derivatives [50], which find application as coupling components. The resulting 3-arylazo-4-hydroxynaphthalimides can be used to dye polyesters [51], polyacrylonitrile [52], polyamides and wool [53], and cellulose [54] in intense, lightfast red tones. Compound **21** [*42358-39-4*] serves as an example of this type of structure.

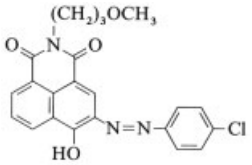

21

Innumerable examples have been published for the application of 3′-naphthalimido-disazo pigments as the charge-creating layer in electrophotography [55–57].

The 4-arylazonaphthalimides are prepared by diazotization of 4-aminonaphthalimide and coupling to dialkylanilines or heterocycles [47, 58, 59].

The azomethine pigments are obtained by formylation of the reactive 4-hydroxynaphthalimides in the 3-position, followed by condensation with aromatic amines [60, 61].

11. Perylene-3,4,9,10-tetracarboxydiimides

The oxidative dimerization of naphthalimide in molten alkali yields perylene-3,4,9,10-tetracarboxydiimide **22** [62, 63].

Depending on the particle size and the crystalline modification, the product is a violet or red-brown covering or transparent pigment that is principally employed for automotive coatings [64, 65]. In order to synthesize further substituted diimides, perylene-3,4,9,10-tetracarboxylic dianhydride [*128-69-8*] is first prepared by solvolysis in concentrated sulfuric acid [66].

Controlled reaction with primary amines yields monoanhydride-monoimides [67, 68] and the symmetrically and unsymmetrically substituted diimides [69, 70]. The symmetrically substituted diimides have achieved the greatest commercial importance. Depending on their properties they find application as vat dyes of indanthrene quality, highly fast pigments for plastics and coatings, fluorescent dyes with high photostability or as charge-creating layers in electrophotography.

The condensation of perylene-3,4,9,10-tetracarboxylic dianhydride with aliphatic, aromatic, or heterocyclic amines is preferably performed in quinoline with zinc-salt catalysis [71, 72], but is also possible under pressure in aqueous solution [73].

Compounds **22** to **28** have achieved importance as high-quality red pigments. Indanthrene Red LGG (**24**) and Indanthrene Scarlet R (**25**) [74] are valuable vat dyes.

In the case of the perylene-3,4,9,10-tetracarboxydiimides, the crystal structure has a decisive influence on the color. Changing the crystal modification can often bring about improvements in chroma and color intensity [75, 76]. Considerable overlap of molecular stacks, as in **29**, leads to a black coloration. The relationship between crystalline structure and color has been investigated on the basis of several examples [77].

22 R = H (C.I. 71 129) [*81-33-4*]
23 R = CH_3 (C.I. 71 130) [*5521-31-3*]

24 R = 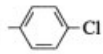 (C.I. 71 135) [*2379-77-3*]

25 R = 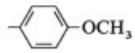 (C.I. 71 140) [*6424-77-7*]

26 R = 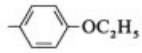 (C.I. 71 145) [*24108-89-2*]

27 R = 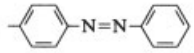 (C.I. 71 155) [*3049-71-6*]

28 R = 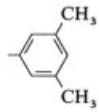 (C.I. 71 137) [*4948-15-6*]

29 R = 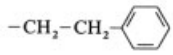 [*67075-37-0*]

30 R = 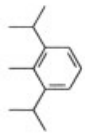 [*82953-57-9*]

31 R = 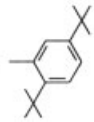 [*83054-80-2*]

The introduction of bulky alkyl substituents results in the formation of soluble fluorescent dyes such **30** [78] and **31** [79]. On account of their high quantum yields and good photostability, they are suitable, when incorporated into glass-clear plastics, for the surface concentration of light and can therefore be employed for fluorescent solar collectors [80]. Good solubility combined with a bathochromic

shift is promoted by donor substitution at positions 6, 7, 12, and 13 of the perylene nucleus [81]. On the other hand, 1,6,7,12-tetrachloroperylene-3,4,9,10-tetracarboxydiimide is a brilliant, weatherproof, red pigment for paints and plastics [82].

The reaction of perylene-3,4,9,10-tetracarboxylic dianhydride with 1,2-diaminobenzenes leads to the corresponding bisbenzimidazoles. They are increasing in importance also in the form of azo pigments in combination with 4-hydroxynaphthalimides as electrophotographic dyes [83–85].

Reaction of perylene-3,4,9,10-tetracarboxylic acid with 4-(2′-hydroxyethoxy)aniline gives a pigment that can be copolymerized with bis(2-hydroxyethyl)terephthalate [86]. Fluorescent red-orange fibers can be produced in this manner.

12. References

1. AS Kaza Chem. Scie., SU 535 306, 1975.
2. Bayer, DE-OS 3 719 850, 1987.
3. W. Stewart, *J. Am. Chem. Soc.* **103** (1981) 7615 – 7620.
4. M. Sadrai, C. Bird, *Opt. Commun.* **51** (1984) 62.
5. Dainippon, US 4 728 592, 1986.
6. Mitsubishi, JP 61 281 220, 1984.
7. BASF, DE-AS 2 331 307, 1973.
8. BASF, US 4 139 532, 1979.
9. BASF, DE 3 642 621, 1986.
10. Hoechst, EP 28 777, 1979.
11. Ugine Kuhlmann, US 3 932 413, 1973.
12. Hoechst, DE 3 618 458, 1986.
13. Nippon Kayaku, JP 59 114 546, 1982.
14. Hoechst, US 4 081 446, 1976.
15. Mitsubishi, JP 7 209 106, 1986.
16. IG Hoechst, *Friedländer* **16,** 850.
17. ICI, DE-OS 2 422 466, 1973.
18. Bayer, US 4 051 134, 1974.
19. A. Bistrzycki, J. Risi, *Helv. Chim. Acta* **8** (1925) 810.
20. Yamamoto, JP 51 130 424, 1975.
21. A. T. Peters, M. J. Bide, *Dyes Pig.* **6** (1985) 417 – 433.
22. Ugine Kuhlmann, GB 1 159 651, 1969.
23. Hoechst, DE-AS 2 253 553, 1972.
24. G. Heywang et al., *Angew. Chem.* **101** (1989) 462 – 464.
25. *Friedländer* **18,** 1501; **19,** 2198.
26. Hoechst, DE-AS 1 049 821, 1957.
27. *Friedländer* **18,** 1497 – 1499.
28. *Friedländer* **19,** 2188 – 2189.
29. B I O S, Final Report 987, 107.
30. *Friedländer* **19,** 2189 – 2192.
31. W. S. Masloch et al., SU 386 967, 1971.
32. L. Carpino, A. Santilli, R. Murray, *J. Am. Chem. Soc.* **82** (1960) 2728 – 2731.
33. Hoechst, US 1 434 743, 1972.
34. Hoechst, DE-OS 2 238 378, 1972.
35. Bayer, DE-OS 2 726 902, 1977.
36. Mitsubishi, JP 51 142 048, 1975.
37. A. T. Peters, Y. S. S. Behesti, *J. Soc. Dyers Colour.* **105** (1989) no. 1, 29 – 35.
38. P. H. Grayshan, A. M. Kadhim, A. T. Peters, *J. Heterocycl. Chem.* **11** (1974) 33 – 38.
39. Hoechst, DE-OS 1 936 612, 1 936 613, 1 936 614, 1969.
40. Hoechst, DE-OS 2 017 765, 1970.
41. Hoechst, US 1 435 955, 1972.
42. A. M. Kadhim, K. H. Mak, A. T. Peters, *J. Soc. Dyers Colour.* **98** (1982) 56 – 58.
43. Hoechst, DE-OS 2 447 024, 1974.
44. Hoechst, DE-AS 1 297 259, 1964.
45. Mitsubishi, JP 60 119 624, 1984.
46. Mitsubishi, JP 61 281 220, 1984.
47. K. Wojciechowski, J. Szadowski, *J. Ther. Anal.* **31** (1986) 297.
48. A. T. Peters, N. J. Bide, *Dyes Pigm.* **7** (1986) 237.
49. K. Wojciechowski, *Dyes Pigm.* **9** (1988) 401 – 417.
50. Nippon, JP 5 100 522, 1976.
51. Mitsubishi, DE-OS 2 258 545, 1971.
52. Mitsubishi, JP 5 246 184, 1973.
53. Mitsubishi, JP 5 045 827, 1973.
54. Mitsubishi, JP 5 132 216, 1975.
55. Canon, JP 108 188 908, 1981.
56. Canon, JP 108 190 036, 1981.
57. Minolta Camera, JP 63 287 859, 1987.
58. Nippon, JP 5 112 822, 1974.
59. BASF, DE-OS 2 503 732, 1975.
60. Hoechst, DE-OS 2 446 543, 1974.
61. Mitsubishi, JP 2 091 027, 1976.
62. Hoechst, EP 0 123 256, 1984.
63. BIOS, 1 484, 21.
64. BASF, DE 2 837 731, 1978.
65. BASF, DE-OS 2 316 536, 1978.
66. Hoechst, EP 0 205 980, 1986.
67. Hoechst, DE-OS 3 017 185, 1980.
68. Y. Nagao, T. Misono, *Bull. Chem. Soc. Jpn.* **54** (1981) 1191 – 1194.
69. Y. Nagao, T. Misono, *Dyes Pigm.* **5** (1984) 171 – 188.
70. Hoechst, DE-OS 3 436 206, 1984.
71. Nippon, JP 50 105 721, 1974.

72. BASF, DE-OS 2 545 633, 1975.
73. BASF, DE-OS 2 252 461, 1972.
74. *Friedländer* **14,** 485.
75. BASF, DE-OS 2 546 266, 1975.
76. Ciba-Geigy, EP 0 023 191, 1980.
77. F. Graser, E. Hädicke, *Liebigs Ann. Chem.* 1984, 483 – 494.
78. BASF, DE-OS 3 049 215, 1980.
79. H. Langhals, *Chem. Ber.* **118** (1985) 4641 – 4645.
80. H. Langhals, *Nachr. Chem. Tech. Lab.* **28** (1980) 716.
81. BASF, EP 260 648, 1986.
82. BASF, DE-OS 3 235 526, 1982.
83. BASF, DE-OS 3 339 540, 1983.
84. Xerox, US 069 544, 1987.
85. Alps Electric, JP 63 287 963, 1987.
86. A. Le Pape, E. Marechal, *Bull. Soc. Chim. Fr.* 1978, 263 – 270.

Naphthenic Acids → Metallic Soaps
Naphthenic Acids → Oil Refining
Naphthol → Naphthalene Derivatives

Naphthoquinones

Johann Grolig, Bayer AG, Leverkusen, Federal Republic of Germany

Rudolf Wagner, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. 1,4-Naphthoquinone 1
2.1. Properties 1
2.2. Production 1
2.3. Uses 2
2.4. Derivatives 2
3. Other Naphthoquinones 3
4. Naphthodiquinones 4
5. Polynaphthoquinone 4
6. References 4

1. Introduction

Of the five possible naphthoquinones, the isomers (**1**) – (**3**) have been synthesized in a pure form. Of these, 1,4-naphthoquinone (**1**) is produced on an industrial scale as an intermediate for anthraquinone. 1,5-Naphthaquinone (**4**) is known only in the form of its derivatives. The existence of 2,3-naphthoquinone (**5**) was confirmed by trapping it as its Diels – Alder adduct with cyclopentadiene. The synthesis of the unstable 1,7-isomer or of its derivatives has to date not been possible. In contrast, polynaphthoquinone (**8**) is a stable derivative of 1,7-naphthoquinone. The naphthoquinones (**6**) and (**7**) have also been synthesized. The industrially most important compounds are 1,4-naphthoquinone and 2-methyl-1,4-naphthoquinone.

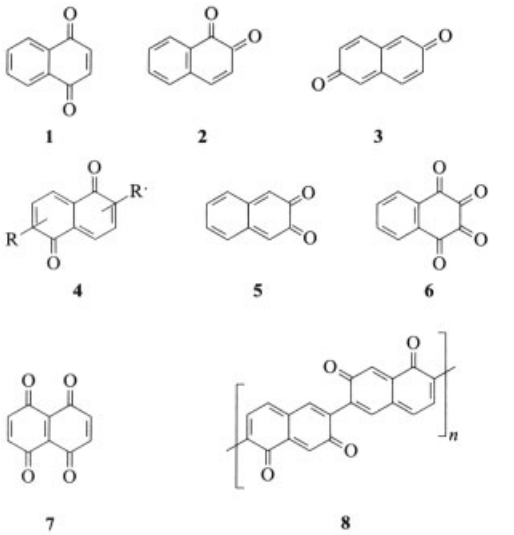

2. 1,4-Naphthoquinone

2.1. Properties

1,4-Naphthoquinone [*130-15-4*], α-naphthoquinone, 1,4-naphthalenedione, M_r 158.16, *mp* 128.5 °C, ϱ 1.42 g/cm^3, crystallizes as yellow needles. It has a penetrating odor, is steam volatile, and sublimes readily. It is insoluble in water but soluble in warm alcohols and hydrocarbons. Phenols and arylamines react at the 2-position. In acids and upon heating it condenses to form oligomers [1] and polymers. Its most important reaction is with butadiene to give tetrahydroanthraquinone, which is oxidized to anthraquinone (see Section 2.3).

2.2. Production

Gas-Phase Oxidation of Naphthalene. The most important process for the production of 1,4-naphthoquinone is the oxidation of naphthalene with air or gases containing molecular oxygen at 250 – 450 °C and 0.1 – 1 MPa, over a catalyst containing vanadium pentoxide.

Kawasaki Kasei Chemicals has been operating an anthraquinone plant with this process in Japan since 1980. Besides vanadium pentoxide, the catalyst also contains potassium sulfate, potassium pyrosulfate, and silicon dioxide [2], [3]. The latter also serves as a carrier and is obtained by treatment of aluminosilicates with acid [4]. As the naphthalene is converted nearly quantitatively, a recycling step is not necessary; 36 mol % of 1,4-naphthoquinone and 54 mol % of phthalic anhydride are obtained.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_067

The reaction is best performed under a slightly positive pressure of 0.3 – 0.8 MPa [5]. Although the oxidation is normally carried out in a fixed bed, a flow-bed variant has also been described [6]. Lifetime and selectivity of the catalysts are increased by the addition of sulfur-containing compounds [7]. Phthalic anhydride is always obtained in substantial amounts as a byproduct.

In the Kawasaki process the oxidation product is first treated with water in order to separate the phthalic anhydride as phthalic acid from the reaction mixture [8]. The naphthoquinone is extracted with xylene and the extract washed with dilute alkali to remove residual phthalic acid [9]. Then the naphthoquinone is crystallized from the solvent. High-purity naphthoquinone can be obtained by recrystallizing from xylene [10], or by sublimation [11]. Purification of the quinone can also be achieved by treating the hot xylene solutions with lime [12]. In some applications it is possible to use the crude naphthoquinone in form of the oxidation mixture.

Liquid-Phase Oxidation. 1,4-Naphthoquinone can alternatively be obtained in the liquid phase in greater than 90 % yield by oxidation of naphthalene or its derivatives. The following oxidizing agents have been used on a preparative scale: chromic acid [13], chromium trioxide [14], lead dioxide [15], manganese(II) sulfate [16], ammonium peroxodisulfate [17–19], tetrabutylammonium nitrate in nitric acid [20], and cerium(IV) compounds [21–27]. The oxidation with cerium(IV), in particular, gives high naphthoquinone selectivities, with yields as high as 98 % [23]. Phase-transfer catalysts increase the rate of reaction [19], [20], [27]. Chemical oxidizing reagents can be regenerated electrochemically. The regeneration of cerium(IV) compounds has been extensively investigated in a pilot plant [26]. 1-Naphthol can be catalytically oxidized to 1,4-naphthoquinone with oxygen in the presence of cobalt complexes (Salcomin) in yields of 90 % [28].

2.3. Uses

1,4-Naphthoquinone is industrially significant as an intermediate in the anthraquinone synthesis by Kawasaki (→ Anthraquinone, Chap. 3.3.) [29,30]. Starting from naphthalene, naphthoquinone is obtained by gas-phase oxidation, as described in Section 2.2. In the second step it is converted to 1,4,4a,9a-tetrahydroanthraquinone [*56136-14-2*] with butadiene in a Diels – Alder reaction [29], [30]. The Diels – Alder adduct is extracted as its sodium salt with dilute alkali. Unconverted naphthoquinone in the organic phase is fed back into the Diels – Alder reaction. In the third step the basic tetrahydroanthraquinone solution is oxidized with air to anthraquinone, which forms a precipitate and is filtered off. Kawasaki produces 3000 t/a anthraquinone with this process (→ Anthraquinone, Chap. 3.3.).

In a process developed by Bayer, the 1,4-naphthoquinone is not isolated after the naphthalene oxidation step. Instead, it is reacted with butadiene and oxidized to anthraquinone in the mixture of naphthalene and phthalic anhydride. The products are then separated by distillation [31–34].

1,4-Naphthoquinone has been described as a polymerization regulator [35] and as a hardener for photochemically cross-linked polyesters [36]. It is used as a corrosion inhibitor [37] and as a stabilizer for transformer oils [38]. It reacts with cyclic aryl phosphinic acid esters to give 1,4-naphthalenediol adducts, which can be used for the production of thermally stable, flame-resistant polyesters [39].

2.4. Derivatives

2-Methyl-1,4-naphthoquinone [*58-27-5*], menadione 2-methyl-1,4-naphthalenedione, vitamin K3, M_r 172.20, *mp* 105 – 106 °C is two to three times more effective as a blood coagulator than vitamin K1 (2-methyl-3-phytyl-1,4-naphthoquinone, see → Vitamins). It is obtained from 2-methylnaphthalene in 70 – 90 % yield by oxidation with chromic acid [40], [41]. Slightly lower yields are obtained with chromium trioxide in conjunction with crown ethers [42]. The gas-phase oxidation of 2-methylnaphthalene results in maximally 40 % yields of the quinone [43].

2-Chloromethyl-1,4-naphthoquinone [*43027-41-4*], 2-chloromethyl-1,4-naphthalenedione, and its derivatives are highly potent cytostatic compounds. It is obtained by the oxidation of 2-chloromethylnaphthalene with chromium trioxide [44].

2,3-Dichloro-1,4-naphthoquinone [*117-80-6*], 2,3-dichloro-1,4-naphthalenedione, dichlone, phygon, M_r 227.05, *mp* 193 °C, is an important intermediate in the synthesis of dyes. It can also be converted to pigments of exceptional stability, which are suitable for integration into high performance membranes [45]. It is also used for conserving grain, as an algicide [46], and as an additive in dye binders [47]. This quinone is obtained by oxidation of 4-hydroxynaphthalene-1-sulfonic acid [*84-87-7*] with potassium chlorate in hydrochloric acid [48], or by direct chlorination of 1,4-naphthoquinone [49–51]. Yields of up to 96 % can be obtained by the latter method if the chlorination is carried out in carbon tetrachloride at 50 °C [51].

5-Nitro-1,4-naphthoquinone [*17788-47-5*], 5-nitro-1,4-naphthalenedione, M_r 203.15, *mp* 167 °C, is obtained in 70 % yield by nitration of 1,4-naphthoquinone with a mixture of highly concentrated nitric and sulfuric acid [52]; 6-nitro-1,4-naphthoquinone [*58200-82-1*] is formed as a byproduct. Nitronium tetrafluoroborate ($NO_2^+BF_4^-$) is also a suitable nitrating reagent [53]. The 5-nitro-1,4-naphthoquinone is also accessible by the electrochemical oxidation of 1-nitronaphthalene in a mixture of a mineral acid and a water-soluble organic solvent [54]. Nearly quantitative yields can be obtained through the reduction of 5-nitro-1,4-naphthoquinone monooxime with sulfur dioxide and phosphorus in the presence of hydrogen iodide, followed by oxidation of the resulting isomeric 1-aminonitronaphthols [55].

The reaction of 5-nitro-1,4-naphthoquinone with butadiene and subsequent oxydehydrogenation leads to 1-nitroanthraquinone [*82-34-8*] [56–58], which is the precursor of the important dye intermediate 1-aminoanthraquinone [*25620-59-1*].

Hydroxy Derivatives. A few monohydroxy derivatives of 1,4-naphthoquinone occur in nature as dyes in plants: for example, 5-hydroxy-1,4-naphthoquinone [*481-39-0*] (juglone) and 2-hydroxy-1,4-naphthoquinone [*83-72-7*] (lawsone). 2-Alkyl-3-hydroxy-1,4-naphthoquinones are used as antimalarial agents [59] and insecticides [60]. Derivatives of naphthazarin (5,8-dihydroxy-1,4-naphthoquinone) [*475-38-7*] are dyes (vat dyes, alkannins). Naphthazarin can be made from 1,5-dinitronaphthalene [*605-71-0*] by conversion with sulfur and sulfuric acid [61] via the intermediate 5-hydroxy-8-amino-1,4-naphthoquinone [*68217-36-7*]. Alternatively, it can be prepared from the reaction of dichloromaleic anhydride [*1122-17-4*] with 1,4-dimethoxybenzene [62]. The resulting 2,3-dichloronaphthazarin [*14918-69-5*] is converted to 2,3-dihydronaphthazarin [*4988-51-6*] with tin(II) chloride in hydrochloric acid in the second step and then converted to naphthazarin in alkaline solution with an overall yield of 66 %.

3. Other Naphthoquinones

1,2-Naphthoquinone [*524-42-5*], α-naphthoquinone, 1,2-naphthalenedione, M_r 158.16, *mp* 145 – 147 °C, crystallizes as yellow needles, is odorless, is involatile in steam, and is a relatively weak oxidizing agent. It is obtained in 80 % yield when 2-naphthol is oxidized by air in the liquid phase with copper(I) chloride as catalyst [63]. Slightly lower yields are obtained when periodic acid and dimethyl formamide [64], iodosobenzene [65], or benzene seleninic anhydride is used as oxidant [66]. A 95 % yield can, however, be obtained with tetramethyloxopiperidinium chloride [*26864-01-7*] at low temperature [67]. The quinone can also be obtained by the oxidation of 1-amino-2-hydroxynaphthalene hydrochloride [*1198-27-2*] with iron(III) chloride in hydrochloric acid [68].

Only the derivative 6-bromo-1,2-naphthoquinone [*6954-48-9*] (bonaphthon), which exhibits activity against type A influenza viruses, is of any practical interest [69].

2,6-Naphthoquinone [*613-20-7*], M_r 158.16, *mp* 135 °C (decomp.), forms red prisms. It is a stronger oxidizing agent than the other naphthoquinones. It is obtained by the oxidation of 2,6-naphthalenediol [*581-43-1*] with lead dioxide in benzene [70].

1,5-Naphthoquinone [*51583-62-1*] is not known in the pure form. Its derivative, 3,7-di-*t*-butyl-1,5-naphthoquinone [*54532-93-33*], is obtained in virtually quantitative yield by oxidizing the corresponding hydroquinone with dichlorodicyano-*p*-benzoquinone [71]. The naphthazarin isomer 4,8-dihydroxy-1,5-naphthoquinone [*6251-03-2*] is a starting material for thermally stable dyes [72].

2,3-Naphthoquinone [*4939-92-8*] has also not been isolated. The existence of this unstable quinone was verified by trapping it as the Diels – Alder adduct with cyclopentadiene during the oxidation of 2,3-naphthalenediol with potassium periodate [73].

1,7-Naphthoquinone [*46001-16-5*] is also very unstable and has not been isolated. The derivative, 3,6-di-*t*-butyl-8-methyl-1,7-naphthoquinone [*83021-64-1*], can be obtained in impure form in solution during the oxidation of the corresponding hydroquinone with dichlorodicyano-*p*-benzoquinine [74].

4. Naphthodiquinones

1,2,3,4-Naphthodiquinone [*30266-58-1*], 1,2, 3,4-naphthalenetetrone, oxolene, M_r 188.10, *mp* 131 °C, ϱ 1.45 g/cm^3, forms colorless crystals. It is obtained from 2,3-dihydroxy-1,4-naphthoquinone through oxidation with nitric acid in glacial acetic acid [75].

1,4,5,8-Naphthalenetetrone [*23077-93-2*], M_r 188.10, *mp* 220 °C, ϱ 1.45 g/cm^3, is obtained as pale yellow prisms by the oxidation of 5,8-dihydroxy-1,4-naphthoquinone with lead tetraacetate [76] or with iodophenyl-bis(trifluoroacetate) in 81 % yield [77].

5. Polynaphthoquinone

Polynaphthoquinone [*38830-94-3*], 1,7-naphthalenedione polymer, is a black powder which melts above 360 °C and sublimes at 450 °C. It is made by oxidizing 1,7-naphthalenediol [*575-38-2*] with dilute nitric acid in ethanol, followed by heating in a stream of air at 350 °C [78], [79]. The polyquinone is suitable as a catalyst for the oxydehydration of alkanes, alkenes, and alcohols. Its reactivity can be greatly enhanced through the addition of Lewis acids [78], [79].

6. References

1. H. E. Höberg, *Acta Chem. Scand.* **26** (1972) 309.
2. Kawasaki, DE-OS 3 033 341, 1979 (R. Matsuura et al.).
3. Kawasaki, JP 8 259 827, 1980.
4. Kawasaki, JP 7 822 559, 1973 (R. Matsuura et al.).
5. Bayer, DE-OS 2 437 221, 1974 (M. Martin, G. Scharfe); DE 2 453 232, 1974 (M. Martin, W. Schmidt, G. Scharfe).
6. Sumikin Coke, JP 8 805 051, 1986 (T. Nishizaki, R. Minami).
7. Bayer, DE-OS 2 234 306, 1972 (H. Dohm, K. Morgenstern); DE-OS 2 312 838, 1973 (R. Wiemers, K. Morgenstern, L. Müller, H. Dohm); Japan Distillation Co., DE-OS 2 506 809, 1974 (Y. Yokoyama, J. Yoshikawa, H. Ota, C. Ichikawa).
8. Kawasaki, JP 7 783 443, 1975 (R. Matsuura et al.).
9. Kawasaki, GB 2 039 897, 1978.
10. Kawasaki, JP 8 385 835, 1981.
11. Rütgerswerke, DE 1 232 943, 1965 (K. Rühl, L. Rappen).
12. Kawasaki, JP 86 155 348, 1984 (A. Matsura, T. Komatsu, T. Sumino).
13. Great Lakes Carbon Corp., US 3 681 401, 1966 (L. A. Joo, L. A. Bryan).
14. Braude, Fawcett, *Org. Synth. Coll.,* vol. 4 (1963) 698.
15. Nippon Kakayu Co., JP 7 375 555, 1972.
16. M. Periasamy, M. V. Bhatt, *Tetrahedron Lett.* 1978, no. 46, 4561.
17. ICI, DE-OS 2 301 803, 1972 (R. A. Rennie, R. Campbell).
18. J. Skarzewsky, *Tetrahedron* **40** (1984) 4997.
19. E. V. Dehmlow, J. K. Makrandi, *J. Chem. Res. Synop.* 1986, 32.
20. D. Pletcher, E. M. Valdes, *J. Chem. Res. Synop.* 1987, 386.
21. ICI, DE-OS 1 804 727, 1967 (R. A. Rennie, L. Campbell).
22. M. Periasamy, M. V. Bhatt, *Synthesis* 1977, 330.

23. Kawasaki, US 4 536 377, 1983 (T. Komatsu et al.).
24. Diamond Shamrock Corp., EP-A 75 828, 1981 (E. A. Mayeda, D. W. Abrahamson).
25. Grace, US 4 670 108, 1986 (R. P. Kreh, R. M. Spotnitz).
26. I. M. Dalrymple, J. P. Millington, *J. Appl. Electrochem.* **16** (1986) 885 – 893.
27. D. Pletcher, E. M. Valdes, *Electochimica Acta* no. 4, 509 – 515 (1988).
28. Shell Int. Res., DE-OS 2 460 665, 1974 (A. J. de Jong, R. van Helden).
29. Kawasaki, JP 5 108 256, 1974; JP 5 108 257, 1974; JP 5 322 559, 1978.
30. Kawasaki, US 4 412 954, 1981 (T. Kamatsu, K. Usui).
31. Bayer, DE-OS 2 532 388, 1975 (J. Priemer, N. Schenk, J. Krekel, W. Schwerdtel).
32. Bayer, DE-OS 2 218 316, 1972 (J. Grolig, G. Scharfe).
33. Bayer, DE-OS 2 245 555, 1972 (G. Scharfe, J. Grolig).
34. Bayer, DE-OS 2 532 450, 1975 (J. Priemer, N. Schenk, M. Martin, W. Swodenk).
35. US Rubber Co., US 2 457 701, 1945 (Ch. D. McCleary). F. A. Bovey, J. M. Kolthoff, *Chem. Rev.* **42** (1948) 491.
36. Akzo GmbH, DE-OS 2 325 179, 1972 (L. Roskott, M. A. A. Groenendaal).
37. Rohm & Haas, US 3 960 671, 1974 (J. S. Clovis, J. Dohling).
38. S. N. Litvinenko, L. K. Mushkalo, A. V. Stetsenko, A. K. Tyltin, *Khim. Tekhnol. Topl. Masel* **18** (1973) no. 8, 41.
39. Japan Exlan Co., JP 86 261 320 A2, 1985 (T. Imamura et al.); JP 86 268 691 A2, 1985 (T. Imamura, T. Matsumoto, K. Matsuzawa).
40. Velsicol Corp., US 2 402 226, 1943 (J. Hyman, C. F. Peters).
41. BASF, DE-OS 2 952 709, 1979 (H. Puetter, W. Bewert).
42. M. Juaristi, J. M. Aizpurua, B. Lecea, C. Palomo, *Can. J. Chem.* **62** (1984) 2941.
43. S. K. Ray, G. S. Murty, H. S. Rao, *Proc. Symp. Chem. Oil Coal* 1969, 320 – 329.
44. A. J. Lin, A. S. Sartorelli, *Biochem. Pharmacol.* **2** (1976) no. 2, 206; *J. Med. Chem.* **1** (1976) no. 11, 1336; *J. Med. Chem.* **16** (1973) no. 11, 1268.
45. NEC Corp., JP 8 515 458, 1983.
46. US-Rubber Co., GB 657 977, 1949 (T. A. Clayton); *Chem. Eng. News* 1956, 3706.
47. Asahi Chem. Ind. Co., JP 7 447 426, 1969.
48. J. Ullmann, M. Ettich, *Ber. Dtsch. Chem. Ges.* **54** (1922) 259.
49. Österr. Stickstoffwerke, AT 190 921, 1955 (A. Wagner).
50. Kawasaki, JP 7 898 943, 1977 (M. Matsuura et al.).
51. Seitetsu Chem. Ind. Co., JP 76 113 859, 1975 (H. Watanabe et al.).
52. Mitsui Toatsu, DE-OS 2 509 819, 1974 (A. Iwamura et al.); Nippon Shokubai, DE-OS 2 436 157, 1974 (A. Fukui et al.).
53. S. Chatterjee, W. D. Kwalwasser, *J. Chem. Soc. Perkin Trans. 2* 1974, no. 8, 873.
54. Sumitomo Chem. Co., JP 8 076 083, 1978.
55. Bayer, DE-OS 2 321 003, 1973 (F. Dürholz, R. Pütter, A. Vogel); DE-OS 2 359 950, 1973 (F. Dürholz, R. Pütter).
56. Ciba Geigy, BE 8 22 754, 1973; BE 8 22 755, 1973.
57. Mitsui Toatsu, DE-AS 2 539 631, 1975 (Y. Torisu et al.).
58. Mitsui Toatsu, GB 1 468 480, 1974.
59. G. Fawaz, L. F. Fieser, *J. Am. Chem. Soc.* **72** (1950) 996, 3999.
60. Du Pont, US 2 572 946, 1949 (M. Paulshock).
61. FIAT-Berichte 1313 II, 208, 1948.
62. J. R. Lewis, J. Paul, *Z. Naturforsch. B: Anorg. Chem., Org. Chem.* **32 B** (1977) no. 12, 1473.
63. G. Wurm, B. Goessler, *Arch. Pharm. (Weinheim, Ger.)* **320** (1987) no. 6, 564.
64. A. V. Pinto, V. F. Ferreira, M. do. C. F. R. Pinto, *Synth.Commun.* **15** (1985) no. 13, 1177.
65. D. H. R. Barton et al., *Tetrahedron Lett.* **23** (1982) no. 9, 957.
66. D. H. R. Barton et al., *J. Chem. Soc. Perkin Trans 1* 1981, 1473.
67. D. H. Hunter, D. H. R. Barton, W. S. Motherwell, *Tetrahedron Lett.* **25** (1984) no. 6, 603.
68. L. F. Fieser, *Org. Synth. Coll.,* vol. 2 (1943) 430.
69. M. P. Zykov et al., *Vopr. Virusol* 1975, no. 1, 58 – 62. G. N. Pershin et al., *Farmakol. Toksikol. (Moscow)* **38** (1975) no. 1, 69 – 73.
70. R. Kuhn, J. Hammer, *Chem. Ber.* **83** (1950) 413.
71. H. L. K. Schmand, P. Boldt, *J. Am. Chem. Soc.* **97** (1975) no. 2, 447.
72. Mitsubishi Chem. Ind. Co., JP 85 151 098, 1984 (T. Niwa, Y. Murata, S. Maeda).
73. V. Horak et al., *Tetrahedron Lett.* **22** (1981) no. 37, 3577.
74. P. Boldt et al., *J. Org. Chem.* **48** (1983) 2814.
75. T. H. Zincke, A. Ossenbeck, *Justus Liebigs Ann. Chem.* **307** (1899) 1.
76. K. Zahn, P. Ochwat, *Justus Liebigs Ann. Chem.* **462** (1928) 86.

77. S. Yoshino, K. Hayakawa, K. Kanematsu, *J. Org. Chem.* **46** (1981) 3841.
78. Y. Iwasawa et al., *J. Chem. Soc. Faraday Trans 1* **68** (1972) 1617; **70** (1974) 193; *J. Catal.* **31** (1973) 444.
79. Y. Iwasawa, S. Ogasawara, *J. Catal.* **37** (1975) 148–157.

Naphthylamines → Naphthalene Derivatives

Natural Gas

Georg Hammer, (Mobil Erdgas – Erdöl GmbH), Celle, Federal Republic of Germany (Chap. 1)

Torsten Lübcke, Mobil Erdgas – Erdöl GmbH, Celle, Federal Republic of Germany (Chap. 2)

Roland Kettner, Mobil Erdgas – Erdöl GmbH, Celle, Federal Republic of Germany (Chap. 2)

Mark R. Pillarella, Air Products and Chemicals, Inc., Allentown, PA, United States (Chap. 3)

Herta Recknagel, Ruhrgas AG, Essen, Federal Republic of Germany (Chap. 4)

Axel Commichau, Mobil Europe Gas Inc., The Hague, Netherlands (Chap. 5)

Hans-Joachim Neumann, German Petroleum Institute, Clausthal – Zellerfeld, Federal Republic of Germany (Chap. 6)

Barbara Paczynska-Lahme, Consultant, Osterode, Federal Republic of Germany (Chap. 6)

Related Articles → Liquefied Petroleum Gas and → Methane are separate keywords. For Upstream Processing of Natural Gas, see → Oil Refining and → Gas Production; for Substitute Natural Gas, see → Gas Production.

1. Introduction . . . 2
1.1. Composition . . . 2
1.2. Physical Properties . . . 3
1.3. Combustion Properties . . . 7
1.3.1. Calorific Value . . . 7
1.3.2. Wobbe Index . . . 8
2. Treating . . . 8
2.1. Objectives of Natural Gas Treating . . . 9
2.2. Natural Gas Wells and Field Installations . . . 10
2.3. Processing Steps During Production . . . 10
2.3.1. Use of Sulfur Solvents . . . 10
2.3.2. Removal of Mercury . . . 12
2.3.3. Dehydration . . . 13
2.3.4. Removal of Hydrocarbons . . . 16
2.4. Removal of Carbon Dioxide and Sulfur Components . . . 17
2.4.1. Absorption Processes . . . 17
2.4.1.1. Physical Absorption Process . . . 17
2.4.1.2. Chemical Absorption Processes . . . 18
2.4.1.3. Physical – Chemical Absorption Processes . . . 19
2.4.2. Liquid Oxidation Processes . . . 20
2.4.3. Adsorption Processes . . . 22
2.4.3.1. Adsorption of Carbon Dioxide . . . 22
2.4.3.2. Adsorption of Hydrogen Sulfide . . . 22
2.5. Membrane Processes . . . 23
2.6. Cryogenic Natural Gas Separation . . . 23
2.7. Recovery of Sulfur . . . 23
2.7.1. Claus Process . . . 23
2.7.2. Tail-Gas Cleanup Processes . . . 25
2.7.3. Incineration . . . 26
2.8. Example of an Integrated Plant . . . 26
3. Liquefaction . . . 29
3.1. Physical Properties of LNG . . . 30
3.2. History . . . 30
3.3. Production . . . 31
3.3.1. Baseload Liquefaction Processes . . . 31
3.3.2. Peak-Shaving Liquefaction Processes . . . 34
3.4. Storage . . . 35
3.5. Transportation . . . 35
3.5.1. Peak Shaving . . . 35
3.5.2. Baseload . . . 36
3.6. Safety Aspects . . . 37
4. Transmission, Storage, and Distribution . . . 37
4.1. Transportation . . . 38
4.1.1. Pipeline Transmission . . . 38
4.1.2. LNG Transportation . . . 39
4.1.3. Gas Transmission Pipeline Design . . . 39
4.1.4. Increase in Capacity . . . 40
4.1.5. Transient Flow in Gas Transmission Systems . . . 41
4.2. Storage . . . 42
4.2.1. Determination of Storage Requirements . . . 42
4.2.2. Storage Systems . . . 43
4.2.3. Optimized Storage Planning . . . 45
4.3. Distribution . . . 46
5. Economic Aspects . . . 47
6. Testing and Analysis . . . 48
6.1. Analysis of Natural Gas Composition . . . 49
6.1.1. Sampling . . . 49
6.1.2. Gas Chromatography . . . 49
6.1.3. Analysis of Nonhydrocarbons . . . 49

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_073.pub2

6.2. Determination of Physical Properties 51
6.2.1. Density 51
6.2.2. Calorific Value 51
6.2.3. Molecular Mass 51
6.2.4. Dew Point 51
7. Acknowledgement 52
8. References 52

Natural gas is a mixture of gaseous hydrocarbons with impurities mainly made of nonhydrocarbons. It occurs in subsurface deposits of many geologic formations. Natural gas is recovered with wells that are drilled into these deposits and is produced at the well site in compositions of considerable variety. The most important method for chemical analysis of natural gas is gas chromatography. The *Wobbe index* is the most important physical characteristic for the classification of natural gases. Natural gas is primarily employed as clean heating fuel in industry, power plants, commercial installations, and households. The objectives of natural gas treating are the adjustment to the required quality standards and the recovery of byproducts. Removal of the sour gas components hydrogen sulfide and carbon dioxide is one of the most common aims of natural gas treating. Of the gas traded across international borders, 75 % is shipped by pipelines, and 25 % is carried by liquified natural gas tankers. A large share of internationally traded natural gas is normally sold by producers under long-term contracts of up to 35 years' duration in order to compensate for the heavy front-end investment in production, purification, and transportation facilities.

1. Introduction

Natural gas is a mixture of gaseous hydrocarbons with varying quantities of nonhydrocarbons, which are normally considered impurities. Natural gas is often associated with liquid hydrocarbons (i.e., condensate and crude oil).

Natural gas formed during many epochs in the history of earth and accumulated after migrating from its place of origin to subsurface deposits in porous or fractured layers of many geologic formations (→ Oil Refining). Natural gas is recovered with wells that are drilled into these deposits. Worldwide annual production amounts to ca. 2.8×10^{12} m^3. Estimated gas reserves are on the order of 150×10^{12} m^3 (see Chap. 5).

1.1. Composition

Usual Composition. Natural gas is composed mainly of methane, ethane, propane, and butane. In addition, it usually contains minor quantities of heavier hydrocarbons and varying amounts of gaseous nonhydrocarbons such as nitrogen, carbon dioxide, and hydrogen sulfide. A typical composition is given in Table 1.

Table 1. Typical composition range of natural gas

Compound	Molar fraction
Hydrocarbons	
Methane	0.75 – 0.99
Ethane	0.01 – 0.15
Propane	0.01 – 0.10
n-Butane	0.00 – 0.02
Isobutane	0.00 – 0.01
n-Pentane	0.00 – 0.01
Isopentane	0.00 – 0.01
Hexane	0.00 – 0.01
Heptane plus higher hydrocarbons	0.00 – 0.001
Nonhydrocarbons	
Nitrogen	0.00 – 0.15
Carbon dioxide	0.00 – 0.30
Hydrogen sulfide	0.00 – 0.30
Helium	0.00 – 0.05

Main Impurities. In many areas the natural gas recovered is *sour*, which means that it contains appreciable quantities of hydrogen sulfide and carbon dioxide. Although larger quantities of carbon dioxide are undesirable because they make the transport of natural gas more costly, diminish its combustion qualities, and have a corrosive effect under certain conditions, carbon dioxide can be tolerated in percentage levels. In contrast, hydrogen sulfide concentrations in salable gas must be reduced below 3 ppm because of its high toxicity. Furthermore, it is very corrosive in combination with water and, when burned, produces sulfur dioxide, which is also toxic and corrosive. Therefore, sour gas is generally desulfurized or *sweetened* to reduce the hydrogen sulfide content to traces (clean gas).

Higher Hydrocarbons and Water Vapor. Natural gas may contain larger quantities of higher hydrocarbons than shown in Table 1 (so-called *rich* or *wet* gas) and, in addition, is saturated with water vapor under natural conditions. These constituents can condense at lower temperature and higher pressure and thereby impair the transport and use of natural gas. Under certain conditions, water and gas can form icelike solids, so-called hydrates (see Section 1.2). Therefore, the proportion of condensable constituents is normally decreased substantially by refrigeration, adsorption, or absorption processes before the natural gas (then called *lean* or *dry* gas) is transmitted through pipelines or liquefied.

1.2. Physical Properties

Molar Mass. Table 2 shows the molar mass of the commonly occurring components of natural gases. The apparent molar mass M_a of a gas mixture such as natural gas is equal to the sum of the molar fraction y_i times the molar mass M_i of each component:

$$M_a = \Sigma y_i M_i \tag{1}$$

Molar Volume and Ideal-Gas Law. The volume occupied by an ideal gas depends on the temperature, pressure, and number of molecules present. It does not depend on the type of molecules (ideal-gas law).

The volume of 1 kmol of an ideal gas at STP (273.15 K, 0.101325 MPa) can be calculated, by using the ideal-gas law, to be 22.414 m^3. The effective molar volumes at STP for common natural gas components are shown in Table 2. The calculation may also be employed for gas mixtures if the apparent molar mass, $M_a = \sum y_i M_i$, is used in computing the number of kilomoles n.

Density and Relative Density. The density of several components of natural gas is given in Table 2. The density of a gas mixture can be calculated by the same procedure described in Equation (1). For deviating temperature and pressure conditions, or when the ideal-gas law cannot be applied, the density of a gas under any conditions can be calculated from the real-gas law:

$$pV = znRT \tag{2}$$

where z is the so-called compressibility factor (see page 5).

With Equation (2) the density of a gas can be calculated from the relationship

$$\varrho = pM/zRT \tag{3}$$

where V and n in Equation (2) are replaced by

$$\varrho = nM/V \tag{4}$$

For a gas mixture such as natural gas, the molar mass M in Equation (3) is replaced by the apparent molar mass, M_a, and z is the compressibility factor of the gas mixture at the temperature and pressure conditions of interest.

The *relative density d* (see Table 2) (traditionally designated specific gravity) of a gas is the ratio of its density to the density of dry air at standard conditions:

$$d = \frac{\varrho_{gas}}{\varrho_{air}} \tag{5}$$

For a gas mixture, the relative density can also be calculated from the apparent molar mass M_a of the gas mixture divided by the molar mass of air (28.963):

$$d = \frac{M_a}{28.963} \tag{6}$$

Critical Temperature and Pressure of Natural Gas Components. The critical temperature T_c is the temperature above which a gas cannot be liquefied even if high pressure is applied. At the critical temperature a gas can be liquefied by application of a certain minimum pressure, the critical pressure p_c. The volume of 1 kg of a gas at this critical point is the critical volume V_c; the reciprocal value of V_c is the critical density ϱ_c.

The critical constants for the primary constituents of natural gases are given in Table 3. The critical temperature and pressure are constants for each species.

Principle of Corresponding States. Properties of gases can be correlated if they are compared at "corresponding" values of absolute temperature and pressure. The critical state of a gas is used as reference point, and so-called reduced values are obtained. The following terms are employed:

$$p_r = p/p_c, T_r = T/T_c, \text{ and } V_r = V/V_c$$

Table 2. Physical properties of natural gas components at STP [4]

Compound	Molar mass M, kg/kmol	Molar volume V, m^3/kmol	Density ϱ, kg/m^3	Relative density d (air = 1)
Methane	16.043	22.360	0.7175	0.5549
Ethane	30.069	22.191	1.355	1.048
Propane	44.096	21.928	2.011	1.555
n-Butane	58.123	21.461	2.708	2.094
Isobutane	58.123	21.550	2.697	2.086
n-Pentane	72.150 *	20.90 *	3.452 *	2.670 *
Isopentane	72.150 *	21.06 *	3.426 *	2.650 *
n-Hexane	86.177 *	20.10 *	4.29 *	3.315 *
n-Heptane	100.203 *	18.3 *	5.48 *	4.235 *
Nitrogen	28.0134	22.403	1.2504	0.9671
Carbon dioxide	44.0098	22.261	1.9770	1.5290
Hydrogen sulfide	34.076	22.192	1.5355	1.1875
Helium	4.0026	22.426	0.17848	0.1380

* Compound is in liquid state at STP.

Table 3. Critical constants of natural gas components [5]

Compound	p_c, MPa	T_c, °C	V_c, 10^{-3} m^3/kg	ϱ_c, kg/m^3
Methane	4.60	−82.55	6.1709	162
Ethane	4.88	32.25	4.9218	203
Propane	4.24	96.65	4.6035	217
n-Butane	3.80	152.05	4.3872	228
Isobutane	3.65	134.95	4.5248	221
n-Pentane	3.37	196.45	4.2134	237
Isopentane	3.38	187.25	4.2411	236
n-Hexane	2.97	234.25	4.2934	233
n-Heptane	2.73	267.05	4.3112	232
Nitrogen	3.39	−146.95	3.1949	313
Carbon dioxide	7.38	31.05	2.1359	468
Hydrogen sulfide	8.94	100.05	2.8903	346
Helium	0.227	−267.96	14.3143	69.9

where T_c, p_c, and V_c are the critical point conditions, and T, p, and V are the values of the system, respectively.

If the principle of corresponding states is applied to a mixture of gases, "pseudocritical" temperatures and pressures are defined as the molar average critical temperatures and pressures of the components:

$$p_{pc} = \sum y_i p_{ci}$$

$$T_{pc} = \sum y_i T_{ci}$$

where the subscript pc designates the pseudocritical values of the gas mixture, *ci* the critical values of the components, and y_i is the molar fraction of component *i*. The pseudocritical pressure and temperature are then used to determine the pseudoreduced conditions:

$$p_{pr} = p/p_{pc}$$

$$T_{pr} = T/T_{pc}$$

Although they are convenient for correlating purposes the pseudocritical values have no physical significance.

Viscosity. For an ideal gas, the pressure and temperature dependence of viscosity is opposite to that of liquids; the viscosity of an ideal gas increases with increasing temperature and is independent of pressure. Actual hydrocarbon gases, however, deviate from ideal behavior and approach that of liquids; their viscosity increases with increasing pressure, and decreases with increasing temperature at intermediate or high pressure.

Figure 1 is employed for estimating the dynamic viscosity u_{ga} of natural gases composed

primarily of hydrocarbons at atmospheric pressure [6]. The presence of carbon dioxide, hydrogen sulfide, and nitrogen increases the viscosity. The viscosity u_g at any desired temperature and pressure can be obtained from Figure 2, where the ratio u_g/u_{ga} is plotted as a function of the pseudoreduced temperature and pressure.

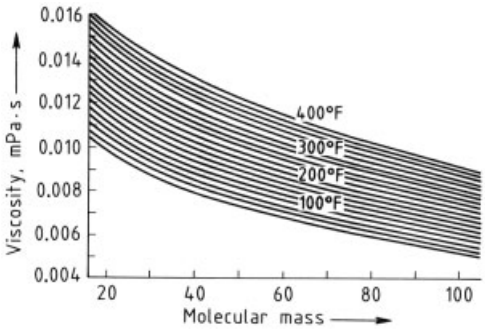

Figure 1. Viscosity of paraffin hydrocarbon gases at 101.3 kPa as a function of molar mass [6]
[(°F−32)/ 1.8 = °C] (Copyright Society of Petroleum Engineers, 1954)

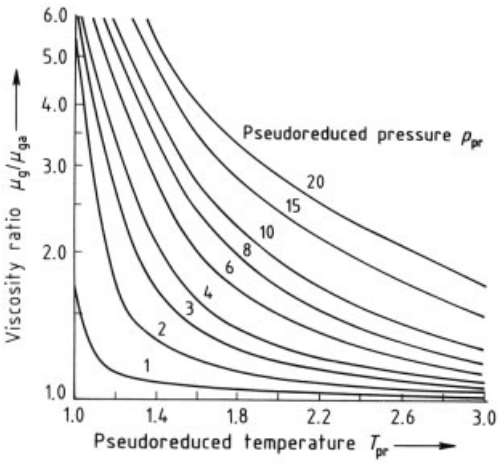

Figure 2. Viscosity ratio as a function of pseudoreduced temperature [6] (Copyright Society of Petroleum Engineers, 1942)

Compressibility of Real Gases. At higher pressure the attractive or repulsive forces between gas molecules become important because the molecules possess finite volumes. The assumptions for ideal-gas behavior are then no longer valid; consequently, the real-gas law (Eq. 2) was formulated by introducing a proportionality term (the compressibility factor) to the ideal-gas law. Compressibility factors for natural gases are determined by laboratory measurements, estimated by correlations, or calculated by equations of state.

Figure 3, which was developed by Standing and Katz [7] from data collected on methane and natural gases in the range of up to 56 MPa and 121 °C, presents a correlation between compressibility factors and pertinent pseudoreduced temperatures and pressures. Correlations are also available for pressures and temperatures above these ranges [2, 3].

The specific $p - V - T$ relationship of fluids is described mathematically by equations of state. These equations number in the hundreds, ranging from those for a specific pure compound to generalized forms that claim to relate the properties of multicomponent mixtures. Well-known examples are the *van der Waals equation* [8] with the modifications of Redlich and Kwong [9], Soave [10], and Peng and Robinson [11] and the *virial equation*. The van der Waals type of equation of state can be solved directly for the compressibility factor z. The most reliable methods for typical natural gases are those of Robinson and Jacoby [12], and Hall and Yarborough [13].

Well-known *virial equations* are the *Benedict – Webb – Rubin* equation [14] and the *Starling – Hon* [15] extension of the Benedict – Webb – Rubin equation. The European Gas Research Group (GERG) and the van der Waals Institute at Amsterdam carried out an extensive research project [16] to develop a virial equation of state for the accurate prediction of compressibility factors for natural gas mixtures. They used the virial equation solved for the compressibility factor z and truncated after the third term:

$$z = 1 + B(T)\,\varrho_M + C(T)\,\varrho_M^2$$

where ϱ_M is the molar density, $\varrho_M = 1/V_M$, and B and C are the second and third virial coefficients, which depend on temperature and gas composition. With this equation, compressibility factors can be predicted within the range 0 to 12 MPa and −8 to 62 °C, which deviate $<$ 0.1 % from experimental data.

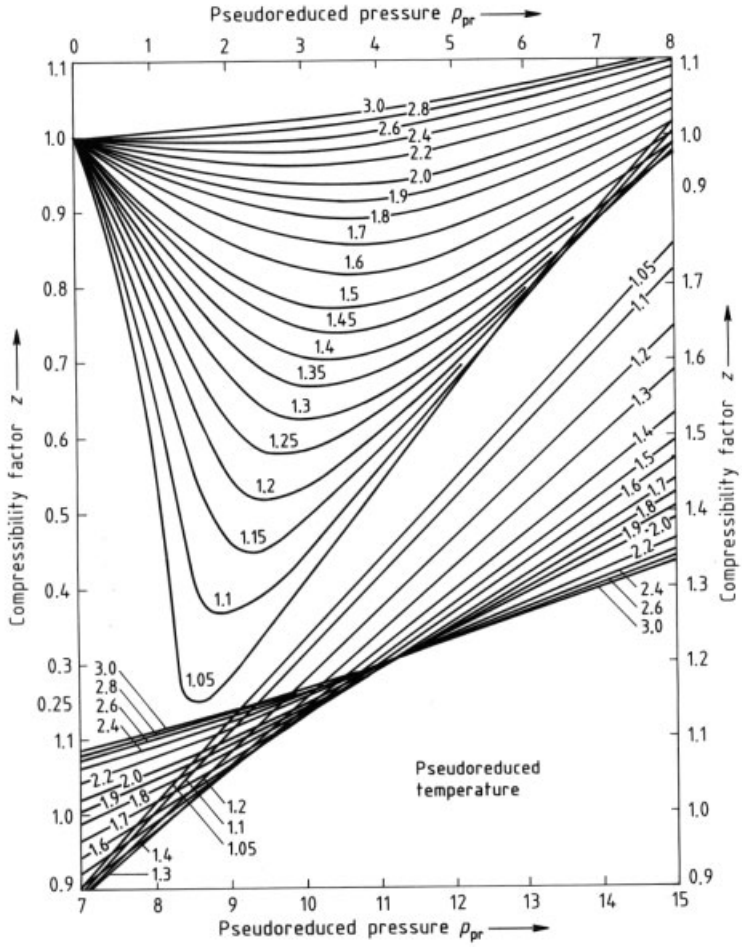

Figure 3. Compressibility factor of natural gases as a function of pseudoreduced pressure and temperature [7] (Copyright Society of Petroleum Engineers, 1942)

Condensation Behavior of Heavier Hydrocarbons.

Phase Diagram. Figure 4 shows the phase diagram of a natural gas system. It represents the phase changes with temperature and pressure for a particular constant composition, and designates the pressure and temperature regions in which only gas or liquids can occur and those in which they coexist at equilibrium. The two-phase region or "envelope" is enclosed on the left side by the bubble point curve up to the critical point C and from there on, via the points of maximum pressure C_p and maximum temperature C_T, by the dew point curve.

The bubble point curve designates the points at which the first infinitesimal bubble of gas is formed from the liquid phase, and the dew point curve indicates the points at which the first infinitesimal drop of liquid condenses from the gaseous phase. The critical point C (juncture of bubble and dew point curves) is the point at which all intensive properties of gas and liquid phases are identical. The pressure p_c and the temperature T_c at this point are the critical values of the gas mixture. The *cricondenbar* point C_p is defined as the maximum pressure at which two phases can exist in equilibrium. In natural gas systems this point lies right of the critical point C. The *cricondentherm* point C_T is defined as the maximum temperature at which two phases can exist in equilibrium.

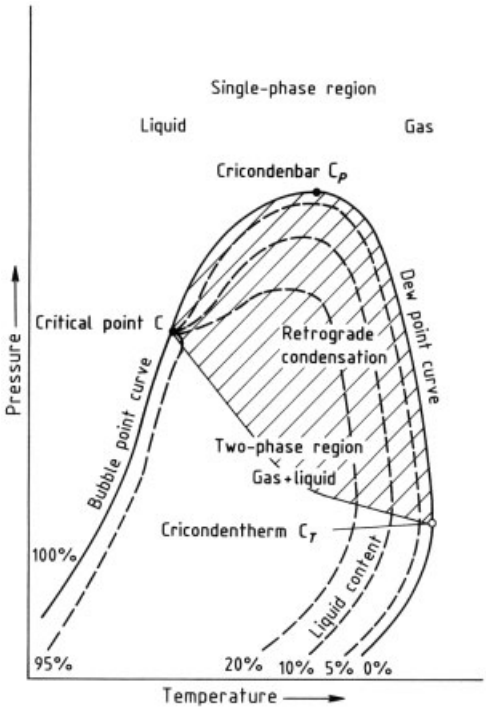

Figure 4. Phase diagram of a hydrocarbon system of fixed composition

Retrograde Condensation. The dashed lines within the two-phase envelope of Figure 4 are liquid-content lines commonly represented by volume percent liquid lines. They originate at the critical point C and are intersected by a line connecting the critical point C with the maximum temperature point C_T. This line separates the two-phase envelope in two regions. In the lower region the gas mixture behaves as expected: At constant or decreasing temperature, the liquid content in the gas mixture increases with increasing pressure. The behavior of the gas mixture in the two-phase region above the connection line $C - C_T$ is quite different. At constant temperature, the liquid content in the gas mixture increases with decreasing pressure, and in the range between the critical pressure and the maximum pressure, the liquid content increases at a pressure drop even when the temperature increases.

This phenomenon, called *retrograde condensation*, may seriously impair the transport of natural gas through pipelines because liquid slugs can be formed if the higher hydrocarbons have not been removed sufficiently. The condensation behavior of natural gas can be determined by employing equations of state or by laboratory measurements [17].

Gas Hydrates. Gas hydrates are icelike solids composed of water and hydrocarbons. Hydrocarbons ranging from methane to cyclopentane are known to form hydrates. The formation of hydrates depends on pressure, temperature, molecular size, and concentration of the component. Generally, hydrates form at high pressure and low temperature, but they can occur as high as 30 °C and below 0.7 MPa. Hydrate formation can be predicted from empirical vapor – solid equilibrium ratios, empirical correlations, and laboratory measurements [18].

To avoid the formation of gas hydrates in pipelines, the water vapor content of natural gas is commonly reduced by dehydration before transport. Lowering the water dew point to −5 to −8 °C in relation to the maximum transmission pressure, is a common stipulation in gas specifications.

1.3. Combustion Properties

1.3.1. Calorific Value

The calorific or heating value of natural gas is defined as the heat evolved by complete combustion of a unit quantity of dehydrated gas with pure oxygen at 0.101325 MPa. The initial temperature of gas and oxygen is 25 °C. The flue gas resulting from the combustion process is cooled to the same temperature.

In determining the *total calorific value*, also called the *gross calorific value* (higher heating value), all the water formed in the combustion process is condensed to the liquid state. For determination of the net calorific value (lower heating value), all the water formed by the combustion reaction remains in the vapor state: i.e., the gross and net calorific values differ in the condensation enthalpy of water. The calorific value is expressed in megajoules per cubic meter or, occasionally, kilowatt-hours per cubic meter.

Calorific Value of Gas Mixtures. The calorific value of natural gas can be calculated from the molar composition and the calorific values of the components:

$$H = \sum y_i H_i$$

where y_i is the molar fraction of component i, H_i is the gross or net calorific value of component i. Table 4 shows calorific values for commonly occurring components of natural gas.

Table 4. Calorific values of natural gas components at STP [19]

Compound	Gross calorific value, MJ/m^3	Net calorific value, MJ/m^3
Methane	39.819	35.883
Ethane	70.293	64.345
Propane	101.242	93.215
n-Butane	134.061	123.810
Isobutane	133.119	122.910
n-Pentane	169.19 *	156.56 *
Isopentane	167.53 *	154.99 *
n-Hexane	208.70 *	193.38 *
n-Heptane	265.22 *	245.99 *
Hydrogen sulfide	25.336	23.353

* Compound is in liquid state at STP.

1.3.2. Wobbe Index

When burned at the same pressure, natural gases with different compositions generate the same amount of heat per time unit if their calorific values divided by the square root of their relative densities (Wobbe index) are equal. The Wobbe index can be referred to the gross or net calorific value:

$$W_{\text{gross,net}} = H_{\text{gross,net}} / \sqrt{d}$$

Since gas burners tolerate only limited fluctuations in the heating value of fuel gas, these fluctuations must be restricted. Therefore, natural gas specifications generally stipulate ranges that are defined by Wobbe index rather than heating values [20]. Consequently, certain gas streams may require blending before sales, or separate pipeline networks may be operated on different gas streams with deviating Wobbe index levels.

2. Treating

Natural gas is produced at the well site in compositions of considerable variety. The objectives of natural gas treatment are correspondingly diverse. The composition range that must be encompassed is shown in Table 5 for selected examples.

Besides its primary constituent methane, natural gas may contain various amounts of impurities, which lead to varying problems during production and transportation:

Water. Condensed water forms solid hydrates with hydrocarbons or hydrogen sulfide, or in addition leads to liquid slugs in pipelines and to corrosion [21].

Higher Hydrocarbons. If natural gas contains higher concentrations of C_{2+} hydrocarbons, the recovery of liquid petroleum gas (LPG) and gas condensate is economically important. The desired Wobbe index (see Section 1.3.2) and calorific value could also require a reduction in the concentration of higher hydrocarbons. Even very low concentrations of higher hydrocarbons can condense in pipelines (retrograde condensation). These condensed hydrocarbons can attack plastic pipes and control devices.

Hydrogen Sulfide. Hydrogen sulfide together with free water can cause corrosion, particularly stress corrosion and hydrogen-induced cracking. Removal of hydrogen sulfide is performed almost exclusively in centralized treatment plants.

Carbon Dioxide. Carbon dioxide together with free water causes pitting corrosion in carbon and low-alloy steels. Some natural gases that contain significant amounts of CO_2 must be treated to increase the methane concentration prior to sales.

Sulfur. Gases containing hydrogen sulfide can also contain elementary sulfur as vapor. Some gas fields, primarily in Canada, Germany, and the United States, contain such high quantities that the sulfur, depending on pressure, temperature, and gas composition, can precipitate and plug the production pipe, which then becomes blocked. Furthermore, elemental sulfur and free water are corrosive.

Mercury. Natural gas can contain mercury in concentrations up to several milligrams per cubic meter, the bulk of which exists in elemental form. Separated liquid mercury causes mercury-induced corrosion in pipes and fittings, corrosion damage to aluminum heat exchangers in cryogenic plants, and damage to measuring and control valves containing nonferrous metals by amalgam formation. Mercury must also be removed because of its toxicity.

Table 5. Composition of natural gas

Source	Methane, vol %	Ethane, vol %	Propane, vol %	Butane, vol %	Higher hydrocarbons, vol %	Nitrogen, vol %	Carbon dioxide, vol %	Hydrogen sulfide, vol %	Elemental sulfur, g/m^3	Organic sulfur, ppm	Helium, vol %	Mercury, mg/m^3
Netherlands, Groningen [a]	81	2.8	0.4	0.2	0.1	14.3	0.9				0.03	0.002
Norway, Ekofisk [b]	85	8.4	2.9	1	0.15	0.4	2	10 ppm		1		
Federal Republic of Germany, Söhlingen[c]	85	1.5	0.07	0.03	2 g/m^3 C_{8+}	12.5	0.5				0.03	1 – 5
Indonesia, Arun[d]	75	5.5	2.3	1.1	0.75	0.3	15	100 ppm				0.2
United States, Panhandle, Tx.[e]	73.2	6.1	3.2	1.6	0.6	14.3	0.3				0.7	
Federal Republic of Germany, Wustrow[f]	43	0.6		0.1		56	0.3				0.04	
Federal Republic of Germany, Süd-Oldenburg[g]	77	0.1				7	8	8	0.5 – 3	200		
former states of USSR, Tenguiz[h]	42	8.5	5.2	3.3	22	0.8	2.6	16		1000 ppm		
Canada, Bearberry[i]	4					1	5	90	80			

[a] Sweet gas with low calorific value, only adjustment of water dew point.
[b] Natural gas with high calorific value, H_2S conversion to COS at molecular sieve.
[c] Mercury removal, C_{3+} separation by low temperature separation and adsorption.
[d] Liquefaction, separation of gas condensate, $CO_2 - H_2S - Hg$ removal.
[e] Helium production.
[f] Nitrogen separation, CO_2 removal.
[g] Sour gas, use of sulfur solvent, H_2S, CO_2 scrubber, sulfur production.
[h] Condensate, LPG, sulfur recovery, removal of organic sulfur compounds.
[i] Ultrasour gas, only for sulfur production.

Other Components. In Alabama, for example, *cyclic hydrocarbons* such as adamantane and diamantane are found in natural gas; these compounds are inclined to sublime and must be removed by injection of oil.

Some gases contain high amounts of *nitrogen*, which must be separated to obtain a sales gas with the required combustion properties.

Recovery of Components with Economic Value. Another important factor is whether additional components must be removed to meet product specifications, or whether the recovery of these products is of economic interest. Examples of the latter include the recovery of helium, the production of sulfur from sour gas, and the recovery of LPG and hydrocarbon condensates. A special case is so-called *ultrasour gas* (see line 9 of Table 5). In this case the main interest is not to produce a fuel gas but rather to provide a feedstock in the production of sulfur. In Canada, production of gas with a hydrogen sulfide content of 90 % began. In addition, the removal of carbon dioxide for enhanced oil recovery in crude oil production should be mentioned.

By contrast, if these components are present only in trace amounts they are not removed because of economic reasons, but rather to avoid corrosion, environmental, or production problems.

The individual process by which undesired gas contaminants can be eliminated depends on the composition of the gas, the amount and type of contaminant, and ultimately whether a salable pure gas must be delivered at the well or whether the gas is piped to a processing plant.

2.1. Objectives of Natural Gas Treating

Two objectives must be met in treating natural gas:

- Adjustment to the required quality standards
- Recovery of byproducts

The main objective worldwide is production of a sales gas of pipeline standard. In this respect, both the extent of removal of undesirable components and the range of conditioning measures are predetermined.

Natural gas treating can take place both directly *at the well* and in *centralized plants*. Treating at the well is always necessary when the gas cannot be transported to a processing plant by pipeline without hazard. To avoid corrosion, sour gas is often dehydrated at the well so that free water can no longer condense in the pipeline. The recovery of sulfur and the separation of nitrogen, carbon dioxide, or helium are carried out in centralized plants, to which streams of different wells are fed. The same holds true for liquefaction of natural gas, recovery of LPG, and removal of higher hydrocarbons.

2.2. Natural Gas Wells and Field Installations

Installations at individual gas wells must fulfill the following objectives:

to guarantee trouble-free gas flow in the production pipes and surface equipment;
to separate or, if necessary, treat coproduced liquids;
to condition the gas for transportation to the processing plant;
to meet the required feed quality for the pipeline distribution network; and
to protect the production pipes and field installations from corrosion.

The type and design of plants are determined by:

type and concentration of accompanying substances in the gas;
behavior of the gas during pipeline transportation; and
applicable sales specifications.

Basic Flow Diagram of Natural Gas Treating. A flow diagram for the most important treatment steps is shown in Figure 5. To prevent deposition of *elemental sulfur*, which can occur, for example, in some sour gas wells, solvents are pumped into the production pipe. Corrosion inhibitors can also be introduced in the same way. On the surface, liquids condensed under the well-head flow conditions are separated. After being depressurized and cooled to collection pipe conditions, further condensate is precipitated. If necessary, separated condensate must be treated further (phase separation, stabilizing, stripping). If large amounts of formation water occur or considerable amounts of liquids (inhibitors, sulfur-dissolving agents) are injected during sour gas production, hydrogen sulfide may be released during depressurization and transportation of these liquids. Emissions can be minimized by high-pressure liquid stripping or by recompression of the flashed gas. When no further condensable hydrocarbons are present, the gas is dehydrated to a dew point equal to or below the dew point specified for pipeline transportation.

Whereas *sweet gas* can be fed directly from the well site to the distribution network, *sour gas* is usually treated further in a centralized plant serving several gas fields. If pressure at the production well is low, compressors are installed upstream of the pipeline network or treatment plant.

For wet gases (see Section 1.1), both the dew point of hydrocarbons and that of water must be lowered. After the injection of hydrate inhibitors (e.g., glycols, methanol), water vapor and higher hydrocarbons are condensed by reducing the temperature [22].

If reduced pressure in the downstream pipeline system causes retrograde hydrocarbon condensation, adsorption processes are used for further removal of higher hydrocarbons [23].

Various adsorbents are available to remove traces of mercury; activated charcoal impregnated with sulfur is particularly useful.

2.3. Processing Steps During Production

2.3.1. Use of Sulfur Solvents

If elemental sulfur is present in the reservoir, sour natural gases are enriched with sulfur to a greater or lesser extent. The ability of the gas to absorb sulfur depends strongly on pressure, temperature, and gas composition [24]. Cooling and pressure reduction during production minimize the solubility of sulfur and lead to its precipitation; at low temperature, solidification and blockage of the pipe may occur. These problems are generally overcome by injection of suitable solvents into the production equipment which permits sulfur to be removed with the injected

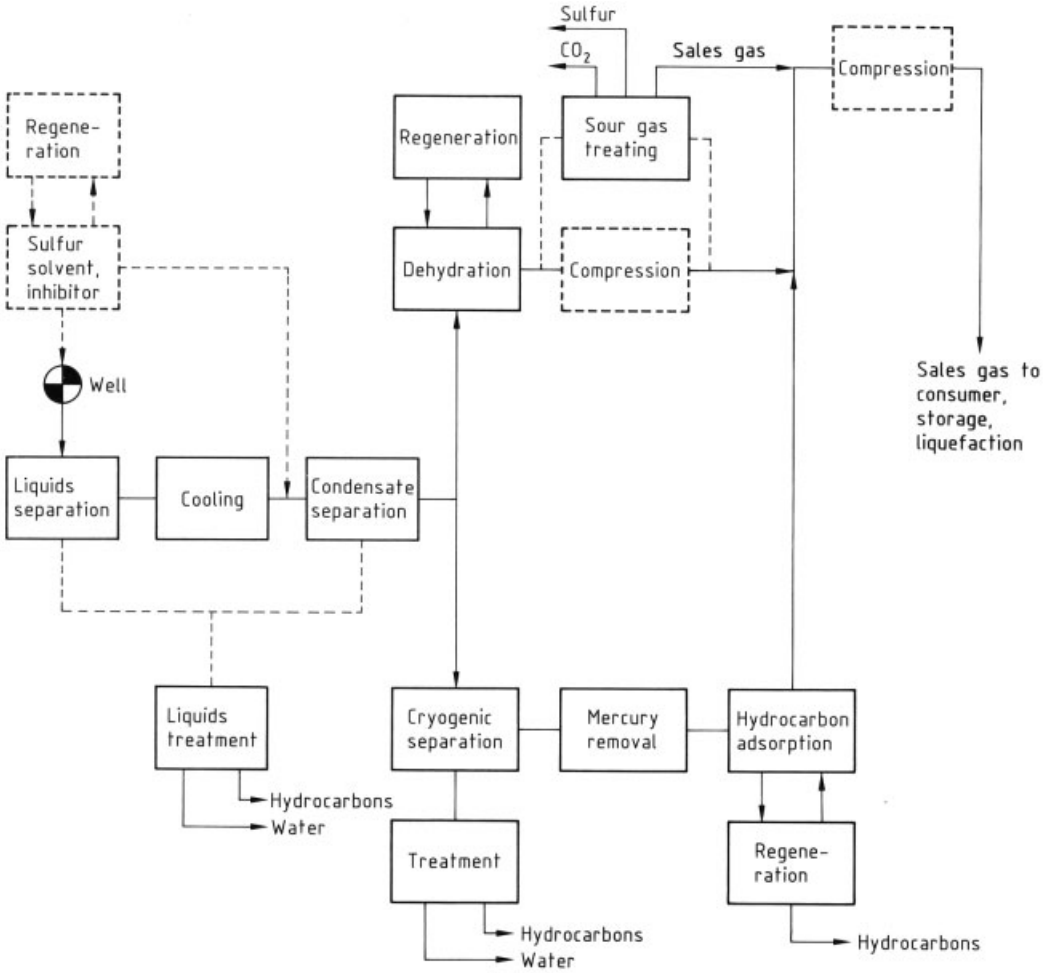

Figure 5. Treating of natural gas
—— Flow sheet for treatment of sweet gas
– – – Additional steps required for treatment of sour gas

liquid, either chemically bound or physically dissolved. Two different reliable processes are applied:

Polysulfide Process [25]. By the addition of aqueous ammonia or monoethylamine solutions, sulfur is dissolved, in the presence of hydrogen sulfide, as polysulfide and separated at the surface in a liquid separator. The sulfur-laden liquid can be regenerated at ca. 150 °C. Ammonia or amine, water, hydrogen sulfide, and carbon dioxide are dissolved in the condensed water and reinjected into the well. Sulfur is precipitated as a liquid and removed from the bottom of the regenerating column.

With this process, only a single liquid phase occurs above ground. Traces of ammonia or amine cause no problems in the downstream treatment. For gases with a high carbon dioxide content the process can be employed only under certain conditions. In this case the alkaline media cannot dissolve sufficient hydrogen sulfide for the formation of polysulfide. This solvent also cannot be employed in the presence of formation water containing alkaline-earth metal chlorides because the calcium carbonate and magnesium hydroxide formed cause blockages that can be dissolved only by acid treatment. Specific measures are necessary to prevent down-hole tubing corrosion.

Physical Solvents [26]. In the production of sour gas, sulfur may also be physically dissolved by means of suitable hydrocarbons. A suitable solvent must have the following properties:

no reaction with formation water;
stability under reservoir conditions;
no reaction with sulfur;
high solvent power for sulfur;
very low vapor pressure; and
noncorrosiveness.

Spindle oil and mixtures of spindle oil with special aromatic coal tar fractions, which have a very high solvent power for sulfur, are the most common solvents for this purpose. These oils simultaneously serve as carriers for corrosion inhibitors.

Sulfur-laden oils can be regenerated, for example, by extraction with amine solutions that also contain hydrogen sulfide. After removal of sulfur by formation of polysulfides these solutions can be regenerated as already mentioned. Phase separation may be hindered by formation of emulsions of the oil with the simultaneously separated water, making specific measures necessary.

Typical concentric piping for sour gas production is shown in Figure 6, in which the solvent is pumped into the well through the outer concentric pipe and returns to the well head with the gas production stream. A sour gas well with solvent injection and the necessary regeneration plant is depicted schematically in Figure 7.

2.3.2. Removal of Mercury

Natural gas from predominantly "Rotliegenden" reservoirs often contains mercury in vapor and/or aerosol form [27]. Natural gas fields containing mercury are known in Algeria, Indonesia, The Netherlands, and the Federal Republic of Germany. Mercury has not resulted in any problems in production from sour gas fields. To avoid health hazards during treatment and use, the mercury content is lowered in separate mercury-removal plants from 5 mg/m^3 to < 10 µg/m^3.

A considerable portion of the mercury has often already been removed by low-temperature separation (LTS) in the dehydration unit at the well. Mercury levels of < 5 µg/m^3 are attained in downstream chemisorption fixed-bed reactors with, for example, sulfur-impregnated activated carbon. Condensable components, such as water or hydrocarbons, reduce the loading ability for mercury or even deactivate the adsorbent. The reactors are not regenerated; rather, the laden adsorbent is fed to an external treatment plant. Regenerative amalgamation processes are currently being tested.

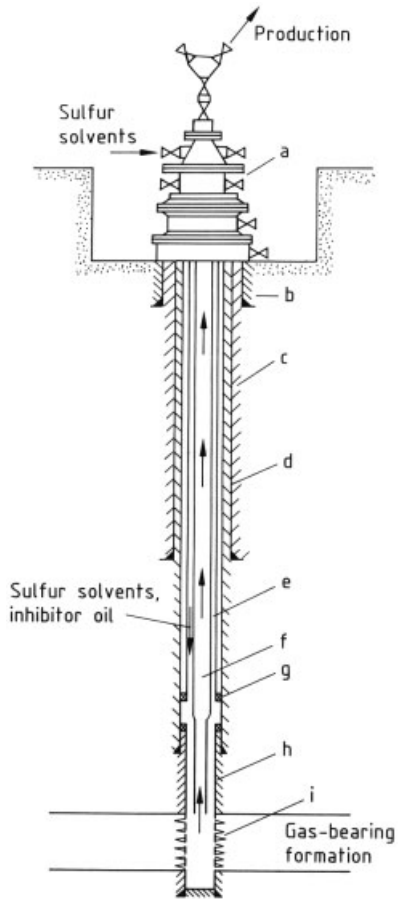

Figure 6. Schematic of a typical sour gas well completion a) Well head; b) Anchor casing; c) Cement; d) Intermediate casing; e) Injection annulus; f) Tubing; g) Production packer; h) Liner; i) Perforations

Mercury is also collected from the liquid hydrocarbon stream by phase separation. Further treatment with solid adsorbents reduces the mercury content from 300 to 20 ng/g.

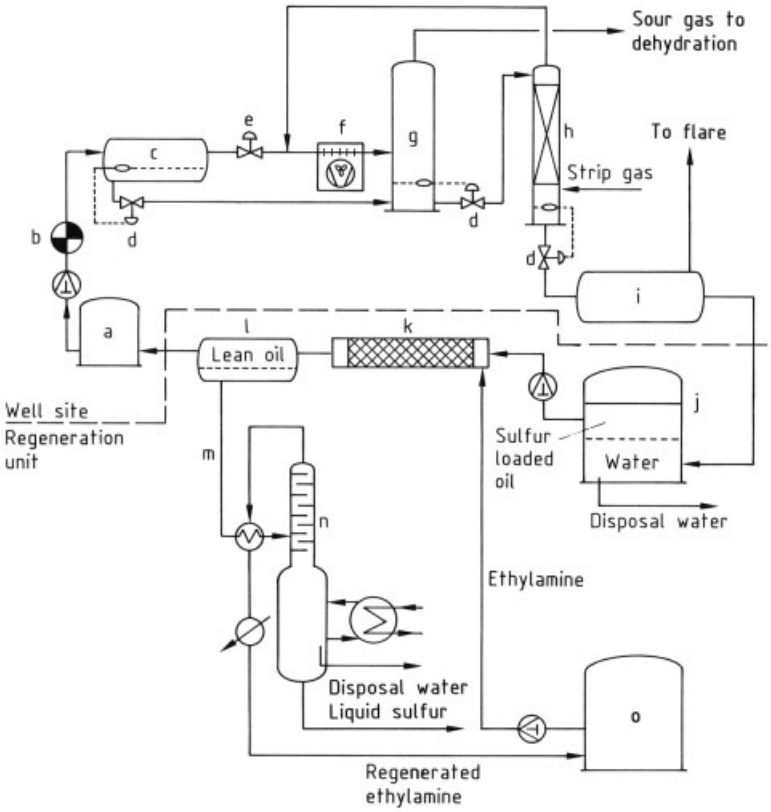

Figure 7. Flow diagram of a sulfur solvent inhibitor system
a) Lean oil storage; b) Sour gas well; c) Free water knockout; d) Liquid control valves; e) Flow control valve; f) Cooler; g) Separator; h) Stripper; i) Liquid storage; j) Phase separation; k) Extraction unit; l) Phase separator; m) Ethylamine polysulfide; n) Ethylamine regeneration; o) Ethylamine storage

2.3.3. Dehydration

In most cases, natural gas is dehydrated close to the gas wells for safety reasons. The formation of gas hydrates in high-pressure gas pipelines can thus be avoided and problems arising from liquids in the transportation system minimized. In general, a water dew point of −8 °C at the pipeline pressure is adequate. In particular, in the presence of carbon dioxide, hydrogen sulfide, and chlorides, corrosion can be minimized by decreasing the water content. In some gas fields, dehydration is not employed; production is maintained by the use of specific corrosion-resistant materials together with corrosion inhibitors and thermally insulated equipment.

Cryogenic Dehydration. To decrease the higher hydrocarbon content in addition to maintaining a specific water dew point, cryogenic dehydration of natural gas is often employed (see Fig. 8). The wet gas is cooled until the components to be removed precipitate by condensation or formation of hydrates. Methanol, glycol, or a paraffin solvent is often also injected. The reduction in temperature can be achieved by Joule – Thomson expansion. This route is employed particularly in the first phase of production in high-pressure wells. When the production pressure is close to that of the transportation system, external cryogenic units are used.

Dehydration by Absorption Processes [28]. Standardized dehydration plants using glycol absorption are employed most widely. Triethylene glycol (TEG) is used in preference to other glycols (mono- and diethylene glycols) because of its high absorption capacity for water vapor, its low vapor pressure (small losses from evaporation), and its high thermal stability. This type of dehydration plant is shown in Figure 9.

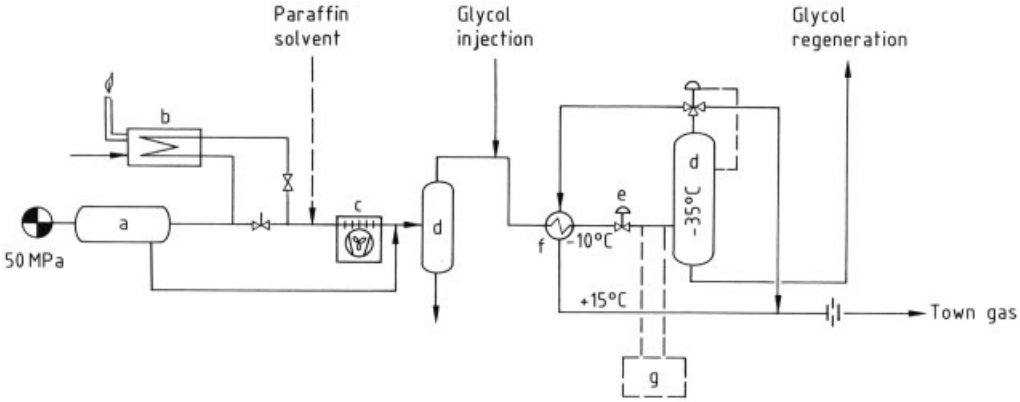

Figure 8. Dehydration with a hydrate inhibitor and cooling at a high-pressure well
a) High pressure separator (free water knockout); b) Heater; c) Air cooler; d) Separator; e) Joule – Thomson valve; f) Gas – gas heat exchanger; g) External cooling system for low-pressure wells

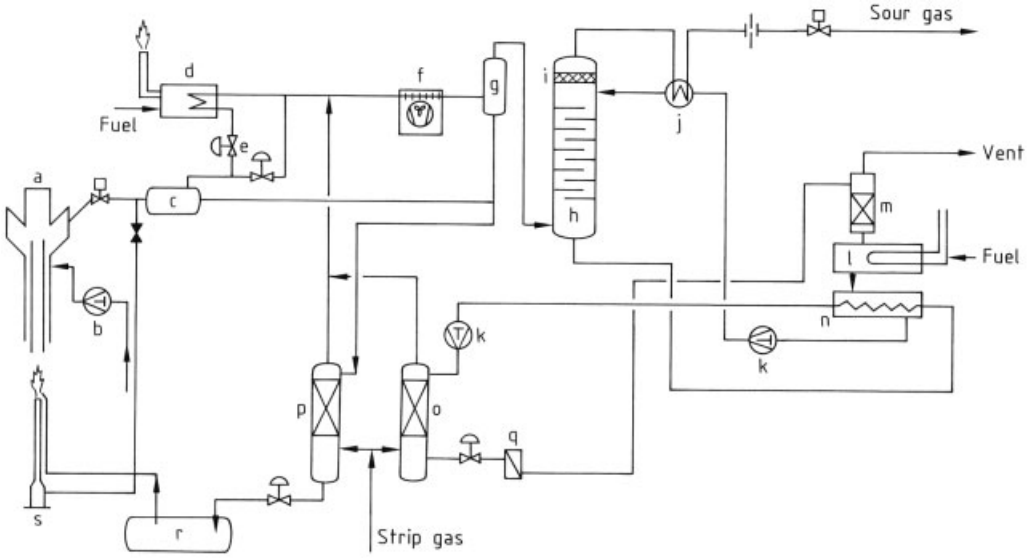

Figure 9. Glycol dehydration
a) Well head; b) Corrosion inhibitor – sulfur solvent injection pump; c) High-pressure separator (free water knockout); d) Heater; e) Flow control valve; f) Air cooler; g) Separator; h) Glycol absorber; i) Demister; j) Glycol – gas heat exchanger; k) Glycol pump; l) Glycol reboiler; m) High-concentration stripper; n) Glycol storage tank; o) Glycol stripper; p) Water – solvent stripper; q) Glycol filter; r) Produced liquids storage tank; s) Flare

In the absorber (h), glycol and gas are brought in contact countercurrently. Plant pressures of 2 – 10 MPa are typical.

In the approximately pressureless *regeneration* step, water-laden glycol is heated to 200 °C. The energy requirement is reduced by extensive heat exchange between laden and regenerated glycol in a coiled tube in the storage vessel and in the top of the regeneration column. Both direct gas-fired regenerators and those with steam or electric heating are used. The residual water content (ca. 1 %) in the glycol after thermal regeneration can be reduced still further by stripping with dry gas or applying a vacuum. The stripping medium, often sweet gas, is passed through the storage vessel to be heated in the spiral tube before the desorption of the water vapor in the distillation section is increased. A residual water content in the glycol of $< 0.3\,\%$ can be achieved [29]. The dew points attainable

for natural gas in equilibrium with a triethylene glycol – water mixture are given in Figure 10 as a function of contact temperature. The plants are self-controlled; the dehydration process is maintained by indirect control of the glycol circulation rate, the regeneration temperature, and the stripping gas rate [30].

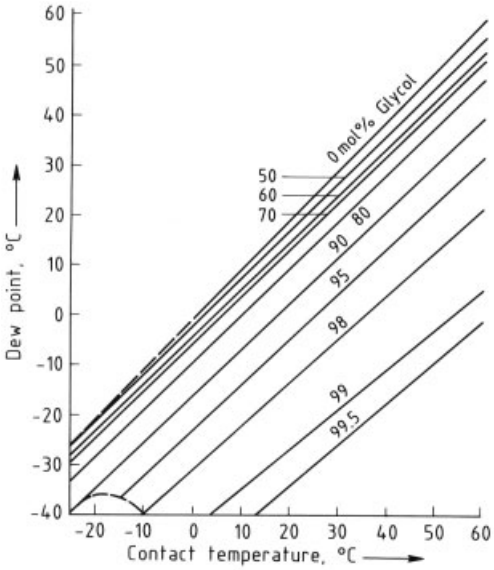

Figure 10. Attainable dew points of natural gas as a function of the (gas) contact temperature with triethylene glycol – water solutions

Enrichment of hydrocarbons in the glycol can be limited by the use of oil skimmers. Degradation products resulting from entry of oxygen, thermal decomposition, or influx of acids from bore hole treatment render glycol corrosive. Corrosion control is usually carried out by the addition of inhibitors that increase the pH. The glycol is usually filtered through activated charcoal. The service life of glycol can be several years [31].

Further measures are necessary for dehydration of sour gas, as shown by the dashed lines in Figure 5. At the usual plant pressure of 10.0 MPa, glycol absorbs considerable amounts of hydrogen sulfide, in addition to water vapor. The former is released during regeneration of the glycol, thereby leading to corrosion and emission problems. High-pressure strippers are therefore installed. At a pressure slightly above that of the glycol absorber, hydrogen sulfide is removed from glycol by stripping with heated sweet gas. The sweet gas often originates from neighboring fields or plants. The strip gas, enriched with hydrogen sulfide (representing $< 2\,\%$ of the dehydration unit capacity), is recycled to the absorber. Glycol dehydration units are constructed of normal carbon steel.

In the final phase of gas production from a well, considerable amounts of salt-containing formation water are coproduced. During flashing and transportation of these laden liquids, gases are released, including hydrogen sulfide and benzene – toluene – xylene (BTX). Emissions can be minimized and risks reduced by additional stripping or by recompression of these flash gases.

Adsorptive Dehydration. In adsorptive dehydration the gas is brought in contact with molecular sieves, silica gel, or SORBEAD (i.e., Na_2O-containing SiO_2). Plants with only two adsorbers, working alternately, are possible. Dew points $< -70\,°C$ are attainable with adsorption plants. This is particularly necessary for cryogenic plants and liquid natural gas (LNG) plants, where traces of water and carbon dioxide can lead to blockage by ice formation.

For *regeneration*, ca. 5 % purge gas is drawn from the main gas flow and fed through the laden adsorber at plant pressure and 200 – 330 °C (Fig. 11). Removed water vapor from purging of the fixed bed is condensed so that the purge gas, which is at plant pressure, can be recycled. Regeneration by depressurization is seldom employed in natural gas treating.

Under the normal requirements of natural gas treating dehydration units of this type are economically less attractive than glycol and cryogenic dehydration plants described above. However, they are nevertheless employed when, in addition to simply removing water, adjustment to a specific hydrocarbon dew point is required, removal of hydrogen sulfide or carbon dioxide is necessary, or an extremely low dew point is required.

The adsorbent lifetime depends mainly on the amount of impurities (e.g., salt, inhibitors, sulfur compounds, and waxy hydrocarbons) that originate from well production. Conversion of sulfur compounds to hydrogen sulfide or carbonyl sulfide in the adsorber has been observed. Ad-

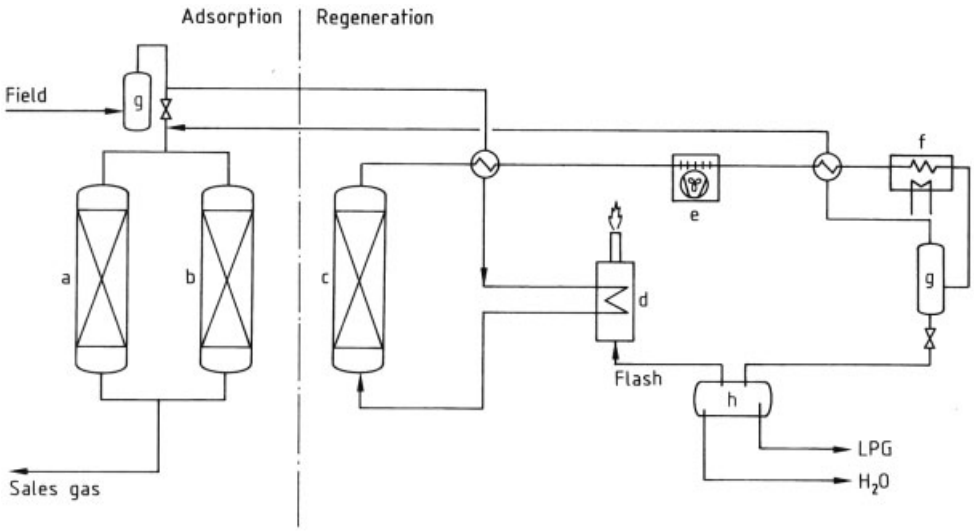

Figure 11. Hydrocarbon removal with adsorption and regeneration under pressure
a) – c) Adsorbers; d) Heater; e) Cooler; f) Chiller; g) Separator; h) Phase separator

sorption plants are often used in offshore fields to minimize the need for scrubbing solvents.

2.3.4. Removal of Hydrocarbons

Recovery of hydrocarbons is mainly done for economic reasons. However, removal may also be necessary to avoid uncontrolled appearance of hydrocarbon condensates in long-distance pipelines and low-pressure distribution systems in cities [32, 33]. Such condensates in pipelines may produce liquid slugs, especially during varying loads, which lead to errors in downstream measuring and controlling devices. In addition, plastic pipelines or plastic components in meters are adversely affected. Furthermore, condensates impair leak detecting by use of odorants. Therefore, in addition to a water dew point, a hydrocarbon dew point is often specified. Thus, condensation of hydrocarbons is avoided at ground temperature and typical pressure ranges of the transportation system (10 – 0 MPa) [34].

Higher hydrocarbons strongly influence the dew point curves of natural gas and also determine the conditions under which retrograde condensation can occur (see Section 1.2). As an example, the dew point lines of an untreated and a treated natural gas are shown in Figure 12. The vapor pressure curve of pure methane is also shown.

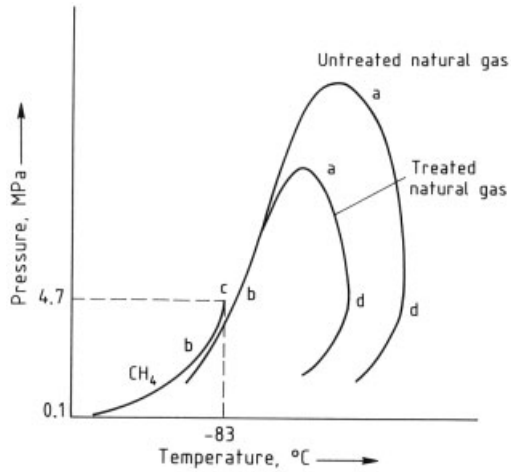

Figure 12. Phase diagram for natural gas containing higher hydrocarbons and for pure methane
a) Dew point curve; b) Boiling point curve; c) Critical point; d) Cricondentherm point

The required hydrocarbon dew point cannot be attained with cryogenic processes. Industrially, adsorption plants with activated charcoal or SORBEAD are used, in which the higher hydrocarbons are removed and recovered preferentially at elevated regeneration temperatures and short cycle times (see Fig. 11). Presence of water causes problems and is usually removed first in upstream dehydration plants [32].

2.4. Removal of Carbon Dioxide and Sulfur Components

Removal of the sour gas components hydrogen sulfide and carbon dioxide is one of the most common aims of natural gas treating. Furthermore, other sulfur compounds (e.g., COS or mercaptans) are frequently present and must also be removed. The quality demands on purified gas and the anticipated type of processing or further use of the separated components are quite varied. As a result, a choice of many different processes is available [35].

2.4.1. Absorption Processes

The method with the broadest application is gas scrubbing. Here, the sour gas components are reversibly bound to a solvent by chemical or physical absorption. In the regeneration step the components are desorbed unchanged and the solvent is recycled to the scrubber.

2.4.1.1. Physical Absorption Process [35 – 37]

If *selective removal of sulfur compounds* is required for natural gases with a high content of sour components, physical absorption processes are preferred. In physical absorption, carbon dioxide and hydrogen sulfide are only physically dissolved in the solvent. The process is reversible: during regeneration under strongly reduced pressure, absorbed gaseous components are released unchanged from the solvent (→ Absorption, Chap. 4.1) (Fig. 13).

Important processes include the following: (1) the Purisol process, which operates with *N*-methylpyrrolidone (NMP); (2) the Selexol process, in which the solvent is a mixture of poly(ethylene glycol dimethyl ether)s; (3) the Sepasolv process, with poly(ethylene glycol methyl isopropyl ether); (4) the Fluor solvent process, with propylene carbonate; and (5) the cold methanol process (Rectisol). With the exception of the Rectisol process (which operates at −10 to −30 °C), all operate at ambient temperature and employ high-boiling organic solvents (Table 6). The absorption properties for carbon dioxide are similar; the selectivity of NMP for hydrogen sulfide is higher, which is often desirable.

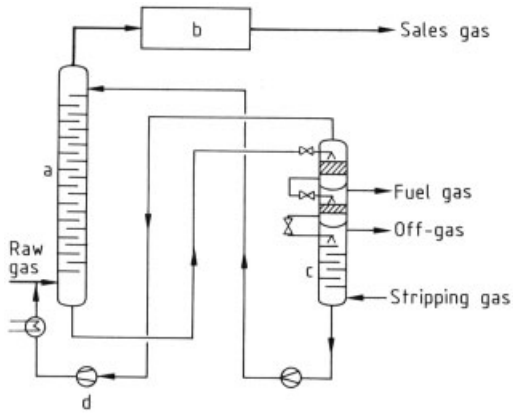

Figure 13. Physical scrubbing with regeneration by depressurization and stripping
a) Absorber; b) Solvent recovery; c) Stripper; d) Recompressor

Physical scrubbing agents are usually anhydrous; thus both dehydration and removal of carbon dioxide and/or hydrogen sulfide can be carried out in a single step; the separate dehydration step usually required in chemical absorption process is superfluous. At a water content of ca. 3 – 5 % in the scrubbing agent (e.g., Selexol), it is also possible to regenerate without a strip gas compressor; the stripping agent is then the water vapor present.

To reduce losses and prevent condensation of the scrubbing agent in the pipeline, small post-treatment units are often installed (e.g., with glycol). Washers, activated carbon adsorbers, or refrigeration systems, are also possible. Because of the good solubility of higher hydrocarbons in these physical solvents, these processes also allow the recovery of LPG.

Amount of Scrubbing Agent. The circulating amount of scrubbing agent required is almost independent of the concentration of the components to be removed but does depend on pressure and raw gas throughput. *Loading* of the physical scrubbing agent increases with the partial pressure of the components to be removed: Thus, physical absorption is used advantageously at higher concentration of these components and higher process pressure. A large amount of coabsorbed methane can be released in the first de-

Table 6. Characteristic properties of physical solvents

Property	Purisol	Selexol	Propylene carbonate	Sepasolv	Rectisol
Vapor pressure (25 °C), Pa	53.2	0.097	11.3	0.049	13 000
Solubility of CO_2 (25 °C), m^3/m^3	3.54	3.6	3.38	3.38	13.43 *
Solubility relative to CO_2					
CH_4	0.072	0.067	0.038	0.066	0.051
C_{3+}	1.07	1.02	0.51		2.35
COS	2.72	2.33	1.88	2.54	3.92
H_2S	10.2	8.53	3.29	6.86	7.06

* at −25 °C.

pressurization step by flashing (usually multistage flashing) and can be fed profitably into internal fuel-gas systems.

A high selectivity for hydrogen sulfide compared to carbon dioxide leads to process streams with high hydrogen sulfide concentration, which can be used in the Claus process. A certain residual amount of carbon dioxide in the treated gas may be desirable to adjust its calorific value to required specifications. To match quality specifications, physical scrubbing processes can be operated with variable selectivity by adjusting circulation of the solvent and temperature.

Physical scrubbing processes are noncorrosive, and regular carbon steel is generally used as construction material. Complete removal of carbonyl sulfide from the crude gas requires sophisticated equipment.

2.4.1.2. Chemical Absorption Processes [35, 38 – 40]

Alkanolamine scrubbing, known since 1930, is currently one of the most frequently used processes (→ Gas Production). Aqueous solutions of monoethanolamine (MEA), diethanolamine (DEA), diisopropylamine (DIPA), diglycolamine (DGA), and methyldiethanolamine (MDEA), in particular, are used as absorbents. The choice of solvent is determined by different factors.

Monoethanolamine allows a high final purity but requires a considerable amount of energy for regeneration, and absorption of carbonyl sulfide leads to degradation products that must be eliminated in a reclaimer (see Section 2.8). Selective removal of hydrogen sulfide is not possible. *Diglycolamine*, like MEA a primary amine, is suitable for the complete absorption of both hydrogen sulfide and carbon dioxide and can also be used for scrubbing of carbonyl sulfide and mercaptans. The secondary amines *DEA* and *DIPA* show a lower heat of reaction for absorption – desorption of sour gas components; as a result, they require less regeneration steam than MEA. At higher carbon dioxide levels, various degradation compounds are formed that cannot be regenerated, and require use of a reclaimer (see Section 2.8). A certain selectivity for hydrogen sulfide is possible, particularly with DIPA, because its reaction with carbon dioxide is slower. Tertiary amines such as *MDEA* have considerably better selectivity for hydrogen sulfide. The heat of reaction with MDEA is lower, degradation products are not formed with carbonyl sulfide and carbon dioxide, and only a few corrosion problems arise. If gases at low pressure are purified, the normal pipeline specification for hydrogen sulfide is not attained. Comparative data for characterization of various alkanolamines are presented in Table 7.

Addition of *corrosion inhibitors* and special chemically active components leads to advanced processes that allow higher alkanolamine concentration and higher loading, for example, or better selectivity for hydrogen sulfide and, therefore, lower circulation rates. A lower corrosion tendency and a reduction of energy required for regeneration can thereby be obtained. Examples are the Amine Guard, Econamine, activated MDEA (BASF), Gas/Spec, and Ucarsol processes [41, 42]. Selectivity can be increased, and further savings in energy can be obtained by the use of sterically hindered amines (Flexsorb, Ucarsol Innovator Solvent 111). The energy consumption for the hydrogen sulfide/carbon dioxide or carbon dioxide scrubbing system is also lower with these amines [43, 44].

Table 7. Comparison of chemical solvents

Absorbent	Heat of reaction, kJ/kg		Relative CO_2 absorption,	Capacity *	
	H_2S	CO_2	%	H_2S	CO_2
MEA	1905	1920	100	0.07	0.50
DEA	1190	1510	40	0.09	0.32
DIPA	1140	2180	20		
MDEA	1050	1420	10	0.10	0.12

* Moles of sour gas per mole of amine for solutions with 1.3 mol of amine per kilogram water.

Activated and inhibited *potassium carbonate* solution (Benfield, Catacarb) is widely applied for purification of natural gas (→ Gas Production) [45]. The process can in principle be used to remove carbon dioxide, hydrogen sulfide, carbonyl sulfide, and carbon disulfide. However, it is used mostly in treating gases with high carbon dioxide content. The steam requirement for regeneration can also be reduced by addition of hindered amines.

2.4.1.3. Physical – Chemical Absorption Processes

The advantages of chemical absorption (e.g., removal of COS, low coabsorption of methane) are combined with those of physically effective organic solvents (e.g., high loading at high partial pressure). As amines, DIPA and MDEA are preferentially employed. The physical absorbents are high-boiling solvents (e.g., Sulfolan or Selexol). In addition, 5 – 30 % water is present. The Sulfinol process has found widespread use [46, 47] and has proved useful in the United States, Canada, and the Federal Republic of Germany for simultaneous removal of hydrogen sulfide, carbon dioxide and organic sulfur compounds. However, the coabsorption of hydrocarbons is also increased by a physical solvent, so that an intermediate flash step with reabsorber for the recovery of sulfur-free fuel gas (shown as a dashed line in Fig. 14) is indispensable. *Selective* hydrogen sulfide scrubbing can be attained with this process by proper choice of the amine.

Other physical–chemical absorption processes include Selefining, Flexsorb PS [44], and Amisol. With the Amisol process a mixture of methanol and alkanolamines or aliphatic alkylamines is used [48]. Like in all chemical scrubbing processes, the gas is saturated with water vapor in the absorption step, making a downstream dehydration unit necessary.

Selection Criteria for Absorption Processes [35, 38, 40, 49]. Selection of the best suited process is essentially determined by the pressure and composition of the raw natural gas, the trace components, and the desired quality of marketable natural gas and separated sour gas components. Investment and operating costs are also important. The design of the absorber and the amount of solvent required are determined by the absorption behavior of the least soluble compound to be removed. Of the investment cost, 50 – 70 % are directly related to the required amount of scrubbing agent in circulation. In contrast, operating costs are determined to about 70 % by the energy requirements for regeneration and recirculation of the scrubbing agent. In view of the many parameters to be considered, a simple selection of a process by means of a diagram or a table is not possible. Only a few factors can be mentioned here.

Raw Gas Composition. Physical absorbents are generally not used for the purification of natural gas that contains higher hydrocarbons. The solubility of butane, for example, in most solvents is similar to that of hydrogen sulfide. Selective removal of the latter, therefore, is not possible. Hydrocarbon solubility in aqueous chemical scrubbing agents is low. Unfortunately, some chemical absorbents react with gas components (e.g., COS, CO_2, and CS_2) to form corrosive products, whereas physical solvents remain unchanged.

Gas Throughput. Physical scrubbing systems are in general more complex and, therefore, require higher investment. For this reason, chemical scrubbers are often more economical for small gas streams if no side reactions with the chemical absorbent occur. For large plants,

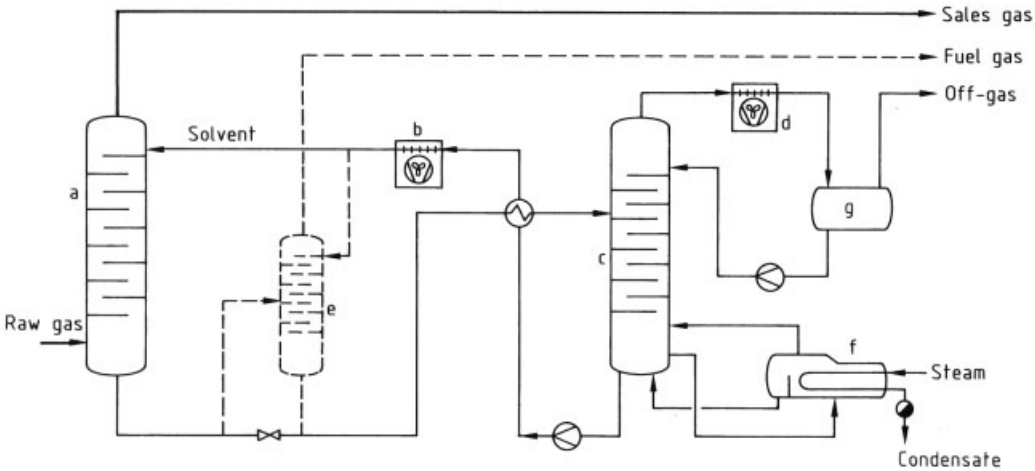

Figure 14. Schematic of physical-chemical scrubbing with hot regeneration
a) Absorber; b) Air cooler; c) Regenerator; d) Air condenser; e) Intermediate flash stage with reabsorber; f) Reboiler; g) Phase separator; h) Expansion turbine

the low operating costs of physical absorption processes outweigh the higher investment costs.

Pressure. In *chemical absorption processes* the capacity of the solvent is determined by reaction of the sour gas components with the active solvent component; i.e., saturation is reached at relatively low pressure. The circulation rate is then proportional to the quantity of the components to be removed and thus also to the total pressure. For *physical scrubbing agents*, loading increases with pressure, and the circulation rate is inversely proportional to the total pressure.

Claus Gas. In most cases, the sour gas removed is fed to a Claus plant for recovery of elemental sulfur (→ Sulfur). For this purpose, the gas should have as high a content of hydrogen sulfide and as low a content of hydrocarbons as possible. If the gas to be purified contains a high level of carbon dioxide in addition to hydrogen sulfide, a selective absorption process is preferable. This requirement is better met by physical than chemical processes. If strip gas is used for regeneration, the hydrocarbon content of the Claus gas should be lowered by intermediate flashing stages and recompression.

Sales Gas. When hydrogen sulfide and carbon dioxide must be simultaneously and almost entirely removed, chemical or physical – chemical scrubbing processes can be attractive. The presence of organic sulfur compounds excludes some solvents. If only partial removal of carbon dioxide is required, selective amine absorption processes are used at low pressure, and physical processes are preferred at high pressure.

Special solvent properties such as vapor pressure, viscosity, toxicity, heat of absorption – desorption, and corrosion behavior must be considered, along with foaming tendency, water-absorbing capacity, pricing, and environmental impact. The decomposition products that must be disposed of can often be reduced by a reclaim and recovery unit.

2.4.2. Liquid Oxidation Processes

The conventional absorptive removal and separate processing of hydrogen sulfide is often too expensive, particularly when the gas stream to be treated is relatively small, the hydrogen sulfide content is low, or the pressure is low. In this case selective liquid oxidation processes can be used [50].

The following steps are generally involved in this process:

Absorption of hydrogen sulfide in alkaline solution
Oxidation of dissolved hydrogen sulfide ions to elemental sulfur with simultaneous reduction of a chemical oxygen carrier or microoganisms

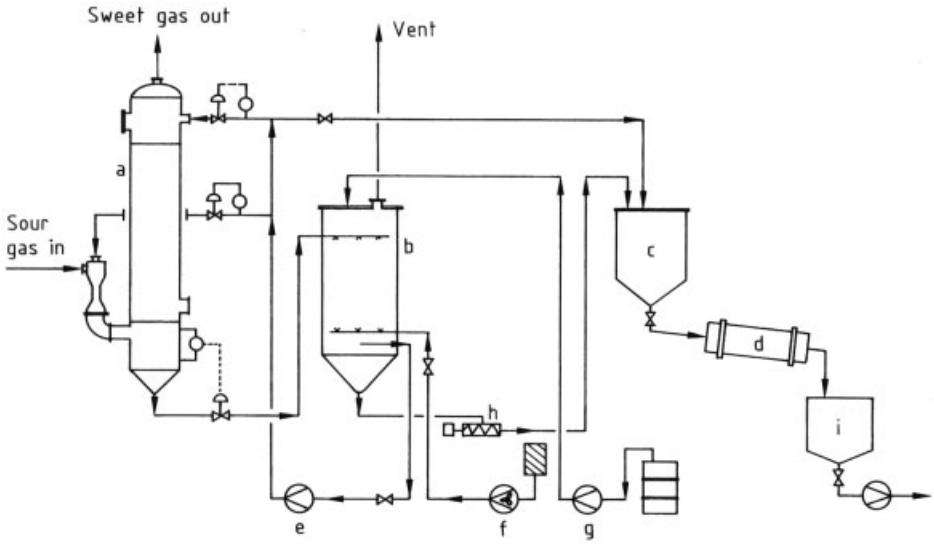

Figure 15. Schematic of a liquid oxidation process
a) Absorber; b) Oxidizer; c) Settling tank; d) Centrifuge; e) Recirculating pump; f) Air blower; g) Chemical additive pump; h) Sulfur slurry pump; i) Sulfur melter

Reoxidation of the active component with air, in special processes with the aid of bacteria
Separation of elemental sulfur by filtration, centrifugation, flotation, or sedimentation and, if necessary, further purification.

The processes are differentiated from each other only by the specific redox system employed. A flow sheet is shown in Figure 15. The sour gas to be purified is brought in contact with the absorption solution in a venturi scrubber or an absorption column. The liquid, which may now contain elemental sulfur in colloidal form, is then treated with air to complete the formation of elemental sulfur and to reoxidize the reduced oxygen carrier which is then recycled to the absorption step. The sulfur sludge is collected and treated further.

The following processes are used:

1) *Stretford, Sulfolin, Unisulf*
Vanadates are employed for oxidation of the absorbed hydrogen sulfide. Organic oxygen carriers (e.g., anthraquinonedisulfonic acids) are required to reoxidize the vanadium compounds. Optimized formulations suppress side reactions (e.g., formation of sulfate and thiosulfate) and lower the consumption of chemicals.

2) *Takahax, Hiperion*
For these processes, naphthoquinonesulfonates or chelates are used as oxidizing agents. Some environmental problems associated with the use of vanadates can thus be avoided. In both groups of processes, hydrogen sulfide is absorbed very selectively; they are also suitable for gases with a high carbon dioxide content.

3) *Lo-Cat, SulFerox, Sulfint, Bio-SR*
After absorption in a solution of $NaHCO_3$ – Na_2CO_3, hydrogen sulfide is oxidized with iron(III) aminopolycarboxylic acid chelates. The resulting iron(II) chelates are reoxidized with air. Differences in composition of the solutions employed in individual processes influence the absorption capacity for hydrogen sulfide, the side reactions, the selectivity of absorption in the presence of carbon dioxide, and the ease of removal of sulfur particles. The Bio-SR process works at pH 2 and without chelates. Iron(II) sulfate is oxidized by *Thiobacillus Ferroxidans*[81]

4) *Sulfa-Check*
In this batchwise process, hydrogen sulfide is oxidized to elemental sulfur with a nitrite solution. The solution used and the sulfur-containing sludge must be disposed of.

5) *Thiopaq*
Thiopaq is a biological process for SO_2 reduction, employing absorption in caustic solution and a sulfur formation step by biological oxidation. It requires a sulfate bleed stream [82],[83]

2.4.3. Adsorption Processes

Adsorption processes for the removal of hydrogen sulfide and carbon dioxide are used in natural gas treatment predominantly when only small amounts of acidic components must be removed [51, 52].

Activated charcoal and zeolitic molecular sieves are used as adsorbents. Zeolites, in particular, are suitable for selective separations because of their ordered crystal lattice with extremely uniform pore size and their high adsorptive capacity. Typical equilibrium loading curves are shown in Figure 16.

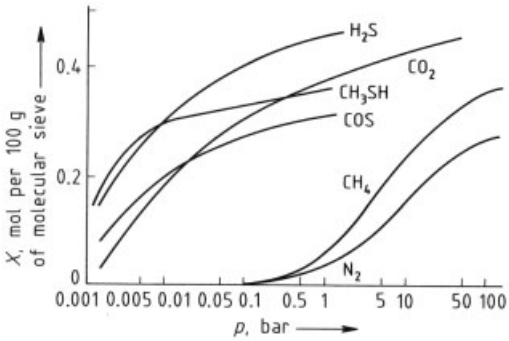

Figure 16. Loading of molecular sieve 5A

Adsorption processes (→ Adsorption) are generally run continuously in such a way that the raw gas to be purified passes through at least one adsorber, while the other adsorber(s) undergo a regeneration cycle. Regeneration can be carried out by temperature swing, pressure swing, inert gas stripping, or a combination of these.

2.4.3.1. Adsorption of Carbon Dioxide

Molecular sieves are used widely for the purification of natural gas when it is used as feedstock for a cryogenic plant for production of liquefied gas. In these plants, even low levels of carbon dioxide cause problems because the gas freezes in the low-temperature unit and can lead to blockage. The simultaneous removal of traces of water by molecular sieves is also an advantage here. A gas quality of < 50 ppm of carbon dioxide and 1 ppm of water can be attained in this way.

The main elements of an adsorption plant with three towers based on the temperature swing process are shown in Figure 11. The first tower (a) is in the adsorption stage for purification of the raw gas; the second (b) is cooled with the treated gas which, after being heated, regenerates the adsorbent in the third tower (c). After completion of this adsorption – regeneration cycle the function of the three identical towers is changed by automatically controlled valves. The regeneration gas, which is enriched in carbon dioxide, can be used for heating. In large plants, however, the volume of regeneration gas may become so large that absorptive workup and recirculation should be considered.

2.4.3.2. Adsorption of Hydrogen Sulfide

Various adsorptive processes are available for the removal of hydrogen sulfide; these are differentiated by the type of binding to the adsorbent and the method of regeneration. *Activated charcoal, iron oxide*, and *zinc oxide* are generally used only for the removal of smaller amounts of hydrogen sulfide because of the limited regeneration capability of these adsorbents [53].

Molecular sieves display a high affinity for sulfur compounds. These materials can be selected to permit simultaneous dehydration and desulfurization without removal of carbon dioxide from the gas. Plants are in service which can purify gas containing 10 – 6000 ppm of hydrogen sulfide and 0.1 – 50 % carbon dioxide. The valuable but hydrogen sulfide-rich regeneration gas is usually treated further [54] and reused.

With special molecular sieves, hydrogen sulfide can be removed and, in the presence of carbon dioxide, converted to carbonyl sulfide:

$$H_2S + CO_2 \rightarrow COS + H_2O$$

For gas with low hydrogen sulfide content the permitted level of carbonyl sulfide in marketable natural gas can be exploited by this means. Re-

generation and further treatment of the hydrogen sulfide-containing desorption gas are then no longer necessary. A survey on adsorption processes is given in Table 8.

Table 8. Adsorption processes

Basic process	Examples	Product
Adsorption	adsorption from oxygen-free gases on active carbon or molecular sieves	H_2S rich gas
Adsorption and chemical reaction	chemical reaction with ZnO	ZnS
Adsorption and oxidation	regeneration by extraction or desorption	
Chemical reaction and oxidation	reaction of H_2S with iron to Fe_2S_3 regeneration by oxidation to Fe_2O_3 and elemental sulfur	S/Fe_2S_3
Adsorption, catalysis	oxidation of H_2S with O_2 to elemental sulfur under active carbon catalysis	S

2.5. Membrane Processes

Gas separation with membranes has developed into an economically interesting process in the last decade. Membranes can be used for almost all aspects of natural gas treating. The following are of particular interest:

Separation of carbon dioxide and methane
Removal of carbon dioxide and hydrogen sulfide
Dehydration
Recovery of higher hydrocarbons [55 – 57]

Poreless, dense polymer films are used for the separation of gas mixtures. Individual components of the gas dissolve in the polymer and move through it by diffusion. Material transport can be described by the following steps: absorption from the gas phase into the membrane matrix, diffusion through the membrane, and desorption out of the membrane into the gas phase. The general rule is that a larger available membrane surface area, thinner effective membrane separating layer, greater difference in partial pressures, higher membrane-specific flow rate of the components to be separated (permeation coefficient), and better selectivity for individual gases result in more complete separation of a mixture. A multitude of polymer membranes are connected in a module to reach as high a packing density as possible. Hollow fiber, spiral-wound, and envelope-type modules have become generally accepted. The principle of a two-step module plant for the treatment of sour natural gas is shown in Figure 17.

The most widespread membrane processes are those for the separation of carbon dioxide from natural gas. An overview of the membrane and module types used in this area is given in Table 9.

A large number of membrane plants are in service in the United States. They are used to reduce the carbon dioxide, hydrogen sulfide, and water content of natural gas directly at the well. Membrane separators are simple to operate and require no electrical energy, coolant, or chemicals. They function without moving parts and can be operated with a minimum of supervision and maintenance. The permeate containing the separated components can, depending on its composition, be used as low-calorific fuel gas employed in the recovery of sulfur, flared, or vented directly to the atmosphere.

The service life of membranes, the pressure difference required, and the loss of gas to the permeate must all be taken into consideration for process selection.

2.6. Cryogenic Natural Gas Separation

Cryogenic processes are always employed when very low-boiling components are to be separated because none of the processes thus far described is suitable. Examples are reduction of the nitrogen content in natural gas and recovery of helium. In these processes, distillation steps are combined in which the considerably different boiling ranges of the primary components methane, nitrogen, and helium can be employed for separation (see → Noble Gases).

2.7. Recovery of Sulfur

2.7.1. Claus Process

Sulfur is recovered directly in elemental form from treated natural gas only by liquid oxidation processes (see Section 2.4.2). With all other absorption and adsorption processes, the regeneration gas stream, containing mainly carbon dioxide and hydrogen sulfide is processed further,

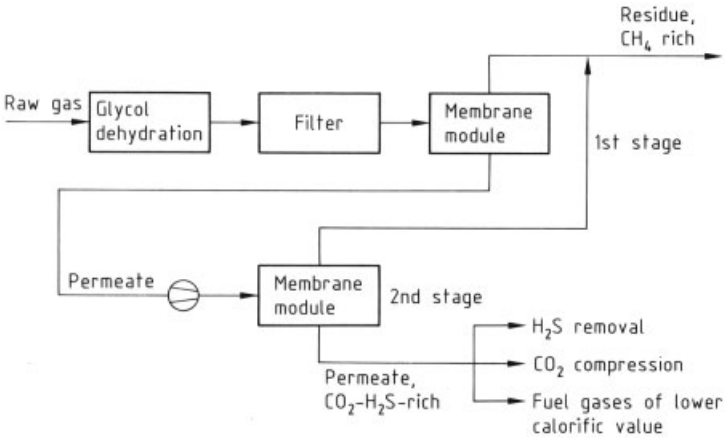

Figure 17. Two-stage membrane process

Table 9. Types of membranes and modules for carbon dioxide separation and dehydration [76 – 80]

Type of membrane and polymer	Type of module	Company	Selectivity $\alpha^{a}_{CO_2/CH_4}$, $\alpha^{b}_{H_2O/CH_4}$	Permeability of CO_2^a, H_2O^b m^3(STP) m^{-2} h^{-1} bar^{-1}
CO_2 Separation				
Asymmetric cellulose acetate	spiral-wound	Separex	25 – 30[a]	0.25 – 0.4[a]
Asymmetric cellulose triacetate	hollow fiber	Cynara, Dow	25 – 30[a]	0.25 – 0.4[a]
Asymmetric cellulose acetate	spiral-wound	Grace	25 – 30[a]	0.25 – 0.4[a]
Composite polysulfone/silicone (PRISM)	hollow fiber	Permea, Monsanto	17 – 24[a]	≈ 0.1[a]
Polyaramide	hollow fiber	Medal (Du Pont/Air Liquide)		
Asymmetric polyimide (6 FDA/DMB [79])	flat sheet (spiral-wound)	Grace	19[a]	0.4[a]
H_2O Separation				
Polysulfone, polyphenylene oxide (PRISM CACTUS)	hollow fiber	Permea		
Composite polyetherimide–ethylcellulose ether	flat sheet (envelope type)	GKSS	500[b]	35[b]
Asymmetric cellulose acetate	spiral-wound	Grace	300[b]	3.2[b]

predominantly by the Claus process. This process route to liquid sulfur is used widely in the natural gas industry [58].

High conversion rates of hydrogen sulfide are mandated worldwide by clean air regulations. Furthermore, the recovery of sulfur is also of economic interest. Liquid sulfur is stored and transported easily, and solid sulfur can be marketed in the form of slates, prills, and pills.

In the Claus process, hydrogen sulfide conversion occurs in a thermal and a catalytic step. The Claus reaction is based on the exothermic overall reaction

$$2\,H_2S + O_2 \rightarrow 2\,S + 2\,H_2O \qquad (1)$$

The first exothermic reaction step (Eq. 2) takes place in the combustion chamber at ca. 900 – 1100 °C:

$$2\,H_2S + 3\,O_2 \rightarrow 2\,SO_2 + 2\,H_2O \qquad (2)$$

Subsequently, sulfur dioxide reacts with additional hydrogen sulfide to form elemental sulfur (Eq. 3). This reaction occurs in the presence of catalyst in several consecutive steps at a stepwise reduction of temperature from 350 to 190 °C.

$$2\,H_2S + 2\,SO_2 \rightarrow 4\,S + 2\,H_2O \qquad (3)$$

Cooling of the process gas from the combustion chamber occurs in the waste-heat boiler, which is combined with a brick-lined combustion chamber. For each tonne of sulfur produced,

more than 2 t of steam is coproduced. Condensers are arranged between individual stages, in which the evaporation of water is used to cool the reaction gases as much as possible; the aim is to attain as high a degree of separation of elemental sulfur as possible.

Prior to the catalytic reaction, reheating of the gas occurs to the activation temperature of the catalyst (ca. 180 – 220 °C). Inactivation caused by condensation of sulfur must be excluded. In larger plants, both gas – gas heat exchangers and in-line burners are used. The latter must be run exactly and substoichiometrically because the Claus reaction is very sensitive to disturbances and catalysts are deactivated by reaction with oxygen. Attainable yields depend on the hydrogen sulfide content of the gas and the number of catalytic steps. The choice of catalyst and the heater – separator concept are also important. The pressure is predetermined: it is approximately the atmospheric regeneration pressure of the upstream gas scrubbing units. The pressure drop in sulfur recovery plants is 20 – 50 kPa.

In the combustion chamber and the waste-heat boiler, 50 – 70 % of the hydrogen sulfide is converted to elemental sulfur. The combustion chamber conditions and the construction of the burner—both multijet burners and single burners have been built for plants generating sulfur at up to 1500 t/d—influence the course of combustion and the formation of compounds that influence the process. Significant amounts of carbonyl sulfide, carbon disulfide, carbon monoxide, hydrogen, and sulfur trioxide are formed [59, 60].

Catalysts for the Claus process are generally based on aluminum oxide, which may also contain other metal oxides. Traces of sulfur trioxide, which may be formed by reaction of sulfur dioxide and traces of oxygen during start-up – shut-down, lead to rapid deactivation of the catalyst by formation of aluminum sulfate. Catalysts based on titanium dioxide are also available that are not sensitive to oxygen and hydrolyze more than 90 % of the COS – CS_2 formed in the combustion chamber, which would otherwise require additional measures in downstream off-gas purification plants [61, 62].

To achieve a high degree of sulfur recovery in a Claus plant, an exact amount of combustion air is required. This amount cannot be calculated precisely in advance because side reactions, such as the formation of hydrogen, carbon monoxide, carbon disulfide, and carbonyl sulfide, also have some effect on the optimal air quantity.

Continuously operating analytical devices for determination of hydrogen sulfide and sulfur dioxide concentration have been established. A sulfur recovery degree of 94 – 96 % can be attained with two catalytic reaction stages; with three reactors, > 97 % recovery can be achieved.

2.7.2. Tail-Gas Cleanup Processes

The sulfur recoveries attainable in Claus plants often do not meet the emission standards in force (exceptions are very small plants). Usually, various standards are set, depending on plant capacity. Mandated sulfur recoveries for plants in the Canadian province of Alberta are shown in Figure 18. Specific emission limits based on both the plant capacity and the hydrogen sulfide content of the sour gas are applied in the United States. In the Federal Republic of Germany a recovery of at least 99.5 % is demanded for all plants with a capacity of > 50 t/d [63 – 65].

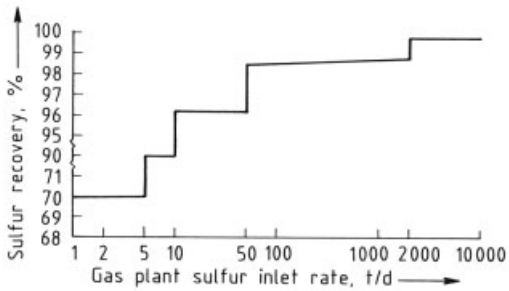

Figure 18. Sulfur recovery requirements for Alberta sour gas plants (August 1988)

Posttreatment of Claus tail gases is therefore generally required. A series of tail-gas purification processes is available which also produce elemental sulfur. The most important industrially applied tail-gas treating processes are described below [41].

Processes Utilizing the Claus Reaction. In the first group the formation of sulfur according to the Claus reaction from hydrogen sulfide and sulfur dioxide is performed in an additional catalytic step. This step is operated at a temperature below the sulfur dew point, thereby shifting the

equilibrium in the direction of sulfur formation. In the so-called *dry-bed* processes the reaction and precipitation of sulfur occur over aluminum oxide catalysts. When the maximum sulfur load is reached the catalyst is regenerated with hot hydrogen sulfide-containing recycle gas, and then tail gas is passed to an already desorbed and cooled reactor. Sulfur recoveries of 98 – 99 % can be reached in this way. Established processes include the following: Sulfreen, CBA, MCRC, and Maxisulf.

In the *IFP Clauspool* process the reaction is carried out in a high-boiling solvent that contains a catalyst; the sulfur formed is drawn off as a liquid. Yields correspond to those of the dry processes.

Sulfur recoveries of 99 % can also be attained with the *Superclaus 99 process*, first introduced commercially in 1989. A conventional Claus plant is run with a substoichiometric amount of air, so that practically no sulfur dioxide appears in the tail gas. Unreacted hydrogen sulfide is converted to sulfur by reaction with air in a downstream reactor using a special iron – chromium catalyst with a small specific surface area [66, 67].

Processes with Catalytic Conversion to Hydrogen Sulfide. In the second group, all sulfur-containing components in the Claus tail gas are catalytically converted to hydrogen sulfide. Further treatment takes place after cooling and condensation of a large part of the water vapor [68]. The following treating processes exist:

1) Absorption processes with regeneration. The hydrogen sulfide removed is returned to the Claus plant. The Scot, Sulften, and BSR–MDEA processes (→ Sulfur) are representatives of this group. To reduce energy consumption, absorbents are being developed which show better selectivity for hydrogen sulfide and lower energy requirements for regeneration (Flexsorb, Ucarsol).
2) Liquid oxidation processes by which absorbed hydrogen sulfide is converted in solution directly to elemental sulfur (Stretford, Sulfolin, and Unisulf process, see Section 2.4.2). Sulfur of lower quality is produced than with the other methods.
3) Catalytic processes in which hydrogen sulfide is converted to sulfur in the gas phase by addition of air, oxygen, or sulfur dioxide. The sulfur is recovered by condensation. Representatives of this group are MODOP [69], Selectox, and Ultra (→ Sulfur).

All processes in this group can be designed so that sulfur recoveries of 99.5 % are attainable. With absorption processes, a recovery of > 99.7 % is possible.

Processes with Catalytic Oxidation to Sulfur Dioxide. This group encompasses processes in which all sulfur-containing components are oxidized to sulfur dioxide. After cooling, sulfur dioxide is removed by absorption. The sulfur dioxide-rich gas arising from regeneration of the laden solvent is recycled to the Claus plant. The Wellman – Lord and Solinox processes, in which a sulfur recovery of > 99.5 % can be achieved, is commercially important.

2.7.3. Incineration

With only a few exceptions, the off-gas from Claus plants or tail-gas cleanup plants must be fed to an incinerator in which the residual sulfur-containing components are oxidized to sulfur dioxide. For thermal incinerations 600 – 800 °C is necessary, whereas catalytic plants require only 300 – 350 °C. All these processes lead to an off-gas with < 10 mg/m^3 of hydrogen sulfide. If the off-gas to be treated contains other sulfur compounds such as carbonyl sulfide or carbon disulfide, then a value of 10 mg/m^3 can also be obtained in a thermal incineration plant, provided the temperature is sufficiently high. For the catalytic process in this case, however, up to 80 mg/m^3 of sulfur must be expected in the off-gas. In principle, energy recovery by waste-heat boilers or gas – gas heat exchangers is possible for both processes. Nevertheless, traces of sulfur trioxide can lead to condensation of sulfuric acid in the heat exchanger. To avoid corrosion, heat exchange must be limited which means that less energy can be saved [70].

2.8. Example of an Integrated Plant

A flow diagram of an operating plant for the purification, dehydration, and conditioning of two different raw natural gases is shown in Figure 19. The design capacity amounts to a maximum of

Table 10. Balance of a sour gas treating plant (nonselective chemical – physical scrubbing)

Quantity	Sour gas 600 000 m^3/h	Pure gas 473 200 m^3/h	Flash gas[a] 4500 m^3/h	Claus gas 127 000 m^3/h	Sulfur[b] 86.5 t/h
H_2S	10 vol %	max. 5 mg/m^3	ca. 10 ppm	47.2 vol %	
CO_2	10 vol %	ca. 100 ppm	4 vol %	47.2 vol %	
N_2	5 vol %	6.3 vol %	3.7 vol %	0.1 vol %	
CH_4	balance	balance	balance	1.7 vol %	
C_{2+}	0.2 vol %		0.3 vol %	0.1 vol %	
Organic sulfur compounds	220 ppm	max. 5 mg/m^3		400 ppm	
Elemental sulfur	1 g/m^3				
H_2O dew point, °C	saturated	−8 °C	saturated	saturated	

[a] Fuel gas for internal use.
[b] Marketable product (highly pure liquid sulfur).

14×10^6 m^3/d for raw gas 1 and 5×10^6 m^3/d for raw gas 2. Characteristic data for the units processing raw gas are presented in Table 10.

Production of sulfur is set at 2100 t/d. The sour gases are predehydrated in the field (see Fig. 5) and are low in higher hydrocarbons. Treatment is thereby reduced to the removal of sulfur components and the adjustment of the inert-gas content. This special plant design has been chosen because of the following criteria [71]:

The high degree of carbonyl sulfide removal, which is necessary for underground storage of the pure gas, can be achieved with the chemical – physical Sulfinol scrubbing process. The latter is unselective but, depending on the ratio of hydrogen sulfide to carbon dioxide in the crude gas, may provide a sufficient Claus gas quality. The high energy consumption of solvent regeneration is also taken into account. The high-calorific pure gas can be stored in a nearby underground storage, but, after addition of nitrogen, can also be fed directly to the sales pipeline as low-calorific gas. The gas must be dehydrated because a water-containing scrubbing agent is used. The three-stage Claus plant attains a sulfur recovery of 97.5 %; the total recovery is increased to 99.8 % in a downstream tail-gas cleanup plant using the Scot process. After degassing, the sulfur produced is stored as a liquid and transported by rail. Sometimes the sulfur is solidified to prills or pellets outside the plant. Steam generated in a Claus plant is used in the reboiler for heating the Sulfinol solvent. Expansion turbines are employed for depressurization of the rich Sulfinol, which produce surplus electricity. In many cases the energy recovered from expansion, heat exchangers, and Claus waste-heat boilers is sufficient for regeneration and running the pumps. The total energy requirement of the sour gas treatment described here is ca. 3 % of the feed gas. One of the greatest consumers of energy is the tail gas purification plant.

Degradation of the DIPA content of the solvent to oxazolidone occurs because of the high carbon dioxide content of crude gas. The oxazolidone is regenerated to reusable scrubbing agent in a secondary plant. In another secondary plant a sulfur-dissolving agent, which prevents sulfur plugging of the wells (see Section 2.3.1), is regenerated for further use.

Lean raw gas 2, which contains a low level of hydrogen sulfide, is treated in a Purisol scrubber. Carbon dioxide remaining in the raw gas helps in adjusting the calorific value in the total pure gas stream from the Sulfinol unit. The flash gases are so low in hydrogen sulfide that they can be burned in a boiler.

Auxiliary Plants. Natural gas treating plants are connected to the public energy distribution network. Reliable auxiliary plants are required to guarantee a high level of availability and independence.

Steam Generation. Waste heat from the Claus process is used in the processing plant to drive process pumps and turbine-driven generators, and to heat solvents. The high-pressure steam generated is first expanded in turbines. The saturated low-pressure steam obtained can be used reboiling of the solvent at constant temperature. The reboiler steam pressure is 0.5 – 0.7 MPa and is adjusted to match the thermal stability of the solvent. Waste-heat boilers in the Claus plant operate at steam pressures of 1.6 – 4.0 MPa; the sulfur condensers at 0.1 – 0.6 MPa. By proper choice of driving equipment

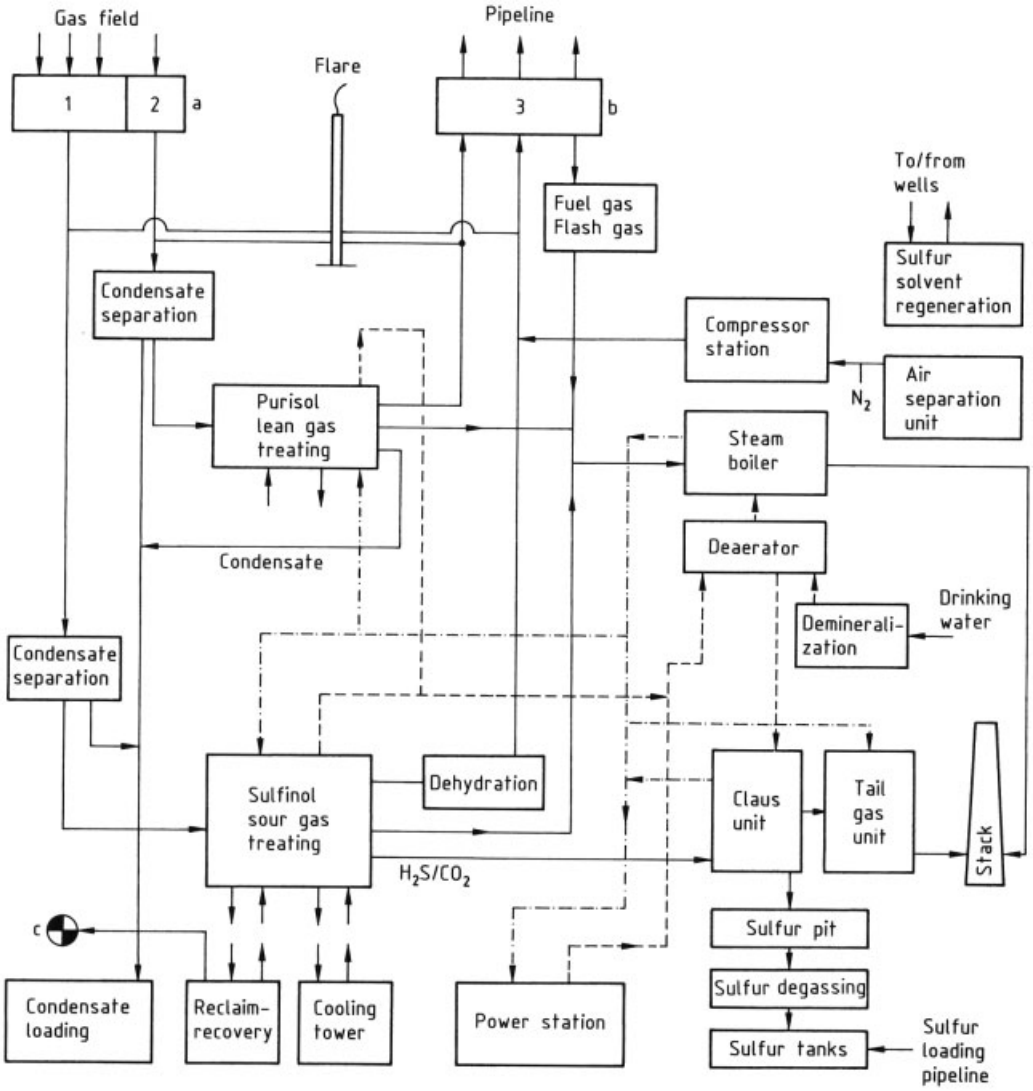

Figure 19. Flow diagram of a natural gas treating plant in the Federal Republic of Germany
a) Sour gas manifold; b) Town gas manifold; c) Disposal well
– · – · – Steam; – – – – Condensate

and steam pressure, energy requirements can be balanced so that only a little external energy is needed. Small amounts of blow-down streams must be continuously removed from boilers and condensers to avoid corrosion. These streams must be replenished by water from a treatment plant and a boiler feedwater degassing plant.

Electrical Power. Large compressors, circulating pumps, and blowers are often driven by steam turbines to achieve a better control of their speed. Frequently, the plant's own power station is only used for start-up and shutdown, whereby both steam- and gas turbine-driven generators are employed.

Inert Gas. Occasionally, purified natural gas must be mixed with inert natural gas to meet specified quality standards. When the admixing of air is not possible because of limitations of the oxygen content, inert gas plants are required. Inert gases produced from both catalytically purified flue gases and air separation are used. The dry gases are generally produced at low pressure

and brought to pipeline pressure (ca. 10 MPa) by multistage compressors. Inert gases are also employed for safe start-up and shutdown of plants and for regeneration of the Claus catalyst [72].

Tank Storage. Both higher hydrocarbons and sulfur are stored in tanks as liquids. They are transported further by road, rail, and frequently, pipeline. With special thermal insulation, liquid sulfur transportation through pipes of more than 8 km in length is possible without additional heating [73].

Wastewater. Natural gas treating plants produce little noxious wastewater. The boiler- and cooling tower blow-down water is fed to public sewage systems. Waste waters from piping and equipment cleaning are collected and treated in a reclaimer. Remaining residues are injected with the salt-containing formation water in separate disposal wells.

Gas Mixing. Different field qualities are mixed in the so-called raw gas manifold, so that the absorbers can be charged with gas of constant composition. Special separators minimize the entrainment of liquids and impurities that would damage the solvent or cause foaming. Pipelines from the field can be charged from the manifold with pigs for cleaning purposes. Mixing of different pure gases to adjust the Wobbe index and the net calorific value is possible in the pure gas manifold.

Reclaimer, Filter. Through side reactions, chemical scrubbing solvents in particular become enriched with degradation products, frequently of unknown nature. These substances are corrosive and may cause foam formation in the absorbers. Solvent quality is stabilized by removing a small side stream and addition of fresh solvent. This side stream is processed in a *reclaimer*, which is usually an auxiliary distillation plant. Purified absorbent can be reused in the process [74].

With physical scrubbing agents, removal of undesired substances (e.g., hydrocarbons) is frequently achieved by addition of water and phase separation. Mechanical filters are used in a side stream.

Special Construction Features. High partial pressures of carbon dioxide and hydrogen sulfide in a wet environment lead to serious corrosion. In particular, the stress corrosion cracking (SCC) caused by hydrogen sulfide corrodes high-strength steel from inside: atomic hydrogen can be formed by acid corrosion in connection with water, and can diffuse into the steel because of its small size of the atom. The hydrogen atoms recombine in the area of microcracks and local inhomogenities; the resulting pressure, which can reach 10^5 MPa, leads to cracks and blistering (hydrogen-induced cracking).

High-alloy hard steels are particularly susceptible; the necessary construction and testing regulations must be adhered to. Consequently, the carbon content of the steel is limited. Equipment and pipeline welding joints must be annealed. A hardness H_{RC} of < 22 must be maintained. Excessive stresses (particularly at plate edges) and thickness changes which are caused by the fabrication techniques must be avoided. Specially designed forged nozzles based on finite element analysis are customary [75].

3. Liquefaction

The principal reason for liquefying natural gas is a 600-fold reduction in volume that occurs with the vapor-to-liquid phase change. In its naturally occurring vapor state, natural gas is a bulky energy source, which is sometimes cumbersome to handle. Storage of the vapor requires huge underground caverns or large telescoping storage tanks. Transporting natural gas from production sources to points of consumption necessitates large pipeline networks. Thus, only overland or somewhat shorter undersea routes can be considered. Finally, natural gas at atmospheric pressure is too bulky to be considered as a fuel for transportation purposes; at the very least, it must be compressed.

Liquefaction of natural gas serves to overcome these obstacles, and permits transport over larger distances and more diverse application of liquefied natural gas (LNG) as an energy source. For example, LNG produced in the Middle East and throughout the Pacific Rim supplies ca. 10 % of Japan's primary energy consumption. Storage of LNG near urban areas permits peak demands for natural gas to be satisfied without building additional pipelines that would be underutilized most of the time.

Compressed natural gas is a feasible transportation fuel for truck and bus fleets. However, more recent aircraft and railroad applications

[87] require the higher energy density of LNG to be commercially viable.

3.1. Physical Properties of LNG

In the nomenclature of the energy industry, LNG is on the light end of the liquefied gas spectrum that also includes natural gas liquids (NGL, primarily ethane with some propane) and liquefied petroleum gas (LPG; → Liquefied Petroleum Gas, which consists mainly of propane with a significant proportion of butane). Common to each of these products is the lack of a standard set of physical properties. Nevertheless, certain characterizations of LNG are possible and useful.

Because methane is its primary component, the boiling point of LNG at atmospheric pressure is ca. −160 °C. Liquefied natural gas is odorless and colorless. Trace contaminants, especially mercury, can render LNG potentially corrosive if they are not removed to sufficiently low levels during the prepurification (see Chap. 2) and liquefaction processes. Liquefied natural gas is a commercial product, so the need to avoid solidification of trace contaminants is buttressed by rigid specifications for an environmentally acceptable fuel. Therefore, contaminants such as water vapor, carbon dioxide, and sulfur compounds, which may be found in vapor-phase natural gas, generally do not exist in LNG. Similarly, the relatively low boiling point of LNG limits the maximum concentration of pentane and heavier hydrocarbons that can exist in the liquid phase.

3.2. History

Early attempts to commercialize LNG failed. Barge transport between stockyards in Chicago, Illinois and gas fields in Louisiana did not employ adequately designed storage [88]. Similarly, a prototype peak-shaving storage tank in Cleveland, Ohio failed in 1944 because of poor material selection. The resultant spill from the tank vaporized, traveled through an adjacent sewer system, and caused a number of remote casualties upon ignition. These early setbacks, as well as the subsequent explosion of a poorly designed LNG tank undergoing repair on Staten Island, New York in 1973, did much to hinder the commercialization of LNG in the United States. Much of the early development centered around Europe and North Africa.

The first *baseload* LNG trade began in 1964 from Algeria to Canvey Island in England and to Le Havre in France. Subsequent projects were built in Libya and Algeria to supply LNG to Europe and its expanding pipeline grid. Japan became a purchaser of LNG in 1969, with the receipt of its first cargo from Kenai, Alaska. Since that time, Japan has become the preeminent buyer of LNG in the world. Major production facilities now also exist in Indonesia, Qatar, Abu Dhabi (United Arab Emirates), Brunei, Malaysia, Australia, Oman, Nigeria, Trinidad, and Egypt. As of 2005, more than 120 million metric tons per year of LNG is being produced, and facilities under construction will produce an additional 60 million metric tons per year.

Peak-shaving applications of LNG expanded rapidly in the late 1960s and throughout the 1970s in both Europe and North America. This growth mirrored the increased use of natural gas in urban areas and, in particular, the preponderant use of natural gas for heating. Virtually no growth of the peak-shaving business occurred in the 1980s because of a number of diverse factors, ranging from energy conservation and environmental siting restrictions to changes in governmental regulations.

Experiments in LNG *fueling of transportation vehicles* have been carried out for more than 20 years. Several driving forces exist for this work, including the environmental advantages of natural gas combustion compared to conventional fuels. However, no major infrastructure for LNG fueling exists to support the conversion of automobiles and trucks. Current development for railroad and marine vessels is, therefore, focusing on closed-loop applications in which the vehicle returns to the fueling station as part of its normal routing. Fishing vessels may also profit from the low temperature of LNG, which can be used to refrigerate the fish that are caught. Similarly, to best utilize LNG as a fuel for high-performance aircraft, the leading edges of the wings may have to be cooled with LNG to reduce skin-friction drag, while the LNG is being vaporized for use in the aircraft's engines.

3.3. Production

Because of the high critical pressure of methane, LNG must be produced commercially by refrigeration rather than pressurization. A number of processes have been commercialized for LNG production.

Common to all natural gas liquefaction processes is the need to pretreat the gas feed for removal of components that would freeze out (thus restricting the flow of the process), cause corrosion, or lead to pollution upon combustion of the revaporized LNG (see Chap. 2).

3.3.1. Baseload Liquefaction Processes (→ Refrigeration Technology,→ Cryogenic Technology)

Large-capacity or baseload liquefaction of natural gas is performed with emphasis on process efficiency. The scale of operations means that production with lower installed power capacity and lower fuel consumption is economically favored. Because the heat that must be removed from natural gas to cool it to −160 °C is eventually rejected to ambient air or water, rather elaborate systems have resulted.

Cascade Refrigeration Processes. The earliest liquefaction processes employed cascades of single-component refrigerants in series (→ Refrigeration Technology) (see Fig. 20). Each refrigerant is established as a separate closed-loop refrigerator that supplies refrigeration at discrete temperature levels. Typically, propane, ethylene, and methane are used to provide a wide, balanced range of refrigeration. After compression (a) three temperature levels for each of the three refrigerants form a nine-stage cascade. Each of these temperature levels corresponds to a preset pressure letdown (d) for evaporating the refrigerant in heat exchange with the natural gas feed and a separate refrigerant stream that requires cooling (e). In this fashion, heat is removed from the natural gas at successively lower temperatures; i.e., the refrigerant is boiled at successively lower pressure. Heat is rejected to ambient air or water via the warmest refrigerant, generally propane, and compressor aftercoolers (b). The methane refrigerant loop is open in that it is combined with the natural gas feed and after the final pressure letdown the liquid methane forms part of the LNG product. Cascade processes allowed use of single-component refrigerants at a time when thermodynamic correlations and thermophysical data bases were not as well developed as they are at present. In addition, the processes could be made very efficient (i.e., the amount of irreversibility could be reduced) by increasing both the number of refrigerants employed and the number of boiling stages for each refrigerant. This improvement in efficiency, however, had significant cost penalties. Each refrigerant requires its own compressor, compressor driver, vessels, and heat exchangers, together with attendant piping, insulation, and instrumentation. Each additional stage of boiling adds to the number of heat exchangers and vessels and to the number of sidestreams entering the compressor.

All-Mixed-Refrigerant Processes. The problems of complex layout, relatively high capital cost, and limited train capacities inherent in cascade processes were addressed by the mixed-refrigerant processes that succeeded them. As equipment and analytical capabilities advanced, the refrigerants could be combined in a single refrigeration loop (see Fig. 21). In such processes, a single refrigerant mixture of pentane, butane, propane, ethane, methane, and nitrogen, for example, is constituted to match the cooling curve of the natural gas to be liquefied; i.e., the temperature – enthalpy warming curve of the mixed refrigerant closely tracks the natural gas cooling curve in (g). This reduces the irreversibility of the process. The same refrigerant is then condensed, a part against air or water at (d) and the balance at lower temperature in heat exchange against itself in (f). In the latter case, incoming gaseous refrigerant is cooled and condensed at elevated pressure in countercurrent flow against evaporating liquid refrigerant fractions, each flashed (h) to a much lower pressure and, hence, temperature. The all-mixed-refrigerant process has performed well in a number of plants and represents a simplification compared to the cascade process. However, it is not thermodynamically efficient enough to be economical in the face of rising energy prices. To match the cooling curve of the natural gas feed over the wide range from the temperature of cooling water or ambient air down to the liquefaction temperature, compromises in the

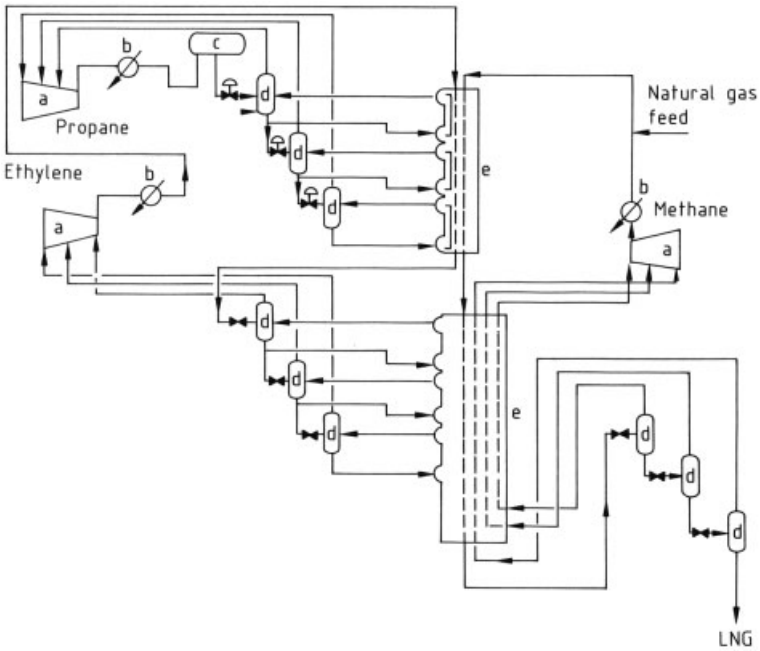

Figure 20. Nine-stage cascade liquefaction process
a) Compressor; b) Condenser; c) Accumulator; d) Phase separator; e) Heat exchanger

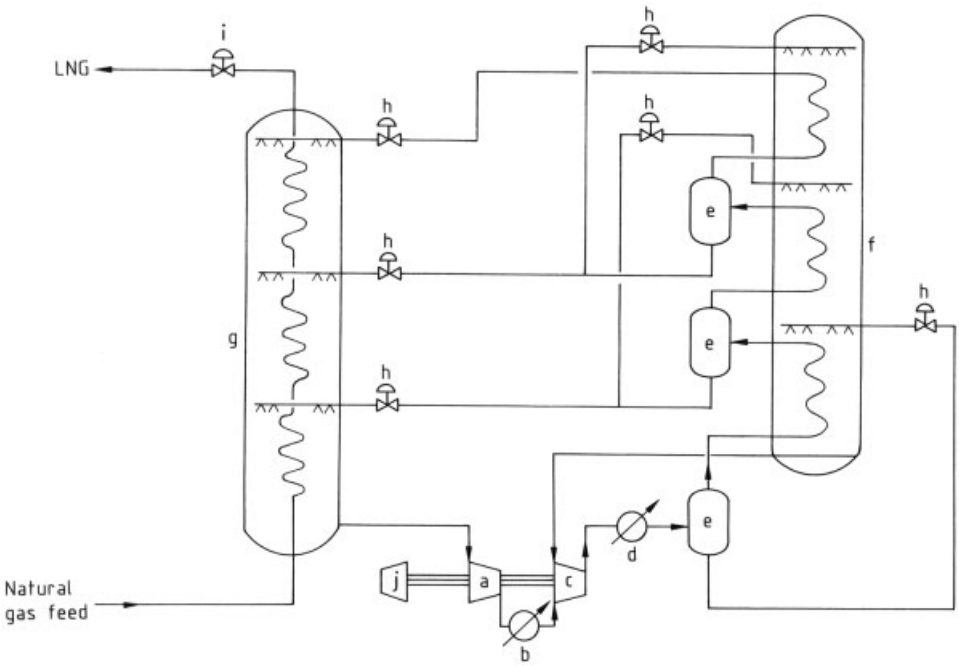

Figure 21. All-mixed-refrigerant (MR) liquefaction process
a) First stage MR compressor; b) Cooler; c) Second stage MR compressor; d) Cooler; e) Phase separator; f), g) Heat exchangers; h) Joule – Thomson valve; i) Product letdown valve; j) Driver

mixed-refrigerant composition are necessary. The wide range of boiling points for the refrigerant components also means that some of the heavier components are compressed to higher pressures than actually required for their condensation to ensure condensation of the lighter, lower boiling components such as nitrogen and methane. Such a recompression penalty cannot be avoided without a certain degree of separation of the refrigerant components such as occurs in precooled mixed refrigerant processes.

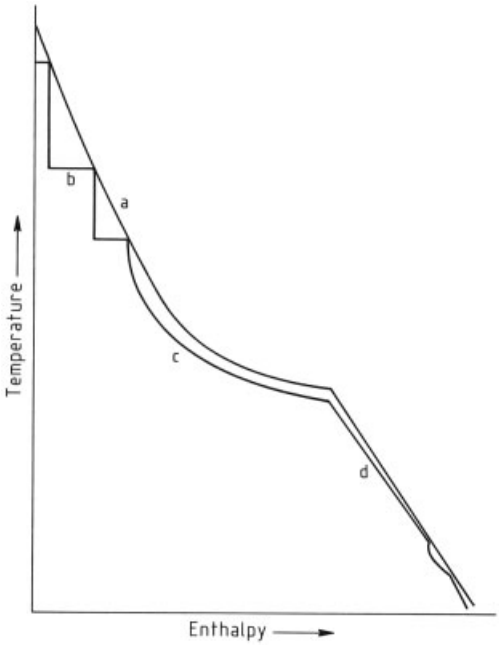

Figure 22. Cooling curve for a propane precooled mixed-refrigerant liquefaction process
a) Cooling curve of natural gas; b) Propane cascade; c) Mixed refrigerant (stage 1); d) Mixed refrigerant (stage 2)

Precooled Mixed-Refrigerant Processes. In the early 1970s, a third generation of processes, precooled mixed-refrigerant processes, developed from direct combination of the other two. The most widely used process employs two separate refrigeration systems, a propane cascade refrigerant loop in series with a mixed-refrigerant cycle that incorporates propane, ethane, methane, and nitrogen as components [89] (Fig. 22 shows the cooling curves for this process). The *propane loop* precools the natural gas and serves as an intermediate refrigerant for heat rejection from the mixed-refrigerant section to cooling air or water. The propane section can be fabricated from lower cost carbon steel, whereas the lower temperature sections require aluminum or nickel steels. Thus breaking the cooling process at the propane stage is compatible with economical material selection. Finally, by reducing the range of cooling that must be achieved by the mixed refrigerant, its composition can be optimized and energy losses due to recompression significantly reduced.

The increase in worldwide demand for LNG has prompted an increase in the capacity of liquefaction trains in order to take advantage of the economies of scale. This is being achieved by improvements in the liquefaction process as well as increases in the equipment capacity. For the most widely used processes, capacity increases of refrigerant compressors, gas turbines, and cryogenic heat exchangers have led to train capacities of nominally 5 million metric tons per year.

For larger train capacities, a hybrid process cycle (known commercially as AP-X^{TM}) has been developed [90, 91]. In this process, a cycle similar to the precooled mixed-refrigerant process is used to precool and liquefy the LNG (see Fig. 23A). However, the liquefied LNG is subcooled in a nitrogen refrigeration closed-loop process cycle (Fig. 23B). Nitrogen gas is compressed, cooled to near ambient conditions with cooling water or ambient air, and then further cooled to cryogenic conditions by expansion to lower pressure. The gaseous nitrogen is then used to subcool LNG, after which it is returned to be recompressed, completing the refrigeration cycle. By employing the nitrogen refrigeration cycle to subcool the LNG, the mixed-refrigerant only has to cool the LNG to about −115 °C (versus −150 °C to °160 −C for the typical precooled mixed-refrigerant process). This allows for a larger production capacity without a substantial increase in equipment size.

Additional process schemes have been developed that fall in the category of precooled mixed-refrigerant processes. If ethane is added to the precooling refrigerant to form a *dual mixed-refrigerant* process, then the precooling temperature can be lowered in accordance with the operating conditions of the loop and the proportion of ethane added [92]. With this process the re-

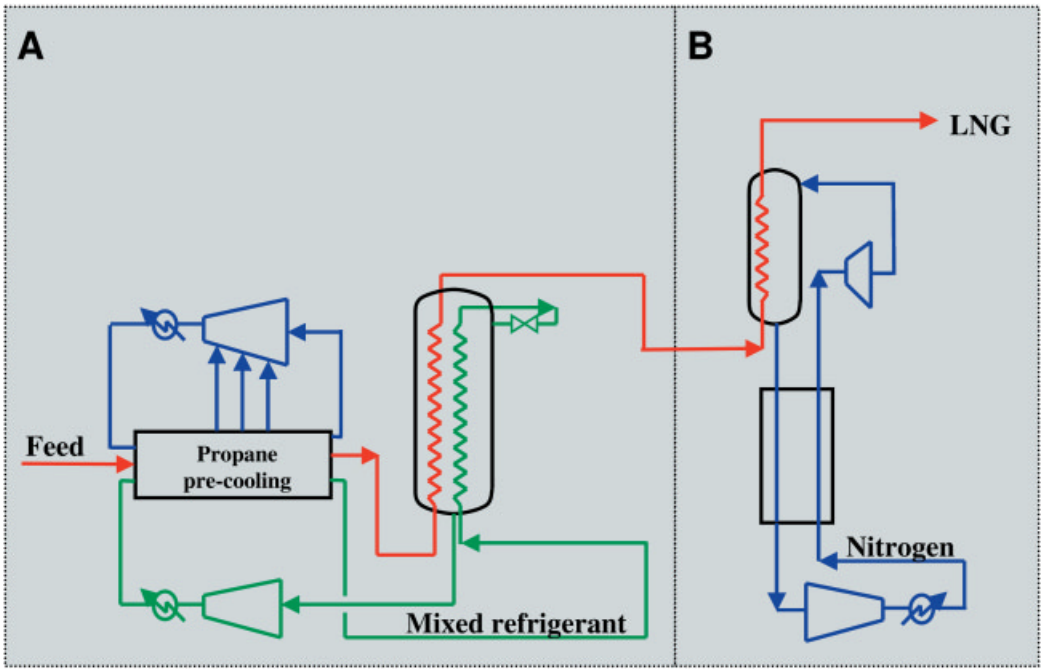

Figure 23. AP-X^{TM} process cycle
A) Propane precooled mixed-refrigerant process; B) Nitrogen refrigeration closed-loop process

frigeration duty can be shifted between the two refrigeration loops – a capability that may be useful in the face of changing feed gas conditions or changing power availability. The latter can occur because of ambient temperature changes that affect the power output of gas turbine drivers. A binary precooling refrigerant is somewhat more complicated to operate because simple pressure control is no longer adequate, but it may be preferred in certain scenarios.

Another precooled process has been proposed for plants that employ gas turbine drivers in the mixed-refrigerant section. In this configuration, waste heat from the gas turbines is used to separate ammonia and water in an *ammonia absorption refrigerator* [93]. This system could provide precooling duty while eliminating the need for costly compressors and drivers. This system has not been employed in an operating facility.

3.3.2. Peak-Shaving Liquefaction Processes

Peaking-shaving LNG plants differ from baseload facilities in several significant aspects that affect plant design. Peak-shaving plants are much smaller, operate only seasonally, and are often located near the point at which the pressure in gas trunk lines is decreased and the lines branch into lower pressure local gas distribution systems. The design emphasis for peak shavers is thus on capital cost minimization rather than high thermodynamic efficiency. *All-mixed-refrigerant* liquefaction cycles have therefore been employed in a large proportion of peak-shaving facilities. If a local distribution pipeline is available at a pressure substantially below that of the main gas trunk line, *expander processes* can be utilized to take advantage of the pressure difference as shown in Figure 24. Several variations exist, but the principle of these plants is to expand the inlet gas nearly isentropically through a turbine, thus rapidly lowering the temperature and partially liquefying the gas. The liquid is sent to storage, and the residual gas is compressed in a compressor mechanically linked to, and driven by, the expander. This gas is then sent to the lower pressure pipeline for distribution outside the plant.

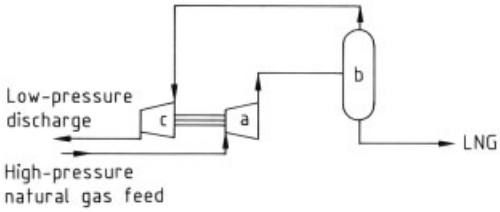

Figure 24. Schematic of an expander liquefaction process a) Turbine expander; b) Separator; c) Compressor

3.4. Storage

Storage tanks for LNG are an important part of both baseload and peak-shaving LNG facilities. In addition, they constitute the major portion of the investment in import receiving terminals. Because of the high cost of these units and their importance in the overall safety of LNG facilities, a great deal of attention has focused on LNG tank design.

One of the earliest efforts in tank design was directed at proper *material selection*. The Cleveland, Ohio tank failure in 1944 was attributable to the use of 3.5 % nickel steel which became embrittled at LNG temperature. Subsequent to this, large-scale programs have proved the suitability of 9 % nickel steel, stainless steels, and certain aluminum alloys (5000 series) for LNG tank fabrication. Designs with the latter have been somewhat limited in size, because the coefficient of thermal expansion for aluminum is approximately double that for steel; in large tanks, such thermal movement during cooldown could result in tank failure.

Tank designs have also evolved as more sophisticated safety analysis has been applied to LNG installations [94]. Early designs featured an internal cryogenic liquid tank enclosed by an outer tank that contained the insulation system for the inner one. In some designs the outer tank contained gaseous nitrogen and was, in turn, connected to a variable-volume or bladder tank, which compensated for changes in the nitrogen volume due to changes in ambient temperature while avoiding pressurization or depressurization of the outer tank. In other designs, the roof of the inner tank was not gastight but merely supported insulation, and the outer tank served as a natural gas holder (see Fig. 25). In both designs, the outer tanks were fabricated of carbon steel and surrounded by a low dike to contain any LNG spills.

Analytical studies indicated that the prime safety risk with an LNG spill is the formation of a large product vapor cloud that can drift, ignite, and cause widespread damage. Subsequent designs incorporated outer tank materials less prone to failure at cryogenic temperatures and taller dikes built closer to the tanks. These measures lead to less free surface for any LNG spill from a tank failure and therefore a lower feed rate to the ensuing vapor cloud.

Additional studies reviewed the consequence of an external impact, such as a crashing aircraft, leading to tank failure and the effect of a catastrophic inner tank failure on outer tank containment. These studies all focused on the need for much more secure outer containment. Resultant tank designs have double-integrity tankage; i.e., a liquid spill from a failure of the inner tank is contained by a second concentric tank that is structurally independent of the first [95]. The outer tanks have been built of reinforced concrete with a stainless steel liner. Finally, full-height earthen berms have been built to protect the tankage from external damage and to contain the postulated asymmetric loading on the outer tank resulting from catastrophic failure of the inner tank. Where the site is too small to permit full development of earthen berms, inground storage or additional reinforced concrete outer containment is employed.

3.5. Transportation

3.5.1. Peak Shaving

Peak-shaving LNG facilities typically vaporize LNG from storage and send it out via a local distribution network. In some installations, however, LNG is merely stored and vaporized for send-out. These satellite facilities receive LNG from other production sites, which is transported in over-the-road trailers. The trailers are similar to those employed in transporting liquid nitrogen and oxygen, and consist of an inner cryogenic vessel and an outer protective vessel of noncryogenic material. The space between vessels is typically filled with expanded perlite insulation and evacuated to subatmospheric pressure to further decrease heat in-leak.

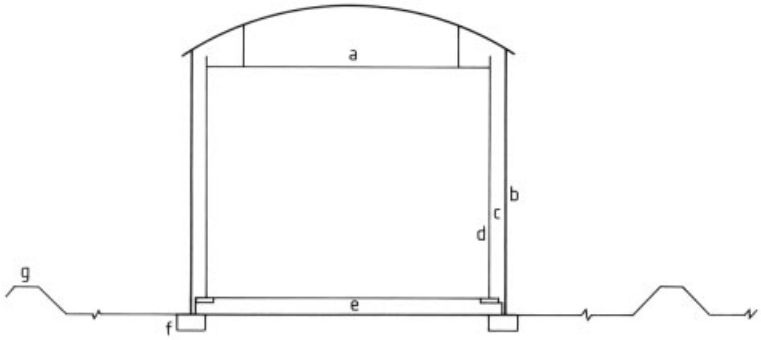

Figure 25. Simplified cross-sectional drawing of single-integrity LNG storage tank
a) Suspended deck; b) Outer tank; c) Sidewall insulation; d) Inner tank; e) Load-bearing insulation; f) Foundations; g) Dike

3.5.2. Baseload

Baseload LNG facilities are linked to their corresponding receiving terminals by fleets of large oceangoing tankers. Modern tankers typically can hold 125 000 m^3 of LNG. Each tanker employs one of several containment systems, which can be divided into three principal types. One design employs *free-standing spherical tanks* supported on cylindrical skirts (see Fig. 26). Each tank is insulated and located between bulkheads in a separate section of the ship. These compartments are blanketed with inert nitrogen gas, which is sampled periodically to detect any LNG leakage.

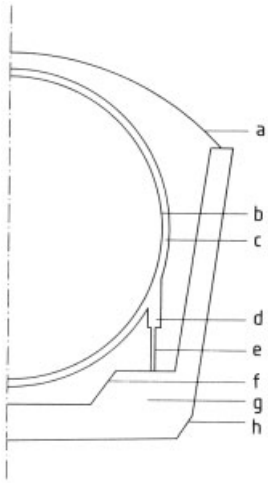

Figure 26. Spherical ship tank
a) Weather cover; b) Tank; c) Insulation; d) Support skirt; e) Thermal break; f) Inner hull; g) Ballast tank; h) Outer hull

A second family of designs is the *membrane tank*. In ships using this type, the insulation and a secondary barrier are applied to the hull of the ship (Fig. 27). The insulation also transmits the load from the LNG cargo to the hull and its structure. A metallic membrane containment barrier is then installed over the surface of the insulation. The membrane may consist of Invar (highly alloyed steel with an extremely low coefficient of thermal expansion) or of a "waffle" structure (made of nickel steel), where the cross section of the waffle allows two-dimensional thermal movement while maintaining intimate contact with the load-bearing insulation. Membrane tanks are also compartmentalized and enable more efficient storage use of the hull volume. This manifests itself in a flattop deck and a lower center of gravity, whereas spherical tanks extend up through the deck and require domed covers.

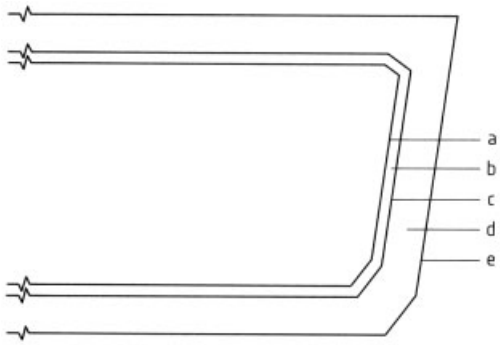

Figure 27. Membrane ship tank
a) Membrane liner; b) Insulation; c) Inner hull; d) Ballast tank; e) Outer hull

A third type of LNG tanker employs *free-standing prismatic* tanks. This design seeks to combine efficient use of hull space with the structural independence of a self-supporting structure. A number of earlier ships were built with this design type, but it is relatively expensive and, with the exception of two smaller ships for the Alaska – Japan trade, is not employed in newer tankers.

3.6. Safety Aspects

Liquefied natural gas may be viewed as a creation of the modern era, and as such, it has been subjected to more thorough scrutiny than traditional fuels such as gasoline and fuel oil. The hazards associated with LNG are that it is a low-temperature fluid, is flammable, and can create vapor clouds dense enough to cause asphyxiation. Because none of these is a problem when LNG is contained in properly designed storage tanks or piping and equipment, safety studies have generally focused on LNG spills. In LNG facilities, extensive safety systems exist to detect gas leaks, to detect and counter fires, and to detect smoke. Spill prevention and containment are implicit in the design of plant facilities, particularly of storage tanks because they contain the greatest quantity of LNG.

If an LNG spill is ignited soon after it occurs, a pool fire results. This fire is certainly a problem, but it is contained in a facility staffed and equipped to deal with such emergencies. A greater concern is that the same spill may be given sufficient time to evaporate and form a vapor cloud that can travel for some distance before ignition. A number of variables such as the nature of the surface underneath the spill, the wind velocity, and the presence of obstructions to the cloud act to shape and direct it [96]. Mechanically, the cloud is formed as LNG boils from the surface of the spill [97]. The evolving vapors are much colder than air and initially form a dense, low-lying cloud. As the cloud mixes with air and is warmed by its surroundings, it begins to rise. Wind serves to add a horizontal component to the cloud's motion. As air mixes with the natural gas, the mixture becomes flammable (i.e., local compositions are between the lower and upper flammability limits of LNG). The flammability limits vary with LNG composition, particularly the proportion of propane. This cloud is then potentially hazardous to areas beyond the battery limits of the LNG facility.

A number of large-scale tests have demonstrated certain characteristics of LNG vapor cloud fires. These fires are deflagrations-rapid combustions in which the flame front that moves through the cloud is preceded by a weak, decoupled shock wave [98]. An accompanying radiation hazard exists, but not the pressure damage of a shock wave that occurs when mixtures of other hydrocarbons and air are detonated.

4. Transmission, Storage, and Distribution

More than 50 % of world gas reserves are located in remote areas. For example, part of the gas used in Western Europe is produced in the harsh permafrost environment of Siberia or in the North Sea.

In most cases, producers ship the gas from the fields to the borders of the countries in which it is used. Importers buy the gas at these points under long-term contracts and resell the fuel to local utilities as well as to industrial users and power stations connected directly to the transmission system. Residential and commercial consumers are normally served by the local distribution companies. The European natural gas transmission system is shown in Figure 28.

Natural gas is used chiefly for heating. Therefore, gas demand varies substantially between winter and summer, weekdays and week-ends, or day and night. The ratio between summer and winter loads in Europe is between 1 : 5 and 1 : 10. Production, transmission, storage, and distribution facilities must be designed and sized to handle these load variations.

Of the gas traded across international borders, 75 % is shipped by pipelines, and 25 % is carried by LNG tankers.

Gas field development and the construction of transportation systems from distant areas of production to gas-importing nations are highly capital-intensive projects. Thus, the load factors at which gas is imported are nearly always very high. For load equalization, gas is stored in underground storage facilities during the off-peak

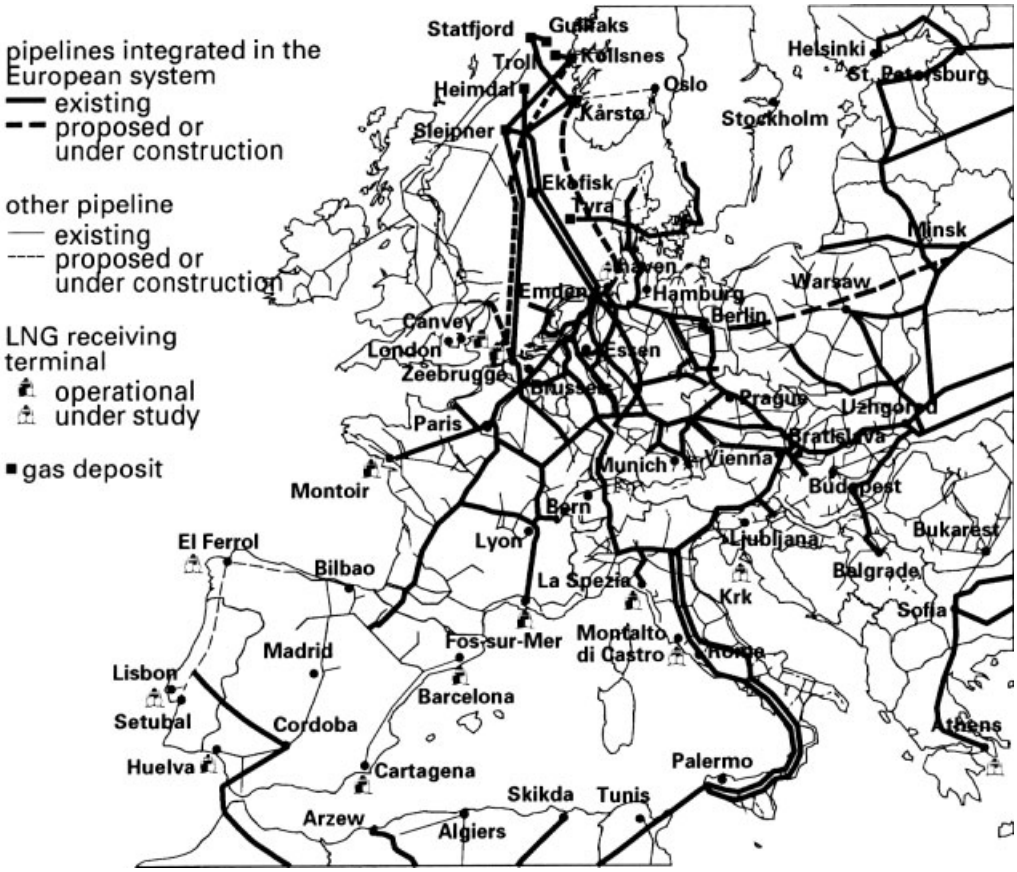

Figure 28. European natural gas transmission system

season and produced from storage during periods of peak demand in winter. Pipeline pressurization (line packing) in the transmission and distribution systems themselves, as well as peak-shaving installations, also help to handle load variations.

4.1. Transportation

4.1.1. Pipeline Transmission

Transportation of large gas volumes is best achieved by large-diameter pipelines operating at high pressure. Currently, the pipeline diameter may be up to 1400 mm and working pressure up to 8 MPa. Such pipelines carry gas at a flow rate of 2×10^6 m^3/h (STP) over a distance of ca. 1000 km. This capacity is insufficient, however, to ship gas from distant fields to markets. Compressor stations must therefore be built to increase the pressure of the gas. Gas compressors are normally driven by turbines or engines fueled by their own pipeline gas for reliability and lower cost. The number of compressors required for a given carrying capacity depends on the distance over which the gas must be shipped and on the compression ratio, which is given by

$$\pi = \frac{p_o}{p_i} \tag{1}$$

where p_o is the maximum compressor outlet pressure (bar) and p_i the minimum compressor inlet pressure (bar).

As the compression ratio increases, the number of compressors needed to operate a gas transmission line decreases. Hence, transmission engineers must seek to optimize capital outlay and follow-up operating expenses for pipelines and compressor stations.

For large-diameter pipelines, the optimum compression ratio is generally relatively low; it is between $\pi = 1.25$ and $\pi = 1.4$ [99]. Offshore pipelines are currently built with diameters up to 1000 mm. Pipelines from the North Sea to continental Europe are 1000-mm-diameter lines that are laid at a water depth of 150 m; Italy and North Africa are connected by 500-mm lines that cross 600-m-deep water. The very high working pressures at which offshore lines can be operated partly make up for the loss in capacity due to the smaller diameters. If necessary, pipeline bundles are laid [100].

Offshore compressor stations are extremely costly structures because they must be built on platforms. For this reason, the optimum compression ratio of offshore pipelines is higher than that of onshore pipelines. Depending on flow, it may be as high as $\pi = 4$.

4.1.2. LNG Transportation

Although the energy input required for liquefaction of natural gas is substantial (see Section 3.3), the volume advantage makes liquefaction economically viable. Cryogenic LNG is shipped by LNG tankers at atmospheric pressure. The cargo tanks are insulated to minimize evaporation. Nevertheless, the boil-off rate is between 0.2 and 0.25 % of the cargo per day. The gas that evaporates is used to fuel the ships engine. Reliquefaction is presently not viable.

The transportation of LNG is often the only way of bringing gas from distant gas-producing countries to market outlets. Any cost comparison between LNG transportation and pipeline transmission must, of course, be geared to the circumstances of the project. Generally speaking, an LNG chain is the only answer if pipeline transmission is not feasible for technical or other reasons or if the distance is sufficiently long. The cost of LNG transportation is lower than that of offshore pipeline transmission even for distances of several hundred kilometers, whereas onshore transmission is nearly always less expensive than LNG transportation unless the distance is extremely long [99, 101 – 103]. A comparison of costs of LNG transportation and pipeline transmission is given in Figure 29.

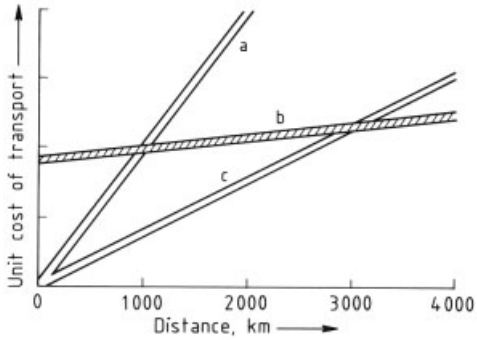

Figure 29. Cost of LNG transportation and cost of pipeline transmission as a function of transportation distance
a) Offshore pipeline (900-mm diameter); b) LNG transportation with liquefaction onshore (125 800 m^3); c) Onshore pipeline (1000-mm diameter)

4.1.3. Gas Transmission Pipeline Design

Pipelines for gas transmission must be designed so that, under peak load conditions, gas is allowed to flow from the point of delivery at the inlet of the system to the outlet of the system at a pressure between the maximum and minimum allowable operating pressure. Generally, pipeline capacity is somewhat lower in summer than in winter because of the following reason: The atmospheric temperature increases soil temperature and hence gas temperatures, and power output from the compressor engine decreases because of the higher temperature of the intake air.

According to Darcy and Weisbach, the pressure drop in a gas transmission pipeline is given by the following equation:

$$p_1^2 - p_2^2 = f \cdot \frac{16}{\pi^2} \cdot \frac{l}{d^5} \cdot \varrho_n \cdot q_n^2 \cdot \frac{T}{T_n} \cdot p_n \cdot K \qquad (2)$$

where p_1 and p_2 are absolute pressures (Pa); p_n is the normal pressure, 1.01325×10^5 Pa; $p_1^2 - p_2^2$ represents the pressure drop (Pa^2); T is the absolute temperature (K); T_n the normal temperature, 273.15 K; ϱ_n the gas density under normal conditions (kg/m^3); f the pipe friction factor; q_n the flow rate under normal conditions (m^3/s); l the length of the pipeline (m); d the diameter of the pipeline (m); K the gas flow deviation coefficient; and subscript 1 denotes the pipeline inlet and subscript 2 the pipeline outlet.

Equation (2) is usually transformed into the following equation in which the pressure is ex-

pressed in bar and the flow rate in cubic meters per hour:

$$p_1^2 - p_2^2 = 4.6401 \cdot 10^{-15} \cdot f \cdot \frac{l}{d^5} \cdot Q_n^2 \cdot \varrho_n \cdot T \cdot K \quad (3)$$

where Q_n is the hourly flow rate under normal conditions (m^3/h); p_1 the absolute pressure at the pipeline inlet (bar); and p_2 the absolute pressure at pipeline outlet (bar).

Equation (3) may be rewritten using the following terms

$$p_1^2 - p_2^2 = K \cdot c \cdot Q_n^2 \quad (4)$$

where

$$c = 4.6401 \cdot 10^{-15} \cdot f \cdot \frac{l}{d^5} \cdot \varrho_n \cdot T \quad (5)$$

$$c\ (\mathrm{bar}^2 \mathrm{h}^2 \mathrm{m}^{-6}) \quad (5)$$

According to NIKURADSE, the pipe friction factor f for *turbulent flow* and rough-surface pipes is given by [104].

$$f = \frac{1}{\left(2 \cdot \log \frac{d}{k} + 1.14\right)^2} \quad (6)$$

According to the Colebrook equation, the following applies to the friction factor in the transition regime between a hydraulically smooth and a hydraulically rough pipe surface:

$$\frac{1}{\sqrt{f}} = -2 \cdot \log \left(\frac{2.52}{\mathrm{Re} \cdot \sqrt{f}} + \frac{k}{3.71 \cdot d} \right) \quad (7)$$

where k is the surface roughness factor.

The *limit* between hydraulic roughness and the transition regime is given by

$$f_1 = \left(\frac{200}{\mathrm{Re}} \cdot \frac{d}{k} \right)^2 \quad (8)$$

where the Reynolds number *Re* can be calculated by using the following Equation:

$$\mathrm{Re} = 353.678 \frac{Q_n^2 \cdot \varrho_n}{d \cdot \eta} \quad (9)$$

where η is the dynamic viscosity of the fluid (10^6 P a · s).

If the pipe friction factor is f_1, then the flow is in the transition regime.

For the design of new lined large-diameter pipelines, the pipe roughness factor k (in mm) is used in flow calculations:

Hydraulically rough regime (Nikuradse) 0.012 mm

Transition regime (Colebrook) 0.006 – 0.008 mm

The gas flow deviation factor K describing the real-gas behavior of natural gases is another important term in flow equations. For most natural gases, K may be approximated by

$$K = K\ (p_m, T_m) \quad (10)$$

$$p_m = \frac{2}{3} \cdot \frac{p_1^3 - p_2^3}{p_1^2 - p_2^2} \quad (10)$$

4.1.4. Increase in Capacity

If the capacity of an existing pipeline must be increased, (1) new compressor stations can be installed or (2) pipeline sections can be looped.

Installation of New Compressor Stations. The increase in capacity that can be achieved by the installation of compressor stations is given by

$$\frac{Q_n}{Q_0} = \sqrt{n+1} \cdot 100\% \quad (11)$$

where Q_n is the capacity of the system with compressor stations (m^3/h); Q_0 is the carrying capacity without compressor stations (m^3/h) and n is the number of compressor stations.

To maximize the increase in carrying capacity achieved by the installation of compressor stations, the compressor station sites must be optimized. The optimum site depends on whether flow is constant throughout the pipeline system or varies because of offtakes between the inlet and the outlet of the system (see Table 11) [105].

Looping of Pipeline Sections. The increase in capacity due to the installation of compressor stations increases the system cost in steps [106]. For this reason, additional compressor horsepower is not a cost-efficient method of increasing capacity for all additional capacity requirements. Alternatively, pipeline sections may be looped to obtain exactly the additional capacity that is needed. The increase in capacity achieved by a loop line of the same diameter as the main line is given by the following formula (if no gas is taken off between the inlet and the outlet of the pipeline system):

$$\frac{Q_L}{Q_0} = \frac{1}{\sqrt{1 - \frac{3}{4} \cdot \frac{L}{L_1}}} \cdot 100\% \quad (12)$$

Table 11. Optimum compressor station sites

Number of compressor stations	Site, % of total length		
	Pipeline without offtake	Pipeline with offtake	Maximal transport capacity, %
0			100
1	50	21	141
2	33 / 67	13 / 31	173
3	25 / 50 / 75	9 / 21 / 38	200
4	20 / 40 / 60 / 80	7 / 16 / 26 / 42	224

where Q_L is the capacity after addition of a loop line of length L_1 (m^3/h); Q_0 is the capacity without looping (m^3/h); L is the length of pipeline between inlet and outlet (m); and L_1 the length of the loop line (m). The length of the loop line required to provide a specified additional capacity depends critically on whether or not gas is tapped from the pipeline system between the inlet and the outlet points [105].

The increase in capacity as a function of loop line length is plotted in Figure 30. If a loop line of the same diameter is assumed, 75 % of the main line must be looped to increase capacity by 50 % if all gas flows from the inlet to the outlet of the transmission system. If gas is taken off regularly between the inlet and the outlet (as in the case of a regional line), only 40 % of the original pipeline has to be looped.

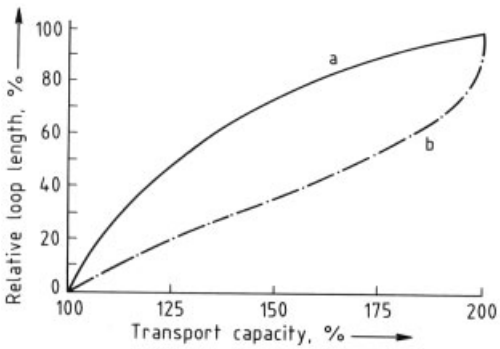

Figure 30. Increase in carrying capacity as a function of loop length
a) Pipeline without offtake; b) Pipeline with offtake

4.1.5. Transient Flow in Gas Transmission Systems

Flow rate and pressure in pipelines are a function not only of distance but also of time. Therefore, line packing effects play an important role. Simulation of transient flow, of course, produce different results than steady-state simulation. In a typical gas system, for instance, an hour of peak demand is usually preceded by a phase of low demand. As pressure increases when demand is low, a large gas volume is packed in the line when the demand peaks. During the period of peak load offtakes, the pressure in the line decreases because of the pressure drop resulting from the higher gas flow velocity through the line. Part of the gas packed in the line is therefore discharged from the system, covering peak demand.

Figure 31 shows the flows through a pipeline system into which gas is delivered at location 1 and from which gas is taken off at location 2 at different load factors. If the *flow rate* at the pipeline inlet (location 1) is constant (Fig. 31 A), inlet and outlet pressures will vary considerably and line packing will contribute substantially to covering peak load demand. As gas is normally delivered into a pipeline system at a continuous rate, the line will be repacked during periods of low demand, for example, during the night. If delivery into the pipeline system is at *constant pressure* (Fig. 31 B), the delivery rate will vary. However, the response at the point where gas is delivered into a pipeline system is always delayed with respect to the system behavior at the point where gas is taken off from the system; for this reason, delivery into the system and delivery from the system never coincide exactly.

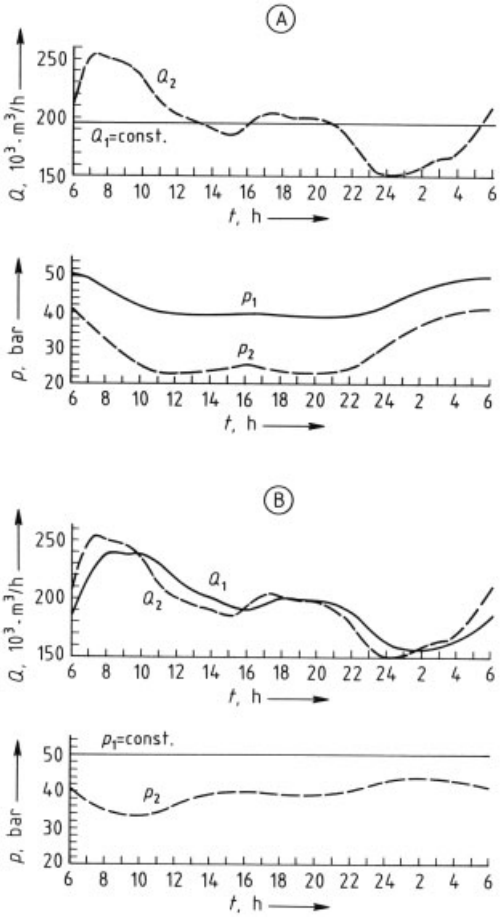

Figure 31. Transient flow in gas transmission systems
A) Constant flow rate at pipeline inlet; B) Constant pressure at pipeline inlet

4.2. Storage

Gas storage facilities are integrated into the pipeline system to match the high load factor at which gas is produced and delivered by suppliers to the pattern of end-user demand, which varies between winter and summer, over the week, and from hour to hour.

4.2.1. Determination of Storage Requirements

Gas demand is influenced by many different physical and economic parameters. Because substantial quantities of gas are used for space heating, ambient temperature is the most critical one. Typical average daily deliveries Q_{d_i} in Northwestern Europe as a function of temperature are shown in Figure 32. Gas demand increases as a linear function of temperature decrease and is constant above a certain limit temperature. The correlation is represented by the following equation:

$$Q_{d_i} = Q_{d_g} + m\,(t_l - t_i) \tag{13}$$

where Q_{d_i} is the daily gas demand at the daily average temperature t_i (kW · h/d); Q_{d_g} the daily baseload gas demand (kW · h/d); t_l the limit temperature for space heating (°C); t_i the average temperature on a given day (°C); and m the increase in deliveries per °C decrease in temperature (kW · h d^{-1} °C^{-1}). This equation allows sufficiently accurate predictions of daily gas offtakes [100].

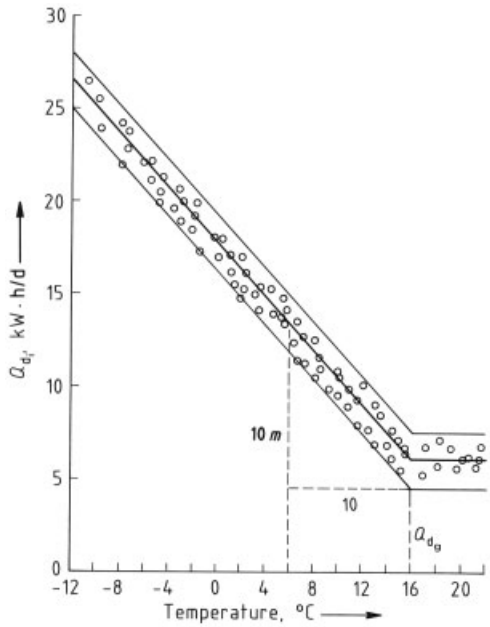

Figure 32. Daily average delivery as a function of temperature
10 m (slope) = increase in delivery per degree temperature decrease

Transmission, storage, and distribution facilities must, of course, be sized for the design day peak load on the system, determined by Equation (13). The ratio between system capacity speci-

fied for peak load conditions and capacity utilization is a crucial commercial and economic factor. Utilization time or system load is expressed differently in various countries.

In *Central Europe*, the concept of utilization time expressed in annual utilization hours or annual utilization days is frequently employed. *Utilization time* is given by Equations (14) and (15):

$$\text{Utilization hours per year} = \frac{\text{Annual gas volume}}{\text{Peak hourly volume}} \tag{14}$$

$$\text{Utilization days per year} = \frac{\text{Annual gas volume}}{\text{Peak daily volume}} \tag{15}$$

In the United States and the United Kingdom, the *load factor* expresses the same relationship. It correlates with the quantities given in Equations (14) and (15) as follows:

$$\text{Load factor} = \frac{\text{Annual volume}}{8760 \times \text{Peak hourly flow rate}}$$

$$\text{Load factor} = \frac{\text{Annual utilization hours}}{8760} \tag{16}$$

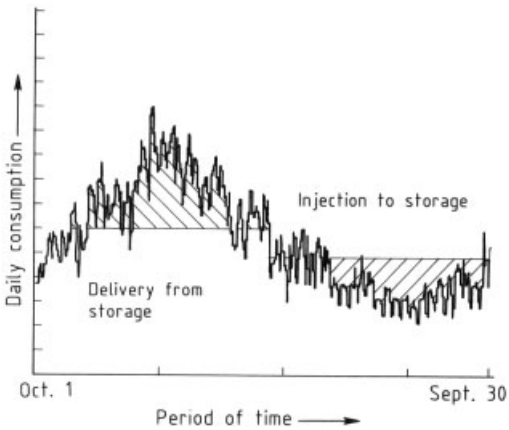

Figure 33. Daily load curve showing gas delivered from storage and gas added to storage

Irrespective of the actual load of the system, transmission, storage, and distribution facilities must be designed for adequate gas supplies regardless of ambient temperature or other conditions. With regard to storage facilities, gas deliveries from storage must at all times be sufficient to satisfy any demand in excess of gas supplies from gas fields into the transmission system. The volume of gas that is stored must be adequate to maintain this delivery throughout the period for which demand exceeds field supply. The necessary storage capacity is thus determined by the difference between the field supply load pattern and the demand load pattern, as well as by ambient temperature during the period in which supply is lower than demand. The volume of gas needed to satisfy peak demand must be stored from excess field supplies during off-peak periods. A typical daily load curve for a period of one year (for Northwestern Europe) is shown in Figure 33.

4.2.2. Storage Systems

Different systems are used by the gas industry to store gas during off-peak periods and to deliver gas from storage for peaking service. Various storage systems serve different peaking needs (see Table 12). The storage of gas in *low-pressure* containers has decreased substantially in importance because the quantity of gas stored in such a facility is insufficient.

Storage pipe arrays or *spherical gas tanks* in which gas is stored at high pressure provide peaking service for diurnal swings and short peaks within a week. Spherical gas tanks operate between 0.6 and 1.6 MPa and can have a diameter as large as 30 m. Pipes used for storage may hold gas at a pressure up to 8 MPa.

Line packing in *high-pressure* gas transmission pipelines is a method of handling diurnal peaks. For line packing, the pipeline must have a large diameter and must be designed for high working pressure. Line packing is feasible only if at least temporary spare capacity is available in the system. Normally, the transmission system design engineer plans a line packing capacity approximately eight times the difference between the hourly delivery rate and the peak hourly sendout into the system. This capacity corresponds to ca. 6 – 7 % of daily sendout.

Any demand peaks not provided for at the regional or the local level must be handled at the supraregional level. The quantities of gas that must be stored for seasonal peak shaving are substantial, and large underground storage facilities are practically the only system that will hold sufficiently large volumes of gas. Figure 34 shows the storage capacity that must be created to handle seasonal space heating load peaks as a function of supply load factor.

Table 12. Suitability of various storage systems for different peak needs

	Peaks			
	Daily	Weekly	Seasonal	System operator
Low-pressure container	×			local distribution companies
High-pressure container	×	×		
Line packing	×	limited		regional transmission companies
Cavern, fields	limited	×	×	supraregional transmission companies

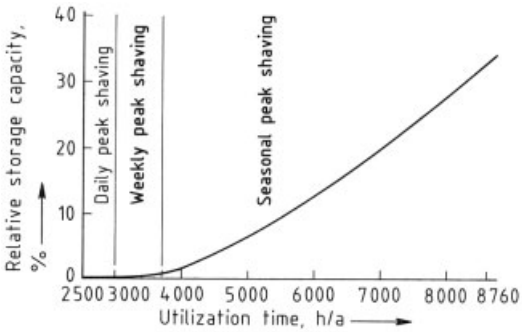

Figure 34. Storage capacity required for seasonal peak shaving as a function of the supply pattern

Gas is stored below ground in appropriate rock formations (see Fig. 35). Suitable formations for underground storage of gas include porous or fissured rock as well as caverns mined in salt rock, for example. These formations must be covered with impermeable caprock to prevent gas from escaping.

The following types of underground storage facility exist:

1) depleted gas fields, oil fields, and aquifer storage facilities (porous rock storage);
2) salt cavern storage facilities; and
3) mined cavern storage facilities (only of minor importance).

Underground storage facilities hold both working gas and cushion gas. *Working gas* is the gas that may be produced from storage for peak-shaving operations.

Cushion gas, sometimes referred to as base gas, is the gas required to ensure sufficient reservoir pressure for recovery of the working gas. It provides the pressure needed to cycle the normal working storage volume.

Porous Rock Storage. Gas can be stored below ground in rock pores. The rock should have sufficient porosity and permeability, and the storage horizon (which must be at sufficient depth) should be covered by tight caprock and be a structure that entraps gas injected into the rock. Structures used for porous rock storage include depleted oil and gas fields as well as aquifer structures. Development of a gas storage facility in a *depleted gas field* does not pose a major risk because data on rock properties and information on the tightness of the caprock are available. The working gas capacity, the maximum reservoir pressure, and the delivery rates as a function of reservoir pressure are known from production history. For the storage of gas in *depleted oil fields*, gas – oil separators are needed; however, secondary oil production by gas injection often makes a depleted oil field a commercially attractive storage site [107].

In *aquifer storage* reservoirs, interstitial water in the pores is displaced by the gas injected at high pressure. Considerable geological and reservoir engineering effort is required to decide whether an aquifer structure is suitable for underground storage of gas: i.e., the structure of the potential storage horizon must be evaluated; the thickness, porosity, and permeability of the storage rock must be determined; and the thickness of the caprock must be verified.

Salt Cavern Storage. Gas may also be stored in caverns leached into salt rock. Unlike the volume of a natural storage rock reservoir, the volume of these cylinder-shaped caverns can be controlled within certain limitations. The diameter of such a cavern may be as large as 80 m and its height 400 m. The cavern volume can reach 500 000 m^3.

The capacity of cavern storage facilities can be adapted to increased storage needs by adding further caverns in a given formation as required. The cushion gas volume required for operation of a storage cavern is much lower than the base gas volume in a porous rock reservoir and may amount to only 20 % of total capacity. Storage caverns allow the production and injection of gas

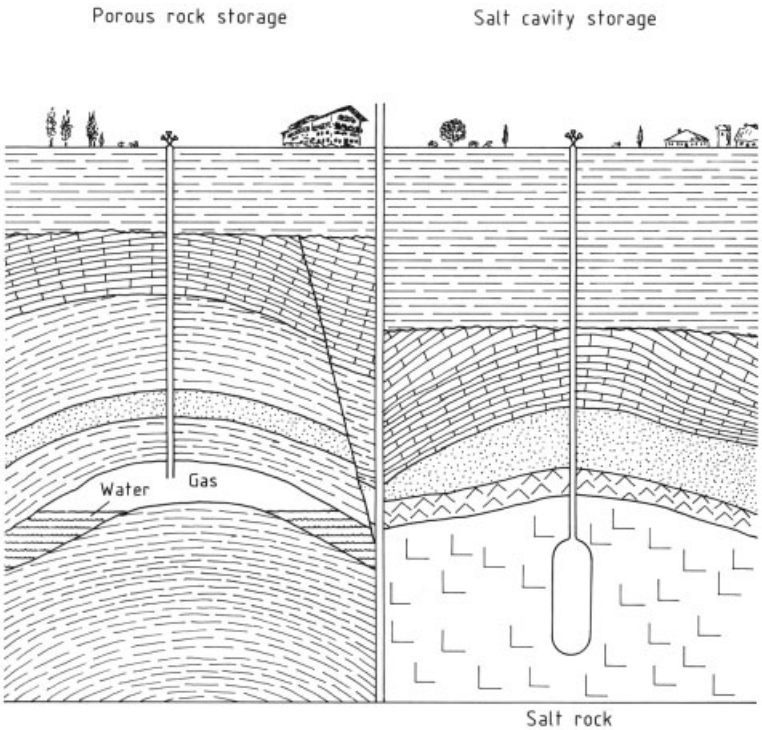

Figure 35. Underground storage facilities for natural gas

at very high rates that are limited chiefly by the capacity of the above-ground installations.

In salt storage caverns, convergence can occur because cavern leaching operations and pressure cycling after commissioning produce strain in the rock. As a result, salt flows into the cavern, reducing its volume. Within limits, convergence may be controlled by retaining a minimum quantity of gas in the cavern [104, 107 – 109].

4.2.3. Optimized Storage Planning

As the load curve in Figure 33 shows, storage facilities produce gas for only relatively short periods of time. Depending on supply and demand load patterns, gas withdrawal is required for peak shaving on 80 – 150 days per year. If the ambient temperature rises, the withdrawal phase may be interrupted on days in which demand is lower than supply thereby allowing more gas to be injected into a partially depleted storage facility.

With *porous rock* storage facilities, production rates are limited by the resistance of the rock to gas flow to the wells. Depending on depth, reservoir pressure, and permeability, wells may produce gas at rates between 1000 and 70 000 m^3/h (STP). At *salt cavern* storage sites, the withdrawal rate is limited only by pressure loss in the tubing and the design of above-ground facilities. Storage caverns today are normally designed for production rates of up to 250 000 m^3/h (STP). However, because gas cools rapidly during depressurization, production at such a high rate is feasible only for a relatively short duration up to 60 h in a limited number of production phases during the year. Similar restrictions do not exist for porous rock storage facilities, which normally hold substantially larger quantities of working gas that is produced at lower rates. The working gas volume at a salt cavity storage site, on the other hand, may be increased by leaching an appropriate number of caverns.

Transmission system operators optimize their peak-shaving operations by employing both

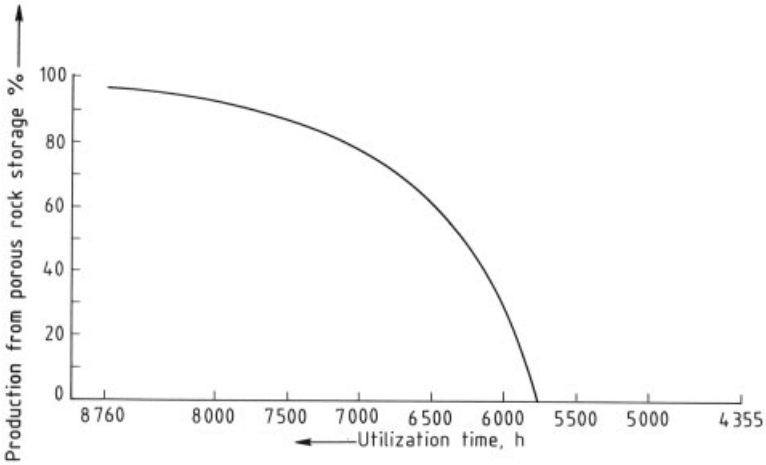

Figure 36. Optimum production of gas from porous rock storage and salt cavern storage for peak shaving

porous rock and salt cavern storage facilities. The peaking service baseload is provided by the production of gas from porous rock facilities at relatively constant load, whereas caverns are used in periods of maximum demand. The optimum employment of storage caverns and storage reservoirs depends, of course, on supply and demand load patterns (see Fig. 36).

4.3. Distribution

Distribution systems take gas from regional supply centers and transport it to residential, commercial, and industrial users. These systems consist of high-pressure, medium-pressure, and low-pressure grids. Distribution systems tend to be operated at low load factors because they serve a market in which demand varies considerably. If most gas is used for heating, the grid is designed for peak load on a winter day.

Distribution grids have grown in most areas over the years because gasworks were built in many cities long before natural gas became available. For this reason, distribution grids consist of mains of different diameters and different materials laid in different years.

Old distribution lines often operate at 2 – 8 kPa, whereas new distribution mains are frequently designed for a working pressure of 100 kPa. Low-pressure distribution grids are normally tied into medium-pressure or high-pressure distribution systems from which they receive gas at special supply stations.

Since local distribution networks consist of mains cross-connected at numerous points, computers are required to predict flow in distribution lines. Early models and methods for solving these nonlinear patterns were proposed by Cross in 1936 and Mac Illroy in 1949 [110]. Substantial computing time is required for such a nonlinear model, which originally consisted of very few nodes in view of the limited performance of early computers. As data-processing technology progressed during the 1950s, network analysis techniques become more widely used. They are now state of the art and are employed as a matter of course in distribution system design. Modern software models all the details of distribution systems.

Many of the input data required for network analysis can be determined only by statistical methods and field tests. For example, statistics are necessary to predict system loads controlled by the demand of numerous gas users whose meters are usually read only once each year. Pipe roughness, which is crucial for the hydraulic behavior of the grid, must be measured in many cases because pipe properties depend on the year in which the distribution line was constructed.

Distribution networks must also be designed to handle peak loads that can be 10 or 15 times higher than the baseload. To optimize the economics of gas distribution, distribution system

capacity is regularly adapted to growing demand. Methods of increasing capacity include the following:

1) replacement of mains or laying of new lines;
2) increasing distribution system pressure often to a maximum of 10 kPa in old systems and installation of service regulators to lower the pressure to 2 kPa at which most appliances are operated; or
3) installation of new medium-pressure or high-pressure system stations to feed gas into the low-pressure grid.

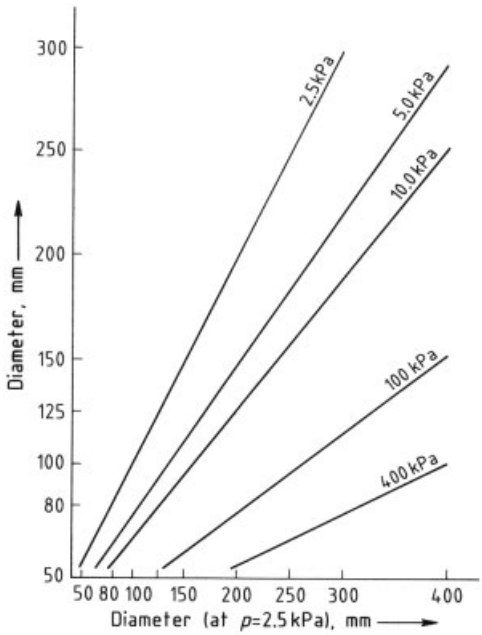

Figure 37. Pipeline diameter as a function of working pressure for constant pipeline capacity (in relation to pipeline diameter for 2.5-kPa working pressure)

New governor stations to interconnect ring mains and distribution lines constitute a particularly effective method of increasing distribution grid capacity [111, 112]. Local distribution companies normally combine all three methods to optimize the economics of gas distribution. The new distribution lines that are increasingly being built to develop service in rural regions are designed mainly for a pressure of 100 kPa, because the higher distribution pressure substantially reduces pipeline diameter for the same pipeline capacity (see Fig. 37). In fact, high distribution pressures are often necessary to serve rural areas because 100-kPa distribution lines can transport gas over much longer distances.

5. Economic Aspects

Natural gas is the world's third most important energy source, behind oil and coal, covering ca. 23 % of the 1995 world energy demand.

Reserves. Proven world gas reserves at the end of 1995 totaled about 150×10^{12} $m^3(V_{st}$ i.e., standard cubic meter of natural gas at 15 °C and 101.3 kPa with a net calorific value of 35.2 MJ or 9.776 kW · h) (equivalent to 5171×10^{18} J or 134×10^9 t of oil), corresponding to ca. 95 % of proven oil reserves [113]. A geographical breakdown is shown in Table 13. These reserves equal about 55 years of current annual gross production.

In addition, ca. 222×10^{12} $m^3(V_{st})$ (equivalent to ca. 7770×10^{18} J or 200×10^9 t of oil [114]) has been identified as probable gas reserves. Including this figure, the total 1995 gas inventory equals approximately 135 years of current annual gross production.

Production. The world's 1995 gross gas production amounted to 2.75×10^{12} $m^3(V_{st})$ (95.15×10^{18} J or 2.4×10^9 t of oil), of which about 80 % was marketed. The rest covers reinjection, flaring, losses, etc. The total 1995 marketed volume of 2.21×10^{12} $m^3(V_{st})$, broken down into major geographic zones, is shown in Table 14.

Consumption. Consumption patterns in various regions of the world generally follow production profiles, as shown in Table 14. The Former Soviet Union and Eastern Europe as well as North America both produce and consume ca. 30 % of the world's annual natural gas availability.

International Trade. Although marginal compared to total consumption, a sizable trade over national borders has developed since 1960. Initial international deliveries were mainly to neighboring countries through pipelines, whereas shipments of LNG developed in the 1970s. Table 15 summarizes the natural gas volume traded internationally in 1995. Of the total of 395×10^9 $m^3(V_{st})$ (equivalent to 13×10^{18}

Table 13. World's proven natural gas reserves as of January 1, 1996 [113]

Area	10^{12} m^3(V_{st})	10^{18} J	%
North America	6.5	222	4.3
Latin America	8.0	277	5.4
Western Europe	6.3	220	4.3
Former Soviet Union, Eastern Europe	59.2	2037	39.3
Africa	10.0	346	6.7
Middle East	45.3	1558	30.1
Asia, Oceania	14.8	512	9.9
Total	150.2	5171	100.0

J or 352×10^6 t of oil) traded, approximately 92×10^9 m^3(V_{st}) or 23 % was received in the form of LNG.

Table 14. World natural gas in 1995 [10^9 m^3(V_{ST})] [113]

Area	Marketed production	Consumption
North America	689	686
Latin America	101	103
Western Europe	241	350
Former Soviet Union, Eastern Europe	734	657
Africa	84	46
Middle East	145	145
Asia, Oceania	213	1220
Total	2207	2207

Pricing. Except for a minor amount used as chemical feedstock, natural gas is primarily employed as clean heating fuel in industry, power plants, commercial installations, and households. To maintain its competitiveness versus alternative energies, natural gas prices for the final consumer are generally determined by prices of competitive fuels such as light heating oil, heavy fuel oil, coal, LPG, or electricity, with due consideration of the inherent environmental, handling and financial advantages of gas utilization.

Since gas transportation costs significantly exceed those of oil or other fuels, gas netbacks at the wellhead are generally lower than those for crude oil. To compensate for the heavy front-end investment in production, purification, and transportation facilities, a large share of internationally traded natural gas is normally sold by producers under long-term contracts of up to 35 years' duration, which oblige sellers to make predetermined annual quantities available, whereas buyers are obliged to offtake certain minimum quantities per year (take-or-pay obligation: usually 80 % of availability). Prices of such long-term gas deliveries are generally tied to official international crude or fuel oil quotations.

6. Testing and Analysis

Natural gas analysis can be divided into two areas: chemical analysis of the composition, and

Table 15. International gas trade in 1995 [10^9 m^3(V_{ST})]

Supplies from	Imports to								
	Japan	Germany	United States	France	Italy	United Kingdom	Belgium	Others	Total
Former Soviet Union		32		13	14			58	117
Canada			79						79
Netherlands		25		6	4		5	1	41
Norway		11		7		2	3	5	28
Algeria			1	8	17		4	8	38
Indonesia	23							10	33
Malaysia	11							2	13
Brunei	7							1	8
United Arab Emirates	5								5
United States	2							2	4
Others	9	1						19	29
Total	57	69	80	34	35	2	12	106	395

determination of the physical properties. Chemical analysis methods for natural gas are summarized in [115, 116]. Instructions for both the chemical analysis and determination of physical properties for natural gases are given in [116].

For the analysis and quality control of gases, see also → Gas Production.

6.1. Analysis of Natural Gas Composition

The most important method for chemical analysis of natural gas is gas chromatography [117 – 123]. Sampling must be carried out before the gas can be analyzed.

6.1.1. Sampling

Glass equipment is mainly used for sampling because glass is not attacked by any component of natural gas and does not affect the composition of the gas [124]. Vessels with two taps (see Fig. 38) are used, which must be completely dry. The gas to be analyzed is first passed through the sampling vessel to completely displace the air from it. Both taps are then closed so that the sample is contained in the vessel. The sample should always be under slight overpressure, so that no air is drawn into the vessel even if the taps are not closed tightly.

Figure 38. Sampling vessel

6.1.2. Gas Chromatography

Gas chromatography has developed since 1952 and has replaced the Orsat method, previously used for analysis of natural gas (→ Gas Production, Chap. 8.2.2.1). Gas chromatography is the preferred method for most gas analyses. The analysis of natural gas by gas chromatography has been standardized in ASTM D-1945. In gas chromatography, the gas to be analyzed is transported by a carrier gas stream (usually helium or argon) through a separating column. Carbon dioxide may also be used as carrier gas according to the Janak (→ Gas Production, Chap. 8.2.2.1), [125]. Figure 39 shows the setup for gas chromatography [120] and Figure 40 an example of a gas chromatogram of natural gas [120]. The separating column is filled with a molecular sieve or with Chromosorb.

Table 16. Standard reference mixtures for the chromatographic analysis of natural gas

Component	Lean gas, mol %	Rich gas, mol %
Methane	85.0	74.0
Ethane	6.0	10.0
Propane	4.0	7.0
n-Butane	2.0	3.0
Isobutane	2.0	3.0
n-Pentane	0.5	1.0
2,2 Dimethylpropane	0.5	1.0
Isopentane	0.5	1.0
Hexanes and higher hydrocarbons	0.1	0.2
Nitrogen	4.0	0.5
Carbon dioxide	1.0	1.0
Helium	1.0	0.5
Hydrogen	3.0	0.5

Thermal conductivity detectors and flame ionization detectors are most commonly used, mass spectrometers may also be used for detection [119]. In some cases, two columns must be used – one for the separation of hydrocarbons, the other (for a new sample) mainly for determination of nitrogen and oxygen. Standard reference mixtures are used for two gas qualities: lean gas (i.e., gas containing few or no condensable hydrocarbons and thus having a lower calorific value) and rich gas (i.e., gas containing a higher amount of condensable hydrocarbons and thus having a higher calorific value) (see Table 16).

6.1.3. Analysis of Nonhydrocarbons

Natural gases can contain various gases apart from hydrocarbons: nitrogen, carbon dioxide,

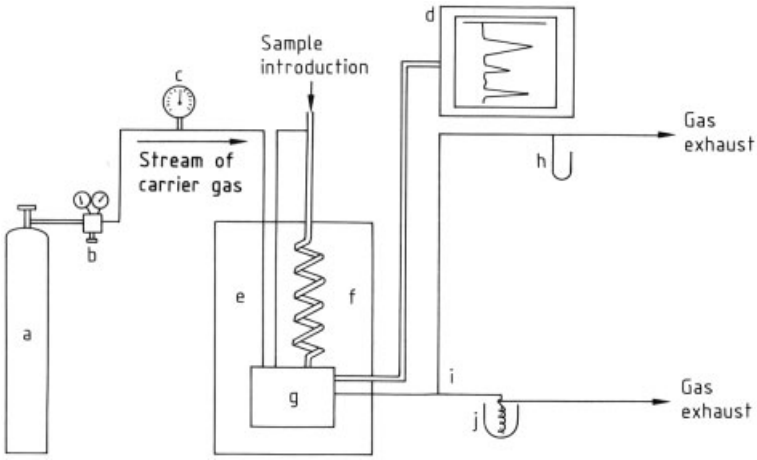

Figure 39. Gas chromatograph
a) Carrier gas; b) Regulating valve; c) Manometer; d) Chart recorder; e) Air thermostat; f) Separating column; g) Thermal conductivity detector; h) Flow meter; i) Three-way tap; j) Trap

and noble gases can be determined during gas chromatographic analysis of the hydrocarbons. A gas chromatographic analysis method for non-condensable gases in light hydrocarbons is given in [126]. The content of water, sulfur compounds, and mercury may also be determined.

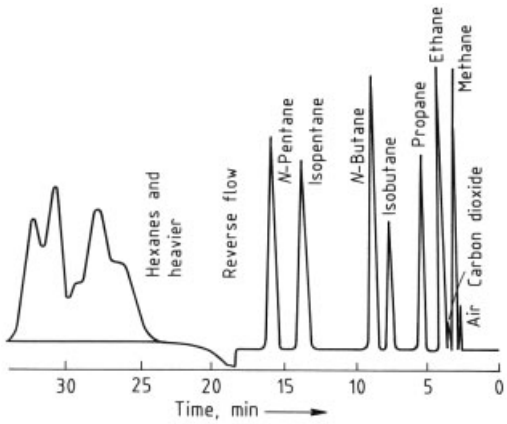

Figure 40. Gas chromatogram of natural gas

Determination of Water. Water must be removed from natural gas to prevent its condensation and to avoid the formation of gas hydrates [127] (see Section 1.2). Low water content of the order of milligrams per cubic meter can be titrated according to the classical *Karl – Fischer method*, in which water reacts with iodine in an alcoholic solution of sulfur dioxide and pyridine [128]. The iodine can also be determined coulometrically. Water can be determined gravimetrically by absorption using a desiccant such as magnesium perchlorate. A variety of *hygrometers* [129] measure, for example, the variation in the conductivity of glycerol with water content. Dew point measurements (see Section 6.2.4) indirectly give the water content.

Determination of Sulfur Compounds. Natural gases can contain hydrogen sulfide and occasionally mercaptans (see Section 2.4). The total sulfur content can be determined by combustion with excess air or oxygen. Either the sulfur dioxide formed can be determined, e.g., by means of a coulometric titration, or the combustion products can be absorbed in an oxidizing solution and the sulfuric acid formed is determined.

Hydrogen sulfide can be detected qualitatively using lead acetate producing a black coloration due to the formation of lead sulfide.

In the ASTM standard test method a measured volume of gas is bubbled through two solutions – first a neutral cadmium sulfate solution and then a basic cadmium sulfate solution [130]. Hydrogen sulfide is found in the neutral solution and the mercaptans in the basic solution. Hydrogen sulfide and mercaptans are then determined iodometrically.

Determination of Trace Components. Heavy metals, especially mercury (see Section

2.3.2), are the most important trace components. There are at present no standard determination methods; atomic absorption spectroscopy (AAS) is usually used.

6.2. Determination of Physical Properties

Natural gases are characterized and classified according to their physical properties. The *Wobbe index* (W) (see Section 1.3.2) is the most important characteristic for the classification of natural gases and they are divided into two groups: group L (low Wobbe index) and group H (high Wobbe index). The Wobbe index is calculated according to

$$W_S = H_S/\sqrt{d}\ (\mathrm{MJ/m^3}),$$

where H_S is the gross calorific value and d the relative density (based on air). In addition to the more commonly used superior Wobbe index W_S, the inferior Wobbe index W_I can be calculated, whereby the inferior calorific value H_I is used instead of H_S. The Wobbe indices indicate gas quality for combustion.

Usually the Wobbe index is calculated from H_S and d. Automated calorimeters are available for its direct determination (Wobbe recorders) which use a gas stream that is adjusted so that the flow rate is proportional to $1/\sqrt{d}$.

Important physical properties of natural gas include the following:

Density
Calorific value
Molecular mass
Dew point

6.2.1. Density

Density can be measured with a *gas density balance*. The method is based on the different buoyancy of a nitrogen-filled glass sphere suspended from a balance in various test gases [131]. The gas to be measured and nitrogen should have the same temperature and pressure. Instruments of this type also allow continuous measurement and recording of gas density in a gas stream.

A particularly reliable method is to weigh a specific volume of gas in a glass flask (e.g., a *Chancel flask* according to IP 59). Either the density of the gas (mass per unit volume) or the relative density d (the ratio of the density of the gas to that of dry air) is determined. The density and relative density can also be calculated from the composition determined by gas analysis [132].

6.2.2. Calorific Value

The calorific value in MJ/m^3 can be determined in three ways: (1) directly by combustion of the gas at constant pressure in a gas calorimeter, (2) by stoichiometric combustion and measurement of the gas : air ratio, and (3) by calculation from the values obtained from gas analysis.

The principle of *continuous recording gas calorimeters* is combustion of a controlled stream of gas, followed by heat exchange between the combustion gases and a liquid, whose temperature increase is measured and recorded [133]. This method is the most widely used. *Discontinuous gas calorimeters* involve combustion of a specific amount of gas and the resulting temperature increase is measured directly.

6.2.3. Molecular Mass

The molecular mass of natural gas is calculated as an average value from the concentrations determined by gas analysis and the molecular masses of the components (see Chap. 1).

6.2.4. Dew Point

In order to avoid condensation of water, the gas temperature should not drop below the dew point of the water contained in the gas. The dew point can be determined directly by passing a constant stream of gas over a polished metal mirror that is slowly cooled. The dew point is the temperature at which the first signs of condensation appear on the mirror. These methods can be used for gases under pressures of up to 10 MPa. Automated instruments using photocells are also employed [134]. If the gas stream contains easily condensable hydrocarbons, then these must be washed out by passing the gas through high-boiling liquid paraffin prior to determination.

7. Acknowledgement

The entire topic was coordinated by Axel Commichau.

8. References

General References

1. H. B. Bradley: *Petroleum Engineering Handbook*, Society of Petroleum Engineers, 1987, Chap. 20.
2. D. L. Katz et al.: *Handbook of Natural Gas Engineering*, McGraw-Hill, New York 1959.
3. G. G. Brown et al.: *Natural Gasoline and the Volatile Hydrocarbons*, Natural Gas Association of America, Tulsa 1948.

Specific References

4. *DIN-Norm 1871:* "Gasförmige Brennstoffe und sonstige Gase," Beuth Verlag, Berlin – Köln 1980.
5. *VDI-Wärmeatlas*, VDI-Verlag, Düsseldorf 1988.
6. N. L. Carr, R. Kobayashi, D. B. Burrows: "Viscosity of Hydrocarbon Gases Under Pressure," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **201** (1954) 264 –278.
7. M. B. Standing, D. L. Katz: "Density of Natural Gases," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **146** (1942) 140 – 144.
8. J. D. van der Waals, *Proc. Acad. Sci. Amsterdam* **3** (1901) 515.
9. O. Redlich, J. N. S. Kwong: "On the Thermodynamics of Solutions. V—An Equation of State. Fugacities of Gaseous Solutions," *Chem. Rev.* **44** (1949) 233 – 244.
10. G. Soave: "Equilibrium Constants from a Modified Redlich – Kwong Equation of State," *Chem. Eng. Sci.* **27** (1972) 1197 – 1203.
11. O. Y. Peng, D. B. Robinson: "A New Two-Constant Equation of State," *Ind. Eng. Chem. Fundam.* **15** (1976) no. 1, 59 – 64.
12. R. L. Robinson, Jr., R. H. Jacoby: "Better Compressibility Factors," *Hydrocarbon Process.* **44** (1965) no. 4, 141 – 145.
13. K. R. Hall, L. Yarborough: "A New Equation of State for Z-Factor Calculations," *Oil Gas J.* **71** (1971) June 18, 82 – 91.
14. M. Benedict, G. B. Webb, L. C. Rubin: "An Empirical Equation for Thermodynamic Properties of Light Hydrocarbons and Their Mixtures," *Chem. Eng. Prog.* **47** (1951) 419.
15. K. E. Starling, M. S. Hon: "Thermo Data Refined for LPG," *Hydrocarbon Process.* **52** (1972) no. 5, 129 – 133.
16. M. Jaeschke et al.: "Accurate Prediction of Compressibility Factors by the GERG Virial Equation," *SPE Paper 17 766, SPE Gas Techn. Symp.*, Dallas, TX, 1988.
17. ISO 6570: "Natural Gas—Determination of Potential Hydrocarbon Liquid Content," 1983/1984.
18. D. L. Katz: "Prediction of Conditions for Hydrate Formation in Natural Gases," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **160** (1945) 140.
19. DIN 51 850: "Brennwerte und Heizwerte gasförmiger Brennstoffe," Beuth Verlag, Berlin-Köln 1980.
20. Deutscher Verein des Gas- und Wasserfaches: Techn. Regeln "Gasbeschaffenheit," Arbeitsblatt G 260/I, Eschborn 1983.
21. D. L. Katz: *Handbook of Natural Gas Engineering*, McGraw-Hill, New York 1959.
22. G. Drewniok, D. Huber: "Neue Ergebnisse bei der Erdgasbehandlung im Feld," *Lecture at the 26th DGMK-Haupttagung*, DGMK-Compendium 78/79, Berlin, Oct. 4 – 6, 1978.
23. J. H. z. Siederdissen, F. Wundram, *GWF Gas Wasserfach: Gas Erdgas* **127** (1985) no. 1, 3 – 11.
24. F. Brunner, W. Woll, *JPT J. Petr. Technol.* **10** (1980) 377 – 384.
25. O. Buchholz, S. Peter, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) 268 – 272.
26. G. Wilken, *SPE Paper 19 813, 64th SPE Technical Conference*, San Antonio, Oct. 8 – 11, 1989.
27. W. Tunn, *Erdöl Kohle Erdgas Petrochem.* **26** (1973) no. 9, 498 – 500.
28. Dow Chemical Comp.: *Gas Conditioning Fact Book*, Midland, Michigan 1972.
29. NGPSA (Nat. Gas Proc. Suppl. Ass.): *Engineering Data Book*, Tulsa, Oklahoma, 1972.
30. A. Misra, *Erdöl Erdgas Z.* **91** (1975) Apr., 99 – 105.
31. D. Ballard, *Hydrocarbon Process.* **45** (1966) June, 171 – 180.
32. J. Kaliner et al.: "High Pressure Separation of Higher Hydrocarbons From Natural Gas With a Continuous Regeneration of the Absorbent By Natural Gas," *17th World Gas Conference*, Washington, D.C., June 5 – 9, 1988.
33. J. McAllister, P. Parsons, *The Chemical Engineer* 1987, April, 24 – 26.
34. J. Voogd, *GWF Gas Wasserfach: Gas Erdgas* **128** (1987) 4: 192 – 200.

35. A. L. Kohl, F. C. Riesenfeld: *Gas Purification*, Gulf Publ., Houston 1979.
36. K. Thormann: *Absorption*, Springer Verlag, Berlin 1959.
37. C. W. Sweeney et al., *Chem. Eng. (N.Y.)* **95** (1988) no. 9, 119 – 125.
38. J. Polasek, J. Bullin, *Energ. Prog.* **4** (1984) no. 3, 146 – 150.
39. J. A. Lagas, *Sulphur* **159** (1982) Mar./Apr., 35 – 39.
40. J. A. Lagas: "Selection of Gas Sweetening Processes," *Comprimo Gas Sweetening and Sulphur Recovery Seminar*, Amsterdam, Nov. 1982.
41. *1989 Gas Process Handbook* issued by Hydrocarbon Process. Gulf Publishing Co, Houston, 1989.
42. R. E. Meissner, U. Wagner, *Oil Gas J.* **81** (1983) Feb. 7, 55 – 58.
43. J. C. Thomas: "Improved Selectivity Achieved With Ucarsol Innovator Solvent 111," *38th Annual Laurance Reid Conference*, Norman, Oklahoma, March 7, 1988.
44. W. Weinberg et al.: "New Gas Treating Alternatives For Saving Energy in Refining and Natural Gas Treating," *11th World Petroleum Congress*, London, Aug. 31, 1983.
45. S. A. Newman: "The Benfield Process For Acid Gas Removal," *AIChE Symposium Acid And Sour Gas Treating*, Atlanta, Mar. 12, 1984.
46. J. Frazier, *Hydrocarbon Process.* **49** (1970) no. 4, 101 – 102.
47. W. G. J. van Dijk, J. G. Tönis, *Erdöl Kohle Erdgas Petrochem.* **19** (1966) June, 404 – 405.
48. M. Kriebel, *Energ. Prog.* **4** (1984) no. 3, 143 – 145.
49. S. A. Newman: *Sour Gas Treating Processes*, Gulf Publ., Houston 1985.
50. "Sweeting Natural Gas II," *Sulphur* **193** (1987) Nov./Dec., 26 – 31.
51. K. Hedden, R. Rao, *GWF Gas Wasserfach: Gas Erdgas* **123** (1982) no. 1, 1 – 6.
52. P. J. Parsons et al.: "Improvement in Adsorption Processes For Testing Natural Gas," *51st Autumn Meeting Institute of Gas Engineers*, London, Nov. 13, 1985.
53. K. D. Henning et al., *GWF Gas Wasserfach: Gas Erdgas* **126** (1985) no. 1, 19 – 24.
54. M. R. Cines et al., *Chem. Eng. Prog.* **72** (1976) Aug., 89 – 93.
55. R. W. Spillman et al., "Gas Membrane Process Optimization," *AIChE Meeting*, New Orleans, March 9, 1988.
56. E. W. Funk,, N. N. Li: "Purification of natural gas by membranes," *Gas Separation Technology*, Elsevier Sci. Publ., Amsterdam, 1990, 355 - 372.
57. K.. Ohlrogge: "Membranes for Dewpointing of Gas Streams," *13th Ann. Membr. Techn./Sep. Plann. Conf.*, Newton, Massachusetts, Oct. 23 - 25, 1995.
58. U. Sander et al.: *Sulphur, Sulphur Dioxide and Sulphuric Acid*, The British Sulphur Corp. Ltd., London 1984.
59. *Sulphur* **187** (1986) Nov./Dec., Suppl. 1 – 20.
60. J. B. Hyne, *Sulphur* **198** (1988) Sept./Oct., 24 – 39.
61. M. J. Pearson, *Hydrocarbon Process.* **94** (1973) Feb., 81 – 85.
62. J. Nougayrede et al., *Oil Gas J.* **85** (1987), Aug. 10, 65 – 70.
63. Energy Resources Conservation Board (ERCB): *Sulphur Recovery Guidelines for Sour Gas Plants in Alberta*, Aug. 1988.
64. Environmental Protection Agency 40 CFR Part 60, *United States Federal Register,* vol. 50, no. 190, Part IV, Oct. 1985.
65. *VDI-Guideline 3454:* "Emission Control Claus Units," Part 1, Beuth-Verlag, Berlin, June 1989.
66. J. E. Meuler: "Claus Tailgas Treatment," Part 1, *Comprimo Gas Sweetening and Sulphur Recovery Seminar,* Amsterdam, Nov. 1985.
67. B. G. Goar: "Emerging New Sulfur Recovery Technologies," *Laurence Reid Gas Conditioning Conference*, Norman, Oklahoma, wMarch 6 – 8, 1989.
68. T. J. van Pol: "Claus Tailgas Treatment," Part 2, *Comprimo Gas Sweetening and Sulphur Recovery Seminar,* Amsterdam, Nov. 1985.
69. R. Kettner, N. Liermann, *Oil Gas J.* **86** (1988) Jan. 11, 63 – 66.
70. R. Kettner, T. Lübcke, *VDI Ber.* **730** (1989) 255 – 274.
71. R. Müller, *Erdöl Ergas Kohle Petrochem.* **28** (1975) 132 – 137.
72. R. Kettner, N. Liermann, *Erdöl Erdgas* **102** (1986), Oct., 448 – 453.
73. S. Lechler, A. Misra, A. Feizlmayr: *New Technology Required for West German Sulfur Line*, Pipe Line Industry, June 1978.
74. T. Lübcke et al.: "The Mobil Alkanolamine Waste Recovery Process," *Comprimo Sulphur Workshop*, Amsterdam, Nov. 1987.
75. R. N. Tuttle, R. D. Kane, *National Association of Corrosion Engineers Publication*, Houston 1981.
76. W. J. Schell, C. D. Houston, *Chem. Eng. Prog.* **78** (1982) Oct. 33 - 37.

77. S. S. Kulkarni, E. W. Funk, N. N. Li, R. L. Riley, *AIChE Symp. Ser. No.229*, **79** (1983) 172 - 178.
78. T. E. Sulpizio, B. S. Minhas, H. S. Meyer, *AIChE National Meeting*, New Orleans, Louisiana, April 2, 1992.
79. L. S. White, T. A. Blinka, H. A. Kloczewski, I. Fan Wang, *J. Membr. Sci.* **103** (1995) Oct., 73 - 82.
80. W. J. Koros, G. K. Fleming, *J. Membr. Sci.* **83** (1993) 1 - 80.
81. T. Shiratori, H. Sonta: "Application of Iron-Oxidizing Bacteria to Hydrometallurgical Flue Dust Treatment and Hydrogen Sulfide Desulfurization," *FESM Micro Biol. Rev.* **11 (1 - 3)** 1993, 165 - 174.
82. C. J. N. Buisman, W. L. Prins: "New Process for Biological (Flue) Gas Desulphurisation," *Biologische Abgasreinigung, VDI-Bericht 1104*, VDI-Verlag Düsseldorf, 1994.
83. H. Satoh, J. Yoshizava, S. Kametani: "Bacteria Help Desulfurise Gas," *Hydrocarbon Process* **67** (1988) no. 5, 76D - 76F.

Specific References

84. J. L. Birnbaum: "The Reemergence of LNG," vols. 1 and 2, Energy Research Assoc., New York 1989.
85. "LNG World Overview," Gotaas-Larsen Shipping Corp., New York, updated annually.
86. R. N. Davis, Y.-N. Liu: "Baseload LNG: An International Energy Source," *1987 Spring National Meeting of the American Institute of Chemical Engineers*, Houston, USA, Mar. 29 – Apr. 2, 1987.

Specific References

87. R. N. Davis, L. E. Olson: "The Future of Natural Gas as a Fuel in Railroad Operations," *16th Energy Technology Conference (ET'89)*, Washington, D.C., USA, Feb. 28 – Mar. 2, 1989.
88. D. Rooke: "LNG Industry—A Retrospective," *9th International Conference on Liquefied Natural Gas (LNG-9)*, Nice, France, Oct. 17 – 20, 1989.
89. L. S. Gaumer, C. L. Newton, US 3 763 658, 1973.
90. M. J. Roberts, R. Agrawal, US Patent 6,308,531 B1, 2001.
91. M. J. Roberts, Y. Liu, J. C. Bronfenbrenner, J. M. Petrowski, "Reducing LNG Capital Cost in Today's Competitive Environment", PS2-6, The 14th International Conference and Exhibition on Liquefied Natural Gas (LNG14), Doha, Qatar, March 21-24, 2004.
92. Y.-N. Liu, J. W. Pervier, US 4 545 795, 1985.
93. I. Aoki, Y. Kitsukawa, US 3 817 046, 1974.
94. N. J. Cuperus: "Developments in Cryogenic Storage Tanks," *6th International Conference on Liquefied Natural Gas (LNG-6)*, Kyoto, Japan, Apr. 7 – 10, 1980.
95. A. C. Timmers, Y. Lev, C. R. McClune, M. J. Pikaar: "Development of Fire Prevention Measures for Large LNG Storage Installations," *6th International Conference on Liquefied Natural Gas (LNG-6)*, Kyoto, Japan, Apr. 7 – 10, 1980.
96. D. R. Jenkins, A. C. Timmers: "LNG Plant Safety in the Light of Recent Experimental Trials and Theories," *7th International Conference on Liquefied Natural Gas (LNG-7)*, Jakarta, Indonesia, May 15 – 19, 1983.
97. J. A. Eyre, M. J. Pikaar, J. L. J. Rosenfield: "The Assessment of Combustion Related Hazards Associated with the Spillage of LNG," *8th International Conference on Liquefied Natural Gas (LNG-8)*, Los Angeles, USA, June 15 – 19, 1986.
98. M. C. Parnarouskis, C. D. Lind, P. P. K. Raj, J. M. Cece: "Vapor Cloud Explosion Study," *6th International Conference on Liquefield Natural Gas (LNG-6)*, Kyoto, Japan, Apr. 7 – 10, 1980.
99. F. Tuppeck: "Perspektiven für den interkontinentalen Transport von Energieträgern—dargestellt am Beispiel des Erdgastransportes," *Erdöl Erdgas Z.* **95** (1979) no. 6.
100. F. Tuppeck: "Die energiewirtschaftliche Bedeutung natürlicher und wirtschaftlicher Einflußgrößen für die Ferngaswirtschaft," (Dissertation, TH Karlsruhe), *Gesammelte Ber. Betr. Forsch. Ruhrgas AG* Essen 1959.
101. K. Liesen: *Der internationale Erdgashandel und die Bedeutung des Erdgases für ein ressourcenarmes Land*, Jahrbuch für Bergbau, Energie, Mineralöl und Chemie 83/83, Verlag Glückauf, Essen 1982.
102. K. Schwier: "Technik der LNG-Kette am Beispiel der deutschen Flüssigerdgas-Importverträge," *GWF Gas-Wasserfach: Gas/Erdgas* **121** (1980) no. 10.
103. H. G. Graf (ed.): "Inline-Kompressorstation in Ferngasleitungen DVGI-Symposion," *Erdöl Erdgas Z.* **98** (1982) no. 2.
104. F. Herning: *Stoffströme in Rohrleitungen*, 4th ed., VDI-Verlag, Düsseldorf 1966.
105. R. Eberhard, R. Hüning: *Handbuch der Gasversorgungstechnik*, R. Oldenbourg Verlag, München – Wien 1984.

106. H. Oermann: "Der Fernleitungstransport von Erdgas in technischer und wirtschaftlicher Sicht," *GWF Gas- Wasserfach: Gas/Erdgas* **112** (1971) no. 1.
107. H. Lübbeck, G.-J. Gralla: "Kriterien und Möglichkeiten für die Speicherung von Gas in deutschen Erdöl- und Erdgaslagerstätten," *Erdöl Erdgas Z.* **91** (1975) no. 11, 381 – 387.
108. W. F. Burke: "Simultaneous Underground Gas Storage and Secondary Oil Recovery," *J. Technol.* (1960) no. 22/26.
109. E. Schindewolf: "Untertagespeicherung – Bindeglied zwischen Erdgasbezug und Erdgasmarkt," *GWF Gas- Wasserfach: Gas/Erdgas* **121** (1980) no. 10, 455.
110. H. Cross: *Analysis of Flow in Networks of Conduits or Conductors*, Engineering Experiment Station Bulletin no. 286, University of Illinois, Nov. 1936.
M. S. McIllroy: "Pipeline Network Flow Analysis Using Ordinary Algebra," *J. Water Works Assoc.* **41** (1949) no. 5, 422 – 428.
111. F. Tuppeck, H. Schachelhuber: "Erfahrungen bei der Berechnung von Ortsverteilungsnetzen," *Gasverwendung* **18** (1967) no. 8, 331 – 337.
112. F. Tuppeck: "Optimaler Aufbau der Transportkapazität von Ortsgasnetzen mit Hilfe systematischer Rohrnetzberechnungen," *GWF Gas- Wasserfach: Gas/Erdgas* **105** (1964) no. 19, 505 – 509.
113. CEDIGAZ (Centre international d'information sur le gaz naturel, Paris) Le gaz naturel dans le monde en 1995, July 1996.
114. BGR (Bundesanstalt für Geowissenschaften und Rohstoffe), *Reserven, Ressourcen und Verfügbarkeit von Energierohstoffen*, Hannover 1995, p. 163.
115. A. W. Drews (ed.): *ASTM Manual on Hydrocarbon Analysis, ASTM Manual Series MNL 3*, 5th ed., Philadelphia, PA, 1992.
116. DVGW-Regelwerk, Arbeitsblatt G 261: Prüfung der Gasbeschaffenheit, ZfGW-Verlag, Frankfurt 1985.
117. K. H. Altgelt, T. H. Gouw: *Chromatography in Petroleum Analysis*, Marcel Dekker, New York 1979.
118. S. H. Kägler: *Neue Mineralölanalyse*, Hüthig Verlag, Heidelberg 1987.
119. T. A. Norris: "Gas Analysis: Analysis of C5 and Lighter Hydrocarbons," in [115], pp. 1 – 5.
120. ASTM D 1945 - 91 in [115], pp. 326 – 337.
121. C. J. Cowper, A. J. De Rose: *The Analysis of Gases by Chromatography*, Pergamon Press, Oxford 1984.
122. DIN 51 405, ed. May 1987.
123. E. Leibnitz, H. J. Struppe: *Handbuch der Gaschromatographie*, 3rd ed., Verlag Chemie, Weinheim 1984.
124. H. Laurien: *Taschenbuch Erdgas*, R. Oldenburg Verlag, München – Wien 1970.
125. DIN 51 872 Part 3, ed. Feb. 1987.
126. ASTM D 2504–83 in [115], pp. 466 – 469.
127. DVGW-Regelwerk, Arbeitsblatt G 285: Hydratinhibierung in Erdgasen mit Methanol, ZfGW-Verlag, Frankfurt 1974.
128. ASTM E 700–79,
DIN 51 777, ed. March 1983.
129. M. A. Berliner: *Feuchtemessung*, VEB Verlag Technik, Berlin 1980.
130. ASTM 2385–81 in [115] , pp. 428 – 431.
131. ASTM D 1070–85.
132. ASTM D 3588–81,
DIN 51 858, ed. Nov. 1982,
ISO 6976–1983.
133. ASTM D 1826–88.
134. ASTM D 1142–90,
ISO 6327–1981,
DIN 51 871, ed. Dec. 1985.

Nematicides

Jürgen Hartwig, Bayer AG, Leverkusen, Germany

Herbert Sommer, Bayer AG, Leverkusen, Germany

Franz Müller, (formerly Novartis Crop Protection, Basel), Allschwil, Switzerland

1. Introduction 1
2. Plant-Parasitic Nematodes 1
2.1. Characteristics of Plant-Parasitic Nematodes 1
2.2. Plant Damage 2
3. Nematode Management 3
3.1. Nonchemical Measures 3
3.2. Chemical Measures 3
4. Individual Nematicides 4
4.1. Fumigants and Other Agents 4
4.2. Carbamates 6
4.3. Organophosphorus Esters 7
4.4. Other Nematicides 8
5. Toxicology 9
5.1. Fumigants and Other Agents 9
5.2. Carbamates 10
5.3. Organophosphorus Esters 11
5.4. Other Nematicides 12
6. References 12

1. Introduction

Nematicides (also called nematocides) are plant protection agents (i.e., pesticides) used to control plant-parasitic nematodes. The growing use of intensive cropping and production methods in agriculture means that plant-parasitic nematodes are becoming an increasingly important yield-limiting factor in many crops. Nematicides are an indispensable part of nematode management in modern agriculture.

This article describes plant-parasitic nematodes and their control, with particular regard to nematicides. For agents used against human-parasitic and animal-parasitic nematodes, see → Anthelmintics, Vetinary Drugs.

2. Plant-Parasitic Nematodes

Nematodes are small, wormlike, unsegmented organisms. On account of their body shape and undulatory movements, nematodes are also termed eelworms.

2.1. Characteristics of Plant-Parasitic Nematodes

Classification. Nematodes belong to the phylum Nematoda. The majority of the currently described nematode species (ca. 15 000) live in freshwater or marine habitats, or as bacterial or fungal feeders in the soil. Numerous species attack humans and animals; about 4300 species are plant parasites.

Morphology and Anatomy. Plant-parasitic nematodes are small, transparent, cylindrical organisms (length ca. 300 – 4000 µm, width 15 – 35 µm). All plant-parasitic nematodes have a stylet, generally in the form of a hollow needle (exception: *Trichodorus* sp.), which is used to puncture the plant cells, inject saliva, and suck out part of the cell contents. Nematodes only have longitudinal muscles and can therefore only execute an undulatory movement. They do not have a vascular system or special respiratory organs.

Life Cycle. The life cycle of all plant-parasitic nematodes is essentially the same. After development of the embryo, the first molt generally occurs in the egg. The hatched larvae are mobile and search for a host plant. Nematodes pass through four larval stages before developing into the adult form. Both sexual and parthenogenetic reproduction occur. The reproduction rate is species dependent, but also depends on environmental factors (e.g., soil, temperature, host plant, competition between and among species). A root-knot nematode female can, for example, produce a total of 500 eggs, several generations can develop per year. Other

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_125

nematode species such as *Globodera rostochiensis* produce only one generation per year. Many nematode species have a dormant phase, which enables them to survive for very long periods without a host plant. Under favorable conditions, some cyst-forming nematodes can remain viable for up to ten years in the absence of suitable host plants.

Mode of Parasitism. The parasitic behavior of plant nematodes varies widely. For example, they can damage plant parts that grow aboveground, (e.g., flowering parts by *Anguina tritici*, leaves and buds by *Aphelenchoides* sp., as well as stalks and leaves by *Ditylenchus* sp.).

Of greater importance, however, are nematodes that damage plant parts below ground. Nonsedentary root nematodes migrate from one feeding site to the next. Ectoparasites normally pierce the roots from outside, whereas endoparasitic root nematodes can enter the root tissue and feed from within.

Sedentary root nematodes generally penetrate the root tissue. They induce a feeding cell system, on which they depend until they are sexually mature and discontinue feeding.

Population Density and Distribution. The soil nematode population constitutes a significant proportion of the mesofauna (animals of medium size), particularly in tropical and subtropical regions. The nematode population can account for up to 5 % of the biomass of the soil (i.e., 100×10^6 nematodes per square meter), but often only a very small proportion is plant parasitic.

It is not possible to give any generally valid information on damage thresholds because they vary according to the nematode species, host plant, temperature, and many local parameters. For example, although only 20 *Meloidogyne arenaria* larvae per 100 mL of soil may produce severe yield losses in tomatoes, under certain conditions about 300 *Xiphinema* sp. per 100 mL of soil do not cause visible damage.

Since nematodes can only migrate over very small distances, active propagation is limited. Passive propagation (i.e., dispersion by harvest residues, soil cultivation, wind, contaminated machinery, water, seeds, and harvested crops) plays an important role in establishing a new infection.

2.2. Plant Damage

Physiological Basis. The primary sucking damage caused by nematodes and the associated removal of the cell contents are of comparatively minor importance. Damage is primarily due to the subsequent plant reactions: to the secretion of saliva such as root galling (swelling), breakdown of root tissue, and increased formation of lateral roots. Altered root growth as a result of nematode infestation leads in turn to reduced water and nutrient uptake. Young plants react particularly sensitively to nematode infestation. The mechanical tissue injury caused by some migrating nematodes as well as the commonly occurring interactions with fungal and bacterial pathogens can also lead to considerable yield losses. Some nematode species (e.g., *Xiphinema*, *Longidorus*, *Paralongidorus*, *Trichodorus*, *Paratrichodorus*) also act as vectors for viruses responsible for yield losses and cause damage as a result of feeding.

Symptoms. Plant-parasitic nematodes only rarely produce characteristic symptoms on above-ground plant parts; typical symptoms can be more readily recognized on and in the roots:

root galls (*Meloidogyne*)
cysts (*Heterodera, Globodera*)
lesions (*Pratylenchus, Radopholus*)
stubby root (*Trichodorus*)

For an accurate diagnosis, special extraction and detection methods are often necessary [16], [17, pp. 13 – 46] .

Damage to Cultivated Plants and Crops. It is difficult to estimate the crop losses caused by nematodes; results differ considerably. Worldwide annual losses due to nematodes were estimated at 12 % in 1987, corresponding to a value of $\$ 77 \times 10^9$ (averaged over all important cultivated plants) [18]. The necessity for nematode control can be assessed comparatively accurately on the basis of estimates of nematicide use: the overall nematicide market, including fumigants, was ca. $\$ 485 \times 10^6$ in 1989 [19] – only about 2 % of the total agrochemical market. Important areas of use of nematicides are the protection of vegetables, potatoes, tobacco, bananas, berries, soft fruits, ornamental plants, peanuts, beet, citrus fruits, and cotton.

The gall-forming, sedentary *Meloidogyne* species are by far the most important nematodes in terms of their effects on crop losses, especially in warmer climates. These roughly 70 different species can infest all important crops; they produce root galls that affect the growth of the root system.

The cyst-forming, sedentary nematode genera *Globodera* and *Heterodera* are primarily of importance in temperate climatic zones, and in some species have a narrow host spectrum. The females are transformed into tough, egg-filled cysts. The eggs can survive in the soil for a very long time. Important examples are *Heterodera avenae* (e.g., cereals), *H. glycines* (e.g., soybean), *H. schachtii* (e.g., sugarbeet), *H. tabacum* (e.g., tobacco), *H. trifolii* (e.g., clover), *Globodera rostochiensis*, and *G. pallida* (e.g., potatoes, tomatoes).

The "citrus nematode" (*Tylenchulus semipenetrans* in citrus fruits), the "lesion nematode" (*Pratylenchus* in cereals, vegetables, fruit, ornamental plants), and the "burrowing nematode" (*Radopholus* in bananas) are further examples of important endoparasitic root nematodes.

"Stem and bulb nematodes" (*Ditylenchus*) cause delayed plant germination. In cases of serious infestation they also cause the death of young plants and lesions on the leaves and stems of onions, strawberries, and many types of ornamental plants.

"Leaf nematodes" (*Aphelenchoides* sp.) and "seed – gall nematodes" (*Anguina* sp.) are of only minor importance.

3. Nematode Management

The object of nematode management is to restrict the spread of plant-parasitic nematodes or keep their population below the threshold for economically significant damage. Control strategies include both chemical and nonchemical measures.

3.1. Nonchemical Measures

Quarantine measures and hygiene are the only preventive measures of any great importance because nematodes are mainly spread and transmitted by passive transport. In modern agriculture with international exchange of plant material and harvested crops, such measures may be insufficient.

Cultivation measures include the maintenance of fallow land or intensified cultivation of nematode nonhost plants or plants with nematicidal properties in crop rotation. These measures are only effective for controlling nematodes with a narrow host spectrum, they are seldom economical under intensive cultivation conditions.

Physical measures are successful in some cases: flooding of arable land; heat sterilization of soils by steam, ultrasound, or "solarization" in which polyethylene tarps are used to trap solar energy; and hot-water treatment of seeds and plant material. On account of their high costs, these measures are used only for a few high-value crops.

Varieties of crops such as potatoes, tomatoes, tobacco, and soybeans exist that are considered to be resistant to certain nematode species. Although the development of *resistant varieties* involves a great deal of time and effort, it is one of the most cost-effective control measures for the user. However, the frequent use of resistant varieties entails the long-term danger of loss of resistance due to nematode pathotypes or the rapid proliferation of nematode species that are not affected by the resistant plant.

A large number of *nematode antagonists* are known (fungi, bacteria, viruses, and protozoa). Their widescale use is, however, limited by a narrow host spectrum, problematic mass production, lack of storage-stable formulations, high application rates, and unpredictable action.

3.2. Chemical Measures

Since the nonchemical measures for nematode control are currently only of limited use, chemical control is unavoidable in many cultivated plants and crops.

Although chemical nematode control had already been used to a limited extent at the end of the nineteenth century (e.g., use of carbon disulfide), it was only from about the mid-1940s that it achieved greater importance with the use of chloropicrin, a mixture of 1,2-dichloropropane and 1,3-dichloropropene (D-D), 1,2-dibromo-3-

chloropropane (DBCP), and 1,2-dibromoethane (EDB).

The nematicides used nowadays can be subdivided into four groups:

1) Halogenated aliphatic hydrocarbons
2) Compounds related to methylisothiocyanates
3) Organophosphates
4) Carbamates

Groups 1 and 2 constitute the first generation of nematicides. On account of their low boiling points they are classified as *fumigants*. Compounds such as methyl bromide, chloropicrin, and methyl isothiocyanate also have a pronounced soil-insecticidal, fungicidal, or herbicidal action in addition to their nematicidal action. Fumigants are usually injected as a gas or liquid into the soil (required amount 200 – 1000 L/ha). On account of their high vapor pressure, the substances are distributed in the top soil layer and kill both free-living nematode stages as well as the eggs. Nematodes in deeper soil layers are not completely killed, and the nematode population may recover. Plant compatibility of these substances is often low and should be borne in mind. Use on established crops is not possible in many cases. Waiting periods of 10 – 14 d are generally necessary. Fumigants have a simple chemical structure and are therefore inexpensive. Despite the high doses required and the technically complicated mode of application, economical use is often possible. In quantitative terms fumigants account for more than 50 % of nematode control worldwide.

Organophosphates and carbamates have been used for nematode control since the early 1960s (second-generation nematicides). They are employed as granulate or liquid formulations in amounts of ca. 2 – 10 kg/ha. The active ingredients are generally water-soluble, and can thus be distributed in the soil capillary system. They often have supplementary insecticidal action. They are well tolerated by plants and can thus be used on seeds, seedlings, and established crops. These nematicides are applied as a band, in furrows, or as a broadcast application; incorporation of the products into the upper layer of the soil promotes their distribution. With some crops, application in a low-pressure irrigation system (e.g., drip irrigation) or treatment of nursery seedlings or bulbs may also be employed. Nematodes are often not killed, but only paralyzed or disoriented depending on the concentration of the active substance in the soil. The effect is ascribed mainly to cholinesterase inhibition which, in turn, interferes with nervous transmission.

Although fumigants are currently still extremely important worldwide in chemical nematode control, their use will probably decrease. Possible reasons for this include the very high amounts applied, the relatively expensive and sophisticated application techniques, translocation into the groundwater (e.g., DBCP, EDB, D-D), and toxicological considerations (e.g., for DBCP: male sterility, carcinogenesis; EDB is banned in the USA). The carbamates and phosphorus esters also exhibit disadvantages: high acute mammalian toxicity, in some cases high avian toxicity, and translocation into the groundwater are becoming increasingly significant.

Third-generation nematicides are subject to stringent requirements, which means that their development is becoming increasingly difficult and expensive. The wide usage requires selective agents that act solely against nematicides. New groups of active substances, possibly with new modes of action, should exhibit a favorable toxicological profile and should also have advantages as regards application and environmental compatibility. For example, use of nematicides as seed treatments would reduce the required dose; after foliar application, it is highly desirable that the nematicide is transported basipetally to the roots (reduction of leaching).

4. Individual Nematicides

In the following only nematicides still used frequently are given, others, for reasons stated above, are only mentioned briefly.

4.1. Fumigants and Other Agents

Carbon disulfide [*75-15-0*], CS_2, has been used as a fumigant against insects and nematodes.

Chloropicrin [*76-06-2*], trichloronitromethane, CCl_3NO_2, M_r 164.4, *mp* − 64 °C, *bp* 112.4 °C,

vapor pressure 3.2 kPa (25 °C), is a colorless liquid with lachrymatory action which is fairly soluble in water, and miscible with most organic solvents [20, p. 226].

Cl_3CNO_2

Chloropicrin is formed by chlorinating nitromethane or picric acid with alkali metal or alkaline earth hypochlorites or by treating acetone with concentrated nitric and hydrochloric acids [21].

Chloropicrin has insecticidal and nematicidal properties. It is used as a soil fumigant to control nematodes and other soil-dwelling pests. It is highly toxic to both plants and animals. It has an extremely irritating effect on the mucous membranes and has been used as tear gas.
Trade Names. Chlor-O-Pic (Great Lakes); Dorochlor (Mitsui Toatsu).

Dazomet [*533-74-4*], 3,5-dimethyl-1,3,5-thiadiazinane-2-thione, $C_5H_{10}N_2S_2$, M_r 162.3, *mp* 104 – 105 °C, vapor pressure 0.37 mPa (20 °C), forms colorless crystals which are fairly soluble in water and diethyl ether, and readily soluble in cyclohexane, chloroform, acetone, benzene, and ethanol [20, p. 335].

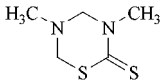

Dazomet is produced by reaction of methylammonium *N*-methyldithiocarbamate with formaldehyde [22]. Dazomet is sensitive to moisture and heat (> 35 °C); acid hydrolysis yields carbon disulfide. In the ground Dazomet decomposes into methyl(methylaminomethyl)dithiocarbamic acid, followed by methyl isothiocyanate, which is the compound responsible for the fungicidal, bactericidal, and nematicidal action [23].

Dazomet is used as a soil sterilant to control nematodes, and as a slimicide in pulp and paper manufacture, and as a preservative in adhesives and glues. Dazomet is also used as a herbicide, fungicide, and insecticide.
Trade Names. Basamid (BASF), Dazoberg (Diachem).

1,2-Dibromo-3-chloropropane [*96-12-8*] was used as a nematicide before sowing or planting.

Ethylene dibromide [*106-93-4*], 1,2-dibromoethane, $C_2H_4Br_2$, M_r 187.9, *mp* 9.3 °C, *bp* 131.5 °C, vapor pressure 1.5 kPa (25 °C), is a colorless liquid which is fairly soluble in water, soluble in diethyl ether, ethanol, and most common organic solvents [20, p. 490].

$BrCH_2CH_2Br$

Ethylene dibromide is produced by the uncatalyzed addition of bromine to ethylene (→ Bromine Compounds, Chap. 1.4.1.)[24]. Ethylene dibromide is used as a fumigant to control nematodes, wireworms, and other soil pests, and in fumigation of mills, warehouses, and households. It is also used as an insecticide.
Trade Name. Dibrome (United Phosphorus).

1,3-Dichloropropene [*542-75-6*], $C_3H_4Cl_2$, M_r 111.0, *bp* 104.3 °C (*Z* isomer), 112 °C (*E* isomer), 108 °C (isomer mixture), is a colorless-to-amber liquid which is fairly soluble in water, miscible with hydrocarbons, halogenated solvents, esters, and ketones [20, p. 366].

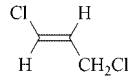

E isomer

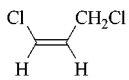

Z isomer

1,3-Dichloropropene is obtained as a byproduct in the chlorination of propene (→ Allyl Compounds, Chap. 1.3. [25].

1,3-Dichloropropene is marketed as a mixture containing the inactive 1,2-dichloropropane, or as the pure active substance (mixture of *E* and *Z* isomers). Dichloropropene is used purely as a soil fumigant nematicide (the *Z* isomer is more effective than the *E* isomer) and is used pre-planting to control most species of nematode in deciduous fruit and nuts, citrus fruit, berry fruit, vines, strawberries, and many other crops.
Trade Names. D-D (Cyanamid, Hokko); Telone (DowElanco); Nematox (Siapa).

Metam [*144-54-7*], methyldithiocarbamic acid, $C_2H_5NS_2$, M_r 107.2 and *metam sodium* [*137-42-8*], sodium methyldithiocarbamate, $C_2H_4NNaS_2$, M_r 129.2, sodium metam forms a colorless crystalline dihydrate which decomposes on melting. It is readily soluble

in water, and moderately soluble in acetone, ethanol, and xylene [20, p. 797].

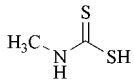

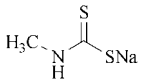

Metam sodium is produced by reaction of methylamine with carbon disulfide and sodium hydroxide [26]. The most common application form of metam sodium is a ca. 30 wt % solution. The action of metham is based on the slow formation of methyl isothiocyanate in the ground, and controls nematodes, soil fungi, weeds, weed seeds, and soil insects.
Trade Names. Vaparn (Zeneca); Arapam (Aragonesas); Trimaton (Elf Atochem); Unifume (Unicrop).

Methyl bromide [*74-83-9*], bromomethane, CH_3Br, M_r 94.9, *mp* −93 °C, *bp* 4.5 °C, vapor pressure 227 kPa (25 °C), is a colorless gas which is readily soluble in many organic solvents [20, p. 824]. Methyl bromide is manufactured from methanol and hydrogen bromide (→ Bromine Compounds, Chap. 1.4.7. [27]. Methyl bromide is a fumigant insecticide and nematicide, and as a multi-purpose fumigant used for insecticidal, acaricidal, and rodenticidal control in mills, warehouses, grain elevators, ships, and in stored products.
Trade Names. Brom-O-Gas, Met-O-Gas 100 (Great Lakes); Haltox (Degeseh).

Methyl isothiocyanate [*556-61-6*], isothiocyanatomethane, C_2H_3NS, M_r 73.1, *mp* 35–36 °C, *bp* 118–119 °C, vapor pressure 2.7 kPa (20 °C), forms colorless crystals which are fairly soluble in water and in many organic solvents [20, p. 827].

CH_3NCS

Methyl isothiocyanate is produced by reaction of methylamine with thiophosgene [28]. Methyl isothiocyanate is a multi-purpose soil fumigant used for control of nematodes, soil fungi, soil insects, and weed seeds.
Trade Name. Trapex (AgrEvo).

4.2. Carbamates

Aldicarb [*116-06-3*], 2-methyl-2-(methylthio)propanal *O*-(methylcarbamoyl) oxime, $C_7H_{14}N_2O_2S$, M_r 190.3, *mp* 98–100 °C, vapor pressure 13 mPa (25 °C), consists of colorless crystals which are fairly soluble in water, readily soluble in acetone, dichloromethane, benzene, xylene, practically insoluble in heptane, and in mineral oils [20, p. 26].

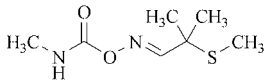

Aldicarb is prepared by treating isobutene with nitrosyl chloride, treating the resulting dimeric adduct with sodium methyl sulfide, and converting the resulting oxime to the carbamate with methyl isocyanate. **Warning:** the dimeric isobutene nitrosochloride is unstable and decomposes explosively at 94 °C. The oxime and end product also decompose vigorously above 140 and 160 °C, respectively. Aldicarb is rapidly oxidized to the corresponding sulfoxide (water solubility > 330 g/L) and slowly to the sulfone [29, 30].

Aldicarb is used as a nematicide, acaricide, and systemic insecticide; it is particularly effective against aphids, white flies (aleyrodids), and leaf minersin a variety of crops.
Trade Name. Temik (Rhône–Poulenc); Sanacarb (Sanachem).

Aldoxycarb [*1646-88-4*] has both insecticidal and nematicidal activity and has been used against a range of pests.

Carbofuran [*1563-66-2*], (2,3-dihydro-2,2-dimethylbenzofuran-7-yl)methylcarbamate, $C_{12}H_{15}NO_3$, M_r 221.3, *mp* 150–152 °C, vapor pressure 2.6 mPa (33 °C), forms colorless crystals which are slightly soluble in water but readily soluble in dichloromethane, isopropyl alcohol, and toluene [20, p. 186].

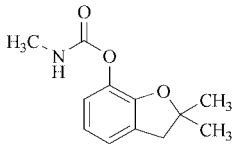

Carbofuran is prepared by the Claisen rearrangement and cyclization of 2-methallyloxy-1-nitrobenzene to 2,3-dihydro-2,2-dimethyl-7-nitrobenzofuran. Reduction of the nitro compound gives the corresponding aniline, which is diazotized and concentrated by boiling to form the phenol. Reaction with methyl isocyanate gives carbofuran [31]. Carbofuran has insecticidal and nematicidal properties and is used in a wide range of crops.
Trade Names. Furadan, Agrofuran (Sanonda); Curaterr (Bayer); Carbodan (Makhteshim-Agan); Carbosip (Sipcam); Cekufuran (Cequisa); Terafuran (Sanachem).

Cloethocarb [*51487-69-5*] was mainly used as an insecticide.

Oxamyl [*23135-22-0*], *N*,*N*-dimethyl-2-methylcarbamoyloxyimino-2-(methylthio)acetamide, $C_7H_{13}N_3O_3S$, M_r 219.3, *mp* 100–102 °C, vapor pressure 31 mPa (25 °C), consists of colorless crystals which are readily soluble in water, methanol, ethanol, acetone, and fairly soluble in toluene [20, p. 909].

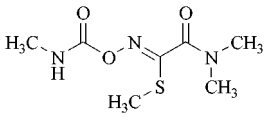

Oxamyl is prepared by chlorination of the oxime of methyl glycolate, reaction with methanethiol and alkali, followed by reaction with dimethylamine, and conversion to the carbamate with methyl isocyanate [32]. Oxamyl is used as an insecticide, acaricide, and nematicide in a wide range of crops. Used as a leaf spray, the compound is also nematicidally active on account of basipetal translocation.
Trade Name. Vydate (Du Pont).

4.3. Organophosphorus Esters

Cadusafos [*95465-99-9*], *S*,*S*-di-*sec*-butyl *O*-ethyl phosphorodithioate, $C_{10}H_{23}O_2PS_2$, M_r 270.4, *bp* 112–114 °C (0.10 kPa), vapor pressure 0.12 Pa (25 °C), is a pale yellow liquid which is moderately soluble in water, and completely miscible with acetone, acetonitrile, dichloromethane, ethyl acetate, toluene, methanol, isopropyl alcohol, and heptane [20, p. 171].

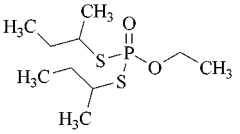

Cadusafos is prepared by treating ethyl dichlorophosphate with *sec*-butanethiol in the presence of a base and under anaerobic conditions [33]. Cadusafos is used as an insecticide and nematicide in bananas, citrus fruits, maize, potatoes, sugar cane, tobacco, and vegetables.
Trade Names. Rugby, Apache (FMC).

Diamidfos [*1754-58-1*] was uses as a nematicide.

Diazinon [*333-41-5*] is mainly used as an insecticide (see → Acaricides, Chap. 3.1.).

Dichlofenthion [*97-17-6*] was used as a nematicide and insecticide.

Ethoprophos [*13194-48-4*], *O*-ethyl *S*,*S*-dipropyl phosphorodithioate, $C_8H_{19}O_2PS_2$, M_r 242.3, *bp* 86–91 °C (25 Pa), vapor pressure 46.5 mPa (26 °C), pale yellow liquid which is moderately soluble in water, readily soluble in acetone, ethanol, xylene, dichloromethane, diethyl ether, ethyl acetate, cyclohexane, and petroleum spirit[20, p. 480].

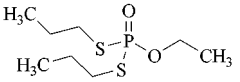

Ethoprophos is prepared by treating ethyl dichlorophosphate with propanethiol in the presence of sodium hydroxide and under anaerobic conditions [34]. Ethoprophos is a nonsystemic insecticide and nematicide used in a wide variety of crops.
Trade Name. Mocap (Rhône–Poulenc).

Fenamiphos [*22224-92-6*], ethyl 4-methylthio-*m*-tolyl isopropylphosphoramidate, $C_{13}H_{22}NO_3PS$, M_r 303.4, *mp* 49.2 °C, vapor pressure 0.133 mPa (30 °C), forms colorless crystals which are moderately soluble in water, and readily soluble in dichloromethane, isopropyl alcohol, toluene, fairly soluble in hexane [20, p. 503].

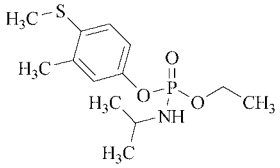

Fenamiphos is prepared by condensing 4-methylthio-*m*-cresol with *O*-ethyl *N*-isopropyl phosphoramidochloride, or with ethyl dichlorophosphate followed by reaction with isopropylamine [35]. Fenamiphos is a broad-spectrum, systemic nematicide used in a wide range of crops.
Trade Name. Nemacur (Bayer).

Fensulfothion [*115-90-2*] was used as an insecticide and nematicide.

Fosthiazate [*98886-44-3*], (*RS*)-*S*-*sec*-butyl *O*-ethyl 2-oxo-1,3-thiazoliden-3-ylphosphonothioate, $C_9H_{18}NO_3PS_2$, M_r 183,.3, *bp* 198 °C (0.5 mm Hg), light brown liquid which is fairly soluble in water, and hexane [20, p. 631].

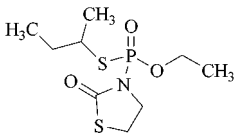

Fosthiazate is produced by treating 2-oxo-3-thiazolidine with *S*-*sec*-butyl *O*-ethyl phosphorochloridothioate in toluene [36]. Fosthiazate is used to control nematodes, aphids, thrips, etc.
Trade Name. Nemathorin (Ishihara Sangyo).

Isazofos [*42509-80-8*], *O*-(5-chloro-1-isopropyl-1*H*-1,2,4-triazol-3-yl) *O*,*O*-diethyl phosphorothioate, $C_9H_{17}ClN_3O_3PS$, M_r 313.7, *bp* 120 °C (130 mPa), vapor pressure 4.3 mPa (20 °C), yellow liquid which is moderately soluble in water, but miscible with organic solvents [20, p. 762].

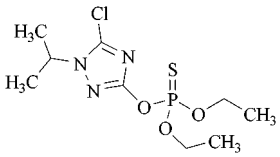

Isazofos is synthesized by treating ethyl 3-isopropylcarbazate with cyanogen chloride, followed by cyclization with phosgene. Phosphorylation of the triazole with *O*,*O*-diethylthiophosphoryl chloride yields isazofos [37]. Isazofos is a soil-applied nematicide and insecticide with contact and systemic properties for use on bananas, citrus fruits, cotton, leeks, maize, rice, vegetables, sugar cane, and in turf against soil-dwelling insects.
Trade Name. Miral (Novartis).

Phorate [*298-02-2*], see → Acaricides, Chap. 3.1.

Terbufos [*13071-79-9*], *S*-(*tert*-butylthio)-methyl *O*,*O*-diethyl phosphorodithioate, $C_9H_{21}O_2PS_3$, M_r 288.4, *bp* 69 °C (1 Pa), *mp* −29.2 °C, vapor pressure 34.6 mPa (25 °C), is a colorless to pale yellow liquid which is sparingly soluble in water but readily soluble in most common organic solvents [20, p. 1165].

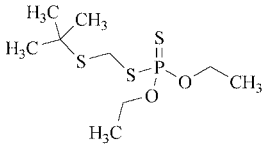

Terbufos is prepared by treating *O*,*O*-diethyldithiophosphoric acid with formaldehyde and *tert*-butanethiol [38]. Terbufos is a nernaticide and insecticide used on bananas and sugar beet.
Trade Name. Counter (Cyanamid); Contraven (Cyanamid, BASF, DowElanco); Cyanater (Siaga); Hunter (United Phosphorus); Testin (Sanonda).

Thionazin [*297-97-2*] was used as a soil insecticide and nematicide.

Triazophos [*24017-47-8*], see → Acaricides, Chap. 3.1.

4.4. Other Nematicides

DCIP [*108-60-1*], bis(2-chloro-1-methylethyl) ether, $C_6H_{12}Cl_2O$, M_r 171.1, *bp* 187 °C (760 mm Hg), light brown oily liquid which is moderately soluble in water [20, p. 330].

$[ClCH_2CH(CH_3)]_2O$

DCIP is produced by treating isopropyl alcohol with anhydrous acid chloride at low temperature [39]. DCIP is used for the control of nematodes and citrus ground mealybug in vegetables, fruit, tobacco, ornamentals, tea, and mulberries.
Trade Name. Nemamort (SDS Biotech).

GY-81 [*7345-69-9*], sodium tetrathio(peroxocarbonate), $CNaS_4$, M_r 186.2, is a hygroscopic orange crystalline solid which is very soluble in water [20, p. 653].

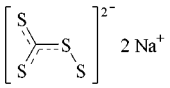

GY-81 is produced by reaction of carbon disulfide with sulfur in alkali ethoxide [40]. GY-81 decomposes in the soil to the broad-spectrum biocide carbon disulfide, it is used as a contact fumigant for the management of grape phylloxera, plant parasitic nematodes and various soil-borne pathogens causing root rot.
Trade Name. Enzone (Unocal).

5. Toxicology

5.1. Fumigants and Other Agents

Chloropicrin [*76-06-2*] [41]

LD_{50} (rat, oral)	250	mg/kg
LC_{50} (mouse, 10-min inhalation)	1.60	mg/L air
LC_{50} (cat, 20-min inhalation)	0.8	mg/L air
LC_{50} (dog, 20-min inhalation)	0.8	mg/L air
LC_{50} (rabbit, 20-min inhalation)	0.8	mg/L air
Rabbit: severe skin irritant		

Dazomet [*533-74-4*] [42]

LD_{50} (rat, oral)	550 – 710	mg/kg
LD_{50} (mouse, oral)	430 – 455	mg/kg
LD_{50} (mouse, dermal)	2400 – 2530	mg/kg
LD_{50} (rat, dermal)	2260 – 2600	mg/kg
LC_{50} (rat, 4-h inhalation)	7.29 to > 8.4	mg/L air
Rabbit: mild eye irritant		
Rabbit: not a skin irritant		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL:		
Reproduction.		
Rat (2-generation feed)	0.5	$mg\,kg^{-1}\,d^{-1}$ (parental)
	18	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Teratogenicity.		
Rat (gavage)	3	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	12.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	0.3 – 1.3	$mg\,kg^{-1}\,d^{-1}$
Mouse (78-week feed)	4 – 6	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	0.5 – 1.5	$mg\,kg^{-1}\,d^{-1}$

1,3-Dichloropropene [*542-75-6*] [20, p. 366]

LD_{50} (rat, oral)	110 – 713	mg/kg
LD_{50} (rat, dermal)	> 1211	mg/L air
LD_{50} (rabbit, dermal)	333 – 540	mg/kg
LC_{50} (rat, 4-h inhalation)	2.70 – 3.07	mg/L air
Rabbit: skin and eye irritant		
NORL/NOAEL:		
Inhalation studies:		
Subchronic.		
Rat (90-d)	0.05	mg/L air
Mouse (90-d)	0.05	mg/L air
Reproduction.		
Rat	0.45	mg/L air
Chronic.		
Rat (2-year)	0.099	mg/L air
Mouse (2-year)	0.025	mg/L air

Ethylene dibromide [*106-93-4*] [43, p. 683]

LD_{50} (rat, oral)	146 – 420	mg/kg	[20, p. 490]
LD_{50} (mouse, oral)	420	mg/kg	[20, p. 490]
LD_{50} (rabbit, oral)	55	mg/kg	
LD_{50} (guinea pig, oral)	110	mg/kg	
LD_{50} (chick, oral)	79	mg/kg	
LD_{50} (rat, dermal)	450	mg/kg	
LD_{50} (rabbit, dermal)	450	mg/kg	
LC_{50} (rat, 7-h inhalation)	0.35	mg/L air	
Rabbit: skin and eye irritant			
ADI: $1.0\,mg\,kg^{-1}\,d^{-1}$			

Metam sodium [*144-42-8*] [20, p. 797]

LD_{50} (rat, oral)	1700 – 1800	mg/kg
LD_{50} (mouse, oral)	285	mg/kg
LD_{50} (rat, oral)	97	mg/kg for isothiocyanate formed in soil
LD_{50} (rabbit, dermal)	1300	mg/kg
LC_{50} (rat, 4-h inhalation)	> 4.7	mg/L air

Rabbit: skin irritant
Rabbit: severe eye irritant
NOEL/NOAEL:
Subchronic.

Rat (65-d (6-h d^{-1}, 5-d $week^{-1}$))	0.045	mg/L air

Methyl bromide [*74-83-9*] [20, p. 824]

LC_{50} (rat, 4-h inhalation)	3.03	mg/L air
LC_{50} (rat, 8-h inhalation)	2.1	mg/L air
LC_{50} (mouse, one-h inhalation)	8.4	mg/L air

Rabbit: skin and eye irritat

Methyl isothiocyanate [*556-61-6*] [44]

LD_{50} (rat, oral)	72 – 175	mg/kg
LD_{50} (mouse, oral)	90 – 104	mg/kg
LD_{50} (rat, dermal)	2780	mg/kg
LD_{50} (mouse, dermal)	1870	mg/kg
LD_{50} (rabbit, dermal)	263	mg/kg
LC_{50} (rat, 1-h inhalation)	1.9	mg/L air

Rabbit: moderate skin irritant
Rabbit: severe eye irritant
Guinea pig: weak skin sensitizer
Nonmutagenic
NOEL/NOAEL:
Reproduction.

Rat (2-generation, drinking water)	1	mg kg^{-1} d^{-1} (parental)
	5	mg kg^{-1} d^{-1} (developmental)

Teratogenicity.

Rat (gavage)	5	mg kg^{-1} d^{-1} (maternal and developmental)
Rabbit (gavage)	3	mg kg^{-1} d^{-1} (maternal)
	5	mg kg^{-1} d^{-1} (developmental)

Chronic.

Rat (2-year, drinking water)	0.514 – 0.746	mg kg^{-1} d^{-1}
Mouse (106-week, drinking water)	3.30 – 3.66	mg kg^{-1} d^{-1}
Dog (one-year, gavage)	0.4	mg kg^{-1} d^{-1}

5.2. Carbamates

Aldicarb [*116-06-3*] [45]

LD_{50} (rat, oral)	0.46 – 1.20	mg/kg
LD_{50} (mouse, oral)	0.38 – 1.5	mg/kg
LD_{50} (guinea pig, oral)	1.0	mg/kg
LD_{50} (rabbit, oral)	1.3	mg/kg
LD_{50} (rat, dermal)	> 10	mg/kg
LD_{50} (rabbit, dermal)	> 10	mg/kg

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL:
Reproduction.

Rat (feed)	0.3	mg kg^{-1} d^{-1}

Teratogenicity.

Rat (gavage)	0.125	mg kg^{-1} d^{-1} (maternal)
	0.25	mg kg^{-1} d^{-1} (developmental)

Chronic.

Rat (93-d feed)	0.1	mg kg^{-1} d^{-1}
Mouse (103-week feed)	0.9	mg kg^{-1} d^{-1}
Dog (one-year feed)	0.05	mg kg^{-1} d^{-1}
Human (single oral dose)	0.025	mg kg^{-1} d^{-1}

ADI: 0.003 mg kg^{-1} d^{-1}

Carbofuran [*1563-66-2*] [46]

LD_{50} (rat, oral)	6.0 – 14.4	mg/kg
LD_{50} (mouse, oral)	14.4	mg/kg
LD_{50} (dog, oral)	15 – 19	mg/kg
LD_{50} (cat, oral)	2.5 – 3.5	mg/kg
LD_{50} (chick)	2.5 – 38.9	mg/kg
LD_{50} (rabbit, oral)	7.5	mg/kg
LD_{50} (guinea pig)	9.2	mg/kg
LD_{50} (rat, dermal)	> 500	mg/kg
LD_{50} (rabbit, dermal)	> 46.4	mg/kg
LC_{50} (rat, 4-h inhalation)	75 – 88	µg/L air

Rabbit: mild skin and eye irritant
Nonmutagenic
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed)	1.2	mg kg^{-1} d^{-1} (parental and developmental)

Teratogenicity.

Rat (feed)	1.5	mg kg^{-1} d^{-1} (maternal)
	4.4	mg kg^{-1} d^{-1} (developmental)
Rabbit (gavage)	0.5	mg kg^{-1} d^{-1} (maternal and developmental)

Chronic.

Rat (2-year feed)	1	mg kg^{-1} d^{-1}
Mouse (2-year feed)	28	mg kg^{-1} d^{-1}
Dog (one-year feed)	0.25	mg kg^{-1} d^{-1}

ADI: 0.002 mg kg^{-1} d^{-1}

Oxamyl [*23135-22-0*] [47]

LD_{50} (rat, oral)	2.5 – 16	mg/kg
LD_{50} (mouse, oral)	2.3 – 3.3	mg/kg
LD_{50} (guinea pig, oral)	7.1	mg/kg
LD_{50} (rat, dermal)	> 1200	mg/kg
LD_{50} (rabbit, dermal)	130	mg/kg
LC_{50} (rat, 4-h inhalation)	0.12 – 0.17	mg/L air

Rabbit: mildly irritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL:
Reproduction.

Rat (3-generation feed)	2.5	mg kg^{-1} d^{-1} (parental and developmental)

Teratogenicity.		
Rat (gavage)	ca. 4	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	ca. 20	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (feed)	ca. 1	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	2	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	2.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	3.6	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	2.5	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$

5.3. Organophosphorus Esters

Cadusafos [*95465-99-9*] [48]

LD_{50} (rat, oral)	39 – 132	mg/kg
LD_{50} (mouse, oral)	71	mg/kg
LD_{50} (rabbit, dermal)	11	mg/kg
LC_{50} (rat, 4-h inhalation)	0.032	mg/L air

Rabbit: eye irritant
Rabbit: slightly skin irritant
Nonmutagenic
NOEL/NOAEL:

Reproduction.		
Rat (2-generation feed)	30	$\mu g\,kg^{-1}\,d^{-1}$ (parental and developmental)
Teratogenicity.		
Rat (gavage)	2	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	6	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	0.3	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Chronic.		
Rat (2-year feed)	50	$\mu g\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	89	$\mu g\,kg^{-1}\,d^{-1}$
Dog (capsule)	0.1	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.0003\,mg\,kg^{-1}\,d^{-1}$

Ethoprophos [*13194-48-4*] [49]

LD_{50} (rat, oral)	33 – 62	mg/kg
LD_{50} (mouse, oral)	18 – 31	mg/kg
LD_{50} (rabbit, oral)	55	mg/kg
LD_{50} (rat, dermal)	226	mg/kg
LD_{50} (pig, dermal)	327	mg/kg
LD_{50} (rabbit, dermal)	26	mg/kg
LC_{50} (rat, 4-h inhalation)	0.123	mg/L air

Rabbit: skin and eye irritant
Nonmutagenic
NOEL/NOAEL:

Chronic.		
Rat (2-year feed)	0.05	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	0.035	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	0.025	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.0003\,mg\,kg^{-1}\,d^{-1}$

Fenamiphos [*22224-92-6*] [50]

LD_{50} (rat, oral)	2.4 – 19.4	mg/kg
LD_{50} (mouse, oral)	22.7	mg/kg
LD_{50} (guinea pig, oral)	56 – 100	mg/kg
LD_{50} (rabbit, oral)	10 – 17.5	mg/kg
LD_{50} (cat, oral)	10	mg/kg
LD_{50} (god, oral)	ca. 10	mg/kg
LD_{50} (chick, oral)	5.3 – 12	mg/kg
LD_{50} (rat, dermal)	72 – 92	mg/kg
LD_{50} (rabbit, dermal)	178 – 225	mg/kg
LC_{50} (rat, 4-h inhalation)	110 – 150	µg/L air

Rabbit: mild eye irritant
Rabbit: minimal skin irritant
Guinea pig: mild skin sensitizer
Nonmutagenic
NOEL/NOAEL:

Reproduction.		
Rat (2-generation feed)	0.17	$mg\,kg^{-1}\,d^{-1}$ (parental)
	0.64	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Teratogenicity.		
Rat (gavage)	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	3	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	0.1	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	0.3	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	0.46	$mg\,kg^{-1}\,d^{-1}$
Mouse (20-month feed)	0.3	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	0.083	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.0008\,mg\,kg^{-1}\,d^{-1}$

Fosthiazate [*98886-44-3*] [20, p. 631]

LD_{50} (rat, oral)	57 – 73	mg/kg
LD_{50} (rat, dermal)	861 – 2396	mg/kg
LC_{50} (rat, 4-h inhalation)	0.588	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: skin sensitizer

Isazofos [*42509-80-8*] [20, p. 726]

LD_{50} (rat, oral)	40 – 60	mg/kg
LD_{50} (rat, dermal)	118 – > 3100	mg/kg
LC_{50} (rat, 4-h inhalation)	0.24	mg/L air

Rabbit: mild skin irritant
Rabbit: very slight eye irritant

Subchronic.		
Rat (90-d feed)	0.2	$mg\,kg^{-1}\,d^{-1}$
Dog (90-d feed)	0.05	$mg\,kg^{-1}\,d^{-1}$

Phorate [*298-02-0*], see→ Acaricides

Terbufos [*13071-79-9*] [51]

LD_{50} (rat, oral)	1.6	mg/kg
LD_{50} (mouse, oral)	5.0 – 9.2	mg/kg
LD_{50} (rabbit, oral)	1.6	mg/kg
LD_{50} (rat, dermal)	0.97 – 9.8	mg/kg
LD_{50} (rabbit, dermal)	0.81 – 1.0	mg/kg
LC_{50} (rat, 4-h inhalation)	6.1	µg/L air
Nonmutagenic		
NOEL/NOAEL:		
Reproduction.		
Rat (3-generation feed)	0.016 – 0.035	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	> 0.20	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	0.2	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Chronic.		
Rat (2-year feed)	0.01 – 0.013	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	0.9	$mg\,kg^{-1}\,d^{-1}$
Dog (one-year feed)	0.06	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.0002\,mg\,kg^{-1}\,d^{-1}$		

Triazophos [*24017-47-8*], see→ Acaricides

5.4. Other Nematicides

DCIP [*108-60-1*] [20, p. 339]

LD_{50} (rat, oral)	503	mg/kg
LD_{50} (mouse, oral)	530 – 599	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	12.8	mg/L air
LC_{50} (mouse, 4-inhalation)	12.8	mg/L air
NOEL/NOAEL:		
Chronic.		
Rat (2-year feed)	13.4	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.13\,mg\,kg^{-1}\,d^{-1}$		

GY-81 [*7345-69-9*] [20, p. 653]

LD_{50} (rat, oral)	631	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	3.17 – 4.73	mg/L air
Rabbit: moderate eye irritant		
Rabbit: severe skin irritant		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
ADI: $6\,mg\,kg^{-1}\,d^{-1}$		

6. References

General References

1. G. N. Agrios: *Plant Pathology,* Academic Press, San Diego-London 1988, pp. 703 – 746.
2. R. H. Brown, B. R. Kerry (eds.): *Principles and Practice of Nematode Control in Crops,* Academic Press, Sydney-New York-London-Toronto 1987.
3. J. A. Bunt: *Effect and Mode of Action of Some Systemic Nematicides,* H. Veenman & Zonen B. V., Wageningen 1975.
4. W. R. Jenkins, D. P. Taylor: *Plant Nematology,* Reinhold Publishing Corp., New York-Amsterdam-London 1967.
5. M. Luc, R. A. Sikora, J. Bridge (eds.): *Plant Parasitic Nematodes in Tropical and Subtropical Agriculture,* CAB International, Wallingford, UK, in press.
6. A. Maggenti: *General Nematology,* Springer-Verlag, New York-Heidelberg-Berlin 1981.
7. D. Mulder (ed.): *Soil Disinfestation,* Elsevier Scientific Publishing Co., Amsterdam-Oxford-New York 1979.
8. W. L. Nicholas: *The Biology of Free-Living Nematodes,* Clarendon Press, Oxford 1984.
9. W. M. Nickle (ed.): *Plant and Insect Nematodes,* Marcel Dekker, New York-Basel 1984.
10. J. E. Peachey (ed.): *Nematodes of Tropical Crops,* Commonwealth Agricultural Bureaux, St. Albans, Herts., UK, 1969.
11. M. R. Siddiqi: *Tylenchida Parasites of Plants and Insects,* Commonwealth Agricultural Bureaux, Farnham Royal, Slough, UK, 1986.
12. J. F. Southey (ed.): *Laboratory Methods for Work with Plant and Soil Nematodes,* Ministry of Agriculture, Fisheries and Food, Reference Book 402, London 1986.
13. J. A. Veech, D. W. Dickson (eds.): *Vistas on Nematology,* Society of Nematologists, Inc., Hyattsville, Maryland, 1987.
14. J. M. Webster (ed.): *Economic Nematology,* Academic Press, London-New York 1972.
15. B. M. Zuckerman, W. F. Mai, R. A. Rhode (eds.): *Plant Parasitic Nematodes,* vol. II, Academic Press, New York-London 1971.

Specific References

16. W. F. Mai, H. H. Lyon: *Pictorial Key to Genera of Plant-Parasitic Nematodes,* Cornstock Publishing Associates, Cornell University Press, Ithaca, London 1982.
17. R. Mc Sorley: "Extraction of Nematodes and Sampling Methods," in R. H. Brown, B. R. Kerry (eds.): *Principles and Practice of Nematode Control in Crops,* Academic Press, Sydney-New York-London-Toronto 1987.

18. J. N. Sasser, D. W. Freckman: "A World Perspective on Nematology: The Role of the Society" in J. A. Veech, D. W. Dickson (eds.): *Vistas on Nematology: A Commemoration of the Twenty-Fifth Anniversary of the Society of Nematologists,* Society of Nematologists, Hyattsville, Maryland 1987.
19. Wood Mackenzie & Co.: *Agrochemical Service*, Edinburgh, October 1998,
20. C. D. S. Tomlin: *The Pesticide Manual* 11th ed., British Crop Protection Council, Farnham 1997.
21. R. L. Datta, N. R. Chatterjee, *J. Am. Chem. Soc.* **37** (1915) 567.
22. Bayer, DE 2 227 338, 1972 (R. Schubert, U. Holzer, E. Roos, Th. Kempermann).
23. N. Drescher, S. Otto, *Residue Rev.* **23** (1968) 49.
24. Associated Ethyl Co., GB 804 995, 1958 (W. J. Read, D. B. Clapp, D. R. Stephus, J. E. Russel).
25. E. C. Williams, *Trans. Am. Inst. Chem. Eng.* **37** (1941) 157.
26. Zeneca, US 2 766 554, 1954 (St. C. Dorman, A. B. Lindquist).
27. Degussa, GB 768 893, 1957.
28. Bayer, DE 1 172 257, 1963 (E. Mühlbauer, E. Meisert).
29. M. H. J. Weiden et al., *J. Econ. Entomol.* **58** (1965) 154.
30. Union Carbide Corp., US 3 217 037, 1962 (L. K. Payne, Jr., M. H. J. Weiden).
31. Bayer, DE 1 493 646
32. Du Pont, S. Afr. 6 803 629, 1968 (J. B. Buchanan).
33. Rhône Poulenc, EP 235 056, 1987 (J. M. Brochard, F. Frison, P. Le Roy).
34. Mobil Oil, US 3 268 393, 1965 (J. H. Wilson, Jr.).
35. Bayer, US 2 978 479, 1959, DE 1 121 882, 1958 (H. Kayser, G. Schrader).
36. Nippon, JP 3 261 792, 1991 (R. Nasu et al.).
37. Ciba-Geigy, GB 1419 t31, 1 419 132, 1971 (D. Dawes, B. Böhner).
38. Cyanamid, US 2 586 655, 1952 (E. O. Hook, Ph. H. Moss).
39. L. Brandsma, J. F. Arens, *Rec. Trav. Chim.* **81** (1962) 33.
40. E. W. Joeman, *J. Soc. Chem.* **119** (1921) 38.
41. P. J. Gehring, R. J. Nolan, T. G. Watanabe, A. H. Schuman in W. J. Hayes, Jr., R. R, Laws, Jr., (eds): *Handbook of Pesticide Toxicology,* Academic Press, San Diego 1991, p. 675.
42. BASF, *J. Pesticide Sci.* **17** (1992) 327.
43. "Pesticide Residues in Food-1966," *FAO Plant Prod. Protect. Paper* **67.32** (1967) 97.
44. Nihon Schering, *J. Pesticide Sci.* **15** (1990) 297.
45. "Pesticide Residues in Food-1992," *WHO/PCS* **93.34** (1993) 33.
46. "Pesticide Residues in Food-1996," *WHO/PCS* **97.1** (1997) 23.
47. "Pesticide Residues in Food-1984," *FAO Plant Prod. Protect. Paper* **62** (1985) 59.
48. "Pesticide Residues in Food-1991," *WHO/PCS* **92.52** (1992) 97.
49. "Pesticide Residues in Food-1987," *FAO Plant Prod. Protect. Paper* **86/2** (1988) 85.
50. "Pesticide Residues in Food-1997," *WHO/PCS* **98.6** (1998) 59.
51. "Pesticide Residues in Food-1990," *WHO/PCS* **91.47** (1991) 151.

Neodymium → Rare Earth Elements
Neon → Noble Gases
Neopentylglycol → Alcohols, Polyhydric
Neptunium → Transuranium Elements
Neuroleptanalgesics → Anesthetics, General
Neuroleptics → Psychopharmacological Agents

Neuropharmacology

VICTOR P. WHITTAKER Arbeitsgruppe Neurochemie, Max-Planck-Institut für biophysikalische Chemie, Göttingen, Federal Republic of Germany Johannes Gutenberg Universität , Mainz, Federal Republic of Germany

1. Introduction 1
2. Types of Neural Cells 2
2.1. Neurons 2
2.2. Glia 3
3. Evolution and Organization of the Vertebrate Nervous System 4
3.1. Evolution and Main Subdivisions . 4
3.2. Patterns of Neuronal Integration . . 5
3.2.1. The Reflex Arc 5
3.2.2. Voluntary Movement 7
3.2.3. The Limbic System 7
3.2.4. The Ascending Reticular Activation System 8
4. Techniques for Investigating the Nervous System 8
4.1. Neuroanatomy 8
4.2. Neurophysiology 10
4.3. Neurobiochemistry (Neurochemistry) 11
4.4. Neuropharmacology 13
5. Synaptic Transmission 14
5.1. Historical Background 14
5.2. Survey of Transmitters 15
5.2.1. Low Molecular Mass Transmitters . . 15
5.2.2. Neuropeptides 16
5.2.3. Evolutionary Significance of the Two Types of Transmitter 18
5.3. Transmitter Action 18
5.3.1. Fast Ionotropic Action 18
5.3.2. Slow Metabolotropic Action 19
5.3.3. Receptors of Importance in Drug Action 20
5.4. Axonal Transport 21
6. References 22

1. Introduction

Most drugs or toxins act on the nervous system in a highly specific way [1–7]. In many cases the particular subsystem or component (e.g., enzyme, ion channel, or transmitter receptor) with which the drug interacts is known precisely. In other cases, the target may be unidentified, but research into the mode of action of the drug, toxin, or toxic substance is best planned on the assumption that the action is specific, and a clue will often be provided by the symptoms that the agent produces in animals or humans. The nonmetabolized gaseous anesthetics (→ Anesthetics, General) are one of the few classes of pharmacological agents to act in a more generalized way on nerve cells but here too their mode of action is to selectively induce biophysical changes in nerve cell membranes which block their excitability, rather than to exert a general effect on the whole cell.

The purpose of this article is to summarize what is known about the organization and function of the nervous system, its chief subsystems and components down to the molecular level in order to provide a background for the following articles on specific topics in neuropharmacology:

Analeptics
Analgesics and Antipyretics
Anesthetics, General
Anti-Obecity Drugs
Antiarrhythmic Drugs
Antiasthmatic Agents
Antiemetics
Antiepileptics
Antihypertensives
Antihypotensives
Calcium Antagonists
Cardiac Glycosides and Synthetic Cardiotonic Drugs
Coronary Therapeutics
Cough Remedies
Hypnotics
Local Anesthetics
Parasympatholytics and Parasympathomimetics
Parkinsonism Treatment
Psychopharmacological Agents
β-Receptor Blocking Agents
Sedatives
Skeletal Muscle Relaxants
Spasmolytics

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_135

Sympatholytics
Sympathomimetics

2. Types of Neural Cells

The cells making up nervous tissue can be divided broadly into two types: *nerve cells* (*neurons*) whose task is the transmission and processing of information, and *glia*, to which an insulating and supporting role is assigned (Fig. 1). Within these two categories are many subtypes whose classification is based both on structure and function.

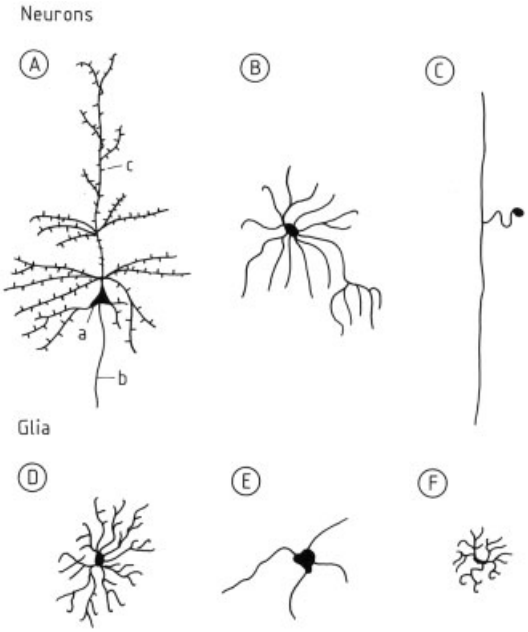

Figure 1. Types of neurons (A – C) and glia (D – F)
Neurons (Golgi staining): A) Pyramidal neuron with cell body (a), axon (b), and dendrites (c); B) Basket cell; C) Bipolar cell
Glia: D) Astrocyte; E) Oligodendrocyte; F) Microglia

2.1. Neurons

A primary distinction can be made between sensory, motor, and interneurons. In the peripheral nervous system (Section 3.1), *sensory neurons* are linked to a variety of sensors (sensory receptors) by which different sensations (touch, pain, heat, cold, pressure, tension, sound, light, smell, taste) and certain internal parameters (such as blood oxygen and carbon dioxide tension) are converted into trains of nerve impulses that convey information about the internal and external environment to the central nervous system (CNS), often through ascending sensory or afferent pathways involving several neurons. By contrast, *motor neurons*, usually under the control of descending or efferent pathways, convey signals to and stimulate effector cells (muscle or gland cells) causing muscles to contract or gland cells to secrete physiologically active substances (hormones, enzymes).

The constant flow of information conveyed by afferent impulses needs to be processed before it can result in effective action. This is achieved by *interneurons* that are often arranged in complex repetitive patterns which resemble the logic units of computer circuits. However, although the nervous system may be thought of as a biological computer regulating and optimizing the responses of the organism to changes in its environment, any close comparison with man-made computers reveals a fundamental difference: the nervous system is not digitalized. It performs numerical computations slowly and with difficulty but is excellent at pattern recognition, association, learning and memory, though again, not of the computational type.

The morphology of nerve cells is elegantly displayed at the light-microscopic level by silver staining, a technique originally devised by C. Golgi (1843 – 1926). In this method only a few cells are stained in a given section of tissue, but the stain spreads until the whole neuron is displayed with all its ramifications (Fig. 1). Cells can then be classified according to their morphology: *pyramidal cells* (Fig. 1 A) are a type of efferent cell with prominent triangular cell bodies; *basket* (Fig. 1 B) and *stellate* (star-shaped) cells are interneurons with bushy processes resembling the wickerwork of a basket or the rays of a star, respectively.

All neurons possess a cell body, which contains a prominent *nucleus* surrounded by *cytoplasm* (perikaryon); in addition almost all neurons send out several or many processes known as *dendrites* (from their resemblance to the branches of trees) and a single process known as an axon. The dendrites and cell body collectively form a receptive field which receives contacts from the axons of neighboring cells. The axon conveys impulses arising from the

convergence of this receptive field to the target cell–another nerve cell or effector (muscle or gland) cell. In sensory cells a dendrite and axon may fuse to form a bipolar cell; impulses generated by receptors attached to the dendrite are picked up by the axon, bypassing the peripherally placed cell body (Fig. 1 C). Towards its terminal the axon may branch several times before terminating on another cell; it may also form swellings or *varicosities* along its length which are in close contact with the target cells. Within the neuron, information is conducted by nerve impulses–self-propagating waves of ion flux that can be detected by minute accompanying fluctuations in membrane potential (Section 4.2). The transmission of information between neurons is brought about by the release of chemical transmitter substances (*neurotransmitters*) from stores in the axon terminal or varicosity (Chap. 5). Neurotransmitters are released by the arrival of a nerve impulse and are of two types–low molecular mass compounds such as acetylcholine and noradrenaline (Section 5.2.1) or neuropeptides (Section 5.2.2); they interact with *receptors* on the cell membrane of the target cell and induce nerve impulses in it. Many drugs owe their effect to their ability to selectively bind to these receptors (Section 5.3.3). The complex of nerve axon termination together with the region of the cell membrane of the target cell immediately apposed to it is known as a *synapse*, a word coined by C.S. Sherrington (1857 – 1952) from two Greek words meaning a "clasping together".

Because so many drugs and toxins act by interfering with synaptic transmission, the synapse has sometimes been called "the Achilles heel" of the nervous system.

2.2. Glia

Glial cells can also be classified on the basis of morphology and function. Three main types are recognized in the CNS: astrocytes, oligodendrocytes, and microcytes. In the PNS, Schwann cells are the peripheral counterparts of oligodendrocytes.

Glial cells were originally regarded as the neural counterpart of connective tissue and as the glue (Greek glía) holding the tissue together. Embryological studies showed that astrocytes and oligodendrocytes, like neurons, are of ectodermal, not mesodermal origin; microglia, however, are derived from mesoderm, like connective tissue.

Astrocytes are star-shaped cells with many radiating processes (Fig. 1 D). The processes may form end-feet applied to brain capillaries and so contribute to the structural and functional barrier between the brain and the blood known as the *blood – brain barrier* or they may surround nerve cells in a manner suggesting a supportive or service function. Much evidence indicates that astrocytes regulate the environment of nerve cells by controlling the extracellular concentrations of hydrogen and potassium ions and chemical transmitters, and by releasing trophic factors essential for the survival of neurons.

Oligodendrocytes. The primary function of these cells (Fig. 1 E) in the CNS and of their counterparts, the *Schwann cells*, in the PNS is to provide an insulating wrapping around axons. This wrapping known as *myelin* is applied in segments (Fig. 2); the short, poorly insulated spaces between segments are known as nodes of Ranvier after their discoverer, L.A. Ranvier (1835 – 1922). The segmental character of this insulation confers an important advantage on the neuron; propagated nerve impulses are forced to jump from one node to the next, a phenomenon known as *saltatory conduction*, and this greatly speeds up nerve conduction.

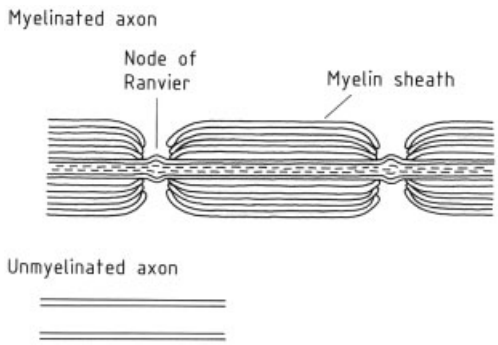

Figure 2. Myelinated and unmyelinated axons

Microcytes (Fig. 1 F) are wandering cells that invade injured or inflamed regions of the nervous system from the blood stream and remove dead cells and tissue debris by phagocytosis. They are not present in healthy nervous tissue in significant numbers.

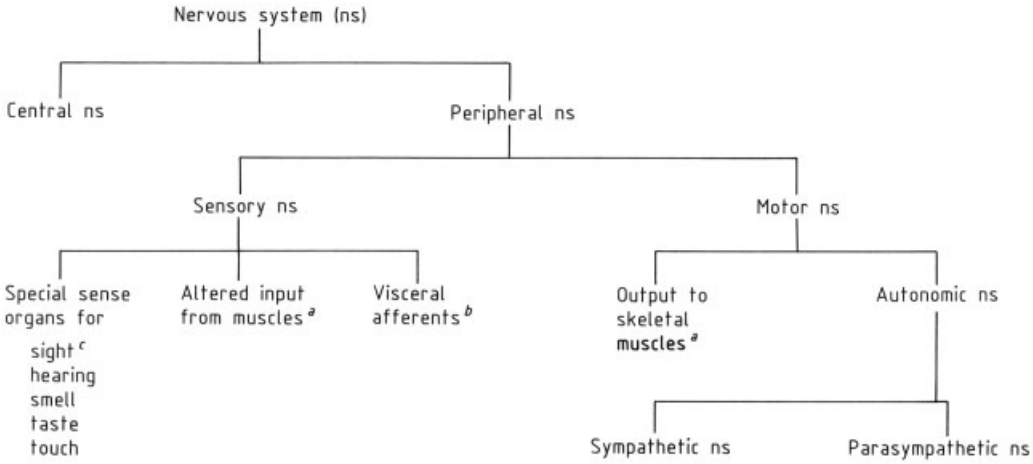

Figure 3. Subdivisions of the vertrebrate nervous system
[a] These portions of the peripheral nervous system are collectively referred to as the somatic nervous system. [b] Often classified as the afferent portion of the autonomic nervous system. [c] The retina is often regarded as a peripheral extension of the CNS.

3. Evolution and Organization of the Vertebrate Nervous System

3.1. Evolution and Main Subdivisions

The simplest vertebrate nervous system is seen in fish, amphibia, and reptiles. Birds and mammals represent separate developments from primitive reptilian forms, mainly characterized by a relative expansion of the parts of the brain concerned with sensory processing and motor agility. Nevertheless, the basic plan of all vertebrate nervous systems is similar (Fig. 3) being divided into the *central nervous system* (CNS), comprising the brain and spinal cord, and the *peripheral nervous system* (PNS), comprising the *somatic nerves* (those innervating the voluntary skeletal musculature), the *autonomic nerves* (those innervating glands and the smooth involuntary musculature of the viscera and blood vessels), and the nerves from the organs of special senses. An alternative, more functional classification of the PNS distinguishes between sensory (afferent) and motor (efferent) nerves. The *sensory nerves* mediate information from the special senses, the somatic (skeletal) musculature, and the viscera. The *motor nerves* induce the contraction of skeletal muscle and control the hollow organs of the body and the glands, the latter being termed *autonomic nerves*. These nerves are further subdivided into *sympathetic* and *parasympathetic*. The *sympathetic nervous system* prepares the body for emergencies: the blood supply to the muscles, heart, and brain is increased by dilatation of their blood vessels, while that to the skin and viscera is reduced by vasoconstriction. The heart beats faster, the urinary and rectal outlets are blocked by contraction of their sphincters, intestinal movement and alimentary secretion are stopped, and blood sugar is raised. By contrast the *parasympathetic system* restores bodily function when the crisis is over. Wastes are eliminated by the relaxation of the appropriate sphincters, blood pressure and heart-rate fall, and alimentary function is resumed. Many drugs block (antagonists) or stimulate (agonists) the action of the transmitters involved (→ Parasympatholytics and Parasympathomimetics ; → Sympatholytics and Sympathomimetics).

The anatomy of the sensory and somatic motor systems is relatively simple (Fig. 4 A). Sensory nerves derived from bipolar neurons (Section 2.1) enter the spinal cord through the dorsal (posterior) roots; their cell bodies are located in the dorsal root ganglia. The motor nerves leave the spinal cord through the ventral (anterior) roots; their cell bodies are located in the ventral columns of the spinal cord, protuberances of neuron-rich grey matter jutting out from the central mass of grey matter in the spinal cord.

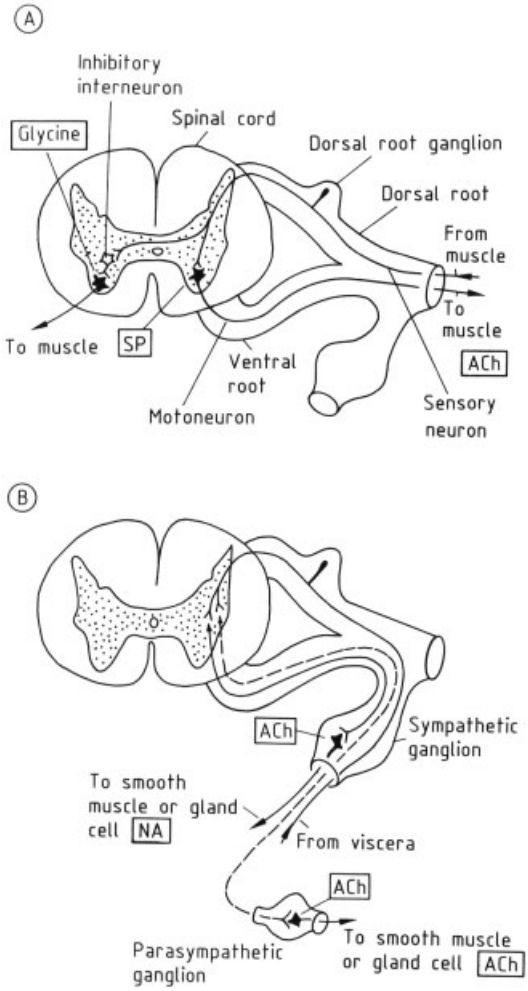

Figure 4. The relationship of the peripheral nervous system to the spinal cord
A) Somatic system: sensory nerves from muscle receptors form a two-neuron arc with a motor neuron (motoneuron) on the same side and inhibit a motor neuron on the opposite side via an inhibitory interneuron, thus forming a three-neuron arc.
B) Autonomic system: continuous lines indicate visceral afferents, preganglionic sympathetic, and postganglionic; dashed lines indicate preganglionic parasympathetic.
The transmitters utilized (Section 5.2) are shown in boxes; ACh = acetylcholine, NA = noradrenaline, SP = substance P. In reality both systems run together in each root and bilaterally; they are shown separately and unilaterally for clarity.

The autonomic outflow is more complex (Fig. 4 B). Motor nerves issuing from the ventral roots enter ganglia arranged in chains on either side of the backbone. Some (the *preganglionic sympathetic* nerves) make contact with nerve cell bodies in these ganglia. The axons of these ganglion cells (the *postganglionic sympathetic* nerves) provide the sympathetic innervation to the heart, viscera, and blood vessels. Other preganglionic nerves (the *preganglionic parasympathetic* nerves) travel through the ganglia, usually accompanying the postganglionic sympathetic innervation to the heart, viscera, and blood vessels, and eventually making contact with ganglion cells nearer or in the target organs themselves. These cells provide the *postganglionic parasympathetic* innervation. How these connections were worked out is explained in Section 5.1. Brain weight is not a good guide to the degree of evolution of the nervous system since large animals have large brains but may show a behavior no more complex than that of their smaller congeners. The larger brains of large members of a particular family or order contain more glial cells, not necessarily more neurons. To accommodate the increased surface area of larger brains, the outer layers (the *cortex*) are thrown into folds; this is particularly obvious in the human brain, where, compared to other animals, there is in addition a net increase (though not as large as is often supposed) in cortical mass relative to whole brain mass.

Another criterion of development is the relative volume of the hind-brain (cerebellum), a part of the brain concerned with the performance of skilled acts. Originally evolved to control swimming movements in fish, it shows marked development in birds and mammals, especially primates.

3.2. Patterns of Neuronal Integration

3.2.1. The Reflex Arc

The unit of integrated action in the nervous system is the *reflex*. The study of reflexes was begun by SHERRINGTON [8] in decapitated (spinal) cats, maintained by artificial respiration and warmth. Such preparations show integrated responses to stimuli of a quasi-mechanical kind– the foot is sharply withdrawn in response to a prick on the foot pad while the contralateral limb extends to support the body (nociceptive reflex), a limb offers resistance to being pulled (stretch reflex) or jerks when a tendon at the knee or ankle is suddenly stretched by tapping (knee or ankle

jerk); stepping movements can be elicited. Muscular movements involve the relaxation of antagonists (muscles opposing a limb movement) when agonists (muscle causing a movement) contract and, very often, compensating movements on the opposite side.

SHERRINGTON grasped that the simplest integrative unit consisted of a *two-neuron reflex arc*, comprising an afferent sensory neuron synapsing with a motor neuron (Fig. 4 A). Thus, in the stretch reflex, stretching a muscle generates impulses from muscle stretch receptors which are conducted to synapses on motor neurons in the spinal cord and cause a reflex contraction of the muscle from which the impulses originate.

To explain the relaxation of antagonists, SHERRINGTON introduced the concept of *central inhibition*, whereby some synapses activated by incoming impulses inhibit firing of a motor neuron. Normally all motor nerves to muscles generate impulses at a low rate, thus generating the tension (muscle tone) needed to maintain posture; inhibition of a motor neuron to an antagonist muscle thus causes it to relax. Reciprocal excitation and inhibition were invoked to explain the compensatory movements of limbs on the opposite side of the body in the nociceptive and stepping reflexes: sensory impulses would activate agonists and inhibit antagonists on one side of the cord, but activate antagonists and inhibit agonists on the other side.

Careful analysis of the speed of reflex activation and inhibition of motor neurons showed that inhibitory reflex action involves a three-neuron arc with an inhibitory interneuron inserted between the sensory input and motor output (Fig. 4 A). This interneuron is activated by the afferent sensory signal and inhibits the motor neuron. The rate of firing of stretch receptors in muscle can be adjusted by the amount of contraction occurring in the fine muscle fibers (muscle spindles) to which they are attached. These can be adjusted independently of surrounding muscle fibers by the so-called γ-motor neurons (Fig. 5).

Such systems are now familiar to engineers as servomechanisms embodying negative feedback loops, but were first discovered by physiologists.

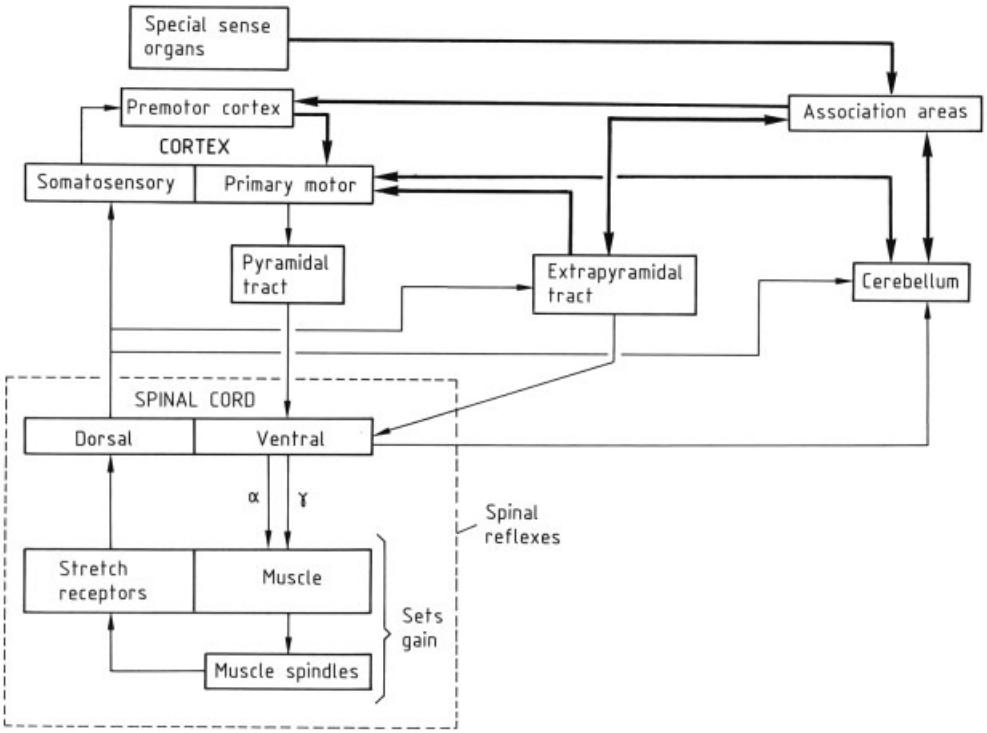

Figure 5. Simplified scheme showing main systems involved in posture and movement

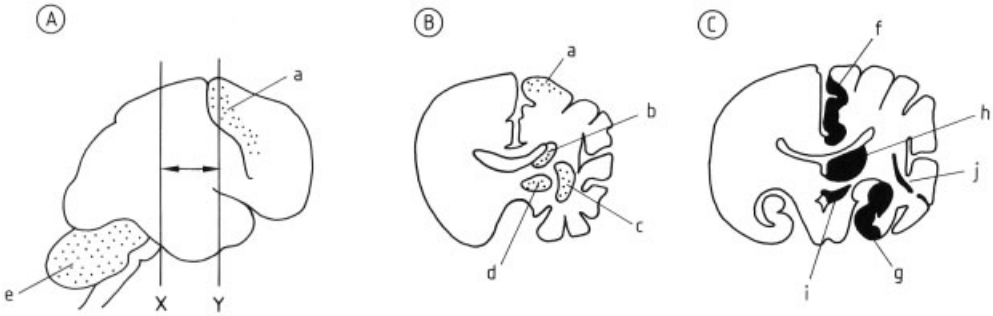

Figure 6. Sketches of human brain anatomy to illustrate the location of pyramidal, extrapyramidal, and limbic systems described in text
A) Side view of the human brain; B, C) Transverse sections along the lines in A marked X and Y, respectively
A, B) Shaded areas show structures involved in voluntary movement: a) Motor cortex containing pyramidal cells; (b – d) Basal ganglia (extrapyramidal tract); b) Caudate nucleus; c) Putamen; d) Globus pallidus; e) Cerebellum
C) Black regions are structures involved in the limbic system: f) Olfactory cortex; g) Hippocampus; h) Thalamus; i) Hypothalamus; j) Part of basal ganglia (pallidum)
For clarity, structures are shown on one side only but are actually on both sides. Arrows in A show view-points of sections.

3.2.2. Voluntary Movement

In the conscious animal muscular movement is under voluntary control, and occurs smoothly and effortlessly in spite of the vast number of adjustments that have to be made. Figures 5 and 6 summarize the main pathways involved. Playing onto the repertoire of spinal reflexes in this systems analysis are three pathways:

1) the direct pathway of the *pyramidal tract* (Fig. 6 B, a) (so called because it originates in large neurons in the part of the cortex controlling movement whose cell bodies are shaped like a pyramid; Fig. 1 A)
2) the *extrapyramidal tract* operating through a set of well defined nuclei (conglomerations of neurons) below the cortex (Fig. 6 B, b – d), and
3) the hindbrain or *cerebellum* (Fig. 6 A, e), which performs the "staff work" for a skilled movement.

Broadly, the extrapyramidal tract is involved in the control of posture and the cerebellum in the programming of skilled action. Volition seems to act through the premotor cortex. The unit of integrated activity is the spinal reflex (Fig. 4 A). Thick lines in Figure 5 indicate the circuits involved in the preparation for voluntary movement, thin lines the evolving muscular response. Diseases (Parkinsonism, Huntington's chorea) which affect the extrapyramidal system cause postural defects, difficulty in initiating movements, tremor at rest (Parkinsonism), or uncontrolled movements at rest (chorea). Damage to the cerebellum causes intention tremor and clumsy execution of movement.

Variations in the repertoire of skilled acts between animals (e.g., a cat and a chimpanzee) are largely caused (1) by the differing extents to which individual muscles or muscle groups are represented in the cortex, and (2) the degree of control over individual spinal motor neurons exerted by the cortical tracts. Strychnine is a specific blocker for inhibitory synapses (which are described in detail in Section 5.2). The painful cramps caused by its injection are due to the failure of central inhibition as antagonists pull against agonists.

The extrapyramidal tract seems particularly sensitive to degenerative diseases and drugs, some of which mimick the symptoms of such diseases while others alleviate them. Such phenomena require an understanding of the synaptic transmitters involved and will be discussed in Section 5.2.1. The transmitter dopamine plays a key role in this system.

3.2.3. The Limbic System

The limbic system is the term often used to denote those (often phylogenetically old) parts of the brain concerned with feeling and emotion (Fig. 6 C). It is much less clearly defined and

well understood than the motor system described in Sections 3.2.1 and 3.2.2. Two well-defined cortical areas, the olfactory cortex (a phylogenetically old part of the cortex concerned with smell) and the hippocampus (also concerned with memory), are involved, as well as various subcortical nuclei located in the thalamus, hypothalamus (also concerned with nervous control of many bodily functions), and the basal ganglia (also concerned with movement). Probably all the brain structures concerned with visceral control, together with their links to the parts of the brain concerned with cognition, learning, memory, and voluntary movement can be regarded as part of the limbic system. Integrated reactions indicative of rage, fear, aggression, or pleasure can be elicited in animals by local stimulation of certain subcortical limbic structures, particularly the hypothalamus, thus providing experimental justification for the concept of the limbic system. Derangements in the limbic system are believed to be involved in psychoses such as schizophrenia and in various forms of depression (→ Psychopharmacological Agents).

3.2.4. The Ascending Reticular Activation System

This system is phylogenetically old and is located in the brain stem with projections to almost all the subcortical nuclei and the brain cortex. It is believed to "activate" higher centers involved in attention, awareness, and the conscious state but is also thought to be active in the rapid eye movement phase of sleep (REM sleep). Many of the neurons forming this system use acetylcholine as their transmitter and they activate largely by suppressing inhibition (Section 5.2.1). Other important central cholinergic pathways are those arising in the basal nucleus of Meynert (nucleus basalis magnocellularis) projecting to the cortex, and the cholinergic input to the hippocampus which arises in the septum [4]. In several degenerative diseases of the nervous system (e.g., Alzheimer's disease, Parkinsonian dementia) the cholinergic neurons of the nucleus basalis are among the first to degenerate.

4. Techniques for Investigating the Nervous System

The nervous system may be investigated morphologically, physiologically, biochemically, and pharmacologically. Each approach has its advantages and limitations and only a combined approach can give the detailed knowledge needed for complete understanding.

4.1. Neuroanatomy

Neuroanatomy provides a basically static view of the nervous system; only with difficulty can series of static images be assembled into sequences from which the underlying dynamic events can be inferred.

Dissection. Gross dissection of the nervous systems of vertebrates reveals many similarities and homologies of structure with an increasing complexity of organization as the evolutionary tree is ascended, culminating in humans. The macroscopic components of the brain with their areas of grey (neuron-rich) and white (myelin-rich) matter are given fanciful names based on their appearance (e.g., hippocampus, caudate nucleus) or more logical ones based on location (e.g., temporal lobe) or known function (e.g., motor cortex).

Light Microscopy. Histology at the light microscopic level reveals individual neuronal and glial cells and gives some clues as to connections. Histochemical methods, based on staining for specific enzymes or antigens, reveal differences within the broad classes of neurons or glia which may have important functional significance. Here, a powerful new technique, *in situ* hybridization with a DNA preparation (cDNA) complementary to a messenger RNA (mRNA) coding for a specific functional protein, is providing valuable information unobtainable in other ways. Microinjection of a strongly fluorescent dye (e.g., procyon yellow) into a neuron often vividly reveals connexions, thus supplementing older methods based on sectioning fiber tracts and tracing their subsequent degeneration.

Electron Microscopy. In the CNS, presynaptic nerve terminals often appear under the electron microscope as club-like expansions (ca.

0.5 µm in diameter) of the fine (0.2 µm diameter) terminal axonal branches (Fig. 7 A). There is a distinct electron-lucent gap between the pre- and postsynaptic membranes, the purpose of which may be to shunt the potentials associated with the arrival of a nerve impulse and to prevent them invading the postsynaptic membrane, thus altering its excitability. This allows synaptic transmission to be purely chemical in character. In some rare synapses where there is evidence of electrotonic transmission, this "synaptic gap" does not appear and pre- and postsynaptic membranes are closely apposed.

The presynaptic terminal cytoplasm is packed with vesicular organelles of equal size (ca. 0.05 µm in diameter). These are referred to as *synaptic vesicles*. Their function is to store transmitter and to release it when the nerve terminal is invaded by a nerve impulse, by temporarily fusing with the terminal plasma membrane and opening out to discharge their contents, a process known as *exocytosis*. This release process is Ca^{2+}-dependent. Synaptic vesicles that have released their contents are recovered by the reverse process, known as *endocytosis*, pulled back into the terminal and recycled. The process is similar to and is possibly an evolutionary adaptation of the general process of cell secretion, with the important difference that whereas in normal secretion vesicles are used only once before being returned to the Golgi region of the cell body for reprocessing, synaptic vesicles are recovered, refilled, and reused.

The synthesis, storage, and release of transmitter demands considerable amounts of metabolic energy so it is not surprising that presynaptic nerve terminals contain local concentrations of the organelles responsible for cellu-

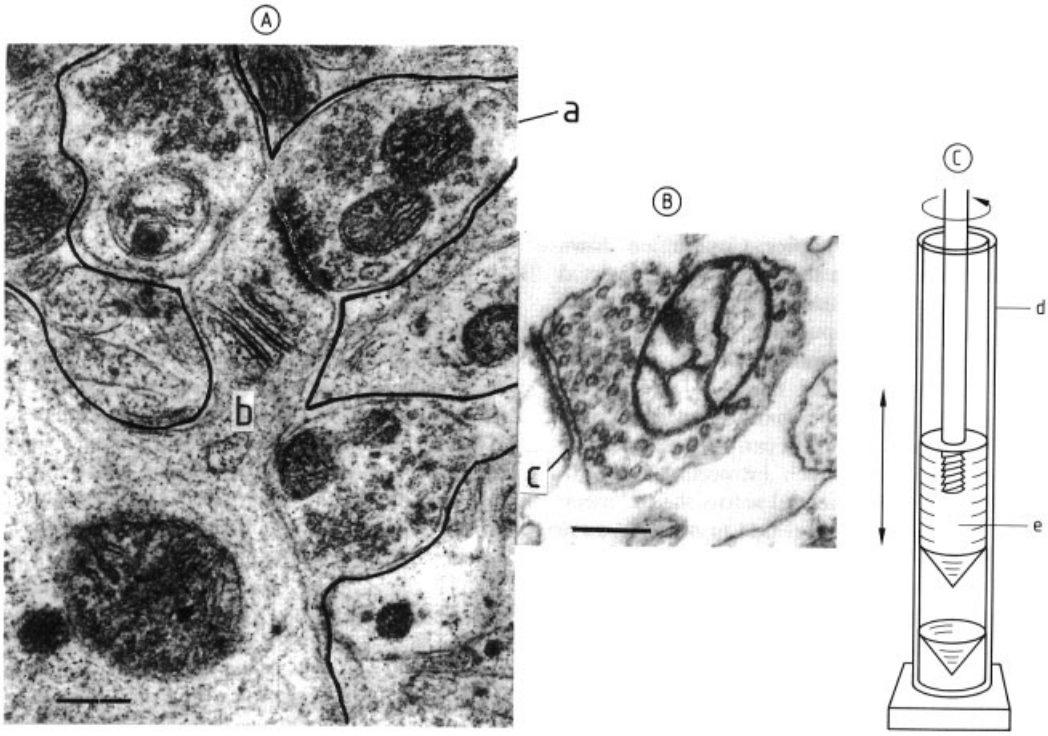

Figure 7. An axodendritic cortical synapse in situ (A) and a synaptosome derived from it (B) with the aid of an homogenizer (C)

A) Cortical synapse, three synapses are outlined: a) Presynaptic terminal bulb filled with synaptic vesicles and containing mitochondria and presynaptic dense projections; b) Dendritic spine with dense postsynaptic membrane and spine apparatus; Bar in lower left-hand corner represents 0.2 µm

B) Synaptosome: c) Postsynaptic membrane still adhering to the synaptosome; Bar in lower left-hand corner represents 0.2 µm

C) Homogenizer: d) Precision-bore glass tube (mortar); e) Rotating (900 rpm) plastic pestle, tissue suspended in iso-osmotic sucrose is forced between the pestle and mortar.

lar respiration and energy production, the *mitochondria*.

The postsynaptic region of the synapse sometimes shows specializations and provides a basis for classification. Thus, in the CNS a broad distinction can be made between *axoaxonic*, *axosomatic*, and *axodendritic* synapses according as to whether the presynaptic axonal termination is made with another nerve terminal, or the cell body or dendrites of the target cell, respectively. Axodendritic synapses are often made, not directly onto the dendrite, but onto a small projection from it called a spine (Fig. 7 A). The numerous spines on dendrites of large neurons of the CNS give them a "furry" appearance in silver staining (Fig. 1 A). There is some evidence that spines increase in size with continued use of the synapse concerned. This is one of the few morphological indications of the elusive synaptic remodeling that must accompany learning and memory.

4.2. Neurophysiology

Physiological techniques have the advantage that they can be applied to the intact animal or to surviving tissue so that the results are often easier to interpret than those obtained by other methods. Against this, physiological methods often leave us with a model of a vital process in which basic mechanisms are concealed within "black boxes".

The simplest preparations for the physiological investigation of the nervous system are single neurons or single synapses. Conduction down axons or transmission across synapses is detected by recording the transient electrical signals that accompany nerve impulses with fine microelectrodes coupled to amplifiers and oscilloscopes (Fig. 8 A). Use is often made of invertebrate preparations (e.g., the giant axon and giant synapse of the squid) that can be penetrated by the microelectrode. Such intracellular recordings, in which the potential across the cell membrane is measured, reveal that in neurons, as in other cells, a resting potential exists across the membrane whose magnitude can be inferred from the difference in K^+ concentrations (strictly, activities) between the inside and outside of the cell. Living cells generally maintain K^+ concentrations in their cytoplasm higher, and Na^+ concentrations lower than those of the extracellular medium. If the membrane is permeable solely to potassium ions, a potential E_m arises across the membrane of a magnitude close to that predicted by the Nernst equation

$$E_m = \frac{RT}{F} \ln \frac{[K_0^+]}{[K_i^+]} = 58 \log_{10} \frac{[K_0^+]}{[K_i^+]} \text{ at } 20\,^\circ\text{C}$$

where R is the gas constant, T the absolute temperature, F the Faraday constant, and $[K_0^+]$ and $[K_i^+]$ the outside and inside K^+ concentrations respectively; under normal resting conditions E_m is close to -75 mV. When a nerve impulse is generated, usually by stimulating the nerve cell electrically, an intracellular electrode registers a transient reversal of the membrane potential due to a short-lasting rise in permeability to Na^+. The Na^+ potential temporarily superimposes itself on the resting potential. Recovery ensues as the Na^+ channel closes and is assisted by a transient increase in K^+ permeability.

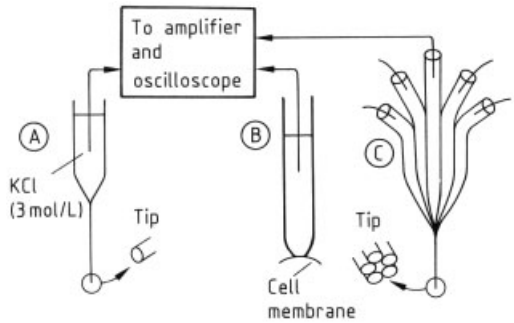

Figure 8. Types of glass electrodes
A) Simple microelectrode for intracellular recording; B) Patch-clamp electrode; C) Multibarrelled electrode (not to scale)

Permeability to ions is brought about by selective, voltage-sensitive ion channels embedded in the membrane. These are complex membrane-spanning proteins whose structure is now known as a result of their isolation and sequencing by the complementary DNA (cDNA) technique. However the fine conformational changes that accompany the opening and closing of the channels are not yet fully understood. Local anesthetics (e.g., cocaine) act by blocking influx of Na^+ (→ Local Anesthetics).

The opening and closing of single ion channels in the membrane can be detected by means

of "patch-clamp" electrodes (Fig. 8 B). These are polished glass electrodes with a small hole in the tip onto which a patch of cell membrane is drawn by suction. A good seal is essential for the detection of the very small electrical signals caused by the spontaneous opening and closing of single ion channels and requires a "clean" membrane. The method works well with certain isolated cell preparations and a large range of ion channels, some of them "ligand-gated" (i.e., opened by specific substances such as transmitters interacting chemically with the channel protein), have been detected and studied.

At the synapse, the signal is transformed from an electrical to a chemical one by the inflow of Ca^{2+} ions through voltage-sensitive calcium channels that are activated by the arrival of a nerve impulse. This causes the exocytosis of numerous vesicles and the release of transmitter, which rapidly diffuses across the 25 to 100-nm synaptic gap and triggers a response in the postsynaptic target cell. Responses are of several kinds but are easiest to detect when they involve the opening of transmitter-controlled ion channels in the postsynaptic membrane and the generation of an excitatory postsynaptic potential (EPSP). This in turn, if large enough, will generate a propagated impulse in the postsynaptic cell. In muscle cells, this propagated impulse is accompanied by muscle contraction.

Analysis of the magnitude of EPSPs reveals them to be made up of subunits or quanta, whose release is synchronized by the arrival of the nerve impulse [9]. The morphological counterparts of these quanta are the synaptic vesicles, the exocytosis of a single vesicle generating a single quantum.

An example of a more complex electrophysiological investigation is the use of the multibarrelled electrode (Fig. 8 C) which may be inserted into the CNS with its tip in proximity to a nerve cell. One "barrel" of the electrode records the response of the cell to chemical substances or drugs released from other barrels by iontophoresis. A complete pharmacological experiment can be performed in this way on a single central neuron. This technique has been important in revealing the responses of central neurons to drugs and putative transmitters.

Intracellular recordings from neurons in specialized areas of the CNS, often using indwelling electrodes in animals free to move, has led to the discovery that neurons are involved in a hierarchy of organization and has given important insights into the functioning of the CNS. This is a complex subject which can only be touched upon in this article. To give one striking example: certain neurons in the striatal vocal center of birds respond when the bird hears the call of its own species but not when it hears those of other species or sounds in general. Such cells are at the apex of a complex process of neural analysis.

Physiological methods have also revealed the existence of conditioned reflexes. The pioneer in this field was I. P. Pavlov (1849 – 1936) who noted that dogs who regularly heard a bell each time they were fed quickly salivated at the sound of the bell alone, in the absence of the sight or smell of food. This approach with its many variants and model systems, ranging from invertebrates to primates, has shed light on such important higher neural functions as learning and memory.

Perhaps the chief failure of neurophysiology is to provide a convincing theory of how the "logic units"–repetitive groups of neurons whose connections and individual properties are known in some detail– combine to achieve the pattern recognition, control of skilled movements, or other quasi-computer-controlled activities of the nervous system. This may well have to await a better understanding of the properties of networks of interacting components.

4.3. Neurobiochemistry (Neurochemistry)

The biochemical investigation of the nervous system has given rise to the subdiscipline of neurobiochemistry (neurochemistry) which includes the investigation of the nervous system by cell- and molecular-biological, and molecular genetical methods. Such methods may reveal little about the way the nervous system functions as a whole, but they can help in prying open the physiologists' "black boxes" to reveal what is going on inside. They are also of great importance in understanding long-term phenomena such as development and aging, and the interaction between the nervous system and the immune and hormonal control systems.

Such investigations often begin by the application of *tissue fractionation techniques*. These involve dispersion of the tissue and fractionation by centrifugal density gradient separation or particle exclusion chromatography. The technique was originally developed for use with tissues with a relatively homogeneous and easily disrupted cellular structure like liver, but in the early 1960s the author of this article and his colleagues pioneered their application to the much more complex mammalian brain [10]. The application of mild liquid shear in a "homogenizer" (Fig. 7 C) breaks open nerve and glial cells, fragments axons and dendrites, and "pinches off" the club-like presynaptic nerve terminals to form sealed structures called *synaptosomes* (Fig. 7 B). A combination of moving-boundary (differential) centrifugation, in which particulate structures are sedimented in progressively more intense centrifugal fields, and density gradient centrifugation, in which such structures are centrifuged until they reach a region of the gradient whose density matches their own, allows synaptosomes to be separated from other organelles and cell fragments such as nuclei, mitochondria, myelin fragments, and vesiculated membrane fragments (microsomes) derived from external and internal cell membranes.

Synaptosomes are sealed structures and conserve to a remarkable degree the properties of the synapse. Morphologically, they retain the characteristic structure of the presynaptic nerve terminal, with its small mitochondria and synaptic vesicles (Fig. 7 B). Physiologically, they are capable of respiring and generating ATP when incubated in a balanced saline medium containing an energy-yielding substrate such as glucose. Such metabolizing synaptosomes take up glucose, transmitters or their precursors by carrier-mediated mechanisms and provide a convenient in vitro system on which to test the action of drugs on such processes. Anorexiants are an example of a class of drugs which may act by controlling monoamine uptake (→ Anti-Obecity Drugs). Synaptosomes also release transmitter in a Ca^{2+}-dependent manner when depolarized with high concentrations of K^+ or Na^+ plus veratridine, a drug that depolarizes by opening up Na^+ channels. Fluorescent probes that report membrane potentials confirm that metabolizing synaptosomes possess a resting membrane potential of the same magnitude as that of intact neurons and analysis of their internal K^+ and Na^+ concentrations shows that they are able to maintain ionic gradients of the requisite magnitude.

Synaptosomes disrupted by osmotic shock release their contents: density gradient separation of these yields fractions of presynaptic terminal cytosol, synaptic vesicles, presynaptic plasma membranes and intraterminal mitochondria for further biochemical analysis. Transmitters are specifically associated with the synaptic vesicles, in the case of acetylcholine, in the amount required by the vesicle theory [11].

Milder disruptive procedures, involving teasing tissue through stainless steel or nylon sieves of decreasing mesh size, yield intact cell bodies. Astrocytes, oligodendrocytes and neurons can be isolated in this way, usually after a short further fractionation of the initial cell dispersion. If cells are prepared from embryonic or neonatal brain, they are often viable in tissue culture.

One class of neuronal cytoplasmic constituents of great importance is comprised by the mRNAs (Section 4.1) that code for the numerous proteins needed for neuronal structure and function. In general, neurons are rich in mRNA, because their cell bodies have to maintain an enormously expanded peripheral cytoplasm enclosed within the dendrites and axon (Section 5.4). A mRNA coding for a functional protein (e.g., an ion channel or receptor) may be tested for activity by transfecting a frog oocyte (egg cell) with it. The transfected oocyte is not only able to synthesize the appropriate protein, it can effect any necessary posttranslational modification and the required protein is then expresssed on the oocyte's cell membrane. The protein can be sequenced by making cDNA (Section 4.1) to the neuron's mRNA, inserting this into the DNA of the bacterium *Escherichia coli* and detecting bacterial clones bearing the appropriate cDNA sequence either by hybridization with a synthetic oligonucleotide whose sequence is based on a partial sequence of the protein in question or by immunochemical means. In the latter procedure, the cDNA is inserted into the host cell along with an "expression vector" which directs the synthesis of protein according to the code supplied by the cDNA insert.

Although cellular or subcellular preparations are preferred for neurochemical studies, a variety of more integrated preparations are available.

Tissue slices, thin enough (0.3 – 0.5 mm) to allow for adequate oxygenation and rapid penetration and diffusion of metabolites, are viable for several hours in a suitable oxygenated medium and permit metabolic processes, electrophysiological recordings, and drug responses to be studied. In the form of a *sterile explant*, tissues and cells may survive for days or weeks; neurites may grow out from the neurons present and form synapses with neighboring cells, thus providing a test system for detecting endogenous *neuronotrophic factors* released by glial or target cells that are essential for neuronal survival and differentiation.

An even higher level of organization is made experimentally accessible by *perfusion of the brain*. "Push – pull cannulae" allow perfusion of a restricted brain region by injecting fluid through the annular space between concentric cannulae and withdrawing it through the inner cannula, or perfusion of portions of the hollow spaces (ventricles) within the brain. Such cannulae have proved useful in detecting transmitter release or investigating the metabolism of transmitter precursors associated with specific brain subsystems. In a later variant (microdiffusion) the cannula ends in a semipermeable membrane.

The highest organizational level of all is reached by the introduction of metabolites or markers into the blood just before it enters the brain. Neural activity in response to stimuli causes local increases in blood circulation or metabolism which can be followed by the local accumulation of label. In an earlier form of the technique, radioactive deoxyglucose was used; the brains were fixed, sectioned, and subjected to autoradiography to determine the local accumulation of the label. In *positron-emission tomography* (PET), which can be applied to humans, the label is a short-lived positron – emitting radionuclide (^{15}O, ^{13}N, ^{11}C, or ^{18}F) generated in a nearby cyclotron. An array of radiation detectors picks up the γ-rays generated by the annihilation of the positrons by electrons in the environment, and computer tomography is used to convert the response of the array to an image of the positron distribution in the tissue in real time. Cerebral blood flow may be detected with $[^{15}O]CO_2$ or $[^{15}O]H_2O$ and the distribution of a variety of metabolites and drugs may also be studied. In spite of the huge capital and running costs of the equipment, this method is now widely used in neuropharmacological research and in clinical investigation. The dynamic character of the results obtained with living patients gives the method a unique power.

Another noninvasive technique, less well developed but of great potential, is *magnetic resonance imaging* (MRI). In its simplest form nuclear magnetic resonance (NMR) signals from hydrogen atoms are transformed by computer tomography to provide an image of the distribution of the resonating atom in any required plane. Most of the signals come from the hydrogen of water, but the image of the tissue so obtained is superior in contrast and resolving power to that obtained by X-ray computed tomography (CT); brain cancers, plaques, vascular damage and other brain abnormalities can be detected. In principle, MRI could also be used to detect the compartmentation of metabolites and the kinetics of their reactions, changes in intracellular pH and the distribution of drugs in small tissue volumes using ^{31}P, ^{13}C, or fluorine labeling.

In spite of the advantages of these new noninvasive techniques, the picture of the nervous system furnished by neurochemistry remains stubbornly molecular. It has much to tell us about development, aging and the function and modus operandi of the units which make up the system but does not reveal how it works as a whole.

4.4. Neuropharmacology

The ideas and concepts of pharmacology, especially that of specific tissue receptors for drugs and toxins, have greatly influenced the understanding of the nervous system. Transmitters and other endogenous neuroactive substances may be described as *autopharmacological agents*. Until chemical or immunochemical methods are devised they must be estimated by *bioassays*, in which the response of the assay system – often the contraction of a muscle registered by a recorder coupled to a strain gauge – is matched by a "standard". The standard may be the endogenous substance itself (if known), a stable preparation of the endogenous substance whose activity is expressed in arbitrary units, or another, known substance with pharmacologically similar or identical properties. Further characterization may involve parallel assays in which comparisons are made between unknown and

standard in two or more bioassay systems with different ligand specificities. Bioassay–although sensitive, rapid, and accurate in experienced hands and requiring inexpensive equipment–has largely been superseded by other methods such as fluorimetry, gas chromatography, high performance liquid chromatography (HPLC), or radioimmunoassay.

The existence of a drug with a powerful action on the nervous system prompts the search for a receptor. This is greatly helped by the use of radiolabeled drugs and a theoretical analysis of the binding data which leads to the calculation of affinity constants and other parameters. At this stage a crude membrane preparation may be all that is required as a source of receptor. The uncritical application of the technique has, however, led to the pejorative description "grind and bind" research.

If well-founded evidence for a specific receptor is obtained, its presence suggests the possible existence of endogenous ligands. Thus, when morphine-specific receptors were found in the brain, a search was instituted for endogenous morphine-like substances (opioids). Some 14 such substances are now known and include short-chain peptides (the *enkephalins*) and the longer chain *endorphins*, all of which contain the common sequences tyr-gly-gly-phe-leu or tyr-gly-gly-phe-met. Centrally acting analgesics and antitussives are among those pharmaceuticals interacting with the brain opioid system (→ Analgesics and Antipyretics ; → Cough Remedies).

Specificity studies of receptors for endogenous ligands, using synthetic analogues, often enable apparently identical receptors to be resolved into a number of subtypes (Chap. 5).

Drugs and toxins often have profound effects on the nervous system in extremely low doses. Thus as little as 10 µg of the hallucinogen lysergic acid diethylamide (LSD) causes disturbance of consciousness in humans, while 1 ng doses of botulinum toxin evoke toxic symptoms. Such extreme sensitivity to drugs illustrates the importance of chemical processes in the working of the nervous system.

5. Synaptic Transmission

5.1. Historical Background

The preceding chapters have shown that the synapse is a specialized region of contact and chemical secretion between the presynaptic neuron and its target cell (neuron, muscle, or gland cell), that transmitters are stored in the synaptic vesicles of the nerve terminals or varicosities, that release is brought about by vesicle exocytosis, that vesicles can be recovered and reutilized, that release is quantized, and that the response of the postsynaptic membrane to transmitter release is frequently the opening of a ligand-gated ion channel in the postsynaptic membrane induced by the transient binding of the transmitter to the channel at a specialized receptor site. Estimates of the number of different transmitters vary, but may well exceed 20.

The observation by C. Bernard (1813–1878) that the South American arrow poison, curare, blocked neuromuscular transmission without affecting the excitability of either the nerve or the muscle, showed that excitation changes its character at the junction between nerve and muscle and becomes particularly sensitive to chemical blockade. It did not however immediately lead to the concept of chemical transmission at synapses: at the turn of this century, resemblances were seen on the one hand between the physiological effects of adrenaline and stimulation of sympathetic nerves and on the other between the effects of muscarine and stimulation of the parasympathetic system. This led to the suggestion that adrenaline, or something similar, mediated sympathetic stimulation and muscarine, or an endogenous analogue, mediated parasympathetic stimulation. Another alkaloid, nicotine, was found first to stimulate and then to block synaptic transmission through autonomic ganglia and this property enabled the anatomy of the autonomic system (Section 3.1 and Fig. 4) to be worked out. U. S. von Euler (1905–1983) discovered that the transmitter at most sympathetic nerve terminals is noradrenaline rather than adrenaline and the work of O. Loewi (1873–1961) and H. H. Dale (1875–1968) showed that the transmitter at autonomic ganglia, most parasympathetic nerve endings, and at neuromuscular junctions is acetylcholine. Dale coined the convenient–if not quite accurate–terms adrenergic and cholinergic to describe the

two types of synapse and this terminology has been extended subsequently to cover other transmitters.

The fact that muscarine and nicotine do not have identical actions on cholinergic synapses, indicates that the acetylcholine receptors in the target cells have markedly different *chemical specificities* as well as differing physiological roles. This was the first example of an important development in pharmacology: the classification of receptors having a common endogenous ligand by means of agonists or antagonists of known but different structures. Thus acetylcholine receptors may be classified as muscarinic (these are found at postganglionic parasympathetic nerve terminals and in the CNS) and nicotinic (found in autonomic ganglia, at the neuromuscular junction, and also in the CNS). That there is more than one subtype of nicotinic receptor is shown by the fact that the muscular relaxant decamethonium is an excellent blocker of neuromuscular transmission, whereas hexamethonium, not very effective at this synapse, is much more so in autonomic ganglia. There are at least two subtypes of muscarinic receptor (designated M_1 and M_2) and at least four types of adrenergic receptors (α_1, α_2, β_1, and β_2). Multiple types of glutamate (probably three) and γ-aminobutyrate (at least two, $GABA_A$ and $GABA_B$) are also known.

The isolation and sequencing of receptors from different parts of the nervous system has fully confirmed the existence of subtypes. In situ hybridization shows that these subtypes are not mingled together indiscriminately but may be specific for particular classes of neuron. This opens up the possibility of very specific drug targeting if variants can be synthesized which specifically recognize the subtypes thus far known only from their molecular composition.

5.2. Survey of Transmitters

5.2.1. Low Molecular Mass Transmitters

The first transmitters to be identified were the low molecular mass compounds noradrenaline and acetylcholine.

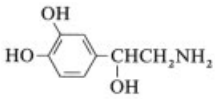

Noradrenaline
[*51-41-2*]

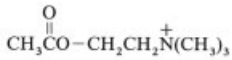

Acetylcholine
[*51-84-3*]

Others followed: dopamine, 5-hydroxytryptamine (5-HT, also known as serotonin), glutamate, γ-aminobutyrate (GABA), glycine, and, less certainly, histamine, aspartate, and taurine. Such transmitters are often nowadays referred to as "classical" transmitters, in contradistinction to the neuropeptides (Section 5.2.2).

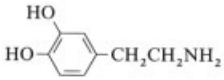

Dopamine
[*51-61-6*]

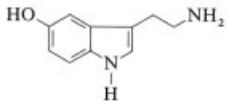

5-Hydroxytryptamine
[*50-67-9*]

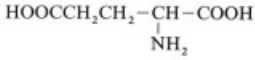

Glutamate
[*56-86-0*]

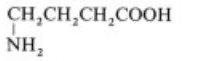

γ-Aminobutyrate (GABA)
[*56-12-2*]

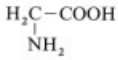

Glycine
[*56-40-6*]

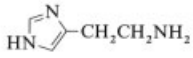

Histamine
[*51-45-6*]

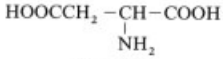

Aspartate
[*56-84-8*]

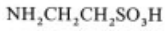

Taurine
[*107-35-7*]

The classical transmitters are usually synthesized in the nerve terminal cytoplasm and are stored in synaptic vesicles. Uptake by synaptic vesicles is carrier mediated and driven by an inwardly directed proton gradient across the vesicle membrane. Release is via Ca^{2+}-activated exocytosis of synaptic vesicles; the vesicles are recovered and refilled with the locally synthesized or recovered transmitter, a process known as vesicle recycling. Transmitter release is thus not immediately dependent on axonal transport (Section 5.4). The released transmitter is salvaged by a high affinity, carrier-mediated uptake system in the presynaptic terminal, either as such or in the form of its hydrolysis products. Thus acetylcholine is hydrolyzed to choline and acetate by the enzyme acetylcholinesterase, present in high concentration in the synaptic

cleft; both hydrolysis products are taken up and resynthesized to acetylcholine.

The hydrolysis of acetylcholine is also an inactivating mechanism, since choline has only one ten-thousandth of the activity of acetylcholine on acetylcholine receptors and acetate is not a ligand for such receptors. Inactivation mechanisms involving oxidation, methylation, and other reactions also exist for other transmitters, but these usually come into play only when the concentration of transmitter in the synaptic cleft is very high; the usual form of inactivation is the presynaptic uptake of the transmitter. This applies to the catecholamines, 5-HT, and the amino acids, but not histamine. In the case of the amino acids, glial cells may also play a part in the uptake process.

Acetylcholine and noradrenaline are transmitters at numerous peripheral and central synapses. Glutamate is an excitatory transmitter widely distributed in the CNS of mammals; in many invertebrates (arthropods, cephalopods) it replaces acetylcholine as the excitatory transmitter at the neuromuscular junction. Dopamine and 5-HT are excitatory transmitters in the mammalian central nervous system; GABA and glycine are inhibitory transmitters in the brain and spinal cord, respectively. In arthropods inhibition is brought about not centrally but by specific inhibitory endings in the muscles where GABA is a transmitter; indeed, GABA was first discovered by using the inhibitory nerve to the stretch receptor of the crayfish as an assay system.

Inhibition is a widespread and important phenomenon in the CNS. Without it, the smooth relaxation of an antagonist muscle when an agonist contracts would be impossible, and painful cramps would accompany any movement. Thoughts and behavior would become wild and disordered. Inhibitory synaptic transmission involves the release of transmitter, the transmitter opening chloride channels instead of channels for sodium and potassium; chloride ions move into and transiently *hyperpolarize* the postsynaptic neuron. Hyperpolarization raises or stabilizes the neuron's resting potential, thereby partially or completely neutralizing excitatory postsynaptic potentials and preventing them from achieving the membrane depolarization which would trigger off a propagated impulse. Epilepsy has been defined as "a disorder of brain function characterized by the episodic recurrence of paroxysmal neurological or behavioral manifestations caused by abnormal synchronous and excessive discharges of large groups of neurons" [1]. The cause is obscure but may be related to an abnormal accumulation of the excitatory transmitter glutamate, a reduced availability of the inhibitory transmitter GABA, or an excessive accumulation of extracellular K^+ which would cause massive depolarization of neurons and extensive discharges (→ Antiepileptics). Two other classes of drugs exert a general effect on neuronal excitability in the CNS, the analeptics, which stimulate (→ Analeptics) and the sedatives (→ Sedatives) which depress it. A fourth class, the neuroleptics, which include some sedatives, seem to act primarily on D_2 dopamine receptors (Section 5.3.3).

5.2.2. Neuropeptides

The discovery of neuroactive peptides in the nervous system grew out of the discovery of peptide and protein hormones (→ Peptide and Protein Hormones). The first hormone to be discovered was secretin which is released from the intestine and evokes pancreatic secretion. Many others followed; among the most important are insulin and glucagon both of which are released from the pancreas and regulate blood sugar, but in opposite directions. The intestine continued to yield hormones, among them substance P, vasoactive intestinal polypeptide (VIP) and cholecystokinin (CK).

Another potent source of physiologically active peptides is the pituitary, or hypophysis, a gland at the base of the brain. Here are found hormones controlling diuresis and water balance (vasopressin), milk ejection (oxytocin), growth (growth hormone), reproductive function (the two gonadotropins, namely luteinizing hormone and follicle-stimulating hormone), lactation (prolactin), fat mobilization (β-lipotropin), thyroid activity (thyrotropin), and adrenal cortical activity (adrenocorticotropic hormone). These in turn are controlled by peptides located in the region at the base of the brain known as the hypothalamus. Such hormones either stimulate (releasing hormones) or inhibit (releaseinhibiting hormones) the release of pituitary hormones.

Releasing hormones for the gonadotropins, thyrotropin, adrenocorticotropic hormone, and growth hormone, and a release-inhibiting hormone for growth hormone (somatostatin) have been characterized. The hypothalamus also links up with the autonomic nervous system; thus the hypothalamic – hypophyseal axis is extremely important for central nervous control of bodily function and metabolism.

Peptides are found in bewildering variety at all levels of the phylogenetic tree down to *Hydra*, a simple animal related to the sea anemone. Some peptides have definite functions, thus one peptide in *Hydra* is essential for the differentiation of the head region and has been detected in the rat hypothalamus; another, in *Aplysia* (a sea-slug) regulates egg laying; others may act as neurotoxins. Frog skin has yielded at least a dozen peptides, many resembling those found in the mammalian hypothalamus and elsewhere.

Immunological techniques have been extremely important in studying the neuropeptides. Specific antisera or monoclonal antibodies (mABs) raised to them have enabled them to be estimated in tissues and subcellular fractions by means of *radioimmunoassay*. Such sera or antibodies have also been extensively used to localize the peptides they recognize in tissues, including nervous tissue, by applying the techniques of immunocytochemistry. Usually, the specific antiserum or mAB is applied after permeabilizing tissue sections with a detergent. After washing out unbound antibody, a second antibody, labeled with a fluorescent group and directed against the immunoglobulin fraction of the animal providing the antiserum or mAB is applied. The section is washed again and the location of the original antigen is revealed by the pattern of fluorescence observed under the microscope.

Such studies have given two remarkable results: (1) peptides including many already identified as hormones are widely but not uniformly distributed in the nervous system; and (2) the neurons staining positively for peptides often utilize low molecular mass, classical transmitters. Certain peptides tend to be associated with certain classical transmitters such as GABA, acetylcholine, and noradrenaline. Thus it has been estimated that 50 % of all central cholinergic terminals also contain VIP and that 75 % of all VIP-containing terminals are also cholinergic [12]. Although peptides (neuropeptides) are present throughout the neuron, they tend to be highly concentrated in the terminals and can often only be identified in cell bodies by using colchicine to block axonal transport (Section 5.4), the process whereby substances or organelles synthesized in the neuronal cell body are conveyed to the terminal.

It thus seems reasonable to conclude that neurons primarily utilize neuropeptides as transmitters. When a classical transmitter coexists with the neuropeptide, the latter may be regarded as a cotransmitter. Neuropeptides are indeed released from nerve terminals on stimulation in a Ca^{2+}-dependent manner and can induce or modify electrical responses in postsynaptic membranes. A good example is substance P which is released on stimulation of afferent neurons entering the spinal cord of the type mediating pain. Neuropeptides differ considerably from classical transmitters in the dynamics of their synthesis, storage, and release [13, 14]. Thus they are not stored in synaptic vesicles as the classical transmitters are but in larger vesicles, which sometimes have electron-dense cores (i.e., have black-looking cores in electron micrographs). They cannot be synthesized in the nerve terminals or salvaged after release so there is no recycling vesicular pool. This means that release is completely dependent on axonal transport.

These differences in intracellular dynamics suggest that the pattern of release of neuropeptide cotransmitters reflects the past intensity and duration of stimulation of the neurons containing them. This could result in a subtle regulation of the rate of synthesis and pattern of release of the classical cotransmitter, especially if the neuropeptide combines with receptors to release "second messengers" (Section 5.3.2) able to regulate intracellular transmitter metabolism.

Neuropeptides are rarely if ever synthesized as such; they are the products of the posttranslational modification of much larger molecules, known as propeptides (prohormones) or prepropeptides (preprohormones). These contain the sequences of neuroactive peptides as segments separated by lysine – lysine or lysine – arginine, points at which the propeptide is easily cleaved by proteolytic enzymes. Thus the hypophyseal hormones adrenocorticotropic hormone, γ-melanotropin and β-lipotropin are synthesized as parts of a single much larger prohormone; adrenocorticotropic hormone itself con-

tains the sequence of α-melanotropin while β-lipotropin breaks down to yield the opioid peptide β-endorphin and γ-lipotropin; the latter in turn yields β-melanotropin.

The discovery of the neuropeptides has opened up a baffling yet exciting new era in neurobiological research. In many cases the function of these peptides in the brain is not known; some may act as transmitters, but others may have a modulating role or link the nervous system with the autonomic system, the immune system or tissue repair after injury. The implications for neuropharmacology are also immense, if stable analogues with antagonistic or agonist properties can be successfully prepared.

5.2.3. Evolutionary Significance of the Two Types of Transmitter

The low molecular mass transmitters are not found in the simplest living organisms that possess nervous systems; thus acetylcholine is not found in coelenterates and first appears in the echinoderms and platyhelminths. Their appearance may coincide with an evolutionary demand for fast, sustained movement. It is probable that neuropeptides were the first transmitters and that chemical transmission is an adaptation of humoral secretion. Neuropeptide secretion is dependent on de novo synthesis and axonal transport, these processes are wasteful and relatively inefficient especially as the size of organisms increased and axons grew longer, separating the sites of synthesis and release by ever greater distances.

Chemical transmission, especially when speeded up by recycling, remained however an attractive mechanism. The store of transmitter in the presynaptic terminal guarantees unidirectional synaptic transmission–an essential prerequisite for the orderly processing of information. If an antidromic impulse (i.e., one travelling in the reverse direction) should arise in an axon, it would be unable to effect reversed transmission due to lack of a store of transmitter in the neuronal cell body or dendrites.

The existence of more than one transmitter and specific receptors for them may be a device to suppress "cross-talk" between interdigitated but functionally distinct systems. The retention of neuropeptides as transmitters or cotransmitters alongside the classical transmitters is harder to explain. Some may represent the vestigial survival of phylogenetically old transmitters which evolutionary pressures were not strong enough to eliminate. Others clearly retain a transmitter function in circuits where short synaptic delays and continuous action may not be so important, as in pain afferents. Yet others have a regulating role or have important hormonal functions linking the nervous system to other body systems.

All chemical synapses– even those using fast-tacting, easily mobilized transmitters–impose a short delay of not less than 0.5 – 1 msec. It is significant that where fast action has evolutionary survival value (e.g., the escape reactions of decapods), chemical transmission has been abandoned for the even faster electrical (electrotonic) transmission in the corresponding neuronal circuits. Electrical transmission is also less temperature dependent than chemical transmission, which may explain why electrotonic synapses are more frequently encountered in cold-blooded animals. Such synapses show apposition of pre- and postsynaptic membranes and are devoid of synaptic vesicles.

5.3. Transmitter Action

5.3.1. Fast Ionotropic Action

The term ionotropic action denotes the action of transmitters which operate by opening channels in the postsynaptic membrane. Such action is swift and terminated by spontaneous closing of the channel and the inactivation of the transmitter by hydrolysis or reuptake.

Excitatory Ionotropic Action . Both acetylcholine (when acting on nicotinic receptors) and glutamate are examples of fast-acting excitatory transmitters. The nicotinic acetylcholine receptor (nAChR) was first isolated in a pure form from an exceptionally rich source: the electric organ of the electric ray, *Torpedo*. Such organs consist of a honeycomb array of columns of flattened cells, called electrocytes, which are embryologically derived from muscle with loss of the contractile machinery; each electrocyte is 80 % covered on its under surface with cholinergic nerve terminals. Electric organ contains

weight for weight, 500 – 1000 times more synaptic material and nAChR than muscle and isolation of the latter becomes a feasible proposition.

Isolation of the receptor was made possible by the discovery that one of the main proteins of cobra venom is a powerful neuromuscular blocking agent and has a high affinity for the receptor. The receptor was solubilized in detergent and adsorbed onto columns of immobilized venom; after washing the column, the receptor could be eluted in concentrated form. It consists of four peptides of molecular masses 40 (α), 50 (β), 60 (γ), and 65 (δ) kDa; the complete receptor is a complex of all four in which the α-peptide is repeated twice. In negative staining the complex appears as a pentagon with a central hole presumed to be the ion channel.

All four peptides have been sequenced [15]: each contains four hydrophobic transmembrane segments and the complete molecule is an assembly of 20 such segments forming an ionic channel through the membrane. When two molecules of acetylcholine bind to the two α-subunits an allosteric change takes place and the channel opens. The channel opening time varies with the ligand. Each package (quantum) of transmitter opens 5000 – 6000 such channels and generates a transient change of about 0.2 mV in the resting membrane potential of the postsynaptic membrane. The *Torpedo* electrocyte is not electrically excitable and cannot conduct an impulse but in an excitable nerve or muscle cell, the synchronized release of several hundred such quanta will generate a postsynaptic potential change sufficient to initiate a conducted response. Reversible neuromuscular blocking agents (muscular relaxants) usually contain two quaternary nitrogen atoms separated by a bridge of carbon atoms instead of one as in acetylcholine. Such compounds block by straddling the two α-subunits of the receptor. One such muscular relaxant, succinylcholine, is readily hydrolyzed and provides a transient, easily controlled block very useful in abdominal surgery. By contrast the snake toxins work by blocking the ion channel directly.

The characterization of the nAChR has been followed by that of the glutamate receptor, which operates on similar lines.

Inhibitory Ionotropic Action. The nature of spinal inhibition was worked out by J.C. Eccles (b. 1903) using microelectrodes inserted into motor neurons and the reflex activation of inhibitory interneurons (Section 3.2.1) [2]. His observation that minute amounts of a chemical substance, strychnine, could block inhibition convinced him that an inhibitory chemical transmitter, subsequently identified as glycine, must be involved. To identify the ligand-gated ionic channel involved, different ions were injected into motor neurons and the effect on the inhibitory (hyperpolarizing) postsynaptic potential (IPSP) generated by the reflexly activated interneuron was observed. The IPSP could be abolished and then converted to a depolarizing potential by increasing the internal Cl^- concentration of the cell, showing that the ionic channel opened by the transmitter is a Cl^- channel. Hydrated univalent anions smaller or only a little larger than Cl^- also reduce the IPSP and are thus also able to pass through the Cl^- channel; other ions (e.g. organic anions, di- or trivalent inorganic anions) are ineffective.

The high affinity of strychnine for the glycine receptor has enabled the latter to be isolated and sequenced using affinity chromatography on immobilized strychnine.

In the brain and the dorsal horn of the spinal cord, GABA (not glycine) is the inhibitory transmitter. It acts in a similar way to glycine but its receptor has a different pharmacological specificity: it is blocked by picrotoxin and bicuculline (ineffective at glycine receptors) but not strychnine. Presynaptic inhibition (i.e., inhibition exerted on excitatory presynaptic nerve terminals by inhibitory synapses present on them) is also mediated by this transmitter.

5.3.2. Slow Metabolotropic Action

Acetylcholine at muscarinic receptor sites, the monoamine transmitters (noradrenaline, dopamine, 5-HT, histamine), and probably some neuropeptides do not operate directly through ligand-gated ion channels but via *second messengers*. These messengers are intracellular components capable of regulating a variety of intracellular functions (including that of ion channels) which in turn are mobilized by binding the transmitter (or an agonist analogue) to its receptor. An intricate series of reactions has been implicated (Fig. 9). The basic features are:

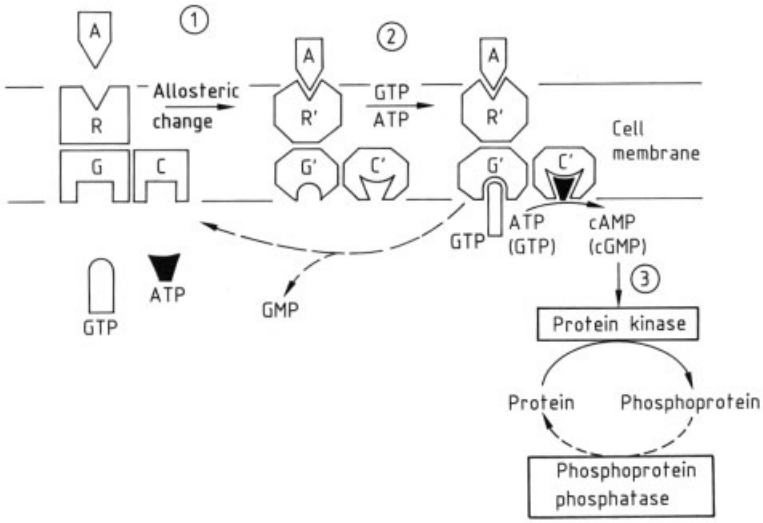

Figure 9. Scheme showing sequence of events regulated by metabolotropic receptor (R)
Allosteric changes are shown schematically by different geometric figures.
1) Agonist A combines with complex of R, G-protein (G) and adenylate (guanidylate) cyclase (C).
2) The ligand-induced conformational change in R communicates itself to G which binds GTP and through G to C which converts ATP (GTP) to cyclic AMP (cyclic GMP).
3) Cyclic AMP (cyclic GMP) activates a protein kinase which regulates functional proteins by phosphorylating them.

1) the identification of a family of G (or N) proteins which link the receptor to adenylate (or guanidylate) cyclase,
2) the regulation of protein kinases by the cyclic adenosine 3′,5′-monophosphate (cAMP) or cyclic guanosine 3′,5′-monophosphate (cGMP)
3) the regulation of functional proteins by phosphorylation catalyzed by such kinases.

Most receptor-regulated cyclases are adenylate cyclases but those regulated by H_1 (histamine receptor) and mAChR (acetylcholine receptor) are guanidylate cyclases. In some cases the functional proteins so regulated are ion channels. Thus an important effect of the activation of α_2 receptors by noradrenaline is a long-lasting inhibition achieved by an increase in K^+ conductance following channel phosphorylation. In other cases, proteins controlling gene expression may be so regulated. A phospholipase C capable of generating triphosphoinositide and diacylglycerol from a membrane phospholipid is yet another protein to be thus regulated. Both of these substances can release Ca^{2+} from intracellular membranes. The Ca^{2+} ions can react with another protein calmodulin to activate another functionally important class of protein kinases. Calcium antagonists are used therapeutically in coronary heart disease, arrhythmia, and hypertension (→ Calcium Antagonists). However, in many cases the link between protein phosphokinase activity induced by the second messenger–cAMP, cGMP–or Ca^{2+} and the physiological response has yet to be worked out.

The involvement of so many allosteric and enzymic steps may explain the relatively slow onset and long duration of such responses. This justifies the classification of these transmitter actions as metabolotropic.

5.3.3. Receptors of Importance in Drug Action

Drugs may act on receptors as *agonists* (mimetics) that produce the same action as the transmitter, or as *antagonists* (lytics, blockers, blocking agents). Several examples have already been given, notably the neuromuscular blocking agents (→ Skeletal Muscle Relaxants) which act by blocking nicotinic acetylcholine receptors in muscle subtypes (Section 5.3.1). The receptors for glutamate, GABA, noradrenaline, and dopamine are also important targets for drug ac-

tion and pharmacological analysis has revealed many subtypes.

Many clinically important drugs are used to regulate the autonomic nervous system by acting as either agonists or antagonists at the acetylcholine or noradrenaline receptors in smooth and cardiac muscle or glands. The reader is referred to other articles, for example → Antiarrhythmic Drugs, → Antiasthmatic Agents, → Antihypotensives, → Antihypertensives, → Cardiac Glycosides and Synthetic Cardiotonic Drugs, → Coronary Therapeutics, → Parasympatholytics and Parasympathomimetics, → Spasmolytics, → Sympatholytics and Sympathomimetics. One of the most widely used class of such drugs are the adrenergic β-blockers (→ β-Receptor Blocking Agents) such as propanolol which reduce blood pressure (→ Antihypertensives, → Coronary Therapeutics) and heart rate by blocking adrenergic vasoconstrictor and cardioaccelerator receptors. β-Blockers do however gain entrance to the CNS and produce depression apparently by blocking adrenergic pathways in the limbic system. By contrast, iproniazid and other monamine oxidase inhibitors, raise the intracellular concentration of noradrenaline and are antidepressive. This strongly suggests that depression involves a disturbance in CNS (possibly limbic) function mediated by adrenergic neurons, which could result either from a failure to synthesize enough noradrenaline or from a desensitization of noradrenergic receptors.

Dopaminergic transmission though not widespread is also of critical importance in the CNS. Dopaminergic cell bodies are present in the substantia nigra, a nucleus of dark hue between the hypothalamus and thalamus (Fig. 5 C) and project to the caudate nucleus (Fig. 5 B). In Parkinson's disease these neurons lose their ability to produce enough dopamine. Administration of a precursor, L-dopa, or of dopamine agonists is therapeutic (→ Parkinsonism Treatment). Conversely, agents which displace dopamine, e.g., the neurotoxin *N*-methylphenyltetrahydropyridine, induce Parkinsonian symptoms.

Dopamine pathways appear to be affected in schizophrenia. All three main classes of neuroleptic (antipsychotic) drugs, phenothiazines, butyrophenones, and thioxanthenes (→ Phenothiazine and Derivatives, → Psychopharmacological Agents), have side effects indicating the involvement of extrapyramidal neurons: Parkinsonian rigidity and tremor or dyskinesia (convulsive movements). D_2 receptor blockade correlates well with neuroleptic activity and one of the symptoms of schizophrenia, catatonia, may be related to Parkinson rigidity.

A third type of receptor of great importance for drug action is the $GABA_A$ receptor which is the main target for the anxiolytic and anticonvulsant benzodiazepines. These compounds bind to a receptor which allosterically modifies the GABA receptor to increase the efficacy of its activator GABA, an inhibitory transmitter. There is an excellent correlation between binding affinity and potency in animal and clinical tests.

5.4. Axonal Transport

The presynaptic nerve terminal cannot synthesize structural and functional proteins and lipoprotein membranes; it is entirely dependent on the synthetic activity of the cell body (perikaryon). The electromotor cell body has been estimated to maintain over 30 times its volume of presynaptic nerve terminal; similar figures must apply to other neurons. Most neuronal perikarya show corresponding signs of intense synthetic activity: they are rich in polysomes (collections of ribosomes synthesizing proteins from mRNA) and membranes, the latter often budding off synaptic vesicle-like structures.

The process whereby soluble proteins and membrane-bound organelles are conveyed from the cell body to the terminals is referred to as axonal transport. It involves more than one mechanism and rate of transport. Paradoxically, synaptic vesicles are conveyed at the fastest rate (400 mm/d), probably with the participation of microtubules (a kind of scaffolding) using a mechanism akin to muscular contraction. Larger particles (e.g., mitochondria) travel more slowly, apparently as a result of increased mechanical resistance, but soluble cytoplasmic proteins travel at the slowest rate (ca. 1 mm/d). Here the mechanism of transport is uncertain but may involve the streaming of cytoplasm within spaces between bundles of microtubules. Colchicine stops transport by causing the disassembly of microtubules.

There is also a need for retrograde transport–the conveyance of spent or worn-out organelles to the cell body for salvage–but the mechanisms involved are still unclear.

6. References

General References

1. G. Adelman (ed.): *Encyclopedia of Neuroscience,* Birkhäuser, Boston – Basel – Stuttgart 1987.
2. P. L. McGeer, J. C. Eccles, E. G. McGeer: *Molecular Biology of the Mammalian Brain,* Plenum Press, New York – London 1978.
3. F. Hucho: *Neurochemistry, Fundamentals and Concepts,* VCH Verlagsgesellschaft, Weinheim 1986.
4. R. Nieuwenhuys: *Chemoarchitecture of the Brain,* Springer Verlag, Berlin 1985.
5. J. R. Cooper, E. F. Bloom, R. H. Roth: *The Biochemical Basis of Neuropharmacology,* Oxford University Press, New York 1986.
6. S. Iversen, L. L. Iversen: *Behavioral Pharmacology,* Oxford University Press, Oxford 1981.
7. E. R. Kandel, J. H. Schwartz: *Principles of Neural Science,* 2nd ed., Elsevier, New York – Amsterdam – Oxford 1985.

Specific References

8. C. S. Sherrington: *Integrative Action of the Nervous System,* University Press, Cambridge 1948.
9. B. Katz: *Nerve, Muscle, and Synapse,* McGraw Hill, New York 1966.
10. V. P. Whittaker in A. Lajtha (ed.): *Handbook of Neurochemistry,* 2nd ed., **vol. 7,** Plenum Press, New York – London 1984, p. 1.
11. V. P. Whittaker in [10]p. 41.
12. D. V. Agoston, E. Borroni, P. J. Richardson, *J. Neurochem.* **50** (1988) 1659-1662.
13. D. V. Agoston, J. M. Conlon, V. P. Whittaker, *Exp. Brain Res.* **72** (1988) 535-542.
14. D. V. Agoston, V. P. Whittaker, *J. Neurochem.* **52** (1989) 1474-1480.
15. M. Noda et al., *Nature (London)* **302** (1983) 528-532.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Nickel

Standard Article

Derek G. E. Kerfoot[1]
[1]Sherritt Gordon Limited, Fort Saskatchewan, Alberta, Canada

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a17_157
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. History
1.2. Physical Properties
1.3. Chemical Properties
2. Occurrence
2.1. Nickel Minerals
2.2. Sulfide Ores
2.3. Oxide Ores
2.4. Economic Trends
3. Beneficiation of Nickel Sulfide Ores
4. Pyrometallurgy of Nickel Concentrates
4.1. Roasting
4.2. The Smelting Process
4.2.1. Primary Smelting
4.2.2. Matte Converting
4.3. Environmental Aspects of Nickel Smelting
4.4. Treatment of Converter Matte
4.4.1. Inco Matte Separation Process
4.4.2. Fluidized-Bed Roasting of Nickel Sulfide
5. Hydrometallurgy of Nickel Concentrates and Mattes
5.1. Ammonia Pressure Leaching
5.2. Atmospheric Acid Leaching
5.3. Acid Pressure Leaching
5.4. Chloride Leach Processes
5.5. Treatment of Nickeliferous Pyrrhotite
6. Nickel Refining
6.1. Electrorefining
6.1.1. Refining of Nickel Metal Anodes
6.1.2. Refining of Nickel Matte Anodes
6.1.3. Solution Purification
6.1.4. Electrorefining Operations
6.2. Electrowinning
6.2.1. Electrowinning from Sulfate Electrolytes
6.2.2. Electrowinning from Chloride Electrolytes
6.3. Carbonyl Refining
6.3.1. Atmospheric Pressure Carbonyl Process
6.3.2. High-Pressure Carbonyl Process (BASF)
6.3.3. Inco Pressure Carbonyl Process

6.4. Hydrogen Reduction to Nickel Powder
7. Beneficiation of Oxide Ores
7.1. The Nippon Yakin Oheyama Process
7.2. Segregation Processes
8. Pyrometallurgy of Oxide Ores
8.1. Ore Pretreatment
8.2. Smelting to Ferronickel
8.2.1. The Rotary Kiln - Electric Furnace Process
8.2.2. The Ugine Ferronickel Process
8.2.3. The Falcondo Ferronickel Process
8.2.4. Refining of Ferronickel
8.3. Smelting to Nickel Matte
8.3.1. Blast Furnace Smelting
8.3.2. Matte Production from Ferronickel
8.3.3. The Inco Selective Reduction Smelting Process
8.4. Energy Consumption in Laterite Smelting
9. Hydrometallurgy of Oxide Ores
9.1. The Caron Process
9.2. Pressure Leaching with Sulfuric Acid
9.2.1. The Moa Bay Process
9.2.2. The Amax Acid Leach Process
10. The Recovery of Byproduct Cobalt
11. Nickel Recovery from Secondary Sources
12. Market Products
13. Uses
14. Economic Aspects
15. Toxicology

[Top of Page]

1. Introduction

Nickel, [7440-02-0], Ni, A_r 28, atomic mass 58.71, is a silver white metal with typical metallic properties. It occurs in the periodic table in the first triad of Group VIII (IUPAC Group 10), after iron and cobalt, to which it is closely related. The electronic configuration of nickel is $1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 3d^8\ 4s^2$.

Nickel occurs naturally as a mixture of five stable isotopes of mass number 58 (67.84 %), 60 (26.23 %), 61 (1.19 %), 62 (3.66 %), and 64 (1.08 %). So far seven artificial radioactive isotopes have been identified with mass numbers of 56, 57, 59, 63, 65, 66, and 67.

The great importance of nickel lies in its ability, when alloyed with other elements, to increase a metal's strength, toughness, and corrosion resistance over a wide temperature range. Nickel is essential to the iron and steel industry, and nickel-containing alloys have played a key role in the development of materials for the aerospace industry. World production of nickel increased from 20×10^3 t/a in 1920 to 750×10^3 t/a in 1976 at an average annual rate of increase of 3 %. However, no significant increase in either consumption or production occurred between 1976 and 1986, with world nickel consumption falling as low as 650 000 t/a in 1981 and 1982. The known world reserves of nickel are more than adequate to maintain this level of production for many years.

1.1. History [1-4]

It is believed that the word “Nickel” or “Kupfer Nickel” is derived from a derogatory term applied by medieval miners in Saxony to a mineral which they mistakenly believed to be a copper ore, but from which they could extract no copper. It was in fact the nickel arsenide ore niccolite (NiAs). The Saxon miners not unnaturally assumed that evil spirits, or “Old Nick” himself, were responsible for their problems.

Nickel itself was first isolated by the Swedish mineralogist AXEL CRONSTEDT in 1751, when studying agersdorffite (NiAsS) ore from Los in Sweden. Nickel's status as an independent element was proved in 1775 by BERGMANN and coworkers, but it was not until 1804 that RICHTER produced a relatively pure sample of the metal and described its essential properties.

Nickel-containing alloys were in use long before CRONSTEDT's discovery. For centuries the Chinese had been making “pai thung” or “white copper” (40 % Cu, 32 % Ni, 25 % Zn, and 3 % Fe) which resembled silver in appearance. This material appeared in Europe in small quantities in the late 1700s. Since it cost only one quarter the price of silver, the alloy had obvious commercial potential as a silver substitute. By the 1830s, copper – nickel – zinc alloys, by then known as German Silver and later as nickel-silver, were being produced commercially in both Germany and England in significant amounts. In addition to its silver color, this alloy was easy to cast and fabricate, was resistant to tarnishing, and was economical to produce. The next significant development occurred in 1857 when the United States issued a cupronickel coin containing

12 % nickel; other countries soon followed suit.

In the mid-1800s nickel was produced in small quantities from sulfide ores mined in Germany, Norway, Sweden, and Russia. Although refined metallic nickel was first produced in Germany in 1838, world production of nickel remained below 1000 t/a until 1876.

The demand for nickel increased abruptly in 1870 – 1880 when PARKES, MARBEAU, and RILEY demonstrated its use in alloy steels, as FLEITMANN succeeded in making malleable nickel, and as electrolytic nickel plating was successfully developed. The first nickel-steel armor plate was made in France in 1885 and soon afterwards in Italy, England, and the United States. In the early 1890s the adoption of nickel steels by the world's navies led to a steep rise in the demand for nickel.

In 1863, GARNIER discovered the nickel oxide ores of New Caledonia, and from 1875 this French island became the world's main producer of nickel, which position it retained until displaced by Canada in 1905. Mining of the sulfide ores of the Sudbury Basin in Ontario, Canada, started in 1886, and these ore bodies were to remain the most important source of the world's nickel supply for most of the twentieth century. Even in 1950 the Sudbury area still supplied 95 % of the nickel for the Western world. Since then Canada's preeminent position has been largely eroded, and by 1980 nickel was being mined, smelted, and refined in more than 20 countries.

Up to the end of World War I nickel was used almost exclusively for military purposes, but intensive research, between the world wars, into potential industrial uses soon led to a variety of new applications. In the 1990s there are thousands of nickel alloys in use with nickel contents ranging from 99 % in wrought nickel to as little as 1 % as hardening additives for various alloys.

1.2. Physical Properties [9-11]

The physical properties of nickel depend largely on its purity, on the physical state of the metal, and on its previous treatment. The wrought metal is highly malleable, moderately strong, tough, ductile, and highly resistant to corrosion in many media. Its good strength and ductility persist to subzero temperatures. Some important physical properties of nickel are as follows:

mp	1455 °C (1728 K)
bp	2730 °C (3003 K)
Relative density (25 °C)	8.9
Volume increase on melting	4.5 %
Heat of fusion (*mp*)	302 J/g
Heat of sublimation (25 °C)	7317 J/g
Heat of vaporization (T_{crit})	6375 J/g
Standard entropy	29.81 J/K
Thermal conductivity (0 – 100 °C)	88.5 W m^{-1} K^{-1}
Heat capacity (0 – 100 °C)	0.452 J g^{-1} K^{-1}
Electrical resistivity (20 °C)	6.9 μΩ/cm
Temperature coefficient of electrical resistivity (0 – 100 °C)	6.8×10^{-3} K^{-1}
Thermal expansion coefficient (0 – 100 °C)	13.3×10^{-6} K^{-1}
Modulus of elasticity	199.5 GPa
Brinell hardness	85

The density of molten nickel in g/cm^3 is given by

$$\varrho = \varrho_0 - 1.16\times10^{-3}(t - t_0)$$

where

$$\varrho_0 = 7.905\,,\ t_0 = 1455;$$

the surface tension in mN/m, by

$$\sigma = \sigma_0 - 0.38(t - t_0)$$

where

$$\sigma_0 = 1778\,,\ t_0 = 1455;$$

and the viscosity in mPa · s, by

$$\eta = \eta_0 \exp(E/RT)$$

where

$$\eta_0 = 0.1663\,,\ E = 855\ \mathrm{J/g}.$$

Some mechanical properties of nickel are listed in Table 1.

Table 1. Mechanical properties of nickel *

Property	Pretreatment							
	Tempered			Hot-rolled				
Temperature, °C	–200	–75	20	20	200	400	600	800
0.2 % proof stress, MPa	230	170	160	170	150	140	110	
Tensile strength, MPa	710	560	500	490	540	540	250	170
Elongation, %	54	58	48	50	50	50	60	60
Impact value (Charpy)	230	230	230					

* Nominally 99.0 % Ni.

Nickel normally forms a face-centered cubic crystal lattice, although a hexagonal modification is known, which changes to the cubic form at 250 °C. Cubic nickel is ferromagnetic below its Curie point of 357 °C, although less so than iron. The hexagonal form of nickel is not ferromagnetic.

1.3. Chemical Properties [12]

In its chemical properties nickel resembles iron and cobalt, as well as copper. The + 2 oxidation state is much the most important in nickel chemistry, although the + 3 and + 4 oxidation states are also known. The oxidation of nickel(II) salt solutions with chlorine, bromine, or persulfate in aqueous alkaline solution produces the insoluble nickel(III) oxide, β-NiO(OH). The nickel(IV) oxide $NiO_2 \cdot n\,H_2O$ is prepared by persulfate oxidation of nickel(II) hydroxide. In contrast to cobalt and iron, nickel is normally only stable in aqueous solution in the + 2 oxidation state.

Nickel(II) ions form complexes very readily. The maximum coordination number is six. The green color of most hydrated nickel(II) salts and their aqueous solutions is due to the octahedral $Ni(H_2O)_6^{2+}$ cation. The strong tendency of the nickel(II) ion to form complexes with ammonia, for example $Ni(NH_3)_6^{2+}$, or $Ni(H_2O)_2(NH_3)_4^{2+}$, is utilized in a number of extraction processes (Sections Ammonia Pressure Leaching and The Caron Process).

The most unusual property of nickel is its ability to react directly with carbon monoxide to form a binary carbonyl complex. When carbon monoxide reacts with metallic nickel at 60 °C it forms the volatile nickel tetracarbonyl, $Ni(CO)_4$ (see ⟶ Nickel Compounds). This reaction is reversible, with the carbonyl compound decomposing to carbon monoxide and nickel at higher temperatures (180 °C).

$$\mathrm{Ni} + 4\,\mathrm{CO} \underset{180\,^\circ\mathrm{C}}{\overset{60\,^\circ\mathrm{C}}{\rightleftharpoons}} \mathrm{Ni(CO)_4}$$

This reaction is the basis for the vapometallurgical refining of nickel (Section Carbonyl Refining). No other metal forms similar carbonyl compounds under such mild conditions at atmospheric pressure.

At moderate temperatures nickel has a high corrosion resistance against air, seawater, and nonoxidizing acids. Nickel is fairly electropositive with a standard electrode potential of – 0.25 V, which is similar to cobalt (– 0.28 V).

$$\mathrm{Ni} \rightarrow \mathrm{Ni^{2+}} + 2\,\mathrm{e^-}\,(E = -0.25\ \mathrm{V})$$

An outstanding property of nickel is its corrosion resistance to alkalis. The metal is therefore frequently used in the production and handling of caustic soda. In contrast, nickel is attacked by aqueous ammonia solutions.

Nickel absorbs hydrogen, especially when finely divided; hydrogen absorption increases with increasing temperature. Defined hydride compounds have not been characterized. Nitrogen is not absorbed by nickel nor does it combine directly with it, although there is evidence for a nitride Ni_3N.

[Top of Page]

2. Occurrence

Nickel ranks twenty-fourth in the order of abundance of elements in the earth's crust with an estimated concentration of 0.008 %. Although nickel is more abundant than copper, zinc, and lead there are relatively few nickel ore bodies of

commercial significance.

The economically important ores can be divided into two types, sulfide and oxide (or silicate). Nickel does not occur as native metal. World reserves of nickel (1988) were estimated at about 50×10^6 t by the U.S. Bureau of Mines. This estimate includes only ores which are economically recoverable at prevailing metal prices. Identified world resources in deposits containing 1 % Ni or more amount to over 130×10^6 t of nickel. About 80 % of the nickel in these resources is in laterite ore bodies, with only 20 % in sulfide deposits. The distribution of high – grade nickel reserves are detailed by country in Table 2, together with the corresponding 1988 mine production rates [13].

Table 2. World reserves of nickel [13]

Country	Reserves, 10^3 t	1988 Mine production, 10^3 t
Albania	180	10
Australia	1 270	74
Botswana	354	16
Brazil	666	15
Canada	8130	190
China	726	25
Colombia	558	21
Cuba	18 140	34
Dominican Republic	517	23
Finland	80	11
Greece	454	12
Indonesia	3 200	58
New Caledonia	4 535	56
Philippines	408	8
Rep. of South Africa	2 540	34
Former Soviet Union	6 620	186
Former Yugoslavia	158	4
Zimbabwe	77	11
World total	48 980	788

Nickel-containing manganese nodules, which occur in vast quantities on the seabed at depths between 3500 and 4000 m, typically contain 0.5 – 1.5 % Ni. While not currently economically viable as a source of nickel, these deposits will undoubtedly become an increasingly important nickel resource in the future.

The United States has significant quantities of low-grade nickel resources such as the laterite ores of Oregon and northern California. However, none of these ores was considered as an economic nickel reserve by the Bureau of Mines at 1988 metal prices, and consequently they are omitted from Table 2.

2.1. Nickel Minerals

The most important nickel-containing minerals occurring in nickel deposits are listed in Table 3, together with their chemical compositions [1]. Some are relatively uncommon, and only pentlandite, garnierite, and nickeliferous limonite are of economic significance. The term garnierite is commonly used as a generic name for a series of mixed nickeliferous silicates with a wide range of nickel contents, and can include colloidal mixtures of silica and nickel hydroxide. Nickeliferous limonite is the term used to describe poorly crystalline nickel-bearing ferric oxides of which the main constituent is goethite (α-FeO · OH).

Table 3. Principal nickel minerals [1]

Mineral	Ideal formula	Nickel content, %
Sulfides		
Pentlandite	$(Ni,Fe)_9S_8$	34.22
Millerite	NiS	64.67
Heazlewoodite	Ni_3S_2	73.30
Polydymite	Ni_3S_4	57.86

Siegenite	$(Co,Ni)_3S_4$	28.89
Violarite	Ni_2FeS_4	38.94
Arsenides		
Niccolite	NiAs	43.92
Rammelsbergite	$NiAs_2$	28.15
Gersdorffite	NiAsS	35.42
Antimonides		
Breithauptite	NiSb	32.53
Silicates and oxides		
Garnierite	$(Ni,Mg)_6Si_4O_{10}(OH)_8$	≤ 47 %
Nickeliferous limonite	$(Fe,Ni)O(OH) \cdot nH_2O$	low

Pyrrhotite ($Fe_{n-1}S_n$) is not included in the list of nickel minerals because it is an iron sulfide, and nickel is not essential to its composition. In pyrrhotites from different ore bodies the composition varies from FeS to Fe_7S_8. Small amounts of nickel may substitute for the iron, making some pyrrhotite nickeliferous, independently of any particles of pentlandite they may contain.

2.2. Sulfide Ores

Sulfide nickel ores consist principally of nickeliferous pyrrhotite (Fe_7S_8), pentlandite $(Ni, Fe)_9S_8$, and chalcopyrite ($CuFeS_2$). Other minerals which occur in small but significant amounts include magnetite (Fe_3O_4), ilmenite ($FeTiO_3$), pyrite (FeS_2), cubanite ($CuFe_2S_3$), and violarite (Ni_2FeS_4). Sulfide ores typically grade 0.4 – 2.0 % nickel, 0.2 – 2.0 % copper, 10 –30 % iron, and 5 – 20 % sulfur. The balance consists of silica, magnesia, alumina, and calcium oxide.

Pentlandite, the commonest of the sulfide minerals, probably accounts for 60 % of world nickel production. Nickeliferous pyrrhotite is normally the most abundant phase in a nickel ore. It contains nickel in solid solution (0.2 – 0.5 % Ni) in addition to very finely divided pentlandite inclusions. Chalcopyrite is the most important copper mineral, although cubanite is often present in lesser amounts.

Although the major minerals are the same in all large sulfide ore bodies, the relative amounts of pyrrhotite, pentlandite, and chalcopyrite vary widely. The relative amounts of the three minerals largely determine the choice of processing route.

In addition to nickel production, a large number of valuable byproducts can be recovered from sulfide ores, including copper, cobalt, the platinum group metals, gold, silver, selenium, tellurium, sulfur, and iron. Cobalt occurs principally in solid solution in pentlandite, at concentrations ranging from 1 to 5 % of the nickel content. Discrete particles of gold, silver, and platinum group metals are very rare, and these metals are normally found in solid solution in the sulfide phases. The treatment processes are nornally designed to yield optimum overall recovery of metal values.

Major sulfide ore bodies occur in Canada, the former Soviet Union, the Republic of South Africa, Australia, Zimbabwe, and Finland.

2.3. Oxide Ores

The oxidic ores of nickel are formed by a chemical concentration process that occurs as a result of the lateritic weathering of peridotite rock. Peridotite consists mainly of olivine, a magnesium iron silicate containing up to 0.3 % nickel. In many rocks the peridotite has been altered to serpentine, a hydrated magnesium silicate, prior to exposure to weathering. Olivine and serpentine are decomposed by groundwater containing carbon dioxide to form soluble magnesium, iron, and nickel, and colloidal silica. The iron rapidly oxidizes in contact with air and precipitates by hydrolysis to form goethite and hematite, which remain near the surface of the deposit. The dissolved nickel and magnesium, and the colloidal silica, percolate downwards in the laterite deposit, remaining in solution so long as the solution is acidic. When the solution is neutralized by reaction with rock and soil, nickel, silica, and some of the magnesium precipitate as hydrated silicates. This process is illustrated in Figure 1 [14].

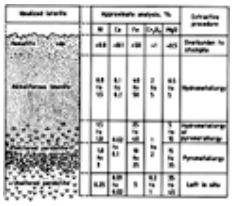

Figure 1. Typical section through alaterite deposit

Complete separation of iron and nickel into distinct zones is never achieved. Some or even most of the nickel may in fact remain in the upper layer, which is thus enriched in iron and nickel, but depleted in magnesium and silica. Such material, consisting mainly of ferric oxide minerals, is described as limonitic. This type of ore is typically relatively rich in cobalt and

chromium. By contrast, in the silicate minerals, the separation of iron and nickel is more complete, and the nickel is present in silicate minerals with a high magnesia content. Both types of material are almost always found in a laterite deposit, but in widely differing proportions.

The differences between limonitic and silicate minerals influence the methods by which they are treated for nickel recovery (Fig. 1). The silicates, which are chemically and mineralogically heterogeneous and require a very flexible processing method, are normally treated by pyrometallurgical processes. The limonites on the other hand are relatively homogeneous in composition and mineralogy and are well suited to hydrometallurgical treatment.

The type of weathering which dissolves silica and metallic elements from rock to produce limonitic and silicate nickel ores occurs most frequently in tropical climates with high rainfall, and with decomposing vegetable matter to provide organic acids and carbon dioxide in the groundwater. Oxide nickel mineral deposits are currently exploited in Australia, Brazil, Cuba, the Dominican Republic, Greece, Indonesia, the Philippines, New Caledonia, Yugoslavia, the United States, and the former Soviet Union. The laterite deposits in Oregon and California were formed when the climate was subtropical in past geologic epochs.

2.4. Economic Trends

Although 80 % of the world's known nickel resources on land are in oxide ore bodies, only 40 % of nickel production comes from this source. The greater part of the nickel produced is recovered from sulfide ores. This situation has arisen from the influence of political, geographic, technical, and economic factors. Sulfide ore deposits lie mostly in politically stable countries and close to the major nickel markets. The nickel content of sulfide ores can be concentrated relatively easily and cheaply by physical methods, while oxide ores must be treated in bulk. The potential for recovering the more numerous and valuable byproducts also favors the exploitation of sulfide ore. However, the grades of sulfide ores are continually declining as the richer ores are consumed, and the cost of underground mining, particularly labor costs, is continually increasing. Inevitably therefore, oxide ore bodies, which are all relatively uniform in grade and can be surface mined, must in future support a larger share of nickel production.

Unfortunately, the production of nickel from oxide ores by the established processes consumes two to three times as much energy as the processing of sulfide ores. Nickel oxide ore treatment is therefore very sensitive to fuel oil and electrical power costs. A major expansion in the exploitation of oxide ores occurred in the decade prior to 1973 when fuel oil prices were low. Many of the new laterite treatment plants built in this period became uneconomic after the oil price increase in 1973, and as a result there was little or no increase in the proportion of nickel produced from oxide ores between 1975 and 1990.

Several international mining industry consortia evaluated the feasibility of recovering nickel from deep sea manganese nodules during the 1970s, and potential mining techniques were developed and tested (see also ⟶ Manganese and Manganese Alloys – Processing of Manganese Nodules and Encrustations). However, processing of the nodules to separate and recover the metals proved to be significantly more expensive than for land-based ores [15], [16]. The international legal aspects of ownership of these deep sea resources were the subject of the 11th Session of the Third U.N. Conference on the Law of the Sea, which ended in April 1982 with only a partial agreement which several major participants refused to sign [17]. These developments, in conjunction with low nickel prices, essentially killed any serious consideration of deep sea nodules as a source of nickel during the 1980s. Development of this resource on a significant scale is therefore unlikely before the next century.

[Top of Page]

3. Beneficiation of Nickel Sulfide Ores [1], [18]

The low metal content of present-day nickel sulfide ores renders them unsuitable for either direct smelting or direct hydrometallurgical processing. Because the sulfide minerals usually occur as distinct grains in the rock matrix, these ores are amenable to a mechanical upgrading in which much of the rock content can be rejected. The metal content of the ore is concentrated by a physical treatment such as comminution to liberate the metal sulfide grains, followed by froth flotation or magnetic separation to recover a metal-rich concentrate [19].

In principle it is possible to produce separate nickel (pentlandite), copper (chalcopyrite), and iron (pyrrhotite) concentrates. However, a clean separation of pentlandite from pyrrhotite is difficult in practice since pyrrhotite typically contains intergrown inclusions of pentlandite as well as nickel in solid solution in the pyrrhotite. In fact pyrrhotite often contains 0.5 – 1 % Ni that cannot be separated by physical methods. The relatively high nickel content in pyrrhotite makes it difficult to achieve high yields of nickel from the ore unless the pyrrhotite is processed to recover its nickel content, and conversely the presence of nickel makes it difficult to obtain a marketable iron product from the pyrrhotite.

Most nickel producers, outside Canada, make no attempt to obtain separate nickel, copper, and iron concentrates and produce a bulk concentrate that contains pentlandite, pyrrhotite, and chalcopyrite as smelter feed. In Canada, both the major producers separate a large part of the pyrrhotite from pentlandite and chalcopyrite, and Inco further separates the pentlandite and chalcopyrite into separate nickel and copper concentrates. Even where the bulk of the pyrrhotite can be separated from pentlandite, residual pyrrhotite in the nickel concentrate still contributes a major part of the sulfur content.

In response to ever tighter constraints on sulfur dioxide emissions from smelters during the 1970s and 1980s, the Canadian companies have made increased pyrrhotite rejection from smelter feed a major priority, and substantial improvements have been achieved. A pyrrhotite concentrate was treated by Inco for many years for nickel and iron oxide recovery. However, environmental constraints on sulfur dioxide emissions have rendered the process obsolete, and pyrrhotite concentrates are now largely stockpiled, since none of the known treatment methods that avoid sulfur dioxide production is economic.

Pyrrhotite can be separated from pentlandite and chalcopyrite either by using its ferromagnetic properties or by flotation. Not all pyrrhotite is ferromagnetic, although in ores from the Sudbury, Ontario area the magnetic monoclinic form predominates

over the paramagnetic or “nonmagnetic” hexagonal form [20]. In general, pyrrhotite containing less sulfur than $Fe_{0.87}S$ is nonmagnetic. This material can usually be separated from pentlandite by flotation.

Nickel or nickel – copper concentrates typically range in grade from 5 to 15 % Ni + Cu, depending on the degree of pyrrhotite rejection achieved. An exceptionally high-grade concentrate, containing 28 % Ni and consisting largely of pentlandite, is produced in limited amounts at Inco's Thompson, Manitoba operation.

One of the more complex nickel concentration process flow sheets is that operated by Inco at its Sudbury operation [21]. In the two primary concentrators, the Clarabelle mill and the Frood – Stobie mill, 50 000 t/d of ore, grading 1.2 % Ni and 1.2 % Cu, is ground to a particle size of ca. 200 µm and treated by magnetic separation and flotation with a sodium amyl xanthate reagent to produce 5000 t/d of a bulk nickel – copper concentrate and 8000 t/d of a pyrrhotite concentrate. Metallurgical data for the Clarabelle mill are given in Table 4.

Table 4. Metallurgical data for the magnetic separation and flotation of nickel ore (Clarabelle mill, Inco) [21]

	wt %	Analysis, %					
		Cu	Ni	Pyr *	Cu	Ni	Pyr *
Ore	100	1.32	1.40	18.9	100	100	100
Magnetic concentrate	26	0.62	1.65	62.7	12	30	84
Flotation concentrate	8	14.4	11.6	15.9	84	64	7
Tails	66	0.07	0.13	2.5	4	6	9

* Pyr = Pyrrhotite.

The two primary concentrates are further upgraded in the Copper Cliff mill. The bulk nickel – copper concentrate is separated by flotation into separate nickel and copper concentrates by using lime and sodium cyanide at 30 – 35 °C to depress pentlandite and pyrrhotite. Under these conditions about 92 % of the copper is recovered to a concentrate grading 29 % Cu and 0.9 % Ni. About 98 % of the nickel and 92 % of the pyrrhotite are rejected to the separator tails, which grade 13 % Ni and 2.0 % Cu.

The Clarabelle rougher pyrrhotite concentrate is treated to separate residual chalcopyrite and pentlandite by grinding to ca. 74 µm followed by magnetic separation and flotation. A magnetic, cleaner pyrrhotite concentrate, a bulk nickel – copper concentrate, and a nonmagnetic pyrrhotite tail are produced. The nonmagnetic tails stream (1 % Cu, 2 % Ni, and 30 % pyrrhotite) is further treated, by regrinding and flotation in the pyrrhotite rejection circuit, for final recovery of pentlandite and chalcopyrite before it is discharged to the tailings pond. Overall, about 80 % of the pyrrhotite content of the ore is rejected from smelter feed to the pyrrhotite concentrate and tails. Metallurgical data for the Copper Cliff mill are given in Table 5.

Table 5. Metallurgical data for the upgrading of pyrrhotite and copper – nickel concentrate (Copper Cliff mill, Inco) [21]

	wt %	Analysis, %					
		Cu	Ni	Pyr *	Cu	Ni	Pyr *
Feed	100	4.58	4.42	49.6	100	100	100
Copper concentrate	13	29.2	0.91	8.4	83	3	2
Nickel concentrate	29	2.28	12.8	43.0	14	85	25
Pyrrhotite concentrate	33	0.11	0.96	84.4	1	7	57
Tails	25	0.36	0.91	32.3	2	5	16

* Pyr = Pyrrhotite.

Falconbridge at its Strathcona mill in the Sudbury basin treats ore containing 1.5 % Ni and 1.1 % Cu by grinding and flotation to produce a nickel – copper concentrate grading 6.7 % Ni and 5.6 % Cu at an average recovery of 83 % Ni and 93 % Cu [22-24]. A large part of the pyrrhotite content of the ore is rejected directly to the tailings pond. The separation of the nickel and copper in the bulk concentrate is carried out in the matte refining process after smelting.

The ore mined by Inco at Thompson, Manitoba consists of pentlandite and pyrrhotite with only a small amount of chalcopyrite (2.7 % Ni, 0.2 % Cu, 22 % pyrrhotite). Since the pyrrhotite contains 78 % of the sulfur but only 5.5 % of the nickel, a high degree of pyrrhotite rejection is feasible without seriously decreasing the nickel recovery [25]. The ore is ground and floated to recover a nickel – copper concentrate and reject 90 % of the gangue and 60 % of the pyrrhotite to tails. The nickel – copper concentrate is treated by flotation to separate first a copper concentrate, and then by magnetic separation and flotation to recover a high-grade (28 % Ni) pentlandite concentrate. The tailings stream from these steps forms the main nickel concentrate smelter feed (11 % Ni, 0.4 % Cu). The overall recovery of nickel to the nickel and pentlandite concentrates is about 87 %, while another 5 % of the nickel reports to the copper concentrate. Over 60 % of the pyrrhotite is rejected to

tailings with only 8 % of the nickel.

At the BCL nickel – copper mine in Botswana the pyrrhotite/pentlandite ratio in the ore averages about 13 : 1 and the pyrrhotite contains over 30 % of the nickel either in solid solution or as small pentlandite inclusions. Consequently, high nickel recoveries can only be achieved by recovering the nickel from pyrrhotite. This is achieved by producing a bulk pentlandite – pyrrhotite – chalcopyrite concentrate, with a typical composition of 2.8 % Ni and 3.2 % Cu [26].

Modern nickel concentrators are largely automated and are equipped with process control systems based on on-line X-ray analysis of process streams [27].

[Top of Page]

4. Pyrometallurgy of Nickel Concentrates [1], [28-32]

Over 90 % of the world's nickel sulfide concentrates are treated by pyrometallurgical processes to form nickel-containing mattes. Production capacities and composition of the feed concentrate and product matte are listed in Table 6 for several major nickel smelters.

Table 6. Nickel sulfide smelters [30-32]

Smelter	Feed concentrate, %				capacity, t/a Ni
	Ni	Cu	Ni	Cu	
Inco, Copper Cliff	13	2.7	63	15	110 000
Inco, Thompson	10.5	0.4	73	3	50 000
Western Mining	11.4	0.85	72	5	60 000
Falconbridge	6.7	5.6	40	34	45 400
Outokumpu	8 – 9	2 – 4	64	26	18 000
BCL, Botswana	2.8	3.2	38	40	20 000
Rustenburg	3.2	2.0	47	27	19 000
Bindura	10.6	3.4	70	20	14 400

The pyrometallurgical treatment of nickel concentrates includes three types of unit operation: roasting, smelting, and converting. In the roasting step sulfur is driven off as sulfur dioxide and part of the iron is oxidized. In smelting, the roaster product is melted with a siliceous flux which combines with the oxidized iron to produce two immiscible phases, a liquid silicate slag which can be discarded, and a solution of molten sulfides which contains the metal values. In the converting operation on the sulfide melt, more sulfur is driven off as sulfur dioxide, and the remaining iron is oxidized and fluxed for removal as silicate slag, leaving a high-grade nickel – copper sulfide matte.

In several modern operations the roasting step has been eliminated, and the nickel sulfide concentrate is treated directly in the smelter.

4.1. Roasting [33]

Roasting is basically a process in which the nickel sulfide concentrates are heated in an oxygen-containing gas, usually air, to a temperature (600 – 700 °C) at which oxygen oxidizes sulfide to sulfur dioxide and reacts with the metal values to form solid oxides, termed the calcine. The roasting step may also serve to preheat the charge for smelting.

The thermodynamic relationships between the metal sulfides and oxides provide the basis for the separation of iron from nickel, copper, and cobalt contained in the concentrate. The sulfides of iron, nickel, cobalt and copper are all in the same thermal stability range at the smelting temperatures (1200 – 1300 °C), but, in the presence of oxygen, each sulfide is unstable with respect to its corresponding oxide. Iron has the greatest affinity for oxygen, followed by cobalt, nickel, and copper. Consequently, if the nickel concentrate is roasted with a deficiency of oxygen the iron is oxidized preferentially, while virtually all of the nickel and copper remain as sulfides. Thus the degree of oxidation of the charge and the degree of elimination of sulfur can be controlled by regulating the supply of air to the roaster.

The degree of sulfur elimination achieved in the roasting step largely determines the grade of the matte produced as well as the losses of nickel and copper to the slag phase in the subsequent smelting operation. All the sulfur present in the calcine in excess of the amount needed to combine with the nickel, copper, and cobalt in the matte combines with iron, which therefore remains in the matte phase and lowers its nickel – copper – cobalt grade. The major constraint on achieving a high-matte grade in smelting is the increasing loss of nickel, copper, and cobalt to the slag phase. If all of the iron were to be oxidized in the roasting step so that it could be slagged off in the smelting stage and a high-grade matte produced directly, an increased amount of nickel, copper, and cobalt oxides would be formed in the roaster, and metal losses to the slag in smelting would increase substantially. Hence, each smelting operation seeks to operate at an optimum matte grade which gives the highest metal value recovery to the matte. The balance of the iron is removed in the converting step.

Normally only about 50 % of the pyrrhotite content of the nickel concentrate is oxidized in roasting, and under these conditions less than 5 % of the nickel and copper are oxidized, with the balance remaining as sulfides. The oxidation of

pyrrhotite to magnetite can be represented by the following reaction:

$$3\,Fe_7S_8 + 38\,O_2 \rightarrow 7\,Fe_3O_4 + 24\,SO_2$$

This reaction is strongly exothermic, and autogenous roasting, without the use of an external heat source, is practiced whenever possible.

In order to achieve effective roasting the furnace must provide good contact between the sulfide particles and the oxygen-bearing gas. The furnace system must also be able to provide close control of the degree of sulfur elimination. Both multihearth roasters and fluidized-bed roasters are in industrial use for nickel concentrate roasting. The use of traveling grate sintering machines, which were formerly widely used to prepare an agglomerated feed for blast furnace smelting, is now practiced only in the former Soviet Union.

The multihearth roasting furnace consists of a series of vertically superimposed circular hearths. Wet concentrate and smelting flux are charged to the upper hearth. A central rotating shaft drives rabble arms that extend across the roaster hearths, turning over the concentrate charge and transferring it via dropholes to the next lower hearth. Air, preheated by the combustion of natural gas, passes upwards through the hearths, heating the concentrate to ignition temperature and providing the oxygen required to burn the concentrate. The temperature is controlled by regulating the flow of air into the roaster.

Since the area of contact between the furnace gas and concentrate particles on the hearths is very limited, the total time required for roasting in this type of furnace is long, resulting in low throughput rates. A further disadvantage of the multihearth roaster is that low oxygen utilization and high air infiltration rates produce an off-gas with a low concentration of sulfur dioxide which is unsuitable for sulfuric acid production.

Multihearth roasters are still used in the Copper Cliff smelter of Inco in Ontario, which was built in 1930. A battery of 24 roasters can each treat 300 t/d of wet concentrate mixed with 75 – 90 t/d of siliceous flux. About 40 % of the sulfur content of the concentrate (13 % Ni, 2 – 3 % Cu, 30 % S, and 40 % Fe) is eliminated at a roasting temperature of 800 °C. Less than 5 % of the nickel and copper are oxidized. This operation is not thermally self-sustaining, due to high heat losses from the roaster, and natural gas is used to provide supplementary heating. The off-gas is cleaned to recover dust and is vented to a stack. The roasted calcine discharge from the bottom of the furnaces is conveyed directly to the reverberatory smelting furnaces [34].

The fluidized-bed roaster is a more recent development and has an improved gas – solids contact efficiency and a simpler construction (Fig. 2). The roasting furnace is a vertical chamber of circular cross section with a grate hearth through which the oxygen-bearing gas is blown. As nickel concentrate and smelting flux are fed to the chamber, the particles are suspended in the vertical air flow. Since each sulfide particle is in constant turbulent motion in the oxidizing atmosphere, roasting is uniform and rapid, with efficient heat transfer and oxygen utilization. Solids are discharged from the furnace either by overflow through a side port or by entrainment in the off-gas stream. The dust-laden gas may be passed through waste-heat boilers to recover its sensible heat. The cleaned off-gas stream contains a high concentration of sulfur dioxide, which can be converted to sulfuric acid.

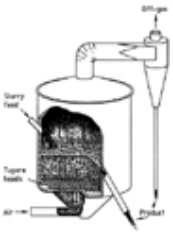

Figure 2. Cutaway view of a fluidized-bed roaster

Fluidized-bed roasting of nickel concentrates is practiced by Inco at the Thompson smelter in Manitoba and at Copper Cliff, where 10 % of the nickel concentrates are processed in fluidized-bed roasters, and by Falconbridge at its smelter in Ontario. The Thompson smelter operates five fluidized-bed roasters with a combined capacity of 3800 t/d of concentrate, while the Falconbridge smelter operates two roasters with a combined capacity of 2000 t/d of concentrate [22], [35].

In the Falconbridge smelter a slurry (75 % solids) of nickel – copper concentrate (6.7 % Ni, 5.6 % Cu, 30 % S, 40 % Fe) in water is fed directly to the roaster. Siliceous flux is added separately. Sulfur elimination ranges from 50 to 60 % at a roasting temperature of 600 – 650 °C. The operation is thermally autogenous, and the water contained in the feed slurry serves as a heat sink to control the bed temperature. Oxygen utilization is essentially 100 %. About 85 % of the calcine leaves the roaster in the off-gas stream which is treated in a series of cyclones, gas coolers, and electrostatic precipitators for solids recovery. The solids recovered from the gas cleaning system are combined with the coarser roaster bed overflow calcine, for charging to the electric smelting furnaces [24].

4.2. The Smelting Process

In the treatment of nickel sulfide ores and concentrates the function of the smelting process is to eliminate the gangue minerals and most of the iron sulfide and to concentrate the metal values into a high-grade nickel – copper matte containing 0.5 – 3.0 % Fe and 6 – 22 % S. Unlike copper, nickel metal cannot be produced by the oxidation of the sulfide at a practical smelting temperature (up to ca. 1400 °C).

When a mixture of metal sulfides, iron oxide, gangue, and siliceous flux are melted together, the iron oxide, gangue, and silica form a slag layer that floats on the heavier, molten sulfide or matte phase. The two immiscible layers are separated and

the slag is discarded.

The smelting process for nickel sulfide concentrates is normally carried out in two stages: primary smelting and converting. The oxidizing, slagging, and removal of iron in stages is critical to the efficiency and economics of smelting operations. If most of the iron were to be oxidized in the roaster and then slagged in a single smelting step, a matte high in nickel and low in iron would be produced. However, as noted above, significant amounts of the nickel and copper sulfides would also be oxidized in the roaster, and these oxides would be lost to the slag in the smelting step. Slag can also mechanically entrain droplets of matte and these losses will also increase with matte grade. Thus if a high-grade matte were to be produced in one step, the overall recovery of nickel would be low.

Provided that only about half the iron is oxidized in the initial roasting step, only 5 % of the nickel and copper sulfides are oxidized. In fact only a small amount of these oxidized metals will be lost to the slag because the oxides react with residual iron sulfides in the smelting process to produce the corresponding sulfides, which return to the matte phase.

$$9\,NiO + 7\,FeS \rightarrow 3\,Ni_3S_2 + 7\,FeO + SO_2$$

$$Cu_2O + FeS \rightarrow Cu_2S + FeO$$

Thus a low-grade, high-iron furnace matte is produced and only a small amount of nickel is lost to the slag as oxide when the nickel concentrate is treated in separate roasting and smelting steps.

The iron remaining in the furnace matte is oxidized and slagged in the secondary smelting step (converting). Since the slag produced in the converting stage, in which a high-grade, low-iron matte is produced, has a high nickel and copper content, it is returned to the primary smelting stage where the metal values are recovered by the reaction of the oxides with iron sulfide and by the settling of entrained droplets of matte. Thus the two-stage smelting process allows the production of a high-grade, low-iron nickel matte, and a high-iron slag of low metal grade which can be discarded without significant loss of metal values.

4.2.1. *Primary Smelting*

The primary smelting step is carried out in a variety of furnaces, including reverberatory, electric, and flash furnaces, which provide a wide variety of smelting conditions. Chemical losses of nickel and copper to the slag increase with the oxidation potential of the system. Thus metal losses to slag are much higher in the flash smelting furnace, in which the roasting step is combined with the primary smelting step, and smelting conditions are strongly oxidizing, than in the electric furnace, where smelting is carried out under reducing conditions. Consequently, slag from the flash smelting process must be treated in a separate step to recover the contained metal values, while slag from the electric furnace or reverberatory furnace smelting processes can be discarded without further treatment.

A typical furnace matte containing 10 % Ni, 65 % Fe, and 25 % S melts at ca. 950 °C, and nickel sulfide itself melts at ca. 790 °C. The melting point of a fayalite slag with an iron to silica ratio of about 7 : 3 is 1200 °C, although the slag melting point can be reduced to 1100 °C by the addition of a few percent of lime and alumina. With this type of slag, smelting is carried out at ca. 1200 °C. However, when significant levels of magnesia are present in the nickel concentrate or ore the melting point of the slag may be as high as 1260 °C, and smelting is then carried out at ca. 1400 °C [36-38].

Magnetite, which is formed in the partial roasting of the concentrates, reacts in the primary smelting step with the remaining iron sulfides and silica flux to form fayalite:

$$3\,Fe_3O_4 + FeS \rightarrow 10\,FeO + SO_2$$

$$2\,FeO + SiO_2 \rightarrow 2\,FeO \cdot SiO_2$$

The presence of magnetite in slags and mattes not only affects the chemistry of the smelting operation, but has an important effect on the smelting furnace. Magnetite is formed under strongly oxidizing conditions and its formation is favored by low operating temperatures and by silica-deficient slags. The magnetite content of slags from the more oxidizing reverberatory and flash smelting furnaces is usually above 10 %, with the content increasing with matte grade, while slags from reduction smelting electric furnaces typically contain less than 5 % magnetite. Magnetite tends to settle out in the bottom of the smelting vessel and accretions can severely reduce the operating capacity of the furnace. Consequently it is always desirable to minimize the formation of magnetite in the smelting process.

Reverberatory Furnace Smelting. In the reverberatory furnace, fossil fuel (pulverized coal, oil, or natural gas) is burned separately from the material being smelted. The furnace is a long rectangular structure with an arched roof and burners at one end and in the roof. A large part of the heat from the combustion gas radiates directly to the charge lying in the hearth below. A further substantial part of the heat is reflected from the walls and roof of the furnace into the charge. Solid charge is fed through vertical fettling pipes mounted in the roof along each side of the furnace. As the charge melts it flows towards the center of the furnace. Molten converter slag is introduced into the furnace at the firing end.

Slag is tapped from the end of the furnace opposite the burners, while matte is tapped from the side. This type of furnace has a low energy efficiency since the off-gas stream carries nearly 50 % of the heating value of the fuel. Waste-heat boilers or heat exchangers are used to recover heat from the off-gases.

The reduction of the magnetite formed in the roaster to ferrous oxide is not complete in the reverberatory furnace, and slags contain up to 7 % magnetite. Scrap iron or ferrosilicon is often added to control the formation of magnetite to minimize metal losses to the slag and to prevent magnetite buildup in the furnace hearth.

Inco uses two reverberatory furnaces equipped with roof-mounted oxyfuel burners for the primary smelting of nickel concentrates at its smelter in Copper Cliff, Ontario. Each furnace treats 1500 t/d of partially oxidized concentrate (13 % Ni,

2.7 % Cu) which is transferred at 700 °C from the multihearth roasters, together with 700 t/d of recycled molten converter slag. Oxygen enrichment of the combustion air supplied to the oxyfuel burners not only increases fuel utilization efficiency but also substantially increases the capacity of the furnace [34], [39-43].

The furnace matte contains 30 – 35 % nickel plus copper, together with the precious metals content of the charge. The reverberatory slag, which normally analyzes 0.2 % Ni, 0.1 % Cu, 0.1 % Co, 38 % Fe, and 36 % SiO_2, is discarded. The matte is tapped at 1150 °C and blown with oxygen-enriched air (30 % O_2), in one of 14 Peirce – Smith converters, to a high-grade, sulfur-deficient matte containing 78 % Ni and Cu and 20 % S. The converter slag, typically containing 2.3 % Ni and Cu, 50 % Fe, and 25 % SiO_2, is recycled to the reverberatory furnace for recovery of the contained nickel and copper [42].

Electric Furnace Smelting. The energy required for smelting may be obtained from electrical power as an alternative to burning fossil fuels or sulfides in air or oxygen. The use of an electric furnace for nickel sulfide smelting is favored when the cost of electrical energy is low or when the required smelting temperature is high, for example, when the concentrate contains high levels of magnesia.

Heating of the bath is achieved by passing a three-phase electrical current through a circuit consisting of carbon electrodes immersed in the slag layer, which has a high electrical resistance (submerged arc technique). Electric furnaces used for nickel – copper matte smelting are either rectangular with six electrodes in line or circular with three electrodes in triangular configuration. The consumable electrodes are made of carbon and are either prebaked or are of the self-baking Soederberg type.

The charge of concentrate (or calcine) and flux is fed through the roof. Thus a layer of unsmelted charge covers the slag giving it a “cold top”. The concentrate or calcine gradually settles into the slag as it melts and then separates into slag and matte layers. The slag and matte are tapped intermittently, as required. Since there is no fuel combustion in an electric furnace the quantity of off-gas is much less than from a reverberatory furnace and heat and dust recovery from the off-gas is therefore easier.

Electric furnace smelting of nickel concentrates is practiced by Inco at its Thompson, Manitoba, operation; by Falconbridge at Sudbury, Ontario; by the Severonikel Combine in the former Soviet Union; by Rustenburg Platinum, Impala Platinum, and Western Platinum in the Republic of South Africa; and by Bindura Nickel in Zimbabwe. The Canadian operations treat partly roasted concentrate calcines, while Bindura and the platinum producers treat nickel – copper concentrates directly. The Severonikel electric furnace smelter treats high-grade ore from Norilsk (3.5 % Ni, 3.5 % Cu) directly [44]. Current industrial practice is illustrated by the following examples.

In the Rustenburg Refiners smelter [45] in the Republic of South Africa, pelletized nickel – copper concentrate, containing 3.2 % Ni, 2 % Cu, 15 % Fe, 39 % SiO_2, 15 % MgO, 3 % CuO, 6 % Al_2O_3, 9 % S, and ca. 130 g/t precious metals, is mixed with a lime flux and smelted directly in a 19.5 MVA electric furnace. About 98 % of the nickel and 97 % of the copper and most of the precious metals contained in the concentrate are recovered to a matte containing about 17 % Ni, 10 % Cu, 40 % Fe, and 27 % S. The electric furnace slag, containing 0.1 % Ni and 0.06 % Cu, is tapped at 1350 °C and discarded. The matte is tapped at 1180 °C and converted in Peirce – Smith converters to a product analyzing 48 % Ni, 27 % Cu, 1.5 % Fe, and 20 % S, again with high recovery of the precious metals to the matte.

The Falconbridge smelter in Sudbury, Ontario, which was built in 1978 on similar lines to Inco's Thompson smelter, is a good example of the modern large-scale application of fluidized-bed roasting and electric furnace smelting of nickel – copper concentrates (Fig. 3). The smelter contains two roaster – electric furnace lines, each of which was designed to treat up to 1250 t/d of nickel – copper concentrate (6.7 % Ni, 5.6 % Cu). The electric furnaces are each rated at 33 MVA and have a six-in-line electrode configuration. The electrodes are of the Soederberg self-baking type. Typical furnace operating conditions are as follows [24], [46], [47]:

Concentrate feed rate	900 t/d
Power consumption	1.2 GJ/t solid charge
Slag temperature	1230 °C
Matte temperature	1130 °C
Off-gas temperature	700 °C
Calcine depth	500 mm
Slag depth	1500 mm
Matte depth	700 mm
Bottom buildup	150 mm
Electrode immersion	250 mm
Off-gas volume	28 300 m^3/h (STP)

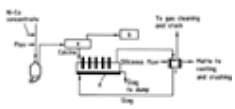

Figure 3. Flow sheet of the Falconbridge smelter

a) Gas scrubbing and calcine recovery; b) Acid plant; c) Converter; d) Electric furnace

Roaster calcine together with the flux added prior to roasting, at about 600 °C, is conveyed directly to the furnace and added

to the bath through fettling tubes mounted in the furnace roof. Recycle dusts and scrap together with reductants are added to the calcine conveyors. Converter slag is fed to the furnace through a launder above the matte-tapping end of the furnace.

Coke is added to the charge as a reductant, at a rate of 2.5 wt % of the concentrate, to improve metal recoveries and to reduce magnetite formation. The use of coke reductant and operation of the furnace under “black top” conditions results in a significant power saving compared with open-bath operation under oxidizing conditions. (1.2 GJ/t versus 1.5 GJ/t of solid charge).

The electric furnace slag, typically containing 0.12 % Ni, 0.20 % Cu, and ca. 4 % magnetite, is tapped at 1230 °C and is transferred in slag pots to a disposal area. The furnace matte, which normally contains 24 % Ni + Cu, is tapped at 1130 °C and is transferred to one of four Peirce – Smith converters where it is blown with air to produce a matte containing 74 % Ni + Cu, 1 % Co, 2.5 % Fe, and 23 % S. The converter matte is cast, crushed, and shipped to the Falconbridge refinery in Norway. The converter slag is recycled to the electric furnace, and the converter off-gases are cleaned in electrostatic precipitators and vented to a stack.

Flash Smelting. In flash smelting, a finely ground sulfide concentrate is smelted by burning some of its sulfur and iron content while the sulfide particles are suspended in the oxidizing atmosphere. The concentrate is roasted and smelted in a single process step. The predried concentrate and flux materials are injected with preheated oxygen-enriched air or commercial oxygen into the reaction shaft of a specially designed furnace. The exothermic heat of reaction of the iron sulfide with oxygen provides the energy to heat the particles to smelting temperature. The smelting temperature and the grade of matte formed are both controlled by adjusting the ratio of the amount of oxygen supplied to the furnace to the throughput rate of concentrate. In current practice, which uses oxygen-enriched air, the process can be thermally autogenous, and the addition of a fossil fuel is not generally required.

The flash smelting process was developed independently by Inco in Canada and Outokumpu in Finland, shortly after World War II. In both cases it was applied commercially to the treatment of copper concentrates. The Inco process uses commercial oxygen (95 %) while the Outokumpu process initially used air, augmented by fuel oil, as the oxidant and energy sources. The Outokumpu process was first applied commercially to the treatment of nickel concentrates in 1959. Oxygen enrichment of the combustion air was introduced by Outokumpu in 1971 and resulted in several major improvements in furnace performance. The Inco flash smelting process was successfully tested on a full production scale for the treatment of nickel concentrate in 1976, and will be utilized in the modernization of the Copper Cliff smelter scheduled for the early 1990s (see Section Environmental Aspects of Nickel Smelting). In addition to its own installation at Harjavalta in Finland, the Outokumpu flash smelting process for nickel has been adopted by the Western Mining Corporation in Australia, by BCL in Botswana, and by the Norilsk Combine in the former Soviet Union. The Outokumpu flash smelting process for copper has been adopted in about 30 operations worldwide [48], [49].

The Outokumpu flash smelting furnace, essentially an upright U in shape, consists of a vertical reaction shaft, a horizontal reverberatory settler hearth, and a vertical gas uptake shaft (Fig. 4). Predried nickel concentrates and fluxes are mixed with oxygen-enriched air and injected into the furnace by means of the concentrate burner mounted at the top of the reaction shaft. The resulting suspension spreads over the area of the reaction shaft. At operating temperatures, ignition of the iron sulfides occurs instantly, and the following exothermic reaction with oxygen provides all the heat required for smelting:

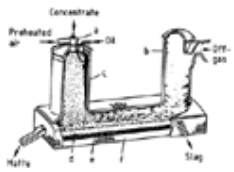

Figure 4. Cutaway view of an Outokumpu flash smelting furnace

a) Concentrate burner; b) Uptake; c) Reaction shaft; d) Slag; e) Matte; f) Settler

$$2\,FeS + 3\,O_2 \rightarrow 2\,FeO + 2\,SO_2$$

The nickel and copper sulfides melt to form a matte, while the iron oxide combines with silica flux to form a fayalite slag. Separation of the molten particles from the combustion gases occurs as the gas stream leaves the reaction shaft. The melt collects in the settler, where the matte and slag form separate layers and are tapped at opposite ends of the furnace.

Nickel flash furnace mattes typically grade 50 – 55 % Ni + Cu, and the grade is controlled by adjusting the ratio of oxygen to concentrate in the furnace. The furnace mattes are treated in Peirce – Smith converters with oxygen-enriched air to produce a high-grade matte product. The converter slag is either recycled to the flash furnace or treated directly in a slag cleaning furnace to recover the contained metal values. The composition of converter matte varies between smelters. The final sulfur content is adjusted in accordance with the requirements of the subsequent refining process. Thus at Harjavalta the matte is blown to a sulfur content of only 4 – 7 %, while the other smelters produce matte containing about 20 % S.

When oxygen-enriched air is used to produce a high-grade furnace matte, operating temperatures are sufficiently high to reduce magnetite formation and prevent significant magnetite accretions building up in the furnace. Inevitably, when a high-grade matte is produced under oxidizing conditions the slags contain too high a level of metal values to be discarded. Nickel flash furnace slags are generally cleaned by reduction and settling in an electric furnace, with coke as the reductant. The Western Mining (WMC) flash furnace in Australia has been modified by the extension of the settler beyond the gas uptake shaft [50], [51]. This extension to the settler hearth contains two sets of electrodes and serves as a continuous slag cleaning furnace. The furnace slag tapped from the extension is sufficiently low in metal values to be discarded. At the BCL smelter in Botswana, which treats a very low-grade concentrate and consequently has a high slag to matte ratio, slag losses are controlled in the settler itself by the addition of lump coal, which floats on the surface of the slag and reduces the metal oxides [26], [52]. The slag is further cleaned in two electric furnaces connected in series before being discarded. Metallurgical data for three flash smelting operations are presented in Table 7.

Table 7. Metallurgical data for flash smelters [50-53]

Smelter		Outokumpu	BCL	WMC
Capacity, t/a				
concentrate		190 000	800 000	600 000
Concentrate, %	Ni	8 – 9	2.8	11.4
	Cu	2 – 4	3.2	0.9
	Fe	24 – 27	43.5	38
	S	19 – 22	31.0	32
Furnace matte, %	Ni	33	15	44
	Cu	19	17	3
	Fe	18		27
	S	25		22
Converter matte, %	Ni	58	36	72
	Cu	32	40	5
	Fe	0.5	< 1	< 1
	S	6	20	21
Discard slag, %	Ni	0.15	0.15	0.28
	Cu	0.32	0.34	0.08
Smelter recovery, %		95	91.5	96

Off-gases leave the flash furnace through the uptake shaft and enter a waste-heat boiler where the heat content of the gas is recovered as high-pressure steam. The gas, which has a high sulfur dioxide content, is cleaned in electrostatic precipitators to recover entrained dust. The cleaned gas stream can be used as a feed to a sulfuric acid plant, for liquid sulfur dioxide recovery, or can be reduced to elemental sulfur with a fossil fuel.

The flash smelting process is particularly suited to the application of automated process control. The major control parameter is the ratio of oxygen in the combustion air to concentrate throughput. This parameter not only controls the degree of oxidation of the sulfides but also the grade of the matte and the operating temperature. The process is thus controlled by adjusting the flow rates of both combustion air and commercial oxygen, in relation to the concentrate feed rate. The slag composition is controlled by the rate of addition of flux materials to the concentrate in the furnace feed blend. Control of the process can thus be fully computerized [53].

4.2.2. *Matte Converting [54-56]*

In the converting step, iron sulfide is removed from the molten, low-grade furnace matte by oxidation and slagging. The slag, which contains high levels of nickel and copper, is returned to the primary smelting furnace for recovery of metal values. The high-grade, low-iron nickel – copper matte, which typically contains 20 % S, less than 1 % Fe, and also any precious metals present in the original concentrates, is the final product of the smelting process.

Converting is a batch operation. The horizontal side blown converter, known as the Peirce – Smith converter (Fig. 5), is normally used for the treatment of nickel furnace mattes, although top blown rotary converters are used in some modern plants, particularly for low production rates. Air or oxygen-enriched air is blown through the molten matte to form iron oxides and remove sulfur as sulfur dioxide. The iron oxides combine with added silica flux to form an iron silicate slag:

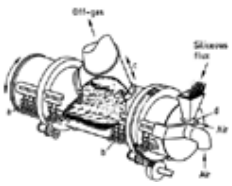

Figure 5. Cutaway view of a Peirce – Smith side blown converter

a) Tuyere pipes; b) Pneumatic punchers; c) Hood; d) Flux gun

$$2\,FeS + 3\,O_2 \rightarrow 2\,FeO + 2\,SO_2$$

$$2\,FeO + SiO_2 \rightarrow 2\,FeO \cdot SiO_2$$

A substantial amount of magnetite is also formed in the conversion process.

Converter slags contain 20 – 35 % silica, with the balance being iron oxides. Silica in the slag combines with iron(II) oxide, thus preventing further oxidation to magnetite. Slags low in silica contain large amounts of magnetite. Typically, up to one-third of the iron may be in the iron(III) state in a 20 % silica slag, while a slag with 35 % silica contains less than 15 % magnetite. Magnetite formation is useful in converting since it can be made to coat the inside of the brick-lined furnace and thus protects it from erosion and corrosion by the slag [57], [58].

The oxidation of iron sulfide is strongly exothermic and much of the heat generated in the conversion process can be utilized

to melt additional ore or concentrate feed or recycled scrap materials. Since the throughput of a converter is almost directly proportional to the amount of oxygen blown through the charge, its capacity can be raised substantially by enriching the air with oxygen. In addition to increasing the converting rate, oxygen enrichment of the air permits the treatment of additional cold charge, and in fact the addition of cold charge is essential to control the bath temperature when oxygen-enriched air is used. Up to 3 t of cold charge materials can be smelted for each tonne of oxygen supplied.

4.3. Environmental Aspects of Nickel Smelting

During the 1970s and 1980s the nickel industry responded to increasing public concern over smelter emissions of sulfur dioxide to the environment by replacing outdated process equipment, such as multihearth roasters, sinter machines, blast furnaces, and reverberatory furnaces, by processes such as fluidized-bed roasting, electric furnace smelting, and flash furnace smelting, which are more energy efficient and environmentally cleaner.

However, while the technology is available to produce sulfuric acid economically from sulfur dioxide in the high-strength off-gas streams produced by the modern smelting techniques, there is not always a market for large amounts of acid within an economic shipping distance of the nickel smelter. Thus, although the Inco and Falconbridge smelters in the Sudbury area of Ontario do convert a substantial amount of their sulfur dioxide to acid, more remote operations, including Inco's smelter at Thompson, Manitoba, Western Mining's smelter in Western Australia, and the BCL smelter in Botswana, vent SO_2-containing off-gases to the atmosphere. In fact the BCL smelter was designed with the capability to convert sulfur dioxide to elemental sulfur by reduction with fuel oil. However, this process proved uneconomic when oil prices increased in the early 1970s and it has never been operated. A similar circuit that uses natural gas to reduce sulfur dioxide to sulfur is in operation at the Norilsk flash furnace smelter in the former Soviet Union, which was commissioned in 1981 [58].

A more cost-effective method of reducing sulfur dioxide emissions is to reject as much pyrrhotite as possible from the smelter feed by separating pyrrhotite from pentlandite in the concentrator. As much as 50 – 60 % of the pyrrhotite content of the ore is now rejected in the major Canadian nickel operations at Sudbury and Thompson. At Thompson, Inco reduced SO_2 emissions by 45 % between 1976 and 1986 simply by increasing pyrrhotite rejection, but incurred significant metallurgical penalties by doing so [25]. Not only was the overall recovery of nickel reduced but the absence of much of the pyrrhotite removed a major energy source for the smelting process. Since the roasting and converting steps were designed to be autogenous using the iron sulfide as fuel, when over 50 % of the pyrrhotite is rejected prior to smelting insufficient iron is available to sustain the converting operation, and extra fuel must be supplied.

Maximum pyrrhotite rejection is practiced by both Inco and Falconbridge at their Sudbury operations. In these ores the nickel content of the pyrrhotite mineral is higher than in the Thompson ore and so pyrrhotite rejection results in significant losses in nickel recovery. A process to treat pyrrhotite concentrate separately to recover nickel, iron ore, and sulfuric acid, was operated by Inco for many years but the operation was abandoned in the mid 1980s when it became uneconomic due to increasingly stringent sulfur emission limits [32].

Falconbridge rejects over 55 % of the pyrrhotite in its ore and recovers sulfuric acid from the sulfur dioxide in the fluidized-bed roaster off-gas stream. Electric furnace and converter off-gases are vented to the atmosphere as a low-strength sulfur dioxide stream. The distribution of sulfur in the Falconbridge operation in 1983 is detailed in Table 8 [24], [52].

Table 8. Sulfur distribution at Falconbridge, Sudbury, Ontario in 1983 [24]

	Sulfur, t/d	Distribution, %
Mill tailings, incl. pyrrhotite	540	58.5
Smelter slag	11	1.2
Nickel – copper matte	44	4.8
Sulfuric acid production	200	21.7
Stack emissions	104	11.3
Copper concentrate, etc.	23	2.5

Particulate emissions from the Falconbridge smelter were also reduced by 80 % when the fluidized-bed roasting – electric furnace process replaced the old sintering – blast furnace process in 1978.

The Outokumpu flash smelting process is probably the most environmentally compatible of the current commercial nickel smelting processes because it permits the capture of virtually all the sulfur dioxide produced prior to the conversion step. The recovery of sulfur dioxide from the off-gas stream from Peirce – Smith converters is not generally feasible due to the low sulfur content of the off-gas and intermittent nature of the operation.

New Developments in Nickel Smelting In contrast to the copper industry, in which a number of continuous smelting processes have been commercialized, there has been little interest in developing a process for nickel concentrates in which the primary smelting and converting steps are combined into a single continuous operation. One such process was proposed by QUENEAU and SCHUMANN in 1976 but has not been adopted commercially [59].

The major development in nickel smelting during the 1990s will be the modernization of Inco's aging nickel – copper smelter at Copper Cliff, Ontario. Ontario Government regulatory measures to reduce acid rain require that Inco reduce its emissions of sulfur dioxide from 685×10^3 t/a in 1990 to 265×10^3 t/a by 1994, corresponding to an increase in the fixation of the sulfur in

the ore from 70 % to 90 %. In the existing operation, 55 % of the sulfur is rejected to tailings as pyrrhotite, and a further 15 % is converted either to acid or to liquid sulfur dioxide. The remaining 30 % is emitted to the environment as low sulfur dioxide content off-gases from the Herreschoff roasters, reverberatory furnaces, and the nickel matte and copper converters.

Several potential smelting options were examined by Inco, including the roast-reduction – smelting process, which was tested on a plant scale at Thompson in 1979 [32], and the oxygen flash smelting of nickel concentrate, which was tested with the copper flash furnace at Copper Cliff in 1976 [60]. However, none of these options was acceptable since they did not eliminate the copper converting operation, which itself produces more than the 1994 sulfur dioxide emission limit of 265 000 t/a [61].

The strategy eventually selected for the modernization is based on the oxygen flash smelting of bulk nickel – copper concentrate followed by converting the high-grade furnace matte in the existing nickel converters. The converter matte will be separated into metallics, nickel sulfide, and copper sulfide fractions using the existing matte separation process, and the iron-free copper sulfide (Cu_2S) will be oxygen flash smelted to semiblister copper. In this way the weak gas streams from the roasters, reverberatory furnaces, and copper converters will be totally eliminated, and the volume of gases from the nickel converters will be halved. All the sulfur dioxide in the high-strength (80 % SO_2) off-gas streams from the flash furnaces will be fixed as sulfuric acid or liquid sulfur dioxide. The oxygen flash smelting process is very energy efficient, and energy consumption in the smelter will be halved.

Since the entire nickel – copper separation will in future be carried out in the matte separation plant, Inco's milling operations will be simplified. Operations will be consolidated in the Clarabelle mill, and the Copper Cliff and Frood – Stobie mills will be closed. The function of the modernized Clarabelle mill will be to reject pyrrhotite and produce a bulk nickel – copper concentrate. A large new semiautogenous grinding mill and new flotation cells, more than ten times larger than the existing cells, will be installed. These changes are projected to increase the rejection of sulfur from the ore as pyrrhotite to tailings to 67 % [61].

In the smelter, two large new flash smelting furnaces, each capable of treating 1800 t/d of bulk concentrate, and a smaller flash furnace capable of treating 725 t/d of Cu_2S concentrate, will be constructed. A new oxygen plant will produce 450 t/d to supplement the 1300 t/d currently available.

A new sulfuric acid plant with a design rate of 2300 t/d will handle the off-gas from both large flash furnaces. Off-gas from the copper flash furnace will form the feed to the existing liquid sulfur dioxide plant. Installation of the new acid plant will permit the closing of the roaster operation and acid plant at the iron ore recovery plant.

The first large flash furnace together with the new oxygen plant and acid plant are scheduled for completion in 1992, while the second large flash furnace will be operating by the end of 1993.

4.4. Treatment of Converter Matte

Nickel – copper converter matte is essentially a melt of nickel, copper, and sulfur with small amounts of cobalt, iron, and oxygen, and traces of the noble metals and various undesirable impurities, including selenium, tellurium, and entrained slag. In the South African platinum industry, where nickel – copper mattes containing over 1000 g/t platinum group metals are common, the noble metal content of the matte far exceeds the base metal content in value. The platinum group metal contents of the Sudbury mattes, although much lower than in the South African mattes, are still of major economic significance.

Further processing of nickel – copper converter mattes is carried out to separate and recover nickel, copper, cobalt, and the noble metals. Mattes may be treated by hydrometallurgy (the Sherritt, Outokumpu, and Falconbridge leach processes), by electrometallurgy if they are low in copper (Inco's direct electrorefining of matte anodes), or by pyrometallurgy (Inco's slow-cooling matte separation process).

4.4.1. *Inco Matte Separation Process [1]*

The controlled slow-cooling matte separation process was developed by Inco in the 1940s to replace the obsolete Orford Process which depended on the different solubilities of copper and nickel sulfides in molten sodium sulfide. Other operations which had used the Orford Process, Rustenburg Platinum in the Republic of South Africa and the major Soviet nickel producers, have subsequently adopted the slow-cooling technique.

This process is based on the fact that when molten nickel – copper mattes freeze slowly, the constituents segregate into grains that are separate and distinct chemical phases. The major phases are Cu_2S and Ni_3S_2, but if the matte contains less sulfur than is required to form these compounds, the excess nickel forms a magnetic alloy phase with copper, which also preferentially collects the noble metals. The slow-cooled matte can then be comminuted and separated into nickel sulfide, copper sulfide, and nickel – copper alloy fractions by normal mineral beneficiation techniques. The nickel – copper alloy, which is separated magnetically, contains virtually all the platinum group metal and gold values at a substantially increased grade compared to the converter matte.

In the Inco matte separation plant at Copper Cliff (Fig. 6) [34] molten sulfur-deficient converter matte (50 % Ni, 26 % Cu, 0.5 % Fe, 22 % S) is poured into molds at 1000 °C and cooled slowly over a period of four days to about 200 °C. The solid matte is crushed and ground and is then separated into its components by magnetic separation and flotation. The copper sulfide concentrate (73 % Cu, 5 % Ni), which also contains most of the silver, selenium, and tellurium, is converted to blister copper in Inco's copper smelting circuit. The silver, selenium, and tellurium are recovered from the copper refinery anode slimes. The magnetic nickel – copper alloy (65 % Ni, 20 % Cu) forms the major feedstock to Inco's Copper Cliff nickel refinery, where the nickel is refined by the pressure carbonyl process to high-grade nickel powder and pellets, and copper and the precious metals are recovered from the refining residue. The nickel sulfide concentrate (70 % Ni, 0.6 % Cu), which is essentially depleted in precious metals, forms the feed to the fluidized-bed roasting plant.

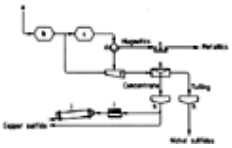

Figure 6. Flow sheet of the Inco matte separation process

a) Converter matte from Copper Cliff smelter; b) Rod mill; c) Ball mill; d) Magnetic separator; e) Centrifuge; f) Rake classifier; g) Flotation cell; h) Thickener; i) Disc filter; j) Dryer

The three large sulfide nickel operations in the former Soviet Union at Pechenga, Monchegorsk, and Norilsk, all use similar slow-cooling matte separation processes, producing separate nickel and copper sulfide concentrates. Rustenburg Platinum in the Republic of South Africa uses the slow-cooling technique to separate a magnetic nickel – copper alloy with a high noble-metal content for processing to recover and refine the platinum group metals, but does not separate the nickel and copper sulfides. The bulk nickel – copper concentrate is treated by pressure hydrometallurgy to separate the nickel and copper.

4.4.2. *Fluidized-Bed Roasting of Nickel Sulfide [62]*

Nickel sulfide concentrates are roasted in fluidized-bed roasters to produce nickel oxide at Inco's matte treatment plant in Copper Cliff and at the major nickel operations in the former Soviet Union. The dead roasting of finely ground nickel sulfide concentrates in a fluidized-bed roaster presents difficulties because nickel sulfide softens and becomes sticky at about 650 ° C, while sulfur removal must be carried out at a temperature above 1100 °C, at which the sulfide is molten, to achieve the required degree of sulfur elimination at an economic rate. A further problem is the very fine nature of the concentrates. Typically 99 % of the particles are less than 75 µm, with about 20 % in the 10 µm range. It is therefore difficult to feed such fine material into a fluidized-bed roaster without either incurring excessive dust carryover from the bed, or causing the bed to fuse with consequent loss of fluidization.

In Inco's practice (Fig. 7) [34] the concentrate is pelletized in order to provide enough mass to carry the sulfide particles down through the rising gas stream into the bed. Agglomeration is carried out by adding recovered roaster dust (about 10 % of the total) to the nickel sulfide concentrate slurry received from the matte flotation circuit. The green nickel sulfide pellets are charged into four fluidized-bed roasters which operate under optimum conditions at a superficial gas velocity of about 2 m/s with a bed expansion of about 150 %. The average retention time of a particle in the bed is ca. 8 h. Roasting under these conditions is autogenous at 1100 °C, but the roaster air is enriched with oxygen to increase the elimination of sulfur and increase the concentrate throughput rate. A coarse granular nickel oxide (NiO) containing ca. 0.2 % S is drawn off from the bed overflow. A single roaster can treat ca. 350 t/d of nickel concentrate.

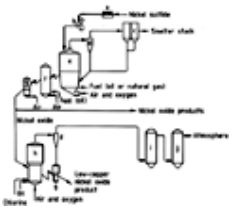

Figure 7. Flow sheet of the Inco nickel matte roasting process

a) Disc filter; b) Pelletizing disc; c) Electrostatic precipitator; d) Cyclone; e) Roaster; f) Reroaster; g) Cooler; h) Chlorination reactor; i) Splash tower; j) Packed tower

The roaster calcine is either cooled in a fluidized cooler with water sprays from 1100 °C to 90 °C and shipped to one of Inco's nickel refineries (Copper Cliff or Port Colborne, Ontario; or Clydach, Wales), or is further desulfurized in a second fluidized-bed roaster. In this second roasting step the sulfur content is reduced to levels as low as 0.005 %. This material is either marketed directly as ‘Sinter 75' (75 % Ni, 0.7 % Cu, and 0.006 % S), which is used in alloy steel making, or is treated by chlorination and reduction to lower its copper content. The removal of copper and other impurities is carried out by the injection of chlorine gas into a fluidized-bed reactor at 1200 °C, at a chlorine addition rate of about 3 wt % of the reactor feed. The bed is heated by light fuel oil and oxygen-enriched air. The reactor product typically contains 0.3 % Cu and less than 0.01 % S. It was formerly treated with a gas containing 75 % hydrogen at 510 °C in another fluidized-bed reactor to yield Incomet, a metallic nickel product, which, however, is no longer made.

[Top of Page]

5. Hydrometallurgy of Nickel Concentrates and Mattes [63]

Several hydrometallurgical processes are in commercial operation for the treatment of nickel – copper mattes to produce separate nickel and copper products. In addition, the hydrometallurgical process developed by Sherritt Gordon in the early 1950s for the direct treatment of nickel sulfide concentrates, as an alternative to smelting, is still commercially viable and competitive, despite very significant improvements in the economics and energy efficiency of nickel smelting technology [64].

In a typical hydrometallurgical process, the concentrate or matte is first leached in a sulfate or chloride solution to dissolve nickel, cobalt, and some of the copper, while the sulfide is oxidized to insoluble elemental sulfur or soluble sulfate. Frequently, leaching is carried out in a two-stage countercurrent system so that the matte can be used to partially purify the solution, for example, by precipitating copper by cementation. In this way a nickel – copper matte can be treated in a two-stage leach process to produce a copper-free nickel sulfate or nickel chloride solution, and a leach residue enriched in copper. The copper-rich residue is treated by pressure leaching, or by roasting and leaching, to solubilize the copper as copper sulfate so that it can be recovered from solution electrolytically as cathode copper. Nickel is recovered from the purified nickel sulfate or chloride solution either electrolytically as a pure nickel cathode or by chemical reduction with hydrogen to give pure nickel powder.

The range of compositions for mattes treated by hydrometallurgical processes is shown in Table 9 [63]. The nickel contents of the mattes treated hydrometallurgically are in the range 35 – 75 %, while the copper content varies from 0 to 52 %, and sulfur from 6 to 24 %. Low-copper mattes (< 10 % Cu) can be treated by ammoniacal pressure leaching, by ferric chloride leaching, or by direct electrorefining of the matte. High-copper mattes can be treated by high-pressure or atmospheric pressure sulfuric acid leaching or hydrochloric acid leaching.

Table 9. Hydrometallurgical nickel refineries [63]

Company/operation	Capacity, t/a Ni	Feed composition, %				
		Ni	Cu	Co	Fe	S
Falconbridge, Norway	54 000	35 – 40	30 – 35	1.0	2.5	22 – 24
Inco, Thompson	46 000	73	3	0.8	0.6	20
Western Mining	33 000	44 – 72	3 – 5	0.8	1 – 25	20 – 23
Sherritt Gordon	24 000	15 – 70	1 – 5	0.5	2 – 30	18 – 30
Sumitomo, Niihama	22 000	72	5	0.7	0.5	21
Rustenburg Refiners	20 000	47	27	0.5	3.0	21
Outokumpu, Harjavalta	17 000	55 – 65	25 – 35	0.6	0.2	6
SLN, Le Havre	16 000	75		1.5	2.5	20
Jianchuan, China	15 000	68	7	1.0	2.0	21
Bindura, Zimbabwe	14 400	66	25	0.6	0.2	6
Empress, Zimbabwe	5000	40	52	0.2	0.2	6
Western Platinum	2000	48	28	0.5	1.0	21

The sulfur contents of the mattes are adjusted during the conversion process in the smelter to the optimal level for the subsequent leaching process. Sulfur levels are typically 20 – 23 %, except when the matte is to be treated by the Outokumpu atmospheric pressure sulfuric acid leaching process, for which the sulfur content is reduced to about 6 %.

5.1. Ammonia Pressure Leaching [1], [65]

The first successful commercialization of a hydrometallurgical process for the treatment of a nickel sulfide concentrate was realized by Sherritt Gordon Mines at Fort Saskatchewan, Alberta, Canada in 1954. This process was developed specifically to treat a pentlandite concentrate from Sherritt's Lynn Lake mine in northern Manitoba by ammonia pressure leaching. The original plant was designed to produce 8000 t/a of nickel powder and briquettes by the chemical reduction of nickel in ammoniacal solution with hydrogen at high pressure. The same plant, with only minor process and equipment modifications, can now refine 24 000 t/a of nickel from a wide range of feed materials, including pentlandite concentrates, low-copper nickel mattes, and nickel sulfates and carbonates. The ammonia pressure leach process is also operated by Western Mining at Kwinana, Western Australia. Originally, this refinery processed nickel sulfide concentrates, but now treats a blend of converter and flash furnace nickel mattes to produce powder and briquettes at a rate of 33 000 t/a Ni. A block flow diagram of the ammonia pressure leach process as operated by Sherritt Gordon is shown in Figure 8.

Figure 8. Flow sheet of the Sherritt ammonia pressure leach process

The finely ground nickel sulfide concentrates and mattes are leached in ammoniacal ammonium sulfate solution at 80 – 95 °C under 850 kPa air pressure in a two-stage countercurrent process in eight four-compartment horizontal autoclaves (Fig. 9). Nickel, cobalt, and copper are dissolved as soluble ammonia complexes, while sulfide is oxidized to form a series of soluble thio salts. Reactive iron is oxidized and reports to the leach residue as a hydrated oxide together with pyrite which does not react in the leach:

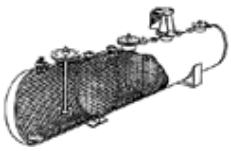

Figure 9. Cutaway view of a horizontal leaching autoclave

$$NiS \cdot FeS + 3\,FeS + 7\,O_2 + 10\,NH_3 + 4\,H_2O \rightarrow Ni(NH_3)_6SO_4 + 2\,Fe_2O_3 \cdot H_2O + 2\,(NH_4)_2S_2O_3$$

The leach solution typically contains 50 – 60 g/L Ni, 1 – 2 g/L Co, 5 – 10 g/L Cu, 130 g/L NH_3 and varying amounts of thio salts, including thiosulfate and thionates, as well as ammonium sulfamate ($H_2NSO_3NH_4$), which is formed by the reaction of

thionate with oxygen and ammonia.

The leach solution is boiled to distil off the free ammonia, and elemental sulfur and sulfur dioxide are added to augment the thio salts in precipitating copper as copper(II) sulfide:

$$Cu(NH_3)_4SO_4 + S + SO_2 + 2\,H_2O \rightarrow CuS + 2\,(NH_4)_2SO_4$$

The copper sulfide is sold to a copper smelter. The ammonia to nickel molar ratio is adjusted to 2 : 1 and the copper free ammoniacal nickel sulfate solution is oxidized at 235 °C under 4 MPa air pressure to convert the sulfur compounds to sulfate and to hydrolyze the sulfamate ion to sulfate. If these sulfur compounds were not fully converted to sulfate, the nickel powder product would contain unacceptably high levels of sulfur.

In the final step of the process, nickel is precipitated from solution as a metal powder by reduction with hydrogen under 3.6 MPa pressure at 200 °C in a four-agitator horizontal autoclave:

$$[Ni(H_2O)_4(NH_3)_2]SO_4 + H_2 \rightarrow Ni + (NH_4)_2SO_4 + 4\,H_2O$$

Details of the reduction technique are given in Section Hydrogen Reduction to Nickel Powder . The nickel powder is either compacted into briquettes for use in stainless steel production or is rolled into nickel strip for coinage production.

Under the specified process conditions cobalt is not precipitated during nickel reduction and the reduction end solution contains 1 – 2 g/L each of nickel and cobalt as well as about 400 g/L ammonium sulfate. This solution is treated with hydrogen sulfide to precipitate the nickel and cobalt as a mixed metal sulfide intermediate product, leaving a pure ammonium sulfate solution which is evaporated to crystallize ammonium sulfate for sale as a fertilizer.

The mixed metal sulfides produced in both the Sherritt and the Western Mining nickel refineries are processed in Sherritt's cobalt refinery in Canada [65]. The sulfides are pressure leached in sulfuric acid under 1.15 MPa air pressure at 150 °C in a lead- and brick-lined, six-compartment horizontal autoclave to form a solution of nickel and cobalt sulfates:

$$NiS + CoS + 4\,O_2 \rightarrow NiSO_4 + CoSO_4$$

Iron is removed from solution in a second autoclave by adjusting the pH to 2.5 with aqueous ammonia solution. At this stage the solution typically contains 40 g/L Ni, 30 g/L Co, and less than 10 mg/L Cu and Fe. The solution is then treated with aqueous ammonia solution under air pressure to oxidize the cobalt(II) ion to the cobalt(III) state which is stable as the soluble cobalt(III) pentammine complex ion $[Co(NH_3)_5 \cdot H_2O]^{3+}$. Nickel is removed in a two-stage process by adding sulfuric acid to selectively precipitate nickel(II) ammonium sulfate, which is recycled to the nickel refinery. The cobalt/nickel ratio in the purified solution is typically 3000 : 1.

The cobalt(III) ion is reduced to cobalt(II) by treating the purified solution with cobalt powder and sulfuric acid. Cobalt powder is precipitated from solution by reduction with hydrogen at 180 °C and 4.3 MPa.

The pentammine cobalt refining process has also been operated by Outokumpu at Kokkola in Finland since 1968.

5.2. Atmospheric Acid Leaching

The atmospheric acid leach process for the treatment of high-copper, low-sulfur nickel matte was developed by Outokumpu [1] in 1960 to process material from its nickel mine in Finland, and was later adopted by the Bindura and Empress nickel refineries in Zimbabwe [66]. In this process the leaching operations are carried out in trains of air-agitated pachuca tanks, in contrast to the mechanically-agitated horizontal autoclaves used in the pressure leaching process.

The Outokumpu process is outlined in Figure 10. Granulated matte, typically analyzing 55 – 65 % Ni, 25 – 35 % Cu, 0.6 % Co, 0.2 % Fe, and 6 % S, is ground to 90 % less than 54 µm and leached at 80 °C in acidic sulfate solution in a three-stage countercurrent leach circuit. Nickel, which is largely present in the low-sulfur matte as metallic nickel, dissolves in the first stage leach by reaction with sulfuric acid and with soluble copper, which is cemented from solution as copper metal. Copper sulfide in the matte does not react:

Figure 10. Flow sheet of the Outokumpu atmospheric acid leach process

$$Ni + H_2SO_4 + 1/2\,O_2 \rightarrow NiSO_4 + H_2O$$

$$Ni + CuSO_4 \rightarrow NiSO_4 + Cu$$

Most of the remaining nickel and some of the copper is leached in the second and third leach stages, which operate under increasingly oxidizing and acidic conditions. The solution from the first leach stage is a copper-free nickel sulfate solution, while the residue from the third stage is enriched in copper relative to nickel. Metal extractions in the atmospheric leach are typically about 90 % Ni and 50 % Cu.

The atmospheric leach residue, which contains about 15 % Ni, 50 % Cu, and 20 % S, was originally recycled to the Outokumpu nickel smelter to recover the metal values. This residue is now treated in a fourth stage, a pressure leach under mildly oxidizing conditions at 160 °C, in which the nickel is extracted selectively, leaving a low-nickel, high-copper sulfide residue that can be treated directly in the copper smelter [67].

The Bindura and Empress refineries do not have a pressure leaching step and operate in the original Outokumpu mode, with copper recovery by electrowinning from the second-stage leach liquor, and recycle of the leach residue to a nickel smelter. At the Empress refinery, good quality copper cathode is produced, despite an unfavorable copper to nickel ratio in the solution, by using air-sparged cells at relatively low current density [66].

The first stage atmospheric leach step is primarily a solution purification step designed to produce a copper-free nickel sulfate solution. Air addition rates are reduced in the later pachuca leach tanks in the first stage to avoid redissolution of copper, and the pH is allowed to rise to 6.0 to ensure precipitation of iron by hydrolysis. Lead is precipitated by adding barium hydroxide or carbonate to the solution, either in the first stage leach or in a separate process step, to precipitate lead sulfate, which coprecipitates with barium sulfate. Cobalt is separated from the nickel sulfate solution in a later step, in which a slurry of nickel(III) hydroxide is added to precipitate cobalt as cobalt(III) hydroxide:

$$Ni(OH)_3 + CoSO_4 \rightarrow NiSO_4 + Co(OH)_3$$

The nickel(III) hydroxide is produced by electrolytic oxidation of nickel(II) hydroxide, formed by neutralizing purified nickel sulfate solution with sodium hydroxide. The cobalt(III) hydroxide precipitate is refined to cobalt metal and cobalt salts by the pentammine process at Outokumpu's cobalt refinery at Kokkola, Finland.

Nickel is recovered from the purified nickel sulfate solution, which contains about 75 g/L Ni, as a high-grade cathode by direct electrowinning (see Section Electrowinning). Anolyte solution containing about 50 g/L Ni and 50 g/L H_2SO_4, is recycled for use as a leachant in the leaching circuits.

5.3. Acid Pressure Leaching

The acid pressure leaching process for the treatment of high-copper nickel mattes was developed by Sherritt Gordon during the 1960s and is currently operated by all four platinum producers in the Republic of South Africa: Rustenburg Refiners, Impala Platinum, Western Platinum, and Barplats [68], [69]. A similar process was also operated by the Amax nickel refinery in Louisiana, United States, from 1974 to 1986 [70], [71].

Acid pressure leaching is normally carried out in brick- and lead-lined horizontal multicompartment autoclaves at 135 – 160 °C and oxygen partial pressures of 140 – 350 kPa. The major advantage of this process for the platinum producers is that very high extractions of nickel, copper, and sulfur can be achieved. Typically, the metal and sulfur extractions are greater than 99.9 %, so that the leach residue is a high-grade platinum group metal concentrate that is virtually free from nickel, copper, and sulfur and can be treated directly in a platinum refinery. An upgrading factor for the platinum group metals of over 100 from the matte to the leach residue is readily attainable [69].

The process flows heet differs significantly for the various operations. Rustenburg and Impala use two countercurrent stages of continuous pressure leaching at 135 – 150 °C, while the smaller Western Platinum and Barplats refineries operate an atmospheric leach at 85 °C followed by a single-stage continuous pressure leach at 160 °C. The nickel products also differ widely. Rustenburg Refiners recover nickel as a high-purity cathode by direct electrowinning, whereas Impala produces nickel powder and briquettes by hydrogen reduction, and Western Platinum and Barplats produce crystalline nickel sulfate, which is processed to metal or plating salts elsewhere.

The Impala Platinum flow sheet [72] is outlined in Figure 11. The nickel – copper matte is leached in two stages at 135 °C under 1 MPa air pressure in spent copper electrolyte solution.

Figure 11. Flow sheet of Impala Platinum acid pressure leach process

$$Ni_3S_2 + H_2SO_4 + 1/2\,O_2 \rightarrow NiSO_4 + 2\,NiS + H_2O$$

$$2\,NiS + 4\,O_2 \rightarrow 2\,NiSO_4$$

$$NiS + CuSO_4 \rightarrow NiSO_4 + CuS$$

Solution from the first stage leach, which contains a high concentration of nickel (typically 100 g/L Ni) and low levels of copper and iron, is contacted with nickel powder and sodium hydrosulfide to precipitate copper, and iron is then removed as an ammonium jarosite precipitate using a horizontal multicompartment autoclave operating at 150 °C:

$$CuSO_4 + Ni \rightarrow NiSO_4 + Cu$$

$$3\,Fe_2(SO_4)_3 + (NH_4)_2SO_4 + 12\,H_2O \rightarrow 2\,NH_4Fe_3(SO_4)_2(OH)_6 + 6\,H_2SO_4$$

The jarosite precipitation step also effectively removes lead and arsenic from the nickel sulfate solution. The purified nickel

sulfate solution is adjusted with aqueous ammonia solution to form the nickel(II) diammine complex ion, and nickel is recovered as a powder by reduction with hydrogen under pressure:

$$Ni(NH_3)_2SO_4 + H_2 \rightarrow Ni + (NH_4)_2SO_4$$

Cobalt and residual nickel are recovered from the reduction end solution as the nickel sulfate – cobalt sulfate – ammonium sulfate double salts, which are refined to produce cobalt metal powder by the Sherritt pentammine process and hydrogen reduction.

The residue from the first stage leach is leached under more oxidizing and more acidic conditions in copper spent electrolyte in the second-stage pressure leach, which completes the extraction of nickel and copper sulfides:

$$NiS + 2\,O_2 \rightarrow NiSO_4$$

$$Cu_2S + 1/2\,O_2 + H_2SO_4 \rightarrow CuS + CuSO_4 + H_2O$$

$$CuS + 2\,O_2 \rightarrow CuSO_4$$

The solution from the second-stage pressure leach, which contains both nickel and copper sulfate with traces of iron and selenium, is treated with sulfur dioxide to precipitate selenium as copper(I) selenide, and copper is recovered as cathode by electrowinning. The spent electrolyte, which contains high levels of nickel as well as 20 g/L Cu and 50 – 70 g/L H_2SO_4, is recycled to the pressure leach circuits.

The second-stage leach residue contains the platinum group metals, with only minor levels of nickel, copper, and sulfur.

5.4. Chloride Leach Processes

Nickel matte refining processes based on leaching in chloride solution in the presence of chlorine gas have been developed and successfully commercialized by Falconbridge at Kristiansand in Norway [73] and by Societé Le Nickel (SLN) at Le Havre-Sandouville in France [74]. Leaching of the matte is carried out at atmospheric pressure at the boiling point of the leach solution (ca. 110 °C). Chlorine, which is recovered from the electrolysis cells, is sparged directly into the leach tanks. In the SLN process the leachant is the iron(III) chloride solution recovered from an iron removal step, whereas in the Falconbridge process the leachant is a solution of copper and nickel chlorides recycled from the copper residue treatment circuit. The flow sheet of the Falconbridge process is shown in Figure 12.

Figure 12. Flow sheet of the Falconbridge refinery process

Granulated nickel – copper mattes from the Falconbridge smelter in Ontario, Canada and from BCL in Botswana, with an average composition of 40 – 45 % Ni, 25 – 30 % Cu, 20 – 22 % S, 2 – 3 % Fe, and 1 – 1.5 % Co, are ground in a ball mill [73]. The ground matte is leached with chlorine in a nickel – copper chloride solution in a train of agitated tanks. The leaching conditions are controlled by means of redox potential measurements to ensure selective leaching of nickel:

$$2\,CuCl + Cl_2 \rightarrow 2\,CuCl_2$$

$$Ni_3S_2 + 2\,CuCl_2 \rightarrow 2\,NiS + NiCl_2 + 2\,CuCl$$

$$NiS + 2\,CuCl_2 \rightarrow NiCl_2 + 2\,CuCl + S$$

$$Cu_2S + S \rightarrow 2\,CuS$$

Essentially all the copper and sulfur present in the matte remain in the leach residue as CuS. The leach residue still contains about 8 % Ni as unreacted NiS, and the leach solution still contains high concentrations of copper. The leach slurry is therefore heated to 140 – 145 °C in two vertical autoclaves, connected in series, to promote the precipitation of copper and the dissolution of nickel by a series of metathesis reactions:

$$Ni_3S_2 + S + 2\,CuCl \rightarrow Cu_2S + NiCl_2 + 2\,NiS$$

$$Ni + S + 2\,CuCl \rightarrow NiCl_2 + Cu_2S$$

$$Cu_2S + S \rightarrow 2\,CuS$$

$$S + 2\,CuCl \rightarrow CuS + CuCl_2$$

The discharge from the autoclaves, which still contains about 7 g/L Cu in solution, is treated with fresh matte to reduce the copper concentration to less than 0.5 g/L. Overall nickel extraction is about 90 %, and the final residue contains 6 – 8 % Ni.

Iron, which is largely dissolved in the chlorine leach, is precipitated by oxidation with chlorine and hydrolysis with nickel carbonate. Cobalt is extracted from the nickel chloride solution by solvent extraction with triisooctylamine (TIOA) and, after purification of the strip solution, is recovered as high-purity cathode by electrowinning. The nickel chloride raffinate is diluted and purified further by treatment with chlorine and nickel carbonate to remove the remaining traces of cobalt, iron, copper, arsenic, lead, and manganese. High-purity nickel is recovered by direct electrowinning as cathode, and as crowns by direct plating onto a suitably masked blank.

Copper is extracted from the chlorine leach residue by roasting it in a fluidized-bed roaster to convert the copper sulfide and elemental sulfur to copper oxide and sulfur dioxide, which is converted to sulfuric acid, and leaching the calcine in sulfuric acid to selectively leach the copper. The copper is recovered from the sulfate solution by conventional electrowinning.

The calcine leach residue, which contains 50 % Ni and 18 % Cu as well as the platinum group metals (PGM), undergoes further chloride leaching to recover the nickel and copper in solution for recycling to the leach circuit, leaving the platinum group metals in an upgraded concentrate. Typical process stream compositions for the Falconbridge operation are shown in Table 10.

Table 10. Process data for the Falconbridge chlorine leach process [73]

Stream	Ni	Co	Cu	Fe	S	Cl
Matte, %	45	1	28	2	22	
Feed solution, g/L	60		25	1		90
Sulfide leach residue, %	7	0.5	56	2	33	0.3
Pregnant solution, g/L	230	5	0.5	6		270
Ni electrolyte, g/L	60					60
Roaster calcine, %	9	0.5	63	2	0.5	
Cu electrolyte, g/L	60	6	60	2.5		
PGM concentrate, %	20	1	25	5	15	5

In the SLN process, in which the feed matte produced from laterite ore in New Caledonia contains virtually no copper, the pregnant leach liquor contains 200 g/L Ni with 5 g/L Co and 10 g/L Fe. Iron is removed as ferric chloride by solvent extraction with tributyl phosphate (TBP). Part of the ferric chloride strip solution is recycled to the chlorine leach, while the balance is concentrated and sold. Cobalt chloride is extracted with TIOA and sold as solution for cobalt recovery elsewhere. Nickel is recovered from the purified solution as high-purity cathode by electrowinning [74].

5.5. Treatment of Nickeliferous Pyrrhotite [1]

Large tonnages of pyrrhotite concentrates containing up to 1 % Ni are currently rejected in the milling and concentrating operations of the major Canadian nickel producers as a means of reducing the iron and sulfur inputs to their smelters, and hence reducing sulfur dioxide emissions. These pyrrhotite concentrates represent a major potential source of nickel and several attempts have been made to develop economically viable processes to exploit them. The primary requirement for a pyrrhotite treatment process is that it should produce an iron ore suitable for use in the steel industry and yield a good recovery of nickel. Such processes are typically based on roasting to eliminate all or part of the sulfur and hydrometallurgical treatment of the calcine to recover the nickel.

In 1956 Inco commissioned a plant at Copper Cliff, Ontario, to treat 1.2×10^6 t/a of pyrrhotite concentrate analyzing 0.9 % Ni, 58 % Fe, and 35 % S. The iron sulfide concentrate was first oxidized to a low-sulfur calcine in fluidized-bed roasters, producing an off-gas containing sulfur dioxide from which sulfuric acid was produced. The calcine was then passed through a reducing kiln to reduce the nickel, cobalt, and copper oxides selectively to metallic form, so that they could be dissolved selectively in ammoniacal ammonium carbonate solution. The solution was purified to remove copper as the sulfide and nickel was precipitated as a basic nickel carbonate. The basic carbonate was calcined to a marketable oxide product containing 77 % Ni. The leached iron oxide residue was pelletized and indurated for market. This plant operated until the early 1980s, when marginal economics and increasing constraints on sulfur emissions resulted in its closure.

From 1960 until the early 1970s, Falconbridge operated a process in which pyrrhotite concentrate, analyzing 1.1 % Ni, 57 % Fe, and 36 % S, was sulfation roasted at 680 °C to convert the nickel, copper, and cobalt to their water-soluble sulfates. The calcine was then water-leached to extract the metals, and nickel, cobalt, and copper were precipitated from solution by cementation on iron turnings for recycling to the smelter. The iron oxide product analyzed 66 % Fe, 0.13 % Ni, and 0.5 % S.

[Top of Page]

6. Nickel Refining [1], [63]

Nickel differs from copper in that it is largely used as an alloying element in stainless steel and alloy steel production. Thus, whereas most copper is used in electrical and thermal energy applications, which require a highly refined metal, a high-purity metal is not required for many of nickel's most important markets. Nickel products such as nickel oxide, metallized nickel oxide, and ferronickel, which are designated Class II products, are produced directly by smelting and roasting and are sufficiently pure, without refining, for many nickel applications such as stainless steel production. These Class II products may contain significant levels of cobalt and copper and a range of trace impurities.

In other applications, high-purity nickel is essential and Class I nickel products, which include electrolytic cathode, carbonyl powder, and hydrogen-reduced powder, are made by a variety of refining processes. In general nickel is refined to decrease to an acceptable level only those impurities which would have an adverse effect on subsequent processing or on the properties of the refined metal product. These elements include antimony, arsenic, bismuth, cobalt, copper, iron, lead, phosphorus, sulfur, tin, and zinc. Cobalt, which resembles nickel closely in physical and chemical properties, is difficult to remove completely and since it is not often a serious contaminant in most applications for nickel, many Class I products contain quite high levels of cobalt.

The object of the refining process is not only to remove the impurities from the nickel, but also to recover those that have an economic value. Obvious examples are the precious and platinum group metals (Pt, Pd, Os, Rh, Ir, and Ru), cobalt, and copper; sulfur, selenium, and tellurium can sometimes also be recovered economically.

Historically the first successful nickel refining operation was established by the Mond Nickel Company (later absorbed by Inco) at Clydach, Wales, in 1902, using the carbonyl refining process invented by MOND and LANGER in 1889. This refinery is still in operation using a modern version of the carbonyl process. A large refinery that uses a high-pressure version of the carbonyl process was established by Inco at Copper Cliff, Ontario in 1972 [75].

The first electrolytic nickel refinery, treating nickel metal anodes, was built by Hybinette in Kristiansand, Norway, in 1910, and this plant was acquired by Falconbridge in 1928. The Kristiansand Refinery was modernized in the late 1970s by replacing the Hybinette process with the present chlorine leach process [73]. Inco established its first nickel refinery in Canada at Port Colborne, Ontario, in 1926, using a version of the Hybinette refining process. This refinery operated until 1984. In 1989 the only electrolytic nickel in North America was produced by Inco's refinery at Thompson, Manitoba, which electrorefines matte anodes [76]. This process is also operated in Japan [77] and China. The electrorefining of nickel metal anodes, which is effectively obsolete in Western Europe and North America, is still practiced in the former Soviet Union [44].

Outokumpu, SLN, and Falconbridge in Europe and Bindura, Empress, and Rustenburg Refiners in Southern Africa all recover electrolytic nickel by direct electrowinning from purified solutions produced by the leaching of nickel or nickel – copper mattes. Sherritt Gordon in Canada, Western Mining in Australia, and Impala Platinum in the Republic of South Africa recover refined nickel powder from purified ammoniacal solution by reduction with hydrogen.

6.1. Electrorefining [1]

Nickel electrorefining is carried out in a tank house containing a large number of electrolytic cells, each filled with a mixed nickel sulfate – nickel chloride solution, and containing 30 – 40 anodes (slabs of metallic nickel or nickel matte), interspersed with cathodes that are thin sheets of pure nickel. The anodes and cathodes in each cell are connected electrically in parallel, and the cells are connected in series. When electrical current is passed through the cells the anodes dissolve, and pure nickel deposits on the cathodes. Cathodes and anodes are replaced periodically to maintain continuity of production.

Crude nickel anodes are made by reducing nickel oxide with coke at temperatures up to 1540 °C and casting the molten metal in molds. The pure nickel starting sheets are made in special electrolytic cells in which nickel is deposited onto permanent stainless steel or titanium cathode blanks, from which the thin sheet is stripped after 24 or 48 h of deposition. The sheets are fitted with suspension loops and placed in the production cells. The electrolytic cells are normally made of concrete and are lined with an acid-resistant membrane such as glass fiber reinforced plastic material.

Nickel refining can be carried out with pure sulfate, mixed sulfate – chloride, or pure chloride nickel electrolytes. In practice, the mixed sulfate – chloride system is normally used for the electrorefining of either nickel metal or nickel matte. Sulfate and chloride electrolytes are used in the nickel electrowinning processes. The sulfate – chloride electrolyte has two major advantages over the sulfate system, in having a higher conductivity which results in a lower cell voltage, and in permitting the use of chlorine for electrolyte purification.

6.1.1. *Refining of Nickel Metal Anodes*

In the refining of nickel metal anodes the principal reaction at the anode is the dissolution of nickel metal as nickel ions:

$$\mathrm{Ni(anode)} \rightarrow \mathrm{Ni^{2+}(soln)} + 2\,\mathrm{e^-(anode)}$$

The principal cathodic reaction is the reduction of nickel ions from solution:

$$\mathrm{Ni^{2+}(soln)} + 2\,\mathrm{e^-(cathode)} \rightarrow \mathrm{Ni(cathode)}$$

At the normal cell operating voltage (ca. 1.6 V for metal anodes) the principal impurities from the anode such as iron, cobalt, lead, arsenic, and copper go into solution with the nickel. Silver, gold, the platinum group metals, sulfur, selenium, and tellurium fall to the bottom of the cell as an insoluble slime. If the soluble impurities reach the cathode surface they will codeposit with the nickel. Consequently, to avoid contamination of the refined nickel cathodes, the impure anolyte solution is taken through a purification process. The purified solution is then returned to the cathode compartment of the cell, which is segregated from the anode by a porous diaphragm in the form of a bag or box enclosing the cathode. The purified solution is fed into the cathode compartment at a rate sufficient to maintain the rate of nickel deposition. The diaphragm cloth is tightly woven so that a hydrostatic head of solution builds up relative to the rest of the electrolyte. This hydrostatic head forces nickel-depleted electrolyte through the diaphragm, from the cathode compartment into the anolyte, at a rate sufficient to prevent the ingress of either anolyte itself, or impurity ions, by diffusion into the cathode compartment.

6.1.2. *Refining of Nickel Matte Anodes*

Nickel sulfide anodes can be electrorefined directly to pure nickel cathode without the intermediate process steps of roasting the sulfide to oxide and reducing the oxide to produce metallic nickel anodes. Nickel sulfide anodes can be cast directly from low-copper converter matte or from melted nickel sulfide concentrate produced by the matte separation process. Controlled

cooling is necessary to produce anodes with the required mechanical properties. The anodes consist essentially of Ni_3S_2 and nickel alloy, but can contain up to 3 % copper as Cu_2S and other impurities such as iron, cobalt, lead, zinc, arsenic, selenium, and platinum group metals. Control of the sulfur content of the anodes is important because at sulfur contents of less than 15 %, the metallic nickel alloy phase dissolves preferentially, leaving nickel sulfide in the slime. At sulfur contents above 15 % , the matte anodes corrode uniformly to yield a slime containing elemental sulfur.

The principal anodic reactions are the oxidation of sulfide to elemental sulfur with simultaneous release of metal ions to the solution and of electrons to the anode:

$$Ni_3S_2 \rightarrow 3\,Ni^{2+}(soln) + 2\,S(slime) + 6\,e^-(anode)$$

The cathodic reaction is the same as in the metal anode refining process:

$$Ni^{2+}(soln) + 2\,e^-(cathode) \rightarrow Ni(cathode)$$

The net cell reaction for the nickel sulfide phase is thus

$$Ni_3S_2 \rightarrow 3\,Ni + 2\,S$$

The metallic nickel phase dissolves electrolytically as in the refining of nickel metal anodes:

$$Ni(anode) \rightarrow Ni^{2+}(soln) + 2\,e^-(anode)$$

The actual cell voltage, including the potential drops across the contacts and through the electrolyte, is about 2.8 V for newly charged matte anodes. As the anodes corrode, the sludge, which consists mainly of elemental sulfur, forms a strongly adherent porous surface layer which causes an increase in volume and effectively doubles the bulk of the anode. As the sludge layer becomes thicker, the electrical resistance between the electrodes increases, and the cell voltage increases to about 4.0 V by the end of the anode cycle. The increase in cell resistance also results in increased heating of the electrolyte. The solution temperature rise in the cell increases from about 15 °C at the start of the anode cycle to about 25 °C at the end of the cycle.

The anode potential is sufficiently high that some water is decomposed:

$$2\,H_2O \rightarrow O_2 + 4\,H^+ + 4\,e^-$$

This unwanted reaction, together with the electrolytic dissolution of impurities from the anode, lowers the anode nickel dissolution efficiency so that nickel is removed from solution at the cathode faster than it dissolves at the anode. The imbalance is corrected during electrolyte purification by dissolving matte in the electrolyte to neutralize the free acid and replenish the nickel content.

Operation of the nickel matte refining cell is similar to that of the metal anode cell except that the anodes are enclosed in bags to collect the voluminous sludge. The anode spacing in the matte refining cell is greater than in the metal refining cell, to allow for the increase in volume of the matte anode as it corrodes.

6.1.3. *Solution Purification [1], [78]*

For the production of high-quality nickel cathode it is essential that the anolyte solution be purified before it is supplied to the cathode compartment. The purification methods used depend on the impurities present in the anode.

Copper is often removed by cementation with activated metallic nickel powder, causing copper ions to precipitate as metallic copper while metallic nickel dissolves:

$$Ni + CuSO_4 \rightarrow Cu + NiSO_4$$

Alternatively copper may be precipitated as the sulfide with hydrogen sulfide:

$$CuSO_4 + H_2S \rightarrow CuS + H_2SO_4$$

This procedure, which is used in the matte electrorefining process, also precipitates arsenic:

$$2\,H_3AsO_3 + 3\,H_2S \rightarrow As_2S_3 + 6\,H_2O$$

Careful control of the redox potential is necessary to ensure complete precipitation of copper and arsenic while minimizing the coprecipitation of nickel.

Iron is normally removed by treating the solution with chlorine and nickel carbonate to oxidize the iron to iron(III) and precipitate iron(III) hydroxide by hydrolysis:

$$Fe_2(SO_4)_3 + 6\,H_2O \rightarrow 2\,Fe(OH_3) + 3\,H_2SO_4$$

$$3\,NiCO_3 + 3\,H_2SO_4 \rightarrow 3\,NiSO_4 + 3\,H_2O + 3\,CO_2$$

Iron(III) hydroxide coprecipitates other impurities from the anolyte, particularly lead and arsenic. In the pure sulfate electrolyte systems, oxidation of iron is accomplished by using air instead of chlorine. Iron can be removed from pure chloride

electrolytes by solvent extraction with tributylphosphate.

Cobalt is precipitated as cobalt(III) hydroxide by treating the solution with chlorine or nickel(III) hydroxide to oxidize the cobalt (II) to cobalt(III). Nickel carbonate is added to control the pH.

$$2\,CoSO_4 + Cl_2 + 3\,NiCO_3 + 3\,H_2O \rightarrow 2\,Co(OH)_3 + 2\,NiSO_4 + NiCl_2 + 3\,CO_2$$

Most of the cobalt precipitates at pH 4 at an oxidation potential of + 1 V. Chlorine cannot be used in pure sulfate electrolyte, and in this system, nickel(III) hydroxide is normally used. Other oxidants which can be used to remove cobalt from the sulfate system include peroxide, persulfate, and ozone.

Cobalt is normally removed from pure chloride nickel electrolytes by solvent extraction with a tertiary amine such as triisooctylamine. Phosphine-based solvent extraction reagents such as di-2-ethylhexylphosphoric acid (D2EHPA) are used to extract cobalt from nickel sulfate solutions.

In practice the purified solutions used in nickel electrorefining are not simple mixtures of nickel sulfate and nickel chloride. The electrolyte is a weakly acidic aqueous solution of nickel, sulfate, sodium, and chloride ions, and boric acid. The sodium and chloride ions carry about 30 % of the current through the electrolyte. Most of the remaining current is carried by the slower moving divalent nickel and sulfate ions, since the concentration of hydrogen ions is negligible. The boric acid, which exists largely as undissociated molecules, acts as a buffer to maintain the pH of the solution below 6. The presence of boric acid thus prevents the precipitation of nickel(II) hydroxide at the surface of the cathode by the reaction between nickel ions and hydroxide produced in the minor cathodic reaction involving the electrolytic decomposition of water.

6.1.4. *Electrorefining Operations*

Although it is not currently operated in North America or Western Europe, the refining of metallic nickel anodes is still practiced in the former Soviet Union. The Severonikel complex at Monchegorsk in the Kola peninsula has the capacity to produce 142 000 t/a of electrolytic nickel [44]. The Monchegorsk smelter treats a blend of high-grade nickel – copper ores, concentrates, mattes, and scrap from the nickel mining operations at Norilsk and Pechenga. Separate nickel and copper sulfide concentrates are produced by the slow-cooling matte separation process. The nickel sulfide is dead roasted at 1170 °C with oxygen-enriched air (25 % O_2), and the oxide calcine is reduced with coke in a rotary kiln to form a metallized oxide, which is melted at 1550 °C in an electric arc furnace for anode casting. About 3 % of the metallized oxide powder from the reduction kiln is retained for use in electrolyte purification. The anodes typically weigh 350 kg and analyze (wt %): 86 – 88 Ni, 2 Co, 4.5 Fe, 4 Cu, and 0.2 S.

The anodes are refined in an electrolyte containing 75 – 78 g/L Ni, 130 – 150 g/L SO_4^{2-}, 14 g/L Cl^-, and 6 g/L H_3BO_3, at pH 2. Each electrolytic cell contains 43 anodes and 42 cathodes, and operates at a current density of 250 A/m^2. The cathodes are enclosed in a synthetic fiber bag supported on a polypropylene frame. Starting sheets are produced by plating nickel on to titanium blanks for 24 h. The sheets are stripped manually. The anode cycle is 30 d, while the cathode cycle is four or six days. Anode slimes are recovered by washing adhering material off the corroded anodes, and from the bottom of the electrolytic cells, at the end of each anode cycle, for subsequent processing to recover precious metal values.

Impure anolyte solution is treated to remove iron, copper, and cobalt, and to replenish the nickel content of the electrolyte, before the purified solution is returned to the cathode compartments. Iron is removed by controlled oxidation with chlorine and neutralization with nickel carbonate. Copper is precipitated by cementation on nickel powder from the reduction kiln, and cobalt is precipitated by oxidation with chlorine and hydrolysis with nickel carbonate under highly oxidizing conditions. The nickel content of the solution is replenished by pressure leaching nickel sulfide concentrate (Ni_3S_2) in sulfuric acid solution under 1 MPa air pressure in vertical titanium autoclaves.

The electrorefining of nickel sulfide anodes is carried out by Inco at its Thompson operation in Canada, by Sumitomo Metal Mining Company at Niihama in Japan, and at Jianchuan in China.

The Thompson smelter [76] produces sulfide anodes by casting molten converter matte (73 % Ni, 0.8 % Co, 2.5 – 3.0 % Cu, 0.5 % Fe, and 20 % S) directly into anodes weighing ca. 240 kg. The cast anodes are cooled slowly over about 36 h to improve their mechanical handling characteristics, but are nevertheless more brittle than metal anodes and require careful handling.

The sulfide anodes are electrolyzed in diaphragm cells in a sulfate – chloride electrolyte. Nickel, cobalt, and copper dissolve from the anode, while sulfur, selenium, and the noble metals form an insoluble sludge or slime, from which they can be recovered. Nickel is plated on to pure nickel starting sheets, made by manually stripping 24-h deposits from stainless steel blanks. The anode cycle is 15 d and the cathode cycle is 10 d. Plating is carried out at a current density of 240 A/m^2. Details of the tankhouse operating parameters are as follows [76]:

Capacity	45 000 t/a Ni
Cells	608
Cell dimensions	1.6 m×0.9 m×5.8 m
Cell construction	fiber reinforced plastic lined precast concrete
Anode weight	238 kg
Anode dimensions	1.1 m×0.7 m×63 mm

Anode spacing	210 mm
Anodes/cell	27
Anode cycle	15 d
Cathode dimensions	1.0 m×0.7 m×13 mm
Cathode weight	88.5 kg
Cathodes/cell	26
Cathode cycle	10 d
Electrolyte flow	0.64 m^3/h per cell
Electrolyte temperature	50 °C
Current density	240 A/m^2
Cell voltage	3 – 6 V

The anolyte purification circuits at Thompson remove copper and arsenic by precipitation with hydrogen sulfide. The nickel content of the electrolyte is then replenished by reacting the solution with ground matte anode scrap under oxidizing conditions. For every 100 kg of nickel plated out in the tankhouse only 87 kg dissolves from the anode. This discrepancy is caused by the dissolution of 4 kg of copper, 1 kg of cobalt, 0.5 kg each of iron and arsenic, and by the formation of 0.2 kg of hydrogen ions. In the final stage of purification cobalt and iron are oxidized and hydrolyzed by the addition of chlorine and nickel carbonate. The cobalt(III) hydroxide precipitate is worked up for cobalt recovery. The purified electrolyte typically contains 75 g/L Ni, 120 $g/L\ SO_4^{2-}$, 28 g/L Na, 50 g/L Cl^-, and 8 g/L boric acid.

Until 1984, the principal products of the Thompson refinery were full nickel cathodes and cut shapes, made by shearing cathodes grown on nickel starting sheets. Now about 40 % of the refinery's production is “S” and “R” Rounds for the plating industry, which are made by plating nickel directly on to stainless steel blanks, covered by a heat-set epoxy mask, with circular areas of steel exposed to plating. Metallurgical data for the Thompson refinery, including the cathode nickel analysis, are given in Table 11.

Table 11. Metallurgical data for the Inco Thompson refinery [76]

Element	Anode, %	Anolyte, g/L	Purified electrolyte, g/L	Electronickel, g/t
Ni	73		75	
Co	0.8	0.13	0.02	600
Cu	3.0	0.50	0.0007	25
As	0.2	0.03	0.0005	30
Fe	0.6	0.03		5
Pb	0.003			2
S	20			7

6.2. Electrowinning

Electrowinning differs from electrorefining in that it is used to recover nickel from a leach liquor and can therefore be carried out with an insoluble anode, since the only function of the anode is to transfer electrons from the electrolyte to the external circuit. The electrowinning of nickel is practiced commercially with both pure sulfate and pure chloride electrolytes. In the sulfate system the anodes are made of chemical lead or antimonial lead, which have a relatively long life in sulfate solution, although soluble in chloride solutions. In the highly corrosive chloride system, dimensionally stable anodes (DSA), made by coating a titanium substrate with a platinum group metal oxide, are used.

The principal anodic reaction in the sulfate system is the decomposition of water to produce gaseous oxygen:

$$2\,H_2O \rightarrow O_2 + 4\,H^+ + 4\,e^-$$

The main cathodic reaction, as in electrorefining, is the deposition of nickel ions to form nickel atoms on the cathode:

$$Ni^{2+} + 2\,e^- \rightarrow Ni$$

The overall cell reaction is:

$$2\,NiSO_4 + 2\,H_2O \rightarrow 2\,Ni + 2\,H_2SO_4 + O_2$$

The sulfuric acid generated at the anode is utilized by recirculating the anolyte to the leach circuits.

In the chloride system the principal anodic reaction is the oxidation of chloride ion to chlorine gas:

$$2\,Cl^- \rightarrow Cl_2 + 2\,e^-$$

The overall reaction is thus the decomposition of nickel(II) chloride to metallic nickel and chlorine gas:

$$NiCl_2 \rightarrow Ni + Cl_2$$

The chlorine liberated at the anode is recycled for use in the leach circuits. In practice some decomposition of water also occurs to yield small quantities of hydrogen at the cathode and oxygen at the anode.

6.2.1. *Electrowinning from Sulfate Electrolytes*

In a sulfate electrowinning system the cathodes are suspended in bags, even though the nickel sulfate solution is purified before it enters the tankhouse. The cathode must still be protected from the anolyte, which is strongly acidic due to anodic generation of hydrogen ions. A high concentration of hydrogen ions at the cathode would result in the formation of hydrogen gas which is formed in preference to nickel deposition, thus seriously reducing the current efficiency of the cell.

Details of the purification methods used to remove copper, iron, lead, and cobalt from nickel sulfate leach liquor prior to electrowinning are given in Chapter Hydrometallurgy of Nickel Concentrates and Mattes . Treatment of electrowinning anolyte or spent electrolyte varies depending on the type of leaching circuit used. In the Outokumpu atmospheric leach, in which very little sulfur is oxidized, almost all the anolyte can be recycled to the leaching circuits, where the acid is neutralized in leaching nickel from the matte, thus replenishing the nickel content of the solution. Only a small solution bleed is taken out of the circuit to control the concentration of sodium sulfate, which would otherwise build up to unacceptable levels.

In the Sherritt acid pressure leaching process, as practiced by Rustenburg Refiners, in which all the sulfur in the matte is oxidized to sulfate, a much larger bleed of solution is required to maintain a sulfur balance. Typically 25 – 35 % of the anolyte stream must be removed from the circuit and treated to recover sodium sulfate. This bleed stream is reacted with sodium hydroxide or sodium carbonate to neutralize the acid and precipitate the remaining nickel as nickel(II) hydroxide or as basic nickel carbonate:

$$H_2SO_4 + 2\,NaOH \rightarrow Na_2SO_4 + 2\,H_2O$$

$$NiSO_4 + 2\,NaOH \rightarrow Ni(OH)_2 + Na_2SO_4$$

The nickel-free barren solution is evaporated to crystallize sodium sulfate. The precipitated nickel hydroxide or basic nickel carbonate is redissolved in a second portion of the anolyte stream to neutralize the acid content and replenish the nickel content:

$$Ni(OH)_2 + H_2SO_4 \rightarrow NiSO_4 + 2\,H_2O$$

This solution is recycled directly as feed solution to the electrowinning circuit. The balance of the anolyte is recycled for use in the matte leach circuits, where the acid is neutralized and the nickel content replenished, and passes through the leach solution purification steps before returning to the tankhouse.

The Outokumpu nickel refinery at Harjavalta in Finland contains 126 electrowinning cells, each having 40 insoluble lead anodes and 39 cathodes [1], [48], [67]. The bagged cathodes are nickel starting sheets made by the deposition of nickel for 48 h onto a stainless steel mother blank. The deposits are stripped manually and automatically made up into starting sheets. The purified solution is fed to the cathode compartments at a rate sufficient to maintain a hydrostatic head of about 30 mm. The current density is 200 A/m^2, giving a cell voltage of about 3.6 V. The cathode cycle is seven days. The cells are hooded to collect and remove the oxygen and acid mist generated at the anode.

The feed solution at pH 3.2, containing 75 g/L Ni as well as sodium sulfate and boric acid, is heated to 60 °C and fed separately to each cathode compartment. The solution passes through the diaphragm cloth into the anolyte, which overflows from each end of the cell. The anolyte, which contains 50 g/L Ni and 40 g/L H_2SO_4, is recycled to the leaching circuits. The cathodes typically weigh ca. 75 kg and contain 99.95 % Ni.

6.2.2. *Electrowinning from Chloride Electrolytes*

In the Falconbridge chloride electrowinning system [73], the dimensionally stable anodes are enclosed in a polyester diaphragm bag, to contain the chlorine gas generated in the anode reaction, and channel it to a vacuum hood fitted to the top of the anode. Ducts, connecting the top of the hood with a manifold that runs alongside the tank, allow the withdrawal of both chlorine and anolyte by suction. Purified solution is fed in at one end of the cell, and the solution level is held constant by maintaining a small overflow at the other end, while the greater portion of the electrolyte is withdrawn as anolyte with the chlorine gas.

The Falconbridge refinery at Kristiansand, Norway, which has a capacity of 54 000 t/a Ni, contains 326 electrolytic cells, of which 24 are devoted to starting sheet production. Each cell contains 46 anodes and 45 cathodes with an anode spacing of 145 mm. Typical operating parameters are as follows:

Inlet flow	4.0 m^3/h per cell
Total current	24 000 A
Current density	220 A/m^2

Current efficiency	98 – 99 %
Electrolyte temperature	60 °C
Catholyte Ni	60 g/L
Anolyte Ni	54 g/L

The refinery produces both regular nickel cathodes and Falconbridge crowns. The regular cathodes are produced by plating onto nickel starting sheets and are sheared to a variety of sizes for the nickel plating industry. Crowns weighing 30 – 60 g are produced by plating nickel onto specially designed cathode blanks, from which they are stripped and polished for market. The refinery produces a limited amount of exceptionally pure superelectro nickel, with less than 100 g/t of impurities including oxygen, and less than 5 g/t Co, by using a specially purified electrolyte.

6.3. Carbonyl Refining [1], [79], [80]

The reaction of carbon monoxide at atmospheric pressure with active nickel metal at 40 – 80 °C to form the gaseous nickel tetracarbonyl was discovered by LANGER and MOND in 1889 [75]. The reaction is readily reversible, with nickel tetracarbonyl decomposing to metallic nickel and carbon monoxide at 150 – 300 °C.

Nickel tetracarbonyl is a volatile liquid that melts at –19.3 °C and boils at 42.5 °C.

Under the mild conditions employed for reaction at atmospheric pressure, the carbonyl-forming impurities in crude nickel metal do not volatilize. Iron forms a volatile carbonyl, iron pentacarbonyl, $Fe(CO)_5$, with a freezing point of –20.5 °C and a boiling point of 103 °C, but the rate of formation is slow. Cobalt forms $Co_2(CO)_8$, which melts at 51 °C and decomposes at 52 °C to form $Co_4(CO)_{12}$, but both are solids with low volatility. Copper, like most other elements, does not form carbonyls directly with carbon monoxide. Thus the extraction of nickel as a carbonyl from a crude metal feed is a highly selective process.

The formation of metal carbonyls by the interaction of carbon monoxide with reduced metals can be catalyzed. In fact if it were not for the large increase in reaction rate resulting from the addition of a controlled amount of catalyst, the tonnage extraction of nickel and iron as carbonyls from reduced metals would probably not be feasible. With the aid of catalysts or inhibitors it is possible to produce the carbonyls of nickel, cobalt, and iron either selectively or together.

Although catalysts such as ammonia, mercury, selenium, and tellurium have been used to increase the rate of nickel carbonyl formation, sulfide activation is normally used. Sulfur is a good, low cost catalyst which can be added in the form of hydrogen sulfide, carbonyl sulfide, or sulfur dioxide. It may be added during the reduction step or prior to volatilization. The carbonyl extraction process is notable not only for its selectivity in the volatilization of the metals as carbonyls but also for the relative ease with which the carbonyls can be separated and decomposed under mild conditions to produce high-purity metals. Nickel and iron carbonyls can be separated by simple fractional distillation as a result of the large difference in their boiling points.

The carbonyl process makes it possible to produce nickel of very high purity with an exceptionally low cobalt content, which makes it particularly useful for nuclear energy applications.

The thermal decomposition of nickel tetracarbonyl is carried out by two different methods. Very rapid heating of the gas phase produces a fine nickel powder. Alternatively if the carbonyl contacts a hot metal surface, nickel deposits on the surface. This effect is used to produce nickel granules by the successive deposition of layers of nickel onto a seed particle, until the required size is attained.

6.3.1. *Atmospheric Pressure Carbonyl Process [1], [75]*

The Inco nickel refinery at Clydach, Wales, began operation in 1902 using the Langer – Mond atmospheric pressure carbonyl process. Originally the plant treated nickel – copper matte, but it now processes a granular nickel oxide, produced by fluidized-bed roasting of nickel sulfide at Inco's Copper Cliff smelter, which typically analyzes 74 % Ni, 2.5 % Cu, 1.0 % Co, 0.3 % Fe, and 0.1 % S.

The refinery still uses the basic Langer – Mond process but the operation has been greatly increased in efficiency over the years. The first three steps in the process—reduction, sulfide activation, and volatilization—are now carried out in rotary kilns. The nickel oxide is first reduced to metal at 425 °C by countercurrent contact with preheated hydrogen gas. The metal is then sulfided in a smaller kiln to activate it, before being contacted countercurrently with carbon monoxide at 50 – 60 °C and atmospheric pressure in a third kiln. About 95 % of the nickel volatilizes under these conditions but iron and cobalt are not carbonylated. The residue from the volatilizers is returned to Canada and forms part of the feed to the converters at the Copper Cliff nickel refinery.

The off-gas from the volatilizer kiln, which contains about 16 vol % nickel carbonyl, is fed to a pellet decomposer (Fig. 13) where it is contacted with preheated nickel granules at 200 °C. The nickel carbonyl decomposes, releasing carbon monoxide and depositing a layer of metallic nickel on the granules, which steadily increase in size.

Figure 13. Cutaway view of a pellet decomposer

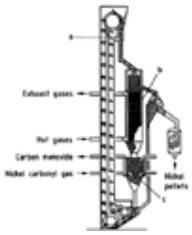

a) Bucket elevator; b) Preheater chamber; c) Reaction chamber

The decomposer is filled with about 30 t of pellets which flow downwards by gravity and are recirculated to the top of the vessel by an enclosed bucket elevator. As the pellets pass downwards they are heated to 200 °C before contacting the carbonyl-containing gas introduced at the bottom of the decomposer. The motion of the pellets prevents them from adhering together. Nickel powder is added periodically to nucleate new pellets, and market-size pellets (8 mm) are automatically segregated from the smaller material and discharged. The off-gas from the decomposer contains ca. 93 vol % carbon monoxide and 0.1 % nickel carbonyl, and is recycled to the volatilization kiln. Nickel pellets from the carbonyl process typically contain 99.97 % Ni and less than 5 g/t Co. In addition to nickel pellets, carbonyl nickel powder is also produced at Clydach in electrically heated powder decomposers and accounts for ca. 20 % of the refinery nickel output.

6.3.2. *High-Pressure Carbonyl Process (BASF)*

Extraction of nickel with carbon monoxide is also used in processes that operate at high pressure and produce liquid nickel carbonyl. The volume change that occurs in the carbonyl reaction:

$$Ni + 4\,CO \rightarrow Ni(CO)_4$$

with four molecules of carbon monoxide forming one molecule of nickel carbonyl, means that an increase in pressure accelerates the formation of the carbonyl. The increased pressure stabilizes the carbonyl and thus permits the process to be carried out at high temperature, which further increases the rate of reaction. As a result, activation of the nickel feed material is no longer necessary, and a wider range of feeds can be processed.

BASF in Germany operated a high-pressure carbonyl nickel refining process from 1932 until 1964. The feed materials were nickel – copper matte or nickel scrap and residues. The feeds were melted and the composition was adjusted to provide just enough sulfur to combine with copper as Cu_2S and with part of the iron. The molten feed was granulated and reacted with carbon monoxide at 230 °C and 20 MPa. Over 95 % of the nickel was extracted as nickel carbonyl in a three-day batch treatment. Some iron was also carbonylated. Pure nickel powder was obtained by fractionation of the liquid carbonyls, and rapid heating of nickel carbonyl vapor to 280 –300 °C.

6.3.3. *Inco Pressure Carbonyl Process [80], [81]*

The newest application of carbonyl refining to the treatment of a nickel- containing feed material is Inco's nickel refining complex in Copper Cliff, Ontario, which was commissioned in 1973. This plant has a nominal capacity of 57 000 t/a of refined nickel.

The refinery consists of two operating plants: the converter plant and the pressure carbonyl plant. The converter plant produces granulated metallic nickel with a controlled sulfur content, using two top-blown rotary converters (TBRC's). The pressure carbonyl plant produces 45 000 t/a of nickel pellets, 9000 t/a of nickel powder, and 2200 t/a of ferronickel byproduct from the granulated, sulfided nickel feed.

The principal feed materials supplied to the refinery by the Copper Cliff smelter are the metallics fraction from the matte separation process (Section Inco Matte Separation Process) and the less pure portions of the nickel oxide made by roasting the nickel sulfide concentrate (Section Fluidized-Bed Roasting of Nickel Sulfide) as well as various other nickel-bearing residues and intermediates. The novel feature of this process is the preparation of a nickel feed material which could be carbonylated at moderate pressure without the need for a separate activation step.

Figure 14 shows the process flow sheet for the preparation of the feed material in the converter plant. The feed materials are blended and smelted in two 50 t capacity top-blown rotary converters (Fig. 15). Smelting is a batch operation carried out in three phases: charging/melting, blowing (sulfur removal), and reduction. In the charging/melting phase, blended feed materials are charged as the furnace rotates and is fired with a natural gas – oxygen flame from a watercooled lance. Coke is blended with the feed to reduce its oxide content, and the oxygen content of the flame is adjusted depending on the amount of sulfur to be removed. Charging is finished when the oxides have been reduced and the temperature of the melt is about 1500 °C. During the charging and blowing phase the furnace is rotated to minimize dust emissions.

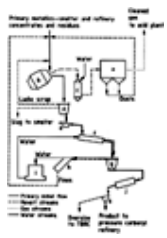

Figure 14. Flow sheet of the Inco pressure carbonyl converter plant

a) Electrostatic precipitator; b) Top-blown rotary converter (TBRC); c) Evaporation chamber; d) Bottom teeming ladle; e) Tundish; f) Granulating sluice; g) Dewatering bins; h) Lamella thickener; i) Cooling tower; j) Rotary dryer; k) Screen

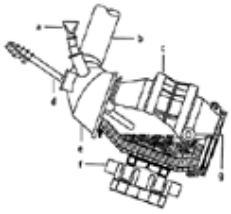

Figure 15. Cutaway view of a top-blown rotary converter

a) Water-cooled charging chute; b) Flue; c) Track ring; d) Water-cooled oxygen/natural gas lance; e) Hood; f) Rotate drive motor; g) Thrust roller

In the blowing phase, a precalculated amount of high-pressure oxygen is blown at supersonic velocities into the converter charge to oxidize sufficient sulfur to produce a matte with a Cu/S ratio of about 4 : 1 (Cu_2S). In the reduction phase, oxides formed during blowing and oxygen dissolved in the bath are reduced with petroleum coke. Some sulfur removal also occurs when oxygen dissolved in the matte reacts with remaining sulfides. The converter is rotated at 15 rpm during reduction. Off-gases are drawn out through a movable hood and cleaned in electrostatic precipitators before being vented to a stack. The recovered dusts are recycled to a subsequent converter charge.

Molten matte at 1600 °C is granulated with high-pressure water jets at a rate of 1 t/min. The granulated matte is dried and transferred to the pressure carbonyl plant. The matte typically analyzes 75 – 80 % Ni, 12 – 17 % Cu, 2 % Co, 2 % Fe, and 3 – 4 % S, with a granule size of less than 9 mm.

Figure 16 shows the process flow sheet for the pressure carbonyl refining operation. Matte granules are carbonylated in 150-t batches at temperatures up to 180 °C and pressures of 7 MPa in three horizontal cylindrical rotating reactors (Fig. 17). Carbon monoxide is continually circulated through the reactor. Nickel carbonyl formation is very exothermic and the heat of reaction is removed by means of water- cooled heat exchanger tubes inside the reactors, which also serve as lifters to improve gas – solids contact as the vessel rotates. Micro-metallic filters at each end of the reactor prevent loss of solids entrained in the gas stream. The direction of gas flow is reversed periodically to prevent filter blinding.

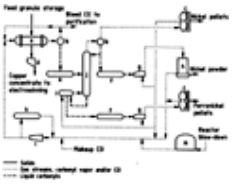

Figure 16. Flow sheet of the Inco pressure carbonyl refining process

a) Reactors; b) Condenser; c) Pellet decomposers; d) Slurry; e) Crude liquid storage; f) Storage; g) Vaporizer; h) Powder decomposers; i) Reboiler; j) Distillation columns; k) High-pressure CO storage; l) Main compressors; m) Low-pressure CO storage

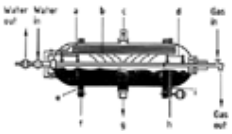

Figure 17. Cutaway view of a pressure carbonyl reactor

a) Coils; b) Flights; c) Feed chute; d) Metallic filters; e) Tires; f) Rollers; g) Discharge chute; h) Drive pinion; i) Spur gear

Nickel and iron carbonyls are carried out of the reactor with the carbon monoxide stream, are liquified by cooling, and stored as liquids at atmospheric pressure. The carbonylation reaction normally extracts about 97.5 % of the nickel and 30 % of the iron, leaving a residue typically containing 57 % Cu, 8.5 % Ni, 9 % Co, 6 % Fe, and 15 % S, as well as the precious metals. The residue is treated hydrometallurgically to recover copper, cobalt, nickel, and the precious metals [82].

The crude, liquid nickel – iron carbonyls typically contain 98.9 % Ni and 1.1 % Fe. This liquid is fractionated in two columns, each with twenty trays, to separate nickel carbonyl (*bp* 43 °C) from iron carbonyl (*bp* 103 °C). Distillation is carried out at below boiling temperatures with carbon monoxide as a carrier gas. Nickel carbonyl, which is obtained as a mixture of vapor and liquid from the overhead partial condensers, is used to make extremely pure nickel powder and pellets. The bottoms from the reboilers, which contain 85 % iron carbonyl and 15 % nickel carbonyl liquids, are decomposed to form ferronickel pellets.

The nickel carbonyl vapor from the distillation column passes directly to pellet decomposers (see Fig. 13) where it is contacted with preheated nickel pellets at 240 – 270 °C. The operation of the decomposers is essentially as described in Section High-Pressure Carbonyl Process (BASF) for the Inco refinery at Clydach. The condensed portion of the nickel carbonyl distillation overhead stream is revaporized with carbon monoxide carrier gas to avoid premature decomposition and passed to electrically heated powder decomposers (Fig. 18). Decomposition by thermal shock at 300 °C produces pure nickel powder, which can be varied in shape, size, morphology, and density by adjusting the operating conditions.

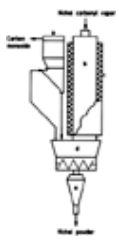

Figure 18. Cutaway view of a powder decomposer

a) Baghouse; b) Decomposer; c) Electric heaters; d) Connecting box; e) Discharge container

The iron carbonyl-containing bottoms product from distillation is treated by revaporizing the liquid at a sub-boiling temperature with carbon monoxide as carrier gas, and feeding it to a pellet decomposer to produce ferronickel pellets.

The pure nickel pellets, with a typical diameter of 10 mm, analyze over 99.97 % Ni with 60 g/t C, 40 g/t O_2, and less than 0.2 g/t Co. The nickel powder, with a particle size in the range 3 – 7 µm, has significantly higher carbon and oxygen contents (650 g/t C and 570 g/t O_2).

6.4. Hydrogen Reduction to Nickel Powder [1], [65]

Nickel is recovered as a pure metal powder from nickel sulfate leach solutions by chemical reduction with hydrogen under pressure in several hydrometallurgical nickel refineries. Precipitation of the metal is carried out with an ammoniacal nickel sulfate solution having a molar ratio of ammonia to nickel of 2 : 1, so that the acid generated in the reduction reaction is neutralized by reaction with ammonia to form ammonium sulfate:

$$[Ni(NH_3)_2(H_2O)_4]SO_4 + H_2 \rightarrow Ni + (NH_4)_2SO_4 + 4\,H_2O$$

The reduction process is operated at 200 °C at a hydrogen pressure of 3 MPa in horizontal single-compartment autoclaves, each equipped with four agitators. The solution is processed batchwise, with the powder particles growing as successive layers of nickel are reduced, onto seed particles, from up to sixty batches of nickel-bearing feed solution. After each reduction or densification, which takes up to thirty minutes, the powder is allowed to settle, and the depleted solution is decanted off and replaced by fresh feed solution.

The nickel precipitation reaction can occur on any metal surface, and in the absence of suitable seed particles most of the metal deposition occurs on the vessel walls as a solid plate. When suitable seed particles such as fine nickel powder are available, most of the nickel deposits on these particles, but even so, a minor but significant amount of nickel still plates on the walls, and must be releached periodically. Normally this plating leach is carried out after each cycle of 50 – 60 densifications, and the leach liquor produced is used to prepare the seed particles for the subsequent densification cycle.

Seed particles are produced by adding an aqueous solution of iron(II) sulfate to the plating leach liquor to form a finely divided precipitate of ferrous hydroxide, and applying a hydrogen overpressure of 2.5 MPa at 115 °C. Under these conditions the nickel deposits onto the nuclei of ferrous hydroxide to form very fine particles of nickel metal, which provide the seed particles for densification. Once the nickel seed particles are formed, subsequent nickel deposition occurs autocatalytically on the surface of the particles, in preference to the vessel walls.

The reduction reaction is selective for nickel from a solution containing 50 g/L Ni and 1 g/L Co until the nickel content drops to about 1 g/L. Reduction is normally stopped at this point, leaving a solution containing about 1 g/L Ni and 1 g/L Co. This solution is reacted with hydrogen sulfide to precipitate a mixed nickel – cobalt sulfide, which is treated hydrometallurgically to recover refined cobalt (Section Ammonia Pressure Leaching), while the barren solution is evaporated to crystallize ammonium sulfate for sale as fertilizer.

At the end of a densification cycle when the powder has reached the desired particle size and density, the whole contents of the autoclave are discharged under pressure to a cone-bottomed flash tank. The solution is drawn off and the nickel powder is withdrawn from the bottom of the tank as a slurry containing 95 % solids. The powder is washed with water and collected on a vacuum filter. The filter cake is dried in a rotary dryer fired with natural gas. The powder is either marketed as such or is converted to briquettes by mixing with an organic binder and pressing in briquetting rolls. The “green” briquettes are sintered and desulfurized at 950 °C in a reducing atmosphere.

The powder typically analyzes 99.8 % Ni, 0.08 % Co, < 0.001 % Cu, 0.002 % Fe, 0.001 % C, 0.025 % S, and 0.02 % O_2.

This process for recovering nickel powder from solution can be used in combination with the ammonia pressure leaching of nickel concentrates and mattes (as at Sherritt Gordon and Western Mining Corporation), with the acid pressure leaching of nickel – copper mattes (as at Impala Platinum and in the former Amax operation at Port Nickel), and with the Caron process for treating nickel laterites (as at Nonoc in the Philippines).

[Top of Page]

7. Beneficiation of Oxide Ores [1]

As described in Chapter Occurrence, nickel oxide ores occur in two types: a limonitic type which occurs in the upper zone of a lateritic deposit, in which ferric oxide materials are predominant, and a silicate, or “garnierite” type, occurring at greater depth in the deposit, which has a lower iron content, but is usually richer in nickel than the limonite, and contains high levels of magnesia and silica. In nickeliferous limonite $(Fe,Ni)O(OH) \cdot nH_2O$ the nickel oxide is mainly present in solid solution with the iron oxides. In nickeliferous silicate, nickel, iron, and cobalt oxides occur in varying proportions, generally replacing part of the magnesium oxide in the serpentinite $[Mg_6Si_4O_{10}(OH)_8]$.

Most deposits contain both types of ore. Usually it is possible to separate the ore by selective mining and screening into an iron-containing limonite fraction and a magnesium silicate fraction enriched in nickel. The metallurgy of nickel oxide ores differs from that of sulfides in that oxide ores are not amenable to most of the standard mineral beneficiation methods due to the chemical dissemination of the nickel in the oxide minerals. Screening may be used to reject the oversize, less weathered fragments that contain less nickel, but this provides only a minor degree of upgrading.

In most cases the ore is treated directly by a pyrometallurgical or hydrometallurgical process. However, two processes have been developed in which the nickel is first converted by high-temperature treatment to a metallic phase, which can then be

separated from the bulk of the ore by standard mineral beneficiation techniques. These processes are the Nippon Yakin Oheyama process (which was based on the Krupp – Renn process developed in Germany in the 1930s to produce iron from otherwise uneconomic low-grade iron ores) and the MINPRO – PAMCO segregation process.

7.1. The Nippon Yakin Oheyama Process [83]

Nippon Yakin Kogyo produces 12 000 t/a nickel as low-nickel ferronickel from silicate ores (2.3 – 2.6 % Ni, 12 – 15 % Fe), imported from New Caledonia, the Philippines, and Indonesia, at its Oheyama works in Japan. The ore is ground and mixed with anthracite, coal, and flux, and briquetted. The briquettes are fed continuously through a rotary kiln, countercurrent to a flow of hot gas produced by the combustion of coal. As the temperature rises from 700 °C to 1300 °C, the charge is successively dried, dehydrated, and reduced. As the charge begins to melt, smelting occurs in the semifused state to form coalesced particles of spongy ferronickel (known as luppen, from the Krupp – Renn process) interspersed in a viscous slag phase. The kiln discharge is quenched in water and the luppen are separated from the slag by grinding, screening, jigging, and magnetic separation. Nickel recovery from ore is typically 95 %.

A major advantage of this process is that the ferronickel product is recovered in the form of 2 – 3 mm particles containing 22 % Ni + Co, 0.45 % S, 0.03 % C, and 0.02 % P which can be used directly in the AOD steel making process, without further upgrading or refining. Although the process is energy intensive, the major energy sources used are cheap anthracite and coal, and no high cost electrical energy or oil is used for heating. Consequently, energy costs are relatively low.

7.2. Segregation Processes [84], [85]

The segregation process was developed to recover copper from refractory minerals such as silicates, and has been applied with limited success on a commercial scale in Africa. Research has shown that the technique can be applied to the production of ferronickel from several types of nickel laterite ore, but the process has not yet been successfully applied on a commercial scale.

In the segregation process the ore is treated with a halide such as calcium chloride and a reductant such as coke at temperatures well below the melting range (950 – 1100 °C). Under these conditions the halide reacts with silicates and water vapor to form hydrogen chloride.

$$CaCl_2 + SiO_2 + H_2O \rightarrow CaSiO_3 + 2\,HCl$$

Iron(III) oxide is reduced to iron(II) oxide by hydrogen and carbon monoxide, which are formed in the reaction between coke and water.

$$C + H_2O \rightarrow CO + H_2$$

$$Fe_2O_3 + CO \rightarrow 2\,FeO + CO_2$$

$$Fe_2O_3 + H_2 \rightarrow 2\,FeO + H_2O$$

The gaseous hydrogen chloride chloridizes both nickel oxide and iron(II) oxide to form the volatile metal chlorides.

$$FeO + 2\,HCl \rightarrow FeCl_2 + H_2O$$

$$NiO + 2\,HCl \rightarrow NiCl_2 + H_2O$$

The metal chlorides are reduced to metal by reaction with hydrogen at the surface of the coke particles,

$$FeCl_2 + NiCl_2 + 2\,H_2 \rightarrow FeNi + 4\,HCl$$

releasing hydrogen chloride to react with further nickel and iron oxides. The ferronickel collects on the particles of coke and can be separated from the silicate phases by grinding and magnetic separation.

A major obstacle to the commercial development of this process has been the lack of a suitable reactor. The most promising approach, which was successfully tested at a rate of 1 t of ore per hour in 1981 – 1983, is the MINPRO – PAMCO segregation process. The reactor employed was a 850 kW former autogenous grinding mill, which was relined with two courses of refractory brick to permit operation at 950 °C.

In the pilot testwork, a nickel ore containing 2.3 % Ni and 13.2 % Fe was preheated in a rotary kiln to 950 °C, and treated in the segregation reactor with 40 kg calcium chloride and 20 kg coke per tonne of ore. The reactor discharge was cooled and ground prior to magnetic separation to yield a ferronickel product grading 59 % Ni and 29 % Fe, and a tailing with 0.45 % Ni. Further upgrading to 68 % Ni was shown to be possible with additional grinding and magnetic separation steps. Energy savings of 25 – 30 % over the conventional smelting process have been projected by the developers of the process.

[Top of Page]

8. Pyrometallurgy of Oxide Ores [1], [7], [8], [29]

Nickel laterite ores have historically been treated almost exclusively by smelting processes, and except for Cuba, where nickel ore treatment is all essentially hydrometallurgical, the industry is still predominantly pyrometallurgical. The development of nickel oxide ore smelting has drawn heavily on iron and steel metallurgy for ferronickel production and on copper metallurgy for nickel matte production.

The separation of nickel from the refractory oxides is relatively simple because there are large differences in the free energies of formation of nickel oxide and the gangue components such as silica and magnesia. Adjustment of the reduction conditions permits the complete reduction of nickel oxide while limiting the degree of reduction of iron oxide, but a total separation of nickel from iron by selective reduction is not possible. Two major process routes have been developed to overcome this problem.

In ferronickel production, the nickel oxide and part of the iron oxide are reduced to metal in an electric furnace. The fraction of iron reduced to metal is a function of the reduction potential of the system, so that ores with a low Ni/Fe ratio can be smelted to yield an acceptable ferronickel grade.

In nickel matte production, the ore is smelted with a reductant, a flux, and a source of sulfur, which may be gypsum, a sulfide mineral, or elemental sulfur. Most of the nickel and a large portion of the iron are reduced and sulfidized to form a low-sulfur matte phase. Iron removal can then be accomplished by fluxing and selective oxidation of the iron in a converter to produce a nickel matte containing over 75 % Ni, 20 – 22 % S, and less than 1 % Fe.

The smelting of nickel silicate ores in blast furnaces was developed by Societé Le Nickel (SLN) in New Caledonia in the 1870s [86]. Initially, attempts were made to produce ferronickel, but after encountering difficulties in eliminating sulfur from the product, the operation switched to the production of nickel – iron sulfide matte. The furnace matte was blown to give a low-iron converter matte, which was shipped to Europe or North America where it was dead roasted to remove sulfur and reduced in briquette form with charcoal to metal rondelles. The SLN introduced electric furnace smelting of nickel silicate ores to ferronickel in 1958, using a furnace design developed by Elkem – Spigerverket of Norway, and most present-day ferronickel smelters use processes based on this technology [87].

Several of the world's major nickel laterite smelters are listed in Table 12 together with details of their ore and product nickel grades and production capacities.

Table 12. Nickel laterite smelters

Smelter	Ore grade, % Ni	Product		1988 production, t/a Ni
		Type	Grade, % Ni	
SLN, New Caledonia	2.70	Fe – Ni	25	37 000
		matte	78	10 000
P.T. Inco Indonesia	1.97	matte	79	28 000
Pacific Metals, Japan	2.40	Fe – Ni	14 – 22	25 000 *
Falconbridge, Dominican Republic	1.75	Fe – Ni	38	29 300
Cerro Matoso, Colombia	2.90	Fe – Ni	45	19 000
LARCO, Greece	1.25	Fe – Ni	24 – 30	13 000
Hyuga Smelter, Japan	2.40	Fe – Ni	20 – 25	12 000 *
Orsk, Soviet Union	1.02	matte	77	16 000
Nippon Mining, Japan	2.40	Fe – Ni	20	10 000 *
Glogovac, Yugoslavia	1.32	Fe – Ni	24	5 500
P.T. Aneka Tambang, Indonesia	2.25	Fe – Ni	25	4 800

* 1985 production data.

8.1. Ore Pretreatment

Nickel laterite ores have high moisture contents (typically up to 45 %) as well as chemically bound water in the hydroxide form. All ore treatment plants use dryers, usually direct-fired rotary units operating at about 250 °C, which reduce the moisture content only to 15 – 20 % to minimize dusting during drying and subsequent smelting. Calcining to dehydrate the ore and prereduction prior to electric furnace smelting are generally carried out in countercurrently fired rotary kilns. Chemically bound water is released above ca. 400 °C, and reduction of the oxides to metal starts at 500 – 600 °C. The maximum temperature reached in the kiln is generally limited to 800 – 900 °C because of the tendency of the ore to agglomerate and form accretions on the furnace walls. Dehydration and prereduction of the ore in the kiln prior to smelting, optimizes the utilization of energy available from the reductants and fuel, thus reducing energy consumption in the smelting operation.

8.2. Smelting to Ferronickel

The rotary kiln – electric furnace smelting process is now used almost universally for the production of ferronickel from oxide ores. Nippon Mining shut down the last blast furnace producing ferronickel, at Saganoseki in Japan, in 1985 [88]. Variations on the electric furnace ferronickel process, which were developed in the 1950s by Ugine (France) and by Falconbridge in the 1960s, made it possible to achieve a more selective reduction of nickel, and thereby produce higher nickel grade products from lower grade ores.

The first commercial scale electric furnaces to produce ferronickel were the 10.5 MVA units installed by SLN at its Doniambo smelter in New Caledonia in 1958. The largest units (as of 1989) in operation are a 51 MVA unit at Cerro Matoso in Colombia and a 60 MVA furnace at Pacific Metals at Hachinohe in Japan.

8.2.1. *The Rotary Kiln – Electric Furnace Process [87]*

In the earliest plants, the function of the rotary kiln was limited to drying and dehydrating the ore, and furnace operating temperatures did not exceed 700 °C. The hot calcine was then blended with reductant, usually coal, and transferred to the electric furnace where reduction of the metals occurred.

In the electric furnace the calcined ore was smelted with the reductant to form immiscible layers of slag and ferronickel. The New Caledonian ores, with their high magnesia and silica contents (2.5 % Ni, 10 – 15 % Fe), require no additional flux for slag formation. Virtually all the nickel and 60 – 70 % of the iron in the ore are reduced to metal to yield a ferronickel grading about 20 % Ni; the slag contains only 0.1 % Ni. The energy consumption in the electric furnace for this mode of operation is 2.0 – 2.2 GJ per tonne of dry ore (550 – 600 kWh/t).

As the process was developed and applied to a wider range of nickel laterite ores, several improvements in design and operating technique were made. The operating temperature of the rotary kiln was increased to 900 – 1000 °C, and reductant was added to the kiln feed to ensure a more complete dehydration of the ore and to allow a partial reduction of the metals to occur in the kiln. As a result energy consumption in the electric furnace was reduced to 1.8 – 2.0 GJ/t (500 – 550 kWh/t).

The degree of reduction achieved in the kiln depends on the composition of the ore and on the reactivity of the reductant. Iron and nickel silicate minerals are generally less reactive than nickel oxide minerals. Usually lignites and charcoal are the most reactive reductants, and high-volatile coals are more reactive than low-volatile coals, anthracite, or coke. Typically, under optimum conditions, up to 40 % of the nickel is reduced to metal in the kiln, while the iron oxides are reduced to iron(II) oxide (FeO).

In the electric furnace the charge must be heated to 1400 – 1650 °C to permit the separation of distinct slag and metal phases. The energy input to the charge is provided by radiation from the electric arcs around the tips of the electrodes and by the heat generated in the slag by resistance heating due to the flow of current through the slag layer between the electrodes. Since the electrical resistivity of the metal layer is much lower than that of the slag, very little heat is generated in the metal layer, which must therefore receive heat by conduction and convection transfer from the overlying slag layer. For satisfactory heat transfer, the temperature differential between slag and metal layers should be at least 100 °C.

The mode of operation of the electric furnace is largely determined by the melting temperature of the slag. Except for the limonitic ores, where the alumina and lime contents have a significant effect, the slag melting temperatures for a particular ore can be estimated from the FeO – MgO – SiO_2 equilibrium phase diagram (Fig. 19) [89]. For SiO_2/MgO weight ratios of less than 2 : 1, the slag melting temperatures are essentially independent of the FeO content. At high SiO_2/MgO weight ratios the slags are close to saturation with SiO_2, and the slag melting temperatures are then a function of the FeO content (Fig. 20) [29]. High iron contents, which normally occur with low-magnesia ores, result in the formation of highly corrosive low-melting (1350 – 1450 °C) slags. Since the use of magnesia or silica fluxes is usually uneconomic, the problem of low-melting slags must be overcome in the design and operating mode of the furnace [90].

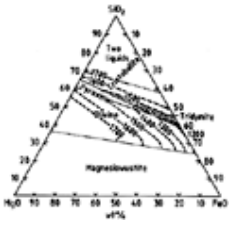

Figure 19. The FeO – MgO – SiO_2 phase diagram

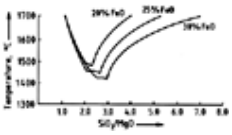

Figure 20. Slag melting point as a function of FeO content

Operation of the electric furnace is simplest when the slag melting temperature is higher than the metal melting temperature (1300 – 1400 °C). For such a system the furnace is operated with a covered bath (Fig. 21 A). The hot ore charge is allowed to build up on top of the molten slag, and the electrodes are not immersed in the slag layer. Under these conditions much of the reduction reaction occurs in the hot charge layer before it melts.

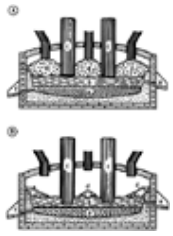

Figure 21. Electric furnace operation

A) Covered bath; B) Open bath

a) Metal pool; b) Slag bath; c) Electrodes; d) Solid charge; e) Slag tap hole; f) Metal tap hole

The high-iron limonite ores, which produce slags with melting points well below the melting point of the metal phase, can be smelted if the distance between the electrode tips and the slag-metal interface is reduced significantly. The electrodes must therefore penetrate deeply into the molten slag layer (Fig. 21 B). A vertical temperature gradient is then set up in the slag layer, with the highest temperature around the electrode tips. Under these conditions it is possible to operate the furnace with a metal tapping temperature higher than the slag tapping temperature. However the deep immersion of the electrodes in the slag results in a significant reaction between the carbon of the electrodes and metal oxides in the slag, with the evolution of carbon monoxide around the electrodes. The evolution of gases makes covered bath operation of the furnace impractical, and necessitates open bath operation (Fig. 21 B) with the slag surface exposed around the electrodes [87].

The ferronickel grade produced in the electric furnace smelting operation can be closely controlled by the amount of solid reductant added to the process. For a given addition of reductant, nickel recovery is highest, followed by cobalt, and iron recovery to the metal phase is lowest. Generally as much as 60 – 70 % of the iron is reduced from silicate ores, while only 10 – 15 % of the iron is reduced from high iron limonitic ores.

Normal operating practice is to produce up to 25 % Ni ferronickel from silicate ores and 15 – 20 % Ni ferronickel from limonitic ores. Nickel losses to the electric furnace slag under these conditions are quite low (0.1 % Ni in the slag). Nickel losses to slag increase with nickel grade, although grades with 35 – 45 % Ni can normally be produced from most ores without excessive losses (0.15 – 0.20 % Ni in the slag). Above 45 % Ni in ferronickel, slag losses increase rapidly. Where higher grades are required, the ferronickel electric furnace product can be upgraded by removing iron by oxidation and slagging in an oxygen-blown converter, and the high Ni content slag is recycled to the electric furnace. Ferronickel grades as high as 90 % nickel can be achieved with two stages of converter upgrading.

Current industrial practice is illustrated by the following descriptions of specific operations. In the Doniambo smelter of SLN in New Caledonia, which was built in 1958, the wet ores are carefully blended and screened to produce a smelter feed with an average composition of 2.5 – 3.0 % Ni and Co, 20 – 28 % MgO, 15 – 20 % Fe_2O_3, and 30 – 40 % SiO_2. The ore is blended with anthracite as a reductant and heated in a rotary kiln to 950 °C. Only a minor degree of reduction occurs in the kiln. The hot calcine is transferred to an electric furnace (one of eight 10.5 MVA or three 33 MVA units). Smelting is carried out with a charge covered bath, although the electrodes dip about 30 mm into the slag layer. Energy consumption is 90 MJ/kg Ni (2.35 GJ/t ore or 650 kWh/t ore) and electrode consumption is 1.5 kg/kg Ni (39 kg/t ore). Nickel and iron recoveries from ore to ferronickel product are 95 and 50 %, respectively. The slag tapping temperature is 1550 – 1600 °C, while the metal is tapped at 1450 °C. The slag, which typically contains 0.1 – 0.2 % Ni, is granulated and discarded [91], [92].

The crude ferronickel contains 24 % Ni, 69 % Fe, 2 % C, 3 % Si, 1.5 % Cu, and 0.25 % S. It is either cast into ingots for market, converted to nickel matte (Section Smelting to Nickel Matte), or desulfurized and refined to higher quality ferronickel. Ferronickel is marketed by SLN as 25 kg ingots and as granules.

Much the same process is used by the three Japanese ferronickel smelters, Pacific Metals at Hachinohe [93], Sumitomo at Hyuga [94], and Nippon Mining at Saganoseki, by Morro do Niquel in Brazil [95], by PT Aneka Tambang in Indonesia, and by Cerro Matoso in Colombia [96], [97]. All these operations process ores with relatively low Fe/Ni ratios of 6 : 1 or less. Representative ore and slag compositions are given in Table 13.

Table 13. Silicate ore and slag compositions [29]

Smelter	Ore composition, %				Slag composition, %		
	Ni	Fe	MgO	SiO_2	FeO	SiO_2	MgO
SLN	2.7	14 – 15	20 – 28	35 – 40	12	45	30
Hyuga	2.4	9 – 14	22 – 29	35 – 43	8	52	36
Morro do Niquel	1.25	6.5	30	43	7	50	30
Aneka Tambang	2.25	13	24	45	5	56	29
Cerro Matoso	2.9	14	15	46	22	59	20

The Cerro Matoso plant (built 1982) treats an ore with the exceptionally high SiO_2/MgO ratio of 3 : 1. The resulting acid slags have proved to be highly corrosive to furnace refractories. The dried ore (12 % moisture) is ground, blended with coal, and pelletized prior to entering the rotary kiln (Fig. 22). The kiln is operated at 900 – 950 °C with an oxidizing atmosphere and a reducing bed to achieve prereduction of 20 % of the nickel to metal, and up to 95 % of the iron oxide to the divalent state

(FeO). Smelting is carried out in a single 51 MVA electric furnace. The ferronickel, grading over 40 % Ni, is tapped at 1420 – 1440 °C, while the slag, containing 0.2 % Ni, is tapped at 1600 – 1630 °C. The ferronickel is refined and upgraded to 43 % Ni prior to being cast into 22 kg ingots, or granulated [96].

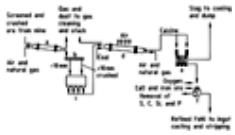

Figure 22. Flow sheet of the Cerro Matoso ferronickel smelter

a) Drying kiln; b) Screen; c) Batch bins; d) Reduction kiln; e) Electric furnace; f) ASEA/SKF Fe – Ni refining

The operation of LARCO at Larymna in Greece is an example of ferronickel production from a very low grade limonitic ore (1.2 – 1.3 % Ni, 31 – 35 % Fe). In this case a high degree of prereduction in the rotary kiln is essential. The ground ore is blended with lignite, coal, and heavy fuel oil, and heated to 920 °C in one of two countercurrently fired rotary kilns. The kiln discharge still contains about 3 % carbon, but is reported to contain no metallic nickel, while most of the iron oxides are reduced to FeO. The calcine is smelted in one of four electric furnaces (13 – 16 MVA) to produce a metal phase containing 15 % Ni, 0.02 % C, and 0.25 % S. The metal is tapped at 1520 °C and upgraded to 28 % Ni in an oxygen-blown converter. This treatment also reduces the sulfur content to 0.06 %. Energy consumption for the electric furnace is 125 MJ/kg Ni (1.6 MJ/t ore or 440 kWh/t ore) and electrode consumption is 0.4 kg/kg Ni (5 kg/t ore) [98].

8.2.2. *The Ugine Ferronickel Process*

In the rotary kiln – electric furnace process described in Section The Rotary Kiln – Electric Furnace Process excess carbon is normally added to the furnace charge to ensure quantitative reduction of nickel oxide. As a result more iron oxide is reduced to metal than is desirable, decreasing the ferronickel grade and leaving a high residual carbon level, typically 1 – 3 %, in the product. The excess iron and carbon are subsequently removed by oxidation with air or oxygen. Consequently the process is not particularly economical in terms of energy consumption.

A more selective method of nickel oxide reduction was developed by the French company Ugine in the early 1950s and applied commercially by Hanna Mining in their plant in Riddle, Oregon, in 1954. In this process the ore is dried and calcined in a rotary kiln and melted in an electric furnace without addition of reductant to produce a melt of iron and nickel oxides. No slag – metal separation occurs in the electric furnace. The molten ore is transferred to a ladle furnace where it is reduced by reaction with ferrosilicon at 1650 °C. The two phases are mixed by repeated pouring of the mixture from one ladle furnace to another. The nickel oxide and some of the iron oxide are reduced to metal. The balance of the iron is removed as slag.

$$2\,Fe_2O_3 + FeSi \rightarrow 4\,FeO + SiO_2 + Fe$$

$$2\,NiO + FeSi \rightarrow 2\,Ni + SiO_2 + Fe$$

$$2\,FeO + FeSi \rightarrow 3\,Fe + SiO_2$$

The reduced ferronickel contains 30 – 50 % Ni and very low levels of carbon and sulfur. The major impurity is phosphorus, which is removed by oxidation with iron ore and slagging with lime.

8.2.3. *The Falcondo Ferronickel Process [99]*

The Falconbridge ferronickel operation in the Dominican Republic treats a low-grade ore (1.5 % Ni, 17 % Fe, 24 % MgO, and 35 % SiO_2) by means of a specially developed process designed to provide a more selective reduction of nickel relative to iron than is possible in the rotary kiln – electric furnace process. Reduction of briquetted ore is carried out in shaft furnaces fired with a reducing gas, generated by cracking low-sulfur naphtha. No solid or liquid reductants are added to the furnace feed. A high degree of reduction of the ore is achieved in the shaft furnace, and the subsequent electric furnace step is required to do little more than melt the reduced calcine to allow separation of the metal and slag phases. The ferronickel produced is high grade (38 % Ni) and contains low impurity levels (< 0.04 % C and Si).

The Falcondo plant, which has a capacity of 35 000 t/a Ni, was commissioned in 1971. The ore is carefully blended to maintain a constant chemical composition, particle size, and moisture content, so that it can be briquetted without a binder. The briquettes are calcined and reduced in twelve open-top shaft furnaces. Each furnace is equipped with its own gasification reactor in which low-sulfur naphtha is cracked by combustion with a deficiency of air to produce reducing gas ($CO + H_2$). The hot reducing gas at 1250 °C is first cooled to 1150 °C before being supplied to the shaft furnace through the primary tuyeres. The gas flows upwards, countercurrent to the ore briquettes which are fed to the open top of the furnace. The gas reduces the nickel oxides to metal and the iron oxides to FeO. A portion of the partly reacted reducing gas is taken off from the top of the furnace, where it is mixed with a controlled amount of air and fed back into the furnace through secondary tuyeres located 1.8 m above the primary tuyeres. This stream provides the combustion fuel to dry and calcine the ore feed in the upper section of the shaft.

The hot reduced briquettes are discharged from the bottom of the shaft furnace at 800 °C and are transferred to one of three 55 MVA electric furnaces. The furnace operation consumes only 1.6 GJ/t (440 kWh/t) of charge. The crude ferronickel, which is tapped at 1475 – 1500 °C, contains 32 – 40 % Ni. The slag, containing about 0.15 % Ni, is tapped at 1500 – 1600 °C. The crude ferronickel, which typically contains only 0.15 % S, 0.03 % P, 0.04 % Si, and 0.02 % C, is refined to remove sulfur and phosphorus.

8.2.4. *Refining of Ferronickel*

The refining of ferronickel resembles that of steel in that the principal impurities are sulfur, carbon, silicon, phosphorus, and oxygen. Typical impurity levels in crude and refined ferronickel are shown in Table 14.

Table 14. Impurities in ferronickel [29]

Smelter	Crude Fe – Ni, %				Refined Fe – Ni, %			
	C	S	Si	P	C	S	Si	P
SLN	2.5	0.40	1.50	0.150	0.03	0.02	0.02	0.02
Cerro Matoso	0.17	0.45	1.45	0.024	0.03	0.03	0.70	0.03
Falcondo	0.02	0.15	0.04	0.030	0.06	0.08	0.30	0.01

Crude ferronickel produced by the conventional rotary kiln – electric furnace process usually contains high levels of carbon and sulfur. These ferronickels are first treated under reducing conditions to remove sulfur and are then refined sequentially under oxidizing conditions with suitable fluxes to remove carbon, silicon, and phosphorus. Ferronickel produced by the selective reduction processes, such as the Ugine or Falcondo processes, which typically contain very low levels (< 0.04 %) of carbon and silicon, are normally treated first for phosphorus and then for sulfur removal. The Falcondo ferronickel is not refined to remove carbon or silicon, and in fact the levels of these impurities increase during refining.

A variety of equipment is used for the refining of ferronickel, including electric arc furnaces, shaking ladles, and low-frequency induction furnaces for desulfurization, and oxygen-blown converters (both L-D and side-blown) for silicon, carbon, and phosphorus removal. The ASEA – SKF ladle furnace, which incorporates an electric arc furnace roof carrying three carbon electrodes and induction stirring of the melt, is used in a number of plants including Cerro Matoso and Falcondo.

Sulfur is generally removed under reducing conditions by adding soda ash, lime, or calcium carbide to the molten ferronickel. Good agitation of the melt is essential to ensure effective mixing of the reagent with the metal phase, and this can be achieved by adding the reagent to the ladle during furnace tapping, by using a shaking ladle, or by providing electrical induction stirring. In some plants poling with greenwood poles is used to provide additional agitation and reductant. Sulfur reacts with the fluxes to form sodium or calcium sulfides, which are slagged off with the silicates formed by reaction of flux with silicon:

$$3\,Na_2O + 2\,S + Si \rightarrow 2\,Na_2S + Na_2SiO_3$$

$$CaC_2 + S \rightarrow CaS + 2\,C$$

The melt cools during this treatment and oxygen is blown into the melt to reheat it to tapping temperature. Two stages of fluxing are often necessary to desulfurize the ferronickel to the required level (0.02 %). The Falcondo smelter uses a complex combination of ferrosilicon, aluminum, silicocalcium, lime, and fluorspar to desulfurize its low-carbon ferronickel.

Removal of silicon, carbon, and phosphorus is normally conducted in an oxygen-blown converter or an oxygen lanced ladle. Carbon is removed simply by oxidation with oxygen to carbon monoxide. Silicon is removed by oxidation with oxygen and fluxing with lime and fluorspar. Phosphorus is oxidized either by blowing the melt with oxygen or by adding iron ore and fluxing the phosphorus(V)oxide with lime:

$$6\,P + 5\,Fe_2O_3 + 12\,CaO \rightarrow 10\,Fe + 3\,Ca_4P_2O_9$$

The slags are removed by skimming after each refining step. If necessary the ferronickel can be deoxidized by addition of ferrosilicon, although this practice increases the silicon content.

Ferronickel is not usually treated to remove cobalt, copper, or arsenic, which may be present in significant amounts. Refined nickel can be produced by the electrorefining of 90 % ferronickel in a chloride – sulfate electrolyte, but this process has not found wide application [98].

8.3. Smelting to Nickel Matte

Low-iron nickel sulfide matte has long been produced from oxide nickel ores by reduction and sulfidization in a blast furnace. This process is still practiced in the former Soviet Union at three smelters in the Urals, but the blast furnace has given way to the electric furnace smelting process in New Caledonia and in the P. T. Inco operation in Indonesia.

8.3.1. *Blast Furnace Smelting*

The blast furnace production of nickel matte was practiced by SLN at its Doniambo smelter in New Caledonia from 1880 until 1972. The nickel silicate ore was blended with calcium sulfate and coke and reduced directly to matte in the blast furnace. The low-grade furnace matte was then blown to 78 % Ni and 20 % S in a converter [91].

At the Yuzhuralnikel Combine at Orsk in the former Soviet Union, nickel silicate ores, averaging 1.02 % Ni, 15.4 % Fe, 10 % MgO, and 49 % SiO_2, are blended with ground coke, limestone, pyrite (0.1 % Cu content), and recycled converter slag and sintered at 1100 – 1200 °C in one of five 75 m^2 downdraft sinter machines. The resulting sinter is crushed and smelted in one of eleven low (5 m height) shaft furnaces using 24 % oxygen enriched air. The furnace matte, containing 17 % Ni, 54 % Fe, and 21 % S is tapped at 1200 °C and blown in Peirce – Smith converters to a high-grade matte (77 % Ni, 2.5 % Cu, 0.2 % Fe, 21 % S).

The converter slag is cleaned to reduce its nickel and cobalt contents in a two-stage contact with molten furnace matte, and

the cleaned slag is recycled to the shaft furnaces. The converter matte is cast, ground, and roasted in fluidized-bed roasters to eliminate sulfur and is then treated in a chloridizing roast in a rotary kiln, followed by acid leaching to remove copper. The leached nickel oxide is reduction smelted with coke and the metal is granulated. The Orsk combine produces 16 000 t/a of nickel granules analyzing 98.5 % Ni, 0.5 % Co, 0.5 % Cu, and 0.5 % Fe from its exceptionally low-grade nickel laterite ore feed [49].

8.3.2. *Matte Production from Ferronickel*

SLN now produces nickel matte from crude ferronickel (24 % Ni + Co, 69 % Fe, 2 % C, and 0.25 % S) by means of a sulfidizing treatment with elemental sulfur in a Peirce – Smith converter [100]. After sulfidizing, the matte is blown with air to oxidize iron, which is slagged off with silica flux in two stages, to produce a matte containing 75 – 80 % Ni, 0.1 to 4 % Fe, and 20 % S. The cobalt content ranges from 0.4 % in matte containing 0.1 % Fe, to 1.2 – 1.7 % Co in matte containing 4 % Fe. This matte is shipped to the SLN nickel refinery at Le Havre-Sandouville in France, where it is refined to high-grade electrolytic nickel. Cobalt is recovered separately in the refining process.

8.3.3. *The Inco Selective Reduction Smelting Process [101-105]*

Inco's process for the production of nickel matte from laterite ores was developed for projects in Indonesia and Guatemala which started operation in 1977. The ore is dried and then calcined and partially reduced in a rotary kiln at 750 – 800 °C. Reduction is accomplished by adding coal to the kiln feed and by lancing high-sulfur oil into the ore bed. The hot calcine is smelted in an electric furnace. The bulk of the sulfur required to form a matte is added as molten elemental sulfur either to the kiln discharge, or directly to the electric furnace discharge. The resulting furnace matte is finally upgraded to the product matte in an air-blown converter.

In the P.T. Inco Indonesia operation on the island of Sulawesi, two ore types are blended to maintain a SiO_2/MgO ratio of 1.9 : 1, and the moist ore is crushed and screened (Fig. 23). The oversize ore is rejected and the fines, which typically contain 2.0 % Ni, 19.3 % Fe, 17 % MgO, and 33.3 % SiO_2 are dried to 20 % moisture. The dried ore is fed to one of three 100-m long rotary kilns, each of which is divided into three zones, for calcining, sulfidizing, and reduction. Reduction is achieved by injecting high-sulfur oil into the hot calcine at the discharge end of the kiln. The combustible cracking products of the injected oil are burned gradually as they pass along the kiln, countercurrent to the ore, by injecting controlled amounts of combustion air. The calcine is discharged at 770 °C and is transferred to one of three 45 MVA electric smelting furnaces. Power for the smelting furnaces is supplied from a captive hydroelectric facility, which has proved a major economic advantage in comparison with most other nickel laterite operations, which are heavily dependent on the use of fossil fuels. Energy consumption in smelting is 2.1 – 2.2 GJ/t ore (580 – 600 kWh/t). Tapping temperatures are 1360 °C for the 32 % Ni, 10 % S matte and 1550 °C for the slag, which typically contains 0.16 % Ni, 20 % Fe, 45 % SiO_2, and 23.5 % MgO.

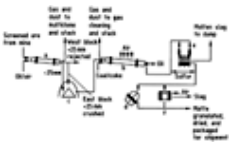

Figure 23. Flow sheet of the P.T. Inco Indonesia matte smelter

a) Drying kiln; b) Reduction kiln; c) Dried ore storage; d) Top-blown rotary converter; e) Electric furnace (melting); f) Peirce – Smith converter

The furnace matte is upgraded to converter matte (79 % Ni, 19.5 % S, 0.5 % Fe) in one of two side-blown Peirce – Smith converters. These converters were installed to replace the top-blown rotary converters originally provided for this purpose, which had proved to be expensive to operate due to high refractory wear and high maintenance costs. The matte is granulated for shipment to the Japanese market.

The P.T. Inco Indonesia operation is to be expanded from its 1988 production rate of 34 000 t/a Ni to 47 000 t/a Ni in the early 1990s. A smaller Inco operation in Guatemala (Exmibal) which utilized the same process, but was dependent on fuel oil for power generation, closed in 1980 after three years of uneconomic operation.

8.4. Energy Consumption in Laterite Smelting

The pyrometallurgical production of ferronickel and nickel matte from oxide ores is extremely energy intensive and energy costs can account for 50 % of the production cost of ferronickel. Energy is consumed in drying the moist ore, in calcining the dried ore to remove chemically bound water, in heating the charge to melting temperature, and in smelting. Typical fuel and electrical energy consumption data collected in a 1986 survey [30] are presented in Table 15.

Table 15. Smelter fuel and power consumption data [30]

Operation	Ore grade, % Ni	Fe – Ni grade, % Ni	Coal, t/t Ni	Oil, 10^3 L/t Ni	Gas, m^3(STP)/t Ni	Electricity, GJ/t Ni
Falcondo	1.75	38		15.7		
P.T. Inco	2.25	matte	1.1	6.9		140
Hyuga	2.4	20 – 25	7.5			115
Nippon	2.4	20	6.0	0.6		79
Aneka Tambang	2.25	25	2.3	3.8		108
Cerro Matoso	2.9	45	2.7		3.1	94

These figures reflect the trend to the use of coal in preference to oil in the decade following the increase in oil prices in 1973. The oil consumption reported for the Falcondo operation includes both the naphtha used in the metallurgical process and the heavy oil used for the production of electrical power in the on-site generating station.

The distribution of energy consumption in a typical ferronickel smelter is illustrated by the following data (in GJ/t ore):

Drying	1.3 – 2.0
Preheating and prereduction	3.3 – 4.2
Electric smelting	1.4 – 2.2
Other electrical energy	0.22 – 0.29
Total	6.22 – 8.69

Electrical energy for the electric furnace actually accounts for only 25 – 30 % of the total energy consumption in a smelter, while the rotary kiln operation consumes about 50 % of the total energy required.

Total energy consumption at the P. T. Inco Indonesia operation was decreased from 620 GJ/t Ni in 1980 to 460 GJ/t Ni in 1985 by continuing process optimization [103]. During this period the percentage of energy provided by fuel oil was also decreased from 73 % to 62 % of the total by a partial substitution of oil by coal, which provided 8 % of the energy used in 1985. Overall, therefore, these changes resulted in a 37 % reduction in oil consumption per tonne of nickel produced. These economies in fuel oil usage, combined with a low cost source of hydroelectric power, have made the P. T. Inco operation in Indonesia one of the world's cheaper nickel producers [104].

[Top of Page]

9. Hydrometallurgy of Oxide Ores [1], [63]

The chemical and mineralogical homogeneity of limonitic nickel ores, allied with the high value of potential byproducts such as cobalt, chromium, and iron, make them ideal feed materials for hydrometallurgical treatment. Most of the nickel in iron-rich laterite ores is contained in goethite (α-FeO · OH), while the cobalt is almost always associated with manganese oxides. The recovery of nickel and cobalt thus essentially consists of separating them from iron and manganese.

A variety of potential hydrometallurgical or combined pyrometallurgical – hydrometallurgical process routes have been proposed for laterite treatment. Approaches considered include sulfate or chloride roasting followed by water leaching to recover solubilized nickel and cobalt sulfates or chlorides, with soda roasting of the residue and water leaching to recover chromium, as well as the direct leaching of nickel and cobalt with sulfuric, nitric, or hydrochloric acid. Only two hydrometallurgical treatment processes have been commercialized, and although both recover cobalt and nickel, neither is operated for chromium or iron recovery. The two commercial processes, the Caron process, in which partially reduced ore is leached in ammoniacal solution, and the Moa Bay process, using direct sulfuric acid pressure leaching of ore, were both developed to treat the very large, high-iron laterite deposits of Cuba.

9.1. The Caron Process

The reduction roast and ammonium carbonate leach process was patented by M. H. CARON in 1924, but the first commercial application of the process was not initiated until World War II, when a plant was constructed at Nicaro, Cuba. The ore is first dried to about 5 % moisture and is then selectively reduced in a multihearth roaster. The reduced calcine is cooled under nonoxidizing conditions and leached in ammoniacal ammonium carbonate solution to selectively extract nickel and cobalt. The pregnant leach liquor is then boiled to strip ammonia and carbon dioxide, causing the nickel and cobalt to precipitate as basic carbonates. The mixed nickel – cobalt basic carbonate can be calcined to a nickel oxide product or can be refined to nickel powder or cathode. Details of a number of operations using the Caron process are given in Table 16.

Table 16. Commercial Caron process plants

Operation	Design capacity, t/a Ni	Feed ore, %		Product type
		Ni	Co	
Nicaro, Cuba	22 500	1.40	0.10	NiO sinter
Queensland Nickel, Australia	18 000	1.35	0.11	NiO rondelles, NiS – CoS
Nonoc, Philippines	30 000	1.22	0.10	Ni powder, NiS – CoS
Punta Gorda, Cuba	30 000			NiO sinter
Tocantins, Brazil	5 000	1.60	0.14	Ni cathode
Elbasan, Albania	3 000			$NiCO_3$

The original plant at Nicaro, built and operated by the Freeport Sulfur Company [1], [86], for the U.S. Government, was

commissioned in 1944. It was closed down in 1947 as uneconomic under peacetime conditions, despite having achieved a production rate of 14 500 t of nickel as nickel oxide in 1946. However nickel recovery from the ore had never exceeded 70 %. It was reopened in 1952 for the U.S. Government under the operating management of a joint venture of the National Lead Company and Fomento de Minerales Cubanos. Its production capacity was increased to 22 500 t/a Ni in 1957, and in 1960 it was nationalized by the Cuban Government. Production is believed to have ranged from 16 000 – 20 000 t/a Ni during the 1970s and 1980s [106].

Freeport was also a member of the joint venture which constructed the Queensland Nickel refinery near Townsville, Australia in 1974, which drew heavily on the design and operating experience gained at Nicaro [107]. A second Caron plant in Cuba was built by the Cuban Government with Soviet assistance in the early 1980s at Punta Gorda, near Moa Bay. The first of three processing lines was commissioned in 1986 and by the end of 1988 production was reported to have reached 10 000 t/a Ni as nickel oxide. A third Cuban Caron plant is under construction in the same area at Las Camarioca, scheduled to start-up in 1995 [108].

The other large Caron plant was built by Marinduque Mining and Industrial Corporation on Nonoc Island in the Philippines with technical assistance from Sherritt Gordon [109], [110]. The Nicaro process was adapted by Sherritt to increase nickel and cobalt extractions and to produce hydrogen-reduced powder from the basic nickel carbonate. The Nonoc nickel refinery was commissioned in 1974 and reached a production rate of around 22 000 t/a Ni in 1977, at a nickel recovery of 78 % Ni. This plant was closed in 1986 after several years of declining operation due to adverse economic conditions. There are currently (1989) plans to refurbish the process plant and resume operation in the early 1990s. Both the Queensland Nickel and the Marinduque nickel operations were designed to recover cobalt separately as a byproduct by precipitating a mixed cobalt – nickel sulfide prior to precipitation of basic nickel carbonate.

The Companhia Niquel Tocantins plant in Brazil began production of nickel cathode from basic nickel carbonate in 1981 and was scheduled to increase capacity to 10 000 t/a Ni by 1990 [111]. The small Caron plant at the Elbasan metallurgical complex in Albania ships its basic nickel carbonate product to the Severonikel Combine in the former Soviet Union for refining to metal.

Industrial practice is illustrated by the following description of the Queensland Nickel process, which closely resembles the original Nicaro process [1] except for the addition of a cobalt recovery step. The flow sheet of this plant, which was designed to treat 5700 t/d of ore from the Greenvale ore body is outlined in Figure 24 [107], [63].

Figure 24. Flow sheet of the Queensland Nickel laterite refinery

CCD = Countercurrent decantation; BNC = Basic nickel carbonate

The blended ore (1.35 % Ni, 0.11 % Co, 25 % Fe, 10 % MgO) is first dried in one of three rotary dryers to about 5 % moisture content. The dried ore is then ground in a ball mill and blended with fuel oil as a reductant, before being fed to one of ten seventeen-hearth Herreschoff roasters, where it is roasted under reducing conditions at 700 – 760 °C with a retention time of ca. 90 min. The ore is heated by hot reducing gas ($CO + H_2$) formed by the combustion of fuel oil in a deficiency of air in the ten combustion chambers on each roaster. The reduction potential of the gas phase is regulated so that most of the nickel and cobalt in the ore are reduced to the metallic state (as a ferronickel alloy) and the iron(III) oxides are reduced to magnetite without significant reduction of the iron oxides to iron(II) oxide or to the metal.

$$NiO + H_2 \rightarrow Ni + H_2O$$

$$3\,Fe_2O_3 + H_2 \rightarrow 2\,Fe_3O_4 + H_2O$$

As a result the nickel and cobalt are rendered amenable to ammoniacal leaching while the iron remains as insoluble oxides or silicates. Typically 80 – 85 % of the nickel content of the ore is converted to the metallic form. Any nickel oxide which is not reduced to metal is rendered insoluble under the roasting conditions by incorporation into the silicate minerals. Good control of the roasting conditions is therefore essential for optimal nickel recovery in the process.

The reduced ore is cooled in a nonoxidizing atmosphere, to avoid reoxidation of the nickel, and is quenched and leached in ammoniacal ammonium carbonate solution containing 7 % NH_3 and 4 % CO_2. Leaching is carried out in three trains of four mechanically agitated, air-sparged leaching tanks at a pulp density of ca. 20 % solids. The leaching reactions can be represented by the following equation:

$$FeNi + O_2 + 8\,NH_3 + 3\,CO_2 + H_2O \rightarrow Ni(NH_3)_6^{2+} + Fe^{2+} + 2\,NH_4^+ + 3\,CO_3^{2-}$$

The iron(II) is further oxidized and precipitates as iron(III) hydroxide:

$$4\,Fe^{2+} + O_2 + 2\,H_2O + 8\,OH^- \rightarrow 4\,Fe(OH)_3$$

However, significant losses of cobalt occur by coprecipitation with the iron hydroxide.

The pregnant liquor is separated from the leached ore in a seven-stage countercurrent decantation thickener circuit using ammoniacal solution as the washing medium. The washing efficiency is typically 99 %. Ammonia is recovered from the washed leach residues by steam stripping, and the barren solids are discarded. The pregnant liquor recovered from the wash

circuit, which contains 10 g/L Ni and 0.5 g/L Co, was formerly treated with ammonium hydrogen sulfide in a pipeline reactor to precipitate virtually all of the cobalt and about 10 % of the nickel. The mixed sufide product (15 % Co, 40 % Ni) was thickened, washed, and spray dried, and sold for cobalt and nickel recovery elsewhere. Ammonia and carbon dioxide were then stripped from the cobalt-free solution by direct stream injection to precipitate basic nickel carbonate.

Since 1988, the pregnant liquor has been treated by solvent extraction to separate nickel and cobalt [112]. An organic reagent (LIX 87 QN) is used to extract nickel selectively from the pregnant liquor. The nickel is recovered from the solvent by stripping with a more concentrated ammoniacal ammonium carbonate solution. Basic nickel carbonate can then be precipitated from the strip liquor by steam injection to remove ammonia and carbon dioxide. Copper sulfide is precipitated from the raffinate and refined elsewhere.

The basic nickel carbonate is recovered by thickening and filtration, and the wet filter cake, containing about 50 wt % moisture, is dried and calcined at 1300 °C in an oil-fired rotary kiln to yield nickel oxide. The nickel oxide calcine (77 % Ni) is upgraded by partial hydrogen reduction to yield nickel oxides containing 85, 88, or 90 % Ni, which are compacted to rondelles and sintered for market.

By the late 1980s the Greenvale nickel ore body, which had provided the sole feed to the Queensland Nickel refinery since 1974 was rapidly becoming depleted, and the operation started to treat laterite ores imported from New Caledonia and Indonesia.

The major difference between the process flow sheets used by Queensland Nickel and Marinduque / Nonoc was the production of refined nickel powder in the latter [110]. Basic nickel carbonate was redissolved in ammoniacal ammonium sulfate solution and metallic nickel powder was precipitated by reduction with hydrogen under pressure at 200 °C (Section Hydrogen Reduction to Nickel Powder). The nickel powder typically analyzed 99.8 % Ni, 0.08 % Co, 0.01 % Fe, 0.007 % Cu, 0.006 % C, and 0.004 % S. By comparison the nickel oxide sinter produced prior to 1960 in the Nicaro plant, which had no cobalt separation step, analyzed 88 % Ni, 0.7 % Co, 0.04 % Cu, 0.3 % Fe, 0.05 % C, 0.05 % S, and 7.5 % O_2. In the Tocantins refinery a crude basic nickel – cobalt carbonate is redissolved in spent sulfuric acid electrolyte, and after purification by sulfide precipitation to remove cobalt and copper, the solution is treated by electrowinning to yield nickel cathode [63].

The Caron process is very energy intensive, and this factor proved a major economic handicap for the Queensland Nickel and Marinduque/Nonoc plants, which were both designed prior to 1973, with their heavy reliance on fuel oil for ore drying, reduction, and power generation. The fossil fuel energy requirements for the process for ore drying and reduction, and power and steam generation amount to ca. 6.7 GJ/t of dry ore, with an additional 200 – 250 MJ/t of electrical power.

The Queensland Nickel plant originally consumed 900 t/d of fuel oil to treat 5700 t of ore. The distribution of oil consumption is as follows [113]:

Ore drying	14 %
Steam and power generation	35 %
Ore reduction roasting	47 %
Nickel carbonate calcining	4 %

Both Queensland Nickel and Marinduque/Nonoc converted their ore drying and steam and power generation facilities to use coal in place of oil in the early 1980s, with substantial economic benefit. However, conversion of the roasters to coal firing was not considered to be technically feasible. The Tocantins plant in Brazil makes extensive use of local eucalyptus wood as fuel in its ore treatment plant.

9.2. Pressure Leaching with Sulfuric Acid [1]

Because of the naturally high moisture content of limonite minerals (20 – 50 % H_2O), it would be advantageous if the ore drying and reduction roasting steps, which together account for over 50 % of the energy consumed in the Caron process, could be avoided, and the ore leached directly. However direct attack by the common mineral acids under ambient conditions is unselective and total dissolution of the ore results, with uneconomically high acid consumption. Fortunately, leaching of the ores in sulfuric acid at high temperature (> 250 °C) and pressure is highly selective since nickel and cobalt are stable as soluble sulfates, while iron(III) sulfate hydrolyzes and precipitates as a hydrated iron(III) oxide. Offsetting the energy savings in this type of process are the increased costs of construction materials capable of withstanding the highly corrosive environment of the high-temperature acid leach.

The applicability of the acid leach process is limited to those ores which do not contain substantial levels of other acid consuming minerals such as magnesia. It was for this reason that Freeport did not select an acid leaching process when seeking a process to treat the Nicaro ores (8 % MgO). However when the same company subsequently developed the low-magnesia (1.7 % MgO) limonitic ore body at Moa Bay, Cuba, the sulfuric acid leach process was selected since it offered much higher (90 % or more) recoveries for both nickel and cobalt, with relatively low acid consumption. Although this process has been applied on a commercial scale only in the one plant at Moa Bay, considerable process development effort was devoted during the 1970s to improving the economics and extending the applicability of the process to higher magnesia content ores, principally by Amax.

9.2.1. *The Moa Bay Process*

The plant at Moa Bay was designed and built by Freeport in 1959 to produce 22 700 t/a Ni and 2000 t/a Co in the form of a mixed sulfide concentrate. The concentrate was to be refined to the pure metals in a separate plant which Freeport built at Port Nickel, Louisiana, where a cheap supply of natural gas was available. The Moa Bay operation was nationalized in 1960

by the Cuban Government before construction was complete. It was recommissioned in mid-1961 with Soviet assistance, and since then the nickel – cobalt sulfide concentrate has been shipped to the former Soviet Union for further processing. The current production rate (1988) is believed to be 18 000 t/a Ni and 1600 t/a Co.

Although the original plant design is well documented, information on subsequent modifications is fragmentary. The following description is therefore based largely on accounts dating from 1960 [1], [106], [114-116]. The process flow sheet is shown in Figure 25. The ore, which consists mainly of goethite (α-FeO · OH), is slurried in water and screened to remove low-grade, oversize material, which is discarded. The fines fraction (< 1 mm) that forms the feed to the acid pressure leach averages 1.35 % Ni, 0.15 % Co, 48 % Fe, 1.7 % MgO, 9 % Al_2O_3, 3 % Cr_2O_3, and 4 % SiO_2 in composition. It is leached at 45 % solids pulp density with 98 % sulfuric acid (240 kg H_2SO_4/t dry ore) at about 245 °C and 4.3 MPa, with a retention time of 1 – 2 h. The acid is made on-site from elemental sulfur, and the waste heat from the acid plant is used in the leaching circuit. Leaching is carried out in four leaching trains each containing four large vertical autoclaves, which are lead- and brick-lined with titanium internal fittings. Agitation is achieved by injecting high-pressure steam into a draft tube at the bottom of the vessel. Heat is recovered from the autoclave discharge slurry for use in preheating the cold feed slurry. Under these conditions, about 94 % of the nickel and cobalt dissolve, while iron solubility is limited to about 1 g/L.

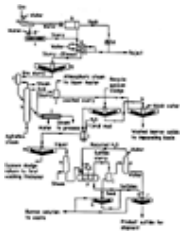

Figure 25. Flow sheet of the Moa Bay nickel laterite plant

a) Shaking grizzly; b) Cylindrical scrubber; c) Log washer; d) Jaw crusher; e) Screens; f) Slurry heaters; g) Storage thickener; h) Leaching autoclaves (4 in series); i) Slurry cooler; j) Flash tank; k) Washing thickeners (6-stage CCD); l) Neutralizing tanks (4-stage cascade); m) Gypsum thickener; n) Heaters; o) Gas cooler; p) Precipitation autoclave; q) Sulfide thickener; r) Washing thickeners (2-stage CCD)

$$NiO + H_2SO_4 \rightarrow NiSO_4 + H_2O$$

$$CoO + H_2SO_4 \rightarrow CoSO_4 + H_2O$$

The leach liquor typically contains 6 g/L Ni, 0.5 g/L Co, 2 g/L Mn, 0.8 g/L Fe, 0.4 g/L Cr, 2.5 g/L Mg, 2 g/L SiO_2, 3 g/L Al, and 30 g/L H_2SO_4. The liquor is separated from the leach residue (0.08 % Ni, 0.01 % Co, and 51 % Fe) by countercurrent decantation (CCD) washing with water in a six-stage thickener circuit. The washed residue is stockpiled for possible future use as iron ore.

The pregnant solution is treated with coral mud (calcium carbonate) to neutralize the acid and precipitate most of the dissolved iron, aluminum, and manganese by hydrolysis. The precipitated sludge is recycled to the wash circuit and joins the leach residue for disposal. The neutralized solution typically contains (g/L): 3.5 – 4.0 Ni, 0.3 – 0.4 Co, 1.0 Mn, 0.3 Fe, 0.4 Cr, 2.5 Mg, 1.0 SiO_2, and 3.0 Al at pH 2.5.

Nickel and cobalt are recovered from solution in 98 – 99 % yield by precipitation at 120 °C with hydrogen sulfide at a pressure of 1.1. MPa in a three-compartment, brick-lined autoclave.

$$NiSO_4 + H_2S \rightarrow NiS + H_2SO_4$$

The barren solution typically contains (g/L): 0.05 Ni, 0.01 Co, 1.0 Mn, and 7 H_2SO_4. The nickel – cobalt concentrate contains 55 % Ni, 5 % Co, 0.7 % Fe, 0.2 % Cu, 1.0 % Zn, and 35 % S.

The refinery built by Freeport in Louisiana to treat the concentrate employed pressure hydrometallurgy and hydrogen reduction to separate and refine nickel and cobalt. This plant was being commissioned when the Cuban nickel industry was nationalized in 1960, and was never used for the purpose for which it was designed. Currently (1989) the total nickel – cobalt concentrate production from Moa Bay is shipped to the former Soviet Union, where it is treated at a smelter in the Urals to produce nickel – cobalt anodes analyzing 87 % Ni, 8 % Co, 2 % Fe, and 0.7 % Cu. The anodes are electrorefined at the Yuzhuralnikel Combine at Orsk in a conventional sulfate – chloride electrolyte system. The nickel is recovered as high-grade electrolytic nickel, and cobalt is separated from anolyte by chlorine precipitation (Section Electrorefining) for subsequent processing to metal [49].

Energy consumption in the Moa Bay acid leach process is relatively low because over half the steam requirement for the leach is provided by waste heat recovered from the sulfuric acid plant. The fuel oil requirement for the generation of power and steam has been estimated at about 40 kg/t of dry ore, which is only 20 % of the fuel requirement for the Caron process. However, this energy benefit is frequently outweighed by the high consumption of sulfuric acid which depends directly on the magnesia content of the ore [117].

9.2.2. *The Amax Acid Leach Process [118-120]*

The Amax acid leach process for the treatment of laterite ores was developed, on the basis of the Moa Bay process, with the object of increasing metal recoveries while decreasing acid and energy consumption. The process was tested extensively in pilot plant and prototype plant campaigns between 1975 and 1981, for a potential new project in New Caledonia which was subsequently abandoned.

In the Amax process (Fig. 26) the ore is divided by selective mining or screening into a low-magnesia limonitic fraction and a smaller fraction of high-magnesia serpentinic ore. The finer limonitic fraction is leached directly in the pressure leach at 270 °

C, while the high-magnesia fraction is first calcined and then preleached at atmospheric pressure with the leach liquor from the countercurrent decantation wash circuit, before it goes to the pressure leach. Essentially stoichiometric utilization of the acid is achieved by this countercurrent leaching of the high-magnesia ore. The overall acid requirement is thus reduced substantially, and the need for a separate alkaline neutralizing agent (e.g., coral mud at Moa Bay) is eliminated. Metal extractions of 94 – 97 % are normally achieved with a 20-min retention time in the pressure leach at 270 °C.

Figure 26. Flow sheet of the Amax nickel laterite process

A major saving (up to 70 %) in energy consumption in the leaching step is achieved by using indirect heating of the feed slurry with flash steam recovered from the autoclave discharge slurry. By using this technique in place of direct steam sparging, the concentration of metals in solution is substantially increased, and the size and cost of subsequent processing steps is reduced. Overall the requirement for fossil fuel is estimated to be only 50 % of that for the Moa Bay operation. A further simplication of the process was achieved by improving mass transfer in the sulfide precipitation step so that it could be operated at a lower pressure of 0.2 MPa compared with 1.1 MPa at Moa Bay. The Amax acid leach process is claimed to be economically viable over a wide range of laterite ore compositions, with magnesia contents as high as 15 % MgO.

[Top of Page]

10. The Recovery of Byproduct Cobalt

All nickel ores, whether sulfide or oxide, contain significant amounts of cobalt that can potentially be recovered as a valuable byproduct. Overall, however, the recovery of cobalt from nickel ores is very low, and the nickel industry provides less than 30 % of world cobalt production. The greater part of world cobalt production arises from the copper operations in Zaire and Zambia.

No cobalt recovery at all is possible in ferronickel smelting, where the cobalt is distributed between the discard slags and the ferronickel product, in which it merely represents an impurity. The same comments apply to class II nickel products such as nickel oxide sinter, made by the smelting of sulfide concentrates or by the Caron process without a cobalt separation circuit.

Cobalt is normally recovered in high yield as a byproduct of carbonyl, electrolytic, or hydrometallurgical refining of nickel metal or matte. However, even in these cases the overall recovery of cobalt from the ore is low, since most of the cobalt is lost to the slag phase in the prior smelting processes. Cobalt recovery in the converting operation depends critically on the final iron content of the matte. Typically if the matte is blown to less than 1 % Fe, over 75 % of the cobalt reports to the slag phase and is effectively lost. If the conversion process is stopped at 3 – 5 % Fe, however, over 75 % of the cobalt remains in the matte, and can be recovered during subsequent refining.

In practice few nickel smelters are operated to maximize the recovery of cobalt. The SLN in New Caledonia and the Soviet smelters producing matte from laterites are notable exceptions. Slags are cleaned by contact with a high-iron matte, which becomes enriched in cobalt. The cobalt is then recovered by hydrometallurgical treatment of the enriched material.

Even in the ammonia pressure leaching of nickel sulfide concentrates, where the smelting step is bypassed, cobalt recovery is much lower than that of nickel, due to a strong affinity of the iron oxide residue for soluble cobalt. Probably the highest overall recovery of cobalt from nickel ores is achieved in the sulfuric acid pressure leaching of laterite ores, in which a cobalt recovery of 90 % is possible.

[Top of Page]

11. Nickel Recovery from Secondary Sources

Nickel-containing scrap from forming processes in fabricating plants and machine shops is an important source of feed material for the production of nickel-containing stainless steel. Typically, in the United States about 40 % of the nickel used for stainless steel production is provided in the form of stainless steel scrap from outside sources. Recycled inhouse scrap is not included in this estimate [121], [122]. Nickel-containing scrap falls mainly into the categories of stainless steels, superalloys, copper – nickel alloys, and nickel itself. Usually scrap is recycled to a plant which produces the same type of alloy.

Inco established The International Metals Reclamation Company (INMETCO) at Ellwood City, Pennsylvania in 1978, to collect and convert nickel-containing waste materials from specialty steel companies into a stainless steel remelt material. This 40 000 t/a capacity plant treats various grades of scrap materials including flue dusts, scale and turnings, as well as stainless steel scrap, to produce 23 kg remelt pigs containing ca. 18 % Ni and 8 % Cr [123]. Falconbridge also started to recycle nickel- and cobalt- containing alloys to its primary nickel smelter in 1985 [124].

[Top of Page]

12. Market Products

Unlike copper, nickel is rarely used without further processing and so the mechanical or physical properties of nickel products are of no significance, and nickel is marketed solely on the basis of its chemical composition. Nickel is marketed according to its chemical purity as either a Class I or a Class II product. Class I includes the high-purity products, electrolytic cathode, carbonyl refined granules and powder, and hydrogen-reduced nickel in powder or briquette form. Commercial nickel of this

class is normally 99.7 % pure or better [125].

Class II includes the various grades of ferronickel, nickel oxide, and metallized nickel oxide. Typical compositions for Class I and II products are given in Table 17. Class II nickel products, which normally sell at a discount to Class I nickel, became more popular during the 1970s, since the new stainless steel making techniques such as the AOD process can tolerate higher impurity levels in the feed materials. However this trend was reversed during the 1980s, during the period of oversupply and low metal prices, particularly for alloy steel production where Class I nickel was preferred [122].

Table 17. Commercial forms of nickel [125]

Type	Composition, %						
	Ni	Co	Cu	Fe	C	S	O
Class I:							
Cathodes	> 99.90		0.005	0.002	0.01	0.001	
Pellets	> 99.97	5×10^{-5}	0.001	0.0015	< 0.10	0.0003	
Powder	99.74			< 0.10	< 0.10	< 0.010	< 0.15
Briquettes	99.90	0.03	0.001	0.01	0.01	0.0035	
Rondelles	99.25	0.37	0.046	0.022	0.022	0.004	0.042
Class II:							
Ferronickel	20 – 50	1 – 2		rest	1.5 – 1.8	< 0.3	
Nickel oxide	76.0	1.0	0.75	0.30		0.006	rest

Some nickel products are designed for specific applications, such as the 0.02 % S content “S Nickel Rounds” produced by Inco as “active” anode material for the plating industry, and the special powders used in Ni – Cd batteries. Traditionally, electrolytic nickel cathodes are cut into 2.5 cm squares for market. Both the major producers of electrolytic nickel, Inco and Falconbridge, now produce electrolytic nickel in the form of “rounds” (Inco) or “crowns” (Falconbridge), by plating nickel onto specially masked cathode blanks, to produce discrete round deposits. These deposits can be stripped from the blanks and drummed directly without shearing, and are easier for the consumer to handle than sheared cathode.

[Top of Page]

13. Uses [122], [126-128]

The major first uses of nickel are summarized in Table 18. Stainless steel production accounts for over 50 % of nickel consumption, followed by ferrous alloys, and nickel-base alloys. Electroplating takes only 11 % of nickel consumption [129].

Table 18. Western-world nickel consumption: distribution by first use in 1987 [129]

First use	Distribution, %
Stainless steel	51
Alloy steels	12
Nickel-base alloys	15
Copper-base alloys	3
Plating	11
Foundry	5
Other	3

A more detailed breakdown of nickel consumption by end use is given in Table 19. The consumer product category, which includes the stainless steel used in kitchen sinks, laundry equipment, furniture and tableware, as well as nickel-plated parts for bicycles, motorcycles, jewelry, eyeglass frames, and musical instruments, accounts for one quarter of nickel consumption. The high strength of nickel stainless steels combined with light weight and low maintenance have contributed to their growing use in railway cars and road tank trailers in the construction and machinery category [128].

Table 19. Western-world nickel consumption: distribution by end use in 1987 [129]

End use	Distribution, %

Consumer products	25
Construction and machinery	13
Automotive industry	11
Chemical industry	9
Building and construction	6
Process industry	6
Electronics	6
Power industry	5
Petroleum industry	5
Aerospace	3
Nickel chemicals	2
Marine applications	2
Coinage	2
Other	5

Nickel consumption dropped significantly during the 1970s in the automotive industry with the amount of nickel used in the average U.S. car falling from about 2 kg in 1970 to only 0.5 kg by 1985 [122]. The chemical industry is the fourth largest market for nickel, mainly in stainless steels for a multitude of applications.

Building and construction, the process industries, and the electronics industries each accounted for about 6 % of nickel consumption in 1987. Stainless steel usage in architectural applications is increasing, as it is in environmental control equipment and food processing equipment. The growth rate of nickel usage in the electronics industry is greater than 10 % per annum. Alloy 42, a nickel – iron alloy, is widely used in lead frames, while a copper – nickel – tin alloy C72500, is used in springs, clips, and terminals.

In the power industry, nickel-containing stainless steels are widely used in nuclear power stations and are increasingly being applied in scrubber systems for the removal of sulfur dioxide from the flue gases of oil- and coal-fired power stations. Nickel-containing materials are used in the petroleum industry in applications ranging from drilling bits to piping and process vessels in petrochemical plants, and to the construction of off-shore production platforms. In the marine category, nickel-containing materials are used in ships and in desalination plants. In the aerospace industry nickel is a key element in superalloys that resist stress and corrosion at temperatures of 1000 °C and above. These materials are used in gas turbine engines.

Nickel chemicals are used in the production of nickel catalysts for use in the hydrogenation of vegetable oils, in heavy oil refining, as nickel salts in electroplating, and for color enhancement in ceramics. Coinage alloys containing nickel have been used since ancient times [3]. Today the 75 Cu – 25 Ni cupronickel alloy is the most widely used, although Canada, the Netherlands, and the Republic of South Africa use pure nickel coinage, which is particularly durable.

In 1984, after several years of decreasing world demand for nickel, sixteen of the major nickel producers cooperated to set up the Nickel Development Institute (NiDI) with headquarters in Toronto, Canada. The NiDI has a mandate to increase the use of nickel. In this function NiDI has taken over the market development role which was performed very effectively by Inco from the 1920s until the mid 1970s, when a much reduced share of world nickel production and the depressed nickel market forced Inco to cut its activity in this area. The NiDI, which is international in scope, has set up technical information offices in North America, Europe, Japan, India, and Central and South America, backed by a worldwide network of consultants. The organization publishes a periodical applications oriented magazine, “Nickel”, which has wide circulation, as well as sponsoring technical papers and a series of reference books on selected topics. The NiDI has initiated a wide range of market development and research programs to promote the use of nickel, and provides expertise and advice to users of nickel and nickel-containing products [126], [128].

[Top of Page]

14. Economic Aspects [126], [129]

Western-world nickel consumption increased by an average of 6.5 % per annum from 1945 to 1974, when it reached a level of 576 000 t (Fig. 27). Consumption dropped sharply in 1975, and the demand for nickel remained flat for the next ten years, with annual rates fluctuating in the range 450 000 – 600 000 t/a. Although demand approached the 600 000 t/a consumption level in 1979 it was not exceeded until 1987 [130]. Western-world nickel consumption in 1987, 1988 and 1992 was 625 000 t, 640 000 t and 813900 t, respectively.

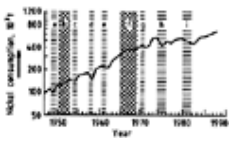

Figure 27. Western-world nickel consumption

a) 1949 recession; b) Korean war; c) 1954 recession; d) 1958 recession; e) 1961 recession; f) Vietnam war; g) 1971 recession; h) 1974 – 1976 recession; i) 1980 – 1982 recession

The slowdown in the growth of nickel consumption in the mid 1970s coincided with a major increase in production capacity, when several laterite projects, which had been under development for five to ten years, came on stream. Many of these projects were heavily dependent on fuel oil and in consequence suffered a severe penalty when oil prices increased dramatically in 1973. During this period, nickel production by the Eastern-bloc countries also regularly exceeded consumption

by 15 000 to 40 000 t/a, and substantial exports were made to Western markets.

By 1980 there were over forty nickel producers in 26 different countries, and the nickel price collapsed as the demand for nickel dropped in 1981 and 1982, and the producers fought for the shrinking market. Even the most cost efficient suffered heavy loses, and several higher cost producers were eventually forced to close permanently. The surviving operations sought to cut operating costs in every facet of nickel production, and as a result the nickel industry was much leaner and more cost efficient by the end of the 1980s.

When Western-world stainless steel production increased by 15 % in 1987, the demand for nickel increased by over 10 % from the relatively constant levels of 1984 to 1986. After more than ten years of depressed metal prices and production rates, the industry was unable to respond by increasing production levels since producer inventories of metal had been run down to a minimum and most marginal capacity had been eliminated in the drive to reduce operating costs. The result was a two year period of inflated nickel prices, with the London Metal Exchange (LME) refined nickel cash price, which had averaged US $ 1.60/lb in January 1987, peaking at $ 10.84 in March 1988. The average price for the whole of 1988 was $ 6.25/lb, and high prices continued through most of 1989 with an average price for the year of $ 6.05. This development restored the financial health of the industry after many years of marginal or uneconomic operation.

The high nickel prices of 1988/89 raised concerns among both stainless steel and nickel producers that substitution of nickel-bearing materials by cheaper materials would be encouraged. Average industry operating costs in 1989 were estimated to be around US $ 2/lb, while estimates of the capital expenditure required for a new integrated nickel project range upwards from US $ 10/lb of annual capacity. A long-term nickel price of US $ 4 – 5/lb, which would be needed to justify investment in new production capacity can be expected to depress demand and render new capacity unnecessary. Future increases in production thus appear more likely to come from the expansion of existing operations than from totally new projects.

Details of world nickel mine and smelter production by country for 1985 are given in Table 20 and the corresponding consumption breakdown is shown in Table 21. Table 22 provides a listing of the world's major nickel-producing operations.

Table 20. Production of nickel in 1989 [131]

Country	Production, 10^3 t	
	Mine	Smelter/refinery *
Albania	8.2	4.5
Australia	64.0	42.1
Botswana	18.7	
Brazil	13.6	13.6
Canada	196.0	136.5
China	28.6	30.0
Colombia	16.0	16.0
Cuba	45.4	25.0
Czechoslavakia		3.8
Former Democratic Republic of Germany	2.0	3.5
Dominican Republic	27.2	26.1
Finland	10.4	13.0
France		8.1
Greece	14.5	14.7
Indonesia	63.5	5.0
Japan		106.4
New Caledonia	98.6	36.3
Norway	0.4	54.9
Philippines	15.2	
Rep. of South Africa	34.8	27.2
Former Soviet Union	205.0	215.0
Taiwan		10.0
United Kingdom		26.1
United States		0.7
Former Yugoslavia	5.6	5.6
Zimbabwe	11.6	16.2
Total	879.3	840.3

* Including ferronickel and nickel oxides as well as refined nickel.

Table 21. Consumption of nickel in 1989 [131]

Country	Consumption, 10^3 t
Africa	10.0
Australia	3.5
Brazil	13.6
Canada	13.6
United States	107.6
Other America	6.6
China	27.5
India	15.0
Japan	163.0
Taiwan	17.6
Other Asia	14.9
Belgium	17.0
Czechoslavakia	11.5
Democratic Republic of Germany	13.0
Federal Republic of Germany	86.2
Finland	16.3
France	37.5
Italy	27.6
Poland	8.0
Soviet Union	130.0
Spain	14.2
Sweden	16.3
United Kingdom	29.0
Other Europe	25.5
Total	825.0

Table 22. Major nickel producers (1990)

Country	Company or location	Operations *	Estimated capacity, t/a
Australia	Western Mining	M, S, R	60 000
	Queensland Nickel	M, R	20 000
Botswana	BCL Limited	M, S	20 000
Brazil	Morro do Niquel	M, S	2 500
	Cia. Niquel Tocantins	M, R	5 000
Canada	Inco Limited	M, S, R	160 000
	Falconbridge	M, S	35 000
	Sherritt Gordon	R	25 000
	Hudson Bay Mining	M	9 000
China	Jianchuan	M, S, R	30 000
Colombia	Cerro Matoso	M, S	20 000
Cuba	Cubaniquel	M, R	45 000
Dominican Rep.	Falcondo	M, S	35 000
Finland	Outokumpu Oy	M, S, R	18 000
France	Soc. Le Nickel	R	10 000
Greece	LARCO	M, S	20 000
Indonesia	P.T. Inco Indonesia	M, S	40 000
	Aneka Tambang	M, S	5 500
Japan	Pacific Metals	S	44 000
	Hyuga Smelter	S	15 000

	Nippon Yakin Kogyo	S	12 000
	Niihama Nickel Refinery	R	29 000
	Nippon Mining	S	10 000
Korea	Korea Nickel	R	12 000
New Caledonia	Soc. Le Nickel	M, S	45 000
Norway	Falconbridge	R	55 000
Rep. of South Africa	Rustenburg Platinum	M, S, R	20 000
	Impala Platinum	M, S, R	15 000
Soviet Union	Norilsk	M, S, R	120 000
	Pechenga	M, S	20 000
	Monchegorsk	S, R	140 000
	Orsk	M, S, R	40 000
	Pobuzhye	M, S	10 000
United Kingdom	Inco Europe, Clydach	R	54 000
Yugoslavia	Glogovac	M, S	6 000
Zimbabwe	Bindura	M, S, R	14 000
	Empress	R	5 000

* M = mine; S = smelter; R = refinery.

For many years nickel was sold on the basis of posted producer prices, set to a large extent by Inco as the major producer. In 1979 the LME started trading in Class I nickel, and even though the LME handles only a relatively small percentage of total nickel production, it has since become the major factor in determining nickel price levels. Prices are quoted in US $ per metric tonne of Class I metal.

[Top of Page]

15. Toxicology

The toxicology of nickel and its compounds and legislation pertaining to nickel and its compounds in the environment and workplace are discussed in detail in ⟶ Nickel Compounds – Toxicology.

[Top of Page]

References

General References

1. J. R. Boldt, P. E. Queneau: *The Winning of Nickel*, Longmans Canada, Toronto 1967.
2. *Gmelin*, **system no. 57, part A11**, p. 1.
3. F. B. Howard-White: *Nickel, An Historical Review*, Methuen, London 1963.
4. Gesellschaft Deutscher Metallhütten- und Bergleute e.V.: *Symposium über Nickel*, Wiesbaden 1970.
5. G. P. Tyroler, C. A. Landolt (eds.): *Extractive Metallurgy of Nickel and Cobalt*, The Metallurgical Society of AIME, Warrendale, PA, 1988.
6. E. Ozberk, S. W. Marcuson (eds.): *Nickel Metallurgy*, vol. **I:** ""Extraction and Refining of Nickel,"" The Metallurgical Society of CIM, Montreal, Quebec, 1986.
7. D. J. I. Evans, R. S. Shoemaker, H. Veltman (eds.): *International Laterite Symposium*, SME-AIME, New York 1979.
8. Y. Ogura, I. Doi (eds.): "Proceedings of International Symposium on Laterite," *Int. J. Miner. Process.* **19** (1987) 1 – 4.

Specific References

9. K. E. Volk: *Nickel und Nickellegierungen*, Springer Verlag, Berlin 1970.
10. W. Betteridge: *Nickel and its Alloys*, MacDonald and Evans Ltd., Plymouth, United Kingdom 1977.
11. C. J. Smithells: *Metals Reference Book*, 5th ed., Butterworths, London 1976.
12. D. Nicholls: *The Chemistry of Iron, Cobalt and Nickel, Pergamon Texts in Inorganic Chemistry,* vol. 24, Pergamon Press, Oxford 1975.
13. W. S. Kirk: "Mineral Commodity Summaries 1989: Nickel," *U.S. Bur. Mines, Miner. Inf. Office*, Washington, D.C., 1989.
14. H. J. Roorda, P. E. Queneau, *Trans. Inst. Min. Metall. Sect. C* **82** (1973) 79 – 87.
15. R. A. Alcock in [5] pp. 67 – 89.
16. Inco Ltd., *First Quarter Report*, Toronto 1978.
17. T. D. Ellison: *Mining Annual Review – 1983*, London 1984, p. 68.
18. H. J. Roorda in [4] pp. 41 – 48.
19. H. S. Mashanyare, M. J. Storey in [6] pp. 13 – 36.
20. J. M. Toguri, *Can. Metall. Q.* **14** (1975) no. 4, 323 –338.
21. A. N. Kerr, W. B. Kipkie: "Recent Developments at Inco's Copper Cliff Milling Complex," *TMS-AIME Extractive and Process Metallurgy Meeting,* San Diego, CA, 1985.

22. R. Fish, *Can. Min. J.* **99** (1978) no. 5, 25 – 69.
23. Falconbridge Ltd., *Annual Report*, Toronto 1983.
24. R. R. Hoffman, G. H. Kaiura: “Smelting Process Update at Falconbridge Limited-Sudbury Operations,” *114th TMS-AIME Annual Meeting*, New York 1985.
25. T. J. Bruce, R. G. Orr in [6] pp. 57 – 94.
26. B J. Elliot, K. Robinson, B. V. Stewart in H. Y. Sohn, D. E. George, A. D. Zunkel (eds.): “Advances in Sulfide Smelting,” vol. 2, *TMS-AIME,* Warrendale, PA 1983, p. 875.
27. S. Heimala, M. Himmi, S. Jounela, M. Saari in [6] pp. 37 – 56.
28. C. S. Simons, *J. Met.* **23** (1971) no. 11, 48 – 58.
29. C. M. Diaz *et al.* in [5] pp. 211 – 239. C. M. Diaz *et al.*, *J. Met.* **40** (1988) no. 9, 28 – 33.
30. E. Ozberk, S. A. Gendron, G. H. Kaiura in [6] pp. 304 – 319.
31. C. A. Landolt, J. C. Taylor in J. C. Taylor, H. R. Traulsen (eds.): “World Survey of Non-Ferrous Smelters,” *TMS-AIME*, Warrendale, PA, 1987.
32. Canada's Non Ferrous Metals Industry: *Nickel and Copper*, Energy Mines and Resources Canada, Ottawa 1984.
33. T. R. Ingraham, R. Kerby, *Can. Metall. Q.* **6** (1967) no. 2, 89.
34. W. Schabas, *Can. Min. J.* **98** (1977) no. 5, 10 – 81.
35. R. G. Orr, A. E. M. Warner: “Fluid Bed Roasting in the Thompson Smelter,” *13th CIM Annual Conference of Metallurgists*, Toronto, Ontario, 1974.
36. B. R. Conard et al. in [6] pp. 222 – 246.
37. Y. Austin Chang, Ker-Chang Hsieh in [6] pp. 248 –276.
38. S. L. Lee, J. M. Larrain, H. H. Kellogg, *Metall Trans. B* **11B** (1980) June, 251 – 255.
39. J. A. Blanco, T. N. Antonioni, C. A. Landolt, G. J. Danyliw: “ Oxyfuel Smelting in Reverberatory Furnaces at Inco's Copper Cliff Smelter,” *50th Congress of the Chilean Inst. Min. Metall. Eng.,* Santiago, Chile 1980.
40. T. N. Antonioni, J. A. Blanco, C. A. Landolt, W. J. Middleton: “Energy Conservation at Inco's Copper Cliff Smelter,” *114th TMS-AIME Annual Meeting,* New York 1985.
41. J. A. Blanco, T. N. Antonioni, C. A. Landolt, C. M. Mitchell: “Productivity Improvements at Inco's Copper Cliff Smelter,” *115th TMS-AIME Annual Meeting,* New Orleans, LA, 1986.
42. T. N. Antonioni, A. Dutton, C. A. Landolt, A. Vahed: “Recent Developments in the Operation of Nickel Converters at Inco's Copper Cliff Smelter,” *23rd CIM Annual Conference of Metallurgists,* Quebec City, Quebec 1984.
43. J. W. Matousek, *CIM Bull.* **76** (1983) no. 856, 86 – 90.
44. R. G. Telewiak: “Canadian Nickel Mission to the U.S.S.R.,” (Sept. 1988), *Energy Mines and Resources Canada,* Ottawa 1990.
45. J. C. Mostert, P. N. Roberts, *TMS-AIME Pap.* **A 73 –48** (1973).
46. A. L. McKague, G. E. Norman, *CIM Bull.* **77** (1984) no. 866, 86 – 92. Links
47. G. E. Norman, R. E. Michelutti, *CIM Bull.* **77** (1984) no. 866, 93 – 98. Links
48. *Outokumpu News* **15** (1978) no. 3, 3 – 6.
49. “Outokumpu's Nickel Technology,” *Met. Bull. Monthly* **157** (1984) Jan, 53 – 60.
50. C. J. D. Williams in *Extraction Metallurgy '81*, Inst. Min. Metall. London 1981 , p. 214.
51. C. W. Hastie, D. R. Hall, C. A. Hohnen, J. M. Limerick: “Kalgoorlie Nickel Smelter: Integration of Flash Smelting and Slag Cleaning within one Process Unit”, *Aust. IMM Symposium on Extractive Metallurgy,* Melbourne 1984.
52. T. M. Young, W. P. Imrie: “Energy Developments in Nickel-Copper Smelting for Minimum Dependence on Oil Fuel,” *109th TMS-AIME Annual Meeting*, Las Vegas, Nevada 1980.
53. J. Asteljoki, J. Sulanto, T. T. Talonen in M. J. Jones, P. Gill (eds.): *Mineral Processing and Extractive Metallurgy*, Inst. Min. Metall., London 1984, pp. 171 –185.
54. J. S. Diakow, Y. F. Mak, R. G. Orr: “Metallurgy of the Converting Process in the Thompson Smelter,” *14th CIM Annual Conference of Metallurgists*, Edmonton, Alberta 1975.
55. R. J. Neal, R. A. Reyburn in R. E. Johnson (ed.): “Copper and Nickel Converters,” *TMS-AIME*, New York 1979, pp. 167 – 184.
56. A. A. Bustos et al. in [5] pp. 335 – 354.
57. S. S. Wang, N. H. Santander, J. M. Toguri, *Metall. Trans.* **5** (1974) Jan, 261 – 265.
58. H. H. Kellogg in [6] pp. 95 – 128.
59. P. E. Queneau, R. Schumann, Jr., US 3 941 587, 1976.
60. M. Y. Solar, R. J. Neal, T. N. Antonioni, M. C. Bell, *J. Met.* **31** (1979) no. 1, 26 – 31.
61. M. C. Bell, J. A. Blanco, H. Davies, P. Garritsen, *CIM Bull.* **83** (1990) no. 933, 47 – 50.
62. W. Curlook: “Fluid Bed Roasting of Nickel Sulfide,” *62nd CIM Annual General Meeting*, Toronto, Ontario 1960.
63. D. G. E. Kerfoot, D. R. Weir in [5] pp. 241 – 270.
64. F. A. Forward, US 2 576 314, 1951.
65. V. N. Mackiw, H. Veltman: “Recovery of Metals by Pressure Hydrometallurgy - The Sherritt Technology,” *Int. Min. Exhibition and Conference*, Calgary, Alberta 1980.
66. R. J. E. Stewart, J. L. Nixon, J. D. G. Groom: “Recovery of Nickel in Zimbabwe from Mixed Nickel Copper Sulfides by Outokumpu Technology,” in *Mining and Metallurgical Operations in Zimbabwe,* Harare, Zimbabwe 1983, vol. IV, 457 – 485.
67. P. Koskinen, M. Virtanen, H. Eerola in [5] pp. 355 –372.

68. R. P. Plasket, S. Romanchuk, *Hydrometallurgy* **3** (1978) 135 – 151. Links
69. C. F. Brugman, D. G. Kerfoot in [6] pp. 512 – 531.
70. J. L. Blanco, *World Min.* **31** (1978) March, 80 – 82.
71. Z. R. Llanos, P. E. Queneau, R. S. Rickard, *CIM Bull.* **67** (1974) no. 742, 75 – 81.
72. R. P. Plasket, G. M. Dunn in J. E. Dutrizac, A. J. Monhemius (eds.): *Iron Control in Hydrometallurgy*, Ellis Harwood, Chichester, United Kingdom 1986, pp. 695 – 718.
73. E. O. Stensholt, H. Zachariasen, J. H. Lund, P. G. Thornhill in [5] pp. 403 – 412.
74. P. Lenoir, A. van Peteghem, C. Feneau, *Proc. Int. Conf. Cobalt: Metall. Uses* 1981, 51 – 62.
75. S. Walker, *Int. Min.* **6** (1989, Nov.) 18 – 21.
76. M. L. Goble, J. A. Chapman in [6] pp. 464 – 480.
77. T. Inami, Y. Ishikawa, N. Tsuchida, T. Sugiura in [5] pp. 413 – 427.
78. D. G. E. Kerfoot in [6] pp. 426 – 441.
79. P. E. Queneau, C. E. O'Neill, A. Illis, J. S. Warner, *J. Met.* **21** (1969) no. 7, 35 – 45.
80. M. D. Head, V. A. Englesakis, B. C. Pearson, D. H. Wilkinson: “Nickel Refining by the TBRC Smelting and Pressure Carbonyl Route,” *105th TMS-AIME Annual Meeting*, Las Vegas, Nevada 1976.
81. L. G. Wiseman, R. A. Bale, E. T. Chapman, B. Martin in [5] pp. 373 – 390.
82. P. M. Tyroler, T. S. Sanmiya, E. W. Hodkin in [5] pp. 391 – 402.
83. T. Watanabe, S. Ono, H. Arai, T. Matsumori in [8] pp. 173 – 187.
84. A. A. Dor (ed.): *Nickel Segregation*, TMS-AIME, New York 1972.
85. A. S. Ericksen, J. Svensson, K. Ishii in [8] pp. 223 –236.
86. C. S. Simons in [5] pp. 91 – 134.
87. A. A. Dor, H. Skretting in [7] pp. 459 – 490.
88. K. Ishii in [8] pp. 15 – 24.
89. E. F. Osborn, A. Muan in M. K. Reser (ed.): *Phase Diagrams for Ceramacists*, American Ceramic Society, Columbus, Ohio 1964, p. 236.
90. B. Wasmund, G. G. Hatch, “Systematic Approach to the Design of a High Power Electric Matte Furnace,” *World Electrotechnical Congress*, Moscow 1977.
91. M. de Vernon in J. N. Anderson, P. E. Queneau (eds.):*Pyrometallurgical Processes in Non-Ferrous Metallurgy, TMS,* vol. 39, Gordon and Breach, New York 1967.
92. *Min. Mag.* **130** (1974) May, 336 – 349.
93. K. Yamada, T. Hiyama in [8] pp. 215 – 221.
94. T. Ogura, K. Kuwayama, A. Ono, Y. Yamada in [8] pp. 189 – 198.
95. E. Langer in [7] pp. 397 – 411.
96. S. C. C. Barnett, I. Patino, F. A. Perez, J. G. Schofield in *Extraction Metallurgy '85*, Inst. Min. Metall., London 1985.
97. R. R. Robertson, I. P. Vargas, *Eng. Min. J.* **186** (1985) no. 5, 18 – 22.
98. *Min. Mag.* **129** (1973) July, 12 – 19.
99. I. H. Keith et al. in [7] pp. 123 – 242.
100. *SLN Bull.* **1** (1976) 1.
101. T. T. Toomver in [7] pp. 252 – 271.
102. R. Musu, J. A. E. Bell in [7] pp. 300 – 322.
103. J. D. Guiry, A. D. Dalvi in [8] pp. 199 – 214.
104. M. C. Bell, W. P. Clement, *CIM Bull.* **83** (1990) no. 933, 57 – 59.
105. M. D. Sopko in [7] pp. 272 – 291.
106. R. Hoppe, *Eng. Min. J.* **178** (1977) no. 11, 123 – 126.
107. J. G. Reid in [7] pp. 368 – 381.
108. *Min. J. (London)* **313** (1989) July 7, 3.
109. D. R. Weir, V. B. Sefton in [7] pp. 325 – 345.
110. N. Colvin, J. W. Gulyas in [7] pp. 346 – 356.
111. *Met. Week* **58** (1987) August 31, 2.
112. Henkel Corp., *Tech/News* **1** (1988) 1.
113. J. G. Reid in [5] pp. 325 – 334.
114. L. C. Rodriquez, *Min. Cuba* **2** (1976) no. 3, 42 – 49.
115. S. L. Sobol *et al.*, *Min. Cuba* **3** (1977) no. 2, 2 – 18.
116. M. C. Pestana, G. G. Lahens, *Revista Tecnologica* **19** (1989) no. 1, 52 – 56.
117. P. T. O'Kane in [7] pp. 503 – 523.
118. W. P. C. Duyvesteyn, G. R. Wicker, R. E. Doane in [7] pp. 553 – 570.
119. G. R. Wicker, M. C. Jha in [6] pp. 566 – 577.
120. D. Urbain, J. P. Duterque, P. Palanque, P. Rey in [6] pp. 578 – 596.
121. E. Ozberk, S. W. Marcuson (eds.): *Nickel Metallurgy,* vol. II: “Industrial Applications of Nickel, ” The Metallurgical Society of CIM, Montreal, Quebec 1986.

122. P. G. Chamberlain in [121] pp. 193 – 215.
123. Inco Ltd., *Annual Reports*, Toronto 1977 and 1978.
124. Falconbridge Ltd., *Annual Report*, Toronto 1986.
125. I. D. Corrick: *Nickel, Mineral Commodity Profiles* (May 1979), U.S. Bureau of Mines, Washington, D.C., 1979.
126. M. O. Pearce in [8] pp. 5 – 14.
127. J. P. Schade in [121] pp. 1 – 16.
128. E. A. Thiers in [5] pp. 21 – 31.
129. J. P. Schade, M. O. Pearce in [5] pp. 53 – 63.
130. *Mining Annual Review*, 1989, C 73.
131. *Non Ferrous Metal Data 1989*, Amer. Bur. of Metal Statistics, Secausus, N. J., 1990, pp. 123 – 125.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Nickel Alloys

F. Werner Strassburg, International Nickel GmbH, Düsseldorf, Federal Republic of Germany

1. **Introduction** 1
2. **Wrought Nickel** 2
3. **Nickel – Iron Alloys** 3
4. **Nickel – Copper Alloys** 5
5. **Nickel – Chromium Alloys** 6
6. **Nickel Casting Alloys** 10
7. **Welding Filler Metals for Nickel Alloys** 11
8. **Nickel Brazing Alloys** 12
9. **Fabrication of Nickel-Base Alloys** . . . 12
10. **References** 13

1. Introduction

Nickel forms mixed crystals over a wide range with copper, manganese, iron, chromium, and other metals. It is mainly used as an alloying element in many nonferrous materials and steels (approximately 2000) to improve their strength, toughness, corrosion resistance, and high-temperature properties. About 60 % of the primary nickel produced in the Western World is used in the production of stainless steels, and a further 13 % is used in the production of nickel-base alloys, including pure nickel semi-finished products.

In 1988 primary nickel consumption by first use in the Western World broke down as follows [2]:

Alloy steels	6 %
Stainless steels	60 %
Nonferrous nickel-base alloys	13 %
Plating applications	10 %
Foundry industry	6 %
Other applications	5 %

Nickel alloys have been standardized in most industrial countries. Table 1 lists the most important materials standards in the USA. In the Federal Republic of Germany, the chemical compositions of nickel-base alloys are standardized in:

DIN 17 740	Wrought nickel
DIN 17 741	Wrought nickel alloys (low alloyed)
DIN 17 742	Wrought nickel alloys with chromium
DIN 17 743	Wrought nickel alloys with copper
DIN 17 744	Wrought nickel alloys with molybdenum and chromium
DIN 17 745	Wrought alloys of nickel and iron with other alloying elements

and the mechanical properties in:

DIN 17 750	Plate, sheet and strip of wrought nickel and nickel alloys
DIN 17 751	Tubes of wrought nickel and nickel alloys
DIN 17 752	Rod and bar of wrought nickel and nickel alloys
DIN 17 753	Wire of wrought nickel and nickel alloys

ISO Standards for nickel and nickel alloys have been under negotiation by the industrial national standard organizations for many years. Two basic ISO Standards have been issued so far: ISO 6372-01:1989 “Nickel and Nickel Alloys – Terms and Definitions, Part 1, Materials” and ISO 6372-03:1989 “Wrought Products and Castings”.

The following ISO Standards (still in the form of draft proposals) are expected to be published in 1991: ISO 9721, 9722, 9723, 9724, 9725, 6207, and 6208.

Nickel alloys are mainly used as corrosionresistant materials or for their excellent high-temperature strength or special physical properties. Accordingly, they can be divided into two principal groups:

1) Single-phase, solid-solution alloys that are primarily corrosion resistant at room and elevated temperature. These have face-centered cubic structures, moderate strength, and good weldability. Their applications often depend on specific physical properties.
2) Precipitation hardening alloys, most of which have high-temperature strength and creep resistance (up to 1050 °C). These alloys have a face-centered cubic matrix with precipitations. They have limited weldability (resistance welding or electron beam welding) or are unweldable.

Developments in manufacturing techniques (argon oxygen degassing and vacuum oxygen degassing methods) have enabled new nickel alloys with improved properties to be produced.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_221

Table 1. ASTM standards for common nickel alloys [3]

Material		Product form				
US designation (ASTM or UNS)	Comparable German grade or alloy composition	Plate, sheet, strip	Rod, bar	Seamless tube and pipe	Welded tube and pipe	Forgings
N 02 200	Ni 99	B 162	B 160	B 161/163	B 725/730	
N 02 201	LC-Ni99	B 162	B 160	B 161/163	B 725/730	
N 04 400	NiCu30Fe	B 127	B 164	B 163/165		B 564
N 04 405	(NiCu30FeS)		B 164			
N 06 600	NiCr15Fe	B 168	B 166	B 163/167	B 516/517	B 564
N 06 690	NiCr29Fe	B 168	B 166	B 167		
N 06 102	(NiCr15Mo3Nb3W3)	B 519	B 518	B 445		
N 06 625	NiCr21Mo9Nb	B 443	B 446	B 444	B 704/705	
N 06 007	(NiCr22Fe20Mo6.5CuNb)	B 582	B 581	B 622	B 619/626	
N 06 002	(NiCr22Fe18Mo9W)	B 435	B 572	B 622	B 619/626	
N 06 110	(NiCr31Mo9W3)	B 755	B 756	B 759	B 757/758	
N 06 001	(NiCr22Mo6.5Nb)	B 436				
N 06 975	(NiCr24Mo6Ti)	B 582	B 581	B 622	B 619/626	
N 06 985	(NiCr22Mo7)	B 582	B 581	B 622	B 619/626	
N 08 800	X10NiCrAlTi3220	B 409	B 408	B 163/407	B 514/515	B 564
N 08 330	X10NiCr3519	B 536	B 511		B 546/535/710	(B 515)
N 08 825	NiCr21Mo	B 424	B 425	B 163/423	B 704/705	
N 08 221	(NiCr21Mo6)	B 424	B 425	B 423		
N 08 020	(X6NiCrMoCuNb3520)	B 463	B 473	B 464/468/729	B 474	B 472
N 08 245	(NiCr20Mo2Mn5Nb)	B 590				
N 08 320	(X3NiCrMo26.22.5)	B 620	B 621	B 622	B 619/626	
N 10 665	NiMo28	B 333	B 335	B 622	B 619/626	
N 06 455	NiMo16Cr16Ti	B 575	B 574		B 619/626	
N 10 003	(NiMo17Cr7)	B 434	B 573			
N 10 276	NiMo16Cr15W	B 575	B 574		B 619/626	
N 06 022	NiCr21Mo14W	B 575	B 574		B 619/626	
N 07 001	(NiCr20Co14Mo4Ti3)		A 637			A 637
N 07 252	(NiCr19Co10Mo10TiAl)		A 637			A 637
N 07 500	(NiCr18Co16Mo4TiAl)		A 637			A 637
N 07 718	NiCr19NbMo		A 637			A 637
N 07 750	(NiCr15Fe7TiAl)		A 637			A 637
N 07 080	NiCr20TiAl		A 637			A 637

Oxide dispersion strengthened alloys with very high creep strengths are manufactured by powder metallurgy.

All nickel alloys may be further alloyed with aluminum, titanium, and/or niobium to render them age-hardenable. The age-hardened, creep-resistant Ni – Cr alloys are of particular importance.

The standard specifications of the United States (Table 1) or the Federal Republic of Germany are used to identify alloys when available; trade names are used to identify newer materials that have yet to be standardized.

The solid solution alloys, which are mostly corrosion resistant, are fusion welded in the construction of equipment, pipework, and vessels by using standardized matching filler metals.

Age-hardenable and oxide dispersion strengthened alloys cannot be fusion welded, and are either resistance welded or joined by using high-temperature nickel-base brazing alloys [1–6].

2. Wrought Nickel

Wrought nickel (DIN 17 740 and ASTM B160, B161, B162, B725, and B730) is used in the construction of equipment and vessels, for electrical and electronic components, and for coins.

The grades covered by DIN 17 740 are listed in Table 2. Weldable grades contain max. 0.02 % C to avoid precipitation of graphite in the heat-affected zones of the welds, as this would have a detrimental effect on strength properties. Pure nickel is ferromagnetic up to its Curie temperature (353 °C). Figure 1 shows the coefficient of thermal expansion of nickel as a function of temperature; other properties are given in [8]. Purity has a great effect on properties. The strength is

Table 2. Wrought nickel grades, standardized in DIN 17 740 [7]

Material		Composition, max %								
Designation	Number	Ni, min.∗	C	Cu	Fe	Mg	Mn	S	Si	Ti
Ni99.6	2.4060	99.6	0.08	0.1	0.2	0.15	0.3	0.005	0.1	0.1
LC-Ni99.6	2.4061	99.6	0.02	0.1	0.2	0.15	0.3	0.005	0.1	0.1
Ni99.2	2.4066	99.2	0.1	0.25	0.4	0.15	0.3	0.005	0.2	0.1
LC-Ni99	2.4068	99.0	0.02	0.25	0.4	0.15	0.3	0.005	0.2	0.1

∗ Includes max. 1 % Co.

Table 3. Tensile properties of plate (up to 50 mm thickness), sheet and strip (up to 2 mm thickness), DIN 17 750

Material designation	Material no.	Tensile strength, MPa	0.2 % Proof strength, MPa	Elongation δ_5, %	Brinell hardness
Ni99.6F370	2.4060.10	370	100	40	max. 100
Ni99.6F490	2.4060.26	490	290	20	ca. 130
Ni99.6F590	2.4060.30	590	490	2	ca. 200
LC-Ni99F340	2.4068.10	340	100	40	max. 100
LC-Ni99F430	2.4068.26	430	150	20	ca. 130
LC-Ni99F540	2.4068.30	540	430	8	ca. 180

low in the annealed condition: elongation and toughness are high (Table 3).

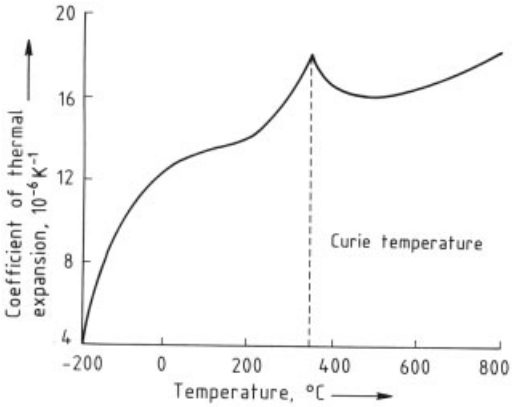

Figure 1. Coefficient of thermal expansion of pure nickel [8]

Since treatment of pure nickel does not produce any structural change, only cold working can increase its strength, the extent of the improvement depending mainly on the carbon content. The increase in strength disappears partly or completely upon recrystallization annealing [7–9].

3. Nickel – Iron Alloys

Nickel-base alloys with iron and other alloying elements (Co, Cr, Cu, and Mo) are solid solution alloys with face-centered cubic lattices. Owing to their physical properties, they are used as soft magnetic materials, as glass-to-metal seals, and as materials with defined thermal expansion properties. The nickel – iron phase diagram (Fig. 2) shows that, above 35 % nickel, the alloy is wholly in the γ-state, in which the magnetic transition has a temperature maximum at about 70 % Ni. The compositions and applications of some typical Ni – Fe alloys are given in Table 4.

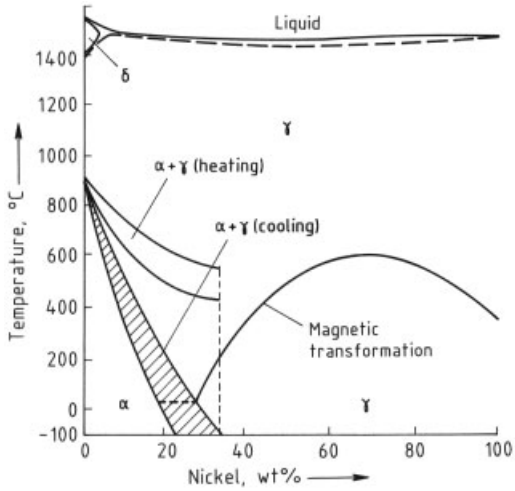

Figure 2. Binary phase diagram nickel – iron [12]

As well as the pure nickel – iron alloys, grades with cobalt and chromium additions are used, especially for metal-to-glass seals (Table 4, see next page).

Table 4. Nickel – iron alloys for main applications [12]

DIN designation	Material number [a]	Composition, wt %			
		Ni	Cr	Fe	Others
Soft magnetic alloys					
Ni48	1.3922	min. 46		49 – 53	
NiFe44	2.4420	min. 53		43 – 46	
NiFe10CuMo	2.4520	min. 75		13 – 17	4 – 6 Cu 3 – 5 Mo
NiFe16CuCr	2.4500	min. 75	1.5 – 2.5	15 – 18	4 – 6 Cu
NiFe15Mo	2.4540	min. 78		14 – 17	3 – 5 Mo
Glass-to-metal seals					
NiFe45	2.4472	min. 53		45 – 47	
NiFe46	2.4475	min. 51		46 – 48	
NiFe47	2.4478	min. 50		47 – 49	
NiFe47Cr [b]	2.4486	min. 47	5.5 – 6.5	47 – 49	
NiFe48Cr	2.4480	min. 50	0.7 – 1.0	47 – 49	
Ni49Cr	1.3921	min. 48	0.7 – 1.0	49 – 51	
Ni29Co18 [c]	1.3981	min. 28		53 – 55	17 – 18 Co
Ceramic-to-metal seals					
Ni42	1.3917	min. 41		56 – 58	
Ni28Co23	1.3982	min. 27		47 – 49	22 – 24 Co

[a] Materials with numbers in group 1 are iron-base alloys.
[b] Especially for soft lead glasses.
[c] Especially for borosilicate glasses.

The pure Ni – Fe alloys do not rust in air and have limited corrosion resistance. Their notched bar impact toughness values, fracture elongations, and reduction in area are high, even at low temperatures. The physical properties of the binary Ni – Fe alloys are given in [12].

Alloys with Controlled Thermal Expansion and for Glass-to-Metal Seals. Alloys having a nickel content of 35 – 55 % Ni have low coefficients of thermal expansion within the temperature range 0 – 100 °C (Fig. 3). For nickel contents above 55 % the expansion coefficients at 0 °C are higher than that of the alloy with 36 % Ni; however, they increase less rapidly with increasing temperature (Fig. 3). Typical applications of these alloys include tape measures and precision equipment, thermostats, temperature regulators, and bimetallic strips.

Whereas addition of other metals usually increases the thermal expansion coefficient of iron – nickel alloys, at room temperature the addition of cobalt lowers it. The cobalt/nickel ratio must be fixed so that the transition temperature Ar_3 is not above −86 °C and the material is completely in the γ-area. Most commercial Ni – Fe – Co alloys contain 27 – 35 % Ni and 16 – 28 % Co, the remainder being Fe. Their main application is for making glass-to-metal seals. Seals for types of glass with higher expansion coefficients are made of Ni – Fe – Cr alloys.

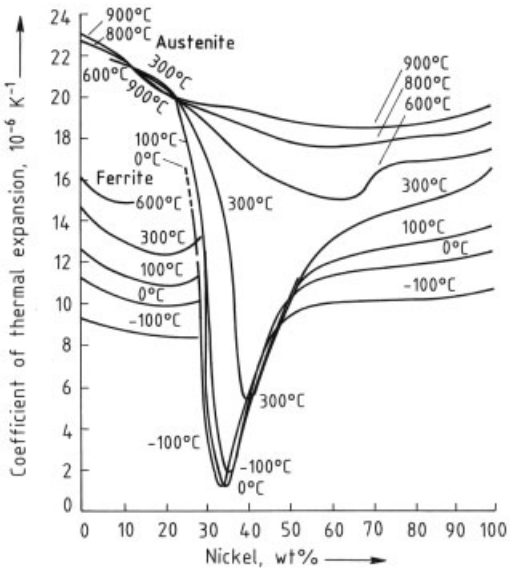

Figure 3. Isotherms of the coefficient of thermal expansion of annealed nickel – iron alloys [12]

Alloys with Controlled Modulus of Elasticity. The Young's modulus of Ni – Fe alloys increases with nickel content (Fig. 4). The moduli of elasticity of magnetically saturated and cold-

strengthened materials are higher than those of demagnetized or annealed materials [12]. The Ni – Fe alloys with temperature-dependent Young's moduli are used mainly for springs, diaphragms, and instrument components.

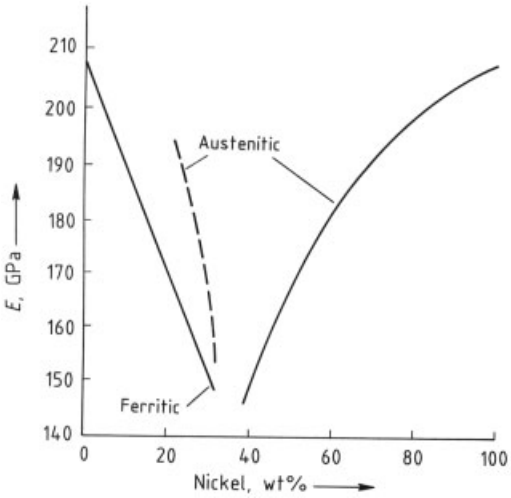

Figure 4. Modules of elasticity of nickel – iron alloys [12]

Soft Magnetic Nickel – Iron Alloys (see also → Magnetic Materials). Alloys containing 72 – 83 % Ni are crystalline materials with the best soft magnetic properties. Higher permeability values can be obtained by heat treatment. These alloys are used for magnetic shielding, medium-frequency transformers, chokes, and relays.

Alloys containing 45 – 68 % Ni have high saturation induction properties and are used for electrical measurement and leakage current protection in the 300 – 500 mA range. These alloys have a round hysteresis loop. A suitable combination of alloying and heat treatment can produce Ni – Fe alloys with rectangular or flat hysteresis loops. The latter also have very low remanence and small coercive field strengths.

Other soft magnetic materials are the amorphous alloys based on Ni, Fe, and Co, which contain additional alloying elements – boron and silicon – in amounts of 5 – 15 %. These inhibit crystallization. These alloys are called metallic glasses (→ Metallic Glasses) and are usually made as thin strips by extremely rapid cooling of a melt on a rotating metal roller. A typical amorphous alloy (Vibrovac 4040) with 40 % Ni, 40 % Fe, the remainder metalloids, has the following typical properties:

Saturation induction	T = 0.8
Saturation magnetostriction	8×10^{-6}
Static coercive field strength	10 mA/cm
Specific electrical resistance	$1.35\ \Omega \cdot mm^2/m$
Tensile strength	1500 – 2000 MPa

Fabrication of Nickel – Iron Alloys. The fully austenitic Ni – Fe alloys can be cold or hot worked. Cold hardening can be eliminated or reduced by annealing at 800 – 870 °C. Owing to the sensitivity of nickel towards sulfur (formation of a nickel – nickel sulfide eutectic), a sulfur-free annealing atmosphere is necessary. For etching of metallographic samples, a mixture of 10 % nitric acid and 10 % hydrochloric acid at ca. 70 °C is used. For machining of Ni – Fe alloys machining data and tool design used for austenitic Cr – Ni steels can be approximately applied [10–12].

4. Nickel – Copper Alloys

The binary phase diagram for nickel and copper shows mixed crystals for all ratios of nickel to copper (Fig. 5). There are therefore many technically used Ni – Cu alloys, the most important being those covered by DIN 17 743 (Table 5). The strength values for the various semi-finished products are given in DIN 17 750 – 17 753 and ASTM B127, B163, B164, B165, and B564.

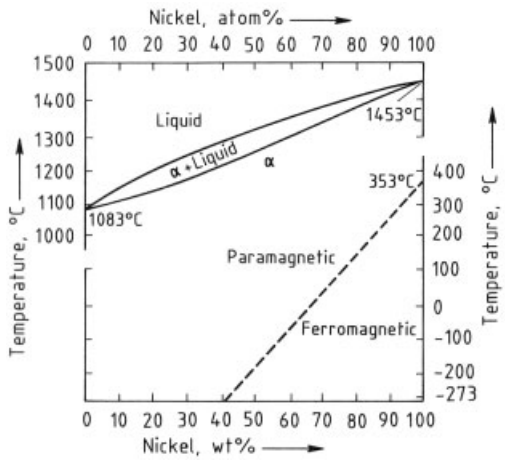

Figure 5. Binary phase diagram nickel – copper [13]

Table 5. Composition of wrought nickel – copper alloys [14]

Designation	Material number	Chemial composition, wt %									Typical application
		Ni + Co ∗	Al	C	Cu	Fe	Mn	S	Si	Ti	
NiCu30Fe	2.4360	min. 63			28	1.0					corrosion-resistant
		max.	0.5	0.15	34	2.5	2.0	0.02	0.5		components
LC-NiCu30Fe	2.4361	min. 63			28	1.0					corrosion-resistant
		max.	0.5	0.04	34	2.5	2.0	0.02	0.3		components
NiCu30Al	2.4375	min. 63	2.2		27	0.5				0.3	age-hardenable
		max.	3.5	0.2	34	2.0	1.5	0.015	0.5	1.0	alloy for corrosion-resistant components

∗ Max. 1 % Co.

Nickel – copper alloys do not diffuse readily. Rapid cooling from the liquid state produces inhomogeneous single-phase crystals in which the center is rich in nickel. Hot working and annealing produce a homogeneous structure in wrought semi-finished products.

Nickel – copper alloys are very resistant to corrosion by alkaline solutions, nonoxidizing salts, and seawater. The addition of 1 – 2 % Fe improves corrosion resistance. The precipitation-hardenable alloy NiCu30Al is important as a high-strength, corrosion-resistant material for components, such as propeller shafts of ships, that remain in contact with seawater for long periods.

Welding materials for Ni – Cu alloys are specified in DIN 1736, parts 1 and 2 (see also Chapter 7) [13, 14].

5. Nickel – Chromium Alloys

Nickel and chromium form solid solutions over a wide range. Nickel – chromium alloys are characterized by their high corrosion resistance at normal and elevated temperatures (resistance to scaling), good high-temperature strength, and high electrical resistance. There are basically three groups of alloys:

1) Ni – Cr – (Fe) alloys with high electrical resistance for heating elements
2) Ni – Cr alloys (with Fe and other alloying elements) with good corrosion resistance
3) Ni – Cr alloys with high-temperature strength and creep resistance, mostly age-hardenable.

Nickel – Chromium – (Iron) Electrical Resistance Alloys. Nickel-based alloys are used for heating elements and electrical shunts: their compositions are given in Table 6 and their physical properties in [15]. Since these alloys are used as heating elements at temperatures which can exceed 1000 °C and are sometimes required to undergo rapid cooling, they must be resistant to high temperatures and rapid temperature changes as well as to oxidizing, neutral, and reducing atmospheres [15]. The electrical resistance depends on temperature and heat treatment [15].

Corrosion-Resistant Nickel – Chromium – Iron Alloys. These alloys have significantly better corrosion resistance than austenitic Cr – Ni

Table 6. Composition and density of some electrical resistance alloys [18]

DIN designation	DIN material number		Composition, wt %								Density, g/cm^3
			Ni + Co ∗	Al	C	Cr	Cu	Fe	Mn	Si	
NiCr8020	2.4869	min.	75			19				0.5	
		max.		0.3	0.15	21	0.5	1.0	1.0	2.0	8.3
NiCr7030	2.4658	min.	60			29				0.5	
		max.		0.3	0.10	32	0.5	5.0	1.0	2.0	8.2
NiCr6015	2.4867	min.	59			14		19		0.5	
		max.		0.3	0.15	19	0.5	25	2.0	2.0	8.2
NiCr20AlSi	2.4872	min.	73	2.5		19				0.5	
		max.		4.0	0.05	21	0.1	1.0	0.7	2.0	8.0

∗ Co max. 1.0 %.

Table 7. Composition and density of corrosion-resistant Ni – Cr and Ni – Cr – Mo alloys [18]

ASTM/UNS designation	DIN designation	DIN material number		Composition, wt %													Density, g/cm^3
				Ni	Al	C	Co	Cr	Cu	Fe	Mn	Mo	Nb/Ta	Si	Ti	Others	
N 06 600	NiCr15Fe	2.4816	min.	72				14		6							
			max.		0.3	0.10	1.0	17	0.5	10	1.0			0.5	0.3		8.5
	LC-NiCr15Fe	2.4815	min.	72				14		6							
			max.		0.3	0.025	1.0	17	0.5	10	1.0			0.5	0.3		8.5
	NiCr23Fe	2.4851	min.	58	1.0			21									
			max.	63	1.7	0.10	1.0	25	0.5	18	1.0			0.5	0.5		8.2
N 06 690	NiCr29Fe	2.4642	min.	58				27		7							
			max.			0.05	1.0	31	0.5	11	0.5			0.5			8.1
N 06 455	NiMo16Cr16Ti	2.4610	min.	bal.				14.0				14.0					
			max.			0.01	2.0	18.0	0.5	3.0	1.0	18.0		0.08	0.7		8.6
N 10 665	NiMo28	2.4617	min.	bal.								26.0					
			max.			0.01	1.0	1.0	0.5	2.0	1.0	30.0		0.08			9.2
N 06 007	NiCr22Mo6Cu	2.4618	min.	bal.				21.0	1.5	18.0	1.0	5.5	1.75				
			max.			0.05	2.5	23.5	2.5	21.0	2.0	7.5	2.50	1.0		1.0 W	8.3
N 06 985	NiCr22Mo7Cu	2.4619	min.	bal.				21.0	1.5	18.0		6.0					
			max.			0.015	5.0	23.5	2.5	21.0	1.0	8.0	0.50	1.0		1.5 W	8.3
	NiCr21Mo6Cu	2.4641	min.	39.0				20.0	1.5	rem.		5.5			0.6		
			max.	46.0	0.20	0.025	1.0	23.0	3.0		1.0	7.0		0.5	1.0		8.3
N 10 276	NiMo16Cr15W	2.4819	min.	bal.				14.5		4.0		15.0				3.0 W	
			max.			0.015	2.5	16.5	0.5	7.0	1.0	17.0		0.08		0.35 V 4.5 W	8.9
N 06 625	NiCr22Mo9Nb	2.4856	min.	bal.				20.0				8.0	3.15				
			max.		0.40	0.10	1.0	23.0	0.5	3.0	0.5	10.0	4.15	0.5	0.40		8.4
N 08 825	NiCr21Mo	2.4858	min.	38.0				19.5	1.5	rem.		2.5			0.6		
			max.	46.0	0.2	0.025	1.0	23.5	3.0		1.0	3.5		0.5	1.2		8.1
	NiCr23Co12Mo **	2.4663		54.0	1.0	0.07	12.5	22.0		0.13	0.03	9.0		0.04	0.35		8.2
N 06 022	NiCr21Mo14W	2.4602	min.	bal.				20.0		2.0		12.5				2.5 W	
			max.			0.015	2.5	22.5		6.0	0.5	14.5		0.08		0.35 V 3.5 W	8.6

* Approximate values.
** Typical values.

Table 8. Typical composition of some creep-resistant superalloys

DIN designation	DIN material number	Trade name	Composition, %								Other elements, remarks
			Ni	Co	Cr	Mo	C	Al	Ti	Nb	
NiCr20Ti	2.4630	Nimonic 75	79		19.5		0.10		0.4		
NiCr20TiAl	2.4631	Nimonic 80A	74		19.5		0.07	1.4	2.3		
NiCr20Co18Ti	2.4632	Nimonic 90	55	18	19.5		0.10	1.4	2.4		
NiCo20Cr15MoAlTi	2.4634	Nimonic 105	52	20	14.5	5	0.16	4.7	1.2		
NiCr13A16MoNb	2.4670	Alloy 713LC	74		12	4.5	0.05	5.9	0.6	2.0	0.1 Zr, cast alloy
NiCr19NbMo	2.4668	Alloy 718	52.5		19	3	0.04	0.5	0.9	5.0	
NiCr23Co19TiNb		Alloy 739	48	19	22.5		0.15	1.9	3.7	2.4	2 W, 0.1 Zr, cast alloy
X1NiCrNb38 13		Alloy 909	38		13		0.01	0.03	1.5	4.7	balance Fe
NiCo20Cr20MoTi	2.4650	Alloy 263	51	20	20	5.9	0.03	0.5	2.15		Zr addition
(NiCr18Co14Mo7AlTi)		Alloy PK33	56	14	18	7	0.04	2.1	2.2		
(NiCr22Mo9Fe)		Alloy PE13	49	1.5	22	9	0.07				

steels and continue the alloy series Ni – Cr – Fe to higher nickel contents. The many alloys are resistant to wet corrosion at normal and elevated temperature and to corrosive gases and slags at high temperature, and the most important of them have been given standard specifications in the industrialized countries. An overview is given in Table 7 (see previous page). Alloys for rolling and forging with increased chromium contents up to 30 % have been developed to improve resistance to corrosion and scaling (Table 7). Strength properties are specified in DIN 17 750 – 17 753 and ASTM (Table 1). If good strength properties at high temperatures are required, the age-hardenable Ni – Cr alloys are used [16–20].

Not only are Ni – Cr – Fe alloys more resistant to general corrosion than stainless steels, but they are also resistant to stress corrosion cracking and pitting in many environments.

Nickel – Chromium – Molybdenum Alloys and Nickel – Molybdenum Alloys. If even better resistance to wet corrosion is required, nickel – chromium alloys with molybdenum additions of 3 – 16 % and sometimes other alloying elements (e.g., Cu) are specified. The Ni – Cr – Mo alloys are particularly resistant to oxidizing acids and gases, whereas the Ni – Mo alloy is resistant to reducing acids. Compositions are given in Table 7 [21–24].

Creep-Resistant, Age-Hardenable Nickel Alloys with High-Temperature Strength [25–29]. With the elements Ti, Al, and Nb, nickel forms intermetallic compounds of the type Ni_3Ti (Al, Nb) which on age hardening (e.g., 16 h at 700 °C) precipitate in a finely dispersed state as a γ'-phase within the basic austenitic structure and at the grain boundaries. The resistance to slipping so produced gives good high-temperature strength and creep resistance, defined as creep rupture strength or creep strength for a defined stress and duration (e.g., 10 000 h). Although wrought nickel and all nickel-base alloys may be artificially age-hardened if alloyed with Ti, Al, and Nb, this process is mainly used with Ni – Cr – (Fe) alloys. The degree of creep resistance is determined by the total content of Ti, Al, and Nb. Vacuum melting and control of grain size and structure also contribute to improvement of creep resistance. The strength of the matrix can be improved by alloying with Co, V, and other elements.

From the large number of age-hardenable superalloys, some widely used typical alloys are summarized in Table 8. Creep rupture strength curves are given in Fig. 6 (see next page). More recently developed grades contain preferentially niobium as an ageing element, if fusion weldability is required (e.g., for gas turbine components made of sheet). Age-hardenable alloys with a reduced Ni content of 38 % (e.g., X1NiCrNb 38 13) have low coefficients of thermal expansion (Fig. 7), which enables very close tolerances of turbine components (higher power, reduced fuel consumption).

Although higher Cr contents have a negative influence on the creep strength of superalloys, they significantly improve the corrosion resistance (Fig. 8).

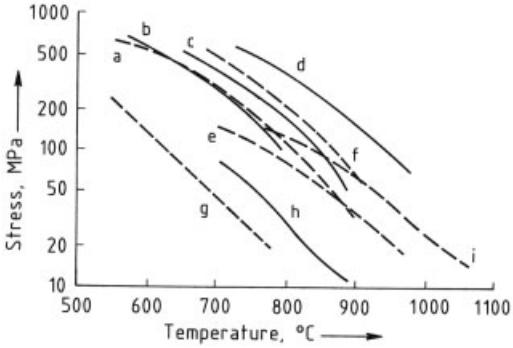

Figure 6. Creep rupture strength $\sigma_{B.1000}$ of age-hardenable nickel base alloys
a) Alloy 263; b) NiCr20TiAl (Nimonic 80A); c) NiCr20 Co18Ti (Nimonic 90); d) NiCo15Cr15MoAlTi (Nimonic 115); e) Alloy PE 13; f) NiCr18Co14Mo7AlTi (Alloy PK 33); g) NiCr20Ti (Nimonic 75) sheet; h) NiCr20Ti (Nimonic 75) wrought; i) NiCr23Co12Mo (Alloy 617)
—— Forged components; – – – – – Sheet alloys

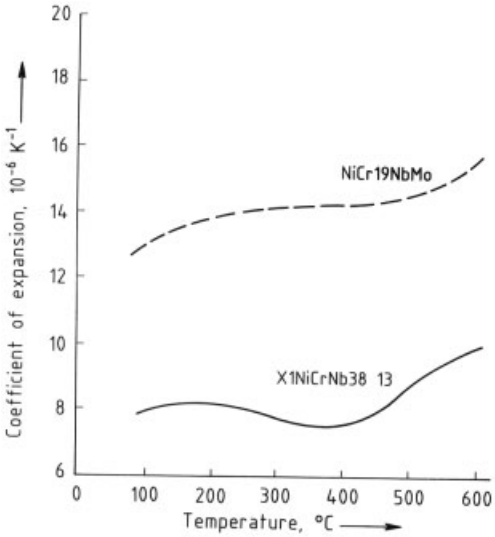

Figure 7. Coefficient of thermal expansion of two creep resistant superalloys [30]

Oxide Dispersion Strengthened and Mechanically Alloyed Superalloys (→ High-Temperature Materials). Nickel alloys with very good high-temperature strengths at operating temperatures above 1100 °C can be produced by powder metallurgical processes. The matrix powder, which usually consists of an atomized precipitation-hardenable alloy powder, e.g., NiCr20Ti, is mixed with a metal oxide powder (e.g., Y_2O_3) in an attritor. The mechanically alloyed powder mixture is then hot hydrostatically pressed into blooms and then forged or hot rolled.

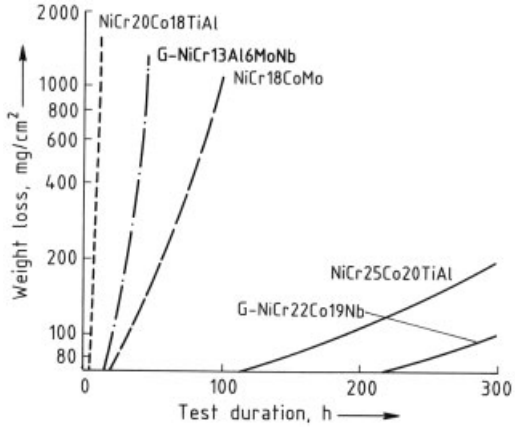

Figure 8. Corrosion resistance against 75 % Na_2SO_4/25 % NaCl at 900 °C

Precipitation hardening by Ni_3Ti and dispersion hardening by metallic oxide are thus combined and establish the superior creep resistance of the alloys over a wide range of temperatures (Fig. 9). The chemical compositions

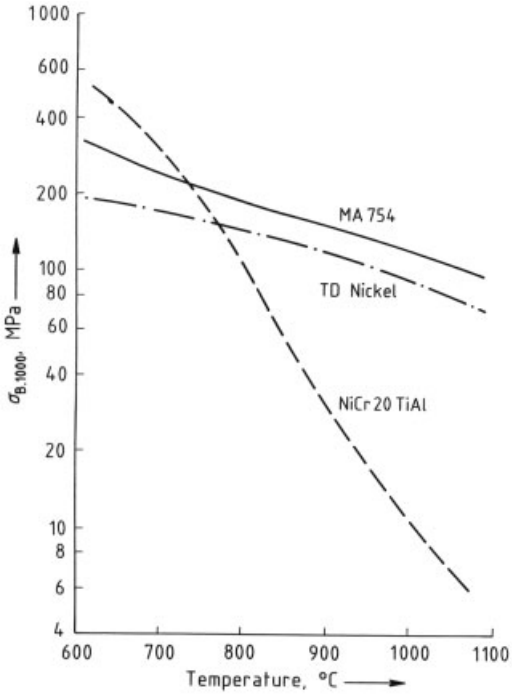

Figure 9. Creep rupture strength of oxide dispersion hardened superalloy MA 754 compared with TD nickel and NiCr20TiAl (Nimonic 80A) [33]

Table 9. Composition of two commercial oxide dispersion strengthened superalloys [33,34]

Alloy	Composition, wt %							
	Ni	Cr	Al	Ti	B	Zr	Y_2O_3	Others
Inconel MA 754	bal.	20	0.3	0.5		0.6		
Inconel MA 6000	bal.	15	4.5	2.5	0.01	0.15	1.1	2.0 Ta, 2.0 Mo, 4.0 W

of such alloys, also known as ODS (oxide dispersion strengthened superalloys), and used in industry as high-temperature construction materials, are given in Table 9. Above 900 °C, creep rupture strengths of ODS alloy MA 6000 are superior to those of the high-temperature superalloy Mar M200 + Hf made by melting and forging.

6. Nickel Casting Alloys

All nickel alloy groups include foundry alloys whose composition is optimized for casting. Most nickel alloys for casting contain 1.5 – 2 % Si, some grades up to 6 % Si, to improve erosion and abrasion resistance (Table 10). Resistance to seizing is also improved by silicide formation by aging (soft annealing, then treatment at 800 °C for 4 – 6 h). With pure nickel (91 – 96 % Ni),

Table 10. Composition of nickel and nickel alloy castings [37,38]

ASTM ∗ designation	DIN ∗∗ designation	Composition, wt %									
		C	Cr	Cu	Fe	Si	Mn	Ni	Mo	Others	
CZ 100		min.						95			
		max. 1.0		1.25	3.0	2.0	1.5				
	G-Ni97	min.				0.2		95			
		max. 1.0		1.0	1.0	1.0	1.5				
	G-Ni94C	min. 1.0				1.0		93			
		max. 2.5		1.0	1.0	2.0	1.5				
	G-Ni92Si	min.				5.5		90			
		max. 1.0		1.0	1.0	6.5	1.5				
M-35-1		min.		26.0				bal.			
		max. 0.35		33.0	3.5	1.25	1.5				
M-35-2		min.		26.0				bal.			
		max. 0.35		33.0	3.5	2.0	1.5				
M-30H		min.		27.0		2.7		bal.			
		max. 0.30		33.0	3.5	3.7	1.5				
	G-NiCu30Si3	min.		27.0	1.0	2.7	0.5	64		0.5	
		max. 0.25		33.0	2.5	3.7	1.5	68		1.5 Al	
M-25S		min.		27.0		3.5		bal.			
		max. 0.25		33.0	3.0	4.5	1.5				
	G-NiCu30Si4	min.		27.0	1.0	3.5	0.5	60			
		max. 0.25		31.0	2.5	4.5	1.5	68			
M-30C		min.		26.0		1.0		bal.			
		max. 0.30		33.0	3.5	2.0	1.5				
	G-NiCu30Nb	min.		26.0	1.0	0.5	0.5	62		1.0	
		max. 0.15		33.0	2.5	1.5	1.5	68		1.5 Nb	
N-12MV		min.			4.0			bal.	26.0	0.20	
		max. 0.12	1.0		6.0	1.0	1.0		30.0	0.60 V	
Ni-7M		min.						bal.	30.0		
		max. 0.07	1.0		3.0	1.0	1.0		33.0		
CY-40		min.	14.0					bal.			
		max. 0.40	17.0		11.0	3.0	1.5				
CW-12MV		min.	15.5		4.5			bal.	16.0	3.75	0.20
		max. 0.12	17.5		7.5	1.0	1.0		18.0	5.25 W	0.40 V
CW-6M		min.	17.0					bal.			
		max. 0.07	20.0		3.0	1.0	1.0				
CW-2M		min.	15.0					bal.	15.0		
		max. 0.02	17.5		2.0	0.80	1.0		17.0	1.0	
CW-6MC		min.	20.0					bal.	8.0	3.15	
		max. 0.06	23.0		5.0	1.0	1.0		10.0	4.15 Nb	

∗ ASTM A 494/A 494 M
∗∗ DIN 17 730.

higher carbon contents lead to graphite precipitation and improvement of galling resistance; mechanical and physical properties are given in [37,38]. With Si contents up to 1.6 %, fabrication welding is possible to a limited extent using filler metals of similar type, as standardized in DIN 1736 (see Chap. 7). At higher Si contents these materials are not weldable. The Ni – Cu and Ni – Cr – Fe grades with 1.2 – 2 % Nb have good weldability. Construction welding can be applied as well as joining with components made of wrought alloys.

Castings with good resistance to hot gases and ash from crude oil combustion can be produced with Ni – Cr alloys containing 40 – 60 % Cr. Alloying with ca. 1.5 % Nb gives improved high-temperature strength and elongation. Although these alloys cannot be hot worked owing to their high chromium content, there are filler metals of similar composition produced using a special metallurgical treatment process, so that coated electrodes, wires, and cast rods can be manufactured for fusion welding. Compositions and properties of the NiCr50 alloy with and without Nb addition are given in [42]. The addition of niobium improves the high-temperature strength [43] and high-temperature corrosion resistance of NiCr50.

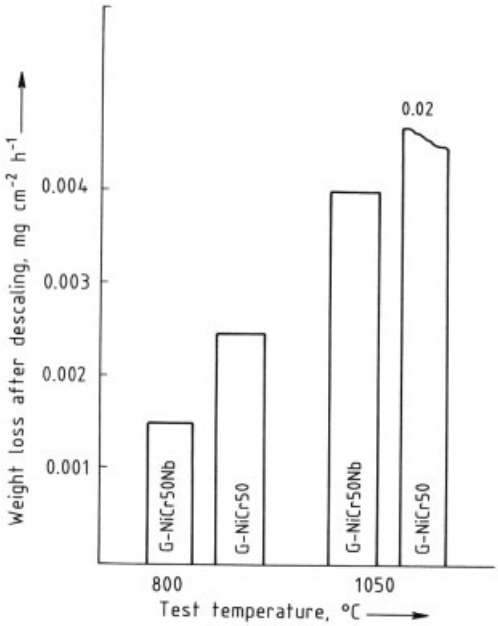

Figure 10. Oxidation resistance of G-NiCr50 and G-NiCr50Nb (duration of test: 1000 h) [43]

A typical application is for tube support hangers for chemical furnaces and power station boilers fired with heavy crude oil (Table 11 and Figure 10) [37–43].

Table 11. Corrosion rate of various heat resistant alloys in a crude oil heater [43] ∗

Composition, wt %			Corrosion rate, mm/a
Cr	Ni	Fe	
60	40		0.150
50	50		0.135
35	65		0.135
25	20	bal.	0.650
25	12	bal.	0.910

∗Fuel analysis: 3.2 % S, 250 – 400 ppm V, and 180 ppm Na; Test duration: 40 months; Tube wall temperature: ca. 900 °C.

7. Welding Filler Metals for Nickel Alloys

For solid solution alloys, which are very suitable for welding, the weld material is supplied in the form of bare wires, wire electrodes, and coated electrodes for all fusion welding processes with the exception of gas welding (oxy-acetylene). These materials are standardized in the following standards: Germany: DIN 1736, parts 1 and 2 [46,47]; United States: AWS specifications A 5.11 and A 5.14 [49,50]; United Kingdom: BS 2901, parts 1 and 2.

The strength properties of the weld metals listed in DIN 1736, part 2 exceed those of the solid solution alloys in the annealed base material. The components are normally welded in the soft-annealed condition. During welding, pure nickel and the Ni – Cu alloys are sensitive to all forms of sulfur, which forms a nickel – nickel sulfide eutectic on heating above 400 °C. This penetrates along the grain boundaries and causes hot and cold cracking under tensile stress. To prevent nitrogen pores in the weld metal, the welding materials contain Ti, Al and/or Nb, which combine with nitrogen to form nitrides; they also react with oxygen to form oxides and with carbon monoxide to give carbides.

The Ni – Cr – Fe alloys that contain 30 – 40 % Ni (e.g., X10NiCrAlTi3220 and NiCr21Mo) are rather susceptible to hot cracking (Fig. 11, see next page) and are preferably welded with filler metals of higher nickel content. The Ni – Cr

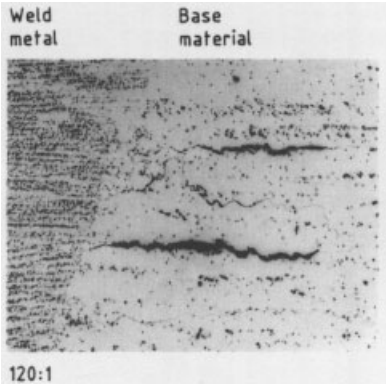

Figure 11. Hot cracks in the heat affected zone of alloy 800 (X10NiCrAlTi3220)

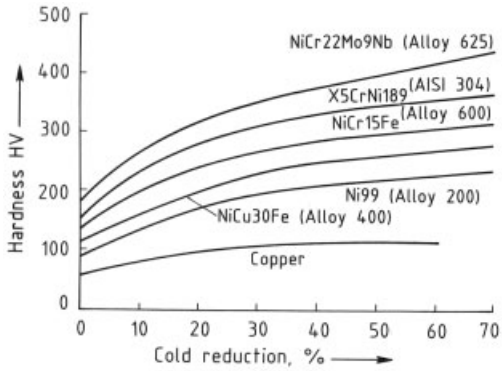

Figure 12. Effect of cold work on hardness of nickel alloys [60]

welding materials are also specified for joining different base materials (e.g., austenitic to ferritic steels) and for joining low-temperature nickel steels.

Age-hardenable high-temperature alloys have limited weldability or cannot be fusion welded. They are brazed (Chap. 8), resistance welded, or if possible welded by the MIG or WIG processes, with or without a matching filler metal. The number of layers is limited because of partial precipitation hardening from the welding heat. Fusion welded joints in age-hardened alloys have a creep strength that is generally 10 – 15 % lower than that of the base material [44–52].

8. Nickel Brazing Alloys

Nickel brazing alloys are used for high-temperature brazing of nickel, cobalt, and superalloys if these materials are not suitable for welding. The nickel-base brazing alloys are specified in DIN 8513 part 5 and in corresponding the ISO Standard. Owing to their composition, they combine low melting temperature with high-temperature strength.

They are supplied as pastes, powders, strip, and wires. Brazing is usually carried out in vacuum, under argon, or sometimes in a reducing atmosphere.

For vacuum brazing, the carefully cleaned component is positioned in the furnace which is evacuated to 0.1 Pa. The component is then heated to the brazing temperature and, after a short holding time, cooled in the vacuum.

The brazing process is used for bonding engine and gas turbine parts, in nuclear technology, and for fabrication of high-tech components [53–59].

9. Fabrication of Nickel-Base Alloys

Hot Forming. Wrought nickel-base alloys are generally supplied in the soft annealed or solution annealed condition. Hot forming is usually performed by the producers of the alloys or by specialized fabricators of forgings, tube bends, dished ends, etc. Single-phase nickel alloys have to be heated for both annealing and hot forming in sulfur-free furnace atmospheres to temperatures between 1130° and 1180 °C, depending on the type of alloy. Hot forming must be performed between 950 and 1000 °C. Rapid cooling through the 750 to 550 °C temperature range is recommended for Ni – Cr alloys, to avoid carbide precipitation. Soft annealing is advisable after all hot forming operations.

Forging of age-hardenable alloys is more critical because of the narrow temperature range in which hot forming can be performed and due to the high strength of such alloys at the lower limit of the forging temperature range.

Cold Forming. All single-phase alloys with a face-centered cubic structure have excellent cold formability due to the relatively large difference between the yield strength and the ultimate tensile strength. During cold forming, nickel-base alloys work harden more rapidly than unalloyed steel or copper (Fig. 12). Due to

the high modulus of elasticity, similar to that of austenitic stainless steel, (180 – 220 GPa), a certain amount of spring back occurs during cold forming. The strain rate influences the work hardening: fast forming produces a higher proof strength than slow forming. For cold forming operations, heavy duty lubricants have to be applied to avoid tool to metal seizing and wear.

Heat Treatment. Heat treatment may be necessary after cold forming to achieve a fine-grain structure (recrystallisation) and improved ductility. For single-phase alloys, soft annealing is normally sufficient. Precipitation hardening alloys must be solution annealed before aging.

If the higher strength resulting from cold forming is to be maintained, stress relieving can be applied. Heat treatment is performed at the following temperatures:

	stress relieving	soft annealing
Ni99	400 – 500 °C	820 – 920 °C
NiCu30Fe	500 – 650 °C	870 – 980 °C
NiCr15Fe	650 – 750 °C	950 – 1100 °C

Pickling. Pickling procedures differ for nickel alloys according to the following groups:

1) pure nickel and Ni – Cu alloys
2) Ni – Fe alloys
3) Ni – Cr and Ni – Cr – Mo alloys.

The pickling treatment depends also on the surface condition of the metal. Producers of nickel alloy semifinished products apply descaling by shot blasting and pickling in a solution containing 20 – 25 % nitric acid and 3 – 5 % hydrofluoric acid at ca. 40 °C. For fabricators the removal of tarnish and oxides from the surface is dominant: a solution of water, sulfuric acid, and salt (NaCl) can be applied at 20 – 40 °C for a few seconds. For Ni – Cu alloys, tarnish can be removed with an aqueous solution of nitric acid and salt at 20 – 40 °C within a few seconds. The Ni – Cr – (Fe) alloys require more sophisticated procedures; use of fused salts for pretreatment is recommended. For continuous operations, electrolytic baths are also applied. Special formulas have been developed for removal of lead and zinc (residues from tools), iron inclusions in the surface, and machining fluids, which can cause discoloration. Details on suitable pickling solution formulations can be obtained from producers of the alloys [59, 60].

10. References

General References

1. K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970.
2. J. P. Schade, *Stahl Eisen* **109** (1989) 539 – 542.
3. *Annual Book of ASTM Standards,* **vol. 02.04,** Sect. 2, "Non-Ferrous Metals–Nickel, Lead, Tin Alloys, Precious Metals", Philadelphia 1989.
4. *Landolt-Börnstein,* **V 2 b,** 307 – 567.
5. Metallgesellschaft AG (W. Bauer, ed.): *Metallstatistik 1977 – 1989,* vol. 77, Frankfurt 1990.
6. S. J. Rosenberg (ed.): "Nickel and its Alloys," *NBS Monogr.* **106,** Washington 1968.

Specific References

7. DIN 17 740: Nickel in Halbzeug, Zusammensetzung, 1983.
8. Nickel-Informationsbüro GmbH: *Nickel – Eigenschaften, Verarbeitung und Verwendung,* **Publ. No. 32,** Düsseldorf 1964.
9. Huntington Alloys: *Nickel 290,* **Publ. No. S-67,** Huntington 1984.
10. Henry Wiggin & Co. Ltd.: *Wiggin-Nickellegierungen für Elektronik und Instrumente,* **Publ. No. 3612,** Hereford 1973.
11. Vacuumschmelze GmbH: *Our Products, Technical Data,* **Publ. No. 1B-A03,** Hanau 1987.
12. International Nickel GmbH: *Eisen-Nickel-Legierungen,* **Publ. No. 68,** Düsseldorf 1973.
13. Nickel-Informationsbüro GmbH: *Nickel-Kupfer-Legierungen,* **Publ. No. 20,** Düsseldorf 1963.
14. DIN 17 743: Nickel-Knetlegierungen mit Kupfer, Zusammensetzung, 1983.
15. Henry Wiggin & Co. Ltd.: *Wiggin Electrical Resistance Alloys,* **Publ. No. 23018E,** Hereford 1978.
16. R. Köcher, *Metall (Berlin)* **43** (1989) 731 – 740.
17. E.-M. Horn, M. Renner, *Metall (Berlin)* **43** (1989) 720 – 727.
18. DIN 17 742: Nickel-Knetlegierungen mit Chrom, Zusammensetzung, 1983.
19. Henry Wiggin & Co. Ltd.: *Wiggin Corrosion Resistant Alloys,* **Publ. No. 3682,** Hereford 1988.

20. Wiggin Alloys Ltd.: *Inconel Alloy 601,* **Publ. No. 3662,** Hereford 1981.
21. Huntington Alloys: *Resistance to Corrosion,* **Publ. No. S-37,** Huntington 1978.
22. Hungtington Alloys: *Alloy G.,* Data Sheet, Huntington 1978.
23. Inco Alloys International: *Alloy G-3,* Data Sheet, Huntington 1986.
24. Wiggin Alloys Ltd.: *Inconel Alloy 625,* **Publ. No. 3700,** Hereford 1981.
25. Inco Ltd.: *High Temperature High Strength Nickel Base Alloys,* **Publ. No. A 393,** Toronto 1977.
26. Henry Wiggin & Co. Ltd.: *Super Alloys for Gas Turbine Engineering,* **Publ. No. 3755,** Hereford 1980.
27. W. Betteridge, J. Heslop: *The Nimonic Alloys,* 2nd ed., New York 1974.
28. Henry Wiggin & Co. Ltd., Hereford, *Nimonic Alloy 75,* **Data Sheet No. 3731,** 1979. *Nimonic Alloy 80A,* **Data Sheet No. 3613,** 1982. *Nimonic Alloy 115,* **Data Sheet No. 3732,** 1979. *Nimonic Alloy 90,* **Publ. No. 25038E,** 1978. *Nimonic Alloy 263,* **Publ. No. 3463,** 1970. *Nimocast Alloy 739,* **Publ. No. 6389,** 1977. *Nimonic Alloy PE 13,* **Data Sheet No. 3695,** 1981. *Nimonic Alloy PE 16,* Publ. No. 3349, 1976. *Alloy 713 C,* **Publ. No. 4015,** 1966.
29. Huntington Alloys: *Inconel Alloy 617,* **Publ. No. T-46,** Huntington 1979.
30. Inco Alloys International: *Incoloy Alloy 909,* **Publ. No. 18,** Huntington 1987.
31. Henry Wiggin & Co. Ltd.: *Super Alloys by Power Atomisation,* **Publ. No. 3712,** Hereford 1978.
32. J. S. Benjamin, *Metall Trans.* **1** (1970) 2943 – 2951.
33. J. S. Benjamin, F. W. Strassburg, *Metall. (Berlin)* **36** (1982) 833 – 856.
34. R. C. Benn, L. R. Curwick, G. A. J. Hack, "Scale-up Thermomechanical Processing of MA 6000 E", *Met. Soc. Conf. Proc.,* "High Performance Materials by Powder Metallurgy", 1980.
35. Inco Alloys International: *Mechanically Alloyed Products,* **Publ. No. IAI-9,** Huntington 1988.
36. Inco Alloys International: *Inconel Alloy MA 6000,* **Publ. No. IAI-51,** Huntington 1988.
37. *Annual Book of ASTM Standards,* **vol. 01.02,** "Castings, Nickel and Nickel Alloys," ASTM A494/A494M-87a.
38. ASTM A560-74: Chromium Nickel Alloy Castings. DIN 17 730: Nickel und Nickel-Kupfer, Beuth Verlag, Berlin 1971.
39. International Nickel Inc.: *Engineering Properties of Cast Nickel 210 and Related Casting Alloys,* **Publ. No. A516,** New York 1969.
40. International Nickel Inc.: *Engineering Properties of some Nickel Copper Casting Alloys,* **Publ. No. A515,** New York 1969.
41. International Nickel Inc.: *Engineering Properties of Nickel Chromium Alloy 610 and Related Casting Alloys,* **Publ. No. A517,** New York 1969.
42. International Nickel Ltd.: *IN-657 Cast Nickel – Chromium – Niobium Alloy for Service Against Fuel Ash Corrosion,* **Publ. No. 4320c/234,** London 1975.
43. Inco Ltd., *High Chromium Cr-Ni Alloys to Resist Fuel Ash Corrosion,* **Publ. No. 4299b,** Toronto 1975. (A Review of Developments and Experience 1955 – 1975).
44. G. Rudolph, T. Hoffmann, *Metall (Berlin)* **41** (1987) 988 – 997 u 1210 – 1221.
45. F. W. Strassburg (ed.): *Schweißen nichtrostender Stähle,* DVS-Verlag, Düsseldorf 1982, pp. 44 – 48, 82 – 84.
46. DIN 1726: Schweißzusätze für Nickel und Nickellegierungen, Zusammensetzung, Verwendung und Technische Lieferbedingungen, part 1, 1985.
47. DIN 1736: Schweißzusätze für Nickel und Nickellegierungen, Prüfstück, Proben, Mechanische Gütewerte, part 2, 1985.
48. BS 2901: *Further Rods and Wires for Gas-shielded Arc Welding, Nickel and Nickel Alloys,* part 5, 1983, British Standards Institution, London, 1983.
49. AWS A 5.14: *Specification for Nickel and Nickel Alloy Bare Welding Rods and Electrodes,* Amer. Weld. Soc. Inc., Miami 1983.
50. AWS A 5.11: *Specification for Nickel and Nickel Alloy Covered Welding Electrodes,* Amer. Weld. Soc. Inc., Miami 1983.
51. L. Griffing (ed.): "Metals and their Weldability," *Welding Handbook,* 8th ed., Sect. 4, Chap. 67, Amer. Weld. Soc. Inc., Miami 1988, pp. 67.1 – 67.45.
52. Inco Alloys International: *Joining Huntington Alloys,* **Publ. No. IAI-14,** Huntington 1986.
53. DIN 8513: Hartlote, Nickelbasislote zum Hochtemperaturlöten, Verwendung, Zusammensetzung, Technische Lieferbedingungen, Part 5, 1983.
54. R. Jockschat, D. Pilhan, P. Zeisser, *DVS-Ber.* **75** (1982) 136 – 143.

55. E. K. Luegscheider, K.-D. Partz, R. Lison, *Weld J. (Miami)* **61** (1982) 3295 – 3335.
56. I. Amato, F. Bandrocco, M. Ravizza, *Weld. J. (Miami)* **51** (1972) 3415 – 3458.
57. V. Ruza, *DVS-Ber.* **75** (1982) 306 – 311.
58. A. DePaoli, Dissertation, Universität Saarbrücken 1975.
59. Inco Alloys International: *Heating and Pickling Huntington Alloys,* Publ. No. T-20, Huntington 1968.
60. Henry Wiggin & Co. Ltd.: *Fabricating of Wiggin Nickel Alloys,* Publ. No. 3590, Hereford 1974.

Nickel Compounds

Keith Lascelles, INCO Europe Ltd., Swansea, United Kingdom (Chap. 1, 2, 3, 6 and 7)

Lindsay G. Morgan, INCO Europe Ltd., Swansea, United Kingdom (Chap. 1, 2, 3, 4, 5, 6 and 7)

David Nicholls, University of Liverpool, Liverpool, United Kingdom (Chap. 4 and 5)

Detmar Beyersmann, University of Bremen, Bremen, Germany (Chap. 8)

1. Introduction 1
2. Inorganic Nickel Compounds 1
3. Nickel Tetracarbonyl 5
4. Nickel Salts of Organic Acids 7
5. Organonickel Compounds 8
6. Analysis 9
7. Economic Aspects 9
8. Toxicology 10
8.1. Distribution in the Environment . . . 10
8.2. Mammalian Toxicology 10
9. References 13

1. Introduction

Of the worldwide production of nickel, only about 3% is in the form of compounds if ferronickel, nickel oxide, and partially reduced nickel oxides used in the metallurgical industry are excluded.

The main uses of nickel compounds fall into four areas:

1) *Electroplating*. This includes electroforming and electroless nickel plating as well as conventional nickel electroplating.
2) *Catalysis*. Nickel, its alloys and compounds serve as catalysts for numerous chemical reactions. The main uses of nickel catalysts, prepared from nickel compounds, are in steam reforming, hydrocracking, and in hydrogenation reactions, including methanation.
3) *Batteries*. There are various routes to the nickel hydroxide used as the active mass in nickel – cadmium batteries.
4) *Pigments*. These include colorants in enamels, glassware, ceramics, and plastics.

Like all transition metals, nickel forms numerous complexes. These are of academic rather than industrial importance. Most nickel compounds are green; some are blue.

Much information is available from the standard reference works [1 – 3] and a review [4]. There is an extensive review of the coordination chemistry of nickel in [5].

2. Inorganic Nickel Compounds

Nickel(II) amidosulfate [*13770-89-3*], nickel sulfamate, $Ni(NH_2SO_3)_2 \cdot 4\ H_2O$, M_r 322.94 is a highly soluble salt which is used in solution in the electrolyte for nickel electroforming processes, giving low stress deposits.

Commercially, it is available not as a solid but as a solution containing about 11% nickel. This is usually prepared by dissolving nickel powder in a hot solution of sulfamic acid ($HSO_3 \cdot NH_2$), but soluble nickel oxide, nickel hydroxide, or nickel carbonate can also be used. Short reaction times are necessary because of the hydrolysis of sulfamic acid to sulfuric acid.

Nickel(II) ammonium sulfate [*7785-20-8*], $NiSO_4 \cdot (NH_4)_2SO_4 \cdot 6\ H_2O$, M_r 395.00, can be crystallized as blue-green crystals from a solution containing the two components in the appropriate ratio. It has low solubility in water, ca. 25 g/L at 20°C. The salt was formerly used in electroplating, but now its uses are very limited.

Nickel(II) Carbonate. Although a hydrated carbonate $NiCO_3 \cdot 6\ H_2O$ and the anhydrous form $NiCO_3$ exist, the industrially important compound is the bright-green, rhombic basic nickel carbonate approximating to $2\ NiCO_3 \cdot 3\ Ni(OH)_2 \cdot 4\ H_2O$ [*29863-10-3*], M_r 587.57. It is virtually insoluble in water, but dissolves in acids with evolution of carbon dioxide. Heating

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_235.pub2

Table 1. Nickel halides

Compound	CAS registry no.	M_r	d	$-\Delta H^\circ$ (25°C), kJ/mol	Color
NiF_2	[*10028-18-9*]	96.69	4.63	651	light-green
$NiCl_2$	[*7718-54-9*]	129.60	3.55	305	pale- yellow
$NiBr_2$	[*13462-88-9*]	218.50	5.10	212	yellow
NiI_2	[*13462-90-3*]	312.50	5.83	78	black
$NiF_2 \cdot 4\ H_2O$	[*13940-83-5*]	168.75	2.22		light-green
$NiCl_2 \cdot 6\ H_2O$	[*7791-20-0*]	237.69	1.92	2103	green
$NiBr_2 \cdot 3\ H_2O$	[*7789-49-3*]	272.55		1146	yellow-green
$NiI_2 \cdot 6\ H_2O$	[*7790-34-3*]	420.60			blue-green

to above ca. 450°C in air gives a reactive nickel oxide with a high specific surface area.

Commercial basic nickel carbonate is made by precipitation from a nickel solution, usually the sulfate, with sodium carbonate. The exact composition depends on the temperature and the concentrations of the components in solution.

Basic nickel carbonate is used for pH adjustment in nickel electroplating baths, in catalyst production, and in the manufacture of some nickel pigments and specialty nickel compounds.

Demand for basic nickel carbonate has recently increased due to the use of zinc – nickel plating of steel by the automotive industry. In one electroplating method, the nickel in solution is replenished by basic nickel carbonate.

It is also possible to prepare nickel carbonate by dissolving nickel in ammonium carbonate solution in the presence of air and carbon dioxide. Boiling off the ammonia and carbon dioxide precipitates basic nickel carbonate. This method forms part of the Caron process for the refining of nickel (→ Nickel). The carbonate is further processed to nickel oxide or nickel metal.

Nickel(II) Halides. Nickel forms dihalides with all four halogens, and each dihalide forms one or more hydrates. Only the chlorides are of industrial importance, particularly the hexahydrate $NiCl_2 \cdot 6\ H_2O$, which is available either as a solid or as a concentrated solution.

Table 1 lists some properties of the dihalides and the hydrate usually encountered. Known hydrates are $NiF_2 \cdot 4\ H_2O$, $NiCl_2 \cdot x\ H_2O$, (x = 2, 4, 6, 7), $NiBr_2 \cdot x\ H_2O$ (x = 2, 3, 6, 9), $NiI_2 \cdot x\ H_2O$ (x = 4, 6).

Nickel dichloride hydrates are stable over the following temperature ranges:

$NiCl_2 \cdot 7\ H_2O$	low temperature
$NiCl_2 \cdot 6\ H_2O$	up to 36.25°C
$NiCl_2 \cdot 4\ H_2O$	36.25 to 75°C
$NiCl_2 \cdot 2\ H_2O$	above 75°C

The difluoride is only slightly soluble in water, but the other halides are very soluble (e.g., 117 g $NiCl_2 \cdot 6\ H_2O$ per 100 mL solution at 20°C).

The difluoride has a tetragonal structure, and the other dihalides have hexagonal layer structures. They are relatively stable, and can be melted or sublimed in a vacuum or under an inert atmosphere.

Production. The anhydrous halides can be made by direct reaction of the elements at high temperature or in nonaqueous solution, although the fluoride is better made indirectly. They are also formed by dehydration of the hydrates in a stream of the hydrogen halide gas to prevent formation of nickel oxide.

The hydrates can be made by dissolving nickel metal, oxide, hydroxide, or carbonate in an aqueous solution of the halogen acid HX, followed by crystallization. Commercially, nickel dichloride hexahydrate is usually produced by dissolving nickel metal in hydrochloric acid, and then crystallizing. Nickel dichloride solutions are important in chloride-based processes for the extraction of nickel, but nickel metal is usually recovered by electrowinning.

Uses. Nickel dichloride hexahydrate is important in electroplating; it is used together with nickel sulfate in the Watts plating bath. Table 2 lists the purity specifications for its use in electroplating. It is also used in the production of some catalysts. Uses of the other dihalides are very minor.

Table 2. Specifications of nickel chloride hexahydrate for electroplating

	BS 564 (1970), %	DIN 50970, %
Ni	24.0 (Ni + Co) min.	23.7 min.
Co	1% of total metal, max.	0.5 max.
Cu	0.005 max.	0.002 max.
Fe	0.01 max.	0.01 max.
Pb	0.002 max.	0.002 max.
Zn	0.002 max.	0.006 max.
Insolubles	0.05 max.	0.05 max.
pH (40% w/v)	2.0 min.	

Nickel(II) hydroxide [*12054-48-7*], $Ni(OH)_2$, M_r 92.71, d 4.15, $\Delta H°$ (25°C) −514 kJ/mol. Treating a nickel salt solution with an alkali-metal hydroxide solution gives a green gelatinous precipitate of nickel hydroxide, also known as nickel hydrate. Filtering, washing, and drying give the pure compound. The precipitation conditions affect the physical characteristics of the nickel hydroxide produced.

Nickel hydroxide is virtually insoluble in water, but it is soluble in dilute acids, giving a nickel salt solution, and in ammoniacal solutions a blue ammine complex is formed. Heating to ca. 250°C drives off water, leaving nickel oxide.

Nickel hydroxide has some use as an intermediate in catalyst production, but its main use is as the active mass in the positive electrode in nickel – cadmium batteries. In pocket-plate electrodes, pellets of nickel hydroxide are used. In sintered electrodes, the nickel hydroxide is prepared within the electrode either by precipitation or electrochemically.

There are various hydrated oxides or hydroxides containing Ni(III) and possibly Ni(IV), some of which are not well characterized. The compounds are black and are referred to as nickel peroxide, nickelic hydroxide, or nickel black.

The β-NiO·OH, obtained by the oxidation of nickel salts in alkaline solution with an oxidizing agent such as sodium hypochlorite, is well characterized. It is also formed in batteries when Ni(II) hydroxide in a nickel electrode is charged.

Persulfate oxidation of nickel(II) hydroxide gives a substance formulated as $NiO_2 \cdot x\ H_2O$.

A commercial "nickel peroxide" corresponded to $Ni_2O_3 \cdot 2\ H_2O$, and gave $Ni_2O_3 \cdot H_2O$ (equivalent to NiO·OH) on partial dehydration [6].

Nickel(II) Nitrate. The commercially available compound is the green hexahydrate [*13478-00-7*], $Ni(NO_3)_2 \cdot 6\ H_2O$, M_r 290.8, d 2.05, $\Delta H°$ (25°C) −2212 kJ/mol. It is very soluble in water. On heating, the hexahydrate dissolves in its own water of crystallization at 56°C. Progressive heating drives off the water of crystallization, giving basic nitrates and finally nickel(II) oxide.

The stability ranges of the hydrates are:

$Ni(NO_3)_2 \cdot 9\ H_2O$	below −3°C
$Ni(NO_3)_2 \cdot 6\ H_2O$	−3 to 54°C
$Ni(NO_3)_2 \cdot 4\ H_2O$	54 to 85.4°C
$Ni(NO_3)_2 \cdot 2\ H_2O$	85.4 to ca. 105°C

Nickel nitrate solution can be produced by dissolving nickel in nitric acid. The reaction is vigorous, and without careful control loss or breakdown of the nitric acid occurs, forming, among others, ammonium nitrate. The dissolution of nickel(II) oxide in nitric acid is easier to control; the hexahydrate is produced by crystallization of the resulting solution.

It is difficult to prepare the pale-green, anhydrous nickel(II) nitrate by dehydration of the hexahydrate. It can be obtained by dehydration with dinitrogen pentoxide in fuming nitric acid, or from nickel tetracarbonyl and dinitrogen tetroxide.

Nickel(II) nitrate hexahydrate is widely used in the production of nickel catalysts, by impregnating an inert carrier and decomposing the nitrate. It is also used in the formation of the nickel(II) hydroxide active mass in sintered nickel electrodes for nickel – cadmium batteries.

Nickel(II) Oxide. Nickel oxide [*1313-99-1*], NiO, M_r 74.69, d 6.8, $\Delta H°$ (25°C) −240 kJ/mol, is green, with a cubic rock-salt structure. It occurs in nature as bunsenite. The exact color depends on factors such as purity and crystallite size. In general, the properties of nickel oxide depend on the method of preparation, especially temperature.

Nickel(II) oxide has a defect structure $Ni_{1-y}O$, which is responsible for its p-type semiconductor properties. In addition, nickel oxide can absorb surface oxygen to give higher O : Ni ratios. There is little evidence for Ni_2O_3

as a bulk compound, although it may exist as a surface compound on NiO.

World production of "nickel oxide" is ca. 40 000 t/a from INCO in Canada, Greenvale in Australia, and from Cuba, but only about 4000 t of this is chemical nickel oxide. The majority is metallurgical nickel oxide, or partially reduced nickel oxide containing 85 – 90% nickel, used mainly in the stainless steel industry. Chemical nickel oxide is available in two forms: green or black.

Production. Green nickel oxide can be made by calcining almost any nickel compound in air, for example, the nitrate, oxalate or hydroxide, at ca. 1000°C. It can also be made by oxidation of carbonyl nickel powder, but some care is needed to ensure complete oxidation.

Black, or soluble, nickel oxide is made by calcining basic nickel carbonate at ca. 550°C; it contains a slight excess of oxygen.

The green oxide is rather refractory, and dissolves only slowly in mineral acids. The black oxide is more reactive and dissolves readily. Both forms are insoluble in water and are easily reduced, for example by hydrogen, to nickel metal.

Uses. Nickel oxide is used in the production of catalysts, in the enamel industry for making frit, in the glass and ceramics industries, and in the manufacture of ferrites for the electronics industry.

Black nickel oxide is particularly suitable for dissolving in acids as a first step in the production of specialty nickel salts.

Higher oxides such as Ni_2O_3, and NiO_2 have been proposed but there is little evidence for their existence free of water.

Nickel(II) Sulfate. Two forms of nickel(II) sulfate are commercially available. The most common is the blue-green, tetragonal hexahydrate [*10101-97-0*], α-$NiSO_4 \cdot 6\ H_2O$, M_r 262.84, d 2.07, ΔH° (25°C) −2683 kJ/mol, nickel content 22.3%, which crystallizes from solution between 30.7 and 53.8°C. The other form is the green, orthorhombic heptahydrate [*10101-98-1*] $NiSO_4 \cdot 7\ H_2O$, M_r 280.85, d 1.95, nickel content 20.9%, which crystallizes from solution below 30.7°C. Crystallization above 53.8°C gives green β-$NiSO_4 \cdot 6\ H_2O$.

The hydrates are readily soluble in water. The solubilities in grams per liter of solution are (the temperature in °C is given in parentheses):

$NiSO_4 \cdot 7\ H_2O$	578	(10)
$NiSO_4 \cdot 7\ H_2O$	775	(30)
α-$NiSO_4 \cdot 6\ H_2O$	770	(40)
α-$NiSO_4 \cdot 6\ H_2O$	851	(50)
β-$NiSO_4 \cdot 6\ H_2O$	959	(60)
β-$NiSO_4 \cdot 6\ H_2O$	1258	(100)

Data also exist for the solubility of nickel sulfate in sulfuric acid [7].

Heating the hydrates causes dehydration. Hydrates with 4 or 1 molecule of water are stable. The bright yellow, cubic, anhydrous salt $NiSO_4$ [*7786-81-4*], M_r 154.75, d 3.68, ΔH° (25°C) −873 kJ/mol, is obtained above 330°C. This decomposes above ca. 800°C to give nickel oxide and sulfur trioxide. The lower hydrates can also be obtained by crystallization from nickel sulfate solutions containing excess sulfuric acid. They dissolve only slowly in water. Basic sulfates also exist.

Production. The worldwide production capacity is about 40 000 t/a, from producers in Europe, Japan, China, and the former Soviet Union. Most of the nickel sulfate is a byproduct of electrolytic copper refining. A bleed stream is purified, finally giving a solution from which nickel sulfate is crystallized.

Table 3. Specifications of nickel sulfate hexa- or heptahydrate for electroplating

	BS 564 (1970) (6 or 7 hydrate), %	DIN 50970 (6 hydrate), %
Ni + Co	20.7 min.	22.0 min.
Co	1% of total metal, max.	0.5 max.
Cu	0.002 max.	0.002 max.
Fe	0.005 max.	0.005 max.
Pb	0.002 max.	0.002 max.
Zn	0.002 max.	0.006 max.
Cd		0.005 max.
As		0.001 max.
Insolubles	0.05 max.	0.05 max.
pH (40% w/v)	2 min.	

Nickel sulfate is also produced by dissolving nickel or nickel oxide in sulfuric acid and

crystallizing. Several nickel refineries produce nickel sulfate solution by leaching nickel mattes with sulfuric acid, but the nickel is usually recovered as nickel metal by electrowinning.

Uses. The main use of nickel sulfate is as the electrolyte in nickel electroplating baths. Table 3 lists the standards covering nickel sulfate for use in electroplating. It is also used in electroless nickel plating, in catalyst manufacture, and in the production of other nickel compounds.

Nickel Sulfides. Nickel(II) sulfide [*1314-04-1*], NiS, M_r 90.75, *d* 5.65, $\Delta H°$ (25°C) −82 kJ/mol, *mp* 797°C.

Neutral nickel(II) salt solutions react with ammonium sulfide initially to give black α-NiS, a hydrous, amorphous precipitate of the type Ni $(SH,OH)_2$. At low pH, this crystallizes to give rhombohedral γ-NiS, whereas in more alkaline solution hexagonal β-NiS is formed.

The γ-form occurs in nature as millerite. This shows little deviation from stoichiometry, but a high-temperature form has a stoichiometry range.

Other nickel sulfides exist, some of which occur in nature. These are Ni_3S_2, Ni_7S_6, Ni_3S_4, and NiS_2. All can be prepared by fusing the correct proportions of the elements.

Sulfide minerals are a major source of nickel, and nickel sulfide is an important intermediate in refining processes. The main use of nickel sulfides is as catalysts, for example, in the hydrogenation of sulfur compounds. The catalyst can be prepared in situ by sulfiding a supported nickel catalyst.

Nickel(II) tetrafluoroborate [*14708-14-6*], $Ni(BF_4)_2 \cdot 6\,H_2O$, M_r 340.41 can be made as apple green needles by reacting fluoroboric acid with basic nickel carbonate, and crystallizing from solution. It has limited use in specialty high-speed nickel plating process, and other minor uses.

Nickel Complexes. In common with all transition metals nickel forms a vast number of inorganic complexes with different oxidation states and geometries. Only the ammine complexes, which are important in ammonia-based leaching processes, have industrial significance.

The structural, spectral, and magnetic properties of nickel complexes are important in studies of transition-metal complex chemistry. In particular, there are concentration and temperature dependent anomalies in properties due to polymerization and equilibria between different geometries.

The most common complex ion is octahedral Ni(II), for example, the green $Ni(H_2O)_6^{2+}$ in aqueous solution. The electronic structure of these complexes is [Ar]$3d^8$ with two unpaired electrons, and hence they are paramagnetic. Other species can displace water molecules, for example, ammonia molecules in ammine complexes $[Ni(H_2O)_x(NH_3)_{6-x}]^{2+}$, $x = 0 - 6$, which are blue or purple. The complex ion $[Ni(H_2O)_4\,(NH_3)_2]^{2+}$ is the required ion in solution for the production of nickel powder by pressure reduction of solutions by hydrogen.

Four-coordinate Ni(II) complexes are quite common, either diamagnetic square planer complexes such as the yellow $Ni(CN)_4^{2-}$ or paramagnetic tetrahedral complexes such as $NiCl_4^{2-}$. Five-coordinate Ni(II) complexes also exist, and are either trigonal bipyramidal or square pyramidal.

Complexes of nickel in lower oxidation states are mostly based on carbonyl ligands. An example of Ni(0) is $Ni(PF_3)_4$ and of Ni(I), $Ni(PPh_3)_3X$ (X = Cl, Br, I).

Higher oxidation states are stabilized by electronegative elements, e.g., Ni(III) in the violet K_3NiF_6 and Ni(IV) in red K_2NiF_6, both of which are octahedral.

3. Nickel Tetracarbonyl

Nickel tetracarbonyl [*13463-39-3*], $Ni(CO)_4$, M_r 170.75, *bp* 42.2°C, *mp* −19.3°C, *d* 1.31, is a very toxic colorless liquid. It possesses a significant vapor pressure at ambient temperature (44 kPa at 20°C; 65 kPa at 30°C).

Nickel tetracarbonyl is virtually insoluble in water, but soluble in many organic solvents. It does not react with dilute mineral acids. It is thermally unstable, decomposing to nickel and carbon monoxide. It burns in air with a luminous flame, giving nickel oxide and carbon dioxide, and forms explosive mixtures with air (3 – 34 vol% $Ni(CO)_4$). The molecule is tetrahedral, with linear Ni−C−O bonds. The bonding consists of both a Ni−C σ-bond and π bonding.

Production. Nickel tetracarbonyl is formed by direct reaction of carbon monoxide and finely divided nickel at relatively low temperatures [8]:

$$Ni_{(g)} + 4\,CO_{(g)} \rightarrow Ni(CO)_{4(g)}$$

$$\Delta G = -160410 + 418.27\,T\ \mathrm{J/mol}$$

At atmospheric pressure, the maximum rate of formation of nickel tetracarbonyl is at 130°C for pure nickel. The temperature at which the rate of formation is a maximum decreases in the presence of a catalyst such as sulfur, and increases with pressure. The reverse reaction begins above ca. 180°C. This reversible reaction is the basis of the atmospheric-pressure Mond process and the INCO pressure process for the production of high-purity nickel pellet and powder (→ Nickel, Chap. 6.3). Nickel tetracarbonyl can also be prepared in solution by a variety of methods [9]. Users of nickel tetracarbonyl frequently produce their own supply, but it is commercially available in the United States.

The conditions under which nickel tetracarbonyl is formed are important because of the possibility of corrosion or transfer of nickel within a system, and in view of its high toxicity. The mere coexistence of carbon monoxide and nickel in some form does not mean that nickel tetracarbonyl will form. Other criteria must be met: (1) a fully reduced nickel-containing surface, (2) a reducing gas containing carbon monoxide, (3) the formation must be thermodynamically possible, which generally means low temperature (ambient to 150°C) and high carbon monoxide partial pressures. In addition, the presence of a catalyst such as sulfur accelerates nickel tetracarbonyl formation.

One situation where a significant amount of nickel tetracarbonyl can form is from a finely divided reduced nickel catalyst and carbon monoxide at low temperature. This is well known to users of nickel catalysts, and such conditions are avoided. Other than this, it is rare that significant amounts of nickel tetracarbonyl are formed. Environmentally, precautions preventing contamination of the workplace by the carbon monoxide will also prevent contamination by any nickel tetracarbonyl.

The presence of nickel tetracarbonyl has been suggested but not demonstrated in cigarette smoke and in gases from the combustion of fossil fuels containing nickel. Attempts to detect it in welding fume failed [10].

Uses. Apart from being an intermediate in the carbonyl refining of nickel, nickel tetracarbonyl can be thermally decomposed to nickel plate other materials, for example, in mold production or in a fluidized bed. It is also used as a carbonylating agent or catalyst in organic chemistry.

Analysis. Nickel tetracarbonyl can be analyzed by decomposition and conventional analysis of the nickel, by gas chromatography, UV or IR spectroscopy. There is a highly sensitive method based on the chemiluminescent reaction of nickel tetracarbonyl with ozone [11]. Commercial instruments based on infrared or chemiluminescent analysis are available.

Reactions of Nickel Tetracarbonyl. Nickel tetracarbonyl undergoes oxidation, reduction, and substitution reactions [12, 13]. These are normally carried out in organic solvents below ca. 50°C to prevent thermal decomposition of the nickel tetracarbonyl.

Reaction with various oxidizing agents gives Ni(II) compounds. Concentrated nitric acid gives nickel nitrate. Solutions of nickel tetracarbonyl in organic solvents are oxidized by air to basic nickel carbonate and by halogens to the corresponding nickel dihalide. Decomposition of nickel tetracarbonyl with bromine water is useful as a means of disposal or for analysis.

Reduction reactions, generally with alkali metals, give polynuclear anions formulated as $[Ni_2(CO)_6]^{2-}$, $[Ni_3(CO)_8]^{2-}$, $[Ni_4(CO)_9]^{2-}$, $[Ni_5(CO)_9]^{2-}$, and $[Ni_6(CO)_{12}]^{2-}$. Reduction of nickel tetracarbonyl by alkali metals in liquid ammonia gives a carbonyl hydride $[NiH(CO)_3]_2$, isolated as a tetra-ammoniate.

Interest in substitution compounds of nickel tetracarbonyl blossomed following the publication in 1948 of work by Reppe and coworkers showing that an effective class of catalysts for the trimerization of acetylene compounds could be formed by substituting CO groups in $Ni(CO)_4$ by donor ligands such as triphenylphosphine.

Thousands of substitution compounds of nickel tetracarbonyl have now been prepared. Most are with ligands containing the group 15 elements phosphorus, arsenic, or antimony as

electron donor, but carbon, nitrogen, and unsaturated organic molecules can also serve as ligands.

Some of the simpler substitution compounds with phosphorus ligands are $Ni(CO)_n(PX_3)_{4-n}$, $n = 0 - 3$, X = H, F, Cl, CH_3, C_2H_5, C_6H_5 (substituted phosphines) and X = OCH_3, OC_2H_5, OC_6H_5 (phosphites). The degree of substitution is controlled by steric and electronic effects. For example, with PF_3 and $P(C_6H_5)_3$, only mono- and disubstituted compounds are formed, whereas with PCl_3 and $P(OC_6H_5)_3$ the carbon monoxide molecules can be completely replaced. The tetrakis(ligand) compounds $Ni(PF_3)_4$ and $Ni[P(C_6H_5)_3]_4$ can be prepared by other means. Substitution by chelating ligands is also possible, e.g., $(CO)_2NiL'$ and NiL'_2, where $L' = o\text{-}C_6H_4\text{-}[P(C_2H_5)_2]_2$.

Fewer substitution compounds based on arsenic and antimony have been prepared. Examples include $Ni(CO)_3AsX_3$ (X = CH_3, C_2H_5, C_6H_5, OCH_3, OC_2H_5, OC_6H_5) and $Ni(CO)_3SbX_3$ (X = Cl, C_2H_5, C_6H_5, OC_6H_5).

4. Nickel Salts of Organic Acids

Nickel(II) acetate tetrahydrate [*6018-89-9*], $Ni(CH_3COO)_2 \cdot 4\,H_2O$, M_r 248.86, *d* 1.744, can be obtained as green crystals when solutions of nickel(II) hydroxide or carbonate in acetic acid are evaporated at room temperature. It is readily soluble in water but is insoluble in ethanol.

Anhydrous nickel(II) acetate $Ni(CH_3COO)_2$, *d* 1.798, can be prepared by dehydration of the tetrahydrate in vacuum or by heating the tetrahydrate under reflux with acetic anhydride. When heated in the absence of air these acetates decompose to give a residue of nickel and Ni_3C, evolving, among other products, carbon monoxide and carbon dioxide. Nickel acetate is a suspect carcinogen: the LD_{50} (rat) is 350 mg/kg [14].

Nickel(II) acetylacetonate [*3264-82-2*], $Ni(CH_3COCHCOCH_3)_2$, M_r 256.91, *mp* 230°C (dec.), is obtained as emerald green crystals by dehydration of the dihydrate at 50°C in vacuum; it has an unusual trimeric structure in the solid state [15]. The dihydrate $Ni(CH_3COCH_3)_2 \cdot 2\,H_2O$ is prepared by the addition of acetylacetone to solutions of nickel(II) salts in the presence of a weak base such as sodium acetate. Nickel(II) acetylacetonate is soluble in organic solvents and finds use in the synthesis of organometallic compounds such as nickelocene and bis(cyclooctadiene) nickel. It is also industrially important as a catalyst component in the oligomerization of alkenes and in the conversion of acetylene to cyclooctatetraene [16, 17].

Nickel(II) dimethylglyoximate [*13478-93-8*] bis(dimethylglyoximato) nickel(II), $Ni(C_4H_7N_2\,O_2)_2$, M_r 288.94, is obtained quantitatively as a bright-red precipitate when an alcoholic solution of dimethylglyoxime is added to a neutral or slightly alkaline solution of a nickel(II) salt. This reaction is used in the gravimetric estimation of nickel [18] and is also useful as a spot test for the presence of nickel ions. The structure of this diamagnetic solid [18] shows the nickel(II) ion to be in a square-planar environment; the dotted lines represent strong intramolecular hydrogen bonds.

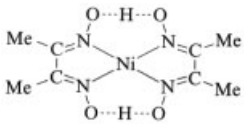

Nickel(II) formate dihydrate [*15694-70-9*], $Ni(HCOO)_2 \cdot 2\,H_2O$, M_r 184.76, *d* 2.154, is obtained as green crystals by evaporation of a solution of nickel carbonate in formic acid. The anhydrous salt is obtained when the dihydrate is heated to 140°C; at 250°C, finely divided nickel metal is produced. Hence, nickel(II) formate is used in the production of nickel catalysts.

Nickel(II) oxalate dihydrate [*6018-94-6*], $Ni(COO)_2 \cdot 2\,H_2O$, M_r 182.76, precipitates as a green powder when oxalic acid or alkali-metal oxalate solutions are added to nickel(II) salt solutions. When heated in vacuum, it evolves water at 150°C and decomposes completely at 300°C, giving finely divided nickel suitable for use in catalysis or in the production of nickel carbonyl.

5. Organonickel Compounds

Nickel Alkyls and Aryls. Simple σ-bonded alkyls and aryls of nickel are too unstable to exist in the absence of stabilizing ligands. Alkyls can be stabilized, however, by π-bonded aromatic systems, by phosphite ligands, and by bipyridyl. Thus the dimethylnickel bipyridyl adduct exists as green crystals that can be recrystallized from boiling benzene.

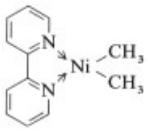

Stable aryls can be prepared by reaction of aryl Grignard reagents ArMgX with the bis(phosphine) nickel(II) halide complexes $NiX_2(PR_3)_2$. The yellow or brown products of the types $ArNiX(PR_3)_2$ and $Ar_2Ni(PR_3)_2$ are particularly stable if the aryl group is ortho-substituted (e.g., Ar = *o*-tolyl or mesityl), and such compounds are air stable and can be heated to 80°C without decomposition. Many such compounds have been prepared [19].

π-Bonded Complexes. There is a very extensive chemistry of nickel π-bonded to unsaturated organic molecules. The most important classes of compounds are the η^2-alkene complexes, the η^3-allyls, and the η^5-cyclopentadienyl complexes. Alkene complexes of nickel are of two basic types, both of which contain nickel in zero oxidation state. There are the trigonal, 16-electron complexes and the less common tetrahedral 18-electron complexes. The two most important alkene complexes are described below together with the parent allyl and cyclopentadienyl complexes; these are the starting materials for the preparation of many other organometallic complexes [20].

(Cyclododeca-1,5,9-triene)nickel [*39330-67-1*], $NiC_{12}H_{18}$, M_r 220.96, is prepared by the reduction of nickel(II) acetylacetonate with diethyl-aluminum ethoxide in the presence of cyclododeca-1,5,9-triene (CDT). It forms red, air-sensitive crystals, which can be sublimed in vacuum and are soluble in most organic solvents. The complex normally isolated has all-*trans*-CDT with the nickel atom surrounded trigonally and in a propeller-like fashion by the double bonds.

The two important reactions shown by Ni (CDT) are ligand addition, for example:

$$Ni\,(CDT) + PMe_3 \rightarrow Ni\,(CDT)\,(PMe_3)$$

and ligand substitution reactions by which other tris(alkene)nickel complexes are prepared, for example:

$$Ni\,(CDT) + 3\,CH_2 = CH_2 \rightarrow Ni(CH_2 = CH_2)_3 + CDT$$

Bis(cycloocta-1,5-diene)nickel [*33221-58-8*], $Ni(C_8H_{12})_2$, M_r 275.05, is prepared by reduction of nickel(II) acetylacetonate with triethylaluminum in the presence of cycloocta-1,5-diene (COD):

$$Ni(acac)_2 + 2\,COD + 2\,AlEt_3 \rightarrow$$
$$Ni(COD)_2 + 2\,Al\,(acac)\,Et_2 + C_2H_4 + C_2H_6$$

It can also be obtained by electrolytic reduction of nickel salts in pyridine in the presence of COD, by reduction of nickelocene with lithium in the presence of COD, and by cocondensation of nickel vapor and COD. It crystallizes as orange-yellow crystals from aromatic solvents and can be handled for a short time in air without appreciable decomposition. The structure of $Ni(COD)_2$ has nickel atoms tetrahedrally surrounded by four alkenic double bonds.

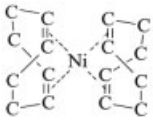

In its reactions with ligands, substitution of either one or, more frequently, two COD molecules occurs, and reaction with excess ligand is a widely used method for preparing tetrakis(ligand) complexes for example:

$$Ni(COD)_2 + 4\,PH_3 \rightarrow Ni(PH_3)_4 + 2\,COD$$

Bis(cycloocta-1,5-diene)nickel has found extensive use not only as a starting point for the preparation of many organonickel complexes but also as a catalyst or cocatalyst in a variety of organic reactions such as the polymerization and cyclotrimerization of butadiene, the oligomerization of other alkenes, the isomerization of strained cycloalkanes, and the polymerization of ethene [16]. The polymerization of ethene with a $Ni(COD)_2$ catalyst in the presence of acids such as thiolactic acid occurs at 65°C and 4 MPa and does not require an aluminum component in the catalyst [21].

Bis(η^3-allyl)nickel [*12077-85-9*], Ni $(C_3H_5)_2$, M_r 140.83, is prepared by reaction of allylmagnesium bromide with nickel(II) bromide in ether at −10°C.

It is soluble in ether and crystallizes below 1°C, forming yellow needles. Ethereal solutions are unaffected by deoxygenated water but the solid is pyrophoric in air. It is the simplest example of a whole range of nickel η^3-allyls. The sandwich structure has been established by X-ray analysis on the more stable bis(η^3-methyl allyl)nickel complex [22].

Bis(η^3-allyl)nickel is a very active catalyst for the room temperature cyclotrimerization of butadiene to cyclododeca-1,5,9-triene:

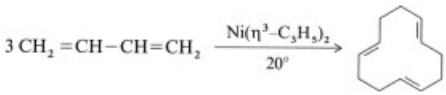

η^3-Allyls of nickel find many uses in catalysis, especially in the oligomerization and isomerization of alkenes [2].

Bis(η^5-cyclopentadienyl)nickel(II) [*1271-28-9*], nickelocene, $Ni(C_5H_5)_2$, M_r 188.88, *d* 1.47, forms dark emerald-green crystals that can be sublimed in vacuum above 80°C and melt in a nitrogen atmosphere with some decomposition at 173 – 174°C. The crystals are only mildly air sensitive, but solutions in organic solvents must be handled under an inert atmosphere. Nickelocene is neither attacked by water nor dissolves in it. The solid is paramagnetic (μ_{eff} = 2.86 BM) and has the sandwich structure as in ferrocene with the nickel atom between planar cyclopentadienide rings.

The simplest laboratory preparation of nickelocene involves the addition of a solution of hydrated nickel(II) chloride in dimethylsulfoxide to a solution of potassium hydroxide and cyclopentadiene in 1,2-dimethoxyethane:

$$NiCl_2 \cdot 6\,H_2O + 2\,C_5H_6 + 2\,KOH \rightarrow Ni(C_5H_5)_2 + 2\,KCl + 8\,H_2O$$

Many other methods are available, however, starting from anhydrous nickel salts with Grignard reagents (or alkali-metal cyclopentadienides) or from nickel metal and cyclopentadiene [23].

Upon oxidation, the orange-yellow nickelocinium ion $(\eta^5\text{-}C_5H_5)_2Ni^+$ is produced; it can be isolated as the tetraphenylborate. The reactions of nickelocene with donor molecules provide a major route to NiL_4 molecules by displacement of both rings. In some reactions only one cyclopentadiene ring is displaced as in, for example, the reaction with nitrogen monoxide, in which the relatively inert $(\eta^5\text{-}C_5H_5)NiNO$ is produced.

Nickelocene has been shown to be an effective catalyst for the oligomerization of alkenes [24], for the isomerization of alkenes [25], for the polymerization of butadiene [26], for the hydrosilylation of alkenes [27], and for the telomerization of butadiene [27].

Nickelocene is carcinogenic with an LD_{50} (rat) of 490 mg/kg [14].

6. Analysis

The classical analytical method for nickel in solution is gravimetric, by precipitation of the red complex with butanedione dioxime (dimethylglyoxime). Other methods include electrodeposition, EDTA titration, and the standard spectroscopic methods of atomic absorption, flame emission, inductively coupled plasma optical emission, and X-ray fluorescence. There is a standard analytical method for nickel in biological tissue [28].

7. Economic Aspects

In general, nickel compounds cost the London Metal Exchange (LME) price for the contained nickel, plus a premium based on the production cost. However, companies producing

nickel compounds, particularly nickel sulfate, as a byproduct of other processes are more flexible in their pricing.

The LME three-month nickel price decreased during the first part of the 1980s from about $ 3 per pound in 1980 to below $ 2 per pound in 1986. The price then rose to over $ 5.50 per pound in 1988, but has since decreased.

8. Toxicology

8.1. Distribution in the Environment

Air. The nickel content of ambient air in general is very low and ranges from below 1 ng/m^3 in the Antarctic [29] up to 5 ng/m^3 in urban air [30]. Higher concentrations are found near fuel incineration sites and in nickel producing or processing industries.

Soil. The topsoil in woodlands contains between 4 and 40 mg Ni/kg dry matter (d.m.) in sandy and clay soils, respectively [30]. The nickel content of agricultural soils varies greatly depending on the mineral matrix and the type of fertilizer administered. Nickel concentrations in German agricultural soils varied between 4 and 400 mg/kg d.m. Use of sewage sludge, which in Germany on average contains 23 mg Ni/kg d.m., further enhances the amount of nickel in soils. Nickel in soils is mainly complexed to inorganic (clay) and organic (humic matter) material and is only soluble in acid environment.

Water. Nickel is mobilized into surface waters mainly by weathering of rocks. River waters in Germany contain nickel predominantly bound to particulate matter (47 – 63 mg/kg) and to a lesser extent in the filterable fraction (15 – 20 µg/L) [30]. The total nickel concentration in coastal waters of the North Sea was found to be 1 – 11 µg/L and that of the Atlantic Ocean only 0.1 – 0.5 µg/L [29]. Sediments may accumulate higher concentrations, for example, sludge in the port of Hamburg contains 50 – 100 mg/kg d.m.

Foodstuffs. The nickel concentration in *drinking water* usually is very low. In Germany, drinking water concentrations are below 5 µg Ni/L [31]. Drinking water supplies on the average 20 µg Ni/d to human nickel intake, of which only 2% is resorbed (see below).

Meat from game and pasture animals contains only traces of nickel close to the detection limit. Beef contained only 1 – 13 µg/kg d.m. with the exception of meat from cattle grazing close to nickel processing facilities [31].

The major sources of nickel supply to humans are *vegetables*, which usually contain 0.05 – 0.34 mg/kg d.m. Higher nickel concentrations are found in cocoa (8 – 12 mg/kg), hazel nuts (0.7 – 2.3 mg/kg), oats (0.3 – 4.8 mg/kg), and soy products (1.1 – 7.8 mg/kg) [32]. In industrialized countries, additional nickel (about 10 mg/d) is taken up from meals cooked in stainless steel pans, which release varying amounts depending on the acidity of the food [33]. The total supply of nickel by nutrition in Germany with mixed diet was estimated to be 90 – 97 µg/d, and with vegetarian diet 185 – 196 µg/day [34].

8.2. Mammalian Toxicology

Nickel has been shown to be an essential trace element in the nutrition of various animal species [35], but in humans no nickel-deficiency symptoms are known. Toxic effects of major concern are sensitization of the skin (contact dermatitis) and carcinogenicity of inhaled nickel compounds.

Resorption, Distribution, and Elimination. The main path of nickel uptake in the general population is by nutrition. Low in quantity but critical for induction of allergic dermatitis is uptake by skin contact with nickel-plated objects. In the nickel producing and processing industry workers are exposed predominantly to breathable nickel aerosols and dusts. Extraordinarily high uptake of nickel through biomembranes occurs by exposition to nickel tetracarbonyl because of its lipophilicity [36].

Oral Resorption. From the estimated average daily human intake of 150 – 250 µg with nutrition, only 1 – 2% is resorbed in the intestine [36].

Inhalative Resorption. Inhalation of nickel compounds in occupational exposures results in nickel accumulation in the lung by about 50-fold for soluble and by up to 500-fold for oxides and

sulfides of low solubility [37]. Even sparingly soluble nickel compounds are bioavailable after deposition in the lung. Inhalation of all types of nickel metal and compounds has resulted in increased concentrations of nickel in blood and urine. It is estimated that a concentration of 90 µg Ni/m^3 results in a concentration of 5 µg Ni/L in blood and about 15 µg Ni/L in urine [38].

Skin Resorption. Cutaneous resorption of soluble inorganic nickel compounds is slow, with the exception of the lipophilic nickel tetracarbonyl, which is easily resorbed and exerts high acute toxicity after penetration through the skin [36].

Distribution. The average nickel content of the human body is about 7 µg Ni/kg body mass. The average concentrations in human body fluids are 0.2 – 0.3 µg Ni/L in serum, 0.34 – 2.3 µg Ni/L in whole blood, about 2 – 3 µg Ni/L in urine, and about 14 mg Ni/kg d.m in feces [36]. In autopsies, human organs contained between 20 and 180 µg Ni/kg d.m. with the highest concentrations in the lung and thyroid and adrenal glands [39].

Elimination. Excretion after oral uptake of soluble nickel salts occurred with a half-life of 10 – 20 h [40]. After accidental oral intake of high doses of soluble nickel a half-life of 60 h was observed [29], and after occupational inhalation by stainless steel welders a half-life of 53 h was estimated [42]. Following inhalative exposure of rats to sparingly soluble nickel compounds, a mean half-life of 90 d was observed [39].

Acute Toxicity. Acute toxic effects of nickel compounds have been observed after high exposures in accidents. Erroneous drinking of nickel salts caused headache, vomiting, abdominal pain, diarrhea, and other symptoms, which lasted for about one week [41]. Accidental inhalation of soluble nickel compounds caused headache, cough, and, at higher exposure, also lung edemas and fibroses. Symptoms were more pronounced after acute exposure to nickel tetracarbonyl [36]. With experimental animals inhalation of nickel carbonyl also caused kidney failure [29]. Lethal doses of various nickel compounds derived from animal experiments are listed in Table 4. The sparingly soluble species nickel metal powder, nickel oxide, and nickel sulfides exert a very low toxicity in short-term exposures. However, this does not hold for chronic toxicity (see below).

Table 4. Acute oral toxicity of soluble nickel salts in experimental animals (data compiled from [29, 43, 44])

Compound	Species	LD_{50} (mg Ni/kg body mass)
Nickel acetate	rat	350
	mouse	420
Nickel chloride	rat	200
Nickel sulfate	dog	25 – 50
	pig	250
	chicken	300
	rat	300
Nickel oxide	rat	> 5000
Nickel subsulfide	rat	> 5000
Nickel sulfide	rat	> 5000
Nickel metal powder	rat	> 9000

Chronic Toxicity.

Skin. Dermal exposure to nickel metal and nickel salts results in a specific nickel contact dermatitis which was first observed in 1899 in workers in electrolytic plating facilities, and later in hairdressers and medical employees using nickel-plated tools like scissors, forceps, and other medical instruments and in dental technicians abrading dental nickel alloys [45]. A high incidence of nickel allergic dermatitis is observed in the general population after skin contact with nickel-plated jewelry, jeans buttons, or watches. Highest incidences are found in people wearing pierced jewelry [46]. For some decades, 5% of all eczemas registered in the general population were attributed to dermal nickel exposure, but due to preventive measures in the last decade the prevalence of this disease is declining considerably [47]. Nickel dermatitis is caused by pure nickel and high-nickel alloys only, but not by dermal contact with objects made from stainless steel with its lower nickel content.

Respiratory Organs. Whereas the nickel concentrations in ambient air usually are very low, high air concentrations have been experienced in nickel producing and processing industries. Exposure to elevated concentrations of sparingly soluble nickel sulfides and oxides was frequent in nickel mining and smelting, and exposure to soluble nickel compounds was high in nickel refining and nickel electrolytic plating plants. Two types of health effects were registered with nickel workers: inflammatory

symptoms and cancer, both of the respiratory tract.

Nonmalignant Effects. The highest incidence of inflammatory diseases was observed with workers exposed to soluble nickel salts. The symptoms comprise inflammation and dysplasia of the nasal epithelium, pneumonia, and lung fibrosis [29, 43].

Carcinogenicity. Nickel compounds are consistently classified as human carcinogens by various national and international agencies, whereas nickel metal is regarded as carcinogenic to humans only by some organizations (Table 5).

Nickel Metal. The epidemiological evidence for the carcinogenicity of nickel metal is blurred by coexposure to other nickel compounds. Cohorts of workers exposed to high concentrations of nickel metal dust but coexposed to oxidic and sulfidic nickel exhibited elevated incidences of nasal and lung cancer [43]. A large number of workers manufacturing high-nickel alloys in the USA showed an enhanced mortality from lung cancer when the total U.S. population was taken as control, but this correlation was not significant if the local population served as control [49]. In experimental studies, nickel metal powder caused local carcinomas in rats after intraperitoneal injection and intratracheal instillation [50, 51].

Soluble Nickel Salts. The classification of soluble nickel compounds as carcinogenic to humans is based on elevated incidences of nasal and lung cancer in nickel refinery workers who had been exposed predominantly to soluble and to a lesser extent to insoluble nickel compounds. In a Norwegian factory where the fraction of soluble nickel in the air was 90%, high incidences of lung cancer were observed, which were related to the atmospheric concentration of soluble nickel in the range between 1 and 15 mg Ni/m^3 [52]. The epidemiological data are seemingly contradicted by the negative result of an experimental animal study in which no tumors were induced by inhalation of nickel sulfate up to a concentration of 0.11 mg Ni/m^3 in rats and 0.22 mg Ni/m^3 in mice [53]. This disagreement may be interpreted by taking into account that the concentrations in the animal study were confined to much lower values than the human exposures because of the onset of severe inflammatory effects at the lower experimental doses [53].

Sulfidic and Oxidic Nickel. Although in industry there never has been pure exposure to either sulfidic or oxidic nickel, some cohorts of nickel smelter workers were predominantly exposed to nickel sulfides and other cohorts predominantly to nickel oxides. Both cohorts showed significantly enhanced incidences of nasal and lung carcinomas [54], and therefore both sulfidic and oxidic nickel have been classified as human carcinogens. This classification is supported by the carcinogenicity of nickel subsulfide and nickel oxide in animal experiments [53].

Toxicity to Other Organs.

Nephrotoxicity. There is limited evidence for nephrotoxicity of soluble nickel compounds in highly exposed workers [55] and after oral uptake of soluble nickel salts by experimental animals [29]. Inhalation of nickel tetracarbonyl caused kidney failure in rats [29].

Reproductive Toxicity. In experiments with rats, oral and intraperitoneal administration of soluble nickel salts interfered with reproduction. Nickel sulfate caused diminished fertility, decreased number and birth weight of pups, and a pattern of anomalies (malformation of eyes, lung, kidney, bones, and palatal clefts) [43]. Inhalation of nickel tetracarbonyl induced teratogenic effects in rats [56].

Genotoxicity. Nickel compounds are only weakly genotoxic. Nickel salts were inactive in nearly all bacterial mutagenicity assays and only weakly mutagenic in mammalian cell tests. Only elevated (millimolar) concentrations of soluble nickel salts induced chromosomal aberrations, sister chromatid exchanges, and DNA damage in mammalian cells in vitro [43]. However, in cellular systems nickel ions enhanced the genotoxicity of alkylating agents and UV irradiation, because they inhibit the repair of DNA damage induced by these agents [57].

Environmental Toxicology.

Effects on Microorganisms. Nickel is an essential element for anaerobic, hydrogen-producing bacteria [58]. The toxicity of nickel

Table 5. Classification of nickel and nickel compounds as carcinogens

Agency	Substances	Classification
IARC 1990 [43]	nickel compounds	1 (carcinogenic to humans)
	nickel metal	2B (possibly carcinogenic to humans)
DFG 2001 [48]	nickel and nickel compounds	1 (causes cancer in humans)
EU	nickel metal	3 (possibly carcinogenic to humans)
	nickel carbonate	3 (possibly carcinogenic to humans)
	nickel dioxide	1 (carcinogenic to humans)
	nickel monoxide	1 (carcinogenic to humans)
	nickel sulfate	3 (possibly carcinogenic to humans)
	nickel sulfide	1 (carcinogenic to humans)
	nickel tetracarbonyl	3 (possibly carcinogenic to humans)
ACGIH 2003	nickel subsulfide	A1 (confirmed human carcinogen)
EPA 2003	nickel refinery dust	A (human carcinogen)
	nickel carbonyl	B2 (probable human carcinogen)
	nickel subsulfide	A (human carcinogen)
NTP (National Toxicology Program), USA 2002	nickel metal	reasonably anticipated to be human carcinogen
	nickel compounds	known human carcinogens

Table 6. Synopsis of limit values for nickel and nickel compounds (as elemental nickel).

Description	Limit values	Source
Occupational limit values, nickel metal dust	0.5 mg/m^{3}1.0 mg/m^3	France, Germany, UKOSHA 2003
Occupational limit values, insoluble compounds (oxidic and sulfidic nickel)	0.5 mg/m^{3}1.0 mg/m^3	France, Germany, UKOSHA 2003
Occupational limit values, soluble nickel salts	0.05 mg/m^{3}0.1 mg/m^{3}1.0 mg/m^3	GermanyFrance, UKOSHA 2003
Occupational limit values, nickel tetracarbonyl	0.15 mg/m^{3}0.3 mg/m^{3}0.007 mg/m^3	GermanyFranceUK, OSHA 2003
General ambient air	0.5 mg/m^3	TA Luft 2002
Effluent discharge	8 – 100 µg/L	EC directive 76/464
Sewage sludge	300 – 400 mg/kg d.m.200 mg/kg d.m.	EC directive 86/278German KlärschlammVO 1992
Soil	30 – 75 mg/kg d.m.15 – 70 mg/kg d.m., depending on soil type	EC Directive 86/278German BbodSchV 1999
Drinking water	0.05 mg/L0.02 mg/L0.1 mg/L	EC Directive 80/778Germany TrinkwasserVO 2001US EPA 2002
Food, recommended limit for uptake of soluble nickel	1 mg/d	US National Academy of Sciences 2001 [62].

to bacteria, yeasts, and fungi differs considerably [59]. The growth of eubacteria is inhibited by nickel concentrations of 5 - 30 µg Ni/L, whereas filamentous fungi are inhibited by 5 - 1000 µg Ni/L. In acid environment the toxicity of nickel compounds to microorganisms is elevated because of enhanced solubilization from complexes.

Effects on Plants. Nickel is essential for the growth of those plants which have been investigated in this respect. Terrestrial plants accumulate nickel and on average contain between 0.3 and 6 mg/kg d.m. Soil concentrations of more than 70 mg/kg d.m. are generally phytotoxic with the exception of plants adapted to nickel-rich soils. Aquatic algae tolerate higher concentrations of nickel and may accumulate the element up to 1000-fold, corresponding to 150 mg/kg d.m. [60].

Regulations and Guidelines. The qualitative classifications of nickel and nickel compounds with respect to carcinogenicity are summarized in Table 5. The quantitative risk of cancer incidence was estimated by the WHO [61] by a unit risk of 4×10^{-4} per µg Ni/m^3. A synopsis of regulations, guidelines an legislation for nickel is given in Table 6.

9. References

1. J. Mellor: *A Comprehensive Treatise on Inorganic and Theoretical Chemistry*, vol. **15**, Longmans, Greens and Co., London 1936, pp. 1 – 497.

2. *Gmelin*, **system no. 57, part B** 2.
3. D. Nicholls in J. C. Bailar, H. J. Emeleus, R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorganic Chemistry*, vol. **3**, Pergamon Press, Oxford 1973, pp. 1109 – 1162.
4. E. R. Braithwaite in R. Thompson (ed.): *Speciality Inorganic Chemicals*, RSC special publication 40, 1981, pp. 375 – 402.
5. L. Sacconi, F. Mani, A. Bencini in G. Wilkinson, R. D. Gillard, J. A. McCleverty (eds.): *Comprehensive Coordination Chemistry*, vol. **5**, Pergamon Press, Oxford 1987, pp. 1 – 347.
6. G. C. Bond, J. B. P. Tripathi, *J. Chem. Soc. Faraday Trans.* **73** (1977) 545 – 552.
7. T. E. Girich, A. K. Buchinskii, *Zh. Prikl. Khim. (Leningrad)* **59** (1986) 884 – 886.
   T. E. Girich et al., *Zh. Neorg. Khim.* **31** (1986) 1575 – 1577
   T. E. Girich et al., *Vopr. Khim. Khim. Tekhnol.* **82** (1986) 60 – 63.
8. Y. Monteil, P. Raffin, J. Bouix, *Thermochim. Acta* **125** (1988) 327 – 346.
9. F. Boix et al., *Synth. Commun.* **17** (1987) 1149 – 1153.
10. L. G. Wiseman, *Weld. J. (Miami)* **68** (1989) 192 – 197.
11. P. M. Houpt, A. Van der Waal, F. Langeweg, *Anal. Chim. Acta* **136** (1982) 421 – 424.
12. P. W. Jolly, G. Wilke: *The Organic Chemistry of Nickel,* vol. 1, Academic Press, New York 1974.
13. P. W. Jolly in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry*, vol. **6**, Pergamon Press, Oxford 1982, pp. 1 – 36.
14. *The Sigma-Aldrich Library of Chemical Safety Data*, ed. II, vol. **II**, Sigma Aldrich Corporation, 1988.
15. D. Nicholls in: *Comprehensive Inorganic Chemistry,* vol. 4, Pergamon Press, Oxford 1973, p. 1149.
16. P. W. Jolly in: *Comprehensive Organometallic Chemistry,* vol. 8, Pergamon Press, Oxford 1982.
17. P. W. Jolly, G. Wilke in: *The Organic Chemistry of Nickel,* vol. 2, Academic Press, New York 1975, p. 6.
18. A. I. Vogel: *A Textbook of Quantitative Inorganic Analysis*, 4th ed., Longman, 1978, p. 447.
19. P. W. Jolly in: *Comprehensive Organometallic Chemistry,* vol. 6, Pergamon Press, Oxford 1982, pp. 52 – 64.
20. L. Sacconi, F. Mani in: *Comprehensive Co-ordination Chemistry,* vol. 5, Pergamon Press, Oxford 1987, pp. 98 – 102.
21. Shell Int. Res. Maatsch. N.V., DE-OS 2 159 370, 1972 (D. M. Singleton, P. W. Glockner, W. Keim).
22. R. Uttech, H. Dietrich, *Z. Kristallogr.* **122** (1965) 60.
23. P. W. Jolly, G. Wilke in: *The Organic Chemistry of Nickel,* vol. 1, Academic Press, New York 1974, pp. 421 – 422.
24. Shell Oil Co., US 3 424 816, 1969 (J. D. McClure, K. W. Barnett).
25. Union Carbide Corp., US 3 375 287, 1968 (S. W. Tinsley, E. A. Rick, J. E. McKeon).
26. Toyo Rayon Co., JP 797 ('67), 1967 (K. Tani, S. Yuguchi).
27. Y. Kiso, K. Tamao, M. Kumada, *J. Organomet. Chem.* **76** (1974) 95.
28. F. W. Sunderman Jr., "Selected Methods of Analysis," *IARC Monogr. Environ. Carcinog.*, (1986) pp. 79 – 92.
29. IPCS (International Programme on Chemical Safety): *Nickel*. Environmental Health Criteria, vol. **108**, World Health Organization, Geneva 1991.
30. UBA (Umweltbundesamt): *Daten zur Umwelt* 2004. www.uba.de.
31. M. Stoeppler, P. Ostapczuk: "Nickel and Cobalt" in M. Stoeppler (ed.): *Hazardous Metals in the Environment*, vol. 12, Elsevier, Amsterdam 1992, pp. 405 – 450.
32. G. D. Nielsen: "Oral Challenge of Nickel-Allergic Patients with Hand Eczema" in E. Nieboer, J. O. Nriagu (eds.): *Nickel and Human Health: Current Perspectives*, Wiley, New York 1992, pp. 201 – 210.
33. M. Accominotti et al.: "Contribution of Chromium and Nickel Enrichment During Cooking of Foods in Stainless Steel", *Contact Dermatitis* **38** (1998) 305 – 310.
34. M. Anke: "Essential and Toxic effects of Macro, Trace, and Ultratrace Elements in the Nutrition of Man" in E. Merian, M. Anke, M. Ihnat, M. Stoeppler (eds.): *Elements and Their Compounds in the Environment*, Wiley-VCH, Weinheim 2004, pp. 343 – 367.
35. M. Anke: "Essential and Toxic Effects of Macro, Trace, and Ultratrace Elements in the Nutrition of Animals", in E. Merian, M. Anke, M. Ihnat, M. Stoeppler (eds.): *Elements and Their Compounds in the Environment*, Wiley-VCH, Weinheim 2004, pp. 305 – 341.

36. F. W. Sunderman: "Nickel", in E. Merian, M. Anke, M. Ihnat, M. Stoeppler (eds.): *Elements and Their Compounds in the Environment*, Wiley-VCH, Weinheim 2004, pp. 841 – 865.
37. K. B. Svenes, I. Andersen: "Distribution of Nickel in Lungs from Former Nickel Workers", *Int. Arch. Occup. Environ. Health* **71** (1998) 424 – 428.
38. J. Angerer, R. Heinrich-Ramm, G. Lehnert: "Occupational Exposure to Cobalt and Nickel. Biological Monitoring", *Int. J. Environ. Anal. Chem.* **35** (1989) 81 – 88.
39. W. N. Rezuke, J. A. Knight, F. W. Sunderman: "Reference Values for Nickel Concentrations in Human Tissues and Bile", *Am. J. Ind. Med.* **11** (1987) 419 – 426.
40. F. W. Sunderman: "Nickel Absorption and Kinetics in Human Volunteers", *Proc. Soc. Exp. Biol. Med.* **191** (1989) 5 – 11.
41. F. W. Sunderman, B. Dingle, S. M. Hopfer, T. Swift: "Acute Nickel Toxicity in Electroplating Workers who Accidentally Ingested a Solution of Nickel Sulfate and Nickel Chloride", *Am. J. Indust. Med.* **14** (1988) 257 – 266.
42. J. Angerer, G. Lehnert: "Occupational Chronic Exposure to Metals. II. Nickel Exposure of Stainless Steel Workers – Biological Monitoring", *Int. Arch. Occup. Environ. Health* **62** (1990) 7 – 10.
43. IARC: "Nickel and Nickel Compounds" in IARC (ed.): *Monographs on the Carcinogenic Risk to Humans*, vol. 49 *(Chromium, Nickel and Welding)*, International Agency for Research on Cancer Geneva 1990, pp. 257 – 445.
44. E. Mastromatteo: "Nickel", *Am. Ind. Hyg. Assoc.* **47** (1986) 589 – 601.
45. F. W. Sunderman: "A Pilgrimage into the Archives of Nickel Toxicology", *Ann. Clin. Lab. Sci.* **19** (1989) 1 – 16.
46. T. Menné: "Nickel Contact Hypersensitivity" in: E. Nieboer, J. O. Nriagu (eds.): *Nickel and Human Health: Current Perspectives*. Wiley, New York 1992, pp. 193 – 200
47. A. Schnuch, J. Geier, H. Lessmann, W. Uter: "Rückgang der Nickelkontaktallergie in den letzten Jahren. Eine Folge der „Nickel-Verordnung"?" *Der Hautarzt* **7** (2003) 626 – 632.
48. Deutsche Forschungsgemeinschaft: "Nickel und Nickelverbindungen" in Deutsche Forschungsgemeinschaft (ed.): *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*, 32. Lieferung, Wiley-VCH, Weinheim 2001.
49. V. C. Arena, N. B. Sussman, C. K. Redmond, J. P. Costantino, J. M. Trauth: "Using Alternative Comparison Populations to Assess Occupation-Related Mortality Risk", *J. Occup. Environ. Med.* **40** (1998) 907 – 916.
50. F. Pott et al.: "Carcinogenicity Studies on Fibres, Metal Compounds, and some other Dusts in Rats". *Exp. Pathol.* **32** (1987) 129 – 152.
51. F. Pott, M. Rippe, M. Roller, M. Csicsaky, M. Rosenbruch: "Carcinogenicity of Nickel Compounds and Nickel Alloys in Rats by Intraperitoneal Injection" in E. Nieboer, J. O. Nriagu (eds.): *Nickel and Human Health: Current Perspectives*, Wiley, New York 1992, pp. 491 – 502.
52. A. Andersen, S. R. Berge, A. Engeland, T. Norseth: "Exposure to Nickel Compounds and Smoking in Relation to Incidence of Lung and Nasal Cancer among Nickel Refinery Workers", *Occup. Environ. Med.* **53** (1996) 708 – 713.
53. J. K. Dunnick et al.: "Comparative Carcinogenic Effects of Nickel Subsulfide, Nickel Oxide, or Nickel Sulfate Hexahydrate Chronic Exposures in the Lung", *Cancer Res.* **55** (1995) 5251 – 5256.
54. R. Doll: "Report of the International Committee on Nickel Carcinogenesis in Man". *Scand. J. Work Environ. Health* **16** (1990) 9 – 82
55. W. E. Sanford, E. Nieboer: "Renal Toxicity of Nickel in Humans", in: E. Nieboer, J. O. Nriagu (eds.): *Nickel and Human Health: Current Perspectives*, Wiley, New York 1992, pp. 123 – 134.
56. F. W. Sunderman, M. C. Reid, S. K. Shen, C. B. Kervokian: "Embryotoxixity and Teratogenicity of Nickel Compounds" in T. W. Clarkson, E. F. Nordberg (eds.): *Reproductive and Developmental Toxicity of Metals*, Plenum Press, New York 1983, pp. 399 – 416.
57. A. Hartwig et al.: "Interference by Toxic Metal Ions with DNA Repair Processes and Cell Cycle Control: Molecular Mechanisms", *Environ. Health Perspect.* **110** (2002) Suppl. 5, 797 – 799.
58. R. K. Thauer: "Nickel to the Fore", *Science* **293** (2001) 1264 – 1265.
59. H. Babich, G. Stotzky: "Nickel Toxicity to Microbes: Effects of pH and Implications for Acid Rain", *Environ. Res.* **29** (1982) 335 – 350.

60. D. P. Spencer, R. W. Greene: "Effects of Nickel on Seven Species of Freshwater Algae", *Environ. Pollut.* **A25** (1981) 241 – 247.
61. World Health Organization: "Air Quality Guidelines for Europe", WHO Regional Publications, European Series 23, WHO, Copenhagen 1987.
62. National Academy of Sciences (USA): "Dietary Reference Intakes for Vitamin A, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc", Report of the Food and Nutrition Board, National Academy of Sciences, Washington, DC 2001, pp. 197 – 223.

Niobium and Niobium Compounds

JOACHIM ECKERT, Hermann C. Starck Berlin, Werk Goslar, Goslar, Federal Republic of Germany

1.	**Introduction**	1
2.	**Properties**	1
2.1.	**Physical Properties**	1
2.2.	**Chemical Properties**	2
3.	**Occurrence**	2
4.	**Processing of Niobium Ores**	3
4.1.	**Winning of Pyrochlore Concentrates**	3
4.2.	**Treatment of Pyrochlore Concentrates**	3
4.3.	**Production of Niobium Oxide from Columbites and Tantalites**	3
4.3.1.	Extraction Processes	3
4.3.2.	Chlorination	4
5.	**Niobium Compounds**	5
5.1.	**Oxides**	5
5.2.	**Halides**	6
5.3.	**Carbides and Hard Materials**	6
6.	**Niobium Metal**	8
6.1.	**Reduction of Niobium Pentoxide**	9
6.2.	**Reduction of Halides**	9
6.3.	**Refining**	9
6.4.	**Uses**	10
7.	**Ferroniobium**	10
7.1.	**Production**	11
7.2.	**Uses of Ferroniobium**	11
8.	**Analysis**	11
9.	**Economic Aspects**	12
10.	**References**	12

1. Introduction

Niobium [*7440-03-1*], Nb, atomic number 41, A_r 92.91, is also known as columbium, Cb, in the United States. There are many artificial radionuclides, but only one known natural nonradioactive isotope, ^{93}Nb. The electronic configuration of the ground state is $4s^2p^6d^3\ 5s^2$, which explains the existence of the oxidation states +2 to +5.

Niobium was discovered by HATCHETT in 1801 in the mineral columbite and was named columbium. In 1844 the name niobium was proposed by ROSE.

2. Properties

2.1. Physical Properties

Some important physical properties of niobium metal are as follows [1–3]:

Density at 20 °C	8.57 g/cm^3
Crystal structure	body-centered cubic ($a = 3.3 \times 10^{-10}$ m)
mp	2468 ± 10 °C
bp	4927 °C
Linear coefficient of thermal expansion	6.892×10^{-6} K^{-1}
Specific heat	0.26 kJ kg^{-1} K^{-1}
Latent heat of fusion	290 kJ/kg
Latent heat of vaporization	7490 kJ/kg
Thermal conductivity at 0 °C	0.533 Wcm^{-1} K^{-1}
Electrical resistivity at 0 °C	15.22 μΩ · cm
Temperature coefficient 0 – 600 °C	0.0396 K^{-1}
Electrochemical equivalent	0.19256 mg/C
Standard electrode potential E° Nb/Nb^{+5}	−0.96 V
Magnetic susceptibility at 25 °C	2.28×10^{-6}
Superconductivity T_C	9.13 K
Spectral emissivity at 650×10^{-10} m (at 2003 K)	2.28×10^{-6}
Ionisation potential	6.67 eV
Work function	4.01 eV

The mechanical properties, like those of most refractory metals, are influenced by the purity of the metal, the production method, and the mechanical treatment. Even small amounts of interstitial impurities increase the hardness and strength but reduce the ductility. Some important mechanical properties of commercial niobium are as follows [4]:

Annealed niobium	
Ultimate tensile strength	195 MPa
Yield strength	105 MPa
Elongation	30 % +
Reduction in area	80 % +
Hardness	60 HV
Poisson's ratio	0.38
Strain hardening exponent	0.24
Elastic modulus	
tension	103 GPa
shear	17.5 GPa

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_251

Table 1. Niobium deposits [6]

Location	Capacity, 10^6 kg Nb_2O_5	Ore grade	Reserves, 10^6 t Nb_2O_5	Production, 10^6 kg Nb_2O_5
Araxa, Brazil	24.75	3.0	500+	9.45
St. Honore, Quebec	3.15	0.7	11	3.15
Catalao, Brazil	2.48	1.5	50	2.48
Nigeria	0.9			
Thailand	0.45 – 0.9			
Zaire	pilot scale	2 – 3	extensive	0.23

Ductile – brittle transition temperature	< 147 K
Recrystallization temperature	800 – 1000 °C
Cold worked niobium	
Ultimate tensile strength	585 MPa
Elongation	5 %
Hardness	150 HV

2.2. Chemical Properties

Niobium is very resistant to most organic and inorganic acids, with the exception of HF, at temperatures up to 100 °C [1]. Concentrated sulfuric acid above 150 °C causes embrittlement. The resistance towards alkaline solutions is lower. Because niobium has a marked tendency to form oxides, hydrides, nitrides, and carbides, its use in air is limited to temperatures up to ca. 200 °C.

3. Occurrence

Niobium occupies the 33rd place in order of natural abundance, being present in the earth's crust at 24 µg/g. It is thus more common than cobalt, molybdenum, or tantalum.

The most important niobium mineral is pyrochlore, a compound with the general formula $(Ca,Na)_{2-m}Nb_2O_6\ (O,OH,F)_{1-n} \cdot x\ H_2O$. The lattice positions of Na and Ca can also be occupied by Ba, Sr, rare earths, Th, and U. The latter two elements are responsible for the radioactivity of some pyrochlore concentrates.

Two types of niobium ore deposits are known. In primary deposits, the pyrochlore is always interstratified in carbonatites. This is so in the Canadian deposits at Niobec and Oka, in which calciopyrochlore is interstratified in dolomite. The ore contains 0.5 – 0.7 % niobium pentoxide. In the important secondary Brazilian deposits at Araxa and Catalao, the niobium content of the carbonate minerals has been considerably enriched by weathering. Here, the ore is present in combination with apatite, iron oxide, and barite and contains about 3 % Nb_2O_5, 46 % Fe_2O_3, 17 – 18 % BaO, and 1.5 % P as apatite [5]. Table 1 shows the average Nb_2O_5 capacity, the production, and the reserves of various deposits [6]. In addition to the sources listed in Table 1, other, potential sources exist in Canada, Africa, Brazil, China, the United States, and the former Soviet Union.

The most commercially important deposits are in Brazil, Canada, Nigeria, and Zaire. World reserves are estimated to be 4.1×10^6 t Nb, of which 78 % are in Brazil. The Canadian reserves are estimated at 0.12×10^6t.

The second most important niobium mineral is columbite, $(Fe,Mn)(Nb,Ta)_2O_6$, in which niobium is nearly always present with tantalum. These ores are referred to as columbites if the Nb_2O_5 content is greater than that of Ta_2O_5, otherwise as tantalites. Columbites and tantalites contain at least 60 % combined pentoxides. These minerals occur as primary deposits in granites and pegmatites, or in alluvial secondary deposits. Table 2 lists the chemical compositions of various niobium and tantalum minerals [7].

Columbites are mined in Australia, Brazil, Nigeria, Malaysia, and Zaire. Since the discovery of the enormous deposits of pyrochlore in Brazil, world production of columbite has decreased considerably. This trend could change when the extraction of columbite as a by-product of tin extraction is started with the Pitinga deposits in Brazil [8]. The discovery of large deposits of pyrochlore in South Greenland (Motzfeld Center) was reported in 1986.

A further source of niobium is provided by tantalum – niobium slags from tin production, since columbites and tantalites are often associated with cassiterite; niobium and tantalum

Table 2. Chemical composition of the principal niobium-bearing minerals, % [7]

Mineral	Nb_2O_5	Ta_2O_5	TiO_2	Fe	MnO	SnO_2
Pyrochlore, $NaCaNb_2O_6F$	40 – 65	2	1 – 6	2		
Columbite, $(Fe,Mn)(Nb,Ta)_2O_6$	40 – 75	1 – 40	0.5 – 3	10 – 20	2.6	2
Tantalo-columbite, $(Fe,Mn)(Nb,Ta)_2O_6$	25 – 60	20 – 50	0.5 – 3	10 – 20	2.6	2
Tantalite, $(Fe,Mn)(Ta,Nb)_2O_6$	2 – 40	42 – 84	0.5 – 3	10 – 20	2.6	2
Microlite, $Ca_2(Ta,Nb)_2O_6(OH,F)$		60 – 70				

concentrate in the slags as oxides. The niobium and tantalum pentoxide content of various tin slags is given in Table 3 [9]. Niobium and tantalum are also found in rare minerals such as stibiocolumbites $(Sb,Nb)TaO_4$, fergusonites $(Re^{3+})NbO_4$, and euxenite $Y(Nb,Ti)_2O_6$. The last two of these, however, are of no commercial significance.

Table 3. Niobium and tantalum pentoxide content of various tin slags, % [9]

Country	Nb_2O_5	Ta_2O_5
Malaysia	4	4
Nigeria	14	4
Portugal	7	7
Singapore	3	2
Thailand	8	12
Zaire	5	9

4. Processing of Niobium Ores

4.1. Winning of Pyrochlore Concentrates

Concentrates containing 50 – 60 % Nb_2O_5 are obtained from pyrochlore-containing carbonatites or weathered ores by conventional beneficiation processes such as crushing, grinding, magnetic separation, and flotation. Brazilian concentrates (Araxa) are also treated chemically to remove lead, phosphorus, and sulfur. The raw concentrate is roasted in the presence of calcium chloride and calcium oxide in a rotary furnace at 800 – 900 °C, and the product is leached with hydrochloric acid. In the case of Brazilian pyrochlore, this procedure also leads to replacement of barium by calcium [10].

4.2. Treatment of Pyrochlore Concentrates

Whereas pyrochlore from Araxa can be converted directly to ferroniobium, pyrochlore concentrates from other sources must first be chemically pretreated before niobium can be obtained. Some well-known processes are treatment with concentrated sulfuric acid or with molten soda. Reductive chlorination of pyrochlore at ca. 1000 °C produces volatile chlorides of niobium and other metals. These processes are, however, generally regarded as uneconomical. This is also true of the process in which digestion of pyrochlore with mixtures of hydrofluoric and sulfuric acids is followed by solvent extraction [11].

4.3. Production of Niobium Oxide from Columbites and Tantalites

4.3.1. Extraction Processes

Niobium and tantalum always occur together in columbites and tantalites and must be separated not only from the other elements present but also from each other. The industrial separation of tantalum from niobium has long been carried out by the Marignac process of fractional crystallization of potassium heptafluorotantalate and potassium heptafluoroniobate [12]. This expensive process of precipitation and crystallization, which is also environmentally unacceptable, has been abandoned, together with the long established Fansteel process [12], in favor of processes based on solvent extraction.

Tantalite and columbite, either naturally occurring or synthetically produced as concentrates from tin slags [13, 14], are digested with hydrofluoric and sulfuric acids at elevated temperature. The accompanying elements are dissolved along with the tantalum and niobium, which form the complex heptafluorides H_2TaF_7 and H_2NbOF_5 or H_2NbF_7. After filtering off the insoluble residue (fluorides of alkaline earth and rare earth metals), the aqueous solution of

Ta – Nb in hydrofluoric acid is extracted in several continuously operated mixer-settlers with an organic solvent, e.g., methyl isobutyl ketone (MIBK) [15–17]. The complex fluorides of niobium and tantalum are extracted by the organic phase, whereas most of the impurities and other elements, such as iron, manganese, titanium, etc., remain in the aqueous phase. In practice, $Nb_2O_5 + Ta_2O_5$ concentrations of 150 – 200 g/L in the organic phase are used. The organic phase is washed with 6 – 15 N sulfuric acid and then reextracted with water or dilute sulfuric acid to obtain the niobium. The aqueous phase takes up the complex fluoroniobate and free hydrofluoric acid, while the complex fluorotantalate remains dissolved in the organic phase. The aqueous niobium solution is reextracted with a small amount of MIBK to remove traces of tantalum. The resulting organic phase is returned to the combined tantalum – niobium extraction stage. Gaseous or aqueous ammonia is added to the aqueous niobium solution to precipitate niobium oxide hydrate. Crystallization of K_2NbF_7 can only be achieved in strong hydrofluoric acid solution; therefore, it is only carried out on a small scale because of the high costs arising from the increased consumption of hydrofluoric acid.

The tantalum is reextracted from the organic phase with water or dilute ammonia solution, and tantalum oxide hydrate is precipitated by ammonia, or potassium salts are added to produce K_2TaF_7, which is used in the production of tantalum metal.

The oxide hydrates are collected by filtration, dried, and calcined at up to 1100 °C. Variation of the conditions of precipitation, drying and calcination produces different particle sizes, giving oxides suitable for various applications. Depending on the quality requirements, the calcination is carried out in directly or indirectly heated chamber or rotary furnaces. The nature of the furnace lining has considerable influence on purity.

Sophisticated process control and optimization enable niobium and tantalum to be produced with high yield (> 95 %) and purity (> 99.9 %).

A number of alternative extraction media have been reported in the literature, most of which have never been used in industry, except for tributyl phosphate (TBP) [18] and tri-*n*-octylphosphine oxide (TOPO) [19].

Figure 1 (see next page) shows the flow diagram of an industrial installation for the processing of tantalum –niobium raw materials [20].

4.3.2. Chlorination

The chlorination process is a modern alternative to the extraction process. There are two versions: reductive chlorination of natural and synthetic raw materials, and chlorination of tantalum – niobium ferroalloys.

In reductive chlorination, the ore or concentrate is pelletized with coal/coke and pitch, dried, and reacted in a stream of chlorine at 900 °C. The nonvolatile alkaline earth metal chlorides remain behind, while the readily volatilized tetrachlorides of silicon, tin, titanium, the pentachlorides $NbCl_5$ and $TaCl_5$, and $WOCl_4$ are distilled off and fractionated. The waste gas, which contains large amounts of phosgene and chlorine, must be rigorously purified.

The chlorination of ferroalloys is much simpler and more economical [21,22]. Ferroniobium or ferroniobium – tantalum are produced by the aluminothermic or electrothermic process, size reduced, and fed together with sodium chloride into a $NaCl - FeCl_3$ melt. The chlorinating agent is $NaFeCl_4$. Chlorine is passed into the melt, continuously regenerating $NaFeCl_4$. The following overall reactions take place:

$$FeTaNb + NaCl + 7\ NaFeCl_4 \rightarrow Ta/NbCl_5 + 8\ NaFeCl_3$$

$$8\ NaFeCl_3 + 4\ Cl_2 \rightarrow 8\ NaFeCl_4$$

The reaction temperature of 500 – 600 °C is much lower than that required for reductive chlorination. The volatile chlorides are evolved from the molten salt bath. The boiling points of $NbCl_5$, $TaCl_5$, and $WOCl_4$ lie between 228 and 248 °C, and these compounds must therefore be separated by means of a distillation column.

The chlorination of ferroalloys produces very pure niobium pentachloride and tantalum pentachloride in tonnage quantities. The $NbCl_5$ contains less than 30 µg/g Ta, and other metallic impurities only amount to 1 – 2 µg/g. Niobium pentachloride is an intermediate in the production of oxides and organometallic compounds, and is the starting material for production of niobium powder and for chemical vapor deposition (CVD) of niobium coatings.

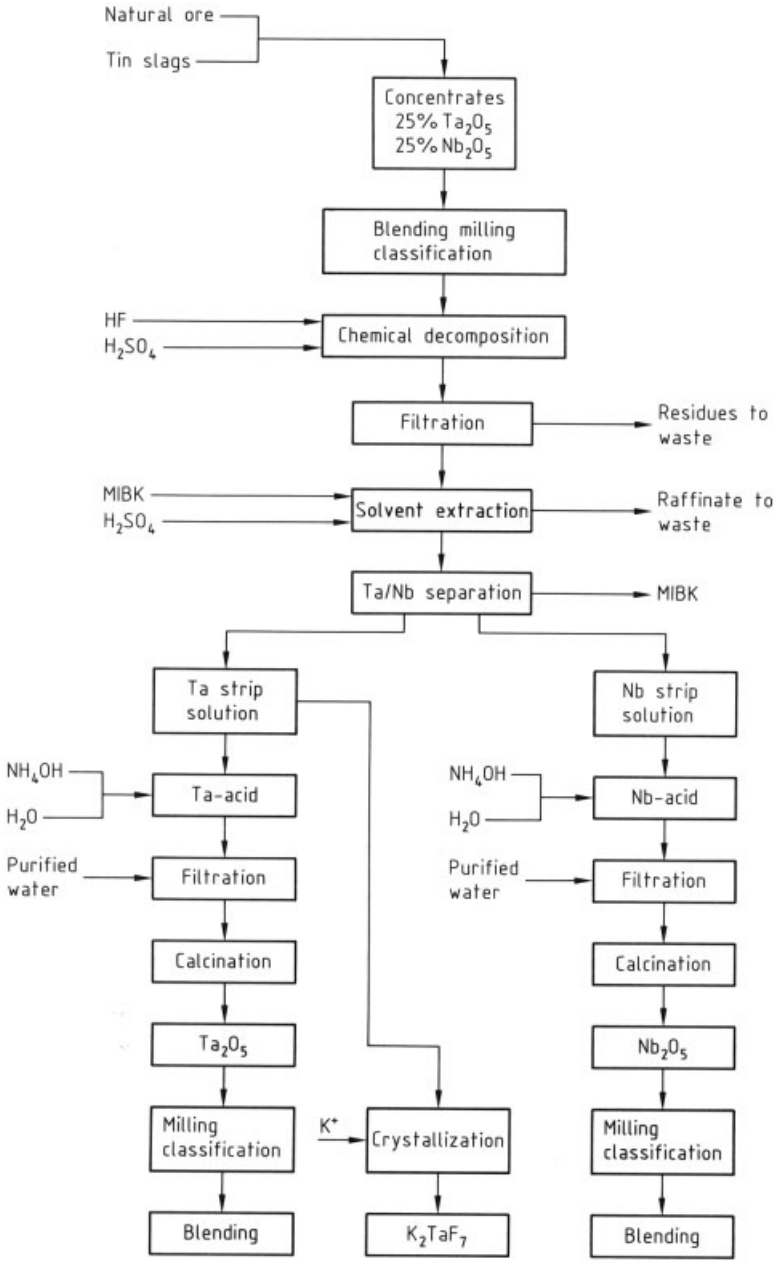

Figure 1. Flow diagram for the processing of tantalum – niobium raw materials [20]

5. Niobium Compounds

5.1. Oxides

Niobium pentoxide [*1313-96-8*], niobic acid, Nb_2O_5, *mp* 265 °C, *bp* 1495 °C, is a colorless powder that can only be dissolved by fusion with acidic or alkaline fluxes such as NaOH or $KHSO_4$, or in hydrofluoric acid. It is prepared by hydrolyzing solutions of alkali-metal niobates, niobium alkoxides (e.g., $Nb(OC_2H_5)_5$), or niobium pentachloride, or by precipitation from hydrofluoric acid solutions with alkali-metal hy-

droxides or ammonia. Depending on the method used, the oxide hydrate formed is either a gel that is difficult to filter or is flocculent. The oxide hydrate is filtered, washed, and calcined at 800–1100 °C. The temperature and treatment time determine which of the various crystalline modifications is formed. Nearly all the phase changes are irreversible [23–25].

Uses. Niobium pentoxide is used in metallurgy, for the production of hard materials, in optics, and in electronics. Hydrated Nb_2O_5, which can be regarded as an isopolyacid, has a high surface acidity, and catalyzes the polymerization of alkenes, e.g., propylene [26].

Various applications require different qualities of Nb_2O_5. For the electrothermic and metallothermic manufacture of niobium metal or its alloys, technical quality niobium pentoxide ($>98-99\,\%$) is suitable. For higher quality requirements, chemically pure niobium pentoxide ($>99.7\,\%$) is used.

In optics, niobium pentoxide is used as an additive to molten glass to prevent devitrification and to control properties such as refractive index and light absorption [27]. Optical grade niobium pentoxide, with a purity of $>99.9\,\%$, must be free from colored impurities such as chromium, nickel, iron, manganese, etc.

Extremely stringent purity requirements apply to Nb_2O_5 used in the manufacture of $LiNbO_3$ [*12031-63-9*] or $KNbO_3$ [*12030-85-2*] single crystals (Nb_2O_5 ultra pure grade, $>99.995\,\%$). These compounds are used for electroacoustic and electrooptical components such as modulators, frequency doublers, and wave filters [28, 29].

Ceramic grade niobium pentoxide, used for making dielectric materials, must be manufactured with a special particle size distribution. This market sector will demand greatly increased quantities of niobium pentoxide if the development of the new class of ferroelectric perovskites (relaxors) is successful, e.g., $Pb(Mg_{1/3}\,Nb_{2/3})O_3$ for the manufacture of ceramic capacitors [30–32].

Details of the various commercial qualities of niobium pentoxide are given in Table 4 (see next page) [20].

5.2. Halides

Niobium pentachloride [*10026-12-7*], $NbCl_5$, M_r 270.2, *mp* 209.5 °C, *bp* 249 °C, is now produced exclusively by chlorination of ferroniobium, niobium metal, or niobium scrap (see Section 4.3.2.).

Niobium pentachloride forms strongly hygroscopic yellow crystals that react with water to form $NbOCl_3$ or $Nb_2O_5 \cdot x\ H_2O$. The pentachloride is very soluble in dry ethanol, tetrahydrofuran, and benzene. Alcoholic solutions of niobium pentachloride are used in the production of niobium alkoxides such as the pentaethoxide, $Nb(OC_2H_5)_5$, from which micronized niobium pentoxide is produced [33]. Pure niobium pentachloride is used for large scale production of niobium pentoxide and niobium metal.

Niobium tetrachloride [*13569-70-5*], $NbCl_4$, sublimes between 350 and 400 °C. Niobium trichloride [*13569-59-0*], $NbCl_3$, disproportionates between 900 and 1000 °C. Niobium trichloride and niobium dichloride [*13777-09-8*], $NbCl_2$, are formed by reduction of $NbCl_5$ with hydrogen, but are of no industrial importance.

Niobium oxychloride [*13597-20-1*], $NbOCl_3$, M_r 215.28, is a colorless, crystalline compound that sublimes at 400 °C and partly decomposes into niobium pentoxide and niobium pentachloride on further heating.

Niobium pentafluoride [*7783-68-8*], NbF_5, M_r 187.91, *mp* 72 °C, *bp* 236 °C, can be prepared by fluorination of niobium pentachloride or niobium metal with fluorine or anhydrous HF. The reaction of Nb_2O_5 with aqueous HF yields fluoroniobic acids of various compositions (e.g., H_2NbF_7 or H_2NbOF_5), depending on the acid concentration. They are soluble in organic solvents and consequently play an important part in the separation of tantalum from niobium by solvent extraction.

5.3. Carbides and Hard Materials

The carbides, borides, silicides, and nitrides of niobium are metallic hard materials [34]. Some of their physical properties are listed in Table 5. Only the carbides are commercially important.

Table 4. Commercially available niobium oxides [20]

Specification	Technically pure grade	Chemically pure grade	Ceramic grade	Grade HPO (high purity optical)	Grade UP (ultra pure)
Composition					
Nb_2O_5	min. 99 %	min. 99.7 %	min. 99.8 %	min. 99.95 %	min. 99.997 %
Ta	max. 1600 ppm	max. 800 ppm	max. 500 ppm	max. 100 ppm	max. 8 ppm
Al				max. 2 ppm	max. 0.5 ppm
B					max. 0.2 ppm
Bi					max. 0.5 ppm
Ca	max. 700 ppm	max. 200 ppm		max. 10 ppm	max. 1 ppm
Cl					max. 10 ppm
Co				max. 2 ppm	max. 0.2 ppm
Cr				max. 2 ppm	max. 0.2 ppm
Cu				max. 2 ppm	max. 0.2 ppm
F					max. 1 ppm
Fe	max. 1000 ppm	max. 200 ppm	max. 100 ppm	max. 10 ppm	max. 1 ppm
K					max. 1 ppm
Mg					max. 0.5 ppm
Mn				max. 2 ppm	max. 0.2 ppm
Mo					max. 2 ppm
Na	max. 400 ppm	max. 200 ppm			max. 1 ppm
Ni			max. 100 ppm	max. 2 ppm	max. 0.2 ppm
Pb					max. 0.5 ppm
Rare earths					max. 1 ppm
S			max. 30 ppm		max. 1 ppm
Si	max. 1000 ppm	max. 250 ppm	max. 50 ppm	max. 50 ppm	max. 1 ppm
Sn					max. 0.5 ppm
Ti	max. 300 ppm	max. 100 ppm	max. 100 ppm		max. 1 ppm
V				max. 2 ppm	max. 0.5 ppm
W			max. 50 ppm		max. 1 ppm
Zr			max. 50 ppm		max. 1 ppm
Alkali			max. 200 ppm		
Loss on ignition	max. 1 %	max. 0.5 %	max. 0.4 %	max. 0.2 %	max. 0.005 %
Average particle size (FSSS)	1 – 10 µm	0.5 – 10 µm	0.7 – 0.8 µm	max. 5 µm	1 – 3 µm
Tap density			0.8 – 1 g/cm^3		0. 8 – 1.25 g/cm^3
Apparent density (Scott)			0.5 – 0.7 g/cm^3		
Microtac analysis			90 % < 1.2 µm 50 % 0.7 – 0.8 µm 10 % < 0.5 µm		
Grain size				1. < 600 µm HPO 600 2. < 400 µm HPO 400 3. < 150 µm HPO 150	< 255 µm
Crystal structure			orthorhombic (T-phase)	monoclinic (H-phase)	monoclinic (H-phase) + orthorhombic (T-phase)

Niobium Carbides (→ Carbides, Chap. 2.4.). The niobium – carbon phase diagram [35] indicates the existence of a face-centered cubic compound, NbC, which melts without decomposition at 3600 °C with a broad region of homogeneity, and of Nb_2C, which has a peritectic melting point. Only NbC has practical applications. In addition to its original use as a grain growth inhibitor in tungsten carbide – cobalt hard materials, it is now used mainly as solid solutions with titanium carbide, tantalum carbide, and tungsten carbide for cutting tools. The precise effect of niobium on the composition, microstructure, and properties of these hard materials is not yet fully understood [36]. In practice, up to 50 % of the more expensive TaC can be replaced by NbC without appreciably affecting the hardness or fracture strength [37].

Production, like that of TaC, is by carburizing the oxide, hydride, or metal at 1500 °C in a carbon tube furnace, vacuum furnace, or in the

Table 5. Physical properties of borides, carbides, nitrides, and silicides of niobium [34]

	Nb_3B_2	NbB	Nb_3B_4	NbB_2	Nb_2C	NbC	Nb_2N	NbN	Nb_4Si	Nb_5Si_3	$NbSi_2$
CAS registry number	[*12045-55-5*]	[*12045-19-1*]	[*12045-89-5*]	[*12007-29-3*]	[*12011-99-3*]	[*12069-94-2*]	[*12033-43-1*]	[*24621-21-4*]	[*12164-59-9*]	[*12060-34-3*]	[*12034-80-9*]
Crystal structure	tetragonal, U_3Si_2 type	orthorhombic, CrB type	orthorhombic, Ta_3B_4 type	hexagonal, AlB_2 type	hexagonal	face-centered cubic, NaCl type	hexagonal, W_2C type	face-centered cubic, NaCl type	hexagonal, $ZrSi_4$ type	α-tetragonal β-tetragonal	hexagonal, $CrSi_2$ type
Density, g/cm^3	8.0			6.6	7.83	7.78	8.33	8.2	7.74	6.26	5.45
Hardness*	2060	2200	2290	2600	2123	2400	2123	8**	550	600	700
mp, °C	1860 (decomp.)	2280	2700 (decomp.)	3000	3100 (decomp.)	3600			1950 (decomp.)	2480	1950
Electrical resistivity, $\mu\Omega\cdot$cm		64.5		12/34		35		60	10	40	50.4
Superconductivity transition temperature, K		8.25	1.27	1.27	9.18	6		9.5	15.2	1.27	1.2

* Vickers hardness at 0.5 N loading. ** Mohs hardness

presence of a molten metallic menstruum such as iron or aluminum. Some of the physical properties of niobium carbides are given in Table 5.

Niobium Borides. The niobium – boron phase diagram [38] shows the existence of Nb_3B_2, NbB, Nb_3B_4, and NbB_2. However, these high-melting, very hard materials, have no industrial uses. Table 5 lists their physical properties.

Niobium silicides, Nb_4Si, Nb_5Si_3, and $NbSi_2$, are produced by the silicothermic reduction of Nb_2O_5 or from the elements by sintering, pressure sintering, or fusion in an electric arc furnace. Some physical data are given in Table 5. These compounds are not used industrially.

Niobium Nitrides. Various nitride phases, with regions of homogeneity varying in distinctness, are obtained by heating niobium metal or mixtures of niobium oxide and carbon in a stream of ammonia or nitrogen [39]. Niobium metal can be produced by thermal decomposition of NbN in high vacuum. The pure nitrides have not as yet found any industrial application. Table 5 lists some of the physical properties of niobium nitrides.

Niobium Hydrides. On heating niobium in a pure hydrogen atmosphere to 350 – 500 °C, hydrogen (ca. 44 atom %) is absorbed, causing expansion of the lattice and embrittlement. Apart from the $NbH_{0.9}$ [*13981-86-7*] thus formed, the hydride Nb_4H_3 [*56941-03-8*] [40,41] and the unstable dihydride NbH_2 [*13981-96-9*] [42] are also known.

The reversible uptake and release of hydrogen by NbH is used in industry as a means of producing niobium powder from compact niobium, e.g., scrap or ingots. The metal is first hydrogenated, then ground, and finally dehydrogenated in a vacuum furnace or under an inert gas.

6. Niobium Metal

The industrial production of niobium is usually carried out by the reduction of the pentoxide or halides. The crude metal is then refined electrothermally.

6.1. Reduction of Niobium Pentoxide

Niobium pentoxide is reacted with carbon, aluminum, or silicon at high temperature. In the carbothermic process, niobium pentoxide is mixed with carbon black, and the mixture is pelletized and reduced in a vacuum furnace in a two-stage process at 1950 °C:

$$Nb_2O_5 + 7\,C \rightarrow 2\,NbC + 5\,CO$$
$$5\,NbC + Nb_2O_5 \rightarrow 7\,Nb + 5\,CO$$

The reaction takes place with the formation of NbC, Nb_2C, NbO_2 and NbO [43,44]. A crude product containing oxygen and carbon is produced, which must be refined in further high-temperature processes.

The following one-stage carbonitrothermic reduction of niobium pentoxide has been successfully carried out on the pilot-plant scale. A pelletized mixture of niobium pentoxide and carbon black is reacted in a stream of ammonia in an induction furnace at 1570 ± 20 °C to form niobium nitride, NbN, which is then thermally decomposed in high vacuum at 2000 – 2100 °C. However, the process does not yet appear to have been used on an industrial scale [45,46].

Over 90 % of niobium metal is produced by aluminothermic reduction:

$$3\,Nb_2O_5 + 10\,Al \rightarrow 6\,Nb + 5\,Al_2O_3$$

Niobium metal produced by this process is designated ATR niobium. In general, an excess of aluminum is used, producing a niobium – aluminum alloy which is melted in a vacuum, electric arc, or electron beam furnace to produce low-oxygen, carbon-free niobium [47,48]. The amount of excess aluminum determines the yield of niobium and its oxygen content [49].

The silicothermic reduction of niobium pentoxide can lead to silicide formation, and the thermodynamics of this reaction are less favorable. This process is therefore not in regular use.

6.2. Reduction of Halides

The reduction of niobium pentachloride can be achieved with hydrogen at 600 – 650 °C [50]. Alternatively, niobium pentachloride vapor is reduced with hydrogen in a fluidized bed furnace at 750 – 1050 °C [51] or in a hydrogen plasma above 2000 °C [52]. Metallothermic reduction with sodium or magnesium has also been reported [53]. The company TOHO Titanium Corporation has commissioned a plant with a capacity of 30 t/a niobium in which niobium pentachloride is reduced by magnesium [54]. A reduction process with zinc using the auxiliary metal bath technique is also known [55]. In contrast to the production of tantalum, the reduction of the fluoroniobates K_2NbF_7 and K_2NbOF_5 has not become industrially significant.

Very pure niobium can be obtained by electrowinning from oxygen-free molten salt systems with double fluorides or chlorides as the source of metal [56]. These molten salts are very corrosive and the current efficiencies are low. For these reasons, molten salt electrowinning of metallic niobium is not carried out commercially.

6.3. Refining

Crude niobium must be refined in order to remove impurities introduced either from the raw materials or during the treatment stages.

Purification can be achieved by high-temperature treatment because the melting point of niobium (2497 °C) is so high that most other elements can be removed by vaporization. Niobium is extremely reactive towards all but the inert gases [57], and melting must therefore always be carried out in a high or ultra-high vacuum or under a pure inert gas to minimize the concentration of harmful interstitial foreign atoms.

In comparison to the classical sintering and electric arc melting processes, electron beam melting (EBM) has considerable technical advantages for the production of pure metal.

The vaporization of the impurities can be controlled in each melting cycle by[58,59]:

1) Optimizing the melting rate,
2) Maintaining the bath in a liquid state for a long period, and
3) Controlled superheating of the molten metal.

Sufficient refinement cannot be achieved with one melting cycle only, so that in practice the solid ingot produced must be remelted. The melting rate (kg/h Nb) for the first melt (main refining step) depends on the raw material and

is lower for ATR niobium because of the high content of aluminum and NbO that must be evaporated compared with compacted granular niobium. Further development of the plasma melting technique would enable a simple plasma furnace to be used for the first melting operation [60, 61]. In the second and third melting cycles, the melting rate can be increased by a factor of 3 – 4, depending on the specification of the material.

Optimization of the equipment used in the electron beam melting process, leading to reduction in residual gas pressure and leakage rate, has enabled standard quality niobium to be produced from ATR niobium on a large scale with only two melting cycles [62].

The electron beam melting and remelting technique enables niobium metal with less than 50 µg/g interstitial impurities to be produced for high-frequency superconductors. Figure 2 shows the concentration of interstitial impurities (O, N, and C) as a function of the number of melting cycles [63].

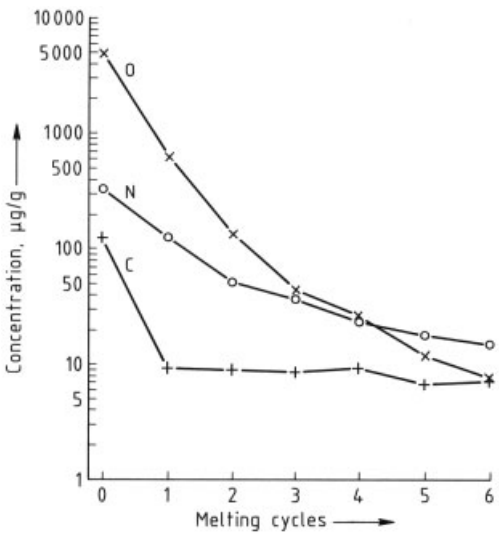

Figure 2. Contents of interstitial impurities (O, N, and C) of an electron beam melted niobium ingot as a function of the number of melting cycles [63]

Ultra-high purity niobium can be obtained from metal melted by the EBM method by a series of additional sophisticated process steps, including electrorefining and electron beam zone melting, followed by high-temperature treatment in high vacuum. By these methods, niobium metal can be prepared with metallic impurities in the parts per billion range and with interstitial impurities < 1 ppm [64].

6.4. Uses

Niobium has good resistance towards corrosive chemicals [4, 65] even at high temperatures, and is therefore used in the construction of chemical equipment, though it is not quite so resistant as tantalum. The process of oxide dispersion hardening of niobium with titanium dioxide gives a material which can be used both for chemical equipment and for medical implants subjected to high mechanical loading [66]. High-niobium alloys such as KBI 40/41 [67] can also be used under these conditions, for which formerly the more expensive metal tantalum had to be used.

Because niobium has a low neutron capture cross-section and is unusually resistant towards corrosion by liquid sodium, it is used in the pure state or as the alloy NbZr1 in the nuclear industry for the production of fuel element cans. This alloy is also used for the sealing caps of sodium vapor lamps.

World consumption of niobium for the production of superconducting materials, such as Nb_3Sn and NbTi is estimated to be 60 – 70 t/a.

The addition of niobium to titanium – aluminum alloys imparts the ductility needed to fabricate this material for the aircraft and space industry [68].

Other newly developed alloys C 103 (NbHf10Ti1) and Nb 752 (NbW10Zr25) combined with a special coating technique enable niobium to be used in the manufacture of nozzles and combustion chambers for rocket propulsion [69].

7. Ferroniobium

Approximately 85 – 90 % of the total niobium production is used in the steel industry in the form of iron – niobium alloy (ferroniobium) containing 40 – 70 % niobium. Depending on the application, the alloy can also contain small amounts of Ta (ferroniobium – tantalum), e.g., FeNb65Ta2, which contains 65 % Nb and 2 % Ta. Other alloy specifications are given in [70].

7.1. Production

Ferroniobium is usually produced by aluminothermic reduction of niobium oxide ores, with the addition of iron oxides if the niobium ore used contains insufficient iron. The starting materials are mainly columbites and pyrochlore concentrates.

The enthalpy of the reaction between Nb_2O_5 and Al is −276.1 kJ/mol Al, which is lower than the threshold value for self-sustaining aluminothermic reactions. The mixture must therefore either be preheated or mixed with oxygenreleasing compounds such as BaO_2, CaO_2, $KClO_4$, $KClO_3$ or $NaNO_3$.

Concentrates with lower percentages of niobium (ca. 40 %) can also be treated by the aluminothermic process in an electric arc furnace. Also, a two-stage electroaluminothermic process for the production of ferroniobium from columbite has been developed [71].

The method of operation is to charge the mixture of niobium concentrate with the additives to refractory lined reaction vessels. Either the whole mixture is reacted, or a small amount is set aside, ignited with a special exothermic mixture, and added to the bulk mixture. The molten reaction product is allowed to solidify in the furnace, and the block of metal separates from the slag. After cooling, it is broken into pieces of the required size.

In the aluminothermic process operated at Araxa, Fe – Nb blocks of metal up to 11 t in weight are produced. The yield of niobium metal is 96 – 97 %. The typical composition of a reaction charge is as follows:

Pyrochlore concentrate	18 000kg (60 % Nb_2O_5)
Iron oxide	4000 kg (68 % Fe)
Aluminum powder	6000 kg
Fluorspar	750 kg
Lime	500 kg

The reaction gives ca. 11 000 kg of ferroniobium of composition:

Nb	66.0 %
Fe	30.5 %
Si	1.5 %
Al	0.5 %
Ti	0.1 %
P	0.1 %
S	0.04 %
C	0.08 %
Pb	0.02 %

and ca. 20 000 kg of slag containing:

Al_2O_3	48.0 %
CaO	25.0 %
TiO_2	4.0 %
BaO	2.0 %
Re_2O_3	4.0 %
Nb_2O_5	trace
ThO_2	2.0 %
U_3O_8	0.05 %

Special niobium alloys, e.g., nickel – niobium, cobalt – niobium, aluminum – niobium and chromium – niobium, are also manufactured by the aluminothermic process. For these alloys, which have various niobium contents [72], niobium oxide is the only raw material used.

7.2. Uses of Ferroniobium

Niobium has very marked carbide and nitride forming properties and is therefore mainly used in the production of steels and gray iron [73]. The addition of niobium prevents intercrystalline corrosion in stainless austenitic chromium – nickel steels and improves corrosion resistance, weldability, ductility, and toughness in ferritic chromium steels [74]. Ferroniobium is also used as a trace additive in large quantities for the production of high-strength low-alloy (HSLA) steels used in automobile body manufacture, structural steels, concrete reinforcing steel, and pipelines. Superalloys containing up to 5 % niobium are used for gas turbine components. These materials have high hot strength [75, 76], and mostly have a nickel or copper base. The niobium is added in the form of nickel – niobium or cobalt – niobium.

8. Analysis

Extensive literature on the analytical determination of niobium is available in handbooks and monographs [77–80].

The determination of niobium in raw materials is carried out by X-ray fluorescence [81–83]. For this purpose, test pieces in tablet form are produced by reacting the niobium-containing material with borate to produce a melt, or by compressing it with binders such as wax or boric acid.

For ferroniobium, X-ray fluorescence analysis of HF solutions [84] or borate tablets is used.

For umpire assay, it is usual to separate the niobium from the other material using ion-exchange resins with HF solutions, followed by gravimetric determination [85]. Small niobium contents in steels are determined photometrically [86].

Metallic impurities in niobium pentoxide or niobium metal are determined by atomic absorption spectrometry (AAS) or atomic emission spectrometry (ICP-OES, DCP-OES) in HF solution [87]. Emission spectrum analysis in a d.c. current plasma arc is also used. The nonmetals oxygen, nitrogen, hydrogen, carbon, and sulfur are determined by extraction at high temperature either with a carrier gas or under vacuum, or by combustion analysis in a stream of oxygen [88]. The hydride-forming elements arsenic, antimony, bismuth, selenium, and tellurium can be converted to their hydrides and detected with high sensitivity by the AAS method. The anions Cl^- and F^- are separated by distillation and determined photometrically, by ion-selective electrodes, or by ion chromatography.

In all the above-mentioned methods, the usual limits of detection are from the low µg/g region down to the ng/g region in routine quality testing. For the analysis of ng/g trace impurities it is necessary to separate and concentrate the material [89, 90]. The final determination is carried out using AAS with a graphite furnace, ICP-OES, DCP-OES, or voltammetry. Recently, ICP mass spectrometry has become a powerful additional method for many metallic impurities. Niobium metal is analyzed simply and quickly for metallic and nonmetallic impurities down to the lower ng/g region by glow discharge mass spectrometry (GDMS) [91]. The additional techniques of neutron activation analysis [92, 93] or proton activation analysis [94] are used for verification.

9. Economic Aspects

The world consumption of niobium has, since 1980, reached $16-20\times10^6$ kg Nb_2O_5. The annual rate of increase is ca. 2 %. More than 90 % of total niobium production goes to the steel industry in the form of ferroniobium or niobium alloys, and therefore the demand for niobium is determined by the world market for iron and steel [95]. Figure 3 shows the clear relationship between steel production and niobium demand. Between 1984 and 1988, technical quality niobium pentoxide was marketed in the price range 14.3 – 15.2 \$/kg. The price of standard grade ferroniobium was 14.5 \$/kg Nb. In mid 1987, the price of niobium concentrate (> 65 % $Nb_2O_5 + Ta_2O_5$, 10 : 1) was 5.5 – 6.4 \$/kg Nb_2O_5 CIF Europe. Canadian pyrochlore concentrate was quoted at 5.9 \$/kg Nb_2O_5 FOB.

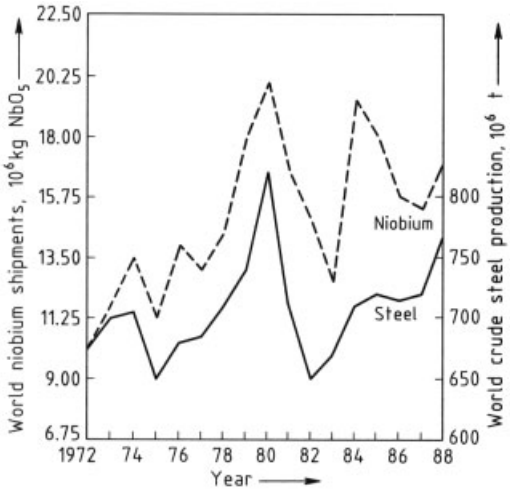

Figure 3. World niobium consumption compared to world crude steel production from 1972 – 1988 [95].

10. References

1. R. Kieffer, H. Braun: *Vanadin-Niob-Tantal,* Springer Verlag, Berlin-Göttingen-Heidelberg 1963.
2. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag und Verlag Stahleisen GmbH, Düsseldorf 1973.
3. C. English: "The Physical, Mechanical and Irridation Behavior of Niobium and Niobium-Base Alloys" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* The Metallurgical Society of AIME, New York 1984, pp. 239 – 324.
4. R. W. Balliett, M. Coscia, F. J. Hunkeler, *J. Met.* **38** (1986) no.9, 25 – 27.
5. G. Perrault, E. A. Manker: "Geology and Mineralogy of Niobium Deposits" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 3 – 77.

6. Tantalum-Niobium International Study Centre Brussels: *Int. J. Refract. Hard Met.,* **7** (1988) no.1, 3 – 7.
7. Roskill: *The Economics of Niobium,* 5th ed., Roskill Information Services, London 1986, p. 1.
8. Bureau of Mines, *Miner. Yearbook, Columbium and Tantalum,* 1987, p. 8.
9. Roskill: *The Economics of Niobium,* 5th ed., Roskill Information Services, London 1986, p. 3.
10. O. de Souza Paraiso, R. de Fuccio, *Min. Mag.* **146** (1982) no. 2, 134 – 147.
11. G. Gabra: *CIM Bull.* **78** (1985) no. 873, 80 – 85.
12. *Ullmann,* 4th ed., **17,** 305, 306.
13. H. C. Starck, DE 2 733 193, 1977 (B. Krismer, H. Pungs).
14. W. Rockenbauer: "Production of Niobium-Metal and Compounds from Tantalite, Columbite Natural Ores and Synthetic Tantalum-Niobium Concentrates" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 133 – 152.
15. K. B. Higbie, J. R. Werning: Bureau of Mines, Report of Investigations 5239 (1956).
16. Gesellschaft für Elektrometallurgie, H. C. Starck, DE-AS 1 018 036, 1957 (F. Brenthel, H. Rothmann, W. Keil).
17. Fansteel Metallurgical Corp., US 3 117 833, 1958 (J. A. Pierret).
18. T. H. Tunley, M. Fieberg: National Institute for Metallurgy South Africa, Report 1154 (1976).
19. Gesellschaft für Elektrometallurgie, DE 3 241 832, 1984 (G. Bauer, J. Eckert).
20. Hermann C. Starck Berlin GmbH & Co. KG, Product Information, *Tantalum and Niobium Compounds,* 1986.
21. H. C. Starck, US 3 407 031, 1970 (H. Fürer, W. Rockenbauer).
22. W. Rockenbauer: *Metall (Berlin)* **38** (1984) no. 2, 156 – 159.
23. H. Schäfer, R. Gruehn, F. Schulte, *Angew. Chem.* **78** (1966) 28 – 41.
24. H. J. Goldschmidt, *J. Inst. Met.* **87** (1958 – 59) 235 – 239.
25. H. Schäfer, A. Dürkop, M. Jori, *Z. Anorg. Allg. Chem.* **275** (1954) 289 – 296.
26. K. Tanabe, *Mater. Chem. Phys.* **17** (1987) 217 – 225.
27. T. Ichimura: "Niobium Oxide in Optical Glass Manufacture" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 603 – 614.
28. A. Räuber: *Current Topics in Materials Science,* **vol. I,** E. Kaldis, North-Holland Publishing Company, Amsterdam 1978, pp. 481 – 601.
29. T. Wearden, *New Electronics,* 20th Mar. 1984, 29 –31.
30. D. Hennings, *Keram. Z. Beilage* **41** (1989) 1 – 6.
31. S. L. Swartz, T. R. Shrout, *Mater. Res. Bull.* **17** (1982) 1245 – 1250.
32. TDK Electronics Co., Ltd.; US 4 265 668, 1981 (S. Fujiwara et al.)
33. N. Sato, M. Nassju, *High Temp. Mater. Processes* **8** (1988) no. 1, 39 – 46.
34. *Ullmann,* 4th ed., **12,** 523.
35. E. K. Storms: "Phase Carbides and Nitrides" in L. E. J. Roberts (ed.): *MTP International Review of Science,* vol. 10, Solid State Chemistry, Butterworths, London 1972.
36. B. Aronsson, L. J. Aschan: "Niobium in Cemented Carbide" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, 637 – 652.
37. R. Kieffer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963, pp. 127 – 135.
38. R. Kieffer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963, p. 405.
39. N. P. Lyakishev, N. A. Tulin, YU. L. Pliner: *Niobium in Steels and Alloys,* Companhia Brasileira de Metalurgia e Mineracao-CBMM, Sao Paulo 1984, pp. 38 – 40.
40. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* 2nd ed., McGraw-Hill, New York 1958.
41. R. P. Elliott: *Constitution of Binary Alloys,* 1st Suppl., McGraw-Hill, New York 1965.
42. G. Brauer, H. Müller, *J. Inorg. Nucl. Chem.* **17** (1961) 102 – 107.
43. J. Krüger: "Use of Vacuum Techniques in Extractive Metallurgy and Refining of Metals" in O. Winkler, R. Bakish (eds.): *Vacuum Metallurgy,* Elsevier, Amsterdam-London-New York 1971, pp. 145 – 173.
44. N. P. Lyakishev, N. A. Tulin, Yu. L. Pliner: *Niobium in Steel and Alloys,* Companhia Brasileira de Metalurgia e Mineracao-CBMM, Sao Paulo 1984, pp. 87 – 101.
45. N. Krishnamurthy, R. Venkataramani, S. P. Garg, *Int. J. Refract. Hard Met.* **4** (1984) no. 1, 41 – 45.
46. C. K. Gupta, D. K. Bose, N. Krishnamurthy, *J. Less Common Met.* **139** (1988) 189 – 202.

47. N. P. Lyakishev, N. A. Tulin, Yu. L. Pliner: *Niobium in Steel and Alloys,* Companhia Brasileira de Metalurgia e Mineracao-CBMM, Sao Paulo 1984, pp. 101 –114.
48. H. A. Wilhelm, F. A. Schmidt, T. G. Ellis, *J. Met.* **18** (1966) 1303 – 1308.
49. G. R. Kamat, C. K. Gupta, *Metall. Trans.* **2** (1971) 2817 – 2823.
50. Rio Algam Mines, GB 1 054 163, 1967.
51. Du Pont de Nemours and Comp., GB 863 310, 1957.
52. R. M. Haire, E. Hade, D. M. Hiller, H. W. Jacobsen in R. M. Kibby (ed.): *The Design of Metal Producing Processes,* AIME, New York 1969, pp. 300 – 308.
53. T. T. Campell, F. E. Block, G. B. Robidart, J. L. Schaller, Bureau of Mines Report of Investigations 6080, 1962.
54. T. Sugai, R. Watanabe: "Extraction of Niobium Metal by Chlorine Process," *Proceedings of International Symposium on Tantalum and Niobium,* Tantalum-Niobium International Study Center, Brussels, Belgium 1988, pp. 241 – 254.
55. G. Jangg, R. Kiefer, P. Topic, *Monatsh. Chem.* **100** (1969) 379 – 384.
56. W. Rockenbauer, *Chem. Ing.-Tech.* **41** (1969) no. 4, 159 – 162.
57. K. K. Schulze, H. A. Jehn, G. Hörz, *J. Met.* **40** (1988) no. 10, 25 – 31.
58. W. W. Albrecht, D. P. Ingals: *Raffinationsverfahren in der Metallurgie,* VCH Verlagsgesellschaft, Weinheim 1983, pp. 230 – 241.
59. K. Schulze, O. Bach, D. Lupton, F. Schreiber "Purification of Niobium" in H. Stuart (ed.): *Niobium Proceedings of the International Symposium,* New York 1984, pp. 163 – 223.
60. M. P. Schlienger, R. C. Eschenbach: "Recent Developments in Plasma Melting of Reactive and Refractory Metals" in R. Bakish (ed.): *Proceedings of the Conference Electron Beam Melting and Refining,* New York 1987, pp. 149 – 156.
61. K. Mimura, M. Nanjo, *High Temp. Mater. Processes* **8** (1988) no. 1, 29 – 38.
62. E. Drost, M. Hörmann, *Metall (Berlin)* **41** (1987) no. 4, 382 – 385.
63. M. Hörmann, *J. Less Common Met.* **139** (1988) 1 – 14.
64. K. K. Schulze, *J. Met.* **33** (1981) no. 5, 33 – 41.
65. D. Lupton, F. Aldinger, K. Schulze: "Niobium in Corrosive Environments" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 533-560.
66. U. Gennari, E. Kny, T. Gartner: "Niobium-Titanium Oxide-Alloy," *Proceedings of the International Plansee Seminar 89,* vol. 3, 1989, pp. 587 – 614.
67. Tantalum Niobium International Study Center Brussels, Belgium: TiC Bulletin no. 46 (1986) 6 – 7.
68. H. Inouye: "Niobium in High Temperature Applications" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* 1984, pp. 615 – 636.
69. J. Lambert, "Aerospace brightens for Columbium, Tantalum + Columbium," Supplement to American Metal Market, New York 1988.
70. R. Durrer, G. Volkert in G. Volkert, K.-D. Frank (eds.): *Metallurgie der Ferrolegierungen,* Springer Verlag, Berlin-Göttingen-Heidelberg 1972.
71. Vanadium Corp., US 2 909 427, 1958 (H. W. Rathman, J. O. Staggers, H. K. Bruner)
72. Gesellschaft für Elektrometallurgie mbH, Düsseldorf, Lieferprogramm 1989.
73. A. J. DeArdo, J. M. Gray, L. Meyer: "Fundamental Metallurgy of Niobium in Steel" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 685 – 759.
74. S. R. Keown, F. B. Pickering: "Niobium in Stainless Steels" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 1113 –1141.
75. C. T. Sims: "A Perspective of Niobium in Superalloys" in H. Stuart (ed.): *Niobium-Proceedings of the International Symposium,* New York 1984, pp. 1169 – 1220.
76. J. K. Tien, J. P. Collier: "Current Status and Future Developments of the Niobium and Tantalum Containing Superalloys," *Proceedings of International Symposium on Tantalum and Niobium* 1988, Tantalum-Niobium International Study Center, Brussels, Belgium, pp. 657 – 679.
77. R. Fresenius, G. Jander: *Handbuch der analytischen Chemie,* **Part 3, "Quantitative Bestimmungs- und Trennungsmethoden," vol. 5b, "Elemente der 5. Gruppe. Vanadin, Niob, Tantal",** Springer Verlag, Berlin 1957.
78. O. G. Koch, G. A. Koch-Dedic: *Handbuch der Spurenanalyse,* Part 2, 2nd ed., Springer Verlag, 1974.
79. R. W. Moskier: *Analytical Chemistry of Niobium and Tantalum,* Pergamon Press, New York 1964.

80. I. M. Gibaldo: *Analytical Chemistry of Niobium and Tantalum,* Ann Arbor, 1970.
81. In [79]
82. A. Knight: *TiC Bulletin,* no. 45, Tantalum-Niobium International Study Center, Brussels, Belgium, 1964.
83. A. M. E. Balaes, K. Dixon, G. J. Wall, *Appl. Spectros.* **41** (1987) no. 3, 509 – 512.
84. *Analyse der Metalle,* "First Supplementary Volume: I Schiedsanalyse, II Betriebsanalyse," Springer Verlag, Berlin 1980.
85. *Annual Book of ASTM Standards,* "Metal Test Methods and Analytical Proceedures," vol. 03.05, "Chemical Analysis of Metals and Metal Bearing Ores," Easton 1987.
86. *Handbuch für das Eisenhüttenlaboratorium,* vol. 5, Verlag Stahleisen GmbH, Düsseldorf 1973.
87. H. M. Ortner, P. Wilhortitz in Tantalum-Niobium International Study Center, Brussels (ed.): *TiC Bulletin,* no. 49 (1987).
88. D. Hirschfeld (ed.): "Gase in Metallen," *Vortragstexte der Tagung Gase in Metallen,* Deutsche Gesellschaft für Metallkunde, Darmstadt 1984.
89. R. Caletka, V. Krivan, *Fresenius Z. Anal. Chem.* **311** (1982) 177 – 182.
90. H. N. Hecker, *Fresenius Z. Anal. Chem.* **293** (1978) 110 – 114.
91. N. E. Sanderson et al., *Microchim. Acta (Wien),* (1987) 275 – 290.
92. R. Caletka, W. G. Faix, V. Krivan, *J. Radioanal. Chem.* **72** (1982) no. 1 – 2, 109 – 130.
93. W. G. Faix, R. Caletka, V. Krivan, *Anal. Chem.* **53** (1981) 1594 – 1598.
94. V. Krivan, *Anal. Chem.* **47** (1975) no. 3, 469 – 478.
95. H. Stuart, F. Heisterkamp: "The Niobium Market – A Perspective," *Proceedings of International Symposium on Tantalum and Niobium,* Tantalum-Niobium International Study Center, Brussels, Belgium, 1988, pp. 71 – 89.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Nitrates and Nitrites

Standard Article

Wolfgang Laue[1], Michael Thiemann[2], Erich Scheibler[3], Karl Wilhelm Wiegand[4]

[1]Gewerkschaft Victor Chemische Werke (now BASF), Castrop – Rauxel, Federal Republic of Germany
[2]Uhde GmbH, Dortmund, Federal Republic of Germany
[3]Uhde GmbH, Dortmund, Federal Republic of Germany
[4]Uhde GmbH, Dortmund, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a17_265
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Sodium Nitrate
1.1. Properties
1.2. Occurrence
1.3. Production
1.3.1. Chile Saltpeter
1.3.2. Synthetic Sodium Nitrate
1.4. Product Forms, Storage, and Transportation
1.5. Uses
1.6. Economic Aspects
2. Potassium Nitrate
2.1. Properties
2.2. Occurrence
2.3. Production
2.3.1. Bacterial Production of Saltpeter
2.3.2. Converted Saltpeter
2.3.3. Potassium Nitrate from Calcium Nitrate
2.3.4. Potassium Nitrate from Ammonium Nitrate
2.3.5. Potassium Nitrate from Potassium Chloride and Nitric Acid
2.4. Product Forms, Storage, and Transportation
2.5. Uses
2.6. Economic Aspects
3. Calcium Nitrate
3.1. Properties
3.2. Occurrence
3.3. Production
3.3.1. Calcium Nitrate from Limestoneand Nitric Acid
3.3.2. Calcium Nitrate as a Byproduct of the Odda Process
3.4. Product Forms, Storage, and Transportation
3.5. Uses
3.6. Economic Aspects
4. Sodium Nitrite
4.1. Properties
4.2. Production
4.3. Product Forms, Storage, and Transportation
4.4. Uses

4.5. Economic Aspects
5. Potassium Nitrite
5.1. Properties
5.2. Production
5.3. Uses
6. Ammonium Nitrite
6.1. Properties
6.2. Production
6.3. Uses
7. Calcium Nitrite
7.1. Properties
7.2. Production
7.3. Uses
8. Analysis of Nitrates and Nitrites
9. Environmental Aspects
10. Toxicology and Occupational Health

[Top of Page]

1. Sodium Nitrate

Sodium nitrate in the form of Chile saltpeter was the most important inorganic nitrogen fertilizer until about 1920; it is the only abundantly occurring mineral nitrate.

1.1. Properties [1]

Sodium nitrate [7631-99-4], Chile saltpeter, Chilean nitrate, $NaNO_3$, M_r 85.00, has an enthalpy of formation ΔH = –466.8 kJ/mol [2], an enthalpy of fusion ΔH = + 15.7 kJ/mol [3] and *m p* 306.8 °C [3]. For pressure dependence of the melting point see [4]. At 25 °C, anhydrous sodium nitrate forms colorless trigonal crystals of space group D_{3d}^{6}; ϱ_{298} = 2.261 g/cm^3 [3]; $C_{p,\,298}$ = 93.1 J mol^{-1} K^{-1} [2]; refractive index at 589 nm n_{ω} = 1.587, n_{ϵ} = 1.336 [3]. This modification forms trigonal crystals of space group D_{3d}^{5} at a transition point of ca. 275 °C.

Sodium nitrate does not form hydrated solid phases (Fig. 1). The salt is hygroscopic and readily dissolves in water; at 25 °C, 92.1 g dissolves in 100 g of water [3]; enthalpy of solution at infinite dilution ΔH_{298} = + 20.5 kJ/mol [3]. Sodium nitrate is soluble in liquid ammonia and forms $NaNO_3 \cdot 4NH_3$ below –42 °C. The solubility of sodium nitrate in anhydrous methanol is 2.8 wt % at 25 °C [5].

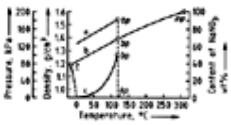

Figure 1. Properties of a saturated solution of sodium nitrate in water

a) Density; b) Content of sodium nitrate; c) Vapor pressure *bp* = boiling point of saturated solution containing 68.5 wt % sodium nitrate at 101.3 kPa and 119 °C; E = cryohydric point: –17.5 ± 0.5 °C, 38.0 % sodium nitrate.

Sodium nitrate is an oxidant and releases oxygen on heating:

$$NaNO_3 \rightleftharpoons NaNO_2 + 1/2\,O_2\ \Delta H_{298} = +109\,\mathrm{kJ/mol} \qquad (1.1)$$

Molten sodium nitrate remains stable in air up to 500 °C; reaction (1.1) reaches equilibrium with the evolution of oxygen between 600 and 750 °C [6]. The product (sodium nitrite) decomposes (see Section Properties). Alkali nitrates contain small amounts of nitrites at temperatures as low as 300 °C [7]. In aqueous solution, alkali nitrates have little oxidizing power; they are reduced to ammonia by nascent hydrogen.

1.2. Occurrence

Deposits of sodium nitrate (the mineral nitratine) occur chiefly in arid regions such as Peru, Mexico, Egypt, and South Africa. The only commercial deposits are, however, in Chile. The origin of these deposits has not been determined conclusively. A number of theories have been proposed since the early 1960s [8-10].

The Chilean deposits lie between latitudes 19 ° and 26 ° at ca. 800 – 1600 m above sea level. The uppermost bed is 20 – 40 cm thick and consists of unconsolidated weathering products containing only a little salt (“chuca”). The next bed, up to 3 m thick, is called “costra”; it is usually a coarse breccia cemented with clay and salts of the saltpeter deposits. The “caliche” or salt bed proper lies beneath and has a thickness of 0.5 – 5.0 m. It is also brecciated and contains various salts (Table 1) and widely varying amounts of clay, sand, and gravel.

Table 1. Analyses of caliche samples (composition in wt %)

Salt	1	2	3	4	5
$NaNO_3$	34.2	34.4	43.3	28.5	53.5
KNO_3	1.6				17.3
Na_2SO_4	8.4	1.6	25.3	5.4	1.9
$CaSO_4$	6.3	1.6	30.9	2.7	0.5
$MgSO_4$	2.0	5.4		3.4	1.4
NaCl	32.0	4.0		17.2	21.3

In addition to nitratine ($NaNO_3$) and generally a little niter (KNO_3), the rock often contains a wide variety of components: halite (NaCl), gypsum ($CaSO_4 \cdot 2H_2O$), anhydrite ($CaSO_4$), bloedite ($Na_2SO_4 \cdot MgSO_4 \cdot 4H_2O$), glauberite ($Na_2SO_4 \cdot CaSO_4$), polyhalite ($K_2SO_4 \cdot MgSO_4 \cdot 2CaSO_4 \cdot 2H_2O$), and darapskite ($NaNO_3 \cdot Na_2SO_4 \cdot H_2O$). In addition to the salts listed in Table 1, borates, iodates, other iodine compounds, and perchlorates may be present in amounts up to a few tenths of a percent.

Reserves may well support several more centuries of working. In Chile alone, reserves of commercial caliche containing 7 wt % sodium nitrate are estimated at 2.5×10^9 t. Reserves of lower quality are about ten times greater [11].

1.3. Production

1.3.1. *Chile Saltpeter*

The winning of sodium nitrate from caliche is very complicated. The raw material contains only 7 – 40 wt % sodium nitrate (corresponding to 2 – 7 % N) and requires multistage processing to achieve an acceptable yield. The saltpeter industry in Chile produces ca. 600 000 t of sodium nitrate, 50 000 t of sodium sulfate, and more than 3000 t of iodine yearly [12]; it is one of the world's leading producers of iodine (⟶ , ⟶ Iodine and Iodine Compounds – Economic Aspects).

As early as 1810, low yields of impure saltpeter were produced by small boileries in Chile. The first improvement was the *Shanks hot-leaching process*, introduced in 1876 [13], [14]. Caliche with a minimum sodium nitrate content of 13 wt % was extracted with mother liquor at ca. 110 °C; after crystallization of sodium chloride at < 5 °C, the product was obtained by crystallization at 22 °C. The yield was 65 – 80 %, and the energy cost was high. The 200 or so plants employing the Shanks process ranged in capacity from 10 000 to 100 000 t/a of sodium nitrate. High energy costs made the Shanks process uncompetitive with synthetic saltpeter after World War I.

The *Guggenheim low-temperature leaching process* was developed in 1923 [8], [15]. By the beginning of the 1930s, two Guggenheim plants with sodium nitrate capacities of 500 000 and 600 000 t/a had been built. The Guggenheim process features relatively low leaching temperatures (40 – 45 °C), 90 % yields, low energy consumption due to good heat utilization, rational methods of caliche mining, and very large plant units. The process can economically handle caliche containing as little as 7 wt % sodium nitrate. Old spoil heaps of spent earth can be processed along with residues from leaching the coarse caliche fraction containing > 7 wt % sodium nitrate.

The isolation of sodium nitrate is based on differences between the solubility curves of the salts contained in the caliche. For example, the solubility of sodium nitrate is four times greater than the solubility of sodium chloride in water at 45 °C (Fig. 2). Operating conditions must be avoided under which insoluble or sparingly soluble double salts containing sodium nitrate are stable. For example, if the Mg^{2+} and Ca^{2+} concentration in the recycle liquor is sufficiently high, darapskite ($NaNO_3 \cdot Na_2SO_4 \cdot H_2O$) forms the insoluble “protective compounds” blödite ($Na_2SO_4 \cdot MgSO_4 \cdot 4\,H_2O$) and glauberite ($Na_2SO_4 \cdot CaSO_4$). Darapskite is a sparingly soluble component that accounts for 30 % of the sodium nitrate in caliche and is stable below 55 °C.

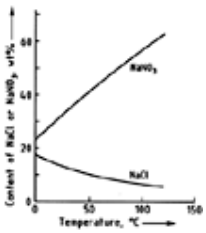

Figure 2. Solubility curves of sodium nitrate and sodium chloride in a saturated aqueous solution containing solid sodium nitrate and sodium chloride [16]

The production of Chile saltpeter from caliche is summarized in Figure 3. The warm (40 – 45 °C) mother liquor recycled to leaching contains ca. 330 g/L of sodium nitrate. The leaching process is performed in a series of vats, each having a capacity of ca. 1000 t of crushed ore. The underflow of each vat is heated before passing to the next one since the dissolution of sodium nitrate is endothermic. Sodium nitrate therefore dissolves preferentially. The product solution obtained after leaching contains 450 g/L of sodium nitrate and is cooled to 5 – 10 °C in heat exchangers and with ammonia refrigerating equipment. The mother liquor obtained after centrifugation of crystallized sodium nitrate is heated to 40 – 45 °C with heat from the condensers in the ammonia chilling loop and then with cooling water from the diesel engines used to generate electric power. Crystalline sodium nitrate obtained after centrifugation contains 3.0 – 3.5 % moisture; it is generally melted at ca. 340 °C and

prilled. The granulated product contains only about 1 % sodium chloride. Table 2 gives a detailed product analysis [17]. Large modern plants also produce coarse crystals of Chile saltpeter, which do not require further processing by melting.

Table 2. Analysis of saltpeter products

Component	Granulated sodium nitrate, * wt %	Granulated potassium nitrate,* wt %	Synthetic saltpeter, wt %
$NaNO_3$	98.32	67.37	99.70
KNO_3		30.36	0.06
NaCl	0.67	0.97	0.09
$Na_2B_4O_7$	0.24	0.23	
$NaIO_3$	0.03	0.03	
Moisture	0.17	0.21	0.10
Undetermined	0.57	0.83	0.05

* From Chile saltpeter [17].

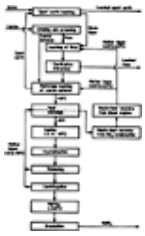

Figure 3. Guggenheim process for the production of Chile saltpeter

To recover more than 80 % of the sodium nitrate present in caliche, the “spent earth” containing 1.0 – 1.5 wt % sodium nitrate must be leached with water. The resulting dilute liquor is either used for leaching fine caliche (Fig. 3) or placed in large open pans and evaporated by the sun and the hot dry winds of the salt desert; precipitated crystals are removed periodically. Greater utilization of solar energy in the future will allow economical processing of caliche containing as little as 4 % sodium nitrate.

Byproducts. Iodine recovery from the mother liquor begins when the sodium iodate content reaches 6 – 9 g/L. The enriched solution is treated with sodium hydrogen sulfite to precipitate free iodine:

$$2\,NaIO_3 + 5\,NaHSO_3 \rightarrow I_2 + 3\,NaHSO_4 + 2\,Na_2SO_4 + H_2O \quad (1.2)$$

Iodine is separated by filtration and purified by sublimation. The side stream after iodine recovery is returned to the circulating mother liquor.

The recovery of Chile niter (KNO_3) as a byproduct is profitable if the caliche contains > 2 – 3 wt % potassium nitrate. This product also accumulates in the mother liquor. To reduce processing costs, the product is refined to a potassium nitrate content of only 30 % (Table 2).

1.3.2. *Synthetic Sodium Nitrate*

The most important method for producing synthetic sodium nitrate is the reaction of tail gases from nitric acid plants with sodium hydroxide or sodium carbonate solution. This method is, however, intended for tail-gas clean-up (removal of nitrogen oxides) rather than for nitrate and nitrite production.

Tail-gas absorption initially yields some sodium nitrate and larger amounts of sodium nitrite. The reactions that produce nitrite alone

$$2\,NaOH + NO_2 + NO \rightarrow 2\,NaNO_2 + H_2O \quad (1.3)$$

$$Na_2CO_3 + NO_2 + NO \rightarrow 2\,NaNO_2 + CO_2 \quad (1.4)$$

proceed faster than the reactions that produce both nitrite and nitrate.

$$2\,NaOH + 2\,NO_2 \rightarrow NaNO_3 + NaNO_2 + H_2O \quad (1.5)$$

$$Na_2CO_3 + 2\,NO_2 \rightarrow NaNO_3 + NaNO_2 + CO_2 \quad (1.6)$$

The countercurrent absorption process takes place in 18–8 stainless steel columns connected in series. In sodium carbonate absorption, the columns are packed with ceramic rings; iron rings are used for sodium hydroxide absorption.

Fresh alkaline solution is circulated into the last column. The liquor is forwarded batchwise to successive columns and finally withdrawn with a content of ca. 500 g/L of sodium nitrate and sodium nitrite. The product liquor contains both nitrate and nitrite, even if nitrite removal is employed (see Chap. Sodium Nitrite). Nitrite is decomposed by inversion with nitric acid at 50 °C in a column made of 18–8 steel prior to nitrate recovery:

$$3\,NaNO_2 + 2\,HNO_3 \rightarrow 3\,NaNO_3 + 2\,NO + H_2O \qquad (1.7)$$

The highly concentrated nitrogen monoxide can be used to synthesize other compounds (e.g., hydroxylamine) or can be returned to the nitric acid plant.

The residual nitrogen monoxide is stripped from the inverted solution with steam or air at ca. 100 °C. The solution is then neutralized with sodium carbonate or sodium hydroxide, passed through a precoat filter, and concentrated batchwise in multistage iron evaporators. The first stage is operated at 100 – 120 °C (p = 120 – 150 kPa in the vapor space). The resulting solution should not crystallize and should still dissolve any salt deposits formed in the second stage. The second stage is operated under vacuum (p = 10 – 20 kPa) at 65 – 75 °C; the product is concentrated to form a slurry containing ca. 20 % precipitated sodium nitrate. The crystals are separated in a centrifuge without cooling. Depending on the quality required, they are washed more or less intensively with cold water. Sodium nitrate from the centrifuge has a moisture content of 2 – 3 wt % and is dried to ca. 0.1 wt % moisture in drums heated with flue gas. The product is dried, cooled, and stored in bins.

The mother liquor from the centrifuge is recycled to the first evaporator stage along with the wash water. Discharge of the liquor is only necessary if the alkali used contains more chloride than permitted in sodium nitrate by quality standards. Accordingly, the process generally does not produce any wastewater.

1.4. Product Forms, Storage, and Transportation

Chile saltpeter is marketed as both a “granular” product (mainly used as a fertilizer) and a coarse crystalline product. Technical-grade synthetic sodium nitrate is a fine, crystalline white powder with a bulk density of about 1.36 kg/L. It is sold as an untreated product and also as a free-flowing product containing up to 0.1 % alkyl aryl sulfonate as an anticaking agent [18]. Table 2 presents analyses of sodium nitrate.

Sodium nitrate can be transported in bulk form. The technical grade, especially the untreated product, is generally shipped in polyethylene bags. As an oxidizing agent, sodium nitrate must not come in contact with readily oxidizable products during transport and storage.

1.5. Uses

Inorganic nitrates and nitrites overlap in their areas of application. Mixtures of the compounds are sometimes employed. More than half of the sodium nitrate produced worldwide (including Chile saltpeter) is used as a fertilizer for crops such as cotton, tobacco, and vegetables. In contrast to other nitrogen fertilizers, sodium nitrate does not overacidify the soil. In Europe and the United States, sodium nitrate is of minor importance compared to other fertilizers.

The major industrial use of sodium nitrate is in the explosives industry. It is employed along with ammonium nitrate in the production of water-containing slurry and gel explosives, whose consistency is stabilized with rubber additives. Sodium nitrate, like potassium nitrate, is added to ammonium nitrate melts (see Section Uses). Sodium nitrate is also used in pyrotechnics, for example, in flare mixtures.

Large amounts of sodium nitrate are used in the glass and enamel industry as a refining agent for removing air bubbles from melts.

In metallurgy, molten salt mixtures containing sodium nitrate are employed as heat-transfer baths for quench hardening and tempering of steel, light alloys, and copper alloys. For the range 260 – 550 °C, $NaNO_3$ – KNO_3 melts are used; melts containing ca. 45 % sodium nitrite along with potassium nitrate and sodium nitrate can be used from 150 to 300 °C [19]. The heat of fusion of the salt mixture can be utilized to keep the temperature constant; the solidification point of the ternary system KNO_3 – $NaNO_3$ – $NaNO_2$ can be adjusted precisely by altering the proportion of components. Metal parts with oxide scales can be pretreated by briefly dipping them in molten $NaNO_3$ – NaOH so that they can be electrically descaled in a sodium sulfate solution without the use of acid (more environmentally friendly); similar methods employ potassium nitrate.

Molten mixtures of sodium nitrate, other alkali nitrates, and nitrites function as heat-transfer and heat-storage media in solar technology. The maximum temperature for this application is ca. 600 °C. The thermophysical properties of such melts are described in [20-23].

Mixtures of sodium nitrate and borax also serve as an auxiliary in soldering and welding. Sodium nitrate is recommended as a corrosion-inhibiting additive in antifreezes and cooling brines. In the regeneration of spent sulfuric acid by distillation in cast-iron vessels, passivation with sodium nitrate is used to prevent corrosion.

Sodium nitrate is used to promote combustion; it is added to activated carbon when sulfur dioxide is oxidized to sulfur trioxide in exhaust-gas combustion devices and is also used to improve the combustibility of tobacco. Cleaning agents for plugged drain pipes are made from sodium nitrate, aluminum filings, and an alkali; hydrogen produced by the action of water is oxidized by sodium nitrate, and this reaction provides the heat needed for effective cleaning. Molten sodium nitrate serves as a reaction medium, for example, in the electrolytic production of chromium dioxide for magnetic tape. A melt of sodium nitrate and potassium nitrate provides a suitable medium for the thermal decomposition of dissolved ammonium nitrate to dinitrogen monoxide. The same melt can be used for the conversion of urea to cyanuric acid in a method devised by Stamicarbon. Pickling salt contains sodium nitrate.

Other applications of sodium nitrate include the production of dyes, pharmaceuticals, charcoal briquetts, and other nitrates.

1.6. Economic Aspects

The output of Chile saltpeter has decreased continually since the late 1920s (Table 3). Despite competition from synthetic sodium nitrate, output seems to have reached a stable level. In Chile, production capacity increased in 1984 [24], and another plant is planned [25]. The cost of producing Chile saltpeter is far less sensitive to rising energy prices than that of synthetic sodium nitrate, which requires ammonia and caustic soda as feedstocks. Most Chile saltpeter is exported; the most important consumers are Western Europe (120 000 t/a), especially Spain and The Netherlands, and the United States (100 000 t/a) [17].

Table 3. Production of Chile saltpeter (10^3 t)

Year	Production	Year	Production
1830	1	1940	1380
1850	30	1950	1660
1880	220	1960	930
1900	1650	1970	700
1910	2470	1975	800
1920	2530	1979	650
1928	3160	1984	710
1933	530	1987	590

Few production figures are available for synthetic sodium nitrate. Sodium nitrate has not been manufactured in the United States since 1988. Further stagnation of sodium nitrate production is expected in the near future. This forecast could change if solar energy development is stepped up [26].

[Top of Page]

2. Potassium Nitrate

Potassium nitrate was manufactured in China with the aid of bacteria as early as the Middle Ages. Up to the beginning of the 20th century, its only important use was in the production of gunpowder.

2.1. Properties [27]

Potassium nitrate [7757-79-1], saltpeter, KNO_3, M_r 101.10, has an enthalpy of formation ΔH_{298} = –492.8 kJ/mol [2], an enthalpy of fusion ΔH = +11.9 kJ/mol [3], and *mp* 334 °C; for pressure dependence of the melting point, see [4]. At 25 °C, anhydrous potassium nitrate forms colorless rhombic crystals of space group $D_2^{16}{}_h$; ϱ_{298} = 2.109 g/cm^3; $C_{p,\,298}$ = 96.36 J mol^{-1} K^{-1} [2]; refractive index at 589 nm n_α = 1.335, n_β = 1.5056, n_γ = 1.5064 [2]. This modification forms rhombic crystals of space group D_{3d}^5 at a transition point of ca. 127.7 °C (enthalpy of formation ΔH = +5.1 kJ/mol) [28], [29]. High-pressure allotropes account for the caking tendency of potassium nitrate [30].

Potassium nitrate does not form solid phases containing water (Fig. 4). Its enthalpy of solution in water at infinite dilution is ΔH_{298} = +34.9 kJ/mol [3]. Potassium nitrate is soluble in liquid ammonia; no ammoniates are known. The solubility of potassium nitrate in anhydrous methanol at 25 °C is 0.34 wt % [5].

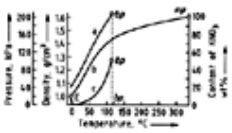

Figure 4. Properties of a saturated solution of potassium nitrate in water

a) Density; b) Content of potassium nitrate; c) Vapor pressure

bp = boiling point of saturated solution containing 77.0 wt % potassium nitrate at 101.3 kPa and 115.6 °C; E = cryohydric point: –2.85 °C, 9.66 % potassium nitrate.

Like sodium nitrate, potassium nitrate liberates oxygen on heating. The endothermic reaction

$$KNO_3 \rightleftharpoons KNO_2 + 1/2\,O_2 \quad \Delta H_{298} = +124\,\mathrm{kJ/mol} \tag{2.1}$$

is reversible. In contact with air, pure molten potassium nitrate is stable up to about 530 °C; reaction (2.1) comes to equilibrium between 650 and 750 °C. Above 750 °C, the nitrite product decomposes with the formation of nitrogen oxides [6]. The KNO_3 – $NaNO_3$ system forms a eutectic with 45 % sodium nitrate, *mp* ca. 218 °C. Solid solutions exist above 175 °C with no miscibility gap; at room temperature the system has a miscibility gap from 0.5 to 99.9 wt % sodium nitrate. The composition of the aqueous solution saturated with both salts is described in [31].

2.2. Occurrence

Potassium nitrate (in the form of the mineral niter) occurs as efflorescences on soils. Deposits in China and the East Indies were economically important; in the first half of the 19th century, more than 10 000 t/a of potassium nitrate were produced by

leaching such soils (current annual production is still several thousand tonnes). Before the large-scale manufacture of converted saltpeter began, East Indian saltpeter enjoyed a virtual monopoly.

Potassium nitrate is present in tobacco plants, from which it can be extracted with water and recovered to diminish nitrogen oxide evolution during smoking.

2.3. Production

For older processes, see [14], [27].

2.3.1. *Bacterial Production of Saltpeter*

Bacterial production of potassium nitrate was employed from the 17th to the 19th century in Europe. This was necessary to avoid having to import this raw material, which was indispensable for military purposes during wartime. Nitrogen-rich organic wastes, especially animal feces and urine, were placed in loose, air-permeable piles of earth together with lime and potash; the nitrogen compounds were converted to nitrates by bacterial nitrification. After several years, the mass was leached with water. Potash was added to the crude liquor to convert calcium and magnesium nitrates to potassium nitrate and sparingly soluble calcium and magnesium carbonates. The filtered liquor was evaporated to obtain saltpeter, which was purified by recrystallization.

2.3.2. *Converted Saltpeter*

From the mid-19th century until the 1950s, the conversion of Chile saltpeter with potassium chloride was the most important potassium nitrate production process (Fig. 5). Potassium chloride was supplied in large quantities after 1860 by the German potash industry.

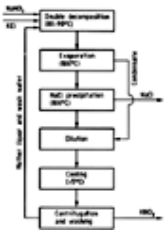

Figure 5. Conversion of sodium nitrate to potassium nitrate

$$NaNO_3 + KCl \rightleftharpoons KNO_3 + NaCl \tag{2.2}$$

The process is based on the slight increase in solubility of sodium chloride with temperature. The mother liquor from potassium nitrate crystallization in the preceding cycle (point C in Fig. 6) is heated and reacted with stoichiometric quantities of sodium nitrate and potassium chloride. Impure starting materials are generally used in this step. The reaction mixture (point A, Fig. 6) is concentrated at 100 °C with the addition of sodium carbonate; sodium chloride and impurities (e.g., $MgCO_3$) precipitate and are filtered out. The filtrate (point B, Fig. 6) is again diluted with the evaporation condensate (to prevent sodium salts from precipitating on cooling), clarified, cooled to 5 °C to precipitate potassium nitrate, and centrifuged. The mother liquor (point C, Fig. 6) is recycled. The potassium nitrate product is recrystallized for technical purposes. Because of high energy costs and other factors, the conversion process can no longer compete with the processes described in Sections Potassium Nitrate from Calcium Nitrate, Potassium Nitrate from Ammonium Nitrate, Potassium Nitrate from Potassium Chloride and Nitric Acid; nevertheless, older plants continue to use it. Physical and chemical fundamentals are described in [32-35]; process design may be found in [36], [37].

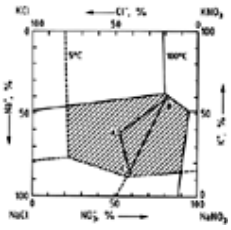

Figure 6. The $NaNO_3$ – KNO_3 – KCl – NaCl – H_2O system

Jänecke's method [34] is used to plot the relative content of cations and anions in percent equivalents in the salt mixture corresponding to the saturated solution. Absolute concentrations or quantities of water are not shown. Values for 5 °C are according to [32], for 100 °C according to [33].

The lines labeled 5 and 100 °C represent the boundaries of fields in each of which one of the four salts occurs as the solid phase in the saturated solution. In the shaded region, sodium chloride precipitates from the saturated solution at 100 °C but potassium nitrate precipitates at 5 °C. (See text for explanation of the working triangle ABC.)

2.3.3. *Potassium Nitrate from Calcium Nitrate*

A number of processes have been developed for obtaining potassium nitrate from calcium nitrate, which is a byproduct of other processes (e.g., digestion of rock phosphate with nitric acid).

With potassium sulfate as a starting product, the reaction

$$Ca(NO_3)_2 + K_2SO_4 + 2\,H_2O \rightarrow 2\,KNO_3 + CaSO_4 \cdot 2\,H_2O \tag{2.3}$$

yields sparingly soluble gypsum, which is easily separated. This is the basis for the relatively simple Victor process [38], which was employed after 1951 in Germany for several decades to manufacture potassium nitrate [39]. Other methods employ potassium chloride as raw material but have no industrial importance [39].

In another proposed process, limestone is first reacted with nitric acid to form calcium nitrate. Potassium sulfate is then added to the resulting liquor at ca. 100 °C. The gypsum formed is removed and the reaction mixture is cooled. Potassium nitrate crystals are isolated with a purity of 99.5 %, and the mother liquor is recycled to the beginning of the process [40].

A process based on the use of calcium nitrate and ion exchangers has also been proposed [41].

2.3.4. *Potassium Nitrate from Ammonium Nitrate*

Potassium nitrate is made from ammonium nitrate (available in high purity) and potassium chloride by double decomposition in aqueous solution:

$$NH_4NO_3 + KCl \rightarrow KNO_3 + NH_4Cl \tag{2.4}$$

The ammonium chloride byproduct is easily separated.

The solubility relationships (Fig. 7) in the system of reciprocal salt pairs described in reaction (2.4) [42] lead to a simple reaction design (Fig. 8) [43]. The starting reagents are added in alternate stages. Ammonium chloride precipitates at a higher temperature and from a higher concentration than potassium nitrate. Evaporation is therefore necessary prior to ammonium chloride crystallization, the condensate being recycled before potassium nitrate crystallization. Because the salt solution is aggressive, molybdenum-alloyed carbon steel is used for process equipment. Fertilizer-grade product containing 95 wt % potassium nitrate and 5 wt % ammonium nitrate is manufactured by this process [24], [44].

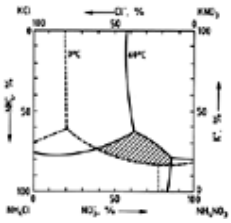

Figure 7. The NH_4NO_3 – KNO_3 – KCl – NH_4Cl – H_2O system (see Fig. 6 for the method of plotting)

In the shaded region, ammonium chloride precipitates from the saturated solution at 60 °C but potassium nitrate precipitates at 0 °C. Absolute concentrations are not shown (see [41]).

Figure 8. Potassium nitrate production from ammonium nitrate

2.3.5. *Potassium Nitrate from Potassium Chloride and Nitric Acid*

Production without Chlorine Formation. In an aqueous medium containing 50 % nitric acid, the reaction

$$HNO_3 + KCl \rightleftharpoons KNO_3 + HCl \tag{2.5}$$

proceeds without oxidation of chloride ion to chlorine and/or nitrosyl chloride only if the temperature is below 40 °C (or below 70 °C at 40 % HNO_3). A number of process design options are available: removing hydrogen chloride by absorption, decreasing potassium nitrate solubility by the addition of alcohol, or using cation exchangers. Only the Israel Mining Industry (IMI) process has been introduced on a large scale (Fig. 9) [45]. Solid potassium chloride is reacted with nitric acid at 5 – 10 ° C (c) in the presence of an organic solvent (*n*-butanol or isopentanol). The solvent extracts the hydrogen chloride, and potassium nitrate is precipitated quantitatively from the aqueous phase. The advantage of the low digestion temperature is that hydrogen chloride is extracted from the organic solvent in preference to nitric acid; furthermore, the aqueous solution contains less potassium nitrate. The three phases present (potassium nitrate, aqueous phase, organic phase) are then separated. The potassium nitrate is centrifuged and washed with water (e). The aqueous phase is returned to the reactors (c). The purity of the product (> 95 %) [46] is sufficient for use as fertilizer; recrystallization is necessary for the technical grade. Hydrogen chloride is separated as a 22 % aqueous solution from the alcohol (which is not miscible with water in all proportions) by countercurrent extraction with water in three mixer – settler units (f). After extraction, the alcohol, which contains nitric acid, is returned to the conversion reactor. Separation of hydrochloric acid from the solvent has been improved by use of a different solvent that extracts nitric acid preferentially and extracts virtually no hydrochloric acid [47]. Haifa Chemicals uses tributyl phosphate for this purpose. Another proposal for improving the process is to replace the alcohol by an organic acid and an amine [48]. Rigid poly(vinyl chloride) and titanium are used as construction materials, with graphite for distillation equipment. A substantial portion of investment is for solvent conditioning.

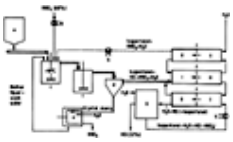

Figure 9. Simplified flow sheet of the IMI process (based on literature data)

a) Potassium chloride supply; b) Cooler; c) Reactor; d) Decanter; e) Centrifuge; f) Extractor (I = mixer, II = separator); g) Multistage evaporator (details not published)

Dashed lines indicate streams containing mainly isopentanol.

This process is employed in two potassium nitrate plants in Haifa. Potassium nitrate output in 1987 was 250 000 t, of which ca. 30 000 t was of technical purity and the remainder fertilizer grade [49]. The hydrochloric acid byproduct is utilized for the digestion of rock phosphate.

Production with Chlorine Formation. The South West Potash Corporation (SWP) process is employed on a large scale for

the production of potassium nitrate and chlorine [44], [50]. The most important steps follow (Fig. 10) [14]:

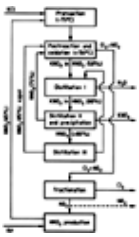

Figure 10. South West Potash Corporation process

1. Potassium chloride is reacted with 65 % nitric acid in an autoclave at 75 °C and ca. 200 kPa to yield potassium nitrate, chlorine, and nitrosyl chloride.

$$4\,HNO_3 + 3\,KCl \rightarrow 3\,KNO_3 + Cl_2 + NOCl + 2\,H_2O \quad (2.6)$$

2. Nitrosyl chloride is oxidized with 80 % nitric acid to yield chlorine and nitrogen dioxide.

$$NOCl + 2\,HNO_3 \rightarrow 1/2\,Cl_2 + 3\,NO_2 + H_2O \quad (2.7)$$

3. Nitrogen dioxide is oxidized with oxygen and absorbed in water to yield 65 % nitric acid, which is recycled to the potassium chloride conversion step.

$$4\,NO_2 + O_2 + 2\,H_2O \rightarrow 4\,HNO_3 \quad (2.8)$$

(Alternatively, nitrogen dioxide can be withdrawn as end product.)

4. The solution of potassium nitrate and nitric acid is concentrated to > 80 % nitric acid by crystallization of potassium nitrate in a vacuum crystallizer. Nitric acid can be obtained in 75 % or 85 % concentration because the azeotropic concentration in the HNO_3 – H_2O system at 80 kPa increases to 82 % nitric acid if the liquid phase contains 40 % potassium nitrate and to 87 % nitric acid at 50 % potassium nitrate [51].
5. The potassium nitrate is centrifuged, dried, melted at 340 – 350 °C, and prilled.

Intense corrosion dictates the use of titanium reactors with acid-resistant linings, whereas tower internals are made of glass or Teflon.

At present, Cedar Chemical Corp. (Trans-Resources) has an annual capacity of 100 000 – 120 000 t of potassium nitrate in an SWP plant at Vicksburg, Mississippi; ca. 20 000 t/a of the output is technical grade. Chlorine, dinitrogen tetroxide, and 65 % nitric acid are obtained as byproducts.

2.4. Product Forms, Storage, and Transportation

All technical-grade and most fertilizer-grade potassium nitrate is sold as a fine crystalline product with a bulk density of 1.1 kg/L. Technical grades may be untreated or treated with fatty amines or stearates to prevent caking [18]. Fertilizer-grade potassium nitrate is also sold as prills (bulk density 1.3 kg/L, obtained by atomizing a melt at 340 °C), because prills have better storage qualities and are easier to apply as fertilizer. Table 4 gives the analyses of typical grades. As a rule, the technical product is shipped in plastic or multi-ply paper bags. As an oxidant, potassium nitrate must not be stored or transported with easily oxidized products.

Table 4. Composition (wt %) of potassium nitrate

Component	Technical grade		Fertilizer grade	
	Requirements for black powder *	Sample	Product I	Product II
KNO_3	min. 99.0	99.8	95.0	96.5
Cl^-	max. 0.03	0.015	0.2	1.8
NO_2^-	not specified		0.002	0.2
SO_4^{2-}	not specified	0.02	3.4	0.05
Na^+	max. 0.1	0.02	0.2	0.7
Ca^{2+} and Mg^{2+}	max. 0.07	0.03	1.0	0.1
Heavy metals	max. 0.1		0.02	0.02
Insolubles	max. 0.1	0.05	0.05	0.2
H_2O	max. 0.25	0.06	0.1	0.1

* Not all requirements are shown.

2.5. Uses

About 75 % of potassium nitrate is manufactured with a purity of only 90 % for use as a nonhygroscopic fertilizer. The absence of chlorine is advantageous for growers of tobacco, citrus fruits, and vegetables (especially for hothouse crops). Potassium nitrate can also be used in the production of clear liquid fertilizers and is an important constituent of multinutrient fertilizers.

Almost half of the technical-grade product is employed in metallurgy, mainly in heat-transfer baths in mixtures with other nitrates and nitrites (see Section Uses). Potassium nitrate is also a component of soldering fluxes and welding electrodes. Large amounts are used in the glass, enamel, and ceramics industries. Glass devices and optical lenses can be improved by techniques based on Na^+ – K^+ ion exchange (treatment of sodium-containing glasses in molten KNO_3).

About 10 – 20 % of technical-grade potassium nitrate is consumed in the manufacture of explosives and pyrotechnics. Along with sodium nitrate and ammonium nitrate, potassium nitrate is an oxygen supplier in safety explosives. Because of its low hygroscopicity, potassium nitrate remains an important raw material for the production of matches and primer cords, as well as primer compositions with controlled burning rate for local heating. Another important use in the explosives industry is the addition of 2 – 10 % potassium nitrate to ammonium nitrate to lower the melting point, which facilitates prilling and improves the stability of the resulting granules.

Small amounts of potassium nitrate are employed in food preservation, cheese processing, and for improving the quality of tobacco.

Because of its oxidizing properties, potassium nitrate can be used for desooting in combustion processes. It is also recommended as an agent for controlling noxious odors during the application of sewage sludge or manure (at the same time, it improves the fertilizing action of the waste).

Potassium nitrate has been proposed as an oxidizing component in an acid-based gas generator system for the rapid inflation of air bags used to prevent injuries in motor vehicle accidents.

Potassium nitrate is also employed as an oxidant in chemical syntheses (e.g., alizarin synthesis).

2.6. Economic Aspects

Except for the small output from deposits still worked in India and the production of Chile saltpeter (which contains only 20 – 30 % KNO_3, see Section Chile Saltpeter), the only source of potassium nitrate for more than 100 years has been synthetic production. Although use of the technical-grade product can be expanded only slightly, a variety of processes have been developed to reduce the cost of manufacturing impure saltpeter and to increase its use as a fertilizer.

World production capacity in 1987 for synthetic potassium nitrate is about 425 000 t/a, 75 % of which is fertilizer grade and only 25 % technical grade (see Table 5). The largest producer is Trans-Resources with 340 000 t/a (Cedar Chemical and Haifa Chemicals, Israel) [52]. Israel is the largest exporter; its most important customers are the United States, Western Europe, and Latin America. By 1992, Haifa Chemicals will boost its capacity to 300 000 t/a [53]. The annual output of Chile saltpeter is 60 000 – 75 000 t of potassium nitrate. Soquimich, the sole manufacturer in Chile, is the third largest potassium nitrate producer in the world [54].

Table 5. Potassium nitrate capacities in 1987 (10^3 t)

Country	Fertilizer grade	Technical grade
Israel	250	30
United States	80	20
FRG	0	30
Other West European countries	20	20
Eastern bloc	*	30
India	*	*
Chile (natural saltpeter)	60 – 75	0

* Capacity unknown.

[Top of Page]

3. Calcium Nitrate

3.1. Properties [55]

Calcium nitrate [10124-37-5], $Ca(NO_3)_2$, M_r 164.09, has an enthalpy of formation of ΔH_{298} = –937.4 kJ/mol [2]. Anhydrous calcium nitrate crystallizes in the cubic system, space group T^6_h; ϱ_{298} = 2.504 g/cm^3 [3]; $C_{p,\ 298}$ = 149.5 J mol^{-1} K^{-1} [2]; refractive index n = 1.595 at 589 nm; $m\,p$ 561 °C (incipient decomposition).

Anhydrous calcium nitrate is highly hygroscopic and readily soluble in water when heated (enthalpy of solution ΔH_{298} = –18.7 kJ/mol at infinite dilution [56]); see also Figure 11 and Table 6. In addition to the anhydrous salt, three hydrates are known [57], [58]. The tetrahydrate [13477-34-4], $Ca(NO_3)_2 \cdot 4H_2O$, is the most important industrially: enthalpy of formation ΔH_{298} =+2130 kJ/mol; monoclinic crystals of space group C^5_{2h}; ϱ_{298} = 1.896 g/cm^3; refractive index at 589 nm n_α = 1.465, n_β = 1.498, n_γ = 1.504. The tetrahydrate dissolves in water with strong cooling (enthalpy of solution ΔH_{292} = + 32.7 kJ/mol at infinite dilution).

Table 6. Data for the system $Ca(NO_3)_2 - H_2O$ (see also Fig. 11)

Transition point and temperature, °C		Content of $Ca(NO_3)_2$, wt %	Range	Stable hydrate formula	*mp*, °C
E *	–28.7	42.9			
			E – A	$Ca(NO_3)_2 \cdot 4H_2O$	42.7
A	+42.6	70.8			
			A – B	$Ca(NO_3)_2 \cdot 3H_2O$	51.1
B	+50.6	77.2			
			B – C	$Ca(NO_3)_2 \cdot 2H_2O$	
C	+51.6	78.1			
			above C	$Ca(NO_3)_2$	561
bp **	+151	79.0			

Water vapor pressure over saturated solution with solid $Ca(NO_3)_2 \cdot 4H_2O$

Temperature, °C	0	10	20	30	40
Pressure, kPa	0.36	0.69	1.25	1.99	2.63

* E = cryohydric point.
** Boiling point (101.3 kPa) of the saturated solution.

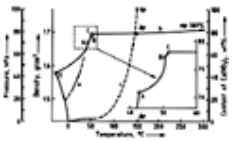

Figure 11. Properties of a saturated solution of calcium nitrate in water (see Table 6 for explanations)

a) Density; b) Content of calcium nitrate; c) Vapor pressure

Calcium nitrate dissolves readily in liquid ammonia, methanol, and ethanol. A number of ammoniates are known.

Thermal decomposition of calcium nitrate

$$Ca(NO_3)_2 \rightarrow CaO + 2\,NO_2 + 1/2\,O_2\ \Delta H = +369\ \mathrm{kJ/mol} \tag{3.1}$$

begins at 500 °C [6]. For the reaction mechanism, see [59].

Double Salts. Several weakly hygroscopic double salts are formed with ammonium nitrate [55]. The double salt $NH_4NO_3 \cdot 5Ca(NO_3)_2 \cdot 10H_2O$ has a high melting point (102 °C) and is important in the production of granular calcium nitrate. The double salt with urea, $Ca(NO_3)_2 \cdot 4CO(NH_2)_2$, is nonhygroscopic and can be employed as a fertilizer. The system $Ca(NO_3)_2 - KNO_3 - H_2O$ includes the phase $Ca(NO_3)_2 \cdot 4KNO_3$ [60].

3.2. Occurrence

Only small amounts of $Ca(NO_3)_2 \cdot 4H_2O$ occur (nitrocalcite) as efflorescences in limestone caves, for example, in Kentucky.

3.3. Production

Calcium nitrate became the first synthetic nitrogen fertilizer under the name Norge saltpeter. It was made by dissolving limestone in nitric acid and neutralizing the resulting liquor with powdered limestone or lime. This process is still in use, but ammonia is now employed for neutralization.

Calcium nitrate is also an important byproduct of nitric acid digestion of rock phosphates by the Odda process (

⟶ Phosphate Fertilizers). It is either used directly for fertilizer production or converted to ammonium nitrate.

A calcium nitrate solution is also obtained when tail gases from nitric acid plants are absorbed in milk of lime. The initial nitrate – nitrite solution is converted to pure calcium nitrate solution by a reaction analogous to reaction (1.7).

3.3.1. *Calcium Nitrate from Limestoneand Nitric Acid*

In this process, limestone is mixed with 50 –60 % nitric acid (Fig. 12). Calcium nitrate is formed by the reaction

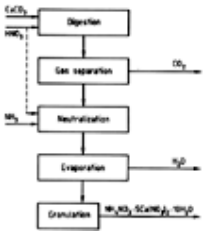

Figure 12. Production of calcium nitrate from limestone

$$CaCO_3 + 2\,HNO_3 \rightleftharpoons Ca(NO_3)_2 + H_2O + CO_2 \tag{3.2}$$

Water vapor and carbon dioxide are withdrawn from the process. Any remaining free nitric acid is then neutralized with ammonia to give ammonium nitrate; further ammonium nitrate can be added to adjust the content of this compound and to improve the storage qualities of the product. The calcium nitrate solution is then evaporated. Solid calcium nitrate can be subjected to any of the usual fertilizer processes (e.g., granulation, flocculation, or prilling).

The end product contains ca. 15.5 wt % nitrogen (0.9 % ammonium N, 14.6 % nitrate N). The main constituent (63 %) is the easily crystallized phase $NH_4NO_3 \cdot 5Ca(NO_3)_2 \cdot 10H_2O$; also present are 12 % $Ca(NO_3)_2 \cdot 2H_2O$ and 25 % calcium nitrate.

3.3.2. *Calcium Nitrate as a Byproduct of the Odda Process*

The Odda process is based on the digestion of rock phosphate with nitric acid to give phosphoric acid and calcium nitrate (Fig. 13) [61], [62]:

Figure 13. The Odda process

$$Ca_3(PO_4)_2 + 6\,HNO_3 \rightarrow 2\,H_3PO_4 + 3\,Ca(NO_3)_2 \tag{3.3}$$

The Ca^{2+} in the digester liquor is treated as a ballast. The CaO : P_2O_5 ratio serves as a measure of the ballast content and is decreased from 1.5 to 0.3 by removing the sand, with subsequent crystallization and removal of calcium nitrate. The remaining liquor (nitrophosphoric acid) is neutralized with ammonia, evaporated, granulated, classified, and conditioned (coated).

Digestion of Rock Phosphate. The digestion of rock phosphate with nitric acid is an exothermic reaction and proceeds as follows at ca. 70 °C:

$$Ca_5[(PO_4)_3F] + 10\,HNO_3 \rightarrow 5\,Ca(NO_3)_2 + HF + 3\,H_3PO_4 \tag{3.4}$$

Usually 50 – 60 % nitric acid is employed in an excess up to 20 % over the stoichiometric amount, depending on the rock phosphate grade and process.

The digestion reaction is governed by three interrelated factors: digestion temperature, particle size, and residence time in the reactor. If the temperature is too high, nitric acid may decompose to form nitrogen oxides. If the temperature is too low and the particles are too large, a longer residence time is necessary.

Crystallization of Calcium Nitrate. The CaO : P_2O_5 ratio of the digester liquor is lowered by crystallizing and removing calcium nitrate tetrahydrate $Ca(NO_3)_2 \cdot 4H_2O$ to enhance the water solubility of phosphorus pentoxide in the fertilizer. The quantity of calcium nitrate tetrahydrate obtained depends on the crystallization temperature, which is usually below 0 °C. The tetrahydrate crystals are washed with cold nitric acid and calcium nitrate – ammonium nitrate solution to recover as much of the phosphorus pentoxide content as possible. The crystals should contain less than 0.3 wt % residual phosphorus pentoxide. Experiments have shown that ca. 60 wt % nitric acid is suitable for washing. The acid is precooled to < 10 °C, sometimes as low as –5 °C, to minimize the risk of redissolving the crystals.

Any acid remaining in the crystal filter cake would interfere with the subsequent conversion of calcium nitrate tetrahydrate. It

is therefore eliminated by treatment with a solution of calcium nitrate and ammonium nitrate.

The tetrahydrate crystals are redissolved in ammonium nitrate solution. The cooling effect produced in this endothermic process and that resulting from the evaporation of ammonia are used to cool the crystallization liquor.

Either the calcium nitrate – ammonium nitrate solution is processed directly to fertilizers (see Section Calcium Nitrate from Limestoneand Nitric Acid), or the calcium nitrate is treated with ammonium carbonate to yield ammonium nitrate.

Conversion of Calcium Nitrate. Figure 14 gives the flow sheet for calcium nitrate conversion. Ammonium carbonate is obtained by reacting ammonia and carbon dioxide in a 65 % ammonium nitrate solution in an absorption tower:

Figure 14. Calcium nitrate conversion

$$2\,NH_3 + CO_2 + H_2O \rightarrow (NH_4)_2CO_3 \qquad (3.5)$$

The reaction occurs in the liquid phase and the product contains < 38 wt % ammonium carbonate.

Calcium nitrate and ammonium carbonate are then reacted at ca. 50 °C in the conversion reactor:

$$(NH_4)_2CO_3 + Ca(NO_3)_2 \rightarrow CaCO_3 + 2\,NH_4NO_3 \qquad (3.6)$$

The reactor operates with a small excess of ammonium carbonate. If the excess is too great, the precipitated lime becomes too fine and is difficult to process. With an excess of calcium nitrate, the precipitate becomes slimy and cannot be filtered. Another reason for avoiding an excess of calcium nitrate is to keep the calcium ammonium nitrate obtained by reacting lime with ammonium nitrate from becoming hygroscopic. Calcium ammonium nitrate, with lime as a filler, contains ca. 26 wt % nitrogen.

3.4. Product Forms, Storage, and Transportation

Calcium nitrate for fertilizer use is marketed as a noncaking grade that is usually treated with aryl alkyl sulfonates. Airtight bins are required for the storage of hygroscopic calcium nitrate. Coated or plastic bags are suitable for transportation. Like other nitrates, calcium nitrate must be kept away from oxidizable substances.

3.5. Uses

Calcium nitrate is employed as a nitrogen fertilizer either directly (see Section Calcium Nitrate from Limestoneand Nitric Acid) or indirectly as calcium ammonium nitrate (see Section Calcium Nitrate as a Byproduct of the Odda Process).

The combined use of calcium nitrate and urea in fertilizers has also been proposed [63], [64]. Other nitrates are obtained by reacting calcium nitrate with salts of other cations, preferably sulfates (see also Section Potassium Nitrate from Calcium Nitrate).

Calcium nitrate is employed as a granulation aid for ground basic slag, ammonium and potassium sulfates, ammonium nitrate, and other fertilizer salts. Because of its high solubility, it is used as an oxidizing additive in water-containing explosives (along with ammonium nitrate).

Addition of calcium nitrate has been suggested as a way of improving the combustion qualities of heating oil and tobacco. It is also used in the preparation of cooling brines and in wastewater treatment (see Chap. Environmental Aspects). Because of its highly endothermic heat of solution, the tetrahydrate $Ca(NO_3)_2 \cdot 4H_2O$ can be used for cooling, for example in medical applications.

3.6. Economic Aspects

The use of calcium nitrate as a fertilizer is declining worldwide. In 1978, only 170 000 t of nitrogen was consumed in this form, most of it in Scandinavia (where the important producers are located), followed by Egypt. No statistics are available on the use of calcium nitrate in the industrial sector.

[Top of Page]

4. Sodium Nitrite

Sodium nitrite is the most industrially important salt of nitrous acid.

4.1. Properties [65]

Sodium nitrite [7632-00-0], $NaNO_2$, M_r 69.00, has an enthalpy of formation of ΔH_{298} = – 358 ± 2 kJ/mol. The β allotrope is stable at 25 °C and in the pure state forms colorless rhombic crystals of space group C_{2v}^{20}; ϱ_{298} = 2.168 g/cm^3 [3]; C_p

, 298 = 63.64 J mol^{-1} K^{-1}; refractive index at 589 nm n_α = 1.35, n_β = 1.46, n_γ = 1.65. At 163 °C, this modification adapts a rhombic form of type $D_2^{25}{}_h$ (enthalpy of crystallization +1.2 kJ/mol) [66], *m p* 271 °C.

Sodium nitrite is hygroscopic (Fig. 15) [65]. The anhydrous salt has an enthalpy of solution of ΔH_{298} = +13.9 kJ/mol at infinite dilution [3]. The hydrate $NaNO_2 \cdot 0.5H_2O$ exists below –5 °C. Sodium nitrite is soluble in liquid ammonia; $2NaNO_2 \cdot NH_3$ forms below –64 °C. The solubility of sodium nitrite in 95 % ethanol at 25 °C is ca. 1.4 %.

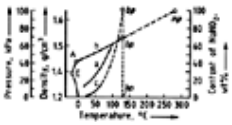

Figure 15. Properties of a saturated solution of sodium nitrite in water

a) Density; b) Content of sodium nitrite; c) Vapor pressure

b p = boiling point of saturated solution containing 68.7 wt % sodium nitrite at 101.3 kPa and 128 °C; A = transition point: –5.1 °C, 41.6 % sodium nitrite; E = cryohydric point: –19.0 °C, 28.6 % sodium nitrite.

Sodium nitrite oxidizes slowly in air. Above 330 °C it decomposes:

$$2\,NaNO_2 \rightarrow Na_2O + NO + NO_2 \tag{4.1}$$

If the resulting nitrogen oxides are not removed, they react further with sodium nitrite:

$$NaNO_2 + NO_2 \rightleftharpoons NaNO_3 + NO \tag{4.2}$$

$$2\,NaNO_2 + 2\,NO \rightarrow 2\,NaNO_3 + N_2 \tag{4.3}$$

so that the end products of decomposition between 330 and 400 °C are sodium nitrate, sodium oxide, and nitrogen. Above 600 °C, the sodium nitrate product decomposes to oxygen and sodium nitrite (reaction 1.1), so that the end products at this temperature are sodium oxide, nitrogen, and oxygen [6].

Aqueous solutions of alkali nitrites reduce strong oxidizing agents (e.g., potassium permanganate solution), but in acidic solution they oxidize potassium iodide to iodine and Fe^{2+} to Fe^{3+}. They oxidize ammonium salts in weakly acidic solution to nitrogen above 50 °C. Urea and amidosulfonic acid are likewise oxidized to nitrogen:

$$CO(NH_2)_2 + 2\,NaNO_2 + 2\,HCl \rightarrow CO_2 + 2\,N_2 + 3\,H_2O + 2\,NaCl \tag{4.4}$$

$$SO_3H - NH_2 + NaNO_2 + HCl \rightarrow N_2 + H_2O + H_2SO_4 + NaCl \tag{4.5}$$

In anhydrous melts at 210 – 220 °C, excess sodium nitrite oxidizes urea in an exothermic reaction that yields sodium carbonate and nitrogen:

$$CO(NH_2)_2 + 2\,NaNO_2 \rightarrow Na_2CO_3 + 2\,N_2 + 2\,H_2O \tag{4.6}$$

Equimolar quantities under the same conditions form sodium cyanate:

$$CO(NH_2)_2 + NaNO_2 \rightarrow NaCNO + N_2 + 2\,H_2O \tag{4.7}$$

The $NaNO_2$ – $NaNO_3$ system forms a eutectic at 37.5 mol % sodium nitrate, *m p* 227 °C. Mixed crystals precipitate from the melt at all concentrations [67].

Nitrites do not occur in mineral form. Nitrite is formed as a metabolic product by microorganisms that oxidize organic nitrogen-containing substances; small amounts of nitrite are thus found in soil and groundwater.

4.2. Production

Until the beginning of the 20th century, sodium nitrite was produced by the reduction of sodium nitrate with lead, iron, zinc, manganese dioxide [which is oxidized to $BaMnO_4$ in the presence of Ba $(OH)_2$], coke, sulfur dioxide, and calcium carbide. Of these processes, the only one of industrial significance was the reduction of molten sodium nitrate with lead at ca. 400 °C. Other earlier processes are described in [14], [68].

Large-scale production of sodium nitrite is now based on the reaction of nitrogen oxides with sodium carbonate or sodium hydroxide solution. The molar ratio of nitrogen oxides is called the “oxidation degree” (OD):

$$\frac{[NO_2]}{[NO] + [NO_2]} \cdot 100 = OD$$

At OD < 50 %, the reaction is described by

$$2\,NaOH + NO_2 + NO \rightarrow 2\,NaNO_2 + H_2O \tag{4.8}$$

Excess nitrogen monoxide does not react but is oxidized to nitrogen dioxide by ambient oxygen. At OD > 50 % and a high partial pressure of oxygen, the following (slower) reactions also take place:

$$2\,NaOH + 2\,NO_2 \rightarrow NaNO_3 + NaNO_2 + H_2O \tag{4.9}$$

$$2\,NaOH + 3\,NO_2 \rightarrow 2\,NaNO_3 + NO + H_2O \quad (4.10)$$

The reaction with sodium carbonate proceeds analogously. In practice, nitrate formation according to reactions (4.9) and (4.10) cannot be completely prevented even at OD < 50 %.

To obtain nitrite solutions with a minimum nitrate content, the OD must be kept below 50 % [69] and the pH above 8. In an acidic environment, nitrous acid is formed and decomposes to form nitric acid:

$$3\,HNO_2 \rightarrow HNO_3 + 2\,NO + H_2O \quad (4.11)$$

If a pure nitrite solution is desired, the gas leaving the absorption columns of the nitric acid plant must first be freed of entrained nitric acid. Because nitrogen dioxide is more thoroughly and rapidly absorbed in the columns than nitrogen monoxide, the gas mixture subjected to alkaline absorption has OD < 50 % and contains 3 – 4 vol % oxygen. Alkaline absorption is described in Section Synthetic Sodium Nitrate [70].

Nitrogen oxides produced by catalytic combustion of ammonia can also be used directly for the production of sodium nitrite. If the oxygen excess is kept small and an intermediate “acid washing” step is used, the OD can be kept lower than 50 %.

4.3. Product Forms, Storage, and Transportation

Sodium nitrite is sold as the salt and in solution. The finely crystalline, slightly yellowish salt is marketed in untreated form and also after treatment with aryl alkyl sulfonates [18]. The salt contains ca. 99.0 % sodium nitrite, 0.6 % sodium nitrate, < 0.1 % sodium chloride and sodium sulfate, and < 0.1 % water.

Because of its hygroscopic nature and toxicity, sodium nitrite can be transported only in packaged form (plastic bags and metal tanks). It must be stored and shipped separately from oxidizable substances, ammonium salts, urea, food, and animal feeds. Sodium nitrite solution is marketed at a density of about 1.3 kg/L; it contains ca. 500 g/L sodium nitrite, 10 – 25 g/L sodium nitrate, and 5 g/L sodium hydrogen carbonate and has a pH of ca. 8. The solution can be stored in iron tanks.

4.4. Uses

Large amounts of sodium nitrite are consumed in the chemical and pharmaceutical industries for the production of nitroso and isonitroso compounds, diazotization reactions (especially for dyes), and the synthesis of pharmaceuticals (e.g., caffeine) and agricultural pesticides (e.g., pyramin).

Other applications of sodium nitrite are in metallurgy and corrosion prevention. It is used as an accelerator in phosphating, added to alkaline pickling solutions for aluminum, applied to steel as a descaling agent, added to drilling oils and abrasives, used for passivating metal surfaces and for creating protective coatings [e.g., of the type $Fe(NH_3)_x(NH_2)_3 \cdot yH_2O$] on carbon steels [71], and added as a corrosion inhibitor to lubricating oils and greases as well as to cooling water loops. Sodium nitrite serves as an oxidizing agent in the alkaline detinning of tinplate scrap [72] and a corrosion inhibitor in heat-transfer baths (see Section Uses).

Sodium nitrite is used in lubricants for glass-forming equipment [73]. Like sodium nitrate and potassium nitrate, it is used as a heat-transfer medium.

Sodium nitrite is recommended as an additive in concrete and gypsum to improve strength and combat the corrosion of iron reinforcement.

In curing salts used in the food industry, sodium nitrite is still the most reliable agent for protecting against botulism, a dangerous bacterial contaminant of meats. High nitrite concentrations can, however, lead to the formation of carcinogenic nitrosamines; legislative restrictions have, therefore, been imposed on the use of nitrite. Methods have been developed for achieving a satisfactory effect with minimal quantities [74]. The simultaneous addition of other substances (e.g., ascorbic acid) is reported to prevent the formation of nitrosamines [75], [76].

4.5. Economic Aspects

Sodium nitrite is produced as an important intermediate by the chemical industry in many countries, usually in conjunction with nitric acid production. Production statistics are not available.

[Top of Page]

5. Potassium Nitrite

5.1. Properties [77]

Potassium nitrite [7758-09-0], KNO_2, M_r 85.10, has an enthalpy of formation of ΔH_{298} = –375 ± 5 kJ/mol. The anhydrous product forms colorless monoclinic crystals of space group C^3_s, which are stable at 25 °C, ϱ_{298} = 1.915 g/cm^3 [3], *mp* 440 °C [3] (decomposition begins below the melting point).

Potassium nitrite is highly hygroscopic (Fig. 16) [78]; the enthalpy of solution of the anhydrous salt is ΔH_{298} = +13.4 kJ/mol at infinite dilution [2]. The hydrate $KNO_2 \cdot 0.5H_2O$ exists below –8.9 °C.

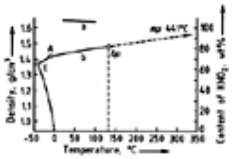

Figure 16. Properties of a saturated solution of potassium nitrite in water

a) Density; b) Content of potassium nitrite

bp = boiling point of saturated solution containing 82.0 wt % potassium nitrite at 100.9 kPa and 132 °C; A = transition point: –8.9 °C, 71.9 % potassium nitrite; E = cryohydric point: –40.2 °C, 64.9 % potassium nitrite.

Thermal decomposition of potassium nitrite yields nitrogen oxides and is analogous to that of sodium nitrite but begins at ca. 410 °C [6].

The system KNO_2 – KNO_3 has a eutectic point at 316 °C and 22.5 mol % potassium nitrite [79]. In the quaternary system KNO_2 – KNO_3 – $NaNO_3$ – $NaNO_2$, the composition with the lowest freezing point (142 °C) is 37.5 mol % potassium nitrite, 37.5 mol % sodium nitrate, and 25 mol % sodium nitrite [80], [23].

5.2. Production

Potassium nitrite can be obtained by reduction of potassium nitrate (similarly to sodium nitrite). A variety of other reactions between sodium nitrate and potassium salts have also been described [77]. Production of potassium nitrite by absorption of nitrogen oxides in potassium hydroxide or potassium carbonate solution is not employed on a large scale because of the high price of these alkalies. Furthermore, the fact that potassium nitrite is highly soluble in water makes the solid salt difficult to recover. The solubility relationships in the system KNO_2 – KNO_3 – H_2O are described in [81].

5.3. Uses

Potassium nitrite is used only in very small quantities, for example, as an oxidizing agent and as a corrosion inhibitor in washing with hot potash for removal of carbon dioxide from gas streams (e.g., synthesis gas). Potassium nitrite is more soluble than sodium nitrite and does not form sparingly soluble bicarbonate in carbon-dioxide-rich solutions.

[Top of Page]

6. Ammonium Nitrite

Ammonium nitrite is a very unstable compound that is virtually always handled in aqueous solution.

6.1. Properties

Ammonium nitrite [13446-48-5], NH_4NO_2, M_r 64.04 [3], has an enthalpy of formation of ΔH_{298} = –260 ± 5 kJ/mol and forms cubic crystals, ϱ_{298} = 1.69 g/cm^3 [3]. The melting point is not well defined because of decomposition.

Ammonium nitrite is highly hygroscopic and readily soluble in water [78] (enthalpy of solution ΔH_{285} = +19.3 kJ/mol at infinite dilution [3]). The saturated solution contains 42.5 wt % ammonium nitrite at the cryohydric point (–27.9 °C), 55.4 % at 0 °C, and 65.0 % at 20 °C. Hydrated solid phases are not known.

Solid ammonium nitrite is stable only in the pure, dry state. Exothermic decomposition begins at 60 °C and may occur explosively:

$$NH_4NO_2 \rightarrow N_2 + 2\,H_2O\ \Delta H = -300\,\mathrm{kJ/mol} \tag{6.1}$$

The saturated solution is stable only up to 30 °C; dilute solutions release nitrogen according to reaction (6.1) above 60 °C. Decomposition involves nitrous acid as an intermediate [82]; the stability of the aqueous solution is enhanced by the addition of ammonia, calcium hydroxide, or potassium iron(II) cyanide.

Dilute ammonium nitrite solutions can be inverted with nitric acid below 30 °C [83]:

$$3\,NH_4NO_2 + 2\,HNO_3 \rightarrow 3\,NH_4NO_3 + 2\,NO + H_2O \tag{6.2}$$

Mixed crystals of ammonium nitrite and ammonium nitrate precipitate from an aqueous solution of NH_4NO_2 – NH_4NO_3 containing > 47 % ammonium nitrite at 20 °C [84]. The most highly concentrated solution in this system at 20 °C contains only 6 % water, 48 % ammonium nitrite, and 46 % ammonium nitrate.

6.2. Production

Ammonium nitrite is produced and utilized industrially solely in the form of a weakly ammoniacal solution with pH > 7.5. When ammonium nitrite solution is prepared, the simultaneous formation of ammonium nitrate must be avoided because the two compounds cannot be separated from one another.

Ammonium nitrite is produced on an industrial scale from nitrogen oxides and ammonia or ammonium carbonate. The reaction is similar to that used for the formation of sodium nitrite. Because of the instability of ammonium nitrite (reaction 6.1), however, care must be taken to keep the pH between 8 and 9 and the temperature between 0 and + 5 °C.

The heat balance of ammonium nitrite formation depends strongly on the feed ammonium compound:

$$2\,NH_3 + NO_2 + NO + H_2O \rightarrow 2\,NH_4NO_2\ \Delta H_{298} = -77.5\,\mathrm{kJ} \tag{6.3}$$

$$(NH_4)_2CO_3 + NO_2 + NO \rightarrow 2\,NH_4NO_2 + CO_2\ \Delta H_{298} = -47\,kJ \quad (6.4)$$

$$2\,NH_4HCO_3 + NO_2 + NO \rightarrow 2\,NH_4NO_2 + CO_2 + H_2O\ \Delta H_{298} = -13.5\,kJ \quad (6.5)$$

Use of ammonium bicarbonate appears advantageous because no water is formed in the reaction to dilute the solution. However, since it has a low solubility the use of ammonium carbonate is generally preferred.

The process is usually carried out under an overpressure of 0.1 – 0.6 kPa [85]. The NO_2 – NO gas mixture is first cooled to 0 °C to decrease the degree of oxidation (see Section Production) to 35 – 40 % by condensation of dilute nitric acid, and then fed into absorption columns connected in series and packed with ceramic Raschig rings or fitted with sieve trays. The circulating solution is held at ca. 0 °C by coolers located at the bottoms of the columns and outside them. To prevent crystallization of ammonium bicarbonate, a metered amount of aqueous ammonia solution is added along with the ammonium carbonate. The final product solution has a pH of 8 and contains ca. 150 – 200 g/L ammonium nitrite, 20 g/L ammonium nitrate, 20 g/L ammonium carbonate, and 2 – 5 g/L ammonia. Ammonium nitrite mist is formed in the absorption step, a final gas scrub is therefore used to recover ammonium carbonate solution for reuse [86].

The ammonium nitrite yield relative to the nitrogen oxides inlet to ammoniacal absorption is less than 90 %. Losses include nitrate formation, nitrite decomposition (reaction 6.1), and tail-gas and mist losses.

6.3. Uses

Ammonium nitrite can be used to passivate steam boiler equipment and chemical plants because it does not leave any nonvolatile residues behind. It is also used as a blowing agent in the manufacture of hollow rubber articles. In the laboratory, very pure nitrites can be prepared from pure ammonium nitrite and the hydroxides of other cations.

[Top of Page]

7. Calcium Nitrite

7.1. Properties [87]

Calcium nitrite, [13780-06-8], $Ca(NO_2)_2$, M_r 132.09, has an enthalpy of formation of ΔH_{298} = –750 ± 10 kJ/mol, ϱ_{273} = 2.26 g/cm^3 [78], *mp* 398 °C (decomposition begins below melting point).

Anhydrous calcium nitrite is hygroscopic and dissolves readily in water (Fig. 17) [78]; the enthalpy of solution for 1 mol in 800 mol of water is ΔH_{293} = –8.5 kJ/mol. The monohydrate crystallizes from solution below 129 °C to form silky hexagonal needles (ϱ_{293} = 2.23 g/cm^3).

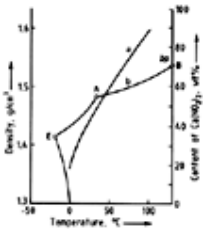

Figure 17. Properties of a saturated solution of calcium nitrite in water

a) Density; b) Content of calcium nitrite

bp = boiling point of saturated solution at 101.3 kPa and 133 °C with decomposition; A = first transition point: $Ca(NO_2)_2 \cdot 4H_2O + Ca(NO_2)_2 \cdot H_2O$ at + 34.6 °C and 55.1 % $Ca(NO_2)_2$; B = second transition point: $Ca(NO_2)_2 \cdot H_2O + Ca(NO_2)_2$ at + 129 °C and 71 % $Ca(NO_2)_2$; E = cryohydric point: ice + $Ca(NO_2)_2 \cdot 4H_2O$ at –20.0 °C and 34.2 % $Ca(NO_2)_2$.

Below 34.6 °C, the tetrahydrate $Ca(NO_2)_2 \cdot 4H_2O$ crystallizes in tetragonal form: ϱ_{293} = 1.72 g/cm^3, enthalpy of solution for 1 mol in 800 mol of water ΔH_{293} = + 33.5 kJ/mol.

Calcium nitrite decomposes to calcium oxide and a mixture of nitrogen monoxide and nitrogen dioxide at temperatures as low as 250 °C [6]. The hydrates are decomposed by atmospheric carbon dioxide, with the formation of nitrogen oxides. A saturated solution of calcium nitrite boils at 133 °C (101.3 kPa) with decomposition.

7.2. Production

Calcium nitrite is formed when a mixture of gaseous nitrogen dioxide and nitrogen monoxide (OD < 50 %) is absorbed in a suspension of calcium hydroxide. This process is used to clean up the final tail gases from low-pressure nitric acid plants when no demand exists for sodium nitrate or sodium nitrite. The spent absorption liquor contains calcium nitrite and generally some calcium nitrate. The nitrite can be recovered as the tetrahydrate $Ca(NO_2)_2 \cdot 4H_2O$ by concentration and cooling. This product has a lower solubility than calcium nitrate tetrahydrate $Ca(NO_3)_2 \cdot 4H_2O$. The nitrite is not, however, recovered on a large scale; the calcium nitrite solution is usually inverted with nitric acid to obtain calcium nitrate.

7.3. Uses

No large-scale application exists for calcium nitrite. Calcium nitrite is reacted with the sulfates of other cations. The corresponding nitrite solutions are obtained after removal of the precipitated gypsum by filtration. Calcium nitrite mixed with calcium nitrate and calcium chloride is recommended as a noncorrosive salt for thawing ice on roads (along with $NaNO_2$). Calcium nitrite is proposed as a corrosion-inhibiting additive in concrete that also improves workability at low temperature.

[Top of Page]

8. Analysis of Nitrates and Nitrites

Nitrate Determination [88], [89]. For *volumetric determination*, nitrate is reacted in strongly alkaline solution with Devarda's alloy (50 % Cu + 45 % Al + 5 % Zn) or in neutral solution with Arnd's alloy (65 % Cu + 35 % Mg) and thus converted to ammonia. The ammonia is distilled into a receiver containing sulfuric acid and determined by acidimetric methods. Ammonium and nitrite are also detected and must, therefore, be determined separately. The method is not applicable in the presence of urea, urea – aldehyde condensates, cyanamide, or protein.

Gravimetrically, nitrate is determined as nitron nitrate after precipitation with nitron in acetic acid solution. Perchlorate, chlorate, nitrite, and other anions precipitate along with nitrate. The method cannot be used in the presence of urea – aldehyde condensates and protein.

At low concentrations (ppm range), nitrate is determined *photometrically* with 2,4-dimethylphenol, brucin, or sodium salicylate. Nitrite interferes and must be removed.

The *potentiometric method* with a nitrate ion-sensitive electrode is suitable for the determination of nitrate in single samples and for automated continuous monitoring (e.g., of wastewater) [90]. Perchlorate, iodide, chlorate, and bromide ions as well as high nitrite concentrations interfere with the method.

Nitrate can be determined in the presence of nitrite by *titrimetry* after reduction with iron(II) sulfate; the end point is indicated potentiometrically [91].

Prior to nitrate determination, nitrite can be destroyed in a weakly acidic medium with ammonium salts, urea, sodium azide, amidosulfonic acid, or hydroxylamine (cf. reactions 4.4 and 4.5).

Ion chromatography is an effective method for determining and separating nitrate [92]. The nitrate ion is detected either with a conductivity detector or a UV detector (excitation wavelength 254 nm). This technique has the advantages of high precision, rapidity, and independence of flow rate [93].

Nitrite Determination. High nitrite concentrations are determined by titration at ca. 35 °C with potassium permanganate in sulfuric acid solution. To avoid loss of nitrite, the solution is allowed to flow with stirring into acidified potassium permanganate solution until it becomes colorless [88]. Other ions that can be oxidized by potassium permanganate in sulfuric acid solution must be considered in evaluating the results.

Low contents of nitrite are determined photometrically with sulfanilamide and *N*-(1-naphthyl)ethylenediammonium dichloride as a red azo dye [94]. The reaction is specific for nitrite and can be automated [95]. This method is recommended for series analyses (e.g., of water samples).

Determination of Cations.
Sodium as the primary constituent can be determined gravimetrically by precipitation with uranyl magnesium acetate. This method can be employed even with a large excess of potassium [89]. Flame photometry and atomic absorption spectroscopy can be used for lower concentrations and in the trace range.

Potassium is precipitated as sparingly soluble potassium tetraphenyl borate and determined as such by weighing after treatment with sodium tetraphenyl borate in weakly alkaline or acidic solution [96-98]. Volumetric determination of potassium with sodium tetraphenyl borate and an electrometric end point indication is also important [98-100]. The chloroplatinate method is employed in the United States [89], [100].

Thermometry measures the heat (ca. –50 kJ/mol) released when potassium is precipitated as the perchlorate [101]. Potassium is also determined by conductometric titration with sodium tetraphenyl borate [98].

The potassium content can be determined quickly in process analysis by flame photometry [89] and atomic absorption spectroscopy. In the plant, β or γ radiation of ^{40}K solid salts or a solution can be utilized for analysis, product inspection, and process control [102].

Ammonium is determined acidimetrically by distillation of ammonia from alkaline solution. Trace amounts can be determined by photometry after formation of indophenol or reaction with Nessler's reagent [94]. Potentiometric determination with an ammonia electrode is suitable for all concentration ranges [90].

Calcium and magnesium in nitrates and nitrites are determined volumetrically by titration with the demineralized disodium salt of ethylenediaminetetraacetic acid. Other divalent cations must first be masked or precipitated [103].

Determination of Impurities. *Water content* can be measured either by the Karl Fischer method or by determination of the weight loss at 105 °C [89].

High concentrations of *chloride* are determined volumetrically by titration with silver nitrate [89]; trace amounts are measured by titration with mercury(II) nitrate and diphenylcarbazone as indicator [100]. A photometric technique is based on the determination of iron(III) rhodanide formed after liberation of rhodanide ion from mercury(II) rhodanide by chloride ion [95]. Very low concentrations can be determined turbidimetrically as silver chloride at 420 nm. This has the advantage that toxic mercury is not involved in the analysis.

Chlorate can be determined iodometrically in strong hydrochloric acid solution. *Perchlorate* is determined gravimetrically as potassium perchlorate [89]. *Iodate* can be determined iodometrically in weakly acidic solution. (Nitrite ion must be destroyed before iodometric analyses are performed.)

Boric acid can be determined by acidimetric titration, either directly or after isolation as methyl borate in the presence of mannitol or glycerol [89].

[Top of Page]

9. Environmental Aspects

Nitrate Removal. Nitrates are undesirable in wastewater for two reasons: uncontrolled denitrification can interfere with the final clarification step in biological wastewater treatment plants, and high nitrate concentrations can cause contamination of drinking water.

In 1970 the World Health Organization (WHO) issued guidelines recommending that the maximum limit for nitrate concentrations should be 50 mg/mL, this was later reduced to 45 mg/mL. The WHO also recommended a maximum acceptable concentration of 100 mg/L [105]. In the United States the EPA proposed regulations for drinking water on May 22, 1989 [106]. Public comment closed August 21, 1989, and the final rule was scheduled for December 1990. A maximum containment level for total nitrate and nitrite of 10 mg/L (as nitrogen) is proposed. As early as July 1980, the Council of the European Community issued a directive concerning the quality of water for human consumption [107]. The maximum permissible nitrate concentration was set at 50 mg/L, with a guide level of only 25 mg/L. Member countries agreed that drinking-water quality would meet this standard within five years (for the members Portugal, Spain, and Greece, who joined the EC later, the appointed five-year period started in 1988). To achieve this, some waterworks dilute their water with low-nitrate water; the trend to equip waterworks with nitrate removal facilities is also increasing. Because this equipment is very expensive, nitrate removal from wastewater before it reaches the treatment plant and reduction in nitrate loading of surface waters (agricultural) are desirable goals; see also ⟶ Fertilizers – Nitrogen. Three methods have been described for the removal of nitrates from wastewater: reverse osmosis, anion exchange, and denitrification.

Reverse Osmosis. The principles of reverse osmosis are described elsewhere (⟶ Membranes and Membrane Separation Processes – Principle of Reverse Osmosis). Pressure is used to force water through a membrane that is impermeable to any dissolved solutes or suspended solids. Despite its high cost, reverse osmosis is becoming more and more important in wastewater treatment; the cost depends on the nature of the inlet wastewater stream (pretreatment). If the wastewater has a high suspended solid content or the dissolved salts might crystallize, reverse osmosis can only be used to a limited extent because the membranes may become plugged. This technique results in the production of a concentrated nitrate solution, which is then dealt with by other methods.

Anion Exchange. Units based on anion-exchange resins [108] can only be used in special cases, because of their inadequate selectivity for nitrate versus chloride ions. The same is true for a process developed especially for the removal of nitrate from wastewater in cellulose nitrate production, which utilizes water-insoluble secondary amines in the chloride form as liquid ion exchangers [109]. In contact with wastewater, these substances exchange chloride for nitrate. In the next step, the amines exchange their nitrate ions for chloride ions from a potassium chloride solution to yield pure potassium nitrate as a byproduct. This technique, however, requires virtually chloride-free wastewater. A similar procedure has been described in which the amine is used in the sulfate form and ammonium nitrate is obtained as byproduct after stripping with ammonia gas [110].

Denitrification. Denitrification is currently the most important process for nitrate removal [111-114]. Nitrate is reduced to nitrogen by facultatively anaerobic bacteria (e.g., *Achromobacter*, *Denitrobacillus, Nitrococcus, or Spirillum*) under anaerobic conditions. These bacteria can use nitrate and nitrite as a source of oxygen.

$$2\,NO_3^- + 10\,H^+ + 10\,e^- \rightarrow 2\,OH^- + 4\,H_2O + N_2 \qquad (9.1)$$

$$2\,NO_2^- + 6\,H^+ + 6\,e^- \rightarrow 2\,OH^- + 2\,H_2O + N_2$$

To maintain metabolism, an adequate supply of carbohydrates must be present or added. The carbohydrates are oxidized to carbon dioxide and water.

Denitrification is the sole metabolic route for converting fixed nitrogen to the molecular form. The required anaerobic conditions are considered satisfied if the BOD_5 value (five-day biological oxygen demand) is more than 1.5 times the oxygen fixed in the nitrate.

Nitrite Removal. Nitrites must be removed from wastewater because they are toxic (see Chap. Toxicology and Occupational Health) and also because they interfere with the oxygen economy of natural waters and water treatment plants. Nitrite ions can be eliminated by oxidation to nitrate, by reduction to elemental nitrogen, or by denitrification (see Nitrate Removal).

Oxidation to Nitrate. Nitrite is reacted with, for example, sodium hypochlorite in a weakly acidic medium (pH 3 – 4):

$$NaNO_2 + NaOCl \rightarrow NaNO_3 + NaCl \qquad (9.2)$$

This process has the advantage of a high reaction rate with virtually 100 % conversion. It can be used as a continuous method through relatively simple redox potential measurements and automation. It has the drawback, however, of increasing the sodium chloride content of water [115].

Another option is oxidation with hydrogen peroxide:

$$NaNO_2 + H_2O_2 \rightarrow NaNO_3 + H_2O \qquad (9.3)$$

This reaction occurs at pH 3 – 3.5 and can be effected only in batch operation because a stepwise measurement and hydrogen peroxide metering cycle must be used. The process has the advantages of low cost and no salting (Table 7) [115].

Table 7. Comparison of nitrite removal processes using hydrogen peroxide and sodium hypochlorite [110]

	Hydrogen peroxide	Sodium hypochlorite
Batch operation	yes	yes
Continuous operation	no	yes
pH	2.0 – 4.0	3.0 – 3.5
Concentration of oxidizing agent, wt %	35	12.5
Oxidizing agent required, kg/kg nitrite	2.25	13.5
Cost of oxidizing agent, DM/kg nitrite	3.85	6.48
Salt produced, kg/kg nitrite	0	2.5 NaCl
Loss of active agent on storage	no	yes

Nitrite is also oxidized to nitrate by *Nitrobacter* species under aerobic conditions:

$$NO_2^- + 1/2\,O_2 \rightarrow NO_3^-$$

Reduction to Nitrogen. Nitrites can be reduced to nitrogen with amidosulfuric acid:

$$NaNO_2 + NH_2SO_2OH \rightarrow N_2 + NaHSO_4 + H_2O \qquad (9.4)$$

This reaction is of value for high nitrite levels and is carried out at ca. pH 4. Urea is also employed as reducing agent (reactions 4.4. and 4.5).

[Top of Page]

10. Toxicology and Occupational Health [116]

Nitrates and nitrites cause relaxation of smooth muscle cells [117, pp. 224 – 244], which can be so drastic that the contractile system is totally inhibited (protein phosphorylation). These compounds have a vasodilative action, allowing blood to pool in the veins so that the heart is not completely filled. The systolic blood pressure therefore drops, and the pulse rate increases. Dilation of blood vessels in the brain leads to headaches. Acute circulatory collapse and fainting are possible. Dilation of vessels in the retina leads to vision disorders.

Acute, severe poisoning is associated with a critical drop in blood pressure, vomiting, and cyanosis, and ends in death due to respiratory and circulatory failure. Methemoglobin produced by reaction of inorganic nitrite or amyl nitrite with hemoglobin is responsible for this effect. Nitrates are converted to nitrites by microbes in the intestines. Methemoglobin prevents oxygen transport because its central heme group contains a Fe^{3+} ion which is incapable of binding oxygen. Only hemoglobin (Hb) with its central Fe^{2+} ion can bind oxygen reversibly:

$$Hb \cdot Fe^{2+} \cdot O_2 \rightleftharpoons Hb \cdot Fe^{3+} \cdot OH^-$$

In the presence of nitrite, coupled oxidation occurs [117, p. 650]. Oxygen is transferred from the hemoglobin to nitrite with the formation of nitrate and methemoglobin (Fe^{3+}). In intact erythrocytes methemoglobin is reduced by enzymes (diaphorase, methemoglobin reductase).

Methemoglobin formation occurs more readily in premature and newborn babies than in adults. The intestinal flora of babies differs from that of adults and has a higher reducing capacity; nitrate is therefore more easily reduced to nitrite. In the fetus hemoglobin is more readily oxidized than in adults and the erythrocyte reductase system is not fully developed [118]. Vegetables fertilized with nitrate or tap water containing nitrate can cause methemoglobinemia in babies.

Severe intoxication is expected at an oral dosage level > 2 g. A single oral dose of 4 g of nitrite is lethal in humans [119]. In the event of poisoning, the stomach should be pumped and a physician notified [117, p. 250], [120], [121].

Nitrosamines cause tumor formation in animals that depends on dosage and route of administration. Nitrosamines are classified as risk factors for tumor formation in humans.

In the stomach secondary and tertiary amines can react with nitrite to form nitrosamines. Secondary amines are formed during cooking (roasting, frying) of proteins and alcoholic fermentation, and are constituents of flavoring substances. Low nitrosamine concentrations are found in pickled meat, alcoholic beverages, and tobacco smoke. Tetracycline contains groups that can be nitrosylated. The drug aminophenazone has been withdrawn from the market on account of its nitrosamine content (*N*-nitrodimethylamine).

[Top of Page]

References

1. *Gmelin*, **21**, 982 – 1048.
2. J. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances*, Springer Verlag, Berlin-Göttingen-

Heidelberg-New York 1973.
3. R. C. Weast, M. J. Astle: *Handbook of Chemistry and Physics*, 63rd ed., CRC Press, Boca Raton, Florida 1982/83.
4. R. Schamm, K. Tödheide, *High Temp. High Pressures* **8** (1976) no. 1, 65 – 71.
5. F. Winkler, H.-H. Emons, W. Klauer, H. Kühn, *Z. Chem.* **12** (1972) no. 5, 190 – 191.
6. K. H. Stern, *J. Phys. Chem. Ref. Data* **1** (1972) no. 3, 747 – 772.
7. R. N. Kust, J. D. Burke, *Inorg. Nucl. Chem. Lett.* **6** (1970) 333 – 335.
8. J. J. Lehr: *Handbuch der Pflanzenernährung und Düngung*, **vol. 2**, Springer Verlag, Wien 1968, pp. 1003–1013.
9. G. E. Ericksen: "Geology and Origin of the Chilean Nitrate Deposits," *U.S. Geol. Surv. Prof. Paper* **1188** (1981) 41.
10. D. E. Garret: "The Chemistry and Origin of the Chilean Sodium Nitrate Deposits," *Proceedings of the Sixth International Symposium on Salt*, Toronto, May 24 – 28, 1983.
11. US-Bureau of Mines Minerals: Commodity Summaries 1979.
12. *Mining Annual Review* (1988) 296.
13. *Gmelin*, **21**, 95 – 118.
14. *Ullmann*, 3rd ed., **15**, 52 – 67.
15. F. A. Henglein, *Chem. Ztg.* **82** (1958) 287 – 294.
16. *Gmelin*, **21**, 372.
17. *Nitrogen*, **148** (1984) March-April, 19 – 21.
18. F. Wolf, J. Holzweissig, *Chem. Techn. (Leipzig)* **20** (1968) 477 – 480.
19. P. Lesage, M. Goetz, *Trait. Therm.* **114** (1977) 29–36.
20. D. A. Nissen, *J. Chem. Eng. Data* **27** (1982) 269 – 273. Links
21. D. J. Rogers, G. J. Janz, *J. Chem. Eng. Data* **27** (1982) 424 – 428. Links
22. T. Asahina, M. Kosaka, K. Tajiri, *Sol. World Congr., Proc. Bienn. Congr. Int. Sol. Energy Soc., 8th*, **vol. 3**, Pergamon Press, Oxford 1983, pp. 1716 – 1720.
23. Y. Takahashi, R. Sakamoto, M. Kaminoto, *Int. J. Thermophys.* **9** (1988) no. 6, 1081 – 1090. Links
24. C. Wahba (ed.): *World Directory of Fertilizer Manufacturers*, 6th ed., The British Sulphur Corporation, London 1986.
25. *Chem. Mark. Rep.* **236** (1989) no. 7, 7.
26. M. Smart, G. Russell, M. Tashiro: "Nitrogen Products", *Chem. Economics Handbook*, SRI-International, August 1987, 757.8002T.
27. *Gmelin*, **22**, 269 – 323.
28. A. Mustajoki, *Ann. Acad. Sci. Fenn. Ser. A 6* **99** (1962) 1 – 11.
29. P. Weidenthaler, *J. Phys. Chem. Solids* **25** (1964) 1491 – 1493.
30. E. Rapport, G. C. Kennedy, *J. Phys. Chem. Solids* **26** (1965) 1995 – 1997.
31. *Gmelin*, **22**, 1111.
32. W. Reinders, *Z. Anorg. Allgem. Chem.* **93** (1915) 202 – 212.
33. E. Cornec, H. Krombach, *Ann. Chim. (Paris)* **12** (1929) 235 – 253.
34. E. Jänecke, *Z. Anorg. Chem.* **71** (1911) 1 – 18. Links
35. A. M. Babenko, *Zh. Prikl. Khim. (Leningrad)* **48** (1975) no. 8, 1752 – 1756.
36. F. Chemnitus, *Chem. Ztg.* **53** (1929) no. 9, 85 – 86.
37. Yu. N. Dulepov, V. P. Napol'skikh, V. A. Postnikov, *Khim. Promst. (Moscow)* 1978 no. 9, 699 – 701.
38. Gewerkschaft Victor, DE 974 061, 1952 (H. Schmalfeldt, W. Vollmer).
39. *Ullmann*, 4th ed., **20**, 342 – 344.
40. S. J. Gohil, V. P. Mohandas, J. M. Joshi, J. R. Sanghavi, *Res. Ind.* **31** (1986) no. 3, 238 – 240.
41. U. K. Tipnis, B. T. Mandalia, *Res. Ind.* **32** (1987) no. 2, 82 – 84.
42. E. Jänecke, *Angew. Chem.* **41** (1928) 916 – 924.
43. Allied Chemical Corp., US 3 595 609, 1969 (J. L. Beckham).
44. *Phosphorus Potassium* (1971) no. 51, 50 – 52.
45. *Phosphorus Potassium* (1971) no. 51, 52 – 55.
46. N. P. Finkelstein, S. H. Garnett, L. Kogan: "Process for the Manufacture of Chloride Free Potassium Fertilizers," *Int. Potash Technol. Conf., 1st*, (1983) 571 – 576.
47. Haifa Chemicals, US 4 364 914, 1982 (S. Manor, M. Bar-Gori, A. Alexandrow, M. Kreisel).
48. A. Eyal, J. Mizrahi, A. Baniel, *Ind. Eng. Chem. Process Des. Dev.* **24** (1985) no. 2, 387 – 390.
49. *Phosphorus Potassium* (1988) no. 153, 14 – 16.
50. *Chem. Eng.* **74** (1967) no. 24, 118 – 119.
51. American Metal Climax, Inc., US 3 211 525, 1959 (J. D. Buehler et al.).
52. Securities and Exchange Commission S1-SEC-Registration, Feb. 6, 1989.
53. *Chem. Week* **142** (1988) no. 4, 34.
54. *Phosphorus Potassium* (1986) no. 141, 16.
55. *Gmelin*, **28**, 341 – 384. Gmelin, 28, 1318 – 1321.
56. J. Boerio, P. A. G. O'Hare, *J. Chem. Thermodyn.* **8** (1976) 725 – 729. Links
57. W. W. Ewing, N. L. Krey, H. Law, E. Lange, *J. Am. Chem. Soc.* **49** (1927) 1958 – 1973. Links

58. W. W. Ewing, A. N. Rogers, J. Z. Miller, E. Mc Govern, *J. Am. Chem. Soc.* **54** (1932) 1335 – 1343. Links
59. M. Doumeng, *Rev. Chim. Miner.* **7** (1970) 897 – 925.
60. L. V. Opredelenkova *et al.*, *Ukr. Khim. Zh. (Russ. Ed.)* **37** (1971) no. 2, 131 – 136.
61. *Phosphorus Potassium* (1988) no. 155, 28 – 33.
62. *Phosphorus Potassium* (1990) no. 166, 21 – 24.
63. S. Lanyi *et al.*, *Bul. Inst. Politech. Bucure ̦ti Ser. Chim.* **50** (1988) 55 – 60.
64. A. Crispoldi, A. Moriconi, M. Chiappafreddo, EP 278 562 A2, 1988.
65. *Gmelin*, **21**, 960 – 982.
66. M. I. Kay, *Ferroelectrics* **4** (1972) 235 – 243.
67. V. I. Kosyakov, Z. G. Bazarova, A. N. Kirgintsev, *Izv. Akad. Nauk SSSR Ser. Khim.* **197** (1971) no. 3, 643 – 645.
68. *Gmelin*, **21**, 263 – 266.
69. S. D. Deshpande, S. N. Vyas, *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1979) no. 1, 69 – 71.
70. G. C. Inskeep, T. H. Henry, *Ind. Eng. Chem.* **45** (1953) 1386 – 1395.
71. A. P. Aholzin, Yu. Ya. Kharitonov, S. G. Kovalenko, *Werkst. Korros.* **38** (1987) 417 – 421.
72. Goldschmidt, DE 1 621 581, 1966.
73. Glass. Eng. Res. Inst., SU 729 150, 1978.
74. Monsanto, DE 2 713 259, 1977.
75. J. Sander, *Umsch. Wiss. Tech.* **74** (1974) 780.
76. Unilever, GB 1 440 183, 1972.
77. *Gmelin*, **22**, 263 – 268.
78. J. Bureau, *Ann. Chim. (Paris)* **8** (1937) 89 – 139.
79. A. N. Kruglov, O. M. Lutova, *Zh. Neorg. Khim.* **21** (1976) 2767 – 2769.
80. S. Berul *et al.*, *Izv. Sekt. Fiz. Khim. Anal. Inst. Obsch. Neorg. Khim. Akad. Nauk SSSR* **25** (1954) 218 – 235.
81. Y. S. Shenkin, S. A. Ruchnova, E. V. Gorozhankin, *Zh. Neorg. Khim.* **20** (1975) 1717 – 1719.
82. E. Abel, *Monatsh. Chem.* **81** (1950) no. 4, 539 – 542. Links
83. B. V. Stamicarbon, NL 7 608 957, 1976.
84. P. I. Protsenko, A. A. Ugryumova, T. E. Kutsenko, *Zh. Prikl. Khim (Leningrad)* **44** (1971) no. 1, 30 – 34.
85. M. A. Miniovich, N. I. Smalii, *Khim. Promst. (Moscow)* **50** (1974) no. 5, 349 – 353.
86. Bayer, DE 2 015 156, 1970.
87. *Gmelin*, **28**, 338 – 341.
88. W. Fresenius, G. Jander: *Handbuch der analytischen Chemie*, Part III: Quantitative Analyse, vol. 5a: Stickstoff, Springer Verlag, Berlin-Göttingen-Heidelberg 1957.
89. Verband Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (VDLUFA): “Die Untersuchungen von Düngemitteln,” *Methodenbuch (Loseblattsammlung)*, **vol. 2**, J. Neumann-Neudamm, Melsungen 1976.
90. Gesellschaft deutscher Chemiker: *Vom Wasser*, **vol. 49**, Verlag Chemie, Weinheim, Germany 1978, pp. 139–171.
91. W. D. Treadwell, H. Vontobel, *Helv. Chim. Acta* **20** (1937) 573 – 589. Links
92. J. L. Veuthey, J. P. Senn, W. Haerdi, *J. Chromatogr.* **445** (1988) no. 1, 183 – 188. Links G. Schwedt, C. Byung Seo, S. Dreyer, E. U. Ruhdel: “Nitrite, Nitrate in Waters. Photometric Rapid Testing and Ion Chromatography in Comparison,” *Chromatographie, Spektroskopie*, Vogel-Verlag, Würzburg 1987, pp. 46 – 47.
93. *Nitrogen* (1984) no. 151, pp. 39 – 40.
94. Gesellschaft deutscher Chemiker: *Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlamm-Untersuchung (Loseblattsammlung)*, Verlag Chemie, Weinheim, Germany 1979.
95. Technicon Industrial Systems: *Industrial Method*, No. 99–70 W/B$^+$ Tarrytown 1974; No. 102–70 W/C$^+$, Tarrytown 1978.
96. P. Raff, W. Brotz, *Fresenius Z. Anal. Chem.* **133** (1951) 241 – 248.
97. ISO Standards 2051, 2485.
98. DIN-Norm 10 400.
99. H. J. Schmidt, *Fresenius Z. Anal. Chem.* **157** (1957) 321 – 337.
100. H. Tollert: “Analytik des Kaliums,” *Die chemische Analyse*, **vol. 51**, Enke Verlag, Stuttgart 1962.
101. G. Peuschel, F. Hagedorn, *Kali Steinsalz* **6** (1972) 4–11.
102. K. C. Scheel, *Angew. Chem.* **66** (1954) 102 – 106.
103. G. Schwarzenbach, H. Flaschka: *Die Komplexometrische Titration*, Enke Verlag, Stuttgart 1965.
104. D. J. de Renzo: *Nitrogen Control and Phosphorus Removal in Sewage Treatment*, Noyes Data Corp., Park Ridge, N.Y. 1978.
105. *Nitrogen* (1987) no. 167, 14 – 17.
106. National Primary and Secondary Drinking Water Regulations, Fed. Reg. 54 : 97 : 22062, May 22nd, 1989.
107. EEC Directive July 15th, 1980, 80/778/EEC, Aug. 30th, 1980.
108. A. B. Mindler: “Removal of Ammonia and Nitrates from Wastewaters,” in C. Calmon, H. Gold (eds.): *Ion Exchange and Pollution Control,* vol. 1, CRC, Boca Raton, Fla, 1979, pp. 217 – 223.
109. T. K. Mattila, T. K. Letho, *Ind. Eng. Chem. Process Res. Dev.* **16** (1977) no. 4, 469 – 472.
110. Kemira Oy, DE 2 350 962, 1973.
111. N. Matsche, *Muench Beitr. Abwasser- Fisch- Flussbiol.* **43** (1989) 308 – 330.

112. J. P. van der Hoek, A. Klapwijk, *Water Supply* **6** (1988) 57 – 62.
113. O. Novak, K. Svardal, *Wien Mitt.: Wasser-Abwasser Gewaesser* **81** (1989) F1 – F54.
114. B. P. Gayle, G. D. Boardman, J. H. Sherrard, R. E. Benoit, *J. Environ. Eng. (N.Y.)* **115** (1989) no. 5, 930 – 943.
115. M. Rodenkirchen, *Galvanotechnik* **74** (1983) no. 8, 924 – 928.
116. N. J. Sax: *Dangerous Properties of Industrial Materials*, 6th ed., Van Nostrand Reinhold Comp., New York 1984, pp. 2002 – 2005.
117. W. Forth, D. Henschler, W. Rummel (eds.): *Allgemeine und spezielle Pharmakologie und Toxikologie*, 4th ed., B.J.-Wissenschaftsverlag, Mannheim 1983.
118. W. Wirth, G. Hecht, C. Gloxhuber: *Toxikologie-Fibel*, Thieme, Stuttgart 1967.
119. G. Fülgraff, D. Palm (eds.): *Pharmakotherapie/Klinische Pharmakologie*, 4th ed., Fischer-Verlag, Stuttgart 1982, p. 368.
120. H. Thaler, *Dtsch. med. Wochenschr.* **101** (1976) 1740 – 1742. Links
121. H.-H. Roth, *Dtsch. Apoth. Ztg.* **119** (1979) 1097.

[Top of Page]

Related Articles

For ammonium nitrate, see ⟶ Ammonium Compounds – Ammonium Nitrate; nitrates and nitrites not discussed here are dealt with under the respective cations (e.g., barium nitrate, ⟶ Barium and Barium Compounds – Barium Nitrate; cadmium nitrate, ⟶ Cadmium and Cadmium Compounds; chromium nitrate, ⟶ Chromium Compounds; cobalt nitrate, ⟶ Cobalt and Cobalt Compounds; and copper nitrate, ⟶ Copper Compounds).

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Nitric Acid, Nitrous Acid, and Nitrogen Oxides

Nitrates and Nitrites is a separate keyword.

Michael Thiemann, Uhde GmbH, Dortmund, Federal Republic of Germany

Erich Scheibler, Uhde GmbH, Dortmund, Federal Republic of Germany

Karl Wilhelm Wiegand, Uhde GmbH, Dortmund, Federal Republic of Germany

1. Nitric Acid 1
1.1. Introduction 1
1.2. Properties 2
1.3. Industrial Production 3
1.3.1. Oxidation of Ammonia 4
1.3.2. Oxidation and Absorption of Nitrogen Oxides 8
1.3.3. Equipment 15
1.3.3.1. Filters and Mixers 15
1.3.3.2. Burners and Waste-Heat Boilers . . 16
1.3.3.3. Compressors and Turbines 18
1.3.3.4. Heat Exchangers and Columns . . . 21
1.3.3.5. Construction Materials 22
1.3.4. Processes 24
1.3.4.1. Weak Acid Processes 24
1.3.4.2. Concentrated Acid Processes 30
1.4. Environmental Protection 35
1.4.1. Wastewater 35
1.4.2. Stack Gas 35
1.4.2.1. Emission Limits 35
1.4.2.2. Analysis 36
1.4.2.3. Control of NO_x Emissions 36
1.5. Storage and Transportation 40
1.6. Uses and Economic Aspects 40
2. Nitrous Acid 40
3. Nitrogen Oxides 41
3.1. Dinitrogen Monoxide 42
3.2. Nitrogen Monoxide 43
3.3. Nitrogen Dioxide and Dinitrogen Tetroxide 44
3.4. Dinitrogen Trioxide 45
3.5. Dinitrogen Pentoxide 45
4. Toxicology and Occupational Health 46
5. References 47

1. Nitric Acid

1.1. Introduction

Nitric acid is a strong acid that occurs in nature only in the form of nitrate salts. When large-scale production of nitric acid began, sodium nitrate (soda saltpeter, Chile saltpeter) was used as the feedstock. At the beginning of the 20th century the reserves of Chile saltpeter were thought to be nearing exhaustion, so processes were developed for replacing nitrogen from natural nitrates with atmospheric nitrogen. Three techniques were used industrially:

1) Production of nitrogen monoxide by reacting atmospheric nitrogen and oxygen at > 2000 °C (direct processes)
2) Production of ammonia by hydrolysis of calcium cyanamide under pressure
3) Production of ammonia from nitrogen and hydrogen

The direct combustion of air in an electric arc, developed by Birkeland and Eyde, was abandoned because of its poor energy efficiency. Later direct processes, such as thermal nitrogen monoxide synthesis with fossil fuels or nitrogen monoxide synthesis in nuclear reactors, did not gain widespread acceptance. Ammonia production from calcium cyanamide was of only transitory value. Ammonia produced from nitrogen and hydrogen by the Haber – Bosch process (→ Ammonia) is, however, still used as feedstock for nitric acid production.

The crucial step in nitric acid production, the catalytic combustion of ammonia, was developed by Ostwald around the turn of the century. The most important design parameters for a nitric acid plant were determined first in laboratory tests and later in a pilot plant. The first production facility employing the Ostwald process came on stream in 1906 at Gerthe in Germany [1–3].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_293

Most of the nitric acid produced is used to form inorganic fertilizers (→ Phosphate Fertilizers); it is mostly neutralized with ammonia to form ammonium nitrate (→ Ammonium Compounds, Chap. 1.2.1.).

1.2. Properties

Nitric acid was known to the ancient Egyptians because of its special ability to separate gold and silver. Many well-known alchemists in the Middle Ages experimented with the acid. In the middle of the 17th century, GLAUBER reported its preparation from saltpeter and sulfuric acid.

Physical Properties. Nitric acid [*7697-37-2*], M_r 63.013, is miscible with water in all proportions. At a concentration of 69.2 wt %, it forms a maximum-boiling azeotrope with water. The azeotropic mixture boils at 121.8 °C. Pure anhydrous nitric acid boils at 83 – 87 °C; the reason a range of boiling points are cited in the literature is that the acid decomposes on heating [4]:

$$4\,HNO_3 \longrightarrow 2\,H_2O + 4\,NO_2 + O_2$$

The nitrogen dioxide formed on decomposition colors the acid yellow or even red at higher concentrations. Because the vapors can also absorb moisture, the term "red fuming nitric acid" is used. In the pure anhydrous state, nitric acid is a colorless liquid.

The most important physical properties of pure nitric acid follow:

fp	−41.59 °C
bp	82.6 ± 0.2 °C
Density, liquid	
at 0 °C	1549.2 kg/m^3
at 20 °C	1512.8 kg/m^3
at 40 °C	1476.4 kg/m^3
Refractive index n_D^{24}	1.3970
Dynamic viscosity	
at 0 °C	1.092 mPa · s
at 25 °C	0.746 mPa · s
at 40 °C	0.617 mPa · s
Surface tension	
at 0 °C	0.04356 N/m
at 20 °C	0.04115 N/m
at 40 °C	0.03776 N/m
Thermal conductivity (20 °C)	0.343 W m^{-1} K^{-1}
Standard enthalpy of formation	
Liquid	2.7474 J/g
Gas	2.1258 J/g
Heat of vaporization (20 °C)	626.3 J/g
Specific heat	
at 0 °C	1.7601 J g^{-1} K^{-1}
at 20 °C	1.7481 J g^{-1} K^{-1}

The decomposition of nitric acid makes its physical properties difficult to determine at higher temperature. Up to ca. 50 °C, conventional methods of measurement are possible; beyond this, indirect thermodynamic calculations or special short-time measuring methods must be employed. Figure 1 illustrates the vapor pressure curve of pure nitric acid. Table 1 lists important physical properties of aqueous nitric acid. A comprehensive review of the physical and chemical properties of nitric acid is given in [6].

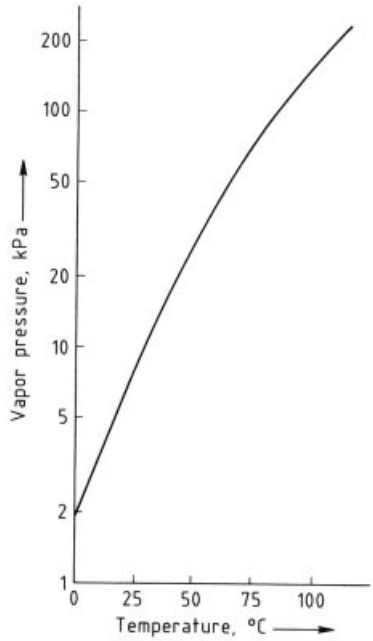

Figure 1. Vapor pressure curve for pure nitric acid

Chemical Properties. Concentrated nitric acid, with nitrogen in the +5 oxidation state, acts as a strong oxidizing agent. The reaction

$$NO_3^- + 4\,H^+ \rightleftharpoons NO + 2\,H_2O$$

goes to the right for all substances with oxidation potentials more negative than +0.93 V [7]. For example, copper (+0.337 V) and silver (+0.799 V) are dissolved by nitric acid, whereas gold (+1.498 V) and platinum (+1.2 V) are resistant. In practice, 50 % nitric acid (aqua fortis)

Table 1. Physical properties of aqueous nitric acid as a function of composition [5]

HNO_3 concentration, wt %	Density (20 °C), g/cm^3	*mp*, °C	*bp*, °C	Partial pressure (20 °C), kPa	
				HNO_3	H_2O
0	0.99823	0	100.0		2.23
10	1.0543	−7	101.2		2.26
20	1.1150	−17	103.4		2.02
30	1.1800	−36	107.0		1.76
40	1.2463	−30	112.0		1.44
50	1.3100	−20	116.4	0.03	1.05
60	1.3667	−22	120.4	0.12	0.65
70	1.4134	−41	121.6	0.39	0.35
80	1.4521	−39	116.6	1.4	0.12
90	1.4826	−60	102.0	3.6	0.03
100	1.5129	−42	86.0	6.0	0

is used for separating gold from silver. Some base metals, such as aluminum and chromium, are attacked by nitric acid only at their surfaces because a thin oxide layer is formed, which protects (i.e., passivates) the core against further oxidation. As a result of this passivation, alloyed steel equipment can be used in nitric acid technology.

Highly diluted nitric acid is almost completely dissociated

$$HNO_3 + H_2O \rightleftharpoons H_3O^+ + NO_3^-$$

and does not attack copper and more noble metals. Due to its acid nature, however, it reacts with base metals, liberating hydrogen and forming nitrates.

A mixture (volume ratio 3 : 1) of concentrated nitric acid and concentrated hydrochloric acid (aqua regia) also dissolves gold.

1.3. Industrial Production

The industrial production of nitric acid by the Ostwald process is described in this section. The process involves three chemical steps (Fig. 2):

1) Catalytic oxidation of ammonia with atmospheric oxygen to yield nitrogen monoxide:

$$4\,NH_3 + 5\,O_2 \longrightarrow 4\,NO + 6\,H_2O \quad (1)$$

2) Oxidation of the nitrogen monoxide product to nitrogen dioxide or dinitrogen tetroxide:

$$2\,NO + O_2 \longrightarrow 2\,NO_2 \rightleftharpoons N_2O_4 \quad (2)$$

3) Absorption of the nitrogen oxides to yield nitric acid:

$$3\,NO_2 + H_2O \longrightarrow 2\,HNO_3 + NO \quad (3)$$

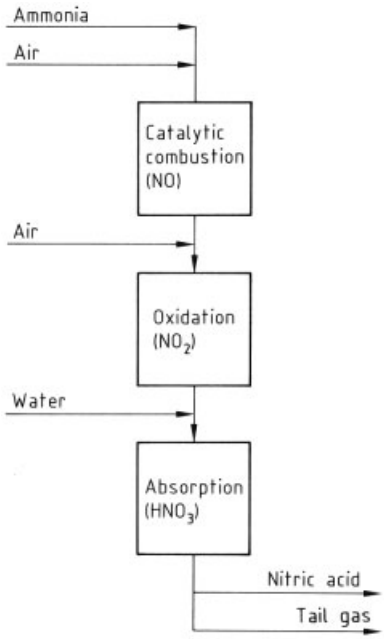

Figure 2. Ostwald process

The way in which these three steps are implemented characterizes the various nitric acid processes. In *monopressure (single-pressure) processes*, ammonia combustion and NO_x absorption take place at the same working pressure. These include medium-pressure (230 – 600 kPa) and high-pressure (700 – 1100 kPa) processes. Very few plants currently employ low pressure (100 – 220 kPa) for both combustion and absorption.

In *dual-pressure (split-pressure) processes*, the absorption pressure is higher than the combustion pressure. Modern dual-pressure plants feature combustion at 400 – 600 kPa and absorp-

tion at 900 – 1400 kPa. Some older plants still employ atmospheric combustion and medium-pressure absorption.

Intensive work on ammonia combustion as well as the oxidation and absorption of nitrogen oxides has been well documented since the early 1920s. Up to now the combustion of ammonia has been considered as a heterogeneous catalytic surface reaction with undesirable side reactions. The oxidation of nitrogen monoxide is a rare example of a homogeneous third-order gas-phase reaction. The absorption of NO_x is a heterogeneous reaction between fluid (i.e., gas and liquid) phases.

1.3.1. Oxidation of Ammonia

The oxidation of ammonia to nitrogen monoxide is described by the equation

$$4\,NH_3 + 5\,O_2 \longrightarrow 4\,NO + 6\,H_2O \quad \Delta H = -904\,\text{kJ/mol} \quad (4)$$

Over a suitable catalyst, 93 – 98 % of the feed ammonia is converted to nitrogen monoxide. The remainder participates in undesirable side reactions leading to dinitrogen monoxide (nitrous oxide, laughing gas)

$$4\,NH_3 + 4\,O_2 \longrightarrow 2\,N_2O + 6\,H_2O \quad \Delta H = -1105\,\text{kJ/mol} \quad (5)$$

and to nitrogen

$$4\,NH_3 + 3\,O_2 \longrightarrow 2\,N_2 + 6\,H_2O \quad \Delta H = -1268\,\text{kJ/mol} \quad (6)$$

Other undesirable reactions are the decomposition of the nitrogen monoxide product

$$2\,NO \longrightarrow N_2 + O_2 \quad \Delta H = -90\,\text{kJ/mol} \quad (7)$$

and its reaction with ammonia

$$4\,NH_3 + 6\,NO \longrightarrow 5\,N_2 + 6\,H_2O \quad \Delta H = -1808\,\text{kJ/mol} \quad (8)$$

Reaction Mechanism. Early theories of the mechanism for the oxidation of ammonia differed as to which intermediate product was formed. The classical theories were proposed by Andrussov (nitroxyl theory, 1926), Raschig (imide theory, 1927), and Bodenstein (hydroxylamine theory, 1935) [3]. According to the hydroxylamine theory, ammonia reacts with atomic oxygen adsorbed on the catalyst to yield hydroxylamine. This in turn reacts with molecular oxygen to form nitrous acid, which then dissociates to nitrogen monoxide and hydroxyl ions:

$$O_2 \rightleftharpoons 2\,O\ (\text{adsorb.})$$

$$O\ (\text{adsorb.}) + NH_3 \longrightarrow NH_2OH$$

$$NH_2OH + O_2 \longrightarrow HNO_2 + H_2O$$

$$HNO_2 \longrightarrow NO + OH^-$$

$$2\,OH^- \longrightarrow H_2O + O$$

In oxidation experiments under high vacuum, Bodenstein and coworkers detected both hydroxylamine and nitrous acid as intermediate products on the platinum catalyst. Later studies with a mass spectrometer showed that no such intermediate products are found above 400 °C [8]. In these tests, both molecular and atomic oxygen was found on the catalyst surface. A direct reaction between oxygen and ammonia was therefore postulated according to the following equations:

$$NH_3 + O_2 \longrightarrow NO + H_2O + H$$

$$NH_3 + O \longrightarrow NO + H_2 + H$$

In fast reactions such as the combustion of ammonia (the reaction time for Eq. 4 is ca. 10^{-11} s), however, the possible intermediates most likely cannot be detected because they are very short-lived and present in very low concentrations. Nowak [9] has compared the numerous theories of ammonia combustion. Pignet and Schmidt present a classical Langmuir – Hinshelwood treatment of the process [10]. The possibility that the combustion of ammonia may be considered as a homogeneous catalytic reaction is now being discussed.

At low temperature (200 – 400 °C), the rate of ammonia oxidation is limited by the reaction kinetics. The byproducts nitrogen, and especially dinitrogen monoxide, are formed preferentially. Between 400 and 600 °C, the reaction rate becomes limited by mass transfer, which is then dominant above 600 °C. Control by mass transfer relates to the diffusion rate of ammonia through the gas film to the catalyst surface. The product in this reaction regime is nitrogen monoxide.

The catalyst surface experiences a very low partial pressure of ammonia and a diminished partial pressure of oxygen due to mass-transfer limitations. The catalytic side reaction between ammonia and nitrogen monoxide is accelerated by an increase in the partial pressures of both species on the catalyst surface. If catalyst poisoning occurs,the number of active centers decreases and the partial pressure, particularly that of ammonia, increases at the remaining active centers. The result is increased formation of undesirable byproducts. In practice, this shift is offset by raising the temperature, i.e., the reaction is again displaced into the region limited by gas-film diffusion.

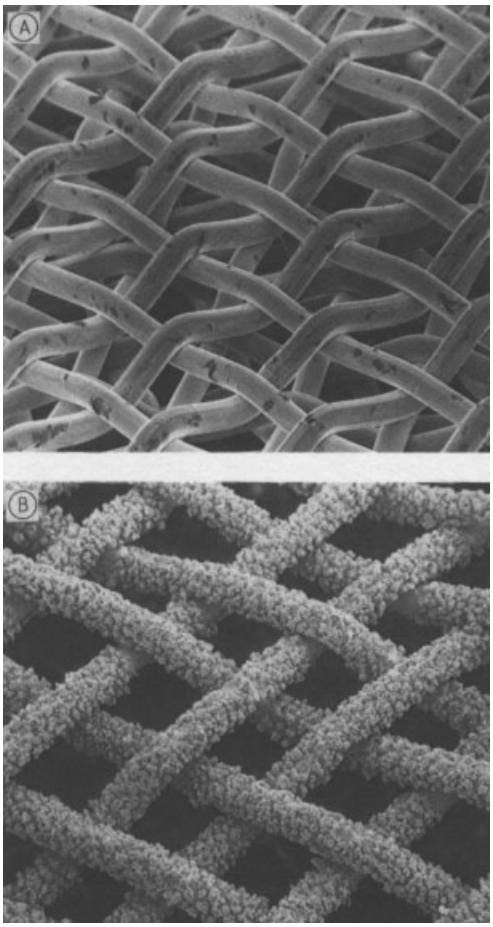

Figure 3. Photograph of platinum – rhodium gauze (Degussa, FRG) taken with a scanning electron microscope (enlargement 100 : 1)
A) Initial stage; B) Highly activated stage

Precious-metal catalysts are usually employed in the form of closely stacked, fine-mesh gauzes. At the beginning of a campaign they have a smooth surface and therefore only slightly limit mass transfer in the gas phase. This gives an initially lower yield of nitrogen monoxide. After a short time in service, however, the surface area of the catalyst increases greatly because of microstructural changes and interactions with volatilizing constituents of the catalyst (Fig. 3). Catalyst growth increases the limitation of mass transfer in the gas phase and thus the yield of nitrogen monoxide.

Reaction Engineering. From an engineering standpoint, the combustion of ammonia is one of the most efficient catalytic reactions (maximum possible conversion 98 %). According to the stoichiometry of Equation (4), the ammonia – air reaction mixture should contain 14.38 % ammonia. A lower ammonia – air ratio is, however, employed for a variety of reasons, the most important being that conversion decreases with too high a ratio; ammonia and air also form an explosive mixture. Because the mixing of ammonia and air in industrial practice is incomplete, locally higher ratios may occur. A ratio that includes a safety margin below the explosion limit is therefore necessary. The explosion limit declines with pressure so that high-pressure burners can operate with up to 11 % ammonia, whereas 13.5 % ammonia is possible in low-pressure systems. The explosion limit also depends on the flow velocity (Table 2, see also [3]) and the water content of the reaction mixture [11].

Table 2. Explosion limits of ammonia – air mixtures at atmospheric pressure

Flow velocity, m/s	Lower explosion limit, % NH_3	Upper explosion limit, % NH_3
0	15.5	27.5
3	28	38
5 – 8	30	40
12	32	37
14	none	none

According to Le Chatelier's principle, the increase in volume in Equation (4) implies that conversion declines as pressure rises (Fig. 4, see next page).

If the gas velocities are too high or the number of catalyst gauzes is insufficient, conversion to

nitrogen monoxide decreases because of the slip (leakage) of ammonia which reacts with it according to Equation (8). If the gas velocities are too low or too many gauzes are used, decomposition of nitrogen monoxide according to Equation (7) is promoted. Table 3 gives typical gas velocities.

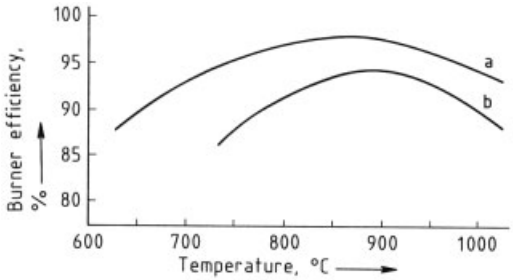

Figure 4. Conversion of ammonia to nitrogen monoxide on a platinum gauze as a funtion of temperature [5]
a) 100 kPa; b) 400 kPa

Table 3. Typical design data for ammonia burners

Pressure, MPa	Number of gauzes	Gas velocity, m/s	Reaction temperature, °C	Catalyst loss, g/t HNO_3
0.1 – 0.2	3 – 5	0.4 – 1.0	840 – 850	0.05 – 0.10
0.3 – 0.7	6 – 10	1 – 3	880 – 900	0.15 – 0.20
0.8 – 1.2	20 – 50	2 – 4	900 – 950	0.25 – 0.50

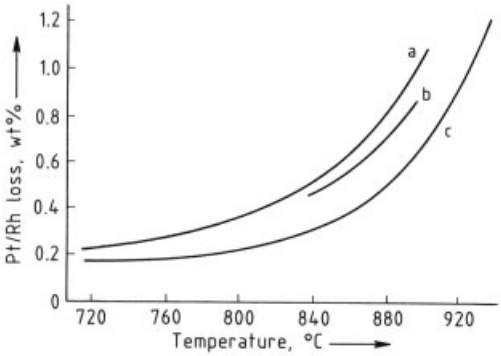

Figure 5. Losses of precious metals in the combustion of ammonia to nitrogen monoxide as a function of temperature and catalyst composition [5]
a) Pt; b) Pt – Rh 98/2; c) Pt – Rh 90/10

A high reaction temperature promotes ammonia combustion (Fig. 4), but high temperature decreases the conversion. Maximum temperatures up to 950 °C have been realized, but catalyst losses, chiefly due to vaporization, increase greatly (Fig. 5). The usual way to maintain the conversion toward the end of a campaign in spite of catalyst deactivation and losses, is to raise the temperature. The ammonia – air ratio is directly related to the reaction temperature: a 1 % increase in the proportion of ammonia increases the temperature by ca. 68 K. If the temperature is too low, conversion to nitrogen monoxide decreases. The reaction temperature also depends on the temperature of the inlet gas mixture. The following formula provides an approximate value in °C:

$$t_{\text{reactor}} = t_{\text{gas mixture}} + r \cdot 68$$
$$\text{where } r = \text{NH}_3 / (\text{NH}_3 + \text{air}) \quad (9)$$

Catalysts [12]. Since the introduction of the industrial production of nitric acid by the Ostwald process, several hundred materials have been tested as catalysts for ammonia combustion. Platinum catalysts have proved most suitable and are used almost exclusively today. Oxides of non-noble metals can also be employed; these cost less than platinum catalysts, but conversion is lower. Another disadvantage of non-noble-metal catalysts is that they are more quickly poisoned [13].

The usual form of catalyst is a fine-mesh gauze (standardized at 1024 mesh/cm^2). Manufacturers offer wire diameters of 0.060 and 0.076 mm. The Johnson Matthey "Tailored Pack" combines gauzes made of different wire diameters [14]. The upstream gauzes have a wire diameter of 0.076 mm and the downstream ones 0.060 mm to optimize the catalyst surface area.

High-pressure plants also utilize "fiber packs" in which a fraction of the fibrous catalyst material is held between two gauzes [15]. The packs have the advantage of reduced platinum loss and thus longer run times. " Fiber packs" have not yet been used in medium-pressure ammonia combustion because fewer gauzes are used in medium-pressure than in high-pressure processes (see Table 3), so the lower pressure drop results in poorer distribution.

To reduce catalyst consumption, particularly in gauzes at the bottom of the reactor, ceramic grids coated with platinum – rhodium have also been developed [16]. Platinum is usually alloyed with 5 – 10 % rhodium to improve its strength and to reduce catalyst loss. During combustion,

the metal surface becomes enriched in rhodium, thus improving catalytic activity [17]. Because rhodium is more expensive than platinum, a rhodium content of 5 – 10 % has proved optimal. Figure 6 shows the efficiency of conversion to nitrogen monoxide and the platinum loss as a function of the rhodium content of the alloy [18].

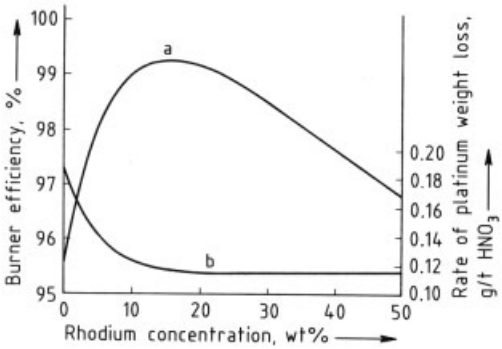

Figure 6. Conversion of ammonia to nitrogen monoxide on platinum gauze (a) and platinum losses (b) as a function of catalyst composition

If a low reaction temperature (< ca. 800 °C) is unavoidable, a pure platinum catalyst should be employed, because otherwise rhodium(III) oxide would accumulate at the catalyst surface and decrease catalytic activity [18]. Palladium is also used in catalyst alloys. Laboratory studies on the catalytic activity of 90 % Pt – 10 %Rh and 90 %Pt – 5 %Rh – 5 %Pd gauzes reveal no great differences in nitrogen monoxide yield [19]. The ternary catalyst is of economic interest because palladium costs much less than platinum or rhodium.

Platinum gauzes are pretreated by the manufacturer (Degussa and Heraeus, FRG; Johnson Matthey and Engelhard, UK) so that they do not need to be activated after installation in the burner. The reaction is ignited with a hydrogen flame until gauzes glow red. After the reaction has been initiated, crystal growth rapidly increases the surface area. The highest conversion on the catalyst is obtained after a few days use. Specially pretreated platinum gauzes that reach maximum activity after just a few hours are also on the market [20]. Start-up problems with platinum gauzes are discussed in [21]. The prolonged decline in activity after the conversion maximum is due to catalyst poisoning and platinum loss. Catalyst poisons include iron oxide (rust) and dust in the process air, which is especially concentrated in the vicinity of fertilizer plants. Any material deposited on the catalyst surface acts as a poison. A multistage filter is used to prevent these substances from reaching the catalyst (see Section 1.3.3.1).

Platinum loss during operation is caused by vaporization and abrasion. Vaporization loss predominates, but if heavy vaporization loss weakens the structure then abrasion can substantially lower burner efficiency. Platinum loss due to vaporization is thought to involve the formation of short-lived platinum dioxide:

$$Pt + O_2 \longrightarrow PtO_2(g)$$

Mechanical losses are accelerated by impurities in the reaction mixture, which can lead to the formation of whiskers at the grain boundaries. The whiskers are more easily broken off by the flowing gas.

Recovery of Precious Metals. Precious-metal losses from the catalyst have a significant effect on the operating costs of a nitric acid plant. The approximate losses and resulting costs can be determined from the ammonia combustion temperature and the loading. Losses increase with temperature and loading; however, smaller losses may be measured at higher loadings because a more uniform incoming flow causes less movement of the gauzes relative to one another and thus reduces mechanical losses. Table 3 lists typical loss rates. Platinum recovery systems are installed in most nitric acid plants to improve process economics.

Mechanical Filters. Mechanical filters are made of glass, mineral wool, or ceramic or asbestos fibers. They are normally installed where the gas temperature is < 400 °C. Their most important drawback is their large pressure drop, which can reach 25 kPa between filter changes.

Mechanical filters recover 10 – 20 % platinum, but recovery rates up to 50 % have been reported [22]. These filters have the advantage of low capital costs and the disadvantage of high operating costs resulting from the large pressure drop across the filter. For economic reasons, they are only employed in mono-high-pressure plants.

In low-pressure burners, marble chips 3–5 mm in length were used immediately downstream of the catalyst gauzes [23]. At high reaction temperature, the marble is converted to calcium oxide, whose high absorption capacity for platinum and rhodium permits up to 94 % recovery. During reactor shutdown, however, the calcium oxide must be protected against moisture.

Recovery Gauzes. A technically elegant approach to platinum recovery is to install, just downstream of the catalyst gauzes, a material that is stable at high temperature and that can absorb platinum oxide vapor and form an alloy with it. The first recovery gauzes in production plants (supplied by Degussa in 1968) consisted of a gold – palladium alloy (20/80). Nowadays most such gauzes are gold-free but have a low content of nickel (ca. 5 %) to enhance their strength. The gauzes have 100 – 1350 mesh/cm^2 and are made of wire with a diameter of 0.2 – 0.06 mm. Several recovery gauzes with different specifications are usually installed. Loss of platinum from the catalyst is greater than that of rhodium; for example, the calculated loss composition for a 90/10 Pt – Rh gauze is 95/5. A "recovery factor" can be determined for each recovery gauze, indicating how much of the available platinum is retained by the gauze. This recovery factor depends on the specifications of the gauze, the operating pressure, and the loading. Figure 7 shows the dependence of the recovery efficiency on wire diameter and mesh fineness [24]. A method for selecting the optimum platinum recovery gauze system for a particular nitric acid plant is described in [25].

The captured platinum diffuses into the palladium gauze and forms an alloy. The topmost gauze is often designed to reach saturation in one campaign, usually 180 d. Determining the economic optimum for a recovery system (the most important criterion is the weight of palladium required) necessitates taking into account the fact that palladium is lost. This loss corresponds to about one-third of platinum recovery.

"Getter" systems can recover more than 80 % of the platinum lost from the catalyst. The mechanism of rhodium recovery has not yet been fully elucidated. Up to 30 % of rhodium from the catalyst is captured by recovery systems.

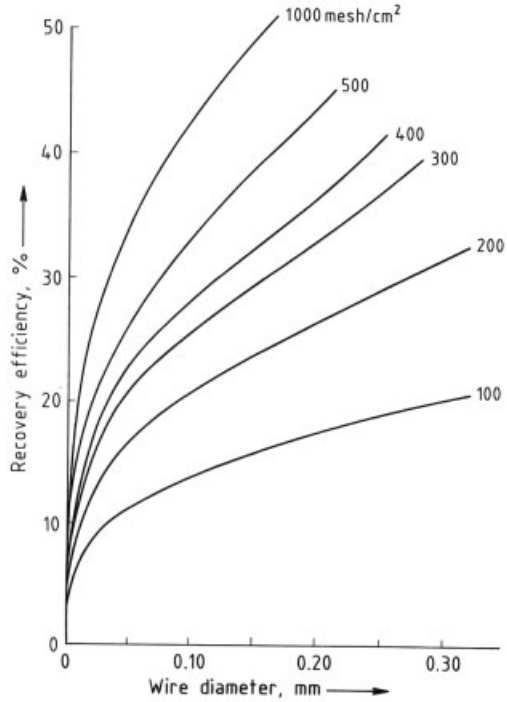

Figure 7. Performance of a palladium-based recovery gauze as a function of mesh size and wire diameter

1.3.2. Oxidation and Absorption of Nitrogen Oxides

After the combustion of ammonia, the nitrogen monoxide product gas is cooled en route to absorption and, if necessary, compressed. As a result, part of the nitrogen monoxide is oxidized to nitrogen dioxide or dinitrogen tetroxide (Eq. 2), which is converted to nitric acid by absorption in water (Eq. 3).

Absorption is usually performed in plate columns where the nitrous gases are converted to nitric acid in the liquid phase on the plates. Nitrogen monoxide is constantly reformed in this step and prevents complete absorption of the inlet gases. Oxidation of nitrogen monoxide occurs mainly in the gas phases between the plates, but also in the gas phase of the bubble layer on them.

Gas Phase. The gas phase is the site of nitrogen monoxide oxidation

$$2\,NO + O_2 \longrightarrow 2\,NO_2 \quad \Delta H = -1127\,\text{kJ/mol} \qquad (10)$$

as well as nitrogen dioxide dimerization

$$2\,NO_2 \rightleftharpoons N_2O_4 \;\; \Delta H = -58.08\,\text{kJ/mol} \tag{11}$$

and dinitrogen trioxide formation

$$NO + NO_2 \rightleftharpoons N_2O_3 \;\; \Delta H = -40\,\text{kJ/mol} \tag{12}$$

Reaction of the oxidized nitrogen oxides with water vapor leads to formation of nitrous and nitric acids. These reactions are of secondary importance from an engineering standpoint.

To describe the kinetics of nitrogen monoxide oxidation (Eq. 10), a third-order rate equation is used [26]:

$$r = \frac{k_p}{RT} p_{NO}^2 p_{O2} \tag{13}$$

where

r	= reaction rate, kmol m^{-3} s^{-1}
k_p	= reaction rate constant, atm^{-2} s^{-1}
R	= universal gas constant, m^3 atm kmol^{-1} K^{-1}
T	= temperature, K
p	= partial pressure, atm

The rate constant is defined by

$$\log k_p = \frac{652.1}{T} - 1.0366 \tag{14}$$

This reaction is unusual because it goes to the right faster at low temperature than at high temperature, i.e., the reaction rate has a negative temperature coefficient.

Because equilibrium in the nitrogen dioxide dimerization system is reached very quickly, an equilibrium formula

$$r = \frac{k_p}{RT}\left(p_{NO2}^2 - \frac{p_{N2O4}}{K_p}\right) \tag{15}$$

can be assumed where K_p is the equilibrium constant. The dimerization rate is virtually independent of temperature [27], so the rate constant at 25 °C

$$k_p = 5.7 \times 10^5\,\text{atm}^{-1}\text{s}^{-1} \tag{16}$$

can be used at all temperatures of technical interest.

Hoftyzer and Kwanten [28] give the following equation for the $NO_2 - N_2O_4$ equilibrium constant:

$$K_p = \frac{p_{N2O4}}{p_{NO2}^2} = 0.698 \times 10^{-9} \exp\left(\frac{6866}{T}\right) \tag{17}$$

An equilibrium formula can also be written for dinitrogen trioxide formation:

$$r = \frac{k_p}{RT}\left(p_{NO}\,p_{NO2} - \frac{p_{N2O3}}{K_p}\right) \tag{18}$$

According to [27], the equilibrium constant can be determined as

$$K_p == \frac{p_{N2O3}}{p_{NO}\,p_{NO2}} = 65.3 \times 10^{-9} \exp\left(\frac{4740}{T}\right) \tag{19}$$

Liquid Phase. The large number of reactive components in the gas phase (NO, NO_2, N_2O_3, N_2O_4) means that the reaction model for absorption is complicated. Figure 8 (see next page) shows a sophisticated model devised by Hoftyzer and Kwanten [28].

The main route for the formation of nitric acid in this model involves two steps in the liquid phase. First, dissolved dinitrogen tetroxide reacts with water, yielding nitric and nitrous acids:

$$N_2O_4 + H_2O \longrightarrow HNO_3 + HNO_2 \quad \Delta H = -87.0\,\text{kJ/mol} \tag{20}$$

Nitrous acid then dissociates to nitric acid, water, and nitrogen monoxide, the latter being transported across the interface into the bulk gas:

$$3\,HNO_2 \longrightarrow HNO_3 + H_2O + 2\,NO \quad \Delta H = -15.3\,\text{kJ/mol} \tag{21}$$

According to Andrew and Hanson [29], a first-order equation can be written for the rate of dinitrogen tetroxide hydrolysis:

$$r = k\,c_{N2O4} \tag{22}$$

where $c_{N_2O_4}$ is the concentration of dinitrogen tetroxide in kmol/m^3. The rate constant is given by

$$\log k = -\frac{4139}{T} + 16.3415 \tag{23}$$

The kinetics of nitrous acid dissociation were studied by Abel and Schmid [30], with the following result:

$$r = k\,\frac{c_{HNO2}^4}{p_{NO}^2} = \frac{k}{H_{NO}^2}\,\frac{c_{HNO2}^4}{c_{NO}^2} \tag{24}$$

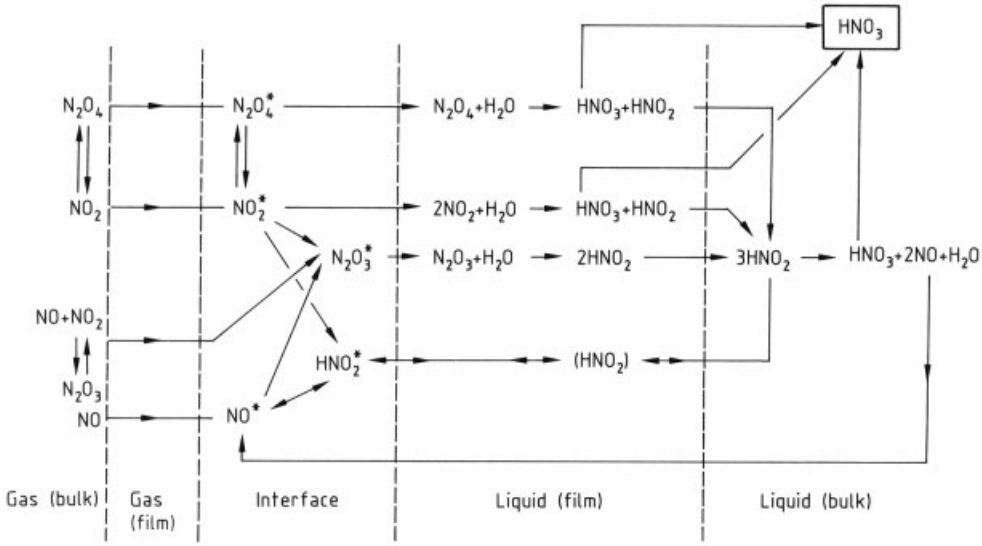

Figure 8. Nonstoichiometric model of absorption of nitrogen oxides in water [28]

where H_{NO} is the Henry coefficient for nitrogen monoxide in m^3atm/kmol and c ist the concentration in kmol/m^3. The rate constant is defined by

$$\log k = -\frac{6200}{T} + 20.1979 \tag{25}$$

Equations (23) and (25) were derived by Hoffmann and Emig [31] and are based on experimental data from [32].

Mass Transfer. For more than 50 years scientists have studied the transport of nitrogen oxides in water and in dilute and concentrated nitric acid. A variety of mechanisms have been proposed depending on the gas composition and the acid concentration. These are partly in agreement with the absorption model of Figure 8.

In the NO_x absorption region important for acid formation, dinitrogen tetroxide transport is considered to be the rate-limiting step. The quantity of dinitrogen tetroxide transported from the bulk gas to the interface depends chiefly on the $NO_2 - N_2O_4$ equilibrium [28, 33]:

$$J_{N2O_4} = \frac{k_{gNO2}}{2RT}\left(p^{o}_{NO2} - p^{i}_{NO2}\right) + \frac{k_{gN2O_4}}{RT}\left(p^{o}_{N2O_4} - p^{i}_{N2O_4}\right) \tag{26}$$

where

J = absorption rate, kmol m^{-2} s^{-1}
k_g = gas-side mass-transfer coefficient, m/s
p^o = partial pressure in the bulk gas, atm
p^i = partial pressure at the interface, atm

At higher NO_2 /N_2O_4 concentrations in the reaction gas, primarily dinitrogen tetroxide crosses the interface and reacts in the fast first-order reaction (Eq. 20) to form nitric and nitrous acids. Nitrous acid dissociates (Eq. 21), and the resulting nitrogen monoxide is transported back into the gas space.

The following equation can be written for the absorption rate of dinitrogen tetroxide [34]:

$$J_{N2O_4} = H_{N2O_4} p_{N2O_4} \sqrt{k D_{N2O_4}} \tag{27}$$

where

H = Henry coefficient, m^3 atm/kmol
k = rate constant, s^{-1}
D = diffusion constant, m^2 /s

Numerous measurements in laboratory absorbers support this mechanism. Interpretations of measurements in absorption columns have given the following correlations for the mass-transfer coefficients [35]:

1) Bubble-cap trays

$$H_{N_2O_4}\sqrt{k\,D_{N_2O_4}} = \exp\left(-\frac{1500}{T} - 4.3790 - 23.279\,W_{HNO_3} + 130.42\,W^2_{HNO_3} - 370.87\,W^3_{HNO_3} + 486.94\,W^4_{HNO_3} - 236.54\,W^5_{HNO_3}\right) \text{ kmol m}^{-2}\text{ kPa}^{-1}\text{ s}^{-1} \tag{28}$$

2) Sieve trays

For $W_{HNO_3} > 0.05$

$$H_{N_2O_4}\sqrt{k\,D_{N_2O_4}} = \exp\left(-\frac{1500}{T} - 2.7648 - 39.614\,W_{HNO_3} + 181.98\,W^2_{HNO_3} - 429.65\,W^3_{HNO_3} + 496.99\,W^4_{HNO_3} - 223.24\,W^5_{HNO_3}\right) \quad \text{kmol m}^{-2}\,\text{kPa}^{-1}\,\text{s}^{-1} \tag{29a}$$

For $W_{HNO_3} < 0.05$

$$H_{N_2O_4}\sqrt{k\,D_{N_2O_4}} = \exp\left(-\frac{1500}{T} + 0.2548 - 315.73\,W_{HNO_3} + 9256.2\,W^2_{HNO_3} - 95.602\,W^3_{HNO_3}\right) \quad \text{kmol m}^{-2}\,\text{kPa}^{-1}\,\text{s}^{-1} \tag{29b}$$

where

H = Henry coefficient, m^3kPa/kmol
T = temperature, K
W = mass fraction

The absorption of $NO_2 - N_2O_4$ in concentrated nitric acid can be treated as pure physical absorption [36]. Deviations from the described chemisorption also occur at low gas-phase $NO_2 - N_2O_4$ concentrations [29].

Absorption Equilibrium. Equilibrium considerations allow determination of the maximum possible acid concentration for a given composition of the inlet gas. For example, suppose the overall reaction is

$$3\,NO_2\,(g) + H_2O\,(l) \rightleftharpoons 2HNO_3(l) + NO\,(g)$$
$$\Delta H = -72.8\,\text{kJ/mol} \tag{30}$$

and the equilibrium constant is defined as

$$K_p = \frac{p^2_{HNO3}\,p_{NO}}{p^3_{NO2}\,p_{H2O}} \tag{31}$$

with the partial pressures of nitrogen monoxide (p_{NO}) and nitrogen dioxide (p_{NO_2}) in the gas phase and the vapor pressures of nitric acid (p_{HNO_3}) and water (p_{H_2O}) over the dilute acid. In industrial absorption the equilibrium constant depends only on temperature. The equilibrium constant is conventionally split into two terms:

$$K_p = K_1 \cdot K_2$$
$$\text{where}\, K_1 = \frac{p_{NO}}{p^3_{NO2}} \text{ and } K_2 = \frac{p^2_{HNO3}}{p_{H2O}} \tag{32}$$

K_2 depends on the acid concentration and the temperature. Gases (N_2O_4, N_2O_3, HNO_2) dissolved in the dilute acid affect the activities of

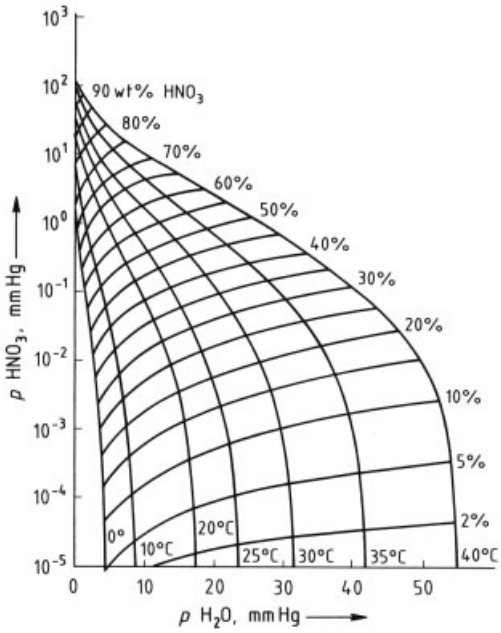

Figure 9. Vapor pressures of nitric acid and water as a function of acid strength and temperature [28]
1mm Hg = 133.2 mPa

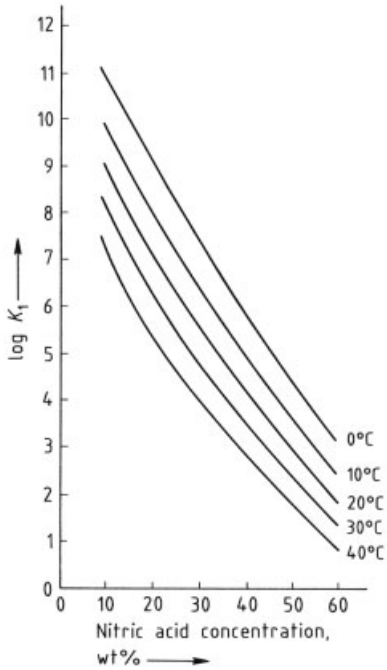

Figure 10. Value of K_1 from Equation (32) as a function of acid concentration and temperature [28]

nitric acid and water. Up to a nitric acid concentration of 65 wt % and an inlet nitrogen oxide partial pressure of 100 kPa, however, the vapor pressures from the binary system $HNO_3 - H_2O$ can be used to determine K_2 (Fig. 9).

Figure 10 shows how K_1 depends on the acid concentration and the temperature. This parameter is particularly interesting from an engineering standpoint because it allows the acid strength to be determined if the gas concentration is known.

If the equilibrium is based on the overall reaction

$$3/2\ N_2O_4(g) + H_2O(l) \rightleftharpoons 2\ HNO_3(l) + NO_3\ (g) \qquad \Delta H = -75\ \text{kJ/mol} \tag{33}$$

the equilibrium constant is

$$K_p = K_1 \cdot K_2 \quad \text{where } K_1 = \frac{p_{NO}}{p_{N_2O_4}^{3/2}} \text{ and } K_2 = \frac{p_{HNO_3}^2}{p_{H_2O}} \tag{34}$$

Again Figure 9 gives the K_2 values. In this formulation, K_1 is a function of the acid concentration and does not depend on temperature (Fig. 11). The curve of Figure 11 can be fitted by an empirical equation [37]:

$$\log K_1 = 7.412 - 20.29 W_{HNO_3} + 32.47 W_{HNO_3}^2 - 30.87 W_{HNO_3}^3 \tag{35}$$

where W denotes the mass fraction.

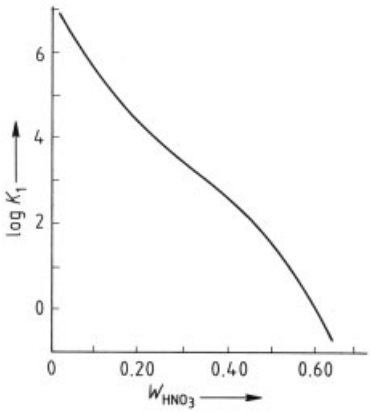

Figure 11. Value of K_1 from Equation (34) as a function of acid concentration

Another way of describing the absorption equilibrium is to start with the hypothetical component NO_2 denoted as "chemical NO_2":

$$p_{\tilde{NO}_2} = p_{NO_2} + 2p_{N_2O_4} \tag{36}$$

The nitrogen dioxides, considered as completely dissociated, are assumed to react according to

$$3\ \tilde{NO}_2(g) + H_2O(l) \rightleftharpoons 2\ HNO_3(l) + NO(g) \tag{37}$$

This hypothetical reaction leads to the equilibrium relation

$$K_p = K_1 \cdot K_2 \quad \text{where } K_1 = \frac{p_{NO}}{p_{\tilde{NO}_2}^3} \text{ and } K_2 = \frac{p_{HNO_3}^2}{p_{H_2O}} \tag{38}$$

The equilibrium constant K_p can be estimated with the following equation [38]:

$$\log K_p = -7.35 + \frac{2.64}{T} \tag{39}$$

Then K_1 can be determined from K_p, and the K_2 values obtained from Figure 9. The parameter K_1 determines the equilibrium curves for constant temperature and constant acid concentration when p_{NO_2} is plotted against p_{NO} (Fig. 12). The K_1 values are necessary to calculate the absorption of nitrous gases according to Toniolo [39]. The application of the Toniolo diagram is described in [40].

Modeling of the Absorption Tower. The literature offers many methods for calculating the absorption of nitrous gases; they can be classified according to the number of reactions, the reaction kinetics, or the type of mathematical method:

1) *Number of Reactions*. The simplest calculation model uses Equation (30) to describe the events on an absorption plate and a kinetic formula to describe nitrogen monoxide oxidation between two oxidation plates [41]. Extended models [42–45] also take into account the dimerization of nitrogen dioxide to dinitrogen tetroxide. In this way it is possible to describe nitric acid formation not just via nitrogen dioxide but also via dinitrogen tetroxide [46–49].
   Models that include the formation and dissociation of nitrous acid in the liquid phase give the best description of known chemical events [31, 50]. A model that permits nitrous acid formation in the gas phase has also been proposed [51].

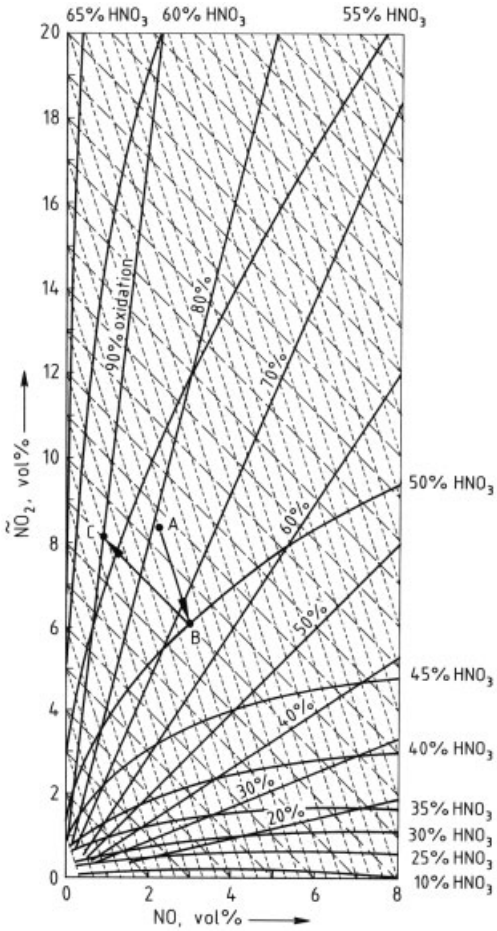

Figure 12. Toniolo diagram for an absorption tower [39]
A → B absorption process; B → C oxidation process

2) *Reaction Kinetics*. Most design computations use equilibrium equations for the basic reactions. These static methods have the drawback of requiring the introduction of efficiency factors [51] that take into account limitation of transport of the reactants across the interface. Dynamic methods do not need empirical efficiency factors because they describe mass transfer in terms of physical and chemical laws [28,31,50,52].
3) *Mathematical Methods*. In view of the large number of nonlinear equations to be solved in both equilibrium and dynamic models, plate-to-plate calculations are generally described in the literature. The relaxation method [31] and a special stage-to-stage calculation [52], however, have also been used for the design of absorption towers.

By way of example, an equilibrium model and a transport model are described in more detail.

Equilibrium Model.

The equilibrium model is based on Equations (11) and (33). If a gas of known composition is fed to the plate and an acid of known strength leaves the plate, the calculation begins at the bottom of the tower. If the gas throughput is taken to be constant (because of the high content of inerts), the following mass balance is obtained for the n-th plate:

$$3\left(y_{n,\mathrm{NO}} - y_{n-1,\mathrm{NO}}\right) = \left(y_{n-1,\mathrm{NO_2}} + 2y_{n-1,\mathrm{N_2O_4}}\right) - \left(y_{n,\mathrm{NO_2}} + 2y_{n,\mathrm{N_2O_4}}\right) \tag{40}$$

where y is the mole fraction in the vapor phase.

Equilibrium Equations (34) for K_{p^1} and (17) for K_{p^2} give the equation

$$\begin{aligned} &a y_{n,\mathrm{NO_2}}^3 + b y_{n,\mathrm{NO_2}}^2 + y_{n,\mathrm{NO_2}} + d = 0 \\ &\text{with } a = 3\left(K_{p_1}\sqrt{p}\right)\left(K_{p_2} p\right)^{1.5} \\ &b = 2\left(K_{p_2} p\right) \\ &d = -\left(3y_{n-1,\mathrm{NO}} + y_{n-1,\mathrm{NO_2}} + 2y_{n-1,\mathrm{N_2O_4}}\right) \end{aligned} \tag{41}$$

This formula allows the compositions of the gas leaving the n-th plate and the acid arriving at the n-th plate to be determined.

Before the $(n+1)$-th plate is calculated, the oxidation of nitrogen monoxide must be determined. The flow behavior of the nitrous gases in an absorption tower is not known exactly, so the oxidation space can be described as either a tubular or a stirred-tank reactor. Actual conditions probably lie somewhere between the two [28], but trends can be identified on the basis of plate type. The stirred-tank model is preferred for bubble-cap trays [47], but flow-through conditions can be expected in the case of sieve plates. Because virtually all absorption towers today are fitted with sieve plates, only these are discussed.

When the reaction is carried out isothermally, integration of the rate equation

$$-\frac{\mathrm{d}p_{\mathrm{NO}}}{\mathrm{d}t} = k p_{\mathrm{NO}}^2 p_{\mathrm{O2}} \tag{42}$$

gives a solution

$$\theta = \frac{2}{k\left(2p_{\mathrm{O2}} - p_{\mathrm{NO}}\right)} \cdot \left\{\frac{x}{p_{\mathrm{NO}}(1-x)} - \frac{1}{2p_{\mathrm{O2}} - p_{\mathrm{NO}}} \ln\left[\frac{2p_{\mathrm{O2}} - p_{\mathrm{NO}}x}{2p_{\mathrm{O2}}(1-x)}\right]\right\} \tag{43}$$

where

θ = residence time, s
k = rate constant, $atm^{-2}\,s^{-1}$
p = partial pressure, atm
t = time, s
x = fraction of oxidized nitrogen monoxide

in which x must be determined by iteration from the residence time θ and the inlet partial pressures of nitrogen monoxide (p_{NO}) and oxygen (p_{O_2}) [47]. The partial pressures of the other components leaving the reactor can be calculated from the fraction x of oxidized nitrogen monoxide.

If an excess of oxygen is assumed to be present, the rate equation can be simplified [53] as

$$-\frac{dp_{NO}}{dt}=\bar{k}p_{NO}^2 \tag{44}$$

so that the partial pressure of nitrogen monoxide at the reactor outlet (p'_{NO}) can be determined:

$$\frac{1}{p'_{NO}}=\frac{1}{p_{NO}}+(2p_{O2}-p_{NO})\frac{k\theta}{2} \tag{45}$$

Because the oxidation reaction is highly exothermic, the tubular reactor can be designed for isothermal operation only when conversion is very low (e.g., when the tail gas contains little NO). Otherwise, adiabatic reaction conditions must be assumed. To describe the reaction under adiabatic conditions the oxidation space is divided into n layers, in each of which the reaction is isothermal. The temperature increase from one layer to the next is given by [33]

$$\Delta T=1/2\frac{C_{NO,0}\,(-\Delta H_R)}{\varrho_0\bar{C}_p}\frac{G_{NO,0}-G_{NO}}{G_{NO,0}} \tag{46}$$

where

$C_{NO,0}$ = initial concentration of NO, $kmol/m^3$
$\bar{C}_p$ = molar heat capacity, $kJ\,kmol^{-1}\,K^{-1}$
G_{NO} = molar gas flow rate, kmol/s
$G_{NO,0}$ = initial gas flow rate, kmol/s
ΔH_R = reaction enthalpy, kJ/kmol
ϱ_0 = initial density, kg/m^3

Transport Model. A transport model for the absorption tower is based on a series of units, each containing a bubble- column reactor and an adiabatic plug flow reactor (Fig. 13) [52]. The bubble-column reactor, modeled as two ideal stirred-tank reactors, simulates the liquid with gas bubbles on a tray (absorption space). The adiabatic plug flow reactor simulates oxidation between plates (oxidation space).

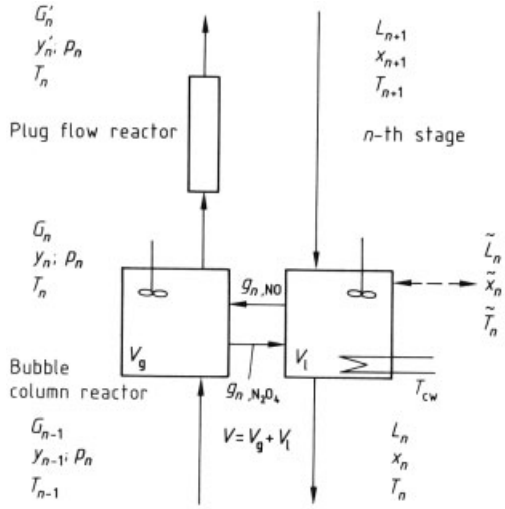

Figure 13. Hydrodynamic model of absorption tower
g = molar gas flow rate of a component; G = molar flow rate of gas; L = molar flow rate of liquid; p = pressure; T = temperature; T_{cw} = temperature of cooling water; V = volume; x = mole fraction in liquid phase; y = mole fraction in gas phase

The balances are carried out with the equations

$$G_n y_{n,i}-G_{n-1}y_{n-1,i}-k_{g,i}aV\frac{p_n}{RT_n}\left(y^*_{n,i}-y_{n,i}\right)=V_g\sum_{j=1}^{M_g}\nu_{g,ij}r_{g,j} \tag{47}$$

$i = N_2,\ O_2,\ NO,\ NO_2,\ N_2O_4$ and $\sum_{i=1}^{k} y_{n,i}=1$ for the components in the gas phase (subscript g) where

a = interfacial area per unit volume, m^2/m^3
G = molar gas flow rate, kmol/s
k_g = gas-side mass-transfer coefficient
M_g = number of reactions in the gas phase
n = plate number
r = reaction rate, $kmol\,m^{-3}\,s^{-1}$
V = volume of bubble layer
V_g = volume of gas phase in bubble layer
y_i = mole fraction of i-th component in liquid phase

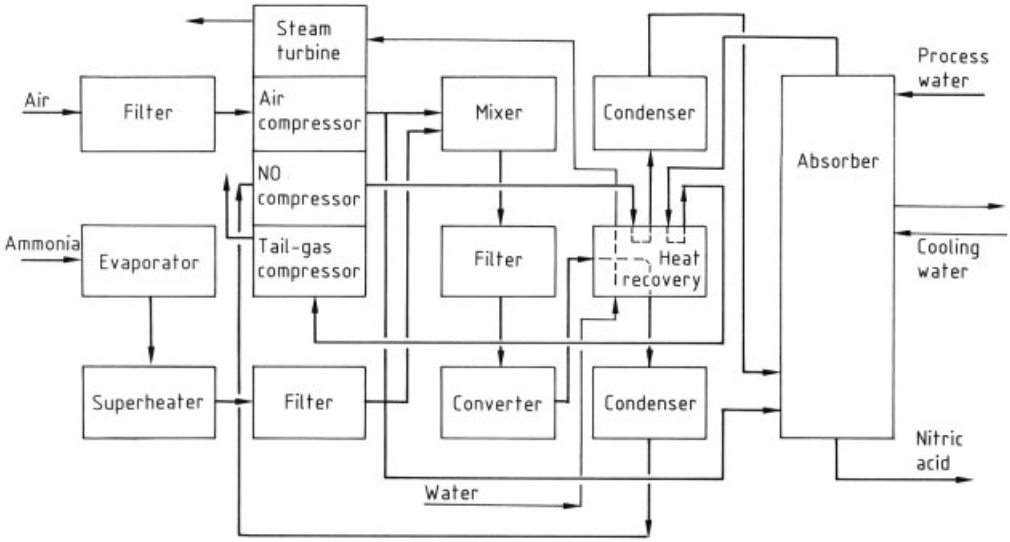

Figure 14. Equipment needed in a nitric acid plant

$y*_i$ = mole fraction of i-th component in liquid phase at the phase boundary
ν = stoichiometric coefficient

Analogously for the liquid phase (subscript l)

$$L_n x_{n,i} - L_{n+1} x_{n+1,i} + k_{l,i} a V c_n \left(x^*_{n,i} - x_{n,i}\right) \pm \tilde{L}_n \tilde{x}_{n,i} = V_l \sum_{j=1}^{M_l} \nu_{l,ij} r_{l,j} \quad (48)$$

i = NO, N_2O_4, HNO_2, HNO_3, H_2O and $\sum_{i=1}^{k} x_{n,i} = 1$
where

c = molar concentration, kmol/m^3
L = molar liquid flow rate, kmol/s
x = mole fraction in liquid phase
$\tilde{L}$ = side stream
M_l = number of reactions in the liquid phase

Thermodynamic equilibrium is assumed at the interface,

$$y^*_{n,i} = H_{n,i} x^*_{n,i} \quad (49)$$

where H is the Henry coefficient and the formula

$$k_{g,i} \frac{p_n}{RT_n} \left(y_{n,i} - y^*_{n,i}\right) = k_{l,i} c_n \left(x^*_{n,i} - x_{n,i}\right) \quad i = \mathrm{NO}, \mathrm{N_2O_4} \quad (50)$$

is used for dynamic mass transfer between phases. A separate heat balance is performed for each phase, so that the temperature may differ in both phases of a single stage. The mathematical treatment of this model is described in [52].

1.3.3. Equipment

Figure 14 summarizes the equipment needed to implement the three chemical steps involved in nitric acid production (Fig. 2). In addition to producing nitric acid, a nitric acid plant also generates steam or mechanical energy from the chemical energy liberated.

1.3.3.1. Filters and Mixers

To obtain the highest possible conversion of ammonia on the catalyst and a long catalyst life, the starting materials (air and ammonia) must be purified. The air must be very clean to prevent catalyst poisoning (page 6).

Air Filter. A multistage filter is normally used on the air intake to a nitric acid plant and should remove 99.9 % of all particles larger than 0.5 µm.

Typical filter media are plastic and glass fibers. Filter frames should be made of stainless steel. Air filters must be replaced regularly because filter bags can tear or filter mats can be-

come overloaded and cause an excessive pressure drop. Filter life depends on the particulate load in the air (typically 0.8 mg/m^3). In areas where sandstorms may occur (maximum loading 500 mg/m^3), the use of a sand separator (a centrifugal collector) as a prefilter is recommended.

Ammonia Filters. The liquid ammonia filter removes solid contaminants, especially small rust particles; 99.9 % of all particles larger than 3 µm should be eliminated. Selection of a liquid ammonia filter requires consideration of the fact that traces of oils and chlorides are also present in ammonia. Proven filter materials are Teflon and sintered metals. Polypropylene and ceramic filter candles (cartridges) are also employed. Magnetic filters have found some use but have limited capacity.

Filtration of ammonia gas should remove 99.9 % of oil and solid particles larger than 0.5 µm. The principal filter media are glass fibers, sintered metals, and ceramics. The pressure drop is up to 10 kPa, depending on filter type.

Filters for the Mixed Gas. The mixed-gas filter provides final cleaning of the ammonia–air mixture and improves the mixing of the air and ammonia; it should remove 99.8 % of all particles larger than 1.5 µm. Contaminants in the gas mixture not only originate externally (in the process air and ammonia) but are also formed by corrosion inside the system (rust). The gas mixture filter should therefore be installed as near as possible to the burner. Ceramic filter cartridges are generally used in which silicon dioxide is the dominant constituent. For a filter cartridge with a surface loading of 250 m^3/m^2, the maximum pressure drop in the clean state is 10 kPa.

Mixers. Static gas mixers are used for two purposes in nitric acid plants:

1) To mix ammonia and air in a ratio of ca. 1 : 10 for catalytic oxidation (see Section 1.3.1)
2) To mix ammonia and tail gas in a ratio of 1 : 100 for the catalytic reduction of NO_x as part of tail-gas treatment (see Section 1.4.2.3).

From an economic standpoint, homogeneous mixing of the gas streams upstream of the reactor is important in both cases. Local ammonia excesses in the burner are a risk for plant safety (explosion limit) and may also cause overheating of the catalyst gauze. Poor mixing lowers the conversion of ammonia to nitrogen monoxide and increases platinum loss from the catalyst.

The efficiency of a given mixing operation is described by the standard deviation σ of measured samples from the theoretical value $\bar{x}$ a certain distance downstream from mixing. In nitric acid plants, the deviation should be < 1 %. If the theoretical ammonia concentration in a mixture is 10 %, for example, measured values fluctuate between 9.9 and 10.1 % [54]. Only static mixers are used (e.g., the Uhde tubular mixer or the Sulzer three-element mixer).

1.3.3.2. Burners and Waste-Heat Boilers

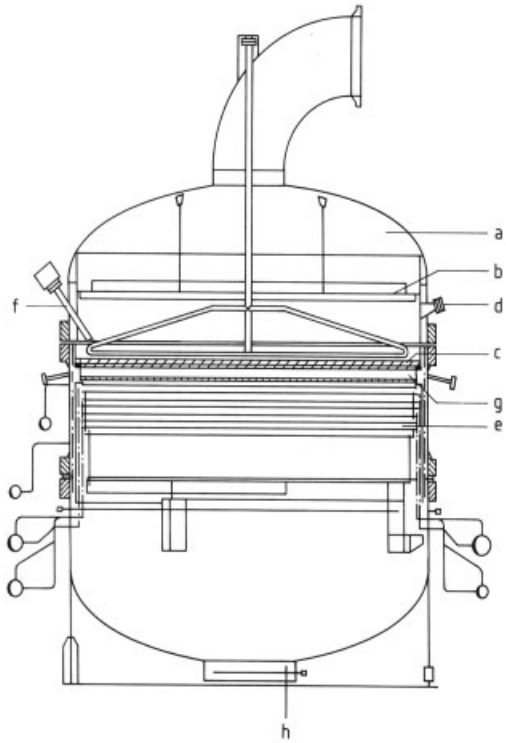

Figure 15. Reactor for catalytic ammonia oxidation with waste-heat recovery system (Lentjes)
a) Burner head; b) Perforated plate; c) Platinum gauzes and platinum recovery gauzes; d) Inspection glass; e) Superheater and evaporator tubes; f) Hydrogen ignition; g) Refractory packing; h) Nitrous gas outlet

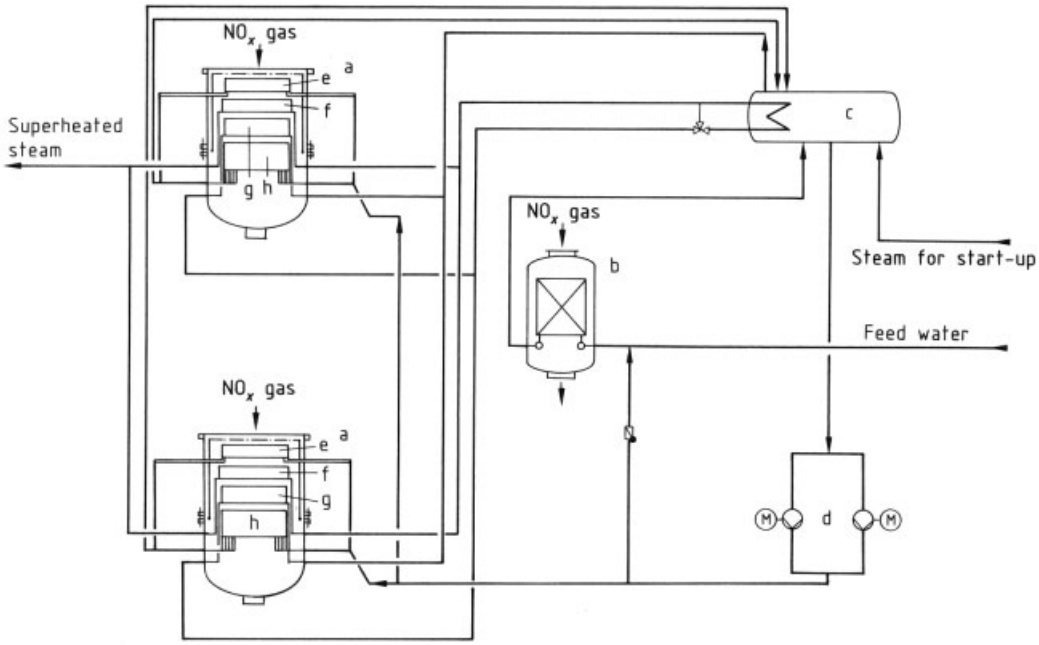

Figure 16. Water and steam connections of heat exchangers (Lentjes) downstream of catalysis
a) Reactor; b) Economizer; c) Steam drum; d) Boiler feedwater pumps; e) Preevaporator; f) Upper superheater; g) Lower superheater; h) Main evaporator

The heat of reaction liberated during ammonia combustion is utilized to produce steam and preheat the tail gas. Steam is produced in a waste-heat boiler located immediately below the burner (Fig. 15). The burner consists of a burner basket packed with filling material to ensure uniform distrubution of the downward flowing reaction gas. Platinum recovery gauzes (page 7) are located above the filling material. The catalyst gauzes (page 6) are located above the recovery gauzes. The burner basket is clamped between the flanges of the burner head and the waste-heat boiler. To ensure better distribution of the reaction mixture, the burner head also contains a perforated plate or honeycomb grid. Hydrogen burners rotating above the surface are often used for ignition. The gauze temperature is measured, and the space above the gauze can be observed through inspection glasses. The gauzes glow bright red during catalysis. The throughput per element can produce up to 1200 t of 100 % nitric acid per day.

Waste-heat boilers up to 6 m in diameter are used. Figure 15 shows a waste-heat boiler for medium-pressure burning. The preevaporator is located directly below the burner basket and protects the downstream superheater against excessive temperature. The superheater is followed by the main evaporator. Evaporator tubes are also installed at the wall to cool it and protect it from excessive temperature.

Figure 16 shows how heat from the waste-heat boiler is used to raise steam. Boiler feedwater is led through the economizer (b) into a steam drum (c). Water at thermal equilibrium with the steam is pumped from the drum through the evaporators. Steam from the drum goes first to the lower superheater (g) and is then cooled in the drum so that the desired temperature can be attained in the upper superheater (f).

Systems with an economizer integrated in the waste-heat boiler have also proved serviceable (Fig. 17, see next page). Here, as in the waste-heat boiler of Figure 15, the heat-exchanger tubes can be arranged spirally in disk form or in a square pack.

A shell-and-tube evaporator with natural circulation can also be used (Fig. 18, see next page). The superheater tubes (e) are again arranged as a flat spiral. These waste-heat boilers are particularly suitable for smaller plants or those using low steam pressure.

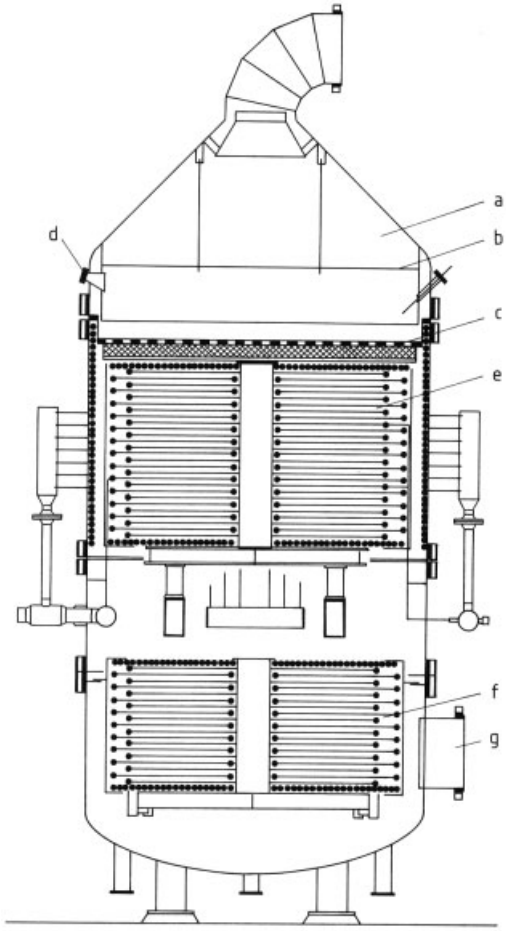

Figure 17. Reactor for catalytic ammonia oxidation with integrated waste-heat recovery system (Steinmüller)
a) Burner head; b) Perforated plate; c) Platinum gauzes and platinum recovery gauzes; d) Inspection glass; e) Superheater and evaporator tubes; f) Feedwater preheater; g) Nitrous gas outlet

Stress cracking corrosion is a special threat to economizer tubes. Start-up always leads to the formation of ammonium salts; nitrous and nitric acids may also condense. To prevent very rapid corrosive attack, the equipment is therefore heated before start-up and the feedwater is preheated to a high temperature.

Waste-heat boilers can be designed for pressures up to 10 MPa and temperatures up to 550 °C in the superheater. The steam is used to drive a turbine or is exported.

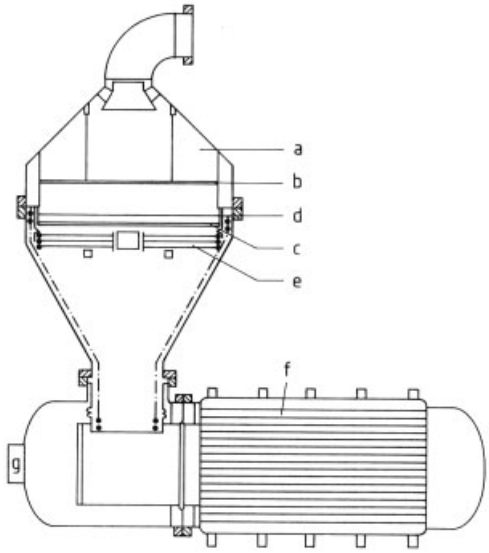

Figure 18. Reactor for catalytic ammonia oxidation with integrated waste-heat recovery system (Steinmüller)
a) Burner head; b) Perforated plate; c) Platinum gauzes; d) Packing; e) Superheater tubes; f) Evaporator; g) Nitrous gas outlet

1.3.3.3. Compressors and Turbines

Machines on a common shaft are used to deliver and compress the gases and supply the drive power needed for this purpose. Power sources are the tail-gas turbine and a steam turbine or electric motor. The tail-gas turbine can supply 35 – 100 % of the compression energy required for the process, depending on the degree of preheating; the remainder comes from the steam turbine. The machinery used depends on the nitric acid process and may consist of an air compressor, a nitrous gas compressor, a tail-gas turbine, and a steam turbine or an electric motor.

Figure 19 shows the nitric acid process with atmospheric combustion and medium-pressure (400 – 600 kPa) absorption. The nitrous gas compressor (b) sucks the reaction mixture through the burner and simultaneously provides the pressure needed for absorption. This compressor does not have interstage cooling and can be of radial or axial design. The tail gas from the absorption column is expanded in a tail-gas turbine (d), usually a single-stage device. The rest of the power required to drive the nitrous gas compressor is supplied by the steam turbine or an electric motor.

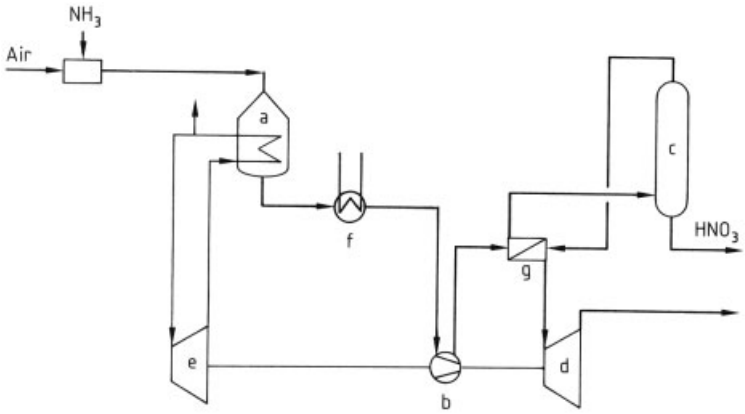

Figure 19. Flow sheet of a nitric acid process with atmospheric combustion and medium-pressure absorption
a) Reactor with waste-heat recovery system; b) Nitrous gas compressor; c) Absorption tower; d) Tail-gas turbine; e) Steam turbine; f) Cooling system; g) Heat exchanger network

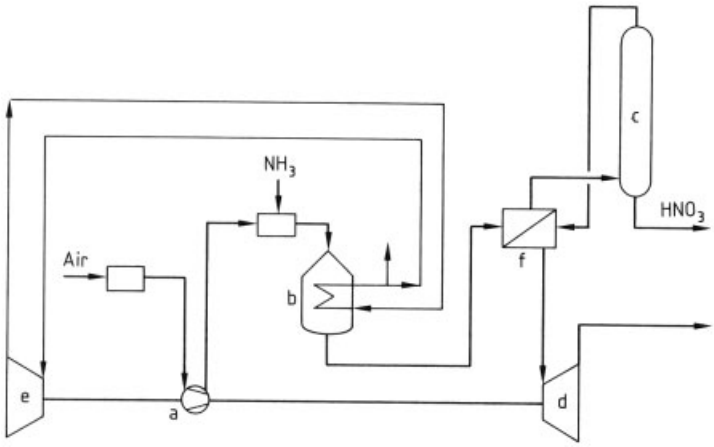

Figure 20. Flow sheet of a nitric acid process with medium-pressure combustion and medium-pressure absorption
a) Air compressor; b) Reactor with waste-heat recovery system; c) Absorption tower; d) Tail-gas turbine; e) Steam turbine; f) Heat exchanger network

Figure 20 shows a medium-pressure process (400 – 600 kPa), as preferred for small- or medium- capacity plants ($\leq$ 600 t/d 100 % HNO_3). The machinery for this process includes a radial or axial air compressor (a) without interstage cooling, a single-stage tail-gas turbine (d), and a steam turbine (e) or electric motor.

A dual-pressure process with medium-pressure combustion and high-pressure absorption calls for both an air compressor and nitrous gas compressor (Fig. 21). The air for combustion is delivered at 400 – 600 kPa by an uncooled radial or axial air compressor (a). The radial nitrous gas compressor (c) produces the 1.0 – 1.2 MPa pressure needed for absorption. The drive power for these two devices comes from a multistage tail-gas turbine (e) and a steam turbine (f) or electric motor. These systems are preferred for large- capacity plants ($\geq$ 400 t/d 100 % HNO_3), especially in Europe.

High-pressure plants (Fig. 22) are preferred in the United States because of lower investment costs. The machinery includes an air compressor with interstage cooling. Radial devices are generally used, but the low-pressure section in larger machines may be axial. The tail-gas turbine is usually a multistage reaction machine.

The *nitrous gas compressors* are virtually always of radial design, but axial machines are available. The design and operation of the com-

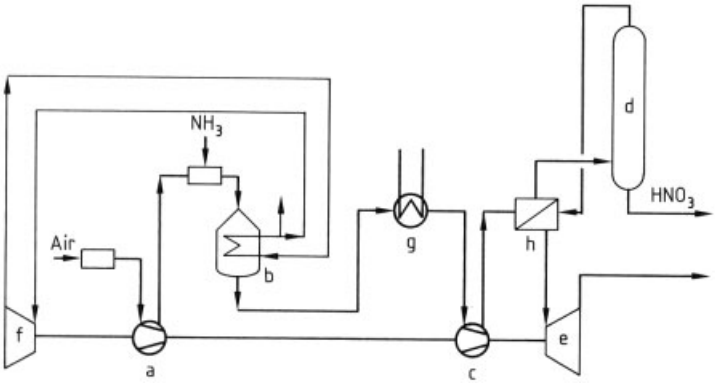

Figure 21. Flow sheet of a nitric acid process with medium-pressure combustion and high-pressure absorption
a) Air compressor; b) Reactor with waste-heat recovery system; c) Nitrous gas compressor; d) Absorption tower; e) Tail-gas turbine; f) Steam turbine; g) Cooling system; h) Heat exchanger network

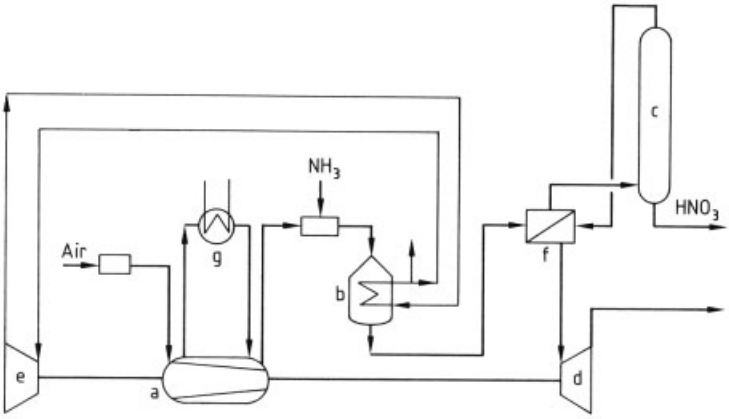

Figure 22. Flow sheet of a nitric acid process with high-pressure combustion and high-pressure absorption
a) Air compressor; b) Reactor with waste-heat recovery system; c) Absorption tower; d) Tail-gas turbine; e) Steam turbine; f) Heat exchanger network; g) Interstage cooler

pressors must take particular account of the following properties of the gas stream:

1) Nitrites and nitrates may be formed
2) The nitrous gas is toxic and highly corrosive
3) Chemical reactions take place when nitrous gas is compressed

When the nitric acid plant is started up or in continuous operation, unreacted ammonia is present in the burner exit gas. The level in continuous operation is ca. 30 ppm of ammonia. As a result, nitrates and nitrites would accumulate in the nitrous gas compressor if they were not removed regularly. If the temperature exceeds 240 °C, combustion or explosion is possible. To avoid an explosion, water or steam is briefly sprayed into the suction of the nitrous gas compressor every 8 h [55]. Because nitrous gas is highly toxic, it must not be allowed to escape from the compressor. The machine is therefore equipped with air-purged labyrinth seals. Chemical reactions must be considered in the design of nitrous gas compressors. The endothermic dissociation of dinitrogen tetroxide to nitrogen dioxide removes heat from the stream [56] and can cool the gas by 20 °C or more.

Because of their large capacities, modern nitric acid plants usually employ axial *air compressors*. The possibility of surging in the compressor, especially during plant start-up, requires the installation of appropriate surge protection. Adjustable inlet guide vanes allow operation at various plant capacities.

Tail-gas turbines can be single-stage or multistage devices of axial or radial design. For part-

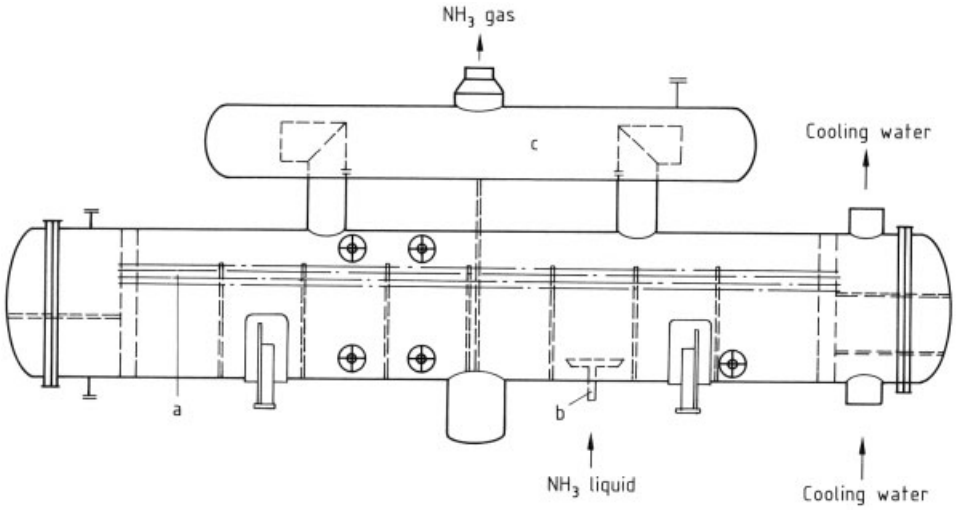

Figure 23. Ammonia evaporator
a) Evaporator tubes; b) Sparger; c) Drop separator

load operation, these machines are equipped with a group of valve-controlled nozzles or adjustable guide vanes in the inlet.

Steam turbines may be condensing or back-pressure machines with or without an extraction/side stream.

1.3.3.4. Heat Exchangers and Columns

Heat Exchangers. The shell-and-tube design predominates for tail-gas heat exchangers, with hot nitrous gas on the tube side and cool tail gas on the shell side. The pressure drop should be kept as low as possible to secure a favorable energy balance for the plant as a whole.

Ammonia Evaporator. Various types of equipment can be used for ammonia evaporation. Figure 23 shows a shell-and-tube heat exchanger with mist collector. Cooling water on the tube side is cooled on evaporation. In medium-pressure plants, the heat of vaporization is also taken from chilled water used as coolant in the absorption step. The shell-side heat-transfer coefficients for ammonia can be calculated according to [57].

Gas Cooler – Condenser. In the cooler – condenser, the temperature is lowered below the dew point of the inlet nitrous gas. The nitric acid condensate has a concentration of, for example, ca. 40 wt % at medium pressure. Suitable materials of construction are discussed in Section 1.3.3.5. The thermal design of a gas cooler – condenser is difficult because the cooling step involves chemical reactions in both the gas and the condensed liquid phases. In the apparatus shown in Figure 24, nitrous gas passes into the shell side through two inlets; each of the partial streams reverses direction four times. Heat is absorbed by cooling water.

Absorption Tower. Absorption tower design is governed by the calculations for the oxidation process (i.e., tower height, see also Section 1.3.2) and thermal design. Because oxidation proceeds more slowly as the NO_x content in the tower decreases, the spacing between plates increases. Figure 25 shows an absorption tower for a plant with a daily production capacity of 1830 t of 100 % nitric acid. Acid formation takes place chiefly in the bottom third of the tower, whereas NO_x is reduced in the upper two-thirds. As a result, most of the heat must be withdrawn in the lower third; this is done with cooling coils on the plates. The coils on the upper plates primarily cool the nitrogen monoxide gas against chilled water, because oxidation proceeds faster at lower temperature. Most modern absorption towers have sieve plates. Compartments built into the bottom of the tower separate acid of medium strength from weak acid in case of plant

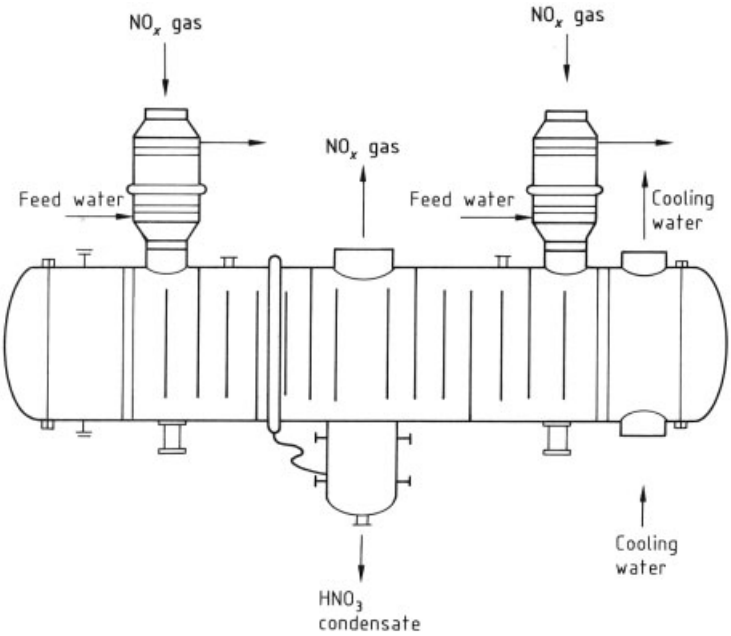

Figure 24. Cooler – condenser with feedwater preheaters attached to nitrous gas inlets

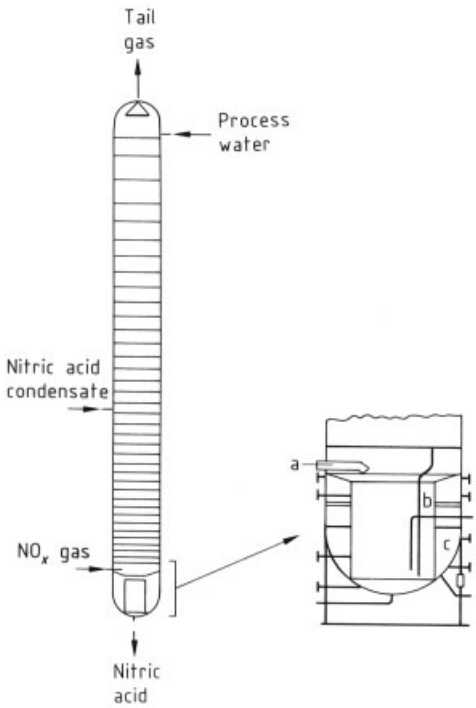

Figure 25. Absorption tower
a) Nitrous gas inlet; b) Inner compartment; c) Outer compartment

shutdown. The collected acids are pumped back to the tower at restart so that a steady state is reached more quickly. Before the tail gas leaves the tower, entrained acid droplets are collected in the demister.

Bleacher. The NO_x gases contained in the acid are stripped out with secondary air in the bleacher. A distribution tray delivers the acid evenly onto the Raschig ring packing. The packing rings are made of stainless steel, material no. 1.4301 (AISI 304) or 1.4306 (AISI 304L).

1.3.3.5. Construction Materials

Because of the special behavior of nitric acid toward metals (see Section 1.2), materials of construction for nitric acid plants must be selected very carefully. When industrial production began, acid-resistant masonry linings were used for plant components that came in contact with the product. By the 1920s, materials technology had advanced to the point that high chromium contents could be incorporated into alloy steels. Today, the principal materials of construction in nitric acid plants ($< 70\,\%$ HNO_3) are austenitic chromium – nickel steels containing 18 % chromium and 10 % nickel. The corrosion resistance

Table 4. Chemical analyses (wt %) of standard and special stainless steels [59]

Type of steel		C max.	Si max.	P max.	S max.	Cr	Ni	Other
AISI	304L	0.035	0.75	0.040	0.030	18 – 20	8 – 13	
	321	0.08	0.75	0.040	0.030	17 – 20	9 – 13	Ti
	347	0.08	0.75	0.040	0.030	17 – 20	9 – 13	Nb
ISO	1.4306	0.03	1.0	0.045	0.030	17 – 20	10 – 12.5	
	1.4541	0.10	1.0	0.045	0.030	17 – 19	9 – 11.5	Ti
	1.4550	0.10	1.0	0.045	0.030	17 – 20	9 – 11.5	Nb
Sandvik	3R12 ∗	0.030	0.60	0.030	0.030	18.5	10.5	
	2R12 ∗	0.020	0.10	0.015	0.010	18.5	11	
	2RE10 ∗∗	0.020	0.30	0.020	0.015	24.5	20.5	

∗ ISO material no. 1.4306. ∗∗ISO material no. 1.4335.

of these steels decreases with increasing temperature. Normal austenitic steels are generally not stable above 70 %. The chromium content determines the corrosion resistance; its anticorrosive action is influenced by the carbon content of steel. The chromium in the alloy forms a carbide (Cr_3C_2) at the grain boundaries; only small amounts of free chromium remain at high carbon contents, and the steel is therefore less corrosion resistant. If excess chromium remains after the amount needed for carbide formation is consumed, only this excess is responsible for the anticorrosion properties.

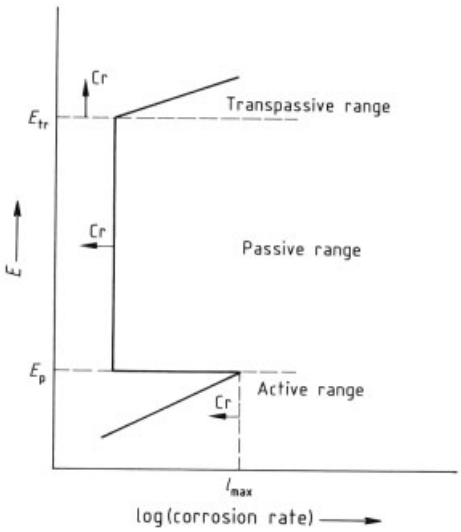

Figure 26. Anodic polarization curve for stainless steel Cr denotes the effect of chromium on the anodic polarization rate

The effect of chromium on the corrosion rate is shown in the anodic polarization curve of Figure 26. When chromium is present in the alloy, the passive region E_p (see Section 1.2) is reached sooner. Furthermore, chromium lowers the corrosion rate in the passive state and raises the transpassive potential E_{tr}. The last point is important because the corrosion potential of stainless steel in strong (highly concentrated) nitric acid often reaches the limit of the transpassive potential. In the transpassive range the passive layer starts to dissolve. A steel with too little chromium or with elements that lower the transpassive potential tends to corrode at a high rate [58].

Steels of a given specification behave differently under nitric acid attack because of permitted batch-to-batch variations in composition. The only way to be sure that a selected steel is suitable for an apparatus in contact with nitric acid is to perform tests. The best known method is the Huey test, in which a clean, polished piece of steel is immersed in boiling 65 % nitric acid at ambient pressure [59]. The specimen is exposed to these conditions for five successive 48-h periods, the acid being renewed at the beginning of each test period. The Huey test gives a good indication of how individual alloy constituents affect intergranular corrosion.

To ensure good weldability, the carbon content is limited to 0.03 %; alternatively, titanium or niobium is added as a stabilizer.

The stainless steels generally used in nitric acid plants are those classified unter material no. 1.4306 (AISI 304L); in the Federal Republic of Germany, however, the corresponding stabilized qualities, material no. 1.4541 (AISI 321) and 1.4550 (AISI 347), are also employed (Table 4). These steels are usually specified for towers and heat-exchanger shells, and also for many heat-exchanger tubes. The 1.4335 steel (AISI 310L) has a very low corrosion rate and is used

particularly where corrosive attack is severe, for example, if the temperature is below the dew point of nitrous gas, even at higher temperatures [59]. These conditions normally occur at the inlet to the tail-gas preheater, in the dew point region of the cooler – condenser, and at the outlet of the feedwater preheater. Corrosion at these locations is caused mainly by reevaporation of nitric acid condensate, which brings the acid concentration up to the azeotropic level (69 % HNO_3). The acid is then very aggressive, even in the vapor phase. Corrosion can take place in the boiler feedwater preheater if condensate forms at the outlet. Reheating of the condensate by the hot process gas can lead to critical conditions. Process design parameters should normally be selected so as to avoid condensation; material no. 1.4306 is then adequate. If, however, condensation or reboiling occurs and this steel cannot be used, 1.4335 should be employed [60].

In plants producing concentrated nitric acid, aluminum (99.8 %), ferrosilicon, tantalum, and special austenitic steels are used. Because tantalum is very expensive, it is employed only with boiling concentrated nitric acid. Ferrosilicon can be used only in castings because of its brittleness.

1.3.4. Processes

Figure 14 illustrates the steps necessary for implementation of the Ostwald process. Industrial processes differ in the sequence and design of these steps. This section describes processes for the production of weak acid [61, pp. 61 – 98] and strong acid [61, pp. 99 – 130].

1.3.4.1. Weak Acid Processes

The first industrial plants for the production of weak acid employed atmospheric combustion and low-pressure absorption [3]. This type of plant is no longer built. It was followed by a process employing atmospheric combustion and absorption at medium pressure. This process has the advantage of lower absorption costs because the marked improvement in absorption obtained with the pressure generated by a nitrous gas compressor allows the apparatus to be downsized. The energy of compression can be partially recovered with a tail-gas turbine. To circumvent the difficulty of operating a compressor for nitrous gas and obtain a further decrease in plant volume, to ammonia was also oxidized at medium pressure. This type of process is particularly economical for smaller capacities and is thus described first.

As plant capacities continued to grow and lower tail-gas levels of NO_x were necessary the absorption pressure was increased still further. Lower energy costs and shorter depreciation periods in the United States have led to a preference for the high-pressure process, whereas more favorable consumption and production figures in Europe have favored the dual-pressure process with medium-pressure combustion and high-pressure absorption.

The principle of medium- and high-pressure (monopressure) plants is that air for ammonia oxidation and NO_x absorption is compressed to the desired pressure. The "primary" air for oxidation is then mixed with ammonia and forced through the burner. The "secondary" air for stripping dissolved NO_x out of the raw acid and oxidizing the intermediate nitrogen monoxide is supplied upstream of the absorption tower. The tail gas is heated and then expanded in a turbine (on a common shaft with the compressor) to produce mechanical energy. The remaining drive power is supplied by a steam turbine running on process steam.

Medium-Pressure Process. Figure 27 is a flow sheet of a medium-pressure process (ca. 550 kPa). Liquid ammonia is evaporated at ca. 700 kPa in the ammonia evaporator (a). The cold generated can be used in the absorption step. Water is removed from liquid ammonia in an ammonia stripper (b). The ammonia – water blowdown mixture is evaporated batchwise with steam. Residual aqueous ammonia can be used in fertilizer production. The ammonia vapor is heated with steam to ca. 90 °C in the ammonia preheater (c) and then filtered (d). If necessary, a small stream of ammonia is diverted to tail-gas treatment.

All the air needed for the process is sucked in through the air filter (f) by the air compressor (g). The compressed air is divided into two substreams in a tail-gas preheater (h). The secondary air goes to the bleacher (x) for stripping raw acid.

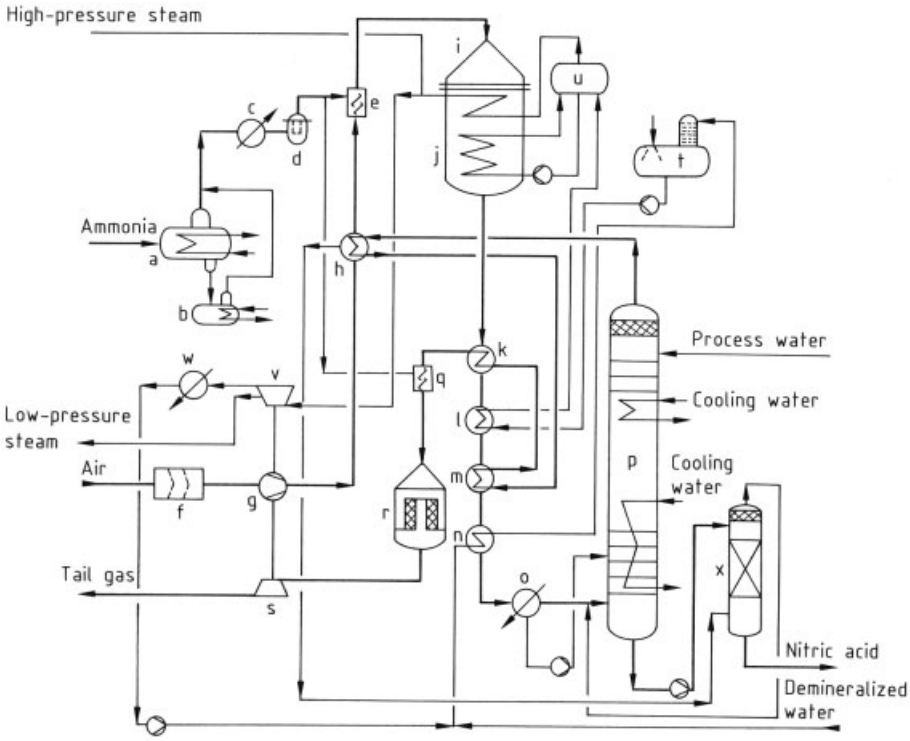

Figure 27. Simplified flow sheet of a medium-pressure process
a) Ammonia evaporator; b) Ammonia stripper; c) Ammonia preheater; d) Ammonia gas filter; e) Ammonia – air mixer; f) Air filter; g) Air compressor; h) Tail-gas preheater III; i) Reactor; j) Waste-heat boiler; k) Tail-gas preheater I; l) Economizer; m) Tail-gas preheater II; n) Feedwater preheater; o) Cooler – condenser; p) Absorption tower; q) Ammonia – tail-gas mixer; r) BASF catalytic tail-gas reactor; s) Tail-gas expansion turbine; t) Feedwater tank with deaerator; u) Steam drum; v) Steam turbine; w) Steam turbine condenser; x) Bleacher

The secondary airstream exiting the bleacher is laden with NO_x and is added to the nitrous gas before it enters the absorption tower (p). The greater part of the process air, the primary air, is mixed with the superheated ammonia in the ammonia – air mixer (e); it then flows through the mixed-gas filter (not illustrated) and, now cleaned, enters the reactor (i) where ammonia reacts with oxygen in the primary air over the platinum – rhodium catalyst, yielding nitrogen monoxide and water. The reaction is exothermic and proceeds at ca. 890 °C. The nitrous gas stream is cooled in the waste-heat boiler (j), raising steam. Nitrogen monoxide is oxidized to nitrogen dioxide in the downstream piping and equipment, heating the nitrous gas stream. Further heat recovery takes place in the tail-gas preheater (k) and the economizer (l). After exiting the economizer, the nitrous gas stream is cooled in a second tail-gas preheater (m). Condensate can form in the next unit downstream, the feedwater preheater (n), which is located directly on the inlet of the cooler – condenser (o). Additional water of combustion condenses out in the cooler – condenser before the nitrous gas stream is mixed with the NO_x-laden secondary air and admitted to the absorption tower (p). The acid condensate from the feedwater preheater and the cooler – condenser flows directly into the bottom of the gas cooler – condenser. This acid condensate is then delivered by an acid condensate pump to a tray with the appropriate concentration in the absorption tower.

In the lower part of the tower, the nitrous gas is oxidized further, thus reaching the necessary degree of oxidation before entering the absorption section. The tower contains sieve trays through which the gas flows countercurrent to the acid. The product acid concentration is attained on the lowermost absorption tray of the tower. The acid is then pumped to the bleacher (x), where dissolved NO_x is stripped, before the

acid is stored in tanks. Chlorides in the absorption tower feed accumulate on trays where the acid concentration is ca. 21 wt %. Because of the danger of corrosion, these trays must be emptied into the product acid tank from time to time. Cooling water flowing through coils on the sieve trays removes heat generated by the oxidation of nitrogen monoxide to nitrogen dioxide and its further conversion to nitric acid. Part of the heat absorbed by the cooling water is utilized to evaporate ammonia. A demister in the head of the tower collects liquid droplets entrained in the tail gas; the condensate runs back to the last tray of the absorption tower.

The tail gas absorbs heat from the secondary air and the nitrous gas in three tail-gas preheaters. In the BASF catalytic tail-gas treatment (r; see also Section 1.4.2.3), ca. 60 vol % of the NO_x in the tail gas is selectively reacted with ammonia. Hot tail gas containing < 200 ppm NO_x goes to the tail-gas expansion turbine (s), in which mechanical energy is produced to drive the air compressor. Finally, the tail gas is discharged to the atmosphere via the stack.

The tail-gas expansion turbine supplies part of the power needed to drive the air compressor. The rest is generated by a condensing steam turbine supplied with product steam. The turbine may be bled. The steam turbine condensate goes to the boiler feedwater preheater (n) where it absorbs heat from the NO_x gas and then flows through the deaerator into the feedwater tank (t). Nondeaerated feedwater from the battery limitalso flows via a preheater into the feedwater tank. The boiler feedwater pump raises the pressure of the deaerated feedwater to the requisite boiler pressure. Product steam is raised with heat from ammonia combustion in the waste-heat boiler (j).

Most of the steam generated in the waste-heat boiler drives the condensing steam turbine. The remainder can be utilized in the ammonia preheater and stripper and for deaerating the feedwater; any excess is exported as product steam.

High-Pressure Process. Figure 28 shows a high-pressure process running at ca. 1 MPa. Liquid ammonia is fed into the evaporator (a), where it is evaporated at ca. 1.15 MPa against warm water. The evaporation temperature rises slightly above 35 °C as the water content builds up in the evaporator. The ammonia – water blowdown mixture is evaporated with low-pressure steam in the ammonia stripper (b). The residual ammonia concentration is ca. 2.5 %.

Ammonia exiting the evaporator system is heated to ca. 130 °C in the ammonia preheater (c), and contaminants are removed in the ammonia gas filter (d).

Air is sucked in through the air filter (f) and compressed to ca. 1 MPa in an air compressor (g) with interstage cooling (h). The primary air (ca. 80 % of the total) is heated to ca. 180 °C against nitrous gas in the air heater (m) and then fed to the ammonia – air mixer (e). The mixed gas stream has a temperature of ca. 175 °C and contains 10.7 vol % ammonia. It passes through a filter (not illustrated) into the reactor (i) where the ammonia and atmospheric oxygen react over the platinum – rhodium catalyst at ca. 900 °C to form mainly nitrogen monoxide and water.

The reaction gases next pass through the waste-heat boiler (j), where they are cooled to ca. 400 °C. The heat is used to raise high-pressure steam and superheat it to 500 °C in the preevaporator, superheater, and reevaporator of the boiler.

The boiler system is fed with condensate from the steam turbine condenser (w) plus demineralized water from the battery limit. Both streams are led to a tank and from there go through the feedwater preheater (n) to the deaerator mounted on the feedwater tank (t) and operating at a slight gauge pressure. The tank serves as suction vessel for the feedwater pumps, which deliver water through the economizer into the boiler drum (u).

The water – steam mixture in the evaporator is circulated by the boiler circulation pump. Saturated steam and water are separated in the boiler drum (u).

After leaving the waste-heat boiler (j), the nitrous gas is cooled to ca. 260 °C in a tail-gas preheater (k). After passing through a run of piping for oxidation, nitrous gas gives up more useful heat in the economizer (l), and its temperature falls to ca. 210 °C. After another oxidation run, the stream is led into the air heater (m) and cooled to ca. 180 °C. The heat recovered is used to preheat the primary air.

Before the nitrous gas is led into the gas cooler – condenser (o), it passes through feedwater preheaters and through the warm-water heater (n) both mounted on the two inlets of the cooler – condenser. The nitrous gas is cooled to

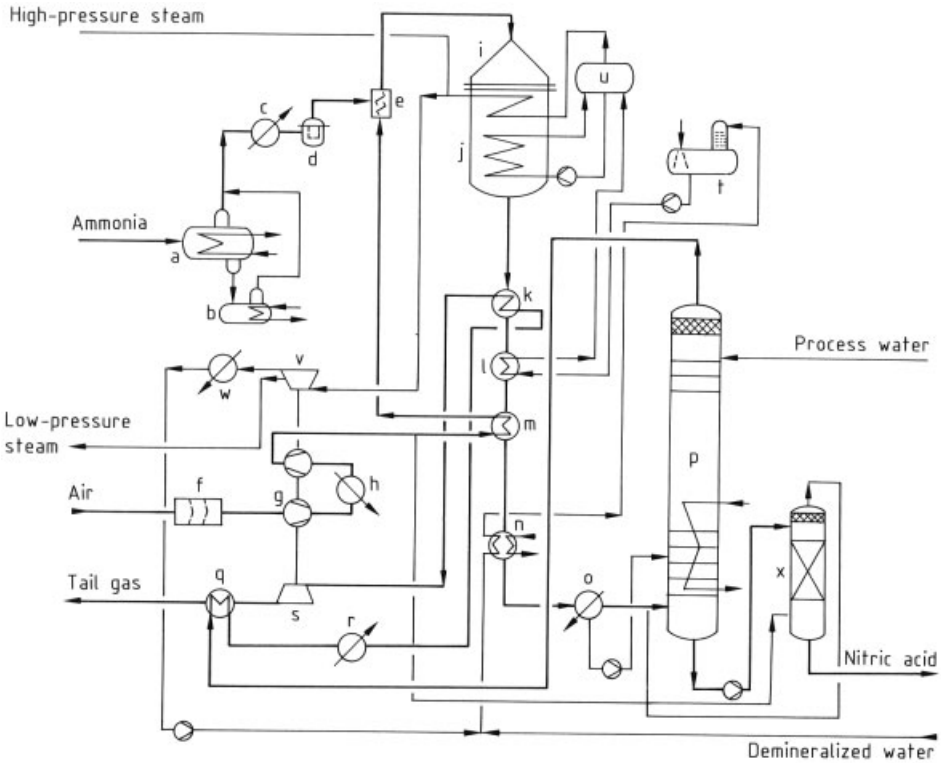

Figure 28. Simplified flow sheet of a high-pressure process
a) Ammonia evaporator; b) Ammonia stripper; c) Ammonia preheater; d) Ammonia gas filter; e) Ammonia – air mixer; f) Air filter; g) Air compressor; h) Interstage cooler; i) Reactor; j) Waste-heat boiler; k) Tail-gas preheater; l) Economizer; m) Air preheater; n) Feedwater and warm-water preheaters; o) Cooler – condenser; p) Absorption tower; q) Tail-gas preheater; r) Tail-gas preheater; s) Tail-gas expansion turbine; t) Feedwater tank with deaerator; u) Steam drum; v) Steam turbine; w) Steam turbine condenser; x) Bleacher

ca. 115 °C and partly condensed before entering the cooler – condenser.

In the water-cooled cooler – condenser (o), the gas is cooled to < 50 °C so that water formed during ammonia oxidation condenses to produce ca. 45 % acid. This product is delivered to the appropriate tray of the absorption tower (p) by the acid condensate pump. The NO_x-laden secondary air recycled from the bleacher (x) is mixed with the nitrous gas stream before it is fed into the tower (p).

In the sieve-tray absorption tower (p) the nitrous gas flows countercurrently to the process water fed at the column head, and nitric acid is formed. Heat from the absorption tower is transferred to the cooling water. Raw acid from the absorption tower is treated with secondary air in the bleacher (x), then delivered to battery limit with the aid of the system pressure.

Tail gas exits the absorption tower at ca. 20 – 30 °C. It is heated to ca. 80 °C in a tail-gas preheater (q) against the tail-gas stream from the expansion turbine. The tail gas is further heated to 140 °C against low-pressure steam (r) and then to ca. 375 °C in a tail-gas preheater (k). It is subsequently expanded in the tail-gas turbine (s); this step produces ca. 70 % of the energy required to drive the air compressor.

Tail gas exits the turbine at ca. 135 °C and is led to a tail-gas preheater (q), where it is cooled to ca. 90 °C before being discharged through the stack.

The rest of the power needed to drive the air compressor can be supplied by a pass-out condensing steam turbine (v). The superheated steam is delivered to the steam turbine or exported. Part of the steam supplied to the turbine is withdrawn as low-pressure steam before admission to the condensing section. This low-pres-

sure steam is used in the plant or exported. The rest of the steam supplied to the turbine passes through the condensing section and is condensed in the steam turbine condenser (w).

Dual-Pressure Process. The dual-pressure process combines the favorable economics of medium-pressure combustion with the efficiency of high-pressure absorption. Figure 29 is a flow sheet of such a plant with ammonia combustion at 0.5 MPa and absorption at 1.1 MPa.

Liquid ammonia is fed to the ammonia evaporator (a), which is under a pressure of ca. 0.65 MPa. The entire stream then goes to ammonia evaporator I, where ca. 80 % of the ammonia is evaporated against cold water at a constant temperature of ca. 12 °C. Residual liquid ammonia is fed to ammonia evaporator II (not illustrated) and evaporated at varying temperatures against cooling water. The evaporation temperature increases from 12 °C to ca. 20 °C as water content builds up in the evaporator. Ammonia evaporator II is designed so that the entire ammonia stream can be evaporated in it, if necessary, at a maximum temperature of 20 °C. If the water content or the evaporation temperature in ammonia evaporator II is too high, the ammonia – water mixture is blown down and fed to the ammonia stripper (b). Ammonia is stripped to a residual content of ca. 2.0 % as the temperature is raised to 150 °C with low-pressure steam. The residue, greatly depleted in ammonia, is discharged.

Gaseous ammonia from the evaporator system is passed through the ammonia gas filter (c) to remove contaminants, then heated to ca. 150 °C in the ammonia preheater (d).

Air is sucked through the filter (f) and compressed to 0.5 MPa in the compressor (g). About 86 % of the air is fed as primary air to the ammonia – air mixer (e). The mixed gas stream has a temperature of ca. 220 °C and contains ca.

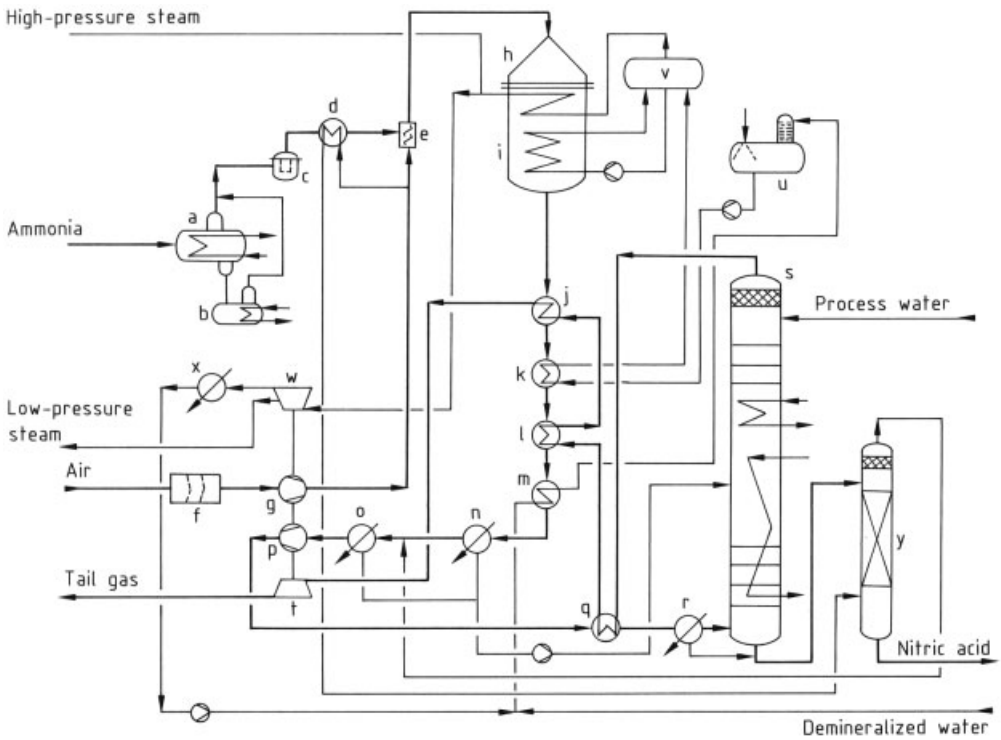

Figure 29. Simplified flow sheet of a dual-pressure process
a) Ammonia evaporator; b) Ammonia stripper; c) Ammonia gas filter; d) Ammonia preheater; e) Ammonia – air mixer; f) Air filter; g) Air compressor; h) Reactor; i) Waste-heat boiler; j) Tail-gas preheater III; k) Economizer; l) Tail-gas preheater II; m) Feedwater preheater; n) Cooler – condenser I; o) Cooler – condenser II; p) Nitrous gas compressor; q) Tail-gas preheater I; r) Cooler – condenser III; s) Absorption tower; t) Tail-gas expansion turbine; u) Feedwater tank with deaerator; v) Steam drum; w) Steam turbine; x) Steam turbine condenser; y) Bleacher

10 vol % ammonia. It is transported through the gas-mixture filter (not illustrated) to the reactor (h). Here, ammonia reacts on the platinum–rhodium catalyst with atmospheric oxygen at 890 °C to produce a 96.5 % yield of nitrogen monoxide and water.

The reaction gases next pass through the waste-heat boiler (i) and are cooled to 355 °C. The heat is used to raise high-pressure steam in the preevaporator, superheater, and reevaporator of the boiler.

The boiler system is fed with condensate from the steam turbine condenser (x) plus demineralized water from the battery limit. Both streams are led to a tank and then pass through the feedwater preheater (m) to the deaerator mounted on the feedwater tank (u). The tank serves as a suction vessel for the feedwater pumps, which deliver the water through the economizer (k) into the steam drum (v).

The water–steam mixture in the evaporator is circulated by the boiler circulation pump. Saturated steam and water are separated in the steam drum (v).

After exiting the waste-heat boiler and passing through a run of piping for oxidation, the nitrous gas is cooled from 390 to 280 °C in tail-gas preheater III (j). After further oxidation, it enters the economizer (k), where it is cooled from 310 to 200 °C. The nitrous gas gives up further useful heat to the tail gas in tail-gas preheater II (l) and its temperature falls to ca. 170 °C.

The gas passes through the feedwater preheater (m), mounted on the cooler–condenser inlet and thereby cools to ca. 100 °C. In the water- cooled cooler–condenser I (n), the gas is cooled to < 50 °C so that the water formed during ammonia oxidation condenses. A ca. 43 % acid results and is pumped to the appropriate tray of the absorption tower (s). The secondary air laden with NO_x and recycled from the bleacher (y) is now mixed with the nitrous gas stream. The mixed stream is then led into cooler–condenser II (o), where it is cooled to ca. 45 °C against chilled water. More water condenses and a ca. 52 % acid is formed, which is forwarded for absorption along with acid formed in cooler–condenser I. The gas is passed through a mist collector (not illustrated) upstream of the nitrous gas compressor (p).

The nitrous gas is compressed to 1.1 MPa in the compressor (p) and has an outlet temperature of ca. 130 °C. It passes to tail-gas preheater I (q), where it is cooled to ca. 75 °C. Downstream of the preheater is cooler–condenser III (r), which the nitrous gas leaves at 50 °C. The resulting acid condensate has a concentration of ca. 58 wt % and is mixed with the raw acid from the absorption column.

From cooler–condenser III the nitrous gas, now at least 90 % oxidized, enters the absorption tower (s). The gas first passes through oxidation trays that receive only a slight flow of liquid, to attain the degree of oxidation needed for equilibrium with 68 % acid. In the absorption section, gas and liquid move countercurrently. Required process water is pumped to the tower head. The 68 % raw acid is withdrawn from the first absorption tray and then treated with secondary air in the bleacher (y). It is delivered to the battery limit with the aid of the system pressure.

Approximately 80 % of the heat generated by absorption is transferred to the cooling water and the remainder to the chilled water employed in the last part of the absorption step where very little heat of reaction is liberated.

The tail gas contains $< 150-200$ ppm NO_x and exits the absorber at 20 °C. It is heated to 110, 200, and finally 350 °C in the three downstream tail-gas preheaters (q, l, and j) before being expanded in the tail-gas turbine (t). The turbine supplies ca. 67 % of the power required by the air and the nitrous gas compressors. The expanded tail gas has a temperature of ca. 100 °C and is discharged through a stack.

The remaining power required to drive the rotating machinery can be obtained from a pass-out condensing steam turbine (w). Superheated steam is fed to the turbine or exported. Part of the steam supplied to the turbine may be withdrawn as low-pressure steam before it enters the condensing section. This low-pressure steam may be used in the plant or exported. The remainder of the steam fed to the turbine passes through the condensing section and is condensed in the steam turbine condenser (x).

Comparison of Medium- and High-Pressure Processes [62]. Two trends can be seen in the worldwide development of weak acid processes. First, from the process-engineering standpoint, a progression occurs from low- through medium- to high-pressure processes [5]. Second, capacities continue to increase; single-

Table 5. Comparison of significant specific consumption figures for nitric acid plants (values are given per tonne of 100 % HNO_3, the tail gas contains < 200 ppm of NO_x)

Parameter	Monopressure processes		Dual-pressure process
	Medium pressure	High pressure	
Operating pressure, MPa	0.55	1.080	0.45/1.1
Ammonia, kg	282 ∗	283	279
Electric power, kW · h	8.5	8.0	9.0
Platinum, g	0.14	0.30	0.11
Cooling water, t ($\Delta t = 10\,^\circ$C) ∗∗	120	125	130
Process water, t	0.3	0.3	0.3
Low-pressure heating steam, t	0.1	0.1	0.1
High-pressure excess steam, t (2.5 MPa, 400 °C)	0.87	0.78	0.81

∗ Includes 1 kg NH_3 for NO_x reduction from 600 to < 200 ppm. ∗∗Includes water for steam turbine compressor.

train plants producing up to 2000 t of nitric acid per day are now being built. In Europe, the dual-pressure design is preferred for larger plants, whereas smaller ones employ the monopressure design. Where feedstock and energy prices are low, monopressure operation offers special advantages; low investment costs ensure a quick payout, particularly in North America. If, on the other hand, feedstock and energy prices are very high (as in Europe), yield and energy efficiency must be maximized, so higher investment costs are acceptable. Table 5 compares consumption figures per unit product for monopressure and dual-pressure processes.

New pollution control regulations and the energy crisis of the mid-1970s have also led to the development of new processes or the improvement of existing ones. To avoid the need for catalytic tail-gas treatment in large-tonnage plants (and thus an increase in specific ammonia consumption), new facilities employ higher absorption pressures.

1.3.4.2. Concentrated Acid Processes

Industrially produced nitric acid contains 50 – 70 wt % HNO_3. This is high enough for fertilizer production, but nitration processes in industrial organic chemistry call for concentrated acid (98 – 100 %). Distillation, the simplest way to concentrate dilute acid, fails because nitric acid and water form an azeotrope (68.4 % HNO_3 at atmospheric pressure).

Concentrated nitric acid is manufactured directly or indirectly [63]. In the *direct process*, the water generated in ammonia combustion is withdrawn by rapid cooling to give a nitrous gas mixture from which concentrated acid can be produced directly in one of two ways. First, the completely oxidized NO_x can be separated in liquid form by absorption in concentrated nitric acid and then led to a reactor where it is reacted with oxygen and water (or weak acid) under pressure to yield concentrated acid. Second, nitrous gases can be reacted with azeotropic acid to form a concentrated acid that can be converted easily to concentrated and azeotropic acid by distillation. The latter product is either completely recycled or used in the production of ordinary weak acid.

In the *indirect processes*, concentration is based on extractive distillation and rectification with sulfuric acid or magnesium nitrate.

Direct Process. The essential feature of the classical direct process (Fig. 30) is that liquid dinitrogen tetroxide is produced and reacted under pressure with pure oxygen and a certain quantity of dilute nitric acid. A detailed account of concentrated acid production via azeotropic nitric acid is given in [61, pp. 99 – 130].

Ammonia Oxidation. As in weak acid processes, feed ammonia is evaporated, superheated, and filtered (see Section 1.3.4.1). Evaporation takes place with the aid of water heated with process heat. Because traces of water and oil accumulate in the ammonia evaporator the sump must be drained into the steam-heated stripper from time to time. Air for ammonia combustion is sucked through a filter (a) and mixed with purified ammonia in the mixer (b). The ammonia content of the mixture is held in the range 11.5 – 12.3 vol %, depending on the temperature. The optimal combustion temperature is 830 – 850 °C. Ammonia oxidation takes place over conventional platinum – rhodium cat-

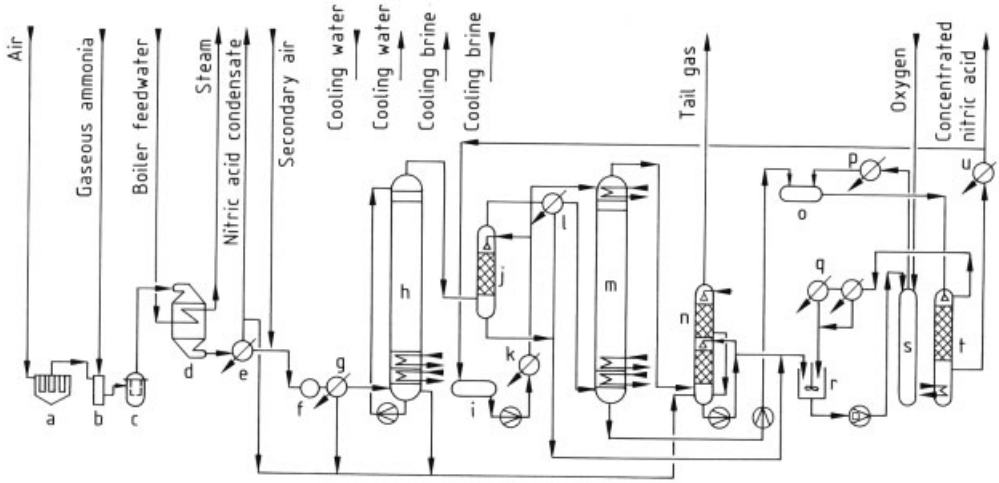

Figure 30. Simplified flow sheet of a process for direct production of strong nitric acid with oxygen
a) Air Filter; b) Ammonia – air mixer; c) Gas mixture filter; d) Reactor; e) Cooler; f) Compressor; g) Cooler; h) Oxidation tower; i) Recirculating acid tank; j) Postoxidizer; k) Cooler; l) Cooler; m) Absorption tower; n) Final absorber; o) Raw acid tank; p) Cooler; q) Precondenser; r) Stirred tank; s) Reactor; t) Bleacher; u) Cooler

alysts in the reactor (d). The heat of reaction is used to raise steam in a waste-heat boiler and an economizer. The process gas exits the economizer at ca. 170 °C and is then cooled to 25 °C in a cooler – condenser (e), giving an acid containing ca. 2 – 3 wt % nitric acid. The cooler – condenser is specially designed to condense as much water as possible without much NO_x absorption. Part of the process heat is also transferred to the warm water used for ammonia evaporation. The weak condensate generated in the cooler – condenser is used as scrub liquor for final absorption.

Oxidation of Nitrogen Monoxide. Secondary air is mixed with the cooled process gas to oxidize nitrogen monoxide to nitrogen dioxide. The amount of secondary air must be sufficient to ensure that the tail gas has a minimal oxygen content (2.5 – 3.5 vol %) before it is discharged to the atmosphere. The concentrated nitric acid venting system is connected to the secondary air intake in such a way that oxygen and nitrogen oxides liberated on venting are recycled to the process. The nitrogen monoxide gas and secondary air are compressed together to ca. 0.14 MPa (f). After further cooling (g) to remove the heat of compression, the mixture enters the oxidation tower (h), which is usually equipped with sieve trays. Tube coils on the trays remove the heat of oxidation. To maintain a liquid level on all trays, acid from the tower bottom is pumped (recirculated) to the head.

Final Oxidation of Nitrogen Monoxide. The nitrous gases leaving the oxidation tower are approximately 95 % oxidized and are led to a final oxidizer (postoxidizer) (j) where they are completely oxidized by desorption of concentrated nitric acid. The acid moves countercurrently to the nitrous gases. Water produced in this reaction dilutes the concentrated acid; acid leaving the postoxidizer contains ca. 75 % HNO_3. Flow control of the concentrated nitric acid feed into the postoxidizer is important because too high a rate lowers the output of concentrated acid, whereas too low a rate displaces final oxidation into the absorber, with the result that the concentrated nitric acid entering the absorption tower is too dilute. The product gas is saturated with nitric acid vapor. For the next step, absorption, the process gas is cooled to −10 °C against brine in a cooler (l). Virtually all the nitrogen dioxide is dimerized to dinitrogen tetroxide at this temperature.

Absorption of Dinitrogen Tetroxide. The nitrous gases are completely dimerized to dinitrogen tetroxide and, in this condition, are fed into the absorber (m). The absorber consists of four sections, each packed with ceramic Raschig rings. The dinitrogen tetroxide is absorbed in concentrated nitric acid that comes from the bleacher and is cooled to ca. −10 °C (k, u). The

acid leaving the absorption tower contains ca. 25 – 30 wt % dinitrogen tetroxide and is pumped to the raw acid tank (o).

Final Absorption. The gas exiting the absorption tower is practically free of nitrogen dioxide but saturated with nitric acid vapor. The vapor is removed by scrubbing in the final absorber (n), which consists of two sections. The scrub liquor (weak nitric acid from the cooler – condenser) is recirculated or used to make up the water balance in the stirred tank.

Dinitrogen Tetroxide Production. The laden concentrated nitric acid is sent from the raw acid tank (o) to the bleacher (t), which consists of a stripping section and a reboiler. The acid enters the head of the tower and flows downward countercurrently to the nitric acid vapor, which strips nitrogen dioxide from the concentrated nitric acid. The bleached concentrated acid goes to the reboiler section of the bleacher, where it is partially evaporated and partially withdrawn as product acid. The vapor exiting the bleacher contains ca. 95 % nitrogen dioxide and 5 % nitric acid. Most of the vapor condenses in a downstream precondenser (q). Dinitrogen tetroxide is liquefied in the liquefier and fed to a stirred tank (r).

Production of Concentrated Nitric Acid. The formation of concentrated nitric acid in the reactor (s) is described by the following equation:

$$4\,NO_2(\rightleftharpoons 2\,N_2O_4) + O_2 + 2\,H_2O \rightleftharpoons 4\,HNO_3$$

This requires a $N_2O_4 - H_2O$ molar ratio of 1 : 1 and a weight ratio of 5.11 : 1. The reaction time can be considerably shortened if dinitrogen tetroxide is present in excess. On the other hand, an excess of dinitrogen tetroxide lowers the production rate of the reactor and increases the dinitrogen tetroxide recycle rate, resulting in higher steam consumption for bleaching and a greater cold requirement for dinitrogen tetroxide liquefaction.

The $N_2O_4 - H_2O$ ratio, denoted the R factor, is the most important criterion for monitoring reactor operation. In practice, it fluctuates between 6.5 and 7. A product concentration of > 98 wt % nitric acid requires a higher R factor. The economically most favorable R factor can be found only by empirical means; the ratio can be varied to achieve optimal operating conditions. The makeup acid used to maintain the water balance of the raw acid mixture is taken from the bottom of the final absorber (n).

Concentrated nitric acid is formed in the reactor (s) at ca. 5.0 MPa and 60 – 80 °C. High mechanical strength and corrosion resistance are needed, but these two requirements cannot be satisfied by any one material. The reactor thus has a carbon steel jacket to provide mechanical strength and a pure aluminum inner shell to provide corrosion resistance. Unreacted oxygen is supplied to the suction of the nitrous gas blower via the venting system. The concentrated nitric acid product is led to the raw acid tank (o) and then to the bleacher. The bleached concentrated acid is cooled to + 30 °C (u). Most of the concentrated nitric acid is returned to the absorption tower and final oxidation.

Indirect Processes. Two types of indirect process are used to produce concentrated nitric acid (i.e., product containing > 97 wt % HNO_3):

1) Sulfuric acid process
2) Magnesium nitrate process

Both concentration techniques are based on extractive distillatfication. Weak acid is first produced by a conventional process (Section 1.3.4.1), then a third component is added to extract the water so that up to ca. 99 % nitric acid can be distilled from the ternary mixture.

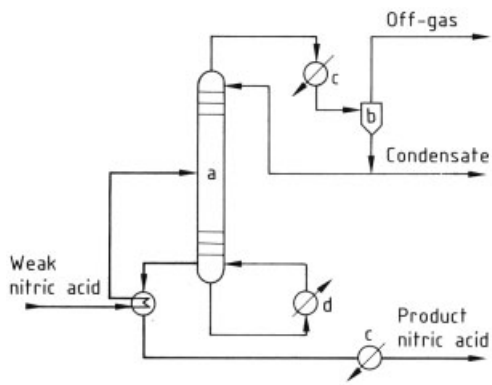

Figure 31. Simplified flow sheet of a process for preconcentration of nitric acid
a) Preconcentrating tower; b) Separator; c) Recirculating evaporator system; d) Cooler

The starting product is ordinary commercial nitric acid (ca. 55 – 65 % HNO_3). Weaker acids

can be preconcentrated in a single-stage or multistage process to give ca. 68 wt % nitric acid. Figure 31 illustrates a single-stage apparatus. Preconcentration takes place in a continuous countercurrent distillation tower (a), which can be of either the packed or the bubble-tray type. The system includes one or, if appropriate, two preheaters, a recirculating evaporator system for the tower bottom, a stripping and rectifying section for the tower, an overhead condenser, and a reflux separator (b).

Sulfuric Acid Process. Sulfuric acid can be used as extracting agent (see also → 4. Distillation and Rectification). The concentrated nitric acid product is clear and colorless; it contains 98 – 99 wt % nitric acid and < 0.05 % nitrogen dioxide.

Consumption figures for the production of 1 t of 99 % nitric acid from dilute nitric acid are given in Table 6.

Table 6. Consumption figures for the production of 1 t of 99 % HNO_3 from dilute nitric acid by the sulfuric acid process∗

Parameter	Starting HNO_3 concentration, wt %		
	55	60	65
Heating steam (1.0 – 1.8 MPa), t	2.0	1.75	1.45
Cooling water, m^3	80	60	50
Electric energy, kW · h	17	14	11
Water evaporated, t	0.82	0.66	0.53

∗ Source: Plinke – NASAC plant data.

The nitric acid concentration step now operates with indirect heating, an advance over earlier technologies that reduces steam consumption by as much as 50 %.

The materials of construction are highly corrosion resistant. Towers are made of borosilicate glass, enameled steel, and steel – polytetrafluoroethylene. Heat exchangers are made of glass, polytetrafluoroethylene, stainless steel, tantalum, titanium, high-purity aluminum, and special alloys.

Figure 32 is a flow sheet of a plant making concentrated nitric acid by the sulfuric acid process. The feed nitric acid, concentrated to about 68 wt %, is preheated (e) and then fed into the distillation tower (a). At least 80 wt % sulfuric acid is fed to the head of the tower. The part of the tower above the nitric acid inlet, which is irrigated with sulfuric acid, can be regarded as the rectifying section, with the sulfuric acid also functioning as reflux. The part of the tower below the nitric acid inlet is the stripping section.

A circulating evaporator system takes care of bottom heating. The ca. 70 % sulfuric acid leaving the bottom of the tower goes to the concentrator (c), which operates under vacuum (8 kPa). The overhead vapor product from the tower is condensed to form 99 % nitric acid and then deaerated (b). The tail gases, which still contain nitric acid vapor, are scrubbed with dilute nitric acid (d).

Magnesium Nitrate Process. In this process a magnesium nitrate solution is used to extract water from the nitric acid. The resulting dilute magnesium nitrate solution is restored to the de-

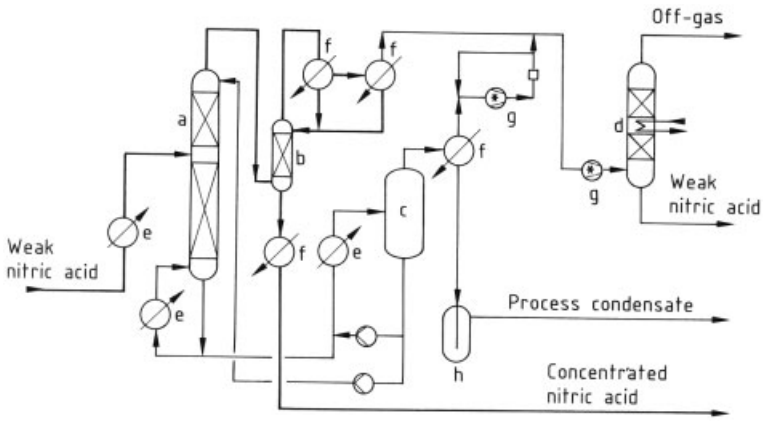

Figure 32. Simplified flow sheet of a process for concentrating nitric acid with sulfuric acid (Plinke – NACSAL process)
a) Concentrating tower; b) Condenser – deaerator; c) Sulfuric acid concentrating tower; d) Tail-gas treatment; e) Preheater; f) Cooler; g) Blower; h) Separator

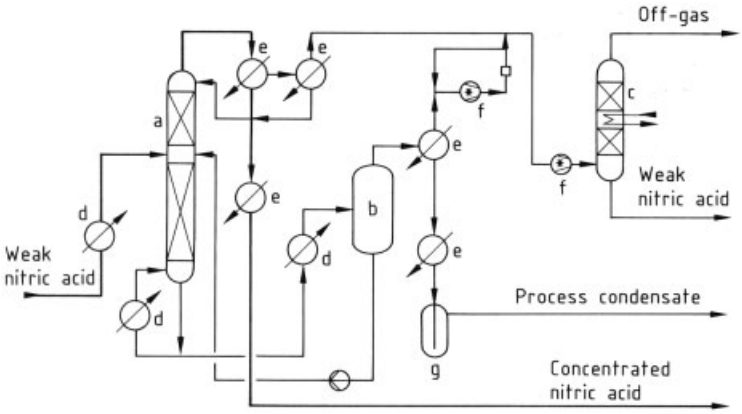

Figure 33. Simplified flow sheet of a process for concentrating nitric acid with magnesium nitrate (Plinke – MAGNAC process) a) Concentrating tower; b) Vacuum evaporator; c) Tail-gas treatment; d) Preheater; e) Cooler; f) Blower; g) Separator

sired working concentration of about 72 % in a vacuum concentrator before being returned to the nitric acid concentration step.

Consumption figures for the production of 1 t of 99 % nitric acid from dilute nitric acid are given in Table 7.

Table 7. Consumption figures for the production of 1 t of 99 % HNO_3 from dilute nitric acid by the magnesium nitrate process ∗

Parameter	Starting HNO_3 concentration, wt %		
	55	60	65
Heating steam (1.0 – 1.8 MPa), t	2.0	1.75	1.45
Cooling water, m^3	80	70	60
Electric energy, kW · h	10	9	8
Water evaporated, t	0.82	0.66	0.53

∗ Source: Plinke – MAGNAC plant data.

Figure 33 is the flow sheet of a plant making concentrated nitric acid by the magnesium nitrate process. Weak acid is fed to the dewatering tower (a). Extractive distillation with a concentrated (72 wt %) solution of magnesium nitrate at 140 °C gives an overhead product containing 99 % nitric acid. A small fraction of the condensed overhead product is refluxed to the dewatering tower. The bottom of the tower is heated, and the dilute magnesium nitrate solution, containing < 0.1 % nitric acid, is led to the vacuum evaporator (b). Condensed weak acid is processed in the recovery tower (c) and recycled if appropriate. The tail gas still contains nitric acid vapor and can be scrubbed with dilute nitric acid (c).

Comparison of Indirect Processes. Which concentration process is better suited to a given task must be decided case by case. From a process-engineering standpoint, the magnesium nitrate process has the following advantages:

1) The installation is very compact. No holding tanks are needed, which reduces energy losses.
2) Because the reconcentration step for magnesium nitrate is carried out in stainless steel vessels, virtually no critical construction materials are required.
3) Magnesium nitrate can be transported in sacks, and is therefore much easier to handle than sulfuric acid.
4) The weak nitric acid condensate, which is produced when the magnesium nitrate solution is reconcentrated, is partly reused in NO_x absorption. Losses of nitric acid are very slight.

As in weak acid production, development of concentrated acid production has followed different paths in the United States and the Federal Republic of Germany (and, in part, in Europe). In the United States estimated operating costs are significantly lower for concentration with magnesium nitrate than with sulfuric acid. The U.S. view is largely that, even after the change from autoclaves to continuous direct production, the equipment cost (i.e., investment cost) is too high for the latter process to be competitive.

1.4. Environmental Protection

In recent years, more stringent water pollution regulations have taken effect and legislative provisions for air pollution have been amended. New processes have been developed and introduced to significantly reduce pollutant emissions from nitric acid plants. The problem of nitrogen oxide pollution is treated in more detail in → Air.

1.4.1. Wastewater

Wastewater problems can be overcome by appropriate design of the nitric acid plant. An especially simple form of wastewater treatment can be employed when the nitric acid is processed directly for the production of mineral fertilizers.

The solution from ammonia stripping contains up to 10 % ammonia; it can be neutralized with nitric acid and then subjected to absorption along with process water. The acid product then contains a small amount of ammonium nitrate, but this is not a problem when the acid is processed into mineral fertilizers.

Leaks from pumps, vessels, etc., are pumped into a separate acid drain tank, and then processed directly or indirectly. The apparatus used for this purpose is completely separate from the sewage system and thus prevents contamination of wastewater.

If heat is removed by recooling systems, cooling water blowdown is fed into the wastewater system to limit thickening of the cooling water. Fresh water must be supplied to compensate for the blowdown and evaporation losses. Corrosion inhibitors, hardness stabilizers, and in some cases, biocides are also carried out of the system by blowdown.

1.4.2. Stack Gas

Obsolescent nitric acid plants can be recognized by their strongly colored yellow to reddish brown plumes of stack gas. The coloration is due to nitrogen dioxide. Some stack gases contain up to 3000 ppm NO_x; very old plants may surpass these values [64].

The absorption of nitrous gases with water forms part of the nitric acid process. However, the laws of nature prevent complete absorption, and some residual emission cannot be avoided (see Section 1.3.2). These emissions can be minimized by optimizing process conditions, increasing the efficiency of absorption, or using special tail-gas treatment methods.

Crucial parameters for the absorption of nitrous gases in water are the following [65]:

1) Pressure
2) Temperature
3) Reaction volume
4) The efficiency of the absorption tower
5) The partial pressures of nitrogen oxides and oxygen

Emissions in the Federal Republic of Germany in 1986 totaled ca. 3×10^6 t NO_x [66]. Transportation, households, and small consumers accounted for ca. 66 % of this; power plants and district heating plants for ca. 25 %; and industry, including furnaces, for ca. 9 %. Nitrogen oxides from nitric acid production were responsible for < 1 % of the total emissions.

1.4.2.1. Emission Limits

Heightened sensitivity to environmental concerns, coupled with the need for larger and more efficient production systems, will lead to further reductions in NO_x emissions. Legislators are following this trend and modifying pollution limits in accordance with changing technical capabilities.

At the beginning of the 1980s the first general administrative regulation was issued in the Federal Republic of Germany under the Federal Pollution Control Act: the Technische Anleitung zur Reinhaltung der Luft (TA-Luft; Technical Instructions for Air Pollution Control) [67]. This regulation and the corresponding VDI guidelines [68] form the basis for issuing operating permits. Regulations in many other countries are based on these guidelines.

The designation NO_x is commonly used to specify emission and concentration values because stack gases always include mixtures of nitrogen oxides. For the purpose of standardization, the oxides are calculated in terms of nitrogen dioxide. The acceptable concentrations from TA-Luft 1986 [67] are

1) Mass of substances emitted as mass concentration in units of g/m^3 or mg/m^3, referred to the volume of stack gas at standard conditions (0 °C, 101.3 kPa) after deducting the moisture content of water vapor, or to the volume of stack gas at standard conditions (0 °C, 101.3 kPA) before deducting the moisture content of water vapor; and
2) Mass of substances emitted, referred to time, as mass velocity in units of kg/h, g/h, or mg/h; and
3) Ratio of the mass of substances emitted to the mass of products generated or processed (emission factors), as mass ratio in units of kg/t or g/t.

The usual concentration unit ppm (mL/m^3) can be converted to the prescribed mass concentration (mg/m^3) with the following factor [69]: 1 ppm = 1.88 mg/m^3 (based on monomeric NO_2 at 101.3 kPa and 298 K and ideal-gas behavior).

The TA-Luft [67] states that the emission of nitrogen monoxide and nitrogen dioxide in the stack gas of nitric acid plants must not exceed 0.45 g/m^3 as nitrogen dioxide. Furthermore, stack gases may be discharged only if colorless; as a rule, this is the case if the mass concentration of nitrogen dioxide in the stack gas does not exceed the value given by the following formula:

$$\text{Mass concentration of nitrogen dioxide}\left(\text{mg/m}^3\right) = \frac{1200}{\text{Internal diameter of stack orifice (dm)}}$$

The requirement of a colorless discharge sets a practical limit of < 200 ppm NO_x. Older low- and medium-pressure plants should comply with these standards by March 1, 1996.

Plants whose emission of nitrogen monoxide and nitrogen dioxide (as nitrogen dioxide) exceeds a rate of 30 kg/h must be equipped with devices that continuously measure the mass concentration of nitrogen oxides. Quantitative relationships between emission and ground-level concentration are also described, thus setting guidelines for stack design.

If ground-level concentration limits are exceeded, the TA-Luft emission limits are further reduced by the relevant authorities. The ground-level concentration limits for gaseous pollutants in air are as follows [67]: nitrogen dioxide, long-term exposure (IW1), 0.08 mg/m^3; short-term exposure (IW2), 0.2 mg/m^3.

1.4.2.2. Analysis

The guidelines for the measurement and monitoring of NO_x emissions from nitric acid production are specified in [67, 70, 71]. *Photometry* without auxiliary chemical reaction and *chemiluminescence* [71, c – e] are suitable for the continuous measurement of NO_x emissions [65].

Difficulties in gas sample preparation due to condensation are discussed in [71, c – g]. Additonal apparatus-related problems occur when chemiluminescence is used to measure emission of moist stack gases at higher NO_x levels. A chemiluminescence analyzer with a maximum range of 0 – 10 000 ppm has been developed [72].

For stack gases containing traces of ammonia, precipitation of ammonium salts in the measuring cells has prevented continuous measurements. Two methods useful for the discontinuous measurement of nitrogen oxide emissions [65] are acidimetric titration and the photometric phenoldisulfonic acid technique [71, a, b].

1.4.2.3. Control of NO_x Emissions

Four basic approaches are used to reduce tail-gas NO_x levels: improved absorption, chemical scrubbing, adsorption, and catalytic tail-gas reduction. Recent decades have seen intense research and development effort invested in these methods. More stringent environmental restrictions have lent special urgency to development in this area. The most important techniques are covered in this section.

Improved Absorption. Absorption efficiency depends chiefly on the absorption pressure, the number of stages, and the temperature. The temperature of the gas between stages is especially important because it governs the progress of oxidation, which is the limiting quantity for the entire absorption process (see Section 1.3.2). In an already existing nitric acid plant, the options are to expand the absorption volume and/or lower the absorption temperature. In the first approach, large additional volumes

only result in small reductions of tail-gas NO_x levels because the oxidation of nitrogen monoxide to nitrogen dioxide proceeds very slowly when the NO_x concentration is low. The drawback to this method is that added absorption volume in stainless steel is very expensive. The advantage is that it does not require any new technology. The absorption volume is added in the form of a second tower; new absorption tower designs with very few stages have been devised [73].

The use of cold energy in the absorption process greatly accelerates oxidation of nitrogen monoxide to nitrogen dioxide. The disadvantage of this method is that the necessary refrigeration equipment and piping demand further investment. Another technique involving the use of cold to lower the NO_x level is to cool the nitrous gas so that more dinitrogen tetroxide than nitrogen dioxide is formed. The dinitrogen tetroxide is then scrubbed with nitric acid at ca. 0 °C; it can be stripped out, and converted to nitric acid by the reaction

$$N_2O_4 + 1/2\ O_2 + H_2O \longrightarrow 2\ HNO_3$$

in an absorption reactor at ca. 60 – 80 °C. The technique is similar to that in concentrated nitric acid production. The method is, however, more suitable for investment in a plant being designed than for the expansion or upgrading of an existing plant.

Chemical Scrubbing. A number of patents and scientific publications deal with options for scrubbing NO_x out of tail gas. Problems related to the scrub liquor (cost, regeneration, quantity, and environmental impact) are encountered in all methods. These problems are always easier to manage if the nitric acid plant is part of an integrated chemical plant, which is usually the case. However, in some instances a nitric acid plant or sections of a fertilizer complex must operate independently. The following scrub liquors have been proposed: aqueous suspension of magnesium carbonate and magnesium hydroxide [74]; solution of vanadium in nitric acid [75]; ammonium sulfide and bisulfide [76]; milk of lime [77]; ammonia [78]; hydrogen peroxide [79]; and urea [80].

Ammonia scrubbing is used in the United States. Goodpasture (Texas) has developed a safe process based on scrubbing the tail gas with ammonium nitrate solution; nitrite formation is suppressed by aeration and the presence of free acid. The tail gas is then led through an ammoniacal scrub liquor, pH 7.5 – 8.5. The scrubbing product is a 30 – 50 % ammonium nitrate solution. The tail gas has a residual level of ca. 200 ppm NO_x.

Hydrogen peroxide scrubbing is based on the following overall reactions:

$$NO + NO_2 + 2\,H_2O_2 \longrightarrow 2\ HNO_3 + H_2O$$

$$2\,NO_2 + H_2O_2 \longrightarrow 2\ HNO_3$$

The reactions are carried on sieve trays or in packed towers with recirculation of the hydrogen peroxide solution. The advantage of this scrubbing process is that the reaction time is very fast; the disadvantage, that the hydrogen peroxide scrub liquor is expensive.

A solution containing 20 % urea and 10 % free nitric acid has also been suggested for scrubbing tail gas to remove NO_x. The process is carried at ca. 50 °C. Both nitrogen oxides are selectively reduced to nitrogen by urea, which decomposes to yield nitrogen and carbon dioxide. The process, developed by Norsk Hydro in Norway, has the advantage that urea is readily available and relatively cheap. The resulting stack gas plume is colorless but heavily laden with water vapor.

Adsorption Processes. The adsorption of NO_x by molecular sieves has long been known but has not yet been tested intensively in full-scale nitric acid plants.

The Pura-Siv-N process (Union Carbide) is claimed to reduce stack gas levels below 50 ppm nitrogen dioxide [81], but high investment costs and other problems have prevented its adoption. A patent granted to Kernforschungsanlage Jülich GmbH has a similar object [82], but again industrial experience is slight. A "wet" adsorption system has also been developed by Cofaz (France). A carbon- containing adsorbent is sprayed with water or dilute nitric acid and brought in contact with the tail gas. This process has been installed in three monopressure plants [83].

Catalytic Reduction. Catalytic reduction was the first method used to reduce NO_x emis-

sions; as early as 1924, a patent for such a process was issued to FAUSER [84]. A fuel is added, at a level below the flash point, to the nitrogen oxides which then react on the catalyst surface to form nitrogen and water vapor. Most fuels (e.g., hydrocarbons and hydrogen), however, react preferentially with the free oxygen that is always present in tail gases from nitric acid production. As a result, the nitrogen oxides are not destroyed until free oxygen has been consumed. Under these conditions the reactions with methane can be described as

$$CH_4 + 2\,O_2 \longrightarrow CO_2 + 2\,H_2O$$

$$CH_4 + 4\,NO \longrightarrow CO_2 + 2\,H_2O + 2\,N_2$$

$$CH_4 + 2\,NO_2 \longrightarrow CO_2 + 2\,H_2O + N_2$$

Ammonia can also be used as a reducing agent, preferentially reducing the nitrogen oxides:

$$6\,NO + 4\,NH_3 \longrightarrow 5\,N_2 + 6\,H_2O$$

$$6\,NO_2 + 8\,NH_3 \longrightarrow 7\,N_2 + 12\,H_2O$$

$$NO + NO_2 + 2\,NH_3 \longrightarrow 2\,N_2 + 3\,H_2O$$

$$4\,NO + O_2 + 4\,NH_3 \longrightarrow 4\,N_2 + 6\,H_2O$$

Catalytic reduction processes are accordingly classified as nonselective or selective.

Catalysts for *nonselective reduction processes* are usually based on platinum, vanadium pentoxide, iron oxide, or titanium. The fuel requirement is the stoichiometric amount needed to reduce all the oxygen present (free and in nitrogen oxides) plus a small excess (ca. 0.5 vol % CH_4).

Unfortunately, the principal byproducts of this process are carbon monoxide (≤ 1000 ppm) and hydrogen cyanide. When hydrocarbon fuels are used, the tail gas must be preheated before the reaction on the catalyst proceeds at all. The preheat temperature depends directly on the fuel selected:

Natural gas	450 – 480 °C
Propane	340 °C
Butane	340 °C
Naphtha	340 °C
Hydrogen	250 °C

The use of hydrogen allows the preheat temperature to be lowered to 150 – 200 °C but is generally too expensive.

If the quantity of fuel is not enough to reduce all the oxygen, nitrogen dioxide is reduced only to nitrogen monoxide. The stack gas is then decolorized; in some countries, this is sufficient. Complete NO_x removal generally requires preheating, multiple fixed-bed catalysts (with cooling between beds to prevent overheating), and careful heat recovery to offset part of the fuel cost.

Most nitric acid plants constructed or modified by Weatherly use nonselective catalytic NO_x abatement units with tail-gas exit temperatures of 650 – 675 °C. Some plants have gas exit temperatures > 815 °C, which is higher than the maximum for admission to the tail-gas expansion turbines. The tail gas must then be cooled, either against tail gas entering the catalytic unit or with the aid of waste-heat boilers.

In the process marketed by Du Pont the tail gas is dried, preheated, and then heated in two stages. The gas is subsequently divided into two streams. One substream is heated further, mixed with methane, and led over the first fixed-bed catalyst. The second substream is mixed with methane without further heating and led over the second fixed-bed catalyst, along with the first substream. Hot tail gases exit the bottom of the reactor and are led through the tail-gas expansion turbine, which recovers part of the work of compression.

Ammonia is the only economically relevant reducing agent for *selective catalytic processes*. The consumption of reducing agent in the selective treatment of NO_x is much smaller than in nonselective processes. Furthermore, the tail-gas temperature after reduction is significantly lower, allowing the use of simpler, cheaper construction materials. The optimal catalyst service temperature is 250 – 350 °C, but operation at up to 500 °C is possible. The costs due to the use of expensive ammonia must, however, also be considered. To prevent formation of ammonium nitrate (a potential explosion hazard) in the expansion turbine or downstream, emissions must be monitored for ammonia (10 – 20 ppm).

When BASF, Ludwigshafen [85] began work on catalytic NO_x reduction in the early 1960s, existing patents indicated the use of ammonia together with noble-metal (Pt, Rh, Ru, Pd) and iron-group (Fe, Co, Ni) catalysts [86, 87]. Alternatives were found: vanadium pentoxide, tungsten oxide, and molybdenum oxide gave excellent results. Vanadium pentoxide on an alumina support proved to be most economical.

Selective processes also require oxygen, which oxidizes part of the nitrogen monoxide to nitrogen dioxide to ensure roughly equal levels of these two oxides (the best condition for reduction). Because the BASF process operates at 200 – 350 °C, side reactions between ammonia and oxygen can be neglected. The reaction is not affected by dinitrogen monoxide, carbon dioxide, or water vapor.

For a plant with a tail-gas stream of 37 000 m^3 /h (STP) operating at a pressure of 730 kPa with an inlet NO_x concentration of 500 –1000 ppm, this process (270 °C) gives an exit concentration of 50 – 150 ppm NO_x. The treated tail gas contains < 20 ppm ammonia (usually ca. 5 ppm), and the nitrogen dioxide level is 30 –50 ppm. The stack gas is generally colorless. Careful temperature monitoring at the tail-gas expansion turbine and measurement of the ammonia level can prevent the deposition of ammonium nitrate. To minimize the pressure drop, the reactor features an annular design with inward flow (Fig. 34).

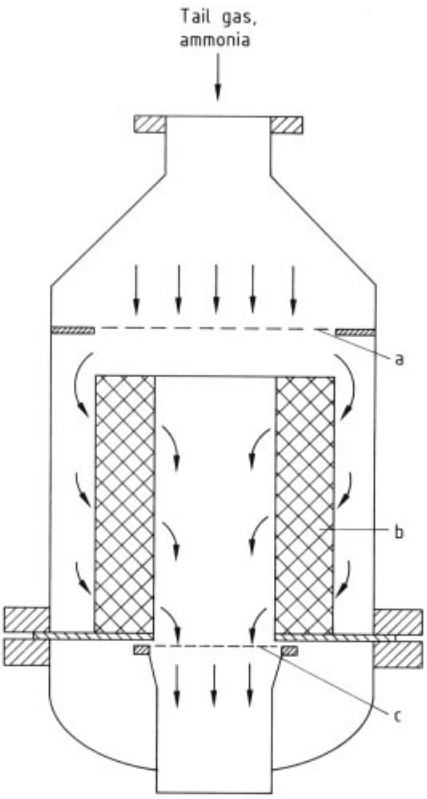

Figure 34. BASF tail-gas reactor for NO_x reduction
a) Hole plate; b) Catalyst; c) Wire gauze

Ammonia is uniformly mixed with the tail-gas stream in static mixers. The performance of the treatment process is monitored by measuring the temperature rise during reduction (ca. 10 °C/1000 ppm NO_x).

Uhde, Dortmund, as licensee, has retrofitted 14 plants with the BASF process. The tail gas, usually containing 500 – 1000 ppm NO_x, is heated to about 260 °C. Ammonia gas is added to the tail-gas stream in a mixer and then led through the reactor. The pressure drop is ca. 25 kPa. The effluent meets the requirement of colorlessness.

The HGW – Didier process, developed jointly by Hamburger Gaswerke (HGW) and Didier, Essen uses a chromium oxide catalyst with a service temperature of 265 – 350 °C. The nitric acid plant planned for DSM Miststoffen (Geleen, Netherlands) is designed for a tail-gas stream of 76 000 m^3/h (STP) with a maximum NO_x level of 2200 ppm. The exit concentration will not exceed 100 ppm NO_x. The tail gas is preheated before entering the reactor. To meet the strict emission limit of < 100 ppm, the catalyst is located in three beds. Ammonia can be admitted separately over each. The following distribution has proved most efficient: 70 % over the first bed, 20 % over the second, and 10 % over the third catalyst bed.

Other companies also have selective catalytic reduction processes for nitric acid tail gases. For example, Mitsubishi Chemical has developed a process operating at 400 – 500 °C under pressure [88, 89]. The "DN-Cat" catalyst is installed between the heat exchanger and the tail-gas expansion turbine in the conventional flow sheet.

The Weatherly selective catalytic process operates at 250 – 310 °C. The noble-metal catalysts are the same ones employed in nonselective processes, except that the fuel is replaced by ammonia.

The Bergbauforschung – Uhde process [90] is chiefly employed for stack-gas cleanup at power plants but may later find use in emission abatement at nitric acid plants. The Bergbauforschung – Uhde process for simultaneous sulfur and NO_x removal from stack gases is utilized at the cold end of the power-plant boiler, i.e., downstream of the air preheater and electrostatic filter. It employs the adsorptive and catalytic properties of "activated coke" at 50 – 150 °C. The reducing agent is again gaseous am-

monia. The technique combines a selective catalytic process for NO_x and an adsorptive one for sulfur dioxide. The catalyst is therefore installed in a moving-bed reactor.

1.5. Storage and Transportation

As a rule, nitric acid is stored in stainless steel tanks and transported in stainless steel containers.

The regulations for transport by rail [91,92] and road [93,94] differentiate among > 70 % nitric acid, 55 – 70 % nitric acid, and < 55 % nitric acid, and also between mixtures with sulfuric acid ("mixed acid") containing > 30 % nitric acid and those containing < 30 % nitric acid. Similar regulations [95] apply to transport by ship. Important transport regulations for > 70 % nitric are as follows:

IMDG Code	Class 8, D 8261, E – F 8186
UN No.	2032
RID/ADR	Class 8, no. 2 a
CFR	49: 172.101, Cor. M

Regulations for < 70 % nitric acid are

IMDG Code	Class 8, D 8260, E – F 8185
UN No.	2031
RID/ADR	Class 8, no. 2 b
CFR	49: 172.101, Cor. M

Provisions range from details of technical equipment to labeling and give instructions in case of accidents. Regulations for the road transport of nitric acid ($\leq 70\,\%$ HNO_3 emphasize the associated hazards (poisoning by inhalation of vapors, danger of chemical burns on contact with flammable substances, formation of nitrous gases). Protective equipment includes respiratory protection apparatus, goggles, and clothing affording complete coverage, as well as an eyewash bottle with pure water. In case of accident, the fire and police departments are to be notified immediately. Spilled nitric acid must not be absorbed with sawdust or other flammable material (because of the fire hazard); instead, its spread must be prevented by the construction of earth barriers.

1.6. Uses and Economic Aspects

The principal use of nitric acid is as a starting material in the manufacture of nitrogen fertilizers; large amounts are reacted with ammonia to yield ammonium nitrate (→ Ammonia Compounds, Chap. 1.2.1.). Weak acid (ca. 60 % HNO_3) is most suitable for this purpose. Smaller amounts of weak acid are used to digest crude phosphates. About 75 – 85 % of the nitric acid produced is used in the fertilizer sector. In addition, porous ammonium nitrate prills are still an important component of explosives (→ Explosives, Chap. 7.2.).

Nitric acid is used as a nitrating agent in the preparation of explosives (→ Explosives) and organic intermediates such as nitroalkanes (→ Nitro Compounds, Aliphatic) and nitroaromatics (→ Nitro Compounds, Aromatic). It is also used in the production of adipic acid (→ Adipic Acid, Chap. 4.1.).

Other applications include use as a chemical in metallurgy (e.g., as an etchant and pickling agent for stainless steels) and in rocket fuel production.

Worldwide output of nitric acid in 1987 was 26.882×10^6 t, representing a slight increase of ca. 4 % over 1976. Production in industrialized countries has stagnated or even declined, but in 1976 – 1985 increases in capacity in less industrialized countries occurred [96]. Production figures for nitric acid in 1987 follow (10^6 t):

Europe, total	16.472
United States	6.553
Eastern Europe	4.289
Asia	1.544
Federal Republic of Germany	2.112
Poland	2.136
Belgium	1.470
Spain (1986)	1.249
Italy	1.195
Hungary	1.018

2. Nitrous Acid

Nitrous acid [*7782-77-6*], HNO_2, M_r 47.01, is a moderately strong to weak acid. Its dissociation constant K in highly dilute solutions at 18 °C is 4.5×10^{-4}. The acid is stable only in cold dilute aqueous solution. It can be prepared as follows:

$$Ba(NO_2)_2 + H_2SO_4 \longrightarrow 2\,HNO_2 + BaSO_4$$

On heating in the presence of sand, glass splinters, or other sharp-edged objects, or even at low temperature, it disproportionates as

$$3\,HNO_2 \rightleftharpoons HNO_3 + 2\,NO + H_2O$$

This disproportionation reaction influences the properties of nitrous acid solutions and is important in the production of nitric acid. Nitrous acid occurs as an intermediate product when NO_x gases are absorbed in water.

By virtue of the above reaction, nitrous acid can act as both a reducing and an oxidizing agent. Strong oxidizing agents such as potassium permanganate oxidize nitrous acid to nitric acid. It can be reduced to less oxygen-rich compounds (NO_2, NO, $H_2N_2O_2$, NH_3) by alkali amalgams or by electrolytic means. Its salts (nitrites) are not stable in aqueous solution because of hydrolysis.

Nitrous acid in aqueous solutions is assumed to be in dynamic equilibrium with its anhydride:

$$2\,HNO_2 \rightleftharpoons N_2O_3 + H_2O$$

Aqueous solutions of the acid appear colorless but take on an increasingly blue coloration at 5 % dinitrogen trioxide.

Figure 35 shows the phase diagram of the $N_2O_3 - H_2O$ system with its sizable miscibility gap. Mixtures break up into a lighter layer with a high water content and a heavy layer with a high dinitrogen trioxide content.

Nitrous acid exists in two isomeric forms:

These structures lead to two series of industrially important organic derivatives:

Nitrite esters Nitrocompounds

An especially important property of nitrous acid is its ability to diazotize organic amines. With primary amines, the acid forms diazonium salts:

$$R{-}NH_2 + HNO_2 + HCl \longrightarrow [R{-}N{\equiv}N]\,Cl + 2\,H_2O$$

These salts are used widely in the manufacture of azo dyes (→ Azo Dyes).

With secondary amines, nitrosamines are formed:

$$R_2{-}NH + HNO_2 \longrightarrow R_2{-}N{-}NO + H_2O$$

Ammonia and urea react with nitrous acid to yield nitrogen:

$$NH_3 + HNO_2 \longrightarrow N_2 + 2\,H_2O$$

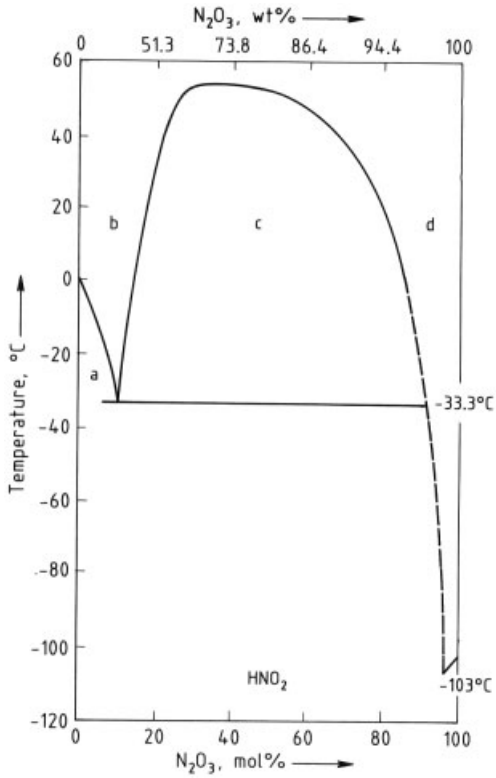

Figure 35. The system $N_2O_3 - H_2O$
a) Ice; b) Homogeneous liquid, water + dinitrogen trioxide; c) Liquid phase, two layers; d) Homogeneous liquid, dinitrogen trioxide + water

3. Nitrogen Oxides

Compounds of oxygen with nitrogen are considered as a class and called nitrogen oxides (often denoted as NO_x).

The known oxides and their equilibrium reactions are as follows:

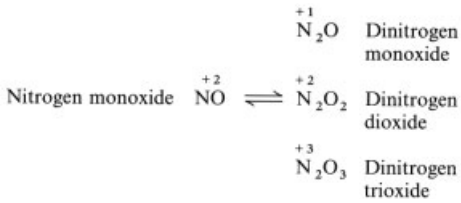

Table 8. Physical properties of nitrogen oxides

Compound	N_2O	NO	NO_2/N_2O_4	N_2O_3	N_2O_5
Oxidation state					
	+1	+2	+4/+4	+3	+5
T_{Cr}, °C	36.41	−93	157.85		
p_{Cr}, MPa	7.245	6.485	10.132		
ϱ_{Cr}, kg/m^3	452	520	550		
mp, °C	−90.86	−163.65	−11.2	−100.7	32.4 ∗
bp, °C	−88.48	−151.77	21.15	−40 to +3	
Specific heat c_p, kJ kg^{-1} K^{-1}	0.879	0.996	1.326	0.862	0.778
Standard enthalpy of formation					
ΔH_F^0, kJ/kg	1864.190	3007.684	721.199	1101.435	104.589
Heat of vaporization at *bp*, kJ/kg	376.07	459.031	414.257	517.416	
Density, kg/m^3					
Gas (0 °C, 101.3 kPa)	1.9775	1.3402	3.4 (20 °C)	1.447(2 °C)	2.05
Liquid (20 °C, 101.3 kPa)	793		1446.8		(solid)
Dynamic viscosity, mPa · s					
Gas (25 °C, 101.3 kPa)	14.874	19.184	12.838		
Thermal conductivity, W m^{-1} K^{-1}					
Gas (25 °C, 101.3 kPa)	0.01718	0.02573	0.1124		
Liquid (20 °C, 101.3 kPa)			0.1336		

∗ Sublimation point.

Nitrogen dioxide $\overset{+4}{NO_2} \rightleftharpoons \overset{+4}{N_2O_4}$ Dinitrogen tetroxide

$\overset{+5}{N_2O_5}$ Dinitrogen pentoxide

Nitrogen trioxide $\overset{+6}{NO_3} \rightleftharpoons \overset{+6}{N_2O_6}$ Dinitrogen hexoxide

Very little is known about nitrogen trioxide (NO_3) or its dimeric form (N_2O_6).

The compounds listed above are described in more detail in order of their oxidation state. Table 8 lists their most important physical properties.

3.1. Dinitrogen Monoxide

Properties (see Table 8). Under normal conditions (i.e., room temperature and atmospheric pressure), dinitrogen monoxide [*10024-97-2*], also called nitrous oxide, N_2O, M_r 44.01, is a colorless gas with a weak, pleasant odor and a sweetish taste. If inhaled, it can bring about a spasmodic inclination to laugh and a condition resembling drunkenness–hence, its historic name, laughing gas [97]. Two mesomeric forms in resonance with each other describe the electron distribution in the molecule [7]:

$$\overset{\delta-}{N}=\overset{\delta+}{N}=O \rightleftharpoons N\equiv\overset{\delta+}{N}-\overset{\delta-}{O}$$

Production. Dinitrogen monoxide occurs as a byproduct of denitrification (especially in the heavy use of nitrogen fertilizers) and, to some extent, in the nitrification of nitration waste acids. The compound was first prepared by PRIESTLEY, who carefully heated ammonium nitrate:

$$NH_4NO_3 \longrightarrow N_2O + 2\,H_2O$$

This is still a common route for industrial production.

In earlier processes [98], high-purity ammonium nitrate was used (contamination by organic material and chlorides had to be avoided), and temperatures up to 260 °C were reached. Today, in contrast, ammonium nitrate can be decomposed in aqueous solution containing chlorides and nitric acid at temperatures as low as 100–160 °C. This synthesis was introduced in 1970 by Hoechst [99] (Fig. 36).

The decompositon flask (a) is supplied with gaseous ammonia and nitric acid (30–100 %). Ammonia gas forms the atmosphere for the entire system, including the tower (b). A small amount of chloride ion (> 0.2 %) is maintained in the aqueous nitric acid solution in the decomposition flask (a) to accelerate the reaction. If appropriate, catalytically useful amounts of metal ions (manganese, copper, cerium, lead, bismuth, cobalt, or nickel) may also be added. Heat liberated during the reaction raises the tempera-

ture to 100 – 160 °C. The solution contains 15 – 35 % nitric acid, 30 – 50 % ammonium nitrate, 15 – 55 % water, and 0.01 – 0.1 % chloride ion. The impure product dinitrogen monoxide ascends from the decomposition flask (a) into the tower (b). Ammonia in the gas space neutralizes any entrained acidic constituents, and this reaction heats the lower part of the tower (b). The hot gas is cooled and scrubbed by an aqueous ammonium nitrate solution fed through a distributor (c). The solution collected in the bottom of the tower (b) has a temperature of 50 – 110 °C; most of itis recycled through the heat exchanger (d). The thermal energy recovered serves to heat the evaporator (e), which concentrates a small part of the solution stream for recycling to the decomposition flask (a). If the temperature profile in the system is not too low, the process does not need a heat supply.

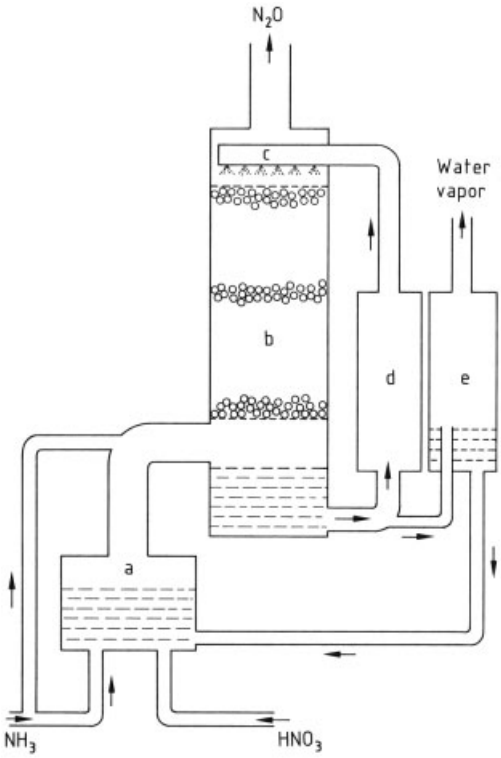

Figure 36. Industrial production of dinitrogen monoxide [99]
a) Decomposition flask; b) Scrubbing tower; c) Ammonium nitrate distributor; d) Heat exchanger; e) Evaporator

This process has two main advantages over earlier ones: it offers a substantial energy saving and produces dinitrogen monoxide free of harmful byproducts.

The dinitrogen monoxide can easily by liquefied under pressure and is available in steel cylinders. Transport regulations for dinitrogen monoxide (compressed) are as follows:

IMDG Code	Class 2.2
UN No.	1070
RID/ADR	Class 2, no. 5 a
CFR	49: 172.102 Nonfla. G

Uses. The principal uses of dinitrogen monoxide are as follows:

1) In medicine, as an anesthetic
2) In the munitions – explosives industry, as a propellant
3) In the food industry, as a foaming agent (whipped cream)

Production data for dinitrogen monoxide are given in [100].

3.2. Nitrogen Monoxide

Nitrogen monoxide [*10102-43-9*], also called nitric oxide, NO, M_r 30.01, is a colorless, toxic, nonflammable gas at room temperature. As soon as it comes in contact with atmospheric oxygen, it is oxidized to nitrogen dioxide, a brown vapor. For physical properties, see Table 8.

Nitrogen monoxide is at dynamic equilibrium with its dimer:

$$2\,NO \rightleftharpoons N_2O_2 \Delta H = -10.5\ \text{kJ/mol}$$

At high temperature, however, the equilibrium is shifted all the way toward nitrogen monoxide. Liquid nitrogen monoxide is completely dimerized, probably as angular ONNO molecules. The nitrogen monoxide molecule contains a double bond and a three-electron bond [101]

This makes the molecule more stable than the resonant structure

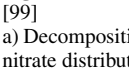

so that the heat of dimerization is slight; the reaction therefore does not occur in the gas phase. The nitrogen monoxide molecule contains 11

valence electrons and thus, like all gases with an odd number of electrons, is paramagnetic.

Laboratory preparation usually involves the reaction of moderately concentrated nitric acid with copper:

$$8\,HNO_3 + 3\,Cu \rightleftharpoons 3\,Cu(NO_3)_2 + 4\,H_2O + 2\,NO$$

Nitrogen monoxide is very reactive and can, as a strongly endothermic compound, be prepared from nitrogen and oxygen in an electric flame arc:

$$N_2 + O_2 \longrightarrow 2\,NO \quad \Delta H = 90.4\,\text{kJ/mol}$$

In acidic solution (e.g., chromic acid, acidified permanganate solution, hydrochloric acid), nitrogen monoxide is oxidized to nitric acid. The combustion of nitrogen (Birkeland – Eyde, Schönherr, Pauling processes) was formerly an important method of industrial nitric acid production; however, these processes are no longer used due to the high energy cost of the arc. Nitric acid production is now based on the combustion of ammonia; nitrogen monoxide is the first reaction product and has thus gained some importance as an intermediate.

Another use of nitrogen monoxide is in nitrosation, where chlorine or bromine reacts with it to yield nitrosyl halides. In the lead- chamber process for sulfuric acid production, nitrosyl hydrogen sulfate is treated with sulfurous acid in the Glover tower, to yield nitrogen monoxide and sulfuric acid.

With aqueous iron sulfate solutions, nitrogen monoxide forms dark brown nitroso iron (II) sulfate [$(FeNO)SO_4$], thus affording a detection method for nitric acid and nitrates.

Transport regulation are as follows:

IMDG Code	Class 2, D, E – F 2095
UN No.	1660
RID/ADR	Class 2, no. 1 ct
CFR	49: 172.101 Nonfla. G

3.3. Nitrogen Dioxide and Dinitrogen Tetroxide

Nitrogen dioxide [*10102-44-0*], NO_2, M_r 46.01, is a brownish red, toxic gas with a pungent odor; for physical properties, see Table 8. It is in equilibrium with its dimer, dinitrogen tetroxide [*10544-72-6*], also called nitrogen peroxide, N_2O_4, M_r 92.01. The equilibrium is strongly temperature dependent; at higher temperature (ca. 100 °C), it is shifted almost all the way toward nitrogen dioxide.

$$2\,NO_2 \rightleftharpoons N_2O_4 \quad \Delta H = -57.2\,\text{kJ/mol}$$

Above 150 °C, nitrogen dioxide begins to dissociate:

$$2\,NO_2 \rightleftharpoons 2\,NO + O_2 \quad \Delta H = 144.1\,\text{kJ/mol}$$

This reaction goes to completion at 650 °C. Figure 37 illustrates the dissociation of dinitrogen tetroxide and nitrogen dioxide as a function of temperature.

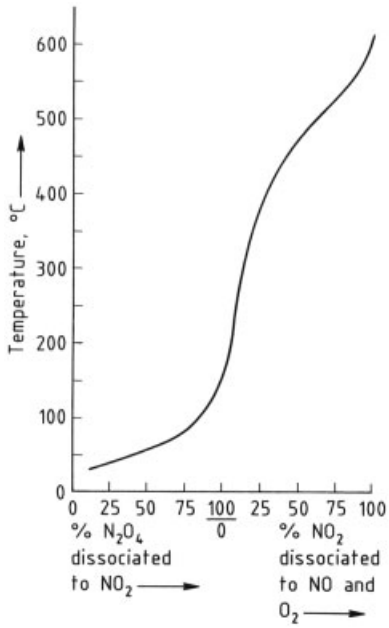

Figure 37. Dissociation of dinitrogen tetroxide and nitrogen dioxide

The following values for the equilibrium constant of the bidirectional reaction are given in [102]:

Temperature, K	273.2	298.5	323.5	359.7
K_p (101.3 kPa)	0.017513	0.1458	0.8492	7.247

According to PAULING, the nitrogen dioxide molecule probably exists in two mesomeric forms with two ionic structures:

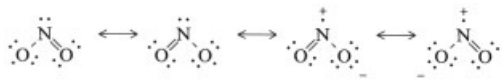

Dinitrogen tetroxide can be prepared in two structures, **1** exists at room temperature and is in equilibrium with **2** at low temperature:

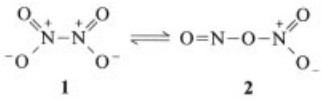

1 **2**

Nitrogen dioxide, like nitrogen monoxide, is an odd-electron compound and therefore paramagnetic; dinitrogen tetroxide is diamagnetic. To reach an even electron number, nitrogen dioxide can give up an electron (NO_2^+, nitronium ion) or gain one (NO_2^-, nitrite anion).

Industrial production of $NO_2-N_2O_4$ employs the Ostwald process (catalytic combustion of ammonia) and constitutes the initial step in the production of nitric acid (see Section 1.3).

Other commercial processes for producing $NO_2-N_2O_4$ are the oxidation of nitrosyl chloride yielding chlorine, and the treatment of sodium nitrite with nitric acid and oxidation of the liberated nitrogen monoxide to nitrogen dioxide.

High-purity $NO_2-N_2O_4$ is obtained in the production of sodium nitrate from sodium chloride and nitric acid. The dioxide and tetroxide (oxidation state 4) are strong oxidizing agents, comparable in strength to elemental bromine. They form explosive mixtures with hydrogen; mixtures with ammonia explode even at low temperature.

Transport regulations for liquefied nitrogen dioxide are as follows:

IMDG Code	Class 2, D 2213, E – F 2099
UN No.	1067
RID/ADR	Class 2, no. 3 at
CFR A	49

Nitrogen dioxide is involved in the production of " photochemical smog" and is a major air pollutant because it is produced in all combustion processes. It is therefore subject to strict emission limits (see Section 1.4.2.1) [67].

Dinitrogen tetroxide has many applications because of its strong oxidizing action. For example, it is employed as a catalyst in oxidation reactions and as an inhibitor in the distillation of acrylates. It is also used in the manufacture of explosives, as a rocket propellant, and as a bleach.

For details on the chemistry of liquid dinitrogen tetroxide, see [103].

3.4. Dinitrogen Trioxide

Dinitrogen trioxide [*10544-73-7*], also called nitrous anhydride, N_2O_3, M_r 76.01, is stable only below 0 °C, taking the form of a deep-blue liquid. When cooled further, it solidifies as pale blue crystals. For physical properties, see Table 8.

Even below 0 °C, dinitrogen trioxide dissociates to a considerable extent:

$$N_2O_3(g) \rightleftharpoons NO + NO_2 \quad \Delta H = 40.58\,\text{kJ/mol}$$

$$2\,N_2O_3(g) \rightleftharpoons 2\,NO + N_2O_4 \quad \Delta H = 12.97\,\text{kJ/mol}$$

At 10 °C and ambient pressure, only 10 % of the molecules are in the form of undissociated dinitrogen trioxide (as $ON-NO_2$ [104]). This dissociation means that a well-defined boiling point cannot be given for dinitrogen trioxide (*bp* range −40 to + 3 °C). The compound is prepared by saturating liquid nitrogen dioxide with nitrogen monoxide:

$$NO + NO_2 \rightleftharpoons N_2O_3$$

However, because nitrogen dioxide is present mainly as the dimer, a eutectic of nitrogen monoxide and nitrogen dioxide can be postulated [105] with the empirical composition $NO_{1.56}$. The equilibrium constant for the reaction has been measured at various temperatures [106]. Because the equilibrium is easily shifted, a mixture of nitrogen monoxide and nitrogen dioxide behaves chemically like dinitrogen trioxide.

Dinitrogen trioxide occurs chiefly in nitric acid production. As much as 4 vol % of the NO_x may be present as dinitrogen trioxide. The trioxide is used in preparing high-purity alkali nitrites from alkali. It also finds use as an oxidizing agent in special fuel systems.

3.5. Dinitrogen Pentoxide

Dinitrogen pentoxide [*10102-03-1*], also called nitrogen pentoxide, N_2O_5, M_r 108.01, is the anhydride of nitric acid. It is a colorless crystalline substance that melts under pressure but sublimes

at atmospheric pressure. For physical properties, see Table 8.

In the solid state, dinitrogen pentoxide has a salt-like structure $[NO_2]^+[NO_3]^-$. It is not very stable and can undergo spontaneous explosive decomposition [107]:

$$2\,N_2O_5 \longrightarrow 4\,NO_2 + O_2$$

Dinitrogen pentoxide can be prepared by treating nitric acid with phosphorus pentoxide. It has strong oxidizing properties and reacts vigorously with organic substances. With water, the hygroscopic crystals form nitric acid:

$$N_2O_5 + H_2O \longrightarrow 2\,HNO_3 \quad \Delta H = -37.66\,\text{kJ/mol}$$

With boron trifluoride, it forms a binary compound $N_2O_5 - BF_3$ that is stable at room temperature. This compound is a good nitrifying agent. Nitrogen pentoxide has not become important in industry.

4. Toxicology and Occupational Health

Nitric Acid. Nitric acid has a caustic effect on the skin and mucous membranes. Industrial safety regulations stipulate the following workplace concentrations [108, 109]: MAK value 10 ppm (25 mg/m^3); TLV-TWA 2 ppm (5.2 mg/m^3); TLV-STEL 4 ppm (10 mg/m^3). Nitric acid reacts with and causes spontaneous ignition of organic substances such as wood shavings, saw dust, cotton, and cellulose. Spilt acid should therefore never be absorbed with such substances.

Nitric acid is classified as a hazardous substance [109, 110]. Indications of special hazards (R warnings) and safety advisories (S warnings) are given in [109].

Nitric acid > 70 %	
R 8:	Danger of fire when in contact with flammable substances
R 35:	Causes severe burns
S 23:	Do not inhale vapor
S 26:	In case of contact with eyes, flush eyes thoroughly with water and call physician
S 36:	Wear appropriate protective clothing when working
Nitric acid 20 – 70 %	
R 35:	Same as above
S 2:	Must be kept away from children
S 23, S 26:	Same as above
S 27:	Immediately remove contaminated or saturated clothing
Mixtures of nitric acid (> 30 % HNO_3) and sulfuric acid (nitrating acid)	
R 8, R 35:	Same as above
S 23, S 26, S 36:	Same as above
S 30:	Never add water to acid

Protective clothing should be worn to prevent contact with the skin. The acid must be kept away from children. In case of burns, the affected area should be washed with copious amounts of water and a doctor should be consulted. Skin burns itch and become yellow due to the formation of xanthoproteins.

Death may occur if large areas of skin are burned. Since regulation of the body temperature is disturbed, the patient must be kept warm. Protein enters the bloodstream and the patient may suffer from shock, appropriate preventive measures must therefore be taken. In case of contact with the eye, the eyelids must be opened and the eye washed with copious amounts of water. A doctor must be consulted.

Nitrous gases are released from nitric acid at room temperature, a fume cupboard should therefore be used when working with the acid. Use of respiratory protection equipment with a B, E, or NO filter is recommended.

Oral consumption of nitric acid leads to attack of the mucous membranes of the mouth, the esophagus, and the stomach. Shock may result. If concious, the patient should drink lukewarm water to dilute the acid.

Nitrous Acid. Nitrous acid is toxic because it decomposes to form nitric acid and nitrogen monoxide.

Nitrogen Monoxide. Pure nitrogen monoxide does not have any irritating effects. It reacts, however, with hemoglobin to form methemoglobin, resulting in cyanosis and possibly death. The TLV-TWA value is 25 ppm (31 mg/m^3).

Nitrogen Dioxide. Nitrogen dioxide is an irritant gas. Its MAK value is 5 ppm (9 mg/m^3), TLV-TWA 3 ppm (5.6 mg/m^3), TLV-STEL 5 ppm (9.4 mg/m^3). Inhalation of nitrogen dioxide causes pulmonary edema which may result in death (lethal dose 200 ppm). The substance is only slightly water-soluble but highly lipid-

soluble. It therefore penetrates the alveoli where it damages the capillary walls resulting in exudative inflammation. The respiratory tract is obstructed due to formation of foam. Concentrations exceeding 60 – 150 ppm produce coughing and a burning sensation in the chest. Pulmonary edema becomes apparent after 2 – 24 h. The patient suffers respiratory distress and insomnia. Chronic exposure to low doses results in coughing, headache, loss of appetite, and gastrointestinal disorders. Patients should be kept under clinical observation. Inhalation of ammonia from ammonium hydrogen carbonate is recommended.

If the presence of nitrous gas is to be expected a gas mask should be worn (NO filter).

Dinitrogen Monoxide. Dinitrogen monoxide (laughing gas) does not irritate the mucous membranes. It has a powerful analgesic action but is only weakly narcotic. The gas displaces nitrogen from air-filled body cavities (middle ear, sinuses, intestines, brain ventricles) resulting in an increase in pressure. After chronic exposure, polyneuropathy and myelopathy have been observed. TLV-TWA value is 50 ppm (90 mg/m^3).

5. References

1. J. Brunborg, P. B. Holmesland in C. Keleti (ed.): *Nitric Acid and Fertilizer Nitrates,* Marcel Dekker, New York-Basel 1985, pp. 1 – 164.
2. H. Chilton: *Strong Water,* The M.I.T. Press, Cambridge, Mass. 1968.
3. *Ullmann,* 3rd ed., **15,** 3 – 67
4. E. M. Horn, H. Keiser, K. Schoeller, *Monatsh. Chemie* **118** (1987) 1205 – 1218.
5. *Ullmann,* 4th ed., **20,** 305 – 332.
6. S. A. Stern, J. T. Mullhaupt, W. B. Kay: "The Physicochemical Properties of Pure Nitric Acid," *Chem. Rev.* **60** (1960) no. 2, 186 – 207.
7. A. F. Hollemann, E. Wiberg: *Lehrbuch der anorganischen Chemie,* 81st – 90th ed., De Gruyter, Berlin 1976.
8. Y. Fogel et al., *Kinet. Katal.* **5** (1964) 496 – 504.
9. E. J. Nowak, *Chem. Eng. Sci.* **21** (1966) 19 – 27.
10. T. Pignet, L. D. Schmidt, *J. Catal.* **40** (1975) 212 –225.
11. G. Nettesheim, *Chem. Ing. Tech.* **41** (1969) 773 – 775.
12. *Nitrogen* **183** (1990) 27 – 32.
13. S. P. S. Andrew in C. Keleti (ed.): *Nitric Acid and Fertilizer Nitrates,* Marcel Dekker, New York-Basel 1985, pp. 31 – 40.
14. Johnson Matthey, DE 3 042 362, 1988 (D. J. Stephanson, A. E. Heywood, G. L. Selman).
15. Johnson Matthey, DE 3 039 286, 1981 (A. G. Knapton, G. L. Selman).
16. Johnson Matthey, EP 0 275 681, 1988 (J. R. Handley).
17. A. R. McCabe, G. D. W. Smith, *Platinum Met. Rev.* **27** (1983) no. 1, 19 – 25.
18. J. K. Bradley, G. Drake: "Nitric Acid Technology," Paper read before the Fertiliser Society of London, 15th October 1981.
19. R. M. Heck, J. C. Bonacci, W. R. Hatfield, T. H. Hsiung, *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 73 – 79.
20. A. J. Fischer, M. A. Lang: "Engelhard's Role in Advancing Nitric Acid Catalyst Technology," Paper read before The British Sulphur Society, 14th Int. Conference, Caracas, 9th – 12th April 1989. Engelhard Corporation, EP 0 259 966, 1988 (H. C. Lee, R. J. Farrauto, R. W. Hatfield).
21. H. C. Lee, R. J. Farrauto, *Ind. Eng. Chem. Res.* **28** (1989) 1 – 5.
22. A. E. Heywood, *Platinum Met. Rev.* **17** (1973) no. 4, 118 – 129.
23. *Nitrogen* **66** (1970) 40 – 42.
24. Engelhard Corporation, EP 0 077 121, 1983 (R. W. Hatfield et al.).
25. *Nitrogen* **159** (1986) 28 – 32.
26. M. Bodenstein, *Z. Phys. Chem.* **100** (1922) 68 – 123.
27. W. T. Richards, J. A. Reid, *J. Am. Chem. Soc.* **54** (1932) 3014 – 3015.
28. P. J. Hoftyzer, F. J. G. Kwanten in G. Nonhebel (ed.): *Gas Purification Processes for Air Pollution Control,* Butterworths, London 1972.
29. S. P. S. Andrew, D. Hanson, *Chem. Eng. Sci.* **28** (1961) 105 – 114.
30. E. Abel, H. Schmid, *Z. Phys. Chem.* **132** (1928) 55 –63; **134** (1928) 279 – 301.
31. U. Hoffmann, G. Emig, *Chem. Ing. Tech.* **51** (1979) no. 5, 516 – 517.
32. M. M. Wendel, R. L. Pigford, *AJChE J.* **4** (1958) no. 3, 249 – 256.
33. W. Weisweiler, *Chem. Eng. Tech.* **13** (1990) no. 2, 97 –101.
34. M. Baerns, H. Hofmann, A. Renken: *Chemische Reaktionstechnik,* Thieme Verlag, Stuttgart-New York 1987.

35. D. N. Miller, *AIChE J.* **33** (1987) no. 8, 1351 – 1358.
36. J. B. Lefers, P. J. v.d. Berg, *Chem. Eng. J.* **23** (1982) 211 – 221.
37. J. J. Carberry: *Chemical and Catalytic Reaction Engineering,* McGraw-Hill, New York 1976, pp. 289 –293.
38. W. Brötz, F. Schnur: "Chemische Reaktionstechnik," *Chem. React. Eng. Eur. Symp, 1ct 1957,* Amsterdam 1957, p. 1.
39. C. Toniolo, G. Giammarco: *C. R. Congr. Int. Chim. Ind.* **4** (1932) 829 – 859.
40. *Gmelin* **4, "Stickstoff",** 953.
41. L. Hellmer, *Chem. Ing. Tech.* **55** (1972) no. 6, 420 –424.
42. R. W. King, J. C. Fielding, *Trans. Inst. Chem. Engr.* **38** (1960) 71 – 83.
43. M. Koukolik, J. Markek: *Chem. React. Eng. Proc. Eur. Symp, 4th 1986,* Brüssel Sept. 1968, pp. 347 –359.
44. G. Boy, B. Koehret, A. Piquereau: *Routine Programs and Use of Computers in Chemical Engineering,* International Congress, Paris, March 1978, C 18, 81 – 85.
45. V. I. Glushchenko, E. D. Kirichuk, *Int. Chem. Eng.* **22** (1982) no. 1, 181 – 186.
46. H. Hesky, *Chem. Ing. Tech.* **33** (1961) no. 1, 27 – 31.
47. T. H. Chilton, *Chem. Eng. Prog. Monogr. Ser.* **56** (1960) no. 3, 1 – 22.
48. K. Asperger, *Chem. Tech. (Leipzig)* **14** (1962) no. 10, 582 – 589.
49. L. R. Roudier et al., *Inf. Chim.* **88** (1970) 63 – 70.
50. R. M. Counce, J. J. Perona, *Ind. Eng. Chem. Process Des. Dev.* **19** (1980) 426 – 431.
51. H. Holma, J. Sohlo, *Comput. Chem. Eng.* **3** (1979) 135 – 141.
52. K. W. Wiegand, E. Scheibler, M. Thiemann, *Chem. Eng. Tech.* **13** (1990) no. 5, 289 – 297.
53. J. J. Carberry: *Chemical and Catalytic Reaction Engineering,* McGraw-Hill, New York 1976, pp. 286 –293.
54. R. Wetter, *Chem. Tech. (Heidelberg)* **13** (1984) no. 3, 59 – 60.
55. W. Hänggeli: "Steam Washing of Nitrous Gas Compressors," *Techn. Inf. Nr. T8/8,* Sulzer, Zürich.
56. R. Staude, *Chem. Prod.* **4** (1980) 32 – 41.
57. *Kältetechnik* **10** (1958) no. 8; DKV-Arbeitsblatt 2 – 26.
58. J. Eimers in C. Keleti (ed.): *Nitric Acid and Fertilizer Nitrates,* Marcel Dekker, New York-Basel 1985.
59. ASTM A 262-85a Practice C: Nitric Acid Test for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels.
60. G. Berglund, *Chem. Ind. (Düsseldorf)* **34** (1982) 669 – 673.
61. G. H. Honti in C. Keleti (ed.): "Various Processes for the Production of Commercial – Grade Nitric Acid," pp. 61 – 98, and "Various Processes for the Production of Concentrated Nitric Acid," pp. 99 – 130 in *Nitric Acid and Fertilizer Nitrates,* Marcel Dekker, New York-Basel 1985.
62. *Nitrogen* **154** (1985) 28 – 37.
63. *Nitrogen* **129** (1981) 32 – 39.
64. W. Freitag, M. W. Packbier: "Solving Nitric Acid Plant Pollution Problems," *Ammonia Plant Saf.* **20** (1978) 11 – 16.
65. VDI-Richtlinie 2295, VDI-Verlag, Düsseldorf 1983.
66. Umweltbundesamt, *Statistisches Jahrbuch,* StBA, Berlin 1988, Tab. 25.
67. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft- TA Luft) 1986.
68. *VDI-Handbuch Reinhaltung der Luft,* VDI-Verlag, Düsseldorf 1974.
69. VDI-Richtlinie 2310, Blatt 12, VDI-Verlag, Düsseldorf 1985.
70. VDI-Richtlinie 2066, Blatt 1, VDI-Verlag, Düsseldorf 1975.
71. VDI-Richtlinie 2456, VDI-Verlag Düsseldorf: a) Blatt 1, 1973; b) Blatt 2, 1973; c) Blatt 3, 1975; d) Blatt 4, 1976; e) Blatt 5, 1978; f) Blatt 6, 1978; g) Blatt 7, 1981; h) Blatt 8, 1986; i) Blatt 9, 1989; j) Blatt 10 (Entwurf) 1989.
72. G. A. Heymann, G. S. Turner, *ISA Trans.* **3** (1976) 209 – 213.
73. "Additional Absorption Capacity Stems Nitric Acid Plant NO_x Pollution," *Nitrogen* **91** (1974) 36.
74. Österreichische Stickstoffwerke, US 3 034 853, 1962 (A. Schmidt, F. Weinrotter).
75. "Oxidative Tailgas Scrubbing Technique," *Nitrogen* **94** (1975) 471.
76. Office National Undustriel de Lazote, US 3 329 478, 1967 (R. Garlet).
77. Potasse et Engrais Chimiques, US 3 348 914, 1967 (B. Quanquin, H. Trimbach).
78. "Pollution Control," *Nitrogen* **95** (1975) 44.
79. J. C. Adrian, J. Verilhac: "A Process for Reduction of NO_x Content in Flue Gas," *2nd International Conference on Control of Gaseous Sulphur and Nitrogen Compound Emissions,* University Salford, April 1976.

80. "Urea as a Pollution Control Agent," *Nitrogen* **95** (1975) 32.
81. B. J. Buck, W. G. Matthews, *Proc Environ. Symp.* 1976, 157 – 168.
82. Kernforschungsanlage Jülich, DE 3 226 840 A1, 1984 (H. Ringel).
83. L. Rodier, N. Durand, B. Mouille: "Nitric Acid Recovery from Waste Nitrous Gases by the COFAZ Process," *Fert. Nitrogen: Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol. 4th 1981,* **2** (1982) 561 – 569.
84. *Nitrogen* **171** (1988) 25 – 34.
85. H. Dittmar: "Katalytische Reduktion von nitrosen Gasen bei der Salpetersäurefabrikation," *VDI-Bericht 525,* VDI-Verlag, Düsseldorf 1985.
86. W. Weisweiler, B. Retzlaff, B. Hochstein, *Staub Reinhalt. Luft* **48** (1988) 119 – 126.
87. W. Weisweiler, B. Hochstein, *Staub Reinhalt. Luft* **49** (1988) 37 – 43.
88. A. More: "The Mitsubishi Process," *Fert. Acids Proc. Br. Sulphur Corp. Int. Conf. Fert. 3rd* (1979) Paper VI, p. 536.
89. M. Yamaguchi, K. Matsushita, K. Takami: "New Catalytic Process for Removing Nitrogen Oxides from Nitric Acid Tail Gas," *Fert. Acids Proc. Br. Sulphur Corp. Int. Conf. Fert. 3rd* **1** (1979) no. 6, 20.
90. R. Erath: "Das Bergbau-Forschung/UHDE-Verfahren," *Brennstoff-Wärme-Kraft* (1985) no. 9; *Staub Reinhalt. Luft* (1985) no. 9; *Umwelt* (1985) no. 4.
91. "Gefahrgutverordnung Eisenbahn GGVE," *Bundesgesetzblatt I,* 23.8.1979, p. 1502.
92. "Règlement international concernant le transport des marchandises dangereuses par chemins de fer –RID," *Bundesgesetzblatt II,* 26.10.1978, p. 1285.
93. "Gefahrgutverordnung Straße – GGVS," *Bundesgesetzblatt I,* 23.8.1979, p. 1509.
94. "Gesetz zu dem europäischen übereinkommen über die internationale Beförderung gefährlicher Güter auf der Straße (ADR)," ADR Rn. 2801, *Bundesgesetzblatt II,* 1969, p. 1489.
95. "Verordnung über die Beförderung gefährlicher Güter auf dem Rhein (ADNR)," ADNR. Rn. 6500/6501.
96. *Industrial Statistic Yearbook,* vol. 2, Commodity Production Statistics, United Nations, 1989.
97. J. S. Lundsgaard et al., *Acta Anaesthesiol. Scand.* **21** (1977) no. 4, 308 – 313.
98. *Ullmann,* 3rd ed., **15,** 44.
99. Hoechst AG, DE 1 921 1815 1970 (H. T. Baechle, R. Kohlhaas).
100. N. M. Levenson: "Nitrous Oxide – U.S. Salient Statistics" in *Chemical Economic Handbook,* SRI International, 1983, pp. 7436000A–7436000E.
101. L. Pauling: *The Nature of the Chemical Bond and the Structure of Molecules and Crystals,* Verlag Chemie, Weinheim 1976.
102. F. H. Verhoek, F. Daniels, *J. Am. Chem. Soc.* **53** (1931) 1254 – 1259.
103. C. C. Addison: *Chemie der nichtwäßrigen ionischen Lösungmechanismen,* Vieweg Verlag, Braunschweig 1967, pp. 1 – 78.
104. A. H. Brittain et al., *Trans. Faraday Soc.* **65** (1969) 1963.
105. I. R. Beattie, *Prog. Inorg. Chem.* **5** (1963) 1 – 26.
106. I. C. Tracy, F. Daniels, *J. Am. Chem. Soc.* **77** (1955) 2033.
107. R. A. Ogg, *J. Chem. Phys.* **15** (1947) 337 – 338.
108. G. Sorbe: *Sicherheitstechnische Kenndaten chemischer Stoffe,* 15. Erg.-Lfg, ecomed Verlagsgesellschaft, Landsberg 1989.
109. "Verordnung über gefährliche Stoffe (Gefahrstoffverordnung – GefStoffV)", *Bundesgesetzblatt I,* 1986, p. 1470; *Bundesgesetzblatt I,* 1987, p. 2721.
110. *GefStoffV: Gefahrstoffverordnung,* 3rd ed., ecomed Verlagsgesellschaft, Landsberg 1988. *Erste Verordnung zur änderung der Gefahrstoffverordnung vom 16. Dezember 1987,* Deutscher Bundes-Verlag GmbH, 5300 Bonn 1.

Nitrides

PETER ETTMAYER, Institut für Chemische Technologie anorganischer Stoffe, Technische Universität Wien, Vienna, Austria

WALTER LENGAUER, Institut für Chemische Technologie anorganischer Stoffe, Technische Universität Wien, Vienna, Austria

1. Introduction 1
2. Ionic Nitrides 3
2.1. Lithium Nitride 3
2.2. Nitrides of Alkaline-Earth Metals . 4
2.3. Rare-Earth Nitrides, Earth Nitrides, and Actinide Nitrides 4
3. Covalent Nitrides 4
3.1. Boron Nitride 4
3.2. Aluminum Nitride 7
3.3. Gallium Nitride 8
3.4. Silicon Nitride 9
3.5. Miscellaneous Covalent Nitrides . . 12
4. Transition-Metal Nitrides 12
4.1. Production 12
4.2. General Properties 13
4.3. Uses 14
4.3.1. Wear-Resistant Layers 14
4.3.2. Cermets 16
4.3.3. Other Uses 17
4.4. Multicomponent Nitrides 18
5. References 18

1. Introduction

Nitrides occur rarely in nature. Titanium nitride, TiN, has been detected in silicate meteorites as the mineral osbornite [1,2]. Titanium nitride and titanium carbonitride, Ti (C, N), sometimes occur in the form of golden-yellow crystals in pig iron (blast furnace cubes). Titanium carbonitride was first described by WOLLASTON in 1822 and was identified by chemical analysis as Ti $(C_{0.2}N_{0.8})$ in 1850 by WÖHLER [3].

At the beginning of the 20th century transition-metal nitrides attracted particular interest because of their special properties such as high degree of hardness, high melting point, metallic luster, and superconductivity [4,5]. However, the first industrial application of a nitride was the use of calcium cyanamide as a fertilizer. Calcium cyanamide is produced by the reaction of nitrogen with calcium carbide; on contact with moisture in the soil it gradually releases nitrogen in the form of ammonia with the formation of several intermediates.

Classification. Nitrides can be classified according to their bonding character. Four main groups can be distinguished [6]:

1) The *ionic or saltlike nitrides* of the alkali metals, alkaline-earth metals, and of metals belonging to group 3 of the periodic table. These nitrides react with hydrogen ions to form ammonium ions.
2) The *covalent nitrides* are formed with elements of groups 13 – 15 (e.g., BN, Si_3N_4). These compounds are usually very hard; they are nonconductors or semiconductors.
3) The *transition-metal nitrides* are classified as interstitial nitrides; the bonding character is primarily metallic.
4) The *molecule-forming nitrides* of the main group elements 15 – 17 are often volatile and have low melting points. They can be explosive and decompose easily.

Synthesis. Chemical reactions most frequently used to prepare nitrides include:

1) The nitriding of metals (in powder or compact form) or metal hydrides with molecular or atomic (activated) nitrogen or in a stream of ammonia, possibly in the presence of a liquid metal (auxiliary metal technique) or at high pressure;
2) The reaction of chlorides or organometallic compounds with a nitrogen – hydrogen mixture or ammonia;
3) The reaction of metal oxides with carbon, combined with simultaneous nitriding in the presence of nitrogen or ammonia in the presence of a reducing agent such as carbon; and

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_341

4) The reaction of metal atoms, metal ions, or atomic clusters with nitrogen atoms or nitrogen ions in the gas phase.

Since nitrogen is a noncondensed phase, nitride synthesis can be controlled by the nitrogen partial pressure. A specific nitrogen equilibrium pressure, which depends on the temperature and the composition of the nitride, can be assigned to each nitride phase. The higher the stability of a nitride phase, the lower is its nitrogen equilibrium pressure. The dependence of stability [7] (free energy of formation) on temperature is shown in Figure 1.

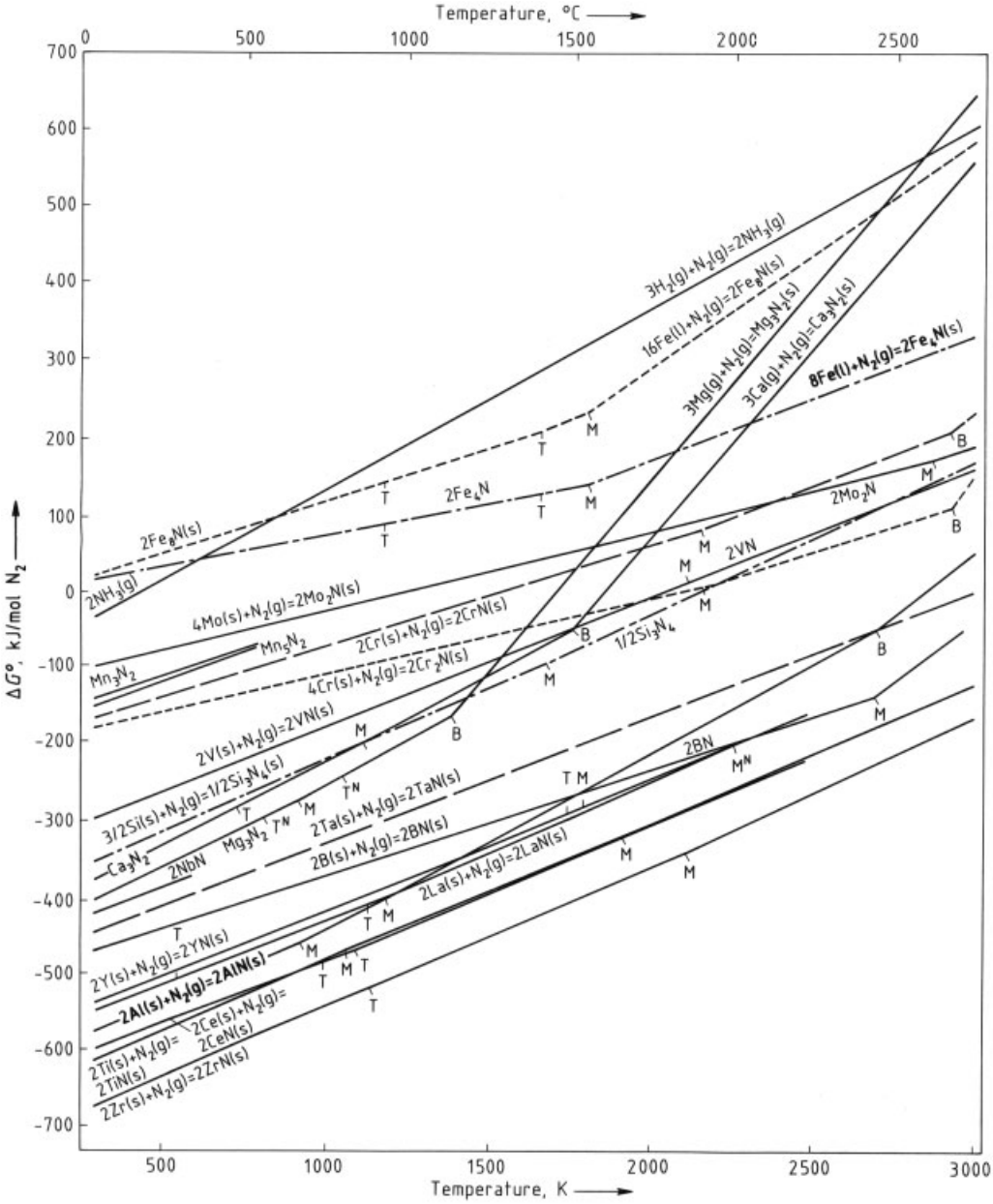

Figure 1. Free energy of formation of various nitrides as a function of temperature (Richardson diagram) [7]
B = *bp* of metal; M = *mp* of metal; M^N = *mp* of nitride; T = transition point of metal; T^N = transition point of nitride

2. Ionic Nitrides

2.1. Lithium Nitride

The alkali metals (M) react with nitrogen to form compounds of the formula M_3N. The compound with the highest stability, and technically speaking the most interesting, is α-lithium nitride [*26134-62-3*], Li_3N [8]. This compound is a fast ionic conductor.

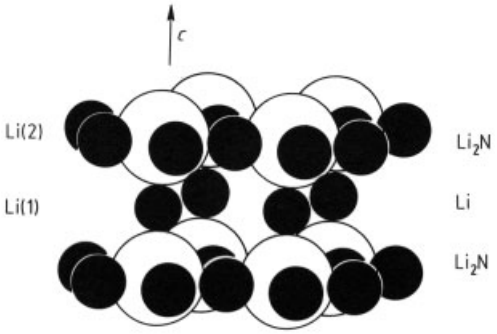

Figure 2. Crystal structure of lithium nitride [ionic conductivity can be attributed to the mobility of the Li (2)$^+$ ions]

Lithium nitride can be obtained by reacting lithium metal with nitrogen at >800 K. The structure of α-Li_3N (Fig. 2) is characterized by hexagonal Li_2N layers, perpendicular to the *c* axis. A lithium defect concentration of 1 – 2 % is responsible for the high mobility of Li^+ ions Li (2) in these Li_2N layers [9]; the positions of the bridging lithium atoms Li (1) between the layers are fully occupied and do not contribute to ionic conductivity. Defect formation is due to the incorporation of hydrogen and NH^{2-} groups [10]. Because of compositional differences, it is very difficult to give precise values for the ionic conductivity of lithium nitride. At 300 K, conductivity values of 1.2×10^{-3} $\Omega^{-1}cm^{-1}$ have been measured perpendicular to the *c* axis for single crystals, and 6.6×10^{-4} $\Omega^{-1}cm^{-1}$ for polycrystalline materials [11]. Important properties of lithium nitride are listed in Table 1. Compact lithium nitride can be handled easily in air but, when finely dispersed, is inflammable.

α-Lithium nitride melts congruently at 1086 K. Ruby-red single crystals can be formed by reacting molten lithium with nitrogen at ca. 0.1 MPa and 800 K, according to the Czochralski method [12]. Single crystals with a volume of ca. 10 cm^3 can be produced in this way for solid-state studies [12, 13].

Simple sintering processes can be employed to produce dense material that is used as a solid-state electrolyte in lithium batteries (Fig. 3) [14]. Lithium serves as the anode; it is separated from the cathode by lithium nitride. The cathode can be formed from a sulfide phase ($TiLi_xS_2$ or $TaLi_xS_2$) that has a high absorbing capacity for Li^+ ions. When the battery is discharged, Li^+ ions migrate from the lithium metal via the solid electrolyte lithium nitride into the sulfide phase.

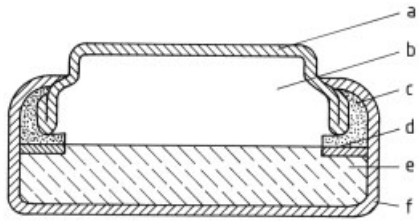

Figure 3. Schematic structure of a lithium battery with a Li_3N diaphragm [11]
a) Cap; b) Anode (Li); c) Isolation; d) Electrolyte (Li_3N); e) Cathode; f) Package

Lithium nitride is also used as a catalyst for the conversion of hexagonal boron nitride (α-BN) to the cubic high-pressure modification (β-BN) [15, 16].

Table 1. Properties of lithium nitride and important covalent nitrides

Property	Li_3N	α-BN	β-BN	AlN	GaN	α-Si_3N_4	β-Si_3N_4
Structure (Pearson)	hP4	hP4	cF8	hP4	hP4	hP28	hP14
Cell dimensions, nm *a*	0.3648	0.2504	0.36155	0.3110	0.3190	0.7818	0.7595
c	0.3875	0.6661		0.4980	0.5189	0.5591	0.2902
Density, g/cm^3 [a]	1.29	2.28	3.49	3.26	6.08	3.15	3.21
mp, K	1086	ca. 3200 [b]	ca. 1700 [c]	ca. 2300 [b]	900 [b]	ca. 1400 [c]	2170 [b]

[a] X-ray density.
[b] Decomposes, *mp* is dependent on the nitrogen pressure.
[c] Transforms.

2.2. Nitrides of Alkaline-Earth Metals

Alkaline-earth metals react with nitrogen to form nitrides of the composition M_3N_2. They are used as sintering additives for special ceramics and as catalysts for the production of cubic boron nitride [17–19]. Beryllium nitride [*1304-54-7*], Be_3N_2 [20, 21], strontium nitride [*12033-82-8*], Sr_3N_2, and barium nitride [*12047-79-9*], Ba_3N_2, have been used as sintering additives for silicon nitride (Si_3N_4) and aluminum nitride (AlN) in isolated cases; however, the addition of magnesium nitride [*12057-71-5*], Mg_3N_2 [22–24], and calcium nitride [*12013-82-0*], Ca_3N_2, to aluminum nitride [*24304-00-5*] [25–27] has gained some importance. Undesirable amorphous grain boundary phases can be almost entirely prevented in Si_3N_4 ceramics by the use of Be_3N_2 or $BeSiN_2$. These compounds are, however, toxic [28]. Calcium nitride [29] and magnesium nitride [30] have been proposed as additives for electrical, copper-based contact materials, thus replacing Ag – CuO and Cu – TiN composites. These materials are used in switches and have good contact welding properties and a high abrasion resistance.

2.3. Rare-Earth Nitrides, Earth Nitrides, and Actinide Nitrides

Nitrides of the rare-earth metals and yttrium display ionic bonding; however, scandium nitride [*25764-12-9*], ScN, has properties that lie between those of the ionic nitrides and the transition-metal nitrides. Scandium nitride [31] is the only earth-metal nitride that is relatively stable under humid atmospheric conditions [32]. It has semiconductor properties [33] and has been proposed for use in solar cells [34].

Other solid-state properties of these nitrides–such as the magnetism of gadolinium nitride [*25764-15-2*], GdN [35]–have not yet been exploited technically. Some rare-earth nitrides such as cerium nitride [*25764-08-3*], CeN, have been used as sintering additives for aluminum nitride and silicon nitride [36]. The actinide mononitrides–actinium nitride, uranium nitride, plutonium nitride, and neptunium nitride–crystallize in the cubic NaCl-type lattice like the lanthanide nitrides. In contrast to rare-earth nitrides, actinides also form phases with a higher nitrogen content (e.g., U_2N_3 [*12033-85-1*] [37], Th_3N_4 [*12033-90-8*] [38]). Phases with even higher nitrogen concentrations (UN_{2-x} [*12033-83-9*] can be synthesized with pressurized nitrogen [39]. The actinide mononitrides have been proposed as nuclear fuels in high-temperature breeder reactors.

3. Covalent Nitrides

For poly(sulfur nitride), see → Inorganic Polymers, Chap. 6.

3.1. Boron Nitride

Boron nitride [*10043-11-5*], BN, crystallizes in three modifications (α, β, and γ) with a stoichiometric ratio of boron to nitrogen of 1 : 1 (→ Boron Carbide, Boron Nitride, and Metal Borides, Chap. 2.) [40–42].

1) *White, hexagonal boron nitride* (α-BN, Fig. 4 A) has a crystal structure that is related to the graphite lattice (Fig. 4 B). The stacking order of the hexagonal layers is, however, different.
2) *Cubic boron nitride* (CBN, β-BN) is a high-pressure modification. It can be produced from α-boron nitride under conditions used for diamond synthesis by employing high pressure (4 – 6 GPa and 1400 – 1700 °C) in the presence of metal or metal nitride catalysts (see Section 2.2) [43]. Pure β-boron nitride is crystal clear. At room temperature it has a high thermal conductivity but no electrical conductivity. With a Knoop hardness of ca. 4700, β-boron nitride is the second hardest solid after diamond (Knoop hardness 7000). Although β-boron nitride is metastable at atmospheric pressure, it is thermally more stable than diamond and can be used for short periods at temperatures up to 1700 K. In air, β-boron nitride is less rapidly oxidized than diamond because a boron oxide (B_2O_3) layer is formed that protects the substance from further oxidation. The structure of β-boron nitride (zinc blende structure) is shown in Figure 4 C.
3) *Hexagonal boron nitride* (γ-BN) crystallizes in the wurtzite structure. It is also a high-pressure modification.

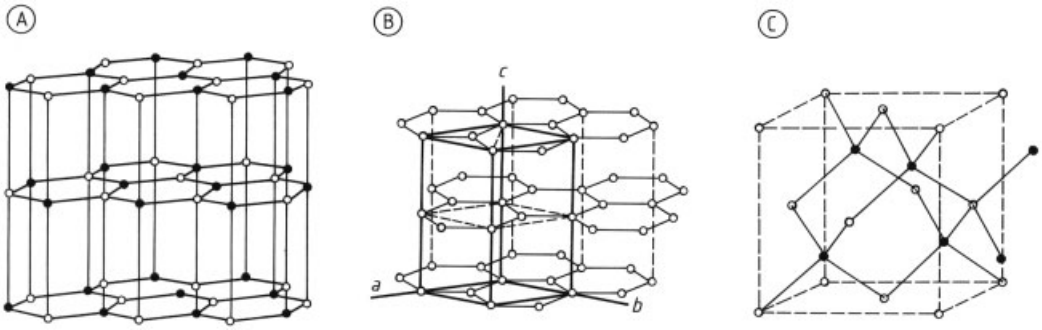

Figure 4. Crystal structure of hexagonal α-boron nitride (A), graphite (B), and β-boron nitride (C)

The phase diagram of the boron nitride system is shown in Figure 5 [44]. Properties of α- and β-boron nitride are listed in Tables 1 and 2.

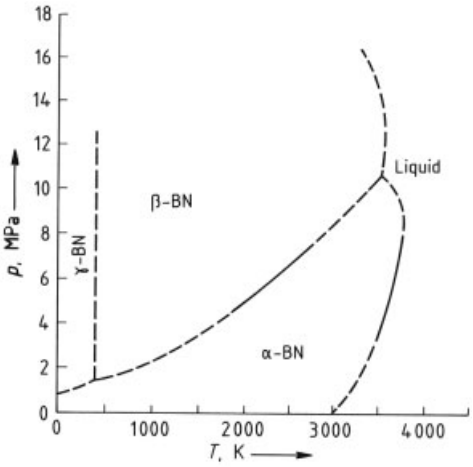

Figure 5. Phase diagram of boron nitride [44]

β-Boron nitride was first synthesized in 1957 and introduced commercially in 1969 [45,46]. It is used mainly as an abrasive grit in grinding wheels and saws (→Abrasives, Chap. 3.7.). Apart from its extreme hardness, the high thermal conductivity of β-boron nitride is also important because it increases grinding depth and speed [42]. Diamond reacts with carbide-forming elements such as iron, chromium, tantalum, and tungsten at > ca. 1200 K and is thus not suitable for grinding or machining steel; β-boron nitride is much more stable toward these elements.

Both β- and γ-boron nitride can be made into dense, polycrystalline materials (polycrystalline cubic boron nitride, PCBN) by applying high pressure and using sintering additives such as aluminum or silicon (similar to diamond). Polycrystalline cubic boron nitride compacts are somewhat softer than single crystals but have a higher fracture toughness and less orientational anisotropy. Polycrystalline β-boron nitride can be made into cutting tools with a defined cutting edge. These tools are used to work quenched and tempered steels that could previously be worked only by grinding [47].

α-Boron Nitride Layers. Various chemical vapor deposition (CVD) processes have been

Table 2. Properties of nitride ceramics.

Property	α-BN (HIP)∗	β-BN	AlN	Si_3N_4(SSN)
Hardness (Vickers) (ca. 300 K) ∗, GPa	0.3 ∗∗	35	12 – 13	16 – 18
Thermal conductivity (ca. 300 K), $W\,m^{-1}\,K^{-1}$	50	100	100 – 170	30 – 40
Coefficient of thermal expansion, $10^{-6}\,K^{-1}$	4.1	4.9	3.5 (300 – 600 K)	3.2 (0 – 1600 K)
Electrical resistivity (ca. 300 K), $\Omega \cdot cm$	$> 10^9$	$> 10^{10}$	$> 10^{11}$	$> 10^{12}$
Relative permittivity	6.7		8.8	7.9
Bending strength (ca. 300 K), MPa	50 – 60	500 – 800	250 – 350	400 – 750
Modulus of elasticity (ca. 300 K), MPa	32	680	310	250 – 310

∗ HIP = hot isostatic pressing.
∗∗ Knoop hardness.

proposed for the production of α-boron nitride layers [48]. The starting materials are mainly volatile boron compounds (e.g., as B_2H_6 [49–51], BCl_3 [52–58], and BF_3 [59]) as well as higher boranes and organoboron compounds. These compounds are reacted with ammonia [60, 61]. Depending on reaction temperature (ca. 700 – 1500 K) and deposition conditions, the properties of the resulting α-boron nitride films can vary considerably, especially in terms of crystalline state, structure, stoichiometry, thermal conductivity, electrical resistivity, and size of the optical band gap [48]. Thus, colorless transparent, amorphous, or crystalline boron nitride films can be produced with band gaps of 3.8 – 6.2 eV and electrical resistivities ranging from 10^9 to 10^{17} $\Omega \cdot$ cm.

The chemical reaction can be assisted by use of a microwave plasma at a moderate temperature (500 – 1200 K). Amorphous boron nitride films can usually be produced with a wide range of physical properties. Physical vapor deposition (PVD) processes, including sputtering [62–64], electron-beam evaporation [65], ion-beam deposition [66–70], and the pulsed plasma method [71–73], have also been used to produce boron nitride films. The range of physical properties of these films is much greater than that of CVD films. Significant stoichiometric variations are possible, allowing production of nontransparent boron nitride layers with a boron : nitrogen ratio up to 2.5 : 1 [63, 74].

For a detailed description of both CVD and PVD, see → Thin Films.

α-Boron nitride films are important primarily in the microelectronics industry. They can be used as a source of boron for doping silicon wafers in diode production (see Fig. 6) [48]. *n*-Silicon is thermally oxidized, then locally etched (A), and coated with a boron nitride film (B). Heating in a nitrogen atmosphere causes the boron to diffuse into the *n*-silicon, which thus becomes locally *p*-conductive (C).

Electrically insulating α-boron nitride layers are also used in metal insulator semiconductor diodes [51]; as thin, transparent substrates in X-ray lithography [75–77]; as coatings to improve oxidation resistance [78] and surface hardness [67–70, 74] (see below); for electrical insulation; and for surface modification of field effect transistors as sensors for hydrogen detection [79].

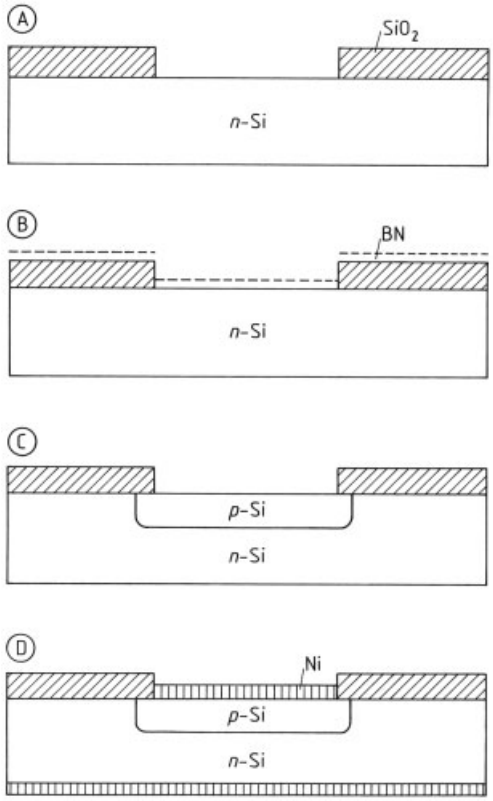

Figure 6. Production of boron nitride films for integrated circuits [48]
A) Thermal oxidation of silicon; B) Coating with boron nitride film; C) Heating in a nitrogen atmosphere with diffusion of boron and decomposition of boron nitride; D) Nickel plating

Extremely Hard Boron Nitride Layers. Hard boron nitride layers have been produced by ion-beam deposition [66–70]. Such layers, however, cannot be designated as β-boron nitride because, in addition to tetrahedrally coordinated boron and nitrogen atoms, they also contain a number of sp^2 hybrid bonds. In analogy with the ion-beam deposition of carbon, these layers have been termed i-boron nitride [69]. Vickers hardnesses up to 30 GPa were obtained.

Pressureless deposition of β-boron nitride has been achieved but it is more difficult than that of diamond [80–82]. Activated hydrogen seems to be essential for the production of β-boron nitride from B_2H_6 and ammonia and for the production of i-boron nitride films [83]. Reaction of activated hydrogen with $(BH_2NH_2)_3$ initially leads to the deposition of β-boron nitride and amorphous boron nitride; however,

the latter is preferentially removed by the ion sputtering effect and β-boron nitride is left behind [83]. Methods used to produce β-boron nitride layers include triode ion plating: a combination of tungsten filament activation plus radiofrequency plasma CVD [80, 83–87]. Results vary significantly, however; in particular, the boron : nitrogen ratio often differs markedly from 1 : 1 [88].

β-Boron nitride layers should be used particularly when both a high degree of hardness and a high thermal conductivity are required. Applications include cutting tools for machining and milling hard alloys (particularly quench-hardened ferrous materials), for nozzles, etc. In contrast to diamond, β-boron nitride can also be used above 1100 K. In microelectronic circuits, β-boron nitride can be employed as a heat sink.

3.2. Aluminum Nitride

Production. Industrial production of aluminum nitride [*24304-00-5*], AlN, is carried out primarily in three ways [89, 90]:

1) Direct nitriding of aluminum powder at 1100 K:

$$2\,Al + N_2 \longrightarrow 2\,AlN$$

an exothermic reaction.

2) The self-propagating high-temperature synthesis (SHS) method [91] in which the heat produced during the reaction generates a self-propagating reaction front. The reaction is started by ignition at one end of the reaction and proceeds through the sample.
3) Carbothermic nitriding by endothermic reduction of Al_2O_3:

$$Al_2O_3 + N_2 + 3\,C \longrightarrow 2\,AlN + 3\,CO$$

In direct nitriding, the product must be ground to a defined grain size (this may adversely affect sintering due to the introduction of impurities). In carbothermic nitriding, grain growth is largely avoided; excess carbon must, however, be removed in a second reaction step via the formation of carbon monoxide by reaction of carbon with mildly oxidizing gases such as water vapor or carbon dioxide.

A less important production method is the reaction of aluminum chloride or organoaluminum compounds with ammonia [89] or mixtures of nitrogen and hydrogen (R = Cl or organic group):

$$R_3Al + NH_3 \longrightarrow AlN + 3\,RH$$

This reaction is used mainly in the production of aluminum nitride layers by CVD.

Properties. Aluminum nitride , the only compound known in the Al – N system, crystallizes with a wurtzite-type structure; it is nontoxic. In powder form, aluminum nitride is pale gray (due to the presence of nonreacted aluminum). In sintered or hot-pressed form it is colorless and transparent. Aluminum nitride has a melting point of ca. 3030 K at a nitrogen pressure of 5 MPa [92]. At atmospheric pressure, it evaporates quickly above 2000 K. Hot mineral acids hydrolyze aluminum nitride powder and attack compact aluminum nitride slightly. The compound is completely stable in air up to ca. 1000 K, but above this temperature it reacts with air and oxygen. Dense aluminum nitride forms a protective, amorphous aluminum oxide layer that recrystallizes at higher temperature and protects the nitride from oxidation up to ca. 1800 K [93].

Aluminum nitride has a favorable combination of physical properties that make it particularly suitable for electronic devices (Tables 1 and 2): unusually high thermal conductivity [94] combined with high electrical resistance and a thermal expansion coefficient similar to that of silicon (to avoid thermal stress) [95]. Because of the continuing miniaturization of electronic components, efficient aluminum nitride heat sinks are necessary as substrates (packages) for heat dissipation; conventional aluminum oxide ceramics can no longer meet the requirements. In addition, chip capacities are continually increasing, which necessitates the appropriate choice of materials with thermal expansion coefficients that closely correspond to that of silicon.

Production of Aluminum Nitride Ceramics. Depending on the particle size of the aluminum nitride powder, sintering without pressure produces a material with a porosity of ca. 15 – 30 % [96–98]. Complete compaction can

be achieved only with pressure sintering processes such as hot axial pressing or hot isostatic pressing, as well as liquid-phase sintering with additives (mainly CaO, Y_2O_3, CeO_2, and ZrO_2) [25, 99, 100]. With aluminum nitride, these additives form lower-melting phases at the grain boundaries that accelerate pore removal and compaction and allow full densification exceeding 99 %. With calcium oxide, the aluminates $CaO \cdot 2\,Al_2O_3$ and $3\,CaO \cdot 5\,Al_2O_3$ can be detected at the grain boundaries of aluminum nitride. However, an oxinitride phase also seems to be formed as an intermediate in the sintering process. Magnesium oxide is unsuitable for the dense sintering of aluminum nitride [92]. The additives usually produce slight changes in its desired properties. If these cannot be tolerated, pressure sintering methods without additives must be used (see Section 3.4). Completely dense aluminum nitride can be obtained at 2100 K under a static axial pressure of 20 MPa in 30 min [92]; however, oxygen present in the starting powder reacts to form an Al – O – N phase with a spinel structure [101]. The aluminum oxinitride has interesting isotropic physical properties (transparency) as well as good chemical stability [101–103]. Aluminum nitride components can be used in air up to 1700 K. They are not wetted by metal melts. Sheet casting is a technologically important production method for aluminum nitride ceramics [104]. It is particularly suitable for the production of thin sintered parts with complicated pierced shapes and for the production of multilayer ceramic composites. In the "doctor blade" technique, aluminum nitride powder is suspended together with plasticizers in an organic solvent (e.g., gasoline, a ketone, toluene, or a mixture of these), drawn to a sheet, and cut to the desired length at a speed of ca. 30 cm/s.

Uses. The following properties of aluminum nitride are important for the electronics industry: high thermal conductivity, a thermal expansion coefficient similar to that of silicon, and high strength (Table 2). These properties are necessary for the development and use of efficient microelectronic circuits such as very large-scale integrated circuits (VLSI). Aluminum nitride substrates (also known as packages) are also used as coolers for power transistors, integrated circuits, laser diodes, thyristors, and ceramic dual inline packages (Fig. 7) [105].

The fact that aluminum nitride is not wetted by metals is taken advantage of in the melting and casting of dental alloys. The crucible surface is coated with aluminum nitride to prevent the diffusion of impurities from crucible materials into the alloys. The high modulus of elasticity of fully dense aluminum nitride combined with its low density make it a suitable material for lightweight personal armor (bullet-proof vests).

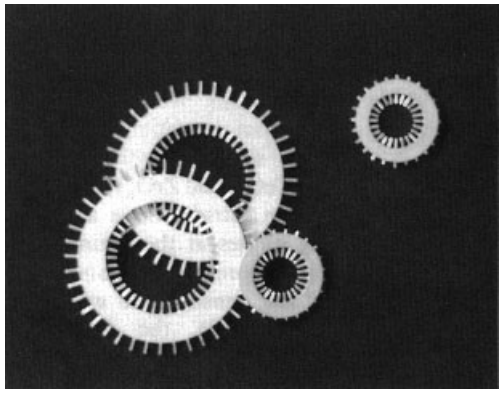

Figure 7. Aluminum nitride substrate ceramics with gold pins for silicon sensors (ANCERAM, Bindlach, FRG)

3.3. Gallium Nitride

Production and Properties. Gallium nitride [*25617-97-4*], GaN, can be produced by heating gallium in a stream of ammonia at 900 – 1200 K [106]:

$$2\,Ga + 2\,NH_3 \longrightarrow 2\,GaN + 3\,H_2$$

Heating in molecular nitrogen at atmospheric pressure fails to give the desired product; very high pressure would have to be applied because of the high nitrogen equilibrium pressure of gallium nitride [107, 108]. Hence, the production of compact samples and single crystals (important for solid-state investigations, particularly electronics) is very difficult. The nitriding of gallium and gallium nitride with ammonia yields crystals with dimensions of ca. 5×0.3 mm and 5×1 mm, respectively [109, 110]. Thin gallium nitride layers have been produced by CVD [111–113], liquid-phase epitaxy [114], and molecular-beam

epitaxy. Chemical vapor deposition is based on the following reaction [111–113, 115–118]:

$$GaCl_3 + NH_3 \longrightarrow GaN + 3\,HCl$$

Satisfactory crystal nucleating substrates on which gallium nitride can grow epitaxially have not yet been found [119]. Either the crystal lattice fit is not good enough (e.g., Al_2O_3, Si) or the chemical resistance is unsatisfactory (e.g., HCl attack in the case of ZnO).

Gallium nitride crystallizes in the wurtzite-type structure (a cubic modification has also been described [120]) and in its compact form is colorless; it can, however, assume orange to yellow shades depending on the impurities or doping agents (e.g., Mg, Zn) [111]. Gallium nitride is almost completely stable to acids but dissolves in hot alkaline solutions. It is stable to oxidation in air up to ca. 1100 K.

Gallium nitride is a semiconductor with a band gap of 3.48 eV at 4.2 K [121]. However, its electronic properties such as carrier concentration and electrical resistance are highly dependent on the method of preparation [122–124].

Uses. Gallium nitride has been used commercially as a light-emitting diode (LED) only since 1985, although its suitability for this purpose was discovered in the early 1970s [123, 125–127]. Gallium nitride LEDs are used in medicine; they emit blue light, which is used to induce fluorescence in blood pigments for analytical purposes.

Other potential uses of gallium nitride (e.g., in microwave amplifiers and as a material for switching contacts [128]) have not been exploited commercially. Applications of gallium nitride are likely to increase in the near future, particularly if pure gallium nitride materials with defined properties can be produced.

3.4. Silicon Nitride

During the last twenty years, silicon nitride [*12033-89-5*], Si_3N_4, has become an extremely important material for structural ceramics owing to its excellent high-temperature properties, its stability to oxidation, its high fracture toughness, and its low thermal expansion coefficient (Table 2) [129–135]; see also → Ceramics, Advanced Structural Products, Chap. 3.3. Heat engines with silicon nitride parts show improved efficiency because a higher operating temperature can be employed.

Structure. Silicon nitride crystallizes in two modifications (α- and β-Si_3N_4) that differ with regard to the stacking order of the SiN_4 tetrahedral layers [136]. The β-form has a phenacite-type structure (Be_2SiO_4) and the stacking order AB, AB, etc. In the α-form the AB order is followed by a mirror image, which results in the stacking pattern ABCD, ABCD, etc., and an elementary cell twice as large as that of β-silicon nitride. Both modifications can exchange silicon and nitrogen atoms for other atoms without structural changes.

α-Silicon nitride is usually formed at < 1700 K; above this temperature, it is irreversibly converted to the β-form [137]. α-Silicon nitride is obtained mainly in the form of fibers and whiskers during the nitriding of silicon at lower temperature. β-Silicon nitride is formed primarily by reaction of nitrogen with pure silicon or silicon alloys at high temperature [138].

α-Silicon nitride forms mixed phases by the partial substitution of Si^{4+} with Al^{3+} and simultaneous incorporation of metal cations (e.g., Li^+, Ca^{2+}, Mg^{2+}, and rare-earth metals [139, 140] or by the partial replacement of N^{3-} with O^{2-}. The phases that are isotypic to α-silicon nitride (α'-phases) can be defined by the following formula:

$$M_xSi_{12-(m+n)}Al_{(m-n)}O_nN_{16-n}$$

where $x < 2$; purely nitridic phases (without oxygen) are also possible [141].

In β-silicon nitride, Si^{4+} can be replaced by Al^{3+} (or Y^{3+} and Be^{2+}); to ensure charge compensation, N^{3-} can simultaneously be replaced by O^{2-}. The substitution of aluminum for silicon occurs fairly easily because of the close coordination symmetry of the two elements [142] and yields the sialons (SiAlONs) [142–146] with the general formula:

$$Si_{6-x}Al_xO_xN_{8-x}$$

$x < 2$ at 1670 K and 4.2 at 1970 K [142].

The physical properties of sialons resemble those of β-silicon nitride, but the chemical properties resemble those of corundum. The thermal expansion and thermal conductivity of sialons are lower than those of β-silicon nitride; their stability to oxidation and high-temperature strength are greater. Sialons can be sintered without pressure [147]; at 2000 – 2100 K they react with sintering additives such as yttrium oxide (Y_2O_3) or magnesium oxide (MgO) to form liquid phases that are less viscous than silicon nitride. Thus, sintering to full density is easier than with silicon nitride ceramics.

The production methods and applications for sialon parts are similar to those of silicon nitride. Sialons were the first nitride ceramics to be used commercially [142]. High-temperature ball bearings, indexable inserts (for the machining of cast-iron and nickel-based superalloys), and high-temperature dies are made from sialons. The applications of sialons are expected to increase substantially.

Production of Silicon Nitride Powder. Silicon nitride is a white powder that can be produced in four ways [90]:

1) Nitriding of silicon powder [148, 149]:

$$3\,Si + 2\,N_2 \longrightarrow Si_3N_4$$

2) Chemical vapor deposition of silicon chloride or chlorosilanes in the presence of ammonia or an $H_2 - N_2$ mixture [90, 149–154]:

$$3\,SiCl_4 + 4\,NH_3 \longrightarrow Si_3N_4 + 12\,HCl$$

3) Carbothermic reduction of silicon(IV) oxide in the presence of nitrogen [149, 155–157]:

$$3\,SiO_2 + 6\,C + 2\,N_2 \longrightarrow Si_3N_4 + 6\,CO$$

4) Precipitation and thermal decomposition of silicon diimide [90]:

$$SiCl_4 + 6\,NH_3 \longrightarrow Si(NH)_2\downarrow + 4\,NH_4Cl$$
$$3\,Si(NH)_2 \longrightarrow Si_3N_4 + 2\,NH_3$$

These methods yield products with widely different properties. Silicon nitride powder that is produced by nitriding silicon must be comminuted and the impurities introduced during comminution must be removed by leaching. Important factors for further processing are the specific surface area (generally regarded as a criterion for the sintering activity of the powder) and the content of nonmetallic (O, C) and metallic (Fe, Al, Ca) impurities. Other criteria include the crystallinity, the ratio of α- and β-forms, and the morphology of the powder particles. An oxygen content of ca. 1 – 2 wt % is advantageous because the silicon dioxide layer on the silicon nitride particles substantially promotes sintering. In contrast, the presence of carbon leads to the decomposition of oxides and impairs sintering.

Sintering of Silicon Nitride. Silicon nitride is a covalently bonded compound with very strong interatomic bonding forces, which makes nitrogen diffusion a very slow process [158]. Thus, sintering of silicon nitride without pressure by self-diffusion alone is virtually impossible. The temperature necessary to accelerate diffusion is markedly higher than the decomposition temperature of silicon nitride at normal pressure (2200 K at 0.1 MPa N_2) [28]. decomposition must be counteracted by a high nitrogen pressure [28, 159].

Four important processes for the production of silicon nitride sintered parts are:

1) Reaction bonding with and without final sintering (reaction-bonded silicon nitride, RBSN);
2) Pressureless sintering (sintered silicon nitride, SSN);
3) Hot pressing (hot-pressed silicon nitride, HPSN); and
4) Hot isostatic pressing (isostatically pressed silicon nitride, IPSN).

Reaction bonding with sintering makes use of the fact that preforms made of silicon powder can be converted into reaction-bonded silicon nitride parts (RBSN) without dimensional changes. The preforms are heated in nitrogen until complete nitriding of silicon has occurred. The preforms are produced by conventional methods such as slip casting, injection molding, or isostatic pressing. After evaporation of plasticizing and pressing additives, the preform is heated in a nitrogen atmosphere. Nitriding begins at ca. 1300 K, and the temperature must not exceed 1680 K (the melting point of silicon).

A 22 % volume increase occurs during nitriding and the pores are thus partially filled [160]. The advantage of this method is that the dimensions of the preforms are closely retained, with no danger of distortion or crack formation. The high temperature strength of the product is advantageous because additives are not required, but stability to oxidation is reduced by the large surface area. Its primary disadvantage is that the maximum obtainable density is only ca. 90 % of the theoretical value [161].

Reaction-bonded silicon nitride can be cheaply redensified by resintering (the sintered reaction-bonded silicon nitride SRBSN process) in which sintering agents are admixed to the silicon powder. The other processing steps are analogous to the production of RBSN; however, additional thermal treatment is necessary.

Pressureless sintering has a potentially wide range of industrial applications [162–165]. Additives can be added to the starting powder. The additives react with the silicon dioxide layer on the silicon nitride particles to form a lower-melting silicate phase in which redissolution leads to the precipitation of β-silicon nitride in the form of rods [166]. The morphology of the precipitated β-phase has a favorable influence on the mechanical properties of the product. Oxide sintering additives include MgO, Y_2O_3, CeO_2, ZrO_2, and Al_2O_3. Nonoxide additives are mainly nitrides composed of two elements (e.g., Mg_2N_3, AlN, TiN, ZrN, and CrN). Sintering is carried out at ≤ 2150 K and atmospheric pressure, or at 2400 K and increased pressure (1 – 2 MPa) [163–165, 167].

Sintered silicon nitride (SSN) contracts by ca. 15 % on sintering but has no structural anisotropy. The strength of the sintered part decreases at > 1300 K due to formation of a glass phase. Subsequent thermal treatment leads to partial crystallization of the amorphous grain boundary phase and thus increases the strength [168].

Hot pressing of silicon nitride is carried out mainly in graphite dies at $\leq$ 2100 K and ≤ 50 MPa. Silicon nitride raw material with additives is used (see above), and fully dense sintered parts can be obtained after a pressing time of 10 min. The elimination of pores by the capillary forces acting during the sintering and diffusion processes is enhanced by the applied pressure. The disadvantages of hot pressing are that only geometrically simple forms can be produced and expensive finishing processes with diamond wheels are necessary. Hot-pressed silicon nitride has a directionally dependent strength, with typical values of ca. 850 MPa.

Hot isostatically pressed (HIP) silicon nitride is produced by subjecting powder compacts or presintered silicon nitride parts to the simultaneous action of gas pressure (< 200 MPa) and high temperature (> 2000 K). This process effectively eliminates closed porosity [169]. In the case of open porosity, the sintered part must be encapsulated before being pressed. Refractory metals, highly viscous glasses [170], or sintered layers of ceramic powders (possibly with intermediate BN layers) [171] are used as encapsulating materials and must be removed after the HIP process. The disadvantages of this method lie in the extreme process conditions and the difficulty of encapsulating geometrically complicated parts. In the sinter-HIP process the material is presintered at atmospheric pressure to a density of up to 94 %, tantamount to closed porosity. A completely dense material is obtained in the same furnace run by subsequently increasing the pressure to ca. 10 MPa. The production costs of this sinter-HIP process are lower than those of the HIP process.

Properties and Uses of Silicon Nitride Ceramics. In 1988 ca. 200 – 250 t of silicon nitride ceramic parts were produced worldwide. These ceramics are characterized by relatively high strength and toughness values, a low thermal expansion coefficient, high thermal conductivity, and high thermal shock resistance (Table 2). The strength remains unchanged at high temperature and surpasses that of the best superalloys above 1400 K. Stability to oxidation is excellent. The materials are also stable to combustion gases containing sulfur and halogens.

The properties of silicon nitride ceramics depend on their microstructure, e.g., grain boundary content, glass phase content, type and amount of additive, production method, and porosity.

Because of its properties, silicon nitride is used in heavy-duty components in automotive engines (e.g., precombustion chambers, cylinder linings, turbochargers, and piston heads) [172]. As early as 1985, silicon nitride turbochargers were used in automobiles in Japan [173]. They

permit the use of higher temperature and thus improve the efficiency of fuel utilization. They also reduce engine weight and structural complexity because cooling equipment is unnecessary. Steel ball bearings can be operated only up to 620 K [174]. By contrast, silicon nitride bearings can be used up to 1100 K, even in highly corrosive atmospheres and media (e.g., HCl, H_2SO_4) [175, 176]. Silicon nitride ceramics are also used in high-performance pressure valves and diffusion nozzles, e.g., in carbon hydrogenation and ammonia synthesis [177]. In addition, they have applications in metalworking (e.g., cutting tools, drawing dies, and cutting torch nozzles) [178].

Silicon nitride layers are used as diffusion barriers and passive coatings in semiconductor components [179]. They are produced by reaction of silicon tetrachloride with hydrazine [180] or by physical vapor deposition [181].

3.5. Miscellaneous Covalent Nitrides

The remaining nitrides of the main group elements are interesting primarily because of their semiconductor properties. They can be produced either by nitriding the elements with nitrogen at high pressure or by reaction in a stream of ammonia. Indium nitride [*25617-98-5*], InN [108, 182, 183], has been proposed as a semiconductor layer for field effect transistors (FETs) and metal – alumina – nitride – semiconductor (MANS) diodes [184, 185]. Thermal nitriding of indium phosphide (InP) wafers by reaction with ammonia at ca. 800 K gives a thin (ca. 5 – 10 nm) indium nitride layer.

Germanium nitride [*12065-36-0*], Ge_3N_4, is thermally unstable [186]. Semiconductor layers of germanium nitride on gallium arsenide (GaAs) are important for the production of metal insulator semiconductors [187, 188]. Amorphous tin [*55574-97-5*], lead [*52036-89-2*] [189–191], and antimony [*12333-57-2*] [192] nitrides have been studied as transparent, electrically conducting films for electrooptical purposes.

Layers of phosphorus nitride [*12136-91-3*], P_3N_5, can be produced by the reaction of phosphine (PH_3), ammonia, and nitrogen at < 1200 K; they are of interest for electronic microtransistors [193].

4. Transition-Metal Nitrides

Transition-metal nitrides attracted interest at an early stage because of their striking properties such as hardness, metallic luster, and color [5]. They are less thermally stable than carbides but resemble them in many ways (→Carbides). Only the nitrides of the group 4 metals titanium, zirconium, and hafnium melt at a nitrogen pressure of 1 MPa without decomposition.

4.1. Production

Transition-metal nitrides can be produced according to the methods specified in Chapter 1; however, precise knowledge of the thermodynamic data of nitrides is necessary [194]. The thermodynamic stability of nitrides decreases with increasing group number; hence, the nitrides of higher group numbers ($\geq$ group 6) can be prepared only by nitriding with nitrogen at high pressure or by nitriding in a stream of ammonia gas. Thus, titanium nitride, zirconium nitride, and hafnium nitride are the most stable nitrides with the lowest nitrogen equilibrium pressures.

Nitride formation with group 4 and 5 metals is exothermic. The heat released during the reaction of metal powder with nitrogen gas can be sufficient to reach the melting temperature of the metal. This is undesirable for the industrial production of nitride powders because the dense-sintered or molten metal reguli (metallic mass formed at the bottom of the vessel) do not react completely with nitrogen; hence, the desired fine-grained nitride powder is difficult to obtain from the reaction products by comminution.

Reaction of metal oxides with carbon and nitrogen yields nitrides containing oxygen and carbon. Reaction of a very fine-grained stoichiometric mixture of titanium(IV) oxide and carbon in the presence of nitrogen can yield titanium nitride with a low oxygen and carbon content and with properties favorable for the use of titanium carbonitride hard metals (cermets) [195].

High-temperature annealing, usually at increased nitrogen pressure, is employed to produce granular or single-crystal nitrides [196, 197]; zone drawing and zone melting methods

are used for the production of large single crystals [198–200]. In zone melting the nitrogen pressure should be kept at levels high enough to prevent both decomposition of the nitride and formation of heterogeneous compounds with a low stoichiometry.

Single crystals of transition-metal nitrides can also be produced by crystallization from fluxes [201–203] and CVD [204]. Flux growth exploits the fact that nitrides are soluble in metal melts and that coarsely crystalline samples and single crystals can be obtained by cooling or crystal extraction.

Volatile chlorides of the transition metals can be employed to advantage in the CVD reactions:

$$TiCl_4 + 1/2\,(1-x)\,N_2 + 2\,H_2 \longrightarrow TiN_{1-x} + 4\,HCl$$

The CVD process is used mainly to coat cutting tools with thin nitride layers (see Section 4.3.1).

Physical vapor deposition and ion-assisted deposition (IAD) utilize the recombination of metal atoms or metal ions with nitrogen in the gas phase to deposit nitride layers [205]. The metal targets are either subjected to cathode sputtering or to evaporation in an arc plasma or by an electron beam. The metal ions or atoms recombine with nitrogen at low pressure, preferentially on the surface of solid "substrates". The properties of the resulting nitride layers are highly dependent on sputtering conditions [206–211].

4.2. General Properties

Data for transition-metal nitride systems are still lacking, the known phases are listed in Table 3. In addition to the technically important mononitride phases with their very wide range of phase composition with respect to nitrogen content (made explicit in the notation by a free parameter, e.g., TiN_{1-x} [*11116-16-8*]), several other phases often exist with a lower nitrogen content and a narrower range of composition. Some nitride phases may contain more nitrogen atoms than metal atoms (e.g., Nb_4N_5 [*12594-34-2*] [212] and Ta_3N_5 [*12033-94-2*] [213]).

Within the homogeneity range, virtually all physical, chemical, and mechanical properties of the nitride phase are subject to continuous change. Deviations from "ideal" stoichiometry can, however, lead to the formation of states of order in the partial metalloid lattice [214, 215] that cause distortion and hence a reduction of symmetry. The crystal structures of binary transition-metal nitrides are generally close-packed (face-centered cubic, hexagonal close-packed, tetragonal close-packed) arrays of the larger metal atoms, with the smaller nitrogen atoms in the interstitial voids [41].

Table 3. Known phases of binary transition-metal nitrides ∗

Nitride	Phases
Ti – N	ε-Ti_2N (t), η-Ti_3N_{2-x} (r), ζ-Ti_4N_{3-x} (r), δ'-$TiN_{0.5}$ (t), δ-TiN_{1-x} (c)
Zr – N	δ-ZrN_{1-x} (c), Zr_3N_4 (c)
Hf – N	ε-Hf_3N_{2-x} (r), ζ-Hf_4N_{3-x} (r), δ-HfN_{1-x} (c)
V – N	β-V_2N (h), δ-VN_{1-x} (c), δ'-$VN_{0.8}$ (t), VN (t)
Nb – N	β-Nb_2N (h), γ-Nb_4N_{3-x} (t), η-NbN (h), δ'-NbN (h), δ-NbN_{1-x} (c), Nb_5N_6 (h), Nb_4N_5 (t)
Ta – N	β-Ta_2N (h), ε-TaN (h), η-TaN (r), δ-TaN_{1-x} (c), Ta_5N_6 (h), Ta_4N_5 (t), Ta_3N_5 (o)
Cr – N	Cr_2N (h), δ-CrN (c), CrN (o)
Mo – N	γ-MoN_{1-x} (c), β-Mo_2N (t), δ-MoN (h)
W – N	W_2N (c), WN (t)
Mn – N	Mn_4N (c), Mn_2N_{1-x} (o), Mn_3N_2 (t), Mo_6N_5 (t), MnN_{1-x} (c)
Tc – N	
Re – N	Re_2N (c)
Fe – N	Fe_4N (c), Fe_2N (o), Fe_2N_{1-x} (t)
Ru – N	
Os – N	
Co – N	Co_4N (c), Co_3N_{1-x} (h), Co_3N (t), Co_2N (o)
Rh – N	
Ir – N	
Ni – N	Ni_4N (c), Ni_4N (t), Ni_3N (h)
Pd – N	
Pt – N	

∗ Crystal system: c, cubic; h, hexagonal; t, tetragonal;o, orthorhombic; r, rhombohedral (trigonal).

The technically useful "mono" nitrides, TiN_{1-x}, ZrN_{1-x}[*119173-61-4*], and HfN_{1-x} [*11115-87-0*], crystallize in the sodium chloride-type lattice. The high-temperature phases of the mononitrides niobium nitride (NbN) and tantalum nitride (TaN) also crystallize with the sodium chloride structure [216–219]. However, these mononitrides undergo phase transitions to modifications with more complicated structures in which nitrogen may not be octahedrally coordinated.

Transition-metal nitrides exhibit metallic character but, in contrast to metals, are brittle and usually very hard. They have high thermal and electrical conductivities as well as a metallic luster. The mononitrides even display superconductivity.

Transition-metal nitrides can form solid solutions with other isotypic nitrides, carbides, and oxides. They are wetted by liquid metals, particularly those belonging to the iron group, if the oxide surface layer that inhibits wetting can be decomposed or removed. Transition-metal nitrides are highly resistant to acid corrosion at room temperature; however, they are slowly dissolved by hot acid and rapidly dissolved by hydrofluoric acid. Stability to oxidation is comparable to that of carbides. The technically most important properties of transition-metal nitrides are listed in Table 4 [220–234].

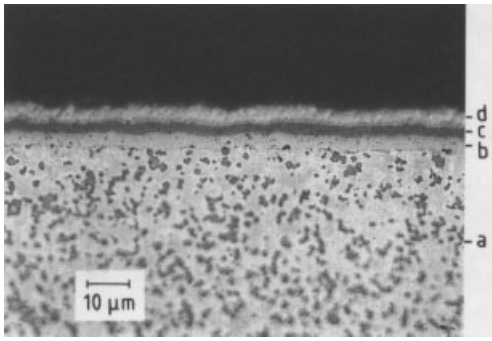

Figure 8. Multilayer coating (CVD) of a hard metal (courtesy of Plansee Tizit, Reutte, Austria)
a) WC – TiC – TaC – Co hard metal; b) TiC; c) Ti (C, N); d) TiN

4.3. Uses

The uses of transition-metal nitrides exploit their properties (hardness, wear resistance, chemical resistance, stability to oxidation, resistance to diffusion of metals, and color).

4.3.1. Wear-Resistant Layers

Most tools for metal cutting are coated with hard materials to reduce wear on the edge of the cutting tool, to prolong tool life, or to increase the cutting speed. These thin ($< 10\,\mu m$) but effective coatings (Fig. 8) should be hard and chemically inert to the material of the running chip moving across the rake face of the cutting tool. They should also adhere well to the substrate and have a thermal expansion coefficient similar to that of the substrate.

Transition-metal nitrides, especially titanium nitride but also zirconium and hafnium nitrides, are particularly suitable for reducing the wear on cutting tools (mainly steel or hard metals).

Use is made of the fact that the nitrides have a lower frictional welding tendency at the tool – workpiece interface compared to that of the uncoated cutting tool (e.g., steel). The adhesive properties of transition-metal nitrides are similar to those of carbides; however, the latter are superior in terms of hardness.

Three processes are employed for applying wear-resistant layers:

1) Chemical vapor deposition (CVD)
2) Physical vapor deposition (PVD)
3) Case hardening in a nitriding bath or with a plasma

CVD and PVD Nitride Layers. For a detailed description of CVD and PVD, see → Thin Films. The CVD of nitrides primarily employs the reaction of volatile chlorides with nitrogen – hydrogen mixtures in the gas phase at a substrate temperature of ca. 1300 K. In PVD processes, however, a temperature of ca. 800 K is used [205, 235].

In ion plating (PVD), the metal component is evaporated (e.g., with an electron beam or electric arc) and reacts with nitrogen ions formed in a low-pressure glow discharge; the resulting nitride is precipitated on the substrate. Because of the low substrate temperature ($< 560\,°C$) high-speed steel tools can be coated below their softening temperature. Sputtering and ion plating are carried out at low pressure where the reactive particles are mainly propagated in straight lines. Development of shadowing effects must therefore be prevented by turning the substrate parts during the coating process.

Both CVD and PVD processes produce layers [often a combination of different hard-material layers such as TiN, TiC, Ti(C,N), and Al_2O_3] [236, 237] with a thickness of ca. 5 – 10 µm. Thicker layers on tools are used for continuous cutting (e.g., carbide tips for turning on a lathe); thinner layers, for interrupted cutting (milling machine).

Layers deposited by CVD are much closer to thermodynamic equilibrium than those obtained by PVD [238]. Thus, the physical and mechanical properties of the PVD layers often differ substantially from those of compact nitrides [239,

Table 4. Properties of metal nitrides in bulk form (groups 4 – 6)

Nitride	Homogeneity range $1-x$ at 1700 K	Structure, Bravaislattice [a]	Cell dimensions (300 K), nm	Density, g/cm^3	Color	Microhardness, *mp* [b], K GPa		Thermal conductivity, W m^{-1} K^{-1} [b]	Thermal expansion coefficient (1000 K), 10^{-6} K^{-1} $(1-x)$	Electrical resistivity (300 K), $\mu\Omega\cdot$cm [b]	Superconductivity transition temperature, K $(1-x)$
TiN_{1-x}	0.50 –	fcc	0.4215		gray	18					<1
	1.00		0.4242	5.39 [c]	golden yellow	14	3320	29.1	9.4 ($TiN_{0.95}$)	25	5.8
ZrN_{1-x}	0.54 –	fcc	0.4585		yellow						
	1.00		0.4570	7.32 [c]		15	3250	10.9	8.2 – 9.1 ($ZrN_{0.92}$)	21	10.47
HfN_{1-x}	0.70 –	fcc	0.4525		olive yellow						8.7 ($HfN_{0.85}$)
	1.00		0.4510	13.83		16.5	3600	11.1	8.6	33	6.92
VN_{1-x}	0.70 –	fcc	0.4060	6.05	brownish yellow	13.0			11.0		2.7 ($VN_{0.76}$)
	1.00		0.4138	6.04	brownish yellow	5.7	2620	11.3	10.8	85	8.9
NbN_{1-x}	0.84 –	fcc	0.4380	8.24 ($NbN_{0.92}$)	pale yellow	13.0			9.5 ($NbN_{0.92}$)		13.8 ($NbN_{0.84}$)
	1.00		0.4392	8.18 ($NbN_{0.95}$)	pale yellow	11.0	[d]	3.8	10.2 ($NbN_{0.95}$)	60	17.2
TaN_{1-x}	0.72 (2200 K)	fcc	0.4345								
	1.00		0.4328	15.9 [c]	grayish yellow	32	[d]		8.0		9.1 [e]
TaN	>0.99 –	h	*a*: 0.2936	14.3	dark gray	11		9.54		128	
	1.00		*c*: 0.2885				[d]				
CrN	>0.97 [f] – 1.00	fcc	0.4148	6.14	dark brown	11	1300 [d]	11.7	2.3	640	Not superconductive
CrN_{1-x} (Cr_2N)	0.31 [f] –	h	*a*: 0.4750 [e]	6.51	gray	12 [e]					Not superconductive
			c: 0.4430 [e]								
	0.50		*a*: 0.4796			14 [b]	1800			81	
			c: 0.4470								
MoN_{1-x} (Mo_2N)	0.40 –	fcc	0.4139								
	0.54 (1200 K)		0.4162	9.48 [c]	gray	17 [b]			9.3		5.0
MoN		h	*a*: 0.5745	8.1	gray						15.1
			c: 0.5622	9.1 [c]							

[a] fcc = face-centered cubic; h = hexagonal. [b] Exact composition unknown, stoichiometry probably close to 2 : 1 (CrN_{1-x}, MoN_{1-x}) or 1 : 1 (all others). [c] X-ray density. [d] Decomposes. [e] Exact composition unknown. [f] Calculated.

240], whereas CVD layers have comparable properties [241, 242]. The PVD layers are more finely grained, are usually subject to compressive stresses, and have a higher degree of hardness up to ca. 1300 K; these factors improve wear resistance. Above ca. 1300 K the hardness values of PVD and CVD layers become similar due to thermally induced rearrangement and recrystallization. The room-temperature Vickers hardnesses (50-g load) of PVD and CVD titanium nitride layers are ca. 2400 and 1800 kg/mm^2, respectively; corresponding values for hafnium nitride layers are 3000 and 1200 kg/mm^2 [239].

Layers produced by PVD have large lattice defect concentrations that affect their mechanical properties [243]. The adhesive strength of CVD layers is usually excellent, due to the high substrate temperature and the reactive components in the gas phase (HCl). However, in coating WC – Co hard-metal tools decarburization may occur at the interface between the carbide substrate and the nitride layer, resulting in a deterioration of edge strength. The cutting edges of hard-metal indexable inserts are therefore rounded or chamfered prior to CVD coating.

Layers of TiN and Ti (C, N) are used particularly for indexable inserts in steel cutting tools and for dies. The trend is toward multilayer coating for metal cutting tools. Because of their striking color, layers of TiN, Ti (C, N), ZrN, and Hf N are used for decorative purposes and for increasing hardness (scratch resistance) of jewelry (e.g., watch cases).

Internationally, ca. 70 % of all hard-metal tools are currently used in the as-coated condition. In Europe, ca. 650×10^6 indexable inserts coated with hard materials are produced, most of them being coated with titanium nitride.

Nitride Layers Produced by Case Hardening. Nitride hardening of materials [244, 245] is carried out by reacting the material surface with nitrogen to produce a hard, wear-resistant layer (case) (→ Metals, Surface Treatment, Chap. 4.2.2.). Hard, corrosion-resistant Fe_3N and Fe_4N layers as well as ternary iron carbonitrides are formed on steel surfaces by methods such as ammonia, plasma, and salt-bath nitriding. The formation of various nitrides and carbonitrides of the alloy components of alloyed steels has been reported [e.g., VN, Cr_2N, CrN, Ni_3(C, N), Mn_4N, and Co_3N] [246]. These improve the mechanical properties of the surface of the material (precipitation hardening). Case hardening has the advantage that the treated items do not become distorted, unlike quench hardening which usually results in small dimensional changes or even distortions. In the nitriding of titanium alloys, the strength and increased hardness of titanium nitride are exploited [247, 248].

Flame spraying and plasma spraying methods are not used to produce wear-resistant nitride layers because the nitrides decompose at high temperature [249, 250].

4.3.2. Cermets

Since the 1970s, titanium nitride and titanium carbonitride have become particularly important as hard components in tungsten carbide-free hard metals [251] (cermets [252]).

The term "cermet" originally denoted all composites consisting of a ceramic material embedded in a metal matrix. Its meaning has recently been narrowed down, at least in the hardmetal industry, to denote sintered composite materials consisting of hard ceramic component(s) [TiC, TiN, or Ti(C, N), with Mo_2C, WC, and other carbides as additives] and a metal binder (Ni, Ni – Mo, or Ni – Co – Mo) [253–256]; see also → Ceramics, Ceramic – Metal Systems, Chap. 6.1.

Solid solutions of (Ti, Mo) (C, N) or (Ti, W) (C, N) tend to decompose on cooling into two isostructural phases α' and α''; the α'-phase is rich in nitrogen and titanium, whereas the α''-phase is rich in carbon and molybdenum (tungsten). In liquid-phase sintering [257], this effect (spinodal separation) produces hard particles consisting of a molybdenum-rich (Ti, Mo)C rim and a nitrogen rich Ti (C, N) core (Fig. 9). The rim is easily wetted by the binding metal, whereas the core has excellent wear resistance.

Sintering and diffusion mechanisms in cermets are still not well understood [258]. The specific grain structure can be obtained not only by spinodal decomposition of a homogeneous solid solution but also if titanium nitride (TiN), titanium carbide (TiC), molybdenum carbide (Mo_2C), or tungsten carbide (WC) are used as starting materials in cermet production. Nitrogen seems to displace molybdenum from the

core of the hard grain to the periphery and thus improves adhesion of the grain in the binder. A development is the addition of ternary nitrides and carbides to the powder mixture [256].

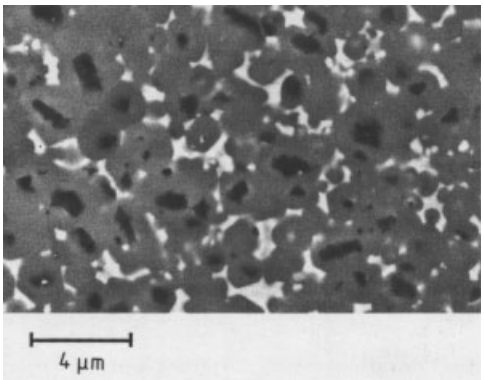

Figure 9. Typical core and rim structure in a Ti (C, N) cermet (TTI, Krupp WIDIA, Essen, FRG)
Dark areas (core) are composed of Ti (C, N), gray areas (rim) of (Ti, Ta, Mo, W) C_{1-x}, and light areas of the matrix alloy.

In Japan, 30 % of the conventional hard metals for metal cutting have been replaced by cermets; the annual growth rate is ≤ 5 %. In the machining of steel with low chip cross-sections, cermets are superior to the conventional WC – TiC – TaC – Co hard metals in terms of holding time and cutting speed and especially when high surface quality is required. Furthermore, cermets are only ca. one-third as dense as WC – Co hard metals.

4.3.3. Other Uses

Some of the more stable transition-metal nitrides such as titanium nitride, TiN [259–261], zirconium nitride, ZrN [262, 263], hafnium nitride, HfN [264], and tantalum nitride [*12033-62-4*], TaN [265, 266], are used as diffusion barriers in semiconductor technology [267, 268]. Their suitability for this application was recognized when the addition of nitrogen [269] improved the properties of Ti–W alloys [270]. The metallization of integrated circuits is required for attaching beam leads [260] to semiconductor components [268], in solar cells [271, 272], for heat sinks, and for protection against atmospheric effects. Semiconductors such as silicon [265] and gallium arsenide (GaAs) [273, 274] are coated with metals such as aluminum, gold, nickel, titanium, and palladium. Diffusion barriers are required to prevent the undesirable formation of intermetallic phases between the semiconductor and the metallized layer [275]. The material of the diffusion barrier must meet the following requirements [265, 276]:

1) High thermodynamic stability, no reaction with metals such as silicon, copper, gold, nickel, palladium, and platinum;
2) High diffusion resistance to these metals;
3) Good adhesive properties;
4) Low ohmic resistance; and
5) Good thermal conductivity.

The diffusion barriers are formed with a thickness of ca. 100 nm by reactive sputtering. Compositions (e.g., nitrogen and oxygen content) can be varied and affect the properties of the barrier [261, 277].

Other potential uses of transition-metal nitrides are found mainly in electronics and optics. Titanium nitride has been proposed for the surface treatment of windows in microwave technology [278] and as a metallization layer on aluminum nitride ceramic substrates; titanium nitride, tantalum nitride, zirconium nitride, and hafnium nitride have been suggested as heat reflectors [279, 280]; and tantalum nitride has potential use as a protective layer for moisture sensors [281]. Attempts to apply iron nitrides [282–284] (Fe_4N has a Curie temperature of > 800 K [285]) and cobalt nitrides [286, 287] as magnetizable compounds in powder form or as layers for magnetic storage are also promising.

The superconductive properties of cubic phases [288], especially of zirconium nitride, ZrN [289–291], vanadium nitride, VN [228], niobium nitride [*24621-21-4*], NbN [292, 293], and molybdenum nitride [*12033-19-1*], MoN [234, 294–300], have been studied. Mixed crystals of nitride pairs can have minimum (e.g., NbN – VN) or maximum (e.g., NbN – TiN) transition points [288].

The use of ZrN, Zr_3N_4, and NbN has been suggested for Josephson junctions in fast computers [289]; the special resistance of the NaCl-type nitrides to radiation [301, 302] could lead to applications in nuclear fusion reactors [303].

4.4. Multicomponent Nitrides

Metal atoms, which otherwise only form covalent nitrides, have been integrated into binary transition-metal nitrides by using PVD to synthesize substances that are not in thermodynamic equilibrium.

An interesting development in the area of multicomponent nitrides is the production of (Ti, Al)N and (Ti, V, Al)N layers [304–306]. Magnetron sputtering (a PVD process) has been used to obtain metastable ternary Ti – Al – N and quaternary Ti – Al – V – N layers with good wear resistance (e.g., for indexable inserts) [307]. The face-centered cubic structure of these metastable materials remains unchanged even after annealing.

A similar technological potential can also be anticipated for Ti – B – N and Ti – C – B – N layers [308] with degrees of hardness up to 50 GPa on the Vickers scale. Very hard sandwich films can be obtained with alternating titanium nitride and vanadium nitride layers (thickness 0.8 – 16 nm) [309].

An interesting modification of β-boron nitride formation is the deposition of quaternary Ti – C – B – N layers [87, 88]. A comparison of Ti – B – N and Ti – C – B – N layers with Ti – B and Ti – C – B layers indicates a marked increase in the degree of hardness (Vickers hardness up to 60 GPa) due to the incorporation of nitrogen.

Rapid development may be expected in the field of PVD because of the large range of available techniques and the possibilities of combining different materials.

5. References

1. *Gmelin,* **System No. 41.**
2. J. M. Herndon, H. E. Suess, *Geochim. Cosmochim. Acta* **40** (1976) 395.
3. F. Wöhler, *Ann. Chim. Phys.* **28** (1850) no. 3, 382.
4. R. Kieffer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963.
5. L. E. Toth: " Transition Metal Carbides and Nitrides," *Refractory Materials Series,* **vol. 7,** Academic Press, New York–London 1972.
6. *Kirk-Othmer,* **15,** 871.
7. K. A. Gschneidner Jr., N. Kippenhan: *Thermochemistry of the Rare Earth Carbides, Nitrides and Sulfides for Steelmaking,* Rare Earth Information Center, Iowa State University, Ames, Iowa 1971.
8. A. Rabenau, in J. Treusch (ed.): "Lithium Nitride", *Festkörperprobleme,* vol. 18, Verlag Vieweg, Braunschweig 1978, p. 77.
9. H. Schulz, K. H. Thiemann, *Acta Crystallogr. Sect. A:* **A 35** (1979) 309.
10. J. Wahl, *Solid State Commun.* **29** (1979) 485.
11. A. Rabenau, *Solid State Ionics* **6** (1982) 277.
12. E. Schönherr, G. Müller, E. Winkler, *J. Cryst. Growth* **43** (1978) 469.
13. E. Schönherr, A. Köhler, G. Pfrommer, *Inorg. Synth.* **22** (1983) 48.
14. U. v. Alpen, *J. Solid State Chem.* **29** (1979) 379.
15. H. J. Beister et al., *Angew. Chemie* **100** (1988) no. 8, 1116.
16. R. H. Wentorf, *J. Chem. Phys.* **34** (1961) 809.
17. De Beers Ind., EP-A 192 391, 1985 (G. J. Davies, L. K. Hedges).
18. S. Yamaoka et al., *Physica B + C (Amsterdam)* **139 – 140 B + C** (1986) no. 1 – 3, 668.
19. H. Lorenz, B. Lorenz, U. Kuehne, C. Hohlfeld, *J. Mater. Sci.* **23** (1988) 3254.
20. H. A. Wriedt, H. Okamoto, *Bull. Alloy Phase Diagrams* **8** (1987) 136.
21. G. Wötting, G. Ziegler, *Sprechsaal* **119** (1986) no. 4, 265.
22. G. R. Terwilliger, F. F. Lange, *J. Am. Ceram. Soc.* **57** (1974) 25.
23. C. D. Greskovich, C. R. O'Clair, US 4 124 402, 1978.
24. R. Becker, *Kernforschungszent. Karlsruhe, Ber KfK* 2771 (1979).
25. J. C. Schuster, *Rev. Chim. Miner.* **24** (1987) 676.
26. M. P. Borom, G. A. Slack, J. W. Szymanszek, *Am. Ceram. Soc. Bull.* **51** (1972) 852.
27. G. A. Slack, *J. Phys. Chem. Solids* **34** (1973) 321.
28. S. Prochazka, C. D. Greskovich, AMMRC TR 78-32 (1978) SRD-77-178.
29. Tanaka Noble Metal Industrial Co., JP-Kokai 58 207 344, 1982.
30. Tanaka Nobele Metal Industrial Co., JP-Kokai 58 207 345, 1982.
31. W. Lengauer, *J. Solid State Chem.* **76** (1988) 412.
32. J. P. Dismukes, W. M. Yim, J. J. Tietjen, R. E. Novak, *RCA Rev.* Dec. 1970, 680.
33. G. Travaglini et al., *Phys. Rev. B, Condens. Matter* **34** (1986) 3876.
34. J. C. François et al., *Thin Solid Films* **127** (1985) 205.

35. P. Wachter, E. Kaldis, *J. Magn. Magn. Mater.* **15 – 18** (1980) 305 – 306.
36. Hitachi, JP 61 215 266, 1986 (T. Fujita, S. Iwanaga, T. Nima).
37. C. L. Hoenig, *J. Am. Ceram. Soc.* **54** (1971) 391.
38. A. L. Bowman, G. P. Arnold, *Acta Crystallogr. Sect. B* : B 27 (1971) 243.
39. H. J. Berthold, H. G. Hein, H. Reuter, *Ber. Dtsch. Keram. Ges.* **50** (1973) 111.
40. K. A. Schwetz, A. Lipp, *Ber. Dtsch. Keram. Ges.* **56** (1979) no. 1, 1.
41. P. Villars, L. D. Calvert (eds.): *Pearson's Handbook of Crystallographic Data of Intermetallic Compounds,* vol. 1 – 3, Metals Park, Ohio 1985.
42. C. F. Gardinier, *Am. Ceram. Soc. Bull.* **67** (1988) no. 6, 1006.
43. F. P. Bundy, R. H. Wentorf, *J. Chem. Phys.* **38** (1963) 1144.
44. R. C. DeVries: "Cubic Boron Nitride," *Handbook of Properties,* **General Electric CRD Rept. No. 72CRD178,** 1972.
45. R. H. Wentorf, *J. Chem. Phys.* **26** (1957) 956.
46. R. H. Wentorf, *J. Chem. Phys.* **34** (1961) 809.
47. H. Kolaska (ed.): *Schneidstoffe, Reihe Pulvermetallurgie in Wissenschaft und Praxis* vol. 4, Verlag Schmidt, Freiburg i. Br. 1988.
48. S. P. S. Arya, A. D'Amico, *Thin Solid Films* **157** (1988) 267 – 282.
49. M. J. Rand, J. Roberts, *J. Electrochem. Soc.* **115** (1968) 423.
50. S. P. Murarka, C. C. Chang, D. N. K. Wang, T. E. Smith, *J. Electrochem. Soc.* **126** (1979) 1951.
51. M. Hirayama, K. Shohno, *J. Electrochem. Soc.* **122** (1975) 1671.
52. W. Baronian, *Mater. Res. Bull.* **7** (1972) 119.
53. M. Sano, M. Aoki, *Thin Solid Films* **83** (1981) 247.
54. S. Motojima, Y. Tamura, K. Sugiyama, *Thin Solid Films* **88** (1982) 269.
55. T. Takahashi, H. Itoh, A. Takeuchi, *J. Cryst. Growth* **47** (1979) 245.
56. T. Takahashi, H. Itoh, M. Kurdoda, *J. Cryst. Growth* **53** (1981) 418.
57. A. Zunger, A. Katzir, A. Halperin, *Phys. Rev. B* **13** (1976) 5560.
58. C. F. Powell, J. H. Oxley, J. M. Blocher: *Vapor Deposition,* Wiley-Interscience, New York 1966, p. 663.
59. H. O. Pierson, *J. Compos. Mater.* **9** (1975) 228.
60. K. Nakamura, *J. Electrochem. Soc.* **132** (1985) 1757.
61. A. C. Adams, *J. Electrochem. Soc.* **128** (1981) 1379.
62. A. J. Noreika, M. H. Francombe, *J. Vac. Sci. Technol.* **6** (1969) 722.
63. M. D. Wiggins, C. R. Aita, F. S. Hickernell, *J. Vac. Sci. Technol. A* **2** (1984) 322.
64. P. D. Davidse, L. I. Maissel, *J. Appl. Phys.* **37** (1966) 574.
65. E. H. Lee, H. Poppa, *J. Vac. Sci. Technol.* **14** (1977) 223.
66. C. Weissmantel in K. J. Klabunde (ed.): *Thin Film Formation from Free Atoms and Particles,* Chap. 4, Academic Press, New York 1985.
67. C. Weissmantel et al., *Thin Solid Films* **72** (1980) 19.
68. C. Weissmantel et al., *Thin Solid Films* **96** (1982) 31.
69. B. Rother et al., *Thin Solid Films* **142** (1986) 83.
70. C. Weissmantel, *J. Vac. Sci. Technol.* **18** (1981) 179.
71. M. Sokolowski, *J. Cryst. Growth* **46** (1979) 136.
72. M. Sokolowski et al., *J. Cryst. Growth* **52** (1981) 165.
73. M. Sokolowski et al., *Thin Solid Films* **80** (1981) 249.
74. S. Shanfield, R. Wolfson, *J. Vac. Sci. Technol. A* **1** (1983) 323.
75. D. Maydan et al., *J. Vac. Sci. Technol.* **16** (1979) 1959.
76. R. E. Acosta, J. R. Maldonado, R. Fair, *J. Vac. Sci. Technol. B* **4** (1986) 240.
77. L. Hanlon, M. Greenstein, W. Grossmann, A. Neukermans, *J. Vac. Sci. Technol. B* **4** (1986) 305.
78. L. Guzman et al., *Thin Solid Films* **117** (1984) L63.
79. W. Torbicz et al., *Phys. Status Solidi A* **86** (1984) 453.
80. Sumitomo Electric Industries, JP 59 80 775, 1982 (Y. Doi).
81. Sumitomo Electric Industries, JP-Kokai 63 699 723, 1988 (K. Fukushima, T. Masaaki).
82. Mitsubishi Heavy Industries, JP-Kokai 63 18 051, 1986 (T. Odohira et al.).
83. H. Saito, T. Ishiguro, Y. Ichinose: *8th Int. Symp. on Plasma Chemistry,* **vol. 2,** Tokyo 1987, p. 1148.
84. C. Mitterer, P. Rödhammer, H. Störi, F. Jeglitsch, *J. Vac. Sci. Technol.* **A 7** (1989) 2646.
85. K. Watanabe, K. Saitoh, Y. Yochi, K. Sekino: *Proc. 11th Symp. of ISIAT '87,* Tokyo 1987, pp. 347 – 350.

86. Applied Coatings International, US 4 415 420, 1983 (H. A. Beale).
87. H. Karner, J. Laimer, H. Störi, R. Rödhammer: *Proc. 12th Int. Plansee Seminar,* **vol. 3,** Reutte 1989, pp. 495 – 509.
88. C. Mitterer, M. Rauter, P. Rödhammer, *Proc. 12th Int. Plansee Seminar,* **vol. 3,** Reutte 1989, p. 411.
89. D. G. Brunner, *Sprechsaal* **121** (1988) 181.
90. G. Boden, R. Irmisch, G. Himpel, T. Reetz, *Sprechsaal* **122** (1989) 224.
91. A. G. Merzhanov, *Combust. Flame* **13** (1969) 143.
92. M. Billy, J. Mexmain, *Sprechsaal* **118** (1985) no. 3, 245.
93. T. S. Verkhoglyadova, T. V. Dubovik, G. V. Samsonov, *Poroshk. Metall. (Kiev)* **1** (1961) 9.
94. M. Billy et al., *Rev. Chim. Miner.* **19** (1982) 673.
95. T. Quiehl, *Sprechsaal* **121** (1988) no. 12, 1176.
96. G. Long, L. M. Foster, *J. Am. Ceram. Soc.* **42** (1959) 53.
97. K. Komeya, M. Inoue, *Yogyo Kyokaishi* **77** (1969) 884.
98. S. Yefsah, M. Billy, J. Jarrige, M. Mexmain, *Rev. Int. Hautes Temp. Refract.* **18** (1981) 167.
99. K. Komeya, *Am. Ceram. Soc. Bull.* **63** (1984) no. 9, 1158 – 1159, 1164.
100. K. Komeya, H. Inoue, *J. Mater. Sci.* **4** (1969) 1045.
101. J. W. McCauley, N. D. Corbin, *J. Am. Ceram. Soc.* **62** (1979) 476.
102. P. Goursat, P. Goeuriot, M. Billy, *Mater. Chem.* **1** (1976) 131.
103. G. Ado, M. Billy, P. Guillo, P. Lefort, FR 83 05 017, 1983.
104. J. Heinrich in H. Kolaska, H. Grewe (eds.): *Moderne Formgebungsverfahren, Pulvermetallurgie in Wissenschaft und Praxis,* vol. 1, Verlag DKG, 1985.
105. N. Iwase, K. Anzai, K. Shinozaki, *Toshiba Rev. (Int. Ed.)* 1985 no. 153, 49 – 53.
106. W. C. Johnson, J. B. Parson, M. C. Crew, *J. Phys. Chem.* **36** (1932) 2651.
107. J. Karpinski, J. Jun, S. Porowski, *J. Cryst. Growth* **66** (1984) 1.
108. R. D. Jones, K. Rose, *CALPHAD Comput. Coupling Phase Diagrams Thermochem.* **8** (1984) 343.
109. I. G. Pichugin, D. A. Yaskov, *Izv. Akad. Nauk SSSR Neorg. Mater.* **6** (1970) 1973.
110. R. B. Zetterstrom, *J. Mater. Sci.* **5** (1970) 1102.
111. H. P. Maruska, J. J. Tietjen, *Appl. Phys. Lett.* **15** (1969) 327.
112. M. Ilegems, *J. Cryst. Growth* **13 – 14** (1972) 360.
113. S. S. Liu, D. A. Stevenson, *J. Electrochem. Soc.* **125** (1978) 1161.
114. R. A. Logan, C. D. Thurmon, *J. Electrochem. Soc.* **119** (1972) 1727.
115. V. S. Ban, *J. Electrochem. Soc.* **119** (1972) 761.
116. M. Sano, M. Aoki, *Jpn. J. Appl. Phys.* **15** (1976) 1943.
117. V. P. Chegnov, A. S. Melkova, A. S. Pashinkin, E. B. Sokolov, *Inorg. Mater. (Engl. Transl.)* **15** (1979) 363.
118. R. Fremunt et al., *Cryst. Res. Technol.* **16** (1981) 1257.
119. P. J. Born, D. S. Robertson, *J. Mater. Sci.* **15** (1980) 3003.
120. J. Lang, Y. Laurent, M. Maunaye, R. Marchand, *Prog. Cryst. Growth Charact.* **2** (1979) 207.
121. H. G. Grimmeis, B. Monemar, *J. Appl. Phys.* **41** (1970) 4054.
122. D. Elwell, M. M. Elwell, *Prog. Cryst. Growth Charact.* **17** (1988) 53.
123. J. I. Pankove, *J. Lumin.* **7** (1973) 114.
124. J. I. Pankove, S. Bloom, G. Harbeke, *RCA Rev.* **36** (1975) 163.
125. Y. Morimoto, *J. Electrochem. Soc.* **121** (1974) 1383.
126. G. Jacob, M. Boulou, M. Furtado, D. Bois, *J. Electron. Mater.* **7** (1978) 499.
127. Matsushita, JP 81 59 699, 1981.
128. Tanaka Noble Metal Ind., JP-Kokai 58 204 143, 1982.
129. G. Wötting, G. Ziegler, *Sprechsaal* **119** (1986) 555.
130. G. Wötting, G. Ziegler, *Sprechsaal* **120** (1987) 96.
131. G. Wötting, G. Ziegler, *Sprechsaal* **120** (1987) 521.
132. G. Wötting, G. Ziegler, *Sprechsaal* **122** (1989) 45.
133. G. Wötting, G. Ziegler, *Sprechsaal* **122** (1989) 210.
134. H. Knoch, J. Heinrich, *Z. Werkstofftechn.* **11** (1980) 361.
135. F. F. Lange, *J. Am. Ceram. Soc.* **56** (1973) 518.
136. K. H. Jack in F. L. Riley (ed.): *The Crystal Chemistry of the SiAlONs and Related Nitrogen Ceramics,* Proc. of the NATO Advanced Study Institute on Nitrogen Ceramics, Canterbury 1976, p. 109.
137. P. E. D. Morgan, *J. Mater. Sci.* **15** (1980) 791.
138. A. J. Moulson, *J. Mater. Sci.* **14** (1979) 1017.
139. S. Hampshire, K. H. Park, D. P. Thompson, K. H. Jack, *Nature (London)* **274** (1978) 880.

140. M. Mitomo, H. Tanaka, K. Muramatsu, *J. Mater. Sci. Lett.* **15** (1980) 2661.
141. K. H. Jack, *Sci. Ceram.* **11** (1981) 125.
142. K. H. Jack, *Met. Technol.* **9** (1982) 297 – 301.
143. Y. Oyama, O. Kamigaito, *Jpn. J. Appl. Phys.* **10** (1971) 1637.
144. K. H. Jack, W. I. Wilson, *Nature (London)* **238** (1972) 28.
145. M. H. Lewis, R. J. Lumby, *Powder Metall.* **26** (1983) no. 2, 73.
146. K. H. Jack in E. P. Glasser, P. E. Potter (eds.): *Sialons and Related Nitrogen Ceramics, The Chemical Society,* Burlington House, London 1977, pp. 204 – 221.
147. M. H. Lewis, A. R. Bhatti, R. J. Lumby, B. North, *J. Mater. Sci.* **15** (1980) 103.
148. G. Schwier in H. Kißler (ed.): *Entwicklung und Herstellung verbesserter Alpha-Siliciumnitridpulver für das Heißpressen und Sintern,* Springer Verlag, Berlin 1984.
149. F. K. van Dijen, R. Metselaar, C. A. M. Siskens, *Sprechsaal* **117** (1984) 627.
150. K. S. Mazdiyasni, C. M. Cooke, *J. Am. Ceram. Soc.* **56** (1973) 628.
151. T. Yamada, T. Kawakito, T. Iwai, *J. Mater. Sci. Lett.* **2** (1983) 275.
152. S. Prochazka, *Am. Ceram. Soc. Bull.* **57** (1978) p. 579.
153. D. L. Segal, *Brit. Ceram. Trans. J.* **85** (1986) 184.
154. D. P. Stinton, T. M. Besmann, R. A. Lowden, *Am. Ceram. Soc. Bull.* **67** (1988) 350.
155. H. Inoue, K. Komeya, A. Truge, *Toshiba Rev.* **39** (1984) 488.
156. S.-C. Zhang, W. Cannon, *J. Am. Ceram. Soc.* **67** (1984) 691.
157. M. Hoch, K. M. Nair, *Am. Ceram. Soc. Bull.* **58** (1979) 187.
158. K. Kijima, S. Shirasaki, *J. Chem. Phys.* **65** (1976) 2668.
159. M. Mitomo, *J. Mater. Sci.* **11** (1976) 1103.
160. J. A. Mangels, *US-DOE-Report,* DEN 3-167, 1980.
161. A. Giachello, P. Popper, *Ceramurt. Int.* **5** (1979) 110.
162. J. A. Mangels, *Ceram. Eng. Sci. Proc.* **2** (1981) 589.
163. C. D. Greskovich, J. H. Rosolowski, *J. Am. Ceram. Soc.* **59** (1976) 336.
164. G. R. Terwilliger, F. F. Lange, *J. Mater. Sci.* **10** (1975) 1169.
165. M. Mitomo, M. Tsutsumi, E. Bannai, T. Tanaka, *Am. Ceram. Soc. Bull.* **55** (1976) 313.
166. W. Braue, H. J. Dudek, G. Ziegler in P. Vincenzini (ed.): *Analytical Investigations of Powder Particle Surfaces and Grain Boundaries of $Si_3N_4$by Means of Auger Electron and Photoelectron Spectroscopy,* Elsevier, Amsterdam 1983, p. 661.
167. H. F. Priest, G. L. Priest, G. E. Gazza, *J. Am. Ceram. Soc.* **60** (1977) 80 – 81.
168. P. Greil, G. Petzow in H. Kißler (ed.): *Gefügecharakterisierung von Si_3N_4-Mischkeramiken und Möglichkeiten der Verbesserung der Hochtemperatureigenschaften,* Springer Verlag, Berlin 1984.
169. G. Ziegler, G. Wötting, *Int. J. High Technol. Ceram.* **1** (1985) 31.
170. H. T. Larker: "On Hot Isostatic Pressing of Shaped Ceramic Parts" in S. Somiya, E. Kanai, K. Ando (eds.): *Proc. 1st Symp. on Ceramic Compounds for Engines,* KTK Scientific Publishers, Tokyo 1983.
171. J. Heinrich, M. Böhmer, *Powder Metall. Int.* **16** (1984) 233 – 235, 283.
172. M. H. van de Voorde, C. A. M. Siskens, W. Betteridge, *Sprechsaal* **116** (1983) 178.
173. *Adv. Mater. Proc.* **135** (1989) 29.
174. R. Cundill, *Fachber. Metallbearb.* **62** (1985) 52.
175. H. Knoch, D. Steinmann, *Metall (Berlin)* **43** (1989) no. 3, 205.
176. D. Steinmann, *Chemical Plants & Processing,* April 1988.
177. D. Steinmann, *Chemie-Anlagen Verfahren* **2** (1988) 61.
178. A. Lipp, D. Steinmann: Company brochure of Elektroschmelzwerk Kempten, FRG.
179. K. T. Ho, M. A. Nicolet, *Thin Solid Films* **127** (1985) 313.
180. K. Bean, P. Gleim, R. L. Yeakley, *J. Electrochem. Soc.* **114** (1967) 733.
181. A. C. Adams, *Solid State Technol.* **26** (1983) 135.
182. J. B. MacChesney, P. M. Bridenbaugh, P. B. O'Connor, *Mater. Res. Bull.* **5** (1970) 783.
183. R. D. Jones, K. Rose, *J. Phys. Chem. Solids* **48** (1987) 587.
184. M. Yamaguchi, *Jpn. J. Appl. Phys.* **19** (1980) L401.
185. Y. Hirota, M. Okamura, T. Kobayashi, *J. Appl. Phys.* **53** (1982) 536.
186. G. D. Bagratishvili, R. B. Dzhanelidze, I. G. Nakhutsrishvili, *Izv. Akad. Nauk Gruz. SSR Ser. Khim.* **12** (1986) no. 4, 272.
187. K. P. Pande, *Solid State Electron.* **25** (1982) no. 2, 145.

188. D. B. Alford, L. G. Meiners, *J. Electrochem. Soc.* **134** (1987) no. 4, 979.
189. Semiconductor Energy Laboratory Co., JP-Kokai 6 240 378, 1985 (T. Shiraishi, S. Yamazaki).
190. Ricoh Co., JP-Kokai 6 122 509, 1984 (Y. Kageyama, Y. Ide, T. Onuma).
191. Ricoh Co., DE-OS 3 438 437, 1983 (S. Nitta, Y. Ide).
192. T. Shiraishi, Y. Arai, S. Yamazaki, *J. Non Cryst. Solids* **77-78** (1985) 1313.
193. T. Kobayashi, Y. Hirota, *Electron. Lett.* **18** (1982) no. 4, 180.
194. M. Hoch in P. S. Rudman, J. Stringer, R. I. Jaffee (eds.): *Phase Stability in Metals and Alloys,* McGraw-Hill, New York 1967, p. 419.
195. K. Müller et al., *Neue Hütte* **30** (1985) 409.
196. P. Ettmayer, H. Priemer, R. Kieffer, *Metall (Berlin)* **23** (1969) 307.
197. W. Lengauer, P. Ettmayer, *Monatsh. Chem.* **117** (1980) 275.
198. A. N. Christensen, S. Fregerslev, *Acta Chem. Scand. Ser. A:* **A31** (1977) 861.
199. A. N. Christensen, *Acta Chem. Scand. Ser. A:* **A32** (1978) 89.
200. Y. Kumashiro et al., *J. Cryst. Growth* **52** (1981) 597.
201. A. Kato, N. Tamari, *J. Cryst. Growth* **49** (1980) 199.
202. R. Madar, G. Jacob, J. Hallais, R. Fruchart, *J. Cryst. Growth* **31** (1975) 197.
203. N. Tamari, A. Kato, *J. Cryst. Growth* **46** (1979) 221.
204. T. Wolf, H. Schneider, C. Politis: *Report No. 2881,* Kernforschungszentrum Karlsruhe 1979, p. 128.
205. N. Savvides, *Thin Solid Films* **163** (1988) 13.
206. N. Savvides, B. Window, *J. Vac. Sci. Technol. A* **4** (1986) 504.
207. N. Savvides, B. Window, *J. Appl. Phys.* **64** (1988) no. 1, 225.
208. B. O. Johansson, H. T. G. Hentzell, J. M. E. Harper, J. J. Cuomo, *J. Mater. Res.* **1** (1986) 442.
209. E. O. Ristolainen, J. M. Molarius, A. S. Korhonen, V. K. Lindroos, *J. Vac. Sci. Technol. A* **5** (1987) no. 4, 2184.
210. K. Salmenoja, A. S. Korhonen, M. S. Sulonen, *J. Vac. Sci. Technol. A* **3** (1985) 2364.
211. K. Salmenoja, A. S. Korhonen, *Vacuum* **36** (1986) 33.
212. N. Terao, *J. Less Common. Met.* **23** (1971) 159.
213. J. Strähle, *Z. Anorg. Allg. Chem.* **402** (1973) 47.
214. C. H. DeNovion, J. P. Landesmann, *Pure Appl. Chem.* **57** (1985) 1391.
215. W. Lengauer, P. Ettmayer, *High Temp.-High Press.* **22** (1990) 13.
216. G. Brauer, W. Kern, *Z. Anorg. Allg. Chem.* **507** (1983) 127.
217. J. Gatterer, G. Dufek, P. Ettmayer, R. Kieffer, *Monatsh. Chem.* **106** (1975) 1137.
218. G. Brauer, H. Kirner, *Z. Anorg. Allg. Chem.* **328** (1964) 34.
219. R. Kieffer, P. Ettmayer, M. Freudhofmeier, J. Gatterer, *Monatsh. Chem.* **102** (1971) 483.
220. G. V. Samsonov, I. M. Vinitskii: *Handbook of Refractory Compounds,* IFI/Plenum, New York 1980.
221. H. J. Goldschmidt: *Interstitial Alloys,* Butterworths, London 1967.
222. E. K. Storms in D. W. A. Sharp (ed.): *M.T.P. International Review of Science, Inorganic Chemistry Series One,* vol. 5, Butterworths, London 1972.
223. Th. Wolf, Dissertation, Karlsruhe 1981.
224. W. Lengauer, P. Ettmayer, *Mater Sci. Eng.* **A105/106** (1988) 257.
225. C. R. Houska, *J. Phys. Chem. Solids* **25** (1964) 359.
226. W. Lengauer, P. Ettmayer, *Monatsh. Chem.* **117** (1986) 713.
227. W. Lengauer, *Surface and Interface Analysis* **5** (1990) 377.
228. W. Lengauer, *J. Phys. Chem. Solids* **49** (1988) 59.
229. C. Politis, G. Rejman: "Progress Report," *Kernforschungszent. Karlsruhe, KfK-Ext.* 6/78-1, July 1978.
230. R. Kieffer, P. Ettmayer, T. Dubsky, *Z. Metallk. (Berlin)* **58** (1967) 560.
231. P. Ettmayer, R. Kieffer, F. Hattinger, *Metall (Berlin)* **28** (1974) 1151.
232. L. E. Toth, C. P. Wang, C. M. Yen, *Acta Metall.* **14** (1966) 1403.
233. P. Ettmayer, *Monatsh. Chem.* **101** (1970) 127.
234. A. Bezinge et al., *Solid State Commun.* **63** (1987) 141.
235. A. Matthews, A. R. Lefkow, *Thin Solid Films* **126** (1985) 283.
236. J. R. Peterson, *J. Vac. Sci. Technol.* **11** (1974) 715.
237. W. Schintlmeister et al., *Powder Metall. Int.* **13** (1981) 71.
238. N. Lindström, R. T. Johannesson, *J. Electrochem. Soc.* **123** (1976) 555.
239. A. J. Perry, *Thin Solid Films* **146** (1987) 165.
240. A. J. Perry, M. Georgson, W. D. Sproul, *Thin Solid Films* **157** (1988) 255 – 265.

241. D. T. Quinto, *J. Vac. Sci. Technol. A* **6** (1988) no. 3, 2149.
242. D. T. Quinto, A. T. Santhanam, P. C. Jindal, *Refract. Hard Met. Int. J.* **6** (1989) 95.
243. V. Valvoda, R. Kuzel, R. Cerny, J. Musil, *Thin Solid Films* **156** (1988) 53.
244. R. Chatterjee-Fischer, *Hermsdorfer Tech. Mitt.* **38** (1983) 1.
245. G. Spur (ed.): *Handbuch der Fertigungstechnik,* **vol. 4/2,** Carl Hanser Verlag, München-Wien.
246. Z. Morita, T. Tanaka, T. Yanai, *Metall. Trans. B* **18 B** (1987) 195 – 202.
247. K. T. Rie, S. Eisenberg, *Hermsdorfer Tech. Mitt.* **42** (1987) 344.
248. K. T. Rie, T. Lampe, S. Eisenberg, *Surf. Eng.* **1** (1985) 198.
249. H. R. Eschnauer, *Metall (Berlin)* **34** (1980) 232.
250. N. J. Archer, *Phys. Technol.* **10** (1979) 152.
251. R. Kieffer, P. Ettmayer, M. Freudhofmeier, *Metall (Berlin)* **25** (1971) 1335.
252. P. Ettmayer, W. Lengauer, *Powder Metall. Int.* **21** (1989) 37.
253. P. Ettmayer, *Annu. Rev. Mater. Sci.* **19** (1989)145.
254. K. Nishigaki, H. Doi, *Jap. Soc. Powd. Powd. Met.* **27** (1980) 39.
255. H. Doi, *Conf. Ser. Inst. Phys.* **75** (1986) 489.
256. P. Ettmayer, H. Kolaska: *Proc. 12th Int. Plansee Seminar,* **vol. 2,** Reutte 1989, p. 771.
257. E. Rudy, *J. Less Common Met.* **33** (1973) 43.
258. B. Roebuck, M. G. Gee: *Proc. 12th Int. Plansee Seminar,* **vol. 2,** Reutte 1989, p. 1.
259. H. Norström, T. Donchev, M. Östling, C. S. Petersson, *Physica Scripta* **28** (1983) 633 – 636.
260. W. J. Garceau, P. R. Fournier, G. K. Herb, *Thin Solid Films* **60** (1979) 237.
261. S. Kanamori, *Thin Solid Films* **136** (1986) 195.
262. M. Ostling et al., *J. Vac. Sci. Technol. A* **2** (1984) 281.
263. W. Sinke, G. P. A. Frijlink, F. W. Saris, *Appl. Phys. Lett.* **47** (1985) 471.
264. I. Suni, M. Mäenpää, M. A. Nicolet, M. Luomajärvi, *J. Electrochem. Soc.* **130** (1983) 1215.
265. M. Wittmer, *Appl. Phys. Lett.* **36** (1980) no. 6, 456 – 458.
266. S. Ashok, M. Crookshank, M. Chu. G. Sellers, *J. Vac. Sci. Technol. A* **6** (1988) 1599.
267. M. A. Nicolet, *Thin Solid Films* **52** (1978) 415.
268. J. R. Shappirio, *Solid State Technol.* **28** (1985) 161.
269. V. Hoffman, *Solid State Technol.* **26** (1983) 119.
270. M. A. Nicolet, M. Bartur, *J. Vac. Sci. Technol.* **19** (1981) 786.
271. H. von Seefeld, N. W. Cheung, M. Mäenpää, M. A. Nicolet, *IEEE Trans. Electron Devices ED* **27** (1980) 873.
272. M. B. Chamberlain, *Thin Solid Films* **91** (1982) 155.
273. M. F. Zhu, A. H. Hamdi, M. A. Nicolet, J. L. Tandon, *Thin Solid Films* **119** (1984) 5.
274. J. R. Waldrop, *Appl. Phys. Lett.* **43** (1983) 87.
275. R. Beyers, R. Sinclair, M. E. Thomas, *J. Vac. Sci. Technol. B* **2** (1984) 781.
276. M. A. Nicolet, M. Bartur, *J. Vac. Sci. Technol.* **19** (1981) 786.
277. N. Kumar et al., *Thin Solid Films* **153** (1987) 287.
278. NEC Corp., JP-Kokai 61 230 398, 1985 (Y. Shimada, Y. Kurokawa, K. Uchiumi, H. Takamizawa).
279. B. Karlsson, C. G. Ribbing, *Proc. SPIE-Int. Soc. Opt. Eng.* **324** (1982) 52.
280. M. A. Farooq, S. P. Murarka, C. C. Chang, F. A. Baiocchi, *J. Appl. Phys.* **65** (1989) 3017.
281. E. Jiang, C. Sun, J. Li, Y. Liu, *J. Appl. Phys.* **65** (1989) 1659.
282. Sumitomo Special Metals Co., JP-Kokai 58 36 908, 1981.
283. Nippon Telegraph and Telephone Public Corp. JP-Kokai 63 86 111, 1986 (S. Matsuoka, K. Ono, K. Hoshino).
284. E. P. Wohlfarth, *Pure Appl. Chem.* **57** (1985) 1403.
285. H. A. Wriedt, N. A. Gokcen, R. H. Nafziger, *Bull. Alloy Phase Diagrams* **8** (1987) 355.
286. M. Matsuoka, K. Ono, T. Inukai, *Appl. Phys. Lett.* **49** (1986) 977.
287. Nippon Telegraph and Telephone Public Corp., JP 63 86 111, 1988 (S. Matsuoka, K. Ono, K. Hoshino).
288. N. Pessall, R. E. Gold, H. A. Johansen, *J. Phys. Chem. Solids* **29** (1986) 19.
289. K. Schwarz et al., *Phys. Rev. B Condens. Matter* **32** (1985) 8312.
290. V. M. Pan et al., *IEEE Trans. Magn.* **23** (1987) 1436.
291. K. Schwarz, D. S. Yee, J. J. Cuomo, J. M. E. Harper, *Phys. Rev. B Condens. Matter* **32** (1985) 5489.
292. A. Shoji in H. O. Hahlbohm, H. Luebbig (eds.): *SQUID '85 Proc. Int. Conf. 3rd,* de Gruyter, Berlin 1985, p. 631.
293. M. Dietrich, B. Milich, L. Paterok, B. Runtsch, *KFK Nachr.* **16** (1984) no. 4, 183 – 187.

294. K. Kawaguchi, S. Shin, *Mater. Res. Symp. Proc.* **103** (1988) 281.
295. H. Ihara et al., *Adv. Cryog. Eng.* **32** (1986) 603.
296. G. Linker et al., *J. Phys. F.* **16** (1986) 2167.
297. H. Yamamoto, T. Miki, M. Tanaka, *Adv. Cryog. Eng.* **32** (1986) 671.
298. N. Terada, M. Naoe, Y. Hoshi, *Adv. Cryog. Eng.* **32** (1986) 663.
299. N. Savvides, *J. Appl. Phys.* **62** (1987) 600.
300. K. Saito, K. Yabe, O. Nishimura, *Nucl. Instrum. Methods Phys. Res. Sect. B* **39** (1989) 623.
301. N. Kobayashi, H. Tanoue, G. Linker, *Nucl. Instrum. Methods Phys. Res. Sect. B* **33** (1988) 795.
302. T. J. Ellis, *Energy Res. Abstr.* **12** (1987).
303. D. W. Capone II, K. E. Gray, R. T. Kampwirth, H. L. Lo, *J. Nucl. Mater.* **141 – 143** (1986) 73.
304. O. Knotek, W. D. Münz, T. Leyendecker, *J. Vac. Sci. Technol. A* **5** (1987) 2173.
305. O. Knotek, T. Leyendecker, F. Jungblut, *Thin Solid Films* **153** (1987) 83.
306. H. Freller, H. Haessler, *Thin Solid Films* **153** (1987) 67.
307. O. Knotek, M. Atzor, F. Jungblut, H. G. Prengel: *Proc. 12th Int. Plansee Seminar,* **vol. 3,** Reutte 1989, p. 49.
308. P. Rödhammer, *Proc. 12th Int. Plansee Seminar,* **vol. 3,** Reutte 1989, p. 661.
309. U. Helmersson et al., *J. Appl. Phys.* **62** (1987) 481.

Nitriles

Acrylonitrile is a separate keyword; benzonitrile → Benzoic Acid; malononitrile → Malonic Acid and Derivatives

PETER POLLAK, Lonza AG, Basel, Switzerland (Chap. 1, 2)

GÉRARD ROMEDER, Lonza AG, Basel, Switzerland (Chap. 1, 2)

FERDINAND HAGEDORN, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 3.1 – 3.3)

HEINZ-PETER GELBKE, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 3.4)

1. Introduction 1
2. Aliphatic Nitriles 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.3. General Production Processes . . 2
2.4. Selected Aliphatic Nitriles 4
2.4.1. Saturated Mono- and Dinitriles . . . 4
2.4.2. Unsaturated Mono- and Dinitriles . 5
2.4.3. Substituted Nitriles 6
2.4.4. β-Iminonitriles 8
2.4.5. β-Ketonitriles 8
2.5. Toxicology and Occupational Health 8
3. Aromatic and Araliphatic Nitriles 9
3.1. Properties 9
3.2. General Production Methods . . . 9
3.3. Selected Araliphatic and Aromatic Nitriles 10
3.3.1. Araliphatic Nitriles 10
3.3.2. Aromatic Nitriles 11
3.3.2.1. Cyanobenzenes 11
3.3.2.2. Cyanonaphthalenes (Naphthalene Carboxylic Acid Nitriles) 12
3.4. Toxicology 13
4. References 13

1. Introduction

Organic compounds containing the – CN group are generically called nitriles. The term "nitrile" refers to the triply bound nitrogen atom, ≡N, and not to the carbon atom attached to it; therefore numbering of the aliphatic chain starts with that carbon atom (e.g., 4-chlorobutyronitrile describes the chemical formula $ClCH_2CH_2CH_2CN$). Nitriles R – CN are usually viewed as derivatives of the corresponding acids R – COOH. Consequently, they are also named by changing the ending "ic acid" or "oic acid" to "onitrile" (e.g., glutaric acid → glutaronitrile; dodecanoic acid → dodecanonitrile, propiononitrile, as an exception, is commonly referred to as propionitrile). Another way of naming nitriles is to state the name of the group R and use the prefix "cyano" or the suffix "cyanide" (e.g., cyanoethane and ethyl cyanide both describe H_3CCH_2CN). Nitriles that are obtained by addition of hydrogen cyanide to ketones or aldehydes and thus have both a hydroxy and a cyano group attached to the same carbon atom are commonly termed cyanohydrins (e.g., acetone cyanohydrin is identical to 2-hydroxy-2-methylpropionitrile). The same nomenclature applies to 3-hydroxypropionitrile which is commonly named ethylene cyanohydrin.

2. Aliphatic Nitriles

Aliphatic nitriles are important starting materials for polymers as well as for the synthesis of e.g., pharmaceuticals and pesticides.

2.1. Physical Properties

The carbon – nitrogen bond is extremely polar, which results in nitriles having a high dipole moment. This large dipole moment leads to intramolecular association; hence, nitriles have higher boiling points than would be expected from their molecular mass. Most of the lower molecular mass aliphatic nitriles (up to $C_{13}H_{27}CN$) are liquids at room temperature.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_363

Table 1. Physical properties

Nitrile	M_r	*mp*, °C	*bp* (101.3 kPa), °C	n_D^{20}	d_4^{20}
Acetonitrile, CH_3CN	41.05	− 45	81.6	1.3441	0.7138
Propionitrile, CH_3CH_2CN	55.08	− 93	97.2	1.3670	0.782
Butyronitrile, $CH_3CH_2CH_2CN$	69.10	−111.9	116 – 117	1.3838	0.7936
Valeronitrile, $CH_3CH_2CH_2CH_2CN$	83.13	− 96	141.3	1.3971	0.8008
Tetradecanonitrile, $CH_3(CH_2)_{12}CN$	209.38	19.25	226 (10 kPa)	1.4392 [a]	0.8281 [d]
Malononitrile, $NCCH_2CN$	66.06	30 – 31	218 – 219	1.4146 [b]	1.0494 [e]
Succinonitrile, $NCCH_2CH_2CN$	80.09	57	265 – 267	1.4173 [c]	0.9867 [c]
Glutaronitrile, $NC(CH_2)_3CN$	94.12	− 29	286	1.4295	0.9911 [f]
Allyl cyanide, $CH_2{=}CH-CH_2CN$	67.09	− 84	119	1.4060	0.8329
Methyleneglutaronitrile, $NCC_2H_4C(=CH_2)CN$	106.12	− 9.0	113 (0.66 kPa)	1.4558	0.9831 [f]

Measuring temperature: [a] 25 °C; [b] 34 °C; [c] 60 °C; [d] 19 °C; [e] 35 °C; [f] 15 °C.

Simple nitriles such as acetonitrile, propionitrile, glycolonitrile, and malononitrile are miscible with water; the latter two having a higher solubility because of the presence of two polar groups. Nitriles with higher molecular mass are sparingly water-soluble. Nitriles are good solvents for both polar and nonpolar solutes. The physical properties of selected nitriles are listed in Table 1.

2.2. Chemical Properties

The chemical properties of aliphatic nitriles are determined by two reactive centers:

1) The electrophilic cyano group which can react with nucleophiles
2) The activated carbon adjacent to the nitrile group which is prone to base-catalyzed substitution reactions

Typical reactions involving the cyano group and the activated carbon adjacent to it are illustrated in Figures 1 and 2, respectively.

2.3. General Production Processes

Of the many synthetic pathways that can be envisaged for producing aliphatic nitriles, the following four are widely used:

Reaction of Nitrogen-Free Precursors (such as Alkanes, Olefins, Alcohols, Aldehydes, or Acids) with Ammonia. Gas phase reaction of *olefins with ammonia* in the presence of oxygen (ammoxidation) and oxidation catalysts (based on, e.g., vanadium or molybdenum) has attained the greatest industrial importance for the production of acrylonitrile from propene and methacrylonitrile from isobutene. This process is known as the *Sohio process* → Acrylonitrile.

$$H_2C{=}CH-CH_3 + NH_3 + 3/2\,O_2 \xrightarrow{\text{Catalyst}} H_2C{=}CH-CN + 3\,H_2O$$

Acetonitrile and hydrogen cyanide are formed as byproducts in typical quantities of 30 – 40 kg and 140 – 180 kg, respectively per 1000 kilograms of acrylonitrile [1]. Propane can also be used as feedstock (the first step being the in situ dehydrogenation to propene) as well as acrolein [2]. Similarly, the simplest aliphatic nitrile, hydrogen cyanide can be obtained from methane, ammonia, and oxygen (*Andrussow process,* see → Cyano Compounds; Inorganic, Chap. 1.2.1.).

The ammoxidation process can also be advantageously carried out with *aldehydes* as starting materials, when they are readily available. This is especially the case for butyraldehyde and isobutyraldehyde which yield butyronitrile and isobutyronitrile, respectively [3]. The ammonolysis of *alcohols* or *aldehydes* can also be conducted under dehydrogenation conditions [4]:

$$(H_3C)_2CH-CH_2OH + NH_3 \xrightarrow[290\,^\circ C]{Cu/SiO_2} (H_3C)_2CHCN + H_2O + 2\,H_2$$

Acids when reacted with ammonia over a dehydration catalyst such as Al_2O_3 at typical temperatures above 360 °C give nitriles without isolation of the intermediary amides [5].

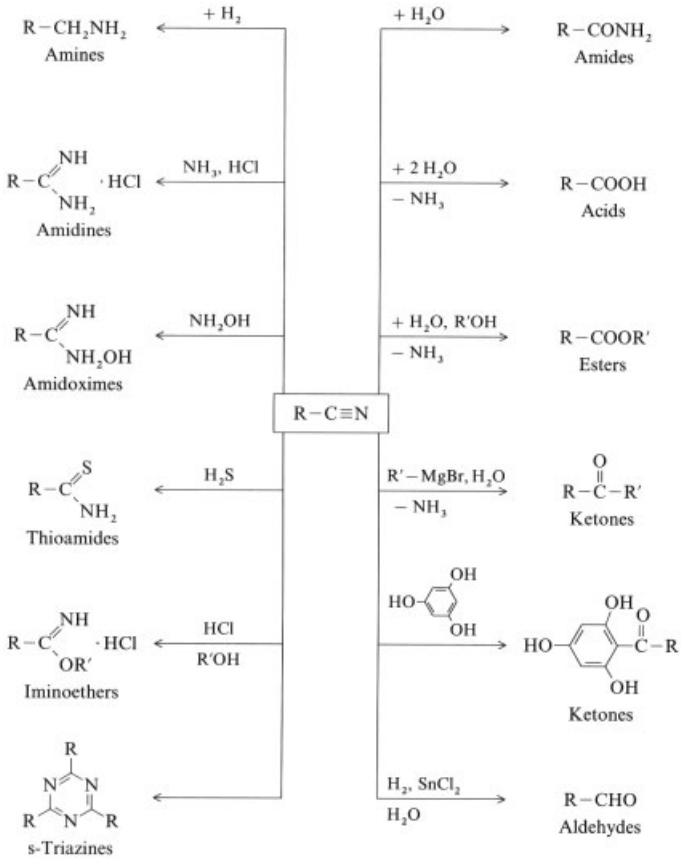

Figure 1. Typical reactions involving the cyano group

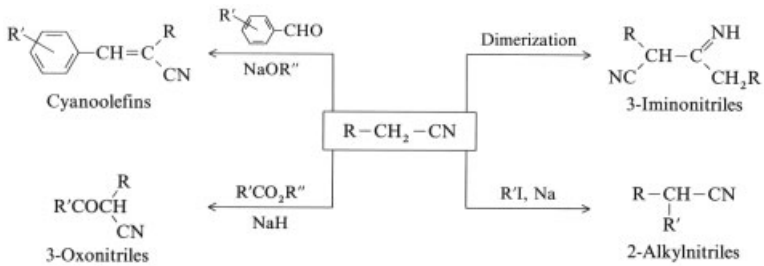

Figure 2. Typical reactions involving the activated carbon atom adjacent to the cyano group

$$(H_3C)_3CCOOH + NH_3 \xrightarrow[380\,^\circ C]{Al_2O_3} (H_3C)_3CCN + 2\,H_2O$$

In addition to ammonia other nitrogen-containing reagents can be used, for example, hydroxylamine which easily yields nitriles upon reaction with aldehydes.

$$R-CHO + NH_2OH \xrightarrow{-H_2O} [R-CH=N-OH] \xrightarrow{-H_2O} R-CN$$

A multitude of dehydration agents is available.

Other nitrogen-containing reagents such as sulfonamides [6], hydroxylamine-*O*-sulfonic acid [7] and chlorosulfonylisocyanate [8] have also been investigated. However, none of those methods have gained industrial importance.

Formation of the Cyano Moiety from Nitrogen-Containing Precursors such as Amines, Amides, or Formamides. If an amine

is employed as starting material, it is usually produced in situ from the corresponding alcohol and ammonia, the dehydrogenation step is then performed catalytically [4]. An electrochemical dehydrogenation with a nickel hydroxide anode has also been investigated [9].

$$R-CH_2-NH_2 \xrightarrow[KOH,\ H_2O]{Ni(OH)_2/Anode,} R-CN + 2H_2$$

Dehydration of *amides* with agents such as aluminum chloride or thionyl chloride is a widely applicable method, although not competitive in most cases. The direct catalytic conversion of acids with ammonia is preferred.

Formamides are also valuable precursors [10]:

$$HCONHC(CH_3)_3 \xrightarrow[t > 250\,°C]{\text{Acetic anhydride}} (H_3C)_3CCN + H_2O$$

Reaction of Hydrogen Cyanide or Cyanide Ions with Double Bonds, Carbonyl Compounds, Hydrogen or Halogens. A typical industrial example of the hydrogen cyanide addition to *double bonds* is the direct hydrocyanation of butadiene to adiponitrile [11].

Reaction with aldehydes and ketones yields the corresponding cyanohydrins.

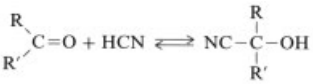

The reaction is usually carried out at 10–15 °C and a pH of 6.5–7.5. The cyanohydrin is continuously removed from the reaction mixture to shift the equilibrium toward cyanohydrin formation, which is usually stabilized with mineral acids or directly chemically converted to a more stable form [12].

A displacement reaction of hydrogen with a halogen cyanide, which has gained great industrial importance, is the synthesis of malononitrile from acetonitrile and cyanogen chloride above 700 °C (→ Malonic Acid and Derivatives):

$$H_3C-CN + ClCN \longrightarrow NC-CH_2-CN + HCl$$

The order of halogen reactivity towards cyanide displacement is Cl < Br < I, fluoride being essentially inert. Examples are the production of:

4-Chlorobutyronitrile from 1-bromo-3-chloropropane [13]

Adiponitrile via prior chlorination of butadiene [14]

Cyanoacetates from chloroacetic acid (→ Malonic Acid and Derivatives, Chap. 4.)

Another important reaction is the synthesis of acyl cyanides in the presence of copper(I) cyanide [15]:

$$H_3C-\overset{O}{\overset{\|}{C}}-Cl + NaCN \xrightarrow{CuCN} H_3C-\overset{O}{\overset{\|}{C}}-CN + NaCl$$

In addition to the readily available hydrogen cyanide or cyanide salts (e.g., sodium cyanide) other cyanation agents such as tetraalkylammonium cyanides [16] or trimethylsilylcyanide [17] have been investigated.

Reaction of the Activated Carbon in Nitrile-Containing Precursors. In addition to the examples highlighted in Figure 2 the electrohydrodimerization of acrylonitrile to adiponitrile, known as "EHD process" is also significant [18]:

$$2H_2C=CH-CN \xrightarrow[-2OH]{2e,\ H_2O} NC-(CH_2)_4-CN$$

2.4. Selected Aliphatic Nitriles

2.4.1. Saturated Mono- and Dinitriles

Acetonitrile [*75-05-8*], methylcyanide, cyanomethane, ethanenitrile, CH_3CN, is a colorless liquid miscible with water, ethanol and many organic solvents but immiscible with many saturated hydrocarbons (e.g., petroleum fractions). The physical properties are listed in Table 1. Characteristics of acetonitrile are its high dipole moment (3.84 D) and dielectric constant (38.8 at 20 °C).

Production. Acetonitrile and hydrogen cyanide are the principal byproducts from the ammoxidation of propylene to acrylonitrile (Sohio process). Some acrylonitrile producers recover and purify acetonitrile, but most companies burn the byproducts as plant fuel.

Economic Aspects. In 1986 the comsumption of acetonitrile amounted to ca. 13 000 t in the United States, Western Europe, and Japan [19]. Along with *N*-methylpyrrolidone and *N,N*-dimethylformamide, acetonitrile represents one of the major solvents for extraction of butadiene from crude C_4 streams. Acetonitrile is also a valuable reaction solvent, especially in the synthesis of antibiotics. Highly purified acetonitrile finds application as a solvent in high-performance liquid chromatography.

The major use of acetonitrile in *organic synthesis* is in the production of malononitrile through reaction with cyanogen chloride (→ Malonic Acid and Derivatives).

Other applications include the synthesis of trimethyl orthoacetate [$H_3C-C(OCH_3)_3$] and acetamidine hydrochloride [$H_3C-C(=NH)-NH_2 \cdot HCl$]. The former is used in the synthesis of 3-(2,2-dichlorovinyl)-2,2-dimethylcyclopropanecarboxylic acid (also known as DV-acid) which is a key precursor for pyrethroides [20]. The latter yields N-containing heterocycles such as 4-amino-5-cyano-2-methylpyrimidine used in the synthesis of thiamine [21].

Transport Classification: RID/ADR 3, no. 11B

Propionitrile [*107-12-0*], ethyl cyanide, CH_3CH_2CN, is a colorless liquid which is miscible with water, ethanol, dimethylformamide, and diethyl ether. The physical properties are listed in Table 1. Propionitrile is obtained either by hydrogenation of acrylonitrile [22] or by the gas-phase reaction of propanal or propanol with ammonia [4]. It is also obtained as a byproduct in the electrohydrodimerization of acrylonitrile to adiponitrile (EHD process) [18].

Hydrogenation of propionitrile yields *N*-propylamines [23]. It is also used as an organic intermediate, for example, in the synthesis of the pharmaceuticals flopropione (via Houben–Hoesch reaction) [24] and ketoprofen [25].

Transport Classification: RID/ADR 3, no. 11B

Butyronitrile [*109-74-0*], propyl cyanide, butanenitrile, $CH_3(CH_2)_2CN$, is a colorless liquid slightly miscible with water, miscible with ethanol and diethyl ether. The physical properties are listed in Table 1. Butyronitrile is usually obtained by the catalytic gase-phase reaction of butanol or butyraldehyde with ammonia [3,4]. Its major use is the manufacture of the poultry drug amprolium [26].

Transport Classification: RID/ADR 3, no. 11B

Isobutyronitrile [*78-82-0*], 2-methylpropanenitrile, isopropyl cyanide, 2-methylpropionitrile, $(CH_3)_2CHCN$, M_r 69.11, *mp* −71.5 °C, *bp* 117 °C (101.3 kPa), n_D^{20} 1.374, d_4^{20} 0.770, is a colorless liquid miscible with ethanol and diethyl ether, slightly miscible with water.

Isobutyronitrile is usually obtained by the catalytic gas-phase reaction of isobutyraldehyde or isobutanol with ammonia [3,4]. Its major use is the synthesis of the insecticide diazinon [27].

Transport Classification: RID/ADR 3, no. 11B

Succinonitrile [*110-61-2*], ethylene dicyanide, ethylene cyanide, dicyanoethane, butanedinitrile, $NCCH_2CH_2CN$, is a colorless, waxy solid slightly soluble in water and ethanol. The physical properties are listed in Table 1. Succinonitrile is obtained by addition of hydrogen cyanide to acrylonitrile [28]. Subsequent hydrogenation of succinonitrile yields 1,4-diaminobutane, which reacts with adipic acid to form the new polyamide 46. The polymer, developed by DSM (Netherlands), is marketed under the trade name Stanyl; the latter is reported to have excellent mechanical properties at high temperature [29].

Transport Classification: RID/ADR 6.1, no. 12C

2.4.2. Unsaturated Mono- and Dinitriles

Methacrylonitrile [*126-98-7*], 2-methyl-2-propenenitrile, $CH_2{=}C(CH_3)CN$, M_r 67.09, *mp* −35.8 °C, *bp* 90.3 °C, d_4^{20} 0.7998, n_D^{20} 1.4003, is a liquid which is miscible with ethanol, diethyl ether and acetone; it is immiscible with water.

Methacrylonitrile can be produced by ammoxidation of isobutene. Its copolymerization with methacrylic acid gives poly(methacrylimide), an engineering plastic commercialized under the trade name Rohacell by Röhm (Federal Republic of Germany) [30].

2-Methyleneglutaronitrile [*1572-52-7*], acrylonitrile dimer, 2,4-dicyano-1-butene, 2-methylenepentanedinitrile, NCC_2H_4C (=CH_2) CN, is a colorless liquid slightly miscible with water, ethanol, acetone and benzene, immiscible with hexane. The physical properties are listed in Table 1.

2-Methyleneglutaronitrile is obtained by dimerization of acrylonitrile in the presence of a metal halide and a tertiary amine [31]. Its hydrogenation yields 2-methylglutaronitrile, which in turn can be converted into 3-methylpyridine [32].

Tetracyanoethylene [*670-54-2*], ethenetetracarbonitrile, TCNE, $(NC)_2C=C(CN)_2$, M_r 128.09, *mp* 197 – 198 °C, *bp* 223 °C, n_D^{25} 1.560. The preferred synthetic preparation of TCNE involves the debromination of the KBr complex of dibromomalononitrile. Tetracyanoethylene is a reactive compound that undergoes a variety of reactions including addition, replacement and cyclization. The chemistry of TCNE is exhaustively reviewed in [33].

2.4.3. Substituted Nitriles

Chloroacetonitrile [*107-14-2*], chloromethanenitrile, chloromethyl cyanide, $ClCH_2CN$, M_r 75.50, *bp* 124 – 126 °C, d_4^{20} 1.1896, n_D^{20} 1.426, is a colorless liquid with pungent odor. It is miscible with hydrocarbons and ethanol; immiscible with water. Chloroacetonitrile can be selectively obtained by the photochemical chlorination of acetonitrile with chlorine in the presence of, for example, $SnCl_4$ [34]. Another method is based on the dehydration of chloroacetamide with, e.g., phosphorous pentoxide [35]. Chloroacetonitrile is used as an organic intermediate, e.g., in the synthesis of the cardiovascular drug guanethidine [36] and the insecticide fenoxycarb [37].

Transport Classification : RID/ADR 6.1, no. 11B

Trichloroacetonitrile [*545-06-2*], trichloroethanenitrile, trichloromethyl cyanide, CCl_3CN, M_r 144.40, *mp* – 42 °C, *bp* 85.7 °C (101.3 kPa), d_4^{25} 1.4403, n_D^{20} 1.4409, is a colorless liquid immiscible with water.

Trichloroacetonitrile can be obtained by dehydration of trichloroacetamide with phosphorous pentoxide [38] or by chlorination of acetonitrile with chlorine. Vapor phase chlorination in the presence of water [39] and photochemical chlorination in the presence of catalysts such as $HgCl_2$ or $AlCl_3$ have been reported [40]. Trichloroacetonitrile is an organic intermediate used, for example, in the synthesis of the fungicide etridiazole [41].

Transport Classification : RID/ADR 6.1, no. 11B

Glycolonitrile [*107-16-4*], hydroxyacetonitrile, formaldehyde cyanohydrin, $HOCH_2CN$, M_r 57.05, *bp* 183 °C (slight decomp.), d_4^{15} 1.104, n_D^{20} 1.4117. Anhydrous glycolonitrile is a colorless liquid which is miscible with water, ethanol and diethyl ether.

Glycolonitrile is obtained by reaction of aqueous formaldehyde and aqueous hydrogen cyanide, in the presence of catalysts such as sodium hydroxide [42]. It is usually handled as an aqueous solution. Anhydrous glycolonitrile tends to be unstable at high pH values and therefore must be stabilized by addition of small amounts of, e.g., phosphoric acid.

The bulk of the glycolonitrile produced is not isolated but further upgraded by reaction with amines. Depending on the molecular ratio, reaction with ammonia yields aminoacetonitrile, iminodiacetonitrile or nitrilotriacetonitrile. Reaction with ethylenediamine gives ethylenediaminetetraacetonitrile which, in turn, is hydrolyzed to ethylenediaminetetraacetic acid (EDTA), a well known sequestrant.

Aminoacetonitrile [*540-61-4*], glycinonitrile, cyanomethylamine, H_2NCH_2CN, M_r 56.07, *bp* 58 °C (20 kPa, partial decomp.).

Aminoacetonitrile is readily obtained as an aqueous solution by reacting an aqueous solution of glycolonitrile with liquid ammonia [43]. The solution is stabilized by lowering the pH with small amounts of sulfuric acid [44]. Aminoacetonitrile can also be isolated as the hydrochloride or sulfate.

Aminoacetonitrile is usually further hydrolyzed by sodium hydroxide to the sodium salt of glycine, which in turn is neutralized to glycine [45].

Lactonitrile [*78-97-7*], acetaldehyde cyanohydrin, 2-hydroxypropanenitrile, $H_3CCH(OH)CN$, M_r 71.08, *mp* −40 °C, *bp* 182–184 °C (decomp.), d_4^{20} 0.9877, n_D^{18} 1.4058, is a colorless or straw-colored liquid, which is miscible with water and ethanol; it is miscible with diethyl ether and carbon disulfide.

Lactonitrile is obtained by reaction of acetaldehyde with hydrogen cyanide in the presence of sodium hydroxide. It is mainly used in the synthesis of lactic acid and lactates through hydrolysis with sulfuric acid and esterification [46].

3-Aminopropionitrile, [*151-18-8*], 3-aminopropanenitrile, β-aminopropionitrile, $H_2N-CH_2-CH_2CN$, M_r 70.09, *bp* 185 °C (101.3 kPa), n_D^{20} 1.4396, is a colorless liquid which tends to slowly polymerize if stored in the presence of air. 3-Aminopropionitrile is prepared by the reaction of acrylonitrile with ammonia [47] and is mainly used as an intermediate in the manufacture of β-alanine and pantothenic acid [48].

3-Chloropropionitrile [*542-76-7*], 3-chloropropanenitrile, β-chloropropionitrile, $ClCH_2CH_2CN$, M_r 89.53, *mp* −51 °C, *bp* 175–176 °C (101.3 kPa), n_D^{20} 1.4360, d_4^{20} 1.1573, is a colorless liquid miscible with acetone, benzene, ethanol and diethyl ether. It can be obtained by hydrochlorination of acrylonitrile with hydrochloric acid [49] and is used as an organic intermediate in the synthesis of the H_2-receptor famotidine [50].

Transport Classification : RID/ADR 6.1, no. 11C

3-Hydroxypropionitrile [*109-78-4*], ethylene cyanohydrin, hydracrylonitrile, 3-hydroxypropanenitrile, $HOCH_2CH_2CN$, M_r 71.08, *mp* −46 °C, *bp* 228 °C (101.3 kPa, slight decomp.), d_4^{25} 1.0404, n_D^{20} 1.4240, is a straw-colored liquid, miscible with water, acetone, and ethanol; immiscible with benzene, carbon tetrachloride, and naphtha. 3-Hydroxypropionitrile is obtained by reacting ethylene oxide with hydrogen cyanide or from the hydration of acrylonitrile [51]. It was used as feedstock in the production of acrylic acid and acrylates.

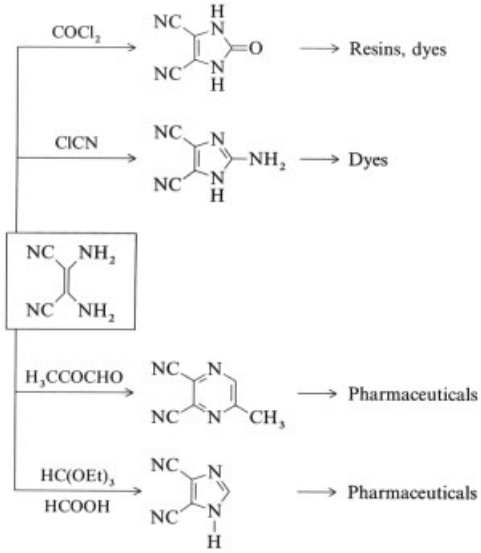

4-Chlorobutyronitrile [*628-20-6*], 4-chlorobutanenitrile, $Cl(CH_2)_3CN$, M_r 103.55, *bp* 189–191 °C (101.3 kPa), d_4^{15} 1.0934, n_D^{20} 1.4413, is a liquid, miscible with ethanol and diethyl ether; it is immiscible with water.

4-Chlorobutyronitrile is readily obtained by reacting 1-bromo-3-chloropropane with sodium cyanide [13]. It is used in the synthesis of the pharmaceuticals buflomedil [52] and buspirone [53]. 4-Chlorobutyronitrile also yields cyclopropanecarbonitrile upon cyclization with base [54].

Diaminomaleonitrile [*1187-42-4*], HCN tetramer, "DAMN", 2,3-diamino-2-butenedinitrile, $NC(NH_2)C{=}C(NH_2)CN$, M_r 108.10, *mp* 183 °C (decomp.), n_D^{20} 1.5898, forms pale-brown needle shaped crystals. They are soluble in methanol, ethanol, and *N,N*-dimethylformamide; slightly soluble in water and dioxane; and insoluble in acetone, benzene, xylene, and carbon disulfide.

Diaminomaleonitrile is obtained by tetramerization of hydrocyanic acid in the presence of catalysts such as alkylaluminum compounds [55]. Nippon Soda runs a 200 t/a plant in Japan and intends to expand the production capacity in the future [56]. Diaminomaleonitrile is a versatile starting material for synthesizing nitrogen-containing heterocycles [57].

2.4.4. β-Iminonitriles

2-Amino-1-propene-1,1,3-tricarbonitrile [*868-54-2*], 2-amino-1,1,3-tricyanopropene, malononitrile dimer, $(NC)_2C{=}C(NH_2)CH_2CN$, M_r 132.13, *mp* 171 – 173 °C, is a light brownish powder, insoluble in water.

Malononitrile dimer is obtained from the base-catalyzed dimerization of malononitrile [58]. It has been investigated as a potential precursor for new dyes but it has no current industrial significance [59].

3-Amino-3-methylacrylonitrile [*1118-61-2*], β-iminobutyronitrile, β-aminocrotononitrile, diacetonitrile, 3-amino-2-butenenitrile, $NCCH{=}C(NH_2)CH_3$, M_r 82.11, *mp* (cis-form) 79 – 83 °C, *mp* (trans-form) 52 – 53 °C, d_4^{77} 0.9519, is a yellowish solid soluble in ethanol and diethyl ether; it is sparingly soluble in benzene and water.

3-Amino-3-methylacrylonitrile is obtained by the base-catalyzed dimerization of acetonitrile [60]. It is mainly used for the synthesis of pyrazolimines.

2.4.5. β-Ketonitriles

For details on *Cyanoacetic Acid* and its *Esters* see Malonic Acid and Derivatives, Chap. 3. Malonic Acid and Derivatives, Chap. 4.

Acetylacetonitrile [*2469-99-0*], β-ketobutyronitrile, 3-oxo-butanenitrile, $H_3CCOCCH_2CN$, M_r 83.09, *bp* 120 – 125 °C (101.3 kPa), can be obtained by hydrolysis of β-aminocrotononitrile [61]. A potential application is the synthesis of 3-amino-5-methylisoxazole, which is an intermediate for various drugs.

3,3-Dimethoxypropanenitrile [*57597-62-3*], cyanoacetaldehyde dimethyl acetal, 3,3-dimethoxy-propanenitrile, $(CH_3O)_2CHCH_2CN$, M_r 115.15, *bp* 195 °C, d_4^{20} 1.001, is a colorless to slightly yellow liquid.

3,3-Dimethoxypropanenitrile can be obtained from acrylonitrile by reaction with methanol in the presence of nitrite and platinum- or palladium-catalysts [62]. It is a valuable starting material for the synthesis of 4-aminopyrimidines such as cytosine.

2.5. Toxicology and Occupational Health

Occupational exposure to aliphatic nitriles mainly occurs by inhalation of vapors and/or by absorption through the skin. Pertinent toxicity and occupational health data relating to the most widely used aliphatic nitriles are summarized in Table 2.

The general toxicity of nitriles in humans includes gastric distress and vomiting, bronchial irritation, respiratory distress, convulsions, and coma. Animal toxicity studies also show gastrointestinal tract, liver, and pancreas irritations.

Whereas the old literature ascribes almost all the actions of nitriles to the liberation of cyanide ions, more recent reports tend to confirm that both the cyanide group and the entire molecule are important for biologic actions. While many of the neural effects are probably due to the cyanide moiety, the irritating and necrotic effects on the skin, liver, and kidney may be associated with the nature of the alkyl group. Although most aliphatic nitriles have cyanide release in common, there are significant differences in the amounts and duration exposure necessary to cause poisoning and toxicity effects. Signs of toxicity following oral administration of aliphatic nitriles to rats have highlighted the fact, that saturated nitriles affect the central nervous system in the same way as potassium cyanide, whereas unsaturated nitriles cause cholinomimetic effects such as salivation, diarrhea, and vasodilatation. Thus, cyanide release from unsaturated nitriles seems to play a minimal role in causing toxicity. Nitriles have tentatively been divided into three groups, with respect to toxicity, namely:

1) Those extensively metabolized to cyanide ions such as malononitrile and α-cyanohydrins
2) Those moderately metabolized to cyanide ions such as propionitrile and butyronitrile
3) Those poorly metabolized to cyanide ions such as acetonitrile, tetramethylsuccinonitrile, aminonitriles and acrylonitrile.

Table 2. Toxicological data

Nitrile	LD_{50}, mg/kg (oral, rats)	Inhalation, rats	LD_{50}, mg/kg (skin, rabbit)	Skin/eye irritation (rabbit)	Recommended workplace exposure limits, mg/m^3/type of limit
Saturated nitriles					
Acetonitrile	3800	LC_{50}: 7551 ppm/8h	1250	mild/severe	34/TWA; 70/MAK
Propionitrile	230	LCLo: 500 ppm/4h	210	mild/moderate	14/TWA
Butyronitrile	50	LCLo: 1000 ppm/4h	500	mild/mild	22/TWA
Isobutyronitrile	102	LCLo: 1000 ppm/4h	310	mild/mild	22/TWA
Saturated dinitriles					
Malononitrile	14			mild/mild	8/TWA
Succinonitrile	450				20/TWA
Adiponitrile	300				18/TWA
Unsaturated nitriles					
Acrylonitrile	78	LCLo: 500 ppm/4h	250	mild/severe	5/TWA
Allyl cyanide	115	LCLo: 500 ppm/4h	LDLo: 1410 (rabbit)	mild/mild	
Substituted nitriles					
Chloracetonitrile	220	LCLo: 250 ppm/4h	71	mild/moderate	
Cyanhydrins					
Acetoncyanhydrin	17	LCLo: 63 ppm/4h	17		4/C ∗
Glycolonitrile	16	LCLo: 250 ppm/4h	5		5/C ∗
Lactonitrile	87	LCLo: 125 ppm/4h	20		
3-Hydroxypropionitrile	3200		5000	mild/mild	

∗ C = Ceiling concentration not to be exceeded during a 15 min. sampling period.

3. Aromatic and Araliphatic Nitriles

Aromatic nitriles can be represented by the general formula Ar−C ≡ N. Further cyano groups or other substituents can be bonded to the aromatic nucleus. In *araliphatic* nitriles the aromatic nucleus is separated from the CN group by a $(CH_2)_n$ group, which can itself bear other substituents. Like the aliphatic nitriles the aromatic and araliphatic nitriles are important intermediates for organic synthesis.

3.1. Properties

Physical Properties. Aromatic and araliphatic nitriles are liquids or crystalline solids, mostly sparingly soluble in water (some heteroaromatic cyano-compounds have a higher water solubility) with sometimes considerable thermal stability. Benzonitrile and a range of substituted benzonitriles have a characteristic odor resembling bitter almonds. Some substituted benzonitriles sublime readily.

Chemical Properties. Reactions involving the cyano group are similar to the corresponding reactions of aliphatic nitriles (see Section 2.2). Ortho-substituents generally lower the reactivity of the cyano group. Functional groups attached to the aromatic ring normally react in the expected fashion. The bonds of halogens ortho and para to nitrile groups are weakened, i.e., these groups are prone to nucleophilic ipso-substitution. Ring-closure reactions involving the cyano-group are often possible. The nitrile group is a meta-directing substituent. The straight form of the cyano group allows even a hexacyano substitution at the benzene ring, in this case all six cyano groups being coplanar with the ring.

3.2. General Production Methods

Araliphatic nitriles can be produced by the methods similar to those described for aliphatic nitriles (see Section 2.3). The most important *cyanoaromatics*, e.g., benzonitrile, 2- and 4-chlorobenzonitrile, phthalonitrile, isophthalo- and terephthalonitrile, toluonitriles, and cyanopyridines are produced industrially by ammoxidation of the corresponding methylaromatics. Some newer special syntheses for aromatic nitriles also exist. The most important synthetic routes are shown schematically in Figure 3.

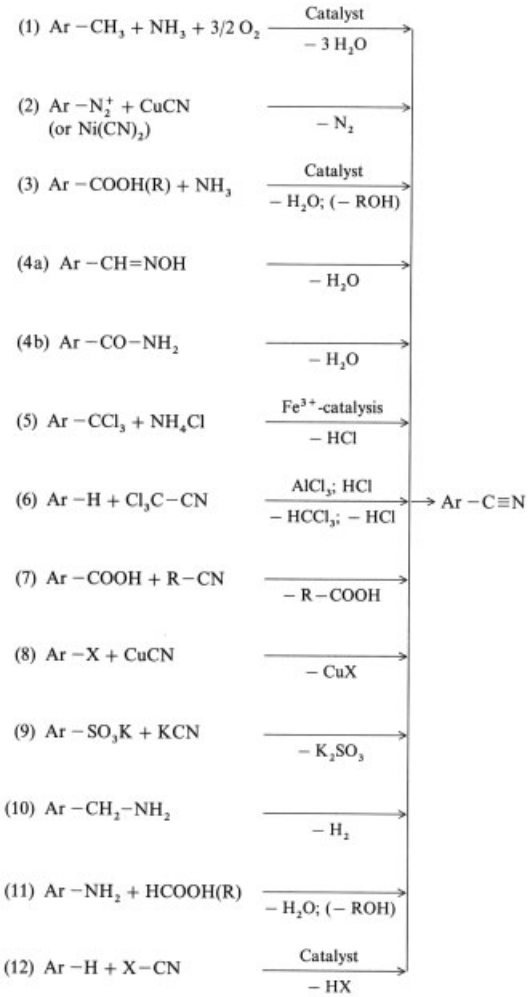

Figure 3. Synthetic routes to aromatic nitriles

Method 1 represents the ammoxidation, by which mono- and dinitriles are obtained [72]. For this reaction mixtures of aromatics (e.g., xylenes) can be also used, in this case nitrile mixtures are formed [73]. Liquid-phase ammoxidations in the presence of manganese (II) bromide are also known [74].

The acid-catalyzed treatment of aromatic carboxylic acids or their esters with ammonia, (method 3) at elevated temperatures in the gas phase, leads to the corresponding nitriles, often with good yields [75].

Elimination of water from aromatic aldoximes (method 4 a) is also performed with high yield. Many variants of this reaction can be found in the literature [76]. The dehydration of aromatic carboxamides is carried out similarly to that of aliphatic amides [77].

Reaction 5 (trichloromethylaromatics with ammonia or ammonium chloride) is catalyzed by iron (III) chloride [78].

The Houben – Fischer method (method 6) known for a long time, uses trichloroacetonitrile (see page 6) and thus leads to the unavoidable production of trichloromethane. In reaction 7 the pK value of the aromatic carboxylic acid should be higher than that of the acid from which the nitrile R – CN was formed. The conversion requires a temperature of 200 – 300 °C (pressure vessel) and frequently gives the nitrile in good yield. Acetonitrile, for example, is a suitable starting nitrile [79]. In place of the carboxylic acids, their alkali salts or esters can be used in some cases [80].

Formylation of an aromatic amine, subsequent thermal dehydration of the formanilide (reaction 11), and reaction via the isonitrile yields nitriles (analogous to the corresponding conversion of aliphatic amines) [81].

Some ether-substituted aromatics (e.g., anisole) react with cyanogen bromide in the presence of catalysts to form nitriles [82]:

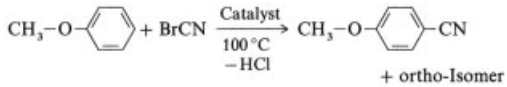

Further special syntheses of aromatic nitriles include reactions of aromatic carboxylic acids with urea and amidosulfonic acid at 200 – 400 °C [12] or with chlorosulfonyl isocyanate and subsequent cleavage of the adduct with triethylamine [84].

3.3. Selected Araliphatic and Aromatic Nitriles

3.3.1. Araliphatic Nitriles

Phenylacetonitrile [*140-29-4*], α-cyanomethylbenzene, benzyl cyanide, M_r 117.14, *mp* − 26 °C, *bp* 233.5 °C, d_4^{20} 1.0157, n_D^{17} 1.5243, flash point 102 °C, is a colorless, oily, toxic liquid.

Phenylacetonitrile is produced by the reaction of benzyl chloride with alkali cyanide in alcohol or aqueous solution under phase trans-

fer catalysis with *N,N*-dialkylbenzylammonium chloride [85].

Phenylacetonitrile is used as an intermediate in the production of synthetic penicillins or barbiturates, in the synthesis of optical bleaches for fibers, in the production of insecticides, and for perfumes and flavors (via phenylacetic acid to "honey-type" ester compounds).

Labeling and transport: RID/ADR 6.1, no. 12C

1,2-Phenylenediacetonitrile [*613-73-0*], 1,2-bis-(cyanomethyl)benzene, M_r 156.18; *mp* 59 – 60 °C; forms colorless crystals.

The nitrile is used as an intermediate for the synthesis of optical brighteners and is produced by the reaction of 1,2-bis(chloromethyl)benzene with sodium cyanide in aqueous suspension under phase transfer catalysis [86].

Phenylglyoxylonitrile [*613-90-1*], benzoyl cyanide, M_r 131.13, *mp* 34 °C, *bp* 206 – 208 °C (101.3 kPa), *bp* 96 °C (2 kPa), *bp* 82 °C (1.2 kPa), n_D^{26} 1.5303, can be used for selective reactions at the carbonyl and cyano groups. It is produced from the reaction of benzoyl chloride or benzoic anhydride with sodium cyanide (or hydrogen cyanide in the presence of alkali cyanide) in practically quantitative yield with copper (I) cyanide as catalyst [87]. Benzoyl cyanide is used as an intermediate in the synthesis of plant protection agents.

Labeling and transport: RID/ADR 6.1, no. 24B

d,l-Mandelonitrile [*532-28-5*], benzaldehyde cyanohydrin, α-cyano-α-hydroxymethylbenzene, $C_6H_5-CH(OH)-CN$, M_r 133.15, *mp* – 10 °C, d_4^{20} 1.1165 – 1.120, is miscible with ethanol and diethyl ether, immiscible with water.

Mandelonitrile occurs naturally as the β-glycoside of gentiobiose (amygdalin). It decomposes on heating into benzaldehyde and hydrogen cyanide. It is produced from benzaldehyde and alkali cyanide in the presence of mineral acid [88]. Mandelonitrile is used as an intermediate in the production of mandelic acid (→ Hydroxycarboxylic Acids, Aromatic, Chap. 4.4.).

Benzothiazolyl-2-acetonitrile [*56278-50-3*], 2-cyanomethylbenzothiazole, M_r 174.21, *mp* 102 °C, forms colorless crystals.

This nitrile is produced by the reaction of 1-amino-2-mercaptobenzene (*o*-aminothiophenol) with the monoiminoether hydrochloride of malononitrile [89] and is used as a condensation component for methine dyes (→ Methine Dyes and Pigments, Chap. 7.3.3.).

3.3.2. Aromatic Nitriles

3.3.2.1. Cyanobenzenes

2-Chlorobenzonitrile [*873-32-5*], 2-chloro-1-cyanobenzene, M_r 137.57, *mp* 44.8 °C, *bp* (101.3 kPa) 232 °C, *bp* (1.6 kPa) 111.5 °C, forms colorless crystals that readily sublime. The crystals are readily soluble in organic solvents, only sparingly soluble in water, and steam volatile.

2-Chlorobenzonitrile is produced industrially by ammoxidation of 2-chlorotoluene [90]. It is used as an intermediate in the production of 2-amino-5-nitrobenzonitrile, which in turn is used for the synthesis of azo dyes (see below).

Labeling and transport: RID/ADR 6.1, no. 12C

4-Chlorobenzonitrile [*623-03-0*], 4-chloro-1-cyanobenzene, $Cl-C_6H_4-CN$, M_r 137.6, *mp* 97 °C, forms colorless crystals that can be sublimed. The compound is steam volatile, readily soluble in organic solvents, and only sparingly soluble in water. 4-Chlorobenzonitrile is produced industrially by the ammoxidation of 4-chlorotoluene [91]. It is used in the synthesis of a red pigment for plastics [92].

2,6-Dichlorobenzonitrile [*1194-65-6*], 1-cyano-2,6-dichlorobenzene, M_r 172.02, *mp* 144.5 – 146.5 °C, is a colorless, crystalline substance. Its solubility in water is 10 ppm at 25 °C. 2,6-Dichlorobenzonitrile is produced industrially by the ammoxidation of 2,6-dichlorotoluene [93]. In addition the nitrile is obtained by oxidation or side-chain-chlorination of 2,6-dichlorotoluene via 2,6-dichlorobenzoic acid and 2,6-dichlorobenzaldehyde, respectively. 2,6-Dichlorobenzonitrile shows herbicidal activity (Casoron, Duphar) and is used in fruit and vine cultivation. It is also used as

an intermediate in the production of 2,6-difluorobenzonitrile and of 2,6-dichlorothiobenzamide (Prefix, Shell) which shows herbicidal activity.

2,6-Difluorobenzonitrile [*1897-52-5*], 1-cyano-2,6-difluorobenzene, M_r 172.02, *mp* 29 – 30 °C, forms colorless crystals. The nitrile can be produced by treatment of 2,6-dichlorobenzonitrile with potassium fluoride in sulfolane [23] and is used as an intermediate in the synthesis of insecticides (e.g., Dimilin, Shell).

2-Chloro-4-nitrobenzonitrile [*28163-00-0*], 2-chloro-1-cyano-4-nitrobenzene, M_r 182.57, *mp* 81 °C, is produced by the reaction of the diazonium salt derived from 2-chloro-4-nitroaniline with copper cyanide (Sandmeyer reaction) [95]. 4-Amino-2-chlorobenzonitrile can be obtained from 2-chloro-4-nitrobenzonitrile by reduction of the nitro group with iron (Béchamp method): it is used as an intermediate for the production of azo dyes.

4-Chloro-2-nitrobenzonitrile [*34662-32-3*], 1-chloro-4-cyano-3-nitrobenzene, M_r 182.57, *mp* 100 – 101 °C, is obtained by the Sandmeyer reaction of 4-chloro-2-nitrobenzene diazonium salt with copper cyanide [95]. Reduction of the nitro group with iron yields 2-amino-4-chlorobenzonitrile: the latter is an intermediate for the production of azo dyes.

2-Chloro-5-nitrobenzonitrile [*16588-02-6*], 1-chloro-2-cyano-4-nitrobenzene, M_r 182.57, *mp* 108 – 109 °C, forms yellow needle-shaped crystals. The chlorine atom ortho to the nitrile group is prone to a range of nucleophilic substitution reactions, of which that with ammonia has achieved industrial importance [96]. The nitrile is produced by nitration of 2-chlorobenzonitrile (see above) at 0 °C [97].

2-Amino-5-nitrobenzonitrile [*17420-30-3*], 2-cyano-4-nitroaniline, M_r 163.13, *mp* 210 °C, is readily soluble in acetone, chloroform, and benzene; itis sparingly soluble in water and ethanol. As already mentioned, the reaction of 2-chloro-5-nitrobenzonitrile (see above) with ammonia gives 2-amino-5-nitrobenzonitrile [98], an intermediate for azo dyes. A newer synthetic route starts from 6-nitroisatoic anhydride [2H-3,1-benzoxazine-6-nitro-2,4 (1H)-dione]: the latter is converted into 2-amino-5-nitrobenzonitrile by reaction with ammonia, subsequent treatment with phosgene in dimethylformamide or *N*-methylpyrrolidone, followed by hydrolysis of the formamidine [77].

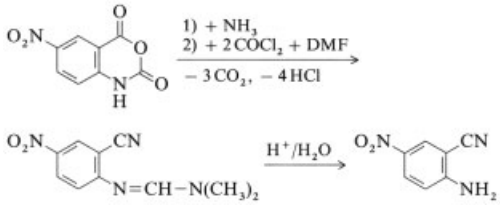

4-Hydroxybenzonitrile [*767-00-0*], 4-cyanophenol, 1-cyano-4-hydroxybenzene, M_r 119.2, forms colorless crystals of *mp* 113 °C. 4-Hydroxybenzonitrile is produced by reaction of 4-bromophenol with copper (I) cyanide [99], by ammoxidation of *p*-cresol [100], by reaction of 4-chlorobenzonitrile with sodium methoxide [101], or by the catalyzed gas phase reaction of ethyl 4-hydroxybenzoate with ammonia [75]. The nitrile is used for the production of the herbicides 3,5-dibromo- and 3,5-diiodo-4-hydroxybenzonitrile (Bromoxynil, Ioxynil, May & Baker).

3.3.2.2. Cyanonaphthalenes (Naphthalene Carboxylic Acid Nitriles)

In principle cyanonaphthalenes can be produced by the general methods given in Section 3.2. Additional production methods involving ring-closure condensation reactions are described under 1,4-dicyanonaphthalene. The industrial importance of cyanonaphthalenes is low.

1-Cyanonaphthalene [*86-53-3*], α-naphthonitrile, 1-naphthoic acid nitrile, M_r 153.17, *mp* 35.5 – 36 °C, *bp* 296 – 299 °C, forms colorless crystals that are readily soluble in ethanol. It is used in organic synthesis.

2-Cyanonaphthalene [*613-46-7*], β-naphthonitrile, 2-naphthoic acid nitrile, M_r 153.17, *mp* 66 °C, *bp* 303 °C, forms colorless crystals that are readily soluble in ethanol. It is

used in organic synthesis and suitable as a reaction medium for obtaining propylene from ethylene and 2-butene by olefin metathesis [102].

1,4-Dicyanonaphthalene [*3029-30-9*], naphthalene-1,4-dicarboxylic acid dinitrile, M_r 178.18, *mp* 208 °C, crystallizes in long, colorless needles. 1,4-Dicyanonaphthalene is produced by condensation of 1,2-phenylenediacetonitrile (see page 11) with glyoxal or its bisaldimines [103]. It is used as an intermediate for the synthesis of optical brighteners and dyes from naphtholactams.

2,6-Dicyanonaphthalene [*31656-49-2*], naphthalene-2,6-dicarboxylic acid dinitrile, M_r 178.18, *mp* 294 – 297 °C, can be obtained by ammoxidation of 2,6-dimethylnaphthalene on sodium vanadate bronzes doped with Fe_2O_3 [104]. It is used for the production of 2,6-bis(aminomethyl)naphthalene, an intermediate for dyes.

6-Cyano-β-naphthol [*52927-22-7*], 6-cyano-2-hydroxynaphthalene, 6-hydroxy-2-naphthoic acid nitrile, M_r 169.17, *mp* 166 – 167 °C. This nitrile is used as an intermediate for the production of liquid crystals [105] and monoamino-oxidase-inhibiting naphthalene-2-oxyalkylamines, which are used as antidepressive agents [106].

3.4. Toxicology

Mandelonitrile is rather toxic with an LD_{50} value (rat, oral) of 116 mg/kg [107]. This high toxicity may be due to the equilibrium of the cyanohydrin with benzaldehyde and hydrogen cyanide.

Phenylacetonitrile has an oral LD_{50} value (rat) of 270 mg/kg [108].

Benzonitriles with Methyl Substituents. The various methyl-substituted derivatives (2-, 3-, and 4-methylbenzonitrile and 2,3-, 2,4-, and 2,5-dimethylbenzonitrile) are less toxic than benzonitrile itself (→ Benzoic Acid, Chap. 9.). Their LD_{50} values (rat, oral) are in the range 3000 – 4500 mg/kg [109]. The monomethylbenzonitriles are reported to be metabolized by oxidation of the methyl group [109]. Mono- and dimethylbenzonitriles generally show a stronger irritating effect on the skin and mucous membrane than benzonitrile [109].

Miscellaneous Substituted Benzonitriles. The oral LD_{50} values of some substituted benzonitriles are as follows:

2-Chlorobenzonitrile	435 mg/kg [109]
4-Chlorobenzonitrile	887 mg/kg [107]
2,6-Dichlorobenzonitrile	2710 mg/kg [107]
2-Chloro-5-nitrobenzonitrile	3200 mg/kg [109]
4-Nitrobenzonitrile	30 mg/kg [109]

The LD_{50} values (rat, oral) for nitriles which are used as herbicides are as follows [110]:

3,5-Diiodo-4-hydroxybenzonitrile	305 mg/kg
3,5-Dibromo-4-octanoyloxybenzonitrile	420 mg/kg
Trichlorophenylacetonitrile	1050 mg/kg
Diphenylacetonitrile	200 mg/kg

4. References

1. K. Weissermel, H.-J. Arpe: *Industrial Organic Chemistry,* Verlag Chemie, Weinheim 1978, pp. 268 – 269.
2. Distillers Co., GB 723 003, 1955 (D. J. Hadley, C. A. Woodcock).
3. Eastman Kodak Co., US 2 786 867, 1957 (H. J. Hagemeyer, B. Thompson, C. W. Hargis).
4. Ruhrchemie AG, DE 3 014 729, 1981 (G. Horn, D. Frohning, H. Liebern).
5. Bayer, DE 3 216 382, 1983 (F. Hagedorn et al.).
6. A. Hulkenberg, J. J. Troost, *Tetrahedron Lett.* 1974, 3187.
7. C. Fizet, J. Streith, *Tetrahedron Lett.* 1982, 1505.
8. G. Lohaus, *Chem. Ber.* **100** (1967) 2719.
9. U. Feldhues, H. J. Schäfer, *Synthesis* 1982, 145.
10. Bayer, EP 58 333, 1982 (D. Arlt, G. Klein).
11. in [1]pp. 217 – 218.
12. W. Nagata, M. Yoshioka, *Org. React. (N.Y)* **25,** 1977 255 – 476.
13. Ethyl Corp., DE 2 618 744, 1976 (E. G. Woods, L. H. Shephard, E. P. Breidenbach).
14. In [1]pp. 216 – 217.
15. Degussa, EP 36 441, 1981 (A. Kleeman, B. Lehmann, H. Klenk).
16. M. Kobler, K. H. Schuster, G. Simchen, *Liebigs Ann. Chem.* 1978, 1946.

17. T. M. Reetz, *Tetrahedron* **39** (1983) 961.
18. Asahi Chemical Industry Co., JP 84 185 788, 1984.
19. SRI International: *Chemical Economics Handbook,* Menlo Park, Feb. 1988.
20. Sagami, BE 847 864, 1977 (K. Kondo et al.).
21. Takeda, US 2 592 930, 1952.
22. Lonza AG, CH 641 155, 1984 (H. Althaus).
23. Mitsubishi Chemicals Ind. Co., Ltd., JP 74 47 303, 1974 (Y. Kageyama, Y. Fukai).
24. Canter et al., *J. Chem. Soc.* 1931, 1245.
25. E. R. Biehl, H. M. Li, *J. Org. Chem.* **31** (1966) 602.
26. Merck and Co., Inc., US 3 020 200, 1962 (S. Rogers).
27. J.R. Geigy AG, US 2 754 243, 1956 (M. Gysin).
28. Standard Oil Co., Ohio, JP 82 67 551, 1982.
29. *Chem. Eng. News* (1984, May 7) 19.
30. H. G. Elias: *New Commercial Polymers 1969 – 1975,* Gordon and Breach Science Publishers, New York 1977, p. 21.
31. Mitsubishi Petrochemicals Co., Ltd., JP 82 158 750, 1982.
32. BASF, EP 57 890, 1982 (W. Rebafka, G. Heilen, K. Halbritter, W. Franzischka).
33. A. J. Fatiadi, *Synthesis* 1987, 749, 959.
34. Bayer, EP 9 788, 1980 (R. Schubart, R. Braden).
35. Mitsubishi Chemical, JP 84 212 459, 1984.
36. Ciba, US 2 928 829, 1960 (R. P. Mull).
37. Hoffmann-La Roche, EP 4 334, 1979 (H. Fischer, F. Schneider, R. Zurflüh).
38. Carpenter, *J. Org. Chem.* **27** (1962) 2085.
39. Dow Chemical Co., US 3 923 860, 1975 (S. H. Ruetman).
40. Du Pont, US 3 418 228, 1968 (P. L. Bartlett).
41. Olin Mathieson, US 3 260 588, 1966.
42. Dow Chemical, US 2 890 238, 1959 (A. R. Sexton).
43. Mitsubishi Gas Chemical Co., Inc., JP 8 280 346, 1982.
44. Showa Denko K.K., JP 79 46 720, 1979 (K. Nakayasu, O. Furuya).
45. Showa Denko K.K., US 4 299 978, 1981 (K. Nakayasu, O. Furuya, Y. Hosaki).
46. Asahi Chemical Ind. Co., JP 82 82 345, 1982.
47. Merck and Co., US 2 742 491, 1956 (J. Weijlard, A. P. Sullivan).
48. Pfizer, US 2 957 025, 1960 (J. O. Brooks).
49. Hoechst, DE 2 555 043, 1977 (H. Erpenbach, K. Gehrmann, H. Joest, W. Lork).
50. Yamanouchi, DE 2 951 675, 1979 (Y. Hirata, I. Yanagisawa, Y. Ishii, M. Takeda).
51. Nitto Chemicals Ind. Co. Ltd., JP 84 196 851, 1984.
52. Orsymonde, GB 1 325 192, 1971 (L. Lafon).
53. Bristol-Myers Co., DE 2 057 845, 1971 (Y. H. Wu, J. W. Rayburn).
54. Ciba-Geigy Co., US 3 853 942, 1974 (Y. S. Surey, H. C. Grace).
55. Nippon Chemicals Co., Ltd., US 4 066 683, 1978 (S. Hideo).
56. *Japan Chem. Week* **Dec. 1** (1988).
57. Nippon Soda, DAMN company brochure.
58. M. Mittelback, *Monatsh. Chem.* **116** (1985) no. 5, 689 – 691.
59. Nippon Kayaku, JP 62 220 557, 1987 (Tada, Shoji).
60. Lentia GmbH, DE 3 231 052, 1984 (G. Stern, F. Grossgut, M. Joos, H. Weismann).
61. Sogo Pharmaceuticals Corp., Ltd., US 4 152 336, 1979 (K. S. Masakata, K. M. Sadamitsu, S. Y. Kazuyuki).
62. Ube Industries Ltd., EP 55 108, 1980 (K. Matsui, S. Uchimumi, A. Iwayama, T. Umeza).
63. *NIOSH Criteria for a Recommended Standard,* US Department of Health, Education and Welfare, Washington 1978.
64. G. D. Clayton, F. E. Clayton: *Patty's Industrial Hygiene and Toxicology,* 3rd revised ed., **vol. 2 C,** pp. 4845 – 4900.
65. N. I. Sax: *Dangerous Properties of Industrial Materials,* 6th ed., Van Nostrand Reinhold Comp., New York 1984.
66. A. E. Ahmed, Y. M. Faroqui, *Toxicol. Lett.* **12** (1982) 157 – 163.
67. E. H. Silver, S. H. Kuttab, T. Hasan, M. Hassan, *Drug. Metab. Dispos.* **10** (1982) 495 – 498.
68. A. E. Ahmed, N. M. Trieff, *Prog. Drug Metab.* **7** (1983) 229 – 294.
69. S. Szabo, G. T. Gallagher, *Surv. Synth. Pathol. Res.* **3** (1984) 11 – 30.
70. H. Tanii, K. Hashimoto, *Arch. Toxicol.* **55** (1984) 47 – 54.
71. H. Tanii, K. Hashimoto, *Arch. Toxicol.* **57** (1985) 88 – 93.
72. BASF, DE-OS 1 643 722, 1967 (C. Palm et al.). BASF, EP 0 222 249, 1985 (H. Engelbach, R. Krokoszinski, W. Franzischka, M. Decker).
73. Showa Denko, *Hydrocarbon Process.* **45** (1969) no. 11, 252.
74. Rhône Poulenc, DE-OS 2 141 657, 1970 (Y. Colleuille, R. Gardon).
75. Ueno, EP 0 074 116, 1981 (R. Ueno, K. Sakota, K. Kawata, Y. Naito).
76. E. Vowinkel, J. Bartel, *Chem. Ber.* **107** (1974) 1221 – 1227.

77. BASF, DE-OS 2 628 055, 1976 (D. Schneider, H. Scheuermann).
78. Hoechst, DE-OS 1 668 068, 1967 (A. Hansel). Bayer, DE-OS 2 550 262, 1975 (F. Hagedorn, E. Klauke, K. Wedemeyer).
79. F. Becke, T. Burger, *Justus Liebigs Ann. Chem.* **716** (1968) 78 – 82.
80. Sumitomo, DE-OS 2 047 160, 1969 (S. Sagawa, T. Nakagawa, S. Okamoto).
81. BASF, EP 0 073 326, 1981 (P. Magnusson, W. Trauzettel).
82. Mobil Oil, US 3 433 821, 1965 (L. A. Hamilton, P. S. Landis).
83. J. Lücke, R. E. Winkler, *Chimia* **25** (1971) 94.
84. H. Vorbrüggen, *Tetrahedron Lett.* **13** (1968) 1631 – 1634. Schering, DE-OS 1 643 009, 1967 (H. Vorbrüggen).
85. K. Fukunaga, *Yuki Gosei Kagaku Kyokaishi* **33** (1975) no. 10, 774. Bayer, DE-OS 1 668 034, 1967 (H. Leuchs, K. H. Schmidt).
86. Ciba-Geigy, DE-OS 2 743 094, 1976 (L. Guglielmetti, A. C. Rochat, I. J. Fletcher).
87. Bayer, DE-OS 2 614 240, 2 614 241, 2 614 242, 1975 (K. Findeisen). Degussa, DE-OS 2 624 891, 1975 (H. Klenk et al.).
88. Bayer, DE-OS 2 914 091, 1980 (H. H. Schwarz, H. P. Wirges, A. Judat, G. Zumach).
89. Ciba-Geigy, DE-OS 2 632 402, 1975 (P. Loew, H. Schwander, H. Kristinsson).
90. Nippon Kayaku, DE-OS 2 711 332, 1978 (H. Hayami, H. Shimizu, G. Takasaki).
91. Nippon Shokubai, EP 0 290 996, 1987 (A. Inoue et al.).
92. Ciba-Geigy, EP 0 094 911, 1982 (A. C. Rochat, L. Caesar, A. Iqbal).
93. Nitto, EP 0 158 928, 1986 (Y. Kiyomiya, Y. Yamaguchi, M. Ushigome, H. Murata). Süddeutsche Kalkstickstoffwerke, EP 0 158 928, 1984 (H. Bayerl, W. Pollwein).
94. I.S.C. Chemicals, DE 2 902 877, 1978 (G. Fuller); *J. Agri. Food Chem.* **21** (1973) no. 6, 997.
95. C. W. N. Holmes, J. D. Loudon, *J. Chem. Soc.* 1940, 1521 – 1523.
96. J. Baudet, *Rec. Trav. Chim. Pays Bas* **43** (1924) 708 – 715.
97. B. B. Dey, Y. G. Doraiswami, *J. Indian Chem. Soc.* **10** (1933) 309.
98. B. Hartmanns, *Rec. Trav. Chim. Pays-Bas* **65** (1946) 468 – 469.
99. Sumitomo, DE-OS 2 047 161, 1969 (S. Sagawa, T. Nakagawa, S. Okamoto).
100. Akademie d. Wissenschaften d. DDR, DD 232 693, 1984 (A. Martin et al.).
101. J. Prachensky et al., CS 242 141, 1984.
102. Pittsburg Petroleum Gas, US 3 637 894, 1969 (J. C. Crano, E. K. Fleming).
103. Ciba-Geigy, DE-OS 2 649 167, 1975 (H. Schwander, C. Zickendraht, L. Guglielmetti). Hoechst, DE-OS 2 503 321, 1975 (L. Heiss).
104. Suntech, US 4 064 072, 1976 (R. D. Bushick).
105. RCA Corp., US 3 925 237, 1974 (D. L. Ross, D. M. Gavrilovic).
106. Kyowa Hakko, JP 52 083 533, 52 023 055, 1976 (N. Nakamizo, T. Hirata, T. Takahashi).
107. J. V. Marhold: *Sbornik Vysledku Toxikologickeho Vysetreni Latek A pripavku,* Institut pro Vychovu Veducicn Prakovniku Chemickeho Prumyclu, Praha, Czechoslovakia, 1972, p. 161.
108. G. P. Galibin, V. J. Fedorova, N. M. Karamzina, *Gig. Sanit.* **32** (1967) no. 8, 20.
109. H. Zeller, H. Th. Hoffmann, A. M. Thiess, W. Hey, *Zentralbl. Arbeitsmed. Arbeitsschutz* **19** (1969) 225.
110. A. Fischer: BASF, *Herbizide,* 1971, p. 154.

Nitrilotriacetic Acid

Charalampos Gousetis, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Hans-Joachim Opgenorth, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Properties 1
3. Production 2
4. Analysis 3
5. Uses 3
6. Economic Aspects 4
7. Toxicology and Environmental Aspects 4
8. References 5

1. Introduction

Nitrilotriacetic acid [*139-13-9*], NTA, *N,N*-bis(carboxymethyl)glycine, along with ethylenediaminetetraacetic acid (EDTA, → Ethylenediaminetetraacetic Acid and Related Chelating Agents) is one of the most important aminopolycarboxylate chelating agents.

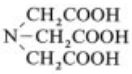

Its synthesis from ammonia and chloroacetic acid was accomplished by Heintz as early as 1861 [1]. Industrial production first took place in 1936 in Ludwigshafen. Since that time, NTA has become an established chelating agent and is used in many industries where multivalent metal ions, chiefly alkaline-earth ions, must be "masked" in order to modify their properties or prevent typical reactions such as the formation of insoluble precipitates.

2. Properties

Physical Properties. Nitrilotriacetic acid, $C_6H_9NO_6$, M_r 191.14, is a tribasic acid, crystallizing from water as colorless needles. The most important physical data are as follows [2]:

mp (decomp.)	242 °C
Solubility in water, g/100 g	
5 °C	0.13
22.5 °C	0.13
80 °C	0.95
100 °C	3.3
Dissociation constants [3] (25 °C, ionic strength = 0.1 mol/L)	
pK_1	1.80
pK_2	2.48
pK_3	9.65

The pH of a saturated solution (ca. 0.13 g NTA per 100 g of solution at room temperature) is ca. 2.3. The alkali-metal salts are the most soluble NTA salts. For example, ca. 640 g of trisodium nitrilotriacetate [*5064-31-3*] dissolve in 1 L of water.

Chemical Properties. The most important property of NTA is its ability to form watersoluble chelates with multivalent metal ions. Not only the carboxyl groups but also the tertiary nitrogen atom function as ligands. Unoccupied coordination sites of the metal ion can be saturated by water molecules. Although the chelates of NTA with metal ions are often represented as octahedral (coordination number 6) for the sake of simplicity, recent structural studies (chiefly by X-ray methods) have shown that higher coordination numbers are also possible for some metal ions. For example, Ca^{2+} in the chelate $CaNTA^-$ has a coordination number of 7 [4].

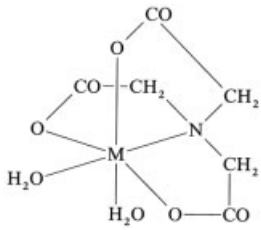

As a rule, complexes are 1 : 1, i.e., 1 mol of metal ions reacts with 1 mol of NTA. If NTA is present in excess, 1 : 2 complexes with 1 mol of metal ions complexed by 2 mol of NTA are formed.

The stability of the metal complex is described by stability constants (K_1 for 1 : 1 complexes, K_2 for 1 : 2 complexes; see Table 1).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_377

Further, because of the polarization of the H−O bond in the chelate, the 1 : 1 complexes behave as weak acids and dissociate. This effect is described by the dissociation constant K_d:

$$K_1 = \frac{[\mathrm{MNTA}]}{[\mathrm{M}]\,[\mathrm{NTA}]}$$

$$K_2 = \frac{[\mathrm{MNTA_2}]}{[\mathrm{M}]\,[\mathrm{NTA}]^2}$$

$$K_d = \frac{[\mathrm{MNTAOH}]\,[\mathrm{H^+}]}{[\mathrm{MNTA}]}$$

where M is a multivalent metal ion, NTA represents the anion $N(CH_2COO^-)_3$, and the charges have been omitted for simplicity except for H^+.

Table 1. Stability constants for NTA chelates∗ [3]

Metal ion	$\log K_1$	$\log K_2$	pK_d
Al^{3+}	11.4		5.09
Ca^{2+}	6.39	8.76	
Cd^{2+}	9.78	14.39	11.25
Co^{2+}	10.38	14.33	10.80
Cu^{2+}	12.94	17.42	9.14
Fe^{2+}	8.33	12.80	10.60
Fe^{3+}	15.90	24.30	4.1/7.8∗∗
Hg^{2+}	14.60		
Mg^{2+}	5.47		
Mn^{2+}	7.46	10.94	
Ni^{2+}	11.50	16.32	10.86
Pb^{2+}	11.34		
Zn^{2+}	10.66	14.24	10.06

∗ Ionic strength (25 °C) = 0.1 mol/L.
∗∗ Behaves as a dibasic acid.

Complexing of metal ions with NTA competes with other secondary reactions. For example, metal ions react with anions such as carbonate, sulfide, sulfate, or oxalate to form sparingly soluble precipitates, and hydrogen ions compete with metal ions for sites on the NTA trianion. This means that several equilibria must be taken into consideration in the analysis of NTA systems.

Nitrilotriacetate chelates are stable over relatively wide pH ranges that depend on the concentration of the chelate and the excess of the complexing agent: Ca^{2+}, pH 9 – 12; Mg^{2+}, pH 7 – 10; Cu^{2+}, pH 3 – 12; Fe^{3+}, pH 1.5 – 3. Assuming a stoichiometric ratio for 1 : 1 complexing, in the optimal pH range 1 g of Na_3NTA binds 156 mg of Ca^{2+}, 94 mg of Mg^{2+}, 247 mg of Cu^{2+}, or 217 mg of Fe^{3+}.

3. Production

The original synthesis of NTA from ammonia and chloroacetic acid has only historical significance [1]. The oxidation of triethanolamine is likewise of no industrial importance [5]. The one-stage alkaline and two-stage acid processes in use are based on the cyanomethylation of ammonia (or ammonium sulfate) with formaldehyde and sodium cyanide (or hydrogen cyanide).

Alkaline Process. The alkaline process was long the established method for NTA production. Trisodium nitriloacetate is synthesized as follows:

$$NH_3 + 3\,HCHO + 3\,NaCN \rightarrow N(CH_2CN)_3 + 3\,NaOH$$

$$N(CH_2CN)_3 + 3\,NaOH + 3\,H_2O \rightarrow N(CH_2COONa)_3 + 3\,NH_3$$

The reaction can be carried out batchwise or continuously, but the continuous process is more economical.

In a typical process, aqueous NaCN solution is fed into a cascade reactor system along with formaldehyde solution [6]. Ammonia is liberated during the synthesis and need not be supplied. The reaction takes place at 80 – 100 °C. Because of the high pH (ca. 14), triscyanomethylamine $N(CH_2CN)_3$ [*7327-60-8*] is hydrolyzed in situ to Na_3NTA. This process generates three times as much ammonia as it consumes and the ammonia concentration must be limited to suppress the production of substances with low degrees of carboxymethylation (glycine, iminodiacetic acid). This is largely achieved by continuously distilling off the ammonia with steam or air throughout the process. However, formation of byproducts (chiefly glycolic acid, hexamethylenetetramine, and the above-mentioned amino acids) cannot be completely prevented.

The resulting solution is sold directly as a 40-wt % solution, or used in the production of $Na_3NTA \cdot H_2O$ in powder form, or acidified to pH 1 – 2 to yield the acid (H_3NTA).

Acid Process [7]. The significant yield of byproducts in the alkaline process has led in recent years to the construction of plants based on the acid process, which features much lower byproduct levels. The acid process is associated

with stringent safety requirements due to the use of hydrogen cyanide; corrosion can also be a problem.

In the first stage, ammonia is reacted with formaldehyde to give hexamethylenetetramine, which is then reacted with hydrogen cyanide in sulfuric acid solution to yield triscyanomethylamine. The solid triscyanomethylamine is sparingly soluble in the acidic solution and is filtered off, washed, and saponified with NaOH to give Na_3NTA. The resulting solution has a far lower byproduct content than the solution from the alkaline method. It is also sold as 40 % product or used in the production of $Na_3NTA \cdot H_2O$ or H_3NTA (see above).

Trade Names. The following trade names apply to the most important NTA products: Dissolvine A grades (Akzo), Hampshire NTA grades (W. R. Grace), Masquol NTA grades (Protex), Nervanaid NTA grades (ABM Chemicals), Rexene NTA grades (Rexolin Chemicals), Trilon A grades (BASF), and Versene NTA grades (Dow).

4. Analysis

A number of methods are suitable for the quantitative determination of NTA and its salts. One widely used technique is potentiometric titration with iron(III) chloride under acidic conditions [8, 13]. Ion-pair chromatography [9], GC analysis [10], and polarographic determinations [11] can also be employed in the trace range.

5. Uses

The uses of NTA and its salts are based on their complexing properties for metal ions [12, 13]. It is employed in many areas to eliminate the interfering effects of metal ions by complexing, to dissolve precipitates, to modify the redox potentials of metal ions, or to prepare metal-ion buffers. The most important application areas are described below.

Water Softening. Nitrilotriacetic acid softens water by complexing Ca^{2+} and Mg^{2+} ions; complexation occurs in the neutral to basic range. The softening of water with hardnesses of 1 °d, 1 °f, 1 °e, and 1 mmol Ca^{2+}/L requires the following amounts of chelating agent: 34, 19, 27, and 191 mg NTA, or 46, 26, 36, and 257 mg Na_3NTA, respectively. [1 °d (German hardness) corresponds to 0.01 g CaO/L water, 1 °e (English hardness) to 0.01 g $CaCO_3$/0.7 L water, and 1 °f (French hardness) to 0.01 g $CaCO_3$/L water.]

Water softening with NTA is used in the paper and textile industries, soap manufacture, cosmetics and toiletries, the production of detergents and cleansers, the treatment of boiler feedwater, and in many chemical process industries (e.g., photographic materials and electroplating industries).

Phosphate Replacement in Detergents. Na_3NTA can be used as a substitute for pentasodium triphosphate in the formulation of low phosphate detergents or, with zeolite and sometimes polycarboxylates, in phosphate-free detergents (→ Laundry Detergents, Chap. 3.2.4.). Substitution of phosphates can help reduce the eutrophication of surface waters.

In comparison with pentasodium triphosphate, Na_3NTA has the advantage of resisting hydrolysis, so that it remains fully effective after spray drying of the detergent slurry and also during washing. The equivalence ratio of Na_3NTA to pentasodium triphosphate is about 0.6 to 1 and there is no detrimental effect on primary and secondary detergency (anti-redeposition).

Replacement of Tetrapotassium Diphosphate in Cleansers. The equivalence ratio of Na_3NTA to tetrapotassium diphosphate is favorable (ca. 1 : 3) and only low concentrations of Na_3NTA are needed in cleansers. This often makes the use of hydrotropes or solubilizers superfluous. Synergetic builder effects occur with partial substitution of the phosphate.

Masking of Heavy-Metal Ions. Although inferior to EDTA in many applications, NTA is also frequently employed for masking heavy-metal ions because of its greater specific complexing power for interfering ions such as Fe^{3+}, Cu^{2+}, and Mn^{2+}. Examples of such applications are in soap manufacture, textile bleaching, the cleaning of hard surfaces, and in many areas of the chemical process industries.

Special applications include the separation of rare-earth elements, complexometric titrations, and the preparation of micronutrients for agriculture.

6. Economic Aspects

It is difficult to state figures for production capacity of NTA mainly because all existing plants can also be used for producing other aminopolycarboxylates. A worldwide capacity of ca. 100 000 t seems realistic however [15].

World consumption of Na_3NTA (as 100 % product) was ca. 40 000 t in 1985 and about 70 000 t in 1989 [15]. There continues to be excess production capacity, built when a rapid increase in demand was expected for the formulation of phosphate-free and low-phosphate detergents, chiefly in the United States. The expected increase has only taken place in Canada, where ca. 40 000 t of Na_3NTA were used in 1988. In Western Europe, consumption is ca. 20 000 t/a; NTA is only used in detergents in a few European countries (Switzerland, The Netherlands). Consumption in the United States has stagnated at 5000 – 7000 t/a for industrial and institutional applications [14].

Important manufacturers are Monsanto (USA), W. R. Grace (USA, UK), BASF (FRG), ABM Chemicals (UK), Akzo (Netherlands), Aminkemi (Sweden), SA Dabeer (Spain), and Protex (France).

It is hard to predict market development for the coming years. Sizable increases should be expected in Western Europe if a number of countries (especially the FRG) adopt the use of NTA in detergents.

7. Toxicology and Environmental Aspects

The results of extensive studies on the environmental behavior and toxicology of NTA have been summarized [16–18].

After adaptation of relevant microorganisms (i.e., production of necessary enzymes), NTA is readily biodegradable. It is completely degraded to inorganic end products (i.e., mineralized) without any measurable concentration of organic metabolites being produced. Environmental studies performed after the introduction of NTA-containing detergents in Switzerland, also confirm the good biodegradability: $> 98\,\%$ of the NTA is biodegraded in biological water treatment plants [19–21]. The half-life in the receiving water is < 1 d [22]; virtually no NTA could be detected in ground water after infiltrating a few meters [23, 24].

As a powerful complexing agent, NTA can form heavy-metal complexes under certain conditions. Whether such an effect occurs in the environment depends on the NTA concentration. Since NTA is rapidly biodegraded, remobilization of heavy metals from the activated sludge of water treatment plants is negligible in comparison with the normal heavy-metal levels in the effluent [22, 23]. The same applies to the remobilization of heavy metals from river sediments, where under realistic conditions no measurable changes in the heavy-metal concentrations can be expected.

The acute toxicity to aquatic organisms is slight. Data for more than 50 freshwater and marine species are known; LC_{50} values vary from ca. 100 to $> 10\,000$ mg/L. The numerical values depend highly on the hardness of the test medium. Numerous studies have also been performed on the chronic toxicity of NTA in microorganisms, macroinvertebrates, amphibians, and fish. No toxic effects are seen at NTA concentrations below 1 mg/L. In most organisms, perceptible toxic effects only occur at very much higher levels.

The acute toxicity in mammals is species dependent. For example, LD_{50} (oral) values of ca. 2000 mg/kg are found for the trisodium salt ($Na_3NTA \cdot H_2O$) in rodents; vomiting occurs in dogs and monkeys. After inhaling 5 mg of NTA dust per liter of respired air over 4 h, rats displayed no clinical symptoms. The irritant action of NTA on the skin, eyes, and respiratory tract is slight; indications of sensitization and allergies are not known. Nitrilotriacetic acid is not teratogenic either by itself or in combination with heavy metals such as cadmium and mercury. Numerous mutagenicity tests have not shown indications of genotoxicity.

Nitrilotriacetic acid is not metabolized by mammals; any absorbed NTA is excreted unaltered by the kidneys. Tissue concentrations are highest in the cells of the renal tubules and in the

ureters. Subchronic and chronic studies therefore show toxic lesions of the kidneys, which can be attributed to disorders of electrolyte and iron metabolism. High doses (e.g., 2000 mg per kilogram body weight per day) result in injury to the epithelial cells of kidneys, ureters, and bladder, and as a consequence to tumor formation. There is, however, a threshold concentration below which tumors cannot develop. The threshold concentration is much higher than the NOEL (no adverse effect level) of 14 mg per kilogram body weight per day.

8. References

1. W. Heintz, *Ann. Chem. Pharm.* **122** (1862) 257 – 294.
2. *Beilstein,* **H** 369, **E I** 482, **E II** 801, **E III** 1180.
3. A. E. Martell, R. M. Smith: *Critical Stability Constants,* vol. 1, Plenum Press, New York 1974.
4. G. Wilkinson (ed.): *Comprehensive Coordination Chemistry,* **vol. 3,** Pergamon Press, Oxford 1987, p. 33.
5. Carbide & Carbon Chemical, US 2 384 816, 1945 (G. Curme, Jr., H. Chitwood, J. Clark); US 2 384 818, 1945 (H. Chitwood).
6. BASF, DE 837 999, 1952 (E. Ploetz); DE 2 062 435, 1970 (E. Hartert); DE 2 625 974, 1976 (K.-L. Hock). W. R. Grace, DE 1 813 718, 1969 (G. Busch, B. Elofsson).
7. BASF, DE 3 029 205, 1980 (H. Distler, K.-L. Hock). W. R. Grace, US 3 463 811, 1967 (J. Goldfrey, J. Sykes, J. Harper, C. Morgan). Hampshire Chemical, US 2 855 428, 1958 (J. Singer, M. Weisberg). Monsanto Co, EP 148 146, 1985 (C. Y. Shen). Stauffer Chemical US 3 984 453, 1966 (S. Chaberek).
8. G. Schwarzenbach, H. Flaschka: *Die komplexometrische Titration,* 5th ed., Enke Verlag, Stuttgart 1985. S. Siggia, D. Eichlin, R. Rheinhart, *Anal. Chem.* **27** (1955) 1745 – 1749.
9. J. Weiss: *Handbuch der Ionenchromatographie,* Dionex GmbH, Weiterstadt, 1985, p. 253.
10. N. T. de Oude, *Vom Wasser* **64** (1985) 283 – 292.
11. DIN 38 413, Part 5.
12. *Ullmann,* 4th ed., **17,** 340.
13. BASF, Technical Bulletin, *Trilon Chelating Agents,* Ludwigshafen 1989.
14. M. Salaices: "Chelating Agents," in *Chemical Economics Handbook Product Review,* SRI International, Menlo Park 1987.
15. BASF, personal communication, Ludwigshafen 1989.
16. H. Bernhardt (ed.): "NTA, Studie über die aquatische Umweltverträglichkeit von Nitrilotriacetat," Verlag H. Richarz, Sankt Augustin 1984.
17. R. L. Andersson, W. E. Bishop, R. L. Campbell, *CRC Crit. Rev. Toxicol.* **15** (1985) 1 – 102.
18. E. Bayer (ed.): "Nitrilotriessigsäure, BUA-Stoffbericht 5," Beratergremium für umweltrelevante Altstoffe (BUA) der Gesellschaft Deutscher Chemiker, VCH Verlagsgesellschaft, Weinheim – New York 1987.
19. W. Giger, M. Ahel, M. Koch, H. U. Laubscher, C. Schaffner, J. Schneider, *Water Sci. Technol.* **19** (1987) 449 – 460.
20. A. Alder et al. in *Eidgenössische Anstalt für Wasserversorgung, Abwasserreinigung und Gewässerschutz Jahresbericht 1987,* pp. 3.5 – 3.9.
21. H. Siegrist, A. Alder, W. Gujer, W. Giger, *Gas Wasser Abwasser* **68** (1988) 101 – 109.
22. P. Reichert et al., in *Eidgenössische Anstalt für Wasserversorgung, Abwasserreinigung und Gewässerschutz Jahresbericht 1987,* pp. 4.1 – 4.2.
23. E. Kuhn, M. van Loosdrecht, W. Giger, R. P. Schwarzenbach, *Water Res.* **21** (1987) 1237 – 1248.
24. L. Schaffner, M. Ahel, W. Giger, *Water Sci. Technol.* **19** (1987) 1195 – 1196.

Nitrites → Nitrates and Nitrites
Nitro Alcohols → Nitro Compounds, Aliphatic

Nitro and Nitroso Dyes

Roderich Raue, Bayer AG, Leverkusen, Federal Republic of Germany

John F. Corbett, Clairol, Stamford, CT 06 922, United States

1. Structure and Color 1
2. History 2
3. Uses 2
4. Hydroxynitro Dyes 2
5. Aminonitro Dyes 2
5.1. Alkylaminonitro Dyes 3
5.2. Arylaminonitro Dyes 3
5.2.1. Arylaminonitro Disperse Dyes 4
5.2.2. Anionic Arylaminonitro Dyes 6
5.2.3. Cationic Arylaminonitro Dyes 7
5.3. Heterocyclic Nitro Dyes 7
6. Azonitro Dyes 8
6.1. Azonitro Disperse Dyes 8
6.2. Anionic Azonitro Dyes 8
6.3. Cationic Azonitro Dyes 10
7. Nitro Dyes with Other Color Systems 10
8. Nitro Dyes in Hair Coloring 10
8.1. Nitroaniline Hair Dyes 12
8.2. Diaminonitrobenzene Hair Dyes . . 12
8.3. Aminonitrophenol Hair Dyes 13
8.4. Nitrodiphenylamine Hair Dyes . . . 14
8.5. Formulation with Nitro Hair Dyes . 14
9. Biologically Active Nitro Dyes 14
10. Nitroso Dyes 15
11. References 16

1. Structure and Color

Nitro dyes are characterized by the fact that a nitro group is conjugated with an electron donating group via an aromatic system. In all industrially important dyes the nitro and donor groups are ortho to each other. A further donor group can also be present.

The donors are usually hydroxyl and amino groups. The hydroxynitro dyes used industrially always contain several nitro groups. In the aminonitro dyes the nitrogen can bear alkyl or aryl groups. In heterocyclic nitro dyes the amino nitrogen is part of a heterocyclic ring system which has been condensed with the aromatic ring bearing the nitro group.

There are also nitro dyes that contain a second color-producing system, e.g., azonitro dyes, anthraquinone nitro dyes, and quinophthalone nitro dyes.

The simple nitroaniline and nitrophenol derivatives are yellow dyes. Substitution of the amino nitrogen gives a bathochromic shift towards orange-yellow color shades.

The 2,5-, 2,4-, and 3,4-didonor-substituted nitrobenzene derivatives extend the range of color shades considerably (Table 1).

The 2,5-disubstituted compounds have the longest wavelength absorption, and the 2,4-disubstituted ones the shortest.

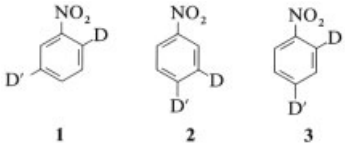

Table 1. Color – structure relationship among nitro dyes

Electron donors		Longest wavelength maxima (ethanol)		
D	D′	2,4-D,D′	3,4-D,D′	2,5-D,D′
NH_2	NH_2	389	408	471
NHR	NH_2			495 – 497 ∗
NH_2	NHR		431	490 – 492 ∗
NHR	NRR			506 – 536 ∗
OH	NH_2	394	394	446
NH_2	OH	394	370	454
OH	NHR	403	412	477
OH	OH	376	390	400

∗ R = methyl at low end of range; R = 2-hydroxyethyl at high end.

The color range of the didonor-substituted nitrobenzene derivatives extends from the greenish yellow of 2,4-diaminonitrobenzene [*5131-58-8*] (**3**) ($D = D' = NH_2$) to the blue-violet of N^2,N^5,N^5-tris(2-hydroxyethyl)-2,5-diaminonitrobenzene [*33229-34-4*] (**1**) ($D = NHCH_2CH_2OH$, $D' = N(CH_2CH_2OH)_2$). 2,5-Diaminonitrobenzene [*5307-14-2*] (**1**) ($D = D' = NH_2$) and 4-(2-hydroxyethylamino)-3-nitrophenol [*65235-31-6*] (**1**) ($D = NHCH_2CH_2OH$, $D' = OH$) are red dyes. Sub-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_383

stituted nitrodiphenylamine derivatives give orange to magenta shades.

Nitroso dyes always contain a hydroxyl group as donor. They can be formulated as tautomers of the quinone oxime and are capable of forming metal complexes.

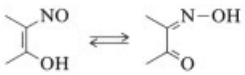

The iron complexes are green pigments.

2. History

The oldest nitro dye is picric acid [*88-89-1*], which was first obtained in 1771 by WOULFE by treating indigo with nitric acid. In 1843 LAURENT determined its structure and prepared it by nitrating phenol. MARTIUS nitrated α-naphthol in 1864 to give the dye Martius yellow. Naphthol Yellow S [*483-84-1*], the first acid hydroxynitro dye, was prepared by CARO in 1879. In 1898 aminonitrobenzene dyes were used for the first time as toners in hair dyeing. The first acid aminonitro dye, Amido Yellow E [*6373-74-6*], was obtained by SCHMIDLIN in 1911. In 1908, Lithol Fast Yellow GG [*2955-16-0*] was the first nitro dye to be used as a pigment. Nitrodiphenylamine disperse dyes achieved importance with the development of cellulose acetate fibers. Important contributions to their development were made by MEYER, HOPFF, and ELLIS in 1924. GMELIN and FISCHER prepared the first 2-nitrodiphenylamine-4-sulfonamides in 1936, and this work was continued by KOPP and GANGNEUX in 1941. Later it became apparent that these dyes were particularly suitable for lightfast dyeing of polyester fibers. More recently, nitro dyes have also been used for transfer printing on polyester woven fabrics and in modern reproduction processes (ink-jet printing, optical data recording, and thermal transfer printing).

1-Nitroso-2-naphthol was prepared by FUCHS in 1875; its iron complex is Pigment Green B. In 1883 HOFFMANN obtained the soluble iron complex Naphthol Green B [*6636-34-6*] from 1-nitroso-2-naphthol-6-sulfonic acid.

3. Uses

Acid nitro dyes were first used to dye wool. More recently they have, like the acid azonitro dyes, been used for dyeing polyamide fibers. Water-insoluble nitro dyes are used either as disperse dyes for polyester materials or as pigments, depending on their structure. Low molecular mass dyes are suitable for transfer printing on polyester materials [4]. Alkylaminonitro dyes with long-chain alkyl groups are suitable for dyeing fats and oils. Cationic nitro dyes, in particular nitroazo dyes, are occasionally used for dyeing polyacrylonitrile. Nitro dyes are also used in ink-jet printing [5] and optical data storage [6]. Apart from dyeing, nitro dyes are used as herbicides or fungicides, and picric acid and hexanitrodiphenylamine [*2844-92-0*] as explosives.

4. Hydroxynitro Dyes

Picric acid (**4**) was formerly used for dyeing wool, silk, and leather. It has been recommended for the colored electrophoretic coating of aluminum as a mixture with alkyd resins [7].

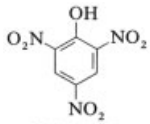

4 [*88-89-1*]

Martius Yellow, 2,4-dinitro-1-naphthol, protects woven woolen material against moths, but has now been replaced by colorless compounds. A sulfonic acid derivative, Naphthol Yellow S (**5**), is obtained by treating 1-naphthol-2,4,7-trisulfonic acid with nitric acid. It can be used to dye wool and silk yellow.

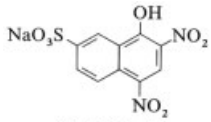

5 [*483-84-1*]

5. Aminonitro Dyes

Aminonitro dyes have low color strength. The molar extinction coefficient is 4000

$L\,mol^{-1} \cdot cm^{-1}$ for *N*-alkylaminonitro dyes and 6000 to 7000 $L\,mol^{-1} \cdot cm^{-1}$ for *N*-arylaminonitro dyes. Only on the introduction of an azo chromophore does the molar extinction coefficient rise to 14 000 $L\,mol^{-1} \cdot cm^{-1}$ [8]. Since some of these dyes have high lightfastness and low production costs, they are nevertheless still in use.

5.1. Alkylaminonitro Dyes

Aminonitro dyes with tert-alkyl groups on the amino nitrogen are suitable for dyeing polyester, triacetate, and acetate rayon materials [9].

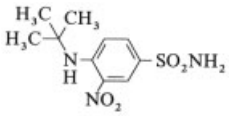

6 [*28799-71-5*]

The dye **6** colors these fibers greenish yellow. Dyes of this type are also suitable for dyeing organic solvents, oils, fats, waxes, or resins. Dyes with longer alkyl groups, e.g., the compound **7**,

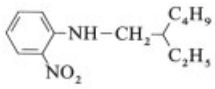

7

are used to color gasoline and mineral oils [10]. An example is C.I. Solvent Yellow 96 (Sudan Yellow G, BASF).

Hydroxyalkylaminonitro dyes that have been esterified with unsaturated carboxylic acids such as acrylic acid can be polymerized and used in polymer form for dyeing leather coatings [11]. A commercial product: Corial Yellow H3G (BASF).

Yellow pigments are obtained by condensation of nitroanilines with aqueous formaldehyde. Lithol Fast Yellow GG (**8**) is used to color paints [12].

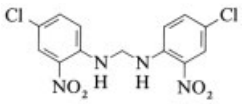

8 [*2955-16-0*]

An acid dye **9**, with two alkylaminonitro dye groups, is obtained by treating 2 mol 2-nitrochlorobenzene-4-sulfonic acid with 1 mol 1,4-bis(aminomethyl)benzene [13].

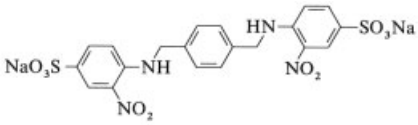

9

Silicon-containing reactive dyes are derived from 2,4-dinitroaniline. The dye **10** is suitable for dyeing natural and synthetic fiber materials and glass fibers [14].

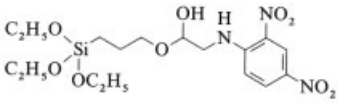

10

Cationic nitro dyes such as **11** are obtained by treatment of an aliphatic triamine with 2-nitrochlorobenzene [15]. The dye **11** colors polyacrylonitrile yellow.

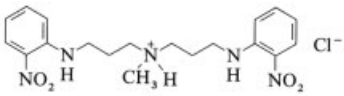

11

Compounds that are quaternized at the central nitrogen atom are also recommended for dyeing polyacrylonitrile and acid-modified polyamide fibers [16].

Nitro dyes with two alkyl groups on the amino nitrogen, e.g., the 2-nitro-*N*,*N*-dimethylaniline-4-sulfonamides, are also suitable for dyeing synthetic fibers. The sulfonic acid-4-ethoxyanilide is known commercially as C.I. Disperse Yellow 86 [*67338-59-4*]. The alkylamides are recommended for transfer printing on woven polyesters [17].

5.2. Arylaminonitro Dyes

Among the arylaminonitro dyes, the disperse dyes and those with sulfonic acid groups are particularly important. Cationic dyes are also known within this class of compounds.

5.2.1. Arylaminonitro Disperse Dyes

The simplest member of this series of dyes is 2-nitrodiphenylamine [*119-75-5*]. It is produced from 2-nitrochlorobenzene and aniline in the presence of copper oxide and *N*-methylpyrrolidone as catalyst. Potassium carbonate is added to bind the hydrochloric acid formed [18]. The reaction is accelerated by the addition of a quaternary tetraalkyl- or benzyltrialkylammonium salt. The corresponding phosphonium salts are also suitable catalysts [19].

Sterically hindered aromatic amines are treated with 2-fluoronitrobenzene in the presence of potassium fluoride at 180 °C [20].

Another process for the production of nitrodiarylamine compounds involves the treatment of 2-nitrochlorobenzene with an alkali-metal salt of a formanilide at 130 – 180 °C in a higher boiling alcohol. During the reaction, the formyl group is cleaved as carbon monoxide [21].

Finally, 2-nitrophenols are also suitable starting materials for the production of 2-nitrodiphenylamines. For example, 2,4-dinitrophenol is treated with phenyl isocyanate in the presence of potassium acetate in nitrobenzene for 6 h at 210 °C. 2,4-Dinitrodiphenylamine [*961-68-2*] is formed with loss of CO_2 [22]. The commercial name of the dye is Disperse Yellow 14.

From 2,4-dinitrochlorobenzene, Disperse Yellow 9 [*6373-73-5*] is obtained with *p*-phenylenediamine, and Disperse Yellow 1 [*119-15-3*] (**12**) with 4-aminophenol. Production of the latter is accelerated in the presence of a surface-active agent and a compound that forms metal complexes, such as ethylenediaminetetraacetic acid [23].

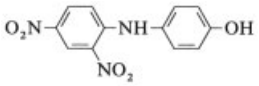

12

The dye is known under the following commercial names: Celliton Fast Yellow RR (BASF), Dispersol Fast Yellow A (ICI), Medurin Fast Yellow A (Mevisa), Navicet Fast Yellow RR (IDI), Serisol Fast Yellow A (Yorkshire Chemicals), Supracet Fast Yellow RR (Holliday), Tar Disperse Yellow B (Tarsus, Turkey).

A yellowish orange dye **13** is obtained by condensation of 1,4-dichloronitrobenzene with 4-anisidine [24]. The commercial name is Amacel Golden Orange 3 G (American Aniline).

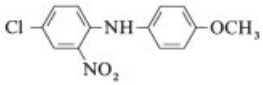

13 [*6373-69-9*]

Mass spectrometric investigations show that 2-nitrodiphenylamines form more stable molecular ions than 4-nitrodiphenylamines. In the latter the nitro group is cleaved more readily. This is also the reason for the lower photostability of the 4-nitrodiarylamines [25]. Spectroscopic investigations show that the nitro dyes are mainly present in the quinoid form in the excited state [26].

The action of picryl chloride on a crown-ether-substituted aniline gives nitrodiphenylamines which are suitable for the colorimetric determination of potassium [27] and lithium ions [28].

Sublimation-resistant nitro disperse dyes are obtained by introduction of the sulfonamide group para to the amino nitrogen. The dye **14** colors acetate rayon and polyester materials greenish yellow.

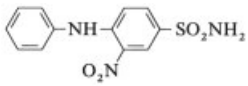

14 [*1565-46-4*]

Commercial products: Amacel Yellow GLF (Koppers), Cibacet Yellow GL (Ciba-Geigy), Serisol Fast Yellow GGL (Yorkshire Chemicals).

The corresponding *N*-butyl sulfonamide [*70660-56-9*] is supplied commercially as Serisol Fast Yellow YGL (Yorkshire Chemicals) and as Supracet Fast Yellow AFG (Holliday).

If 1-chloro-2-nitrobenzenesulfonamide is treated with *p*-phenetidine, an ethoxy group is introduced into the position para to the amino nitrogen. The absorption maximum is thus shifted bathochromically.

Commercial products: Amacel Yellow 4 RLF (Koppers) Amacron Yellow RLS (American Aniline), Cibacet Yellow 4 RL (Ciba-Geigy), Samaron Fast Yellow RRL (Hoechst), Vondanyl Yellow 2 R (Vondelingenplaat).

The industrially most important nitro dye **15** is produced by treatment of 1-chloro-2-nitroben-

zene-4-sulfonyl chloride with two equivalents of aniline. The aromatic chlorine substituent is substituted and a sulfanilide is formed.

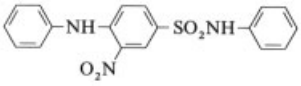

15 [*5124-25-4*]

An improved process for the production of **15** involves the treatment of 4-chloro-3-nitrobenzene-1-sulfonic acid with thionyl chloride at 70 – 80 °C to give the sulfonyl chloride, and its subsequent reaction with aniline in the presence of a surfactant at pH 7.5 – 12.5 in two stages at 5 – 50 °C and at 80 – 100 °C [29]. Calcium and magnesium carbonate have been proposed as acid-binding agents [30]. To improve the dyeing properties the dye has been converted into the heat-stable α-modification [31]. The resulting dye, Disperse Yellow 42, is contained in almost all product ranges for dyeing polyester. Dye mixtures that contain Disperse Yellow 42 as the yellow component have a particularly good exhaustion capacity on acetate fibers [32]. In 1990, 1100 t Disperse Yellow 42 were used for dyeing worldwide.

Commercial products include: Amacron Yellow LS (American Aniline), Cibacet Yellow GWL (Ciba-Geigy), Dianix Fast Yellow YL (Mitsubishi), Dispersol Fast Yellow T (ICI), Eastman Polyester Yellow ALSR (Eastman), Etalon Yellow GWL (Colorom), Foron Yellow SE-FL (Sandoz), Kayalon Polyester Yellow YL-SE (Nippon Kayaku), Miketon Polyester Yellow YL (Mitsui), Palanil Yellow GE (BASF), Polynal Yellow GSFD (National Aniline), Polyspers Yellow BT (Tenneco Chem.), Proxester Yellow GRL (Protex), Resolin Yellow GGLS (Bayer), Samaron Yellow 4 G (Hoechst), Serilen Yellow WLS (Yorkshire Chemicals), Sodyecron Yellow WLS (Southern Dyestuffs), Sumifix Yellow YGLN (Sumitomo), Terasil Yellow GWL (Ciba-Geigy), Tertranese Yellow P-GR (CKC/Tetre), Vondisperse Yellow ALS (Vondelingenplaat).

The dye produced analogously from *p*-toluidine has a color strength about 20 % higher [33]. A nitro dye with an oxyethylsulfonyl group para to the amino nitrogen is obtained by treatment of 1,4-phenylenediamine with 4-chloro-3-nitrophenyloxyethylsulfone [34].

Bifunctional nitro dyes are produced by reacting 1-chloro-2-nitrobenzene-4-sulfonyl chloride with ethylenediamine, followed by substitution of the aromatic chlorine by an aniline group [35].

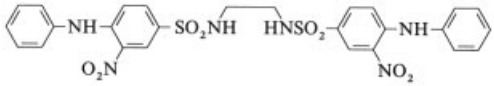

16 [*25960-13-8*]

The reaction of dianisidine with 2 mol 2,4-dinitrochlorobenzene leads to the pigment **17** [36], known commercially as Mikrolith Brown 2 R-N (Ciba-Geigy).

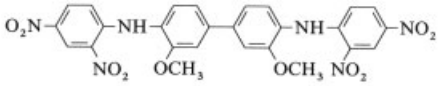

17 [*29398-96-7*]

The 2,4-dinitrophenylhydrazone of 2,4-dihydroxyacetophenone **18** is also used as a nitro dye [37].

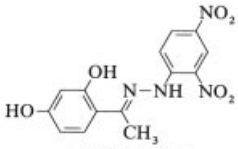

18 [*1166-12-7*]

Commercial product: Polysynthren Golden Orange 2 GSL (Cassella).

A hydrazone dye of the same type is obtained by coupling diazotized 2,4-dinitroaniline with α-acetyl-α-(4-bromophenyl)acetonitrile in aqueous ethanol, whereby the acetyl group is cleaved [38]. Yellow hydrazone dyes for polyesters are also obtained by coupling diazotized 2-nitroanilines, substituted by chlorine, methyl, or methoxy in the 4-position, with 4-nitrobenzylcyanide [39]. A nitro dye **19** of the hydrazone series with a heterocyclic end group is produced by coupling diazotized 2-nitroaniline with benzimidazolyl-2-acetamide [40].

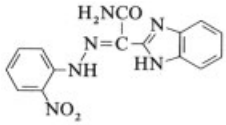

19 [*2328-07-6*]

Commercial products: Palanil Yellow 4 GE (BASF), Ostacet Yellow SE-2 G (Chemapol), Rathilen Yellow 4 GE (Rathidye Chem).

5.2.2. Anionic Arylaminonitro Dyes

The first anionic nitro dyes were produced by treating picryl chloride with aromatic amino-sulfonic acids. However, only the product of the reaction of 2,4-dinitrochlorobenzene with 4-aminodiphenylamine-2-sulfonic acid (**20**, R = NO_2 [*15347-52-1*]) became important industrially [41].

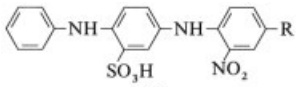

20

This dye colors wool and polyamide fibers in lightfast yellow shades and has good leveling capacity. It is also used for one-bath dyeing and tanning of furs and fleeces [42].
Commercial products: Acidol Yellow-Brown GTE (BASF), Lurazol Fur Yellow E (BASF), Anthralan Yellow RRT (Hoechst), Coranil Brown HEPS (Hoechst), Apollo Acid Yellow-Brown EL (Taiheung), Lissamine Fast Yellow AE (ICI), Xylen Fast Yellow ES (Sandoz).

For dyeing wool and polyamide in yellow-brown shades, other derivatives of **20** are also suitable. For example the mixture of the two dyes with R = SO_2NH_2 [*72829-30-2*] and R = $SO_2NHC_6H_5$ [*72829-29-9*] is commercially available as Nylosan Yellow-Brown E-GL (Sandoz) [43]. Other dyes have been proposed for the same purpose, in which R is a sulfonate ester [44], carbonyl [45], or carboxylic acid amide group [46]. A commercial product from this series of dyes is Sella Acid Brown 5 GL (CibaGeigy). 2-Nitro-4′-aminodiphenylamine-4-sulfonic acid [*135-11-5*] is produced by condensation of the sodium salt of 4-chloro-3-nitrobenzenesulfonic acid with 1,4-diaminobenzene in saturated sodium chloride solution [47]. Inorganic salts of divalent metals (e.g., FeS) have also been proposed as catalysts [48]. The compound is commercially available as Coranil Brown HEPM new (Hoechst). Sulfonation of this compound affords 2-nitro-4′-aminodiphenylamine-2′,4-disulfonic acid, an intermediate for the production of acid nitrodiphenylaminoazo dyes (see Chap. 6.2) [49].

A bifunctional acid nitro dye **21** (C.I. 10 410, Acid Brown 13) can be produced by reacting 4,4′-dichloro-3,3′-dinitrodiphenyl sulfone with 4-aminodiphenylamine-2-sulfonic acid [50].

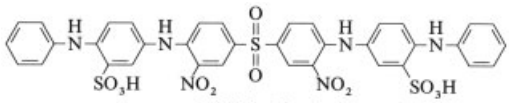

21 [*6373-79-1*]

Commercial product: Wofacid Fast Yellow-Brown 3 GL (Chemiekombinat Wolfen).

In the nitro dye **22**, which colors polyamide a brilliant greenish yellow, the conjugation is interrupted by CH_2 bridges between the aromatic rings [51].

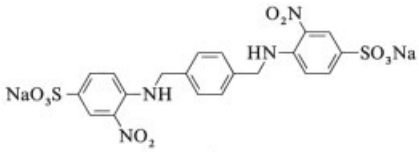

22

Some nitro dyes are classified as reactive dyes. An example is the dye **23**, which gives a violet-brown shade on cellulose [52].

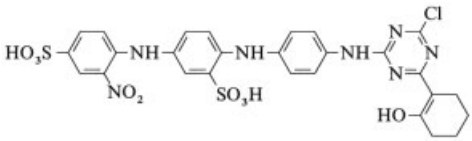

23 [*57975-30-1*]

For food dyeing it is desirable that the dye is not absorbed through the gastrointestinal tract. Polymeric nitro dyes such as **24** fulfil this requirement [53].

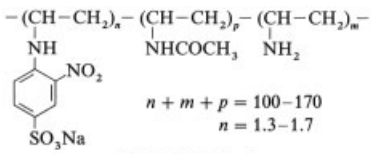

$n + m + p = 100–170$
$n = 1.3–1.7$

24 [*109563-84-0*]

5.2.3. Cationic Arylaminonitro Dyes

2-Nitrodiphenylamine-4-sulfonamides whose amide groups bear a cationic group are suitable for dyeing polyacrylonitrile [54].

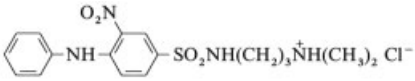

25 [*4261-62-5*]

5.3. Heterocyclic Nitro Dyes

Polycyclic ring systems containing an NH group in one of the rings acquire the character of a nitro dye if they have a nitro group in the position neighboring the NH group. 4-Nitroacridone (**26**) is formed on heating 2-nitrodiphenylamine-2′-carboxylic acid with concentrated sulfuric acid in a ring-closure reaction [55].

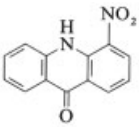

26 [*4261-62-5*]

This dye is known commercially as Polynal Yellow NLA (National Aniline). Its resistance to sublimation is low. It can therefore be used to dye synthetic fibers by the transfer printing process [56].

The resistance to sublimation of nitroacridone dyes can be increased by the introduction of arylthioether groups [57]. The dye **27** colors

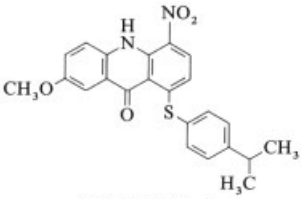

27 [*21810-30-0*]

polyester and polyamide yellow. Sublimationresistant acridone dyes are also obtained by introducing sulfonamide [58], sulfonate ester [59], and aryloxyalkylether groups [60].

Commercial products: Intrasil Yellow 2 GL (Intracolor), Terasil Yellow 2 GL (Ciba-Geigy).

Nitrophenoxazine and nitrophenothiazine dyes contain oxygen or sulfur, respectively, as the ring-closing member instead of the carbonyl group. They are also suitable for dyeing polyester fibers [61]. 2-Chloro-4-nitrophenoxazine dyes are produced by treating 1,4-dichloro-2,6-dinitrobenzene with 2-aminophenol derivatives in a polar organic solvent under alkaline conditions at 60 – 120 °C [62]. Under the same reaction conditions the 2,4-dinitrophenoxazine dyes are obtained with picryl chloride [63]. In both reactions a nitro group is cleaved during ring-closure.

Anionic heterocyclic nitro dyes are produced by reacting 2-aminophenol-4-sulfonic acids, which can be substituted in the 6-position by chloro, nitro, carboxyl, or sulfonic acid groups, with 2,4- [64] or 2,6-dinitrochlorobenzenes [65].

Nitrophenothiazine dyes are formed in the condensation of 2-aminothiophenol derivatives with 2,6-dinitrochlorobenzenes. Substitution at sulfur occurs initially, then a diphenylamine compound is formed by a Smiles rearrangement, and finally the phenothiazine ring closes with elimination of nitrous acid [66]. Nitro dyes of the phenazine series can be produced by reaction of 2,4,6-trinitroanisole or picryl chloride with 2,6-diaminodiphenylamine-4-sulfonic acids [67]. The dye **28** colors wool a lightfast brown. It was supplied commercially as Guinea Brown GRL.

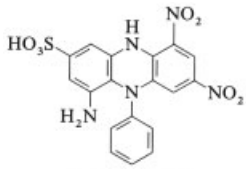

28 [*6373-81-5*]

Some indazolones also belong to the heterocyclic nitro dyes. The dye **29**, for example, colors wool in fast brown shades.

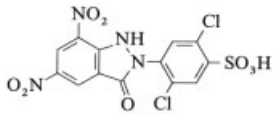

29 [*70209-94-8*]

Commercial products: Kiton Fast Brown 2 RL, Neonyl Printing Brown 2 R (Ciba-Geigy).

6. Azonitro Dyes

Azonitro dyes are still important industrially because of their increased color strength compared to the dyes described above.

6.1. Azonitro Disperse Dyes

1,4-Phenylenediamine, after treatment with one equivalent of a halonitrobenzene, can then be diazotized and coupled with aromatic hydroxy compounds [68]. Besides 2,4-dinitrochlorobenzene, 2-nitrochlorobenzene compounds bearing a sulfonamide [69], carboxylic acid ester, or a carbonyl group [70] in the 4-position can be used as starting materials. Commercial products of this series are Serilen Golden Yellow R-FS (**30**) (Yorkshire Chemicals):

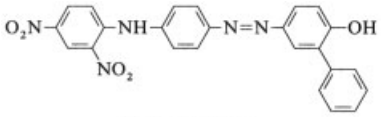

30 [*12217-97-9*]

and Sumikaron Yellow RS (**31**) (Sumitomo):

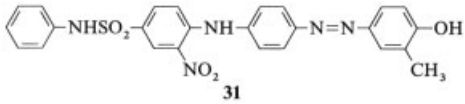

31

The hydroxyl group of the coupling component may be etherified [71] or esterified with benzenesulfonyl chlorides [72]. If the hydroxyl group is reacted with a polyether, a dye is obtained that colors polyester fibers from an aqueous liquor without previous dispersion [73].

1-Phenyl-3-methylpyrazol-5-one [74] and 2,6-dihydroxypyridine-4-carboxylic acid [75] are also possible coupling components.

The azo group can also be introduced by the reaction of 4-aminoazobenzene with halonitrobenzene compounds [76]. A further route to the introduction of nitro and azo components involves treating 2-nitrodiphenylamine-4-sulfonyl chloride with hydroxyazo dyes [77]. The nitro dye component can also be added on in the last stage by diazotizing and coupling 3-nitro-4-chloroaniline, and subsequently substituting the chlorine for an arylamino group [78].

6.2. Anionic Azonitro Dyes

For the production of anionic azonitro dyes, which are particularly suitable for dyeing polyamide materials, 1-chloro-2-nitrobenzene-4-sulfonic acid is treated with 1 mol 1,4-phenylene-diamine, diazotized, and coupled with aromatic hydroxy compounds [79]. The resulting dye (**32**, R = H, [*38455-53-7*]) is known commercially as Nylomine Yellow A-4 R (ICI) and Tertroxyl Fast Yellow 5 R (Crompton Knowles).

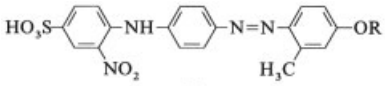

32

If the phenolic coupling component has a carboxyl group ortho to the OH group, the dye formed is particularly suitable for dyeing chromed leather, as it forms a chromium complex [80].

By alkylation of dyes which have been obtained by coupling with phenols [81] or naphthols [82], yellow or orange dyes are produced which color polyamide materials in very lightfast shades.

Esterification with *p*-toluenesulfonyl chloride gives a dye (**32**, R = $O_2S-C_6H_4-CH_3$, [*12220-06-3*]) which is contained in many product ranges for dyeing polyamide [83].
Commercial products: Kayanol Yellow N-3 R (Nippon Kayaku), Lanaperl Fast Yellow C-3 R (Hoechst), Nylanthren Yellow 3 RL (Althouse), Nylomine Orange C-G (ICI), Sandolan Golden Yellow MF-RL (Sandoz).

Besides the aromatic hydroxy compounds, amino compounds have also been reported as coupling components [84].

Anionic azonitro dyes with a sulfonic acid group on the central benzene ring are produced by treating *p*-phenylenediaminesulfonic acid with one equivalent of a halonitrobenzene, diazotization, and coupling to aromatic hydroxy compounds [85]. With *m*-cresol as the coupling component, the dye **33** is obtained which colors polyamide reddish yellow.

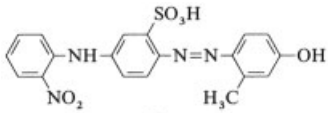

33

By etherification [86] or esterification [87] of the hydroxyl group, other dyes for polyamides are obtained.

A diazo component with two sulfonic acid groups can be produced by sulfonation of the product of the reaction of 2-nitrochlorobenzene-4-sulfonic acid with 1,4-phenylenediamine [88]. Diazotization and coupling with *o*-cresol gives an azonitro dye which colors polyamide fibers reddish yellow [89].

For synthetic polyamides, dyes are also used that have a sulfimide [90] or a sulfonamide group instead of a sulfonic acid group in the 4-position of the diphenylamine ring system. In the latter group of dyes the introduction of a sulfonic acid group either on the aromatic ring [91] or on the alkyl ether group is necessary to achieve sufficient solubility. This can involve either an alkanesulfonic acid group [92] or the sulfate ester of a hydroxyalkyl group [93].

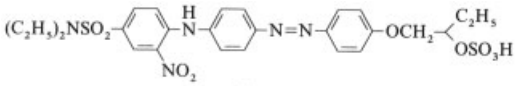

34

Finally, the sulfonic acid can also be bonded via the sulfonamide group [94]. The dye **35** colors polyamide yellow.

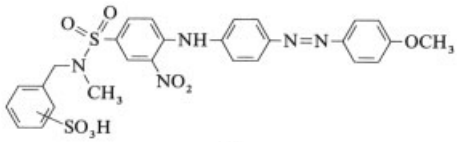

35

Nitrodisazo dyes are produced by diazotization of 4-amino-2′-nitrodiphenylamine-4′-sulfonic acid and coupling with aminohydroquinone dimethyl ether, diazotization of the monoazo dye, and coupling with naphthol or naphthylaminesulfonic acids [95]. The following dye colors nylon black:

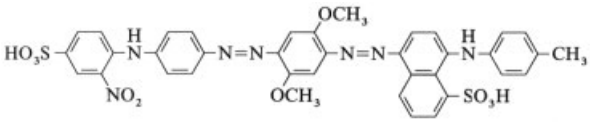

36 [*65612-07-9*]

1-Naphthylamine-6-sulfonic acid is also used as the middle component and 2-methylindole as the end component [96].

1-Chloro-2-nitrobenzene-4-hydroxyethylsulfone [34] is a starting material for the production of azonitro reactive dyes. After treatment with an aromatic [97] or aliphatic [98] diamine and conversion into the sulfate ester, the nitro dye is bonded to an azonitro reactive dye by treatment with an azo dye bearing a 2,4-dichloro-triazinyl group. The dye **37** contains two reactive groups and is used for level dyeing of mixtures of cotton with regenerated viscose fibers in a red shade.

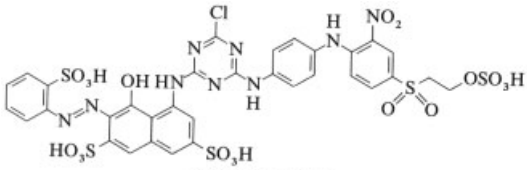

37 [*105926-57-6*]

Disazonitro reactive dyes are obtained by bonding the nitro dye components described above once [99] or twice [100] with a disazo dye via the triazinyl ring. The dye **38** colors cellulose blue with a green cast. It has a high color strength and a very good exhaustion capacity.

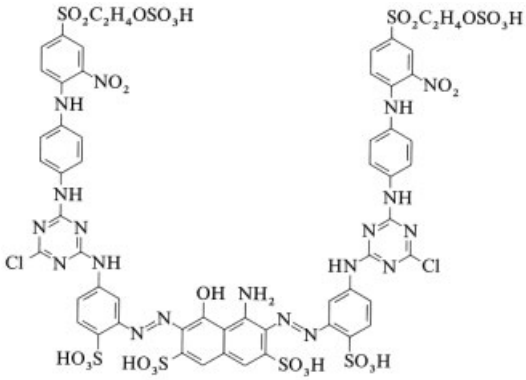

38 [*107471-33-0*]

Nitro dyes have also been converted into nitrodisazo dyes by diazotization and coupling with monoazo dyes [101].

Nitroazo metal complex dyes are produced by diazotizing the product of the reaction between 2-nitrochlorobenzene-4-sulfonic acid and 1,4-phenylenediamine and coupling it with the monoazo dye from 2-amino-4-nitrophenol and resorcinol. After conversion to the 2 : 1 chromium complex, a brown leather dye is obtained [102]. To produce yellow-brown dyes for polyamide fibers and wool, the nitro dye from 2-nitrochlorobenzene-4-sulfonic acid and 1,4-phenylenediamine is treated with diketene to give the acetoacetamide, which is coupled with diazotized 2-amino-4-nitrophenol. The 2 : 1 cobalt complex dye **39** dyes wool with very good wet fastness [103].

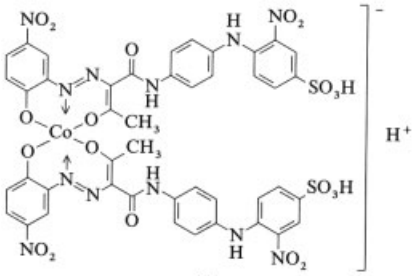

39

6.3. Cationic Azonitro Dyes

Cationic azonitro dyes are obtained by treating a 1-chloro-2-nitrobenzene-4-sulfonamide, bearing a cationic substituent on the amide group, with an aminoazo compound. The cationic group can be a dialkylamino [104] or a trialkylammonium group [105]. The dye **40** colors polyacrylonitrile yellow with a red cast.

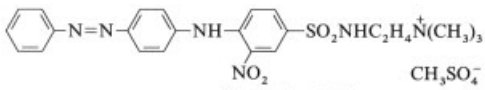

40 [*50567-78-7*]

7. Nitro Dyes with Other Color Systems

Anthraquinone nitro dyes have very low solubilities; most are pigments. By treatment of 3-nitro-4-chlorobenzoic acid amides with 1-aminoanthraquinone, pigments for coloring bulk plastics are formed [106]. The dye **41** colors PVC yellow.

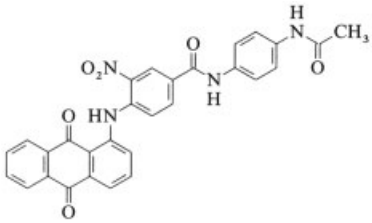

41

A yellow pigment for dyeing polyolefins, polystyrene, and PVC is the nitroquinophthalone dye obtained by treating 7-amino-2,4-dimethylquinoline with cyanuric chloride and 2 mol 2-nitroaniline, followed by condensation with tetrabromophthalic anhydride [107].

8. Nitro Dyes in Hair Coloring

The major commercial use of nitrobenzene dyes, bearing amino or amino and hydroxy substituents, is in the dyeing of human hair.

The importance of nitro dyes for this use is shown by their accounting for about 30 % of the more than 4000 hair dye patents issued worldwide since 1900.

The development of semi-permanent hair colorants as a commercial category can be traced to the exploitation of the *N*-hydroxyalkyl derivatives of nitroanilines and diamino-nitrobenzenes which were first described in 1951 by Unilever [108].

Today, nitrobenzene dyes are used mainly in the so-called semipermanent hair dyes. These products employ a variety of derivatives of nitroanilines, diaminonitrobenzenes, and aminonitrophenols (see Table 2) together with certain aminoanthraquinones and azo dyes [109]. The early semipermanent colorants employed nitro dyes dissolved in an anionic shampoo base. Today, the dyes are generally dispersed or dissolved in a nonionic lotion which is applied directly to the hair and left for 20 – 40 minutes before rinsing. The dyes diffuse into the hair to give annular dyeing which withstands six to

Table 2. Nitro dyes used commercially in the United States

Nitro dye	CAS registry no.	Color	Reference
N-(2-Hydroxyethyl)-2-nitroaniline (**43**)	[*4926-55-0*]	yellow	[108]
1,2-Diamino-4-nitrobenzene (**2**, D = D′ = NH_2)	[*99-56-9*]	yellow	
N^1-(2-Hydroxyethyl)-1,2-diamino-4-nitrobenzene (**2**, D = NH_2, D′ = $NHCH_2CH_2OH$)	[*56932-44-6*]	orange-yellow	[108]
N^1-Tris(hydroxymethyl)methyl-1,2-diamino-4-nitrobenzene (**2**, D = NH_2, D′ = $NHC(CH_2OH)_3$)	[*56932-45-7*]	orange-yellow	[108]
2-Amino-3-nitrophenol	[*603-85-0*]	yellow	[122]
2-Amino-4-nitrophenol	[*99-57-0*]	orange-yellow	
4-Amino-2-nitrophenol	[*119-34-6*]	orange	
2-Amino-5-nitrophenol	[*121-88-0*]	orange-yellow	
O,*N*-bis(2-Hydroxyethyl)-2-amino-5-nitrophenol (**61**)	[*59820-43-8*]	yellow	[122]
N-(2-Hydroxyethyl)-2-amino-5-nitroanisole (**62**)	[*66095-81-6*]	yellow	[125]
4-Amino-3-nitrophenol	[*610-81-1*]	orange-red	
N-(2-Hydroxyethyl)-4-amino-3-nitrophenol (**63**, R = H)	[*65235-31-6*]	red	[126]
N-(2-Hydroxyethyl)-4-amino-3-nitroanisole (**63**, R = CH_3)	[*57524-53-5*]	orange-red	[127]
1-(3-Methylamino-4-nitrophenyl)-propane-2,3-diol (**64**, R = $CH_2CHOHCH_2OH$)	[*80062-31-3*]	yellow	[128]
3-Methylamino-4-nitrophenoxyethanol (**64**, R = CH_2CH_2OH)	[*59820-63-2*]	yellow	[129]
2,5-Diaminonitrobenzene (**49**, R = H)	[*5307-14-2*]	red	
N^2-(2-Hydroxyethyl)-2,5-diaminonitrobenzene (**49**, R = CH_2CH_2OH)	[*2871-01-4*]	bluish red	[113]
N^5-(2-Hydroxyethyl)-2,5-diaminonitrobenzene (**56**, R = H)	[*24905-87-1*]	bluish red	[114]
N^2-Methyl-2,5-diaminonitrobenzene (**49**, R = CH_3)	[*2973-21-9*]	bluish red	[113]
N^2,N^5,N^5-tris(2-Hydroxyethyl)-2,5-diaminonitrobenzene (**44**, $R^1 = R^2 = R^3 = CH_2CH_2OH$)	[*33229-34-4*]	blue-violet	[108,113]
N^5-(2-Hydroxyethyl)-N^2,N^5-dimethyl-2,5-diaminonitrobenzene (**58**, R = CH_2CH_2OH)	[*10228-03-2*]	blue-violet	[117]
N^5-(2,3-Dihydroxypropyl)-N^2,N^5-dimethyl-2,5-diaminonitrobenzene (**58**, R = $CH_2CHOHCH_2OH$)	[*93633-79-5*]	blue-violet	[118]
2,4-Diaminonitrobenzene (**3**, D = D′ = NH_2)	[*5131-58-8*]	green-yellow	
Picramic acid (**60**)	[*96-91-3*]	magenta	
N-Methyl-isopicramic acid (**59**)	[*20291-98-9*]	violet	[123]
4-Amino-2-nitrodiphenylamine (**65**)	[*2784-89-6*]	red	[130]
4-Hydroxy-2′-nitrodiphenylamine (**66**)	[*54381-08-7*]	yellow	[131]

ten subsequent shampooings. The poor lightfastness, relative to normal textile dyes, is a consequence of their small molecular size. The latter property is of particular importance since it is necessary that the dyes should diffuse into the hair fiber at ambient temperature. Further, the poor wash fastness is of little disadvantage since the hair requires redyeing after four to six weeks due to the growth of new hair at the scalp.

In 1988, factory sales of hair colorant products in the United States exceeded $ 600 million. Of this, approximately 20 % was represented by semipermanent colorants containing nitro dyes.

The plethora of patents on nitrobenzene dyes for hair results in part from the fact that they represent the lowest molecular weight colored organic molecules, making them particularly suitable for dyeing at room temperature, and from the number of molecular structures that can be derived by N, O, or ring substitution of the six isomeric diaminonitrobenzenes and the ten isomeric nitroaminophenols.

In general, as can be seen from Table 2, the derivatives of *o*-nitroaniline and those of 2,5-didonor and 3,4-didonor substituted nitrobenzenes have been found to be the most useful. This can be ascribed to their ease of synthesis and to their superior lightfastness [110] compared to other isomers.

While a wide variety of O- and N-substituents have been examined, the most useful are evidently the hydroxyalkyl and methyl groups, appropriate combinations of which can be used to control solubility in the dye base and partition between the base and the hair.

A wide variety of reaction types, including direct nitration, selective reduction, electrophilic subsitution, alkylation, and molecular rearrangement are used in the manufacture of the commercially used nitro dyes.

8.1. Nitroaniline Hair Dyes

N-(2-hydroxyethyl)-2-nitroaniline (**43**) is employed as a yellow hair dye and is manufactured by heating 2-chloronitrobenzene (**42**) with monoethanolamine.

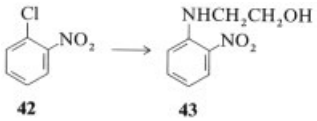

The 4-trifluoromethyl derivative of **43** has also found use and is similarly obtained from 2-chloro-5-trifluoromethylnitrobenzene [111].

8.2. Diaminonitrobenzene Hair Dyes

Three of the diaminonitrobenzenes, the 2,5-; 3,4-; and 2,4-isomers, have been used as hair dyes for about 90 years, although the last has rather poor lightfastness.

The red 2,5-isomer (**44**, $R^1 = R^2 = R^3 = H$) is obtained by nitration of *N,N′*-diacetyl-1,4-diaminobenzene, followed by hydrolysis, and is an important dye in the formulation of red and auburn shades.

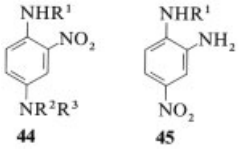

The yellow 3,4-isomer (**45**, $R^1 = H$) is manufactured from 2,4-dinitrochlorobenzene by reaction with ammonia, followed by selective reduction of the 2-nitro group of the resulting dinitroaniline with sodium sulfide. The orange-yellow N^4-hydroxyalkyl derivatives of **45**, where R is 2-hydroxyethyl or tris(hydroxymethyl)methyl, are similarly obtained by using monoethanolamine or tris(hydroxymethyl)methylamine in place of ammonia.

The first of the blue-violet nitro dyes to be described was N^2,N^5,N^5-tris(2-hydroxyethyl)-2,5-diaminonitrobenzene (**48**, $R = CH_2CH_2OH$) and was originally obtained, in a somewhat impure state, by direct hydroxyethylation of 2,5-diaminonitrobenzene with 2-chloroethanol or ethylene oxide [108].

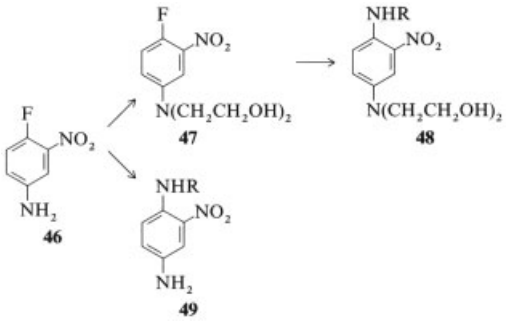

The development of a commercial synthesis for 4-fluoro-3-nitroaniline (**46**), involving low-temperature nitration of 4-fluoroaniline with fuming nitric acid in oleum [112], made possible the manufacture of **48** and other similar trisubstituted derivatives in a pure state. Hydroxyethylation of **46** to give the tertiary amine **47**, followed by reaction with a primary amine, is a convenient route to a variety of blue-violet dyes **48** [113]. The dyes **48** where R is 2-hydroxyethyl, methyl, ethyl, or 2,3-dihydroxypropyl have all found commercial use.

The bluish red dye **49** ($R = CH_2CH_2OH$) is obtained by reaction of 4-fluoro-3-nitroaniline with monoethanolamine, while reaction with methylamine gives the similar dye **49** ($R = CH_3$) [113].

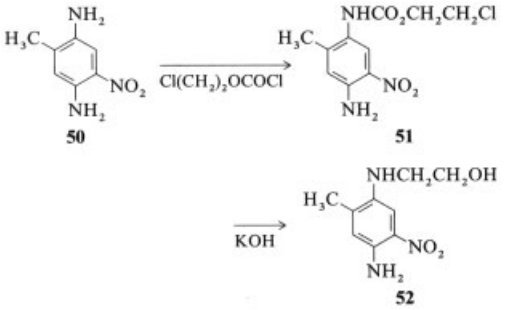

A method for the introduction of a single *N*-hydroxyethyl group onto the 5-amino group of a 2,5-diaminonitrobenzene, e.g., **50**, involves the reaction of the amine with 2-chloroethylchloroformate and treatment of the resulting carbamate **51** with potassium hydroxide in aqueous ethanol [114] to give the bluish red dye **52**. A similar reaction can be used to introduce a 2-hydroxyethyl group onto the 2-amino group, provided that the 5-amino group is first blocked by acetylation [115].

These reactions involve the intermediate formation of the aminonitrophenyloxazolidone and have been used in the synthesis of N^2,N^5-disubstituted-2,5-diaminonitrobenzenes [116].

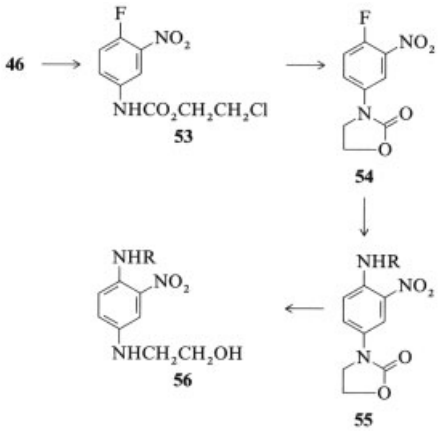

Reaction of 4-fluoro-3-nitroaniline (**46**) with 2-chloroethylchloroformate gives the carbamate **53**, which is cyclized to **54** by the action of sodium methoxide in methanol. The oxazolidone **54** is then reacted with a primary amine to give the intermediate **55** which, on alkaline hydrolysis, gives the violet diaminonitrobenzene dyes **56**, where R is alkyl or hydroxyalkyl.

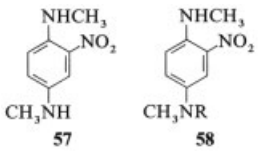

Two other commercially used blue-violet dyes **58**, where R = 2-hydroxyethyl or 2,3-dihydroxypropyl, are obtained by direct hydroxyalkylation of N^2,N^5-dimethyl-2,5-diaminonitrobenzene (**57**) with 2-bromoethanol in the presence of calcium carbonate [117], or with glycidol [118], respectively. The starting material **57** is made by reaction of the N^5-acetyl-N^5-methyl-N^2-tosyl compound with dimethyl sulfate, followed by conventional detosylation and hydrolysis [119]. It can also be obtained by methylation of the N^2,N^5-ditosyl derivatives of 2,5-diaminonitrobenzene with dimethyl sulfate, followed by detosylation with concentrated sulfuric acid [120], or by direct methylation with a mixture of formaldehyde and sulfuric acid at 40 – 120 °C [121].

8.3. Aminonitrophenol Hair Dyes

Five of the ten isomeric aminonitrophenols are used in hair dyes. Four of these, 2-amino-4-nitrophenol, 2-amino-5-nitrophenol, 4-amino-2-nitrophenol, and 4-amino-3-nitrophenol are commodity chemicals used in textile dye synthesis. The fifth, 2-amino-3-nitrophenol, is prepared by nitration of 2-acetoxyacetanilide with nitric acid in acetic anhydride at low temperature [122], followed by hydrolysis.

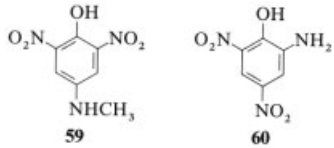

The first N-substituted aminonitrophenol to be employed in hair dyeing was *N*-methylisopicramic acid **59** which can be made by nitration of *N*-acetyl-*N*-methyl-4-aminophenol or, more conveniently, by nitration with nitric acid of a cold solution of 4-*N*-methylaminophenol in concentrated sulfuric acid [123]. Picramic acid (**60**) is made by the selective reduction of picric acid with sodium sulfide.

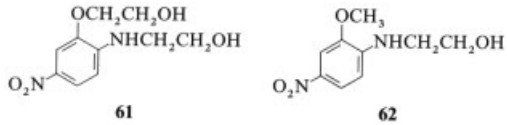

O,*N*-Bis(2-hydroxyethyl)-2-amino-5-nitrophenol **61**, which was originally listed in official compendia as the *N*,*N*-bis(2-hydroxyethyl) compound [124], is prepared by direct hydroxyethylation of 2-amino-5-nitrophenol with 2-chloroethanol in the presence of sodium hydroxide as an acid binder. A less convenient multistep synthesis has also been described [122].

A similar yellow dye, *N*-(2-hydroxyethyl)-2-amino-5-nitroanisole **62** is obtained by the action of 2-bromoethanol on 2-amino-5-nitroanisole [125].

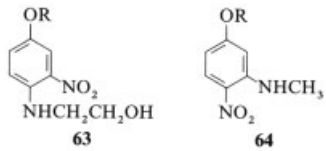

The red dye *N*-(2-hydroxyethyl)-4-amino-3-nitrophenol (**63**, R = H) is manufactured by the reaction of 4-amino-3-nitrophenol with 2-chloroethylchloroformate in dioxane with calcium carbonate as an acid binder. The intermediate carbamate is converted to **63** (R = H) by the action of aqueous potassium hydroxide, from which the potassium phenolate precipitates. The free phenol is precipitated by neutralizing an aqueous solution of the phenolate with hydrochloric acid [126].

The methyl ether **63** ($R = CH_3$) has also found use as an orange-red dye and is obtained by the action of ethylene oxide on 4-amino-3-nitroanisole in aqueous acetic acid at ambient temperature [127]. 1-(3-Methylamino-4-nitrophenoxy)propane-2,3-diol (**64**, $R = CH_2CHOHCH_2OH$) is reported to be more soluble in typical semipermanent dye bases than the previously used anisole derivative ($R = CH_3$). The yellow dye is made by the slow addition of 3-*N*-methylamino-4-nitrochlorobenzene to a solution of sodium hydroxide in glycerol at about 120 °C [128]. A similar reaction with ethylene glycol gives the analogue **64** ($R = CH_2CH_2OH$) [129].

8.4. Nitrodiphenylamine Hair Dyes

Two nitrodiphenylamine dyes have been found to be particularly useful in semipermanent hair colorants.

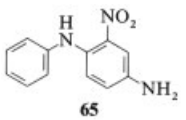

65

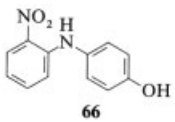

66

The red dye 4-amino-2-nitrodiphenylamine (**65**) is manufactured by the reaction of 4-fluoro-3-nitroaniline with aniline in the presence of calcium carbonate as an acid scavenger [130].

A yellow dye, 4-hydroxy-2′-nitrodiphenylamine (**66**), is obtained by the reaction of 2-chloronitrobenzene with 2-aminophenol [131].

8.5. Formulation with Nitro Hair Dyes

While relatively unreactive, the nitro dyes employed in semipermanent hair colorants cannot be regarded as completely inert if an adequate shelf-life for the product is to be attained.

A wide range of nonionic or anionic surfactants have been found to be compatible with the dyes, particularly fatty alkanolamides and ethoxylated fatty alcohols.

Dyeing is best achieved at alkaline pH, and the choice of alkalizing agent has been found to be important. Dyes having a secondary amino group ortho to the nitro group cannot be used in formulations containing ammonia due to a facile amine exchange reaction; e.g., the conversion of **68** to **67** occurs within weeks at ambient temperature [132].

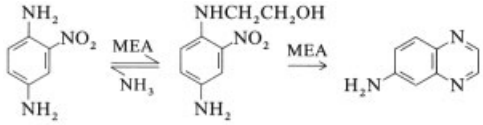

67 **68** **69**

A more remarkable reaction has been found to occur when diaminonitrobenzenes are stored in the presence of monoethanolamine (MEA). An amine exchange reaction is followed by a series of oxidation/reduction steps resulting in the formation of 6-aminoquinoxalines **69** [132]. The reaction appears to be quite general and occurs with 2,4- and 3,4-diaminonitrobenzenes as well as the 2,5-isomers.

These reactions can be avoided by the use of other alkanolamines such as diethanolamine and aminomethylpropanol [133].

The problem of poor solubility of the nitro dyes in cosmetically acceptable lotion bases has been met in part by the use of dyes bearing hydroxyalkyl substituents. In particular, the dye N^2,N^5-dimethyl-N^5-(2,3-dihydroxypropyl)-2,5-diaminonitrobenzene (**58**) ($R = CH_2CHOHCH_2OH$) can cosolubilize similar dyes such as those of structure **48** where R is alkyl or hydroxyalkyl [134].

Some nitro dyes have been found to be subject to air oxidation. Antioxidants, particularly ascorbic acid or erythorbic acid, which do not reduce the nitro group, are generally added to the formulations [135].

9. Biologically Active Nitro Dyes

A whole series of compounds with nitro dye character are used in plant protection or have

pharmaceutical activity. Of the hydroxynitro compounds, 3,5-dinitro-*o*-cresol is used as a herbicide. Alkylanilines containing nitro groups (e.g., **70**) [136] and *N*-alkyl derivatives of nitro-*m*-phenylenediamines (e.g., **71**) [137] also exhibit herbicidal effects.

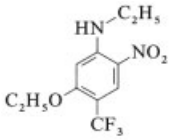

70

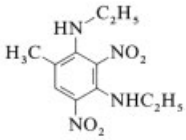

71

Nitro-*m*-phenylenediamine compounds with *N*-(aminoalkyl) substituents such as **72** are suitable for the therapy and prophylaxis of various viral infections [138].

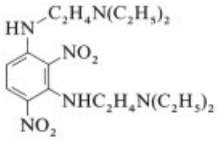

72

2-Nitro-4-trifluoromethyldiphenylamines bearing a substituted alkyl ether group in the 4′-position are used as herbicides [139], for example **73**.

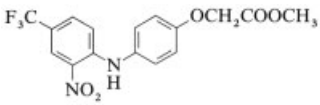

73

The introduction of a thiophosphate ester into the 4′-position gives diphenylamine compounds with insecticidal and miticidal properties [140]. 2,4-Dinitro-6-trifluoromethyldiphenylamine compounds with fluorine-containing substituents in the second phenyl ring are used as algicides for the protection of ships' hulls [141], as herbicides (e.g., **74**) [142], and as insecticides [143].

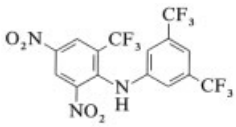

74

2,6-Dinitro- and 2,4,6-trinitrodiphenylamine compounds have fungicidal properties [144]. Pyridylnitrophenylamines bearing trifluoromethyl groups are used as insecticides [145]. An example is:

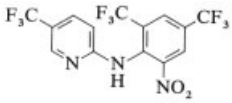

75

The corresponding pyrimidyl compounds are suitable for the same purpose [146].

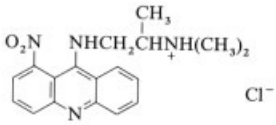

76

Nitroacridine compounds such as **76** inhibit tumor formation [147].

10. Nitroso Dyes

Hydroxynitroso compounds are formed by the action of nitrous acid on phenols and naphthols. These nitroso compounds are not themselves dyes, but are capable of forming metal complexes that are either pigments or, if the starting compound bears hydrophilic groups, are acid dyes.

The iron(II) complex of 1-nitroso-2-naphthol (Pigment Green B) is used for wallpaper and disperse paints because of its lightfastness [148].

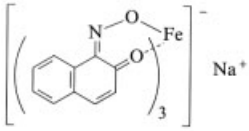

77

The pigment properties of **77** are improved by the addition of dimethylphenylbenzylammonium chloride in an alkaline soda medium before complex formation [149]. Pigment Green B absorbs light of wavelength 650 – 850 nm, λ_{max} = 730 – 750 nm. It is therefore suitable for the production of recording inks that are read

by using IR radiation [150] and for data storage with a semiconductor laser [151]. Pigment Green B is also used for coloring rubber [152]. The toxicity of Pigment Green B is low: LD_{50} (oral) = 5000 mg/kg. Skin and eye irritation have not been observed.

The nickel complex of 1-nitroso-2-naphthol is also in use [153]. It is known commercially as Monarch Gold X 3306 (Hercules). Naphthol Green B, the sodium salt of the iron complex of 1-nitroso-2-naphthol-6-sulfonic acid, is suitable for dyeing wool, silk, and paper. Its commercial name is Basovit Green 971 (BASF). The dye is also used for the production of IR-absorbing inks [154] and polarizing filters [155]. When precipitated with barium salts, Naphthol Green B is used for coloring paints.

The iron complex of 2-nitroso-1-naphthol-4-sulfonamide is used for dyeing and printing polyamide and polyurethane fibers in green shades [156].

1-Nitroso-2-hydroxynaphthalene-3-carboxylic acid amides whose amide groups have cationic substituents can be transformed by metallization into lightfast cationic complex dyes, which are suitable for dope dyeing of polyacrylonitrile fibers and for dyeing leather [157].

Heterocyclic compounds are also used for the production of nitroso dyes, examples of which include iron complexes of 5-nitroso-6-hydroxy-quinoline [158], 6-hydroxy-7-nitrosobenzothiazole [159], 1-methyl-4-nitroso-5-hydroxybenzimidazole [160], and 3-chloro-6-hydroxy-7-nitrosoindazole [161].

Open-chain compounds obtained by nitrosation of acetoacetic acid arylamides also belong to the nitroso dyes. Particularly useful pigments are obtained if the nitrosated acetoacetic acid arylamide is first converted to the oxime with hydroxylamine and the nickel complex is then formed [162]. The complex **78** is a lightfast yellow pigment for printing inks, varnishes, or plastics.

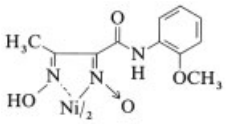

78

Commercial products: Paliotol Yellow 3 R, Luconyl Yellow 3 R (BASF).

11. References

General References

1. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 1,** Academic Press, New York 1952, pp. 401 – 408.
2. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin 1964, pp. 242 – 246.
3. H. Zollinger: *Color Chemistry, Syntheses Properties and Applications of Organic Dyes and Pigments,* Verlag Chemie, Weinheim, New York 1987, pp. 83 – 84.

Specific References

4. Bayer, DE-OS 2 342 723, 1973 (A. Brack). Ciba-Geigy, DE-OS 2 945 686, 1978 (R. Bäuerle).
5. Canon, JP 55 07 856, 1978; JP 58 174 466, 1982.
6. Japan Broadcasting, JP 62 32 090, 1985.
7. Honey Chem., JP 73 05 085, 1970.
8. E. Merian, *Angew. Chem.* **72** (1960) 766.
9. ICI, GB 1 212 992, 1967 (A. H. Berrie, R. Windle).
10. BASF, DE-OS 1 569 673, 1967 (G. Lange).
11. BASF, DE-AS 1 599 360, 1966 (W. Fabian et al.).
12. Hoechst, DE 158 543, 1904 (B. Homdke, J. Erber). BASF, DE 212 594, 1908 (E. Fussenegger); 220 630, 1909.
13. Hoechst, DE-OS 2 613 425, 1976 (M. Hähnke, T. Papenfuhs).
14. VEB Chemiekombinat Nünchritz, DE-OS 2 908 204, 1978 (B. Hirsch, G. Horn, M. Reuther); DE-OS 2 908 205, 1978 (B. Hirsch, G. Horn, H. Reuther).
15. Hoechst, DE-OS 2 413 515, 1974 (R. Mohr, K. Löhe).
16. Hoechst, DE-OS 1 941 376, 1969 (K. Löhe).
17. K. Nishida, T. Ohtsu, S. Tanaka, *J. Soc. Dyers Colour.* **89** (1973) 335 – 336.
18. Bayer, DE-OS 2 633 811, 1976 (J. George, J. Repplinger).
19. Sumitomo, JP 57 40 445, 1980.
20. J. J. Kulagowski, C. W. Rees, *Synthesis* 1980, no. 3, 215.
21. Monsanto, EP 148 145, 1983 (H. L. Merten).
22. Bayer, DE-AS 2 148 223, 1971 (J. Blahak).
23. Zaklady Przemyslu Barnikow "Organika Boruta," PL 129 902, 1981 (M. Effenberg et al.).

24. British Celanese, GB 237 943, 1923; 239 470, 1923. Hoechst, DE-OS 2 553 566, 1975 (W. Mohler, H. Wagner); DE-OS 2 648 944, 1976 (W. Mohler, H. Wagner). M. Day, A. T. Peters, *J. Soc. Dyers Colour.* **83** (1967) 137 – 143.
25. A. T. Peters, *J. Soc. Dyers Colour.* **89** (1973), 385 – 388. R. S. Asquit, H. S. Blair, A. C. Hamill, *J. Soc. Dyers Colour.* **90** (1974) 185 – 188.
26. R. S. Asquit, I. Bridgeman, A. T. Peters, *J. Soc. Dyers Colour.* **81** (1965) 439 – 441.
27. U. Keihei, M. Takagi, JP 54 14 994, 1977. K. Ueno, JP 54 151 994, 1978.
28. G. E. Pacey, Y. P. Wu, B. P. Bubnis, *Synth. Commun.* **11** (1981) no. 4, 323 – 328.
29. Hoechst, DE-OS 3 618 219, 1986 (W. Brodt, T. Papenfuhs).
30. V. L. Plakidin, SU 833 956, 1977.
31. F. Vyskocil, J. Simandl, CS 166 881, 1970. F. Vyskocil et al., CS 191 522, 1976. Martin Luther Universität Halle, DD 251 359, 1986 (U. Koch et al.).
32. Moriroku Shoji, JP 58 31 183, 1981; JP 58 126 382, 1982.
33. F. Vyskocil, CS 164 655, 1973.
34. Hoechst, DE-OS 3 530 338, 1985 (W. Brodt, T. Papenfuhs).
35. Eastman Kodak, US 3 729 493, 1967 (D. H. Wilcox).
36. Ciba, DE-AS 1 143 483, 1959 (E. Reich).
37. Cassella, DE-AS 1 135 416, 1961 (K. Mix, C. Heid).
38. Produits Chimiques Ugine Kuhlmann, DE-OS 3 001 266, 1979 (M. Rau).
39. Produits Chimiques Ugine Kuhlmann, FR 2 471 400, 1979 (M. Rau).
40. J. Dvorak, T. Kudrnova, CS 161 477, 1973.
41. Hoechst, DE 263 655, 1911.
42. Ciba-Geigy, EP 22 067, 1979 (K. Albiez).
43. Sandoz, DE-OS 2 035 075, 1969 (D. van Assche, J. Frauenknecht).
44. Sandoz, DE-OS 2 121 304, 1970 (D. van Assche).
45. Ciba-Geigy, DE-OS 2 218 446, 1971 (J. Brünisholz, F. Beffa).
46. Ciba-Geigy, DE-OS 2 235 250, 1971 (J. Brünisholz, F. Beffa).
47. Hoechst, DE-OS 2 844 610, 1978 (O. Arndt, W. Tronich).
48. Osrodek Badawczo-Rozwojowy Przemyslu Barnikow Organika, PL 109 498, 1978 (M. Muszynski).
49. ICI, GB 1 517 105, 1976 (M. Yelland).
50. Hoechst, DE 838 045, 1949 (H. Tampke); 838 046, 1949 (H. Tampke)
51. Hoechst, DE-OS 2 613 425, 1976 (M. Hähnke, T. Papenfuhs).
52. Ciba-Geigy, DE-OS 2 516 557, 1974 (R. Begrich).
53. Dynapol, DE-OS 2 754 485, 1976 (J. D. Dawson et al.).
54. Sandoz, DE-OS 1 644 334, 1964 (R. Altersmatt, R. Entschel, C. Müller).
55. F. Ullmann, *Justus Liebigs Ann. Chem.* **355** (1907) 328.
56. Bayer, DE-OS 2 342 723, 1973 (A. Brack).
57. Geigy, BE 692 928, 1966.
58. Ciba-Geigy, CH 520 746, 1968 (C. Zickendraht)
59. Nippon Kayaku, JP 14 70669, 1966.
60. Ciba-Geigy, DE-OS 2 516 440, 1974 (K. Burdeska, H. Schwander).
61. BASF, DE-AS 1 045 361, 1956 (W. Seibert, K. G. Roessler). Hoechst, DE-OS 2 214 699, 1972 (A. Sieber).
62. VEB Chemiekombinat Bitterfeld, DD 236 105, 1985 (G. Wallaschek et al.).
63. V. N. Knyazev, V. N. Drozd, T. YA. Mozhaeva, *Zh. Org. Khim* **16** (1980) no. 4, 876 – 882.
64. VEB Chemiekombinat Bitterfeld, DD 150 622, 1980 (E.-A. Jauer et al.).
65. VEB Chemiekombinat Bitterfeld, DD 200 890, 1981 (E.-A. Jauer et al.).
66. R. R. Gupta, G. S. Kalwania, M. Kumar, *J. Heterocycl. Chem.* **21** (1984) no. 3, 893 – 896.
67. IG Farben, DE 545 859, 1929 (O. Siebert, W. Benade); DE 694 565, 1937 (R. Fleischhauer); DE 728 376, 1939 (R. Fleischhauer); DE 737 436, 1937 (R. Fleischhauer).
68. Bayer, DE-OS 2 630 991, 1976 (H. Hugl et al.). Osrodek Bradawczo-Rozwojowy Przemyslu Barnikow Organika, PL 112 469, 1979 (G. Przybylski, J. Jablonski).
69. Yorkshire Chem., GB 1 364 291, 1971 (J. F. Dawson, K. Bramham).
70. ICI, GB 1 423 648, 1972 (C. W. Greenhalgh et al.).
71. Bayer, DE-OS 2 630 989, 1976 (H. Hugl et al.); DE-OS 2 630 987, 1976 (K. Trautner, K.-H. Schündehütte).
72. Bayer, DE-OS 2 630 990, 1976 (H. Hugl et al.).
73. Bayer, DE-OS 2 630 988, 1976 (K. Trautner, K.-H. Schündehütte).
74. Sandoz, US 3 884 990, 1969 (R. Altermatt).
75. ICI, GB 1 423 999, 1972 (C. W. Greenhalgh et al.).
76. Sandoz, DE-OS 2 719 828, 1976 (R. Altermatt).
77. Sandoz, DE-OS 2 050 900, 1969 (R. Altermatt).

78. ICI, GB 1 429 179, 1972 (C. W. Greenhalgh, D. A. S. Phillips).
79. ICI, DE-OS 2 224 116, 1971 (D. R. A. Ridyard); DE-OS 2 635 820, 1975 (W. F. Skuce); DE-OS 2 710 465, 1976 (D. R. A. Ridyard, M. Yelland). Osrodek Badawczo-Rozwojowy Przemyslu Barnikow Organika, PL 119 751, 1979 (C. Przybylski, J. Gmaj).
80. Colour Chem., IN 147 672, 1977.
81. Bayer, DE-OS 2 307 178, 1973 (K.-H. Schündehütte, K. Trautner); Bayer, DE-OS 3 048 212, 1980 (K. Trautner).
82. ICI, GB 2 021 617, 1978 (D. R. A. Ridyard).
83. Osrodek Badawczo-Rozwojowy Przemyslu Barnikow Organika, PL 119 756, 1979 (C. Przybylski, J. Gmaj). J. Szadowski, C. Przybylski, *Dyes Pigm.* **5** (1984) no. 1, 49 – 64.
84. Osrodek Badawczo-Rozwojowy Przemyslu Barnikow Organika, PL 108 921, 1978 (C. Przybylski).
85. Bayer, DE-OS 2 307 179, 1973 (K. H. Schündehütte, K. Trautner, H. Nickel). ICI, DE-OS 2 619 886, 1975 (D. R. A. Ridyard, D. Thorp).
86. ICI, DE-OS 2 620 121, 1975 (D. R. A. Ridyard, D. Thorp).
87. ICI, DE-OS 2 619 885, 1975 (D. R. A. Ridyard, D. Thorp).
88. ICI, GB 1 517 105, 1976 (M. Yelland).
89. ICI, DE-OS 2 421 579, 1973 (D. R. A. Ridyard).
90. Bayer, DE-OS 2 422 465, 1974 (R. Sommer et al.).
91. Bayer, DE-OS 2 307 179, 1973 (K.-H. Schündehütte et al.).
92. Sandoz, DE-OS 2 249 643, 1971 (D. van Assche, H. Uehlinger).
93. Bayer, DE-OS 2 557 504, 1975 (H. Hugl et al.); DE-OS 2 536 489, 1975 (H. Hugl et al.).
94. Bayer, DE-OS 2 536 490, 1975 (H. Hugl et al.). ICI, DE-OS 2 714 739, 1976 (D. Brierley, D. R. A. Ridyard).
95. ICI, GB 1 483 909, 1974 (B. Lamble, D. R. A. Ridyard).
96. ICI, GB 1 558 067, 1976 (D. Brierley, D. R. A. Ridyard).
97. Hoechst, DE-OS 3 510 179, 1985 (M. Segal, M. Kunze); DE-OS 3 510 180, 1985 (M. Segal, M. Kunze); DE-OS 3 528 049, 1985 (H.-U. von der Eltz).
98. Hoechst, EP 271 883, 1986 (H. Springer et al.).
99. Hoechst, DE-OS 3 517 655, 1985 (M. Segal, M. Kunze).
100. Hoechst, EP 201 868, 1985 (M. Segal, M. Kunze); DE-OS 3 517 547, 1985 (M. Segal, M. Kunze).
101. Bayer, DE-OS 3 317 383, 1983 (H. Jäger et al.). Hoechst, DE-OS 3 801 997, 1988 (R. Hähnle et al.).
102. Colour Chem., IN 148 676, 1978.
103. Hoechst, DE-OS 2 616 405, 1976 (H. Fuchs, K. Filzinger); DE-OS 2 618 020, 1976 (H. Fuchs, K. Filzinger).
104. Sandoz, DE-OS 2 055 962, 1969 (R. Altermatt, W. Steinemann).
105. Sandoz, DE-OS 2 055 963, 1969 (R. Altermatt, W. Steinemann).
106. Sandoz, DE-OS 2 210 047, 1971 (X. Pfister).
107. Teijin K. K., JP 53 47 439, 1976.
108. Unilever, GB 707 618, 1952 (W. M. J. Eckardt); GB 741 334, 1954 (W. M. J. Eckardt).
109. C. Zviak: *The Science of Hair Care,* Marcel Dekker, New York 1986, p. 235.
110. J. F. Corbett, *J. Soc. Cosmet. Chem.* **35** (1984) 297.
111. Wella AG, US 4 835 314, 1987 (E. Konrad, T. Clausen, H.-J. Braun, H. Mager).
112. Clairol, US 3 586 719, 1968 (M. Bil); US 3 944 612, 1970 (M. Bil).
113. Clairol, US 3 632 582, 1968 (M. Bil).
114. L'Oreal, GB 2 085 484, 1981 (C. Monnais, J. Cotteret, A. Bugaut).
115. L'Oreal, GB 1 104 970, 1965 (G. Kalopissis, A. Bugaut).
116. L'Oreal, GB 2 165 258, 1985 (G. Lang, A. Junino).
117. L'Oreal, FR 1 454 313, 1965 (G. Kalopissis, A. Bugaut, A. Ormancey).
118. L'Oreal, GB 2 150 148A, 1984 (J. F. Grollier, J. Cotteret, A. Bugaut, A. Genet).
119. Clairol, US 3 274 249, 1964 (W. H. Brunner, A. Halasz).
120. Gillette, GB 1 228 604, 1965 (J. F. Corbett, M. W. Parsons, J. M. Bird).
121. Clairol, US 3 591 638, 1968 (A. Halasz).
122. L'Oreal, US 4 337 061, 1976 (A. Bugaut, P. Andrillon).
123. Gillette, GB 1 172 635, 1968 (J. F. Corbett).
124. CTFA: *Cosmetic Ingredient Dictionary,* Cosmetic Toiletry and Fragrance Assoc., Washington, D.C. 1982.
125. L'Oreal, US 4 420 637, 1979 (A. Bugaut, J. J. Vanderbossche).
126. L'Oreal, US 4 125 601, 1977 (A. Bugaut, M. Andrillon).
127. A. J. A. Medici, L. Owen, C. Sflomos, *J. Chem. Soc. Perkin Trans. I* 1977, 2258.

128. L'Oreal, GB 2 081 735, 1981 (A. Bugaut, P. Andrillon).
129. L'Oreal, FR 2 349 325, 1974 (A. Bugaut, P. Andrillon).
130. Clairol, US 3 959 377, 1972 (M. Bil).
131. Clairol, US 3 950 127, 1972 (A. Halasz, D. Cohen).
132. M. Bil, J. F. Corbett, *Dyes Pigm.* **2** (1981) 215.
133. Clairol, US 3 951 589, 1975 (G. Alperin, R. Feinland, W. Howard).
134. L'Oreal, GB 2 150 148, 1984 (J. F. Grollier, J. Cotteret, A. Bugaut, A. Genet); GB 2 164 959, 1985 (J. F. Grollier, J. Cotteret, A. Bugaut, A. Genet).
135. Clairol, US 3 488 138, 1962 (S. Iscowitz).
136. US Borax, DE-OS 2 136 617, 1971 (L. P. Strong, D. L. Hunter, C. W. Le Fevre).
137. US Borax, DE-OS 2 136 616, 1971 (D. L. Hunter et al.).
138. Hoechst, DE-OS 1 545 868, 1965 (E. Winkelmann, M. Schorr).
139. ICI Australia, EP 13 144, 1978 (A. Serban et al.). Nippon Soda, JP 55 149 234, 1979. Kanesho, FR 2 481 273, 1980 (S. Someya et al.). Toyo Soda, JP 57 206 644, 1981. Nippon Soda, JP 58 65 254, 1981.
140. Kanesho, JP 57 185 296, 1981.
141. ICI, GB 1 523 884, 1976 (A. F. Hawkins, R. D. Bowden).
142. ICI, GB 1 544 078, 1975 (D. J. Collins et al.).
143. Bayer, DE-OS 2 823 168, 1978 (A. Hartmann et al.). Eli Lilly, EP 33 580, 1980 (A. J. Clinton); EP 60 951, 1981 (B. A. Dreikorn, K. E. Kramer); SU 984 402, 1981 (A. J. Clinton); US 4 366 168, 1981 (A. J. Clinton, G. O. P. O'Doherty).
144. Eli Lilly, US 4 018 896, 1974 (R. G. Jones); US 4 152 460, 1978 (B. A. Dreikorn). Ishihara Sangyo Kaisha, JP 58 113 151, 1981.
145. Ishihara Sangyo Kaisha, JP 57 58 667, 1980; JP 57 126 475, 1981; JP 57 126 476, 1981.
146. Ishihara Sangyo Kaisha, JP 56 65 804, 1979.
147. Politechnika Gdanska, DE-OS 2 713 103, 1976 (A. Ledochowski et al.).
148. BASF, DE 356 973, 1921. H. Pfitzner, *Angew. Chem.* **62** (1950) 243.
149. J. P. Kovalenko et al., SU 297 660, 1966.
150. Mitsui Toatsu, JP 58 13 676, 1981.
151. Nippon Columbia, JP 60 36 188, 1983.
152. Toa Gomu Kogyo K.K., JP 31 54485; 35 04185; 47 04285, 1985.
153. Hercules, CH 565 842, 1972.
154. Mitsui Toatsu Chem., JP 58 11 568, 1981.
155. Universität für Angewandte Mathematik, SU 1 211 276, 1982.
156. Bayer, DE-OS 1 569 746, 1967 (W. Scholl).
157. Bayer, DE-AS 1 089 102, 1957 (H. Kleiner, O. Bayer).
158. Leningrad Technological Inst., SU 617 465, 1976 (T. M. Timoleeva et al.); SU 870 423, 1979 (A. V. Eltsov et al.).
159. Leningrad Technological Inst., SU 759 515, 1978 (A. V. Eltsov et al.).
160. Leningrad Technological Inst., SU 789 521, 1978 (L. A. Aleksandrova et al.).
161. FMC, DE-AS 1 098 654, 1954 (R. F. M. Sureau, G. R. H. Muigasson).
162. BASF, DE-AS 1 252 341, 1966 (P. Dimroth); DE-OS 2 159 440, 1971 (G. Henning).

Nitroaniline → Nitro Compounds, Aromatic
Nitrobenzene → Nitro Compounds, Aromatic
Nitrocellulose → Explosives

Nitro Compounds, Aliphatic

SHELDON B. MARKOFSKY, W. G. Grace & Co., Columbia, Maryland 21044, United States

1. **Introduction** 1
2. **Properties** 1
3. **Production** 2
4. **Quality Specifications and Analysis** . 4
5. **Storage and Transportation** 5
6. **Uses of Nitroalkanes and their Derivatives** 5
   6.1. **Nitromethane and Derivatives** 5
   6.2. **Nitroethane and Derivatives** 6
   6.3. **1-Nitropropane and Derivatives** 7
   6.4. **2-Nitropropane and Derivatives** 7
7. **Toxicology and Occupational Health** . 8
8. **References** 9

1. Introduction

Aliphatic nitro compounds (RNO_2), also called nitroalkanes and nitroparaffins, are isomeric with the chemically different alkyl nitrites (RONO). As shown by the following equilibrium involving a secondary nitroalkane, the nitroalkane and nitronic acid are tautomers, sharing a common nitronate anion.

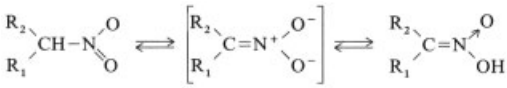

The nitroalkane is less acidic than the corresponding nitronic acid and, in most cases, the equilibrium lies far to the left. The acidity of the primary and secondary nitroalkanes is due partly to the electron-withdrawing effect of the nitro group and the resonance stability of the nitronate anion.

2. Properties

Physical Properties. The four nitroalkanes of greatest industrial significance, nitromethane [*75-52-5*], CH_3NO_2; nitroethane [*79-24-3*], $CH_3CH_2NO_2$; 1-nitropropane [*108-03-2*], $CH_3CH_2CH_2NO_2$; and 2-nitropropane [*79-46-9*], $CH_3CH(NO_2)CH_3$ are all colorless liquids when pure. Nitroalkanes are only slightly soluble in water. Due in part to its polarity, the nitro group causes a large increase in flash point and boiling point compared to the corresponding hydrocarbon. Nitroalkanes are often useful because they readily form azeotropes with many organic solvents such as alcohols, ketones, and hydrocarbons [1]. The physical properties of the four basic nitroalkanes are listed in Table 1 (see next page).

Chemical Properties. Primary and secondary nitroalkanes are excellent building blocks for the synthesis of more complex molecules, since the readily formed nitronate anion undergoes many useful condensation reactions. Especially important is the condensation of these anions with aldehydes and ketones (Henry reaction) [3–8], and the Michael reaction, in which the nitronate anion reacts with α,β-unsaturated carbonyl compounds, nitriles, esters, and other activated alkenes [9–12]. The following is an example of the Henry reaction:

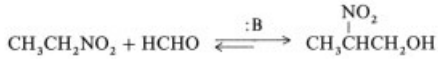

A typical Michael addition is shown as follows:

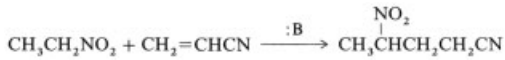

Figure 1 shows how nitroalkanes can be employed to make more complicated molecules.

The nitro group can subsequently be transformed into a variety of other functional groups, as illustrated in Figure 2. Thus, nitroalkanes allow the precise placement of these functional groups into complex structures. The chemistry depicted in Figures 1 and 2 (see next pages) is discussed in detail in [2, 13–18].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_401

Table 1. Physical properties of the four common nitroalkanes [2]

Property	Nitromethane	Nitroethane	1-Nitropropane	2-Nitropropane
M_r	61.041	75.068	89.095	89.095
bp (101.3 kPa), °C	101.20	114.07	131.18	120.25
Aqueous azeotrope				
bp (101.3 kPa), °C	83.59	87.22	91.63	88.55
wt % nitroalkane	76.4	71.0	63.5	70.6
mp (101.3 kPa), °C	−28.55	−89.52	−103.99	−91.32
Density, g/cm^3				
at 20 °C	1.138	1.051	1.001	0.988
at 30 °C	1.124	1.039	0.991	0.977
n_D^{20}	1.38188	1.39193	1.40160	1.39439
Vapor pressure at 25 °C, MPa	4.89	2.79	1.36	2.29
Vapor density (air = 1)	2.1	2.58	3.06	3.06
Evaporation rate (BuOAc = 1)	1.39	1.21	0.88	1.10
Evaporation number (diethyl ether = 1)	9	11	16	10
Heat of combustion, kJ/mol (liq.) at 25 °C	−709	−1363	−2017	−2001
Heat of formation, kJ/mol (liq.) at 25 °C	−113	−142	−168	−181
Heat of vaporization, kJ/mol (liq.)				
at 25 °C	38.3	41.6	43.4	41.4
at *bp*	34.4	38.0	38.5	36.8
Specific heat capacity at 25 °C, Jmol^{-1} K^{-1}	106	138.6	175.7	175.3
Solubility in water, wt %				
at 25 °C	11.1	4.7	1.5	1.7
at 70 °C	19.3	6.6	2.2	2.3
Solubility of water in nitroalkane, wt %				
at 25 °C	2.1	1.1	0.6	0.5
at 70 °C	7.6	3.0	1.7	1.6
Lower flammability, wt %	7.1	3.4	2.2	2.6
Upper flammability, wt %				11
Flash point, °C				
Tag open cup	44.4	41.1	48.9	37.8
Tag closed cup	35	30.6	35.6	27.8
Ignition temperature, °C	418	414	421	428
Critical temperature t_c, °C (calculated)	315	324	340	326
Critical pressure p_c, MPa (calculated)	6.30	4.98	4.33	4.49
Critical density ϱ_c, g/cm^3 (calculated)	0.353	0.329	0.314	0.318
Surface tension, (N/cm) × 10^{-5} at 20 °C	37.48	32.66	30.64	29.87
Viscosity, mPa · s				
at 10 °C	0.731	0.769	0.972	0.883
at 20 °C	0.647	0.677	0.844	0.770
at 30 °C	0.576	0.602	0.740	0.677
Dielectric constant, at 30 °C	35.87	28.06	23.24	25.52
pH of 0.01 M aqueous solution at 25 °C	6.4	6.0	6.0	6.2

3. Production

For more than forty years, the four principal nitroalkanes have been produced by the high-temperature vapor-phase nitration of propane [19]. The process, which employs nitric acid as the nitrating agent, is based on a free radical reaction, in which the active species is the NO_2 radical. This procedure produces a nitroalkane mixture rich in nitropropanes. The process is used by Angus Chemical Company for bulk production of nitromethane, nitroethane, 1-nitropropane, and 2-nitropropane.

The reaction is carried out at 350 – 450 °C. At this temperature all participants in the reaction are gaseous. The reaction is fast but not too fast, so that good temperature control is still assured. The pressure is adjusted to 0.8 – 1.2 MPa, so that the reaction product leaving the reactor can be condensed without liquefying the hydrocarbon. The temperature of the exothermic reaction can be controlled in various ways:

1) By using an excess of propane to remove the heat, the molar ratio of propane to nitric acid being at least 4 : 1;

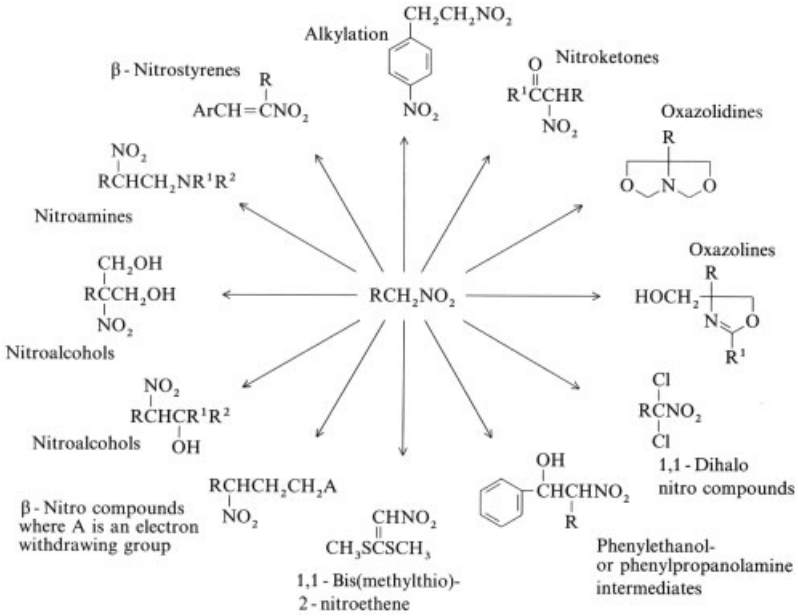

Figure 1. Building new chemicals with nitroalkanes

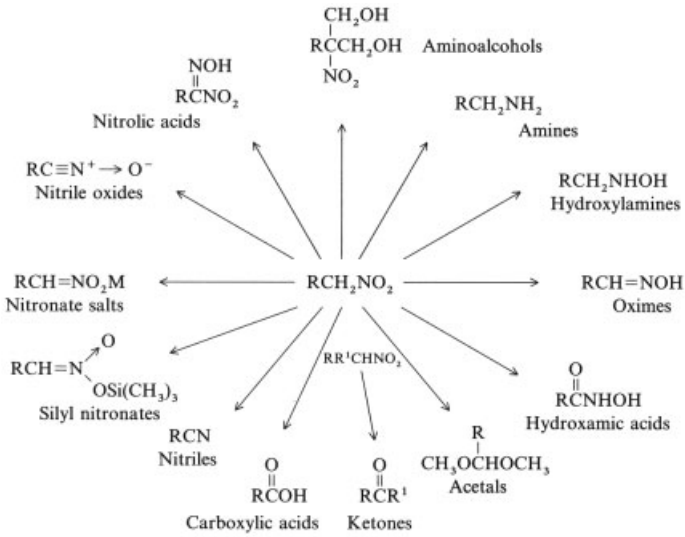

Figure 2. Functional group transformations

2) By spraying liquid nitric acid into the heated propane, the heat of reaction being used to vaporize nitric acid and to produce nearly adiabatic conditions;
3) By using 60 – 70 % nitric acid, thus producing a large volume of steam, which acts as an inert medium in the reactor; or
4) By controlling the temperature in the reactor via the residence time, the shorter the residence time the lower being the temperature. Short residence time is also important for maximum formation of nitroalkane and minimum formation of byproducts.

The vapor-phase nitration of propane by nitric acid proceeds mainly by a free radical mechanism. The decomposition of the nitric acid to ·OH and ·NO_2 or ·ONO, the postulated initiating step, requires temperatures above 350 °C.

By reaction of the OH radicals with alkanes, alkyl radicals can be formed, which then react further with NO_2 radicals to give nitroalkanes. However, the alkyl radicals can also react with HNO_3 to form nitroalkane and •OH and in that way continue the chain. The most important side reaction of the alkyl radical is the formation of nitrite, which is unstable at the reaction temperature and decomposes to nitric oxide and an alkoxy radical.

$$R\cdot + \cdot ONO \longrightarrow RONO$$

$$RONO \longrightarrow RO\cdot + NO$$

Alkoxy radicals can decompose with cleavage of the carbon – carbon bond and formation of alkyl radicals of lower molecular weight.

$$R'CH_2O\cdot \longrightarrow R'\cdot + CH_2O$$

This can explain the formation of nitroethane and nitromethane during the nitration of propane. Other byproducts of this process are alcohols, aldehydes, ketones, and related oxygen-containing derivatives.

The conversion of nitric acid to nitroalkanes in vapor-phase nitration is less than 50 %. An improved conversion can be achieved by adding a small amount of oxygen or halogen. This is not done in the industrial process, however, since addition of oxygen promotes the formation of large amounts of oxygen-containing byproducts, whereas with halogen addition, corrosion problems or difficulties with the processing of the reaction mixture can occur.

Most of the nitric acid is converted to NO, NO_2, N_2O, and N_2; NO and NO_2 are recovered. The overall loss of nitric acid is 20 – 40 %.

Excess propane is sometimes also recovered from the reactor liquid and reused, so that about 60 – 80 % of the propane reacts to give nitroalkanes.

Rapid cooling immediately after the nitration leads to liquefaction of the nitroalkanes and the oxygen-containing byproducts. The liquid and gas phases are separated, and after propane and nitrogen oxides have been recovered from the gas phase they are returned to the reactor. Some of the byproducts, e.g., formaldehyde, are removed by separation of the aqueous phase. The low-boiling oxygen-containing byproducts such as acetaldehyde or acetone, and nitrogen-containing byproducts such as acetonitrile, can be separated as distillation first runnings from the organic phase. The nitroalkanes so obtained are subjected to a chemical washing process (to remove dissolved higher-boiling oxygen-containing impurities) and then washed with water. The remaining water is removed by passing through a drying tower, and the individual nitroalkanes are then obtained as pure products by fractional column distillation.

More recently, a number of publications have been issued describing improved methods for the syntheses of nitroalkanes [20–25]. However, in the past few years only one major new nitroalkane plant has been placed in operation, the W. R. Grace & Co. facility in Deer Park, Texas. The new Grace process employs a mixture of ethane and propane for the hot-tube, free-radical process, and uses N_2O_4 as the source of NO_2. This process produces a nitroalkane mixture richer in nitromethane and nitroethane.

4. Quality Specifications and Analysis

The specifications for the four nitroalkanes produced at the Grace plant in Texas are given in Table 2. Purity and total nitroalkane content are assayed by capillary gas chromatography; water content is measured by typical Karl Fischer procedures. The acidity is determined by dissolving

Table 2. Product specifications of the four common nitroalkanes (W. R. Grace & Co.)

Specification	Nitromethane	Nitroethane	1-Nitropropane	2-Nitropropane
Purity, wt % (min.)	97	97∗	94	94
Total nitroalkanes, wt % (min.)	99	99	99	99
Acidity, wt % acetic acid (max.)	0.1	0.1	0.2	0.1
Water, wt % (max.)	0.1	0.1	0.1	0.1
Color, APHA (max.)	20	20	20	20
Relative density at 25 °C	1.124 – 1.129	1.042 – 1.047	0.997 – 0.999	0.984 – 0.988

∗ A special grade of nitroethane is available with > 99 % purity.

the nitroalkane in methanol and titrating with dilute sodium hydroxide with bromocresol green as indicator.

5. Storage and Transportation

Nitroalkanes can be safely stored and transported, but certain precautions should be observed with respect to their shock sensitivity. It is necessary to protect nitroalkanes from extremes of temperature and pressure, e.g., fire and shock, and from chemical contamination which could further sensitize them to these extremes. Although the four commercial nitroalkanes are relatively insensitive to detonation by shock at ordinary temperatures, their sensitivity increases with increasing temperature. The tendency of the nitroalkanes toward detonation is inversely related to chain length. Nitromethane presents the most serious shock-sensitivity hazard of the four common nitroalkanes.

Adiabatic compression also poses a special problem in the handling of nitromethane. Thus, the transfer of nitromethane via pumps and piping requires special conditions designed to avoid adiabatic compression [1].

In addition, special care must be taken to avoid forming dry alkali-metal salts of the nitroalkanes, especially in the case of nitromethane. For example, in the presence of a strong base, such as sodium hydroxide, nitromethane can form the sodium salt of methazonic acid (2-nitroacetaldehyde oxime) which, when dry, is very shock sensitive and can explode. Mixtures of nitromethane and an amine and/or heavy metal oxides, such as those of silver, lead, and mercury, can lead to violent decompositions and should be avoided.

Nitromethane is packaged in 55-gallon drums to a net weight of 500 lb. The drum headspace is filled with nitrogen to exclude moisture and to reduce the hazard of adiabatic compression during transportation. The drums are thin-walled so that they rupture easily in the event of high energy impact. Nitromethane drums should not be stacked, but stored on-end in a single layer.

Nitroethane, 1-nitropropane, and 2-nitropropane are also transported in 55-gallon drums under nitrogen. However, because these nitroalkanes are not as shock sensitive as nitromethane, there are no special restrictions on stacking drums.

Nitroethane, 1-nitropropane, and 2-nitropropane are also available neat in bulk quantities transported in tank wagons or rail cars. Bulk nitromethane presents special hazards and, in the United States, may not be transported undiluted in containers having a capacity greater than 110 gallons (ca. 416 L). Nitromethane, however, can be transported in bulk when mixed with any of a variety of approved diluents. The following diluents, along with their minimum content in wt %, have been approved for the transportation of nitromethane: 1,2-butylene oxide (40 %), cyclohexanone (25 %), 1,4-dioxane (35 %), methanol (45 %), 1,1,1-trichloroethane (50 %), 1-nitropropane (48 %), and 2-nitropropane (47 %). These bulk nitromethane mixtures are safely transported by road and ocean freight. All of the nitroalkanes should be kept away from oxidizing, corrosive, or sensitizing materials.

For more details on storage and transportation procedures, refer to a nitroalkane safety guide [1].

6. Uses of Nitroalkanes and their Derivatives

6.1. Nitromethane and Derivatives

One of the most important uses for nitromethane in which it is used without chemical modification is the stabilization of halogenated hydrocarbons. For example, small amounts of nitromethane (and sometimes nitroethane and/or 1-nitropropane) are widely used in industry to form stable noncorrosive mixtures with 1,1,1-trichloroethane that are used in vapor degreasing, dry cleaning, and for cleaning semiconductors and lenses. These nitroalkanes are useful for inhibiting corrosion on the interiors of tin-plated steel cans containing water-based aerosol formulations. Nitromethane is also employed to stabilize the halogenated propellants for aerosols.

Nitromethane is frequently employed as a polar solvent for cyanoacrylate adhesives and acrylic coatings. Nitromethane is also used for cleaning electronic circuit boards; nitroethane and nitropropanes are also used for this purpose. Nitromethane alone, and in mixtures with

methanol and other nitroparaffins, is used as a fuel by professional drag racers and hobbyists. Hobbyists use the nitromethane-based fuel primarily for radio-controlled aircraft. The explosives industry utilizes nitromethane in a binary explosive formulation. The liquid (nitromethane) and solid (inorganic nitrate) ingredients are safely transported and stored separately as standard commercial products. When mixed, an explosive more energetic than 60 % dynamite is formed. There are also commercial and military applications for nitromethane in shaped charges, which are often used for targeted undersea explosions and line trenching.

Nitromethane Derivatives. *Chloropicrin* [*76-06-2*], trichloronitromethane, Cl_3CNO_2 can be prepared by the reaction of nitromethane with sodium hypochlorite. This chlorinated nitro compound is an effective fungicide and nematocidal fumigant.

Tris(hydroxymethyl)nitromethane [*126-11-4*], [2-(hydroxymethyl)-2-nitro-1,3-propanediol], $(CH_2OH)_3CNO_2$ is obtained if three moles of formaldehyde react with nitromethane via the Henry reaction. This derivative is used as a biocide.

Tris(hydroxymethyl)aminomethane [*77-86-1*], [2-amino-2-(hydroxymethyl)-1,3-propanediol], $(CH_2OH)_3CNH_2$ is prepared by reduction of tris(hydroxymethyl)nitromethane. It is used industrially as a buffer and as a component in adhesives and resins.

The addition of two moles of formaldehyde to nitromethane by the Henry reaction gives di(hydroxymethyl)nitromethane $(CH_2OH)_2CHNO_2$, which can be converted to the corresponding amino compound, $(CH_2OH)_2CHNH_2$ (2-amino-1,3-propanediol) [*1794-90-7*]. This amine is employed in the synthesis of the X-ray contrast agent Iopamidol [26].

The brominated product $(CH_2OH)_2CBrNO_2$ (2-bromo-2-nitro-1,3-propanediol) [*52-51-7*], which can be derived from di(hydroxymethyl) nitromethane is a widely used biocide (Bronopol).

The reaction between benzaldehyde and nitromethane does not stop at the corresponding nitroalcohol; instead, dehydration to β-nitrostyrene, $C_6H_5CH{=}CHNO_2$ [(2-nitroethenyl)-benzene] [*102-96-5*], occurs. This nitroalkene has been used as a chain transfer agent, i.e., to lower the molecular weights of polymers in their free radical initiated synthesis.

Treatment of β-nitrostyrene with bromine, followed by dehydrobromination, gives bromonitrostyrene, $C_6H_5CH{=}CBrNO_2$ [(2-bromo-2-nitroethenyl)benzene] [*7166-19-0*], which is employed as a slimicide.

An important use of nitromethane is in the synthesis of the anti-ulcer drugs Nizatidine [27] and Ranitidine [28]. A commercial route to Ranitidine is shown below [29]:

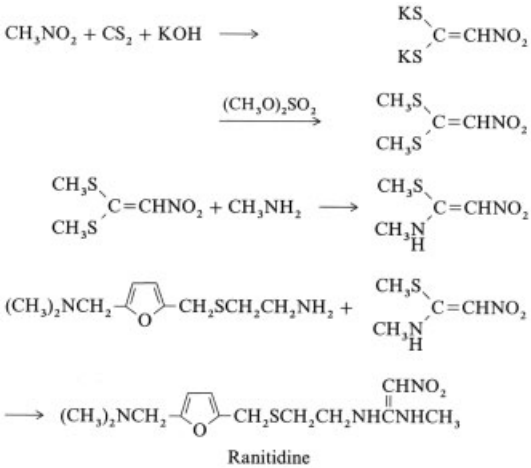

Finally, nitromethane is used in the preparation of the psychotropic agent Sulpiride [30].

6.2. Nitroethane and Derivatives

Nitroethane is employed as a solvent due to its excellent wetting properties and its ability to form azeotropes with many industrial solvents. For example, some printing inks contain nitroethane.

Nitroethane is used by the commercial blasting industry as either a fuel or sensitizer of water-based blasting agents. Nitroethane can be added to nitromethane to reduce its tendency to detonate when used as a fuel in internal combustion engines.

Nitroethane Derivatives. The Henry reaction between two moles of formaldehyde and nitroethane gives 2-methyl-2-nitro-1,3-propanediol [*77-49-6*], $CH_3CNO_2(CH_2OH)_2$ which is utilized as a biocide in cutting oils.

Catalytic hydrogenation of 2-methyl-2-nitro-1,3-propanediol affords the corresponding amine, 2-amino-2-methyl-1,3-propanediol [*115-69-5*]. After reaction with oleic acid, this aminoalcohol forms an oxazoline [31], which is used as a specialty cationic surfactant.

Nitroethane is used to make a plasticizer for solid rocket fuels [32]. The molecuar structure for this plasticizer is:

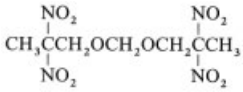

The antihypertensive drug Aldomet (methyl DOPA) is made using nitroethane. The synthesis follows the general route [33, 34]:

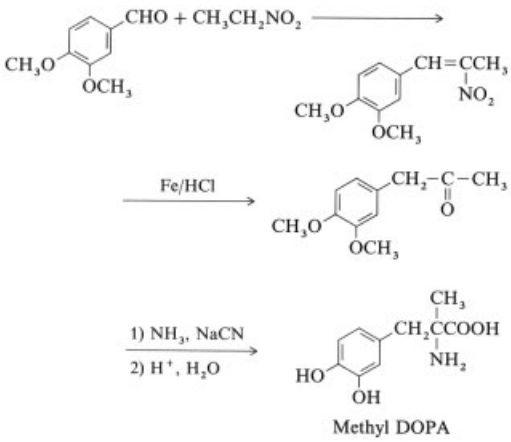

6.3. 1-Nitropropane and Derivatives

Like the other basic nitroalkanes, 1-nitropropane is employed as a solvent or cosolvent.

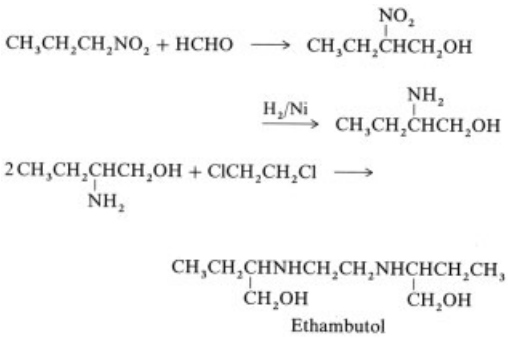

The most important use for 1-nitropropane is in the synthesis of Ethambutol, an antituberculosis drug [35].

Only the D-enantiomer of 2-amino-1-butanol is used in the final synthetic step.

Relatively small quantities of 1-nitropropane are used in the preparation of biocides [36].

6.4. 2-Nitropropane and Derivatives

A major use of 2-nitropropane is as an industrial solvent. It is used in vinyl inks that are employed in printing, flexography, and photogravure. The 2-nitropropane, often mixed with alcohols, dissolves a large number of resins such as epoxy, polyurethane, polyester, vinyl, urea–formaldehyde, and phenolic. These solvent–resin mixtures are used for coatings. For example, can-coating varnishes contain the nitroalkane because of its excellent wetting properties. Other applications of these 2-nitropropane formulations are in adhesives and in electrostatic paints. Some of the properties of 2-nitropropane that make it so versatile in many of these applications are: high polarity and flash point, appropriate evaporation rate, good wetting and azeotropic properties, and satisfactory resistivity.

2-Nitropropane Derivatives. The Henry reaction of 2-nitropropane with formaldehyde yields 2-methyl-2-nitro-1-propanol (NMP) [*76-39-1*].

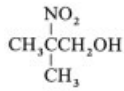

This nitroalcohol has been employed as an adhesive component to improve tire cord bonding [37].

Catalytic hydrogenation of NMP affords 2-amino-2-methyl-1-propanol (AMP) [*124-68-5*]. This aminoalcohol is a useful organic base for neutralizing and solubilizing applications. Other applications of AMP are:

1) AMP is used in toiletries, cosmetics, and in hair sprays, in which it neutralizes and solubilizes the carboxyl-containing polymers used in their formulation.
2) In conjunction with fatty acids, AMP is an excellent dispersing agent for powders and pigments, especially titanium dioxide.

3) AMP has been employed as a formaldehyde scavanger in melamine – formaldehyde, urea–formaldehyde, and phenol – formaldehyde resins.
4) AMP has also been used as a wetting agent, as an emulsifier in polishing waxes, and in textiles for permanent pleats.

Finally, 2-nitropropane has been utilized in the preparation of the experimental β-blocking drug, Bucindolol, [38–40], whose synthesis scheme shows how 2-nitropropane can be used to introduce a *tert*-butylamine functionality into a molecule:

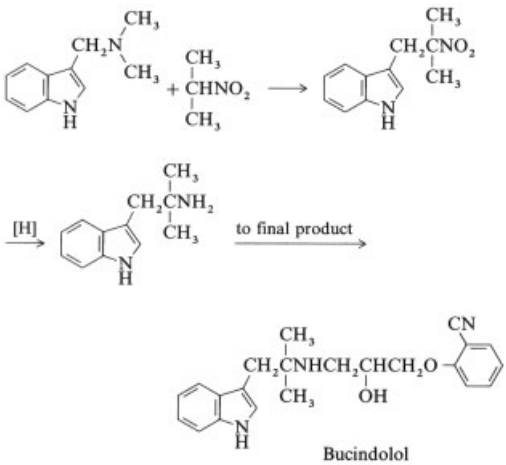

7. Toxicology and Occupational Health

The two most important health controls for nitroalkane exposure are adequate ventilation and prevention of skin contact.

In the United States, OSHA and ACGIH have set PE's and TLV's for the nitroalkanes [41,42]. The OSHA and ACGIH limits are shown in Table 3, together with MAK values.

Table 3. Industrial exposure limits for nitroalkanes, mL/m^3 (ppm)

Compound	PEL (OSHA)	TLV (ACGIH)	MAK
Nitromethane	100	100	100
Nitroethane	100	100	100
1-Nitropropane	25	25	25
2-Nitropropane	25	10	*

* Identified as a suspected human carcinogen; safe exposure limit has not been established.

Inhalation is the major industrial hazard of nitroalkane exposure. The reported effects of overexposure to vapors are headache, nausea, vomiting, and convulsions. Although chronic exposure to animals indicates some liver and kidney injury, no such injury has been reported in humans when exposures were maintained below the TLV levels.

Nitroalkane vapors can cause eye irritation at levels above the recommended TLVs. Since the nitroalkanes' odor detectabilities are poor, odor detection does not serve as a warning for overexposure. Therefore, prolonged exposure to vapors above the TLV necessitates the use of respirators. Such exposure may occur in operations that require entry into tanks or closed vessels and in emergency situations. Acute exposure to extremely high concentrations of 2-nitropropane vapors has resulted in serious injuries and, in some cases, death when workers were exposed in enclosed spaces without using recommended respiratory protection or adequate ventilation. For respiratory protection, supplied-air or self-contained breathing apparatus with a full facepiece should be used.

Nitroalkanes are mild skin irritants due to their solvent action but are not absorbed through the skin. Although irritation can occur from prolonged or repeated skin contact, no allergic or sensitization reactions have been reported.

The acute LD_{50} (rat, oral) of the nitroalkanes is as follows [43]:

Nitromethane	1210 ± 322 mg/kg
Nitroethane	1620 ± 193 mg/kg
1-Nitropropane	455 ± 75 mg/kg
2-Nitropropane	725 ± 160 mg/kg

The National Toxicology Program of the U.S. Department of Health and Human Services, The International Agency for Research on Cancer, and the American Conference of Governmental Industrial Hygienists list 2-nitropropane as a suspect carcinogen. The classification is based on studies in which prolonged exposure to 2-nitropropane was found to cause liver neoplasms in laboratory rats [44].

In 1979, an epidemiological study of workers exposed to 2-nitropropane was reported by the International Minerals and Chemical Corporation. The authors concluded that analysis of these data does not suggest any unusual cancer or other disease mortality pattern among this

group of workers [45]. There is no evidence that 2-nitropropane causes cancer in humans.

8. References

1. *Grace Nitroparaffins–Safety and Handling Guide,* W. R. Grace & Co., Lexington, MA.
2. G. O'Neill, S. Markofsky: *Grace Nitroparaffins Chemistry Guide,* W. R. Grace & Co., Lexington, MA.
3. S. Kanbe, H. Yasuda, *Bull. Chem. Soc. Jpn.* **41** (1968) 1444.
4. R. H. Wollenberg, *Tetrahedron Lett.* 1978, 3219.
5. T. I. Gubino, *Tr. Molodykh Uch. Sarat. Univ.* 1971, 177.
6. T. D. Zheved, K. V. Altukov in G. V. Nekrasova (ed.): *Sint. Issled. Nitrosoedin. Aminokislot* 1983, 3 – 5.
7. O. I. Rosumov et al., *Zh. Obsch. Khim.* **47** (1977) no. 3, 567.
8. Henkel and Cie. GmbH, DE 1 954 173, 1969 (R. Wessendorf).
9. D. A. White, M. M. Baizer, *Tetrahedron Lett.* 1973, 3597.
10. E. D. Bergmann, R. Corett, *J. Org. Chem.* **21** (1956) 107; **23** (1958) 1507.
11. E. A. Parfenov, A. R. Bekker, G. F. Kostereva, *Zh. Org. Khim.* **17** (1981) no. 8, 1591.
12. G. A. Smirnov, T. A. Klimova, V. V. Sevost'yanova, *Izv. Akad. Nauk SSSR, Ser. Khim.* 1981, no. 11, 2624.
13. D. Seebach, E. W. Colvin, F. Lehr, T. Weller, *Chimia* **33** (1979) 1.
14. N. Ono, A. Kaji, *Yuki Gosei Kagaku Kyokaishi* **38** (1980) no. 2, 115.
15. G. Rosini, R. Ballini, *Synthesis* 1988, 833.
16. M. Braun, *Nachr. Chem. Tech. Lab.* **33** (1985) no. 7, 598.
17. H. Feuer (ed.): *The Chemistry of the Nitro and Nitroso Groups,* Part 1, R. E. Krieger Co., New York 1981.
18. K. Torssell: *Nitrile Oxides, Nitrones, and Nitronates in Organic Synthesis,* VCH Publishers, New York 1988.
19. *Ullmann* 4th ed., **17,** 373.
20. Société Chemique de la Grande Paroisse, FR 2 272 975, 1971 (M. Lucquin, J. Dechaux).
21. Société Chemique de la Grande Paroisse, US 4 260 838, 1978; US 4 313 010, 1978 (P. Lhonore, B. Jacquinot, J. Quibel).
22. W. R. Grace & Co., US 4 469 904, 1983; US 4 524 226, 1983; US 4 517 393, 1983 (M. Sherwin, P. Wang).
23. W. R. Grace & Co., US 4 476 336, 1984; EP 85 328A, 1983 (M. Sherwin).
24. W. R. Grace & Co., EP 174 600A, 1988 (P. Wang).
25. Dow Chemical, US 4 421 940, 1982; US 4 431 842, 1982 (W. V. Hayes).
26. Savac AG, US 4 001 323, 1975 (E. Felder, R. S. Vitale, D. E. Pitre).
27. Eli Lilly and Co., US 4 375 547, 1980 (R. P. Pioch).
28. Allen and Hansbury Ltd., GB 1 565 966, 1976 (B. Price, J. Clitherow, J. Bradshaw).
29. Allen and Hansbury Ltd., US 4 128 658, 1977 (B. J. Price, J. W. Clitherow, J. Bradshaw).
30. Etudes Scientifiques et Industrielle Fr., FR 2 019 350, 1969 (M. Hashimoto, T. Kamiya).
31. J. A. Frump, *Chem. Rev.* **71** (1971) no. 5, 483.
32. E. Hamel et al., *Ind. Eng. Chem. Prod. Res. Dev.* **1** (1962) no. 2, 108.
33. Merck and Co. Inc., US 2 868 818, 1959 (K. Pfister, G. Stein).
34. Merck and Co. Inc., US 3 158 648, 1964 (R. Jones, K. Krieger, J. Lago).
35. R. G. Wilkinson, R. G. Shepherd, J. P. Thomas, C. Baughn, *J. Am. Chem. Soc.* **83** (1961) 2212.
36. IMC Chemical Group, Inc., US 4 088 817, 1977 (J. Hunsucker, R. Shelton).
37. Uniroyal Inc., US 3 598 690, 1967 (A. Danielson).
38. H. R. Snyder, L. Katz, *J. Am. Chem. Soc.* **69** (1947) 3140.
39. Mead Johnson & Co., US 4 234 595, 1979 (W. Kreighbaum, W. Comer).
40. Bristol Myers Co., DE 3 421 252, 1984 (W. Kreighbaum).
41. General Industry Safety and Health Standards in Occupational Safety and Health Act (OSHA), 29 CFR 1910.1000, 1989.
42. Threshold Limit Values and Biological Exposure Indices for 1986 – 87, American Conference of Governmental Industrial Hygienists.
43. E. Bingham, A. Robbins: "Health . . . 2-NP . . ." *DHHS (NIOSH) Publ. (U.S.)* **80-142** (1980) Oct.
44. Huntington Research Center, *HEW Publ. (NIOSH) (U.S.)* **210-75-0039** (1975).
45. M. Miller, G. Temple: *2-NP Mortality Epidemiological Study of the Sterlington, LA, Employees,* International Minerals and Chemical Corporation, Mundelein, IL 1979.

Nitro Compounds, Aromatic

Nitroanthraquinones → Anthraquinone Dyes and Intermediates ; Nitrobenzaldehydes → Benzaldehyde ; Nitrobenzoic acids → Benzoic Acid and Derivatives ; Nitro dyes → Nitro and Nitroso Dyes ; Nitronaphthalenesulfonic acids → Naphthalene Derivatives

GERALD BOOTH, Booth Consultancy Services, Thorpe House, Uppermill, Oldham OL3 6DP, United Kingdom

1. Introduction . . . 1
2. Nitration . . . 2
2.1. Nitrating Agents . . . 3
2.2. Reaction Mechanisms . . . 4
3. Nitro Aromatics . . . 5
3.1. Nitrobenzenes . . . 5
3.2. Nitrotoluenes . . . 9
3.3. Nitroxylenes . . . 13
3.4. Nitronaphthalenes . . . 14
3.5. Other Nitro Aromatics . . . 16
4. Nitrohalo Aromatics . . . 16
4.1. Chloronitrobenzenes . . . 16
4.2. Dichloro- and Polychloronitrobenzenes . . . 20
4.3. Chloronitrotoluenes . . . 21
4.4. Fluoronitrobenzenes and Fluoronitrotoluenes . . . 23
5. Nitroamino Aromatics . . . 24
5.1. Nitroanilines . . . 25
5.2. Nitrotoluidines . . . 28
5.3. Halogenonitroanilines . . . 29
5.4. Cyanonitroanilines . . . 29
6. Nitroaromatic Sulfonic Acids and Derivatives . . . 30
6.1. Nitrobenzenesulfonic acids and Derivatives . . . 30
6.2. Nitrotoluenesulfonic Acids and Derivatives . . . 32
6.3. Chloronitrobenzenesulfonic Acids and Derivatives . . . 34
6.4. Chloronitrotoluenesulfonic Acids . . . 36
7. Nitrohydroxy and Alkoxy Aromatics 37
7.1. Nitrophenols and Derived Ethers . . . 37
7.2. Nitroalkylphenols . . . 40
8. Nitroketones . . . 41
9. Nitroheterocycles . . . 42
9.1. Pyridine Derivatives . . . 43
9.2. Quinoline Derivatives . . . 43
9.3. Imidazole Derivatives . . . 43
9.4. Furan Derivatives . . . 43
9.5. Thiophene Derivatives . . . 44
9.6. Thiazole Derivatives . . . 44
10. References . . . 45

1. Introduction

The earliest aromatic nitro compounds were obtained by MITSCHERLICH in 1834 by treating hydrocarbons derived from coal tar with fuming nitric acid. By 1835 LAURENT was working on the nitration of naphthalene, the most readily available pure aromatic hydrocarbon at that time. DALE reported on mixed nitro compounds derived from crude benzene at the 1838 annual meeting of the British Association for the Advancement of Science. Not until 1845, however, did HOFMANN and MUSPRATT report their systematic work on the nitration of benzene to give mono- and dinitrobenzenes by using a mixture of nitric and sulfuric acids.

The first small-scale production of nitrobenzene was carefully distilled to give a yellow liquid with a smell of bitter almonds for sale to soap and perfume manufacturers as "essence of mirbane."

Bechamp's iron reduction process, which made aniline more readily available, was published in 1854, and the discovery of aniline mauve by PERKIN in 1856 started the European aniline dye industry that became the basis for a worldwide synthetic colorant industry estimated to have had sales of $\$ 6 \times 10^9$ in 1988.

Process development and scaleup of the nitration and reduction processes, begun by PERKIN and continued by many others, resulted in the trade price of aniline dropping from 50 Fr/kg in 1858 to 10 Fr/kg in 1863; by 1871, the European production of aniline had reached 3500 t/a [23].

In 1985 European production of aniline was 500 000 t/a, and it is still the single largest product based on a nitration process. Its use in colorants now accounts for only 4 % of aniline pro-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_411

duction because most of the growth has been due to rubber chemicals and isocyanates, the latter now consuming well over 50 % of nitrobenzene production for the manufacture of methylene diphenyldiisocyanate.

The number of naturally occurring nitroaromatic compounds is small; the first to be recognized was chloramphenicol (**1**), an important compound extracted from cultures of a soil mold *Streptomyces venezuelae* and identified in 1949.

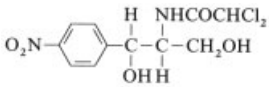

1

This discovery stimulated investigations into the role of the nitro group in pharmacological activity, following the earlier (1943) discovery of the antibacterial activity of nitrofuran derivatives. Many synthetic pharmaceuticals and agrochemicals contain nitroaromatic groups, although the function of the nitro group is often obscure.

Nitration is the dominant method for introducing the nitro group into aromatic systems. Several so-called indirect methods also exist:

1) Oxidation of nitroso or amino compounds,
2) Replacement of diazonium groups (nitro-Sandmeyer reaction),
3) Rearrangement of nitramines, and
4) Nucleophilic displacement reactions.

Although useful in overcoming problems of substituent orientation or substrate sensitivity to nitration conditions [14], they have very little industrial significance. Of more than 200 nitroaromatic compounds described in this article, only two require synthesis by indirect methods. Many important nitro compounds are produced by applying unit processes (e.g., sulfonation, halogenation, or amination) to primary nitro starting materials, most of which are derived from the key primaries nitrobenzene, nitrotoluenes, and nitrochlorobenzenes.

Nucleophilic displacement of activated nitro groups is of minor industrial significance but has considerable synthetic potential. The nucleofugicity may rival that of the corresponding fluorine substituent [24], and the nitro precursor may be more readily available than its halogen analogue. For example, the readily available 1,3-dinitrobenzene reacts with sodium methoxide to form 3-nitroanisole or with potassium fluoride (180 °C, DMF) to form 3-fluoronitrobenzene.

The choice of nitro compounds covered here is influenced strongly by their commercial application, and a good guide to this is the registration of compounds in the 1981 European Core Inventory (ECOIN). Most nitro compounds, or their derivatives, are intermediates for colorants, agrochemicals, pharmaceuticals, or other fine chemicals with a few major volume outlets for synthetic materials and explosives.

For many compounds, CAS nomenclature conventions differ from the more important names in common usage. The latter are given here, with the CAS name and number where meaningful.

2. Nitration

The unit process of nitration can be defined simply as the irreversible introduction of one or more nitro (NO_2) groups into an aromatic nucleus by replacement of a hydrogen atom. O-Nitration to give nitrates and N-nitration to give nitramines are much far less important commercially for aromatic compounds and are not dealt with here.

Nitration, an electrophilic substitution reaction, is represented by the equation:

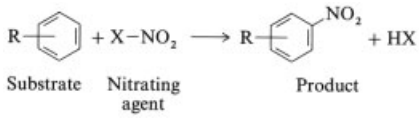

Introduction of the nitro group so deactivates the ring to further electrophilic substitution that dinitration rarely occurs under the conditions used for mononitration. The more vigorous conditions required for dinitration (i.e., excess, stronger acid and higher temperature) must usually be applied to isolated mononitro compounds, rather than carrying out stepwise reaction in situ.

The nitration reaction is always strongly exothermic, as exemplified by the mononitration of benzene ($\Delta H = -117$ kJ/mol) and naphthalene ($\Delta H = -209$ kJ/mol) and is probably the most potentially hazardous industrially operated

unitprocess. This is due fundamentally to the heat generated, which can trigger the power of nitric acid to degrade organic materials exothermically to gaseous products with explosive violence [19]. Nitroaryl compounds, especially those with more than one nitro group, are potentially hazardous due to their very high oxygen contents. Some polynitro compounds (e.g., trinitrotoluene and picric acid) are detonable and have a long history of use as explosives [8]. A less publicized hazard with nitroaromatic compounds is occasioned by the violent decomposition reactions that may occur on heating with alkali [19].

The nitration process may be carried out on either a batch (discontinuous) or a continuous basis. Lower tonnage requirements are met by batch reaction in reactors (nitrators) designed to accommodate a variety of products. Maximum flexibility is obtained by using the more expensive enamel-lined mild-steel reactors in place of the traditional cast iron or stainless steel. The exothermic process requires efficient cooling of the reactor by jacket, internal cooling, or both. For safety reasons, the vessel is usually limited to about 6000-L capacity and is fitted with an efficient agitator.

Continuous reaction for large-tonnage intermediates (e.g., nitrobenzene and nitrotoluenes) is attractive for reasons of safety as well as the obvious economies of a dedicated unit. Whether continuous operation is achieved by a continuous mixing device or by a cascade of small agitated reactors, the material inventory is much lower than for batch reactors so the reaction temperature is easier to control. The continuous reactor requires closely defined operating parameters and management.

Typical nitroaromatic production is based on high-yield processes, with more than 80 % of the total cost being represented by the raw materials. This means that only a small proportion of the cost is available for operational savings. Two major areas that have yielded improvements are (1) the sulfuric acid recycle that has been an integral requirement of all efficient nitration processes and (2) isomer control and separation. Both these points are highlighted in more detail, together with information on plant design, in the descriptions of processes for individual products.

2.1. Nitrating Agents

For reasons of practicability and economics, industrial-scale nitration is usually carried out with a mixture of nitric and sulfuric acids (mixed acid), and occasionally with aqueous nitric acid, nitric acid in acetic acid, or nitric acid in acetic anhydride. Use of alternative component acids such as perchloric acid, hydrofluoric acid, or boron trifluoride is limited to important supporting studies. These are sometimes carried out in inert organic solvents such as chlorohydrocarbons or sulfolane to give homogeneous reaction mixtures.

The strength of the nitrating agent ($X-NO_2$) decreases with decreasing electronegativity of X [25]: nitronium ion (e.g., $BF_4^-NO_2^+$) > nitracidium ion ($OH_2^+-NO_2$) > nitronium chloride ($Cl-NO_2$) > acetyl nitrate ($AcO-NO_2$) > nitric acid ($HO-NO_2$) > ethyl nitrate ($C_2H_5O-NO_2$).

The nitronium ion, NO_2^+, is considered to be the active species in all of these systems. In the most common system the overall equation

$$HNO_3 + 2\,H_2SO_4 \rightleftharpoons NO_2^+ + H_3O^+ + 2\,HSO_4^-$$

is a composite of many equilibria present in $HNO_3-H_2SO_4-H_2O$ mixtures. These must all be taken into account when considering the reactivity of the substrate and the extent of nitration required.

A typical nitrating agent for large-scale aromatic mononitration consists of 20 % nitric acid, 60 % sulfuric acid, and 20 % water; this is referred to as 20/60/20 mixed acid. An alternative definition of the same acid, used in some situations, is 15 mol % nitric acid, 30 mol % sulfuric acid, and 55 mol % water. Commonly, the liquid aromatic substrate and the nitrated product together form a separate phase from the aqueous mixed acid. Efficient agitation is therefore required to maximize contact with the organic phase and minimize resistance to mass transfer. Solid substrates are best dissolved in the sulfuric acid phase. Stronger nitric acid leads to oxidative side reactions, whereas higher temperature leads to decreased nitronium ion concentration. Much detailed developmental work is required for each individual nitration to optimize these and other variables, to maximize formation of

the required isomer, and to minimize side reactions.

Free-radical reactions (typically in the vapor phase with nitric acid or nitrogen dioxide as nitrating agent) tend not to be employed in aromatic nitration, being more appropriate for aliphatic nitration (→ Nitro Compounds, Aliphatic). However, claims have been made that a radical mechanism makes some contribution, even when ionic mechanisms clearly predominate [26].

2.2. Reaction Mechanisms

Much experimentation has been applied to aromatic nitration since the nitronium ion, NO_2^+, was confirmed as the active species by the Ingold – Hughes school in 1950 through study of Raman spectra. Temperature and degree of mixing are important parameters for all nitrating agents, but other variables often lead to inconsistent results [27, 28, 16]. In commercial nitration with HNO_3– H_2SO_4 the relative rate constants cover a wide range, with reaction times varying from several seconds for active substrates to many hours for inactive substrates. It has been argued that under the nonideal conditions of large-scale production, nitration is controlled more by mass transfer than by kinetics, and that many laboratory studies wrongly assumed that mass-transfer resistance is negligible [29]. Another early assumption that the rate of nitration parallels the equilibrium formation of the nitronium ion was also challenged because of the realization that different mechanisms may apply at differing acid strengths [30].

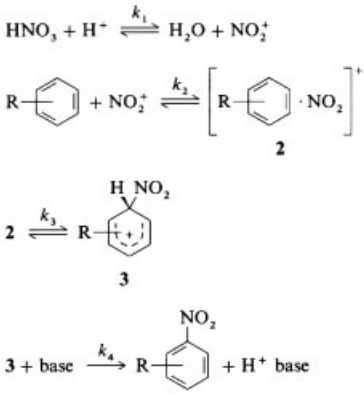

Figure 1. Basic mechanism of aromatic nitration

The mechanism that can accommodate all the facts available on the nitration of simple substrates (i.e., benzene and toluene, but not phenols or polyalkylbenzenes) is shown in Figure 1 [28]. The possible involvement of four rate constants and three significant reversible reactions makes the kinetics extremely complex even if the reaction is totally kinetically controlled.

The encounter complex, or encounter pair (**2**) [27], is the novel feature of this mechanism; it is a molecular complex of less specific nature than the earlier concept of an intermediate π-complex. The important point is that **2** has a kinetic role and may indeed be rate determining. Conversion of the σ-complex (**3**), sometimes called a Wheland intermediate, to the product is not rate limiting because the nitration reaction does not exhibit a primary kinetic isotope effect. Although under certain conditions formation of the nitronium ion may be rate determining, for reactive substrates in aqueous sulfuric acid the formation of **2** is predominately the rate-determining step, whereas nitration of toluene, benzene, and other less reactive aromatic compounds obeys an overall rate law, for which formation of **3** is predominately the rate-determining step [28]; that is

$$\text{rate} = k_{\text{obsd}}\,[HNO_3]\,[ArH]\,a_{H+}$$

where k_{obsd} depends on the structure of the aromatic compound.

When the ring is highly activated towards electrophilic attack, as in the case of phenol, the nitrosonium ion (NO^+), arising from a catalytic amount of nitrous acid, comes into play, with formation of an intermediate nitroso compound, which is oxidized by nitric acid to the nitro derivative with regeneration of nitrous acid.

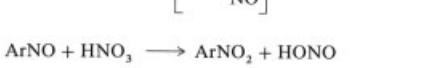

Ipso-electrophilic attack at a substituent position represents another exception to the general mechanism. With di- and polyalkylbenzenes the *ipso*-nitroarenium ion (**4**) may be formed instead of **3**. Although rearrangement of **4** to the conventional Wheland intermediate does occur, it is accompanied by side reactions such as

nitrodealkylation, thus giving polyalkylbenzene nitration potentially more variable mixture products than other nitrations [31]. Increased understanding of

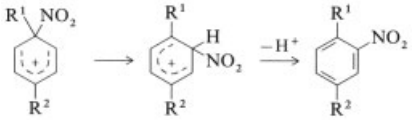

4

the formation of ipso-Wheland intermediates has led to their exploitation in other reactions, especially with nucleophiles [16].

3. Nitro Aromatics

3.1. Nitrobenzenes

Mono-, di-, and symmetrical trinitrobenzenes are readily available by sequential nitration of benzene. Neither *o*- and *p*-dinitrobenzene nor unsymmetrical trinitrobenzenes have any significant industrial importance.

Nitrobenzene [*98-95-3*] (**5**), oil of mirbane, $C_6H_5NO_2$, M_r 123.1, is a pale yellow liquid with an odor of bitter almonds, which is readily soluble in most organic solvents and miscible with benzene in all proportions. It is a good solvent for aluminum chloride and is therefore used as a solvent in Friedel – Crafts reactions. It is only slightly soluble in water (0.19 % at 20 °C; 0.8 % at 80 °C) and is steam volatile. Binary azeotropes of nitrobenzene are listed in Table 1. Some other physical properties of nitrobenzene are as follows:

mp	5.85 °C
mp (tech. spec.)	5.5 °C (min.)
bp	
at 101 kPa	210.9 °C
at 13 kPa	139.9 °C
at 4 kPa	108.2 °C
at 0.13 kPa	53.1 °C
$d_4^{1.5}$ (solid)	1.344
d_4^{10}	1.213
$d_4^{15.5}$ (tech. spec.)	1.208 – 1.211
d_4^{25}	1.199
Viscosity (15 °C)	2.17×10^{-2} mPa · s
Surface tension (20 °C)	43.35 mN/m
Dielectric constant	
at 20 °C	35.97
at 170 °C	18.15
Specific heat (30 °C)	1.418 J/g
Latent heat of fusion	94.1 J/g
Latent heat of vaporization	331 J/g
Flash point (closed cup)	88 °C
Autoignition temperature	482 °C
Explosive limit in air (93 °C)	1.8 vol%
Vapor density	4.1
Refractive index n^{15}	1.55457

Table 1. Nitrobenzene-containing binary azeotropes

Second component	*bp* (101 kPa), °C	Nitrobenzene, wt %
Acetamide	202.0	76
Benzotrichloride	210.7	98.5
Benzyl alcohol	204.2	38
4-Chlorophenol	219.9	92
N,*N*-Diethylaniline	210.7	97
Ethyl benzoate	210.6	81
Ethylene glycol	185.9	41
Propionamide	205.4	76

Reactions of nitrobenzene involve either electrophilic meta substitution of the aromatic ring by nitration, sulfonation, or halogenation (Fig. 2, see next page), for example, or reduction of the nitro group to yield a variety of products, as shown in Table 2. The primary reduction products are nitrosobenzene, *N*-phenylhydroxylamine, and aniline, with the other products being formed by interactions and rearrangements, as shown in Figure 3 (see next page). Very specific reduction conditions are required to obtain pure products, and intermediate stages (e.g., nitrosobenzene) are not usually isolable.

Table 2. Nitrobenzene reduction products

Reducing agent	Product
Fe, Zn, or Sn/HCl, H_2 – catalyst	aniline
Zn – H_2O	*N*-phenylhydroxylamine (rearranges with acid to 4-aminophenol)
Na_3AsO_3	azoxybenzene
Zn – NaOH	azobenzene, hydrazobenzene (rearranges with acid to benzidine)
$LiAlH_4$	azobenzene
$Na_2S_2O_3$ – Na_3PO_4	sodium phenylsulfamate ($C_6H_5NHSO_3Na$)
Electrolytic reduction	4-aminophenol

Production. The traditional batchwise nitration process, in which mixed acid (27 – 32/56 – 60/8 – 17 wt %) is added to a slight excess of benzene (to avoid nitric acid in the spent acid), with the temperature controlled at 50 – 55 °C and the

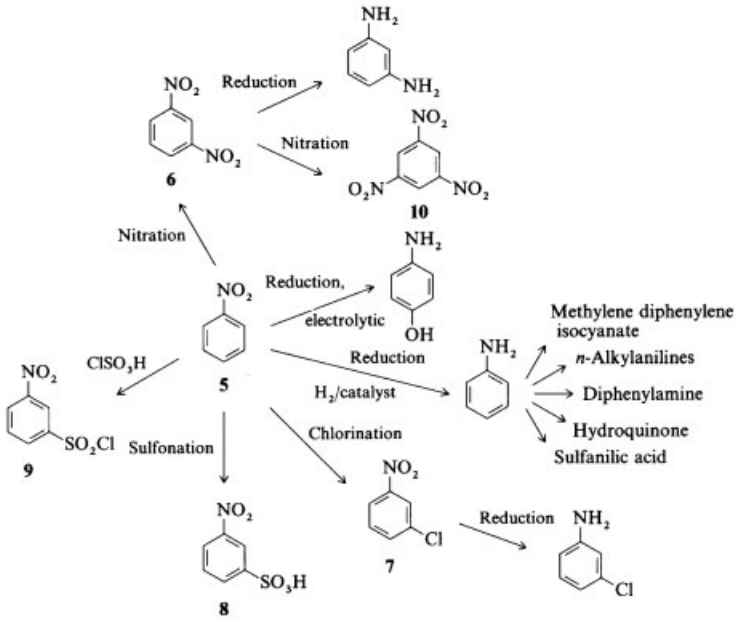

Figure 2. Key intermediates derived from nitrobenzene

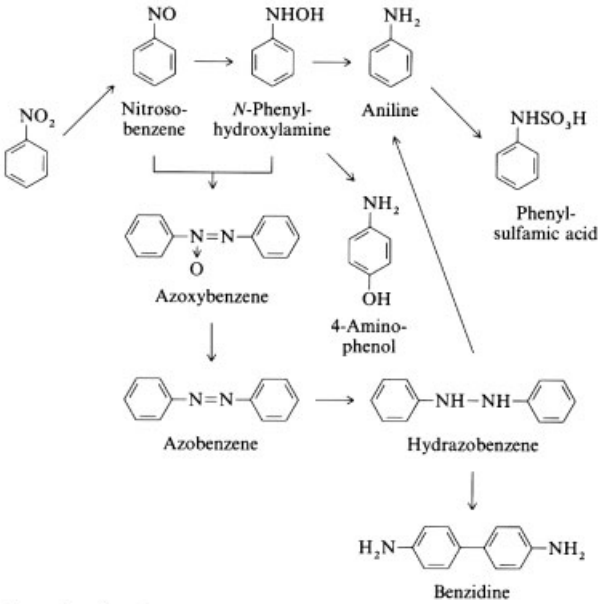

Figure 3. Nitrobenzene reduction products

reaction completed by heating to 80 – 90 °C, has been replaced by continuous processes that operate under similar conditions. Both economy and safety, as summarized in Section 2.1, result from the more efficient mixing and higher reaction rates possible in smaller reactors. An 120-L continuous reactor has been reported to give the same output of nitrobenzene as a 6000-L batch reactor. The variety of reactor configurations employed is typified by a stirred cylindrical reactor operating at 50 – 100 °C. The critical reuse of spent acid is achieved by continu-

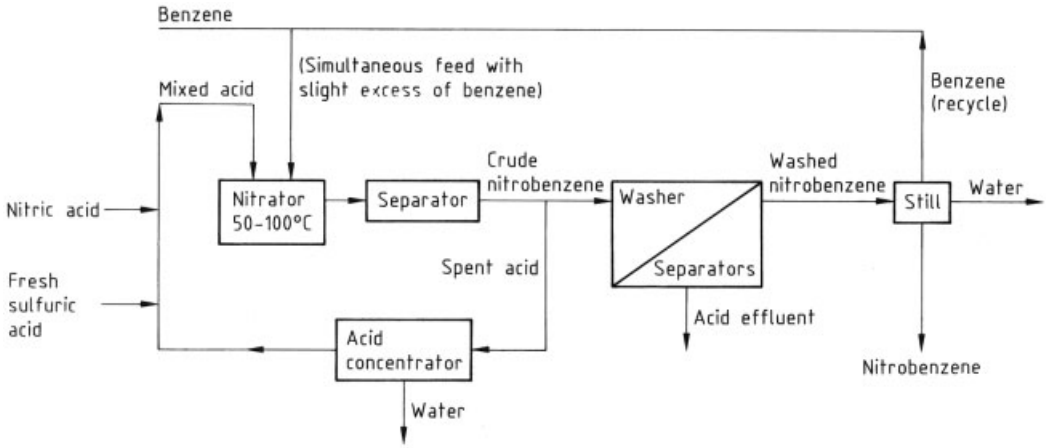

Figure 4. Production of nitrobenzene – continuous process

ous concentration and addition of fresh acid, as shown in Figure 4. This differs from the concept of batchwise reuse although, in each case, addition of fresh acid or SO_3 results in the disadvantage of an equivalent amount of spent acid for disposal.

Many of the world's current production facilities are package units based on years of design experience. For example, Meissner units, with capacities up to 12 t of nitroaromatics per hour, have built-in nitrogen blanketing for additional safety. Each output stream passes through purging steps; thus, spent acid is extracted with incoming benzene to remove both residual nitrobenzene and nitric acid, while residual waste gases are scrubbed by a mixed acid loop to meet environmental regulations [32].

An alternative development from the turbulent flow tube reactor is the pump nitration circuit of Nobel Chematur. When nitration actually takes place in the pump itself, reaction times can be less than a second due to the intensive mixing, and many consequential advantages are claimed [33, 34].

The concept of an adiabatic process that uses the heat of reaction for concentration of spent acid has been around since 1941 [35]. The American Cyanamid adiabatic process [36], jointly developed with CLR, forms the basis for one of the world's largest nitrobenzene plants, rated at 159 000 t/a. This process takes cocurrent streams of benzene (1.1 mol per mole of HNO_3) and mixed acid (6 – 8/62 – 68/24 – 32 wt %) through a vigorously agitated tubular reactor, with entry at 60 – 80 °C (below the boiling point of benzene) and exit at ca. 120 °C, so engineered that the residence time is around 4 min. The requirement for much weaker sulfuric acid makes acid recycle much more economical, especially because process heat is virtually sufficient for evaporation if this is carried out under vacuum.

Another option for eliminating the need to reconcentrate the sulfuric acid in a separate step is to carry out the nitration at higher temperature (120 – 160 °C), with excess water being distilled from the nitrator as an azeotrope with benzene. In a duplex process the benzene azeotrope resulting from a high-temperature first-stage partial reaction is used in a lower temperature second stage to complete the reaction [37]. Azeotropic nitration offers considerable energy savings, but it is not thought to be competitive with adiabatic nitration.

In all these processes the work-up streams are very similar (cf. Fig. 4). The reaction mixture passes from the nitrator to a separator or centrifuge, and aqueous spent acid is fed to the recycle loop via a concentrator or strength adjustment stage as appropriate. Crude nitrobenzene, as the top layer, is passed through a series of washer – separators in which residual acid is removed by washing first with dilute alkali and then with water. The washed product is topped in a still to remove water and benzene (recycled) and, if required, finally vacuum distilled to give pure product in ca. 96 % overall yield. The washing stages give rise to an effluent problem, and wastewater treatment is necessary, principally to remove nitrobenzene.

World capacity for nitrobenzene in 1985 was ca. 1.7×10^6 t/a, with about one-third located in Western Europe and one-third in the United States [20].

Uses. The use of nitrobenzene as a processing solvent in specific chemical reactions is minor but important. Most (95 % or more) nitrobenzene produced is converted to aniline, which has hundreds of downstream products. Lower volume, but nevertheless important, industrial outlets include electrolytic reduction to 4-aminophenol, nitration to give 1,3-dinitrobenzene (**6**), chlorination to give 3-chloronitrobenzene (**7**), sulfonation to give 3-nitrobenzenesulfonic acid (**8**), and chlorosulfonation to give 3-nitrobenzenesulfonyl chloride (**9**). The last three products are consumed mainly as their reduction products, 3-chloroaniline, metanilic acid, and 3-aminobenzenesulfonamide, respectively.

Nigrosin (C.I. Solvent Black 5) survives as the crude mixture obtained by reacting nitrobenzene with aniline and aniline hydrochloride at 200 °C in the presence of iron or copper.

Toxicity. Nitrobenzene (TLV 5 mg/m^3) is very toxic and can cause acute poisoning because itis readily absorbed via the skin or respiratory tract. The primary effects are lowering of the hemoglobin level, methemoglobinemia, cyanosis, and breathlessness. Chronic exposure can lead to spleen and liver damage, jaundice, and anemia. Ingestion of alcohol may speed up and exaggerate the effects. Working conditions must be tightly specified and controlled [22].

1,3-Dinitrobenzene [*99-65-0*] (**6**), *m*-dinitrobenzene, $C_6H_4N_2O_4$, M_r 168.1, crystallizes from ethanol as pale yellow rhombic – bipyramidal plates; the compound is steam volatile. Physical properties are listed in Table 3, together with those of the 1,2- and 1,4-isomers for comparison.

Chemical reduction of 1,3-dinitrobenzene gives 3-nitroaniline or *m*-phenylenediamine, depending on conditions. Under mild conditions, 1,3-dinitrobenzene does not react with aqueous sodium hydroxide or sodium sulfite solution, whereas the 1,2- and 1,4-isomers react to form alkali-soluble nitrophenols or nitrobenzene sulfonic acids. These reactions provide the basis for the removal of small amounts of isomeric impurities from 1,3-dinitrobenzene.

Table 3. Physical properties of the three dinitrobenzenes

Property	Isomer		
	Ortho	Meta	Para
mp, °C	118	91	178 ∗
bp, (101 kPa) °C	318	307	298
Density, g/cm^3 (t, °C)	1.3119 (120)	1.3729 (100)	1.617 (20)
Dipole moment, D		3.88	
Solubility in H_2O (25 °C), %	0.013	0.081	0.065
Solubility in C_2H_5OH (20 °C), %	1.9	3.5	0.4
Solubility in benzene (18 °C), %	5.66	39.45	2.56
Solubility in chloroform (18 °C), %	27.1	32.4	1.82

∗ After sublimation.

Production. Nitrobenzene is further nitrated with strong mixed acid (HNO_3 33 wt %, H_2SO_4 67 wt %) by controlled addition and reaction at 60 – 90 °C. After completion of the reaction and addition to water the crude oil is separated while still hot. The ortho isomer (10 %), the para isomer (2 %), and nitrophenol (2 %) byproducts are removed by treatment with an aqueous NaOH – $NaHSO_3$ mixture. The final yield is ca. 82 % of 99 % pure product (*mp* 88 – 89 °C).

Uses. Catalytic hydrogenation of 1,3-dinitrobenzene yields *m*-phenylenediamine; smaller quantities are reduced with aqueous sodium sulfide to give 3-nitroaniline (Section 5.1). The I.G. Farbenindustrie developed an interesting alternative isomer-directing process for 1,3-dichlorobenzene by passing chlorine into boiling 1,3-dinitrobenzene; a 79 % yield of distilled product was reported [1]. Historically, 1,3-dinitrobenzene has been used in the formulation of explosives, lubricants, polymerization inhibitors, and corrosion inhibitors.

Toxicity. 1,3-Dinitrobenzene and its isomers are highly toxic [8] by inhalation or skin absorption. The TLV dermal is 1 mg/m^3. Prolonged exposure to small amounts can lead to accumulation in fatty tissue, with the additional danger that ingestion of alcohol can flush out these residues to produce the same severe toxic symptoms as massive exposure. Progressive symptoms of cyanosis, headaches, and vomiting can lead to severe poisoning which is signaled by

yellowing of the hair and nails in prolonged cases.

1,3,5-Trinitrobenzene [*99-35-4*] (**10**), *s*-trinitrobenzene, $C_6H_3O_6N_3$, M_r 213.1, crystallizes as yellow dimorphic crystals from ethanol or nitric acid (*mp* 61 and 122.5 °C, respectively). 1,3,5-Trinitrobenzene is inert to further ring substitution. Reduction with sodium sulfide gives 3,5-dinitroaniline and with Fe – HCl, 1,3,5-triaminobenzene. The latter is an intermediate in the production of phloroglucinol. Reaction with potassium methoxide results in the formation of 3,5-dinitroanisole via a classical Meisenheimer complex. An important property is the formation of π-complexes with hydrocarbons and other donors, which are useful in their characterization or isolation. Although trinitrobenzene has more explosive power than trinitrotoluene (TNT), it is too expensive in comparison [8].

Production. Although 1,3-dinitrobenzene may be further nitrated with strong mixed acid, extensive (dangerous) side reactions occur and the yield of 1,3,5-trinitrobenzene is low. The preferred synthesis is by decarboxylation of 2,4,6-trinitrobenzoic acid (obtained from TNT by oxidation with chromic acid) by heating in aqueous medium.

3.2. Nitrotoluenes

All three isomeric mononitrotoluenes, 2,4-dinitrotoluene (together with mixed isomeric dinitrotoluenes), and 2,4,6-trinitrotoluene (see → Explosives) are industrially important products obtained from the sequential nitration of toluene. The main outlets for mono- and dinitrotoluenes depend on their catalytic hydrogenation to the derived amines, but minor outlets also exist for sulfonated and chlorinated derivatives. Figure 5 shows key intermediates derived from nitrotoluenes.

2-Nitrotoluene [*88-72-2*] (**11**), *o*-nitrotoluene, ONT, 1-methyl-2-nitrobenzene, $C_7H_7NO_2$, M_r 137.1, is a clear yellow liquid that crystallizes at lower temperatures to solid α- or β-forms, depending on conditions. The compound is steam distillable at 100 g of ONT per kilogram of distillate. 2-Nitrotoluene is soluble in most organic solvents but only slightly soluble in water (0.065 % at 30 °C). Some physical properties are listed in Table 4, together with those of 3- and 4-nitrotoluenes for comparison.

Table 4. Physical properties of the three nitrotoluenes

Property	Isomer		
	Ortho	Meta	Para
mp, °C	− 9.27 (α-form) − 3.17 (β-form)	16.1	44.5 (unstable) 51.9 (stable form)
bp, °C			
at 101 kPa	222.3	231.9	238.3
at 0.13 kPa	50.0	50.2	53.7
d_4^{20}	1.163	1.1571	1.286
Vapor density (air = 1)	4.72	4.72	4.72
Refractive index n_D^{20}	1.5474	1.5470	1.5460
Surface tension, mN/m	42.3 (15 °C)	39.9 (30 °C)	36.8 (60 °C)
Dielectric constant (58 °C)	21.61	22.2	22.2
Flash point (closed cup), °C	91	102	106

Reduction of 2-nitrotoluene with (1) hydrogen or iron borings and acid gives *o*-toluidine; (2) zinc dust and ammonia in aqueous ethanol gives *o*-tolylhydroxylamine, (3) iron powder and aqueous caustic soda gives *o*-azoxy- and *o*-azotoluene; and (4) zinc dust and alcoholic sodium hydroxide yield *o*-hydrazotoluene (cf. Table 2).

Oxidation of 2-nitrotoluene with manganese dioxide in sulfuric acid gives 2-nitrobenzaldehyde [*552-89-6*] (**14**) or 2-nitrobenzoic acid [*552-16-9*] (**15**), depending on the strength of the sulfuric acid. When potassium permanganate, potassium dichromate, or nitric acid is used as oxidizing agent, only 2-nitrobenzoic acid is obtained. Treatment of 2-nitrotoluene with boiling aqueous sodium hydroxide produces anthranilic acid through the unusual combination of autoxidation and reduction. However, the yield is too low for industrial use.

Reaction of 2-nitrotoluene with diethyl oxalate gives 2-nitrophenylpyruvic acid [*5461-32-5*] (**27**), which on reductive cycliza-

Figure 5. Key intermediates derived from nitrotoluenes

tion and decarboxylation converts to indole – the model example of the Reissert indole synthesis. This is one of the many reactions of 2-nitrotoluenes with neighboring group interaction to give cyclized products [38].

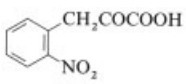

27

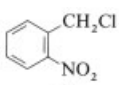

28

Nitration of 2-nitrotoluene yields a mixture of 2,4-dinitrotoluene and 2,6-dinitrotoluene. Chlorination in the presence of iron or Friedel – Crafts catalysts yields a mixture of 4-chloro-2-nitrotoluene (**16**) and 2-chloro-6-nitrotoluene (**17**), whereas in the absence of iron, 2-nitrobenzyl chloride [*612-23-7*] (**28**) or 2-chlorotoluene is formed, depending on reaction

Table 5. Nitration products of toluene with various agents

Nitrating agent	Temperature, °C	2-Nitrotoluene, %	3-Nitrotoluene, %	4-Nitrotoluene, %	Ortho – para ratio
85 % HNO_3	−30	55.6	2.7	41.7	1.33
85 % HNO_3	60	57.5	4.0	38.5	1.49
23.8 % HNO_3, 58.7 % H_2SO_4, 17.5 % H_2O	50	62.2	4.3	33.5	1.86
5.4 % HNO_3, 71.4 % H_2SO_4, 21.5 % H_2O	60	58.2	5.5	36.2	1.61
90 % HNO_3 + aromatic sulfonic acid	50	51.0	3.6	45.4	1.12
10 % HNO_3, 90 % H_3PO_4	30	45.1	4.3	50.2	0.9

conditions. 2-Nitrotoluene reacts with formaldehyde in concentrated sulfuric acid to give dinitroditolylmethane.

Production. 2-Nitrotoluene is the dominant isomer in conventional mixed acid nitration of toluene. Table 5 shows a 60 ± 5 % ortho, 38 ± 5 % para mix at different sulfuric acid strengths, as well as the products obtained by using aromatic sulfonic acid or phosphoric acid, which contain lower ortho – para ratios. This is a desirable feature because the ratio of ca. 1.6 obtained under practical mixed acid conditions does not normally match the requirements and leads to a surplus of 2-nitrotoluene. However, the phosphoric acid alternative with the desirable ratio [39] shown in Table 5 has, apart from the problem of phosphoric acid recovery, a tendency to increase the amount of dinitrotoluene formed, which leads to an apparent reduction in the amount of ortho product. Under conditions in which dinitrotoluene formation is eliminated, one evaluation of the phosphoric acid process could only achieve a reduction in the ortho – para ratio to 1.35 [40]. Many claims for alternative processes and reaction media have been patented [17] without becoming established.

Large-scale nitration is still carried out under the usual conditions of mixed acid (17 – 32/67 – 52/ca. 16 %) at 25 – 40 °C and a nitric acid to toluene molar ratio close to one. The inherent speed of reaction (17 times faster than with benzene) makes continuous reaction with low residence times most suitable. An ortho – para ratio of 1.6 and an isolated yield of 96 % total isomers are typical in a plant similar to that described for nitrobenzene but operating at as low a temperature as possible to avoid the more readily formed byproducts. The completed nitration mix is fed to a continuous, usually centrifugal, separator while hot, and the crude isomer mix is washed with dilute alkali and water (cf. nitrobenzene production). The washed product is transferred to a still, and the water and residual toluene are topped off before fractional distillation under vacuum (96 – 97 °C, 1.6 kPa) to give fairly pure 2-nitrotoluene [8].

The 1984 capacity for mononitrotoluene (all isomers) production in the Western world was ca. 200 000 t/a [20].

Uses. 2-Nitrotoluene derivatives are used principally as colorant intermediates. For example, 2-toluidine, 2-amino-4-chlorotoluene [Fast Scarlet TR Base, by reduction of 4-chloro-2-nitrotoluene (**16**)], 2-amino-6-chlorotoluene [Fast Red KB Base, by reduction of 2-chloro-6-nitrotoluene (**17**)], and 2-toluidine-4-sulfonic acid [by reduction of 2-nitrotoluene-4-sulfonic acid (**18**)] are all diazo components of azo dyes.

A more recent outlet for 2-toluidine (which is important in consuming the normal surplus of 2-nitrotoluene) is its conversion to 2-ethyl-6-methylaniline, an agrochemical intermediate.

Reduction to *o*-hydrazotoluene followed by rearrangement to 2-tolidine has virtually been phased out based on analogy with the proven carcinogen benzidine.

3-Nitrotoluene [*99-08-1*] (**12**), *m*-nitrotoluene, 1-methyl-3-nitrobenzene, $C_7H_7NO_2$, M_r 137.1, is a yellow liquid at ambient temperature; it is soluble in most organic solvents, but only slightly soluble in water (0.05 % at 30 °C). Physical properties are listed in Table 4.

Reduction of 3-nitrotoluene with hydrogen or strong reducing agents produces 3-toluidine, but treatment with Zn – H_2O or Zn – EtOH – KOH yields *m*-tolylhydroxylamine and *m*-azotoluene, respectively. Oxidation with chromic acid gives 3-nitrobenzoic acid [*121-92-6*]. Further nitration leads to a mixture of 2,3-, 2,5-, and 3,4-dinitrotoluenes (Table 6, see next page).

Production. The middle "meta fraction" distilled from the mononitrotoluene isomer mix after 2-nitrotoluene has been removed is predom-

inantly 3-nitrotoluene, but with a large 4-nitrotoluene content. This is accumulated to form the basis for more rigorous fractional distillation to give technical quality 3-nitrotoluene.

Table 6. Dinitrotoluene isomeric mixtures obtained by nitration

Starting material	Dinitrotoluene, wt %					*cp* ∗, °C
	2,3-	2,4-	2,5-	2,6-	3,4-	
Toluene	1.3	78	0.5	18	2.4	55 – 57
2-Nitrotoluene		67		33	< 0.5	50 – 51
3-Nitrotoluene	25		20		55	
4-Nitrotoluene		99				69

∗ Crystallization point of the isomeric mixture.

Uses. Catalytic hydrogenation or reduction with iron converts 3-nitrotoluene to *m*-toluidine, which provides important coupling components for azo dyes, usually after N-alkylation.

4-Nitrotoluene [*99-99-0*] (**13**), *p*-nitrotoluene, PNT, 4-nitro-1-methylbenzene, $C_7H_7NO_2$, M_r 137.1, forms colorless to light yellow rhombic – bipyramidal crystals, which are soluble in most organic solvents but only slightly soluble in water (0.044 % at 30 °C). Physical properties are listed in Table 4.

Reduction of 4-nitrotoluene with hydrogen, iron, and acid, or electrolytically in aqueous hydrochloric acid produces *p*-toluidine. Treatment with zinc dust in hot aqueous – alcoholic calcium chloride gives *p*-tolylhydroxylamine. Oxidation with chromic oxide in acetic acid produces 4-nitrobenzaldehyde [*555-16-8*]. Electrolytic oxidation in the presence of manganese salts or treatment with other strong oxidizing agents yields 4-nitrobenzoic acid, whereas controlled electrolytic oxidation in acetic acid – sulfuric acid results in 4-nitrobenzyl alcohol (**22**). Air oxidation (MeOH – KOH medium) leads to 4,4′-dinitrodibenzyl and 4,4′-dinitrostilbene.

Chlorination in the presence of Friedel – Crafts catalysts gives 2-chloro-4-nitrotoluene (**23**), whereas without catalyst 4-nitrobenzyl chloride (**24**) is produced. Further nitration or sulfonation of 4-nitrotoluene gives highly selective 2-substitution.

Production. After distillation of the meta fraction from the isomer mixture, the still residues are cooled in a crystallizer to separate technical quality 4-nitrotoluene. Further distillation of nitrotoluene residues and fractions should be implemented with great care because this has been reported to cause explosions. Holding residues at 150 – 200 °C results in an undefined "aging" process that can lead to unpredictable evolution of heat, especially if air is introduced [41].

Uses. 4-Nitrotoluene derivatives are used principally as intermediates for colorants and related products; for example, *p*-toluidine, 4-nitrobenzoic acid [*62-23-7*] (by oxidation of 4-nitrotoluene with 15 % HNO_3 at 175 °C), 4-amino-2-chlorotoluene (by reduction of 2-chloro-4-nitrotoluene), and 4-nitrotoluene-2-sulfonic acid (**25**), which is of great importance in forming stilbene intermediates for fluorescent whitening agents (see Section 6.2).

2,4-Dinitrotoluene [*121-14-2*] (**19**), DNT, 2,4-dinitro-1-methylbenzene, $C_7H_6N_2O_4$, M_r 182.1, *mp* 71 °C, *b p* 300 °C (decomp., 101 kPa), d_4^{15} 1.521, d_4^{71} 1.321, vapor density 6.27, forms yellow needles from ethanol or carbon disulfide, which are moderately soluble in most organic solvents but only sparingly soluble in water (0.03 % at 22 °C).

Reduction of 2,4-dinitrotoluene with iron – acid or H_2 – catalyst gives 2,4-diaminotoluene smoothly. Selective reduction of either nitro group may be achieved under carefully controlled conditions by using the stoichiometric amount of hydrogen with a platinum black catalyst to obtain 4-amino-2-nitrotoluene [*119-32-4*], or an equimolar quantity of tin(II) chloride in alcoholic hydrochloric acid to obtain the isomeric 2-amino-4-nitrotoluene [*99-55-8*].

Oxidation of 2,4-dinitrotoluene with aqueous nitric acid or potassium permanganate yields 2,4-dinitrobenzoic acid [*610-30-0*].

Production. 4-Nitrotoluene is nitrated with mixed acid (containing equimolar HNO_3) under controlled conditions and continuous operation to give a 96 % yield of 2,4-dinitrotoluene. In Meissner units, for tighter control of the whole process, cocurrent flow is used in each step to separate mono- and dinitration, but countercurrent flow is used between them to prevent under- or overnitration; this is also done in one-step nitration plants [32]. After nitration, if separation of the resulting emulsion is difficult, the introduction of pure product can give surprisingly rapid separation [42].

Alternative processes yield mixed isomer products. When toluene is nitrated directly under similar conditions with 2.1 equivalents of nitric acid, the product is a ca. 80 : 20 mixture of 2,4-dinitrotoluene and 2,6-dinitrotoluene [*606-20-2*]. Nitration of 2-nitrotoluene, which is sometimes present in surplus, gives a ca. 67 : 33 mixture of 2,4- and 2,6-dinitrotoluene (Table 6).

Uses. Most of the 2,4-dinitrotoluene produced is hydrogenated (nickel catalyst) to 2,4-diaminotoluene (*m*-tolylenediamine) [*95-80-7*] for conversion to toluene diisocyanate (TDI), which is a component of a polyurethane. A much smaller amount is used in explosives and for further nitration to TNT (**20**).

Crude mixtures of 2,4- and 2,6-dinitrotoluene are also used to produce mixed diamines (80 : 20 or 67 : 33, depending on the nitration process used), which in turn are converted to mixed isocyanates for polyurethane production, the use of isomer mixtures being considerably cheaper.

Toxicity of Nitrotoluenes. The toxic effects of mononitrotoluenes are similar to those described for nitrobenzene. Generally, the toxicity of alkylnitrobenzenes decreases with increasing number of alkyl groups and increases with increasing number of nitro groups. Mononitrotoluenes (TLV 30 mg/m^3) are methemoglobin formers and may be absorbed through the skin or the respiratory tract. Dinitrotoluenes (TLV 1.5 mg/m^3) and TNT (TLV 1.3 mg/m^3) are progressively more potent.

3.3. Nitroxylenes

All three xylenes can be nitrated readily, with *o*- and *m*-xylene giving mixtures of products. Even under simple conditions with mixed acid, isomer ratios can be changed significantly by varying the reaction conditions.

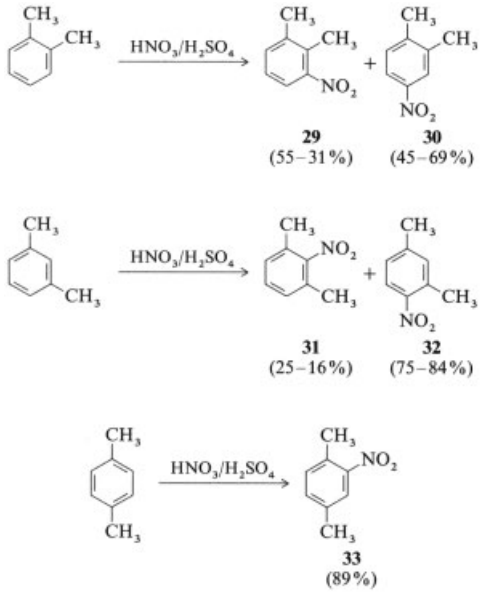

All five important mononitroxylenes are pale yellow oils or low-melting solids, $C_8H_9NO_2$, M_r 151.17, which are soluble in most organic solvents but only slightly soluble in water. Some physical properties are listed in Table 7.

The nitroxylenes (**29**–**33**) are all readily reduced to the corresponding aminoxylenes (xylidines). Oxidation can result in conversion of one or both of the methyl groups to a carboxylic acid group, depending on the reagent and reaction conditions. For example, **30** with aqueous nitric acid yields 2-methyl-4-nitrobenzoic acid and **31** with potassium permanganate gives 2-nitroisophthalic acid, whereas **32** results in a mixture of 4-nitroisophthalic acid and 3-methyl-4-nitrobenzoic acid under similar conditions. Further nitration of nitroxylenes in mixed acid at 35 – 50 °C with a slight excess of nitric acid gives high yields of dinitroxylenes.

Production. Batchwise nitration of *p*-xylene in mixed acid at 25 – 35 °C with a slight excess of nitric acid gives yields of 85 – 95 %. Conven-

Table 7. Physical properties of nitroxylenes

Compound	CAS registry number	*mp*, °C	*bp* (101 kPa), °C	ϱ, g/cm^3
1,2-Dimethyl-3-nitrobenzene (**29**)	[*83-41-0*]	15	246	1.1402
1,2-Dimethyl-4-nitrobenzene (**30**)	[*99-51-4*]	28 – 30	258	1.139
1,3-Dimethyl-2-nitrobenzene (**31**)	[*81-20-9*]	15	222	1.112
1,3-Dimethyl-4-nitrobenzene (**32**)	[*89-87-2*]	9	244	1.126
1,4-Dimethyl-2-nitrobenzene (**33**)	[*89-58-7*]	25	240	1.132

tional separation and workup lead to fairly pure 1,4-dimethyl-2-nitrobenzene (**33**). The mixture of **31** and **32** obtained from *m*-xylene is distilled to remove **31** and leave the main product 1,3-dimethyl-4-nitrobenzene (**32**). The mixture of **29** and **30** obtained from *o*-xylene is much more difficult to separate because the ratio is nearly 1 : 1 and their boiling points are closer. Initial fractionation by vacuum distillation is followed by a sweating process (dependent on melting point differences) to complete the purification [1].

Uses. The sole outlet for nitroxylenes is via their reduction to the corresponding xylidines. 3,4-Dimethylaniline [*95-64-7*] (3,4-xylidine, **34**), and 2,6-dimethylaniline [*87-62-7*] (2,6-xylidine, **35**) are starting materials for the production of riboflavin and agrochemicals, respectively. Other isomers are used as colorant intermediates; the most important of these are 2,4-dimethylaniline [*95-68-1*] (2,4-xylidine, **36**), used as a diazo component (e.g., in C.I. Solvent Orange 7 and C.I. Acid Red 26) and 2,5-dimethylaniline [*95-78-3*] (2,5-xylidine, **37**) used as a coupling component (e.g., in Dispersol Orange 7 and C.I. Solvent Red 26). Conveniently, mixed xylidines can be used for some colorants.

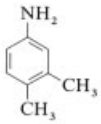

34

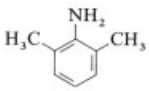

35

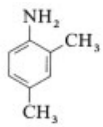

36

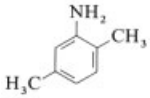

37

3.4. Nitronaphthalenes

Under moderate conditions, naphthalene is nitrated to give 95 % 1-nitronaphthalene (**38**) and 5 % 2-nitronaphthalene (**39**). FIERZ – DAVID [43] obtained this ratio, together with traces of dinitronaphthalene and dinitronaphthol, under a variety of conditions: (1) 20 % nitric acid at 95 – 98 °C, (2) 95 % nitric acid in acetic acid and acetic anhydride at 50 – 70 °C, and (3) HNO_3 – H_2SO_4 at 60 °C. The last conditions form the basis of large-scale production. Mixed acid nitration at higher temperature (80 – 100 °C) results in a mixture of 1,5-dinitronaphthalene (**40**) and 1,8-dinitronaphthalene (**41**) in ca. 2 : 3 ratio. This demonstrates deactivation of the ring containing the first-entering group, with the peri position then being favored in the second ring. Unlike the sulfonic acid substituent, the nitro group does not undergo migration in the naphthalene ring, and isomer ratios are therefore more constant and controllable.

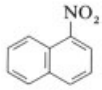

38

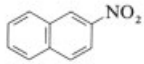

39

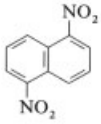

40

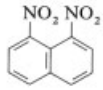

41

All ten dinitronaphthalene isomers and many trinitro and tetranitro derivatives have been described in the literature, but only the commercially important 1,5- and 1,8-isomers are described here, together with 1- and 2-nitronaphthalenes.

Nitration of naphthalene mono- and disulfonic acids, leading to an important series of aminonaphthalenesulfonic acids, is usually performed with the appropriately formed sulfonation mass (→ Naphthalene Derivatives).

1-Nitronaphthalene [*86-57-7*] (**38**), α-nitronaphthalene, $C_{10}H_7NO_2$, M_r 173.17, *mp* 52 (metastable) – 61 °C, *bp* 304 °C (101 kPa), forms pale yellow, odorless needles that sublime at 30 – 50 °C (1.33 Pa). It is soluble in ethanol, chloroform, or ether, but insoluble in water.

Halogenation of 1-nitronaphthalene occurs in the 5- and 8-positions, and sulfonation results in 5-nitronaphthalene-1-sulfonic acid, although the latter is produced on a large scale by nitration of naphthalene-1-sulfonic acid. When chlorine gas is passed into molten 1-nitronaphthalene, the nitro group is replaced, and a mixture of chloronaphthalenes is formed.

Hydrogenation or iron reduction results in 1-aminonaphthalene; intermediate azoxy, azo, and hydrazo stages can be obtained by reduction with Zn – NaOH in alcohol. Controlled re-

duction with aqueous ammonium sulfide under mild conditions yields 1-naphthylhydroxylamine. Oxidation with air in aqueous $Ca(OH)_2$–KOH at 140 °C produces 1-hydroxy-4-nitronaphthalene. Reaction with NH_2OH – NaOEt yields 2-amino-4-nitronaphthalene.

Production. Naphthalene is charged to sulfuric acid alternately with mixed acid at 40 – 50 °C over 8 h, with the final temperature being 55 °C. The acid strength used allows the molten product to be separated as an oil and washed. Continuous operation at 50 – 60 °C with mixed acid (33/48/19) gives a comparable product. The crude reaction product is obtained in 90 – 95 % yield and is purified by vacuum distillation.

Uses. Virtually all of the 1-nitronaphthalene produced is catalytically reduced to 1-naphthylamine (→ Naphthalene Derivatives). Historically, ithas also been used as a debloming agent for petroleum and oils and as a component in the formulation of explosives.

Toxicology [22]. Vapors of 1-nitronaphthalene are poisonous and lacrimatory, being skin, eye, and mucous membrane irritants.

2-Nitronaphthalene [*581-89-5*] (**39**), β-nitronaphthalene, $C_{10}H_7NO_2$, M_r 173.17, *mp* 79 °C, *bp* 312 °C (97.4 kPa), *bp* 165 °C (2 kPa), does not sublime and is volatile in steam, in contrast to 1-nitronaphthalene. It is soluble in ethanol, ether, or chloroform, but insoluble in water.

Halogenation and sulfonation take place mainly in the 5-position, and further nitration gives a mixture of 1,6- and 1,7-dinitro- and 1,3,8-trinitronaphthalenes.

Preparation is carried out by indirect methods, for example, a nitro-Sandmeyer reaction with 2-naphthylamine.

Toxicology. 2-Nitronaphthalene is moderately toxic by ingestion and intraperitoneal routes; it is also a skin and lung irritant. 2-Nitronaphthalene is treated as a human carcinogen because it is thought to be metabolized to 2-aminonaphthalene in the human body. This is most relevant at the purification stage in the production of 1-nitronaphthalene.

1,5-Dinitronaphthalene [*605-71-0*] (**40**), $C_{10}H_6N_2O_4$, M_r 218.17, *mp* 219 °C, forms colorless needles from benzene, acetic acid, or acetone; the compound is potentially explosive [19, 22].

Reduction with iron – acid or hydrogen – catalyst yields 1,5-naphthalenediamine, but Zn – NH_4Cl in ethanol produces 5,5′-dinitro-1,1′-azoxynaphthalene, and reduction with ammonium sulfide gives 5-nitro-1-aminonaphthalene. Nitration of **40** produces 1,4,5-trinitronaphthalene, whereas sulfonation with 20 % oleum results in 1,5-dinitronaphthalene-3-sulfonic acid. Vigorous oxidation with nitric acid at 150 °C gives a mixture of 3,5-dinitrobenzoic acid, 3-nitrophthalic acid, and picric acid.

Production. Naphthalene is added slowly to mixed acid (22/58/20) at 40 °C, and the temperature is raised to 80 °C over 4 h to give a mixture of 1,5- and 1,8-dinitronaphthalenes. The isomers are separated by fractional crystallization (e.g., from ethylene dichloride) or, preferably, by solvent extraction [44]. The 1,5-isomer can be extracted with toluene, leaving 99 % pure 1,8-dinitronaphthalene. After evaporation of the toluene, the residue is extracted with a strongly polar solvent (e.g., sulfolane) to leave 99 % pure 1,5-isomer.

World production of dinitronaphthalenes occurs mainly in the Federal Republic of Germany and Japan, with the capacity of the latter reported to be 1200 t/a.

Uses. 1,5-Dinitronaphthalene is an intermediate in the production of naphthazarin (5,8-dihydroxy-1,4-naphthoquinone) and 1,5-naphthalenediamine, which is mainly converted to naphthalene 1,5-diisocyanate. As a sensitizing agent for ammonium nitrate explosives the use of mixed isomers is adequate.

1,8-Dinitronaphthalene [*602-38-0*] (**41**), $C_{10}H_6N_2O_4$, M_r 218.17, *mp* 172 °C, forms colorless or pale yellow plates from ethanol or chloroform; the compound is potentially explosive. It is sparingly soluble in organic solvents and insoluble in water.

Chlorination yields 1,4,5-trichloronaphthalene, and nitration gives 1,3,8-tri- and 1,3,6,8-tetranitronaphthalenes. Sulfonation with sulfuric acid yields 4,5-dinitronaphthalene-2-sulfonic acid, but treatment with oleum at 40 – 50 °C results in 4-nitroso-5-nitro-1-naphthol. Selective hydrogenation to give 8-nitro-1-aminonaphthalene can be achieved with a platinum black catalyst.

Production. 1,8-Dinitronaphthalene is produced by separation from the mixed isomers obtained on nitration of naphthalene (see above).

Uses. 1,8-Dinitronaphthalene is catalytically hydrogenated to 1,8-naphthalenediamine for use mainly as a colorant intermediate for naphthperinones. If demand for the 1,5-isomer as an isocyanate precursor increases, one outlet for the surplus 1,8-isomer would be direct conversion to sulfur dyes.

3.5. Other Nitro Aromatics

Although ethylbenzene and cumene are readily available and can be nitrated under similar conditions to toluene, their nitro derivatives very rarely supplant the nitrotoluenes as precursors. A startling technical advantage of, for example, an ethyl homologue over a toluene derivative would be required to justify the establishment of large-scale nitro isomer separation. Table 8 records the isomer trends on nitration of higher alkylbenzenes under comparable conditions. The ortho – para ratio decreases incrementally due to steric hindrance [45].

Table 8. Isomer ratios obtained from nitration of monoalkylbenzenes with mixed acid

Starting	Nitro isomers, wt %			Ortho – para
material	Ortho	Meta	Para	ratio
Toluene	58.5	4.5	37	1.58
Ethylbenzene	45.0	6.5	48.5	0.93
Cumene	30.0	7.7	62.3	0.48
tert-Butylbenzene	15.8	11.5	72.7	0.22

Nitration of polyalkylbenzenes (e.g., mesitylene and durene) requires extra care because of oxidation and other side reactions that can easily result from nitrosation and ipso substitution. Trinitromesitylene is prepared readily from mesitylene and mixed acid, and has been used as an explosive [8]. Two nitro derivatives of mesitylene are, however, of more recent importance as intermediates for colorants. Mesidine [*88-05-1*] (2,4,6-trimethylaniline, **42**) is produced by reduction of 1,3,5-trimethyl-2-nitrobenzene [*603-71-4*], which is obtained by nitration of mesitylene under special conditions, e.g., low temperature in the presence of acetonitrile and sulfamic acid [46]. 3,5-Diamino-2,4,6-trimethyl-benzenesulfonic acid [*32432-55-6*] (**43**) is produced by dinitration of the reaction mixture from the monosulfonation of mesitylene followed by reduction of the dinitro mesitylenesulfonic acid.

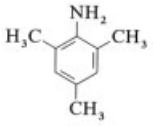

42

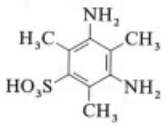

43

4. Nitrohalo Aromatics

4.1. Chloronitrobenzenes

The three isomeric monochloronitrobenzenes and 2,4-dinitrochlorobenzene are industrially important; all except 3-chloronitrobenzene are obtained by nitration of chlorobenzene. The primary outlets for chloronitrobenzenes depend on nucleophilic reactions (e.g., amination and alkoxylation), reduction of the nitro group, and electrophilic ring substitution (e.g., sulfonation). This makes them highly versatile primary intermediates with a wide range of derived products, as shown in Figures 6 and 7, for use in the production of colorants and other effect chemicals [47].

2-Chloronitrobenzene [*88-73-3*] (**44**), OCNB, 2-nitrochlorobenzene, $C_6H_4ClNO_2$, M_r 157.56, crystallizes as light yellow monoclinic needles. The compound is very soluble in ether, benzene, or hot ethanol, but insoluble in water. Physical properties are listed in Table 9, together with those of 3- and 4-chloronitrobenzene for comparison.

Table 9. Physical properties of the three monochloronitrobenzenes

Property	Isomer		
	Ortho	Meta	Para
mp, °C	33	46 (stable) 24 (labile)	83
bp, °C (kPa)	245.5 (100) 119 (1.1)	235.5 (101)	242 (101) 113 (1.1)
Density, g/cm^3 (°C)	1.368 (22) 1.305 (80)	1.534 (20) 1.310 (80)	1.520 (22) 1.303 (85)
Dielectric constant (°C)	37.7 (50)	20.9 (50)	12.7 (83.5)
Flash point (closed cup), °C	127	127	

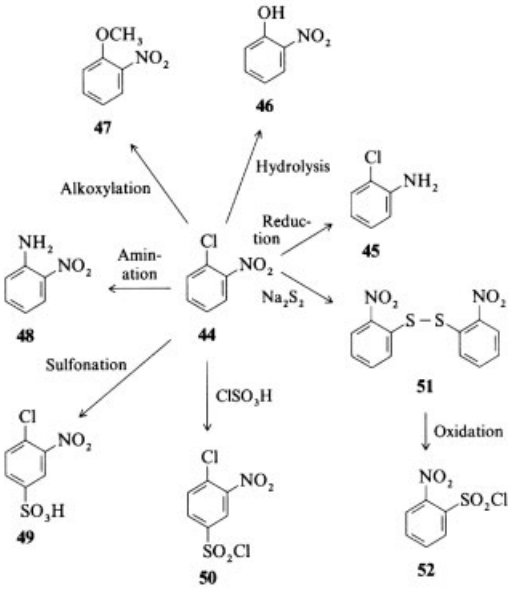

Figure 6. Key intermediates derived from 2-chloronitrobenzene

Reduction of the nitro group with Fe – HCl gives 2-chloroaniline, with alternative reducing agents giving intermediate reduction products analogously to nitrobenzene (Fig. 3). The chlorine atom is readily replaced by nucleophilic attack to give hydroxy, alkoxy, and amino derivatives (Fig. 6). Electrophilic ring substitution (e.g., by halogenation, nitration, or sulfonation) takes place mainly in the position meta to the nitro group.

Production. Nitration of chlorobenzene with mixed acid (30/56/14) typically gives an isomer mix in 98 % yield consisting of 34 – 36 % 2- chloronitrobenzene, 63 – 65 % 4- chloronitrobenzene, and only ca. 1 % 3- chloronitrobenzene. The ortho – para ratio of about 0.55 is in sharp contrast to the 1.6 obtained with nitrotoluene isomers (Section 3.2). As with the nitration of toluene, much work has been done on isomer control so that the producer might have some flexibility towards the balance of isomer demand. Although no major change in ratio has been achieved, the situation is better for nitrochlorobenzene than nitrotoluenes, because the favored isomer is in much greater demand. In further contrast to the nitration of toluene, the nitration of chlorobenzene in the presence of phosphoric acid decreases the proportion of 4- chloronitrobenzene [40],and the use of phosphoric acid in the presence of a transition-metal catalyst is said to increase the ortho – para ratio to ca. 0.8 [48].

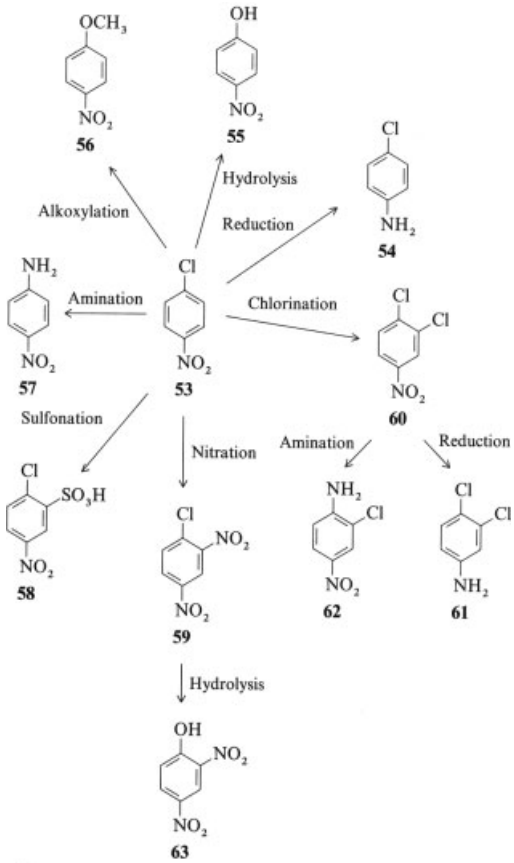

Figure 7. Key intermediates derived from 4-chloronitrobenzene

Even though the rate of nitration of chlorobenzene is an order of magnitude slower than that of benzene, comparable temperatures (40 – 70 °C) are adequate, and the techniques and equipment are very similar to those described for benzene (Section 3.1). In Meissner units the slower reaction rate is compensated by placing additional reactors in series, thereby facilitating flexible operation of this plant type as a basis for multipurpose installations [32].

The mixed product output stream is the same whether produced by a batch or continuous process, and the isomers are separated by a combination of fractional crystallization and distillation. For a simple first separation the isomer mixture is held at a temperature slightly above its crystallization point (15 °C), whereby much

of the 4-chloronitrobenzene crystallizes and can be separated. Fractional distillation gives a para-rich distillate containing all the meta isomer and an ortho-rich still residue. Each of these is crystallized, separated, and the liquid component is refractionated to gradually accumulate high-purity ortho and para products, together with intermediate fractions for continual recycle.

Estimated 1985 product figures for 2- and 4-chloronitrobenzene are for the Federal Republic of Germany 60 000 t/a, the United States 40 000 t/a, and Japan 30 000 t/a.

Uses. 2-Chloronitrobenzene derivatives (Fig. 6) find many outlets in the synthesis of colorants and effect chemicals. Reduction with iron produces 2-chloroaniline [*95-51-2*] (**45**, Fast Yellow G Base), and electrolytic reduction followed by rearrangement of the resulting hydrazo derivative leads to 3,3′-dichlorobenzidine [*91-94-1*], both of which are important diazo components. In the alternative, more economical hydrogenation process, a modified catalyst is required to inhibit dechlorination as a side reaction. Typically, a platinum on carbon catalyst, together with a small quantity of an inorganic acid acceptor such as magnesium oxide, is used; however, up to 2 % aniline may still be formed. The alternative use of morpholine as a dechlorination suppressor, rather than just an acid acceptor, is recommended for 2- and 4-chloronitrobenzene, this reduces the extent of dechlorination to 0.5 % [49]. The use of modified catalyst systems based on platinum, rather than the conventional nickel catalyst used in the production of most aniline derivatives, is essential for hydrogenation of all nitrohalo aromatics.

Treatment of 2-chloronitrobenzene with aqueous sodium hydroxide at 130 °C produces 2-nitrophenol (**46**; cf. Chap. 7), whereas treatment with MeOH – NaOH gives 2-nitroanisole (**47**) or with EtOH – NaOH 2-nitrophenetole (Chap. 7), all of which are used as precursors of the derived amines and many other products. Treatment with aqueous ammonia at 175 °C under pressure yields 2-nitroaniline (**48**; cf. Chap. 5). Sulfonation yields 4-chloro-3-nitrobenzenesulfonic acid (**49**), and chlorosulfonation gives the corresponding 4-chloro-3-nitrobenzenesulfonyl chloride (**50**). Reaction with sodium disulfide gives 2,2′-dinitrodiphenyl disulfide (**51**), which on oxidation with chlorine gives 2-nitrobenzenesulfonyl chloride (**52**). The three derivatives **49**, **50**, and **52** are all precursors to further series of intermediates (cf. Chaps. 6 and 7).

3-Chloronitrobenzene [*121-73-3*] (**7**), 3-nitrochlorobenzene $C_6H_4ClNO_2$, M_r 157.56, forms pale yellow prisms (from ethanol), which can exist as a stable or a labile modification in the solid state. The compound is insoluble in water, readily soluble in benzene or ether, and soluble in acetone or hot ethanol. Physical properties are listed in Table 9 together with those of 2- and 4-chloronitrobenzenes for comparison.

Reduction of the nitro group with Fe – HCl gives 3-chloroaniline, and ring substitution with electrophiles yields mixtures of products. Unlike the 2- and 4-isomers, the chlorine atom of 3-chloronitrobenzene is not activated towards nucleophilic substitution.

Production. Nitrobenzene is chlorinated at 35 – 45 °C in the presence of sublimed iron(III) chloride to give an isomer mixture containing 86 % of the desired 3-isomer. A continuous process has been described that uses a series of reactors operating at 35 – 55 °C, with a residence time of 5 h [50]. Purification is achieved by a combination of distillation and crystallization [51]. Disposal of the byproducts, 10 % 2-chloronitrobenzene and 4 % 4-chloronitrobenzene, can be achieved by feeding them to the mixed isomer stream in the chlorobenzene nitration plant. The 1 % crude 3-chloronitrobenzene from this plant may similarly be fed to the purification stream of the chlorination plant. Final purification of 3-chloronitrobenzene may be achieved chemically by caustic hydrolysis of the residual 2- and 4-chloronitrobenzenes and washing them out as nitrophenols.

Uses. Reduction of 3-chloronitrobenzene to 3-chloroaniline [*108-42-9*] (Orange GC base) is its primary outlet, with minor uses in other fields [51]. Crude 3-chloronitrobenzene can be used for exhaustive chlorination to give pentachloronitrobenzene (Section 4.2).

4-Chloronitrobenzene [*100-00-5*] (**53**), PCNB, 4-nitrochlorobenzene, $C_6H_4ClNO_2$, M_r 157.56, crystallizes as light yellow monoclinic prisms, which are insoluble in water and very soluble in toluene, ether, acetone, or hot ethanol. Physical properties are listed in Table 9 together

with those of 2- and 3- chloronitrobenzenes for comparison.

Reactions of 4- chloronitrobenzene are similar to those of 2- chloronitrobenzene: nitration gives 2,4-dinitrochlorobenzene, chlorination gives 3,4-dichloronitrobenzene, and sulfonation gives 6- chloro-3-nitrobenzenesulfonic acid.

Production. Nitration of chlorobenzene and separation of the isomers have already been described for 2- chloronitrobenzene.

Uses. 4-Chloronitrobenzene derivatives find important outlets in many key intermediate chains (Fig. 7). The first stages in these chains are 4- chloroaniline [*106-47-8*] (**54**), 4-nitrophenol (**55**), 4-nitroanisole (**56**) (cf. Chap. 7), 4-nitroaniline (**57**; cf. Chap. 5), 6-chloro-3-nitrobenzenesulfonic acid (**58**; cf. Chap. 6), 2,4-dinitrochlorobenzene (**59**), and 3,4-dichloronitrobenzene (**60**).

Reaction with sodium sulfide can be controlled to give 4-amino-4′-nitrodiphenyl sulfide (**64**), which after N-acetylation and oxidation with hydrogen peroxide forms **65**: subsequent reduction and hydrolysis yields 4,4′-diaminodiphenyl sulfone [*80-08-0*] (dapsone, **66**), a valuable drug in the treatment of leprosy. Dapsone is also used as an analytic reagent and can alternatively be obtained via the reaction of 4- chloronitrobenzene with 4-acetylaminobenzenesulfinic acid to give (**65**).

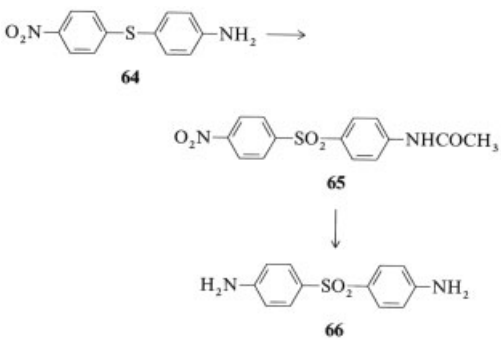

Condensation of 4- chloronitrobenzene with 2,4-dichlorophenol gives 2,4-dichloro-4′-nitrodiphenyl ether [*1836-75-5*] (**67**) which is used as a herbicide (nitrofen). There are many related products of this type; for example, the more recently developed preemergent herbicide Fluoronitrofor [*13738-63-1*] (**68**) – derived by reaction of 4- chloronitrobenzene with 2,4-dichloro-6-fluorophenol.

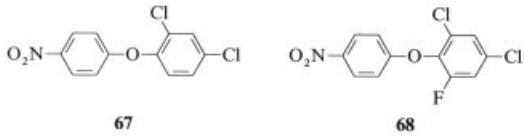

A major outlet for 4- chloronitrobenzene exploits the condensation with aniline to give 4-nitrodiphenylamine [*836-30-6*] (PNDPA, **69**), which on reductive N-alkylation gives important antioxidants for rubber, e.g., IPPD [*101-72-4*] (**70**).

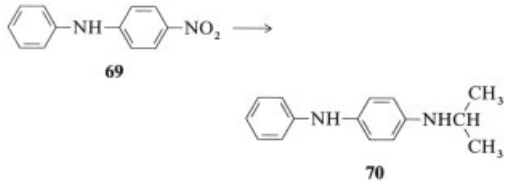

1-Chloro-2,4-dinitrobenzene [*97-00-7*] (**59**), 2,4-dinitrochlorobenzene, DNCB, $C_6H_3ClN_2O_4$, M_r 202.56, *mp* 53.4 °C (α), 43 °C (β), 27 °C (γ), *bp* 315 °C (101 kPa, decomp.), d_4^{75} 1.4982, flash point (closed cup) 194 °C. The stable α-form is obtained as yellow rhombic crystals from ether; two other unstable crystal modifications are also known. The compound is soluble in benzene or ether, and insoluble in water.

Reduction of **59** with Fe – HCl gives 2,4-diaminochlorobenzene, but selective reduction is difficult because of the high reactivity of the chlorine atom toward basic reagents. Pyridine reacts readily to form a stable pyridinium salt. Nucleophilic substitution reactions are similar to those of 2- chloronitrobenzene, except that the greatly increased activity of the chlorine atom (usually several thousandfold) means that much milder reaction conditions suffice. Further substitution on the aromatic ring is difficult, although nitration under forcing conditions yields 2,4,6-trinitrochlorobenzene [*88-88-0*] (picryl chloride).

Production. Although chlorobenzene can be dinitrated directly, this results in unnecessary isomer problems. 4-Chloronitrobenzene is usually nitrated with mixed acid (35/65) at 60 °C to give the pure dinitro isomer. However, 2-chloronitrobenzene can be nitrated to produce

1-chloro-2,4-dinitrobenzene, together with ca. 10 wt % of the isomeric 2-chloro-1,3-dinitrobenzene. This may be separated for disposal, but the mixed isomers are preferably used directly if tolerated by the end product (e.g., sulfur dyes).

Uses. Industrially, the most important derivatives of 1-chloro-2,4-dinitrobenzene are obtained by nucleophilic reactions in aqueous media at moderate temperature. Ammonia gives 2,4-dinitroaniline (Chap. 5), alkali gives 2,4-dinitrophenol (Chap. 7), and methanolic sodium hydroxide gives 2,4-dinitroanisole (Chap. 7). Refluxing with hydrazine in ethanol yields 2,4-dinitrophenylhydrazine [*119-26-6*], a reagent used for the characterization of carbonyl compounds. Reaction with ammonium thiocyanate in aqueous medium at 80 °C gives dinitrophenylrhodanate, which was used as an insecticide. Reaction with substituted anilines gives 2,4-dinitrodiphenylamine derivatives that are used as yellow disperse dyes (→ Nitro and Nitroso Dyes). Reaction with pyridine gives the reactive (2,4-dinitrophenyl)pyridinium chloride [*4185-69-7*], an intermediate in the preparation of pentamethine dyes.

Nitrochlorobenzenes, especially dinitrochlorobenzene, have traditionally been used to produce sulfur dyes; for example, C.I. Sulfur Black 1 is obtained from 2,4-dinitrochlorobenzene by prolonged refluxing with sodium polysulfide liquor.

Toxicity of Nitrochlorobenzenes. The mononitrochlorobenzenes (TLV 1 mg/m^3) are considered more toxic than nitrobenzene. They are absorbed through the skin and lungs, with lowering of the hemoglobin level and its consequential effects. 2,4-Dinitrochlorobenzene is similarly toxic and is also an extreme irritant to the skin, frequently leading to skin eruptions and allergic dermatitis [22]. These allergic properties are probably due to the antigen–antibody reaction resulting from bonding with natural proteins (cf. the Sanger test with 2,4-dinitrofluorobenzene).

4.2. Dichloro- and Polychloronitrobenzenes

Both 1,2- and 1,4-dichlorobenzene are nitrated on a large scale to give important intermediates for colorants via reduction or nucleophilic substitution reactions analogous to those used for 2- and 4-chloronitrobenzene and shown in Figures 6 and 7.

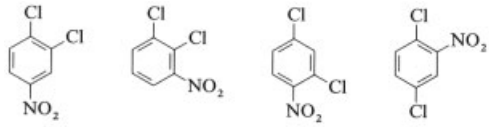

60 **71** **72** **73**

1,2-Dichloro-4-nitrobenzene [*99-54-7*] (**60**), DCNB, 3,4-dichloronitrobenzene, $C_6H_3Cl_2NO_2$, M_r 192.01, crystallizes as yellow needles from ethanol. It is insoluble in water, and soluble in benzene, ether, or hot ethanol. The stable α-form has a *mp* of 43 °C; the unstable β-form is a liquid that reverts to the α-form at 15 °C, *bp* 263 °C (101.3 kPa), d_4^{75} 1.4588.

Production. Nitration of 1,2-dichlorobenzene with mixed acid at 35–60 °C results in a mixture of 3-nitro (10 %) and 4-nitro (90 %) isomers, which are separated by crystallization. A more expensive process (two stages from chlorobenzene) is based on chlorination of molten 4-chloronitrobenzene (90–100 °C) by a process similar to that used for 3-chloronitrobenzene production. Chlorination has the advantage of giving a pure product, but it cannot be used if the isomeric 1,2-dichloro-3-nitrobenzene is also required.

Uses. Iron reduction of **60** yields 3,4-dichloroaniline [*95-76-1*], an important agrochemical intermediate. Amination produces 2-chloro-4-nitroaniline, which is an important diazo component and intermediate (Section 5.3).

Reaction of 1,2-dichloro-4-nitrobenzene with potassium fluoride in an aprotic solvent gives 3-chloro-4-fluoronitrobenzene (**74**); the derived amine is used as a herbicide intermediate.

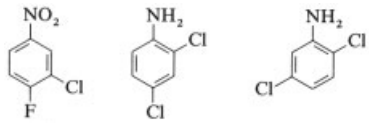

74 **75** **76**

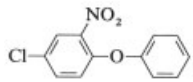

77

1,2-Dichloro-3-nitrobenzene [*3209-22-1*] (**71**), 2,3-dichloronitrobenzene, 3-nitro-1,2-dichlorobenzene, $C_6H_3Cl_2NO_2$, M_r 192.01, *mp* 60 °C, *bp* > 230 °C (decomp.), d_4^{80} 1.449, crystallizes as yellow monoclinic needles, which are moderately soluble in most organic solvents and insoluble in water.

Production. 1,2-Dichloro-3-nitrobenzene is a byproduct of the nitration of 1,2-dichlorobenzene.

1,3-Dichloro-4-nitrobenzene [*611-06-3*] (**72**), 2,4-dichloronitrobenzene, $C_6H_3Cl_2NO_2$, M_r 192.01, *mp* 34 °C, *bp* 101 °C (0.5 kPa), d_4^{78} 1.551, crystallizes as pale yellow needles from ethanol.

Production. 1,3-Dichlorobenzene is nitrated with mixed acid, and the product is separated from the resulting isomer mixture.

Uses. 1,3-Dichloro-4-nitrobenzene is reduced to the intermediate 2,4-dichloroaniline [*554-00-7*] (**75**). Preferential amination in the ortho position can be effected to give 5-chloro-2-nitroaniline [*1635-61-6*] which is, however, much less important than the other chloronitroanilines (see Section 5.3).

1,4-Dichloro-2-nitrobenzene [*89-61-2*] (**73**), 2,5-dichloronitrobenzene, 2-nitro-1,4-dichlorobenzene, $C_6H_3Cl_2NO_2$, M_r 192.01, *mp* 56 °C, *bp* 266 °C (101 kPa), d_4^{75} 1.4390, crystallizes as pale yellow prisms or plates from ethanol. The compound is insoluble in water and soluble in benzene, ether, or hot ethanol.

Production. Nitration of 1,4-dichlorobenzene with mixed acid at 35 – 65 °C results in a 98 % yield of essentially pure 1,4-dichloro-2-nitrobenzene.

Uses. Derivatives of **73** are used mainly as colorant intermediates. Reduction yields 2,5-dichloroaniline [*95-82-9*] (Fast Scarlet GG Base, **76**), which is sulfonated to give the equally important 2,5-dichloroaniline-4-sulfonic acid [*88-50-6*] (Fast Red FR base). Amination produces 4-chloro-2-nitroaniline (Section 5.3); hydrolysis, 4-chloro-2-nitrophenol (Section 7.1); and methoxylation, 4-chloro-2-nitroanisole (Section 7.1). Condensation with phenol gives 4-chloro-2-nitrodiphenyl ether (**77**), which is reduced to 4-chloro-2-aminodiphenyl ether (Red FG Base); a series of similarly prepared aminodiphenyl ethers are used as diazo components.

Condensation with arylamines gives nitrodiphenylamines that are used as yellow to orange disperse dyes (→ Nitro and Nitroso Dyes).

1,2,4-Trichloro-5-nitrobenzene [*89-69-0*] (**78**), 1-nitro-2,4,5-trichlorobenzene, $C_6H_2Cl_3NO_2$,

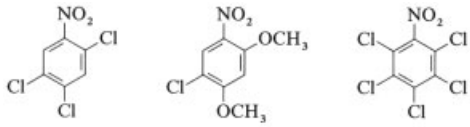

78 **79** **80**

M_r 226.46, *mp* 57 °C, *bp* 288 °C (101 kPa), d_4^{80} 1.554, crystallizes as yellow prisms from ethanol.

Production. 1,2,4-Trichlorobenzene is nitrated to give mainly the 5-nitro compound, which is purified by fractional crystallization.

Uses. Reduction of **78** gives 2,4,5-trichloroaniline [*636-30-6*], and methoxylation yields 5-chloro-2,4-dimethoxynitrobenzene (**79**) [*119-21-1*]; the latter is reduced to 5-chloro-2,4-dimethoxyaniline [*97-50-7*].

Pentachloronitrobenzene [*82-68-8*] (**80**), $C_6Cl_5NO_2$, M_r 295.36, *mp* 144 °C, *bp* 328 °C (101 kPa, decomp.), d_4^{20} 1.718, forms yellow crystals, which are moderately soluble in organic solvents and insoluble in water.

Production. Pentachloronitrobenzene is produced by nitration of pentachlorobenzene.

Uses. The early use of **80** as a fungicide (terrachlor) has led to several series of nitro-containing agrochemicals. Reduction gives pentachloroaniline [*527-20-8*], which has limited application. Reaction with ethanol – potassium hydroxide yields pentachlorophenetole, exemplifying an easy route to pentachlorophenyl ethers.

Toxicity of Di- and Polychloronitrobenzenes. All di- and polychloronitrobenzenes are moderately toxic [22] and should be handled as potential eye, nose, and skin irritants.

4.3. Chloronitrotoluenes

Of the ten chloronitrotoluene isomers available, only five are of potential technical interest. Both

nitration of chlorotoluenes and chlorination of nitrotoluenes result in isomer problems, but the latter is more substituent directive.

Side-chain chlorination of nitrotoluenes is mentioned in Section 3.2, but only 4-nitrobenzyl chloride (**24**) is technically important.

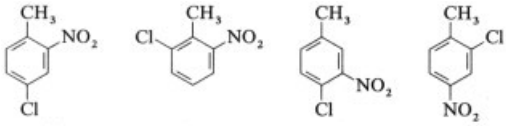

16 **17** **81** **23**

4-Chloro-2-nitrotoluene [*89-59-8*] (**16**), 2-nitro-4-chloro-1-methylbenzene, $C_7H_6ClNO_2$, M_r 171.58, *mp* 38.5 °C, *bp* 240 °C (96 kPa), d_4^{80} 1.2559, crystallizes as yellow monoclinic needles.

Production. 4-Chlorotoluene is nitrated with mixed acid (39/59/2) at 25 °C to give a mixture of 2-nitro (65 %) and 3-nitro (35 %) isomers. The major component (**16**) is separated as the lower boiling fraction on vacuum distillation. 4-Chloro-3-nitrotoluene [*89-60-1*] (**81**) can be obtained, if required, from the residue by distillation and sweating.

Uses. Reduction of **16** gives 4-chloro-2-toluidine [*95-81-8*] (Fast Red KB Base, **82**). The byproduct **81** can be methoxylated and reduced to produce cresidine [*120-71-8*] (**83**), which can, however, be obtained by several alternative routes.

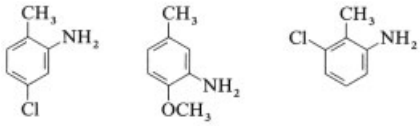

82 **83** **84**

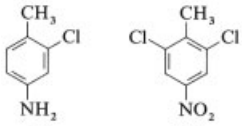

85 **86**

2-Chloro-6-nitrotoluene [*83-42-1*] (**17**), 6-chloro-2-nitrotoluene, 2-nitro-6-chloro-1-methylbenzene, $C_7H_6ClNO_2$, M_r 171.58, *mp* 37.5 °C, *bp* 238 °C (101 kPa).

Production. Chlorination of 2-nitrotoluene, in a process similar to the chlorination of nitrobenzene, gives a mixture of 6-chloro (80 %) and 4-chloro (20 %) isomers. The major component **17** is separated by vacuum distillation.

Uses. Reduction of **17** gives 6-chloro-*o*-toluidine [*87-63-8*] (Fast Scarlet TR Base, **84**) which, apart from forming azo dyes, is used to produce 2,6-dichlorotoluene, by Sandmeyer reaction, for conversion to 2,6-dichlorobenzaldehyde.

2-Chloro-4-nitrotoluene [*121-86-8*] (**23**), 4-nitro-2-chloro-1-methylbenzene, $C_7H_6ClNO_2$, M_r 171.58, *mp* 68 °C, *bp* 264 °C (101 kPa), crystallizes as yellow needles from ethanol. The compound is soluble in most organic solvents, only slightly soluble in hot water, and steam volatile.

Production. Chlorination of 4-nitrotoluene at 50 °C in the presence of iron(III) chloride catalyst gives good conversion to 2-chloro-4-nitrotoluene, the most readily obtained of all the chloronitrotoluene isomers.

Uses. Reduction of **23** gives 2-chloro-4-toluidine [*95-74-9*] (**85**), which undergoes sulfonation to give 4-amino-6-chlorotoluene-3-sulfonic acid [*88-51-7*]. Further chlorination of **23** gives 2,6-dichloro-4-nitrotoluene [*7149-69-1*] (**86**) for reduction to 2,6-dichloro-*p*-toluidine [*56461-98-4*].

4-Nitrobenzyl chloride [*100-14-1*] (**24**), 1-chloromethyl-4-nitrobenzene, $C_7H_6ClNO_2$, M_r 171.58, *mp* 71 °C, forms pale yellow plates or needles from ethanol.

Production. Chlorination of 4-nitrotoluene at 190 °C without catalyst gives **24**. The alternative process of nitration of benzylchloride coproduces 2-nitrobenzyl chloride (**28**) and requires a difficult separation stage.

Uses. 4-Nitrobenzyl chloride is used as an alkylating agent (e.g., in a biosynthetic penicillin) and in the production of 4-nitrobenzyl cyanide [*555-21-5*] (**87**), which is hydrolyzed to 4-nitrophenylacetic acid [*104-03-0*] (**88**), an acylating agent.

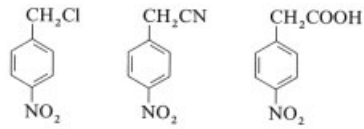

24 **87** **88**

Toxicity of Chloronitrotoluenes. In the absence of detailed information on individual products, all chloronitrotoluenes should be assumed to have a toxicity similar to that of chloronitrobenzenes and should be handled accordingly.

4.4. Fluoronitrobenzenes and Fluoronitrotoluenes

The two most technically important classes of fluoronitrobenzenes and fluoronitrotoluenes are (1) those fluoronitrobenzenes that can be manufactured conveniently by halogen-exchange reactions, with the readily available chloronitrobenzenes and (2) nitration products from benzotrifluoride and chlorobenzotrifluorides.

Nitration of fluorobenzene with mixed acid gives a 92 : 8 mixture of 4-fluoronitrobenzene [*350-46-9*] (**89**) and 2-fluoronitrobenzene [*1493-27-2*] (**90**). However, production of these two compounds and of 2,4-dinitrofluorobenzene [*70-34-8*] (**91**) is always based on fluorine-exchange reaction of the corresponding chloro compound by treatment with potassium fluoride in a polar solvent such as dimethylformamide or dimethyl sulfoxide [52]. These compounds and their homologues form the basis for many substituted fluoro aromatics by further substitution of the nitro compounds or their derived amines. Fluoro aromatics often justify the additional costs over the corresponding chloro compounds in agrochemicals because of the lower dosage required and their greater selectivity and safety. Examples are the fungicide fluoroimide [*41205-21-4*] (**92**), derived from 4-fluoroaniline, and the herbicide flampropisopropyl [*52756-22-6*] (**93**), from 3-chloro-4-fluoroaniline (the reduction product of **74**).

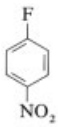

89

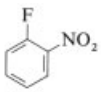

90

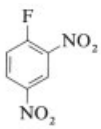

91

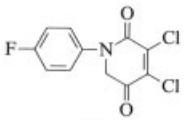

92

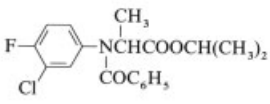

93

Ring-substituted fluoronitrotoluenes have no particular advantage over the available fluoronitrobenzenes, but side-chain fluorinated toluenes (i.e., substituted benzotrifluorides) are of great importance. The I.G. Farbenindustrie produced a large range of substituted aminobenzotrifluorides as speciality fast salts that give Naphtol azo derivatives of high lightfastness. Most of these amines were derived from the corresponding nitro compounds, and some products have survived as speciality dye components. More important, the chemistry of these substituted benzotrifluorides has been exploited in the production of agrochemicals containing the trifluoromethyl group.

Benzotrifluoride is nitrated to give 3-nitrobenzotrifluoride [*98-46-4*] (**94**) in high yield, even after separation of the coproduced 6 % ortho and 3 % para isomers. Under similar conditions, 2-chlorobenzotrifluoride yields 2-chloro-5-nitrobenzotrifluoride [*777-37-7*] (**95**), and 4-chlorobenzotrifluoride yields 4-chloro-3-nitrobenzotrifluoride [*121-17-5*] (**96**). Forcing nitration of **94**, **95**, or **96** at 100 °C gives the 3,5-dinitro derivatives, with 4-chloro-3,5-dinitrobenzotrifluoride [*393-75-9*] (**97**) being the most important example.

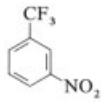

94

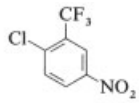

95

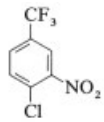

96

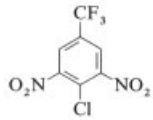

97

Nitration of 3-chlorobenzotrifluoride gives a mixture of isomers from which the desired 5-chloro-2-nitrobenzotrifluoride [*118-83-2*] (**98**) is separated by crystallizing it at 5 °C and centrifuging off the unwanted isomer as an oil.

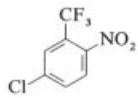

98

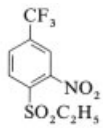

99

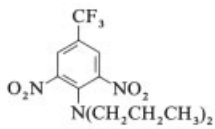

100

Iron reduction of the mononitro compounds **94**, **95**, **96**, and **98** gives a series of trifluoromethylanilines used as diazo compounds, e.g., 2- chloro-5-trifluoromethylaniline [*121-50-6*] from **96** as Fast Orange RD Base and 4- chloro-2-trifluoromethylaniline [*445-03-4*] from **98** as Scarlet VD Base. Reaction of **96** with sodium ethyl sulfinate gives the nitrosulfone (**99**), which on reduction gives 3-amino-4-ethylsulfonylbenzotrifluoride [*382-85-4*]; used as Fast Golden Orange GR Base. 3-Aminobenzotrifluoride [*98-16-8*], obtained by hydrogenation of **94**, is more important as an intermediate for drugs and crop protection chemicals than for colorants.

Reaction of **97** with dialkylamines produces a range of 2,6-dinitro-4-trifluoromethyl-*N*,*N*-dialkylanilines used as preemergent herbicides, typified by trifluralin [*1582-09-8*] (**100**).

The fluorine atom in dinitrofluorobenzenes is readily displaced, as evidenced by the use of 2,4-dinitrofluorobenzene (**91**) in the Sanger method for determination of the N-terminal amino group of a peptide chain.

5. Nitroamino Aromatics

All the isomeric mononitroanilines and many of their derivatives are industrially important. Production as colorant intermediates has made them available as starting materials for many other outlets. The most important additional substituents are nitro (i.e., dinitroanilines), chloro (i.e., chloronitroanilines), and methyl (i.e., nitrotoluidines).

The aromatic nitroamino compounds are weakly basic amines due to the strongly electron-withdrawing nature of the nitro group, especially when it is in a position ortho or para to the amino group. Widely varying p*K* values have been reported, and a comparison of their ionization constants relative to 4-nitroaniline (Table 10) is useful to indicate the extent of protonation in different strengths of sulfuric acid at different temperatures [53]. This is important when nitroanilines undergo reaction (e.g., nitration or diazotization) in sulfuric acid, although solubility must also be considered.

An alternative measure of the basicity of nitroanilines is to consider their polarity as exhibited by the main absorption maximum in a comparable series of monoazo dyes, for example, as exemplified by **101**.

Table 10. pK_a values for the ionization of primary nitroanilines in aqueous H_2SO_4

Compound	25 °C	90 °C
4-Nitroaniline	1.00	0.60
2-Nitroaniline	− 0.30	− 0.51
4-Chloro-2-nitroaniline	− 1.06	− 1.15
2,5-Dichloro-4-nitroaniline	− 1.75	− 1.76
2-Chloro-6-nitroaniline	− 2.38	− 2.35
2,6-Dichloro-4-nitroaniline	− 3.27	− 3.12
2,4-Dinitroaniline	− 4.27	− 3.81
2,6-Dinitroaniline	− 5.39	− 4.72
2-Bromo-4,6-dinitroaniline	− 6.69	− 5.81
2,4,6-Trinitroaniline	− 10.03	− 8.67

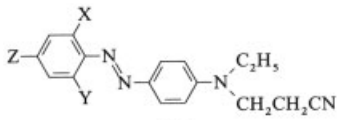

101

Table 11 shows the trend of bathochromic shift with increasing polarity [54]. These absorption maxima are not necessarily related directly to the p*K* values of the nitroanilines because the steric effects of the ortho substituents may differ in each series. Crowding due to the two chlorine substituents in example 4 results in loss of molecular planarity and broadening of the absorption band. The *o*- cyano group (examples 6 and 8) does not exert steric effects in the same way and therefore has a more powerful effect than a nitro group in the same position. 2,6-Dicyano-4-nitroaniline (example 8) is the most bathochromic nitroaniline in this series for comparison with nitro heterocyclic amines used in monoazo dyes (cf. Chap. 9).

Table 11. Substituent effect on absorption in a series of nitroaniline-based dyes (**101**)

Example	X	Y	Z	λ_{max} (in MeOH), nm
1	NO_2	H	H	425
2	H	H	NO_2	453
3	Cl	H	NO_2	475
4	Cl	Cl	NO_2	417
5	NO_2	H	NO_2	491
6	CN	H	NO_2	504
7	NO_2	NO_2	NO_2	520
8	CN	CN	NO_2	549

Nearly all the primary nitroanilines have found use as diazo components, mainly in pig-

ments or disperse dyes, and the most weakly basic of them (e.g., examples 4–8 in Table 11) require diazotization with nitrosylsulfuric acid. Isolated diazo compounds have long been known to be unstable, and often detonatable. However, not until a serious accident in 1969 during diazotization of 6-chloro-2,4-dinitroaniline with nitrosylsulfuric acid were concentrated solutions of certain diazotized nitroanilines in sulfuric acid found to be detonatable [55].

Although aniline and *C*-alkylanilines can be nitrated, this is not a viable route for production of the more important nitroanilines. However, nitrations of this type are important in studying the effects of cationic species (positive poles) on aromatic nitration. Careful analysis of the classical work claiming complete meta nitration of PhN^+Me_3 showed that 89 % of the meta isomer and 11 % of the para isomer are produced, and that the proportion of para substitution increases for PhN^+HMe_2 (22 %), PhN^+H_2Me (30 %), and PhN^+H_3 (38 %), all nitrated at 25 °C in 98 % sulfuric acid. Direct nitration of aniline is sensitive to sulfuric acid concentration, with para nitration increasing to 52 % in 89 % sulfuric acid, but this does not correlate with the calculated amount of protonated amine, so both meta and para substitution must result from nitration of the anilinium ion [56].

Nitration of *o*- and *p*-toluidines at low temperature gives desirable isomers by more controlled meta substitution than with aniline. Nitrochlorotoluenes with the required substitution pattern are less readily available than nitrochlorobenzenes or nitrodichlorobenzenes, which are used predominantly as precursors in the manufacture of nitroanilines and chloronitroanilines by nucleophilic amination of the active chlorine atom. When 3-nitroaniline configurations are required, partial reduction of the dinitro compound is an option, but difficulties arise if the compound is unsymmetrically substituted.

The primary reactions undergone by nitroamino aromatics are diazotization, reduction to the corresponding diamine, and various N-acylation reactions. As diazo components, the nitroanilines and many of their derivatives found early use as fast bases for azoic dyes. The fast base name indicates the hue derived on diazotization and coupling with Naphtol AS.

5.1. Nitroanilines

2-Nitroaniline [*88-74-4*] (**48**), *o*-nitroaniline, ONA, $C_6H_6N_2O_2$, M_r 138.13, crystallizes as orange-yellow plates from water. The compound is moderately soluble in organic solvents and sparingly soluble in water (0.13 %). Physical properties are listed in Table 12 alongside those of *m*- and *p*-nitroanilines for comparison.

Table 12. Physical properties of the three mononitroanilines

Property	Isomer		
	Ortho	Meta	Para
mp, °C	71.5	114	148
bp, °C	284 (decomp.)	307 (decomp.)	331 (decomp.)
Density (15 °C), g/cm^3	1.442	1.398	1.437
Dipole moment (in benzene), D	4.23	4.68	6.2
Flash point, °C	167		199
Refractive index (°C)	1.5362 (100)	1.595 (111)	1.5401 (160)

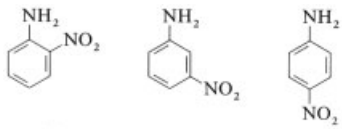

48 **102** **55**

Production. 2-Chloronitrobenzene (Section 4.1) is heated with excess (10 mol/mol) strong aqueous ammonia in an autoclave. The temperature is gradually increased to 180 °C over 4 h and held there for 5 h more. The pressure builds up to around 4 MPa and is released to an ammonia recycle loop before the product is isolated by filtration and washing. The reaction is extremely exothermic ($\Delta H = -168$ kJ/mol), and too rapid heating or inadequate temperature control can result in a runaway reaction. Because of this hazard, I.G. Farbenindustrie [1] developed a continuous amination unit [17] for amination of chloronitrobenzenes; the process is summarized under 4-nitroaniline.

Uses. 2-Nitroaniline is reduced with Fe–acid or H_2–catalyst to *o*-phenylenediamine [*95-54-5*]. Acetylation followed by reduction gives 2-aminoacetanilide [*555-48-6*]. 2-Nitroaniline is used as a diazo component (Fast

Orange GR Base) in azo dyes and pigments (e.g., C.I. Disperse Yellow 10 and C.I. Pigment Yellow 5) and as an intermediate for vat dyes (e.g., C.I. Vat Red 14).

3-Nitroaniline [*99-09-2*] (**102**), *m*-nitroaniline, MNA, $C_6H_6N_2O_2$, M_r 138.13, crystallizes as yellow needles from water. It is moderately soluble in organic solvents and sparingly soluble in water (0.11 %). Its physical properties are listed in Table 12.

Production. 1,3-Dinitrobenzene (Section 3.1) is added to warm water containing magnesium sulfate. An aqueous solution of sodium hydrogen sulfide (6 molar equivalents) is added gradually to the vigorously stirred emulsion, and reduction is completed by heating to 90 °C. The 3-nitroaniline produced solidifies on cooling and is separated by filtration.

Uses. Acetylation of 3-nitroaniline followed by reduction gives 3-aminoacetanilide [*102-28-3*]. Diazotization, followed by reduction of the diazosulfonate with ammonium bisulfite and subsequent hydrolysis, gives (3-nitrophenyl)hydrazine [*619-27-2*]; an intermediate in the production of pyrazolone azo coupling components. 3-Nitroaniline is used as a diazo component (Fast Orange R Base) in azo dyes (e.g., C.I. Disperse Yellow 5 and C.I. Acid Orange 18).

4-Nitroaniline [*100-01-6*] (**55**), *p*-nitroaniline, PNA, $C_6H_6N_2O_2$, M_r 138.13, crystallizes as yellow needles from water. 4-Nitroaniline is moderately soluble in organic solvents and sparingly soluble in water (0.08 %). Its physical properties are listed in Table 12.

Production. Batchwise amination of 4-chloronitrobenzene (Section 4.1) is carried out at 195 °C (4.5 MPa) in a process similar to that for 2-nitroaniline.

In the alternative continuous process [1], molten 4-chloronitrobenzene and 40 % ammonia liquor in a molar ratio of 1 : 17 are pumped at 20 MPa through a steel reaction coil held at 237 – 240 °C by oil bath heating. Inflow is controlled to give a contact time of ca. 10 min, and outflow is directed into an expansion chamber with a tangential flow of hot water to give, in quantitative yield, a suspension of product that is readily isolated and washed.

The alternative historical process based on nitration of acetanilide is not economical compared with amination, but nitration of formanilide followed by alcoholysis and recycle of ethyl formate brings the costs closer [57].

Uses. The volume outlet for 4-nitroaniline is in iron reduction or catalytic hydrogenation to produce *p*-phenylenediamine (PPD) [*106-50-3*]. This is still the major route to *p*-phenylenediamine despite more recent alternatives [21]. Reaction of 4-nitroaniline with acetic anhydride gives 4-nitroacetanilide [*104-04-1*] (**103**) which, on reduction, yields 4-aminoacetanilide [*122-80-5*], an important colorant intermediate. Chlorination (aqueous $HCl - H_2O_2$) gives 2,6-dichloro-4-nitroaniline [*99-30-9*] (**104**), a diazo component specifically for yellow-brown disperse dyes (e.g., C.I. Disperse Brown 1). Bromination (aqueous $HBr - H_2O_2$) gives 2-bromo-4-nitroaniline [*13296-94-1*] (**105**), which is occasionally used as an alternative diazo component to 2- chloro-4-nitroaniline.

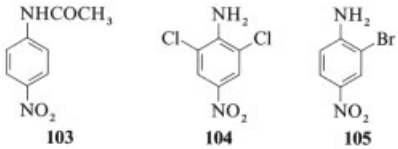

Diazotization of 4-nitroaniline followed by reduction with $(NH_4)_2SO_3 - NH_4HSO_3$ and hydrolysis gives (4-nitrophenyl)hydrazine [*100-16-3*], an intermediate in the production of pyrazolone azo coupling components. Reaction of 4- chloro-1-nitrobenzene with hydrazine is not a viable alternative.

4-Nitroaniline is used as a diazo component (Fast Red GG Base) in the production of disperse monoazo dyes (e.g., C.I. Disperse Orange 25, example 2 in Table 11) and pigments (e.g., C.I. Pigment Red 1). It is also used in dyes such as C.I. Direct Black 19, where the 4-nitroaniline monoazo dye is reduced with sodium sulfide to the aminoazo compound for further diazotization and coupling. Coupling of diazotized 4-nitroaniline with 2,5-dimethoxyaniline produces Black K Base (**106**).

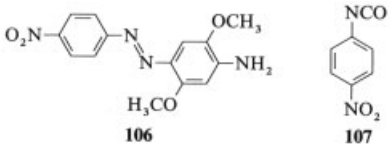

Phosgenation of 4-nitroaniline in aqueous medium gives *N,N′*-bis(4-nitrophenyl) urea [*587-90-6*], whereas phosgenation in toluene yields 4-nitrophenyl isocyanate [*100-28-7*] (**107**). The latter is used for the production of unsymmetrical ureas by reaction with amines. For example, with 3-picolylamine the urea [*53558-25-1*] (**108**) is formed, which has been used as a rodenticide [58]. An analogous process with guanidine hydrochloride leads to Nitroquanil [*51-58-1*] (**109**), an antimalarial drug which, although it is less active against malaria than its analogue Paludrine (derived from 4-chloroaniline), is also less toxic.

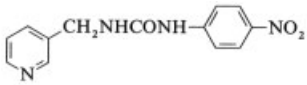

108

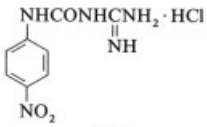

109

2,4-Dinitroaniline [*97-02-9*] (**110**), 2,4 DNA, $C_6H_5N_3O_4$, M_r 183.12, *mp* 188 °C, d_4^{15} 1.615, flash point 224 °C, forms yellow needles from aqueous acetic acid. The compound is sparingly soluble in ethanol and insoluble in water.

Production. 2,4-Dinitrochlorobenzene (Section 4.1) is aminated with aqueous ammonia initially at 70 °C; then the temperature is allowed to rise gradually to 120 °C to complete the reaction. Only a low-pressure autoclave is required, compared with the amination of 2- chloro- and 4- chloronitrobenzene.

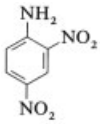

110

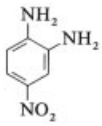

111

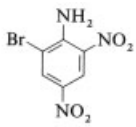

112

Uses. Reduction with iron – acid gives 1,2,4-triaminobenzene [*615-71-4*] and with sodium sulfide, 3,4-diaminonitrobenzene [*99-56-9*] (**111**), an intermediate for the preparation of nitroheterocyclics. Acetylation followed by reduction gives 2,4-diaminoacetanilide [*6373-15-5*]. Bromination of a sulfuric acid solution yields 6-bromo-2,4-dinitroaniline [*1817-73-8*] (**112**).

As a diazo component, 2,4-dinitroaniline typically gives violet monoazo disperse dyes when coupled with arylamines (e.g., C.I. Disperse Violet 12, **113**). The derived 6- chloro- and 6-bromo-2,4-dinitroanilines are much more important in the production of navy disperse dyes for polyester, (e.g., C.I. Disperse Blue 79, **114**) [59].

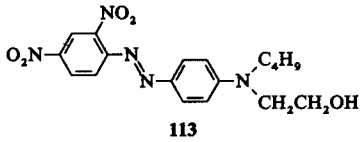

113

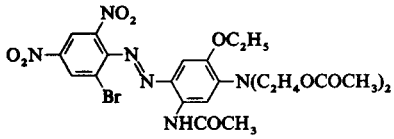

114

2,6-Dinitroaniline [*606-22-4*] (**117**), 2,6 DNA, $C_6H_5N_3O_4$, M_r 183.12, *mp* 142 °C, forms yellow needles from ethanol.

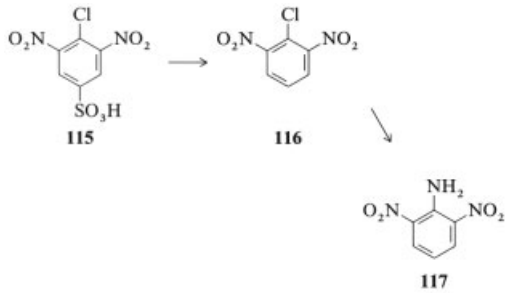

Production. Chlorobenzene-4-sulfonic acid is dinitrated to give 4- chloro-3,5-dinitrobenzenesulfonic acid [*88-91-5*] (**115**), which is desulfonated by heating the diluted sulfuric acid solution from the nitration mix to give 2- chloro-1,3-dinitrobenzene [*606-21-3*] (**116**); the latter is aminated to **117**.

Uses. 2,6-Dinitroaniline and its derivatives are reported to have algicidal and heribicidal activity; for example, heating with phthalic anhydride yields *N*-(2,6-dinitrophenyl)phthalimide [*57491-99-3*] (**118**) for use as a herbicide.

2,4,6-Trinitroaniline [*489-98-5*] (**119**), picramide, $C_6H_4N_4O_6$, M_r 228.12, crystallizes as yellow prisms from aqueous acetic acid; these melt at 192 – 195 °C and explode on further heating.

Production. Picric acid (Section 7.1) is converted to 2,4,6-trinitrochlorobenzene by treatment with PCl_5, or a similar chlorinating agent; this is then aminated under mild reaction conditions.

Uses. 2,4,6-Trinitroaniline is used in explosives [8] and as a diazo component for disperse dyes; in the latter, 2,4-dinitroaniline is nitrated to give picramide, and the resulting sulfuric acid solution is diazotized in situ [60]. Dyes based on picramide are reportedly improved by introduction of the nitro group by a final nucleophilic displacement reaction [$NaNO_2$ – $Cu(OAc)_2$ – DMF] [61] on the dye derived from 6-bromo-2,4-dinitroaniline (e.g., **114**).

Toxicity of Nitroamino Aromatics. All of the readily available nitroanilines are classified as highly toxic on inhalation or skin absorption. The TLV for 4-nitroaniline is 1 ppm [22].

5.2. Nitrotoluidines

The commercially available nitrotoluidine isomers, historically important as diazo bases, are obtained by low-temperature nitration of toluidines or conventional nitration of acylated toluidines. Certain isomers can be obtained by mono reduction of dinitrotoluenes, but mixtures arise unless the molecule is symmetrical (e.g., 2,6-dinitrotoluene). Nomenclature is a problem because the old usage, which still prevails in much of industry, numbers toluidines with 1-methyl, whereas many reference books adopt a convention of 1-amino.

4-Nitro-*o*-toluidine [*99-55-8*] (**120**), 2-methyl-5-nitroaniline, $C_7H_8N_2O_2$, M_r 152.15, *mp* 107 – 109 °C, d_4^{15} 1.365, crystallizes as yellow prisms from ethanol.

Production. *o*-Toluidine is nitrated at – 10 °C by slow addition of nitric acid to a solution of the amine in 98 % sulfuric acid. On quenching the nitration mixture in water and allowing the temperature to rise, a solution is obtained from which the sulfate of the product crystallizes on cooling, leaving most of the 6-nitro-*o*-toluidine byproduct in solution.

Uses. 4-Nitro-*o*-toluidine is used as a diazo component (Fast Red RL Base) in C.I. Pigment Red 12, for example.

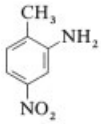

120

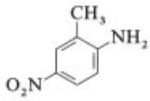

121

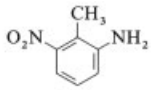

122

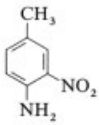

123

5-Nitro-*o*-toluidine [*99-52-5*] (**121**), 2-methyl-4-nitroaniline, $C_7H_8N_2O_2$, M_r 152.15, *mp* 132 – 133 °C, d_4^{15} 1.366, crystallizes from ethanol as yellow needles.

Production. *N*-Benzenesulfonyl-*o*-toluidine is dissolved in chlorobenzene at 40 – 50 °C, and 62 % nitric acid is added gradually to give 5-nitro-*N*-benzenesulfonyl-*o*-toluidine, which is isolated and hydrolyzed in sulfuric acid to give 5-nitro-*o*-toluidine in 80 % yield.

Uses. 5-Nitro-*o*-toluidine is used as a diazo component (Fast Scarlet G Base), for example, in C.I. Pigment Red 162 and C.I. Acid Black 28.

6-Nitro-*o*-toluidine [*603-83-8*] (**122**), 2-methyl-3-nitroaniline, $C_7H_8N_2O_2$, M_r 152.15, *mp* 92 °C, *bp* 305 °C, d_4^{15} 1.378, crystallizes from ethanol as yellow leaflets.

Production. 6-Nitro-*o*-toluidine is produced by sulfide reduction of 2,6-dinitrotoluene in a process similar to that for 3-nitroaniline (Section 5.1) or by controlled hydrogenation of 2,6-dinitrotoluene with palladium on a carbon catalyst [62].

3-Nitro-*p*-toluidine [*89-62-3*] (**123**), *m*-nitro-*p*-toluidine, MNPT, 4-methyl-2-nitroaniline, $C_7H_8N_2O_2$, M_r 152.15, *mp*

117 °C, *bp* 134 °C (0.26 kPa), d^{15} 1.31, forms orange leaflets from aqueous ethanol.

Production. *N*-Acetyl-*p*-toluidine is nitrated with mixed acid, and the quenched reaction mixture is heated and hydrolyzed to 3-nitro-*p*-toluidine.

Uses. 3-Nitro-*p*-toluidine is used as a diazo component (Red GL Base) in C.I. Pigment Yellow 1 and C.I. Pigment Red 3 (**124**), among others.

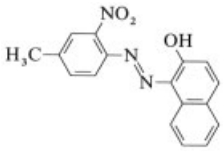

124

5.3. Halogenonitroanilines

Many halogenonitroaniline isomers exist [17] but only those obtained from the readily accessible dichloronitrobenzenes have large-scale use.

4-Chloro-2-nitroaniline [*89-63-4*] (**125**), $C_6H_5ClN_2O_2$, M_r 172.57, *mp* 116 – 117 °C, flash point 191 °C, forms orange needles from water, in which it is sparingly soluble (0.05 % at 20 °C, 0.65 % at 100 °C).

Production. 1,4-Dichloro-2-nitrobenzene (Section 4.2) is aminated at 165 °C by using a process similar to that described for 2-nitroaniline.

Uses. 4-Chloro-2-nitroaniline is used as a diazo component (Red 3GL Base) in C.I. Pigment Yellow 3 and C.I. Pigment Red 6, for example.

2-Chloro-4-nitroaniline [*121-87-9*] (**62**), $C_6H_5ClN_2O_2$, M_r 172.57, *mp* 108 °C, flash point 205 °C, crystallizes from water as yellow needles.

Production. 1,2-Dichloro-4-nitrobenzene (Section 4.2) is aminated by using a similar process to that described for 4-nitroaniline.

Uses. 2-Chloro-4-nitroaniline is used as a diazo component in disperse dyes (e.g., C.I. Disperse Red 65, **126**, example 3, Table 11) for polyester or in modified basic dyes (e.g., C.I. Basic Red 18, **127**) for acrylic fibers.

Bromination of (**62**) by treatment of a sulfuric acid solution with bromine or by treatment of a hydrobromic acid solution with hydrogen peroxide gives 2-bromo-6-chloro-4-nitroaniline [*99-29-6*] (**128**), which is used occasionally as an alternative to 2,6-dichloro-4-nitroaniline (**104**) in yellow-brown dyes for synthetic fibers.

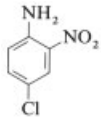

125

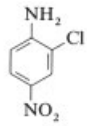

62

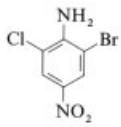

128

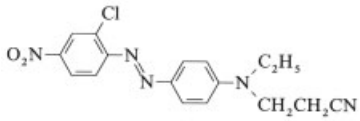

126

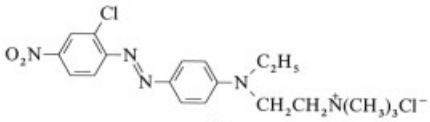

127

Acylation of **62** with the acid chloride derived from 5-chlorosalicylic acid gives *N*-(2-chloro-4-nitrophenyl)-5-chlorosalicylanilide (niclosamid) [*50-65-7*] (**129**), which is a molluscicide and anthelminthic and typifies a series of related nitrophenylsalicylanilides used for similar outlets [63].

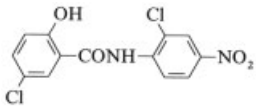

129

5.4. Cyanonitroanilines

The cyano group as a more bathochromic (Table 11) pseudohalogen substitutent in nitromonoazodisperse dyes has found increasing use in dyeing polyester because it also confers greater lightfastness.

2-Cyano-4-nitroaniline [*17420-30-3*] (**130**) is produced by nitration of 2-chlorobenzonitrile, separation of the required 2-chloro-5-nitrobenzonitrile [*16588-02-6*], and amination of this with ammonia.

Of the many monoazo disperse dyes derived from **130**, C.I. Disperse Red 73 (**134**, example 6 of Table 11) is typical [64]. Bromination of **130** gives 6-bromo-2-cyano-4-nitroaniline (**131**), which is used in blue dyes (e.g., C.I. Disperse Blue 183, **135**) [65]. 2-Cyano-4,6-dinitroaniline (**132**) and 2,6-dicyano-4-nitroaniline (**133**) are difficult to prepare and to diazotize. The derived bright blue dyes (e.g., **136** and **137**) are best prepared by introduction of the cyano group (cuprous cyanide in dimethylformamide) into the preformed monoazo dyes obtained from 2-bromo-4,6-dinitroaniline [66] and 6-bromo-2-cyano-4-nitroaniline [67], respectively.

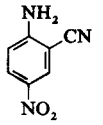
130

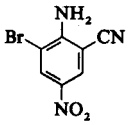
131

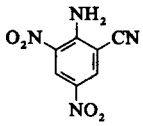
132

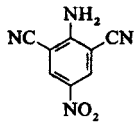
133

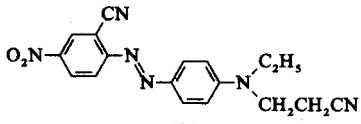
134

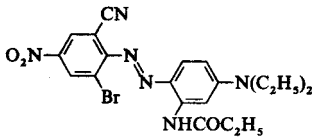
135

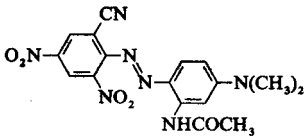
136

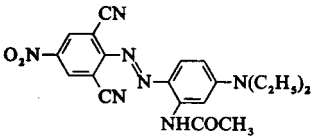
137

C. I. Disperse Blue 165

6. Nitroaromatic Sulfonic Acids and Derivatives

Most of the sulfonic acids derived from nitrobenzene, nitrotoluenes, chloronitrobenzenes, and related products are reduced to yield anilinesulfonic acids, for use in the synthesis of water-soluble dyes or pigments insolubilized as lakes. Some of the derived sulfonamides and sulfones are also important intermediates for colorants and fine chemicals; they are produced from the corresponding sulfonyl chlorides. The sulfonyl chloride group is more economically introduced by direct chlorosulfonation of the nitroaromatic rather than by conversion of the sulfonic acid.

6.1. Nitrobenzenesulfonic acids and Derivatives

2-Nitrobenzenesulfonic acid [*80-82-0*] (**138**), *o*-nitrobenzenesulfonic acid, $C_6H_5NO_5S$, M_r 203.18, *mp* (anhydride) 85 °C, crystallizes as very hygroscopic pale yellow crystals that form several hydrates.

Production. 2-Chloronitrobenzene is reacted with sodium disulfide at 80 °C in aqueous ethanol to give 2,2′-dinitrodiphenyl disulfide (**51**), which is isolated and treated with chlorine in mineral acid medium in the presence of nitric acid to produce 2-nitrobenzenesulfonyl chloride (**52**) by oxidative chlorination. This is hydrolyzed in dilute alkali, and the liquor of the resulting **138** is reduced with Fe – acid or H_2 – catalyst to orthanilic acid [*88-21-1*] (**139**), an important diazo component in reactive dyes for cotton.

2-Nitrobenzenesulfonyl chloride [*1694-92-4*] (**52**), $C_6H_4ClNO_4S$, M_r 221.62, crystallizes from ether, *mp* 68 – 69 °C.

Production. 2-Nitrobenzenesulfonyl chloride is isolated as an intermediate in the production of 2-nitrobenzenesulfonic acid. When 2,2′-dinitrodiphenyl disulfide is chlorinated in an organic solvent the product is 2-nitrophenylsulfenyl chloride [*7669-54-7*] (**140**).

Uses. Reduction with zinc – acid gives 2-aminothiophenol [*137-07-5*], although this is more readily produced from aniline. The main use of **52** is via derived sulfonamides; reac-

tion with ammonia produces 2-nitrobenzenesulfonamide [*5455-59-4*] (**141**), and methylamine gives the corresponding *N*-methylsulfonamide [*23530-40-7*], both of which have been patented for use as herbicides. Reaction with *N*-ethylaniline produces a nitrosulfonamide that, on reduction, gives 2-amino-*N*-ethylbenzenesulfonanilide [*81-10-7*] (**142**), a valuable diazo component for acid wool dyes.

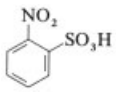

138

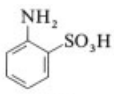

139

140

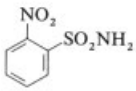

141

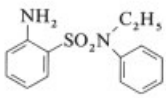

142

Many other sulfonamides derived from aminoheterocyclics or *o*-aminobenzoic esters can be ring- closed after reduction to give compounds with a variety of reported pharmacological properties.

3-Nitrobenzenesulfonic acid [*98-47-5*] (**8**), *m*-nitrobenzenesulfonic acid, $C_6H_5NO_5S$, M_r 203.18, crystallizes in yellow plates that are slightly deliquescent in air. Both the free acid and the sodium salt are potentially unstable to heat.

Production [1]. 3-Nitrobenzenesulfonic acid is produced by dissolving nitrobenzene in 98 % sulfuric acid and heating to 80 °C. Then 65 % oleum is added at this temperature, and the reaction is completed by heating for 9 h at 105 °C (higher temperatures are unsafe). After quenching the sulfonation mass in water, neutralizing with lime, and filtering off gypsum, the calcium salt of the product is converted to the sodium salt with sodium carbonate. After calcium carbonate has been removed by filtration, the sodium salt solution can be used directly or evaporated.

The formation of 3,3′-dinitrodiphenylsulfone [*1228-53-1*] as a byproduct is claimed to be reduced from 8 to 4 % by addition of sulfur trioxide [68].

Although benzenesulfonic acid can be nitrated to give mainly 3-nitrobenzenesulfonic acid (54 %), the process is not economical.

Uses. 3-Nitrobenzenesulfonic acid is used primarily as a mild oxidizing agent or as a precursor for colorant intermediates. In the former case it is used in processing certain anthraquinone intermediates (e.g., amination of anthraquinone-1-sulfonic acid to give 1-aminoanthraquinone) or as a dye-printing auxiliary to obtain resist effects with, for example, vat dyes. Reduction of 3-nitrobenzenesulfonic acid (usually as a liquor) to the important metanilic acid is achieved with iron or by catalytic hydrogenation. The alternative reduction with zinc and alkali or with sodium amalgam yields hydrazobenzene-3,3′-disulfonic acid, which rearranges to benzidine-2,2′-disulfonic acid [*117-61-3*] on treatment with acid.

3-Nitrobenzenesulfonyl chloride [*121-51-7*] (**9**), $C_6H_4ClNO_4S$, M_r 221.62, *mp* 64 °C, is crystallized from petroleum ether.

Production. Nitrobenzene is added to excess chlorosulfonic acid, and the temperature is raised gradually to 100 °C and held there for 6 h. After cooling, the mixture is poured into ice and water, and 3-nitrobenzenesulfonyl chloride is filtered off in 75 % yield for use as a paste, after being washed (to remove acid) with cold water.

Uses. Halide exchange gives 3-nitrobenzenesulfonyl fluoride [*349-78-0*]. Reduction with aqueous sodium sulfite yields 3-nitrobenzenesulfinic acid [*13257-95-9*] (**143**), which is readily alkylated to form sulfones. When alkylation is carried out with ethylene chlorohydrin or ethylene oxide, the product is 3-nitrophenyl-2-hydroxyethylsulfone [*41687-30-3*], which is reduced to 3-aminophenyl-2-hydroxyethylsulfone [*5246-57-1*] (**144**), a valuable intermediate in the production of dyes (e.g., C.I. Reactive Blue 19).

3-Nitrobenzenesulfonyl chloride reacts with amines forming 3-nitrobenzenesulfonamide [*121-52-8*] with ammonia, and 3-nitrobenzenesulfonanilide [*80-37-5*] (**145**) with aniline. Reduction of these nitrosulfonamides yields the corresponding amines for use as diazo components.

4-Nitrobenzenesulfonic acid [*138-42-1*] (**146**), *p*-nitrobenzenesulfonic acid, $C_6H_5NO_5S$, M_r 203.18, forms hygroscopic crystals from ethyl acetate – benzene, *mp* (anhydrous) 106 – 108 °C.

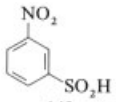

143

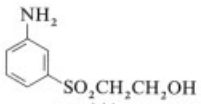

144

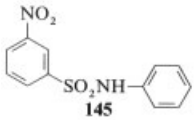

145

Production. 4-Nitrobenzenesulfonic acid is produced by hydrolysis of 4-nitrobenzenesulfonyl chloride.

Uses. 4-Nitrobenzenesulfonic acid is less important than the ortho and meta isomers because sulfanilic acid and its derivatives are readily available from aniline and acetanilide.

4-Nitrobenzenesulfonyl chloride [*98-74-8*] (**147**), $C_6H_4ClNO_4S$, M_r 221.62, *mp* 80.5 °C, crystallizes from petroleum ether as colorless needles.

Production. 4,4′-Dinitrodiphenyldisulfide [*100-32-3*] (**148**), produced by reaction of 4-chloronitrobenzene with sodium disulfide, is treated with chlorine in $HCl-HNO_3$ by a process similar to that used for 2-nitrobenzenesulfonyl chloride.

Uses. Nitrosulfonamides derived from **147** have been used as herbicides. Sulfa drugs are rarely prepared from **147** because the route via *N*-acetylsulfanilyl chloride is much more economical. However, when *N*-acylsulfonamides are required, the nitro route must be used, e.g., reduction of **149** [derived from reaction of **147** with a heterocyclic amine (RNH_2), followed by acylation with acetic anhydride] under mild conditions to give the required product.

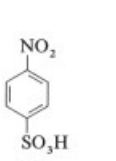

146

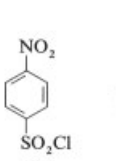

147

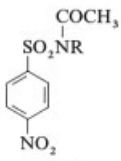

149

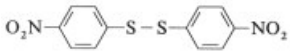

148

2,4-Dinitrobenzenesulfonic acid [*89-02-1*] (**150**), $C_6H_4N_2O_7S$, M_r 248.17, crystallizes from water as colorless needles of the trihydrate (*mp* 108 °C), which dehydrate above 130 °C.

Production. 2,4-Dinitrochlorobenzene is warmed with aqueous sodium sulfite to give an 80 % yield of the sodium salt of **150** [*885-62-1*] as yellow needles after salting out. The lower reactivity of 2- and 4- chloronitrobenzene requires much higher temperature for this nucleophilic sulfitation, with too many side reactions to be practicable.

Uses. Reduction of **150** gives the colorant intermediate 1,3-diaminobenzene-4-sulfonic acid [*88-63-1*] as an alternative to the sulfonation of

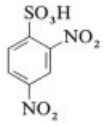

150

m-phenylenediamine. The sulfonic acid group of **150** can be replaced by amino and hydroxyl groups in the same way as the chlorine atom of 2,4-dinitrochlorobenzene, and sometimes the use of a water-soluble alternative may be advantageous. Treatment of **150** with chlorosulfonic acid gives 2,4-dinitrobenzenesulfonyl chloride [*1656-44-6*] which is used as a photographic chemical together with the parent sulfonic acid.

6.2. Nitrotoluenesulfonic Acids and Derivatives

Nitrotoluenesulfonic acids have properties similar to 3-nitrobenzenesulfonic acids and are reduced to toluidinesulfonic acids, which sometimes offer advantages over anilinesulfonic acids as colorant intermediates. The methyl group introduces an additional functionality, leading to alternative derivatives, mainly via oxidation reactions.

2-Nitrotoluene-4-sulfonic acid [*97-06-3*] (**18**), 3-nitro-4-methylbenzenesulfonic acid $C_7H_7NO_5S$, M_r 217.20, crystallizes from water as pale yellow hygroscopic needles of the dihydrate (*mp* 92 °C, decomp. < 245 °C).

Production. 2-Nitrotoluene is sulfonated by heating with 25 % oleum at 80 °C for 3 h, then quenching and liming to give a neutral solution of the sodium salt of **18** [1].

Uses. The total liquor of **18** is hydrogenated in situ to give *o*-toluidine-4-sulfonic acid [*618-03-1*] (**151**). Isolated 2-nitrotoluene-4-sulfonic acid may be converted to 2-nitro-*p*-toluenesulfonyl chloride [*54090-41-4*] (**152**), but the preferred process for **152** is chlorosulfonation of 2-nitrotoluene. Reaction of **152** with dimethylamine followed by reduction gives 3-amino-4-methyl-*N,N*-dimethylbenzenesulfonamide [*6331-68-6*] for use as an azoic diazo component.

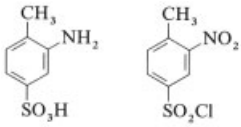

151 **152**

Heating a solution of 2-nitrotoluene-4-sulfonic acid with sodium hydroxide produces the diazo component 4-sulfoanthranilic acid [*98-43-1*], and the autoxidation can be better controlled than with 2-nitrotoluene (cf. Section 3.2) to give a viable manufacturing process.

4-Nitrotoluene-2-sulfonic acid [*121-03-9*] (**25**), 5-nitro-2-methylbenzenesulfonic acid, *p*-nitrotoluene-*o*-sulfonic acid, PNTOS, $C_7H_7NO_5S$, M_r 217.20, crystallizes from water as plates of the dihydrate (*mp* 133.5).

Production. 4-Nitrotoluene is sulfonated with 25 % oleum at 60 °C and worked up to give a solution of the sodium salt of **25**.

Uses. The most important outlets for 4-nitro-toluene-2-sulfonic acid exploit the readiness with which it forms stilbene derivatives. Selfcondensation products obtained by heating in aqueous sodium hydroxide are used as yellow water-soluble cotton dyes and consist of mixtures containing 4,4′-dinitroso-2,2′-stilbenedisulfonic acid [*58058-72-3*] (**153**), 4,4′-dinitro-2,2′-stilbenedisulfonic acid [*128-42-7*] (**154**), and 4,4′-dinitrodibenzyl-2,2′-disulfonic acid [*6268-17-3*] (**155**).

The entire class of stilbene dyes (→ Azo Dyes, Chap. 4.1.2.1.) is based on this chemistry via further reaction of the condensation products, especially **154**, with aromatic amines or, more usually, with aminoazo compounds to give cheap, water-soluble orange and brown dyes.

Manufacture of pure **154** is achieved by controlled air oxidation of **25** in the presence of a catalyst. I.G. Farbenindustrie [1] used manganese sulfate and a specially designed unit to optimize air distribution and agitation. A simpler alternative is the use of sodium hypochlorite as oxidant in a process involving the simultaneous addition of oxidant and sodium hydroxide to an aqueous solution of **25** at 75 – 80 °C [69].

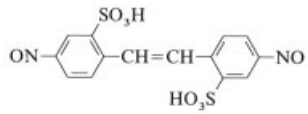

153

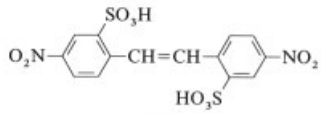

154

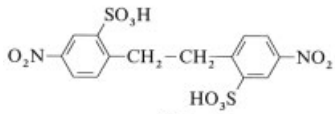

155

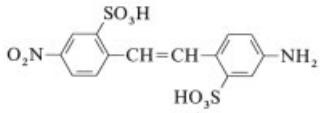

156

Apart from its use in the production of stilbene dyes, **154** is even more important as the precursor, by Fe – HCl reduction at 100 °C, of 4,4′-diamino-2,2′-stilbenedisulfonic acid [*81-11-8*], which is the basis for 80 % of all fluorescent whitening agents (→ Optical Brighteners).

Reduction of **154** with sodium sulfide results in production of the disodium salt of 4′-nitro-4-aminostilbene-2,2′-disulfonic acid [*6634-82-8*] (**156**) for use as a diazo component; for example, in polyazo dyes for which the nitro group of **156** is reduced to an azoxy group by treatment with glucose and sodium hydroxide [70].

Hydrogenation of **25** gives *p*-toluidine-2-sulfonic acid (5-amino-2-methylbenzenesulfonic acid) [*118-88-7*], which can be used as an alternative to metanilic acid. 4-Nitrotoluene-2-sulfonyl chloride [*121-02-8*] (**157**), the acid chloride of **25**, is best obtained by chlorosulfonation of 4-nitrotoluene. Reaction of **157** with aniline followed by reduction gives 4-aminotoluene-2-sulfonanilide [*79-72-1*] (**158**) for

use as a diazo component. The sulfonamide derived from **157** is oxidized and ring closed in one step to give 6-nitrosaccharin (**159**) by treatment with $CrO_3 - H_2SO_4$ at 65 – 70 °C [71].

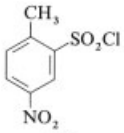

157

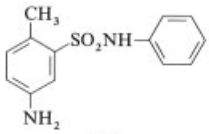

158

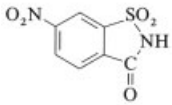

159

2,6-Dinitrotoluene-4-sulfonic acid [*88-90-4*] (**160**), 3,5-dinitro-4-methylbenzenesulfonic acid, $C_7H_6N_2O_7S$, M_r 262.20, crystallizes as a hydrate in the form of pale yellow crystals, which soften at 110 °C, dehydrate at 140 °C, and melt at 165 °C.

Production. 2-Nitrotoluene is sulfonated to form 2-nitrotoluene-4-sulfonic acid; the reaction mixture is diluted with 90 % sulfuric acid, and sodium nitrate is added over 6 h at 80 °C. After a further 1 h at this temperature the reaction mass is poured into water and the sodium salt of the product is salted out by addition of sodium sulfate. After being washed acid-free with brine, the yield of **160** is ca. 80 %.

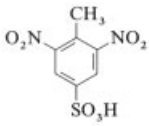

160

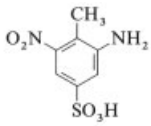

161

Uses. Reduction of **160** with sulfide gives 2-amino-6-nitrotoluene-4-sulfonic acid (**161**), and reduction with iron – acid gives 2,6-diaminotoluene-4-sulfonic acid [*98-25-9*]. 2,6-Dinitrotoluene-4-sulfonyl chloride [*80198-19-2*] is obtained from **160** by treatment with thionyl chloride. Reaction of this sulfonyl chloride with an alkylamine followed by hydrogenation leads to a readily accessible series of symmetrical diaminotoluenesulfonamides [72].

6.3. Chloronitrobenzenesulfonic Acids and Derivatives

The commercially available chloronitrobenzenesulfonic acids are usually obtained by sulfonation or chlorosulfonation of chloronitrobenzenes or by sulfition of dichloronitrobenzenes. Although the corresponding substituted anilines can be obtained by reduction, the main outlets for this group of products are via nucleophilic displacement reactions of the chlorine substituent.

2-Chloro-5-nitrobenzenesulfonic acid [*96-73-1*] (**58**), 4-chloronitrobenzene-3-sulfonic acid, PN salt, $C_6H_4ClNO_5S$, M_r 237.62, crystallizes as the dihydrate from water (*mp* 169 °C, decomp.). The compound undergoes explosive decomposition when heated above 200 °C, and all reactions, including its production, must be undertaken with great caution.

Production. 4-Chloronitrobenzene is sulfonated with 65 % oleum at 115 °C; special precaution must be taken to control temperature [41]. Work-up is by quenching and isolation of the sodium salt after it has been salted out.

Uses. Amination of **58** gives 4-nitroaniline-2-sulfonic acid [*96-75-3*] (**162**), which is used as a diazo component or reduced to the important colorant intermediate 2,5-diaminobenzenesulfonic acid [*88-45-9*]. Chlorination of **162** in aqueous solution is a more facile route to 2,6-dichloro-4-nitroaniline (**104**) than chlorination of 4-nitroaniline.

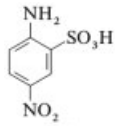

162

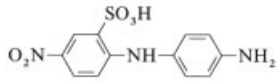

163

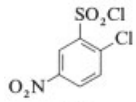

164

Monocondensation of **58** with *p*-phenylenediamine gives 4′-amino-4-nitrodiphenylamine-2-sulfonic acid [*91-29-2*] (**163**), which can be reduced to 4,4′-diaminodiphenylamine-2-sulfonic acid [*119-70-0*], a tetrazo component in direct

cotton dyes. Monoaminodiphenylaminesulfonic acid diazo components are obtained by condensation of **58** with aniline or toluidine, followed by reduction. 2-Chloro-5-nitrobenzenesulfonyl chloride [*4533-95-3*] (**164**) may be obtained from **58** but is best prepared by direct chlorosulfonation of 4-chloronitrobenzene. Amination of **164** gives 4-nitroaniline-2-sulfonamide [*54734-85-9*].

4-Chloro-2-nitrobenzenesulfonic acid (**165**), $C_6H_4ClNO_5S$, M_r 237.62, *mp* 114–115 °C (anhydrous), forms hygroscopic crystals. The sodium salt is water soluble; the potassium salt, less so (ca. 1 % at 25 °C).

Production. 4-Chloro-2-nitrobenzenesulfonic acid is obtained by alkaline hydrolysis of 4-chloro-2-nitrobenzenesulfonyl chloride [*4533-96-4*] (**166**), which is obtained from bis(4-chloro-2-nitrophenyl) disulfide (the latter is formed from 2-nitro-1,4-dichlorobenzene) by analogy with the process for 2-nitrobenzenesulfonyl chloride. The primary outlets for **166** are amination to 4-chloro-2-nitrobenzenesulfonamide [*13852-81-8*] (**167**) and further chlorosulfonation to the disulfonyl dichloride (**168**), an intermediate for a series of 4-amino-6-chlorobenzene-1,3-disulfonamides that are useful diuretics.

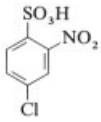

165

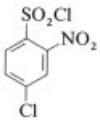

166

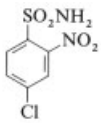

167

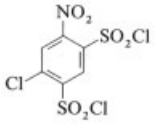

168

4-Chloro-3-nitrobenzenesulfonic acid [*121-18-6*] (**49**), $C_6H_4ClNO_5S$, M_r 237.62, crystallizes as needles from water (decomp. > 200 °C). The potassium salt forms pale yellow crystals in water, *mp* 325 °C (decomp.). The solubility of the potassium salt in water is 2 % at 30 °C versus 19 % for the sodium salt.

Production. 2-Chloronitrobenzene can be sulfonated with 65 % oleum analogously to 4-chloronitrobenzene, but the preferred process is to sulfonate chlorobenzene at 100 °C and nitrate the sulfonation mass in situ by the addition of nitric acid or potassium nitrate at 20–30 °C. The product is isolated in 90 % yield as its sodium or potassium salt by salting out from the quenched reaction mixture after addition of lime and removal of the precipitated calcium sulfate. Purification, if required, is achieved by recrystallization of the potassium salt.

Uses. Hydrolysis of **49** gives 2-nitrophenol-4-sulfonic acid [*616-85-3*] (**169**) for reduction to the diazo component 2-aminophenol-4-sulfonic acid [*98-37-3*]. Amination yields 2-nitroaniline-4-sulfonic acid [*616-84-2*] (**170**), and iron reduction produces 2-chloroaniline-5-sulfonic acid [*98-36-2*], both of which are used as diazo components. Heating **49** with aqueous sodium sulfite results in nitrobenzene-2,5-disulfonic acid [*119-00-6*] (**171**), an intermediate for an alternative route to aniline-2,5-disulfonic acid [*98-44-2*].

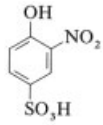

169

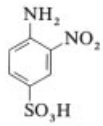

170

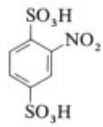

171

The acid chloride of **49** is best produced by direct chlorosulfonation of 2-chloronitrobenzene. 4-Chloro-3-nitrobenzenesulfonyl chloride [*97-08-5*] (**50**) is especially important as the intermediate for 4-chloro-3-nitrobenzenesulfonamide [*97-09-6*] (**172**); the latter leads to the diazo component 2-aminophenol-4-sulfonamide [*98-32-8*]. 4-Chloro-3-nitrobenzenesulfonyl chloride is also used for the production of nitro dyes, e.g., Disperse Yellow 42 (**173**) by condensation of **50** with two equivalents of aniline (→ Nitro and Nitroso Dyes).

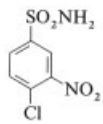

172

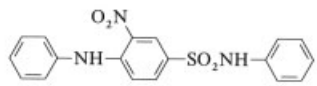

173

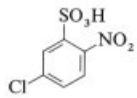

174

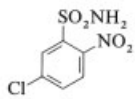

175

5-Chloro-2-nitrobenzenesulfonic acid [*54481-12-8*] (**174**), $C_6H_4ClNO_5S$, M_r 237.62, forms hygroscopic crystals of the monohydrate (*mp* 93 °C).

Production. 5,5′-Dichloro-2,2′-dinitrodiphenyldisulfide (obtained from the reaction of 2,4-dichloronitrobenzene with Na_2S_2) is oxidatively chlorinated with chlorine in $HCl-HNO_3$ to give the sulfonyl chloride, which hydrolyzes to **174**.

Uses. Reduction of **174** gives 4-chloroaniline-2-sulfonic acid, and amination results in 4-nitroaniline-3-sulfonic acid, both of minor importance as intermediates. The derived sulfonyl chloride leads to 5-chloro-2-nitrobenzenesulfonamide (**175**), an intermediate in the production of the benzothiadiazine derivative diazoxide, used as a vasodilator.

2-Chloro-3,5-dinitrobenzenesulfonic acid [*4515-26-8*] (**176**), $C_6H_3ClN_2O_7S$, M_r 282.62, crystallizes from water as needles.

Production. 4-Chloronitrobenzene is sulfonated by gradual addition of 25 % oleum at 60 °C; the temperature is raised slowly to 125 °C, and held there for 8 h. The sulfonation mixture is cooled to 70 °C, nitric acid is added over 5 h, and the reaction is completed by heating for 12 h at 70 °C followed by 4 h at 100 °C. After cooling and quenching, the sodium salt of the product is salted out by addition of sodium sulfate and isolated in 80 % yield.

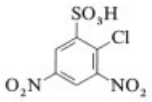

176

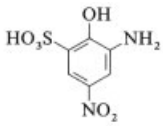

177

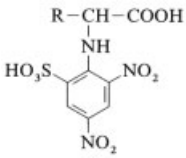

178

Uses. Hydrolysis of **176** with aqueous alkali readily yields 2,4-dinitrophenol-6-sulfonic acid, which is reduced with sodium sulfide to 4-nitro-2-aminophenol-6-sulfonic acid [*96-67-3*] (**177**), a diazo component of metallizable azo dyes. Under very mild conditions, **176** is aminated to 2,4-dinitroaniline-6-sulfonic acid, and **176** reacts with sodium sulfite to give 3,5-dinitrobenzene-1,2-disulfonic acid.

As a reagent for the identification of α-amino acids ($RCHNH_2COOH$), **176** forms derivatives (**178**) that are more readily hydrolyzed in acid solution to give the parent α-amino acid than the corresponding 2,4-dinitrophenyl derivatives [73].

4-Chloro-3,5-dinitrobenzenesulfonic acid [*88-91-5*] (**115**), $C_6H_3ClN_2O_7S$, M_r 282.62, forms hygroscopic crystals in water (*mp* 293 °C).

Production. 4-Chloro-3,5-dinitrobenzenesulfonic acid is produced by a process based on the sulfonation and nitration of chlorobenzene that uses 2 mol/mol of potassium nitrate, in place of the 1 mol/mol employed in the production of 4-chloro-3-nitrobenzenesulfonic acid. The product is isolated from the quenched reaction mixture as the potassium salt, after addition of potassium chloride.

Uses. Conversion of **115** to 2,6-dinitroaniline is described in Section 5.1. More important is the use of the derived sulfonyl chloride to form the sulfonamide (**179**) (a fungicide) and the sulfone (**180**). Reaction of **179** with di-*n*-propylamine forms oryzalin [*19044-88-3*], and reaction of **180** with di-*n*-propylamine forms nitralin [*4726-14-1*], both of which are important members of the large class of substituted 2,6-dinitroaniline herbicides.

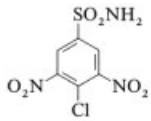

179

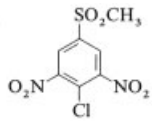

180

6.4. Chloronitrotoluenesulfonic Acids

6-Chloro-2-nitrotoluene-4-sulfonic acid [*68189-28-6*] (**181**), 3-chloro-4-methyl-5-nitrobenzenesulfonic acid, $C_7H_6ClNO_5S$, M_r 251.65.

Production. 6-Chloro-2-nitrotoluene is sulfonated with 20 % oleum at 60 °C with careful temperature control to avoid a runaway reaction. Work-up is by conventional processes.

Uses. Reduction of **181** gives 2-amino-6-chlorotoluene-4-sulfonic acid [*6387-27-5*], which is used as a diazo component.

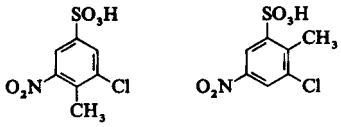

181 **182**

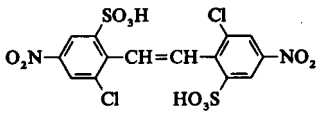

183

6-Chloro-4-nitrotoluene-2-sulfonic acid (**182**), 3-chloro-2-methyl-5-nitrobenzenesulfonic acid, $C_7H_6ClNO_5S$, M_r 251.65, crystallizes from water as hygroscopic needles.

Production. 2-Chloro-4-nitrotoluene is sulfonated with 30 % oleum at 75 °C to produce 6-chloro-4-nitrotoluene-2-sulfonic acid.

Uses. Oxidation of **182** with sodium hypochlorite gives 6,6′-dichloro-4,4′-dinitro-2,2′-stilbene disulfonic acid (**183**), which is reduced to the corresponding diamine; the latter is used in experimental fluorescent whitening agents, analogous to those obtained from 4,4′-diamino-2,2′-stilbene disulfonic acid.

Toxicology of Nitrohydrocarbon Sulfonic Acids. As a general rule the water solubility conferred by the sulfonic acid group, as well as the hydrophilicity conferred by the sulfonamide and sulfone groups, renders this group of products less toxic than their unsulfonated parent compounds. The products are less volatile, penetrate the skin less readily, and are more readily excreted if ingested. This is demonstrated by the sodium salt of 3-nitrobenzenesulfonic acid, which has an oral LD_{50} (rat) of 11 000 mg/kg, compared to 640 mg/kg for nitrobenzene.

7. Nitrohydroxy and Alkoxy Aromatics

Phenol can be nitrated with aqueous nitric acid in the presence of nitrous acid (cf. Section 2.2) to give a mixture consisting mainly of 2- and 4-nitrophenol. Although these can be separated (the ortho isomer is steam volatile), the economics of this route compares unfavorably with hydrolysis of 2- and 4-nitrochlorobenzene. Hydrolysis is also used for 2,4-dinitrophenol. Nitrocresols are more frequently obtained by nitrosation or nitration, although mixtures of isomers often result.

Nitrophenyl ethers are also produced from reactive nitrochloro compounds by reaction with alcohols or phenols. The less common *meta*-nitrophenyl ethers are prepared by methylation of the 3-nitrophenol precursor. Nomenclature with substituted anisidines and their homologues is confusing. Common practice (used here) maintains the methoxy group at the 1-position, in line with substituted phenols, but many references use the convention of the amino group in the 1-position. Only the formula or the CAS name is unambiguous for compounds such as 4-nitro-*o*-anisidine.

The most general use of both nitrophenols and nitro ethers involves reduction to the corresponding hydroxy- and alkoxyanilines.

7.1. Nitrophenols and Derived Ethers

2-Nitrophenol [*88-75-5*] (**46**), *o*-nitrophenol, $C_6H_5NO_3$, M_r 139.11, *mp* 45 °C, *bp* 216 °C (101 kPa), forms yellow needles from ethanol. It is slightly soluble in water (1 % at 100 °C) and steam volatile (intramolecular hydrogen bonding, unlike the para isomer).

Production. 2-Chloronitrobenzene in 8.5 % sodium hydroxide solution is heated gradually (exothermic) to 170 °C in an autoclave and held there under pressure for 8 h. The resulting solution is cooled and acidified to give the product in 95 % yield.

Uses. Catalytic hydrogenation (Pd–C) of 2-nitrophenol gives 2-aminophenol [*95-55-6*], which is used as a photographic developer and, in larger amounts, as a versatile intermediate for dyes and fine chemicals. 2-Aminophenol reacts with acetic anhydride to give 2-methylbenzoxazole, which is nitrated with mixed acid to give the 6-nitro derivative for subsequent hydrolysis to 5-nitro-2-aminophenol [*121-880-0*] (**184**).

2-Nitroanisole [*91-23-6*] (**47**), 2-methoxynitrobenzene, 2-nitrophenyl methyl ether, $C_7H_7NO_3$, M_r 153.13, *mp* 10 °C, *bp* 272 °C (101 kPa), is steam volatile. The compound is sparingly soluble in water, but readily soluble in ethanol, ether, or benzene.

Production. Methanolic sodium hydroxide is added slowly to a solution of 2- chloronitrobenzene in methanol at 70 °C; the mixture is then heated gradually to 95 °C under pressure to complete the reaction. After dilution with water the product is separated as an oil, in 90 % yield, and methanol is recovered from the aqueous layer.

Uses. 2-Nitroanisole is reduced (H_2– catalyst or iron – formic acid) to *o*-anisidine [*90-04-0*] or (benzidine-type reaction) to *o*-dianisidine [*119-90-4*], both of which are important as dye intermediates. Low-temperature nitration of *o*-anisidine, either directly or by addition of *o*-anisidine nitrate to sulfuric acid, gives 4-nitro-*o*-anisidine (Fast Scarlet base) [*99-59-2*] (**185**) in good yield. Nitration of *N*-acetyl- *o*-anisidine with mixed acid at 25 °C gives mainly the 5-nitro derivative, which can be separated from the 4-nitro isomer after hydrolysis to the amine because the more weakly basic 5-nitro-*o*-anisidine (Fast Red B base) [*97-52-9*] (**186**) is precipitated first on dilution of the sulfuric acid hydrolysate. Chlorosulfonation of **47** yields 2-nitroanisole-4-sulfonyl chloride [*22117-79-9*], from which *o*-anisidine-4-sulfonamide derivatives are obtained.

2-Ethoxynitrobenzene [*610-67-3*] (**187**), 2-nitrophenyl ethyl ether, 2-nitrophenetole,$C_8H_9NO_3$, M_r 167.16, *cp* 4.5 °C, *bp* 275 °C (101 kPa), is a greenish yellow oil.

Production. 2-Chloronitrobenzene in ethanol solution is treated with ethanolic sodium or potassium hydroxide in a process similar to that for 2-nitroanisole, except that the reaction mixture is aerated to avoid reduction to the potentially explosive azoxy derivative. The nitrophenol byproduct is formed in larger amounts than with nitroanisole and requires washing with alkali or post-ethylation with ethyl chloride under pressure [1].

Uses. Hydrogenation of **187** gives *o*-phenetidine [*94-70-2*] (2-ethoxyaniline), which is nitrated to 4-nitro-*o*-phenetidine.

4-Chloro-2-nitrophenol [*89-64-5*] (**188**), $C_6H_4ClNO_3$, M_r 173.56, *mp* 87 °C, forms yellow prisms from ethanol. The compound is sparingly soluble in water and is steam volatile.

Production. 2,5-Dichloronitrobenzene is hydrolyzed with 8 % aqueous sodium hydroxide at 120 – 130 °C under pressure [1].

Uses. Reduction of **188** gives 2-amino-4-chlorophenol [*95-85-2*], and chlorination yields 2,4-dichloro-6-nitrophenol [*609-89-2*] (**189**), which is reduced to 2-amino-4,6-dichlorophenol [*527-62-8*]. These substituted aminophenols are members of a large class of diazo components used for hydroxyazo dyes, which are important for wool and polyamide when converted to their chromium complexes. For the amine derived from the methyl ether of **188**, 4- chloro-*o*-anisidine [*93-50-5*], the preferred route is methoxylation of 2,5-dichloronitrobenzene followed by reduction. 2-amino-4-chlorophenol reacts with phosgene to give 5-chlorobenzoxolinone which nitrates to the 6-nitro compound (**190**); the latter undergoes alkaline hydrolysis to 2-amino-4-chloro-5-nitrophenol.

3-Nitrophenol [*554-84-7*] (**191**), *m*-nitrophenol, $C_6H_5NO_3$, M_r 139.11, *mp* 97 °C, *bp* 194 °C, forms yellowish crystals from dilute hydrochloric acid. The compound is moderately soluble in water (1 % at 20 °C, 13 % at 90 °C).

Production. 3-Nitroaniline is diazotized in aqueous sulfuric acid and then hydrolyzed by being added gradually to boiling dilute sulfuric acid. The crude product solidifies on cooling and is filtered off in 90 % yield.

Uses. Reduction of **191** gives 3-aminophenol [*591-27-5*], and methylation to 3-nitroanisole [*555-03-3*] followed by reduction gives *m*-anisidine [*536-90-3*]. Carboxylation of 3-aminophenol gives the tuberculostatic 4-aminosalicylic acid [*65-49-6*].

4-Nitrophenol [*100-02-7*] (**55**), *p*-nitrophenol, $C_6H_5NO_3$, M_r 139.11, forms yellow needles (mixed α- and β-forms), *mp* 114 °C, *bp* 216 °C (101 kPa). The (metastable) α-form crystallizes from toluene above 63 °C, and the yellow, prismatic β-form crystallizes from toluene below 63 °C. 4-Nitrophenol is not steam volatile and is much more soluble in water (30 % at 100 °C) than the ortho isomer. Its flash point is

105 °C, and the decomposition point (explosive) of the sodium salt is 156 °C.

Production. 4-Chloronitrobenzene is hydrolyzed by a process identical to that described for 2-nitrophenol.

Uses. Reduction of **55** to 4-aminophenol [*123-30-8*] has largely been superseded by processes based on direct reduction of nitrobenzene. Reaction of sodium 4-nitrophenoxide with dialkylthiophosphoric chlorides (125 °C in chlorobenzene) gives a series of insecticides typified by parathion [*56-38-2*] (**192**; R = ethyl), which has largely been displaced by the less toxic methylparathion [*298-00-0*] (**192**; R = methyl) and related products. 4-Nitrophenol reacts with 4- chloro-3-nitrobenzotrifluoride (**96**) to form fluorodifen (**193**)–related to nitrofen (**67**), mentioned in Section 4.1; it is used as a herbicide specifically on drilled rice crops.

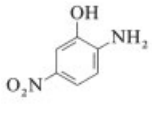
184

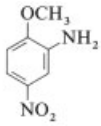
185

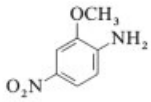
186

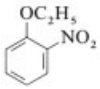
187

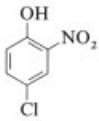
188

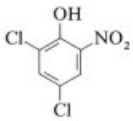
189

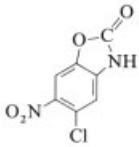
190

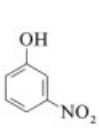
191

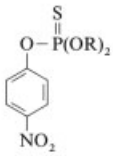
192

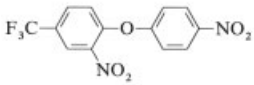
193

4-Nitroanisole [*100-17-4*] (**56**), 4-methoxynitrobenzene, 4-nitrophenyl methyl ether, $C_7H_7NO_3$, M_r 153.13, *mp* 54 °C, *bp* 274 °C (101 kPa), forms prisms from ethanol. 4-Nitroanisole is insoluble in water, moderately soluble in ethanol or benzene and steam volatile.

Production. 4-Chloronitrobenzene is dissolved in methanol and treated with methanolic sodium hydroxide, which is added slowly at 80 – 90 °C. After completion of reaction and neutralization, the excess methanol is removed by distillation, and after cooling, the product is filtered off in 95 % yield.

Uses. Reduction of **56** with sulfide gives *p*-anisidine [*104-94-9*]. *N*-Acetyl- *p*-anisidine is nitrated with 62 % nitric acid in chlorobenzene solution to give the 3-nitro derivative for hydrolysis to 3-nitro- *p*-anisidine (Fast Bordeaux GP base) [*96-96-8*] (**194**).

4-Ethoxynitrobenzene [*100-29-8*] (**195**), 4-nitrophenyl ethyl ether, 4-nitrophenetole, $C_8H_9NO_3$, M_r 167.16, *mp* 60 °C, *bp* 283 °C (101 kPa), crystallizes from ethanol as prisms, which are insoluble in water and moderately soluble in ethanol or ether.

Production. 4-Ethoxynitrobenzene is obtained from 4- chloronitrobenzene, by a process analogous to that described for 2-ethoxynitrobenzene.

Uses. Reduction of **195** gives *p*-phenetidine [*156-43-4*], which undergoes similar reactions to *p*-anisidine and is also the intermediate for the analgesic phenacetin (4-ethoxyacetanilide [*62-44-2*]) and the sweetener dulcin (4-ethoxyphenylurea [*150-69-6*]).

2,4-Dinitrophenol [*51-28-5*] (**63**), $C_6H_4N_2O_5$, M_r 184.11, *mp* 114 °C (decomposes explosively at higher temperature), pK_a 4.89, forms yellow needles. 2,4-Dinitrophenol is sublimable, steam volatile, sparingly soluble in water (1.3 % at 100 °C), and readily soluble in acetone, ethanol, or benzene.

Production. 2,4-Dinitrochlorobenzene is hydrolyzed by heating with 6 % aqueous sodium hydroxide at 95 – 100 °C for 4 h. The product is precipitated by addition of acid, filtered off, and washed to remove acid and also a small quantity of the more soluble 2,6-isomer. The yield is 95 %.

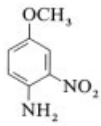
194

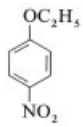
195

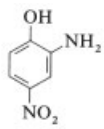
196

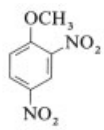
197

Uses. 2,4-Dinitrophenol is used in antiseptics, as an explosives additive, as a pest control agent, and as an intermediate for black sulfur

dyes. Sulfide reduction gives 4-nitro-2-aminophenol [*99-57-0*] (**196**).

2,4-Dinitroanisole [*119-27-7*] (**197**), 1-methoxy-2,4-dinitrobenzene, $C_7H_6N_2O_5$, M_r 198.13, *mp* 83 °C, forms yellow crystals from ethanol. 2,4-Dinitroanisole is potentially explosive on heating.

Production. 2,4-Dinitrochlorobenzene is treated with methanolic sodium hydroxide in a process similar to that described for 4-nitroanisole. 2,4-Dinitrophenetole [*610-54-8*] is similarly manufactured by using ethanol.

Uses. Catalytic hydrogenation of **197** yields 2,4-diaminoanisole [*615-05-4*], which is converted by N-acylation and N′-alkylation to azo coupling components. Sulfide reduction gives 4-nitro-*o*-anisidine (**185**) as an alternative route to nitration of *o*-anisidine, but the product from this selective reduction is less pure.

2,5-Dimethoxynitrobenzene [*89-39-4*] (**198**), 2-nitrohydroquinone dimethyl ether, 1,4-dimethoxy-2-nitrobenzene, $C_8H_9NO_4$, M_r 183.13, *mp* 72 – 73 °C, forms yellow needles from aqueous ethanol.

Production. 1,4-Dimethoxybenzene is nitrated with 34 % nitric acid at 35 – 85 °C to give **198**.

Uses. 2,5-Dimethoxynitrobenzene is reduced to 2,5-dimethoxyaniline [*102-56-7*] for use as an azo coupling component or as an intermediate for Fast Blue RR Base by benzoylation, nitration, and reduction.

2-Chlorohydroquinone dimethyl ether is similarly nitrated to give 4-chloro-2,5-dimethoxynitrobenzene [*6940-53-0*] (**199**) for reduction to 4-chloro-2,5-dimethoxyaniline [*6358-64-1*].

2,5-Diethoxynitrobenzene [*119-23-3*] (**200**), 2-nitrohydroquinone diethyl ether, 1,4-diethoxy-2-nitrobenzene, $C_{10}H_{13}NO_4$, M_r 211.16.

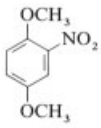

198

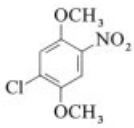

199

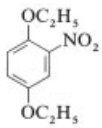

200

Production. 1,4-Diethoxybenzene is nitrated in a manner similar to the dimethyl ether to give **200**.

Uses. Reduction of **200** yields 2,5-diethoxyaniline [*94-85-9*], which is used as an azo coupling component or as the intermediate for Fast Blue BB Base by consecutive benzoylation, nitration, and reduction.

7.2. Nitroalkylphenols

Many mono- and dinitrocresols are available, but only those that can be produced economically by nitrosation or nitration of readily available cresols are significant. The next most accessible group of nitrocresols is obtained by diazotization of nitrotoluidines. Phenols substituted with higher alkyl groups are readily available, and certain of their nitro derivatives are used as agrochemicals.

4-Nitro-*m*-cresol [*2581-34-2*] (**201**), 1-hydroxy-3-methyl-4-nitrobenzene, $C_7H_7NO_3$, M_r 153.13, *mp* 128 – 129 °C, *bp* > 200 °C (101 kPa, decomp.), flash point ca. 110 °C, ignition temperature 455 °C, crystallizes as yellow needles, which are sparingly soluble in water and readily soluble in alcohol, chloroform, or benzene.

Production. *m*-Cresol is nitrosated in isopropanol solution and the isolated 4-nitroso-*m*-cresol is treated with aqueous nitric acid to oxidize it to **201**.

Uses. The derived insecticide metathion (fenitrothion) [*122-14-5*] (**202**), an analogue of methylparathion, is less toxic than parathion (**192**).

2-Nitro-*p*-cresol [*119-33-5*] (**203**), 1-hydroxy-4-methyl-2-nitrobenzene, $C_7H_7NO_3$, M_r 153.13, *mp* 33 °C, *bp* 125 °C (2.9 kPa), forms yellow needles, which are sparingly soluble in water and readily soluble in alcohol or benzene.

Production. *p*-Toluidine is diazotized and hydrolyzed to *p*-cresol, which in the presence of nitrous acid forms 2-nitroso-*p*-cresol; the latter is readily oxidized in situ to **203** by addition of aqueous nitric acid [1].

Uses. Treatment of an aqueous alkaline solution of 2-nitro-*p*-cresol with methyl chloride under pressure gives the methyl ether (**204**), which is reduced (Fe – formic acid) to cresidine (**83**). This azo coupling component is also available by an alternative process from the less readily

available 4-chloro-3-nitrotoluene (**81**; see Section 4.3).

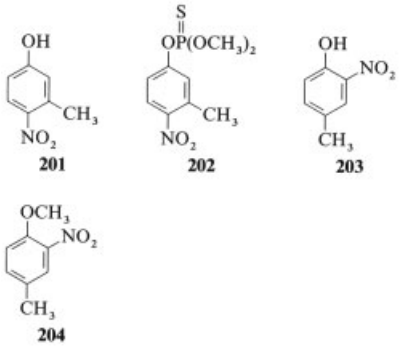

4,6-Dinitro-*o*-cresol [*534-52-1*] (**205**), DNOC, 2-methyl-4,6-dinitrophenol, 2-hydroxy-1-methyl-3,5-dinitrobenzene, $C_7H_6N_2O_5$, M_r 198.13, *mp* 86.5 °C, forms yellow crystals. The compound is sparingly soluble in water and readily soluble in alcohol, ether, or benzene.

Production. *o*-Cresol is sulfonated in excess 75 % sulfuric acid to give the disulfonic acid. The sulfonation mass is diluted with water, and 2 equivalents of nitric acid are added at 70 °C to form the dinitro derivative. The product is separated while molten and washed with hot water.

Uses. 4,6-Dinitro-*o*-cresol has been used since the 1930s as a contact herbicide, being gradually displaced by improved products, some of which have related structures. Examples are dinoseb (2-*sec*-butyl-4,6-dinitrophenol) [*88-85-7*] (**206**; R = *sec*-butyl) and dinoterb (2-*tert*-butyl-4,6-dinitrophenol) [*1420-07-1*] (**206**; R = *tert*-butyl), which are produced by controlled nitration of the corresponding alkylphenol. Members of this class of dinitro compounds have also been used as insecticides and acaricides, with dinocap [*39300-45-3*] (**206**, crotonate, R = 1-methylheptyl) specifically claimed to be effective against fruit mildew.

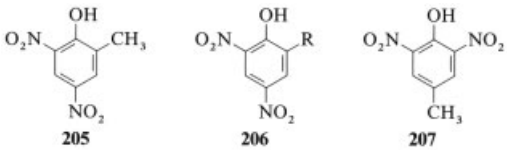

2,6-Dinitro-*p*-cresol [*609-93-8*] (**207**), 4-methyl-2,6-dinitrophenol, 4-hydroxy-1-methyl-3,5-dinitrobenzene, $C_7H_6N_2O_5$, M_r 198.13, *mp* 82 °C, forms yellow crystals.

Production. *p*-Cresol is dinitrated directly in dilute sulfuric acid solution by careful addition of nitric acid to yield 2,6-dinitro-*p*-cresol.

Uses. 2,6-Dinitro-*p*-cresol and analogous alkyl compounds have been used as agrochemicals similarly to the 4,6-dinitro-*p*-cresol.

Toxicity of Nitrophenols and Derivatives. Nitrophenol compounds are all at least moderately toxic by inhalation and skin absorption. Nitro- and, especially, dinitrophenols and cresols are pseudoacids and must be handled accordingly. 2,4-Dinitrophenol and the dinitrocresols are more poisonous by ingestion than mononitrophenols. Detailed toxicity data are available on most of the compounds, especially the known active agents (e.g., pesticides) [22].

8. Nitroketones

Acyl substituents on the benzene ring make nitration difficult, and the isomer ratio is very sensitive to temperature and nitration conditions. Introduction of an alkyl or alkoxyl group makes nitration easier, and para substitution leads to uniform products. For example, nitration of 4-methylacetophenone gives 4-methyl-3-nitroacetophenone (**208**) in good yield. Similarly, 4-methoxypropiophenone gives the 3-nitro derivative (**209**), provided that nitration is carried out at low temperature (−5 °C). Above 0 °C, ipso nitration takes place with the formation of 2,4-dinitroanisole.

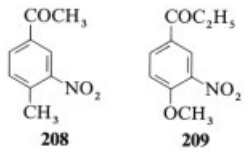

Difficult accessibility of nitroaromatic ketones together with the problem of selfcondensation of derived anilines has eliminated this class of compound from the historical colorant usage common to all other classes. Special requirements, such as pharmaceutical precursors, have a small but important place.

2-Nitroacetophenone [*577-59-3*] (**210**), $C_8H_7NO_3$, M_r 165.13, *mp* 24.5 °C, *bp* 134 °C (0.5 kPa).

Production. *O*-Acetyl-1-phenylethanol is nitrated to give a mixture of *o*- and *p*-nitrophenylethyl acetates, which are hydrolyzed with acid and fractionally distilled to yield approximately equal amounts of 1-(2-nitrophenyl)ethanol (**211**) and the corresponding 4-nitro isomer. Oxidation of **211** with dichromate – H_2SO_4 gives an 85 % yield of **210**, which can be further purified by vacuum distillation.

An alternative process involves oximination of 2-nitroethylbenzene (Section 3.5) followed by acid hydrolysis to the required product [74].

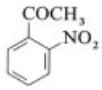

210

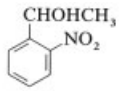

211

212

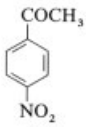

213

214

3-Nitroacetophenone [*121-89-1*] (**212**), 1-(3-nitrophenyl)ethanone, $C_8H_7NO_3$, M_r 165.13, *mp* 81 °C, *bp* 167 °C (2.4 kPa), crystallizes as needles from ethanol. The compound is steam volatile.

Production. Acetophenone is carefully dissolved in concentrated sulfuric acid at 0 °C, and mixed acid is added slowly below 0 °C. The crude product is precipitated by quenching into ice, filtered, and washed acid free. Purification, mainly from the ortho isomer, is achieved by recrystallization from ethanol with significant lowering of the overall yield to ca. 55 %.

4-Nitroacetophenone [*100-19-6*] (**213**), 4-nitrophenyl methyl ketone, $C_8H_7NO_3$, M_r 165.13, *mp* 80 – 81 °C, *bp* 138 °C (0.2 kPa), forms yellow prisms.

Production. 1-(4-Nitrophenyl)ethanol, coproduced with the 2-nitro compound, is oxidized with dichromate to give **213**, which is purified, if required, by recrystallization from methanol.

Uses. One of several alternative routes to synthetic chloroamphenicol (**1**) uses **213** as starting material with a first-stage bromination to α-bromo-4-nitroacetophenone [*99-81-0*] (**214**) followed by buildup of the side chain. The intermediate **214** is the key precursor in the synthesis of the bronchodilator clenbuterol [*37148-27-9*].

9. Nitroheterocycles

The chemistry of nitro derivatives of many heterocyclic systems has long been established, but their commercial exploitation has developed only since the 1930s. The more important of the diverse outlets are summarized here. Many of the compounds mentioned are likely to be covered in more detail under alternative headings (e.g., antibiotics or dyes) but not, however, compared in either synthesis or mode of action. Given the large number of nitroaryl derivatives that are used as agrochemicals, surprisingly few examples of nitroheterocyclics are reported in this field. Availability in bulk for use in one area often leads to outlets in other fine chemicals.

The nitrofuran group of broad-spectrum antibacterial drugs, based on 5-nitro-2-furaldehyde (**215**), was developed in the 1940s [75]. The less important nitroimidazole group of antibiotics was studied in the 1950s, following isolation of the naturally occurring azomycin (**216**). Also about this time, synthetic nitrothiazoles, typified by 2-acetylamino-5-nitrothiazole [*140-40-9*] (**217**), were found to have antitrichomonal activity.

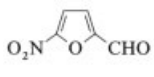

215

216

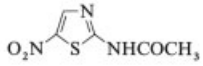

217

The 1950s witnessed a change in the field of colorants with the introduction by Eastman Kodak [76] of the nitroheterocyclic-based blue disperse dye (**218**) for cellulose acetate fibers. Since then, a succession of useful nitroheterocyclic diazo components have been developed in the dyemakers continual search for high strength and more bathochromic azo dyes [77].

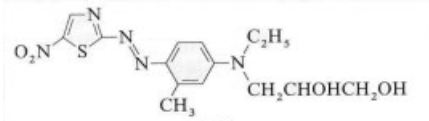

218

In this brief review, only those nitroheterocycles of known commercial interest are considered. This limits the systems to pyridine, quinoline, imidazole, furan, thiophene, and thiazole.

9.1. Pyridine Derivatives

Pyridine, in common with many other nitrogen heterocycles, is not rapidly nitrated because of the deactivation deriving from the protonated ring nitrogen. High-temperature nitration in oleum gives a low yield of 3-nitropyridine [*2530-26-9*]. However, pyridine *N*-oxide is nitrated under moderate conditions (mixed acid at 100 °C) [78] to give 4-nitropyridine *N*-oxide [*1124-33-0*] (**219**), which requires purification from a small quantity of the 2-isomer.

Hydrogenation of **219** under different conditions yields either the amine oxide or 4-aminopyridine [*504-24-5*]. Compound **219** is also a useful intermediate in the synthesis of many 4-substituted pyridines by facile nucleophilic displacement of the nitro group.

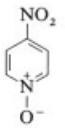

219

220

221

9.2. Quinoline Derivatives

Quinoline is readily nitrated in the isocyclic ring with mixed acid at 0 °C, yielding a 1 : 1 mixture of 5-nitroquinoline [*607-34-1*] (**220**) and 8-nitroquinoline [*607-35-2*] (**221**). By analogy with pyridine *N*-oxide, quinoline *N*-oxide nitrates to give the 4-nitro derivative, which is reduced to 4-nitroquinoline [*3741-15-9*] or 4-aminoquinoline [*578-68-7*], depending on the reducing agent. The 5- and 8-nitroquinolines are reduced to 5-aminoquinoline [*611-34-7*] and 8-aminoquinoline [*578-66-5*], respectively. All of these derivatives are patented in a variety of uses, many of which depend on the modified metal–ligand behavior compared to quinoline and 8-hydroxy quinoline in, for example, corrosion inhibition applications.

9.3. Imidazole Derivatives

2-Nitroimidazole [*527-73-1*] (**216**) has low toxicity and exhibits broad antibiotic activity. Many substituted nitroimidazoles have been synthesized in an attempt to clarify the structure–activity relationships of such apparently simple molecules. Synthesis of **216** is difficult and requires reaction of 2-aminoimidazole sulfate with sodium nitrite in the presence of excess copper sulfate [79]. 4-Nitroimidazole [*3034-38-6*] (4-nitroglyoxaline, **222**), containing a proportion of the 5-nitrotautomer, is more readily obtained by direct nitration of imidazole [80].

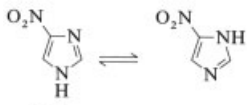

222

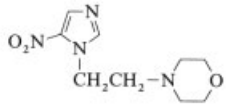

223

Alkylation of the sodium derivative of **222** with 2-morpholinoethyl chloride gives nimorazole [*6506-37-2*] (**223**). Similarly, 2-methylimidazole is nitrated [81] to give 2-methyl-5-nitroimidazole [*696-23-1*] (**224**), which on alkylation with 2-chloroethanol gives metronidazole [*443-48-1*] (**225**). Both of these drugs are used in the treatment of amoebiasis and trichomoniasis.

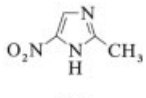

224

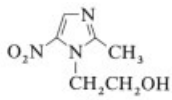

225

The nitro group is essential for the activity of imidazole compounds but may, however, be the cause of recently suspected carcinogenic potential of certain derivatives (see also → Imidazole and Derivatives, Chap. 7.2.).

9.4. Furan Derivatives

The use of nitrofurans as antibacterial agents was established in 1944 for the treatment of war wounds. In 1947 the same drugs were found to be effective in the prevention of coccidiosis in chickens and enteritis in swine. These and other veterinary applications led to large-volume usage in animal feed supplements.

5-Nitro-2-furaldehyde [*698-63-5*] (**215**), 2-formyl-5-nitrofuran, $C_5H_3NO_4$, M_r 141, *mp* 35 – 36 °C, *bp* 159 – 161 °C (1.3 kPa), forms crystals in petroleum ether; it is moderately soluble in water.

Production. Furfural is nitrated in acetic anhydride solution at 25 – 40 °C to give the diacetate [*92-55-7*] (**226**), which is purified by recrystallization from isopropanol and then used directly for the production of derived drugs [82].

Uses. Derivatives of **215** used as antibacterial agents include the semicarbazone nitrofurazone [*59-87-0*] (**227**) and the other hydrazones nitrofurantoin [*67-20-9*] (**228**) and furazolidone [*67-45-8*] (**229**), all containing the same active center. The mode of action is not known but probably involves an unstable reduction product. High activity is offset by the general toxicity of the class, and reports of mutagenic and carcinogenic activity have discouraged their use.

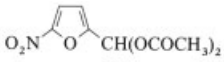
226

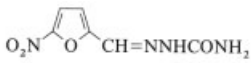
227

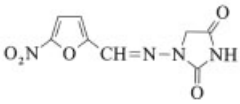
228

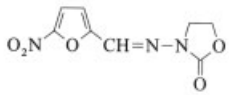
229

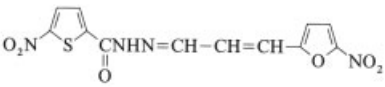
230

9.5. Thiophene Derivatives

Thiophenes are used less commonly than furans as chemotherapeutic agents, although nifurzide [*39978-42-2*] (**230**), obtained by condensation of 5-nitrothiophene-2-carboxylic acid hydrazide with 3-(2-nitro-5-furyl) acrolein, contains both a nitrofuran and a nitrothiophene residue.

Thiophene itself is nitrated to 2-nitrothiophene by gradual addition of a solution of thiophene in acetic anhydride to a mixture of nitric and glacial acetic acids at 10 °C. Attempts to circumvent acetyl nitrate processes are reported to be both difficult and dangerous. Substituted thiophenes more amenable to nitration are limited. 2-Acetylthiophene is nitrated with acetyl nitrate to give 2-nitro-5-acetylthiophene (**231**), the derived amine of which has been compared with 2-amino-3-nitro-5-acetylthiophene (**232**) as a diazo component in azo disperse dyes [83]. The latter is prepared by amination of 2-chloro-3-nitro-5-acetylthiophene; the derived dyes would have been of commercial interest if a more economical route had existed [77].

The Gewald synthesis of substituted 2-aminothiophenes [84] made 2-acetylamino-3-thiophenecarboxylic acid ethyl ester readily available for mononitration, with mixed acid at 0 – 5 °C, to give 2-amino-3-ethoxycarbonyl-5-nitrothiophene [*42783-04-0*] (**233**) [85] after deacylation. Similarly, the *N*-acetyl derivative of the readily available 2-amino-3-thiophenecarboxylic acid is dinitrated with mixed acid at 0 °C, with decarboxylation and deacylation to give 2-amino-3,5-dinitrothiophene [*2045-70-7*] (**234**) [86]. Monoazo dyes derived from **234** yield important blue and green shades on polyester, e.g., C.I. Disperse Green 9 (**235**) with λ_{max} 614 nm (cf. Table 11).

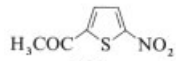
231

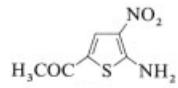
232

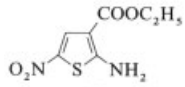
233

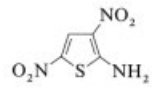
234

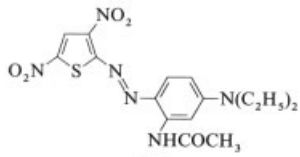
235

9.6. Thiazole Derivatives

2-Amino-5-nitrothiazole [*121-66-4*] (**236**), 5-nitro-2-thiazolamine, Enheptin, Entramin, $C_3H_3N_3O_2S$, M_r 145, *mp* 200 °C.

Production. Early production of **236** was based on nitration of 2-acetylaminothiazole and careful hydrolysis [87]. Direct nitration of 2-aminothiazole gives a cleaner product and is

best achieved by adding 2-aminothiazole nitrate [*57530-25-3*] to concentrated sulfuric acid at 0–10 °C [88]. This process and the nitration of *o*-anisidine are the only large-scale processes based on rearrangement of an amine nitrate, and both are potentially hazardous due to the possibility of runaway exothermic reactions [89].

Uses. Although **236** has been used as a selective chemotherapeutic agent, the *N*-formyl derivative, forminitrazole [*500-08-3*] and the *N*-acetyl derivative, acinitrazole [*140-40-9*] (**217**) are more important, being used in the treatment of trichomoniasis and as a veterinary antibacterial, respectively. The derivative niridazole [*61-57-4*] (**237**), used in the treatment of schistosomiasis, is obtained by reaction of **236** with 2-chloroethyl isocyanate followed by ring closure [88]. The antiparasitic thenitrazole [*3810-35-3*] (**238**) is obtained by acylation of **236** with 2-thenoyl chloride in pyridine.

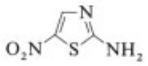

236

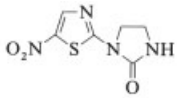

237

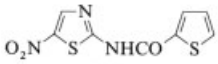

238

Production of dye **218** (C.I. Disperse Blue 106) has resulted in the synthesis of many analogous dyes [76,77,90], but with only marginal improvement in specific properties. Diazotization of **236** with nitrosylsulfuric acid, followed by coupling with tertiary arylamines, gives azo products in yields of 50–60 %. Improvements in both yield and purity are claimed to result from using displacement coupling, for example, with 4-formyl- or 4- carboxylarylamine coupling components [91].

Benzothiazole and Benzoisothiazole Derivatives. Nitroamino derivatives of benzothiazole and benzoisothiazole have both been used as diazo components. For example, **239**, derived from 2-amino-6-nitrobenzothiazole [*6285-57-0*], is a valuable bluish red dye for polyester. It is bathochromic to the extent of 68 nm in the visible spectrum, compared with the corresponding dye obtained from 4-nitroaniline [92], thus demonstrating the comparatively low polarizing effect of a nitro substituent in an aromatic ring compared with a nitro substituent in a heterocyclic ring, e.g., in dyes derived from **236**. The acylic substituted products are, however, more thermally stable and have superior lightfastness. Blue and greenish blue dyes based on 7-amino-5-nitrobenz [3,4-*c*] isothiazole (**240**) are more important because the *o*-quinonoid structure leads to more bathochromic colors than those based on **236**, while having good fastness properties [93]. Synthesis of **240** occurs by treatment of 2-cyano-4-nitroaniline (**130**) with sulfide, followed by oxidative cyclization of the derived thioamide with hydrogen peroxide [94].

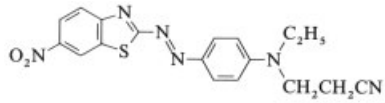

239

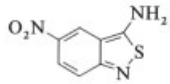

240

10. References

General References

1. B.I.O.S., F.I.A.T. reports: *German Dyestuffs and Dyestuffs Intermediates, Including Manufacturing Processes, Plant Design and Research Data 1945–1948.*
2. H. E. Fierz-David, L. Blangey: *Fundamental Processes of Dye Chemistry,* Interscience, New York 1949.
3. K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. 1, Academic Press, New York 1952.
4. P. H. Groggins (ed.): *Unit Processes in Organic Synthesis,* McGraw Hill, New York 1958.
5. P. B. P. De La Mare, I. H. Ridd: *Aromatic Substitution, Nitration and Halogenation,* Butterworths, London 1959.
6. A. V. Topchiev: *Nitration of Hydrocarbons,* Pergamon Press, London 1959.
7. M. J. Astle: *Industrial Organic Nitrogen Compounds,* Reinhold Publ. Co., New York 1961.
8. T. Urbanski: *Chemistry and Technology of Explosives,* vol. 1, Pergamon Press, Oxford 1964.

9. H. R. Schweizer: *Künstliche organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin 1964.
10. P. A. S. Smith: *The Chemistry of Open- Chain Organic Nitrogen Compounds* vol. 2, Benjamin, New York 1966.
11. N. V. Sidgwick, I. T. Millar, H. D. Springall: *Organic Chemistry of Nitrogen,* 3rd ed., Oxford 1966.
12. R. O. C. Norman, R. Taylor: *Electrophilic Substitution in Benzeneoid Compounds,* Elsevier, Amsterdam 1969.
13. H. Feuer (ed.): *The Chemistry of the Nitro and Nitroso Groups,* part 1, Interscience, New York 1969, part 2, 1970.
14. *Houben-Weyl,* **X/1,** 463 – 889.
15. E. G. Hancock (ed.): *Benzene and its Industrial Derivatives,* Benn, London 1975.
16. K. Schofield: *Nitration and Aromatic Reactivity,* University Press, Cambridge 1979.
17. *Ullmann,* 4th ed., **17,** 383 – 416.
18. S. Patai (ed.): *The Chemistry of Amino, Nitroso and Nitro Compounds and their Derivatives,* parts 1, 2, Wiley-Interscience, New York 1982.
19. L. Bretherick: *Handbook of Reactive Chemical Hazards,* 3rd ed., Butterworths, London 1985.
20. H.-G. Franck, J. W. Stadelhofer: *Industrial Aromatic Chemistry,* Springer-Verlag, Berlin 1988.
21. G. Booth: *The Manufacture of Organic Colorants and Intermediates,* Society of Dyers and Colourists, Bradford 1988.
22. N. I. Sax, R. J. Lewis: *Dangerous Properties of Industrial Materials,* 7th ed. **vol. 3,** Van Nostrand Reinhold, New York 1989.

Specific References

23. T. Travis: "Early Intermediates for the Synthetic Dyestuffs Industry," *Chem. Ind. London* 1988, 508 –514.
24. J. R. Beck: "Nucleophilic Displacement of the Aromatic Nitro Group," *Tetrahedron* **34** (1978) 2057 –2068.
25. R. J. Gillespie, D. J. Millen, *Rev. Chem. Soc.* **2** (1948) 277 – 306.
26. A. I. Titov, *Tetrahedron* **19** (1963) 557 – 580.
27. G. A. Olah: "Preparative and Mechanistic Aspects of Electrophilic Nitration in Industrial and Laboratory Nitrations," *ACS Symp. Ser.* **22** (1976).
28. L. M. Stock: "Classic Mechanism for Organic Nitration," *Prog. Phys. Org. Chem.* **12** (1976) 21 – 47.
29. C. Hanson, J. G. Marsland, K. G. Wilson: "Macro-kinetics Applied to Large-Scale Nitrations," *Chem. Ind. (London)* 1966, 675 – 683.
30. R. C. Miller, D. S. Noyce, T. Vermeulen, *Ind. Eng. Chem.* **56** (1964) no. 6, 43 – 53.
31. R. C. Hahn, H. Shosenji, D. L. Strack: "Recent Developments in *ipso*-Nitration" in *Industrial and Laboratory Nitrations, ACS Symp. Ser.* **22** (1976).
32. Josef Meissner GmbH & Co., DE 2 921 487, 1979.
33. Nobel Chematur AB (Sweden); personal communication.
34. G. Erlandsson, *Chem. Ztg.* **104** (1980) 353 – 358.
35. DuPont, US 2 256 999, 1941 (J. B. Castner).
36. Am. Cyanamid, US 4 021 498, 1975 (V. Alexanderson, J. B. Trecek, C. M. Vanderwaart).
37. DuPont, US 3 981 935, 1975 (R. McCall).
38. P. N. Preston, G. Tennant, *Chem. Rev.* **72** (1976) 627 –678.
39. Bofors, GB 1 207 384, 1966 (C. L. Hakansson et al.).
40. G. F. P. Harris: "Isomer Control in the Mononitration of Toluene" in *Industrial and Laboratory Nitrations, ACS Symp. Ser.* **22** (1976).
41. G. Booth, *A. I. Chem. Eng. Progress Manual,* **vol. 10,** (1976).
42. Mobay, US 3 350 466 , 1966 (W. K. Menke).
43. H. E. Fierz-David, R. Sponagel, *Helv. Chim. Acta* **26** (1943) 98 – 111.
44. Bayer, DE 2 517 437-9, 1975 (H. U. Blank et al.).
45. H. C. Brown, W. G. Bonner, *J. Am. Chem. Soc.* **76** (1954) 605 – 606.
46. Sumitomo, JP 7 505 339, 1973 (T. Ikeda, T. Hadano).
47. N. R. Ayyanger, A. G. Lugade, *Colourage* **29** (1982) no. 13, 3 – 9.
48. Monsanto, GB 1 420 733; 1 421 107, 1973 (I. Schumacher).
49. J. R. Kosak: "Catalytic Hydrogenation of Aromatic Halonitro Compounds," *Ann. N. Y. Acad. Sci.* **172** (1970) 175 – 186.
50. Hoechst, DE 2 156 285, 1971 (K. Baessler).
51. N. R. Ayyanger, A. G. Lugade, *Colourage* **29** (1982) no. 26, 3 – 9.
52. G. C. Finger, C. W. Kruse, *J. Am. Chem. Soc.* **78** (1956) 6034 – 6037.
53. C. D. Johnson, A. R. Katritsky, S. A. Shapiro, *J. Am. Chem. Soc.* **91** (1969) 6654 – 6662.
54. P. F. Gordon, P. Gregory: *Organic Chemistry in Colour,* Springer-Verlag, Berlin 1983.

55. P. Bersier, L. Valpiana, H. Zubler, *Chem. Ing. Tech.* **43** (1971) 1311 – 1315.
56. J. H. Ridd in T. Urbanski (ed.): *Nitro Compounds,* Pergamon Press, Warsaw-Oxford, 1964.
57. C. S. Rondestvedt, *Ind. Eng. Chem. Prod. Res. Dev.* **16** (1977) 177 – 179.
58. Röhm and Haas, DE 2 409 686, 1973 (J. E. Ware, D. L. Pearden, E. E. Kilbourn).
59. Sandoz, GB 952 468, 1959.
60. ICI, GB 1 220 448, 1967 (T. D. Baron, B. R. Fishwick).
61. ICI, GB 1 226 950, 1967 (T. D. Baron, B. R. Fishwick).
62. M. Lounasmaa, *Acta Chem. Scand.* **22** (1968) 2388 –2390.
63. Bayer, US 3 079 297, 1960 (E. Schraufstaetter, R. Goennert).
64. Sandoz, GB 855 488, 1958.
65. Sandoz, GB 1 051 264, 1963.
66. Kuhlmann, FR 1 458 333, 1965 (M. Jirou, J. Leroy).
67. Bayer, GB 1 125 683 – 4, 1966 (A. Gottschlich et al.).
68. Witco Chem., DE 2 328 574, 1972 (A. Benson, M. L. Mausner).
69. M. Zahradnik: *Production and Application of Fluorescent Brightening Agents,* Wiley-Interscience, Chichester 1982.
70. Bayer, DE 2 910 458, 1979 (H. Nickel).
71. N. C. Rose, *J. Heterocycl. Chem.* **6** (1969) 745 – 746.
72. Bayer, DE 3 012 800, 1980 (R. Kopp et al.).
73. R. J. Pollitt, *J. Chem. Soc.* 1965, 6198 – 6201.
74. A. H. Ford-Moore, H. N. Rydon, *J. Chem. Soc.* 1946, 679 – 681.
75. K. Miura, H. K. Reckendorf, *Prog. Med. Chem.* **5** (1967) 320 – 381.
76. Eastman Kodak, US 2 659 719, 1953; US 2 683 708-9, 1954; US 2 730 523, 1956; US 2 746 953, 1956 (J. B. Dickey, E. B. Towne).
77. M. A. Weaver, L. Shuttleworth, *Dyes and Pigm.* **3** (1982) 81 – 121.
78. H. J. Denttertog, J. Overhoff, *Rec. Trav. Chim. Pays Bas* **69** (1950) 468 – 473.
79. A. G. Beaman et al., *J. Am. Chem. Soc.* **87** (1965) 389 – 90.
80. BASF, DE 2 208 924, 1972 (H. Spaenig et al.).
81. Zaklady Chem., GB 1 418 538, 1973.
82. H. J. Sanders et al., *Ind. Eng. Chem.* **47** (1955) 358 –367.
83. J. B. Dickey et al., *J. Soc. Dyers Colour.* **74** (1958) 123 – 132.
84. K. Gewald, *Chem. Ber.* **98** (1965) 3571 – 3577.
85. ICI, GB 1 394 365, 1972.
86. ICI, GB 1 394 367-8, 1972 (D. B. Baird et al.).
87. Monsanto, US 2 573 641, 1951; US 2 573 656-7, 1951 (H. L. Hubbard, G. W. Steahley).
88. M. Wilhelm et al., *Helv. Chim. Acta* **49** (1966) 2443 –2452.
89. L. Silver, *Chem. Eng. Prog.* **63** (1967) no. 8, 43 – 49.
90. J. B. Dickey et al., *J. Org. Chem.* **24** (1959) 187 – 196.
91. Eastman Kodak, US 4 247 458, 1981 (L. Shuttleworth).
92. M. F. Satori, *J. Soc. Dyers Colour.* **83** (1967) 144 – 146.
93. BASF, GB 1 112 146, 1965 (M. Seefelder et al.).
94. H. G. Wippel, *Melliand Textilber.,* **50** (1969) 1090 –1096.

Nitrogen

PETER HÄUSSINGER, Linde AG, Werksgruppe VA, Höllriegelskreuth (München), Federal Republic of Germany
PAUL LEITGEB, Linde AG, Werksgruppe VA, Höllriegelskreuth (München), Federal Republic of Germany
BERND SCHMÜCKER, Linde AG, Werksgruppe VA, Höllriegelskreuth (München), Federal Republic of Germany

1. History 1
2. Properties 1
3. Occurrence 5
4. Production 5
4.1. Cryogenic Air Separation 6
4.2. Air Separation by Adsorption 10
4.3. Nitrogen Production by Membrane Separation 11
4.4. Oxygen Removal from Air 15
5. Storage and Transportation 15
6. Quality Specifications and Analysis . 17
7. Uses 18
8. Economic Aspects 24
9. References 24

1. History

Nitrogen was discovered ca. 1770, independently by SCHEELE and PRIESTLEY, as a constituent of air that does not support respiration or combustion. LAVOISIER gave it the name azôte (from the Greek, "lifeless"). After niter and nitric acid were recognized as compounds of the element, the name nitrogenium (Latin, "niter former") was introduced. In 1784 CAVENDISH produced nitrogen oxides from nitrogen and oxygen using an electrical discharge. Not until the 1900s did atmospheric nitrogen come into large-scale use for the production of calcium nitrate (FRANK and CARO, 1902), nitric acid (BIRKELAND and EYDE, 1905), ammonia (HABER and BOSCH, 1908), and nitric acid by ammonia oxidation (OSTWALD and BAUER, 1914). Nitrogen was first liquefied by CAILLETET in 1877; large-scale nitrogen production by the rectification of liquid air began in 1902 (VON LINDE).

2. Properties

Nitrogen, N, atomic number 7, A_r 14.0067, has the electron configuration $1s^2 2s^2 2p^3$. It is the lightest element in group 15 of the periodic table. In compounds, nitrogen occurs in the oxidation states +5 to −3, with +5, +3, and −3 being preferred.

Elemental nitrogen forms a stable diatomic molecule, known as molecular nitrogen or dinitrogen [*7727-37-9*].

Physical Properties. At atmospheric pressure and room temperature, nitrogen is a colorless, odorless, noncombustible gas. At 0 °C and 101.3 kPa, 1 L of nitrogen weighs 1.2505 g. Nitrogen condenses to a colorless liquid at −195.80 °C and 101.3 kPa and forms a white solid at −209.86 °C.

Nitrogen is diamagnetic in all states of aggregation; the magnetic susceptibility at room temperature is -12.0×10^{-6} cm^3/mol. Some physical properties of nitrogen (comprehensive data are given in [1–12]) are as follows:

M_r	28.0134
N – N bond length	109.4 pm
Triple point	
T	63.15 K
p	12.463 kPa
ϱ_g; ϱ_l; ϱ_s	0.6803; 867.7; 947 g/L
heat of fusion	25.8 kJ/kg
Boiling point (101.3 kPa)	77.35 K
ϱ_g; ϱ_l	4.61; 808.6 g/L
heat of vaporization	199 kJ/kg
Critical point	
T_{crit}	126.2 K
p_{crit}	3.39908 MPa
ϱ_{crit}	314.03 g/L
1st ionization energy	15.58 eV
Properties at 0 °C, 101.3 kPa:	
ϱ	1.2505 g/L
Relative density (air = 1)	0.967
Specific heat capacity	1.039 J g^{-1} K^{-1}
Dynamic viscosity	15.9×10^{-6} Pa · s
Thermal conductivity	23.86 mW m^{-1} K^{-1}
Real gas factor pV/RT	0.9995

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_457

Table 1. Selected thermophysical properties of nitrogen along the saturation line [a] and of supercritical nitrogen at 0.1 MPa [5] [b]

Temperature, K	Pressure, MPa	Density ϱ, kg/m^3	Enthalpy H, J/mol	Entropy S, J mol^{-1} K^{-1}	Specific heat capacity C_p, J mol^{-} K^{-1}	Thermal conductivity, W m^{-1} K^{-1}	Viscosity, μPa · sy,
63.15 [c]	0.01246	867.9	−4212.0	68.01	54.68	0.151	283
		0.6803	1809.0	163.5	29.64	0.00568	4.22
80	0.1367	796.4	−3248.0	81.47	58.00	0.130	138.0
		6.081	2218.0	149.9	32.02	0.00793	5.48
100	0.7777	688.6	−2033.0	94.73	64.75	0.0981	72.6
		31.87	2466.0	139.7	41.25	0.0115	7.21
120	2.513	525.2	−504.6	107.8	121.7	0.0646	36.5
		124.5	2103.0	129.6	115.8	0.223	10.4
126.24 [d]	3.398	314.1	864.3	118.4			
200	0.1	1.688	5800	179.9	29.22	0.0186	12.9
300	0.1	1.123	8717.0	191.8	29.15	0.026	18.0
500	0.1	0.6737	14570	206.7	29.56	0.0388	26.1
1000	0.1	0.3369	30100	228.1	32.63	0.0680	41.5

[a] At saturation, the first figures are properties of the liquid, the second figures are properties of the coexisting gaseous state.
[b] The values for the enthalpy H_0 and entropy S_0 at the reference state temperature $T = 298.15$ K are $H_0 = 8669.0$ J/mol and $S_0 = 191.502$ J mol^{-1} K^{-1}.
[c] Triple point.
[d] Critical point.

Table 2. Physical properties of solid nitrogen

Modification	Existence range, K	Crystal structure	ϱ, g/L	Vapor pressure, kPa
β	35.6 – 63.15	hexagonal	948 [a]	6.3 [a]
α	0 – 35.6	cubic	1027 [b]	1.47×10^{-11} [b]

[a] At 60 K.
[b] At 21.2 K.

Table 1 lists properties of gaseous and liquid nitrogen as a function of temperature, while some properties of solid nitrogen are given in Table 2. Figure 1 shows the vapor-pressure curve between the triple point and the critical point; Figure 2 is the enthalpy – temperature diagram.

Nitrogen has little solubility in liquids. Table 3 lists some values; for a comprehensive summary, see [15]. The solubility of nitrogen in molten metals is important in the production of some steels.

The diffusion constants for nitrogen in gases at ambient temperatures lie in the range of 0.1 – 1 cm^2/s (20 °C); in liquids, $< 10^{-4}$ cm^2/s; in solids, $< 10^{-10}$ cm^2/s. Table 4 (see page 4) lists some values.

For equilibrium data on the adsorption of nitrogen on various adsorbents, see [20, 21].

The boiling equilibria of O_2–N_2 and O_2–N_2–Ar mixtures form the basis for the low-temperature distillation of air [22, 23]; the separation of nitrogen from natural gas depends on those of nitrogen – hydrocarbon mixtures [22].

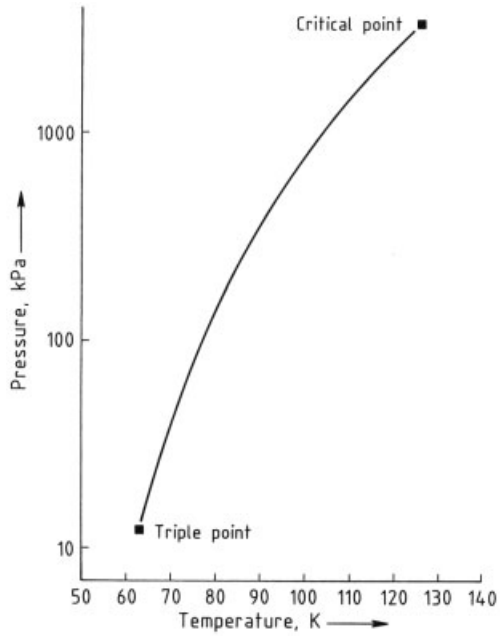

Figure 1. Vapor pressure of nitrogen

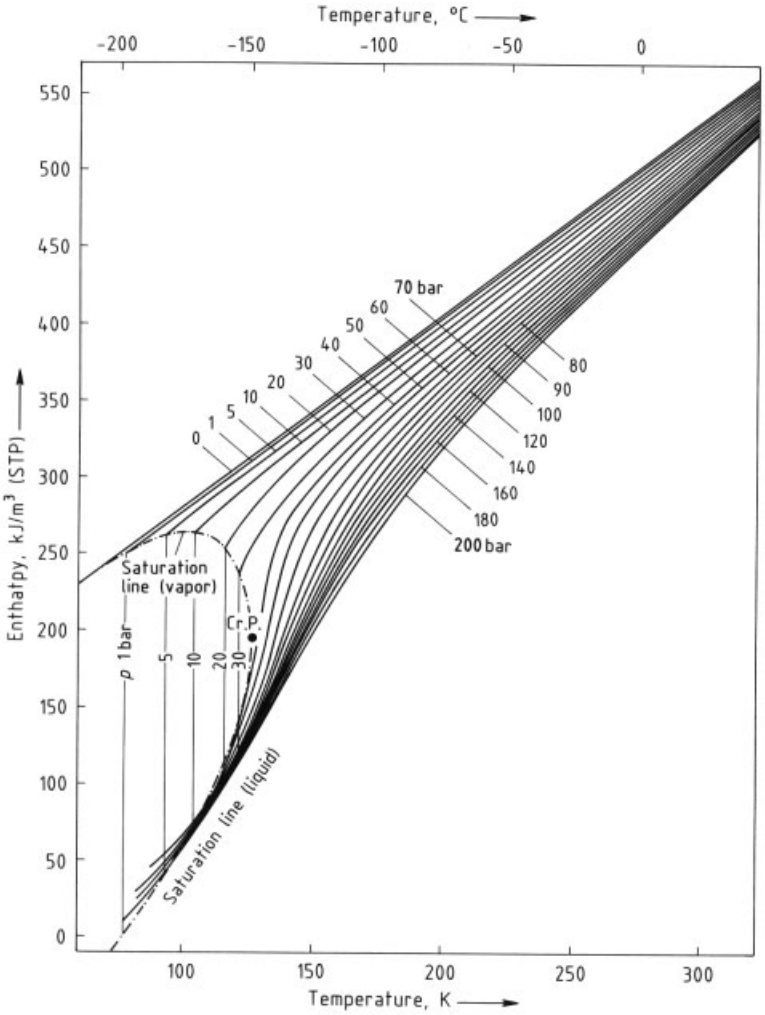

Figure 2. Enthalpy – temperature diagram of nitrogen
The heat absorption of liquid nitrogen on vaporization and subsequent heating of the cold gas can be determined from this equilibrium diagram. For example, at a pressure of 1 bar, the heat of vaporization is ca. 248 kJ/m^3 , and the heat required to warm the cold gas from 77 K to 288 K (15 °C) is 518 − 248 = 270 kJ/m^3 (1 kJ = 0.239 kcal)

Table 3. Solubility of nitrogen (further values: [13–15])

Solvent	T, °C	pN_2, MPa	Solubility, mol N_2/kg
Water	0	0.1	1.04×10^{-3}
	25	0.1	6.2×10^{-4}
	25	10	5.6×10^{-2}
	25	100	0.32
Ethanol	20	0.1	6.8×10^{-3}
Ethyl acetate	20	0.1	8.6×10^{-3}
Blood	38	0.1	6.2×10^{-4} mol/L
Olive oil	22	0.1	2.3×10^{-3} mol/L
γ-Iron	1000	0.1	8×10^{-3}

Table 4. Diffusion constants of nitrogen (further values: [16–18])

Diffusion medium	T, °C	D, cm^2/s
Nitrogen [a]	0	0.174
Hydrogen	20	0.77
Oxygen	20	0.219
Water vapor	20	0.23
Water	18	1.6×10^{-5}
Seawater	23.5	1.9×10^{-5}
α-Fe [b]	500	7.3×10^{-9}
Molten iron [b] [19]	1600	7.5×10^{-5}

[a] Self-diffusion.
[b] Diffusion as atomic N.

Table 5. Typical specific energy consumptions for the production of nitrogen gas at 8 bar (abs.) in a single- and double-column process (for standard purities)

	Single-column process, kWh/m^3 (STP) N_2	Double-column process, kWh/m^3 (STP) N_2
Air compression	0.25	0.13
N_2 compression		0.09
Total compression	*0.25*	*0.22*

Isotopes. Natural nitrogen contains the isotopes ^{14}N (A_r 14.00307) and ^{15}N (15.00011); the abundances in atmospheric nitrogen are 99.64 vol % and 0.36 vol %, respectively. The isotopic concentrations may differ in other gases, such as natural gas.

The artifical nitrogen isotopes have short half-lives: 0.011 s (^{12}N), 10 min (^{13}N), 7.2 s (^{16}N), 4.16 s (^{17}N), and 0.63 s (^{18}N). The most important processes for enriching the natural isotopes are countercurrent distillation of nitrogen monoxide and the $NO-HNO_3$ exchange process (Spindel – Taylor process) [24, 25]. Other methods are the laser-induced decomposition of tetrazine $C_2H_2N_4$ to give N_2 and HCN, which enriches $^{15}N^{14}N$ (and $H^{13}CN$) [26], and the continuous isotope separation by use of an ionexchange resin [27].

Chemical Properties. Nitrogen has an extremely high heat of dissociation:

$$N_2 \rightleftharpoons 2\,N \quad \Delta H_0 = 943.8\ kJ/mol$$

No marked dissociation takes place even at 3000 °C and standard pressure. The equilibrium constant for the thermal dissociation of molecular nitrogen is $K \approx 10^{-120}$ at 25 °C and $K \approx 10^{-6}$ even at 3000 °C. The $N\equiv N$ triple bond is nonpolar and has a significantly higher ionization potential than other comparable triple-bond systems. The strength of the $N\equiv N$ bond is responsible for the inertness of N_2.

Nitrogen reacts with lithium at room temperature, and also with calcium and magnesium at elevated temperatures, to form nitrides. Many other metals, such as Al, Ti, V, and Cr, combine directly with nitrogen at red heat, also forming nitrides. With boron, the product is boron nitride (→ Boron Carbide, Boron Nitride, and Metal Borides, Chap. 2.).

Important reactions of nitrogen with nonmetals are those with hydrogen, yielding ammonia (→ Ammonia), and with oxygen. The latter, an endothermic reaction, gives nitrogen monoxide:

$$N_2 + O_2 \rightleftharpoons 2\,NO \quad \Delta H_0 = 180\ kJ/mol$$

and has long been utilized for the production of nitric acid (→ Nitric Acid, Nitrous Acid, and Nitrogen Oxides). The reaction of nitrogen with calcium carbide leads to calcium cyanamide.

Nitrogen appears in a variety of forms in organic chemistry, such as amines, amides, nitriles, oximes, and also N-heterocyclic compounds. Naturally occurring polymers containing nitrogen are proteins and nucleic acids; those of synthetic origin include polyamides, polyimides, polyacrylonitrile, polyurethanes, and melamine and urea resins.

Compounds of trivalent nitrogen with coordination number 3 usually have a pyramidal molecular structure, one orbital of the sp^3 hybrid being occupied by an electron lone pair. Molecules of this type undergo inversion in which the N atom moves back and forth through the plane formed by the three covalently bonded ligands. The inversion barrier in NH_3 is only 25 kJ/mol.

The lone pair imparts donor properties to covalently bonded trivalent nitrogen. Since nitrogen is one of the most electronegative elements, it readily forms hydrogen bonds in its compounds with hydrogen.

The nitrogen molecule, often referred to as dinitrogen, is isoelectronic with carbon monoxide and acetylene. Transition metal dinitrogen complexes were first prepared in 1965. By now more than 100 are known [28]. The transition metals such as Mo, Cr, Co, Ni, Ti, and the platinum group metals can fix nitrogen in dinitrogen complexes (chemical nitrogen fixation). Most of

these exhibit end-on bonding of the dinitrogen molecule, but examples of side-on complexes have also been found [29].

Molecular nitrogen can also be fixed biologically by microorganisms. Both routes have been studied extensively in recent research (→ Nitrogen Fixation).

Active or atomic nitrogen can be prepared by silent electrical discharge in flowing nitrogen and at low pressures. Its persistent characteristic yellow afterglow (Lewis – Rayleigh afterglow) indicates that the recombination of the atomic to the molecular form is slow. In contrast to the largely inert N_2, active nitrogen reacts with numerous elements, as well as with inorganic and organic compounds. Most of the reactions are attributed to nitrogen atoms, but electronically excited molecules are also involved. A summary is given in [30].

3. Occurrence

The Earth's atmosphere contains ca. 78 vol % nitrogen, which is partly dissociated to atomic nitrogen at high altitudes by ultraviolet radiation. Electrical discharges (in storms) generate nitrogen oxides, which are transported to the soil by rain (ca. 1 kg of nitrate nitrogen per hectare of soil per year). Nitrogen is also present in natural gas, inclusions in rocks, and water (dissolved). Because of the high solubility of most nitrates, minerals containing nitrogen are rare; important reserves include the nitrate deposits in Chile (Chile saltpeter). The estimated total nitrogen abundance in the outer crust is 0.03 wt %.

Organically fixed nitrogen is present in the proteins of all living things and is liberated by biodegradation and excretion. The human body contains ca. 3 wt % N.

Nitrogen and its compounds also occur extraterrestrially. Nitrogen has been found in meteorite inclusions. Nitrogen has been detected in the atmosphere of Mars (2.7 vol %), Venus (3.5 vol %), on Pluto, and possibly in traces on Mercury.

4. Production

Nitrogen is obtained from air. The processes used to separate nitrogen from air can be classified as cryogenic (low-temperature) processes, adsorption processes, and membrane separation. Older methods, such as the chemical removal of oxygen from air, are of second importance.

The selection of separation technology for a particular application is dependent on the specific use requirements for the gas. While cryogenic separation can meet almost all product requirements, this is not the case for adsorption and membranes. The present relative cost effectiveness of various technologies for nitrogen at smaller capacities is illustrated by a purity versus capacity plot in Figure 3. The pace of technological development has been rapid since the 1970s for both cryogenic and noncryogenic technologies.

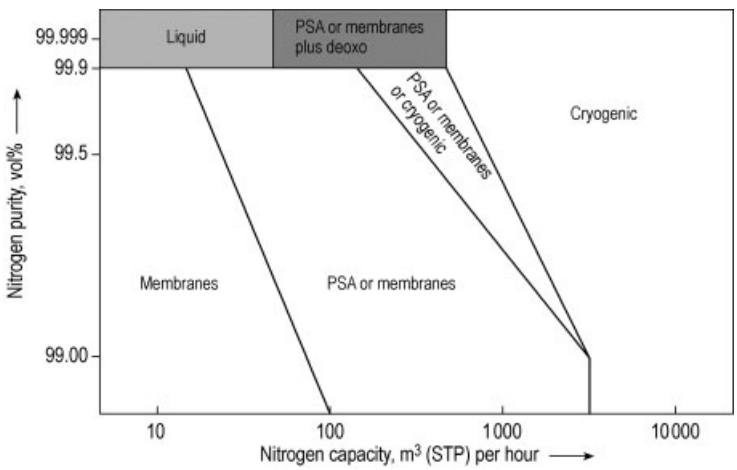

Figure 3. Areas of cost effective nitrogen production in a purity versus capacity plot [31]

4.1. Cryogenic Air Separation

Principles. To separate air by rectification, part of it must be liquefied. This can only be achieved below the critical point $T_{crit} = 132.5$ K and $p_{crit} = 37.7$ bar. The vapor pressure curves of air constituents indicate that air condenses under ambient pressure (1 bar) at a temperature of 81.5 K (− 192 °C), whereas at 6 bar only 101 K (− 172 °C) is required. To cool 1 kg of air from + 15 to − 196 °C requires ca. 415 kJ. The refrigeration required can be provided by expansion in a throttle, by the Joule – Thomson effect, or by using a turbo expander, which produces a greater amount of refrigeration energy. Starting at ambient temperature, neither effect is in itself sufficient to produce liquid air; therefore, countercurrent precooling must be used to lower the temperature to a suitable level.

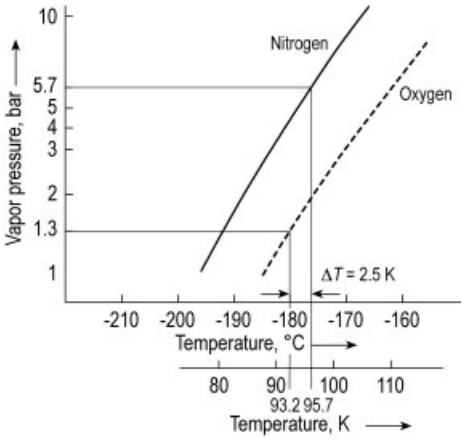

Figure 4. Vapor pressure curves of nitrogen and oxygen
The temperature difference ΔT required for heat transfer determines the operating pressures of the high- and low-pressure columns [33]

Separation of air components is performed by cryogenic rectification (→ Cryogenic Technology, Chap. 6.4.). The simplest system is the single-column process. It is used to produce gaseous nitrogen and small amounts of liquid nitrogen. If both nitrogen and oxygen products are required, the double-column process is used. The principle of double-column rectification combines the condenser of the high-pressure column (HP column) with the reboiler of the low-pressure column (LP column) in a single heat-transfer apparatus. To provide a sufficient heat-transfer rate, a temperature difference of about 2 K is required. Therefore, the columns are operated at different pressures. Figure 4 outlines the relations between the vapor pressures of oxygen and nitrogen, temperature difference and resulting operating pressures at which most double-column processes are operated. Comprehensive information is given in [32, 33].

Single-Column Process. Figure 5 shows a flow sheet of a cryogenic air separator for nitrogen with a single rectification column and molecular sieve adsorbers for front-end purification. The air separator described here is suitable for the production of both gaseous and liquid nitrogen.

Air is filtered (a) and compressed in a compressor (b), then cooled against water and/or chilled water in a tube-type or direct-contact cooler (c). After the condensed water has been removed by a scrubber or a demister (d), the air enters the zeolite adsorbers (e), where water, carbon dioxide, and other impurities (e.g., hydrocarbons) are removed. These units are usually designed as duplex adsorbers and automatically switched from adsorption to regeneration and vice versa. In the heat exchanger (g), the air is cooled almost to its liquefaction temperature in countercurrent to gaseous nitrogen product and residual gas. The air enters the rectification column (i) with an operating pressure of about 6 – 10 bar, where it is separated into pure nitrogen at the top and oxygen-enriched liquid at the bottom. Part of the gaseous nitrogen is product nitrogen at column pressure. The gaseous nitrogen that is not withdrawn at the top of the rectification column is fed to the evaporator/condenser (k). In the evaporator/condenser oxygen-enriched liquid at low pressure is evaporated from the liquid, the heat required is provided by nitrogen condensing at column pressure. Part of the liquid nitrogen is used to maintain the necessary reflux for air separation in the distillation column, part of it can be withdrawn as liquid nitrogen product.

The evaporated oxygen-enriched liquid leaves the evaporator/condenser as residual gas, is warmed in the heat exchanger, and enters the expansion turbine (h). The refrigeration maintained in the expander covers the cold losses as well as the refrigeration requirement for nitrogen liquefaction. After expansion of the residual gas to nearly atmospheric pressure, it is used as

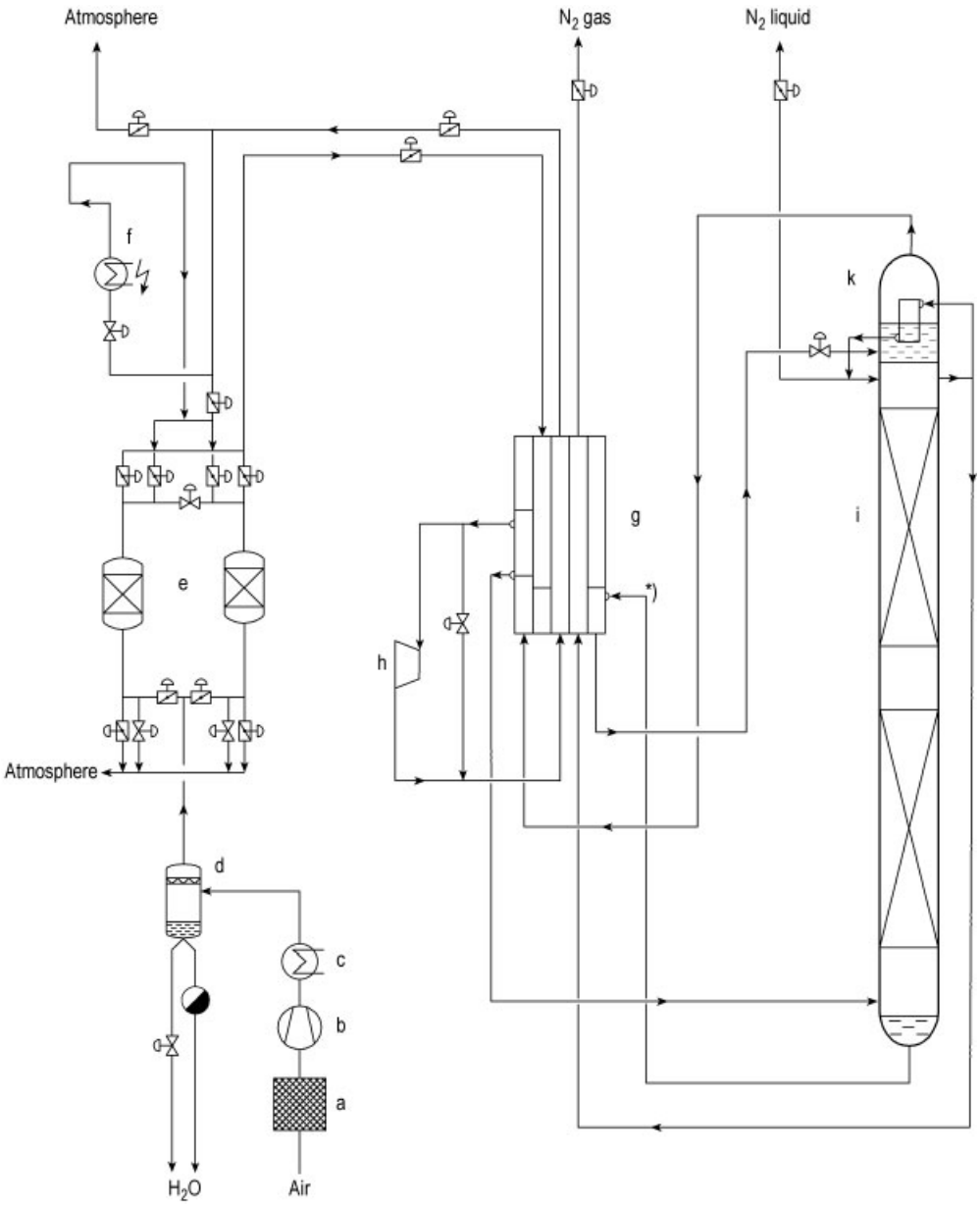

Figure 5. Flow sheet of a cryogenic air separator for production of gaseous and liquid nitrogen, single column process
a) Air filter; b) Air compressor; c) Air cooler; d) Scrubber or demister; e) Zeolite adsorbers; f) Heater; g) Heat exchanger; h) Expansion turbine; i) Rectification column; k) Evaporator/condenser

the regeneration gas for the molecular sieve adsorbers. This single-column process is normally selected for small and medium-sized units for producing nitrogen gas and liquid at a pressure suitable for final use, so that no additional product compressor is required. Commercial plants produce nitrogen containing about 1 ppm O_2 (by volume), argon, and, depending on the contents in the feed air, hydrogen, and carbon monoxide.

Rectification columns are equipped with internal trays or, more recently, structured packings. Because of the reduced pressure drop, this design results in increased flexibility in plant load.

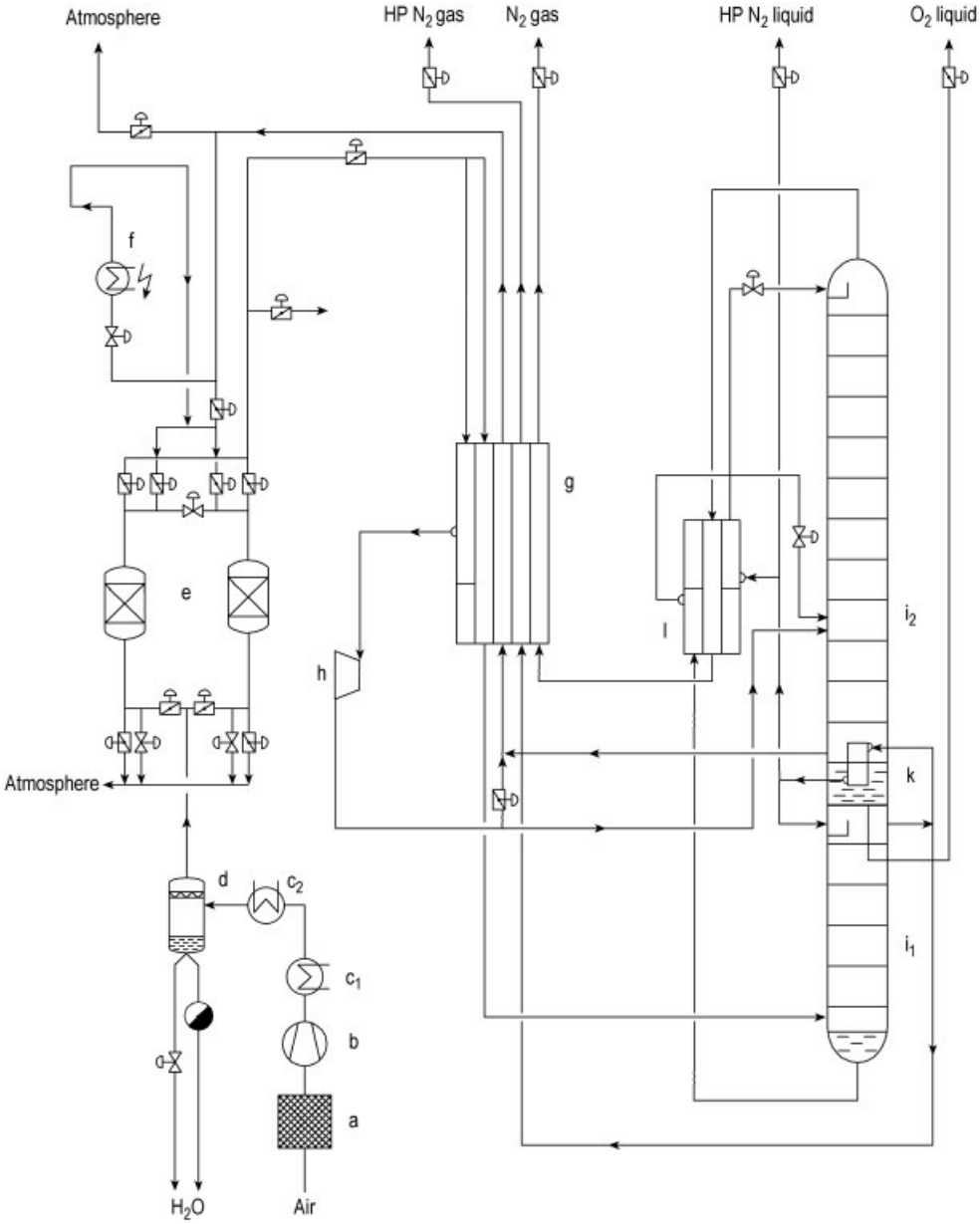

Figure 6. Flow sheet of a cryogenic air separator for production of gaseous and liquid nitrogen, double column process
a) Air filter; b) Air compressor; c_1, c_2) Air cooler; d) Scrubber or demister; e) Adsorber; f) Heater; g) Heat exchanger; h) Expansion turbine; i) Rectification columns with sieve trays (i_1) High-pressure column; i_2) Low-pressure column); k) Evaporator/condenser; l) Subcooler

Double-Column Process. Figure 6 (see next page) shows the flow sheet of a cryogenic air separator for nitrogen production by a double-column rectification process. The compression and purification of the air is as for the single-column process.

In the high-pressure column (HP column, i_1) the air is preseparated into pure nitrogen and oxygen-enriched liquid. In the low-pressure column (LP column, i_2), final separation into pur oxygen as bottom product and pure nitrogen at the top occurs. The gaseous nitrogen at the top

of the HP column is condensed in the evaporator/condenser (k). Part of it serves as reflux for the HP column, and the remainder is subcooled, expanded, and uses as reflux for the LP column. Liquid nitrogen product can be withdrawn from the top of the HP column (at high pressure) or, alternatively, from the top of the LP column (at low pressure). Pure gaseous nitrogen is withdrawn from the top of the LP column and warmed in the subcooler (l) and the heat exchanger (g). In addition, pure gaseous high pressure nitrogen is withdrawn from the HP column. Oxygen-enriched liquid at the bottom of the HP column is subcooled and expanded in the middle section of the LP column as feed flow.

Figure 7. Double-column process – three-dimensional inside view of a cold box

To maintain the refrigeration, a sidestream of air is drawn from the heat exchanger to the expansion turbine (h), expanded, and fed to the middle section of the LP column. Impure oxygen gas from the LP column and expanded air from the turbine are mixed and used to regenerate the molecular sieve adsorbers (e) after being heated in the heat exchanger.

All low-temperature equipment is assembled in an insulated cold box construction in order to minimize cold losses.

Figure 7 shows a three-dimensional inside view of a cold box structure.

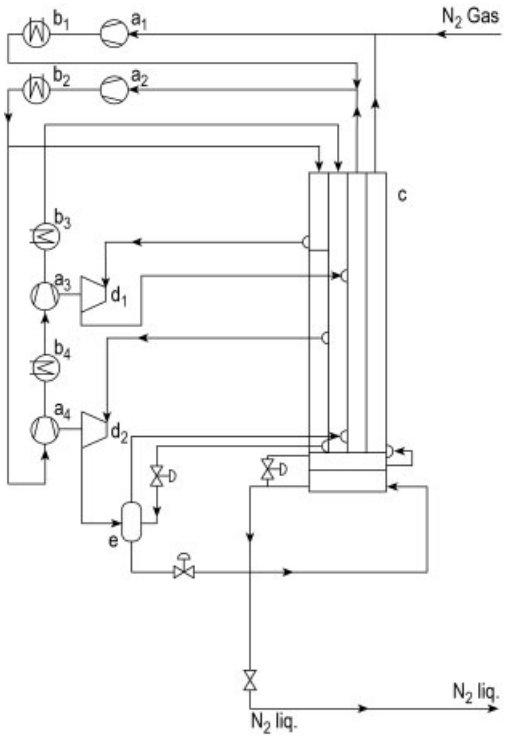

Figure 8. Nitrogen liquefaction unit with two expansion turbines for gas
$a_1) - a_4)$ compressors; $b_1) - b_4)$ Water coolers; c) Heat exchanger and subcooler; $d_1)$, $d_2)$ Expansion turbines; e) Liquid nitrogen vessel

Power Consumption. The specific energy consumptions of nitrogen generators range from 0.1 kWh/m^3 N_2 to 0.3 kWh/m^3 N_2 for the gas and depend on capacity and pressure. Table 5 lists typical specific energy consumptions in a single- and a double-column process. To achieve the same pressure level as the nitrogen gas at the

output, the nitrogen gas in the double-column process has to be further compressed by a product compressor. The total specific energy consumption is lower in the double-column process. The choice of process depends on the nitrogen capacity and pressure required, and a comparison of investment and operation costs is necessary.

Nitrogen Liquefaction. To obtain the entire nitrogen product in liquid form, additional refrigeration and liquefaction equipment is required. Several types of liquefaction units are possible. Integrated liquefiers are designed as part of the air-separation unit. In many cases liquefiers are built separately. Standard liquefier types include a process with a single turbine. Nitrogen is admitted as gaseous feed. Such liquefiers are suitable for small or medium amounts.

Figure 8 shows a schematic of a two-turbine nitrogen liquefier. Gaseous nitrogen is compressed from ambient pressure to about 600 kPa in a_1, mixed with medium-pressure recycle nitrogen, boosted to about 2800 kPa in a_2, and then further boosted isothermically in compressors a_3 and a_4 driven by expansion turbines, with aftercoolers to 5800 kPa. Refrigeration is obtained by expanding sidestreams of high-pressure nitrogen in the turbines d_1 and d_2 to a medium pressure level. The main nitrogen stream, which after exchanger c is well below its critical temperature, is expanded by a throttle valve and forms a vapor and liquid phase. Phase separation is performed in vessel e, from where product liquid is withdrawn.

Cold cascades with separate or mixed loops are used in large liquefaction units. The advantage of a cascade is that it has a specific energy consumption for liquid nitrogen (ca. 0.5–0.6 kWh/m^3 N_2) lower than that of the nitrogen loop described above (ca. 0.8 kWh/m^3 N_2) [34, 35].

Purification Methods. Nitrogen production by cryogenic distillation of air inherently produces a product of high purity. It removes virtually all bulk and trace gases whose boiling points are well above that of nitrogen. Methane, water, and carbon dioxide are removed to less than ppb levels by volume. The very low boiling gases helium, hydrogen, and neon are not removed by standard single-column cryogenic processes, and nor is carbon monoxide, which is difficult to separate by cryogenic fractionation because of its boiling point of 81.6 K. Several purification systems have been developed to produce ultrahigh-purity nitrogen to comply with specifications in the sub-ppb level. Front-end purification prior to the cryogenic nitrogen generator as well as downstream purifier systems are used.

CO and H_2 can be removed from the compressed air before distillation by oxidation in an upstream catalytic reactor with a noble metal catalyst. The oxidation products water and carbon dioxide are removed by the conventional air-purification system. The distillation section is then designed to remove oxygen (and Ar if necessary) down to less than 1 ppb [36].

Downstream purifier systems operate by oxidation over metal oxides, primarily manganese oxide, which is capable of oxidizing carbon monxide, and chemisorption by transition metals. Thus, nickel at ambient temperature eliminates oxygen as the oxide and hydrogen as the hydride. Higher pressures favor more effective contaminant removal. Finally, a filter system is installed to remove particles. Typically, very high performance, series/parallel canister filters of electropolished stainless steel are used. Particulate loadings on the order of 0.35 particles per liter with a diameter greater than 0.01 µm are achieved. Purifier systems are described in [37] and [38].

New purification concepts using systems of the getter or cryoadsorber type are under development [39].

4.2. Air Separation by Adsorption

Nitrogen production by pressure-swing adsorption (PSA) employs carbon molecular sieves or zeolitic molecular sieves. Carbon molecular sieves were developed by Bergbauforschung of Germany. Today modified carbon molecular sieves are produced by a number of other firms after expiration of the original patents. Zeolitic molecular sieves were first produced on an industrial scale by Union Carbide in the United States. These materials are employed in various processes for the separation and purification of gases and liquids; however, up to now they have not been used extensively for the production of

nitrogen. Separation of air components with zeolites is based on differences in adsorption capacity; with carbon molecular sieves, on differences in diffusion rate. Nitrogen is produced mainly on carbon molecular sieves, whereby the unadsorbed N_2-rich product gas is withdrawn at the discharge end of the adsorber during the adsorption phase [40–42]. In the zeolite process, previously adsorbed nitrogen is recovered by lowering the pressure during the regeneration phase.

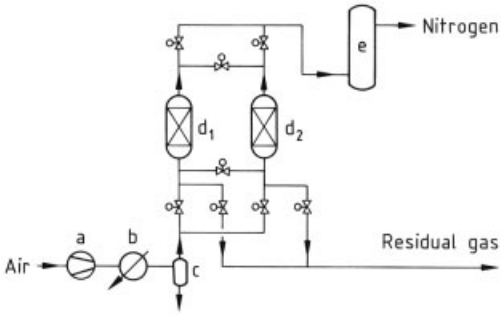

Figure 9. Schematic of a PSA plant for the production of nitrogen from air on carbon molecular sieve
a) Compressor; b) Heat exchanger; c) Condensate separator; d_1), d_2) Adsorbers; e) Buffer vessel

Figure 10. Pressure swing adsorption unit for production of 1400 m^3 nitrogen (STD) (Linde AG, Höllriegelskreuth)

Figure 9 shows schematically a PSA nitrogen unit. The feed air is compressed to 5 – 12 bar, passed through a water-cooled heat exchanger and a condensate separator, and then fed to the adsorbers (d), which alternate between adsorption and regeneration. Water, carbon dioxide, and oxygen are removed by adsorption on carbon molecular sieves. The unadsorbed nitrogen leaves the adsorber during the adsorption phase, goes to a buffer vessel (e) at 4 – 11 bar (i.e., ca. 1 bar below the selected inlet pressure) and is delivered from the vessel. At the same time, the second adsorber is being regenerated by pressure reduction. In principle, the pressure level during the adsorption and regeneration phase can also be lowered; the compressor is replaced by a blower, and a vacuum pump is used to reduce the pressure for regeneration. Figure 10 shows a unit with a capacity of 1400 m^3 nitrogen (STP) containing 1 vol % residual O_2.

The PSA process can economically produce nitrogen with a residual oxygen content of 100 ppm by volume without further catalytic treatment. By employing newly developed carbon molecular sieves, it is even possible to reduce the residual oxygen content in the nitrogen product gas to less than 10 ppm. The dew point is − 60 to − 70 °C, and the CO_2 content is below 5 ppm by volume [43].

The production capacity depends on the required purity and the permissible residual O_2 content in the product. For example, if the O_2 level in the product is raised from 0.5 to 2 vol %, the capacity doubles. The specific energy requirement for producing nitrogen by the BF process at 3 vol % residual oxygen is 0.2 kWh/kg N_2; at 1000 ppm oxygen by volume, 0.5 kWh/kg N_2. By utilizing newer carbon molecular sieves with improved separation properties, the above values can be reduced by as much as 20 %.

4.3. Nitrogen Production by Membrane Separation

Membranes can also be used for the separation of oxygen and nitrogen. For detailed physical and design principles, see → Membranes and Membrane Separation Processes.

Physical Principles. Gases permeate through dense nonporous membranes by solution diffusion. The permeability coefficient P_i, which is a measure of the capacity of the membrane to transport gas species i, is the product of the Henry's law constant S_i and the diffusivity coefficient D_i for component i in the membrane material. The volumetric flux of the component

Table 6. Oxygen and nitrogen permeabilities and selectivity of selected polymers for air separation [44–47]

Polymer	Temperature, °C	Permeability, Barrer *		O_2/N_2 selectivity α
		O_2	N_2	
Microporous Vycor glass	25	53 000	56 700	0.935
Poly(1-trimethylsilyl)propyne	23	10 040	6745	1.49
Poly(1-trimethylsilyl)propyne, fluorinated	23	1963	413	5.5
Silicone rubber	35	933	466	2.0
Polydimethylsiloxane	23	602	281	2.14
Polyethylene (density 0.935 g/cm^3)	23	1.3	0.45	2.89
Polybutadiene (butadiene rubber)	23	19.1	6.45	2.96
Polytetrafluorethylene	23	4.2	1.4	3.0
Ethylcellulose (50 % ethoxycellulose)	30	11.2	3.3	3.4
Poly(4-methyl-1-pentene)	35	27.3	6.5	4.2
Polycarbonate (LEXAN)	23	1.41	0.3	4.7
Cellulose acetate	30	1.0	0.18	5.5
Polysulfone (aromatic)	25	1.06	0.17	6.2
Aromatic polyamide	25	3.1	0.45	6.8
Polytriazoles [poly(*p*-,*m*-phenylene-1,2,4-triazoles)]	40	1.2	0.13	9.0
Polyimide UBE	30	0.5	0.045	11.0

* $1\,\mathrm{Barrer} = \frac{\mathrm{cm}^3\,(\mathrm{STP})\,\mathrm{cm}}{\mathrm{cm}^2\,\mathrm{s\,cm\,Hg}} \cdot 10^{-10}$.

through a membrane of surface area A and thickness l, whose two faces are kept at pressures p_1 and p_2 is given by

$$J_i = P_i A / l \Delta p_i$$

where Δp is the difference in partial pressures $(p_1 - p_2)$ across the membrane. The ability of a membrane to separate a pair of gases is described by the ratio of permeabilities $\alpha = P_i/P_j$, which is called the selectivity or separation factor for gaseous components i and j. Table 6 (see next page) lists the permeability coefficients and O_2/N_2 selectivities for a variety of polymers. Virtually all membranes are more permeable to oxygen than they are to nitrogen. Hence, the nitrogen product from a membrane separation stage is the retentate on the high-pressure feed side. For a monography on polymeric gas-separation membranes, see [45].

Status of Membrane Technology. Economical membrane operation requires a high flow rate (of oxygen, in the case of nitrogen production) and a O_2/N_2 selectivity of at least 4. Furthermore, a thin membrane is essential to give a high gas flux. Thin polymeric films by themselves are too weak to withstand the high differential gas pressures required. The commercial breakthrough of membranes in gas separation came with the fabrication of asymmetric and thin-film composite membranes. The separation membrane can be arranged either in spiral wound modules or in the shape of hollow fibers; the latter are normally mounted together in bundles. For the fabrication of hollow fibers see [48]. The state of the art 1990 are membrane fibers with properties as listed below [48]:

O_2/N_2 selectivity	6.0
O_2 permeability, Barrer	2.0
Skin thickness, nm	50
Membrane cost, \$/m^2	180
Feed pressure, bar	11
Permeate pressure, bar	1
Fiber geometry, µm	250 × 500
Length, m	2

The thin membrane must be virtually defect free because a higher defect area is detrimental to the selectivity α owing to Knudsen diffusion. Surface defects can be eliminated by overcoating the membrane with a 500 nm layer of a highly selective material such as silicone rubber [50].

Process Cycle Modes. Early membrane systems used single-stage processes for N_2 production, which are best suited for low-purity nitrogen. More sophisticated, multistage processes are better suited for higher purities and larger capacities. Four processes that are used for nitrogen production are shown in Figure 11. Nitrogen containing less than 10 ppm by volume of oxygen can be produced economically by using a membrane/Deoxo hybrid system; oxygen is removed by reducing it to water over a catalyst

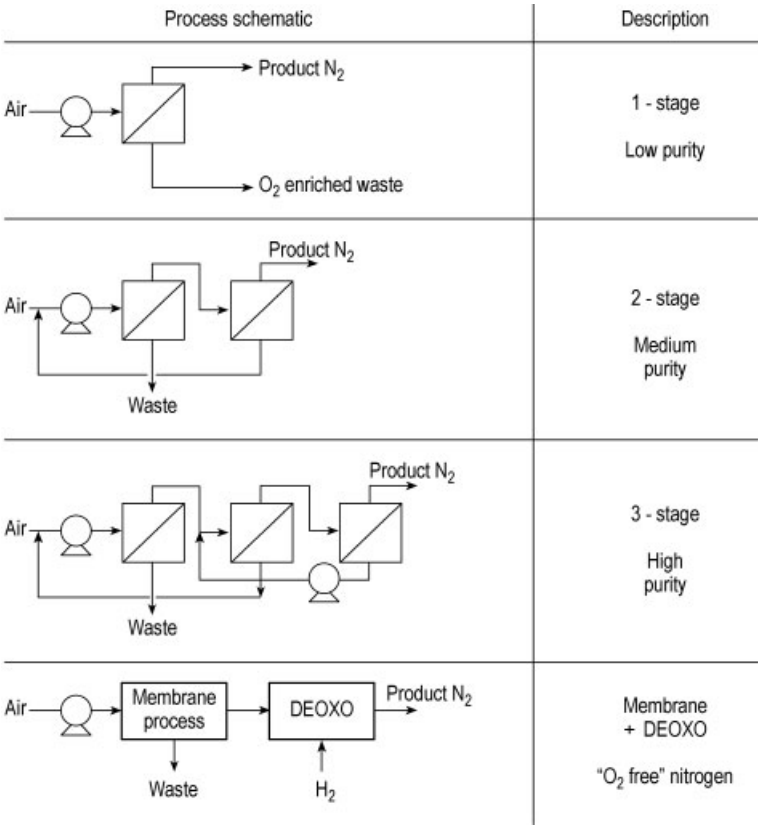

Figure 11. Configurations of membrane processes for nitrogen production [49]

with hydrogen. The gas leaving the catalyst bed is cooled, dried, and delivered as product.

A number of hybrid processes which combine membrane and nonmembrane processes such as membrane/cryo and membrane/PSA processes have been studied and patented.

Commercial Membrane Air Separation Modules. The main components of a membrane system include the compressor, the feed-conditioning package, the membrane modules and the system controls. A schematic of a typical arrangement of these components is shown in Figure 12. Air is compressed to 6 – 15 bar, and conditioned to remove contaminants and to provide temperatures and humidity levels for optimal membrane performance. The separation unit typically comprises one or more separation stages, each consisting of one or more membrane modules in parallel.

Prepackaged systems for producing 98 % N_2 range in nominal capacity from < 10 to $> 800\,m^3$ (STP) per hour. The largest nitrogen membrane units produce 600 t/d of nitrogen, which corresponds to an hourly capacity of $20\,000\,m^3$ (STP) [51]. Commercial experience indicates unscheduled down times below 1 – 1.5 %. Given protection from organic contaminants and liquid water, membrane lifetimes may exceed 10 years or more.

A detailed method for economic evaluation of membrane based air separation is given in [52]. Prediction of membrane performance along with approximation of membrane costs and lifetime are used to calculate nitrogen enrichment costs for membrane systems. Experimental data obtained during testing various small nitrogen generators from different vendors gave electrical consumption figures from 0.7 to 0.26 kWh per m^3 (STP) of nitrogen, depending on feed air quantity, purity and nitrogen recovery. See also Table 7.

Leading suppliers of membrane nitrogen units are Permea (a Monsanto company, USA),

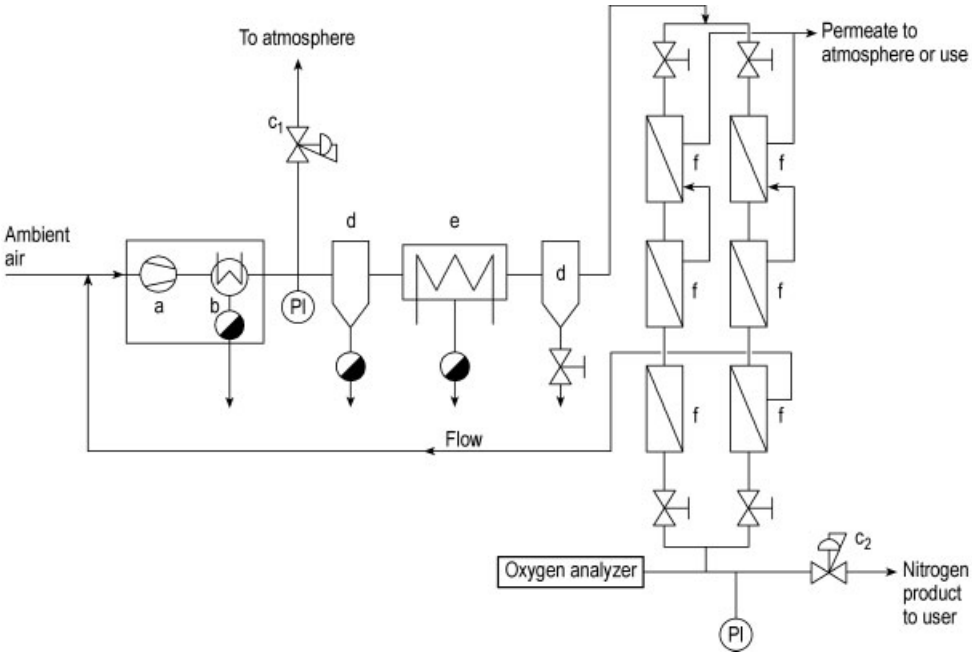

Figure 12. Schematic of a membrane nitrogen generator
a) Compressor; b) Cooler with condensate drain; c_1), c_2) Pressure control valves; d) Filter with condensate drain or valve; e) Refrigerant cooler with condensate drain; f) Membrane modules

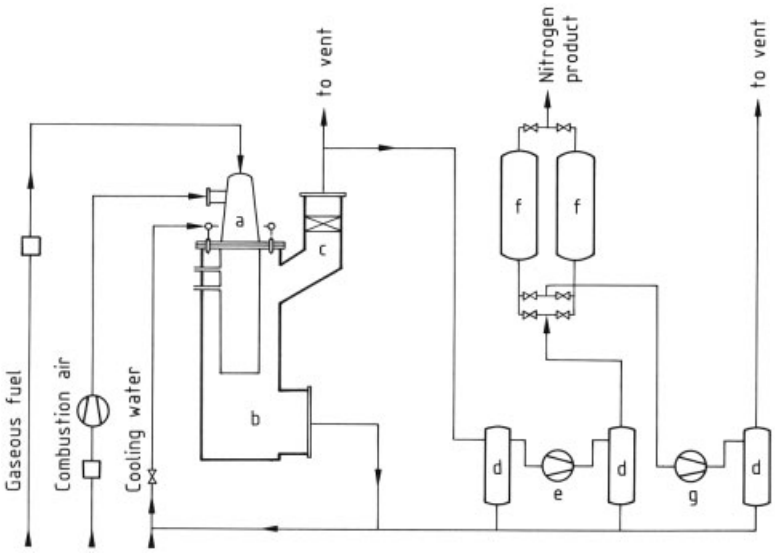

Figure 13. Nitrogen generator with combustion chamber for oxygen removal from air and CO_2 and water separation by adsorption (Calorie, Germany)
a) Combustion chamber; b) Water quench; c) Droplet demister; d) Water condensate drums; e) Blower; f) CO_2 and H_2O adsorber; g) Vacuum pump for adsorber regeneration

Table 7. Specific energy consumption per m^3 (STP) of nitrogen as a function of air feed and nitrogen recovery and purity

Air feed [a], m^3 (STP)/h	N_2 production, m^3 (STP)/h	O_2 content, mol %	N_2 recovery, %	Specific power consumption, kWh/m^3 (STP)
190	32	0.5	21.7	0.73
200	43	1.0	27.3	0.58
230	62	2.0	35.0	0.45
270	97	4.0	46.1	0.34
320	140	6.0	55.2	0.29
375	186	8.0	63.6	0.26

[a] Air feed inlet pressure 12.35 bar.

Praxair (USA), Generon, Delair (The Netherlands) and Ube (Japan).

4.4. Oxygen Removal from Air

As an alternative to air separation, which yields not only nitrogen but also an oxygen-rich gas, an inert gas can be obtained by removing all or part of the atmospheric oxygen. In the system shown in Figure 13, fuel (natural gas, propane, NGL) is burned in a specially designed burner supplied with the stoichiometric quantity of air. The hot combustion gases exit at the bottom of the reaction chamber and are directly cooled by atomized water in countercurrent. The water is separated, the gas is dried by adsorption on silica gel, and carbon dioxide is removed on molecular sieves. Apart from nitrogen and the noble gases initially present in the air, the product gas contains less than 0.1 vol % CO_2; it has $CO + H_2 < 0.01$ vol % and $O_2 < 0.5$ vol % for oxidizing burner operation, or $CO + H_2 < 0.7$ vol % and $O_2 < 0.01$ vol % for reducing operation; the dew point is $-50\,°C$. The unit shown in Figure 13, with a capacity of 1000 m^3/h N_2 (STP), consumes 117 m^3/h of methane or 46 m^3/h of propane (STP), 75 m^3/h of cooling water, and 9 kW of power [59].

In a similar way, atmospheric oxygen can be removed through the combustion of hydrogen or ammonia, the latter according to:

$$4\,NH_3 + 3\,O_2 \rightarrow 2\,N_2 + 6\,H_2O$$

with subsequent separation of water [60]. Nitrogen or a nitrogen – hydrogen mixture containing 0 – 75 vol % H_2 can also be obtained by catalytically cracking ammonia to give nitrogen and hydrogen (nickel catalyst, 800 °C; for more details → Hydrogen, Chap. 4.5.2.) and then burning the hydrogen. The cracked ammonia atmosphere is used, for example, in the surface treatment of steels.

5. Storage and Transportation

Major consumers usually produce nitrogen on site (captive nitrogen). With the development of PSA and membrane processes on-site production of smaller amounts of nitrogen is becoming increasingly feasible.

Merchant nitrogen can be stored and transported in gaseous form or as liquid nitrogen at cryogenic temperatures. In Germany, the latter is often designated LIN. The principle logistics for nitrogen (and oxygen) storage and transportation are shown in Figure 14 (see next page).

Small quantities of gaseous nitrogen are usually stored and shipped in high pressure cylinders or cylinder clusters. The charging pressure of a standard cylinder (50 L in Germany) is 200 bar, corresponding to 10 m^3 (STP) of gas. For shipping of small quantities of liquid nitrogen, insulated transportation and storage vessels are available. Large quantities are transported in road and rail tankers. Typical vacuum insulated tanker trucks can carry 10 – 22 t of liquid nitrogen with a maximum pressure of 1.7 bar. The higher cost of liquefaction is partly offset by lower transportation costs due to smaller volume and weight.

At the consumer site nitrogen is usually stored in vacuum-insulated storage tanks. The available capacities range from 3 to 80 m^3 carrying 4.6 – 95 t of liquid nitrogen and giving 2000 – 52 000 m^3 (1 bar, 15 °C) of gaseous product upon vaporization. Depending on the pres-

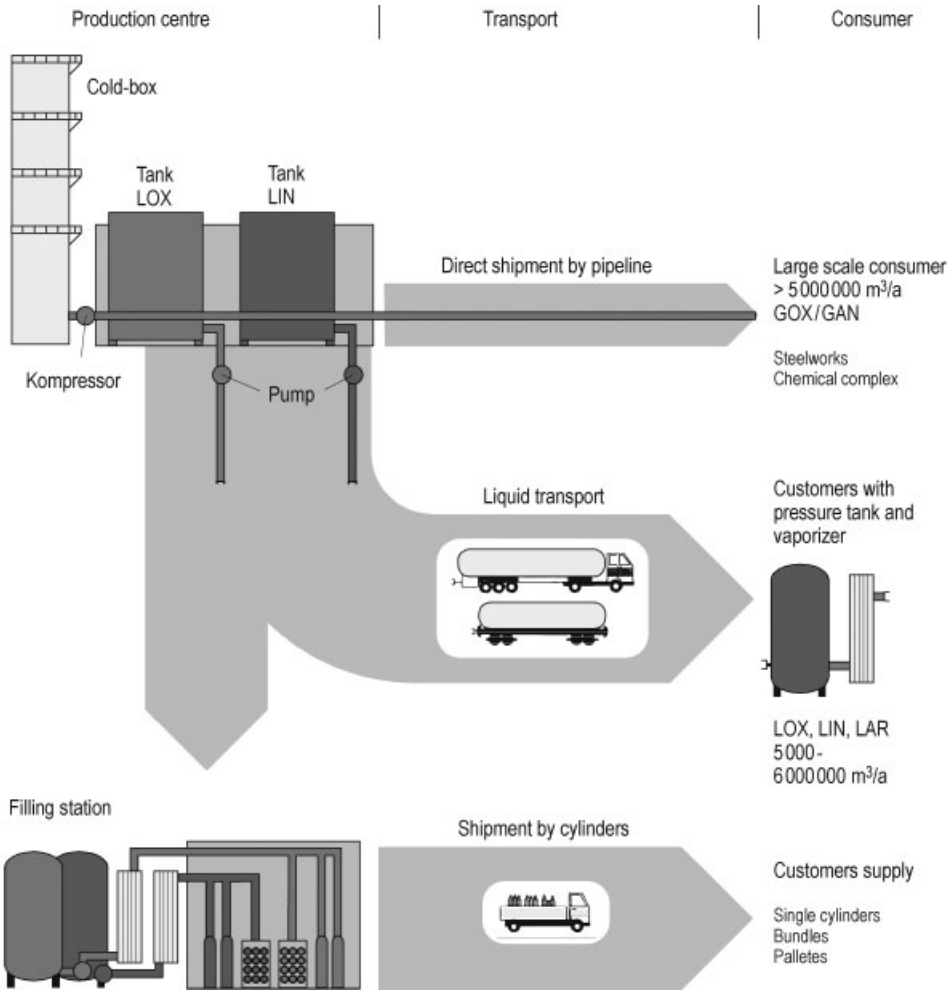

Figure 14. Nitrogen distribution from production to consumer

sure required for use, tank vessels with different operating pressures are available (e.g., 18 or 36 bar). Evaporation losses are between 0.6 and 0.15 % per day, depending on the capacity. Figure 15 shows a medium-sized tank vessel and a liquid nitrogen road tanker during filling operation.

Liquid nitrogen can be consumed directly as a liquid to utilize the cold; for users of gaseous nitrogen liquid nitrogen is vaporized in fin-tube heat exchangers heated by ambient air. Figure 16 illustrates schematically the design of such a cryogenic gasifier for liquid nitrogen.

Figure 15. Liquid nitrogen storage tank and transportation truck during filling operation

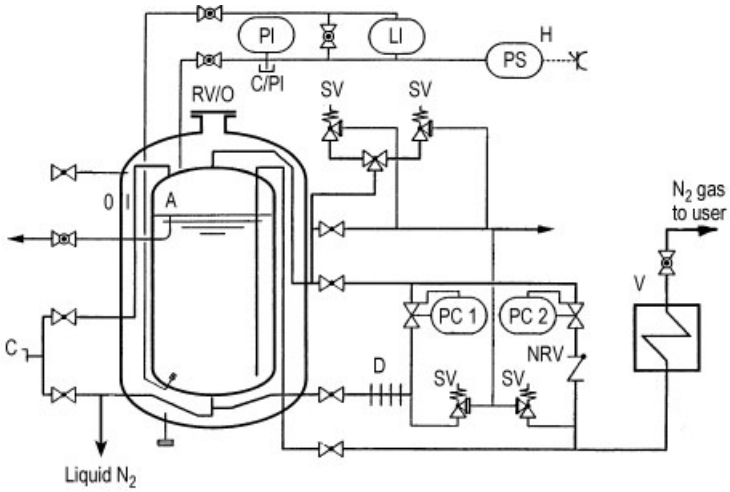

Figure 16. Schematic of a liquid storage vessel with cryogenic gasifier unit
A) Inner vessel; C) Filling coupling; D) Evaporator for pressure build up; I) Insulation; L) Liquid nitrogen line; O) Outer vessel; PC) Pressure control valve; PS) Pressure switch; RV/O) Safety flange outer vessel; SV) Safety valves; V) Evaporator

Safety precautions. Nitrogen is nontoxic but can act as an asphyxiant by displacing the amount of air necessary to sustain life; for humans this is the case for more than 88 vol %. Inhalation of nitrogen in excessive concentrations can result in dizziness, nausea, vomiting, and loss of consciousness. In pure nitrogen, fainting occurs immediately without warning followed by death after a few seconds.

The rules governing the handling of liquid and pressurized nitrogen in the country of use (in Germany Technische Regeln Gase TRG 280 and 369) and safety data sheets (e.g., from Kühn Birett S023) must be complied with. Recommendations for means by which to prevent such situations are given in CGA P-14, and CGA Safety Bulletin SB-2 "Oxygen Deficient Atmospheres" (CGA: Compressed Gas Association, USA).

Liquid nitrogen is extremely cold; contact causes freezing of human tissue or burns and severe eye damage. Liquid nitrogen spills should be removed by flushing with water and airblowing. CGA P-12 gives information on safe handling of cryogenic liquids. For further information, see [54].

6. Quality Specifications and Analysis

Quality Specifications. Nitrogen produced by cryogenic air separation (see Section 4.1) usually contains less than 5 ppm of oxygen by

Table 8. Commercial grades of nitrogen

Designation	Purity, vol % (min.)	Other constituents, vol ppm (max.)				
		O_2	H_2O	C_nH_m	Ar	CO
Nitrogen 4.6	99.996	5	5	0.5	30	
Nitrogen 5.0	99.999	3	5	0.2	3	
Nitrogen, CO-free	99.999	3	5	0.2	3	0.1
Nitrogen 5.3	99.9993	2	2	0.2	3	
Nitrogen 5.6	99.9996	0.5	1	0.1	3	0.1
Nitrogen 6.0	99.9999	0.3	0.5	0.1	0.3	0.1
Nitrogen, on-site purified *		2 **	10 **	10 **		1 **
Nitrogen, liquid	99.995					

* Other specifications = $H_2 < 5$, $CO_2 < 1$ ppb by volume; particles (0.01 µm) $< 100/m^3$ (STP) [56]

volume. Argon and carbon monoxide can also be removed down to residual levels of a few ppm. The air constituents with lower boiling points than nitrogen (helium, 5.3 ppm and neon, 18 ppm) as well as the impurity hydrogen (0.5 to 0.6 ppm) from the feed air can end up in the nitrogen product, depending on the configuration of the cryogenic unit (e.g., single-column nitrogen unit, see Section 4.1).

Table 8 outlines the commercial grades of nitrogen in Germany. In the United States, the Compressed Gas Association lists nine grades of nitrogen, differentiated by oxygen content, dew point total hydrocarbon content, and other contaminant concentrations [55].

For the use in semiconductor manufacturing ultrahigh-purity nitrogen with purity 7.0 or even higher is required. To obtain such a low impurity level, on-site purification units are used (see Section 4.1).

Liquid nitrogen normally has a purity of > 99.995 vol %. Higher purities are delivered on request.

Table 9. Trace gas analyzers used in UHP nitrogen monitoring [57]

Impurity	Analytical method	Limit of detection
O_2	Hersch galvanic cell	10 ppb
H_2O	piezoelectric quartz crystal	10 ppb
CH_4	nondispersive infrared	10 ppb
CO_2	nondispersive infrared	10 ppb
H_2	gas chromatograph, reducing gas detector	10 ppb
CO	gas chromatograph, reducing gas detector	1 ppb
Particles	laser scattering	0.1 µm
Particles	Condensation nuclei	0.01 µm
O_2, H_2O, CH_4, CO_2, H_2	APIMS *	< 0.05 ppb 2.0 ppb

* Atmospheric pressure ionisation mass spectrosmetry.

Analysis. No simple qualitative reaction is known for the detection of nitrogen. Volumetric techniques or thermal conductivity measurements are suitable for the quantitative determination of N_2 in gases. Nitrogen, along with other components, can be determined with a gas chromatograph and a thermal conductivity detector; modern high-sensitivity detectors determine N_2 down to the volume ppm level. Traces of nitrogen in, e.g., helium or hydrogen are concentrated by cryogenic adsorption and analyzed in the usual manner. For impurities in ultrahigh-purity nitrogen, as required in the electronics industry, sophisticated gas-monitoring systems are used, which can monitor ppb levels of impurities continuously. Table 9 lists the analytical methods and limits of detection [57].

The Kjeldahl technique is used to determine fixed nitrogen in organic and inorganic materials. The matter is treated with concentrated sulfuric acid over various catalysts to convert nitrogen to NH_4^+. This is treated with NaOH to liberate ammonia, which is determined by titration. In the elemental analysis of organic substances, the nitrogen oxides produced by combustion are reduced to N_2, which is determined volumetrically or by thermal conductivity. Nitrogen, fixed as amino group in peptides, proteins etc. is determined after hydrolysis and ion-chromatography by colorimetric detection; normally, these analyses are conducted automatically by amino acid analyzers.

Volatile compounds containing nitrogen can be determined in a high-sensitivity gas chromatograph with a thermionic phosphorus–nitrogen (PN) detector: CN radicals are generated in a reducing flame and reduced to CN anions by an alkali bead [61]. For elucidation of the structure of nitrogen compounds by NQR, NMR, photoelectron spectroscopy, and ESCA, see [62].

7. Uses

The largest-volume use of nitrogen is for ammonia synthesis (→ Ammonia).

Metallurgy. In metallurgy, nitrogen is an alloy constituent and is used for heat treatment [63,64] and as a scavenging and protective gas.

As an alloying element in steel, nitrogen serves to improve hardness (nitriding) and corrosion resistance [65,66] (→ Steel).

In the scavenging-gas and vacuum treatment of steel melts, nitrogen is used to remove oxygen and hydrogen, and for homogenization of the melt. The scavenging gas is introduced either through lances or porous linings. The nitrogen consumption is $0.01-0.08\ m^3$ per tonne of steel, depending on the process. Nitrogen mixed with argon is employed as a blanket gas during remelting and over the casting stream and molds; it is simultaneously used for nitriding.

The consumption of a plasma remelter ranges from 0.01 to 0.05 m^3 of N_2 and 5 to 10 m^3 of Ar per tonne of steel [67]. In continuous casting, a nitrogen gas purge is maintained over the casting stream, or the stream is blanketed with liquid nitrogen delivered by an annular nozzle [68, 69]. Nitrogen is used in place of chlorine and Freons when aluminum melts are degassed to remove hydrogen and suppress oxide formation [68, 70]. Liquid nitrogen is employed for cooling the dies in extrusion systems, for quenching steels (to lower the residual austenite content), and in the fabrication of casting molds [68, 73]. Deep cryogenic treatment with liquid nitrogen at − 196 °C enhances wear resistance and hardness of tool steels [71].

Inert-Gas Shielding, Fire Fighting. Nitrogen, as an inert gas, is used to lower the oxygen concentration below the levels required to initiate or support combustion or explosion. To eliminate the risk of fire or explosion with a combustible gas, liquid or dust, the concentration of O_2 in the gas or vapor phase must be well below the explosive range of the product vapor to oxygen ratio. Table 10 lists the O_2 concentration for several reactive materials for the minimum oxygen content needed for combustion as well as the maximum recommended oxygen content that must be maintained for the inerting with nitrogen [72]. For other values and information on the pressure and temperature dependencies, see [74, 75].

Table 10. Critical and recommended oxygen contents in nitrogen for the inerting of combustible gases and dusts [72]

Flammable gas, liquid or dust	Minimum oxygen needed for combustion, vol %	Maximum recommended oxygen, vol %
Hydrogen	5	4
Methane	12	9.5
Ethane	11	8
Ethylene	10	8
Methanol	10	8
Gasoline	12	10
Coal dust	14	10
Wood dust	11	8
Sulfur	12	9
Light-metal dusts	2 – 6	0 – 2

Nitrogen atmospheres and purges are used in pipelines, vessels, and silos where flammable gas mixtures or dusts occur. An inert atmosphere can be produced by blanketing or by continuous or intermittent purging. In most cases, continous monitoring of the oxygen concentration or flammability of the atmosphere is necessary. When nitrogen blankets are used in tanks, slight overpressures are sufficient; the exhaust gas is led to a flare system [76–78].

Fires of fire class B (liquid materials, e.g., gasoline, alcohols, oils) and C (gaseous material, e.g., methane, propane, hydrogen) can be extinguished by using nitrogen. The choice of nitrogen as a fire extinguisher depends on the cost and local conditions. In firefighting in coal mines, liquid nitrogen extinguishes the fire, renders explosive atmospheres inert, and leads to cooling. Large colliery fires were combatted successfully [79].

Food Storage and Food Processing. During processing, cooling, storage, and transportation of food, liquid nitrogen or vaporized liquid nitrogen is used. Liquid nitrogen provides the cryogenic cold necessary for freezing, and nitrogen vapor as an inert medium inhibits oxygen-dependent ripening and decay processes. Further advantages of nitrogen are its sterility; the absence of taste, odor, and color; and its nontoxicity.

The flash freezing of foods with liquid nitrogen involves a shorter time than conventional deep freezing. The formation of larger ice crystals in cells and in intercellular spaces is reduced; thus, damage of cell walls is avoided. The concentration of salts, proteins, sugars and other substances in the cell liquor therefore does not increase and so there is less water diffusion driven by osmotic pressure difference across the cellwalls. With increasing freezing velocity the drip loss of cellular liquor during thawing of the product is diminished.

Tunnel, spiral, or cabinet freezers are usually employed in cryogenic food freezing. High-speed freezing in immersion freezers is used in special cases only. Figure 17 shows the various freezer types. In the tunnel freezer, the product is precooled in countercurrent against vaporized nitrogen, then exposed to a liquid spray. The capacity of tunnel freezers varies from 0.5 to 3 t/h of frozen product. Figure 18 shows an in-line tunnel freezer during loading. In the spiral freezer, the product moves along a spiral belt through the annular nitrogen spray. Cryogenic

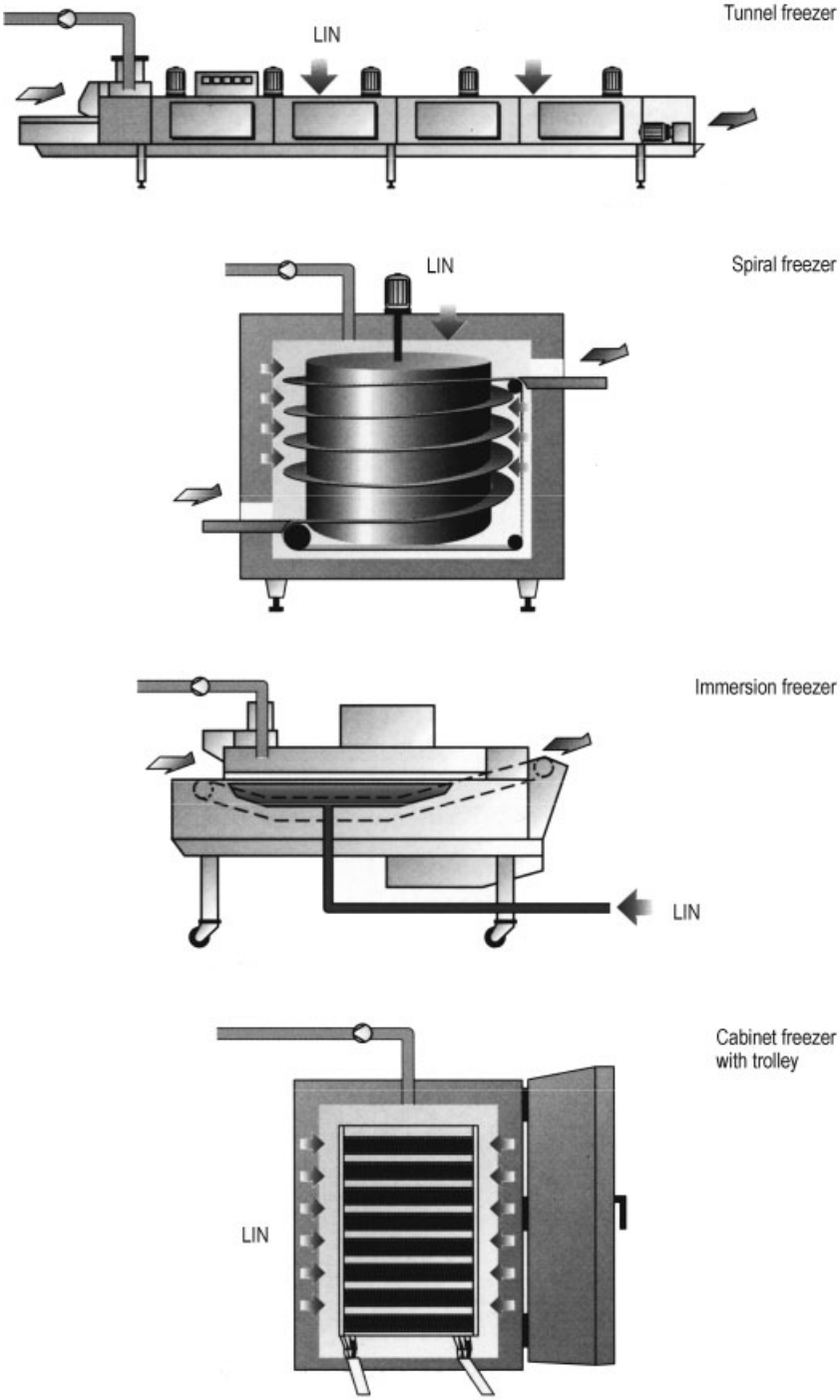

Figure 17. Schematics of the various freezer types (Linde Cryoline system)

flash-freezing requires ca. 0.8 – 1.2 kg liquid nitrogen per kilogram of product [73, 80, 81].

Goods in transit can be kept cold with liquid nitrogen instead of refrigerating equipment. A controlled quantity of liquid nitrogen is sprayed into the cargo space or vaporized from an open vessel whose insulation is adapted to the desired vaporization rate.

Food is stored under controlled atmospheres or in cold-storage houses. In controlled atmospheres (CA), the oxygen and carbon dioxide contents are usually kept at 3 – 4 vol % and max. 2 vol %, respectively, at a high relative humidity and low temperature. Thus, ripening of fruits and vegetables can be postponed until sale without appreccable quality degradation. Nitrogen can be provided by noncryogenic processes or from liquid nitrogen, which is advantageous because of its low temperature.

Figure 18. An in-line tunnel freezer in operation (Linde Cryoline)

Modified atmospheres are special gas mixtures rich in nitrogen and are increasingly used in the packaging of foods (modified atmosphere packing, MAP). Quality, texture and appearance are greatly influenced by the absence of oxygen.

In cold-storage depots a temperature of – 24 °C must be maintained. In the case of power loss or failure of the refrigeration system, liquid nitrogen can be utilized for emergency cooling. Often, liquid-nitrogen emergency cooling manifolds are installed [82].

Nitrogen is increasingly used in the brewing and soft-drinks industries. Excess oxygen is removed from brewery liquors and beer by stripping with nitrogen. Sparge deoxygenation typically reduces the oxygen content from 10 ppm to 0.1 ppm. Nitrogen serves as an inert gas for canning and bottling lines [83]. Noncarbonated beverages are pressurized during canning by injection of a small amount of liquid nitrogen so that traditional methods of packing, stacking, and shipping can be used without the presence of CO_2 [84].

Manufacturing Applications. In the plastics industry, liquid nitrogen is used for flash removal, size reduction, shrink-fitting, and extrusion blow molding. High-purity nitrogen is used extensively throughout the electronics industry.

Flash Removal. Flash must be removed from molded plastic parts. Brief cooling with liquid nitrogen renders the flash brittle while leaving the part still flexible. The result is a weak point at the transition from flash to part, so that the flash can easily be removed in a tumbler or a vibration or blasting machine (e.g., MSA and TSA flash removal, Linde, Höllriegelskreuth) [85, 86]. The consumption rate is ca. 0.2 – 1.5 kg of liquid nitrogen per kilogram of plastic.

Cryogenic Grinding and Disintegration. Embrittlement of grinding feed with liquid nitrogen makes it possible to handle adherent, soft, or tough materials. In addition to better grindability, the advantages include easier heat removal during grinding and the prevention of dust explosions by the inert atmosphere generated. Also the possibility of efficient separation is provided in cases in which some materials in a mixture become brittle, while others do not. Cryogenic grinding or cryopulverizing is employed for plastics, rubber, spices, coffee, cocoa, soybeans, waxes, etc. The method is also useful in processing pharmaceutical products that have low softening points or are heat sensitive [87]. Cryogenic disintegration or debonding is used during recycle operations, e.g., reclaiming rubber from used tires, separation of ferrous and nonferrous scrap, stripping polymeric insulation from cables, and for recycling also for batteries, electronic scrap, composite materials, and color concentrates [88–90].

Figure 19 (see next page) shows a schematic of a cryogenic grinding system, in which a cold stream of nitrogen is led in cocurrent to the grinding feed. The temperature of the cooling screw or tunnel is controlled by the injection of liquid nitrogen and is adapted to the feed material (e.g., ca. – 30 °C for spices but – 160 °C for

nylon). The consumption of a cryogenic grinding unitis 0.5 – 1.5 kg of liquid nitrogen per kilogram plastic or 0.25 kg/kg for spices; that of a disintegrator for old tires 0.5 kg per kilogram of rubber processed [91].

Shrink-Fitting. When parts are assembled with shrink fits, the inner part is immersed in liquid nitrogen so that its volume is reduced. Then it is inserted in the outer part. The heating of the inner part leads to a press fit between the parts. When cryogenic shrink-fitting is used instead of heating the outer part, metallurgical changes are avoided. Cryogenic shrink-fitting requires ca. 0.35 (brass) to 0.80 (aluminum) kg of liquid nitrogen per kilogram [114].

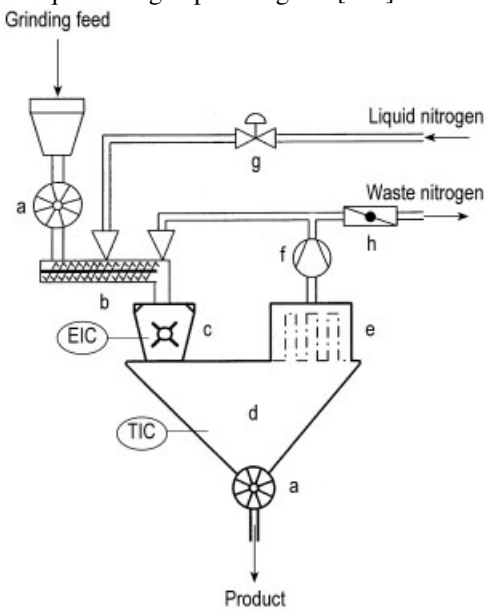

Figure 19. Flow sheet of a cryogenic grinding unit for suitable thermoplastic material [90]
a) Star feeder for feed and product; b) Cooling screw; c) Impact grinder; d) Product bin; e) Dust-removal filter; f) Recirculation blower; g) Liquid nitrogen regulating valve; h) Waste nitrogen throttle valve

Extrusion Blow Molding. Hollow plastic bodies are made by extruding the soft plastic into a hollow mold in an extrusion machine, then allowing the material to cool in the mold. The working fluid is compressed air. If liquid nitrogen is used instead of compressed air, the cycle time on the extruder can be greatly reduced. The nitrogen is injected into the hollow mold in liquid form. The pressure generated as it vaporizes forces the plastic against the mold and cools it. The use of nitrogen diminishes the tendency to stress cracking and enhances the transparency of PVC moldings. Nitrogen-assisted injection molding also results in significant weight savings in structural plastic components. The consumption is 0.1 – 0.8 kg of liquid nitrogen per kilogram of plastic [92, 93].

In the plastics industry the cold of liquid nitrogen is also used in the fabrication of high-pressure rubber hoses and for easy polishing of plastic car parts (cold polishing). Liquid nitrogen also acts as a propellant for foaming plastics.

Electronics Manufacturing. Nitrogen is used extensively throughout the electronics industry and in semiconductor production. Blanketing by standard pure nitrogen atmosphere is advantageous during soldering of electronic parts by wave-soldering processes [94] and infrared reflow soldering [95].

High-purity or ultrahigh-purity nitrogen is used in the semiconductor and chip-manufacturing industry as a blanketing medium and as an inert gas carrier for reactive process gases in chemical vapor deposition and doping processes.

Oil and Gas Recovery. High-pressure nitrogen (up to 600 bar) is used in many working processes in oil and gas drilling. For well-completion drilling, nitrogen gas is used as a drilling medium in under-pressured reservoirs. When hydrocarbon zones are passed, nitrogen minimizes the risk of downhole fires. Other effects of nitrogen are to create aerated nonflammable drilling fluids, removing liquids from the feed drill, perforating casings, and breaking up deposit materials. When liquid nitrogen is used, special high-pressure vaporisation units are required. Costly cryogenic equipment is being replaced by on-site membrane systems [96].

Nitrogen injection for enhanced oil and gas recovery has become an efficient method for oil reservoirs under initial conditions or for those depleted by water or steam flooding [97]. Nitrogen has advantages over CO_2 because of is lower solubility in water and hydrocarbons. Nitrogen for enhanced oil recovery is produced cryogenically. A typical large reservoir may require up to 100 000 m^3/h (STP) at a pressure of 41 MPa. The air separation plants at the field site are modified to produce large quantities with improved efficiency [98] (see Chapter 4.1).

Cryomedical and Cryobiological Applications. Liquid nitrogen freezing has many applications in medicine and biology. Cryogenic treatment with liquid nitrogen allows the preservation of stored blood past the maximum four-week life of conventional units. Before deep-freezing, the plasma is separated and a stabilizer (e.g., glycerol or dimethyl sulfoxide) is added to reduce osmotic damage. After immersion in liquid nitrogen, the frozen unit of blood is stored below $-110\,^{\circ}C$ in a cold N_2 gas atmosphere over liquid nitrogen. The technique gives erythrocyte survival rates in excess of 98 % [99]. Deep-freezing with liquid nitrogen also helps to preserve bone marrow, skin tissue, corneas and human sperm.

Cryosurgery is an alternative to excision surgery and is useful in the treatment of hypertrophic and malign neoplasms. Liquid nitrogen is applied by cryoprobes or cryospray. Ice balls are generated within the area to be distroyed, whereas freeze spraying is an effective way of freezing large tumors. Further clinical applications are in oral surgery, ophthalmology and rheumatism treatment. A survey is given in [100].

For preserving biological substrates, ready-made systems are available for yeast, bacteria, or proteins. For this purpose liquid nitrogen is sterilized by steril-filtration. So-called cryobanks are operated for the storage of blood conserves, livestock semen for cattle breeding, preserving environmental samples for documentation, etc. [101]. The digestion of cells and microorganisms containing active enzymes or heat-sensitive proteins can be carried out by cryogenic freeze cleaving. After immersion into liquid nitrogen, pellets of the material are formed. By cryogrinding, purification, and drying the active substances (i.e., enzymes or related biologically active material) are recovered. The yield of active substance is increased by 50 % compared to conventional digestion [102].

Other uses. Gaseous or liquid nitrogen is also used for various applications:

Ground Freezing. Freezing of water contained in unstable ground temporarily improves the strength and eliminates or reduces water movement. Ground freezing is used for making vertical shafts, sewerage, subway tunnels, etc. [103, 104]. *Pipe freezing* generates temporary plugs in pipelines to allow local cutting or repairs [105]. The recovery of tar and oil spills from contaminated soils is facilitated by freezing and embrittlement.

Cryocondensation. Waste-gas streams containing organic or chloro-organic impurities can be purified by indirect condensation of the compounds using cryogenic cold. Thus, 99 % purification can be achieved by countercurrent cooling with liquid nitrogen; final purification can be performed cold adsorption on molecular sieve adsorbers. Regeneration of the adsorbers together with the condensate from cooling yields a liquid containing 99.9 % of the impurities [106].

Tire Curing. In the tire manufacturing process, nitrogen together with steam is used to create heat and pressure required for proper vulcanization ("curing") of the tire rubber in a press mold. The curing step shapes the tire into its final form; the tire mold is fitted with a collapsible rubber bladder into which the curing media are injected. The use of oxygen-free nitrogen is advised to increase the bladder life which is a critical factor in tire curing [107].

Concrete Cooling. Concrete hardening is an exothermic process, and the large amounts of heat produced in the concrete body can cause cracking. With large concrete volumes, high quality demands on the concrete, or in hot climates, cooling with liquid nitrogen is recommended. This can be done either by precooling the water, aggregate, and cement or by directly cooling the concrete mixture by injecting liquid nitrogen [108].

Chemical Industry. In the chemical and petrochemical industry, nitrogen is used for simultaneous blanketing and cooling of reactors. Cooling times and thus idle times in, for example, hydrogenation or reforming reactors can be shortened by the injection of liquid nitrogen. Nitrogen also ensures safe operation due to its inertness [109]. The cold of liquid nitrogen is used to control or to stop chemical reactions. For example, exothermic polycondensations can be terminated by quenching the product resin in liquid nitrogen [110]. An alternative to removing coke from thermal reactor tubes by decoking with air is the Sandjet process. A blasting material is blown down the reactor pipes at high speed using nitrogen as carrier gas. Coke deposits are removed quickly and without emissions to the atmosphere. Corrosion products can be removed

by using sharp-edged cinders as blasting material [111].

There are many other applications for which nitrogen can be used; they include:

1) Filling high-performance shock absorbers or pneumatic spring suspensions to avoid oxidation and degradation of the filling oil.
2) Conditioning of natural gas to sales gas specifications from LNG storage facilities by adding nitrogen (i.e., to adjust the heating value) before compression and pipeline distribution [112].
3) The liquid nitrogen wash process used to purify hydrogen, produced from crude oil or coal, for use in ammonia synthesis (→ Gas Production, Chap. 5.6.3.).

8. Economic Aspects

In Europe nitrogen is the second most important industrial gas after oxygen. Total captive and merchant nitrogen production in 1996 amounted to $8.5 \times 10^9\,m^3$ (STP) in Germany and to $21.5 \times 10^9\,m^3$ (STP) in Europe. An annual growth of ca. 5 % per annum is expected up to the year 2001. In the United States about $22 \times 10^9\,m^3$ (STP) of nitrogen was produced in 1992, and this rose to $28.3 \times 10^9\,m^3$ (STP) in 1996. Production in Japan in 1996 was $9.8 \times 10^9\,m^3$ (STP), approximately 40 % of which was gaseous or liquid merchant nitrogen.

Gaseous nitrogen production is growing more rapidly than liquid nitrogen production for two reasons: First, the largest applications have always been most economically served by gas from multi-user or over-the-fence pipelines. Second, noncryogenic production technologies such as membranes and PSA are making gas production at customer sites economic under more and more circumstances [113].

9. References

1. *Gmelin,* 8th ed., **Lieferung 1,** Verlag Chemie, Weinheim 1936.
2. *Encyklopédie des Gaz, L'Air Liquide Division Scientifique,* Elsevier, Amsterdam 1976.
3. N. B. Vargaftik: *Tables on the Thermosphysical Properties of Liquids and Gases,* Wiley & Sons, New York, 1975.
4. R. T. Jacobsen, R. B. Stewart, *J. Phys. Chem. Ref. Data* **2** (1973) 757.
5. B. A. Younglove, *J. Phys. Chem. Ref. Data,* **11** (1982) Supplement no. 1, 1 – 170.
6. "Nitrogen" in "International Thermodynamic Tables of the Fluid State," vol. 6, *IUPAC Chem. Data Ser.* **20** (1979).
7. *Handbook of Compressed Gases,* 4th ed., CGA, Kluver Academic Pub., New York 1999.
8. J. D. Cox, D. D. Wagmann, V. A. Medvedev: *KODATA Key Values for Thermodynamics,* Hemisphere Publishing Corp., New York 1984.
9. "International Thermodynamic Tables of the Fluid State-Nitrogen", S. Angus, K. M. de Reuck, B. Armstrong (eds.): *IUPAC,* Pergamon Press, Oxford 1979.
10. B. A. Younglove, "Thermophysical Properties of Fluids," *J. Phys. Chem. Ref. Data* **11** (1982) (Suppl. No 1.)
11. R. T. Jacobsen, R. B. Stewart, M. Jahangiri, *J. Phys. Chem. Ref. Data* **15** (1986) no. 2 735.
12. NIST Standard Reference Database 69, Nov. 1998 Release: NIST Chemistry WebBook, see also http://webbook.nist.gov/chemistry/.
13. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* Pergamon Press, Oxford – New York 1979.
14. W. F. Linke: *Solubilities of Inorganic and Metal-Organic Compounds,* 4th ed., Amer. Chem. Soc., Washington 1965.
15. R. Battino (ed.) : *Solubility Data Series,* vol. 10, "Nitrogen and Air", Pergamon Press, Oxford 1982.
16. *Landolt-Börnstein* **5 a,** Springer Verlag, Berlin 1969.
17. W. Jost: *Diffusion,* D. Steinkopff Verlag, Darmstadt 1972.
18. E. L. Cussler: *Diffusion: Mass Transfer in Fluid Systems,* Cambridge University Press, Cambridge 1984, Chaps. 3,7.
19. G. Svjazin, T. El Gammal, *Arch. Eisenhüttenwes.* **46** (1975) 181.
20. D. P. Valenzuela, A. L. Myers: *Adsorption Equilibrium Data Handbook,* Prentice Hall, Englewood Cliffs, N. J. 1989.
21. D. M. Ruthven: *Principles of Adsorption and Adsorption Processes,* Wiley-Interscience, New York 1984.
22. D. Behrens, R. Eckermann (eds.): *Vapour-Liquid Equilibria for Mixtures of Low Boiling Substances,* DECHEMA Chemistry Data Series vol. VI, Schön & Wetzel, Frankfurt 1981.

23. E. Bender, *Cryogenics* **13** (1973) 11.
24. D. Staschewski, *Chem. Ing. Tech.* **43** (1971) 810.
25. D. Staschewski, *Chem. Techn. (Heidelberg)* **4** (1975) 296.
26. R. R. Karl, K. K. Innes, *Chem. Phys. Lett.* **36** (1975) no. 3, 275.
27. A. Kruglov, B. Andreev, I. Pozhidaev, *Sep. Sci. Technol.* **31** (1996) 471.
28. R. A. Henderson, G. J. Leigh, C. J. Pickett, *Adv. Inorg. Chem. Radiochem.* **27** (1983) 197.
29. B. E. Smith et. al. in H. J. Evans, P. J. Bottomley, W. E. Newton (eds.): *Nitrogen Fixation Research Progress,* Martinus Nijhoff, Dordrecht 1985, p. 597.
30. G. R. Brown, C. A. Winkler, *Angew. Chem.* **82** (1970) 187.
31. F. Notaro: "Advances in ambient temperature air separation", Munich Meeting on Air Separation Technology, Munich, Oct. 10 – 11, 1996.
32. W. H. Isalski: *Separation of Gases,* Clarendon Press, Oxford 1989.
33. *Principle of air separation*, Company brochure, Linde AG Process Engineering and Contracting Division, Höllriegelskreuth 1998.
34. H. Springmann, *Linde Reports Sci. Tech.* **28** (1978) 23.
35. B. A. Hands (ed.): *Cryogenic Engineering,* Academic Press, London 1986.
36. D. Gary, *Sep. Purif. Tech.* **11** (1977) 47.
37. M. Murakami, M. Nomura, *Gas Sep. Purif.* **2** (1988) 95.
38. H. Cheung, J. H. Royal, "Cryogenic Processes and Machinery," *AIChE Sympos. Ser.* **89** (1993) no. 294, 71.
39. D. Gary, *Sep. Purif. Technol.* **11** (1997) 47.
40. K. Knoblauch, *Chem. Eng.* (1978) Nov. 6, 87.
41. H. Jüntgen, K. Knoblauch, K.-B. Harder, *Fuel* **60** (1981) 817.
42. K. Knoblauch, *Gas Wärme Int.* **32** (1983) 78.
43. Unpublished experimental results, Linde AG, Höllriegelskreuth, 1988.
44. A. K. Fritzsche, R. S. Narayan, *CEER Chem. Econ. Eng. Rev.* **19** (1987) no. 1, 2, 3, 19.
45. D. R. Paul, Y. P. Yampolskii (eds.): *Polymeric Gas Separation Membranes,* CRC Press, Boca Raton, Fla., 1994.
46. S. T. Hwang, C. K. Choi, K. Kammermeyer, *Sep. Sci.* **9** (1974) 461.
47. M. Langsam, M. Anand, E. J. Karwack, *Gas. Sep. Purif.* **2** (1988) 162.
48. P. S. Puri, *Gas Sep. Purif.* **4** (1990) 29.
49. R. Prasad, F. Notaro, D. R. Thomson, *J. Membr. Sci.* **94** (1994) 225.
50. US 5 091 216, 1992 (O. M. Ekiner, R. A. Hayes, P. Manos).
51. F. Notaro: "Advances in ambient temperature air separation", Munich Meeting on Air Separation Technology, Oct. 10 – 11, 1996, p. 227.
52. R. Jain, *Gas. Sep. Purif.* **3** (1989) 123.
53. LINDE International Report, 1992.
54. Compressed Gas Association: *Handbook of Compressed Gases,* 3rd ed., Van Nonstrand Reinhold, New York 1990.
55. Kirk Othmer, 4th ed., **17**, p. 153.
56. H. Cheung, J. H. Royal, *AIChE Sympos. Ser.* **89** (1993) no. 294, 71 – 79.
57. J. B. Goddard, M. L. Malczewski, *AIChE Sympos. Ser.* **89** (1993) no. 294, 62 – 70.
58. H. J. Schröter, H. Jüntgen in A. E. Rodriguez, M. D. LeVan, D. Tondeur (eds.): *Adsorption: Science and Technology,* Kluwer Academic Publishers, Dordrecht 1989. NATO ASI Ser, Ser **E** 158 (1989) 269 – 283.
59. CALORIC Apparatebau, Company Broschure, Gräfelfing b. München, 1976.
60. Zimmer, US 3 215 503, 1965, (F. Nessler).
61. B. Kolb, M. Linder, B. Kempken, *Angew. Chromatogr.* **21** (1975).
62. H. G. Fitzky, D. Wendisch, R. Holm, *Angew. Chem.* **84** (1972) 1037.
63. R. G. Bowes, *Metallurgia* **46** (1979) March, 163.
64. C. Rademacher, *Gas Aktuell (Messer Griesheim)* **12** (1976) 9.
65. *Gemeinfaßliche Darstellung des Eisenhüttenwesens,* 17th ed., Verlag Stahleisen, Düsseldorf 1971.
66. V. P. Nemchenko et al., *Steel USSR* 1976, Oct., 547.
67. G. Böhm, *Techn. Mitt.* **69** (1976) 614.
68. G. Verannemann, *Linde Ber. Wiss. Tech.* **47** (1980) 47.
69. M. Digerens, *Concast News* **15** (1976) 3.
70. G. Leconte, K. Buxmann, *Aluminium (Düsseldorf)* **55** (1979) 387.
71. P. Stratton, *BOC Technol. Magazine* (1997) no. 6, 29.
72. P. Blakey, G. Orlando, *Chem. Eng.* (1984) May 28, 97.
73. G. Verannemann, *Linde Ber. Wiss. Tech.* **44** (1978) 3.
74. W. Bartknecht: *Explosionen,* Springer Verlag, Berlin 1978.
75. R. Hirst, K. Booth, *Fire Technol.* **13** (1977) 296.
76. *VDI-Nachr.* **32** no. 28 (1978) 7.
77. *Gefährliche Ladung* **24** (1979) no. 7, 37.

78. G. Holocomb, S. Laschober, *Hydrocarbon Process.* **58** (1979) no. 9, 163.
79. A. Ramaswamy, P. S. Katiyar, *Mine Safety* **36** (1988) 415.
80. A. Fleig, *Lebensmitteltechnik* **24** (1992) no. 5, 14.
81. W. Hall, *Food Rev.* **12** (1985) no. 2, 15.
82. G. Dinglinger, *Die Kälte u. Klimatechnik* **1982,** no. 4, 142.
83. N. Kuzniarski, *BOC Technol. Magazine* (1987) no. 7, 30.
84. J. Wikstrom, *CryoGas Int.* **31** (1993) no. 8, 16.
85. D. Rebhahn, *Industrieanzeiger* **7** (1996) 24.
86. D. Rebhahn, *Kautschuk-Gummi-Kunststoffe* **48** (1995) 34.
87. J. Wary, R. B. Davies, *Chemtech* 1976, March, 200.
88. Proc. European Symposium Particle Technology 1980, DECHEMA Frankfurt/M 1980, pp. 110 – 136.
89. A. Baumann, P. Belger, W. Duesberg, *Gas aktuell (Messer Griesheim)* **56** (1999) 4.
90. Company brochure, Linde AG, Werksgruppe TG, Höllriegelskreuth 1996.
91. W. Biddulph, *Chem. Eng.* **1980,** Feb. 11, 93.
92. M. Corbett, *BOC Technol. Magazine* (1992) 17, 53.
93. *European Plastic News* **5** (1991) 66.
94. H. Kuhn, *Elektronik Produktion u. Prüftechnik* **1994,** no. 3, 40.
95. K. McKean, M. Mapes, F. Rotman, *Proceedings Technical Program-National Electronic Packaging and Production Conference* **V 3** (1995) 1295.
96. P. D. Allan, *Proceedings SPE Annual Technical Conference, New Orleans, USA 1994,* Society of Petroleum Engineers, Richardson, TX, pp. 291 – 295.
97. A. Kanzas, I. Chatzis, F. A. L. Dullien, *Proc. SPE/DOE Sixth Symposium on Enhanced Oil Recovery,* Tulsa, OK, 1998.
98. R. F. Prahade, J. H. Ziemer, B. C. Price, Conference Paper, AIChE *National Meeting,* New Orleans, LA, 1988.
99. P. Mazur, *Science* **168** (1970) 939.
100. B. J. Fuller, B. Grout (eds.): *Clinical Applications of Cryobiology,* CRC Press, Boca Raton 1991.
101. A. Donnerhack, K. Pfeil-Schneider, K. Thoma, *Gas aktuell (Messer Griesheim)* **41** (1991) 16.
102. J. Buchmüller, S. Zerwas, *Chem. Ing. Tech.* **65** (1993) 553.
103. *Cryogenics* **26** (1986) 572.
104. B. Marx, *Gas aktuell (Messer Griesheim)* **41** (1991) 23 – 63.
105. M. Biederbick, R. Schaub, *Techn. Überw.* **29** (1988) 182.
106. *Umwelt Magazin (Ort)* 1988 (March) 100.
107. K. A. Kuberka, *Elastomerics* **7** (1990) 16.
108. D. Rebhahn, *Tiefbau-Ingenieurbau Straßenbau* **5** (1988) 250.
109. G. Holocomb, S. Laschober, *Hydrocarbon Process.* **58** (1979) no. 9, 163.
110. F. Schlömer, *Verfahrenstechnik* **24** (1990) no. 7/8, 12.
111. *Verfahrenstechnik* **27** (1993) no. 7/8, 34.
112. *Gas aktuell (Messer Griesheim)* **55** (1998) 31.
113. B. Heydorn, *CryoGas Int.* 1993, Dec., 10.
114. R. M. Hall, *Metall (Berlin)* **31** (1977) 43.

Nitrogen Fixation

J. Elaine Barclay, AFRC Institute of Plant Science Research Nitrogen Fixation Laboratory, University of Sussex, Brighton, BN1 9RQ United Kingdom

G. Jeffery Leigh, AFRC Institute of Plant Science Research Nitrogen Fixation Laboratory, University of Sussex, Brighton, BN1 9RQ United Kingdom

1. Introduction 1
2. Biological Nitrogen Fixation 1
2.1. Microorganisms 1
2.2. Nitrogenases 2
2.3. Reaction Stoichiometry and Mechanism 2
2.4. Genetic Control of Nitrogenases 3
3. Reactions of Dinitrogen 4
3.1. Reaction with Metals and Metal Complexes 4
3.2. Activation of Dinitrogen for Ammonia Formation 4
3.3. Structures of Dinitrogen Complexes . 4
4. Protonation and Alkylation of Dinitrogen Complexes 5
4.1. End-on Dinitrogen Complexes 5
4.2. Bridging Dinitrogen Complexes 6
5. Mechanisms of Reactions of Coordinated Dinitrogen 7
6. Fixing Systems Involving Chemical Reduction of Dinitrogen Before N–H Bond Formation 8
6.1. Nitriding Systems 8
6.2. Cyclopentadienyltitanium Systems . . 9
7. Nitrogen Fixation in Protic Solvents . 10
8. Other Aqueous Fixation Systems . . . 11
8.1. Approaches to Models for Nitrogenases 11
8.2. Mechanisms 12
9. Fixation of Nitrogen Involving Metal Clusters 12
10. Photochemical Fixation Processes . . 12
11. Oxidative Fixation Processes 13
12. References 13

1. Introduction

Since 1941, industrial nitrogen fixation has been carried out almost entirely by the Haber process (→ Ammonia). The Norwegian arc process was last exploited commercially at that time, and other processes, such as the cyanamide process, ceased to be significant even earlier. Apart from various minor efforts to revive the arc process to operate under rather special, restricted circumstances (see Chap. 11), hopes for new nitrogen-fixation processes have stemmed from the observation that biological nitrogen fixation takes place under mild conditions in the presence of dioxygen and water, and that it should therefore be possible to reproduce such processes in a completely synthetic system in vitro. The requirement for molybdenum or vanadium first demonstrated in the 1930s also supported the belief that the biological processes involve metalloenzymes [1]. This prompted research, successful first in 1965, into the synthesis of dinitrogen complexes [2]. The two related but different areas of research, biological and chemical, have been the subject of increasing effort since ca. 1960. The first indications of application of this research to industrial nitrogen-fixation processes are only now becoming evident.

Some of the material discussed here is also dealt with briefly in an earlier article, see → Ammonia, Chap. 4.7.

2. Biological Nitrogen Fixation

2.1. Microorganisms

All known biological nitrogen-fixation processes are performed by microorganisms. Some, such as the rhizobia, fix nitrogen in obligate symbiotic relationship with plants, principally but not exclusively with legumes, but other non-obligate symbioses–for example, *Azotobacter paspalum* with a tropical grass–seem to occur accidentally. The basis of the symbiosis appears to be the exchange of carbohydrate for fixed nitrogen. Other, non-symbiotic systems include

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_471

anaerobic and aerobic bacteria (e.g., *Clostridium pasteurianum* and *Azotobacter vinelandii*, respectively), blue-green algae or cyanobacteria (e.g., *Anabaena cylindrica*), and phototrophic bacteria (e.g., *Rhodospirillum rubrum*).

In certain parts of the world, nitrogen-fixing organisms living on the leaves of the water fern azolla have been used as a "green fertilizer," for example, by ploughing the ferns into the soil. However, the aim of recent research is the construction of nitrogen-fixing plants by incorporation of appropriate bacterial DNA into the plant genome rather than the exploitation of natural nitrogen-fixing systems [3].

2.2. Nitrogenases

It is recognized that there are at least three varieties of nitrogenase, possibly all expressed in nature under appropriate conditions [4]. Until recently, the existence of only a molybdenum-containing nitrogenase was recognized, but this may be a consequence of artificial selection forced by culture conditions rather than an expression of the state of affairs in nature. All molybdenum nitrogenases are very similar, whatever their source. The recognized vanadium nitrogenases are very similar to the molybdenum nitrogenases [5], but the third nitrogenase seems to differ in some respects because its component proteins do not form active hybrid nitrogenases with proteins from the other two nitrogenases [6].

The molybdenum nitrogenases consist of two air-sensitive metalloproteins. The oxygen sensitivity means that organisms which use nitrogenases have developed techniques to separate nitrogen fixation either temporally or spatially from activities that involve oxygen, e.g., photosynthesis and respiration. The larger protein (the molybdenum – iron protein) contains 2 molybdenum atoms, ca. 25 iron atoms, and an equivalent amount of "inorganic" sulfide and has a molecular mass of ca. 230 000. Because the protein comprises four subunits, two each of two kinds, its constitution is designated $\alpha_2\beta_2$. An iron – molybdenum cofactor (FeMoco), which contains no peptide residues and has inorganic constituents with approximate stoichiometry Fe_6MoS_4, has been extracted from this protein. Some of the iron atoms are apparently contained in Fe_4S_4 clusters. The smaller, iron protein (M_r ca. 60 000) has two subunits (δ_2), and contains four iron atoms in a single Fe_4S_4 cluster [3].

The vanadium nitrogenase is in most respects similar except that vanadium replaces molybdenum; it possesses an iron – vanadium cofactor, FeVaco.

2.3. Reaction Stoichiometry and Mechanism

To obtain fixation in vitro, the two proteins are mixed, generally with considerable excess of the smaller, together with a reducing agent (normally sodium dithionite), adenosine 5′-triphosphate (ATP, an energy source), a buffer, and dinitrogen (nitrogenase reduces many substrates, most notably acetylene; for details, see [3]). The ability to reduce acetylene to ethylene has long been regarded as characteristic of nitrogen-fixers, and it has even been used diagnostically because of the relative ease of testing for it. The reduction of dinitrogen to ammonia by a molybdenum nitrogenase is believed to be most efficient in use of electrons when its stoichiometry corresponds to the reaction

$$N_2 + 8\,H^+ + 16\,ATP + 8\,e^- \longrightarrow 2\,NH_3 + H_2 + 16\,ADP + P^i$$

where P^i represents inorganic phosphate [3].

Evolution of dihydrogen always occurs and, at its minimum, represents 25 % of the electron flux. The evolution of dihydrogen is believed to be associated with its displacement by dinitrogen at a polyhydridic metal site, presumably molybdenum. Although precedents for this kind of displacement exist, it is not clear why nitrogenase should adopt it.

$$[CoH_3(PPh_3)_3] + N_2 \rightleftharpoons [CoH(N_2)(PPh_3)_3] + H_2$$

The stoichiometry of dinitrogen reduction by vanadium nitrogenase suggests that 50 % of the electron flux is used in dihydrogen evolution. If this is a minimum loss of electrons, a very different mechanism is indicated for vanadium-mediated reduction compared to molybdenum-mediated reduction.

The pre-steady-state kinetics of dinitrogen reduction by *Klebsiella pneumoniae* wild-type

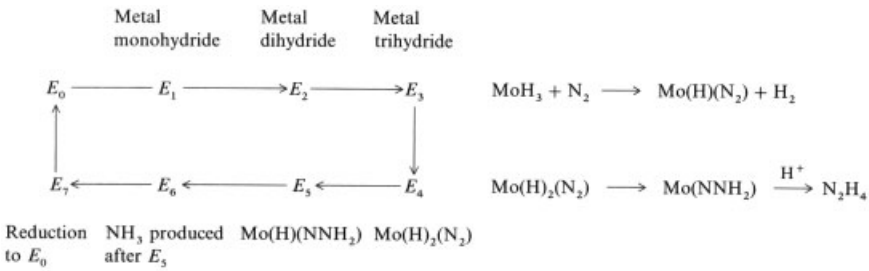

Figure 1. Simplified scheme for the function of nitrogenase
$E_0, \ldots, E_n$ represent different states of the enzyme system, each arising as one electron (of the total of eight electrons for the complete cycle) is taken up. For each state, the intermediate metal species or the reaction believed to occur is indicated. These are inferred from chemical studies.

nitrogenase have been studied by stopped-flow spectrophotometry, and the mechanistic scheme in Figure 1 has been proposed to accommodate the data [7]. In Figure 1, E_0, E_1, ..., E_7 represent eight different states of the enzyme system corresponding to the eight individual electron transfers of the stoichiometric reaction described above. For the transfer of a single electron ($E_n \rightarrow \mathrm{E}_{n+1}$), the smaller, iron protein is believed to accept an electron from the reducing agent and then form a "complex" with the larger protein, to which it transfers the electron, and from which it finally dissociates before accepting another electron. Dinitrogen binds to the larger protein, and accepts electrons stepwise while simultaneously taking up protons from solution. The proposed identity of the dinitrogen or hydride species involved in each stage is indicated in Figure 1; further details are given in [8,9]. None of this analysis can reveal the dinitrogen pattern of reactivity at an atomic level, which requires detailed chemical studies.

2.4. Genetic Control of Nitrogenases

The successful expression of bacterial nitrogen fixation activity in a plant would seem to require a good understanding of the bacterial genome. The 20 or so genes required for the synthesis, control, and function of the molybdenum nitrogenase in *Klebsiella pneumoniae* are shown in Figure 2; they are contiguous and are found in a single cluster. In the azotobacters, in contrast, the genes are scattered throughout the chromosome, with gaps of as in 1984 unidentified function [10]. Azotobacter chromosomes also contain apparently duplicate genes which are concerned with the vanadium and third nitrogenases [11]. There are also some additional regulatory genes, and in the rhizobia, for example, genes concerned with infection. However, there seems a remarkable degree of conservation between genes fulfilling the same function, whatever their origin.

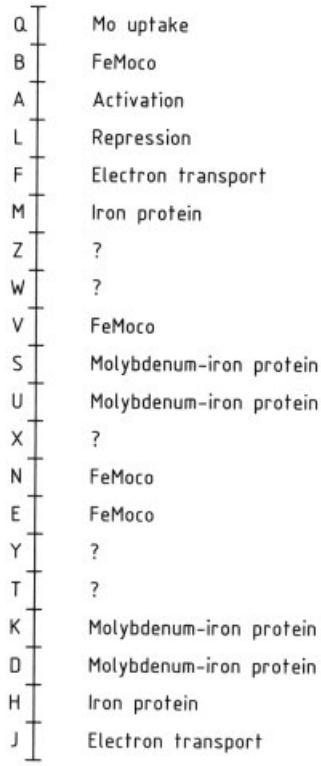

Figure 2. The linear arrangement of the nif gene cluster of *Klebsiella pneumoniae*
The letters Q, B, A, L, etc. are internationally agreed gene designations and have no further significance. The proposed functions of the gene products are shown in the right-hand column.

Attempts are being made to express the products of some of these genes, such as nifH, in plant chloroplasts as a preliminary to attempting the transfer and expression of the whole genome. Success in this endeavor would be a great technical achievement. Its value in agricultural terms, however, would remain to be proven.

3. Reactions of Dinitrogen

3.1. Reaction with Metals and Metal Complexes

Dinitrogen interacts with most transition elements, but the intensity of interaction varies between the extremes of weak physisorption (detectable only under very specific conditions) to strong chemical interaction resulting in strong metal – nitrogen binding and breaking of the N – N bond, i.e., nitride formation; Figure 3 summarizes the reactions. A rough division, indicated by the stepped line, can be made between the strong interactions resulting in nitride formation and weak interactions which allow the dinitrogen molecule to remain intact. In contrast, dinitrogen complexes stable at room temperature are formed by almost all transition elements, given an appropriate arrangement of co-ligands (Fig. 4).

3.2. Activation of Dinitrogen for Ammonia Formation

A major emphasis in nitrogen fixation research has been to find conditions under which coordinated dinitrogen can be converted to ammonia. However, strongly reduced systems (analogous to the nitride-forming metals) also exist which, under anhydrous conditions, can apparently reduce and even split dinitrogen molecules. Addition of protons then yields ammonia directly. In addition, there is a third major area of fixation research in which conversion of dinitrogen to ammonia is observed. These have been referred to as "soups," and the involvement of dinitrogen complexes is inferred from kinetic and spectral measurements (see Chaps. 7 and 8).

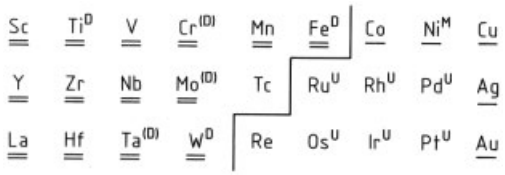

Figure 3. Room temperature adsorption behaviour of dinitrogen on various metals
Symbols doubly underlined indicate exothermic nitrides; singly underlined indicate endothermic nitrides. The thick line is the borderline between dissociative and molecular adsorption at room temperature.
U = Binary nitride unknown; M = molecular adsorption; D = dissociative adsorption; (D) = probable dissociative adsorption

3.3. Structures of Dinitrogen Complexes

Considerable structural variety exists among dinitrogen complexes [8, 9]. The largest group of dinitrogen complexes contains one, occasionally two, and, in one case, three dinitrogen molecules bound linearly end-on to a metal atom, with a nitrogen – nitrogen separation of ca. 112 pm, little longer than in free dinitrogen. Some of these complexes, generally those of transition elements of groups 4, 5, and 6, can be converted to ammonia and/or hydrazine by the action of protic acids. However, dinitrogen complexes with elements of groups 7 – 10 have similar geometries and dimensions, but little or no tendency to undergo protonation. Clearly, geometry is no criterion of activation.

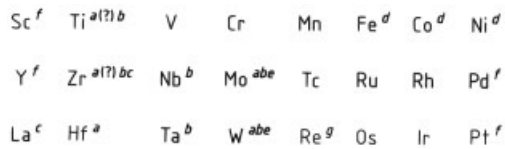

Figure 4. Binding and activation of coordinated dinitrogen
[a] Terminal, end-bound N_2, activated to protonation, yielding ammonia and/or hydrazine.
[b] Bridging, end-bound N_2, activated to protonation, yielding hydrazine. In some cases, it is not clear whether end-on or side-on dinitrogen (or both) is activated.
[c] Side-on N_2 activated to protonation, yielding hydrazine (observed for a single Sm complex).
[d] Other activated N_2; in these systems additional reductant is required.
[e] Terminal end-bound N_2, activated to alkylation and acetylation, yielding diazenido complex.
[f] All the elements shown in this figure form dinitrogen complexes stable at 20 °C except for these, for which stable dinitrogen complexes have not yet been unequivocally characterized.
[g] Acetylation of coordinated N_2 observed.

These complexes contain dinitrogen that can function as a base and is polarized in the sense $M-N^{\delta+}-N^{\delta-}$. As Lewis bases they form adducts with Lewis acids (including, often, a proton). Such bridging dinitrogen adducts rarely contain activated dinitrogen. In a second group of complexes containing bridging dinitrogen, bond length considerations imply that the dinitrogen may be formally regarded as $(N{=}N)^{2-}$ or even $(N-N)^{4-}$. Electron counting generally leans towards the latter formulation. These complexes are generally homonuclear and contain transition elements of groups 5 and 6. They react with proton acids to yield hydrazine, consistent with the presumptive reduced nature of the dinitrogen. As routes from molecular dinitrogen to hydrazine, they are, however, not very promising, since syntheses directly from molecular nitrogen generally give low yields, and alternative syntheses based on sources of fixed nitrogen are recommended for most purposes.

A number of other structural types of dinitrogen complex exist. These are often uncommon, single examples and are not discussed here. For a general review, see [8, 9].

4. Protonation and Alkylation of Dinitrogen Complexes

4.1. End-on Dinitrogen Complexes

Although reactions of dinitrogen have been known for many years, such reactions have involved extreme conditions or highly reduced chemical species that are generally unstable in the presence of water.

The first dinitrogen complex, $[Ru(NH_3)_5(N_2)]^{2+}$, was reported in 1965 [2]. Though unreactive toward protonation, it opened the way to synthesis of many other dinitrogen complexes.

Protonation Reactions. The first reported protonation of a dinitrogen complex was of $[W(N_2)_2(dppe)_2]$ to form $[WBr(NNH_2)(dppe)_2]^+$ ($dppe = Ph_2PCH_2CH_2PPh_2$) [12]. Shortly afterwards, it was shown that complexes of the type $[M(N_2)_2P_4]$ (M = Mo or W; P = tertiary phosphine) can be protonated directly to ammonia when P is a tertiary monophosphine [13, 14].

The highest yield of ammonia (90 % based upon $W^0 \rightarrow W^{VI}$) is obtained from *cis*-$[W(N_2)_2(PMe_2Ph)_4]$ and sulfuric acid in methanol at room temperature. One mole of dinitrogen is evolved rapidly, together with a trace of dihydrogen, while the second mole of dinitrogen yields ammonia and a small quantity of hydrazine. Typical data are listed in Table 1. The reasons for the wide variation in the amounts of products are not known. Hydrazido(2−) (NNH_2^{2-}) species have been isolated from many of these systems, and hydrazidium(1−) (NNH_3^-) complexes have also been characterized [15].

Table 1. Yields of ammonia and hydrazine from the reaction of $[M(N_2)_2P_4]$ ($P = PMe_2Ph$ or $PMePh_2$) with sulfuric acid at room temperature [14]

Complex	Solvent	NH_3, % [a]	N_2H_4, % [a]	N_2, % [b]
cis-$[W(N_2)_2(PMe_2Ph)_4]$	MeOH	90	2	94
cis-$[W(N_2)_2(PMe_2Ph)_4]$	THF	46	13	59
trans-$[W(N_2)_2(PMePh_2)_4]$	THF	36	15	54
cis-$[Mo(N_2)_2(PMe_2Ph)_4]$	MeOH	36	< 1	40
cis-$[Mo(N_2)_2(PMe_2Ph)_4]$	THF	35	< 1	34
trans-$[Mo(N_2)_2(PMePh_2)_4]$	THF	4	< 1	ca. 5

[a] Based upon six electrons available from each metal atom.
[b] Based upon one dinitrogen molecule expected from each metal atom.

Stepwise protonation of coordinated dinitrogen, as shown below, has been proposed as a basis for understanding these reactions.

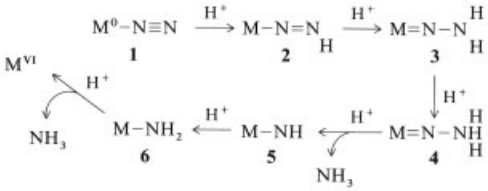

Although all the species **1**–**6** have been characterized, only **1**–**4** have been isolated from protonation reactions. For discussion of protonation mechanisms, see Chapter 5.

It is likely that a similar mechanism holds for the protonation of dinitrogen in $[Mo(N_2)_2(triphos)(PPh_3)]$ ($triphos = Ph_2PCH_2CH_2PPhCH_2CH_2PPh_2$), at least in the initial stages. The final stages may well involve a disproportionation [16].

$$2\ \textit{trans}\text{-}[Mo(N_2)_2(triphos)(PPh_3)] \xrightarrow{4\,HBr} 2\,N_2 + 2[MoBr(NNH_2)(triphos)(PPh_3)]Br$$
$$\xrightarrow{4\,HBr} 2[MoBr_3(triphos)] + 2[PPh_3H]Br + N_2 + 2\,NH_3$$

Table 2. Production of ammonia by an electrochemical cycle [20] ∗

Complex	Amount of complex, µmol	Amount of NH_3, µmol	Yield of NH_3, mol % per mole of complex
$[W(NNH_2)(O_3SC_6H_4Me)(dppe)_2]^+$	145	33	23
$[W(NNH_2)(O_3SC_6H_4Me)(dppe)_2]^+$	341	99	29

∗ Conditions: 0.2 mol/dm^3 $[NBu^{n4}][BF_4]$ in THF (30 cm^3); Hg pool cathode at − 2.6 V vs. ferrocinium/ferrocene; cathode compartment purged with N_2.

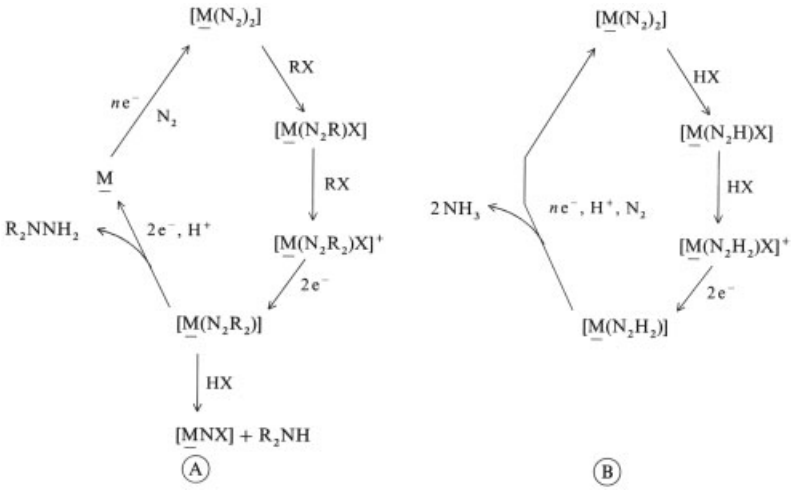

Figure 5. Cyclic electrochemical systems for producing ammonia, amines, and hydrazines [$\underline{M}$ = M(dppe)$_2$, M = Mo or W, R = alkyl] [20]
(A) X = Br or I; (B) X = 4-$CH_3C_6H_4SO_3$

Alkylation Reactions. Dinitrogen in $[M(N_2)_2(diphosphine)_2]$ can also be alkylated and silylated in high yields. Alkylation reactions (for mechanism, see Chap. 5) yield products that formally resemble the protonated species M−NNH and M=NNH_2 discussed above [17]. Typical examples are $[MoBr(NNMe)(dppe)_2]$ and $[WBr(NNMe_2)(dppe)_2]Br$. Such compounds can be destructively reduced (for example, by lithium aluminum hydride) to give the corresponding amine and ammonia [18].

Cyclic Electrochemical Systems. More interesting is the controlled two-electron chemical or electrochemical reduction of the dialkylhydrazide to yield, for example $[W(NNMe_2)(dppe)_2]$ [19]. Treatment of these species with acid results in the cleavage of the N−N bond, as shown in Figure 5 A. Normally, electrochemical reduction of the NNH_2^{2-} analogues of the alkylhydrazides yields dihydrogen, but in the specific case when X is 4-$CH_3C_6H_4SO_3$ the system undergoes reduction in a manner similar to the alkylated systems (Fig. 5 B) and thus constitutes a cyclic system to generate ammonia from dinitrogen. Typical data are given in Table 2.

4.2. Bridging Dinitrogen Complexes

Much less information is available on protonation of dinitrogen-bridged compounds. Complexes containing formal N_2^{2-} or N_2^{4-} (e.g., $[\{Ta(CHBu^t)(CH_2Bu^t)(PMe_3)_2\}_2N_2]$) give hydrazine in high yield on treatment with protic acids [21]. The N−N bond length here is ca. 120 pm, intermediate between a triple and a single bond. However, $[\{Zr(C_5Me_5)_2N_2\}_2N_2]$ also yields hydrazine and (with HCl) $[Zr(C_5Me_5)_2Cl_2]$, although the N−N bond length is only 118.2 pm [22]. In the latter case, poorly understood bridging – terminal dinitrogen exchange reactions occur and the system requires further research.

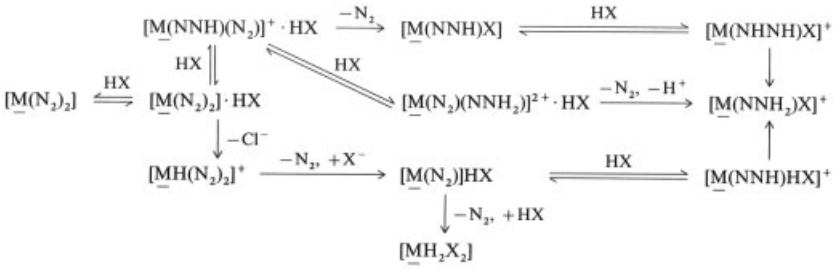

Figure 6. Mechanism of protonation of coordinated dinitrogen in complexes $[M(N_2)_2(\text{diphosphine})_2]$ by acids HX (X = Cl or Br) in tetrahydrofuran [24]

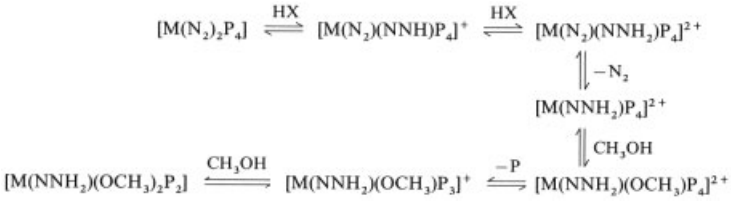

Figure 7. Mechanism of protonation of *cis*-$[M(N_2)_2(PMe_2Ph)_4]$ (M = Mo or W) by acids HX (X = Cl, Br, or HSO_4^-) in methanol [25]

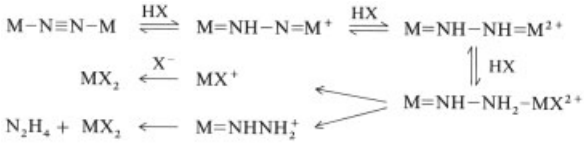

Figure 8. Mechanism of protonation of bridging dinitrogen in $[\{M(S_2CNEt_2)_3\}_2N_2]$ (M = Nb or Ta) by acids HX (X = Cl or Br) in acetonitrile [26]

Protonation of Complexes Yielding Ammonia. Many of the less common and highly reduced dinitrogen complexes containing bridging or side-on dinitrogen also protonate at dinitrogen. For example, protonation of $[\mu_3\text{-}N_2\{(\eta^5,\eta^5\text{-}C_{10}H_8)(\eta^5\text{-}C_5H_5)_2Ti_2\}\{(\eta^1,\eta^5\text{-}C_5H_4)(\eta^5\text{-}C_5H_5)_3Ti_2\}]\cdot[(\eta^5\text{-}C_5H_5)_2Ti(C_6H_4O_3)]$, which has an N–N bond length of 130.1 (12) pm, in aqueous ether gives ammonia (> 90 % yield) and some hydrazine [23].

5. Mechanisms of Reactions of Coordinated Dinitrogen

Protonation Reactions. The mechanisms of protonation and alkylation of coordinated dinitrogen have been investigated in detail, principally by kinetic studies. The reaction schemes outlined in Figures 6 and 7 summarize the conclusions and refer generally to the reactions discussed in Chapter 4. The reader is referred to the original publications for more details of the methods and of the analysis [24–27]. The actual route taken in any given case depends on the diphosphine, the acid, and the solvent. Although these reactions stop normally at the hydrazide(2−) (NNH_2^{2-}) stage, NNH_3^- has also been recognized in two cases.

The protonation of dinitrogen complexes with at least one tertiary monophosphine as ligand can yield ammonia. Only the initial stages have been kinetically established (Fig. 6). The solvent is an integral part of this mechanism, although this may not be evident in the final products.

The sole kinetic study on the protonation of bridging dinitrogen involve a species containing a "reduced" form of dinitrogen (Fig. 8).

Alkylation Reactions. Alkylation reactions of dinitrogen complexes have not been success-

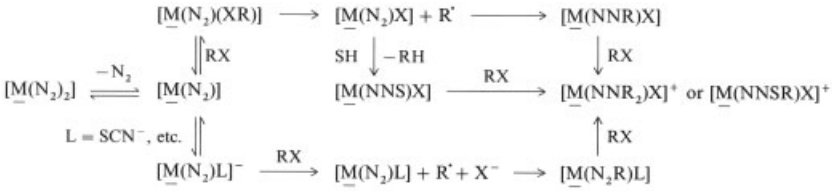

Figure 9. Alkylation of coordinated dinitrogen in $[M(N_2)_2(\text{diphosphine})_2]$ (M = Mo or W) by alkyl halides RX (X = Br or I) in various solvents (SH) [8,27]

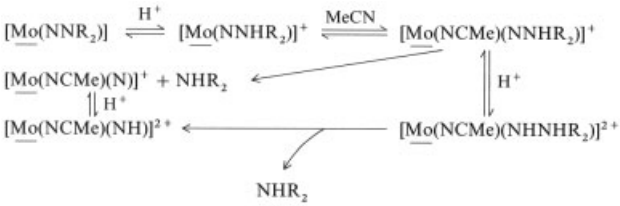

Figure 10. Protonation of the five-coordinate hydrazide species $[Mo(NNR_2)(dppe)_2]$ by HBr in acetonitrile [27]

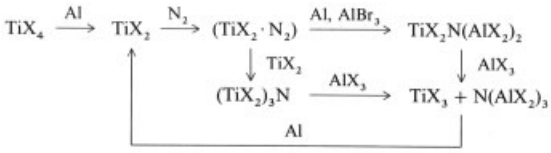

Figure 11. Proposed scheme for the catalytic nitridation of aluminum (X = Cl, Br, or OBu) [31]

ful when the ligands are all tertiary monophosphines. The free-radical scheme (Fig. 9) applies to complexes with two tertiary diphosphines and is the only mechanistic scheme that is pertinent, with the possible exception of two examples: one of nucleophilic attack and one of electrophilic attack on coordinated dinitrogen. The reactions with alkyl halides that convert the diazenide (NNR^-) to hydrazide (NNR_2^-) probably involve nucleophilic attack (S_N2) of the exo-nitrogen on the α-carbon atom of the alkyl group.

Splitting of the N−N bond has been studied kinetically only in the case of dialkylhydrazide complexes, though essentially the same mechanism is believed to hold in the case of the species involved in the electrochemical cycle (see Section 4.1) (Fig. 10). The intermediacy of the hydrazidium ($NNHR_2^-$) species is particularly noteworthy, because it may prove to be a general feature of N−N bond-breaking mechanisms [27].

The arylation of coordinated dinitrogen to give an aryldiazenido complex is restricted at present to a single example [28].

6. Fixing Systems Involving Chemical Reduction of Dinitrogen Before N−H Bond Formation

6.1. Nitriding Systems

Typically, these nitrogen-fixing systems consist of a transition metal compound that binds and reduces dinitrogen, along with a strong reducing agent. They are normally oxygen- and water-sensitive, and release ammonia and/or hydrazine upon addition of acids or protic solvents. Work in this area was stimulated by Vol'pin, whose publication in 1964 was the forerunner of all modern work on nitrogen fixation [29]. His initial systems were $CrCl_3/LiAlH_4/N_2$ at 15 MPa, which gave 0.02 mol of ammonia per mole of $CrCl_3$, and $CrCl_3/EtMgBr/N_2$ at 15 MPa, which gave 0.17 mol of ammonia per mole of $CrCl_3$. Since then, many metal halides have been used, some of which yield appreciable quantities of dinitrogen at even 101.3 kPa nitrogen pressure. Typical data for such systems, often referred to, perhaps inaccurately, as nitriding systems, are presented

Table 3. Fixation of nitrogen by typical nitriding systems [30]

System	Molar ratio of complex to reductant	Solvent	N_2 Pressure, MPa	Reaction time, h	Yield of NH_3, mol/mol complex
$[Ti(C_5H_5)_2Cl_2]$/EtMgBr	9	Et_2O	15	31	0.93
$[Ti(C_5H_5)_2Cl_2]$/Li(Hg)	10	THF	0.1	72	0.70
$[Ti(C_5H_5)_2Cl_2]$/Li($C_{10}H_8$)	6	THF	0.1	16	0.96
$[Ti(C_5H_5)Cl_3]$/Mg	2	THF	0.1	24	0.46
$TiCl_4$/Na	excess	THF	0.1	144	0.44
$CrCl_3$/Pr^nMgBr	9	Et_2O	15	11	0.30
$MoCl_5$/$LiAlH_4$	9	Et_2O	15	12	3.03
WCl_6/EtMgBr	9	Et_2O	15	11	0.15
$FeCl_3$/Mg,MgI_2	14 : 2.5	$Et_2O-C_6H_6$	8 – 10	7	0.26
$[VO(acac)_2]$/$LiBu^n$	9	$n-C_7H_{16}$	7.5	10	0.32
$Ti(OPr^i)_4$/Na($C_{10}H_8$)		THF	0.1	0.5	1.3

in Table 3 [30]. An electric current can also be used as a reducing agent [31].

Although the necessary final hydrolysis destroys a nitriding system and renders it noncyclable, the catalytic nitridation of aluminum has been achieved with a mixture of a titanium(IV) compound, aluminum, and aluminum tribromide in benzene (see Fig. 11 and Table 3) [30]. Hydrolysis of the resultant product releases ammonia (Table 4).

Table 4. Yields of ammonia via catalytic nitridation of aluminum in the system $TiCl_4$/Al/$AlBr_3$, $p_{N_2} \approx 10$ MPa [31]

Molar ratio $TiCl_4$: Al : $AlBr_3$	Reaction time, h	NH_3, mol per atom of Ti	NH_3, % *
1 : 12 : 33	14	11	92
1 : 150 : 200	30	95	64
1 : 600 : 1000	30	286	48

* With respect to reducing agent.

The mechanisms of such reactions have not yet been elucidated in detail. Nevertheless, some indication of what occurs may be given by the characterized reaction products in the nickel – dinitrogen – phenyllithium [32] and cobalt – dinitrogen – potassium systems [33]. In these systems, which yield hydrazine and/or ammonia on hydrolysis, the dinitrogen is bound in a complicated end-on/side-on fashion by both kinds of metal atom; however, no nitrides have been detected.

6.2. Cyclopentadienyltitanium Systems

The most studied systems are those based upon cyclopentadienyltitanium compounds, although even here the reactions are not well understood. Several dinitrogen complexes in this series are known. Compounds $[\{Ti(C_5H_5)_2R\}_2(N_2)]$ (R = certain aryl or alkyl groups) are well characterized [34]. Such compounds might well arise from reduction of $[Ti(C_5H_5)_2Cl_2]$ by a Grignard or alkyllithium reagent under dinitrogen. Although $[\{Ti(C_5H_5)_2N_2\}_2N_2]$ has been established [35], the existence of a related species $[\{Ti(C_5H_5)_2\}_2(N_2)]$ is less certain. Two different routes have been proposed for the reduction of these dinitrogen compounds.

TEUBEN has suggested titanium – cyclopentadienyl ring cleavage [36]:

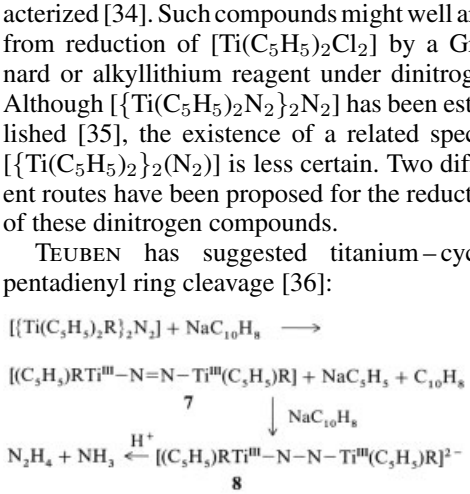

Intermediate **7** is considered to be a precursor of diazene upon hydrolysis, whereas **8** is a hydrazine derivative. With correct choice of reagents, yields of ammonia can reach 1 mol per atom of titanium.

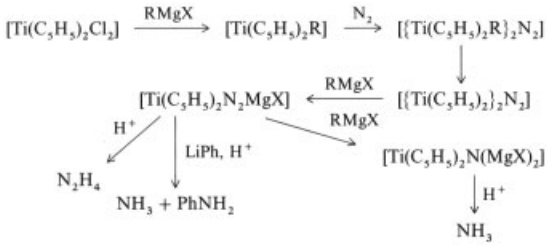

With RMgCl or RLi as reductant (R = Me, Et, or Pr^i), SHILOV [37] envisages a different reac-

Table 5. Some systems that reduce dinitrogen in aqueous methanol ($p_{N_2} \approx 10$ MPa) [38]

Principal metal ion	Reducing agent	Products	Yield [a]	Temperature, °C
Mo^{III}	$Ti(OH)_3$	N_2H_4, NH_3	1	60
Mo^{III}	$Ti(OH)_3 + Mg(OH)_2$	N_2H_4, NH_3	170	110
	$Ti(OH)_3 + Mg(OH)_2$ without Mo	N_2H_4, NH_3	0.005	180
Mo^{III}	$CR(OH)_2$	N_2H_4, NH_3	0.80	90
	$Cr(OH)_2$ without Mo	N_2H_4, NH_3	0.015	90
V^{II}	$V(OH)_2 + Mg(OH)_2$	N_2H_4, NH_3	0.65	20 [b]
V^{II}	$V(OH)_2 + Mg(OH)_2$	N_2H_4, NH_3	0.35	20 [c]
Nb^{III}	$Nb(OH)_3$	N_2H_4, NH_3	0.09	35
Ta^{III}	$Ta(OH)_3$	N_2H_4	0.02	35

[a] Moles per mole of principal metal ion.
[b] In water, pH > 12.5.
[c] In water, pH < 12.

tion sequence in which the Ti−R bond is ruptured and the $Ti(C_5H_5)_2$ moiety remains intact.

Which of these mechanisms is the more likely under any given conditions is not yet clear.

7. Nitrogen Fixation in Protic Solvents

Metal Hydroxide Systems. It is convenient to classify systems that fix nitrogen in aqueous solution by their principal discoverers. By far the largest group of such systems must be attributed to A. E. SHILOV and his collaborators.

Many hydroxo species can interact with dinitrogen at high pH to yield hydrazine and/or ammonia; some of these function at 101.3 kPa of nitrogen pressure. These systems are all heterogeneous, which makes detailed mechanistic analysis difficult.

In a typical experiment, a gel hydroxide of Cr^{II}, V^{II}, Ti^{III}, Nb^{III}, or Ta^{III} is produced by addition of strong base to a solution of an appropriate salt. In the case of Cr^{II} and Ti^{III}, the presence of Mo^{III} increases the yield, but this is not necessary for the other metal ions. The product is pH-dependent, hydrazine predominating over ammonia as the pH increases. In many cases, such as V^{II}, the presence of an excess of Mg^{2+} is required for maximum activity. Yields and products are listed in Table 5 [38].

Homogeneous systems related to the hydroxide gel systems have also been developed. Thus, V^{II} and a polyphenol (by far the best yields are obtained with catechol) reduce dinitrogen to ammonia only in the narrow pH range 9 – 11. Typically, at a nitrogen pressure of 10.13 MPa in aqueous methanol, the yield of ammonia is 0.75 mol per mole of vanadium [39].

Electrochemical Reduction of Dinitrogen. An electrochemical system has been developed starting with Ti^{III}/Mo^{III} and utilizing a mercury cathode as a source of electrons [40]. In the presence of guanidine this is also homogeneous [41]. Under optimum conditions, the current yield with respect to hydrazine reaches 43 %, but only 4 % with respect to ammonia. Note that the solutions used are often very dilute and that the amount of products are therefore often very small (Table 6) even where they indicate high turn-over numbers [38].

Table 6. Electrochemical reduction of dinitrogen [38]

System*	Products	Yield**
$Mo^{III} + Na(Hg) + Mg^{2+}$	N_2H_4	2.5
Mo^{III} + Hg cathode + $Ti(OH)_3$	N_2H_4, NH_3	0.5
Mo^{III} Hg cathode + $Ti(OH)_3$ + guanidine	N_2H_4	0.15

* In aqueous methanol at 20 °C: $p_{N_2} \approx 10$ MPa.
** Moles per mole of molybdeum.

Catalytic Fixation in Aqueous Systems. The electrochemical system shows that genuine catalytic fixation under aqueous conditions should be possible. This has actually been achieved with the $Ti^{III}/Mg^{II}/Mo^{III}$ system at 110 °C (see Table 5). In the presence of sodium amalgam (or a mercury cathode), molybdenum (+ Mg^{II}) can mediate the conversion of dinitrogen to hydrazine and ammonia. The presence of the phospholipid phosphatidylcholine enhances the yield, as does the presence of tertiary phosphines (see Table 7, which includes data for molybdenum nitrogenase for comparison) [42, 43].

Table 7. Catalytic nitrogen fixation in aqueous systems [41,42]

System	Nitrogen pressure, MPa	Specific rate, moles of N_2 reduced per atom of Mo per second	Yield (N_2H_4 + 1/2 NH_3) Per atom of Mo (turnover number)	Per equivalent reducing agent, %
Mo^{III} + Na(Hg)	7	< 0.0001	0.5	0.004
+ Mg^{2+}	7	0.003	2.5	0.02
+ Mg^{2+} + PC∗	7		25	0.25
+ Mg^{2+} + PC∗	0.1	0.05	3.5	0.03
+ R_3P	0.1	0.4	200 – 1000	5 – 10
+ R_3P	7		10 000	30
Nitrogenase	0.1	0.6		

∗ PC = phosphatidylcholine.

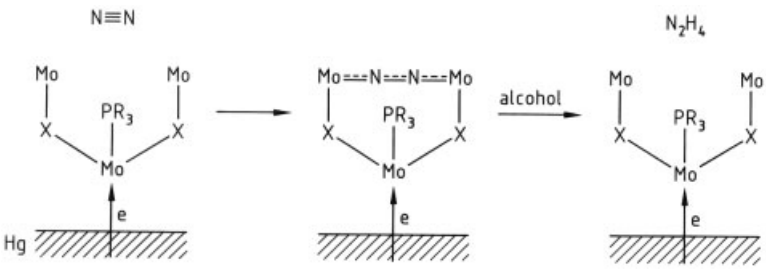

Figure 12. Proposed mechanism for the reduction of dinitrogen by molybdenum – lipid – phosphine systems [37]

Phosphatidylcholine is believed to coat the surface of the amalgam, forming a bridge between the electron donor and the molybdenum catalytic center. It also presumably hinders hydrogen evolution.

SHILOV has discussed the mechanisms of these reactions at some length [38]. Metal ions with three *d*-electrons are believed to be the most favorable for dinitrogen reduction, probably in polynuclear arrangements to facilitate multielectron transfers. A bridging complex of dinitrogen is envisaged as the key intermediate. These ideas are supported by considerable circumstantial physical evidence, but they cannot be regarded as fully proven. The mechanism proposed for these reactions is shown in Figure 12.

8. Other Aqueous Fixation Systems

8.1. Approaches to Models for Nitrogenases

The other considerable body of work on aqueous systems is associated with SCHRAUZER. He showed initially that suitably complexed and buffered molybdenum compounds are capable of mediating the conversion of dinitrogen to ammonia and of ethyne to ethene and ethane, typical reactions of nitrogenase. These studies form part of a series of about 20 papers devoted to the development of a model for nitrogenase function [44].

Molybdocysteine Models. Several molybdenum species have been shown to reduce dinitrogen. One is a derivative of cysteine, formulated as shown below in its reduced form, although this may not be absolutely correct in detail [45].

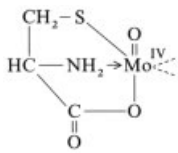

It is believed to bind dinitrogen side-on and to reduce it to the level of diazene as the molybdenum is oxidized. Diazene is released and then disproportionates or is reduced by more reduced

catalyst. A typical suggested stoichiometry for the system is

$$6e^- + 6H^+ + 3N_2 \longrightarrow$$
$$3N_2H_2 + 2H^+ + 2e^- \rightarrow 2N_2 + H_2 \quad / \quad 2NH_3$$

and is consistent with the stoichiometry of dinitrogen reduction by nitrogenase.

Cyano Complexes. A more efficient system starts from $K_4[MoO_2(CN)_4] \cdot 2\,H_2O$ dissolved in an acetate buffer. With a solution containing 42 µmol of this salt in 4 cm^3 of 0.25 % acetate buffer (pH 4.3, 65 °C, nitrogen pressure 101.3 kPa), the yield of ammonia approached 0.91 µmol. The intermediate formation of hydrazine was also observed [46].

Reduced Peptides as Ligands to Molybdenum. Reduced peptides such as bovine insulin can also give rise to efficient fixing systems, and in this case the rate of reduction is claimed to reach 2.5 mol of dinitrogen per atom of molybdenum per minute, which is about 5 % of the rate exhibited by nitrogenase [47].

8.2. Mechanisms

All of Schrauzer's systems are suggested to function through side-on binding of a dinitrogen to a Mo^{IV} species, followed by conversion of Mo^{IV} to Mo^{VI} and generation of diazene. This is contentious. SCHRAUZER also studied the V^{II}/Mg^{II} [48] and the Mo^{III}/Ti^{III} [49] gel systems of Shilov and concluded that the characteristics of these systems are more compatible with his side-on mechanism than with Shilov's bridging mechanism. The situation is not clear.

9. Fixation of Nitrogen Involving Metal Clusters

Although molybdenum is believed both to be the active site and to be contained in a metal–sulfur cluster in molybdenum nitrogenase, nitrogen fixation by metal clusters has not been definitively proven. Some clusters may be efficient electron-transfer agents to catalysts, but the few claims of fixation by clusters either have been disproved or have yet to be confirmed. One of the most complete studies is of the reduction of dinitrogen by $[Fe_4S_4(SPh)_4]^{2-}$ and $[Mo_2Fe_6S_8(SPh)_9]^{3-}$ under conditions of controlled potential electrolysis in methanol–tetrahydrofuran [50]. The 3− and 5−cluster anions are said to be responsible for binding and reducing dinitrogen, respectively. Yields of up to 3.4 µmol of ammonia were produced by 16 µmol of cluster after 24 h electrolysis at 25 °C. In aqueous solutions, 3.5 µmol of molybdenum cluster can produce 7 µmol of ammonia in 4 d, which may indicate a genuine electrocatalysis.

10. Photochemical Fixation Processes

Photochemical processes can be used to overcome the activation energy barrier to ammonia synthesis. The production of dihydrogen by photochemical splitting of water has been extensively investigated. In particular, iron-doped anatase or iron-doped rutile (TiO_2) under mercury-light irradiation is a good catalyst for the production of dihydrogen and dioxygen. In the presence of dinitrogen, ammonia is also produced, together with some hydrazine [51]. Yields of ammonia and hydrazine vary with time of reaction, temperature, and nitrogen pressure. The overall equations can be written as follows.

$$N_2 + 3\,H_2O \xrightarrow[TiO_2]{h\nu} 2\,NH_3 + 3/2\,O_2$$
$$N_2 + 2\,H_2O \xrightarrow[TiO_2]{h\nu} N_2H_4 + O_2$$

Both reactions are endothermic.

Iron-doped TiO_2 was prepared by impregnating anatase with solutions of iron(III) sulfate and heating the resulting powder to 1000 °C in air for 1–4 h. The product was exposed to mercury arc irradiation for 1–3 h in the presence of the reactants. In a typical experiment, 6.0 ± 2.5 µmol of ammonia and 0.19 ± 0.07 µmol of hydrazine were produced by 3-h irradiation of 0.2 g of catalyst (0.2 wt % Fe_2O_3) at 40 °C under a nitrogen pressure of 101.3 kPa. This fixation process is also exhibited by certain desert sands [52], and it has been suggested that similar processes

may have contributed to the terrestrial stock of fixed nitrogen. The mechanism of this process has not been established. SCHRAUZER believes that diazene is involved as an intermediate but does not discuss the reaction route in great detail. However, it has also been postulated that the N_2/H_2O reaction over rutile proceeds via an oxidative route involving formation of hydrogen peroxide:

$$\begin{array}{lll} & OH^-_{ads} + \text{positive hole} & \xrightarrow{TiO_2} OH^{\bullet}_{ads} \\ & 2OH_{ads} & \longrightarrow H_2O_{2_{ads}} \\ & H_2O_{2_{ads}} & \longrightarrow H_2O_{ads} + O^*_{ads} \\ & N_{2(g)} + O^*_{ads} & \longrightarrow NO^*_{ads} + N_{ads} \\ & N_{ads} + O^*_{ads} & \longrightarrow NO^*_{ads} \\ \text{or} & N_{ads} + O_{2_{ads}} & \longrightarrow NO^*_{ads} + O_{ads} \end{array}$$

where the asterisked species are either neutral or anionic [53]. Although this process may also have significance terrestrially under certain circumstances, the species involved have only been detected spectroscopically.

11. Oxidative Fixation Processes

Oxidative processes for the fixation of nitrogen are no longer commercially viable, the arc process for the conversion of dinitrogen to nitrogen(II) oxide having ceased production in the early 1940s [54]. The synthesis of calcium cyanamide by reaction of calcium carbide with dinitrogen in an electric furnace is part of a commercial route to cyanide but is is not generally competitive for fertilizer production.

At relatively high temperature, metal oxides can oxidize dinitrogen to oxides of nitrogen, but no commercial applications have resulted. An attempt is currently being made to use off-peak electricity to energize a mobile arc system that could be used to supply fixed nitrogen directly to agricultural land. The whole apparatus is planned to moved on a railway track, for example [55]. It is not clear whether this is commercially viable. A very imaginative low-technology application of the arc process based on renewable energy resources (such as sun, wind, and water) has been developed. This is intended to provide nitrate fertilizer to farmers in remote areas and should cost only a few hundred dollars. The prototypes appear to function satisfactorily. Whether this approach will be successful in the field is still not clear [56].

12. References

1. H. Bortels, *Arch. Mikrobiol.* **1** (1930) 333 – 342.
2. A. D. Allen, C. V. Senoff, *J. Chem. Soc. Chem. Commun.* 1965, 621 – 622.
3. J. R. Postgate: *The Fundamentals of Nitrogen Fixation,* Cambridge University Press, 1982, p. 1.
4. R. N. Pau, L. A. Mitchenall, R. L. Robson, *J. Bacteriol.* **171** (1989) 124 – 129.
5. R. R. Eady, *BioFactors* **1** (1988) 111 – 116.
6. R. R. Eady, R. N. Pau, *AFRC News,* April 1988, p. 4. R. R. Eady et al., *Biochem. J.* 244 (1987) 197 – 207.
7. D. J. Lowe, R. N. F. Thorneley, *Biochem. J.* **224** (1984) 895 – 901.
8. R. A. Henderson, G. J. Leigh, C. J. Pickett, *Adv. Inorg. Chem. Radiochem.* **27** (1983) 197 – 287.
9. J. Chatt, J. R. Dilworth, R. L. Richards, *Chem. Rev.* **78** (1978) 589 – 625.
10. R. A. Dixon, *J. Gen. Microbiol.* **130** (1984) 2745 – 2755.
11. C. Kennedy et al., *Mol. Gen. Genet.* **205** (1986) 318 – 325.
12. J. Chatt, G. A. Heath, R. L. Richards, *J. Chem. Soc. Dalton Trans.* 1974, 2074 – 2082.
13. J. Chatt, A. J. Pearman, R. L. Richards, *J. Chem. Soc. Dalton Trans.* 1977, 1852 – 1860.
14. J. Chatt, A. J. Pearman, R. L. Richards, *Nature (London)* **253** (1975) 39 – 40.
15. A. Galindo, A. Hills, D. L. Hughes, R. L. Richards, *J. Chem. Soc. Chem. Commun.* 1987, 1815 – 1816.
16. G. E. Bossard et al., *Inorg. Chem.* **22** (1983) 1968 – 1970.
17. J. Chatt, R. A. Head, G. J. Leigh, C. J. Pickett, *J. Chem. Soc. Dalton Trans.* 1978, 1638 – 1647.
18. P. C. Bevan, J. Chatt, G. J. Leigh, E. G. Leelamani, *J. Organomet. Chem.* **139** (1977) C 59 – C 62.
19. W. Hussain, G. J. Leigh, H. Mohd.-Ali, C. J. Pickett, *J. Chem. Soc. Dalton Trans.* 1986, 1473 – 1477.
20. C. J. Pickett, J. Talarmin, *Nature (London)* **317** (1985) 652 – 653; C. J. Pickett, K. S. Ryder, J. Talarmin, *J. Chem. Soc. Dalton Trans.* 1986, 1453 – 1457.
21. S. M. Rocklage, H. W. Turner, J. D. Fellman, R. R. Schrock, *Organometallics* **1** (1982) 703 – 707.

22. R. D. Sanner, J. M. Manriquez, R. E. Marsh, J. E. Bercaw, *J. Am. Chem. Soc.* **98** (1976) 8351 – 8357.
23. G. Pez, *J. Am. Chem. Soc.* **98** (1976) 8072 – 8078.
24. R. A. Henderson, *J. Chem. Soc. Dalton Trans.* 1982, 917 – 925.
25. R. A. Henderson, *J. Organomet. Chem.* **208** (1981) C 51 – C 54.
26. R. A. Henderson (1988), unpublished observations.
27. R. A. Henderson, G. J. Leigh, C. J. Pickett, *J. Chem. Soc. Dalton Trans.* 1989, 425 – 430.
28. H. M. Colquhoun, *Acc. Chem. Res.* **17** (1984) 23 – 28.
29. M. E. Vol'pin, V. B. Shur, *Dokl. Akad. Nauk SSSR* **156** (1964) 1102 – 1104.
30. M. E. Vol'pin, *J. Organomet. Chem.* **200** (1980) 319 – 334.
31. E. E. van Tamelen, *Acc. Chem. Res.* **3** (1970) 361 – 367.
32. C. Kruger, Y.-H. Tsai, *Angew. Chem. Int. Ed. Engl.* **12** (1973) 998 – 999.
33. R. Hammer et al., *Angew. Chem. Int. Ed. Engl.* **15** (1976) 612 – 613.
34. J. H. Teuben, *J. Organomet. Chem.* **57** (1973) 159 – 167.
35. J. E. Bercaw, E. Rosenberg, J. D. Roberts, *J. Am. Chem. Soc.* **96** (1974) 612.
36. F. W. Van der Weij, J. H. Teuben, *J. Organomet. Chem.* **120** (1976) 203 – 207.
37. Y. G. Borodko et al., *J. Chem. Soc. Chem. Commun.* 1973 169.
38. A. E. Shilov, *J. Mol. Catal.* **41** (1987) 221 – 234.
39. L. P. Luneva, L. A. Nikonova, A. E. Shilov, *Kinet. Katal.* **21** (1980) 1041 – 1045.
40. G. W. Nikolaeva, O. N. Efimov, Kh-M. A. Brikenshtein, A. E. Shilov, *React. Kinet. Catal. Lett.* **6** (1977) 349 – 356.
41. S. I. Kulakovskaya, V. N. Tsarev, O. N. Efimov, A. E. Shilov, *Kinet. Catal.* **18** (1977) 1045 – 1047.
42. L. P. Didenko, A. K. Shilova, A. A. Shilov, *Dokl. Akad. Nauk SSSR* **254** (1980) 643.
43. A. K. Shilova, V. D. Makhaev, A. E. Shilov, *Dokl. Akad. Nauk SSSR* **277** (1984) 1414.
44. G. N. Schrauzer, *Angew. Chem. Int. Ed. Engl.* **14** (1975) 514 – 522.
45. P. R. Robinson et al., *J. Am. Chem. Soc.* **99** (1977) 3657 – 3663.
46. E. L. Moorehead, P. R. Robinson, T. M. Vickrey, G. N. Schrauzer, *J. Am. Chem. Soc.* **98** (1976) 6555 – 6561.
47. B. J. Weathers, J. H. Grate, N. A. Strampach, G. N. Schrauzer, *J. Am. Chem. Soc.* **101** (1979) 925 – 928.
48. S. I. Zones, T. M. Vickrey, J. G. Palmer, G. N. Schrauzer, *J. Am. Chem. Soc.* **98** (1976) 7289 – 7295.
49. N. L. Liu, N. Strampach, J. G. Palmer, G. N. Schrauzer, *Inorg. Chem.* **23** (1984) 2772 – 2777.
50. K. Tanaka, Y. Hozumi, T. Tanaka, *Chem. Lett.* 1982, 1203 – 1206.
51. G. N. Schrauzer, T. D. Guth, *J. Am. Chem. Soc.* **99** (1977) 7189 – 7193.
52. G. N. Schrauzer et al., *Proc. Natl. Acad. Sci. USA* **80** (1983) 3873 – 3876.
53. R. I. Bickley, V. Vishwanathan, *Nature (London)* **280** (1979) 306 – 308.
54. S. A. Topham in J. R. Anderson, M. Boudart (eds.): *Catalysis Science and Technology,* Springer Verlag, Berlin 1985 p. 1.
55. New York Times, Wednesday, August 26, 1987, D 6.
56. R. W. Treharne et al., *Proc. Int. Solar Energy Meeting,* Atlanta, Georgia 1979.

Nitroparaffins → Nitro Compounds, Aliphatic
Nitrophenols → Nitro Compounds, Aromatic
Nitrophosphates → Phosphate Fertilizers
Nitrosamines → Nitrates and Nitrites
Nitrotoluenes → Nitro Compounds, Aromatic
Nitrous Acid → Nitric Acid, Nitrous Acid, and Nitrogen Oxides
Nitrous Oxide → Nitric Acid, Nitrous Acid, and Nitrogen Oxides
Nobelium → Transuranium Elements

Noble Gases

PETER HÄUSSINGER, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Chaps. 1 – 3, Sections 4.2 and 4.3, Chap. 9)

REINHARD GLATTHAAR, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Section 4.1.1)

WILHELM RHODE, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Sections 4.1.2 and 4.1.3)

HELMUT KICK, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Section 4.2)

CHRISTIAN BENKMANN, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Section 4.2.2)

JOSEF WEBER, Linde AG, Process Engineering and Contracting Division, Höllriegelskreuth, Federal Republic of Germany (Chap. 5)

HANS-JÖRG WUNSCHEL, Linde Technische Gase GmbH, Höllriegelskreuth, Federal Republic of Germany (Chap. 6)

VIKTOR STENKE, Linde Technische Gase GmbH, Höllriegelskreuth, Federal Republic of Germany (Chap. 7.1)

EDITH LEICHT, Linde Technische Gase GmbH, Höllriegelskreuth, Federal Republic of Germany (Sections 7.2 and 7.3)

HERMANN STENGER, Linde Technische Gase GmbH, Höllriegelskreuth, Federal Republic of Germany (Chap. 8)

1. Introduction 2
2. Properties 3
2.1. Physical Properties 3
2.1.1. Single-Component Systems 3
2.1.2. Two- and Multicomponent Systems 8
2.2. Chemical Properties 12
2.2.1. Chemistry of Helium, Neon, and Argon 13
2.2.2. Chemistry of Krypton 13
2.2.3. Chemistry of Xenon 14
2.2.4. Radon Compounds 17
2.2.5. Molecular Structures of Noble Gas Compounds 17
2.2.6. Noble Gas Clathrate Compounds . 18
3. Occurrence and Availability 18
3.1. Terrestrial Occurrence 18
3.2. Occurrence of Radioactive Isotopes 22
3.3. Lunar and Solar Occurrence . . . 23
4. Extraction and Isolation 23
4.1. Isolation From Air 24
4.1.1. Helium and Neon 24
4.1.2. Argon 28
4.1.3. Krypton and Xenon 30
4.2. Helium Extraction from Natural Gases 31
4.2.1. Crude Helium Extraction 31
4.2.1.1. Low-Temperature Fractionation . . 31
4.2.1.2. Crude Helium Extraction by Permeation Processes 34
4.2.2. Purification of Helium 35
4.2.2.1. Purification of Helium by Pressure Swing Adsorption 36
4.2.2.2. Other Processes for Purification of Helium 37
4.2.3. Example of a Complete Process . . 37
4.3. Isolation of Argon from Ammonia Purge Gases 39
5. Liquefaction of Helium 41
5.1. Processes for Liquefaction of Helium 41
5.1.1. Recuperative Processes 42
5.1.2. Regenerative Processes 43
5.2. Gas Purification for Helium Liquefiers 44
5.3. Examples of Helium Cryo Plants 45
5.3.1. Liquefier Using the Claude Process 45
5.3.2. Liquefier Using the Stirling Process 47
5.3.3. Helium Cryo Plant Using the Gifford – McMahon (GM) Process . . . 48
5.4. Process Developments 48
6. Quality Specifications and Analysis 50
7. Uses 51

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_485

7.1. Welding 51
7.2. Illumination 51
7.3. Other uses 52
8. Storage and Transportation 53
9. Economic Aspects 53
10. References 54

1. Introduction

The noble gases helium (He), neon (Ne), argon (Ar), krypton (Kr), xenon (Xe), and radon (Rn) form group 18 of the periodic table. They are mixtures of numerous isotopes. Table 1 lists the stable natural isotopes and their abundances.

Table 1. Stable isotopes of the noble gases

Element	Atomic number	A_r	Mass number	Abundance, %
Helium	2	4.0026	3	0.000137
			4	99.999863
Neon	10	20.179	20	90.92
			21	0.257
			22	8.82
Argon	18	39.948	36	0.337
			38	0.063
			40	99.600
Krypton	36	83.80	78	0.354
			80	2.27
			82	11.56
			83	11.55
			84	56.90
			86	17.37
Xenon	54	131.30	124	0.096
			126	0.090
			128	1.919
			129	26.44
			130	4.08
			131	21.18
			132	26.89
			134	10.44
			136	8.87
Radon	86	222		

Only radioactive isotopes of radon, with mass numbers 199 – 226, are known, several of which were formerly referred to by the archaic term emanation. The most important radioactive isotopes of the noble gases, together with their half-lives, are as follows (∗ denotes a metastable state of the isotope):

^{19}Ne	18 s
^{23}Ne	38 s
^{24}Ne	3.38 min
^{37}Ar	34.3 d
^{41}Ar	1.83 h
^{77}Kr	1.2 h
^{81}Kr	2.1×10^5 a
^{85}Kr∗	4.4 h
^{85}Kr	10.76 a
^{87}Kr	76 min
^{125}Xe	18 h
^{131}Xe ∗	12 d
^{133}Xe ∗	2.3 d
^{133}Xe	5.27 d
^{135}Xe	9.1 h
^{199}Rn – ^{226}Rn	0.27 μs – 3.32 d

All of the noble gases exist as gases under standard conditions. Because their structures consist of very stable closed shells of electrons, the noble gases are chemically extremely unreactive. Only krypton, xenon, and radon form chemical compounds, and these only with the most electronegative elements (F, Cl, O). General literature can be found in [1–5].

History. The first investigations of solar protuberances were performed during the solar eclipse of 1868. One of the most prominent emission lines was a bright yellow line at 587.6 nm, which in 1869 was also found in the spectrum of the chromosphere and was originally attributed to hydrogen. LOCKYEAR believed that the line was due to a new element, which was given the name helium by FRANKLAND and LOCKWOOD because of its occurrence in the sun (*helios*, the sun) [6]. In 1885 RAMSAY and CROOKES identified terrestrial helium, formed by dissolution of the uranium mineral cleveite in acids. In the same year KAYSER demonstrated by means of spectroscopy that helium also occurs in the earth's atmosphere. The other noble gases were discovered between 1884 and 1889 by RAYLEIGH, RAMSAY, and TRAVERS, partly using the air liquefaction process developed by VON LINDE. The initiative for this work was provided by precise measurements of the density of atmospheric nitrogen obtained by absorbing the oxygen with copper, and of nitrogen prepared chemically from nitrogen compounds such as nitrous oxide, nitric oxide, ammonia, or urea. Having removed the oxygen from air, RAYLEIGH and RAMSAY treated the nitrogen with red-hot magnesium turnings. The spectrum of the residual gas exhibited a series of previously unknown red and green lines. The ratio of the

specific heats ($C_p/C_v = 1.66$) indicated the presence of a monoatomic gas. They attributed these lines to a new element, which they later named argon on account of its low reactivity (the idle). The discovery of argon was followed by krypton (the hidden), neon (the new), and xenon (the strange). The unstable isotopes of the last noble gas radon were discovered as products of various radioactive decay series in 1900 – 1904 by RUTHERFORD et al. Summaries are given in [7, 8].

KAMERLINGH ONNES achieved the liquefaction of helium in 1908, thereby establishing a new temperature range which lay a factor of ten closer to absolute zero than was attainable with liquid hydrogen. Solid helium was first obtained in 1926 by KEESOM by using high pressure.

Although the high helium contents of U.S. natural gas had been known since 1805, up to 1917 helium was prepared only in laboratory quantities from uranium and thorium minerals. The production of helium from natural gas in the United States by means of cryogenic technology was initiated by SEIBEL.

Shortly after its discovery, the bright red light emitted by neon in the Geissler tube suggested its use in illumination. Although the first neon sign, containing a neon mixture, was manufactured by Griesheim – Elektron in 1913, it was only much later that pure neon was produced. In 1960, 10 L of liquid neon was transported for the first time. Also in 1913, LANGMUIR discovered that the light yield and life of incandescent bulbs could be increased when the bulb was filled with an argon-rich argon – nitrogen mixture instead of being evacuated. Thus the use of argon for illumination commenced. Large-scale production of argon first became feasible with the patenting of processes by L'Air Liquide (1920) and the Gesellschaft für Linde's Eismaschinen (1914). Up to 1930 krypton and xenon were prepared in small amounts by rectification of liquid air. Only after these gases became important in lighting applications were they prepared on a large scale.

Noble gases are now of major importance in modern technology. Production of pure argon increased greatly in the United States after 1943, in the Federal Republic of Germany from 1950, and subsequently worldwide due to its increasing use as a protective gas for the production and processing of metals. Compounds of the noble gases were predicted, among others, by PAULING [9]. In 1962 BARTLETT [10] and HOPPE [11] independently prepared xenon compounds. These discoveries led to a rapid development of the chemistry of xenon.

2. Properties

2.1. Physical Properties

2.1.1. Single-Component Systems

The noble gases are colorless, tasteless, odorless substances that exist in the gas state under standard conditions. The electron distribution in the atoms is spherically symmetrical. Therefore, the noble gases, particularly ^{3}He and ^{4}He, are often used as model substances for testing physicochemical theories. The interactions between noble gas atoms consist of weak van der Waals and London forces whose magnitude is directly proportional to the polarizability and inversely proportional to the ionization potential of the atoms. These relationships determine the gradation of properties among the noble gases. For example, the boiling points are proportional to the square root of the polarizibility, and the critical temperatures are proportional to the reciprocal ionization potentials.

Some physical properties of the noble gases are listed in Table 2. Since the isotopes ^{3}He and ^{4}He differ greatly in some of their properties, they are listed separately. Owing to the rarity of radon, its properties have not been studied extensively, and the values are less accurate than for the other noble gases. More recent listings of the thermophysical properties are given in [12] for helium, and in [13] for neon, argon, krypton, and xenon.

Solid and Liquid ^{3}He and ^{4}He [14–17]. The vapor pressure of ^{3}He and ^{4}He is listed in Table 3. Liquid and solid ^{3}He and ^{4}He exhibit properties that are not observed for any other substances. They are the only elements that do not possess a triple point, they remain liquid down to absolute zero upon cooling under their own vapor pressure, and freeze only under high pressure. For example, the minimum pressure for the solidification of ^{3}He at 1 K is 4.011 MPa and at 3 K, 12.9 MPa. The corresponding values

Table 2. Physical properties of the noble gases

	^{3}He	^{4}He	Ne	Ar	Kr	Xe	Rn
Triple point							
T, K			24.54	83.798	115.96	161.35	ca. 202
p, mbar			433.0	688.92	732	816	
ϱ_{sol}, g/cm^3			1.442	1.623	2.826	3.540	
ϱ_{liq}, g/cm^3			1.247	1.418	2.451	3.076	ca. 5
mp, K	0.319 ∗	0.775 ∗	24.57	84.0	116.2	161.4	ca. 202
Enthalpy of fusion, kJ/mol	0.000 ∗	0.000 ∗	0.335	1.17	1.64	2.30	ca. 2.5
Boiling point							
bp at 101.3 kPa, K	3.1905	4.224	27.09	87.295	119.79	165.03	211
Enthalpy of vaporization, kJ/mol	0.0255	0.0837	1.761	6.56	9.03	12.63	18.4
ϱ_{liq}, g/cm^3	0.0586	0.1248	1.206	1.410	2.415	3.06	4.4
Critical values							
T, K	3.3093	5.2014	44.448	150.7	209.40	289.777	377.5
p, bar	1.1459	2.275	26.60	48.7	55.1	58.8	63.2
ϱ, g/cm^3	0.04119	0.06945	0.4835	0.535	0.909	1.105	1.528
Gas							
Statistical entropy at *bp*, kJ kmol^{-1} K^{-1}	28.238	37.598	96.389	129.217	145.030	157.287	(168.9)
Liter weight (0 °C, 101.3 kPa), g/L	0.13448	0.17847	0.9000	1.7839	3.7493	5.8971	9.73
Specific heat (constant pressure), J mol^{-1} K^{-1}	20.79	20.79	20.79	20.79	20.79	20.79	20.79
Thermal conductivity (0 °C), μW cm^{-1} K^{-1}	1419	1430	461	164	87.8	51.0	21.4
Refractive index							
(0 °C, 101.3 kPa), $(n-1)\cdot 10^6$							
$\lambda = 480.0$ nm		35.04	67.31	283.8	431.8	712.8	
$\lambda = 576.9$ nm		34.90	67.10	281.7	427.6	702.9	
$\lambda = 643.8$ nm		34.86	67.02	280.9	425.8	698.7	
Molar refraction ∗∗ (D line), cm^3		0.521	1.004	4.203	6.397	10.435	
Gas kinetic diameter (0 °C, 101.3 kPa), nm	25.6	21.9	25.9	36.6	39.5	46.4	
1st Ionization energy, kJ/mol		2368.7	2079.9	1520.4	1350.5	1170	1037

∗ Minimum of melting point curve: ^{3}He 29.315 bar [8,9]; ^{4}He 25.316 bar [10].

∗∗ $R = \frac{n^2-1}{n^2+2} \cdot \frac{A_{\mathrm{r}}}{\varrho}$.

Table 3. Vapor pressures and saturation temperatures of ^{4}He and the atmospheric noble gas isotope mixtures. The horizontal lines indicate the phase transition He-I/He-II for ^{3}He and solid/liquid for the other noble gases

Pressure, bar	Temperature, K						
	^{3}He	^{4}He	Ne	Ar	Kr	Xe	Rn ∗
10^{-7}	0.2280	0.549	9.95	34.35	48.9	68.60	94
10^{-6}	0.2757	0.643	10.95	37.66	53.3	74.74	101
10^{-5}	0.344	0.771	12.17	41.67	58.74	82.10	110
10^{-4}	0.451	0.953	13.68	46.61	65.37	91.11	121
10^{-3}	0.631	1.226	15.51	52.77	73.70 ∗∗	102.39	134
10^{-2}	0.966	1.666	18.13	60.70	83.70 ∗∗	116.96	152
10^{-1}	1.657	2.483	21.60	71.41	99.61	137.54	176
0.25	2.123	3.020	23.36	76.96	107.02	146.46	188
0.50	2.589	3.544	24.95	81.51	113.19	154.79	198
1.00	3.178	4.201	27.06	87.17	120.12	164.79	210.7
2.00		5.018	29.56	94.25	129.26	177.78	227.0
5.00				105.76	144.68	198.98	253.2
10.00				116.57	159.17	218.63	278.2
20.00				129.72	176.77	242.48	309.9
30.00				138.68	188.79	258.82	328.3
40.00				145.71	198.18	271.59	346.0
50.00					205.92	282.15	360.4
60.00							372.9

∗ The log $p-(1/T)$ curves from various sources exhibit differing positions and gradients. Here, the formula $\log(p/\mathrm{Torr}) = 7.4955 - 884.41/(250.0+t)$ [33] was used for liquid radon, and for solid radon, Table 33 from Ref. [33].

∗∗ Ref. [34] gives the following values: 10^{-3} bar at 73.84 K and 10^{-2} bar at 84.39 K.

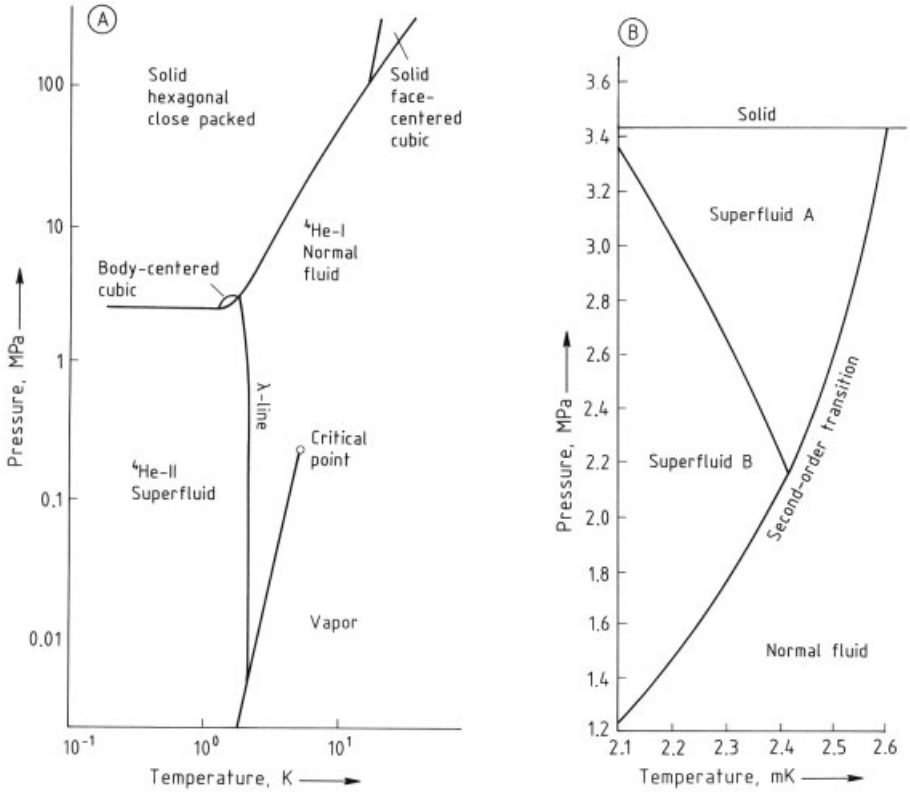

Figure 1. A) Phase diagram of ^{4}He; B) Phase diagram of ^{3}He in the existence range of the superfluid modification [27]

for ^{4}He are 2.5347 MPa (1 K) and 7.996 MPa (3 K). The melting point – pressure curves of ^{3}He [18–20] and ^{4}He [21–23] have been measured up to very high pressures; extrapolated values are given in [24]. The curves exhibit a minimum at low temperature. At the temperature of the minimum, the heat of melting is zero. Below this temperature, heat is liberated upon melting; i.e., heat must be supplied for helium to freeze. Both helium isotopes form three solid allotropes (see Figs. 1 A [17,24–26] and B [27]).

Helium I and Helium II. If boiling ^{4}He (^{4}He-I) is cooled by lowering the pressure, it does not solidify. Instead, it undergoes a second-order transformation at 2.177 K and 5.04 kPa to give a new liquid phase, ^{4}He-II. The existence ranges of the two phases are separated by the λ-line. Whereas ^{4}He-I behaves as a normal fluid, ^{4}He-II exhibits unusual properties, so that anomalies are observed upon the phase transition ^{4}He-I $\rightarrow$ ^{4}He-II.

Thus, upon pumping down boiling ^{4}He-I, the liquid suddenly becomes calm when the λ-point is reached. The smooth surface of the liquid forms an extremely sharp meniscus at the wall of the vessel, although in fact the ^{4}He-I continues to evaporate. The reason for this behavior is that at the λ-point the viscosity undergoes a millionfold decrease while the thermal conductivity increases by a factor of a million. Near absolute zero, the viscosity becomes zero, and ^{4}He-II is a frictionless superfluid. ^{4}He-II flows over obstacles that lie above the miniscus to reach the lowest attainable level. If a vessel is filled by dipping into a bath of ^{4}He-II and then raised above the surface of the bath, the ^{4}He-II flows as a film up the inner wall of the vessel, over the edge, down the outer wall, and drips from the bottom of the vessel back into the bath.

In saturated ^{4}He-II vapor, solid surfaces are covered with a film of ^{4}He-II about 100 atomic diameters thick. The film moves across the surface completely without friction up to a critical

velocity of 30 cm/s. The frictionless flow of ^{4}He-II is also observed in capillaries up to a diameter of 0.001 cm. The superfluidity of ^{4}He-II gives rise to unusual thermomechanical effects. When two vessels containing ^{4}He-II are linked by a small opening, a capillary, or a surface coated by ^{4}He-II, ^{4}He flows from the colder vessel to the warmer vessel. The flow of cold ^{4}He-II through a capillary into a warmer vessel is best demonstrated by the so-called fountain effect. A periodic addition of heat to ^{4}He-II is propagated as a wave (second sound), whereas in other substances a thermal wave is immediately damped and the energy dispersed [28]. The thermal conductivity of ^{4}He-II is ca. 3×10^6 greater than that of ^{4}He-I and ca. 200 times greater than that of copper at 20 °C.

The anomalous behavior results from quantum effects, although it has not been explained completely; ^{4}He-II is often referred to as a quantum fluid [29]. Many of the properties of ^{4}He-II can be explained qualitatively by the two-liquid model [14,30], which regards ^{4}He-II as being a mixture of two interpenetrating liquids, one a normal fluid like ^{4}He-I and the other a perfect superfluid. The concentration of the superfluid is taken as 0 % at the λ temperature and 100 % at 0 K. Superfluidity has also been observed for ^{3}He, but only below 2.6×10^{-3} K. Two superfluid phases, designated A and B, were detected. The intersection of the λ-line with the melting point – pressure curve is at 2.6 ± 0.1 mK and 3.44 MPa [31] (cf. Fig. 1 B); a summary is given in [27,32].

Solid and Liquid Neon, Argon, Krypton, Xenon, and Radon. In contrast to the trimorphic helium isotopes, all other noble gases crystallize in face-centered cubic structures. Vapor pressures are given in Table 3; the melting behavior is described in [35–38]. Solid or liquid radon phosphoresces due to its radioactivity.

Properties of the Gases. After hydrogen, helium is the lightest gas. At 15.5 °C and 101.3 kPa, 1 3 of ^{3}He weighs 0.1211 kg, 13 ^{4}He weighs 0.1607 kg, 13 hydrogen weighs 0.0810 kg, and 13 air weighs 1.16307 kg. Hence the buoyancy of helium in air is 1.00237 kg/m^3 and that of hydrogen 1.08207 kg/m^3. The slightly lower buoyancy and higher cost of helium as a balloon gas are balanced out by its nonflammability and by its rate of diffusion through balloon envelopes– 65 % lower than that of hydrogen.

Helium-3. The $H-T$ $(0<p<6.6\,\text{bar})$ and $H-p$ diagrams can be found in [39]; further thermodynamic properties are given in [40].

Helium-4. The temperature dependence of the properties of ^{4}He is shown in Table 4. Further thermophysical properties for temperatures of 4 – 3000 °R and 1500 psia are given in [23]; for 2 – 1500 K and up to 10^3 atm in [22,41]. Numerical coefficients of the functions $T(p,H)$, $\varrho(p,H)$, $S(p,H)$, $H(p,S)$, and $H(p)$ at the phase boundaries, based on the data in [41], can be found in [43]. Older values for 3 – 300 K and 0.5 – 100 atm, and for 5 – 540 °R and 10 – 1500 psia are given in [44,45]. Thermodynamic properties up to 50 000 K can be found in [46].

Neon. The temperature dependence of some physical properties of neon is shown in Table 5. Values of the coefficient of compressibility in the range 80 – 130 K and 75 – 2072 bar are given in [48]. Calculated values of the specific entropy and enthalpy are listed in [49].

Argon. The temperature dependence of some properties of argon is shown in Table 6. Values in the range 90 – 600 K and 1 – 5000 atm can be found in [51], while calculated values for 300 – 35 000 K and $\log p$ from -4.9 to $+4.2$ are given in [52].

Krypton. Values of p, V, T from 130 to 240 K are given in [53]; values for 0 – 150 °C and 0 – 1000 bar can be found in [54], together with calculated thermodynamic properties as a function of the Amagat density (the relative density divided by the relative density under standard conditions $\varrho_{T,P}/\varrho_{273.15,\,101.3\,\text{kPa}}$). Equations of state for 120 – 423 K and 0 – 1000 bar are given in [55], and for 120 – 1300 K and 0 – 1000 bar in [56].

Xenon. The p, V, T isotherms for the range 0 – 150 °C and 0 – 2800 atm are given in [57], and calculated values as a function of the Amagat density in [58] (see also [33]).

Like polyatomic gases, the noble gases do not follow the ideal gas laws (Fig. 2). For helium and neon, repulsive forces between the atoms dominate down to relatively low temperatures. Therefore, they are the only gases, apart from hydrogen and deuterium, that undergo a temperature increase upon expansion at normal temperatures.

Table 4. Properties of gaseous helium at 0.1 and 20 MPa and various temperatures [42] ∗

Temperature, °C	Compressibility coefficient	Density, kg/m^3	Enthalpy ∗∗ $H-H_0$, kJ/mol	Entropy ∗∗ $S-S_0$, J mol^{-1} K^{-1}	Thermal conductivity λ, WK^{-1}m^{-1}	Viscosity $\eta \cdot 10^5$, Pa · s
Pressure 0.1 MPa						
0	1.0005	0.17615	0.00131	0.04646	0.1426	1.864
20	1.0005	0.16414	0.4170	1.4688	0.1501	1.953
50	1.0004	0.14891	1.0406	3.4939	0.1612	2.083
100	1.0004	0.12896	2.0798	6.4842	0.1790	2.292
200	1.0003	0.10172	4.1583	11.4194	0.2126	2.687
500	1.0001	0.062256	10.394	21.626	0.3023	3.745
1000	1.0001	0.037809	20.787	31.994	0.4188	5.264
3000	1.0000	0.014707	62.357	51.622	0.7911	10.072
Pressure 20 MPa						
0	1.1023	32.013	0.262	− 43.961	1524	1.891
20	1.0947	30.028	0.677	− 42.492	1592	1.977
50	1.0851	27.475	1.300	− 40.470	1694	2.103
100	1.0725	24.068	2.338	− 37.483	1859	2.307
200	1.0553	19.286	4.414	− 32.555	2179	2.697
500	1.0308	12.081	10.639	− 22.364	3054	3.749
1000	1.0163	7.441	21.014	− 12.015	4206	5.265
3000	1.0044	2.9287	62.529	7.585	7918	10.072

∗ More comprehensive tables can be found in [42].
∗∗ Reference values: H_0 at $t = 0\,^\circ$C and $p = 0$ MPa; S_0 at $t = 0\,^\circ$C and $p = 0.1$ MPa.

Table 5. Properties of gaseous neon at 0.1 MPa and various temperatures [47] ∗

Temperature, °C	Compressibility coefficient	Density, kg/m^3	Enthalpy ∗∗ $H-H_0$, kJ/mol	Entropy ∗∗ $S-S_0$, J mol^{-1} K^{-1}
80	0.9941	3.0398	2.958	91.09
100	0.9947	2.4284	3.376	95.74
120	0.9952	2.0224	3.793	99.54
140	0.9965	1.7329	4.209	102.76
160	0.9981	1.5160	4.626	105.53
180	0.9993	1.3475	5.042	107.99
200	1.0002	1.2127	5.458	110.18
220	1.0007	1.1024	5.874	112.16
240	1.0009	1.0105	6.290	113.97
260	1.0009	0.9328	6.706	115.63
280	1.0008	0.8662	7.122	117.18
300	1.0005	0.8085	7.538	118.61

∗ Ref. [47] lists values for 30 – 300 K and 0.1 to 200 atm.
∗∗ Reference values: $H_0 = 0$ and $S_0 = 0$ for the liquid at the triple point.

Table 6. Properties of gaseous argon at 0.1 MPa and various temperatures [50]

Temperature, K	Compressibility coefficient	Density, kg/m^3	C_p, J/mol	C_p/C_v	Enthalpy ∗ $H-H_0$, kJ/mol	Entropy, $S-S_0$, J mol^{-1} K^{-1}	Thermal conductivity λ, μW cm^{-1} K^{-1}
100	0.9785	4.9105	21.67	1.706	2.030	131.85	64.4
200	0.99710	2.4092	20.91	1.674	4.142	146.52	124.7
300	0.99938	1.60240	20.84	1.670	6.229	154.98	176.6
400	0.99998	1.20108	20.81	1.668	8.311	160.97	222.7
500	1.00018	0.96068	20.80	1.668	10.392	165.61	264.0
600	1.00025	0.80049	20.80	1.667	12.472	169.40	301.6
800	1.00028	0.60037	20.79	1.667	16.631	175.38	368.3
1000	1.00026	0.48030	20.79	1.667	20.789	180.03	426.6
2000	1.00016	0.24017	20.79	1.667	41.579	194.43	
3000	1.00011	0.16012	20.78	1.667	62.367	202.86	

∗ H_0 = ideal gas enthalpy at zero pressure; S_0 = ideal gas entropy at 1 bar.

Table 7. Coefficients of the thermal conductivity λ [mW/mK] and the viscosity η [μP] (1 μP = 10^{-5} Pa · s) [59], for $T \leq 260$ K [60] and ≥ 300 K [61]

Temperature, K	He		Ne		Ar		Kr		Xe		Rn	
	λ	η	λ	η	λ	η	λ	η	λ	η	λ	η
180	107.2	139.5	34.7	222.4	11.31	144.6	5.93	157.0	3.52	142.1		
220	122.8	159.5	40.1	256.4	13.55	174.2	7.13	190.9	4.24	172.6	2.66 *	180 *
260	137.4	178.9	44.9	287.6	15.67	202.4	8.30	223.9	4.96	202.9	3.15 *	220 *
300	152.0	199.1	49.3	318.5	17.72	227.4	9.49	255.4	5.69	232.8	3.64 *	250 *
350	169.7	221.8	54.4	354.4	20.13	259.0	10.90	294.3	6.59	270.8	4.22 *	290 *
400	187.0	243.4	59.0	387.5	22.33	288.5	12.26	331.0	7.45	307.1	4.80 *	330 *
500	220.2	283.7	68.5	448.0	26.38	341.6	14.7	397.7	9.05	374.7	5.86 *	400 *
600	251.5	321.5	77.1	504.0	30.1	389.3	16.9	457.8	10.55	436.6	6.90 *	475 *
800	308	392.8	94.5	607.7	36.9	475.1	20.9	565.3	13.5	548.5	8.74 *	600 *
1000	357	459.9	109.1	703.5	42.7	552.9	24.4	661.7	16.1	649.1	10.42 *	720 *
1400	447	584.2	137.2	876.8	53.5 *	692.8	30.6	832.4	20.8	827.3		
1800	528	697.3	161.8 *	1031	64.1 *	817.1	35.9	982.4		983.4		
2000	566	750.3	172.7 *	1102	69.2 *	874.7	38.2	1052		1055		
2500	657		200 *		81.5 *							
f_1	2.50 ± 0.05		2.54 ± 0.04		2.52 ± 0.03		2.52 ± 0.04		2.61 ± 0.05		(2.62) *	

* Estimated or extrapolated.

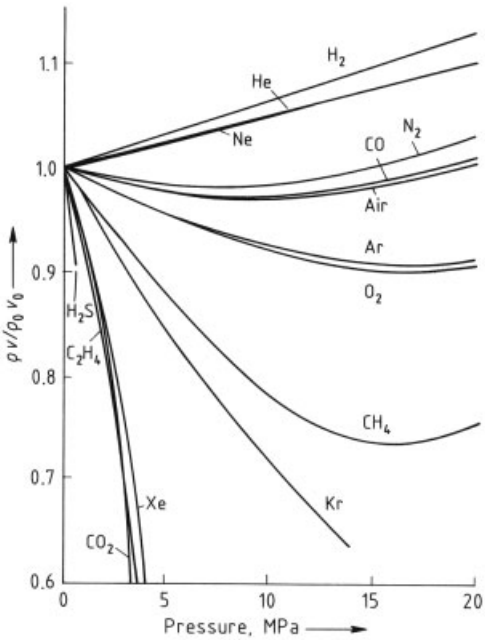

Figure 2. Compressibility coefficients of the noble gases compared to those of several polyatomic gases at 0 °C

The specific heat of the noble gases is independent of temperature and pressure over a wide range. The ratio C_p/C_v has the limiting value

$$\frac{\frac{3}{2}R + R}{\frac{3}{2}R} = 1.667.$$

Because they are monoatomic, the noble gases do not exhibit sound dispersion. The coefficients of the thermal conductivity λ, viscosity η, and diffusion D are given by

$$\lambda M_r/\eta C_v = f_1 \text{ and } D\varrho/\eta = f_2$$

where M is the relative molecular mass, C_v the molar heat capacity at constant volume, and ϱ the density. The function f_1 is a constant over a wide temperature range and has the value 2.5 corresponding to that expected for a monoatomic gas consisting of rigid balls. For ^{4}He, f_1 decreases below 100 K: at 81.4 K it is 2.44 and at 20.9 K, 1.99. The value of f_2 is ca. 1.3; it decreases slowly with increasing temperature. Table 7 lists values for λ and η. The values of λ and η increase with increasing pressure [62–65]. In the range 80 – 300 K, the viscosity of ^{3}He is ca. 15 % lower than that of ^{4}He [66].

2.1.2. Two- and Multicomponent Systems

Mixtures of Liquid ^{3}He and ^{4}He. The large differences in the physical properties of the helium isotopes is also demonstrated by their limited miscibility below 0.87 K (Fig. 3). Theoretically, phase separation is expected near absolute zero. The finite solubility of ^{3}He in ^{4}He

Table 8. Solubility of helium in vacuum pump oils [84]

Temperature, °C	Solubility ∗		
	Silicon DC 702	Silicon DC 200	Apiezon oil GW
20	0.61	1.36	0.68
30	0.72	1.58	0.87
40	0.81	1.78	1.03
60	0.94	2.12	1.29
80	1.02	2.38	1.46

∗ mmol He per kg oil at $p_H^e = 100$ kPa.

of 6.37 mol % was, however, unexpected. This anomalous behavior of $^3He-^4He$ mixtures is exploited in $^3He-^4He$ refrigerators for temperatures in the range $<<$ 100 mK to 1 mK [68–72].

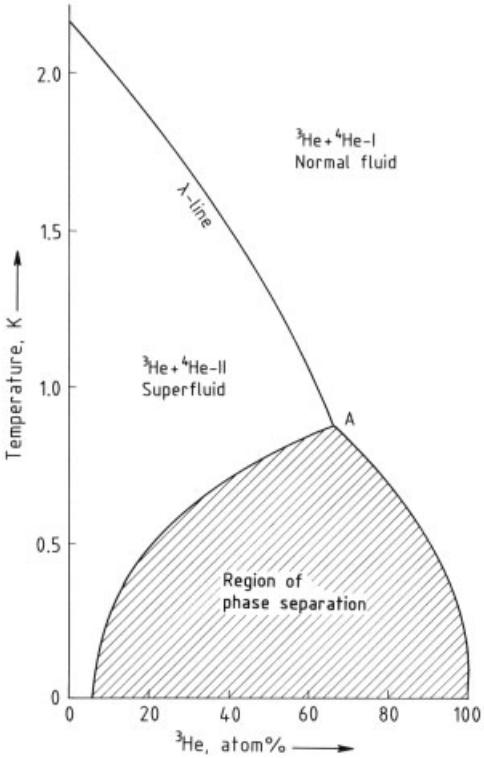

Figure 3. $^3He-^4He$ phase diagram [67, 70]

The melting behavior of binary mixtures containing argon, krypton, and xenon is described in [73] ($Ar+F_2$); [74, 75] ($Ar+CH_4$); [76] ($Ar+Kr$); [77] ($Ar+N_2$); [78] ($Ar+Xe$); and [79] ($Kr+CH_4$).

Solubility. The solubility of noble gases in liquids depends on the size (formation of interstices in the solvent) and deformability of the noble gas atoms (interaction between the dissolved atoms and the solvent). The very small helium atoms, in particular, can readily find interstices for their absorption, so that helium has a finite solubility, despite the extremely low deformability of its atoms. The solubility of helium is in most cases lower than that of all other gases. The solubility coefficient mostly decreases at high pressures. With increasing atomic volume, the interactive forces are dominant, and the solubility increases greatly from helium to radon. The solubility in nonpolar organic solvents is usually higher than in water. Dissolved electrolytes generally lower the solubility; exceptions are LiCl and LiI for helium, and $HClO_4$ for helium and argon. Solubility values for helium and neon are given in [80], for argon in [81], and for krypton, xenon, and radon in [82]. The solubilities of helium in some low-volatility oils are listed in Table 8; further values can be found in [83, 84].

Phase Equilibria. Because of theoretical interest and their importance for the separation of industrial gas mixtures, the solid – gas and liquid – gas equilibria of numerous binary, ternary, and multicomponent systems containing noble gases have been measured over wide ranges of temperature and pressure [85, 86]. Helium-3 and helium-4 exist in the supercritical state in the existence range of all other liquified gases. Therefore, as is the case with other supercritical gases the lens-shaped area enclosed by the boiling point curve and the dew point curve in the P, x, y or T, x, y diagrams does not reach the limiting value of 100 mol % He. The solubility of helium in liquid nitrogen and liquid methane [76] is given in Table 9 (see next page).

Table 9. Solubility of helium in liquid nitrogen and liquid methane at various temperatures [87]

Total gas pressure,	Solubility, mol %					
	in N_2			in CH_4		
MPa	64.9 K	95 K	120 K	95 K	124.9 K	184.8 K
1.01	0.140			0.034		
2.03	0.275	0.730		0.069	0.225	
4.05	0.505	1.650	2.60	0.135	0.490	
6.08	0.678	2.520	5.94	0.200	0.750	2.84
10.13		4.235	11.06	0.312	1.19	6.65
14.18		5.785	14.26	0.410	1.58	10.17
18.24				0.503	1.92	12.70

Gas – Gas Equilibria. Phase separation in the gas state has been observed at high pressure in many systems containing noble gases as components (first- and second-order gas – gas equilibria). Therefore, the critical curve and the miscibility gap extend well beyond the critical temperatures and pressures of the components [88]. Further work can be found in the literature: $He + H_2$, 0 – 10 kbar [89,90]; $He + N_2$, 0 – 4.3 kbar [91]; He + Ne [92]; H_2 + Ar [93]; Ne + Ar, 0 – 1 kbar [94]; Ne + Kr, 0 – 1.9 kbar [95].

Permeation. In contrast to hydrogen, noble gases are insoluble in solid and molten metals. Therefore, metals are impermeable up to high temperatures (red heat), even for the small helium atom. The permeation coefficient P for unalloyed and high-alloy creep-resistant steels free of lattice defects at 1160 – 1290 °C and a pressure difference of ca. 11 bar is less than 10^{-13} mmol $s^{-1} \cdot cm^{-1} \cdot bar^{-1}$ [96]. The extremely low permeation of helium is, in addition to the fact that only its very small content of ^3He can be activated by neutrons, of major importance for the cooling of nuclear reactors. Single crystals such as beryl (P_{He} 10^{-14} at 20 °C), fluorspar ($< 3 \times 10^{-15}$), mica ($< 10^{-16}$), calcite and quartz ($< 5 \times 10^{-16}$), selenite ($< 10^{-16}$), and rock salt ($< 10^{-13}$) are also practically impermeable to helium. The differing permeabilities of minerals are important in the formation of helium deposits and for determining the age of rocks. The permeability of inorganic glasses depends on their composition and pretreatment (Fig. 4). The permeability for helium increases with increasing proportion of glass-formers (SiO_2, B_2O_3, P_2O_5). Even at low temperatures, the permeation of helium through glass is sufficiently high that Dewar vessels containing liquid helium must be periodically or continuously re-evacuated. The permeation through glass decreases with increasing atomic radius of the noble gases.

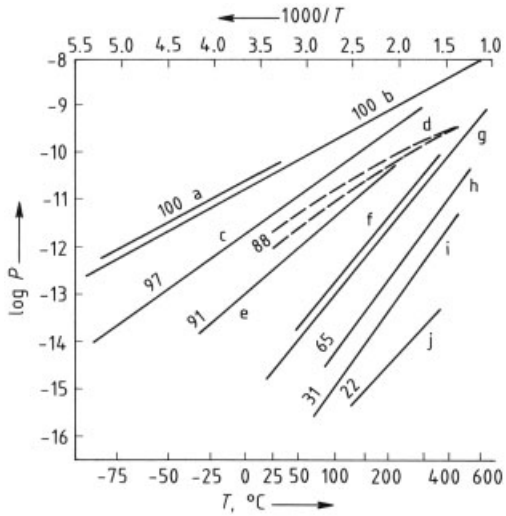

Figure 4. Permeation of helium through various types of glass [97,98] (the figures on the curves indicate the content of $SiO_2 + B_2O_3 + P_2O_5$ in wt %)
a) Vycor C; b) Quartz glass; c) Pyrex 97; d), e) Borosilicate glass; f) Phosphate glass; g) Soda glass; h) Pyrolysis tube; i) X-ray protective glass; j) Lead borate glass
—- according to [97] – – – according to [98]

Helium has a high permeation rate in organic materials. Permeability coefficients for various organic and inorganic polymer membranes are listed in Table 10. The coefficients are mostly considerably higher than those of methane and nitrogen, so that high selectivities can be attained. The high permeability and selectivity of membranes for helium are exploited in its separation from natural gas (see Section 4.2).

Table 10. Permeability coefficients of helium, nitrogen, and methane in inorganic and organic membrane materials

Membrane material	Temperature, °C	Permeability coefficient $P \cdot 10^{10}$, $\frac{cm^3 (STP \cdot cm)}{cm^2 \cdot s \cdot cm\,Hg}$			Selectivity α	
		He	N_2	CH_4	P_{He}/P_{N_2}	P_{He}/P_{CH_4}
Inorganic membranes						
Porous glass, Vycor type	25	140×10^3	56×10^3	79×10^3	2.5	1.8
Quartz glass	400	0.6		0.00002		$>10^5$
Molecular sieve carbon (3–5 Å) [100]	25	5200	240		20	
Porous Al_2O_3 (100–200 Å)					3.2	
Organic membrane materials						
Ethyl cellulose	20	31	2.8	11	11	2.8
Cellulose acetate	20	13.6	0.28		49	
Polytetrafluoroethylene (PTFE)	20	62	2	1.5	31	42
Poly(vinyl fluoride)	20	1.87	0.099	0.0064	19	292
Polystyrene (PS)	20	3.51	2.27	2.27	1.5	1.5
Polydialkylsiloxane	20	230	150	590	1.5	0.4
Polydimethylsiloxane (PDMS)	25	360	280	900	1.3	0.39
Polycarbonate	20	67	4.7	3.6	14.3	18.6
Aromatic polyimide (Pyrrolone) [101]	35	89	1.2	0.54	74	165

In organic nonporous polymer membranes normally the solubility increment of the permeability prevails. Therefore, in contrast to inorganic membranes, the permeability coefficient of the noble gases increases together with their solubility, which is higher at higher atomic masses. Thus xenon and krypton have higher permeability coefficients than the lighter noble gases argon and helium and the other constituents of air; e.g., with polydialkylsiloxane, $P_i \cdot 10^{10}$ [cm^3 (STP) $\cdot$ cm/($cm^2 \cdot s \cdot cm$ Hg)] at 20 °C is as follows: He, 230; N_2, 184; Ar, 401; Kr, 750; Xe, 1660 [99].

Kryptonates are solids whose surfaces contain intercalated krypton atoms. The intercalation can occur by bombardment of the solid by krypton ions, through diffusion at high temperature and pressure, or by tribosorption (e.g., working of metals in a krypton atmosphere). Kryptonates are stable at high temperatures, regardless of how they were prepared. The krypton occupies only the outer lattice positions up to a depth of 10^3 nm. The use of kryptonates in nuclear technology for the retention of radioactive fission gases has been suggested [102, 103].

Adsorption. Examples of the general adsorption behavior of the noble gases are shown in Figures 5 and 6 (see next page). Older results for activated carbon are given in [106, 107]. The adsorption of krypton and xenon on activated carbon, silica gel, and zeolites has been studied extensively because of its possible applications in reducing the pressure of heavy noble gases formed in nuclear fuel elements and for retaining radioactive fission gases [110–116].

The adsorption enthalpy decreases with increasing loading of the adsorbent (Fig. 7); it is usually 1.5 – 3 times larger than the heat of vaporization [117]. The lines of constant loading in the diagrams log p_i or log f_1 versus $1/T$ are linear over wide ranges. Therefore, like the gas – liquid phase equilibrium, the adsorption equilibrium follows the Clausius – Clapeyron law. The specific heat of the adsorbed gases is about one-third that of the solid noble gases.

Krypton and xenon cannot penetrate narrow-pored Na – K A zeolites at room temperature and are therefore not adsorbed. However, at high temperatures the gas atoms can be forced into the zeolites. When the resulting krypton-loaded molecular sieve (Na/K, 60/40) is cooled, the pores become smaller again, and the krypton remains trapped with almost zero leakage [118]. The gas can be liberated by heating.

The noble gases are used for testing theories of adsorption and, in particular the lighter members of the group, for determining the surface area of adsorbents, catalysts, and powders (BET method).

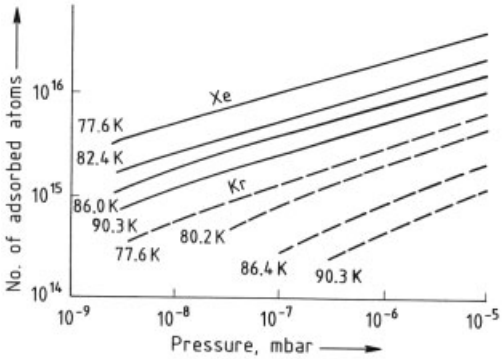

Figure 5. Adsorption isotherms for krypton and xenon on pyrex glass [104]

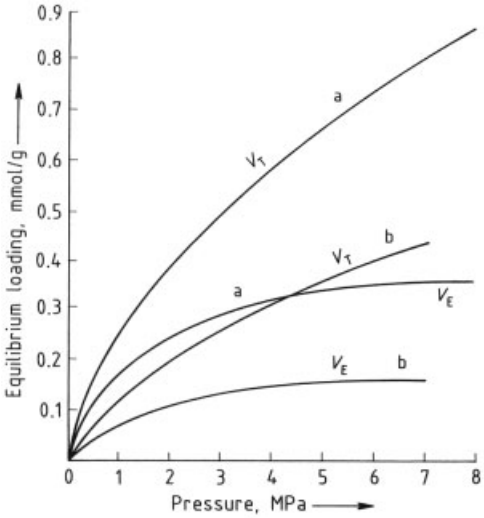

Figure 6. Adsorption isotherms of helium on activated carbon (a) and zeolites (b) at 76 K [105]
V_T is the excess of gas present in the pores of the adsorbent relative to the amount that would be present at normal density and equilibrium pressure. V_E is the total amount of gas held in the pores by adsorptive forces

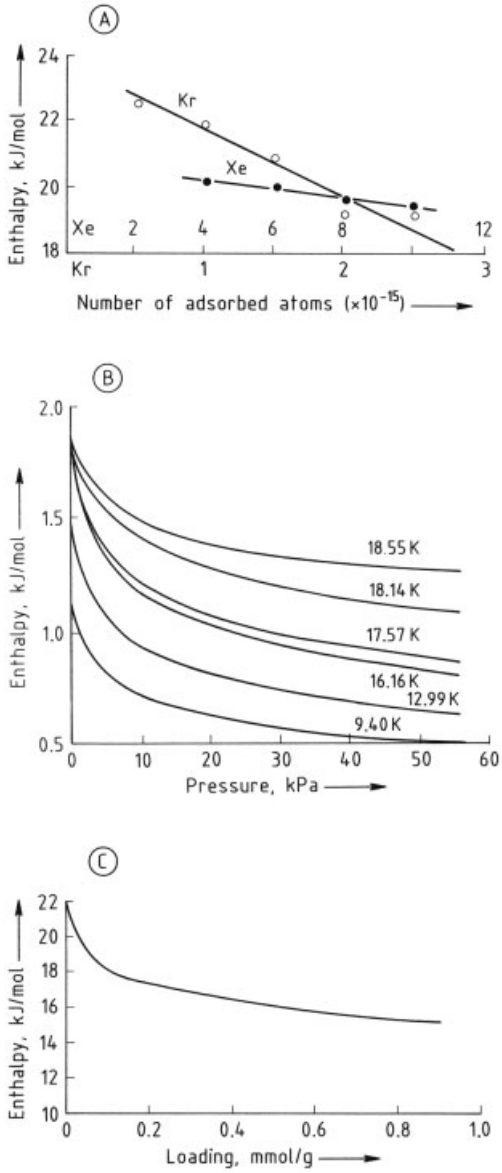

Figure 7. Dependence of adsorption enthalpy on loading
A) Krypton and xenon on pyrex glass [104]; B) Helium on carbotox [108]; C) Argon on activated carbon [109]

2.2. Chemical Properties

The electronic configurations of the noble gases, which are characterized by filled shells of electrons, are as follows:

He		$1s^2$			
Ne	[He]	$2s^2$	$2p^6$		
Ar	[Ne]	$3s^2$	$3p^6$		
Kr	[Ar]	$3d^{10}$	$4s^2$	$4p^6$	
Xe	[Kr]	$4d^{10}$	$5s^2$	$5p^6$	
Rn	[Xe]	$4f^{14}$	$5d^{10}$	$6s^2$	$6p^6$

The stability of the noble gases with respect to loss or acceptance of electrons is due to their high ionization potentials and the highly negative values of their electron affinities. These effects are essentially responsible for the chemical inertness of these elements.

It was long believed that the noble gases were incapable of forming chemical compounds.

Nevertheless, in 1962 three groups succeeded independently in preparing noble gas compounds. In June 1962 BARTLETT prepared the orange-yellow, moisture-sensitive compound "xenon hexafluoroplatinate" by reaction of Xe with PtF_6. In July 1962 HOPPE obtained the first binary compound of xenon, XeF_2, and this was followed one month later by the synthesis of XeF_4 by CLAASEN, SELIG, and MALM. These discoveries prompted a rapid growth in noble gas chemistry. A summary of the discovery of the noble-gas compounds and a discussion of priority claims is given in [119].

A surprising result was that the valence compounds of krypton and xenon do not involve a new type of chemical bonding. The atoms are bound in the same manner as in the long-known interhalogen compounds, such as IF_7, and in TeF_6. For the formation of noble gas compounds, an electron must be promoted from the p shell to the d shell, after which formation of bonding electron pairs with another atom can occur with hybridization of the type spd^n ($n = 1, 2, 3, 4$). A detailed description of the bonding is given in [120].

Noble gas compounds can be divided into three general types: (1) short-lived molecules containing noble gas atoms, (2) valence compounds, and (3) inclusion compounds (clathrates and intercalation compounds); comprehensive reviews are given in [120–123]. Valence compounds are only formed with the most electronegative elements (till now, F, Cl, Br, N, and C), whereby thermodynamically stable compounds with fluorine are formed only by xenon and radon. The thermodynamic stability of noble gas compounds, for example, the halides, follows the following general rules:

1) The stability of the compounds EX_2 increases with increasing atomic number of the noble gas and with decreasing atomic number of the halogen:

$$ArF_2 < KrF_2 < XeF_2 < RnF_2$$

$$XeF_2 > XeCl_2 > XeBr_2$$

2) The stability of the compounds decreases as the oxidation state of the noble gas increases.

2.2.1. Chemistry of Helium, Neon, and Argon

Extensive theoretical, radiochemical, and mass spectroscopic studies on short-lived ions and molecules have been carried out [124, 125]. Short-lived He_2^{2+} and He^{2-} ions have been theoretically predicted and experimentally detected [126]. The existence of a stable molecule He – Be – O, with a linear geometry, has been predicted on the basis of ab initio calculations [127], as has the formation of covalent bonds between neon or argon and carbon dications (C^{2+}, CC^{2+}) or BeO [128].

Up to now, attempts to react helium, neon, and argon with other elements have failed. Thus, attempts to prepare ArF_2 (estimated decomposition energy 105 kJ/mol) or a compound of the presumably more stable cation ArF^+ were unsuccessful.

2.2.2. Chemistry of Krypton

The chemistry of krypton is limited to the detection of the ions KrF^+ and $Kr_2F_3^+$ and the KrF_2 radical and the synthesis of KrF_2 and its complexes $KrF_2 \cdot 2\,SbF_5$ and $KrF_2 \cdot x\,AsF_5$. A report of the detection of KrF_4 proved to be erroneous. The existence of a compound with a Kr – N bond is claimed in [129].

Krypton (II) fluoride, KrF_2, is only stable below 0 °C. Therefore, it cannot be prepared thermally from the elements, and its synthesis requires strong activation at low temperature. It can be synthesized by:

1) Irradiation of a mixture of Kr and F_2 at − 150 °C with a 1.5 MHz electron beam [130], or at − 140 °C with 10 MeV protons or 40 MeV α-particles [131]
2) Electrical discharge through a 1: 1.1 mixture of Kr and F_2 in a prefluorinated recycling reactor at − 196 °C and 2.5 kPa [132]

Krypton (II) fluoride is now prepared straightforwardly by a photochemical reaction [133]; it is purified by sublimation.

Properties. Krypton (II) fluoride forms colorless, tetragonal crystals that can be stored without decomposition in dry pyrex or PTFE

vessels at $-40\,^{\circ}C$. The vapor pressure at $-40\,^{\circ}C$ is 0.2 – 0.25 kPa; the compound is readily sublimable below $0\,^{\circ}C$ at 4 kPa. The standard enthalpy of formation for the gas is + 60 kJ/mol, and for the solid it is + 15.5 kJ/mol; thus, KrF_2 is metastable. The KrF_2 molecule is linear with a Kr – F bond length of 189.9 pm in the gas phase.

Reactions. Krypton (II) fluoride is hydrolyzed by water and dilute acids and alkalies:

$$KrF_2 + H_2O \rightarrow Kr + 1/2\,{}_2O + 2\,HF$$

Treatment of KrF_2 with SbF_5 gives the adduct $KrF_2 \cdot 2\,SbF_5$, which is composed of $[KrF_2]^+[Sb_2F_{11}]^-$ ion pairs. The compound is colorless and dissolves in excess SbF_5 to form a clear, yellow solution. The $[KrF]^+$ ion can be stabilized by other anions such as AsF_6^-, TaF_6^-, $Ta_2F_{11}^-$, and $Nb_2F_{11}^-$. Hydrolysis of the adducts $KrF_2 \cdot MF_5$ in alkaline or weakly acidic solutions gives krypton, oxygen, and fluorine monoxide. The adducts react explosively with organic compounds. The first compound of two noble-gas fluorides, the adduct $KrF_2 \cdot XeF_6$, which is more stable than XeF_2, was synthesized in 1975 [134]. Krypton – oxygen compounds have not yet been prepared. The most promising attempt– the reaction of KrF_2 with the $OTeF_5$ ligand– was unsuccessful [135]. At $-100\,^{\circ}C$ the reaction gave only elemental krypton. Graphite – KrF_2 intercalation compounds of varying compositions have been obtained by prolonged contact of KrF_2 with graphite [136]. A compound with a krypton – nitrogen bond, $[HC{=}NKrF]^+$ AsF_6^-, has been obtained by reacting KrF_2 with $[HC{=}NH]^+AsF_6^-$ in HF at $-60\,^{\circ}C$; the compound decomposes explosively above $-50\,^{\circ}C$ [137]. Krypton (II) fluoride is one of the most powerful fluorinating agents known. Fluorination reactions with KrF_2 are ca. 50 kJ/mol more exothermic than those with elemental fluorine. It can be used to prepare fluorides of elements in their highest oxidation states, converting Au to Au (V), Xe to XeF_6, and iodine to IF_6.

2.2.3. Chemistry of Xenon

Compounds of xenon are known in the oxidation states II – VIII, some of which are remarkably stable. Xenon (II) fluoride is even commercially available. The most important xenon compounds and some of their properties are listed in Table 11. The cation Xe^+ has been observed in the solid state, whereas Xe^{2+} is observed only in solution in SbF_5 [138].

Xenon Halides. The ions Xe_2F^+, XeF^+, $Xe_2F_3^+$, $Xe_2F_5^+$, XeF_3^+, and $Xe_2F_7^+$ have been detected during mass-spectroscopic studies of xenon fluorides. The compounds $XeCl_2$, $XeCl_4$, and $XeBr_2$ were detected as highly unstable products of the β-decay of $^{129}ICl_2$ and $^{129}ICl_4$. Also, $XeCl_2$ has been detected as a component of the low-temperature condensate obtained after electrical discharge through a mixture of xenon and chlorine.

Table 11. Important xenon compounds

Oxidation state	Compound	Form	*mp*, °C	Structure	Comments
II	XeF_2	colorless crystals	129	linear	hydrolyzes to $Xe + O_2$; very soluble in liquid HF
IV	XeF_4	colorless crystals	117	square planar	stable, $\Delta H_B^{298\circ} = -284\,kJ\,mol^{-1}$
VI	XeF_6	colorless crystals	49.6	complex	stable, $\Delta H_B^{298\circ} = -402\,kJ\,mol^{-1}$
VI	$CsXeF_7$	colorless solid			decomp. $> 50\,^{\circ}C$
VI	Cs_2XeF_8	yellow solid		archimed. antiprism [a]	stable up to $400\,^{\circ}C$
VI	$XeOF_4$	colorless ligand	−46	octahedral [b]	stable
VI	XeO_2F_2	colorless crystals	31	tbp	metastable
VI	XeO_3	colorless crystals		tetrahedral [b]	explosive: $\Delta H_B^{298\circ} = +402\,kJ\,mol^{-1}$; hygroscopic, stable in solution
VI	$n\,K^+[XeO_3F^-]_n$	colorless crystals		qpy F-bridges	very stable
VIII	XeO_4	colorless gas		tetrahedral	explosive
VIII	XeO_6^{4-}	colorless salt		octahedral	the anions $HXeO_6^{3-}$, $H_2XeO_6^{2-}$, $H_3XeO_6^-$ also exist

[a] In the salt $(NO^+)_2\,[XeF_8]^{2-}$, prepared from NOF and XeF_6.
[b] Molecule has an electron lone pair.

Xenon (II) fluoride, XeF_2, is formed in high yield in an exothermic reaction when a mixture of xenon and fluorine (Xe: F_2, 2: 1) is passed through a nickel tube at 400 °C and 2 bar. The XeF_2 is condensed at − 50 °C.

$$Xe + F_2 \rightarrow XeF_2 \quad \Delta H_0 = 164\ \text{kJ/mol}$$

For thermodynamic data and spectroscopic and crystallographic studies, see [120]. The molecule is linear with a bond length of 197.7 pm in the gas phase and 200 pm in the crystal. Xenon (II) fluoride is soluble in water at 0 °C up to 15 M concentration. It undergoes hydrolysis slowly in dilute acid, and rapidly in bases according to:

$$XeF_2 + 2\,OH^- \rightarrow Xe + 1/2\,O_2 + 2\,F^- + H_2O$$

Characteristic for XeF_2 is its strong oxidizing power. The oxidation potential for the reaction

$$XeF_2\ (aq) + 2\,H^+ + 2\,e^- \rightarrow Xe + 2\,HF\ (aq)$$

has been estimated to be 2.64 V. Xenon (II) fluoride acts as a fluoride ion donor towards Lewis acid metal fluorides MF_n (M = As, Sb, B, Al, or Ga), forming the complexes $[XeF]^+[MF_{n-1}]^-$. Xenon (II) fluoride forms a number of covalent derivatives, mainly of the type F – Xe – OL and Xe $(OL)_2$, where L is, for example, TeF_5, SO_2F, or $COCF_3$. Except for the stable pentafluorotellurates, most of these compounds decompose readily and must even be regarded as explosive [123].

Xenon (IV) fluoride, XeF_4, can be prepared by heating xenon with excess fluorine (molar ratio Xe: F_2, 1: 6) for several hours at 400 °C in a nickel vessel

$$Xe + 2\,F_2 \rightarrow XeF_4 \quad \Delta H_0 = -278\ \text{kJ/mol}$$

or through the reaction of XeF_2 with fluorine. Pure XeF_4 can be obtained in quantitative yield by irradiation of XeF_6 [139]. Thermodynamic, spectroscopic, and crystallographic data for XeF_4 can be found in [120]. The compound is stable in the solid state and can be stored in thoroughly dried nickel or glass vessels. The solubility in hydrogen fluoride is 0.18 mol XeF_4/kg HF; the solution is electrically nonconducting. In neutral or acidic solution, XeF_4 undergoes hydrolysis according to:

$$3\,XeF_4 + 6\,H_2O \rightarrow XeO_3 + 2\,Xe + 3/2\,O_2 + 12\,HF$$

In alkaline solutions the yield of XeO_3 is lower due to decomposition. Xenon (IV) fluoride is a stronger oxidizing and fluorinating agent than xenon (II) fluoride. It selectively fluorinates the ring in substituted aromatics such as toluene and reacts rapidly with hydrogen to give xenon and hydrogen fluoride.

Xenon (VI) fluoride, XeF_6, can be synthesized by heating a mixture of xenon and fluorine (molar ratio 1: 20) in a nickel or monel vessel to 700 °C at 200 bar. Thermodynamic data for XeF_6 are given in [120]. Solid XeF_6 is trimorphic:

$$\text{phase III (cubic)} \xleftrightarrow{-19.35\,^\circ C} \text{phase II (cubic)} \xleftrightarrow{18.65\,^\circ C} \text{phase I (triclinic)} \xleftrightarrow{49.47\,^\circ C} \text{liquid}$$

The compound is colorless below 43 °C and yellow above this temperature; the liquid is yellow and the vapor yellow-green. When XeF_6 is sublimed slightly below room temperature, the cubic phase IV is formed. Xenon (VI) fluoride reacts rapidly with quartz

$$2\,XeF_6 + SiO_2 \rightarrow 2\,XeOF_4 + SiF_4$$

and dissolves in anhydrous HF with solvolysis:

$$XeF_6 + HF \rightarrow XeF_5^+ + HF_2^-$$

The yellow solution in HF becomes colorless on cooling, is electrically conducting, and attacks all materials. Xenon (VI) fluoride also acts as a fluoride ion acceptor, as demonstrated by its reactions with alkali-metal fluorides to give hepta- or octafluoroxenate(VI) salts:

$$XeF_6 + RbF \rightarrow RbXeF_7$$

The rubidium and cesium salts are well characterized; they decompose at 20 °C and 50 °C, respectively, according to the equation:

$$2\,MXeF_7 \rightarrow XeF_6 + M_2XeF_8\ (M = Rb\ \text{or}\ Cs)$$

The rubidium and cesium octafluoroxenates are pale yellow solids and are thermally the most stable xenon compounds known, being stable up to almost 400 °C. They react vigorously with water and hydrolyze in air to give xenon-containing products with oxidizing action. Although the compound XeF_8 has been reported itappears not to exist [140].

Xenon Oxides and Oxyhalides. The following xenon oxides and oxyhalides are known:

Oxidation state	IV	VI	VIII
Oxide		XeO_3	XeO_4
Oxyhalides	$XeOF_2$	XeO_2F_2	XeO_3F_2
	$XeOF_4$	XeO_2F_4	

All xenon – oxygen compounds are synthesized from the xenon fluorides.

Xenon(VI) oxide, XeO_3, can be obtained by careful hydrolysis of XeF_6:

$$XeF_6 + 3\,H_2O \rightarrow XeO_3 + 6\,HF$$

or through disproportionation of XeF_4 or $XeOF_4$

$$3\,XeF_4 + 6\,H_2O \rightarrow Xe + 2\,XeO_3 + 12\,HF$$

Colorless crystals of XeO_3 are obtained by evaporating the solution. With a standard heat of formation of 402 kJ/mol, the compound is highly unstable; it is an explosive with an explosive power comparable to that of TNT. The crystals contain isolated XeO_3 molecules with a pyramidal structure, a Xe – O bond length of 176 pm, and a O – Xe – O bond angle of 103°. Xenon (VI) oxide is strongly hygroscopic. Solutions in pure water are stable; xenon concentrations up to 11 mol/L have been attained. Aqueous solutions of XeO_3 are colorless, weakly acidic, and possess a very low conductivity. They contain unchanged XeO_3 and small amounts of H_2XeO_4. The addition of alkali leads to formation of xenate (VI), $HXeO_4^-$, which slowly hydrolyzes to give xenon(VIII) and xenon:

$$XeO_3 + OH^- \rightarrow HXeO_4^-$$

$$2\,HXeO_4^- + 2\,OH^- \rightarrow XeO_6^{4-} + Xe + O_2 + 2\,H_2O$$

Aqueous solutions of XeO_3 are strongly oxidizing [E_o (acid) = + 2.10 V, E_o (alkaline) = + 1.24 V].

Xenon(VIII) oxide. The slow action of sulfuric acid on xenonates, preferably Ba_2XeO_6, yields gaseous XeO_4, which can be condensed at – 196 °C as a pale yellow solid and purified by careful sublimation at – 70 °C. The melting and boiling points of the compound are unknown. The vapor pressure at – 35 °C is 0.4 kPa and at 0 °C, 3.3 kPa. Xenon (VIII) oxide is highly unstable (ΔH_f 642.5 kJ/mol) and extremely explosive. Even at – 40 °C, it can explode unexpectedly. The XeO_4 molecule is tetrahedral and has a Xe – O bond length of 173.6 pm.

Aqueous solutions of xenon (VIII) can be obtained by disproportionation of xenate (VI) or by passing ozone through dilute alkaline solutions of XeO_3. Stable, insoluble perxenate salts such as $Na_4XeO_6 \cdot 8\,H_2O$, $Na_4XeO_6 \cdot 6\,H_2O$, or $Ba_2XeO_6 \cdot 1.5\,H_2O$ can be precipitated. The barium salt can also be obtained by oxidizing XeO_3 with ozone in a solution of barium hydroxide. Perxenate solutions are strong oxidizing agents. They are alkaline due to hydrolysis:

$$XeO_6^{4-} + H_2O \rightarrow HXeO_6^{3-} + OH^-$$

Apart from $H_2XeO_6^{2-}$, the predominant species in solution is $HXeO_6^{3-}$. Free perxenic acid cannot be isolated because it readily loses O_2, forming xenic (IV) acid.

Xenon oxide difluoride, $XeOF_2$, is synthesized by careful hydrolysis of XeF_4 at – 80 °C. It is a pale yellow solid that is metastable up to ca. – 25 °C.

Xenon dioxide difluoride, XeO_2F_2, and *xenon dioxide tetrafluoride,* XeO_2F_4, can be prepared by careful hydrolysis of XeF_6, by reaction of SiO_2 with XeF_6, or by comproportionation of XeF_6 with XeO_3. The former forms pale yellow crystals that are sensitive to hydrolysis, while the latter is a colorless, readily hydrolyzed liquid.

The oxide fluorides of octavalent xenon are highly unstable. The reaction of XeO_4 with XeF_6 affords XeO_3F_2 and $XeOF_4$; however, the latter has only been detected by mass spectroscopy. Thus, apart from the stable perxenates, the + 8 oxidation state of xenon is not readily accessible.

Other Xenon Compounds. The synthesis of a compound with a xenon – nitrogen bond was achieved in 1973. The reaction of XeF_2 with $HN(SO_2F)_2$ in CF_2Cl_2 gave $FXeN(SO_2F)_2$. The unstable compound $Xe(CF_3)_2$ has been reported to be formed by the action of CF_3 radicals on xenon [141]. The presence of a xenon – carbon bond in the $[C_6F_5Xe]^+$ ion has been proved by an X-ray structural analysis on $[CH_3CN - Xe - C_6F_5]^+$. Species of this type are obtained by reacting XeF_2 with $B(C_6F_5)_3$ at – 40 °C in a noncoordinating solvent [142].

2.2.4. Radon Compounds

Since the first ionization energy of radon is lower than that of xenon (see Table 2), it should be the most reactive noble gas. However, ^{222}Rn has a half-life of 3.82 d and is an intensive β-emitter, so that investigations of its chemistry are extremely difficult and can only be performed as tracer experiments. Radon(II) fluoride, RnF_2, is formed by direct fluorination of radon or through reaction with fluorinating agents such as O_2F_2, KrF_2, and ClF_3. Since the resulting radon compound sublimes only at 230 – 250 °C, it is assumed that radon fluorides are ionic (e.g., $Rn_2^+F_2^{2-}$). Cationic radon (probably RnF^+ or Rn^{2+}) has been detected in $C_2F_3Cl_3$ solution by ion exchange on a dry cation-exchange resin [143]. The hydrolysis of RnF_2 affords elemental radon. The fluorination of radon in BrF_5 as solvent in the presence of NiF_2 as catalyst at 250 °C gives a higher radon fluoride (RnF_4 or RnF_6), hydrolysis of which affords RnO_3 [144]. Like xenon, radon forms complex salts, which presumably have the general formula $RnF^+MF_6^-$ (M = Sb, Bi).

2.2.5. Molecular Structures of Noble Gas Compounds

The structures of the fluorides, oxyfluorides, and oxides of xenon follow the rules of the valence shell electron pair repulsion model (VSEPR) [145]. When the lone pairs are taken into account, XeF_2, $XeOF_2$, and XeO_2F_2 have trigonal bipyramidal structures with a linear F – Xe – F axis in the gas phase. In XeF_4 and its oxyfluorides, the four fluorine atoms occupy equatorial positions, while the electron lone pairs or oxygen atoms occupy the axial positions of the octahedral structure.

The structure of XeF_6 has been the subject of controversy [146]. The structure of monomeric XeF_6 in the gas phase is almost certainly not octahedral. Most likely it is a structure in which the nonbonding electron pair occupies an edge or vertex of the octahedron, distorting the structure. Since the structure is highly fluxional, the atoms around the xenon atom can undergo rapid rearrangement, so that on average the xenon atom is almost spherically surrounded (see Fig. 8 A) [147].

The structures in solution and in the solid state are shown in Figures 8 B and 8 C, respectively. The reaction of binary xenon fluorides with fluoride acceptors gives cations of the type XeF^+ or $Xe_2F_3^+$. Their salts often do not have simple ionic structures; in many cases fluoride bridging between the ions occurs, as exemplified by the structure of $XeF \cdot RuF_6$ (Fig. 9 A) [148]. Figure 9 B shows the crystal structure of a covalent oxygen-bonded derivative of XeF_2 [149].

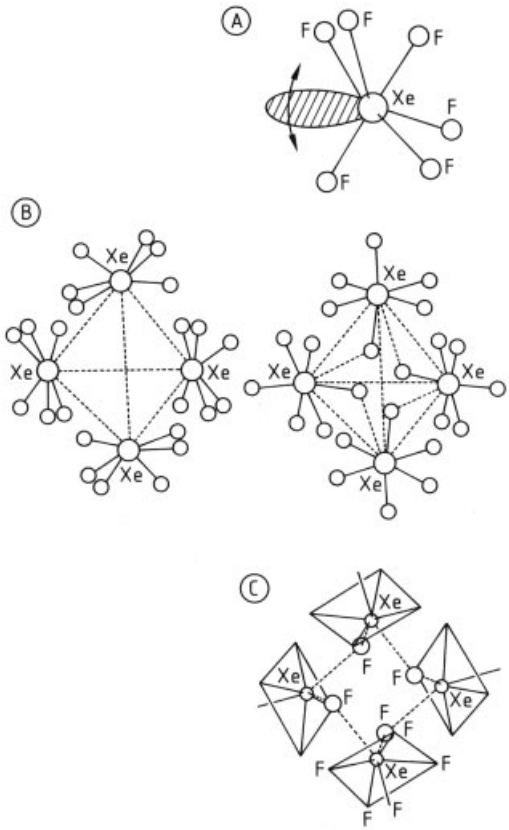

Figure 8. Molecular structure of XeF_6 [147]
A) In the gas phase: dynamic, distorted octahedral; B) In solution: In nonbridging solvents (CF_2Cl_2, CF_3Cl, SO_2ClF, F_5SOSF_5), XeF_6 forms Xe_4F_{24} tetramers by fluoride bridging, and perhaps also through weak Xe – Xe interactions. The 24 F atoms are magnetically equivalent, suggesting that they are involved in a rearrangement process; C) In the crystal: The tetrameric Xe_4F_{24} units are "frozen" in the solid state and are best described as $(XeF_5^+F^-)_4$

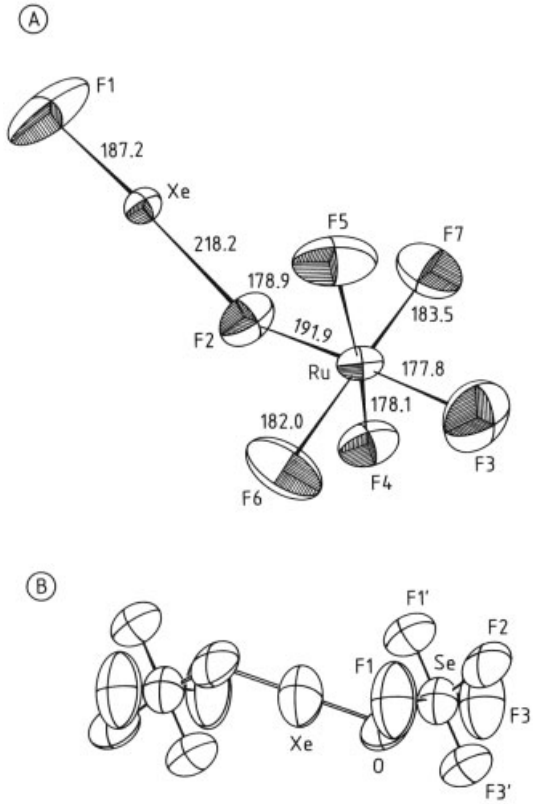

Figure 9. Crystal structures of xenon compounds
A) $[XeF^+][RuF_6^-]$; B) $Xe(OSeF_5)_2$

2.2.6. Noble Gas Clathrate Compounds

Clathrate compounds are nonstoichiometric compounds in which noble gas atoms are held by van der Waals forces in the cavities in the crystal lattice of a host substance [150–152]. Host molecules include water, heavy water, hydroquinone, and phenols. The stability of the clathrate compounds increases with increasing polarizibality of the noble gas atoms; helium does not form clathrates. Characteristic data for some noble gas clathrates is listed in Tables 12 and 13 [120]. The hydroquinone clathrates of krypton and xenon are stable in air and undergo no change when powdered. When the hydroquinone framework is destroyed by solvents such as water, methanol, or ether, the noble gases are released violently. The clathrates also decompose upon heating to 130 °C. The hydroquinone clathrates can be stabilized by irradiation with X-rays [153].

The structures of the clathrates have been the subject of numerous investigations. Figure 10 shows a section of the hydroquinone cage structure with the formula $3\,C_6H_4(OH)_2$ [154]. Hydroquinone clathrates have been used for the purification and storage of the heavier noble gases. The clathrates of ^{85}Kr and ^{133}Xe are among the safest β- and γ-sources, respectively. Clathration of ^{85}Kr has been suggested as a method for trapping and storing radioactive fission gases [155].

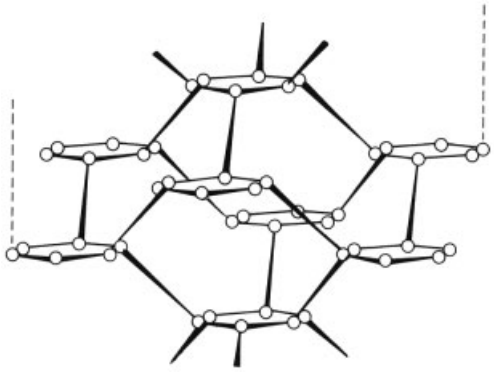

Figure 10. Mode of cage formation of hydroquinone molecules
The hydroxyl groups are arranged in hydrogen-bonded hexagons. The longer lines connecting the hexagons denote the O – O axis and the benzene ring of the hydroquinone molecule

3. Occurrence and Availability

Helium is the second most abundant element in the universe after hydrogen and accounts for more than 25 mol % of the total cosmic material. In contrast, the other noble gases, particularly the lighter ones, are rare in the solar system.

3.1. Terrestrial Occurrence

The content of noble gases in the atmosphere is listed in Table 14. The composition of the atmosphere with respect to the noble gases has presumably been constant for the last 10^9 years and will remain so for the foreseeable future; only the content of argon is slowly increasing. There

Table 12. Some thermodynamic properties of noble gas hydrates and deuterohydrates with the theoretical composition $8\,X \cdot 46\,H_2O$ or $8\,X \cdot 46\,D_2O$ (X = noble gas) [109] [a]

Noble gas	Composition (found)	Dissociation pressure at 0 °C, MPa	Decomposition temperature at 100 kPa	Enthalpy of formation (kJ/mol) from gaseous noble gas (101.3 kPa) and ice	H_2O_{liq} or D_2O_{liq}
Ar	$\leq 7.99\,Ar \cdot 46\,H_2O$	9.67	230	12	50 – 55
Kr	$\leq 7.2\,Kr \cdot 46\,H_2O$	1.46	245	17 – 22	58.2
Kr	$\leq 7.2\,Kr \cdot 46\,D_2O$	15	250	31.1	68.2
Xe	$\leq 7.6\,Xe \cdot 46\,H_2O$	0.154	269	24.3	64.9 ± 0.3
Xe	$\leq 7.6\,Xe \cdot 46\,D_2O$	1.4	270	33.9	72.1
Rn		0.04	273.5 – 273.9		

[a] Hydrates and deuterohydrates of He and Ne are unknown. Estimated decomposition pressures: He hydrate 7.6×10^4 MPa, Ne hydrate 1.6×10^3 MPa.

Table 13. Composition and decomposition pressure of organic noble gas clathrates [109]

Clathrate	Temperature, °C	Ar	Kr	Xe
Hydroquinone clathrates				
Composition				
theoretical		$Ar \cdot 3\,C_6H_4(OH)_2$	$Kr \cdot 3\,C_6H_4(OH)_2$	$Xe \cdot 3\,C_6H_4(OH)_2$
found		$\leq 1\,Ar \cdot 3\,C_6H_4(OH)_2$	$\leq 0.9\,Kr \cdot 3\,C_6H_4(OH)_2$	$\leq 0.96\,Xe \cdot 3\,C_6H_4(OH)_2$
Decomposition pressure, kPa	25	340	40	5.9
Phenol clathrates				
Composition				
theoretical		$Ar \cdot 3\,C_6H_5OH$	$Kr \cdot 3\,C_6H_5OH$	$Xe \cdot 3\,C_6H_5OH$
found		$\leq 0.73\,Ar \cdot 3\,C_6H_5OH$	$\leq 0.54\,Kr \cdot 3\,C_6H_5OH$	$Xe \cdot 3\,C_6H_5OH$
Decomposition	0	3.3	0.6 – 1.0	ca. 0.1
pressure, MPa	25	13.8	2.0	0.3
	40	33.9	5.6	0.66

Table 14. Contents of the noble gases in dry tropospheric air

Gas	Content in atmosphere, mol %	Content in atmospheric noble gas mixture	
		mol %	wt %
^3He	6.8×10^{-10}		
^4He	0.000524	0.056	0.0056
Ne	0.0018	0.193	0.098
Ar	0.9300	99.738	99.869
Kr	0.000114	0.012	0.025
Xe	0.0000086	0.0009	0.003
Rn	$(0.06 – 18) \times 10^{-23}$		
Total	0.932443	100.00	100.00

are no mechanisms known that bind noble gases. The small amounts isolated from the atmosphere are eventually released to the atmosphere. Although, with the exception of argon, the noble gases are only trace components of the atmosphere, they are present in almost inexhaustible quantities (He, 3.68×10^9 t; Ne, 6.46×10^{10} t; Ar, 6.79×10^{13} t; Kr, 1.59×10^{10} t; Xe, 1.97×10^9 t).

Air layers close to the ground, particularly those above extensive water bodies, have lower contents, especially of the more soluble, heavy noble gases, than the free atmosphere [156]. For the same reason, air dissolved in rain, rivers, seawater, groundwater, and in the bodily fluids of animals and plants is enriched in the heavier noble gases relative to the free atmosphere. The composition of the atmospheric noble gas mixture is constant up to an altitude of 59 km, above which the content of helium and neon increases while that of the heavier noble gases decreases. Helium and hydrogen are the major components of the thermosphere and ionosphere between 300 and 600 km.

Most terrestrial helium is generated by α-decay of uranium and thorium in the earth. A stationary equilibrium exists between helium gen-

Table 15. Helium contents of helium bearing natural gases and some estimated reserves

Location	Field site	Composition, mol %				Estimated
		He	C_mH_n	N_2	CO_2	reserves (1984), m^3
United States						
Wyoming	Tip Top Field	0.4 – 0.8				1.2×10^9
New Mexico, San Juan	Beautiful Mountain	4.05	49	45	0.9	
Alaska, North Slope	South Barroweast	2.54	90.2	6.8	0.3	
Texas, Young	Young Regular	1.17	66.2	31.1	0.1	
Kansas, Grant	Hugoton (Compr. station)	0.44	81.8	17.6	trace	
Canada						
Alberta	Worsley	0.53	93	6.0	0.5	1.27×10^9
Ontario	Norfolk	0.36	91.5	8.1	0.1	
The Netherlands	Groningen	0.05	8	14	0.9	
	De Wijk	0.05	94.3	5.6	trace	
Poland	Ostrow	0.4	56	43	0.3	
North Sea, British sector	Indefatigable	0.05	96	3.3	0.5	
Germany, Niedersachsen	Apeldorn	0.12	25.9	73.3	0.2	
Former Soviet Union	Urengoi	0.055	94.2	5	0.03	6×10^9
	Orenburg					3.4×10^9
Orange Tree State	O.F.S. Goldfields	2.91	86.8	9.9	trace	
Algeria	Hassi R'Mel	0.19	93.6	5.8	0.2	2.4×10^9
Australia, Northern Territory	Palm Valley	0.21	97.5	2.3	0.1	0.9×10^9

erated in the earth's crust, helium released from the oceans [157], helium captured from the solar wind, and that which escapes to space from the atmosphere. The loss of helium is estimated at 5×10^8 t/a, about 50 % of the annual world requirement [158]. The total terrestrial inventory of helium (atmosphere, hydrosphere, and lithosphere) is estimated to be 900×10^9 t [159]. The isotopic ratio $^3He/^4He$ of terrestrial helium is ca. 10^{-6}. Therefore, air is uninteresting as a source of ^3He, and this isotope is synthesized by the nuclear reaction $^6Li\ (n, \alpha) \rightarrow {}^3He\ (\beta)$.

Air is the only source for the industrial production of neon, argon, krypton, and xenon. These gases, and sometimes also helium, are byproducts of air rectification and ammonia synthesis. Tetrafluoromethane, which is present in the atmosphere in trace amounts, has similar thermodynamic properties to krypton and is therefore concentrated in high-purity krypton, in which it was first detected in 1974 [160].

Natural Gas. Natural gas contains varying amounts of helium, ranging from negligible to over 0.8 mol %. Natural gas that contains > 0.3 mol % He is termed helium-rich. Helium contents of natural gases are listed in Table 15 [161–163]. The occurrence of helium in natural gas depends on the presence of particular geological conditions. First, the country rock and sedimentary rocks must have a high content of uranium and thorium and overlying porous rock to act as a gas reservoir. This gas reservoir must in turn be covered by a gas-tight, tectonically undisturbed layer. The helium is presumably carried into the reservoir by nitrogen and methane from neighboring oil fields or is desorbed by convection cells in hydrothermal fields. The helium content increases with the geological age of the formation. Most of the helium-containing fields lie at a depth of about 2500 m. It is believed that, independent of temperature and pressure, above a critical depth of 2100 m insufficient coverage is present to retain helium reserves [164].

The economically exploited helium reserves are concentrated in the United States (see Table 16 [165]). Most of them are clustered in the midcontinent and Rocky Mountain regions, including Cliffside Field, Texas; Hugoton Field, Kansas; Keyes Field, Oklahoma; Panhandle Field, Kansas, Oklahoma, Texas; and Tip Top Field, Wyoming. Total U.S. helium resources in natural gas are estimated by DOI at 1.2×10^{10} m^3. Notable helium fields outside the United States are Worsley Field, Alberta, Canada; Hassi R'Mel Field, Algeria; Orenburg Field, former Soviet Union; Sichuan Field, China; and Ostrow Field, Poland.

Table 16. World helium resources as of January 1996 (in 10^6 m^3) [165]

Country of region	$m^3 \times 10^6$	Share, %
United States		
Stored in Cliffside field	950	
Helium in helium-rich natural gas	5800	
Helium in helium-lean gas	5100	
Total (U.S.)	11 850	43.5
Former Soviet Union	8900	32.7
Algeria	2000	7.3
Canada	2000	7.3
China	1000	3.7
Poland	800	2.9
Netherlands	672	2.5

The content of the other noble gases in natural gas varies. The concentrations of argon, krypton, and xenon, relative to the accompanying nitrogen, are mostly lower than in air. Xenon from natural gas is enriched in the lighter isotopes compared to atmospheric xenon.

U.S. Helium Act Amendment. Helium is lost when natural gas is burned. These losses amounted to ca. 2×10^9 mol in the United States in 1972. Due to its rareness and use in certain strategically important applications, helium has been classified as a strategic resource by the U.S. Government. To ensure that U.S. helium requirements would be met by helium from natural gas up to and beyond the year 2000, the Helium Conservation Programme was introduced in 1960. This planned the extraction of ca. 7.5×10^{10} mol of helium from natural gas in the form of 70 % raw helium by 1986. The helium was stored in empty gas fields. The resulting costs were supposed to be covered by the sale of pure helium by the U.S. Bureau of Mines. The Bureau signed 22-year contracts with four private firms, which thereupon built helium extraction plants and together sold 3.5×10^9 mol of helium to the Bureau at a fixed price of \$ 28/kmol. By 1973, ca. 4.8×10^{10} mol of helium had been stored at Cliffside Field. Because of increasing deficits caused by private companies offering lower prices, the government cancelled the entire program under the announcement, that "the purpose of Helium Act Amendment has been accomplished."

In the mid-1970s, the Bureau of Mines began accepting privately owned crude helium for storage in the Cliffside gas field under long-term contracts. This privately owned crude helium is returned to owners as needed for purification to supply private demand. For example, during 1996, 35×10^6 m^3 of private helium was delivered to the storage system, and 20.6×10^6 m^3 was withdrawn. Then, in 1996 helium operations were transferred to the Bureau of Land Management (BLM) after the Bureau of Mines was closed. In October the *Helium Privatization Act* was signed. The legislation specifically calls for the Federal Government to exit the helium business, to begin selling its crude helium reserves by or before 2005, and to complete those sales by 2015. A Helium Reserve Committee presently is charged with estimating the effect of the sale of federal helium reserves on market and industry [166].

Hot Springs. Many hot springs in regions of current or former volcanic activity are rich in helium and neon. They contain up to 10 % helium, 1.2 % neon, as well as argon, some krypton, xenon, and traces of radon. The Ne/N_2 ratio is generally lower than that of air. Gases in the highly concentrated thermal brines of Ruhrkarbon (Central Europe) contain 10 – 15 vol % methane, 10 – 22 % carbon dioxide, 60 – 70 % nitrogen, 4 – 5 % helium, smaller amounts of neon, argon and krypton, and some radon. The average rate of gas production of these sources is ca. 1 kmol/d. Nitrogen-rich, noble-gas-containing gases are also found in coal mines. The gas quantities produced are small, typically 5 mol/h. These pit gases contain only traces of noble gases, which originated mostly from air dissolved in the sap of the plants from which the coal was formed.

Minerals. Because of the wide distribution of α-emitting isotopes, numerous minerals and rocks, particularly primitive rocks, contain helium. High helium contents (mmol He/kg) have been found in thorianite (150 – 470), cleveite (35 – 350), monazite (35 – 110), broeggerite (45), fergusonite (45 – 80), and pitchblende (ca. 12). Some minerals contain more helium than can be accounted for by their content of radioactive elements, for example, gadolinite (up to 110), beryl (up to 0.5), and sylvine (up to 2.5×10^{-4}). Argon occurs as an accompanying element in all potassium minerals. The heavy noble gases are rare in terrestrial minerals and occur in extremely small amounts. For example, a xenon content between 4×10^{-8} and 5×10^{-9} mmol/kg has been found in granite. In

comparison to atmospheric xenon, xenon in minerals is enriched in the heavier isotopes.

3.2. Occurrence of Radioactive Isotopes

A new source for stable and unstable isotopes of krypton and xenon are nuclear reactors and nuclear fuel reprocessing plants. Krypton isotopes with mass numbers 83 – 86 and xenon isotopes with mass numbers 131 – 136 accumulate as fission products in fuel rods at a rate of ca. 0.01 mol/h Kr and ca. 0.1 mol/h Xe per 1000 MW. They are released when the fuel rods are reprocessed [167]. Of the isotopes formed, it is mainly ^{85}Kr (half-life 10.86 a) that poses a radiological hazard. The others are short-lived nuclides with half-lives of at most a few days. The noble gas content of 1 t of fuel after a burnup of 33 GW · d/t and a cooling time of one year is 1 3 Xe and 0.13 Kr, of which 0.6 % is ^{85}Kr and the rest stable isotopes [168].

The noble gases are distributed rapidly in the entire atmosphere; apart from their low solubility in the oceans, they have no sinks. The ^{85}Kr activity in the atmosphere has been monitored since 1965. Measuring stations are located at Freiburg (FRG), Cracow (Poland), and Miami (USA). A steady increase in atmospheric ^{85}Kr concentration on a global scale is evident in Figure 11. The rate of increase is about 25 mBq/m^3 per year (ca. 3 %). The ground level in the northern hemisphere is, however, small, and amounted to 1.3 Bq/m^3 of air in 1998 [169].

The naturally occurring isotopes of radon are short-lived intermediates in the radioactive series of uranium, thorium, and actinium and are steadily generated and in decay. After plutonium, radon is the rarest element in the earth's crust with an estimated abundance of 6.2×10^{-16}%. Traces of radon are found in all radioactive material, particularly minerals. Radon readily escapes from the sites where it was formed and is disseminated in water and air. The concentration of radon, which is mainly ^{222}Rn, in lower atmosphere layers is 2 – 20 Bq/m^3, and in potable water 1 – 10 Bq/L with a maximum of up to 200 Bq/L. High radon concentrations are observed in spring waters near radioactive deposits such as those at Hot Springs, Arkansas or in the air of poorly ventilated mining tunnels.

The release of radon increases with the formation of surface fissures during geological events such as earthquakes and volcanic eruptions. Therefore, radon is permanently monitored in seismic regions as a useful instrument to forecast earthquakes. An activity estimated at 1.1×10^{17} Bq of radon was released during the eruption of the Mount St. Helens volcano in 1980.

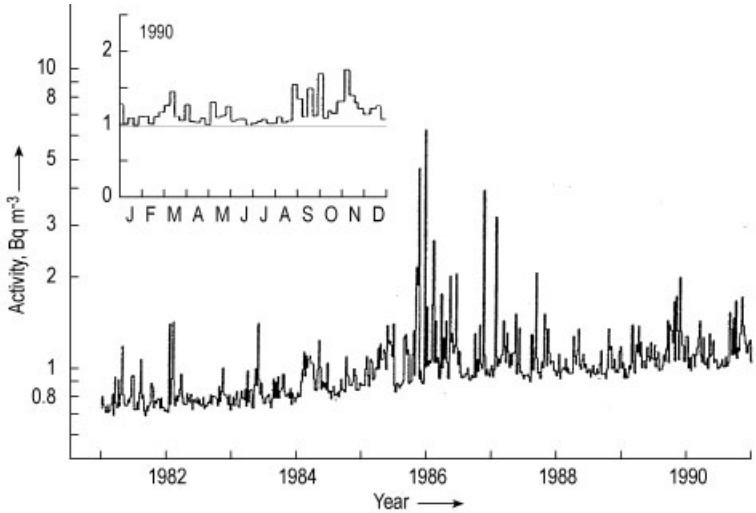

Figure 11. Increase of the ^{85}Kr concentration in the atmosphere from 1981 to 1991 [169] Measuring stations: Freiburg (points) and Schauinsland (histogram), Germany

3.3. Lunar and Solar Occurrence

Moondust and gas-rich meteorites contain relatively large amounts of hydrogen and noble gases. The gases, which apparently were adsorbed from the solar wind, are found mostly in thin layers on the surfaces of the minerals. Typical gas contents of moon dust in mmol/kg are: hydrogen, 45; helium, 9; neon, 0.09; argon, 0.02; krypton, 9×10^{-6}; and xenon, 4.5×10^{-6}. Moon rock that has not been loaded by the solar wind also contains small amounts of noble gases [171–173].

The relative abundances of the noble gases with respect to one another and the distribution of abundances of the isotopes of the individual gases vary according to origin. These variations allow conclusions to be made about the origin of the elements and, therefore, the evolution of the solar system. Thus, for example, investigation of the solar-wind noble gases in moon dust has given accurate information on the abundances of the noble gases in the sun. The xenon isotopes are particularly important in these studies because their mode of formation is well known and they can be determined quantitatively with high accuracy (xenology) [173].

4. Extraction and Isolation

The sources for industrial production of noble gases are nowadays exclusively air and natural gas, or else residual gases from processes where air or air products and/or natural gas are processed, such as ammonia synthesis. The concentrations of the noble gases in these sources are low, which has two consequences as far as their separation is concerned:

1) The reversible, i.e., minimum, specific work for separating a unit quantity of noble gas is relatively high compared with other industrial separations. It is shown in Figure 12 for two different sets of conditions, namely for infinitely large gas throughput corresponding to Equation (1) and for complete separation of the noble gases, i.e., 100 % yield, in accordance with Equation (2).

$$W_{\mathrm{rev},\infty} = -RT_{\mathrm{a}} \ln x_0 \qquad (1)$$

$$W_{\mathrm{rev},100} = -RT_{\mathrm{a}} \left[\ln x_0 + \frac{1 - x_0}{x_0} \ln (1 - x_0) \right] \qquad (2)$$

where T_{a} is ambient temperature.

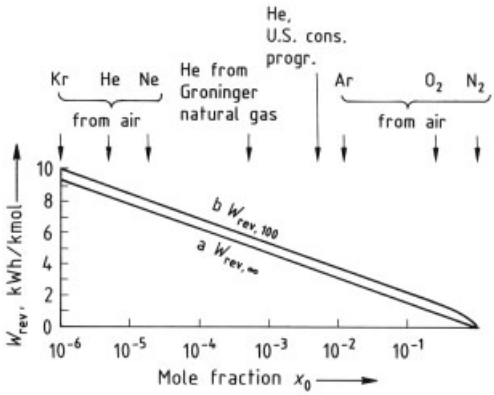

Figure 12. Reversible work for separation of ideal mixtures as functions of initial concentration x_0
a) For infinitely large gas throughput (Eq. 1); b) For complete separation of the noble gases (Eq. 2)

2) The actual separation work is much higher than the reversible work and increases relatively sharply with falling initial concentration of the component to be extracted because of the wasteful simultaneous treatment of large quantities of accompanying components. Because of these energy considerations and because of the high specific capital costs involved with low initial concentrations the noble gases are always produced as byproducts from fractionation plants for oxygen and ammonia synthesis residual gas, where they have already become enriched during the separation of the main products. Table 17 shows the concentrations in various feed gases and the quantities obtainable in the various processes.

The demand for noble gases, especially for argon, krypton and xenon, has risen continuously over the years. More and more plants for fractionating air are therefore being fitted with auxiliary equipment for separating noble gases.

Table 17. Concentration and quantities of noble gases in various sources

Gas	Concentration (mL/m^3) in				Potential production rate, mol/h		
	Air	Groninger natural gas	NH_3 synthesis gas	Synthesis residual gas	1000 t/d O_2 plant	1000 t/d NH_3 plant	1000 MW ^{235}U reactor
He	5.24	~400	~90	~1100	33	490	
Ne	18.18	trace	6	70	115	31	
Ar	9320		3000	34 000	5910	15 600	
Kr	1.14		0.4	4	7.2	2.0	0.02
Xe	0.086		0.03	0.3	0.55	0.14	0.19

4.1. Isolation From Air

Separation of noble gases from air is described here based on the example of an oxygen plant using the low-pressure process. A schematic of the process is shown in Figure 13. Atmospheric air is compressed to approximately 6 bar and freed from H_2O, CO_2, and hydrocarbons in molecular sieve adsorbers. The majority of the air is cooled to almost dew-point temperature in countercurrent flow to nitrogen and oxygen and fed to the bottom of the medium-pressure column. After the molecular sieve adsorbers a smaller quantity of the air is compressed in a booster, which is coupled directly to the expansion turbine, and is also cooled in countercurrent flow to nitrogen and oxygen; it is then expanded in an expansion turbine to the pressure of the low pressure column and injected into it.

4.1.1. Helium and Neon

The concentration of helium in air is ca. 5 mol ppm and that of neon is 20 mol ppm. Air represents the only economically exploitable source of neon, but helium is only extracted from air in exceptional cases as many natural gases contain helium in concentrations which are a hundred to a thousand times higher.

In the distillative fractionation of air as shown in Figure 13, the noble gases helium and neon become concentrated in the reflux condenser of the medium-pressure column together with the hydrogen contained in the air. In noble gas production there is an enrichment of up to about a thousand times the input value. Further concentration of the helium and neon to ca. 50 mol % is carried out in a small secondary column whose overhead condenser is cooled at ca. 80 K with evaporating liquid nitrogen. A fraction can be drawn from the top of this column with a typical composition of 11 mol % He, 35 mol % Ne, 5 mol % H_2 and 49 mol % N_2 with a noble gas yield of over 90 %. Further separation of this fraction is discussed below.

Helium – Neon Separation by Condensation. An older publication [174] describes a condensation method for separating helium and neon in which a helium refrigerator is required for liquefying the neon. In the Linde process shown in Figure 14 (see page 26), the helium refrigerator is replaced by an Ne – He cryoprocess with nitrogen precooling and a Joule – Thomson valve. The hydrogen and nitrogen components are removed from the crude gas before the low-temperature separation is performed. Hydrogen is converted catalytically to water on a noble metal catalyst (a_1) with the addition of oxygen. The gas, which is now saturated with moisture, is transferred to the gasholder (b_1), from where it is transferred to the high-pressure reservoir (b_2) by the compressor (c_1) for intermediate storage. Even in large air fractionation plants the crude He – Ne flow is only ca. 0.1 mol/s and is therefore too small for a continuously-operated condensation system operating on an industrial scale. The intermediate storage allows the subsequent process steps to be carried out in batches.

The crude gas from the high-pressure reservoir (b_2) is expanded to about 30 bar, freed from water by adsorption at room temperature in the drier (d_1), which is filled with molecular sieves, and then cooled to ca. 66 K in the heat exchangers (e_1) and (e_2). This condenses most of the nitrogen contained in the crude gas. The liquid phase is separated from the gaseous phase in the separator (f_1). The remainder of the nitrogen in the gas phase (< 2 mol %) is retained by the silica-gel-filled adsorber (d_2) to give a mixture

Mass balance for Figure 13: throughput [kmol/h or mol/h] and proportion [%] of the components in positions ⟨1⟩ to ⟨6⟩ and the associated thermodynamic data (bp = boiling point at 1.013 bar, α = relative volatility, $k_0 = y/x$ = distribution ratio of the components in the liquid and vapor phases at pressure p [bar] and infinite dilution).

Thermodynamic data						⟨1⟩ Air		⟨2⟩ O_2		⟨3⟩ N_2		⟨4⟩ Crude Ar		⟨5⟩ He–Ne*		⟨6⟩ Kr–Xe*	
	bp, K	α	k_0	in	p, bar	kmol/h	or %	kmol/h	%	kmol/h	%	kmol/h	%	kmol/h	%	mol/h	%
N_2	77.4	3.7/3.5		O_2	1.3	78.11				78.10	99.2	0.01	1	0.00	30		
O_2	90.2					20.96		20.53	99.5	0.41	0.5	0.02	3			536.95	99.435
Ar	87.3	1.6/1.1		O_2	1.3	0.932		0.10	0.5	0.23	0.3	0.60	96			0.537	0.1
Σ								20.63		78.74		0.63		0.00		540	
						mol/h	vol ppm	mol/h	vol ppm	mol/h				mol/h	%	mol/h	%
He	4.2		250	N_2	6	0.524	5.24			0.05				0.47	16		
H_2	20.4					0.05	0.50			0.00				0.05	2		
Ne	27.1		55	N_2	6	1.818	18.18			0.25				1.55	52		
Kr	119.8		0.12	O_2	1.5	0.114	1.14	0.01	0.5							1.08	0.2
Xe	165.0					0.0086	0.086	0.00	0.0							0.081	0.015
CH_4	111.7		0.30	O_2	1.5	0.15	1.5	0.02	ca. 1							1.35	0.25

* Concentrates.

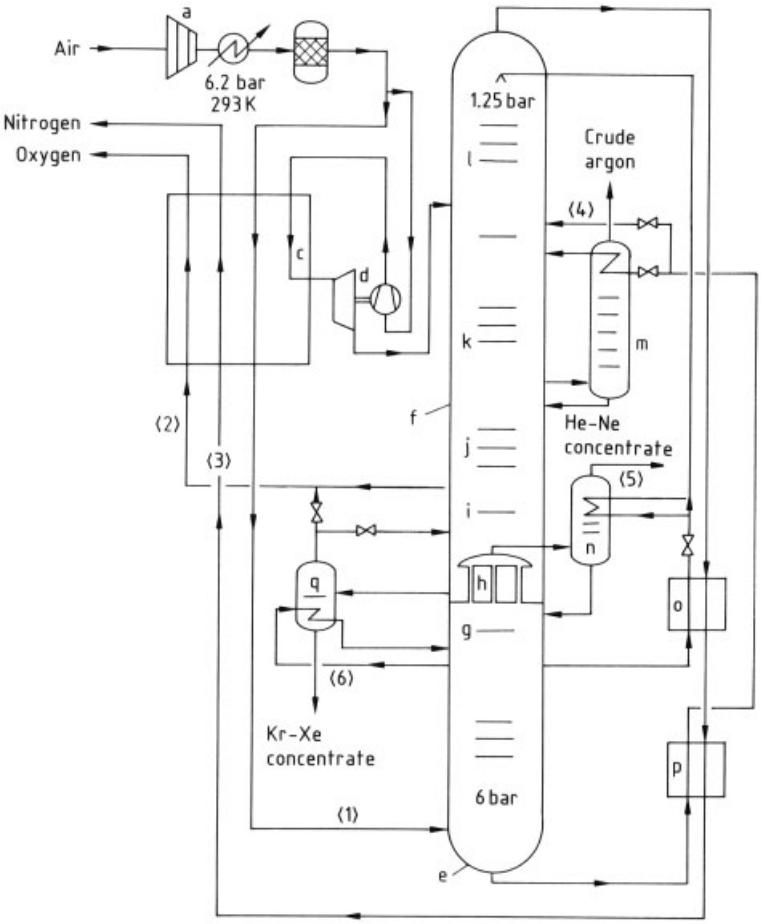

Figure 13.
Schematic of a low-pressure oxygen plant with a noble gas production system
a) Compressor; b) Molecular sieve adsorber; c) Main heat exchanger; d) Booster turbine; e) Medium-pressure column; f) Low-pressure column; g) Neon trays; h) Main condenser; i) Krypton trays; j) Oxygen section; k) Argon section; l) Nitrogen section; m) Crude argon column; n) Helium – neon concentration column; o) Wash nitrogen subcooler; p) Crude oxygen subcooler; q) Krypton – xenon concentration column.

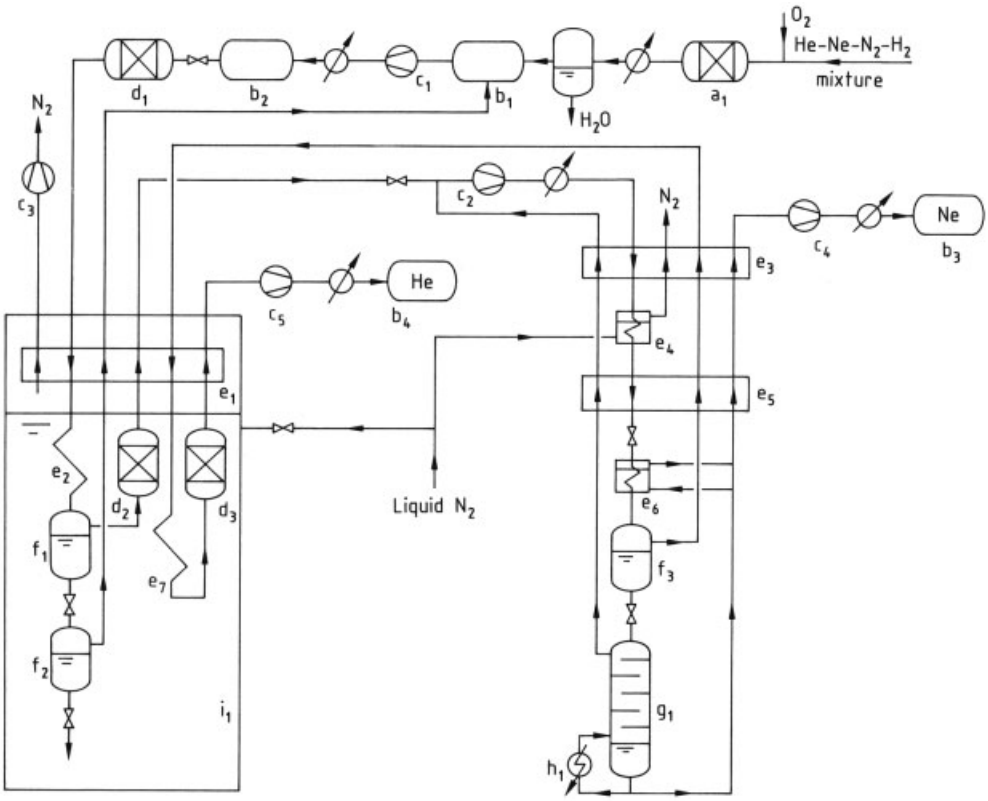

Figure 14. Linde process for production of helium
a) Catalyst; b) High-pressure reservoirs; c) Compressors; d) Adsorbers; e) Heat exchangers; f) Separators; g) Stripping column; h) Heating system; i) Nitrogen bath

of approximately 76 mol % neon and 24 mol % helium, which, after it has been heated to room temperature and expanded to ambient pressure, is transferred to the compressor (c_2).

The liquid phase from the nitrogen separator (f_1) is expanded to almost ambient pressure in the second separator (f_2), and the majority of the neon and helium dissolved in the liquid is released; it is returned to the gasholder (b_1) after heating to room temperature. After its pressure has been reduced to about 0.2 bar, the remaining condensate from the second separator (f_2) is vaporized together with the liquid nitrogen supplied as refrigerant, heated to room temperature, and released to atmosphere through the vacuum pump (c_3).

The compressor (c_2) compresses the helium – neon mixture to ca. 180 bar. The gas stream is cooled to about 80 K in the heat exchanger (e_3) and the nitrogen evaporator (e_4) and then further to about 50 K in the heat exchanger (e_5). After expansion to ca. 25 bar the majority of the neon condenses in the heat exchanger (e_6) and the heat of condensation is dissipated by pure neon evaporating at virtually ambient pressure. The condensate is separated from the gas phase in the separator (f_3). The liquid is then expanded to ca. 1.5 bar. The majority of the dissolved helium passes into the gas phase. The liquid neon is then freed from the remaining helium in the stripping column (g_1). After evaporation in the heat exchanger (e_6) the purified neon is heated to room temperature again. The diaphragm compressor (c_4) transfers the extracted neon, with an impurity level of less than 10 mol ppm, to the high-pressure reservoir (b_3).

The neon-rich gas leaving the top of the column (g_1) is heated to room temperature in the heat exchangers (e_5) and (e_3) and is fed to the suction side of the He – Ne compressor (c_2).

The gas leaving the separator (f_3) contains ca. 85 % helium and 15 % neon. This gas flows

through heat exchangers (e_5), (e_3), (e_1) and (e_7) and the neon adsorber (d_3), which contains activated carbon. After warming to room temperature, helium with an impurity level of less than 10 mol ppm can be transferred to the high-pressure reservoir (b_4) by the diaphragm compressor (c_5).

The gas produced when regenerating the adsorbers (d_2) and (d_3) is transferred to the gasholder (b_1). This process achieves a yield of helium and neon of more than 98 %.

The process described above is carried out industrially in a variety of ways. The heat exchangers and adsorbers were arranged as shown in Figure 14 for practical reasons. In principle the heat exchangers (e_1) and (e_3) can form one structural unit, (e_4) can be accommodated in the nitrogen bath (i_1), and the neon adsorber (d_3) can be positioned directly after the heat exchanger (e_5).

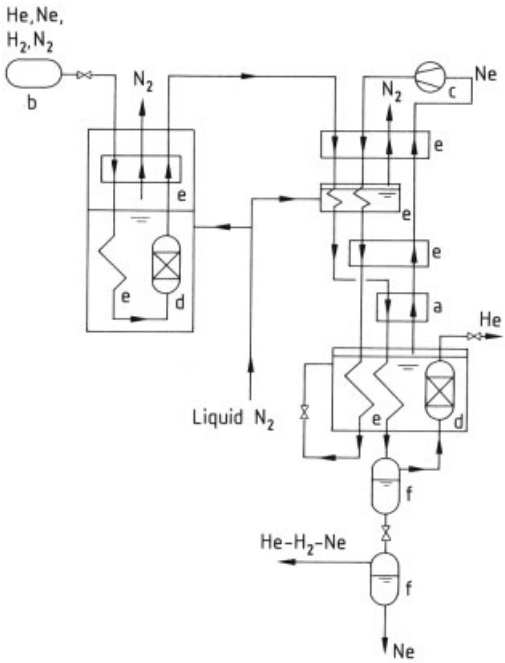

Figure 15. Leyarovski process for production of helium and neon
a) Freezing heat exchanger; b) High-pressure reservoir; c) Compressor; d) Adsorber; e) Heat exchanger; f) Separator

Figure 15 shows a process [175] which dispenses with catalytic hydrogen conversion. Nitrogen is adsorbed at 80 K and neon condenses at approximately 30 K under a pressure of 10 bar. After expansion to ambient pressure the dissolved components helium and hydrogen escape, with the result that neon can be obtained with an impurity level of less than 1000 mol ppm. The gas produced during expansion is released to atmosphere so the neon yield is only about 90 %. The fraction which is noncondensible at 30 K and 10 bar contains 80 mol % helium. After adsorption of neon on active carbon at 30 K, helium is produced with a purity of 99.8 mol %. Liquid nitrogen and a closed neon circuit with nitrogen precooling, countercurrent heat exchanger, and Joule–Thomson valve are used for cooling.

Helium – Neon Separation by Absorption. The solubility of neon in liquid nitrogen is higher than that of helium. A laboratory apparatus is described in [176] with which neon is washed out of an Ne – He – N_2 mixture with pure liquid nitrogen. This produces a gas containing ca. 90 mol % He, 3 mol % Ne, and 7 mol % N_2 and a liquid with 0.2 mol % He, 2.8 mol % Ne, and 97 mol % N_2. The additional process for preparation of pure helium and neon is only outlined.

Helium – Neon Separation by Adsorption. Neon and helium can be separated from an Ne – He – N_2 mixture by condensation and multistage adsorption by using the process shown in Figure 16 [177]. The mixture is cooled to 67 K at ca. 5 bar to condense the nitrogen. The subsequent N_2 adsorption of the residual nitrogen from the gas is performed at 77 K. Neon adsorption is carried out in two stages at 67 K. In the first stage neon is adsorbed at 5 bar, which increases the purity of the helium to 99.9 mol %. After the adsorbent is exhausted the pressure in this first adsorption stage is lowered to 3 bar at low temperature, and the gas produced is transferred to a second stage for renewed neon adsorption, again producing pure helium. By lowering the pressure in the second adsorption stage to 2 bar at low temperature ca. 8 % of the total neon is produced with a purity of 98.5 mol %. Pure neon at 99.98 mol % is liberated during the final expansion in the two adsorption stages to ambient pressure and indirect heating with air at 470 K. The neon yield is 82 % and that of helium 90 %. All adsorption stages are fitted with two reversible units.

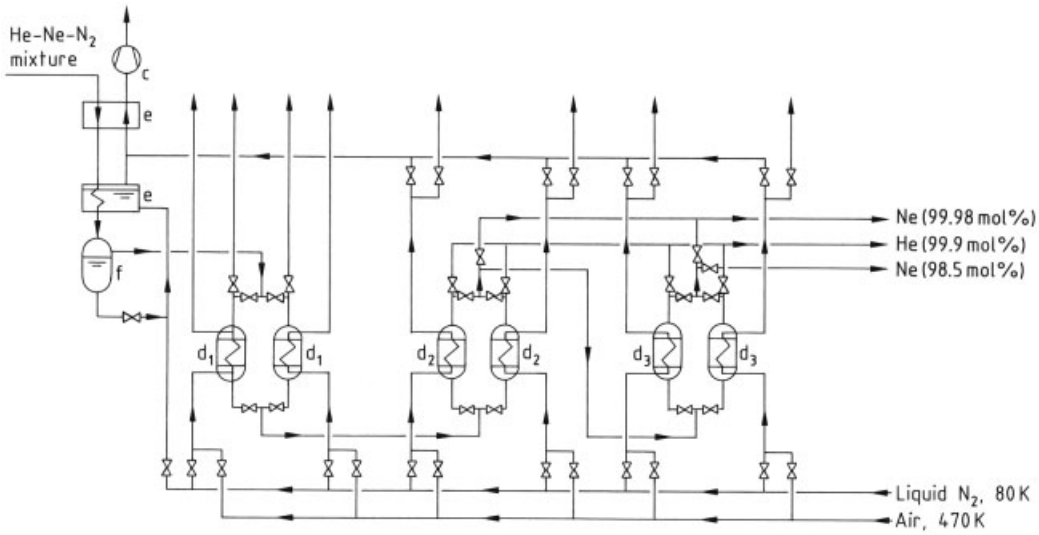

Figure 16. Leyarovski adsorption process for producing helium and neon
c) Compressor; d_1) N_2 adsorber; d_2) Ne adsorber, first stage; d_3) Ne adsorber, second stage; e) Heat exchanger; f) Separator

A more recent adsorptive method operates with a pressure-swing system at 80 K and is claimed to produce neon with impurities of less than 100 mol ppm [178].

4.1.2. Argon

The volatility of argon lies between those of nitrogen and oxygen; for example, the ratio of the volatilities of nitrogen, argon, and oxygen in the low-pressure column with little nitrogen and 15 % argon in the vapor is 3.6: 1.5: 1, cf. the mass balance in Figure 13. As a result of this equilibrium behavior, when the usual reflux ratios are applied in the low-pressure column (where pure oxygen is produced) the argon accumulates above and below the crude oxygen feed point. In the lower part of the column, argon concentrations of up to 15 % in the vapor phase can be attained.

This argon maximum– or somewhat below it to reduce the nitrogen content–is the appropriate point for drawing off the so-called argon side stream, which is then rectified in the argon auxiliary column (crude argon column, see Fig. 13). Crude argon, with 3 – 5 % O_2 and up to 1 % N_2, is drawn off at the top of this column.

Energy Balance and Argon Yield. In spite of the high liquid/vapor ratio in the column the separation of crude argon theoretically requires no additional energy. This is because the addition of the auxiliary column to the system reduces the amount of the materials processed in the low-pressure column. In practice, however, low-pressure columns with argon extraction have more trays than those without argon extraction and consequently exhibit higher pressure losses; they therefore require a higher initial feed air pressure. Against this, however, can be set the higher oxygen yield in columns with argon extraction, so the "extra" energy is not significant.

Argon yield. High argon yields require two loss terms to be minimized: the equilibrium loss that would occur if the argon content in the nitrogen being drawn off were in equilibrium with the argon content of the nitrogen reflux, and the rectification loss caused by incomplete approach to equilibrium during rectification.

Minimization of the equilibrium loss requires a small argon content in the reflux nitrogen. In air fractionation plants with oxygen as main product, no better level of purity than about 1 % O_2 in the reflux nitrogen is required. This is normally obtained with a liquid vapor ratio in the medium-pressure column of ca. 0.53. The equilibrium constant of argon at 6 bar is 0.54 with nitrogen as opposed to 0.65 in equilibrium with air, so for a liquid/vapor ratio of 0.53 there is only slow upward enrichment of the argon. Only

fairly large reflux ratios and low oxygen content in the reflux nitrogen lead to smaller argon contents. Compromises are required for the sake of a good oxygen yield.

However, a high liquid/vapor ratio in the medium-pressure column means less reflux nitrogen and therefore a lower liquid/vapor ratio in the nitrogen section of the low-pressure column. This eventually leads to increasing disruption of the argon equilibrium at the top of the low-pressure column and so to increasing rectification loss. The total amount of argon produced from the feed air quantity, i.e., the argon yield, is mainly dependent on the turbine flow to the low-pressure column (see Fig. 17).

Purification of Crude Argon. After the crude-argon column the crude argon passes through the following process stages:

1) Crude argon is compressed to 4 – 6 bar after heating to ambient temperature.
2) After addition of hydrogen the oxygen is reduced to a residual content of less than 1 vol ppm in the deoxo, a container filled with noble metal catalyst. If the crude argon has a fairly high oxygen content it must be recycled to cool the adiabatically operated deoxo.
3) The oxygen-free crude argon is dried adsorptively at ambient temperature.
4) In the subsequent pure-argon distillation column, which is heated with medium-pressure nitrogen (or crude argon, with recovery of the excess hydrogen) and is cooled with liquid nitrogen, the top product is nitrogen, the excess hydrogen and some argon, while argon with 99.9999 % purity is drawn off at the bottom, normally in liquid form for ease of storage and transport. It is necessary to ensure, by using a sufficiently high pressure ($\geq$ 2 bar) for the nitrogen evaporating in the overhead condenser, that the condensing argon (triple point 84 K) does not solidify.

In principle, it is also possible to produce a crude argon intermediate product with an oxygen content of 1 vol ppm by complete rectification of the side stream in the auxiliary column. This, however, requires a very large number of trays. To ensure that condensation is still possible by cooling with evaporating crude oxygen, the pressure loss in the argon auxiliary column must not be greater than, for example, in a tray column containing about 50 sieve trays. This is possible with a column which has random or structured packing for the separating stages instead of sieve trays. This arrangement (see Fig. 18) makes the addition of hydrogen to the crude argon and subsequent combustion of oxygen in a deoxo unnecessary [179].

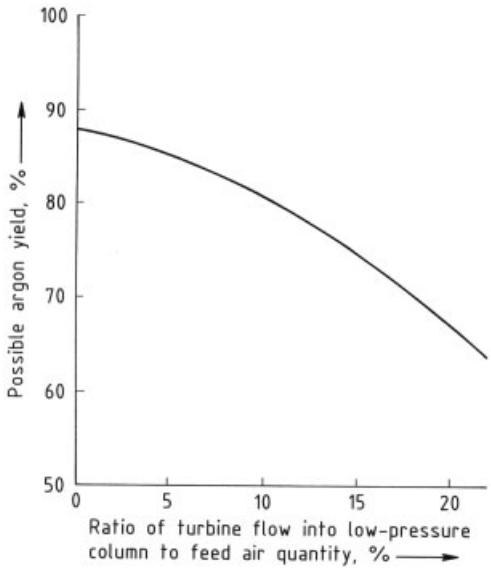

Figure 17. Argon yield as a function of the turbine flow to feed air ratio

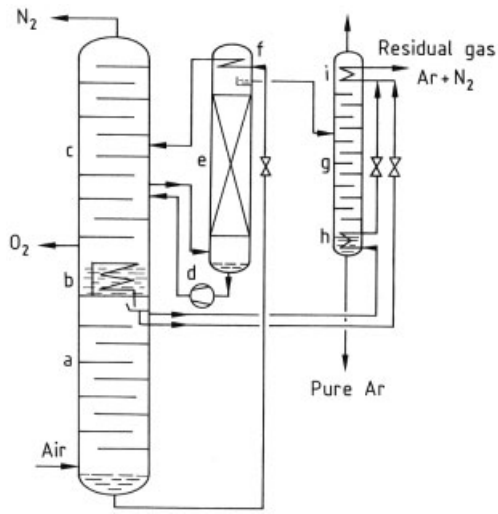

Figure 18. Production of pure argon
a) Medium-pressure column; b) Main condenser; c) Low-pressure column; d) Liquid pump; e) Crude argon column; f) Crude argon condenser; g) Pure argon column; h) Pure argon evaporator; i) Pure argon condenser

4.1.3. Krypton and Xenon

Because of their low volatilities, krypton and xenon collect in the oxygen bath of the main condenser. In plants with gaseous oxygen take-off without krypton–xenon recovery the krypton concentration is approximately 50 mL/m^3 (50 vol ppm) in gaseous oxygen. To recover the noble gases, part of the liquid oxygen from the oxygen bath is fed to the krypton – xenon concentration column, where a krypton – xenon concentrate with ca. 0.3 % Kr + Xe is produced.

Further concentration of krypton and xenon should not result in any great loss of yield. For this purpose, extra trays (krypton trays) are installed between the main condenser and the take-off point for gaseous oxygen product (see Fig. 13) to wash back krypton and xenon by oxygen reflux.

Undesirably high hydrocarbon contents must be avoided at all points in the system. The methane level in the main condenser can be regulated by the quantity of feed to the concentration column. The methane level is normally kept to a few hundred vol ppm CH_4, i.e., well below the ignition limit of 5 % CH_4, to take account of the fact that higher concentrations could occur as a result of phase changes, local concentration, and plant stoppage. In the concentration column the content of methane and other hydrocarbons, for safety reasons, must not exceed 0.5 %.

The methane level in the main condenser can be reduced, albeit with loss of krypton, by feeding the oxygen vapor formed as top product of the concentration column directly into the product oxygen, instead of into the low-pressure column.

An effective additonal measure to reduce the methane level in the oxygen bath around the main condenser is to use a methane bypass column, which is installed in parallel to the krypton barrier trays and in which gaseous product oxygen is made. The methane level in the main condenser is lowered to about one-third, with negligible krypton losses, if the column is operated with a liquid/vapor ratio of 0.2 (between $k_{Kr} = 0.12$ and $k_{CH_4} = 0.30$).

Safety while concentrating the preconcentrate can be further increased by replacing the oxygen by argon before the concentration [180]. This step, which exploits the anomalous argon – oxygen equilibrium behavior, can be integrated particularly well, and without argon loss, in plants that are also intended for argon production.

Purification of the Krypton – Xenon Concentrate. In addition to krypton and xenon the krypton – xenon concentrate still contains oxygen (or argon), hydrocarbons, and possibly traces of other impurities, e.g., carbon dioxide.

The concentrate is, for example:

1) Brought to a pressure of 55 bar (supercritical) with a liquid pump and heated in a hot water heater to ambient temperature so that it can then be expanded again to 6 – 10 bar. Compression to a supercritical state is necessary on safety grounds, so that no dangerous enrichment of hydrocarbons can take place during the heating such as would be possible in an evaporating liquid;
2) Freed from hydrocarbons in a platinum contact furnace at 600 K;
3) Freed from carbon dioxide and water in an adsorber filled with molecular sieves;
4) Freed from oxygen and argon in a column at ca. 4 bar. The oxygen and argon stream is returned to the krypton – xenon concentration column. The increased pressure reduces the solidification of krypton and xenon.

The krypton – xenon mixture containing 10 – 50 % oxygen which is obtained at the column bottom is placed in intermediate storage and processed further, after another hydrocarbon combustion step, in a batch distillation process (Fig. 19). The crude krypton – xenon – oxygen mixture is condensed in a column, which is then heated at the sump and cooled at the top. At the top a small gas flow is drawn off which consists initially of almost pure oxygen and is returned to the first concentration stage. After this comes an oxygen – krypton mixture which is placed in intermediate storage, and finally pure krypton is obtained with less than 100 vol ppm impurities. When xenon also appears in the krypton the batch is finished and a krypton – xenon mixture predominantly consisting of xenon collects in the sump. This is freed from krypton in a further batch distillation process after removal of trace impurities in combustion furnaces and adsorbers. The residual gas is also collected here and is returned later to a suitable point in the process. Pure xenon with less than 100 vol ppm

impurity collects in the sump of this column. Because of the small quantities, the complicated procedure, and the laborious analysis, the production process for pure krypton and pure xenon is normally housed in a laboratory.

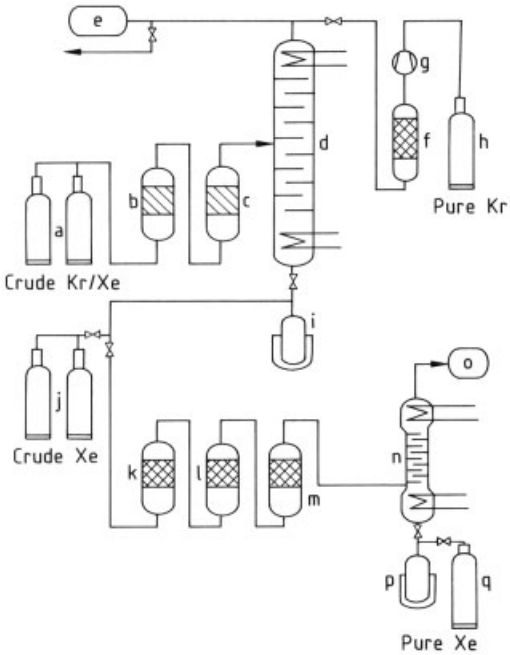

Figure 19. Kr – Xe separation
a) Crude Kr – Xe collecting bottles; b) C_mH_n combustion system; c) CO_2/H_2O removal; d) Kr – Xe separation column; e) Crude Kr – O_2 – Ar gas collector; f) O_2 oxidation; g) Pure Kr compressor; h) Pure Kr collecting bottle; i) Crude Xe pressure evaporator; j) Crude Xe collecting bottle; k) C_mH_n combustion; l) CO_2 – H_2O removal; m) N_2O oxidation; n) Pure Xe column; o) Crude Xe gas collector; p) Pure Xe pressure evaporator; q) Pure Xe collection bottle

4.2. Helium Extraction from Natural Gases

Of the noble gases contained in natural gas, only helium is suitable for extraction [181]. The other noble gases are present in such small concentrations that they are not worth separating, but with a content of 0.005 – 8 % the helium offers a good source for helium production [182, 183]. Extraction of helium by itself is only viable at contents above 1 % [184], and it is only worth extracting helium at lower concentrations if other products are also obtained from the natural gas.

In industrial plants a crude helium stream is first separated from the natural gas. This is then followed by one or more purification stages if highly pure helium is required. The crude helium extraction and the purification systems are often interconnected by recycle streams.

4.2.1. Crude Helium Extraction

Crude helium is usually produced by partial condensation at low temperatures [185]. Increasing consideration has also recently been given to the membrane process, or to a combination of cryogenic and membrane processes.

4.2.1.1. Low-Temperature Fractionation

Partial condensation exploits the extremely low liquefaction temperature of helium. Natural gas is cooled in stages until practically all the components with the exception of helium have condensed. The residual gas phase is the crude helium product, with a helium concentration which is determined by the pressure and the lowest condensation temperature. Maximum helium concentrations of ca. 99 % can be reached at temperatures close to the melting point of nitrogen [181].

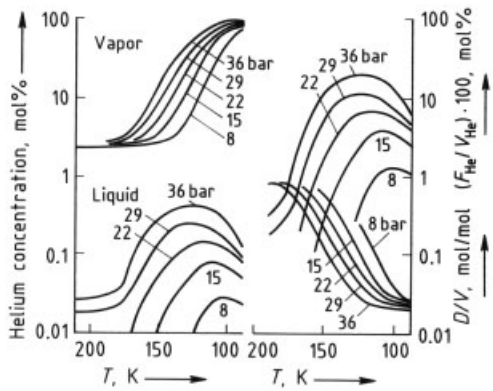

Figure 20. Behavior of helium during condensation of natural gas
D = vapor quantity, V = total quantity, F_{He} = dissolved helium, V_{He} = total helium

Figure 20 shows the temperature and pressure dependence of:

1) The equilibrium compositions of the liquid and vapor phases,

2) The ratio of vapor quantity to total quantity, and
3) The quantity of helium dissolved in the liquid as a percentage of the total quantity of helium

for natural gas from Keyes, Oklahoma (2.1 % He, 26.3 % N_2, 71.6 % CH_4 and heavier components [186]). Although the concentration of helium in the liquid phase of the natural gas is very low, the dissolved helium can reach a substantial proportion of the total quantity of helium contained in the natural gas because of the large quantity of condensate formed. In practice the dissolved helium is therefore recovered by stripping, rectification, or expansion of the condensate in stages, which substantially improves the helium yield [187].

Industrially, helium extraction is usually combined with other processing steps for natural gas, particularly with partial condensation. Helium separation then becomes a subprocess of a main process which determines the process structure and this naturally makes a significant improvement in the cost-effectiveness of the extraction process. Very often this is carried out during the separation of nitrogen from nitrogen-rich natural gases to improve the calorific value, because a high helium content is usually associated with a high nitrogen content [182].

In natural gas liquefaction, the helium collects during condensation in a gas residue, which can be readily concentrated and purified. The first stages of the gas treatment and partial condensation are also already performed during the extraction of ethane and heavier hydrocarbons from natural gas, thus relieving any subsequent helium separation of this task.

An important factor in all processes is the initial concentration of helium in the natural gas. With low helium concentrations great care has to be exercised in the recovery of the helium dissolved in the condensates in order to achieve good yields.

An important part is also played by the natural gas impurities, especially CO_2, H_2S, and heavy hydrocarbons. These components can solidify at low temperatures and lead to deposits. If their concentrations exceed certain limits they must therefore be removed before they enter a low-temperature section. Great efforts are made to use skilful process technology to pass the highest possible percentage of impurities safely through the system and to avoid preceding purification steps. It is, however, impossible to avoid removing the water vapor.

The required output pressure of the helium-free natural gas is the decisive factor determining the energy requirement of the system. If the treated natural gas can be delivered under reduced pressure, then the refrigerating requirements can sometimes be covered by the Joule–Thomson effect alone. In other cases additional refrigerating cycles or recompression of the natural gas is necessary.

A simple, but typical modern process for helium recovery combined with nitrogen removal from natural gas is shown in Figure 21. The clean natural gas is cooled against cold product streams in heat exchanger (a) and is fed to the bottom of the high-pressure column (b). The condenser (c) provides reflux for this column and for the low-pressure column (d). That part which is not condensable contains most of the helium. It is drawn from the crude helium separator (e) and leaves the plant as crude helium product after warming.

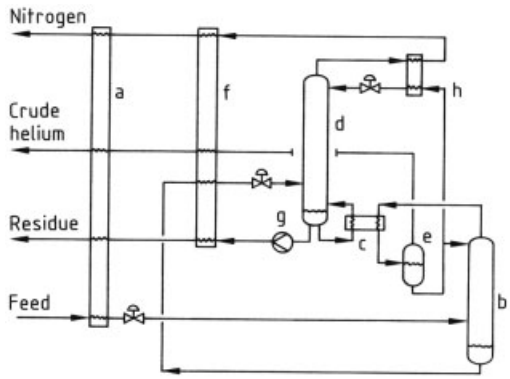

Figure 21. Nitrogen rejection and crude helium recovery from natural gas
a) Feed/product exchanger; b) High-pressure column; c) Reboiler/condenser; d) Low-pressure column; e) Crude helium separator; f) Feed subcooler; g) Residue pump; h) Reflux subcooler

The bottoms from the high-pressure column, having some of the nitrogen removed, form the feed stream for the low-pressure column (d), where the final nitrogen–methane separation is performed. The overhead is the nitrogen stream, which is rewarmed against the incoming feed stream. A reboiler in the bottom of the column

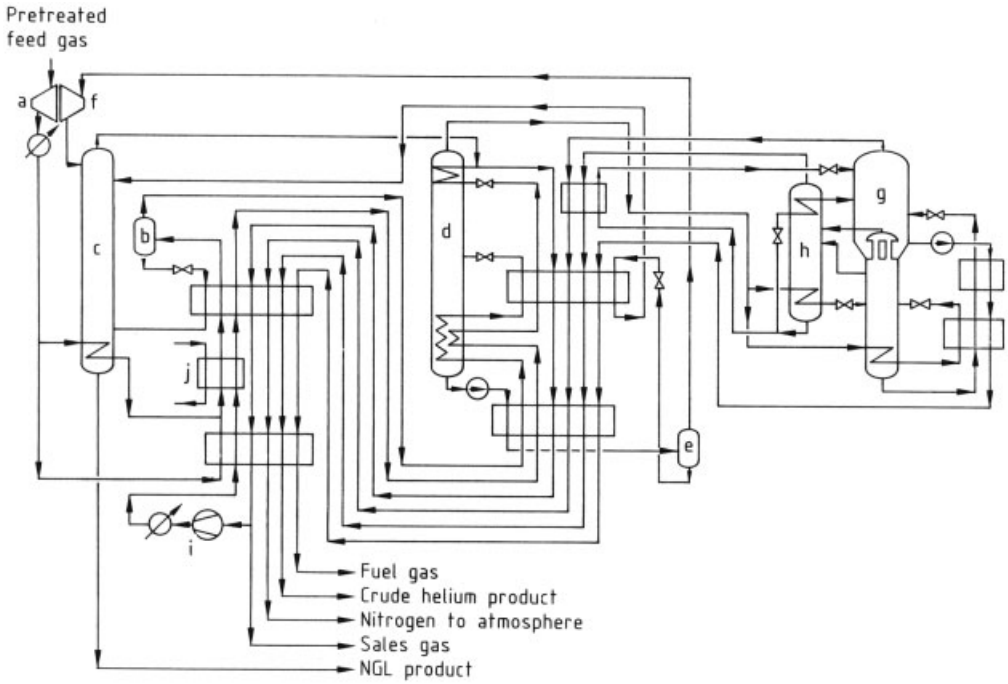

Figure 22. Schematic of a plant for recovery of helium from natural gas, with integrated separation of nitrogen, methane, and a C_{2+} fraction
a) Expander booster compressor; b) Cold separator; c) Demethanizer; d) Nitrogen enrichment column; e) Expander suction scrubber; f) Turbine; g) Low pressure column; h) Helium column; i) Recycle compressor; j) Propane refrigeration

provides a liquid, methane-rich residual product, which is pumped to an elevated pressure, vaporized and rewarmed against the incoming feed.

The reboiler duty in the low-pressure column necessary to provide a residue with low nitrogen content is provided by condensing nitrogen at the top of the high-pressure column. This is possible because of the different pressures of the two columns. The condensation temperature of nitrogen in the overhead of the high-pressure column (b) is higher than the boiling temperature of the methane-rich liquid in the low pressure column (d).

A significantly more complicated process is used in a plant in the United States [188]. The following products, besides a crude helium fraction, are recovered from the natural gas: a C_{2+} fraction, a methane sales gas, a fuel gas fraction, and a nitrogen fraction (Fig. 22). The composition of the feed gas and the products is given in Table 18.

Table 18. Composition of the feed gas and products for the plant in Figure

	Feed gas	NGL	Sales gas	Fuel gas	He	N_2
Composition, mol %						
Helium	0.50		0.01		56.94	0.17
Nitrogen	10.20		2.99	23.48	42.99	99.55
Methane	73.02	0.70	96.52	76.51	0.07	0.28
Ethane	6.94	41.21	0.47	0.01		
Propane	5.16	32.07	0.01			
C_{4+}	4.18	26.02				
Flow, kmol/h	2988	480.1	2019.4	299.1	25.4	164
Pressure, bar	44	9	8	8	27	1
Temperature, K	311	262	306	306	306	306
Gross wet calorific value, MJ/kmol	989.7	2154.5	852.7	670.2		

Pretreated natural gas is compressed to 50 bar in an expander booster compressor (a) and is cooled to 230 K by heat exchange with low-temperature propane and with cold products. The condensate is separated in the separator (b) and conveyed to the demethanizer (c).

After partial condensation, the gas from the separator (b) enters the nitrogen enrichment column (d), which operates at 31 bar. The bottom product from this column is pumped to 40 bar and partially vaporized, and the resulting phases are separated in the expander suction scrubber (e). Both phases are fed to the demethanizer (c) – the liquid phase after partial vaporization, the gas phase after expansion in the turbine (f). The liquid produced in the turbine serves as a reflux for the column.

The bottom product from the demethanizer (the C_{2+} fraction) and the top product (the sales gas, mainly methane), leaves the system, the top product having been warmed up again against the feed gas and the recycle stream.

The top product from column (d), enriched with nitrogen, is fed to a double column (g). After it has been subcooled, the bottom product from the lower, high-pressure section of the column forms the feed for the upper, low-pressure section of the column. The gaseous top product of the high-pressure column, in which almost all the helium has collected, and part of the top condensate are fed into the helium column (h) where the crude helium is concentrated to 57 % and, after being warmed up, leaves the system.

After subcooling, the bottom product from the helium column (h) serves as a reflux for the low-pressure section of the double column (g), from the top of which the nitrogen fraction is drawn off and discharged to the atmosphere after warming up. The bottom product is pumped to about 9 bar and, after warming up, used as fuel gas.

The necessary heating and cooling of the nitrogen enrichment column (d) is provided by a recycle with part of the sales gas via the recycle compressor (i).

4.2.1.2. Crude Helium Extraction by Permeation Processes

Also suitable for separating helium from natural gas are processes based on the principle of permeation. Gas permeation through homogeneous membranes occurs by solution – diffusion. The permeability coefficient is equal to the product of Henry's law constant S_i, and the diffusivity coefficient D_i, of component i in the membrane material. A membrane's ability to separate a gas mixture can be obtained from the ratio of the permeation coefficients: $\alpha_{ij} = P_i/P_j$, where α is the selectivity or separation factor of gaseous components i and j. Homogeneous membranes have a high permeability for helium compared to the remaining components of natural gas, i.e., nitrogen, methane, carbon dioxide and others (see Table 10; further literature in [189] and [190]).

The first applications of membranes for the extraction of helium date from 1958 when helium was extracted from natural gas by means of tubular silicate and quartz glass membranes. Commercial use of membrane processes for extracting hydrogen began in the early 1970s and since 1985 they have also been used for helium separation from natural gas for the production of crude helium [191].

The heart of a membrane process is the separation module. Commercially available separation modules contain either hollow fiber bundles (Permea, Ube) or spiral-wound gas permeation elements (Grace, Separex). For the structure of these modules see → Hydrogen, Chap. 5.5.2.

Figure 23 shows a schematic of a helium concentration process. In this case the two stages of the membrane system consists of a row of separators each containing a hollow fiber bundle physically similar to a shell and tube heat exchanger. The membrane is an aromatic polyimide polymer membrane with a membrane area of 9000 m^2 per m^3 of volume. The membrane works satisfactorily in the temperature range − 80 to + 150 °C and tolerates natural gas accompanying compounds such as water (up to saturation), hydrogen sulfide (up to 5 vol %) as well as C_{5+} (up to saturation). Table 19 gives some operating figures of the first stage of a membrane helium recovery system for helium-rich natural gas [192].

By their nature membrane processes are more suitable for coarse separation of helium from natural gas; i.e., it is not possible to achieve a high level of helium purity combined with a high helium yield in a simple membrane process. If this is required it is necessary to change to a system of several membrane process steps

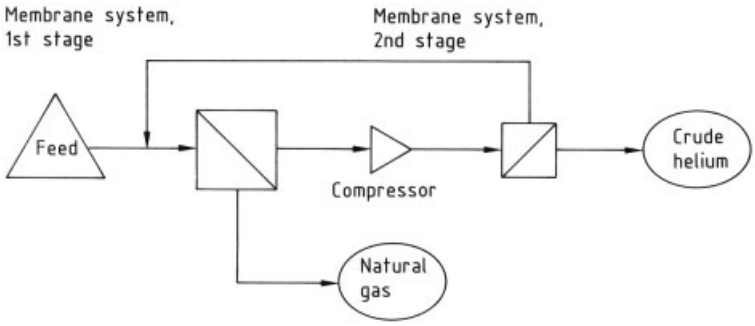

Figure 23. Two-stage membrane helium recovery system [192]

connected in series. One or more compressor stations would then be needed between the process steps and for returning residual gases which are produced during the membrane process from retentates and permeates to be recycled. This would overextend the entire process and make it uneconomical. In such cases there is much to recommend a combination of cryo-process and membrane process, or of membrane process and pressure swing adsorption (PSA) process.

Table 19. Operating figures of the first stage of a membrane helium-recovery unit

	Feed	1st Stage permeate	Crude He
Pressure, bar	35	1.05	1.05
He recovery, %			95
Composition, mol %			
He	2.40	11.2	59.53
CO_2	0.07	0.3	0.76
Ar			
N_2	24.54	22.3	10.02
CH_4	72.71	66.1	29.68
C_{2+}	0.28	0.1	0.01

Several hybrid membrane – cryogenic or membrane – cryogenic – adsorption processes have been suggested [193]. The most advantageous seems to be the hybrid process with the following characteristics (Fig. 24). Helium-containing natural gas (ca. 2 % He, 26 % N_2, 64 % CH_4) at ca. 32 bar is cooled to ca. − 50 °C to condense all the heavier hydrocarbons. Then the gas is further cooled to ca. − 150 °C to condense all the methane and some of the nitrogen. The gas from these separation stages contains 30 – 35 vol % helium. This helium-rich stream is fed to the membrane system, in which a number of membrane stages can be configured to maximize the separation efficiency and to produce the desired 95 % stream. Any unrecovered helium is recycled to the cryogenic unit. To obtain grade A helium (purity 4.5), a PSA unit may be integrated. The process can be optimized by varying the helium content of the helium-enriched stream leaving the cryogenic stage and hence the membrane area to He content ratio. It is claimed that the power required by the hybrid process is only 54 % of that of a purely cryogenic process.

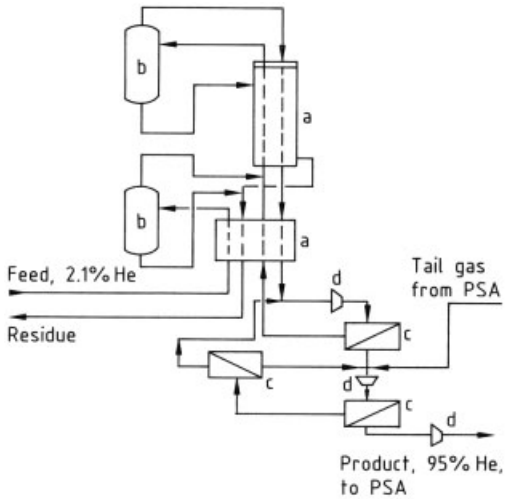

Figure 24. Cryogenic – membrane hybrid process
a) Heat exchangers; b) Separators; c) Membrane stages; d) Compressors

4.2.2. Purification of Helium

Besides adsorptive processes also cryogenic and chemical processes are used for the purification of crude helium. Different process steps are frequently linked with one another in case the required purity cannot be achieved in a single step.

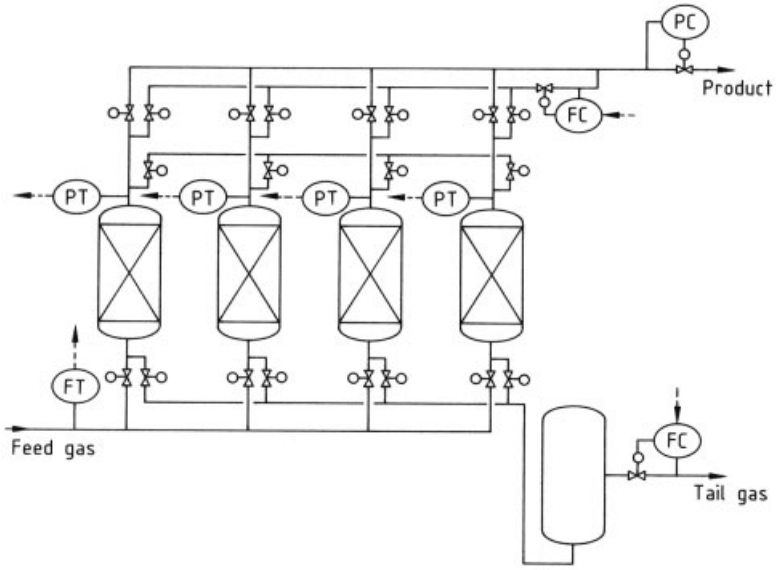

Figure 25. Schematic of a typical four-adsorber PSA unit

4.2.2.1. Purification of Helium by Pressure Swing Adsorption

Helium is retained only weakly by adsorbents so it is easy to purify by adsorption processes. In principle this involves processes with thermal regeneration or pressure swing adsorption (PSA) processes. In processes with thermal regeneration heat is supplied to the process externally for desorption, but in PSA processes the regeneration is carried out without the supply of external heat, merely by pressure release and purging at low temperature. Generally, processes with thermal regeneration are confined to the removal of trace impurities or readily adsorbed components. At ambient temperature these processes are used for removing water and carbon dioxide, and at low temperatures also for removing, for example, traces of nitrogen, oxygen, neon, and hydrogen. With PSA systems it is possible to remove less readily adsorbed components at ambient temperature and even at quite high concentrations.

During the extraction of helium from natural gas the hydrocarbons and the majority of the nitrogen are separated in the low-temperature system. Nitrogen is virtually the only helium impurity requiring secondary purification in the PSA system. Zeolitic molecular sieve and narrow-pored activated carbon are suitable adsorbents for nitrogen. These adsorbents give final purities of 99.999 – 99.9999 vol % He without additional purification steps.

Pressure swing adsorption systems for helium purification are comparable in principle with systems for hydrogen purification; in this process helium behaves in a similar way to hydrogen (→ Hydrogen, Chap. 5.2.3.). Adsorption systems with four or less adsorber vessels are normally used because the feed gas flows to be processed are smaller than in the hydrogen systems (see Fig. 25). In such a system with 4 adsorbers the individual adsorbers pass cyclically through the following process steps: adsorption, pressure equalization with another adsorber, further pressure release in parallel flow to the adsorption step to output purge gas for another adsorber, counter-current pressure release to output tail gas, purging with decompression gas from another adsorber, pressure build-up by pressure equalization with yet another adsorber, and further pressure build-up to the adsorption pressure with helium product. A helium yield of about 80 % can be achieved with an adsorption pressure of 15 – 20 bar.

The tail gas produced in PSA processes still contains considerable quantities of helium so, unlike the procedure in the majority of hydrogen PSA systems, it is compressed again because of the substantially higher value of helium and re-

turned to the cryogenic unit, resulting in a helium yield of practically 100 % for the entire system.

In principle a PSA system can remove nitrogen contents of up to 50 vol % and produce ultrapure helium. The N_2 loading at the adsorption pressures normally used is virtually linear over a fairly large partial pressure range so the adsorber size only increases slightly with increasing N_2 concentration in the crude helium, while the helium yield falls with increasing adsorber size.

Due to the recycling of tail gas, the quantity of crude helium fed to the PSA system increases in inverse proportion to the helium yield of the PSA. The capital costs and the energy consumption for recompression of the tail gas to the cryogenic unit shows the same behavior. On economic grounds it has therefore been found that an N_2 content of 5 – 15 vol % in the crude helium to the PSA is the most cost-effective solution for combination with low-temperature separation.

4.2.2.2. Other Processes for Purification of Helium

Except for hydrogen, virtually all components can be separated from helium by pressure swing adsorption. Normally there is no hydrogen in helium, but if hydrogen is present, e.g., through the addition of industrial gases, then other processes have to be used.

Hydrogen is removed by catalytic combustion. The crude helium is mixed with air and the mixture is passed over a catalyst. If the resulting heat of reaction is too high then a recycle system can be used to reduce the hydrogen content in the feed gas to the reactor. The water formed is condensed out by cooling (provided that this is possible from the point of view of the partial pressures) and the remaining water can be separated together with the other impurities in a pressure swing adsorption system.

Purification processes based on hydride-forming metal alloys are also suitable for producing high-purity helium and for purification of the other noble gases. Noninert gas constituents (H_2O, CO_2, O_2, CO, C_mH_n) are combined on the surface of highly active getter alloys, or, as in the case of hydrogen, they diffuse into the metal lattice. When the partial pressure reaches the so-called hydride formation plateau the hydrogen is combined as metal hydride (→ Hydrogen, Chap. 9.3.1.2.).

Highly effective getter alloys consist of Ti, V, Fe, and Mn in varying molar ratios depending on the application; the H_2 partial pressure can be lowered to 5×10^{-6} bar by $TiV_{1.5}Fe_{0.4}Mn_{0.1}$. The gas flow is passed directly through the getter beds (pressure drop < 1 bar), which can be regenerated by raising the temperature (200 – 500 °C) or returned to the manufacturer after use (e.g., HWT, Germany). A helium purity of 6.0 (99.9999 vol %) is achieved; a purity of 7.0 can be attained if the gas is passed over a high-temperature getter bed at 350 – 450 °C, which allows practically no foreign gas constituents to pass. This results in an oxygen content, for example, which can no longer be detected by current analytical methods [194].

4.2.3. Example of a Complete Process

The two plants at Alfortville in France where nitrogen is separated from Groningen natural gas and helium is extracted are examples of plants which perform crude helium extraction as well as helium purification [195].

The helium content in the Groningen natural gas is relatively low compared with many natural gases in the United States or in Poland for which systems have also been built for helium extraction [196–198]. For this reason, particular emphasis was placed here on minimizing the helium losses in the other product streams. Another particular feature is that the CO_2 is not removed in spite of its relatively high content of 0.9 % in the crude gas. Figure 26 (see next page) shows the process layout.

After removing aromatic hydrocarbons and water in the adsorber (a), CO_2-free top product and a CO_2-rich bottom product is obtained by rectification of a partial stream of the natural gas. The bottom product and the untreated natural gas stream condensed in the reboiler (d) are fed into the lowest section of the medium-pressure column (e). After condensation in the reboiler (d) and in the heat exchanger (l), the CO_2-free top product from column (c) is fed into a higher section of column (e) to extract the CO_2 from the vapor rising from the lower section. In this way all the CO_2 accumulates in the bottom product of column (e) which, due to the presence of C_{2+}

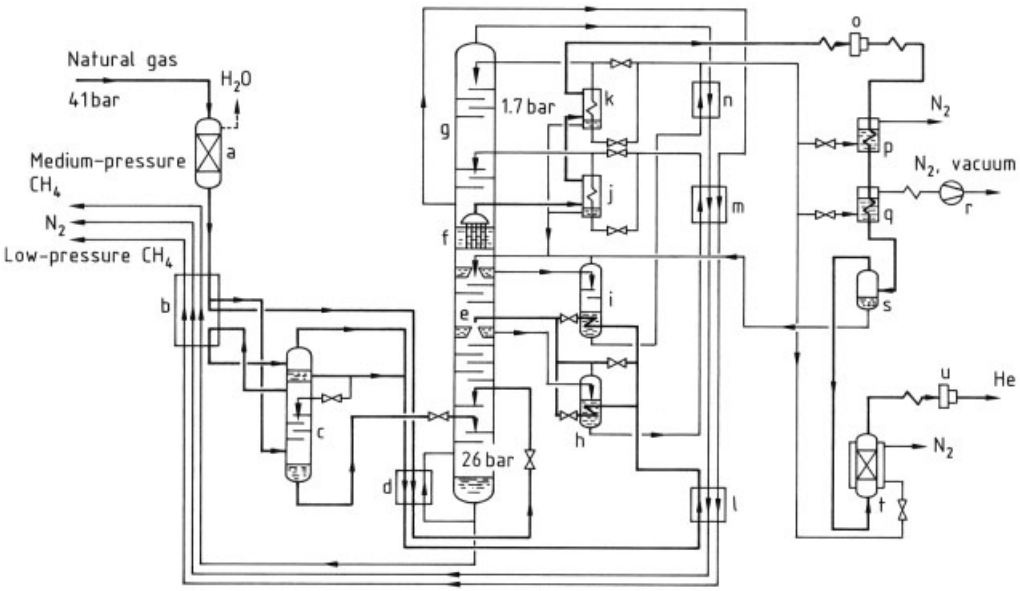

Figure 26. Plant for the extraction of pure helium from Groningen natural gas (Alfortville, France)
a) Adsorber; b) Main heat exchanger; c) Carbon dioxide column; d) Reboiler; e) Medium-pressure column; f) Condenser; g) Low-pressure column; h), i) Helium strippers; j), k) Helium concentrators; l), m), n) Low-temperature coolers; o) Helium compressor; p) Nitrogen evaporator; q) Vacuum nitrogen evaporator; r) Vacuum pump; s) Nitrogen separator; t) Adsorber; u) Helium compressor

hydrocarbons concentrated here, has sufficient dissolving capacity for CO_2 to carry it out of the system without causing any problems.

All the helium introduced by the natural gas flows with the nitrogen to the top of column (e) and is enriched in condenser (f) during the nitrogen liquefaction. Part of the helium is dissolved in the liquid trickling down and is lost with the liquid fractions which are fed from column (e) to the final nitrogen/methane separation stage in the low-pressure column (g). In order to prevent this and to reduce helium losses to a minimum the liquids drawn off are stripped in the reboiler (h) or the column (i). The helium driven off during this procedure is returned to the column (e). In this way, practically all the helium can be extracted at the top of condenser (f) at a concentration of ca. 10 %. In the two following heat exchangers (j) and (k), the helium concentrators, the helium is cooled by evaporating liquids (crude methane and nitrogen) and is enriched to 86 to 90 %.

For further concentration, the crude helium is first compressed by the diaphragm compressor (o) to 200 bar and then cooled until close to the melting point of nitrogen. The refrigeration is provided by liquid nitrogen, which vaporizes in heat exchanger (p) under atmospheric pressure and in heat exchanger (q) under vacuum. The condensate produced is separated in vessel (s) and returned together with the condensates from the concentrators (j) and (k) to the medium-pressure column (e) for recovery of dissolved helium. The helium, which now only contains 0.5 % N_2, is transferred to the final purification stage through the nitrogen-cooled molecular sieve adsorber (t). It leaves this with less than 5 vol ppm impurities. In compressor (u) it is finally compressed to the storage pressure of 350 bar.

During normal operation the system requires no external energy. The fractionation work and refrigeration requirements are covered simply by lowering the pressure of the natural gas. Nevertheless, 72 % of the combustible components can be produced as a medium-pressure fraction at 25 bar.

4.3. Isolation of Argon from Ammonia Purge Gases

In the production of ammonia synthesis gas from natural gas, water vapor, and air in the steam reforming process, the noble gas constituents of the air and those of the natural gas pass into the synthesis gas. Enrichment of those gases together with residual methane, mainly from CO methanization, takes place in the synthesis cycle. In order not to exceed a given inert gas level a portion of the synthesis gas is continuously purged from the cycle and, under normal circumstances, is used as a fuel gas (→ Ammonia, Chap. 4.5.3.5.).

The composition of this purge gas varies according to process type and catalyst age. The average purge gas composition, determined by statistical evaluation of purge gas analyses from about 50 projects, is as follows:

H_2	63.02 mol %
N_2	21.01 mol %
Ar	4.89 mol %
CH_4	11.08 mol %

Trace amounts of NH_3 and H_2O may also be present.

The argon content varies from plant to plant and can range from 3 to 6 mol %. The concentration of the helium depends on the helium content of the natural gas converted in the steam reformer. However, even when processing Groningen natural gas (Netherlands) with only 0.04 – 0.05 % helium appreciable concentrations are obtained in the residual gas, i.e., ca. 0.11 mol %. It can reach values of up to 1 mol %. The other noble gases are of lesser importance.

Hydrogen-depleted purge gas, i.e., the off-gas stream of hydrogen recovery units, which is produced either by membrane separation systems or cryogenically, contains 6 – 12 mol % argon with admixtures of hydrogen, nitrogen, and methane.

Suitable cryoprocesses [199–201] are available for production of noble gases from synthesis purge gas. Initially, extensive separation of nitrogen, argon, and methane is carried out by partial condensation. Purge argon is then recovered from the condensate in a two-stage condensation process. If helium is to be recovered it can be concentrated by liquefaction of the hydrogen in the hydrogen-rich gas phase, followed by purification. The heavier noble gases, krypton and xenon, pass into the methane fraction.

A modern industrial plant for producing high-purity argon, synthesis gas for re-use in ammonia synthesis, as well as liquid nitrogen as by-product was built by Linde for Rashtriya Chem. & Fertilizer Ltd. in Trombay, India in 1993. The process is shown schematically in Figure 27 (see next page).

The purge gas at pressures of up to 70 bar is first introduced into the adsorber (a) where traces of water and ammonia are removed. It is then transferred to the heat exchangers (b_1) and (b_2) for cooling. The gas is then fed into separators (c_1) and (c_2) where separation of liquefied nitrogen, argon, and methane from gaseous hydrogen takes place, and dissolved hydrogen flashes off. The liquid bottom product is fed into the fractionating column (d_1) where methane (bottom) is separated from the nitrogen fraction (top).

A methane-free nitrogen – argon mixture (liquid) is withdrawn from the middle of this column and fed as reflux into the Ar purification column (d_2), where nitrogen is separated from argon. The bottom product is argon of product purity and is transferred to a vacuum-insulated storage tank.

Both columns (d_1) and (d_2) are operated within an N_2 cycle in which cold is produced by expanding high- or medium-pressure nitrogen.

The CH_4 bottom stream from (d_1), compressed by liquid pump (g), is evaporated against feed gas and normally led to battery limits as fuel gas. The higher boiling noble gases krypton and xenon are contained in this stream and, in principle, can be isolated. The process described (operating data are given in Table 20) is the result of continuous improvements in cryogenic argon purge gas recovery, namely the omission of one of the three cryogenic columns normally used and a reduction of the required N_2 cycle pressure from 150 bar to 45 bar at the most. There are three main factors influencing the economy of such a process.

1) N_2/Ar ratio: this affects the cost of argon recovery; the lower the argon concentration the higher the cost of separation.
2) Purge gas pressure: the purge gas pressure also plays an important role in process energy consumption. The higher the pressure,

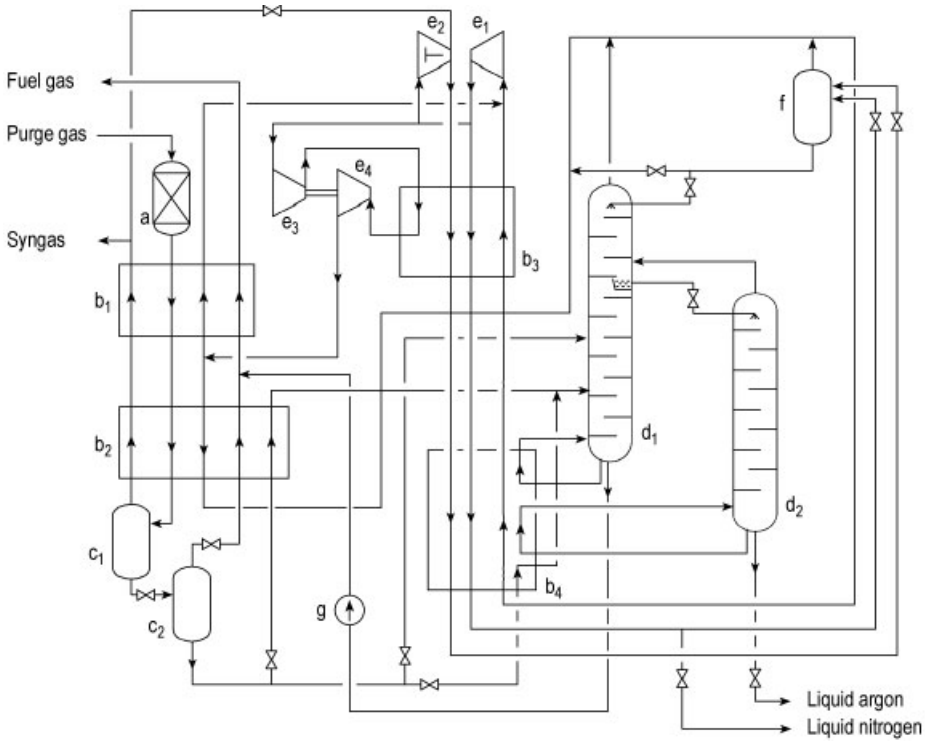

Figure 27. Flow sheet of a cryogenic argon recovery unit for ammonia purge gas
a) Adsorber for water and ammonia removal; b_1), b_2) Heat exchangers for the reheating of separation products and cooling of feed gas; b_3) Heat exchanger for the cooling and reheating of recycled nitrogen; b_4) Heat exchanger for the condensation and subcooling of recycled nitrogen, reboiling for columns and the reheating of recycled nitrogen; c_1) Separator for the separation of liquefied nitrogen, argon, and methane from gaseous hydrogen; c_2) Separator for the flashing of dissolved hydrogen; d_1) Column for splitting methane (bottom) from nitrogen/argon; d_2) Column for separating nitrogen (top) from pure argon (bottom); e_1), e_2) Centrifugal compressors for compressing low pressure recycle; e_3), e_4) Expanders with brake blowers for compression of the partial stream of compressed recycled nitrogen to 10 – 17 bar and expansion of this stream to 1 – 2 bar to produce refrigeration; f) Tank for storing liquid nitrogen; g) Pump for pressurizing the methane fraction to fuel gas pressure

Table 20. Technical data of a unit for the recovery of liquid argon from ammonia synthesis purge gas (products: syngas, liquid argon, liquid nitrogen, and fuel gas)

Material balance	Feed gas	Syngas	Fuel gas	Liquid N_2	Liquid Ar
H_2, mol%	58.5	75.8	28.1	0.6	–
N_2, mol%	20.7	23.5	2.6	99.2	< 2.0 mol ppm
Ar, mol%	5.5	0.6	0.4	0.2	99.9997
CH_4, mol%	15.3	0.1	68.9	–	< 1 mol ppm
kmol/h	405	278	104.6	2.2	20.2
T, K	283	311	311	83	91
p, bar	27.5	25.5	5.4	9.8	1.4

Utilities consumption: electric power 796 kW, cooling water ($\Delta T = 10$ K) 70 m^3/h.

the lower the cost of separation. This is due to a high Joule – Thomson effect, which in turn relieves the expansion turbine and permits a purge gas dew point that allows almost complete evaporation of the purge gas condensate prior to delivery to the methane separation column, thus reducing the reboiler heat load. The effect of pressure on the system has been deemed so important that the designs are adapted to ensure purge-gas delivery at up to 70 bar.

3) Precooling: an ammonia plant normally provides liquid ammonia at ambient pressure (240 K), which can be used to lower the front-

end entry temperature from cooling water level (305 K) to ca. 249 K. This provides a remarkable level of relief to the expansion turbine cycle.

The so-called hybrid argon recovery process (HARP, Airco, BOC) is a process for argon recovery which has been developed as an alternative to all-cryogenic processes. It is designed to handle hydrogen-depleted ammonia purge gases and make effective use of a combination of pressure swing adsorption (PSA) and cryogenic technology. The main objective of the PSA unit is bulk separation of hydrogen and nitrogen and removal of methane down to a fractional ppm level. A downstream cryogenic section separates and purifies the argon. A schematic representation of the HARP PSA unit is shown in Figure 28. This process has been undergoing extensive laboratory testing [202].

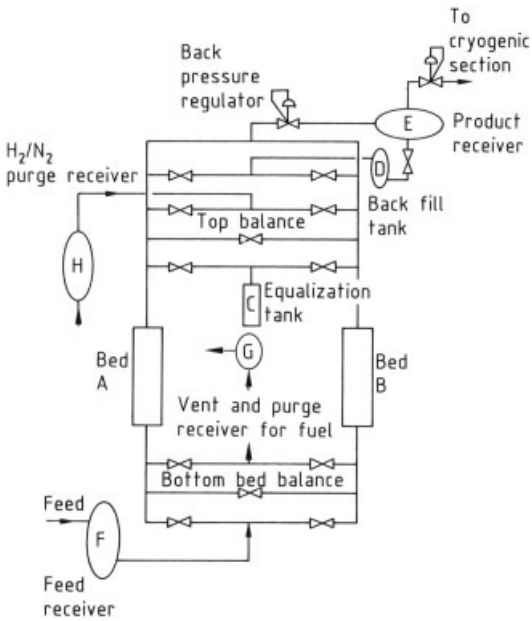

Figure 28. Schematic of the HARP PSA separation

Ammonia synthesis gas that has been produced from heavy oils by partial oxidation is normally purified in a liquid nitrogen wash to such an extent that extraction of a purge gas stream from the synthesis loop becomes unnecessary. The argon carried in with the oxygen in the gas production process is found in the wash liquid; it can be separated from the wash liquid by rectification in a similar manner to the way it is separated from the condensate of a synthesis purge gas. However, the separation is more complicated because of the presence of carbon monoxide with a vapor pressure close to that of argon [203].

Nowadays it is only in exceptional cases that helium is produced from the H_2–N_2 fraction of ammonia purge gas separators or from the hydrogen fraction of plants for hydrogen enrichment from ammonia purge gas. See [204] for these processes.

The quantities of the heavier noble gases, krypton and xenon, in the methane fraction of a purge gas separator are small. This normally contains krypton and xenon in a volumetric ratio of 1: 0.07 with a total content relative to methane (in the case of synthesis gas production by steam reforming) of 30 – 50 vol ppm, depending on the mode of operation of the plant and the age of the catalyst. In a relatively simple process the krypton and xenon can be enriched by rectification and drawn off as a 50 % concentrate. In a subsequent secondary purification process the methane is burnt with the addition of pure oxygen, the combustion product is removed, and the raw krypton – xenon obtained is processed to give pure krypton and xenon [205]. A plant of this type was supplied by Linde in 1988 to Chem-Co, Wratza, Bulgaria. It is intended to produce 34 m^3 xenon and 420 m^3 krypton per year.

5. Liquefaction of Helium

Unlike the other noble gases, helium remains in a gaseous state during the production process. However, helium is increasingly required as a liquid coolant for cooling at very low temperatures. It is also often transported and stored in liquid form.

5.1. Processes for Liquefaction of Helium (see also → Cryogenic Technology)

In the majority of cases helium is liquefied by isenthalpic throttling in a valve (Joule – Thomson valve, JT valve); the proportion of liquid is determined largely by the pressure and

temperature before the throttle valve. The temperature and pressure range in which the temperature drop of the gas needed to produce liquid at each throttling stage takes place in order to liquefy helium by the Linde process is limited by the inversion curve (Fig. 29). The maximum values are 43 K and 38.5 bar; if helium is cooled to a state within this curve then liquefaction is possible. Temperatures below 20 K and pressures of less than 20 bar are usually used for industrial operation.

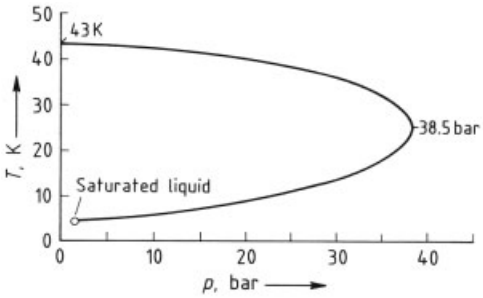

Figure 29. Joule – Thomson inversion curve

Processes for helium liquefaction can be divided into two categories:

1) Recuperative processes; i.e., the helium is cooled and heated in heat exchangers (recuperators)
2) Regenerative processes; i.e., the helium is cooled and heated in heat accumulators with high thermal capacities and transmission surfaces (regenerators)

Typical representatives of the two categories are described below. The thermodynamic properties for the process calculations can be found in [206].

5.1.1. Recuperative Processes

Compressed helium gas is cooled in heat exchangers to the required precooling temperature. A partial stream is liquefied by throttling at the cold end of the plant. In liquefaction plants the liquid is collected in a tank. The fraction which remains gaseous is heated in countercurrent to the high-pressure gas and transferred to the suction side of the compressor.

Linde Process. Liquefaction of helium by the Linde process is a multi-stage process (cascade).

Liquid hydrogen is suitable for precooling compressed helium to approximately 20 K. If the hydrogen cascade stage also operates on the Linde principle then it also requires precooling as the hydrogen inversion curve is far below the ambient temperature. This precooling is achieved with the aid of liquid nitrogen. The basic layout is shown in Figure 30. The main disadvantage of such a liquefying apparatus is the number of refrigerating cycles required with separately-supplied refrigerants. When hydrogen is used as a refrigerant its combustibility also has to be taken into account, which requires safety precautions.

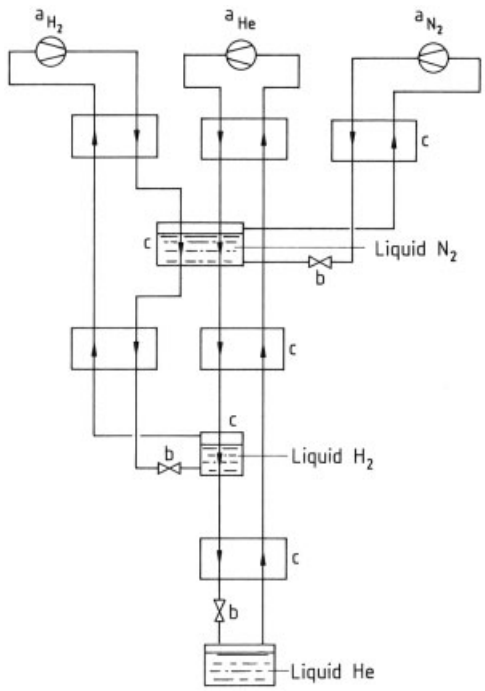

Figure 30. Cascade arrangement for helium liquefaction
a) Compressors; b) Joule – Thomson valves; c) Heat exchangers

Claude Process (→ Cryogenic Technology, Chap. 3.5.). The number of refrigerants is reduced and the use of hydrogen avoided in the Claude cycle in which helium acts as refrigerant.

In the system designed by KAPITZA and MEISSNER (Fig. 31) helium gas is cooled to a temperature below 20 K while performing work in an expander, so that the precooling of compressed helium required for the Linde process is achieved by countercurrent heat transfer. The first precooling stage at about 80 K with liquid nitrogen is retained as this is generally readily available and causes no problems in relation to plant safety.

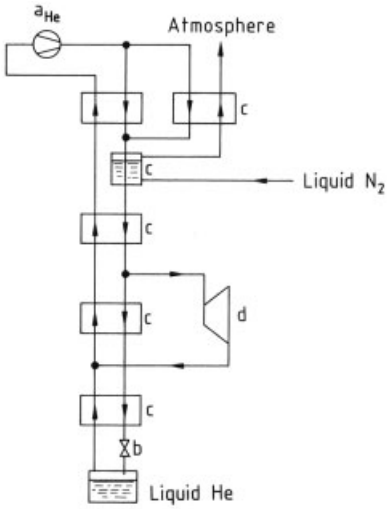

Figure 31. Claude cycle for helium liquefaction (Kapitza – Meissner)
a) Compressor; b) Joule – Thomson valve; c) Heat exchangers; d) Expander

Collins Cycle (→ Cryogenic Technology, Chap. 3.7.). An improvement in the energy aspect of helium liquefaction is achieved by incorporating another expander in the closed circuitin the temperature region between 80 and 20 K (Fig. 32). This process was first used by COLLINS and is therefore named after him.

Nowadays the majority of helium liquefaction and refrigeration plants in the lower and medium capacity range use this process. With only one stage the specific energy consumption is 4.8 kWh/L liquid He, a value which is about 22 times as large as the theoretical expenditure (lossless process operation) of about 0.22 kWh/L liquid He under the specified conditions. Introduction of a second precooling stage reduces the expenditure of energy by more than 50 % to 2.3 kWh/L. Additional precooling stages are chiefly of interest for plants with high refrigerating capacities and correspondingly high energy costs. The energy consumption with three stages is 1.8 kWh/L, and with four stages, 1.6 kWh/L.

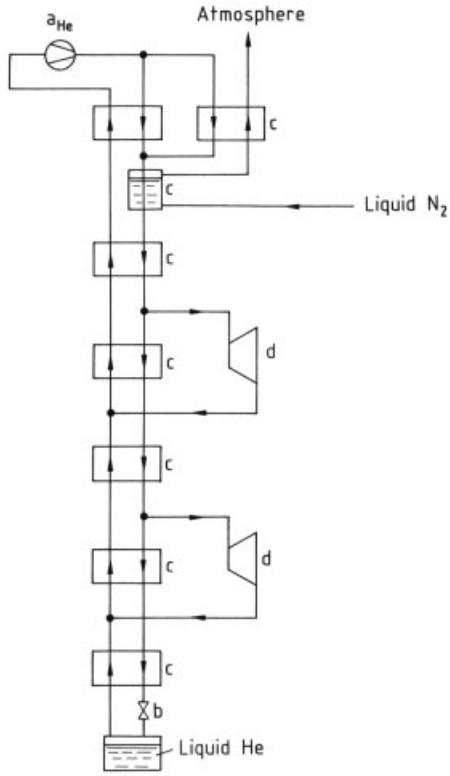

Figure 32. Collins cycle for helium liquefaction
a) Compressor; b) Joule – Thomson valve; c) Heat exchangers; d) Expanders

5.1.2. Regenerative Processes

The characteristic mode of operation of regenerative processes is explained with the aid of Figure 33. The displacer (a) divides the cylinder into a low-temperature operating chamber (b) and a high-temperature operating chamber (c). The displacer (a) can be moved freely within the cylinder. The two operating chambers are connected with each other via the regenerator (d) – a heat accumulator of high thermal capacity and transfer surface area. During downward movement of the displacer the gas flows at high temperature via the regenerator into the operating chamber (b) and is cooled; as a system with

constant volume is involved, the pressure in the system decreases. If the displacer is moved upwards then cold gas flows to (c) via (d); it is heated and the system pressure increases.

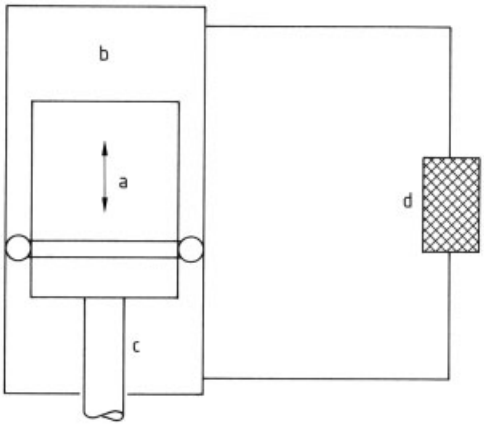

Figure 33. Principle of the regenerative process
a) Displacer; b) Low temperature operating chamber; c) High temperature operating chamber; d) Regenerator

If the displacer is driven by a crank shaft the two volumes and the pressure change periodically with the angle of rotation. However, this change would take place in opposing phases so the resulting refrigerating capacity would be zero. Some work must therefore be applied in the form of pressure variations which have a suitable phase displacement in relation to the changes of the plunger position. This can be carried out in the following possible ways:

1) In the Stirling process (Philips design) a compression piston is located in the hot chamber and has a phase displacement of $-90°$ in relation to the plunger (→ Cryogenic Technology, Chap. 4.4.).
2) In the Gifford – McMahon process pressure changes are generated by changing the gas content of the system for which purpose the system is connected alternately to the high and low pressure sides of a compressor (→ Cryogenic Technology, Chap. 4.7.).
3) In the Vuilleumier process (→ Cryogenic Technology, Chap. 4.5.) two displacer systems are linked. The hot piston (thermal compressor) is housed in a heated cylinder, and the cold piston forms the actual refrigerator. In contrast to the Stirling process the hot (compression) piston here has a phase displacement of $+90°$ relative to the cold (expansion) piston.

These processes have been shown to be industrially feasible for helium liquefaction. However, in most cases they are used for precooling compressed helium gas to the temperature required for liquefaction.

5.2. Gas Purification for Helium Liquefiers

While it is being cooled to liquefaction temperature the helium must be free from all impurities which cause operational problems by freezing out. With the exception of oil most of these impurities are picked up while the liquid helium is being used and must be removed in various purification stages.

The oil which passes into the helium during the compression process is removed in several stages. A typical arrangement is shown in Figure 34. Coarse separation of the oil from the helium takes place in the gravity separator (a) (cyclone separator, demister). Fairly large droplets are retained in the subsequent separator (b) so that gas containing approximately 50 ppm enters the very fine filter combination (c). Separation takes place here typically to less than 0.5 ppm. A residual oil content of less than 10^{-2} ppm is finally achieved in the following active carbon adsorber (d).

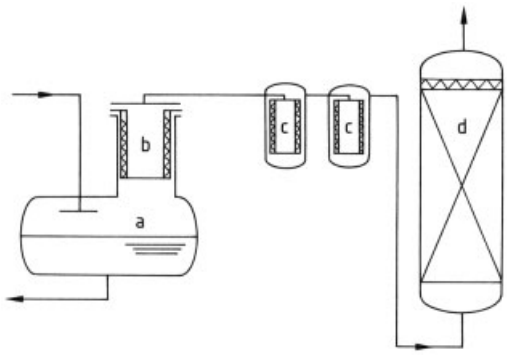

Figure 34. Typical design of an oil separation system
a) Gravity separator; b) Separator; c) Very fine filter; d) Active carbon adsorber

Water and carbon dioxide are normally removed at room temperature by alternating drying stations. While one string is in operation

the second string is regenerated by, for example, heating and flushing. This allows the liquefaction to take place on a continuous basis. Silica gels and molecular sieves are suitable for the adsorption of water and carbon dioxide.

At high concentrations (> 1 %) nitrogen and oxygen are first removed by condensation at approximately 65 K. In individual cases this is followed by adsorption at room temperature. The fine purification is then carried out either adsorptively with silica gel at approximately liquid nitrogen temperature (ca. 80 K) or by freezing out in specially designed countercurrent equipment. Adsorbers are normally designed with two strings to achieve continuous operation of the plant, while freezing-out equipment is usually designed with only a single string. This means that during the operational period of the heat exchanger more gas has to be purified than is actually liquefied in order to be able to maintain plant operation during the regeneration period. Both processes provide gas purities of less than 1 ppm.

Neon and hydrogen are normally only present in traces and are removed adsorptively at ca. 30 K. Activated carbon is a suitable absorbent for this purpose. In most cases, a single adsorber has a sufficiently long service life.

5.3. Examples of Helium Cryo Plants

With helium cryo plants it is necessary to differentiate in principle between two modes of operation–the liquefaction operation in which liquid helium is taken from the cryo plant, and the refrigerator operation in which a load is built into the circuit and is cooled by evaporating helium, the saturated vapor being returned to the cryo plant. Many plants are designed for the two modes of operation, either singly or in combination.

5.3.1. Liquefier Using the Claude Process

Cryo plants using the Claude process have the widest range of applications. The capacity range starts at a few watts of refrigerating capacity or liters of liquefying capacity and extends far into the kilowatt range. In practice, only plants using the Claude process are used above ca. 150 W refrigerating capacity.

An example of a small cryo plant is shown in Figure 35 [207]. It was developed specifically to provide liquid helium for nuclear spin tomographs. This requires a superconducting magnet operating at ca. 4.4 K which is surrounded with a helium-cooled radiation shield to reduce the incidence of external heat.

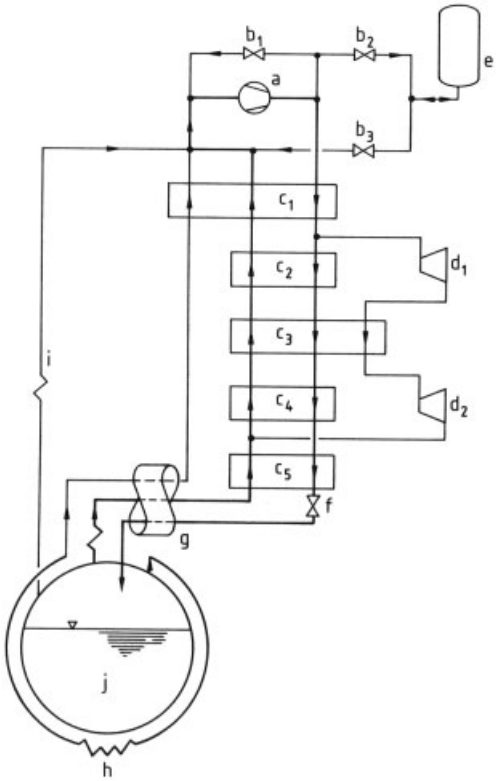

Figure 35. Schematic of a small helium cryo plant
a) Compressor; b) Control valves; c) Heat exchangers; d) Expansion turbines; e) Buffer vessel; f) JT valve; g) Transfer line; h) Radiation shield; i) Current leads j) Magnet with liquid helium bath

The plant operates in a mixed cooling and liquefaction mode. In steady state operation there is an input of 2.9 W to the 4.4 K helium bath; to this is added 1.8 W heat loss of the transfer line from the cryo plant to the magnet. The incidence of heat on the radiation shield is about 60 W. The current leads require 0.012 g/s liquid helium for cooling. The above figures give a 4.4 K equivalent cryo capacity of 14 W. For a drive capacity

of approximately 20 kW this means a specific energy consumption of about 1400 W per watt.

A single-stage oil-injection screw compressor (a) compresses approximately 7 g/s helium from 1.05 bar to 7.4 bar. The high-pressure gas is cooled in countercurrent flow to low-pressure gas in the aluminum plate heat exchangers (c). After the first heat exchanger, ca. 75 % of the high-pressure gas is transferred to the two expansion turbines (d), connected in series. After the last heat exchanger the remaining high-pressure flow is expanded in valve (f) to ca. 1.2 bar, whereby ca. 15 % of the gas stream is liquefied and supplied to the magnet via the transfer line. Two partial streams are taken from the liquid:

1) 0.12 g/s for cooling the radiation shield (h). This stream is returned to the first heat exchanger of the cryo plant at ca. 100 K.
2) 0.012 g/s for cooling the currents lead (i). This part is fed to the suction side of the cycle compressor at ambient temperature.

The remaining liquid is vaporized by the heat input from the magnet. The resulting vapor, together with the helium that remained gaseous in the expansion stage, is warmed in the heat exchanger and fed to the suction side of the cycle compressor.

The plant operates in a closed circuit; it therefore does not require any continuously operating gas purifying equipment with the exception of the oil-retaining system. All components are designed for a maintenance-free operating period of one year.

Figure 36 shows a schematic of a large-scale helium cryo plant used to cool a 6.4 km long ring of superconducting magnets, part of the HERA project for high energy research at DESY, Hamburg [208]. This plant is also used in a mixed cooling and liquefaction operation; 6500 W at 4.3 K are needed to compensate for the heat input into the magnets (g). The heat input to the vaporizing helium in the transfer lines is 275 W. To this is added 20 kW between 40 K and 80 K for cooling the magnet and transfer line radiation shields (e) and 20.5 g/s liquid helium for cooling the current leads (f) of the magnets. This gives a 4.3 K equivalent cryo capacity of approximately 10 kW or a 3000 L/h liquefier. The electrical power consumption of the compressors is ca. 2.9 MW, resulting in a specific energy consumption of approximately 280 W per Watt.

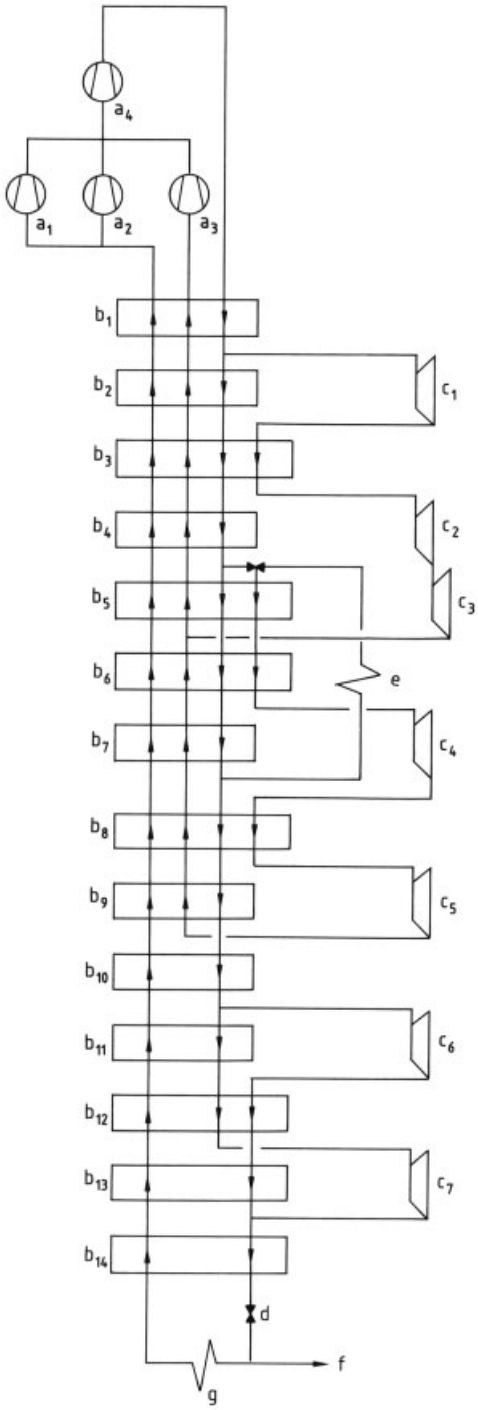

Figure 36. Flow diagram of a large-scale helium cryo plant a) Compressor; b) Heat exchanger; c) Expansion turbines; d) JT valve; e) Radiation shield; f) Current leads g) Magnet

This specific power consumption, which is very low for helium cryo plants, is achieved by several precooling circuits with a total of 7 turbines. The in-series-connected turbines (c_1)–(c_3) cover the temperature range 300 K to ca. 60 K. The turbines (c_4) and (c_5) operate in parallel to (c_1)–(c_3), but are also connected in series with one another, as a second precooling stage with ca. 20 K as the lowest temperature.

The two precooling stages are supplied with high-pressure helium at ca. 18 bar and 16 bar respectively; the outlet pressure from the turbines (c_3) and (c_4) is ca. 2.5 bar. The turbine return gas therefore requires its own passages in the heat exchangers. On the other hand the cryo capacity of the plant can be adapted very simply to the actual requirements by varying the inlet and outlet pressures of the turbines while keeping the pressure ratio constant.

At least 50 % of the total mass flow of approximately 870 g/s is supplied to the cold section of the plant. Unlike conventional designs, two parallel turbines (c_6) and (c_7) are used here before the throttling stage. This configuration permits the use of very small temperature differences between high- and low-pressure streams in the five cold heat exchangers, minimizing the exchange losses between the two streams.

The gas liquefaction takes place at the throttle valve (d) after the last heat exchanger. Because the turbines are positioned upstream the throttling does not take place from compressor final pressure but from ca. 4 bar.

The quantity required for cooling the current leads is drawn off from the resulting liquid and, after heating to ambient temperature, is transferred to the suction side of the compressor. The remaining liquid is vaporized by the heat load and is heated, together with the helium which remained in the gaseous state during the throttling stage, at approximately atmospheric pressure in parallel to the turbine flow.

The helium circuit is compressed in two stages in four machines. The 1 bar stream is compressed with two parallel compressors (a_1), (a_2) and the turbine return gas stream is compressed with one compressor (a_3), all to the same intermediate pressure. From this state the total mass flow is compressed in a second stage (a_4) to the process end pressure.

An external helium purifier can be added if required. It retains air components and operates at ca. 80 K. Continuous operation is achieved with two interchangeable adsorbers filled with silica gel. Liquid nitrogen is used as the refrigerant.

5.3.2. Liquefier Using the Stirling Process

Gas cryo machines using the Stirling process have often been used for the precooling stages in helium liquefaction plants. In more recent developments the Stirling principle is being used increasingly for very small helium cryo plants down to liquid nitrogen temperatures. Figure 37 shows a laboratory scale helium liquefaction plant [209]. The liquefaction capacity is 0.34 g/s (10 L/h) and the specific energy consumption is 2.8 kWh/L helium. This value, which is very good for the size of plant under consideration, is the result of four-stage precooling in the two two-stage gas cryo machines.

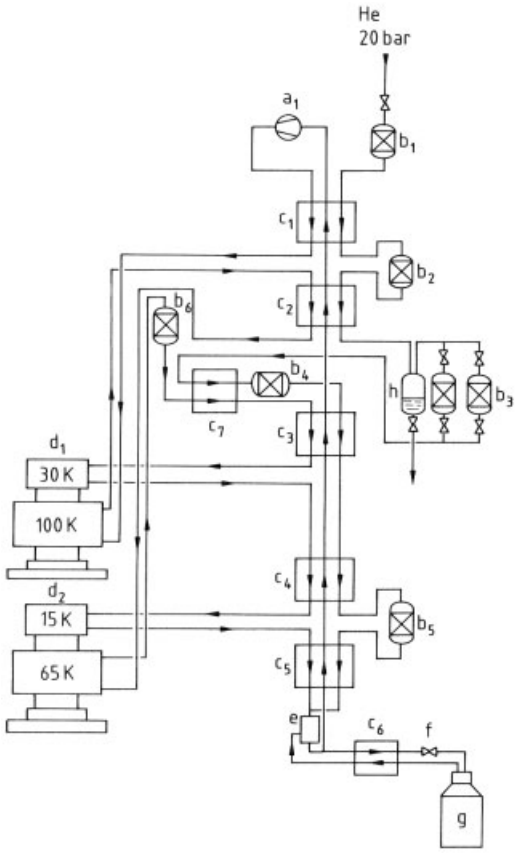

Figure 37. Helium liquefier using Stirling machines
a) Compressor; b) Adsorbers; c) Heat exchangers; d) Gas cryo machines; e) Ejector; f) Joule – Thomson valve; g) Helium tank; h) Condensate separator

Compressor (a) compresses ca. 2.5 g/s helium from 2.5 bar to 20 bar. This is a two-stage machine with the pistons sealed by roller diaphragms, which ensures oil-free transport. The high-pressure gas is precooled to ca. 15 K in the heat exchangers (c) by the gas cryo machines (d). After heat exchanger (c_5) it is throttled to 2.5 bar in the ejector (e). The gas flow is then divided: after (c_6) the smaller part is expanded in (f) to 1.1 bar and transferred to the tank (g) in a partially liquefied state. The vapor fraction of the throttled gas after (c_6) is recompressed to 2.5 bar in (e). The larger part of the flow which is subdivided after the ejector is fed to the suction side of the compressor after heating in the heat exchangers.

As a result of extensive segregation of the refrigeration circuit from the gas to be liquefied the gas cleaning system can be confined predominantly to the latter. The gas is dried in adsorber (b_1) and carbon dioxide is retained in adsorber (b_2). Liquid air is separated in the condensate separator (h); the remaining air impurities are retained in (b_3) and (b_4), while neon and hydrogen are removed in (b_5). Only adsorber (b_6) is used for purifying the gas compressed by the compressor.

5.3.3. Helium Cryo Plant Using the Gifford – McMahon (GM) Process

The GM process is used predominantly for small cryo capacities. Although liquid helium temperatures can be achieved with a GM machine, an additional separate Joule – Thomson (JT) circuit should be used for reasons of energy and operational reliability (Fig. 38) [210].

At 4.3 K the maximum cryo capacity of the plant shown is 3.5 W; the drive capacity is 6.4 kW resulting in a specific energy consumption of ca. 1800 W per Watt. The GM machine (d) used for precooling the helium supplies 35 W at 80 K in the first stage and 10 W at 20 K in the second stage. To produce this cryo capacity, 3.9 g/s He is compressed by a reciprocating compressor (a_1) to 23 bar. The compressor has oil injection and is provided with a downstream oil separator system. The two-stage oil-injection reciprocating compressor (a_2) for the JT circuit-compresses 0.36 g/s He from 1 bar to 18 bar. A multistage downstream oil separating system is also used here. Because of the strict separation of the GM and JT circuits the two systems can be operated at different pressures (better efficiency). Impurities in the JT circuit have no effect on the GM machine and vice versa (higher operational reliability).

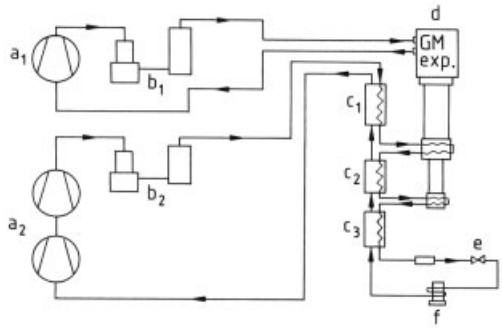

Figure 38. Simplified flow sheet of a cryo plant using the GM process
a) Compressors; b) Oil separators; c) Heat exchangers; d) Gas cryo machine; e) JT valve; f) Consumer

5.4. Process Developments

In addition to the applications of the classical methods for cryo generation already described there has also been investigation and development work into many other principles. Until now these have practically all been used for very small capacities and very low temperatures. Very often the application is in the forefront of the sphere of space cryotechnology. Some examples taken from the great number of possible methods will be described below.

Adsorption coolers are being developed for producing small capacities in the range 150 K to 0.3 K. This method utilizes heat as the driving energy and requires no moving mechanical parts. Figure 39 [211] shows a schematic of a typical design. The refrigerator consists of several compressors, heat exchangers, precoolers, and a JT valve. The compressors are pressure vessels with adsorbents and they are provided with heating and cooling devices. The functional principle on which this process is based relies on the fact that certain porous materials and sintered metals have a high, temperature-dependent storage capacity for specific gases. This fact is utilized for absorbing, compressing, and expelling

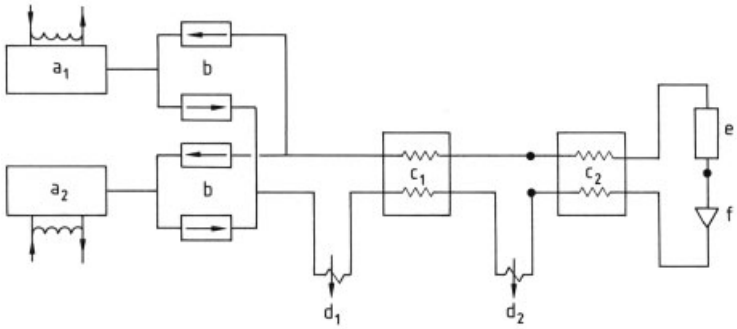

Figure 39. Typical design of an adsorption cooler
a) Compressors; b) On-off valves; c) Heat exchangers; d) Precoolers; e) Consumer; f) JT valve

the gas. A continuous gas flow can be generated in this way with alternate heating and cooling of the adsorbent (compressor).

By heating the compressor (a_1) the gas is compressed and discharged via the corresponding valve (b). The gas is cooled in the heat exchangers (c) and precoolers (d); the isenthalpic throttling takes place in the JT valve (f). After absorbing heat from the consumer (e) the gas returns via the heat exchangers (c) to the cooled compressor (a_2) where it is adsorbed (sucked in).

Magnetic refrigerators (→ Cryogenic Technology, Chap. 4.8.) were originally developed to generate temperatures in the microkelvin range, but are now also finding applications at higher temperatures. The functional principal is based on the magnetocaloric effect. In magnetic substances the atoms behave like very small bar magnets. Due to their thermal movements the atoms have completely irregular alignments. When an external magnetic field is applied at very low temperature the magnetic dipoles become aligned. If the magnetic field is then switched off again the dipoles attempt to return to the disordered equilibrium state. The energy required for this is withdrawn from the lattice vibrations; the lattice and magnetic moments achieve thermal equilibrium again. This procedure cools the magnetic material.

Figure 40 shows the principle design of a magnetic refrigerator for recondensation of helium [212] as an example of the application of this principle. If a high intensity magnetic field is generated by the pulsed magnets (e) the temperature of the magnetic material (a) increases. To reduce the temperature again heat is removed via the cooling circuit (b). After the heat has been removed the magnet is switched off; there is a further drop in temperature of the magnetic material. Heat can now be taken from the helium bath (c) via the transfer tube (d), during which helium vaporized from the bath is recondensed.

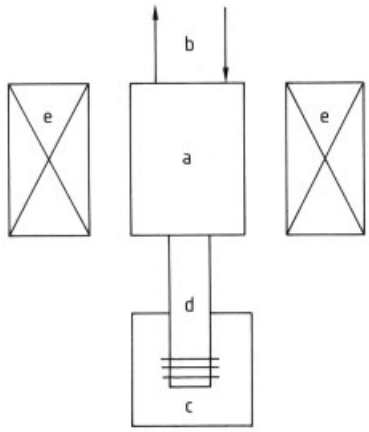

Figure 40. Example of a magnetic refrigerator, principle design
a) Magnetic material; b) Cooling circuit; c) Helium bath; d) Transfer tube; e) Magnets

^{3}He – ^{4}He Dilution Refrigerators (→ Cryogenic Technology, Chap. 4.9.). All the plants described so far use helium-4. However, helium-3 has much to recommend it for achieving very low temperatures down to the mK region because of its significantly higher vapor pressure at the same temperature. At very low temperatures ^{3}He is not soluble in all proportions in ^{4}He; there is an extensive dilution zone. Below 0.1 K the light phase consists of almost 100 % ^{3}He, which, however, accounts for only

approximately 6 % of the heavy phase. During its transition from the light into the heavy phase with the associated dilution the ^{3}He is cooled. This permits the supply of heat from a load at constant temperature.

Figure 41 shows the principle design of a dilution refrigerator [213]. Precooled helium at a pressure in the millibar region is throttled in capillary (a_1). The helium is cooled further in the distillation vessel (b) and the heat exchanger (c) between which is positioned a second throttling capillary (a_2).

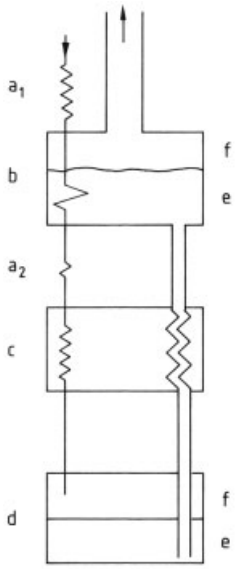

Figure 41. Basic design of a segregation cryo station
a) Throttles (capillaries); b) Distillation vessel; c) Heat exchanger; d) Mixing chamber; e) Aqueous phase; f) Concentrated phase

The helium is now introduced into the mixing chamber (d) where it passes from the light to the heavy phase. After absorbing heat from a load the helium flows from the mixing chamber, through the heat exchanger (c), and into the distillation vessel (b) because of the difference in concentration. There it is vaporized and fed to the compressor.

Although the concentration of ^{3}He in the liquid in the distillation vessel is very low, the vapor consists of virtually 100 % ^{3}He. The helium recirculating in the closed loop is therefore almost pure ^{3}He. The ^{4}He in the system is virtually stationary in the distillation vessel, heat exchanger, and mixing chamber.

6. Quality Specifications and Analysis

New technical advances are leading to increasingly stringent requirements for the purity of noble gases. For example, even traces of impurities in pure gases could adversely affect the operation of gas lasers, and exceeding a given isotopic ratio would limit their use in the nuclear industry. Most gas suppliers use the so-called point notation, where, for example, He 5.8 means a minimum purity of 99.9998 % and Ar 6.0 denotes argon of minimum purity 99.9999 %. Contents of impurities are usually given in vol ppm or ppb [214]; mg/m^3 and mg/kg are also used. Typical purity grades of commercially available noble gases are between 4.0 and 6.0.

The following impurities are found in noble gases:

1) Oxygen and nitrogen, which enter during transfer of noble gases with gas transfer systems that are not gas-tight
2) Hydrocarbons, from contaminated pipes or compressors
3) Water, desorbed from the walls of insufficiently pretreated gas cylinders
4) Carbon dioxide from various sources
5) Perfluorinated hydrocarbons, which are enriched during production of the crude fraction in air separation plants
6) Hydrogen, which can result from the process for the purification of noble gases.

Empirical methods for the analysis of noble gases, such as the lamp test formerly used in the lighting industry [215], have since been superseded by physical or physicochemical analytical methods. Methods such as measurement of the density, vapor pressure, and refractive index have declined in importance, partly because the quantities of gas required are too large [216]. The following are the most widely used methods:

1) Gas chromatography, with various detectors. Low-boiling hydrocarbons can be determined in all noble gases with a detection limit of 0.1 vol ppm by using a flame ionization detector. The so-called total hydrocarbon analyzer also gives the total hydrocarbon content (C_nH_m) with a detection limit of 0.1 vol ppm. The helium ionization detec-

tor can be used to detect traces of H_2, O_2, N_2, CH_4, CO, and other noble gases. Similar sensitivity is exhibited by the argon detector [217]. This is a discharge detector in which the excitation is produced by high-frequency a.c. current. It has proved very sensitive for the detection of impurities in argon.

2) Dispersive and nondispersive (NDIR) infrared spectroscopy allows the quantitative and qualitative determination of heteroatomic molecules having characteristic absorptions. The limit of detection for hydrocarbons is < 1 vol ppm. Similar sensitivity is attainable for CO_2 due to the intense absorption at 2349 cm^{-1} [218, 219]. Due to ease of use, NDIR instruments have become well established in the analysis of gases [220].
3) Moisture in noble gases is determined by using electrolytic or capacitive hygrometers. In laboratories, water content is sometimes determined by measurement of the dew point in a Peltier-cooled apparatus. The moisture content is usually expressed as a dew point. Tables for conversion to vol ppm, mg/L, and other units are available in the literature.
4) Oxygen is determined electrochemically in a further development of the oxygen-specific cell patented by Hersch [221, 222].

7. Uses

7.1. Welding

Argon and helium are used in the welding, cutting, and spraying of metals. Depending on the welding process, the noble gases are used pure, as a mixture, or in combination with oxygen, hydrogen, or carbon dioxide [224]. Typical applications are as follows:

1) Ar, Ar – He mixtures, and He for tungsten inert gas (TIG) and metal inert gas (MIG) welding of aluminum and its alloys [225]
2) High-purity Ar for TIG welding of gas-sensitive metals such as titanium, tantalum, and tungsten
3) Mixtures of Ar with oxygen (up to 8 vol %) and/or carbon dioxide (up to 3 vol %) for MIG welding of high-alloy steels [226]. Part of the argon can be replaced by helium
4) Argon – hydrogen mixtures for TIG welding of austenic steels
5) Argon-based mixtures with high contents of oxygen (up to 12 %) and carbon dioxide (usually 18 %, sometimes more) for MIG welding of construction steels. Part of the argon can be replaced by helium
6) Argon as a plasma gas for plasma welding and argon-hydrogen mixtures (> 5 % H_2) as protective gases for plasma welding

Protective gases for welding are standardized in DIN 32 526. A corresponding EN standard is being prepared. The purity of the gases is very important, usually exceeding 99.99 % with a dew point < -50 °C.

7.2. Illumination

Neon, argon, krypton, and xenon are used as filler gases for lamps, sometimes as constituents of gas mixtures. High-purity gases are required for these applications.

Neon is used as the filler gas for high-voltage tubular lamps. This application is, however, of minor importance. These crimson lamps are used for advertising signs. In addition, neon, with small additions of other gases, is used in low-voltage neon-glow lamps and in starters for fluorescent tubes.

Argon, as a mixture with nitrogen, is the most widely used filler gas for conventional light bulbs.

Krypton is used as the filler gas for high-quality light bulbs. Compared to argon or nitrogen, krypton reduces the rate of evaporation of the tungsten filament and, due to its low thermal conductivity, lowers its rate of cooling. Krypton-filled lamps have higher light efficiencies and color temperatures than lamps filled with other gases. Halogen lamps are filled with krypton that contains controlled amounts of iodine- or bromine-containing compounds. The use of krypton for lamps has increased with the introduction of energy saving lamps.

Xenon is used in gas-discharge lamps. The xenon discharge has an almost continuous spectrum, resembling daylight. Therefore,

Xenon is highly suitable for lamps where the quality of the color is of particular importance. Furthermore, high light efficiencies can be achieved. Xenon lamps are used for projectors, for illumination of film and television recordings, in floodlights, as flashlights, and as navigation lights for runways.

7.3. Other uses

Argon is used in metallurgy as a protective gas for annealing at temperatures at which nitrogen would react with the metal, as a scavenger gas for the degassing of molten metals, and in the AOD process for the production of chromium steels.

High-purity argon is used in the production of silicon for the semiconductor industry. In the production of chips, it is used as a protective gas and as a carrier gas for dopants. It is also used as a protective and carrier gas in spark and ICP spectrometry and gas chromatography.

Helium is used on account of its special properties when use of another gas is impossible or only partially feasible.

European consumption of helium in 1997 broke down as follows:

Low-temperature technology	36 %
Welding, cutting	14 %
Optical fibers	8 %
Breathing mixtures, diving	6 %
Analysis	14 %
Leak detection	9 %
Balloons	7 %
Other uses	6 %

Liquid helium is used in cryogenic technology as a refrigerant, for example, for superconducting magnets:

1) For magnetic resonance imaging tomography in medical diagnosis,
2) In NMR spectrometers,
3) In particle accelerators,
4) In magnetic separators, and
5) For research in solid-state and high-energy physics.

Other uses include cryopumps for ultrahigh vacuum, superconducting sensors (SQUID's), superconducting memories and circuit elements in computers, and low noise amplifiers. Interesting potential uses are the cooling of superconducting power station generators, electromotors for ships, and nuclear fusion reactors.

Gaseous helium is used in rocketry for conveying liquid fuel and for inerting rocket motors.

In diving, helium is used instead of nitrogen to avoid problems of decompression. In medicine it is used to improve the ease of breathing of asthma sufferers, and helium-containing gas mixtures are used to investigate the functioning of the lungs.

Helium is used as a carrier gas in gas chromatography, as a filler gas for thermometers, and in counter tubes for the measurement of high-energy radiation. Helium is of importance in leak detection in plant construction and vacuum technology. It can be used to detect even the smallest leaks. This application exploits the high rate of diffusion of helium and the high sensitivity of helium mass spectrometers.

Helium is used for filling balloons and airships. Its high buoyancy and nonflammability make it especially suitable for this purpose.

Helium is used as a coolant in high-temperature nuclear reactors. Helium-4 is the only element that gives no radioactive products upon high-energy irradiation. The high thermal conductivity of helium is important in this application.

Helium is essential to many high-technology processes, an example being the production of optical glass fibers, in which it serves as a coolant and as a carrier gas for the chlorine used for the removal of hydroxyl groups. Finally, helium is the major component in the gas mixture used in the CO_2 laser.

Neon is used in the excimer laser, in which ittakes part in the energy transfer process. It is also used in the He – Ne laser. Liquid neon is now commercially available and is finding application as an economical cryogenic refrigerant. It has over 40 times more refrigerating capacity per unit volume than liquid helium and more than three times that of liquid hydrogen.

Xenon shows promise for use in medicine and medical applications. Xenon-enhanced computerized tomography and xenon magnetic resonance imaging (MRI) are used for investigating blood flow in the brain and in the lungs. Xenon can also be used as anesthetic in place

of the more familiar nitrous oxide. The physiological impact is enhanced because of the good solubility of xenon in fat tissue [223].

The use of xenon as a fuel for ion drives for space vehicles could allow the replacement of chemical fuels by nuclear or solar energy in space flight. For this purpose xenon is ionized by electrical energy and accelerated to high velocity by electrodes. Ion drives are also suitable for stabilization of missiles. Commercial use can be expected in the near future. In high-energy physics, xenon is used in radiation counter tubes.

8. Storage and Transportation

Gaseous noble gases are transported in steel cylinders comparable to those used for other industrial gases. In Germany this is usually a 50 L steel cylinder with 20 MPa overpressure. Also available are 10 L cylinders, and 1 L and 2 L cylinders for the expensive noble gases neon, krypton, and xenon. Small amounts of noble gases are delivered in glass flasks (1 or 2 L). Disposable pressure cans with a volume up to 1 L and an overpressure of 1.2 MPa are also available. Large quantities of gases are transported in 600 L cylinder bundles or road tankers with 13 – 20 m^3 volume and 20 MPa overpressure. For high-purity requirements, special withdrawal fittings are required that have minimal leak rates and a small dead volume and can be heated and evacuated.

Large amounts of argon are transported in liquid form. Stationary and transport tanks for liquid argon are powder vacuum insulated and resemble tanks for liquid oxygen and nitrogen in design.

Containers for the transport of liquid helium are made of high-alloy stainless steels. The inner container in which the liquid helium is located is connected to the outer container only by a neck. Heat absorption from the surroundings is minimized by superinsulation, high vacuum, and thermal radiation shielding. The superinsulation is a multilayer coating of insulating and reflecting foils around the inner container. The radiation shields between the inner and outer containers carry heat to the tube, which is cooled by evaporating helium. The structure of a liquid helium dewar is shown in Figure 42.

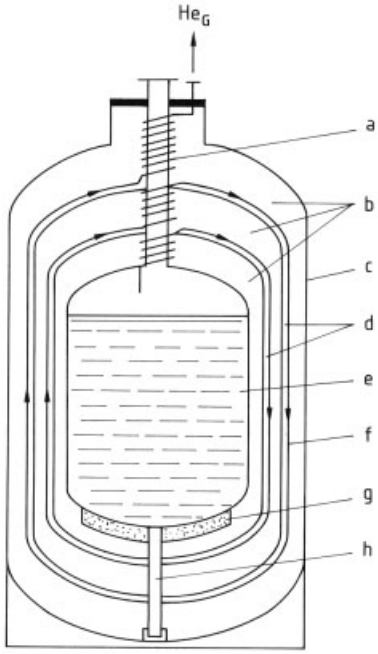

Figure 42. Helium dewar, cooled by evaporating helium
a) Tube; b) Superinsulation; c) Outer jacket; d) Radiation shielding; e) Container for liquid helium; f) Cooling tube system; g) Adsorbent; h) Support

Independent of the filling level, boil-off losses amount to 1 % per day relative to the full container. This means that in a 1000 L container the heat absorption from the surroundings is only 0.3 W/d.

Stationary helium tanks are built with volumes up to 100 m^3. Small amounts of liquid helium are transported in containers with 30 – 500 L volume. For long distance transport, road tankers with 3.7 or 41 3 volume are used.

Liquid helium is usually transferred by pressurization with gaseous helium. The liquid helium passes through a vacuum-insulated siphon into the container to be filled. Helium suppliers collect the resulting evaporated helium, which is then either reliquefied or filled as gaseous helium into cylinders. Such a system is, however, not economically viable for most helium users.

9. Economic Aspects

Helium. In the 1990s helium demand worldwide steadily increased at a growth rate of approximately 7 – 10 % per annum. In 1993 helium

Table 21. World helium production for 99.995 purity (Grade A) or better, in 10^6 m^3 [227]

Country or region	1990	1992	1994	1995	1996	1997	1998
United States	84.9	94.4	100.4	96.1	94.7	107	108
U.S. exports	24.7	30.7	25.0	27.7	22.8	29.5	29.7
Algeria			12	16	16	16	16
Poland	4	4	2	1.9	1.4	1.4	1.4
Former Soviet Union			4.0	4.0	4.2	4.2	4.2
Other countries							
World total	89	98.4	118.4	118	116.3	128.6	129.6

production amounted to some 100×10^6 m^3, 95 % of which was supplied by U.S. private and Bureau of Mines plants. About 30 % of U.S. helium production is exported, and in 1997, for instance, ca. 55 % of U.S. exports were shipped to Asia, mostly Japan, and 20 % to Europe. From 1994 on Poland, Russia and Algeria gradually accounted for 20 % of annual world production. Table 21 shows world helium production statistics. As per December 1997 the world helium annual production plant capacity including plants on standby was 160 million cubic meters. One-third of the helium is delivered in liquid form as a refrigerant.

The price for helium depends on quantity ordered, liquid or gaseous, type and size of cylinders, and regional factors. The following are approximate prices at given conditions: United States: Bulk price f.o.b. helium plant in 1997 \$ 1.98 per m^3 (STP) gaseous, \$ 2.5 per m^3 (STP) as liquid. Europe (1999): € 4 to 10 per liter (in transport cans), He, compressed in steel cylinders from € 30/m^3 (STP) (purity 4.6) to € 80/m^3 (STP) at purity 6.0 (99.9999 %).

Argon. Argon is the only noble gas used industrially in bulk quantities. Most of the argon is produced in air separation plants with an argon rectification column; only a minor amount comes from ammonia tail gas separation units. More than 80 % is used for metallurgy and metal processing. In 1998 ca. 2×10^9 m^3 of argon was used worldwide. Total U.S. argon production amounted to 3500 t/d equal to 0.7×10^9 m^3 per year. One-third of the argon is used in the United States, one-third in Europe and one-third in the rest of the world. Liquid argon shipped by truck costs approximately € 0.8 – 1.5 per m^3 (gaseous), whereas cylinder fillings in laboratory quality costs appr. € 15 per m^3 (STP).

Neon, Krypton and Xenon. These gases are produced exclusively in air separation plants. Increasing amounts are being used for lighting, lasers, and medical application. Given high growth rates and new applications, the availability of these gases, particularly the extremely rare gas xenon, could become a problem in the future: 100 m^3 of pure xenon corresponds to the annual production of an air separation plant producing 50 000 m^3/h of oxygen. World production of xenon in 1998 was estimated to be 5000 – 7000 m^3 per annum.

The prices of these rare gases depend on quantity ordered, the purity required, type and size of cylinders, and regional factors. The following are approximate prices for the purchase of small quantities in Europe (1999):

Neon: € 0.20/L
Krypton: € 1/L
Xenon: € 10/L

10. References

1. *Gmelin,* System no. 1, Edelgase, 1 – 251.
2. A. G. Cook: *Argon, Helium and the Rare Gases,* J. Wiley & Sons, New York 1961.
3. R. E. Stanley, A. A. Moghissi (eds.): *Noble Gases,* CONF 730 915 (Library of Congress 75-27055) ERDA TIC 1973, p. 680.
4. *Kirk-Othmer,* 3rd ed., **12,** 249.
5. M. L. Klein, J. A. Venables (eds.): *Rare Gas Solids,* **vols. I and II,** Academic Press, London, New York, San Francisco 1976, 1977.
6. N. J. Lockyer, *Nature (London)* **3** (1871) 221.
7. Helium Symposia Proceedings in 1968 – A Hundred Years of Helium, *Inf. Circ. U.S. Bur. Mines* **8471** (1969) 199 – 228.
8. M. W. Travers: *Discovery of the Rare Gases,* E. Arnold, London 1928.
9. L. Pauling, *J. Am. Chem. Soc.* **55** (1933) 1895.

10. N. Bartlett, *Proc. Chem. Soc. London* 1962, 218.
11. R. Hoppe, *Angew. Chem. Int. Ed. Engl.* **1** (1962) 599.
12. S. Angus, K. M. Renck (eds.): *International Thermodynamic Tables of the Fluid State. Helium 4,* Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt 1977.
13. V. A. Rabinovich, A. A. Vassermann, V. I. Nedostup, L. S. Veksler (eds.): *Thermophysical Properties of Neon, Argon, Krypton, and Xenon,* Hemisphere Publishing Corporation, Washington, New York, London 1987.
14. K. R. Atkins: *Liquid Helium,* Univ. Press., Cambridge 1959.
15. C. T. Lane: *Superfluid Physics,* McGraw-Hill, New York, San Francisco, Toronto, London 1962.
16. J. Wilks: *Liquid and Solid Helium,* Clarendon Press, Oxford 1967.
17. W. E. Keller: *Helium-3 and Helium-4,* Plenum Press, New York 1969.
18. E. R. Grilly, *J. Low Temp. Phys.* **4** (1971) 615 – 635.
19. R. A. Scribner, M. F. Panczyk, E. D. Adams, *J. Low Temp. Phys.* **1** (1969) 313 – 340.
20. J. M. Dundon, J. M. Goodkind, *J. Low Temp. Phys.* **18** (1975) 315 – 319.
21. E. R. Grilly, *J. Low Temp. Phys.* **11** (1973) 33 – 52.
22. R. D. McCarty: "Thermophysical Properties of Helium-4 from 2 to 1500 K with Pressures to 1000 Atmospheres," *NBS Tech. Note (U.S.)* **631** (1972) Nov., 161.
23. R. D. McCarty: "Thermodynamic Properties of Helium-4 from 4 to 3000 °R with Pressures to 15 000 psia," *NBS Tech. Note (U.S.)* **622** (1972) Sept., 135.
24. J. S. Dugdale, F. E. Simon, *Proc. R. Soc. London Ser. A* **218** (1953) no. 7, 291 – 310.
25. J. S. Dugdale, J. P. Franck, *Philos. Trans. R. Soc. London, Ser. A* **1076** (1964) no. 257, 1 – 29.
26. R. L. Mills, A. F. Schuch, *J. Low Temp. Phys.* **16** (1974) 305 – 308.
27. J. C. Wheatley, *Physica (Utrecht)* **69** (1973) 218 – 244.
28. K. Mendelssohn: "The World of Liquid Helium" in Helium Symposia Proceedings in 1968, *Inf. Circ. U.S. Bur. Mines* **8417** (1969) 256 – 268.
29. K. Dransfeld: *Naturwissenschaften* **58** (1971) 183 –188.
30. F. London: *Superfluids,* vol. 11, Dover Publications Inc., New York 1964.
31. T. A. Alvesalo, H. K. Collan, M. T. Loponen, M. C. Veuro, *Phys. Rev. Lett.* **32** (1974) 981 – 985.
32. J. Kendall, *New Sci.* **64** (1974) no. 918, 100 – 102.
33. *Landolt-Börnstein,* **IV 4 a,** 172 – 205.
34. Y. Larher, *J. Chim. Phys. Phys.-Chim. Biol.* **65** (1968) 1683 – 1685.
35. D. W. Robinson, *Proc. R. Soc. London Ser. A,* **225** (1954) no. 1162, 393 – 405.
36. R. L. Mills, E. R. Grilly, *Phys. Rev.* **99** (1955) no. 2, 480 – 486.
37. J. S. Dugdale, *Nuovo Cimento Suppl.* **9** (1958) no. 10, 27 – 31.
38. J. C. Stryland, J. E. Crawford, M. A. Mastoor, *Can. J. Phys.* **38** (1960) 1546 – 1547.
39. J. Kraus, E. Uhlig, W. Wiedemann, *Cryogenics* **14** (1974) 29 – 35.
40. R. Radebaugh, *NBS Tech. Note (U.S.)* **362** (1967).
41. R. D. McCarty, *J. Phys. Chem. Ref. Data* **2** (1973) no. 4, 923 – 1041.
42. N. V. Tsederberg, V. N. Popov, N. A. Morozova: *Thermodynamic and Thermophysical Properties of Helium,* Israel Program for Scientific Translations, U.S. Department of Commerce, National Technical Information Service, Springfield/Va. 22 151, Jerusalem 1971.
43. V. Arp, *Cryogenics* **14** (1974) 593 – 598.
44. D. B. Mann: "The Thermodynamic Properties of Helium from 3 to 300 K between 0,5 and 100 Atmospheres," *NBS Tech. Note (U.S.)* **154** (1962) Jan.
45. D. B. Mann: "The Thermodynamic Properties of Helium from 6 to 540 °R between 10 and 1500 psia," *NBS Tech. Note (U.S.)* **154 A** (1962) Jan.
46. W. J. Lick, H. W. Emmons: *Thermodynamic Properties of Helium to 50 000 K,* Harvard Univ. Press, Cambridge, Massachusetts, 1962.
47. R. D. McCarty, R. B. Stewart, *Am. Soc. Mech. Eng. New York* 1965, 84 – 97.
48. W. B. Streett, *J. Chem. Eng. Data* **16** (1971) no. 3, 289 –292.
49. R. Runge, *VDI-Forschungsh.* **529** (1968) 5 – 44.
50. J. Hilsenrath et al.: *Tables of Thermodynamic and Transport Properties of Air, Argon, Carbon Dioxide, Carbon Monoxide, Hydrogen, Nitrogen, Oxygen and Steam,* Pergamon Press, Oxford, London, New York, Paris 1960, pp. 75 – 137.

51. F. Din: *Thermodynamic Functions of Gases,* **vol. 2,** Butterworths, London 1962, pp. 146 – 199.
52. H. S. Brahinsky, C. A. Neel: "Tables of Equilibrium Thermodynamic Properties of Argon," *U.S.C.F.S.T.I. AD Rep.* (1961) 19.
53. F. Theeuwes, R. J. Bearman, *J. Chem. Thermodyn.* **2** (1970) no. 2, 179 – 185.
54. N. J. Trappeniers, T. Wassenaar, G. J. Wolkers, *Physica (Amsterdam)* **32** (1966) 1503 – 1520.
55. J. Juza, O. Sifner, *Acta Techn. CSAV* **17** (1972) 380 –401.
56. M. P. Vukalovitsch, V. V. Altunin, G. A. Spiridonov, *Inzh. Fiz. Zh.* **16** (1969) no. 3, 504 – 509.
57. A. Michels, T. Wassenaar, P. Louwerse, *Physica (Amsterdam)* **20** (1954) no. 2, 99 – 106.
58. A. Michels, T. Wassenaar, G. J. Wolkers, J. Dawson, *Physica (Amsterdam)* **22** (1956) no. 1, 17 – 18.
59. C. Y. Ho, R. W. Powell, P. E. Liley, *J. Phys. Chem. Ref. Data* **1** (1972) no. 2, 279 – 421.
60. H. J. M. Hanley, *J. Phys. Chem. Ref. Data* **2** (1973) no. 3, 619 – 642. H. J. M. Hanley, G. E. Childs, *NBS Tech. Note (U.S.)* **352** (1967) March, 23, 24.
61. J. T. R. Watson: *Viscosity of Gases in Metric Units,* Department of Trade and Industry, Nat. Engng. Laboratory, Fluidsgroup, HMSO, Edinburgh 1972, p. 47.
62. C. Y. Ho, W. Leidenfrost: *Progress in Heat and Mass Transfer,* **vol. 1,** Pergamon Press, Oxford, London, New York, Paris 1969, pp. 56 – 58.
63. J. R. Moszynski, B. P. Singh: *Proc. Symp. Thermophys. Prop., 6th,* Amer. Soc. Mech. Engrs., Atlanta, Georgia, August 6 – 8, 1973, pp. 22 – 31.
64. *Landolt-Börnstein,* **II 5 b,** 53 – 54, 57 – 59.
65. *Landolt-Börnstein,* **IV 4 b,** 595 – 596.
66. *Landolt-Börnstein,* **II 5 a,** 2 – 20; 62 – 69; 517; 524; 528 – 540.
67. E. H. Graf, D. M. Lee, J. D. Reppy, *Phys. Rev. Lett.* **19** (1967) 417.
68. T. A. Alvesalo et al., *Cryogenics* **14** (1974) no. 7, 384 – 390.
69. G. von Minnigerode, *Naturwissenschaften* **55** (1968) 482 – 487.
70. O. V. Lounasmaa: *New Methods for Cooling to the Milli-Degree and Sub-Milli-Degree Region,* Techn. Univ. Helsinki, Finland, May 1969, p. 84.
71. O. V. Lounasmaa: *Experimental Principles and Methods below 1 Kelvin,* Academic Press, London 1974.
72. M. Schinkmann, *Vak.-Tech.* **21** (1972) no. 8, 207 – 214.
73. C. S. Barrett, L. Meyer, J. Wasserman, *J. Chem. Phys.* **47** (1967) no. 2, 740 – 743.
74. S. C. Greer, L. Meyer, C. S. Barrett, *J. Chem. Phys.* **50** (1969) no. 10, 4299 – 4304.
75. P. Van't Zelfde, M. H. Omar, H. G. M. Le Pair-Schroten, Z. Dokoupil, *Physica (Amsterdam)* **38** (1968) no. 2, 241 – 252.
76. R. Heastie, *Proc. Phys. Soc. London* **73** (1959) no. 3, 490 – 500.
77. H. M. Long, F. S. Dipaolo, *Chem. Eng. Prog. Symp. Ser.* **59** (1963) no. 44, 30 – 35.
78. R. Heastie, C. Lefebure, *Proc. Phys. Soc. London* **76** (1960) no. 2, 180 – 184.
79. H. Veith, E. Schröder, *Z. Physik. Chem., Abt. A* **179** (1937) 16 – 22.
80. H. Lawrence Clever (ed.): *Solubility Data Series,* vol. 1, "Helium and Neon," Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt 1979.
81. H. Lawrence Clever (ed.): *Solubility Data Series,* vol. 4, "Argon," Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt 1980.
82. H. Lawrence Clever (ed.): *Solubility Data Series,* vol. 2, "Krypton, Xenon and Radon," Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt 1979.
83. *Landolt-Börnstein,* **II 2 b,** 1 – 210.
84. *Landolt-Börnstein,* **IV 4 c,** 1.
85. *Landolt-Börnstein,* **II 2 a,** 336 – 408, 723 – 724; IV 4 a, 11 – 15, 77 – 80.
86. *Landolt-Börnstein,* **IV 4 c,** 2.
87. K. F. Liu: *Phase Equilibria in the Helium-Carbon Dioxide, -Argon, -Methane, -Nitrogen, and -Oxygen Systems,* Ph. D. Thesis, Georgia Inst. Technology, Atlanta, July 1969.
88. D. S. Ciklis: *Phasentrennung in Gasgemischen,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1972.
89. W. B. Streett, *Astrophys. J.* **186** (1973) no. 3, 1107 –1125.
90. M. J. Hiza, *Fluid Phase Equilib.* **6** (1981) 203.
91. W. B. Streett, J. L. E. Hill, *J. Chem. Phys.* **52** (1970) no. 3, 1402 – 1406.
92. C. Heck, P. Barrick, *Adv. Cryog. Eng.* **12** (1967) 714.
93. J. C. Mallins, W. T. Ziegler, *Adv. Cryog. Eng.* **10** (1965) 171.
94. N. J. Trappeniers, J. A. Schouten, *Physica (Amsterdam)* **73** (1974) 539 – 545.

95. N. J. Trappeniers, J. A. Schouten, *Physica (Amsterdam)* **73** (1974) 546 – 555.
96. H. Schenk, K. W. Lange, *Arch. Eisenhüttenwes.* **44** (1973) no. 5, 389.
97. F. J. Norton, *J. Appl. Phys.* **28** (1957) 34 – 39.
98. L. C. Waters, *J. Am. Ceram. Soc.* **53** (1970) no. 5, 288 – 289.
99. S. Egli, A. Ruf, A. Buck, *Swiss. Chem.* **6** (1984) 90.
100. J. E. Konsh, A. Sofu, *Sep. Sci. Tech.* **19** (1983) no. 8, 723.
101. R. Walker, W. Korns, *J. Membr. Sci.* **55** (1991) 99.
102. F. Schumann: *Bericht SZS,* Staatl. Zentrale für Strahlenschutz, Juli 1972, p. 34.
103. C. M. Slansky, *At. Energy Rev.* **9** (1971) 423 – 440.
104. B. G. Baker, P. G. Fox, *Trans. Faraday Soc.* **61** (1965) 2001 – 2012.
105. A. J. Kidnay, M. J. Hiza, *Adv. Cryog. Eng.* **12** (1967) K-5, 730 – 740.
106. *Landolt-Börnstein,* **II 3,** 510 – 511.
107. *Landolt-Börnstein,* **IV 4 b,** 122; 130 – 133.
108. A. van Itterbeek, W. van Dingenen, *Physica (Utrecht)* **5** (1938) 529 – 540.
109. W. Müller, M. Maul, *Z. Phys. Chem. (Wiesbaden)* **83** (1973) 111 – 117.
110. B. G. Aristov, V. Bosacek, A. V. Kiselev, *Trans. Faraday Soc.* **63** (1967) 2057 – 2067.
111. R. M. Barrer, R. Papadopoulos, *Proc. R. Soc. London Ser. A* **326** (1972) 315 – 330. R. M. Barrer, R. Papadopoulos, J. D. F. Ramsay, *Proc. R. Soc. London Ser. A* **326** (1972) 331 – 345.
112. J. M. Fernbacher, L. A. Wenzel, *Ind. Eng. Chem. Fundam.* **11** (1972) 457 – 465.
113. S. Khani, J. Takada, *J. Nucl. Sci. Technol.* **2** (1965) no. 2, 51 – 56.
114. J. L. Kovach: *Krypton-Xenon Adsorption on Nacar Carbons,* Research and Development Department of the North American Carbon, Inc., Columbus, Ohio, Report NACAR-010004, March 1970; NACAR-010005, June 1970.
115. E. Lejarovski, L. Zekov, M. Sakharieva, *Izv. Fiz. Inst. ANEB, Bulg. Akad. Nauk.* **15** (1966) 19 – 31.
116. E. Wirsing, Jr., L. P. Hatch, B. F. Dodge: *Brookhaven Nat. Lab. Rep. BNL* **50 254 (T-586),** June 1970, February 1971.
117. K. F. Chackett, D. G. Tuck, *Trans. Faraday Soc.* **53** (1957) 1652 – 1658.
118. W. E. Clark, R. E. Blanco: *Oak Ridge Nat. Lab. Rep. ORNL (U.S.)* **4473** (1970).
119. P. Laszlo, G. Schrobilgen, *Angew. Chem.* **100** (1988) 495.
120. *Gmelin,* vol. 1, Supplement, Edelgasverbindungen (1970).
121. J. H. Holloway: *Noble Gas Chemistry,* Halstedt, New York 1973.
122. N. Bartlett, F. O. Sladky: "The Chemistry of Krypton, Neon and Radon" in *Comprehensive Inorganic Chemistry,* vol. 1, Pergamon, New York 1973, pp. 213 – 330.
123. K. Seppelt, D. Lentz: "Novel Developments in Noble Gas Chemistry," *Progress. Inorg. Chem.* **29** (1982) 167 – 202.
124. D. Cremer in Z. B. Maksic (ed.): *Theoretical Models of Chemical Bonding,* part 2, Springer Verlag, Berlin 1990.
125. G. von Bünan: "Edelgasreaktionen in der Strahlenchemie," *Fortschr. Chem. Forsch.* **5** (1965) no. 2, 347 – 393.
126. R. Janoschek, *Nachr. Chem. Tech. Lab.* **36** (1988) 496.
127. G. Frenking, D. Cremer, *J. Am. Chem. Soc.* **109** (1987) 5917.
128. G. Frenking, W. Koch, J. Gauss, D. Cremer, *J. Am. Chem. Soc.* **110** (1988) 8007.
129. G. J. Schrobilgen, *J. Chem. Soc. Chem. Commun.* 1988, 863.
130. D. R. Mackenzie, *Science* **141** (1963) 1171. *Brookhaven Nat. Lab. Rep. BNL* **7307** (1963) 1 – 3.
131. D. R. Mackenzie, J. Faber, *Inorg. Chem.* **5** (1966) 699 – 700.
132. F. Schreiner, J. G. Malm, J. C. Hindman, *J. Am. Chem. Soc.* **87** (1965) 25.
133. J. Slivnik et al., *J. Fluorine Chem.* **5** (1975) 273.
134. V. D. Klimov et al., *Nachr. Chem. Tech.* **23** (1973) 393.
135. E. Jacob, D. Lentz, K. Seppelt, A. Simon, *Z. Anorg. Allg. Chem.* **472** (1981) 7.
136. H. Selig, P. K. Gallagher, L. B. Ebert, *Inorg. Nucl. Chem. Lett.* **13** (1977) 427.
137. G. J. Schrobilgen, *J. Chem. Soc. Chem. Commun.* 1988, 863.
138. L. Stein, W. W. Henderson, *J. Am. Chem. Soc.* **102** (1980) 2856.
139. K. Seppelt, H. H. Rupp, *Z. Anorg. Allg. Chem.* **409** (1974) 331.
140. B. Weinstock, E. E. Weaver, C. P. Knop, *Inorg. Chem.* **5** (1966) 2189.
141. L. J. Turbini, R. E. Aikmann, R. J. Lagow, *J. Am. Chem. Soc.* **101** (1979) 5833.
142. H. J. Frohm, S. Jakobs, G. Henkel, *Angew. Chem.* **101** (1989) 1534.
143. L. Stein, *J. Chem. Soc. Chem. Commun.* 1985, 1631.

144. V. V. Avrorin et al. *Sov. Radiochem. (Engl. Transl.)* **23** (1981) 708.
145. R. J. Gillespie, *Angew. Chem.* **79** (1967) 885.
146. S. Y. Wang, C. C. Lohr, Jr., *J. Chem. Phys.* **60** (1974) 3901. W. H. E. Schwarz, *Angew. Chem.* **86** (1974) 505; *Angew. Chem. Int. Ed. Engl.* **13** (1974) 454. S. Y. Wang, C. C. Lohr, Jr., *J. Chem. Phys.* **61** (1974) 4110. K. S. Pitzer, C. S. Bernstein, *J. Chem. Phys.* **63** (1975) 3849. W. E. Falconer, M. J. Vasile, F. A. Stevie, *J. Chem. Phys.* **66** (1977) 5335.
147. K. Seppelt, *Angew. Chem.* **91** (1979) 199.
148. N. Bartlett et al., *Inorg. Chem.* **12** (1973) 1717.
149. L. K. Templeton, D. H. Templeton, K. Seppelt, N. Bartlett, *Inorg. Chem.* **15** (1976) 2718.
150. L. Mandelcorn, *Chem. Rev.* **59** (1959) 827.
151. *Kirk-Othmer,* 3rd ed., **6,** 179.
152. M. Hagan: *Clathrate Inclusion Compounds,* Van Nostrand, New York 1962.
153. K. O. Lindquist, W. S. Diethorn, *Int. J. Appl. Radiat. Isot.* **19** (1968) 333 – 344.
154. D. E. Palin, H. M. Powell, *J. Chem. Soc.* 1947, 208.
155. W. E. Clark, R. E. Blanco: *Oak Ridge Nat. Lab. Rep. ORNL (U.S.)* **4473** (1970) Feb.
156. *Gmelin,* **1,** 1 – 251.
157. R. Bieri, M. Koide, E. D. Goldberg, *J. Geophys. Res.* **71** (1966) 5243.
158. R. Bieri, M. Koide, E. D. Goldberg, *Science* **146** (1964) 1035.
159. E. F. Hammel: "Helium, Its Past, Present and Future," *Adv. Cryog. Eng.* **25** (1980).
160. M. Gassmann, *Naturwissenschaften* **61** (1974) 127.
161. National Technical Information Service: *Analyses of Natural Gases 1917 – 74,* Springfield, USA 1976.
162. B. J. Moore, *Inf. Circ. U.S. Bur. Mines.* **8870** (1982).
163. NTIS, U.S. Department of Interior Bureau of Mines: Analyses of Natural Gases, 1917 – 1997, CD-ROM: Order No. PB 98-502206INC. Internet: http://www.ntis.gov/fcpc/cpn4186.htm
164. H. Bartenstein, *Erdöl Kohle Erdgas Petrochem.* **35** (1982) 560.
165. J. B. Peterson: Helium, U.S. Geological Survey, Mineral Commodity Summaries, Jan. 1998.
166. *Cryo Gas Int.* **1999,** no. 3, 30.
167. C. L. Bendixsen, G. F. Offutt, B. R. Wheeler, *Chem. Eng. (N. Y.)* **78** (1971) 55.
168. R. von Ammon, *GIT Fachz. Lab.* **87** (1987) no. 10, 906.
169. H. Sartorius: BfS Jahresbericht 1998, 67.
170. H. Sartorius: BfS Jahresbericht 1990, 89.
171. T. Kirsten, *Naturwissenschaften* **57** (1970) 236.
172. W. Kiesl, H. Malissa, Jr. (eds.): *Analyse extraterrestrischen Materials,* Springer Verlag, Berlin 1974.
173. H. Hintenberger, *Naturwissenschaften* **59** (1972) 285.
174. G. P. Peer: "Kontinuierlich arbeitende Anlage zur Gewinnung von reinstem Neon," *Kältetechn. Klim.* **20** (1968) 179 – 181.
175. E. I. Leyarovski, J. K. Georgiev, B. V. Nicolov: "Method for Extracting Neon and Helium from Air," *Cryogenics* **26** (1986) 431 – 434.
176. E. I. Leyarovski, J. K. Georgiev, B. V. Nicolov, A. L. Zahariev "Continuous Process for the Separation of Ne-He Mixtures," *Cryogenics* 28 (1988) 599 –604.
177. E. I. Leyarovski, "A Method for Obtaining of Neon and Helium from Air on an Industrial Scale," *Cryogenics,* 11 (1970) 48 – 52.
178. Japan Oxygen Co., JP-Kokai 63 69703 A2, 88 69703, 1988 (Y. Takashi, W. Akira).
179. W. Rohde, *Linde Berichte aus Technik und Wissenschaft* **71** (1994) Achema ed., 3.
180. M. Streich, P. Daimer: "Recovery of Rare Gases in Air Separation and Ammonia Plants," *Cryotech 73 Proceedings,* IPC Science and Techn. Press, Guildford 1974, pp. 31 – 35; *Linde Reports on Science and Technology* **21** (1975) no. 5, 44 – 48.
181. *Ullmann,* 4th ed., **10,** 290 – 295.
182. R. F. Pahade, J. J. Maloney, T. F. Fisher, *Energy Prog.* **8** (1988) 196.
183. *Kirk-Othmer,* 3rd ed., **12,** 249 – 287.
184. P. Grimm, U. Rathmann, *Linde Ber. Tech. Wiss.* **51** (1982) 15.
185. H. Kranz, *GWF Gas Wasserfach: Gas/Erdgas* **114** (1973) 345.
186. W. M. Deaton, R. D. Haynes, *Pet. Refiner* **40** (1961), no. 3, 205.
187. W. Förg, P. Wirtz, *Linde Ber. Tech. Wiss.* **28** (1970) 44.
188. P. Butz, G. Bengeser: Presentation, *Jahrestreffen der Verfahrensingenieure,* Munich, Sept. 1984.
189. *Kirk-Othmer,* 3rd ed., **15,** 92.
190. S. L. Matson, J. Lopez, J. A. Quinn, *Chem. Eng. Sci.* **38** (1983) 503.
191. H. Knapp, D. Bender, *Verfahrenstechnik (Mainz)* **9** (1975) 161.
192. Company Brochure: Ube Gas Separation System by Polyimide Membrane, Tokyo 1988.

193. J. S. Choe, R. Agrawal, S. R. Anvil, T. R. White, *Presentation 67th CPA Annual Convention,* March 14 – 16th, 1988, Dallas, Texas.
194. Press Release, HWT Mühlheim/Ruhr, FRG, 1986.
195. M. Streich, *Hydrocarbon Process.* **49** (1970) no. 4, 86.
196. J. K. Jones, J. M. Stacey, *Cryogenics* April 1974, 198.
197. D. I. Limb, *Chem. Eng.* (1974) Dec. 9, 80.
198. C. A. Pruitt, J. V. O'Brian, *Oil Gas J.* **87** (1989) Oct. 9, 78.
199. *Winnacker Küchler,* 4th ed., **3,** 623 – 624.
200. W. S. Hwang, D. J. Kamrath: *Proceedings of Int. Conference Cryogenics and Refrigeration,* May 22 – 26, Southampton 1989, p. 44.
201. R. Fabian, D. Tillmann, *Linde Reports on Science and Technology* (1987) no. 42, 12.
202. R. Krishnamurthy, S. L. Lerner, D. L. MacLean, *Gas Sep. Purif.* **1** (1987) 16.
203. G. Kobelt, W. Otto, *Chem. Techn. (Leipzig)* **19** (1967) 449.
204. *Ullmann,* 4th ed. **10,** 269.
205. G. Kobelt, *Luft- Kältetech.* (1980) no. 3, 146.
206. R. D. McCarty, *NBS Tech. Note (U.S.)* **631** (1972).
207. J. Clausen, A. Patzelt, A. Stephan, M. Wanner: *The Linde Turborefrigerator for MR-Tomographs,* Paper FC 08, CEC (1989).
208. H. R. Barton Jr. et al., *Adv. Cryog. Eng.* **31** (1986) 631 – 645.
209. G. J. Haarhuis, *Philips Tech. Rundsch.* **29** (1968) 202 –209.
210. Y. M. Kang, K. Sakitani, S. Taneya, K. Miura: *Proceedings of the 12th International Cryogenic Engineering Conference,* Southampton 1988, pp. 582 – 586.
211. L. Wade, P. Sywulka, M. Hattar, J. Alvarez: *High Efficiency Sorption Refrigerator Design,* Paper HB 04, CEC (1989).
212. H. Nakagome, N. Tanji, O. Horigami, H. Ogiwara, *Adv. Cryog. Eng.* **29** (1983) 581 – 588.
213. D. Petrac, U. E. Israellson, H. W. Jackson, D. M. Strayer: *JPL Research to Develop a $^3He/^4He$ Dilution Refrigerator for Space Applications,* Paper DB 04, CEC (1989).
214. Linde TG GmbH, Cataloge "Spezialgase", 1999.
215. G. A. Cook: *Argon, Helium and the Rare Gases,* **vol. 2,** Interscience, New York 1961, pp. 524 f.
216. K. Jaeger, *Linde Ber. Techn. Wiss.* **19** (1965) 38 – 45.
217. D. Jentzsch, E. Otte; *Detektoren in der Gas-Chromatographie,* Akad. Verlagsges, Frankfurt 1970, pp. 373 ff.
218. Perkin-Elmer, Infrared Bulletin 36, Oct. 1973.
219. H. Gürzler, H. Böck, IR-Spektroskopie, Physik Verlag, Weinheim 1975, p. 222.
220. "Betriebliche Meßtechnik," *VDI Ber.* **199** (1974) 41 –48.
221. P. Hersch, GB 707 323, 1953; 750 254, 1954.
222. A. Verdin: *Gas Analysis Instrumentation,* Macmillan, London 1973, pp. 176 – 179.
223. *BOC Technology* **Nov.** (1996) no. 4, 20.
224. D. Böhme: *DVS Berichtsband 127,* DVS Verlag, Düsseldorf 1989.
225. B. Haas, *Schweiz. Alum. Rundsch.* **5** (1982).
226. H. Geipl: *DVS Berichtsband 74,* DVS Verlag, Düsseldorf 1982.
227. U.S. Geological Survey, Mineral Commodity Summaries, Jan. 1998.

Nonlinear Optics

Fiber Optics is a separate keyword

Hyun N. Yoon, Hoechst Celanese, Summit, New Jersey 07901, United States (Chaps. 1 and 5)

Robert A. Norwood, Hoechst Celanese, Summit, New Jersey 07901, United States (Chaps. 2 and 4)

Hong-Tai Man, Hoechst Celanese, Summit, New Jersey 07901, United States (Chap. 3)

1. Introduction 1
2. Nonlinear Optical Phenomena . . . 2
2.1. Macroscopic Nonlinear Optics . . . 3
2.1.1. Maxwell Equations and Constitutive Relations 3
2.1.2. Second-Order Nonlinear Optical Phenomena 3
2.1.3. Third-Order Nonlinear Optical Phenomena 5
2.2. Microscopic Nonlinear Optics 8
2.2.1. Constitutive Relations and Local Fields 8
2.2.2. General Classical Models 8
2.2.3. Quantum Mechanical Treatment of Nonlinear Optics 9
3. Second-Order Nonlinear Optical Materials 9
3.1. Inorganic Materials 9
3.2. Organic Materials 11
3.3. Nonlinear Optical Polymers 11
3.4. Examples of SHG Materials 13
3.5. Examples of Linear Electro-Optical Materials 13
3.6. Thin Film Applications 14
4. Third-Order Nonlinear Optical Materials 15
4.1. Fundamental Requirements for $\chi^{(3)}$ Materials 15
4.2. Measurement of $\chi^{(3)}$ Properties . . 16
4.3. Inorganic Materials 17
4.4. Organic Materials 18
5. Nonlinear Optical Devices and Applications 20
5.1. Electro-Optic Devices 20
5.2. Frequency Doubling and Parametric Conversion 22
5.3. Third-Order Nonlinear Optical Devices 22
5.4. Nonlinear Optics in Optical Fibers 23
6. References 24

Abbreviations used in this article:

ADP ammonium dihydrogen phosphate
BBO barium borate
DAN 2-(*N,N*-dimethylamino)-5-nitroacetanilide
DCSHG d.c.-induced second harmonic generation
FWM four-wave mixing
GVD group velocity dispersion
KDP potassium dihydrogen phosphate
KD∗P deuterated potassium phosphate
KTP potassium titanyl phosphate
LAP L-arginine phosphate
LBO lithium triborate
MNA 2-methyl-4-nitroaniline
MQW multiple quantum well
NLO nonlinear optics
OKE optical Kerr effect
OPO optical parametric oscillator
POM 3-methyl-4-nitropyridine-*N*-oxide
SHG second harmonic generation
SLM spatial light modulators
SOS sum-over-states
SRS stimulated Raman scattering
THG third harmonic generation

1. Introduction

Traditional linear optics are based on optical phenomena arising from interaction between light and materials whose optical properties (e.g., refractive index) are independent of the applied light intensity (→ Optical Materials). In nonlinear optics the optical properties of the material are a function of the electrical field strength of high-intensity light (i.e., a laser) or an externally applied electrical field.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_541

Expansion of research activity in nonlinear optics (NLO) has been largely stimulated by the rapid development of laser technology and increasing importance of optical technology in telecommunication and information processing. Advances in NLO science span broadranging frontiers, covering new phenomena (e.g., frequency doubling in optical fibers), optical materials (e.g., NLO polymers), and optical devices (e.g., soliton lasers).

The demonstration of NLO properties required the availability of a high-power, coherent light source (i.e., the laser, → Lasers). Modern NLO owes its beginning to Franken [1] who demonstrated second harmonic generation (i.e., frequency doubling) of a ruby-laser beam with a quartz crystal. In the 1960s much of the fundamental theoretical framework was established by Bloembergen [2]. With the advent of optical fiber telecommunications in the late 1970s (→ Fiber Optics), the industrial importance of NLO was fully appreciated. This resulted in intensive research in NLO and its applications during the 1980s. Information is now often carried on a laser beam in communication processes, data storage, retrieval, sensing, or display. Increasing effort is being made to obtain even greater data manipulation by use of ultrafast optoelectronic interactions and by harnessing the inherent parallelism of optical signals.

In parallel with the development of the understanding of NLO phenomena is the increasing variety and sophistication in NLO materials. Early materials were largely based on inorganic dielectric crystals, and atomic and molecular gases. The possibility that organic materials could provide more efficient NLO properties was first realized in 1969 [3]. In organic materials, NLO properties are related to molecular electronic transitions, and electron density variations are easily obtained; fast-response, large-magnitude NLO properties can therefore be obtained.

The first reported work on NLO polymers was by Meredith [4], in which *N,N′*-dimethylaminonitrostilbene dyes were dissolved in a side- chain liquid crystal polymer and an asymmetric material structure was created by applying an external electrical field. This concept was further extended by Hoechst Celanese to polymeric materials: the NLO chromophores are attached as side chains to methacrylic chains thereby increasing the chromophore concentration in the materials. Furthermore, the electrically poled structure is stabilized and exhibits the long relaxation time of glassy polymer materials.

In polymeric third-order materials, a molecular structure is required that allows a long electron delocalization length. Investigations in the 1980s were focused on polydiacetylenes, polyacetylenes, and other modifications of electrically conducting polymers. Since third-order NLO effects are relatively small in magnitude and the materials tend to absorb too strongly at most wavelengths suitable for optical applications, progress has lagged behind that in the secondorder materials area.

Since the early 1980s the range of NLO materials has been expanded by the production of new artificial semiconductor materials, e.g., multiple quantum wells, quantum wires, quantum dots, and superlattices. Multiple quantum wells are fabricated by growing alternate layers of two different semiconductors, typically GaAs and AlGaAs [5]. This produces an alternating succession of potential hills and valleys for electrons and holes; the carriers are confined in the potential wells, thereby exhibiting quasi-two-dimensional behavior that leads to large optical nonlinearities.

The physics of NLO are reviewed in Chapter 2. Surveys of recent developments in second-order and third-order materials are given in Chapters 3 and 4. Potential applications of the materials and material property requirements are discussed in Chapter 5.

2. Nonlinear Optical Phenomena

Electromagnetic radiation can be regarded as an oscillating electrical field. When a light beam (a form of electromagnetic radiation) is incident on, for example, an inorganic crystal, the ions in the crystal lattice are polarized and oscillate with the radiation. This polarization of the ions in turn induces a small electric field which alters the direction and diminishes the magnitude of the field of light (→ Optical Materials). In linear optics the polarization is proportional to the incident radiation and is responsible for phenomena such as refraction and reflection of light. A laser, however, may have a peak field strength

of 10^7 V/m which is comparable to the forces binding the ions together. Thus polarization induced by lasers can induce highly nonlinear oscillations, i.e., the polarization of the ions is not proportional to the incident radiation. This gives rise to nonlinear phenomena such as optical mixing and harmonic generation.

2.1. Macroscopic Nonlinear Optics

2.1.1. Maxwell Equations and Constitutive Relations

Maxwell equations describe classical electromagnetic phenomena and contain terms to account for nonlinearity. It is through the electric polarization $\boldsymbol{P}$ that nonlinearity enters the Maxwell equations, according to the expansion in the dipole approximation

$$\boldsymbol{P} = \chi^{(1)}\boldsymbol{E} + \chi^{(2)}\boldsymbol{E}\boldsymbol{E} + \chi^{(3)}\boldsymbol{E}\boldsymbol{E}\boldsymbol{E} + \ldots \tag{1}$$

where $\boldsymbol{E}$ is the electric field, $\chi^{(1)}$ the linear optical susceptibility, $\chi^{(2)}$ the second-order NLO susceptibility, etc. The Maxwell wave equation is then written as

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\mu \frac{\partial^2}{\partial t^2}\left[\chi^{(1)}\boldsymbol{E} + \chi^{(2)}\boldsymbol{E}\boldsymbol{E} + \chi^{(3)}\boldsymbol{E}\boldsymbol{E}\boldsymbol{E} + \ldots\right] \tag{2}$$

where

$$\nabla = \left(i\frac{\partial}{\partial x} + j\frac{\partial}{\partial y} + k\frac{\partial}{\partial z}\right)$$

∇ is the spatial derivative of the three cordinates *x, y,* and *z*, $\times$ refers to the cross product of the vectors, and μ is the magnetic permeability. The essence of NLO is embodied in Equation (2). When ultrashort (subpicosecond) light pulses are considered, Equation (1) must be modified since $\boldsymbol{P}$ at time t is a result of electric fields present at times t' ($t' < t$). Advances in ultrashort pulse lasers have now allowed observation of a host of subpicosecond phenomena based on this principle.

2.1.2. Second-Order Nonlinear Optical Phenomena

In the majority of optical materials (i.e., linear optical materials) the second-order term $\chi^{(2)}$ is zero. This is due to fundamental material symmetry considerations; macroscopic second-order NLO phenomena do not occur in centrosymmetric media such as gases, liquids, and most solids. Only materials lacking a center of inversion can exhibit bulk second-order [$\chi^{(2)}$] phenomena; examples are lithium niobate, potassium dihydrogen phosphate (KDP), gallium arsenide, gallium phosphide, lithium tantalate, barium titanate, quartz, copper(I) chloride, 2-methyl-4-nitroaniline, and electrically poled NLO polymers (Chap. 3); these materials also exhibit piezo- and pyroelectric effects. Since interfaces between media are noncentrosymmetric, $\chi^{(2)}$ phenomena are also induced at interfaces and $\chi^{(2)}$ effects have been used to study interfacial phenomena [6]; $\chi^{(2)}$ effects are also found in chiral media [7].

Low-Frequency $\chi^{(2)}$ Phenomena. Several $\chi^{(2)}$ phenomena (e.g., electro-optics, optical rectifications, and magneto-optics) involve the application or production of a low-frequency electromagnetic field. Extensive research in second-order NLO has been directed toward electro-optic materials, waveguides, and devices since materials exhibiting large electro-optic effects under low applied fields are already available. The effects are rapid ($\tau < 1$ ns), making high-speed switching possible. Only low-power light sources (e.g., semiconductor laser diodes) are required and in the early years of NLO a viable waveguide technology was available for lithium niobate, a perovskite crystal electro-optic material.

The *electro-optic effect* is a change in the refractive index n of a medium induced by an applied electric field according to

$$\Delta\left(\frac{1}{n^2}\right)_i = \sum_{j=1}^{3} r_{ij}\boldsymbol{E}_{\mathrm{m}j} \tag{3}$$

where $\boldsymbol{E}_{\mathrm{m}_j}$ is the j-th component of the modulating field $\boldsymbol{E}_{\mathrm{m}}$, r_{ij} the electro-optic tensor, and the index i of the two-dimensional tensor is an integer between 1 and 6.

The electro-optic effect is described by a second-order polarization $\boldsymbol{P}^{(2)}$ resulting from the interaction of an applied low-frequency electrical field $\boldsymbol{E}(\omega_m)$ and an optical field $\boldsymbol{E}(\omega_o)$ such that

$$\boldsymbol{P}^{(2)}(\omega_o \pm \omega_m) = \chi^{(2)}(-(\omega_o \pm \omega_m); \omega_o, \pm \omega_m)\, \boldsymbol{E}(\omega_o)\, \boldsymbol{E}(\pm \omega_m) \quad (4)$$

where $(\omega_o \pm \omega_m)$ is the frequency of light leaving the material, ω_o is the frequency of light entering the material, and ω_m the frequency of the applied electrical signal. The total electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ is given as a continuous function of the position vector $\boldsymbol{r}$ and time t by

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E}_0 + \frac{1}{2}\sum_{i=1}^{N}\left[\boldsymbol{E}(\omega_i)\, e^{-\omega_i t} + \text{c.c.}\right] \quad (5)$$

and similarly for $\boldsymbol{P}^{(2)}(\boldsymbol{r}, t)$; the term c.c. denotes complex conjugate. Equation (4) describes the frequency dependence of $\chi^{(2)}$. The induced polarization frequency is the negative of that on the right-hand side of the equation. Equation (4) shows that the electro-optic effect corresponds to the creation of an induced polarization and thereby an electric field through Equation (2), at sideband frequencies $\omega_o - \omega_m$ and $\omega_o + \omega_m$. Usually ω_m is approximated to zero; this is not, however, reliable at high switching speeds (gigahertz) or when the source bandwidth is less than the modulation frequency ω_m.

In *optical rectification* (an inverse electro-optic effect) two optical waves interact to create a low-frequency electrical field. The advent of subpicosecond pulses of large bandwidth has allowed the use of optical rectification to create terahertz pulses of "low-frequency" radiation in a controlled manner in electro-optic crystals such as lithium tantalate [8].

Magnetic fields and optical fields interact in some media to produce *magneto-optic effects*; optical properties can be changed by applying a magnetic field. Typical magneto-optic media are rare-earth doped glasses and crystals. These materials are being applied to optical isolation, optical data storage, and in modulating devices. Magneto-optics are used for optical isolation because the direction in which an optical beam's polarization is rotated is independent of its direction of propagation.

Interesting effects have been observed in multiple quantum well (MQW) structures (Fig. 1) that involve interplay of absorbed light, generated currents, and applied electric fields [9]. An example is observed in self electro-optic effect devices (SEEDs) [10]. These devices work on the principle that changes in the optical absorption of a MQW can be induced by changes in the electric field perpendicular to the quantum well stacks. When used in conjunction with conventional semiconductor structures such as *p-i-n* diodes, this effect can be used to make optical switches; small switching energies (4 pJ) have been demonstrated [11].

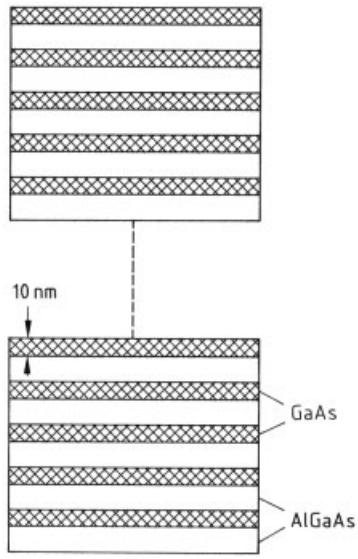

Figure 1. Multiple quantum well (MQW) structure made from alternating layers of GaAs and AlGaAs

All-Optical $\chi^{(2)}$ Phenomena. The first observed second-order all-optical effect was second harmonic generation (SHG) of ruby-laser light in crystalline quartz, i.e., some of the incident light at 694 nm was converted to light with half this wavelength (347 nm) [1]. The induced polarization at the second harmonic frequence is given by

$$\boldsymbol{P}^{(2)}(2\omega) = \chi^{(2)}(-2\omega; \omega, \omega)\, \boldsymbol{E}(\omega)\, \boldsymbol{E}(\omega) \quad (6)$$

The input fields are at ω and ω and have a positive sign on the right-hand side of the semicolon. The output field is at frequency 2ω on the left-hand side of the semicolon; by convention output fields have a negative sign.

The intensity at the second harmonic frequency $I_{2\omega}$ depends on the intensity at the fundamental frequency I_ω according to (no pump depletion)

$$I_{2\omega} = \frac{8\omega^2 d_{\text{eff}}^2 L^2}{n_{2\omega} n_\omega^2} \left(\frac{\mu_0}{\varepsilon_0}\right)^{3/2} (I_\omega)^2 \operatorname{sinc}^2\left(\frac{\Delta k L}{2}\right) \quad (7)$$

where d_{eff} is the effective nonlinearity, L the length of the medium, n_ω and $n_{2\omega}$ the refractive indices at the fundamental and the second harmonic frequencies, μ_0 the vacuum magnetic permeability, ε_0 the vacuum electric permittivity, $\Delta k = 2\omega\,(n_{2\omega} - n_\omega)/c$, and sinc $\xi = \sin \xi/\xi$. (ξ is a dummy variable). The result illustrates two features of second-order NLO: firstly the SHG intensity is quadratic in d_{eff}, L, and I_ω, and secondly it depends on the function $\operatorname{sinc}^2(\Delta kL/2)$.

When phase-matching conditions are satisfied, $\Delta k = 0$ and $\operatorname{sinc}^2(0) = 1$ but when $\Delta k \neq 0$, $\operatorname{sinc}^2(\Delta kL/2)$ reaches its largest value at one coherence length defined by $l_{\text{coh}} = \lambda/4\,(n_{2\omega} - n_\omega)$. The coherence length is the distance over which the beams of fundamental frequency (ω) and the second harmonic (2ω) travel coherently. If phase matching is perfect, the coherence length becomes infinite and a large second-harmonic intensity can be obtained. In most materials l_{coh} is ca. 20 – 30 µm and the amount of SHG is small without phase matching. Phase matching can be accomplished by using crystal birefringence or anomalous dispersion in bulk materials or through modal dispersion in waveguides. Quasi-phase matching in which the optical medium incorporates a nonlinearly varying periodicity equal to l_{coh} has been demonstrated in bulk crystals and waveguides [12, 13].

Second harmonic generation is a special case of the general second-order interaction between optical fields at frequencies ω_1, ω_2, and ω_3, where $\omega_3 = \omega_1 + \omega_2$. For SHG $\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$. In *sum-frequency generation* $\boldsymbol{E}(\omega_2)$ is converted to $\boldsymbol{E}(\omega_3)$ through interaction with an intense pump field $\boldsymbol{E}(\omega_1)$. Sum-frequency generation is useful for converting IR signals that are difficult to detect into visible signals for which detectors with high quantum efficiency (i.e., high sensitivity) are available. Sum frequency generation has been used to generate 450 – 500 nm radiation with a diode laser and a diode pumped Nd : YAG (yttrium aluminum garnet) laser as sources in a $KTiOPO_4$ (KTP) crystal [14]. *Difference frequency mixing* involves generating $\boldsymbol{E}(\omega_2 = \omega_3 - \omega_1)$ from $\boldsymbol{E}(\omega_1)$ and $\boldsymbol{E}(\omega_3)$ through suitable NLO interactions in a medium. Difference frequency mixing can be used to generate tunable IR radiation at wavelengths which are difficult to reach with existing laser sources. If a nonlinear medium is placed in an optical cavity between two appropriate mirrors, injection of an intense pump beam $\boldsymbol{E}(\omega_3)$ into the cavity can result in the spontaneous creation of photons at different frequencies (ω_1, ω_2) with conservation of energy. Those photons fed back into the nonlinear medium can stimulate more emission at their respective frequencies. This device is known as an *optical parametric oscillator* (OPO) and is a source of tunable radiation over large wavelength ranges (hundreds or even thousands of nanometers) [15]. Difference frequency mixing can also be applied to amplify a weak beam (parametric amplification) and for generating sub-Gaussian-noise laser pulses (squeezed states) [16]. The latter are laser pulses which have uncertainties in their amplitudes or phases that are less than the standard limit set by the Heisenberg uncertainty principle.

2.1.3. Third-Order Nonlinear Optical Phenomena

Third-order [$\chi^{(3)}$] NLO phenomena are symmetry allowed in all media. However, they are usually observed at laser powers that are several orders of magnitude higher than those required for $\chi^{(2)}$ effects. This is because the ratio of successive terms in Equation (1) is given by $|\boldsymbol{P}^{(n+1)}/\boldsymbol{P}^{(n)}| \approx |\boldsymbol{E}/\boldsymbol{E}_{\text{atomic}}|$, where $\boldsymbol{E}$ is the applied electric field and $\boldsymbol{E}_{\text{atomic}}$ is the average atomic electric field (10^4 V/µm) on an electron. An optical field of equivalent strength would correspond to intensities of 5×10^{14} W/cm^2. If 100 kW/cm^2 is assumed to be a reasonable intensity for NLO interactions $|\boldsymbol{E}| \approx 0.4$ V/µm, so that $|\boldsymbol{E}/\boldsymbol{E}_{\text{atomic}}| \approx 10^{-5}$. Large $\chi^{(3)}$ effects can, however, be observed when there is resonance between the sum of the applied field frequencies and a material energy level.

Macroscopic $\chi^{(3)}$ phenomena can result from a variety of microscopic manifestations, one of which is nonlinear distortion of electron clouds (the primary mechanism for $\chi^{(2)}$ effects). Two other important mechanisms leading to $\chi^{(3)}$ responses are based on orientational and thermal effects. Orientational effects are due to nuclear

motion in response to applied optical or low-frequency electrical fields resulting in a field-dependent anisotropy of the refractive index. Thermal effects involve absorption of light by the system and dissipation of some of the absorbed energy as heat. The resulting temperature rise produces local changes in density which are manifested as refractive index changes.

Intensity-Dependent Refractive Index Phenomena. These are NLO effects in which the refractive index n of a medium is changed by an intense optical or electrical field, the change in n is proportional to the square of the applied field

$$n = n_0 + \frac{1}{2} n_2 c \left(\varepsilon\varepsilon_0\right)^{1/2} \left|\boldsymbol{E}_{\mathrm{pump}}\right|^2 \tag{8}$$

where c is the speed of light, $|\boldsymbol{E}_{\mathrm{pump}}|$ the pump field amplitude, n_0 the linear index of refraction, n_2 the nonlinear index of refraction, ε the dielectric constant of the medium, and ε_0 the vacuum permittivity. For an electric field with two main Fourier components, $\boldsymbol{E}(\omega_0)$ (the pump field) and $\boldsymbol{E}(\omega)$ (the probe field) where $|\boldsymbol{E}(\omega_0)| >> |\boldsymbol{E}(\omega)|$, the polarization at ω is

$$\boldsymbol{P}^{(3)}(\omega) = \chi^{(3)}(-\omega;\omega,\omega_0,-\omega_0)$$

$$\boldsymbol{E}(\omega_0)\,\boldsymbol{E}(-\omega_0)\,\boldsymbol{E}(\omega)$$

where ω_0 is the frequency of the electrical field perturbing the material and ω the frequency of the beam used to test the effect of the pump field on the properties of the material. When $\omega_0 = 0$, the *d.c. Kerr effect* results, in which birefringence is induced by an electric field. To detect this the sample is placed between crossed polarizers and the increase in transmission is measured on application of a large d.c. field [17] (see also Section 4.2). The amount of probe light transmitted is quadratic in the intensity of the d.c. field (Eq. 8).

When ω_0 is an optical frequency, optically induced birefringence (i.e., the *optical Kerr effect*, OKE) occurs. This was first observed in 1964 in several molecular liquids (carbon disulfide, benzene) with a Q-switched ruby laser [18]. Studies using 100 fs pulses show a variety of new molecular motion phenomena in molecular liquids [19]. In solid media the OKE is due solely to electron cloud distortion and is the basis for many proposed ultrafast all-optical devices.

Self action effects are $\chi^{(3)}$ phenomena that require only one laser beam and do not generate any new frequencies. Examples are self focusing and self phase modulation. *Self focusing* (Fig. 2) occurs because all laser beams have a lateral extension with the peak intensity at the center. In $\chi^{(3)}$ materials the refractive index may increase with laser intensity (Eq. 8). The laser beam creates a refractive index profile for itself that mimics its own intensity profile. When $n_2 > 0$, the beam is focused and its intensity increases until either dielectric breakdown occurs or a stable self-guided beam is formed in which positive self focusing balances diffraction. Self focusing

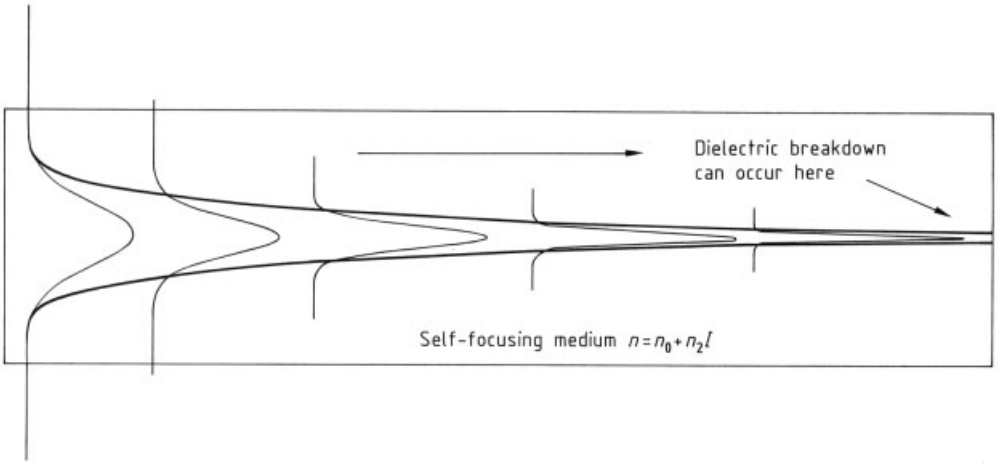

Figure 2. Self focusing in a material with an intensity-dependent refractive index

has implications for areas as diverse as optical computing and laser fusion. Self-guided waves (spatial solitons) have been observed in glass [20] and thermal defocusing effects have been seen in semiconductors and liquids.

Self phase modulation denotes that the phase of a beam depends on its intensity and has important implications for ultrafast optical signal processing and pulse propagation. It is observed in optical media with an intensity-dependent refractive index. The uncertainty principle states that subpicosecond optical pulses should have very large frequency bandwidths. Most materials have a positive refractive index dispersion causing a short pulse to spread out (the red parts of the pulse travel faster than the blue parts). This group velocity dispersion (GVD) can be made negative (i.e., the blue parts of the pulse travel faster than the red parts) in wavelength regions where the material or waveguide exhibits anomalous dispersion, while self phase modulation works in the same direction as positive GVD. With an appropriate temporal intensity profile these effects can cancel and temporal solitons result, i.e., pulses that do not spread out. Solitons have been demonstrated in optical fibers [21] and the same principle has been used to shape and compress pulses [22] (see also Section 5.4).

Frequency Generation Phenomena. Many $\chi^{(3)}$ phenomena result in the creation of new frequencies and, as in the $\chi^{(2)}$ case, phase matching and material transparency are difficult to achieve. The simplest frequency-generating $\chi^{(3)}$ effect is *third harmonic generation* (THG), in which part of a laser beam is converted into a beam whose wavelength is one third of the original beam; a field with a frequency three times the fundamental frequency is generated. As in SHG, $\boldsymbol{E}(3\omega)$ builds up over a coherence length $l_{\mathrm{coh}} = \lambda/6\,(n_{3\omega} - n_{\omega})$, which is of the order of $10-20\,\mu\mathrm{m}$. Phase matching is not possible unless the material is highly birefringent or anomalous dispersion effects are used. THG is a useful characterization technique for $\chi^{(3)}$ materials (see Section 4.2).

Since $\chi^{(2)}$ is a third-order tensor it behaves very much like a vector (a first-order tensor). In a symmetric material these odd-order tensors cannot exist because they are cancelled. If a high d.c. field is applied to a solution of polar molecules molecular alignment occurs breaking the center of symmetry and allowing SHG to occur; this is known as d.c.-induced second harmonic generation (DCSHG) or electric-fieldinduced second harmonic. Since three input fields are involved (the applied d.c. field and the input optical beam which is counted twice), this is a $\chi^{(3)}$ phenomenon with frequency dependence $\chi^{(3)}(-2\omega;\omega,\omega,0)$. This technique is used for measuring molecular hyperpolarizabilities β (corresponding to $\chi^{(2)}$) and γ (corresponding to $\chi^{(3)}$) of molecules in solution [23].

Stimulated Raman scattering refers to the generation of a coherent light beam of a new frequency when an incoming laser beam interacts with a material that is undergoing a vibrational transition. It is another frequency-generating NLO effect that is automatically phase-matched and can be used to achieve $10-20\,\%$ conversion of light to Raman-shifted wavelengths. With optimal focusing and control over polarization, discrete tunability with good conversion efficiencies can be achieved over many hundreds of nanometers in the visible and near infrared. The best media for stimulated Raman scattering are molecular gases and liquids.

Four-wave mixing occurs when three incident beams with wave vectors $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, and $\boldsymbol{k}_3$ interact to create a fourth beam at $\boldsymbol{k}_4$. Conservation of momentum gives

$$\boldsymbol{k}_4 = \boldsymbol{k}_1 \pm \boldsymbol{k}_2 \pm \boldsymbol{k}_3 \tag{9}$$

where $\pm$ refers to the electric field. When $\omega_4 = \omega_1 = \omega_2 = \omega_3 = \omega$, degenerate four-wave mixing results and is automatically phase-matched. Degenerate four-wave mixing can occur through nonresonant electronic $\chi^{(3)}$ effects, light-induced concentration gratings, and thermal effects (thermal gratings). It can be used to create phase- conjugate mirrors that "undo" image distortions and reflect the incident beams back on themselves (Section 5.3). Spectacular effects are observed in crystals such as barium titanate and lithium niobate where the microscopic origin of the effect (photorefraction) is the motion of photo-induced electrical changes coupled to an electro-optic effect [24].

Resonant and Nonresonant $\chi^{(3)}$ Effects. The treatment of resonant NLO effects includes the analysis of populations of electronic excited

states. Resonance occurs when some algebraic combination of applied frequencies is equal to an eigenfrequency (inherent frequency) of the system and large enhancement of $\chi^{(2)}$ or $\chi^{(3)}$ can result. With a single applied field of frequency ω_0, one-photon enhancement occurs when ω_0 is close to Ω (a material eigenfrequency), and two-photon enhancement can occur when ω_0 is close to $\Omega/2$. Energy conservation alone does not guarantee large enhancement since, in the case of multiphoton processes, symmetries of the excited states and the ground state of the system are involved. Saturable absorption is an important NLO effect in which large populations of excited state energy levels result in a dramatic increase in the transparency of a system when enough light is present near the resonance frequency. Disadvantages of resonant NLO effects for device applications include high absorption and long recovery times.

2.2. Microscopic Nonlinear Optics

Macroscopic NLO phenomena are based on various microscopic mechanisms.

2.2.1. Constitutive Relations and Local Fields

The constitutive relation between the microscopic dipole moment $\boldsymbol{p}$ and the microscopic electric field $\boldsymbol{e}$ is

$$\boldsymbol{p} = \mu + \alpha \boldsymbol{e} + \beta \boldsymbol{e}\boldsymbol{e} + \gamma \boldsymbol{e}\boldsymbol{e}\boldsymbol{e} + \ldots \tag{10}$$

where μ is the microscopic permanent dipole moment, α the linear polarizability tensor, β the first hyperpolarizability tensor, and γ the second hyperpolarizability tensor. Equation (10) can only be used when a well-defined source of microscopic hyperpolarizability exists. The same symmetry rules as for the macroscopic properties then result, so that only noncentrosymmetric microscopic entities may have nonvanishing β (and nonvanishing μ) while α and γ are nonvanishing for all entities.

The microscopic NLO susceptibilities β and γ are related to the macroscopic NLO susceptibilities $\chi^{(2)}$ and $\chi^{(3)}$ by an averaging procedure. Thermodynamic, orientational, and local field averaging are required along with superposition of a macroscopic number of microscopic units. Orientational averaging is trivial for a gas or a weakly interacting liquid since all orientations may be assumed to be equally probable and the averaging relies solely on geometric factors. For crystals, liquid crystals, and oriented glasses or polymers, the positions and orientations of the microscopic elements must be included in the averaging. It is difficult to perform local field averaging in dense media. The Lorentz–Lorenz local field theory is adequate for gases and weakly interacting liquids and solids. In this theory the relationship between the microscopic field $\boldsymbol{e}$ and the macroscopic field $\boldsymbol{E}$ is $\boldsymbol{e} = (\varepsilon + 2)\boldsymbol{E}/3$, where ε is the dielectric constant of the medium. Connecting the $(n-1)$-order microscopic hyperpolarizability to the nth-order macroscopic NLO susceptibility requires $(n+1)$ local field factors; local field effects can be significant for higher-order NLO phenomena.

2.2.2. General Classical Models

In linear optics many useful results are obtained by modeling the motion of an electron in an atom as a forced simple harmonic oscillator where the forcing field is the applied optical field. The model can be extended to describe NLO effects by including an anharmonic term in the equation of motion:

$$\frac{d^2x}{dt^2} + G\frac{dx}{dt} + \omega_0^2 x + ax^2 = \frac{F}{m} \tag{11}$$

where x is the position of the electron, Γ is a damping term, $a << \omega_0^2/A$ (A is the amplitude of the electron's motion), $\boldsymbol{F}$ is the electric field force ($\boldsymbol{F} = -e\boldsymbol{E}$), e is the electron charge, and m the electron mass. Let $\boldsymbol{E}$ have two primary frequency components

$$\boldsymbol{E} = \left[\frac{1}{2}\boldsymbol{E}(\omega_1)\, e^{-i\omega_1 t} + \text{c.c.}\right] + \left[\frac{1}{2}\boldsymbol{E}(\omega_2)\, e^{-i\omega_2 t} + \text{c.c.}\right] \tag{12}$$

and attempt to solve Equation (11) perturbatively. The zeroth-order solution gives classic Lorentzian oscillators, with two solutions $x^{(1)}(\omega_1)$ and $x^{(1)}(\omega_2)$, corresponding to the two frequencies ω_1 and ω_2. Then

$$\left(x^{(1)}\right)^2 = \left(\frac{a}{2m}\sum_{i=1}^{2}\frac{\boldsymbol{E}(\omega_i)\,e^{-i\omega_i t}}{(\omega_0^2 - iG\omega_i - \omega_i^2)}\right)^2 \quad (13)$$

and the expansion of this quadratic gives terms with frequency dependences $2\omega_1$, $2\omega_2$, $\omega_1+\omega_2$, $\omega_1-\omega_2$, and zero frequency terms. Each of these terms acts as a source for a different frequency component of the motion. Equation (11) can be solved for $x^{(2)}(2\omega_1)$, i.e., microscopic SHG

$$x^{(2)}(2\omega_1) = \frac{\frac{a}{2}(e/m)^2\,\boldsymbol{E}^2(\omega_1)}{(\omega_0^2 - iG\omega_1 - \omega_1^2)\,(\omega_0^2 - 2iG\omega_1 - 4\omega_1^2)} \quad (14)$$

This solution exhibits a quadratic dependence on $\boldsymbol{E}(\omega_1)$, as well as frequency denominators having resonances at $\omega_1 = \omega_0$ and $\omega_1 = \omega_0/2$. As $x^{(2)}$ solutions have more frequency dependence than $x^{(1)}$, higher order terms have even more dispersion. Difficulties in relating NLO measurements on the same material at different light frequencies are due to dispersion factors.

2.2.3. Quantum Mechanical Treatment of Nonlinear Optics

The correct way to describe the interaction of matter with radiation is by using quantum mechanics. The first step in the quantum mechanical treatment is to determine the Hamiltonian of the system, H. Usually $H = H_0 + H'$, where H_0 is the Hamiltonian of the system in the absence of radiation and H' is the interaction between the system and the applied fields, usually taken to be $H' = -e\boldsymbol{r}\cdot\boldsymbol{E}$ for a one-electron system. In the sum-over-states (SOS) approach, the time-independent Schrödinger equation is solved through a self-consistency scheme, such as the Hartree–Fock theory. Time-dependent perturbation theory is then used to compute corrections to the zeroth order wave functions resulting from the presence of H′. The polarization density is calculated as the expectation value of the polarization operator, and the microscopic susceptibilities can be extracted. This procedure gives the following for $\beta_{ijk}(-\omega_3;\omega_1,\omega_2)$:

$$\beta_{ijk}(-\omega_3;\omega_2,\omega_1) = \frac{8\pi^3 e^2}{h^3}P(\omega_1,\omega_2,\omega_i,\omega_j)\left[\sum_{mn}{}' \frac{\langle r_i\rangle_{gm}\langle \bar{r}_j\rangle_{mn}\langle r_k\rangle_{ng}}{(\Omega_{ng}-\omega_j)(\Omega_{mg}-\omega_i-\omega_j)} + \frac{\langle r_j\rangle_{gm}\langle \bar{r}_i\rangle_{mn}\langle r_k\rangle_{ng}}{(\Omega_{mg}^*+\omega_i)(\Omega_{ng}-\omega_j)} + \frac{\langle r_k\rangle_{gm}\langle \bar{r}_j\rangle_{mn}\langle r_i\rangle_{ng}}{(\Omega_{ng}^*+\omega_j)(\Omega_{mg}^*+\omega_i+\omega_j)}\right] \quad (15)$$

where ω_3 is the frequency of the induced polarization wave, ω_1 and ω_2 the applied frequencies, e the electron charge, h Planck's constant, $P(\omega_i, \omega_j, \omega_1, \omega_2)$ a permutation operator, the sum is over all states except the ground state, $\Omega_{ng} = \Omega_n - \Omega_g$ is the frequency difference between states n and g of the system, and the barring operator is defined by $\langle\bar{A}\rangle_{mn} = \langle A\rangle_{mn} - \langle A\rangle_{gg}$. Expressions such as Equation (15) have been used to calculate the microscopic NLO susceptibilities β and γ [25,26].

The above discussion applies only to the motion of electrons, but similar approaches apply for contributions to the microscopic NLO susceptibilities from nuclear motion (vibrational and rotational effects). In general, the macroscopic NLO susceptibilities do not stem directly from their microscopic counterparts; exceptions are SHG, parametric mixing, THG, and nondegenerate four-wave mixing. The macroscopic NLO susceptibility normally depends upon several different microscopic NLO susceptibilities (e.g., DCSHG, $\chi^{(3)}$ depends on μ, β, and γ).

3. Second-Order Nonlinear Optical Materials

3.1. Inorganic Materials

Inorganic Crystals. As discussed in Chapter 2, a material must be noncentrosymmetric for a second-order NLO effect to occur. In addition, it should also be transparent to all frequencies involved in the nonlinear interaction. Quartz satisfies these requirements and is readily available as large, perfect single crystals. Quartz has therefore played an important role in the development of NLO materials. The efficiency of SHG (i.e.,

frequency doubling) in quartz is very small because of the low nonlinearity of the material and the fact that the SHG cannot be phase-matched.

The discovery of phase matching [27] came shortly after the discovery of SHG with the use of ammonium dihydrogen phosphate (ADP) [*7722-76-1*] and potassium dihydrogen phosphate (KDP) [*7778-77-0*] crystals. Unlike quartz these two materials have large birefringences so that an incidence angle of the optical beam can be adjusted to provide phase matching between the fundamental and second harmonic waves. With this new phase-matching technique and the crystals' reasonably high resistance to laser damage, SHG efficiences approaching 10 – 50 % were reported with pulsed lasers. The deuterated form of KDP, designated KD∗P, is standard in commercially available laser systems for frequency doubling from the near infrared into the visible.

The next major advance came with the discovery of lithium niobate [*12031-63-9*], $LiNbO_3$, as an NLO material [28]. Lithium niobate offers several advantages over ADP and KDP. It is nonhygroscopic, hard, can be polished to a good finish, and has a large nonlinear coefficient. In addition, the temperature sensitivity of its birefringence is such that phase matching can be achieved by varying the temperature. Furthermore, soon after the discovery of lithium niobate as an electro-optic material, researchers discovered ways to make controlled waveguides by diffusion of titanium into the lithium niobate when heated at high temperature to increase its refractive index. Lithium niobate is, however, easily damaged when illuminated by a continuous wave laser. This photorefractive effect manifests itself as a track of slightly altered refractive index in the crystal along the path of the laser beam. For many practical purposes this effect is not detrimental but is devastating for efficient phase-matching interactions in which the relative phases of the waves are crucial. Despite this disadvantage, lithium niobate remains a very important NLO material.

New, efficient materials are emerging that show promise for a wide range of optical technologies. β-Barium borate (BBO) [*13701-59-2*], β-BaB_2O_4 [29], contains ionic groups with planar conjugated π-orbitals (as in organic molecules) whose electron clouds are polarized more efficiently by incoming radiation. This results in greater nonlinearity and thus in more efficient SHG. The material is transparent in the UV and exhibits an effective $\chi^{(2)}$ 4 – 6 times that of KDP.

Other materials for frequency-doubling lasers include lithium triborate (LBO) [*12007-41-9*], LiB_3O_5, and L-arginine phosphate (LAP) [*49720-39-0*]. Study of certain ferroelectric materials consisting of octahedral crystals containing oxygen and bonds with large nonlinear polarizabilities (TaO, TiO, etc.) has led to the development of potassium titanyl phosphate (KTP) [*12690-20-9*] [30] which is stable and may replace KD∗P in many applications.

Semiconductors. Semiconductors (e.g., GaAs [*1303-00-0*] and ZnS [*1314-98-3*]) play an important role in electro-optic modulators (devices that switch light partially or totally on or off at a very fast speed) in the IR region; most conventional electro-optic materials (e.g., KDP) are opaque in this region (although $LiNbO_3$ transmits from 0.4 to 5 μm). Apart from the need for a large electro-optic coefficient, the criteria for choosing electro-optic semiconductors include an energy gap larger than the photon energy of the light being modulated. This insures freedom from interband transitions that render the material opaque. Conducting currents from the modulating field result in electrical losses and therefore a good material for electro-optic applications should have high electrical resistivity as well as a low free-carrier absorption (free carriers are the electrons and "holes" in the semiconductor). Optical absorption by free carriers can lead to attenuation of the optical signal and additional heating.

Multiple Quantum Wells. A semiconductor quantum well is a thin heterostructure of a low-band-gap material (e.g., GaAs) sandwiched between two layers of high-band-gap material (e.g., AlGaAs). The thickness of the low-band-gap material is ca. 10 nm or about 30 – 40 atomic layers (Fig. 1). A typical MQW structure might consist of approximately 100 such layers totaling about 1 μm.

Interest in the use of MQW materials in optical modulators has recently increased because the electroabsorptive effect [31] (one of the phenomena responsible for light modulation

in such devices) is ca. 50 times larger than in bulk semiconductors. In addition, the technology of layered semiconductor growth is well established because it has long been used in many optoelectronic devices such as laser diodes (→ Electroluminescent Materials and Devices, Chap. 3.).

3.2. Organic Materials

Origin of Optical Nonlinearity in Organic Materials. Interest in organic materials has steadily grown since the 1980s due to their potentially superior properties over inorganic compounds. In inorganic crystals, optical nonlinearity originates from oscillations of positive and negative ions which have relatively large masses. This restricts both the speed of oscillation and displacement. The origin of nonlinear phenomena in organic materials lies in their conjugated π-electrons; this involves transport of electrons, responses are therefore both fast and large in amplitude.

Organic molecules such as benzene have centrosymmetric structures and therefore do not show second-order NLO properties in the electric-dipole approximation. The necessary acentric structure for a second-order material can be provided in the benzene ring by distorting the π-electron cloud. This can be achieved by replacing some of the hydrogen atoms on the ring with functional groups that act as strong electron donors or acceptors. In 4-nitroaniline, the molecular ground-state dipole moment points nearly axially across the ring from the amino group to the nitro group. Since the amino group is an efficient electron donor and the nitro group an efficient electron acceptor, the excited state is relatively stable and has a strong nonlinear response.

Organic Crystals. A number of organic single crystals that show considerable optical NLO effects have been identified and synthesized. These include urea [*57-13-6*] [32], 2-methyl-4-nitroaniline (MNA) [*99-52-5*] [33], 2-(*N,N*-dimethylamino)-5-nitroacetanilide (DAN) [*5367-36-2*] [34], 3-methyl-4-nitropyridine *N*-oxide (POM) [*1074-98-2*] [35], (2,4-nitrophenyl)-aminopropanoate (MAP) [*10420-63-0*] [36], and *N*-(4-nitrophenyl)-(*S*)-prolinol (NPP) [*88422-19-9*] [37]. Research into organic crystals for second-order applications primarily involves the ability to grow asymmetric crystals of sufficient optical quality and size. Structural tailoring of the molecular units sometimes permits the realization of asymmetry and is an important feature of organic materials.

The disadvantages of many organic crystals, however, are that they lack mechanical strength and are highly suceptible to chemical attack, since the molecular lattice is preserved only by relatively weak Van der Waals forces.

Langmuir – Blodgett Films. Certain organic molecules exhibit amphiphilic behavior because they contain both hydrophilic and hydrophobic regions which can be preferentially ordered as a monomolecular layer at an air–water interface. By compressing the film to nanometer dimensions, and employing a dipping process, this film can be transferred onto a suitable substrate. Multilayered structures (Langmuir – Blodgett films) can be formed over relatively large areas by repeated dipping. An example of an organic molecule used for this technique is the merocyanine dye described in [38].

A high degree of noncentrosymmetric polar order can theoretically be achieved using this method. In practice, however, stable noncentrosymmetry can only be maintained over a few layers. For waveguide applications, ca. 1 μm thick guiding layers are required, corresponding to more than 500 layers. It is therefore unlikely that the Langmuir – Blodgett technique will find applications in the fabrication of NLO waveguides.

3.3. Nonlinear Optical Polymers

Organic Guest – Host Systems. Large areas can be covered with nonlinear optical films by using techniques such as spin coating or dip coating. Thicknesses of up to 20 μm can easily be obtained. The spin solution can be prepared by dissolving a NLO-active guest molecule with an inactive host polymer [e.g., poly(methyl methacrylate)] in a suitable solvent. After spinning, drying, and solvent evaporation a film of the nonlinear molecules remains embedded in the host polymer.

For a material to possess second-order nonlinear effects, it must be noncentrosymmetric. The asymmetry in a spin-coated film can be achieved with electric field poling described below.

Poling of NLO Polymers. Electric field poling involves heating the polymer to a temperature at which it is in a viscous state [39]. A strong electrical field is then applied, and active NLO chromophores align as a result of interaction of their dipole moment with the electric field. This process usually takes place in the vicinity of the polymer glass transition temperature (T_g) and is counteracted by thermal disordering effects. When equilibrium is reached (usually in a few minutes) the polymer is cooled to room temperature before the electric field is removed and the alignment is thus frozen in a metastable state. The electro-optic coefficient r of the resulting poled polymer can be written as

$$r \sim N f \beta \frac{\mu E}{kT} \tag{16}$$

where N is the number density, f the local field correction, β the intrinsic molecular second-order nonlinearity, μ the dipole moment of the NLO chromophore, E the poling field, k the Boltzmann constant, and T the temperature. Therefore, to achieve large macroscopic second-order nonlinear effects the $\mu\beta$ product can be increased or the applied poling field can be maximized, the latter is limited by the dielectric strength of the polymer. Poling efficiencies of 10 – 20 % are typically achieved.

Thermal Stability of Poled Polymers. Long-term stability of the poled polymer is critically important because the poled structure is inherently in a nonequilibrium state. Relaxation in polymers arises from thermal reorientational effects whose rate is governed by the molecular mobility. The mobility is in turn determined by environmental parameters including storage temperature and molecular properties such as the glass transition temperature T_g and the free-volume content. The higher the T_g of the polymer, the more stable is its poled structure [40].

Cross-Linked Polymer Matrix. In some poled NLO polymers the chromophores are covalently bonded in a cross-linked polymer matrix [41]. The NLO chromophore (e.g., 4-nitro-1,2-phenylenediamine [41]) is reacted with an optically passive bifunctional epoxy monomer to form a soluble prepolymer that can be spin coated onto a flat substrate. Since this involves a chemical reaction at elevated temperature (ca. 140 °C), poling must be accomplished before curing is complete. A mild precure produces a gel that can withstand reasonable poling electric fields without breakdown before completion of the chemical reaction.

Main-Chain and Side-Chain Polymers. Polymeric solid solutions have several disadvantages in practical applications. They suffer from low packing densities of the NLO chromophores due to phase separation, opacity, low breakdown fields, and poling stabilities that are below commercial requirements. Polymers now being investigated generally utilize chemical bonding between the "guest" NLO and the inactive "host" material. Polymers have been synthesized in which the NLO chromophores are incorporated in the main polymer chain, so the dipoles all point in the same direction (head-to-tail). This molecular configuration offers a very high chromophore density and a potentially more stable alignment than guest – host systems. However, it is questionable whether effective poling alignment can be achieved in such systems.

Covalent attachment of the active NLO unit to the polymer backbone as a pendant side chain (Fig. 3) [42] not only increases the NLO density but also prevents the phase separation observed in guest – host systems. This is because one end of the moiety is permanently attached to the polymer backbone. Furthermore, the side-chain architecture gives sufficient mobility to respond to electrical poling, while providing enough poling stability for commercial applications.

Extensive research has been carried out on these systems. For example, materials such as 4-nitroaniline, disperse red, and dimethylamino-4-nitrostilbene have all been covalently attached to backbones such as polymethacrylate, poly(methyl methacrylate), polystyrene, polysiloxanes, and polycarbonates. The resulting materials show good SHG (frequency doubling) efficiencies and electro-optic coefficients comparable with or exceeding those of lithium niobate, as well as good poling stability. In one experiment, polymers were made with a predicted

80 % retention of the poling over a period of 5 years at 70 °C [40].

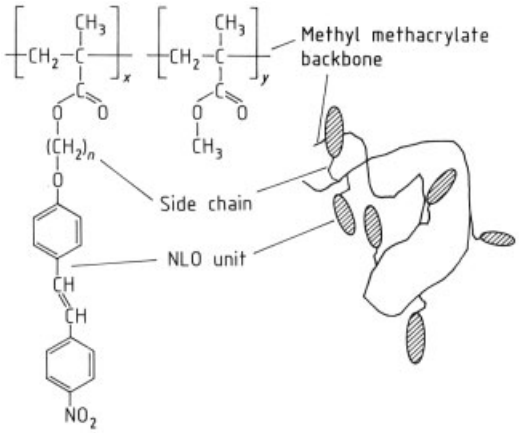

Figure 3. Methyl methacrylate polymer with NLO units attached to the polymer backbone via side chains

3.4. Examples of SHG Materials

The choice of nonlinear material for SHG depends on the application and the performance required. The key requirements in selecting a nonlinear crystal are a large phase-matching transparency range, high conversion efficiency, good optical homogeneity, good environmental stability, and ease of production. Organics generally have very high optical nonlinearities and high damage thresholds, but a transparency cut-off near 450 nm. Potassium titanyl phosphate does not suffer from a transparency cut-off but is very difficult to grow with sufficient optical quality. β-Barium borate and L-arginine phosphate are easier and cheaper to grow, have high transparency and wide phase-matching ranges; the latter has an extremely high optical damage threshold. Concerted efforts in improving crystal growth will make them attractive for high-power applications such as laser fusion, and low-power applications such as integrated optical devices and frequency conversion of miniature diode-pumped lasers. Table 1 shows SHG coefficients of some inorganic and organic materials.

3.5. Examples of Linear Electro-Optical Materials

As described in Section 2.1.2, the electro-optic effect arises from an anisotropic change in the refractive index of the material upon application of a d.c. electric field, quantified by the electro-optic tensor r_{ij}. If a linear polarized beam is incident on such a material, its two resolved orthogonal components will travel at different velocities through the crystal as a result of the induced birefringence. Upon emerging from the crystal a phase difference exists between these two components, the amount of retardation depending on the crystal, the crystal symmetry, dimensions, and the applied voltage. If the crystal is placed between crossed polarizers, the change in phase between the two polarizations causes a change in the transmittance of the beam.

Table 1. Second harmonic generation (SHG) coefficients of some inorganic and organic materials

Material	Wavelength transparency range, nm	d_{eff} [a], pm/V	Figure of merit [b]
Inorganic materials			
Potassium dihydrogen phosphate (KH_2PO_4)	200 – 1500	0.6	0.11
Deuterated potassium phosphate (KD_2PO_4)	200 – 1500	0.4	0.05
Potassium titanyl phosphate ($KTiOPO_4$)	350 – 4000	7.4	9.40
β-Barium borate (β-BaB_2O_4)	190 – 3500	1.6	0.60
L-Arginine phosphate	220 – 1950	1.1	0.30
Organic materials			
2-(*N,N*-dimethylamino)-5-nitroacetanilide	430 – 2000	27	148
N-(4-Nitrophenyl)-(*S*)-prolinol	500 – 2000	102	2000
3-Methyl-4-nitropyridine	414 – 2000	8	13
2-Methyl-4-nitroaniline	400 – 2000	31	170
		$d_{11} = 250$ [c]	

[a] $_{\mathrm{eff}}$ is the effective second-order nonlinear coefficient for a fundamental wavelength of 1.064 µm.
[b] Figure of merit is defined as d^2_{eff}/n^3 in units of $(\mathrm{pm/V})^2$ and is a measure of the second harmonic generation efficiency; n = refractive index.
[c] The d_{11} coefficient of 2-methyl-4-nitroaniline is not phase-matchable in bulk crystals.

Table 2. Linear electro-optic coefficients of some organic and inorganic materials

Material	Symmetry of crystal structure	$n^3 r$ [a]	Mode of operation	Reduced V_π (633 nm) [b], kV
Inorganic materials				
Lithium niobate ($LiNbO_3$)	3m	$1/2\,(n^3{}_e r_{33} - n_0^3 r_{13}) = 112$	transverse	2.94
Deuterated potassium phosphate (KD∗P)	42m	$n_0^3\, r_{63} = 90$	longitudinal	3.50
SBN : 60 [c]	4mm	$1/2\,(n^3{}_e r_{33} - n_0^3 r_{13}) = 1300$	transverse	0.25
β-Barium borate (β-BaB_2O_4)	R3c	$n_0^3\, r_{yyy} = 11.6$	transverse	56.0
Organic materials				
3-Nitroaniline	mm2	$1/2\,(n^3{}_z r_{33} - n^3{}_y r_{23}) = 63$	transverse	5.0
2-Methyl-4-nitroaniline	Cc	$1/2\,(n_1^3\, r_{11} - n_3^3 r_{31}) = 270$	transverse	1.3
DCNP [d]	Cc	$1/2\,(n^3{}_z r_{333}) = 860$	transverse	0.37

[a] n = refractive index; r = linear electro-optic coefficient. [b] The reduced V_π is calculated when the ratio of the thickness to the length of the crystal is normalized to 1. [c] Strontium barium niobate $Sr_{0.6}Ba_{0.4}Nb_2O_6$ [43].
[d] 3-(1,1-Dicyanoethenyl)-1-phenyl-4,5-dihydro-1*H*-pyrazole [44].

The efficiency of an electro-optic device can be defined in terms of the voltage required to induce a phase retardation of π radians and hence a switch from full extinction to full transmission. This is referred to as the half-wave voltage V_π. Such an electro-optic modulator has two modes of operation: the transverse configuration (electric field applied perpendicular to the direction of light propagation), and the longitudinal configuration (electric field applied parallel to the direction of propagation). The half-wave voltage required depends on which configuration and hence which components of r_{ij} are sampled. Table 2 shows the values of electro-optic parameters for some inorganic and organic materials.

3.6. Thin Film Applications

Thin films for integrated optical waveguides have a number of advantages over bulk materials. The small cross section of the waveguide structure can produce very high optical intensities which, when combined over long interaction lengths, can give rise to efficient nonlinear processes even with low-power lasers. Their compatibility with optical communication systems makes this class of devices very useful.

Electro-optic devices are generally fabricated using titanium-diffused $LiNbO_3$. Recent studies include efficient frequency doubling and difference frequency generation in Ti : $LiNbO_3$ and proton exchanged $LiNbO_3$ waveguide structures.

Various methods can be used for depositing organic polymers onto substrates. Spin coating, dipping etc. can result in large area coverage at relatively low temperatures (150 °C). Techniques such as photolithography can then be used to define the waveguide region. Since the resulting films are centrosymmetric, they need to be electrically poled. Results obtained using this poling technique are encouraging, achieving electro-optic coefficients greater than those of $LiNbO_3$. The low dielectric constant of these films compared to $LiNbO_3$ also makes them attractive for high-speed applications. Table 3 shows a comparison of properties of electro-optic materials for thin film applications.

Table 3. Comparison of some properties of electro-optic materials for thin film applications

Material	ε [b]	n	r, pm/V	$n^3 r$, pm/V	Figure of merit [c]
P2ANS/MMA [a]	3.2	1.63	> 40	> 170	> 54
$LiNbO_3$	16.5	2.2	30	328	20
GaAs	6.5	3.4	1.6	60	9

[a] Copolymer of diaminonitrostilbene and methyl methacrylate in a 50 : 50 ratio [13].
[b] The value given is the effective dielectric constant for coplanar waveguide electrodes on the material's surface.
[c] Figure of merit is defined as $n^3 r/\varepsilon$.

4. Third-Order Nonlinear Optical Materials

A number of second-order NLO materials have been used in commercial products and others show promise. No developed products based on third-order NLO effects exist apart from some liquid crystal display devices based on the d.c. Kerr effect (see Section 2.1.3). Reasons for this include the lack of well-defined systems applications and the high cost of powerful lasers, but the main difficulty is materials development.

4.1. Fundamental Requirements for $\chi^{(3)}$ Materials

The third-order NLO susceptibility, $\chi^{(3)}$ relates the third-order nonlinear polarization $\boldsymbol{P}^{(3)}$ and the applied electric field $\boldsymbol{E}$ through $\boldsymbol{P}^{(3)} = \chi^{(3)} \boldsymbol{E}\boldsymbol{E}\boldsymbol{E}$, or in indicial form (Chap. 2)

$$\boldsymbol{P}_i^{(3)} = \chi_{ijkl}^{(3)} \boldsymbol{E}_j \boldsymbol{E}_k \boldsymbol{E}_l \tag{17}$$

For many device applications an intensity-dependent refractive index is desired, often expressed as the nonlinear refractive index n_2, where

$$n = n_0 + \frac{1}{2} n_2 c \left(\varepsilon\varepsilon_0\right)^{1/2} \langle \boldsymbol{E}\boldsymbol{E} \rangle \tag{18}$$

with n_0 the linear refractive index, ε the dielectric constant of the material, ε_0 the permittivity of free space, c the speed of light, and $\boldsymbol{E}$ the monochromatic applied field. The relationship between n_2 and $\chi^{(3)}{}_{xxxx}$ is given by

$$n_2 = \frac{1}{c} \left(\frac{1}{n_0 \varepsilon_0} \right) \chi_{xxxx}^{(3)} \tag{19}$$

where $\chi^{(3)}{}_{xxxx}$ is the completely diagonal component of $\chi^{(3)}$ with all fields and polarization taken to be in the x-direction.

The other quantity often quoted is the *microscopic third-order NLO susceptibility*, γ. There is some confusion in the literature concerning the definition of γ. In Section 2.2, γ was defined in terms of the microscopic field and represented only electronic NLO processes. It is reasonable to extend this viewpoint so that $\gamma = \gamma_{\mathrm{el}} + \gamma_{\mathrm{lat}}$, where γ_{el} is the electronic contribution of γ and γ_{lat} is that due to lattice motions. However, γ is often defined as the microscopic average of $\chi^{(3)}$ (i.e., $\gamma = \chi^{(3)}/NF$, where N is the number of microscopic elements per unit volume and F a local field factor).

Another crucial requirement for $\chi^{(3)}$ materials is *optical transparency*. In linear optics the attenuation of light is governed by Beer's law $I(z) = I_0\, e^{-\alpha z}$, where $I(z)$ is the intensity at position z along the direction of propagation, and α the frequency-dependent absorption coefficient. Since many NLO effects and devices depend on a high power of I, absorption artifacts are amplified. If the relevant device variable is the phase shift $\Delta\Phi$ and the maximum length L is limited by Beer's law to be of the order of $1/\alpha$, then

$$\Delta F_{\max} = \frac{2\pi n_2 \bar{I}}{\lambda\alpha} \tag{20}$$

with $\bar{I}$ a mean intensity that takes account of attenuation. The figure of merit is $W = n_2/\lambda\,\alpha$, which should be maximized (not just n_2). A given material has a maximum in W at a wavelength λ_{opt} owing to the wavelength dependence of n_2 and α.

Two other important qualities are *fabricability* and *stability*. For single-mode optical interactions, as in silica optical fibers, materials must be amenable to channel waveguide fabrication. Silica optical fiber is an example of an optimized linear optical signal delivery system, for which an advanced waveguide technology has been developed (→ Fiber Optics, Chap. 3.5.). Similar developments are needed to make NLO materials viable for signal processing unless focal plane switching and free space interconnection approaches develop rapidly. Techniques that have been investigated include evaporation, metal oxide chemical vapor deposition, molecular beam epitaxy, ion sputtering, reactive ion etching, spin coating, Langmuir – Blodgett casting, laser ablation, and photobleaching. The stability of materials for $\chi^{(3)}$ applications is of concern, since these materials are subjected to high light intensities. For polymers and organic crystals, stability to oxidation and high light intensity is uncertain especially under the demanding conditions that devices can experience (high temperature, high pressure, etc.).

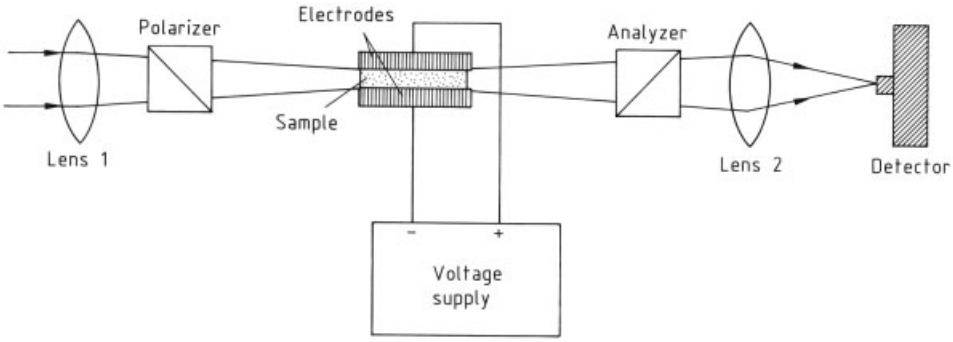

Figure 4. Apparatus for performing d.c. Kerr effect measurements on materials
The applied voltage results in a small rotation of the polarization vector of the light and thus in transmission through the analyzer.

4.2. Measurement of $\chi^{(3)}$ Properties

Studies of $\chi^{(3)}$ materials are of two types, frequency domain (dispersion) and time domain (dynamics). $\chi^{(3)}$ spectroscopy is used to study states of a system that are inaccessible to linear optical spectroscopy, owing to quantum electrodynamical selection rules. Studies on organic molecules have been performed with $\chi^{(3)}$ spectroscopy to investigate the symmetries of excited states [45]. Practically, the dispersion of $\chi^{(3)}$ coupled with the dispersion of $\chi^{(1)}$ determine the most useful spectral region of a given material.

Dynamic studies of $\chi^{(3)}$ materials have become prevalent since the advent of subpicosecond and femtosecond laser sources. Measurements of the speed of $\chi^{(3)}$ effects are important because they help to earmark materials for appropriate applications and devices. Third-order materials with large subpicosecond responses are sought for all-optical computing.

Since the d.c. Kerr effect does not rely on the presence of laser light, it was the earliest method for measuring third-order NLO properties [17]. The d.c. Kerr effect measures the off-diagonal elements $\chi^{(3)}_{1221}$, $\chi^{(3)}_{1212}$, and $\chi^{(3)}_{1122}$ that produce polarizations which are orthogonal to one or more of the applied fields. The experimental setup requires a high-voltage source, two polarizers, and a light source (Fig. 4).

Third harmonic generation (THG) is a direct, single-beam method for measuring the purely electronic part of $\chi^{(3)}$, usually $\chi^{(3)}_{1111}$ $(-3\,\omega;\omega,\omega,\omega)$ or equivalently the microscopic third-order nonlinear susceptibility, γ. This makes it useful for fundamental studies of materials, and THG measurements do not require state-of-the-art lasers or multibeam alignments. The Maker fringe technique is used, which determines the two unknowns, $\chi^{(3)}_{1111}$ and the coherence length. Difficulties include the requirement for a large transparency window in the material, and connection between THG measurements and n_2.

The optical Kerr effect (OKE) has been used to measure n_2 of materials [46], and the same tensor components can be measured as in the d.c. Kerr effect. Pump – probe techniques are used which are useful for time domain studies. In orientable media most of the OKE response comes from induced birefringence resulting from the linear polarizability anisotropy of the molecules or atoms in the medium; only a small portion of the response comes from γ_{el}, except for subpicosecond pulses.

Four-wave mixing, nonlinear interferometry, self focusing, and saturable absorption have all been used to measure $\chi^{(3)}$ or n_2. Degenerate four-wave mixing is used in both phase-conjugate and forward scattering geometries and is popular because it is self-phase-matching and signals are large due to resonant enhancement in most materials. This technique also lends itself readily to time domain studies. Because of the difficulty of analyzing the data, self focusing has seldom been used to measure n_2, but improvements in laser beam shapes have led to some research in this area [47]. Nonlinear interferometry has been used to measure third-order NLO properties [48]; elimination of the effects of artifacts (e.g., sample heating, air currents, and

vibrations) is difficult. Through the Kramers–Kronig relations, saturable absorption measurements can be related to resonant n_2 or $\chi^{(3)}$. Typically, resonant $\chi^{(3)}$ values estimated in this way can be six to ten orders of magnitude larger than their nonresonant counterparts.

4.3. Inorganic Materials

Glasses and Ionic Crystals. Some of the earliest NLO measurements were performed on glasses to find materials with low n_2 values that would not self focus in a laser cavity. The microscopic motion responsible for n_2 in glasses and in the majority of ionic crystals (e.g., LiF and NaCl) is mainly electronic and ultrafast. Furthermore, most of these materials have high transparency, making them suitable for high-power optical systems. A waveguide technology exists (optical fiber) for fused silica that makes it a viable material for optical device demonstrations.

Table 4 shows n_2 and n_0 values for some glasses and ionic crystals [48,49]. There is a rough correlation between the values of n_2 and n_0, indicating a monotonic relationship between $\chi^{(3)}$ and $\chi^{(1)}$. Although fused silica has a small value of n_2, its absorption in the near infrared (1.3 µm) is so small ($\alpha \approx 10^{-8}$ m) that its W parameter (Section 4.1) is large.

Research continues to identify materials with low n_2 values for ultrahigh-power applications like laser fusion. A low n_2 is important for minimizing self focusing. Some of the initial triumphs in all-optical processing were made in optical fibers since they constitute an ideal optical waveguide capable of supporting high peak power pulses over long distances. An example of an all-optical switch in silica optical fiber is a directional coupler in which switching between two cores of a dual core fiber occurs < 500 fs [20].

Table 4. Values of n_2 for glasses and inorganic crystals at 1.06 µm

Material ∗	n_0	n_2, 10^{-19} m^2/W	Method of measurement ∗∗
Glass (SF-6)	1.77	2.2	NI
Silicate glass (FR-5)	1.686	0.52	NI
Borosilicate glass (BK-7)	1.517	3.42	NI
PbF_2	1.769	17	NI
KH_2PO_4	1.468	0.29	NI
Fused silica	1.458	0.27	NI
NaCl	1.5312	0.43	FWM
KBr	1.5435	0.80	FWM
Fused silica	1.4496	0.25	FWM
SiO_2 (quartz(o))	1.5342	0.31	FWM
ZnO(o)	1.99	5.30	FWM
TiO_2	2.48	9.5	FWM
$KTiOPO_4$	1.74	1.38	FWM
ZnS(o)	2.29	8.65	FWM
CdS(o)	2.33	54.6	FWM

∗ (o) denotes the ordinary optical axis of the material.
∗∗ NI = nonlinear interferometry; FWM = four-wave mixing.

Semiconductors. Semiconductors attracted intense interest as all-optical materials after optical bistability and differential optical gain were observed in thin etalons of GaAs [50]. (An etalon is formed by placing two partially reflecting surfaces parallel to one another to form a Fabry–Perot interferometer). Refractive indices of semiconductors (n_0 and n_2) are listed in Table 5. Gallium arsenide has many advantages for optical processing because it is simultaneously an electronic material, an electro-optic material, a lasing medium, and an all-optical material. However, GaAs has drawbacks such as instability and a large absorption coefficient when n_2 is large. Some of these advantages and diffi-

Table 5. Values of n_0 and n_2 for semiconductors

Material	λ, µm	n_0	$\lvert n_2 \rvert$, 10^{-17} m^2/W
Nonresonant			
GaAs	1.06	3.2	3.0
AlGaAs	0.81	3.0	40
InAs	10.6	ca. 3.0	5.0
Ge	10.6	ca. 3.0	0.5
CdS (glass filter)	0.69	2.33	0.02
CS 3-68 (glass filter, Corning)	0.7	ca. 2.5	0.05
CS 3-67 (glass filter, Corning)	0.7	ca. 2.5	3.0
InSb	9.24	3.75	90.0
Si	1.3–7	3.44	1.0
Resonant			
AlGaAs–GaAs multiple quantum well	0.84	3.0	2×10^9
RG66S (Schott glass $CdS_{0.3}Se_{0.7}$)	0.63	2.5	1.67
CdS_xSe_{1-x}	0.50	2.5	1×10^5
InP–InGaAsP DREAM ∗	1.06	3.0	ca. 2×10^{12}
CuCl	0.4	2.5	7.5×10^4
GaAs (at 10 K)	0.82	3.0	4×10^9
InAs	3.1		3×10^8
InSb	10.6		2×10^6

∗ DREAM = depletion region electroabsorption modulator

culties also apply to multiquantum well (MQW) GaAs – AlGaAs structures (see Section 3.1).

The n_2 values of nonresonant semiconductors are not unusually large. However, when semiconductors are probed near the band gap (the energy difference between the conduction band and the valence band), huge enhancements in $\chi^{(3)}$ are achieved. Mechanisms leading to these enhancements include phase – space filling (an exclusion principle effect), the Burstein – Moss shift, and excitonic nonlinearities. An exciton is a hydrogen-atom-like entity consisting of an electron in the conduction band bound to a "hole" in the valence band of a semiconductor. Excitonic nonlinearities are not present at room temperature in bulk semiconductors since the thermal agitation breaks up the excitons. In MQWs the confinement in the direction perpendicular to the plane of the MQW stabilizes the excitons, so that excitonic nonlinearities are possible at room temperature.

Several other semiconductors besides GaAs have demonstrated large NLO effects. The value of n_2 for InSb is very large in the infrared (10.6 μm), and has been used to make a multistable device (a device that has more than one stable output state for a given input state) [51]. Increasing interest has been devoted to semiconductor microcrystallites embedded in glass (e.g., CS 3-68 and CS 3-67, Table 5), since these systems are good approximations to quantum dots [52]. Quantum dots are three-dimensional confined structures and can be treated theoretically with the particle-in-a-box model of quantum mechanics.

Semiconductors are now being engineered with confinement effects and doping (as in GaAs – AlGaAs *n-i-p-i* structures [53]) to optimize nonlinearity with respect to absorption and thermal effects. The electronic capabilities of semiconductors make hybrid systems like self electro-optic devices (SEEDs) possible.

4.4. Organic Materials

In the 1980s research into the NLO properties of organic media grew rapidly due to the large, ultrafast NLO effects observed in organic polymers, the ability to modify molecules and enhance their nonlinearities through chemical synthesis, and the compatibility between waveguide technology and organic thin film formation techniques. Organics are promising materials for $\chi^{(3)}$ applications, but electronic low-intensity bistability and optical switching have to be observed in organic media. Generally, organic media are desired that are readily synthesized and fabricated, and have an ultrafast response time. Values of n_2 for various organic substances are listed in Table 6 (see next page). Improvements in lasers and chemical synthesis have brought all-optical processing devices based on organic materials closer to realization.

Origin of Nonlinearity. The $\chi^{(3)}$ effects in organic molecules are primarily attributable to the large delocalization of π-electrons in conjugated molecules. Delocalized electrons can produce nonlinear effects because they constitute an interacting system. For example, anharmonic motion in one electron induces anharmonic motion in other conjugated electrons. The theoretically predicted superlinear dependence of $\chi^{(3)}$ on conjugation length for a one-dimensional conjugated chain has been observed. Since the electrons in the π-orbitals are delocalized over the entire conjugation length, the electron cloud can be readily distorted and at high speeds. Nonlinear optical effects in organic molecules do not involve real excitation of the molecule into one of its allowed states, but rather "virtual" excitations that decay in femtoseconds.

Organic media exhibit large resonantly enhanced n_2 values when a combination of applied frequencies approaches an eigenfrequency. This is important because it allows appreciable NLO effects with modest laser input powers. However, the response time also increases and the tradeoff between enhancement of n_2 and an attendant increase in absorption (which may damage the material) is not as good as for semiconductors. This is because the cooperative effects associated with the band gap in semiconductors (e.g., exciton and biexciton formation and propagation, Coulomb screening, and phase space filling) are weak in most organic media. Resonant n_2 values for semiconductor MQWs (2×10^{-8} m^2/W) are three to four orders of magnitude larger than the largest observed resonant electronic n_2 value in an organic molecule (silicon naphthalocyanine, 10^{-11} m^2/W).

Table 6. Values of n_2 for organic materials

Material	λ, µm	n_2, 10^{-18} m^2/W	Method [a]
Carbon disulfide	1.06/0.53	5.2	OKE
Nitrobenzene	0.694	3.7	OKE
Benzene	1.06/0.53	1.7	OKE
PTS – PDA [b]	1.91/0.64	178	THG
Poly-TCDU – PDA [c]	1.91/0.64	45	THG
2-Methyl-4-nitroaniline	1.06/0.532	58	DCSHG
Poly-4BCMU – PDA [d]	0.63	10	DFWM
Polyacetylene	1.91/0.64	2000	THG
Poly(3-methyl-4′-octyl-2,2′-bithiophene-5,5′-diyl)	1.91/0.64	380	THG
Poly-2,5-thienylene vinylene	1.85/0.62	306	THG
Ferrocene	1.06	10	SF
BBL [e]	1.06	148	DFWM
Platinum phthalocyanine	1.06	1580	DFWM
Polysilane	1.9/0.64	18	THG
Fluoroaluminum phthalocyanine	1.06/0.36	439	THG
Azo dye 25/MMA 75 [f]	2.05/0.7	12.4	THG

[a] OKE = optical Kerr effect; THG = third harmonic generation; DCSHG = d.c. second harmonic generation; OFWM = degenerate four-wave mixing; SF = self focusing. [b] *p*-Toluenesulfonate – polydiacetylene. [c] Poly(5,7-decadiyne-1,12-diol-bisphenylurethane) with the side group – $(CH_2)_4OCONH\text{-}C_6H_5$ – polydiacetylene. [d] Poly(4-butoxycarbonylmethylurethane) – polydiacetylene. [e] Poly{(7-oxo-7,10*H*-benz[de]imidazo [4',5′:5,6] benzimidazo [2,1-a] iso-quinoline-3,4:10,11-tetrayl)-10- carbonyl}. [f] Copolymer of an azo dye with methyl methacrylate in a 25 : 75 ratio.

Small Molecules. Liquids consisting of small molecules (e.g., carbon disulfide, benzene, and nitrobenzene) have fairly large n_2 values. Many NLO effects were first observed in these systems and they continue to be of interest for studying new phenomena and as standards. Carbon disulfide is one of the few materials whose absolute n_2 value (Table 6) is known with $> 30\,\%$ accuracy. The optical Kerr effect is used to study the structure and dynamics of organic liquids at short time scales (femtosecond pulses). With pulses longer than a few tens of picoseconds, the n_2 of organic liquids is mainly due to nuclear orientation and only a small part is attributable to electronic motion since the conjugated bond lengths of these materials are short.

Large Molecules and Polymers. The first large organic molecules to be studied for NLO effects were phthalocyanines [54] and biological molecules such as *trans*-retinal [55]. Phthalocyanines are good saturable absorbers and are used to Q-switch ruby lasers; however, in their pure form they are not very soluble in any solvent. This is a frequent problem with highly conjugated molecules, and progress in organic NLO depends on increasing the solubility of the chromophoric units.

Interest has centered on polyacetylenes [*25067-58-7*] and polydiacetylenes because of their large n_2 values in the near infrared (Table 6), which make them suitable for optical fiber applications. Single crystals of the polymer made from *p*-toluenesulfonate and 2,4-hexadiyne-1,6-diol (PTS – polydiacetylene [*32535-60-7*]) have been studied by third harmonic generation [56], and techniques have been developed for forming Langmuir – Blodgett films from polydiacetylenes. Some soluble polydiacetylenes are now available, and all-optical phase modulation has been demonstrated in a single crystal of PTS – polydiacetylene [57].

Polyacetylene

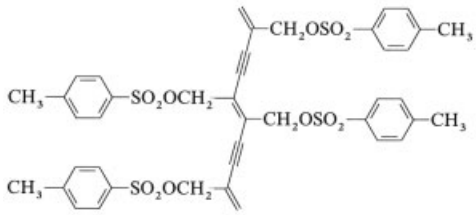

PTS – polydiacetylene

The values of n_2 have been measured in a variety of conjugated structures, including phthalocyaninines [58], electro-optic chromophores [59], polysilanes, polygermanes [60], polythiophenes [61], ladder polymers [62], and block copolymers [63] (Table 6). Most of these materials are at an earlier devel-

opment stage than the polydiacetylenes, and have not generally been incorporated in devices. However, electronic optical bistability in a Fabry – Perot cavity containing a thin film of silicon naphthalocyanine oligomers (resonant $n_2 = 10^{-11}$ m^2/W) has been observed [64].

5. Nonlinear Optical Devices and Applications

5.1. Electro-Optic Devices

The linear electro-optic (or Pockels) effect in which the refractive index changes linearly with an applied field originates from a second-order NLO property. It is the simplest and one of the most important NLO effects. The basic physics of electro-optics and its relationships to NLO are described in Chapter 2 (for further details see [65, 66]). Various electro-optic devices utilize the change in refractive index of a dielectric medium with application of an external field as discussed in Chapter 3. The index change can be used to modulate both phase and amplitude of optical waves passing through the medium. An electro-optic modulator is designed to obtain a required depth of modulation with the lowest possible modulation power over the largest possible modulation bandwidth.

A variety of electro-optic devices based on waveguide and through-the-plane designs have been developed. Technical requirements for two representative devices, a waveguide electro-optic interferometric modulator and an electro-optic spatial light modulator (SLM), are described below.

Waveguide Electro-Optic Modulators. Depending on its geometry, a waveguide can allow a beam to travel at a finite number of angles relative to its axis. Most electro-optic devices are based on a single-mode waveguide architecture in which there is only one angle for beam propagation. The ever-increasing need for faster data communication systems necessitates lower frequencies for optical-to-electrical signal conversions and development of optical switching technology to replace traditional electronic switching.

Moreover, telecommunications networks are based on single-mode fibers and any network components must closely match the optical characteristics of these fibers. It is easier to control guided optical waves by electro-optic effects in the single-mode waveguide. Also, because the waveguides have small diameters (a few micrometers) low modulation voltages can generate a large modulating electrical field. This, in turn, generates larger electro-optic effects, decreases the device length, and thus lowers the capacitance and increases the modulation speed. It also allows integration of the devices making them commercially more attractive. The theory of optical waveguides is described in [67, 68]. Fabrication methods of inorganic waveguides are described in [69], see also → Fiber Optics, Chap. 3.

A large number of waveguide electro-optic devices have been developed and are described in [69]. They include the electro-optic directional coupler switch [70], polarization controller [71], frequency shifter [72], polarization multiplexer and demultiplexer [73], and tunable wavelength filters [74].

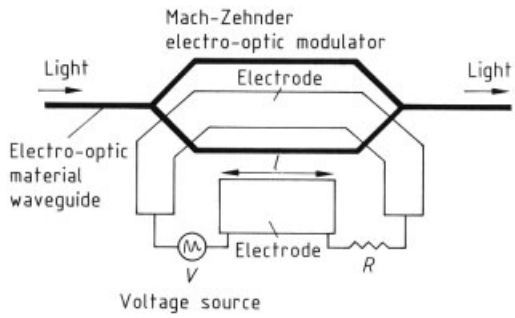

Figure 5. Schematic diagram of a Mach – Zehnder type electro-optic modulator, where l is the effective length over which switching can occur

The Mach–Zehnder interferometric modulator best illustrates the design principles and material property requirements of electro-optic devices (Fig. 5). The incoming optical beam is divided in two at a Y-junction and then recombined at a second Y-junction. If the phase shift of the divided beams in both arms is identical, they interfere constructively and all the power reappears at the output. If the phase-shift difference between the two arms is π, the modes interfere destructively and power radiates out into the substrate and is lost. Thus, the optical output power can be controlled by an electrical field applied to the modulator via the electro-optic properties of the

waveguide material. Because the electrode spacing can be made very small (ca. 5 µm), relatively small switching powers are needed. The device capacitance is small and the modulator can be operated at very high frequencies (> 1 GHz).

For a $LiNbO_3$ Mach – Zehnder modulator cut perpendicular to the x-axis of the crystal, the switching voltage V_π is related to the operating wavelength λ, the electrode gap w, the electrode length l, and the material properties such as the Pockels constant r_{33} and refractive index n_e through

$$V_\pi = \lambda \frac{w}{n_e^3 r_{33} l G} \tag{21}$$

Here Γ is the overlap constant between the optical beam and the modulating electrical field and is typically about 0.4 for transverse electrodes.

In high-frequency light modulation, the transit time τ of light through the modulator must be taken into account. If the transit time is not negligible compared to the modulation period, the modulation efficiency is reduced due to mismatch in phase velocity between the modulating microwave signal and the optical wave. The transit time limitation is usually overcome by using a traveling wave electrode design where the switching voltage $V_\pi(f)$ or the bandwidth f is determined by the magnitude of the mismatch. The switching voltage $V_\pi(f)$ of a traveling wave modulator with a length l is increased at high frequencies by a factor

$$\frac{V_\pi(f)}{V_\pi} = \frac{\pi T_\omega f}{\sin(\pi T_\omega f)} \tag{22}$$

where $T_\omega = (l/c)(\sqrt{\varepsilon} - n)$ is the transit time mismatch between the copropagating microwave and optical signals, ε is the dielectric constant of the electro-optical medium, and c is the speed of light.

For efficient high-speed modulator applications (i.e., a low modulation power) electro-optic materials must therefore possess a high Pockels constant and a low dielectric constant. Moreover, the square root of the dielectric constant must be as close as possible to the refractive index to produce a large bandwidth. Compared to traditional inorganic materials such as $LiNbO_3$, newly developed electro-optic polymers (Section 3.3) offer a significant advantage in this respect. For example, $|\sqrt{\varepsilon} - n|$ for $LiNbO_3$ is typically about 3.4 and that for polymers is about 0.17, indicating that the potential bandwidth for polymer modulators can be almost 20 times larger than that of $LiNbO_3$ modulators.

A number of commercial $LiNbO_3$ modulators are available with a switching voltage of ca. 5 V and a bandwidth of ca. 4 GHz. Many companies are developing polymer electro-optic modulators; switching voltages as low as 12 V and bandwidths as high as 1 GHz have been reported [75].

Spatial light modulators (SLM) are devices that modify the phase, polarization, amplitude, or intensity of a one- or two-dimensional readout light distribution as a function of space and time [76]. They exploit the inherent parallelism of light and are a key component of optical information processing systems. The modulators vary according to operational principles, information processing functions, and light modulating materials. Materials include membranes, magneto-optic materials, liquid crystals, and electro-optic crystals.

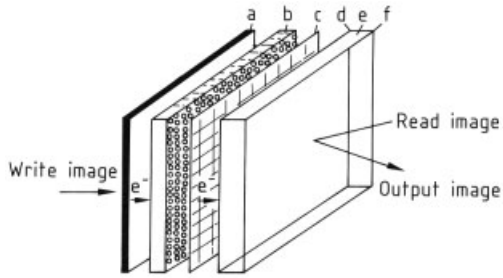

Figure 6. Optically addressed microchannel spatial phase modulator (MSLM)
a) Photocathode; b) Microchannel plate; c) Grid; d) Dielectric mirror; e) Electro-optic modulating material ($LiNbO_3$); f) Transparent electrode

The optically addressed microchannel SLM (MSLM) utilizes a design in which the applied electric field is parallel to the direction of light propagation in $LiNbO_3$ (Fig. 6) [77]. The magnitude of this longitudinal electro-optic effect depends solely on the applied electrical voltage; a large voltage drop (ca. 2 kV) is needed to achieve sufficiently large effects for device operation. A unique feature of the MSLM is the microchannel plate (b) which provides a large electron gain. It contains a number of fine capillaries through which electrons pass. The latter can be accelerated by adjusting the potential at the grid (c). The electrons are then deposited onto the dielectric

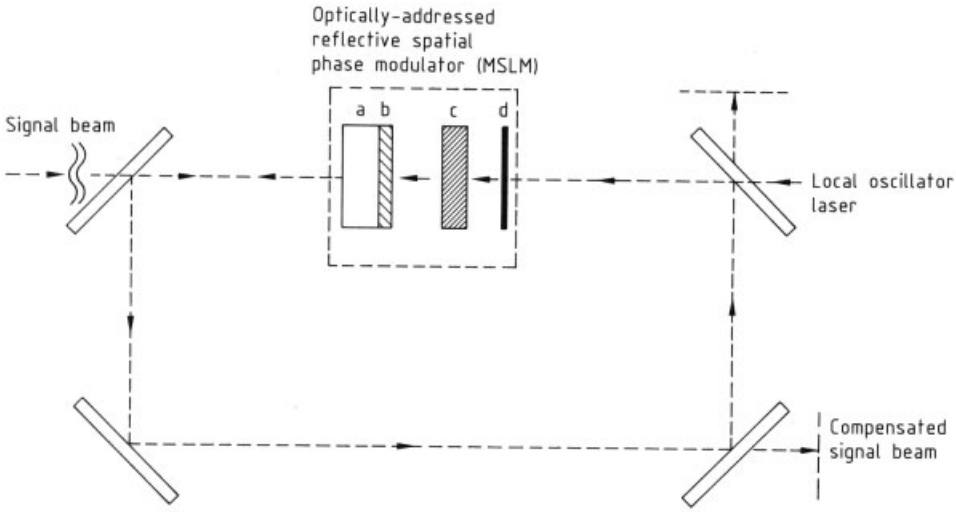

Figure 7. Interference phase loop using a microchannel spatial light modulator
a) Electro-optic crystal; b) Dielectric mirror; c) Microchannel plate; d) Photocathode

mirror (d) and generate a two-dimensional electrical field across the modulating material (e). This creates a two-dimensionally distributed optical path in the modulating material which can be read with an optical beam. Figure 7 shows an application of MSLM called an interference phase loop in which the signal beam interferes with itself through a feedback to sharpen its image [78]. The local oscillator laser creates a fixed phase difference in the feedback so that the system can continuously estimate, compensate and conjugate the wave-front phase over π radians. In operation, a signal image is converted into a spatial electron image by the photocathode. The electron image is amplified by the microchannel plate and deposited on the surface of the $LiNbO_3$ crystal. The charge distribution on the surface of the crystal modulates the readout beam via the longitudinal electro-optic effect. The grid between the microchannel plate and the crystal is used as a secondary electron source to erase the image frame.

5.2. Frequency Doubling and Parametric Conversion

Wave interaction of two input beams in a second-order nonlinear medium leads to generation of a third output beam (three-wave mixing). When an optical wave interacts with itself, its frequency can be doubled (second harmonic generation). If the interacting waves have different wavelengths, sum- and difference-frequency generation can result, which are termed parametric processes (Chaps. 2 and 3). With advances in tunable lasers, frequency doubling and parametric conversion processes have become widely used NLO effects for extending the tunable range of coherent sources to UV and IR wavelengths. The source beams generally have a high intensity and a high power conversion can be achieved if the phase-matching condition is met.

Frequency doubling of low-power diode lasers is drawing intense interest. Since data density in optical recording is diffraction-limited, a compact short-wavelength source would allow higher recording density. A 10-mW source is usually required for writing and a 2-mW source for reading in commercial optical data storage systems. To maintain a relatively high optical power density and thus to improve conversion efficiency, the diode laser beam can be confined in a waveguide and phase matching obtained through a number of methods including periodic modulation of the $\chi^{(2)}$ [12, 13].

5.3. Third-Order Nonlinear Optical Devices

Third-order NLO effects are relatively small. Although phenomena such as third harmonic gen-

eration, the optical Kerr effect, self focusing, stimulated Raman scattering [79], and optical phase conjugation [80] have been known since the early stages of NLO, practical devices are still in concept stages and require significant improvement in materials.

Nonlinear Waveguide Coupler. A manifestation of third-order NLO effects is the light-intensity-dependent refractive index n_2 defined in Section 2.1.3. This phenomenon has been used in a nonlinear waveguide directional coupler based on GaAs [81], which has a fairly large nonresonant n_2 (Table 5).

In waveguide couplers, two identical waveguides are in proximity (Fig. 8). The optical beam launched in one guide is transferred to the other guide at an interaction distance characteristic of the device. If the refractive index of the waveguide is intensity-dependent, then the coupler is nonlinear. If the optical beam intensity in one guide is large, the refractive index is changed to such an extent that the mode pattern changes and transfer becomes incomplete or does not occur at all. With GaAs waveguides of $10\,\mu m^2$ cross section and 1.7 mm coupling length, a critical power (the input optical power at which the output power is equally divided between the two guides) of $500\,\mu W$ was demonstrated for $\lambda = 0.85\,\mu m$. A detailed analysis of waveguide NLO device concepts is given in [82].

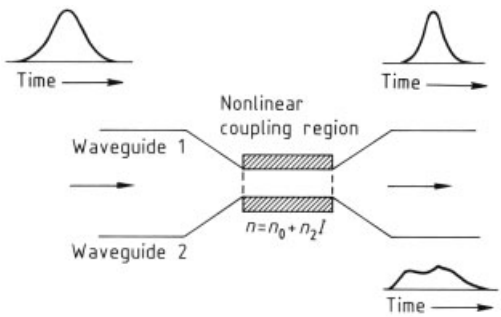

Figure 8. Two coupled channel waveguides with a nonlinear material in the coupling region, this structure can operate as an all-optical switch

Optical Phase Conjugation. In optical phase conjugation a NLO effect is used to reverse the direction of propagation of a light beam and its overall phase factor. Optical phase conjugation is produced by two standard methods; stimulated Brillouin scattering and four-wave mixing in a NLO medium [83]. When a light source is reflected from a phase-conjugate mirror, the reflected wave carries all the information of the original beam, but retraces its original path regardless of its incidence angle. A general phase-conjugate medium can conjugate a beam with an arbitrary wavefront back to its original path, point by point, irrespective of its incident angle. A phase-conjugate wave is also termed "time-reversed" because it runs along the same trajectory as the original beam but in the opposite direction.

An application of optical phase conjugation is to "anti-distort" a distorted wave in a so-called double pass geometry. A high-quality optical beam can, for instance, be transmitted through a turbulent atmosphere and be made to retrace its own path after reflection from a phase-conjugate mirror. When the beam returns to its point of origin it is therefore free of distortion. A variation of this is removal of distortions in fiber optic cables. In an optical fiber, the image travels in many optical modes. These modes travel different paths, and are out of step by the time they reach the end of the cable. The image can be unscrambled if it is phase conjugated and sent back through the same fiber or another identical link.

In self targeting of radiation, a broad beam from a relatively low-power source is directed in the vicinity of the object that is to be targeted. The object scatters the radiation in all directions, some of it falls on an amplifier placed near the object. As the radiation travels through the amplifier its power is magnified. A phase conjugator at the end of the amplifier creates an "anti-distorted" beam, reflects it back through the amplifier, and a powerful directed beam is thus focused on the object.

Optical phase-conjugate beams can also be used in pointing and tracking of moving objects, image processing, optical computing, interferometry, laser gyroscopes, fiber and satellite communication systems, laser weapon systems, and noncontact photolithography.

5.4. Nonlinear Optics in Optical Fibers

Because silica is a symmetric molecule, silica glass does not normally exhibit second-order

NLO effects. The lowest-order nonlinear effects in silica optical fibers originate from the third-order susceptibility $\chi^{(3)}$. A prominent third-order effect is the intensity-dependent refractive index (Eq. 8). Although n_2 of silica glass is very low ($\approx 10^{-20}$ m^2/W), the glass fiber is an excellent third-order material because the NLO effects can occur over a large length due to its outstanding optical transparency.

Optical Solitons. The term "optical soliton" refers to a special optical wave that can propagate undistorted over long distances and remains unaffected after collision with other solitons.

To form solitons, the optical fiber must have an intensity-dependent refractive index n_2 and an anomalous group velocity dispersion (Section 2.1.3). A light pulse propagating in such an optical fiber is spectrally broadened due to the nonlinearity. However, because of the anomalous dispersion, the broadened pulse is sharpened back to its original shape. Thus, the interplay between n_2 and dispersion maintains the pulse shape over a long propagation distance [84].

An interesting application of solitons in fiber is the soliton laser. Here, an optical fiber cavity is coupled to a synchronously pumped, mode-locked dye laser. The pulse width of the laser can be precisely controlled by varying the fiber length [85]. Pulses as short as 50 fs have been produced by this technique.

The performance of high-speed fiber optic telecommunication systems is generally limited by pulse broadening due to dispersion. Since solitons maintain their pulse widths over long distances, soliton-based communication systems have been proposed. Transmission of 55 ps solitons over 400 km in optical fibers has been demonstrated [86].

Second Harmonic Generation. Silica optical fibers can generate a significantly large second harmonic of a 1.06 µm pump beam if the fiber is illuminated with the pump radiation for several hours [87]. The mechanism is not fully understood; it is generally believed that a periodic ordering of some entity such as color centers or defects along the fiber length occurs resulting in a phase-matching condition [84].

6. References

1. P. A. Franken, A. E. Hill, C. W. Peters, G. Weinreich, *Phys. Rev. Lett* **7** (1961) 118.
2. N. Bloembergen: *Nonlinear Optics,* Benjamin, New York 1965.
3. B. F. Levine, *Phys. Rev. Lett.* **22** (1969) 78.
4. G. R. Meredith, J. Van Dusen, D. J. Williams, *Macromolecules* **15** (1982) 1385.
5. D. S. Chemla, D. A. B. Miller, *J. Opt. Soc. Am. B Opt. Phys.* **2** (1985) 1155.
6. T. F. Heinz, C. K. Chen, D. Ricard, Y. R. Shen, *Phys. Rev. Lett.* **48** (1982) 478.
7. E. Meijer, E. E. Havinga, G. L. J. A. Rikken, *Phys. Rev. Lett.* **65** (1990) 37.
8. D. H. Auston, *Phys. Rev. Lett.* **53** (1984) 1555.
9. D. A. B. Miller, D. S. Chemla, D. J. Eilenberger, P. W. Smith, *Appl. Phys. Lett.* **41** (1982) 679.
10. D. A. B. Miller et al., *Appl. Phys. Lett.* **45** (1984) 13.
11. A. L. Lentine et al., *Appl. Phys. Lett.* **52** (1988) 1419.
12. E. J. Lim, M. M. Fejer, R. L. Byer, *Electr. Lett.* **25** (1989) 731.
13. G. Khanarian et al., *Appl. Phys. Lett.* **57** (1990) 977.
14. W. P. Risk et al., *J. Opt. Soc. Am.* **A 4** (1989) 128.
15. L. K. Cheng, W. R. Bosenberg, C. L. Tang, *Appl. Phys. Lett.* **53** (1988) 175.
16. J. H. Kimble, D. F. Walls, *J. Opt. Soc. Am. B Opt. Phys.* **4** (1988) 1449.
17. J. Kerr, *Philos. Mag.* **50** (1875) 337.
18. G. Mayer, F. Gires, *C. R. Acad. Sci. Ser. A* **258** (1964) 2039.
19. W. T. Lotshaw, D. McMorrow, C. Kalpouzos, G. A. Kenney-Wallace, *Chem. Phys. Lett.* **136** (1987) 323.
20. J. S. Aithchison et al., *Opt. Lett.* **15** (1990) 471.
21. L. F. Mollenauer, R. H. Stolen, J. P. Gordon, *Phys. Rev. Lett.* **45** (1980) 1095.
22. B. Nikolaus, D. Grischkowsky, *Appl. Phys. Lett.* **42** (1983) 1.
23. B. F. Levine, C. G. Bethea, *J. Chem. Phys.* **63** (1975) 2666.
24. J. Feinberg, *Physics Today* **41** (1988) 46.
25. S. J. Lalama, A. F. Garito, *Phys. Rev. A* **20** (1979) 1179.
26. J. R. Heflin, K. Y. Wong, O. Zamani-Khamiri, A. F. Garito, *Phys. Rev. B Condens. Matter* **38** (1988) 1573.
27. P. D. Maker, R. W. Terhune, M. Nisenoff, C. M. Savage, *Phys. Rev. Lett.* **8** (1962) 21.

28. G. D. Boyd et al., *Appl. Phys. Lett.* **5** (1964) 234.
29. C. Chuangtian, W. Bochang, J. Aidong, Y. Guiming, *Scientia Sinica (B)* **28** (1985) 235.
30. F. C. Zumsteg, J. D. Bierlein, T. E. Gier, *J. Appl. Phys.* **47** (1976) 4980.
31. D. A. B. Miller et al., *Phys. Rev. B Condens. Matter* **32** (1985) 1043.
32. S. K. Kurtz, T. T. Perry, *J. Appl. Phys.* **39** (1968) 3798.
33. G. F. Lipscomb, A. F. Garito, R. S. Narang, *Appl. Phys. Lett.* **38** (1981) 663.
34. P. A. Norman, P. V. Jones, R. J. Jones, S. R. Hall, *J. Opt. Soc. Am. B Opt. Phys.* **4** (1987) 1013.
35. J. Zyss, D. S. Chemla, J. F. Nicoud, *J. Chem. Phys.* **74** (1981) 4800.
36. R. Twieg, A. Azema, K. Jain, Y. Y. Cheng, *Chem. Phys. Lett.* **92** (1982) 208.
37. J. Zyss, J. F. Nicoud, M. Coquillay, *J. Chem. Phys.* **81** (1984) 4160.
38. I. R. Girling et al., *Optics Communications* **55** (1985) no. 4, 289.
39. K. D. Singer, J. E. Sohn, S. L. Lalama, *Appl Phys. Lett.* **49** (1986) 248.
40. H. T. Man et al., *Proc. SPIE Int. Soc. Opt. Eng.* **1213** (1990) 7.
41. M. Eich et al., *J. Appl. Phys.* **66** (1989) 3241.
42. R. Demartino, H. N. Yoon, US 4 808 332, 1989.
43. R. R. Neurgaonker, W. K. Cory, *J. Opt. Soc. Am. B Opt. Phys.* **3** (1986) 274.
44. S. Allen et al., *J. Appl. Phys.* **64** (1988) 2583.
45. R. M. Hochstrasser, H. N. Sung, J. E. Wessel, *J. Am. Chem. Soc.* **95** (1973) 8179. J. J. Song, M. D. Levenson, *J. Appl. Phys.* **48** (1977) 3496.
46. T. Y. Chang, *Opt. Eng.* **20** (1981) 220.
47. M. Sheik-bahae, A. A. Siad, E. W. Van Stryland, *Opt. Lett.* **14** (1989) 955.
48. D. Milam, M. J. Weber, A. J. Glass, *Appl. Phys. Lett.* **31** (1977) 822.
49. R. Adair, L. L. Chase, S. A. Payne, *Phys. Rev. B Condens. Matter* **39** (1989) 3337.
50. H. M. Gibbs et al., *Appl. Phys. Lett.* **35** (1979) 451.
51. D. A. B. Miller, S. D. Smith, A. Johnston, *Appl. Phys. Lett.* **35** (1979) 658.
52. J. Yumoto, S. Fukushima, K. Kubodera, *Opt. Lett.* **12** (1987) 832.
53. H. Ando, H. Iwamura, H. Oohashi, H. Kanbe, *IEEE J. Quantum Electron* **QE-25** (1989) 2135.
54. J. A. Armstrong, *J. Appl. Phys.* **36** (1965) 471.
55. J. P. Hermann, D. Ricard, J. Ducuing, *Appl. Phys. Lett.* **23** (1973) 178.
56. C. Sauteret et al., *Phys. Rev. Lett.* **36** (1976) 956.
57. M. Thakur, D. M. Krol, *Appl. Phys. Lett.* **56** (1990) 1213.
58. J. S. Shirk et al., *Appl. Phys. Lett.* **55** (1989) 1287.
59. C. W. Dirk, M. G. Kuzyk, *Chem. Mater.* **2** (1990) 4.
60. J. C. Baumert et al., *Appl. Phys. Lett.* **53** (1988) 1147.
61. B. P. Singh, M. Samoc, H. S. Nalwa, P. N. Prasad, *J. Chem. Phys.* **92** (1990) 2756.
62. J. R. Lindle et al., *Appl. Phys. Lett.* **56** (1990) 712.
63. S. A. Jenekhe, W.-C. Chen, S. Lo, S. R. Flom, *Appl. Phys. Lett.* **57** (1990) 126.
64. J. W. Wu et al., *J. Opt. Soc. Am. B Opt. Phys.* **6** (1989) 707.
65. A. Yariv: *Quantum Electronics,* 3rd ed., Wiley, New York 1989.
66. H. Haus: *Waves and Fields on Optoelectronics,* Prentice Hall, Englewood Cliffs, NJ, 1984.
67. T. Tamir (ed.): *Integrated Optics,* Springer Verlag, New York 1979.
68. L. D. Hutcheson (ed.): *Integrated Optical Circuits and Components,* Marcel Dekker, New York 1987.
69. H. Nishihara, M. Haruna, T. Suraha: *Optical Integrated Circuits,* McGraw-Hill, New York 1989.
70. M. Papuchon et al., *Appl. Phys. Lett.* **27** (1975) 289.
71. R. C. Alferness, L. L. Buhl, *Appl. Phys. Lett.* **38** (1981) 655.
72. F. Heisman, R. Ulrich, *Appl. Phys. Lett.* **45** (1984) 490.
73. O. Mikami, *Appl. Phys. Lett.* **36** (1980) 491.
74. R. C. Alferness, J. J. Veselka, *Electron. Lett.* **21** (1985) 466.
75. G. R. Möhlmann et al., *Proc. SPIE Int. Soc. Opt. Eng.* **1337** (1990) 215 – 224.
76. C. Warde, A. D. Fisher in J. L. Horner (ed.): *Optical Signal Processing,* Chap. 7.2, Academic Press, New York, 1987, pp. 477 – 523.
77. C. Warde, J. Thackara, *Opt. Eng.* **22** (1983) 695.
78. A. D. Fisher, C. D. Warde, *Opt. Lett.* **8** (1983) 533.
79. Y. R. Shen, N. Bloembergen, *Phys. Rev.* **137** (1965) A 1787.
80. A. Yariv, *IEEE J. Quantum Electron* **QE-14** (1978) 650.

81. P. Li Kam Wa et al., *Electron Lett.* **21** (1985) 26.
82. G. I. Stegeman, J. J. Burke, C. T. Seaton, in L. D. Hutcheson (ed.): *Integrated Optical Circuits and Components,* Chap. 9, Marcel Dekker, New York 1987, pp. 317 – 396.
83. D. M. Pepper, D. Fekete, A. Yariv, *Appl. Phys. Lett.* **33** (1978) 41.
84. G. P. Agrawal: *Nonlinear Fiber Optics,* Academic Press, New York 1989.
85. L. F. Mollenauer, R. H. Stolen, *Opt. Lett.* **9** (1984) 13.
86. L. F. Mollenauer, K. Smith, *Opt. Lett.* **13** (1988) 675.
87. U. Oesterberg, W. Margulis, *Opt. Lett.* **12** (1986) 585.

Nonwoven Fabrics

Felts and Leather Imitates are separate keywords.

Erich Fahrbach, Carl Freudenberg, Weinheim, Federal Republic of Germany

Gerhard Schaut, Carl Freudenberg, Weinheim, Federal Republic of Germany

Arnold Weghmann, Carl Freudenberg, Weinheim, Federal Republic of Germany

1. Introduction 1
2. Raw Materials 3
2.1. Fibers 3
2.2. Binders 5
3. Production Processes 7
3.1. Web Formation 8
3.1.1. Dry Web Formation 8
3.1.2. Wet (Hydrodynamic) Web Formation 10
3.1.3. Spunbondeds 12
3.2. Bonding of Nonwovens 14
3.2.1. Mechanical Bonding 14
3.2.2. Adhesive Bonding 16
3.2.3. Thermal Bonding 18
3.3. Finishing and Posttreatment 18
4. Economic Aspects and Areas of Application 20
5. Testing 21
6. References 21

1. Introduction

Definition. An unambiguous definition of nonwovens has not yet been found. In the broadest sense nonwovens are understood to mean textiles that are manufactured by nonconventional technologies. They are made of fiber webs which are bonded together as a result of inherent fiber-to-fiber friction (entanglement), mechanical treatment, heat, or chemical methods.

Nonwovens (Fig. 1 A, see next page) differ from conventional textiles manufactured from yarns [e.g., wovens (Fig. 1 B) and knitted fabrics (Fig. 1 C)] in that they are formed directly from individual fibers. Webs are the starting material for nonwovens (DIN 61 210), and are formed primarily from fibers, possibly with admixture of other fibrous constituents of various types and fiber lengths (e.g., cellulose pulp). The fibers are normally of a limited length, mainly staple fibers (length ca. 30 – 100 mm) that are suitable for spinning into yarn. They may, however, also be only a few millimeters long or of unlimited length (i.e., continuous filaments).

Nonwovens may be produced from staple fibers on a dry basis (dry-laid nonwovens) using methods based on conventional textile technology, or on a wet basis (wet-laid nonwovens) using methods based on conventional paper technology. Spunbondeds are not formed from staple fibers, but from continuous filaments that are produced by melt spinning.

The expression "bonded fabrics" was initially adopted in the United Kingdom. However the broader term "nonwoven fabrics" used in the United States has now found widespread usage. The terms "nonwoven textiles" or simply "nonwovens" are also encountered. These terms all denote textiles that are not woven. In Eastern bloc countries the term "textile composites" is also used to denote nonwovens that include textiles manufactured from fibers by sewing methods. Textile composites are, however, otherwise understood to be fabrics composed of at least two layers or plies bonded together. Nonwoven composites may include yarns as an additional structural element, although the nonwoven character should predominate. Fabrics made of fibrillated plastic films are not included in DIN 61 210.

Classification. The classification of nonwovens according to their method of manufacture has been adopted worldwide (Fig. 2). DIN 61 210 also specifies characteristic properties of nonwovens that can also be used for their classification.

Historical Development. Although nonwovens had been known since the 1920s, they only became important during World War II. The lack of high-quality cotton and viscose fibers during

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_565

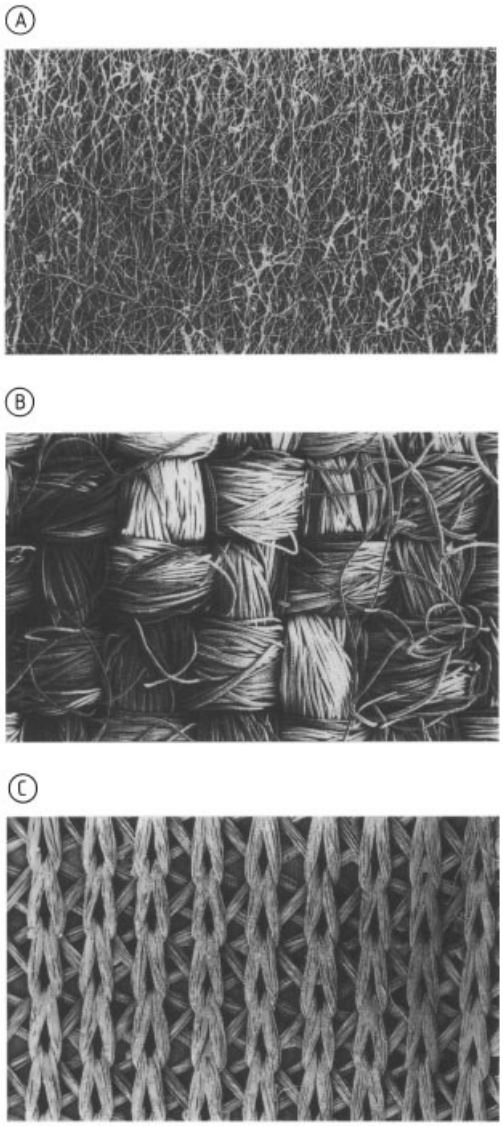

Figure 1. Comparison of a spunbonded nonwoven (A), a textile fabric made of yarn from staple fibers (B), and a knitted fabric of filament yarn (C)

the war led to attempts to bond poorer quality fibers that were not suitable for spinning as yarns and to discover new applications for these products. In the United States disposable products were developed (e.g., cleaning cloths, handkerchiefs, dressings and bandages, and sanitary and hygiene articles). However, as early as 1944 the Minnesota Mining and Manufacturing Company began producing nonwovens for electrical insulation applications; compliance with a specific quality standard was therefore essential.

Pioneers in the Federal Republic of Germany included Vorsteher and Bunger, Wuppertal, who claimed a patent for the production of dry-laid nonwovens for the shoe industry (e.g., for interlining, reinforcement, and insoles). " Capama-Stoffe" is the trademark of the equivalent product now manufactured by Franz Kalff, Stotzheim.

A nonwoven leather imitate was developed by Ten Cate and Nottebohm in 1930. The manufacturing process for this product was taken over by Carl Freudenberg, Weinheim in 1936. The fiber raw materials were cottonized hemp and viscose staple; natural rubber latex and poly(vinyl acetate) dispersions were used as binders. The production of " Vlieseline" (an interlining for the clothing industry) began in 1948. The production of durable " Vileda" domestic cloths followed in 1950; " Viledon" nonwovens for technical applications appeared shortly after.

In 1933 C. H. Dexter and Sons in the United States were the first to produce wet-laid nonwovens from vegetable-type fibers on a Fourdrinier paper machine that were used for tea bags. Manila hemp was subsequently used and was processed on more suitable rotary formers or inclined wire machines.

In 1964 Feldmühle (Oberlahnstein) produced wet-laid nonwovens on an inclined wire machine; the products were used as disposable sheets in casualty departments and isolation wards in hospitals.

In 1965 spunbonded nonwovens manufactured by Carl Freudenberg and Du Pont appeared on the market.

Du Pont's most important patent, relating to water entanglement, was granted in 1968. In 1970 ICI began production of thermobonded nonwovens based on the polyamide 6/6.6 Heterofil fiber. In the early 1980s microfiber nonwovens manufactured by the melt-blown method patented by Exxon began to appear on the market.

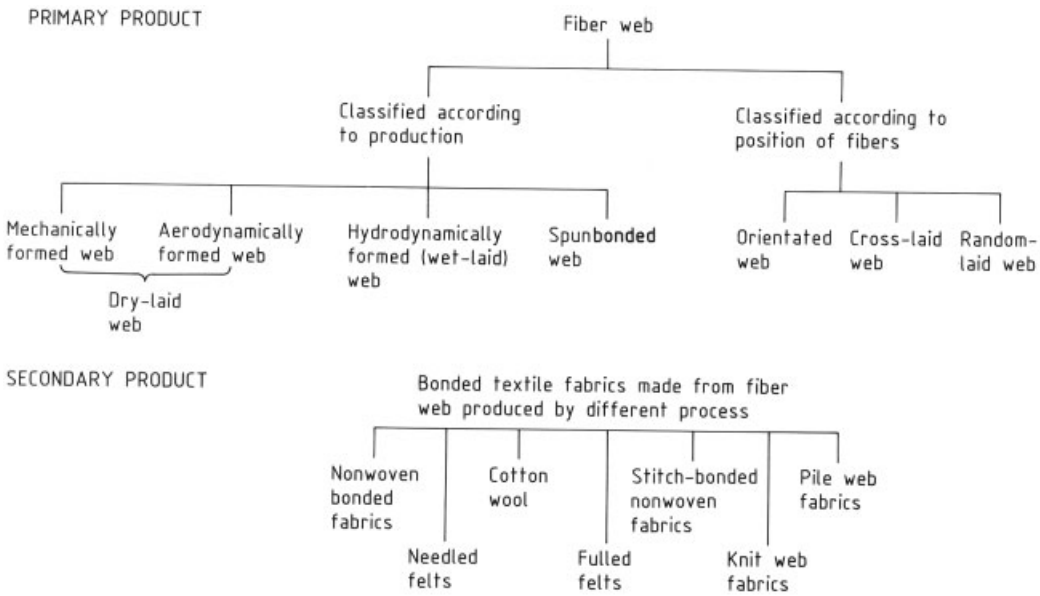

Figure 2. Technological classification of fiber webs – nonwovens

2. Raw Materials

2.1. Fibers [14–19]

The wide range of manufacturing methods for nonwovens permits the use of all commercially available fibers, natural fibers, mineral fibers, as well as man-made inorganic and organic fibers.

Synthetic organic fibers predominate; polypropylene fibers are the most important, followed by polyester, viscose, and polyamide fibers. Synthetic inorganic fibers (e.g., glass or mineral fibers) are far less important. These fibers are described elsewhere in detail (→ Cellulose ; → Fibers, 4. Synthetic Organic; → Fibers, 5. Synthetic Inorganic). The consumption of the most important fibers for the nonwovens industry is summarized in Table 1.

The fibers are incorporated into the nonwovens either in pure form or as mixtures in accordance with the required end-use properties and the manufacturing methods employed. Polypropylene, polyester, and viscose fibers account for more than 80 % of consumption in Western Europe. Polypropylene fibers exhibit the highest growth rate due to sharply rising capacities in spunbonded nonwovens and thermobonded staple-fiber nonwovens for hygiene and medical applications. Polyester also exhibits a marked growth rate due to increased demand for nonwovens in technical applications (e.g., filtration and interlinings for clothing).

Table 1. Summary of fiber consumption for nonwovens in Western Europe, the United States, and Japan (10^3 t/a)

Region	1984	1985	1986	1987
Western Europe [a]				
Polypropylene	73	94	109	132
Polyester	65	68	81	86
Viscose	52	52	49	51
Polyamide	10	10	8	10
Other synthetic fibers	7	5	15	18
Natural fibers	11	10	8	9
Total	218	239	270	306
United States [b]				
Polypropylene	54	62	75	90
Polyester	81	74	79	93
Viscose	57	52	57	57
Total	192	188	211	240
Japan [c]				
Polypropylene	23	25	26	28
Polyester	15	17	18	20
Viscose	14	15	16	17
Polyamide	11	12	13	14
Other synthetic fibers	5	6	6	7
Natural fibers	1	1	1	1
Total	69	76	80	87

[a] Including spunbonded nonwovens, according to the European Disposables and Nonwovens Association (EDANA), Brussels.
[b] Consumption of staple fiber, according to the Fiber Economics Bureau, excluding spunbonded nonwovens.
[c] Consumption of staple fiber, excluding spunbonded nonwovens.

The major proportion of fibers processed into nonwovens in 1987 consisted of regular staple

fibers for dry-laid and wet-laid nonwovens. The proportion in Western Europe was ca. 67 %, in the United States ca. 70 %, and in Japan ca. 65 %. The proportion of spunbonded nonwovens was ca. 31 % in Western Europe, 20 % in the United States, and 24 % in Japan.

Polypropylene fiber has a low density and is hydrophobic. It is used in the form of staple fibers and as filaments in spunbonded nonwovens, particularly nonwovens for diaper top sheets and the medical sector. On account of its resistance to both acid and alkali, it is also used in technical applications such as separators for batteries. On account of its low melting point, polypropylene fiber is ideally suited for thermobonding with heated calender rolls (see Section 3.2.3).

Polyester Fibers. Due to their easy processability as staple fibers, polyester fibers are important in a wide range of nonwovens. They can be bonded by chemical, mechanical, and thermal methods, and are extremely important in applications requiring a durable nonwoven (filters, interlinings, coating substrates, insulation materials). Polyester fibers are also a starting product for high-strength spunbonded nonwovens used as support materials for tufted carpets or bitumen roof substrates.

Polyamide fibers have a very high dry strength, a high flexural elasticity, and an outstanding abrasion resistance. They impart excellent softness to nonwovens in textile applications (e.g., interlinings).

Viscose fibers are derived from the natural polymer cellulose, they are hydrophilic and are used in absorbent nonwovens such as wiping cloths. On account of their nonthermoplastic nature, viscose fibers have become less important because they cannot be processed by modern thermal bonding methods. However, they could attract interest in the future due to increased use in wet-laid nonwovens and in the water entanglement process (see page 14) for manufacturing nonwovens that are environmentally friendly since viscose fibers are readily degradable.

Bicomponent Fibers. Special fibers are becoming increasingly important in nonwovens manufacture for achieving specific properties. Bicomponent fibers composed of different polymers are particularly important in this respect. Figure 3 shows the most common cross-sections of such fibers. Examples of polymer combinations are polyester – copolyester, polypropylene – polyethylene, polyester – polyamide, polyester – polypropylene, and polyamide 6 – polyamide 6.6. One polymer component has a lower melting point than the other. In a sheath – core fiber, for example, the sheath component has a lower melting point than the core. This is exploited in thermally bonded nonwovens. Heating melts the sheath, and bonding is achieved because the fibers adhere to one another and the core remains unchanged as a supporting fiber. In matrix – fibril fibers the matrix components can be chemically dissolved from the nonwoven, leaving the fibril fibers behind. These fibers are used to manufacture synthetic velour materials with a leatherlike character.

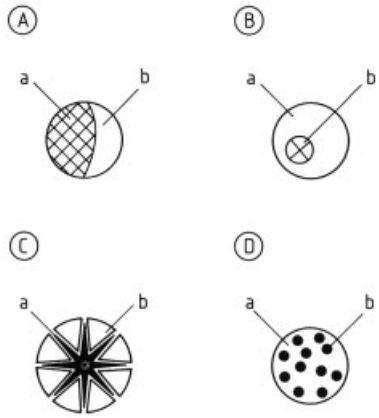

Figure 3. The most common cross sections of heterofilament fibers composed of two different polymers (a and b) A) Side-by-side; B) Core and sheath; C) Orange-type; D) Matrix and fibrils

Other Fibers. Other special fibers are used in small amounts for nonwovens. Polyacrylonitrile fibers are used in the filtration sector to remove dust from air because they are more resistant to hydrolysis than polyester fibers. Polyimide, polytetrafluoroethylene, and poly(phenylene sulfide) fibers are used in applications where high thermal resistance is required (e.g., in protective clothing or for hot gas filtration).

Fiber Properties Depending on the fiber fineness (titer in dtex, 1 dtex = 1 g/10^4 m), the length of the staple fibers varies between 28 and 100 mm: preferably between 40 and 60 mm for dry-laid nonwovens, and between 3 and 20 mm (preferably between 6 and 15 mm) for wet-laid nonwovens. Titers are preferably in the range 1.0 – 6.0 dtex. The most common standard titers are 1.7 – 2.2 – 3.3 dtex. Coarser titers (e.g., 17 to more than 100 dtex) are, however, found in high-bulk fiberfill nonwovens (e.g., for filter mats and abrasive materials) and in spunbonded nonwovens for building construction and civil engineering.

In recent years the category of fine fibers (1.0 – 2.2 dtex) has been extended to include very fine spun fibers (0.2 dtex) which have, for example, been used to produce lightweight nonwovens with good web uniformity and covering power. These very fine fibers are mainly used in mixtures with coarser fibers to improve the appearance of the nonwoven.

Even finer fibers can be obtained by electrostatic spinning and by using the melt-blown process, where the fiber diameter can be adjusted in the range 1 – 8 µm. Spunbonded nonwovens can also be produced with such fine fiber diameters (see page 13).

In the matrix – fibril fibers with a titer of 1.7 – 2.2 dtex, polyamide or polyester fibrils are embedded in a polystyrene matrix. The polystyrene matrix is dissolved, leaving the fibrils with a fineness of ca. 1/100 dtex and a cross section of 0.3 – 1.0 µm. Poromeric velour leather for clothing (Alcantara, Toray; Amaretta, Kuraray) can be produced from these fibers.

Fiber length and fineness are not the only important properties; the cross section, crimp, and shrinkage behavior are also significant.

Synthetic fibers are usually spun with a full, circular cross section, whereas spun viscose fibers have irregular, fissured cross sections with corresponding surface properties. Fiber surface properties affect fiber-to-fiber friction and thus influence the fiber-opening and web-formation processes as well as the shrinkage behavior of the bonded nonwoven. The same applies to synthetic fibers with geometrically simple structures (e.g., hollow fibers or fibers with a triangular, star-, or bean-shaped cross section). A minimum crimp is necessary if the fibers are to be used for production of dry-laid nonwovens. A high crimp is necessary for the manufacture of high-bulk, highly textured nonwovens, with crimps of 20 – 210 per 100 mm and a crimping factor of 10 – 50 (ratio of crimped to drawn-out length). The type of crimp also plays a role in determining the properties of nonwovens; a distinction is made between flat, two-dimensional and spiral, and three-dimensional crimp.

Fiber shrinkage is generally undesirable in nonwoven manufacture; fibers with 1 – 4 % shrinkage are normally used.

Thermoplastic bonding fibers in the homofilament or bicomponent form with widely differing melting points have become extremely important in the manufacture of nonwovens. The fibers should have a relatively broad softening range before the melting point is reached to permit bonding on application of pressure and temperature. In pressure-free bonding in a hot air stream the fibers should not shrink significantly before the adhesion temperature is reached. Fibers with thermoplastic and thus adhesive properties are usually mixed with nonthermoplastic fibers (e.g., viscose) to obtain binder-free nonwovens.

2.2. Binders

Nonwovens can be bonded in different ways. One of the most important methods for dry- and wet-laid nonwovens is the use of adhesive resins, commonly referred to as binders.

According to the European Disposables and Nonwovens Association (EDANA), ca. 31 000 t of binders (in aqueous dispersion, but calculated as dry substance) were used for the production of nonwovens in Western Europe in 1988. Contributions of the relevant binders were as follows: polyacrylates 41 %, styrene – butadiene copolymers 9 %, acrylonitrile – butadiene copolymers 9 %, poly(vinyl chloride) and copolymers 4 %, and poly(vinyl acetate), principally copolymers with ethylene, 23 %. Other binders account for ca. 14 % of the total. Consumption of binders in the United States was ca. 70 000 t in 1988, in Japan consumption was ca. 12 000 t.

The binders are selected according to the required properties of the nonwoven (e.g., softness, wash resistance, absorbency). Practically all film-forming binders can be used. For further

details on the binders described in this section, see → Adhesives.

Polyacrylate Dispersions. Polyacrylates and polyacrylate copolymers are preferred as binders for nonwovens because of the large number of available products and their broad range of properties. The polymers formed from the methyl, ethyl, and butyl acrylates and methacrylates are most important. The hardness and solvent resistance of the polymer decrease with increasing chain length of the alcohol moiety. Polymethacrylates have a higher film hardness than polyacrylates. Polyacrylic acid and polymethacrylic acid are soluble in water and alkali.

Highly cross-linked, water-resistant, solvent-resistant binders are obtained by copolymerizing acrylates with acrylic acid or acrylamide. The acid-containing copolymers can be externally cross-linked by adding zinc oxide, while the amide-containing copolymers can be externally cross-linked with N-methylol compounds such as urea – formaldehyde or melamine – formaldehyde resins. Internal cross-linking occurs if, e.g., *N*-methylolacrylamide is incorporated into the polymer; the addition of a cross-linking agent is then no longer necessary. Self-cross-linking with *N*-methylolacrylamide has become the most important type of cross-linking.

In addition to methyl methacrylate, other monomers (particularly styrene and acrylonitrile) are also copolymerized to modify the handle. In contrast to rubber-type polymers, acrylates do not have carbon – carbon double bonds. They are resistant to light, ageing, and, if well cross-linked, to washing and dry cleaning; they are not, however, as elastic and resilient as rubber-type polymers.

Rubber Latexes. Synthetic styrene – butadiene rubber (SBR) and acrylonitrile – butadiene rubber (NBR) are preferred to natural rubber on account of their high resistance to light, oxygen, and heat. Nitrile rubber polymers are also more resistant to oil, grease, and solvents. They are therefore used for durable household cloths for domestic use and in car care. Styrene rubber latexes are preferred for applications in which discoloration or staining is relatively unimportant and a low consumer price is desired (e.g., disposable articles, carpet backings, and filter mats). Sulfur cross-linking in a vulcanization oven can be avoided by copolymerization of methacrylic acid or acrylamide, which can be cross-linked with zinc oxide or aminoplasts on drying. Copolymers with reactive monomers (e.g., *N*-methylolacrylamide) are also available and self-cross-link on heating.

Of special interest are latexes that are made to coagulate at 40 – 60 °C by adding sensitizing and stabilizing agents; alternatively such latent coagulation properties can be imported to the latex polymer by the manufacturer. They are used to impregnate heavyweight webs and suppress the migration of the binder dispersion to the web surface that would otherwise occur on drying. Depletion of binder in the web interior can lead to delamination of the nonwoven under bending and buckling stress. As a result of coagulation at relatively low temperature (40 – 60 °C) the binder remains uniformly distributed over the web cross section. In the sensitization systems coagulation is initiated by raising the temperature and/or by altering the pH. The method is used particularly in the manufacture of leather imitates (e.g., substrates for coatings or heel-grip linings for shoes).

Poly(Vinyl Ester) Dispersions. The most important poly(vinyl ester) dispersions for web bonding are composed of copolymers of vinyl acetate and ethylene, and of vinyl acetate, vinyl chloride, and ethylene. The hardness of the binder can be adjusted by a suitable choice of the monomer ratios. A high ethylene content leads to soft binders, a high vinyl acetate content gives moderately hard binders, whereas vinyl chloride produces very hard binders.

As described for acrylate dispersions, thermally self-cross-linkable products or products that can be cross-linked with additives are obtained by incorporation of monomers bearing functional groups. These products have a very high water resistance and a satisfactory solvent resistance. Thermosensitive adjustment is also possible by using a suitable choice of dispersion stabilizing system and thermosensitizing agent.

Vinyl chloride copolymer dispersions have a good thermal sealability and a high-frequency weldability. These binders also provide a good basis for flame protection.

Poly(vinyl ester) binders are used for hydrophilic or hydrophobic hygiene and hospital

nonwovens, domestic wiping cloths, weldable fiberfill materials in automobile construction, flame-resistant filter media, and needle-felted flooring materials of low flammability.

Poly(Vinyl Alcohol). Poly(vinyl alcohol) is prepared by partial or total saponification of poly(vinyl esters), especially poly(vinyl acetate); film properties (water solubility, hardness, and strength) vary correspondingly. Cross-linking with urea or melamine – formaldehyde precondensates results in acceptable washfastness.

Polyurethane Dispersions. The impregnation of nonwovens (mostly needled) with polyurethanes dissolved in dimethylformamide, has been used since the 1960s for synthetic shoe upper materials and since the 1970s for synthetic velour leather for clothing. The binder range has been expanded to include aqueous polyurethane dispersions. These surfactant-free dispersions can be thermosensitively adjusted without difficulty. However, their high price has prevented large-scale use in the nonwovens sector.

Aminoplast and Phenoplast Precondensates. The majority of aqueous binders for nonwovens are emulsions and dispersions of elastomers and thermoplasts. Urea – and melamine – formaldehyde resins are used as aqueous precondensates for bonding glass fiber nonwovens and as cross-linking agents for reactive binders. Phenol – formaldehyde resins are used as binders for relatively coarse-fibered abrasive nonwovens with variable abrasive grain composition. In insulation mats for the automobile sector the phenol resin is generally added in powder form.

Articles produced from aminoplast and phenoplast precondensates, and to a lesser extent, self- cross-linking binders that have N-methylol groups contain free formaldehyde. Formaldehyde levels are restricted for health reasons, particularly in the clothing industry. Formaldehyde concentrations are measured by precisely defined methods. Table 2 gives an overview of existing regulations and free formaldehyde limits in various countries. In the Federal Republic of Germany the upper limit for textile fabrics in general is 1500 ppm. The formaldehyde concentration of adhesively bonded nonwovens can be kept well below this limit and is normally 75 – 300 ppm; fiber-bonded, nonfinished nonwovens are formaldehyde-free.

3. Production Processes

Two main processes based on textile processing methods are used to manufacture nonwovens, the primary products are dry-laid fiber webs formed from staple fibers and spunbondeds (Fig. 2). The two processes account for 80 % of the total world nonwovens market.

Table 2. Free formaldehyde limits in textile fabrics (including nonwovens)

Country	Regulation	Maximum concentration, ppm	Test method
Japan	Law 112	Group 1: up to 1000 (outerwear, protective wear linings) Group 2: up to 300 (shirts, blouses, nightgowns) Group 3: up to 75 (underwear, socks) Group 0, baby clothing	acetylacetone
United Kingdom	company-specific regulations (e.g., Marks & Spencer)	(300) – 700, (700) – 1000, new 500	Shirley Bulletin 48, 17, 1975; AATCC ∗ 112, 1978
United States	company-specific regulations (e.g., Levi Strauß);	1000, new 500	AATCC 112, 1978
	pending regulation	400 – 1000 with declaration	AATCC 112, 1978
Federal Republic	Harmful Products Law of Oct. 1st 1986, Amendment 1,	1500	modified De Jong Method or DIN 54 280
of Germany	No. 2.8; company-specific regulations	0 – 1000	acetylacetone
Denmark	File No. 892 (Ministry of Labor)	1000	acetylacetone according to Japanese Law 112
Finland	Regulation 845/1986, Ministry of Economics, effective Sept. 1st 1987	clothing for children under 2 years, 30; underwear, nightwear, 100; clothing with skin contact, 300	Finish Standard SFS 4996 acetylacetone according to Japanese Law 112

∗ AATCC: Association of American Textile Chemists and Colorists.

The wet process yielding wet-laid nonwovens is based on conventional paper technology and accounts for ca. 10 % of the market.

A further manufacturing technique, described particularly in the American specialist literature, is the dry paper process (air-laid pulp). A more recent development is the admixture of short fibers to cellulose pulp, this yields products which, like wet-laid nonwovens, are difficult to distinguish from paper [33].

3.1. Web Formation

3.1.1. Dry Web Formation [20–30]

This section describes dry processes for forming card webs and fibrous webs that start with spun fibers (generally synthetic fibers) of finite length, i.e., staple fibers. Fiber fineness ranges between 1 and 30 dtex, and fiber length between 20 and 120 mm.

In contrast to conventional yarn manufacture, fiber mixtures of different composition are processed to satisfy subsequent application criteria. Fiber blending, involving opening and mixing, is therefore subject to particularly stringent requirements. Two main routes are used:

1) Decentralized: an individual, web-forming line is preceded by a continuously operating opening and mixing line
2) Central: the fiber mixtures are prepared in batches on a central line and pneumatically conveyed to the fiber bins of the individual web-forming lines

The following factors are of importance for the design of the lines: output (in kilograms per hour), batch size, frequency of batch replacement, number of fiber components, quality requirements, and degree of automation.

The considerable increases in output of web-forming lines achieved in recent years have placed more stringent requirements on the degree of opening (loosening) and mixing; a high degree of flexibility and high productivity of the lines are also needed.

Blending involves the following steps:

1) Removal of the bale packaging,
2) Unpacking and opening the individual bales,
3) Weighing and metering the individual fiber components according to the formulation,
4) Mixing,
5) Storage, and
6) Opening

Depending on requirements and machine design, various functions (e.g., mixing and storage) can be combined in a single machine; one function can also be performed by several machines (e.g., opening). A highly uniform mass distribution along the length and in the cross direction is required in the end product, flock feed to the web former must therefore be precisely controlled. Conventional hopper feeders are equipped with electronic scales to ensure that a uniform, precise amount of flock is discharged per weighing cycle onto the conveyor belt of the web former. Volumetric chute feeding has become more important as a result of increasing outputs and machine widths. With volumetric chutes, a uniform specific density is achieved along the length by regulating the charging height and the compaction of the fiber flocks. The flock can be compacted mechanically in a so-called vibrating chute or with air. The latter method has the additional advantage of a self-regulating width distribution of the flocks due to discharge of air from the chute feed. Adjustable segments in the rear wall of the chute feed provide an additional means of modifying the web profile across the machine width.

Mechanical Web Formation on Cards. Carding is still the most economically important web-forming technique and has the largest market share. The carding machine opens, mixes, and homogenizes the flocks. It also orientates the individual fibers and converts them into a coherent web with a defined specific density. The design and construction of a roller card is described in DIN 64 118. In its simplest form it comprises feed rolls, licker-in rolls, a main cylinder, workers, strippers, a doffer, and take-off rolls. In modern machines all rolls have a metallic card clothing; the carding or take-off effect is influenced by teeth geometry, teeth density, teeth settings, relative roll speeds, and roller gaps.

Recent developments have been aimed at increasing productivity (e.g., by incorporating several doffers) and at influencing the degree of orientation of the fibers in the carded or fibrous web (e.g., by incorporating randomizing

rollers and/or compacting rollers). Changing the fiber orientation improves the machine direction – cross direction (MD/CD) ratio of the web strengths and results in high line productivity while maintaining the required quality properties. This primary product is also termed "tangled web", especially in the hygiene applications sector.

Further improvements include an increase of machine widths up to 4 m, and computer-aided, programmable line control. Line speeds of 150 – 200 m/min have already been achieved under continuous operation. The roller card can process practically all synthetic and natural fibers of different fineness. This fact combined with its substantially higher productivity have led to the predominance of the roller card in the nonwovens sector over the conventional flat cards used in cotton mills. The latter machines are only used in a few special fields.

The unidirectionally or randomly oriented fiber web (weight ca. 10 – 20 g/m^2) produced by the carding machine is laid on a conveyor belt; webs from two or more cards are generally laid one on top of the other (Fig. 4). The preferred orientation is in the machine direction. In bonded webs this orientation can produce a tensile strength in the MD that is up to ten times greater than that in the CD.

The carded web is taken up by a cross lapper which is equipped as a camelback with an oscillating laying device (Fig. 5), or with a horizontal laying device (Fig. 6), and is laid in a zig-zag pattern on the apron. The resultant cross orientation of the web can be varied by changing the laying angle and take-off speed.

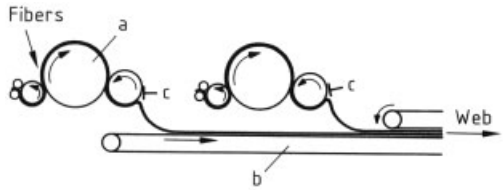

Figure 4. Two carding machines aligned in tandem
a) Card; b) Conveyor belt; c) Comb

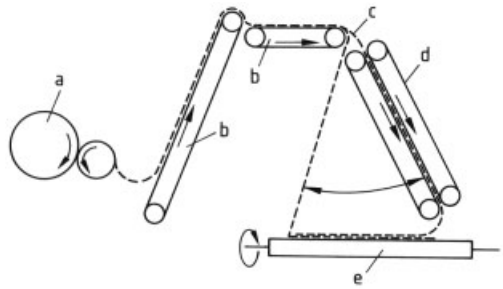

Figure 5. Oscillating cross lapper (camelback)
a) Card; b) Conveyor belt; c) Carded web; d) Oscillating laying device; e) Delivery belt (apron)

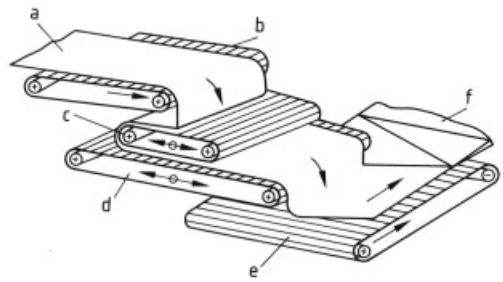

Figure 6. Cross lapper with horizontal laying device
a) Card web; b) Feed transport belt; c) Upper conveyor belt; d) Lower conveyor belt; e) Delivery belt (apron); f) Laid webs

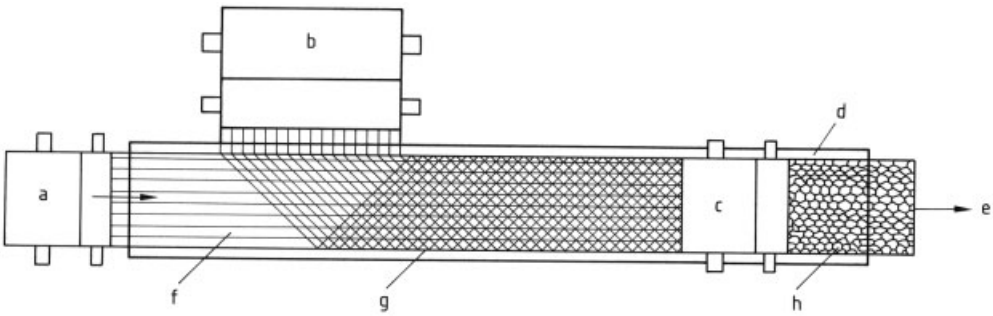

Figure 7. Combination of unidirectional and cross-laid webs
a) Card 1; b) Card 2; c) Card 3; d) Conveyor belt; e) Composite web; f) Multidirectional web; g) Cross-laid card web; h) Unidirectional web

Unidirectional and cross laying (Fig. 7) can be combined to form webs with uniform tensile strengths in all directions.

Aerodynamic Web Formation [31–34]. Pneumatic (aerodynamic) web formers have the same objective as carders, namely the production of randomized webs with isotropic strength properties. Almost all commercial machines are based on the principle shown in Figure 8. The randomized mass of fiber flocks is mechanically opened with a high-speed saw-toothed drum. The centrifugal force and a supporting air current dislodge the fibers from the sawteeth. The air current conveys the fibers to a delivery system (a suction drum or suction belt) where the web is formed under free-fall (with compressed air and/or suction) or in a closed system.

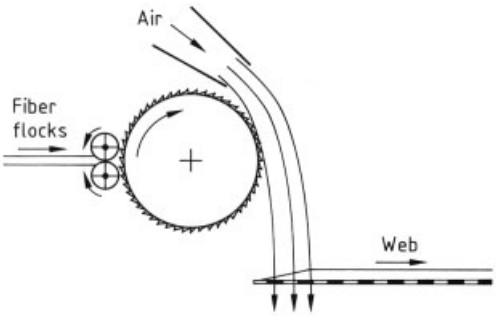

Figure 8. Basic principle of aerodynamic web formation (courtesy of Textilmaschinenfabrik Dr. Ernst Fehrer AG)

The stringent requirements concerning the blending of the fiber flocks and a uniform mass distribution during feeding (see introduction to this Section 3.1.1) are of particular importance because substantial opening and mixing do not occur in the aerodynamic web former.

Aerodynamic web formation is suitable for medium-fine to coarse fibers with a uniform titer. Suitability for fine fibers is limited because fiber mixtures tend to separate in the air current and during take-off. The process is particularly suitable for heavier, voluminous webs. An advantage is the low delamination tendency due to the three-dimensional, rooftile-like arrangement of the fibers in the web.

The aerodynamic web former in Figure 9 consists of four high-speed drums arranged in series. Doffing of the fibers takes place at four points by centrifugal force and is assisted by the suction effect of the screen belt. Additional carding is achieved by pairs of workers/strippers arranged above the carding drums.

3.1.2. Wet (Hydrodynamic) Web Formation [35–40]

The wet web formation process is based on paper manufacture technology. Short staple fibers (length generally < 20 mm) are dispersed in water, frequently together with cellulose pulp to give a dispersion with a density of 0.1 – 0.8 g/dm^3. This concentration is ten times more dilute than that used for paper manufacture. The fibers are deposited onto an inclined (Fig. 10) or rotary screen (Fig. 11) to form the web. Water is then uniformly sucked off over the machine width through suction boxes located be-

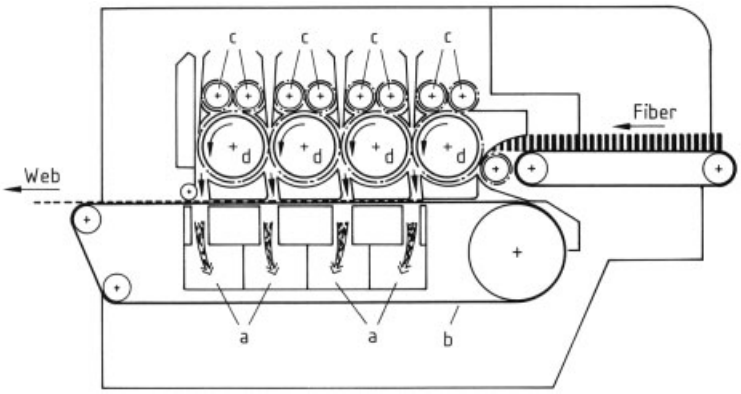

Figure 9. The K21 random card (courtesy of Textilmaschinenfabrik Dr. Ernst Fehrer AG)
a) Suction boxes; b) Screen belt; c) Worker rolls; d) Main cylinder

neath the screen. Deposition of the fibers on the screen is controlled by the filtration resistance of the web that builds up during the process. The inclined screen permits higher outputs and larger machine widths than the rotary screen, in which the layout and arrangement of the suction boxes and the sheet-forming zone are restricted. Inclined screen equipment such as the Hydroformer manufactured by Dorries (FRG), the No-Wo-Former manufactured by Bruderhaus Maschinenfabrik (FRG), and the Delta-Former manufactured by Sandy Hill Corp. (USA) are widely used.

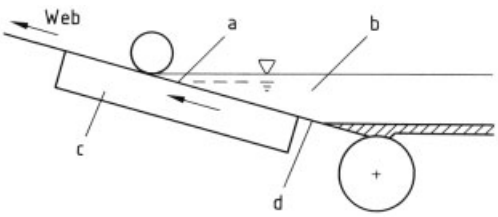

Figure 10. Principle of web forming on an inclined screen a) Inclined screen; b) Suspension; c) Suction box; d) Suction-free zone

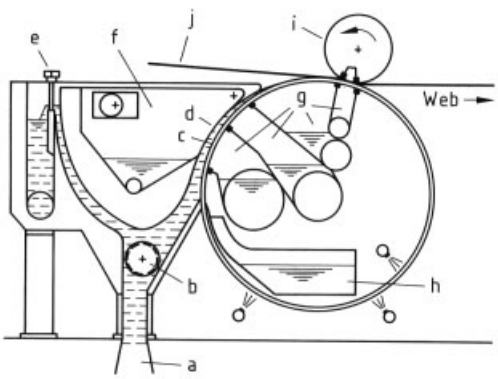

Figure 11. A rotary screen former (Sandy Hill rotoformer) a) Distributor; b) Dispersion roll; c) Sheet-forming zone; d) Adjustable front wall; e) Adjustable overflow; f) Weight regulator; g) Suction boxes; h) Water collection; i) Extraction roll; j) Delivery belt

Fiber orientation can be regulated by adjusting the web-forming speed as well as the suspension rates. A randomized web is formed if the two rates are the same, whereas a difference in the rates favors the formation of a unidirectional web.

The production rate depends on the water drainage and the desired weight. Machine manufacturers cite rates of 30 – 300 m/min for webs of 10 – 100 g/m^2. Up to 1500 g/m^2 are possible depending on the fiber material. The productivity of a 4.5-m wide machine is about 12 000 t/a.

The length of the processed staple fibers depends mainly on the fiber titer. Fine textile fibers of 1.7 – 3.7 dtex should be 6 – 15 mm long; from 5 dtex upwards, lengths of 20 – 25 mm are possible. Use of crimped fibers is limited on account of the danger of entangling; wet-laid nonwovens are therefore less voluminous and "textile-like" than dry-laid nonwovens. Large amounts of cellulose pulp can, however, be added without excessively lowering the strength. Wet processing can also be employed with short, brittle mineral fibers, which are mechanically destroyed on carding machines. Investment costs for a high-performance wet-forming machine are roughly five times higher than those for the dry process. This is, however, offset by the high productivity and lower labor costs in comparison to the dry process, although the economy of the high-performance machine is dependent on good design.

The binder can be precipitated on the fibers (preferably cellulose pulp) prior to web formation by coagulation in the mixing vat. Binders are, however, usually applied by spraying, padding, foaming, screen printing, or saturation; these methods are also used with dry-laid webs (see Section 3.2.2). The webs are dried on a suction drum or can dryer.

Developments are aimed at optimizing web formation. Multilayer webs, in which each layer is formed from different raw materials, can be obtained by altering the feeding system. New types of binder fibers permit thermal bonding on a suction drum dryer for more sophisticated applications. Water entanglement allows production of strong, more drapable, wet-laid textile nonwovens. The high web-forming rates that can be achieved in the wet process permit the admixture of ultrafine, synthetic fibers up to a titer of 0.6 dtex without adversely affecting process economy.

In the medium term, a trend towards the use of biologically degradable fibers (e.g., viscose fibers) can be detected which may also give a new stimulus to wet web-forming.

3.1.3. Spunbondeds [41–53]

In contrast to dry- and wet-laid nonwovens (Sections 3.1.1 and 3.1.2) which are made from staple fibers, spunbondeds are spun directly from thermoplasts (e.g., polyethylene, polypropylene, polyesters, polyamides, or polyurethanes) by the melt spinning process (→ Fibers, 3. General Production Technology, Chap. 6.).

As in conventional melt spinning, the molten polymer is extruded and forced through spinnerets. Thread forming and simultaneous drawing require high drawing rates comparable to those used in high-speed spinning processes for synthetic fiber production (up to 6000 m/min). Aerodynamic processes are now widely used for drawing the filaments. They can be subdivided into primary, primary – tertiary, or tertiary air current processes depending on the relative locations of the spinnerets and the web-forming device. The air from the ambient surroundings used to cool the spun filaments is termed secondary air. A fourth process involves the combination of mechanical drawing over galettes with the subsequent aerodynamic process.

The choice of drawing and cooling systems affects the physical properties of the fibers (i.e., their crystallinity, strength, fineness, and shrinkage). The spun filaments are finally laid on a perforated, endless collecting belt with the help of mechanical or aerodynamical deflection devices. The resultant web is held on the belt by suction and conveyed to the bonding device. The random arrangement of the individual filaments is responsible for the isotropic physical properties of the end product which is an important quality feature of spunbondeds.

Du Pont Process. In the Du Pont process the filaments are bundled, drawn in air currents, and simultaneously cooled. Separate air currents may be used for cooling and drawing (with optional mechanical drawing) by means of drafting rolls. The fiber bundles can be deflected after striking a deflector, spread out, and sucked onto a perforated endless belt. Alternatively, they can be electrostatically charged in an aspirator; the individual filaments then repel one another and are deposited as a web on a grounded (earthed) conveyor belt (Fig. 12).

Two different polymers with different melting points can be spun from separate spinnerets. The filaments are jointly bundled, drawn, and laid to form a web. The lower melting point component acts as a binder fiber either immediately or when heated to a temperature above its melting point but below that of the higher melting component.

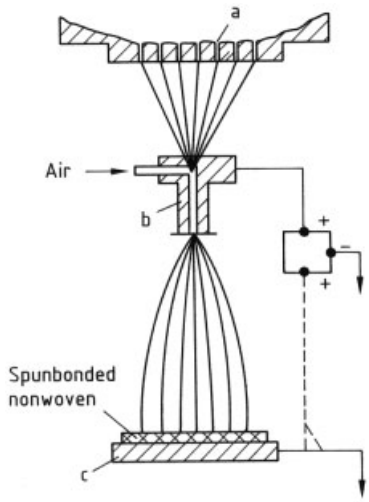

Figure 12. Schematic of the Du Pont spunbonded process
a) Spinning dies; b) Aspirator; c) Counterelectrode

Freudenberg Process. In the Freudenberg process (Fig. 13, see next page) the filaments are spun from rows of spinnerets fed from a spinning beam. The filaments are drawn in air currents in guide channels and laid under reduced pressure as a web on a suction belt. As in the Du Pont process, mixed webs can be produced with a lower melting point component that acts as a binder fiber.

Commercial Processes. In contrast to spunbonded processes developed by Du Pont, Freudenberg, and fiber manufacturers (e.g., Akzo, BASF, Hoechst, Rhône – Poulenc) whose spinning techniques have been developed only for their own use, companies such as Lurgi and Reifenhäuser market their processes and equipment worldwide.

The Lurgi – Docan process is based on the principle of tertiary air spinning and employs a rectangular-multifield die. The Reicofil process was introduced by Reifenhäuser in 1984 for the manufacture of polypropylene and polyethylene spunbondeds. The line uses an electri-

cally heated spinning head with a central melt line, the melt is distributed over the whole width according to the " clothes hanger system" used in flat-film production [48].

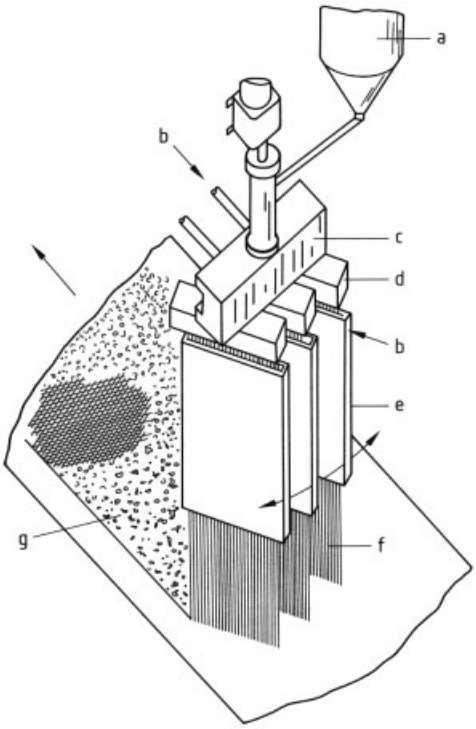

Figure 13. Schematic of the Carl Freudenberg process
a) Polymer melt; b) Air; c) Spinning beam; d) Spinning dies; e) Drawing channels; f) Drawn filaments; g) Web on transporting belt

Wet Spinning. In wet spinning the polymer (e.g., cellulose) is spun into a water bath where the fibers coagulate. The fibers are finally suspended on a suction belt (Asahi, Japan).

Flash spinning (evaporation spinning) developed by Du Pont has also become important. High-density polyethylene is dissolved in Freon or dichloromethane at elevated pressure and temperature in an autoclave to form a single-phase solution. This mixture is released from a nozzle into the atmosphere and expands. The low-boiling organic solvent vaporizes in an explosive manner (flash) and the polymer is scattered into a network of very fine filaments. Investment costs are roughly three times those for a wet line.

Exxon Melt-Blown Process. The melt-blown process developed by Exxon is a special spunbonded process. Molten polymer solution is forced at high pressure through dies (Fig. 14). On leaving the die the melt is entrained and drawn in the direction of the porous laying belt by a high-velocity hot air current. The incident forces must be sufficiently large to break off the fine fibers that are formed at the die exit. The fibers then pass through a current of cold air in which they solidify in their drawn form. They are deposited on the porous belt to form a web. Fibers with finenesses down to 1 µm can be spun as a result of process optimization and the development of special polymers with a high melt flow index (particularly polypropylene, polyester, polyethylene, and ethylene–vinyl acetate copolymer).

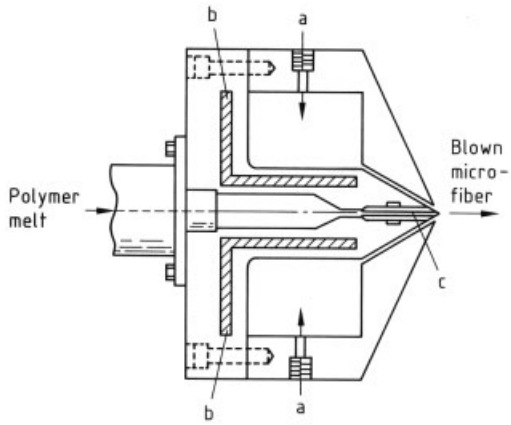

Figure 14. Melt-blown die
a) Hot air inlet; b) Heating plate; c) Capillary tube

Electrostatic Process [54, 55]. Extremely fine, electrostatically formed nonwovens are also classified as spunbondeds (Fig. 15). In order to manufacture extremely fine, bulky webs, a spray electrode is brought into contact with a polymer melt or solution. Charged polymer droplets are formed at the edge of the electrode and move with a threadlike deformation and possibly simultaneous evaporation of the solvent to a roller as counterelectrode. The resulting microfibers are continuously laid as a web (generally on a lightweight nonwoven carrier) and are covered on the upper side with a second lightweight nonwoven. The short deformed fibers in the web are electrostatically charged in

the high-voltage field; this improves the initial efficiency of filter materials.

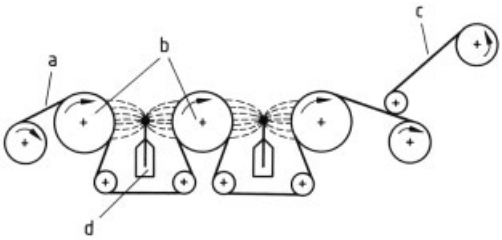

Figure 15. Schematic of electrostatic web formation
a) Carrier material; b) Counterelectrode rolls; c) Covering layer; d) Spray electrode

Production Lines. Production lines for spunbondeds can be designed in widths exceeding 500 cm. In the lightweight web sector, continuous line speeds of up to 400 m/min have been achieved. The webs are bonded (1) by the inherent adhesion of the filaments during web formation, (2) by subsequent heating in thermofusion ovens without pressure, or (3) by application of pressure on smooth or embossed, heated calender rolls. The binders employed to bond dry-laid or wet-laid spun fiber webs can also be used to stabilize spunbondeds. Heavyweight spunbondeds are often compressed mechanically by needling.

High outputs are obtained under continuous operation with relatively high production rates. The long downtimes resulting from the complex cleaning process required in the wet process are avoided. Profitable production is only possible on a large scale, necessitating extensive markets with a high consumer demand. The process is not as flexibile as the dry and wet processes. However, in certain application areas spunbondeds are undisputedly superior to other textile fabrics on account of their economic and qualitative advantages (e.g., tuft backing, geotextiles).

3.2. Bonding of Nonwovens

In rare cases a fiber web can be used directly in the nonbonded state; for example in medical, cosmetic, or industrial waddings, which cohere as a result of the fiber friction produced by the crimp. However, by definition such products are not classified as nonwovens, they are fabrics made of fibers or filaments that have been subjected to a consolidating posttreatment.

3.2.1. Mechanical Bonding [56–68]

Wool or animal hair can be felted by felting and milling under the action of moisture, heat, pressure, and movement. The fibers interlock with one another and become entwined, with simultaneous consolidation (→ Felts).

Needle Punching. A nonfeltable fiber web can be consolidated by needle punching. The web is compressed by rollers and then punctured with a needle board during passage through one or more needle punching machines (→ Felts, Chap. 4.1.). The properties of the nonwoven are primarily determined by the fineness, cross section, number, and shape of the barbs on the needles, as well as by the number of needles (needle density), and penetration depth. The fibers caught by the barbs are bundled, simultaneously drawn into the web on entry of the needle, and become disengaged during the reverse movement of the needle. The web is then moved on a few millimeters and the procedure is repeated (typical rate 1000 – 2200 strokes per minute using 6000 needles per meter of working width). Speeds are 1 – 25 m/min and the working widths are up to 16 m. Bonding can be intensified by turning the web and passing it through the machine again or by needle punching both sides simultaneously. Structured surfaces with a velour character and patterns with a three-dimensional structure (pile nonwovens) can also be obtained with specially designed needles and needle punching machines.

The most important areas of application of needled nonwovens are carpet floor coverings, filter materials, supporting materials for coatings, household and domestic cloths, papermachine felt, velour leather imitates, and medical compresses.

Water Entanglement (Water-Jet Bonding). The so-called spunlace process developed by the U.S. companies Du Pont and Chicopee is a logical development of mechanical needling of webs. In contrast to needle punching, relatively lightweight (20 – 120 g/m^2) webs can be bonded by entanglement produced by water jets to form

nonwovens at high speeds (up to ca. 150 m/min). Although water entanglement is suitable for all types of web-forming processes based on a wide range of synthetic and natural fibers, it is primarily employed for bonding dry-laid nonwovens. Crimped polyester and viscose fibers with titers of 1.7 dtex and a length of 40 mm are mainly used. The fiber webs are prewetted on a perforated or patterned belt and then sprayed with water jets expelled from fine circular nozzles of 100–150 µm diameter at a pressure of 1.5–25 MPa (Fig. 16). The water jets strike the web and the fibers become intertwined and entangled at the points of impact. The degree of entanglement and thus the intensity of the bonding varies depending on the number of rows of jets, the water pressure, the size of the nozzle orifices, and the design of the carrier belt. The patterning of the carrier belt permits combined bonding and structuring of the nonwoven surfaces (e.g., to produce perforated nonwovens). The process is extremely energy-intensive on account of the high water pressures. Since the process water rapidly becomes contaminated with fiber fragments, it must constantly be subjected to multistage fine filtration. The resulting nonwovens have a soft textile hand, but their dimensional stability is poor; the latter can, however, be improved by posttreatment (e.g., additional chemical bonding).

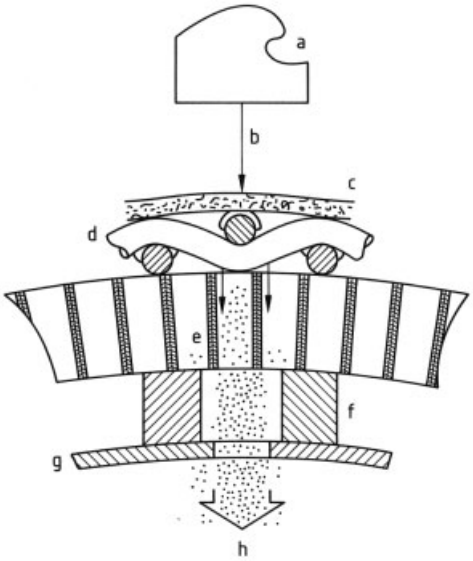

Figure 16. Principle of the Valmet spunlace bonding process a) Jet manifold; b) Water jet; c) Nonwoven; d) Entangling substrate; e) Honeycomb shell; f) Vacuum seals; g) Vacuum box; h) Process water

The main areas of application are medical products (e.g., operating theater drapes and gowns, gauze substitutes), domestic textiles, and wiping cloths. Since additional capacity for this technology is being developed in Europe, Japan, and the United States, a range of new areas of application is anticipated. Manufacturers of water-jet bonding lines include Valmet (USA), Perfojet (France), and Unicharm (Japan).

Stitch Bonding. Stitch bonding, which includes the Maliwatt and the Maliweb processes, is also a form of mechanical bonding for nonwovens.

In *Maliwatt nonwovens* the web is bonded with a stitching thread that is drawn with a sliding needle through a previously formed stitch in the nonbonded web to form a new stitch. The old stitch is knocked over, resulting in warp-stitch seams, which in their simple form have a separable fringe weave, but in their locked form are no longer separable and therefore have a more stable knitted weave. The Maliwatt machine is manufactured in the former German Democratic Republic. The Arachne machine (equipped with closable tubular needles but otherwise of very similar construction) and its product Arabewa are manufactured in former Czechoslovakia. Depending on the stitch length, the working speed is ca. 1–4.5 m/min. Web weights of 100–500 g/m^2 can be processed on up to 560-cm wide lines. *Maliweb nonwovens* are also produced by stitch bonding but without perforated stitching-thread needles. During their return movement the sliding needles pick up small bundles of fibers that are formed into a stitch on the reverse side of the web. Production speeds are ca. 0.25–7.5 m/min, depending on the stitch length. The line width and web weights are similar to those quoted for the Maliwatt machine.

A thread reinforcement technique developed by the Karl Mayer and Liba companies (FRG) combines the special properties of nonwovens (e.g., softness and high bulk) with the high moduli of elasticity of threads. A row of warp threads is worked into the web to produce longitudinal reinforcement. Transverse or diagonal threads can also be laid and combined with the nonwo-

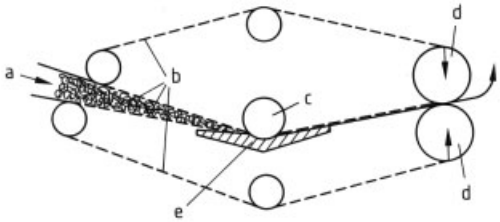

Figure 17. Saturator impregnating line, Rodney Hunt system
a) Web; b) Feed screen; c) Immersion roll; d) Squeeze rolls; e) Impregnating liquid

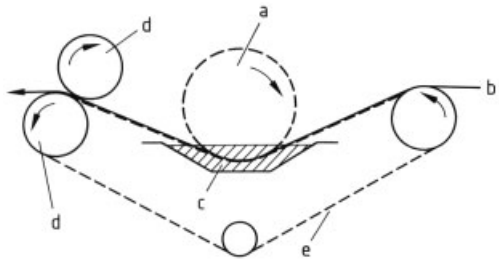

Figure 18. John Waldron impregnating line system
a) Perforated screen roll; b) Web; c) Impregnating liquid; d) Squeeze rolls; e) Screen belt

ven by interlocking with the length reinforcement. The properties of the end product depend on the number, orientation, and choice of thread material; stitch structure; and type of nonwoven.

Ultrasonic Bonding. This process was patented by Crompton and Knowles (USA) at the beginning of the 1970s. The mechanically produced vibration energy used to bond synthetic fibers (especially polypropylene) is generated in the frequency range 20 – 40 kHz. The width of the vibrating metal head (sonotrode) is only 200 – 250 mm. A number of sonotrodes therefore have to be arranged side by side in a calender. An unheated engraving roller of relatively small diameter serves as the anvil. There are few process parameters: the amplitude of the ultrasonic vibrations, the residence time of the web in the gap, the pressure load on the web, and the engraving geometry. The major advantage of the process is its substantially lower energy consumption compared to thermal bonding with calenders. It is used in particular with nonwovens having a weight greater than 100 g/m^2, e.g., in laminating processes (pinsonic process).

3.2.2. Adhesive Bonding [69–79]

The fibers in the web are usually bonded to one another with aqueous dispersions of adhesives (see Section 2.2). The binder content of the nonwoven is 20 – 60 wt %, generally between 20 and 30 wt %. The delicate, unstable web must be carefully handled during impregnation to avoid deformation.

Saturation. Two systems are conventionally used for saturation:

1) Impregnation between two screen belts (Fig. 17)
2) Impregnation between a screen belt and a screen roller (Fig. 18)

Flotation of the fibers and adherence to the screen belt may be a problem and require time-consuming cleaning. Accurate belt guidance is important, especially to ensure uniform suction when a suction device is used.

Impregnation in the Nip of a Padding Machine. This simple process (Fig. 19) can be used only with lightweight, readily wettable fiber webs. Application of the method is limited due to the short available distance for wetting the fiber that is determined by a fixed roller circumference and constant filling height of the trough.

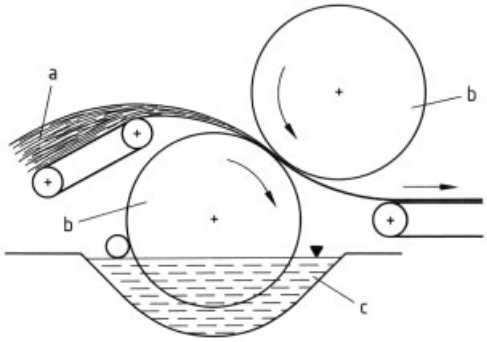

Figure 19. Impregnation of lightweight, nonbonded webs in the nip of a padding machine
a) Web; b) Squeeze rolls; c) Impregnating liquid

Foam Impregnation in the Nip of Squeezing Rolls. In the C. L. Nottebohm process the web is coated on one side with binder foam

which is then allowed to dry. Complete impregnation is achieved by subsequent padding. In the method developed by Carl Freudenberg the loose web is passed directly to the nip of two squeezing rolls; the nip contains a foamed binder dispersion (Fig. 20).

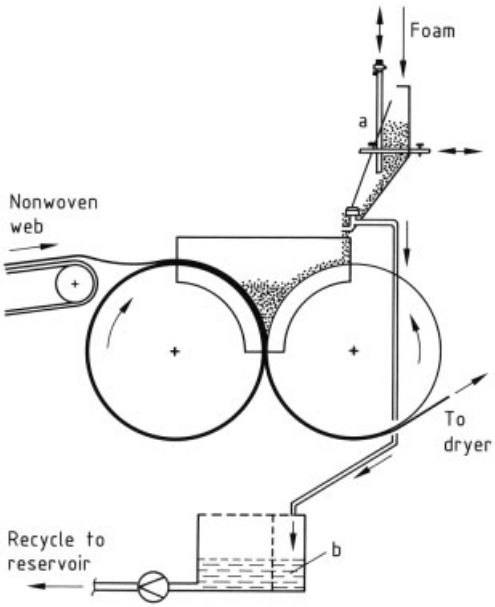

Figure 20. Foam application in the nip of a squeezer
a) Distributor vat; b) Binder dispersion

The binder is distributed at the crossover points of the fibers, in contrast to saturation, in which the fibers are coated from all sides. Foam impregnation produces a textile drape, a soft handle, large bulk, higher air permeability, lower migration tendency, savings in binder and drying energy, and higher production rates. Attention must be paid to the consistency, wetting properties, and stability of the foam. Due to its numerous advantages, foam impregnation is one of the most popular adhesive bonding processes.

Spray Bonding. High-bulk webs are usually bonded by this technique. The webs pass through a chamber where they are sprayed with the binder dispersion from nozzles moving at right angles to the direction of movement of the webs. Oscillating sets of nozzles are used for mean production rates below 30 m/min. Rotating sets of jets have proved suitable for large working widths and/or high production speeds.

Uniform distribution of the sprayed droplets over the cross section of the fiber web is achieved by applying suction from below the web. Thicker webs are normally dried and then passed through a second spray station to bond the reverse side (Fig. 21). Each spraying operation is followed by a drying stage.

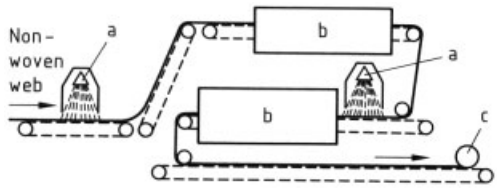

Figure 21. Spray bonding
a) Spray nozzle; b) Dryers; c) Rolling-up

Print Bonding. If the web is not to be bonded over its whole area but only partially in the form of a pattern or design, printing processes commonly used in the textile processing industry are employed (→ Textile Printing). In rotary printing a binder paste is forced with a doctor blade or roller bar through the perforations of the rotary screen onto the web. In relief printing the binder is transferred from the raised parts of a screen roller onto the web. Gravure printing and flexographic printing can also be used, but the web should then be prewetted or prebonded. Since the applied binders are generally colored, web bonding and dyeing can be performed in a single operation.

Drying and Cross-Linking. Binder-impregnated webs are converted into bonded nonwovens by drying and, if necessary, cross-linking and vulcanization. Air drying is generally preferred to contact drying on drying cylinders (can dryer). Infrared and high-frequency drying are rarely used.

In ventilation systems conventionally used in *air drying,* hot air is directed both parallel and perpendicular to both sides of the material. Single- and multi-tier dryers with air jets or parallel airstreams are preferred for relatively dense webs.

Air-through systems (suction drum dryer, Fig. 22) are preferred for the extremely intense drying of highly porous fabrics. Hot air flows inwards through the web. The drums operate under

reduced pressure. Honeycomb-type cylinders allow a higher vacuum and thus more efficient drying than conventional suction drum dryers. Air-through dryers can also be designed as band dryers with suction boxes arranged beneath the guide belt.

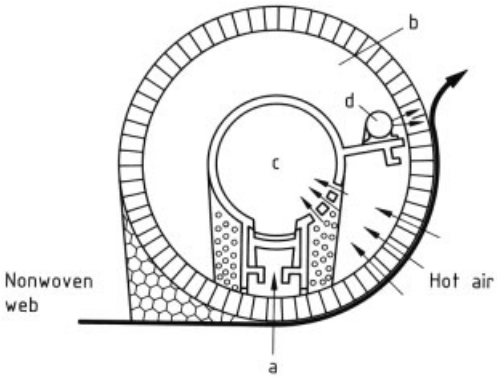

Figure 22. A honeycomb cylinder dryer
a) High vacuum suction; b) Vacuum suction; c) Air channel; d) Blow jet (to detach the web)

In *contact drying* with heated cylinders the impregnated web is dried and simultaneously smoothed in contact with the cylinder surface. Contact drying is only suitable for thin and relatively compact nonwovens.

Infrared radiation is used to coagulate heat-sensitive binders prior to drying to prevent binder migration. *High-frequency drying* also prevents migration, but has not been widely adopted on account of the high energy costs.

3.2.3. Thermal Bonding [80–93]

Thermal bonding of webs containing thermoplastic fibers is extremely important. Suitable fibers include those made of polypropylene, polyester and polyester copolymers, polyamide, modified poly(vinyl chloride), as well as sheath – core bicomponent fibers in which the low-melting component is utilized for bonding (see Section 2.1). Bonding with contact heat and pressure is the most important bonding method. In calender bonding (thermobonding) the web is bonded between two heated rolls with uniform pressure distribution over the machine width. In area bonding, pairs of steel – steel rolls as well as steel – coated rolls (e.g., with cotton or silicon) are used, depending on the weight and the required quality of the end product. In point bonding with engraved rolls, the engraving geometry is extremely important and affects the modulus of elasticity, softness, handle, and abrasion of the end product.

To obtain uniform line pressure, precise temperature control is required over the machine width; compensation of roll deflection in two-roll calenders is also very important in this respect. Calenders with deflection compensation rolls have been developed for temperatures up to 250 °C by Küsters (Hot-S-roll), Ramisch-Kleinewefers (Hydrein), and Escher Wyss/Freudenberg (Nipco).

In addition to calender bonding between two rolls, the Stork Beltbonder has achieved some importance, especially for producing powder-bonded nonwovens. Only a relatively low pressure can be achieved on the web to be bonded via the tension of the endless compression belt.

Thermal activation by means of hot air (thermofusion) is mainly used to manufacture high-bulk nonwovens. The air-through system is most suitable for this (see page 17) and has a continuously adjustable air flow, high temperature uniformity, and high heat-transfer coefficients.

The use of thermoplastic powders for bonding nonwovens had fallen into disuse, but is now enjoying renewed interest. Thermoplastic powders of copolyamides or copolyesters with melting points of 100 – 140 °C are uniformly incorporated into the fiber web and melted by heating with IR radiation or circulating air. The viscosity of the melted powder particles is low enough to ensure that web bonding occurs mainly at the fiber crossover points.

3.3. Finishing and Posttreatment [94–100]

Nonwovens can be finished in the same way as other textile fabrics, and to some extent also like paper or leather. Some treatments are, however, specific to nonwovens. Whenever the desired range of properties can be obtained by a suitable choice of raw materials, process materials, and manufacturing procedure, an attempt will always be made to avoid the finishing stage on grounds of cost.

Shrinking was introduced in the manufacture of nonwoven leather imitates to increase density and give the nonwoven a leatherlike handle. The needled web, containing 20 – 50 % of special, highly shrinkable fibers, is shrunk to ca. 50 % of its original area before further treatment in hot air, hot water, or steam.

Splitting of thick leather imitates into 2 – 6 plies is just as practicable as with natural leather. The same skiving machines with endlessly circulating band knives are used, with a gap accuracy of ± 0.05 mm.

Buffing with buffing machines conventionally used in the leather industry imparts a velourlike appearance to nonwovens.

Calendering as posttreatment (in contrast to calender bonding) can be used to impart a uniform thickness, for compaction, and to smoothen or emboss nonwovens. A more textilelike handle accompanied by loss in strength is obtained by partially destroying the binder adhesion.

Perforating and slitting result in drapable soft products. The water entanglement bonding process (see page 14) produces perforated structures directly. Needle or knife rolls can be used for perforation and notching of finished bonded nonwovens.

Compacting (Clupak process) and *crêping* (Micrex process) also serve to soften nonwovens. As in Sanforizing, the moist, generally adhesively bonded nonwoven is compacted in the longitudinal direction with structural compression between a compacted endless rubber belt and the smooth, steam-heated Clupak cylinder. In the Micrex process a dry nonwoven is compacted in the gap between a rotating grooved roller and a fixed guide plate; in contrast to the Clupak process a visible "crêping" effect can be obtained.

Dyeing is performed in adhesively bonded nonwovens with the application of the aqueous binder. Finely dispersed pigments in the binder liquor are usually sufficient. Binder-free, thermally bonded nonwovens can be produced from dyed fibers and can also be subsequently dyed either batchwise or continuously. Instead of being colored over its whole area, one or both sides of the nonwoven can be color printed.

Washing may be necessary to remove auxiliaries after dyeing and printing, to improve color-fastness, or to wash out surfactants and emulsifiers from nonwovens that are required to have a water-repellent finish. In household cloths and car cloths, washing removes pore-forming agents such as salts or urea. Full-width washing machines and suction drum washing machines are commonly used.

Coating of nonwovens is an important aspect of finishing. Area coating of nonwovens with poly(vinyl chloride) or polyurethane is commonly used in the shoe or leather imitate industry. Area coating with pressure-sensitive adhesives is also frequently used (e.g., in adhesive plaster substrates). Nonwovens coated with thermoplastic synthetic resins play an important role as fusible interlinings in the clothing and shoe industry. These dot-coated interlinings are bonded to the upper textile material or to leather under the action of heat and pressure. The interlinings impart form and shape to articles of clothing and are used in the ready-made garment industry. Compared to sewing or stitching, fusing (on steam presses, electric presses, and nowadays largely on continuous presses) has brought considerable cost advantages and a degree of standardization to ready-made garments. The proportion of fusing interlinings is about 80 – 85 % . The nonwovens are generally dot-coated with aqueous pastes or, occasionally, with anhydrous plastisols by rotary machine printing. The coating is dried in a hot air oven.

Preferred thermoplastic adhesives (hot melts) for fusing interlinings are copolyamides (e.g., of polyamide 6, 6.6, 11, or 12) or mixtures thereof with or without plasticizers. Copolyesters of widely varying monomer composition with melting points ranging from 100 to 140 °C are also important. Typical fusing conditions are a temperature of 100 – 150 °C for 4 – 12 s with minimum application of pressure. Under these conditions an optimum thermoplastic adhesive composition for the clothing industry results; this leads to the formation of an insoluble adhesion between the nonwoven and upper material that is resistant to repeated washing and dry cleaning.

On account of the temperature sensitivity of natural leather, the activation temperature of thermoplastic adhesives for shoe interlinings should not exceed 100 °C. Suitable adhesives include copolyesters, copolymers of vinyl chloride and vinyl acetate or of ethylene and vinyl acetate.

For simple fusing applications (e.g., in small parts of a garment such as waist bands) inexpen-

sive high-density polyethylene in powder form is scattered randomly onto the nonwoven substrate and sintered on.

An interlining nonwoven can also be spun-coated with a thermoplastic adhesive. Such spun webs can be extruded on release paper. After removal of the paper the spun nonwoven can be used for lamination processes. Nonwovens coated with thermoplastic adhesives have also been laminated by means of heat and pressure with membrane foils that are permeable to water vapor but impermeable to water and wind (e.g., Sympatex produced by Akzo, Goretex produced by Gore).

Dressing is often combined with nonwoven manufacture (e.g., by adding the agents to the binder liquor or, in the case of thermally bonded nonwovens, by a postfinishing treatment). Water-repelling agents (e.g., silicones and fluorocarbons), antistatic finishes, and flame retardants are applied in this way. Antibacterial, antimycotic, and fungicidal finishes can easily be applied with the binder liquor, or by postimpregnation in the case of thermally bonded nonwovens.

Heat setting is sometimes used to improve the dimensional stability (e.g., in interlinings). The nonwoven is exposed to hot air under controlled tension in a tenter frame dryer. The nonwoven material is transported through the heating zone of the tenter by holding both edges of the material at a predetermined width with tenter clips or needles.

The hot flue (a loop convection dryer) can also be used for thermal posttreatment. Like the suction drum dryer and tenter, the loop dryer is subdivided into several variable temperature zones, each zone can be traversed at a different speed.

These and other methods commonly used in textile processing can be applied to nonwovens.

4. Economic Aspects and Areas of Application

Economic Aspects. Industrial production of nonwovens (amounting to a few thousand square meters) began in the United States in 1944. After World War II nonwovens production experienced a sharp growth in all industrialized countries.

In 1987 ca. 30×10^9 m^2 of nonwovens were produced in the United States, Japan, and Western Europe. Annual growth rates up to 1987 were 10 %. Nonwovens production in the United States in 1987 was ca. 600 000 t, of which 150 000 t was spunbonded. In Japan production amounted to ca. 132 000 t, of which 30 000 t was spunbonded. Production for Western Europe (EDANA) was 342 000 t, spunbondeds accounting for 111 000 t.

Lightweight nonwovens below 25 g/m^2 (mainly used for diaper top sheets) account for slightly more than one-third of production. Nonwovens weighing between 25 and 70 g/m^2 account for almost a quarter, and nonwovens from 70 to 150 g/m^2 and above account for the remainder.

Areas of Application. Tables 3, 4, 5 summarize the production and use of nonwovens in Western Europe, the United States, and Japan, according to their main areas of application.

Table 3. Nonwovens production (in 10^3 t) in Western Europe according to areas of application (based on data supplied by EDANA)

Area of application	1984	1985	1986	1987
Diaper top sheet	64.4	73.4	85.2	96.9
Medical sector	8.8	7.9	11.7	13.2
Clothing with interlinings	15.8	16.4	17.1	19.2
Wiping cloths	24.6	24.8	27.6	28.0
Supports for shoes and leather goods	15.9	14.2	14.4	13.2
Decorative and domestic textiles	28.6	30.9	38.2	45.4
Automobile sector	5.6	5.0	5.6	8.5
Filtration (liquid – gas)	8.8	8.2	16.0 *	20.4 *
Civil engineering, building	49.2	52.1	61.4	69.0
Miscellaneous	35.0	37.4	34.2	30.0
Total	256.7	270.3	311.4	343.8

* Includes tea and coffee filters.

Table 4. Nonwovens production (10^6 \$) in the United States according to areas of application (various sources)

Area of application	1987
Diaper top sheet	290
Medical sector	240
Clothing with interlinings	215
Wiping cloths	205
Supports for shoes and leather goods	135
Decorative and domestic textiles	100
Automobile sector	65
Filtration (liquid – gas)	190
Civil engineering, building	295
Miscellaneous	280
Total	2015

Table 5. Nonwovens consumption (in 10^3 t) in Japan according to areas of application, excluding spunbondeds (various sources)

Area of application	1984	1985	1986	1987
Diaper top sheet	5.4	9.0	11.2	12.5
Medical sector	4.1	4.1	4.3	4.2
Interlinings	6.6	7.0	7.0	8.3
Wiping cloths	5.0	6.0	8.1	11.0
Supports for shoes and leather imitates	2.4	2.7	2.9	3.0
Automobile sector	11.3	11.4	11.9	13.1
Filters (liquid–gas)	14.4	17.0	18.7	20.9
Miscellaneous	29.9	31.0	31.2	29.8
Total	79.1	88.2	95.3	102.8

5. Testing [101–104]

Nonwovens are tested in accordance with national and international standards.

Federal Republic of Germany

DIN 53 482-79E	Electrical resistance
DIN 53 854	Weight determinations (1975)
DIN 53 855 1 + 2	Determination of the thickness of textile fabrics (1982)
DIN 53 857, T-2-79	Simple strip tensile test
DIN 53 859, T1, 2, 5-77/84	Determination of needle tear resistance
DIN 53 887	Determination of air permeability
DIN 53 890	Determination of crease recovery angle (1979)
DIN 53 878	Determination of dimensional change in textile fabrics (1987)
DIN 54 303 T1, 2-88	Behavior of nonwovens during dry cleaning
DIN 54 304 T1-88	Behavior of nonwovens during washing
DIN 54 305-76	Determination of the compression elasticity behavior
DIN 54 307	Indentation test (1982)
DIN 54 310	Separation of fixed interlining from upper material (1980)
DIN 54 311	Determination of dimensional change in fixed interlinings
DIN 61210 T1, 2-1987	Draft, definition and determination of nonwovens

International Standards Organization

ISO 9073-1	Determination of mass per unit area of nonwovens
ISO 9073-2	Dermination of thickness of nonwovens
ISO 9073-3	Determination of tensile strength and elongation of nonwovens
ISO 9073-4	Determination of tear resistance of nonwovens
ISO 9092	Textiles – nonwovens definition

European Disposables and Nonwovens Association

101-72	Determination of the liquid uptake capacity
201-73	Tensile strength testing
303-78	Determination of thickness
402-77	Determination of weight
502-80	Determination of bending length (cantilever)
601-75	Adjustment to standard climatic conditions
701-75	Determination of tear growth resistance
801-75	Determination of bursting pressure
901-76	Determination of drapeability
1001-78	Determination of brightness (brightness/degree of reflection in the blue part of the visible spectrum)
1101-78	Determination of opacity
1201-80	Water pressure test (water tightness)
1301-80	Sampling
150-0-84	Liquid penetration time through nonwovens
160-0-84	Surface “linting”
1401-81	Air permeability

United States

ASTM D 1117-80	Section
Abrasion resistance	18
Absorbency time and absorptive capacity	5
Air permeability	6
Breaking load and elongation of fabrics	7
Bursting strength	8
Conditioning	4
Dimensional change	9
Dry cleaning	10
Flexural rigidity	11
Recovery from creasing	12
Seam breaking strength	13
Tearing strength of fabrics	14
Trapezoid tear	14
Tongue tear	15
Elmendorf tear	16
Thickness	19
Weight	17

6. References

General References

1. M. S. Casper: *Nonwoven Textiles,* Noyes Data Corporation, Park Ridge, New Jersey 1975.
2. M. T. Gillies: *Nonwoven Materials, Recent Developments,* Noyes Data Corporation, Park Ridge, New Jersey 1979.
3. R. Krčma: “Geschichte und Perspektive der Entwicklung von Vliesstoffen,” *Technische Textilien* **31** (1988) no. 4, 99 ff.
4. H. Kirchenberger: *Vliesstoffe,* Verlag J. L. Boudi & Sohn, Wien-Perchtolsdorf 1984.
5. P. Boettcher: “Zeittafel der Vliesstoffentwicklung,” *Textiltechnik* (Leipzig) 37 (1987) 630.
6. W. Wirth: “Vliesstofftechnik – Neuentwicklungen und Trends,” *Internationales Textil-Bulletin Garnherstellung* **34** (1988) 43 ff.

7. H. Jörder: *Textilien auf Vliesbasis (Nonwovens),* P. Keppler Verlag, Heusenstamm 1977.
8. J. Lünenschloss, W. Albrecht (eds.): *Vliesstoffe,* Thieme Verlag, Stuttgart – New York 1982.
9. P. Sroka: *Handbuch der textilen Fixiereinlagen,* 2nd ed., SINUS Verlag, Krefeld 1983.
10. K. H. Stukenbrock: "Nichtgewebte Textilien, Entwicklung, Herstellung, Eigenschaften und Anwendung," *Melliand Textilber.* **43** (1962) 705, 831.
11. W. Albrecht: "Vliesstoffe: Ihre Bedeutung und Entwicklungstrends," *Chemiefasern Textilind.* **38** (1988) no. 90, 672.
12. M. Jacobsen: "Nonwoven Bonding Technologies, A Primer on the Basic Bonding Techniques," *Nonwovens Ind.* **10** (1989) 43 ff.
13. *Vliesstoff-Technik,* Arbeitgeberkreis Gesamttextil, Frankfurt am Main 1987.
14. W. Loy: "Vliesstoffbinder in Faserform," *Chemiefasern Textilind.* **39** (1989) no. 91, 288.
15. L. Jarecki: "New Fibers in Western Europe," *Am. Text.* **17** (1988) July, 50.
16. M. Salamon et al.: "Synthesefasern für die Herstellung von Vliesstoffen auf nassem Weg," *Chemiefasern Textilind.* **24** (1974) no. 76, 639.
17. V. Kacvinsky: "Faserstoff mit hohen Gebrauchseigenschaften für technische Anwendungsbereiche, Teil 1: Aramidfasern," *Technische Textilien* **31** (1988) 131.
18. W. Lauppe: "Der Einfluß des Rohstoffes Fasern und seiner Eigenschaften auf die Herstellung und Verwendbarkeit von Vliesstoffen," *Melliand Textilber.* **57** (1976) 290.
19. W. Albrecht: "Chemiefasern für die Vliesstoffindustrie," *Textilveredlung* **23** (1988) 47 ff.
20. R. Knaus: "Die aerodynamische Vliesbildung," *Chemiefasern Textilind.* **24** (1974) no. 76, 209.
21. C. L. Nottebohm: "Vliesstoffe auf trockenem Weg," *Chemiefasern Text. Anwendungstech.* **18** (1968) 667.
22. C. L. Nottebohm: "Trocken-Vliesstoffe, Herstellung und Verwendung," *Z. Gesamte Textilind.* **71** (1969) 172 ff.
23. W. Wirth: "Bedeutung der Krempel auf dem Nonwoven-Gebiet," *Melliand Textilber.* **61** (1980) 391 ff.
24. W. Wirth: "Herstellung leichter Vliesstoffe – Neuentwicklungen," *Melliand Textilber.* **62** (1981) 636 ff.
25. W. Wirth: "Vliesstofftechnik – Neuentwicklungen und Trends," *Internationales Textil-Bulletin Garnherstellung* **1** (1988) 43 ff.
26. D. H. Müller, S. Bernhard: "Hohe Produktionsgeschwindigkeiten beim Thermobonding von Nonwovens," *Melliand Textilber.* **67** (1986) 360 ff.
27. F. Morgner: "Neue Entwicklungen für das öffnen und Mischen von Fasern," *INDEX 87 Kongreß, Sitzung A3,* Genf 1987.
28. D. Müller: "Random Orientation in the Case of Carded Nonwovens," *EDANA's UK Nonwovens Symposium,* London 1989.
29. S. Schlichter: "Aktuelle Entwicklungen in der Vorbereitung von Fasermaterial für die Vliesstoffherstellung," *TPI Text. Prax. Int.* **44** (1989) 227 ff.
30. H. Jakob: "Erfahrungen mit der Wirrvliestechnologie," *Melliand Textilber.* **70** (1989) no. 3, 185 ff.
31. T. Miura: "Totally Dry Nonwoven System Combines Air-Laid and Thermobonding Technology," *Nonwovens World,* **73** (1988) March, 46 ff.
32. P. Lennox-Kerr: "Textilien aus Folien," *Z. Gesamte Textilind.* **70** (1968) 331.
33. H. S. Jensen: "Start-Up of the Latest M & J Air-Laid Installations," *Impact* (1987) 93.
34. M. Pool: "Start-Up of Kotka Dry Forming Plant," *Insight Conference,* 1985.
35. Th. Ploetz: "Vliesstoffe nach dem nassen Verfahren," *Chemiefasern Text. Anwendungstech.* **18** (1968) 846 ff.
36. P. Graf: "Die hydrodynamische Vliesformierung und deren Besonderheiten im Vergleich zu anderen Formierungsmethoden," *Melliand Textilber. Int.* **50** (1969) 404.
37. D. A. Shiffler: "Defective Fibers in Wet-Lay Nonwoven Fabrics," *Tappi J.* **71** (1988) June, 117 ff.
38. B. Scholz: "Wet-Lay Nonwovens-Production. Methods and Trends for the Future," *Pap. Trade J.* **34** (1972) 44.
39. S. Schmidt: "über die Herstellung von Vliesen auf nassem Wege," *Wochenbl. Papierfabr.* **97** (1969) 1015 ff.
40. C. F. White: "Wet Forming High Performance Precursor Nonwoven Webs," *Impact 88 High Performance Fabrics Conference,* Stuart, Florida 1988.
41. L. Hartmann: "Vliesstoffe – mit besonderer Berücksichtigung der Spinnvliese," *Chemiefasern Text. Anwendungstech. Textilind.* **22** (1972) no. 74, 231.

42. L. Hartmann: "Spinnvliestechnik und industrielle Anwendungen von Spinnvliesen," *Melliand Textilber. Int.* **54** (1973) 460.
43. L. Gerking: "Spinnvliese – Herstellungsverfahren und Besonderheiten," *Melliand Textilber. Int.* **55** (1974) 326.
44. F. Fourné: "Neue Spinnvliesverfahren," *Chemiefasern Textilind.* **29** (1979) no. 81, 446.
45. F. Fourné: "Spunbond: Processes and Plants," *Man-Made Fiber Year Book,* CTI Fachverlag, Frankfurt/Main 1988, 66 ff.
46. U. Liebscher et al.: "Erzeugniskonstruktion und Herstellung von Elementarfadenvliesstoffen unter Berücksichtigung des Einsatzes von Schlüsseltechnologien," *Technische Textilien* **31** (1988) no. 4, 114 ff.
47. H. Jörder: "Weltweiter überblick über die Hersteller von Spinnvliesstoffen und ihre Erzeugnisse," *Allgemeine Vliesstoff Report* **4** (1975) 200.
48. M. Makita: "Cellulose-Spinnvliesstoffe," *Melliand Textilber.* **57** (1976) 635.
49. M. Wehmann: "New Processes Offer Alternatives for Spunbonded Nonwovens," *Nonwovens World* **11** (1987) 52 – 54.
50. L. C. Wadsworth et al.: "Novel Melt Blown Research Findings," *Nonwovens Ind.* **17** (1986) 17 ff.
51. W. J. McCulloch et al.: "A Systems Approach to Nonwovens Production Using the Melt-Blowing Process," *Nonwovens World* **2** (1987) 67 ff.
52. D. T. Lohkampp, J. P. Kelev, US 3 825 379, 1974.
53. W. Brandt: "Vliesbildung aus Endloskabeln," *Allgemeine Vliesstoff Report* **10** (1974) 228.
54. M. Davis: "Electrostatic Melt Spinning Process Delivers Unique Properties," *Nonwovens World* **2** (1987) 51 ff.
55. A. Weghmann: "Production of Electrostatic Spun Synthetic Microfiber Nonwovens and Applications in Filtration," *Nonwovens Ind.* **13** (1982) 24 ff.
56. R. Knaur: "Neuentwicklungen auf dem Gebiet des Vlies- und Nadelmaschinenbaues," *Melliand Textilber. Int.* **53** (1972) 981.
57. J. Lünenschloss: "Die mechanische Verfestigung von Vliesen aus Chemiefasern durch Vernadeln – einige neue Erkenntnisse," *Lenzinger Ber.* **36** (1974) 136.
58. J. Zocher, W. Reichelt: "Oberflächenstrukturierung mechanisch vorgefestigter Faservliese," *Text. Prax. Int.* **29** (1974) 467.
59. M. Jacobsen: "What's New in Needlepunched Nonwovens," *Nonwovens Ind.* **19** (1988) no. 3, 22 ff.
60. H. C. Bocksrucker: "Eigenschaften genadelter Spinnvliese," *Chemiefasern Textilind.* **23** (1973) no. 79, 226.
61. J. Zocher: "BIPOL-Non-wovens: Mechanisch vorverfestigte und doppelflächig strukturierte Faservliese," *Chemiefasern Textilind.* **25** (1975) no. 77, 349.
62. H. Jörder: "Verfestigung von Faservliesen mit Hilfe von Wasserstrahlen (Spunlaced-Vliesstoffe)," *Allgemeine Vliesstoff Report* **3** (1974) 75.
63. G. K. Moore: "Aspects of Nonwoven Bonding: The Development and Potential of Hydroentanglement," *Tappi J.* **71** (1988) July, 112 ff.
64. W. Scholtis et al.: "Maliwatt- und Malivlies-Nähwirkverfahren – einige neue technologische Erkenntnisse," *Dtsch. Textiltech.* **21** (1971) 513.
65. B. Frenzel, H.-D. Hentschel: "Vliesstoffherstellung auf Malimo-Nähwirkmaschinen," *Dtsch. Textiltech.* **20** (1970) 368.
66. M. Turi: "The Outlook for Spunlaced Nonwovens," *Nonwovens Ind.* **19** (1988) November, 30 ff.
67. J. R. Starr: "Water Jet Entangled Nonwovens Expanding Rapidly," *Nonwovens World* **3** (1988) 62 ff.
68. D. F. Beaumont, K. R. Randall: "Rotary Hydraulic Entanglement of Nonwovens," *Nonwovens World* **1** (1986) 76 ff.
69. "A Guide to Binders and Additives for Nonwovens," *Nonwovens Ind.* **18** (1987) no. 11, 37 ff.
70. K.-A. Weber: "Der Einsatz von Bindemitteln bei der Herstellung von Vliesstoffen," *Textilveredlung* **1** (1966) 447.
71. W. von Langenthal, G. A. M. Butterworth: "Dispersionsbinder und Faser – ein Beitrag zur Vliesverfestigung," *Melliand Textilber. Int.* **53** (1972) 687.
72. G. Sinn, W. von Langenthal: "Synthetische Kautschuklatices für die Bindung von Vliesstoffen," *Chemiefasern Textilind.* **23** (1973) no. 75, 570.
73. H. Moroff: "Acryl-Polymere für die Bindung von Vliesstoffen," *Chemiefasern Textilind.* **23** (1973) no. 75, 574.
74. E. Nischwitz: "Vinyl-Polymere für die Bindung von Vliesstoffen," *Chemiefasern Textilind.* **23** (1973) no. 75, 646.

75. H. Jörder: "Die Migration von Bindemitteln bei der Trocknung von Vliesstoffen," *Melliand Textilber. Int.* **54** (1973) 1217.
76. K. Fischer: "Was leisten Polymer-Dispersionen bei der Vliesbindung?" *Chemiefasern Textilind.* **38** (1988) no. 90, 678.
77. F. Matejcek: "Bindung von Naßvliesen bei physikalisch-chemischer Koagulatbildung," *Chemiefasern Textilind.* **23** (1973) no. 75, 657.
78. J. Troesch: Die Thermomigration wäßriger Binder-Dispersionen," *Allgemeine Vliesstoff Report* **3** (1974) 180.
79. A. Einwiller: "Thermosensible Polymerdispersionen und ihre Anwendung als Bindemittel," *Allgemeine Vliesstoff Report* **5** (1976) 78.
80. W. Specht: "Mischpolymerisatfasern für die Bindung von Vliesstoffen," *Chemiefasern Textilind.* **23** (1973) no. 75, 651.
81. W. Loy: "Vliesstoffbinder in Faserform," *Chemiefasern Textilind.* **39** (1989) no. 91, 3, 228 ff.
82. E. Tudósy: "Pulververfestigung von Vliesen mittels Stork-Beltbonder," *Allgemeine Vliesstoff Report* **6** (1988) 372 ff.
83. D. H. Müller, H. Nacken: "Textilkalender und damit erzielbare Effekte," *ITB Veredelung* **2** (1988) 5 ff.
84. S. Prömpler: "Thermobonding Kalender," *2. Hofer Vliesstoff-Seminar,* Hof Nov. 5th 1987.
85. W. Albrecht: "Copolyamid- und Polyesterfasern für die Bindung von Vliesstoffen," *Chemiefasern Textilind.* **23** (1973) no. 75, 653.
86. A. Watzl: "Faserherstellung und Vliesverfestigung für thermo- und chemisch verfestigte Nonwovens," *Melliand Textilber.* **69** (1988) 714.
87. H. Houben, M. Pabst: "Verkleben von Synthesefasern unter Hochtemperatur-Bedingungen," *Chemiefasern Textilind.* **23** (1973) no. 74, 564.
88. H. Ruef: "Latest Technology in Fibers & Nonwoven Production for Binder and Thermobond Nonwovens," *Impact 88 High Performance Fabrics Conference,* Stuart, Florida 1988.
89. H. Drösler: "Auslösung der Bindung (Verfestigung) durch Trocknung und Trockner," *Chemiefasern Textilind.* **23** (1973) no. 75, 463.
90. P. Ehrler, I. Jamitza: "Wasserlösliche Polyvinylalkoholfasern für Naßvliesstoffe," *Chemiefasern Textilind.* **23** (1973) no. 75, 746.
91. I. W. Schappel: "Cellulosische Bindefasern," *Chemiefasern Textilind.* **23** (1973) no. 75, 740.
92. H. Endler: "Das Prägen und Verfestigen von nichtgewebten Textilien," *Allgemeine Vliesstoff Report* **3** (1974) 244.
93. M. F. Meyer, R. L. McConnell: "Production of Nonwovens by Powder Bonding," *Nonwovens World* **1** (1986) 57 ff.
94. F. Böttger: "Veredlungsprozeß für Vliesstoffe," *Chemiefasern Textilind.* **24** (1974) no. 76, 217.
95. H. Jörder: "Ausrüstungsverfahren für Vliesstoffe und ähnliche nicht gewebte Textilien," *Textilveredlung* **7** (1972) 317.
96. K. H. Stukenbrock: "Chemische Grundlagen von Heißsiegelklebern für Fixiereinlagen," *Textilveredlung* **6** (1971) 459.
97. H. Steukart: "Die Fixiertechnik in der Bekleidungsindustrie. Theorie – Maschinen – Praxis," *Textilveredlung* **6** (1971) 469.
98. U. Einsele: "Flammhemmende Ausrüstungen von Vliesstoffen und Nadelfilzen," *TPI Text. Prax. Int.* **30** (1975) 1559, 1672.
99. K. Klesper: "Antimikrobielle Ausrüstung von Vliesstoffen und Nadelfilzen," *TPI Text. Prax. Int.* **30** (1975) 1011.
100. H. Schmiedgen: "Antistatische Ausrüstung von Vliesstoffen und Nadelfilzen," *TPI Text. Prax. Int.* **31** (1976) 890.
101. H. Jörder: "über die Prüfung von Vliesstoffen I, II, III," *Z. Gesamte Textilind.* **64** (1962) 593; **67** (1965) 22; **71** (1969) 299.
102. H. Jörder: "Prüfung von Vliesstoffen," *Allgemeine Vliesstoff Report* **1** (1972) 28.
103. H. Jörder: "Prüfung von Vliesstoffen und ähnlichen nicht gewebten textilen Flächengebilden," *Allgemeine Vliesstoff Report* **4** (1975) 152.
104. J. W. Palmer: *Textile Processing and Finishing Aids,* Noyes Data Corporation, Park Ridge, New Jersey 1977.

Novoloid Fibers → High-Performance Fibers

Nuclear Technology

Eberhard Teuchert, Leverkusen, Federal Republic of Germany (Chap. 1)

Joachim K. Axmann, Institut für Wissenschaftliches Rechnen, Technische Universität Braunschweig, Braunschweig, Scandpower GmbH, Hamburg, Federal Republic of Germany (Chap. 2)

Bernd Klüver, EnergieSystemeNord GmbH, Kiel, Federal Republic of Germany (Chap. 2)

Peter-Jürgen Meyer, formerly Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Sections 3.1, 3.1.1, 3.1.2.5, and 3.4)

Hans Hollinger, Siemens AG, Unternehmensbereich KWU, Offenbach, Federal Republic of Germany (Sections 3.1 and 3.1.2)

Ivan Kausz, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.1.3)

Werner Debray, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.1.4)

Manfred Erve, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.1.4)

Rolf Riess, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.1.5)

Rudolf Nieder, Interatom GmbH, Bergisch-Gladbach, Federal Republic of Germany (Section 3.2)

Klaus Kotthoff, Gesellschaft für Reaktorsicherheit, Köln, Federal Republic of Germany (Section 3.3)

Dietrich Budnick, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.4)

Ulrich Dörfler, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Section 3.4)

Siegfried Dreyer Interatom GmbH, Bergisch-Gladbach, Federal Republic of Germany (Chap. 4)

Karl-Ludwig Schillings, Bergisch-Gladbach, Federal Republic of Germany (Chap. 4)

Viktor P. Luster, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Chap. 4)

Claus Essig, EDF Septen, Dept. Combustible, Villeurbanne Cedex, France (Chap. 4)

Günter Koch, Kernforschungszentrum Karlsruhe, Institut für Heisse Chemie, Karlsruhe, Federal Republic of Germany (Sections 5.1 and 5.4)

Siegfried Träger, Nukem GmbH, Alzenau, Federal Republic of Germany (Section 5.2)

Arthur Max, Nukem GmbH, Alzenau, Federal Republic of Germany (Section 5.2)

Wolf-Dieter Krebs, Siemens AG, Unternehmensbereich KWU, Erlangen, Federal Republic of Germany (Sections 5.3.1 and 5.3.3)

Wolfgang Stoll, (Siemens AG Brennelementwerk Hanau) Hanau, Federal Republic of Germany (Sections 5.3.1 and 5.3.3)

Werner Heit, Nukem GmbH, Hanau, Federal Republic of Germany (Section 5.3.2)

Ernst Warnecke, Bundesamt für Strahlenschutz, Salzgitter, Federal Republic of Germany (Section 5.5)

Peter Brennecke, Bundesamt für Strahlenschutz, Salzgitter, Federal Republic of Germany (Section 5.5)

Erich Merz, Forschungszentrum Jülich, KFA, Jülich, Federal Republic of Germany (Section 5.6)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a17_589

UWE SCHUMACHER, Universität Stuttgart, Institut für Plasmaforschung, Stuttgart, Federal Republic of Germany (Chap. 6)

HANS HEROLD, Universität Stuttgart, Institut für Plasmaforschung, Stuttgart, Federal Republic of Germany (Chap. 6)

Abbreviations and symbols used in this article

Abbreviations for reactors

AGR	advanced gas-cooled reactor
ATR	advanced thermal reactor
AVR	Arbeitsgemeinschaft Versuchsreaktor
BWR	boiling-water reactor
CANDU	Canadian heavy-water-cooled pressure tube reactor
EFR	European fast reactor
FBR	fast breeder reactor
GGR	gas-cooled graphite-moderated reactor
HTR	high-temperature reactor
HWR	heavy-water reactor
LMFBR	liquid metal fast breeder reactor
LWGR	graphite-moderated light-water-cooled reactor
LWR	light-water reactor
MZFR	Mehrzweck-Forschungsreaktor (multipurpose research reactor)
PHWR	pressurized vessel heavy-water reactor
PWR	pressurized-water reactor
SGHWR	steam-generating heavy-water reactor
SPX	Superphénix
THTR	thorium high-temperature reactor

Further abbreviations

AECL	Atomic Energy of Canada Limited
AISI	American Iron and Steel Institute
ALARA	as low as reasonably achievable
ASME	American Society of Mechanical Engineers
AVT	all volatile treatment
BNFL	British Nuclear Fuel
BR	conversion or breeding rate
COGEMA	Compagnie Générale des Matières Nucléaires
CRDM	control rod drive mechanism
CRT	cathode ray tube
DNB	departure from nucleate boiling
EAR-I	estimated additional resources, cat I
EAR-II	estimated additional resources, cat II
ECL	extra low carbon steel
ECP	electrochemical potential
FIFA	fissions per initial fissile atoms
FIMA	fissions per initial (heavy) metal atoms
hm	heavy metal
HAW	high active waste
HEU	highly enriched uranium and thorium
IAEA	International Atomic Energy Agency
ICF	inertial confinement fusion
ICRP	International Commission on Radiological Protection
IHX	intermediate heat exchanger
I&C	instrumentation and control
ISR	Institut für Sicherheitsforschung und Reaktortechnik
IUREP	international uranium resources evaluation project
KTA	Kerntechnischer Ausschuß
LAW	low active waste
LEU	low-enriched uranium
LPRM	local power range monitoring system
MAW	medium active waste
MCA	maximum credible accident
MOX	mixed oxide
PCRV	prestressed concrete reactor vessel
PUREX	plutonium, uranium, reduction, extraction
RAR	reasonably assured resources
RCC-M	Règles de Conception et de Construction, Part "Mechanical components of PWR nuclear islands"
RHR	residual heat removal
RPV	reactor pressure vessel
SR	speculative resources
StrlSchV	Strahlenschutz-Verordnung
TIG	tungsten-inert gas
WOCA	world outside centrally-planned economic areas

Symbols

A	mass ratio of nucleus to neutron
B	burnup
d_{q}	phase-space element
D	energy dose, distribution coefficient
$\dot{D}$	energy dose rate
D	extent of damage
D_{h}	hydraulic cooling channel diameter
D_{q}	dose equivalent
$\dot{D}_{\mathrm{q}}$	equivalent dose rate
DF	decontamination factor
E	modulus of elasticity
F_{Q}	form factor
f	plant utilization
I	ion dose
k_{eff}	neutron multiplication factor
k_{F}	specific fuel cost
k_{p}	specific cost of the plant
K	activity concentration
L	leakage rate
M_{F}	fuel mass
N	particle concentration
n	neutron density
p_{st}	steam pressure
P_{D}	probability of damage
P_{e}	electrical output
P_{p}	pump or blower power
P_{th}	thermal output
q	quality factor
q'	specific fuel rod power
q''	heat flux density
q'''	power density
q'''_{max}	permissible power density
q_{sp}	average specific power
r	radius
R	reaction rate
R_{D}	risk of damage
T_{c}	coolant warming up span in the reactor
T_{o}	maximum (central) fuel power
T_{st}	steam temperature
V_{c}	flow rate of the coolant
x	distance
α	heat transfer coefficient
β	fraction of delayed neutrons
δ	ultimate elongation
ε_{a}	macroscopic efficient thermal neutron cross-section
η	thermal efficiency factor
η_{c}	viscosity of the coolant
λ	mean free path
λ_{D}	radioactive disintegration constant
λ_{C}	thermal conductivity
ν	neutron velocity
ξ	logarithmic mean energy loss
Ω	solid angle
ϱ_{c}	density of the coolant
Σ_{f}	macroscopic fission cross-section
σ	cross-section; strength values
τ	half-life
Φ	neutron flux density, neutron flux
χ	energy per fission

1. Fundamentals

1.1. Introduction

Early in 1939, Otto Hahn and Fritz Strassmann published their discovery of nuclear fission in the uranium isotope ^{235}U [1]. Their disclosure marked the beginning of nuclear technology [2], whose goal is to utilize the energy liberated in this process.

The fission of an atomic nucleus is caused by penetration of a neutron into the nucleus, which splits into two fission products (i.e., nuclei of lighter atoms). In addition, the fission releases two to three neutrons, on average, along with γ radiation. The kinetic energy of the fission products and neutrons is imparted to the surrounding matter by collisions and – together with the absorbed radiation – heats it. An energy of some 200 MeV (i.e., 3.2×10^{-11} J) is produced in each fission, whereas in conventional energy generation, the formation of a molecule of carbon dioxide liberates about 4.2 eV (i.e., 6.7×10^{-19} J).

For nonfissionable nuclides, addition of a neutron to the nucleus leads to the formation of an isotope with a higher neutron number. In this way, new nuclides – in particular, the fissile materials ^{233}U, ^{239}Pu, and ^{241}Pu – can be artificially "bred."

Utilization of these physical processes occurs in nuclear reactors. Contained in fuel elements, the fission and breeding processes take the form of a chain reaction, which is continuously sustained by the neutrons generated. The heat liberated is used to produce electrical power, and the fissile breeding products serve to extend the operating cycle.

Most nuclides created in nuclear processes are unstable and decay by emitting radioactive radiation. Therefore, reliable means must be provided to keep them out of the biosphere. Nuclear

reactors include highly advanced safety systems, which ensure the containment of radioactive materials in all normal and accident situations. In the same way, the transport of spent fuel assemblies from the reactor, their processing, and the storage of the radioactive waste must all occur under conditions of reliable isolation.

The use of nuclear energy for *power generation* was demonstrated in 1951, when a small reactor at Oak Ridge produced 150 W of electrical power. The first central power station reactor producing 90 MW_e began commercial operation at the end of 1957 in the United States. By 1988, the nuclear industry's contribution to worldwide power output had risen to 17 %; in France, a special case, the nuclear share was 70 %.

The use of nuclear energy is not limited to electrical power. Reactors are used to propel ships [3], especially submarines and icebreakers [4]. The possibility of creating new fissionable nuclides has led to the development of breeder reactors, in which an excess of fissile material can be recovered [5–7]. High-temperature gas- cooled reactors, which heat their coolant to 950 °C, represent a new energy source for chemical processes in the field of high temperature, and for the synthesis of liquid fuels [8–10]. Heat reactors are being devised to supply district heating networks in both large cities and small communities [11], [12]. When located close to these heat consumers, the reactors must be equipped with advanced safety features, and, beyond that, must be fully accepted by the population.

Development work at present is concentrated on reactors employing much-simplified technology for dealing with accidents. Smaller reactor units, up to 600 MW of thermal power, can be designed to stabilize themselves naturally after destabilization, with no possibility of radioactivity escaping [13], [14]. Larger power plants can be assembled from such smaller reactors in modular fashion.

Energy demand worldwide is expected to increase substantially. The consumption of fossil fuels must, however, be limited on ecological grounds. These constraints make it desirable to satisfy the increasing demand for energy through the utilization and further development of nuclear power.

1.2. Design of a Nuclear Reactor

Nuclear power plants use the heat liberated by fission processes for steam generation. The steam drives a turbine generator, which produces electricity. In contrast to power plants using fossil fuels, the heat generation process must be strictly isolated from the environment.

The components essential for the operation and safety of nuclear reactors are illustrated in Figure 1 for a pressurized-water reactor [15–17]. The reactor core is made up of a number of *fuel assemblies*. They consist of an array of fuel rods, which are metallic tubes filled with a stack of uranium dioxide (UO_2) fuel pellets. Here, the heat-producing nuclear processes occur. Cooling water under a pressure of 16.0 MPa flows around the fuel rods and carries the heat to the steam generator.

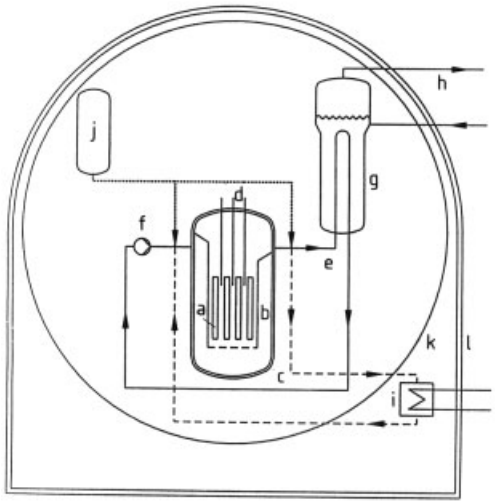

Figure 1. Schematic diagram of a pressurized-water reactor a) Reactor core; b) Reflector; c) Pressure vessel; d) Control rods; e) Main coolant pipe; f) Main coolant pump; g) Steam generator; h) Piping to turbine; i) Decay power removal system; j) Accumulator (borated water); k) Safety vessel; l) Reinforced-concrete containment

The water simultaneously functions as a *moderator*. Neutrons are slowed (moderated) by elastic collisions with nuclei of the water molecules, thus coming to thermal equilibrium with the scattering nuclei (see Section 1.3.5). Outside the reactor core, water serves as a *reflector*, deflecting part of the escaping neutrons back into the core by scattering processes.

To regulate or terminate the chain reaction, *control rods* are moved between the fuel rods. The control rods contain a strong absorber of neutrons, for example, boron. In addition, neutron-absorbing boric acid can be injected into the water.

When the reactor is shut down, heat production in the fuel rods does not stop completely. Decay power output results from the decay of unstable nuclides, especially fission products. To prevent overheating of the fuel rods, a separate cooling system is installed for removal of the decay power. The reactor equipment includes several independent water supply systems, which ensure that the fuel assemblies will remain covered with water and cooled during accidents.

The reactor and its components are surrounded by steel shell containment and a reinforced-concrete containment building in which a slightly subatmospheric pressure is maintained. These structures form an external barrier against the escape of radioactive nuclides during operation or in the case of accident. At the same time, they afford protection against outside events, such as the impact of an aircraft.

The components named relate to the operation and safety of the reactor. They are also present, although in modified form, in other reactor types that have come into use (Table 1).

Modular Reactor Design. The purpose of developing small modular reactors is to simplify engineering design. In these reactors, active measures for ensuring the safety of the reactor in accidents are replaced by passive measures.

In *modular liquid-cooled reactors*, the entire primary coolant circuit including the heat exchanger is located inside the reactor pressure vessel. A containment vessel surrounds the pressure vessel closely enough for the fuel assemblies to remain covered by coolant, even in the event of a leak. Emergency coolant injection is thus superfluous, and decay power is carried away by natural circulation.

In *modular high-temperature reactors*, decay power is absorbed in the graphite of the fuel elements, carried to the pressure vessel wall by radiation and conduction, and from there removed to the surroundings. In this way the temperature of the fuel elements is limited and does not reach a region where fission products can diffuse through the fuel cladding. Even in accidents involving loss of coolant from the reactor, the fission products remain confined in the fuel elements.

1.3. Physical Processes [18–26]

Reactor operation involves closely associated processes of neutron physics, burnup, heat transfer, and fuel handling. For an existing reactor, these steps are monitored and measured. In contrast, facility design, construction, and licensing of a reactor are based on computer simulations that build up the history from repeated calculations of the many partial processes involved. The simulation also offers insight into physical processes that are not accessible for experimental study.

1.3.1. Structure of Atomic Nuclei

The atomic nucleus is made up of about equal numbers of protons and neutrons (both included in the term "nucleons"). Heavy nuclei contain more neutrons than protons. Neutrons and protons are held together by nuclear forces. The analogy with a drop of water led Gamow [27] to introduce the " liquid drop model," which describes the increase in nuclear volume and, approximately, the increase in nuclear binding energy with nucleon number.

Five types of internal forces are distinguished: (1) an extremely short-range (2×10^{-13} cm) attractive term; (2) a term considering the assymmetric distribution of forces at the surface, analogous to surface tension; (3) repulsive forces attributed to Coulombic interaction between protons; (4) weakening of bonds due to unequal proton and neutron numbers; and (5) saturation of nucleon spin states. The binding energy of a nucleon in the nucleus is approximately 8 MeV.

1.3.2. Reactions with Neutrons [28]

The liquid drop model offers an easy way to visualize nuclear fission. The entry of a neutron

Table 1. Reactor types

Type	Abbreviation	Fuel	Moderator	Coolant	Refueling	Burnup, MW · d/kg
Pressurized-water reactor	PWR	LEU [a]	H_2O	H_2O	annual	36
Boiling-water reactor	BWR	LEU [a]	H_2O	H_2O	annual	30
Heavy-water reactor	CANDU	$U^b{}_{nat}$	D_2O	D_2O	continuous	7
Gas- cooled graphite-moderated reactor	GGR	$U^b{}_{nat}$	graphite	CO_2	continuous	7
Advanced gas- cooled reactor	AGR	LEU [a]	graphite	CO_2	continuous	24
Water- cooled graphite-moderated channel	RBMK	LEU [a]	graphite	H_2O	continuous	20
High-temperature reactor	HTR	HEU [c], LEU [a]	graphite	He	continuous	100
Fast breeder reactor	FBR	^{239}Pu, ^{238}U		Na	annual	70

[a] LEU = low enriched uranium.
[b] U_{nat} = natural uranium containing 0.711 % ^{235}U.
[c] HEU = highly enriched uranium and thorium.

adds energy to the nucleus, which becomes deformed and executes surface vibrations (Fig. 2). If the excitation energy is greater than a threshold value, the Coulombic repulsion exceeds the restoring surface tension, and the nucleus splits. If the threshold energy is not reached, the capture of a neutron leads to the formation of the next higher nuclide and the emission of γ radiation [an "(n, γ) process"] (see Tables 2 and 3).

Table 2. Energy release, in MeV, on fission of ^{235}U

Kinetic energy of fission products	169.6
Kinetic energy of fission neutrons	4.8
Prompt γ radiation	7.0
γ Radiation from decay of fission products	6.3
β Radiation from decay of fission products	6.4
Neutrinos	8.7

Figure 2. Nuclear fission in the liquid drop model

If a nucleus with an odd number of neutrons captures a neutron, pair saturation of the neutron spin occurs, resulting in a slight additional increase in excitation energy. This is why odd-neutron nuclides (e.g., ^{233}U, ^{235}U, ^{239}Pu, ^{241}Pu, so-called fissile materials) undergo *fission* upon the addition of neutrons with low kinetic energy (i.e., thermal neutrons). Even-neutron nuclides, such as ^{232}Th and ^{238}U (so-called fissionable materials), mainly exhibit *neutron capture* and are likely to undergo fission only in the case of high-energy (fast) neutrons.

Along with fission and capture, *collision processes* are important. The slowing-down of neutrons by the nuclei of the moderator (H_2O, D_2O, C, Be) involves *elastic scattering*, in which the neutron is again free after its capture and the nucleus remains unexcited. In *inelastic scattering*, the neutron escapes after the collision but leaves the nucleus in an excited state, which decays by the emission of γ quanta. In *inelastic collisions*, a fast neutron knocks one or two additional neutrons out of the nucleus [(n, n′) and (n, 2 n) reactions]. In certain nuclides, high-energy gamma quanta can induce the emission of a neutron [(γ, n) processes].

1.3.3. Decay

A nuclear reaction leaves the nucleus in an unstable state, which decays with a certain half-life.

The fission product nuclei (daughter nuclides) generated in a nuclear fission process have a high excitation energy. In some cases, in a process with a half-life between 0.2 and 55 s, the daughter nuclide emits a *neutron* and changes to the nuclide with the next lower neutron number. These "delayed" neutrons represent only a fraction of about 0.2 – 0.7 %, compared to the neutrons emitted "promptly" upon fission, but the delay makes a significant contribution to stabilizing the chain reaction (see Section 1.3.7).

In α *decay*, a helium nucleus is emitted. The proton and neutron numbers are each reduced

Table 3. Nuclear fission process

Property of the nuclides	^{233}U	^{235}U	^{239}Pu	^{241}Pu	^{232}Th ∗	^{238}U ∗
Neutrons produced, ν	2.50	2.44	2.87	2.93	2.09	2.55
Available energy, MeV	190.3	194.0	199.9	201.7	184.4	194.4
Fission probability on absorption of a neutron	0.92	0.86	0.73	0.74	0.44	0.86

∗ Fission by fast neutrons at 1.5 MeV.

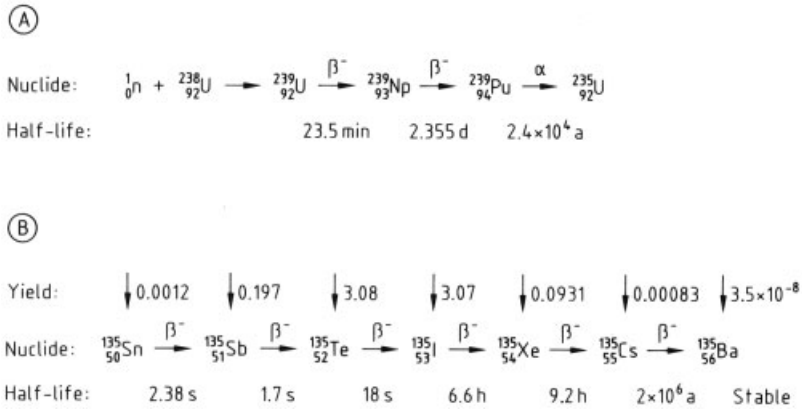

Figure 3. Examples of decay chains
A) In breeding; B) Isobaric fission products with atomic mass 135 (yield = nuclei produced per 100 fissions in ^{235}U)

by two. Alpha radiation is strongly ionizing but it has a short range.

An internal neutron – proton transformation is involved in β *decay*. The β emitters are classified as β^- (electron) and β^+ (positron) emitters.

γ *Emission* represents high-energy electromagnetic radiation in the 1-MeV region. Because of its long range, γ radiation requires thick shielding.

Decays often occur in succession until a stable state is reached. One such chain begins with the capture of a neutron by ^{238}U and ends in the formation of the fissile nuclide ^{239}Pu (Fig. 3). Most fission products decay in a similar way until stable nuclides are formed. Because of their radiotoxic properties, active reactor materials must be kept shielded and isolated from the biosphere.

1.3.4. Cross-Sections

The probability that a neutron approaching a nucleus will react with it is expressed in terms of the cross-section σ. The cross-section of a nucleus represents the area, normal to the direction of the incident neutron, that the neutron must hit to react with the nucleus. The larger the cross-section, the higher is the probability of a neutron – nucleus reaction.

Cross-sections have the dimensions of area and the unit barn: 1 barn = 10^{-24} cm^2. The cross-section depends strongly on the energy of the incident neutron and often exhibits selective maxima that correspond to resonance absorption or scattering at certain neutron energies (analogous to light absorption by atomic electron shells; see Figs. 4 and 5 [29]). The reaction cross-section may exceed the geometric cross-section of the nucleus by a factor of more than 10^6.

Cross-sections are identified by the nuclear reaction in question: σ_f for fission processes, σ_c for neutron capture, σ_s for scattering. The absorption cross-section is $\sigma_a = \sigma_f + \sigma_c$, and the total cross-section is $\sigma_t = \sigma_a + \sigma_s$.

In a medium with N nuclei per cubic centimeter, the product

$$N \cdot \sigma = \Sigma$$

is the macroscopic cross-section. This equation applies to each of the various nuclides and their reactions and also to their sum. In what follows,

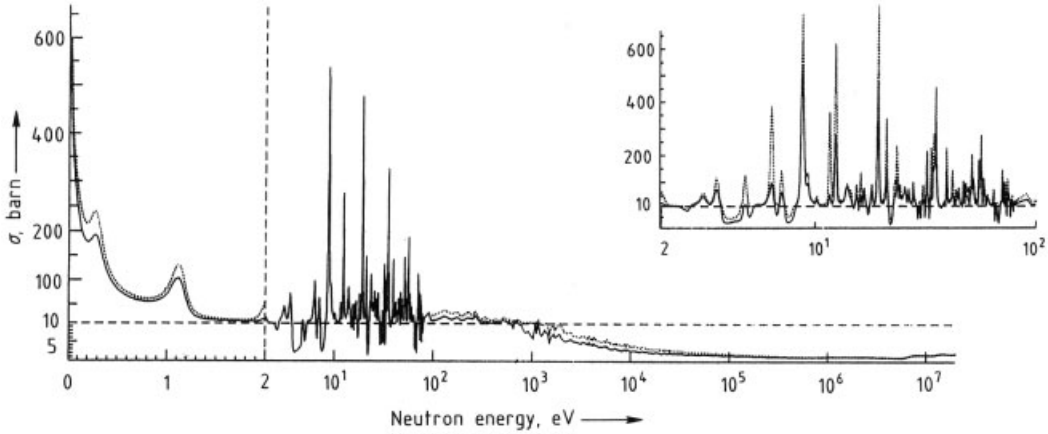

Figure 4. Microscopic cross-sections for ^{235}U at 1273 K
—— σ_f; $\cdots$ $\sigma_a = \sigma_f + \sigma_c$

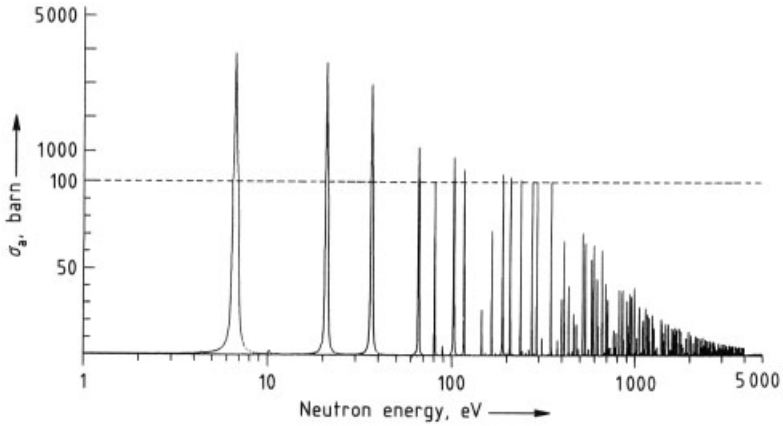

Figure 5. Absorption cross-section σ_a of ^{238}U in the resonance region (1273 K)

Σ is used without identifying information; for example, $\Sigma_f(^{235}U) \rightarrow \Sigma$. The probability that a neutron will travel a distance x without reacting is given by $e^{-\Sigma x}$. The quantity

$$\frac{1}{\Sigma} = L$$

is the mean free path. Its typical value is a few centimeters; near a strong resonance, it will be a fraction of a millimeter.

Absorption resonances are particularly marked in the case of ^{238}U. The energy region between 4 and 4000 eV contains 416 resonances. Out of a continuous neutron energy spectrum, the neutrons with energies near the resonances are selectively absorbed (Fig. 6). The fraction of such neutrons is increased by the thermal motion of the nuclei: since the nuclei move with and against the direction of neutron flight, neutrons whose actual kinetic energies lie outside the natural width of a resonance are also absorbed. This phenomenon is analogous to the Doppler effect as known in optics and acoustics. For computational purposes, the effect is described in terms of temperature-dependent broadening of the resonances [30].

The temperature-dependent resonance absorption of ^{238}U is very important for nuclear reactor control. In normal operation, this effect stabilizes the chain reaction. In accidents where the fuel is heated, the enhanced neutron absorption causes the chain reaction to decay because

of insufficient heat removal. This mode of reactor control through the fuel temperature is independent of the operability of reactor equipment. It is an inherent property of the reactor, because it is based solely on natural laws.

1.3.5. Moderation of Neutrons

Nuclear fission produces neutrons with energies between 0.1 and 10 MeV. These are slowed by elastic collisions with light moderator nuclei (e.g., H_2O, D_2O, C) until they reach thermal equilibrium with them (see Table 4). In the thermal energy range, the neutrons experience ca. 30 – 200 collisions before they are captured or initiate a fission. The residence time at thermal energies is ca. 10^{-3} s, which is the major part of the generation time of a neutron.

The energy given up by a neutron in a collision is determined under a rigid-sphere model. The energy loss is a maximum in a central collision; if the starting energy is E_1, it leads to the energy:

$$E_2 = E_1 \left(\frac{A-1}{A+1}\right)^2$$

where A is the mass ratio of nucleus to neutron. When colliding with a hydrogen nucleus ($A = 1$), a neutron can give up all of its energy in a single collision; this is not possible with heavier nuclei. On the average, the energy E_2 after collision is a constant fraction of the neutron energy E_1 before collision. The logarithmic mean energy loss per collision $\xi = \ln(E_1/E_2)$ is:

$$\xi = 1 + \frac{\alpha}{1-\alpha}\ln\alpha \quad \text{where} \quad \alpha = \left(\frac{A-1}{A+1}\right)^2$$

The logarithmic mean energy loss is a constant for each moderator material.

If no light moderator nuclei are available, the neutrons are moderated in smaller steps. In the fast breeder reactor (FBR), for example, neutrons are scattered on ^{23}Na nuclei. Most of the neutron absorption and fission events in this reactor take place at fast neutron energies. This fact is the reason the FBR is referred to as a "fast" reactor. In "thermal" reactors, on the other hand, fission is mainly initiated by neutrons that have reached the energy range of thermal equilibrium.

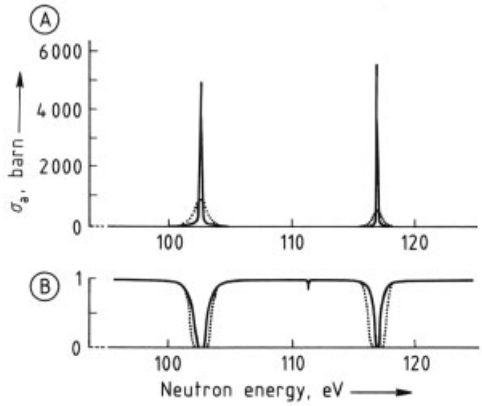

Figure 6. Temperature-dependent neutron absorption at resonances of ^{238}U
A) Absorption cross-section at two resonances; B) Probability of traversing a LWR fuel rod without absorption in ^{238}U
—— 0 K; · · · · 1273 K

1.3.6. Neutron Flux [31–33]

The distribution of neutrons in the reactor is characterized by the neutron density n. The product of neutron density and velocity ν is the neutron flux density ϕ, usually called "neutron flux"

$$\phi = \nu \cdot n \tag{1}$$

This quantity tells how many neutrons per second pass through a surface element of unit area in all directions. It is equal in magnitude to the sum of the distances traveled by all neutrons from a unit volume in a second.

The ratio of this distance ϕ to the mean free path Λ of neutrons in a given medium

$$R = \frac{\phi}{\Lambda} = \Sigma \cdot \phi$$

is the number of reactions per second in the unit volume, that is, the reaction rate R. Reaction rates can be determined separately for various types of nuclear processes, for various materials present in a reactor, and for various positions in space. Determination of reaction rates is a central problem in reactor physics. The macroscopic cross-sections Σ required for this purpose characterize the properties of the materials and their spatial distribution in the reactor; their values are thus set by reactor design. The neutron flux

Table 4. Number of collisions with moderator nuclei ∗

Range of collision	Energy	LWR	HTR	FBR
Thermalization of fast neutrons	2 MeV → 1 eV	16	92	171
In region of ^{238}U resonances	4000 → 4 eV	8	44	82
At thermal energy	$E < 2$ eV	50	200	

∗ LWR = light-water reactor; HTR = high-temperature reactor; FBR = fast breeder reactor.

ϕ describes the distribution of neutrons that can initiate reactions, and this quantity must be determined by computation.

Calculation of Neutron Flux. Neutron flux can be calculated in two ways. The principal way is to write a *balance equation* for the neutron density and solve it numerically. Another method is based on *direct calculation of the fate of individual neutrons* in the real reactor configuration. The point of origin of the neutron is selected on the basis of a statistically generated random number. Analogously, random numbers are applied to select the direction of flight, its energy within the fission spectrum, and all other quantities having the sense of probabilities. In this way, a flight path can be determined for the neutron from birth through the many scattering processes to absorption. When a great number of individual histories have been simulated by a computer, the distribution of the neutron flux is found by averaging. The procedure is called the "Monte Carlo" technique; when detailed information is given on the geometric configuration of the reactor, it can yield very reliable results.

The *balance equation* for the neutron flux is formulated in a similar way to the balance equation for the number of gas molecules per unit volume, as introduced by BOLTZMANN in the kinetic theory of gases. In a detailed treatment, the neutron density, the flux, and the reaction rate are functions of position, velocity, and time. The expression

$$n\left(\vec{r}, E, \vec{\Omega}, t\right) \mathrm{d}V \, \mathrm{d}E \, \mathrm{d}\vec{\Omega}$$

gives the number of neutrons located in the volume element dV at position $\vec{r}$ at time t, having kinetic energy between E and $E + \mathrm{d}E$, and moving in a direction lying in the solid-angle element $\mathrm{d}\vec{\Omega}$ at solid angle $\vec{\Omega}$.

These variables represent coordinates in phase space. The reaction rate

$$R\left(\vec{r}, E, \vec{\Omega}, t\right) \mathrm{d}q = \Sigma\left(\vec{r}, E\right) \cdot \phi\left(\vec{r}, E, \vec{\Omega}, t\right) \mathrm{d}q$$

is the number of reactions per second in the phase-space element $\mathrm{d}q = \mathrm{d}V \, \mathrm{d}E \, \mathrm{d}\vec{\Omega}$.

By analogy with the kinetic theory of gases, the Boltzmann equation for neutron flux is derived from the balance equation for reaction rates. In terms of neutron flux it is given as:

$$\frac{1}{\nu} \frac{\partial \phi\left(\vec{r}, E, \vec{\Omega}, t\right)}{\partial t} \mathrm{d}q = \text{Gains} - \text{Losses} \qquad (2)$$

using the definition of ϕ in Equation (1). Losses include all reaction rates that lead out of the phase-space element dq: absorption, scattering into other phase-space elements, and neutrons moving out of the spatial volume element dV. Gains are in-scattering events from other phase-space elements, newly produced prompt fission neutrons, and delayed neutrons.

For a real reactor, because of its complex structure, this equation cannot be solved numerically. The problem is, therefore, simplified and subdivided in various ways. Normally, integral quantities are formed, and the balance equation is transformed to a balance equation for these quantities. For example,

$$\phi\left(\vec{r}, E, t\right) = \int_{\vec{\Omega}} \phi\left(\vec{r}, E, \vec{\Omega}, t\right) \mathrm{d}\vec{\Omega}$$

represents the neutron flux integrated over all directions of flight. For the energy dependence, an expansion in integral values over wide subintervals or "energy groups" ΔE_{g} is usual:

$$\phi_{\mathrm{g}}\left(\vec{r}, t\right) = \int_{E_{\mathrm{g}}}^{E_{\mathrm{g}} + \Delta E_{\mathrm{g}}} \phi\left(\vec{r}, E, t\right) \mathrm{d}E$$

The position dependence can also be represented by integral values over subregions J formed by volumes ΔV_J.

Most commonly, the quantity of interest is the neutron flux when the reactor state is constant over time. In this case, the problem can be simplified to solving the steady-state equation $\partial\phi/\partial t = 0$.

In practice, the calculation is often performed in two steps. First, a representative cell is formed for each of various local subregions of the reactor. The cell represents the geometry of a fuel rod and the surrounding material. The neutron flux $\phi\left(\vec{r}, E\right)$ is calculated for each cell, and this flux function is used in forming an average for the cross-sections of the various materials located within the cell. These are found by integrating the reaction rates to

$$\sigma_{\mathrm{J,g}} = \frac{\int\limits_{\Delta V_{\mathrm{J}}} \mathrm{d}V \int\limits_{\Delta E_{\mathrm{g}}} \mathrm{d}E \cdot \sigma\left(\vec{r}, E\right) \cdot \phi\left(\vec{r}, E\right)}{\int\limits_{\Delta V_{\mathrm{J}}} \mathrm{d}V \int\limits_{\Delta E_{\mathrm{g}}} \mathrm{d}E \cdot \phi\left(\vec{r}, E\right)} \quad (3)$$

where the denominator is the integrated neutron flux.

Second, the averaged cross-sections of the subregion J are assigned to the respective positions $\vec{R}$ in the reactor $[\sigma_{\mathrm{J,g}} \rightarrow \sigma_{\mathrm{g}}(\vec{R})]$ and used for calculating the global distribution of the neutron flux. Under some simplifying assumptions, Equation (2) takes the form of a diffusion equation, whose solution gives the neutron-flux distribution $\phi_{\mathrm{g}}(\vec{R})$ by energy groups g and over the entire reactor volume. The integral neutron flux is

$$\phi = \sum_{\text{groups g}} \int\limits_{\text{reactor}} \phi_{\mathrm{g}}\left(\vec{R}\right) \cdot \mathrm{d}V \quad (4)$$

This flux is normalized such that the integral fission rate R_{f} formed with it (see Section 1.3.8) equals the power output of the reactor. In real reactor operation, this corresponds to adjusting the neutron flux with control devices to obtain the desired heat output.

At thermal energies, the neutron-flux function is determined by scattering on moderator nuclei. The thermal motion of these nuclei influences the neutron spectrum (Fig. 7), and for this reason the mean cross-sections from Equation (3) depend on the moderator temperature (Table 5). Like the Doppler effect in resonance absorption, this temperature dependence is an inherent property that contributes to automatic control of the reactor. However, the averaged cross-sections for both neutron capture $\sigma_{\mathrm{c,J,g}}$ and fission $\sigma_{\mathrm{f,J,g}}$, (i.e., the rate of neutron loss and the rate of new neutron production) are also temperature dependent. In some reactor concepts, the increase in fission rate with increasing temperature is dominant [34]; these reactors have a positive moderator temperature coefficient. In most reactors, enhanced neutron capture is the dominant effect, and an increase in temperature results in automatic shutdown.

Table 5. Mean thermal absorption cross-sections $\sigma_{\mathrm{a,J,g}}$, barn (averaged with neutron flux in light-water reactor fuel assembly from Fig. 7).

Temperature, °C	^{235}U	^{135}Xe	^{10}B
20	299	13.4×10^6	2300
300	257	11.6×10^6	1950

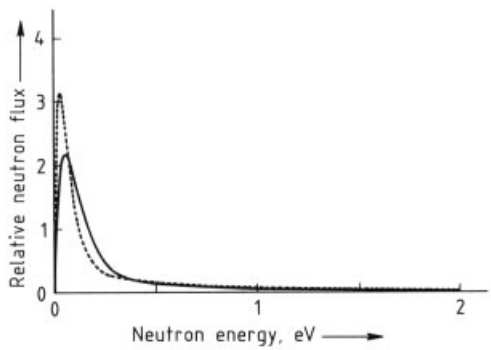

Figure 7. Thermal neutron flux in the LWR fuel rod —— 300 °C; · · · · 20 °C

1.3.7. Time Dependence of Neutron Flux [35], [36]

Reactor control actions and accidental situations cause a change in neutron flux over time. In computational terms, this process is described by solving the time-dependent balance equation (Eq. 2) or the corresponding time-dependent diffusion equation.

With regard to time, the neutron gain due to fissions must be divided into a prompt and a delayed fraction, as described in Section 1.3.3. The fraction of prompt neutrons is between 99.3 and

99.8 %, and their generation time is ca. 10^{-3} s. The delayed neutrons, with a fraction of 0.2 – 0.7 %, have a mean generation time of around 13 s (Table 6). It is also important that the neutron loss at the absorption resonances increases with increasing temperature (Section 1.3.4).

Table 6. Delayed neutrons in six decay groups ∗

Group	^{235}U		^{239}Pu	
i	τ_i, s	β_i, %	τ_i, s	β_i, %
1	0.18	0.018	0.22	0.008
2	0.50	0.088	0.55	0.023
3	2.23	0.281	2.09	0.074
4	5.98	0.130	5.17	0.049
5	21.84	0.140	22.29	0.063
6	54.49	0.026	53.73	0.009
Total β		0.690		0.226
Mean τ	8.81		9.53	

∗ τ = half-life, β = fraction per fission neutron.

In normal reactor operation, the neutron flux is at equilibrium and $\partial\phi/\partial t = 0$. This state is perturbed by a change in neutron losses, for example, when the control rods are moved or the coolant flow rate varies. According to Equation (2), a reduction in neutron losses brings about an increase in the neutron flux over time, $\partial\phi/\partial t > 0$. As a consequence, the neutron gain increases in the subsequent neutron generation, because the rate of production is proportional to the flux. The incremental increase is very different for the prompt and delayed neutrons because of their different generation times.

If the decrease in losses is *high*, the increase in flux is governed by the generation time of the prompt neutrons. The flux increases by orders of magnitude in a fraction of a second. This condition is called "prompt criticality"; it cannot be controlled by technical measures. However, at the same time, any increase in flux is associated with an increase in reactor output, that is, heat production in the fuel assemblies. If the temperature increases by 100 °C, neutron absorption increases by 0.2 – 0.8 %, depending on the reactor type; in this way an excursion can be controlled by the rise in fuel-assembly temperature.

However, if the decrease in neutron losses is *low*, i.e., less than the fraction β of delayed neutrons in the rate of neutron gain, the increase in flux is governed by the generation time of delayed neutrons. The increase takes place slowly, over minutes, and can be detected and counteracted by the reactor control system.

A knowledge of the kinetic processes is essential for the safe design of a reactor. Computational treatment includes the simulation of individual processes and the linking of these over time.

1.3.8. Reaction Rates

The various reaction rates in the reactor are determined with the aid of the normalized neutron flux. They can be calculated as integral quantities or separately per energy group, subregion of the reactor, and nuclide. The reaction rates give an insight into what is happening in the reactor in terms of neutron physics. They also enable computer modeling of the variations in nuclide concentrations during power generation.

For example, the fission rate of the nuclide ^{235}U is:

$$R_{\mathrm{f}}\,(235\mathrm{U}) = \sum_{\text{groups g}} \int_{\text{reactor}} \Sigma_{\mathrm{f,g}}\left(235\mathrm{U},\ \vec{R}\right) \cdot \Phi_{\mathrm{g}}\left(\vec{R}\right) \cdot \mathrm{d}V$$

The sum of the fission rates for all nuclides gives the integral fission rate R_{f}. The ratio of this quantity to the integral flux Φ from Equation (4) defines a mean macroscopic fission cross-section $\Sigma_{\mathrm{f}} = R_{\mathrm{f}}/\Phi$, so that the integral fission rate can be written as $R_{\mathrm{f}} = \Sigma_{\mathrm{f}}\Phi$. The rates of other nuclear processes are calculated in a similar way.

A central reactor characteristic is the *neutron multiplication factor* k_{eff} :

$$k_{\mathrm{eff}} = \frac{\nu\Sigma_{\mathrm{f}} \cdot F}{\Sigma_{\mathrm{a}} \cdot F + L} = \frac{\text{Rate of production of neutrons}}{\text{Rate of loss of neutrons}}$$

This is the ratio of the integral production rate to the rate of loss. The latter is the integral absorption rate plus the leakage rate L of neutrons escaping from the reactor. For a steady-state chain reaction in which the number of neutrons neither increases nor decreases, $k_{\mathrm{eff}} = 1$ (criticality condition).

Since the absorption of a neutron is accompanied by a simultaneous nuclear transformation, the reaction rates also describe how the nuclide concentrations vary. The absorption rate

$\Sigma_a(\text{fiss})\Phi$ describes the consumption of fissile nuclides per second. The production of new fissile nuclides is given by the absorption rate of fertile nuclides (^{232}Th, ^{234}U, ^{238}U, ^{240}Pu) or the decay rate of intermediate product nuclides (^{233}Pa, ^{239}Np; see Fig. 3). The ratio:

$$\text{BR} = \frac{\text{Production rate of fissile nuclides}}{\text{Rate of loss of fissile nuclides}}$$

characterizes the fissile balance of the reactor. In most thermal reactors, BR is in the range 0.5 – 0.8 and is called the *conversion ratio*. In reactors with BR ≥ 1, chiefly fast reactors, BR is called the *breeding ratio* and the reactor is referred to as a breeder reactor.

The physical events in a reactor can be understood by interpreting the reaction rates (Table 7). Approximately half the neutrons are absorbed in the fissile nuclides to sustain the chain reaction. About 30 % are absorbed in fertile material, leading to the formation of new fissile material. About 20 % represent unavoidable neutron losses in the reactor materials (in fission products, moderator, control means, fuel cladding, reflector). The amounts of fissile and fertile materials are fixed so that the relative loss rates add up to 100 %, that is, the criticality condition is satisfied.

The breeding effect can be enhanced by a particularly favorable programming of the neutron balance. This is achieved by closing the fuel cycle, that is, reprocessing unloaded fuel assemblies and reusing fissile material. Thermal reactors benefit from employing ^{232}Th as the fertile material and reprocessing the fuel assemblies frequently [37]. In this way, the favorable fissile nuclide ^{233}U is allowed to accumulate (Table 3), and neutron absorption in the fission products is reduced. Fast reactors are loaded with plutonium as fissile material. Plutonium has good fission properties at high neutron energies. In this energy range, parasitic absorption by reactor materials is slight, and ca. 14 % of the fissions are fast fissions in ^{238}U. Thus, the consumption of fissile nuclides (i.e., the loss rate) is accordingly reduced. Leakage losses are reduced by the use of a fertile blanket in the reflector region.

1.3.9. Burnup and Control

In the narrow sense, burnup is the energy produced in the fuel divided by the mass of heavy metals (fissile and fertile material) used. Fissions of nuclides bred in the reactor are included. Burnup is expressed in units of MW · d/kg$_{\text{hm}}$ (hm = heavy metal). The total energy content of 1 kg of heavy metals is 921 MW · d/kg (available energy in ^{235}U). Various reactor designs provide for 0.7 – 10 % burn-up (i.e., fissioning) of the heavy metals (see Table 1).

More broadly, burnup embraces all nuclear transformations in the reactor: fission, neutron capture, and decay. Some 820 fission products and about 100 heavy-metal nuclides are formed and transformed. Many of these influence physical processes in the reactor and are important for safety. If the spatial distribution of the neutron flux is known, the formation of these nuclides and their decay history in the fuel assemblies can be modeled [38].

For each nuclide, the time dependence of its concentration N is found from the balance equation:

$$\mathrm{d}N/\mathrm{d}t = \text{Rate of production} - \text{Rate of destruction} \tag{5}$$

Through the absorption rate $N\sigma_a\Phi$ or the decay rate λN (λ being the decay constant in s^{-1}), the destruction of nuclides leads to the formation of new ones. Production is due to fission, neutron capture, or the decay of parent nuclides. As a result, the burnup equations (Eq. 5) for the individual nuclides are coupled to one another.

During reactor operation, the fissile content of the fuel assemblies decreases and the content of fission products increases. Some of these fission products have large absorption cross-sections. Two approaches are taken to maintaining the criticality condition $k_{\text{eff}} = 1$: In reactors with *continuous refueling*, old fuel assemblies with high burnup are continuously removed during operation, and fresh assemblies are loaded in their place. In reactors with *periodic refueling*, the fissile content is highest at the start of an operating cycle, and absorber material is added. The absorber is gradually withdrawn as burnup increases. Absorber rods, absorber solution in the coolant, and burnable absorbers are used. An example of a burnable absorber (control poison) is the nuclide ^{10}B, which has a large absorption cross-section (Table 5); it is burned up by neutron capture and transformed by an (n, α) process to ^{7}Li, which is virtually nonabsorbing.

Table 7. Neutron balances per fission neutron, in percent, and the conversion or breeding ratio

Neutron balance	Without recycle of fissiles		Closed fuel cycle	
	LWR	HTR	HTR ∗	FBR
Absorption in fissile material	48.0	52.0	45.9	36.5
Fission in other actinides	2.3	0.2	0.1	6.1
Capture in fertile materials	27.7	30.0	39.5	45.3
In other actinides	1.1	1.3	2.8	0.4
In fission products	6.8	7.7	5.4	2.3
In reactor materials and control poison	10.6	1.7	2.1	5.9
Leakage, control poison in reflector	3.5	7.1	4.2	3.5
Total	100.0	100.0	100.0	100.0
Conversion or breeding ratio	0.58	0.56	0.84	1.24
Slowing down to thermal energy	66.1	60.7	61.1	0.0

∗ ^{232}Th as fertile nuclide.

The differing properties of fission products are a factor in real load- changing operation and must be continuously corrected by the control system. The process can be explained for the most important example, the buildup of ^{135}Xe. In equilibrium operation, 98.6 % of this nuclide is produced by the decay of ^{135}I and only a small fraction directly by fission (Fig. 3). Xenon-135 is destroyed chiefly by neutron absorption (because of its large cross-section; see Table 5) and partly by decay. A reduction of reactor power is achieved by a decrease in the neutron flux. As a result, the rate of destruction of xenon decreases, whereas the rate of production by the decay of iodine initially remains unchanged. After a few hours, the iodine concentration slowly declines and the production rate of ^{135}Xe falls off. The relative absorption rate varies along with the xenon concentration (Fig. 8). To maintain the criticality condition $k_{\text{eff}} = 1$, the varying neutron absorption in the transient period must be compensated by withdrawing and later inserting the control rods. In normal operation, the control rods must be inserted partway to maintain a reactivity reserve so that the reactor can be controlled under load-changing conditions.

Computer simulation of reactor operation involves a frequent alternation of neutron-flux and burnup calculations, adjustment of control poison, and insertion and removal of fuel assemblies. In this way, the burnup and isotopic composition of the discharged fuel are determined. The need for new fuel is calculated, and the load history of the fuel assemblies is determined. In addition, the simulation can reproduce all operating states for detailed study so that reactor safety can be ensured.

The explicit calculation of the burnup history, can be restricted to those nuclides that are of marked importance for neutron absorption and the fissionable balance. These include about 15 actinides and some 50 fission products. On the other hand, all nuclides must be calculated when the radioactivity of the irradiated fuel is to be determined.

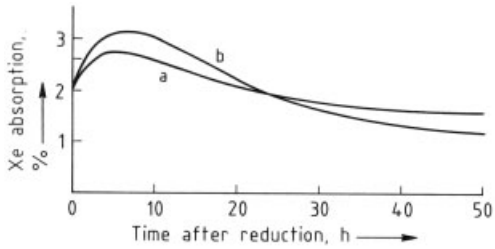

Figure 8. Neutron absorption of ^{135}Xe on a decrease in LWR output (a) to 40 %; (b) to 20 %

1.3.10. Fission Products and Actinides [40]

The threat posed by radioactive waste products is offbalanced by the utility of nuclear energy. However, the isolation of radioactive waste from the environment is an urgent problem that demands a great engineering effort. The concerns involved are reactor safety while the fuel assemblies are in the reactor and waste disposal after they are unloaded. The corresponding effort with fossil energy sources is very slight in comparison because, to a great extent, the waste from combustion processes can simply be discharged to the environment.

Inside the reactor, the containment of fission products, actinides, and activated materials is performed by multiple barriers: the fuel

cladding, the confinement of the coolant loop, and the reactor containments (Fig. 1). The objective of safety engineering is to protect these barriers against attack by mechanical, chemical, and especially thermal action.

On reactor shutdown, whether deliberate or initiated by the control devices, the output of energy from nuclear fission stops, but heat produced by the decay of fission products continues. Averaged over the first hour, this heat output is ca. 2 % of the previous reactor power and it decays slowly (Fig. 9). Integrated over the first 24 h, it matches 11 min of reactor output. This decay power is transferred to the coolant and moderator that surround the fuel assemblies, thereby heating these materials. The decay power must be removed to prevent overheating and failure of the fuel cladding.

In present reactors, decay power removal is accomplished by active technical means that are designed with multiple redundancy. In small reactor units for modular construction, now under development, decay power is removed by natural circulation of the coolant. Here, in the case of a break in the coolant loop (the second of the barriers listed above), decay power can still be removed without the use of active technical means: in liquid-cooled reactor designs by virtue of the enclosing containment vessel; in gas-cooled reactors by conduction in the graphite (see Section 1.2).

After being unloaded from the reactor, the fuel assemblies are placed in sealed containers and held in intermediate storage for some years. This time allows a large fraction of the fission products to decay, generating a decay power of 1 – 10 W/kg_{hm} (Table 8). After this, the fuel is either reprocessed or transferred to an ultimate storage facility. Reprocessing recovers the fissile material; the waste is vitrified, sealed, and also placed in ultimate storage.

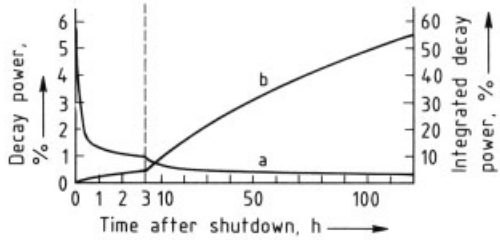

Figure 9. Decay power after reactor shutdown (HTR module)
a) Decay power as a fraction of reactor output in full-load operation; b) Integrated decay power as a fraction of energy produced in 1 h of full-load operation

Table 8. Activity and decay power output of unloaded fuel (referred to 1 kg of light-water reactor fuel)

Activity and decay power	Time after unloading, years										
	0	0.01 (3.6 d)	0.1 (36 d)	1	10	100	1000	10^4	10^5	10^6	10^7
Activity, Ci *											
Fission products	140 000	22 100	10 300	2 300	326	35	0.023	0.022	0.017	0.004	8×10^{-5}
Heavy-metal nuclides	37 000	6 800	150	110	69	6	1.45	0.408	0.038	0.016	0.005
Total	177 000	28 900	10 450	2 410	395	41	1.47	0.430	0.055	0.020	0.005
Decay power output, W	1 780	113	43	10	1.2	0.27	0.046	0.012	0.001	0.0003	0.0001

* 1 Ci = 3.7×10^{10} Bq (see Section 2.3.1).

Most of the fission products decay in the first decades; decay of the long-lived ones, such as ^{137}Cs and ^{90}Sr with half-lives of ca. 30 years, will be completed after a 1000 years. The actinides and their daughter products have comparatively low activities; however, these nuclides require safe confinement over several million years. Thus, ultimate storage facilities are sited, for example, in salt formations that have proved to be stable over geological spans of time. A concept developed in Sweden [40] provides for storage at a depth of 500 m in granite formations. In addition, studies are being conducted on the reuse of actinides in nuclear reactors [41] in order to convert them to short-lived nuclides by neutron-capture reactions. Of course, realization of this concept requires major development in reactor design, reprocessing, and recycling technology.

2. Power Reactors, Survey

2.1. Introduction, Scope of Utilization

Power reactors serve to supply energy. Approximately 200 MeV (3.2×10^{-11} J) of energy is released per nuclear fission (see Chap. 1) [42], [43]. Although this energy is emitted at first as the kinetic energy of the two daughter nuclei (ca. 165 MeV), the kinetic energy of the neutrons generated (ca. 5 MeV), and as β and γ radiation, it is largely converted into heat within the fuel itself. Only a small part of the energy produced (generally < 5 %) leaves the reactor in the form of γ or neutrino radiation.

Today, the heat created in power reactors is used predominantly in nuclear power plants for the production of electrical energy. Other uses include the production of process heat for industry (see Section 3.2.3.7), for the desalination of sea water [44], or for district heating [45] and exploitation of the energy for propulsion purposes (particularly for nuclear-powered ships [44]). A combination of several applications is also possible.

Nuclear-powered ships represent one of the oldest uses of nuclear energy. In fact, the Mark I prototype for the submarine *Nautilus* was put in operation in early 1955. Since then, numerous warships – but only a few civilian ships (NS *Lenin*, NS *Savannah*, NS *Otto Hahn*, NS *Mutsu*,

NS *L. Breschnew*, NS *Sibir*, NS *Rossia*, NS *Taymyr*, and NS *Vaygach*) – have been equipped with nuclear power units. Today old nuclear powered submarine wreckages of the former Soviet Union represent a possible source of radioactive pollution in the polar regions of the Atlantic ocean. Commercially, nuclear-powered ships have not gained acceptance because at the normal driving power of up to 73 500 kW (100 000 metric HP), fuel costs do not compensate for the comparatively high price of the nuclear units. Besides, many legal problems (dockage licenses, liability questions, etc.) remain to be clarified.

Nuclear process heat has also not been widely used. Although industry requires both low-temperature ($T < 300\,°C$) and high-temperature process heat [45], the high-temperature processes are either still being developed or undergoing testing. In addition, efficient heat transport requires that the reactor be located near the factory, which is, in turn, generally close to a city. Until now, the location of nuclear plants close to cities has been avoided as much as possible.

The same applies to the use of nuclear energy or its waste heat for district heating [45], [46]. Since domestic heating accounts for about 50 % of the energy consumption in industrial nations, power-heat coupling is of tremendous importance as far as energy politics are concerned. However, the heating energy to be decoupled can be transported only over relatively short distances (ca. 50 km) because of the losses incurred. This means that nuclear heating plants must be located in populated areas to be economical.

For these reasons, commercial reactors are used mainly for the generation of electricity in stationary nuclear power plants. A schematic representation of a nuclear power plant with a pressurized-water reactor (PWR) is given in Figure 10. In this case, the reactor core consists of square fuel assemblies, which are placed side by side (see Fig. 18). The fuel assemblies are composed of individual fuel rods that are approximately 1 cm thick. The fuel rods contain pellets of uranium dioxide with a ^{235}U enrichment of 3 – 4 %. A zirconium alloy serves as cladding material for the rods. Heat produced in the fuel elements is removed by the coolant in the primary loop and transferred to the secondary loop (working cycle) in the steam generators. Waste heat is dissipated by means of a cooling tower or a river (see Fig. 10).

The economic motivation for building nuclear power plants is the high energy content of the nuclear fuel uranium and its comparatively low cost: 1 kg of uranium (with 3 % ^{235}U) produce 2.5×10^9 kJ and cost about \$ 630 in the form of uranium hexafluoride (UF_6) in 1996. In comparison, 1 t of oil produce about 4×10^7 kJ and cost approximately \$ 157 in 1996. Thus, the specific cost of uranium $\$\,2.5 \times 10^{-7}$/kJ is lower than the corresponding cost of oil $\$\,3.9 \times 10^{-6}$/kJ or other fossil fuels by a factor of ten. Although this does not include the costs involved in fuel processing, the higher costs of erecting nuclear power plants, and the reprocessing or disposal costs, nuclear power plants are more economical for a wide range of output and load than conventional power plants [44], [48], [49].

The first large-scale nuclear power plant went on-line at Calder Hall (United Kingdom) in 1956. It consisted of four graphite-moderated, carbon dioxide-cooled Magnox reactors (see Section 3.2.2.1) with a total electric output of 220 MW_e. The first large pressurized-water reactor (see also Section 3.1.1), the Shippingport reactor in the United States (140 MW_e), was put in operation in December 1957. However, the birth of commercial nuclear power plant technology is considered to be the Oyster Creek event: i.e., the installation of a 640 MW_e boiling water reactor in the United States at the end of 1963, without governmental assistance and according to purely commercial aspects. Although the total electrical output of all nuclear power plants on order, under construction, or in operation, increased almost exponentially (output doubled ca. every 2 years) until the mid-1970s, this value has remained virtually constant since 1975. At the same time, the number of operating nuclear power plants increased due to the completion of plants that were under construction. In contrast, the number of plants on order decreased, especially after the accidents at Three Mile Island (1979) and Chernobyl (1986). However, an increase in construction is anticipated because of increasing electricity demands, the carbon dioxide problem, and the replacement of old nuclear power plants ([49–52]).

In Figure 11, the total electric capacity of nuclear power plants of a particular type on or-

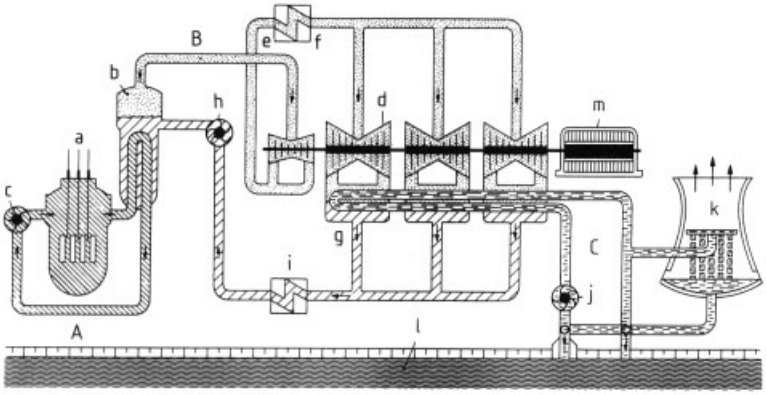

Figure 10. Basic configuration of a nuclear power plant with a pressurized-water reactor [47]
A) Primary system; B) Secondary system; C) Cooling-water system
a) Reactor; b) Steam generator; c) Reactor coolant pump; d) Turbine-generator unit; e) Moisture separator; f) Reheater; g) Condenser; h) Feedwater pump; i) Feedwater heater; j) Cooling-water pumps; k) Cooling tower; l) River; m) Electricity generation

der, under construction, or in operation is plotted logarithmically against the year ([47], [53]). Different types of reactors are described in the following sections.

At the beginning of 1996, 426 nuclear power plant units with a total output of 344 422 MW_e were in operation worldwide, and 39 additional nuclear power plants with a capacity of 32 594 MW_e were under construction or planned [54]. However, the percentage of electric power drawn from nuclear reactors varies considerably from country to country. It ranges from approximately 2 % (China, India), through 30 % to 45 % (Germany, Finland, Japan, Spain, Rep. of Korea, Switzerland and Sweden), up to 75 % and 85 % (France, Lithuania). In 1996 nuclear energy account for about 7 % of the worlds' primary energy consumption.

2.2. Important Technical and Economic Parameters

The costs of electricity generation in a nuclear power plant are related to the following factors (see, e.g., [42], [44]):

1) The *specific cost of the plant* k_p, which depends on the type of reactor and the size of the power plant, is usually expressed in terms of dollars per kilowatt of electrical output. In 1997, a 1300-MW nuclear power plant with a pressurized-water reactor would be expected to have a k_p of about \$ 2000/kW.
2) *Plant utilization* (*load factor*) f which is the ratio of the energy actually produced (in one year) to the maximum possible energy production.
3) The *thermal efficiency factor* η ($\eta = P_e / P_{th}$) is the ratio of electric (P_e) to thermal (P_{th}) output. A distinction is made between the net and the gross efficiency, depending on whether or not the internal electric consumption of the power plant has been deducted. The factor η is determined basically from the condition of the steam in front of the turbine (steam pressure p_{St}, steam temperature T_{St}) and depends largely on the average reactor coolant temperature as a result of heat transfer in the steam generator.
4) The *specific fuel cost* k_F covers the costs of the finished fuel assemblies in dollars per kilogram of uranium used. For reactors that contain thorium or substantial amounts of plutonium, in addition to uranium, this factor is related to 1 kg of heavy metal (hm = U + Th + Pu). The fuel costs are made up of the cost of the natural uranium (or thorium) required, the enrichment costs, and the expenses involved in fabrication of fuel assemblies. In 1997, k_F values of about \$ 1000/kg uranium (3 % ^{235}U) are expected for a pressurized-water reactor. In addition, repro-

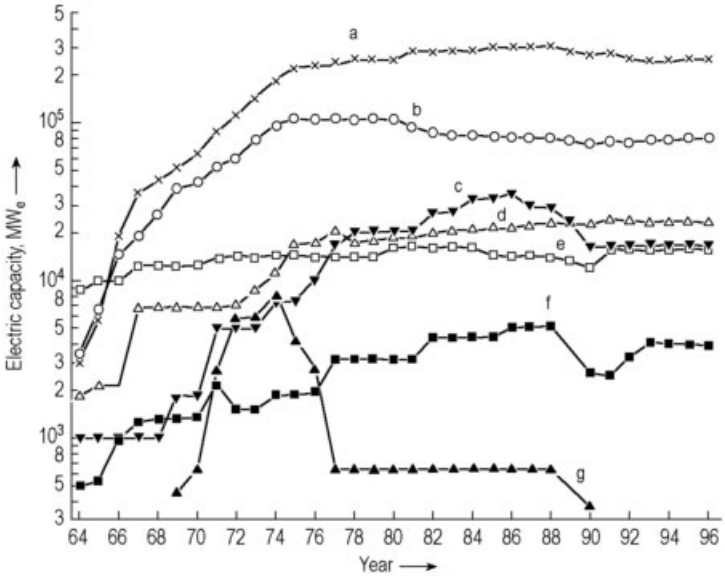

Figure 11. Total worldwide electric capacity of nuclear power plants in operation, under construction, or on order [47], [53] a) Pressurized-water reactors (PWR); b) Boiling-water reactors (BWR); c) Light-water cooled graphite-moderated reactors (LWGR); d) Heavy-water reactors (D_2O); e) Gas-cooled graphite reactors (GGR; including Magnox and advanced gas-cooled reactors); f) Fast breeder reactors (FBR); g) High-temperature reactors (HTR)

cessing and disposal costs ($k_R \approx \$\,2000$/kg uranium, in 1997) have to be considered.

5) The *average specific power* q_{sp} ($q_{sp} = P_{th}/M_F$) is the ratio of the thermal power output P_{th} of the reactor to the fuel mass M_F. In general, q_{sp} is expressed in terms of kilowatts per kilogram of uranium or heavy metal. The average specific power q_{sp} depends on the maximum permissible power density in the fuel q'''_{max} and on the ratio of maximum to average power density (i.e., the power form factor F_Q). If the fuel density is ϱ_F (in the case of UO_2, e.g., 10.35 g/cm^3), then

$$q_{sp} = \frac{q'''_{max}}{F_Q \cdot \varrho_F}$$

$$F_Q = \frac{q'''_{max}}{q'''}$$

The power form factor F_Q is determined by the shape of the neutron-flux density Φ. (For a definition of neutron-flux density, see Section 1.3.6).

The following relation applies:

$$q''' \left[\mathrm{J\,m^{-3}\,s^{-1}}\right] = \chi \Sigma_f \Phi$$

where $\chi = 3.2 \times 10^{-11}$ J is the energy per fission; Σ_f the macroscopic fission cross-section (m^{-1}); and Φ the neutron-flux density (m^{-2} s^{-1}). The maximum power density q'''_{max} is limited by maximum permissible temperatures. The following limits are important:

a) The maximum (central) fuel temperature T_0. In the case of cylindrical fuel rods, as in Figure 12, the heat produced per unit of length is described by the *specific fuel rod power* q':

$$q' = 4\pi \cdot \int_{T_1}^{T_0} \lambda(T)\,\mathrm{d}T$$

where λ is the heat conductivity and T_1 is the fuel pellet surface temperature. The fact that T_0 must remain below the melting temperature of the fuel results in a limiting value for q'. For fuel rods made of uranium dioxide (independent of the radius of the fuel rods and practically independent of the type of reactor), q' must be ≤ 650 W per centimeter fuel rod.

b) The maximum heat flux density $q''_{\max}$ at the surface of the fuel rod. In pressurized-water reactors, a film of steam may develop between the surface of the fuel rod and the coolant if the heat flux density is too high (socalled DNB = departure from nucleate boiling). The transfer of heat would then be impeded, and the cladding tube would be overheated. In pressurized water reactors, the required protection from film boiling limits the maximum possible heat flux density $q''_{\max}$ to approximately 200 W/cm^2.

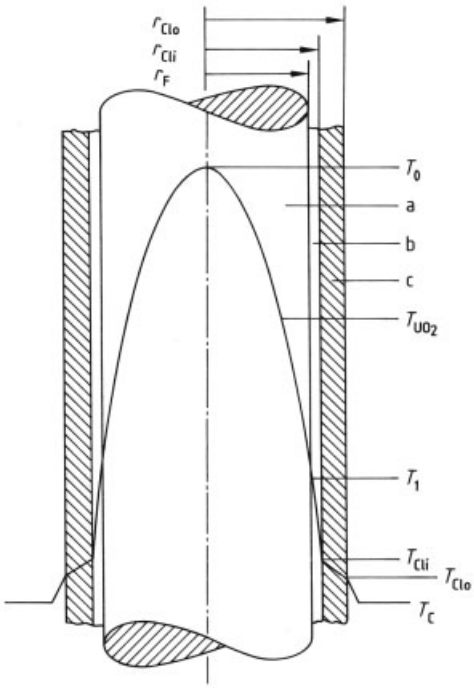

Figure 12. Temperature distribution in a fuel rod [42]
a) Fuel (UO_2); b) Gap; c) Cladding
T = temperature; r = radius; Clo = cladding outside; Cli = cladding inside; F = fuel; C = coolant

For cylindrical fuel rods with radius r_F and a constant power density within the fuel rod, the power density q''', the heat flux density q'', and the specific fuel rod power q' are related as follows:

$$q' = 2\pi r_F \cdot q'' = \pi r_F^2 \cdot q'''$$

Thus, a limit for $q'''_{\max}$ results from the energy output limits for q' and q'' given above.

6) The *burnup B* (see Section 1.3.9) is the thermal energy produced per kilogram of uranium (or heavy metal) during the time of use in the reactor (burnup is usually expressed in the units kW · d/kg U or MW · d/t U). Because each fission releases about 200 MeV (3.2×10^{-11} J) of energy, burnup can also be expressed as FIFA (fissions per initial fissile atoms) or FIMA [fissions per initial (heavy) metal atoms].
The burnup attainable is limited by factors relating both to neutron physics and to the construction materials used. The fuel starts to swell at a higher burnup because of the resulting fission products. This, in turn, can cause damage to the cladding tube (see Section 3.1.4.1). From the standpoint of neutron physics, the burnup of fissile material and the increased poisoning with fission products result in a loss of reactivity (see Section 1.3.9). The material limitation is the limiting factor in most types of reactors. From the point of view of neutron physics, the reactor is generally designed so that the maximum permissible material-dependent burnup is achieved by selecting a suitable initial enrichment (reactors operating with natural uranium are an exception). Because of the location-dependent power density distribution in the reactor, a distinction must be made between the average burnup of a fuel element (or a batch of fuel elements) and the maximum, local burnup.
7) In addition, the costs of electricity generation in a nuclear power plant are influenced by *economic factors* such as depreciation periods and interest rates.

In a nuclear power plant with a pressurized-water reactor, about 70 % of the cost of electricity generation in 1997 was accounted for by investments (depreciation and interest), about 20 % by the fuel (stock and consumption), and about 10 % by the operation.

The following ratios are important for the costs of electricity generation:

k_p/f	for plant costs
$k_F/f\,\eta\,q_{sp}$	for the fuel stock
$k_F/B\,\eta$	for fuel consumption

Table 11 gives a survey of the technical data of some nuclear power plants.

2.3. Radiation Protection and Safety

2.3.1. Radiological Quantities and Units

The α, β, and γ particles generated in nuclear reactors predominantly cause excitation and ionization of atoms or molecules. In this process, electrons are transferred to higher energy levels or removed completely. In complex organic molecules, this leads to the distortion or cleavage of chemical bonds and to the formation of aggressive free radicals, which, in turn, can alter or damage biological macromolecules. Although neutrons cannot interact directly with electrons because of their electrical neutrality, indirectly they are just as active as charged particles via collision and capture processes with the nuclei. In addition, subsequent γ disintegration of the target nucleus or (n, p) and (n, α) reactions can cause chemical transformations, whereby nucleic acids and enzymes are altered.

Radiation damage is approximately proportional to the energy absorbed. Consequently, the most important radiological quantity is the *energy dose D*, which is the radiation energy (of any type) absorbed per kilogram of any substance. The unit of measure is the gray (Gy): 1 Gy = 1 J/kg (the old unit rad is still frequently encountered; 1 rd = 10^{-2} Gy).

However, the biological effect of radiation depends not only on the energy dose, but also on the type of tissue affected and on the manner in which the total energy absorbed per kilogram is distributed spatially and over time. The *energy dose rate* $\dot{D}$, i.e., the absorbed energy per kilogram and second (units: W/kg, Gy/s, mGy/h; or the old units: rd/s, mrd/h, etc.), serves as a measure of radiation exposure per unit time.

The spatial distribution of molecular damage depends on the nature of the radiation. For example, α particles and protons have a substantially shorter range than β particles (electrons) of the same energy because the former have a much larger mass and a correspondingly lower speed. They release their energy along a much shorter path and thus generate a much higher ionization density than β and γ particles. Hence, the molecular damage produced by β or γ radiation is distributed more evenly over the tissue than in the case of an equal energy dose of the heavy corpuscular radiation. Thus, biological damage depends not only on the dose of energy but also on the nature of the radiation. The *dose equivalent* D_q has been introduced as a measure of biological damage:

$$D_q = q \cdot D$$

The dimensionless quality factor q is used to express the relative biological effectiveness of a certain type of radiation compared with γ-rays. Thus according to the definition, $q = 1$ for γ quanta. Quality factors for other types of radiation may be summarized as follows [55], [56]:

β and γ radiation	1
p (energy dependent)	1 – 15
α radiation (energy dependent)	10 – 20
heavy recoil nuclei	20
n, thermal	2.3
n, 5 keV	2.0
n, 20 keV	3.3
n, 100 keV	7.4
n, 1 MeV	10.6
n, 10 MeV	6.8

The dose equivalent D_q is measured in joules per kilogram. However, the special unit sievert (Sv) is also used for D_q, which eliminates any confusion between D and D_q. Therefore,

$$D_q\,(\mathrm{Sv}) = q \cdot D\,(\mathrm{Gy})$$

Correspondingly, the following equation applies to the *equivalent dose rate*:

$$\dot{D}_q = q \cdot \dot{D}$$

$$(\text{units}: \mathrm{W/kg\ or\ Sv/s,\ mSv/h,\ etc.})$$

An old unit for the dose equivalent, which is still encountered frequently in the literature, is the rem (**r**ad **e**quivalent **m**an): 1 rem = 10^{-2} Sv.

In addition to the dosage units mentioned above, the *ion dose J* is a measure of the charge produced per kilogram of air by ionizing γ radiation [units: C/kg or (old) roentgen (R), 1 R = 258×10^{-6} C/kg]. This is the oldest radiological unit, it is applicable only to γ or X-rays and is material related. Because the ionization potential of the air is known (34.0 eV), the ion dose can be converted into the energy dose. The following applies to air: 1 C/kg = 34 Gy (1 R = 0.87 rd), and to body tissues: 1 C/kg = 39 Gy (1 R $\approx$ 1 rd).

Furthermore, the activity A is of considerable importance. It is the number of radioactive disintegration processes per unit time and is measured in becquerel (Bq): $1\,\mathrm{Bq} = 1\,\mathrm{s}^{-1}$. The old unit curie (Ci) is still used frequently: $1\,\mathrm{Ci} = 3.7 \times 10^{10}\,\mathrm{Bq}$.

2.3.2. Radiation Burden, Radiation Damage, Radiation Safety Rules and Regulations

2.3.2.1. Natural and Medical Radiation Doses

Natural exposure results from cosmic radiation (elevation dependent); terrestrial radiation due to radioactive elements in the earth's crust, especially in rocks (location dependent); and the intake of natural radioactive isotopes with air and drinking water (internal radiation burden). For instance, the average value in Germany is approximately 2.0 mSv/a = 200 mrem/a [56], [57]. However, considerable deviation is observed, depending on location; e.g., the external radiation burden arising from cosmic rays at an elevation of 3000 m is 1 mSv/a and at 6000 m more than 6 mSv/a (compared to ca. 0.3 mSv/a at sea level). Local deposits of radioactive minerals and rocks have an even stronger effect. Values up to 120 mSv/a have been measured at individual locations. Apart from the natural radiation burden, the radiation exposure connected with various *medical procedures* is of special importance and interest. For example, these are approximately 0.3 mSv from a lung X-ray, 1 – 4 mSv from a pelvic X-ray, and up to several sieverts from γ radiation in cancer therapy [56].

2.3.2.2. Radiation Damage and Radiation Safety Regulations

The *somatic damage* caused by higher radiation burdens is presented in Table 9 for the worst case (one-time exposure of the entire body). No direct damage can be observed up to a dose equivalent of 0.25 Sv. Clinical symptoms appear at dosages higher than ca. 0.25 Sv, and subjective complaints set in above ca. 1 Sv (100 rem).

Table 9. Medical symptoms following whole-body exposure to different doses of radiation

Dose, Sv	Symptoms and their consequences if no medical treatment is given
0.25	(hazardous dose) changes in blood count detectable; no subjective complaints
1	no definite symptoms detectable at first; loss of appetite, nausea, exhaustion, diarrhea and moderate marasmus after three weeks; recovery of the patient is probable; chance of death in individual cases
4	nausea and vomiting 1 – 2 h after exposure to radiation; symptoms soon disappear, only to reappear in a more severe form after 2 weeks, together with symptoms described above; other complaints: fever, severe inflammation of mouth and throat, and nasal bleeding; 50 % chance of survival
6	course of the disease similar to that described above; interval without marked symptoms is shorter; virtually 100 % chance of death

A typical characteristic of radiation sickness is that irradiation is followed by a longer or shorter latent period without defined symptoms, the duration of which depends on the radiation dose. This period can last a few days, even in the case of the virtually 100 % lethal dose of 6 Sv. The latent period is caused by the fact that radiation damages particularly the cells of hematopoietic organs (bone marrow and lymphatic tissue) and, consequently, interferes with or totally stops the supply of red and white blood cells. However, this lack of replenishment comes to light only after the death of the cells still present in blood. In a similar manner, the otherwise continuous renewal of the intestinal mucosa is prevented. The results are diarrhea, anemia, increasing weakness, and inability of the body to fight additional infections.

Apart from the hematopoietic organs, the intestine, skin, and growing bones (i.e., parts of the body that undergo rapid cell division), the reproductive glands are particularly sensitive to radiation. Chromosomes and genes can be damaged, and genetic traits can be altered (*genetic damage*). Also, the statistical long-term risk of leukemia and other forms of cancer increases slightly with increasing doses of radiation (*long-term damage*). The radiation-induced probability of getting cancer is, however, generally small compared to the natural incidence of this disease. The spontaneous frequency of cancers in the industrialized countries of the Western world is about 20 % (i.e., 20 % of all people die from cancer). On the other hand, the risk of getting cancer from radiation in doses of up to 1 Sv is reported

to be 0.07 per sievert [58], [59]. Therefore, in the case of a total dose of 0.1 Sv (10 rem), the additional radiation-induced incidence of cancer would amount to 0.7 % (compared to 20 %).

The information presented above shows that radiation exposure up to a few hundredths of a sievert per year is harmless to the health of the individual. In contrast, the statistical risk of the occurrence of hereditary defects and certain diseases (long-term damage) increases slightly but continuously with increasing doses of radiation. Although this additional risk generally remains negligibly small for the individual compared with other risks in life, it could, however, become important when larger parts of the population are exposed to an increased radiation burden.

For this reason, the ICRP (International Commission on Radiological Protection) has drawn up recommendations for a *radiation protection law* [55] and is revising them constantly. This law is based on the ALARA (**a**s **l**ow **a**s **r**easonably **a**chievable) principle. These regulations have been accepted by governmental organizations of individual countries [60–63] and represent binding, national law. The German radiation protection regulation (StrlSchV) [64], for example, is based on the principle that the radiation burden of the *entire population* must remain as low as possible and must not exceed the range of the natural radiation burden. On the other hand, *special regulations* apply to persons who are exposed to radiation in the course of their work. These are based on a maximum annual exposure of 50 mSv.

For the general population outside a control area, the annual dose equivalent must not exceed 0.6 mSv, of which the contribution by air or water is limited to 0.3 mSv each. Experience has shown that the actual annual radiation burden in the vicinity of a nuclear power plant is approximately 0.01 mSv.

The permissible doses and dose rates for the irradiation of parts of the body can be larger (the dose rate for occupationally exposed persons is e.g., 300 mSv/a for the thyroid and 500 mSv/a for the hand). Other limiting values for practical radiation protection (e.g., limitation of the occupational lifetime dose to 0.4 Sv) are being discussed continuously and updated in national regulations (see e.g., [64], [65]).

Because of the relationship between dose rate, neutron or γ-flux density, and activity, limiting values for the neutron- and γ-flux densities, for the maximum permissible activity input into a body, and for the concentration of radioactive substances in air and water are also obtained from the maximum permissible dose rates. The activity concentration K is deduced from the particle concentration N and the radioactive disintegration constant λ_{D} according to the formula

$$K = \lambda_{\mathrm{D}} \cdot N$$

Numerical values can be obtained from the corresponding radiation protection regulations.

2.3.3. Emission Protection

All nuclear reactors are surrounded by *concrete shielding* (several meters in thickness), as a protection barrier against direct neutron and γ radiation from the reactor core. Thus, direct radiation plays no role outside the nuclear plant. On the other hand, the emission of gaseous and liquid radioactive substances into the surroundings cannot be completely avoided. Emission protection devices serve to maintain the radioactive environmental burden below the maximum permissible values. This results in important restrictions concerning maximum permissible activity emissions per unit of time.

According to international law, *liquid radioactive waste* may be drained off into the surroundings only if the activity concentration of the water discarded has a dose equivalent rate $\dot{D}_q \leq 1.5$ mSv annually, even if the water is permanently used as drinking water, bathing water, etc. Even lower limiting values are set in some countries, e.g., the 0.6 mSv- concept in Germany [64].

The maximum permissible emission of *activity via the stack* of the power plant is determined by the aforementioned maximum dose rate allowed in the surroundings of the plant and from the dissipation of activity in the atmosphere. Special directives apply to the calculation of activity dissipation [56]. On the basis of these ratios, licensing authorities stipulate maximum values for the annual permissible activity emission for each nuclear power plant.

The main sources of radioactivity in nuclear power plants are:

1) *Fission products* that arise from nuclear fission in the fuel (see Table 59) and their daughter products; and
2) Nuclei activated by neutron capture, the *activation products*, which develop primarily in the coolant and from impurities therein (corrosion products).

Although the fission products remain for the most part within the fuel elements, a small part of the more mobile isotopes (mainly noble gases, iodine, etc.) can diffuse out of the fuel and enter the coolant through defects in the cladding tube. Even though most of the fission and activation products are removed in coolant purification plants, a portion is still released into the surroundings with the wastewater and the exhaust air from ventilation units.

Emission via the Stack. The emission rate of noble gases via the stack is ca. 10^{13} Bq annually (mainly ^{133}Xe and ^{85}Kr) for a 1300-MW_e pressurized-water reactor. Smaller amounts of ^{131}I and aerosols (ca. 2×10^8 Bq/a) are also emitted. However, experience shows that the resulting real radiation burden for the population living in the vicinity of a nuclear plant, is far below the design value and, at ca. 1×10^{-2} mSv annually, amounts, e.g., to only about 1 % of the natural radiation burden in Germany. The path of exposure, pasture – cow – milk – thyroid gland, is considered to be the *critical path*. The intake of radioactivity via the food chain is called ingestion.

Radioactive wastewater is produced by leakage and, to a larger extent, in laboratories, laundries, etc. The methods used in processing wastewater are precipitation, filtration, evaporation, and ion exchange. Indeed, decontamination can be reduced by a factor of $10^4 - 10^5$ (excluding volatile constituents and ^{3}H) with modern evaporation plants. In this way, the activity (without ^{3}H) finally released into the surroundings can be limited to ca. $2\times10^7 - 7\times10^9$ Bq annually, with an additional annual tritium emission maximum of 3.7×10^{13} Bq (1000 Ci/a) in the case of large pressurized-water reactors. Retaining ^{3}H is practically impossible with conventional methods. Experience shows that the effective radiation burden on the population living in the vicinity of the plant due to wastewater is $< 1\times10^{-3}$ mSv/a (i.e., barely 1 ‰ of the natural radiation burden in Germany) [44].

A direct connection between an activity emission rate (Bq/a) and a radiation burden (mSv/a) cannot be made because the radiation burden of a human being depends on the amount of radioactive substances absorbed (which depends on location, weather conditions, etc.) and the special isotopes involved. However, with the help of meteorological dissipation factors a known emission rate (Bq/a) can be used to calculate the activity concentration (Bq/m^3) produced at a certain spot with a particular weather condition. This value, in turn, gives the particular radiation burden for a person at a definite rate of intake, a definite residence time (e.g., permanent residence), and a known isotopic composition. The average (or maximum) values for the annual radiation burden of the population in the vicinity of the plant can then be calculated by taking into account the statistical frequency of various weather situations and living conditions. The calculations are checked and supplemented by means of control measurements of the actual radiation burden in the vicinity of nuclear power plants.

Solid Waste. Apart from gaseous and liquid waste products, solid waste is also produced in the form of used ion-exchange resins, evaporator concentrates, filters, etc. In the case of a large pressurized-water reactor, a total of ca. 200 m^3 of low-, medium- and high-level waste is produced per year. This does not represent an environmental hazard if intermediate and final storage of the waste and spent fuel elements are properly conducted (see Section 5.5.4).

2.3.4. Industrial Safety

During normal operation, the radiation burden from a power plant is negligibly small (barely 1 % of the natural radiation burden). However, the question arises whether malfunctions could pose an environmental threat.

A nuclear explosion, as in the case of an atomic bomb, is basically impossible with a reactor because the reactivity decreases with increasing temperature and decreasing material density. Consequently, the chain reaction

would immediately stop itself (for special problems with fast breeder reactors and graphite-moderated light-water-cooled reactors, see Sections 4.1.3 and 3.3.3, respectively). However, the radioactivity remaining in the fuel elements represents a considerable potential threat. The release of a large part of the radioactivity contained in a reactor can have catastrophic consequences, as shown, for example, by the reactor accident in Chernobyl (see Section 3.3.6).

Therefore, the most important task of the technical safety equipment is to ensure that disasters of this type are impossible and to rule out even smaller malfunctions with adequate certainty. For this purpose, every reactor plant is provided with extensive safety devices, which are based on careful malfunction analyses [66]. A safety report must be submitted to the authorities before a reactor is put in operation. This report must contain an exact description of the plant with its safety facilities and must furnish proof that the plant will pose no threat to the surroundings.

Not only severe mishaps, but all malfunctions must be prevented in a nuclear power plant, hence safety measures are an integral part of a multistage concept. The individual levels of this concept are the operational systems, the protection systems, the safety systems, and the severe accident management.

All systems required for the operation of the plant in accordance with regulations (normal operation) are considered *operational systems*. All parts of the plant that protect the nuclear power plant or component systems from destruction during abnormal operation (see Section 3.1.3) are called *protection systems*. The function of the *safety systems* is to control possible malfunctions and to guarantee containment of radioactive substances. *Severe accident management* includes measures for the utilization of design reserves as well as operational and external facilities that are capable of reducing the consequences of an accident inside or outside the area of the power plant.

2.3.4.1. General Design Principles for Protection Systems

Inherent Safety. The reactor system must be designed in such a way that it returns to a safe state by itself in the case of malfunction. This occurs, e.g., because of the negative power coefficients of the reactivity (see also last paragraph of Section 1.3.4) and control rods that automatically shut down the reactor if there is a power failure.

Single-Error Criterion. A relevant technical safety system must fulfill its function as required even if, at the same time an additional single error causes the failure of components or parts of the system. Hence, essential safety devices are present several times (i.e., redundantly).

Redundancy. The installation of surplus units to increase reliability is known as redundancy. For the measurement of important parameters, for instance, 3 or 4 similar measuring channels are used whose signal techniques are independent of each other, spatially separated, and connected in parallel. One or two of these channels may fail without affecting the efficiency of the system. The same applies to many other components, (e.g., the pumps of the emergency cooling system).

Diversity. The diversity of protective measures refers to the installation of different protective devices, which serve the same purpose but are designed according to different principles and, as far as possible, are triggered by different physical variables (e.g., shutting down the reactor by using absorbing rods or by using liquid boric acid). Diversity is meant to exclude systematic errors that would not be prevented by the multiple use of identical protective devices.

Automation. All technical safety actions are triggered automatically by signals that come from various limiting value indicators. Here, the safety system has priority over all other systems and cannot be influenced by the operator. The components actively operated by the protection system are, in case of demand, independent of an external power supply.

Quality and Control. All safety facilities must meet special quality requirements, which include production controls and regular rechecks (automatically, if necessary).

Object Protection. In addition, the most important components of the individual safety systems are protected against forces from outside, such as an airplane crash, earthquake, explosion, or sabotage.

2.3.4.2. Safety Channels, Safety Limits, and Safety Measures

A safety system consists of (1) measuring elements for the detection of physical forward signals; (2) a signal processing and decision level for the processing of measured variables; and (3) limiting values, as well as sequentially arranged mechanical safety equipment, which is started by appropriate control signals. The measuring and decision parts form a safety channel that triggers one or more protective actions when permissible limiting values are exceeded. These values are selected in such a way that no dangerous operational parameter is reached.

Important *safety channels* cover, for example,

The thermal reactor output P_{th},
The reactor period $T_p = \Phi/\dot{\Phi}$ [especially important in the start up phase (Φ = neutron-flux density)],
The protection against film boiling in water-cooled reactors,
The coolant flow rate,
The system pressure,
The water level in the steam generators, and
The condenser pressure.

The most important *protective actions* are as follows:

The reactor scram (i.e., emergency shutdown),
The turbine emergency shutdown,
The response of various safety valves,
The emergency power supply,
The emergency feeding of the steam generators,
The emergency core cooling, and
containment (closing off the building).

2.3.4.3. Engineered Safeguards

The protective actions mentioned above are triggered by technical safety equipment, which may also include passive safety devices. The most important safety features of pressurized-water reactors are shown in Figure 13 [47]. The *passive safety elements* include (see Fig. 13):

1) the cladding tubes and the closed primary circuit (reactor cooling loop), which form the first and second barriers against the escape of fission products;
2) the biological shield (d) with the protective cylinder (e);
3) the gastight steel containment (g, wall thickness ca. 30 mm);
4) the secondary containment (h) made of concrete (wall thickness ca. 2 m); and
5) the accumulator (m), which contains borated water. In case of a serious loss of coolant, back-pressure valves permit quick and automatic feeding of sufficient quantities of water into the coolant cycle (8×20 % redundancy).

The steel containment is designed to hold the mass and energy of the primary coolant even in the case of a total loss of coolant. The steel containment is enclosed in a secondary concrete containment (h) to protect the surroundings from the direct radiation of the radioactive coolant. The annulus between steel and concrete containment is maintained at a subatmospheric pressure to prevent the escape of radioactive materials in the event of leakages.

The *active safety devices* of a pressurized-water reactor include (Fig. 13) the following:

1) Low-pressure residual heat-removing pumps (j). In the case of loss of coolant, these pumps deliver borated water from the borated water storage tanks (l) or, after they have been emptied, from the building sump via secondary coolers to the primary circuit. In case of a large leak, these pumps ensure cooling of the reactor core even after drainage of the pressure tank (4×50 % redundancy)
2) High-pressure safety feed pumps (i), which can feed against higher pressure and can flood the core if, in the event of a smaller leak, the system pressure drops relatively slowly (4×50 % redundancy)
3) The emergency power supply, consisting of batteries (s) and converters for the uninterrupted supply of important control and locking systems, etc. (2×100 % redundancy), as

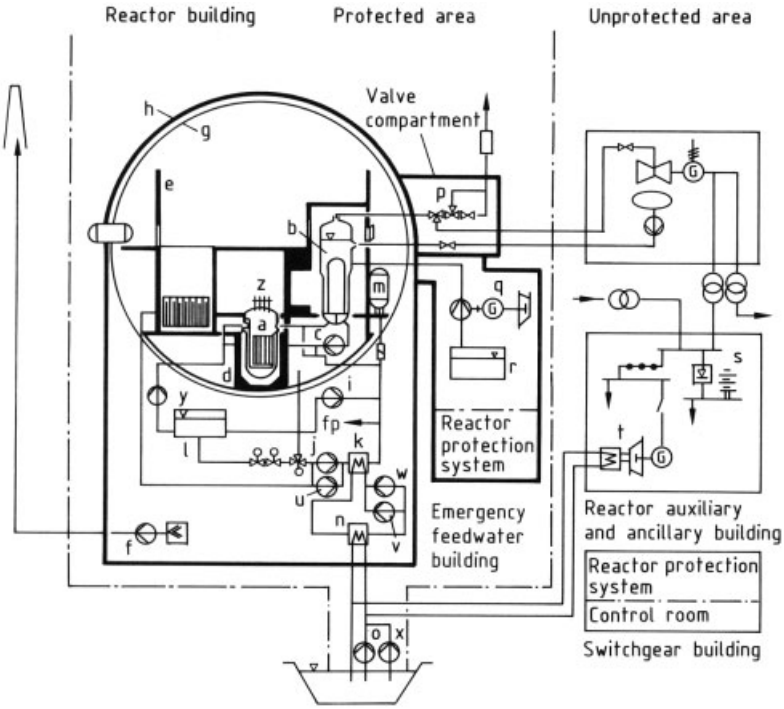

Figure 13. Engineered safety features of Siemens/KWU pressurized-water reactors [47]
a) Reactor pressure vessel; b) Steam generator; c) Reactor coolant pumps; d) Biological shield; e) Internal concrete cylinder; f) Annulus ventilation system; g) Containment; h) Secondary containment; i) Safety injection pump (4×50 % redundancy); j) Residual heat-removal pump (4×50 % redundany); k) Residual heat-exchanger (4×50 % redundancy); l) Borated water storage tank (4×50 % redundancy); m) Accumulator (8×20 % redundancy); n) Intermediate cooler; o) Service cooling-water pumps; p) Main steam and feedwater valve compartment; q) Emergency generator; r) Demineralized water storage tank; s) Batteries (2×100 % redundancy); t) Emergency diesel engine (4×50 % redundancy); u) Fuel pool pump; v) Nuclear component cooling pumps; w) Emergency nuclear component cooling pumps; x) Emergency service cooling-water pumps; y) Extra borating pump; z) Rapid shutdown system
G = generator; fp = fuel pool

well as highly efficient, quick starting diesel engines (t) (4×50 % redundancy)
4) The control rod system for the reactor scram
5) Numerous safety valves, shutoff valves, and similar components that cannot be discussed here in detail

All possible mishaps can be controlled with the help of these generally multiple redundant safety devices. The special safety devices of other types of reactors are dealt with as each reactor is described.

2.3.5. Accidents

2.3.5.1. Classification and Analysis

Classification of the conditions of operation of nuclear power plants involves the following definitions (according to [66], [67]):

1) *Designated operation* includes
   a) Operational processes for which the plant with properly functioning systems (undisturbed state) is designed and to which it is suited (normal operation)
   b) Operational processes that proceed properly when components of the plant fail to function (disturbed state), as long as no

safety reasons exist for discontinuing operation (abnormal operation)

c) Processes involving control, testing, maintenance, and repair

2) *Malfunctions* are a course of events that require plant operation to be discontinued for reasons of safety, even though the plant has been designed for these events.

3) A range of conceivable events, called *accidents*, remains beyond the safety design of nuclear power plants. These include events that are, as far as anyone can judge, so improbable that specific measures for prevention or limitation of their consequences are not usually taken, or events whose occurrence and course are not foreseeable. In the event of accidents, limiting values for the radiation burden stipulated in radiation protection regulations can be exceeded.

The most important malfunctions include the following:

Mishaps dealing with reactivity (e.g., control rod ejection)
Disturbances in the core cooling system (e.g., because of pump failure or loss of coolant)
Disturbances in heat removal (e.g., because of load rejection or emergency shutdown)
Failure of secondary pipes

Conceivable *external causes of damage* are storms, storm tides, floods, earthquakes, explosions of plants or of means of transport in the vicinity (e.g., air crashes), and sabotage. Protection against external damage requires structural measures and precautionary action to be taken in the design of the technical systems. These measures are based on meteorological, geological, and transport-oriented expert opinions and experience.

The safety planning and assessment of a plant is based on an exact analysis of all possible malfunctions, courses, and effects of malfunctions, that is, of all conceivable chains of events including the corresponding probability of occurrence.

The probability of occurrence and the effects of any case of damage can be calculated in advance (*probability concept*) if all the independent individual probabilities are known. Indeed, a quantitative measure for the *risk of damage* $R_D = P_D \cdot D$, can be defined by multiplying the *extent of damage D* (e.g., quantified by the activity released or the radiation burden) by the probability P_D. In any event, this risk must be so low that for the people living in the vicinity of a nuclear power plant, the additional nuclear risk of damage is sufficiently small compared with the risk from other situations in life (illness, traffic accidents, etc.).

Until now, all investigations have shown that the *statistical risk* from nuclear power plants is negligibly small compared to other natural and civilization-dependent risks (see [43], [67–70]). The statistical risk also includes the reactor accident at Chernobyl, as shown in the evaluation of available data [70], [71]. Quantitative details on the relationship between the extent and probability of damage are presented in Figure 14.

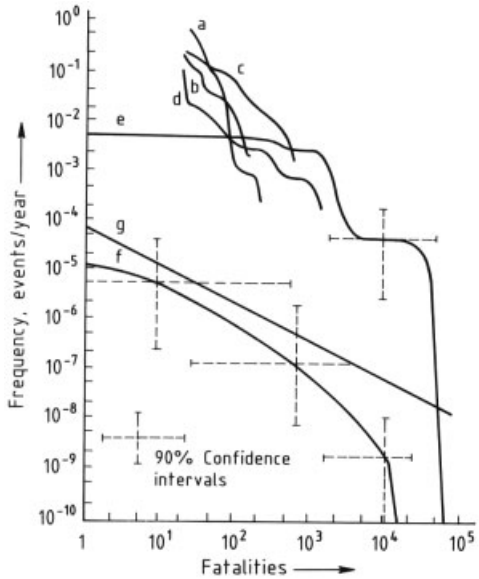

Figure 14. Cumulative distribution function of fatalities for different types of accidents [69]

United Kingdom:	a) Air crash
	b) Mining accidents
	c) Railroad accidents
	d) Fires and explosions
FRG:	e) Late fatalities
	f) Early fatalities
Unites States:	g) Meteors

b) – d): actual occurrence; e), f): expectation for nuclear power plants

2.3.5.2. Design Accident

It is customary to design safety facilities according to the *maximum credible accident* (MCA). The MCA is defined as the accident that produces an effect, measured by the radioactive burden on the surroundings, that cannot be surpassed by any other credible accident during the lifetime of the plant. Frequently, the term *design accident* is preferred because it refers to the largest malfunction, which not only forms the basis for the design of safety devices but is also safely handled by them.

For a long time, a single maximum credible accident was considered relevant for all safety facilities. This was, for most types of reactors, a sudden loss of coolant that was defined, for example, in the case of a pressurized-water reactor, by a double-sided open break of the main coolant pipe. This amounts to a maximum leakage cross-section of about 0.5 m^2 (nominal diameter of the main coolant pipe is 750 mm) in reactors of the 1300-MW category. However, more recent experimental and theoretical studies have shown that a single design accident does not have the same significance for all safety devices. For instance, the double-sided open break in the main coolant pipe is considered to be an MCA only for emergency cooling, whereas otherwise, a malfunction spectrum with varying leakage cross-sections and points of breakage has to be taken into account.

2.3.6. Location Problems

The choice of suitable locations is becoming increasingly difficult because the demands for safety, profitability, and technical realizability (e.g., with respect to waste-heat removal) partly contradict each other. In nuclear power plants that exclusively generate electricity, consumer proximity is not an economic necessity. In fact, until now, remote locations have been preferred for these plants because of the large cooling water requirement. However, because rivers are only allowed to be heated to a limited extent for ecological reasons, cooling towers are increasingly employed.

In wet cooling towers, only the quantity of water that evaporates in the cooling tower must be replaced. Because of the large enthalpy of evaporation of water, the cooling water requirement is barely 2 % of the quantity required for direct river-water cooling. Dry cooling towers require no cooling water at all, but they need large heat- exchanging surfaces (increased construction volume) and a higher condenser temperature (decrease in thermodynamic efficiency).

Efforts made to find suitable locations for nuclear power plants either underground or at sea have not been successful so far.

In general, *consumer proximity* is required for the location of plants with power – heat coupling (e.g., for district heating). The problems that are still unsolved are mainly a result of the licensing procedures.

2.4. Survey of Reactor Types

2.4.1. Classification Features

Commercial reactors can be classified according to various characteristics:

Neutron energy:	thermal or fast
Fuel:	metallic or ceramic
Degree of enrichment:	natural or increased content of fissile material
Fissile material:	^{235}U, ^{239}Pu, ^{241}Pu, ^{233}U
Moderator:	H_2O, D_2O, C, or none
Coolant:	H_2O, D_2O, CO_2, He, liquid Na, etc.
Operational cycle:	direct circulation of the coolant between reactor and turbine or interconnection of a heat exchanger (e.g., a steam generator)

2.4.1.1. Neutron Energy Spectrum

The fast neutrons generated in nuclear fission (ca. 2 MeV average initial energy) are usually slowed down by a *moderator* (see Section 1.3.5) to thermal energies (< 0.1 eV) before they initiate new nuclear fissions. In the thermal range of energy, the effective cross-section for nuclear fission (see Section 1.3.4) is very large. For this reason, thermal reactors can be operated with a relatively small content of fissile material (lightly enriched uranium or natural uranium). The required enrichment depends primarily on neutron absorption in the moderator, coolant and structural materials, and also the desired burnup.

Reactors that work with fast neutrons (fast breeder reactors) do not contain a moderator.

A fast breeder reactor is the only type of reactor that can generate more fissile material in the form of ^{239}Pu than it uses by neutron capture in ^{238}U, which is subsequently converted to ^{239}Pu. Because of these characteristics, the fast breeder reactor is of fundamental importance for the long-term supply of nuclear fuel. Only in these reactors is the *conversion rate* > 1 (see Section 1.3.8). The reason is that the neutron yield per neutron absorption in the fissile material increases with increasing neutron energy and is sufficiently large only in fast breeder reactors. Nevertheless, fast breeder reactors require a substantially higher content of fissile material in the fuel than thermal reactors (ca. 15 – 30 % compared with 3 – 4 % in a pressurized-water reactor), so the probability is sufficiently high that the fast neutrons generated during nuclear fission will initiate new fissions before they are slowed down by collisions with fuel and coolant nuclei. Only substances that have a low slowdown power or a low density, particularly liquid sodium or helium gas, can be used as coolants (for detailed description, see page 164 and Section 4.1.2).

2.4.1.2. Fuels

Uranium dioxide is used as fuel in most types of reactors. It possesses a high melting point (ca. 2750 °C), favorable burnup properties, and good compatibility with cladding materials and most coolants. Although *carbidic fuels* (UC and UC_2) clearly have a higher thermal conductivity (especially UC), they are not yet well researched and are incompatible with water (important if the cladding tubes are damaged). However, UC may be the fuel of the future for sodium-cooled fast breeder reactors (see Section 4.2.1). For graphite-moderated high-temperature reactors (see Section 3.2.3.1), a ^{232}Th – ^{233}U breeding process seems suitable for reasons of neutron physics and materials science. Here, uranium dioxide (UO_2) is employed together with thorium dioxide (ThO_2) (in the future, possibly UC_2 and ThC_2) in the form of small particles that are coated with pyrocarbon and embedded in a graphite matrix. The ^{233}U, is bred from ^{232}Th which, with increasing burnup, steadily assumes the function of the fissile material.

Metallic uranium permits only low fuel temperatures and low burnup, and is used exclusively in Magnox reactors (see also Section 3.2.2.1).

2.4.1.3. Moderators

Only substances with a low mass number (good decelerating power) and a low neutron absorption can be used as moderators. Hence, H_2O, D_2O, and graphite are good moderators. *Heavy water* (D_2O) is contained in natural water to the extent of 0.015 % and is very expensive in the pure state (99.75 %), ca. $ 500 per kilogram (for methods of separation, see → Isotopes, Natural, Chap. 4.2.1.). D_2O also has a poorer slowing-down power than H_2O. For this reason, the volume required is higher by about an order of magnitude. Thus, D_2O-moderated reactors require moderator – fuel volume ratios of about 20: 1 and are, consequently, much larger than H_2O-moderated reactors with the same output. On the other hand D_2O has a lower (desirable) capture cross-section for neutrons than H_2O. That is why D_2O-moderated reactors can be operated with natural uranium (0.72 % ^{235}U). The same applies to *graphite*. Enriched fuel is, however, used in graphite-moderated advanced gas-cooled reactors (AGR) and in high-temperature reactors (HTR) to achieve a higher burnup.

2.4.1.4. Coolants

Coolants must have a high specific heat capacity, high density, good thermal conductivity, low viscosity, low susceptibility to activation, and low neutron absorption. In addition, they must be chemically compatible with materials and fuels. From the standpoint of heat transfer, the most important factor is the heat-transfer coefficient α of the coolant which, in turn, depends on the Reynolds number *Re* and the Prandtl number *Pr*, via the thermal conductivity λ_c, the hydraulic cooling- channel diameter D_h, and the Nusselt number *Nu*:

$$\alpha = \frac{\lambda_c \cdot Nu}{D_h}$$

$$Nu = f\ (Re, Pr)$$

$$Re = \frac{\varrho_c \cdot v_c \cdot D_h}{\eta_c}$$

$$Pr = \frac{c_{p,c} \cdot \eta_c}{\lambda_c}$$

where ϱ_c is the density, v_c the flow rate, η_c the viscosity, and $c_{p,c}$ the specific heat capacity of the coolant. For most coolants,

$$Nu \sim (Re)^{0.8}\ ;\ \alpha \sim \lambda_c \cdot \frac{(\varrho_c \cdot v_c)^{0.8}}{\eta_c}$$

The following proportionality applies to the ratio of the circulation power P_p (pump or blower power) to the thermal reactor power P_{th} [42]:

$$\frac{P_p}{P_{th}} \sim \frac{v_c^2}{c_{p,c} \cdot \Delta T_c} \sim \frac{(q'')^2}{\varrho_c^2 \cdot c_{p,c}^3 \cdot \Delta T_c^3}$$

where ΔT_c is the coolant warm-up span in the reactor (i.e., the difference between the outlet temperature $T_{o,c}$ and the inlet temperature $T_{i,c}$) and q'' is the heat flux density on the surface of the fuel rod.

Large heat-transfer coefficients α require large mass flow densities $\varrho_c \cdot v_c$, whereas large flow rates v_c result in large circulation power P_p. In gas-cooled reactors, the coolant circulation power P_p is acceptable only if the coolant heat-up span $\Delta T_c = T_{o,c} - T_{i,c}$ is large. Because the average coolant temperature $(T_{o,c} + T_{i,c})/2$ is particularly significant for the thermal efficiency of the plant, a loss of efficiency is acceptable here if the outlet temperature is limited by the construction materials used and $T_{i,c}$ must be lowered. In addition to the warm-up span ΔT_c, the combined data of the cooling medium $\varrho_c^2 \cdot c_{p,c}^3$ are particularly important for the relative circulation power P_p/P_{th}.

The thermohydraulic data of four typical power reactors are presented in Table 10. Practically, only water, heavy water, carbon dioxide, helium, and sodium are employed as coolants. In addition, organic liquids (disadvantage: radiation sensitivity), argon ($\varrho_c^2 \cdot c_{p,c}^3$ value is less favorable than that of He), and NaK (a eutectic alloy of approximately 20 % Na and 80 % K) were considered.

Water and heavy water are used both as coolants and as moderators. As a result of the low density of gaseous coolants and of the high mass number of sodium, an additional moderator (generally graphite) is necessary in these cases, allowing the reactor to operate with fast neutrons.

2.4.2. Reactor Types

The various moderators, coolants, and fuels could theoretically be combined to give a large number of different types of reactors. However, only the reactors described below have or had practical importance. Some technical data for typical power reactors are listed in Table 11. Details are given in the specific sections.

2.4.2.1. Light-Water Reactors (LWR)

Pressurized-Water Reactor (PWR). Water under high pressure serves as moderator and coolant, and lightly enriched uranium dioxide in cladding tubes made of zircaloy is used as fuel. The heat created in the reactor is transferred via a closed primary loop, with the help of steam generators, to a nonradioactive secondary cycle (see Fig. 10 and Table 11).

Boiling-Water Reactor (BWR). The boiling-water reactor is also moderated and cooled with water, but it differs from the pressurized-water reactor in that the water in the reactor is boiling and the resulting steam is led directly into the turbine without an interconnecting steam generator. This results in a simplified coolant circuit; the large steam generators and the pressurizer are no longer required, and the circulation power is smaller. The reactor operating pressure is lower than in pressurized-water reactors at about the same operating temperature. However, the entire circulation system is more or less radioactive. The reactor core has a similar design as that of a pressurized-water reactor, but the power densities attainable are lower. Also, more space for steam separators, etc., is required in the pressure vessel, which makes the pressure vessels of boiling-water reactors larger and more expensive than those of

Table 10. Thermohydraulic data of typical power reactors [42], [72]

	PWR Biblis B	AGR Dungeness B	HTR THTR-300	FBR Na Creys – Malville
Power, MW				
Thermal P_{th}	3733	1458	750	3000
Electrical (net) P_e	1240	606.5	296	1200
Average heat flux density q'', W/cm^2	60.6	17.2	24.0	104.0
Coolant	H_2O	CO_2	He	Na
Average temperature, °C	306	498	505	470
Average pressure, MPa	15.5	3.2	4.0	ca. 0.6
Average density ϱ_c, kg/m^3	713	21.7	2.42	837
Average specific heat capacity $c_{p,c}$, kJ kg^{-1} K^{-1}	5.64	1.18	5.20	1.27
$\varrho_c^2 \cdot c_{p,c}^3$, kJ^{3}m^{-6} kg^{-1} K^{-3}	9.12×10^7	7.74×10^2	8.23×10^2	1.44×10^6
$\varrho_c^2 \cdot c_{p,c}^3$ relative to H_2O	1	8.49×10^{-6}	9.03×10^{-6}	1.58×10^{-2}
Mass throughput m_c, kg/s	2.00×10^4	3.38×10^3	3.00×10^2	1.64×10^4
m_c/P_{th}, kg/MW · s	5.36	2.32	0.40	5.47
Average flow rate v_K, m/s	4.9	20	11.6	4.0
Coolant warm-up span $T_{o,c}-T_{i,c}$, °C	32.8	355	490	150
$\varrho_c^2 c_{p,c}^3 (T_{o,c}-T_{i,c})^3$ relative to H_2O	1	0.011	0.030	1.5
$(q'')^2/\varrho_c^2 c_{p,c}^3 (T_{o,c}-T_{i,c})^3$ relative to H_2O	1	7.50	5.21	1.96
Circulation power P_p, MW	26	45	10.7	16.8
P_p/P_{th}	6.96×10^{-3}	3.09×10^{-2}	1.43×10^{-2}	5.60×10^{-3}
P_p/P_e	2.10×10^{-2}	7.42×10^{-2}	3.61×10^{-2}	1.40×10^{-2}

pressurized-water reactors. Altogether, the advantages and disadvantages more or less balance each other, so that the pressurized-water reactor and the boiling-water reactor are the dominating reactor types (see Fig. 11, Table 11, and Section 3.1.2).

2.4.2.2. Graphite-Moderated Reactors

Graphite differs from water, from the point of view of neutron physics, in that it has a lower neutron absorption (desirable) but a poorer neutron moderation. For this reason, graphite-moderated reactors require substantially more moderator volume and are (for the same output) always much larger than water-moderated reactors. Basically, four different types of graphite-moderated reactors are used:

Magnox Reactors. Metallic uranium of natural isotopic composition in cladding tubes made of a magnesium alloy (Magnox) serves as the fuel. Carbon dioxide is employed as the coolant. The cooling channels pass through the graphite moderator. Fuel elements of this type allow only a low burnup and a low specific power (see Table 11). The costs of the plant are relatively high. Reactors of this type are no longer being built (see Section 3.2.2.1).

Advanced Gas-Cooled Reactors. In the AGR, metallic fuel rods are replaced by oxidic elements made of rod bundles (UO_2 pellets in cladding tubes of stainless steel) to improve the thermal and material properties. The fuel rods are similar to those used in H_2O- and D_2O-moderated reactors. Carbon dioxide is used as coolant. The use of steel as cladding material necessitates the use of slightly enriched uranium as fuel. The improvement in efficiency compared to the Magnox reactor is shown in Table 11. Until now, reactors of this type have been built only in the United Kingdom (see Section 3.2.2.2).

High-Temperature Reactors. The HTR is helium-cooled with fully ceramic fuel elements that are suitable for the highest coolant outlet temperatures (see Section 3.2.3). Helium is used as coolant because of its chemical inertness. Small particles of enriched uranium dioxide (or UO_2/ ThO_2 or UC_2/ ThC_2), about 1 mm in size, serve as fuel. These particles are coated with a protective layer of pyrocarbon and are embedded in block-shaped or spherical fuel elements made of graphite. The graphite acts as moderator

Table 11. Technical data for typical power reactors [42]

	PWR Biblis B	BWR Krümmel	D_2O-R. Atucha	Magnox Wylfa	AGR Heysham A	HTR * THTR-300 Uentrop	FBR * Creys – Malville
Power, MW							
Electrical (gross)	1300	1316	340	655	666	307	
Electrical (net) P_e	1240	1260	319	590	622	296	1200
Thermal (core) P_{th}	3733	3690	1100	1876	1510	750	3000
Efficiency (net), %	33.2	34.1	29	31.4	41.2	40.5	41.3
Fuel	UO_2	UO_2	UO_2	U	UO_2	UO_2/ThO_2	UO_2/PuO_2
Heavy-metal (hm) content (U, Th, Pu), t	102.7	155.8	38.6	595	110	6.49 t Th, 0.65 t U	32
Enrichment, %	3.0	2.6	natural	natural	2.1 – 2.6	93	14.5/18.5
Average specific power,							
MW/t_{hm}	36.4	23.7	27.3	3.15	13.7	101	93.8
MW/t fissile material	1212	911	3845	444	536	1157	≈ 550
Core, height × diameter, m	3.90 × 3.64	3.71 × 4.99	5.30 × 4.54	9.10 × 17.4	8.2 × 9.3	5.1 × 5.6	1 × 3.66
Average power density, kW/L	92	50.9	12.8	0.87	2.71	6	275
Average burnup,							
MWd/t_{hm}	32 500	27 500	7500	3500	18 000	100 000	70 000
FIFA	1.19	1.16	1.16	0.54	0.84	1.34	≈ 0.6
Conversion rate *CR*	≈ 0.55	≈ 0.6	≈ 0.8	≈ 0.8	≈ 0.6	≈ 0.53	≈ 1.2
Fuel assembly type and shape	16 × 16 – 20	8 × 8 – 1	bundle 36	single rods	bundle 36	sphere	bundle 271
Fuel assembly, distance of centers, cm	22.9	15.3	27.2	19.7	45.7		17.9
Diameter of fuel rod, cm	1.075	1.25	1.19	2.79	1.53	coated particle	0.865
Material of cladding tube/wall thickness, mm	Zry 4; 0.72	Zry 2; 0.85	Zry 4; 0.55	Magnox	steel; 0.38	graphite	steel; 0.7
Specific fuel rod power, average/maximum, W/cm	206/600	198/440	216/575	42/65	170/260		280/450
Maximum fuel temperature, °C	2430	1900	2320	569	1500	1250	2100
Coolant; pressure, MPa	H_2O; 15.5	H_2O; 7.0	D_2O; 11.5	CO_2; 2.8	CO_2; 4.1	He; 4.0	Na; 0.6
Inlet/outlet temperature, °C	290/323	215/286	272/306	247/414	287/651	260/750	395/545
Throughput, t/h	72 000	55 600	23 400	36 948	13 272	1080	59 040
Moderator	H_2O	H_2O	D_2O	graphite	graphite	graphite	
Mass, t/temperature, °C	285/307	295/286	265/190	3740/340	1643/400	141/583	
Control rods							
Material/number	AgInCd/61	B_4C/205	AgInCd/29	BSt/185	BSt/81	B_4C/78	B_4C/24
Steam pressure, MPa/steam temperature, °C in front of the turbine	5.2/266	6.7/282	4.2/253	4.6/401	16.0/538	19.0/535	17.7/487

and, at the same time, transfers heat generated in the fuel to the coolant, which flows through the fuel elements in cooling channels (or around the spherical elements).

Reactors of this type are of long-term interest. Because of the high temperature of the heat produced, it can be used directly as process heat in chemical engineering (see Table 11). Furthermore, changeover from the normally employed water-steam turbine process (with heat exchanger) to the direct helium-gas turbine process is possible. The higher plant costs, compared with those of an H_2O-moderated reactor, as well as problems related to fuel production (see Section 5.3.2.2) and fuel reprocessing, are disadvantages of this reactor type.

Graphite-Moderated, Light-Water-Cooled Reactors (LWGR). The LWGR is built only in the former Soviet Union. The tubular fuel elements consist of a uranium – molybdenum alloy with approximately 3 % of ^{235}U enrichment. They are located in pressure tubes embedded in the graphite moderator. In more recent types, superheated steam is produced within the fuel elements (see Section 3.3).

2.4.2.3. Heavy-Water Reactors (D_2O – R)

The low neutron absorption of heavy water permits the use of natural uranium (0.72 % ^{235}U) instead of the enriched uranium (ca. 3 % ^{235}U) required for H_2O moderation. The expensive uranium enrichment process (→ Uranium, Uranium Alloys, and Uranium Compounds), which can only be carried out in special isotope separation plants (→ Isotopes, Natural), is avoided here. In contrast, substantially more D_2O than H_2O (similar to graphite-moderated reactors) is required for neutron moderation. For this reason, D_2O-moderated reactors are always considerably larger than H_2O-moderated reactors (with the same output). Plant costs are correspondingly higher.

Because much less heavy water is required for cooling purposes than for moderation, the fuel rods are not distributed evenly over the reactor (as in light-water reactors with their fuel elements directly adjacent to each other) but are accommodated as rod-bundle assemblies in individual cooling channels surrounded by D_2O. This functional separation of moderator (D_2O) and coolant allows the use of other coolants as well. Thus, in addition to D_2O, CO_2, H_2O, and organic liquids can be used as coolants, but in these cases, enriched uranium is generally required as fuel (for details, see Section 3.4).

2.4.2.4. Fast Breeder Reactors

Fast breeder reactors (see Table 11) contain no moderator and require fuel with a relatively high enrichment of fissile material (ca. 20 % ^{235}U or ^{239}Pu). The basic advantage of this type of reactor is that is "breeds" more fissile material in the form of ^{239}Pu from ^{238}U than it consumes (see Section 4.1.1).

The only coolant currently used is liquid sodium. In general, a nonactive sodium intermediate cycle is connected between the radioactive primary loop and the water – steam cycle for reasons of safety. Hence, the heat-transfer system is quite complex and the plant costs are high. At present, whether this disadvantage is compensated by correspondingly lower fuel costs is undetermined.

2.4.3. Future Developments

A prognosis for the development of various types of reactors is accompanied by substantial uncertainties. Over the next 15 to 20 years, *light-water reactors* are expected to dominate. Toward the end of this period, however, the high-temperature reactor and the fast breeder reactor could become important again. Regarding the ratio of pressurized-water reactors to boiling-water reactors, a shift in favor of the PWR can be observed (see Fig. 11). This may be due to some small economical and technical advantages of the pressurized-water reactor.

High-temperature reactors have suffered a few serious setbacks (all orders for commercial HTR nuclear power plants were withdrawn in 1975, and operation of the THTR-300 in Uentrop stopped in 1990). The reasons for this were technical plant problems, high plant costs, and unanswered questions regarding the fuel cycle. In view of the production of high-temperature process heat, of the possibility of a direct helium gas turbine process, and of its safety characteristics, the high-temperature reactor is without

competition and is therefore promising. This is confirmed not only by the great interest in this technology but also by the research efforts in some of the manufacturing countries.

Fast breeder reactors will probably become indispensable in the long term because of their breeding properties and because they are relatively insensitive to possibly dramatic increases in the price of uranium. The marketability of this type of reactor will depend essentially on the time when the price of uranium increases or when a possible shortage becomes apparent.

The D_2O-moderated natural uranium reactor is independent of enrichment plants and, for this reason, is expected to retain its market position, as shown in Figure 11, even in the future.

Research in the field of new reactor concepts includes both economic optimization of present reactors (better fuel utilization, recycling of plutonium, and lower plant costs by standardization) and further development of passive safety properties. The latter includes above all the guarantee of inherent automatic shutdown and residual heat removal. Very different concepts are being pursued in individual countries to achieve these goals. They are, however, all based on the principle of passive interruption of the chain reaction and removal of the residual heat via natural convection [45], [73].

Figure 15 shows the countries that use nuclear energy for peaceful purposes (as of January 1996) and the number and gross electric capacity of these reactors.

3. Thermal Reactors

3.1. Light-Water Reactors

Climatic changes are considered to pose a major impact on the environment. Thus the reduction of the emission of greenhouse gases from electricity generation units is in the focus of attention of energy planners, both at governmental and on utilities side. Nuclear power is one option to solve the current and future energy demand and to reduce the emission of greenhouse gases. Units fueled by renewable sources emit 2 – 5 times and those fueled by fossil fuels 40 – 100 times more greenhouse gases than nuclear power plants. By the end of 1995 nuclear power plants generated ca. 17 % of the electric energy worldwide. The total number of nuclear power plants in operation worldwide was 437, 343 of which were light-water reactors. The two main types installed are the pressurized-water reactor (PWR) and the boiling-water reactor (BWR), the PWR dominate in number and power output. Many countries depend heavily on nuclear power for production of electric energy. In some countries nearly 80 % of the electric energy is produced using nuclear power.

A broad experience of operating nuclear power plants is available worldwide, resulting from ca. 7700 reactor years. These units have proven their viability as an reliable, economic, and safe energy source. The operating records show the high availability of these plants, 80 % availability and above became the normal standard. Nuclear power plants are not only operating in base-load mode, in many countries their excellent load-follow capability could be demonstrated, which is fully compatible with that of fossil-fired power plants.

The economic viability for a reactor of a certain size could be attained and is higher than that of conventional plants. The overall energy situation in many countries, the lack of financing, and the stagnant request for electrical power reduces the need for new plants drastically. Programs for installing new nuclear units really exist in India, Japan, the Republic of Korea, Russia, the People's Republic of China, Taiwan and maybe Turkey. In countries such as the United States several nuclear units will be shut down in the next decade. The probalitiy that extensive new nuclear programs will be launched in the future remains questionable.

Even with that background worldwide development programs for advanced reactors are launched to achieve improvements in the next generation of nuclear power plants. The technology and the design of future nuclear power plants will inevitably change to adopt new requirements and to implement the operating experience from current nuclear power plants. The main goals of that development are improvements in economy, reliability, and availability. The safety of nuclear power plants has already reached a very high standard, consequently the space left for further enhancement is rather limited. Nevertheless, the new plants have to be de-

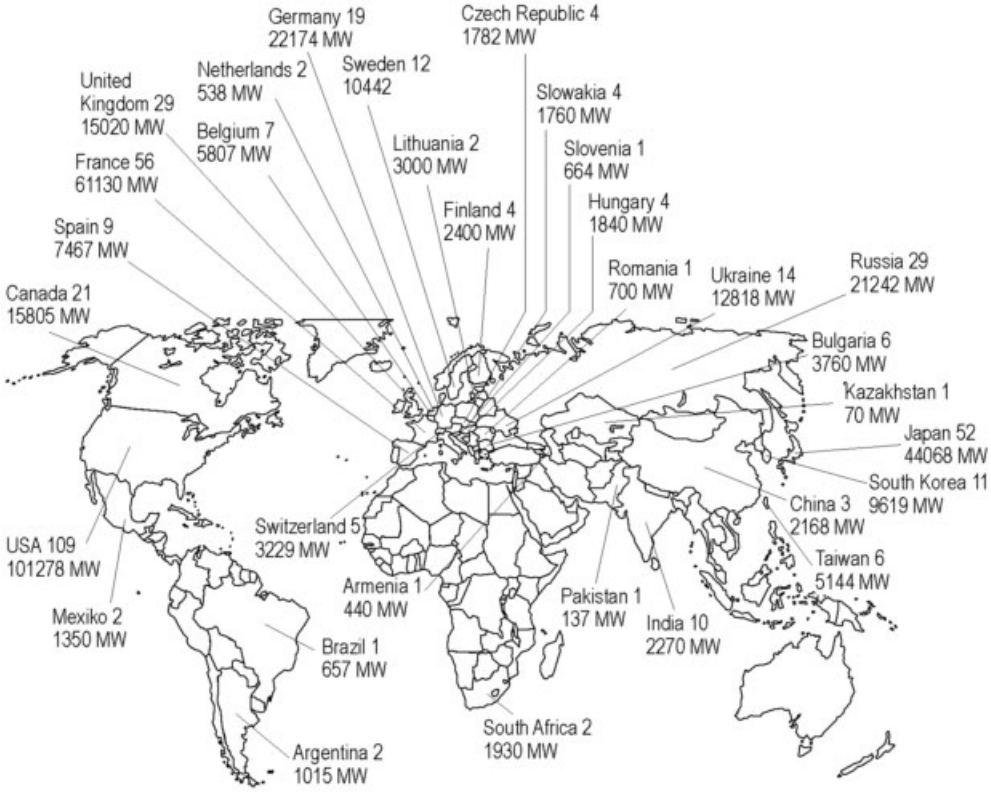

Figure 15. Number and gross electric capacity of nuclear power plants in operation (January 1996) worldwide (for abbreviations see Fig. 11)
Argentina 2 D_2O; Armenia 1 PWR; Belgium 7 PWR; Brazil 1 PWR; Bulgaria 6 PWR; Canada 21 D_2O; China 3 PWR; Czech Republik 4 PWR; Finland 2 PWR, 2 BWR; France 54 PWR, 2 FBR; Germany 13 PWR, 5 BWR; Hungary 4 PWR; India 8 D_2O, 2 BWR; Japan 22 PWR, 28 BWR, 1 GGR, 1 D_2O; Kazakhstan 1 FBR; Korea (South) 10 PWR, 1 D_2O; Lithuania 2 LWGR; Mexico 2 BWR; Netherlands 1 PWR, 1 BWR; Pakistan 1 D_2O; Romania 1 D_2O; Russia 13 PWR, 15 LWGR, 1 FBR; Slovakia 4 PWR; Slovenia 1 PWR; Spain 7 PWR, 2 BWR; South Africa 2 PWR; Sweden 3 PWR, 9 BWR; Switzerland 3 PWR, 2 BWR; Taiwan 2 PWR, 4 BWR; Ukraine 13 PWR, 1 LWGR; United Kingdom 1 PWR, 28 GGR; United States 72 PWR, 37 BWR;

signed to withstand severe accidents (also called beyond-design basis accidents).

The advanced designs can be divided into two different categories. The first type can be called "evolutionary design", these new plants represent descendants of current plant designs. These advanced concepts feature improvements and modifications that are based on operating experience or are aimed at adopting new technological achievements. The second category covers design concepts that incorporate significant departures from current designs and need more engineering efforts, probably including construction and operation of a prototype plant. This category may be named "developmental design". For the evolutionary designs the goals vary from concept to concept and may incorporate features such as:

1) Passive safety features
2) Precautions to cope with severe accidents and its consequences (core melt, hydrogen production, steam explosion, cooling of the melt in the reactor or reactor building, prevention of interactions between core melt with base mat concrete, etc.)
3) Better fuel utilization, increase of discharge burnup, optimization of fuel design, application of burnable absorbers, improvement of neutron economy, use of blanket fuel assemblies

A large variety of advanced reactors are under development and under licensing. Specific

codes and standards specifying the new requirements on top of the utility requirements established in the United States, Europe, Japan, and Korea constitute the basis for the new designs. The safety requirements will mostly correspond to those applied for current designs. The additional or modified safety features incorporated shall prevent and mitigate severe accidents. The further purpose of the utility requirements is to present a clear, complete statement of utilities desire for the next generation of nuclear power plants. Due to the deviations in codes, standards, rules, and regulations a group of European utilities agreed to establish a separate set of requirements for European conditions, which deviate from the EPRI Requirements reflecting the practice in the United States. This may lead to a certain harmonization between the participating European countries. The first joint project established between France and Germany accommodates these European Utility Requirements as a basis.

Table 12 presents an overview of the various concepts currently under development. These efforts enable the future application of nuclear energy on an economic basis together with conventional energy sources such as fossil-fired power plants, renewable sources, and maybe others such as solar energy.

Situation in Germany. In Germany, 19 LWR plants are in operation 1997, with a total capacity of 21 984 MW_e. These 19 plants comprise 13 PWRs and 6 BWRs, with capacities of 15 367 MW_e and 6 619 MW_e, respectively. The PWR has become the preferred reactor type for electricity generation in Germany.

Concurrent with the increase in number of nuclear power plants (NPPs) the need of standardization became evident. Siemens – KWU initiated in 1980, together with the German utilities, experts from the TüV, and the licensing authorities, the Konvoi concept. For a series of successive nuclear power plants, the procedure provided for uniform planning using identical software and hardware for all site-independent equipment. As a consequence, construction schedules and costs could be stabilized, and the construction time reduced. The Konvoi plants went into operation between three and seven months ahead of schedule. This generation of German PWRs already incorporated "advanced" technology achievements that comply with the development targets specified for many advanced or next generation PWR plant designs.

Based on the experience gained from the operation of the German PWRs, in particular with the Konvoi plants, and from the huge PWR program in France, the two countries have agreed upon a joint development program for a PWR called the European Pressurized Water Reactor (EPR) (see Section 3.1.5). This development project was started by Siemens, the French supplier Framatome, and their jointly owned company Nuclear Power International (NPI), together with the French utility Electricité de France (EdF) and a group of German utilities. The respective national authorities were also involved from the very beginning.

The German utilities and Siemens have launched a further project, the development of a 1000 MW_e BWR with passive safety systems. The intent of this project is to establish an advanced BWR design that is based on the experience gained from current BWRs while taking into account new safety requirements, to be available for the utilities as an alternative to the PWR. Traditional safety systems will largely be replaced by passive systems to permit simplification of systems and structures, in order to improve the economic viability.

The following description of the PWR-technology is based on the Konvoi, a four-loop 1300 MW_e PWR, contracted for three independent sites in Germany. Because of the various codes, standards, rules, and regulations applied in the different countries the technology

Table 12. Survey of advanced light-water reactor designs under development

Plant group	Name	Type	Size, MW_e	Supplier/designer	Design status
Evolutionary, large-size plants	APWR	PWR	1300	Westinghouse, United States and Mitsubishi, Japan	basic design [c]?
	ABWR	BWR	1300	General Electric, United States	detailed design
	BWR 90	BWR	1200	ABB Atom, Sweden	detailed design
	EPR	PWR	1500	Nuclear Power International (NPI)	basic design
	ESBWR [a]	BWR	1190	General Electric, United States	preliminary [d] design
	KNGR	PWR	1350	Korea Electric Power Corp., Republic of Korea	basic design
	Sizewell C	PWR	1250	National Nuclear Corp. (NNC), United Kingdom	conceptual design [e]
	System 80+	PWR	1350	ABB Combustion Engineering, United States	detailed design
	SWR 1000	BWR	1000	Siemens, Germany	conceptual design
	VVER-1000 (V-392)	PWR	1000	Atomenergoexport/Gidropress, Russia	basic design
Evolutionary, medium-size plants	AC-600	PWR	600	China National Nuclear Corp. (CNNC), People's Republic of China	conceptual design
	AP-600 [b]	PWR	600	Westinghouse, United States	detailed design
	HSBWR	BWR	600	Hitachi Ltd., Japan	basic design
	MS-600	PWR	600	Mitsubishi, Japan	detailed design
	VVER-640 (V-407)	PWR	640	Atomenergoexport/Gidropress, Russia	detailed design
Design concepts requiring substantial development	ISIS	PWR	300	Ansaldo Spa., Italy	conceptual design
	JPSR	PWR	630	Japan Atomic Energy Research Institute (JAERI), Japan	design study
	PIUS	PWR	650	ABB Atom, Sweden	basic design
	SPWR	PWR	600	Japan Atomic Energy Research Institute (JAERI), Japan	conceptual design
	VPBER-600	PWR	630	OKBM, Russia	conceptual design

[a] ESBWR (and JSBWR in Japan) is enlarged version of GE's SBWR design. [b] Enlarged versions (EPP in Europe & SPWR in Japan [M]) are being developed.
[c] Basic design mainly covers the concept plus general design and layout of the system components. [d] Preliminary design is less than a concept, only the main data such as power output and other general data are fixed. [e] Conceptual design describres the fundamentals (cf. Chap. 1).

Table 13. Selection of pressurized – water reactors in commercial operation (gross capacity $\geq 1000\,MW_e$) in several industrialized countries (source IAEA, April 1996)

Plant	Gross capacity, MW_e	Commercial operation
Federal Republic of Germany		
Biblis A	1204	1975
Biblis B	1300	1977
Brokdorf	1395	1986
Emsland	1363	1988
Grafenrheinfeld	1345	1982
Grohnde	1394	1985
Isar 2	1410	1988
Mülheim-Kärlich	1302	1987
Neckarwestheim 2	1385	1989
Phillippsburg 2	1390	1985
Unterweser	1320	1979
France		
Belleville-1	1363	1988
Belleville-2	1363	1989
Cattenom-1	1362	1987
Cattenom-2	1362	1988
Cattenom-3	1362	1991
Cattenom-4	1362	1992
Flamanville-1	1382	1986

Table 13. (Continued)

Plant	Gross capacity, MW_e	Commercial operation
United States		
Braidwood-1	1175	1988
Braidwood-2	1175	1988
Byron-1	1175	1985
Byron-2	1175	1987
Callaway-1	1232	1984
Catawba-1	1192	1985
Catawba-2	1192	1986
Commanche Peak-1	1181	1990
Commanche Peak-2	1181	1993
Diablo Canyon-1	1124	1985
Diablo Canyon-2	1137	1986
Donald Cook-1	1056	1975
Donald Cook-2	1100	1978
Indian Point-3	1000	1976
McGuire-1	1220	1981
McGuire-2	1220	1984
Millstone-3	1184	1986
Palo Verde-1	1303	1986
Palo Verde-2	1303	1986
Palo Verde-3	1303	1988
Salem-1	1149	1977
Salem-2	1149	1981
San Onofre-2	1127	1983
San Onofre-3	1127	1984
Seabrook-1	1200	1990
Sequoyah-1	1162	1981
Sequoyah-2	1162	1982
South Texas-1	1312	1988
South Texas-2	1312	1989
Trojan	1122	1976
Vogtle-1	1159	1987
Vogtle-2	1163	1989
Waterford-3	1120	1985
Zion-1	1085	1973
Zion-2	1085	1974

used and the safety approach also differ from country to country. The Russian technology of the WWER-plants compared to western PWRs demonstrates the huge differences between the plant designs. A detailed analysis of different designs would be beyond the scope of this contribution.

3.1.1. Pressurized-Water Reactors

In 1995, a total of 250 PWRs were in operation worldwide (see Table 13; including the Russian WWER) with a total capacity of 220 395 MW_e (≈ 64 % of the total nuclear capacity installed). In Germany the nuclear program consist of 19 LWR from which 13 NPP are of the PWR type. The German PWR design was developed by Siemens (KWU) in accordance with German requirements (see Section 3.1).

The following description of the PWR technology is based on the Konvoi, a four-loop 1300 MW_e PWR, contracted for three independent sites in Germany. Because of the various codes, standards, rules, and regulations applied in the different countries, the technology used and the safety approach also differ from country to country. The Russian technology of the WWWer plants compared to western PWRs demonstrates the huge differences between the plant designs. A detailed analysis of different designs would be beyond the scope of this contribution

3.1.1.1. Design [77]

The essential feature of pressurized-water reactors, as distinguished from boiling-water reactors, is the closed reactor coolant system that is separated from the secondary water – steam cycle. The demineralized water used in the reactor coolant system is maintained at a pressure of 15.7 MPa and an inlet temperature of 291.3 °C. Water is heated on flowing through the reactor core to an average outlet temperature of 326.1 °C. Because of the high pressure, the heat transferred from the fuel assemblies to the reactor coolant does not cause coolant evaporation in the reactor core. Depending on the capacity of the plant, the reactor coolant system may have two, three, or four identical loops, each consisting of a steam generator, a reactor coolant pump, and interconnecting piping between the components. A pressurizer is connected to one loop to regulate the pressure in the reactor coolant system. The pressurizer is partly filled with water, the balance being steam. During operation the pressure of the steam can be either raised by means of electrical heaters in the pressurizer bottom or lowered by injecting water through sprays in the head. Overpressures during transients are reduced by relief valves and safety valves, which blow steam down into the pressurizer relief tank. There, the steam condenses by direct contact with the water contained in the tank.

Figure 16 shows the isometric arrangement of the reactor coolant system in a four-loop PWR plant. PWR-Konvoi plants in Germany

are currently operating with a power output of 1450 MW_e. The next generation of advanced PWR plants, accommodates an electric power output of 1500 MW_e, which corresponds to a thermal reactor power of 4250 MW_{th}. To increase the economic viability of nuclear energy in comparsion to conventional power plants, these large units are favored. The increase in power is accepted by the licensing authorities because of the very conservative average linear heat rate. Various measures are incorporated in the new designs, i.e., fuel management procedures and advanced fuel assemblies which lead to higher discharge burnup of $\approx 60\,000$ MW · d/t.

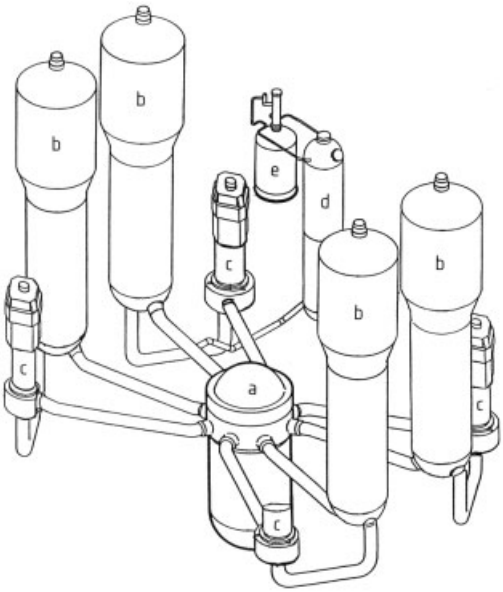

Figure 16. Four-loop primary coolant system of a 1300-MW pressurized-water reactor
a) Reactor pressure vessel; b) Steam generator; c) Reactor coolant pump; d) Pressurizer; e) Pressurizer relief tank

If grid expansion is planned, the new units should not have a larger capacity than ca. 10 % of the total power installed. In addition, a redundancy for the new nuclear unit (i.e., a second nuclear power plant) of the same size should be planned to cover outages for maintenance and other nonavailability reasons. This strategy may limit the installation of large units. For these cases nuclear power plants with an output of 600 MW_e or 1000 MW_e are available. The plant size not only depends on economic aspects, also the forecast for the future energy market is essential. The development of the local industry and the dependency on only a single energy resource may promote the application of small and medium nuclear power plants.

In countries with a total grid capacity $< 15\,000 - 20\,000$ MW, larger units cannot be incorporated into the grid, especially, if only smaller conventional units exist as redundancies. For grid expansion planning it is essential to have sufficient redundancy available in case the larger nuclear unit is shut down for some reasons.

The thermal flow diagram of a 1300 – 1400-MW four-loop plant is shown in Figure 17; characteristic data are given in Table 14. The reactor coolant is heated in the reactor core (a) and transfers its heat in the steam generator (b) to the secondary-side feedwater, which is converted into steam. The saturated main steam thus produced passes at 70 MPa through the main steam line to the saturated steam turbine (f). After releasing its enthalpy content in the turbine, the exhaust steam condenses in the condenser (i). The condensate is pumped through the low-pressure heaters (k) into the feedwater tank (m). Feedwater pumps (n) transfer feedwater through the high-pressure heaters (o) to the secondary side of the steam generator. A moisture separator (g)/reheater (h) between the high-pressure and low-pressure sections of the turbine improves the efficiency of the plant. Furthermore, this vessel prevents droplet impact erosion damage to the low-pressure turbine.

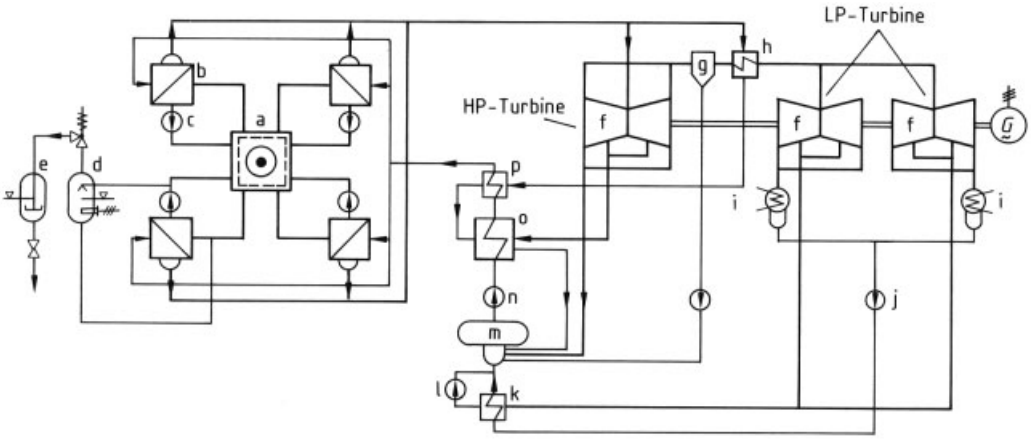

Figure 17. Schematic of the reactor coolant system and water – steam cycle of a four-loop pressurized-water reactor nuclear power plant
a) Reactor; b) Steam generator; c) Reactor coolant pump; d) Pressurizer; e) Relief tank; f) Turbine generator set; g) Moisture separator; h) Reheater; i) Condenser; j) Main condensate pump; k) Low-pressure feedwater heating system; l) Heater drains pump; m) Feedwater tank; n) Main feedwater pump; o) High-pressure feedwater heating system; p) Drains cooler

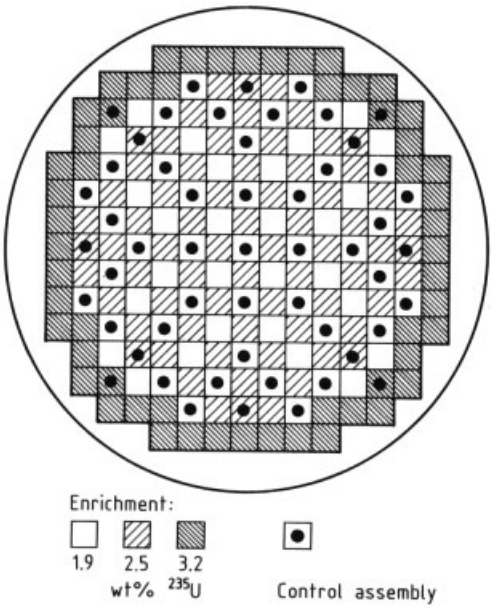

Figure 18. Positions of control assemblies and arrangement of fuel assemblies of different enrichment levels in the first core

Table 14. Characteristic data of a 1300-MW pressurized-water reactor

Overall plant	
Generator terminal output, MW	1390
Net electrical output, MW	1310
Steam generator thermal output, MW	3782
No. of reactor coolant loops	4
Net efficiency, %	34.1
Reactor coolant mass flow, kg/s	18 800
Average coolant temperature, °C	319
Coolant pressure, MPa	15.7
Main steam mass flow rate, kg/s	2056
Main steam temperature, °C	280.3
Main steam pressure, MPa	6.35
Final feedwater temperature, °C	218
Reactor core	
Mean linear heat generation rate, W/cm	162.9
Mean heat flux, W/cm^2	54.5
Mean specific power, kW/kg	37.1
Mean power density, kW/L	93.3
^{235}U enrichment in first core, wt %	1.9/2.5/3.2
^{235}U enrichment in equilibrium core, wt %	3.2/3.4
Uranium mass, t	103
Fuel assemblies	
No. of fuel assemblies	193
Total length, mm	4827
Cross-section, mm	230×230
No. of rods	300
No. of control rod guide thimbles	24
No. of spacers	9
Cladding material	Zircaloy 4
Active core length, mm	3900
Control rods	
No. of control rods	61
No. of control rod fingers per rod	24
Cladding material	X 10CrNiNb 189
Absorber material	Ag80In15Cd5
Positioning rate, mm/min	375/750
Step length, mm	10
Reactor pressure vessel	
Inside diameter, mm	5000
Wall thickness plus cladding, mm	250+6
Total height, mm	12 012
Mass, t	500
Steam generator	
Heat exchange surface, m^2	4×5400
Tube diameter × wall thickness, mm	22×1.5
U-tube material	Incoloy 800
Outside diameter, mm	4812
Total height, mm	21 300
Mass, t	420
Reactor coolant pumps	
Capacity, kg/s	4×4700
Delivery head, MPa	0.65
Motor output (cold), kW	7350
Speed, rpm	1480
Pressurizer	
Total volume, m^3	65
Installed heater output, kW	2000
Inside diameter of vessel, mm	2600
Total inside height, mm	13 200
Shipping mass, t	140
Containment	
Sphere diameter, m	56
Design pressure, MPa	0.53
Wall thickness, mm	38

Separation of the reactor coolant system from the water – steam cycle not only allows the secondary system to be kept free of radioactivity while in operation but, above all, permits independent selectivity of water chemistry and materials for the primary and secondary systems. High water purity is required on the primary side to ensure a low corrosion rate. On the secondary side, proper selection of materials, leaktight components such as the condenser, and correct water chemistry in accordance with the specification are important in preventing the transport of corrosion products into the secondary side of the steam generator.

Reactor Core. The reactor core, which is located in the reactor pressure vessel (RPV), is the heat source for the nuclear power plant. It comprises square fuel assemblies with a square cross-section, which are arranged in an approximately circular array in the RPV. The core of a 1300-MW pressurized-water reactor contains 193 fuel assemblies. The nuclear fuel is usually uranium dioxide with an equilibrium ^{235}U enrichment of 3.2 – 3.4 wt %. Mixed uranium – plutonium oxide (MOX) can be used instead of uranium dioxide. For long-term reactivity changes, the chain reaction is controlled by dissolution of boric acid in the coolant; for short-term changes, control assemblies are used. Gadolinium is incorporated in the fuel of the first core to compensate for a certain amount of excess reactivity so as to maintain a negative fuel temperature coefficient. Since the power density distribution over the cross-section of the first core, and subsequently that of the equilibrium core, should be as even as possible and because there is no burnup at first core, fuel assemblies with different enrichments (e.g., 1.9, 2.5, 3.2 wt % ^{235}U) are used (Fig. 18).

Thermal design of the core must ensure that no damage can occur to the fuel cladding during all feasible modes of operation of the plant. Furthermore, the prevention of film boiling is a critical variable (DNB ratio; see Section 2.2). A theoretical "hot channel" is defined to establish a minimum DNB ratio. In the definition of this channel, all phenomena are considered that increase local heat flux in the core above the average value. The core is then designed in such a way that the hot channel withstands every conceivable adverse operating condition. This design approach ensures a high safety standard, because all phenomena tending to increase heat flux in the core are extremely unlikely to occur at the same place at the same time. Another crucial physical variable with respect to economic operation is the average fuel assembly burnup (i.e., the average total energy generated per unit mass). Today's modern pressurized-water reactors can achieve discharge burnups of 45 000 – 50 000 MW · d per tonne of uranium within 3 to 4 years of operation.

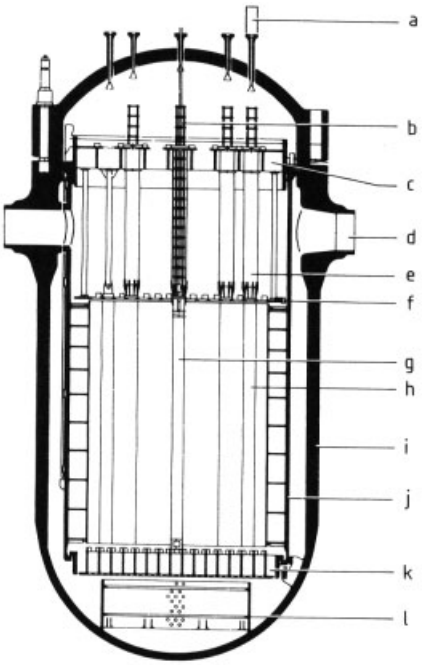

Figure 19. Reactor pressure vessel and internals
a) Control drive mechanism; b) Control rod guide assembly; c) Upper core support; d) Coolant inlet nozzle; e) Support column; f) Upper core plate; g) Fuel assembly; h) Core shroud; i) Pressure vessel; j) Core barrel; k) Lower core support; l) Flow skirt

Fuel Assemblies. For fuel fabrication, powdered UO_2 is pressed into *pellets* of ca. 10 mm in height and diameter. The pellets are then sintered and filled into ca. 4-m long cladding tubes. The *fuel rods* thus formed are pressurized to approximately 2.0 MPa to reduce the mechanical loading during operation and then sealed gastight with an end cap (for details on fuel fabrication, see Section 5.3.1). A number of fuel rods are bundled together to form fuel assemblies; the quantity of fuel rods depends on the power output of the reactor. A square bundle containing 14×14 to 18×18 fuel rods is generally used today. The bundle is open on all sides and forms a structural unit with the control assembly.

The *fuel-assembly structure* consists of several spacer grids that are welded to the control rod guide thimbles at regular intervals between the top and bottom end fittings. The fuel rods are held elastically in this skeleton and are free to expand and contract axially. Mixing vanes on the spacers prevent laminar flow and local superheating at the cladding. They equalize heating of the coolant in its passage through the reactor core, and a high power density can be achieved without approaching critical states. The *first core* configuration is zoned with differently enriched fuel assemblies to attain as broad a peak power distribution (flat Gaussian distribution) as possible. *Reload fuel assemblies* have a uniform enrichment which depend on the length of the fuel cycle (e.g., 3.2 or 3.4 % ^{235}U)

Absorber rods located in the control rod guide thimbles in the fuel-assembly structure are bundled to form *control assemblies*. Each of these is moved up and down by an electromagnetic jack mechanism on the head of the reactor pressure vessel.

Components of the Reactor Coolant System. The *reactor pressure vessel* (Fig. 19) with its internals accommodates the reactor core and channels the coolant. The coolant enters the vessel through the inlet nozzle (d), flows downward through the annular gap between the core barrel (j) and the pressure vessel wall (i), is reversed in the bottom head and flows upward through the reactor core, leaving the vessel through the outlet nozzle without further contact with the wall of the pressure vessel. The coolant inlet and outlet nozzles are the only penetration through the

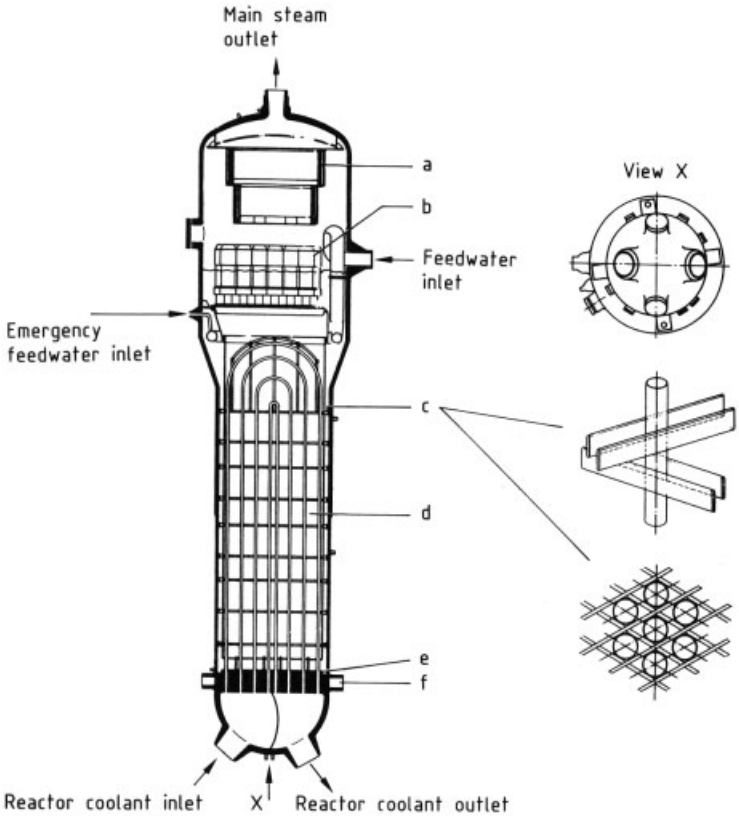

Figure 20. Steam generator
a) Steam dryer; b) Moisture separator; c) Tube spacer grids; d) Heating tube bundle; e) Tube sheet; f) Support

wall in the body of the RPV. Thus, refilling of the reactor core by the emergency core cooling and residual heat removal systems is ensured under all circumstances, even in the event of a reactor coolant pipe rupture. No penetrations are located in the lower part of the vessel, thus no breakage must be considered.

The RPV is the central fixed point of the nuclear steam supply system. Support lugs and horizontal stops that are welded to the shell flange transfer the load of the vessel to the building structure. The latter is designed to absorb all forces and moments during operation and hypothetical accident conditions.

The pressure vessel is made of seamless forged rings. It has a spherical bottom head and a flange ring at the top. This forged flange ring accommodates the coolant nozzles and forms the mating flange for the RPV head. A flange ring and a dome plate are welded together to form the pressure vessel head. This dome plate contains the nozzles for the control assemblies and the in-core instrumentation. Concentric, metal O-rings are used to seal the flange surfaces of the head and body of the reactor pressure vessel; 52 necked down studs hold the two parts together. The studs are uniformly preloaded with a hydraulic or electrical stud tensioner. The base material for the RPV walls is 20 MnMoNi55 steel, which contains a welded two-layer austenitic strip corrosion-resistant cladding. The vessel is surrounded on the outside by self-supporting thermal insulation, which provides unrestricted access from the outside for later in-service inspection. The insulation around the head flange is removable.

The *core structure*, the basic structural framework of the reactor core, consists of (1) the lower core structure and (2) the upper core structure with control rod guide assemblies (b).

The core structure has the following functions:

1) Taking the weight and restraint forces of the fuel assemblies
2) Alignment and location of the fuel and control assemblies
3) Accommodation of the thrust of the control assemblies on the reactor trip
4) Distribution of the reactor coolant flow
5) Support of in-core instrumentation
6) Reduction of the neutron irradiation of the pressure vessel wall
7) Accommodation of irradiation specimen for accelerated monitoring of the properties of the pressure vessel material

The core structure is built in such a way that it can be removed as a whole from the reactor pressure vessel. The upper core structure is removed during refueling. The upper core structure, or even the complete core structure, can be stored in a designated setdown area in the reactor well.

The *lower core structure* consists of the cylindrical core barrel (j). The core barrel houses the polygonal core shroud that surrounds the core.

The lower core structure is closed at the bottom by the lower core support (k) and the flow distribution plate. The core barrel, centered by shims, is suspended from the ledge on the pressure vessel flange. The part of the barrel shell around the core belt line serves simultaneously as a thermal shield. Lateral stops limit radial displacement.

The lower core support (k) is the actual support structure and is integrated into the lower section of the core barrel. The flow distribution plate, together with the flow skirt (l), provides uniform coolant flow through the reactor core.

Irradiation channels for the housing of irradiation specimens for accelerated monitoring of the properties of the RPV material are fastened to the outer wall of the core barrel close to the outlet nozzle.

The *upper core structure* is located in the nozzle section of the core barrel and thus closes the reactor core at the top. It consists of the upper core support and top plate, support columns, control rod guide assemblies (b), and the grid plate (c). The core barrel and the upper core structure are held by hold-down assemblies with Belleville springs between the flanges of the reactor pressure vessel.

On refueling, the upper core structure is removed, together with the disconnected control assembly drive rods (a), with the aid of a lift rig from the flooded pressure vessel. Guide rods in the pressure vessel and centering pins on the core barrel ensure proper alignment during the lifting process. The height of the upper core structure is largely determined by the length of the control assemblies because these are almost completely withdrawn from the core during operation. Austenitic materials are used exclusively for the core internals.

The heat released in the reactor core is transferred by the *steam generator* (Fig. 20) from the reactor coolant to the feedwater – steam cycle. Thus, the steam generator forms the barrier between the radioactive and nonradioactive cycles. It is a vertical U-tube heat exchanger with natural circulation on the secondary side. The reactor coolant enters the hemispherical primary channel head through the inlet nozzle, flows through the ca. 4000 U-tubes where it releases its heat content and then leaves the steam generator through the outlet nozzle.

Feedwater enters the steam generator through an inlet nozzle in the steam dome and is distributed uniformly around the circumference by a sparger ring arranged around the cylindrical tube shroud. The feedwater mixes with the water returned from the moisture separator (b) passes into the downcomer, and from there flows radially into the tube bundle above the tube sheet. An annular flow restrictor in the downcomer produces the pressure drop required for stable natural circulation in the single-phase regime. The steam – water mixture above the tube bundle passes through the moisture separator (b) and steam dryer (a). The water separated from the steam then flows back into the downcomer and mixes with the freshly supplied feedwater. The recirculation coefficient (the ratio of the recirculation mass flow to feedwater mass flow) is ca. 4. Dried steam exits through the outlet nozzle in the steam dome and flows to the turbine along the main steam line.

To attain maximum leaktightness, the steam generator tubes are not only welded to the primary-side cladding but also tightly rolled into the tube sheet. The walls of the vessel are made of 20MnMoNi55X forged steel; material 1.4858 (similar to Incoloy 800) is used for the heat-exchange tubes. This material is especially re-

sistant to intergranular and transgranular stress corrosion cracking.

Flow through the reactor coolant system occurs by forced circulation. One *reactor coolant pump* (Fig. 21) is installed in each loop to overcome pressure losses in the reactor pressure vessel, the steam generator, and the connecting piping.

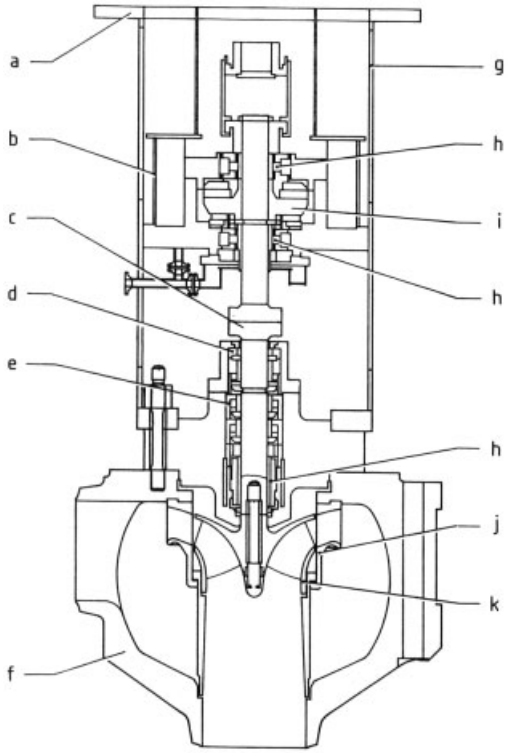

Figure 21. Reactor coolant pump with integrated oil supply a) Motor flange; b) Oil cooler; c) Shaft coupling; d) Standstill seal; e) Hydrodynamic shaft seal (three-stage); f) Pump casing; g) Lantern; h) Radial bearing; i) Axial bearing; j) Diffuser; k) Impeller

A three-stage hydrodynamic shaft seal (e) prevents uncontrolled leakages from the primary system to the reactor building. The system pressure in the pump casing is reduced to atmospheric pressure by this three-stage hydrodynamic shaft seal. Purified barrier water is the medium used for the hydrodynamic seals. The high-pressure leak-off is discharged into the volume control system, and the low pressure leak-off is passed into the equipment drain system.

If all three seal stages fail at the same time, a separate standstill seal (d) is inserted by nitrogen pressure. This seal is able to withstand the full system pressure. Extensive instrumentation is provided for controlling the pump units in operation. The electrical data of the motor, the temperatures and pressures in the bearings and seals, and, lubricating oil and leak-off flows are constantly monitored.

The *pressurizer* (Fig. 22) builds up the pressure required for operation of the reactor coolant system. It also limits and compensates for coolant volume fluctuations without causing substantial pressure changes. Fluctuations in coolant volume occur because of changes in system temperature following load changes. The pressurizer is partly filled with boiling water with a steam cushion above.

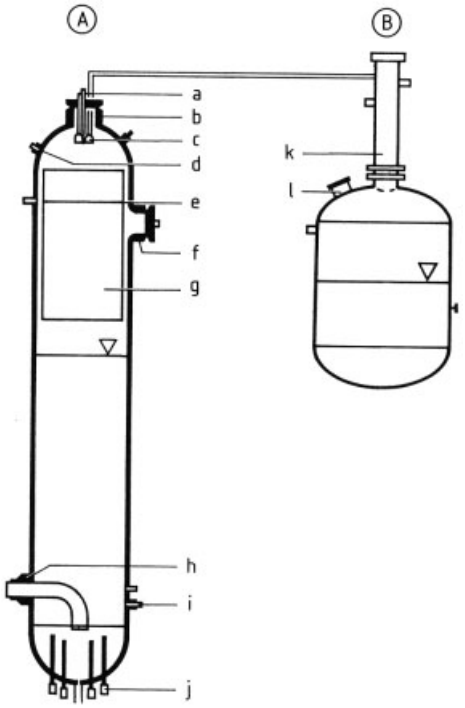

Figure 22. A) Pressurizer; B) Pressurizer relief tank
a) Vent nozzle; b) Central spray nozzle; c) Box-type spray manifold and spray nozzles; d) Nozzle for safety and relief valves; e) Instrumentation nozzle; f) Manhole; g) Spray skirt; h) Surge line nozzle; i) Instrumentation nozzle; j) Heater rods; k) Steam dome; l) Rupture diaphragm

Electric heaters, spray nozzles, and a pressure relief system are used for pressure control. When the pressure drops, the electric heater (j) is switched on so that water evaporates and the pressure returns to its set point. When the pres-

sure rises, either plenum water is sprayed into the steam space to condense steam, or the relief valves (d) are opened to discharge steam, thus lowering the pressure to its set point. The changes in pressure in the reactor coolant system are communicated to the pressurizer by the surge line (h). The steam blown down by the relief valves is condensed in a body of water in the pressurizer relief tank (B) which is connected to the pressurizer. This vessel is not exposed to the reactor coolant system pressure it is kept at a low pressure.

Reactor Building. (Fig. 23) The reactor building contains the reactor system, all other high-pressure and fuel handling systems, parts of the safety systems, ventilation systems, and instrumentation. The entire building is part of the controlled area. The reactor building is a double containment structure that consists of an inner, free-standing steel sphere (the containment proper) surrounded by a concrete shell that provides secondary shielding. The space between the steel sphere and the concrete shell, the *annulus*, accommodates (1) the emergency core cooling and residual heat removal systems, (2) the component cooling systems, (3) the extraborating system, and (4) the high-pressure pumps of the volume control system. With a wall thickness of ca. 2.00 m, the concrete building is designed to withstand airplane crashes, earthquakes, sabotage, gas explosions, flooding, and the radiological loading produced during a design basis accident.

The *containment interior* is divided into service and equipment compartments. The latter are separated from the steel containment by the missile shield (i). A distinction is made between compartments that are and those that are not accessible at any time during operation. The plant compartment is maintained at a slightly subatmospheric pressure compared to the service compartments. These compartments are separated by an airtight seal. Thus, if leakages occur in systems, no radioactivity can escape into the accessible compartments. Four airlocks lead into the containment interior, i.e., the normal personnel airlock connecting the auxiliary building with the service compartments, two emergency airlocks, and an equipment airlock (j). The gantry crane that is used to transport shipping casks for new and spent fuel assemblies into and out of the building is located outside the equipment airlock. The spherical geometry provides sufficient space on the service level for the storage of refueling equipment and for maintenance and repairs, as necessary.

If the design basis accident occurred, pressure equalization ports would equalize the pressures in the equipment and the service compartments. The concrete walls are dimensioned to withstand a certain pressure difference and the jet impingement forces that would occur. The full-pressure containment vessel is designed for ca. 0.53 MPa and 145 °C, i.e., the conditions calculated for a design basis accident. Additionally, a containment venting system consisting of relief valve, filter, piping, and controls is provided to protect the containment in the highly improbable event of a core melt accident.

Reactor Auxiliary and Supporting Systems. Trouble-free operation of the reactor coolant system is not possible without a variety of auxiliary and supporting systems. Besides fulfilling the requirements associated with plant operation (e.g., supply, purification and treatment of coolant), some of these systems are essential to meet the stringent requirements pertaining to the internal and external safety of the overall plant and to environmental protection. The important criteria for design of the auxiliary and supporting systems are dictated by the reactor coolant system; together they all form the nuclear steam supply system.

In particular, the *emergency core cooling and residual heat removal system*, which ensures cooling of the core when the reactor is shut down and in the event of loss- of- coolant accidents, is important for plant safety. The *volume control system* is the link between the hot, high-pressure reactor coolant system and the low-pressure systems of the reactor auxiliary systems. It interacts with the pressurizer to equalize fluctuations in coolant volume caused by temperature and operational changes, and with the *chemical control system* to adjust the coolant chemistry to specified levels. To maintain high water purity, the volume control system continually releases a part-flow of the coolant circulating in the reactor, conveys it to the *coolant purification system* for removal of fission and corrosion products, and subsequently feeds it back into the reactor coolant system. Various *processing systems* pro-

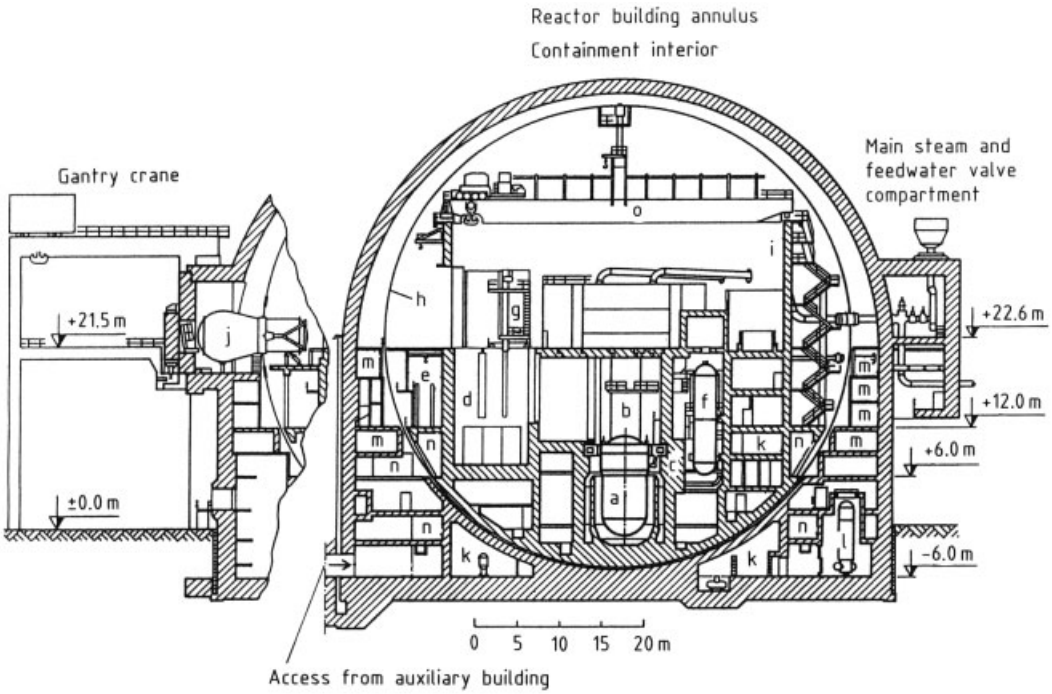

Figure 23. Cross-section of a reactor building
a) Reactor pressure vessel; b) Reactor well; c) Biological shield; d) Fuel pool; e) New fuel store; f) Pressurizer; g) Refueling machine; h) Containment; i) Missile shield; j) Equipment airlock; k) Equipment and valve compartments; l) Component cooling heat exchanger; m) Cable duct; n) Pipe duct; o) Polar crane

tect the environment by ensuring that gaseous, liquid, and solid radioactive releases are kept below stipulated limits.

3.1.1.2. Instrumentation and Control [78], [79]

Instrumentation and control (I & C) functions in nuclear power plants include measurement, control, regulation, monitoring, and protection. The operational functions of the overall plant are based on the physical, chemical, process engineering, and operational requirements placed on the systems and plant equipment. The instrumentation and control system can be subdivided into safety and operational I & C. Plant and environmental protection are assigned to the safety systems. These systems operate predominantly on demand, and most are inactive during normal, trouble-free operation. The operability of this redundantly structured system is continually checked by self-monitoring and periodic function testing. The reactor power controls are usually considered to be part of the safety I & C. The operational I & C includes all systems that ensure trouble-free normal operation of the nuclear power plant and has a decentralized and hierarchical structure. That is why it can be divided into functional groups and complexes in line with the process functions.

Safety I & C. Safety instrumentation and control are shown in Figure 24 and [80].

Reactor Control. The function of the reactor power controls is to match reactor power to the generator output demand at any time. This is carried out by measuring relevant control variables, such as neutron flux, power density, and coolant temperature. The reactor power controls also regulate pressures, water levels (in, e.g., steam generator, pressurizer), flow rates, and temperatures.

The *generator load controller* controls steam extraction by the turbines in accordance with the power demand of the electric grid. Reac-

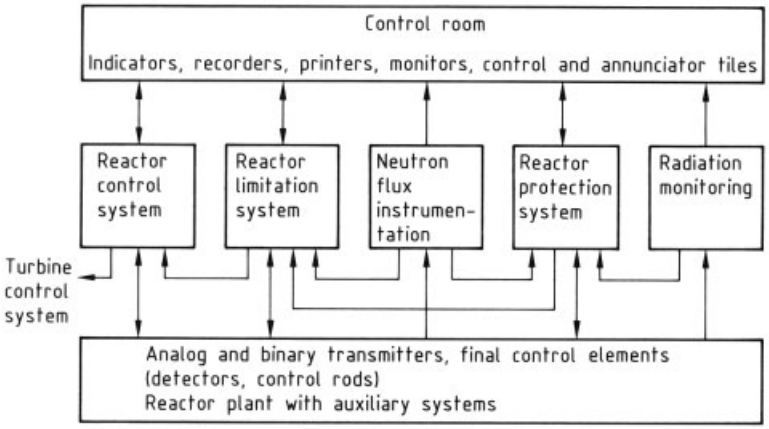

Figure 24. Safety instrumentation and control

tor power controls use the self-regulation of the reactor to match its power to generator output. An increase in steam extraction causes a gradual decrease of coolant temperature in the steam generators. Because of the negative temperature coefficient of the reactor reactivity, which sets in after a certain period in service, the output of the reactor increases. Thus, part of the increased steam demand of the turbine is automatically met. This process is reversed when turbine load is reduced.

The *reactor power controls* immediately intervene in the case of large and rapid load changes in the reactor plant. Two modes of control are possible:

1) Operation at constant main steam pressure. For the water – steam cycle, this has the advantage of a reduction in design pressure and turbine throttling at partial load.
2) Operation at constant average coolant temperature. This is the reactor-oriented mode (small control rod movements because of no additional feedback effect on the steady-state coolant temperature when load is changed; largely constant water level in the pressurizer because of only small changes in coolant volume; possibility of heat storage in the steam generators at partial load because pressure increases with decreasing load).

As shown in the part load diagram (Fig. 25), the Konvoi plants use a combination of the two modes.

In addition to these considerations, the reduction of thermal loadings on the reactor pressure vessel in the normal output range of 50 – 100 % is crucial to the life expectancy of the plant.

Reactor power is controlled by neutron absorbers. Control assemblies variously grouped in banks effect rapid changes in reactivity. Slow changes (such as compensation of xenon buildup and burnup) are achieved by altering the concentration of the boric acid dissolved in the coolant by adding either boric acid or demineralized water. As mentioned in Section 3.1.1.1, addition of gadolinium to the fuel compensates for the initial excess reactivity present in the first core. Gadolinium is burned up in the first operating cycle and removed from the core with the spent fuel assemblies during the first refueling outage.

Reactor Protection. The task of the reactor protection system is to reliably detect all design basis accidents and to initiate highly reliable protective actions that ensure the safety of the plant at any time. Accidents are sequences of events dictating that plant operation must be discontinued for safety reasons. An analog data acquisition system detects accident-specific process variables, and logic gating circuits generate actuation signals when certain limits are exceeded. These actuation signals initiate reactor trip and, taking priority over all operational controls, activate engineered safety features required for accident control. In all cases, accident detection is based on different (diverse) process variables to eliminate uncertainties in the accident detection system and common-mode failures in the data acquisition system. Redundancy is provided to mitigate the effects of initiating events within the reactor protection system. Individual redun-

dancies are physically and spatially separated and are accommodated in buildings designed to withstand external events. The reactor protection system panel gives operating personnel an overview of the state of the reactor protection system at all times during normal operation and accident conditions. Periodic testing is performed by means of special testing equipment. Active and passive failures in the initiation channels are self-reporting.

Table 15. Number of permanently installed radiation monitors in a Konvoi plant

Monitored item	No. of monitors
Local dose rate	15
of which high-dose	2
Plant and systems	41
of which in reactor protection system	24
Area	16
Additional samplers	3
Radioactive releases	10
of which accident instrumentation	3
additional samplers	5
Total continuous monitors	82
of which accident monitors reactor protection system	5
Monitors	24
and continuous samplers	8

Reactor Limitation Functions. In the event of deviation of process variables from their set points that cannot be corrected by the controls, limitation systems trigger certain countermeasures. The latter usually make shutdown unnecessary so that the plant can stay in operation. These limitation systems also include the core protection system and act in tandem with the controls and the reactor protection system to provide extensive safeguards for the I & C system and to increase the safety and availability of the plant.

Radiological Protection and Monitoring [81]. The function of radiological protection is to keep radiation exposure and contamination in the nuclear power plant and the environment as low as possible. The radiation monitoring system measures and records radiation in the various controlled areas. Different radiation monitoring systems are provided to survey local dose rates, systems, and areas, the environment and personal doses, as well as the release of solid, gaseous, and liquid wastes. The monitoring systems include permanently installed instruments that are integrated into the I & C system; stationary monitors (e.g., contamination monitors), which indicate locally, continuous, and intermittent sampling points; and mobile radiation measuring instruments for local or personal monitoring. The samples are evaluated in the health physics laboratory. Table 15 shows the number of permanently installed radiation monitors in a Konvoi plant.

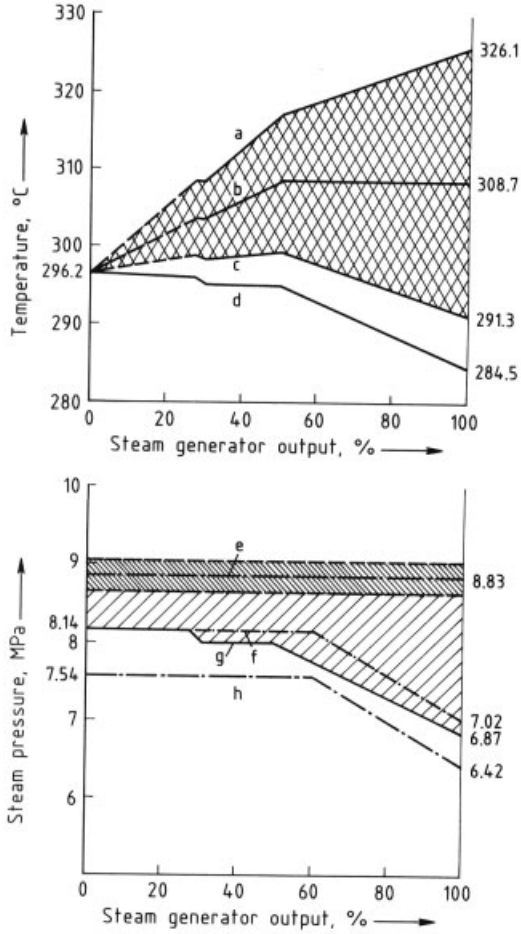

Figure 25. Part load diagram of a pressurized water reactor (Konvoi plant) featuring set point control at constant main steam pressure up to about 50 % power and then constant average coolant temperature up to full load
a) Steam generator inlet temperature; b) Average coolant temperature; c) Steam generator outlet temperature; d) Main steam temperature at steam generator outlet; e) Response pressure of main steam safety valves; f) Response pressure of main steam maximum pressure limit; g) Main steam pressure at steam generator outlet; h) Main steam minimum pressure limit

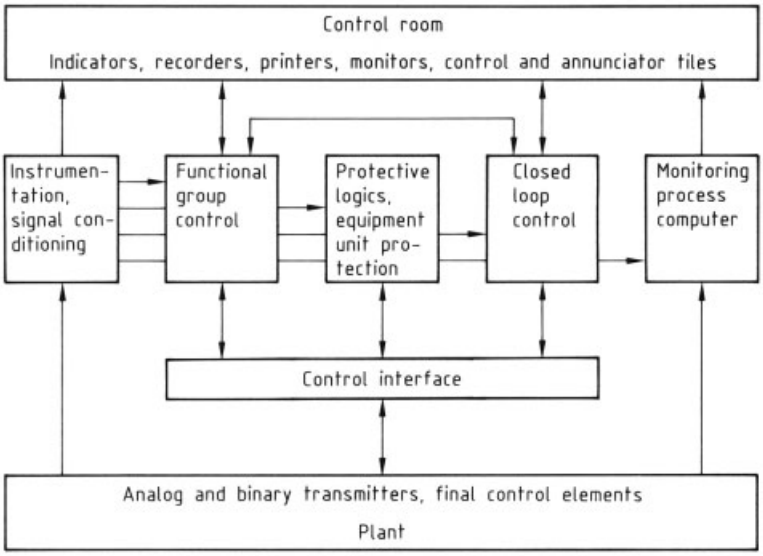

Figure 26. Tasks of the operational instrumentation and control

Neutron Flux Measurement. The neutron flux inside and outside the core is measured to detect the reactor power and power density distribution. *Outside the reactor pressure vessel*, the source, intermediate, and power range channels are installed, which measure the integral reactor power in ionization chambers over approximately 13 decades. The measured data are used for indication, display, control, limitation, and protection, as well as the process computer.

In the reactor core, 48 self-powered detectors measure the power density. The Aeroball system is used for calibration of neutron-flux instrumentation channels and for intermittent neutron-flux mapping. In the reactor core, stacks of vanadium alloy steel balls (diameter 1 mm) are activated by the neutron flux. The activation measurements are transmitted to a computer connected to the Aeroball system which calculates the relative power density distribution in the core and references it to the thermal output.

Operational Instrumentation and Control (Fig. 26). The operational instrumentation measures the various physical, mechanical, and chemical process data and converts measured data into electrical signals. These binary and analog signals are used for control, regulating, and monitoring functions. The normal trouble-free operation of a nuclear power plant is performed by the operational instrumentation and controls.

Manual operation and observation of the overall nuclear power plant process takes place in the central control room. The control room is divided into the master control station and the systems control boards, on which the control tiles, monitors, printers, alarms, and display units are arranged.

Konvoi plants are additionally equipped with the computer-aided process information system PRISCA. This is a problem-oriented decentralized data acquisition computer. The classical tasks of a supervisory computer, such as recording process events on paper, indicating alarms on video display units, and performing nuclear calculations, are supplemented by the processing of high-grade process information and the output of process graphics on color CRT (cathode-ray tube) monitors. Thus, the functions of all systems can be observed during normal operation and accidents; the fulfillment of important safety functions can be monitored and the effectiveness of automatic and manual actions can be controlled.

Further developmental trends in operational I & C are toward programmable and redundant automation systems interconnected by local area networks.

3.1.1.3. Nuclear Power Plant Operation

The operation of a reactor is divided into three phases, start-up, load operation, and shutdown. A typical *start-up* diagram of a pressurized water reactor is shown in Figure 27. When the plant is started up from the cold condition, it is first vented and then pressurized to approximately 3.5 MPa by heating the pressurizer. Subsequently, the reactor coolant pumps start up. The reactor is made critical by control rod withdrawal and deboration of the coolant. Then it is warmed up at a rate of about 50 K/h by nuclear heat generation. The reactor is capable of delivering power to the turbine 4 h after reaching criticality and about 13 h after start-up.

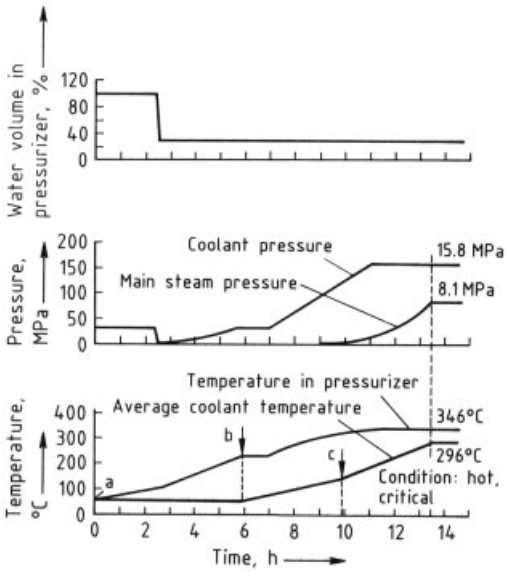

Figure 27. Start-up diagram
a) Pressurizer heaters switched on; b) Reactor coolant pumps started; c) Reactor goes critical

In *load operation* the interactive controls for generator output and the reactor coolant temperature implement load changes ranging from 30 to 100 % nominal load at rates of up to 10 %/min (relative to nominal load) within wide margins about the starting point. The entire load range can be traversed in about 60 min. The stored energy capacity of the steam generators is used for rapid load changes. Hence, pressurized-water reactors are suitable not only for base-load operation but also for load cycling because of their self-regulation features.

After about a year of full-load operation, the core approaches the end of its cycle because all excess reactivity has been burned up. This state is reached when the full power can no longer be maintained even if the control assemblies are almost completely withdrawn from the core. However, the cycle duration of the core can be extended by slightly reducing the coolant temperature and limiting the power output (stretch-out operation). This permits the plant to operate for another 60 – 100 d before output falls below 70 or 50 %. The output averaged over this period is 85 % and 75 %, respectively. Thus, the time at which the reactor is shut down for refueling can be coordinated with operational requirements.

The *shutdown* process is shown in Figure 28. First, the plant output is reduced to 30 % by manually adjusting the generator load set point. The coolant temperature control follows automatically. The output is then further reduced to 15 % by lowering the neutron-flux set point (also manually). At first, the reactor is cooled by the steam generators and the main steam bypass station (steam dumping into the condenser) until the temperature has been decreased sufficiently (≈ 120 °C). Then cooling is taken over by the residual heat removal system.

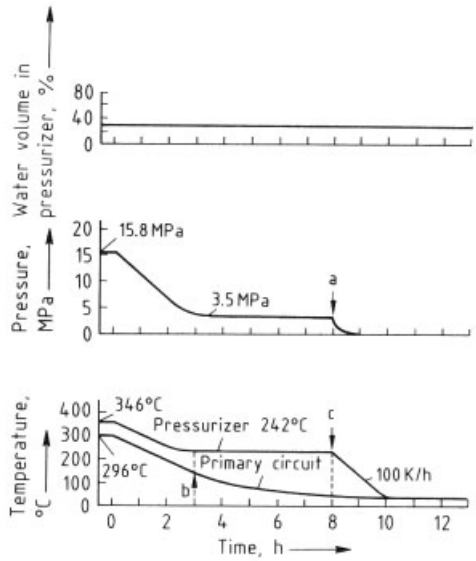

Figure 28. Shutdown diagram
a) Reactor coolant pumps tripped; b) Residual heat removal system in operation; c) Spraying from value control system

3.1.1.4. Radioactive Releases to the Environment

Operational experience with pressurized-water reactor power plants has shown that they emit only very small amounts of radioactivity during operation. As a result of their design (highly efficient processing and retention systems for liquid and gaseous effluents, shielding) and intensive personnel training, the average emissions are far below the statutory limits. The actual levels for individual plants depend on a variety of factors, such as height of stack, method of processing, weather conditions, mode of operation–base load or load cycling–and fuel quality. In a modern 1300-MW plant, *gaseous emissions* from the stack are generally $< 3.7 \times 10^{13}$ Bq/a (predominantly noble gases ^{133}Xe, ^{41}Ar, ^{85}Kr, traces of ^{131}I, and aerosols). The activity of annual *liquid effluent* releases is limited to ca. 3.7×10^{10} Bq/a but actual values are usually less. In addition, a particular feature of pressurized water reactors is that tritium emissions are 2.2×10^{13} Bq/a averaged over a period of several years. However, this is only of secondary importance because the radiotoxicity of tritium is low compared to other isotopes. These emission levels raise the radiation background in the plant environment by 0.5 Sv/a, which is less than 1 % of the radiation emitted by natural sources. Up to now, not one case of radiationinduced impairment of health is known to have occurred in the vicinity of a PWR nuclear power plant.

Annual emissions of radioactive substances in becquerels, from a 1300-MW plant may be summarized as follows:

Gaseous effluents	
Noble gases	4×10^{12}
Aerosols	1×10^{7}
^{131}I	5×10^{6}
Liquid effluents	
Fission and activation products (without tritium)	2.2×10^{9}
Tritium	1×10^{13}

However, these values depend very strongly on individual plant parameters. A comparison of older data with those of recent years reveals that constant efforts are being made to reduce the radiation exposure of both plant personnel and the environment at large.

3.1.1.5. European Pressurized-Water Reactor

The European Pressurized-Water Reactor (EPR) is currently being developed by Nuclear Power International (NPI) and its parent companies, Framatome and Siemens, in cooperation with EdF and German utilities, aiming at achieving a new improved nuclear power plant design that will become an acceptable and attractive alternative for meeting energy demands in the future.

3.1.1.5.1. Description of the Nuclear Systems

Primary Circuit and its Main Characteristics. The primary loop configuration is the same as that of existing designs and can be considered well-proven. The reactor pressure vessel (RPV), steam generator (SG) (especially secondary side) and pressurizer (PZR) are of a larger size than current designs.

In the *reactor pressure vessel* the free water volume between the level of the reactor coolant lines and the top of the active core is increased in order to improve the mitigation of loss of coolant accidents (LOCA) by prolonging the period until beginning of core uncovery in case of LOCA or minimizing the depth of core uncovery, if any. At the same time the increase of this volume contributes to an improvement in the mitigation of accidents at shutdown conditions (in particular mid-loop operation), e.g., with loss of the residual heat removal system (RHR), by providing longer grace periods.

For the *pressurizer* a large volume both in terms of water and steam phase is provided in order to smoothen plant response to relevant operating transients and accidents. The great water volume of the secondary side of the *steam generator* supports smoothing of normal operating transients and reduces the potential for unplanned reactor trips. In case of a total loss of all feedwater supply (including emergency feedwater), the expected dryout time of the SG will be > 30 min. The valve configuration of the primary side overpressure protection aims at avoiding the response of "nonisolatable valves" in plant conditions with a potential for radioactivity release. The valves are mounted to the top of the pressurizer in order to minimize the length

of high pressure piping. Each of the three discharge trains is provided with two safety valves. Automatic opening of the main valve is actuated by pilot actuators dedicated to each individual safety valve.

With the chosen arrangement of pressurizer discharge, the following safety functions can be performed:

1) Overpressure protection of the reactor cooling system by automatically initated discharge of either steam, water or two-phase fluid
2) Depressurization of the reactor cooling system by discharge of steam, water or two-phase fluid in plant conditions when pressurizer spraying is not available or not efficient
3) Discharge of the reactor cooling system to enable residual heat removal in case of unavailablility of the secondary side heat removal (feed and bleed)
4) Discharge of the reactor cooling system in a postulated core melt situation to guarantee depressurization to a sufficiently low level that would rule out the high pressure core melt accident and its severe consequences.

On the secondary side of the SGs, three discharge trains will be arranged on each main steam line at the outlet of the containment. One discharge line will be equipped with a relief valve (normally open) and an isolation valve (normally closed). The two other discharge lines will be equipped with nonisolatable safety valves.

Reactor Core and Fuel Design. The reactor core is built up by 241 mechanically identically designed fuel assemblies; somewhat more than in currently operating units. Each fuel assembly consists of 264 fuel rods and 25 guide tubes arranged in a 17×17 array; their active length is 4.20 m. The fuel rods are made of Zircaloy tubing containing uranium dioxide ceramic pellets (initial enrichment ≤ 5.0 wt %). The core is designed for UO_2 fuel elements. They can be substituted up to about 50 % by MOX fuel assemblies. Some fuel assemblies contain burnable absorber (Gd_2O_3) to suppress high excess reactivity, especially in the first core. The reactor power level is monitored by the ex-core instrumentation. The in-core instrumentation consists of the aeroball system and the self-powered detector system which controls the power distribution in the core. The in-core instrumentation penetrates the reactor pressure vessel head from the top at only very few positions.

Primary Components. The *reactor pressure* vessel is designed for a lifetime of 60 a. A total neutron fluence of 10^{19} *nvt* will not be exceeded. This is achieved by provision of a rather large water gap and a heavy reflector around the core. The upper part of the reactor pressure vessel will be machined from one single forging. The flange is made as an integral part of the nozzle shell. The nozzles themselves are of the on-set type so that nondestructive examination can be easily performed from the inside. The nozzles are located as high as practicable above the upper edge of the core to increase the hydrostatic pressure for reflooding and to avoid the loop seal effect. The reactor pressure vessel is internally clad with two layers of low-carbon stainless steel. Generally, in-service inspection is performed from inside the vessel; an access from the outside, between the outer wall and the thermal insulation, is also provided.

Reactor Internals. The core barrel flange rests on a ledge machined from the flange of the reactor pressure vessel and is preloaded by an elastic system. The fuel assemblies are placed directly on a flat perforated plate, machined from forged stainless steel. The fuel assemblies are centered by two pins each, which are screwed to the core plate. The cooling water flows through the core plate through four holes dedicated to each fuel assembly. The space between the polygonal outer shape of the core and the cylindrical inner surface of the core barrel is filled by a stainless steel structure to reduce leakage of fast neutrons and to flatten the power distribution. The bulk of the internals is made of low-carbon stainless steel.

The core barrel flange rests on a ledge machined from the flange of the reactor pressure vessel and is preloaded by an elastic system. The fuel assemblies are placed directly on a flat perforated plate, machined from forged stainless steel. The fuel assemblies are centered by two pins each, which are screwed to the core plate. The cooling water flows through the core plate through four holes dedicated to each fuel assembly. The space between the polygonal outer

shape of the core and the cylindrical inner surface of the core barrel is filled by a stainless steel structure to reduce leakage of fast neutrons and to flatten the power distribution. The bulk of the internals is made of low-carbon stainless steel.

The *steam generators* feature an axial economizer to provide a steam pressure increase of about 3 bar when compared to a boiler type of the same heating surface. The material for the heating tubes will be chosen from either Incoloy 800 or Inconel 690. The heating tubes are supported by perforated plates.

The *pressurizer* is of conventional design but with an enlarged free volume. The spray systems for normal operation and auxiliary spray are completely separated from each other. In addition, in order to achieve a life time of 60 a and meet the requirements as regards plant power flexibility, several separate spray lines are provided and operated alternatively.

All pressure boundary parts, except for the heater penetrations, are made of ferritic steel; basically the same as used for the reactor pressure vessel. The penetrations are in stainless steel and welded with an Inconel material.

The *reactor coolant pumps* are of well-proven design, as already used in plants in France and Germany. The reactor coolant pumps are provided with a standstill seal in order to assure leak-tightness of the shaft seal without the need of an active seal water supply system under conditions when the pump is at rest, e.g., in the event of a station blackout.

The *main coolant lines* will either be made from forged ferritic steel with austenitic steel cladding or from forged stainless steel. A high quality and surveillance program enables preclusion of the leak-before-break philosophy. However, a mass flow equivalent to a double ended break of a main coolant line is still assumed when designing the emergency core cooling system.

Operating Characteristics. The EPR is a Nuclear Island (i.e., includes all equipment inside the reactor building and the main reactor auxiliary systems) for a power output of about 1500 MW_e. The primary components are enlarged relative to current types, the safety and operating systems have been redesigned or updated to the conditions required.

The EPR can be operated between 20 and 100 % of rated generator power. In the power range between 50 % and 100 % load, the control systems will keep the average coolant temperature constant in accordance with the so-called part load diagram; the main steam pressure will vary between 8.4 and 7.25 MPa. In the lower power range, below 50 % power, the main steam pressure is kept constant at 8.4 MPa and the average coolant temperature will vary with the load. This control strategy results in the lowest demands on the chemical and volume control system, loads on the pressurizer surge line, and the control rod drive mechanisms during load changes in the most frequent operation mode. In the lower power range, a continuation of the constat average coolant temperature mode would penalize the secondary side design (by a design pressure increase) and eventually also the primary side design pressure. The control and operational systems are designed to provide the EPR with a high capability to follow the actual power demands of the grid.

Reactor Auxiliary Systems. The *chemical and volume control system* has safety-related and operational functions. The safety-related functions aim at providing make-up to the reactor coolant system in the event of loss of inventory due to limited leakages and providing boration capacity up to a concentration as required for cold shutdown under operational and certain accident conditions. In addition, the system provides a means for pressure control of the primary circuit by supplying auxiliary spray to the pressurizer (thereby condensing steam and reducing primary system pressure) and seal water to the reactor coolant pumps.

The operational functions of the system are: control of the water inventory, the water quality, and the boron concentration in the primary system. Additionally, the system adjusts the chemical composition of the reactor coolant system and removes dissolved gases by degasification of the letdown flow.

3.1.1.5.2. Instrumentation and Control Systems

I&C Structure. The I&C systems and equipment are divided into three classes (F1A,

F1B, F2) in accordance with their importance for safety and with respect to required reliability, performance, failure behavior, maintenance, testing, and QA; in addition, there is a nonclassified category (NC). The main features of these categories are:

1) The I&C of the F1A class is used for automatic actuation and control of safety-grade systems. It shall withstand a single failure also during periods of maintenance and testing. In addition, spurious actuation upstream of the last voter is prevented by the fourfold redundant and physically separated safety trains.
2) The I&C of the F1B class has a functional structure that can cope with a single failure – without the additional requirements related to maintenance and testing. Special emphasis is laid on the qualification of the back-up control and information means, including the related software.
3) The I&C of the F2 and NC classes is used for normal operation of the plant, including the control of major plant parameters, limiting conditions of operation (LCOs), information and operation and other nonsafety I&C functions.

Safety I&C (F1). The safety I&C functions must have a high reliability so that they will not be a dominant contributor to the unavailability of safety systems. The safety I&C is of redundant architecture and designed to limit the consequences of equipment failures or malfunctions that may result from failure-inducing events within I&C systems (single failure) and their consequential effects. Connections between redundant trains are necessary for exchange of information and commands, but they must not impair the independence. To this end, the redundant trains, or divisions, are installed with physical separation and with a minimum number of interconnections.

Appropriate measures are provided to cope with common cause failures (CCF) in order to meet the overall probabilistic design targets. CCFs and their consequences can result from different sources (e.g., faulty manufacturing, erroneous design, equipment failures, and environmental conditions during the course of accidents).

Operational I&C. The operational I&C encompasses all I&C functions and associated systems and equipment for normal operation. It includes the measurements, the signal conditioning, open- and closed-loop controls, the signal processing, and the data transfer to the man – machine interface.

The tasks of the open- and closed-loop control are: (1) control of the nuclear power generation during normal operation and anticipated operational occurences in such a way that predetermined setpoints for relevant process variables are reached and maintained; (2) protection of mechanical equipment with high investment costs via redundant equipment protection, and (3) provision of information for displaying the plant status for normal, upset, and accident conditions and for documentation of all relevant process data.

Man – Machine Interface Facilities and Control Rooms. The main control room is a screen-based control room with an overview panel. It is used for process control during normal, or accident situations including outages. In addition, the main control room has a safety control area with back-up control means. Further functions of the main control room or of adjacent rooms are security surveillance, fire protection and radiation monitoring, management of maintenance and periodic testing, external and internal communication, and access to documentation and to recorded information. The main control room accommodates operator work stations (all of the same design) which are used for process control in all plant conditions via operational I&C. A plant overview panel is visible from all work places and will be used for the coordination among the operators.

The safety control area (with the back-up control means) in the main control room is used in the event of major losses of the normal control means. It can be used for the safe shutdown (hot or cold) of the plant or to perform post-accident operation. The EPR is provided with a Remote Shutdown Station (RSS) which is designed for transferring the plant to and maintaining it in safe shutdown conditions, in case of unavailability of the main control. The RSS is equipped with internal and external communication means.

3.1.1.5.3. Safety Concept

Safety Requirements and Design Philosophy. The strategy pursued for the EPR is to further enhance the very high safety level of French and German pressurized-water reactors. This strategy implies improving the prevention of accidents – including severe accidents – and adding features, mainly related to the containment, to mitigate the consequences of postulated severe accident scenarios – including core melt situations – to avoid need for stringent off-site countermeasures. The probability of such postulated accidents has been significantly reduced. The plant design is based on a deterministic approach and considers "risk reduction" measures.

Deterministic Design Basis. In the deterministic analysis the different events are categorized in four Plant Condition Categories (PCCs) in accordance with their anticipated frequency of occurrence; PCC1 covers normal operation states, PCC2 to PCC4 envelop transients, disturbed states, and accidents. Stringent radiological limits are applied for normal operation and anticipated operational occurrences as well as for accidents.

Risk Reduction. The EPR design takes into consideration also events beyond the Deterministic Design Basis (i.e., events with multiple failures and corresponding occurrences up to the total loss of safety-grade systems) on a probabilistic basis in order to quantify the residual risk. Release limits of radioactivity in case of severe accidents have been specified on the basis that no stringent off-site emergency response actions (such as evacuation or resettlement) would be needed outside the immediate vicinity of the plant.

The overall safety objectives that have been set for the EPR require that the probability of core damage (CDF) shall be below 10^{-5}/a; and the probability of large releases to the environment shall be below 10^{-6}/a, including all events and all reactor operating states.

Two Risk Reduction Categories (RRCs) have been introduced and representative scenarios defined for both. RRC-A relates to additional features to prevent accidents from progressing to a core melt situation, and RRC-B to prevention of large releases, in order to provide a design basis for risk reduction features. Typical examples on risk reduction features are:

1) Primary system discharge to the in-containment refueling water storage tank, in the event of total loss of secondary side cooling (RRC-A); and
2) Features for spreading and cooling of corium, for hydrogen recombination, and for containment heat removal in the event of a core melt situation (RRC-B).

External and internal hazards are normally not assigned directly to specific plant condition categories or risk reduction categories, in order to avoid the study of numerous sequences. But the main principles behind the deterministic design basis and the risk reduction approach (namely: the more probable the event, the more conservative rules and acceptance criteria) are applied also for dealing with external and internal hazards.

External hazards are to a certain extent site--dependent. Sites with an extreme external hazard potential are not taken into consideration as potential sites.

Safety Systems and Features (Active and Passive).

Safety Systems Configuration. Important safety systems (safety injection, emergency feedwater, main steam relief, cooling chain, emergency electric power) are arranged in a four-train configuration (see Fig. 29). The layout comprises four separate divisions, corresponding to the four trains.

The safety injection systems (see Table 16), feature an In-containment Refueling Water Storage Tank (IRWST) located at the bottom of the containment. During design basis accidents the low-head safety injection (LHSI) system transfers the decay heat to the ultimate heat sink via heat exchangers. The containment spray system is only used for heat removal in case of severe accidents. The delivery head of the medium head safety injection (MHSI) system will be adjusted below the steam generator relief and safety valve set points. In case of a steam generator tube rupture, the affected steam generator will be isolated on the secondary side. After the initial transient, the primary and secondary pressures will equalize at a level below the set points of the safety valves in this steam generator.

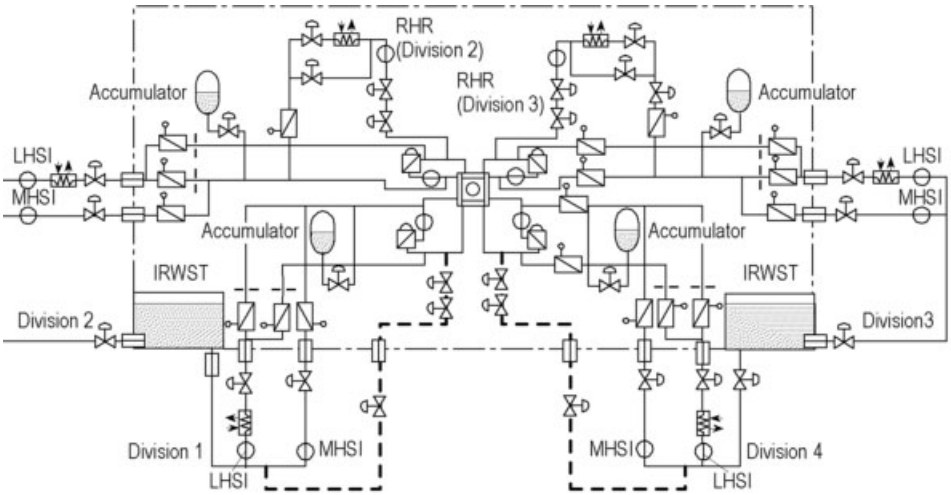

Figure 29. EPR – Configuration of fluid safety system

Table 16. Organization of safety injection systems

Medium head safety injection system	4 trains, cold side injection
Accumulators	4 accumulators, cold side injection
Low head safety injection system	4 trains, cold side injection (short term); combined hot /cold injection (long term)
Residual heat removal system	2 trains, independent from safety injection system
In-containment refueling water storage tank	storage of borated water inside containment

A separate residual heat removal (RHR) system is provided, arranged in two trains and installed inside the containment to minimize risk of containment bypass.

The function of any one of the safety systems can be accomplished by another diverse system (or group of systems) during malfunction. Special design provisions are devoted to preventing the high pressure core melt. Therefore, highly reliable secondary side heat removal systems are implemented.

Safety Injection Systems. The safety injection systems mitigate loss of coolant accidents of all sizes, specific non-LOCA events, such as breaks of main steam lines, and sequences leading to feed and bleed. The systems ensure heat removal, coolant inventory, and reactivity control.

The medium head safety injection (MHSI) system feeds into the cold legs of the reactor coolant system. The shut-off head of the system is 8.0 MPa, which is sufficient to cope with all LOCA related requirements.

In addition to the medium head injection system, cold leg accumulator injection is provided to cope with large and intermediate break sizes. Four accumulators with a volume of 50 m^3 each are provided, each directly assigned to one cold leg. The response pressure of the accumulators is designed to 4.5 MPa.

The *low pressure injection system* feeds initially into the cold leg. In order to stop the core outlet steaming and the steam release into the containment, after 1 – 2 h combined injection into the hot and cold legs starts. The injection pressure of 2.5 MPa offers advantages for feed and bleed operation and supports accumulator injection in an optimum way for a large spectrum of break sizes.

In-Containment Refueling Water Storage Tank. The in-containment refueling water storage tank is located inside the containment between the reactor cavity and the missile protection cylinder on the bottom level of the containment. In the case of loss of coolant accidents, or in feed and bleed situations, the safety injection system draws from the in-containment refueling water storage tank. The water steam mixture escaping through the leak and through the bleed valve, respectively, is returned to the tank. In the case of severe accidents the in-containment water storage tank will provide the cooling water

for flooding the spread molten corium. In addition, the storage tank provides water for the operational function of flooding the reactor pit and the pools during refueling.

Emergency Feedwater System. The emergency feedwater system (only safety grade, see Fig. 30) consists of four separate and independent trains, each providing injection to one of the four steam generators. Each emergency feedwater pump takes suction from an emergency feedwater tank. These tanks and the systems are located in the four divisions of the safeguard buildings.

The emergency feedwater system transfers the residual and latent heat from the reactor coolant system via the steam generators to the atmosphere as long as the steam generator saturation temperature is above 150 °C, following any plant incident or accident other than those reactor coolant boundary ruptures for which complete residual heat removal by the safety injection system is possible (i.e., moderate to large size loss of coolant accidents). Following a LOCA not all core residual heat is released through the break flow and at least a portion of the heat must be removed via the steam generators, the emergency feedwater system ensures sufficient water supplies to the steam generators.

In the case of a steam generator tube rupture, the emergency feedwater system removes the heat via the intact steam generators.

The emergency feedwater system keeps the water inventory of at least one steam generator above an adequate level to maintain primary to secondary heat transfer, assuming a single failure. Safety grade, i.e., normally locked closed headers, can only be opened after a certain grace period for pressure reduction in the unfed steam generator.

After a small break LOCA or a steam generator tube rupture, the emergency feedwater system provides enough cooldown capability that the saturation pressure of the steam generator of 6.0 MPa is reached within a time span consistent with the performance requirements of the safety injection system and the radioactivity release limits for steam generator tube rupture, assuming a single failure.

The emergency feedwater system shall provide sufficient heat removal capacity and autonomy to ensure continued removal of decay heat for 24 h with a final reactor coolant system temperature not exceeding nominal hot shutdown conditions.

Residual Heat Removal System. The residual heat removal system combines operational and safety functions. The safety function implies that the residual heat removal system, in conjunction with the component cooling water system and the service water system, shall maintain the reactor coolant system core outlet and hot leg temperature below 180 °C following a reactor shutdown, assuming a single failure and the maximum design temperature of the service water. The operational function requires the residual heat removal system to be capable of cooling down the reactor coolant system to 50 °C following a reactor shutdown, with at least one reactor coolant pump in operation and considering the maximum design temperature of the service water. The residual heat removal system consists of two separate and independent trains located inside the containment. During normal operation, only the inside RHR trains are used for cooldown and cold shutdown. Back-up trains are not used unless the RCS temperature is below 100 °C.

In case of a total loss of both residual heat removal trains, the residual heat removal function can be taken over by two of the four low head safety injection trains, which function as a back-up for the residual heat removal system.

Severe Accidents (Beyond Design Basis Accidents). The EPR design enhances accident prevention and reduces the residual risk for the public and the environment by reducing possible releases of radioactive material and radiological consequences arising from severe core damage sequences. The design target of the EPR is that off-site emergency response actions (population evacuation or relocation) shall be restricted to the nearby plant vicinity. To this end, maintaining the integrity of the containment is utterly important. this can be achieved by:

Avoiding early containment failure or bypass
Cooling of the corium in the containment and retention of fission products by covering with water
Preservation of containment functions, such as low leak rates, reliable isolation function, prevention of basemat melt-through, and ultimate pressure resistance to cope with energetic events

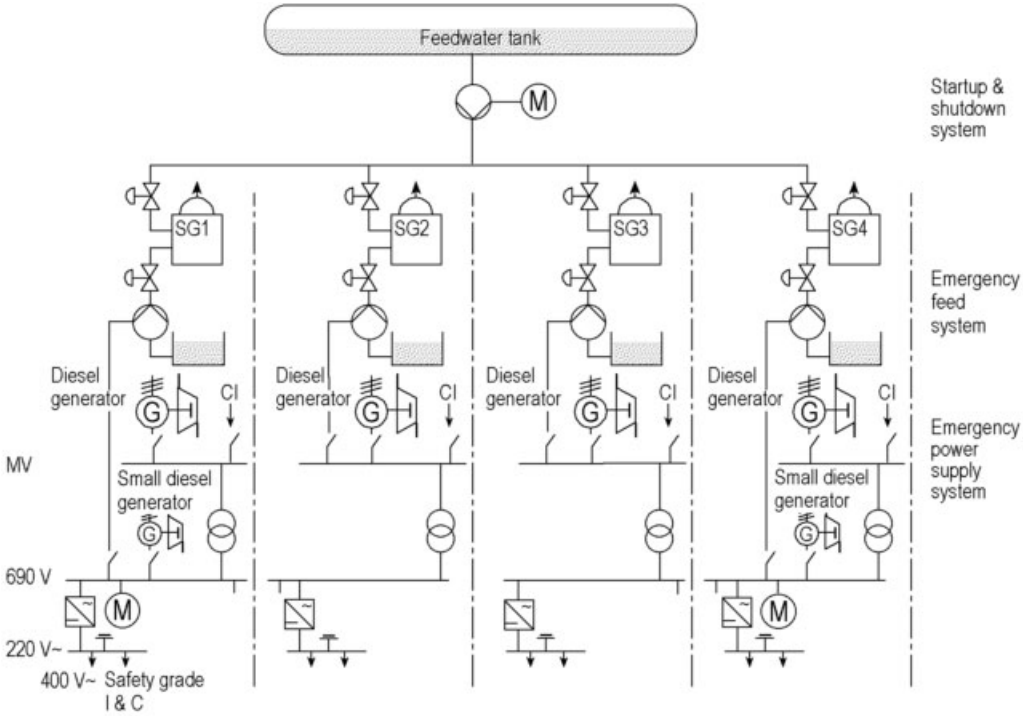

Figure 30. EPR – Configuration of emergency feeedwater system

Pressure reduction inside the containment by dedicated heat removal
Collection of unavoidable containment leakages in the annulus atmosphere and release via stack after filtration

Severe Accident Prevention and Mitigation Features. The EPR strategy includes both preventive measures and mitigating features:

1) Prevention of high pressure core melt situations, by ensuring a high reliability of the decay heat removal systems, complemented by pressurizer relief valves. The depressurization eliminates high pressure failure of the RPV and the danger of direct containment heating with the potential consequence of early containment failure. The consequences of an instantaneous full cross-section break of the RPV at a pressure of about 2.0 MPa are nevertheless taken into account for the pit layout and support design.
2) Prevention of formation or reduction of the hydrogen concentration in the containment by catalytic H_2-recombiners and, if necessary, by selectively arranged igniters. The prevention of molten core–concrete interaction contributes to reducing the amount of hydrogen.
3) Prevention of ex-vessel steam explosions endangering the containment integrity by minimizing the amount of water in the area where the corium is spread.
4) Prevention of a molten core–concrete interaction by spreading the corium in a spreading compartment provided with a protective layer, and
5) Connection of this spreading compartment to the reactor pit via a melt discharge channel which slopes towards the spreading compartment. This channel is closed by a steel plate (possibly covered with refractory material) which will resist melt-through for a certain time, in order to accumulate the melt in the pit.
6) Provisions for connecting the spreading compartment with the In-Containment Refueling Water Storage Tank (IRWST) for water flooding after spreading; these pipe connections are closed during normal operation.

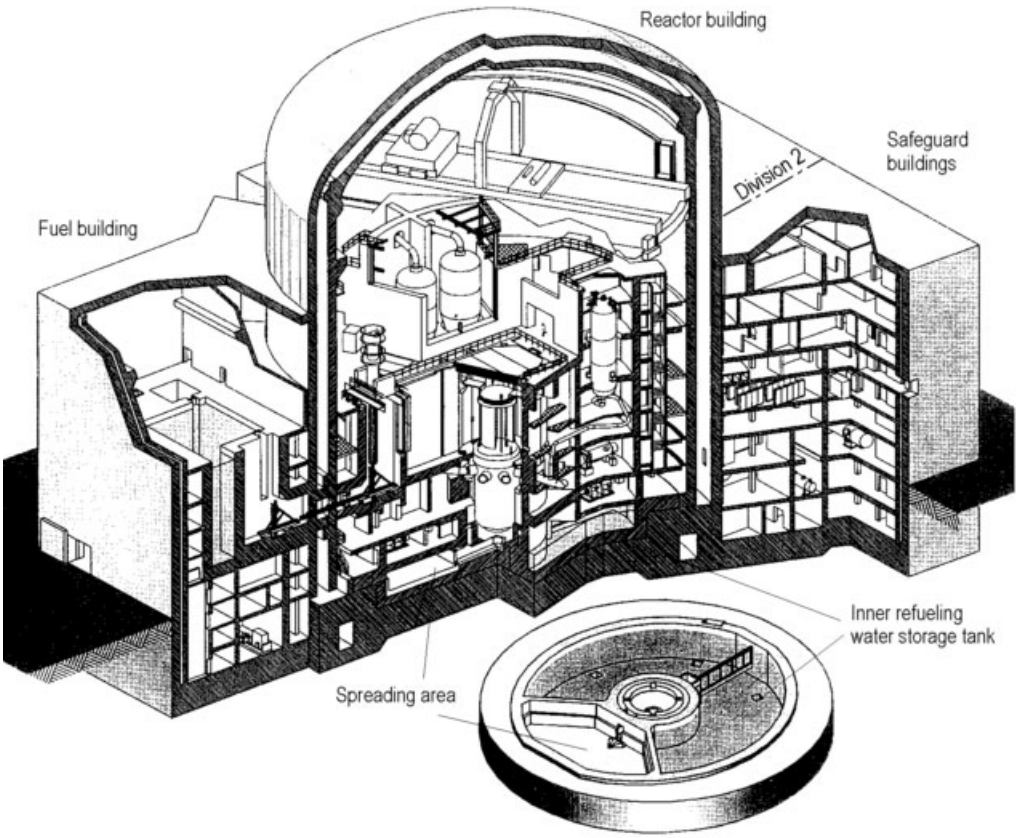

Figure 31. EPR – Plot plan

7) Dedicated basemat cooling system supplied by the containment heat removal system to prevent high temperature loadings in the basemat.

3.1.1.5.4. Plant Layout

Buildings and Structures, Including Plot Plan. The general layout of the EPR plant (plot plan) is shown in Figure 31. The reactor building with the containment is surrounded by the safeguard and fuel buildings which contain safety systems. Most of the safety-grade systems are designed with a four-fold redundancy, arranged in four independent divisions with complete physical separation.

Other plant buildings, such as the access building and the nuclear auxiliary building, are located in close contact with the safeguard and fuel buildings, whereas the turbine building and the associated conventional electrical building are built separate from the reactor building complex and arranged such that the reactor building is located in the projection of the turbine generator shaft.

Design Requirements. The plant is designed to withstand the impacts of internal and external events. With respect to earthquake and explosion pressure waves, the buildings and structures have been strengthened so that the function of safety-grade equipment will not be jeopardized by collapsing structures, and the equipment itself must withstand the dynamic effects inside the buildings. For protection against airplane crash, safety-related equipment is located in bunkers, or redundant portions will be geographically separated from each other so that only one train is affected.

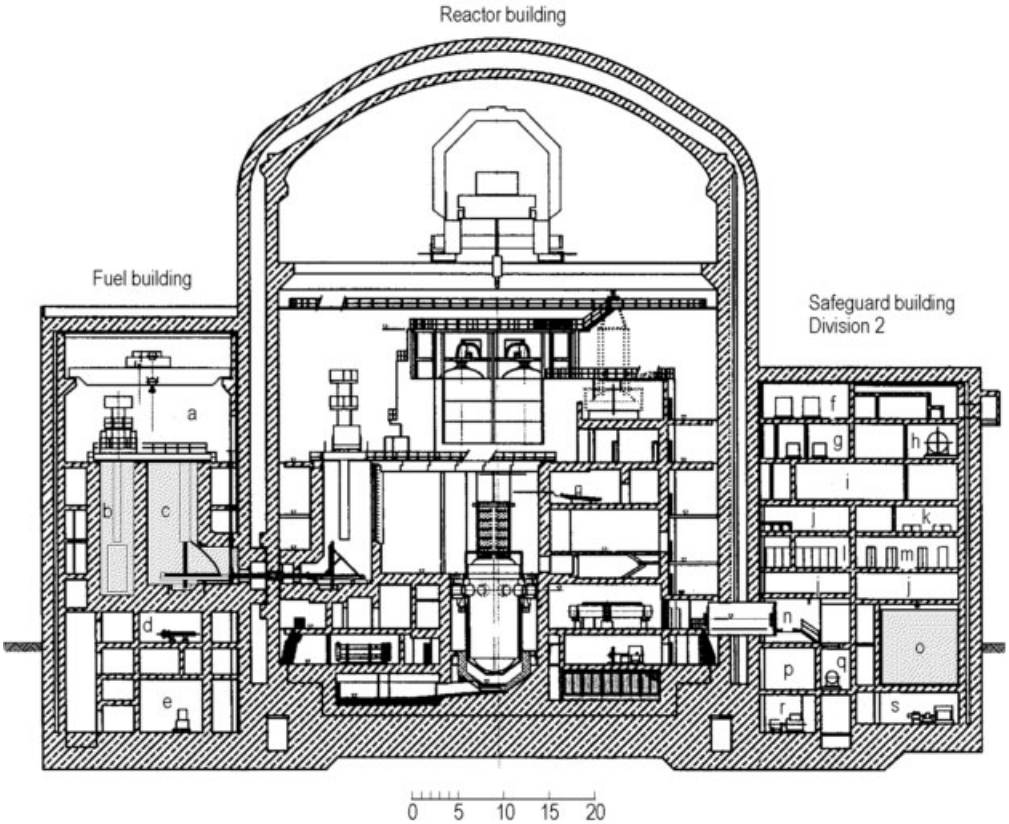

Figure 32. EPR – Building arrangement
a) Fuel pool floor; b) Fuel pool; c) Transfer station; d) Minimum-flow cooler; e) CVCS pump; f) Security chilled water system; g) Air conditioning MCR; h) CCWS tank; i) Main control room; j) Cable floor; k) Batteries, 220 V; l) I&C cabins; m) Switchgears; n) Personnel air lock; o) Demineralized water pool; p) SIS valve room; q) LHSI HX; r) LHSI pump; s) EFWS pump

Table 17. Boiling-water reactors in the Federal Republic of Germany

Plant	Gross output, MW_e	Handover	Decom-missioned
Kahl	16	1962	1985
Gundremmingen A	250	1966	1980
Lingen	267	1968	1977
Würgassen	670	1975	
Brunsbüttel	806	1976	
Isar 1	907	1979	
Phillippsburg 1	900	1980	
Krümmel	1316	1984	
Gundremmingen B	1300	1984	
Gundremmingen C	1308	1985	

The loads from internal events (e.g., fire loads, missile loads, jet impingement loads, flooding effects) are covered by design features (e.g., four-train four-division concept).

Reactor Building (Containment). The reactor building (cf. Figure 32, main building complex) is the central building of the plant complex. It coincides completely with the containment, and thus, the following description of the containment covers also the reactor building.

Adoption of a double concrete containment design was decided for the EPR. The inner containment wall is made of prestressed concrete. The leak-tightness requirement of less than 1 vol % per day can be ensured without provision of a containment liner. The outer wall, in reinforced concrete, completes the double containment arrangement. Leakages through the inner containment wall are released via the annulus air extraction system. Openings for access of personnel or supply of equipment to the inside of the containment are permanently closed hatches or air locks

Table 18. Characteristic data for boiling-water reactors manufactured in the Federal Republic of Germany

	Name/Location				
Characteristic data	Kahl	Lingen	Würgassen	Brunsbüttel	Gundremmingen C
Overall plant					
Start of commercial operation	11/61	10/68	11/75	2/77	1/85
Thermal output, MW_t	60	516 + 216 [a]	1912	2292	3840
Net electric output, MW_e	15	240	640	771	1308
Net efficiency, %	24.8	32.6	33.4	33.6	32.5
Direct cycle			yes	yes	yes
Indirect cycle	yes	yes			
Reactor core					
Uranium mass, t	5	32	81	97	139
Fuel	enriched UO_2, UO_2-PuO_2	enriched UO_2, UO_2-ThO_2	enriched UO_2	enriched UO_2	enriched UO_2
No. of fuel assemblies	88	284	444	532	784
No. of fuel rods per assembly	36	36	63	63	62
Outside diameter of fuel rod, mm	14.3	14.3	12.5	12.5	12.3
Average specific power, kW/kg U	11.3	16.1	23.5	23.5	26.5
No. of control rods	21	69	109	129	193
Nuclear steam supply system					
Coolant core outlet temperature, °C	286	286	286	286	286
Coolant pressure, MPa	7.06	7.06	7.06	7.06	7.06
No. of recirculation pumps	natural circulation	2 external [b]	2 external 16-jet pumps [c]	8 internal [d]	8 internal [d]
Turbine					
Main steam pressure, MPa	4.7	4.2	6.57	6.57	6.7
Main steam temperature, °C	259	530	281.5	281.5	283

[a] Additional fossil-fired superheater (216 MW_t) in secondary cycle.
[b] Reactor water recirculated outside RPV.
[c] Eight jet pumps in RPV receive motive water from one external pump.
[d] Axial-flow recirculation pump inside RPV.

with double sealings on both sides. Fluid systems penetrating the containment are provided with double isolation valves, inside and outside the containment.

Accumulation of combustible gases, especially hydrogen, is controlled. Furthermore, the basemat in the spreading compartment is protected by protective layers and the dedicated cooling system which is fed by the containment heat removal system against elevated temperatures resulting of a core melt.

The severe accident conditions lead to more severe design conditions compared to existing plants. Most important is the increased design pressure of 650 kPa.

Other Buildings.

Safeguard and Fuel Buildings. Connections between the safety systems and the reactor coolant system are made as short as possible. The individual trains of the safety systems are arranged radially to the primary loops.

The four trains of the safety systems are located in four safeguard buildings that surround the reactor building.

The storage pool for spent fuel assemblies is located outside the containment, facilitating loading and unloading of the fuel transport cask for the transport of fresh fuel to the plant and of spent fuel away from the plant. A transfer tube in the containment wall connects the inside of the containment with the fuel building.

The primary system is arranged symmetrically. Concrete walls are provided between the loops and between the hot and cold legs of each loop to provide protection against consequential failures. The pressurizer is located in a separate compartment. A concrete wall around the entire primary system protects the containment from impact of missiles and reduces the radiation from the primary system to the surroundings.

A water pool for storage of the upper core internals during refueling and for the entire core internals during inspection, is located inside the containment for radiation protection reasons.

3.1.1.5.5. Technical Data

General plant data	
Power plant output, gross, MW_e	1500
Reactor thermal output, MW_{th}	4250
Reactor coolant system	
Primary coolant flow rate, kg/s	21900
Reactor operating pressure, MPa	15.5
Coolant inlet temperature, at RPV inlet, °C	291
Coolant outlet temperature, at RPV outlet, °C	327
Reactor core	
Active core hight, m	4.2
Average linear heat rate, kW/m	15.5
Fuel material	sintered UO_2
Fuel assembly total length, mm	4800
Rod array	square, 17×17
Number of fuel assemblies	241
Number of rods per assembly	264
Number of control rod guide tubes	25
Number of spacers	9
Enrichment (range) of first core, wt %	≤ 5.0
Operating cycle length (fuel cycle length), months	18 (12, 24)
Average discharge bunup of fuel, MW · d/t	60000
Cladding tube material	Zr 4
Cladding tube wall thickness, mm	0.725
Outer diameter of fuel rods, mm	9.5
Number of control rods	89
Absorber material	Ag-80, In-15, Cd-5
Drive mechanism	magnetic jack
Positioning rate (10×75 or 750 mm/min), steps/min	75
Primary containment	
Type	prestressed concrete
Overall form	cylindrical
Dimensions (diameter), m	48
Free volume, m^3	90000
Design pressure	
(DBEs), kPa	~650
(severe accident situation), kPa	~650
Design leakage rate, vol % d^{-1}	< 1
Is secondary containment provided?	yes, reinforced concrete (APC protection)

3.1.2. Boiling-Water Reactors

An experimental atomic power plant marked the beginning of nuclear power plant development in Germany. More than 25 years of valuable experience for the further development of the BWR were gained with this 16-MW_e experimental plant at Kahl for which planning started in 1958 and which was decommissioned in 1985. The same applies to the subsequent demonstration plants Gundremmingen A rated at 250 MW_e (1962 – 1980) and Lingen rated at 267 MW_e (1964 – 1977). These two plants, which also

Table 19. Boiling-water reactors in important industrialized countries

Country	Reactor name	Capacity (Gross), MW_e	Commercial Operation
Italy	Caorso	882	1981
Japan	Fukushima-Daiichi-2	784	1974
	Fukushima-Daiichi-3	784	1976
	Fukushima-Daiichi-4	784	1978
	Fukushima-Daiichi-5	784	1978
	Fukushima-Daiichi-6	1100	1979
	Fukushima-Daini-1	1100	1982
	Fukushima-Daini-2	1100	1984
	Fukushima-Daini-3	1100	1985
	Fukushima-Daini-4	1100	1987
	Hamaoka-1	540	1976
	Hamaoka-2	840	1978
	Hamaoka-3	1100	1987
	Kashiwazaki Kariwa-1	1100	1985
	Kashiwazaki Kariwa-5	1100	1990
	Onagawa-1	524	1984
	Shimane-1	460	1974
	Shimane-2	820	1989
	Tokai-2	1100	1978
	Tsugura-1	357	1970
Spain	Cofrentes	990	1985
	Santa Maria de Garona	460	1971
Sweden	Barsebeck-1	615	1975
	Barsebeck-2	615	1977
	Forsmark-1	1005	1980
	Forsmark-2	1005	1981
	Forsmark-3	1150	1985
	Oskarshamn-1	462	1972
	Oskarshamn-2	630	1975
	Oskarshamn-3	1200	1985
	Ringhals-1	780	1976
United States	Big Rock Point	73	1963
	Browns Ferry-1	1098	1974
	Browns Ferry-2	1098	1975
	Browns Ferry-3	1098	1977
	Brunswick-1	815	1977
	Brunswick-2	815	1975
	Clinton-1	985	1987
	Copper	787	1974
	Dresden-2	812	1970
	Dresden-3	812	1971
	Duane Arnold-1	545	1975
	Enrico Fermi-2	1215	1988
	Fitzpatrick	830	1975
	Grand Gulf-1	1372	1985
	Hatch-1	801	1975
	Hatch-2	806	1979
	Hope Creek-1	1117	1986
	Lasalle-1	1078	1984
	Lasalle-2	1190	1984
	Limerick-1	1092	1986
	Limerick-2	1098	1990
	Willstone-1	684	1971
	Monticello	564	1971

Table 19. (Continued)

Country	Reactor name	Capacity (Gross), MW_e	Commercial Operation
	Nine Mile Point-1	630	1969
	Nine Mile Point-2	1150	1988
	Oister Creek	650	1969
	Peach Bottom-2	1098	1974
	Peach Bottom-3	1098	1974
	Perry-1	1250	1987
	Pilgrim-1	690	1972
	Quad Cities-1	813	1973
	Quad Cities-2	813	1973
	River Bend-1	995	1986
	Susquehanna-1	1068	1983
	Susquehanna-2	1068	1985
	Vermont Yankee	535	1972

made valuable contributions, have since been decommissioned. As the operating experience gained with subsequent boiling-water reactors of larger capacities and operated exclusively for commercial power production continued to grow, these demonstration plants became dispensable because both power plants (Gundremmingen A and Lingen) were found to require highly capital-intensive backfitting measures to improve their safety.

These demonstration plants were followed by the boiling-water reactors belonging to the 1969 Product Line which were rated at ca. 700 – 900 MW_e (Würgassen, Brunsbüttel, Philippsburg 1, and Isar 1 Nuclear Power Stations). These, in turn, were succeeded by the 1972 Product Line. The present stage reached in the evolution of the German BWR is characterized by the large- capacity ratings of the 1972 Product Line plants – approximately 1300 MW_e (Krümmel, and Gundremmingen B and C).

On December 31, 1988, there were 87 boiling water reactors in operation worldwide and 14 under construction. Table 17 shows the status of BWR plants in Germany in 1989. Table 18 illustrates various evolutionary stages by giving characteristic data for typical BWR plants. An overview of BWR plants operating in different countries is given in Table 19.

The following description of the BWR concept is based on Gundremmingen Nuclear Power Station. The principal technical data for a plant of this type are presented in Table 20.

Table 20. Technical data for a 1300-MW_e boiling-water reactor

Overall plant	
Reactor thermal output, MW_t	3840
Gross electric output, MW_e	1308
Net efficiency, %	32.5
Nuclear steam supply system	
Pressure at RPV outlet, MPa	7.06
Temperature at RPV outlet, °C	286
Core flow rate, kg/s	14 300
Steam flow at RPV outlet, kg/s	2 076
Steam moisture at RPV outlet, wt %	0.2
Final feedwater temperature, °C	215
Reactor core	
Active core height, mm	3710
Average specific power, kW/kg U	26.5
Average heat flux, W/cm^2	52.9
Average enrichment of initial core, wt %	2.0
Average enrichment of reload fuel, wt %	2.75
No. of fuel assemblies	784
Fuel	UO_2
No. of fuel rods per assembly	62
Outside diameter of fuel rod, mm	12.3
Overall length of fuel assembly, mm	4470
Total mass of fuel assembly including channel, kg	315
Total mass of uranium in core, t	139
Control rods	
No. of control rods	193
Length of absorber, mm	3660
Absorber material	B_4C
Normal speed, cm/s	3
Normal insertion time, s	122
Scram speed, cm/s	ca. 120
Scram insertion time, s	3.2
Reactor pressure vessel	
Inside diameter, mm	6 620
Clear height, mm	22 350
RPV wall thickness + cladding, mm	163 + 8
Design pressure, MPa	8.73
Design temperature, °C	300
Total mass, t	ca. 785
Reactor recirculation pumps	
Type of pumps	axial flow
No. of pumps	8
Recirculation flow per pump, m^3/h	8731
Pump head, m	31.1
Containment	
Inside diameter, mm	2 900
Clear height, mm	32 500
Design pressure, MPa	0.43

∗ RPV = reactor pressure vessel.

3.1.2.1. Design

The generation of energy in the boiling-water reactor occurs in the reactor pressure vessel at a lower pressure and lower temperature (7.06 MPa, 286 °C) than in the PWR. In contrast to the PWR, steam for driving the turbine is

generated directly inside the RPV. The heat flow diagram of the steam, condensate, and feedwater cycle is shown in Figure 33.

In the BWR (see Fig. 34), heat released by the fission process is used to boil demineralized water inside the RPV (a) at a system pressure of 7.06 MPa. This water enters the core (b) in a slightly subcooled state (215 °C), is heated to boiling temperature, and partially evaporates. The steam – water mixture produced in the core area passes through the centrifugal-type steam separators (d) located above the core where most of the water is removed. The water is returned to the bottom of the RPV and the steam is routed upward to the steam dryers (e). The main steam for the turbine exits the RPV through four main steam nozzles in the RPV wall and flows directly to the turbine.

The steam exhausted from the turbine is condensed, and this condensate is cleaned, deaerated, and heated before being returned as feedwater to the core via feedwater nozzles (h) on the reactor pressure vessel located above the core (b). A sparger ensures uniform mixing of the feedwater around the reactor vessel with the water flow from the steam separators. The slightly subcooled water is routed downward past the core through the downcomer annulus between the RPV wall and the core shroud; it flows along the bottom of the RPV toward the center and then upward again in the direction of the core.

Boiling-water reactors with the power densities that are common today require forced circulation of the coolant. In early reactors with forced circulation, water from the downcomer annulus was taken outside the RPV to external recirculation pumps that returned the water at an increased pressure to the bottom of the vessel. In the United States, another type of coolant recirculation pumping system was then introduced, in which a system of jet pumps is installed in the downcomer annulus of the RPV and only a small amount of water is recirculated externally to provide the driving flow for the jet pumps. In the reactor pressure vessel shown in Figure 34, several axial-flow recirculation pumps (k) are installed directly inside the annulus to provide internal forced circulation of the reactor water.

Reactor Core. The steam is produced in the reactor core, which is located in the bottom half of the reactor pressure vessel (see Fig. 34). The reactor core is composed of cells, each of which consists of four fuel assemblies and one cruciform control rod (see Fig. 35). Two grids, one located at the top and one at the bottom of the core, hold the fuel assemblies in a lattice configuration. The fuel assemblies are supported directly at the bottom of the reactor pressure vessel by the control rod guide tubes. Orifices at the top of the control rod guide tubes ensure proper coolant flow distribution to each fuel assembly. The reactor core is contained within the cylindrical core shroud, which supports the two core grids and the steam separators located on top. Table 20 contains the main technical data for a reactor core. For economic reasons, a high rate of energy extraction from the nuclear fuel is required. The average discharge burnup serves as a measure of this energy extraction. At modern plants, this value is about 28 000 MW · d/t of uranium, which corresponds to an average operating period of 4 years.

Fuel Assemblies. Each fuel assembly consists of fuel rods arranged in a 6×6, 7×7, or 8×8 array (see Fig. 35). The fuel rods consist of thin-walled Zircaloy cladding tubes, which are sealed gastight at both ends, have an outside diameter of 12 – 14 mm and a length of ca. 4 m, and contain the nuclear fuel in the form of pellets stacked over a length of about 3.7 m. Burnup of the fuel is optimized by using ^{235}U enrichments between 0.7 and 3.9 %. Upper and lower tie plates, through which the coolant passes, hold the fuel rods together to form fuel assemblies.

Spacers provided at several axial locations maintain the desired distance between fuel rods. A central rod holds the spacers in position at different levels. A channel made of Zircaloy encloses each fuel assembly. The functions of the fuel channel are to channel the steam – water mixture produced in the reactor core and to guide the cruciform control rods when they are moved. The control rods are responsible for reactor control and shutdown. They absorb the excess reactivity in neutron-absorbing boron carbide (B_4C), which is enclosed in steel tubes in the control rods (see Fig. 35). The excess reactivity of the initial core is absorbed by gadolinium, a burnable neutron absorber, which is added to the fuel for this purpose.

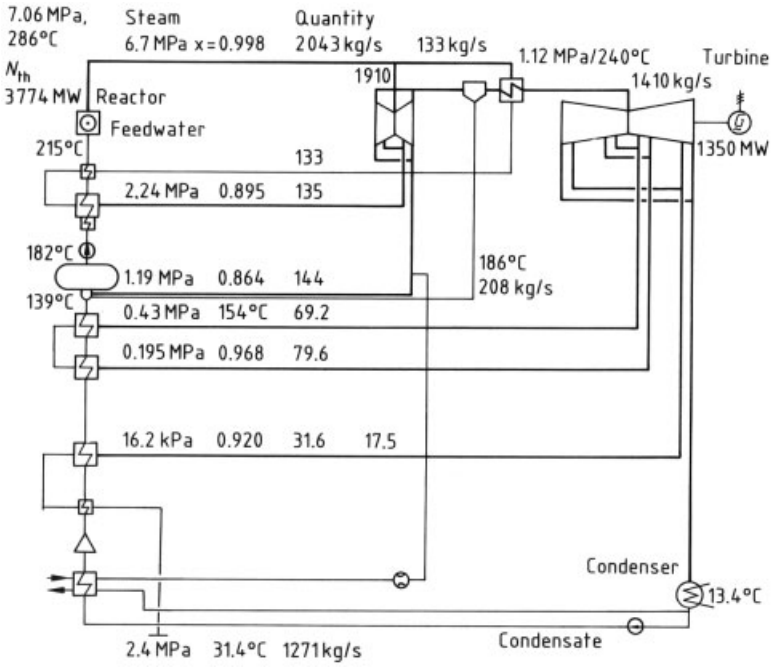

Figure 33. Heat flow diagram

Components of the Reactor Coolant System. In today's boiling-water reactors, the *reactor pressure vessel* encloses all the components required for the generation of steam (i.e., in addition to the reactor core, the steam separators and dryers, as well as the internal recirculation pumps; see Fig. 34). For refueling purposes, the RPV is opened by removing the head. All external piping connections are welded.

A low-alloy, fine-grained steel that is easily workable is an optimum material for use in light-water reactors. Irradiation has only a small effect on its toughness because of the large water gap between the reactor core and the RPV wall, resulting in a service life of more than 40 years. The wall of the reactor pressure vessel is clad on the inside (see Table 20 for main dimensions and design data). Stringent quality controls during manufacture rule out unacceptable material defects. In-service inspections (by the ultrasonic method) are performed to monitor material conditions at specified intervals during the reactor service life.

A *control rod drive mechanism* (CRDM) (see Fig. 34, g) is provided for each control rod. The drive inserts and withdraws the control rod either slowly by electromechanical means or quickly (by means of a hydraulic system) to scram the reactor. The CRDMs are located below the RPV in control rod drive housings that are welded to the reactor bottom head.

Axial-flow pumps located inside the reactor pressure vessel are used as *recirculation pumps* (r) in the reactor (see Fig. 34). Depending on the capacity rating of the reactor, between four and eight pumps are installed in the downcomer annulus. The pump shaft is led downward out of the RPV through a small pump nozzle in the RPV bottom head and is driven by an external variable-speed electric motor. As in a reactor with natural circulation, the entire volume of recirculating water remains inside the reactor.

The *steam separator assembly* comprises individual centrifugal separators. Here, the steam–water mixture that exits the core area is separated into water, which is returned to the downcomer annulus, and low-moisture steam, which flows upward to enter the steam dryer assembly. There, the steam is dried to a residual moisture content of < 0.2 % in accordance with the design of the saturated-steam turbine.

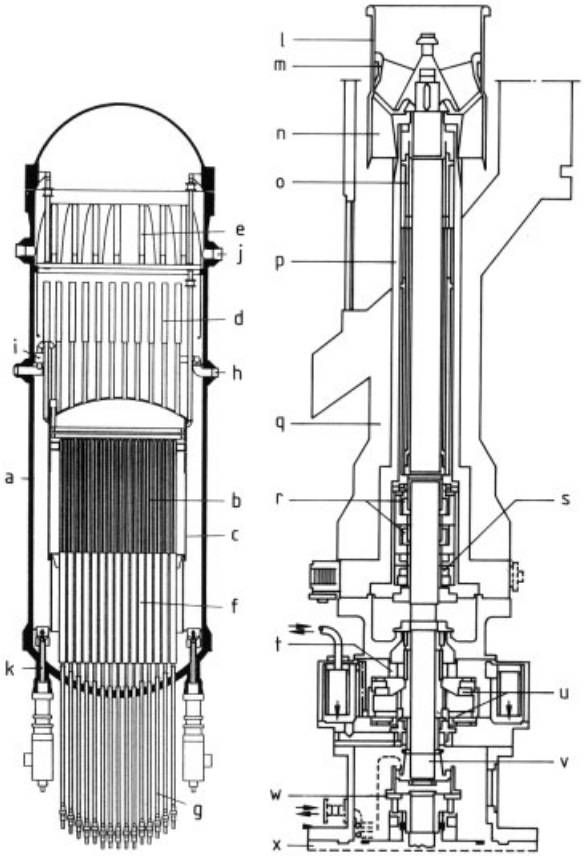

Figure 34. Boiling-water reactor pressure vessel with internals and axial-flow recirculation pumps
a) Reactor pressure vessel; b) Reactor core; c) Core shroud; d) Steam separator assembly; e) Steam dryer assembly; f) Control rod; g) Control rod drive; h) Feedwater nozzle; i) Core spray supply header; j) Main steam nozzle; k) Recirculation pump; l) Inlet pipe; m) Impeller; n) Guide vanes; o) Radial bearing; p) Pump casing; q) RPV pump nozzle; r) Hydrodynamic seals; s) Shaft seal (for plant shutdown); t) Radial bearing; u) Thrust bearing; v) Pump shaft; w) Curved-tooth coupling; x) Motor flange

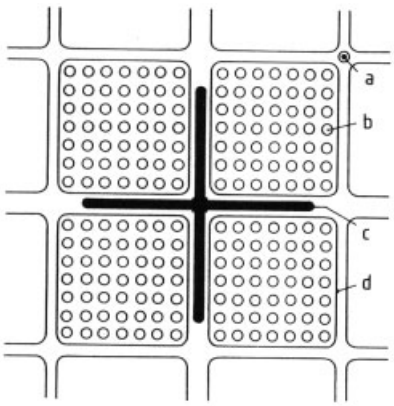

Figure 35. Core cell consisting of four fuel assemblies with 7×7 fuel rods each
a) Neutron detector; b) Fuel rod; c) Control rod; d) Fuel assembly channel

Reactor Building. The reactor building (see Fig. 36) accommodates the systems and components required for reactor operation; protects the plant against natural or man-made external hazards; acts as a final containment barrier to prevent the uncontrolled release of radioactive materials; and protects the environment against nuclear radiation. The building is a cylindrical structure approximately 50 m in diameter and 70 m in height with a wall thickness of approximately 2 m. The cylindrical primary containment with the pressure suppression system is arranged concentrically inside the reactor building. The space between the reactor building wall and the primary containment–the annulus–

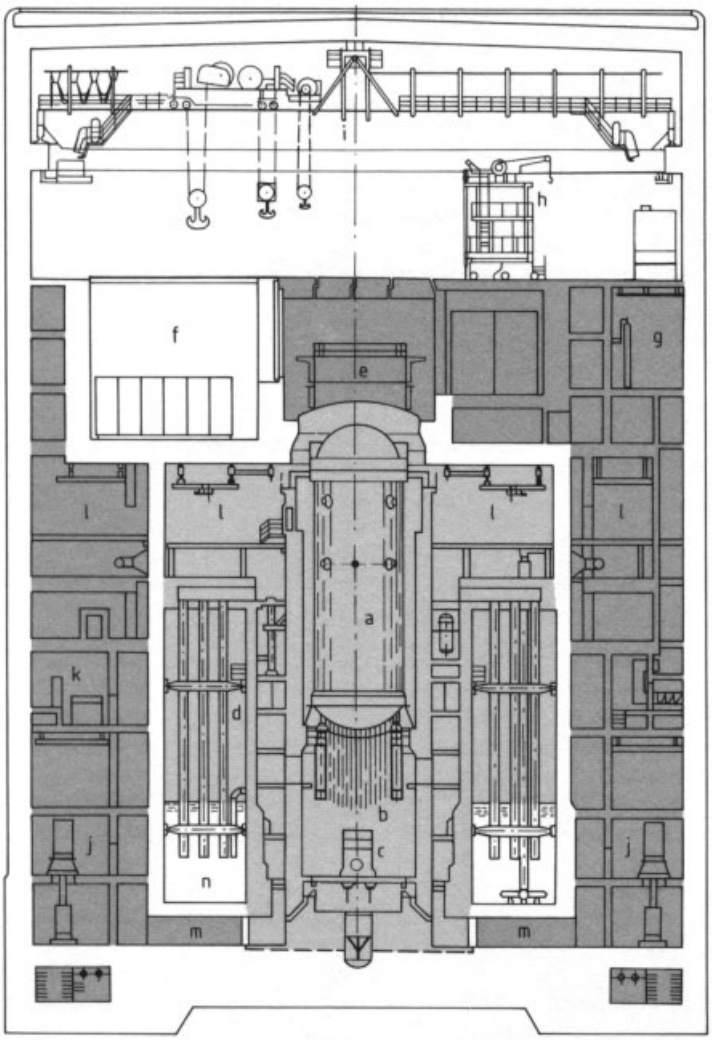

Figure 36. Section through boiling-water reactor building
a) Reactor pressure vessel; b) Control rod drive compartment; c) Assembly platform; d) Wet well; e) Dry well seal; f) Spent fuel storage pool; g) New fuel store; h) Refueling platform; i) Reactor building crane; j) Residual heat removal system; k) Switchgear for recirculation pumps; l) Pipeways; m) Cable ducts; n) Liner

houses the scram system, the residual heat removal, and the nuclear component closed cooling water systems. The reactor well, the fuel pool, and the dryer and separator storage pool are located above the containment.

The *primary containment* is another protective barrier. It is a prestressed concrete structure with a steel lining. The nuclear steam supply system (the RPV, the main steam and feedwater piping up to the first isolation valves, the safety-relief valves, the recirculation pumps, and the CRDMs) is accommodated inside the containment, which is subdivided into the dry well and the wet well. In the event of accidents with or without loss of coolant occurring inside the primary containment, the pressure suppression system serves as a heat sink by condensing the steam in the pool of cold water. A cooling system connected to the pressure supression system provides for heat removal. The pressure suppression system keeps accident pressures to a low level.

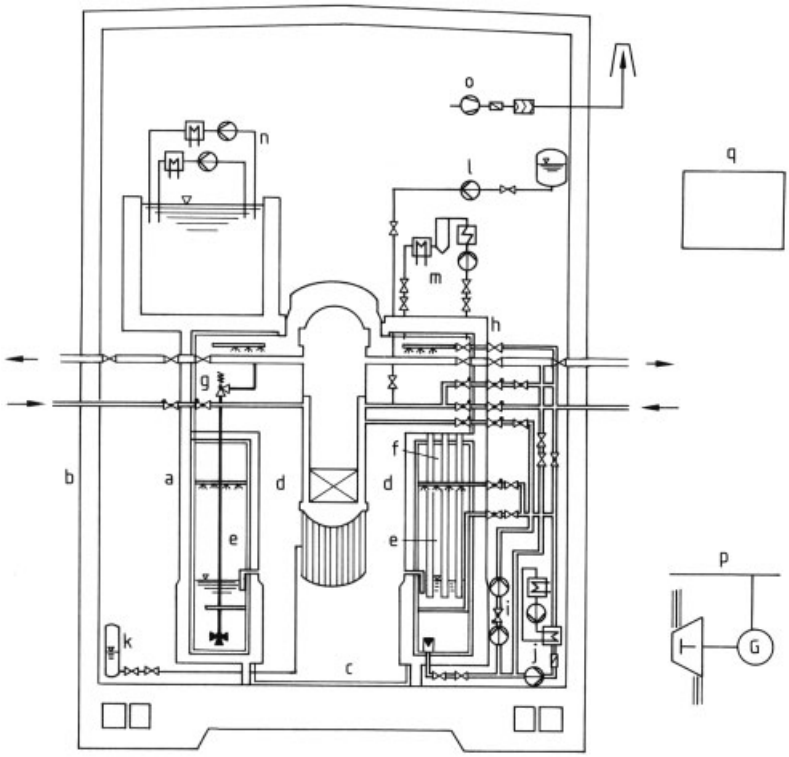

Figure 37. Boiling-water reactor safety-related systems
a) Containment; b) Reactor building outside wall; c) Liner; d) Dry well; e) Wet well; f) Pressure suppression system vent pipes; g) Pressure relief system; h) Isolation valves; i) Residual heat removal system; j) Nuclear component closed cooling water system; k) Scram system; l) Standby liquid control system; m) Hydrogen reduction system; n) Fuel pool cooling system; o) Postaccident subatmospheric pressure control system; p) Emergency power supply system; q) Reactor protection system (control room)

More recent safety scenarios have also considered core melt accidents. To guarantee that the functional integrity of the containment is maintained, it is either inerted with nitrogen or equipped with hydrogen igniters (which ensure timely combustion of any hydrogen released) and provided with an additional venting system for depressurization.

The entire reactor building belongs to the controlled access area. During operation, access is available through airlocks. The containment, however, is normally entered only during an outage when the reactor has been shut down. A slight subatmospheric pressure is maintained at all times inside the containment, and the exhaust air is routed via off-gas filters to the vent stack. All primary system piping that penetrates the containment is provided with two isolation valves.

Reactor Auxiliary and Service Systems. The main safety-related facilities are shown in Figure 37, and the auxiliary and service systems in Figure 38.

The function of the *reactor scram system* is to bring the reactor to a subcritical condition from any state of operation within a few seconds, in the event of accident. This is effected by inserting the neutron-absorbing control rods. Apart from the scram system, a *standby liquid control system* is provided that can be used to shut down the reactor and maintain it in a subcritical condition if the control rod system is inoperable. This system pumps an aqueous solution of boric acid into the reactor pressure vessel.

In the event of an accident, the *residual heat removal system* ensures that the core is adequately covered with water and that decay heat is removed from the reactor after it has been

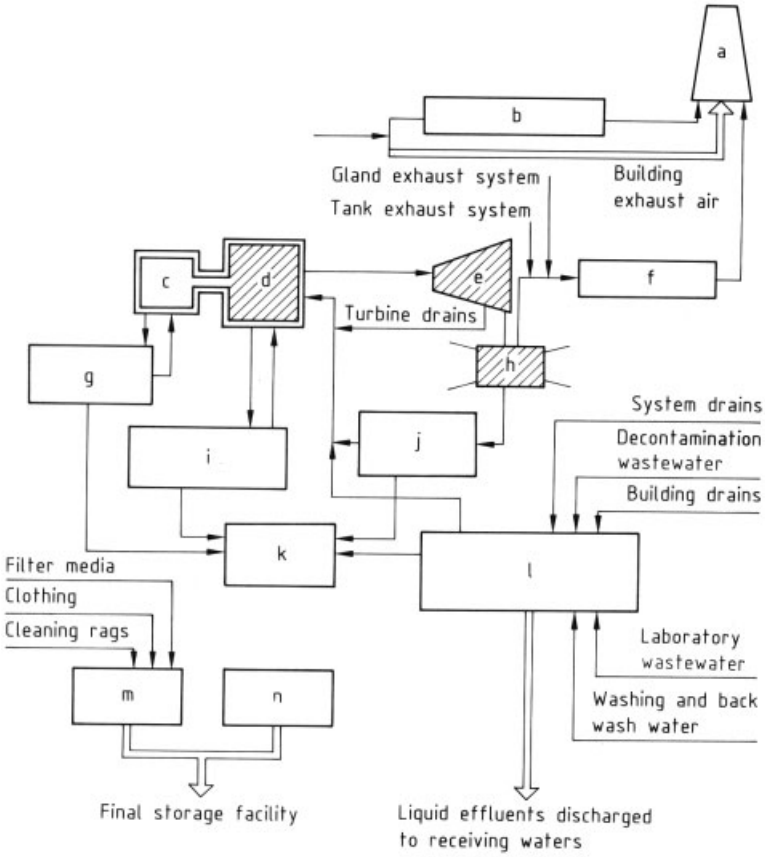

Figure 38. Boiling-water reactor auxiliary and service systems
a) Stack; b) Exhaust air filters; c) Fuel pool; d) Reactor; e) Turbine; f) Off-gas processing system; g) Fuel pool purification system; h) Turbine condenser (air removal system); i) Reactor water cleanup system; j) Condensate demineralizer system; k) Concentrate storage tanks; l) Liquid waste processing system evaporators and filters; m) Drumming station; n) Waste treatment systems

shut down. This system can function in both the high-pressure and the low-pressure range, and is of multiple redundant design. During normal plant operation, this system is used (1) to cool the fuel pool when the complete core has been discharged and (2) to cool the water in the wet well. It also performs certain cooling functions during start-up and shutdown of the plant.

In the *reactor water cleanup system*, a portion of the reactor water flow is continuously purified to reduce the content of soluble and insoluble corrosion products as well as any fission products.

Liquid and solid waste processing facilities are provided to treat liquid and solid radioactive process waste (leakage water, resin and filter backwash water, and water used for decontamination work). Depending on the type of wastewater involved, processing is based on solids filtration, mixed-bed ion exchange, or evaporation. A *gaseous waste processing system* is required for the processing and controlled release of gaseous radioactive wastes. The off-gas system is also connected to the turbine and to the plant heating, ventilating, and air conditioning systems.

3.1.2.2. Instrumentation and Control

The instrumentation and control systems are characterized as follows:

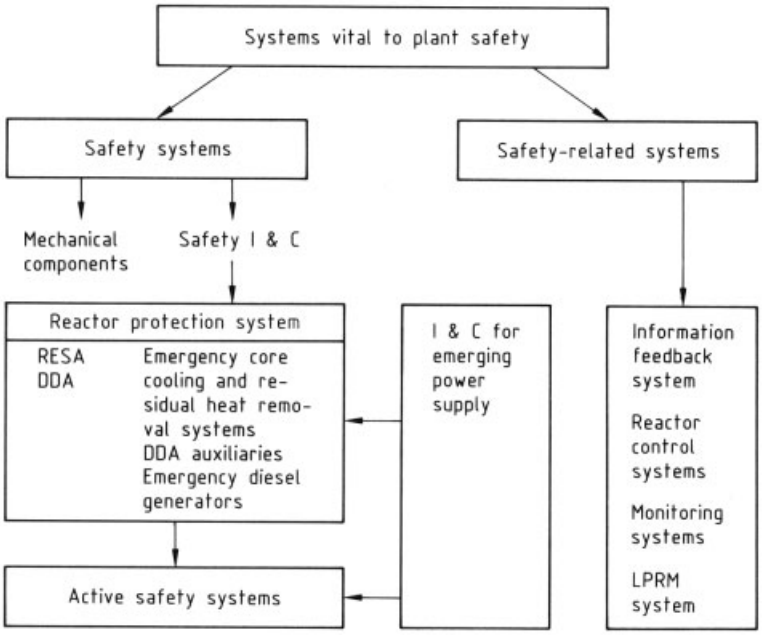

Figure 39. Design of boiling-water reactor safety instrumentation and control
RESA = reactor scram; DDA = containment isolation (closure of double isolation valves); LPRM system = local power range monitoring system

1) Subdivision into functional groups with a redundant, decentralized arrangement, and hierarchical structures
2) Clear separation of protective, monitoring, and control functions
3) Prevention of accident situations by means of equipment and measures based on a multilevel concept of response to prevent unnecessary initiation of a reactor scram and other accident-related safety measures
4) High degree of automation
5) Standardization of signal processing levels and control levels

The instrumentation and control systems are subdivided into safety I & C for the reactor, turbine I & C for the turbine generator set, and non-safety-related process I & C.

Safety Instrumentation and Control. Safety I & C (Fig. 39) consists of electrical and electronic equipment necessary to guarantee safe reactor operation. It includes the instrumentation, process monitoring systems, closed-loop control facilities, reactor protection system, and core monitoring systems. All relevant process variables are measured. The measured signals are processed in transducer compartments that are physically separated from each other.

Reactor Control. The control rods and reactor water recirculation pumps represent the elements for reactor power level control. In control rod control, neutrons are absorbed by the boron carbide contained in the blades of the control rods. In this process, the further the control rods are inserted into the reactor core, the more thermal neutrons are absorbed (causing a power reduction). Conversely, the more the control rods are withdrawn from the reactor core, the fewer thermal neutrons are absorbed (causing a power increase).

Recirculation flow control influences the volume of steam in the reactor core and hence the moderating capability of the coolant. If flow through the core is slowed down by reducing the speed of the recirculation pumps, the steam voids remain longer inside the core, curtail the moderating effect of the water, and hence reduce the power level. If the recirculating water flow is increased, the moderating effect is increased as is the power level.

Two different operating modes are possible (Fig. 40), depending on the intended service of the plant [(1) based-load operation with scheduled load changes or (2) peaking service with unexpected load changes from the partial load range]:

1) Adjustment of the reactor power level by using the control rods with the recirculation pumps operating at minimum speed (Fig. 40, line i) until the 100 % flow control line (line c) is reached (at ca. 60 % reactor power). Above this value, power is adjusted through regulation of the recirculation pump speed.
2) Adjustment of the power level by using the control rods up to approximately 40 % power (line i) and beyond that by using both the control rods and the recirculation pumps in such a manner that the pump speed is maintained constant at about the center of the speed band to enable optimum accommodation of rapid load changes (lines f and e).

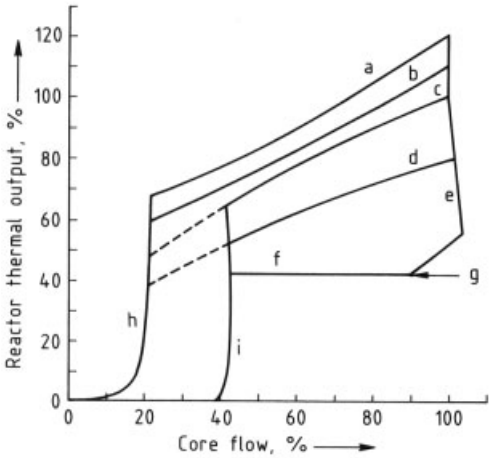

Figure 40. Control of boiling-water reactor by control rod adjustment and recirculation flow control
a) Variable neutron-flux scram limit; b) Variable rod block limit; c) 100 % flow control line; d) 80 % flow control line; e) Pumps at rated speed; f) Pump constant speed band; g) 40 % feedwater flow; h) Natural circulation line (pump speed = 0); i) Pumps at minimum speed

The prerequisite for rapid power control is a nearly constant steam pressure in the reactor. Reactor pressure is controlled by a proportional-integral pressure regulator acting on the turbine control valves or the bypass control valves. From a process control viewpoint, the regulator is assigned to the reactor I & C but the equipment is really part of the turbine I & C system.

A three-element control system is used for reactor water level control. The elements for control during start-up are the start-up valves of the feedwater system and the variable-speed feedwater pumps during power operation.

Reactor Protection. The task of the reactor protection system is to continuously monitor selected process variables that are relevant to the safety of the nuclear power plant and its environment; to detect abnormal conditions; and to process these parameters to form signals for initiating reactor protection system action. Reactor protection must guarantee the following basic objectives:

Reactor scram
Residual heat removal
Prevention of unallowable radioactive release to the environment

If malfunctions escalate into accidents, automatic measures are initiated to control their consequences. The reactor protection system automatically triggers the necessary accident-related measures. No manual intervention by operating personnel is required for the first 30 min after onset of an accident.

Neutron-Flux Monitoring. The local power range monitoring system (LPRM) is used to monitor power distribution in the reactor core. It not only monitors the linear heat generation rate and the margin to film boiling, but also signals alarms on excessive conditions so that appropriate countermeasures can be initiated. The signals used for these purposes are output by LPRM detectors. These LPRM detectors measure the neutron flux, which is proportional to reactor power. Miniature neutron-sensitive ionization chambers are used as detectors, which are distributed uniformly over the core cross-section and the height of the core. Thus, information on the neutron-flux density is available for all fuel assemblies. The fixed in-core ionization chambers are checked and calibrated at regular intervals by traversing in-core probes. The LPRM detector output signals are used (1) by the reactor protection system to monitor the core and (2) by the process computer to monitor and control plant operation.

Turbine Instrumentation and Control. The turbine I & C is integrated into the general I & C concept of the overall plant. In certain cases, however, specific tasks or operational conditions require special solutions. The demands imposed on the dynamic behavior of the closed-loop control circuits regarding control speed, forces, and accuracy require a control system that is specifically tailored to the application involved. This also pertains to the turbine protection system, for which an especially high degree of reliability is required.

Non-Safety-Related Process Instrumentation and Control. The primary function of non-safety-related process I & C is to monitor and control the power plant from the control room. Non-safety-related process I & C forms the interface between operating personnel and power plant processes. It serves to optimize the basic conditions required for plant operation. Personnel are aided by a process computer to achieve this objective. Hard-wired protective logics are used in the control systems to execute simple protective functions. Closed-loop control systems not only provide for automatic process control but also maintain process conditions within specified limits.

3.1.2.3. Operation of the Boiling-Water Reactor

Operation of a reactor plant includes the various modes of plant operation and maintenance activities. Plant operation is divided into start-up, scheduled or unscheduled shutdown, and power operation with or without load following.

Start-up and Shutdown. The rate of plant start-up from the cold condition and run-up to full load is determined primarily by the allowable temperature gradients at the turbine and the reactor pressure vessel. First, the recirculation pumps are run up to a low speed; then the desired reactor power level is set (e.g., heating power by control rod withdrawal). The turbine is run up in the variable-pressure mode. After about 7 h, the required system pressure is attained in the reactor pressure vessel (see Fig. 41).

If reactor shutdown is scheduled, reactor power is reduced and the reactor is shut down by inserting all of the control rods. This causes the turbine to trip. The steam that is still generated by the decay heat in the reactor is dumped to the turbine condenser via the bypass station. The shutdown cooling system goes into operation as soon as the reactor pressure drops below 0.5 MPa.

If, for example, sudden malfunctions occur the plant is designed in such a way that it will automatically initiate an emergency shutdown (reactor scram and turbine trip).

Power Operation. During power operation, the plant is operated at an almost constant power level (base load). In Germany and other countries, however, nuclear power plants are designed to accommodate changes in load demands arising during power operation (load following). The limiting factors for the rates at which load changes may be made are not only the allowable temperature gradients, as in the case of start-up and shutdown, but also the mechanical properties of the fuel assemblies. The Grundremmingen BWR described here is designed for the following load change rates:

$\pm 120\,MW_e/min$	over the entire power range (ramp change)
$\pm 120\,MW_e$	between 70 and 90 % of rated reactor power (step change)
$-120\,MW_e$	between 100 and 90 % of rated reactor power (step change)

If these values are exceeded, a limitation immediately intervenes.

The plant is designed such that the following requirements are fulfilled at all times, not only during power operation but also during start-up and shutdown:

1) *Safety in the Event of Accidents*: If postulated accidents occur, the reactor must be able to be safely shut down and kept in a subcritical condition. The decay heat must be reliably removed over a prolonged period of time and radioactive releases maintained below the allowable limits.
2) *Safe Shutdown*: In case of loss of external load caused by abnormal events in the electrical distribution network, the plant must continue to supply its own auxiliary power requirements and must be available for reconnection to the power system.

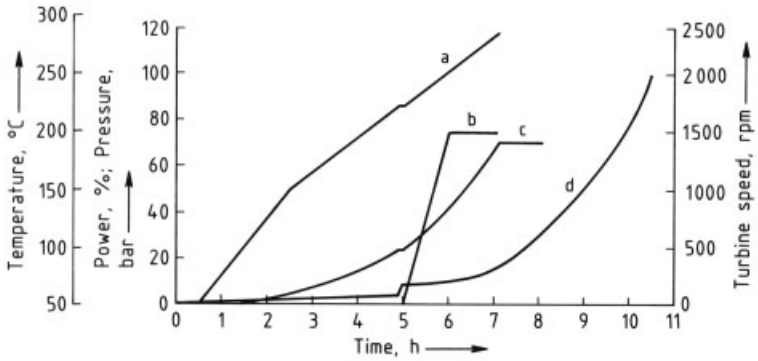

Figure 41. Start-up and loading of boiling-water reactor nuclear power plant from cold condition
a) Temperature; b) Turbine speed; c) Pressure; d) Power

3) *Unit Availability*: In the event of failure of important components (e.g., reactor feed pumps or recirculation pumps), the functions of these components must be taken over immediately by standby components.

The reactor is shut down on the average of once a year for refueling and maintenance.

Refueling. When the fuel assemblies have reached the degree of burnup requiring their discharge from the core, the fuel still contains fissile material in the form of undepleted uranium-235 plus plutonium (plutonium-239 and -241), which is produced in the reactor. Consequently, economics dictates reprocessing of this fuel.

The replacement of about 25 % of the core load with fresh fuel gives the reactor enough excess reactivity to be operated continuously at full load for at least another year. Nuclear fuel developments have, in fact, made operating cycles of more than one year possible. During a one-year cycle, the loss of reactivity due to fuel burnup must be compensated for. This occurs through burnup of the neutron poison incorporated in the fuel and by gradually withdrawing the control rods. At the end of the cycle, all the control rods have been withdrawn from the core. If the reactor is operated beyond the end of the cycle (stretch-out operation), the output will decrease each day by about 0.33 % of the average daily output.

For refueling, the space above the reactor pressure vessel is flooded and the gate connecting to the spent fuel storage pool is opened. For radiation protection and residual heat removal reasons, the fuel assemblies are then discharged or shuffled underwater. The spent fuel assemblies can be kept in interim storage in the fuel pool for a prolonged period of time. Refueling requires approximately 2 to 4 weeks, depending on the size of the plant.

Maintenance. To guarantee reactor safety the plant must be inspected and maintained on a routine basis. This is carried out during the annual refueling outage. Maintenance and inspection requirements must be taken into account in the early planning stages of a plant; these include

1) Adequate accessibility for staff, replacement components, and equipment;
2) Development of automated inspection processes and equipment (e.g., manipulators for ultrasonic and eddy-current examinations);
3) Shielding measures and possibilities for decontamination;
4) Design of components to facilitate inspection and replacement;
5) Installation of monitoring systems (e.g., loose parts monitoring and leak detection systems) for timely identification of defects; and
6) Use of self-checking electronic modules.

The reason for this foresight during planning is to minimize radiation exposure of the personnel and to reduce outage periods and, consequently, costs.

3.1.2.4. Radioactive Releases to the Environment

The many years of experience gained with boiling-water reactors have shown that radioactive releases to the environment are small. The average annual release, (in Bq/a) of radioactive material from a German 1300-MW_e boiling-water reactor may be summarized as follows:

Gaseous effluents	
Noble gases	4×10^{12}
Aerosols	1×10^{7}
^{131}I	1×10^{8}
Liquid effluents	
Fission and activation products	5×10^{8}
Tritium	1×10^{12}

For modern power reactors, radioactive emissions from the stack are ca. 4×10^{12} Bq/a of noble fission gases, particularly ^{133}Xe, and ca. 1×10^{8} Bq/a of ^{131}I.

If normal habits are assumed, persons living in the vicinity of a nuclear power plant are subjected to an additional radiation dose of $< 10^{-2}$ m Sv/a (< 1 mrem/a) (for natural radiation doses see Section 2.3.2.1). Even if the most unfavorable hypothetical conditions are assumed (e.g., an infant is fed for a whole year exclusively on milk from a cow grazing at the most unfavorable receiving point in the immediate vicinity of the power plant), the calculated radiation dose is about 0.5 mSv/a (50 mrem/a) (0.9 mSv/a = 90 mrem/a is permissible for the thyroid gland according to Strahlenschutzverordnung).

The radioactivity released with liquid effluents is less than 5×10^{8} Bq /a, and additional tritium releases are about 1×10^{12} Bq /a, although the latter are unimportant because of the low radiotoxicity of tritium. Under the assumption that a resident receives the entire protein requirement from fish caught downstream of the plant discharge structure and, in addition, that drinking water comes exclusively from this water source, even then the additional radiation dose is $< 10^{-2}$ mSv/a (< 1 mrem/a).

3.1.2.5. SWR 1000

In 1992, a goup of German utilities and Siemens agreed on the development of a new BWR nuclear power plant with emphasis on replacing active safety systems by passive safety features, when possible. The new SWR (Siedewasserreaktor = BWR) will be designed for a net output of 750 MW_e.

The main design goals of this advanced concept are

1) Reliable control of the various design basis accidents
2) Low probalility of beyond-design basis accidents (core melt frequency)
3) Limitation of the impact of core melt accidents to the plant and the environment
4) High plant availability
5) Economic viability of the concept

The new features mainly function because of physical phenomena such as gravity, natural convection, evaporation; or inherent capability of the respective physical phenomena (i.e., temperature coefficient).

At the end of the conceptual phase, it was decided that the new requirements for the advanced SWR, especially economic aspects, were in favor of a concept with an increased power output: i.e., a thermal reactor power of 2778 MW and a net electric power output of 977 MW.

Various features have been modified compared with the existing designs:

1) The power density of the reactor core is reduced
2) Large water storage capacities inside and outside the reactor containment enable long grace periods and make operator interference unnecessary, especially during accidents
3) For transients and for accident control an emergency condenser and an additional containment cooling condenser enable passive decay heat removal from the core and from the containment, respectively
4) The safety functions reactor scram, containment isolation, and primary systems depressurization are backed up by passive systems
5) In case of a postulated core melt accident the installed cooling capacity is sufficiently conservative to prevent an impact on the structures inside the reactor system or on the environment

3.1.2.5.1. Nuclear Systems

Primary Circuit and its Main Characteristics. The nuclear steam supply system is located inside a concrete containment with steel liner and is surrounded by the reactor building. The main dimensions (22.8 m long and 7.0 m in inner diameter) of the reactor pressure vessel correspond to those of the reactor pressure vessel of the 1260 MW_e nuclear power plant in Krümmel (Germany). This means that a comparatively large water inventory is available above the core.

As in the case of large-capacity plants (1300 MW_e), the reactor is depressurized by discharging steam to the pressure suppression pool via safety-relief valves actuated by diverse pilot valves of both active and passive design. A further diverse means for pressure relief is provided by rupture disks. Once the reactor pressure has been sufficiently reduced, water can flow by gravity into the RPV from an elevated pool, the core flooding pool.

To reduce the buildup of pressure inside the drywell in the event of a loss-of-coolant accident, the SWR 1000 has vent pipes located outside the pressure suppression chamber, which terminate at the bottom in inclined ends discharging into the pressure suppression pool water. The routing of the relief lines and the location of the vent pipes outside the air space of the suppression chamber both constitute a further enhancement of safety, because these configurations rule out the possibility that steam can leak into the chamber air space.

The reactor is shut down either by inserting the control rods into the core using the control rod drives, by rapidly inserting all control rods using a hydraulic system, or by rapid injection of boron solution into the reactor water. The accumulators of the latter system, filled with a boron solution, perform two functions: first serving as a redundant backup to the water tanks of the scram system for rapid insertion of the control rods; and second providing a diverse means for stopping the nuclear fission in the core by injecting the boron solution into the reactor water.

Steam is conveyed to the turbine by three main steam lines and the feedwater is supplied to the reactor in two feedwater lines.

Reactor Core and Fuel Design. The main difference between the SWR 1000 core and previous designs is the reduced core height of 2800 mm (usually 3710 mm). Because of the reduced core height, the core can be positioned at a lower level inside the reactor pressure vessel, yielding a larger water volume above the core, which supports the accident control. The lowered core also implies a reduced length of the control rod drives, which means that the RPV can be located at a lower position inside the reactor building.

The reactor core consists of 648 fuel assemblies and 157 control elements. The internal square water channel provides increased and more uniform moderation and leads to a more even power distribution across the ATRIUM fuel assemblies. The ATRIUM design has a large hot-to-cold reactivity range ideal for longe cycles.

The fuel cycle lenght is planned for 24 month with a mean reload burnup of 65 GW · d/t. This requires fuel enrichment in the range of 5.3 wt % ^{235}U.

Fuel Handling and Transfer Systems. Refueling will be performed in a similar way as in the currently operating BWRs in Germany.

Primary Components.

Reactor Pressure Vessel. The reactor pressure vessel (Fig. 42) accommodates the reactor core and the reactor pressure vessel internals. The pressure vessel head is connected to the lower part of the RPV via a flange and fixed by bolts. The complete RPV is supported by brackets/pads in the building's structure. The inner surface of the RPV is lined with stabilized austenitic stainless steel. The pressure vessel is manufactured from 20MnMoNi55 steel. The alloy composition is specified in a very narrow range to optimize the ductility, stresses, and resistance against neutron radiation exposure.

All RPV *internals*, except for the core baffle, are removable to ease the accessibility for in-service inspection.

The six *reactor recirculation pumps* assure sufficient cooling of the reactor core. The power output of the pumps can be controlled by means of the speed control. In case of a failure of one pump, the flow through the core can be held constant by increasing the speed of the remaining pumps. The recirculation pumps are centrifugal pumps driven by wet motors, without mechanical seals and lubricated bearings.

Each of the three *main steam lines* is equipped with three isolation valves, two of which are system fluid-operated and one is motor-operated. One of the fluid-operated valves is located inside the containment, and the other outside it, together with the motor-operated valve.

The two *feedwater lines* are provided with three check valves as isolation valves, one of them is of diverse design. One valve is located inside the containment, while the other two valves are installed outside.

The outer valves of both steam and feedwater lines are located immediately adjacent to the containment.

Reactor Auxiliary Systems.

Safety-Related Systems. The design and mode of operation of the pressure relief system are basically the same as those of the pressure relief systems installed in operating BWR plants (see Section 3.1.2.1).

The application of eight pressure pulse transmitters to initiate reactor scram, containment isolation, RPV depressurization, and passive cooling of the reactor core – dependent on the water level in the RPV – represents a new feature of the design. Further passive systems are the four emergency condensers acting as heat sink for the RPV, as well as a flooding system with four flooding lines working by gravity as a low pressure injection system to control accidents.

Non-Safety-Related Systems. Two fully redundant residual heat removal trains are provided for removing heat from the reactor after shut down and depressurization. Each train consists of components of the RHR-system, closed cooling water and service water systems. Both RHR trains are equipped with pumps and heat exchangers designed to remove heat from the reactor under low-pressure conditions as well as heat from the containment. These systems can also be used to supply coolant to the reactor in case of an accident.

The *reactor water cleanup system*, which is basically a two-train design, continuously removes impurities such as fission and corrosion products from a part flow of reactor water diverted from the reactor. One of the two RHR trains has an additional shutdown filter to enhance reactor water purification while the plant is being shut down. These systems are located entirely inside the containment.

For cooling the control rod drives and also for supplying make-up water to the scram tanks following a reactor scram, a *control rod drive cooling system* is provided, which is equipped with two pumps sucking from the condensate storage tank. This system can be used to supply coolant to the RPV at high reactor pressure levels in case of loss of the normal feedwater supply.

The systems and components provided for reactor refueling, for the storage of new and spent fuel, and for handling RPV internals are not different in their design or mode of operation from those installed in current operating SWR plants. A novel feature is the integration of the spent fuel pool cooling and purification systems inside the pool. This dispenses with external connections to the pool below the water level, thereby reducing the potential for leakage from the pool, and eliminates some potentially highly contaminated circuits in the reactor building.

Operating Characteristics. Operating characteristics include:

1) The average annual availability over the plant service life will be > 87 %.
2) The load-following capability will meet the demands of the German load dispatcher.
3) The collective dose – as long-term average – of in-house and external personnel should be < 0.5 Sv/a (50 rem).
4) The duration of scheduled annual refueling outages shall be < 25 d.
5) The number of forced outages (over the plant service life) should be < 1/a.
6) The average discharge fuel burnup shall be between 50 000 and 65 000 MW · d/t U.
7) The plant shall permit stretch-out operation for 70 full-power days.
8) The reactor shall permit use of mixed-oxide fuel up to 50 % without any modifications to the concept, and up to 100 % with appropriate modifications.
9) The spent fuel storage in the reactor building shall be sufficient for at least 10 years of operation plus one entire core.

3.1.2.5.2. Turbine Generator Plant System

Turbine Generator Plant. The turbine is connected to the RPV by three main steam lines.

Since the bypass station is installed in parallel to the turbine, the entire steam volume generated inside the reactor can be dumped directly to the condenser. The main steam from the reactor is admitted to the high-pressure (HP) section of the turbine via combined isolation and control valves.

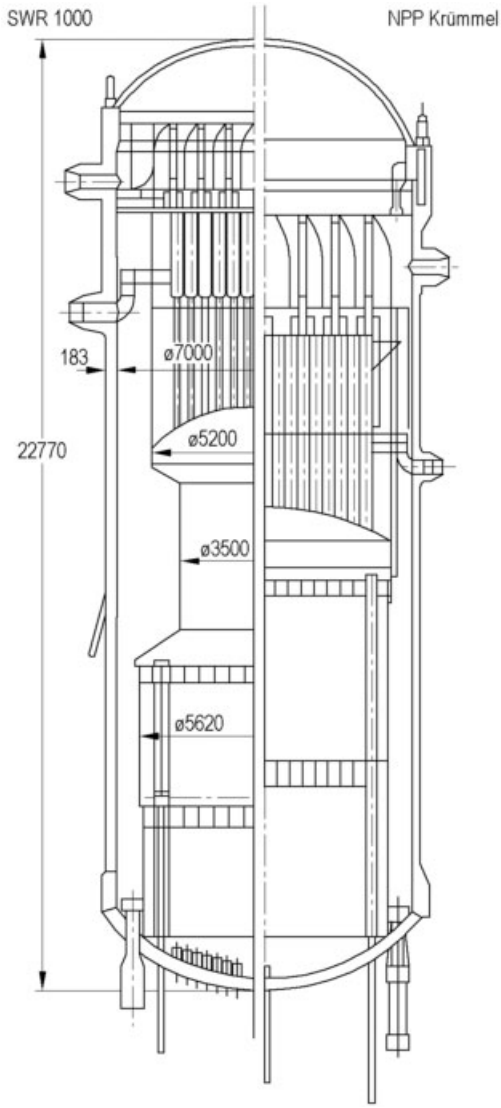

Figure 42. Reactor vessel comparison between SWR 1000 and NPP Krümmel

After partial expansion in the HP turbine section, the steam passes through a moisture separator without reheater to the three double-flow low-pressure (LP) sections of the turbine. The steam expands in the LP sections through several stages of bladings down to the condenser pressure.

Condensate and Feedwater Systems. The condensate collected in the hotwells of the three condensers is discharged into a common header supplying the main condensate pumps. The condensate passes a demineralizing system before being heated up to the final feedwater temperature in the feedwater heaters. The feedwater is injected via two feedwater lines into the reactor using speed-controlled feedwater pumps.

3.1.2.5.3. Instrumentation and Control Systems

The SWR 1000 will use a digital I and C system, covering the following areas:

Operational I and C
Safety-related I and C
Process control systems

Main Control Room. The entire plant process is controlled via the main instrument panel, independent of the mode of operation, i.e., during power operation, start-up, and shutdown of the plant. The process information is gathered in a raised instrument panel. Normal controlling and monitoring results can be adopted from flow diagrams and plasma screens.

Emergency Control Panel. An emergency control panel will be located in one of the emergency power supply buildings. In case of an accident, impact from outside, and non-availability of the main control room, shutdown of the reactor and decay heat removal can be initiated from the emergency control panel if it was not yet initialized by the automatic protection system. The emergency control panel is completely independent from the main control room and houses:

1) Control and supervision of both decay heat removal systems
2) Indication and documentation of special parameters of the incident monitoring system and long range I and C
3) Additional indicators and documentation and especially selected process parameter, necessary for function control of passive safety features.

Local Panels. Systems, that are necessary only for part-time operation and that need not be directly accessible from the control room, are operated from local control panels. Examples are the emergency diesel and sampling system.

Monitoring Equipment. The *neutron flux level* is directly monitored in the reactor core by means of fixed neutron detectors. The neutron flux monitoring system has the following functions:

1) Monitoring of the neutron flux density and periodic changes during shutdown conditions, during start-up from cold, subcritical conditions up to 125 % of nominal power
2) Monitoring of full power core conditions
3) Monitoring of the three-dimensional power distribution in the core

The proper functioning of the reactor depends on reliable monitoring of the following operation parameters: (1) reactor pressure, (2) reactor water level, (3) flow rate through the core, (4) live steam flow rate, and (5) temperature.

For control purposes, redundant measuring channels will be provided enabling the fulfillment of functions during operation, i.e., shutdown and incidents/accidents control.

Reactor Protection System and Other Safety Systems. The safety systems of the SWR 1000 consist of both active and passive safety features, and the latter will normally be actuated directly by process parameters without signal from the safety-related I and C.

The *safety-related I and C* generally detects and records accidents and/or incidents, and initiates active safety systems. In case of nonavailability or disturbances in active safety systems, the passive features take over to bring the nuclear power plant to safe condition.

The detection of an accident is identified by various process parameters, on which the uncertainties and the impact of common cause failures, based on direct measured information, can be governed.

If these requirements can not be fulfilled, a pressure increase inside the reactor vessel and/or a reactor water level decrease will lead to intervention of passive safety systems. A redundant design prevents immanent system failures due to loss of one redundancy because another will directly take over.

The redundancies are protected within the system and the reactor plant against failure-induced events, and each sub-system is protected against the other by physical separation. The redundancies are located in the two emergency diesel buildings, which are protected against external impacts. In each redundancy a control panel is included, from which the present plant status can be assessed, and manual actions to shut down the plant, for pressure reduction, and decay heat transfer can be initiated. A battery-backed energy supply system is incorporated in each of the redundancies. For function control of the passive safety features relevant process parameters are controlled and recorded. For these functions a separate battery-backed energy supply system is installed in the switchgear building.

During shutdown of the plant and opened reactor pressure vessel, repairs may be carried out in the two redundant safety systems. In such cases the passive safety features will be quite inactive, the functions normally dedicated to the safety systems have to be taken over by operational auxiliary systems.

3.1.2.5.4. Electrical Systems

The generator feeds into the 400 kV grid via the main transformer. The *auxiliary power* necessary for plant operation is extracted between the generator and the main transformer and fed into the 10 kV switchgear system via two auxiliary transformers. The auxiliary system is divided into two trains, each consisting of a 10 kV main distribution and a 690 V and 400 V low voltage distribution.

In the event of loss of auxiliary power supply, safe shutdown of the reactor plant requires that certain consumers are started or connected and that others remain in operation. Power supply to safety-related systems therefore can be divided into two categories:

1) Loads that may be de-energized for a period of 20 s during start-up of the diesel generating units are connected to the a.c. power distributions in the emergency power supply system.
2) Loads, which can not accept an interruption in operation or which have to be started

immediately, are connected to the secured power supply system. They are fed by the 220 V d.c. system or via a static d.c. – a.c. inverter and the connected 400 V distribution

With respect to accident control, the following process systems are available, besides passive safety features:

1) 2×100 % decay heat removal systems
2) 2×100 % additional systems with diverse emergency power consuming components

For emergency power supply of these systems a two train emergency system is installed.

3.1.2.5.5. Safety Concept

The safety concept of the SWR 1000 is based on accident prevention and mitigation. Innovative measures are introduced in the accident mitigation concept leading to remarkable higher safety standards.

With respect to prevention of accidents, all proven measures of current safety concepts assuring reliable operation are incorporated in the SWR 1000 plant design.

All requirements mentioned in accident prevention rules for PWRs in Germany have been considered, as far as they would be applicable to a BWR. Further precautions against hypothetical core melt accidents have been taken to limit the impact on the plant.

The new concept advantageously uses the reduced power density in the core and the large water reservoirs in the RPV, condensation chamber, and in the flooding pool. It also takes advantage of the active and passive system functions, which are redundant and diverse relative to each other.

Safety Systems and Features (Active, Passive, and Inherent). The necessary safety functions for accident control are ensured by redundant active systems. The design of the active systems takes into account the existence of additional passive systems to realize the deterministic redundancy concept in connection with the single failure strategy.

For fast shutdown of the reactor, a scram system with two independent trains activates the control rods, each train covering 50 % of the rods. Each train has a normal water and a borated water storage tank. They have two functions: the energy for insertion of the control rods is delivered from the system and the boric acid injection into the core is considered as a diverse shutdown system. In addition, the control rods can be moved into the core using electric drives, which are connected to the back-up energy supply system.

The main steam and feedwater lines are provided with three normally open isolation valves. One valve is located inside, the other two outside the containment.

The I and C system generates the signals for closing of the main steam valves via electromagnetic pilot valves, which operate in fail-safe mode. The check valves in the feedwater lines do not need any control signal.

Following an isolation from the main heat sink, the eight medium-operated safety and relief valves will blow steam down into the condensation chamber with its pressure suppression system to prevent excessive pressure in the RPV. These safety and relief valves also have the function to relieve the pressure in the RPV in a loss-of-coolant accident. To keep the main valves open during decreased RPV pressure they are mechanically locked.

Core flooding can be performed by two existing operational systems. High pressure injection is carried out with the control rod flushing system and low pressure injection with the residual heat removal system. Active components of the two-train systems feed all control rod drives with flushing and cooling water.

The two-train residual heat removal system will also be used as a low-pressure core flooding system during loss-of-coolant accidents. The water is transferred from the condensation chamber in the RPV under a pressure < 1 MPa. The feeding capacity of one train is sufficient for this operating mode.

Under accident conditions (i.e., failure of the main heat sink, loss-of-coolant accident), which lead to a separation from the main heat sink, the residual heat removal cooling system together with the condensation chamber of the wetwell (flooding pit-water cooling system) is used for heat transfer from the containment to an ultimate heat sink.

For passive *initiation* of reactor scram, isolation of main steam lines, and automatic depres-

surization of the RPV a passive pulse transmitter is installed. In case of a reduced RPV water level steam enters the transmitter and heat is transferred to a secondary medium that, in turn, will actuate, e.g., pilot valves.

In the scram system diverse passive pilot valves are installed in addition to the magnetic pilot valves in each train for passive initiation of scram.

The passive trail of the RPV pressure limitation system (see Fig. 43) includes redundant and diverse features. Besides the active magnetic pilot valves one extra passive spring-loaded pilot valve is planned for each of the eight safety and relief valves. In addition to the passive control of the safety and relief valves eight redundant and diverse rupture disks are installed. For pressure relief, which is only necessary during a LOCA, each of the safety and relief valves will be equipped with a diaphragm pilot valve actuated by a passive pulse transmitter.

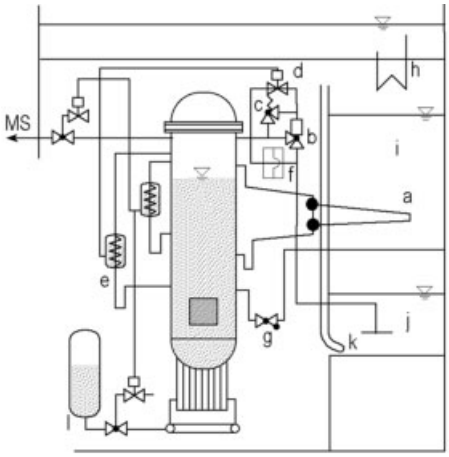

Figure 43. SWR 1000 – Passive System

Pos.		Number
a	Emergency condenser	4
b	Safety-relief valve	8
c	Spring-loaded pilot valve	8
d	Diaphragm pilot valve	8
e	Passive pressure pulse transmitter	2×4
f	Rupture disk	8
g	Flooding line	4
h	Containment cooling condenser	4
i	Core flooding pool	
j	Pressure suppresion pool	
k	Vent pipes	15
l	Scram system	2

The *emergency condenser* has two functions: To release the decay heat to the flooding pool water; and to recycle the condensate back into the RPV is a fully passive component. It is located in the core flooding pool and is connected to the RPV by nonisolatable steam discharge and return lines. It starts operation upon dropping water level in the RPV.

For passive heat transfer from the containment, a separate containment cooling condenser will be installed, divided into 4×25 % units. The steam released into the containment by a LOCA or evaporation of water from the core flooding pool condenses in the containment cooling condenser and is recycled into the flooding pool. The condenser works by natural circulation.

Severe Accidents (Beyond Design Basis Accidents). In case of a core melt (which has a very low probability of occurrence), flooding lines from the core flooding pool will open to supply water to the outside of the RPV, so that the melt can be cooled while still inside the vessel. This ensures retention of most of the radioactive aerosols inside the containment.

The release of the most significant isotope, ^{137}Cs, amounts to approximately 10^{11} Bq. This is far below the level of 10^{14} Bq, which has been internationally accepted. The containment is inerted and the release of hydrogen would not lead to either combustion processes or an explosion. The design of the containment is such as to preclude loss of containment integrity under such pressure loadings.

3.1.2.5.6. Design

The location of the different buildings of the nuclear power plant can be seen from the plot plan (Fig. 44).

The central unit consists of the reactor building with integrated stack and the turbine building with integrated auxiliary building. The reactor auxiliary building with the entrance to the controlled area and the switchgear building with the main control room are located adjacent to these buildings.

At a certain distance from the power plant buildings and the cooling water entrance, there

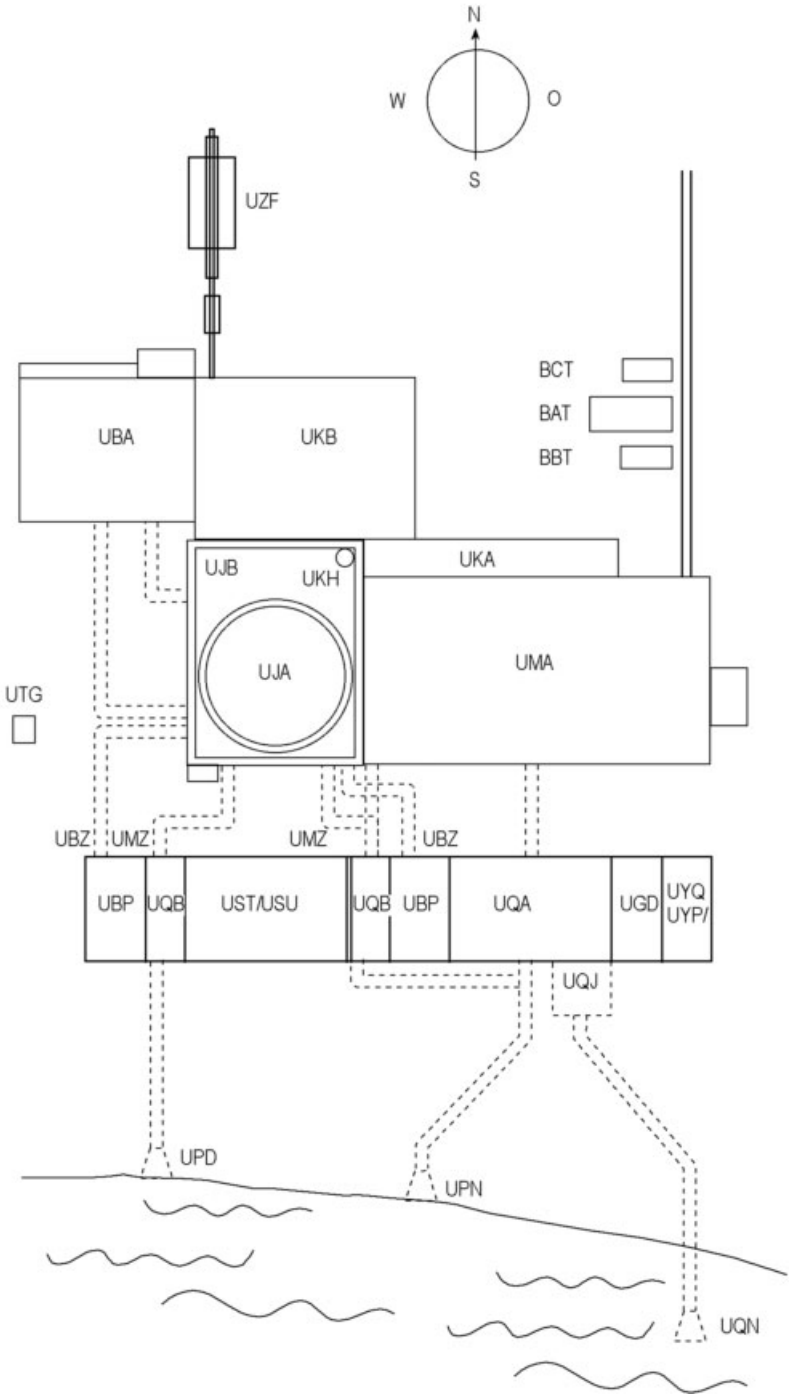

Figure 44. SWR 1000 – Plot plan
UBA = Switchgear building; UJA = Containment; UJB = Reactor building; UKA = Auxiliary system building; UKB = Service building; UMA = Turbine building

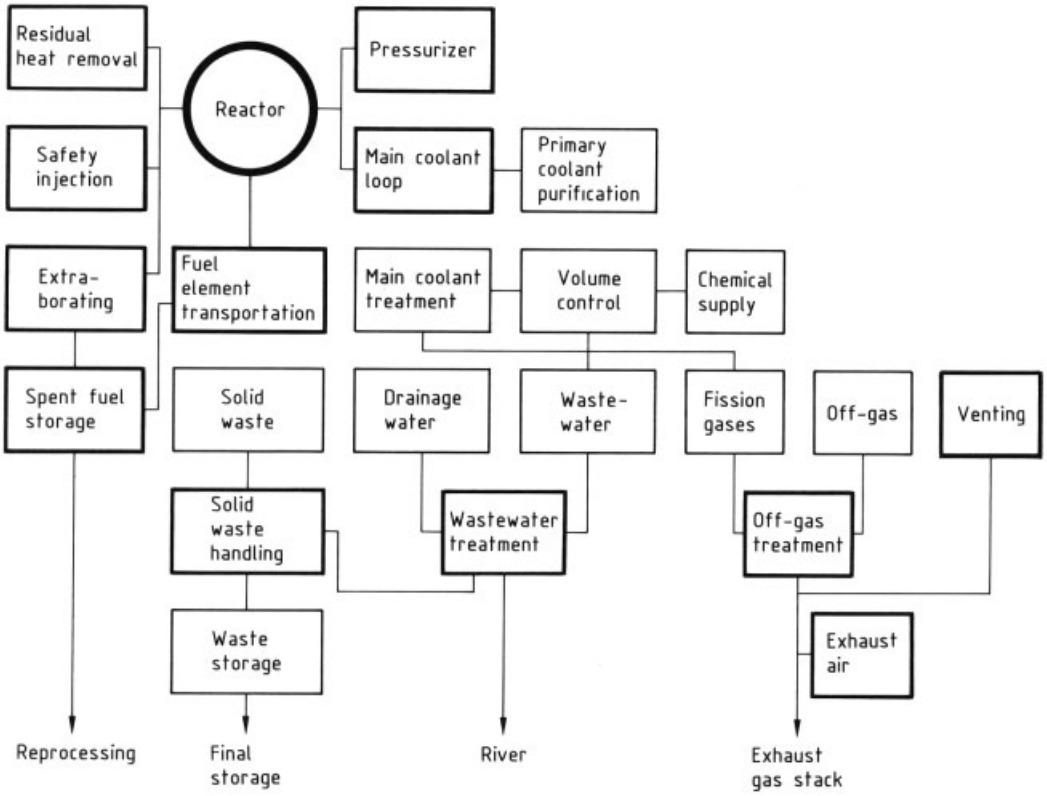

Figure 45. Auxiliary and supporting systems in a nuclear power plant

are the buildings for the pump station, emergency diesel, and the auxiliary systems. Access by the staff during plant operation is possible via a personnel bridge from the social area of the auxiliary building.

Reactor Building. The reactor building accommodates all systems and components necessary for reactor operation and reactor protection. The complete fuel storage and handling facilities are located in the protected area. The other compartments not occupied by the primary system are used for installation of reactor auxiliary systems. Short cable and pipe routes are chosen. Redundancy is reached by physical separation of safety-related systems. The compact arrangement of systems and components enables optimal radiation protection and emergency protection. The reactor building accommodates the following installations:

1) Containment
2) Fuel handling and storage facilities
3) Scram tanks and pumps for control element drive flushing
4) Condesate tank
5) Ventilation and air conditioning systems for the controlled area
6) Radioactive waste concentrate tank
7) Off-gas system (delay beds)
8) Evaporation plant
9) Evaporator feed tank
10) Active waste collection tank
11) Condensate flushing tank of the main condensate system

The dimensions of the reactor building result from the size of the RPV and the condensation chamber, the flooding pool, the laydown area, and the fuel storage pool.

Containment. The pressure-retaining components of the reactor system (mainly the RPV) and the directly connected systems up to the second isolation valve are installed inside the containment.

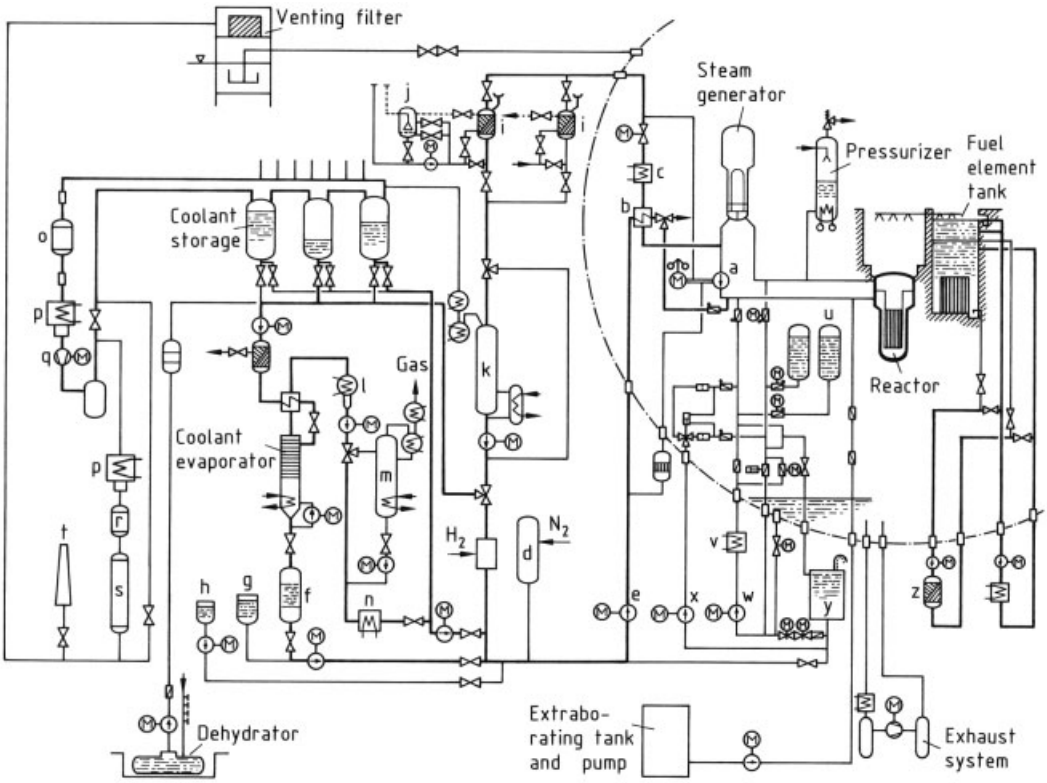

Figure 46. General flow sheet of a PWR auxiliary system
a) Primary coolant pump; b) Regenerative heat exchanger; c) High-pressure cooler; d) Volume control surge tank; e) High-pressure charging pump; f) Boric acid storage tank; g) Boric acid and batch tank; h) Chemicals mixing tank; l) Ion exchanger; j) Spent resin tank; k) Degasifier; l) Condenser; m) Demineralized water degasifier; n) Destillate cooler; o) Recombiner; p) Predryer; q) Off-gas compressor; r) Gas dryer; s) Fission gas delay bed; t) Exhaust-air stack; u) Pressure tank; v) Residual heat exchanger; w) Residual heat removal pump; x) High-pressure safety injection pump; y) Refueling water tank; z) Spent fuel pit cooling water purification

The containment also incorporates some of the auxiliary systems.

The *plant components* are located below the condensation chamber. In this area the following components are installed: (1) residual heat removal pumps and coolers; (2) fast opening tank isolation valves for the control rod drives; (3) reactor water cleanup system pumps; and (4) reactor water cleaning coolers and filters.

The heat transfer out of the compartment is achieved by means of gas circulation coolers. The compartments are vented and can be entered by the staff.

The *drywell* accommodates the flooding pool and inner cylinder and houses the following components:

1) RPV
2) Feedwater and main steam lines and isolation valves
3) Safety and relief valves with exhaust pipe
4) First isolation valves of the residual heat removal system
5) Control rod drives and ring pipes of the Scram system
6) Recirculation pumps
7) Emergency condenser with condensate and steam pipes
8) Containment cooling condenser
9) Reactor water purification cleanup filter
10) Passive pulse transmitters
11) Flooding tubes
12) Measuring pipes
13) Vent pipes
14) Air circulation units (3×50 %)

The *wetwell* is filled with water up to about 40 %. For pressure suppression during incidents the released steam is injected into the water where it condenses. The vent pipes and the quencher tubes enter the water in the condensation chamber.

Turbine Building. The equipment of the water-steam cycle (including turbine and generator) the preheaters, and the feedwater system (including feedwater tank and feedwater pumps) are located in the turbine building. At the lowest floor the main cooling water pipes enter the building. The condensers are arranged transversal and occupy the room up to the turbine operating floor. The cooling water inlet and outlet of the condenser are held short. The preheaters are situated on the intermediate sealing adjacent to the turbine.

Other Buildings. The tanks used for wastewater storage and treatment are located in the *service building*.

The work shops, the waste treatment and storage facilities, and the central entrance to the controlled area are located in the *auxiliary system building*. The chosen design leads to short pipe and cable routes to the systems located in the reactor building. The auxiliary system building accommodates the following installations:

- Solid waste treatment
- Decontamination plant
- Fresh and off-air systems
- Hygiene rooms, especially for the controlled area (showers, lavatories, toilets)
- Laboratory
- Hot and cold work shop
- Office and storage rooms

The complete staff movement is being checked and guided. The entrance to the controlled area is located at about 13.30 m. For the transport of the components into the hot work shop the equipment has to pass the airlock.

The *switchgear building* accommodates two energy supply units, physically separated from each other, and constitutes the redundancy for the switchgear and I and C systems. Each part has its own cable, fresh, and off-air system.

3.1.2.5.7. Technical Data

General plant data	
Power plant output, net, MW_e	977
Reactor thermal output, MW_{th}	2778
Power plant efficiency, net, %	35.2
Reactor coolant system	
Primary coolant flow rate, kg/s	12000
Reactor core	
Active core hight, m	2.8
Average core power density, kW/L	48
Fuel material	sintered UO_2
Rod array	$12\times12-(4\times4)$ in square lattice
Number of fuel assemblies	648
Number of rods per assembly	128
Enrichment (range) of first core, wt %	4.95
Enrichment of reload fuel at equilibrium core, wt %	$\sim$5.3
Operating cycle length (fuel cycle length), months	24
Average discharge bunup of fuel, MW · d/t	65000
Cladding tube material	annealed, recrystallized Zr 2
Fuel channel/box material	Zr 4
Active length of fuel rods, mm	2800
Burnable absorber, strategy/material	Gd_2O_3 in fuel
Number of control rods	157
Drive mechanism	electromechanical
Soluble neutron absorber	Boron
Reactor pressure vessel	
Inner diameter of cylindrical shell, mm	7000
Total height, inside, mm	22800
Base material	
cylindrical shell and RPV head	20MnMoNi55 steel
lining	stainless steel
Design pressure, MPa	8.8

3.1.2.5.8. Project Status and Planned Schedule

The contact between Siemens and the German utilities runs for a period of 4 years, from 1995 to 1999. During this period, the necessary R & D activities shall be completed, the basic engineering work performed, and licensing documents prepared. The BWR shall be licensable in Germany in accordance with German rules, regulations and requirements, and also meet the European Utility Requirements. The conceptual design shall be established in early 1997.

Tests have been performed on the emergency condenser and these tests have confirmed the calculated performance; the full cooling capacity was attained earlier than expected and design margins could be quantified.

3.1.3. Safety and Operating Systems and Process Equipment for Light-Water Reactors

The special physical and technical conditions of energy production by nuclear fission, as well as radiation-related problems, call for special solutions that often differ from those employed in conventional thermal power plants. Requirements going beyond the conventional ones are set forth in pertinent legislation [82], regulations [83], codes [84], [85], and recommendations [86]. An additional standard for safety-relevant equipment in nuclear facilities is that such equipment must correspond to the state of the art. In addition, this equipment must be inspected at regular intervals and replaced when appropriate.

The three most important *safety criteria* in the design of safety systems are [87]:

1) Shutdown of the reactor must be possible from any operating condition.
2) Residual heat removal must be possible over an indefinite period.
3) Release of more than the permitted amounts of radioactive substances into the biosphere must be prevented.

In a pressurized-water reactor, the safety criteria (safety systems) and the requirements for power production (operating systems) are met as follows: grouped around the primary coolant loop are different systems connected directly to it. Depending on their function, these systems can be contaminated with different levels of radioactivity; therefore, they are located in the controlled area. Design of these systems and their components must take into account the following important *objectives* beyond the usual engineering rules:

1) The systems and equipment must be appropriately shielded because of the radioactivity they contain.
2) The components must often be placed in separate compartments so that they will be accessible and reparable during operation.
3) In many cases, processes must be remotely controlled or automated.
4) Extreme safety requirements lead to multiple redundancy in the design of many systems.

Figure 45 gives a general overview of the systems and equipment, with interconnections between related systems indicated. The way in which these systems work together is treated in more detail in the following sections (see also Fig. 46). Systems whose task is to fulfill the *safety criterion* are indicated in heavy outline in Figure 45. Systems relating to safety objectives (1) and (2) are, of course, directly connected to the nuclear reactor. Those intended to prevent the escape of radioactive substances [safety objective (3)] are located peripherally. Systems and equipment in between are needed chiefly for optimal power operation.

The *central system* includes the reactor with the primary loop and the pressurizing system connected directly to it (see page 46). Residual heat removal (RHR) and safety injection systems provide cooling for the reactor core in an operating status to be defined later.

The reactor has two independent shutdown systems. The first consists of a number of absorber rods. The second–the extraborating system–injects a "neutron poison" (concentrated boric acid solution) into the core when needed.

During reactor operation, a side stream is bled continuously from the primary loop, cleaned, and fed to the *volume control system*, from which it is returned to the primary loop. The cleanup system removes unwanted and interfering fission and corrosion products from the coolant. The volume control system evens out volume fluctuations in the primary loop and, together with the *chemical control system*, ensures that the coolant composition is optimal for the requirements imposed on it (see Section 3.1.3.1). Excess water is supplied to the coolant treatment system, where boric acid is removed from the coolant and the concentrated boric acid treated with water is stored separately for reinjection.

Surplus water from the reactor system, after gas stripping, is combined with wastewater to be treated. In the *wastewater treatment system* (see Section 3.1.3.7), the combined stream is cleaned up, decontaminated, and then released into the river water under monitoring. The *fission gases* obtained by stripping are led to the *off-gas treatment system*. After sufficient decay time the gas is diluted with exhaust air of the ventilation system and released to the atmosphere via the stack.

Radioactive solids are collected, compacted, and appropriately packaged. After interim stor-

age at the reactor site, they are shipped to the disposal facility (see Section 5.5.3). A similar procedure applies to *spent fuel assemblies*. After being unloaded from the reactor vessel, these are placed for interim storage in cooled fuel storage pools; after a sufficiently long decay time, they are packed in shielded shipping containers and transported to the reprocessing plant.

The *venting system* prevents the containment (i.e., the final barrier) from failing after a hypothetical accident, after a postulated blackout of all safety systems.

3.1.3.1. Volume Control System

The volume control system has the following tasks:

1) Removal of water from the primary system for purification;
2) Maintenance of the proper water level in the pressurizer during operation, start-up, and shutdown;
3) Provision of water for the sealing-fluid system of the primary coolant pumps;
4) Maintenance of a hydrogen excess in the coolant to prevent water radiolysis in the core; and
5) Injection of demineralized water and boric acid in the proportion required for reactor power control.

The maintenance of optimal coolant conditions required for the existing operating mode is attainable more easily and economically, if an appropriate part of the primary coolant stream is bled off downstream of the steam generator and brought to a low temperature (50 °C) and pressure (1.2 MPa, as shown in Fig. 46). The partial bleed off stream is controlled by the pressurizer water level; the feedwater flow stays constant. Since the mass flow rate in the pressure-reducing station downstream of the high-pressure cooler is controlled by the pressurizer water level, the level in the pressurizer can be held constant. After cleanup, the bypass coolant stream (at a reduced pressure of 1.2 MPa) flows back to the suction side of the high-pressure charging pumps. The proper hydrogen concentration is set either in the volume control surge tank itself or in a special gas saturator device. The latter has the advantage that the quantity of free hydrogen stored in the gas injector is very small compared to that in the surge tank; thus, the danger of fire or explosion is reduced.

Before reinjection into the primary loop, the coolant is preheated to ca. 280 °C in a recuperative heat exchanger. The sealing water for the primary coolant pumps is provided by a partial coolant stream that is led through very fine mechanical filters and then into the stuffing boxes of the pump shaft. Leakage from the shaft seal reenters the system downstream of the reducing station.

3.1.3.2. Chemicals Control System

The chemicals control system has the following tasks:

1) To store and feed in boric acid,
2) To feed in demineralized water, and
3) To inject chemicals for water chemistry control (see Section 3.1.5 for more details).

The excess reactivity of the core, which declines continuously throughout the reactor campaign (run), can be compensated by a variety of measures, including (1) absorber rods that can be inserted in the core as needed (see page 45); (2) burnable poisons that can be integrated into the fuel; and (3) water-soluble neutron poisons such as boric acid (see page 44). *Absorber rods* always have a nonuniform distribution in the core. As a result, the neutron spectrum is distorted. *Soluble poisons* have an optimal action because of their uniform distribution, so that a better burnup of fissile material can be achieved. Since no change of phase (i.e., evaporation) occurs in a PWR, this kind of reactor can be operated with a clearly controlled water chemistry (in contrast to a boiling-water reactor); in other words, the boron concentration can be adjusted to suit existing operational needs.

The required coolant boron concentration is ca. 1050 ppm boron on start-up after refueling and decreases by ca. 3 ppm/d at full power. The boron concentration is matched to the instantaneous reactor power demand. It follows the load cycles and ensures an optimal power density distribution in the core (lower boric acid concentration at higher power output and vice versa). For this purpose, either boric acid or demineralized water is supplied into the coolant on

the suction side of the high-pressure charging pumps (see Fig. 46). Chemical mixing tanks supply lithium [^{7}Li] hydroxide to adjust the pH during operation and hydrazine to bind free oxygen after every refueling or after repairs (see Section 3.1.5.1.1 for more details).

3.1.3.3. Primary Coolant Purification and Degasification System

The primary coolant bypass stream taken out downstream of the steam generator is passed through the coolant purification and degasification system (Fig. 46) to limit the concentration of corrosion and fission products (^{131}I, ^{58}Co, ^{60}Co, ^{137}Cs, ^{133}Xe, ^{85}Kr, etc.). Coolant purification involves a combination of mechanical and ion-exchange filters. The variety and extremely low concentrations of isotopes to be removed, as well as the relatively high concentration of added chemicals, which may be expensive (^{7}Li), call for careful monitoring of this process to achieve the highest possible decontamination factors for radioisotopes and the lowest possible consumption of ^{7}Li.

The ion-exchange resin, with a service life of one or two years, is not regenerated after it becomes exhausted but is disposed of as radioactive waste. Regeneration would again place the radioisotopes in highly dilute solutions, which would be unsuitable for reprocessing.

The water in the fuel storage pool is cleaned in a similar way to the coolant. For practical reasons, these two systems are located close to each other because of common components (e.g., resin flushing pump). Gas strippers connected downstream of the filter system can be employed to remove noble-gas fission products from the coolant. These gases should be removed before opening the closure head of the reactor vessel for refueling.

3.1.3.4. Primary Coolant Storage and Treatment System

The functions of the primary coolant storage and treatment system are as follows:

1) To receive and buffer incoming coolant,
2) To separate boric acid and water (distillation),
3) To store demineralized water and boric acid for reuse, and
4) To prepare excess coolant for discharge.

The coolant resulting from changes in boric acid concentration and the excess water from expansion when the temperature is increased from 50 to 300 °C on reactor start-up are retained in *primary water storage tanks* (see Fig. 46). Water and boric acid are then separated in coolant evaporators. The relatively high steam volatility of boric acid must be considered in evaporator design. Concentrate (4 % H_3BO_3) is taken out continuously of the evaporator's reboiler and led to the *boric acid tank*. The distillate (demineralized water) passes back to the primary water storage tank that was drained for this operation. The tanks are automatically switched from water containing boric acid (coolant) to demineralized water, and the evaporator is automatically turned on and off, so that storage volume and operator effort can be kept to a minimum.

The closed coolant cycle not only offers economic benefit but also minimizes the release of radioactivity. In three special cases, however, water or boric acid may have to be released from the loop:

1) To make up the effective boron isotope (^{10}B), whose concentration declines during operation as a result of (n, γ) reactions
2) To limit unfilterable tritium (^{3}H), which is produced from hydrogen or boron and escapes from the fuel element as a fission product
3) To prevent overfilling of the primary coolant system

Before it is discharged to the environment, the coolant is evaporated, degasified, and condensed to remove all dissolved or gaseous radionuclides; the resulting water is then monitored in the wastewater treatment system and may have to be retreated there.

3.1.3.5. Off-Gas Treatment System

Nitrogen is used as blanket gas in all systems where primary coolant is stored or processed. To prevent noble fission gases from escaping into the building atmosphere through leaks in

the systems, a slight subatmospheric pressure is maintained in all tanks with the aid of the off-gas system (left in Fig. 46). As a result, both fission gases and hydrogen (from conditioning of the coolant; see Section 3.1.5.1) are present in the gas blankets of the tanks; if leaks occur in the system walls, atmospheric oxygen may also be present. Excessive hydrogen levels (danger of explosion) are prevented by using the off-gas system compressors to purge all tanks continuously with recirculating gas (96 – 100 % N_2, 1 – 3 % H_2, < 1 % O_2, $\leq$ 0.01 % noble gases). In this way, the hydrogen concentration is equalized and kept under the ignition limit (4 %) (Fig. 46 shows only the purging of the primary water storage tanks and one of the two compressors). The recirculating compressor draws the gas mixture through a recombiner, where hydrogen is catalytically burned with added oxygen. Protecting the recombiner area against explosion is especially important.

A bypass of the purge gas stream is led through a *delay line*, an adsorption unit operated at saturated equilibrium. In the relatively long path through this unit, individual gas molecules have several opportunities to undergo exchange processes with similar and identical molecules on the activated carbon (at kinetic equilibrium). The holdup unit is sized so that the hold up time of individual radioactive noble gases is roughly 10 times the half-life $t_{1/2}$ of the most important radioactive isotopes (^{133}Xe, $t_{1/2} = 5.65$ d; ^{85}Kr, $t_{1/2} = 4.4$ h). The radioactivity is thus reduced by a factor of $2^{10} \approx 1000$; for practical purposes, this is sufficient. Gas cleaned in this way is returned to the main stream. If maintaining lower pressure in the off-gas system requires the extraction of small amounts of gas because of inward leakages, this gas is automatically bled off downstream of the delay line, mixed with exhaust air, and bled to the stack.

Besides the methods described here, other options exist for the treatment of radioactive noble gases. Examples are storage in cyclically operated tank batteries (charging, decay, release), chemisorption processes, and low-temperature adsorption. Low-temperature adsorption has the advantage of better separation factors and, under some conditions, longer storage (decay) times. For practical use, however, the method described above is quite adequate.

The selective *absorbing* power of certain organic compounds (such as Frigen) can also be used to hold up the noble gases.

The reduction of radioactive noble-gas releases to extremely low levels is limited because some leaks cannot be prevented in systems filled with high pressure primary coolant; that is, a certain amount of noble gases can escape without passing through the off-gas treatment system.

3.1.3.6. Residual Heat Removal and Emergency Cooling System

After reactor shutdown in normal operation and during incidents or accidents, the residual heat removal and emergency cooling system removes the remaining decay heat. The heat source in a nuclear power plant differs from that in conventional steam boiler plants in several ways, one of which is that the heat production of the fuel assemblies cannot be reduced to an arbitrary extent. Some residual heat occurs even after shutdown because the radioactive fission products continue to decay (see Section 1.3.10).

Since this residual heat is an inherent quality of the reactor, core cooling must be ensured by a large safety factor to prevent damage to the fuel assemblies and the release of radioactive fission products. The RHR system is, therefore, highly redundant in design. Because safety regulations require operability of the heat removal system in the event of an additional single failure in passive systems that do not depend directly on the immediate damage, redundant systems must be totally independent of each other.

The importance of these systems in view of safety objective (item 2 in second sequential list in Section 3.1.3) has led to precise statements of their design criteria [88]. Residual heat removal is handled by a four-loop arrangement with four equivalent partial systems (trains), any one of which can deliver 50 % of the rated cooling requirement even in the case of the maximum credible accident. Each of these four trains has a cooling chain independent of all the others. The loop filled with primary coolant–the first link in the cooling chain–gives up its heat to a closed component cooling loop. Only this loop is cooled with water from open sources (rivers, ponds, cooling towers, etc.). The chain arrangement means that radioactivity can-

not escape from the primary coolant loop into the open atmosphere in any operating or accident situations. Safety regulations also mandate closed-component cooling systems for all cooling points in other systems that contain radioactive media.

The extreme high reliability of the RHR system requires that no two cooling chains be situated close together, so that a single failure in one redundancy cannot entail a damage in the nearby redundant train. Naturally, the safety systems have independent emergency power supplies and are also protected against earthquakes. Another essential point is that not only the mechanical parts of the cooling chains, but also the entire electrical power and transmission systems, are physically separated. Appropriate fire protection measures prevent damage to more than one cooling system at a time in case of fire.

An example of a cooling system whose layout corresponds to the state of the art is illustrated in Figure 46 (right of center). The following tasks must be carried out:

1) Removal of residual core heat on normal shutdown; cooling down to near room temperature if necessary
2) Flooding and draining of the reactor well during refueling
3) Residual heat removal from irradiated fuel assemblies stored in the fuel storage pool
4) Flooding and emergency cooling of the core after a loss-of-coolant accident (maximum credible accident: loss of coolant due to total rapture of a primary loop pipe)

Each system is provided with adequate reserves of borated water located in the refueling water tanks. Low-pressure or high-pressure injection pumps draw coolant from these tanks as needed. The suction side of the low-pressure safety injection pump is also connected to the containment sump; this branch is always closed in normal operation.

One train of the *low-pressure safety injection system* includes the refueling water tank, the RHR pump, and the RHR system heat exchanger. Inside the containment, the system is subdivided into the *hot leg*, which joins the primary loop between the reactor pressure vessel and the steam generator, and the *cold leg*, which is located between the reactor pressure vessel and the primary coolant pump. The accumulators are also part of the low-pressure injection system. They store concentrated borated water under higher pressure (e.g., 4.0 MPa). In normal operation, the control elements (check valves) in the hot leg can be opened freely after the reactor has been shut down and the pressure reduced sufficiently. For the residual heat removal from the core, the primary coolant is taken out through this piping and reinjected through the RHR cooler with the RHR pump (also called the low-pressure injection pump) into the cold leg.

A train of the *high-pressure safety injection system* includes essentially the high-pressure safety injection pump with associated piping and fittings. The high/low-pressure and pressurizer feed lines are combined before injection into the primary loop.

As Figure 46 shows, the injection pumps and accumulators are isolated from the reactor core only by check valves (two in series at each connection). In normal operation, the large pressure difference keeps these closed. If, a large leak permits coolant to escape from the primary loop, the primary pressure decreases rapidly and the lower pressure automatically allows injection. In this way, the system is made highly reliable. When the RHR loops (RHR pumps and RHR system heat exchangers) are activated after reactor shutdown, the residual heat of the core in any normal operating status can be removed regardless of whether the fuel assemblies are all in the reactor, all in the fuel storage pool, or some in each.

The borated water stored in the *refueling water tanks* is used to flood the reactor compartment above the pressure vessel during refueling. The diverse functions of this safety system are described by means of its main operation modes. Depending on the size of a possible leakage the pressure in the primary loop falls slowly or rapidly. Then the reserve water from the refueling tank is immediately injected by the high-pressure safety injection pumps (low mass flow rate, high head) or the RHR pumps (high mass flow rate, low head). In the event of a *small* or *medium-sized leak*, the high-pressure safety injection pumps come into play. After the pressure in the primary loop has fallen and the reserve water from the refueling water tank has been injected, the RHR pumps reinject water from the sump. In the *design case*

(the maximum credible accident for which the cooling system is designed e.g., total failure of a primary coolant pipe), the accumulators automatically go into action first to flood the core immediately. Flooding is then continued by the RHR pumps. As soon as the refueling water tank is completely empty, further core cooling occurs in a semiopen loop: containment sump → RHR pump → RHR system heat exchanger → primary injection → reactor pressure vessel → overflow at leak → containment sump.

The fourfold redundancy of the system enables one of the trains to be repaired during operation. Even in this case, one further train can become unavailable because of single failures. The remaining two trains ensure core cooling. The containment, inside which all components in contact with high pressure and primary coolant are located, is designed so that all the steam escaping from the primary loop in an accident is retained. Cooling of the reactor core can thus be continued for any length of time until measures have been taken to remedy the damage.

The approach described above is for any conceivable accident with an extremely high probability. After Three Mile Island (1979), and especially after Chernobyl (1986), hypothetical design accidents more serious than those previously used were defined. One such is the postulated *station blackout*: failure of all electrical power, including all diesel-powered generators, possibly, but not necessarily, combined with a loss-of- coolant accident. Analysis of such accidents yields scenarios involving a core meltdown. To cope with these accidents, it is advisable to provide cooling by passive measures (without the use of energy) involving no pumps (e.g., by natural convection). In addition, there are considerations to maintain the possibly molten core in a coolable configuration even outside the reactor pressure vessel by proper design measures in the containment sump area.

One consequence of such hypothetical accidents would be the formation of *large amounts of hydrogen*; after a meltdown, this could be produced either by metal – water reactions in the reactor vessel, after a possible core melting or from the water contained in concrete structures. In parallel with hydrogen formation or in a later phase of the scenario, decay heat from the molten core could cause an increase in pressure in the containment and, ultimately, uncontrolled failure (leakage) of the containment. Containment failure is prevented by the *pressure relief system*, a new feature of all German reactors that can serve as an example of compliance with the legislated requirement that equipment keep up with the state of the art. The passive pressure relief system consists of a relief line, attached to the containment, that vents the overpressure to the atmosphere through an apparatus with a venturi scrubber and an aerosol filter. The filters are sized to prevent inadmissible contamination of the environment.

Another means to limit the pressure build-up in the containment after such a severe accident is a *containment heat removal system* equipped with an independent cooling chain and electrical power supply. Such an *active system* would offer the possibility to reduce the pressure in the containment to atmospheric level and would thus limit the release of the containment atmosphere into the environment. Together with an appropriate confinement design (e.g., double wall containment) the consequences of even a severe accident would be limited to the plant site itself–a requirement which is part of the new atomic law in Germany and valid for the next generation of nuclear power plants.

A variety of ways are being studied to control the danger of detonation resulting from high hydrogen levels in the containment. The problem is most likely to be solved by some combination involving autarkic igniters and large-area hydrogen oxygen recombiners. Such a system could be adapted to existing plants, such as the pressure relief system or the containment heat removal system . The consistent design of an entire facility against all possible hypothetical accidents will depend on analysis and developmental work that is not yet complete.

3.1.3.7. Waste Treatment

To keep environmental pollution as low as possible, all media leaving the controlled area must be treated. *Wastewater* is intercepted by drains as close as possible to the source and led to a central collecting station. Depending on its activity as well as chemical and physical qualities, the wastewater is subjected to mechanical fil-

tering, evaporation, or ion- exchange treatment. The maximum decontamination factor (ratio of inlet radioactivity to end product radioactivity) varies from ca. 10 for mechanical filtering to $10^3 - 10^5$ for evaporation. A variety of processing steps can also be configured in series.

Decontaminated water is led to monitoring tanks, checked for radioactivity, and released only when below the allowed radioactivity limit. All wastewater from the controlled area must be discharged through a single pipe equipped with recording instruments that measure the mass flow and the radioactivity. If the permissible activity level of 10^{-4} Ci/m^3 (3.7×10^6 Bq/m) is exceeded, a valve in the discharge pipe closes automatically.

Concentrated low- and intermediate-level wastes (filtrate, ion-exchange resins, evaporator sludge) must be converted to storable products. Under regulations in most countries, this waste is incorporated into matrices with low water solubility that can tolerate the expected radiation dose for at least 100 years without substantial physical or chemical changes. Suitable matrix materials are bitumen, concrete, and certain plastics (see Section 5.5.2 for more details).

Solid radioactive wastes (mechanical filter media, contaminated garments, machine parts etc.) are also gathered at central points, compacted (in ram presses), packed in standard radiation-proof and radiation-resistant containers, similar to those used in wastewater treatment, and shipped to a disposal facility such as an old mine (see Section 5.5.4 for more details).

3.1.3.8. Ventilation System

The ventilation system serving the controlled area must perform the following tasks:

1) Maintain the air temperature and humidity at specified levels,
2) Maintain specified pressure differences and properly directed flows, and
3) Lower the airborne radioactivity in the controlled area (to protect operating personnel) and limit the activity discharged through the stack (to protect the environment).

Compartments to be ventilated are grouped in accordance with the following criteria:

1) Compartments to which access is not allowed during normal operation because of intense direct radiation. There is mostly large-scale heat generation in such compartments (especially those containing high-pressure equipment).
2) Compartments to which access is allowed during normal operation (including access only after equipment is shut down and reactivity is sufficiently reduced).
3) Compartments to which access is freely allowed (control rooms, passageways, etc.).

In the first group of compartments, closed air circuits remove the dissipated heat in heat exchangers. Both air circulation and fresh-air cooling can be employed in the second group, depending on equipment and operating costs. Freely accessible compartments are supplied with fresh air. A reduced pressure of ca. 0.05 – 0.2 kPa relative to the outside atmosphere is maintained in all compartments in the controlled area; compartments with higher activity levels are held at lower pressures than those with lower levels.

The mass flow of the exhaust air consists of the fresh air fed into the second and third compartments and the leakage air that originates from the reduced pressure in the entire control area. If necessary, this air must be passed through aerosol filters (fine mechanical filters) or, if radioactive iodine (chiefly ^{131}I) is present, through activated carbon filters.

3.1.4. Materials for Light-Water Reactors

The materials used in the construction of pressurized-water reactors and boiling-water reactors are similar, and they are discussed together here. Any differences are pointed out in the discussion.

The selection of materials for full-scale industrial equipment depends on the qualities required. These attributes vary between different applications, and are also governed by the service conditions and the safety profile.

Material selection for systems and components of light-water reactors differs from that for fossil-fueled power plants because of significant differences in design, service conditions, and postulated safety margins. This counts, above

all, for the primary coolant and some components of the water – steam circuit and in the nuclear auxiliary systems [89].

Table 21. Properties of Zircaloy-2 (Zry 2) and Zircaloy-4 (Zry 4)

Properties	Zircaloy-2	Zircaloy-4
Chemical composition, wt %		
Sn	1.20 – 1.70	1.20 – 1.70
Fe	0.07 – 0.20	0.18 – 0.24
Cr	0.05 – 0.15	0.07 – 0.13
Ni	0.03 – 0.08	
Fe + Cr		< 0.28
Fe + Cr + Ni	0.18 – 0.38	
Modulus of elasticity E, MPa	96 000	
Strength values at room temperature (recrystallized)		
Yield strength $R_{p\,0.2}$, MPa	≤ 241	
Ultimate tensile strength R_m, MPa	≤ 413	
Total elongation, %	≤ 20	
Mean coefficient of thermal expansion α, K^{-1}	ca. 6.5×10^{-6}	
Density ϱ, g/cm^3	6.55	
Macroscopic effective thermal-neutron cross-section ε_a, cm^{-1}	0.0079	

3.1.4.1. Materials for the Core Region

Materials in the core include those used for fuel assemblies (aside from fuel, see Section 5.3.1) and for control and shut-down (both control and shut-down elements have an identical design) elements, which must meet fundamentally different requirements, as well as those for the pressure vessel internals (core structure).

Materials for Fuel Assemblies. Fuel-assembly materials must have the following properties:

1) Low neutron absorption
2) Dimensional stability and microstructural stability under neutron bombardment
3) Corrosion resistance with respect to the water or steam coolant (fuel cladding tubes must also be compatible with the fuel)
4) Mechanical stability (i.e., strength and ductility)

Light-water reactor fuel assemblies consist of cladding tubes for fuel rods (which separate fuel and coolant), top and bottom end pieces, intermediate spacer grids, and guide tubes (see Section 5.3.1).

Virtually the only alloys used for *cladding tubes* in present-day Western LWRs are Zircaloy-2 (BWRs) and Zircaloy-4 (PWRs) (see Table 21). In LWRs of the former USSR design, the principal cladding tube alloy is Zr1Nb.

Other *structural materials* are chiefly Zircaloy-4 plates for spacer grids; Zircaloy-2 and 4 for fuel-rod end plugs; cast or forged and welded austenitic stainless steel (i.e., 1.4550) for top and bottom end pieces and Inconel X (BWRs) or Inconel 718 (PWRs) for spacer springs.

More recently, new zirconium alloys for cladding tubes with improved corrosion resistance are under development. This work is in progress in several countries. They are expected to be commercially available around 1993.

Zirconium technology can serve as a model for the deliberate development of new materials for a new application [90], [91], [92].

The primary industrial process for obtaining zirconium today is from zircon ($ZrSiO_4$), which occurs extensively on the coast of Australia in the required quality. The undesirable companion element hafnium (→ Hafnium and Hafnium Compounds, Chap. 5.2.,→ Zirconium and Zirconium Compounds, Chap. 1.4.2.) is separated; the zircon concentrate is reacted with chlorine to form zirconium tetrachloride ($ZrCl_4$), and this product is reduced with pure magnesium by the Kroll process (→ Zirconium and Zirconium Compounds, Chap. 1.4.3.) [93].

All zirconium alloys are melted in high-vacuum electric-arc furnaces with consumable electrodes; remelting generally follows. Semifinished product is obtained by conventional hot forming (forging, hot extrusion). Since zirconium is sensitive to gas absorption (oxygen, nitrogen) at high temperature, the heat treatment of semifinished and finished products is usually carried out under vacuum. In cold working (rolling, drawing, rocking), the anisotropic deformability of zirconium alloys in the (hexagonal) alpha structure (below ca. 800 °C) must be taken into account. The textures formed can influence service properties (e.g., radiation-induced growth) [94].

Zirconium is sufficiently stable against nuclear reactions initiated by neutron bombardment. The nuclear-grade metal (ASTM B 350 Grade R 60001) has a macroscopic effective

thermal neutron cross-section $\varepsilon_a = 0.0079\,cm^{-1}$ (for comparison, iron has an ε_a of $0.22\,cm^{-1}$).

Corrosion [95] at high pressure (7.0 – 15.0 MPa) in hot water or steam up to 360 °C is characterized by the formation of a usually adherent oxide film (ZrO_2):

$$Zr + 2\,H_2O \rightarrow ZrO_2 + 2\,H_2$$

In *pressurized-water reactors* with high thermal efficiency, this form of cladding corrosion has become a design limiting factor. The rate of corrosion is governed largely by the temperature at the metal – coolant interface. In *boiling-water reactors*, locally accelerated "nodular" corrosion of Zircaloy-2 and Zircaloy-4 cladding and structural parts has led to increased attention internationally. This type of corrosion depends chiefly on irradiation and water chemistry.

Along with increased corrosion, the absorption of hydrogen in zirconium alloys has become increasingly important. This effect results in part from the corrosion reaction and in part from the hydrogen excess (ca. 2 – 4 ppm) present in PWR coolant (see Section 3.1.5). High hydrogen absorption in zirconium alloys leads to the formation of hydrides and thus to embrittlement of the material.

The mechanical properties of Zircaloys can be regarded as satisfactory. They have a texture-dependent tendency to growth, induced by neutron irradiation, although the effect is quite manageable.

Materials for Control and Shut-Down Elements. Absorber materials must satisfy the following requirements:

1) High neutron-absorbing capacity in the thermal and epithermal ranges (< 0.01 MeV)
2) Dimensional stability under (neutron) irradiation
3) Mechanical stability
4) Corrosion resistance if the materials are used without protection

In *boiling-water reactors* the principal absorber material for control and shutdown rods today is boron carbide (B_4C), usually in granulated form (grain size, ca. 100 µm, ^{10}B content 18.9 %). Powdered boron carbide is charged into tubes or hollow drilled plates of austenitic stainless steel (1.4550), which are assembled into cross-shaped control elements.

The effective isotope ^{10}B is transformed to ^{7}Li and helium by (n, α) reactions. The associated swelling of boron carbide must be taken into consideration in the design of the control and shut-down devices. As an alternative some of the boron carbide-filled rods have been replaced by rods or strips of solid hafnium, which is generally used without cladding because of its high corrosion resistance.

For *pressurized-water reactors*, a ternary alloy of silver, indium, and cadmium (80 – 15 – 5 wt %) is used as the absorber material in control, and shut- down elements. The alloy is made into rod shape and inserted into tubes of austenitic stainless steel (i.e., 1.4550), which are closed by tight welding. In contrast to boron carbide, the neutron absorption reaction of this material is based on (n, γ) reactions, which cause little change in the density and mechanical properties of the absorber material.

In addition to absorber materials in the control and shut- down devices of BWRs and PWRs, gadolinium oxide (Gd_2O_3) is added to uranium dioxide fuel as a burnable absorber (poison; see Sections 3.1.1 and 3.1.2).

3.1.4.2. Materials for Components of the Nuclear Steam Supply System

Two essentially different material concepts are employed in the primary loops of pressurized-water reactors:

1) All components, including connecting piping, are made of low-alloy ferritic steels. Component surfaces in contact with the primary coolant water–especially in modern designs–have a weld-deposited cladding of austenitic stainless steel to minimize the ingress of corrosion products in the core region of the reactor pressure vessel.
2) In mixed construction, low-alloy ferritic steels are used for the large components (RPV, steam generator, pressurizer); the surfaces in contact with the media are weld cladded as in item 1. The main coolant piping, main coolant pumps, and any gate valves present in the primary loop are made of solid austenitic stainless steel as structural material.

Table 22. Low-alloy materials

Materials	Specified analysis, %									Remarks,	Standards,	Applications
	C max.	Si	Mn	P max.	S max.	Ni	Cr	Co max.	Mo	miscellaneous	regulations	
SA 508 cl. 2	0.27	0.15 – 0.4	0.5 – 1.0	0.025	0.025	0.5 – 1.0	0.25 – 0.45		0.55 – 0.7	V $\leq$ 0.05	ASME [a]	SG and RPV [d]
22 NiMoCr 37	0.16 – 0.21	0.15 – 0.35	0.5 – 1.0	0.02	0.015	0.5 – 0.9	0.2 – 0.5	0.03	0.45 – 0.7	Cu $\leq$ 0.20 Al 0.01 – 0.04	VdTüV-WB 366 (5.72)	SG and RPV [d] SG and RPV [d]
SA 533 Gr B	0.25	0.15 – 0.4	1.15 – 1.5	0.035	0.04	0.4 – 0.7			0.45 – 0.6		ASME [a]	SG and RPV [d]
16 MND 5	0.2	0.10 – 0.30	1.15 – 1.55	0.015	0.012	0.5 – 0.8	$\leq$ 0.25		0.45 – 0.5	V $\leq$ 0.03, Cu $\leq$ 0.20, Al $\leq$ 0.04	RCC-M [b]	SG and RPV [d]
20 MnMoNi 55	0.17 – 0.23	0.15 – 0.30	1.2 – 1.5	0.012	0.008	0.5 – 0.8	$\leq$ 0.20	0.03	0.4 – 0.55	Cu $\leq$ 0.12, V $\leq$ 0.02	KTA [c] 3201.1	
15 Cr 2 MFA	0.13 – 0.18	0.17 – 0.37	0.3 – 0.6	0.02	0.02	0.4	2.5 – 3.0		0.6 – 0.8	V 0.25 – 0.35, Cu 0.20 $\pm$ 0.05	Soviet regulations	RPV, older VVER 440 series
15 Cr 2 NMFAA	0.13 – 0.18	0.17 – 0.37	0.3 – 0.6	0.01	0.012	1.0 – 1.5	1.8 – 2.3		0.5 – 0.7	V 0.1 – 0.12, Cu $\leq$ 0.10	Soviet regulations	RPV, new VVER 1000 series
15 MnNi 63	0.12 – 0.18	0.15 – 0.35	1.2 – 1.65	0.015	0.005	0.5 – 0.85	$\leq$ 0.15		$\leq$ 0.05	V $\leq$ 0.02, Ti $\leq$ 0.02, N $\leq$ 0.015	VdTüV-WB 427	containment
SA 516 Gr. 70	0.28	0.15 – 0.4	0.85 – 1.2	0.035	0.04						ASME [a]	containment

[a] ASME = Amercian Society of Mechanical Engineers.
[b] RCC-M = Règles de Conception et de Construction, Part "Mechanical components of PWR nuclear islands."
[c] KTA = Kerntechnischer Ausschuß.
[d] SG = Stream generator; RPV = Reaction pressure vessel.

Low-Alloy Ferritic Steels. Forgings, plates, and castings are made mainly of the following steel types: Ni – Mo – Cr (22 NiMoCr 37, SA 508 cl 2) [96] or Mn – Mo – Ni (20 MnMoNi 55, SA 508 cl 3, SA 533 Gr B, 16 MND 5) [97] in the United States and Europe, and Cr – Mo (15 Cr 2 MFA, 15 Cr 2 NMFAA) [98] in Soviet designs (Table 22). The special requirements on weldability (weld cladding and joining), uniformity of chemical composition, and mechanical and technological properties – even for very thick-walled components (e.g., up to ca. 500 mm for the pressure vessel flange, up to ca. 650 mm for the steam generator tube sheet) and extremely heavy forgings (up to 575 t) – have sometimes led to complex and sophisticated qualification measures.

The toughness of the steels is especially important. Notched-bar impact tests, drop-weight tests, and in some cases fracture-mechanics tests are performed to ensure that the materials are tough enough for all service conditions.

For the cylindrical parts of the pressure vessel wall that are located in the neutron flux, the stability of the material against neutron irradiation demands special attention. The radiation sensitivity–in particular, the loss of toughness as a function of neutron irradiation–is substantially enhanced by the alloying elements copper and phosphorus. The first systematic investigation of the effect of copper and phosphorus was published in the United States 1969 [99]. Currently applicable regulations restrict these elements, for example, copper ≤ 0.10 % and phosphorus ≤ 0.012 % in the radiation-affected region of the pressure vessel [100].

Austenitic Stainless Steels. For service at ca. 320 °C and under water chemistry conditions of light-water reactors, the main candidate steels are austenitic 18 % Cr and 10 % Ni types, such as X 6 CrNiNb 18 10 (1.4550), AISI 304, AISI 316, and AISI 347 (see Table 23). To guard against intergranular stress-corrosion cracking in the PWR and BWR environments, those parts or surfaces in the primary loops and auxiliary systems that are in contact with the media are made of either austenitic stainless steels stabilized with niobium or titanium, or extra-low-carbon (ELC) steel (ELC qualities with C ≤ 0.002 %), for which the chemical composition is strictly controlled by special quality specifications (e.g., AISI 316 NG) [101], [102].

The reactor *pressure vessel internals* are the largest individual components in light-water reactors, which are fabricated entirely of highly corrosion-resistant austenitic stainless steel. The reasons for this choice include not only the high corrosion resistance with respect to the coolant and the associated low metal loss rate (< 5 µm/a), but also the low sensitivity of the material properties of this steel for changes related to neutron irradiation.

Both the pressure vessel internals exposed directly to the neutron flux and all other components in contact with the primary coolant (whose corrosion can release corrosion products into the radiation field) call for a steel with a low content of alloying elements and impurities that form long-lived radioisotopes.

Nickel-Rich Alloys The main use of these alloys is in *steam generator tubes* (see Tables 23 und 24), which function as a barrier between the radioactive primary coolant and the activity-free secondary coolant and must therefore meet the most stringent requirement with respect to leak tightness. The principal criterion for tubing material selection is, therefore, corrosion resistance with respect to the high-temperature primary and secondary coolant water. Chemistry guidelines for replacment steam generators of the Siemens – KWU type are listed in Table 25.

The choice of materials for the tube bundles is currently limited to austenitic and nickel-based alloys based on the Fe – Cr – Ni system, which offer the optimum combination of mechanical, thermal, and chemical properties. The materials used for the steam generator tubing in the first PWR plants, up to 1967, were austenitic stainless steels (AISI 304, 316, or 347) and similar grades are still employed in the Russian VVER horizontal steam generators. The corrosion behavior of these generators was generally satisfactory, and in view of the problems subsequently encountered with one of the replacement materials, some people regretted that these grades were not conserved.

Hard Facing Materials. Many years of operating experience have shown that the activation products of the loop materials determine buildup of the radiation dose in the var-

Table 23. High-alloy materials

Materials	Specified analysis, %											Remarks,	Standards,	Applications
	C max.	Si	Mn	P max.	S max.	Ni	Cr	Co max.	Mo	Ti	Nb	miscellaneous	regulations	
Incoloy 800	0.03	0.3 – 0.7	0.4 – 1.0	0.020	0.015	32.0 – 35.0	20.0 – 23.0	0.10		$\leq$ 0.60		balance Fe, Cu $\leq$ 0.75, A1 0.15 – 0.45	KTA ∗∗ 3201.1	steam generator tubes
Inconel 600	0.15	$\leq$ 0.5	$\leq$ 1.0		0.015	$\geq$ 72	14 – 17					Fe 6 – 10, Cu $\leq$ 0.5	ASME ∗ SB 163	
Inconel 690	0.15	$\leq$ 0.5	$\leq$ 1.0		0.015	$\geq$ 58	27 – 31					Fe 7 – 11, Cu $\leq$ 0.5	ASME ∗ SB 163	
Inconel X 750	0.08	$\leq$ 0.35	$\leq$ 0.35	0.015	0.01	$\geq$ 70	14 – 17	1.00		2.25 – 2.75	0.7 – 1.2	Fe 5 – 9	KTA ∗∗ 3204 Cu $\leq$ 0.5	RPV internals (special parts)
Inconel 718	0.08	$\leq$ 0.35	$\leq$ 0.35	0.015	0.015	50 – 55	17 – 21	1.00	2.8 – 3.3	0.65 – 1.15	4.75 – 5.5	Cu $\leq$ 0.3	KTA ∗∗ 3204	
316 L	0.035	$\leq$ 0.75	$\leq$ 2.0	0.040	0.03	10 – 15	16 – 18		2 – 3			balance Fe	ASME ∗ SA 312	piping, valves pumps, RPV internals (according to KTA ∗∗ 3204)
316 NG	0.02	0.25 – 0.5	0.5 – 2.0	0.025	0.005	10 – 14	16 – 18	0.10	2.6 – 3	$\leq$ 0.02	$\leq$ 0.02	N 0.6 – 0.1, V $\leq$ 0.05, Cu $\leq$ 0.3	US utility specification	
304 L	0.035	$\leq$ 0.75	$\leq$ 2	0.040	0.03	8 – 13	18 – 20						ASME ∗ SA 312	
X 6CrNi Nb 18 10 (1.4550)	0.04	$\leq$ 1.0	$\leq$ 2.0	0.035	0.020	9.0 – 12.0	17.0 –19.0	0.2			0.65	Nb/C $\geq$ 8 or $\geq$ 13; C $\leq$ 0.04; Nb $\leq$ 0.65	KTA ∗∗ 3201.1	

∗ ASME = Americian Society of Mechanical Engineers.
∗∗ KTA = Kerntechnischer Ausschuß.

Table 24. Nominal composition of tubing material for steam generators

Element	Alloy 600[a)]	Alloy 690[a)]	Alloy 690[e)]	Alloy 800[c)]	Alloy 800[d)]
C	0.010 – 0.050	0.010 – 0.030	$\leq$ 0.050	$<$ 0.1	$\leq$ 0.03
Mn	$\leq$ 1.00	$\leq$ 0.50	$\leq$ 0.50	$\leq$ 1.50	0.4 – 1.0
Si	$\leq$ 0.50	$\leq$ 0.50	$\leq$ 0.50	$\leq$ 1	0.3 – 0.7
P	$\leq$ 0.025	$\leq$ 0.015			$\leq$ 0.02
S	$\leq$ 0.015	$\leq$ 0.010	$\leq$ 0.015	$\leq$ 0.015	$\leq$ 0.015
Ni	$\geq$ 72	$\geq$ 58	$\geq$ 58	30.0 – 35.0	32.0 – 35.0
Cr	14 – 17	28 – 31	27 – 31	19 – 23	20 – 23
Fe	6 – 10	8 – 11	7 – 11	balance	balance
Cu	$\leq$ 0.50	$\leq$ 0.50	$\leq$ 0.50	$\leq$ 0.75	$\leq$ 0.75
Co	$\leq$ 0.10	$\leq$ 0.035 average $\leq$ 0.018	$\leq$ 0.10		$\leq$ 0.10
Ti	$\leq$ 0.50	$\leq$ 0.50		0.15 – 0.6	$\leq$ 0.6[e)]
AI	$\leq$ 0.50	$\leq$ 0.50		0.15 – 0.6	0.15 – 0.45
B		$\leq$ 0.003			
Nb		$\leq$ 0.10			
N	6	$\leq$ 0.05			

[a] Framatome reactor in France. [a] ASME code Case 20. [a] ASTM 163. [a] Modified Siemens-KWU. [a] Ti/C $\geq$ 12; Ti/C+N $\geq$ 8.

Table 25. Nominal mechanical properties of steam generator tubing

	Yield strength (20 °C), MPa	Ultimate tensile strength (20 °C), MPa	A (20 °C), %	Yield strength (Hot), MPa
Alloy 690 (Code Case N20)	276 – 448	$>$ 552	$>$ 30	–
Alloy 690 EDF Plants	275 – 375	$>$ 630	$>$ 30	$>$ 215 (350 °C)
Alloy 800 ASME	$>$ 207	$>$ 517	$>$ 30	$>$ 163 (350 °C)
Alloy 800 KWU	335 – 470	580 – 700	$>$ 30	$>$ 295(350 °C)
Alloy 600 EDF Plants	275 – 450	$\geq$ 550	$>$ 30	$>$ 215 (343 °C)

ious systems of the reactor. Because of its long half-life and the high energy of the emitted radiation, ^{60}Co makes the principal contribution. Analytical investigations have shown that cobalt-containing hard facings on fittings and of valves and parts of the pressure vessel internals are of prime importance. In addition to optimization of the primary coolant chemistry, the use of cobalt-free hard-facing materials is thus important. Materials based on C – Cr – Fe and C – Cr – Mn – Si – Fe having wear and workability properties comparable to those of cobalt-containing materials are used as substitutes [106].

3.1.5. Water Chemistry of Light-Water Reactors

3.1.5.1. Water Chemistry of Pressurized-Water Reactors

3.1.5.1.1. Primary Coolant System

The primary coolant serves as a moderator and as the medium for transporting heat from the core to the steam generators. Hence, it must not endanger plant operation by the corrosion of materials and consequences thereof. The task of water chemistry can be divided into the following main points:

1) Metal release rates of the structural materials should be minimal
2) The occurrence of localized forms of corrosion should be counteracted

3) The transport and deposition of corrosion products must be influenced in such a manner that contamination of the primary coolant system is kept low
4) The deposition of corrosion products on heat-transfer surfaces, particularly on fuel assemblies, should be prevented as far as possible
5) The radiolytic formation of oxygen should be suppressed

In certain instances, situations may be encountered where conditions of water chemistry are optimum for achieving one goal but are not at optimum with respect to other goals. Thus, the water chemistry specifications must define parameters to achieve a balance among the five objectives, taking into account that highest priority is assigned to materials and fuel integrity. This demonstrates the close connection between choice of materials and water chemistry. Although the other goals are in the second line of priority, they cannot be ignored. The water chemistry conditions specified to fulfill these requirements apply to the materials (see Section 3.1.4), which are in contact with the primary coolant:

1) Austenitic stainless steels for components and piping of the primary system
2) Zircaloy-4 for cladding of fuel assemblies
3) Incoloy 800, Inconel 690, and Inconel 600 for steam generator tubes
4) High-alloy materials (ferritic stainless steels) of low surface area for internals of the primary system

To fulfill the water chemistry requirements, the primary coolant, which contains boric acid as a neutron absorber, is chemically conditioned by the addition of isotopically pure lithium [^{7}Li] in the form of hydroxide as a solid alkalizing agent and of hydrogen.

In this type of "alkaline" conditioning, the additives lithium [^{7}Li] hydroxide and hydrogen are thermally and radiolytically stable so that no decomposition and conversion products develop. The use of lithium [^{7}Li] hydroxide as a pH control agent is advantageous because ^{7}Li is also formed during power operation from boron in the coolant by the nuclear reaction ^{10}B (n,α) ^{7}Li. If lithium of natural isotopic composition is used, tritium is formed as a result of the nuclear reaction ^{6}Li (n,α) ^{3}H. Thus, isotopically pure lithium [^{7}Li] hydroxide is used to prevent the formation of tritium.

At high temperature, boric acid influences the pH of pure water only slightly because it tends to undergo reversible polycondensation with increasing temperature. For this reason, the pH value at the operational temperatures of the primary coolant is determined virtually exclusively by the pH control agent lithium [^{7}Li] hydroxide. The B: Li ratio can lead to different operational conditions which are described in the following.

In the 1970s there was no concern at all about the B: Li ratio. In the early 1980s several units started to employ the *Coordinated Chemistry* where the lithium concentration is coordinated with that of boron to maintain a pH of 6.9 at operating temperature (see Fig. 47). The upper curve in Figure 47 results when additional 0.2 mg/kg Li are added to the coolant. Such considerations are based on the assumption that the transport of metal oxides (corrosion products) in the primary system depends on the pH value of the coolant at operating temperature and should not be influenced by pH fluctuations.

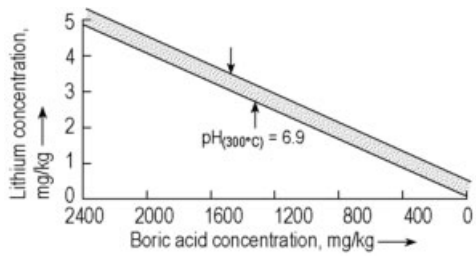

Figure 47. Schematic of the coordinated chemistry regime (at 300 °C)

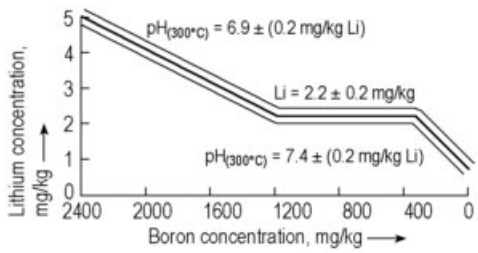

Figure 48. Schematic of modified chemistry regime (at 300 °C)

During the 1980s it became evident that the deposition of metallic impurities in ionic and

colloidal forms and in the form of insoluble oxides (called crud) on the core and the reduction of out-of-core radiation fields may be lower at a pH (T) of 7.4. This is in line with the recognition that nickel ferrite is more important than magnetite for materials transport in the coolant. In the 1990s a modified chemistry regime with ≈ 2.0 or 2.2 mg/kg lithium and pH (T) = 7.2 – 7.4 has been adopted by many plants (Fig. 48).

Materials Integrity Considerations. Water chemistry influences both general corrosion and stress corrosion cracking (SCC) of reactor coolant system (RCS) materials. *General corrosion* is an oxidation process which occurs relatively uniformly over a material surface → Corrosion, Chap. 3.1. Some of the corrosion products released from system surfaces form deposits on the fuel elements. Activation products of species present in such deposits (e.g., ^{58}Co, ^{60}Co) and their release, transport, and deposition on out-of-core surfaces leads to increased plant radiation levels that are of particular concern during outages. General corrosion is no fundamental problem for the operation of PWR systems.

In contrast to this, numerous instances of *stress corrosion cracking* → Corrosion, Chap. 3.3.1. have occured. Primary water stress corrosion cracking (PWSCC) of mill-annealed (MA) Alloy 600 steam generator tubes and plugs, for example, has been found. Apparently coolant chemistry is only a second-order effect in the PWSCC process.

In several plants, PWSCC has occurred after one to ten years of plant operation; whereas in other cases, no leaks could be detected after twenty years of operation.

Fuel Integrity Considerations. Nuclear fuel cladding material must possess a unique combination of properties to function as a barrier for radioactivity between the fuel and the reactor coolant. The cladding material must have a low neutron cross-section, high resistance to corrosion, and adequate mechanical properties (see Section 3.1.4).

Zircaloys develop adherent corrosion films when exposed to pure high-temperature water or steam (see Section 3.1.4).

The important environmental factors affecting the corrosion of specific zirconium alloys are: Temperature, radiation, heat flux, quantities of H_2 and O_2 dissolved in the coolant, and coolant additives.

Examples for the influence of environmental factors are shown in Figure 49 (radiation) and Figure 50 (LiOH as coolant addition).

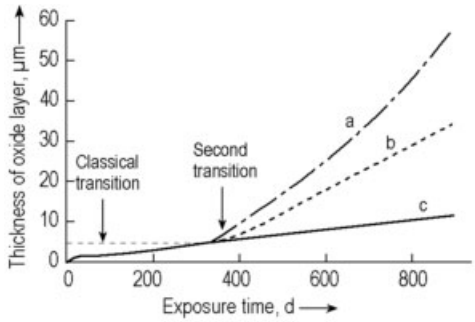

Figure 49. Oxide layer thickness as a function of time
a) PWR fuel rod with constant heat flux $q'' = 75\,W/cm^2$, T_0 340 °C; b) Isothermal inside the PWR; c) Isothermal out-of-file

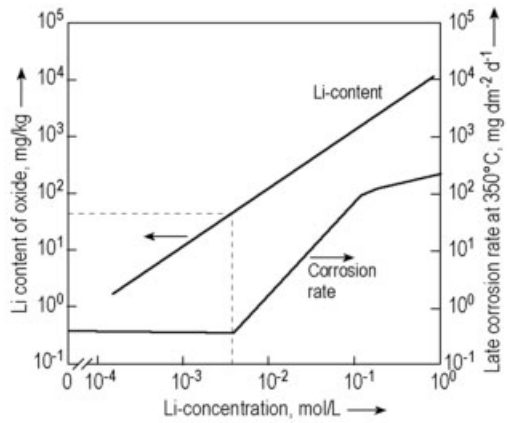

Figure 50. Effect of LiOH on corrosion of Zircaloy at 350 °C and the lithium content in the grown oxide layer

Another factor may be the corrosion products of non-stoichiometric nickel ferrites, $Ni_xFe_{3-x}O_4$ ($0.4 < x < 0.8$), deposited on fuel surfaces which can lead to increased temperatures of the cladding and increased corrosion rates.

Typical corrosion experience data for zirconium cladding materials from different classes of PWRs are shown in Figure 51. The oxide layer thickness increases with increasing burnup of the fuel rods. Since the discharge burnup has been increased from 10 – 20 MW · d/kg to

currently 50 MW · d/kg, fuel integrity was not a concern in the early days but becomes life limiting in the modern operational conditions. The allowable oxide thickness depends on several factors and is for a uniform oxide layer in the range of 100 μm which is the upper limit value.

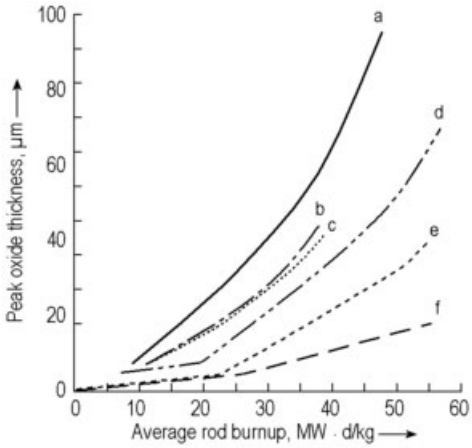

Figure 51. Oxide layer thickness plotted against burnup for different PWRs
a) Siemens-KWU KKGg; b) Fragema 1300 MW; c) Siemens-KWU PWR; d) Westinghouse Millstone and North Anna; e) Westinghouse Zion; f) Westinghouse Zorita

The key chemistry parameters requiring control to minimize problems with fuel integrity are discussed below.

Minimization of *dissolved oxygen* concentrations will minimize corrosion of Zircaloy-4. During power operation, the concentrations of oxygen and oxidizing species are controlled by the addition of hydrogen to the reactor coolant and by minimization of oxygen in the makeup water.

Aluminium, calcium, magnesium, and *silica* may be observed intermittently in the reactor coolant as a result of contaminants in the makeup water, the refueling water, or the boric acid. Many oxides and silicates of *aluminum*, *calcium*, and *magnesium* have negative temperature coefficients of solubility, i.e., they will deposit preferentially at the hottest locations in the RCS, namely, on the fuel rods. Likewise, portions of these deposits may be dissolved during reactor cooldown and thus not be detected or only detected in reduced quantity in fuel surveillance programs. Precipitates can densify the porous core deposit and aggravate corrosion of Zircaloy-4, particularly in plants where some boiling occurs. There are no data that suggest that *silica*, in concentrations $< 1000\,\mu g/kg$ and in the absence of the cationic species, will form these deposits, since the silica solubility is > 200 mg/kg under RCS conditions. However, calculations suggest that higher concentrations of silica will precipitate low concentrations of zeolite-forming cations. If aluminum, calcium, or magnesium are present, precipitation/deposition will occur very rapidly and high concentrations of the cations will not likely be detected in the coolant. Therefore, periodic monitoring of the concentrations of these species in reactor makeup streams should be considered.

Silica may be found in high concentrations in the spent fuel and refueling pool waters. During refueling operations, silica contaminates the refueling water via fuel movement through the transfer canal. Removal of silica can be expensive.

In PWRs with Alloy 600-tubed steam generators the concentration of suspended solids in the coolant can be in the range of 50 to 100 μg/kg and they may form deposits on the cladding material. These deposits can lead to increased temperatures of the fuel cladding and thereby to increased corrosion rates. In plants with Alloy 800-tubed steam generators the corrosion product concentration is roughly one order of magnitude lower than in Alloy 600 tubed plants.

Operation below an average pH (T) of 6.9 is not desirable because of the expected increase in the concentration of suspended solids in the RCS, which will lead to heavier crud deposits. At present, almost all plants operate at an average pH (T) > 6.9.

With the currently prevailing rod discharge burnup of 50 – 60 GW · d/MtU, the end-of-life oxide thickness on Zircaloy-4 cladding can reach or exceed 90 – 120 μm in many PWRs, which is the range considered acceptable by most fuel vendors. It is generally agreed that the probability of *lithium* enhancing Zircaloy oxidation increases with oxide thickness, and hence burnup.

The modified chemistry regime is judged to be acceptable for plant operation despite the fact that existing fuel rod oxide data and laboratory test results indicate that switching primary chemistry from the coordinated chemistry to the modified chemistry regime may result in a small increase in cladding oxide thickness at

high burnups. Adverse effects of crud formation will be minimized under these conditions. Likewise, operation at a minimum pH (T) of 6.9 is recommended, even during longer fuel cycles where boron concentrations are high, to minimize crud effects on fuel corrosion.

Radiation Field Control. Shutdown radiation fields can be affected by numerous factors including operational chemistry, plant configuration, materials of construction, and time of operation.

The contribution of the long-lived radionuclides ^{60}Co and ^{58}Co to shutdown radiation can be minimized by reducing the deposition of corrosion products on the surface of fuel assemblies and minimizing the transport of corrosion products between regions of higher and lower temperature. The accumulation of activated corrosion products on out-of-core surfaces is the main cause for buildup of radiation fields that persist after shutdown. Accumulation proceeds via three essential steps. First, corrosion products are released to the coolant from out-of-core surfaces. Second, some of this material is deposited in the core region, where it is activated by the neutron flux. Third, some of the activated material is released from the core and redeposited out-of-core. Primary coolant chemistry influences each of these steps, and its variation during the fuel cycle has an important effect on the accumulated amount of out-of-core radioactivity.

Corrosion rates of out-of-core materials, corrosion product release rates from these materials, solubility of corrosion products, the properties of corrosion product particles suspended in the coolant, and the thermodynamics and kinetics of in-core deposition, release and redeposition out-of-core are partially related to pH. Corrosion products are carried by the primary coolant in two forms: as solids (i.e., metal oxide crud particles) and as dissolved species such as metal ions and their complexes with hydroxyl ions.

The overal process of activity accumulation on out-of-core surfaces involves a complex set of chemical and physical mechanisms. Nevertheless, operation at coolant pH in the range from 7 to 7.4 at 300 °C can be expected to minimize radiation field buildup, with best results at the high end of the range. The most important parameters for activity accumulation are discussed below.

Corrosion Product Release. Corrosion products containing iron, nickel, and cobalt enter the primary system mainly as a result of general corrosion of out-of-core surfaces. This form of corrosion proceeds so slowly in the primary coolant, that it has no measurable effect on the integrity or performance of components, but it is of concern because of its consequences for activation and radioactivity redistribution.

Particulate Transport. Deposits on fuel and corrosion product particles in the primary coolant are visible evidence of corrosion product migration.

Any attempt to understand the behavior of colloids and larger particles must consider the effects of electrical charges that metal oxides acquire when immersed in aqueous solutions. Electrical charges affect the tendency of particles to deposit on another solid surface or to agglomerate with each other. The net charge on metal oxide surfaces is influenced strongly by pH.

Solution Transport. Based on differences in solubility, dissolved corrosion products (e.g., nickel ferrites) can be transported, dissolved, or deposited by the coolant. Solubility of nickel ferrites depends on temperature, pH, and redox potential, all of which vary with position around the nonisothermal loop. The pH at each location may change with time because boron is removed during the fuel cycle, and redox potential may sometimes be altered if there is a change in the concentration of dissolved hydrogen. These differences cause deposits to grow when concentrations of dissolved corrosion products are greater than the solubility limits and to decrease in amount when concentrations are lower.

Plant Experience. Radioactivity accumulation on out-of-core piping and components is mainly affected by corrosion product release, particulate transport, and solution transport. Effects of all three factors should be mitigated by a coolant chemistry with a high-temperature pH between 7 and 7.4. The mitigation should be most effective at the higher end of this pH range.

The effectiveness of the modified coolant chemistry versus the coordinated B/Li chemistry can be shown by comparing the dose rate development of plant I and plant K (see Fig. 52), that differ mainly in the primary coolant chem-

istry. In plant I, with the higher dose rate, coordinated coolant chemistry was preferred with pH (300 °C) = 6.9, whereas in plant K modified coolant chemistry was applied from the very beginning, in this case pH (300 °C) = 7.4. As a consequence of these measures taken, the radiation fields and the average occupational radiation exposures decreased in the Siemens designed PWRs. When a combination of modified coolant chemistry plus substitution of cobalt-based alloys (either partly or fully) is used, an average occupational radiation exposure of less than 50 rem per plant could be achieved.

Chemistry Guidelines. The chemistry guidelines represent the best compromise between the various parameters as discussed in the previous sections. Primary chemistry guideline parameters are divided into two categories:

1) *Control parameters* require strict control due to materials integrity considerations
2) *Diagnostic parameters* assist in interpreting primary coolant chemistry variations, or may affect radiation field buildup, corrosion performance of system material, or fuel integrity.

In addition, action levels are given. They are chemistry limits which are appropriate for protecting system materials, ensuring fuel performance, and controlling radiation buildup. Three action levels have been defined for remedial actions to be taken when parameters are confirmed to be outside the control values. The remedial actions are defined in the individual specifications of the manufacturers. The reactor coolant specifications of Siemens are given as a typical example in Table 26.

Summary and Outlook. The previous sections have described the state-of-the-art in reactor coolant technology for PWR primary systems. However, the current trend toward longer fuel cycles and/or higher enriched fuel has put an additional burden on the optimization of RCS chemistry.

There may be three areas in the near future where changes/modifications have to be expected:

1) Use of enriched boric acid
2) Injection of zinc (depleted in zinc-64)
3) Start-up and shutdown procedures.

Enriched Boric Acid. Natural boron contains $\sim 20\,\%$ ^{10}B that acts as a neutron absorber in the coolant. If enriched fuel is used in the system (e.g., for extended cycles of 18 – 24 month), a higher reactivity binding capacity must be provided by increasing the boron-10 concentration. This can be achieved by increasing the natural boron concentration or by using enriched boric acid.

Injection of zinc (depleted in zinc-64) is carried out to counteract the activity buildup and/or mitigate PWSCC.

Qualification programs for zinc injection are currently ongoing both in the United States and in Germany. The German program is only aimed at radiation field control, whereas in the United States the SCC consideration is included.

Start-up and shutdown periods may be important when considering radiation field control. This may be relevant especially for those units operating with Alloy 690 as steam generator tubing material.

3.1.5.1.2. Secondary System

In pressurized-water reactors, separation of the primary and secondary cycles means that the secondary cycle needs not to meet the safety requirements of the primary cycle, which are necessitated by the interaction of γ and neutron radiaton with water and with the substances contained in it. However, water and steam must not endanger plant operation by material corrosion and its consequences. The requirements on water chemistry include the following:

1) The metal release rates of the structural materials should be minimal
2) The occurrence of localized forms of corrosion should be counteracted
3) The deposition of corrosion products on heat-transfer surfaces should be prevented as much as possible
4) The formation of environments promoting corrosion in the steam generator should be counteracted

Thus, a close connection exists between the choice of materials and water chemistry. The water specifications necessary to fulfill the above-mentioned requirements apply to the following

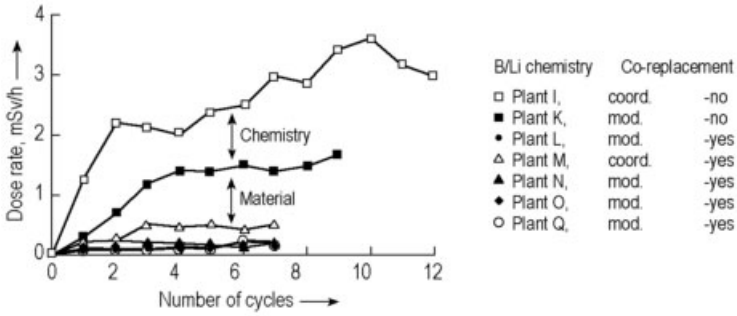

Figure 52. Dose rate development at main coolant piping of seven Siemens PWRs

Table 26. Specification of the reactor coolant during steady state operation

	NV[a]	Al[b] 1	Al[b] 2	AL[b] 3
Control parameters				
Chloride, µg/kg	< 100	> 100	> 200	> 1000
Lithium, mg/kg	corresponding to the plant specific program[c]			
	≤ 2	$\Delta k^d > 0.1$	$\Delta k^d > 0.2$	< 0.2
		> 2.1	$> 2.2^e$	< 5
Hydrogen, mg/kg	2–4	< 2	< 1	< 0.5
		$> 4^e$		> 5
Diagnostic parameters				
Conductivity (25 °C), µS/cm	depending on Li, B			
pH (25 °C)	depending on Li, B			
Boron, mg/kg	parameter determined by reactor physics			
Total sulfur, (µg SO_4/kg)	$< 50^f$			
Fluoride, µg/kg	$< 50^f$			

[a] NV = normal value. [b] AL = action level. [c] It is recommended to use the "modified B/Li-Chemistry". [d] Deviation below the recommended Li/B concentration value. [e] Time duration for deviations: 1 d [f] Extended surveillance program is recommended in case of accelerated fuel cladding corrosion.

materials, which are in contact with water and steam:

1) Incoloy 800, Inconel 690, and Inconel 600 for steam generator tubes
2) Austenitic stainless steels or titanium for condenser pipes
3) Copper alloys for condenser tubes, preheater and moisture-separator-reheater tubing
4) Erosion- and corrosion-resistant materials for parts of the system that transport steam and in which condensation processes occur
5) Unalloyed or low-alloy steels for components and piping of the secondary cycle

The *all-volatile treatment (AVT) method of operation* is used in the secondary cycle (water-steam cycle) to fulfill the demands made on water chemistry. This method of operation is characterized by the fact that chemical conditioning is performed exclusively with the volatile alkalizing agent hydrazine and its thermal decomposition product ammonia. The AVT method can only be used if the secondary system is copper-free.

If the secondary system contains copper alloys amines like morpholine, ethanolamine, and 5-aminopentanol are very often used instead of ammonia. The plant specific selection of a pH control agent depends on:

1) Materials in the secondary cycle system, especially presence of copper alloys
2) Design and use of a condensate-demineralizing system
3) Restrictions regarding the effluents
4) Specific vendor recommendations

An important prerequisite for use of the AVT method of operation is the absence of salts in the secondary cycle to the greatest possible extent, for the following reasons:

Corrosive conditions can arise at the steam generator tubes if salts ingress with the cooling water in the event of condenser leaks. In addition, stagnant steam bubbles can form in certain sectors of the steam generator, e.g., in the deposits of corrosion products on the tube sheet. The ingressed salts are concentrated to saturation in the transition region between the stagnant steam bubbles and the circulating water in the steam generator. Moreover, these salts are hydrolyzed in the steam generator water to produce acid or alkali, depending on the type of salt. These concentration and hydrolysis processes can lead to pH excursions (i.e., strong deviations from the desirable slightly alkaline pH value) and to the formation of aggressive media.

To minimize the entry of salt into the steam generator,

1) Corrosion-resistant materials may be used for condenser tubes;
2) Sensitive methods must be applied to monitor the condenser and locate leaks; and
3) The makeup water for the secondary cycle and the conditioning agent hydrazine must have high purity.

Effective reduction of the ingress of corrosion products into the steam generators can be achieved with a high alkalinity (high pH value) in the secondary cycle. In this way, metal release rates of the structural materials and erosion/corrosion in parts of the system that transport steam can be reduced to a minimum. A prerequisite for the use of the AVT method of operation is the absence of copper alloys in the entire secondary system.

Moreover, reducing water conditions are required to prevent localized corrosion at the tubes of the steam generator. These conditions are guaranteed by conditioning with hydrazine. First, hydrazine binds the residual oxygen:

$$N_2H_4 + O_2 \longrightarrow N_2 + 2\,H_2O$$

Second, the ammonia required to maintain a high pH is formed by thermal decomposition of excess hydrazine:

$$2\,N_2H_4 \longrightarrow 2\,NH_3 + N_2 + H_2$$

Steam Generator Operating Experience (General). Worldwide experience shows that more than 80 % (90 % of those with more than five years operation) of some 250 operating PWRs, including CANDU reactors, have had corrosion or mechanical degradation problems on their steam generators. Such problems often entrain unscheduled or extended outages for preventive and corrective maintenance, which are costly in terms of inspection, repair work, lost power, and personnel radiation protection. Furthermore, the steam generators have been replaced in about 30 PWRs and replacement is planned or underway in more than 20 others.

The size and design of the different types of the predominantly used recirculating steam generators vary widely, and have been changed in the course of time. Nevertheless, the operating conditions remain fairly similar in all cases, leading to virtually identical mechanical loading and chemical environments. The principal characteristics of the steam generators in recent reactors built by Siemens and Framatome are given in Table 27.

The choice of Incoloy 600 was dictated by the following considerations:

1) The 18-10 stainless steels are susceptible to stress corrosion cracking in chloride-containing media. The risk of SCC is increasing, due to the steady increase in size and operating temperature of the generators. Thus, alloys with a high nickel content which are more resistant to this type of attack were considered to be preferable.
2) The thermal expansion coefficient of Alloy 600 is much closer to that of the ferritic steels (used for the shells and tube plates) than that of stainless steel.
3) Satisfactory experience with Alloy 600 in naval reactors suggested that it could also be used in civilian power plants, although size, operating temperature, and thermohydraulic conditions differ significantly.
4) The lower nickel grade Alloy 800 has a lower mechanical strength in the annealed condition. Moreover, the nickel content of Alloy 800 was perceived to be too low to guarantee its corrosion resistance in chloride-containing environments.

Table 27. Operating conditions of recent steam generators

Steam generator	No. of steam generators	Primary temperature	Secondary temperature (pressure)	No. of tubes per steam generator, dimensions
Framatome N4, 1450 MW_e	4	329 °C/292 °C	289 °C/(73 bar)	5599/19.05 × 1.09 mm
Siemens Convoy, 1314 MW_e	4	326 °C/291 °C	280 °C/(64.5 bar)	4118/22 × 1.2 mm

From the late 1960s onwards, many cases of cracking of Alloy 600 occured in the primary medium in regions where the tubes were subjected to the highest mechanical stresses (bends and tube-plate assembly points). In France nowadays, this phenomenon is the major cause of tube plugging and is the principal reason for the replacement of steam generators.

Extensive research work has led to a better understanding of the phenomena involved. It was found that the SCC behavior is extremely sensitive to both the metallurgical structure of the tubes and the testing temperature. The behavior of Alloy 600 tubing could thus be improved significantly in new generators from the early 1980s. An additional vacuum heat treatment at 700 °C at the end of tube manufacture led to a structure with a much higher SCC resistance and reduced the level of residual stresses.

In parallel to this work, an extensive research and development effort was undertaken in a search for new materials. In the early 1980s a recommendation was made to plant constructors to abandon Alloy 600 in new equipment.

Choice of Alloy 800. Reactors with generators equipped with Alloy 800 tubing were developed by Siemens as early as 1972.

The composition of the material employed was modified with respect to the ASTM specifications, in order to improve the corrosion resistance of the tubing and guarantee its long term structural stability.

The alloy is annealed between 960 and 1000 °C. In order to meet the design strength, about 4 % strain is introduced by cold drawing after annealing.

This alloy is also used in the Canadian CANDU power stations commissioned after 1983. In addition, five power stations other than those of the Siemens type have chosen Alloy 800 for replacement steam generators.

Choice of Alloy 690. In virtually all the other steam generators commissioned since 1989 (currently 84 units), the tubing material is Alloy 690, recommended in the early 1980s and adopted in France in 1984. The tubes are given a final heat treatment of a few hours at 715 °C.

Conclusion from Operating Experience. Marked improvements have been made to the steam generators of PWR power plants, and have apparently led to satisfactory solutions. These ameliorations include the overall generator design, the choice of materials for the condenser and reheater sections of the secondary circuit, the water chemistry, the use of either Alloy 800 or Alloy 690 for the tubing, and the tube manufacturing process.

Modifications of the *feedwater circuit* involved the replacement of copper alloys in the condensers and reheaters. If copper is present in the water, only a limited amount of ammonia and hydrazine may be added. This makes it impossible to chemically treat the medium to attain high pH levels that correspond to the specifications for water in the secondary cycle. The use of stainless steels for the exchangers and titanium for the condensers in coastal plants (i.e., seawater-cooled plants) is generally recommended in order to limit the ingress of iron oxides and copper – ammonia complexes into the steam generators.

Currently, the steam generator tubing is best made of Alloy 800, strain hardened by cold drawing (and eventually prestressed by shot peening on the outside surface to improve the SCC resistance in the secondary medium), or Alloy 690, with a final heat treatment at 715 °C to relax residual manufacturing stresses and improve the microstructure.

Since the initial observations of SCC in the *secondary circuit* had been attributed to the presence of sodium hydroxide, extensive testing was carried out in aqueous sodium hydroxide solu-

tions. The results reveal a roughly equivalent behavior for alloys 800 and 690 (Fig. 53), both of which appear to be markedly better than Alloy 600.

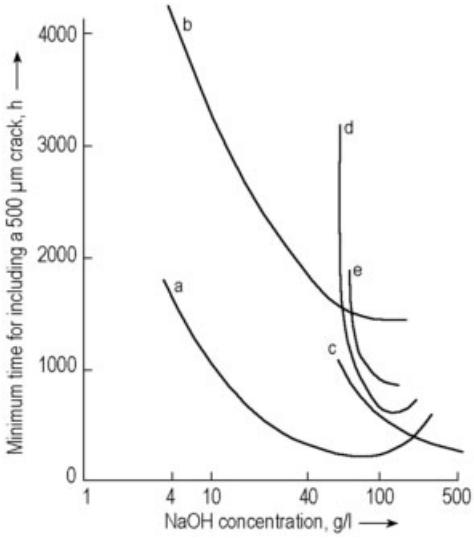

Figure 53. Behavior of C-rings [stressed to yield-strength (ASTM STP 425)] in deaerated caustic soda solution at 350 °C
a) Alloy 600 MA; b) Alloy 600 HT 16 h 700 °C; c) Stainless steel 316; d) Alloy 800; e) Alloy 690

The two alloys have different thermal conductivities and thermal expansion coefficients, which are taken into account in plant design and construction.

The stress relieving treatments performed on the welds in the shell have been adapted to account for thermal expansion of the different materials, particularly in the tube-plate.

Water Chemistry Specifications. The complexity of the operational conditions of steam-water cycles and the close connection between design, material, and water chemistry has been described. Instead of discussing all aspects for setting specifications of chemistry, the values defined by Siemens and Framatome for replacement steam generators are given in Tables 28 and 29.

3.1.5.2. Water Chemistry of Boiling Water Reactors

The coolant in the water-steam-cycle of Boiling-Water-Reactors must meet essentially the same requirements as the primary coolant of PWRs. The water chemistry program has to satisfy the following goals in particular:

1) Avoid crack initiation of SCC and/or minimize crack growth of intergranular stress corrosion cracking (IGSCC) and irradiation-assisted stress corrosion cracking (IASCC)
2) Minimize fuel cladding corrosion
3) Minimize radiation field buildup
4) Reduce or minimize general and selective types of corrosion in the BOP systems and components

The concept of BWRs involves radiolytic decomposition of water in the radiation field of the core by forming a steady-state concentration of hydrogen, oxygen, and hydrogen peroxide. The noncondensable gases are, for the most part, transported with the steam out of the reactor pressure vessel. They are largely removed from the cycle through the condenser exhaust and fed to the waste gas system. The interaction between these processes results in widely varying gas concentrations at different points in the water-steam cycle and, consequently, very different oxygen concentrations.

The reactor water contains an equilibrium O_2 concentration of 200 – 300 µL/m^3 and all the hydrogen peroxide formed. This results in a strongly oxidizing environment, so that even water with a very low conductivity can cause SCC in susceptible materials.

In areas of high neutron flux where the environment is most oxidizing, even components constructed with SCC-resistant materials may undergo IASCC. Radiation can change the microchemistry of the materials (radiation-induced segregation), the water chemistry (radiolysis), and the material properties, e.g., stress. Concerns about IASCC are currently increasing, especially IASCC of components that are not readily replaceable. Options to mitigate both IGSCC and IASCC are:

1) Repair and replacement of core internals and of affected piping systems (very costly and time-consuming).
2) Hydrogen addition to the feedwater. This option protects only certain parts of the reactor pressure vessel and the internals. The addition of metal catalysts can improve the situation. This option is currently under development.

Table 28. Chemistry guidelines for replacement steam generators of the Siemens – KWU type (blowdown sample)

	Normal frequency	Frequency	Expected value	Specified value	Action level 1	2	3
Control parameters							
Cation conductivity, µS/cm	continuous		< 0.5	$\leq 1.0^a$	$> 1^a$	$> 2^a$	$> 7^a$
Sodium, µg/kg	semicontinuous		< 10	< 50	> 50	> 100	> 500
Diagnostic parameters							
pH (25 °C)			> 9.0				
SO_4^{2-}		^{b}	^{c}				
Cl^-		^{b}	^{c}				
F^-		^{b}	^{c}				

^{a} Due to total strong anionic species. The theoretical cation conductivity should be calculated and compared to the measured cation conductivity. Any inconsistencies should be investigated. ^{b} To be measured in case of overriding expected values. ^{c} Strong anions to be measured and their concentrations used to check the cationic conductivity specification.

Table 29. Chemistry guidelines for replacement steam generators of the Siemens – KWU type (feedwater sample)

	Normal freqency	Normal value	Expected value	Specified value	Action Level 1	2	3
Control parameters							
pH (25 °C)	3 per week		< 0.5	$\geq 9.3^a$	< 9.3		
Cation conductivity, µS/cm	continuous		< 0.12	$\leq 0.2^b$	> 0.2		
Oxygen, µL/m^3 (ppb)	continuous		< 1	< 5	> 5	> 20	> 100
Diagnostic Prameters							
Hydrazine, mL/m^3 (ppm)		> 0.02					
Cation conductivity, µS/cm		> 5.5					

^{a} as high as possible, compatible with secondary cycle. ^{b} Due to anions of strong acids.

3) Alternatives to hydrogen water chemistry (HWC).

Zircaloy fuel cladding shows an acceptable corrosion behavior under BWR water conditions. However, in the presence of certain impurities, localized effects may occur. In addition, corrosion product deposits can precipitate on the fuel cladding, which may accelerate general corrosion.

Activated corrosion products are responsible for the radiation fields in BWRs. Major parameters influencing buildup of activity are iron and nickel concentrations in the feedwater, higher water conductivity, and concentration of ^{60}Co, ^{58}Co, and Zinc in the reactor water. Corrosion phenomena in the water-steam cycle are not very frequent.

3.1.5.2.1. Stress Corrosion Cracking

The BWR environment is characterized by very pure water (low conductivity) but strongly oxidizing conditions. Actual operating values for conductivity approach the theoretical value of 0.055 µS/cm at 25 °C. An increase in conductivity may endanger all goals of BWR water chemistry. However, conductivity and oxidizing conditions are only two parameters out of several that may influence crack initiation and/or crack growth. These include:

1) Material conditions
2) Mechanical loading
3) Radiation
4) Properties of the protective layers
   - Presence of zinc
   - pH value
5) Environment
   - Conductivity (SO_4^{2-}, Cl^-, CrO_4^{2-}, NO_3^-, crack electrolyte, see below)
   - Corrosion potential (C_{ox}, C_{red}, pH value)

Conductivity. Crack growth rates can increase dramatically when increasing the conductivity (see Fig. 54). Consequently, the *countermeasure* is to keep conductivity in reactor water below 0.1 µS/cm. To this end, maybe improved

cleaning methods must be developed. In addition, it is important to check which kind of impurity causes the increased conductivity. Sulfate contamination as a result of cooling water inleakage, regenerant chemical inleakage, resin ingress or resin decomposition, is most detrimental with respect to environmentally-assisted cracking of BWR materials. Thus, sulfate concentrations should be $< 5\,\mu g/kg$. The second detrimental species is *chloride*. Besides IGSCC of sentitized stainless steels it may also promote transgranular stress corrosion cracking (TGSCC), pitting, and crevice cracking. Especially the combination of chlorides and oxygen is very dangerous for reactor materials. Since chloride accelerates all forms of corrosion its level should be kept below 5 µg/kg during power operation.

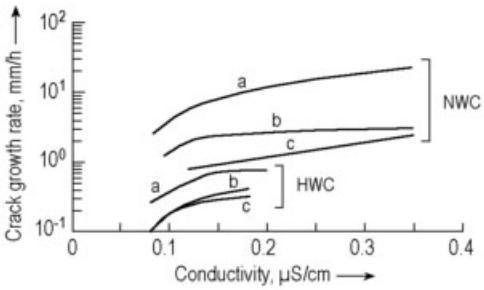

Figure 54. Effect of conductivity on IGSCC in-reactor
a) Alloy 182; b) Ferritic steel type 304; c) Alloy 600
NWC = Nitrogen water chemistry; HWC = Hydrogen water chemistry

The next impurity of concern is *chromate* because it can in some cases contribute to more than 50 % to the conductivity at room temperature. In addition, large chromate spikes have been experienced, when switching from hydrogen water chemistry (HWC) to normal water chemistry (NWC).

Finally, the concentrations of *silicia* in the reactor water are 10 to 100 times higher than those of other impurities. However, the impact of silica on IGSCC is considered to be very small.

Other countermeasures against SCC may be understood by using the SCC model of MacDonald shown in Figure 55. According to Figure 55 the crack propagation occurs by an anodic dissolution of the metal at the crack tip. The electrons migrate to the metal surface where they can reduce oxygen or hydrogen peroxide to water. All measures to slow down this mechanism of metal dissolution can be used as countermeasure against SCC. Beside the impurity (conductivity) reduction an increase in the reactor water pH value may be helpful. Such an increase in pH can be achieved by an optimization of the cation: anion ratio of the ion exchange resin in the reactor cleaning systems. Another contribution to a pH increase is the injection of zinc into the BWR coolant. If zinc is integrated into the surface layers, the quality of these layers will be improved and the reduction of O_2 and H_2O_2 thereby slowed down (see Fig. 55).

At present, the best available control parameter for mitigation of SCC in BWRs is an electrochemical corrosion potential (ECP) control. The ECP in normal BWR water chemistry is determined by the production of O_2, H_2O_2, and H_2 and a resulting ECP value between 100 to 250 mV standard hydrogen electrode. The corrosion potential in operating BWRs can be reduced by injecting hydrogen into the reactor feedwater. This results in lower ECP value, the target value for mitigating SCC being < -230 mV. At this level the crack propagation rate should be low enough that no failures will be produced in a 40 years lifetime.

However, the ECP varies significantly around the BWR circuit. Accordingly, H_2 addition to the feedwater will yield a different ECP at various locations. The values at different location can be calculated by an empirical correlation between ECP value and impurities present in the system. In addition, the variations in O_2, H_2O_2 and H_2 around the systems have to be considered. They can be calculated with existing radiolysis models.

In the future, alternatives to hydrogen injection may arise. These include the use of catalysts inside the reactor system or the generation of more isolating protective layers on the metal surfaces. Another possibility may be the use of less volatile reducing agents such as NH_3 or N_2H_4 instead of hydrogen. However, all such alternatives have to be qualified for the use in BWR systems.

3.1.5.2.2. Radiation Fields

Radiation fields during the steady-state operation of BWRs are determined by the nitrogen-16 activity. The major source for personnel expo-

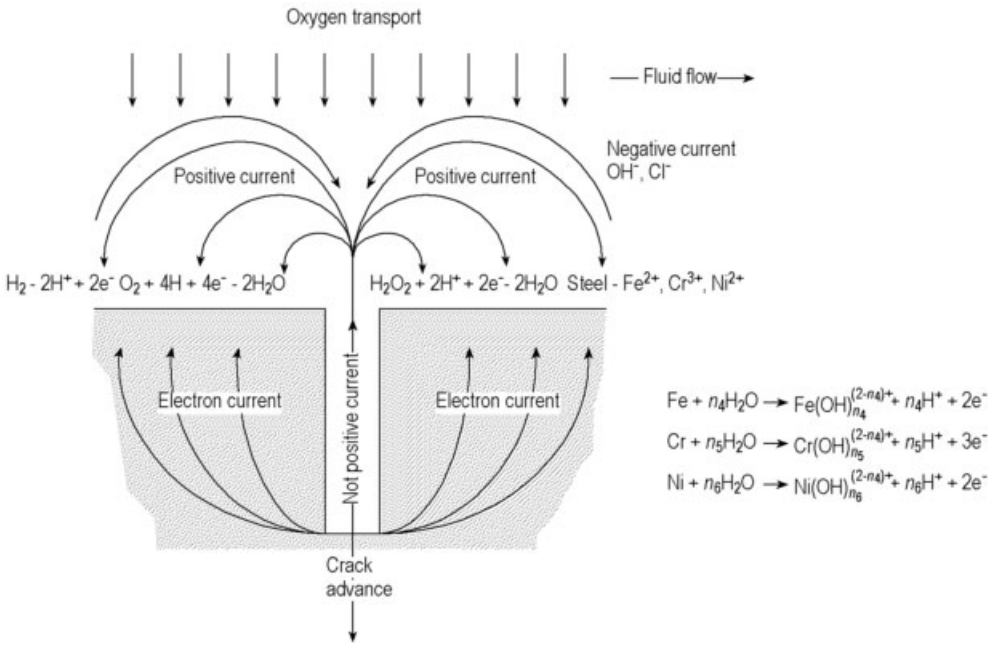

Figure 55. Crack-electrolyte; galvanic coupling effect according to D. D. MacDonald

sure during shutdown is ^{60}Co, which is embedded in the protective layers of the reactor systems.

Nitrogen-16 Radiation Considerations. ^{16}N is produced by the reaction

^{16}O (n, p) ^{16}N

in the core region. Its production cannot be influenced by the chemical environment. However, the behavior of nitrogen-16 in the system, especially its volatility depends on the ECP. Under NWC conditions, the main species of nitrogen are NO_3^- and NO_2^- which are non-volatile. A small fraction, which produces the ^{16}N radiation fields in the main steam area is considered to consist of NO, NO_2, and some NH_3.

Under HWC conditions the nitrogen compounds consist of NH_3 and NH_4^+ which are in equilibrium with one another. Depending on the H_2-concentration the volatility of the nitrogen compounds is increased and thereby the ^{16}N radiation levels. The ^{16}N levels can be increased by a factor of 5 to 6.

Radiation Fields on Out-of-Core Surfaces. The primary source of radiation fields on out-of-core surfaces is ^{60}Co. Most models to describe the transport of activity include the following steps:

1) Crud particles are released to the condensate and feedwater from metal surfaces and subsequently appear in the reactor coolant. In some plants forward-pumped heater drains contribute to the corrosion products and in all plants leakage occurs through the condensate treatment system.
2) Metal ions are also released to the coolant from corroding surfaces in the primary system. Colloidal oxide or hydroxide particles are formed in the coolant when the saturation limit is exceeded.
3) Some interactions between soluble ionic species and crud particles result in adsorption (scavenging) of ionic species on the crud particles suspended in the water.
4) Crud and soluble species deposit on the fuel surface. Some of the loosely attached deposit in the outer layer may be transformed into a tenacious layer on the fuel surface.
5) The deposits on the fuel surface and the in-core structure materials are activated by neutrons. Radioactive species are released from in-core materials by dissolution and wear,

and from fuel surface deposits by erosion and spalling caused by hydraulic shear forces.

6) The radioactive species redistribute into "soluble" and "insoluble" species, depending on the solubililty of the oxide and the concentration of iron crud which acts as a scavenger. Soluble and insoluble radioactive species deposit on out-of-core surfaces by different mechanisms.
7) Formation of double layer surfaces is caused by base metal corrosion and water-borne crud deposition. The oxidation state is changed at the material surfaces (e.g., Fe^{2+} is oxidized to Fe^{3+}) and a small fraction of the corrosion products, both radioactive and nonradioactive, is released mostly from the outer oxidized layer into the coolant.

3.1.5.2.3. Fuel Integrity

Basic consideration regarding fuel integrity are described in Sections 3.1.4 and 3.1.5.1.1. The extent of corrosion and hydriding of Zircaloy in the BWR is determined primarily by the composition of the alloy, the processing history of the Zircaloy cladding, and the presence of water impurities.

In general, the problems with fuel integrity in a BWR environment are smaller than those in the PWR case. This is due to the lower temperature at the oxide metal interface in BWRs.

An anomaly in corrosion behavior, called *nodular corrosion*, can occur in BWRs. This anomaly is characterized by the formation of white pustules, appearing in metallographic cross-sections as lense-shaped oxide deposits with significantly higher thickness than the uniform oxide. Under continued exposure, these pustules grow together to an almost uniform and thick oxide layer. Experience from various BWRs has shown that such corrosion can vary considerably, depending on water chemistry and cladding condition. However, nodular corrosion does not depend on temperature and normally does not limit fuel rod life.

Effects of Specific Coolant Impurities and Additives. The effect of *iron* can be a high crud-loading of the fuel. Operation within the current limits on feedwater iron levels effectively prevents the buildup of iron oxide crud deposits.

A corrosion failure mechanism named crud-induced localized corrosion has been associated with elevated levels of *copper* in BWR feedwater. These failures are characterized by local through-wall corrosion pitting. However, a minimum value for copper to produce this type of failure cannot be given, because the failure depends on the corrosion resistance of the fuel cladding and possible trigger mechanisms.

The addition of *hydrogen* to the feedwater of a BWR effectively mitigates stress corrosion cracking of austenitic materials. In case of moderate H_2 concentrations increased hydrogen pickup need not be worried about. However, hydrogen addition has to be performed with great caution (1) in plants susceptible to crud-induced localized corrosion; (2) in plants operating with cycles > 18 month and a high power density.

Zinc is added to the BWR to adjust a concentration of $5-10\,\mu g/kg$ in the reactor water. When zinc is added, $20-30$ % of the deposited iron oxide in the crud is converted to an iron-zinc spinel. This type of spinel oxide is more adherent to the Zircaloy cladding.

3.1.5.2.4. Chemistry Guidlines

As an example, the present VGB chemistry guidelines for BWRs are given in Table 30.

3.2. Gas-Cooled Graphite-Moderated Reactors

The principal difference between gas-cooled nuclear reactors and other types (e.g., boiling-water and pressurized-water reactors) lies in the coolant and the moderator. The gaseous coolant can be air, carbon dioxide, helium, or another gas. The power densities in the core are low, compared to other reactor types, both because gases exhibit poorer heat-transfer than water or sodium and because graphite is used as moderator. The power density of the THTR-300 (a helium- cooled high-temperature reactor) is $6\,MW_{th}/m^3$, in contrast to the $80\,MW_{th}/m^3$ of water-cooled reactors, which means that the core of a gas-cooled reactor is much larger than that of a water-cooled reactor. This fundamental difference is one of the reasons for the price

Table 30. Present VGB chemistry guidelines for BWRs

	Normal value	Action level 1	2	3
Feedwater				
Conductivity at 25 °C (direct and continuous, µS/cm, measurement at the sampling location)	≤ 0.06	$> 0.07^a$		
Total iron, µg/kg	$\leq 2^b$	> 5		
Total copper, µg/kg	$\leq 0.3^c$	> 0.75		
Oxygen, µg/kg	20 – 200	< 20 and > 200		
Reactor water				
Feedwater Conductivity at 25 °C (direct and continuous, µS/cm, measurement at the sampling location)	≤ 0.15	> 0.25	> 1	> 5
Chloride, µg/kg	≤ 10	> 20	> 100	> 200
Sulfate, µg/kg	≤ 10	> 20	> 100	> 200
Silica, µg/kg	≤ 200			

[a] These are requirements, which may not be fulfilled in single cases. However, they describe the status of the feedwater. [b] Depending on design of the secondary cycle, clean-up procedures and system materials, the parameter "total iron" can have concentrations in the 1 to 5 µg/kg range and this can be tolerated. [c] Depending on the condenser tubing, the parameter "total copper" can have concentrations in 0.2 to 0.5 µg/kg range and this can be tolerated.

advantage of water-cooled reactors over carbon dioxide-cooled reactors. Helium-cooled reactors are competitive with water-cooled reactors.

3.2.1. History

After the discovery of artificial nuclear fission, the first nuclear reactor in the world was built by Fermi in Chicago in 1942. Initially, the reactor power was 0.5 W, later increased to 100 W. Fermi used ca. 46 t of unenriched natural uranium as fuel. About 385 t of graphite served as moderator. The coolant was air.

The first reactor with higher power was also graphite moderated and air cooled. The reactor was commissioned in 1943 in Oak Ridge, Tennessee to generate plutonium for military purposes. The next reactors built in the United States were graphite moderated but water-cooled (Hanford reactors). Thus, in the United States an early changeover to water cooling occurred.

In the United Kingdom gas cooling was retained for plutonium-production because of its greater inherent safety.

3.2.2. Carbon Dioxide-Cooled Reactors

3.2.2.1. Magnox Reactors

The Magnox reactor marked the changeover to peaceful industrial uses of nuclear energy in the United Kingdom and France. The fuel was natural uranium, which meant that only a weak neutron absorber could be used as cladding material (e.g., an aluminum – magnesium alloy). The reactors were named after this Magnox alloy. The choice of coolant was limited to a gas because of (1) the lack of experience with other coolants at the time and (2) the use of natural uranium as fuel, which forced the choice of canning material. The choice of gaseous coolant was further limited by physical, neutron, and chemical properties to either helium or carbon dioxide. Carbon dioxide was chosen because of its more favorable physical properties for this application and because of its low price.

A diagram showing the main components of a gas-cooled reactor (e.g., a Magnox reactor) is given in Figure 56. The design of the various reactors is very different. For example, in French reactors at St. Laurent-des-Eaux, the boilers are below the core, whereas the boilers in British designs are at the side of the core. In general, the French preferred an integral con-

struction whereas the British had a loop design. Figure 57 is a schematic section of the 300-MW_e Magnox station in Oldbury, Great Britain. A section through one of the 500-MW_e Magnox stations in St. Laurent-des-Eaux is shown in Figure 58. Table 31 summarizes the commercial Magnox stations.

Table 31. Magnox reactors

Reactor	Power output, MW_e	Start-up	Shutdown
United Kingdom			
Berkeley	2×166	1962	1989
Bradwell	2×129	1962	
Hunterston A	2×160	1964	
Trawsfyndd	2×235	1965	
Dungeness A	2×212	1965	
Hinkley Point A	2×270	1965	
Sizewell A	2×250	1966	
Oldbury	2×225	1967	
Wylfa	2×495	1971	
France			
Marcoule, G 1 – 3	2×28	1958 – 1960	
Chinon-1	68	1963	
Chinon-2	198	1965	
Chinon-3	476	1966	
St. Laurent A 1	487	1969	1990
St. Laurent A 2	515	1971	
Bugey-1	555	1972	
Italy			
Latina	210	1963	1986
Spain			
Vandellos-1	500	1972	
Japan			
Tokai Mura	166	1966	

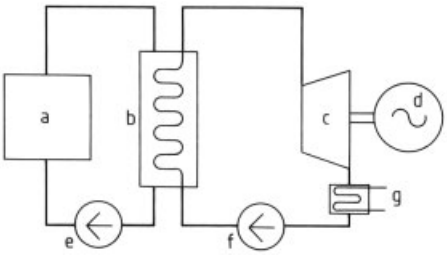

Figure 56. Main elements of a gas-cooled reactor steam-electric plant
a) Reactor; b) Boiler; c) Steam turbine; d) Alternator; e) Gas circulator; f) Boiler feed pump; g) Condenser

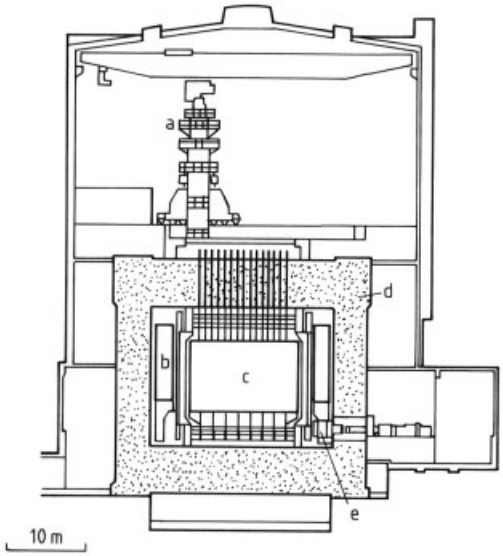

Figure 57. The 300-MW_e Magnox reactor at Oldbury
a) Refueling machine; b) Boiler; c) Core; d) Concrete vessel; e) Gas circulator

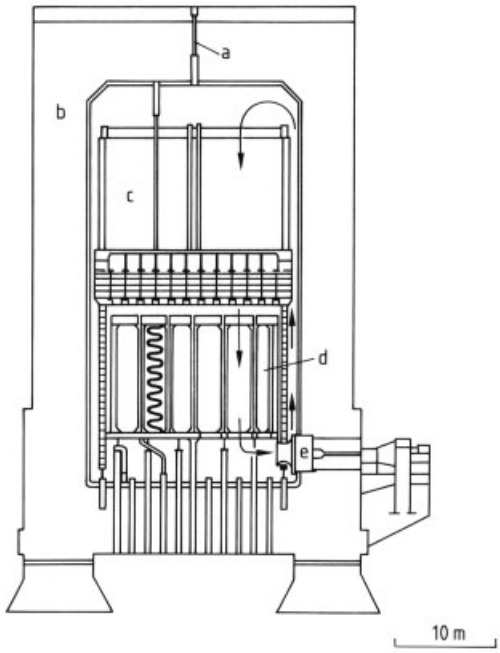

Figure 58. St. Laurent-des-Eaux, France – a 500-MW_e Magnox-type reactor
a) Stand pipes; b) Concrete vessel; c) Core; d) Boilers; e) Gas circulator

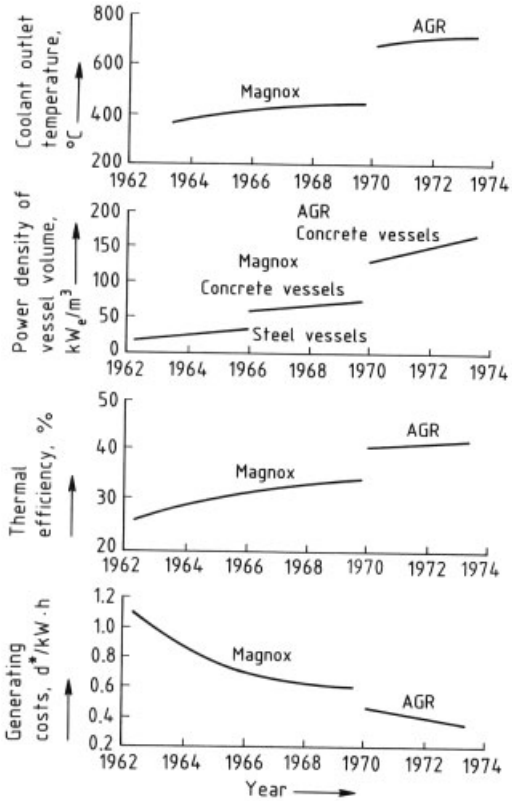

Figure 59. Trend in AGR performance as seen in the mid-1960s
d ∗ = pence

3.2.2.2. Advanced Gas-Cooled Reactors

The advanced gas-cooled reactors were built only in Great Britain. The French converted to light-water reactors after the construction of Bugey-1 (see Table 31) for economic reasons. The AGR is a development of the Magnox reactor. The efficiency was improved and costs were lowered by increasing the power density and raising the coolant outlet temperature, as shown in Figure 59. Because of the higher temperatures than in Magnox stations, the fuel elements were canned in stainless steel. This required the use of lightly enriched uranium as a fuel. Further development of the prestressed concrete reactor vessel (PCRV) allowed an increase in primary circuit pressure to > 4 MPa. Steam temperature and pressure were similar to those in coal- or oil fired-stations, thus resulting in high efficiency. A cross-section through the Hinkley Point B advanced gas-cooled reactor is shown in Figure 60. The *core structural components* are made of graphite. Fuel elements are exchanged during operation.

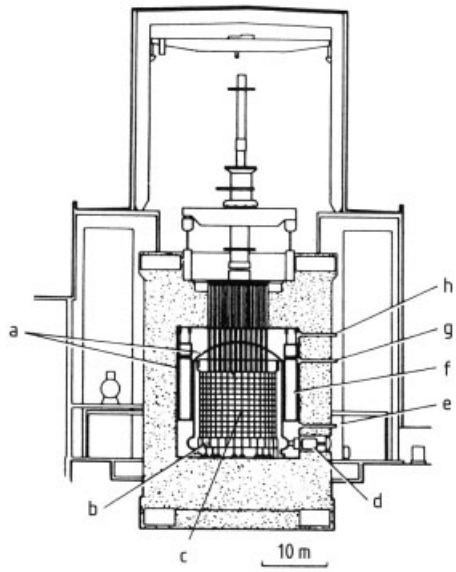

Figure 60. Cross-section through a Hinkley advanced gas-cooled reactor
a) Thermal insulation; b) Supporting grid; c) Core; d) Gas circulator; e) Boiler feed penetration; f) Boiler; g) Main steam penetration; h) Reheat steam penetration

The *boiler materials* are austenitic steels for temperatures $> 350\,°C$ and mild steel for $< 350\,°C$. Table 32 gives details of all the AGRs built.

Table 32. Advanced gas-cooled reactors

Reactor	Power output, MW_e	Start-up
Windscale ∗	34	1963
Hunterston B	2×660	1976 – 1977
Hinkley Point B	2×660	1976 – 1978
Dungeness B-1	508	1983
Dungeness B-2	660	1985
Hartlepool	2×660	1983 – 1985
Heysham I	2×660	1983 – 1985
Heysham II	2×660	1988
Torness	2×701	1988 – 1989

∗ Experimental reactor. Operational experience gained on this reactor was used to design commercial AGRs. Final shutdown occurred in 1981.

3.2.2.3. Advanced Gas-Cooled Reactor Safety

Because the pressure in the secondary (steam) circuits is always higher than in the primary circuit, carryover of radioactivity into the steam can be ruled out. Maintenance work can be carried out on the secondary circuits without radiation dose problems for the staff. The activity of the primary coolant is very low. In the event of leakage, only a small amount of radioactivity will be released. The core radiation is practically shielded by the PCRV. The primary coolant is in the gaseous state under any operating conditions (in contrast to, e.g., water or sodium that can evaporate or solidify, respectively). Thus, in the event of a reactor incident, no discontinuous change in heat-transfer properties can occur. The amount of excess reactivity in the core is always low because of on-line refueling. If the coolant compressors fail, decay heat can be removed by natural convection. Sufficient cooling water supplies for the steam generators must be available in this case. All important reactor components are designed for redundancy. Because of the good safety properties of the AGR, no containment other than the PCRV is necessary. This means that many components and the auxiliary plant can be maintained or repaired easily during operation, if necessary.

3.2.2.4. Primary Circuit Chemistry

The chemistry of the primary circuit is determined by the Boudouard reaction, in which the core structural component material graphite (which is also the neutron moderator) reacts with the coolant carbon dioxide at high temperature as well as under radiation (thermal and radiolytic corrosion):

$$C + CO_2 \rightleftharpoons 2\,CO$$

Because the graphite core structures are not meant to be replaced, corrosion must be kept to a tolerable level.

Thermal Corrosion. The graphite temperature in a *Magnox reactor* is $< 350\,°C$. The graphite temperature in an *AGR* is kept below 500 °C by an extensive system of coolant channels, although the average gas exit temperature is 650 °C. The coolant channels are shown in Figure 61. Thermal graphite corrosion is very low in both cases.

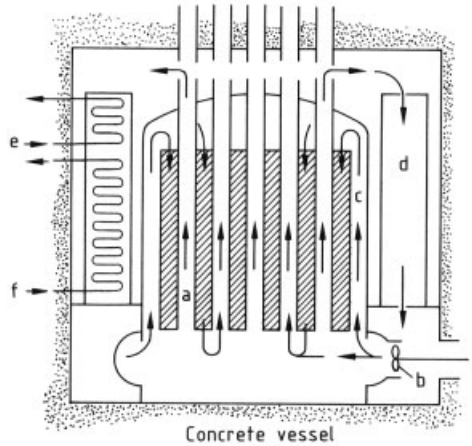

Figure 61. Diagram showing gas flow in the Hinkley AGR a) Fuel channel; b) Gas circulator; c) 50 % flow cooling shield and graphite; d) Boiler; e) Steam reheater; f) Boiler feed

Radiation-Induced Corrosion. Radiation-induced graphite corrosion by carbon dioxide is observed in both types of reactors. Irradiated carbon dioxide forms, among other ions, a $[CO_2 \cdot CO_2]_n{}^+$ complex with a lifetime of 0.1 μs, which is equivalent to a free path length of about 0.5 µm. The graphite surfaces, including the pores, will be attacked to this depth (when $n = 2$ in the reaction intermediate $[CO_2 \cdot CO_2]_n{}^+$) according to:

$$[CO_2 \cdot CO_2]_2^+ + C \rightarrow 2\,CO + 3\,CO_2$$

The reaction rate increases with carbon dioxide pressure and with radiation dose. This reaction is thus more important for AGRs than for Magnox reactors. The reaction is inhibited by its product carbon monoxide. Hence, the carbon monoxide concentration in Magnox stations is maintained at 1 – 1.5 %. This is achieved by use of a gas purification plant. Much higher concentrations would be necessary for AGRs, which would lead to excessive carbon deposition. For this reason a small amount of methane is injected continually into the primary circuit. Laboratory investigations have shown that methane is an order of magnitude more effective than carbon monoxide in suppressing radiolytic corrosion. The average methane concentration is ca. 500 ppm by

volume. Methane inhibits the corrosion reaction in two ways:

1) Radiolytic disintegration products of methane react with the reactive complexes formed by carbon dioxide radiolysis.
2) During methane radiolysis, finely divided carbon is deposited on the graphite surfaces, especially the pores, which protects the graphite against carbon dioxide oxidation. This protective effect is not completely identical with the protective oxide layer on metals. The oxide layer on metals is not attacked, whereas the deposited carbon is preferentially oxidized and functions as "sacrificial graphite." Methane must, thus, be injected continually.

Part of the methane is oxidized under AGR conditions according to:

$$CH_4 + 3\,CO_2 \rightarrow 4\,CO + 2\,H_2O$$

Most of the water produced is removed continually by special dryers to avoid condensation. This prevents the corrosion of metallic components by water and the increased corrosion of metals by carbon dioxide in the presence of water. The concentration of carbon monoxide must also be limited to avoid carbon deposition. Therefore, carbon monoxide is oxidized to carbon dioxide in a bypass catalyst bed.

/enddocument

Because of the differences in operating parameters among various AGRs–especially the pressure–the optimum carbon monoxide and methane partial pressures differ for each reactor, as do the $CO - CH_4$ ratios. An extensive research program was necessary to determine the optimum concentrations for each reactor.

The radiolytic reactions in the primary circuit yield a variety of mostly *organic compounds* (e.g., ethane, ethylene, propane, acetaldehyde). Table 33 shows the typical concentrations of organic compounds formed by radiolysis in the primary circuit of the Windscale AGR [107]. These concentrations are, however, so low that they do not affect reactor operation.

The principal *radioactive impurity* in the primary circuit is ^{41}Ar, which is formed via a (n, γ) reaction from ^{40}Ar. Argon- 40 is an impurity in the makeup carbon dioxide, or it can result from air leakage. Its very short half-life of 1.8 h means that it is of virtually no importance.

The *fission product concentration* is usually zero and, when it is not zero, it is very low. All fuel elements are monitored and replaced in the event of damage.

A chemical problem observed particularly in AGRs is *carbon deposition* on the fuel elements. The carbon is formed by thermal and radiolytic decomposition of methane:

$$CH_4 \rightarrow 2\,H_2 + C$$

and the reverse Boudouard reaction:

$$2\,CO \rightleftharpoons CO_2 + C$$

Another possible reason for carbon deposition is oil leakage from the primary coolant compressors. Evidence for this is given by the presence of carbonyl sulfide. Compressor oil can contain up to 0.6 % sulfur.

A 0.1-mm-thick carbon layer on the fuel elements can reduce heat transfer so drastically that the elements undergo damage. The carbon monoxide and methane concentrations must be kept within limits at which carbon deposition cannot occur.

Table 33. Typical analysis of a Windscale AGR coolant (760 ppm by volume CH_4/0.25 % CO)

Compound	Amount, ppm by volume
Methane	760
Ethane	32.1
Ethylene	3.75
Propane	9.0
Propene	0.59
Isobutane	1.16
n-Butane	1.04
2,2-Dimethylpropane	0.83
Isopentane	0.73
2,2-Dimethylbutane	8.30

3.2.2.5. Secondary Circuit Chemistry

The steam circuits of the early Magnox stations had circulatory boilers; later, once-through boilers were used. All AGRs are equipped with once-through boilers. The design details do not differ from similar circuits in fossil-fired stations. Water chemistry specifications are unchanged. The pH is controlled by ammonia, hydrazine is used to remove oxygen. Because the pressure in the steam circuits is always higher than in

the primary circuit, the transfer of radioactivity to the secondary circuit in case of a steam leak can be ruled out. The steam is lightly contaminated by small amounts of tritium, which diffuse through the steam generator steel into the steam circuits. However, because of the very low tritium concentration in the primary circuit, the tritium transport rate and, hence, the tritium concentration in the steam circuit are very low.

3.2.2.6. Operating Experience

Operating experience with gas-cooled reactors is generally good. Magnox stations have the reputation of being "nuclear cart horses". However, since about 1970, the gas exit temperatures in the Magnox stations had to be lowered, which has reduced their efficiency. The reason for this corrective measure was carbon dioxide corrosion of mild steels, which was unknown during design of the reactors.

In 1969, broken bolts were discovered initially in British and later also in French reactors. The corrosion was called *breakaway oxidation* because of spalled oxide layers. Corrosion tests before construction of the Magnox stations showed that mild steel forms a protective oxide layer below 450 °C in contact with carbon dioxide. The metal loss rates were so low that no effects were anticipated for the expected 30 years of reactor operation. However, testing stopped too soon. Initially, the protective oxide layer grew with parabolic kinetics. After about 10 000 h of operation the protective oxide layer spalled away from the metal surface. A porous oxide appeared with linear formation kinetics under the protective layer (Fig. 62). Chemical investigation showed that in both cases magnetite was formed according to:

$$3\,Fe + 4\,CO_2 \rightarrow Fe_3O_4 + 4\,CO$$

The breakaway oxide also contained 4 % carbon.

Linear, as opposed to parabolic, kinetics means that the oxide layer is not protective. Oxide growth led to high stresses that caused bolt failures. The effect was amplified by water [108].

Fortunately, the experiments to explain the damage revealed a marked temperature dependence of oxide formation. The porous oxide layer is not formed at $< 360\,°C$. The breakaway problem was solved for operational Magnox stations by lowering the gas exit temperature to 350 °C. In addition, the experiments showed that silicon steels and austenitic steels did not exhibit breakaway oxidation. Austenitic steels form a protective chromium oxide layer on the metal surface. Consequently, austenitic steels were used in those regions of AGRs susceptible to breakaway oxidation.

In the early years of operation, the ferritic parts of some of the steam generators exhibited erosion corrosion, caused by construction features that resulted in unfavorable flow patterns. Numerous shut-downs of the AGRs were caused by problems in the turbine sets (i.e., in the conventional part of the plant).

The AGR was capable of only limited development. The physical characteristics of the core do not allow an increase in power density and consequent reduction in core volume. Gas outlet temperatures are limited by graphite corrosion. Fuel burnup is low. Thus, development of this type of reactor ceased. None of the above limitations applies to the helium-cooled high-temperature reactor.

Interestingly enough, Great Britain– the birthplace of the AGR– did not proceed to the HTR. Only an experimental reactor with no steam turbine was built (Winfrith Dragon Reactor). The successor to the AGR was the internationally tried and tested pressurized-water reactor. However, in the early 1990s, prospects for the pressurized-water reactor are doubtful in the United Kingdom.

3.2.3. Helium-Cooled High-Temperature Reactors

The helium-cooled high-temperature reactor is the next stage of development in the gas-cooled reactor line. The use of helium as coolant and of graphite as the core structural material allows much higher operating temperatures than in AGRs and, hence, higher thermal efficiency. Helium has especially favorable characteristics as a coolant: it is chemically inert, has a high thermal conductivity, and has good neutron physics properties.

The cost of helium has fallen sharply in the last few years because it is a byproduct of the nat-

ural gas industry. The price of helium dropped to less than \$ 5/m^3 (STP) in the latter part of the 1980s. Experience with HTRs has shown that helium consumption only constitutes a small fraction of the operating costs.

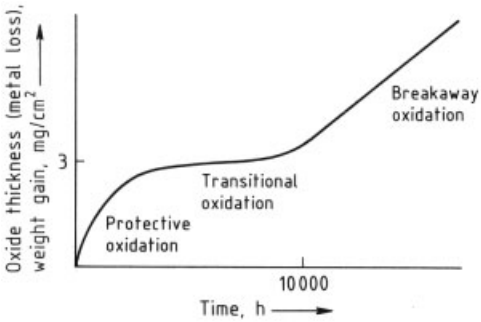

Figure 62. Kinetics of oxidation of mild steels in high-pressure carbon dioxide at 400 °C

3.2.3.1. Fuel Elements

Compared to carbon dioxide reactors, the fuel elements in helium- cooled HTRs are completely new. The fuel element consists of small particles of uranium and thorium oxide or uranium and thorium carbide embedded in a graphite matrix. The fuel particles have a diameter of ca. 500 µm. The particles are coated with one porous and one high-density pyrocarbon layer; newer fuel elements are also coated with a silicon carbide layer. The coated particles have excellent retention properties for fission products. The graphite matrix surrounding the coated particles forms an additional barrier to the release of fission products.

The two basic fuel element designs differ with respect to shape: the first is a sphere with a diameter of 6 cm; the second is a hexagonal block, 80 cm high, with a width over flats of 36 cm. Figure 63 shows a schematic of a coated particle and a spherical fuel element. A spherical fuel element contains 20 000 coated particles. The outer layer of the fuel element is fuel free. Hexagonal prismatic fuel elements have a central borehole for ease of handling and smaller channels for cooling of the coated particles and the burnable poison. Apart from these types of fuel element, rod-shaped fuel elements also exist (Dragon Reactor, Peach Bottom Reactor) but are not being developed at present.

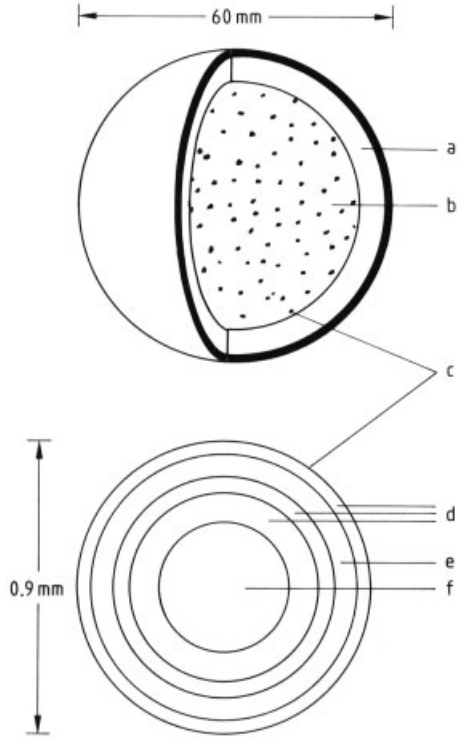

Figure 63. HTR fuel element
a) Graphite shell; b) Graphite matrix; c) Coated particle; d) Carbon layers; e) Silicon carbide layer; f) Uranium dioxide kernel

Because of the quasi-homogeneous fuel distribution and the use of *enriched uranium* much higher power densities can be achieved than in AGRs. For example, a typical AGR, the Hartlepool station, has a power density of 2.8 MW_{th}/m^3; several HTRs achieve power densities of 6 – 8 MW_{th}/m^3 (Table 34). The Dragon Reactor's high power density of 14 MW_{th}/m^3 was atypical because it was an experimental reactor. The AVR reactor (Arbeitsgemeinschaft Versuchsreaktor) had a power density of only 2.5 MW_{th}/m^3 to increase its inherent safety (at low power densities, no provision for removal of decay heat is necessary). For the same reason, the power density in the HTR module design is limited to 3.0 MW_{th}/m^3.

The neutron economics of HTR plants are very favorable: Construction of the core from only fuel and graphite means that the core contains no neutron-absorbing materials. Neutron leakage is also small.

Table 34. High-temperature reactors

	AVR	Dragon	Peach Bottom	THTR-300	Fort St. Vrain	HTR module
Type	experimental reactor	experimental reactor	experimental reactor	prototype	prototype	prototype
Location	Jülich, FRG	Winfrith, UK	Susquehanna river, Pennsylvania	Schmehausen, FRG	Platteville, Colorado	
Start-up	1967	1966	1967	1985	1976	
Final shutdown	1988	1975	1974	1988	1989	
Thermal power, MW_{th}	46	20	115	750	851	200
Electrical power, MW_e	15		40	308	342	70
Average power density, MW_{th}/m^3	2.8	14	8.3	6.0	6.3	3.0
Type of fuel element	spherical	rod	rod	spherical	hexagonal block	spherical
Number of fuel elements	110 000	47	804	675 000	1462	360 000
Coolant pressure, MPa	1.1	2.0	2.4	4.0	5.0	6.0
Maximum temperature in center of fuel elements, °C	1170	1227	1350	980	1260	845
Mean gas outlet temperature, °C	950	750	713	787	785	700
Helium circulation through core, kg/h	47×10^3	36×10^3	200×10^3	1064×10^3	1537×10^3	306×10^3
Steam pressure, ∗ MPa	7.3		10.0	18.6	17.3	19.0
Steam temperature, ∗ °C	505		538	535	538	530

∗ At outlet of final superheater.

The excellent neutron economy can be exploited in two conversion cycles, the uranium – thorium cycle and the uranium – plutonium cycle. In the former, nonfissile thorium is transformed into the fissile isotope ^{233}U:

$$^{232}Th + n \longrightarrow {}^{233}Th \xrightarrow{\beta^-,\,23.2\,min} {}^{233}Pa \xrightarrow{\beta^-,\,27.4\,d} {}^{233}U$$

The latter cycle is:

$$^{238}U + n \longrightarrow {}^{239}U \xrightarrow{\beta^-,\,23.5\,min} {}^{239}Np \xrightarrow{\beta^-,\,2.3\,d} {}^{239}Pu$$

The uranium – thorium cycle is especially suited to the HTR, whereas the uranium – plutonium cycle can be used more advantageously in the fast breeder reactor.

Because only graphite is used as fuel canning material and only helium as coolant, the core gas exit temperature can be raised to more than 1000 °C. The average gas exit temperature of the AVR in Jülich, Germany was increased to 950 °C in 1974. Lower temperatures were set only for experimental purposes. The change to much higher temperatures is limited by the present state of development of the fuel elements, because the retention capacity of the pyrocarbon and silicon carbide layers for solid fission products decreases at higher temperature. Another limitation is set by the steels used in the heat removal components, such as steam generators and other heat exchangers.

The high gas exit temperature offers more than merely better plant efficiency. The suggestion has been made that HTR power stations should be built as single-circuit units, in which the reactor coolant delivers energy to a gas turbine rather than a steam turbine via a steam circuit. High-temperature helium can be used as a source of heat for chemical processing, e.g., coal gasification (see Section 3.2.3.7).

3.2.3.2. Primary Circuit Materials

Graphite is the key material in an HTR primary circuit. Its most obvious property is its stability at high temperature. The temperature of sublimation is about 3620 K. Graphite is a good moderator because of its low neutron absorption. Graphite displays excellent mechanical properties at high temperature and thus can be used with advantage as a construction material. The coefficient of thermal expansion is much lower than that of metallic materials. This results in lower thermal stresses and favorable behavior during temperature transients. Potential energy

arising from neutron irradiation is stored in the graphite crystal structure at $< 300\,^{\circ}\mathrm{C}$. This stored energy can be released suddenly and can lead to overheating (Wigner energy). To avoid this effect, the graphite temperature even in the coolest parts of the primary circuit is kept above a certain level.

Another effect of irradiation on graphite is a temperature-dependent change in dimension. Below $300\,^{\circ}\mathrm{C}$, graphite shrinks parallel to the direction of the graphite grain and expands perpendicular to it. Above $300\,^{\circ}\mathrm{C}$, graphite shrinks both parallel and perpendicular to the grain direction. The extent of shrinkage depends on the dose. A maximum contraction is reached at $> 10^{22}\,\mathrm{n/cm^2}$, at higher doses, graphite exhibits turnaround and starts to expand. Thus, the linear dimensions may be larger than before irradiation.

Recently developed theories of solid-state physics allow the description, in principle, of the effects caused by neutron irradiation. However, because the graphite components have a very complicated structure, the changes in shape cannot be calculated exactly. For this reason, samples of graphite intended for reactor service are irradiated in material testing reactors. The fast-neutron flux in these test reactors is normally one to two orders of magnitude higher than in power reactors. Thus, changes relevant to power reactors can be investigated in a shorter time. Only those graphites that exhibit disturbance-free operating characteristics during irradiation tests are used in power reactors.

Graphite plays another important role as pyrolitically deposited carbon (→ Carbon, Chap. 4.2.2.). The pyrocarbon layer on the fuel particles has the same function as the metallic canning of fuel rods in other types of reactors. Pyrocarbon, a polycrystalline carbon, is formed by thermal decomposition of hydrocarbons. The particles are coated in a fluidized bed. Acetylene, for example, is used to manufacture the *porous buffer zone* and methane, to produce the *dense surface layer*. *Silicon carbide layers* are produced from silanes by using chemical vapor deposition.

The manufacture of pyrocarbon- coated particles is a complicated technical process, which was optimized as a part of HTR development. The mechanism of pyrolysis has yet to be fully explained. Irradiation of pyrocarbon results in changes in properties. As mentioned above, these changes cannot be calculated; the different types of fuel particles intended for reactor service must be irradiated in material testing reactors.

The most important factors in determining the loads on *metallic components* are the design parameters of the steam circuit. High-alloy austenitic steels are used, as in conventional power stations. The boiler superheaters in the Peach Bottom, Fort St. Vrain, THTR-300, and HTR module designs are manufactured from Incoloy 800 H. In the HTR module, the entire boiler is constructed from Incoloy 800 H.

3.2.3.3. Safety of High-Temperature Reactors

The safety of the high-temperature reactor is characterized by a series of inherent properties, which differentiate it from other reactor types and make it extremely safe.

1) The use of graphite as fuel element cladding and core structural material means that a core meltdown can be ruled out, because a very large difference exists between the normal operating temperature and the maximum tolerable temperature.
2) Temperature transients are slow because of the large thermal capacity of the graphite.
3) If the power density is limited (e.g., in the AVR to $2.5\,\mathrm{MW_{th}/m^3}$ or the HTR module design to $3.0\,\mathrm{MW_{th}/m^3}$), the fuel element does not rise above $1600\,^{\circ}\mathrm{C}$, even if direct cooling fails. Fission product release due to fuel particle bursting occurs at much higher temperatures. Decay heat can be removed solely by conductivity and radiation. In the HTR module design, heat is transferred to the extended surface cooler outside the steel vessel; in the case of AVR, to the steam generator. Figure 64 shows the change of maximum fuel element temperatures in the HTR module design without direct removal of decay heat. To demonstrate these safety properties a loss-of-coolant accident (i.e., failure of the core cooling) was simulated on the AVR. Only a small increase in temperature was observed.
4) Use of helium as coolant is another safety feature. Helium is chemically inert and does

not react with graphite or the metallic core components. Helium cannot be activated, with the exception of trace amounts of ^{3}He. The neutron absorption cross-section is small, so that the core reactivity is not increased in the event of a loss-of-coolant accident.

5) The use of coated particles in the fuel elements results in low contamination levels in the primary circuit and, hence, low radiation doses to operating personnel.
6) An important property of neutron physics is the strongly negative temperature coefficient over the entire temperature range of the reactor. Thus, large power excursions can be ruled out.
7) The chemical reactivity of graphite with air and water at high temperature is, to a certain extent, a disadvantage. The risk of steam entering accidentally via a boiler leak can be limited by construction measures; e.g., the boiler in the HTR module design is displaced to one side below the core. Operating experience also shows that the ingress of foreign substances into the core occurs rarely and is easily controlled.

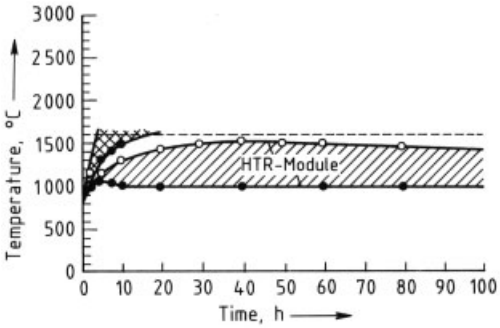

Figure 64. Maximum fuel element temperatures in high-temperature reactor module design without active removal of decay heat
○ = Atmospheric pressure; ● = Operational pressure

3.2.3.4. Types of High-Temperature Reactors

Various possibilities are available for the layout of the primary circuit in HTRs. Each design has its advantages and disadvantages. Table 34 shows layout data for both existing and planned reactors.

The Dragon Reactor. The Dragon Reactor in the United Kingdom was a joint project of the twelve European OECD nations. It was planned as a research reactor and had no turbine. The 20-MW thermal power was dissipated to the atmosphere via heat exchangers. The first reactor core contained 37 rod-shaped fuel elements. The power density was 14 MW_{th}/m^3, which is atypically high for an HTR. The core was contained in a steel pressure vessel. A graphite annulus was built as a neutron reflector around the core. The six heat exchangers were in external loops. The secondary circuit contained ^{6}Li as a neutron poison.

A major part of the research program involved the investigation of various fuels (e.g., UO_2, PuO_2, UC_2– ThC_2) with different coatings and the release and deposition of fission products. The behavior of nonradioactive impurities in helium was investigated by injection experiments.

The reactor was shut down in 1975 because the contributing nations could not reach agreement about further financing of the project.

Peach Bottom Reactor. The Peach Bottom Reactor in the United States was the first HTR power station. It was built by a consortium of 53 electrical utility companies as part of the reactor demonstration program of the U.S. government. Similar to the Dragon Reactor, its core consisted of fuel pins within an annular graphite reflector. The nuclear portion of the station was placed in a concrete containment. The concrete containment was surrounded by a steel containment, which was filled with nitrogen during operation. During its seven and a half years of operation before its final shutdown in 1974, the reactor achieved, in general, very good operating results. The availability was 37.2 % over the entire period of operation.

The biggest operational problem occurred with the first core. The coated particles of the first fuel element charge had only a single pyrocarbon coating, which in view of the dimensional changes caused by neutron irradiation was not an optimal design. The coating cracked. The improved fuel elements of later cores behaved excellently.

Early in the operating life of the reactor, chloride-induced stress corrosion cracking was

observed in the superheater. The superheater was replaced with Incoloy 800 H tubing.

Fort St. Vrain Reactor. The Fort St. Vrain Reactor design differed in some respects from the experimental Peach Bottom Reactor. The core, boilers, and helium compressors were placed in a prestressed concrete pressure vessel. Its main volume was an upright cylinder 23 m high with a diameter of 9.5 m. A section through the reactor is shown in Figure 65. The core contained 1462 hexagonal fuel element blocks, which were shuffled during the annual reactor shutdown and, depending on burnup, replaced.

The four helium compressors had water labyrinth bearings and were driven by steam turbines, fed by a bleed from the steam circuit.

The reactor went critical for the first time in 1974 and delivered electrical power in 1976. Two years later, it reached 70 % of its nominal full power. The start-up was marked by a series of problems.

During operation, serious temperature fluctuations occurred, caused by flow-induced mechanical vibration of the fuel element blocks. These vibrations occurred whenever the reactor reached more than 70 % of nominal full power. For this reason the reactor was not brought to full power before 1981. The problem was solved by an improvement in the attachment of the fuel element heads. The largest problem was caused by the water labyrinth compressor bearings, which leaked time and again to a greater or lesser degree. The drying out of the core was made more difficult and extended by the fact that the ceramic insulation between the liner and PCRV was wet. This moisture damaged other reactor components. Repair of the damage led to long shutdown periods. Long-term, trouble-free operation was obviously impossible, and the reactor was shut down permanently in 1989. The final determinant of the decision to shut down the reactor was the discovery of hairline cracks in the steam feeds to the compressor drives.

The problems that arose were not typical for HTRs. Together with financial difficulties, however, these technical problems were the reason why the reactor construction company General Atomic gave up various contracts for larger HTRs.

AVR. The AVR has demonstrated the longest and most successful operation of all high-temperature reactors. The differences between the AVR and the aforementioned reactors are the spherical fuel elements and the arrangement of core and steam generator (see Figure 66). The steam generator was placed above the core and very close to it, separated by a slotted roof of graphite – kohlestein (ungraphitized carbon bricks). The helium entered the core at the base, was heated, and flowed through the slots to the steam generator. Cooled helium left the steam generator and flowed around the outside of the core to cool both the core shielding and the pressure vessel. This construction avoided the problems caused by hot gas ducts, especially insulation problems.

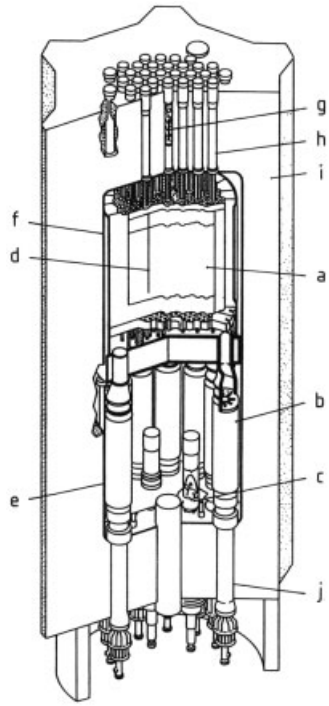

Figure 65. Cross-section of the Fort St. Vrain Reactor
a) Core; b) Steam generator; c) Blower; d) Control rod; e) Liner; f) Thermoshield; g) Control rod drive; h) Top ducts; i) Prestressed concrete pressure vessel; j) Bottom ducts

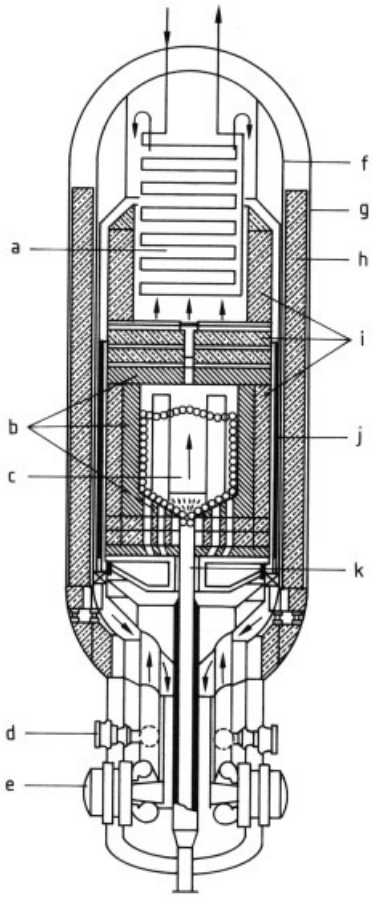

Figure 66. AVR primary system
a) Steam generator; b) Graphite reflector; c) Core; d) Flapper valves; e) Coolant blowers; f) Inner reactor vessel; g) Outer reactor vessel; h) Biological shield; i) Carbon bricks; j) Thermal shield; k) Fuel element discharge pipe

The core, the steam generator, and the two helium blowers (compressors) were placed in a double-walled steel pressure vessel.

The ca. 110 000 fuel elements were circulated continually. Spheres that had reached a certain burnup were removed during this process and replaced by fresh fuel elements. 50 fuel elements were replaced per day, on average. The fuel element breakage rate was very low. (Only about one in 20 000 fuel elements was damaged.)

The reactor temperature was controlled by four boron rods in buttresses in the side reflector. The rods that had not been directly driven into the core also served as a reactor shutdown system. Reactor power output was controlled by the compressors.

Only low-alloy ferritic steels were used in the steam generators to avoid the risk of stress corrosion cracking.

The most important operational aim was to demonstrate the smooth operation of a pebble bed high-temperature reactor. The one operational parameter that stood out above all the others was its excellent operational availability. This increased from 53 % in the first full year of operation (1968) to 92 % in 1976. The average availability over the 21 years of operation (December 1967 to December 1988) was 66 %. Many periods of reactor shutdown were caused by the preparation and operation of reactor experiments. The longest shutdown, about one year, arose after massive water ingress in May 1978. About 30 t of water escaped into the shutdown reactor through a small hole in the boiler. What caused the hole is unknown. Most of the water was collected via a valve in the fuel element reloading circuit; the remainder was removed by repeated evacuation and nuclear heating. The experience gained in the incident showed that no damage of any concern arose because of it.

The low coolant activity and, hence, low radioactive deposition led to an extremely low radiation dose to reactor personnel. The fuel elements showed good operational characteristics with very low fission product liberation rates, although in some cases exceptionally high burnups could be obtained.

During the entire period of operation a large number of experiments were carried out. Stress was placed on large-scale fuel tests and investigation of the liberation and deposition of solid fission products. A loss-of-coolant accident was simulated to demonstrate the inherent safety. The "four-rods-jammed" experiment gave two main results. (1) The reactor would shut down quickly because of the negative temperature coefficient even if the shutdown rods could not be activated. (2) The decay heat would give rise to only a small temperature increase in the core. The behavior of chemical (i.e., nonradioactive) impurities under reactor conditions was investigated by numerous experiments in which impurities were injected into the primary circuit.

The AVR was shut down finally at the end of 1988, having fulfilled its aim, to demonstrate the feasibility of operating a pebble bed reac-

tor for 21 years with an extensive experimental program.

THTR-300. All the components in the THTR-300 (Fig. 67) were placed in a PCRV (inside diameter 15.90 m, inside height 15.30 m, wall thickness 4.45 – 5.10 m). All openings were equipped with reinforced pipes. The inner surface of the pressure vessel was fitted with a liner (gastight insulation) with welded pipes through which cooling water flowed. The liner cooling guarded against overheating of the PCRV.

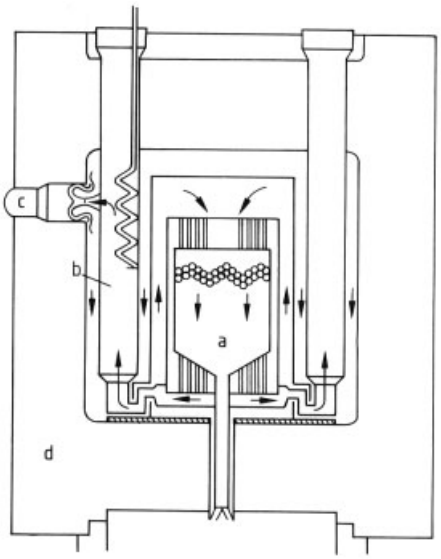

Figure 67. Thorium high-temperature reactor (THTR)
a) Core; b) Steam generator; c) Cooling gas blower; d) Prestressed concrete reactor vessel

Like the AVR, the THTR-300 was a pebble bed reactor: 675 000 spherical fuel elements were circulated continuously and exchanged for fresh elements, if required. As in the case of the AVR the reactor did not have to be shut down to change fuel. To determine when fuel elements should be removed, fuel burnup was determined with a special burnup reactor. Damaged fuel was also removed. The fuel elements contained thorium and highly enriched uranium in a ratio of 10: 1. The absorber elements of the first core contained boron and hafnium. Reactor temperature was controlled by 36 control rods in the reflector. Reactor power was controlled between 40 and 100 % of full power by the compressors. The reactor was shut down by driving 42 shutdown rods into the pebble bed.

Helium was driven down through the core by six electrically driven compressors. Hot helium from the core was collected in a chamber below the core that was equipped with graphite pillars, and then passed upward through one of six steam generators. The helium inventory of 47 000 m^3 (STP) was circulated about twice per minute. The steam generators were once-through boilers with intermediate superheaters. The main cooling water circuit was a closed circuit cooled by a dry cooling tower.

The construction time was greatly exceeded because the reactor was a prototype reactor and the licensing authorities did not possess the proper HTR regulations. During construction, regulatory requirements changed often.

During commissioning, problems with the fuel reloading system occurred, which were solved by slight modification of the equipment. In addition, other problems arose due, for example, to broken bolts in the hot gas duct.

The reactor was not brought up to power again after a planned shutdown in 1988 because agreement could not be reached on the responsibility for financial risks. The decision to shut the reactor down can only be viewed as a political decision of the provincial government. In the second half of the 1980s, public acceptance of nuclear power–after the incidents at Three Mile Island and Chernobyl–was very low.

HTR Module. The HTR module design was developed by Siemens/Interatom and evolved from the AVR design in many aspects, both in reactor philosophy and in component design. The fuel elements are spherical, with a diameter of 6 cm as in the AVR and the THTR-300. Figure 68 shows a section through the reactor and the nuclear heat generation system. The nuclear components and the steam generator are in two separate adjacent steel pressure vessels. Helium is driven downward through the core. It flows with an exit temperature of 700 °C through a horizontal gas duct to the steam generator, rises through the steam generator, and is cooled. Helium then returns through slits around the steam generator tube bundle to the blowers. The cold gas exiting the blower flows in a coaxial pipe around the hot gas duct back into the reactor pressure vessel. It returns to the core via slits in the side reflector. The core diameter is only 3 m. Because of this limited diameter, shut-

down rods that penetrate the pebble bed are not required and the reactor can be shut down with control rods in holes drilled in the side reflectors. A second design limitation is that the fuel element temperature should always be $< 1600\,°C$, even if the active cooling system fails during decay heat removal. The reason is that solid fission products are retained in the coated particles below 1600 °C. This requirement limits the core power density to $3\,MW_{th}/m^3$. To obtain maximum power from the reactor, the core should be as high as possible. The upper limit of 10 m is set by the pressure drop across the core and the problem of xenon swing. The design power of the HTR module is thus limited to $200\,MW_{th}$. For larger stations, a number of HTR modules will be placed together on a site. Auxiliary plants for gas purification, water conditioning, etc., would be common to several module reactors.

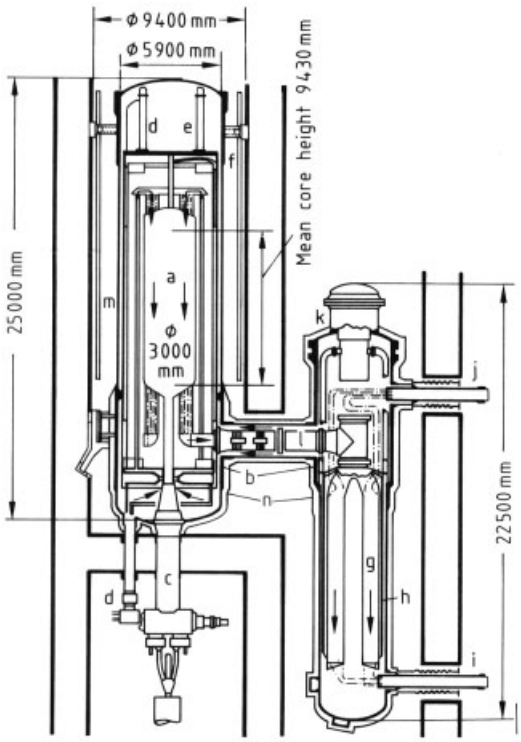

Figure 68. Cross-section of a modular unit with steam generator (primary circuit)
a) Pebble bed; b) Pressure vessel; c) Fuel discharge; d) Small absorber balls; e) Reflector rod; f) Fuel loading; g) Steam generator: pipe assembly; h) Outer shroud; i) Feed line; j) Live steam line; k) Blower; l) Hot gas duct; m) Surface cooler; n) Insulation

In the event of failure of the active cooling in the primary circuit, decay heat is removed simply by conduction and radiation to sheet heat exchangers outside the reactor pressure vessel. Special circuits for decay heat removal are not required. The sheet coolers are also used during normal operation to remove heat leaving the primary circuit.

The displaced layout of steam generator and reactor pressure vessel is advantageous because in the event of steam generator leakage, only a limited amount of water can reach the core. Because the decay heat is removed by the sheet heat exchanger, the steam generator can be decoupled from the reactor if necessary. The displaced layout is also of advantage during periodic reactor inspection.

The HTR module is suited to the generation of process heat for chemical applications. Gas exit temperatures are increased to 950 °C in this case. The many years of operation of the AVR at a gas exit temperature of 950 °C have shown that this is possible with current designs of fuel elements. Various plant concepts are described in Section 3.2.3.7.

3.2.3.5. Chemistry of the Primary Coolant in the High-Temperature Reactor

The primary coolant chemistry in an HTR is determined by the following parameters:

1) The core components and fuel element surface layers are made of graphite. The maximum graphite temperatures are ca. 1100 °C.
2) The coolant is very pure helium.
3) The heat exchangers are the hottest metallic components in the primary circuit. Maximum temperatures at the steam generators in electricity-generating HTR stations are 500 – 550 °C. The process heat units (developmental status in the 1980s) have heat exchanger design temperatures up to 950 °C.

The following *impurities* are generally found in the helium of the primary circuit: H_2O, CO_2, N_2, CO, H_2, CH_4, Ar, noble fission gases, tritium, solid fission products, and dust.

Oxygen cannot be detected above 250 °C because it reacts with, or is chemically adsorbed, by graphite. Nonradioactive impurities are due to small, technically unavoidable air and water

leaks. Hydrogen also diffuses from the steam circuits into the primary circuit. Methane is formed radiochemically.

The concentration of all impurities is in the region of ppm by volume. Table 35 gives the measured concentrations in the AVR, Peach Bottom Reactor, and THTR-300 during normal operations. The values are given in partial pressure (pascal) to allow comparison. The partial pressures were calculated by multiplying the concentration (ppm by volume) by the system pressure.

Table 35. Helium impurities in primary circuit during normal operation (in Pa)

	AVR	Peach Bottom	THTR-300
H_2O	≤ 0.05	< 1.0	≤ 0.05
CO_2	$< 0.5-2.0$	< 0.1	$< 0.4-0.6$
CO	30.0 – 80.0		$< 1.2-1.5$
		1.0	
H_2	5.0 – 10.0	20.0	1.5 – 3.5
CH_4	0.2 – 1.0	2.0	$< 0.4-0.8$
N_2	5.0 – 30.0		< 0.4
		1.0	
Ar	< 0.5		< 0.4
He pressure, MPa	1.1	2.4	4.0

The average concentrations in the Dragon and Fort St. Vrain reactors are not listed in Table 34 because they were high due to external influences. Numerous gas injection experiments into the primary coolant were conducted in the Dragon reactor. In the Fort St. Vrain reactor, numerous leaks occurred from the water labyrinth bearings in the helium compressors.

An important reason for the relatively high concentrations of impurities in the AVR was the presence of carbonate and bicarbonate in the Kohlestein (an ungraphitized carbon material), which decomposed thermally with increasing time of operation. It should be noted that the reactor operation was not influenced by this fact, with the exception of the commissioning phase.

Graphite Corrosion. At the normal operating temperature of 1100 °C the fuel element and core graphite is very reactive toward oxygen, water, and carbon dioxide. Because even trace impurities can lead to unacceptably high corrosion rates, the demands on helium purity are high. In addition, any leakage into the reactor must be avoided as far as possible. Experience so far suggests that both demands can be met. Helium is a byproduct of natural gas production and can be separated from natural gas or from air by liquefaction. Both processes yield a very pure product.

The important graphite corrosion reactions are:

$C + H_2O \rightleftharpoons CO + H_2$	(heterogeneous water gas reaction)
$2C + O_2 \rightleftharpoons 2CO$	
$C + O_2 \rightleftharpoons CO_2$	
$C + CO_2 \rightleftharpoons 2CO$	(Boudouard reaction)
$C + 2H_2 \rightleftharpoons CH_4$	
$CO + H_2O \rightleftharpoons CO_2 + H_2$	(water shift reaction)

The rates of all graphite corrosion reactions are very temperature dependent. The heterogeneous water gas reaction, which is a reaction of special practical importance, is extremely slow at < 500 °C. Thus, an HTR core can be dried without damage at this temperature in case of water leakage. The kinetics of the reaction of graphite with *water* can be divided into three regimes with increasing temperature (Fig. 69):

1) *Chemical regime* (volume reaction). Between 500 and 800 °C the graphite corrodes at the same rate throughout the blocks.
2) *In-pore diffusion regime*. Between 800 and 1200 °C the water concentration decreases with increasing distance from the graphite surface. The corrosion rate decreases in the same direction. The in-pore diffusion regime is the most important reactor corrosion because water leakage occurs at these temperatures.
3) *Mass transport regime*. Above 1200 °C, water reacts so rapidly with graphite that reaction occurs on the graphite surface.

The reaction products hydrogen and carbon monoxide inhibit the graphite – water reaction because of adsorption effects on the graphite surface.

Graphite reacts with *oxygen* under similar conditions much faster than with water. Oxygen is chemically adsorbed on graphite irreversibly above ca. 100 °C. The chemical reaction starts at about 250 °C. The helium must be very pure because of the high chemical reactivity of graphite with oxygen. No air must be allowed to enter the primary circuit during operation. This is, however, unlikely because of the high pressure of the system during operation.

The *Boudouard reaction* (graphite – carbon dioxide) is a secondary reaction of the

graphite – water and graphite – oxygen reactions. Under ca. 700 °C the reaction rate is very low. The reaction is inhibited by carbon monoxide, a reaction product. The rate of reaction of each of the corrosive gases with graphite relative to the rate of the graphite – carbon dioxide reaction (at 800 °C and 10 kPa) may be summarized as follows:

$C + O_2$	1×10^5
$C + H_2O$	3
$C + CO_2$	1

The rate of *radiochemical* corrosion of graphite by hydrogen is several orders of magnitude smaller and depends on reactor power. Experience has shown that this reaction has no practical importance.

In contrast to carbon dioxide-cooled reactors, radiolytic reactions play practically no part in the primary coolant, except for the graphite hydrogen reaction already mentioned.

This was demonstrated by water and carbon dioxide injection experiments in the AVR. The AVR was ideally suited for this purpose because the power could be varied while the outlet temperature was kept constant. No difference was observed in reaction rates at the same temperatures in experiments at half or full power. This does not mean that no radiolytic graphite corrosion by water and carbon dioxide occurs, but that the thermal reactions are much faster and hence determine the reaction rate.

Laboratory investigations have shown that traces of *heavy metals* (e.g., iron) and *fission products* (Ca, Ba, Sr) catalyze all gas – graphite reactions. Operating experience is that iron occurs in the primary circuit as the oxide and is ineffective as a catalyst. The fission product concentrations on the graphite surfaces are so small that postirradiation examination of fuel elements, with a few exceptions, shows no increased graphite corrosion.

Carbon Transport. During the cooling of helium in the primary circuit the aforementioned graphite corrosion reactions can run in the reverse direction on the surface of metal catalysts (e.g., on the metallic parts of the steam generators). This gives rise to carbon deposition. In addition, carbon is formed by decomposition of the methane present in the gas space. This carbon, together with fuel wear, gives rise to a low concentration of dust in the primary circuit. Dust concentrations measured in various HTRs are between ca. 1 and 10 μg/m^3 (STP). Dust collects in various parts of the primary circuit. The dust is of radiological importance because it is the transport medium around the circuit for various solid fission products. Gas – graphite reactions are not influenced by the dust. Carburization of steam generator steels does not occur because of the relatively low temperature.

Gas – Metal Reactions. Carburization and other chemical reactions of trace gases with metallic materials occur, however, if an HTR is used to generate process heat. Numerous experiments have shown that trace contaminants in helium can give rise to corrosion (H_2O, CO_2), carburization (CO, CH_4), and decarburization. The type of process depends on the gas composition, the metal used, and the temperature.

Long-term experiments show that chromium, in particular, determines the alloy behavior [109]. The chromium content of almost all suitable alloys is at least 20 wt %.

Chromium reacts with *water* to yield the oxide:

$$2\,Cr + 3\,H_2O \rightarrow Cr_2O_3 + 3\,H_2$$

Chromium carbides are *decarburized*, e.g.,

$$Cr_{23}C_6 + 6\,H_2O \rightarrow 23\,Cr + 6\,CO + 6\,H_2$$

Carburization reactions with carbon monoxide and methane are

$$9\,Cr + 3\,CO \rightarrow Cr_7C_3 + Cr_2O_3$$

$$7\,Cr + 3\,CH_4 \rightarrow Cr_7C_3 + 6\,H_2$$

The gas composition can be adjusted so that a stable protective chromium oxide layer forms. A minimum partial pressure of carbon monoxide in the primary coolant and a certain $CO - H_2$ ratio are essential. Operating experience with HTRs shows that because of small unavoidable air and water leaks, and because of desorption reactions from graphite, low concentrations of impurities are always present in the primary circuit. The conditions for the stability of the protective oxide layer (a certain amount of CO and H_2O)

are, thus generally fulfilled during normal operation. If the concentration of carbon monoxide is too low during normal operation, the primary coolant can be dosed with carbon monoxide.

If a protective oxide layer is not present, chromium carbides can be formed at $< 850\,°C$. At very high carbon monoxide concentrations, carbon deposition is possible at $\leq 650\,°C$. Long-term experience with the AVR showed, however, that despite continually high carbon monoxide concentrations and the use of unalloyed steels, only a small amount of carbon deposition was observed.

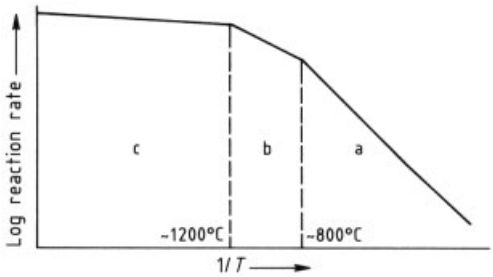

Figure 69. Temperature regimes of graphite corrosion with water vapor
a) Chemical regime; b) In-pore diffusion regime; c) Mass transport regime

Helium Purification Plant. The helium purification plant has the following functions:

1) Removal of the corrosive impurities water and carbon dioxide
2) Limitation of carbon monoxide and methane to minimize dust formation
3) Removal of tritium to reduce the amount that diffuses into the secondary circuits (hydrogen is retained along with tritium)
4) Removal of nitrogen to minimize ^{14}C formation
5) Removal of radioactive noble gases. However, only the concentrations of the long-lived noble gases are decreased, e.g., ^{85}Kr, (half-life 10.8 a) and ^{133}Xe (half-life 5.3 d).

Only small amounts of graphite dust can be removed because graphite is deposited in the circuit and does not reach the helium purification plant. The dust filter upstream serves to protect the helium purification plant.

The various steps in purification are adsorption processes on solid adsorbents. Gas scrubbing, found in industrial gas purification, is not used. Gas purification occurs in a series of steps. First, hydrogen and carbon monoxide are catalytically oxidized to water and carbon dioxide on a copper oxide bed at ca. 300 °C. Second, carbon dioxide and water are adsorbed on a molecular sieve. Methane, nitrogen, argon, and both active and inactive xenon and krypton isotopes are adsorbed on activated charcoal at liquid nitrogen temperature. On the AVR, hydrogen and carbon monoxide are converted with oxygen on a platinum catalyst bed. This process has functioned well.

A helium purification plant consists normally of two parallel streams. One of these is on-line, while the other is regenerated. For economic reasons the plant is designed as a bypass unit.

Fission and Activation Products. Although, in principle, the same fission and activation products occur in high-temperature reactors and light-water reactors, the differences between their release and their behavior are marked. The principal reason can be found in the new HTR fuel concept in which graphite is used as a canning material instead of metal. A second reason is the fine fuel distribution in coated particles.

The gaseous fission products, (e.g., noble fission gases and tritium) are almost completely retained under normal conditions. The retention capability for solid fission products is also very good. If no broken particles are present, fission product release is caused almost exclusively by the uranium contamination that remains outside the coating during fuel fabrication. Depending on the type of fuel, the average uranium contamination is between 10^{-4} and 10^{-5} (ratio of fissile material outside the fuel particles to that inside). This value is one of the most important parameters in fuel element fabrication.

In addition to the coating, the graphite matrix and the fuel-free zone form an additional barrier against fission product release. Depending on their half-lives, a large proportion of *fission product gases* decay during diffusion through the pores of the graphite. The graphite of the core structural components and of the fuel element surfaces forms a very effective sink for all *solid fission products* due to its large adsorptive capacity. The concentrations of solid fission and activation products are so low that they can be

determined only after a preconcentration step. The majority of *tritium* is chemisorbed on the graphite of the fuel elements and the core structural components.

Table 36 shows typical noble gas concentrations measured on the AVR at an average gas exit temperature of 950 °C in September 1983. The R/B value is normally given, in addition to the absolute concentration. The R/B is a dimensionless number–the ratio of the amount of activity released R to the activity formed B. The R/B for ^{85}Kr cannot be measured because it does not reach equilibrium in the primary coolant. ^{85}Kr has a long half-life and accumulates in the fuel particles. Therefore, the ^{85}Kr concentration in the coolant is an important parameter because a marked increase in concentration is an indication of broken fuel particles. In case of air ingress ^{41}Ar is detected in the coolant.

Table 36. Noble gas fission products in the AVR primary circuit (950 °C, 1983)

Noble gas	Half-life	Concentration, Bq/m^3	R/B
^{85m}Kr	4.48 h	7.88×10^7	6.2×10^{-6}
^{85}Kr	10.76 a	0.2×10^7	
^{87}Kr	76.3 min	11.33×10^7	4.4×10^{-6}
^{88}Kr	2.8 h	11.40×10^7	3.6×10^{-6}
^{89}Kr	3.18 min	4.87×10^7	1.2×10^{-6}
^{90}Kr	32.2 s	1.27×10^7	0.3×10^{-6}
^{91}Kr	8.6 s	0.05×10^7	0.02×10^{-6}
^{133}Xe	5.29 d	6.80×10^7	4.3×10^{-6}
^{135}Xe	9.17 h	12.53×10^7	2.6×10^{-6}
^{135m}Xe	15.3 min	2.22×10^7	1.4×10^{-6}
^{137}Xe	3.8 min	4.11×10^7	0.8×10^{-6}
^{138}Xe	14.1 min	2.08×10^7	0.4×10^{-6}
^{139}Xe	39.7 s	0.69×10^7	0.2×10^{-6}

The R/B is a useful method for comparing different HTRs. The given AVR values are of the same order of magnitude as in other reactors.

The fission product concentration in the gaseous coolant is temperature dependent. Quantitative relationships can be investigated in the AVR because the average core gas exit temperature can be altered at full power between 770 and 950 °C. The temperature dependence of the release of gaseous fission products is weak, in contrast to that found for solid fission products.

The concentration of solid fission products in the coolant is an order of magnitude lower than that of gaseous fission products. One reason for this, the deposition on graphite surfaces in the primary circuit, was mentioned above. Deposition also leads to a concentration difference for most solid fission products between the hot and cold regions of the circuit. The concentration in the hotter regions is higher. Table 37 gives the fission product concentrations determined on the AVR at a gas exit temperature of 950 °C in 1984. The activities were determined by preconcentration of 2700 m^3 of helium over a two-week period. Because of the long preconcentration time, both concentration minima and maxima were smoothed out.

Table 37. Solid fission products in the primary circuit of the AVR at 950 °C (1984)

Isotope	Half-life	Concentration in hot gas, Bq/m^3 (STP)	Concentration in cold gas, Bq/m^3 (STP)
^{90}Sr	28.5 a	70	45
^{110m}Ag	250.4 d	350	50
^{131}I	8.04 d	320	17
^{134}Cs	2.06 a	300	30
^{137}Cs	30.1 a	555	92

Activation products, such as ^{60}Co, ^{59}Fe, ^{51}Cr, ^{54}Mn, and ^{46}Sc (scandium is an occasional impurity in graphite), are found in the primary circuit in addition to solid fission products. Their concentrations are in the same range as the fission products.

Tritium. Tritium plays a special role in all gas-cooled reactors because it is the only isotope that can diffuse into the steam circuits under all operational conditions. The tritium in the primary coolant is principally an activation product of ^{3}He, ^{6}Li and ^{10}B. Tritium produced by fission is almost completely retained in the coated particles, as shown by postirradiation examinations.

Tritiated hydrogen (HT) makes up more than 90 % of the gas-phase tritium. Tritium is also found as CH_3T and HTO. The majority of tritium is chemically adsorbed on the graphite components in the primary circuit and can be released by exchange reaction with nonradioactive hydrogen. (This exchange explains the observation that the concentrations of nonradioactive and tritiated hydrogen in the primary coolant are linked). Tritium in the primary coolant is removed, together with nonradioactive hydrogen, in the helium purification plant. A small proportion diffuses as tritium atoms through the boilers into the steam circuits. There it forms tritiated water. The gas- and steam-side oxide films form

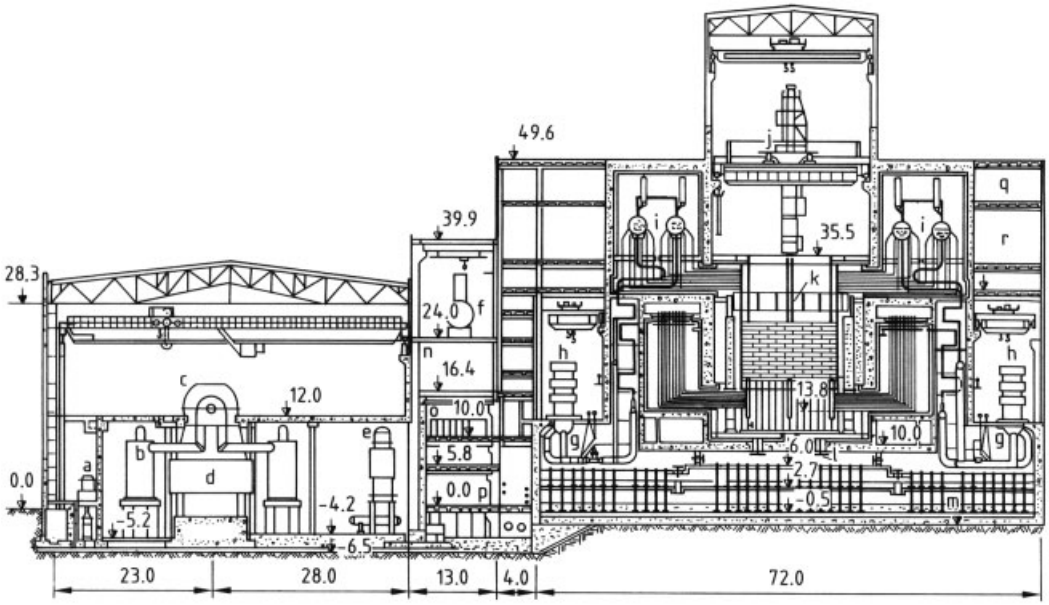

Figure 70. Cross-sectional view of reactor building, switchyard building, and turbine hall at Smolensk power station (dimensions are given in meters)
a) Condensate pump; b) Separator/steam superheater; c) Turbine; d) Condenser; e) Low-pressure heater; f) Deaerator; g) Reactor coolant pump; h) Electric motor of reactor coolant pump; i) Steam separator; j) Refueling mechanism; k) RBMK-1000 reactor; l) Accident localization system check valve; m) Suppression pool; n) Pipe aisle; o) Control room; p) House switchgear location; q) Exhaust ventilation plant location; r) Plenum ventilation plant location

an important permeation barrier. The steam-side oxide layer can be improved by oxygen treatment of the feedwater, if certain conditions are fulfilled.

Tritium concentration in the steam circuit is low. The measured concentration in the steam circuits of the AVR and THTR-300 was 10^4–10^5 Bq/kg.

3.2.3.6. Steam Circuits

Only once-through boilers have been used in HTRs (an exception is the Dragon Reactor, which did not have any turbine circuit). The steam quality is similar to that in modern fossil-fired power stations. Thus, in HTRs the same impurities can be expected as in fossil-fired power stations. These impurities arise from corrosion, condenser leakage, or breakthrough in the feedwater conditioning plants. Thus, chemical impurity guidelines for the once-through boilers in fossil-fired power stations can be used. In general, the steam circuits are alkaline conditioned. Combined feedwater conditioning is used on the AVR. In combined conditioning the feedwater is dosed with oxygen. The oxygen serves as a corrosion inhibitor.

If the steam generator lies in the neutron flux from the core, radioactive corrosion products (e.g., ^{59}Fe, ^{60}Co, ^{65}Zn) are found in the secondary circuits. This was true of the AVR and, to a lesser extent, the THTR-300. Because the pressure in the steam circuits is always much higher than in the primary circuit, fission products cannot pass into the steam circuits.

3.2.3.7. Nuclear Process Heat

The HTR need not be used only to generate electricity; it is also a versatile heat source. The suitability of the HTR pebble bed reactor as a source of process heat was demonstrated in 1974, when the gas exit temperature of the AVR was increased to 950 °C. Because of its inherent safety, an HTR unit can be built as part of a chemical plant or other industrial plant without special protective measures. Coal gasification was investigated as part of a series of studies in the

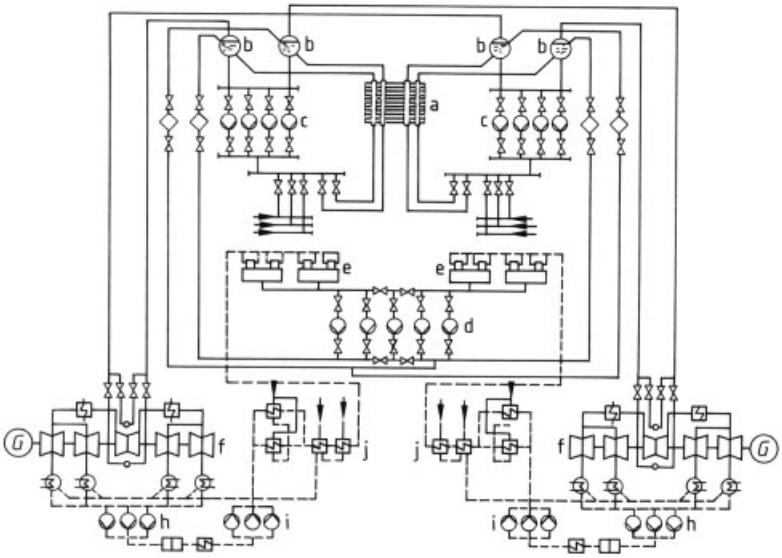

Figure 71. Schematic of the coolant and power conversion systems
a) Reactor; b) Steam separator; c) Reactor coolant pump; d) Feedwater pump; e) Deaerator; f) Turbine; g) Condenser; h) First-stage condensate pump; i) Second-stage condensate pump; j) Preheater

Projekt Nukleare Prozeßwärme (nuclear process heat project). Components were tested for all applications, and suitable materials were chosen in a material testing program.

The following areas of application were investigated among others:

Coal gasification by hydrogenation
Steam reforming of coal
District heating with combined heat and power generation
Process steam for the chemical industry
Desalination of seawater
Tertiary oil production
Natural gas reforming

Coal gasification yields substitute natural gas (SNG) or synthesis gas. The latter is a starting material in the industrial production of hydrogen, methanol, and gasoline, among others.

3.3. Graphite-Moderated, Light-Water-Cooled, Pressure Tube Reactors

The graphite-moderated, light-water-cooled, pressure tube reactor has a unique design, which has been developed and constructed only in the former Soviet Union. With a total of 15 units connected to the grid by the end of 1989 (see Table 38), this type of reactor constitutes an important part of electrical power production in the former Soviet Union.

Table 38. Soviet graphite-moderated, light-water-cooled reactors connected to the grid ∗

Reactor	Gross capacity, MW_e	Year of grid connection	Year of commercial operation
Beloyarsky-2	160	1967	1969
Chernobyl-1	1000	1977	1978
Chernobyl-2	1000	1978	1979
Chernobyl-3	1000	1981	1982
Ignalina-1	1500	1983	1985
Ignalina-2	1500	1987	1987
Kursk-1	1000	1976	1977
Kursk-2	1000	1979	1979
Kursk-3	1000	1983	1984
Kursk-4	1000	1985	1986
Leningrad-1	1000	1973	1974
Leningrad-2	1000	1975	1976
Leningrad-3	1000	1979	1980
Leningrad-4	1000	1981	1981
Smolensk-1	1000	1982	1983
Smolensk-2	1000	1985	1985

∗ Data from Dec. 31, 1989.

The graphite-moderated, light-water-cooled, pressure tube reactor differs in essential design characteristics from other reactors. The first unit

of this type, Beloyarsky-1 (capacity 100 MW_e) began commercial operation in 1964; the second, Beloyarsky-2 with a capacity of 200 MW_e, in 1967. After these two demonstration plants, the former Soviet Union began to construct full-scale 1000-MW_e units (Soviet abbreviation: RBMK-1000). The first RBMK-1000 was connected to the grid at the Leningrad atomic power station in 1973. By the end of 1989, 13 RBMK-1000 were operating. In 1985, the first of two larger units of this reactor type with a capacity of 1500 MW_e (RBMK-1500) began commercial operation at Ignalina. The design and the core size of the RBMK-1500 are similar to those of the RBMK-1000, but its fuel power density is 50 % higher.

Significant differences in design exist between the first-generation RBMK-1000 units, such as Leningrad 1 to 4, and the more recent second-generation units, such as Chernobyl 3 and 4 or Smolensk 1 and 2. Because more information is available on second-generation units, these RBMK-1000 units are described in the following sections.

3.3.1. General Design Characteristics

The RBMK-1000 reactors are constructed as twin units. Auxiliary building, switchyard building, and turbine building are common to both units. The two reactor buildings and the auxiliary building located between them are integrated into one large building. Parallel to this large building is the turbine hall with the four turbine generators (two for each unit). The reactor and auxiliary building are connected by the switchyard building to the turbine hall. The switchyard building contains separate control rooms, switchyards, cable rooms, and venting systems for the two reactor units (Fig. 70).

The RBMK-1000 reactor is a vertical pressure tube, graphite-moderated, boiling-water-cooled reactor with online refueling. The uranium dioxide fuel is located in 1661 vertical pressure tubes. The reactor coolant system consists of two independent loops supplying coolant to separate halves of the core (Fig. 71). Each loop contains two steam separators (b) and four reactor coolant pumps (c) (three in operation and one on standby under normal conditions).

Coolant enters the pressure tubes from the bottom at 270 °C. In flowing through the pressure tubes the coolant is heated and partially evaporates. Coolant pressure and temperature at the core outlet are 7 MPa and 284 °C. The average core outlet fraction of steam is 14.5 % at full power. Steam and water are separated in the separators (b) of the corresponding loop. Reactor coolant pumps (c) supply the water back to the headers, which distribute it via subheaders to the individual pressure tubes of the reactor core (a). Steam is fed to the two turbine generators (f)–with an output of 500 MW_e each–via steam lines connected by headers. After leaving the low-pressure turbine, steam condenses in the main condenser (g). The condensate is pumped back to the steam separators by the condensate (h, i) and feedwater pumps (d). It enters the steam separators at 165 °C and mixes with water from the core outlet to provide recirculation water of 270 °C. Coolant flow through the pressure tubes is adjusted by manually operated flow control valves in the coolant supply lines of the individual pressure tubes. In addition, the overall core coolant flow can be adjusted within a certain range by control valves in the discharge lines of the reactor coolant pumps. The main technical data of RBMK-1000 reactors may be summarized as follows:

General design characteristics

Thermal capacity	3200 MW_{th}
Gross electrical capacity	1000 MW_e
Net electrical capacity	925 MW_e
Total reactor coolant flow rate	37 600 t/h
Operating pressure	7 MPa
Core outlet temperature	284 °C
Core inlet temperature	270 °C
Total steam flow rate	5780 t/h
Feedwater inlet temperature to steam separators	165 °C

Core description

Height/diameter	7 m/11.8 m
Number of fuel channels	1661
Lattice spacing	250 mm × 250 mm
Moderator material	graphite
Moderator mass	1700 t
Maximum allowable moderator temperature	750 °C
Reflector dimensions	
Top and bottom	0.5 m
Sides	0.88 m

Fuel

Number of fuel elements per fuel channel	2, connected in series
Number of fuel rods per fuel element	18
Fuel material	UO_2
Total uranium mass	190 t
Enrichment	2 wt % ^{235}U
Maximum fuel burnup	20 000 MW · d/t
Cladding material	Zr – 1 % Nb
Active core height	6.9 m
Average rod power	150 W/cm

Control rods

Number of control rod assemblies	211
Cladding material	aluminum
Neutron absorption material	B_4C
Control rod spacing	500 mm × 500 mm
Control rod travel	6.25 m
Insertion velocity	
Control mode	0.20 m/s
Shutdown mode	0.40 m/s

Fuel channel

Number	1661
Length of pressure tube	
Total	22 m
Below core zone	5 m (stainless steel)
Core zone	8 m (Zr – 2.5 % Nb)
Above core zone	9 m (stainless steel)
Pressure tube outer diameter	88 mm
Pressure tube wall thickness	4 mm
Average tube power	1.89 MW
Maximum tube power	3.25 MW

Reactor coolant recirculation system

Independent flow loops	2
Steam separators	4 total, 2 per loop
Pumps	8 total (6 operating, 2 standby)
Pump dynamic head	1.96 MPa
Pump capacity	5500 – 12 000 m^3/h

Steam secondary system

Steam collector	primary system steam separators
Number of steam separators	4 total, 2 per loop
Steam separator	
Inner diameter × length	2.6 m × 30.984 m
Power generation	1000 MW_e
Turbine generators	2500 MW_e each
Feedwater inlet temperature to steam separators	165 °C

3.3.2. Reactor Core

The reactor core (Fig. 72) has the shape of a cylinder 11.8 m in diameter and 7 m high. It consists of graphite blocks with a cross-section of 250 mm × 250 mm and a height of 600 mm. The graphite blocks are stacked together. They form the moderator with a total mass of 1700 t. In addition, top and bottom graphite reflectors, each 500 mm high, and a radial graphite reflector 880 mm thick yield a total graphite stack 8 m in height and 13.8 m in diameter. The *pressure tubes* that contain the fuel elements and the *guide tubes* for the absorber rods are located in central vertical holes of the graphite blocks, thus forming a lattice with a pitch of 250 mm. Each graphite block has such a hole except for those blocks that form the radial reflector. A total of 1661 fuel channels and 211 channels for absorber rods can be found.

The *graphite stack* is surrounded by a leak-tight vessel (e) formed by a cylindrical steel shell together with the upper (b) and lower biological shields (h). A gas mixture of 80 % helium and 20 % nitrogen is fed into this vessel to provide a medium for heat transfer from the graphite stack to the coolant channels. The gas mixture is monitored for moisture to detect any leakage from the channels.

The *fuel* and *absorber rod channels* penetrate the upper and lower biological shields. They are connected above and below the reactor core to the corresponding loops of the reactor coolant system and the absorber rod cooling system, respectively. The drives of the absorber rods are located above the core and below the removable operating floor shield structure.

Fuel Channels. A simplified cross-section of a fuel channel is shown in Figure 73. The fuel channels contain *fuel assemblies* consisting of two fuel elements connected to a central support tube. The *fuel elements* are cylindrical 18-rod clusters each 3.644 m long. In the past, 1.8 and 2 % enriched uranium dioxide fuel in the form of sintered uranium dioxide pellets had been used. After the Chernobyl accident the enrichment of the fuel was increased to 2.4 %. The total mass of the fuel is 190 t. For neutron economy the fuel rod cladding tubes and the section of the pressure tubes in the core region are fabricated of zirconium – niobium alloys (Zr – 1 % Nb and Zr – 2.5 % Nb, respectively). Above and below

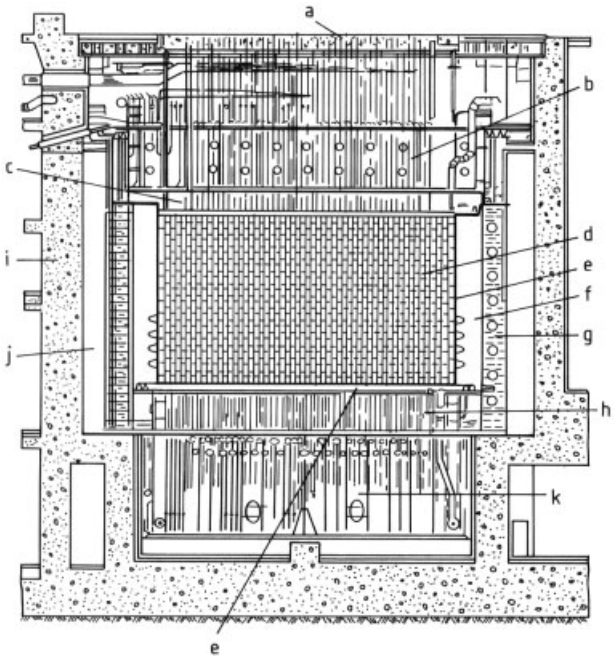

Figure 72. Cross-section of reactor cavity
a) Removable floor; b) Upper biological shield; c) Welded tube ducts; d) Graphite stack, helium – nitrogen filled; e) Reactor vessel; f) Nitrogen filled space; g) Annular water tank; h) Lower biological shield; i) Reactor vault; j) Sand; k) Reactor support

the core the pressure tubes consist of stainless steel that is diffusion welded to the zirconium section. The flow rate of the coolant in an individual fuel channel can be adjusted over a certain range by a manual control valve in the inlet section of the fuel channel.

The fuel channels are designed to allow *refueling during power operation*. For refueling the refueling machine is lowered over the outside of a fuel channel. After establishing a leaktight connection between fueling machine and fuel channel the plug on top of the fuel channel is unsealed. The plug and the spent fuel assembly are lifted from the fuel channel into the fueling machine. The fresh fuel assembly and the plug are then inserted into the fuel channel. Finally, the plug is sealed and the fueling machine is removed from the fuel channel. Cooling water is pumped from the fueling machine into the fuel channel during refueling to cool the discharged fuel assembly.

The 211 absorber rods that move in special channels in the graphite stack are designed for control of overall reactor power and power density, as well as for reactor shutdown. Cooling of these channels and the absorber rods is provided by the control rod cooling system.

3.3.3. Reactor Physics

The physics of the RBMK reactor is characterized by its graphite moderator, the large size of the reactor core, and the large core load of enriched uranium fuel. In contrast to PWRs and BWRs in which coolant and moderator are indentical, the RBMK reactor uses water as coolant and graphite as moderator. This has an important effect on the neutron characteristics of the RBMK reactor. Because of the large graphite – fuel ratio, the graphite moderator produces a sufficient amount of thermal neutrons under most operating conditions. In such a design the neutron moderator function of the coolant water is small, whereas its neutron absorber function plays an important role. Therefore, if the coolant void fraction increases, the number of neutrons absorbed by the water decreases and thus the number of neutrons available for fission increases, resulting in an increase of reactor power.

The dependence of the change of reactivity on a change of coolant void fraction is characterized by the *coolant void coefficient*. The sign and magnitude of this coefficient are functions of control rod position, fuel enrichment, fuel burnup, and supplemental absorber loading. For the RBMK reactor the coolant void coefficient is positive under most operating conditions; it increases with increasing fuel burnup or when the absorber rods are withdrawn from the core. Two other important reactivity feedback coefficients are the *graphite temperature coefficient* and the *fuel temperature coefficient*. The graphite temperature coefficient and the fuel temperature coefficient (Doppler coefficient) characterize the dependence of the reactivity on the temperature of the graphite stack and the fuel temperature, respectively. The graphite temperature coefficient is also positive. The fuel temperature coefficient is the only coefficient that is always negative.

For RBMK-type reactors, incidents or accidents that lead to an increase in coolant void fraction may result in an uncontrolled power increase, if they are not limited in time by the control system or reactor shutdown, respectively. Therefore the design must ensure that the reactor shutdown system is actuated and becomes effective in a sufficiently short time under all circumstances. Examples of incidents or accidents leading to an increase of the coolant void fraction are loss-of-coolant accidents, depressurization of the reactor coolant system, flow reduction in the reactor coolant system, or an increase of core inlet temperature.

The coolant void fraction depends on the reactor power. It is small at low power and increases to 14.5 % at full power. Thus, the total possible amount of changes in coolant void fraction and neutron flux is larger at low power compared to full power. In addition the fuel temperature is low at low reactor power; that is, the stabilizing effect of the fuel temperature (negative temperature coefficient) is less effective than at full power. As a consequence, the RBMK reactor is more sensitive to power transients at low reactor power than at full power. To prevent unacceptable power transients at low reactor power the procedures require immediate manual reactor shutdown by the operators if reactor power drops below 20 %.

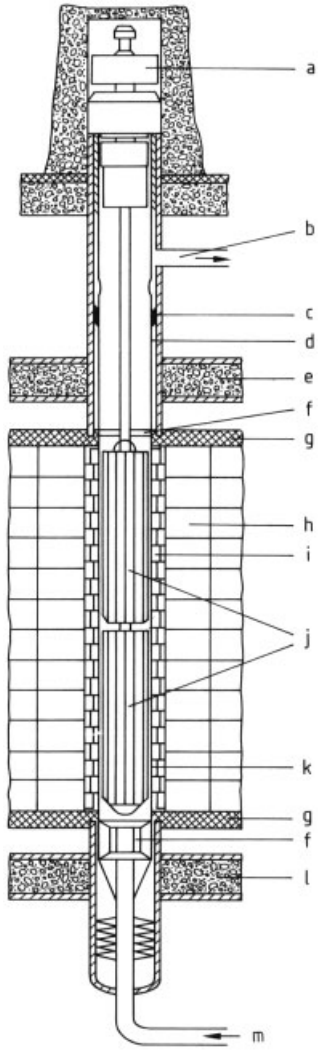

Figure 73. Schematic of a fuel channel
a) Pressure tube sealing plug; b) Steam/water outlet; c) Weld; d) Pressure tube (stainless steel); e) Upper biological shield; f) Zirconium – stainless steel transition; g) Thermal shield; h) Graphite block; i) Graphite ring; j) Fuel assembly; k) Pressure tube (Zr – 2.5 % Nb); l) Lower biological shield; m) Coolant inlet

3.3.4. Instrumentation and Control

The reactor is widely instrumented to guarantee economic and safe operation. The monitoring and control system is composed of two subsystems: the control and protection system and the reactor process monitoring system (SKALA).

The *reactor instrumentation* consists of:

1) Ionization chambers in the reflector region for overall power control;
2) Iission chambers in the reflector and in the core region for reactor start-up and local automatic control;
3) β - Emission sensors in the core region for measurement of axial and radial neutron flux distribution;
4) Thermocouples for monitoring the temperature of the graphite stack;
5) γ-Spectrometer probes for measuring the activity of the steam – water mixture in the core outlet pipes (detection of fuel cladding failure);
6) Measurement of relative humidity and temperature of the helium – nitrogen mixture in the reactor vessel (monitoring the integrity of the fuel channels); and
7) Measurement of the coolant flow rate in the process and control channels.

Control and Protection System. Regulation of reactor power and power density, as well as emergency protection, is performed by the control and protection system. Two independent sets of controllers are used for power regulation of the reactor: *overall power control* and *local automatic power control*. The overall power controller has three modes of operation: manual, automatic, and emergency control. The local power controller operates in two modes: automatic and emergency control.

Power control and emergency shutdown of the reactor are provided by 211 control rods (absorbing rods), which are grouped as follows:

115 manual control rods
12 control rods for overall power control
12 control rods for local automatic power control
24 emergency control rods
24 local emergency control rods
24 short absorbing rods.

Overall Power Control. The overall power controller consists of three identical regulators. Each regulator operates with a set of four control rods that are distributed symmetrically over the reactor core. One of these regulators provides power control for low power (0.5 – 10 %); the remaining two, for the range of 5 – 100 % of full power (one regulator in operation and one on standby).

Local Automatic Power Control. Local automatic power control is provided by 12 control rods with independent regulators. Local power control is the preferred system for power regulation in the range of 10 – 100 % of full power because of its ability to smooth out radial – azimuthal power distortions during transitional states.

Manual Power Control. Manual control of reactor power and control of excess reactivity can be performed by four groups of control rods (one peripheral group plus three centrally located, regularly intermixed lattices of control rods).

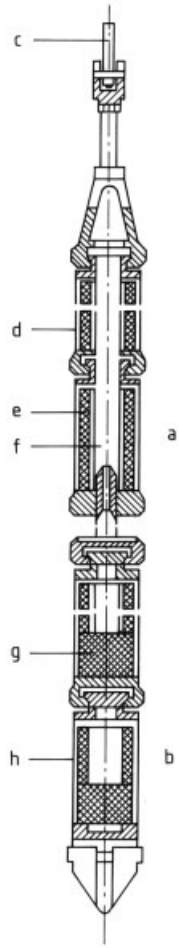

Figure 74. Control rod
a) Absorber section; b) Displacer section; c) Belt; d) Aluminum alloy cladding; e) Boron carbide; f) Void for coolant passage; g) Graphite; h) Aluminum alloy cladding

Control Rods. The control rods can be subdivided into 187 control rods of full length and 24 short absorbing rods of half length. The short absorbing rods are used only for the control of the axial power distribution in the lower part of the reactor core. The control rods (Fig. 74) are composed of six absorbing sections (a), each about 1 m long and containing boron carbide as absorbing material enclosed in sealed annular aluminum alloy elements (short absorbing rods contain three absorbing sections only). Except for the rods for overall power control, all control rods are equipped with five displacer sections (b) of 1-m length each at the bottom (at the top for short absorbing rods). The displacers are filled with graphite to improve neutron efficiency in the reactor when the control rods are partially withdrawn from the core (the strongly neutron absorbing water in the control rod channels is then replaced by the graphite of the displacer elements). If a control rod is fully withdrawn from the reactor core, a core section of 1 m on either end of the displacer element is filled with water (Fig. 75). Thus, when this control rod is inserted, the reactivity increases in the bottom part of the core during the initial phase of rod movement.

The control rods are operated by a d.c. motor drive mechanism via a belt cable. The insertion – withdrawal rate in the control mode is 0.2 m/s. In case of an emergency shutdown the control rods are inserted at a maximum rate of 0.4 m/s. The speed is limited by the flow resistance of the water in the control rod channels. Thus, the insertion time of an absorber rod, which has been fully withdrawn from the core, is about 15 s. In addition, a few seconds are required for an absorber rod that has been fully withdrawn from the core to become effective. Therefore, under all operating conditions a minimum number of absorber rods must be partially inserted in the reactor core to provide immediate reduction of neutron flux in an emergency shutdown.

The five levels of emergency shutdown, depending on the nature of the emergency, are complete reactor shutdown, partial reactor shutdown, and preventive controlled power reduction by 20, 50, and 60 %.

In case of *complete emergency shutdown*, all control rods except the short ones are automatically inserted in the reactor core at maximum speed. Examples of situations actuating complete reactor shutdown are

- Power overshoot of 10 % of full power,
- Reduction of the reactor period to 10 s,
- Reduction or increase of the water level in the steam separators,
- Rressure excess in the accident localization compartments, or
- Trip of both turbine generators or of the last operating turbine generator.

In the event of a local power overshoot, withdrawal of the local regulating rod is stopped. In a local emergency power overshoot the two local emergency rods are lowered into the core until the overshoot clears. In addition, automatic reduction of overall power is initiated.

Reactor Process Monitoring System. The reactor process monitoring system provides information on the current status of parameters important for reactor operation and safety. This information includes data from each channel (e.g., flow rate, type of cell charge, rod position), channel failure detection, physical power monitoring system, fuel cladding failure detection, and graphite stack temperature monitoring.

3.3.5. Safety Systems

Like other nuclear power plants, RBMK-type reactors are equipped with extensive safety systems to cope with incidents and accidents. The extent of safety systems installed in a reactor depends on the set of design basis accidents and transients, considered during the design of the plant, that must be handled successfully by the safety systems without damage to the fuel or the environment. Examples of accidents and transients considered during the design of second-generation RBMK-1000 reactors are:

1) Rupture of the largest pipe in the reactor coolant recirculation loops (i.e., the 900-mm-diameter discharge header of the reactor coolant pumps),
2) Rupture of one pressure tube in the graphite stack,
3) Rupture of a reactor outlet pipe,
4) Reactivity transients,
5) Trip of reactor coolant pumps,
6) Turbine trip,

7) Complete loss of normal in-house power supply, and
8) Internal flooding or fire.

Vital safety systems for the RBMK-1000 reactor are the emergency core cooling system, the accident localization system, and the emergency power supply. Safety systems must perform their tasks even if one component fails in the course of an incident or accident. Therefore, safety systems usually consist of three independent trains, each with 50 % of the capacity required for the most demanding incident or accident. The main characteristics of the safety systems of the second generation RBMK-1000 reactors are as follows:

Emergency core cooling system

Number of subsystems	3 (independent)
Capacity	3×50 %
Accumulators	
Number	2 subsystems, 6 tanks each
Capacity	2×50 %
Total volume	25 m^3 per tank
Water volume	13 m^3 per tank
Pressure	10 MPa
Safety injection pumps (damaged half of reactor)	
Number	3 subsystems, 2 parallel pumps each
Capacity	3×500 t/h (3×50 %)
Safety injection pumps (undamaged half of reactor)	
Number	3 subsystems, 1 pump each
Capacity	3×250 t/h (3×50 %)

Accident localization system

Type	pressure supression system for reactor cavity, inlet piping, and main coolant pumps
Design pressure	0.45 and 0.18 MPa, respectively
Number of bubbler ponds	2
Total water volume	3200 m^3
Bubbler pond cooling system	3 pumps (3×50 %)

Emergency power supply

Number of subsystems	3 (independent)
Capacity	3×50 %
Emergency diesel generators	3 (5.5 MW each)

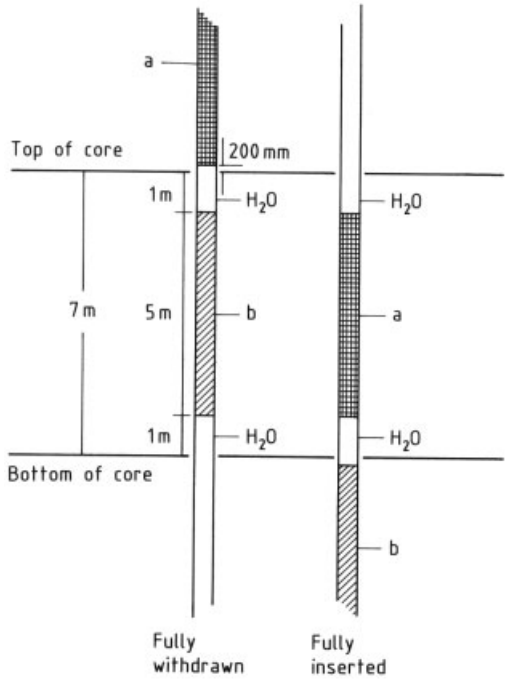

Figure 75. Schematic of fully withdrawn and fully inserted control rod
a) Boron carbide absorber; b) Graphite displacer

3.3.5.1. Emergency Core Cooling System

The emergency core cooling system is designed to provide cooling to the reactor core and to remove the decay heat in the event of a loss-of-coolant accident. In more recent plants such as Chernobyl units 3 and 4, the emergency core cooling system has been designed to cope with the rupture of the largest pipe in the reactor coolant recirculation loops, i.e., the 900-mm-diameter discharge header of the main coolant pumps. (Former units were designed for the rupture of a 300-mm-diameter subheader.) The emergency core cooling system has three independent trains, each with 50 % of the capacity required (Fig. 76). The three trains of the emergency core cooling system are subdivided into (1) a fast-acting section and (2) a section providing prolonged water supply to the reactor core.

Fast-Acting Section. The fast-acting section supplies water to the damaged half of the reactor core in the first phase of a loss-of-coolant accident (3×50 % of the capacity required), i.e.,

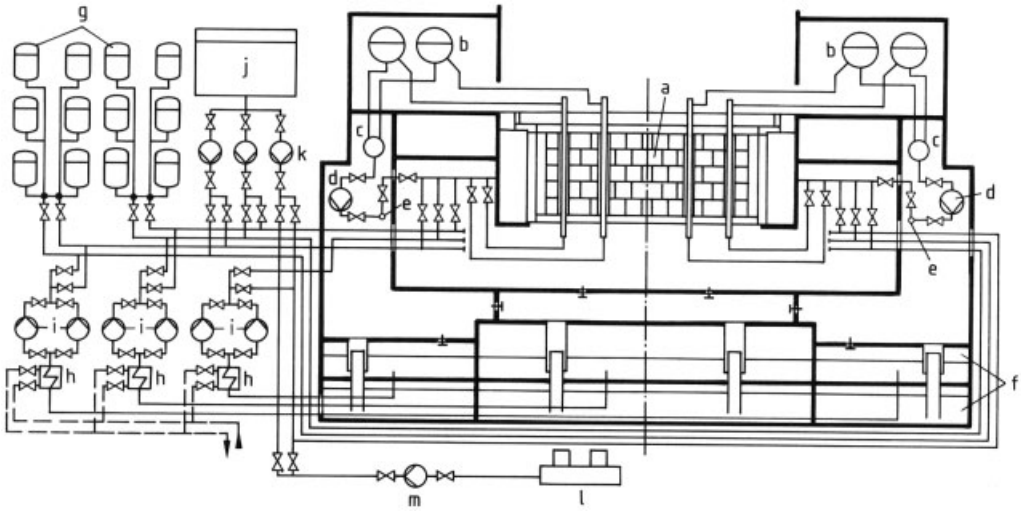

Figure 76. Schematic of emergency core cooling system
a) Reactor; b) Steam separators; c) Reactor coolant pump suction headers; d) Reactor coolant pump; e) Reactor coolant pump discharge header; f) Suppression pool; g) Accumulator; h) Decay power removal cooler; i) Emergency core cooling pump (damaged half of reactor core); j) Condensate storage tank; k) Emergency core cooling pump (undamaged half of reactor core); l) Deaerator; m) Feedwater pump

to the half of the reactor core belonging to the reactor coolant loop that is affected by the loss of coolant. Two trains consist of six accumulators, each filled with water and nitrogen. The nitrogen at a pressure of 10 MPa provides the driving force to inject the water into the damaged half of the core. The total water volume of the accumulators of one train is about 80 m^3. The accumulator systems supply water for at least 100 s during the accident. The third train of the fast-acting section uses the motor operated feedwater pump to inject water from the feedwater system into the damaged half of the core. In case of a loss-of-coolant accident along with a loss of in-house power supply, the feedwater pump is powered by the turbine generators during their rundown for about 45 – 50 s. All three trains of the fast-acting section are connected to the fuel channels of the reactor core via pipes and headers. They are operated by fast-acting valves and automatically switched to the damaged half of the reactor core.

Prolonged Section. The prolonged section of the emergency core cooling system provides cooling to both the damaged and the undamaged halves of the reactor core. Each of the three independent trains consists of two parallel pumps for the damaged half of the core and one pump for the undamaged half. Again, each train provides 50 % of the water required in the most severe design basis loss-of-coolant accident. Immediately after operation of the fast-acting part of the system, the two subsections of the prolonged part are automatically switched to the damaged and undamaged halves of the reactor core, respectively.

3.3.5.2. Accident Localization System

To limit radioactive releases to the environment in a loss-of-coolant accident, parts of the reactor coolant loops of RBMK-type reactors are enclosed by a set of sealed compartments (accident localization system, Fig. 77). The accident localization system consists of:

1) Two compartments, each enclosing the suction and discharge headers of the coolant pumps of one reactor coolant loop;
2) One compartment enclosing the reactor coolant pump discharge subheaders and the piping between the subheaders and the fuel channel inlets;
3) The steam distribution corridor, and
4) The pressure suppression pool (g) at the bottom part of the reactor building with two elevations and a total water volume of 3200 m^3.

Piping not enclosed by sealed compartments includes

1) the core outlet piping between upper biological shield and steam separators;
2) the steam separators;
3) the downcomers from the steam separators to the reactor coolant pump suction headers; and
4) the main steam and feedwater lines.

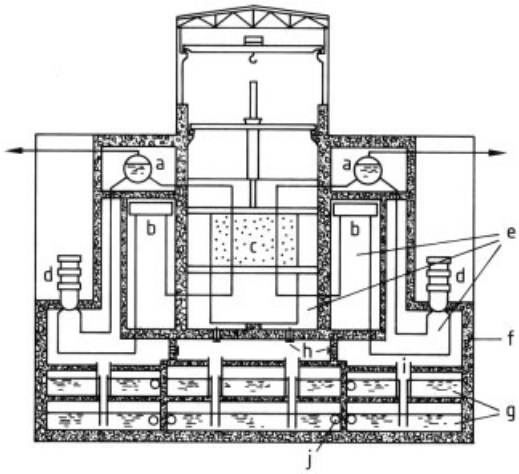

Figure 77. Schematic of the accident localization system a) Steam separators; b) Reactor coolant pump discharge header; c) Reactor; d) Reactor coolant pump; e) Sealed compartment of accident localization system; f) Concrete structure; g) Suppression pool; h) Check valves; i) Double-stage relief channel; j) Overpressure protection relief pipe

The fuel channels are not enclosed by the accident localization system itself but by the *sealed reactor vessel*, which is not connected to the compartments of the accident localization system.

The compartments of the accident localization system are interconnected by check valves. The two compartments enclosing the suction and discharge headers of the reactor coolant pumps, as well as the distribution corridor, are connected to the two levels of the *suppression pool* by double-stage relief channels. If a loss-of-coolant accident occurs in the part of the reactor coolant loops that is enclosed by the accident localization system, steam released to the sealed compartments is suppressed by the water in the suppression pool via the check valves and the relief channels. Thus, overpressure in the sealed compartments and radioactive releases to the environment are prevented.

In addition, the suppression pool serves as:

1) The water supply for the emergency core cooling system,
2) The heat sink for the overpressure protection of the reactor vessel in the event of a fuel channel rupture inside the graphite stack, and
3) The heat sink for the overpressure protection of the reactor coolant circuit.

Overpressure in the sealed reactor vessel could lift the upper biological shield and lead to massive failure of the fuel channels welded to the upper shield. Therefore, the reactor vessel is connected by large-diameter pipes to special sections of the suppression pool. If a leakage or rupture of a fuel channel occurs, the steam released to the reactor vessel is discharged via these pipes to the water in the suppression pool and overpressurization of the reactor vessel is prevented. For most RBMK-1000 units, the reactor vessel discharge pipes are designed to handle the rupture of one fuel channel. In the most recent units the discharge capacity has been increased to prevent overpressurization of the reactor vessel in the event of rupture of three channels.

The *reactor coolant overpressure protection system* is designed to ensure that the permissible pressure in the reactor coolant system is not exceeded, especially in the case of the unavailability of the main condenser. The system consists of eight main steam safety valves with a total capacity of 5800 t/h, the nominal steam output of the reactor at full power. The main steam safety valves discharge steam from the main steam lines to the water in the suppression pool via piping with a large number of submersible nozzles. In addition to the automatic mode, the main steam safety valves can be opened manually from the control room.

A system with three 50 % trains is used to remove heat discharged into the water of the suppression pool.

3.3.5.3. Emergency Power Supply

In a typical RBMK unit, *normal in-house power* is provided by (1) the two turbine generators (d), (2) the main grid, (3) the auxiliary grid, and (4) a cross connection to the other units at the same

site (Fig. 78). The normal in-house power supply provides electrical power to all of the plant equipment.

Safety-related equipment (i.e., equipment necessary for emergencies) is powered by the *emergency power supply*. The emergency power supply has three independent trains (3×50 %). Emergency power is provided by the normal in-house power supply if the latter is available. In case of a complete loss of normal in-house power, three emergency power diesel generators (6 kV, 5500 kW each) provide electrical power to the three trains of the emergency power supply.

Part of the safety equipment does not tolerate any interruption of power. These systems (e.g., the isolation system of the accident localization system, fast-acting valves in the emergency core cooling lines, the reactor protection system, instrumentation, and control of safety systems) are powered by 0.4-kV busbars supplied by batteries via static converters. Some safety equipment must be provided with electrical energy immediately after the occurrence of specific design basis accidents. Because the diesel generators need time to start and to power the emergency busbars in the event of loss of normal in-house power, this equipment is powered by the turbine generators using their inertial energy during run down.

3.3.5.4. Improvements after the Chernobyl Accident

Following the April 26, 1989 accident at the Chernobyl nuclear power station, RBMK-type reactors have been and will continue to be improved substantially to decrease their sensitivity to the core steam void fraction and increase the effectiveness of safety provisions. Important improvements to decrease the influence of the void coefficient are:

1) Greater enrichment of the fuel;
2) Higher number of fixed absorber rods in the reactor core;
3) Increasing the effectiveness of the existing shutdown system by
   - Increasing the number of control rods that must be partially inserted in the reactor core under all operating conditions,
   - Preventing complete withdrawal of control rods from the reactor core by interlocks (i.e., control rods are inserted at least 1.2 m into the reactor core to avoid a reactivity increase in the bottom part of the reactor core during the first phase of control rod insertion), and
   - Preventing power operation at power levels below 700 MW_e by an automatic trip signal;
4) Installation of an additional fast-acting shutdown system,
5) Increased capacity of the overpressure protection of the sealed reactor vault;
6) Improved information and alarms available to operators to follow the status of important operational and safety-related parameters and to recognize any deviation from authorized domains for operation of the reactor core.

3.3.6. The Chernobyl Accident

The RBMK reactors in the former Soviet Union had a good operating history until 1986. More than 100 years of operation of this reactor type had demonstrated a high degree of reliability. On April 26, 1986, Chernobyl 4 suffered the most serious accident in the peaceful use of nuclear energy. The disaster was caused by violation of procedures, combined with specific characteristics of the RBMK reactor at low power.

The Chernobyl nuclear power station is located about 120 km northwest of Kiev on the Pripyat River. By April 1986, four RBMK units were in commercial operation at the site. Two additional units were under construction. On April 25, 1986, unit 4 of the Chernobyl nuclear power station was to be shut down for routine maintenance. Just before shutdown a special test was scheduled to demonstrate improvements in the capacity of the turbine generators that support essential systems during the first phase of a complete loss of the normal in-house power supply. Key steps in the test procedure were (1) power reduction to 700 – 1000 MW_{th}, (2) tripping the last operating turbine, and (3) measuring the electrical power provided by the turbine generator during its inertial rundown, using four of the eight main coolant pumps as electrical load. At 1: 30 a.m. operators began to reduce the power of unit 4. At noon a power level of

50 % (1600 MW_{th}) was reached, and one of the two turbine generators was disconnected from the grid. At the command of the Kiev load dispatcher, further power reduction was delayed until 11: 00 p.m., which lead to a substantial buildup of the neutron poison xenon-135 in the reactor core. At 11: 00 p.m. the operators continued to decrease reactor power to the test target level of 700 – 1000 MW_{th}. Because of an operator error during switchover of the reactor power control system, reactor power dropped to $< 30\,MW_{th}$ and could not be restored to the level required by the test procedure. Nevertheless the operators decided to keep the plant running to perform the test.

In violation of instructions in the procedures, the power was raised to 200 MW_{th}, far below the level required for stable operation of the reactor. After switching on the two standby main coolant pumps as required by the test procedure, the operators blocked several trip signals to prevent automatic shutdown of the reactor before or during the test–again a violation of operating procedures.

Immediately before the start of the test the status of the reactor was characterized by:

- Low reactor power and high coolant flow rate, resulting in low coolant void fraction and high coolant inlet temperature;
- Blocking of essential trip signals; and
- Complete withdrawal of all control rods from the reactor core (except for the 12 control rods for automatic power control).

In summary, the potential for void formation and power increase was large and safety provisions were substantialy reduced; the reactor thus was highly prone to a power excursion.

When the test began at 1: 23.04 a. m. on April 26, 1986, the last operating turbine generator was shut down as required by the test procedure. The reactor continued operation because of the bypassed trip signals. The four main coolant pumps that were connected to the turbine generator as an electrical load for the test coasted down together with the turbine generator, thus lowering the core coolant flow. Because of the reduction in the core flow rate, the core void fraction and the reactor power began to increase.

After the test had been in progress for about 36 s, the shift supervisor ordered a manual scram because the electrical equipment to be tested was operating successfully. At that time a slight power increase was observed, which was actually the starting point of a fast power ramp. Since almost all of the control rods had been withdrawn completely from the core, several seconds were required for the control rods to move into the core and become effective. This time was too long to prevent a large power excursion. The power excursion destroyed the reactor, and large amounts of reactor core and moderator material were ejected from the reactor building. The hot material ejected from the reactor core started more than 30 fires that threatened the three other units of the power station. Fire brigades succeeded in extinguishing these fires after about four hours.

To decrease the release of radioactive fission products from the damaged reactor core to the environment and to extinguish the burning graphite moderator, the reactor core was covered with a variety of materials (total mass ca. 5000 t) by helicopters. This work was finished on May 6, 1986. By October 1986, a tight concrete structure had been built around the damaged reactor to prevent further release of radioactive fission products from the reactor to the environment.

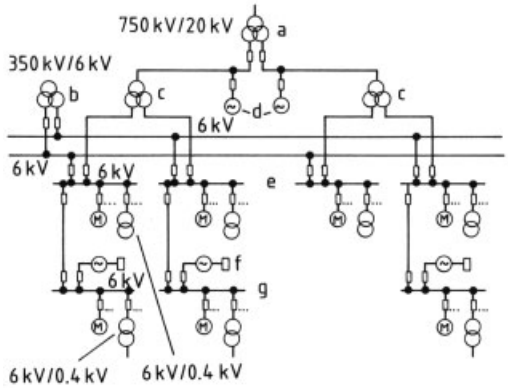

Figure 78. Schematic of in-house and emergency power supply
a) Main transformer; b) Auxiliary transformer; c) In-house service transformer; d) Turbine generator; e) In-house service power busbar; f) Emergency diesel generator; g) Emergency power supply busbar

During and after the accident, a total of ca. 3 % of the radioactive inventory of the reactor core was released to the environment; more than half of this was deposited in the area surrounding Chernobyl and in other parts of the former Soviet Union. Because of the widespread radi-

ological consequences of the accident, an area up to 30 km from the power station had to be evacuated, part of which will not be habitable for many years. The economic consequences of the accident for the former Soviet Union have been very high. In the meantime, the former Soviet government has decided to shut down units 1 to 3 of the Chernobyl nuclear power station, which are still being operated.

3.4. Heavy-Water Reactors

All nuclear fission reactors that use heavy water (D_2O) to moderate the fast neutrons released by nuclear fission are called heavy-water reactors (HWR). Because heavy-water reactors can be operated with natural uranium as fuel, they are of interest to countries that have economically recoverable natural uranium reserves, despite the high price of heavy water in 1996 ($ 280/kg).

3.4.1. History and Development

The suitability of heavy water as a moderator for natural uranium-fueled reactors was indicated by Werner Heisenberg in 1939, shortly after the discovery of nuclear fission [113]. Although experiments with heavy water and natural uranium were conducted in Germany during World War II, the world's first experimental heavy-water reactor (the so-called CP-3) was built in 1944 in the United States. In the ensuing period a number of *experimental* and *plutonium-producing reactors* were built that used heavy water as moderator. The development of heavy-water reactors for *power generation* began in Canada in 1950, in Germany in 1956, and in other countries in the years that followed. Several designs were developed in the beginning whereas most of these prototype reactors are already shut down or decommissioned.

Prototype or Test Reactors. *Pressure Tube Reactors Using Water as Coolant and Heavy Water as Moderator.* Examples: EL-4 in France, Lucens in Switzerland, Niederaichbach in Germany.
Pressure Vessel Reactors Using Carbon Dioxide as Coolant and Heavy Water as Moderator. Example: A 1-Jaslovské Bohunice in Slovakia.
Boiling-Water Reactors Using Carbon Dioxide as Coolant and Heavy Water as Moderator. Examples: Winfrith in the United Kingdom, Gentilly-1 in Canada, CIRENE in Italy, Fugen in Japan.
Pressure Tube Reactors Using Organic Compounds (Santowax, Therphenyl) as Coolants and Heavy Water as Moderator. Example: WR-1 in Canada.

Power Reactors. *Pressure Tube Reactors Using Heavy Water as Coolant and Moderator.* Examples: CANDU reactors in Canada, India, Pakistan, Korea, and Argentina. These reactors are being further developed as advanced PHWR.
Pressure Vessel Reactors Using Heavy Water as Coolant and Moderator. Examples: MZFR (multipurpose research reactor) in Germany, Atucha 1 and 2 in Argentina.

Nowadays, only pressure tube and pressure vessel reactors using heavy water as moderator and coolant are in commercial operation.

Because of the relatively high investment costs, only a few heavy-water reactors have been able to compete with light-water reactors on the world market (see Table 39). Some of the factors dictating the high specific investment costs are (1) the size of the reactor pressure vessel, or calandria vessel, (2) the necessary segregation of moderator and coolant, and (3) heavy-water conditioning, purification, and upgrading.

In contrast to light-water reactors, *refueling* in heavy-water reactors is performed as a rule on-line (i.e., the reactors can be operated at full power for extended periods). Because the differences in neutron physical properties between cold and hot operating conditions are small, the load can be changed rapidly as and when required. Furthermore, no special control features are required because heavy-water reactors are operated only with low excess reactivity.

Future Trends. The trends in future HWR design and technology include (1) *alternative fuel options* and (2) *improvement of plant design* [114].

Despite its good neutron economy the HWR operated with natural uranium achieves low burnup. Therefore, MOX fuel or slightly enriched uranium is used, which offers a higher burnup at nearly the same fabrication costs. There are plans to use the spent fuel of LWRs in HWRs in a tandem mode.

Table 39. Nuclear power plants with heavy-water reactors

	In operation		Under construction, on order, or planned		Summary	
Country	Number	MW_e	Number	MW_e	Number	MW_e
Argentina	2	935	1	692	3	1 627
India	8	1 395	4	808	12	2 203
Japan	1	165			1	165
Canada	21	14 907			21	14 907
South Korea	1	629	3	1950	4	2 579
Pakistan	1	125			1	125
Romania			2	1 300	2	1 300
China			2	1 480	2	1 480
Total	34	18 156	12	6 230	46	24 386

Darlington Nuclear Power Station in Canada has the largest HWR units ($4 \times 935\ MW_e$ gross) in operation, unit No. 2 started commercial operation in 1992. Further development of CANDU reactors covers ratings up to $1300\ MW_e$, including features for life extension, rapid load-following capabilities, and shorter construction periods [114]. The CANDU 3 ($450\ MW_e$ gross) has been developed as an alternative to coal-fired stations for small electrical grids. The latest design, the CANDU-80 with a power output of ca. $100\ MW_e$, is suited to a variety of electricity and cogeneration applications. In India, units with $500\text{-}MW_e$ gross output are being designed.

In Germany the Siemens pressurized vessel heavy-water reactors (PHWRs) originate from the multipurpose research reactor (MZFR), which operated successfully for about 18 years. This HWR line was continued in Argentina with the Atucha 1 and Atucha 2 reactors. Improvements in specific areas include reduction of the tritium content in the coolant, control rod drives, spent fuel storage capacity, and alternative fuel cycles. A $380\text{-}MW_e$ advanced HWR of this line (ARGOS) has been drafted in Argentina.

In Japan the advanced thermal reactor (ATR) was originally developed for the utilization of plutonium, recovered uranium, and depleted uranium. A prototype reactor Fugen ($165\ MW_e$), a heavy-water-moderated boiling-water reactor, has been in operation since 1979. An ATR demonstration plant (Ohma Power Station) was scheduled to go into commercial operation in Japan in March 1999. [115].

3.4.2. Design Features

Arrangement of Fuel Rods in the Reactor. Heavy-water moderated reactors require a $D_2O - UO_2$ volume ratio of 16 – 18. To maintain this ratio, bundles of 19 – 37 relatively tightly packed fuel rods are arranged with ca. 25-cm spacing between the bundles. This produces two clearly defined zones:

1) The fuel bundles themselves are accommodated in coolant channels and are swept by *coolant* to remove heat generated by nuclear fission.
2) The zone outside the coolant channels accommodates the *moderator*.

(In light-water reactors the fuel elements are arranged as a heterogenic core. Light water is used as moderator and coolant which are not separated from each other.)

Design. Heavy-water reactors are either of the pressure vessel or pressure tube type. In a *pressure vessel reactor* the coolant channels and the moderator zone are accommodated in a common pressure vessel. Coolant and moderator are kept at the same pressure–but the moderator at lower temperatures–through pressure equalization openings. The coolant channels are therefore not pressure retaining and can have appropriately thin walls.

In a *pressure tube reactor* the moderator zone and the coolant are clearly separated physically. In the moderator zone, pressure and temperature are lower than in the coolant. The pressure tubes are exposed to the full primary system

Table 40. Characteristic data for heavy-water reactors

	Atucha 1	Atucha 2	Gentilly 2	Darlington	Fugen	ATR
Thermal power, MW_e	1179	2160	2180	2774	557	1930
Net power output, MW_e	345	692	638	881	165	606
Fuel	UO_2	UO_2	UO_2	UO_2	UO_2/PuO_2	MOX
Fuel load in core, t	38.8	85.1	87.5	119	34.2	90
Number of cooling channels	253	451	380	480	224	616
Average power rating, kW/L	13.74	14.1	11.2	18.5	11.7	14.7
Coolant loops/steam generators	2/2	2/2	2/4	4/4	2/0	2/0
Separate moderator loops	2	4	2	2	1	1
Operating pressure, MPa	11.3	11.5	11.0	10.6	6.8	6.9
Coolant temperature (inlet/outlet), °C	262/296	278/312	266/310	265/313	277/284	188.4/284.5
Coolant mass flow, t/h	22 100	37 100	28 000	43 750	7600	24 500
Live steam pressure, MPa	4.41	5.59	4.7	5.07	6.8	6.9
Live steam temperature, °C	254	271	260	264.7	283.5	284.5

pressure of ca. 100 MPa. The coolant channels and pressure tubes that serve as coolant-carrying and pressure-retaining parts, respectively, cover the entire length of the moderator zone. To keep heat losses from coolant to moderator low, the coolant channels and pressure tubes are also provided with thermal insulation. Some characteristic design data for heavy-water reactors are given in Table 40.

Coolant. Segregation of the coolant and moderator zones provides different coolant (gases, liquids) options, particularly with pressure tube reactors. In the past, numerous types of coolant have been used, which account for the large variety of heavy-water reactors built worldwide. The ensuing diversification of developments is, however, one of the reasons why heavy-water reactors have not been as successful as light-water reactors.

For heavy-water reactors fueled with natural uranium, only coolants with low neutron absorption coefficients (D_2O, CO_2, He) are suitable. In addition, coolants must be chemically compatible with the reactor materials and must be stable under irradiation and temperature loads, which further restricts the suitability of coolants. Switching to enriched uranium as fuel extends the spectrum of possible coolants, particularly compounds with a high hydrogen content such as light water or organic liquids.

Materials. To decrease the damage from neutron absorption, materials with the lowest possible neutron absorption coefficients must be used, especially in a natural uranium reactor. Zirconium alloys have become the most accepted standard material in the *reactor core*. The most important of these alloys are:

1) Zr – Sn alloys with small additions of Fe, Cr, and Ni (e.g., the standard alloys Zircaloy-2 and Zircaloy-4);
2) Binary Zr – Nb alloys (e.g., Zr – 2.5 % Nb); and
3) Other zirconium alloys (e.g., MgZr; ZrNb 3 Sn 1; ZrCrFe; ZrCuFe).

The alloy silver – indium – cadmium or hafnium is used as material for the *absorber rods*.

As in LWR nuclear power plants, high temperature-resistant fine-grained steels (e.g., 20 MnMoNi 55) are used for the structural parts and pressure-retaining components of the *coolant loops*; inner surfaces in contact with coolant or moderator are generally clad with austenitic steel to keep corrosion rates low.

3.4.3. Types of Heavy-Water Reactors

3.4.3.1. Heavy-Water-Cooled Pressure Tube Reactor

The Canadian heavy-water-cooled pressure tube reactor (CANDU), jointly developed in Canada by Atomic Energy of Canada Limited (AECL) and the nuclear industry, is today the leading heavy-water reactor. About 85 % of the 18 156 MW_e of all heavy-water reactors in operation, under construction, or planned are Canadian CANDU reactors (Table 41). Reactors with

Table 41. Countries with heavy-water reactors

Country	Plant	Type	In operation, MW_e	Under construction, MW_e	Planned, MW_e	Date of commercial operation
Argentina	Atucha 1	Siemens/PHWR	367			1974
	Atucha 2	Siemens/PHWR		745		1990
	Cordoba	CANDU	648			1984
Canada	Bruce 1	CANDU	904			1977
	Bruce 2	CANDU	904			1977
	Bruce 3	CANDU	904			1978
	Bruce 4	CANDU	904			1979
	Bruce 5	CANDU	915			1985
	Bruce 6	CANDU	915			1984
	Bruce 7	CANDU	915			1986
	Bruce 8	CANDU	915			1987
	Darlington 1	CANDU	935			1992
	Darlington 2	CANDU	935			1990
	Darlington 3	CANDU	935			1993
	Darlington 4	CANDU	935			1993
	Gentilly 2	CANDU	685			1983
	Pickering 1	CANDU	542			1971
	Pickering 2	CANDU	542			1971
	Pickering 3	CANDU	542			1972
	Pickering 4	CANDU	542			1973
	Pickering 5	CANDU	540			1983
	Pickering 6	CANDU	540			1984
	Pickering 7	CANDU	540			1985
	Pickering 8	CANDU	540			1986
	Point Lepreau	CANDU	680			1983
China	Quinshan 3	CANDU			740	2002
	Quinshan 4	CANDU			740	2003
India	Kaiga 1	CANDU		220		1998
	Kaiga 2	CANDU		220		1998
	Kakrapar 1	CANDU	220			1993
	Kakrapar 2	CANDU	220			1995
	Kalpakkam 1	CANDU	170			1984
	Kalpakkam 2	CANDU	170			1986
	Narora 1	CANDU	220			1991
	Narora 2	CANDU	220			1992
	Rajasthan 1	CANDU	100			1973
	Rajasthan 2	CANDU	200			1981
	Rajasthan 3	CANDU		220		1998
	Rajasthan 4	CANDU		220		1999
Japan	Fugen	D_2O-H_2O	165			1979
	ATR	D_2O-H_2O			606	1999
Pakistan	Kanupp	CANDU	137			1972
Republic of Korea	Wolsung 1	CANDU	679			1983
	Wolsung 2	CANDU		700		1997
	Wolsung 3	CANDU		700		1998
	Wolsung	CANDU		700		1999
Romania	Cernavoda 1	CANDU	700			1996
	Cernavoda 2	CANDU		700		2002
	Cernavoda 3	CANDU		700		
	Cernavoda 4	CANDU		700		
	Cernavoda 5	CANDU		700		

ratings up to 935 MW_e are under construction or in operation [116–120]. Most of the reactors built to date have ratings between 550 and 950 MW_e.

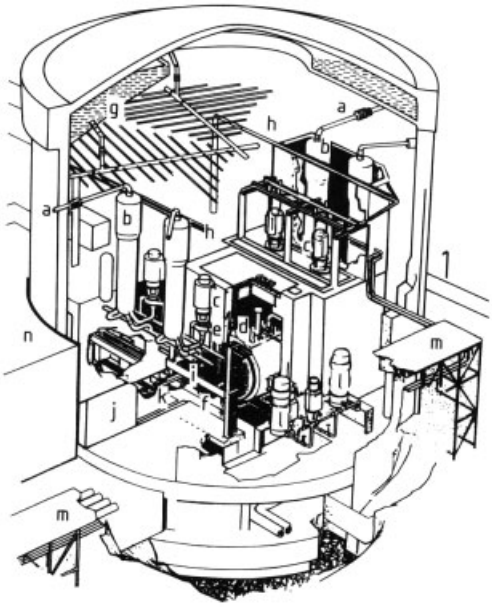

Figure 79. Reactor building of a pressure tube heavy-water reactor (CANDU)
a) Main steam supply piping; b) Steam generators; c) Main primary system pumps; d) Calandria assembly; e) Feeders; f) Fuel channel assembly; g) Dousing water supply; h) Crane rails; i) Fueling machine; j) Fueling machine door; k) Catenary; l) Moderator circulation system; m) Pipe bridge; n) Service building

Concept. The reactor is a large, horizontally oriented cylindrical tank, called a calandria, which contains the heavy-water moderator at low pressure and temperature. This tank is penetrated by a number of horizontal fuel channels that contain the fuel assemblies and are swept by the pressurized, high-temperature, heavy-water coolant. Steam generators and coolant pumps are located at each end of the reactor so that flow is in one direction through one-half of the fuel channels and in the opposite direction through the other half.

All primary and moderator system components are located in the containment, which basically forms an envelope around these nuclear components (Fig. 79). The containment system comprises a prestressed, posttensioned concrete structure with a plastic liner and energy sinks that consist of an automatically initiated dousing system and building air coolers.

The Reactor. The reactor (Fig. 80) consists of a cylindrical stainless steel *calandria structure* containing the heavy-water moderator, the reactivity control mechanisms, and 380 *fuel channel assemblies* (600 MW). Each of the fuel channel assemblies consists of a pressure tube (material Zr – 2.5 % Nb) and channel closures at both ends. The fuel channels contain the fuel and heavy-water coolant and pass through the calandria tubes. The gap between each fuel channel and calandria tube is filled with gas to provide thermal insulation. A cooling system dissipates the heat transferred to the moderator from the high-temperature fuel channels, as well as the heat generated in the heavy water itself through interaction with fission neutrons and γ radiation.

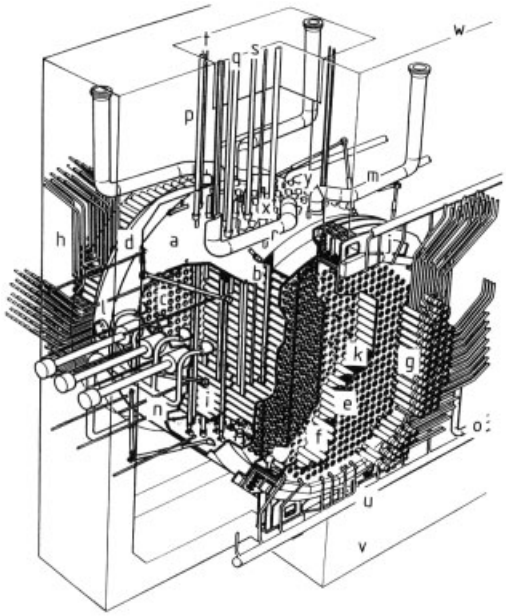

Figure 80. CANDU reactor assembly
a) Calandria; b) Calandria main shell; c) Calandria-side tube sheet; d) Calandria subshell; e) Fueling machine-side tube sheet; f) Lattice tubes; g) End fittings; h) Feeders; i) Calandria tubes; j) Shield tank solid shielding; k) Steel ball shielding (end shield); l) Manhole; m) Emergency discharge pipes; n) Moderator inlets; o) Moderator outlets; p) Shutoff unit; q) Adjuster unit; r) Vertical flux detector; s) Control absorber; t) Liquid zone control unit; u) End shield cooling piping; v) Shield tank; w) Shield tank extension; x) Rupture disk assembly; y) Moderator overflow

The calandria is supported by end shields that provide shielding between the reactor core and the fueling machine operating areas. The calandria and end shields are housed in the *concrete vault*, which is steel-lined and filled with light water that serves as a cooling medium and as biological shielding.

The reactor is fueled with natural uranium in the form of uranium dioxide pellets. Approximately 29 pellets, stacked end to end, are sealed in a zirconium alloy sheath to form a fuel element; 37 of these elements are assembled into a fuel bundle (length 495 mm). Each fuel channel accommodates 12 fuel bundles.

Neutron-absorbing systems, both liquid and solid, assist in controlling reactivity. Fast reactor shutdown is achieved by using shutoff rods or liquid poison injection into the moderator.

Reactor System. The primary coolant system circulates heavy water through the reactor fuel channels to remove the heat produced by fission of the uranium dioxide fuel. The heat is carried to the steam generators where it is transferred to the plant's steam – feedwater cycle. Figure 81 shows a simplified overall flow diagram.

Heat Transport System. The main components of the heat transport system are 380 reactor fuel channels, four vertical steam generators, four pumps, four reactor inlet and four reactor outlet headers, one pressurizer, and all the necessary interconnecting piping and valves. The headers, steam generators, and pumps are located above the reactor. At a pressure of 11 MPa the coolant enters the *coolant channels* at 267 °C and leaves at 310 °C. The total mass flow rate is 27 800 t/h. A common *pressurizer* controls the pressure in the heat transport circuits. The heavy water stored in the pressurizer is heated by electric immersion heaters to provide the proper pressure. The heavy water in the calandria is circulated through the moderator system for cooling, purification, and control of the substances used for reactivity adjustment. The inlet temperature to the calandria is controlled at 43 °C.

Auxiliary Systems. Several auxiliary systems, such as the shutdown cooling system, the purification systems, the liquid poisoning system, and the heavy-water collecting and sampling systems, are provided.

Special Features. Fueling is performed during power plant operation. The fuel bundles are pushed into the reactor channel by a remotely operated fueling machine. Spent fuel bundles are discharged simultaneously into another fueling machine at the opposite end of the reactor and then transferred to a water-filled storage bay in a separate building adjacent to the reactor building. Approximately 16 fuel bundles must be replaced each full-power day.

In operting multi–unit plants (e.g., Darlington) 4 CANDU 900 MW_e reactors are arranged in line. A large fueling duct is located under each reactor and runs the entire length of the station. This duct is used as a passage for two mobile fueling machines, which are mounted on trolleys and serve all four units. In addition, the duct serves as a pressure relief connection between each reactor building and a vacuum building. This is a 75-m-high cylindrical concrete structure that is maintained at subatmospheric pressure. It is isolated from the fueling duct and reactor vaults by large self-actuated pressure relief valves.

The reactor design is somewhat different in terms of station power and building – equipment arrangement. The power output was increased to 935 MW_e, and the calandria contains 480 fuel channels, each with 13 fuel bundles.

Design development of CANDU 9 is aimed at a single-unit adaptation of the 900-MW class.

3.4.3.2. Heavy-Water-Cooled Pressure Vessel Reactor

The heavy-water-cooled pressure vessel reactor was developed in Germany by Siemens [121–123]. It was the basis for the 50-MW_e multipurpose research reactor built at the Karlsruhe nuclear research center between 1961 and 1965. This reactor operated very successfully until 1984 when it was shut down for economic reasons (too low rating).

Between 1968 and 1974 the 367-MW_e nuclear power plant Atucha 1 was built in Argentina. The 745-MW_e Atucha 2, another nuclear plant in Argentina, has been under construction since 1980. The following description is based on the latter plant.

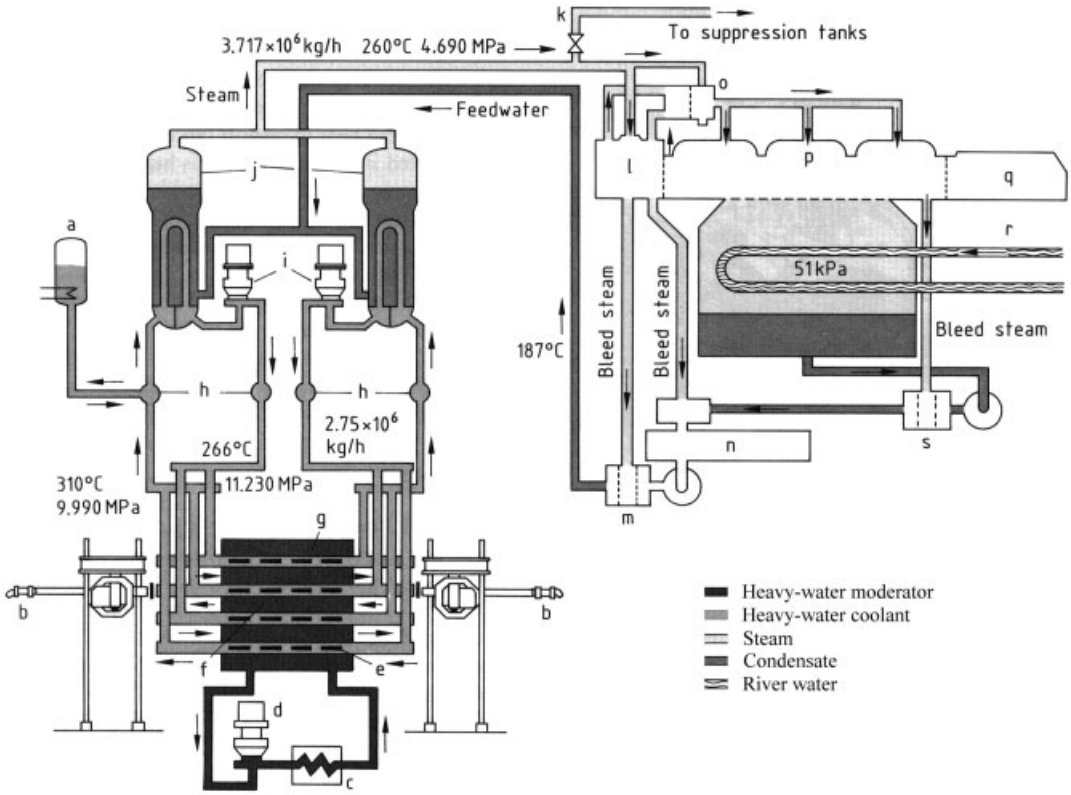

Figure 81. CANDU – simplified flow diagram
a) Pressurizer; b) Fueling machine; c) Moderator heat exchanger; d) Moderator pump; e) Fuel; f) Reactor; g) Calandria; h) Headers; i) Four primary pumps; j) Four steam generators; k) Safety valves; l) High-pressure turbine; m) 2×2 High-pressure heaters; n) Deaerator and storage tank; o) Two moisture separator reheaters; p) Low-pressure turbine; q) Generator; r) Cooling water from rivers; s) 2×3 Low-pressure heaters

Concept. The pressure vessel reactor has many common design features with the pressurized light-water reactor. These features include (1) the reactor pressure vessel; (2) the components of the reactor coolant loops; (3) the pressurizer system; (4) the spherical steel containment; (5) the secondary concrete shielding that forms the annulus building; (6) the safety systems; and (7) the instrumentation and control. The layout of the reactor building with leaktight containment, the concrete shield building including the valve compartment for external impact protection, and the annulus in between to accommodate safety-related systems and components are shown in Figure 82.

The Reactor. The *reactor pressure vessel* that forms the pressure boundary of the reactor core consists of the vessel body and the closure head. These two parts are connected by a flange and are made leaktight by a welded lip seal. The closure head accommodates the nozzles for the fuel assemblies, the control rod drive mechanisms, and core instrumentation (Fig. 83).

The *moderator tank* accommodates all core components and separates the moderator area from the coolant flow. The pressures of coolant and moderator are equalized through pressure equalization openings in the moderator tank head, so that no significant pressure differences act on the RPV internals.

The 451 vertical *coolant channels* (material Zry- 4) arranged in a trigonal lattice (lattice pitch 272 mm) pass through the moderator tank. Each coolant channel contains one 37-fuel-rod bundle with an active length of 5300 mm. Coolant enters the inlet nozzles of the pressure vessel, flows downward in the annulus between RPV wall and

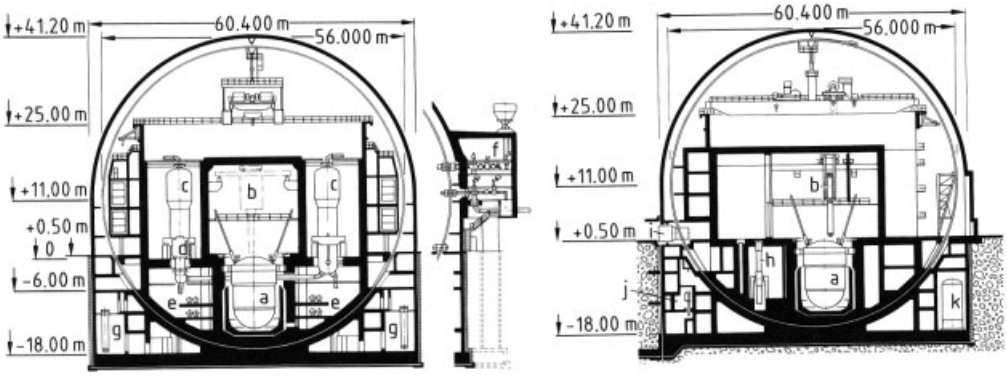

Figure 82. Reactor building of a pressurized vessel heavy water reactor (Siemens/PHWR)
a) Reactor pressure vessel; b) Refueling machine; c) Steam generator; d) Reactor coolant pump; e) Moderator cooler; f) Steam and feedwater valves; g) Component cooling heat exchanger; h) Tilter; i) Personnel air lock; j) Safety injection pump; k) Water storage tank

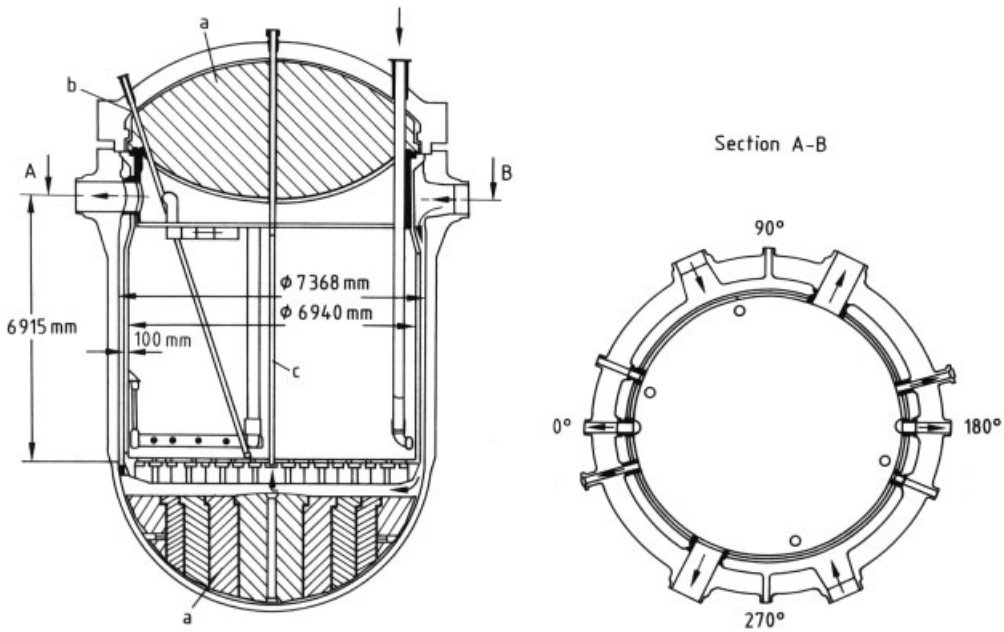

Figure 83. Reactor pressure vessel with internals (Siemens/PHWR)
a) Filler pieces; b) Control element guide tube; c) Coolant channel; d) Control element positions

moderator tank, and enters the coolant channels from below. It leaves the coolant channels above the moderator tank cover and is routed to the vessel outlet nozzles.

The 18 *control rods* are arranged around the periphery of the cover. To provide adequate space for the on-load refueling machine, control rods are inserted into the reactor core at an angle of about 20°. Filler pieces in the closure head area and in the RPV bottom section are used to displace and save heavy water.

Reactor System. The reactor is operated with natural uranium (UO_2) as fuel and heavy water (D_2O) as the primary coolant and moderator. Figure 84 shows the reactor coolant system

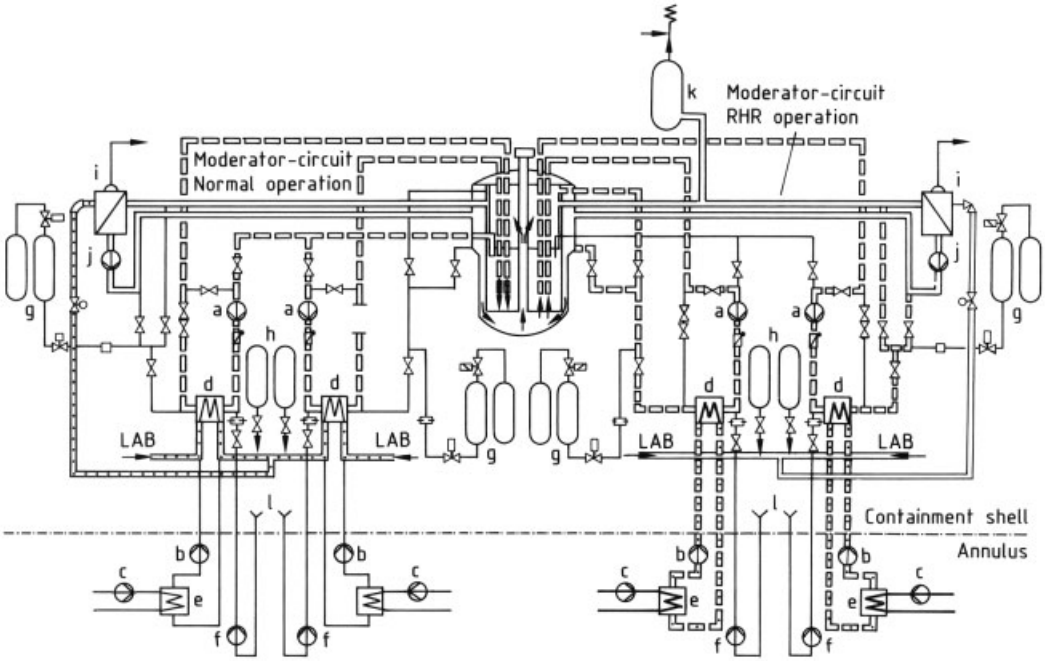

Figure 84. Coolant and moderator system, safety-related system (Siemens/PHWR)
a) Moderator pumps; b) RHR intermediate coolant pumps; c) Secured service cooling water pumps; d) Moderator cooler; e) RHR intermediate cooler; f) Safety injection pumps; g) Accumulators; h) Flooding tanks; i) Steam generator; j) Main coolant pumps; k) Pressurizer; l) Sump [ ] Reactor coolant; □ □ □ Moderator; [· · ·] Feedwater; □ □ □ RHR intermediate coolant
RHR = residual heat removal; LAB = feedwater system

with two parallel reactor coolant loops and four moderator loops.

Coolant Loops. Heat generated in the reactor is passed to the plant's steam and feedwater cycle via the steam generators and moderator coolers. The two reactor coolant loops each contain a steam generator, a reactor coolant pump, a common pressurizer, and interconnecting piping. Heavy water at a pressure of 11.5 MPa is heated from 278 to 312 °C as it flows through the coolant channels. The total flow rate is 37 100 t/h.

To maintain the desired pressure in the primary system and to equalize pressure and volume fluctuations due to changes in temperature, a pressurizer is connected to the reactor coolant loops. It is provided with electrical heating and a spraying system.

Moderator System. The moderator system consists of four identical loops, each with a moderator cooler, a moderator pump, and associated connecting pipings. Heat extracted from the heavy water in the moderator coolers is used to heat the feedwater of the secondary-side steam plant to increase plant efficiency. By changing the feedwater flow, the moderator temperature can be varied between 170 and 220 °C and can thus be used as an additional means of controlling reactivity.

The moderator system must perform several functions. During *normal operation* (left-hand part of Fig. 84) it keeps the moderator temperature lower than that of the coolant. In the *residual heat removal (RHR) mode* the system is reconfigured by changing over certain valves so that residual heat is removed via four independent RHR chains (right-hand part of Fig. 84). In the *emergency cooling mode* the system serves as a high-pressure injection system.

Since altering the moderator temperature greatly affects reactor reactivity, this is used to compensate slow-speed reactivity changes (e.g., xenon poisoning during load-following operation).

Auxiliary and Supporting Systems. Different auxiliary and supporting systems back up reactor system operation. These include a cleaning system for removing corrosion products; a gas-handling system for collecting, treating, and storing gases (deuterium, fission gases) produced in the different parts of the reactor system; and a conditioning and upgrading system for contaminated and downgraded heavy water.

Special Features. To achieve a high burnup, the heavy-water-moderated reactor – with little excess reactivity of the uranium dioxide fuel – requires fuel shuffling and replacement during power plant operation. Refueling is performed automatically by one refueling machine and is monitored from the control room. Normally, approx. two fuel assemblies are replaced in the reactor each full-power day.

This type of reactor is designed for uranium dioxide fuel. However, several different fuel cycles are possible without major plant modifications. These alternatives increase fuel burnup and uranium exploitation, and ultimately result in lower fuel cycle costs.

3.4.3.3. Pressure Tube Boiling-Water Reactor (Decommissioned)

Because expensive heavy water should not be used in the large-volume cooling system of a boiling-water reactor for economic reasons, heavy-water-moderated boiling water reactors [type Winfrith SGHWR (steam-generating heavy-water reactor), Fig. 85] are of hybrid design: light water as the heat-transfer medium flows in the pressure tubes, whereas the heavy-water moderator is accommodated in a tank (calandria vessel). Because of the inevitable neutron losses with this variant through absorption in the light water and structural materials, the reactor can be operated only with slightly enriched nuclear fuel. The ATR concept developed in Japan for commercial application provides for the use of fuel rods containing plutonium (ca. 2.5 % enrichment with fissile material).

Concept. The reactor consists of a calandria vessel filled with heavy water and maintained at low pressure. Numerous, vertically arranged, thermally insulated coolant channels pass through the calandria vessel. The plant also has two steam drums into which the water – steam mixture from the coolant channels is fed (Fig. 86). Each coolant channel contains one fuel assembly from which heat is removed by boiling water.

The Reactor. In the Japanese ATR reactor, 616 coolant channels pass vertically through the calandria vessel. The coolant channels are formed by the calandria tubes, which are firmly connected to the calandria vessel, and by the pressure tubes arranged concentrically inside the calandria tubes. The material used for the pressure tubes is steel (material A 308) outside the reactor core and Zr – 2.5 % Nb in the reactor core. The gap between pressure tube and calandria tube is filled with carbon dioxide for insulation and heat removal.

Each pressure tube contains one fuel assembly, and coolant (H_2O) flows upward through the pressure tubes from below. The fuel assemblies are 36-rod bundles (rod diameter 14.5 mm, length ca. 3700 mm). A Zr – 2.5 % Nb alloy is used as cladding material. Characteristic design data from the Fugen reactor [124] operating in Japan since 1979 and the planned prototype reactor ATR [114] are compiled in Table 40.

Reactor System. Two recirculation pumps transfer the *coolant* at a pressure of ca. 6.7 MPa from the steam drums into two distribution systems from which it is routed into the individual coolant channels. Light water enters the reactor from below at 277 °C and is partially evaporated as it passes through the coolant channels. The steam – water mixture is routed to the steam drums at 284.5 °C. The saturated steam (ca. 6.8 MPa) flows from the steam drums via main steam lines to the turbine and is subsequently condensed in the turbine condenser. Feedwater is pumped back into the steam drums and mixed with the reactor cooling water (Fig. 86).

The *moderator circuit* is separated from the steam – water cycle. Moderator temperature is maintained at ca. 80 °C, i.e., the calandria vessel is at normal pressure. The pressure tubes can be handled individually during operation. The fuel assemblies are inserted and replaced from below by using a remotely controlled fueling machine. Control rods containing boron carbide are in-

Table 42. Breeding ratio of European fast breeder reactors

Reactor	Country	Status	Breeding ratio
PFR	United Kingdom	decommissioned	1.0
Phénix	France	operational	1.16
SNR 300	Germany	decommissioned	0.96 – 1.12
BN 600	Russia	operational	1.3
Superphénix	France	scheduled for decommission	1.24
EFR	France, United Kingdom, Germany	design	1.0 – 1.15

serted into the calandria vessel from above for control purposes.

Special Features. Whereas in the Fugen reactor the level of the heavy-water moderator is normally used for control, the planned ATR is equipped with 13 steel control rods and 76 control rods containing boron carbide, which are inserted into the reactor from above. In addition, a boric acid injection system is also provided. As in light-water reactors, the moderator is poisoned with neutron-absorbing boric acid for long-term control.

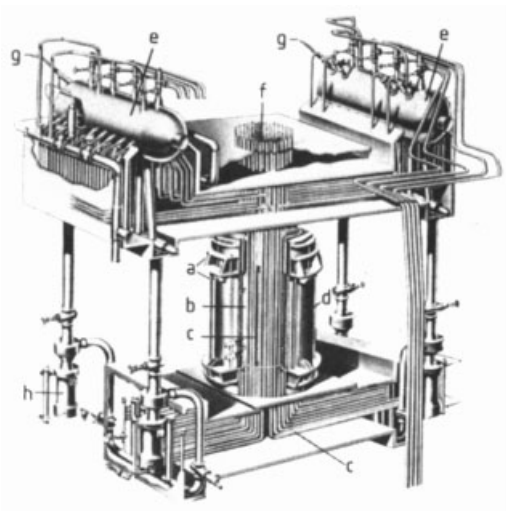

Figure 85. Winfrith steam generating heavy-water reactor (SGHWR)
a) Reactor; b) Calandria vessel; c) Coolant channel with fuel; d) Shielding; e) Steam generator; f) Coolant channel closures; g) Feedwater supply; h) Main coolant pump

4. Fast Reactors

4.1. Function and Design

4.1.1. General Aspects

Fast reactors are employed for electricity generation like light-water reactors, but they may also be used to balance fuel inventories of reactors. In case of high energy demand the economic importance of fast reactors, used as breeder reactors, is based on their ability to produce more new fuel from otherwise nonutilizable waste uranium than they simultaneously consume [125], [126]. This is made possible by the fact that fast breeder reactors, unlike light-water reactors, do not operate with moderated neutrons, but directly with the energy-rich fast neutrons arising from nuclear fission. By the action of fast neutrons, nonfissile ^{238}U (which is contained in natural uranium to the extent of 99.3 %) is converted to fissile plutonium ^{239}Pu:

$$^{238}_{92}\mathrm{U} + ^{1}_{0}\mathrm{n} \longrightarrow ^{239}_{92}\mathrm{U}$$

$$^{239}_{92}\mathrm{U} \longrightarrow ^{239}_{93}\mathrm{Np} + \mathrm{e}$$

$$^{239}_{93}\mathrm{Np} \longrightarrow ^{239}_{94}\mathrm{Pu} + \mathrm{e}$$

The breeding ratio (see Section 1.3.8) is a measure of the excess plutonium produced in the reactor. It is defined as the ratio of production to consumption of fissile isotopes. A reactor is called a breeder if its breeding ratio is > 1; otherwise, it is referred to as a converter.

Light-water reactors are unable to utilize more than 1 % of the uranium fuel for energy production. Therefore, worldwide reserves of uranium (with economically justifiable costs of extraction) could cover present electricity consumption for only a few decades if used exclusively in light-water reactors. On the other hand, fast breeder reactors can utilize uranium 60 times more efficiently; i.e., they can produce

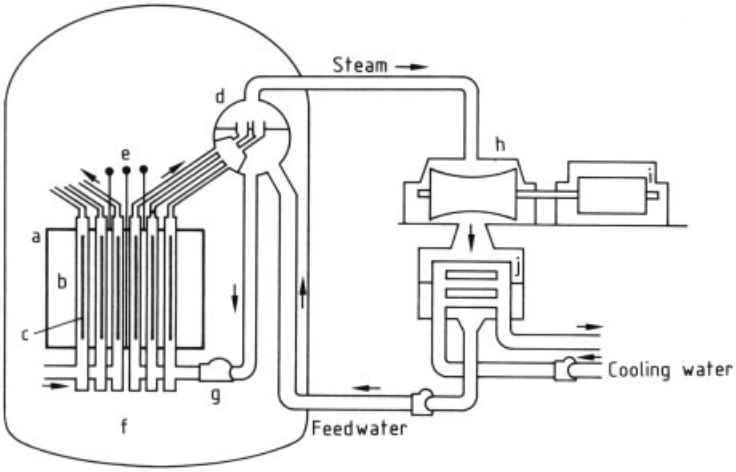

Figure 86. ATR – simplified flow diagram
a) Calandria tank; b) Heavy water; c) Fuel assembly; d) Steam drum; e) Control rods; f) Pressure tube; g) Water pumps; h) Turbine; i) Generator; j) Condenser

60 times more energy from a given amount of uranium than light-water reactors can. The dependence of uranium utilization on the conversion or breeding ratio is shown in Figure 87. At a breeding ratio of 1.0, uranium utilization increases dramatically and reaches its limit at about 60 % [127]. The breeding ratios of some European fast breeder reactors are listed in Table 42.

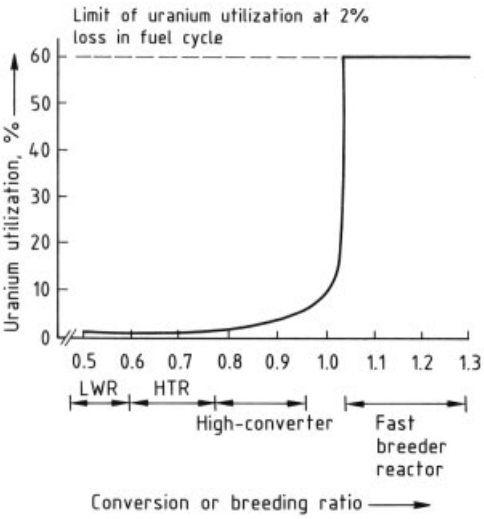

Figure 87. Uranium utilization as a function of conversion or breeding ratio
(High-converters are new types of PWR with highly enriched fuel that are in the developmental stage.)

The total amount of energy that could be produced from the known reserves of uranium worldwide by using fast breeder reactors is substantially higher than the sum of all fossil energy reserves [128] (see Fig. 88). Fast breeder reactors can be employed commercially only if their development continues. In this way, a marked decrease in the specific uranium requirement could be achieved, as well as an efficient utilization of the plutonium obtained in reprocessing spent fuel elements from light-water and high-temperature reactors [130].

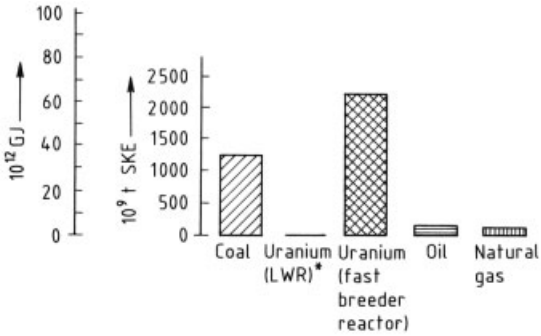

Figure 88. Economical availability of worldwide energy resources (in 1986). By comparison, water power potential $< 1.5 \times 10^6$ t SKE/a. (SKE = German Steinkohleneinheit), 1t SKE=29.3GJ. $*$ Cost of extraction ≤ 80 \$/kg uranium.

The possibility of artificially producing fissile material and being independent of natural re-

sources was one of the early considerations in the use of energy from nuclear fission. In 1951 the EBR-1 (experimental breeder reactor) in Idaho, United States, succeeded for the first time in generating electricity from nuclear energy with a fast liquid-metal (Na – K) reactor [129].

During operation, fast breeder reactors require supplies of only nonfissile waste uranium and produce an excess of plutonium over and above their own requirements. This plutonium can then be used in other fast breeder reactors. Thus, the cost of electricity generated by fast breeder reactors is largely independent of the price of uranium, and safe supply to the power industry is substantially improved [131].

The annual consumption of natural uranium will decrease worldwide if fast breeder reactors are increasingly employed (see Fig. 89). This reasoning is based on the assumption that the rate of increase in nuclear power installations will follow the curve in Figure 90 worldwide and that fast breeder reactors with uranium – plutonium oxide fuel and a breeding ratio of 1.15 – 1.20 will be employed by 2010 [132], [133].

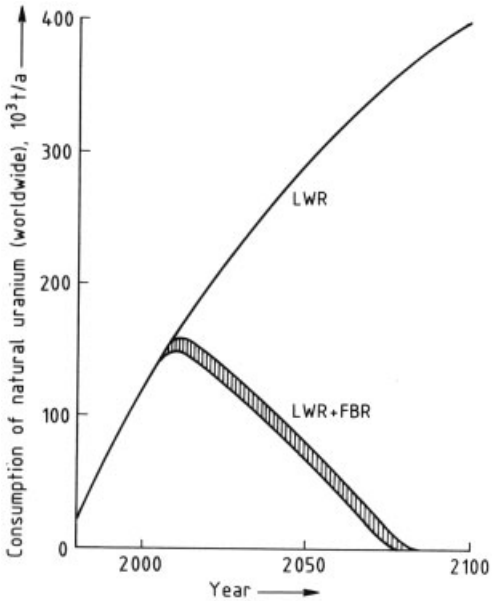

Figure 89. Annual consumption of natural uranium (worldwide) in fast breeder reactors

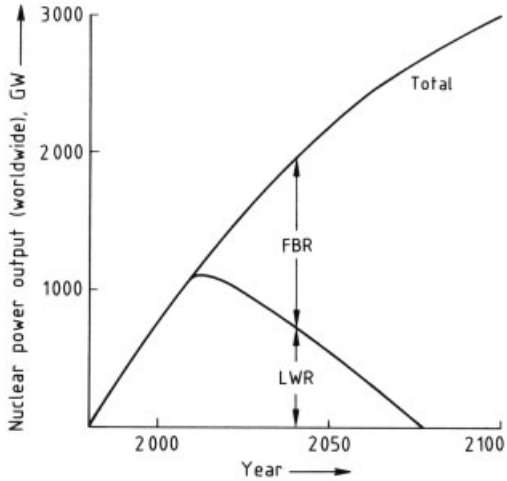

Figure 90. Nuclear power output (worldwide)

Nevertheless, the energy need in the last two decades was less than predicted at the time when the decision was made to built the first industrial size fast breeder reactors such as BN 600 and Superphénix. Therefore, the new fast reactor programs were slowed down or even abandoned. A survey of fast breeder reactors is given in Table 43 [138–147]. On the other hand the problems of the back-end of the nuclear fuel cycle, especially the waste disposal, become more important. In this framework fast reactors gain interest because of their capability to produce various fast neutron spectra. This particularity may be used to transform fission and conversion products from light-water reactors, which need particular precautions for waste disposal, into other products which are easier to handle. These products are mainly plutonium which may be used to produce atomic weapons and minor actinides that are of concern for due to their long half-life. In this framework the role of plutonium may be modified depending on national policies. There may be nations who do no longer use plutonium as a fuel but consider it as waste that is to be treated. Others continue to see plutonium as a potential energy source. At present various methods to adept fast reactors to the new scope of plutonium conversion and actinide transmutation are under investigation: One possibility is to modify the core of fast breeder reactors by introducing plutonium and minor actinides in specific, highly enriched core fuel (CAPRA

Table 43. Survey of fast breeder reactors

Reactor name	Country	Date of criticality	Power MW_{th}	Power MW_e	Fuel	Coolant system
EBR I (shut down)	United States	1951	1.2	0.2	enriched U	1 loop
BR 5	Russia	1958 1973	5 10		PuO_2 and UC	2 loops
DFR (shut down)	United Kingdom	1959	60	14	UO_2	24 loops
Enrico Fermi (shut down)	United States	1963	300	60.9	U	4 loops
EBR II	United States	1963	62.5	18.5	U	pool
KNK II	Germany	1977	58	20	PuO_2-UO_2	2 loops
Rapsodie	France	1966	20		PuO_2-UO_2	2 loops
Sefor (shut down)	United States	1969	20		PuO_2-UO_2	1 loop
BOR 60	Russia	1969	60	12	UO_2	2 loops
BN 350	Kazakhstan	1972	1000	150	PuO_2-UO_2	6 loops
Joyo	Japan	1977	100		PuO_2-UO_2	2 loops
Phénix	France	1973	560	250	PuO_2-UO_2	pool
PFR	United Kingdom	1974	600	250	PuO_2-UO_2	pool
FFTF	United States	1980	400		PuO_2-UO_2	3 loops
BN 600	Russia	1980	1470	600	PuO_2-UO_2	pool
Superphénix	France	1985	2900	1200	PuO_2-UO_2	pool
SNR 300	Germany	abandoned before criticality	736	312	PuO_2-UO_2	3 loops
PEC	Italy	abandoned before end of construction	125		PuO_2-UO_2	2 loops half-pool
Monju	Japan	1994	714	300	PuO_2-UO_2	3 loops
EFR	France, United Kingdom, Germany	design	3600	1500	PuO_2-UO_2	pool

core = "Consommation Accrue de Plutonium dans les Réacteurs Rapides"; English: Increased Plutonium Consumption in Fast Reactors). Another one is to combine a subcritical fast reactor with a spallation source in order to produce hard neutron spectra for efficient actinide transmutation. This idea is now under evaluation in various R & D organizations under the keywords Hybrid Reactors and Accelerator Driven Systems. At present the nuclear industry is beginning to evaluate the industrial relevance of this kind of reactor.

Coolant. *Water* cannot be used for heat transfer because it would moderate the neutrons produced by nuclear fission, as in light-water reactors. Former research and developmental work on *steam cooling* has been discontinued in particular because of the too low breeding ratio. Similarly, cooling with *helium* is no longer being pursued because the high power density in the core requires especially extensive measures for heat removal. The fast breeder reactor operating with liquid *sodium* is the only technically safe concept that can guarantee a low-risk and ecologically beneficial supply of electric power for centuries. Sodium-cooled fast breeder reactors are technically feasible and can be operated reliably with the sodium technology fully under control [134], [135]. In fact, the French Superphénix achieved its full output of 1200 MW_e at the end of 1986 [136].

Status and Future Developments. The present state of fast reactor development in Europe is demonstrated by the EFR (European Fast Reactor) project. In planning since 1988 [137], this project combines different national fast reactor variants in a common concept. The present EFR reference design is the most advanced economic optimization of a fast reactor taking into account current safety requirements, operating experiences gained in the past, and approved updated technological knowledge. Nevertheless, abundance of energy supply that will reduce near-future interest in EFR fast reactor development, leads to a re-evaluation of related R & D long term objectives towards fuel cycle back-end and waste management within fast reactor systems.

4.1.2. Sodium as Coolant for Fast Reactor Power Plants

The low neutron moderation, which enables a high breeding ratio, and the excellent heat-transfer properties of sodium make it a particularly suitable coolant for a fast core. However, a specially designed plant is required if sodium is to be used [148–150].

Thermal Properties. A comparison of the thermodynamic properties of sodium and water (Table 44) shows that sodium has a higher melting point and boiling point than water, excellent heat transport properties, and a relatively low specific heat capacity.

Table 44. Comparison of the thermodynamic properties of sodium and water

	Sodium	Water
Density, kg/m^3	850	1000
mp (0.1 MPa), °C	98	0
bp (0.1 MPa), °C	883	100
Specific heat capacity, $kJ\,kg^{-1}\,K^{-1}$	1.3 *	4.2 **
Thermal conductivity, $kJ\,m^{-1}\,h^{-1}\,K^{-1}$	260 *	2.1 **
Viscosity, Pa · s	3×10^{-4} *	10×10^{-4} **

* At 400 °C.
** At 200 °C.

The concept of the plant leads to the following:

1) The plant can be designed as a low-pressure system even at high temperature. The steam conditions of conventional power plants can be achieved, with the resulting high efficiencies.
2) The sodium level in vessels has to be protected against contact with the atmosphere by an inert cover gas, usually argon or nitrogen. Due to the fact that sodium plant loops operate far below boiling temperature of the coolant, the overpressure of the cover gas is low. Thus the operation pressure in the loop depends mainly on the dynamic pressure in the coolant flow and the geodetic pressure. In real plants the overpressure in operation at the lowest point does not exceed 15 bar. For this reason, large cross-sections of piping are not in danger of opening up due to sudden breakage, and large leaks are unlikely to occur in vessels. In case of a structural failure of the sodium containing structure (leak, crack) no violent pressure build-up in the reactor containment occurs. The pressure transient in the containment atmosphere depends only on thermal effects of the release of the hot coolant (and not on evaporation effects as in water-cooled reactors). Therefore, the design pressures of the containment system are low with correspondingly greater freedom of design. In case of malfunction, excess heat can be absorbed by the heat capacity of the remaining coolant without significant pressure increase because of the large difference between the operating temperature of ca. 550 °C and the boiling point of sodium (ca. 1000 °C) at the reactor pressure. Therefore, the effects can be safely controlled.
3) Although the power density in the core of a fast reactor is high, the good heat exchange properties of sodium allow heat removal throughout the range of permissible temperature gradients between the fuel rod bundle and the sodium. For the same reason, the specific heating surface requirement in heat exchangers is low.
4) The low specific heat capacity of sodium in comparison to water necessitates large coolant temperature ranges (i.e., temperature difference core outlet – core inlet) with approximately the same rate of coolant circulation as in pressurized-water reactors.
5) The great temperature range between ambient temperature and operating temperatures of about 550 °C requires extensive measures for compensation of thermal expansion. Components are designed to withstand high temperatures.
6) The good heat-exchange properties of the coolant, high operating temperatures, and large coolant temperature ranges require extensive temperature stress analyses and components that are designed to be as insensitive to thermal shock as possible.
7) For start-up and maintenance of readiness for operation, the sodium-containing components and pipes must be heated either electrically with a heating cable or with heating gas.

Nuclear Physical Properties. The nuclear physical properties of sodium may be summarized as follows:

1) Sodium has a low moderation capacity; therefore, the fast neutron spectrum required for fast reactors can be obtained. For the same reason fast reactor cores can be designed for multiple purposes as, e.g., burning of plutonium (breeding gain < 1) or transformation of minor actinides.
2) Sodium is strongly activated while flowing through the reactor core by the following reactions:

$$^{23}\mathrm{Na}\ (\mathrm{n}, \gamma)\ ^{24}\mathrm{Na}$$

$$^{23}\mathrm{Na}\ (\mathrm{n}, 2\,\mathrm{n})\ ^{22}\mathrm{Na}$$

For this reason, parts of the plant through which the core coolant flows are accommodated in strongly shielded rooms. These rooms are not accessible during operation and can be entered only after the plant has been shut down and 10 – 14 days have passed to allow for the decay of ^{24}Na.

Chemical Properties. Sodium undergoes a strongly exothermic reaction with water; hot sodium reacts with oxygen. To protect the reactor from reactions caused by the rupture of a pipe in the steam generator and to prevent reactions with highly radioactive sodium, the coolant in the reactor circulation system (primary circuit) does not enter the steam generator directly. Rather, they are separated by a nonradioactive intermediate heat-transfer circuit (secondary circuit). The compartments with the primary system are maintained under a nitrogen atmosphere with a low oxygen content. Therefore, in the event of accidents involving sodium leakage, the possible burning of activated sodium is excluded.

The secondary circuit is provided with a pressure relief system to safely remove reaction products and limit a pressure buildup due to the sodium – water reaction:

$$2\,\mathrm{Na}_{(l)} + \mathrm{H_2O}_{(l)} \longrightarrow \mathrm{Na_2O}_{(s)} + \mathrm{H_2}\quad \Delta H^{\circ}_{298} = -\,134.3\,\mathrm{kJ/mol}$$

Both ferritic and austenitic steels can be employed because sodium with a low oxide content is minimally corrosive. The oxide content is always kept low by use of an appropriate purification system. All free sodium surfaces are covered with argon to prevent stronger oxide contamination or inflammation of sodium in the presence of oxygen.

Mechanical Properties. The coefficient of fricton of metallic surfaces that are wetted with sodium is higher than that of pure metal surfaces. For this reason, special pairing of materials is required for surfaces that slide over each other. This is especially important in the handling of core elements.

4.1.3. Plant Concepts for Fast Reactor Power Stations

Two concepts are employed in the design of the primary circuit system of fast breeder reactors (with sodium as coolant):

1) The *pool concept*, in which the pump and intermediate heat exchanger are integrated into the reactor vessel.
2) The *loop concept*, in which the pump and intermediate heat exchanger are located outside the reactor vessel in shielded and nitrogen-inert rooms. They are connected to the reactor vessel by a piping system.

No serious technical, safety, or cost-related differences exist between these two concepts. The pool concept is the concept of choice for the EFR because with the Superphénix (SPX), a plant having a power range $> 1200\,\mathrm{MW_e}$ has already been realized by using this concept.

Figure 91 shows a schematic of the nuclear power plant SNR 300 (Kalkar), which has a loop reactor [151], [152]. The Japanese demonstration fast reactor plant Monju follows the same design principle. The schematic of the French Superphénix, a fast breeder with a pool reactor, is depicted in Figure 92 [153], [154].

All fast reactor power stations have a primary and a secondary sodium circuit. Sodium in the primary circuit flows through the reactor core and transfers absorbed heat energy to the sodium of the secondary circuit in intermediate heat exchangers (IHX). This secondary sodium circuit transports heat energy to the conventional water – steam cycle via steam generators. The steam thus produced passes to the steam turbine,

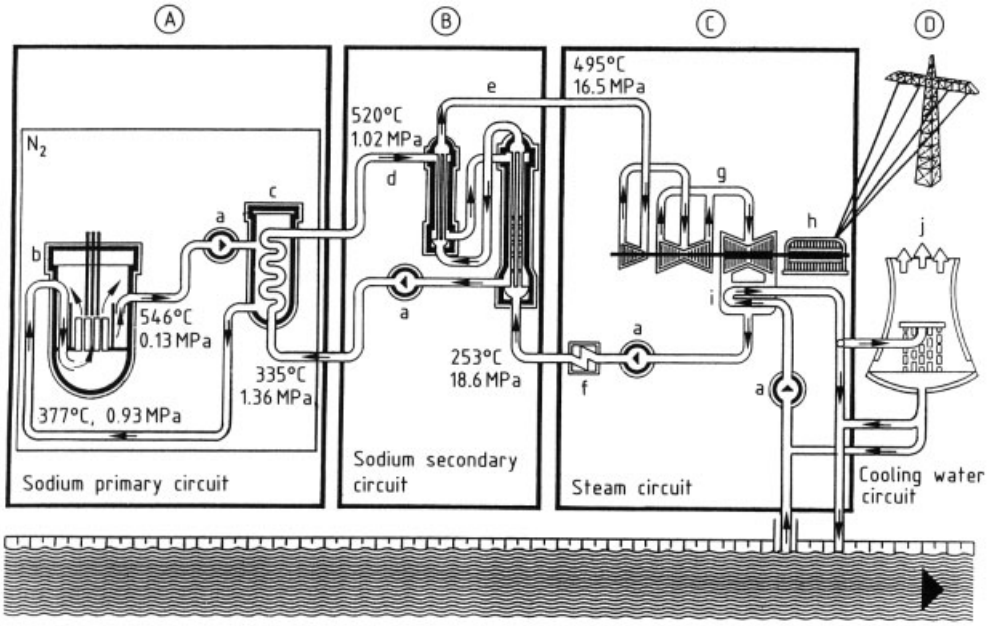

Figure 91. Schematic of SNR300 (loop concept)
A) Reactor building; B) Steam generator building; C) Powerhouse; D) Power supply system
a) Pump; b) Reactor vessel with internals; c) Intermediate heat exchanger; d) Superheater; e) Evaporator; f) Preheater; g) Turbine; h) Generator; i) Condenser; j) Cooling tower

which powers the generator to generate electricity. In this way, a rigorous separation is obtained between the first (nuclear) and the third (conventional) heat transfer system (= water – steam system). The pressure in the secondary circuit is always higher than in the primary circuit. Thus, in the event of leakage in the intermediate heat exchanger, no radioactive sodium from the primary system can enter the practically nonradioactive sodium of the secondary system.

In the *pool concept*, the reactor vessel is enclosed in a second container, the safety vessel, to ensure that the reactor core is always covered with sodium. Therefore, the core can be cooled, even in case of leakage in the reactor vessel or in the primary circuit. In the *loop concept* the primary circuits are generally arranged above the reactor core level in order to prevent the reactor vessel from draining by siphon effect. Those parts of the primary loops that are arranged below the reactor core level are equipped with a protection against loop draining by a double envelope of the structure. Information on core, fuel, sodium temperature, and main steam conditions is provided in Table 45. The principal steam data of fast reactors are similar to those of a conventional steam power plant. For this reason, fast reactor power plants have a thermodynamic efficiency > 40 %.

4.1.4. Plant Arrangement

A longitudinal section through the reactor, fuel handling building, and turbine hall of the Superphénix is shown in Figure 93 [155].

4.1.5. Components and Systems

The most important characteristics of the different systems and components of sodium-cooled fast reactors are described below (using the SNR 300 as an example of a typical loop reactor and the Superphénix as the most powerful pool reactor).

Reactor Core. A cross-section of the core of SNR 300 is shown in Figure 94 [156], [157]. The core consists of hexagonal elements, which are radially subdivided (starting from the inside) into the following three zones:

Table 45. Data on core, fuel, sodium temperature, and steam conditions of fast breeder reactors

	EBR II	BOR 60	KNK II	Phénix	PFR	BN 350	SNR 300	BN 600	Superphénix
Thermal power, MW_{th}	62	60	58	560	600	1000	730	1470	3000
Electric power, MW_e	20	12	21	250	270	150 desalination	327	600	1240
Active part of core									
Height, m	0.36	0.41	0.62	0.85	0.9	1.06	0.95	0.75	1
Diameter, m	0.65	0.40	0.82	1.39	1.45	1.50	1.78	2.05	3.6
Total mass of active fuel, kg	590	176 ^{235}U	700	4300	4100	1050 ^{235}U	5100	1260 ^{235}U	31 500
Type of active fuel	U	UO_2	UO_2 PuO_2	UO_2 PuO_2	PuO_2 UO_2	UO_2 or PuO_2	UO_2 PuO_2	UO_2 PuO_2	UO_2 PuO_2
Fissile zone enrichment	67 % ^{235}U	90 % ^{235}U	30 % Pu	19.2 – 27.2 % Pu	24 – 30 % Pu	23.19 % ^{235}U	23.7 – 33.3 % Pu	21 – 28 % ^{235}U	15 % Pu
Average specific core thermal power, MW	670	800	300	406	500	500	330	550	280
Expected burnup, MW · d/t	42 000	10 % burnout	65 000	100 000	61 000	38 000	56 000	10 % burnout	70 000
Sodium temperature at core inlet/outlet, °C	371/472	360/600	362/527	400/560	400/562	300/500	377/546	377/550	392/545
Steam conditions									
Temperature, °C	423	435	485	510	538	435	495	505	487
Pressure, MPa	8.79	7	8	16.3	16.17	5	16.5	14	18

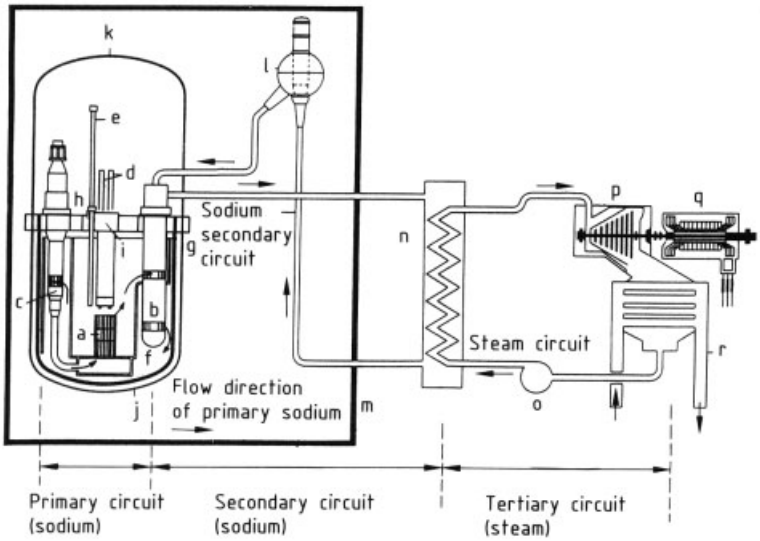

Figure 92. Schematic of Superphénix (pool concept)
a) Core; b) Intermediate heat exchanger; c) Primary pump; d) Control elements; e) Transfer machine; f) Reactor vessel; g) Closure plate; h) Large rotating top shield; i) Small rotating top shield; j) Safety vessel; k) Dome; l) Secondary pump; m) Reactor building; n) Steam generator; o) Pump; p) Turbine; q) Generator; r) Condenser

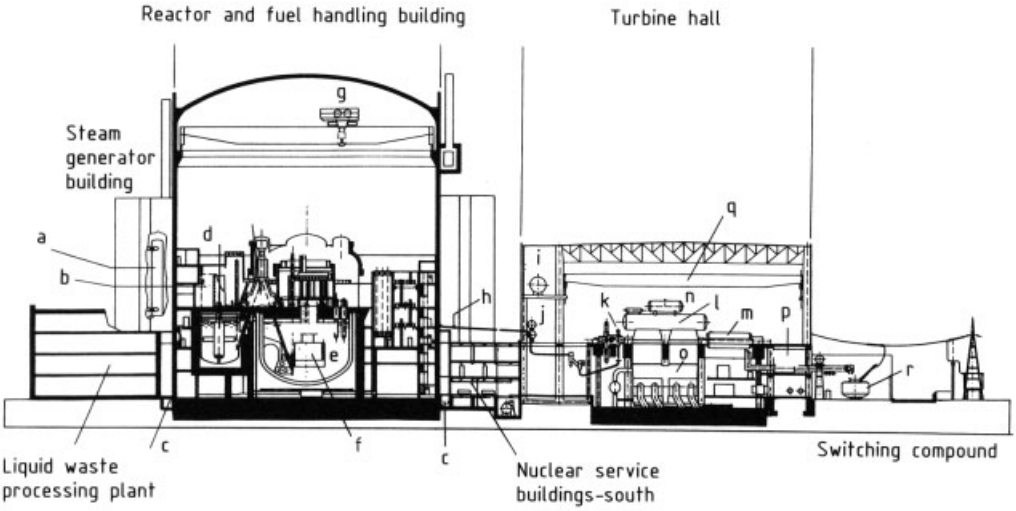

Figure 93. Longitudinal cross-section of Superphénix
a) Steam generator; b) Handling area; c) Peripheral gallery; d) Storage carousel; e) Core; f) Reactor; g) Crane (3500 kN); h) Steam circuit; i) Flash tank; j) Recuperative heat exchanger; k) Steam turbine l) Feedwater tank; m) Generator; n) Deaerator; o) Condenser; p) Coupling switch; q) Crane (1200 kN); r) Generator, transformer

1) The *fission zone* in which almost all the reactor power is produced. This zone is normally subdivided into two enrichment zones (inner and outer zone in Fig. 94).
2) The *blanket*, which contains the blanket subassemblies. Here, ^{238}U is converted to ^{239}Pu. The fission zone is radially and axially surrounded by the blanket.
3) The *reflector zone*. The subassemblies of this zone primarily shield important structures located around the core (core shroud, reactor vessel, plus intermediate heat exchangers and pumps in the pool concept).

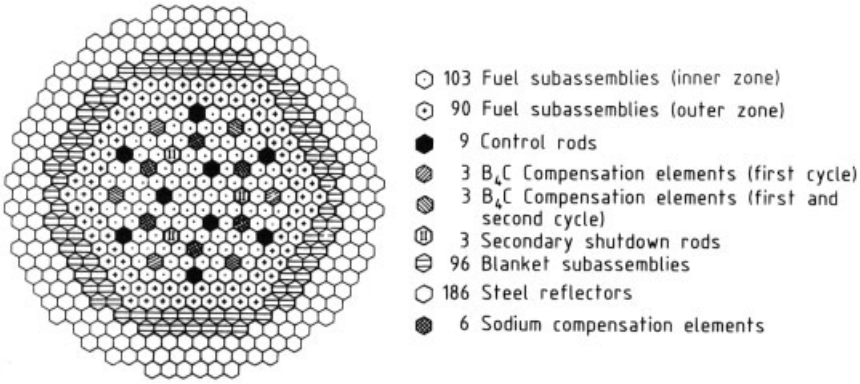

Figure 94. Transverse section of SNR300 with Mark Ia core (first and second cycle)

The *fuel* used in today's fast reactor power plants is a mixture of uranium dioxide and plutonium dioxide. To achieve better power distribution, the inner fission zone has a lower enrichment with plutonium than the outer one. Axially movable absorber rods made of boron carbide are used for control and shutdown. These are located primarily in the inner zone (on fuel element positions) and are uniformly distributed over its cross-section.

The entire core structure is fixed by the diagrid (grid plate) and the core shroud. This diagrid is also used to guide the sodium coolant to different core zones. In this process, sodium flows upward through the core.

Reactor Vessel System. The reactor vessel system consists of the reactor vessel, the safety (guard) vessel, vessel internals (shield tank, guiding system for internal flow, and core support structure), and the rotating top shield. The reactor vessel system of the SNR 300 is shown in Figure 95 [158], [159]. The free surface of the sodium contained in the reactor vessel is always blanketed with argon. The vessel room is both the collecting compartment and the expansion vessel for the primary circuits. The reactor vessel is basically not pressure stressed because of the closed entry plenum below the grid plate. All pipes of the primary heat-transfer system are connected to the vessel above the sodium minimum level in the vessel that is required for decay power (residual heat) removal.

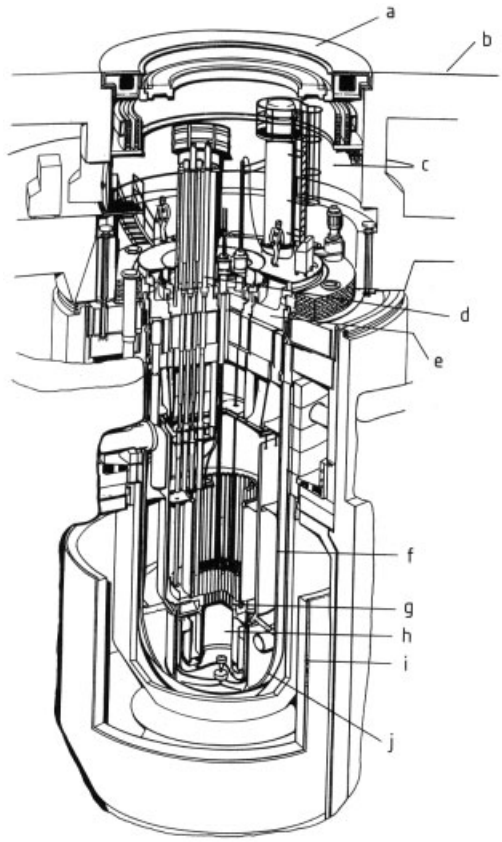

Figure 95. Reactor vessel system of SNR300
a) Pit cover; b) Reactor floor; c) Reactor cell; d) Control rods; e) Rotating top shield; f) Reactor vessel; g) Diagrid; h) Gas bubble separator; i) Core catcher with cooling system; j) Guard vessel

The reactor vessel of a *pool reactor* contains, in addition, the entire primary heat-transfer sys-

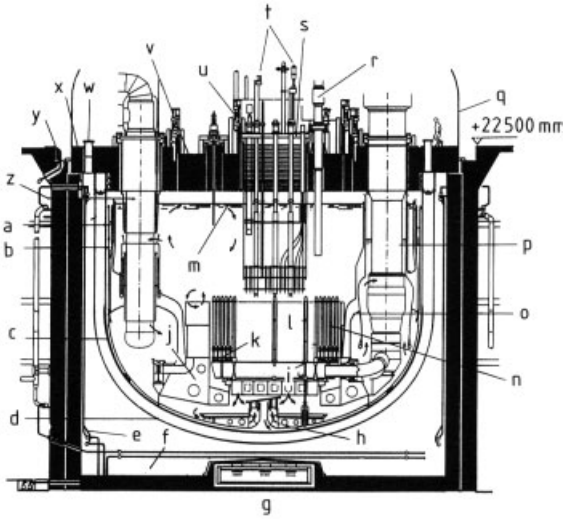

Figure 96. Superphénix reactor
a) Overflow; b) Main vessel cooling baffles; c) Main reactor vessel ($d = 21$ m); d) Safety vessel ($d = 22.45$ m); e) Emergency cooling circuit; f) Insulation; g) Neutron measurement cell; h) Core catcher; i) Core diagrid; j) Core support plate; k) Dummy diagrid; l) Core; m) Sodium level; n) Lateral neutron shielding assemblies; o) Inner tank; p) Primary pump; q) Dome; r) Transfer machine; s) Core closure plate; t) Control rod mechanism; u) Small rotating plug; v) Large rotating plug; w) Inspection hole; x) Roof slab; y) Cover cooling circuit; z) Intermediate heat exchanger

tem, which consists of the primary pumps and the intermediate heat exchangers. A section of the Superphénix reactor is shown in Figure 96.

Fuel Handling. The economics of the fast breeder fuel cycle require that the ex-core period of the fuel (time between unloading of spent fuel and blanket subassemblies from the reactor core and refueling with reprocessed fuel) is kept as short as possible. However, because the decay power of the spent fuel considerably hinders the handling and treatment of these assemblies, newer fast reactors are provided with facilities in the reactor vessel outside the core shielding to store spent fuel for an operational cycle (ca. 1 year). Hence the fuel assemblies remain continually under sodium during this period. After this storage time the decay power decreases to < 10 kW per fuel assembly. The fuel assemblies can then remain without sodium cooling during the handling transfer to the spent fuel storage. In Superphénix the spent fuel elements are cleaned by a special cleaning process, which is a chemical reaction of sodium to yield sodium hydroxide or sodium carbonate, and subsequent washing with water [160]. In Superphénix the washed elements are transferred to a water storage. This facility allows to store several core loadings in order to cover the time span before industrial reprocessing.

To facilitate the *refueling process* (Fig. 97), which is carried out after the reactor has been shut down and the temperature of sodium has fallen to ca. 200 °C, the central part of the reactor roof consists of a rotating top shield. With this shield the structures arranged above the core (control mechanisms and instrumentation) are swung to one side, making the fuel and blanket subassemblies accessible to the handling equipment located in the rotating top shield. The spent and decayed subassemblies, along with the new assemblies, are removed from or inserted into the reactor vessel, respectively, by using a shielded handling flask. The spent fuel and blanket subassemblies are subsequently prepared for transport to reprocessing facilities in a hot cell (i.e., discharge cell for irradiated fuel) and are then loaded into a transport cask. A transport cask provided with adequate shielding and fall protection has a mass of approximately 100 t.

Heat-Transfer Systems. Heat generated in the reactor is transferred to the water – steam cycle by the primary and secondary heat-transfer

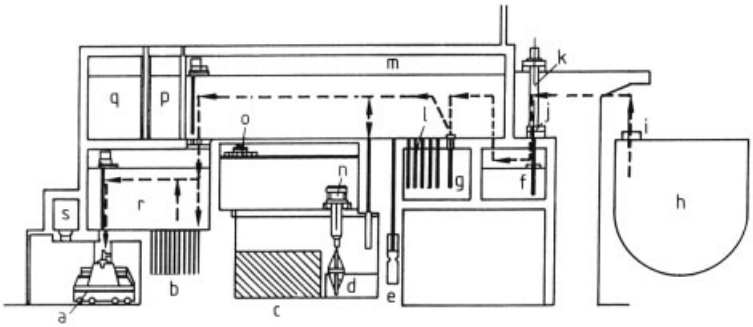

Figure 97. Fuel handling
a) Transport cask; b) Buffer store; e) Subassembly storage in water pond; d) Pond transfer; e) New fuel transfer; f) Subassembly transfer; g) Wash cell; h) Reactor; i) Reactor transfer valve; j) Transfer valve; k) Transfer flask; l) New fuel store; m) New and irradiated fuel cell; n) Pond transporter; o) Maintenance crane; p) Decontamination cell; q) Maintenance cell (fuel cell); r) Irradiated fuel discharge cell; s) Cask plug removal cell
— — — Handling of spent fuel and blanket subassemblies (from reactor to transport cask; intermediate storage in water pond)

systems. The most important components are sodium pumps, intermediate heat exchangers, steam generators, and piping (see Figs. 91 and 92). The principal data of the components of the fast breeders SNR 300 and Superphénix are summarized in Table 46 [161–164].

Sodium Pumps. Single-stage, single-flow, speed-controlled centrifugal pumps are provided for the circulation of sodium in primary and secondary circuits. The free sodium surface in the pump casing is covered with argon gas. The vertical shaft has double bearings – an oiled roller bearing at the upper end and a sodium-lubricated hydrostatic radial bearing at the lower end in the region of the impeller. The use of electromagnetic pumps is limited to sodium auxiliary plants (e.g., purification systems) because of their low efficiency.

Intermediate Heat Exchangers. The intermediate heat exchangers are the link between the primary and secondary circuits. Here, heat produced in the reactor is transferred from radioactive primary sodium to nonradioactive secondary sodium. The design of intermediate heat exchangers is shown in Figure 98. These devices can have very compact structures because of the good heat-transfer properties of sodium and because of the use of thin-walled tubes (made possible by the low pressure of sodium). This is a great advantage for the pool reactor because the intermediate heat exchanger is located in the reactor vessel.

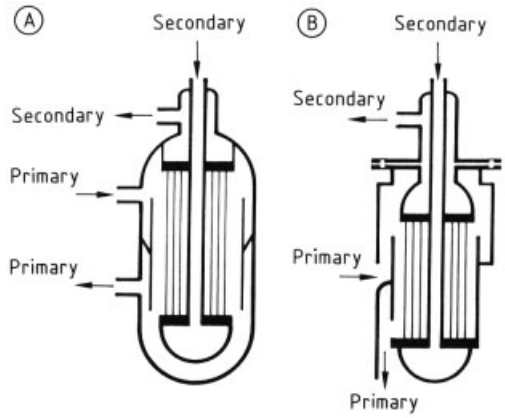

Figure 98. Intermediate heat exchanger (design principle)
A) Loop type; B) Pool type

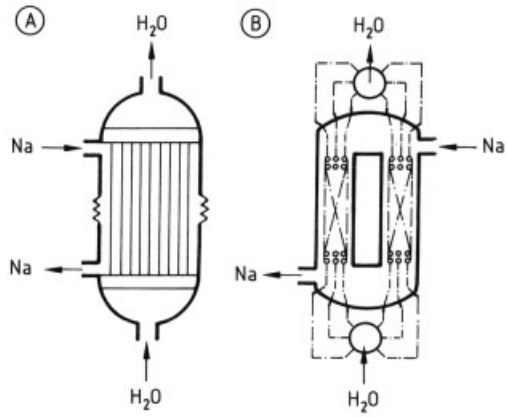

Figure 99. Steam generator
A) Straight tube type; B) Helicoidal tube type

Steam Generators. Extremely stringent quality requirements must be met by the steam generators because the sodium and water flowing through these components are separated only by a tube wall. In the event of a leak, these two media would react chemically with each other. To limit the pressure buildup in the secondary circuit that would then develop, steam generators are provided with fast-responding rupture disks on the sodium side. These allow the reaction products to be removed into a pressure relief system. Different types of steam generators are used, depending on the latest technological developments and the performance required (see Fig. 99).

Auxiliary Systems. Fast reactor power plants that use sodium as coolant require special systems to guarantee safe operation and the functioning of the primary heat-transfer systems. The required purity of sodium is produced, measured, and monitored in *purification systems. Argon* is used in every case when a gas is necessary as operation and control fluid and this fluid is in permanent contact with the sodium coolant. Argon systems are used for the following tasks:

1) As cover gas for all free sodium surfaces in the reactor vessel, sodium pumps, sodium tanks, and some types of steam generators.
2) For buildup and control of the required system pressures at different operating states, and generation of excess or subatmospheric pressure for sodium decanting (operating pressure in reactor vessel ca. 0.12 MPa).
3) To supply handling devices and reactor top shield with sealing, rinsing, and pressurizing gas for seals and compartments.
4) For collecting and treating of radioactive gases and controlled release of treated gas to the environment.
5) For early detection of damages of fuel assemblies in the reactor vessel.
6) For purification of argon from impurities and fission products.

In case of a leak in the steam generator, *pressure relief systems* in the region of the steam generator prevent damage (especially to the intermediate heat exchanger) that results from sodium – water reactions. In addition, these systems limit the spread of reaction products in the secondary circuits.

Table 46. Primary data of SNR 300 and Superphénix components

Power station	SNR 300 (Kalkar)	Superphénix (Creys Malville)
Type	loop	pool
Thermal/electric power, MW_{th}/MW_e	770/327	3000/1240
Number of primary/secondary loops	3/3	0/4
Primary sodium pumps		
Number of units	3	4
Capacity per unit, kg/s	1270	4240
Delivery head, m Na	140	63
Maximum rotation speed, rpm	960	452
Operating temperature, °C	546	395
Shaft power, MW	2.20	4.19
Material	austenitic steel 1.4948	austenitic steel 316 ss
Intermediate heat exchanger		
Number of units	3×3	8
Thermal power per unit, MW	85	395
Primary inlet/outlet temperature, °C	546/377	542/392
Secondary inlet/outlet temperature, °C	335/520	345/525
Design concept	straight tube	straight tube
Number of tubes per unit	846	5380
Diameter×thickness of tubes, mm	21×1.4	14×1
Material	austenitic steel 1.4948	austenitic steel 316 ss
Steam generators		
Number of units	9, evaporator 9, superheater	4, once-through type
Thermal power per unit, MW	85 (54+31)	750
Sodium inlet/outlet temperature, °C	520/335	525/490
Feedwater/steam temperature, °C	253/500	235/490
Design concept	both straight and helicoidal tube	helicoidal
Number of straight/helicoidal tubes per unit	650/77	0/357
Diameter×thickness of tubes, mm		25×2.6
Evaporator (straight/helicoidal)	17.2×2.0/26.9×2.9/	
Superheater (straight/helicoidal)	17.2×2.9/26.9×4.5	
Material for tubes	2.25 Cr – 1 Mo – Ni – Nb	Incoloy 800 2.25 Cr – 1 Mo

The decay power removal after reactor shutdown normally occurs via the cooling-water cycle of the steam turbine plant. If this cooling chain fails, a multifunctional *decay power removal system* is available (see Fig. 100). The decay heat is first transferred to immersion coolers ar-

ranged in the reactor vessel and then passed to a sodium – air heat exchanger via sodium circuits [165], [166].

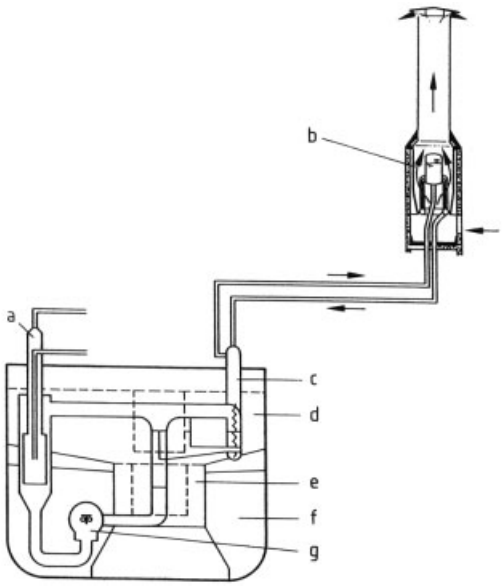

Figure 100. Decay power removal system
a) Intermediate heat exchanger; b) Sodium – air heat exchanger; c) Immersion cooler; d) Hot plenum; e) Core; f) Cold plenum; g) Pump

Other systems characteristic of sodium-cooled fast reactors are *inert gas systems* that maintain the plant rooms of the radioactive sodium systems (mainly in loop reactor systems) under a neutral gas atmosphere in order to prevent sodium fire in case of a leak. In these cases nitrogen is used as inert gas. Another role of the inert gas systems may be cooling and heating of components and systems.

Since sodium has a melting point of 98 °C, all sodium components and piping are provided with *facilities for heating*. Heating is carried out either by an electric trace heater or by hot argon or nitrogen.

Instrumentation and control facilities, as well as power supply, are attached to all systems. Specially developed instruments for measuring pressure, level control, flow, and oxygen content in sodium have been employed for sodium systems.

Steam Turbine Unit. The steam turbine unit is essentially identical to that in conventional units because the steam conditions of a fast reactor power plant are similar to those of conventional power plants. Special facilities include cooling systems for decay power removal.

4.2. Materials [167]

4.2.1. Nuclear Fuels

In principle, the fuels uranium and plutonium can be used in two forms: as pure metals or as ceramic compounds (oxide, carbide, or nitride). Metallic fuel produces a higher rate of burning because of its higher uranium density. However, its disadvantages are a limited operating temperature because of crystallographic transformation, its low melting point, and its radiation properties at high burnup.

Ceramic fuels with an isotropic crystal structure and a production-related porosity for the storage of fission products achieve a high degree of burnup. Today, oxidic nuclear fuel is used almost exclusively. Currently under development are carbidic and nitridic fuels, which give a large breeding ratio in combination with a lower plutonium stock in the reactor core. Oxidic fuel is prepared as pellets by sintering. Great demands are made on the homogeneity of the two components of the mixed oxide (UO_2 and PuO_2) and on their solubility in boiling nitric acid in accordance with the initial stage of the PUREX process.

4.2.2. Cladding and Structural Materials for Fuel Elements

A list of the cladding and structural materials of fast reactors with sodium cooling is presented in Table 47. The stress conditions of the cladding and structural materials are characterized by the sodium-specific corrosion processes that occur at high temperature (up to 700 °C for the cladding material and 560 °C for the structural material). In addition, this material is stressed by the high flux ($> 10^{23}$ n/cm^2) of fast neutrons. The fuel cladding tubes are also subjected to interactions with the fuel. Currently, cold-formed austenitic steels are preferred [e.g., the titanium-stabilized material X10 NiCrMoTiB 15 15 (1.4970) for the SNR 300 and

Table 47. Materials for fast breeder reactors

Reactor	Cladding	Core structure reactor vessel/-internals vessel, pipes, IHX *	Steam generator	
			Evaporator	Superheater
EBR II	type 304 ss	type 304 ss	2.25 Cr – 1 Mo	2.25 Cr – 1 Mo
BOR 60	OKh 16 N 15 M3B	type 304 ss	2.25 Cr – 1 Mo – Nb	2.25 Cr – 1 Mo
KNK II	1.4970/1.4981/1.4988	2.25 Cr – 1 Mo – Ni – Nb	2.25 Cr – 1 Mo – Ni – Nb	2.25 Cr – 1 Mo – Ni – Nb
Phénix	type 316 ss	type 316 ss	2.25 Cr – 1 Mo	type 321 Hss
PFR	M 316/FV 548/PE 16	type 321 ss	2.25 Cr – 1 Mo – Nb	9 Cr – 1 Mo
BN 350	Okh 16 N 15 M3B	1 × 18 N 9 (304 ss)	2.25 Cr – 1 Mo	2.25 Cr – 1 Mo
SNR 300	1.4970	1.4948	2.25 Cr – 1 Mo – Ni – Nb	2.25 Cr – 1 Mo – Ni – Nb
BN 600	Okh 18 N 10 T	type 304 ss	2.25 Cr – 1 Mo	18 Cr – 9 Ni
Superphénix	type 316 modified	type 316 ss	2.25 Cr – 1 Mo	Incoloy 800

* IHX = Intermediate heat exchanger.

type 316 modified for the Superphénix]. Short-term and long-term values are markedly increased by approximately 20 % cold forming. Titanium-stabilized steels have improved long-term strength as well. The short-term strength is increased further by irradiation at 400 – 450 °C, whereas above 500 °C the opposite recovery processes come into effect. Reduction of the ductility of the material due to irradiation is, however, especially important in the design. For this reason, the primarily temperature- and stress-dependent creep (resulting from the inner pressure of fission gases in the fuel rod) must be limited to 0.2 %. The time-dependent creep strength under irradiation is of little importance in design of the core element structures because of the low primary stresses. Swelling and irradiation-induced creeping are caused by the extremely fast neutron flux. The phenomenon of swelling is less pronounced in cold-formed steels than in solution-treated materials. The fuel assemblies are designed to control swelling to an extent of 10

4.2.3. Structural Materials for Components

Austenitic and ferritic steels are employed in the components of a fast reactor in both reactor areas and heat-transfer systems (see Table 47). The behavior of both materials in sodium has been successfully proved. Properties that are important for materials selection are strength, corrosion resistance, and the availability of design data. Relevant regulations must be considered.

4.2.4. Application of Materials in Sodium Systems

The *corrosion effects* resulting from the use of materials in sodium are minor and have been determined by extensive experimentation and operating experience. Erosive corrosion is influenced most strongly by the oxygen content of sodium. Other important parameters are the temperature and flow rate of sodium. After an initial phase, the rate of corrosion remains constant during long-term operation under constant operating conditions. No significant differences exist between austenitic and ferritic steels. Allowances are made for corrosion effects by additions to wall thickness that do not interfere with design.

During reactor operation, the sodium systems operate with an oxygen concentration of ≤ 5 ppm by the use of purification systems. In this way, a sufficiently low corrosion rate is achieved (e.g., 4 µm/a at 550 °C with type 316 ss material). Inside the reactor plant, the small reduction in wall thickness caused by erosive corrosion need be considered only in the case of thin-walled components (e.g., cladding tubes of the fuel elements and tubes of the intermediate heat exchangers).

The material removed from the high-temperature zones of the sodium system is deposited at low temperature regions. An example is the transfer of carbon. Since the solubility of carbon in sodium increases with temperature, carbon is dissolved in sodium in zones of higher temperature, transported with the coolant, and deposited in low temperature regions if the equilibrium

concentration is exceeded. This mass transport (*carbon transport*) causes, for example, a reduction in strength of the material in hot zones due to decarburization and deposition of carbides in colder zones.

The *transfer and deposit of radionuclides* are very important for the accessibility of the primary sodium systems to maintenance and repair. These radionuclides are generated by the activation of corrosion products from the cladding and structural materials, as well as activation of the coolant and its impurities; and by fission products from defective fuel rods. Among activated corrosion products, the dominating nuclide is ^{54}Mn ($t_{1/2} = 312$ d), which is deposited predominantly in the cold part of the primary system. Considerably less radioactivity is generated by ^{60}Co ($t_{1/2} = 5.3$ d), which is deposited primarily in the hot area. In addition, both ^{22}Na ($t_{1/2} = 2.6$ d) and ^{65}Zn ($t_{1/2} = 244$ d) contribute markedly to the radioactivity. In case of operation with defective fuel rods, the dominating nuclide is ^{137}Cs ($t_{1/2} = 30.1$ d). Cesium traps are used to remove cesium from sodium.

4.3. Chemical Engineering

4.3.1. Monitoring Sodium Purity

The measurement and control of impurities in sodium are imperative for the operational safety of sodium-cooled fast reactors. Changes in the "normal" concentration of impurities indicate to the plant engineer certain variations or malfunctions in the operation, in time to take the necessary countermeasures. In fact, determination of the impurities in sodium provides information on corrosion, mass transport, damaged fuel rods, and the penetration of air, water, or oil. Above all, the concentrations of oxygen, hydrogen, and carbon must be constantly monitored.

Specific analytical methods are available for determination of each of the impurities in sodium [168–170]. In principle, a distinction must be made between measuring instruments that record continuously during reactor operation and measurements conducted discontinuously, which involve sampling of sodium and subsequent laboratory analysis.

4.3.1.1. Measuring Instruments

Instruments installed in the plant are increasingly used to monitor sodium purity because, unlike chemical analyses, they provide data continuously and are usually very sensitive.

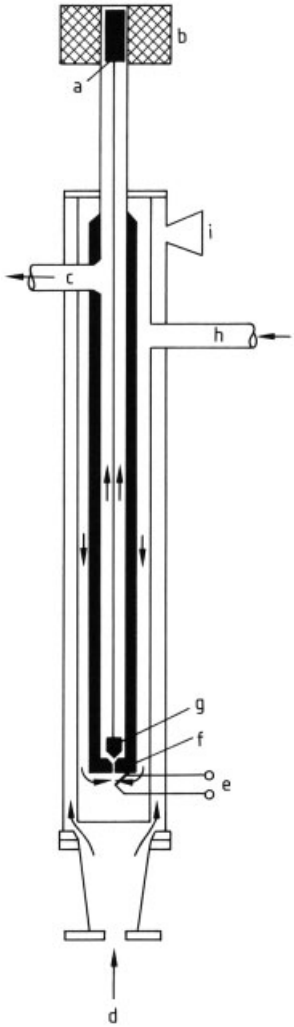

Figure 101. Plugging meter
a) Solid iron; b) Solenoid; c) Sodium outlet; d) Air inlet; e) Thermocouple; f) Orifice; g) Needle; h) Sodium inlet; i) Air outlet

Plugging Meter. Information on the amount of *impurities* that may lead to deposits of reaction products and corrosion effects in the plant, is necessary in order to prevent any disturbance of plant operation. Mainly the amount of oxy-

gen, hydrogen, and carbon is to be monitored. Information is obtained by determining the saturation temperature of the impurities in a sodium bypass, which is led through the orifice of the plugging meter (Fig. 101) at a defined flow rate. In this process, the temperature of sodium is decreased until the saturation concentration of the dissolved impurities is exceeded and they crystallize, plugging the orifice and drastically reducing the flow rate. The temperature at the break in flow rate is defined as the plugging temperature. This is a simple and robust method for the characterization of coolant purity [171]. The plugging temperature is an important input parameter for the operation of the sodium purification systems. The minimum temperature of the cooling systems in operation is maintained at about 20 °C above the plugging temperature.

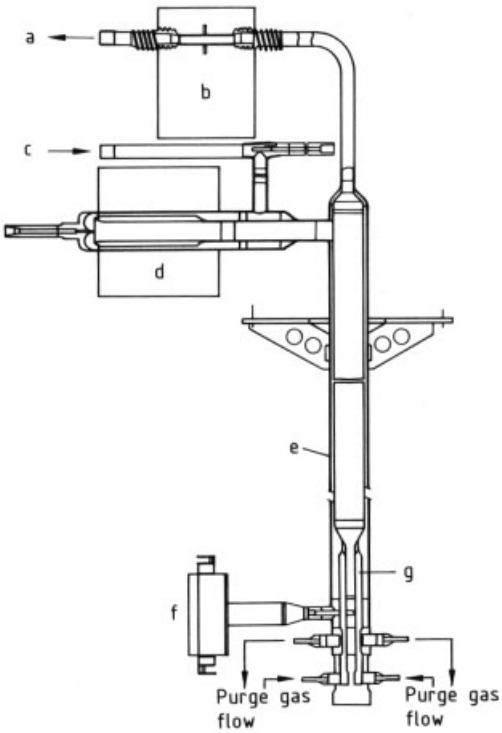

Figure 102. Hydrogen – tritium measuring device
a) Sodium outlet; b) Flow rate meter; c) Sodium inlet; d) Electromagnetic pump; e) Heat exchanger; f) Thermocouple; g) Nickel tubes

Hydrogen Measuring Device. Hydrogen is detected by its diffusion through a nickel membrane. The device used predominantly is an instrument with a nickel finger system dipping into the sodium (Fig. 102). Hydrogen flow through the nickel membrane is measured with either an ionic getter pump or a katharometer using a carrier gas. The lower limit of detection is ca. 0.01 ppm.

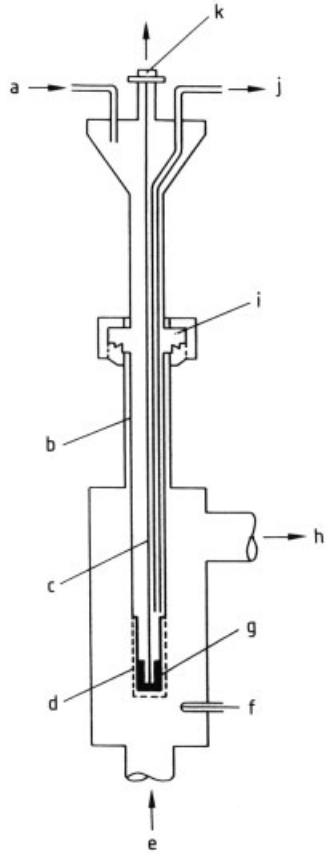

Figure 103. Oxygen meter
a) Reference gas inlet; b) Steel pipe; c) Platinum wire; d) Protective capsule; e) Sodium inlet; f) Thermocouple; g) Solid electrolyte; h) Sodium outlet; i) Screw connection; j) Reference gas outlet; k) Electrical connection

Oxygen Meter. The oxygen dissolved in liquid sodium can be determined by measuring the electromotive force of the oxygen concentration chain:

$$Na_1, [O]_{Na} \left| \begin{array}{c} ThO_2 - Y_2O_3 \\ \text{solid electrolyte} \end{array} \right| \begin{array}{c} O_2 \text{ reference} \\ \text{electrode} \end{array}$$

Various technical models of this measuring device have been developed for use in the plant

(Fig. 103) [172]. The lower limit of detection is < 0.01 ppm.

Carbon Control. The carbon dissolved in liquid sodium is determined by the carbon content of material samples (which either carburize or decarburize) hanging in the sodium. These samples are removed from the sodium system at defined time intervals and analyzed in the laboratory.

4.3.1.2. Chemical Laboratory Analysis

Chemical analyses of the impurities in sodium are employed to measure those impurities that cannot be determined by the operational measuring instruments.

Sodium Sampling. The withdrawal of sodium samples makes great demands on the tightness of the sampling device. Contact with air and the resulting undesirable side reactions must be avoided. Currently, sampling in tube segments [173] or crucibles [174] is used for this purpose. Radioactive samples of sodium from the primary system can be handled only after decay of the ^{24}Na activity (ca. 12 d). Additional contamination of sodium by reaction with air cannot be excluded during transport of the sodium sample from the reactor to the laboratory or during analysis. To avoid reactions with air, preferred devices for the withdrawal of sodium allow direct distilling off of sodium in the sodium cycle [174]. Nonvolatile compounds remaining in the distillation residue are subsequently analyzed [175]. The chemical determination of sodium oxide gives reliable results only if the sodium to be tested is free from particles. If this is not the case, oxygen concentrations measured by this method are usually too high. Volatile compounds can be determined only in sodium samples that have not been distilled. Activation analysis is also employed occasionally.

4.3.2. Control of Cover Gas Purity

The cover gas for sodium surfaces in the cooling systems must have high purity to prevent the interaction of gas impurities with the sodium. The argon cover gas quality is continuously monitored with special gas chromatographs.

4.3.3. Purification of Sodium and Cover Gas

Excessively high concentrations of impurities in sodium can cause malfunctioning of the plant, such as plugging of narrow gaps, corrosion of cladding and structural materials, and transport of radioactivity. Impurities can largely be removed by cold trapping due to their temperature dependent solubility in sodium. A sodium bypass flow enters the cool zone via a regenerative heat exchanger. After the temperature in the cool zone falls below their saturation temperature, dissolved impurities deposit on packings in this zone. A lower operating temperature of 120 °C is technically possible in the cold trap; this corresponds to a residual content of about 1 ppm of oxygen, which is of no importance for corrosion. In addition, packings cause the mechanical deposition of undissolved particles in the cold trap.

Among the radionuclides, ^{131}I and ^{54}Mn (by using nickel packings) can be removed satisfactorily from sodium with the help of cold traps. The nuclides ^{134}Cs and ^{137}Cs can be eliminated by means of a carbon-filled nuclide trap.

Sodium aerosols can be removed from the cover gas with steam traps; fission gases are adsorbed on special activated carbon filters.

4.3.4. Sodium – Water Reactions

During operation of sodium-heated steam generators, the possibility of a sodium – water reaction cannot be completely ruled out. Rates of removal of steam generator materials on the order of several micrometers per second [176] (by a combination of erosion and corrosion) can be determined by experiments in the reaction flame at ca. 1000 °C. With a hydrogen measuring device, even a small leak in the steam generator can be detected early enough to prevent the spread of damage. As soon as a specific critical value for the hydrogen concentration in sodium is reached, the steam generator must be shut off to permit localization and repair of the leak. This alarm value depends on the type and state of the plant.

There are, however, event sequences of a sodium – water reaction that lead to a fast increase of the leakage section by corrosion or even by overheating of neighboring steam generator tubes. In those cases the detection by hydrogen measuring may not be fast enough to trigger the mitigation measures. This is because of the time required for the transport of the reaction products from the leak to the detection device and because of the time it takes to increase the hydrogen concentration in the coolant up to the detection threshold. Acoustic detection devices have been developed in order to analyze the noise spectra of the plant and to identify the specific noise spectrum of the sodium – water reaction. This method eliminates the aforementioned delays. In case of a large leak in the steam generator, the increase in hydrogen pressure is limited by a passive responding pressure relief system (rupture disks). As a back-up, active mitigation measures are triggered (e.g., shut down of the reactor) by detection of the break of the rupture disks.

4.3.5. Sodium Fires

Sodium fires may occur if, as a consequence of a structural material failure (fissure, crack), leakage sodium comes in contact with the ambient atmosphere. Due to the overall plant arrangement and design sodium fires have no radiological consequences and no direct impact on the plant safety. The inherent design principles of sodium plants (low pressure, good material behavior) reduce the risk of sodium leaks. In addition high quality requirements that have been introduced in plant realization, reliable detection systems, and efficient mitigation measures contribute significantly to risk reduction.

Nevertheless, several incidents of this kind occurred in the past. The immediate consequences of sodium fires are energy release, pressure increase in the concerned plant rooms, sodium aerosol release, and chemical reactions with different structures, walls, and system parts. If sodium reaction products are released to the outside of the plant room (which cannot completely be ruled out), the consequences have to be investigated for the plant (e.g., aerosols may be aspired by the decay heat coolers) and for the environment (hazards for human beings, agricultural consequences).

Sodium fires may occur as pool fires, columnar fires, spray fires, or combined fires. The chemical consequences depend on the concerned material, e.g., concrete, steel, insulation material, electrical equipment, and on the temperature produced by the fire event. In pool fires or columnar fires with high leakage sodium flow rates, the temperature is dominated by the operation temperature of the coolant. Another temperature threshold may be the evaporation temperature of the sodium. Accident sequences with high spray fire influence may lead to temperatures higher than 1000 °C. Most of the applied materials and components withstand temperatures in the given order of magnitude only for a limited time. On the other hand the large quantities of sodium aerosols produced by the fire absorb heat radiation and reduce air access to the burning sodium. This limits the propagation of fire. The design goals are:

1) To proof that local destructions in the concerned plant rooms have no consequences for the plant safety. In this framework chemical and thermal resistance of various kinds of material had to be investigated.
2) To investigate consequences of the decay heat removal. This includes investigation of the behavior of the aerosol cloud in the narrow plant environment in order to study the behavior of decay heat removal systems with aerosol loaded atmosphere.
3) To evaluate the aerosol distribution and deposits beyond the plant boundary and potential consequences for humans, plants, and animals.

Particularly the latter aspect has been conscientiously investigated in Germany, France, and Japan. The chemical products of sodium fires are close to those that appear naturally in the environment (all forms of sodium oxide and hydroxide, sodium bicarbonate). The consequences found were not dangerous for the environment, neither theoretically nor practically.

4.3.6. Purification and Decontamination of Components

Components that are wetted with sodium are generally cleaned by using the *steam – nitrogen*

process. In this process the components are heated in nitrogen and then exposed to increasing amounts of steam, which reacts with sodium. Sodium is then withdrawn as an aqueous sodium hydroxide solution. This process is carried out under controlled conditions below 150 °C to avoid corrosion.

For parts with a complicated geometry (narrow gaps, blind holes) the *evaporation of sodium under vacuum* frequently gives better results than the steam – nitrogen method. *Cleaning with alcohols* (methanol, ethanol, isopropanol, etc.) is used only for small parts because of the risk of fire. A higher diethyleneglycol monoethyl ether has been used for this purpose. It is considerably less dangerous than the lower alcohols and can even be heated to improve its cleaning action [177]. Decontamination of the components is always carried out after the removal of sodium. Normal decontamination agents are employed for this purpose. "Soft agents" are used mainly to reduce the danger of corrosion because they are mixtures of mild acids and detergents.

5. Fuel Cycle

5.1. Introduction

In contrast to electricity generation by fossil-fueled power stations, operation of a nuclear reactor requires that the fuel undergoes a cycle in order to make optimum use of the energy carriers, i.e., the fissile isotopes of uranium (^{235}U) and plutonium (^{239}Pu and ^{241}Pu). Spent nuclear fuel elements must be removed from the reactor before all the fissile nuclei contained in them have been fissioned, or "burned," essentially for three reasons: (1) The concentration of fissile nuclei in the fuel decreases due to fission, which lessens the chance for a neutron to hit another fissile nucleus and thus to propagate the fission chain; in other words, the reactivity of the fuel decreases with increasing burnup. (2) Some of the fission product isotopes are strong neutron absorbers, which decrease the fuel reactivity even further. (3) The mechanical strength of the fuel element's structural materials decreases during operation due to radiative, thermal, and hydraulic stresses, thus limiting its mechanical lifetime.

The following rule-of-thumb estimate is aimed at obtaining an idea of the magnitude of the problem. Fresh light-water reactor fuel is enriched to roughly 3.5 % fissile ^{235}U. After a burnup of 35 000 MW · d/t, the fuel, which is now due for removal, still contains ca. 0.9 % ^{235}U, plus ca. 0.6 % fissile plutonium isotopes created by neutron absorption of ^{238}U. The total of ca. 1.5 % fissile isotopes represents more than 40 % of the original fissiles content in the fresh fuel. To make the residual energy carriers available, they must be separated from the spent fuel and recycled into new fuel elements.

Stages of the Fuel Cycle. Figure 104 schematically shows the various stages in the fuel cycle of the light-water reactor. The cycle starts with the *mining* of uranium ore from which uranium is extracted and converted into " yellow cake," U_3O_8. The oxide is further *converted* into uranium hexafluoride (UF_6), a readily volatile compound that is suitable for *isotope enrichment* in a multistage gas separation process such as gaseous diffusion or gaseous centrifugation. In this stage, the ^{235}U isotopic concentration is enriched from its natural 0.72 % to about 3 – 3.5 %. The depleted uranium fraction contains ca. 0.25 % ^{235}U and is excreted from the cycle for eventual reuse in a breeder reactor. Enriched uranium hexafluoride then passes to the *fuel element fabrication plant* where it is converted into uranium dioxide (UO_2), which is pelletized, filled into Zircaloy tubes, and assembled. The fuel assemblies are loaded into the reactor for power generation by neutron-induced fission of the fissile isotopes.

At the end of their useful life, the spent fuel assemblies are unloaded from the reactor and transferred to an on-site storage pond for intermediate storage, or "cooling". This allows the radioactive fission products to decay according to their specific half-lives. The short-lived fission product isotopes, in particular, decay totally during on-site storage, which reduces both the radiation level and the decay heat output substantially. The "cooled" fuel assembly is then loaded into an air-cooled, shielded container and transported to the chemical reprocessing plant, where it is cooled further either under water after unloading into a storage pond, or by air cooling in a dry storage facility. The total cooling period of LWR fuel elements before reprocessing is usu-

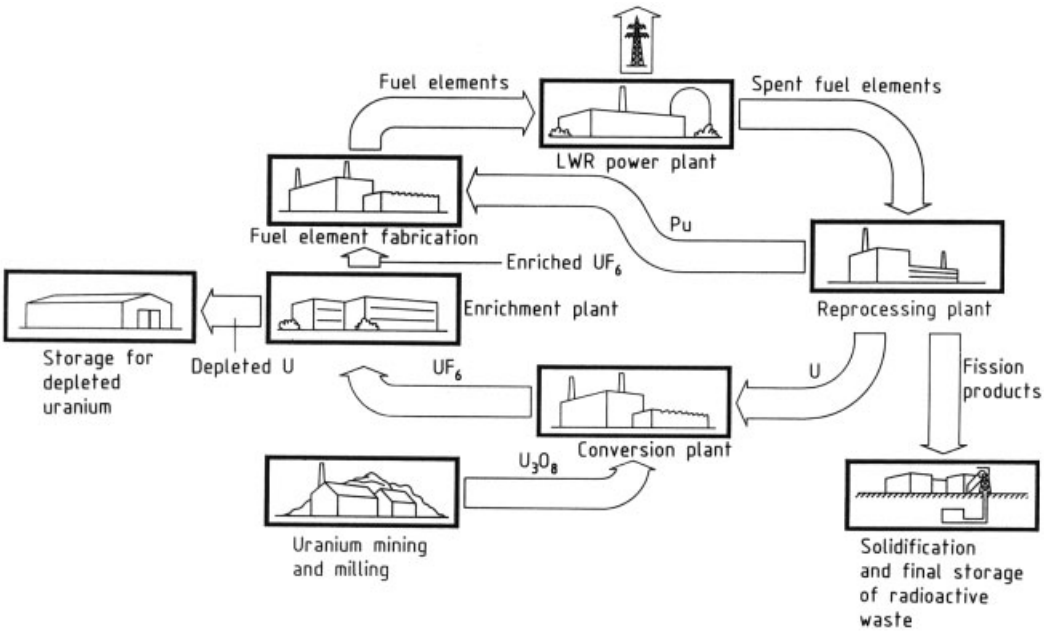

Figure 104. The LWR fuel cycle

ally between three and seven years because a fast recycle of the fissile materials is not essential in the LWR.

The main task of the *reprocessing plant* is to separate and purify the residual energy carrier materials (uranium and plutonium) from the fission products. This is done by the PUREX process, which yields the metal values in the form of concentrated solutions of their nitrates, namely $UO_2(NO_3)_2$ and Pu $(NO_3)_4$, respectively. At the same time the bulk of the fission products are collected as a concentrated "high-active waste" (HAW) nitrate solution and transferred to a *waste solidification plant* and then to a *final storage site*. The uranium product is recycled for conversion and reenrichment, whereas the plutonium product is recycled for conversion and fabrication into a (U, Pu)O_2 mixed-oxide (MOX) fuel.

Fuel Consumption of Different Types of Fuel Cycles. The net effect of the LWR fuel cycle is illustrated in Figure 105 [178] which details the consumption of uranium for the generation of 1 gigawatt-year (GW · a) of electric energy (1 GW · a is roughly the electrical energy produced by a standard 1300-MW power reactor during one year of operation, if an overall load factor of ≈ 80 % is assumed). The upper line illustrates the "throwaway" operation; i.e., spent fuel is not reprocessed and recycled, but instead is disposed in the deep underground as a radioactive waste material. In this case, ca. 205 t of natural uranium is required, which must first be converted into 37 t of 3 % ^{235}U-enriched uranium fuel. The balance of 168 t of depleted uranium is left behind as a waste material that cannot be used further in thermal reactors. Although it could be used in fast breeder systems, these cannot be started because the plutonium needed is not recovered in the throwaway operation mode. Of the fuel uranium, 1.2 t is fissioned in the reactor to produce 1 GW · a of electric energy; the spent fuel, composed of roughly 35 t of uranium (residual enrichment 0.9 %), 0.3 t of plutonium (65 % fissile), and 1.2 t of fission products, is wasted.

The middle line of Figure 105 illustrates the light-water reactor with fuel recycle; the figures given describe the recycle equilibrium state. In this case, the natural uranium requirement for 1 GW · a of electric energy is 120 t, a saving of ca. 40 % compared to the throwaway case. Again, a major part of the material is lost in the form of depleted uranium, but the only other ma-

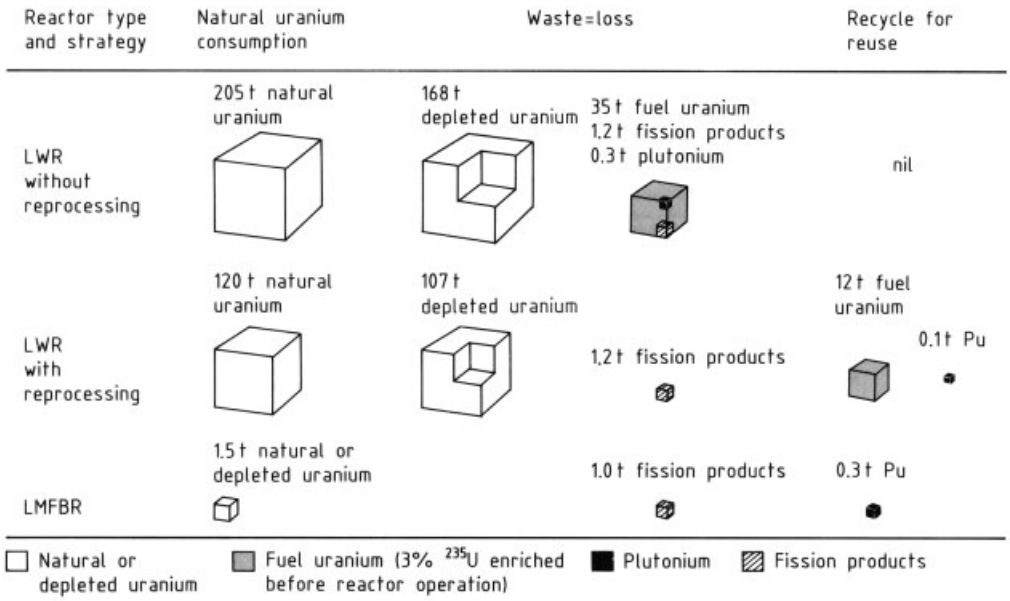

Figure 105. Uranium consumption for generation of 1 GW · a of electricity by light-water reactors with or without fuel recycle, and by fast breeder reactors with inherent fuel recycle [178]

terial loss is the 1.2 t of fission products produced during reactor operation. Both the spent fuel uranium and the plutonium produced by neutron absorption are recycled for additional energy production. If needed, the recovered plutonium can be used as the starting fuel for fast breeder reactors. In this case the stock of depleted uranium could be used as the blanket material for breeding new fuel.

The fuel-saving potential of the fast breeder reactor is illustrated in the lower line of Figure 105. The enormous superiority of this reactor type in terms of fuel economy is due to the fact that it "breeds" more fissile plutonium from nonfissile uranium-238 than it consumes for fission and energy production (see Chap. 4). Consequently, no external supply of fissile isotopes is required after the FBR has reached its recycle equilibrium, and the only material to be replaced is the uranium-238 that is converted into plutonium-239 and higher isotopes. In equilibrium, with inevitable recycle losses taken into account, this reload requirement is roughly 1.5 t of natural or depleted uranium per gigawatt-year of electric energy, less than 1 % of the throwaway LWR. The FBR generates fewer fission products than the LWR, ca. 1 t per GW · a, because it produces steam at a distinctly higher temperature (ca. 500 °C vs. 280 °C) and consequently achieves a higher thermodynamic conversion efficiency (ca. 40 % vs. 32 %). The excess production of plutonium ("breeding gain")/given as 0.3 t/GW · a in Figure 105, but adjustable within wide limits by varying reactor design details, can be used for the start-up of additional FBRs if required. The FBR fuel cycle (see Fig. 106) eliminates the need for uranium enrichment and, hence, the loss of depleted uranium, which is primarily responsible for the high uranium demand of thermal reactor systems.

Nevertheless, with the present oversupply and consequent low prices of uranium, the cost of electricity generation is more favorable for the light-water reactor than for the fast breeder reactor. A French estimate claims a cost difference of some 15 % for the two under current uranium price conditions [179]. This difference in cost is partly due to the reactor itself but also to the higher cost of both FBR fuel element fabrication and spent fuel reprocessing. For the long term, therefore, the question of uranium resources and its relation to demand remains to be answered.

Uranium Resources. Surprisingly enough, uranium is not a particularly rare element. Its average concentration in the earth's crust is 2.1 g/t

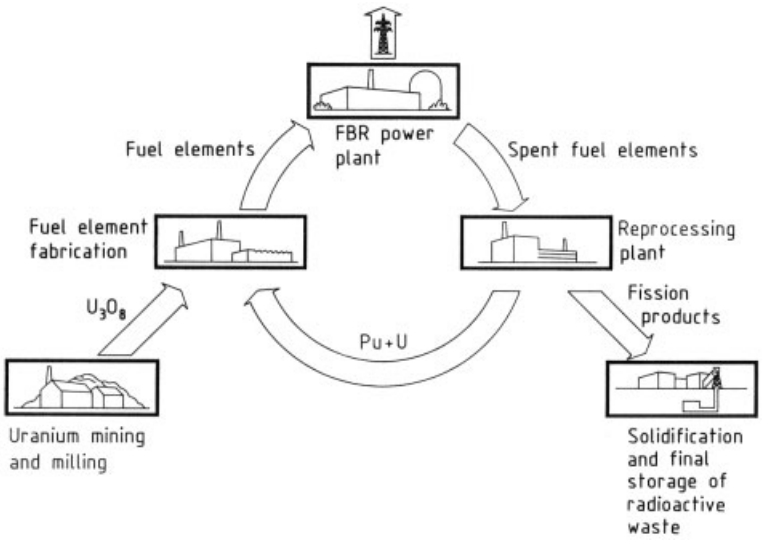

Figure 106. The FBR fuel cycle

[180], making it more abundant than metals such as cadmium (0.15 g/t) or silver (0.1 g/t). In primary rock, its concentration is generally much higher in acid rock such as granites (4.8 g/t) than in basic rock such as basalts (0.6 g/t). Its average concentration in seawater is 0.3 mg/m^3. The total amount of uranium in the earth's crust down to a depth of 20 km has been estimated as 5×10^{13} t, the ocean waters are estimated to contain 5×10^{9} t. Unfortunately, however, due to its chemical properties, uranium does not normally tend to accumulate in large deposits. As a result, most of it is more or less evenly distributed over the earth's surface at the above-mentioned concentration, which obviously does not warrant economical recovery.

The very small fraction of terrestrial uranium that occurs in deposits concentrated enough, large enough, and accessible enough, to make recovery economical, is referred to as "uranium resources." These are quite well-known because uranium is easily detected and analyzed by virtue of its radioactive properties, and they are listed in *Uranium Resources, Production and Demand*. This report, published jointly by the Organization for Economic Cooperation and Development (OECD) and the International Atomic Energy Agency (IAEA) and known as the " Red Book", is periodically updated, the latest available edition at the time of writing having been issued in 1989 [181]. Based on national reports provided by the participating countries, it presents data for the (then) noncommunist world, referred to as the "world outside centrally planned economic areas" (WOCA). Unfortunately, figures from the (then) communist countries, notably the former Soviet Union, China, former East Germany, and former Czechoslovakia, which presumably have major resources, are not available thus far, but this situation can be expected to change in the future. Nevertheless, the accessible data draw a picture clear enough to allow conclusions to be made.

Categories of Resources. The Red Book defines, in decreasing order of certainty of knowledge, four categories of resources:

Reasonably Assured Resources (RAR) refers to uranium that occurs in known mineral deposits of delineated size, grade and configuration such that the quantities which could be recovered within the given production cost ranges with currently proven mining and processing technology, can be specified. Estimates of tonnage and grade are based on specific sample data and measurements of the deposits and on knowledge of deposit characteristics. Reasonably Assured Resources have a high assurance of existence.

Estimated Additional Resources–Category I (EAR-I) refers to uranium in addition to RAR that is inferred to occur, mostly on the basis

of direct geological evidence, in extensions of well-explored deposits or in deposits in which geological continuity has been established but where specific data, including measurements of the deposits, and knowledge of the deposits' characteristics are considered to be inadequate to classify the resource as RAR. Estimates of tonnage, grade and cost of further delineation and recovery are based on such sampling as is available and on knowledge of deposit characteristics as determined in the best known parts of the deposit or in similar deposits. Less reliance can be placed on the estimates in this category than on those for RAR.

Estimated Additional Resources– Category ***II*** (EAR-II) refers to uranium in addition to EAR-I that is expected to occur in deposits for which the evidence is mainly indirect and which are believed to exist in well-defined geological trends or areas of mineralisation with known deposits. Estimates of tonnage, grade and cost of discovery, delineation and recovery are based primarily on knowledge of deposit characteristics in known deposits within the respective trends or areas and on such sampling, geological, geophysical or geochemical evidence as may be available. Less reliance can be placed on the estimates in this category than on those for EAR-I.

Speculative Resources (SR) refers to uranium, in addition to Estimated Additional Resources–Category II, that is thought to exist, mostly on the basis of indirect evidence and geological extrapolations, in deposits discoverable with existing exploration techniques. The location of deposits envisaged in this category could generally be specified only as being somewhere within a given region or geological trend. As the term implies, the existence and size of such resources are speculative.

In addition, the Red Book classifies resources according to their estimated recovery cost into three categories: (1) up to \$ 80/kg of uranium, (2) \$ 80 – 130/kg of U, and (3) \$ 130 – 260/kg of U. Practically all reported resources of types RAR and EAR-I are in the lower cost categories (1) and (2). In contrast, only a limited number of national reports have been prepared for speculative resources, and these were either for recovery cost categories up to \$ 130/kg and \$ 130 – 260/kg or, in some cases, did not differentiate at all among cost levels.

Furthermore, the Red Book defines the combined RAR and EAR-I resources of recovery cost below \$ 130/kg of U as *known conventional resources*, whereas the combined EAR-II and SR categories are termed *undiscovered conventional resources*, indicating the degree of confidence that the respective estimates deserve. All uranium presently mined is in the cost range below \$ 80/kg of U, and in fact far below this mark, as spot market prices (not cost) of roughly \$ 26/kg of U (June 1989) [182] indicate.

From Table 48, a total of ca. 3.4×10^6 t of known conventional uranium resources in 1989 can be inferred for WOCA [181]. Estimates given for the high cost range (\$ 130 – 260/kg U) of RAR and EAR-I were a total of roughly 0.6×10^6 t (RAR: 0.400×10^6 t; EAR-I: 0.185×10^6 t) and, when included in the account, do not essentially change the picture. Undiscovered conventional resources (Table 49) for production costs up to \$ 260/kg, as described in national reports, were between ca. 8.1 and 9.5×10^6 t of uranium, and mainly in the "speculative" category, but only a few national authorities have submitted speculations on undiscovered resources. However, in the International Uranium Resources Evaluation Project (IUREP) carried out by OECD and IAEA in 1977 and revised in 1983, the range of speculative resources at recovery costs up to \$ 130/kg of uranium for 169 countries in WOCA was believed to be between 6.4 and 16×10^6 t of uranium, with the most likely range between 9.6 and 12.1×10^6 t [181].

Table 48. Known conventional resources (in 10^6 t of uranium) for the world outside centrally planned economic areas (WOCA), January 1989 [181]

Resource	Cost range		Total
type ∗	≤ 80/kg U	\$ 80 – 130/kg U	≤ \$ 130/kg U
RAR	1.546	0.655	2.201
EAR-I	0.773	0.391	1.164
Total	2.319	1.046	3.365

∗ RAR = reasonably assured resources; EAR-I = estimated additional resources– category I.

Unconventional Resources. In addition to the conventional resources considered so far, a total of roughly 7×10^6 t of uranium in unconventional (i.e., byproduct) resources has been reported in a few national reports evaluated in the Red Book [181]. By far the most important source in this category is uranium occurring in

marine phosphate rock deposits, notably those in Morocco, where a total of 6.5×10^6 t of uranium is claimed to exist at an average grade of 120 g of uranium per tonne. This uranium might become economically recoverable as a byproduct, i.e., at a production rate determined by the rate of phosphoric acid production from such material, provided that the extra recovery cost can be reduced to competitive levels.

Table 49. Undiscovered conventional resources (in 10^6 t of uranium) for WOCA, January 1989 [181]

Resource	Cost range		Total
type ∗	≤ \$ 130/kg U	\$ 130 – 260/kg U	≤ \$ 260/kg U
EAR-II	1.686	0.649	2.335
SR			5.813 – 7.113
Total			8.148 – 9.448

∗ EAR-II = estimated additional resources– category II; SR = speculative resources.

Uranium Resources and Demand–Future Trends. To relate resources and demand, a very simple approach is made here on the following basis: (1) The resource base is restricted to known conventional resources (3.4×10^6 t U for WOCA in 1989). This condition seems justified in view of the high degree of uncertainty regarding the existence, size, and recovery costs of undiscovered resources, and of the presumably limited production rate of uranium from byproduct resources. (2) Nuclear energy production in WOCA is assumed to increase linearly from its present annual level of ca. 165 GW · a (OECD 1989 [183]; the OECD figure is essentially the same as that for WOCA) to 330 GW · a per year, within 25 years (i.e., with the same average rate as in the first 25 years of nuclear power utilization), and then to remain at this level. This is probably a rather pessimistic assumption, in view of the expected worldwide increase in energy demand and the anticipated requirement limit the combustion of carbon-based fossil fuels because of their environmental impact ("greenhouse effect"). (3) No limitations in uranium availability due to restricted or delayed construction of mining, recovery, enrichment, fuel fabrication, fuel reprocessing, or waste treatment and disposal facilities are expected. With these conditions and the data given in Figure 105, the operational range of WOCA's known conventional resources would be estimated to be:

light-water reactors operated with fuel throwaway	ca. 60 years
light-water reactors operated with fuel recycle	ca. 90 years
fast breeder reactors	ca. 6800 years

For a more reliable forecast, much better knowledge of future trends in world energy demand, supply structures, resource availability, production capacities, prices, and political developments is obviously required. Despite the inherent imponderables many scenarios have been developed and analyzed, and the results have been presented and discussed at World Energy Conferences which are held periodically as an international forum on energy matters. The majority of such scenarios, at least the more serious ones, arrive at results surprisingly close to the primitive rule-of-thumb approach presented here; see, e.g., Beckurts et al. [184] who, on the basis of a very detailed analysis, arrive at an operational range of < 50 years for LWRs without fuel recycle.

The conclusions that must be drawn from this situation are, in principle, self-evident. At the present stage of technical development, human energy requirements are predominantly satisfied by the combustion of carbon-based fossil fuels and by the fission of heavy nuclei. Both are limited in their operational range to at best a few centuries, with the exception of nuclear fission in fast breeder reactors which has the potential to cover several millennia or more if low-grade uranium resources become economical for this low-consumption technique. Other available technologies will contribute only small increments to the world's growing energy demand. As long as no fundamentally new large-scale energy generation process is developed to the stage of technical and economic application, nuclear fission in fast breeder reactor systems is the only long-term energy source available to humans (where "long-term" means, e.g., $> 10^3$ years). Therefore, wasteful techniques must not be used in the consumption of uranium. This is the reason why legal provision has been made in a number of countries to prevent the waste of fissile raw materials. The German Atomic Law [185], for example, incorporates the so-called reutilization requirement (*Wiederverwertungsge-*

bot), which stipulates that spent nuclear material containing fissile residues must be recycled for reuse as long as technically possible and economically feasible.

Feasibility of the Fuel Cycle. The technical feasibility of all stages of the nuclear fuel cycle is demonstrated by the successful operation of large-scale industrial facilities in various countries, with one important exception. Permanent storage facilities for unretrievable disposal of high-level radioactive waste products in deep geologic repositories have not so far been put into operation; instead, these are still in the research or planning stage, while HAW materials are being stored in interim surface storage plants. Nevertheless, permanent disposal facilities are expected to be built and operated successfully.

Fuel Cycle Economics. What remains to be answered is the question of fuel cycle economics. For this purpose the results of a very careful analysis in 1985 by the OECD Nuclear Energy Agency (NEA) Working Group on the Economics of the Nuclear Fuel Cycle [186], and an update based on French industrial experience [187], are discussed. An important feature of the OECD/NEA model is that the complete reactor lifetime, not a single reload, is considered, and that the costs and payment schedules for all parts of the fuel cycle are precisely defined.

Table 50 presents a breakdown of the fuel cycle reference costs for pressurized-water reactors operated with or without fuel recycle, as evaluated by the OECD/NEA [186]. Total fuel cycle reference costs were estimated to be 8.56 millidollars (or "mills", 1 m\$ = 0.1 cent) per kilowatt-hour for fuel recycle and 7.78 m\$/kW · h for the throwaway case, i.e., a cost advantage of 0.78 m\$/kW · h, or 10 %, for the latter. However, the absolute value of 0.78 m\$/kW · h makes clear that this cost difference is marginal compared to the total electricity generation cost, which is on the order of 50 m\$/kW · h, being determined by the investment and operating costs of the reactor itself which account for about 80 % of the total costs of nuclear electricity generation.

To account for uncertainties in the assumptions made for different fuel cycle components and parameters, a thorough *sensitivity analysis* was also carried out. The total fuel cycle costs were found to be highly sensitive to uranium prices but only slightly sensitive to major changes in the costs of uranium conversion, fuel transport and storage, reprocessing, and waste treatment and disposal. Based on sets of reasonable variations of parameters, a "likely range" of fuel cycle costs was evaluated to be between 7.1 and 10.4 m\$/kW · h for fuel recycle and between 6.3 and 9.7 m\$/kW · h for the throwaway option.

A reevaluation by Couture et al. [187] using the OECD/NEA methodology, but applying real reprocessing costs as experienced by the French COGEMA (Compagnie Générale des Matières Nucléaires), has led to total PWR fuel cycle costs of 7.6 m\$/kW · h for fuel recycle and 8.2 m\$/kW · h for the throwaway case, i.e., a slight cost advantage of 0.6 m\$/kW · h (8 %) for recycle. The COGEMA values are well within the "likely cost range" as determined by the OECD experts, and the cost difference is, in absolute terms, again marginal. Any increase of the (highly sensitive) uranium price beyond the present low level will, however, shift the cost advantage further in favor of fuel recycle. Moreover, progress in recycle technologies with inherent cost reductions (for which ample room exists) will have the same effect.

Conclusions. Based on these facts, the following rules can be formulated for a responsible application and management of nuclear power, taking into account both the present state of, and the long-term prospects for, the uranium supply:

1) Operation of nuclear power reactors with permanent disposal of the spent fuel is a wasteful technique that does not correspond to the state of scientific, technological, and economic development and is therefore inadequate. Specifically, this mode of operation would presumably limit the utilization of nuclear energy to less than a century because it would prevent the eventual transition to fast breeder technologies.
2) For the present period of high uranium supply and low prices, the operation of thermal power reactors with fuel recycle is adequate as long as this is more economic than the use of fast breeder reactors.
3) Provided no fundamentally new large-scale energy source can be developed and employed in time, the operation of fast breeder

Table 50. Breakdown of PWR fuel cycle component costs, reference case (source: OECD/NEA 1985) [186]

	Reprocessing cycle		Fuel throwaway ∗	
	m$/kW · h	%	m$/kW · h	%
Uranium	3.48	40.7	3.48	44.7
Conversion	0.17	2.0	0.17	2.2
Enrichment	2.28	26.6	2.28	29.3
Fuel fabrication	0.88	10.3	0.88	11.3
Subtotal for front end	6.81	79.6	6.81	87.5
Transport of spent fuel	0.14	1.6	0.14	1.8
Storage of spent fuel	0.17	2.0	0.65	8.4
Reprocessing and vitrification	2.18	25.5		
Fuel conditioning and disposal			0.18	2.3
Waste disposal	0.08	0.9		
Subtotal for back end	2.57	30.0	0.97	12.5
Uranium credit	−0.54	−6.3		
Plutonium credit	−0.28	−3.3		
Subtotal of credits	−0.82	−9.6		
Total costs	8.56	100.0	7.78	100.0

∗ m$ = millidollar = 0.1 cent.

reactors will eventually become economically competitive due to a decrease in uranium supplies and consequent increase in prices. Over the long term, the application of fast breeder recycle techniques will constitute an absolute necessity under the circumstances assumed because of its very low uranium demand.

5.2. Uranium Production, Conversion, and Enrichment [188–190]

In contrast to other primary energy sources such as coal, natural gas, and petroleum, uranium is currently used only as a fuel for the production of electricity in nuclear power plants. Future nuclear plants might generate not only electrical power but also process heat for the production of secondary energy sources, such as methane obtained by coal gasification. One special application of nuclear energy is the propulsion of ships and submarines.

5.2.1. Occurrence and Classification of Deposits

The continental crust contains relatively high concentrations of uranium. The average abundance is 2.1 ppm. Naturally occurring uranium is a mixture of three isotopes: ^{238}U (99.28 %), ^{235}U (0.71 %), and ^{234}U (0.006 %). Geochemically, uranium is a lithophilic element, as shown by its high affinity for oxygen. Consequently, uranium has the ability to form many complexes that are stable in aqueous solution over wide pH ranges. This ability is responsible for the transport of uranium and its mobility in geologic formations.

Uranium mineralization occurs in nearly all rocks of the earth's crust. Uranium ore deposits can be classified in seven main types with a number of subtypes. Three principal types are of particular economic importance: the unconformity, conglomerate, and sandstone types. Deposits of these types are associated predominantly with certain epochs in the earth's history and often occur in geographic provinces (see Table 51).

The great economic importance of *unconformity deposits* is due to the size of individual deposits and their often very high average ore grade (2 – 10 % U_3O_8, sometimes up to 15 % U_3O_8). These are the most significant deposits at present, accounting for some 27 % of currently known world reserves and more than 30 % of world production. The latter figure will increase as lean-ore deposits such as Canada's Elliot Lake region become exhausted or uneconomical.

Conglomerate deposits are the oldest known deposits. They are mined in only two regions, i.e., the Witwatersrand area in South Africa and the Elliot Lake area in Canada, Uranium is the

Table 51. Uranium ore deposits

Type	Characteristics	Age	Example	Mean U_3O_8 content, %
Unconformity	unconformity between metamorphic rocks (basement) and nonmetamorphic clastic sediments; monomineralic, polymetallic	Proterozoic	Key Lake, Canada	2.5
			Cigar Lake, Canada	10.0 – 14.0
			Jabiluka, Australia	0.30
			Ranger, Australia	0.25
Conglomerate	gold – pyrite-bearing quartz conglomerates	Late Archean – Early Proterozoic	Witwatersrand, South Africa	0.025 – 0.05
			Elliot Lake region, Canada	0.1 – 0.15
Sandstone	stratabound ore bodies; permeability-dependent ore bodies unconformable with beds; basal channel casts	Upper Paleozoic, Mesozoic	Colorado Plateau, United States	0.1 – 0.3
			Grants Mineral Belt, United States	0.1 – 0.3
			Wyoming and Texas basins, United States	0.1 – 0.3
			Akouta, Niger	0.5

primary mineral product in Canada, whereas in South Africa it is a byproduct of gold mining. Because of its low ore grade (0.02 – 0.15 % U_3O_8) the importance of this deposit type will continue to decline; its share is expected to fall from 19 % of world production in 1989 to barely 17 % in 1991.

Sandstone uranium deposits are the most common worldwide, occurring in virtually every epoch of geologic history. At the beginning of the 1980s, this type accounted for some 40 % of world reserves. With the falling price of uranium, the significance of these deposits has decreased sharply because of their low ore grades (0.1 – 0.5 % U_3O_8). Sandstone deposits constituted about 20 % of the recoverable reserves worldwide at the end of 1988 (up to $ 80/kg U) and represented a similar share of world production. The drastic cut in production from these deposits is particularly marked in the United States. More than 40 mines were operating in the Grants Minerals Belt at the end of the 1970s, but conventional uranium mining in sandstone deposits has since come to a complete halt. The last operating mine, Mt. Taylor, closed in late 1989.

Other deposit types are currently of secondary importance in overall production. This situation may change if larger and more economic deposits are discovered. One example is the Roxby Downs deposit in Australia, which is a polymetallic iron ore deposit with economically exploitable quantities of copper, uranium, gold, and silver. The deposit is a brecciated ore body of the Proterozoic age, consisting mainly of granite fragments. Intensive exploration in regions not yet heavily examined may increase the significance of other deposit types.

5.2.2. Production

Prospecting and Exploration. The search for possible deposits (*prospecting*) begins with the evaluation of large areas in accordance with the most recent findings of earth sciences. Geophysical and geochemical methods, in particular, have increased in importance over the past 20 years for prospecting purposes. When an area with large anomalies has been localized, it is then *explored*; i.e., the exact size and extent of the ore bodies are determined, along with the petrographic and mineralogic compositions. Detailed exploration ends with a *feasibility study*, which considers the economics of recovery. Once the feasibility study has been completed and assessed, a decision on implementing the project is made. The time required from prospecting to the start of mining and ore processing is 10 – 15 years: prospecting requires 3 – 5 years; exploration 2 – 3 years; and the approval and construction phase 5 – 8 years [191]. The lifetime of a project, i.e., the actual mining and ore processing may vary from a few months to several decades.

Processing. Because of their relative abundance, only 12 of more than 120 naturally occurring uranium minerals are mined and processed into commercial uranium concentrate (see Table 52).

Uranium ore from the mine is *presorted* by composition into uranium-bearing ore and bar-

Table 52. Important uranium minerals [192]

Mineral	CAS registry no.	Chemical composition
Oxides		
Uraninite pitchblende	[*1317-99-3*]	$(U^{+4}_{1-x}, U^{+6}{}_{x})O_{2+x}$
Becquerelite	[*1317-75-5*]	$7\,UO_3 \cdot 11\,H_2O$
Mixed oxides		
Brannerite	[*12197-37-4*]	$(U, Ca, Fe, Th, Y)\,(Ti, F)_2O_6$
Davidite	[*12173-20-5*]	$(Fe, Ce, U, Ca, Zr, Th)_6(Ti, Fe, V, Cr)_{15}(O, OH)_{36}$
Pyrochlore	[*12174-36-6*]	$(Na, Ca, Fe, U, Sb, Pb, Th, Zr, Ce, Y)_2-(Nb, Ta, Ti, Sn, Fe)_2O_6(O, OH, F)$
Silicates		
Coffinite	[*14485-40-6*]	$U(SiO_4)_{1-x}(OH)_{4x}$
Uranophane	[*12195-76-5*]	$Ca(UO_2)_2(SiO_3)_2(OH)_2 \cdot 5\,H_2O$
Phosphates		
Autunite	[*16390-74-2*]	$Ca(UO_2)_2(PO_4)_2 \cdot 10-12\,H_2O$
Torbernite	[*26283-21-6*]	$Cu(UO_2)_2(PO_4)_2 \cdot 12\,H_2O$
Vanadates		
Carnotite	[*60182-49-2*]	$K_2(UO_2)_2(VO_4)_2 \cdot 1-3\,H_2O$
Tyuyamunite	[*12196-95-1*]	$Ca(UO_2)_2(VO_4)_2 \cdot 5-8\,H_2O$
Hydrocarbons		
Thucholite	[*12178-52-8*]	uraninite, complexed with hydrocarbons

ren rock. Ore separation is basically performed to reduce operating costs. At most ore concentration plants, presorted uranium ore is milled in crushers to obtain fine ore with grain sizes < 10 mm. A subsequent grinding step might follow to yield a fine powder.

Extraction includes leaching of crushed uranium ore with acid (sulfuric, hydrochloric, nitric, or phosphoric) or alkaline solvents (carbonates or hydrocarbonates). The choice of solvent depends mainly on the chemical composition of the uranium mineral and the accompanying rock.

After leaching, solids and leach liquor are separated by *flocculation* and *filtration*. The uranium is precipitated from the solution, usually as a diuranate.

The uranium concentrate obtained in this way is called " yellow cake" because of its color. It generally contains > 70 % uranium.

The time required for subsequent purification, refining, enrichment, and fuel fabrication steps varies greatly and depends on many factors such as the reactor type and the stockpiling policies of the power plant operator.

A rough overview of typical times is given below:

From ore preparation to U_3O_8 (yellow cake)	1 – 6 months
From U_3O_8 concentrate to UF_6	2 – 6 months
From UF_6 to enriched UF_6	6 months
From enriched UF_6 to UO_2	3 months
From UO_2 to fuel assembly	3 – 6 months
From fuel assembly to reactor fueling	0 – 3 months

5.2.3. Output and Demand

Uranium *production* rose steadily from 1974 to 1980. A consolidation phase followed in 1980 to 1982; since that time, output in the Western world mining countries has declined steadily (see Table 53). The lowest level was reached in 1990. The decline in output between 1980 and 1990 resulted from shutdowns of uneconomical uranium mines, political sanctions, and reduced demand. Output will again increase slightly until 2000, but will not reach the 1980 high.

In 1983, uranium *demand* for nuclear power plants exceeded production for the first time. The shortfall has since been made up from power plant operators' existing inventories, but this resource will exist for only a limited time.

Set against the cumulative reactor demand of some 560 000 t of uranium between 1990 and 2000 are reserves of 2.3×10^6 t U (categories RAR and EAR-I recoverable up to \$ 80/kg U); not included are reserves in China and the Eastern Bloc. Even if nuclear energy were to grow faster than expected, a dependable supply would be available until 2030.

5.2.4. Conversion

Conversion is the intermediate process between yellow cake production and enrichment. To in-

Table 53. Uranium production and reactor demand for WOCA in tonnes of uranium (1975 – 2000)

Country	Year					
	1975	1980	1985	1990	1995 ∗	2000 ∗
Africa	4 725	15 330	12 390	9 280	9 450	9 350
Australia		1 560	3 250	3 520	4 550	4 550
Europe	2 155	2 980	3 510	3 240	4 060	3 160
Canada	3 565	7 150	10 880	8 700	13 410	12 410
United States	8 370	15 470	3 980	3 550	5 050	5 140
Other	100	300	565	1 620 ∗∗	2 000 ∗∗	2 450 ∗∗
Production total	18 915	42 790	34 575	29 910	38 520	37 060
Reactor demand	15 500	21 500	39 900	45 800	50 700	55 200

∗ Estimated.
∗∗ Including shipments by the People's Republic of China and the former Soviet Union.

crease the ^{235}U content in natural or reprocessed uranium, a volatile uranium compound is required in the enrichment step. Uranium hexafluoride (UF_6) is the only uranium compound that is suitable for enrichment because of its thermal stability and relatively high volatility. All current enrichment processes, therefore, employ uranium hexafluoride.

Uranium hexafluoride is produced from uranium concentrate by two processes, wet and dry. A schematic of the conversion processes is given in Figure 107.

Wet Process. In the wet or aqueous process, the uranium concentrate is first dissolved in nitric acid and purified by solvent extraction with tributyl phosphate (TBP), diluted with kerosene. Treatment with ammonia precipitates ammonium diuranate [$(NH_4)_2U_2O_7$, ADU]. This product is dried, calcinated, and reduced to uranium dioxide with hydrogen. In another process, the solution remaining after solvent extraction is evaporated to obtain uranyl nitrate hexahydrate [$(UO_2)(NO_3)_2 \cdot 6\,H_2O$, UNH]. The UNH is thermally decomposed to uranium trioxide, which is also reduced to uranium dioxide with hydrogen.

The uranium dioxide is then converted to uranium hexafluoride by reaction with hydrogen fluoride (hydrofluorination) and oxidation with elemental fluorine.

Dry Process. In the dry process, the uranium concentrate is reduced directly to uranium dioxide with hydrogen in a fluidized-bed reactor. The next steps are reaction with hydrogen fluoride and oxidation with fluorine as in the wet process, except that the hexafluoride product is less pure. Residual impurities are removed by fractional distillation. The end product, however, has a higher purity than that produced from the wet process.

For technical reasons, a conversion plant can be operated at only 80 – 90 % of design capacity. Even so, the capacity of existing plants is sufficient to meet international demand for years to come (see Table 54). Although installed conversion capacity is around 54 000 t of uranium (plants in Eastern Europe are not included) the demand for conversion of natural uranium in 1990 was 42 300 t U and will increase to ca. 47 500 t of uranium in 2000.

Table 54. Conversion plants in WOCA

Country	Plant	Uranium capacity, t	Process
United States	Metropolis	12 700	dry
	Sequoyah	9 090	wet
France	Pierrelatte	14 000	wet and dry
United Kingdom	Springfields	9 000	wet
Canada	Port Hope	9 000	wet
Total		53 790	

In addition to large-capacity plants in the United States, France, Canada, and the United Kingdom, whose products are sold internationally, smaller plants serve primarily domestic needs or are intended to do so in the future (Brazil, South Africa, Japan, and India).

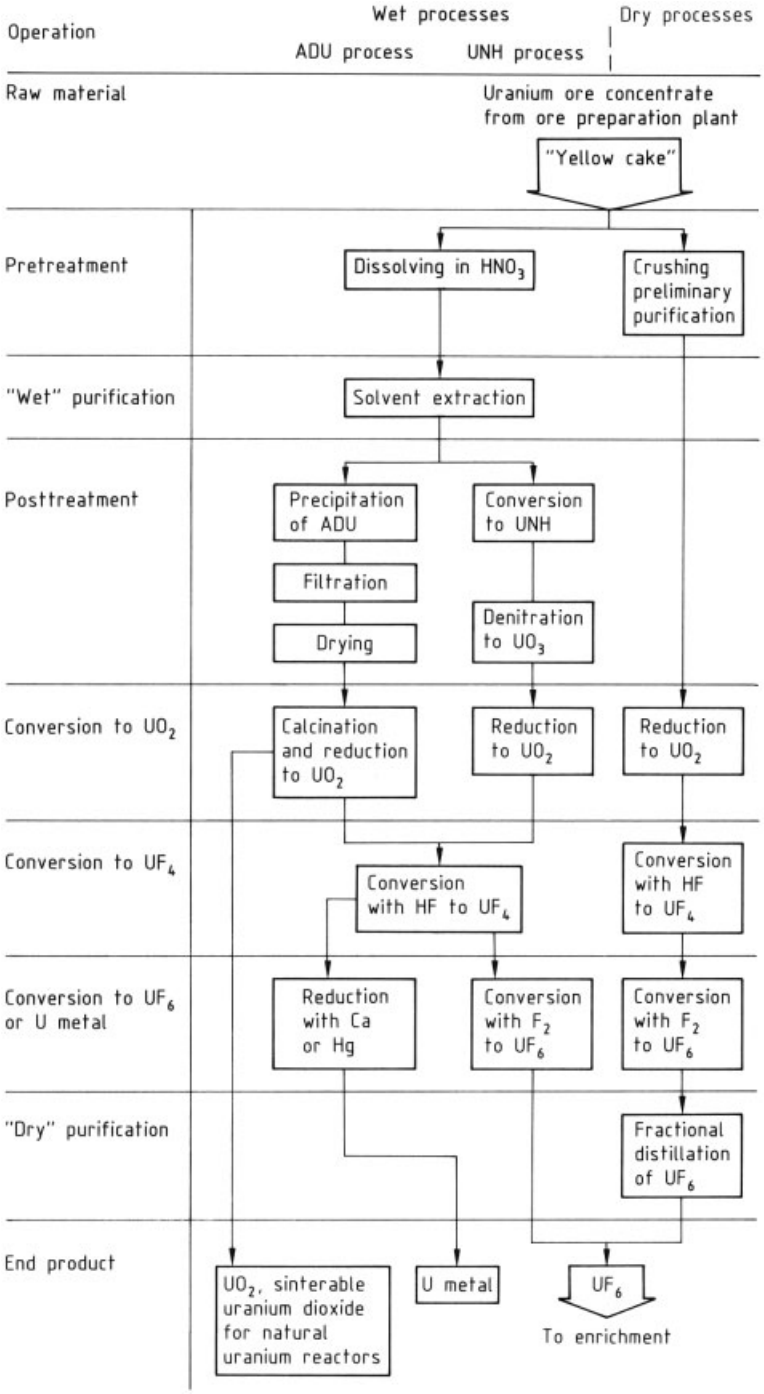

Figure 107. Schematic diagram of conversion processes [193]

5.2.5. Enrichment

Natural uranium contains only 0.71 % of the isotope ^{235}U. This fraction must be increased to > 2.5 % to sustain a chain reaction in light-water reactors. Modern power reactors use uranium with an enrichment up to 5.0 %. Large-scale uranium enrichment is currently achieved by the gaseous diffusion and gas centrifuge processes. About 90 % of the enrichment plants in the Western world employ gas diffusion. The nozzle separation (jet nozzle) process developed in Germany has also come into full-scale use. The laser separation process is still in the developmental stage; a demonstration plant is not expected to be in operation before the mid-1990s.

The gaseous diffusion, gas centrifuge, and nozzle separation processes have in common a gas mixture feed consisting of $^{235}UF_6$ and $^{238}UF_6$, which is inlet into a medium where it is separated into two partial streams that are discharged. The product stream is a gas fraction enriched in $^{235}UF_6$, whereas the tails stream is depleted in $^{235}UF_6$. Along with $^{238}UF_6$, the tails stream also contains $^{235}UF_6$, the so-called tails assay.

Centrifuge Process. In the centrifuge process, the $^{235}UF_6$–$^{238}UF_6$ stream enters into a rotating centrifuge (see Fig. 108 A). The heavier $^{238}UF_6$ molecules collect at the outer wall of the centrifuge, whereas the lighter $^{235}UF_6$ molecules collect near the axis where they are withdrawn as an enriched product stream. The separation factor of a centrifuge depends on the peripheral velocity and the length of the centrifuge tube. Industrial centrifuges are therefore tubes operated in supercritical rotation. A counterflow is generated inside the centrifuge to enhance the isotope separation. Even so, the separating efficiency is so low that many centrifuges must be linked in a cascade. To achieve a 3 % enrichment requires a series of 10 to 15 centrifuges.

Diffusion Process. In the diffusion process, gaseous uranium hexafluoride is pumped under high pressure through fine-pored membranes (barriers). The membrane material must be mechanically strong to withstand the large pressure difference and chemically resistant to UF_6 (see Fig. 108 B). The lighter $^{235}UF_6$ diffuses through the barriers more rapidly and collects on the downstream side. As in the centrifuge process, a single separation step is insufficient. To effect a 3 % enrichment, separation must be repeated 1000 to 1500 times in barrier cascades.

The centrifuge separation process has the advantage of a much lower power consumption in comparison to gaseous diffusion.

Jet Nozzle Process. The jet nozzle process uses primarily centrifugal forces for isotope separation. A gas mixture consisting of 4 % UF_6 and 96 % helium is passed through a slot-shaped nozzle. The gas stream is deflected at an outer wall by 180°. The heavier $^{238}UF_6$ particles collect at the outer wall, whereas the lighter $^{235}UF_6$ particles are deflected with the gas stream (→ Air Classifying, Chap. 4.2.2.). This separation process is increased by a peeling cutter.

Laser Processes. Much effort is being put into the development of a laser process for isotope separation. The *AVLIS* (*atomic vapor laser isotope separation*) *process* developed in the United States and France is based on slight differences in the absorption spectra of various isotopes of the same element. A stream of elemental uranium vapor is irradiated by a laser beam and the ^{235}U isotopes are selectively photoionized. The ^{235}U isotopes are then separated electromagnetically from the other uranium isotopes.

Developmental work on *molecular laser isotope separation (MLIS)* continues in Germany.

In addition to the United States, France, the former Soviet Union, and the three Urenco plants (in The Netherlands, the United Kingdom, and Germany) are the most important vendors of enrichment services (see Table 55). Other plants, usually with smaller capacities, serve domestic consumers.

5.3. Fabrication of Fuel Elements

5.3.1. Fuel Assemblies for Light-Water Reactors [195], [196]

5.3.1.1. Functions of Fuel Assemblies

Of a number of material combinations suitable for fuel assemblies in light-water reactors (uranium metal with aluminum cladding is an early example), one has become the standard: uranium dioxide (UO_2) fuel in the form of a stack of cylindrical pellets surrounded by a cladding

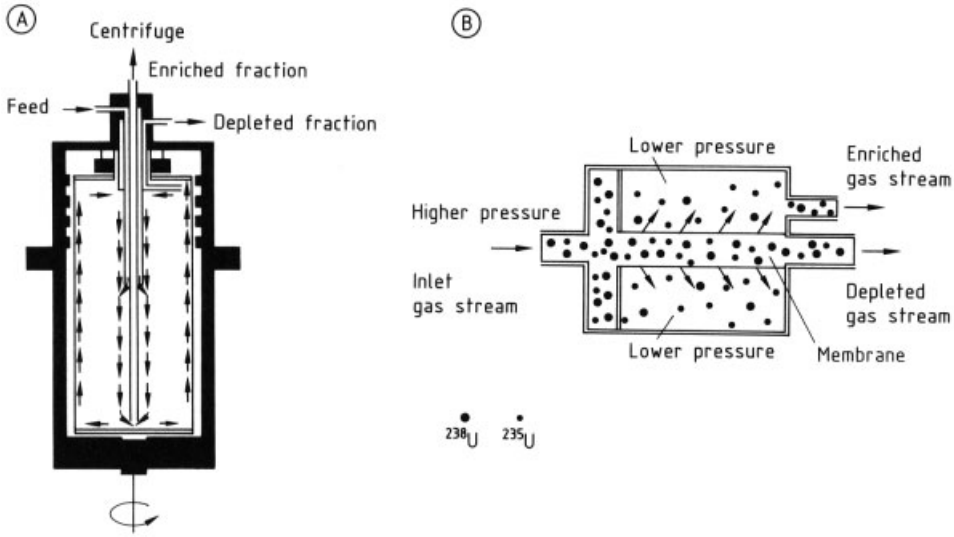

Figure 108. Schematic representation of the most common separation processes [194]
A) Gas centrifuge; B) Gas diffusion stage

Table 55. Uranium enrichment plants

Country	Plant	Capacity, 1000 SWU ∗	Process	On stream
Argentina	Pilcaniyeu	100	diffusion	1988
Brazil	Resende	200	centrifuge	1995
China	Lanzhou	200	diffusion	1980
France	Pierrelatte	10 800	diffusion	1979
Germany	Gronau	500	centrifuge	1985
India	Kamataka	n.r. ∗∗	centrifuge	planned
Japan	Ningyo-Toge	200	centrifuge	1988
	Rokkasho-Mura	1 500	centrifuge	1992
Pakistan	Kahuta	n.r. ∗∗	centrifuge	1984
South Africa	Valindaba	300	South African process	1982
Former Soviet Union	Siberia	10 000	centrifuge	
The Netherlands	Almelo	1 250	centrifuge	1984
United Kingdom	Capenhurst	900	centrifuge	
United States	Paducah	11 300	diffusion	1954
	Portsmouth	8 300	diffusion	1956

∗ SWU = separative work unit, a measure of capacity for separation of uranium into enriched and tails material.
∗∗ n.r. = not reported.

of Zircaloy. A fuel rod or pin, up to 5 m long and 9.5 – 11 mm in diameter, performs two contrasting functions. The rod must offer the least possible resistance to the release of heat produced by fission, but it must also provide total retention of fission products over its entire lifetime in the reactor (up to 4 (in PWR) and 6 (in BWR) years). Although off-gas systems of light-water reactors are designed to handle the amount of fission products formed in reactor operation with up to 1 % defective fuel rods, fuel-rod failure ratios of only ca. 0.001 % are common. The economics of the fuel assembly is governed largely by the burnup achievable at low defect rates. Since the 1970s the reload-batch discharge burnup has been steadily increased to 45 – 50 MW · d/kg for PWR (i.e., for a reload batch size of ca. one-fourth of the fuel assembly per year) and 40 – 45 MW · d/kg for BWR. This increase is a result of the development of cladding material that is more resistant to waterside corrosion, more dimensionally stable pellet ceramic structures, and the use of integral burnable absorbers such as Gd_2O_3.

With ^{235}U enrichments up to 5 wt % (current fabrication limits) burnups of 60 – 65 (PWR) and 55 – 60 (BWR) MW · d/kg are realistic targets for the next 1 – 2 decades [197].

5.3.1.2. Raw Material for Nuclear Fuel

Virtually the only raw material needed for the manufacture of nuclear fuel is uranium hexafluoride because light-water reactors contain enriched uranium and industrial enrichment processes operate with gaseous uranium hexafluoride. Uranium from reprocessing plants is available as uranium oxide, U_3O_8 or UO_2. If it is reused in LWRs it must either be reenriched or blended with medium or high enriched uranium (MEU or HEU). This material is at present available from the disposal of surplus weapon HEU due to nuclear disarmament treaties or from the disposal of surplus amounts of MEU (20 – 23 % ^{235}U) used for military ship reactors. Reenriched material arrives at the fuel fabrication process in the form of uranium hexafluoride. When chemically contaminated scrap and waste are produced in the fabrication process, their uranium content is recovered by extraction as $(UO_2)(NO_3)_2$ solution and recycled in this form.

5.3.1.3. Conversion Processes → Uranium, Uranium Alloys, and Uranium Compounds, Chap. 6.3.10.

Conversion in this case means the transformation of gaseous uranium hexafluoride to sinterable uranium dioxide. The amount of uranium dioxide produced worldwide is some 10 000 t/a. There are four competing conversion processes [198–200]. The demands made on these processes are as follows: they must be simple, quantitative, and able to be carried out in simple apparatus with the smallest possible amount of harmful byproducts. The uranium dioxide must have nuclear-grade purity and must contain the smallest possible amount of residual fluorine (because as little as 20 ppm fluorine in uranium dioxide can cause severe cladding damage). It must also have a consistent processability; i.e., it must be sinterable and free-flowing. The latter property allows quick and uniform mold filling and ensures the structural homogeneity of the sintered material, with no perceptible structure of the original UO_2 grains [201].

From the process-engineering standpoint, the conversion should be assessed in conjunction with the subsequent fabrication of uranium dioxide sintered material, because extra effort to prepare the powder for compaction can be offset against extra effort in the chemical process.

Wet Process. At first, gaseous ammonia or hydrogen peroxide was used to precipitate uranium from an *aqueous nitrate solution* because this operation, which yields ammonium diuranate (ADU), $(NH_4)_2U_2O_7$ and $UO_4 \cdot 2\,H_2O$, also removes impurities. The product compounds are converted by heating to UO_3 and U_3O_8, respectively, and either of these can be quantitatively reduced to uranium dioxide with hydrogen at 600 °C. These processes give better product with respect to residual fluorine content and the sinterability than direct hydrolysis of UF_6 (see below):

$$UF_6 + 2\,H_2O \rightarrow UO_2F_2 + 4\,HF$$

$$UO_2F_2 + H_2O \rightarrow UO_3 + 2\,HF$$

or precipitation of ADU from aqueous UO_2F_2-solution.

Ammonium Uranyl Carbonate Process. A free-flowing powder with disappearing particle structure was formerly produced only in the ammonium uranyl carbonate (AUC) process [202]. In this process, originally developed by NUKEM and now still employed by ABB-Atom, KNFC (Korea) and INB (Argentina), uranium hexafluoride is introduced into an ammonium carbonate solution through nozzles, along with ammonia and carbon dioxide. In contrast to ADU, which forms a very fine-grained precipitate when ammonia alone is added, a precipitate of tetraammonium tricarbonato dioxouranate $[(NH_4)_4 \cdot UO_2(CO_3)_3]$ is formed along the pH gradient (pH 5.5 – 7.5) in the reaction zone. Tetraammonium tricarbonato dioxouranate crystallites of well-defined size (around 40 µm) precipitate when more ammonium carbonate is added (in the pH range between 5.5 and 7.5). This process and the final pH value can be controlled to some extent through the solution temperature and the $NH_3 - CO_2$ ratio. The very uniform salt precipitate is separated from the mother liquor on rotary disk filters, and the

salt cakes are treated with methanol to remove adherent ammonium fluoride residue. Fluorine levels < 0.01 % can be achieved in this way. The moist filter cake is conveyed pneumatically through a system of locks into a fluidized-bed furnace (Fig. 109).

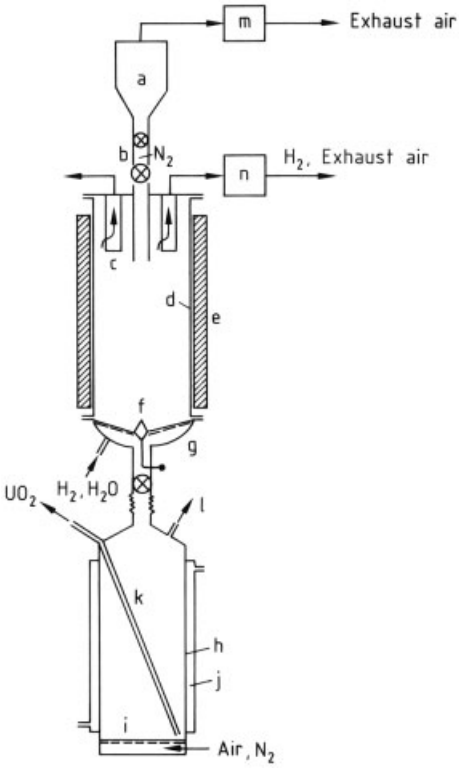

Figure 109. Fluidized-bed equipment for the reduction of AUC, $(NH_4)_4$ $[UO_2(CO_3)_3]$, to uranium dioxide, cooling of the powder, and partial reoxidation (RBU system)
a) Cyclone for separation of AUC; b) Lock and powder inlet; c) Cartridge filters; d) Fluidized-bed furnace; e) Heating jacket; f) Discharge closure cone; g) Metal frits; h) Cooling vessel; i) Metal frits; j) Cooling jacket; k) Discharge of UO_2; l) To particulate filter; m) Fine particulate filter; n) Condensation and absorption of NH_3, CO_2 and H_2O

At ca. 500 °C, the uranate is dried, decomposed, and treated with a mixture of hydrogen and steam both to reduce it and to drive out fluorine to < 100 ppm. Furnace charges of some 300 kg are finally heated to 650 °C to reduce the specific surface of the powder to ca. 5 m^2/g.

Thermal Decomposition Processes. Small amounts of uranium dioxide are obtained by thermal decomposition of $UO_2(NO_3)_2$ in a microwave field, yielding a powder with a particularly large specific surface but only moderate sinterability, and by thermal decomposition of sol-gel particles formed by reaction between peptized uranyl nitrate droplets and ammonia.

Dry Process. Another process is the continuous gas- or vapor-phase conversion of uranium hexafluoride with steam and hydrogen. Here, uranium hexafluoride is diluted with nitrogen as an inert carrier gas (to avoid local overheating where the two reactants come in contact); this gas mixture is admitted into a large powder mass (e.g., in a fluidized bed or rotary kiln; see Fig. 110), and the above reaction is carried out in two steps at temperatures below the melting point of UO_2F_2 to avoid inclusions. Large installations in France, United Kingdom, and the United States employ this process. In the fluidized bed method, the primary uranium dioxide particles grow into onionlike shell structures by absorption of UO_2 formed by conversation of UF_6. These product structures were so stable that they could still be detected in polished sections of compacted and sintered pellets. The rotary kiln method uses a gas-phase reaction that produces soft particles whose particle structure disappears in the compaction step.

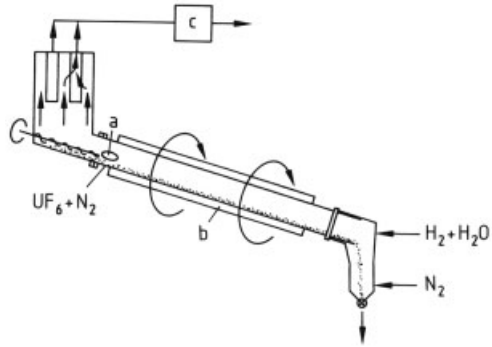

Figure 110. Rotary furnace for dry $UF_6 \rightarrow UO_2$ conversion
a) UF_6 and N_2 inlet; b) Rotary furnace; c) HF and H_2O condensation

More recently, a modified fluidized bed process was developed and implemented in full-scale production plants in Germany and the United States by Siemens. This process produces soft particles and a free flowing powder.

Quality and Testing. The sintering process for the pellets demands the most well-defined raw materials possible. Along with the processing test, the specific surface area (determined

by the BET method), the dumped and vibrated densities, the oxygen – uranium ratio, and scanning electron micrographs are employed to assess product quality.

Choice of Conversion Process. Dry conversion processes have significant environmental advantages over wet processes. Dry processes recover hydrofluoric acid as a by-product and thereby avoid generation of large quantities of contaminated liquid wastes.

5.3.1.4. Production of Uranium Dioxide Sintered Pellets

Ideally, heat of up to 500 W per centimeter of fuel-rod length must be transferred to the water coolant. At the same time the temperature in the center of the fuel pellet must be as far as possible below the melting point of uranium dioxide (2850 °C). For these reasons, the heat transfer and thermal conductivity of the fuel must be optimized. This applies chiefly to the radial conductivity of uranium dioxide and the heat transfer in the annular clearance between the stack of pellets and the cladding tube. This gap must be closed over the largest possible area in the operating condition. Although the precision of the cylindrical geometry of the cladding tube can be ensured by monitoring the tube rolling process, up to now no industrial sintering process can give a cylinder accurate to ± 10 µm in diameter without a subsequent correction step (grinding).

The pellet shape now used is a cylinder 8 – 10 mm in diameter and 30 – 70 µm smaller in diameter than the cladding tube; the length is ca. 1.2 times the diameter. The pellets deform in reactor service as a result of mainly radial temperature differences; the end faces bulge out and press the edges to the inside of the cladding, which can result in the formation of circumferential grooves. To restrict axial elongation of the fuel stack during operation, one or both end faces of each pellet are dished so that pellets make contact only in the relatively cool outer annular zone. In addition, the cylinder edges are beveled to prevent their pressing against the cladding tube. Edge fractures are thus reduced, and the danger of ceramic fragments breaking off during pellet charging is avoided. These fragments could interfere with the charging operation and create point load inside the cladding tubes, leading to deformation and thus to preferred points of stress-corrosion attack causing pellet cladding interaction (PCI) failures.

Competitive fuel-assembly plants must produce several pellets per second. Therefore, high-speed rotary presses have been developed, which demand a very uniform and free-flowing feed material. If the uranium dioxide powder from conversion is not sufficiently free-flowing, it must be modified by the addition of binders, premolding, crushing and granulation, or screening. To prevent uneven compaction of the powder in the green compact, the wall friction in the mold and the internal friction in the powder must be low. This is accomplished with highly polished hard metal mold surfaces, mold lubrication, and sometimes the addition of a few tenths of a percent of lubricant (zinc stearate) to the powder.

In *sintering*, the stresses induced in molding are transformed to work of compaction (about 18 % shrinkage). This process, which involves recrystallization in the uranium dioxide lattice between 1100 and 1700 °C, is accelerated by oxygen. As oxygen is present in uranium dioxide in a higher than stoichiometric quantity, a sufficient amount of it is available. However, because a stoichiometric final product ($UO_{2.00}$) is required, the last step in sintering is always reduction with hydrogen (or sometimes carbon monoxide). The entire process cannot, however, be entirely separated into two steps, because a completely compacted ceramic is no longer porous enough for the oxygen (which is present mainly as the OH radical under sintering conditions) to diffuse away rapidly.

Sintering furnaces are usually operated with dry or partially water-saturated hydrogen. At 1700 °C, the entire operation (of which sintering proper takes only a few minutes) is complete after 2 – 3 h. Aluminum oxide-lined pusher type or walking-beam furnaces with electrical heating by molybdenum coils are employed because of the high temperature. Energy can be conserved, if sintering is performed in carbon dioxide at 1100 °C and the pellets are subsequently reduced with hydrogen at 1100 °C in a furnace zone separated from the first section by a hot-gas lock. Depending on the pellet size and the quality of the edges, the pellets can pass through the sintering process either loose in molybdenum pans or racked with molybdenum spacer plates.

The cylindrical shape is next brought to the desired diameter by centerless *grinding*. The pellets are washed and dried, and defective pellets are removed. Chemical and physical tests are performed (purity, uranium content, O: U ratio, moisture content, density and dimensions, plus sampling inspection of the microstructure by preparation of polished sections), and in this way only high-quality pellets complying with specifications are incorporated into the fuel rods.

$UO_2 - Gd_2O_3$ Pellets. To allow higher target burnups and smooth the neutron flux in the reactor, the initial excess reactivity of the, in this case, more highly enriched fissile material must be compensated by burnable neutron absorbers. For this purpose, uranium dioxide powder is mixed with up to 10 % gadolinium (III) oxide powder before the pellet fabrication operations described above. Interdiffusion of uranium dioxide and gadolinium (III) oxide usually does not yield a homogeneous mixed crystal; thus restrictions on the microstructure, a somewhat lower thermal conductivity, and an altered pore structure must be accepted. Because confusion and contamination of pure with Gd-doped UO_2 would have grave consequences for local heat transfer, the production lines must be strictly separated. Westinghouse uses a pellet surface coating with boron (ZrB_2) instead of the widely used homogeneous Gd doping of the pellet volume.

5.3.1.5. Production of Fuel Rods and Fuel Assemblies

The cladding tube represents the essential barrier against the escape of fissile materials and fission products into the coolant and, in the other direction, against mechanical and chemical attack on the ceramic fuel by the flowing coolant (water or steam). For light-water reactors, Zircaloy has provided a suitable compromise among mechanical strength, chemical stability, and neutron absorption (see Section 3.1.4.1). It is delivered to the fuel-rod plant as complete cladding tubes with some excess length. The material is annealed above 830 °C, the temperature of the $\beta - \alpha$ phase transition (at which Zircaloy becomes ductile); passed through a final train of pilger rolls to accomplish some 25 % of the hardness gain attainable with cold rolling and subjected to ultrasonic inspection for pores and cracks. In this way, the cladding material going into the reactor has about the same strength that would be produced during operation through the opposing actions of annealing at service temperature and radiation-induced embrittlement (resulting from dislocations in the metal lattice). The tube surface is polished and carefully washed to ensure the required cleanliness needed as well as smoothness of surface which is the basis for a uniform oxide film in the reactor. To achieve higher resistance against waterside corrosion claddings with Zr based materials slightly outside the ASTM specification – partly with the addition of vanadium and/or niobium and binary Zr – Nb alloys – are used increasingly [203]. Siemens introduced duplex cladding with a 20 % highly corrosion resistant outer layer on standard Zircaloy material.

Fuel-Rod Fabrication. The first step in fuel-rod fabrication is to close one end of the cladding tube with a welded plug. The pellets are placed end to end in V channels, forming stacks and are mechanically pushed or vibrationally charged into the cladding tube with the least possible damage and with no edge breakage. A coil spring is inserted to hold the stack of pellets in end-to-end contact during operation, because voids would cause hot spots in the fuel rods. Heat transport in the clearance (annulus) between the cladding tube and the pellets is improved by evacuating the space and filling it with helium. A saving in thickness of the cladding is possible if the system pressure in the PWR (15.5 MPa at operating temperature) is counteracted by prepressurizing the rod to 2.1 – 2.7 MPa (measured cold). Rods that are used for BWRs are prepressurized to about 0.5 MPa. The cladding tube is then closed with a second plug. Either this operation is performed under helium pressure (one-step welding), or an axial hole about 1 mm in diameter is left in the plug and helium is injected later; the hole is then closed by subsequent spot-welding (two-step welding). The two-step welding operation is necessary if the end plug is welded in by the tungsten – inert gas (TIG) method, because it is difficult to weld by this method under helium pressure. Because of this complication, flash butt welding (also called "resistance pressure welding") has been adopted. This technique allows one-step application of

the plug. It offers a very low rate of defects because of the high mechanical pressure in the annular melt zone, which prevents pore formation and results in a narrower fused zone of altered microstructure.

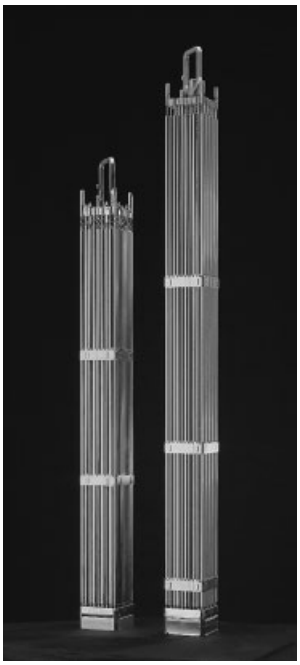

Figure 111. PWR fuel assembly (Siemens photo)

Helium leak testing, which can detect leaks down to $10^{-11}\,\mathrm{bar\,L^{-1}\,s^{-1}}$, is also common. The test involves an evacuated chamber and a mass spectrometer. The final rod check is performed in a rod scanner, where the correct pellet enrichment and axial location is determined.

Fabrication of Fuel Assemblies. The fuel assembly (Figs. 111 and 112) initially consists of a prefabricated skeleton of spacer rods (and control rod guide tubes in the case of PWR assemblies) holding the spacer grids or assemblies. The fuel rods are inserted in layers into these spring-clip grid assemblies; a PWR assembly has up to 300 fuel rods and a BWR assembly up to 100 rods. In a PWR assembly no fuel rods are connected to the upper and lower tie plate, only the guide tubes. Traditionally, in BWR assemblies a few tie fuel rods are connected to both end pieces. However, modern 10×10 designs by Siemens and ABB avoid the high operating loads on these fuel rods by applying a skeleton or a subassembly structure.

Figure 112. BWR fuel assembly (Siemens photo)

The final operation is a careful dimensional check with vertical fuel assembly (not horizontal because the rods bend under their own weight).

5.3.2. Fuel Elements for High-Temperature Reactors

5.3.2.1. General Aspects

The main characteristic of the HTR is its ability to make the energy generated by nuclear fission available at a high temperature. Coolant outlet temperatures of 750 – 950 °C necessitate fuel-element surface temperatures higher than 1000 °C. These conditions almost inevitably lead to the use of helium as coolant and to a fuel element made exclusively of ceramic materials.

The fuel elements currently used in the high-temperature reactor differ greatly in shape. The HTR fuel element developed in the United States [204], [205] and in Japan [206] is a large, hexagonal prism. Germany [207–209] and the former Soviet Union [210–212] use a spherical fuel element. Despite these very different shapes, the elements all have the same basic design. The fuel they contain is in the form of microspheres,

each the size of the head of a pin, coated with several layers of ceramic material. These coated particles are embedded in a graphite matrix. The graphite serves simultaneously as structural and a moderator material.

From the very beginning, development of the high-temperature reactor was closely linked with the uranium – thorium fuel cycle [213–215]. Highly enriched uranium (HEU) with ca. 93 % enrichment is used as fuel for the fuel cycle and thorium is added as fertile material. Neutron capture leads to the conversion of thorium into the fissile isotope uranium-233 with protactinium as the intermediate:

$$^{232}\mathrm{Th} + \mathrm{n} \longrightarrow {}^{233}\mathrm{Th} \xrightarrow{\beta^-} {}^{233}\mathrm{Pa} \xrightarrow{\beta^-} {}^{233}\mathrm{U}$$

Compared to the uranium – plutonium cycle, the HEU cycle is clearly of economic advantage to the HTR. Moreover, with this fuel cycle, large thorium deposits can be used for energy production by nuclear fission. Both prototype plants – Fort St. Vrain (330 MW) in the USA and the THTR (300 MW) in Germany – were operated by using the HEU fuel cycle.

Fuel kernels of the THTR elements contain thorium and highly enriched uranium as a mixed oxide in the ratio 10 : 1. The diameter of these very dense sintered kernels is 400 µm. In the fuel elements of the Fort St. Vrain reactor, the fuel kernels consist of the mixed carbide $(U, Th)C_2$ having a thorium: uranium ratio of 4.25 : 1. The diameter is in the range of 100 to 275 µm. Apart from these fuel kernels, others kernels contain fertile material only in the form of ThC_2 and have a diameter of 300 – 500 µm.

In view of the nonproliferation of materials employed in nuclear weapons, the trend worldwide in the HTR line at the beginning of the 1980s moved toward the uranium – plutonium cycle, [i.e., the fuel cycle with lightly enriched uranium (LEU)] and away from highly enriched uranium [216], [217]. Fuel concepts in Germany, the former Soviet Union, and Japan provide for the use of uranium dioxide fuel kernels with 6 – 10 % uranium enrichment. The diameter of the kernels is 500 µm (Germany, Former Soviet Union) and 600 µm (Japan). The use of thorium as fertile material is still being pursued in the United States. This fuel concept provides for the application of uranium oxide carbide (UC_xO_y) fissile kernels. The uranium has almost a 20 % degree of enrichment, and the kernels have a diameter of 300 µm, whereas the diameter of the kernels of fertile material is 500 µm.

Coated Particles. As with all other reactor concepts, fission products in the HTR should be retained in the fuel element as much as possible and should not enter the coolant cycle. The graphite in which the fuel particles are embedded cannot retain the fission products because of its porosity. This task is carried out by a system of layers, covering each individual fuel kernel [218], [219]. When the HTR fuel element was first developed, this layer system consisted of two pyrolytically deposited carbon layers (BISO coating). However, an additional layer of silicon carbide (TRISO coating) was found not only to markedly increase the mechanical stability of the coating system, but also to improve considerably the retention capability for solid fission products that tend to diffuse at these temperatures. Although the fuel particle of the THTR fuel element still has a BISO coating, all currently used fuel element concepts worldwide provide for a TRISO coating [220]. Therefore, the most highly developed and tested fuel particle with this coating is described here, namely, the particle for the LEU fuel element of the German THTR successor plants.

The innermost layer, the *buffer layer*, consists of very porous pyrocarbon ($\varrho = 1.0$ g/cm^3) and has a thickness of 95 µm. The functions of this layer are:

1) To buffer the swelling of the fuel kernel under irradiation (i.e., to prevent the resulting mechanical strain on the subsequent layers);
2) To attenuate fission products that leave the fuel kernel and originate from fissions near the surface (recoil atoms), thereby preventing damage to subsequent layers;
3) To provide void volume for gaseous fission products and thus prevent a dramatic pressure increase in the particle.

The layer above the buffer layer is made of dense pyrocarbon ($\varrho = 1.9$ g/cm^3) with a thickness of 35 µm. This gastight pyrocarbon layer protects the next very important silicon carbide layer from corrosive attack by fission products. Moreover, this layer prevents the penetration of hydrogen chloride and other chlorine – containing compounds into the buffer layer. These compounds are formed on deposition of the sili-

con carbide layer (see Section 5.3.2.2) and contribute to the corrosion of this layer during reactor operation.

The silicon carbide layer is 35 µm thick and practically reaches the theoretical density of silicon carbide ($\varrho = 3.2\ g/cm^3$). This layer acts, above all, as a diffusion barrier for solid fission products. In addition, it gives the particle a higher mechanical strength. This is especially important because the particle and its coating constitute a pressure vessel that must withstand inner pressures of up to 20 MPa.

The outermost layer is made of pyrocarbon ($\varrho = 1.9\ g/cm^3$) and is 40 µm thick. It protects the brittle silicon carbide layer to a certain extent, contributes somewhat to the stability of the pressure vessel, and acts as a barrier for gaseous fission products if the silicon carbide layer fails.

Fuel Elements. The *spherical fuel element* developed in Germany (see Fig. 113) has a diameter of 60 mm. The actual fuel zone containing the coated particles has a diameter of 50 mm. The surrounding fuel-free shell is 5 mm thick. Both the fuel zone and the fuel-free shell consist of the same type of graphite, which, in contrast to normal graphite, still contains a certain proportion of nongraphitic carbon due to the manufacturing process [221], [222].

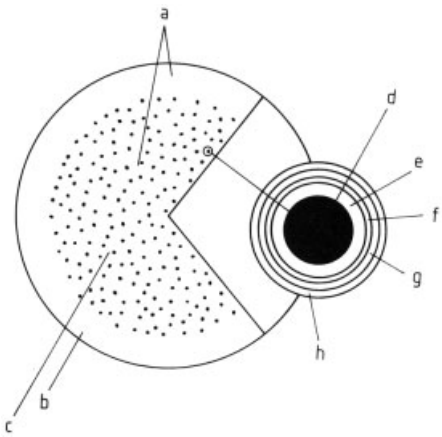

Figure 113. Spherical fuel element (Germany)
a) Matrix; b) Fuel-free shell; c) Fueled zone; d) Kernel; e) Buffer layer; f) Inner pyrocarbon layer; g) Sic layer; h) Outer pyrocarbon layer

The spherical fuel element developed in the former Soviet Union also has a diameter of 60 mm. The diameter of the fueled zone, however, has not yet been stipulated definitely. Diameters of 40 and 50 mm are currently being considered.

The *block-shaped HTR fuel element* developed in the United States (see Fig. 114) is a hexagonal prism made of graphite having a width over flats of 359.1 mm and a height of 793 mm. It is provided with axial holes to receive the fuel rods and to guide the coolant gas. The fuel rods have a diameter of 15.9 mm and a length of ca. 60 mm. They consist of a relatively porous graphite matrix in which the fissile and fertile coated particles are embedded.

The *Japanese fuel element* concept (see Fig. 115) also provides for block-shaped, hexagonal fuel elements with a width over flats of 360 mm and a height of 580 mm. The fuel compacts containing coated particles have the form of hollow cylinders. They are put into graphite tubes, that are, in turn, inserted into corresponding holes in the graphite block. Coolant gas flows through the gap between the graphite block and the graphite tube.

5.3.2.2. Production

Production of Fuel Kernels. The production of the spherical fuel kernels can be carried out by powder agglomeration or by wet chemical processes based on solidification of droplets [223], [224]. A narrow bandwidth of the kernel diameter and an almost ideal spherical shape are essential for good irradiation performance of the coated particles. These quality requirements can be met more easily with wet chemical processes that dry methods are no longer of interest. Many variations of the *wet chemical* process have been developed. However, the gel precipitation process and the hydrolysis process (H process) are practically the only processes of technical importance today.

Gel Precipitation. In the production of *uranium dioxide kernels* by gel precipitation (Fig. 116), a uranyl nitrate solution is the starting material [207], [208]. This solution is preneutralized to a certain degree with aqueous ammonia before the addition of organic additives [poly(vinyl alcohol) and tetrahydrofurfuryl alcohol]. Forming is achieved by the vibrodropping of this feed solution. The droplets are subjected to a lateral stream of gaseous ammonia as they fall. This results in presolidification based

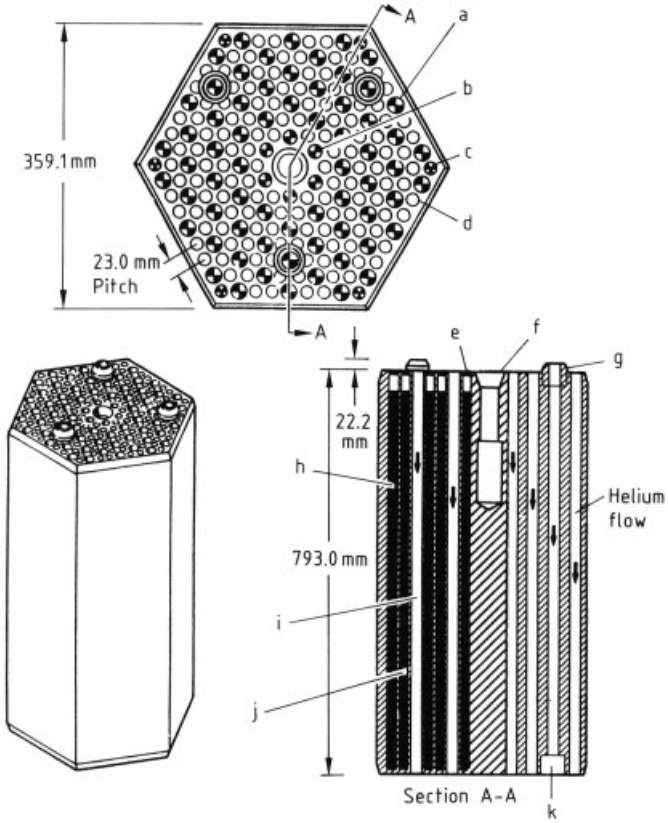

Figure 114. Prismatic fuel element (United States)
a) Coolant hole 21.0 mm diameter (66); b) Coolant hole 18.2 mm diameter (6); c) Burnable poison hole 12.7 mm diameter; d) Fuel hole 15.9 mm diameter (132); e) Cemented graphite plug; f) Fuel-handling pickup hole; g) Dowel pin (3); h) Burnable poison rod; i) Coolant channel; j) Fuel-rod stack; k) Dowel socket (3)

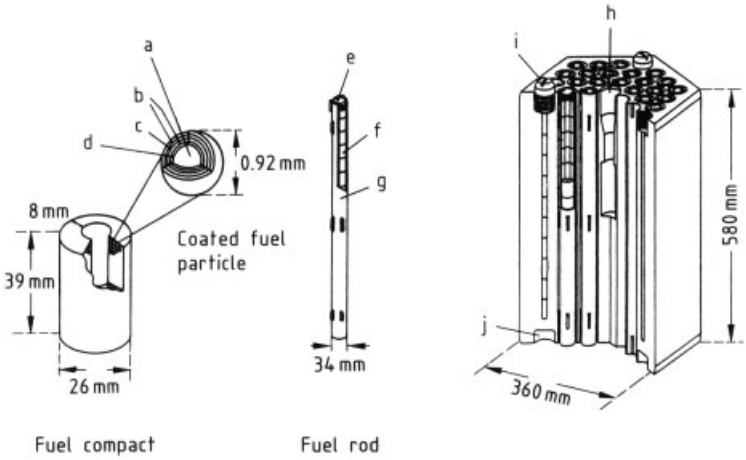

Figure 115. Prismatic fuel element (Japan)
a) Fuel kernel; b) High-density pyrocarbon layer; c) Sic layer; d) Low-density pyrocarbon layer; e) Plug; f) Fuel compact; g) Graphite sleeve; h) Fuel-handling hole; i) Dowel pin; j) Dowel socket

on the fact that ammonia-containing uranium trioxide hydrate is formed, which covers the droplet like an elastic skin. The drop is now so stable that it no longer loses its round form when it hits the surface of the precipitation bath.

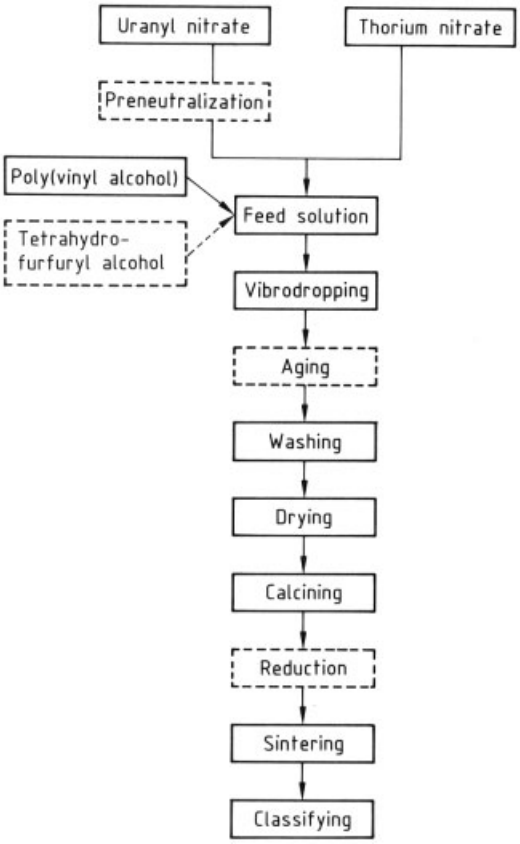

Figure 116. Fuel kernel fabrication

The reaction of uranyl nitrate with ammonia is brought to an end in the precipitation bath, an aqueous solution of ammonia. Small, translucent, gellike kernels are obtained. After aging, the kernels are washed first with water and then with alcohol. The ammonium nitrate formed in the kernels is dissolved in this process. Subsequently, the required process steps at higher temperature can be performed: drying and calcining in air, reduction, and sintering in an atmosphere of hydrogen. The kernels are then sieved, and the small number of highly unround kernels are separated by using a vibration plate inclined against the horizontal (classifying).

The use of the gel precipitation process for the production of kernels of *mixed oxides of uranium and thorium* requires a starting feed solution that contains uranyl nitrate, thorium nitrate, and poly(vinyl alcohol). Tetrahydrofurfuryl alcohol is not required and the steps – preneutralization, aging, and reduction – in the flow diagram are not necessary because of the different chemical behavior of uranium and thorium.

If appropriately modified, this process is also suitable for the production of *carbide and oxide – carbide kernels*. The amount of carbon required for the formation of carbide in each case is added to the feed solution as carbon black. Both carbide formation and sintering are carried out between 1400 and 1800 °C in argon. If uranium carbide kernels are required with a density > 90 % of the theoretical density, further treatment at about 2500 °C (melting) is essential. Here, too, the last two steps of fabrication are sieving and classifying. The gel precipitation process is used in Germany to produce the predominant part of AVR fuel and all fuel kernels for THTR fuel elements.

Hydrolysis Process. In the gel precipitation process, the ammonia that causes gelation is brought in contact with the droplets of feed solution from the outside (external gelation). In the H process, however, the ammonia is produced inside the droplets of feed solution by the chemical decomposition of hexamethylenetetramine, which serves as the ammonia carrier (internal gelation) [225–228]. As a result, the H process can also be used to make uranium dioxide kernels with a diameter of 1 mm or more, which cannot be achieved with the gel precipitation process.

The aqueous feed solution in the H process normally contains uranyl nitrate in relatively high concentrations, urea (to stabilize the solution by complex formation), and hexamethylenetetramine. This feed solution must be kept at < 5 °C because higher temperature initiates the decomposition of hexamethylenetetramine into ammonia and formaldehyde, resulting in early precipitation of the uranium compound.

The cooled feed solution is dripped into a hot precipitation bath of tetrachloroethylene. The ammonia liberated at the high temperature initiates a hydrolysis reaction and thus causes the solidification of the droplets in the precipitation bath. A short temperature treatment then

frees the kernels from the still adhering, freely volatile tetrachloroethylene. The ammonium nitrate formed and the remaining organic substances are removed from the kernels by subsequent washing with ammonia-containing water. After the kernels have been dried in air, they are reduced (at ca. 500 °C) and then sintered (at ca. 1500 °C) in a hydrogen atmosphere.

Coating. Coating of the kernels is carried out by chemical vapor deposition (CVD) of pyrocarbon or silicon carbide from the gas phase in fluidized beds [229]. The core of the fluidized-bed installation is a gastight, water-cooled, cylindrical steel vessel with a built-in graphite tube as heating conductor [207], [230]. This conductor surrounds another graphite tube, the working tube, which has a conical base provided with inlet nozzles through which the coating gases and the carrier gas flow into the reaction chamber. The coating gases are pyrolytically decomposed at the temperatures prevalent here, resulting in the deposition of pyrocarbon and silicon carbide on the fluidized fuel kernels in the reaction chamber. The desired density and structure of the layers can be adjusted by the appropriate choice of process parameters such as temperature, coating gas concentration, and gas flow. The thickness of the layer is then a function of time. Coating is usually carried out in batches of 4 – 10 kg of fuel kernels.

The *buffer layer* of the TRISO coating is deposited by the decomposition of acetylene at about 1200 °C. The starting material for the two *dense pyrocarbon layers* is a mixture of acetylene and propene. The optimal decomposition temperature of this mixture is ca. 1300 °C. The *SiC layer* is deposited from methyl trichlorosilane at ca. 1500 °C.

The coated particles are removed from the fluidized bed and sieved. Nonspherical particles, which have a lower mechanical stability as a result of their shape, are separated with the help of vibration plates [208].

As far as irradiation properties are concerned, the important characteristics of quality are a minimum variation in layer thickness and layer density and, above all, as small an amount as possible of coated particles with a production-related, defective silicon carbide layer. With particles of good quality, the fraction of defect SiC-layers today is less than 10×10^{-6} ; i.e., less than 10 out of 10^6 particles have a defective silicon carbide layer.

Fabrication of Fuel Elements.

Spherical Fuel Elements. In Germany, the fabrication of spherical HTR fuel elements (Fig. 117) begins with coated fuel particles and binder-containing graphite powder, also called " molding powder" [207], [231]. This is prepared by kneading a graphite powder mixture with a solution of the binder (phenolic resin) in methanol, and by subsequently drying and grinding the kneaded material. In an overcoating process, the coated particles are provided with a layer of molding powder of approximately 200-µm thickness. This overcoating layer functions as a spacer, preventing direct point contact between the particles during pressing of the elements and, consequently, resulting breakage of particles. To guarantee the required overcoating quality, these particles are sieved and sorted on a vibration plate, as described for coated particles.

By starting with the overcoated particles and the molding powder, the fueled zone is first premolded and then, in a second step, a fuel-free shell of molding powder is shaped around the fueled zone. In view of the high anisotropy of graphite, both molding steps are carried out in silicone rubber molds. Hence, an approximately isostatic pressure distribution is achieved on pressing and, consequently, sufficiently good isotropy in the element.

The spheres attain their exact shape by mechanical processing (lathing). They are subsequently heated first to 800 °C in the absence of oxygen to carbonize the binder and then to 1950 °C in vacuum for gas removal and purification. Until now, this process has been used in Germany to produce approximately 1.2×10^6 fuel elements for application in AVR and THTR plants.

In the former Soviet Union, the production of spherical fuel elements is still in the developmental stage. One of the methods being investigated is the molding of binder-containing graphite powder and coated particles in steel molds at elevated temperature [232]. The binder used is coal tar pitch. In another production process, the use of a graphite matrix bound with pyrocarbon is being tested. Here, spherical compacts of relatively low density are produced by the molding of graphite powder and coated par-

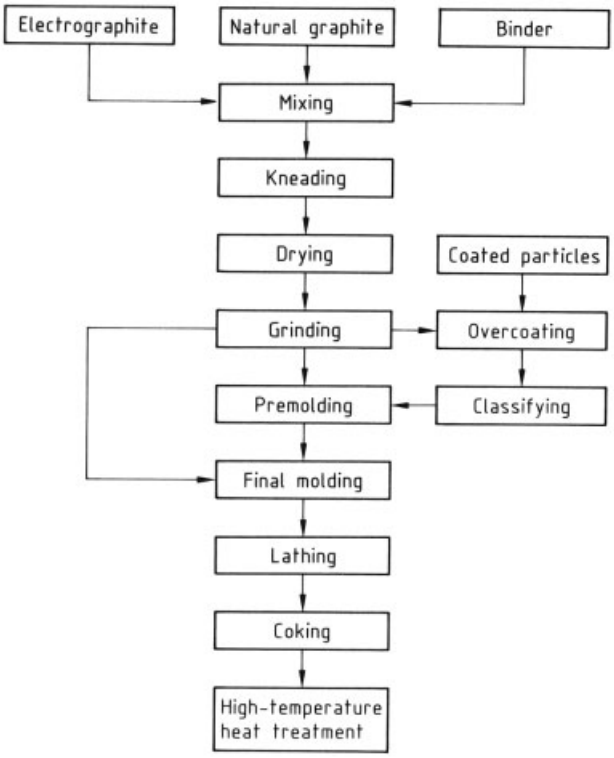

Figure 117. Fabrication of the spherical fuel element (Germany)

ticles in steel molds [233]. The insertion of pyrocarbon, which acts as binder, is achieved by the very slow pyrolytic decomposition of natural gas in the pores of the compacts (impregnation from the gas phase).

Prismatic Fuel Elements. Prismatic fuel elements, such as the elements produced in the United States for use in the Fort St. Vrain reactor, consist essentially of prefabricated electrographite blocks provided with axial holes and cylindrical fuel rods [234]. In the production of these fuel rods, dosing of the number of particles of fuel and fertile material required for each fuel rod is carried out first. After homogenization, the particle mixture is transferred to a steel mold and heated to approximately 170 °C. A heated, plastic matrix of graphite powder and pitch binder is then injected into this preheated particle bed, which has been fixed in shape in the mold. The "green" fuel rods are extruded after the mold is cooled. They are then packed into a bed of aluminum oxide powder and heated in inert gas to about 800 °C to carbonize the pitch binder. The low inherent stability of the green fuel rods in the heating phase necessitates this measure. The carbonized fuel rods are subsequently subjected to gas purification with hydrogen chloride at approximately 1700 °C and then to annealing in inert gas at about 1800 °C. Treatment with hydrogen chloride results not only in the removal of impurities from the graphite matrix but, more important in removal from the fuel rods of uranium and thorium released by particle breakage. The fuel rods are inserted into the corresponding holes in the graphite block, which are then sealed with graphite plugs.

Although the elements contain the same number of fuel rods, they must have different amounts of uranium and thorium, depending on their position in the reactor core. For this reason, two types of particles of fissile and fertile material containing different concentrations of heavy metal in each case – achieved by using the same layer thickness for different kernel thicknesses – are employed in the production of fuel elements for the Fort St. Vrain reactor (i.e., in

total four different types of particles). The different heavy-metal concentrations required in the fuel element were adjusted by suitable mixing of different types of particles. In the future, however, only one type of fissile and one type of fertile particles in a suitable combination with graphite particles for volume compensation will be used. This will greatly simplify the process.

The production of fuel elements for the *Japanese HTR*, presently in the planning stage, has only one exceptional feature, the fuel compacts [235]. The graphite blocks and graphite tubes are delivered prefabricated. The production and coating of uranium dioxide kernels with an average uranium enrichment of 6 % proceeds according to the processes described above.

As with the spherical THTR fuel element, production of the fuel compacts starts with coated particles and resinated graphite powder. In contrast to the production of spheres, the entire amount of resinated graphite required for a certain number of particles is granulated onto this number of particles. The overcoated particles are then molded into appropriate steel molds, at a temperature above the softening point of phenolic resin to the fuel compacts. The compacts are then carbonized and annealed as described for the fabrication of spherical elements. Fuel element assembly occurs in a subsequent step.

5.3.3. Fuel Assemblies made from Reprocessed Plutonium

5.3.3.1. Availability of Plutonium [236]

In the neutron field of the LWR, ^{239}Pu is formed from ^{238}U (by neutron capture and two β decays) and at the same time fissioned with the evolution of heat. At equilibrium between generation and fission, toward the end of burnup, plutonium contributes nearly 40 % of fission-produced heat. In the burnup range of interest, between 30 and 60 MW · d/kg fuel, the discharged fuel contains slightly less than 1 wt % plutonium, and with increasing burnup the isotopic composition of this element is shifted increasingly toward the even-numbered isotopes 238, 240, and 242 (which are barely fissile by thermal neutrons) until, eventually, only a little more than 60 % of the plutonium discharged remains fissile by thermal neutrons (i.e., the isotopes 239 and 241). Weapons grade plutonium, which is produced from very lightly irradiated fuel (less than 0.4 MW · d/kg), has very high concentrations of ^{239}Pu (93 – 95 %). Therefore, the use of light water power reactors for production of weapons grade plutonium is the most uneconomic way due to the inherent batch reload scheme of light-water reactors and has never been employed by the existing and suspected nuclear weapons states.

At present, the global total quantity of plutonium, in all forms, is about 1240 t (Table 56). Of this total, approximately two-thirds is contained in stored spent fuel from civil power reactors. The OECD Nuclear Energy Agency (NEA) estimates that 50 t of plutonium are generated worldwide in the 427 operating reactors (as of January 1998) every year. In 1995 the amount of "civil" plutonium (from reprocessed power reactor fuel) in the separated form, kept in stores, was 140 t. Two industrial factors currently control the transformation of separated plutonium in reactors: MOX fuel fabrication capacity and the number of reactors licensed to use MOX fuel. These parameters will control supply and demand for MOX in the coming years. Since the current capacity of MOX fuel fabrication plants for thermal reactors is limited, the world inventory of separated plutonium continues to increase. In 1995 the stock of separated plutonium increased by some 14 t (22 t were separated by reprocessing, 8 t of which were used as MOX in LWRs and fast reactor development programs).

Table 56. Present and predicted stocks of plutonium globally (in t) [237]

Civil plutonium	1995	2010
In spent reactor fuel	800	1400
Separated and stored plutonium	140	150
In fast reactor fuel cycle	30	550
In thermal MOX fuel cycle	20	
Subtotal	990	2100
Military plutonium	249	not available
Total	1239	

In addition worldwide, some 250 t of military plutonium are held, about a fifth of the total plutonium inventory. The largest quantity, nearly 150 t, is kept in Russia. Most of the remainder

is in the USA (about 100 t) with United Kingdom, France, and China probably holding less than 10 t each [238]. Following agreements between the USA and Russia to dismantle surplus nuclear weapons, the USA has declared 52.5 t of weapons grade plutonium as no longer required for defense purposes. Russia has also declared that it has a similar excess but not yet stated a precise quantity. The United Kingdom, France and China have not declared an excess of plutonium over their military requirements. Reduction of the quantity of surplus weapons plutonium is generally viewed as a higher priority than plutonium reduction of produced by reprocessing from civil nuclear power generation because of the former's greater potential as a security risk. Two main options for the management of surplus military plutonium are considered at present: (1.) Consumption as MOX fuel in reactors in a once-through fuel cycle, which renders surplus weapons plutonium into spent fuel and leads to the degradation of the plutonium's isotopic quality to a nonweapons grade level. (2.) Immobilization (e.g., by vitrification) and disposal in a final repository, this option is required for existing surplus plutonium which is unsuitable for use in MOX fuel because of, e.g., chemical contamination.

5.3.3.2. Utilization Strategies

From the reactor engineering standpoint, ^{235}U and ^{239}Pu (the most important fissile isotope of Pu) differ somewhat. At higher neutron energies plutonium has a favorable ratio of fission to capture cross-sections so that, in the presence of ^{238}U, more plutonium can be generated from this nuclide than is consumed in fission reactions. This breeding process takes place in the breeder reactors with fast neutrons (see Chap. 4). Currently, fuel assemblies of most fast breeder reactors consist of stacked pellets of a mixed UO_2–PuO_2 ceramic containing up to 28 % fissile plutonium and enclosed in steel cladding tubes.

5.3.3.3. Requirement for Mixed-Oxide Thermal Fuel Assemblies

As long as the demand for plutonium breeder fuel assemblies remains lower than the rate of plutonium production, it is desirable to reduce the excess by recycling it to existing LWRs. The recycled mixture of plutonium isotopes has a fission cross-section > 30 % greater than that of ^{235}U, so neutron physics would be optimized by distributing this fissile material as evenly as possible in the reactor to prevent local overheating. However, homogeneous mixing with the total uranium dioxide in all the pellets throughout the entire reactor, is uneconomical because the high radiotoxicity of plutonium would increase processing costs and complexity. Mixed-oxide fuel assemblies are therefore employed. In each row of fuel rods in such an assembly, the plutonium content increases from outside to inside. As many as six different contents of fissile Pu (range 0.4 – 1.7 of average) are used in boiling-water reactors; three (0.5 – 1.2 of average), in pressurized-water reactors [239]. To flatten neutron-flux peaks when fissile contents are increased for higher burnups, hollow or water-filled rods and/or individual UO_2 – Gd_2O_3 rods are included in the fuel assembly. Since the assemblies are repositioned in the reactor several times over their service lives to obtain uniform burnup, MOX assemblies must also be interchangeable with UO_2 assemblies, so that the two kinds have the same geometry and are subject to the same limits on thermal and mechanical loads. Mixed reactor core designs with up to 30 – 50 % MOX fuel assemblies have been licensed [240], [241]. Also full MOX cores are feasible. The corresponding assembly design for PWRs is even simpler than that of mixed cores. For full MOX cores a lower content of fissile plutonium in the peripheral fuel rods is not required which is necessary in mixed cores to accommodate for the spectral shift between MOX and UO_2 fuel assemblies (a "harder", i.e., less moderated neutron spectrum exists in MOX assemblies). MOX manufacturing plants as well as transportation and reactor operations with MOX fuel have to – and can – meet the same safety and environmental standards as applied in the UO_2 fuel cycle. The same amount of electricity can be generated from a MOX fuel assembly as from a substituted UO_2 assembly, provided they are operated to the same burnup.

5.3.3.4. Requirement for Fast Breeder Fuel Assemblies

Since the fuel for a fast breeder consists solely of a mixture of uranium and plutonium, the chemical form and geometry of fuel rod and assembly can in principle be chosen freely, in contrast to the recycled fuel for a thermal reactor which must match the enriched UO_2 fuel. The constraints are that (1) the fuel must be chemically compatible with the sodium coolant and steel cladding at the operating temperature, and (2) the fuel must withstand the high neutron fluences (up to 10^{23} neutrons/cm^3 over the fuel-assembly service life) without massive amounts of strain energy. Deformation due to power excursions has to remain controllable.

Therefore, mixed-oxide fuel pins with austenitic steel cladding have been developed (wall thickness ca. 0.4 mm); the fuel pellets are between 6 and 7.5 mm in diameter. The highest central temperature in the pins (800 – 1000 °C) is far below the melting point of the mixed oxide (2500 °C). The stack of pellets displays inherent self-regulation in power excursions (decrease in reactivity because of axial expansion combined with Doppler broadening of resonance peaks when the temperature rises, see Section 1.3.6); and these facts make the system relatively stable. Ample experience in working with fissile oxides means that processing difficulties need not be feared. However, the drawback – the relatively low thermal conductivity of the oxides – dictates a small pin diameter (6 – 8 mm). Another disadvantage is that conversion of tetravalent fissile materials to mono-, di-, or trivalent fission products leads to more than the stoichiometric quantity of oxygen in the fuel matrix as burnup increases (i.e., the presence of free oxygen). Free oxygen, in combination with fission-product iodine and cesium, can lead to corrosion of the cladding tubes from inside, in addition cesium forms chromates with the cladding tube material. As a result, it has become necessary to fabricate the oxides somewhat substoichiometricly.

In contrast, not only do alternative fuels, such as monocarbides, mononitrides, or carbonitrides of uranium and plutonium, have a threefold better thermal conductivity than the oxides, but also their higher density permits a higher breeding ratio with a given neutron flux. However, advanced methods are required for production and reprocessing of these fuels. Tests on carbide fuel assemblies have been carried out in only a few of the operating breeders. Only one small reactor (BR-5, former Soviet Union) had a complete carbide fuel inventory.

Table 57. Radiation protection

Protection against	Protective measures
Intake of, or contamination by Pu	α-tight enclosure (glove box, welded plastic bags), pressure difference by forced ventilation
Low-energy γ radiation from Pu (low intensity)	lead shielding in glove boxes
Low-energy γ radiation from daughter nuclides, e.g., ^{241}Am (high intensity)	separation of ^{241}Am from Pu before processing of Pu, lead shielding
	use of hot- cell technology (difficulties were encountered during tests in one case)
	use of mechanization or automation to cut down on manual operations
Neutrons from spontaneous fission (mainly ^{249}Pu and α, n reactions)	water walls or plastic shielding with added absorbers (B, Cd)
	automation
Nuclear excursions (criticality)	limitation on quantities with allowance for moderating ratio and density
	"safe geometry" design of plant
	use of B, Cd, Hf as neutron absorbers (poisons)
Environmental effects	design of buildings and facilities in compliance with guidelines of the Reactor Safety Committee (Reaktorsicherheitskommission)
Third-party action or misuse	surveillance and accountability of nuclear materials

5.3.3.5. Problems in Handling Plutonium

Both the fissiability and the high specific α activity of the odd-numbered plutonium isotopes make this element difficult to work with (see Table 57) [242]. Plutonium recovered in LWR fuel-assembly reprocessing has a specific α activity of 5 – 25 GBq/g and minimum critical masses of ca. 13 kg in dry spherical geometry or ca. 600 g when optimally moderated in water. The specific α activity means that plutonium is highly radiotoxic; its chemical toxicity as a heavy metal is not nearly as important. For persons occupationally exposed to plutonium the limit of intake of plutonium oxide powder based on a current metabolic model is only 400 Bq/a in the lung. Dust inhalation is the most important intake route because the range of the inhaled α

particles is sufficient to reach the lung tissue. This limit is reached if only one dust particle weighing 0.04 μg is inhaled. The atmosphere of the area in which plutonium is handled must therefore be reliably isolated from the worker's breathing air. Alpha-tight enclosure is accomplished if gastight gloveboxes are used. In newer processing plants, manual work is reduced as much as possible through automation, i.e., production is carried out automatically, repair and maintenance manually. Plutonium areas are further shielded – by walls containing lead and moderators (water, plastics) – against γ radiation (from daughter products of ^{236}Pu and ^{241}Pu) and neutrons [from even-numbered plutonium isotopes, chiefly ^{240}Pu, and from (α, n) reactions with $^{17}O_2$ and $^{18}O_2$].

So that critical configurations can be totally avoided during processing, moderators are excluded wherever possible [this means that aqueous solutions of plutonium salts can only be processed if the surface – volume ratio is large enough to guarantee sufficient neutron leakage ("geometrically safe equipment")], and the quantities handled in one place are limited to less than half the critical mass. In many cases, neutron absorbers (boron, hafnium, gadolinium, cadmium) are also employed.

Many obstacles exist to the misuse of plutonium from irradiated LWR fuel for weapons purposes to avoid liberation of its potential explosive force. Most important in this respect is the chance of preignition due to neutron-emitting higher isotopes, requiring a highly sophisticated ignition technology. Even then the explosive yield is principally unpredictable. However, reactor plutonium cannot principally be excluded as raw material for weapons. These facts have led to the most stringent security measures to guard against the diversion or misuse of this material. National and international inspectorates monitor the flow of fissile materials. Access to these materials is under strict surveillance. Sensitive monitors prevent undetected removal from the controlled area.

5.3.3.6. Fabrication of Mixed-Oxide Fuel Assemblies

Aside from experimental quantities of metal and carbide fast breeder fuel, the processing of plutonium into fuel assemblies today resembles uranium dioxide technology (see Section 5.3.1.5). Earlier, several attempts were made to introduce simpler techniques (which would better lend themselves to automation and remote control), in place of the classical ceramic operations with many individual steps, which require the handling of fine powders; high geometrical precision in molding, sintering, and grinding; and thus large amounts of manual work. None of these experiments, however, was successful enough for the developers to undertake the costly process of gaining approval for widespread power-reactor use.

Modifications of UO_2 Process for MOX Fuel Assemblies. An important point with regard to fuel use is that 200 times as many fissions per unit time occur in a plutonium dioxide particle as in an adjacent particle of natural uranium dioxide. For this and a number of other reasons (Doppler broadening, self-shielding), the plutonium dioxide must be distributed as finely and uniformly as possible in the uranium dioxide matrix. On the other hand, because the surface of a ceramic in normal air is always coated with adsorbed water, the flowability of powders declines as grain size decreases and the tendency to self-agglomeration increases. Thus, it becomes more difficult to charge the powder into molds, and problems in industrial pellet fabrication are created. Therefore, fine plutonium dioxide powder (mean grain size 1 – 2 μm; e.g., produced by calcination of precipitated plutonium oxalate) was mixed carefully with flowable uranium dioxide (agglomerate size 20 – 40 μm; product from AUC process) in such a way that the fine plutonium dioxide coated the uranium dioxide grains, and the still-flowable powder was subsequently molded. This method was technologically simple and successful. However, this approach had to be abandoned when it was discovered that interdiffusion of uranium dioxide and plutonium dioxide (both in sintering and during later in-pile irradiation) was insufficient for the formation of homogeneous (U,Pu) O_2 mixed crystals. Therefore in reprocessing, up to 15 % of the plutonium dioxide remained in the dissolver sludge, in acid-insoluble form, when the fuel was dissolved in nitric acid. In response to this finding, the fuel assembly, even before irradiation, was required to contain plutonium in HNO_3-soluble

form (i.e., as a homogeneous mixed crystal with less than 40 % Pu). This standard is met in a variety of ways today. With plutonium nitrate as starting product, a mixture of $Pu(NO_3)_4$ and $UO_2(NO_3)_2$ solutions with a maximum plutonium content of 40 % is prepared. The process developed by the Power Reactor and Nuclear Corporation (PNC) in Japan converts this mixture to the oxides by (1) evaporating the solution in a microwave field, (2) calcining the salts, (3) reducing the agglomerate size by milling and sieving, and (4) reducing to the dioxides with hydrogen. A process tested at Risley (BNFL, UK) starts with stabilized mixed nitrate solutions, generates droplets, and solidifies these to sol-gel spheres in gaseous ammonia. The particles are calcined to obtain a moldable charge of spherical homogeneous mixed crystals. Siemens developed the *A(UPu)C process*, a modified AUC process (see Section 5.3.1.3). Here, a mixed solution of the uranium (VI) and plutonium (VI) nitrates, is used to yield $(NH_4)_4[(U, Pu)O_2](CO_3)_3$ as an intermediate; this product is calcined to give homogeneous $(U, Pu)O_2$. This "A (U, Pu)C" process, like the microwave technique, has been tested on several hundred kilograms of plutonium; it yields a powder that requires no further treatment before being mixed with UO_2 (from the AUC process) and molded.

The *Cogema reprocessing plant*, however, produces PuO_2 from the oxalate, and the product must be converted into a completely acid-soluble MOX fuel [243]. Two very similar processes have been developed for this purpose: Mimas at Belgonucleaire (and later improved at Cogema) and OCOM at Siemens. Each is based on the intensive milling and mixing of part of the UO_2 with PuO_2. In the Mimas process, UO_2 agglomerates from very small (1 – 5 μm) primary particles (from the CEA dry conversion plant at Cadarache), with a tendency toward self-agglomeration, or UO_2 from AUC processes are mixed with the PuO_2 and ground for about 24 h in a large tubular ball mill. The powder is then put through a self-agglomeration process after removal of grinding balls, mixed with more UO_2 from the same source, and molded into pellets. In the OCOM process, PuO_2 and UO_2 (from the AUC process) are the feeds. A few tenths of a percent of a grinding aid is added to prevent self-agglomeration, the mixture is ground for 12 h to achieve a very high solubility, and the resulting powder is mixed with more AUC-product uranium dioxide and molded.

The added plutonium modifies the properties of the ceramics because plutonium dioxide, in contrast to uranium dioxide, can form a stable sesquioxide Pu_2O_3 under the reducing conditions in the sintering furnace (the atmosphere contains at least ten times more hydrogen than water). As a result, the product uranium dioxide is slightly substoichiometric. The Pu_2O_3, if present in isolated grains in the sinter, is an interfering species not just because its crystal lattice is different (hexagonal instead of face-centered cubic), but also because it can incorporate hydrogen (formed by the radiolysis of adsorbed water) into interstitial lattice sites in a particularly stable way [244]. To prevent local hydride damage (sunburst defects) on the inside of the Zircaloy cladding, the content of $(H_2 + H_2O)$ in the ceramic matrix should be held below 10 ppm determined as H_2O. This requires that the pore shape, pore size, and residual moisture content in the ceramic be carefully matched. If a large fraction of fission products must be accomodated by the fuel, a fast breeder fuel is required that contains a certain amount of pores of a distinct shape and size, i.e., its density is outside the common density range for sintering (92 – 96 % of theoretical). In this case pore-forming agents such as ammonium uranyl carbonate, zinc stearate, or oxamide are added.

Another feature of MOX fabrication is that the hydrogen used in sintering is diluted with inert gases (N_2, Ar) to below the explosion limit (< 8 %) for safety reasons. Particular care is needed so that traces of plutonium will not be deposited on the outside of the cladding tube when the pellets are charged into it. Only in this way can the welded fuel rods be handled like uranium dioxide rods.

Testing and Analysis. Mixed-oxide fuel rods also undergo several special inspections. In addition to the absence of washable contamination, a uniform plutonium – uranium distribution must be verified, and a γ-spectrometer scan must be performed to make sure that all pellets in a rod belong to a certain production batch.

Assembling. The techniques used for making fuel assemblies from uranium fuel rods can be applied directly. Since a PWR assembly with 500 kg of fuel contains up to 25 kg of plutonium,

the longer-lasting operations must be carried out behind neutron shielding, which usually consists of stacks of borate glass and plastic slabs.

5.3.3.7. Radiation Protection and Safety Aspects

In addition to the legislation and regulations that govern environmental protection and accident prevention in conventional plants, a large number of further regulations and directives apply to the design of a plant for plutonium fuel assemblies. These ultimately bear on the problems listed in Table 57.

5.4. Chemical Reprocessing of Nuclear Fuels

Introduction. The scope of Section 5.4 is the industrial reprocessing of commercial power reactor fuels. The discussion will concentrate on the treatment of fuel elements from light-water reactors because at present these form the most widely used class of commercial nuclear reactors worldwide. A short outlook to the requirements of fast breeder reactor fuel recycle will also be given because if nuclear energy is to become a long-term energy source, the development of FBR systems to industrial maturity will be inevitably required (Section 5.1). In both reactor types, oxide fuels are exclusively (LWR) or preferably (FBR) being used, i.e., slightly enriched uranium dioxide in the LWR, and uranium – plutonium mixed oxide (MOX) in both the LWR and the FBR. As a consequence, this chapter will deal particularly with oxide fuel reprocessing.

History. For a better understanding of the evolution of current reprocessing technology, a short historical review is given [245–248]. Chemical processing techniques for neutron-irradiated uranium fuels were first developed with the aim of producing plutonium for weapons applications. The "production" reactors used as irradiation sources had in common the utilization of uranium metal fuels. These were only slightly irradiated to burnup levels of a few hundred MW · d/t, to obtain a product of high ^{239}Pu isotopic purity, but were usually processed after short cooling periods, down to ca. 100 days. The first generation of commercial power reactors in Britain and France were of the gas – graphite reactor type (see Section 3.2). These also applied uranium metal fuels, which were irradiated to burnup levels of ca. 2500 to 4000 MW · d/t, but the cooling periods before chemical reprocessing were increased to a few years.

The first industrial process for plutonium production, the so- called *Bismuth Phosphate Process* (BIPO), had been developed during World War II in the frame of the U.S. Manhattan Project, and was at first performed in a pilot plant at the Clinton Works, Oak Ridge, Tennessee, which later became the Oak Ridge National Laboratory (ORNL). A full-size industrial plant was then constructed at Hanford, Washington, which operated until 1951. The BIPO Process was a batch precipitation process in which the coprecipitation of Pu^{4+} with bismuth phosphate from a nitric acid solution of the fuel was the essential separation step.

All the later industrial processes, both for defense and commercial purposes, were based on *solvent extraction*, a separation technique highly suitable for continuous multi-stage operation, thus offering the potential for high throughput combined with excellent separation efficiency and product yield. The *Redox Process* performed from 1951 to 1966 in the Redox Plant at Hanford in the United States, applied methyl isobutyl ketone ("hexone") to coextract both hexavalent plutonium (oxidized by chromate) and uranium from a nitrate fuel solution. The separation of the two products was performed after reduction of Pu (VI) (hence the name "Redox") to inextractable Pu (III) with "ferrous sulfamate" [iron (II) amidosulfate]. A similar process was the *Butex Process* developed in the United Kingdom, which was applied between 1952 and 1964 in the Windscale I Plant at Sellafield, Cumbria. The extractant, called " butex" or " dibutyl carbitol" (diethylene glycol dibutyl ether) extracts uranium (VI) and plutonium, both as Pu (IV) or Pu (VI), from a nitric acid fuel solution. For uranium – plutonium separation again the inextractability of Pu (III) was utilized.

The *PUREX-Process* which is the sole reprocessing procedure in industrial use today, applies tri-*n*-butyl phosphate (TBP) in a kerosene-type aliphatic hydrocarbon diluent as the extractant for uranium and plutonium. Again the

process utilizes the highly selective extractability of tetra- and hexavalent uranium and plutonium nitrates to separate them from the bulk of fission products in a nitric acid medium, and the plutonium – uranium partitioning is performed by reduction of plutonium to inextractable Pu (III). The PUREX process was developed in the late 1940s by LANHAM and RUNION at Oak Ridge, who also gave it the name PUREX (Plutonium, Uranium, Reduction, EXtraction) [249]. After successful pilot-plant demonstration at ORNL, the process was applied in two large-scale plants at Aiken, South Carolina (Savannah River Plant, operated from 1954), and Hanford, Washington (operated from 1956), and was later adopted worldwide both for military and commercial application. For reviews, see [245–248], [250], [251].

In addition to *aqueous processes* in which the fuel is first dissolved in nitric acid before some type of high-effi-ciency aqueous separation chemistry is carried out, a variety of *nonaqueous* alternatives has been investigated, with the hope of arriving at a simpler and more compact recycle procedure. Most of these processes achieved only a low separation efficiency, i.e., a low decontamination of uranium and plutonium from radioactive fission products, with the effect that the uranium recovered would have been unsuitable for reenrichment, and fuel refabrication would have to be carried out by remote techniques under high shielding. The only high-efficiency separation procedure among the nonaqueous approaches, the *Fluoride Volatility Process*, was based on the volatility of actinide hexafluorides but was hampered by the chemical instability of PuF_6, the high corrosivity of fluorine compounds, and many other drawbacks. None of the nonaqueous procedures could be developed to industrial maturity. For reviews see [248], [252], [253].

5.4.1. Reprocessing of LWR Fuel Elements

5.4.1.1. General Scheme

After discharge from the reactor and pre-cooling on the reactor site, spent LWR fuel elements are transported in shielded containers to the reprocessing plant where they are transferred into an on-site fuel element storage facility for further cooling. Both underwater storage in a cooling pond, and dry storage in air-cooled shielded containers are in common use.

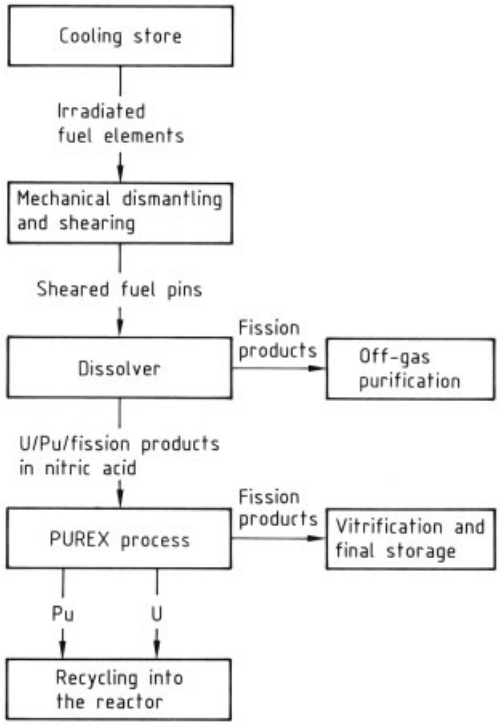

Figure 118. Block scheme of nuclear fuel reprocessing

The chemical processing of LWR fuel elements is usually performed after a total cooling period between 3 and 7 years. This allows the short-lived radioactive fission product and actinide nuclides to decay, while only those with half-lives > 80 d survive to any noteworthy extent. Figure 118 presents a block scheme of the various steps that are carried out in a reprocessing plant. First, the fuel element is disassembled, and the fuel pins are sheared or "chopped" into short pieces ("mechanical head-end"). These pieces are transferred into a dissolver where the uranium oxide fuel matrix is dissolved or "leached" with boiling nitric acid, leaving behind the chopped Zircaloy cladding hulls which are not attacked by the acid ("chemical head-end"). The entire pretreatment is called *shear – leach* or *chop-and-leach process.* Next, the dissolver solution is clarified and transferred to the PUREX solvent extraction plant where uranium and plutonium are separated from fission products, byproduct actinides, and from

each other by countercurrent multistage extraction with 30 % TBP in a kerosene-type diluent. The two products are then further purified, usually by one or two additional cycles of TBP solvent extraction ("tail end"), before they are recycled for conversion into new fuel materials. The fission product waste stream is concentrated by evaporation, and stored as a high-activity waste (HAW) solution in cooled tanks before it is transferred to the waste treatment plant for solidification into a product which can be permanently stored. In a similar manner, solid and liquid medium- (MAW) and low-activity wastes (LAW) are concentrated and solidified for safe long-term disposal. Gaseous effluents from all process stages must be purified from radioactive and chemical pollutants before they are released to the environment.

5.4.1.2. Fuel Composition and Purification Requirements

The composition of a typical irradiated LWR fuel is given in Tables 58 and 59. The fuel selected as example is a Biblis A-type PWR (pressurized water reactor) UO_2 fuel of 3.4 wt % initial ^{235}U enrichment, irradiated to 36 000 MW · d/t burnup, and cooled after reactor discharge for three, five, or seven years. All data were calculated using the KFK KORIGEN code, and are given per tonne initial uranium. Inventories of mass, radioactivity, and thermal power are presented for actinides (Table 58) and fission products (Table 59). Long-term trends for the radioactivity and decay heat of both the fission products and the actinides are shown in Figure 119. The two tables demonstrate that literally "half the periodic table" is formed by fission or neutron capture during fuel irradiation, but they also show that after decay periods of ≥ 3 years, only a relatively small number of nuclides still make an essential contribution to the fuel's radioactivity and thermal power. Whereas for short- and medium-term periods both parameters are dominated by the radioactive fission product nuclides, the long-lived isotopes of plutonium, americium, and neptunium become the dominant factors after prolonged decay times (Fig. 119). This aspect is important with respect to the permanent disposal of the radioactive waste products.

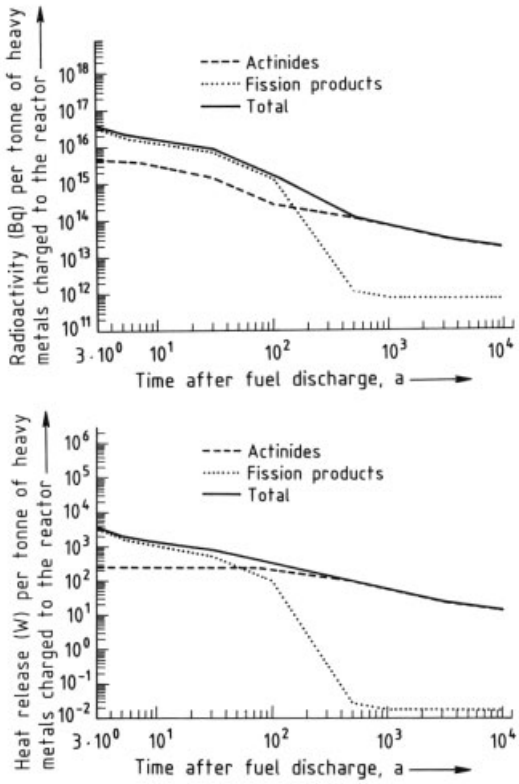

Figure 119. Radioactivity (A) and decay heat (B) of Biblis A-type PWR fuel of 36 GW · d/t burnup as a function of time

The degree of purification which the product values require, is primarily determined by the demand that these can be handled "directly" or "hands-on" in the prereactor fuel cycle facilities into which they are recycled, to avoid the need of heavy shielding and remote operation of these plants. This means that uranium must be decontaminated from both the β,γ-emitting fission products, and the α-emitting transuranium actinides including plutonium, to an activity level equivalent to that of natural uranium in equilibrium with its decay daughters. For plutonium the primary requirement is to reduce the γ emitters to a level corresponding to its own γ and neutron dose rate so that it can be handled in glove-box equipment. In addition, "conventional" impurities such as corrosion products or water and reagent impurities must be restricted in the products. This is particularly true for neutron-absorbing impurities (expressed as the "boron equivalent", i.e., the amount of ^{10}B

Table 58. Actinide inventory of irradiated PWR type-Biblis uranium fuel. Initial enrichment 3.4 wt % ^{235}U, 36 GWd/thm burnup (calculated with KFK-KORIGEN)

Elements concentration, g/thm[a]	Years after discharge			Nuclides concentration, g/thm[b]	Years after discharge		
	3	5	7		3	5	7
U	9.522×10^5	9.522×10^5	9.522×10^5	^{234}U	1.520×10^2	1.546×10^2	1.571×10^2
Np	4.927×10^2	4.935×10^2	4.947×10^2	^{235}U	8.176×10^3	8.176×10^3	8.176×10^3
Pu	9.858×10^3	9.755×10^3	9.661×10^3	^{236}U	4.408×10^3	4.409×10^3	4.409×10^3
Am	3.113×10^2	4.126×10^2	5.042×10^2	^{238}U	9.394×10^5	9.394×10^5	9.394×10^5
Cm	3.055×10^1	2.835×10^1	2.643×10^1	^{237}Np	4.927×10^2	4.935×10^2	4.947×10^2
Total	9.629×10^5	9.629×10^5	9.629×10^5	^{238}Pu	1.688×10^2	1.663×10^2	1.637×10^2
				^{239}Pu	5.632×10^3	5.631×10^3	5.631×10^3
				^{240}Pu	2.401×10^3	2.403×10^3	2.404×10^3
				^{241}Pu	1.113×10^3	1.011×10^3	9.184×10^2
				^{242}Pu	5.433×10^2	5.433×10^2	5.433×10^2
				^{241}Am	2.104×10^2	3.118×10^2	4.034×10^2
				^{243}Am	1.004×10^2	1.004×10^2	1.004×10^2
				^{244}Cm	2.789×10^1	2.584×10^1	2.394×10^1
				^{245}Cm	1.076×10^0	1.076×10^0	1.076×10^0
				^{246}Cm	1.085×10^0	1.085×10^0	1.085×10^0
				Total	9.629×10^5	9.629×10^5	9.629×10^5

Nuclides activity Bq/thm[c]	Years after discharge			Nuclides thermal power, W/thm[d]	Years after discharge		
	3	5	7		3	5	7
^{239}Np	7.411×10^{11}	7.410×10^{11}	7.408×10^{11}	^{238}Pu	9.583×10^1	9.440×10^1	9.292×10^1
^{238}Pu	1.069×10^{14}	1.053×10^{14}	1.037×10^{14}	^{239}Pu	1.089×10^1	1.089×10^1	1.088×10^1
^{239}Pu	1.296×10^{13}	1.296×10^{13}	1.296×10^{13}	^{240}Pu	1.705×10^1	1.707×10^1	1.707×10^1
^{240}Pu	2.025×10^{13}	2.027×10^{13}	2.028×10^{13}	^{241}Pu	3.537×10^0	3.212×10^0	2.918×10^0
^{241}Pu	4.246×10^{15}	3.856×10^{15}	3.502×10^{15}	^{241}Am	2.415×10^1	3.578×10^1	4.629×10^1
^{241}Am	2.673×10^{13}	3.961×10^{13}	5.125×10^{13}	^{242}Cm	1.636×10^1	8.466×10^{-1}	1.480×10^{-1}
^{243}Am	7.411×10^{11}	7.410×10^{11}	7.408×10^{11}	^{244}Cm	7.899×10^1	7.317×10^1	6.778×10^1
^{242}Cm	1.643×10^{13}	8.501×10^{11}	1.486×10^{11}	Total	2.483×10^2	2.369×10^2	2.395×10^2
^{243}Cm	6.850×10^{11}	6.525×10^{11}	6.215×10^{11}				
^{244}Cm	8.353×10^{13}	7.738×10^{13}	7.168×10^{13}				
Total	4.515×10^{15}	4.115×10^{15}	3.765×10^{15}				

thm = tonne heavy metal
[a] Elements < 10 g/thm not included. [b] Nuclides < 1 g/thm not included. [c] Nuclides $< 3.7 \cdot 10^{11}$ Bq = (10 Ci)/thm not included. [d] Nuclides < 1 W/thm not included.

having the same neutron absorption capacity as the elements in question), and in the case of uranium for elements forming nonvolatile fluorides which would disturb the UF_6 conversion stage. Typical product specifications resulting from such considerations are given in Table 60.

The so-called *decontamination factor (DF)*, defined as

$$DF = \frac{\text{(impurity : product) before purification}}{\text{(impurity : product) after purification}}$$

which must be attained in PUREX for the radioactive fuel contaminants, is obtained by relating product specifications with fuel composition data as given in Tables 58 and 59. For instance, the uranium represented in Tables 58 and 59 and cooled for 3 years, requires a *DF* of ca. 2×10^6 for γ-emitting fission products and of ca. 1×10^6 for plutonium to meet the specifications of Table 60, whereas for the plutonium the respective figures are ca. 2×10^6 for fission products and 3×10^4 for uranium.

Table 59. Fission product inventory of irradiated PWR type-Biblis uranium fuel. Initial enrichment 3.4 wt % ^{235}U, 36 GW · d/thm burnup (calculated with KFK-KORIGEN)

Elements	Years after discharge			Nuclides	Years after discharge			Nuclides	Years after discharge		
concentration, g/thm [a]	3	5	7	activity, Bq/thm [b]	3	5	7	thermal power, W/thm [c]	3	5	7
H	4.760×10^{-2}	4.254×10^{-2}	3.803×10^{-2}	^{3}H	1.700×10^{13}	1.520×10^{13}	1.358×10^{13}	^{85}Kr	1.231×10^{1}	1.082×10^{1}	9.509×10^{0}
Se	6.000×10^{1}	6.000×10^{1}	6.000×10^{1}	^{85}Kr	3.042×10^{14}	2.673×10^{14}	2.349×10^{14}	^{90}Sr	8.613×10^{1}	8.212×10^{1}	7.831×10^{1}
Br	2.373×10^{1}	2.373×10^{1}	2.373×10^{1}	^{90}Sr	2.746×10^{15}	2.618×10^{15}	2.496×10^{15}	^{90}Y	4.114×10^{2}	3.923×10^{2}	3.740×10^{2}
Kr	3.934×10^{2}	3.909×10^{2}	3.887×10^{2}	^{90}Y	2.746×10^{15}	2.619×10^{15}	2.497×10^{15}	^{106}Ru	4.140×10^{0}	1.047×10^{0}	2.647×10^{-1}
Rb	3.752×10^{2}	3.777×10^{2}	3.799×10^{2}	^{95}Zr	4.125×10^{11}	1.511×10^{8}	5.535×10^{4}	^{106}Rh	6.698×10^{2}	1.694×10^{2}	4.282×10^{1}
Sr	9.269×10^{2}	9.016×10^{2}	8.775×10^{2}	^{95}Nb	9.158×10^{11}	3.355×10^{8}	1.229×10^{5}	$^{110\,m}$Ag	2.754×10^{0}	3.632×10^{-1}	4.790×10^{-2}
Y	5.010×10^{2}	5.010×10^{2}	5.010×10^{2}	^{99}Tc	5.198×10^{11}	5.198×10^{11}	5.198×10^{11}	^{125}Sb	1.454×10^{1}	8.801×10^{0}	5.327×10^{0}
Zr	3.906×10^{3}	3.932×10^{3}	3.956×10^{3}	^{106}Ru	2.584×10^{15}	6.534×10^{14}	1.652×10^{14}	^{134}Cs	6.797×10^{2}	3.470×10^{2}	1.772×10^{2}
Nb	3.219×10^{-3}	3.302×10^{-3}	4.018×10^{-3}	^{106}Rh	2.584×10^{15}	6.534×10^{14}	1.652×10^{14}	^{137}Cs	1.175×10^{2}	1.122×10^{2}	1.072×10^{2}
Mo	3.670×10^{3}	3.670×10^{3}	3.670×10^{3}	$^{110\,m}$Ag	6.102×10^{12}	8.048×10^{11}	1.061×10^{11}	$^{137\,m}$Ba	3.947×10^{2}	3.769×10^{2}	3.599×10^{2}
Tc	8.283×10^{2}	8.283×10^{2}	8.283×10^{2}	$^{113\,m}$Cd	1.294×10^{12}	1.177×10^{12}	1.070×10^{12}	^{144}Ce	5.480×10^{1}	9.234×10^{0}	1.556×10^{0}
Ru	2.416×10^{3}	2.401×10^{3}	2.397×10^{3}	^{125}Sb	1.721×10^{14}	1.042×10^{14}	6.305×10^{13}	^{144}Pr	6.073×10^{2}	1.023×10^{2}	1.724×10^{1}
Rh	4.984×10^{2}	4.984×10^{2}	4.984×10^{2}	$^{125\,m}$Te	4.200×10^{13}	2.543×10^{13}	1.539×10^{13}	^{147}Pm	2.695×10^{1}	1.589×10^{1}	9.368×10^{0}
Pd	1.448×10^{3}	1.463×10^{3}	1.467×10^{3}	^{127}Te	4.522×10^{11}	4.350×10^{9}	4.184×10^{7}	^{154}Eu	9.772×10^{1}	8.318×10^{1}	7.080×10^{1}
Ag	6.862×10^{1}	6.859×10^{1}	6.859×10^{1}	$^{127\,m}$Te	4.617×10^{11}	4.441×10^{9}	4.271×10^{7}	^{155}Eu	4.001×10^{0}	3.025×10^{0}	2.288×10^{0}
Cd	8.015×10^{1}	8.016×10^{1}	8.015×10^{1}	^{134}Cs	2.471×10^{15}	1.262×10^{15}	6.442×10^{14}	Total	3.186×10^{3}	1.715×10^{3}	1.256×10^{3}
In	1.473×10^{0}	1.487×10^{0}	1.501×10^{0}	^{137}Cs	3.932×10^{15}	3.755×10^{15}	3.585×10^{15}				
Sn	5.121×10^{1}	5.121×10^{1}	5.121×10^{1}	^{137m}Ba	3.720×10^{15}	3.552×10^{15}	3.391×10^{15}				
Sb	1.373×10^{1}	1.196×10^{1}	1.089×10^{1}	^{144}Ce	3.057×10^{15}	5.151×10^{14}	8.679×10^{13}				
Te	4.942×10^{2}	4.960×10^{2}	4.971×10^{2}	^{144}Pr	3.057×10^{15}	5.151×10^{14}	8.679×10^{13}				
I	2.415×10^{2}	2.415×10^{2}	2.415×10^{2}	$^{144\,m}$Pr	3.668×10^{13}	6.181×10^{12}	1.041×10^{12}				
Xe	5.819×10^{3}	5.819×10^{3}	5.819×10^{3}	^{147}Pm	2.780×10^{15}	1.639×10^{15}	9.665×10^{14}				
Cs	2.805×10^{3}	2.724×10^{3}	2.659×10^{3}	^{151}Sm	1.279×10^{13}	1.259×10^{13}	1.240×10^{13}				
Ba	1.683×10^{3}	1.763×10^{3}	1.829×10^{3}	^{154}Eu	4.042×10^{14}	3.441×10^{14}	2.928×10^{14}				
La	1.328×10^{3}	1.328×10^{3}	1.328×10^{3}	^{155}Eu	2.035×10^{14}	1.539×10^{14}	1.164×10^{14}				
Ce	2.630×10^{3}	2.608×10^{3}	2.604×10^{3}	Total	3.088×10^{16}	1.871×10^{16}	1.484×10^{16}				
Pr	1.222×10^{3}	1.222×10^{3}	1.222×10^{3}								
Nd	4.406×10^{3}	4.428×10^{3}	4.431×10^{3}								
Pm	8.102×10^{1}	4.777×10^{1}	2.817×10^{1}								
Sm	8.494×10^{2}	8.824×10^{2}	9.018×10^{2}								
Eu	1.552×10^{2}	1.465×10^{2}	1.394×10^{2}								
Gd	1.310×10^{2}	1.399×10^{2}	1.472×10^{2}								
Total	3.711×10^{4}	3.711×10^{4}	3.711×10^{4}								

[a] Elements with fission yield < 0.1 % not included (exception: H). [b] Nuclides with activity $< 3.7\times10^{11}$ Bq = (10 Ci)/thm not included. [c] Nuclides with thermal power < 1 W/thm not included.

Table 60. Typical product specifications

	U	Pu
Total transuranium element α activity, Bq/g	250	
Total fission product γ activity, Bq/g	1.85×10^4	1.48×10^6
Boron equivalent, [a] ppm ^{10}B	8	10
Metals forming nonvolatile fluorides, [b] ppm	500	
Total metals impurities, [c] ppm		5000
Single elements impurities, ppm		
Ca		300
Ti	1	
Th		200
Sb	1	
V	2	
Nb	1	
Ta	1	
Cr	20	
Mo	3.5	
W	2	
Tc	0.5	
Ru	1	
C	200	
Si	150	
P	150	500
F		100
Cl		100
Br	2	
F + Cl + Br + I	250	
U		3000
Pu	0.01	
Am		2500

[a] Ag, B, Ca, Cd, Co, Cr, Dy, Eu, Fe, Gd, Hf, Li, Mn, Mo, Ni, Sm, Ti
[b] Ag, Al, Ba, Be, Bi, Ca, Cd, Cu, Fe, K, Li, Mg, Mn, Na, Ni, Pb, Sn, Sr, Th, Zn, Zr
[c] Ag, Al, Ca, Cd, Co, Cr, Cu, Eu, Fe, Gd, Li, Mg, Mn, Mo, Na, Ni, Pb, Si, Sn, Th, Ti, U, V, W, Zn

5.4.1.3. Mechanical Head-End

The Zircaloy which is used as the cladding material for LWR fuel pins is not attacked by nitric acid. Therefore, some pretreatment is necessary to remove or at least open the pin cladding to expose the fuel to the acid. Although chemical decladding procedures have been developed and applied on a pilot-plant scale (Eurochemic Plant, Mol, Belgium [254]), the pretreatment in normal industrial use is the mechanical shearing of the fuel pins into short segments of ca. 3 – 5 cm length. The shearing can be carried out either on the fuel element as a whole, or on bundles of fuel pins after previous disassembling of the fuel element. Single-pin shearing is only employed in pilot-scale plants of low throughput.

As a typical example, the head-end facilities of the French UP-3 plant at Cap de La Hague are described in more detail [255]. With a nominal capacity of 800 t uranium per year (4 t uranium per day), UP-3 is, at the time of writing (1991), the largest operating LWR fuel reprocessing plant worldwide. It has two independent head-end lines, each housed in a separate hot-cell complex and equipped with a hydraulic fuel element shear, a dissolver, and a feed clarification unit plus auxiliary equipment such as off-gas purification and solid waste packaging. Each line is designed for the full plant capacity, so that only one unit is in operation at any time while the other is on standby. The fuel element to be processed is transferred by a special conveyor truck from one of the storage ponds via an inclined channel into the reception cell. A bridge crane is used to unload the fuel element from the truck and, after thorough inspection, to carry it over into the shear cell. Here it is loaded into a tipping device connected to the shear, rotated into a horizontal position, and pushed into the shear. The fuel element's foot-end is cut away by the horizontally operated shear blade, and is transferred to a solid waste container. Next, the loosened pin bundle is sheared into segments of ca. 3 cm length which fall down into the dissolver. The fuel element's top-end which is left behind is again discharged to solid waste.

A shear machine must be designed with great care in regard to its maintenance and repair, because most of this must be performed by remote techniques. For instance, the remote exchange of spent shear blades is a frequent routine operation. Direct maintenance can be applied for the hydraulic shear drive which in the UP-3 plant is located in a separate, accessible room outside the hot cell.

5.4.1.4. Fuel Dissolution

The next treatment step is the dissolution of the fuel in nitric acid. The main reason for the use of a nitric acid medium in reprocessing–instead of media such as hydrochloric or sulfuric acids–is the fact that stainless steel can then be used as the construction material for equipment, vessels, and piping, which would not be possible

with the other acids because of their corrosive action.

Chemistry. Uranium dioxide is readily dissolved by nitric acid with the formation of uranyl nitrate and a mixture of gaseous nitrogen oxides. The following reactions contribute to the total turnover [256], [257]:

$$UO_2 + 4\,HNO_3 \longrightarrow UO_2(NO_3)_2 + 2\,NO_2 + 2\,H_2O \quad (6)$$

$$3\,UO_2 + 8\,HNO_3 \longrightarrow 3\,UO_2(NO_3)_2 + 2\,NO + 4\,H_2O \quad (7)$$

$$4\,UO_2 + 10\,HNO_3 \longrightarrow 4\,UO_2(NO_3)_2 + N_2O + 5\,H_2O) \quad (8)$$

The extent to which the two main reactions (6) and (7) contribute is determined by the equilibrium:

$$3\,NO_2 + H_2O \rightleftharpoons 2\,HNO_3 + NO \quad \Delta H = -73\,\text{kJ/mol} \quad (9)$$

Consequently the molar NO_2: NO ratio increases with increasing nitrate concentration and temperature. Reaction (8) is a side reaction which is favored by a high nitric acid (but not nitrate) concentration and a high temperature; under favorable conditions (12.5 M HNO_3, 90 °C) it may contribute as much as 16 % to the total turnover. The dissolution rate is characterized by an induction period of various length, indicating an autocatalytic process. The rate increases strongly with the nitrate concentration and with the temperature up to 90 °C, but it decreases if the temperature is further increased, indicating the loss of a catalyst. The rate also decreases with increased intensity of mixing until a lower rate limit is reached, again indicating the removal of a catalyst from the UO_2 surface. The rate law was found to be:

$$dn/dt = k\,q_0\,(1-n)^{2/3} \quad (10)$$

where n is turnover, t is time, k is rate constant, q_0 is the UO_2 surface at $t = 0$; the term $(1-n)^{2/3}$ accounts for the decrease of the surface with progress of dissolution. Uranium (IV) ions could not be detected as an intermediate in solution despite intense search.

The following reaction mechanism was derived from these observations: during the induction period, an autocatalytic species is formed at the UO_2– HNO_3 interface, which could either be HNO_2, the primary reduction product of HNO_3, or a species in equilibrium with it, such as N_2O_3:

$$2\,HNO_2 \rightleftharpoons N_2O_3 + H_2O \quad (11)$$

This assumption is supported by the fact that the induction period is decreased by the addition of $NaNO_2$, but is substantially increased if a nitrous acid scavenger such as azide ion is added (Eq. 12).

$$HN_3 + HNO_2 \longrightarrow N_2 + N_2O + H_2O \quad (12)$$

Moreover there is direct visual evidence for the formation of N_2O_3 at the UO_2 surface. The catalyst oxidizes U^{4+} in the crystal surface to penta- or hexavalent uranium, and is itself reduced to NO. The higher-valent uranium species formed can escape from the very stable fluorite lattice of UO_2 and is able to transfer into the solution, while NO reacts with HNO_3 under reformation of the catalyst (Eq. 13).

$$2\,NO + HNO_3 + H_2O \rightleftharpoons 3\,HNO_2 \quad (13)$$

This mechanism also explains why PuO_2 is practically insoluble in nitric acid, because the oxidation potential of both HNO_3 and HNO_2 is insufficient to oxidize Pu^{4+} which is bound in the fluorite lattice to a higher oxidation state. However, PuO_2 can be dissolved if a stronger oxidant such as Ce^{4+} or Ag^{2+} is added. A variant is the electrolytic dissolution of PuO_2 in the presence of small amounts of Ce^{3+} or Ag^{+} which act as redox catalysts. Plutonium dioxide can also be dissolved in HNO_3 in the presence of a strong complexant for Pu^{4+} such as fluoride ion, but obviously this reaction proceeds by a totally different mechanism [258], [259].

The dissolution behavior of $(U_{1-x}Pu_x)O_2$ solid solutions in HNO_3 is characterized by a transition from easy solubility at plutonium concentrations $x \lesssim 0.35$ to virtual insolubility at $x \gtrsim 0.40$. As a consequence, the dissolution behavior of UO_2–PuO_2 mixed oxide fuels, the composition of which may vary from a few percent of Pu (LWR – MOX) up to ca. 30 % Pu (MOX for small- to medium-size FBRs), is determined by the chemical state of the material. If the fuel is in the state of a homogeneous solid solution, it will dissolve totally in hot nitric acid; if it is inhomogeneous and contains a major proportion of a plutonium-rich $(U, Pu)O_2$ phase, or of pure PuO_2, it will be only partially soluble,

and a major proportion of the plutonium is retained in the insoluble solids. These plutonium residues could only be dissolved by the aforementioned methods (strong oxidants, complexants). However, for irradiated fuels these would be very inconvenient because fluoride is strongly corrosive to stainless steel, and would require the use of a different construction material, whereas strong oxidants have the drawback that the fission product ruthenium is oxidized to volatile RuO_4 which would escape into the off-gas system. Other potential methods such as dissolution in a strongly reducing acid (e.g., HBr) or in a salt melt (e.g., molten Na_2CO_3) also have obvious severe drawbacks. Thus, it is most important that MOX fuel is fabricated in a way to guarantee that the material is highly homogeneous (see Section 5.3.3.6).

So far only the behavior of unirradiated fuel materials has been considered. In the dissolution of irradiated UO_2 or MOX fuels, small amounts of insoluble, highly radioactive residues are always observed. The mass of these residues increases more than proportionally with fuel burnup, but remains below 1 % of the total fuel mass even with very highly irradiated material. The amount of plutonium retained in these insolubles is generally well below 0.1 % of the total plutonium in the fuel, and thus does not present a loss problem. The composition of the insolubles from as-specified fuels, both UO_2 and MOX, is complicated [260], [261]. Apart from coarse objects such as screws resulting from fuel element shearing, the fines contain materials originating both from the fuel itself, and from the cladding. Thus, particles of metallic or partially oxidized Zircaloy (LWR) or stainless steel (SNR) come from the cladding. The residues originating from the fuel itself are mainly composed (1) of a metallic noble metal phase containing the fission products Mo – Tc – Ru – Rh – Pd which is a primary product formed in the fuel (so-called ε-phase); and (2) of a secondary product generated by reprecipitation of dissolved fission products from the fuel solution in the form of an oxidic material containing mainly Mo – Tc – Ru – Rh – Pd – Te – Zr.

Technical Performance. Fuel dissolution is usually performed as a *batch process*. The dissolver is positioned below the fuel element shear and consists of a vessel fabricated from a highly corrosion-resistant stainless steel or nickel-base alloy, into which a basket, made from the same material, is inserted. The sheared fuel pin segments fall down into the basket which is submersed in nitric acid, and the dissolution is accelerated by heating. At the end of the process the basket is lifted out of the solution, the leached Zircaloy cladding segments retained in it are rinsed with dilute acid, and are then transferred to solid waste. In the design of the dissolver, criticality control plays a dominant role (see Section 5.4.1.11) and is a limiting factor for the dissolver capacity. For large plants, several dissolvers are therefore needed which can be operated in a cycle mode, and a semicontinuous form of operation is thus achieved.

In contrast, each of the two head-end lines of the French *UP-3 plant* is equipped with a *continuous rotary dissolver* (Figs. 120 A and B) [262]. It consists of a stainless steel vessel of critically safe dimensions which holds the nitric acid – uranyl nitrate dissolver solution of average composition 200 – 300 g/L uranium, 2.7 – 3.3 mol/L nitric acid). A Zircaloy wheel with perforated walls, which is subdivided into 12 sectorial compartments, is suspended in the cover of the vessel, such that its lower part submerges into the solution. Both the cover, which also carries all mechanic drives, and the wheel itself, can be remotely removed by a crane for repair or replacement. The fuel element shear which is arranged above the dissolver, delivers the sheared pin segments into a wheel compartment submerged in solution. When the compartment is loaded, the wheel is turned by 30 °, and the next sector is charged with fuel. The speed with which the wheel is turned is rated such that complete dissolution of the fuel in the hot dissolver solution is guaranteed. After several 30 ° turns, the leached cladding segments emerge from the solution, and the residual solution entrained in the hulls can drain out. Finally, when the compartment arrives in an upper position the hulls are discharged via a rectangular gorge into a rinsing vessel, from which they are evacuated by a helicoidal ramp while the rinsing solution is recycled.

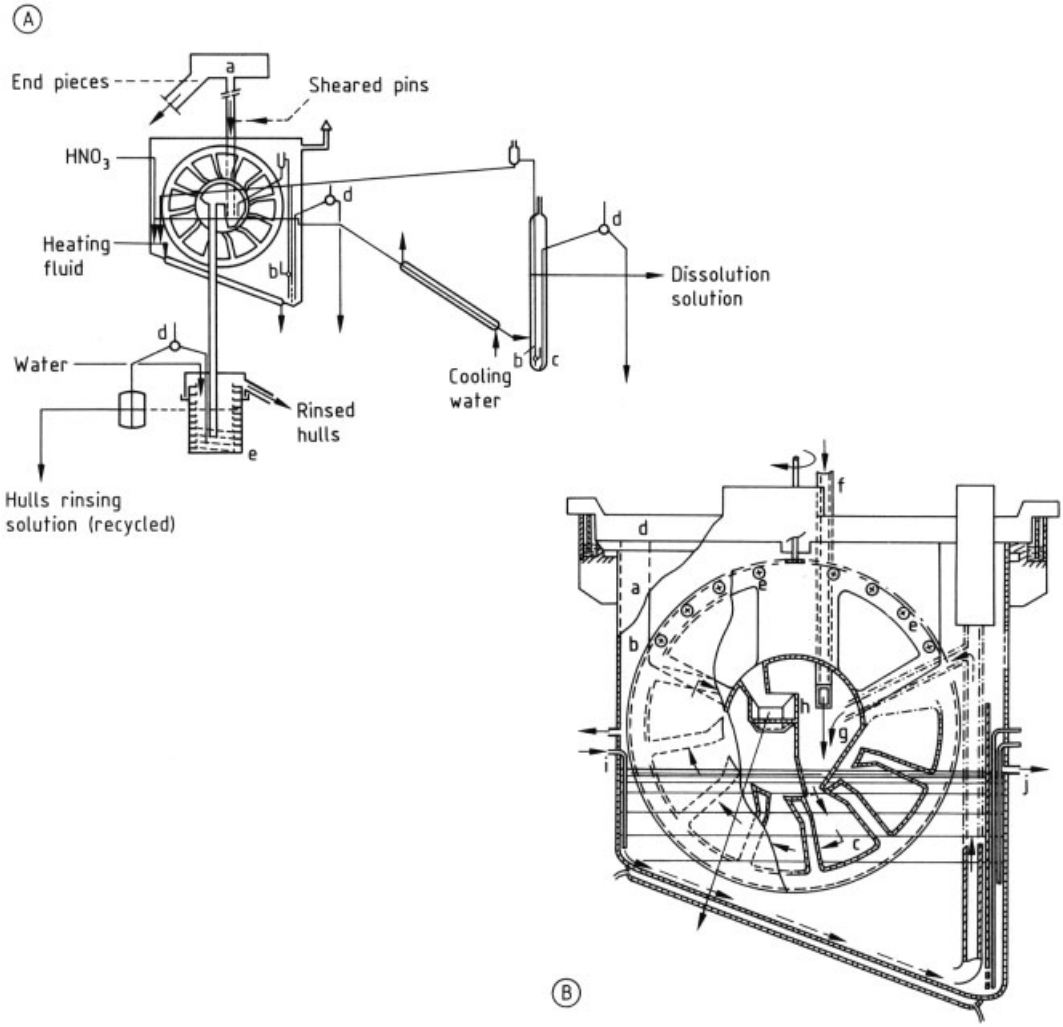

Figure 120. A) Arrangement of shear – dissolver – rinsing vessel at UP-3 (schematic) [262]
a) Shearing machine; b) Air lift; c) Measuring pot; d) Siphon; e) Rinsing vessel
B) Rotary dissolver as used in the UP-3 plant, Cap de la Hague, France [262]
a) Dissolver vessel; b) Rotating wheel; c) Wheel compartment; d) Cover; e) Suspension roller gear; f) Feeder chute; g) Hopper; h) Discharge chute; i) Acid inlet; j) Product outlet

5.4.1.5. Feed Clarification and Make-up

Small quantities of insoluble, highly radioactive fine particles which amount to ca. 5 kg per tonne of fuel remain in the dissolver product solution. These must not pass on into the solvent extraction system because here they would contribute considerably to the radiation decomposition of the organic extractant, and to the consequent formation of interfacial "cruds" which can lead to severe operational problems in solvent extraction equipment. Therefore, a highly efficient clarification process which can retain fines with a size down to ca. 1 µm must be carried out. Both filters and centrifuges have been utilized; in modern plants, the trend is in favor of high-efficiency centrifuges.

Before the feed solution is introduced into the solvent extraction system, its uranium and acid

concentrations must be made up to flow sheet specifications.

5.4.1.6. PUREX Process: Chemistry

The basic extraction chemistry of the elements involved in the PUREX process has been the target of many investigations. Reviews can for instance be found in [250], [263], and [264]. In this chapter, only a very general description is given.

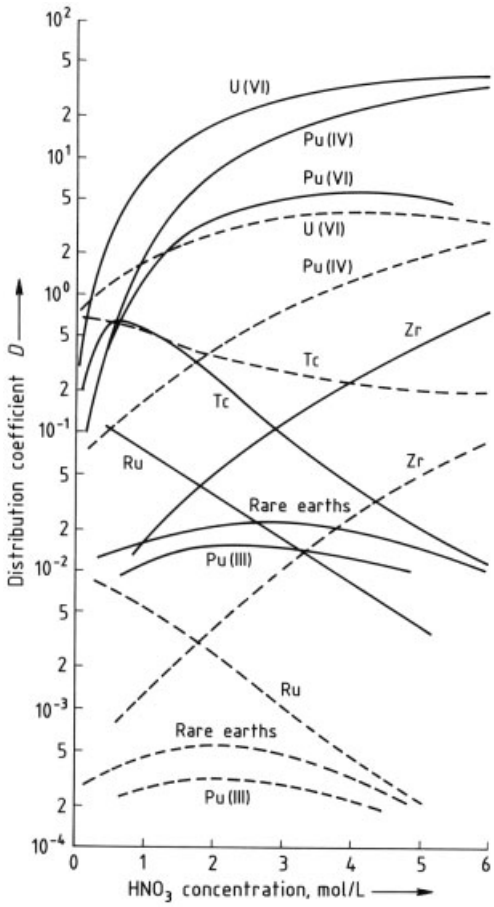

Figure 121. Distribution coefficients for uranium, plutonium, and representative fission products in the system HNO_3 – 30 vol % TBP – kerosene as a function of the HNO_3 concentration
——— Trace metal concentration; – – – – – – – In the presence of 100 g/L uranium in the organic phase

Actinide Extraction Behavior [263], [264]. Tri-*n*-butyl phosphate is a strong and highly selective extractant for tetra- and hexavalent actinides from nitric acid media. This property is based on the formation of stable complexes between actinide nitrate and TBP which are soluble in the organic phase. For hexavalent actinides, and especially for uranyl nitrate, the formation of complexes of the composition $MO_2(NO_3)_2(TBP)_2$ is well established. $UO_2(NO_3)_2(TBP)_2$ has been isolated as a low-melting crystalline solid (*mp* $-6\,°C$), and the molecular structure of the triethyl phosphate (TEP) analogue, $UO_2(NO_3)_2(TEP)_2$, has been determined by X-ray diffraction.

In this compound, the trialkyl phosphate is bound to the central uranium atom via its phosphoryl oxygen atom, and the nitrate ion acts as a bidentate ligand. The identity of the complexes formed by tetravalent actinides is less well established. Based on indirect evidence such as the dependence of distribution coefficients on the extractant concentration, the composition $M(NO_3)_4(TBP)_2$ is most often proposed but direct evidence is lacking. Instead, a thorium model compound $Th(NO_3)_4(TMP)_3$ carrying three instead of the proposed two trimethyl phosphate (TMP) ligands has been isolated, keeping the question of the true composition of the M^{4+} complexes in solution unresolved.

Actinide nitrates in oxidation states other than +4 or +6 form less stable complexes with TBP, and hence are much less efficiently extracted. The same is true for most other metal nitrates, including the fission product metals.

The extent to which a species M is extracted into the organic phase is given by the distribution coefficient (distribution ratio) *D*, defined as

$$D\,(\mathrm{M}) = \frac{[\mathrm{M}]_{\mathrm{org}}}{[\mathrm{M}]_{\mathrm{aq}}} \tag{14}$$

where [ ] denotes equilibrium concentrations, and the indices "org" and "aq" denote the organic and aqueous phases, respectively. The distribution coefficient is a function of many parameters, in particular the composition of the aqueous and organic phases. Under typical PUREX conditions, i.e., 20 – 30 vol % TBP in the organic solvent, the distribution coefficient for U (VI) and Pu (IV), which are the normal oxidation states in nitric acid, is < 1 at low HNO_3 concentrations but increases to > 10 at intermediate acidities

until a maximum is reached at about 6 M HNO_3 (Figure 121). It then decreases slightly with further increase of acidity. The order of extractability at room temperature and low to intermediate acidities is U (VI) > Pu (IV) > Pu (VI); a practical consequence for PUREX is that the plutonium oxidation state must be adjusted to +4 for optimum extraction. The distribution coefficient for trivalent plutonium is ca. three orders of magnitude lower, and is similar to that for the other trivalent actinides (Am, Cm) or lanthanides (Figure 121); this offers the basis for U – Pu separation. Pentavalent actinides such as NpO_2^+ are also extracted to only a small extent.

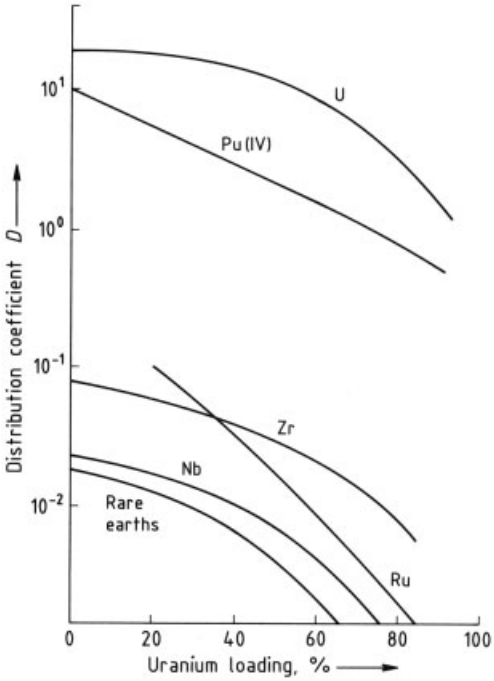

Figure 122. Decrease of distribution coefficients with increased uranium loading of the 30 vol % TBP – kerosene phase.
Aqueous HNO_3 concentration rises from 2 mol/L at 0 % loading to 3 mol/L at 90 % loading [246]

The distribution ratios of the tetra- and hexavalent oxidation states of U and Pu increase with approximately the second power of the TBP concentration in the organic phase if tracer metal concentrations are considered. If the metal concentration is increased to bulk amounts while the TBP concentration is kept constant, a decrease of the distribution ratio is observed because the concentration of "free", i.e., noncomplexed, TBP decreases when the latter is consumed in the formation of metal complexes, see Figure 122, uranium curve. At the same time, and for the same reason, the distribution ratios for foreign metals present in small concentrations will also decrease with increasing "metal loading" of the organic phase, see the plutonium and fission product curves of Figures 121 and 122. If two extractable metals in bulk concentrations are coextracted, they will mutually compete for the extractant, with a mutual decrease of the distribution coefficients. Available data is compiled in [265]. Nitric acid also forms a solvate with TBP, and is therefore extracted into the organic phase to an appreciable extent, where it participates in the mutual competition for the extractant.

A somewhat different presentation of the effect of bulk metal extraction is given in Figure 123 where the organic vs. aqueous equilibrium metal concentrations are plotted. These so-called "equilibrium curves" are a most important tool for the layout of a solvent extraction flow sheet. In the past, such flow sheets were graphically designed by drawing a McCabe – Thiele diagram [267] into the equilibrium curve to determine the number of extraction stages, followed by an iterative refinement, but this procedure suffers from the drawback that only single-component systems can be easily treated with sufficient accuracy. In consequence, computational procedures [268] are currently applied but again the (digitalized) equilibrium curves form the numerical basis for this treatment. An analysis of the uranium curves of Figure 123 shows that at very high aqueous uranium concentrations, the organic phase becomes saturated with metal, and the equilibrium curve levels out. This is to be expected when all TBP has been consumed for complex formation; the upper limit of organic metal loading is termed "maximum loading". In the case of plutonium, however, the theoretical maximum loading cannot be attained; instead, the two-phase equilibrium curve ends before the region of total consumption of TBP is reached. This is due to a phenomenon known as "third phase formation": the organic phase splits into two phases, a heavy one rich in metal and TBP, and a light one rich in diluent, if a certain limit of metal (and acid) loading is exceeded. Third phase formation is quite a common phenomenon in solvent extraction systems;

in particular for metal nitrate – TBP – paraffinic diluent systems, it occurs readily with tetravalent actinides but to a much lesser degree with hexavalent MO_2^{2+} ions. A practical consequence for PUREX is that organic-phase plutonium loadings must be kept well below the third-phase boundary under all conditions of operation because solvent extraction equipment is designed to separate two, not three, liquid phases.

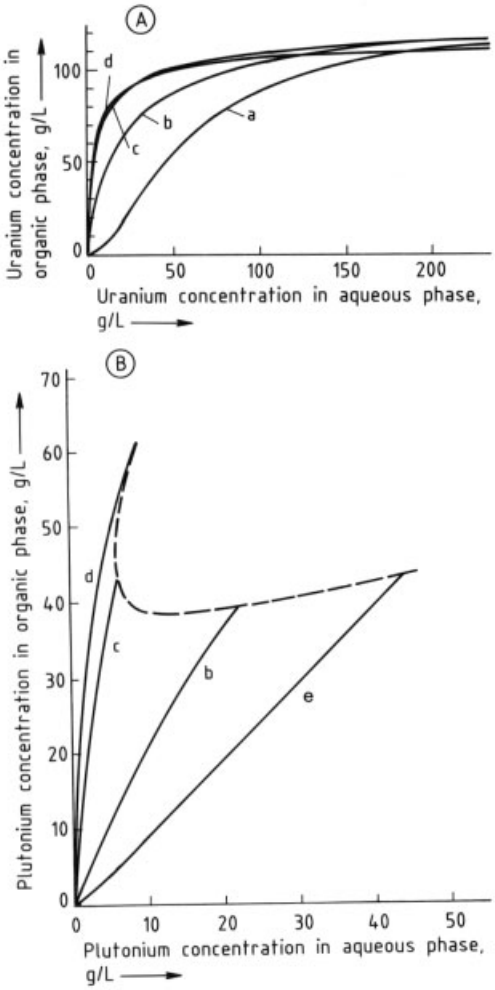

Figure 123. Equilibrium curves for the extraction of uranyl nitrate (A) and plutonium(IV) (B) with 30 vol % TBPdodecane, T = 25 °C
The dashed curve in (B) represents the three-phase limits for plutonium.
a) 0.1 M HNO_3; b) 1 M HNO_3; c) 3 M HNO_3; d) 5 M HNO_3; e) 0.5 M HNO_3

The temperature dependence of the distribution ratios is generally slight; thus, a slight decrease of D for U and Pu is observed with increasing temperature. The tendency to form a third phase decreases with increasing temperature.

Fission Product Extraction Behavior [269]. Particular attention must be paid to those nonvolatile fission products which at the time of reprocessing still contain important amounts of radioactive isotopes (see Table 59). Of these, the alkali and alkaline-earth metals do not form TBP complexes, and are therefore virtually not extracted. The lanthanide(III) nitrates form TBP complexes, similar to their actinide(III) analogues, but these complexes are much weaker than those produced by tetra- or hexavalent actinides. Therefore a high nitrate salt concentration and low acidity combined with a high TBP concentration is needed to extract them efficiently into the organic phase. Under typical PUREX conditions (ca. 3 M HNO_3, 30 % TBP) the trivalent lanthanides and actinides show distribution coefficients nearly three orders of magnitude lower than Pu (IV) or U (VI) (Fig. 121), and are therefore efficiently separated from the latter. The same applies for cadmium and antimony.

More trouble is caused in PUREX by the fission products ruthenium, zirconium – niobium, and technetium, with ruthenium probably the most awkward (for more details, see [269]).

The distribution coefficients of *ruthenium* given in Figures 121 and 122 represent the global extraction behavior of this metal in the extraction section of a first PUREX cycle. Indeed the ruthenium fraction in a fuel solution is not uniform, but consists of a mixture of oxidation states +2, +3, and +4, which all form a variety of nitro and nitrato complexes. Whereas most of these are only slightly extracted by TBP, a few species exist which are highly extractable, notably the neutral or anionic tri- or tetra-nitrato and -nitro complexes of the very stable nitrosylruthenium entity, $RuNO^{3+}$. Since they are preferentially extracted, these species accumulate in the organic TBP phase, and therefore the distribution coefficients during back-extraction operations (scrub and back-wash, see Section 5.4.1.7) are much higher than the "global" values given in Figures 121 and 122. The accumu-

lation of highly extractable species also explains why the ruthenium decontamination achieved in TBP purification cycles is substantially lower than that achieved in the first extraction cycle. Although it is true that these species are in equilibrium with poorly extractable cationic species, the interconversion rates are relatively slow, and equilibrium is not attained during the residence times available in solvent extraction equipment. As can be derived from Figures 121 and 122, a high nitric acid concentration in the aqueous phase and in particular a high uranium loading of the organic solvent help to improve the separation of ruthenium from the product elements. Moreover, the distribution coefficients of most of the ruthenium species decrease with increasing temperature, while the complex interconversion rates increase, with the effect that the formation of poorly extractable complex species is accelerated. Both effects lead to a decrease of the "global" ruthenium distribution ratios with increasing temperature, and therefore it is often recommended to apply a "warm" scrub for improved decontamination from this fission product.

Zirconium-95 (half-life 65 d) and its daughter nuclide *niobium-95* (35 d) are very troublesome in the processing of short-cooled (e.g., FBR) fuel elements, but due to their relatively rapid decay the problems they present diminish greatly with prolonged cooling. The extraction chemistry of zirconium in HNO_3–TBP systems is complicated because Zr^{4+} forms a variety of nitrato – hydroxo – aquo complexes, and nitrato – hydroxo- or oxo-bridged oligomers. The extractable species are believed to be the TBP-disolvates $Zr(NO_3)_4(TBP)_2$ and $ZrOH(NO_3)_3(TBP)_2$ [269]. In contrast to ruthenium, the zirconium distribution coefficient increases with increasing nitric acid concentration (Fig. 121); therefore in a PUREX flow scheme optimized for the decontamination of ruthenium and zirconium, the optimal acidity will be a compromise between the individual requirements of the two [270]. Similar to most other fission products, the extraction of zirconium is suppressed by increased uranium loading of the organic phase (Figures 121 and 122). The factor which dominates the behavior of zirconium in PUREX is however its high extractability by dibutylphosphoric acid, $(C_4H_9O)_2POOH$ or "HDBP," which is formed by radiolytic degradation of TBP. Zirconium is extracted as a bis-dibutylphosphate salt, $ZrX_2(DBP)_2$ (where X^- is NO_3^- or OH^-). This compound has a relatively low solubility both in the organic and in the aqueous phase, and hence it precipitates as a slimy interfacial crud as soon as its solubility limit is exceeded. Secondary radiolysis products of TBP, such as monobutylphosphoric acid, $C_4H_9OPO(OH)_2$ or "H_2MBP," do also contribute to the formation of crud, which has the unpleasant property of absorbing any (radioactive) fine particles present in the solution. Since crud is difficult to remove from extraction equipment, especially from mixer – settlers, it can build up an additional irradiation source for the organic solvent which then accelerates the degradation of TBP and, hence, the formation of more crud, similar to an autocatalytic process. If formed in large amounts, crud can become a serious operational problem by plugging the solvent extraction equipment. This is the reason why fission product zirconium must be closely controlled in PUREX, even though it contains only inactive isotopes after long cooling. Measures to prevent the formation of crud include (1) a highly efficient feed clarification to prevent the transfer of radioactive fine particles from the dissolver into solvent extraction; (2) a high uranium loading of the organic phase and an increase of the aqueous nitric acid concentration to ca. 3 mol/L because it has been found that the formation of Zr – HDBP precipitates diminishes under these conditions; (3) the application of short-residence-time solvent-extraction equipment, i.e., of pulsed columns or centrifugal extractors instead of mixer – settlers, to decrease the irradiation dose to the solvent; and (4) a highly efficient solvent wash after the process, to prevent the build-up and recycle of TBP degradation products with the organic solvent [248].

Technetium, as the long-lived ($t_{1/2} = 2.1 \times 10^5$ a) soft β-emitter ^{99}Tc, the only isotope of practical importance, is not a radiation problem in the PUREX process but instead causes troubles of a different kind. Uranium which is to be reenriched must be highly decontaminated from technetium because it causes operational problems both in the conversion and enrichment processes (see Table 60). Moreover, its control is necessary because due to its long half-life

and high mobility in the environment, it could present a long-term radiation hazard if improperly handled. Finally, if technetium is present in substantial quantities, as is the case with fuels of high burnup, it can cause chemical problems in the solvent extraction process itself. Technetium is stable in nitric acid solution as the heptavalent pertechnetate anion (TcO_4^-). Strong reductants reduce TcO_4^- to Tc^{4+}. Pertechnate alone is moderately extracted by TBP (Figure 121), probably as the trisolvate $HTcO_4(TBP)_3$. The distribution coefficient has its maximum at ca. 0.5 mol/L HNO_3 and decreases with increasing HNO_3 concentration; thus, for typical PUREX extraction conditions (3 mol/L HNO_3, 30 vol % TBP, 30 °C), D(Tc) is ca. 0.1, a value which would not suggest problems in separating it from U and Pu. However, in contrast to other fission products the extractability of TcO_4^- increases substantially if extractable cations such as Zr^{4+}, Pu^{4+}, or UO_2^{2+} are present. The reason is that TcO_4^- can replace NO_3^- in the extracted metal nitrate – TBP complexes, i.e., species of the type $M(NO_3)_3(TcO_4)(TBP)_2$ or $MO_2(NO_3)(TcO_4)(TBP)_2$ are formed. In this way technetium is coextracted with uranium and plutonium in the extraction contactor. In the partitioning contactor a mixture of uranium (IV) nitrate and hydrazonium nitrate is used to reduce Pu (IV) to inextractable Pu (III) (see Section 5.4.1.7) and this reductant also reduces TcO_4^- to inextractable Tc^{4+}, which is thus separated from the uranium and accompanies the plutonium. However, technetium must be prevented from accumulating in the partitioning extractor because Tc^{4+} is easily re-oxidized by nitric acid, and both the nitrous acid and the TcO_4^- formed oxidize the holding reductant hydrazine.

5.4.1.7. PUREX Process: Flow Sheet

The chemical separation process must satisfy two requirements which are contradictory to each other: (1) The recovery of the values, uranium and plutonium, must be very high; a yield of > 99 % is usually demanded. (2) The decontamination of the products from impurities, i.e., the fission products and nonfissile actinides, must be very high; a decontamination factor of $> 10^6$ is usually required. There is no chemical procedure which could fulfill these two requirements in a single step, and this is also true for solvent extraction by TBP (see extraction data in Section 5.4.1.6). The only way to achieve this goal is a multiple repetition of the recovery and purification operations as a multistage, countercurrent process.

Figure 124 presents a basic first-cycle PUREX flow sheet. A "box" in Figure 124 symbolizes a multistage, countercurrent extractor. Auxiliary equipment such as evaporators, tanks, pumps, metering equipment is not shown. The organic solvent stream is symbolized by a dashed line whereas aqueous streams are depicted by solid lines. The designations of the various reagent streams are indicated in ovals. A set of "typical" compositions and flow rates of the various feed and product streams is presented in Table 61. These data do not represent an actual flow sheet as applied in an existing plant, but instead form a sample flow sheet calculated for the treatment of LWR fuel in the basic set of extraction equipment presented in Figure 124. The actual arrangement and sequence of extractors, as well as the compositions and flow rates of reagent streams applied in existing plants, can vary within wide limits [270]. Actual plant flow sheets, as far as they have been published, have been reviewed [271].

The organic solvent stream which cycles through the various contactors, is typically composed of 30 vol % TBP in a paraffinic diluent, but lower extractant concentrations (20 or 25 vol % TBP) are applied in a few plants. Facilities specially designed for the processing of highly enriched uranium fuels (e.g., from research and material test reactors) apply much lower TBP concentrations, typically between 2.5 and 7.5 vol %. The diluent is a paraffinic hydrocarbon mixture of approximately C_{12} average chain length; both straight-chain (e.g., *n*-paraffin mixtures $C_{10} - C_{13}$) and branched-chain formulations (e.g., hydrogenated propylene tetramer, HPT) are applied [270], [272]. These diluents are characterized by relatively high boiling and flash points, and by a low density, making the organic phase the lighter phase of the two-phase extraction system.

The high-activity feed solution has a typical composition of ca. 1 mol/L of heavy metals (i.e., uranium plus plutonium) and 3 mol/L HNO_3. It is metered into the middle of the *first contactor*, designated "A" in Figure 124, and flows

Table 61. Sample flow sheet for a LWR PUREX extraction cycle as outlined in Figure 124 (calculated with KFK-VISCO)

Extractor	Stream	Composition	Relative Flow
A	high-activity feed	238 g/L U 2.4 g/L Pu 7.5 g/L fission products 3 mol/L HNO_3	100
A	solvent	30 vol% TBP/kerosene	267
A	scrub	2 mol/L HNO_3	50
A	high-activity waste	5.3 g/L fission products 2.6 mol/L HNO_3	142
A → B	organic product	85.7 g/L U 0.86 g/L Pu 0.12 mol/L HNO_3 30 vol% TBP/kerosene	278
B	reductant 1	0.2 mol/L $N_2H_5NO_3$ 0.5 mol/L HNO_3	55
B	reductant 2	48 g/L U(IV) 12 g/L U(VI) 0.5 mol/L $N_2H_5NO_3$ 0.5 mol/L HNO_3	20
B	organic scrub	30 vol% TBP/kerosene	63
B	Pu product	3.16 g/L Pu 11 g/L U 0.27 mol/L $N_2H_5NO_3$ 0.8 mol/L HNO_3	76
B → C	organic U product	71.1 g/L U 0.033 mol/L HNO_3 30 vol% TBP/kerosene	340
C	strip	0.05 mol/L HNO_3	300
C	U product	78.6 g/L U 0.084 mol/L HNO_3	308
C → W	solvent	0.001 mol/L HNO_3 30 vol% TBP/kerosene	330
W	wash	0.2 mol/L Na_2CO_3	3.3
W	waste	0.1 mol/L $NaNO_3$ 0.1 mol/L $NaHCO_3$ 0.1 mol/L Na_2CO_3	3.3
W → A/B	solvent	30 vol% TBP/kerosene	330

countercurrently to the organic solvent which is introduced into the lowest extraction stage of the A contactor. At the other end of the contactor an aqueous "scrub" stream is introduced which joins the aqueous feed at the feed stage. The solvent extracts uranium and plutonium in the extraction section, while most of the poorly extractable fission products are rejected. The small fraction of fission products which is co-extracted, is scrubbed out of the organic phase in the scrub section of the A contactor. Flow rates are arranged such that the *organic product stream* which exits from the scrub section is loaded with uranium and plutonium to ca. 70 % of the maximum theoretical loading, i.e., to a heavy-metal concentration of ca. 90 g/L. The *aqueous fission product HAW stream* which exits from the extraction section of the A extractor is sent to an evaporator for concentration, and is then intermediately stored as a liquid HAW concentrate (HAWC) before being solidified into a glass for permanent disposal.

The organic product stream leaving the A contactor is routed into the middle of the *second*, or "*partitioning*", *contactor* (B in Fig. 124), where plutonium is separated from uranium. This is achieved by introducing an *aqueous reductant stream* into the organic exit stage which flows countercurrently to the organic stream. Pu^{4+} is reduced to inextractable Pu^{3+} which transfers into the aqueous phase. The small proportion of uranium which is costripped with Pu^{3+} according to its extraction equilibrium is scrubbed out of the aqueous phase by a small solvent stream (" *organic scrub*"), which is introduced into the aqueous outlet stage of the B contactor.

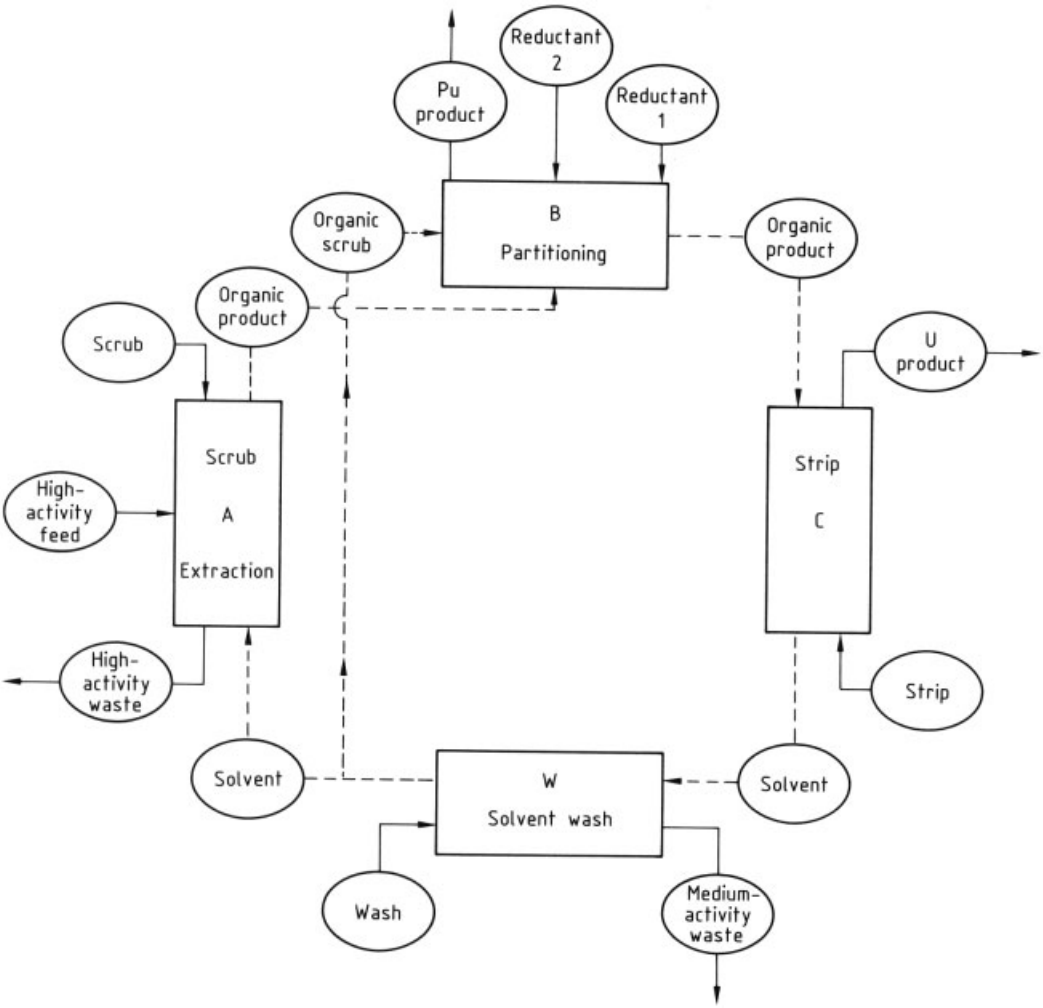

Figure 124. PUREX process flow sheet
For explanations see text.

The reductant originally developed for the PUREX process was iron(II) amidosulfate ("ferrous sulfamate", FeSA) dissolved in dilute nitric acid, where the Fe^{2+} acts as the reductant for Pu^{4+} and the amidosulfate anion serves as a scavanger for nitrous acid (Eq. 15):

$$H_2NSO_3^- + HNO_2 \longrightarrow HSO_4^- + N_2 + H_2O \qquad (15)$$

Nitrous acid which is a strong oxidant for both the Fe^{2+} and the Pu^{3+} is formed by radiolysis of HNO_3 primarily in the A contactor, but it is extracted by TBP and thus transferred into the B contactor. A large molar excess of the reductant, typically 6- to 10-fold, must be used to guarantee a 100 % recovery of plutonium. Since iron(II) is not extracted by TBP, the reductant is introduced into the end stage of the B contactor ("Reductant 1" in Fig. 124). The disadvantages of FeSA are (1) that a large amount of (inactive) iron is introduced into the process which ends up in the medium-active process waste, thus adding to the processing and storage costs of this material; and (2) that the sulfuric acid generated by decomposition of sulfamate (Eq. 15) is highly corrosive to stainless steel equipment.

Currently uranium(IV) nitrate stabilized by hydrazonium nitrate as the nitrous acid scav-

anger (Eq. 16) is employed as reductant in most LWR reprocessing plants. Uranium is a process-inherent material, and hydrazine is easily oxidized to gaseous products, thus both components do not increase the amount of waste generated, and they are not corrosive.

$$N_2H_5^+ + 2\,HNO_2 \longrightarrow N_2 + N_2O + 3\,H_2O + H^+ \qquad (16)$$

Due to the fact that uranium(IV) is extracted by TBP, the reductant cannot be introduced into the end-stage of the extractor; instead, a "split" reductant stream is employed, i.e., the $U(NO_3)_4$ is introduced near the organic feed-stage ("Reductant 2" in Figure 124) while a stream of the stabilizer solution is introduced into the end-stage ("Reductant 1" in Figure 124). As in the case of ferrous sulfamate, a large stoichiometric excess of the reductant is needed for 100 % recovery of plutonium.

The uranium(IV) reductant solution is normally prepared by electrochemical reduction of an aqueous uranyl nitrate – hydrazonium nitrate solution. An elegant alternative is the direct in-line electroreduction of plutonium in the partitioning extractor itself. In this process which was developed by Karlsruhe's Institut für Heisse Chemie [273], the extractor is equipped with cathodes made from an electrochemically inert material, e.g., titanium. For the anodes platinum-plated titanium or tantalum is used. There is no need to separate the electrode compartments by a diaphragm because the anodic reoxidation of U^{4+} is strongly inhibited. Both electrochemical mixer – settlers and pulsed columns have been designed and operated [274] (see Section 5.4.1.9). The aqueous reductant stream is simply a dilute hydrazonium nitrate – nitric acid stabilizer solution which is introduced at the "Reductant 1" position of Figure 124. Current densities are controlled such that all Pu^{4+} is reduced to Pu^{3+}, and an additional small excess of U^{4+} is generated which acts as the holding reductant.

Other reductants, and separation methods based on complexation instead of reduction of Pu^{4+}, have been investigated but have not so far found application in a commercial plant. For a review of the chemistry and performance of U – Pu partitioning, see [275].

The aqueous plutonium product stream leaving the B contactor is routed to a *reoxidation unit* (not shown in Fig. 124) where Pu^{3+} is oxidized to Pu^{4+}, while the $N_2H_5^+$ is at the same time oxidized to N_2 and N_2O. Both chemical oxidation by $NaNO_2$, gaseous NO_2, or liquid N_2O_4, and electrochemical reoxidation by the so-called ROXI process [273], [274] are used. The plutonium is then further purified by a PUREX purification cycle or some other process (see below). The organic uranium product stream coming from the B contactor is routed into the *back-extraction* or *"strip" contactor* (C in Fig. 124) where the uranium is back-extracted into a stream of slightly acidified water. The C contactor is usually operated at ca. 50 °C because the U distribution coefficient decreases with increasing temperature. The aqueous uranium product stream is then sent to final purification.

The stripped solvent leaving the C contactor must be separated from degradation products formed by radiolytic and chemical attack before it is recycled. The accumulation of degradation products would lead to a rapid deterioration of solvent performance, characterized by a decrease of product decontamination, increased retention of actinides and fission products with subsequent product losses to waste streams, poor phase disengagement, and finally, the formation of radiolytic cruds [272], [276].

Dibutyl- and monobutylphosphoric acid can be removed from the solvent by an alkaline wash which converts them into water-soluble alkali salts. The *wash contactor* (W in Fig. 124) is normally designed as a mixer – settler with internal aqueous recycle which permits operation at a high organic to aqueous flow ratio. Sodium carbonate is generally used as the alkaline scrubbing agent because carbonate acts as a complexant for uranium and plutonium and thus prevents the precipitation of retained heavy metal traces. Unfortunately most of the soda is consumed by the residual nitric acid carried by the solvent, and the sodium nitrate so formed adds a serious burden of inactive solid material to the aqueous medium-active waste stream leaving the wash contactor. Therefore "hydrazine carbonate", i.e., a hydrazine solution saturated with carbon dioxide up to a molar ratio of approximately 2 : 1, was later proposed as the washing agent because the $N_2H_5^+$ ion can be easily oxidized to gaseous nitrogen in a ROXI-type electro-oxidation cell [277].

Although di- and monobutylphosphoric acids are the main products of solvent degradation, a

large variety of minor products are formed by radiation-induced reactions of both the TBP and the diluent [272], [276]. Some of these also form complexes with fission product or actinide metals, but cannot effectively be washed out of the solvent by aqueous alkaline solutions because they contain long- chain aliphatic groups, and their alkali salts are therefore not water-soluble [276]. In consequence they accumulate slowly in recycled solvent, and are a major factor limiting its lifetime.

The sequence of operations as presented in Figure 124 has been implemented in this basic form in only a few existing plants while modifications have been applied in most others. In many cases the extraction – scrub contactor (A) is followed by an additional *scrub contactor* (often called "S") in which the metal-loaded solvent undergoes a second fission product scrub before entering into the partitioning system. The scrub section of the A contactor is operated at increased acidity for crud prevention and improved removal of ruthenium while the S contactor is operated at lower acidity, which improves the removal of zirconium and also minimizes the amount of nitric acid which is carried by the solvent into the partitioning system. Decontamination of ruthenium can be further improved if the S contactor is operated at increased temperature. The aqueous scrub stream leaving the S contactor is routed into the scrub section of the A contactor where it joins the A-scrub stream introduced there.

The partitioning contractor B which comprises a reduction section and an organic-scrub section, is often split into two separate contactors. The main reason is that the flow rate of the organic phase is quite different in the two sections, so that two tailor-made contactors can be expected to give better performance than a single unit.

In a few plants, uranium – plutonium partitioning is not performed at all in the first extraction cycle but instead is displaced into the second extraction cycle. In such a case the high-activity cycle is designed as a "codecontamination" process comprising the joint extraction, scrub, and back extraction of plutonium and uranium.

5.4.1.8. PUREX Process: Product Purification

The fission product decontamination factors achieved in the first PUREX cycle are normally of the order of several 10^3. Therefore, uranium and plutonium must undergo additional purification operations in order to meet specifications.

Uranium. Uranium is generally purified by sending it through one or two additional PUREX extraction cycles. However, since most of the plutonium has already been separated from the uranium (except in plants where second-cycle partitioning is performed), and because the fission product radioactivity level is much lower, the uranium purification cycles can be somewhat simplified as compared with the high-activity cycle. Moreover, the purification operations can be tailored for the specific removal of obstinate contaminants which have not been satisfactorily separated in the first cycle, e.g., neptunium [266], [278], [279].

A uranium purification cycle is usually designed as a simple extraction – scrub – backwash cycle comprising two contactors termed 2D (extraction – scrub) and 2E (backwash). The 2D contactor can be operated with either a "concentrated" or a "dilute" flow sheet. In the first case, the first- cycle product solution is concentrated in an evaporator to a uranium concentration of ≥ 1 mol/L before it is routed into the 2D contactor. In the second case, the first- cycle product solution is acidified and then routed directly into this contactor. Flow sheet conditions are in both cases arranged so that a high solvent loading is attained. A small amount of a reductant such as ferrous sulfamate or uranium(IV) nitrate – hydrazonium nitrate is added to the 2D contactor feed to reduce traces of Pu^{4+} to inextractable Pu^{3+}. Neptunium, if present, is rapidly reduced to inextractable Np (V) but only slowly to Np (V) which is moderately extractable. Thus, neptunium decontamination is in most cases only moderate, but tends to be higher with a "dilute" flow sheet.

The second-cycle uranium product is often clean enough to meet product specifications. It is then usually routed through a *fixed-bed column of silica gel* for clarification, and in addition this treatment removes any residual zirconium – niobium by sorption. Other plants,

in particular those treating high-burnup LWR fuels, prefer to incorporate a *third PUREX cycle* for final uranium purification. In this case, residual neptunium can be specifically removed by adding enough nitrous acid (either in the form of $NaNO_2$ or of liquid N_2O_4) into the feed solution so that inextractable Np (V) is stabilized.

The final uranium product solution generated in the PUREX process is concentrated by evaporation to a uranium concentration of ca. 400 g/L. Its further treatment is dependent on the destination of the material – direct recycle into MOX fuel fabrication, or conversion to UF_6 for reenrichment (see Section 5.2.5).

Plutonium. Various methods such as anion exchange or solvent extraction by a long- chain tertiary amine nitrate solution have been used for the final purification of plutonium, but all modern plants have reverted to one or two additional cycles of TBP extraction [251], [270], [271]. After reoxidation to Pu^{4+} and reacidification, the first- cycle plutonium product is routed into the *second-cycle A contactor* for extraction and scrub. Any uranium present is coextracted into the organic solvent. As discussed in Section 5.4.1.6 the flow sheet must be designed so that the plutonium loading of the organic phase remains well below the three-phase boundary under all conditions of operation. Due to the moderate loading of the organic phase which can thus be tolerated, fission products are extracted much better than under high-loading conditions, and consequently the decontamination which can be attained in a plutonium purification cycle is lower than that in a uranium cycle.

The back extraction of plutonium cannot easily be performed with warm dilute nitric acid because plutonium(IV) tends to hydrolyze irreversibly with the formation of polymeric species under these conditions [280]. Therefore, the organic plutonium product stream leaving the 2A contactor is routed to a *B-type contactor* where plutonium is reduced to Pu^{3+} and back-extracted. Reductants used are ferrous sulfamate, uranium(IV) nitrate – hydrazonium nitrate, or as a novel development, hydroxylammonium nitrate stabilized by hydrazonium nitrate [281], [282]. The latter reductant must be operated at increased temperature to overcome the relatively low reaction rate of hydroxylamine with Pu^{4+}. The most advanced technique is, however, the direct electrochemical reduction of plutonium in the 2B contactor itself [273], [274] which has been applied in the WAK pilot reprocessing plant at Karlsruhe, Germany.

The way in which the residual uranium present in the 2A contactor product stream is treated, depends on the specifications determined for the plutonium. Material which is recycled into *power reactor MOX fuel* must not be purified from uranium to a high degree. In this case, the organic scrub section is omitted from the 2B contactor, and the flow sheet is designed so that most of the uranium is costripped with the plutonium. If, however, specifications demand a uranium-free product, the flow sheet must be designed so that the uranium leaves the 2B contactor with the solvent stream, which is most easily achieved by incorporating an organic scrub section. The uranium-bearing solvent is then either stripped in a 2C contactor, analogous to the first-cycle flowsheet presented in Figure 124, or it is routed back into the first-cycle C contactor where it is costripped with the uranium main product. Alternatively, if the uranium content is low and must not be recovered, the solvent can be routed directly into the soda wash system where the uranium is stripped as a carbonate complex.

A third plutonium purification cycle, if present in a plant, is more or less a duplication of the second cycle. The final plutonium product solution is evaporated down to a concentration of ca. 200 g/L, and is then transferred to a MOX fuel refabrication plant either directly as a concentrated $Pu(NO_3)_4$ solution, or as solid PuO_2 after precipitation as the oxalate and subsequent calcination.

5.4.1.9. PUREX Process: Extraction Equipment

For a review of extraction equipment for various applications see Liquid – Liquid Extraction, Chap. 3. Equipment to be used in a reprocessing plant must be simple and robust so that a minimum of maintenance is required. So far, pulsed sieve-plate columns, box-type multistage mixer – settlers, and centrifugal contactors have been applied [283].

Mixer – settlers have given satisfactory performance in plants where low-burnup fuels are treated, but difficulties were encountered in the

processing of high-burnup power reactor fuels. Because of the long residence time in a mixer – settler the organic solvent is exposed to a high radiation dose particularly in first-cycle extraction, which promotes the formation of solvent radiolysis products and, consequently, of interfacial crud. Crud tends to accumulate preferably at the quiescent aqueous – organic interface in the settler chambers, which can ultimately cause hydraulic failures due to clogging of ports and weirs. Although this effect can be counteracted by proper feed clarification and by chemical flow sheet optimization, mixer – settlers are no longer used as the first-cycle A contactor in modern LWR fuel reprocessing plants; they can, however, be used in downstream sections of the process.

Pulsed sieve-plates columns offer the advantage of a lower residence time, in particular of the disperse phase, which makes crud formation less probable. If formed, crud has less possibilities of accumulation because there is only a single quiescent interface, and it can be more easily flushed out. In addition, columns are of a simple and robust design, and they can more easily be adapted to criticality-safety conditions (see Section 5.4.1.11). These properties make the pulsed sieve-plate column a first choice both for the high-activity cycle and for plutonium purification. A certain disadvantage is the large height of a column, and the consequent hot-cell height requirement. The detail design of pulsed columns for PUREX plants is still mostly based on experience; "typical" dimensions are: plate spacing, ca. 50 mm; free plate area, ca. 20 – 30 %, pulse amplitude, ca. 20 – 25 mm, pulse frequency, ca. 40 – 90 min^{-1}. Sieve plates have been designed both with simple holes, or with nozzles, and good results have been reported for a "crown-and-disk," or "disk-and-doughnut," arrangement of unperforated plates [284].

Two designs of *centrifugal contactors* have so far been applied in a reprocessing plant. In the Savannah River 221 F Plant in the United States, the first- cycle A mixer – settler (24 stages) was replaced in 1966 by a bank of 18 single-stage centrifugal contactors developed on-site, which have since given very satisfactory service [281]. Operating characteristics for the centrifugal contactors are: rotational speed, 1745 min^{-1}; centrifugal separation force, 300 – 500 g; dynamic liquid volume, 15 L per stage; total liquid flow, 38 – 250 L/min. Unfortunately, no experience with high-burnup fuel is available so far for this attractive contactor type. For further informations and illustrations, see [281], [283].

The French UP-2 plant at Cap de la Hague, originally designed for the processing of low-burnup gas – graphite reactor fuel, was modified in 1976 for the reprocessing of high-burnup LWR fuel [285]. Among the modifications made in the solvent extraction section was the installation of a *ten-stage Robatel centrifugal extractor* (→ 6. Liquid – Liquid Extraction, Chap. 3.1.5.), [283] in front of the original A mixer – settler, to prevent the main fraction of the fission products from entering into the mixer – settler line. A third uranium purification cycle, equipped with similar centrifugal contactors, was also installed. Although details have not been published, there are indications that clogging of the extractor outlet by accumulated solids has sometimes been a problem [285], which would be in line with the results of inactive tests performed in the United States and Germany [283]. Also the fact that in the new UP-3 plant centrifugal contactors have not been incorporated might be taken as an indication that the relatively complicated multistage design was not fully satisfactory.

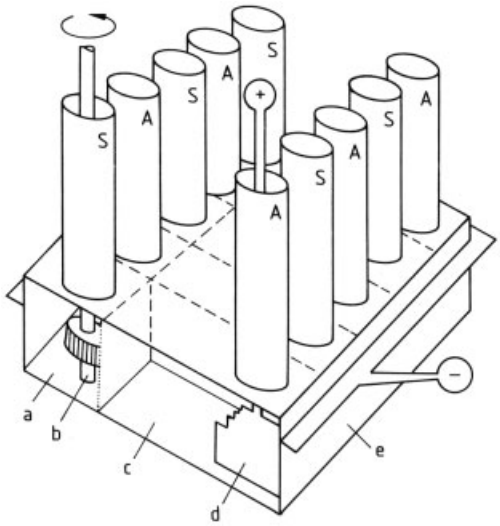

Figure 125. Electro-mixer – settler as operated in the WAK pilot reprocessing plant, Karlsruhe (schematic) [274]
A = Anode duct, S = Stirrer duct
a) Mixer chamber; b) Stirrer; c) Settler chamber; d) Anode; e) Cathode

For the realization of electrochemical PUREX techniques, both mixer – settlers and pulsed sieve-plate columns have been developed [274]. The *electro-mixer – settler* (Figure 125) is constructed from a corrosion-resistant material such as titanium, and the extractor body which acts as the cathode is grounded. The anodes, made from platinized tantalum or titanium, are installed in the settler chambers where the electro-reduction takes place. The anode area is made much smaller than the cathode area to keep anodic re-oxidation of Pu^{3+} low. In 1980 a 12-stage unit of this design was installed in the plutonium purification cycle of the German WAK pilot reprocessing plant at Karlsruhe as the back-extraction (2 B) extractor, and in 1984 a similar 9-stage unit was installed as the first- cycle partitioning (1 B) contactor to replace the former uranium(IV) nitrate partitioning process. Both units have performed very satisfactorily up to the end of WAK's operation in 1990.

The *electro-pulsed column* (Fig. 126) enables a particularly elegant design by using both the column tube and the sieve plates as the cathodes, thus accounting for ample cathode area. The central rod, spaced from the sieve plates by ceramic insulator rings, is made the anode. The same construction materials as for the mixer – settler can be used. Obviously, such a column must be operated in the aqueous- continuous mode to provide electric conductivity. A semi-industrial-scale unit was operated both with first- cycle partitioning and with plutonium purification flowsheets in the Plutonium Test Facility (PUTE) at Institut für Heisse Chemie, Karlsruhe, with a performance even superior to that of the mixer – settler.

An *electrochemical re-oxidation (ROXI) cell* used for the re-oxidation of plutonium(III) and/or hydrazine solutions, is shown in Figure 127. In this very compact design, the anode area must be made large in comparison to the cathode to provide for efficient oxidation. Again, the body of the apparatus is made the cathode which is connected to ground. The anodes are arranged as a compact package of platinized metal grids which are placed into the electrolyte channel. The latter is meander-shaped to suppress axial mixing, and to provide a cross flow of liquid and anode gas resulting from hydrazine oxidation. ROXI cells have been very successfully operated both in the WAK pilot plant and in the PUTE facility.

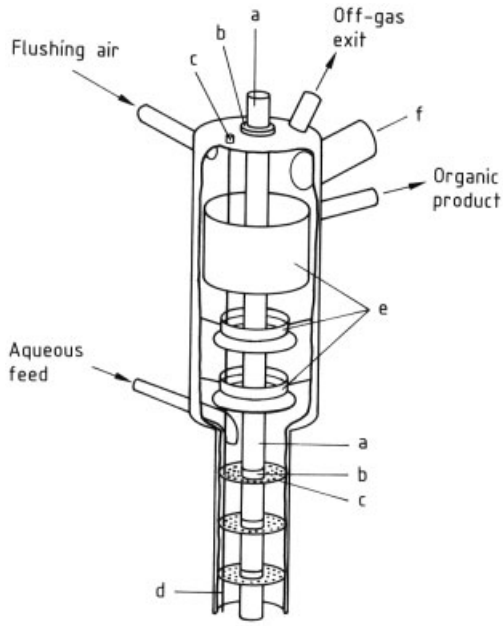

Figure 126. Electro-pulsed column as operated in the semi-industrial plutonium test facility, Karlsruhe [274]
a) Anode; b) Insulator; c) Cathode; d) Electric supply; e) Phase separator; f) Emergency overflow

5.4.1.10. Off-Gas Purification

The retention of radioactive and chemical pollutants from the gaseous effluents of a fuel reprocessing plant is of similar importance as the treatment of liquid and solid radioactive wastes. However, while the latter can be intermediately stored in buffer stores and then treated in more or less independent facilities, the process and cell-ventilation off-gases must be treated directly and unbuffered. Consequently, the off-gas purification procedures must be integrated into the main processes, which means that a particularly high degree of reliability and availability is required for them.

Pollutants carried by the off-gas streams originate from three sources:

1) *Gaseous or highly volatile fission product elements or compounds* are released from the fuel during shearing and dissolution. The element of main concern is *iodine*, with the radioactive isotopes ^{129}I ($t = 1.7 \times 10^7$ a) and

^{131}I ($t = 8.05$ d); the latter isotope is only important if fuel is treated after cooling times of less than 6 months. Iodine and its compounds can distribute between gaseous, aqueous, and organic-liquid phases, and so the element tends to "smear" over most gaseous and liquid streams of a plant if no countermeasure is taken. In consequence, procedures were developed to force the iodine efficiently into the dissolver off-gas stream, and so to prevent it from penetrating into downstream processes. *Krypton*, with the radioactive isotope ^{85}Kr ($t = 10.8$ a), is 100 % released into the head-end off-gas stream. Although ^{85}Kr is one of the major radiation sources in medium-cooled fuel, it is not routinely retained in present reprocessing plants because of its low contribution to the total human radiation burden. The ternary fission product *tritium* ($t = 12.3$ a) is partially released during dissolution as tritiated water, HTO; another fraction remains firmly bound, as zirconium hydride, in the Zircaloy cladding of LWR fuel pins. Part of the HTO enters the off-gas streams as water vapor but due to its low concentration, no special retention measures must be taken. Moreover, the release of tritium with the gaseous effluents can be minimized if tritiated water is confined to the high-activity portion of a plant. To this end, tritiated water that enters the PUREX first-cycle A contactor with the high-activity feed and distributes into the organic solvent which is scrubbed from the latter by using tritium-free acid as the scrub solution. The HTO is thus forced into the HAW stream, and the tritiated acid recovered during HAW concentration is re-used exclusively as dissolver acid, but is not applied for the make-up of downstream reagents. Finally, ^{14}C ($t = 5730$ a) which is produced in LWR fuel in very minor amounts both by ternary fission and by neutron activation of nitrogen and oxygen, is released as $^{14}CO_2$ during fuel dissolution but again no special measures must be taken for its retention.

2) *Radioactive dusts* arise in the fuel element shearing process, and *radioactive aerosols* are produced from liquids by operations such as reflux boiling, evaporation, air agitation, and transfer by air-lift or steamjet.

3) *Chemical pollutants* of concern are the nitrous oxides which are predominantly produced during fuel dissolution.

Head-End Off-Gas Stream. By far the highest radioactivity level is present in the relatively small head-end off-gas stream produced in the shear – leach process. It contains most of the total inventory of volatile fission products, and the aerosols carried by it contain high concentrations of $\beta-\gamma$-emitting fission product nuclides in addition to the α-emitting actinides which form the main radioactivity in downstream off-gases. Consequently it is separately treated in a series of high-efficiency purification steps which are schematically presented in Figure 128.

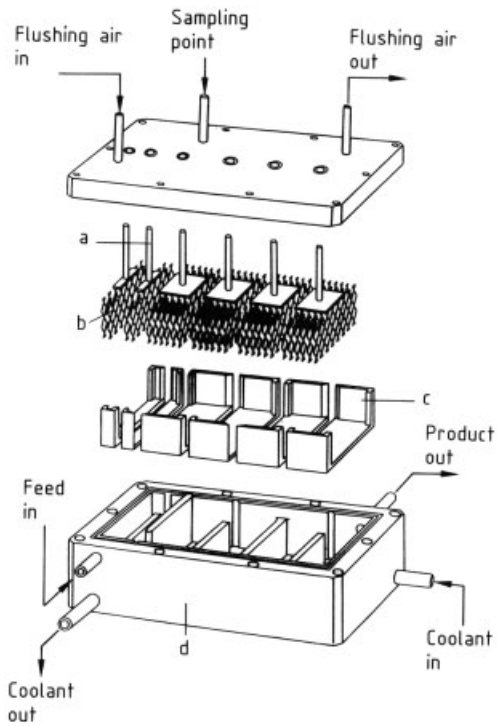

Figure 127. ROXI reoxidation cell as operated in the WAK pilot processing plant, Karlsruhe [274]
a) Electric feed; b) Anode; c) Insulator; d) Cathode

The gas stream is first routed through a washing tower where nitrous oxides, acid vapors, and aerosols are scrubbed by dilute nitric acid at near room temperature. The spent scrub acid is recycled into the main process to avoid an extra waste stream. The gaseous effluent is passed through a sequence of regenerable glass-fiber filters followed by non-regenerable high-efficiency par-

ticulate absorber (HEPA) filters for the precipitation of residual radioactive aerosols. The regenerable filters are regenerated by periodic flushing with nitric acid, while the HEPA filters must be periodically replaced.

Iodine [286] is retained by high-efficiency solid sorption processes based on the formation of silver iodide. The sorbent used is either a silver nitrate-impregnated silica gel or a silver zeolite of high specific surface area. The gaseous effluent from the filter line is preheated and passed through a fixed bed of the sorbent which is operated at ca. 150 °C. A residence time of < 1 s is sufficient for > 99.9 % retention of iodine, including organic iodine species which are formed by reaction of iodine with shear machine grease and with organic impurities in the dissolver acid.

As mentioned before, radioactive *krypton* is not routinely retained in present plants, but in view of potential future requirements, low-temperature separation processes have been developed to the demonstration stage [287]. The most promising procedures are kryogenic rectification as demonstrated at the Tokai Mura Reprocessing Plant in Japan, and low-temperature absorption in coolant R-12 (dichlorodifluoromethane) as developed at Karlsruhe, Germany [288]. In both cases, the inactive fission noble gas xenon is automatically co-separated, and could be recovered as a potentially valuable material. Radioactive ^{85}Kr has also special technical applications.

Downstream Process and Cell-Ventilation Off-Gases. The much larger volumes of downstream process and cell-ventilation off-gases are much less radioactive than the head-end off-gas stream, and can be adequately purified by simpler procedures. They are normally processed through an acid scrubber and on through a filter line for aerosol retention. The need for routine retention of iodine can be avoided if this fission product is prevented from entering into downstream solutions by proper treatment in the dissolution and feed pre-treatment stages. A particularly efficient procedure [289] eliminates iodine traces from the clarified dissolver product by heating it close to the boiling point, adding inactive potassium iodate for isotopic dilution, blowing nitrous oxide gases through the solution to reduce nonvolatile species such as iodic acid to volatile elemental iodine, and combining the off-gas produced in this treatment with the dissolver off-gas stream. The downstream off-gas purification system must then only provide for emergency iodine retention to counteract any failure of the head-end off-gas system.

5.4.1.11. Nuclear Safety

In wide parts of a reprocessing plant, radiation protection requirements are determined by the highly penetrating gamma radiation. Alpha and beta particle radiation is much less penetrating and, hence, much easier to shield, while the neutron radiation, although highly penetrating as well, accounts in most instances for only a minor fraction of the total radiation dose rate. To provide protection against gamma radiation, the respective process equipment is enclosed in a highly shielded hot cell, with the consequence that this equipment must be remotely operated [290], [291].

The shielding material most widely used for industrial hot cells is reinforced heavy concrete. Lead or steel shielding is applied for smaller units, such as analytical cells, and for special hot-cell accessories such as wall penetration plugs used for the conveyance of process media, electric ducts, etc. Viewing is provided by shielding windows made from special high-density lead glass. In-cell operations are performed by remotely operated tools such as master – slave manipulators or crane-supported heavy-duty tools. To prevent the penetration of radioactivity into or through the cell walls, these are often sealed with a dense stainless steel liner. Cell ventilation is provided to remove radioactive gases and dusts, and chemically corrosive or explosive gases and vapors, and a staggered sequence of slight underpressure is maintained inside the hot cells and the surrounding buildings so that no uncontrolled air flow out of the radioactive zones into control areas or into the environment can occur.

Remote operation of the process means that all equipment used must be extraordinarily reliable, must require a minimum of maintenance, and must be controlled remotely. One of the many consequences is, for instance, that apparatus equipped with mechanically moving parts is avoided wherever possible, e.g., air-lifts or steam jets are preferred against mechanical pumps for liquids transfer. If mechanical components can-

not be avoided, these are arranged outside the hot-cell whenever possible.

For *maintenance and repair of hot-cell equipment*, two principal concepts known as "direct" and "remote" are available. *Direct maintenance* means that the repair work is done hands-on by personnel entering into the hot-cell. This is only possible after remote decontamination of the equipment and cell structure from its radioactive inventory so that the radiation dose rate inside the hot-cell is reduced to a level permitting access by man. Cell decontamination is tedious and time-consuming work which normally means that the process area concerned is shut down for many months, if not years. In order to limit the expenditure of manpower and down-time, a hot-cell plant designed for direct maintenance is therefore sub-divided into many single cell units sealed and shielded against each other, each of which houses a selected portion of process equipment. "Critical" equipment is often duplicated to gain redundancy. The equipment must be positioned in a way to enable easy access, and remotely operated tools are provided for auxiliary work and for decontamination.

In a plant designed for *remote maintenance*, defective equipment is replaced exclusively by means of remotely operated tools. Examples of remote-maintenance facilities are the Savannah River PUREX Plant at Aiken, South Carolina, and the Hanford PUREX Plant at Hanford, Washington [292–294]. Radioactive equipment is mounted in one or two large, concreteshielded, window-less cell buildings called "canyons". Connections from vessels to the permanently embedded piping are made via remotely operated connectors or "jumpers". The plant is serviced by a bridge crane travelling along the canyon which is operated from a highly shielded cab. The operator views the canyon equipment through a remote optic system, and the jumpers are disconnected or connected by tools carried by the crane. This technique requires that very close tolerances are maintained on the location of wall nozzles, vessel supports, positioning guides, and other cell appurtenances, and the same precision must be attained in the fabrication of the process equipment. Failed equipment is transported by crane to a remotely operated, shielded maintenance facility where it is first decontaminated, and then either repaired for reuse, or dismantled for waste disposal.

Many improvements of the original concept have been made. An important development was the arrangement of process equipment in standardized modules or "racks," which can be replaced by crane as a whole, as it was first applied in the Plutonium Reclamation Facility at Hanford. Another step forward was the development of remotely operated television equipment which allows the remote control of in-cell cranes and similar devices, and thus eliminates the need of a crane operator present inside the high-level area.

In modern commercial plants, a *combination of remote and direct maintenance* is often applied. The main advantage of remote maintenance in a single large cell is that failed equipment can be readily and rapidly replaced, keeping plant downtime and associated profit losses low. Its disadvantage is its distinctly higher investment cost. Thus for a cost-optimized plant, remote maintenance is restricted to high-level areas where equipment requiring frequent maintenance (e.g., fuel element shear), or known to have a limited lifetime (e.g., dissolvers, evaporators, filters, pumps, etc.), is located, while direct maintenance is chosen for long-life accessories such as pulse columns, and for low-level process areas. A fine example of a modern plant designed along these lines is the UP-3 plant at Cap de la Hague, France [295].

Criticality Control [296], [297]. As outlined in Sections 1.3.8 and 1.3.9, criticality is characterized by a self-sustained fission chain reaction. It is accompanied by the release of a large amount of energy, and by the emission of penetrating radiation, especially neutron and gamma radiation. In a plant where fissile material is handled, protection against the occurrence of a criticality excursion must therefore be provided. Generally speaking, such a plant must be designed and operated such that under all circumstances of normal and abnormal conditions, criticality can never occur. Basically the following principles can be used for criticality prevention:

1) *Mass Control.* The total inventory of fissile materials handled in a closed process, or process section, is limited to an amount smaller than the minimum critical mass of that material, e.g., the minimum critical mass for (hy-

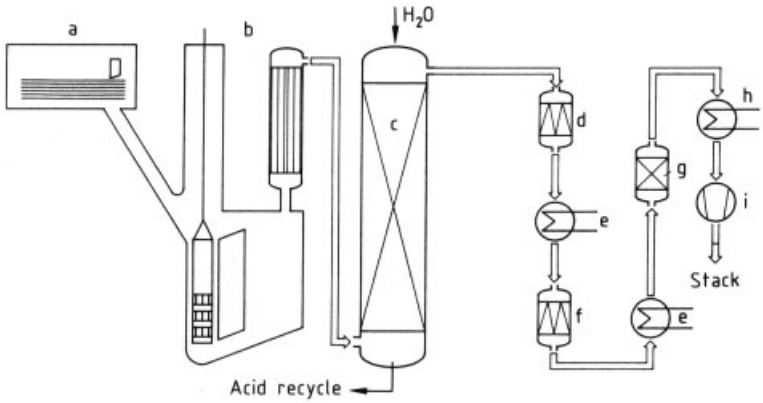

Figure 128. Scheme of a typical head-end off-gas purification system
a) Fuel element shear; b) Dissolver; c) Scrubber; d) Regenerable aerosol filter; e) Heat exchanger; f) HEPA filter; g) Iodine filter; h) Cooler; i) Blower

pothetical) homogeneous $^{239}Pu-H_2O$ mixtures is 531 g plutonium.

2) *Concentration Control.* The fissile materials concentration is limited so that the minimum critical concentration cannot be attained or exceeded under both normal and abnormal operation conditions; e.g., the minimum critical concentration for homogeneous $^{239}Pu(NO_3)_4-H_2O$ mixtures of infinite volume is 7.8 ± 0.3 g ^{239}Pu per liter.
3) *Geometric Control.* The dimensions of a vessel containing fissile material are limited to values below the minimum critical dimensions for said geometric solid containing said fissile material; e.g., the minimum critical diameter of a straight cylindric tube of infinite length, filled with a $^{239}Pu(NO_3)_4-H_2O$ mixture of arbitrary concentration, is 16.28 cm.
4) *Poisoning Control.* In the PUREX process a neutron absorber can either be attached to the solution as a soluble component (*homogeneous poisoning*), or it can be attached to the process vessel holding that solution as a solid construction component (*heterogeneous poisoning*). In either case it is important that the physical presence of the poison is regularly verified, e.g., by continuous monitoring of a homogeneously poisoned solution, or by repeated testing of a heterogeneous poison structure.
5) *Administrative Control.* Administrative control procedures play an important role in criticality prevention, and are manifold in number and character. Consider, for instance, the transfer of a solution which is normally fissile-free but might contain fissile material under abnormal operation conditions, from a critically-safe vessel to a vessel of non-safe geometry. The two vessels will be safety-locked against each other to prevent unintentional transfer, and unlocking for transfer will be authorized only after previous analytical verification of the absence of fissile material.

In an industrial-scale plant where large inventories of fissile material must be handled, criticality control is generally achieved by a combination of the aforementioned principles. Mass limitation alone will only have limited use because of the small size of a critical plutonium mass, but can be applied with advantage in areas where only small batches must be handled, such as the analytical laboratory. Mass limitation in combination with geometric and/or poisoning control is used in areas where batch procedures are operated, such as a batch dissolver. For a continuous process such as solvent extraction, geometric control in combination with poisoning can often be applied with advantage. Gadolinium nitrate has been considered as a *homogeneous poison* for PUREX but its disadvantages are that this expensive material is lost with the high-level waste, and that it is not extracted into the organic solvent phase, so that fissile plutonium and poison would separate in case of an accidental contactor shut-down. *Heterogeneous*

poisoning of a geometrically limited apparatus is therefore often more straightforward. Consider for instance the case of a sieve-plate column for plutonium extraction which is geometrically eversafe up to a diameter of ca. 16 cm (see point 3 in the list). Wider columns as they are required in industrial plants can be made critically safe by incorporating a solid neutron absorber into the structure. A technical solution is, for instance, an annular column with its central tube filled with a neutron poison such as boron carbide. Even more efficient is a design where the sieve plates themselves are made from a corrosion-resistant neutron absorber material, such as hafnium [298].

It is not sufficient to make a single piece of equipment critically safe by its own; the possibility of neutron interaction between neighboring equipment, piping, etc. must be accounted for.

Nuclear Materials Safeguards. The International Atomic Energy Agency (IAEA) and the European Atomic Energy Community (Euratom) were established in 1957 in separate treaties. General aims were to foster and ensure the peaceful uses of atomic energy and to implement safeguards for this purpose. On 1st July 1968, the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) was opened for signature. The IAEA began to develop an NPT-related safeguards system. With 43 ratifying states, the NPT entered into force on March 1970 for an initial period of 25 years. Practical basis for safeguards within the European Community are the Verification Agreement between IAEA and Euratom and its members [299], and the Euratom Regulation 3227/76 [300]. For each nuclear installation Particular Safeguards Provisions are issued.

International nuclear safeguards in general consist of four components: (1) containment, (2) surveillance, (3) accountancy, and (4) inspections. Depending on the type of nuclear facility, or part of it (material balance area, MBA), one or more components are necessary to detect the diversion of nuclear material. *Containment* and *surveillance* using seals and video cameras are typical instruments used in "item facilities" such as nuclear power stations or storage MBAs of reprocessing plants. Here identified and sometimes verified items (e.g., fuel elements) are fixed and observed in such a way that the continuity of knowledge is maintained. For the process area of "bulk handling facilities" (e.g., fuel fabrication, reprocessing) *accountancy* is mostly used. Relying on measurements and records, the operator declares the material balance within a period between two physical inventory takings. The inspectorates verify measurements by observation of operator actions or by independent methods (sample taking, analysis). Because of the great number of actions and data relevant to safeguards, bulk handling facilities need continuous inspections during operation times. Due to measurement uncertainties, differences can always be observed at physical inventory takings between booked and measured inventories. Differences between operator and inspector measurements also occur. The inspectorates evaluate these differences and decide whether they are statistically significant and therefore safeguards relevant. Depending on this decision further investigations might be necessary to fulfill the safeguards obligations.

For industrial-scale reprocessing plants containing higher amounts of "direct use material" (e.g., unirradiated plutonium), the internal IAEA rules lead to difficulties with the goal of timeliness of verification. Since the loss of 8 kg plutonium should be detected within one month, "classic" material balancing (process stop, clean-out, and verification of all nuclear materials) cannot be used. Therefore, an advanced method called " near-real-time accountancy" (NRTA) has been discussed and developed worldwide [301]. The basic idea is to establish short-time material balances from those measurement data on process inventories which are primarily collected for process control purposes. Then, sequential statistical tests are applied to the short-time material balances. The NRTA method is less intrusive to operation than "classic" inventory taking, however, much more information has to be processed and transmitted from the operator to the inspectorates.

5.4.2. Research and Development

Intense research and development work is or has been carried out in a number of countries, with emphasis on two targets. The more immediate of these is the optimization of existing LWR fuel reprocessing technologies, with a specific view on process economy and safety.

There is not a conflict between these aspects; instead, the extremely high safety standards posed in the nuclear industry normally have the effect that safety-related improvements produce economic benefits at the same time, and vice versa. The other target, which is directed more into the future, is the development and qualification of technologies for FBR fuel recycle.

5.4.2.1. Developments in LWR Fuel Reprocessing

A major economic burden in fuel reprocessing is the high cost of radioactive waste treatment, storage, and disposal. Efforts have therefore been made to minimize the production of radioactive waste streams. While it is obviously not possible to decrease the amount of radioactive nuclides to be handled, the possibility exists to decrease the mass and volume of inactive materials which are contaminated by radioactivity. In the chemical separation process, this refers particularly to the liquid medium- and low-activity waste streams, which contain only small amounts of radioactive material but often carry large amounts of dissolved, nonvolatile matter in the form of process reagents and reaction products.

Procedures have therefore been developed in which these nonvolatile reagents are either replaced by electrochemical redox processes, or by compounds such as hydrazine or hydroxylamine which are easily decomposed into gaseous products. Examples for such procedures which are sometimes termed "salt-free", are the plutonium – uranium partitioning by reductants such as uranium (IV) nitrate, hydroxylammonium nitrate, or electro-reduction, and the solvent cleanup by a hydrazine – carbon dioxide solution (see Sections 5.4.1.7, 5.4.1.8, 5.4.1.9). The medium-active waste streams which remain after electrochemical or chemical decomposition of the nitrogen-base reagents are essentially salt-free and can be evaporated down to small volumes. The treatment of these MAW concentrates can be further simplified by routing them into the HAW evaporator and thus combining them with the high-activity waste concentrate, without significant increase of the volume and solids content of the latter.

Another heavy burden on process economy is the requirement of remote operation in heavily shielded building structures equipped with a sophisticated and expensive infrastructure. Any process simplification which decreases the hot-cell space requirement is therefore a step towards lower process cost.

In the chemical separation process, the need for three successive extraction cycles for adequate purification of each product element is a complication which consumes ample hot-cell space. Consequently, efforts have been underway for many years to improve the efficiency of PUREX extraction so that the number of extraction cycles can be decreased. The ultimate outcome is the so- called IMPUREX (IMproved PUREX) process in which one single cycle of TBP solvent extraction combined with a compact nonextractive product refinement is sufficient for efficient purification [302].

The elements of IMPUREX, i.e., the process stages where essential improvements had to be effected, are summarized in Table 62. Core of the process is a PUREX extraction cycle where the solvent is loaded with heavy metals to nearly 100 % in the extraction – scrub section. As discussed earlier, this results in a very high fission product decontamination, but involves the risk of product losses in case of even slight deviations from the specified flowsheet. Two measures were taken to counteract this risk: (1) the uranium extraction front, which in IMPUREX is kept near to the HAW outlet, must be monitored and kept in position by an automatic control system regulating the extractant or feed flow; and (2) conditions had to be found where, in contrast to normal PUREX conditions, plutonium is better extracted than uranium, so that the less valuable product (uranium) instead of the more valuable plutonium would be lost to the HAW in case of a maloperation. Condition (1) could be fulfilled by continuous in-line temperature monitoring for the localization of the extraction front, i.e., by measuring the (exothermic) heat of extraction of uranium. For condition (2), extrapolations of the available data base [265] using the VISCO computer model [303] had indicated that the plutonium (IV) distribution coefficients might exceed those for uranium(VI) at increased temperature and nitric acid concentration, which was then experimentally confirmed (Fig. 129). However, for hexavalent plutonium $D(Pu^{VI})$ is smaller than $D(U^{VI})$ under all conditions, and consequently a pre-reduction

of Pu^{VI} in the feed solution is necessary to avoid plutonium losses. Of great importance is a highly efficient feed clarification, and in the laboratory-scale demonstrations discussed below, a fine-pore sinter metal filter followed by a bed of diatomaceous earth was effective. However, secondary waste is produced by this process, and consequently other possibilities such as an ultra-high-speed centrifuge should be investigated. Uranium – plutonium partitioning is performed using two electro-reduction columns in series instead of a single unit, resulting in a uranium product highly decontaminated from plutonium [304].

Table 62. Elements of IMPUREX

Measure	Purpose	Procedure
Feed clarification	avoid interfacial crud and minimize solvent degradation	sintered metal filter diatomaceous earth filter bed
"100 %" Solvent loading	high decontamination vs. yield	auto- control U extraction front
Feed prereduction	avoid Pu(VI) losses	e.g., electrochemically
Distribution coefficient $D_{Pu} > D_U$	avoid Pu(IV) accumulations	increased acidity and/or elevated temperature
Optimal extraction cycle	increase decontamination	optimize HA/HS 1/HS 2 length; adjust scrub flow ratios
Advanced U – Pu split	improve Pu-decontamination	two serial electroreduction columns
Additional product refinement	moderate maloperations backup	crystallization
N_2H_5OH/CO_2 solvent wash	medium active waste minimization	oxidation → N_2 + evap. → high active waste

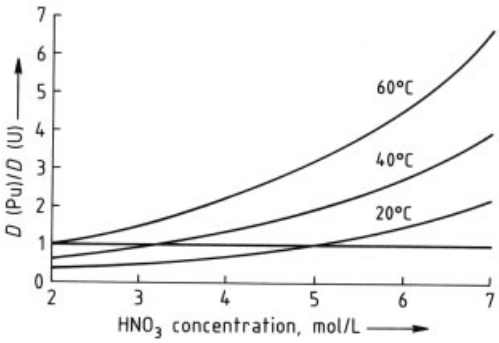

Figure 129. Ratio of distribution coefficients for Pu(IV) and U(VI) as a function of nitric acid concentration and temperature, extractant = 30 vol % TBP – *n*-alkane [298]

A first "hot" demonstration of the one-cycle process was performed in the highly shielded, laboratory-scale extraction facility known as MILLI (because its throughput is 1 kg, or a milliton, of irradiated fuel per day) at the Institut für Heisse Chemie, Karlsruhe, Germany, using a fast-breeder MOX fuel of ca. 12 % Pu content, 110 000 MW · d/t average burnup, and 15 months cooling time which originated from the KNK II power station [305]. Extraction in the A mixer – settler was performed at an aqueous HNO_3 concentration of 4.9 mol/L and at 56 °C, with an organic solvent loading of ca. 93 % at the feed stage. Based on the VISCO model, the position of the extraction front which was monitored by temperature measurement as described above, was used as the reference variable for control of the solvent flow rate. Uranium losses to the HAW were restricted to 100 mg/L (corresponding to 0.1 %) while the plutonium losses, mainly caused by Pu^{VI} traces, were 10 mg/L (0.2 %). The feed filter system worked extremely well; no signs of crud formation could be seen, and despite the higher temperature and acidity, less dibutyl phosphoric acid than in former standard PUREX campains was formed. Fission product decontamination of the purified products was very satisfactory and for most impurities specifications were met. Table 63 summarizes the results of this experiment, and of a second campain in which LWR fuel irradiated to 31 500 MW · d/t burnup in the Swiss Goesgen power station was processed [305]. Final purification of uranium by the crystallization process described below yielded additional decontamination factors of > 100 for the most important fission products.

Freeze crystallization of uranyl nitrate hexahydrate (UNH) was developed as a compact uranium product refinement procedure [306]. The IMPUREX product solution is concentrated by evaporation to 300 – 700 g uranium per liter and adjusted to 4 – 6 mol HNO_3 per liter. Uranyl nitrate hexahydrate is crystallized by lowering the temperature to − 20 to − 30 °C at a rate of 1 to 3 °C/min. The high decontamination achieved for most impurities is based on the unique structure of the UO_2^{2+} cation and, hence, of the $UO_2(NO_3)_2 \cdot 6\,H_2O$ lattice which has no tendency to accept "normal" metal cations. Penta- or hexavalent neptunium and plutonium which would cocrystallize with the uranyl nitrate, are

Table 63. Fission product decontamination factors for uranium attained in the MILLI first cycle for two IMPUREX campains (FBR, LWR) as compared with a conventional PUREX campain (FBR), and with the DF required for 7 years cooled LWR fuel [305]

Fission Product	Approximate requirement for LWR	Conventional PUREX FBR	IMPUREX LWR (1990)	IMPUREX FBR (1988)
^{137}Cs	1×10^6	1.7×10^4	$7.4\times10^5-9.6\times10^7$	$5\times10^5-1.3\times10^6$
^{144}Ce	1×10^6	6×10^4	$1.7\times10^6-5\times10^8$	$1.5\times10^6-4.5\times10^6$
^{106}Ru	1×10^4	1.7×10^2	$3\times10^5-2\times10^6$	$1\times10^5-3\times10^5$
^{95}Zr	1×10^4	6×10^2	$5\times10^4-8\times10^5$	$1\times10^4-1.5\times10^4$
^{99}Tc	1.7×10^3		$>2\times10^4$	5×10^2

rejected after prior reduction to lower valency states. The mother liquor and wash solutions which contain up to 10 % of the uranium are recycled into the solvent extraction feed stage to recover the uranium together with residual plutonium, while the fission products report into the HAW; in this way, uranyl nitrate crystallization is an ideal "no-waste" process.

5.4.2.2. Reprocessing of Fast Breeder Reactor Fuels

The fuel for the sodium-cooled oxide-fuelled FBR is $(U, Pu)O_2$, a material similar to the UO_2 and MOX fuels used in LWRs. Indeed there are many more analogies in the reprocessing requirements for FBRs and LWRs, and it is therefore not surprising that most efforts for elaborating the technology for FBR fuel recycle have concentrated on the successful PUREX process. However, the technical requirements are distinctly more stringent in the FBR case, primarily due to the following reasons:

1) The concentration and radioactivity level of fission products in FBR fuel is higher at the time of reprocessing (Table 64) because the fuel burnup is higher, the cooling time is shorter, and the fission nuclide yield distribution is different. As Table 64 shows, the problem is somewhat eased if the axial blanket is coprocessed with the core fuel; it is further eased if the radial blanket is also coprocessed. The higher radiation level, and the associated higher thermal power have consequences on fuel transport and storage, on the layout of buildings and equipment, and on the radiation dose received by process reagents. It also requires a higher fission product decontamination factor if the same product specifications are applied as in the LWR case. A special problem is that part of the noble metal fission products remain undissolved as an insoluble, highly radioactive residue in the dissolver (ε-phase, see Section 5.4.1.4).
2) The plutonium concentration in FBR fuel is higher, but again the problem is eased if blanket material is coprocessed with the core fuel (Table 64). This has consequences on the criticality layout, and makes adaptations of the PUREX flowsheet necessary. In the dissolver the lower dissolution rate of MOX must be taken into account, and the fabrication of a highly homogeneous MOX is obligatory for proper fuel dissolution (see Section 5.4.1.4).
3) The use of sodium as the FBR reactor coolant influences fuel element transport and storage as well as the head-end treatment in the reprocessing plant because residues of sodium must be deactivated before the fuel element can be exposed to air, water, or nitric acid.
4) The design of FBR fuel elements is characterized by a very close package of thin fuel pin "needles" which are held in position by grid or spiral wire spacers, and by a mechanically stable stainless steel wrapper which holds this pin bundle arrangement together. This has specific consequences on the mechanical head-end treatment.
5) The use of stainless steel instead of Zircaloy as the cladding and structural material in FBR fuel elements is of relatively little influence on reprocessing. A specific chemical effect is that tritium is not chemically bound in stainless steel but instead diffuses through this material in the reactor, so that only a fraction of the tritium generated will still be present in the fuel at the time of reprocessing.

Modifications of existing technologies which have so far been elaborated are shortly reviewed

Table 64. Order-of-magnitude fuel data at the time of reprocessing for FBR and LWR

Parameter	FBR			LWR
	Core only	Core + axial blanket	Core + total blanket	
Average burnup, MW · d/t	75 000	45 000	35 000	35 000
Cooling time, a	1	1	1	5
Average Pu concentration, % of total heavy metals	15	10	7	1
Average fission product concentration, kg/t	77	46	36	36
Average fission product radioactivity, Bq/t	1.5×10^{17}	9×10^{16}	7×10^{16}	2×10^{16}

below. However, FBR reprocessing is not yet an industrially established process and still in the state of research and development.

Fuel Element Storage and Transport. Spent FBR fuel elements are discharged from the reactor into a sodium cooling pond which is usually integrated into the reactor. Due to the short cooling time and consequent higher thermal power output, transport to the reprocessing plant is best made using liquid sodium as the heat transfer medium.

Head-End Treatment. Sodium adhering to the fuel element after unloading from the transport flask must be deactivated by a water vapor – nitrogen treatment before the shear – leach process can be carried out. The stainless steel wrapper makes shearing of the entire FBR fuel element unadvisable and previous mechanical dismantling is preferred. The wrapper can be cut open either by a mechanical tool such as a saw, or by laser cutting. After removal of the wrapper the fuel pin bundle is sheared into segments of ca. 3 cm length which fall down into the dissolver. The dissolution rate is slower, and the amount of insoluble solid radioactive residues, notably the metallic ε-phase, is higher than in the LWR case.

Feed Pretreatment. Procedures used for iodine removal and clarification of the dissolver product are similar to those applied in LWR fuel reprocessing. Due to the high radioactivity of insoluble fines, with $^{95}Zr - ^{95}Nb$ still present, a highly efficient feed clarification is particularly important to avoid difficulties in the extraction process. Adjustment of the plutonium valency to +4, e.g., by careful electrochemical reduction of Pu (VI) and re-adjustment with NO_2, is advisable for the prevention of plutonium losses.

PUREX Process. Flowsheets for separating plutonium and uranium by the conventional three-cycle PUREX process have been elaborated and successfully tested on the laboratory and pilot plant scale. Measures which suppress the formation of dibutyl phosphate cruds, such as a high solvent loading and a high acidity in the extraction-scrub contactors, are of particular importance. For U – Pu partitioning salt-free processes are highly recommended. The British Dounreay Fast Reactor Reprocessing Pilot Plant which processes fuel from the 250 MW_e Prototype Fast Reactor, uses sulfuric acid as a nonreductive complexant for U – Pu partitioning [307]. The choice of highly corrosive sulfuric acid was dictated by capacity considerations when the original, much smaller facility was redesigned and scaled up to PFR throughput, and it is unlikely that this reagent will be applied in a commercial-scale plant.

The one-cycle IMPUREX concept (see Section 5.4.2.1) [305] is especially suitable for reprocessing of FBR fuel.

Other variations of the standard PUREX scheme include the replacement of TBP by trialkyl phosphate esters which show less tendency to form a third phase with plutonium (IV) (see Section 5.4.1.6). Promising results have for instance been reported for tri-isoamyl and tri-isohexyl phosphate extractants [308].

Status of FBR Fuel Reprocessing. In the most advanced countries the scale of FBR fuel reprocessing has reached the pilot-plant level. Thus, in France, the Atelier Pilot de Marcoule is used for pilot-scale demonstrations, and from 1977 to 1983 ca. 9 t of fuel from the 250 MW_e Phénix reactor containing 1.3 t plutonium have been treated. The plant was then refurbished

and extended, and is now available for demonstrations with both Phénix and Superphénix (1200 MW_e) fuel elements [309]. In addition several tonnes of Phénix fuel were coprocessed in mixtures with low-burnup, low-plutonium gas-graphite reactor fuel in the UP-2 plant at Cap de la Hague, demonstrating a way how FBR fuel could be economically processed in a period when FBRs are being introduced into an existing array of thermal power reactors, and when special FBR reprocessing plants are not yet existing.

The British Dounreay pilot plant which is operated in connection with the 250 MW_e Prototype Fast Reactor has already been mentioned [307]. In Japan, a pilot-scale FBR reprocessing plant which will be operated in conjunction with the 280 MW_e MONJU FBR is in the planning stage; its design is based on experience with a laboratory-scale facility at Tokay Mura. Other countries, such as the former Soviet Union [308] and Germany [305], have reported the reprocessing of FBR fuels on the laboratory scale.

Other Processes. A very different FBR reactor and fuel cycle concept is presently under investigation in the United States. The so-called Integral Fast Reactor (IFR) concept involves a pool-type, sodium-cooled reactor using a metal fuel, i.e., a 19 wt % U – 10 wt % Pu – Zr alloy, and a pyrometallurgical procedure for reprocessing and recycle of the fuel and blanket. The objectives of the IFR process are adequate (not necessarily high) removal of fission products, replacement of the plutonium consumed by fission with fresh plutonium generated in the blanket, and a > 98 % recovery of uranium and plutonium.

If successfully developed, the IFR process which is investigated at the Argonne National Laboratory, is planned to be tested on the pilot plant scale in the former EBR II Fuel Cycle Facility at Idaho Falls, Idaho. Modification of this facility which is still existent for IFR fuel cycle demonstrations is in the design stage [311].

The process which is investigated for the treatment of the core fuel is *high-temperature electrorefining* [310]. The fuel element is mechanically disassembled, and the pins are chopped into segments of ca. 6 mm length which are charged in an anode basket into the low-carbon steel electrorefiner vessel. The metal alloy fuel is leached out of the cladding hulls at 450 – 500 °C by liquid cadmium which becomes the anode of the cell. Uranium and plutonium are electrolytically transported through a molten salt electrolyte overlying the cadmium phase, and are deposited at a suitable cathode, e.g., cadmium contained in a ceramic crucible. The salt electrolyte is composed of stable chlorides, i.e., 35.5 mol % $BaCl_2$, 19.5 mol % $CaCl_2$, 37.0 mol % LiCl, and 8.0 mol % NaCl. Fission products are removed in three streams: (1) *gaseous fission products* are released during chopping and dissolution; (2) the *electrolyte salt* which contains electropositive fission products such as alkali, alkaline-earth, and rare earth metals; and (3) the *cadmium anode melt* which contains the noble metals fission products. For the separation of the purified U – Pu product from the cadmium cathode, vaporization of the cadmium is envisaged. For the processing of blanket material, a *halide slagging process* has been devised. The blanket alloy is molten at 1250 – 1300 °C in a beryllium oxide crucible which is not attacked by uranium, and the plutonium is preferentially extracted as $PuCl_3$ into an overlying $BaCl_2$– $CaCl_2$ salt melt which contains UCl_3 as the oxidant for plutonium. The salt extract in which the Pu: U ratio is much higher than in the original blanket alloy, is added to the electrolyte of the core electrorefiner to effect reenrichment of the core material with plutonium.

5.5. Radioactive Waste Management

5.5.1. Classification of Radioactive Waste

5.5.1.1. Generic Classification

Classifications of radioactive waste have often been proposed [312–316], but no generic, universally applicable, and quantitative classification exists for raw waste or for conditioned waste packages. A multitude of subdivisions for radioactive waste has been proposed or applied. Most commonly, a classification according to low active waste (LAW), medium active waste (MAW), and high active waste (HAW) is utilized. This subdivision is useful in handling and treating radioactive waste of a known type and radionuclide composition. Definitions of these classes differ greatly. Activity concentrations,

dose rates, and content of α emitters have, for example, been used as parameters for classification along with definitions according to the origin or type of waste.

Simple classifications according to LAW, MAW, and HAW do not allow a safety-related judgement of waste. They do not include sufficient information to draw conclusions for the release of radionuclides and the dose to humans under normal or extraordinary conditions (defined in the assumed scenarios of safety assessments). The resulting radiation exposure in a given scenario can be higher from a poorly conditioned waste with low activity than from an excellently conditioned high-activity waste. To comply with the requirements for radioactive waste, the simple definitions of LAW, MAW, and HAW must be abandoned and a more informative system of waste description established.

5.5.1.2. Waste Classification for Disposal

The system of waste classification is best developed from a disposal point of view because a repository must handle all types of waste from various waste producers and cannot be tailored to an individual waste source.

Radioactive waste can be divided into *heat-generating waste* and *non-heat-generating waste* [316], because the thermomechanical effects that result from heating the host rock necessitate very specific, additional investigations, layouts, and proofs. Non-heat-generating waste can additionally be subdivided into so called α *waste* and *non-α waste*. This subdivision is important if non-α waste with short-lived radionuclides is intended to be disposed of in shallow ground. Waste with long-lived α-emitting radionuclides must be disposed of in deep geological formations.

For further characterization, various radioactive waste must be categorized [317], [318] and their properties described individually in data sheets that include sufficient information to plan the installation and to perform safety assessments.

As a consequence of site-specific safety assessments, the various waste categories can be condensed into a few, well-defined classes with similar behavior for release of radionuclides. This has been demonstrated for the Konrad repository [319]. As a result, six *waste form groups* and two *waste packaging classes* [319], [320] were defined.

5.5.1.3. Catalog of Waste Types

The definition of waste form groups and waste packaging classes is the starting point for the development of a waste type catalog because different radioactive wastes must be adequately separated to allow appropriate pretreatment and conditioning for interim storage and disposal. The waste type catalog surveys radioactive waste from its generation in a nuclear facility until its disposal in a repository [321], [322]. This catalog specifies the raw waste, pretreatment methods, and resulting intermediate products, as well as the conditioning methods and waste packages required. It thus describes the "flow" of radioactive waste.

The subdivision of *raw waste* is hierarchically organized into *waste groups*, *waste subgroups*, and *waste types*. The assignment of raw waste to a waste group, waste subgroup, or waste type must be performed in accordance with the requirements of waste pretreatment and the conditioning process. If necessary, raw waste must be sorted or specified in more detail, for example, if it is categorized as chemicals, filters, or metals. Raw waste can be *pretreated* to form an *intermediate product* and must be *conditioned* (treatment and packing) to a *waste package* (waste form and packaging).

This waste type catalog, therefore, forms the basis for the description and separation of radioactive waste from the raw waste into a waste package suitable for disposal. It guarantees the necessary clarity in waste management, gives guidance for waste conditioning, allows determination of relevant data for classified wastes, and ensures sufficient supervision of various waste management steps by the authorities.

5.5.2. Conditioning of Radioactive Waste

5.5.2.1. Heat-Generating Radioactive Waste

Heat-generating radioactive waste contains such a high concentration of radionuclides that heat is emitted which necessitates cooling. This waste

includes spent fuel elements (if these are deemed to be a waste), high-level liquid waste and dissolution residues such as hulls and ends from spent fuel reprocessing, and highly activated core components from reactors.

High-Level Liquid Waste. High-level liquid waste results from the reprocessing of spent fuel and represents fission product concentrates that contain about 300 g/L of element nitrates in nitric acid solution. These liquids are stored in cooled stainless steel tanks. The storage time is kept short (e.g., one year), if a conditioning process and an interim storage facility or a repository are available.

Various conditioning concepts have been discussed and investigated. Examples are vitrification in borosilicate and phosphate glasses [323], [324]; embedding in ceramic materials (e.g., Synroc, [325]); and production of an advanced calcination product (e.g., Supercalcine). At present virtrification in borosilicate glasses is the primary choice for solidification of high-level liquid wastes. The following are the principal process steps involved: (1) concentration, (2) denitration, (3) drying, (4) calcination, (5) melting with glass frit, and (6) pouring the melt into a stainless steel canister. The off-gas of a vitrification process passes through an off-gas cleaning system before it is discharged.

Two virtrification processes have been applied in longer, active operation: the AVM and the Pamela processes.

A VM process (Atelier de Virtrification Marcoule, France). The main pieces of equipment in this process are a rotary kiln calciner and a metallic melter [323]. The high-level liquid waste is mixed with sugar as a denitration agent and then fed into the *calciner* where it is denitrated, dried, and calcinated. The calcine, the dissolution fines, and the glass frit are fed into the *melter*, which is operated at ca. 1100 °C. The radioactive waste is dissolved and the fines are embedded in the glass which is poured batchwise through a bottom drain into the canister. The glass solidifies on cooling. Based on experience with the AVM process since June 1978, a vitrification plant has been built to immobilize high-level liquid waste from the reprocessing plant at La Hague, France. The first hot glass block was cast in June 1989.

Pamela Process (acronym for: Pilot-Anlage zur Erzeugung lagerfähiger Abfälle, Belgium). The primary component of this process [324] is an electric, direct-heated ceramic melter. A mixture of liquid waste, glass, and additives is fed directly into the melter. This feed forms a layer of solid product on top of the melt, which reduces the release of volatile radionuclides such as ruthenium. All steps in glass fabrication–denitration, drying, calcination, and melting–are carried out inside the melter. The glass is drained off either through a bottom outlet or via a continuous overflow.

The Pamela facility, run in German – Belgian cooperation, started operation in 1985 and will have solidified high-level liquid waste stored at the Eurochemic reprocessing plant in 1991. After that, the Pamela plant might be used to vitrify high-level liquid waste from the reprocessing plant at Karlsruhe (WAK), if the transport of the liquid HAW is accepted.

Dissolution Residues and Reactor Core Components. The spent fuel elements are chopped up, and the fuel is dissolved in nitric acid solution in a reprocessing plant. Hulls, end pieces, and insoluble fines remain as residue. These require a conditioning technique that yields a product with sufficient radiation and temperature resistance. Cementation has been applied [326], and high-pressure compaction, resulting in the fabrication of pellets, is in the developmental stage [327].

Reactor core components are packed primarily into high-integrity packages without fixation [328].

Spent Fuel. If spent fuel is disposed of as radioactive waste it has to be conditioned in such a way that the volatile radionuclides in the fuel pins are retained in the waste package or released during conditioning.

One possibility is packing of spent fuel elements (or fuel pins after disassembling) into a thick-walled, gastight, self-shielded, high-integrity package that can be stored and/or disposed of [329]. Packaging made of cast iron, steel, copper, or lead has been developed. It can have an overall mass up to 120 t if used for transportation and storage. The weight of waste packages used for the disposal of spent fuel in deep

geological formations is currently limited to 65 t because of the capacity of the hoist cage.

An alternative is packaging cut fuel pins in a steel canister that has the same dimensions as the canister for glass blocks from virtrification of high-level liquid waste [330]. In this case, identical techniques for storage, transportation, and disposal of spent fuel and glass blocks can be used.

Interim Storage of Immobilized Heat-Generating Radioactive Waste. Heat-generating radioactive waste must be stored in such a way that the heat is removed. In principle, air or water cooling can be used, but air cooling is preferred at present. Several technical solutions are possible and the following concepts have been selected:

1) Thick-walled, cast-iron packages of heat-generating radioactive waste, especially spent fuel, are stored in a concrete building with inlets for cooling air in the side walls and outlets in the roof [331]. The packaging is the barrier against radionuclide release, therefore, the building is not considered as protection in safety assessments. Such a facility has been built at Gorleben, Germany. The licensed facility has not yet been taken into operation.
2) An air-cooled store for high-level glass blocks is operated in Marcoule, France [323]. The glass blocks are emplaced with a loading machine in vertical shafts that are cooled by a heat-induced natural air draft supported by the exhaust stack. The storage shafts are plugged, and thus the area on top of the shafts is accessible. The building has no protection against outside factors such as an airplane crash. A similar facility is in operation at La Hague, France.

5.5.2.2. Non-Heat-Generating Radioactive Waste

Minimization and Avoidance of Radioactive Waste. The minimization of radioactive waste is especially important for non-heat-generating waste. This waste consists mainly of inactive materials that are contaminated or activated and contain only small amounts of radioactive isotopes.

In principle, all materials that are introduced into a radioactive process or into a radioactive control area could be released as radioactive waste. Systematic, intelligent process and plant design is of utmost importance for the reduction of waste which can be achieved by:

1) Electrolytic techniques [332];
2) Chemicals that react to form nontoxic gaseous compounds or that can be decomposed by electrolysis or by thermal treatment [332], [333];
3) Use of heterogeneous instead of homogeneous neutron poisons [334];
4) Use of effective decontamination methods with a minimum of waste and re-use of decontaminated materials [333]; and
5) Decay of short-lived radioactivity and disposal as nonradioactive waste.

The minimization of radioactive waste through the avoidance of waste in combination with volume-reducing waste conditioning techniques has already been successfully applied. This results in continuously declining amounts of conditioned radioactive wastes.

Conditioning Techniques. Conditioning of radioactive waste includes processing and packaging of the waste, if required, after pretreatment or sorting. Various techniques and strategies are applied. The selection of a conditioning process depends on factors such as the requirements for storage and disposal; the overall costs of conditioning, storage, and disposal; acceptance of the process; and the volume of the resulting waste packages. Therefore, waste producers in different countries may apply different conditioning techniques to the same type of waste. Furthermore, the need to minimize the volume of conditioned waste, because of repository shortage, stimulates the development of new conditioning techniques.

Pretreatment of Radioactive Waste. Raw waste must be collected and pretreated in such a way that it becomes suitable for the conditioning process. Pretreatment methods include decontamination, crushing, compression, evaporation – distillation – rectification, decantation – dewatering – filtration, and incineration – pyrolysis [321], [322], [333], [335], [336].

Incineration is especially attractive for all types of combustible waste. Solid or liquid waste including alpha contaminated waste are incinerated. The large volume reduction and the inorganic and inert character of the intermediate products (e.g., ash or slag) favor the incineration process [321]. However, off-gas treatment and secondary waste must be taken into account.

Treatment of Radioactive Waste. The *cementation* of radioactive waste is widely applied [333], [335]. It is used for (1) the solidification of liquids; (2) the embedding of solids such as ash, powder, and granules; and (3) the casting of voids in scrap, rubble, or filters. Various cementation techniques are used, and the equipment can be mobile or stationary.

Liquid radioactive waste is mixed with dry cement. This mixing is carried out by continuous mixers, batchwise in-drum mixing, tumbling, or with "lost" stirring blades (i.e., blades that remain in the waste after cementation).

The liquid radioactive waste reacts with the cement and is bound in the product. The compatibility of the cement and the liquid radioactive waste must be checked to guarantee that the product sets. If necessary, special cement formulations are developed or suitable additives are used.

Solid radioactive waste such as drying residue from liquids, ash from the incineration of combustible radioactive waste, or granulated material such as silica gel or ion exchange resins is often embedded in cement by mixing techniques. Reactions with the cement must be taken into account, e.g., reactions with amphoteric metals in the ash under evolution of gas or swelling of dried resins.

Casting of void spaces between solid materials is used to fill the space between wastes, to contain possible surface contamination or residual liquid, and to hinder radionuclide release.

Cementation technology is relatively simple and is widely applied. However, more advanced processes are becoming very attractive because of volume minimization.

Bituminization is a classic process for the solidification of liquid radioactive waste [335]. The liquid and the bitumen are mixed at $\geq 150\,^{\circ}\mathrm{C}$ in an extruder, a film, or a mixer – evaporator. The solids contained in the liquid are incorporated into the bitumen, which cools down after it is poured into packaging. The bituminization process must be carefully planned to avoid, e.g., exothermic reactions or self-ignition of the hot bitumen product. The product must be protected in a suitable packaging or building to avoid unacceptable radionuclide releases, e.g., as a consequence of a fire. Furthermore, the tendency toward swelling due to radiolytic gas generation must be taken into consideration. These facts hamper the application of bituminization.

Polymers are used in only a few cases [336], [337]. The embedding of resins with polystyrene is one example. The solidification of tributyl phosphate in poly(vinyl chloride) is being investigated. In general, the monomers are combined with the radioactive waste. Mixing techniques are comparable to those used in bituminization. Polymerization is initiated by a catalyst or by radiation.

High-pressure compaction with 1500 – 2000-t compactors is a new development for minimizing waste amounts [328]. Solid material in a drum is compacted to a form-stable pellet. This technique is used with metallic materials, paper, plastic and clothes, rubble, and even ash from the incineration of organic radioactive waste. In 1988, unexpected gas generation was detected in compacted waste. Investigation showed that corrosion occurred when the metal came in contact with wet organic materials. This experience demonstrates that waste must either be dry or sorted before compaction.

Drying of liquid radioactive waste is a new development for waste minimization [328]. The liquid is fed into a packaging that is heated. Evaporation is supported by application of a slight vacuum. The resulting dry residue consists of the inactive and active constituents that were dissolved or dispersed in the liquid. The activity concentration in the resulting product is thus higher than in the cemented waste form; therefore, superior shielding (often made of cast iron) is required.

Melting of metallic radioactive materials (e.g., after decontamination) is of special interest for the decommissioning, repair, and maintenance of nuclear facilities. If the radioactivity level in the melt is low the metals can be reused in casting packagings for radioactive waste disposal. If the radionuclide content is too high the melt must be poured into packaging as radioac-

tive waste. The resulting slag also has to be conditioned and disposed of.

Packaging of Radioactive Waste. Radioactive waste must be packaged for handling and transportation. Packagings for disposal may be designed for the following functions:

shielding of the radiation,
retention of volatile radionuclides,
protection against undesired events (e.g., fire or dropping of a waste package), and
acting as a barrier to radionuclide release after an ingress of water.

The necessary quality of a packaging depends on the type of waste and its radionuclide inventory. Packaging requirements can be fulfilled by the packaging alone or by a shipping cask. Sheet steel, reinforced concrete, and cast iron are common packaging materials. Cylindrical and rectangular packagings of different size and weight are being used. Standardization of packaging is necessary to make the equipment, as well as the handling and disposal techniques, compatible [320].

The conditioning of radioactive waste by packaging only is also employed, primarily with superior packaging (e.g., cast iron). This packaging is employed for activated materials and ion-exchange resins that have higher radionuclide inventories [328].

5.5.3. Origin and Amount of Radioactive Waste

A realistic data base is an essential prerequisite for the development of disposal strategies or the planning and construction of repositories. Data on the origin, type, and expected amount of radioactive waste are therefore necessary to estimate the required disposal capacity.

5.5.3.1. Origin of Radioactive Waste

Radioactive waste basically originates from:

1) The operation of light-water reactors and reactors of advanced design (breeder reactors and pebble bed reactors);
2) The reprocessing of spent fuel elements from nuclear power plants;
3) Basic and applied investigations at nuclear research centers;
4) Other research centers, universities, industries, or medical applications of radioisotopes;
5) Uranium enrichment and the production of fuel elements, and industrial research and development;
6) Decommissioning and dismantling of nuclear facilities; and
7) Other waste producers, e.g., the pharmaceutical industry or the army.

In addition, spent fuel elements from light-water reactors and pebble bed reactors must also be considered.

5.5.3.2. Present Amount of Unconditioned and Conditioned Radioactive Waste

The origin and amount of radioactive waste in Germany [338] are representative for a country having a well-developed nuclear technology. Thus, they may be used to provide information on the amount of radioactive waste arisings.

On behalf of the Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU, i.e., Federal Minister for Environment, Nature Protection, and Nuclear Safety), the Physikalisch-Technische Bundesanstalt (PTB) has carried out annual inquiries concerning the amount of unconditioned and conditioned radioactive waste produced in Germany since 1985. [The responsibility for the long-term storage and disposal of radioactive waste was transferred from the PTB to the newly founded Bundesamt für Strahlenschutz (Bf S) on November 1, 1989.] According to the latest inquiry [339], the amount of unconditioned radioactive waste (primary waste volume) was about 13 000 m^3 on December 31, 1989 (including about 1300 m^3 of short-lived waste stored for decay). This was produced mainly by nuclear power plants (ca. 8100 m^3), collecting depots of the federal states (about 2400 m^3; radioactive waste originating from other research centers, universities, industries, or from medical application of radioisotopes is generally handed over to collecting depots), and the Karlsruhe and Jülich nuclear research centers (about 1200 m^3). The volume of conditioned (i.e., processed and packaged) radioactive waste amounted to about 44 400 m^3. Of this,

Table 65. Amount of unconditioned and conditioned radioactive waste in the Federal Republic of Germany (including short-lived waste stored for decay)

Origin	Primary waste, m^3	Conditioned waste, m^3		
	Amount stored up to Dec. 31, 1989	Amount stored up to Dec. 31, 1989	Amount produced from Jan. 1 to Dec. 31, 1989	Estimated amount in 1990
Reprocessing	148	8 955	615	615
Nuclear power plants	8 140	13 827	997	3 017
Collecting depots	2 384	1 699	101	79
Nuclear research centers	1 245	18 105	1 760	2 681
Nuclear fuel cycle industry	878	1 743	266	516
Other waste producers	236	91	8	8
Total	13 031	44 420	3 747	6 916

waste from the research centers contributed ca. 18 100 m^3; waste from the operation of nuclear power plants, ca. 13 800 m^3; and that from reprocessing spent fuel elements ca. 9000 m^3. Details on the origin and amount of unconditioned and conditioned radioactive waste are given in Table 65.

In comparison to the amount of radioactive waste occurring in other countries, the total waste package volume in Germany is rather small. It indicates the use of modern conditioning techniques (i.e., waste product minimization via high-pressure compaction, incineration, and thick-walled cast-iron packaging).

5.5.3.3. Future Amounts of Radioactive Waste

Based on annual surveys and data presented to the BfS by waste producers, future amounts of conditioned radioactive waste can be estimated up to the year 2000 [339]. Taking into account that the volume of arising radioactive waste was always lower than the amount calculated on the basis of model data from the waste producers, two forecasts with different data were made. The forecast on the basis of the model data is assumed to be an upper limit for the occurring waste, and the forecast on the basis of an extrapolation of the number of actually arising waste packages is assumed to be a lower limit for the amount of waste at the end of the year 2000. Thus, the prognosis is based on the following assumptions and boundary conditions: (1) the installation of nuclear power plants with a total electricity-generating capacity of about 24 GW by the year 2000; (2) the return of radioactive waste due to the reprocessing of spent fuel elements from nuclear power plants in Germany from 1993 onwards [reprocessing is carried out by Compagnie Générale des Matières Nucléaires (COGEMA) in France and British Nuclear Fuel plc (BNFL) in Great Britain], and (3) an annual amount of radioactive waste originating from nuclear research centers, the nuclear fuel cycle industry, the collecting depots, and from other waste producers which, based on current amounts, is considered to be constant.

The prognosis predicts an amount of radioactive waste with negligible heat generation at the end of the year 2000 (see Fig. 130) of about 170 900 m^3 (upper estimate) or of about 129 900 m^3 (lower estimate). Taking into account the overestimation of the expected amount of waste by the model data [339], it is interesting to see that the waste quantities estimated by the PTB on the basis of the inquiry for 1986 lie just in the middle of this bandwidth [340].

The distribution of the predicted amount of waste (upper estimate) between the various groups of waste producers at the end of the year 2000 shows the following results:

reprocessing	32.2 %
nuclear power plants	28.8 %
research centers	26.2 %
others	12.8 %

The amount of conditioned heat-generating radioactive waste at the end of the year 2000 is estimated to about 3600 m^3 (upper estimate) or to about 3200 m^3 (lower estimate).

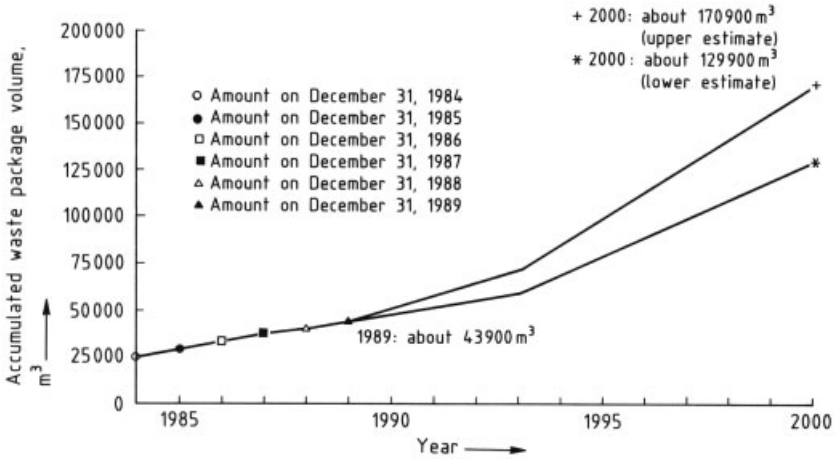

Figure 130. Expected amounts (upper and lower estimates) of conditioned radioactive waste with negligible heat generation accumulated up to the year 2000

5.5.4. Disposal of Radioactive Waste

The overall objective of radioactive waste disposal is to ensure that waste is handled and stored in such a way that, at present and in the future, the protection of humans and the environment from harm caused by the ionizing radiation from these wastes is guaranteed [313], [341]. Effective protection can be achieved by adhering to the basic principles of radiation protection; proper planning; and the adoption of appropriate methods for the conditioning, handling, and disposal of radioactive waste. Important considerations for waste disposal are the selection of an appropriate disposal site with its specific confinement characteristics and the implementation of engineered barriers (e.g., immobilization of material, packaging, disposal room, and shaft sealing) so that radioactive waste is adequately isolated from the biosphere.

Wide international agreement exists that the multibarrier concept (i.e., imposing natural and engineered barriers between radioactive waste and the biosphere) is particularly advantageous and should be adopted in any waste disposal system. Thus, the release of radioactive material can be limited or delayed, its transport retarded, or its concentration sufficiently diluted to ensure that the impact on humans and the environment remains within acceptable limits.

5.5.4.1. Principles of Waste Disposal

Two basic waste management techniques exist that can be used to comply with the objective of disposal:

1) *Disposal to the environment* so that radioactive waste is diluted to a level that posess no harm to humans and the environment
2) *Containment of the waste*, to achieve the required degree of isolation from humans and the environment, until the radioactivity has decayed to insignificant levels

In practice, disposal to the environment has been realized by sea dumping and emplacement in the ocean floor; containment of the waste has been realized by shallow land burial and disposal in geological formations. The safety of sea dumping relies on the vast dilution of the radioactivity as it disperses at the bottom of the ocean. The quantities of activity dispersed by this method are very small compared with the natural activity of the oceans. *Sea dumping* of radioactive waste was successfully practiced under the auspices of the OECD Nuclear Energy Agency [342]. However, this operation was stopped in 1983, and it will probably not be resumed. The research and development program for *emplacing vitrified radioactive waste in the ocean floor*, coordinated by the OECD Nuclear Energy Agency, also suffered a severe setback

by the withdrawal of the United States from international cooperation.

With these techniques out of favor, the primary emphasis in radioactive waste disposal now rests with land-based facilities. The international view is that shorter-lived radionuclides, or waste containing rather low concentrations of long-lived radionuclides can, if appropriate, be *buried* near the surface. Long-lived radioactive waste requires a higher degree of isolation and should be disposed of predominantly in geological formations. Thus, shallow land burial in highly engineered facilities (i.e., near-surface repositories) or disposal in deep-lying repositories has generally been favored.

The objective of waste disposal applies to all types of radioactive waste. To demonstrate that this objective can be achieved, according to a wide measure of agreement, a safety assessment must be performed that takes into account the geological and hydrogeological situation, the technical design of the facility including its scheduled mode of operation, and the type of waste packages to be disposed of.

5.5.4.2. Underground Laboratories

An essential prerequisite for a safety assessment-oriented approach to disposal is the provision of a comprehensive, realistic data base. The necessary data must be provided by site-specific above-ground and underground investigation programs, by information on disposal-relevant characteristics of the waste packages, and by appropriate planning of the disposal facility.

A very important feature in the framework of research and development programs is the operation of underground laboratories. In these laboratories, in situ measurements and experiments can be carried out, and waste handling and emplacement techniques can be demonstrated. Thus, necessary data and experience can be obtained to aid the design of future repositories in geologic formations. Such installations are in operation in Switzerland, Sweden, and Germany.

Switzerland. At the Grimsel pass in the Swiss alps an underground rock laboratory has been established in granite. Since 1984, extensive research programs have been carried out at the Grimsel test site, including methods for nondestructive rock investigations, rock movement measurements, and various tests regarding rock mechanics and hydrogeologic measurements.

Sweden. At Stripa, an abandoned iron ore mine in central Sweden, an international project under the sponsorship of the OECD Nuclear Energy Agency is being conducted. The project is mainly concerned with investigation into the flow and chemistry of groundwater, the migration of radionuclides in granite, and the behavior of bentonite clay as a material for backfilling of cavities.

Germany. In Germany the former Asse salt mine near Remlingen/Wolfenbüttel has been used since 1965 to carry out comprehensive research and development programs. Within the scope of these programs, about 124 500 drums of low-level radioactive waste and about 1300 drums of intermediate-level waste had been disposed of by the end of 1978 [343]. Since then, research activities have continued without further storage. At present, the disposal of heat-generating waste is being investigated under international cooperation.

Information gathered in the different underground laboratories is exchanged so that it can be cross-checked. This information provides a valuable source of knowledge in further planning of deep-lying repositories.

5.5.4.3. Near-Surface Repositories

Short-lived radioactive waste can be sufficiently isolated from humans and the environment in engineered near-surface facilities. Shallow land burial is practiced on an industrial scale in Great Britain, Sweden, and especially, France.

Great Britain. Most of the low-level radioactive waste in Great Britain is disposed of at the Drigg site in Sellafield in Cumbria. The waste is disposed of in large trenches that are covered by soil and a waterproof layer. According to current planning, the disposal concept will be changed, and this type of waste will be disposed of in a deep-lying repository (200 – 1000-m depth) soon after the turn of the century [344].

Sweden. In Sweden the SFR repository is located 1 km offshore of the Forsmark nuclear power plant and consists of a series of caverns and silos ca. 60 m below the seabed. Access

to this repository is by road tunnels from land. The SFR repository is designed to hold about 90 000 m^3 of waste package. The first phase, which was completed in 1988 and has been in operation since then, includes one concrete-walled silo and four caverns. The remaining silos and caverns are scheduled for completion at the end of the century.

France. In France, short-lived radioactive waste is routinely disposed of in surface concrete facilities. The waste packages (concrete blocks consisting of steel drums in cylindrical concrete packaging or concrete boxes) are stacked in concrete trenches. The trenches are then backfilled with concrete and protected by waterproof covers against rainwater infiltration. Via a system of monitoring shafts, the state of this disposal facility can be checked, and any water that has infiltrated can be collected and treated. These techniques are used at the disposal site of La Manche, next to the La Hague reprocessing plant, where almost 400 000 m^3 of waste package has been disposed of since 1969. As the La Manche site is likely to be saturated at the beginning of the 1990s, a new site is being constructed at Soulaines, 200 km southeast of Paris. This 1 000 000-m^3 site is now in operation.

5.5.4.4. Underground Repositories

National programs and time scales for the planning and operation of deep-lying repositories vary widely because of the different economic, political, and technical circumstances, and the different boundary conditions to be considered in different countries. Important technical considerations are the type and volume of radioactive waste to be disposed of, the availability of disposal options, and the need for further research and development prior to detailed planning of repositories.

Many countries are still involved in the discussion of disposal concepts, the performance of research and development programs, and preliminary planning work for disposal facilities in geologic formations. However, progress has been made (e.g., in the United States and Germany).

United States. In the United States, a strict separation exists between radioactive waste originating from commercial applications and that originating from the defense program. The Waste Isolation Pilot Plant (WIPP), which has been constructed in a bedded salt formation near Carlsbad, New Mexico, is designed to receive transuranic radioactive waste from the U.S. defense program [342] [transuranic waste is material containing α- emitters in concentrations $> 3.7 \times 10^3$ Bq/g (> 100 nCi/g)]. A broad research and development program is also being carried out at WIPP.

Germany. In Germany, three sites are being considered for repositories:

The Abandoned Konrad Iron Ore Mine. Here, radioactive waste with negligible heat generation is to be disposed of, i.e., the host rock temperature must not be increased by more than 3 K on an average by the waste package [345].

The Konrad repository will be operated with two shafts. The sequence of strata in the coral oolite at a depth of 800 – 1300 m is available as emplacement formation. A repository cavity of about 1 000 000 m^3 will be excavated, allowing disposal of a waste package volume of ca. 650 000 m^3. At the end of a 40-year operational lifetime a total activity on the order of 10^{18} Bq is anticipated in this facility. The licensing procedure was started in August 1982. Operation of the Konrad repository is expected to begin in the mid-1990s.

Gorleben Salt Dome. The planned repository in the Gorleben salt dome is designed for the disposal of all kinds of solid and solidified radioactive waste, including the emplacement of heat-generating waste from reprocessing [330], [346]. According to plans the Gorleben repository will also operate with two shafts. The emplacement level is at a depth of about 870 m. The cumulative inventory of radioactive waste to be disposed of during an operational lifetime of nearly 70 years, is estimated to be on the order of 10^{21} Bq. At present, after comprehensive aboveground investigation, two exploratory shafts are being sunk. On completion of these shafts, the underground site investigation program will be initiated. The start of operation of the Gorleben repository is scheduled for the year 2008.

Morsleben Repository. In the former German Democratic Republic an abandoned salt mine in Morsleben near Helmstedt was selected

and reconstructed to serve as a repository for radioactive waste originating from the operation of nuclear power plants and from radioisotope production and application [347]. This facility began continuous operation in 1981. Liquid waste is solidified in situ in the disposal cavities by using lignite filter ash as the binder. Solid primary radioactive waste is disposed of in cavities without immobilization; 200-L drums of solid or solidified waste are stacked in disposal chambers.

After German unification in October 1990, a safety assessment on this repository is being performed on behalf of the BMU. The results of this assessment, among other things, will form a basis for decisions concerning the future use of the facility.

5.6. Safety Aspects in the Design of Reprocessing and Waste Treatment Plants

Legislation concerning the safety of reprocessing plants is similar worldwide. Thus safety requirements mandated by the German authorities are described in the following sections as an example. An overview of international legislation is given in Section 5.6.8.

5.6.1. Objectives of Protection

The construction and operation of a reprocessing plant, including facilities for waste treatment, require a permit according to § 7 of the German atomic law (Atomgesetz AtG) [348]. This permit can be granted only if in the design and operation of the plant, all precautions taken against damage represent the state of the art. According to the radiation protection decree (Strahlenschutzverordnung, StrlSchV) the technical design and operation must fulfill the following safety objectives [349]:

1) Any unnecessary exposure or contamination of persons, material goods, or the environment is to be avoided [349], (§ 28, Section 1, No. 1).
2) Any exposure or contamination of persons, material goods, or the environment– even if below stipulated limits–is to be kept as low as possible in terms of both the state of the art and all circumstances of the individual case ([349], § 28, Section 1, No. 2, § 44, § 45, § 49 – 56).
3) The planning of structural or other safety measures for design accidents, notwithstanding requirement (1), must be based at least on the malfunction planning values of § 28, Section 3 of the StrlSchV [349].

These stipulations result in the following basic *safety objectives*:

Enclosure, retention, and shielding of radioactive substances
Minimization and control of the discharge of radioactive substances via exhaust air and wastewater
Guarantee of nuclear subcriticality
Avoidance of fire and explosion or their effective control
Safe removal of radioactive decay heat or other waste heat
Protection against outside influences (plane crash, earthquake, chemical explosion, etc.)
Design compliance with all operational and maintenance requirements so as to avoid contamination and minimize the exposure of operating personnel to radiation
Provisions to be made with regard to closing and decommissioning of the plant

A comprehensive *safety concept* is necessary to fulfill these basic requirements. Important features of this safety concept are:

- Multishelled enclosure of radioactive substances (multiple barrier concept)
- Equipment with monitoring and safety facilities; redundancies for maintenance of important protective functions (i.e., more ready-to-use technical means must be available than are required to carry out scheduled functions)
- Redundancy of safety measures by using auxiliary devices that differ in type and mode of operation
- Safety-oriented quality-securing measures
- Precautions for safe operation, maintenance, and shutdown, safety in the case of malfunction, and reduction and elimination of the effects of malfunction
- Training of operating personnel and safety-oriented organization of the course of operations.

Additional requirements exist with regard to liability in case of damage [350], protection against action from third parties, and control of fissionable materials on the basis of international agreements [351].

5.6.2. Safety Through Multiple Barrier Enclosure

Safe operation and permanently effective enclosure of large amounts of radioactive material can be guaranteed by means of construction, control, and administrative measures [352]. Every future licensing procedure for a reprocessing plant must conform with safety criteria stipulated by the government. In nuclear technology, the multiple barrier principle is applied not only in nuclear power plants, but also in reprocessing plants and storage facilities for radioactive waste. Figure 131 shows the enclosure technology typical for a reprocessing plant. Chemical reactions occur in leaktight containers or apparatus that are interconnected by piping (*first barrier*). The process vessels and piping systems are enclosed in a hot cell, which is cladded with stainless steel to ensure that it is gastight. Radiation shielding is guaranteed by thick concrete walls (*second barrier*). The concrete cells are, in turn, located in a massive building, which takes over the function of the *third barrier* and is simultaneously designed as protection against outside influences.

Special attention must also be directed to the redundant protective effect of the barriers for all media piping employed in supply and removal. This is achieved by the use of double locks that are effective independently of each other, multiple filters, and quick shutting dampers. The rooms and hot cells are maintained at a stepwise decreasing subatmospheric pressure (lowest subatmospheric pressure inside the hot cells) to prevent the uncontrolled escape of radioactivity. Thus, air always flows from the outside into the process building, then into the cells, and finally either via the equipment or directly through a multiple filter system to the outlet air stack. Wastewater is passed first through storage tanks and purification facilitites, and only after careful examination is it allowed to enter the main drainage.

5.6.3. Safety Measures for Protection of Employees

The protection of employees against the dangers of radioactivity is guaranteed not only by closed, controlled, and monitored areas within the plant, but also by constant supervision with the help of personnel dosimetry and regular medical control. Exposure of operating personnel to radiation is always kept below permitted values by efficient shielding, by monitoring both the local dose rate and the personal dose, and by restricting accessibility to rooms with a higher level of radiation. Indeed, § 51 StrlSchV and § 44 StrlSchV must be adhered to at all times for internal and external supervisory areas, respectively. The local dose rate in places where maintenance work is carried out must be low enough that, given the frequency of maintenance, the exposure of personnel meets the demands of § 28 Section 1 StrlSchV [349], after all other contributions to exposure have been taken into account.

Remote-controlled systems or shielded rooms are provided for the removal, repair, and storage of contaminated components with a high local dose rate. Shielding measures, automation, or remote control is to be used for work involving unavoidably high local dose rates.

Special protective measures for personnel must be taken in the event of malfunctions that cannot be ruled out. In particular, immediate help must be available in case of incorporation or contamination by radioactive substances. For areas in which the danger of the incorporation of radioactive substances exists, activity concentrations in the air must be continuously measured. An alarm must be sounded if limits are exceeded. This in turn must be followed by suitable measures for the protection of the operating personnel (e.g., use of a respirator or immediate evacuation of the workspace).

Physical safety is increased further by regular medical examinations. According to § 62, Section 1 StrlSchV, the body dose or individual dose of persons who work in the control area must be determined.

5.6.4. Malfunction and Safety Analysis

An ingenious technical safety system allows all malfunctions that have any probability of occurring, to be controlled in such a way that the

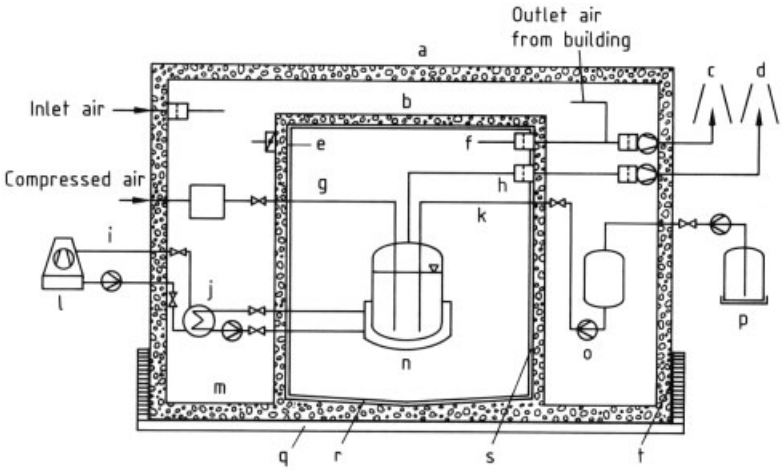

Figure 131. Schematic representation of the barriers in a reprocessing plant
a) Building; b) Hot cell; c) Outlet air stack; d) Off-gas stack; e) Overflow outlet; f) Cell outlet air/off-gas; g) Measuring/ agitation air; h) Container off-gas; i) Secondary cooling system; j) Primary cooling system; k) Chemicals input duct; l) Recooling unit; m) Operating compartment; n) Container with radioactive inventory; o) Batching tank; p) Chemicals supply; q) Contact plate; r) Tank surface with detectors; s) Cell cladding; t) External sealing

population and the operating personnel do not suffer any serious injury to health. Operational disorders and malfunctions that could occur in the plant must be examined in a *malfunction analysis*. Here, conditions in the plant during the course of operations must be analyzed systematically, and experience from comparable plants taken into account. The consequences of possible malfunctions should be examined carefully, and precautionary measures must be taken to control and contain the effects of damage. In the analysis of possible malfunctions, human error must be considered.

The following events must be taken into account as malfunctions in the design of a reprocessing plant:

- Disintegration of a fuel element with too short a decay time and thus a too high level of radioactivity
- Mechanical damage of vessels, equipment, facilities that serve to enclose radioactive substances, and safety facilities, damage due to crashing of loads
- Leakage
- Fire
- Explosion and exothermic chemical reactions
- Criticality events
- Failure of important cooling facilities that are important for safety
- Failure of energy and media supply
- Failure of off-gas and ventilation facilities
- Spontaneous breakdown of pressure vessels for media (i.e., water, gases, coolants, etc.) supply on the plant premises

All natural events that depend on the specific location of the plant, particularly earthquakes, must also be considered as design-determining malfunctions.

The *safety report* submitted in the licensing procedure serves to predict the time-dependent behavior of all barriers by the use of scientific methods. Here, among other things of special importance, changes in functioning of the hold-back mechanism of the barrier must be evaluated. Reliable bases for calculation must be obtained for the events and processes considered so that changes with time in the hold-back function of the barriers used, and thus in the release rates of radionuclides into the biosphere, can be derived quantitatively.

In the *licensing procedure*, the safety report must be displayed for public scrutiny. It serves as an essential document for both experts and authorities in the examination and granting or denying of a license. There are two differenct ways of conducting a safety analysis: the de-

terministic and the probabilistic methods. The most important elements and differences are presented in Figure 132. Currently, the deterministic method is being used almost exclusively for the licensing procedure. As in reactor safety research, an attempt is also being made in the case of reprocessing plants to supplement deterministic analyses with probabilistic methods in the assessment of the compatibility of the concepts involved.

5.6.5. Environmental Protection and Radiological Exposure

After being granted a license, every nuclear plant is subject to observation and supervision by state authorities during its entire lifetime. They must make sure that the operating company adheres to laws, as well as to other rules and regulations, and follows licensing conditions.

Representatives of the supervisory body or experts appointed by them can enter the plant at any time, inspect and examine it, and demand information from the licence holder. Both the supervisory body and the state licensing authority are, to the same extent, subject to supervison by the Federal Minister for Environment, Wildlife and Reactor Safety.

During normal operation, a minute fraction of the radioactive substances handled in the plant is released in a controlled way. This leads to a radiological burden on the plant surroundings, whose limits are stipulated in the radiation protection ordinance (Strahlenschutzverordnung) and the operating license. The calculated emissions, which as a rule are considerably higher than real emissions, result from plant-dependent emission conditions and location-specific meterological diffusion conditions.

For a reprocessing plant, the *radioecology* is determined exclusively by longerlived nuclides because of the preceding cooling time of the fuel elements. Calculation of the maximum possible radiation exposure is based on the following limiting conditions: operation of the plant for 50 years, emission at the approved rates of release, no reduction of activity other than through radioactive decay, and persons considered in the calculation must not only remain at the same location for 50 years (the place that exhibits the highest annual average activity) but also receive all their food from this location. Assumptions of this type are called conservative. They ensure that the result of the entire calculation rests on the safe side even when values used at individual points in this calculation are below those stated in various publications (e.g., the transfer factors for radionuclides).

The following exposure pathways are taken into consideration:

Exhaust air
- Direct irradiation from the exhaust air cloud: external β and γ radiation
- Direct irradiation from deposits on the ground
- Inhalation from the exhaust air cloud
- Deposits from the air on plants; consumption of milk, meat, and plant food

Wastewater
- Direct irradiation: swimming, boating, and sojourn at the waterfront
- Drinking water
- Fish consumption
- Use of river water for the cattle watering tank: consumption of milk, meat, and plant food

5.6.6. Radioactive Residual Substances and Waste

A plan must be prepared for the disposal of residual matter and waste that accumulate in a reprocessing plant. The concept must include a description of the type, nature, and amount of these materials; the concentrations of radioactive substances, stipulation of storage and conditioning; and evaluation of the safety aspects connected with this task (see Section 5.5).

Several measures are taken to achieve the final goal of radioactive waste treatment, namely, safe and long-term isolation from the biosphere. Thus, a process is required which yields minimum quantities of a waste that is as easy to treat as possible. At the same time, the various flows of waste material must be collected separately and subjected to different methods of treatment in accordance with the composition and nature of the materials [353]. Radioactive substances are separated from raw waste flows and enriched in a concentrate. The decontaminated flows are

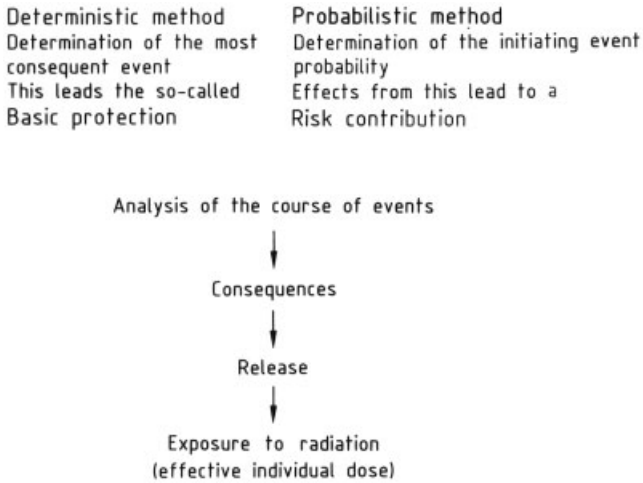

Figure 132. Two different approaches for safety analysis

then returned to the process or released into the environment. For the conditioning of radioactive waste, see Section 5.5.2.

5.6.7. Decommissioning and Dismantling

In the case of planned or forced discontinuation of a reprocessing plant, considerable amounts of radioactive substances can still be contained in the plant even after the removal of the operating materials. Therefore, a reprocessing plant should be designed and built in such a way that, in keeping with radiation protection requirements, it can be secured or totally dismantled after final shutdown. For this reason, aspects of a total decontamination in the future must be taken into consideration when designing a plant. Factors of importance include stipulation of room size; use of building and other materials; and arrangement of hot cells, components, piping, and operating facilities.

5.6.8. International Coordination of Safety Regulations

The operation of reprocessing plants, radioactive waste management facilities, and intermediate and final storage is subject to rather similar criteria in countries using nuclear energy. Most important is the radiation protection of humans. The health hazards associated with the handling of radioactive materials had been recognized in the early days of the nuclear age. Since 1928, in combination with international radiology congresses (held every 3 – 5 years), efforts have been made to found an international expert body with the aim of harmonizing radiation protection. The International Commission on Radiological Protection (ICRP), founded in 1950, developed from these efforts and now dominates the area of radiation protection worldwide. This organization issued two basic recommendations in 1953 and 1985 [354], which increasingly influenced legislation for the protection of employees–and hence the population –from ionizing radiation [355]. Noteworthy is the high degree of international agreement on the basic problem of radiation protection and its biological principles [356]. This agreement is independent of social system; i.e., technical safety criteria for nuclear technology are similar worldwide.

Recommendations of the ICRP are based on expert judgement; radiation protection should be controlled on the basis of joint scientific principles and joint formulated protection aims within certain limits. The EC incorporated these recommendations into its basic norms for the protection of employees and the population, and developed regulations that are mandatory for all member states. The EC countries are also at liberty to issue more rigorous regulations for radiation protection, such as those issued in the German Radiation Safety Ordinance (Strahlen-

schutzverordnung). In countries using nuclear energy, legislation concerning radiation protection leads to the interaction of official and unofficial bodies at regional, national, and international levels with contributions from scientific societies, professional assemblies, administrative bodies, ministry offices, and professional experts from governmental consultant committees. This intermingling is also observed at an international level with the IRCP as highest authority.

The United Nations has also set up a commission of experts from 20 member states [United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR)] to control radiation exposure and describe its effects. The WHO is also active in radiation protection. The International Atomic Energy Agency (IAEA) in Vienna organizes international symposia on the safe use of nuclear energy and publishes regulations for the operation of nuclear power plants. Finally, the work of the Nuclear Energy Agency of the Organization for Economic Cooperation and Development (OECD/NEA), located in Paris, is also available as reference material.

The effect of all these influences leads to laws, guidelines, standards, and regulations that are virtually identical in countries with nuclear energy programs. The outcome of these regulators is always an atomic energy law for the safe use of nuclear energy. Second in rank are laws concerning the handling of radioactive waste.

Examples are Waste Policy Act, Fed. Reg. 10 CFR Part 960, 1982 and Amendments Act of 1987; Regles Fondamentales de Surete, France, 1984; Projekt Gewähr, Switzerland, 1985; Law on activities in the nuclear field, Sweden, 1984; Waste Disposal Policy, Cmnd 8607, United Kingdom, July 1982.

For occupational health the following rules hold [357]:

1) No work is allowed unless a net benefit can be demonstrated.
2) Radiation exposure should be kept as low as possible with regard to scientific and social factors.
3) The dose equivalent for a single person should not exceed the recommended limiting values.

As a consequence, virtually all countries that build or operate reprocessing plants employ the same industrial and administrative safety regulations. Listing specific regulations is beyond the scope of this article.

In addition, official hearings are obligatory in nearly every country. Here, the population has the opportunity to express doubts and concerns to state regulatory authorities. Examples are the Windscale hearing (held in 1977) concerning the construction of the THORP reprocessing plant in England, and the energy policy debate on the expansion of a reprocessing plant at La Hague. This debate took place in the French Parliament in October 1981, on the basis of the so-called Castaing Reports. *Reprocessing plants* are in operation in the United Kingdom, France, Japan, China, the former Soviet Union, the United States, and to a lesser extent, in India. Small reprocessing plants in Germany, Belgium and Italy were closed down years ago.

Laws and safety regulations governing *installations used for waste disposal* are of greater public interest than those concerning reprocessing plants. The basic regulations are similar in every country [358]. The respective national atomic energy laws define the limiting framework sanctions for nuclear plants, their closure, and the disposal of radioactive waste. Differences exist with regard to responsibility for the disposal of waste material. In most countries the producer is responsible for the safe disposal of radioactive waste, but the state always checks that disposal has been carried out correctly. In several countries (e.g., Germany), radioactive waste disposal occurs under state responsibility. In some countries (e.g., Sweden, Switzerland, Germany) the use of nuclear energy depends on proof that permanent and safe disposal of the radioactive waste can be achieved. Other countries have less stringent regulations; normally, appropriate intermediate storage of the spent fuel elements or the radioactive waste is sufficient.

The following societies and organizations in various countries are responsible for the disposal of radioactive waste:

Belgium: ONDRAF, responsible for total waste management, founded in 1981.

Federal Republic of Germany: Bundesamt für Strahlenschutz, Salzgitter, responsible for final waste disposal. The "Deutsche

Gesellschaft zum Bau und Betrieb von Endlagern für Abfallstoffe" (DBE), Peine, is engaged as a so-called third party in the installation of federal facilities for the long term storage and final disposal of radioactive waste.
France: ANDRA, responsible for total waste management, founded in 1979.
United Kingdom: NIREX, responsible for total waste management, founded in 1982.
India: Atomic Energy Authority, state controlled.
Italy: NUCLECO, responsible for total waste management, founded in 1981.
Japan: Power Reactor and Nuclear Fuel Development Corporation (PNC), responsible for all aspects of fuel cycle.
Canada: Atomic Energy Authority (AECL), state controlled.
The Netherlands: COVRA, responsible for total waste management, founded in 1982.
Sweden: Swedish Nuclear Fuel Supply Company (SKBF), responsible for total waste management; KBS, a subsidiary of SKBE, is responsible for all R & D work.
Switzerland: NAGRA, responsible for total waste management, founded in 1972.
Spain: Junta de Energia Nuclear (JEN), state controlled; total waste management was transferred to a state-owned society, ENRESA, in 1985.
The former Soviet Union: Use of Atomic Energy, state committee.
United States: Department of Energy, with industrial subcontractors, based on the Radioactive Waste Policy Act of 1982.

6. Nuclear Fusion

Fusion of light nuclei is by far the most important process of energy release in the universe. The sun and most of the other stars are powered by the fusion of four hydrogen nuclei into one nucleus of helium, which was named after the sun (Greek: *helios*). Every such process releases an energy of 26.72 MeV (6.7 MeV per nucleon). Moreover, fusion processes are responsible for the generation of elements in the stars. This is obvious from the dependence of the nuclear binding energy per nucleon on the atomic mass number (see Fig. 133). Fusion energy is released in the fusion of light elements, while fission energy can be obtained from the fission of very heavy elements. Nuclei of medium atomic mass such as iron and nickel (Ni) have the lowest mass per nucleon (i. e., the highest binding energy) and hence are stable elements, because every rearrangement of their nuclei leads to a state of higher energy.

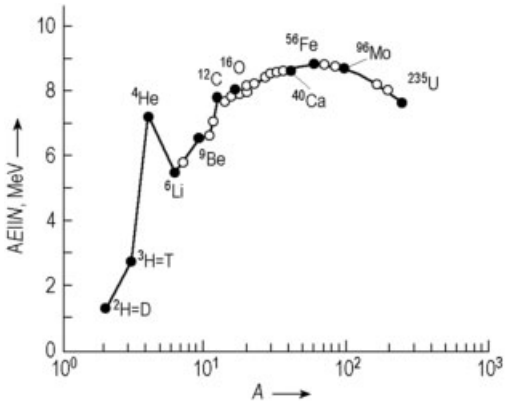

Figure 133. Nuclear binding energy per nucleon $\Delta E/N$ as function of the nuclear mass number A

On earth, up to now fusion energy has only been released in an uncontrolled way by the test explosion of hydrogen bombs. The fusion reactor based on the controlled operation of thermonuclear fusion, however, is a very important option for future energy supply, since the fuel is practically inexhaustible and this new energy source is environmentally acceptable. The enormous research efforts into controlled thermonuclear fusion, which in the early 1950s began in the United States, the United Kingdom and the former Soviet Union, aim at the development of a reactor that produces net energy from the fusion of hydrogen isotopes in a high temperature plasma. Because of the extremely difficult physics and technology problems which still have to be solved, it is expected that it will take more than half a century until this commercially attractive and environmentally acceptable fusion reactor will be available.

6.1. Fusion Physics

Normally, positively charged nuclei interact by their repulsive Coulomb forces unless their rela-

tive velocity is high enough to tunnel through the Coulomb barrier. Hence, fusion reactions occur practically only between light nuclei. Among the nearly 80 possible fusion reactions the following are most important:

$$D + D \longrightarrow {}^{3}_{2}He + n + 3.267\,MeV$$

$$D + D \longrightarrow T + p + 4.032\,MeV$$

$$D + {}^{3}_{2}He \longrightarrow {}^{4}_{2}He + p + 18.351\,MeV$$

$$D + T \longrightarrow {}^{4}_{2}He + n + 17.586\,MeV$$

$$T + T \longrightarrow {}^{4}_{2}He + 2n + 11.327\,MeV$$

$$p + {}^{11}_{5}B \longrightarrow 3\,{}^{4}_{2}He + 8.664\,MeV$$

$$p + {}^{6}_{3}Li \longrightarrow {}^{3}_{2}He + {}^{4}_{2}He + 3.890\,MeV$$

where p denotes proton, n neutron, D deuterium, and T tritium. The energy, which is released as kinetic energy of the reaction products, is also indicated, with 1 MeV = $1.6022 \cdot 10^{-13}$ J. This energy release – normalized to the mass involved – is more than six orders of magnitude higher than that of typical chemical reactions such as carbon oxidation in fossil fuels. This reflects the fact that nature – in releasing thermonuclear fusion energy in the stars – has chosen the process with the minimum mass consumption.

For thermonuclear fusion reactors the D – T reaction is by far the most attractive one. It has the highest fusion cross-section (with a maximum at relatively low energy) and also a very high energy release per reaction. One fifth of the energy is carried away by ${}^{4}_{2}He$ (alpha particle), while 80 % is released as kinetic energy of the neutron. Hence, with this reaction the highest power densities in a fusion reactor can be achieved (see Fig. 134).

Although the resources of deuterium (concentration of 0.015 % in natural hydrogen) are nearly inexhaustible and geographically equally available, tritium does not naturally occur in sufficient quantities, because it is radioactive (half-life 12.35 a). Using the fusion neutrons from the D – T reaction the tritium required is gained in the reactor breeding reactions with lithium:

$${}^{6}_{3}Li + n \longrightarrow T + {}^{4}_{2}He + 4.8\,MeV$$

$${}^{7}_{3}Li + n \longrightarrow T + {}^{4}_{2}He + n' - 2.47\,MeV$$

Natural lithium contains 7.42 % ^{6}Li and 92.58 % ^{7}Li. Lithium is, as deuterium, geographically equally available, also in the oceans. Hence, deuterium and lithium are the actual fuels for the fusion reactor, tritium is only internally produced in the reactor system. The first tritium inventory, however, has to be supplied from fission reactors.

Because of the possible safety and environmental risks of the D – T reaction, which are related to the use of tritium as radioactive fuel component and the activation of structural materials by the fast fusion neutrons, the other fusion reactions listed before are considered, too. As shown in Fig. 134, the other fusion reactions have a much lower fusion power density, and their control involves higher physical and technological problems than those encountered with the D – T reaction.

The Figure 134 moreover, shows that high fusion powers are obtained at high temperature (or energy) of the reaction partners, which is in the range of 15 keV (or about 170×10^{6} K since 1 eV corresponds to 11 604.5 K). At these high temperatures the matter is in the plasma state, i.e. all particles are charged. At sufficiently high plasma densities, the Coulomb collisions between these charged reaction partners are so frequent, that the collisions have a Maxwellian velocity and energy distribution, hence a temperature. Since the fusion cross-section strongly increases with the temperature, "thermonuclear fusion" occurs if the temperature of the plasma has been made sufficiently high by supplying a large quantity of energy.

For a selfsustained "thermonuclear burn" the unavoidable energy losses due to radiation and heat conduction must be balanced by part of the fusion power in order to sustain temperature and density. Bremsstrahlung losses are such unavoidable radiation losses, which are related to the deceleration of electrons by Coulomb interaction with other charged particles in the plasma. Other losses are caused by line and recombination radiation from impurities, by heat conduction, or by plasma turbulence and instabilities. Since for the D – T reaction one fifth of the fusion power is released in the (positively charged) alpha particles, while four fifth are carried away by the neutrons, thermonuclear burn occurs if the *ignition condition* is fulfilled. This criterion

requires power balance between the heating of the plasma by the alpha particles formed and all unavoidable power losses. If it is assumed that the plasma is not deluted by impurities, hence is a pure D – T plasma, the ignition criterion is simply expressed by the product of the plasma pressure p and the energy confinement time τ_E to be greater than a numerical value, which only depends on the temperature T of the plasma. For the D – T reaction a minimum is reached at a temperature of about 15 keV, such that the D – T ignition criterion reads $p \cdot \tau_E \geq 1$ MPa · s.

The energy confinement time τ_E is the characteristic energy loss time constant, which qualifies the thermal insulation of the plasma confinement. It is simply given by the ratio of the plasma energy density and the heating power necessary to sustain the plasma parameters.

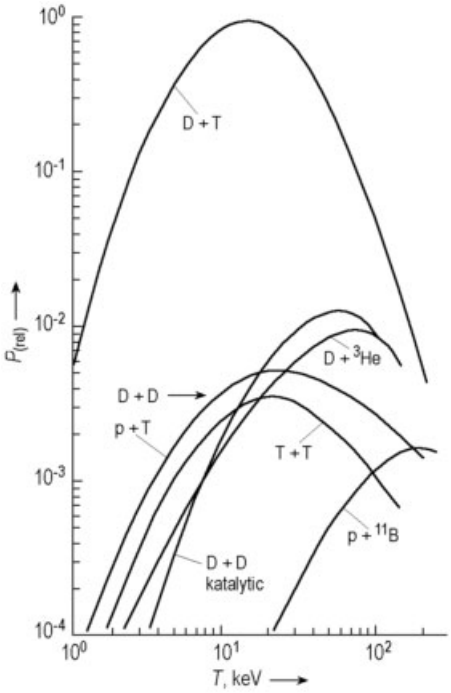

Figure 134. Power densities P of different fusion reactions at pressure limitation versus plasma temperature T, normalized to the maximum of the D – T reaction.

Although the energy of the fusion neutron is not taken into account for the energy balance underlying the D – T ignition criterion, the main aim of worldwide thermonuclear fusion research is to fulfill this criterion. Since the plasma pressure p is given by the plasma density n and the temperature T_i and T_e of ions and electrons, respectively, through $p = n \cdot k_B (T_i + T_e)$, the ignition criterion gives the minimum conditions for the three key parameters of the plasma – the ion temperature T_i, the plasma density n and the energy confinement time τ_E – necessary for a self-sustained "thermonuclear burn" of the plasma. Typical plasma densities are in the range of $n = 10^{20}$ m^{-3}, and temperatures in excess of 15 keV have been reached, hence energy confinement times of more than 2 s are necessary. Although the critical value of each of these parameters was reached and usually substantially exceeded separately in different experiments, the simultaneous fulfillment of all conditions has not been achieved up to now. This is the aim of worldwide investigations in close international cooperation.

6.2. Plasma Confinement and Status of Fusion Research

Two key problems must be solved to achieve net power from any fusion system. A sufficiently dense and pure D – T plasma has to be heated to ≥ 10 keV ($> 10^8$ K) and be confined against energy losses for a certain minimum time. Two very different concepts exist for plasma confinement: (1) *magnetic confinement*, which has been studied from the beginning of fusion research in different configurations, and (2) *inertial confinement* (laser pellet fusion or ion beam pellet fusion).

6.2.1. Magnetic Plasma Confinement

A hot, highly ionized plasma is a good electrical conductor. A plasma with an electron temperature of 10 keV has an electrical conductivity about 20 times as high as that of metallic copper. Correspondingly, plasmas can carry strong currents and can be influenced by magnetic fields. In a microscopic view, charged particles (electrons and ions) gyrate around magnetic field lines. Therefore, they are essentially bound to the field lines but they can freely flow along them. In the linear or "open" magnet configurations, field lines leave the plasma and the plasma particles can escape. Fusion approaches with linear configurations such as mirror machines and

linear pinches suffer from such end losses. In contrast, when the plasma forms a toroid and all the field lines stay within the plasma end losses can be avoided. In the "closed" toroidal machines the magnetic field lines generate an assembly of nested toroidal magnetic structures that can contain the plasma. The poloidal (located around the minor torus circumference) and the toroidal magnetic field components generating these configurations must be tailored to guarantee plasma equilibrium (balance of forces exerted by the magnetic fields and by the plasma) and at least gross plasma stability. In practice, often auxiliary outer magnetic fields or currents, including also a shaping of the plasma cross-section, are applied to achieve these objectives. The basic magnetic field components in toroidal devices can be produced, either *solely by currents in conductors* (magnetic coils) *surrounding* the plasma or with a strong additional *current component in the plasma*. Machines of the first type are the Stellarators, those of the second type are the Tokamaks and Reversed Field Pinches. In a long-lasting selection process that has gone on since the beginning of fusion research 35 years ago, toroidal confinement devices have been essentially successful. The Tokamak particularly is considered the most promising contender for a fusion reactor.

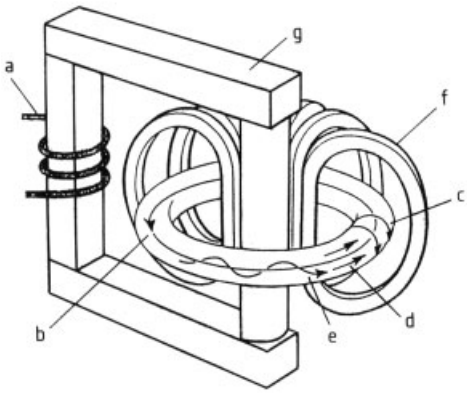

Figure 135. Schematic of the Tokamak principle
The poloidal magnetic field is produced by the toroidal plasma current induced in the plasma by a transformer. The toroidal field is produced by D-shaped coils. The coils for the vertical magnetic field are not shown.
a) Transformer winding (primary circuit); b) Plasma current (secondary circuit); c) Poloidal magnetic field; d) Toroidal magnetic field; e) Resultant helical field (twist exaggerated); f) Toroidal field coils; g) Ion transformer core

In toroidal devices the density of the plasma is rather low (ca. $10^{20}\,\mathrm{m}^{-3}$), which is about 10^5 times lower than that of a gas under atmospheric pressure. At the temperature envisaged ($\geq 10\,\mathrm{keV}$) plasma pressures reach several atmospheres. They must be much smaller (5 – 10 %) than the pressure of the confining magnetic field to achieve magnetohydrodynamic stability. So the ultimate strength of the magnetic field that can be generated in superconducting coils ($\leq 12\,\mathrm{T}$) eventually determines the upper limits of plasma density. According to the ignition criterion, τ_E must typically be a few seconds. Because the fusion power densities in such dilute plasmas are low (some watts per cubic centimeter), huge plasma volumes have to be used in toroidal fusion reactors.

Tokamak. The Tokamak is the most advanced fusion system with respect to plasma physics. The Tokamak principle was developed initially at the Kurchatov Institute in Moscow in the 1960s. Because of the spectacular results, particularly the achievement of high plasma temperatures, Tokamaks spread all over the world in the 1970s. The very successful, and currently largest, fusion system JET (Joint European Torus) in Culham, UK is a Tokamak. A schematic of the Tokamak configuration is shown in Figure 135. The Tokamak uses magnetic fields that are a combination of a strong toroidal field with a weaker poloidal field (field lines circling around the torus core). The *toroidal field* of several teslas is produced by magnetic coils placed outside the toroidal vacuum vessel, whereas the *poloidal field* is generated by a strong current transiently induced in the plasma by the action of a transformer. This current also produces an initial heating of the plasma. The field line pattern resulting from the combination of both field components is a helix with a low pitch angle. A third component, a *vertical magnetic field*, produced in a set of poloidal coils, is necessary to keep the plasma in a well-centered equilibrium position and to shape the cross-section of the plasma for optimum confinement properties. The transformer principle does not allow a continuous stationary plasma current. Therefore, all Tokamaks work in a pulsed mode. For a reactor, steady-state operation appears to be necessary, mainly for technical and economic reasons (simplification of engineering design, improvement of reactor availability, avoidance

of cyclic stresses and of intermediate energy storage). Experiments are in progress to drive a stationary current by means of electromagnetic waves or neutral particle injection.

The current ohmic heating of the Tokamak plasma becomes inefficient at increasing temperature T because the electrical conductivity of the plasma is proportional to $T^{3/2}$. To approach ignition temperatures, additional heating methods such as neutral deuterium beam injection or injection of radio-frequency waves into the plasma are used in Tokamak experiments.

Connected with the energy confinement time, the existence of impurities in the plasma is one of the most serious problems in magnetic confinement devices. Even when sufficient gross plasma confinement is achieved, impurities will be released at the vacuum wall by particles escaping from the plasma by diffusion. Because these impurities are mainly neutral atoms, they can penetrate the plasma. As a result, radiation losses are dramatically increased and energy confinement is thus impaired. Various techniques are used to reduce the impurity influx. In normal Tokamaks, diffusive plasma extension is cut off by a limiter, which may be a material boundary or a magnetic limiter. In more advanced *divertor Tokamaks* the plasma edge sheath is sucked up by a magnetic system, the divertor. The plasma particles and impurities are magnetically guided to an external divertor chamber and deposited there on target plates. Such divertors are necessary anyway in toroidal reactors to remove burned fuel.

Stellarator. In the Stellarator, helical magnetic fields for confinement are produced solely by *external coils*. Because of the lack of both a net toroidal current in the plasma and a transformer, the Stellarator configuration is naturally suited for steady-state operation. Topologically somewhat different magnetic configurations of the Stellarator class can be generated by particular arrangements of coil sets, leading to Torsatrons, Heliotrons, and different classes of Stellarators. In the *advanced Stellarator concept* (Fig. 136), nonplanar coils are used in a toroidal system of large aspect ratio, offering the advantage of a modular system that is more accessible than the Tokamak system. Heating in the Stellarator is performed with nonohmic methods. Otherwise, with respect to plasma properties and impurity problems, the Stellarator has much in common with the Tokamak, to which it may be a promising alternative.

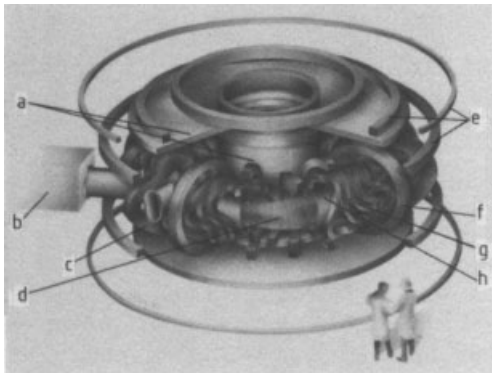

Figure 136. Schematic of the Stellarator W VII – AS (Advanced Stellarator)
a) Support frame; b) Neutral beam injector; c) Extra coil allowing neutral injection; d) Plasma; e) Poloidal field coils; f) Toroidal field coils; g) Twisted coils; h) Vacuum vessel

6.2.1.1. Status of Magnetic Confinement Fusion

Efforts to develop fusion as a future source of energy are underway worldwide. The bulk of the effort is concentrated in four large programs of about equal size: in the EC, in Japan, in the former Soviet Union, and in the United States. In *Western Europe*, research and development on controlled fusion is coordinated by Euratom. The program is shared by the fusion centers of the EC and by Sweden and Switzerland. The program concentrates on the development of the Tokamak system including development of the technological components necessary for a future reactor. The largest single experiment in the European program (and in the world) is the Joint European Torus (JET) at Culham, United Kingdom. Alternative systems, pursued with somewhat lower priority, are the Stellarator (MPI, Garching, Germany) and the Reversed Field Pinch (Padua, Italy). Supplementary research is carried out on rather large Tokamak devices in national laboratories of the EC states (Asdex Upgrade, Garching, Germany; Tore-Supra, Cadarache, France).

Similar programs in fusion research with a concentration on the Tokamak are in progress by the other three parties and have led to large Tokamak devices: TFTR (United States), JT 60

Table 66. Plasma parameters reached in JET, compared with those required in a Tokamak fusion reactor

	η_{eo}, 10^{20} m^{-3}	T_{io}, keV	τ_E, s	$\eta_{io} T_{io} \tau_E$, 10^{21} keV · s/m^3	$\langle\beta\rangle$, ∗ %
Reached in JET	0.4	22	1.1	0.9	2
Needed for a reactor	2 – 3	10 – 20	1 – 2	5	∼6

∗ $\langle\beta\rangle$, the ratio of the plasma pressure to the pressure of the confining magnetic field, characterizes the utilization of magnetic field energy.

(Japan), and T 15 (Russia). More than ten Stellarator experiments, mostly small scale, are in operation worldwide, and five Stellarators are under construction.

The currently operating Tokamaks including JET and TFTR are still pure plasma experiments, with hydrogen or deuterium. The components necessary for a fusion reactor, i.e., blanket, shield, and tritium systems, have not yet been installed. Technological developments only exist with respect to the magnets. Tokamaks with superconducting magnets are operating in France (Tore-Supra) and in Russia (T 7 and T 15).

The status of Tokamak research is well represented by JET (Fig. 137), which became operational in 1983. The goal was to study plasma behavior–heating, stability, and confinement – in a parameter region close to reactor conditions and at plasma dimensions approaching those of a reactor: (plasma diameter horizontally 2.5 m, vertically 4 m; major torus radius 6 m). The JET has routinely operated with currents above 5 MA, which can be maintained for 2 s. In 1986, a plasma ion temperature of 100×10^6 K and an energy confinement time of 0.5 s could be reached at a heating power of 8 MW. These results were remarkable because they indicated progress in thermal insulation of the plasma. In earlier experiments the confinement times were lower and decreased with increasing heating power. These results were made possible by modification of the magnetic structure at the plasma edge, thus reaching a favorable "high-confinement regime" (H-regime). The regime was discovered in the large divertor Tokamak Asdex at Garching (Germany). However, in JET, as in other Tokamaks, the confinement is still dominated by anomalous heat conduction, mainly due to the plasma electrons. Despite long-lasting and extensive research, a basic understanding of this obviously complex phenomenon is not yet available.

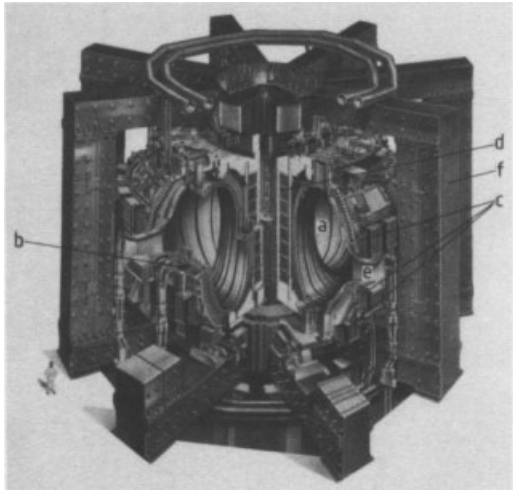

Figure 137. JET (Joint European Torus)
a) Vacuum vessel; b) Toroidal field coil; c) Outer poloidal field coils; d) Inner poloidal field coils; e) Ports for neutral beam injection; f) Transformer limb

Over the last years (1987 – 1997), additional heating with *neutral beam injection* up to 21 MW and with *ion cyclotron resonance* up to 11 MW has been applied. At a total input power of 30 MW, ion temperatures of 32 keV and energy confinement times of 1.1 s have been obtained. With respect to impurity control, great progress was achieved by covering the inside walls of the JET vacuum vessel with beryllium instead of carbon, as done before. Radiation losses were reduced, and the neutron density could be increased almost to reactor-relevant values. In Table 66 the plasma data reached in single discharges are compared with those needed for a fusion reactor. In discharges of the H-regime, for very short times (ca. 0.1 s), record values of the fusion reaction product $\eta_i T_i \tau_E$ of 0.9×10^{21} keV · s/m^3 and a power amplification factor Q of 0.9 (extrapolated for D – T operation) were attained (Q is the ratio of fusion power re-

leased to total heating power; $Q = 1$ means break even and $Q = \infty$ ignition in a D – T reactor). The limitation in time was caused by an unavoidable impurity backflow from the divertor, which should be overcome in the future (1992) by means of additional pumping in the divertor. The (1990) status at JET and at TFTR is shown in Figure 138 ($\eta_i T_i \tau_E / T_i$ diagram), which also illustrates the plasma physics path toward controlled fusion, beginning with the first Tokamak experiments on the T 3 device in the former Soviet Union.

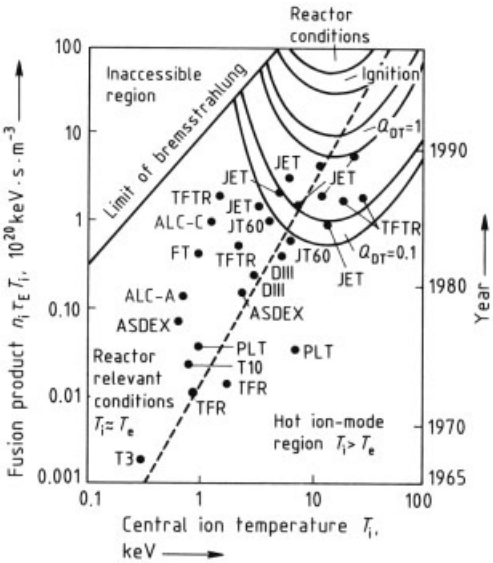

Figure 138. Performance of Tokamaks (1965 – 1990)
The fusion product $n_i \tau_E T_i$ is the product of the central ion density, the confinement time and the central ion temperature reached in single experiments.
JET: Culham, United Kingdom; ALC-A, ALC-C: Boston, United States; ASDEX: Garching, Germany; D III: San Diego, United States; JT 60: Tokyo, Japan; FT: PLT: Princeton, United States; T 3, T 10: Moscow, Former Soviet Union; TFR: Fontenay, France; TFTR: Princeton, United States
Q_{DT} is the ratio of the peak fusion power to the total input power, translated for a D – T mixture. (The experiments were performed in pure deuterium.)

The latest JET results indicate that ignition can be achieved without further complications in a somewhat larger post-JET device such as the Next European Torus (NET) or International Thermonuclear Experimental Reactor (ITER) (see Section 6.2.1.2). Nevertheless, a concluding experimental stage is foreseen on JET in which, for the first time in a large facility, deuterium – tritium fuel has been used, thus initiating a decisive phase for fusion research in the magnetic confinement approach.

6.2.1.2. Further Development of Magnetic Confinement Toward a Fusion Reactor

Based on favorable results with JET, TFTR, and other large Tokamak devices, the final objective is an energy-producing Demonstration Reactor (DEMO) with only one intermediate experimental facility. In the framework of the Euratom program, another large Tokamak facility NET (Next European Torus) was foreseen. Design and construction work for NET has been going on since 1983. In the meantime, stimulated by the Reagan – Gorbachev initiative (Geneva 1985), a joint venture called the International Thermonuclear Experimental Reactor was initiated by the International Atomic Energy Agency in 1987, and subsequently, conceptual design activities began under the auspices of the IAEA (1988). The ITER takes advantage of NET planning as well as of the earlier INTOR concept of the IAEA. Partners in the ITER program are the EC, the United States, Japan, and Russia.

The ITER is expected to confirm fully the scientific feasibility – and to address the technological feasibility – of fusion power via the magnetic fusion approach. Consequently, the machinery must be designed for controlled ignition and extended burn of deuterium – tritium plasmas. It must also demonstrate and perform integrated testing of components required to utilize fusion power for practical purposes. Among other items, ITER will be equipped with superconducting magnets; refueling and ash removal systems; test blankets; and tritium-handling systems, and should already generate 1 GW of fusion power. After definition of the concept, the design phase began in 1989 at the MPI Garching with international teams of scientists and engineers. About 13 years of design and construction work is expected until ITER becomes operational.

The heart of the ITER facility is the Tokamak device represented in Figure 139. Its principal parameters can be summarized as follows:

Plasma major radius R	6.0 m
Plasma half-width at midplane a	2.15 m
Toroidal field on axis B_0	4.85 T
Nominal maximum plasma current I_P	22 MA
Nominal fusion power P_f	1000 MW

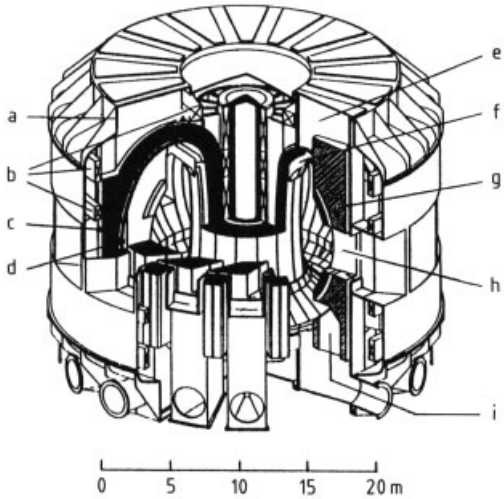

Figure 139. Basic elements of the ITER Tokamak
a) Cryostat; b) Poloidal field coils; c) Toroidal field coils; d) Vacuum vessel; e) Vertical port; f) Divertor; g) Blanket and shield; h) Horizontal port; i) Vacuum pump duct

The design features and parameters are determined by the objectives, such as attainment of ignition and extended burn, plasma stability, impurity control, and current drive, and also by safety considerations (particularly in tritium handling) and cost minimization. A preliminary estimate of the capital cost in 1989 amounted to $\$4.9\times10^9$.

6.2.2. Inertial Confinement Fusion

Since ca. 1965, another approach to controlled fusion power has been pursued, particularly in the United States, Japan, and France. In inertial confinement fusion (ICF) or *pellet fusion*, energy is produced in a series of microexplosions. The basic idea is the rapid implosion of a solid deuterium – tritium fuel pellet by irradiation with laser or ion beams, so that a hot, very dense core develops that fuses before the pellet disintegrates. Generally, magnetic confinement fields are not needed.

For the implosion, two concepts are presently being investigated. In *direct drive*, solid spherical pellets of complicated shell structure (see Fig. 140) are irradiated directly by very intense laser or ion beams. In *indirect drive*, laser light is focused into a hollow shell, where it is converted into X-rays (hohlraum radiation). This radiation, in turn, compresses the pellet which is centrally located within the shell. In both cases a rapid ablation of the outer pellet surface develops. The repulsion of the rapidly expanding plasma corona should first compress the pellet core to about $10^3 - 10^4$ times solid-state density and then adiabatically heat it to initiate an outgoing, self-sustaining fusion burn wave. Provided a sufficient fraction of the fuel is burned before the pellet cools down, the energy gain (ratio of nuclear energy output to driver energy output) may make up for the very low efficiencies of the driver systems. The fuel pellets must be very small to keep the explosive yield within technologically acceptable limits, because the reactor chamber and the blanket must be able to withstand the stress resulting from repetitive pellet explosions.

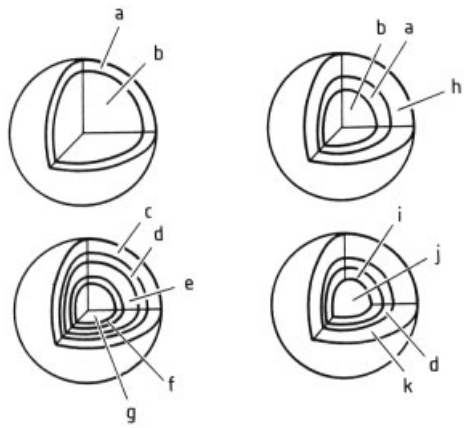

Figure 140. Different designs of inertial fusion pellets
a) SIO pusher; b) D – T gas fuel; c) Beryllium ablator; d) TaCHO pusher; e) Plastic foam; f) Tungsten pusher; g) Liquid D – T fuel; h) Teflon ablator; i) Frozen D – T fuel; j) Void; k) LiH ablator

Laser Fusion. In the current, most advanced laser fusion experiments, neodymium – glass lasers are usually applied. However, their emission at a wavelength of 1.05 μm is not well suited to pellet compression. Thus their infrared light is

generally converted to 0.53 or 0.35 μm by crystal converters, to improve the laser – target coupling and avoid the generation of fast electrons that would cause unfavorable preheating of the pellet core. The largest laser facilities are NOVA (100 kJ in laser pulses at 1.05 μm, Livermore, United States) GEKKO (50 kJ, Osaka, Japan), and PHOEBUS (20 kJ, Limeil, France). A number of medium- and small-sized facilities in the kilojoule to 100-J range also exist (United States, China, UK, France, Germany, Italy).

Many serious problems must be overcome: laser power of tens of terawatts must be focused to submillimeter dimensions to reach peak power densities of $> 10^{16}$ W/cm^2 at the pellet surface. Furthermore, irradiation has to be with spherical symmetry and high uniformity, otherwise instabilities causing disturbances of the pellet surface would impair the symmetry of the compression. In experiments with direct drive, up to 24 or even 60 laser beams, distributed over the pellet surface, have been used. The distribution is achieved by complicated beam splitting and aiming systems. Optical smoothing methods are also successfully applied.

Considerable compression factors up to 100 – 200 times liquid density have been obtained in both drive methods by using multilayered capsules filled with D – T gas. In a high-temperature mode using direct drive, T_i of about 10 keV and high fusion neutrons, yields of 2×10^{13}, but only low compression to liquid density, were attained in NOVA and GEKKO. For indirect drive, the best performances were obtained with NOVA leading to ion temperatures of 1.9 – 2.3 keV, about 100 times the liquid density, and a deduced $\eta\tau$ of $(2-4)\times10^{20}$ s/m^3. The $\eta\tau$ is comparable with that attained in advanced Tokamaks, but in ICF, higher values must be reached, mainly because of the poor driver efficiencies. However, Nd – glass lasers are not well suited for energy production with ICF. Their efficiency, in the 0.1 % range, is too low. New lasers with short pulse duration ($< 10^{-9}$ s) and high efficiency (≈ 10 %) must be developed. A leading candidate for future use is the KrF laser, which should reach efficiencies of 2 – 5 %.

Ion Beam Drivers. Light and heavy ion beams are also being considered as drivers. Multimegajoule pulses of light ions can be produced rather easily, and the efficiency can be made sufficiently high (15 – 30 %), but very high voltages (multimegavolts) are required. Beam transport and beam focusing on the pellet will be difficult, particularly because they must be performed during repeated thermonuclear blasts in future reactors.

The *light-ion technologies* have been developed mainly at Sandia (United States) and Osaka (Japan). In the PPFA-II device at Sandia the use of megajoule pulses of lithium ions for ICF is being developed. Significant advances have been achieved, but the progress is relatively slow. Major problems are the beam transport and the beam focusing onto the pellet surface.

Heavy ion beams as drivers may have better prospects. They have a shorter stopping range in the pellet and so the beam – target coupling should be more efficient. Furthermore, the efficiency of heavy ion accelerators could be made relatively high (about 25 %). Research on heavy ion drivers is still at an early stage. Some work is in progress on the large linear accelerators at Berkeley, United States, at Moscow and at the GSI in Darmstadt, Germany.

6.3. Fusion Reactors as Energy Sources

The development of an economical electricity-producing source based on fusion power is one of the most difficult and lengthy endeavors ever undertaken in the history of science and technology. After about 45 years of fusion research, even the scientific feasibility of a fusion reactor has not yet been proven. In the most advanced fusion devices an improvement in the plasma confinement parameter of about a factor of ten (see Fig. 138) still has to be achieved. Furthermore, difficult technological problems with plasma control and peripheral reactor components must be solved. To prove the scientific and technological feasibility of fusion, the technical practicability, reliability, and safety of all necessary components of a fusion power plant, including the operation of a full demonstration plant (DEMO), must be shown. To achieve this objective, a sequential development strategy based on prototype machines of increasing size has to be followed, which in the Tokamak line may consist of the sequence JET – ITER (or NET – DEMO). Even with this optimistic strategy consisting of

only three steps, fusion power that is commercially available to a significant extent cannot be expected before the middle of the next century.

6.3.1. Fusion Power Plants

Magnetic Confinement Fusion. Preliminary studies of fusion power plants have been made, beginning in the early 1970s, some of which led to rather comprehensive fusion reactor designs, mostly for Tokamak reactors. These studies rely on numerically extrapolated D – T plasma confinement properties and contain the layout of technical reactor components, estimates of material needs, fuel requirements, estimates of investment and energy costs, and initial assessments of environmental and safety aspects and, particularly, of the radiological hazard potential of future fusion reactors.

The basic structure of a fusion reactor is about the same in all approaches to magnetic confinement fusion. Figure 141 shows a simplified schematic of a Tokamak fusion reactor, representing a cross-section of the toroidal arrangement. The processes in a fusion power plant are illustrated in Figure 142. The hot fusion plasma with a temperature of $100-200\times10^6$ K and a density of 10^{20} m^{-3} is surrounded first by a metallic, possibly layered wall. The next element in a toroidal shell structure is the blanket of about 0.5 – 1-m thickness, composed of many closely fitted modules. In the blanket modules, lithium is contained in the form of either liquid lithium metal, solid lithium compounds, or molten lithium salts. The blanket has many functions; it moderates the neutrons that leave the fusion plasma and penetrate the blanket, and the heat produced is extracted by a coolant. Furthermore, according to Equations (21) and (22) (see Section 6.1), tritium is produced, which is used as a fuel after extraction, purification, and processing outside the reactor. Surrounding the blanket, a shield absorbs the remaining neutrons and γ radiation to protect the magnets against excess heat. The location of the primary vacuum containment has not yet been decided. A vessel that directly surrounds the plasma as in present experiments, may not be feasible for a fusion reactor because it is exposed to neutron damage and stresses. Of the different options, a primary vacuum containment just behind the blanket, but inside the toroidal coils, is favored. The *toroidal magnetic field coils* are expected to be fabricated from superconductors, because normal conducting coils would require too large a fraction of the power output. Additional coil systems produce time-varying magnetic fields. These are necessary for plasma equilibrium and initial heating (*poloidal field coils* and *ohmic heating coils*), as well as for ash removal (*divertor coils*).

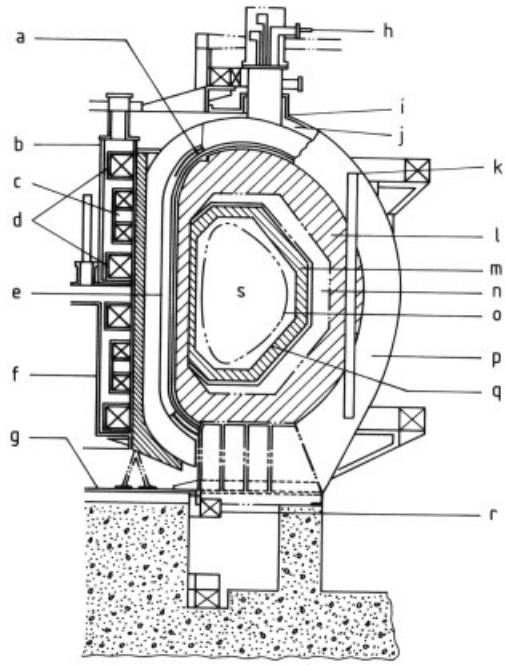

Figure 141. Basic construction of a Tokamak reactor (Starfire cross-section showing the magnetic coil configuration and parts of the vacuum equipment.)
a) Common vactank; b) Upper solenoid He vessel; c) Ohmic heating coils; d) Inner equilibrium field coils; e) Centerpost support cylinder; f) Lower solenoid He vessel; g) Common vactank; h) Helium fill line; i) Vacuum tank; j) Helium vessel; k) Intercoil shear panel; l) Shield; m) Blanket; n) Vacuum plenum; o) Plasma boundary; p) Toroidal field coil and helium vessel; q) First wall; r) Outher equilibrium field and ohmic heating coil; s) Plasma

Since the inherent heating of the Tokamak is insufficient to reach ignition, additional heating methods, such as neutral beam injection or wave heating, must be applied. After ignition is reached, the current must be sustained to allow a continuous "burn" of the plasma, in contrast to the pulsed Tokamak operation in present-day experiments. Methods for current drive envisaged are, again, wave heating or neutral beam injec-

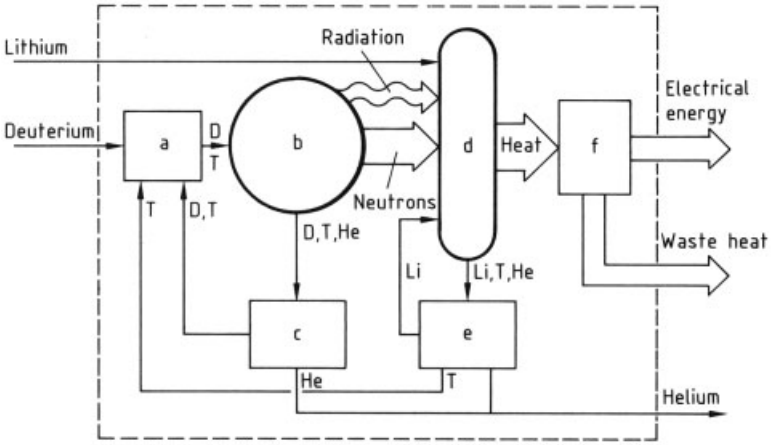

Figure 142. Schematic of functions in a fusion reactor showing very simplified fuel circuits
a) Injector; b) Nuclear fusion; c) Gas purification; d) Blanket; e) Tritium separation; f) Energy transformation

tion. The burned fuel and the fuel lost by diffusion must be replaced continuously. Replacement injection will be performed with pellets made of solid deuterium – tritium or by gas puffing.

The heat that is generated in the first wall, in the blanket, and in the shield (predominantly by fusion neutrons) is conventionally converted to electrical energy by means of a secondary coolant loop, heat exchangers, turbines, and electrical generators.

The *fuel handling* and *fuel circuits* in the fusion power plant are shown in Figure 142. A fusion power plant should be self-supporting with tritium fuel, so a sufficient tritium breeding ratio B (B = rate of tritium production to rate of tritium consumption in fusion) has to be achieved in the blanket. Calculations for different classes of breeding materials (e.g., Li, Li_2O, $Li_{17}Pb_{83}$ with 90 % ^6Li) in design blanket modules give breeding ratios between 1.2 and 1.4. These calculations were carried out for the International Tokamak Reactor (INTOR) design. Such breeding ratios are high enough to reach doubling times of the tritium inventory on the order of months.

Fusion reactors based on magnetic confinement are relatively large in size and in power output. Test reactors of small size are not feasible because some constraints of fusion physics, technology, and economy can be fulfilled only at large torus dimensions. The energy confinement, for example, varies roughly with the square of the minor torus radius. Furthermore, critical limits for the neutron and heat load of the first wall of the reactor do not permit a very compact or small reactor with high wall load. Even with the currently favored first wall (neutron) loads of 1 – 5 MW/m^2, the first wall must be replaced two to four times during the lifetime of the reactor. Thus fusion power densities in the reactor core of only ca. 1 – 5 MW/m^3 can be expected (power densities in fission reactors are about two orders of magnitude higher). From an economic point of view, electrical power outputs of 1 – 2 GW are preferred because the very expensive and complicated components–blanket, shield, superconducting coils–should enclose as high a fusion power as possible. Power plants based on the stellarator may be even larger, because the Stellarator requires higher aspect ratios (major torus radius/minor torus radius) than the Tokamak for stable plasma confinement.

The main characteristics of conceptual commercial Tokamak reactors are:

Thermal power	2 – 5 GW
First wall minor radius (vertical elongation of first wall cross-section ≈ 1.6)	1 – 2 m
Major radius	5 – 8 m
Strength of toroidal magnetic field in plasma center (produced with superconducting D-shaped coils)	3 – 8 T
Plasma current	10 – 20 MA
Neutron loading of first wall	1 – 5 MW/m^2
Total heating power	50 – 90 MW
Overall efficiency	0.3

Figure 143 shows the planned Tokamak power reactor Starfire. It is very important in such a toroidal system with interlinked components that all reactor components be easily accessible for repair, maintenance, and exchange. These requirements pose a serious engineering challenge. Many pertinent design studies for the exchange of reactor components exist, but no convincing solutions have been found. The problem is aggravated by the inaccessibility of the reactor hall, even after short reactor operation, because of radioactivity. All service and maintenance must be performed by remote control.

Construction Materials. Connected with the maintenance problem is the choice of materials for the reactor structure, the first wall, and other components inside the shield. Material requirements are resistance against radiation damage, compatibility with lithium, low neutron absorption cross-section, low long-time activation, low tritium permeation, and good thermal and mechanical properties up to 10^3 K. For experimental devices such as NET or ITER, stainless steels (AISI 316) are very likely to be selected. They have been optimized for fission power plants, but they may not be the best choice for fusion reactors because of the very different neutron spectrum in fusion. Additional material research, which should also include low-activation alloys (i.e., ferritic or austenitic stainless steels with reduced or replaced N, Ni, Mo, Cu, and Nb content; vanadium alloys such as V-15 Cr-5 Ti, V-20 Ti, V-3 Ti-1 Si), is necessary. Research should be devoted especially to understanding material properties when materials are exposed to large fluences of 14-MeV neutrons that are produced in the D – T reaction. Neutron irradiation affects the mechanical, thermal, and electrical properties of materials in many ways; it produces, for example, various defects in the crystal lattice by generation of recoil nuclei, and generates hydrogen and helium via (n, p) and (n, α) reactions. The most serious consequences, pertaining particularly to the first wall and other casings and containing systems (blanket elements, coolant pipes, etc.), are loss of ductility and swelling of the material. Radiation damage is especially critical for the first wall between the plasma and the blanket. Here the flux of primary fusion neutrons with a 14-MeV maximum will be at least $5\times10^{13} - 5\times10^{14}\,\mathrm{cm^{-2}\,s^{-1}}$ corresponding to a power density of 1 – 2 MW/m^2. Furthermore, the first wall is exposed to high fluxes of photons and high-energy ions and atoms that escape from magnetic confinement. By bombardment with fast plasma particles, wall material is sputtered off, leading eventually to contamination of the plasma and deterioration of the energy confinement.

The questions of radiation damage and first-wall durability are considered key problems in fusion reactor development, equivalent to the problem of plasma containment. Present expectations are that the first wall and the blanket will have to be replaced 3 to 4 times during the lifetime of a power reactor.

Inertial Confinement Fusion Reactors. System studies and conceptual designs on inertial confinement fusion reactors have also been performed. In a laser fusion reactor the pellets are placed in the center of a burn chamber, where thermonuclear burns are initiated by laser pulses. A blanket of lithium surrounds the chamber and the exploding pellets. The minimum energy release in one microexplosion, determined by physical and economic constraints, is about 100 MJ (equivalent to 20 kg of TNT). Because the energy is carried by a small mass of high-energy particles (neutrons, α particles, and pellet debris) and by radiation, control of the hydrodynamic blasts may be eased. To protect reactor components, designers have proposed lithium wetted-wall chambers, magnetically protected walls, and liquid lithium curtains. A performance map of the operational parameters for a 1-GW$_e$ power plant with a single reactor chamber is depicted in Figure 144. This figure shows the driver energy required, the fusion energy released per pulse and the repetition rate necessary to reach 1-GW$_e$, if the indicated performances i.e., energy gain in the pellet, and driver efficiency, could be achieved. Energy gain is the ratio of the driver energy absorbed in the pellet to the fusion energy released. According to Figure 144, a 1-GW$_e$ single- chamber power plant in the intermediate gain region (G $\approx$ 100 – 500) would require laser drivers of at least 10 – 20 % efficiency, with megajoules of driver energy and a repetition rate of $10^5 - 10^6$ pellets per day. Each pellet will produce only about 10 kW · h electrical energy. Pellet factories would have to be installed to produce the relatively complicated

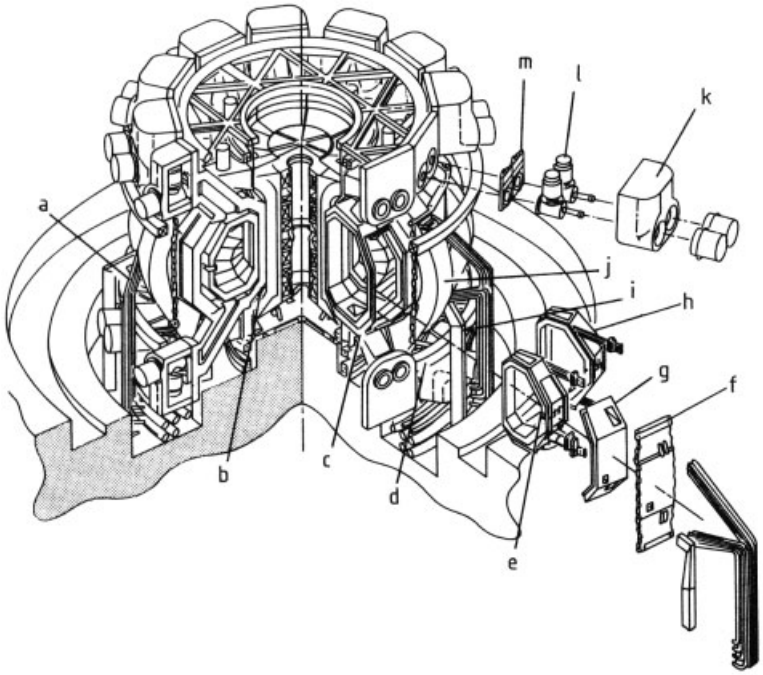

Figure 143. Starfire commercial Tokamak reactor reference design, showing exchange of modules
a) Water coolant inlet and outlet; b) Segmented copper equilibrium field coils; c) Shield; d) Superconducting equilibrium field coils; e) Limiter; f) Antitorque panel; g) Shield access door; h) Blanket sector; i) RF duct; j) Toroidal field duct; k) Vacuum pump shield; l) Pumps; m) Valves

multishell pellets cheaply, but with high precision. The development of efficient drivers, the mass production of pellets, the handling of about 0.5 kg of tritium per day, the design of blast-resistant chambers, and the safety aspects of repeated nuclear blasts in a surrounding of highly radioactive components pose extreme demands that must be met for energy production. These problems in combination with the relationship of ICF research to nuclear weapons development, may have led the EC to concentrate almost exclusively on magnetic confinement fusion.

6.3.2. Energy Resources and some Economic Aspects

Fusion reactors based on the D – T reaction consume lithium and deuterium as primary fuels. *Deuterium* is present in all water. With a supply of about 10^{13} t in the oceans and a specific consumption of about 25 g MW_{th}^{-1} a^{-1} in a fusion reactor, the equivalent energy resources of deuterium are practically unlimited. *Lithium* is widely distributed in the earth's crust. At present, it is extracted from brine and pegmatitic ores. At least about 1.2×10^6 t of lithium is economically minable from the land mass. This amount corresponds to an approximate total energy production of $3.5\times10^4 - 12.3\times10^4$ EJ (1 EJ = 10^{18} J), if it were used up in future fusion reactors. The average lithium content of seawater is about 0.1 ppm, leading to technically available resources of 10^{11} t, with an energy equivalent of 5×10^9 EJ. The total worldwide energy consumption in 1990 was about 350 EJ. Thus with respect to fuel supplies, fusion can be considered an inexhaustible energy source. Lithium and deuterium can be extracted from seawater relatively easily and at low cost. By using actual raw material prices for deuterium and lithium, specific fractional energy costs for fuel (D_2+ Li) of 3×10^{-3} cents/$kW_e\cdot h$ would result.

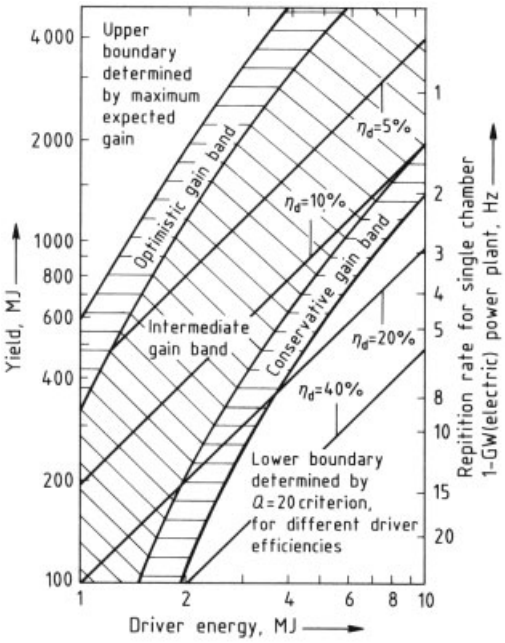

Figure 144. Operational parameters of an W_e 1-GW ICF power plant showing the driver energy and the fusion yield per pellet explosion and the repetition rate required for a single reaction chamber. Conservative and optimistic gain refer to assumptions in the implosion physics. η_d is the driver (photon- or ion-beam) efficiency.

The average energy density in a fusion reactor is very low (some watts per cubic centimeter). Thus, the energy expenditure to build a reactor might be too high, or the energy payback time could be too long. Extensive studies have shown that the *energy expenditure* to build and operate a 1-GW_e fusion reactor will be less than 1 % of the total net energy gained in 30 years of operation. This relation is more favorable for the fusion reactor than for a fission reactor (LWR) because of the low energy expenditure for fuel in the fusion reactor.

Cost estimates of generating electricity by fusion have been made. They are necessarily imprecise for a system that will not be put in general service until the middle of the next century, at the earliest. For a Tokamak reactor the costs of fuel and fuel processing are negligible, but the capital costs of the plant are high, particularly for the unconventional part– the Tokamak reactor. System studies made in 1977 – 1980 estimated the total capital costs for a commercial power plant with a Tokamak reactor to be between \$ 1280 and 1440/kW_e using prices at that time. Energy price projections have been made for the middle of the next century. At that time, energy costs for fusion-generated electricity should be comparable to those of fission reactors and coal power plants but should be still cheaper than those of photovoltaic electricity generation. Thus, from the present point of view, competitive production of electrical energy by nuclear fusion seems possible.

6.3.3. Environmental Impact and Safety-Related Aspects of Fusion Power

Although, presently, the fundamental principles of thermonuclear fusion deserve most of the research effort, several studies on the environmental impact of fusion power were performed in the 1990s. Quantitative statements on these safety and environmental properties of the future fusion reactor are based on the detailed design of a 1 GW_e Tokamak reactor of the FCTR type (European First Commercial Tokamak Reactor).

As a result of these investigations the following favorable statements with respect to the safety and environmetal impacts of the fusion reactor can be made:

1) In a fusion reactor chain reactions and steep power excursions – which would result in a power runaway of the reactor – do not occur.
2) The radioactive materials in a fusion reactor – tritium and the activated material produced by activation of the reactor structure by fusion neutrons – have a relatively low biological hazard potential.
3) The quantity and the activity of the radioactive materials produced in a fusion reactor will be substantially reduced by future material developments.
4) The acquisition of materials for fusion reactors will not despoil their resources. The fusion fuels deuterium and lithium (which serves as breeding material for tritium) are abundant and widely distributed on earth. Exceptions may be beryllium and helium for specific reactor designs, since beryllium may be used as a neutron multiplicator in blanket designs with solid breeders, and helium is needed for the superconducting coils.
5) Fusion reactors do not emit chemicals that might have an impact on the climate.

Environmental Impact. The biological hazard potential of tritium, a β emitter with a half-life of 12.3 a, is relatively low. Nevertheless, protection of the environment from the radiological impact of a fusion power plant depends essentially on the control of tritium emission. Tritium can penetrate hot metallic walls relatively easily. In reactor design, extensive measures for tritium containment must be employed, such as doublewalled tubing, diffusion barriers, low tritium concentration in coolant pipes, detritating of the reactor hall, and outer metallic containment for both the reactor hall and the tritium process building. Provided that leakage is limited by such measures, estimates forecast that tritium losses and their effects can be made smaller than required by current radiological safety regulations. For the FCTR (1.2 GW_e) a release of 2 g of tritium per year (740×10^{12} Bq/a) of atmospheric emissions and 0.15 g of tritium per year (50×10^{12} Bq/a) of aquatic emissions are estimated. These emissions occur mainly because of maintenance work. If it is unfavorably assumed that all emissions are in their oxidized form (HTO; an unfavorable assumption), a person living at the reactor fence (1000-m distance) would obtain a dose of 8 µSv/a (0.8 mrem/a), provided that dilution caused by a chimney or by cooling water is taken into account. In planning nuclear power stations, a dose of < 300 µSv/a must be guaranteed.

In Tokamak reactor designs with stainless steel as the structure material the *radioactive inventory* formed by neutron activation is very high. Its activity (FCTR: ca. 2.8×10^{20} Bq/GW_e) is comparable to that of fission reactors. Most of the activity is found in solid metallic structural components, with no actinides, no plutonium, and not many volatile radioactive materials in addition to tritium.

In routine reactor operation, activated structural materials will be released to the environment only in negligible quantities, mainly as corrosion products. The biological hazard potential (BHP) of a fusion reactor is considerably lower (10 – 100-fold) than that of fission reactors and decays much faster after shutdown.

Furthermore, the BHP depends on the structural materials of the reactor. In current designs, reactor steels are foreseen. The BHP could possibly be reduced further by radiologically tailoring the materials used in construction of the plant, (i.e., choosing low-activation materials that form no long-lived isotopes). Different alloys (e.g., steels without nickel and molybdenum, or vanadium – titanium alloys) have been proposed with which the decay time of radioactivity is drastically reduced and a considerably lower total BHP results.

Radioactive Waste. Assessments indicate that Tokamak fusion reactors produce huge amounts of radioactive waste during their lifetimes. Such waste is caused mainly by maintenance; by first-wall and divertor plate erosion by replacement of the first wall; of plasma facing components, and of the breeding blanket. The high-activity waste stems mainly from the stainless steel structure of the first wall. It must be disposed of in geologic repositories. Indications are that the high- and low-activity wastes are comparable in volume and initial activity to those of a fission reactor with the same power output. The low-activity waste comes mainly from the tritium handling and purification systems, and probably also from solid breeder materials. Again, the possibility exists for developing materials with favorable radiological properties, so that part of the waste may even be recycled after a sufficiently long cooling time.

Reactor Accidents. With respect to accidents during operation and accidental release of *radioactivity*, precise statements cannot be made because fusion reactor designs are still in the planning stage and credible accidents cannot be identified or evaluated very well. Maximum consequences can only be roughly estimated. The reaction chamber in a fusion reactor contains a small amount of fuel, sufficient for some seconds of burn time. The reaction rate is self-limiting, so a power excursion is impossible. Furthermore, the power density of the decay heat and its increase after shutdown of a fusion reactor are much lower than in a fission reactor. Thus, emergency cooling is considerably easier in fusion reactors, and meltdown of materials can be excluded. The energies stored in the reactor components are not high enough to melt or evaporate a considerable fraction of structural material.

Most of the hazard could be expected from the *chemical energy* of the metallic lithium

in the blanket in the event of a lithium fire. Lithium reacts vigorously with air, water (self-ignition), and even concrete. Radioactivity cannot be induced in lithium. In modern reactor designs, breeding blankets are provided that contain solid lithium compounds or lithium – lead alloys. Such materials are not prone to exothermic reactions or self-ignition.

A potential environmental hazard in the event of accident is primarily the *release of tritium*. Even in severe accidents, only up to 200 g tritium is expected to be released into the reactor hall or other buildings because most of the tritium inventory ($\approx$ 5 kg) is rigidly bound in reactor structures in the storage and partially in the process systems, all of which are in different isolated buildings. The release outside the buildings may be kept one or two orders of magnitude lower than the release inside, given the outer containment and tritium removal in the reactor buildings by emergency cleanup systems. Even in an accident that ruptures the building containment (earthquake, airplane crash, sabotage), only the tritium content of the pertinent building (about 100 g) is expected to be released to the environment.

Estimates of *radiation exposure* (*effective dose*) have been made, based on a release of 10 g of tritium as HTO. The effective dose for adults living within 1000-m distance and consuming permanently contaminated food would be about 25 mSv (2.5 rem). According to radiological safety regulations, the limiting dose in accidental radiation exposition (abnormal events of very low probability) is 5 – 15 rem.

In relevant contemporary studies, no sequence of events could be identified in any accident scenarios for considerably higher releases of radiation. Nevertheless, in the extreme limit, hypothetical release of the total mobilizable inventory of about 1.5 kg of tritium and 1 kg of activated structural material may be contemplated. Even then, evacuation would not be required beyond a radius of a few kilometers, if the limits for evacuation are 5 – 25 rem. Therefore, fusion is not expected to cause a major disruption of normal life in the community surrounding the reactor site.

7. References

Acknowledgment. The author of Section 5.4 (G. Koch) wishes to acknowledge contributions by E. Henrich, KfK Institut für Heisse Chemie (Off-gas purification), J. Lausch, WAK Betriebsgesellschaft (Nuclear materials safeguards) G. Petrich, KfK Institut für Heisse Chemie (VISCO flow sheet calculations), and H. W. Wiese, KfK Institut für Neutronenphysik und Reaktortechnik (KORIGEN fuel calculations).

1. O. Hahn, F. Strassmann, *Naturwissenschaften* **27** (1939) 11, 89.
2. *Ullmann,* 4th ed., **14,** 13 – 16.
3. D. Bünemann, M. Kolb: "The Core Design of the Reactor of the Nuclear Ship "Otto Hahn," *Adv. Nucl. Sci. Technol.* **6** (1972) 1 – 44.
4. K. Juurmaa, A. Laukia: "Nuclear-Powered Merchant Ships Today and Tomorrow," *Nucl. Eur.* **9** (1989) nos. 7 – 8, 28.
5. A. E. Waltar, A. B. Reynolds: *Fast Breeder Reactors.* Pergamon Press, New York 1981.
6. IAEA (ed.): "Status of Liquid Metal Cooled Fast Breeder Reactors," *Tech. Rep. Ser. I.A.E.A.* 1985, no. 246.
7. M. Rapin: "World Overview of FBR Development," *Nucl. Eur.* **9** (1989) nos. 9 – 10, 15.
8. K. Kugeler, R. Schulten: *Hochtemperaturreaktortechnik,* Springer Verlag, Berlin 1989.
9. K. Kussmaul (ed.): "The High-Temperature Reactor and Nuclear Process Heat Application," *Nucl. Eng. Des.* **78** (1984) no. 2, 89 – 300.
10. "High-Temperature Gas-Cooled Power Reactors," *Nucl. Eng. Des.* **26** (1974) no. 1, 1 – 230.
11. L. Stengärde: "SECURE – the Heat Only Nuclear Reactor," *Atomkernenerg. Kerntech.* **45** (1984) no. 1, 31.
12. W. Seifritz: *Ein heterogener Heizreaktor kleiner Leistung (10 – 50 MW) für die nukleare Nahwärmeversorgung,* **EIR-526 (Sept. 1984)** Eidg. Inst. für Reaktorforschung, Würenlingen, Switzerland,1984.
13. H. Reutler, G. H. Lohner: "Advantages of Going Modular in HTRs," *Nucl. Eng. Des.* **78** (1984) 129.
14. C. Sundqvist, T. Pedersen: "PIUS the Forgiving Reactor," *Mod. Power Systems* **5** (1985) no. 9, 69.
15. L. S. Tong, J. Weismann: *Thermal Analysis of Pressurized Water Reactors,* American Nuclear Society ANS, Hinsdale, Ill., 1970.

16. W. Oldekop: "Druckwasserreaktoren für Kernkraftwerke," *Thiemig-Taschenb.,* (1974) no. 51.
17. R. T. Lahey Jr., F. J. Moody: *Thermal Hydraulics of a Boiling Water Nuclear Reactor,* American Nuclear Society ANS, Hinsdale, Ill., 1977.
18. A. R. Foster, R. L. Wright: *Basic Nuclear Engineering* 4th ed., Allyn and Bacon, Boston, Mass. 1983.
19. S. Glasstone, A. Sesonske: *Nuclear Reactor Engineering,* 3rd ed., Van Nostrand Reinhold, New York 1981.
20. R. A. Rydin: *Nuclear Reactor Theory and Design,* University Publications, Blacksburg, Va. 1977.
21. L. Massimo: *Physics of High Temperature Reactors,* Pergamon Press, Oxford 1976.
22. J. R. Lamarsh: *Introduction to Nuclear Engineering,* Addison – Wesley, Reading, Mass. 1975.
23. W. Oldekop: "Einführung in die Kernreaktor- und Kernkraftwerkstechnik," *Thiemig-Taschenb.,* (1975) no. 53, 54.
24. D. Emendörfer, K. Höcker: *Theorie der Kernreaktoren,* **B.I.-Hochschultaschenbücher, vol. 411/411 a, 412/412 a,** Bibliograph. Institut, Mannheim 1969.
25. J. R. Lamarsh: *Introduction to Nuclear Reactor Theory,* Addison – Wesley, Reading, Mass. 1966.
26. K. Wirtz, K. H. Beckurts: *Neutron Physics,* Springer Verlag, Berlin 1964.
27. G. Gamow: *Der Bau des Atomkerns und die Radioaktivität,* Hirzel-Verlag, Leipzig 1932.
28. E. Bodenstedt: *Experimente der Kernphysik und ihre Deutung,* **vols. 1 – 3,** Bibliograph. Institut, Mannheim 1972.
29. R. Kinsey (ed.): *ENDF-102 Data Formats and Procedures for the Evaluated Nuclear Data File,* **ENDF/B-V. BNL-NCS-50 496 (ENDF-102), 2nd ed. (ENDF/B-V) revised Nov. 1983,** Brookhaven Natl. Lab., Upton, NY.
30. L. Dresner: *Resonance Absorption in Nuclear Reactors,* Pergamon Press, New York 1960.
31. J. J. Duderstadt, L. J. Hamilton: *Nuclear Reactor Analysis,* Wiley, New York 1976.
32. G. I. Bell, S. Glasstone: *Nuclear Reactor Theory,* Van Nostrand Reinhold, New York 1970.
33. M. Clark, K. F. Hansen: *Numerical Methods of Reactor Analysis,* Academic Press, New York 1964.
34. V. I. Pushkarev, A. D. Zhirnov, A. P. Sirotkin: "Ways of Altering the Coefficients of Reactivity in RBMK Reactors," *At. Energiya* **46** (1979) 386.
35. D. L. Hetrick: *Dynamics of Nuclear Reactors,* Univ. of Chicago Press, Chicago, Ill., 1971.
36. G. R. Keepin: *Physics of Nucelar Kinetics,* Addison – Wesley, London 1965.
37. E. Teuchert, H. J. Rütten, H. Werner, *Nucl. Technol.* **58** (1982) 414.
38. M. J. Bell: *ORIGEN – The ORNL Isotope Generation and Depletion Code,* **ORNL-4628 (May 1973)**, Oak Ridge National Laboratory, Oak Ridge, Tenn., 1973.
39. M. Benedict, T. H. Pigford, H. W. Levi: *Nuclear Chemical Engineering,* Mc Graw-Hill, New York 1981.
40. Swedish Nuclear Fuel Supply, Division KBS: Final Storage Nuclear Fuel – KBS-3, Summary, Stockholm 1977.
41. W. D. Bond, H. C. Clairborne, R. E. Leuze: "Methods for Removal of Actinides from High-Level Waste," *Nucl. Technol.* **24** (1974) 362.
42. W. Oldekop: "Einführung in die Kernreaktor- und Kernkraftwerkstechnik," *Thiemig Taschenb.,* (1975) no. 53, 54.
43. G. Kessler: *Nuclear Fission Reactors,* Springer Verlag, Wien 1983.
44. H. Michaelis, C. Salander: *Handbuch Kernenergie,* VWEW Verlag,Frankfurt a. M. 1995.
45. A. Voß, E. Thöne (eds.): *Kernenergie und die zukünftige Energieversorgung,* TÜV Rheinland, Köln 1987.
46. T. Schwarz: "The Four Steps to Nuclear District Heating," *Nucl. Eng. Int.* **22** (1977) no. 261, 45.
47. W. Oldekop in K. O. Thielheim (ed.): *Primary Energy,* Springer Verlag, Berlin 1982, p. 66.
48. H. Frewer: "Future Challenge of Nuclear Power in the European Community," *Commemorative Anniversary Meeting of the European Nuclear Society,* Karlsruhe 1989.
49. IAEA (ed.): *Nuclear Power; Status and Trends,* IAEA, Vienna 1987.
50. Nuclear Power Oversight Committee (ed.): "Nuclear Energy for the Future: What We Must Do," *ATW Atomwirtsch. Atomtech.* **34** (1989) 305 – 309.
51. C. G. F. Harding: "The Future of Nuclear Energy in Western Industrialized Nations," *ATW Atomwirtsch. Atomtech.* **33** (1988) 543 – 546.

52. IAEA (ed.): "Nuclear Power Performance and Safety," *Proc. Int. Conf. by IAEA held in Vienna,* **vol. 1, 28. Sept. – 2. Oct. 1987,** IAEA, Vienna 1988.
53. W.-M. Liebholz (ed.): *Jahrbuch der Atomwirtschaft 1997,* Verlagsgruppe Handelsblatt, Düsseldorf 1997.
54. *IAEA Bulletin* **38** (1996) no. 3 and (1996) nos. 72, 73, 74.
55. F. D. Sowby (ed.): "Recommendations of the International Commission on Radiological Protection," *ICRP-Publ.* 1977, no. 26.
56. E. Sauter: *Grundlagen des Strahlenschutzes,* 2nd ed., Thiemig Verlag, München 1983.
57. UNSCEAR: "Ionizing Radiation, Sources and Biological Effects," *Report to the General Assembly with Annexes, UN,* New York 1982; "Sources, Effects and Risks of Ionizing Radiation," *Report to the General Assembly with Annexes, UN,* New York 1988.
58. H. G. Paretzke: "Risiko für somatische Spätschäden durch ionisierende Strahlung," *Phys. Bl.* **45** (1989) 16 – 24.
59. IAEA (ed.): "Radiation Protection in Nuclear Energy," *Proc. Int. Conf. by IAEA held in Sydney,* **18 – 22 April 1988,** IAEA, Vienna 1988.
60. IAEA (ed.): "Basic Safety Standards für Radiation Protection," *Saf. Ser. I.A.E.A.* 1982, no. 9.
61. European Community, ABlEG 1980, no. L246.
62. European Community, ABlEG 1984, no. L265.
63. W. Bischof: "Das aktuelle Strahlenschutzrecht," *ATW Atomwirtsch. Atomtech.* **33** (1988) 527 – 531.
64. Verordnung über den Schutz vor Schäden durch ionisierende Strahlen, Version June 30, 1989, last changes: August 18, 1990 (BGBl I p. 59).
65. ICRP (ed.): "1990 Recommendations of the ICRP," *Ann. ICRP* 1991, no. 60.
66. IAEA (ed.): "Basic Safety Principles for Nuclear Power Plants," *Saf. Ser. I.A.E.A.* 1988, no. 75.
67. Bundesministerium für Forschung und Technologie (ed.): *Deutsche Risikostudie Kernkraftwerke,* TüV Rheinland, Köln 1979.
68. U.S. Atomic-Energy Commission (ed.): *Reactor Safety Study,* **WASH-1400 (NUREG-75/014)** 1975.
69. U. Hauptmanns et al.: *Technische Risiken – Ermittlung und Beurteilung,* Springer Verlag, Berlin 1987.
70. J. Rassow: *Risiken der Kernenergie – Fakten und Zusammenhänge im Lichte des Tschernobyl-Unfalls,* VCH Verlagsgesellschaft, Weinheim 1988.
71. K. Köberlein: "Sicherheitsanalysen im Lichte von TMI und Tschernobyl," *ATW Atomwirtsch. Atomtech.* **32** (1987) 237 – 243.
72. Creys-Malville Nuclear Power Station, *Nucl. Eng. Int.* **23** (1978) no. 272, 43.
73. ENS (ed.): "Reactors for 21st Century," *Nuclear Europe Worldscan* 1990 nos. 11 – 12, 10 – 25.

General References

74. Th. Bohn (ed.): *Handbuch Energie,* **vol. 10,** TüV Rheinland, Köln, in press.
75. IAEA-Publications (TECDOC, Conference Proceedings, Guide Books, Special Reports).
76. OECD/NEA Publications.

Specific References

77. *Status of Advanced Technology and Design for Water Cooled Reactors: LWR,* **IAEA TECDOC-479,** 1988.
78. W. Aleite, H. Hofmann, M. Jung: "Leittechnik in Kernkraftwerken," *ATW Atomwirtsch. Atomtech.* **32** (1987) no. 3.
79. W. Aleite, S. Struensee: "Leistungsregeleinrichtungen und Begrenzungen von DWR and SWR," *ATW Atomwirtsch. Atomtech.* **32** (1987).
80. G. Bachmann, W. Sych: "Aufgaben und Konzept des Reaktorschutzsystems," *ATW Atomwirtsch. Atomtech.* **32** (1987).
81. W. H. Dio, W. Sych, M. Grosse-Schulte: "Strahlungs- und Aktivitätsmessungen in Kernkraftwerken," *ATW Atomwirtsch. Atomtech.* **32** (1987) 139 –143.
82. Bundesminister des Innern: Gesetz über die friedliche Verwendung der Kernenergie und den Schutz gegen ihre Gefahren (German atomic law), Version July 15, 1985 (BGBl. I, p. 1565).
83. Bundesminister des Innern: Verordnung über den Schutz durch ionisierende Strahlen, Oct. 13, 1976 (BGBl. I, p. 2905), last amendment May 27, 1989 (BGBl. I, p. 940).
84. *Sicherheitstechnische Regel des KTA (Kerntechnischer Ausschuß),* Carl Heymanns Verlag, Berlin (continuously updated).
85. IAEA: *IAEA Safety Guides,* International Atomic Energy Agency, Division of Publications, Vienna (continuously updated).
86. RSK (Reaktor-Sicherheitskommission): *Leitlinien für Druckwasserreaktoren,* 3rd ed., Oct. 1981 (Beilage no. 19 182, BAnz. no. 69, 14 Oct. 1982).

87. Bundesministerium des Innern: Sicherheitskriterien für Kernkraftwerke (BAnz. no. 206, 3 Nov. 1977).
88. KTA 3301, Nachwärmeabfuhrsysteme von Leichtwasserreaktoren, Carl Hymanns Verlag, Berlin, Nov. 1984.
89. E. Tenckhoff, M. Erve, E. Weiß: "Material Concept – Basis for Reliability and Structural Integrity of Components and Systems," *Proceedings of the Sixth International Conference on Pressure Vessel Technology,* Beijing, 11 – 15 Sept. 1988, pp. 1021 – 1039.
90. B. Lustman, F. Kerze Jr., *The Metallurgy of Zirconium,* McGraw Hill, New York 1955.
91. G. L. Miller: *Zirconium,* Butterworth, London 1957.
92. D. Knödler, S. Reschke, H. G. Weidinger, *Kerntechnik* **50** (1987) 255 – 265.
93. W. J. Kroll, *Z. Metallkunde* **45** (1954) 67.
94. E. Tenckhoff, ASTM STP 551, 1973, 179 – 199.
95. S. Kass, ASTM STP 368, 1964, 3 – 27.
96. W. Debray, H. Cerjak: "Werkstoffeigenschaften des Stahles 22 NiMoCr 37 für Kernreaktorkomponenten," *Sammelband VGB-Werkstofftagung,* 1971, pp. 12 – 23.
97. H. Cerjak, W. Debray, F. Papouschek: "Eigenschaften des Stahles 20 MnMoNi 55 für Kernreaktorkomponenten," in "Werkstoffe und Schweiß-technik im Kraftwerk," *Sammelband VGB-Konferenz,* 1976, pp. 3 – 44.
98. M. Brumovsky, J. Brynda, Z. Hrbek, J. Sule: "Properties of Semiproducts for VVER 440 Reactor Pressure Vessels," *Skoda Review* **1** (1981).
99. U. Potapovs, J. R. Hawthorne: "The Effect of Residual Elements on the Response of Selected Pressure Vessel Steels and Weldments to Irradiation at 550 °F," *Nuclear Applications* **6** (1969) 27.
100. C. Leitz, J. Koban: "Development of Reactor Pressure Vessel Design, Neutron Fluence Calculation and Material Specification to Minimize Irradiation Effects on Radiation Embrittlement of Nuclear Reactor Pressure Vessel Steels," *ASTM Spec. Tech. Publ., 1011 (Radiation Embrittlement Nucl. React. Pressure Vessel Steels: Int. Rev.)* **3** (1989) 130 – 144.
101. W. Debray, L. Stieding, E. Weiß, N. Wieling: "Experience with Stabilized Stainless Steels in Light-Water Reactor Technology," *The 2nd Japanese-German Joint Seminar,* Tokyo, 21 – 24 Feb. 1983.
102. M. Erve et al.: "Planned and Implemented Remedial Measures to Prevent IGSCC and to Ensure a Safe Plant Operation with BWR Piping Systems made of Stabilized Austenitic Steels," *20. MPA-Seminar*, Stuttgart, Oct. 6. and 7. 1994.
103. H. Coriou et al.: "Sensitivity to Stress Corrosion and Intergranular Attack of High Nickel Austenitic Alloys," *Corrosion Houston* **22** (1966) 280 – 290.
104. U. Smith et al.: "Inconel 690, A Material with Improved Corrosion Resistance for PWR Steam Generator Tubes," *2nd International Symposium on Environmental Degradation of Material in Nuclear Power Systems Water Reactors,* Monterey, Ca., Sept. 1985.
105. B. Stellwag, N. Wieling, L. Stieding: "Corrosion Resistance of SG Tubing Materials Alloy 800 and Alloy 690 – A Comparative Study," *Proceedings of the Second International Symposium on Environmental Degradation of Materials in Nuclear Power Systems – Water Reactors,* Monterey, Ca., 9 – 12. Sept. 1985, American Nuclear Society, La Grange Park, Ill., 1986.
106. B. C. Friedrich: "Reduction of Cobalt 60 Inventory in Light Water Reactors," *Kerntechnik* **54** (1989) no. 2, 109 – 113.
107. A. Dyer (ed.): *Gas Chemistry in Nuclear Reactors and Large Industrial Plant,* Heyden, London 1980.
108. D. R. Holmes, *Corros. Steels CO2, Proc. Br. Nucl. Energy Soc. Int. Conf., 1974.*
109. G. Menken et al.: "Review of the Gas/Metal Interactions in HTR Helium up to 950 °C," *Gas Cooled Reactors Today,* BNES-Conference, Bristol, 20 – 24 Sept. 1982.
110. IAEA (ed.): "Summary Report on the PostAccident Review Meeting on the Chernobyl Accident," *Saf. Ser. IAEA* 1986, no. 75.
111. *Report on the Accident at the Chernobyl Nuclear Power Station,* NUREG-1250, U.S. Nuclear Regulatory Commission, Washington, D.C., 1987.
112. W. Kröger, S. Chakraborty: *Tschernobyl und weltweite Konsequenzen,* TüV Rheinland, Köln 1989.
113. W. Heisenberg: *Der Teil und das Ganze,* Piper Verlag, München 1969, p. 230.
114. Status of Advanced Technology and Design for Water Cooled Reactors: Heavy Water Reactors, IAEA-TECDOC-510, 1989.
115. Atoms in Japan, Aug. 1989.

116. CANDU 600, AECL PP-26 (1976).
117. The CANDU Power Plant, AECL-5327 (1977).
118. CANDU 600 MW(e), IAEA, Vienna 1978.
119. *The Nuclear Equipment Arrangement of Darlington Generating Station,* **ASME 78-NE-11** (1978).
120. R. Knox: "A Larger Version of CANDU Designed for World Markets," *Nucl. Eng. Int.* **26** (1981) no. 314, 43 – 52.
121. C. C. Madero et al.: "Building a 745 MWe Pressure-Vessel PHWR in the Argentine," *Nucl. Eng. Int.* **27** (1982) no. 332, 30 – 40.
122. H. Fabian, K. Frischengruber: "Safety Concept and Evaluation for the Pressurized Heavy Water Reactor Atucha II," *Atomkernenerg. Kerntech.* **46** (1985) no. 1, 14 – 19.
123. V. Dörfler, F. Loos, K. Frischengruber, R. C. Vanzulli: "Atucha II and the KWU Pressurized Heavy Water Reactor Line," *Kerntechnik* **52** (1988) no. 2, 83 – 90.
124. S. Sawai, T. Haga, M. Akebi, K. I. Kontani:"Fugen HWR Reaches Commercial Operation", *Nucl. Eng. Int.* **24** (1979) no. 8, 33 – 48.
125. H. Michaelis, H. Mayer: "Schnelle Brutreaktoren," *Handbuch der Kernenergie,* Econ Verlag, Düsseldorf 1986, pp. 105 – 117.
126. *Ullmann,* 4th ed., **14,** 101 – 125.
127. W. Marth: "Die Brutrate bei europäischen Brüterkraftwerken," *ATW Atomwirtsch. Atomtech.* **30** (1985) 623 – 627.
128. U. Klimmek, H. Mayer, U. Wehmann: "Die Bedeutung des Brutreaktors im Energieträgermarkt," *Jahrestagung Kerntechnik 1988,* Deutsches Atomforum.
129. IAEA: *Experimental Breeder Reactor (EBR),* **vol. IV,** Vienna 1962, pp. 313 – 318.
130. W. Marth: "Entwicklungstendenzen beim Schnellen Brüter," *ATW Atomwirtsch. Atomtech.* **29** (1984) 363 – 368.
131. H. Michaelis, J. Rehnelt: "Wirtschaftliche Aussichten der verschiedenen Reaktorlinien," *Handbuch der Kernenergie,* Econ Verlag, Düsseldorf 1986, pp. 554 – 558.
132. J. R. Frisch: "Future Stresses for Energy Resources," *World Energy Conference,* Vienna 1986.
133. H. Michaelis, K. E. Kegel, E. U. Krische: "Die Versorgung mit Natururan und Thorium," *Handbuch der Kernenergie,* Econ Verlag, Düsseldorf 1986, pp. 681 – 706.
134. R. Carle: "Gegenwart und Zukunft der Schnellen Brüter," *ATW Atomwirtsch. Atomtech.* **33** (1988) 478 – 481.
135. R. Lallement, L. Guccia: "Research and Development in Support of Superphénix," *Nucl. Eur.* **5** (1985) no. 11.
136. NERSA: *The Creep-Malville Power Plant (Superphénix),* Company brochure, 1987.
137. W. Marth: "Auf dem Weg zum Euro-Brüter," *ATW Atomwirtsch. Atomtech.* **34** (1989).
138. U.S. Atomic Energy Commission, Comparison of Foreign and U.S. LMBRF Plant Parameters, WASH-1354, 1974.
139. IAEA (ed.): *Enrico Fermi: Atomic Power Plant,* **vol. IV,** IAEA, Vienna 1962, pp. 319 – 324.
140. IAEA (ed.): *DFR (Dounray Fast Reactor),* **vol. IV,** IAEA, Vienna 1962, pp. 301 – 306.
141. IAEA (ed.): *PFR (Prototype Fast Reactor),* **vol. IX,** IAEA, Vienna 1971, pp. 218 – 226.
142. IAEA (ed.): *BOR 60,* **vol. IX,** IAEA, Vienna 1971, pp. 211 – 218.
143. IAEA (ed.): *BN 350,* **vol. VII,** IAEA, Vienna 1968, pp. 311 – 318.
144. A. J. Leipunski et al.: "The BN 600 Fast Reactor Nuclex 69," *Techn. Meeting* 1969, no. 4/5.
145. IAEA (ed.): *Rapsodie,* **vol. VI,** IAEA, Vienna 1966, pp. 233 – 238.
146. IAEA (ed.): *Phénix,* **vol. IX,** IAEA, Vienna 1971, pp. 227 – 234.
147. A. Plessa, A. Mergui, R. Casini: "Superphénix kurz vor seiner Vollendung," *ATW Atomwirtsch. Atomtech.* **30** (1985).
148. W. Marth: "Natrium – immer noch das Kühlmittel der Wahl?" *ATW Atomwirtsch. Atomtech.* **33** (1988) no. 5, 246 – 249.
149. K. Förster, G. Ruloff: "Flüssigmetalle als Wärmeübertragungsmittel," *VGB Kraftwerkstechnik* **65** (1988) no. 5, 455 – 460.
150. L. Auffret: "Superphénix Sodium," *Nucl. Eur.* **5** (1985) no. 11.
151. A. Brandstetter, P. Engelmann, W. Koop: "SNR-300, Fast Reactor Power Station," *Conference High Temperature Reactor Process Application,* Britisch Nuclear Energy Society, London, Nov. 1974.
152. F. H. Morgenstern: "Kalkar Nuclear Power Plant (SNR-300) – a Sodium Cooled Fast Breeder Reactor Prototype," *Nucl. Technol.* **78** (1987) no. 9, 231 – 244.
153. "Superphénix," *Nucl. Eur.* **5** (1985) no. 11.
154. F. Cogne, F. Justin: "Safety of the Creys-Malville Plant (Superphénix)," *Nucl. Eur.* **5** (1985) no. 11.
155. A. Mergui, R. Casini, A. Plessa: "Engineering, Design and Construction of Superphénix," *Nucl. Eur.* **5** (1985) no. 11.

156. W. K. Biermann et. al.: "Fabrication of Fuel Elements for SNR-300," *Nucl. Technol.* **78** (1987) no. 9, 278 – 283.
157. P. Clauzon, F. Crignon, J. Ladet, H. Mikailoff, J. R. Pajot: "The Superphenix Core: Fuel Design and Fabrication," *Nucl. Eur.* **5** (1985) no. 11, 35 – 36.
158. M. Köhler, H. Rothfuß: "Der Reaktortank des SNR-300," *Kerntechnik* **15** (1973) 2.
159. H. Rothfuß, F. Vogt: "Reactor Vessel System (SNR-300)," *Nucl. Technol.* **78** (1987) no. 9, 245 –254.
160. D. Althaus, N. Brahy: "Refueling System Fabrication and Testing for SNR-300, Sodium-Cooled Fast Reactor," *Eng. Proc. Symp.* (1970) 771 – 786.
161. J. Villeneuve, F. Granito: "The Superphenix NSSS (Nuclear Steam Supply System)," *Nucl. Eur.* **5** (1985) no. 11.
162. A. Brandstetter, S. Dreyer: "The Sodium Systems of the SNR-300," *Kerntechnik* **16** (1973) no. 3, 115 – 120.
163. K. Vinzens, H. Kappauf: "Primary and Secondary Main Piping Systems for SNR-300," *Nucl. Technol.* **78** (1987) no. 9, 255 – 261.
164. A. J. van't Hoft: "Fabrication of Intermediate Heat Exchangers, Steam Generators and Sodium Pumps for SNR-300," *Nucl. Technol.* **78** (1987) no. 9, 262 – 277.
165. H. Stehle, G. Hansen, W. Jansing: "Versuchsprogramm zur Wärmesenke der passiven Nachwärmeabfuhr zukünftiger Brutreaktoren," *Jahrestagung Kerntechnik 1988,* Deutsches Atomforum.
166. K. Dumm, W. Jansing, W. Roßbach: "Entwicklung eines Nachwärmeabfuhrsystems mit Naturumlauf für zukünftige Brutreaktoren," *ATW Atomwirtsch. Atomtech.* **33** (1988) 481 – 483.
167. W. Dietz, P. Paetz, K. Kummerer: "Werkstoffe für Schnelle Brutreaktoren," *ATW Atomwirtsch. Atomtech.* **31** (1986) no. 3, 137 – 142.
168. L. Silvermann: *The Determination of Impurities in Nuclear Grade Sodium Metal,* Pergamon Press, Oxford 1971.
169. O. J. Foust (ed.): *Sodium-NaK Engineering Handbook,* Gordon & Breach Science Publ, New York 1972.
170. W. Haubold, K. Ch. Stade, H. H. Stamm, *VGB Kraftwerkstechnik* **55** (1975) 94 – 104.
171. C. C. McPheeters, J. C. Biery, *Nucl. Appl. Technol.* **6** (1969) 573 – 580.
172. J. Jung, *J. Nucl. Mater.* **56** (1975) 213 – 220.
173. L. Burris et al., USAEC-Rep. ANL/ST-6, 1971.
174. J. F. M. Rhode et al., *J. Nucl. Energy* **24** (1970) 503 – 508.
175. *Liquid Metals Handbook,* Sodium-NaK Supplement, USAEC Rep. TID-5277, 1955.
176. Steam Generator for LMFBR's, *IAEA-Study Group Meeting,* Bensberg, FRG, 14 – 17 Oct, 1974.
177. J. Olivier et al., *Proc. 4th Intern. Conf. Liquid Metal Eng. & Technology,* Avignon, 17 – 21 Oct. 1988, p. 303.
178. G. Koch, *Radiochim. Acta* **37** (1984) 205.
179. J. Baumier, L. Duchatelle: "The Case of Developing the LMFBR in France" in L. G. Brookes, H. Motamen (eds.): *The Economics of Nuclear Energy,* Chapman & Hall, London 1984.
180. F. Weigel: "Uranium" in J. J. Katz, G. T. Seaborg, L. R. Morss (eds.): *The Chemistry of the Actinide Elements,* **vol. 1,** Chapman & Hall, London 1986, p. 187.
181. *Uranium Resources, Production and Demand 1989,* OECD and IAEA Report, Paris 1990.
182. *ATW Atomwirtsch. Atomtech.* **34** (1989) 382.
183. *Nuclear Energy Data/Données sur l'Energie Nucléaire 1990,* OECD, Paris 1990.
184. K. H. Beckurts, T. J. Conolly, U. Hansen, W. Jaek: "Nuclear Energy Futures," *11th World Energy Conference,* **vol. 4 B,** München 1980, p. 239.
185. Deutsches Atomgesetz, Bundesgesetzblatt § 9 a. no. 41/1985, Part I, p. 1565.
186. *The Economics of the Nuclear Fuel Cycle,* OECD, Paris 1985.
187. J. Couture, J. C. Guais: "Economic Aspects of the Back End: Reprocessing or Once-Through" in N. Geary (ed.): *Nuclear Technology International 1987,* Sterling Publications, London 1987, p. 137.
188. OECD/IAEA: *Uranium-Resources, Production and Demand,* Paris 1988.
189. OECD/IAEA: *Uranium-Resources, Production and Demand,* Paris 1989, unpublished.
190. OECD/IAEA: *World Energy Conference 1980,* Paris.
191. F.-J. Dahlkamp: "Entwicklung und Zustand der Uranbergbauindustrie," *BHM Berg-Huettenmaenn. Monatsh.* **134** (1989) no. 4.
192. *Gmelin,* **Suppl. vol. A 3,** 1981.
193. Deutsches Atomforum: *Brennstoffkreislauf,* Bonn 1981.

194. Informationskreis Kernenergie: *Das neue Kernenergie-Seminar,* Bonn 1981.
195. KTG Subgroup Fuel Elements: Conference Transactions Fabrication and Quality Assurance of Uranium and Mixes Oxide Fuel for LWRs, 1995 FZK.
196. H. U. Siebert et al.: "Statistische Prozeßregelung und moderne Methoden des Qualitätsmanagements in der Brennelement-Fertigung," Jahrestagung Kerntechnik 1997, KTG/INFORUM, Fachsitzungsband.
197. P. Urban, M. Blank: "Extending fuel burn-ups: status and prospects," *Nuclear Europe Wordscan* **3–4** (1997) 40.
198. S. G. Brandberg, *Nucl. Technol.* **18** (1973) 177–184.
199. M. Benedict, T. Pigford, H. Levi: "Konversion," in: *Nuclear Chemical Engineering,* Chap. 9.8, McGraw Hill, New York 1981, p. 274.
200. R. E. Lerch, R. E. Norman: "Nuclear Fuel Conversion and Fabrication Chemistry," *Radiochim. Acta* **36** (1984) 75.
201. H. Assmann, H. Bairiot: "Process and Product Control of Oxide Powder and Pellets for Reactor Fuel," IAEA STI-DOC 10/221, 1983, p. 149 ff.
202. C. S. Choi et al.: "The Influence of AUC-Powder on Characteristics of UO_2-Pellets," *J. Nucl. Mater.* **153** (1988) 148–155.
203. "IAEA-TECDOC-996:" *Waterside Corrosion of Zirconium Alloys in Nuclear Power Plants* 1998.
204. R. C. Dahlberg, R. F. Turner, W. V. Goeddel: "Core Design Characteristics," *Nucl. Eng. Int.* **14** (1969) 1073.
205. R. F. Turner, W. V. Goeddel, E. O. Winkler: "HTGR Fuel Design GA-9743 (1970)," *Symp. Sol-Gel Process Reactor Fuel Cycles,* Gatlinburg 1970, pp. 31–48.
206. Japan Atomic Energy Research Institute (JAERI): *Present Status of HTGR Research and Development,* Jan. 1989.
207. K. G. Hackstein, W. Heit, W. Theymann, G. Kaiser: "Stand der Brennelementtechnologie für Hochtemperatur-Kugelhaufenreaktoren," *Atomkernenerg./Kerntech.* **47** (1985) 163–169.
208. W. Heit, H. Huschka, W. Rind, G. Kaiser: "Status of Qualification of High-Temperature Reactor Fuel Element Spheres," *Nucl. Technol.* **69** (1985) 44–54.
209. A. W. Mehner et al.: "Spherical Fuel Elements for Advanced HTR Manufacture and Qualification by Irradiation Testing, "*J. Nucl. Mater.* **171** (1990) 9–18.
210. A. S. Chernikov et al.: "The Construction and Main Characteristics of Fuel Elements for HTGR," *Proc. Tech. Comm. Meet. Gas-Cooled Reactors,* Minsk 1981.
211. A. S. Chernikov et al.: "Spherical Fuel Elements and Some Results of Their Tests," *Proc. Conf. Gas-Cooled Reactors Today,* Bristol 1982, pp. 143–148.
212. A. S. Chernikov, L. N. Permyakov, L. I. Mikhailichenko, S. D. Kurbakov: "Characteristics of HTGR Spherical Fuel Elements," *Proc. Tech. Comm. Meet. Gas-Cooled Reactors and Their Applications,* Jülich 1986, pp. 445–460; IAEA-TECDOC-436.
213. E. Teuchert et al.: "Closed Thorium Cycles in Pebble Bed HTR," KFA Report Jul-1569 (1979).
214. R. F. Turner: "Economic Performance of HTGR Fuel Cycles Under Various Recycling and Economic Conditions," GA-A 13 791 (1976).
215. R. C. Dahlberg, R. F. Turner, W. V. Goeddel: "HTR Fuel and Fuel Cycle," Summary Description, GA-A 12801 (1974).
216. M. H. Merrill: "Use of the Low Enriched Uranium Cycle in the HTGR," GA-A 14340 (1977).
217. S. Brandes, G. Lohnert: "On the Flexibility of High Temperature Reactor Cores for High- and Low-Enriched Fuel," *Nucl. Technol.* **58** (1982) 23–28.
218. T. D. Gulden, H. Nickel: "Coated Particle Fuel," *Nucl. Technol.* **35** (1977) 205.
219. H. Huschka, P. Vygen: "Coated Fuel Particles: Requirements and Status of Fabrication Technology," *Nucl. Technol.* **35** (1977) 238.
220. R. J. Price: "Properties of Silicon Carbide for Nuclear Fuel Particle Coatings," *Nucl. Technol.* **35** (1977) 320.
221. M. Hrovat, H. Nickel, H. Koizlik: "über die Entwicklung eines Matrixmaterials zur Herstellung gepreßter Brennelemente für Hochtemperatur-Reaktoren," KFA Report Jül-969 (1973).
222. R. E. Schulze, H. A. Schulze, W. Rind: "Graphitic Matrix Materials for Spherical HTR Fuel Elements," KFA Report Jül-Spez-167 (1982).

223. M. Kadner, J. Baier: "Production of Fuel Kernels for HTR Fuel Elements," *Kerntechnik* **18** (1976) 413 – 420.
224. A. S. Chernikov et al.: "Spherical Coated Particle Fuel for Fuel Elements of HTGR," *IAEA Specialists' Meeting on Gas-Cooled Reactor Fuel Development and Spent Fuel Treatment,* Moscow 1983, pp. 113 – 132.
225. F. W. v. d. Brugghen et al.: "A U (VI) – Process for Microphere Production," *Symp. Sol-Gel Process Reactor Fuel Cycles,* Gatlinburg 1970, pp. 253 –263.
226. R. Förthmann, H. Nickel, A. Naoumidis, W. Burck: "Investigations on the Preparation of UO_2 Microspheres by Internal Gelation of a UO_2 Sol and from a Uranium (VI) Solution," *Symp. Sol-Gel Process Reactor Fuel Cycles,* Gatlinburg 1970, pp. 551 – 566.
227. R. Förthmann: "Die chemischen Grundlagen des Hydrolyseverfahrens zur Herstellung sphärischer Kernbrennstoffteilchen," KFA Report Jül-950 (1973).
228. H. W. H. Lahr: "Fabrication, Properties and Irradiation Behavior of U/Pu Particle Fuel for Light Water Reactors," *Nucl. Technol.* **31** (1976) 183 –190.
229. R. L. Lefevre, M. S. T. Price: "Coated Nuclear Fuel Particle: The Coating Process and its Model," *Nucl. Technol.* **35** (1977) 263 – 278.
230. W. Warzawa, W. Heit, H. Huschka: "Coated Particles – Entwicklungsstand," *Jahrestagung Kerntechnik 1980,* Deutsches Atomforum, pp. 578 – 581.
231. H. Nabielek et al.: "Development of Advanced HTR Fuel Elements," *Nucl. Eng. Des.* **121** (1990) 199 – 210.
232. A. S. Chernikov, Yu. V. Koshelev, L. I. Mikhailichenko, A. A. Kuznetsov: "HTGR Spherical Fuel Elements and Their Experimental Development," *IAEA Specialists' Meeting on Gas-Cooled Reactor Fuel Development and Spent Fuel Treatment,* Moscow 1983, pp. 113 – 132.
233. V. F. Zelenskiy, V. A. Gurin, S. S. Keruchen'ko, Yu. F. Konotop: "HTGR Spherical Fuel Elements of the Monolithic Type Bound by Pyrocarbon," *IAEA Specialists' Meeting on Gas-Cooled Reactor Fuel Development and Spent Fuel Treatment,* Moscow 1983, pp. 106 – 112.
234. R. C. Noren: US HTGR Fuel Manufacturing, GA-A 19 479 (1988).
235. S. Yoshimuta, N. Suzuki, M. Kaneko, K. Fukuda: "Production Process and Quality Control System for the HTTR Fuel," *IAEA Specialists' Meeting on Behavior of Gas Cooled Reactor Fuel Under Accident Condition,* Oak Ridge, 5 – 8 Nov. 1990.
236. OECD Nuclear Energy Agency 1997 Management of Separated Plutonium: The Technical Options.
237. IAEA Transactions International Symposium on Nuclear Fuel Cycle and Reactor Strategies: Adjusting to new Realities 1997.
238. D. Albright, F. Berkout, W. Walker, *Plutonium and Highly Enriched Uranium,* Oxford Universtity Press 1997.
239. P. Urban, D. Bender, G. Schlosser, *The Nuclear Engineer* **37** (1996) 183.
240. G. Schlosser, W. D. Krebs, P. Urban, *Nuclear Technology* **102** (1993) 54.
241. W. Goll, H. P. Fuchs, F. Schlemmer, *Nuclear Technology* **102** (1993) 29.
242. D. Gumprecht et al.: *Strahlenexposition and Strahlengefährdung durch Plutonium,* vol. 14, G. Fischer-Verlag, Stuttgart 1989.
243. D. Vollath et al.: "On the Dissolution of $(U, Pu)O_2$ in Nitric Acid," *Nucl. Technol.* **71** (1985) 240 – 245.
244. H. Elbel, J. Klews, R. Löb: "Dependence of the UO_2-PuO_2-Pellet-Structure on Fabrication Parameters," *J. Nucl. Mater.* **153** (1988) 160 – 168.
245. J. F. Flagg (ed.): *Chemical Processing of Reactor Fuels,* Academic Press, London 1961.
246. S. M. Stoller, R. B. Richards (eds.): "Fuel Reprocessing," in *Reactor Handbook,* vol. II, Interscience, New York 1961.
247. *Gmelin,* **Transuranium Elements, A 1, Suppl. vol. 7 b,** 177.
248. G. Koch, *Chem. Ztg.* **101** (1977) 64.
249. W. B. Lanham, T. C. Runion: "Purex Process for Plutonium and Uranium Recovery," *USAEC Rep* **ORNL-479** (1949).
250. W. W. Schulz, L. L. Burger, J. D. Navratil (eds.): *Science and Technology of Tributyl Phosphate,* **"Applications of Tributyl Phosphate in Nuclear Fuel Reprocessing," vol. III,** CRC Press, Boca Raton 1990.
251. G. Koch: "Existing and Projected Reprocessing Plants: a General Review," *Atomkernenerg./Kerntech.* **33** (1979) 241.
252. *Gmelin,* **Transuranium Elements, A 1, Suppl. vol. 7 b,** 240, 251.
253. R. K. Steunenberg, R. C. Vogel: "Fluoride and other Halide Volatility Processes," in *Reactor Handbook,* vol. II, Interscience, New York 1961, p. 250. H. M. Feder, I. G. Dillon: "Pyrometallurgical Processes," in *Reactor Handbook,* vol. II, Interscience, New York 1961, p. 313.

254. J. van Caeneghem, Y. Marchant, E. Détilleux: "Dissolution du zirconium dans une solution fluorure de sodium – acide citrique," in *Aqueous Reprocessing Chemistry for Irradiated Fuels, OECD Symposium, Brussels 1963,* Paris 1963, p. 313.
255. *Reprocessing News,* 1990no. 15, 5.
256. B. Herrmann: "Auflösung unbestrahlter UO_2-Pellets in Salpetersäure," *Kernforschungszent. Karlsruhe Ber. KfK-3673* (1983).
257. B. Herrmann, W. Bumiller, E. Henrich, R. v. Ammon, *KfK-Nachr.* **16** (1984) 87.
258. J. L. Ryan, L. A. Bray: "Dissolution of Plutonium Dioxide – A Critical Review," in J. D. Navratil, W. W. Schulz (eds.): "Actinides Separations," *ACS Symp. Ser.* **117** (1980) 499.
259. L. A. Bray, J. L. Ryan: "Catalyzed Electrolytic Dissolution of Plutonium Dioxide" in J. D. Navratil, W. W. Schulz (eds.): *Actinide Recovery from Waste and Low-grade Sources,* Harwood, Chur 1982, p. 129.
260. H. Kleykamp: "Der chemische Zustand von AUPuC- und OKOM-Mischoxid in verschiedenen Stadien des Brennstoffkreislaufs," *Kernforschungszent. Karlsruhe Ber. KfK-4430,* 1988.
261. H. Kleykamp, *J. Nucl. Mater.* **171** (1990) 181.
262. B. Lorrain, D. Saudray, M. Tarnero: "Dissolveur continu rotatif pour combustibles eau legère," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage. 1987,* vol. 4, Paris 1987, p. 1471. Commissariat à l'Energie Atomique, Europ., EP 0 030 884,1979 (J. Auchapt-Tavenaux, P. Auchapt, H. Sauvage, M. Tarnero).
263. *Gmelin,* **Uranium, D 2.**
264. *Gmelin,* **Transuranium Elements, D 2, Suppl. vol. 21,** 177.
265. G. Petrich, Z. Kolarik: "The 1981 Purex Distribution Data Index," *Kernforschungszent. Karlsruhe Ber. KfK-3080,* 1981.
266. V. Drake: "Extraction Chemistry of Neptunium," in [250], p. 123.
267. W. O. Haas: "Solvent Extraction: General Principles," in [245], p. 125.
268. A. L. Mills: "Computation and Modelling," in [250], p. 181.
269. D. J. Pruett: "Extraction Chemistry of Fission Products," in K. G. Hackstein, [250], p. 81.
270. G. Koch: "Das PUREX-Schema und seine Variationsbreiten," in F. Baumgärtner (ed.): *Chemie der Nuklearen Entsorgung,* Pt. II, Thiemig, München 1978.
271. J. L. Swanson: "PUREX Process Flowsheets," in [250], p. 55.
272. G. F. Vandegrift: "Diluents for TBP Extraction Systems," in W. W. Schulz, J. D. Navratil, A. E. Talbot (eds.): *Science and Technology of Tributyl Phosphate,* vol. I, CRC Press, Boca Raton 1984, p. 69.
273. F. Baumgärtner, H. Schmieder, *Radiochim. Acta* **25** (1978) 191.
274. H. Schmieder, H. Goldacker: "Technical Scale Electroredox Equipment for Separation of Plutonium," *Proc. Int. Solvent Extr. Conf. ISEC 1986,* vol. I, Dechema, München 1986, p. 117.
275. J. H. Miles: "Separation of Plutonium and Uranium," in [250], p. 11.
276. L. Stieglitz, R. Becker: "Chemical and Radiolytic Degradation in the Purex Process," *Proc. Int. Nucl. Fuel Process. Waste Manage. 1984,* Jackson, Wyo. 1984, p. 1 – 451.
277. H. Goldacker, H. Schmieder, F. Steinbrunn, L. Stieglitz, *Kerntechnik* **18** (1976) 426.
278. W. W. Schulz, G. E. Benedict: "Neptunium-237, Production and Recovery," *USAEC* **Rep. TID-25 955** (1972).
279. *Gmelin,* **Transuranium Elements, A 1, Suppl. vol. 7 b,** 288.
280. A. Brunstad, *Ind. Eng. Chem.* **51** (1959) 38.
281. D. A. Orth, J. M. McKibben, W. C. Scotten: "Progress in Tributyl Phosphate Technology at the Savannah River Plant," *Proc. Int. Solvent Extr. Conf. ISEC 1971,* **vol. 1,** Soc. Chem. Ind., London 1971, p. 514.
282. P. Patigny, J. Regnier, P. Miquel, D. Taillard: "Utilisation of Hydroxylamine Nitrate for the Final Concentration of Plutonium in the Irradiated Fuel Reprocessing Factory at La Hague," *Proc. Int. Solvent Extr. Conf. ISEC 1974,* **vol. 3,** Soc. Chem. Ind., London 1974, p. 2019.
283. H. Goldacker: "Design and Evaluation of Extractors in the First Cycle of a Fast Breeder Reactor Fuel Reprocessing Plant," in F. Baumgärtner, K. Ebert, E. Gelfort, K. H. Lieser (eds.): *Nuclear Fuel Cycle,* vol. 2, Verlag Chemie, Weinheim 1983, p. 177.
284. J. A. Coste, D. Mathieu, G. Bernier: "Colonnes pulsées annulaires – Résultats d'essais," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage., 1987,* vol. 4, Soc. Française d'Energie Nucléaire, Paris 1987, p. 1529.
285. M. Delange: "LWR Spent Fuel Reprocessing at La Hague: Ten Years On," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage. 1987,* vol. 1, Soc. Française d'Energie Nucléaire, Paris 1987, p. 228.

286. W. Hebel, G. Cottone (eds.): *Management Modes for Iodine-129,* Harwood, Chur 1982.
287. W. Hebel, G. Cottone (eds.): *Methods of Krypton-85 Management,* Harwood, Chur 1983.
288. E. Henrich, F. Weirich: "Development of a Cryogenic Absorption Process for Rare Gas Removal From Reprocessing Off-Gas," *Proc. DOE/NRC Nuclear Air Cleaning Conf., 20th 1988,* **vol. 1,** Boston 1988, p. 154.
289. N. Boukis, E. Henrich, *Radiochim. Acta* **55** (1991) in press.
290. *Proc. Conf. Remote Syst. Technol. 1st 1953 – 38th 1990.*
291. R. G. Jaeger (ed.): *Engineering Compendium on Radiation Shielding,* **vol. I – III,** Springer Verlag, Berlin 1968 – 1975.
292. J. M. McKibben, J. B. Starks, J. K. Brown: "The Savannah River Plant Purex Plant – 25 Years of Successful Remote Operation," *Proc. Conf. Remote Syst. Technol., 27th* 1979, p. 173.
293. J. D. McIntosh, G. J. Raab: "Remote Maintenance in a Large Scale Separations Plant," *Proc. Conf. Remote Syst. Technol. 14th* 1966, p. 196. J. F. Kane, W. H. Koontz: "Major Repair and Modification Techniques used on Large Scale Separation Facilities," in *Proc. Conf. Remote Syst. Technol. 14th* 1966, p. 203.
294. Hanford Atomic Products Operation (ed.): "Purex Technical Manual," *USAEC* **Rep. HW-31 000** (1955), Chap. XI.
295. F. Chenevier, C. Bernard, J. P. Giraud: "Design and Construction of the New Reprocessing Plants at La Hague," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage. 1987,* **vol. I,** Soc. Française d'Energie Nucléaire, Paris 1987, p. 97.
296. *Gmelin,* **Transuranium Elements, A 2, Suppl. vol. 8,** 361.
297. W. Thomas, W. Weber: *Handbuch zur Kritikalität,* Universität München, Garching 1970 – 1972.
298. Metallgesellschaft and Kernforschungszentrum Karlsruhe, DE 2 520 870, 1975 (M. Poetzschke, H. Schmieder, E. Warnecke).
299. The Text of the Agreement between Belgium, Denmark, the Federal Republic of Germany, Ireland, Italy, Luxembourg, The Netherlands, the European Atomic Energy Community and the Agency in Connection with the Treaty on the Non-Proliferation of Nuclear Weapons, *INFCIRC/193,* IAEA, Vienna 1973.
300. Commission Regulation (Euratom) no. 3227/76 of 19th Oct. 1976 Concerning the Application of the Provisions on Euratom Safeguards, *Official Journal of the European Community* **19** (1976) no. L 363.
301. *Nucl. Safeguards Technol. Proc. Symp. Nucl. Mater. Safeguards, 1986,* IAEA, Vienna, Nov. 1986.
302. H. Schmieder, G. Petrich, *Radiochim. Acta* **48** (1989) 181.
303. G. Petrich: "The Purex Computer Model VISCO and its Applications," *Proc. Int. Solvent Extr. Conf. ISEC 1986,* **vol. I,** Dechema, München 1986, p. 31.
304. U. Galla, J. Schön, H. Schmieder: "U/Pu Separation in the PUREX Process: Experimental Comparison of Different Procedures," *Proc. Int. Solvent Ext. Conf. ISEC 1990,* Elsevier, Amsterdam 1991.
305. G. Petrich et al.: "Progress in the Development of a One-Cycle PUREX Process (IMPUREX)," *Proc. Int. Solvent Extr. Conf. ISEC 1990,* Elsevier, Amsterdam 1991.
306. E. Henrich, H. Schmieder, K. Ebert: "Combination of TBP Extraction and Nitrate Crystallisation for Spent Nuclear Fuel Reprocessing," in "Extraction 87: The Recovery of High Value Materials," *Inst. Chem. Eng. Symp. Ser.* **103** (1987) p. 191.
307. A. L. Mills: "The Dounreay Prototype Fast Reactor Fuel Reprocessing Scheme – Flowsheet and Chemistry," in: *Fast Reactor Fuel Reprocessing,* Soc. Chem. Ind., London 1980.
308. A. S. Nikiforov et al.: "Modification of Aqueous Reprocessing Technology to Reduce External Nuclear Fuel Cycle," *Proc. Int. Solvent Extr. Conf. ISEC 1988,* **vol. IV,** Former USSR Acad. Sci., Moscow 1988, p. 168.
309. A. Calame-Longjean, G. Revol, J. P. Roux, G. Ranger: "Renovation et extensions pour la R et D de l'Atelier Pilote de Marcoule," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage., 1987,* vol. I, Paris 1987, p. 115.I. M. Farruga, G. Malterre, M. Tachon, G. Polydor: "Remise en état après 20 ans d'exploitation et rédemmarrage de l'Atelier Pilot de Retraitement Marcoule," ibid., vol. 4, p. 1405.
310. L. Burris, R. K. Steunenberg, W. E. Miller: "The Application of Electrorefining for Recovery and Purification of Fuel Discharged from the Intergral Fast Reactor," in R. E. White, R. F. Savinell, A. Schneider (eds.): "Electrochemical Engineering Applications," *AIChE Symp. Ser.,* **83** (1987) no. 254, 135.

311. M. J. Lineberry, R. D. Phipps: "A Demonstration Facility for the IFR Fuel Cycle," *Proc. Int. Conf. Nucl. Fuel Reprocess. Waste Manage., 1987,* **vol. I,** Paris 1987, p. 153.
312. Standardization of Radioactive Waste Categories, *Tech. Rep. Ser. I.A.E.A.* 1970, no. 101.
313. Underground Disposal of Radioactive Wastes – Basic Guidance, *Saf. Ser. I.A.E.A.* 1981, no. 54.
314. Radioactive Waste Management Glossary, IAEA Technical Document 264, 1982.
315. H. Dyroff, F. K. Fleischmann: "Anfall radioaktiver Abfälle in der Bundesrepublik Deutschland," in: *Radioactive Abfälle in der Bundesrepublik Deutschland,* KWA 1214, vol 1, 1976 – 1977.
316. Plan – Endlager für radioaktive Abfälle – Schachtanlage Konrad, Salzgitter, Sept. 1986, Physikalisch-Technische Bundesanstalt, Braunschweig 1986.
317. E. Warnecke, P. Brennecke, G. G. Eigenwilig, F. Heil, *Proc. IAEA Symp. Conditioning of Radioactive Wastes for Storage and Disposal,* IAEA, Vienna 1983, p. 391.
318. P. Brennecke et al., *Energiewirtschaftliche Tagesfragen* **34** (1984) 767 – 770.
319. H. P. Berg, D. Ehrlich, H. Illi, B. Thomauske, *Proc. Semin. Radioactive Waste Products – Suitability for Final Disposal,* Jülich 1985, p. 512.
320. P. Brennecke, E. Warnecke (eds.): *Requirements on Radioactive Wastes for Disposal (Preliminary Waste Acceptance Requirements April 1990) – Konrad Repository*, Bundesamt für Strahlenschutz, BfS ET-4/90, Salzgitter 1990.
321. "Richtlinie zur Kontrolle radioaktiver Abfälle mit vernachlässigbarer Wärmeentwicklung, die nicht an eine Landessammelstelle abgeliefert werden," *Bundesanzeiger* **41** (1989) no. 63.
322. E. Warnecke, P. Brennecke, A. Hollmann: *Einteilung radioaktiver Abfälle und Abfallartenkatalog,* **PTB-SE-IB-33,** Physikalisch-Technische Bundesanstalt, Braunschweig 1988.
323. *Radioactive Glasses – Research and Testing,* Commissariat à l'Energie Atomique Group, Paris 1981.
324. K. D. Kuhn, W. Kunz, W. Grünewald, *International Waste Management Conference,* **23 – 28 Oct. 1989, Kyoto, Japan, vol. 2,** ASME, New York 1989, p. 111.
325. A. E. Ringwood et al.: *Geochem. J.* **13** (1979) 141 – 165.
326. *Management of Cladding Hulls and Fuel Hardware, Tech. Rep. Ser. I.A.E.A.* 1985, no. 258.
327. R. Krämer et al., *Proc. ANS Spectrum '86, Waste Management and Decontamination and Decommissioning,* Seattle 1986, p. 414.
328. E. Passig, H. Hepp, T. F. Kienle, D. Rittscher, *Proc. IAEA-Conf. Radioactive Waste Management,* IAEA, Vienna 1984, p. 129.
329. K. D. Closs et al., *International Waste Management Conference,* **23 – 28 Oct. 1989, Kyoto, Japan, vol. 2,** ASME, New York 1989, p. 417.
330. E. Warnecke, K. Ermisch in K. Ebert, R. v. Ammon (eds.): *Safety of the Nuclear Fuel Cycle,* VCH Verlagsgesellschaft, Weinheim 1989, p. 297.
331. W. Kessels et al., *Proc. IAEA Symp. Patram '86,* **vol. 2,** IAEA, Vienna 1987, p. 11.
332. H. Schmieder, K. L. Huppert in F. Baumgärtner (ed.): *Chemie der Nuklearen Entsorgung,* Part II, Thiemig, München 1978, p. 50.
333. W. Bähr: *Entsorgung in Wiederaufarbeitungsanlagen,* ecomed Verlagsgesellschaft, Landsberg 1984.
334. E. Warnecke, H. Schmieder, W. Comper in F. Baumgärtner (ed.): *Chemie der Nuklearen Entsorgung,* Part II, Thiemig, München 1978, p. 333.
335. *Conditioning of Low- and Intermediate-Level Radioactive Wastes, Tech. Rep. Ser. I.A.E.A.* 1983, no. 222.
336. *Options for the Treatment and Solidification of Organic Radioactive Wastes, Tech. Rep. Ser. I.A.E.A.* 1989, no. 294.
337. "Immobilization of Low and Intermediate Level Radioactive Wastes with Polymers," *Tech. Rep. Ser. I.A.E.A.* 1988, no. 289.
338. P. Brennecke in L. Lamberg (ed.): *Proc. 3rd Finnish – German Seminar on Nuclear Waste Management 1986,* Valtion Technillinen Tutkimuskeskus (VTT), Espoo 1988, p. 49.
339. P. Brennecke, J. Schumacher: *Anfall radioaktiver Abfälle in der Bundesrepublik Deutschland – Abfallerhebung für das Jahr 1989,* BfS-Bericht ET-1/90, Salzgitter 1990.
340. P. Brennecke, J. Schuhmacher, E. Warnecke, *Nucl. Eur.* **8** (1988) 34 – 35.
341. "Sicherheitskriterien für die Endlagerung radioaktiver Abfälle in einem Bergwerk," *Bundesanzeiger* **35** (1983) 45 – 46.
342. K. Kühn, *Kerntech.* **51** (1987) 157 – 160.

343. R. Stippler, H. Kleimann, *Bericht 7. IRPA-Regionalkongress Radioaktive Abfälle,* Würenlingen, Switzerland 1980, p. 261.
344. R. Masters, *Nucl. Eng. Int.* **32** (1987) 50 – 56.
345. H. P. Berg, P. Brennecke: *The Konrad Mine – The Planned German Repository for Radioactive Waste with Negligible Heat Generation,* BfS-Bericht ET-6/90, Salzgitter 1990.
346. Bundesamt für Strahlenschutz: *Fortschreibung des zusammenfassenden Zwischenberichts über bisherige Ergebnisse der Standorterkundung Gorleben vom Mai 1983,* BfS-Bericht ET-2/90, Salzgitter 1990.
347. Staatliches Amt für Atomsicherheit und Strahlenschutz: *Die Entsorgung von radioaktiven Abprodukten aus der Anwendung der Atomenergie in der Deutschen Demokratischen Republik,* Report SAAS-381, Berlin 1990.
348. Bundesminister des Innern: Gesetz über die friedliche Verwendung der Kernenergie und den Schutz gegen ihre Gefahren (German atomic law), version July 15, 1985 (BGBl. I, p. 1565), last amendment Feb. 18, 1986 (BGBl. I, p. 265).
349. Bundesminister des Innern: Verordnung über den Schutz durch ionisierende Strahlen (Strahlenschutzverordnung-StrlSchV), Oct. 13, 1976 (BGBl. I, p. 2905;Report 1977, p. 184, 269), last amendment Jan. 8, 1987 (BGBl. I, p. 114).
350. Bundesminister des Innern: Verordnung über Dekkungsvorsorge nach dem Atomgesetz (Atomrechtliche Deckungsvorsorge-Verordnung-AtDeckV), Feb. 25, 1977 (BGBl. I, p. 220).
351. R. Gerstler, G. Stein, R. Weh: "Internationale Kernmaterialkontrollen für Wiederaufarbeitungsanlagen," *ATW Atomwirtsch. Atomtech.* **30** (1985) 330 – 335.
352. H. Böhm: "Sicherheit der Entsorgung," *ATW Atomwirtsch. Atomtech.* **28** (1983) 518 – 522.
353. E. Merz: *Behandlung radioaktiver Abfälle im Kernbrennstoffkreislauf,* Bericht JüL-Spez-394 (1987) ISSN no. 0343-7639.
354. ICRP (ed.): "Recommendations of the International Commission on Radiological Units (London)," Revised by the International Commission on Radiological Protection at the 6th International Congress of Radiology, London 1950," *Radiology (Easton, Pa)* **56** (1951) 117 – 119; *Br. J. Radiol.* **24** (1951) 46.
355. F. D. Sowby: "ICRP Dose Limits from 1934 up to 1977." *Radiol. Protect. Bull.* (1979) no. 28, 11.
356. ICRP (ed.): *Publication 12: General Principles of Monitoring for Radiation Protection of Workers.* A Report by Committee 4 of the International Commission on Radiological Protection (adopted by the Commission on May 24, 1968) Pergamon Press, Oxford 1968.
357. ICRP (ed.): *Publication 22: Implications of Commission Recommendations that Doses be Kept as Low as Readily Achievable,* A Report of ICRP Committee, Pergamon Press, Oxford 1973.
358. ICRP (ed.): Radiation Protection Principles for the disposal of solid radioactive waste. *Ann. ICRP* **15** (1985) no. 4. NEA (ed.): *Longterm Radiation Protection Objectives in Radioactive Waste Disposal,* OECD Nuclear Energy Agency, Paris 1984.

General References

359. F. F. Chen: *Introduction to Plasma Physics and Controlled Fusion,* Plenum Press, New York 1984.
360. G. Schmidt: *Physics of High Temperature Plasmas,* Academic Press, London 1979.
361. J. Wesson: *Tokamaks,* Clarendon Press, Oxford 1987.
362. R. A. Gross: *Fusion Energy,* John Wiley & Sons, New York 1984.
363. W. M. Stacey, Jr.: *Fusion. An Introduction to the Physics and Technology of Magnetic Confinement Fusion,* John Wiley & Sons, New York 1984.
364. L. A. Artsimovich: *Controlled Thermonuclear Reactions,* Gordon and Breach, New York 1982.
365. T. J. Dolan: *Fusion Research,* Pergamon Press, New York 1982.
366. R. D. Gill: *Plasma Physics and Nuclear Fusion Research,* Academic Press, London 1981.
367. J. Raeder et al.: *Kontrollierte Kernfusion,* B. G. Teubner, Stuttgart 1981.
368. E. Teller (ed.): *Fusion,* **vol. 1**Part A and Part B, Academic Press, New York 1981.

Nucleic Acids

Genetic Engineering is a separate keyword.

HELMUT BURTSCHER, Boehringer Mannheim GmbH, Penzberg, Federal Republic of Germany (Chaps. 1 – 6)

SIBYLLE BERNER, Boehringer Mannheim GmbH, Tutzing, Federal Republic of Germany (Chap. 7)

RUDOLF SEIBL, Boehringer Mannheim GmbH, Penzberg, Federal Republic of Germany (Chap. 8)

KLAUS MÜHLEGGER, Boehringer Mannheim GmbH, Tutzing, Federal Republic of Germany (Chap. 9)

1. Introduction 1
2. Structure 2
2.1. Structure of DNA 2
2.2. Structure of RNA 5
3. Properties 6
3.1. Physical and Chemical Properties . 6
3.2. Interaction with Proteins 7
4. Biosynthesis and Biological Function 7
4.1. DNA Replication 7
4.2. Gene Expression 8
4.2.1. Transcription 9
4.2.2. Translation 10
4.3. Modification and Degradation . . . 10
4.4. Recombination 11
4.5. DNA Repair 11
4.6. Nucleic Acids as Enzymes 11
5. Isolation, Purification, and Transfer 12
6. Analysis of Nucleic Acids 12
7. Chemical Synthesis 14
7.1. Synthesis Strategy 14
7.2. Protecting Groups 16
7.3. Functionalization of the Support . . 17
7.4. Methods of Synthesis 17
7.5. Cleavage of Protecting Groups and Purification of Oligonucleotides . . 19
7.6. Synthesis of Modified Oligonucleotides 19
8. Uses 21
8.1. Hybridization Techniques for Nucleic Acid Detection 22
8.2. Labeling and Detection Systems . . 23
8.3. Amplification Systems 24
8.4. Applications of Probe Technology . 25
9. Nucleosides and Nucleotides 26
9.1. Nucleosides 26
9.2. Nucleotides 29
9.3. Therapeutically Important Nucleoside and Nucleotide Derivatives . . . 31
10. References 31

1. Introduction

Nucleic acids are high molecular mass compounds found in all living cells and viruses. Their name originates from their discovery in the nuclei of eucaryotic cells. They can be chemically degraded to yield phosphoric acid, pentoses, and nitrogen-containing heterocycles (bases). Nucleic acids can be divided into two main classes depending on the sugar they contain: *deoxyribonucleic acids* (DNA) contain 2-deoxy-D-ribose and *ribonucleic acids* (RNA) contain D-ribose.

Nucleic acids are long, unbranched chains of sugar and phosphate (Fig. 1, see next page): the C-3′ atom of each sugar is linked by a phosphodiester bond to the C-5′ atom of the neighboring sugar. Either a purine (adenine, guanine) or a pyrimidine (cytosine and thymine in DNA; cytosine and uracil in RNA) is attached to C-1′ of the sugar by a β-glycosidic bond. For a detailed description of purines and pyrimidines, see → Purine Derivatives, → Pyrimidine and Pyrimidine Derivatives.

Although nucleic acids have been known since the second half of the nineteenth century it was only in the 1940s that their importance as the carrier of genetic information became clear. Genetic engineering and improved physical and biochemical methods of analysis have led to enormous progress in the understanding of the structure of DNA, DNA – protein interactions, and gene organization, expression, regulation, and transfer. The importance of nucleic acids became even more obvious after the discovery that they can have other functions in addition to their ability to store and transfer genetic information. It is widely assumed that in the course

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_001

of evolution first RNA and then DNA came into being [36, 37].

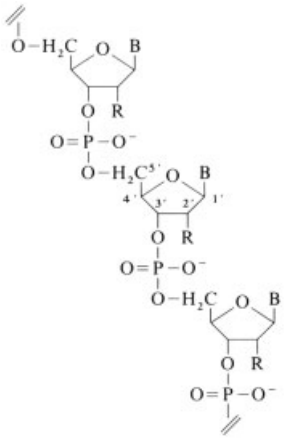

Figure 1. Structure of DNA (R = H) and RNA (R = OH)
B = base (adenine, guanine, thymine or uracil, cytosine)

2. Structure

2.1. Structure of DNA

The joining of the DNA building blocks by 5′- and 3′-phosphodiester bonds gives the molecule polarity (Fig. 1); base sequences are always written starting with the 5′-terminus, i.e., in the $5' \rightarrow 3'$ direction. The specific base sequence of DNA and its ability to form double-stranded structures according to precisely defined rules are of utmost importance for the storage of genetic information and for interactions with other nucleic acids and proteins.

From X-ray analysis data, Crick and Watson proposed a double-stranded structure for DNA in 1953 in which two antiparallel (i.e., $5' \rightarrow 3'$ and $3' \rightarrow 5'$) polynucleotide chains form a right-handed helix (i.e., looking along the axis of the helix, the strands are coiled clockwise). Naturally occurring DNA usually consists of right-handed helices with a major and a minor groove (Fig. 2). The hydrophobic bases are located inside the helix and the sugar – phosphate "backbone" on the outside [38]. Bases that are opposite each other are paired according to defined rules as a result of hydrogen bond formation: adenine always pairs with thymine or uracil and guanine with cytosine. Complementary bases can be bound by the more common Watson – Crick pairing (Fig. 3 A) or by Hoogsteen base pairing (Fig. 3 B). The double-stranded structure is further stabilized by hydration of the phosphate groups and hydrophobic interactions between the aromatic ring systems that result in stacking of the bases.

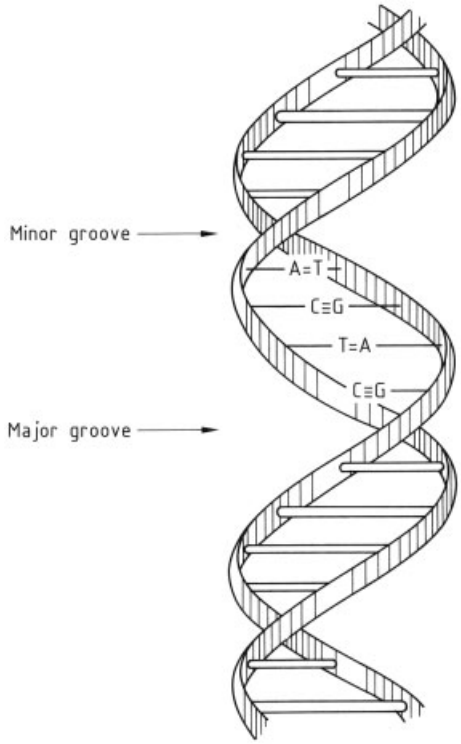

Figure 2. Right-handed double-helix of DNA
A = adenine, C = cytosine, G = guanine, T = thymine

In double-stranded DNA, the bases are densely stacked and there is a cooperative effect between hydrogen bonding and stacking. Internal bases can be continuously paired and unpaired; double-stranded regions open and form single-stranded "bubbles" ("breathing" of DNA). Breathing is more frequent in regions rich in A – T pairs and could be important for interactions with proteins. DNA helices can exist in various forms (A, B, C, D, and Z) [39] some of which are interconvertible depending on the concentration and type of salts present. The helices always exhibit a degree of microheterogeneity that plays an important part in genetic regulation mechanisms.

The DNA helices can exist in linear form (e.g., in the chromosomes of higher organisms)

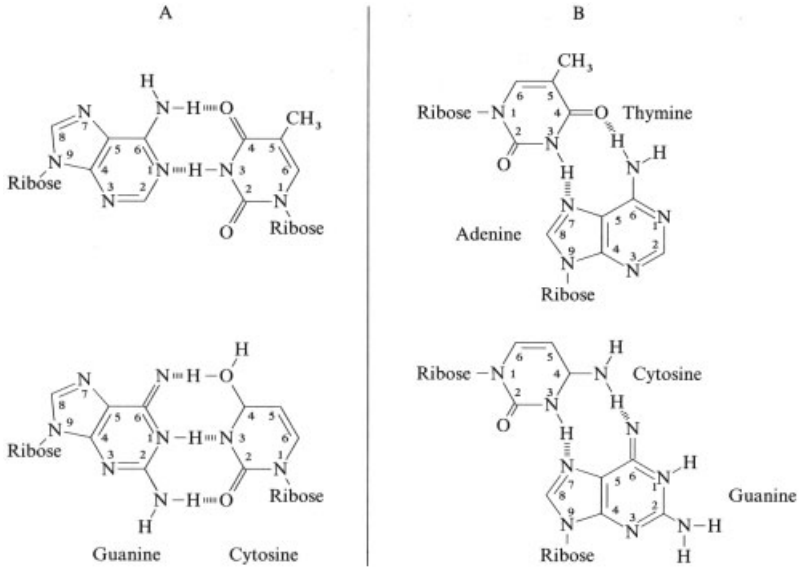

Figure 3. Watson – Crick base-pairs (A), and Hoogsteen base-pairs (B)

or as closed rings (e.g., in *Escherichia coli*); the molecules can also be twisted (superhelicity or supercoiling). In order to accommodate the large amount of DNA present in living cells, it must be packaged as compactly as possible with the help of proteins and RNA. Proteins can recognize specific binding sites on the DNA. The grooves of the DNA helix are large enough to allow proteins to come into contact with the bases [40]. Defined regions in DNA can also be recognized with the help of the methylation pattern of the bases (see Section 4.3).

Forms of DNA.

A-DNA can be observed in X-ray analyses at 66 % relative humidity. It has 11 base pairs per turn of the helix, the planes of the base pairs are tilted away from the vertical helical axis (19°), the helix is right-handed and has a diameter of ca. 2.3 nm.

B-DNA is the classical Watson – Crick form. It represents the structure of DNA at a relative humidity of > 92 % and largely corresponds to that found under physiological conditions. The helix is also right-handed with about 10.2 – 10.4 base pairs per turn and a diameter of ca. 2 nm. Single unpaired bases can be "looped out" of the helix and barely disturb the rest of the structure [41,42]. Protein – DNA interactions usually require recognition of nucleotide sequences in the major groove of the B-DNA double helix.

C-DNA helices can be observed at a relative humidity of 44 – 66 % in the presence of lithium salts. The helix is also right-handed and similar to the B form, but with 9.3 base pairs per turn.

D-DNA occurs in nature only in sequences with alternating adenine and thymine residues and in the DNA of the bacteriophage T 2 (T-DNA). The helix is also right-handed and has 8 base pairs per turn.

The left-handed conformation of *Z-DNA* has an alternating sequence of pyrimidines and purines and is formed in vitro at high salt concentrations (> 2 mol/L NaCl) or in the presence of divalent cations (Mg^{2+} > 0.7 mol/L). Unlike the right-handed helices (which have two grooves), this structure forms a single, very deep groove that penetrates the helix axis. The sugar – phosphate backbone assumes a zig-zag arrangement (therefore Z-DNA) with 12 base pairs per turn of the helix. Z-helices can form in vivo at physiological salt concentrations. They are less stable than B-DNA, but are stabilized by supercoiling, proteins, special ions, and methylation [43]. Torsional stress of DNA in vivo can favor the formation of Z-DNA [44]. Z-DNA and B-DNA are interconvertible; part of a DNA molecule may exist in the B form and another part in the Z form.

Supercoiling. Circular DNA and DNA between fixed sites can be twisted to supercoils. The term *supercoiling* refers to the curvature of the double helix axis. Supercoiled (superhelical) DNA was discovered in the 1960s in polyoma virus [45]. Rotation in the direction of winding is called *positive supercoiling* and rotation in the opposite direction is called *negative supercoiling*.

Torsional stress due to negative supercoiling can be overcome by the formation of DNA structures other than the B form. Negative supercoiling is a strong driving force for the stabilization of Z-DNA. Supercoiling makes DNA more compact, which is very important in DNA packaging. Almost all naturally occurring superhelical DNAs are underwound (i.e., have negative superhelices) but overwound DNAs also exist [46]. The strain produced by over- or underwinding can be accommodated by the formation of local single-stranded regions which tends to increase with increasing temperature. "Breathing" of the DNA (see third paragraph in Section 2.1) plays an important part here too. A sequence with $> 90\,\%$ A – T can exist permanently unpaired in a superhelical molecule. This is important for many reactions of DNA. Supercoiling influences transcription (see Section) and vice versa. Positive supercoils are formed in front of the transcription apparatus and negative supercoils behind it; these supercoils are controlled by enzymes [47].

Bending. The base sequence of DNA is of tremendous importance for its structure [44]. In a right-handed helix, the twist angle between two bases changes depending on the sequence. This may result in the bending of a linear double helix. Bending can also be caused by proteins. Bending is of significance for the packaging of DNA and for many of its biological reactions [48, 49].

Intrinsically bent DNA is formed when special base sequences or structural motives are repeated in phase with the DNA helical repeat; homopolymeric A tracts being the best example [50, 51]. Protein-induced DNA bending plays an important role in recombination, initiation of transcription, and replication [52, 53]. Bends are also important structural features; indeed, regulatory protein binding sites can be replaced by an intrinsic bend [49, 54].

Special Structural Elements. Short sequences are frequently repeated in regulatory regions. *Repeats* can be recognized by DNA-binding proteins. Owing to DNA breathing, double-stranded regions (hairpins or stem-loops and cruciform structures) can be formed at repeats within a single strand (Fig. 4). This rarely happens in double-stranded DNA because stem loops are energetically less favorable than linear double strands. However, it is encountered frequently in single-stranded DNA and RNA. Supercoiling can promote the formation of cruciform structures, whereas transcription inhibits it [55]. Hairpins can play a part in replication, transcription, and RNA processing [56].

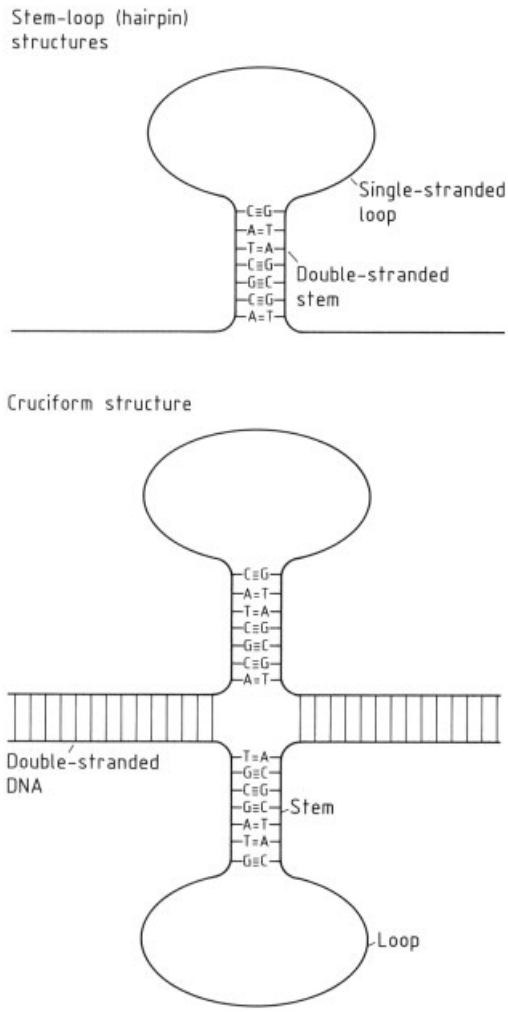

Figure 4. Special structural elements in nucleic acids

Homopyrimidine – homopurine runs are frequently found in regulatory regions of eucaryotic genes and are especially sensitive to nucleases [57]. There is a high tendency to form right-handed structures other than B-DNA in such regions [58].

Although the chains of double-stranded DNA are normally antiparallel, *parallel double-stranded oligomers* have also been found in vitro [59]. They form a right-handed helix and are even recognized by several enzymes. They are less temperature-stable than the corresponding oligomers.

At homopurine – homopyrimidine sections of the DNA, a homopyrimidine oligonucleotide can attach itself parallel to the homopurine strand in the major groove and form a *triple helix* [58]. Structures of this type can be used for specific strand cleavage with the help of coupled ellipticine derivatives or metal chelates [60, 61].

Centromeres are important compact DNA structures of the eucaryotic chromosome that are rich in adenine and thymine. Their exact structure is not known but they are important for the attachment of the spindle fibers during mitosis.

The ends (*telomeres*) of linear chromosomes (as in eucaryotic DNA) pose a special problem. DNA polymerases synthesize DNA from a DNA template and always require an RNA primer to start replication. Cleavage of this primer then results in a small 5′-gap which cannot be closed by the polymerase. Under normal replication conditions, the ends should therefore beocme shorter with every cycle of DNA replication (see Section 4.1). Special enzymes (telomerases) are responsible for adding telomere repeats to the chromosome ends to maintain constant length (Fig. 5): repeats can fold back and provide a 3′-OH group which serves as a primer for copying the last segment of a linear DNA molecule. Disturbances in telomeres can lead to aging phenomena [62, 63] and a role in carcinogenesis is also being discussed. Broken ends of chromosomes that are no longer protected by telomeres are very susceptible to fusion with other DNA ends and to degradation by nucleases [64]. The antiparallel structure and function of telomeres are highly conserved in all eucaryotes and are species specific. They consist of simple, tandemly repeated sequences with clusters of G residues [65, 66]. The G-rich strand is aligned in the 5′ → 3′ direction towards the end of the chromosome and has a single-stranded 3′-end containing 12 – 16 nucleotides. Telomeres can associate to form stable, parallel, four-stranded structures (G 4-DNA) [67].

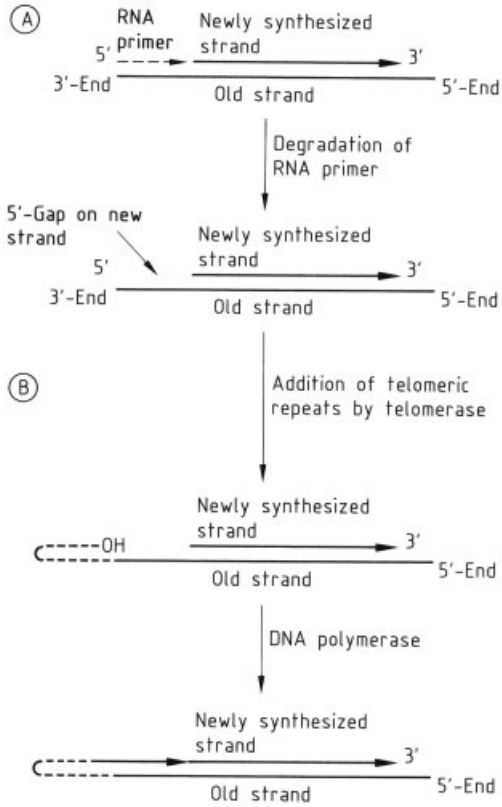

Figure 5. The importance of telomeres
Without telomere addition (A) newly synthesized DNA strands become shorter; with telomere addition by telomerase (B) constant length can be maintained.

2.2. Structure of RNA

RNA is an unbranched single-stranded polymer with many intramolecular double-stranded sections that may account for 50 – 67 % of the molecule. As in DNA, the backbone of RNA consists of 3′,5′-phosphodiester bonds (Fig. 1); however the sugar is ribose (and not deoxyribose) and uracil replaces thymine. Double-stranded RNA cannot form a B-helix because of steric hindrance caused by the 2′-OH groups of ribose; helices of the A type are, however, possible.

The functional groups of the nucleotides in the major groove of the A type of double he-

lix found in RNA are not easily accessible to proteins [68]. Protein binding to RNA probably occurs via interaction with single-stranded regions.

Four functional RNA families exist: messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and small nuclear RNA (only in eucaryotes). The structure of tRNAs has been studied most extensively; about half of the ca. 75 – 90 nucleotides within the tRNA molecule are paired, resulting in a secondary structure with a stem and three loops similar to that of a cloverleaf [69].

RNA has many different biological functions and exhibits a spectrum of flexible structures that more closely resemble those of proteins rather than those of the chemically related DNA [70]. RNAs have secondary structures–double-stranded sections, hairpins, internal loops, and bulged bases. With unpaired nucleotides pronounced tertiary structures are formed in addition to the secondary structure. Examples of tertiary structure motives are pseudo knots [71], produced by folding back in a hairpin and formation of a second stem – loop structure [72].

Formation of DNA – RNA hybrids is of importance in the replication and transcription of DNA and in the reverse transcription of viral RNA. Such hybrids can form secondary structures but they are considerably more polymorphous than DNA alone [73].

3. Properties

3.1. Physical and Chemical Properties

The size of naturally occurring DNA varies from a few thousand to 10^9 base pairs. The length of such molecules (micro- to centimeter range) can easily be measured under the electron microscope.

DNA absorbs UV light at 260 – 280 nm due to its bases. Aqueous DNA solutions are very viscous; viscosity depends on DNA length, DNA concentration, and temperature. Heating to a critical temperature is accompanied by a decrease in viscosity because the hydrogen bonds responsible for base pairing are disrupted and the helix structure collapses. This process is called *thermal denaturation or melting of DNA*. The temperature at which one-half of the base pairs is disrupted is denoted the *melting temperature*. It depends on the base composition (G – C pairs are more stable than A – T pairs). Double-stranded DNA ranging in size from 100 to > 100 000 base pairs melts at ca. 90 °C. In shorter double strands a gradual decrease in the melting temperature is observed. The melting temperature increases with increasing salt concentrations because the solubility of the bases decreases and hydrophobic interactions are increased. Chemicals that compete with hydrogen bond formation, such as urea or formamide, lower the melting temperature of DNA. Methanol has a similar effect; it increases the solubility of the bases and increases the interaction with water. The "melting" of double-stranded DNA is also facilitated by solvents such as ethylene glycol, dimethylformamide, dimethyl sulfoxide; low ionic strength; or extreme pH values. DNA can be denatured at an alkaline pH because the keto – enol equilibria of the bases are shifted preventing these groups from participating in hydrogen bonding.

Since the stacked bases in the double-stranded helix are not as easily excited by UV light as in single strands, absorption at 260 nm is lower for double-stranded DNA than for single strands. Increase in UV absorption can thus be used to measure DNA denaturation. At 260 nm solutions containing 50 µg/mL of double-stranded DNA, 50 µg/mL of single-stranded DNA, and 50 µg/mL of free bases have absorptions of ca. 1.00, 1.37, and 1.60, respectively.

Denaturation can also occur in the presence of proteins that destabilize the helix (melting proteins). Such proteins are required to unwind the helix during replication and to facilitate interaction between single strands during genetic recombination.

The reassociation (renaturation) of thermally denatured DNA is a spontaneous process but only occurs if the solution is cooled slowly below the melting temperature. Renaturation can take several hours, depending on the size of the molecule, because it initially relies on random base pairing (hybridization); it is, however, a cooperative process. Rapid cooling of denatured DNA at salt concentrations > 50 mmol/L produces a very compact molecule in which about two-thirds of the bases are hydrogen bonded or stacked. At salt concentrations below

10 mmol/L the DNA remains denatured even after cooling.

The length of RNA varies greatly: tRNA has a length of 75 – 90 nucleotides and mRNA can be up to several thousand nucleotides long. Denaturation effects are rarely observed because RNA has few truely double-stranded regions; it is most likely to be observed in tRNA.

Because they are extremely long, DNA molecules are extremely sensitive to mechanical influences (shearing forces, e.g., vigorous stirring) and easily break into small fragments (ca. 1000 base pairs). Ultrasonic treatment of DNA in solution produces fragments of ca. 100 – 500 base pairs owing to disruption of hydrogen bonds and single-strand and double-strand breaks in the sugar – phosphate backbone [74]. Nucleic acids are sparingly soluble in water (depending on the molecular mass). They are negatively charged and acidic at physiological pH and form water-soluble alkali and ammonium salts that can be precipitated with ethanol.

RNA and DNA are insoluble in cold acid. DNA is more sensitive to acid hydrolysis than RNA. At pH < 1, however, both DNA and RNA break down into the free bases, phosphoric acid, and (deoxy)ribose. Acid hydrolysis can be used to determine the base composition of nucleic acids (e.g., total hydrolysis can be achieved by heating DNA in 90 % formic acid at 180 °C for 30 min). The β-glycosidic linkage between the N-9 of purines and the C-1 of deoxyribose is selectively cleaved at ca. pH 4, resulting in apurinic sites. Anhydrous hydrazine cleaves the pyrimidine residues.

DNA is stable at pH 13, only 0.2 of 10^6 phosphodiester bonds are broken per minute at 37 °C. In contrast, RNA is rapidly hydrolyzed at alkaline pH.

DNA can be both specifically and nonspecifically cleaved by a variety of enzymes [deoxyribonucleases (DNases)]. RNA is cleaved by ribonucleases (RNases). Some of these cleavage reactions are exploited for sequencing RNA [75, 76]. See also → Enzymes, Chap. 6.4.→ Enzymes, Chap. 6.5.

3.2. Interaction with Proteins

In bacteria, DNA occurs as a complex with RNA and proteins that is bound to but not surrounded by a membrane. The DNA often has a closed circular form and is organized in a series of superhelical loops.

The DNA of higher cells is enclosed within the nuclear membrane as morphologically distinct units of varying size (chromosomes); it is associated with basic proteins called histones. The number and size of the chromosomes are species specific (karyotype). Two full turns of the DNA double helix (146 base pairs) are wound around a histone octamer (diameter ca. 8.6 nm) to form a nucleosome. The width of the grooves varies due to the periodic arrangement of A – T trinucleotides on the inside and G – C trinucleotides on the outside of the nucleosome at intervals of about ten base pairs [53]. Nucleosomes can become condensed into fibers of 10 or 30 nm (super superhelices, solenoids).

Eucaryotic cellular organelles (e.g., mitochondria, chloroplasts) possess closed circular DNA that is not associated with histones.

4. Biosynthesis and Biological Function

4.1. DNA Replication

The genetic information of all cellular organisms is stored in double-stranded DNA (viruses may, however, also have single-stranded DNA or RNA, as well as double-stranded RNA). It is extremely important that the transfer of biological information in DNA (i.e., its base sequence) occurs with a very high degree of accuracy. Because of perfected proofreading and repair mechanisms (see Section 4.5) DNA replication has an error level of $10^{-8} - 10^{-11}$ [77], i.e., for every $10^8 - 10^{11}$ bases in newly synthesized DNA only one is incorrectly incorporated. The replication of DNA is carried out by DNA polymerases which require a single strand of DNA as a template and a short double-stranded piece of nucleic acid (formed with the help of a primer) for initiation (Fig. 6; see also → Enzymes, Chap. 6.2.).

The DNA is synthesized from deoxyribonucleotide triphosphates which are polymerized on the single-stranded DNA template with the release of pyrophosphate; the cleavage of pyrophosphate by a pyrophosphatase provides the

energy required for DNA biosynthesis. The addition of new nucleotides always takes place at the 3′-OH group of the sugar, therefore all biologically synthesized nucleic acids grow in the $5' \rightarrow 3'$ direction. Divalent cations such as Mg^{2+} or Mn^{2+} are important cofactors. Some polymerases contain $3' \rightarrow 5'$ exonuclease activity and can remove bases that have been incorrectly incorporated. The primer in vivo is usually RNA, which is synthesized by a special RNA polymerase (primase) and removed later by an exonuclease (e.g., DNA polymerase with $5' \rightarrow 3'$ exonuclease activity). The various activities are combined in a multienzyme complex.

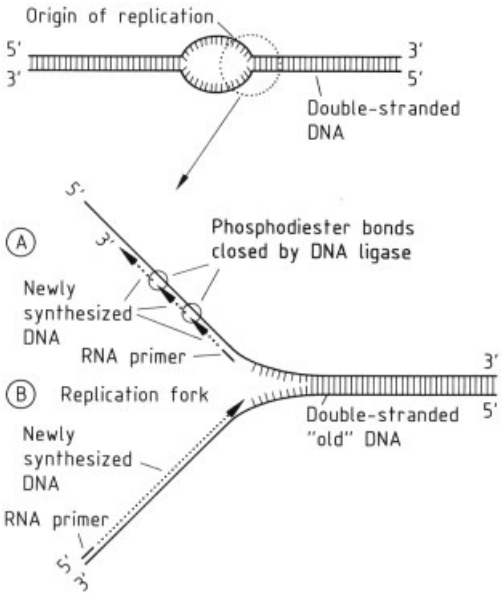

Figure 6. Simplified scheme of DNA replication
A) Lagging strand synthesis: after primer addition by primase a stretch of DNA is synthesized by DNA polymerase, the primer is removed and after synthesis of the adjacent stretch of DNA the phosphodiester bond between the two stretches is closed by a DNA ligase.
B) Leading strand synthesis: after primer addition by primase the new strand is synthesized by DNA polymerase. Arrows denote direction of synthesis.

The replication of DNA is a semiconservative process, i.e., one strand of each of the two new daughter molecules of DNA is an old strand and the other a newly synthesized one [78]. Replication starts with the creation of replication forks at the "origin of replication", and proceeds in opposite directions along the DNA (see Fig. 6).

The DNA of procaryotes is replicated from a single replication origin. The main replication enzyme in *Escherichia coli*, DNA polymerase III, consists of at least ten different subunits. Replication is probably a membrane-bound process [79].

Because DNA strands are antiparallel, one strand (lagging strand) must be synthesized "backwards" to keep the replication complex together. The direction of synthesis of this strand is contrary to the direction of movement of the replication fork and occurs discontinuously. Segments (Okazaki fragments) are formed which are then joined by a DNA ligase. The strand, that is synthesized "forward" is called the leading strand. Replication in eucaryotes starts at several points along the DNA and the replication complexes are fixed to the nuclear matrix [80]. Viroids are the smallest self-replicating structures known. These single-stranded, circular, protein-free RNAs are a few hundred nucleotides long and occur as pathogens in higher plants [81].

4.2. Gene Expression

The base sequence of DNA constitutes the genetic information in the form of discrete units (genes). A sequence of three successive bases (codon) acts as a code for a certain amino acid (e.g., GCT, GCC, GCA, or GCG code for alanine) or as a stop signal (TAG, TGA, or TAA) in protein synthesis. Of the 64 ($=4^3$) possible triplets, 61 code for amino acids and 3 specify stop signals. Since this genetic code is almost universal, eucaryotic proteins can be synthesized in procaryotes from eucaryotic DNA (→ Genetic Engineering). Some organisms and organelles differ in a few codons from the universal code [82]. The genetic information contained in DNA is normally used as a template for the synthesis of mRNA which itself then acts as a template for protein synthesis.

The regulation of gene expression in living organisms is based primarily on the recognition of nucleic acid sequences by proteins. However, nucleic acids can also be involved. "Antisense RNA" for instance, can hybridize with mRNA, leading to a block of translation and to the degradation of the mRNA by RNase. The termination of transcription can occur in a similar way. This principle may possibly be exploited for therapeutic applications [83].

4.2.1. Transcription

The process by which genetic information is transferred from DNA to mRNA is called transcription. This process is catalyzed by DNA-dependent RNA polymerases and involves the synthesis of mRNA from ribonucleotide triphosphates in the presence of Mg^{2+} or Mn^{2+} and a single- or double-stranded DNA template. The direction of synthesis is $5' \rightarrow 3'$ and no primer is required. As in DNA synthesis, energy is obtained from the cleavage of pyrophosphate. The accuracy of transcription is lower than that of DNA replication (ca. 10^{-4}) because there are no repair processes [77]. Most RNA polymerases are complex enzymes consisting of several subunits. Bacteria generally have one, but eucaryotes have three (see also → Enzymes, Chap. 6.3.).

Due to the specific base pairing in the double strand, the base sequence information of DNA determines the base sequence of the mRNA synthesized by RNA polymerases. Messenger RNA acts as an intermediate for conveying the information required for protein synthesis from the DNA to the protein-synthesizing structures, i.e., the ribosomes (see below). For interaction with the ribosome, mRNA possesses a ribosomal binding site shortly before the translation initiation sequence. This site is complementary to rRNA sequences [84]. The average length of *Escherichia coli* mRNA is about 1200 base pairs, eucaryotic mRNA can be much longer. Some types of mRNA are very rapidly degraded and exist in relatively low concentrations. Other types are more stable and accumulate in considerable amounts in the cell. The ends of mRNA are important for its stability.

Modifications in the form of loop structures [85] in procaryotes, and poly(A) at the $3'$-terminus and a cap (formed by addition of 7-methyl-GTP followed by methylation) at the $5'$-terminus in eucaryotes protect the ends from nonspecific degradation by nucleases (Fig. 7). Capping and the addition of poly(A) usually occur before splicing. Random degradation of mRNA is also prevented by cellular ribonuclease inhibitors. The mRNA transcript is finally exported to the cytoplasm where it acts as a template for protein synthesis.

Transcription is increased or decreased by sequence elements on the DNA that act as binding sites for protein factors [86–88]. Transcription also involves initiation and termination factors, some of which are ribonucleoproteins. Structural features also affect transcription, e.g., bent DNA can activate the process [52].

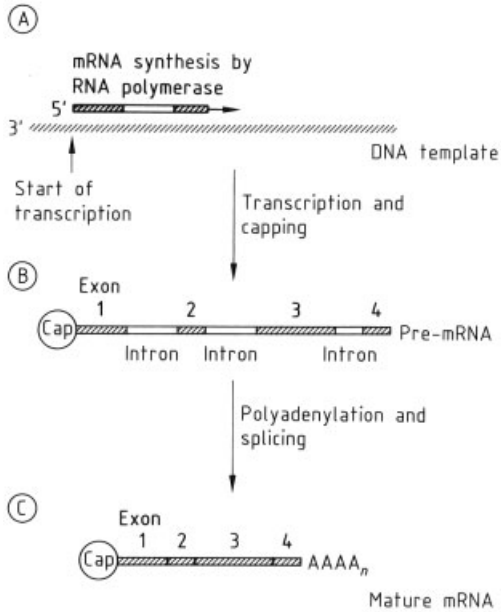

Figure 7. Simplified scheme of transcription in eucaryotes A DNA template is transcribed into pre-mRNA by RNA polymerase (A) and a cap is added to the $5'$-end of the mRNA (B). A stretch of A-residues is then added to the $3'$-end. During splicing intervening sequences (introns) are removed and the sequences carrying the information for protein synthesis (exons) are joined to form mature mRNA (C).

RNA polymerase binds to a promoter region in the DNA template and then transcribes the DNA into mRNA. In most eucaryotic genes protein-coding sequences (exons) are interrupted by intervening noncoding sequences (introns). Many exons code for protein domains; exon rearrangements in DNA can facilitate the exchange of domains (exon shuffling). Lower eucaryotes have a higher content of intron-free genes than higher eucaryotes. Introns are rarely found in bacterial genes [89]. Introns are removed from eucaryotic precursor mRNA and the exons are then joined to give mature mRNA in a process known as splicing (see Fig. 7). The length of introns varies from 30 to 100 000 base pairs, the average exon length is 50 – 300 base pairs. Splicing demands a high degree of accuracy and is carried out by a multicomponent

complex known as a spliceosome that consists of protein and small RNAs (snRNAs = small nuclear RNAs). The main components of the spliceosome are ribonucleoproteins (snRNPs = small nuclear ribonucleoproteins) [90, 91]. Introns usually start with GU (5′-splice site) and end with a pyrimidine-rich section followed by AG (3′-splice site) [92]. The half-life of introns in mRNA varies from a few seconds to 10–20 min. Splicing always takes place in the nucleus. Unspliced RNA remains in the nucleus and is degraded.

The splicing mechanism is highly conserved: a mammalian cell extract can process yeast RNA. The number of possible transcription products can be increased by alternative splicing (i.e., not all available exons of a gene are used to create one mature mRNA).

In protozoa RNA is also "edited", the primary transcript is converted into its functional form not only by post-transcriptional removal of certain bases, but also by insertion of others [93, 94].

In some RNA viruses (retroviruses), replication of the genetic material proceeds via a DNA intermediate in a process known as reverse transcription. The information in RNA is transcribed back to DNA with suitable enzymes (reverse transcriptases). The DNA is then used as a template for RNA synthesis (see also → Enzymes, Chap. 6.2.4.). Reverse transcription can also be used in vitro to produce intron-free complementary DNA (cDNA) from eucaryotic mature mRNA. Telomerases are special reverse transcriptases with an internal RNA template [95, 96].

4.2.2. Translation

Protein synthesis takes place on the ribosomes, which are complexes of rRNAs and more than 50 different proteins. Protein factors are involved in all phases of translation. Various types of RNA are involved in protein synthesis: mRNA, the carrier of genetic information from the DNA; rRNA, component of ribosomes and directly involved in almost all stages of protein synthesis [97]; and tRNA, which makes activated amino acids available for protein synthesis. Translation is an irreversible process with error rates of 10^{-3} – 10^{-4} [77].

The initiation of protein synthesis is a very intricate multistep process [98, 99]. A translation initiation complex is formed from ribosomes, mRNA, a special tRNA, initiation factors, and GTP. The mRNA start codon for protein biosynthesis is usually AUG but GUG, UUG, or AUU may also be used [89]. After initiation of translation, the protein is synthesized step by step (elongation) from amino acids that are delivered and activated by tRNA. Each amino acid has a specific tRNA.

The tRNAs act as adapters between the codons of mRNA and the corresponding amino acids. Apart from the "usual" bases, tRNA contains several rare bases (e.g., 5-hydroxymethylcytosine). Transfer RNAs have an amino acid attachment site and a template recognition site (anticodon) for the mRNA. Amino acids are attached by an ester bond either to the 3′-or 2′-OH group of the terminal adenosine of tRNAs. Loading the tRNAs with the correct amino acids is crucial for exact translation of genetic information. This step is catalyzed by specific enzymes (aminoacyl-tRNA synthetases). Each of these enzymes must be able to recognize a specific amino acid and its associated tRNA.

The amino acid sequence of the protein product is specified by the mRNA codons and thus ultimately by the base sequence of DNA. A stop codon (TAA, TAG, TGA) on the mRNA leads to termination of protein synthesis and the release of the protein. Differences between translation complexes in procaryotes and eucaryotes permit the selective interruption of procaryotic protein synthesis (e.g., by some antibiotics).

4.3. Modification and Degradation

DNA is exposed to many damaging influences and can be modified and cross-linked by radiation and chemicals. Spontaneous hydrolysis of DNA bases may also occur.

Mutations change the normal base sequence of DNA. Erroneous bases can appear as a result of replication errors or chemical modification (→ Mutagenic Agents). Intercalating agents such as benzopyrenes, aflatoxins, or ethidium bromide can cause deletions. Insertion of individual bases is also possible. Insertion of larger DNA segments can occur through mobile DNA elements (transposable elements, jumping

genes). The insertion site is usually arbitrary. This process (transposition) can lead to gene activation or inactivation.

Methylation of DNA is used by cells in different ways [100, 101], e.g., to distinguish between newly synthesized strands (unmethylated) and old strands (methylated). Methylation occurs at defined sites: in bacteria at the N-6 position of adenosine and the C-5 position of cytosine, and in eucaryotes, physiologically only at cytosine. The methylation pattern also differentiates between endogenous and foreign DNA and is recognized by restriction – modification systems (in procaryotes): endogenous DNA is methylated at specific sites and foreign DNA is digested. Methylation can be used to switch genes on or off in eucaryotes. C-Methylation influences the equilibrium between B- and Z-DNA [102] and can lead to altered DNA – protein interactions by the modulation of DNA topology. Methylation patterns are stable and hereditary. Methylation can have a mutagenic effect: the deamination of 5-methylcytidine to thymine occurs spontaneously.

Degradation of nucleic acids can be effected in vivo and in vitro by a series of enzymes. Some DNases are specific for single strands and others for double strands, some cleave bonds located within a double strand (endonucleases) and others require a free DNA end (exonucleases). Exonucleases act either in the $3' \rightarrow 5'$ or the $5' \rightarrow 3'$ direction; exonuclease activity is often associated with DNA polymerases for proofreading in DNA replication. Restriction endonucleases are a special class of nucleases. RNases can also be divided into exo- and endonucleases. RNase II is a bacterial exonuclease; RNase H is an endonuclease that digests RNA in DNA – RNA hybrids. See also, → Enzymes, Chap. 6.

4.4. Recombination

In recombination two DNA molecules interact and exchange segments to produce two new molecules that contain genetic information from both parental DNAs. DNA recombination is by no means limited to genetic engineering, but occurs widely in nature. This process is catalyzed by enzymes and usually requires the presence of homologous regions (homologous recombination). In rarer cases, nonhomologous, illegitimate recombination can take place. Transposition (see Section 4.3) is also a recombination phenomenon [103].

4.5. DNA Repair

Although changes in base sequences create the genetic variety required for evolution, these changes are only rarely beneficial for individual organisms. Despite many damaging influences (see Section 4.3), nucleic acids are replicated with amazing accuracy. This exact conservation of genetic material is one of the fundamental requirements for genetic success. Damaged DNA that is not repaired can lead to mutations and cell death, and give rise to tumors. Living organisms have therefore developed a variety of enzymatic mechanisms for eliminating DNA damage. Most damage is recognized by repair systems in the cell and repaired either by directly reversing the damage or by using the undamaged second strand as a template. *Escherichia coli* has been used as a model for studying repair mechanisms [104]. The best studied mechanism is nucleotide excision repair. This multistep process involves recognition of the error, nicking of the damaged strand, removal of nucleotides at the damaged site, repair DNA synthesis, and ligation. It requires the participation of multienzyme complexes [105, 106].

4.6. Nucleic Acids as Enzymes

Ribozymes are RNA molecules that act as catalysts: sequence-specific RNA endonucleases, nucleotidyl transferases, phosphatases, kinases, and glucanotransferases [37, 107]. The sequence-specific cleavage of oligodeoxyribonucleotides and single-stranded DNA can also be catalyzed by ribozymes [108]. Previously, only proteins (enzymes) were thought to have activities of this kind. Ribozymal properties were discovered in the early 1980s in the self-splicing introns of *Tetrahymena* [109] and ribonuclease P [110], which cleaves tRNA precursors and contains RNA as the catalytic subunit. In eucaryotes, small nuclear RNAs (snRNAs) are required for the splicing of pre-mRNA; they are part of the spliceosome, an RNA – protein complex [90].

The smallest RNA structure that exhibits catalytic (cleavage) activity is 19 nucleotides long and forms a hammer head structure with its target [111]. No pure DNA catalysts have been discovered yet [112]. However, since a few of the ribonucleotides in ribozymes can be replaced by deoxyribonucleotides, the three-dimensional structure that permits nucleic acid catalysis may not necessarily be limited to RNA [113, 114].

In all proven examples of RNA catalysis, the substrate attacked by the ribozyme is itself a nucleic acid. However, there are indications that rRNA also attacks other molecules (i.e., is an RNA enzyme). RNase P is the only known naturally occurring ribozyme that does not cleave itself, but attacks other molecules [72, 115].

5. Isolation, Purification, and Transfer

Chromosomal DNA from bacteria or eucaryotes has a very high molecular mass and is extremely fragile. Thus to isolate DNA, cells must be gently lysed with enzymes (pronase, protease K, lysozyme) and mild detergents. Accompanying RNA can be removed by treatment with enzymes which digest RNA or by gel filtration [116]. Proteins associated with DNA are denatured with phenol and extracted [117].

Plasmid DNA is relatively small and has a closed circular and supercoiled form. It is more stable than chromosomal DNA when subjected to shearing forces and can be renatured much more easily after denaturation; isolation and purification of plasmid DNA is achieved relatively easily (e.g., by equilibrium centrifugation in cesium chloride – ethidium bromide gradients at 150 000 – 400 000 g for 5 – 48 h). Kits are also available which permit the isolation of pure plasmid DNA in a short time without ultracentrifugation using binding to glass powder (glass milk) and/or gel filtration. Similar methods are used for the *isolation of RNA*. However, degradation by ribonucleases must be prevented (RNA is much more sensitive to nucleases than DNA). Either proteases or chloroform are employed to remove protein; with chloroform protein precipitates at the phase boundary. Phenol is not normally used because a small part of the RNA also goes into the phenol phase. Oligo(dT) cellulose (cellulose-bound oligonucleotides containing only thymine) can be used for the purification of eucaryotic mRNA [118] taking advantage of its poly(A) tails.

Nucleic acids can also be purified by ion-exchange chromatography (e.g., by HPLC) [119]. Suitable separating media are hydroxyapatite and anion exchangers. Reversed phase methods and gel filtration are also used.

Microgram amounts of DNA fragments can easily be isolated and purified by techniques such as electroelution from agarose gels, binding to anion-exchange paper, dialysis, or binding to glass milk (powdered glass suspension). After purification, DNA can easily be concentrated by precipitation with ethanol or 2-propanol.

DNA can be transferred relatively easily in vivo between related organisms (especially in bacteria) via endogenous mechanisms (e.g., conjugation, some viruses). Several other methods are available for the transfer of DNA in the laboratory: application as a $CaCl_2$ precipitate, exposure to an electric field (electroporation), and the use of viral vectors (see also → Genetic Engineering).

6. Analysis of Nucleic Acids

Sequencing. Computer data banks such as EMBL [120] and GenBank [121] contain sequence information in the order of millions of nucleotides.

Single or double-stranded DNA can be sequenced [122]. The most widely used methods are the Sanger method [123] and the Maxam and Gilbert method [124] (→ Genetic Engineering, Chap. 3.3.). In the Maxam and Gilbert method the DNA is cleaved with chemicals statistically and specifically at one of the four bases. The Sanger method employs enzymatic synthesis of a complementary copy of a single-stranded DNA template and random chain termination with dideoxynucleotides. Both methods produce DNA fragments that are separated by gel electrophoresis.

Both methods initially made use of radioactively labeled nucleotides and subsequent exposure on X-ray films for identification of the fragments. More recent variations use fluorophores [fluorescein, tetramethylrhodamine, Texas red, 4-chloro-7-nitrobenzo-2-oxa-1-diazole (NBD)]

which are more stable, easier to use, less dangerous, and permit the automatic detection of all four nucleotides on a single track. Sequences of > 450 bases can be determined per gel run and track [125–128]. An automatic sequencing machine allows determination of sequences of up to 10 000 bases per day. Further acceleration by combination with automatic DNA preparation in robotic workstations may be possible [129].

More recent techniques aim at faster sequence determinations (several hundred bases per second [130]) with smaller amounts of DNA. Variations of the Sanger method permit determination of sequences up to 10 000 bases without subcloning [131].

The Sanger method can also be used for sequencing RNA [132] and sequences up to 150 bases can be determined in a single run. Reverse transcriptase is used to synthesize DNA with RNA as a template. The RNA sequence is then deduced from the DNA sequence. Another method for RNA sequencing involves the use of specific RNases (→ Enzymes, Chap. 6.5.2.6.) [75, 76].

X-Ray Analysis. X-ray analysis is of considerable importance for elucidating the structure of nucleic acids and their building blocks [44]. Useable crystals can be obtained from aqueous solutions for molecules up to the size of tRNA (75 – 90 base pairs). Nucleic acids of higher molecular mass are too heterogeneous to produce crystals, quasi-crystalline fibers drawn from concentrated aqueous solutions are used instead.

Nuclear Magnetic Resonance (NMR). The best method for determining the structure of relatively short (10 – 20 bases or base pairs) DNA or RNA fragments in solution is ^{1}H-, ^{13}C-, ^{15}N-, or ^{31}P-NMR (one- or two-dimensional) [133, 134]. NMR is also used to study short triple helices [135], DNA – RNA hybrids in solution [136], DNA – drug complexes, mismatches, and intercalating agents. Torsion angles and the mobility of base pairs and the sugar phosphate backbone can be determined.

Electron Microscopy. Electron microscopy permits the determination of the length of DNA. DNA heteroduplexes, binding of large protein factors [137], and supercoiled DNA [138] can be visualized. A special application is DNA-hybridization electron microscopy [139, 140]: biotinylated DNA probes are hybridized to selected regions of DNA or RNA and then reacted with avidin and visualized by electron microscopy.

Scanning tunneling microscopy (STM) allows the observation of single-stranded DNA adsorbed on graphite with atomic resolution. It opens up the possibility of structural studies on DNA and of direct sequencing [141–143]. The profiles show excellent correlation with the results of X-ray analysis.

Centrifugation. Ultracentrifugation in cesium chloride gradients can be used to measure the molecular mass of DNA and also for the experimental identification of covalently closed circles. Sedimentation at alkaline pH (> 11.3) causes disruption of the hydrogen bonds and unwinding of the DNA molecules. In the case of linear DNA, two single strands are obtained which have a higher sedimentation coefficient than native DNA at salt concentrations > 0.3 mol/L. Covalently closed circles cannot be denatured into single strands, they collapse and have a sedimentation coefficient that is three times higher than that of native DNA.

In equilibrium centrifugation in cesium chloride, ethidium bromide is firmly bound to DNA in salt solutions by intercalating between base pairs. This lowers the density of the DNA by about 0.15 g/cm^3 and the DNA unwinds with increased binding of ethidium bromide. Unlike linear DNA, covalently closed molecules have a limited absorption capacity for ethidium bromide. Linear DNA is consequently lighter than covalently closed circular DNA. Use of a cesium chloride density gradient with ethidium bromide therefore permits the separation of supercoiled and linear (relaxed) DNA of the same size.

Staining and Quantification. Nucleic acids can be stained with ethidium bromide [144], Stainsall, 1-ethyl-2-[3-(1-ethylnaphtho[1,2 *d*] thiazoline-2-ylidene)-2-methyl-1-propenyl]-naphtho[1,2 *d*] thiazolium bromide [145], or silver [146–148]. DNA stained with ethidium bromide can be quantified in electrophoresis gels after being photographed [144, 149]. Silver staining followed by a densitometer scan is

equally sensitive (1 – 2 ng of DNA can be detected) [146]. A quick and sensitive estimation of the amount of nucleic acid can be obtained by measuring the UV absorption at 260 nm: an absorption of 1.00 corresponds to a concentration of ca. 50 µg/mL of double-stranded DNA.

DNA quantification in the picogram range has been reported with a silicon sensor, that responds to the interaction between DNA-binding proteins and DNA [150]. Quantification in the picogram range is also possible after specific hybridization (see Chap. 8).

Fluorescence imaging of DNA in picogram to nanogram amounts is made possible by laser excitation of a DNA – ethidium homodimer [5,5′-diazadecamethylene) bis(3,8-diamino-6-phenylphenanthridinium) dichloride dihydrochloride] complex at 488 nm [151].

Electrophoresis. Since DNA is negatively charged because of its phosphate groups, it migrates in an electrical field (usually at a pH of ca. 7). The electrophoretic mobility of DNA fragments in solution is independent of the molecular mass (constant linear charge density). However, electrophoresis of DNA in gel matrices (agarose or polyacrylamide) at a constant field strength is an effective method of separating DNA molecules according to molecular mass. The mechanism of separation is not well understood [152].

Important experimental parameters are gel concentration, field strength, temperature, and running time. Gels of 0.5 – 2 % are usually employed for the separation of 50 – 20 000 base pairs at field strengths of 2 – 3 V/cm [153]. The mobilities of linear and supercoiled DNA of equivalent molecular mass are different. Estimation of the molecular mass of supercoiled DNA molecules based on linear standards is therefore not possible.

Polyacrylamide gels (4 – 12 %) give a sharp separation of polynucleotides up to a length of about 600 base pairs. Special gels can separate polynucleotides with up to 2500 base pairs [148]. Mobility varies logarithmically with the size of the DNA at field strengths below 1 V/cm [154]. Separation efficiency is limited in the range of 10 000 – 50 000 base pairs, there is no resolution above 50 000 base pairs.

Denaturing gels (e.g., with urea) are used in the sequencing and analysis of single-stranded nucleic acids. Under special conditions, the behavior of individual DNA molecules in an electric field can be observed with a fluorescence microscope [155].

Pulse-field electrophoresis is used to separate larger linear DNA molecules. Here the direction of the electric field is changed at intervals. The DNA molecules therefore have to change their orientation, the time required for this change depends on the size of the DNA [44, 156]. Unlike conventional electrophoresis (duration 30 – 60 min), a pulse-field run takes 20 – 140 h. Usually, 1.5 % agarose gels are used and 10 – 20 µg of DNA can easily be separated at field strengths of 2.5 – 10 V/cm. Separation can be achieved up to ca. 12×10^6 base pairs [157]. Separation is affected by DNA topology and sequence, pulse time (seconds to minutes), field geometry, field strength, gel composition, sample concentration, temperature, and running time [154, 158].

7. Chemical Synthesis

The chemical synthesis of oligonucleotides began in 1955 when A. M. Michelson and A. R. Todd prepared the first dinucleoside monophosphate [159]. They used the so-called *triester method* to show that oligonucleotides of defined sequence can be prepared by purely chemical methods. In 1956 Khorana et al. successfully synthesized the same nucleotide by the *diester method* [160]. This process dominated oligonucleotide synthesis until the mid-1970s, and was also used by Khorana and coworkers for the first total synthesis of a gene (alanine-specific tRNA from yeast) [161]. The importance of synthetic oligonucleotides was generally underestimated because the synthesis was time-consuming and expensive. It was the progress in genetic engineering that resulted in a dramatic increase in the demand for synthetic oligonucleotides, which, in turn, led to the development of new synthesis concepts and to the automation of DNA and RNA synthesis.

7.1. Synthesis Strategy

The chemical synthesis of DNA consists of three steps:

1) Separate synthesis of the two complementary strands
2) Hybridization of the two strands (formation of hydrogen bonds between A and T, and between C and G)
3) Enzymatic joining of the DNA molecules to give larger DNA units

Three standard methods are used for the synthesis of the 3′,5′-internucleotide phosphodiester bond. They differ in the type of monomer building blocks used and are known as the phosphate triester, phosphoramidite (phosphite), and the *H*-phosphonate methods (Fig. 8, Section 7.4).

An oligomer can be synthesized either by stepwise addition of individual monomer building blocks or by joining oligomer blocks (e.g., dimers or trimers). The former concept is employed in solid-phase synthesis, which was developed almost simultaneously by R. B. Merrifield [162] for peptides and by R. L. Letsinger [163] for oligodeoxyribonucleotides. The first nucleoside (sugar and base) is covalently bound to an insoluble support at the 3′-OH group of the sugar. Stepwise synthesis of the chain then proceeds by condensation of the appropriate nucleoside monomers. Since the reaction product remains immobilized on the support, it can easily be freed of other reactants by washing; time-consuming chromatographic purification is not required. Synthesis is usually carried out on the 0.2 – 1 μmol scale and all

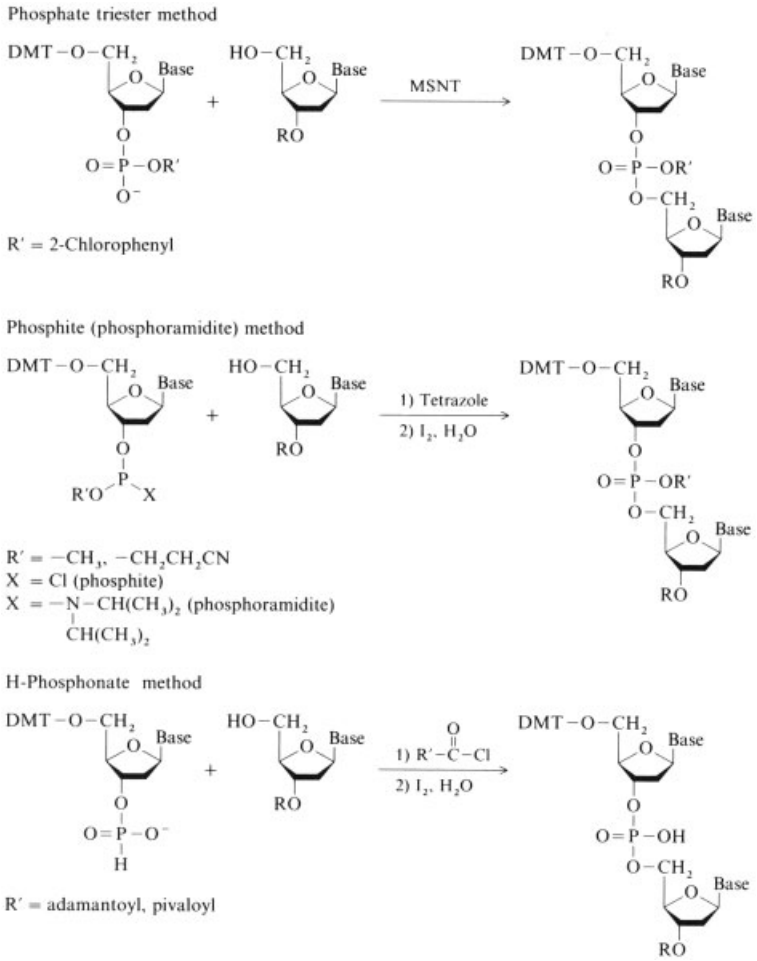

Figure 8. Chemical synthesis of oligodeoxynucleotides
DMT = 4,4′-dimethoxytrityl residue; MSNT = mesitylene sulfonylnitrotriazole; R = polymer support

reagents are applied in large excess. Solid-phase synthesis of oligonucleotides can be divided into the following steps:

1) Preparation of monomer units that are either protected or can be activated
2) Functionalization of the support
3) Stepwise synthesis of the desired nucleotide sequence by one of the three methods listed above
4) Cleavage of the protecting groups and the support
5) Isolation and purification of the desired sequence

7.2. Protecting Groups

In order to avoid unwanted side reactions during oligonucleotide synthesis, functional groups that do not participate in the linking reaction must be suitably protected. Only a few protective groups have gained acceptance and are based on the work of Khorana et al. [164].

Two types of groups can be distinguished: permanent and intermediary protecting groups.

Permanent protecting groups are maintained throughout the synthesis and protect the exocyclic amino functions of adenine, cytosine, and guanine; the OH group of phosphorus; and the 2′-OH group of ribose (in RNA synthesis). *Intermediary protecting groups* protect the 5′-OH group of the sugar and are cleaved after each condensation step.

The most important protecting groups are depicted in Figure 9. Adenine and cytosine are usually converted to the corresponding acid amides with excess benzoyl chloride, the amide of guanine is obtained with isobutyric anhydride. Base-protecting groups are introduced with a method developed by Ti et al. (transient protection) [165]. After termination of the synthesis, the protecting groups are removed with 32 % aqueous ammonia at 50 – 60 °C in 4 – 5 h. In long, guanine-rich sequences, the O-6 position of guanine can also be protected [166–169], usually with a 4-nitrophenylethyl residue [170]. This is especially important for attaining good yields in the triester process because otherwise sulfonation with the condensing agent easily occurs. More labile amidine or acyl base-protecting groups are of interest for the synthesis of phosphate-modified oligonucleotides because they are more sensitive to hydrolysis than their diester analogues and treatment with ammonia at 60 °C can cause considerable strand breakage [171–173]. Important examples of such groups are the 4-phenoxyacetyl residue

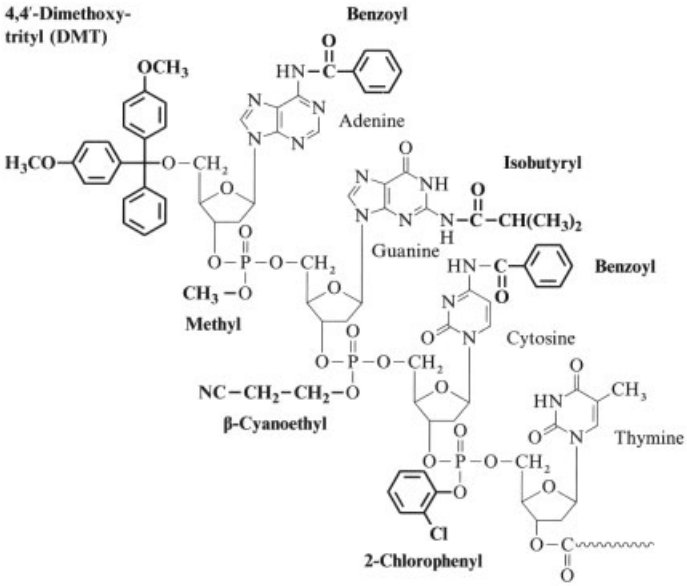

Figure 9. Protecting groups (bold type) used in the synthesis of oligonucleotides

(Pac) for adenine and guanine and the isobutyryl residue for cytosine.

The *tert*-butyldimethylsilyl residue is preferred as the protecting group for the sugar 2′-OH group in the solid-phase synthesis of oligoribonucleotides [174].

The 5′-OH group of deoxyribose or ribose is usually blocked with the 4,4′-dimethoxytrityl group [164], which can easily be removed after each coupling step by mild acid hydrolysis with 3 % di- or trichloroacetic acid in dichloromethane [175].

In the triester method the 2-chlorophenyl group is used to protect the phosphorus [176], while in the phosphoramidite method, the β-cyanoethyl group is preferred to the originally used methyl group [177, 178]. The β-cyanoethyl group can easily be removed at the end of the synthesis by β-elimination with aqueous ammonia [179]. No internucleotide phosphorus-protecting group is required in the *H*-phosphonate process.

7.3. Functionalization of the Support

Polystyrene [180], polyacrylamide [181], cellulose [182], and silica [183] have been used for solid-phase synthesis of oligodeoxyribonucleotides. The preferred support material is controlled pore glass (CPG) [184, 185] functionalized with a long-chain alkylamine spacer. The material is relatively rigid, does not swell, and is inert to all reagents used in the synthesis. In addition, it withstands the high mechanical stress encountered in the large number of reactions performed in automatic synthesizers. The narrow pore-size distribution gives more uniform diffusion parameters and, consequently, more reproducible flow rates and availability of functional support sites.

The support is functionalized with an alkylamine by silanization with 3-aminopropyltriethoxysilane [183, 186]. Heating and subsequent treatment with methanol result in cross-linkage of the silyl groups and hydrolysis of outer, less strongly fixed groups [187]. The nucleoside at the 3′-OH end of the DNA to be synthesized is linked to the support via a succinyl spacer [183, 188, 189]. Amino groups of the support that have not reacted are capped with acetic anhydride [188]. To determine the loading of the support with nucleoside material, an aliquot is treated with strong acid to remove the DMT residue. The resulting orange solution containing the dimethoxytrityl cation [190] can be determined by measuring the absorbance at 495 nm. The average loading is usually 15 – 25 µmol per gram of CPG.

7.4. Methods of Synthesis

The three main methods of oligonucleotide synthesis are the phosphate triester, the phosphoramidate (phosphite), and the *H*-phosphonate methods (see Fig. 8). The phosphoramidite method is the most widely used (Fig. 10, see next page). The reaction cycle involves

1) Cleavage of the 5′-OH DMT group
2) Condensation of the next monomer building block (chain elongation)
3) Masking of 5′-OH groups that have not reacted completely (capping)
4) In the phosphoramidite and *H*-phosphonate processes, an additional oxidation step is required after each condensation step or the end of the synthesis respectively to obtain the desired internucleotide phosphodiester.

Chain elongation in all processes occurs in the $3' \rightarrow 5'$ direction.

Phosphate Triester Method. This method was first used to synthesize oligonucleotides from suitable building blocks in solution [191–194]. It can also be employed for the solid-phase synthesis of nucleotides up to a chain length of about 20 bases. The monomer building block is a protected nucleoside phophodiester derivative (Fig. 8) [193] which reacts with the free 5′-end of a support-bound nucleoside to give a neutral phosphotriester. Various aromatic sulfonic acid derivatives [192, 195, 196], primarily mesitylene sulfonylnitrotriazole (MSNT), serve as condensing agents. The reaction can be accelerated by using catalysts such as *N*-methylimidazole [175]. This method is not widely used any longer due to disadvantages such as long reaction times, a large number of side reactions, and the frequently observed nicks caused by incomplete cleavage of the internucleotide protecting group.

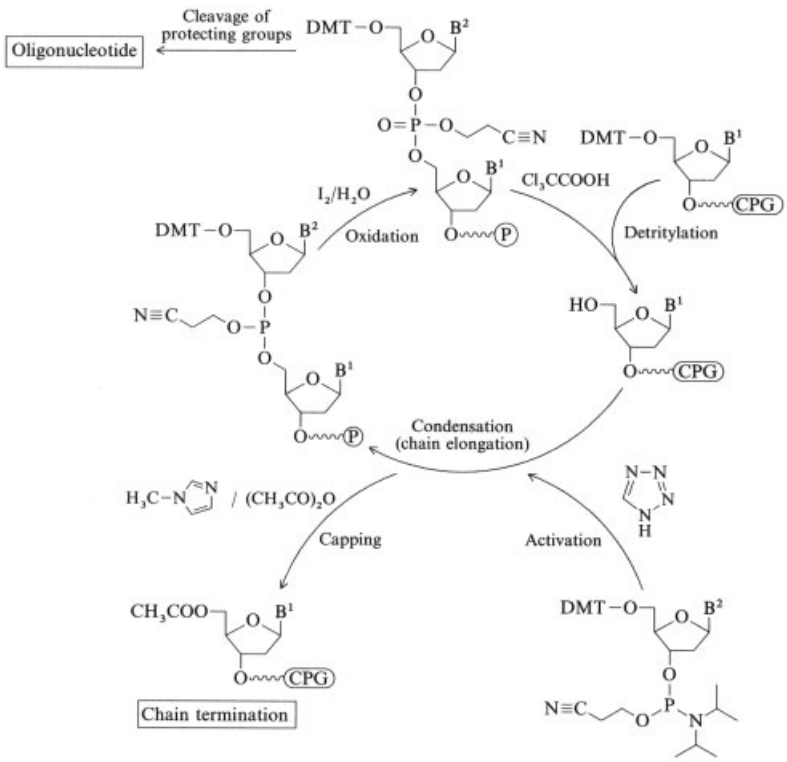

Figure 10. Synthesis cycle according to the phosphoramidite method
B = base; DMTr = 4,4′-dimethoxytrityl residue

Phosphoramidite (Phosphite) Method. This concept is similar to the triester method, but chain synthesis involves trivalent phosphite triester intermediates. The preferred monomer building blocks are the so-called nucleoside phosphoramidites introduced by CARUTHERS et al. [177, 178] that can be activated with tetrazole. After each condensation step, the phosphite triester is converted to the phosphate triester by oxidation with iodine solution. The exceptionally high reactivity of the activated phosphoramidites results in a condensation time of ca. 1 min and $\geq 99\,\%$ yield per chain elongation step.

***H*-Phosphonate Method.** This synthesis method was developed almost simultaneously by GARREGG et al. [197] and by FROEHLER and MATTEUCCI [198]. Condensation is achieved by activation of nucleoside 3′-*H*-phosphonate monomers with pivaloyl or adamantoyl chloride. Since the *H*-phosphonate bond is not cleaved during the synthesis cycle, only one oxidation step is required at the end of the synthesis. This considerably reduces the duration of the cycle. Another advantage over the phosphoramidite method is the higher stability of the monomers. Moreover, excess *H*-phosphonates can be regenerated after the synthesis, thus reducing costs; this makes this method particularly interesting for large-scale synthesis. Since the dinucleoside *H*-phosphonate bond can be easily attacked by nucleophilic reagents [199], oligonucleotide thiophosphates or phosphoamidate analogues can easily be prepared by oxidation with sulfur or amines [199, 200].

Synthesis of Oligoribonucleotides. The solid-phase synthesis of oligoribonucleotides can be conducted by the phosphoramidite [174] or the *H*-phosphonate method [201]. An additional protecting group is required for the 2′-OH group of the ribose [195]. The *tert*-butyldimethylsilyl group has gained acceptance for solid-phase synthesis with ribonucleoside phosphoramidites. It is stable under the acidic conditions required to remove the DMT group and can

be cleaved with tetrabutylammonium fluoride at the end of the synthesis. However, steric hindrance due to the *tert*-butyldimethylsilyl group results in yields of ca. 95 % per chain elongation step. Hence, this method is suitable only for the synthesis of oligomers up to a chain length of about 40 nucleotides.

7.5. Cleavage of Protecting Groups and Purification of Oligonucleotides

Cleavage of Protecting Groups. After synthesis, the oligomer is still completely protected and attached to the support. Cleavage of protecting groups should always begin with the internucleotide protecting group on the phosphorus to prevent strand damage [196, 202].

If the triester method is used, the 2-chlorophenyl group is eliminated by treatment with an "aldoximate" [203]; the base-protecting groups and the support are subsequently removed with 32 % aqueous ammonia at 50 °C. Finally, treatment with 80 % acetic acid cleaves the DMT group, provided it is not required for purification by reversed phase high performance liquid chromatography (HPLC).

If methyl phosphoramidites are used for synthesis, the first deprotection step is treatment with thiophenol to cleave the methyl group [7]. If, however, β-cyanoethyl phosphoramidites are used, then ammonia treatment at 50 °C is sufficient to remove all protecting groups including the support.

Purification of Oligonucleotides. Since condensation does not give 100 % yields per coupling step, short-chain homologues are present in addition to the target sequence. For example, for an oligonucleotide with 70 bases and a 98.5 % yield per condensation step, the target sequence accounts for ≤ 35 % of the total yield. A purification step is therefore required for longer oligomers [204]. High performance liquid chromatography (HPLC) or polyacrylamide gel electrophoresis is commonly used for product isolation. Purification by *reversed phase HPLC* is ideal for the routine separation of oligonucleotides up to ca. 40 – 50 nucleotides [204, 205]. Here, the target sequence that is still tritylated is eluted last. Separation efficiency can be increased with special high-affinity protecting groups allowing purification of chains with 150 bases [206]. If the trityl group is removed before purification, the polyanionic character dominates. It is then more advantageous to separate the desired product by ion-exchange chromatography. *Polyacrylamide gel electrophoresis* under denaturing conditions is the most efficient method for oligonucleotide separation [207, 208]; the homologues are separated according to size. The target sequence, which is usually the longest sequence, has the lowest mobility. The individual fragments become visible on exposure to UV light (λ 254 nm) and are cut out of the gel matrix. Extraction with buffer or electroelution and subsequent desalting (dialysis or gel filtration) yield the desired oligonucleotide.

7.6. Synthesis of Modified Oligonucleotides

Oligonucleotides bind specifically to a defined target sequence under suitable hybridization conditions. Relatively short oligonucleotides ($<$ 20 bases) are used for new drug design strategies involving targeting interference of genetic expression at the level of transcription or translation [209].

For in vivo chemotherapeutic applications based on sequence-specific hybridization (antisense inhibition), nuclease-resistant oligonucleotides are required (the most important are shown in Fig. 11, see next page). They can be obtained by modifying the phosphate backbone of the oligonucleotides, by using α-anomers (**10**) or 2′-O-alkylribosides (**9**). In spite of the changes these oligonucleotides still hybridize with the target sequence and they can also pass through cell membranes more easily due to their lipophilic nature. Applications of these analogues as antisense oligonucleotides are summarized in [209–211].

Phosphothioate analogues (**1**), in which a nonbridging atom of the phosphate group is replaced by a sulfur atom either at a single specific position or at all positions within the chain, can be synthesized by following the phosphoramidite or *H*-phosphonate route [212–214]. The sulfurization reaction is usually carried out with a solution of sulfur in pyridine and carbon disulfide [215]. Other sulfurization reagents that

can also be used in the synthesis are described in [216, 217].

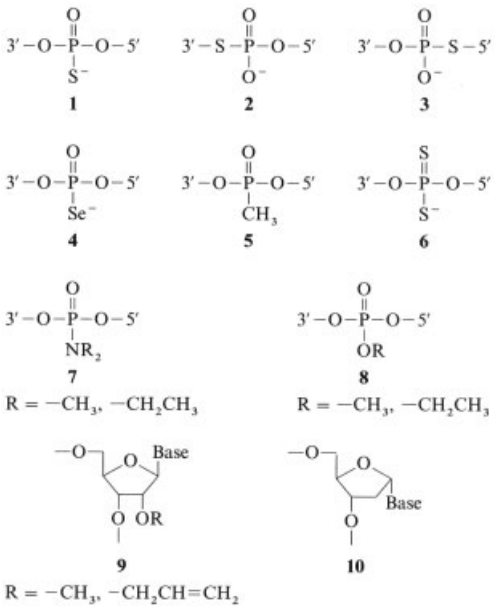

Figure 11. Modified oligonucleotides

Like native DNA, phosphorodithioate analogues (**6**) are stereochemically uniform products and can be prepared by the phosphoramidite method with the corresponding thiophosphoramidite monomers [218].

Methylphosphonates (**5**), phosphoramidates (**7**), and phosphotriesters (**8**) are nonionic and have a more hydrophobic character than naturally occurring nucleic acids. The synthesis of oligonucleoside methylphosphonates is described in [219–222], and is best carried out with nucleoside 3′-methylphosphonous acid imidazolide, which in turn is prepared from methylphosphonous acid bisimidazolide and the fully protected nucleoside. Alternatively, a *H*-phosphonate is reacted with *tert*-butyldimethylsilyl chloride to give silyl phosphite and subsequently oxidized with alkyl or aryl halides [223]. All of the early investigations on the biological activity of these oligonucleotide analogues (which were named "matagen", an acronym for mashing tape for gene expression) were conducted by Ts' O and MILLER [222].

Phosphoramidates (**7**) can be prepared by reacting *H*-phosphonate oligonucleotides with a solution of the appropriate amine in carbon tetrachloride [213].

The 2′-O-methylribonucleotides (**9**) are well established modified nucleosides that are used as monomer building blocks for oligonucleotide synthesis [224]. The synthesis of the corresponding nucleoside phosphoramidites is described in [225]. Oligonucleotides with α-2′-deoxynucleosides (α-DNA) form β-DNA with natural DNA. This form has parallel strands and is also nuclease resistant [226, 227].

To increase the binding of an oligonucleotide to the target, an intercalating agent (e.g., 2-methoxy-6-chloro-9-aminoacridine [228, 229] or phenazine [230]) can be covalently attached to the oligonucleotide. These agents insert "internally" between the base pairs of the double helix, increasing duplex stability.

A second method involves binding of photoreactive groups (e.g., psoralen [231–233], azidophenacyl or azidoproflavin derivatives, proflavins, and porphyrins [234–238]) that are activated by UV or visible light and can cross-link with the target DNA or RNA. "Sequence-specific artificial endonucleases" [239] are oligonucleotides that bear a group which can induce specific cleavage of the target nucleic acid after binding to the target. Such groups are the above-mentioned photoreactive groups: subsequent treatment with piperidine allows cleavage of the target sequence at the cross-linking site. Direct light-induced cleavage at neutral pH is achieved with ellipticine and diacapyrene derivatives [238]. Metal chelates attached to oligonucleotides (e.g., iron – EDTA [240–242], copper – phenanthroline [243–246], or iron – porphyrin [247]) lead to relatively specific cleavage of the target sequence. A nonspecific DNase or RNase bound to a complementary oligonucleotide can also destroy the target nucleic acid [248, 249].

Nonradioactive labeling of synthetic oligonucleotides has become an important technique in molecular biology and several chemical procedures have been developed. Oligonucleotides are covalently attached to "marker groups" such as fluorescent dyes, biotin, and other biologically active groups (see also Section 8.2). These reporter groups can be attached to an oligonucleotide at the 5′-OH terminus, the 3′-OH terminus, the phosphate backbone, or one of the bases. Attachment at the 5′-OH terminus via pro-

tected aminoalkyl or thioalkyl phosphoramidites is most common. These amino- or thiolinkers are added in the last synthesis step and produce a free amino or thiol terminal after cleavage of the protecting groups [250–252]. Subsequent reaction with the activated marker molecule (e.g., biotin-*N*-hydroxysuccinimide) gives the labeled oligonucleotide.

A large number of bifunctional linkers (e.g., maleimidohexanoyl-*N*-hydroxysuccinimide ester) have been described which allow coupling of enzymes [253], peptides, and proteins to linked oligonucleotides via amino groups of the amino acids or thiol groups of cysteine. Attachment at the 3′-OH terminus can be carried out by linking a ribose unit with T 4 RNA ligase, followed by periodate cleavage of the ribose ring, and subsequent reductive amination [254, 255]. Labeling via a 3′-terminal thiol group is described in [256]. Also a modified support is commercially available which produces an aliphatic primary amino group at the 3′-terminus [257]. At the internucleotide bond reporter groups can be attached, e.g., via a phosphoramidate linkage [258–260].

The attachment of reporter groups to modified bases usually occurs at the C-5 position of pyrimidines or the C-8 position of purines [261]. During oligonucleotide synthesis, a fully protected deoxyuridine phosphoramidite [262] modified with a trifluoroacetylaminopropenyl residue is inserted at the desired position in the sequence (usually at C-5); after cleavage of the protecting groups, the group to be coupled is bound via the resulting amino function. Numerous other labeling positions have been described (e.g., the C-4 position [263–265] of 2′-deoxycytidine and the N-6 position of the adenine ring [266] or the exocyclic amino function of the guanine residue [267]).

8. Uses

The direct use of nucleic acids in *nucleic acid probe technology* is becoming increasingly important. Here, labeled nucleic acids are used as probes to detect other nucleic acids by means of hybridization [21, 23]. Identification of the specific nucleotide sequence of a pathogen (virus, bacterium, parasite) can be used to diagnose an infection [268, 269]. More complex analyses are often necessary for the detection of changes in nucleic acids (point mutations, deletions, insertions, translocations, expression level or number of copies of the genetic information) in the diagnosis of hereditary diseases [24], or in the investigation of tumors [25, 26]. This chapter will be limited to describing the uses of nucleic acids in probe technology but other applications will first be briefly summarized.

The second important field of application for nucleic acids is *genetic engineering* (→ Genetic Engineering) which allows the production of proteins in bacteria, fungi, animal or plant cell cultures, or in animals or plants. The nucleic acid that codes for the desired protein is introduced into the organism and expressed.

Antisense technology

is a third field of application of nucleic acids which, however, still only exists in model systems. This technology aims to prevent the expression of nucleic acid sequences that have a negative effect on an organism. This can be achieved by the interception and binding of mRNA with synthetic antisense oligonucleotides (see Section 7.6) or with an antisense RNA from a recombinant expression system. The mRNA forms double strands with the antisense RNA by specific base pairing and cannot be translated [270, 271]. Another method of stopping expression is the specific cleavage of the target nucleic acid by catalytic RNA (ribozymes) [270, 271]. The target nucleic acid is recognized by specific base pairing with the ribozyme. Antisense technology may be used to treat infections caused by viruses, bacteria, or parasites by cleaving or preventing the expression of the genetic information of these pathogens. Another possibility is the treatment of diseases resulting from overexpression of an endogenous gene or its expression in an erroneous target cell.

Finally, nucleic acid technology offers great promise in *research*. As a result of the dramatic improvements in cloning and sequencing large sections of DNA, DNA sequences or genes are increasingly being used as a starting point for the analysis of a biochemical mechanism. Nucleic acid sequences are also used to study evolutionary relationships between different organisms by comparing sequences of conserved genes.

8.1. Hybridization Techniques for Nucleic Acid Detection

Nucleic acid hybridization is based on the fact that two complementary nucleic acid strands (DNA – DNA, RNA – RNA, DNA – RNA) associate to form a double strand according to the rules of base pairing (see also → Genetic Engineering, Chap. 3.1.). The nucleic acid to be detected must be present as a single strand, i.e., denaturation of double strands is required before hybridization. The nucleic acid probe must contain a sequence complementary to that of the nucleic acid under investigation and have a label to permitidentification of the double strands formed by hybridization.

Solid-Phase Hybridization. The labeled hybrid strands must also be distinguishable from unhybridized probes. The simplest way of doing this is to use methods in which the nucleic acid of interest is bound to a solid phase (usually a nitrocellulose or nylon membrane). The potential nonspecific binding sites for the probe on the solid phase are blocked prior to the hybridization reaction with unlabeled nucleic acids of different origins, proteins, and polymers. Consequently, the labeled probe only binds to the solid phase via specific base pairing to the target nucleic acid. The unbound probe is subsequently washed away.

In the *dot blot process* nucleic acid samples are spotted on a membrane [272]. After hybridization the detection of bound label reveals which spots contain the nucleic acid of interest. Quantification is possible by comparison with the signal from a standard [273].

In *Southern blotting* (→ Genetic Engineering) the DNA to be analyzed is separated according to size by gel electrophoresis and subsequently transferred to a membrane without changing the spatial distribution of the DNA fragments [21, 274]. After hybridization the size of the fragment containing the sequences of interest can be analyzed. Analogous separation and hybridization of RNA is referred to as *Northern blotting*.

In *in situ hybridization* the nucleic acid of interest is fixed in such a manner that its natural cellular compartmentalization is maintained [21, 23, 275]. As in histological studies, entire cells are applied and fixed to microscope slides or tissue is cut into ultrathin layers and fixed. Hybridization takes place on the slide between the labeled probes and the nucleic acids in the sample material. Examination under the microscope reveals which and how many cells contain the nucleic acid of interest, as well as its subcellular location. In situ hybridization permits differentiation between individual cells, whereas in solid-phase blotting methods the nucleic acids from many cells are mixed together. In situ hybridization also permits the assignment of a nucleic acid sequence to a chromosome by analyzing cells that are in the metaphase of cell division [276, 277]. The areas of the chromosomes containing the complementary sequences are specifically labeled by the nucleic acid probe.

Hybridization in Solution. Hybridization reactions in which both reaction partners are in solution are more efficient. The hybridized probe, as part of a double strand, can be separated from the unreacted single-stranded probe by *electrophoresis* or *chromatography*.

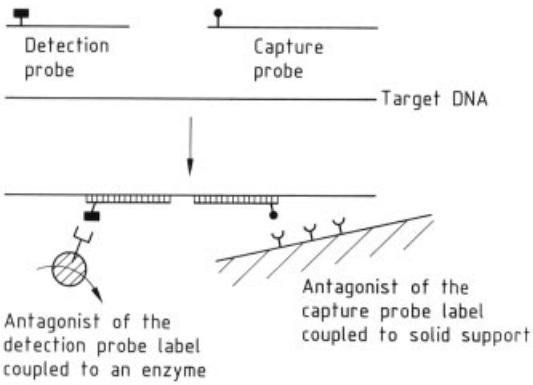

Figure 12. Sandwich hybridization
The oligonucleotide detection and capture probes are labeled with two different labels.

Sandwich hybridization offers an alternative way of separating hybrids formed in solution (Fig. 12) [278, 279]. The nucleic acid of interest is hybridized with two different probes that recognize different sections of the nucleic acid. One of the probes (detection probe) is labeled to allow detection of the hybrid. The second probe acts as a capture probe to bind the hybrid consisting of the nucleic acid of interest and the detection probe to a solid phase. The detection probe only binds to the solid phase if the

nucleic acid of interest is present in the solution and forms a bridge between the two probes. The capture probe can be covalently coupled to the solid phase [280], or can be labeled with a molecule that is distinct from the detection label and is bound to the solid phase via a fixed antagonist. Examples are biotin-labeled probes with a streptavidin- or avidin-loaded solid phase [281] or antigen-loaded probes with an appropriate antibody on the solid phase.

Strand displacement is another method for detecting hybrids formed in solution (Fig. 13) [282, 283]. Here, the probe is a hybrid of a short, labeled oligonucleotide and a longer nucleic acid strand forming a partially single-stranded, partially double-stranded molecule. The longer strand is complementary to the nucleic acid of interest. If the nucleic acid of interest is present in the assay, double strands are formed with the single-stranded region of the probe. This double-stranded region of the unlabeled probe strand and the nucleic acid of interest is longer than the double-stranded region formed between the two probe strands. Therefore complete hybridization occurs between the longer probe strand and the nucleic acid of interest. As a result the short, labeled probe strand is released and can be detected.

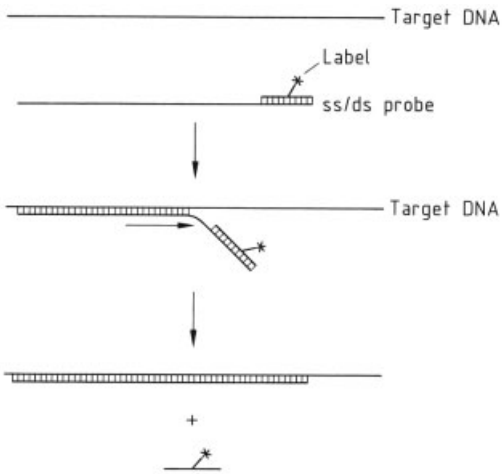

Figure 13. Strand displacement
ss = single-stranded; ds = double stranded

Finally, there are a series of modifications of a hybridization principle which does not require the separation of bound from unbound probe. Here, two labeled probes hybridize with the nucleic acid of interest at adjacent positions. Spatial proximity of the two labels is only guaranteed if the nucleic acid of interest is present and produces a signal that can be measured (e.g., enzymatic activity, light of a certain wavelength) [284].

8.2. Labeling and Detection Systems

Radioactive labeling is losing importance due to the development of sensitive, nonradioactive labels. Nucleic acids are usually radioactively labeled by enzymatic incorporation of nucleotides containing ^{3}H, ^{35}S, or ^{32}P isotopes with polymerases. Alternatively, the 5′-OH terminus of DNA can be labeled with ^{32}P by polynucleotide kinase [21]. Radioactive nucleic acids are detected by exposure to X-ray film or a photoemulsion (in situ hybridization) or by measurement in a scintillation counter.

In nonradioactive methods, the nucleic acid can be directly labeled with a fluorescent or luminescent dye [285], bound to an enzyme (e.g., β-galactosidase, horseradish peroxidase, alkaline phosphatase) or coupled with biotin or a hapten (antigenic determinant) [284]. The hapten (or biotin) is detected by the binding of an appropriate antibody (or streptavidin/avidin). The antibody and streptavidin/avidin are, in turn, directly coupled to a dye molecule or to an enzyme [21]. Additional chemical methods for labeling nucleic acids are described in [21, 286–288] and in Section 7.6.

Photolabeling is a simple and fast labeling method [21, 289]. Biotin or a hapten is coupled to a photoreactive group (photobiotin [289], photodinitrophenyl phosphate (DNP) [290], or photodigoxigenin [291]) which reacts with the amino groups of the nucleotides. On irradiation with light the photoreactive group is activated resulting in the formation of a covalent bond between biotin or the hapten (DNP, digoxigenin) and the nucleic acid.

Nucleic acids from biological sources (e.g., plasmids) are usually labeled with the aid of enzymes. Modified nucleotides are inserted with polymerases [21]; DNA polymerases are used to insert the deoxyribonucleotide triphosphate derivatives (e.g., biotin – dUTP, biotin-dATP, digoxigenin – dUTP, bromo – dUTP, and fluorescein – dUTP) in nick

translation (→ Enzymes, Chap. 6.2.1.) or by random primed labeling. The 3′-terminus of nucleic acids can be labeled with these nucleotides by terminal transferase (→ Enzymes, Chap. 6.2.5.). Modified ribonucleotides (e.g., biotin – UTP and digoxigenin – UTP) are inserted into RNA probes by in vitro transcription with RNA polymerases, particularly with phage-coded RNA polymerases [21, 22] (→ Enzymes, Chap. 6.3.).

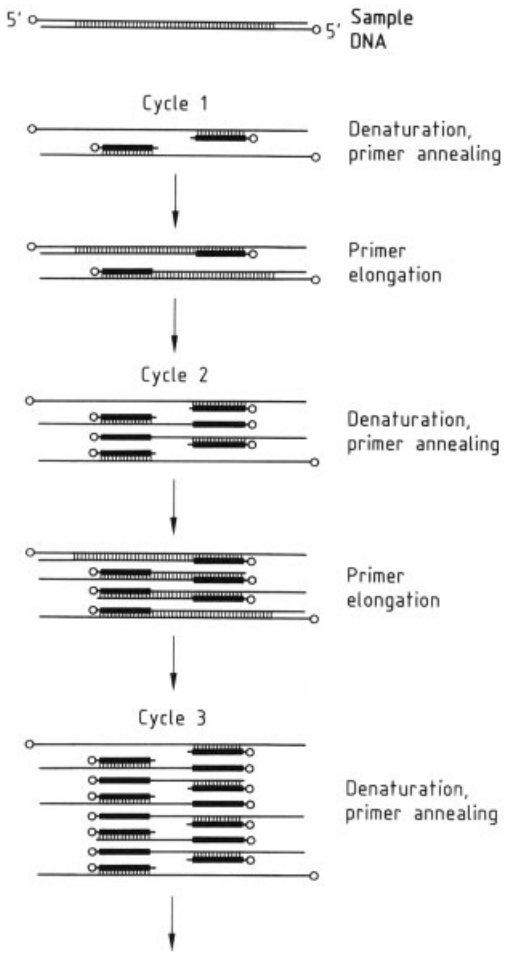

Figure 14. Polymerase chain reaction (PCR)

8.3. Amplification Systems

The sensitivity limit of nucleic acid probe technology can be increased by specifically amplifying (i.e., increasing the amount of) the nucleic acid of interest [21].

Polymerase Chain Reaction (PCR). The polymerase chain reaction is an established technique [15, 27, 292, 293] which is used to amplify a particular nucleic acid from a complex mixture of sequences. The amplified DNA can be subsequently sequenced and/or cloned using genetic engineering techniques, or can be used to increase the sensitivity of pathogen detection. The sensitivity can be increased to such an extent that, for instance, one genome from the human immunodeficiency virus (HIV) integrated in the DNA of a single cell can be detected in the presence of many uninfected cells.

The PCR is based on the repetition of three partial reactions in 20 – 50 cycles (Fig. 14):

1) An oligonucleotide primer is annealed to each of the two denatured complementary strands of the DNA sequence to be amplified.
2) Primer extension catalyzed by a DNA polymerase converts the region to be amplified into two double strands.
3) The double strands are denatured by heating.

Theoretically the number of templates can be doubled in each cycle, so that 2^n templates would be created in n cycles; 25 cycles would result in an amplification rate of 3.4×10^7. However, the partial reactions are not perfectly efficient, and amplification rates of ca. $(1-3) \times 10^6$ are obtained in 25 cycles. Implementation of the PCR is facilitated by the use of thermostable DNA polymerases because otherwise new polymerase would have to be added at the start of each cycle. Typical PCR cycles consist of a few minutes of primer annealing at ca. 55 °C, a few minutes of elongation at ca. 70 °C, and 1 – 2 min of denaturation at 95 °C. These steps can be carried out automatically in programmable "thermocyclers".

RNA can also be amplified by PCR if a DNA template is created in the first step with the help of reverse transcriptase. The PCR product can be detected by electrophoresis with or without previous cleavage with restriction endonucleases, and with or without hybridization with a probe. The PCR product can be labeled directly by using labeled primers [294] or by incorporating labeled nucleotides (Section 8.2) [295, 296]. Con-

tamination is a serious problem in PCR because of the enormous sensitivity of the method. For example, contamination of a sample with a single molecule of the DNA of interest can lead to an incorrect positive result.

Other Amplification Systems. In *transcription-based amplification* oligonucleotide primers are annealed to the nucleic acid of interest which, in addition to complementary sequences to the target, also contain a promoter (binding site) for RNA polymerase [297]. After primer elongation RNA polymerases are used to produce RNA transcripts starting from the promoters. The transcripts can be converted to DNA with reverse transcriptase and used again as targets for the oligonucleotide primers.

The *ligase chain reaction* uses a ligase instead of a polymerase in temperature cycles analogous to the PCR [298, 299]. Two oligonucleotides are annealed adjacent to one another on the template strand and joined by a ligase. The ligated strand is then removed by denaturation and two oligonucleotides can anneal again in the next cycle.

Other amplification systems do not require cycles with different temperatures but proceed isothermally. In *self-sustained sequence replication* (3 SR) [300] and *nucleic acid sequence based amplification* (NASBA) [301], three enzymes act simultaneously at a constant temperature to amplify the target sequence. Reverse transcriptase synthesizes DNA on an RNA template by extending oligonucleotide primers that contain promoter sequences for RNA polymerases. The RNA strand of the resulting DNA–RNA hybrid is digested with RNase H allowing double-stranded DNA to be produced by reverse transcriptase with a second primer. RNA polymerase then transcribes the double-stranded DNA starting at the promoters, and the resulting RNA is again converted to DNA with reverse transcriptase.

Another isothermal amplification technique is based on the replication of RNA genomes with the help of *Qβ replicase*–a RNA-dependent RNA polymerase from phage Qβ [302, 58]. Natural substrates of Qβ replicase are characterized by RNA sequences forming a special secondary structure. For nucleic acid detection these sequences are modified by the insertion of short probe sequences. Unlike the amplification methods described above, the target sequence itself is not multiplied, instead the probe is amplified after it has located the target sequence (signal amplification). Nonspecifically bound and contaminating probe molecules cause background problems that can be alleviated by repeating the hybridization in cycles (reversible target capture) [303].

Another method of signal amplification involves the formation of probe networks in which the target-specific probe has target sequences for secondary probes. A "Christmas tree" of labeled probe sequences is produced for each original target molecule [304–306].

8.4. Applications of Probe Technology

Detection of Infectious Microorganisms. Detection of nucleic acids can be used to identify pathogenic viruses, bacteria, or parasites. In comparison with the detection of proteins or antigens, nucleic acid detection has the advantage that it allows identification of infectious organisms that frequently change their antigens (e.g., HIV) or that exist in a variety of serotypes. Furthermore, viral infections can be detected at the latent stage when no detectable synthesis of viral proteins occurs (e.g., HIV or herpes viruses). DNA probe technology can be more sensitive than immunological techniques due to the possibility of nucleic acid amplification. Although immunological tests are still easier, faster, and cheaper, competitive nucleic acid tests are to be expected in the future.

Detection of the nucleic acid of a pathogen can also serve as a direct measure of the infectiousness of a patient or test material [307]. Quantitative determination of the pathogenic nucleic acid can be used to follow the progress of treatment.

Detection of viruses is possible with a broad spectrum of nucleic acid probes, ranging from synthetic oligonucleotides to the complete cloned viral genome. Infections caused by some virus groups (e.g., papilloma viruses or enteroviruses [308]) can be diagnosed with a probe containing group-specific sequences identical in all members of the group and by using probes with nucleic acid sequences specific for each member.

For the detection of bacteria, the fact that each bacterium contains approximately 10 000 ribosomes and thus 10 000 copies of rRNAs can be used to increase the sensitivity level. The rRNA sequences are relatively well conserved and can used to establish relationships between different organisms in phylogenetic trees. Tests for individual bacterial species, for groups, or even for a wide range of bacteria can be carried out with appropriate oligonucleotide probes [309]. Another group of nucleic acid probes contains sequences of bacterial pathogenicity factors and allows differentiation between pathogenic and apathogenic types of a species (e.g., *Escherichia coli*) [268, 269]. Analogously, detection of resistance genes permits detection of antibiotic resistance and simultaneous identification of the pathogen [310]. Bacterial identification is also applied in the detection of pathogens in food (e.g., *Salmonella*).

Detection of Human DNA Sequences. Nucleic acid probe technology can be used for the *diagnosis of hereditary diseases* and *genetic counseling* [24]. If the genetic basis of the disease is known, the alteration in the nucleic acid can be detected by hybridization with the appropriate region. Such alterations include point mutations (sickle-cell anemia), deletions (Duchenne's muscular dystrophy), or variations in expression (forms of thalassemia). If the genetic cause of the disease is unknown, genetic markers that are inherited together with the defect gene are analyzed, e.g., recognition sites for restriction enzymes (restriction fragment length polymorphism, RFLP) [311].

The analysis of human sequences can also be applied in *tumor diagnosis* [25, 26]. Some human tumors are associated with characteristic chromosomal changes or alterations in genes (oncogenes) [312]. Suitable nucleic acid probes can be used for the highly sensitive detection of DNA sequences of different chromosomes connected by chromosomal translocations or for the detection of mutations in oncogenes that are associated with tumors. For instance, the Philadelphia chromosome (a translocation between chromosome 9 and 22) is characteristic for chronic myelogenous leukemia [313]. Point mutations in the *ras* oncogene are associated with several tumors, and amplification of the *N-myc* gene with neuroblastomas. In the case of recessive oncogenes (tumor suppressor genes) [314, 315], the loss of gene expression results in a tumor; for example, the loss of the retinoblastoma gene results in an eye tumor, retinoblastoma. Another area of tumor diagnosis is the detection of clonality of a leucocyte population caused by the transformation of a white blood cell [316]. Since all lymphocytes have their own characteristic rearrangement of immunoglobulin genes or T-cell receptor genes, the appearance of a defined pattern in many lymphocytes indicates proliferation of a single cell.

Methods that use the genetic diversity of humans to study individuals have wide applicability. Determination of the genes of the major histocompatibility locus (HLA) can be used both for diagnosis (transplantation medicine, prediction of risk of disease, autoimmune diseases) and for differentiation between individuals [317]. Probes which hybridize with repetitive sequences that occur in each individual in a slightly different, genetically determined arrangement can be used for finger printing [28]. Here nucleic acids cleaved by restriction enzymes are separated electrophoretically and hybridized with these probes in the Southern blot technique (Section 8.1). Several fragments of different sizes, each containing these repetitive sequences, produce a signal. Each individual inherits half the sequences from the father and half from the mother, and therefore has a unique fragment pattern [318–320]. This technique is used in forensic medicine to identify criminals by DNA isolated from cells (hair, blood etc.). Other fields of application are in the determination of paternity and family relationships.

9. Nucleosides and Nucleotides

9.1. Nucleosides

The term nucleosides was originally applied to the ribose derivatives of purines that can be isolated from an alkaline RNA hydrolysate. Later, this name was given to all purine and pyrimidine *N*-glycosides of D-ribose and 2-deoxy-D-ribose. Today, this term refers to all natural or synthetic compounds which consist of a heterocyclic nitrogen-containing base (aglycon) and a carbohydrate residue (glycon). The nitrogen atom (*N*-nucleoside) or carbon atom (*C*-

nucleoside) of the heterocycle is linked to the anomeric carbon atom of the sugar residue.

Ribonucleosides and deoxyribonucleosides are obtained from naturally occurring RNA and DNA by enzymatic or chemical hydrolysis. Most nucleosides are stable within a wide pH range. However, under strongly acidic or alkaline conditions at elevated temperature cleavage of the *N*-glycosidic bond may occur. The purine nucleosides are adenosine [*58-61-7*], deoxyadenosine [*958-09-8*], guanosine [*118-00-3*], and deoxyguanosine [*961-07-9*]. The pyrimidine nucleosides are cytidine [*65-46-3*], deoxycytidine [*951-77-9*], thymidine [*50-89-5*], and uridine [*58-96-8*].

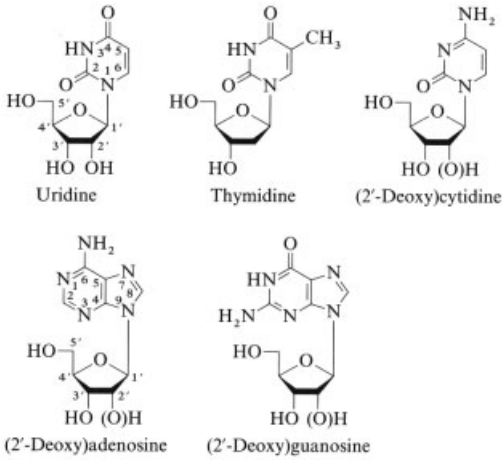

Figure 15. Common nucleosides

The furanoses D-ribose and 2-deoxy-D-ribose are linked to the bases by a β-glycosidic bond at C-1 (Fig. 15). A limited number of nucleosides in the α-D configuration (e.g., 5,6-dimethylbenzimidazole-α-D-ribofuranoside of vitamin B 12) can be isolated from natural substances.

Rare nucleosides and nucleoside antibiotics are found in biopolymers as fermentation products of microorganisms or as building blocks of nucleic acids [29] (see also → Antibiotics, Chap. 3.4.). They exhibit structural modifications of the nucleobase or the glycon residue.

An example of a rare building block of viral DNA is 2′-deoxy-5-hydroxymethylcytidine (**11**) and those of tRNA are the hypermodified nucleosides queuosine [*57072-36-3*] (**12**) and wyosine [*52662-10-9*] (**13**).

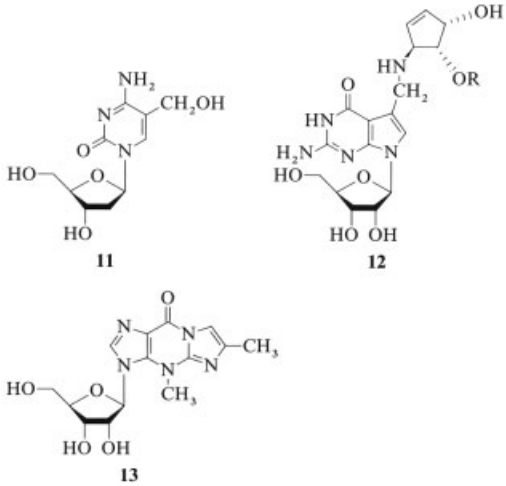

Examples of nucleosides formed by microorganisms are tubercidin [*69-33-0*] (**14**), formycin [*6742-12-7*] (**15**), cordycepin [*73-03-0*] (**16**), blasticidin S [*2079-00-7*] (**17**), vidarabin [*24356-66-9*] (**18**), oxetanocin [*103913-16-2*] (**19**), aristeromycin [*19186-33-5*] (**20**), neplanocin A [*72877-50-0*] (**21**), sinefungin [*58944-73-3*] (**22**), and tunicamycin A [*66081-37-6*] (**23**, $n = 9$), tunicamycin B [*66081-36-5*] (**23**, $n = 10$), tunicamycin C [*66081-37-6*] (**23**, $n = 8$), and tunicamycin D [*66081-38-7*] (**23**, $n = 11$) (Fig. 16).

Naturally occurring nucleosides show a wide variety of biological effects. Accordingly a large number of derivatives have been synthesized in order to find useful chemotherapeutics [30]. The 2′,3′-dideoxyribonucleosides are important synthetic nucleoside derivatives with modified sugar moieties. In the cell the 5′-triphosphate derivatives of these compounds inhibit the reverse transcriptase of retroviruses and thus are potentially important therapeutic drugs (e.g., 3′-azidothymidine [*30516-87-1*] (**24**) (AZT), the first approved drug against HIV infection, see → HIV and AIDS Therapy, Chap. 2.2.).

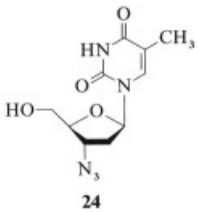

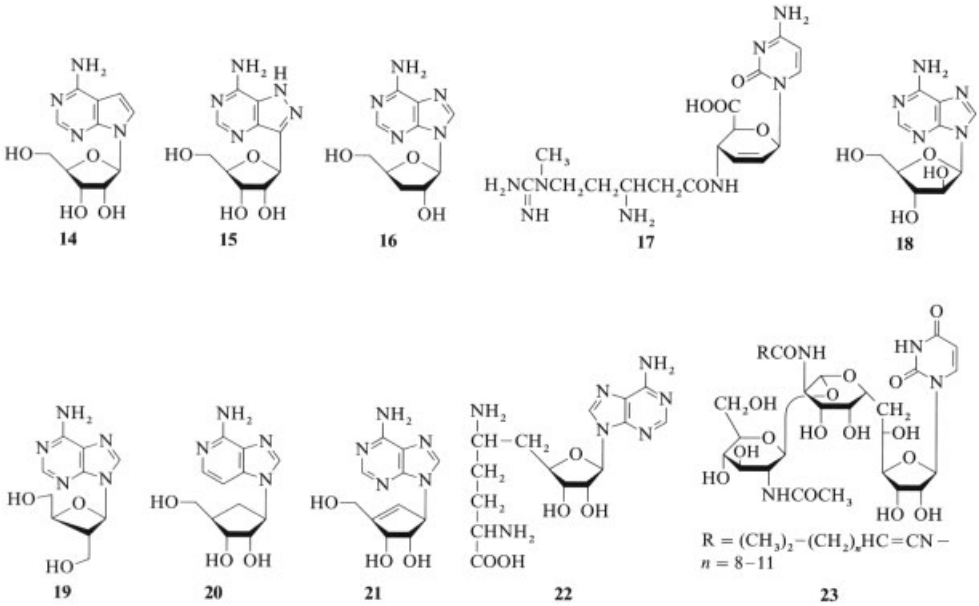

Figure 16. Nucleosides formed by microorganisms

Other derivatives which contain only a part of the ribofuranosyl residue (e.g., acyclonucleosides) also show antiviral activity. This class of substances includes acyclovir (**25**), ganciclovir (**26**) and phosphonylmethyl analogues [e.g., (*S*)-9-(3-hydroxy-2-phosphonylmethoxypropyl) adenine (**27**)]. Nucleosides with a halogen substituent in the heterocyclic moiety are important antiviral drugs (e.g., 5-iodo-2′-deoxyuridine, 5-bromovinyl-2′-deoxyuridine (**28**), or 2-chloro-2′-deoxyadenosine).

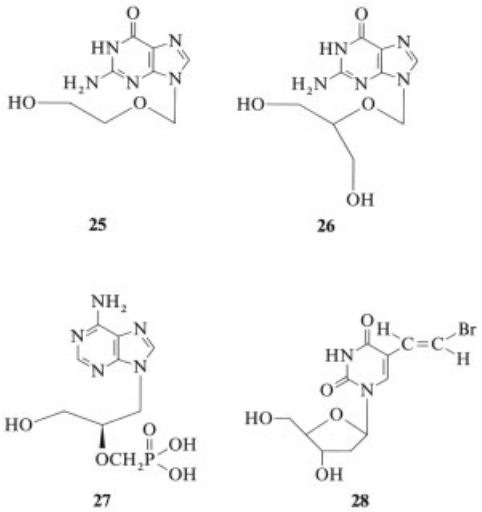

The nucleosides are obtained by modification of existing nucleosides or total synthesis. Total synthesis can be carried out in a linear or convergent manner. Due to the chiral centers on the glycon residue, preparation of optically pure nucleosides requires stereoselective methods of synthesis, especially for the glycosidation of the aglycons. The most important methods of total synthesis will be described briefly. Detailed surveys of the chemical synthesis of and on nucleoside derivatives are given in [31–35].

Synthesis of *N*-Nucleosides. The first syntheses date back to the work of Fischer and Helferich who condensed acetyl-protected α-D-glucopyranosyl halides with the silver salts of heterocycles. The analogous chemical synthesis of the natural nucleosides adenosine and guanosine was described by Todd and coworkers [321]. This method was improved after the introduction of the chloromercury purine derivatives [322] and the use of mercury cyanide (which avoids the tedious preparation of the heavy metal salts of the nucleobases) [323]. The Hilbert–Johnson method was developed for the preparation of pyrimidine nucleosides [324]. This method starts with 2,4-dialkoxypyrimidines, which yield the corresponding nucleosides on reaction with O-acetyl-protected

1-halogenosugars (glycosyl halides) and subsequent ammonolysis or hydrolysis of the 4-alkoxy intermediates [325].

The introduction of the melt process according to SATO [326] represented another advance in nucleoside synthesis. Here, a polyacylated sugar is heated with the purine or pyrimidine base and a catalyst (e.g., 4-toluenesulfonic acid or zinc chloride) in vacuum. This method has the advantage that the free heterocyclic bases (instead of their heavy metal salts) and the 1-O-acetyl sugars (instead of the corresponding glycosyl halides) can be used.

Another synthesis, developed simultaneously by BIRKHOFER and coworkers [327] and NISHIMURA and coworkers [328], involves reaction of trialkylsilylated heterocycles with peracylated 1-halogenosugars in the presence of silver perchlorate as catalyst.

This method was improved by VORBRÜGGEN and is a silyl version of the Hilbert–Johnson method [329]. It is the method of choice for the synthesis of pyrimidine nucleosides. A similar process is based on transglycosylation, i.e., the transfer of the sugar residue of a nucleoside (donor) to a heterocyclic base (acceptor). Apart from chemical transglycosidation, enzymatic sugar transfer is also possible [330].

An elegant method for the synthesis of the nucleosides of purines or purine analogues, developed by SEELA and coworkers, is based on phase transfer catalysis [331]. It gives high yields of the desired β-anomers and the anomeric ratio can be controlled by the choice of catalyst. Finally, reference should be made to the synthesis of the heterocyclic moiety from the C-1 functionalized sugar molecule [332]. This is the method of choice for the synthesis of carbocyclic nucleosides [333].

Synthesis of *C*-Nucleosides. Two methods are available for synthesizing the C–C linkage between the glycosidic and the aglyconic parts of the molecule [334]:

1) Synthesis according to the Fischer–Helferich principle
2) Synthesis via a preformed function on C-1 of the sugar

Unlike *N*-glycosides, the glycosidic bond of *C*-nucleosides is stable to acids and these compounds are not attacked by nucleoside phosphorylases. Consequently *C*-nucleosides are more stable than *N*-nucleosides in biological systems.

9.2. Nucleotides

The term nucleotide generally refers to the phosphate esters of nucleosides. Only mononucleotides (i.e., nucleotides consisting of only one nucleotide unit) are considered here. Mononucleotides can be joined together to form an oligonucleotide (up to 200 nucleotide units) or a polynucleotide (> 200 nucleotide units).

Nucleotides are obtained by the chemical or enzymatic hydrolysis of nucleic acids and by the chemical or enzymatic phosphorylation of nucleosides [335–339]. They may be ribonucleotides (from RNA) or 2′-deoxyribonucleotides (from DNA). Depending on the position at which the sugar moiety is phosphorylated nucleotides can exist as 2′-, 3′-, or 5′-phosphoric acid monoesters. Cyclic 2′,3′- and 3′,5′-phosphoric acid diesters can be formed by intramolecular dehydration.

Cyclic adenosine monophosphate [*60-92-4*] (cAMP) is especially important as a mediator of the action of a large number of hormones (second messenger) [340].

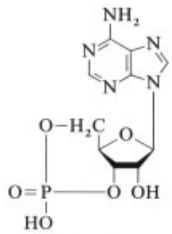

Adenosine
3′,5′-monophosphoric acid
(cyclic AMP)

Cyclic AMP is formed in the cell from ATP by the enzyme adenylate cyclase. It was first synthesized in the laboratory by COOK and coworkers [341]. Numerous attempts have been made to specifically alter the pharmacological effects of cAMP by chemical modification [342].

The nucleotides can contain more than one phosphate group. The 5′-triphosphates of the ribonucleosides and 2′-deoxyribonucleosides of the purines adenine (ATP or dATP) and guanine

(GTP or dGTP) and of the pyrimidines cytosine (CTP or dCTP) and thymine (dTTP) or uracil (UTP) are the most common natural nucleotides. They serve as building blocks for the polymerase-catalyzed synthesis of nucleic acids. Adenosine 5′-triphosphate [*56-65-5*] is also the main source of chemical energy in the living world.

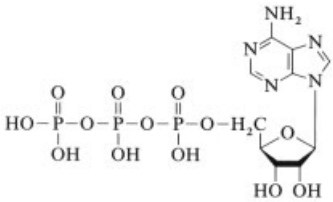

Adenosine
5′-triphosphoric acid
(ATP)

In contrast to many nucleosides, the corresponding phosphate esters are readily soluble in water as their alkali salts. Although the nucleoside monophosphates are stable compounds, the corresponding nucleoside di- and triphosphates undergo slow hydrolytic cleavage of the pyrophosphate bond, even at ambient temperature and neutral pH. Nucleoside mono- and polyphosphates are dephosphorylated by phosphatases (phosphomonoesterases) to yield the corresponding nucleosides.

Analogues containing modified phosphate groups (e.g., thiophosphate [343], imidophosphate [344], or methylene phosphonate functions [345]) are not or only very slowly attacked by phosphomonoesterases. The triphosphate derivatives of those analogues which are only slightly modified compared to the parent nucleotides are valuable tools for studying the reaction mechanisms of ATP-dependent enzyme systems.

The naturally occurring cyclic nucleoside 3′,5′-monophosphates are also very stable. They are converted to acyclic 5′-monophosphates by specific phosphodiesterases. Other important nucleotide building blocks occurring in nature are the adenosine phosphate residues in nicotin-amide adenine dinucleotide phosphate [*53-59-8*] (NADP), coenzyme A [*85-61-0*], and flavine adenine dinucleotide (FAD) [*146-14-5*].

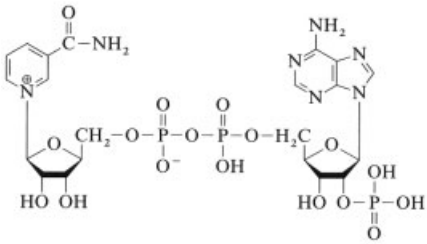

β-Nicotinamide adenine dinucleotide phosphate
(NADP)

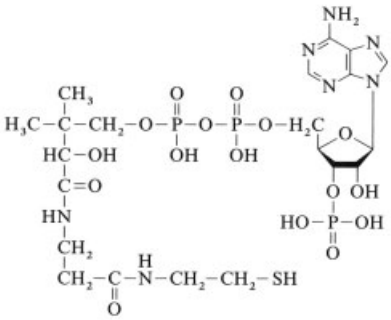

Coenzyme A

Synthesis of Nucleotides. For a review of chemical nucleotide syntheses, see [346]. Nucleoside monophosphates are usually prepared by chemical phosphorylation of the nucleosides with phosphorus oxytrichloride. If the reaction is carried out with a trialkyl phosphate as solvent the 5′-monophosphates are obtained almost exclusively [347]. Protection of the 2′- and/or 3′-OH functions of the sugar is not required. Various methods are available for the chemical synthesis of nucleoside di- and triphosphates from nucleoside monophosphates.

In the *morpholidate method* the nucleoside monophosphate is reacted with morpholine and dicyclohexylcarbodiimide to give the corresponding phosphomorpholidate [348]. The latter is subsequently reacted with orthophosphate to give the nucleoside diphosphate or with pyrophosphate to yield the nucleoside triphosphate.

In the *imidazolidate method* the nucleoside monophosphates are reacted with 1,1′-carbonyldiimidazole to give the corresponding nucleoside monophosphate imidazolidates [349]. These compounds are in turn treated with inorganic ortho- or pyrophosphate to yield the nucleoside di- or triphosphates. In comparison with

the morpholidate method, this process has the advantage of being a "one-pot" reaction.

In the *anhydride method* the nucleoside monophosphates are reacted with diphenyl chlorophosphate [350]. The resulting nucleoside diphenyl pyrophosphates can be subjected to nucleophilic exchange by treatment with anions such as ortho- or pyrophosphate.

These methods can be used analogously to prepare the nucleotides of the coenzyme type by replacing the orthophosphate with, for example, sugar phosphates. The possibility of reacting the nucleoside phosphodichloridate, obtained after treatment with phosphorus oxytrichloride, directly with pyrophosphate in a "one-pot" reaction has been described [351].

Enzymatic syntheses of nucleotides are described in [335–338].

9.3. Therapeutically Important Nucleoside and Nucleotide Derivatives

Antiviral Derivatives. The strategies involved in the development of antiviral nucleosides are reviewed in [352]. For the development and applications of antiviral active nucleotide analogues, see [353]. Of the large number of compounds that have been synthesized since the early 1970s, only a few derivatives have so far been approved. Among them are:

1) 5-iodo-2′-deoxyuridine [*54-42-2*] (IDU)
2) 5-trifluoromethyl-2′-deoxyuridine [*70-00-8*] (trifluorothymidine, TFT)
3) 9-β-D-arabinofuranosyladenine [*24356-66-9*] (Ara-A, vidarabin) (**18**)
4) 3′-azido-3′-deoxythymidine [*30516-87-1*] (AZT, zidovudine) (**24**)
5) 1-β-D-ribofuranosyl-1,2,4-triazole-3-carboxamide [*36791-04-5*] (ribavirin, virazol)
6) 9-[(2-hydroxyethoxy)methyl]guanine [*59277-89-3*] (acyclovir) (**25**)

The following nucleoside analogues are active against HIV infections. However, clinical trials of these compounds are still incomplete or they have not yet been approved partly because of serious side effects:

1) 2′,3′-dideoxycytidine [*7481-89-2*] (ddC), 2′,3′-dideoxyadenosine [*4097-22-7*] (ddA), and 2′,3′-dideoxyinosine [*69655-05-6*] (ddI)
2) 9-(1,3-dihydroxy-2-propoxymethyl)guanine [*82410-32-0*] (ganciclovir, DHPG) (**26**)
3) (*E*)-5-(2-bromovinyl)-2′-deoxyuridine [*73110-56-2*] (BVDU) (**28**)

For further details, see → Chemotherapeutics, Chap. 4.6.1. → Chemotherapeutics, Chap. 4.6.8.

Antitumor Derivatives. For details, see → Cancer Chemotherapy, Chap. 2.4.

10. References

General References

1. L. Stryer: *Biochemistry,* W. H. Freeman, New York 1988.
2. D. Freifelder: *Molecular Biology,* Jones and Bartlett Publishers, Boston 1987.
3. J. Darnell, H. Lodish, D. Baltimore: *Molecular Cell Biology,* Scientific American Books, W. H. Freeman, New York 1990.
4. B. Alberts: *Molecular Biology of the Cell,* Garland Publishing, New York-London 1983.
5. J. D. Watson et al.: *Molecular Biology of the Gene,* The Benjamin/Cummings Publishing Comp., Menlo Park, California 1987.
6. B. Lewin: *Gene,* VCH Verlagsgesellschaft, Weinheim 1988.
7. J. Sambrook, E. F. Fritsch, T. Maniatis: *Molecular Cloning. A Laboratory Manual,* 2nd ed., Cold Spring Harbor Laboratory Press 1989.
8. E.-L. Winnacker: *From Genes to Clones. Introduction to Gene Technology,* VCH Verlagsgesellschaft, Weinheim 1987.
9. A. Kornberg: *DNA Replication,* W. H. Freeman, San Francisco 1980.
10. E. C. Friedberg: *DNA Repair,* W. H. Freeman, New York 1985.
11. W. Saenger: *Principles of Nucleic Acid Structure,* Springer Verlag, New York 1984.
12. J. E. Dahlberg, J. N. Abelson (eds.): *Methods in Enzymology,* vol. 180, Part A, RNA processing, Academic Press, New York 1989.
13. P. M. Wassarman, R. D. Kornberg (eds.): *Methods in Enzymology,* vol. 170, Nucleosomes, Academic Press, New York 1989.
14. H. F. Noller, Jr., K. Moldave (eds.): *Methods in Enzymology,* vol. 164, Ribosomes, Academic Press, New York 1988.

15. M. A. Innis, D. H. Gelfand, J. J. Sninsky, T. J. White (eds.): *PCR Protocols – A Guide to Methods and Applications,* Academic Press, New York 1990.
16. B. S. Sproat, M. J. Gait in M. J. Gait (ed.): *Oligonucleotide Synthesis, A Practical Approach,* IRL Press, Oxford-Washington 1984.
17. H. G. Gassen, A. Lang (eds.): *Chemical and Enzymatic Synthesis of Gene Fragments: A Laboratory Manual,* Verlag Chemie, Weinheim 1982.
18. E. Sonveaux: "The Organic Chemistry Underlying DNA Synthesis," *Bioorg. Chem.* **14** (1986) 274 – 325.
19. J. W. Engels, E. Uhlmann in A. Fichter (ed.): *Gene Synthesis, Advances in Biochemical Engineering/Biotechnology,* vol. 37, Springer Verlag, Berlin – Heidelberg 1988.
20. G. Zon: "Oligonucleotide Analogues as Potential Chemotherapeutic Agents," *Pharm. Res.* 1988, no. 5, 539 – 549.
21. G. H. Keller, M. M. Manak: *DNA Probes,* Macmillan Publishers, New York 1989.
22. H.-J. Hölltke, C. Kessler, *Nucleic Acid Res.* **18** (1990) 5843 – 5851.
23. B. D. Hames, S. J. Higgins (eds.): *Nucleic Acid Hybridization: A Practical Approach,* IRL Press, Oxford 1985.
24. K. E. Davies (ed.): *Human Genetic Diseases: A Practical Approach,* IRL Press, Oxford 1986.
25. J. Cossman (ed.): *The Molecular Genetics and the Diagnosis of Cancer,* Elsevier Science Publishing, New York 1990.
26. M. Furth, M. Greaves (eds.): *Molecular Diagnostics of Human Cancer, Cancer Cells 7,* Cold Spring Harbor Laboratory, New York 1989.
27. H. A. Erlich (ed.): *PCR Technology. Principles and Applications for DNA Amplification,* Stockton Press, New York-London-Tokyo-Melbourne-Hong Kong 1989.
28. L. T. Kirby: *DNA Fingerprinting, An Introduction,* Stockton Press, New York – London – Tokyo –Melborne – Hong Kong 1990.
29. R. J. Suhadolnik: *Nucleoside Antibiotics,* Wiley-Interscience, New York 1970.
30. R. J. Suhadolnik: *Nucleosides as Biological Probes,* Wiley-Interscience, New York 1979.
31. C. A. Decker, L. Goodman in W. Pigman, D. Horton (eds.): *The Carbohydrates,* vol. IIA, Academic Press, New York-London 1970.
32. W. Zorbach, R. S. Tipson: *Synthetic Procedures in Nucleic Acid Chemistry,* vol. I, Interscience Publishers, New York 1968.
33. L. Goodman in P. O. P. Ts'o (ed.): *Basic Principles in Nucleic Acid Chemistry,* **vol. I,** Academic Press, New York-London 1974, p. 93.
34. L. B. Townsend, R. S. Tipson: *Nucleic Acid Chemistry,* Wiley-Interscience, New York, Parts 1 and 2, 1978; Part 3, 1986.
35. Y. Mizuno; *The Organic Chemistry of Nucleic Acids,* Elsevier, Amsterdam 1986.

Specific References

36. G. F. Joyce, *Nature (London)* **338** (1989) 217.
37. A. I. Lamond, T. J. Gibson, *Trends Genet.* **6** (1990) 145.
38. J. D. Watson, F. H. C. Crick, *Nature (London)* **171** (1953) 737.
39. S. B. Zimmerman, *Annu. Rev. Biochem.* **51** (1982) 395.
40. R. Schleif, *Science (Washington, D. C.)* **241** (1988) 1182.
41. L. Joshua-Toret et al., *Nature (London)* **334** (1988) 82.
42. M. Miller et al., *Nature (London)* **334** (1988) 85.
43. A. Nordheim: *ISI Atlas of Science: Biochemistry* **1** (1988) 279.
44. H. R. Drew et al., *Annu. Rev. Cell Biol.* **4** (1988) 1.
45. J. Lebowitz, *TIBS* **15** (1990) 202.
46. J. Grinsted, P. M. Bennett, *Methods Microbiol.* **21** (1988) 129.
47. G. J. Pruss, K. Drlica, *Cell* **56** (1989) 521.
48. A. A. Travers, *Cell* **60** (1990) 177.
49. C. F. McAllister, E. C. Achberger, *J. Biol. Chem.* **264** (1989) 10 451.
50. A. A. Travers, *Annu. Rev. Biochem.* **58** (1989) 427.
51. D. M. Crothers et al., *J. Biol. Chem.* **265** (1990) 7093.
52. T. T. Eckdahl, J. N. Anderson, *Nucleic Acids Res.* **18** (1990) 1609.
53. A. Travers, A. Klug, *Nature (London)* **327** (1987) 280.
54. L. Bracco et al., *EMBO J.* **8** (1989) 4289.
55. M. S. Z. Horwitz, L. A. Loeb, *Science (Washington, D. C.)* **241** (1988) 703.
56. U. R. Müller, W. M. Fitch, *Nature (London)* **298** (1982) 582.
57. R. D. Wells, *J. Biol. Chem.* **263** (1988) 1095.
58. Y. Kohwi, T. Kohwi-Shigematsu, *Proc. Natl. Acad. Sci. USA* **85** (1988) 3781.
59. J. H. van de Sande et al., *Sience (Washington, D. C.)* **241** (1988) 551.

60. L. Perrouault et al., *Nature (London)* **344** (1990) 358.
61. C. Hélène, J.-J. Toulmé, *Biochim. Biophys. Acta* **1049** (1990) 99.
62. C. B. Harley et al., *Nature (London)* **345** (1990) 458.
63. G.-L. Yu et al., *Nature (London)* **344** (1990) 126.
64. V. A. Zakian et al., *Trends Genet.* **6** (1990) 12.
65. W. I. Sundquist, A. Klug, *Nature (London)* **342** (1990) 825.
66. J. D. Boeke, *Cell* **61** (1990) 193.
67. D. Sen, W. Gilbert, *Nature (London)* **344** (1990) 410.
68. F. Baudin, P. J. Romaniuk, *Nucleic Acids Res.* **17** (1989) 2043.
69. J. Normanly, J. Abelson, *Annu. Rev. Biochem.* **58** (1989) 1029.
70. A. Bhattacharyya et al., *Nature (London)* **343** (1990) 484.
71. R. M. W. Mans et al., *Nucleic Acids Res.* **18** (1990) 3479.
72. J. R. Wyatt et al., *BioEssays* **11** (1989) 100.
73. S. Arnott et al., *J. Mol. Biol.* **188** (1986) 631.
74. H. I. Elsner, E. B. Lindblad, *DNA* **8** (1989) 697.
75. H. Donis-Keller et al., *Nucleic Acids Res.* **4** (1977) 2527.
76. G. Krupp, H. J. Gross, *Nucleic Acids Res.* **6** (1979) 3481.
77. J. Parker, *Microbiol. Rev.* **53** (1989) 273.
78. M. Meselson, F. W. Stahl, *Proc. Natl. Acad. Sci. USA* **44** (1958) 671.
79. A. Kornberg, *Biochim. Biophys. Acta* **951** (1988) 235.
80. J. P. Vaughn et al., *Nucleic Acids Res.* **18** (1990) 1965.
81. D. Riesner, H. J. Gross, *Annu. Rev. Biochem.* **54** (1985) 531.
82. T. D. Fox, *Annu. Rev. Genet.* **21** (1987) 67.
83. H. M. Weintraub, *Sci. Am.* **262** (1990) 34.
84. J. Shine, L. Dalgarno, *Proc. Natl. Acad. Sci. USA* **71** (1974) 1342.
85. J. G. Belasco, C. F. Higgins, *Gene* **72** (1988) 15.
86. P. F. Johnson, S. L. McKnight, *Annu. Rev. Biochem.* **58** (1989) 799.
87. A. Baniahmad et al., *Cell* **61** (1990) 505.
88. D. R. Herendeen et al., *Science (Washington, D. C.)* **248** (1990) 573.
89. L. Gold, *Annu. Rev. Biochem.* **57** (1988) 199.
90. J. A. Steitz, *Sci. Am.* **258/6** (1988) 56.
91. M. R. Green, *Annu. Rev. Genet.* **20** (1986) 671.
92. S. M. Mount, *Nucleic Acids Res.* **10** (1982) 459.
93. A. M. Weiner, N. Maizels, *Cell* **61** (1990) 917.
94. V. Volloch et al., *Nature (London)* **343** (1990) 482.
95. D. Shippen-Lentz, E. H. Blackburn, *Science (Washington, D. C.)* **247** (1990) 546.
96. E. H. Blackburn, *J. Biol. Chem.* **265** (1990) 5919.
97. A. E. Dahlberg, *Cell* **57** (1989) 525.
98. L. Gold et al., *Annu. Rev. Microbiol.* **35** (1981) 365.
99. C. O. Gualerzi, C. L. Pon, *Biochemistry* **29** (1990) 5881.
100. W. Doerfler, *Annu. Rev. Biochem.* **52** (1983) 93.
101. E. U. Selker, *TIBS* **15** (1990) 103.
102. W. Zacharias et al., *J. Bacteriol.* **172** (1990) 3278.
103. N. D. F. Grindley, R. R. Reed, *Annu. Rev. Biochem.* **54** (1985) 863.
104. G. C. Walker, *Annu. Rev. Biochem.* **54** (1985) 425.
105. B. van Houten, *Microbiol. Rev.* **54** (1990) 18.
106. G. M. Myles, A. Sancar, *Chem. Res. Toxicol.* **2** (1989) 197.
107. L. A. Grivell, *Nature (London)* **344** (1990) 110.
108. D. L. Robertson, G. F. Joyce, *Nature (London)* **344** (1990) 467.
109. K. Kruger et al., *Cell* **31** (1982) 147.
110. C. Guerrier-Takada et al., *Cell* **35** (1983) 849.
111. O. C. Uhlenbeck, *Nature (London)* **328** (1987) 596.
112. T. R. Cech, *Science (Washington, D. C.)* **236** (1987) 1532.
113. J.-P. Perreault et al., *Nature (London)* **344** (1990) 565.
114. D. Herschlag, T. R. Cech, *Nature (London)* **344** (1990) 405.
115. J. J. Rossi, N. Sarver, *TIBTECH* **8** (1990) 179.
116. G. J. Raymond et al., *Anal. Biochem.* **173** (1988) 125.
117. J. Marmur, *J. Mol. Biol.* **3** (1961) 208.
118. U. Gubler, B. J. Hoffman, *Gene* **25** (1983) 263.
119. J. A. Thompson, R. D. Wells, *Nature (London)* **334** (1988) 87.
120. G. H. Hamm, G. N. Cameron, *Nucleic Acids Res.* **14** (1986) 5 – 9.
121. H. S. Bilofsky et al., *Nucleic Acids Res.* **14** (1986) 1 – 4.
122. E. Y. Chen, P. H. Seeburg, *DNA* **4** (1985) 165.
123. F. Sanger et al., *Proc. Natl. Acad. Sci. USA* **74** (1977) 5463.
124. A. M. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. USA* **74** (1977) 560.

125. M. M. Yang, D. C. Youvan, *Bio/Technology* **7** (1989) 576.
126. J. A. Brumbaugh et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 5610.
127. L. M. Smith, *Genet. Eng.* **10** (1988) 91.
128. W. Ansorge et al., *Nucleic Acids Res.* **15** (1987) 4593.
129. E. R. Mardis, B. A. Roe, *BioTechniques* **7** (1989) 840.
130. J. H. Jett et al., *J. Biomol. Struct. Dyn.* **7** (1989) 301.
131. J. A. Sorge, L. A. Blinderman, *Proc. Natl. Acad. Sci. USA* **86** (1989) 9208.
132. C. D. Carpenter, A. E. Simon, *BioTechniques* **8** (1990) 26.
133. E. P. Nikonowicz et al., *Biochemistry* **29** (1990) 4193.
134. F. J. M. van de Ven, C. W. Hilbers, *Eur. J. Biochem.* **178** (1988) 1.
135. C. de los Santos et al., *Biochemistry* **28** (1989) 7282.
136. S.-H. Chou et al., *Biochemistry* **28** (1989) 2435.
137. E. P. Geiduschek, G. P. Tocchini-Valentini, *Annu. Rev. Biochem.* **57** (1988) 873.
138. J. M. Sperrazza et al., *Gene* **31** (1984) 17.
139. M. I. Oakes, J. A. Lake, *J. Mol. Biol.* **211** (1990) 897.
140. M. I. Oakes et al., *J. Mol. Biol.* **211** (1990) 907.
141. D. D. Dunlap, C. Bustamante, *Nature (London)* **342** (1989) 204.
142. P. K. Hansma et al., *Science (Washington, D. C.)* **242** (1988) 209.
143. R. J. Driscoll et al., *Nature (London)* **346** (1990) 294.
144. C. Malvy, *Anal. Biochem.* **143** (1984) 158.
145. A. E. Dahlberg et al., *J. Mol. Biol.* **41** (1969) 139.
146. S. Peats, *Anal. Biochem.* **140** (1984) 178.
147. C. R. Merril, *Nature (London)* **343** (1990) 779.
148. N. C. Mills, J. Ilan, *Electrophoresis (Weinheim, Fed. Republ. Ger.)* **6** (1985) 531.
149. C. E. Willis, G. P. Holmquist, *Electrophoresis (Weinheim, Fed. Republ. Ger.)* **6** (1985) 259.
150. V. T. Kung et al., *Anal. Biochem.* **187** (1990) 220.
151. A. Glazer et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 3851.
152. J. Maddox, *Nature (London)* **345** (1990) 381.
153. N. C. Stellwagen, *Adv. Electrophor.* **1** (1987) 177.
154. C. R. Cantor et al., *Annu. Rev. Biophys. Chem.* **17** (1988) 287.
155. S. B. Smith et al., *Science (Washington, D. C.)* **243** (1989) 203.
156. S. M. Clark et al., *Sience (Washington, D. C.)* **241** (1988) 1203.
157. E. Lai et al., *BioTechniques* **7** (1989) 34.
158. M. V. Olson, *J. Chromatogr.* **470** (1989) 377.
159. A. M. Michelson, A. R. Todd, *J. Chem. Soc.* 1955, 2632 – 2638.
160. H. G. Khorana, G. M. Tener, J. G. Moffatt, E. H. Pol, *Chem. Ind. (London)* **34** (1956) 1523 – 1531.
161. H. G. Khorana, *Science (Washington, D. C.)* **203** (1979) 614.
162. R. B. Merrifield, *J. Am. Chem. Soc.* **85** (1962) 3821 –3827.
163. R. L. Letsinger, W. B. Lunsford, *J. Am. Chem. Soc.* **85** (1963) 3045 – 3046.
164. H. Schaller, G. Weimann, B. Lerch, H. G. Khorana, *J. Am. Chem. Soc.* **85** (1963) 3821 – 3827.
165. G. S. Ti, B. L. Gaffney, R. A. Jones, *J. Am. Chem. Soc.* **104** (1982) 1316 – 1319.
166. C. B. Reese, A. Ubasawa, *Tetrahedron Lett.* **21** (1980) 2265 – 2268.
167. A. Krazewski, J. Stawinski, M. Wiewiorowski, *Nucleic Acids Res.* **8** (1980) 2301 – 2305.
168. R. T. Pon, N. Usman, M. J. Damha, K. K. Ogilvie, *Nucleic Acids Res.* **14** (1985) 6453 – 6469.
169. K. K. Ogilvie, R. T. Pon, M. J. Damha, *Nucleic Acids Res.* **13** (1985) 6447 – 6465.
170. F. Himmelsbach et al., *Tetrahedron Lett.* **11** (1981) 59 – 72.
171. B. C. Froehler, M. D. Matteucci, *Nucleic Acids Res.* **22** (1983) 8031 – 8036.
172. J. C. Schulhof, D. Molko, R. Teoule, *Nucleic Acids Res.* **15** (1987) 397 – 416.
173. J. C. Schulhof, D. Molko, R. Teoule, *Nucleic Acids Res.* **16** (1988) 319 – 326.
174. N. Usman, K. K. Ogilvie, M.-Y. Jiang, R. J. Cedergren, *J. Am. Chem. Soc.* **109** (1987) 7845 – 7854.
175. B. S. Sproat, M. J. Gait in M. J. Gait (ed.): *Oligonucleotide Synthesis, A Practical Approach,* IRL Press, Oxford – Washington 1984, pp. 83 – 115.
176. C. B. Reese, R. C. Titmas, L. Yau, *Tetrahedron Lett.* **19** (1978) 2727 – 2730.
177. S. L. Beaucage, M. H. Caruthers, *Tetrahedron Lett.* **22** (1981) 1859 – 1862.
178. L. J. McBride, M. H. Caruthers, *Tetrahedron Lett.* **24** (1983) 245.
179. N. D. Sinha, J. Biernat, J. McMagnus, H. Köster, *Nucleic Acids Res.* **12** (1984) 4539 – 4557.

180. H. Ito, Y. Ike, S. Ikuta, K. Itakura, *Nucleic Acids Res.* **10** (1982) 1755.
181. K. Miyoshi, R. Arentzen, T. Huang, K. Itakura, *Nucleic Acids Res.* **8** (1980) 5507.
182. R. Crea, T. Horn, *Nucleic Acids Res.* **8** (1980) 2331.
183. M. D. Matteucci, M. H. Caruthers, *J. Am. Chem. Soc.* **103** (1981) 3171 – 3174.
184. V. A. Efimov, S. V. Reverdatto, O. B. Chakhmakhcheva, *Nucleic Acids Res.* **9** (1982) 6675 – 6694.
185. S. P. Adams et al., *J. Am. Chem. Soc.* **105** (1983) 661 – 663.
186. M. J. Nemer, K. K. Ogilvie, *Tetrahedron Lett.* **21** (1980) 4159 – 4162.
187. T. G. Wadell, D. E. Leyden, M. T. de Bello, *J. Am. Chem. Soc.* **103** (1981) 5303 – 5307.
188. M. H. Caruthers in H. G. Gassen, A. Lang (eds.): *Chemical and Enzymatic Synthesis of Gene Fragments: A Laboratory Manual,* Verlag Chemie, Weinheim 1982, pp. 71 – 79.
189. W. Bannwarth, P. Iaiza, *DNA* **5** (1986) 413 – 419.
190. F. Chow, T. Kempe, G. Palm, *Nucleic Acids Res.* **9** (1981) 2807.
191. J. C. Catlin, F. Cramer, *J. Org. Chem.* **38** (1973) 245 –250.
192. S. A. Narang, H. M. Hsiung, R. Brousseau, *Methods Enzymol.,* **68** (1979) 90 – 98.
193. S. S. Jones et al., *Tetrahedron* **36** (1980) 3075 – 3085.
194. J. H. van Boom et al., *J. Chem. Soc. Chem. Commun.* 1971, 869 – 871.
195. E. Ohtsuka, S. Iwai in S. Narang (ed.): *Synthesis and Application of DNA and RNA,* Academic Press, San Diego – New York – Berkley – Boston 1987.
196. J. H. van Boom, *J. Chem. Soc. Chem. Commun.* 1976, 167.
197. P. J. Garegg, T. Regberg, J. Stawinski, R. Strömberg, *Chem. Scr.* **25** (1985) 280 – 282.
198. B. C. Froehler, M. D. Matteucci *Tetrahedron Lett.* **27** (1986)469 – 472.
199. B. C. Froehler, *Tetrahedron Lett.* **27** (1986) 5575 –5578.
200. B. C. Froehler, P. Ng, M. D. Matteucci, *Nucleic Acids Res.* **16** (1988) 4831 – 4839.
201. J. Stawinski, R. Strömberg, M. Thelin, E. Westman, *Nucleic Acids Res.* **16** (1988) 9285 – 9298.
202. J. H. van Boom et al., *Nucleic Acids Res.* **4** (1977) 1047.
203. E. E. van Tamelen, S. V. Daub, *J. Am. Chem. Soc.* **99** (1977) 3526.
204. L. W. McLaughlin, J. U. Krusche in H. G. Gassen, A. Lang (eds.): *Chemical and Enzymatic Synthesis of Gene Fragments A Laboratory Manual,* : Verlag Chemie, Weinheim 1982.
205. L. W. McLaughlin, N. Piel in M. J. Gait (ed.): *Oligonucleotide Synthesis, A Practical Approach,* IRL Press, Oxford – Washington 1984, pp. 117 –132.
206. G. Schmidt, H. Seliger, *J. Chromatogr.* **397** (1987) 141 – 151.
207. R. Wu in M. J. Gait (ed.): *Oligonucleotide Synthesis, A Practical Approach,* IRL Press, Oxford – Washington 1984, pp. 135 – 151.
208. A. M. Maxam, W. Gilbert, *Methods Enzymol.* **65** (1980) 499 – 503.
209. G. Zon, *Pharmaceutical Research* **5** (1988) no. 9, 539 – 549.
210. C. Hélène, J.-J. Toulmé, *Biochim. Biophys. Acta* **1049** (1990) 99 – 125.
211. C. A. Stein, J. S. Cohen, *Cancer Research* **48** (1988) 2659 – 2668.
212. W. J. Stec, G. Zon, W. Egan, B. Stec, *J. Am. Chem. Soc.* **106** (1984) 6077 – 6079.
213. B. C. Froehler, *Tetrahedron Lett.* **27** (1986) 5565 – 5568.
214. A. Andrus, J. W. Efcavitch, L. J. McBride, B. Giusti, *Tetrahedron Lett.* **29** (1988) 861 – 864.
215. B. H. Dahl, K. Bjergarde, V. B. Sommer, O. Dahl, *Nucleosides & Nucleotides,* **8** (1989) 1023 – 1027.
216. S. L. Beaucage et al., *J. Am. Chem. Soc.* **112** (1990) 1254 – 1255.
217. P. C. J. Kamer et al., *Tetrahedron Lett.* **30** (1989) 6757 – 6760.
218. W. K.-D. Brill, J.-Y. Tang, Y.-X. Ma, M. H. Caruthers, *J. Am. Chem. Soc.* **111** (1989) 2321 – 2322.
219. P. S. Miller, K. N. Fang, N. S. Kondo, P. O. P. Ts'o, *J. Am. Chem. Soc.* **93** (1971) 6657 – 6665.
220. P. S. Miller et al., *Biochimie* **67** (1985) 769 – 776.
221. P. S. Miller et al., *Nucleic Acids Res.* **11** (1983) 6225 – 6242.
222. P. S. Miller et al., *Biochemistry* **25** (1986) 5092 – 5097.
223. E. de Vroom et al., *Recl. Trav. Chim. Pays-Bas* **106** (1987) 65 – 66.
224. H. Inoue et al., *Nucleic Acids Res.* **15** (1987) 6131 – 6148.
225. B. S. Sproat, B. Beijer, A. Iribarren, *Nucleic Acids Res.* **18** (1990) 41 – 49.
226. U. Sequin, *Experientia* **29** (1973) 1059 – 1062.

227. J. L. Imbach, B. Rayner, F. Morvan, *Nucleosides & Nucleotides* **8** (1989) 627 – 648.
228. C. Hélène et al., *Biochimie* **66** (1985) 777 – 783.
229. A. Zerial, N. T. Thuong, C. Hélène, *Nucleic Acids Res.* **15** (1987) 9909 – 9919.
230. V. V. Vlassov et al., *Nucleic Acids Res.* **14** (1986) 4065 – 4076.
231. H. B. Gamper, G. D. Cimino, J. E. Hearst, *J. Mol. Biol.* **197** (1987) 349 – 362.
232. Y. B. Shi, H. B. Gamper, J. E. Hearst, *J. Mol. Biol.* **263** (1988) 527 – 534.
233. P. S. Miller, P. O. P. Ts'o, *Anti-Cancer Drug Des.* **2** (1987) 117 – 128.
234. U. Asseline et al., *Proc. Natl. Acad. Sci USA* **81** (1984) 3297 – 3301.
235. Le Doan et al., *Nucleic Acids Res.* **15** (1987) 7749 – 7760.
236. D. Praseuth, L. Perrouault, T. Le Doan, *Proc. Natl. Acad. Sci USA* **85** (1988) 1349 – 1353.
237. D. Praseuth et al., *Biochemistry* **27** (1988) 3031 – 3038.
238. C. Hélène, T. Le Doan, N. T. Thuong in P. E. Nielsen (ed.): *Photochemical Probes in Biochemistry,* Kluwer Publ., Norwell MA, 1989, pp. 219 – 229.
239. C. Hélène, N. T. Thuong, T. Saison-Behmoaras, J. C. Francois, *Tibtech* **7** (1989) 310 – 315.
240. A. S. Boutorin et al., *FEBS Lett.* **172** (1984) 43 –46.
241. B. C. F. Chu, L. E. Orgel, *Proc. Natl. Acad. Sci USA* **82** (1985) 963 – 967.
242. G. B. Dreyer, P. E. Dervan, *Proc. Natl. Acad. Sci. USA* **82** (1985) 968 – 972.
243. C. H. B. Chen, D. S. Sigman, *Proc. Natl. Acad. Sci. USA* **83** (1986) 7147 – 7151.
244. C. H. B. Chen, D. S. Sigman, *J. Am. Chem. Soc.* **110** (1988) 6570 – 6572.
245. J.-C. Francois et al., *Biochemistry* **27** (1988) 2272 – 2276.
246. J.-C. Francois et al., *J. Biol. Chem.* **264** (1989) 5891 – 5898.
247. T. Le Doan et al., *Nucleic Acids Res.* **15** (1987) 8643 – 8659.
248. D. Corey, P. Schultz, *Science (Washington, D. C.)* **238** (1987) 1401 – 1403.
249. R. N. Zuckermann, P. G. Schultz, *Proc. Natl. Acad. Sci. USA* **86** (1989) 1766 – 1770.
250. Applied Biosystems: DNA Synthesizer, User Bulletin, No. 49, Foster City 1988.
251. Beckman Instruments, Aminomodifiers and Easy Label Kits, User Manual, Palo Alto 1990.
252. B. C. F. Chu, G. M. Wahl, L. E. Orgel, *Nucleic Acids Res.* **11** (1985) 6513 – 6529.
253. E. Jablonski, E. W. Moomaw, R. H. Tullis, J. L. Ruth, *Nucleic Acids Res.* **14** (1986) 6115 – 6128.
254. M. Lemaitre, B. Bayard, B. Lebleu, *Proc. Natl. Acad. Sci. USA* **84** (1987) 648 – 652.
255. M. Lemaitre, C. Bisbal, B. Bayard, B. Lebleu, *Nucleosides & Nucleotides* **6** (1987) 311 – 315.
256. R. Zuckermann, D. Corey, P. Schultz, *Nucleic Acids Res.* **15** (1987) 5305 – 5321.
257. P. S. Nelson, R. A. Frye, E. Liu, *Nucleic Acids Res.* **17** (1989) 7187 – 7194.
258. S. Agrawal, J.-Y. Tang, *Tetrahedron Lett.* **31** (1990) 1543 – 1546.
259. R. L. Letsinger, M. E. Schott, US 4 547 569, 1985.
260. A. Jäger, M. J. Levy, S. M. Hecht, *Biochemistry* **27** (1988) 7237 – 7246.
261. R. P. Langer, A. A. Waldrop, D. C. Ward, *Proc. Natl. Acad. Sci. USA* **78** (1981) 6635 – 6637.
262. A. F. Cook, E. Vuocolo, C. L. Brakel, *Nucleic Acids Res.* **16** (1988) 4077 – 4095.
263. M. S. Urdea et al., *Gene* **61** (1987) 253 – 264.
264. I. C. Gillam, G. M. Tener, *Anal. Biochem.* **157** (1986) 199 – 207.
265. U. Pieles, B. S. Sproat, P. Neuner, F. Cramer, *Nucleic Acids Res.* **17** (1989) 8967 – 8978.
266. G. Gebeyehu et al., *Nucleic Acids Res.* **16** (1987) 4937 – 4534.
267. P. Tchen, R. P. P. Fuchs, E. Sage, M. Leng, *Proc. Natl. Acad. Sci. USA* **81** (1984) 3466 – 3470.
268. J. A. Washington, G. L. Woods in B. Swaminathan, G. Prakash (eds.): *Nucleic Acid and Monoclonal Antibody Probes,* Marcel Dekker, New York 1989, p. 319.
269. F. C. Tenover, *Clin. Microbiol. Rev.* **1** (1988) 82.
270. E. Uhlmann, A. Peyman, *Chemical Rev.* **90** (1990) 543 – 584.
271. C. Hélène, J.-J. Toulmé, *Biochim. Biophys. Acta* **1049** (1990) 99 – 125.
272. J. Brandsma, G. Miller, *Proc. Natl. Acad. Sci. USA* **77** (1980) 6851.
273. P. McIntyre, G. R. Stark, *Anal. Biochem.* **174** (1988) 209.
274. E. M. Southern, *J. Mol. Biol.* **98** (1975) 503.
275. T. R. Moench, *Mol. Cell. Probes* **1** (1987) 195.
276. J. E. Landegent et al., *Nature (London)* **317** (1985) 175.
277. P. Lichter et al., *Science (Washington, D. C.)* **247** (1990) 64.
278. M. Ranki et al., *Gene* **21** (1983) 77.

279. P. J. Nicholls, A. D. B. Malcolm, *J. Clin. Lab. Anal.* **3** (1989) 122.
280. M. Virtanen et al., *J. Clin. Microbiol.* **20** (1984) 1083.
281. A. C. Syvaenen et al., *Nucleic Acids Res.* **14** (1986) 5037.
282. M. S. Ellwood et al., *Clin. Chem.* **32** (1986) 1631.
283. C. P. H. Vary, *Nucleic Acids Res.* **15** (1987) 6883.
284. J. A. Matthews, L. J. Kricka, *Anal. Biochem.* **169** (1988) 1.
285. L. J. Arnold, Jr., et al., *Clin. Chem.* **35** (1989) 1588.
286. R. P. Viscidi et al., *J. Clin. Microbiol.* **23** (1986) 311.
287. A. H. N. Hopman et al., *Exp. Cell. Res.* **169** (1987) 357.
288. G. H. Keller et al., *Anal. Biochem.* **170** (1988) 441.
289. A. C. Forster et al., *Nucleic Acids Res.* **13** (1985) 745.
290. G. H. Keller et al., *Anal. Biochem.* **177** (1989) 392.
291. K. Mühlegger et al., *Biol. Chem. Hoppe-Seyler* **371** (1990) 953.
292. K. B. Mullis, F. A. Faloona, *Methods Enzymol.* **155** (1987) 335.
293. R. K. Saiki et al., *Science (Washington, D. C.)* **230** (1985) 1350.
294. H. Lee et al., *Science (Washington, D. C.)* **244** (1989) 471.
295. D. B. Schowalter, S. S. Sommer, *Analyt. Biochemistry* **177** (1988) 90–94.
296. T. Lion, O. A. Haas, *Analyt. Biochemistry* **188** (1990) 335–337.
297. D. Y. Kwoh et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 1173.
298. D. Y. Wu, R. B. Wallace, *Genomics* **4** (1989) 560.
299. K. J. Barringer et al., *Gene* **89** (1990) 117.
300. J. C. Guatelli et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 1874.
301. J. van Brunt, *Bio/Technology* **8** (1990) 291.
302. P. M. Lizardi et al., *Bio/Technology* **6** (1988) 1197.
303. H. Lomeli et al., *Clin. Chem. (London)* **35** (1989) 1826.
304. H. Wolf et al., *J. Virol. Methods* **13** (1986) 1.
305. M. S. Urdea et al., *Gene* **61** (1987) 253.
306. P. D. Fahrlander, *Bio/Technology* **6** (1988) 1165.
307. M. Berninger et al., *J. Med. Virol.* **9** (1982) 57.
308. H. A. Rotbart et al., *J. Clin. Microbiol.* **26** (1988) 2669.
309. J. J. Hogan in B. Swaminathan, G. Prakash (eds.): *Nucleic Acid and Monoclonal Antibody Probes,* Marcel Dekker, New York 1989.
310. S. Huovinen et al., *Antimicrob. Agents Chemother.* **32** (1988) 175.
311. D. N. Cooper, J. Schmidtke, *Hum. Genet.* **73** (1986) 1.
312. S. Nishimura, T. Sekiya, *Biochem. J.* **243** (1987) 313.
313. P. Benn et al., *Cancer Genet. Cytogenet.* **29** (1987) 1.
314. A. J. Levine, *BioEssays* **12** (1990) 60.
315. R. A. Weinberg, *Cancer Res.* **49** (1989) 3713.
316. M. A. Lovell, *Clin. Chem. (Winston-Salem, N.C.),* Suppl. 35 (1989) 1343.
317. H. A. Erlich et al., *Bio/Technology* **4** (1986) 975.
318. A. J. Jeffreys, *Biochem. Soc. Trans.* **15** (1987) 309.
319. H. Zischler et al., *Hum. Genet.* **82** (1989) 227.
320. A. H. Cawood, *Clin. Chem. (Winston-Salem, N.C.)* **35** (1989) 1832.
321. J. Davoll, B. Lythgoe, A. R. Todd, *J. Chem. Soc.* 1948, 1685.
322. J. Davoll, B. A. Lowy, *J. Am. Chem. Soc.* **73** (1951) 1650.
323. N. Yamaoka, K. Aso, H. Matsuda, *J. Org. Chem.* **30** (1965) 149.
324. G. E. Hilbert, T. B. Johnson, *J. Am. Chem. Soc.* **52** (1930) 4489.
325. J. Pliml, M. Prystas, *Adv. Heterocycl. Chem.* **8** (1967) 115.
326. T. Sato, T. Shimadate, Y. Ishido, *Chem. Abstr.* **56** (1962) 11 692 g.
327. L. Birkhofer, A. Ritter, H. P. Küelthau, *Angew. Chem.* **75** (1963) 209.
328. T. Nishimura, B. Shimizu, I. Iwai, *Chem. Pharm. Bull.* **11** (1963) 1470.
329. U. Niedballa, H. Vorbrüggen, *Angew. Chem., Int. Ed. Engl.* **9** (1970) 461.
330. A. Holy, I. Votruba, *Nucleic Acid Symp. Ser.* **18** (1987) 69.
331. F. Seela, H. D. Winkeler, *J. Org. Chem.* **47** (1982) 226.
332. D. H. Shannahoff, R. A. Sanchez, *J. Org. Chem.* **38** (1973) 593.
333. V. E. Marquez, M. I. Lim, *Med. Res. Rev.* **6** (1986) 1.
334. S. Hanessian, A. C. Parnet, *Adv. Carbohydr. Chem. Biochem.* **33** (1976) 111.
335. K. Ogata, *Adv. Appl. Microbiol.* **19** (1975) 209.
336. K. Ogata, S. Kinoshita, T. Tsunoda, K. Aida: *Microbial Production of Nucleic Acid-Related Substances,* Halsted Press, New York 1976.

337. S. L. Haynie, D. N. Whitesides, *Appl. Biochem. Biotech.* **23** (1990) 205.
338. D. J. Merkler, V. L. Schramm, *Anal. Biochem.* **167** (1987) 148.
339. L. A. Slotin, *Synthesis* 1977, 737.
340. G. A. Robinson, R. W. Butcher, E. W. Sutherland: *Cyclic AMP,* Academic Press, New York – London 1971.
341. W. H. Cook, D. Lipkin, R. Markham, *J. Am. Chem. Soc.* **79** (1957) 3607.
342. M. S. Amer in N. J. Harper, A. B. Simmonds (eds.): *Advances in Drug Research,* **vol. 12,** Academic Press, New York – London 1977, p. 1.
343. F. Eckstein, *Ann. Rev. Biochem.* **54** (1985) 367.
344. R. G. Yount, D. Babcock, D. Ojala, W. Ballantyne, *Biochemistry* **10** (1971) 2484.
345. T. C. Myers, K. Nakamura, J. W. Flesher, *J. Am. Chem. Soc.* **85** (1963) 3292.
346. G. R. Pettit: *Synthetic Nucleotides,* vol. I, Van Nostrand Reinhold, New York 1972.
347. M. Yoshikawa, T. Kato, T. Takenishi, *Tetrahedron Lett.* **50** (1967) 5065.
348. J. G. Moffatt, H. G. Khorana, *J. Am. Chem. Soc.* **83** (1961) 649.
349. D. E. Hoard, D. G. Ott, *J. Am. Chem. Soc.* **87** (1965) 1785.
350. A. M. Michelson, *Biochim. Biophys. Acta* **91** (1964) 1.
351. J. Ludwig, *Acta Biochim. Biophys. Acad. Sci. Hung.* **16** (1981) 131.
352. E. De Clercq, *Int. Congr. Ser. Excerpta Med.* **750** (1987) 631.
353. J. C. Martin: "Nucleotide Analogues as Antiviral Agents," *ACS Symp. Ser.* **401** (1989).

Nylon → Fibers, 4.Synthetic Organic
Nylon → Polyamides

Octane Enhancers

Automotive Fuels and Methyl Tert-Butyl Ether are separate keywords.

Kuno Schädlich, VEBA OEL AG, Gelsenkirchen, Federal Republic of Germany (Chap. 1 – 4, Sections 6.2, 6.3)

Peter Schug, VEBA OEL AG, Gelsenkirchen, Federal Republic of Germany (Chap. 1 – 4, Sections 6.2, 6.3)

Jörg Fabri, Deutsche Shell AG, Hamburg, Federal Republic of Germany (Chap. 6.1)

Werner Dabelstein, Deutsche Shell AG, Hamburg, Federal Republic of Germany (Chap. 6.1)

Arno Reglitzky, Deutsche Shell AG, Hamburg, Federal Republic of Germany (Chap. 6.1)

1. Introduction 1
2. Technical Information 2
2.1. Combustion in Otto Engines 2
2.2. Knock Phenomena 2
2.3. Octane Number Measurement . . . 2
3. History 3
4. Types of Octane Enhancers 3
5. Ashless Octane Enhancers 3
5.1. Blend Octane Numbers 4
5.2. Nitrogen-Containing Components . 5
5.3. Oxygen-Containing Components . . 5
5.3.1. EC Directive on Fuel Oxygenates . . . 5
5.3.2. Blend Octane Numbers of Oxygenates 6
5.3.3. Other Properties of Oxygenates 7
5.3.4. Production 8
5.3.5. Environmental Protection 9
5.3.6. Toxicology and Occupational Health . 10
5.3.7. Storage and Transportation 10
6. Metal-Containing Additives 10
6.1. Alkyllead Compounds 10
6.2. Methylcyclopentadienylmanganese tricarbonyl 12
6.3. Ferrocene 13
7. References 14

1. Introduction

Historically, the quality of motor gasoline and the technical development of the spark ignition Otto engine were closely connected. Since its invention in 1876, the Otto engine technology has been continuously improved, a process which is still going on.

Of particular importance for this development was the parallel improvement in gasoline knock resistance. Knock resistance is one of the most important gasoline properties and its improvement allowed the engine manufacturers to increase the compression ratio. This in turn resulted in higher engine efficiency and improved exhaust emission performance. The knock resistance is measured in terms of gasoline octane numbers in accordance with worldwide standardized test procedures.

Figure 1 (see next page) shows, for Germany as an example, the close relationship between octane number and engine compression ratio from the turn of the century until now. Over this period a variety of measures were applied to improve refinery technologies to meet the increasing engine octane demand. These measures incorporated the integration of complex processing plants, from which a range of gasoline component streams with different octane levels became available. In 1929, *alkyllead compounds* were introduced in the United States as efficient octane enhancers. Thereafter these were used generally to improve the gasoline octane quality until environmental concerns led to their use being stepwise restricted worldwide. The resulting shortfall of octane was compensated by improvements in refinery processes, as well as by the use of alternative octane enhancers such as fuel oxygenates. Starting in the 1970s a number of *fuel oxygenates* became available and were increasingly used for gasoline blending.

More recently, efforts to reduce air pollution led to the enforcement of more restrictive exhaust emission limits for vehicles. In this context, passenger cars with catalytic converters were introduced. Since the addition of lead to gasoline results in the deactivation of the cat-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_037

alytic converters, unleaded gasoline of high octane quality was consequently made available.

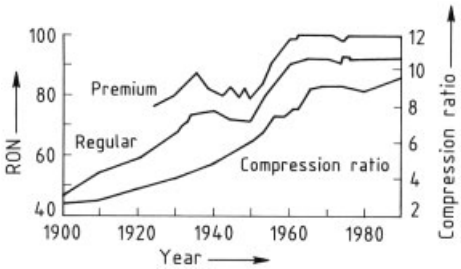

Figure 1. Trend of research octane number and compression ratio in Germany

2. Technical Information

2.1. Combustion in Otto Engines

(→ Automotive Fuels, Chap. 2.1.)

During the compression stroke the air – fuel mixture is compressed to 1.5 – 2.5 MPa and the temperature rises up to 1000 K, depending on the engine compression ratio. The mixture composition and the engine condition are controlled in such a way that the upper flammability limit of the mixture is not exceeded i.e., no self-ignition occurs. Shortly before the piston reaches the upper dead point, combustion is initiated by an electrical discharge between the electrodes of the spark plug. The appropriate spark timing is essential for engine efficiency. Best efficiency is achieved when maximum chemical energy released from the combusted charge is transformed into mechanical energy during the expansion stroke. The spark transports the necessary ignition energy to form a self-sustained flame kernel in the air – fuel mixture. Subsequently this flame kernel expands and flame fronts propagate spherically into the combustion chamber. The speed of flame propagation depends on a variety of parameters.

First of all the process is affected by the initial status of the air – fuel mixture with respect to temperature, pressure, and turbulence. Since the combustion process is exothermic, temperature and pressure increase rapidly. This accelerates the flame propagation, which in turn affects the status and turbulence of the burned gas as well as the unburned charge ahead of the flame. The entire combustion consists of a complex interaction of fluid dynamics, reaction kinetics, radical diffusion, and radiation processes.

2.2. Knock Phenomena

Average flame speeds in the engine combustion process range from 10 to 30 m/s. In contrast, flame speeds of 300 to 500 m/s are measured under knocking conditions. The knock phenomena typically occur at the end of the normal combustion process (Fig. 2). They originate from zones of unreacted charge where the critical conditions in terms of temperature and pressure are exceeded. This causes spontaneous self-ignition which in turn results in a sharp pressure increase.

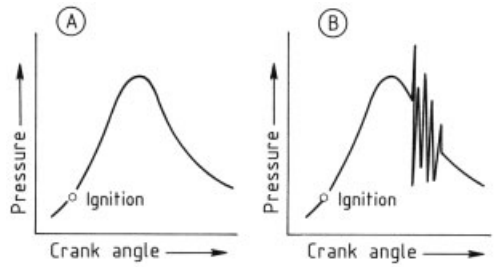

Figure 2. Pressure profile versus crank angle during combustion
A) Normal combustion; B) Knocking combustion

The detailed processes occur very rapidly and involve the complex coupling of fluid dynamic and reaction kinetic processes. Despite a large number of experimental and theoretical studies the phenomenon itself is not fully understood [1].

The phenomenological process, however, is readily detectable. It is characterized by a typical high-frequency noise emitted from engine parts. Knocking combustion results in high mechanical stress and thermal overload of the engine. Finally, it leads to severe damage to pistons, piston rings, seals, and/or cylinder heads.

2.3. Octane Number Measurement

The knock resistance of gasolines is measured by their research octane number (RON) and motor octane number (MON). Both are determined

in single cylinder engines according to standardized test procedures. The design of the engines, their mode of operation, and the calibration and measuring methods are complex; detailed descriptions are given in ASTM D 2699 (RON) and ASTM D 2700 (MON) and in equivalent ISO and EN standards (→ Automotive Fuels, Chap. 3.1.1.). Measurement of MON is performed under higher engine severity than the RON measurement in terms of intake air, temperature, and engine speed. The MON describes high speed and high load knock behavior in productional engines whereas the RON is a measure for knock performance under acceleration conditions.

3. History

Until the mid-1920s there was only one gasoline grade available in most countries, predominantly produced from straight-run crude oil products which were selected according to their composition of aliphatic and aromatic components. In order to meet the increasing octane demand, e.g., indicated by the introduction of a premium grade in 1924, several refinery processes were developed. Today, a typical refinery produces up to twenty individual product streams used for gasoline blending (→ Automotive Fuels, Chap. 4.1.). The octane enhancement of tetraethyllead (TEL) was discovered in 1922 in the United States by T. Midgley, Jr. [2]. It was produced by the Ethyl Gasoline Corporation and first introduced onto the U.S. market in 1929. In Germany it became available shortly before World War II as an octane enhancer. TEL and TML (tetramethyllead) are so-called "lead fluids" and are still used in some parts of the world. However, leadalkyl compounds have disappeared from gasoline in most countries due to environmental concerns and because lead deactivates catalytic converters.

Other octane enhancers–except fuel oxygenates–never achieved any significant degree of application. The current dominance of unleaded gasolines in the market, however, causes octane shortfalls. This situation has led to renewed attention to additional octane sources, including both new and known oxygenates and alternative octane enhancers. Therefore, octane enhancers may play a more pronounced role in the future provided that environmental aspects are satisfied, such as toxicological properties and full compatibility with modern exhaust emission control systems.

4. Types of Octane Enhancers

Octane enhancers are substances which result in octane improvements when added to gasolines. This definition also includes all hydrocarbon-based refinery streams of sufficient octane quality. These streams, however, must be regarded as conventional gasoline blendstocks and are not covered in this article, in spite of the fact that some are traded as octane enhancers e.g., aromatic refinery fractions such as toluene.

Alternative ways to improve octane quality involve the use of special chemical compounds. Since these compounds have to be readily soluble in gasolines they must be hydrocarbon-based. They may, however, contain functional groups with heteroatoms such as oxygen, nitrogen, sulfur, phosphorus, halogens, or metal atoms. Depending on the compound concentration in the fuel, components and additives can be distinguished. *Components* are compounds blended in quantities which also affect gasoline properties other than octane number, such as heating value, density, or vapor pressure. In contrast, *additives* are applied in the ppm-concentration range so that other fuel properties are essentially unaffected. This discrimination reflects the fact that additives are much more efficient than components. Additives consist exclusively of organometallic compounds. When combusted they form solid oxidation products (ash) in addition to gaseous combustion products.

5. Ashless Octane Enhancers

Components are always added in higher quantities, and must, therefore, consist of "ashless" compounds. Only oxygen- or nitrogen-containing functional groups are suitable for octane enhancement. Sulfur-, phosphorus-, or halogen-containing compounds either promote knocking or achieve only insignificant octane improvements.

5.1. Blend Octane Numbers

When a component is blended to a base gasoline, the resulting octane number of the mixture cannot be calculated from the octane numbers of the neat gasoline and the component. Furthermore, the octane number of the mixture is not proportional to the component concentration. In order to characterize the octane efficiency of a component, therefore, "blend octane numbers" (B-ON) are defined according to the following equation:

$$\mathrm{B-ON} = \frac{\mathrm{ON_M} - \mathrm{ON_{BG}}\,(1-x)}{x}$$

where B-ON is the blend octane number of the component; $\mathrm{ON_M}$ the octane number of the mixture; $\mathrm{ON_{BG}}$ the octane number of the base gasoline; and x is the volume fraction of the component in the blend. All parameters on the right-handside of this equation must be determined experimentally. This equation attributes the entire change in octane number to the component. Graphically (Fig. 3), it reflects a linear extrapolation from the base gasoline $\mathrm{ON_{BG}}$, via the mixture $\mathrm{ON_M}$, to a fictive B-ON value of the neat component. If the calculated B-ON is compared with the measured octane number of the neat component ($\mathrm{ON_C}$), two alternatives are possible. If the B-ON is larger than the measured $\mathrm{ON_C}$, the mixture has a "synergistic" property, if the B-ON is smaller than the $\mathrm{ON_C}$, the mixture shows an "antagonistic" behavior.

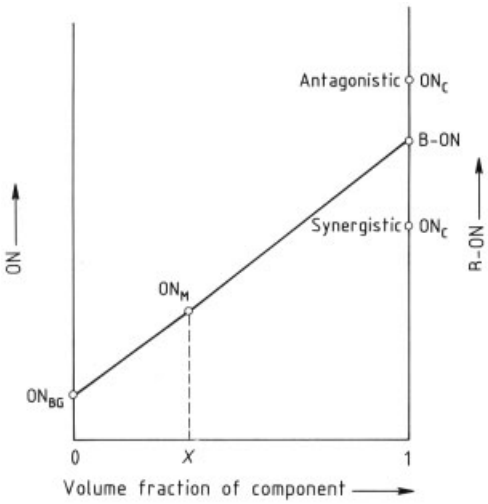

Figure 3. Graphic interpretation of blend formula

An example of the nonlinear MON blend behavior and the determination of the B-MON value is shown in Figure 4 for straight-run gasoline – MTBE mixtures. This example shows that the extrapolated B-MON value based on $x = 0.2$ is higher than the experimentally determined MON value of neat MTBE, indicating a synergistic effect of MTBE.

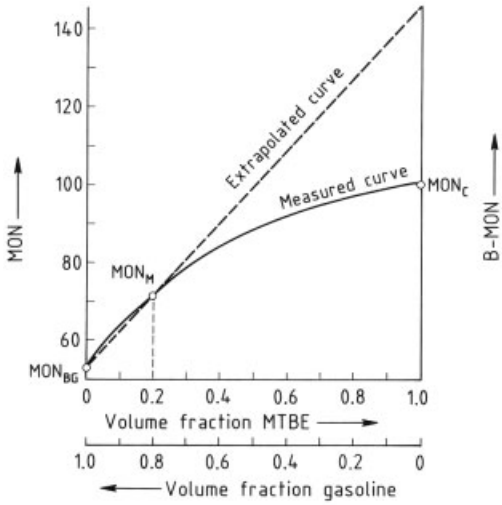

Figure 4. Example of MON blend behavior of straight-run gasoline – MTBE mixtures

In general, B-ON values of octane enhancers and refinery components depend on the component concentration, as well as the octane level and the composition of the base gasoline. These dependencies on the base gasoline are relatively small for refinery components. For some octane enhancers, however, in particular oxygenates such as alcohols, the B-ON values vary significantly with gasoline composition. Therefore, in refinery blending, different B-ON data are used for the same oxygenate when blended to different gasoline grades. Normally, B-ON data derived from paraffinic gasolines are higher than those from gasolines with high olefinic and/or aromatic contents. Most components achieve higher B-RON than B-MON values. Both depend on the component concentration as can be seen in Figure 4 for B-MON. In practice, this effect can be neglected because the concentration range used is normally small.

Some other properties such as density, volatility, and vapor pressure also exhibit non-

linear blend behavior in mixtures. Blend properties are determined analogously to B-ON and play an important role in the evaluation of the economic potential of a component.

5.2. Nitrogen-Containing Components

The influence of nitrogen-containing compounds on octane performance has been studied extensively [3]. Depending on the functional group and the molecular structure, they provide either proknock (i.e., they promote knocking) or antiknock properties.

Whereas *organic nitrates* and *aliphatic amines* were generally found to promote knocking, *aromatic amines* achieve relatively high octane enhancements. Examples of B-ON values determined for aniline and some aniline derivatives are shown in Table 1. All measurements were performed in the same base gasoline with a component addition of 2 vol %. The octane enhancement efficiency varies with molecular structure. Relative to aniline, differences are small when hydrogen atoms of the aromatic ring are substituted by methyl or ethyl groups. Substitution with longer aliphatic chains results in reduced efficiency. Whereas primary and secondary amines have good octane enhancing properties, tertiary amines e.g., *N*,*N*-dimethylaniline are ineffective.

Table 1. Blend-octane numbers of nitrogen-containing components (values determined from 2 wt % addition to premium base gasoline)

Component	B-RON	B-MON
3,4-Dimethylaniline	370	320
3,5-Dimethylaniline	340	310
p-Toluidine	340	305
p-Ethylaniline	320	300
Diphenylamine	310	300
Aniline	310	290
p-tert-Butylaniline	300	260
N-Methylaniline	280	250
Indoline	300	150
N,*N*-Dimethylaniline	95	84

A serious *disadvantage* of nitrogen-containing components results from the fact that nitrogen is partially oxidized in the combustion process. This increases nitrogen oxide emissions in the exhaust gases. There is also experimental evidence that some of the components may cause deposit formations in fuel intake manifolds and/or combustion chambers, when used at the required concentrations. Of the listed components, only *N*-methylaniline was used as an octane enhancer for a short period after World War II in Germany.

5.3. Oxygen-Containing Components

Organic oxygen compounds such as alcohols, ketones, ethers, esters, and phenols are commonly characterized as "oxygenates" or "fuel oxygenates". The majority of the oxygenates act as octane enhancers; their efficiency depends, however, on the functional group and on the individual structure of the molecule. Some derivatives of esters, ketones, and phenols achieve relatively high B-ON values, as indicated in Table 2. Because of other inappropriate properties e.g., high boiling points, high acidity, low material compatibility, and high toxicity they are not used as octane enhancers.

Table 2. Blend octane numbers of selected oxygenates

Component	B-RON	B-MON
3,4-Dimethylphenol	180	180
Phenol	190	160
m-Cresol	160	150
o-Cresol	160	140
Dimethyl succinate	130	130
Acetylacetone	120	110
Acetone	107	98

Of the oxygenates, only alcohols and ethers are used to any significant extent. They are preferred over other octane enhancers because of lower costs, availability, handling, and their physical and chemical properties, which are similar to gasoline components. Names, abbreviations, formulae, and molecular mass of the most important alcohols and ethers suitable for fuel blending are listed in Table 3.

5.3.1. EC Directive on Fuel Oxygenates

The use of oxygenates in gasolines was established during the 1970s and has developed rapidly to a large-scale business. Prior to any official regulation, the German automobile and oil industry reached an interindustrial agreement to restrict the methanol content of fuel to 3 vol% with the obligation to use an adequate cosolvent component (e.g., TBA) to ensure appropiate water tolerance of methanol-containing fuels. For

Table 3. Important fuel oxygenates

Oxygenate	Abbreviation	Formula	M_r	CAS registry no.
Alcohols				
Methanol	MeOH	CH_3OH	32.04	[*67-56-1*]
Ethanol	EtOH	C_2H_5OH	46.07	[*64-17-5*]
1-Propanol	PrOH	C_3H_7OH	60.97	[*71-23-8*]
Isopropanol (2-propanol)	IPA	C_3H_7OH	60.97	[*67-63-0*]
1-Butanol	NBA	C_4H_9OH	74.12	[*71-36-3*]
sec-Butanol (2-butanol)	SBA	$C_2H_5CH(OH)CH_3$	74.12	[*78-92-2*]
Isobutanol (2-methyl-1-propanol)	IBA	$(CH_3)_2CHCH_2OH$	74.12	[*78-83-1*]
tert-Butanol (2-methyl-2-propanol)	TBA	$(CH_3)_3COH$	74.12	[*75-65-0*]
Ethers				
Dimethyl ether (oxybismethane)	DME	CH_3OCH_3	46.07	[*115-10-6*]
Methyl *tert*-butyl ether (2-methoxy-2-methyl-propane)	MTBE	$CH_3OC(CH_3)_3$	88.15	[*1634-04-4*]
Ethyl *tert*-butyl ether (2-ethoxy-2-methylpropane)	ETBE	$C_2H_5OC(CH_3)_3$	102.18	[*637-92-3*]
Diisopropyl ether (2,2′-oxybispropane)	DIPE	$(CH_3)_2CHOCH(CH_3)_2$	102.18	[*108-20-3*]
tert-Amyl methyl ether (2-methoxy-2-methyl-butane)	TAME	$CH_3OC(C_2H_5)(CH_3)_2$	102.18	[*994-05-8*]
Isopropyl *tert*-butyl ether [2-methyl-2-(1-methylethoxypropane)]	PTBE	$(CH_3)_2CHOC(CH_3)_3$	116.28	[*17348-59-3*]
Di-*sec*-butyl ether (2,2′-oxybisbutane)	DSBE	$C_2H_5CH(CH_3)OCH(CH_3)C_2H_5$	130.23	[*6863-58-7*]
sec-Butyl-*tert*-butyl ether (2-(1,1-dimethyl-ethoxybutane)	BTBE	$C_2H_5CH(CH_3)OC(CH_3)_3$	130.23	[*32970-45-9*]

the EC the European Council passed a directive in December 1985, regulating the use of oxygenates in gasolines after several years of intensive negotiations of the ministries of commerce supported by experts [4]. Meanwhile, the technical annex of the directive has been implemented in the gasoline standards of all EC member states.

The purpose of the directive was to ensure:

1) The performance of the existing car population
2) The use of gasoline extender and octane-enhancing components in addition to materials derived from crude oil
3) The compensation of octane shortfall due to lowering of alkyllead concentration
4) A harmonization of gasoline quality in Europe

Table 4. EC directive on fuel oxygenates

Oxygenate	A, ∗ vol%	B, ∗∗ vol%
Methanol (stabilizers to be added)	3	3
Ethanol (stabilizers may be added)	5	5
Isopropanol (IPA)	5	10
tert-Butanol (TBA)	7	7
Isobutanol (IBA)	7	10
Ethers, with five or more C atoms	10	15
Other oxygenates, defined	7	10
Total oxygen content, wt %	2.5	3.7

∗ A = limits which have to be accepted by all EC countries.
∗∗ B = limits which may be adopted by individual member states but above which labeling is mandatory.

The annex lays down definitions and blend compositions of oxygenates. The most important information is summarized in Table 4, providing the maximum volumetric additions of individual oxygenates and the maximum gravimetric oxygen content of the blends.

5.3.2. Blend Octane Numbers of Oxygenates

The octane efficiency of oxygenates cannot be characterized in a simple way because it depends on a variety of factors. Therefore, B-ON ranges are usually defined to account for their possible variation. This makes it difficult to compare octane efficiencies of different oxygenates.

To overcome this problem, Veba Oil conducted a series of investigations with the objective to characterize the octane blend behavior of some oxygenates. In the progress of this work, the gasoline sensitivity (SEN, i.e., the difference between RON and MON) was found to be a suitable correlation parameter [5]. Empirically it may be considered as an approximate gauge for the gasoline composition; a high paraffinic content corresponds to low SEN, high aromatic and/or olefinic contents result in high SEN.

The measurements were performed in a variety of different unleaded gasolines, the compositions of which were widely varied. Oxygenate concentrations ranged from 10 to 20 vol%. Examples of the graphic correlations based on linear regression of B-ON data versus gasoline

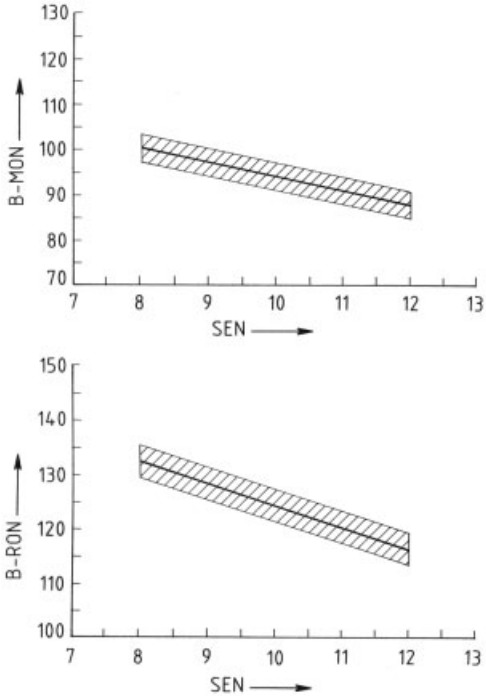

Figure 5. Correlation between B-ON and SEN for methanol

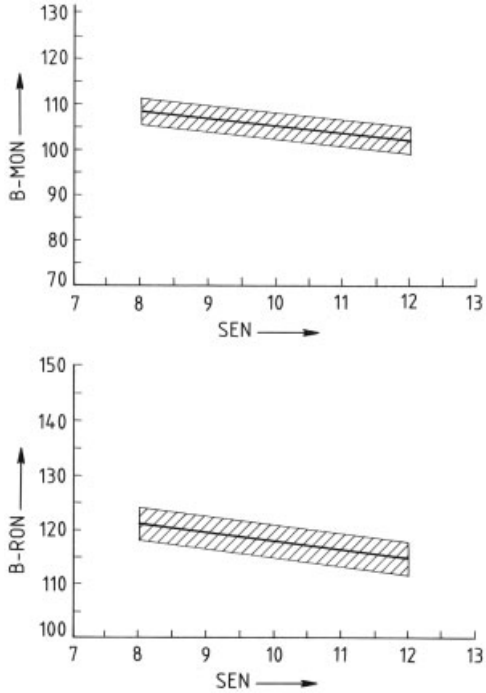

Figure 6. Correlation between B-ON and SEN for PTBE

SEN are shown in Figure 5 for methanol and in Figure 6 for PTBE. Both B-RON and B-MON decrease with increasing gasoline SEN. Similar blend behavior was found for the majority of the oxygenates investigated. Typically, lower alcohols such as methanol and ethanol show a strong dependence of B-ON on SEN. For higher alcohols and ethers this dependence is less pronounced; in a few cases, e.g., for TBA, no significant dependence was observed.

5.3.3. Other Properties of Oxygenates

Beside their octane efficiency, there are other properties of oxygenates which also affect the gasoline quality. Again their blend behavior may not be linear so that blend properties must be determined analoguously to the B-ON procedure. Selected properties of alcohols and ethers used as oxygenates are listed in Table 5.

Table 5. Properties of oxygenates

Oxygenate	*mp*, °C	*bp*, °C	d_4^{20}	n^{20}	Heat of combustion, kJ/mol	RVP ∗	Composition, wt %		
							C	H	O
MeOH	− 97.7	64.70	0.7912	1.3284	− 725.7	0.81	37.48	12.58	49.93
EtOH	−114.1	78.29	0.7894	1.3614	−1367.6	0.70	52.13	13.13	34.73
IPA	− 87.8	82.26	0.7755	1.3772	−2006.9	0.72	59.95	13.42	26.63
SBA	−114.7	100.00	0.8069	1.3972	−2660.6		64.81	13.60	21.59
IBA	− 88.5	107.66	0.8016	1.3958	−2668.5	0.63	84.81	13.60	21.59
TBA	25.6	82.80	0.7866	1.3877	−2644.0	0.64	64.81	13.60	21.59
DME	−141.5	−24.80	0.6689	1.3018	−1460.5		52.14	13.13	34.73
MTBE	−108.6	55.20	0.7406	1.3690	−3362.1	0.65	68.13	13.72	18.15
ETBE	− 94.0	73.10	0.7399	1.3756	−4014.9	0.29	70.53	13.81	15.66
DIPE	− 85.5	68.30	0.7235	1.3681	−4042.6	0.33	70.53	13.81	15.66
TAME		86.30	0.7700	1.3885	−4010.8	0.77	70.53	13.81	15.66
PTBE		87.50	0.7400	1.3791	−4667.2	0.20	72.35	13.88	13.77

∗ RVP Reid vapor pressure ASTM D 323.

Density. Densities of alcohols are generally higher than the specified density region for gasolines, whereas those of ethers range within it. Blend densities of alcohols and ethers show only moderate differences when compared to the densities of the neat components.

Vapor Pressure. Due to their polarity, lower alcohols such as methanol and ethanol possess blend vapor pressures (B-VP) which exceed the vapor pressure of the neat components, because association of alcohol molecules is hindered in the mixture. Therefore, their B-VP values strongly depend on the component concentration. In contrast, B-VP data of ether components hardly show any dependence on concentration. They are significantly lower than those of lower alcohols and are similar to the vapor pressure of the neat components.

Volatility. The addition of oxygenates affects the boiling characteristics of the gasolines. When ethers are blended, only moderate changes in the distillation curves are observed. Blends of lower alcohols, however, cause significant deformations of the distillation curves due to the formation of azeotropes. This rises the front-end volatility, i.e., an increased distillate fraction in the temperature range of the alcohol boiling points is obtained.

5.3.4. Production

The important processing data of alcohols and ethers are summarized in [6]. Further details of production and economics are described in [7].

Alcohols.

Methanol is produced catalytically from synthesis gas (→ Methanol, Chap. 4.). Apart from its use as a gasoline blending component, large amounts of methanol are also employed as feedstock for the etherification with isobutene to form MTBE (→ Methyl Tert-Butyl Ether, Chap. 3.).

Synthetic ethanol is produced by the hydration of ethylene with an acid catalyst such as sulfuric acid and phosphoric acid in reactor columns or in ion-exchanger fixed-bed reactors (→ Ethanol, Chap. 4.1.→ Ethanol, Chap. 4.2.). *Bioethanol* (→ Ethanol, Chap. 5.) is produced by fermentation of monosaccharides such as glucose, fructose, mannose, or xylose derived from molasses, sugarcane, sugar, corn, or potatoes. Costs of bioethanol are higher than those of synthetic alcohol because the process economics depend essentially on the feedstock price.

Higher alcohols i.e., 2-propanol, 1-butanol, 2-butanol, and 2-methyl-2-propanol can also be produced by hydration with acid ion-exchanger resins [8]. *Tert*-butanol (TBA) is formed as a coproduct in the production of propylene oxide by epoxidation of propene with *tert*-butyl hydroperoxide (TBHP) (→ Propylene Oxide). In the case of high MTBE demand in the market, TBA is sometimes dehydrated to isobutene which is then fed to the MTBE process [9].

Ethers. The ethers are chiefly manufactured by exothermic reaction of branched olefins bearing the double bond at the branching point with alcohols such as methanol, ethanol, 2-propanol, or butanol according to:

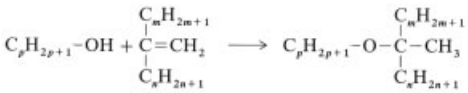

[for MTBE $m = n = p = 1$; for PTBE $m = n = 1$, $p = 3$; for *tert*-amyl methyl ether (TAME) $m = 2$, $n = p = 1$].

MTBE is the most important ether used as an oxygenate. MTBE capacity (see Fig. 7) and plant construction have increased steadily because MTBE blending offers simple and economic flexibility to fill any octane gap in the refiner's gasoline pool [10]. More than 40 % of the total onstream capacity of MTBE is used as merchant MTBE, which is available to supply refiners in need of additional octane quality [11]. For production of MTBE, see → Methyl Tert-Butyl Ether, Chap. 4.

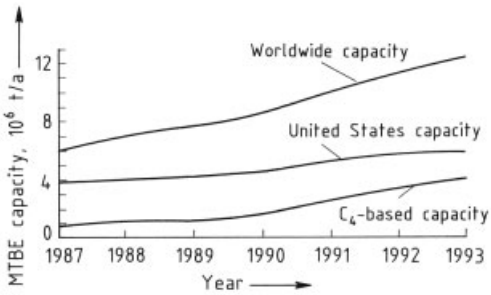

Figure 7. Worldwide MTBE capacity trends (1989 survey)

Reaction of ethanol with isobutene produces ethyl *tert*-butyl ether (ETBE), which has similar properties to MTBE. This process offers an outlet of bioethanol [12, 13]. Reaction of 2-propanol with isobutene yields *PTBE* [14, 15], an oxygenate with excellent blend properties such as low vapor pressure and low oxygen content in combination with high octane quality.

Based on a similar technology methylbutenes (amylenes) of C_4 to C_6 streams can be etherified with methanol to *TAME* [16, 17], a blending ether with a high octane number. The BP Etherol process [18] uses unsaturated C_5 compounds of cracker fractions. Branched olefins react with methanol to form higher ethers. However, most of the branched olefins have terminal double bonds with low reactivity. To increase the yield of 3-methyl-2-butene (isoamylene) the double bond is shifted in the presence of a palladiumimpregnated cation-exchanger resin. The isomerization proceeds according to:

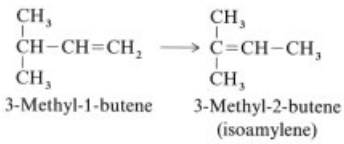

Isoamylene can then be reacted with methanol to form TAME.

5.3.5. Environmental Protection

The compulsory use of oxygenates in gasolines is becoming widespread in the *United States*, because several investigations have shown that oxygenates reduce carbon monoxide emissions. This effect offers a means of bringing CO emissions down to required standards. Depending on the area and season in the United States, the minimum oxygen content in the fuel is in the range of 2.0 – 3.0 wt % oxygen (Clean Air Act Amendment) [19]. Addition of TBA and methanol may reduce the CO emissions up to 40 % and the hydrocarbon emissions up to 20 % in comparison with conventional gasoline [20]. Insufficient supply of oxygen in air – fuel mixtures causes incomplete combustion of the fuel, accompanied by CO formation. When oxygen is supplied by oxygenates that are blended to the gasoline, this chemically available oxygen reduces the formation of carbon monoxide. Governments of several states made efforts to establish the relationship between fuel oxygen content and CO exhaust concentrations. In 1988, the state of Colorado mandated a minimum of 1.5 wt % oxygen in gasoline from January to March; this program was continued from November 1988 until March 1989, this time a minimum of 2.0 wt % oxygen was required. The 1988 program was estimated to lower CO air concentrations by 8 – 10 % along the Front range of the Rocky Mountains [10]. In the United States emission measurements with different vehicles show CO reduction up to 32 %, during the Federal Test Procedure (FTP) cycle (a standardized driving cycle for emission tests) [21]. The tests included vehicles with different exhaust and closed-loop systems and demonstrated a statistically significant CO reduction due to the oxygenated fuels, depending on the system and oxygen content (Fig. 8). The gasolines were blended with MTBE and ethanol. Emissions of unburned hydrocarbons and nitrogen oxides were virtually unaffected.

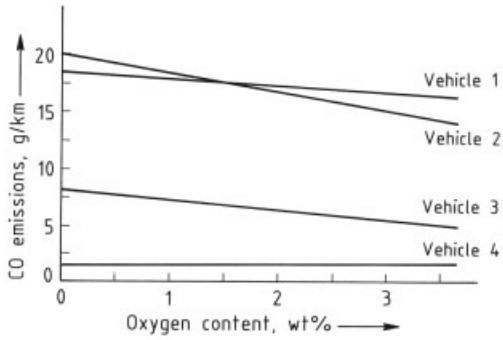

Figure 8. CO emissions FTP 75

Under *European conditions*, studies were made with ethers and alcohols on the influence of oxygen content on the limited pollutants such as carbon monoxide, hydrocarbons, and nitrogen oxide [5]. Blend components were methanol with IPA – TBA as cosolvents and the ethers MTBE, PTBE, and BTBE. Another investigation was conducted by the European Fuel Oxygenates Association (EFOA) using mixtures of ethers and alcohols to study the dependence on oxygenates systematically [22]. Oxygen content was varied up to 3.6 wt %. The results showed that oxygenates reduce CO emissions to a large

degree. The reduction increases with oxygen content but is not a function of the type of oxygenate.

5.3.6. Toxicology and Occupational Health

MTBE. Similar investigations of toxicity and other potential health effects of MTBE were performed. Animal testing to date has indicated that after ingestion, inhalation, or even injection of MTBE, it is mostly excreted unchanged [4].

PTBE. The Ames test does not show any mutagenity. According to an acute toxicity study in rats, PTBE is considered to be nontoxic up to a dose of 2000 mg per kilogram body weight.

5.3.7. Storage and Transportation

The higher ethers MTBE, PTBE, and TAME are stable in the presence of air under atmospheric conditions. Formation of peroxides is not observed during contact with oxygen. Containers, sealing materials, and also elastomers in automobiles have to meet the same requirements as for contact with gasoline.

6. Metal-Containing Additives

A large number of organometallic compounds are known to act as octane enhancers. Generally, compounds containing Pb, Cu, Ni, Th, Fe, and Mn achieve high octane improvements, whereas those containing Co, Zn, Bi, V, Se, and Ce are slightly less effective. For practical use, most of them must be excluded because of high costs, instability, toxicity, impact on catalytic converters, or other undesired properties.

Their use in gasolines also requires an official approval in most European countries and in the United States. In Germany, the application of metal-containing additives–other than lead–is restricted by the gasoline lead act (Benzin – Blei – Gesetz).

The octane enhancement of organometallic additives is explained by a catalytic effect, because during combustion the metal compound is oxidized to form very fine solid particles of metal oxide. These particles have large surface areas with active centers which can react with free radicals formed by partial oxidation of the gasoline. Therefore the spontaneous combustion is delayed and knocking does not occur. The consumed centers are reactivated by oxygen and can react again with free radicals. Due to this cycle only small concentrations of additive are necessary to cause high octane enhancement.

A metal-free component is unable to delay the spontaneous combustion catalytically. The organic compounds are consumed irreversibly by the intermediate reactants which are responsible for knocking. Therefore the effectivity of antiknocking components is limited and high concentrations are needed to get a higher octane quality.

A disadvantage of metal-containing additives results from the formation of solid combustion products that lead to deposit accumulation in the combustion chambers. For the lead fluids this problem was solved by addition of lead scavengers. However, this is unfeasible for most other organometallic additives, thus their use must be restricted to very low concentrations. Additives that achieve sufficient octane improvements in the ppm concentration range and that meet other technical and economical requirements are methylcyclopentadienylmanganese tricarbonyl (MMT) and dicyclopentadienyliron (ferrocene).

6.1. Alkyllead Compounds

The strong octane-boosting effect of tetraethyllead (TEL) was discovered in 1921 in the United States [23]. Its wide use began in the United States in 1929 and in Germany shortly before World War II. Curiously, investigations at that time regarded lead as relatively environmentally friendly.

The effect of lead is based on an intervention in the mechanism of the combustion process. In precombustion reactions, organic peroxides are formed, which decompose to form free radicals. The latter initiate knocking reactions. These highly reactive radicals can be deactivated by very finely divided lead compounds formed from the alkyllead compounds in the combustion chamber [24, 25].

In addition to tetraethyllead, which was used exclusively at first, tetramethyllead (TML)

came into use from 1960 onward. Mixed alkyllead compounds, such as diethyldimethyllead or ethyltrimethyllead, and mixtures of TEL and TML were also employed, because a broad-spectrum effect can be achieved through a combination of different characteristics (Table 6). Thus, the comparatively low boiling point of TML is good for increasing the front octane number. (Fig. 9).

Table 6. Comparison of TEL and TML

Chemical and physical properties	Tetraethyllead	Tetramethyllead
	$(C_2H_5)_4Pb$	$(CH_3)_4Pb$
Molecular mass, g/mol	323.5	267.3
Lead content, %	323.5	267.3
Density, g/cm^3	1.66	1.995
bp, °C	200 (decomp.)	110
Vapor pressure (20 °C), Pa	34	31.5×10^2

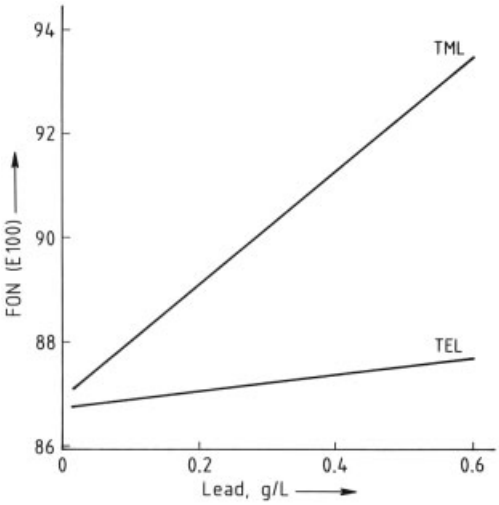

Figure 9. Effect of tetraethyllead and tetramethyllead on the gasoline front octane number (FON)

The increase in octane number that can be achieved with alkyllead compounds depends strongly on fuel composition. The *lead susceptibility,* a measure of the response behavior to the addition of lead, decreases in the following order: paraffins >> cycloparaffins > aromatics > olefins > diolefins (≈ 0). Accordingly, fuel components rich in paraffins, such as straight-run gasoline, alkylate, and isomerate (see → Automotive Fuels, Chap. 4.), react very well to lead addition, whereas olefin-rich catalytically cracked gasoline has a poor response (Fig. 10).

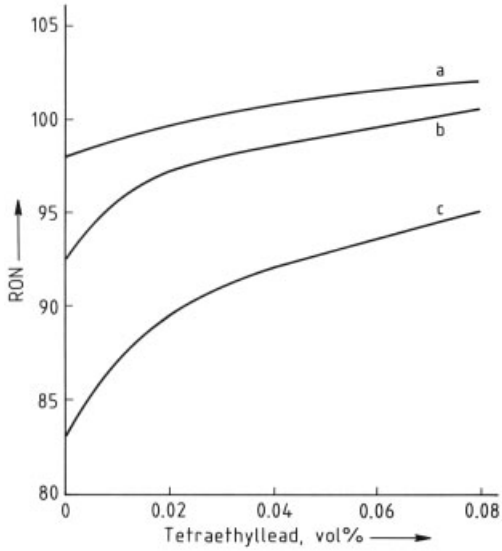

Figure 10. Effect of tetraethyllead on some base fuels (octane quality of the base fuel types are widely varying in practice)a) Olefin-rich fuel; b) Aromatic-rich fuel; c) Paraffin-rich fuel

In connection with the addition of alkyllead compounds to gasoline, the term "lead bonus" was sometimes used. It meant that highly leaded fuels, in comparison with unleaded or low-lead ones, have a disproportionately high road octane number, although the RON and MON determined in test engines are equally high [26–30]. Investigations carried out with the introduction of unleaded fuels showed, however, that low-lead and unleaded fuels behave almost identically with regard to the road octane number [31] in most cases.

Usually, alkyllead compounds have always been used in combination with scavengers. These are organochlorine and -bromine compounds, namely, 1,2-dichloroethane and 1,2-dibromoethane. In the combustion chamber, scavengers cause the formation of lead chloride and lead bromide whose good volatility at high temperature ensures that ca. 70 – 80 % of the lead is not deposited there. However, investigations in 1989 have shown that scavengers can be removed without adverse affects when the lead concentration in the fuel is low (e.g., 0.15 g/L).

On the other hand, treatment of leaded fuels with an antioxidant is necessary. Tetraethyllead, in particular, is prone to oxidation over long storage periods, leading to the separation of solid residues that can lead to blockages in the fuel system.

Originally, the addition of lead to gasoline was considered to be safe to the environment. This opinion has steadily changed, however [32–34], so that a decrease in the maximum permitted lead content has been legislated virtually worldwide. In the United States, Japan, Canada, Australia, Europe, and other areas leaded fuels have practically disappeared from the market. From a toxicological point of view, however, not only the lead but also the scavengers connected with it are suspected to be health risks, because these can form traces of highly toxic halogenated dibenzodioxins and -furans in the combustion chamber [35]. However, not only the poor compatibility of leaded fuels with the environment but also their incompatibility with catalytic converter technology (catalyst poisoning, see → Automobile Exhaust Control) has driven the move to unleaded gasoline in many parts of the world.

Alkyllead compounds protected also unhardened valve seats, present in older engines. Here, lead compounds formed in the combustion chamber inhibited erosion and wear ("recession") of the valve seats in two ways: (1) by avoiding high-temperature corrosion and (2) by their effect as solid lubricants. However, nontoxic alkali-metal additives are now available for this purpose, e.g., Powershield (Lubrizo) Formula Shell (Shell) Valve Master (Du Pont). Practically the only suppliers of alkyllead compounds are Ethyl and Octel.

6.2. Methylcyclopentadienylmanganese tricarbonyl

Methylcyclopentadienylmanganese tricarbonyl (MMT) [*12108-13-3*], is an organometallic additive used as an octane enhancer at low concentration. MMT is a metallocene compound in which the delocalized electrons of the cyclopentadienyl group are partially donated to the manganese atom and form covalent bonds. Carbon monoxide has a lone electron pair on the carbon atom which is donated to the manganese atom to form the carbonyl bond. In Canada it has been blended to unleaded gasoline for many years at a concentration of 0.0165 g Mn per liter of gasoline. In the United States, MMT is used in combination with alkylleads in leaded gasoline. Ethyl Corporation (Orangeburg) is the U.S. producer of MMT.

History. Ethyl Corporation discovered MMT in 1954 and introduced it into the U.S. market as a supplement to TEL in 1958. Evaluation in unleaded gasoline was made in 1969. From 1974 to 1979 MMT was used in unleaded gasoline.

The effect of MMT was tested on a series of in-use three-way catalysts. Catalyst activity was measured and compared to the activity of the catalyst without MMT operation. Steady-state and light-off temperature measurements indicated a reduction in catalyst efficiency as exposure to MMT was increased [37].

The EPA predicted that over the useful life of vehicles, a greater number of MMT-fueled vehicles would not meet their designed hydrocarbon emission standard compared with corresponding vehicles fueled without MMT. Based on these results and in accordance with Clean Air Act Amendments, the EPA banned MMT in 1979 for unleaded gasoline [38].

In 1983 production of MMT was ca. 1800 t/a.

Trade names: AK-33 X, Antiknock – 33, Combustion improver – 2.

Physical Properties [39]. MMT is sparingly soluble in water (70 ppm at 25 °C) and soluble in hydrocarbon solvents. MMT is a dark orange (straw color) liquid with a herbaceous odor; *mp* − 2.2 °C, *bp* 232.8 °C.

Chemical Properties. When heated to decomposition MMT emits toxic fumes of CO.

Production. The methylcyclopentadiene dimer [*26472-00-4*] is added to molten sodium suspended in diethylene glycol dimethyl ether at 185 to 190 °C. At this temperature, the dimer splits into the monomer and reacts with sodium to form sodium methylcyclopentadienide:

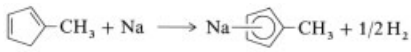

Anhydrous, flaked manganese chloride is added to the reaction mixture. The reaction forms bis(methylcyclopentadienyl)manganese with sodium chloride as byproduct. In the next step carbonylation under a pressure of 4.3 to 4.5 MPa at 190 °C forms MMT. Pure MMT is subsequently separated by fractional distillation.

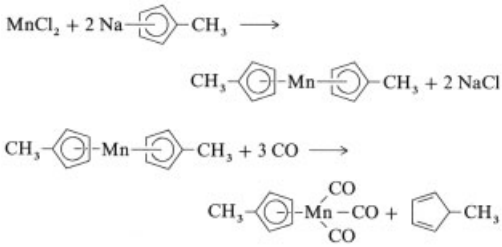

Octane Enhancement. MMT can enhance between 0.5 and 1.0 (RON + MON)/2 per 0.01 g Mn/L, depending on gasoline composition, octane level, and leaded or unleaded basestock.

Toxicology and Occupational Health. MMT is a poison by ingestion, inhalation, skin contact, intravenous, and intraperitoneal routes (Table 7). MMT is included in the EPA's Extremely Harzardous Substances List. Manganese and some compounds are on the Community Right To Know List.

Table 7. Toxicity data of MMT ∗ [40]

Test	Species					
	Rat	Mouse	Dog	Rabbit	Guinea pig	Cat
Oral						
LD_{50}, mg/kg	50	230				
LDLo, mg/kg			620	95	900	
Skin						
LD_{50}, mg/kg				140		
Inhalation						
LD_{50}, mg/m^3 (4 h)	76	58.6				
LCLo, mg/m^3			489 (2 h)	1237 (1 h)	1237 (1 h)	223 (7 h)
Intraperitoneal						
LD_{50}, mg/kg	23	152				
Intravenous						
LDLo, mg/kg			10			
LD_{50}, mg/kg				6600		

∗ American Conference of Governmental Industrial Hygienists–Threshold Limit Values: 0.1 mg Mn/m^3 (skin) time weighted average.

6.3. Ferrocene

In the early 1930s iron pentacarbonyl was used commercially as an octane-enhancing additive in Europe in concentrations similar to those of lead at that time. During combustion in engines iron pentacarbonyl formed excessive amounts of iron oxides which caused engine wear and influenced spark plug operation.

Ferrocene is another iron compound with similar octane-enhancing properties. Ferrocene [*102-54-5*] or dicyclopentadienyliron, $Fe(C_5H_5)_2$, belongs to the metallocenes; the iron atom is centered between two parallel cyclopentadiene rings. Ferrocene was first synthesized by MILLER, TEBOTH and TREMAINE in 1952. The sandwich or doublecone structure was first elucidated by G. WILKINSON and E. O. FISCHER.

Physical Properties. Important physical data of ferrocene are listed below [41]:

M_r	186.04
Density	1.49 g/cm^3
mp	173 °C
bp (at 102.3 kPa)	249 °C
Triple point	183 °C
Vapor pressure	
40 °C	0.0066 kPa
100 °C	0.34 kPa
Heat of fusion (at 175 °C)	17.8 kJ/mol
Heat of vaporization (at 175 °C)	47.2 kJ/mol
Heat of sublimation (at 25 °C)	70.2 kJ/mol
Start of decomposition	465 °C

Ferrocene is readily soluble in all hydrocarbon and hydrocarbon – oxygenate fuel blends and does not alter specified gasoline properties except for octane numbers.

Chemical Properties. Ferrocene is resistant to water, air, light, and heat.

Production. The modern electrochemical synthesis of ferrocene consists of two stages:

1) Electrochemical step (NaBr added as conductivity enhancer)

$$Fe + 2\,C_2H_5OH \rightarrow Fe(OC_2H_5)_2 + H_2$$

2) Chemical conversion

$$Fe(OC_2H_5)_2 + 2\,C_5H_6 \rightarrow Fe(C_5H_5)_2 + 2\,C_2H_5OH$$

Total:

$$Fe + 2\,C_5H_6 \rightarrow Fe\,(C_5H_5)_2 + H_2$$

In the *first stage*, metallic iron is oxidized anodically in ethanol to iron(II)ethoxide. At the cathode, the acidic protons of ethanol are reduced and form gaseous hydrogen which evolves out of the liquid. The iron(II)ethoxide is sparingly soluble and precipitates on the rotating cathode. A fixed scraper separates the iron(II)-ethoxide from the cathode surface and suspends it into the liquid. In the *second stage* the ethoxide reacts with cyclopentadiene to form ferrocene and ethanol, which is recycled to the first stage. Ferrocene is filtered from the solution and purified by sublimation.

Octane Enhancement. The octane improvements depend–as in the case of other octane enhancers–on the ferrocene concentration and on the composition of the base gasoline. Figure 11 represents a compilation of octane measurements performed in about 50 different base gasolines with ferrocene concentrations up to 30 ppm [42]. The octane response of RON is higher than that of MON.

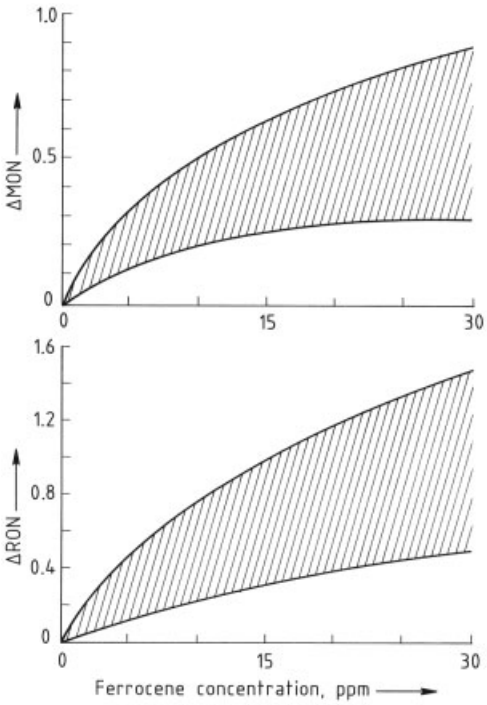

Figure 11. Ferrocene octane enhancement in unleaded gasolines

The composition of the different unleaded fuel grades tested was widely varied. The fuel blends included all major refinery components as well as available oxygenates within their maximum volumetric addition presently specified in Europe. Generally it was found that the lower the octane number of the base gasoline, the higher the octane gain. Several measurements of octane response versus concentration have shown that response is more pronounced in the low ferrocene concentration region and that levelling-off effects are obtained at higher concentrations.

Toxicology. The toxicological data of ferrocene are listed in Table 8.

Table 8. Toxicity data of ferrocene [40]

Test	Species	
	Rat	Mouse
Oral		
LD_{50}, mg/kg	1320	832
Intraperitoneal		
LD_{50}, mg/kg	500	335
Intravenous		
LD_{50}, mg/kg		178

7. References

1. *CEC Symp. Knocking of Combustion Engine,* Wolfsburg, Germany Nov. 1981.
2. T. Midgley Jr., *Ind. Eng. Chem.* **31** (1939) no. 5, 504 – 506.
3. M. H. Graham, L. Hess: Organic Nitrogen Compounds as Antiknock Additives Report IR 52 – 12 Ethyl-Corporation, June 1952.
4. M. Allion, *EFOA 4th Conf.,* Brussels 1990.
5. K. Gottlieb, H.-K. Schädlich, P. Schug: "Erzeugung von oktanzahlsteigernden höheren Ethern und ihre anwendungstechnische Erprobung unter praxisnahen Bedingungen", *BMFT-Forschungsbericht 85 034* (1989).
6. L. F. Hatch, S. Matar: *From Hydrocarbons to Petrochemicals,* Gulf Publ. Co., Houston 1981.
7. SRI reports: P. L. Morse, Rep. no. 53, "Synthetic Ethanol and Isopropanol", (1969); H. W. Scheeline, S. Terada, S. L. Soder, Rep. no. 131, "Methyl-Tertiary-Butyl-Ether and Tertiary-Butyl-Alcohol", (1974); P. L. Morse, Rep. no. 111, "Methanol as a Chemical Raw Material", (1979), Stanford Research Institute, Menlo Park, California.

8. J. Schulze, M. Homann: *C 4-hydrocarbons and Derivatives,* Springer Verlag, Berlin, 1989.
9. H. E. Eilerts de Haan: "MTBE and Methanol Blends in Europe",*EFOA 3rd Conf.,* ARCO Chemical Europe, Frankfurt 1988.
10. G. H. Unzelmann, *Oil Gas J.,* **87** (1989) Apr. 17, 44.
11. Oxygenated Fuels Association, Inc., Feb. 10, 1989, Communication to EFOA.
12. H. Bruderreck, BML-Report FE 83 NR 003 (1985).
13. D. Schliephake (ed.): *Nachwachsende Rohstoffe,* Verlag J. Kordt, Bochum 1986.
14. VEBA OEL, EP 006 315, 1982 (K. Gottlieb, H. Bruderreck, F. H. Wehmeier).
15. T. Kaitale, E. Merikallio, E. Rautianen, *EFOA 2nd Conf.,* Rome 1987.
16. W. J. Piel: *National Conference on Octane & Oxygenates,* San Francisco (1989).
17. *Oil Gas J.* **86** (1988) Nov. 21, 41.
18. H. Wittcoff, *CHEMTECH* 1990, Jan., 48.
19. EFOA, Newsletter no. 4, July 1989.
20. ARCO, *Report on Methanol,* 1983.
21. J. A. Gething, J. S. Welstand, J. C. Horn, SAE-paper 890 216 (1989).
22. H. Guttmann, K. P. Schug: "Results of Exhaust Emission Measurements with Oxygenates Motor Gasolines", SAE-paper no. 900 274, 1990.
23. D. Blackmore, A. Thomas (eds.): *Fuel Economy of the Gasoline Engine,* Macmillan, New York 1977, pp. 79.
24. R. Sharma, J. Bardwell, *Combust. Flame* **9** (1965) 106 – 108.
25. K. Salooja, *Combust. Flame* **9** (1965) 211 – 217.
26. A. Bell, J. Keene, K. Reders, SAE-Paper no. 710 624, 1971.
27. D. Gray, A. Azhari, SAE-Paper no. 720 084, 1972.
28. H. Bigley, J. Benson, SAE-Paper no. 730 013, 1973.
29. N. Mariotti, A. Ramsden, Int. Automob. Tech. Congr., Pap. B 2-1, 1974.
30. D. Hornbeck, R. Labruyere, L. Stinson, G. Unzelmann, SAE-Paper no. 750 936, 1975.
31. K. Reders, W. Dabelstein, A. Reglitzky, W. Lange, *Mineralöltechnik* (1989) no. 13, 38.
32. H. Schellhaus, W. Obländer, K. Neuschwander, *Schriften. Ver. Wasser- Boden- Lufthyg.* **42** (1974) 245 – 258
33. Conference Report on International Symposium, Airborne Lead in Perspective, National Academy of Sciences, Conference Report, Washington, D.C. 1972.
34. EEC Commission: Health Aspects of the Environmental Pollution with Lead, Amsterdam 1972/1973.
35. E. Bessey, Report on the BMFT (Bundesministerium für Forschung und Technologie)– FVV (Forschungsvereinigung Verbrennungskraftmaschinen)– DGMK (Deutsche Gesellschaft für Mineralöl und Kohle) Gemeinschaftsprojekt ″Emission halogenierter KW,″ DGMK-Conference Report 8901, 1989, pp. 85 – 119.
36. G. Schoonveld, R. Riley, S. Thomas, S. Schiff, *Autom. Eng.* **95** (1987) no. 1, 72 – 76.
37. R. G. Williams, W. L. H. Watkins, R. G. Griffins, SAE-paper no. 890 582 (1989).
38. Environmental Protection Agency; Regulation of Fuel and Fuel Additives, *MMT Suspension of Enforcement,* Ann Arbor 1979.
39. R. K. Hinderer, *Am. Ind. Hyg. Assoc. J.* **40** (1979) 164.
40. N. I. Sax, R. J. Lewis: *Dangerous properties of Industrial Material,* 7th ed., van Nostrand – Reinhold, New York 1988.
41. Chemische Betriebe Pluto, *Ferrocene-Product-Information,* Herne, Germany 1989.
42. K. P. Schug, H.-J. Guttmann, A.-W. Preuss, K. Schädlich, SAE-paper no. 900 154 (1990).

Octanol → Alcohols, Aliphatic
Octanol → 2-Ethylhexanol
Odorants → Flavors and Fragrances

Oil Refining

Oil and Gas is a separate keyword.

WALTHER W. IRION, Deutsche Shell AG, Hamburg, Federal Republic of Germany

OTTO S. NEUWIRTH, UK Wesseling, Wesseling, Federal Republic of Germany

1. Introduction 2
1.1. History 2
1.2. General Aspects of Oil Refining . 3
2. Crude Oils and Products 5
2.1. Crude Oil Constituents 5
2.2. Classification of Crude Oils 6
2.3. Oil Products 7
2.3.1. Gas Fuels 7
2.3.2. Liquid Fuels 8
2.3.3. Nonfuel Applications 9
3. Oil Refining Processes 9
3.1. Crude Oil Distillation 9
3.1.1. Atmospheric Distillation 10
3.1.2. Vacuum Distillation 12
3.1.3. Crude Oil Desalting 12
3.2. Catalytic Cracking (Cat Cracking) 13
3.3. Hydrotreating 16
3.3.1. Distillate Hydrotreating 17
3.3.2. Pyrolysis Gasoline Hydrotreating . 18
3.4. Catalytic Reforming 19
3.4.1. Introduction 19
3.4.2. Semiregenerative Reformer 21
3.4.3. Fully Regenerative Reformer 22
3.4.4. Continuously Regenerative Reformer 22
3.5. Hydrocracking 23
3.6. Residue Conversion Processes . . 23
3.6.1. Introduction 23
3.6.2. Fundamentals of Residue Conversion and Process Options 24
3.6.3. Hydrogen Addition ("H-in") Processes 26
3.6.3.1. Residue Hydrotreating (Demetallization, Desulfurization, Denitrification) 26
3.6.3.2. Residue Hydrocracking (Hydroconversion) 27
3.6.4. Carbon Rejection ("C-out") Processes 29
3.6.4.1. Thermal Processes (Visbreaking, Coking) 29
3.6.4.2. Catalytic Processes (Residue Cat Cracking) 31
3.6.5. Other Processes 31
3.6.5.1. Extraction of Asphaltenes 31
3.6.5.2. Partial Oxidation 32
3.6.6. Process Combinations 33
3.7. Gasoline Upgrading Processes . . 34
3.7.1. Alkylation 34
3.7.2. Polymerization 35
3.7.3. Isomerization 35
3.8. Other Processes 36
3.8.1. Gas Treating 36
3.8.2. Sulfur Recovery 36
4. Environmental Protection in Oil Refining 37
4.1. Introduction 37
4.2. Manufacturing Emissions 37
4.2.1. Hydrocarbons (HC) 38
4.2.1.1. Hydrocarbons in Air 38
4.2.1.2. Hydrocarbons in Wastewater 39
4.2.1.3. Hydrocarbons in Soil and Groundwater 39
4.2.2. Sulfur and Nitrogen Compounds . . 40
4.2.2.1. Sulfur Compounds 40
4.2.2.2. Nitrogen Compounds 41
4.2.3. Noise 41
4.3. Consumer Related Emissions . . . 42
4.3.1. Transportation Fuels 42
4.3.1.1. Motor Gasoline 42
4.3.1.2. Diesel Fuel 44
4.3.2. Burning Fuels 44
4.4. Cost of Environmental Conservation 44
5. Integrated Refinery Models 46
5.1. Trends of Refinery Structures . . 46
5.2. Hydroskimming Refinery 46
5.3. Conversion Concepts 47
5.3.1. Cat Cracking – Visbreaking Refinery 47
5.3.2. Hydrocracking – Cat Cracking Refinery 48
5.3.3. Hydrocracking – Coking Refinery . 48
6. Corrosion and Materials 48
7. Testing and Analysis 50
7.1. Crude Oil and Product Properties 50
7.2. Testing Methods and Standards . 50
8. Storage and Transport 51
9. References 51

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_051

Abbreviations used in this article:

BTX	benzene – toluene – xylene
CC	catalytic cracker
CR	circulating reflux
DAO	deasphalted oil
DEA	diethanolamine
DIPA	diisopropyl amine
FCC	fluid-bed catalytic cracking
GO	gas oil
IBP	initial boiling point
ICR	intermediate circulating reflux
HC	hydrocracking, hydrocarbons
HDN	hydrodenitrification
HDS	hydrodesulfurization
HGO	heavy gas oil
HT	hydrotreating
LCR	lower circulating reflux
LGO	light gas oil
LHSV	liquid hourly space velocity
LPG	liquefied petroleum gas
MEA	monoethanolamine
MON	motor octane number
MTBE	methyl *tert*-butyl ether
OPEC	Organization of the Petroleum Exporting Countries
RON	research octane number
SDA	solvent deasphalting
TBP	true boiling point
TIP	total isomerization process
UCR	upper circulating reflux
VB	visbreaker
VGO	vacuum gas oil

1. Introduction

1.1. History

Crude oil or petroleum, and a limited number of oil derivatives such as asphalt, pitch, and naphtha have been known and sporadically used since medieval times in the Orient, Europe, and America. However, it was not until the 19th century that significant oil processing started.

The first commercial oil operations can be dated back to the 1850s, with production commencing in Russia, Romania, and the United States [1]. The first successfully purpose-drilled oil well in 1859 in Pennsylvania was the beginning of the modern oil industry, since it stimulated fast development in the search, production, and refining of oil.

Subsequently, small and simple oil refineries were built and operated in direct vicinity of the Pennsylvanian oil fields and on the east coast of the United States, where former coal distillation processes were modified for oil processing. These first distilling units worked in *batch operation* and produced a small number of oil fractions.

Batch distillation was later combined with a thermal cracking step in order to increase the distillate yield. In this crack distillation, the heavier part of the oil was heated further and cracked until coke remained as residue in the still. The lower sections of the vessels were equipped with fired heating tubes to improve heat transfer.

Important progress in oil processing was achieved in Russia with the *continuous crude distillation* process which was introduced in a refinery in Baku in 1875. The continuous principle worked with a battery of stills, arranged in series, with stepwise gravity overflow of the oil throughout the battery. The stills were heated up to successively increasing temperatures, so that each still could produce a distillate fraction corresponding to its temperature level. This continuous process system offered the possibility for mass production at reduced manpower requirements, and utilization of the residual heat for preheating the crude oil feed.

During the first five decades of oil refining the predominant product was kerosene (lamp oil), and only small amounts of other products, such as fuel oil, pitch, and lubricating oil, were marketable. The lubricating oils were produced by vacuum distillation (United States, in the 1870s) and steam distillation (Russia, in the 1880s). The lightest and heaviest fractions of the crude oil distillation, namely gasoline and residue, were not used at that time and had to be burnt.

From the beginning of the 20th century, a complete shift in product demand occurred because of the fast expansion of electricity and motorization. The availability of electricity greatly diminished the demand for lamp oil, whereas the increase in motor vehicles increased the need for gasoline, which soon became the priority product.

The increasing gasoline demand initiated the development of *thermal cracking processes*. The Burton process, which came into operation ca.

1912, produced up to 30 % of cracked gasoline from paraffinic gas oil feedstocks.

During World War I the requirements for a variety of oil products, in particular gasoline and fuel oils, rose dramatically, and the demand continued to increase in the 1920s and 1930s. Gasoline again showed the highest growth rates. In the same period the bitumens (asphalts) (→ Asphalt and Bitumen) became a desired and important product group, owing to the expansion of road construction.

By the mid 1920s a new generation of *improved thermal cracking processes* emerged, of which the Dubbs process gained supreme importance. The gas oil or residue feed was heated in a cracking furnace, and the cracking reactions were completed in reaction chambers in which coke was formed as residual product. The yield of cracked gasoline could be as high as 50 %. However, with the advent of the *Houdry catalytic cracking process* during the second half of the 1930s, the thermal cracking capacities were largely replaced by catalytic cracking units which took the lead in gasoline production in the next two decades.

The catalytic cracking process proved that the octane quality of cracked gasolines that contain unsaturated hydrocarbons was superior to that of distilled (straight-run) gasolines. Hence, the straight-run naphthas were increasingly processed in *thermal reformers* at high cracking temperatures, to add a higher portion of unsaturated hydrocarbons to the motor gasoline.

In the late 1930s the first *catalytic reforming processes* were developed with improved gasoline quality by simultaneous desulfurization and increase in octane number. These *hydroforming* processes used molybdenum-on-alumina catalysts in the presence of hydrogen at high temperature and pressure.

With the introduction of *platinum catalysts* in the late 1940s, the catalytic reforming process was greatly improved: high quality gasolines were produced in higher yields, and hydrogen was formed as byproduct.

A desulfurization step (*hydrotreating*) was installed upstream of the reformer, using cobalt – molybdenum catalysts and hydrogen in order to remove sulfur from the gasoline feedstock. Sulfur would otherwise poison the noble metal catalyst in the reforming process. Owing to their quality and yield advantages, catalytic reforming units replaced the thermal reforming capacities in a relatively short period of time.

In the 1930s the United States and the Soviet Union were the main oil producers and supplied sufficient gasoline products for their own needs. The consumer countries in the rest of the world had to rely on the oil supplies from a number of producing countries, chiefly from South and Central America (Venezuela, Mexico) and the Middle East (Iraq, Saudi Arabia).

The largest expansion of the oil industry took place within the three decades after World War II. Between 1950 and 1980, a fivefold increase in oil demand occurred worldwide. The formation of political and economic blocs in "eastern" and "western" hemispheres and the foundation of the OPEC (Organization of the Petroleum Exporting Countries) in 1960 were significant for international oil trade and distribution.

Oil refining underwent a rapid development initially in the highly industrialized regions of North America, Western and Eastern Europe, Eastern Asia, and Australia, and later, in the wake of increasing populations and energy demands, also in many developing countries in Latin America, Africa, and Asia. Finally, large refinery plants were also built in the OPEC countries for the export of products at the expense of crude oil exports.

The two worldwide crude oil price increases in 1973 – 1974 and 1979 – 1980 prompted changes in oil consumption, which caused the oil industry to develop processes with increased yields and a reduction of refinery capacities, especially in North America and Western Europe. The refining structures became more complex with the installation of higher capacities of residue conversion processes, such as *hydrocracking*, which increased yields of transportation fuels while crude oil consumption could be reduced.

1.2. General Aspects of Oil Refining

Crude oils are complex mixtures of a multitude of – mostly hydrocarbon-type – constituents with a wide variety of properties, hence they cannot be directly used for industrial or commercial applications. Therefore, crude oils must be worked up in a refinery by suitable processes to produce an assortment of products which can

be marketed according to certain quality requirements.

These products are predominantly used as transportation fuels (gasolines, kerosene, diesel) and for heating purposes (LPG, fuel oils), but also as feedstocks for the petrochemical industry (naphtha, LPG) and special manufacturing routes (lubricating oils, greases, bitumen), as well as specialty products (solvents, wax, coke). All these products consist of hydrocarbon mixtures, and there are only few materials that are produced as individual chemical compounds, e.g., propane, sulfur, or benzene – toluene – xylene (BTX) aromatics.

The crude oils are usually transported by tankers and/or pipelines to the refinery site, and stored in tanks before being processed. Refineries were preferably built in coastal regions near industrial consumer centers, but from the late 1950s on, they were also built in inland regions for area-oriented supplies. In the 1970s and 1980s large refinery complexes have also been erected on the coasts of oil exporting regions.

The processing structure of a refinery depends on the market requirements for the products to be sold and on the available type(s) of crude oil. Both factors must be considered to avoid excessive production of unmarketable or otherwise undesirable product(s) and to determine the refinery integration. Typical integrated refinery structures are dealt with in Chapter 5.

Modern refineries are operated continuously and with a high degree of automation. The processing capacities are high compared with other chemical industries. A medium-sized refinery may have a crude capacity of 10 000 t per day [corresponding to ca. 73 000 barrels (bbl) per day]. The crude distillation determines the refinery capacity. A breakdown of worldwide refining capacities in 1997/1998 is given in Table 1. In this table the countries with oil refining capacities are allocated to seven world regions: North America, Latin America, Europe (excluding Former Soviet Union), Former Soviet Union (FSU), Asia (excluding FSU), and Australia/Oceania.

A selection of the following process units (see next page) may be found in modern integrated refineries, depending on the process routes to be installed.

Table 1. World oil refining capacities 1997/1998, capacity figures are given in 10^6 t/a (derived from [2])

Country/region	Refining capacity, 10^6 t/a
United States	795.3
Canada	92.6
Subtotal North America	**887.9**
Brazil	83.2
Mexico	76.0
Venezuela	58.9
Argentina	33.3
Virgin Islands	27.3
Netherl. Antilles	26.3
Cuba	15.1
Other Latin America	76.1
Subtotal Latin America	**396.2**
Italy	122.7
Germany	109.2
France	93.3
United Kingdom	91.3
Spain	64.7
Netherlands	59.4
Belgium	31.5
Romania	27.1
Sweden	21.4
Greece	19.6
Poland	18.3
Bulgaria	16.2
Norway	15.4
Portugal	15.2
Other Europe (excl. FSU)	104.4
Subtotal Europe	**809.7**
Russia	343.6
Ukraine	62.4
Belarus	27.2
Azerbaijan	22.1
Kazakhstan	21.4
Other FSU	32.1
Subtotal Former Soviet Union	**508.8**
Japan	248.4
China	148.4
South Korea	127.1
Saudi Arabia	82.6
Iran	67.9
Singapore	57.9
India	54.3
Indonesia	46.5
Kuwait	44.3
Taiwan	38.5
Turkey	35.4
Thailand	35.2
Malaysia	18.5
Iraq	17.4
Philippines	16.2
Other Asia (excl. FSU)	87.6
Subtotal Asia (excl. FSU)	**1126.2**
Egypt	27.3
Algeria	24.2
South Africa	23.3
Nigeria	21.9
Libya	17.4
Other Africa	32.7
Subtotal Africa	**146.8**
Australia	38.1
Other AUS/Oceania	4.6
Subtotal AUS/Oceania	**42.7**
Total world	**3918.3**

Figure 1. Aerial photograph of a refining complex

Process	Purpose/Products
Crude distillation	crude oil fractionation into gas, gasoline, kerosene, gas oils, and atmospheric residue
Crude desalter (as required)	crude oil desalination
Vacuum distillation	separation of atmospheric residue into vacuum gas oils and vacuum residue
Naphtha hydrotreater	gasoline desulfurization
Kerosene hydrotreater	kerosene desulfurization
Gas oil hydrodesulfurizer	diesel and light heating oil desulfurization
Catalytic reformer	conversion of desulfurized gasoline into high-octane gasoline, LPG, hydrogen, and fuel gas
Gas treating and separation	purification of LPG, fuel gas (e.g., removal of hydrogen sulfide)
Catalytic cracker	conversion of vacuum gas oils or residues into fuel gas, LPG, gasoline, and gas oils
Visbreaker	thermal cracking of vacuum residue into fuel gas, gasoline, gas oil, and heavy fuel oil
Hydrocracker	conversion of vacuum gas oils or residues into fuel gas, gasoline, kerosene, and gas oils
Coker	thermal cracking of vacuum residue into fuel gas, gasoline, gas oil, and coke
Deasphalting	solvent extraction of vacuum residue for deasphalted oil and asphalt
Alkylation (C_3/C_4)	production of alkylate [1]
Polymerization (C_3/C_4)	production of polymerate [1]
Isomerization (C_5/C_6)	production of isomerate [1]
Aromatics plant	solvent extraction of high-octane reformate to produce BTX aromatics
Lubricating oil plant	extraction, dewaxing, and hydrofinishing of vacuum gas oils to produce lubricating oils
Bitumen blowing	oxidation of vacuum residue to bitumen grades
Sulfur recovery	conversion of hydrogen sulfide (from gas treating) into sulfur
Hydrogen plant	hydrogen manufacture

[1] Gasoline components

A photograph of a typical refining complex is shown in Figure 1.

2. Crude Oils and Products

2.1. Crude Oil Constituents

Crude oils are composed of a great variety of *hydrocarbons* (→ Hydrocarbons). Hydrocarbons with one to four carbon atoms (methane – butane) are gaseous at room temperature, hydrocarbons with more carbon atoms are liquid at room temperature (pentane, hexane, etc.). Hydrocarbons with more than seventeen carbon atoms per molecule are solid at room temperature as pure compounds, but these are usually dissolved in the liquid portion of the crude oil.

Three types of molecular structures are possible for hydrocarbons, i.e., straight-chain, branched-chain, and ring structures. Three possible structures of hexane are given as an example:

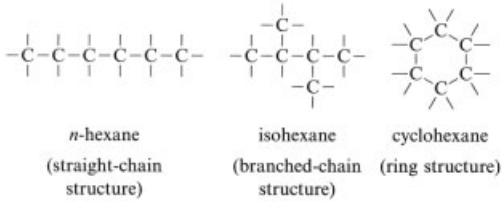

The higher and more complex hydrocarbon molecules consist of various combinations of these structures.

Saturated hydrocarbons are known as paraffins or alkanes (IUPAC name). They can be further classified in normal, iso-, and cycloalkanes. Cycloalkanes are also called naphthenes.

Unsaturated hydrocarbons (olefins or alkenes) are not normally contained in crude oils, but they are formed in various refining processes, particularly during cracking and dehydrogenation.

Crude oils also contain varying amounts of *aromatics*: simple aromatics such as benzene, condensed polynuclear aromatics, and aromatic ring systems with various paraffinic or olefinic side chains. Aromatics are also formed in various conversion processes.

There are also a number of nonhydrocarbons present in crude oils in relatively small concentrations. Examples are sulfur-, nitrogen-, and oxygen-containing compounds. Metal compounds, e.g., vanadium, iron, and nickel compounds occur in trace quantities.

Most of these components are undesirable because they either cause problems in crude oil processing or deteriorate product qualities. Therefore, they must be removed by suitable refining processes.

Sulfur compounds such as hydrogen sulfide, thiols (mercaptans), sulfides (thioethers), disulfides, polysulfides, and thiophenes, may cause corrosion of equipment, bad odor in products, and catalyst poisoning. Moreover, sulfur compounds in fuel products contribute to air pollution by formation of sulfur dioxide and trioxide when burnt, and thus have to be removed from oil products as far as possible. This is achieved by means of desulfurization processes and subsequent conversion of hydrogen sulfide to elemental sulfur as a marketable product.

Nitrogen compounds are more stable than sulfur compounds, and are only partially removed in the desulfurization processes. Specific denitrogenation may be desirable for catalytic cracking or hydrocracking feedstocks to avoid catalyst deterioration, and for fuel oils to reduce formation of nitrogen oxides in burning.

Oxygen compounds may be present as naphthenic acids and phenols. *Naphthenic acids* (mostly mono- or dicyclic cycloalkane carboxylic acids, e.g., cyclopentane carboxylic acid) are constituents of naphthenic crude oils. Being corrosive, they have to be removed from crude oils or distillates by alkali extraction in the distillation process. The separated naphthenic acids are marketed as acidic extractants for metals; naphthenates are used as metallic soaps.

Phenols (which cause odor problems) are formed and recovered in the catalytic cracking process.

Metals act as catalyst poisons and can be removed by a demetallization step upstream of catalytic residue conversion processes. In catalytic cracking metal passivation additives are applied where buildup of vanadium and nickel in the catalyst bed poses a problem.

2.2. Classification of Crude Oils

The quality of crude oils can be classified either according to hydrocarbon type, product fractions, or sulfur content. Although rather unspecific, these classification methods are useful for the selection of suitable crude oils for given refining structures and product demand patterns.

Classification According to Hydrocarbon Types.

Paraffin-based crude oils consist predominantly of paraffinic hydrocarbons. The low and medium molecular mass paraffins are suitable for all kinds of catalytic conversion into gasolines and middle distillates, and for chemical feedstocks. The high molecular mass paraffins are usually good raw materials for the manufacture of lubricating oils and paraffin waxes.

Naphthene-based crude oils mainly contain naphthenic, aromatic, and asphaltic hydrocarbons. The low and medium molecular mass ranges can be used for the production of valuable gasoline components, solvents, and feedstocks for aromatics manufacture. The high molecular mass ranges find use in special types of lubricants and in the production of bitumen.

Mixed-based crude oils have a more or less even distribution of paraffinic and naphthenic/aromatic hydrocarbons over all ranges of molecular mass. They constitute the largest class of crudes.

Classification According to Product Fractions. This characterization commonly distinguishes between "light", "medium", and "heavy" crudes, i.e., the classification is based on the proportions of distillates and residue. This quality distinction takes into account the wide compositional varieties of crude oil fractionation. The percentages of the main fractions of some well-known crudes are listed in Table 2.

Table 2. Yields of distillation fractions of selected crude oils

Type of crude, origin	Gas, LPG, vol %	Gasoline, vol %	Kerosene, vol %	Gas oil, vol %	Atmospheric residue, vol %
Nigerian light, Nigeria	3	26	14.5	27.5	29
Ekofisk, Norway	3.5	31	13.5	21.5	30.5
Forties, United Kingdom	4	22.5	12	22	39.5
Arabian light, Saudi Arabia	1.5	20.5	12	21	45
Alaska, United States	0.5	13.5	11.5	21.5	53
Tia Juana, Venezuela	–	1.5	3.5	14.5	80.5

The *light crudes* which are rich in gasoline and gas oil are, for example, represented by Norwegian and Nigerian crudes with correspondingly low residue contents. The *heavy crudes* such as Venezuelan or Alaskan crude oils contain relatively small amounts of distillates but high portions of residue, thus requiring higher conversion capacities than lighter crudes. A *middle group*, e.g., Saudi Arabian or British crude, has an advantageous (i.e., well-balanced) distribution of fractions.

Classification based on percentages of product fractions may assist in the crude selection for a refinery according to product priorities, taking into consideration the refinery conversion structure.

The lighter crude types are usually more expensive than those with high residual portions.

Classification According to Sulfur Content. An important crude property is the sulfur concentration which determines the processing complexity of a refinery to a considerable extent. The higher the sulfur content in the crude and its various fractions, the stronger is the need for adequate desulfurization and sulfur recovery capacities. Table 3 lists the sulfur contents of a number of crudes and their residues.

Table 3. Sulfur contents of selected crude oils and their atmospheric residues

Type of crude, origin	Sulfur content of crude, wt %	Sulfur content of atmospheric residue, wt %
Ekofisk, Norway	0.1	0.3
Forties, United Kingdom	0.3	0.6
Es-Sider, Libya	0.5	0.9
Alaska, United States	1.0	1.5
West Siberia, Russia	1.5	2.6
Isthmus, Mexico	1.6	3.0
Kuwait, Kuwait	2.5	3.8
Arabian light, Saudi Arabia	3.0	4.5

The sulfur content is an important factor worldwide in pricing crude oil; the group of low-sulfur crudes with <1 wt % of sulfur constitute the most desirable but generally most expensive qualities. Crude oils with 1.0 – 1.5 wt % of sulfur are classified as medium-sulfur, and those having above 1.5 wt % as high-sulfur crudes.

2.3. Oil Products

In general ca. 80 – 90 % of oil products are used as transportation and combustion fuels. The most important products, their applications, and some typical specifiactions are briefly described below.

2.3.1. Gas Fuels

1) *Refinery gas* consists of gas streams that contain mostly methane and ethane which are burnt as fuel gas in refinery furnaces.
2) *Liquefied petroleum gas* (LPG) (→ Liquefied Petroleum Gas) consists of propane, butanes, and mixtures thereof. It is used for heating purposes and special products. In some countries, e.g., Japan, the United States, and The Netherlands, LPG is also used to a considerable extent as motor fuel. LPG specifications are based on national standards. The specified characteristics mainly concern vapor pressure, density, and contents of impu-

rities (e.g., elemental sulfur, COS sulfur, hydrogen sulfide, vaporization residue, hydrogen, nitrogen, oxygen, and methane).

2.3.2. Liquid Fuels

Motor gasolines are blended products with a boiling range between 40 and 200 °C and are used for motor car engines (spark-ignition motors) (→ Automotive Fuels). Typical gasoline components are light and heavy cracker and reformer gasolines, alkylates, isomerates, polymerates, pyrolysis gasoline, and LPG. Alcohols and ethers are added as nonhydrocarbon products. Gasolines have to comply with a large variety of specifications, among which the knock rating (octane number), volatility, boiling characteristics, density, oxidation stability, and lead contents are of prime importance. Typical specifications for unleaded gasolines (regular and premium grades, EN 228) are:

Regular grade	
Knock ratings	
RON	these ratings are specified according to national standards (e.g., for Germany RON min. 91.0, MON min. 82.5)
MON	
Density (at 15 °C)	min. 725 kg/m^3 max. 780 kg/m^3
Vapor pressure	specified according to applicable volatility classes (table in EN 228)
Boiling characteristics	
Lead content	max. 0.013 g/L
Benzene content	max. 5.0 vol %
Sulfur content	max. 0.05 mass %
Oxidation stability	min. 360 min
Premium grade	
Knock ratings	
RON	min. 95.0
MON	min. 85.0
Density (at 15 °C)	min. 725 kg/m^3 max. 780 kg/m^3
Vapor pressure	specified according to applicable volatility classes (table in EN 228)
Boiling characteristics	
Lead content	max. 0.013 g/L
Benzene content	max. 5.0 vol %
Sulfur content	max. 0.05 mass %
Oxidation stability	min. 360 min

The maximum addition rates of oxygen-containing gasoline components (*oxygenates*, e.g., methanol, ethanol, isopropanol, 2–methylpropanol, and ethers) are specified by relevant national standards.

Kerosenes are tailor-made products in the boiling range between 150 and 250 °C.

Lamp kerosene (lighting oil, cooking oil) has the longest tradition of oil products and is still widely used in most areas of the world. Smoke point, flash point, and volatility are important specifications.

Jet fuels, or aviation turbine fuels (→ Aviation Turbine Fuels), which are used worldwide in commercial aviation, have to undergo stringent quality screening. Control of density, freezing characteristics, viscosity, thermal stability, conductivity, corrosion behavior, water content, aromatics, and sulfur content is essential. A selection of Jet A-1 specifications is given below (based on relevant ASTM methods) (see also → Aviation Turbine Fuels, Chap. 2.1.).

Density (at 15 °C)	775 – 840 kg/m^3
Distillation	
10 vol % recovered at	max. 205 °C
End point	max. 300 °C
Residue/loss	max. 1.5 vol % each
Flash point	min. 38 °C
Acid number	max. 0.015 mg KOH/g
Aromatics content	max. 22.0 vol %
Mercaptan sulfur content	max. 0.003 mass %
Sulfur content	max. 0.30 mass %
Freezing point	max. − 47 °C
Viscosity (at − 20 °C)	max. 8.0 mm^2/s
Specific energy	min. 42.8 MJ/kg
Smoke point	min. 25 mm
Copper corrosion (2 h/100 °C)	max. 1
Thermal stability (at 260 °C)	
Tube depositrating	< 3
Filter pressure drop	max. 25.0 mm Hg
Electrical conductivity	50 – 450 pS/m

Further specifications are given for contaminants and additives. Additional domestic specifications have to be observed in various countries.

Gas oils are light and heavy gas oil fractions, and blends thereof, from straight-run and cracked origins, in a boiling range between 200 and 350 °C. The gas oils are predominantly used as automotive diesel fuels (→ Automotive Fuels) and as domestic heating fuels (→ Heating Oil). Typical specifications for automotive diesel fuel and domestic fuel oil are listed below.

Automotive diesel fuel (specifications for moderate climate according to EN 590):

Density (at 15 °C)	min. 820 kg/m^3 max. 860 kg/m^3
Kinematic viscosity (at 40 °C)	min. 2.0 mm^2/s max. 4.5 mm^2/s

Distillation	
Recovery at 250 °C	max. 65 vol %
Recovery at 350 °C	min. 85 vol %
Recovery at 370 °C	min. 95 vol %
Cetane number	min. 49
Cetane index	min. 46
Flash point	min. 55 °C
Cold filtration test (CFPP)	classified between + 5 °C and − 20 °C for regional specifications and interseasonal qualities
Sulfur content	max. 0.20 mass %
Water content	max. 200 mg/kg
Ash content	max. 0.01 mass %
Carbon residue	max. 0.30 mass %
Oxidation stability	max. 25 g/m^3

Domestic fuel oil (specifications for extra light heating oil according to German standard DIN 51603-1 as an example):

Density (at 15 °C)	max. 860 kg/m^3
Kinematic viscosity (at 20 °C)	max. 6.0 mm^2/s
Distillation	
Recovery at 250 °C	max. 65 vol %
Recovery at 350 °C	min. 85 vol %
Flash point	min. 55 °C
Pour point	max. − 9 °C
Calorific value (net)	min. 42.6 MJ/kg
Carbon residue	max. 0.5 mass %
Sulfur content	max. 0.20 mass %
Water content	max. 200 mg/kg
Ash content	max. 0.01 mass %

Heavy fuel oils consist of various grades of mixtures of residual oils from distilling and conversion processes (→ Heating Oil). High density, viscosity, or sulfur content of very heavy oils can be corrected by blending with gas oils. Marine fuels (bunker fuels), power stations and industrial furnaces are major outlets for heavy fuel oils. Specifications of various grades of marine distillative fuels and marine residual fuels are given in ISO 8217.

Important characteristics are density, kinematic viscosity, flash point, pour point (mostly defined for summer and winter qualities), carbon residue, ash, water, and sulfur content, and sediments.

2.3.3. Nonfuel Applications

Nonfuel applications are as follows:

1) LPG and light naphtha fractions are used as feedstocks for the petrochemical industry.
2) Special boiling-range naphthas are produced for various grades of solvents used in the chemical and pharmaceutical industry.
3) Paraffin waxes are used in the paper, chemical, and textile industries, and in medicine.
4) Heavy distillates are employed for the production of lubricants and greases (→ Lubricants and Related Products).
5) Heavy residues are raw materials for the manufacture of road bitumens and asphalts (→ Asphalt and Bitumen).
6) Petroleum coke (→ Petroleum Coke) and high-grade coke are used for various industrial applications and electrode manufacture, respectively.

3. Oil Refining Processes

3.1. Crude Oil Distillation

The fractional atmospheric distillation of crude oil is performed in continuously operated fractionation columns. The necessary heat for the fractionation is introduced through fired tube furnaces. A great variety of trays or decks are applied as column internals. Examples are grid trays, bubble cap trays, box trays, and valve trays which have been developed with the aim of improving the stage-to-stage vapor – liquid separation in the columns. Special types of trays are designed for distributing the incoming product and for collecting the draw-off product.

Figure 2 (see next page) shows the temperature – yield diagram for the boiling curves of four fractions of a crude fractionation, i.e., light and heavy gasoline (naphtha), kerosene, and gas oil. Their boiling ranges between the 5 and 95 vol % yield points are ca. 45 – 105 °C, 110 – 155 °C, 165 – 260 °C, and 240 – 360 °C, respectively. In this example there is obviously a boiling "overlap" between kerosene and gas oil; boiling "gaps" exist between heavy gasoline and kerosene and between the two gasoline fractions. The overlap reveals a rather inferior fractionation, whereas the gaps indicate good fractionation in the respective column sections.

Hydrocarbon fractions boiling above 360 °C should not be recovered by atmospheric distillation because cracking of the heavier products starts above this temperature. Therefore, the atmospheric residue must be subjected to vacuum

distillation. Under the reduced pressure in vacuum columns, the higher-boiling oil fractions can be vaporized under similar temperature conditions as in the gas oil section of atmospheric fractionation. The final boiling point of a full-range vacuum gas oil (feedstock for catalytic cracker) would correspond to ca. 550 °C at atmospheric conditions.

3.1.1. Atmospheric Distillation

Atmospheric distillation is the first step in crude oil processing. The crude oil is separated into fractions of different boiling ranges. The distillation columns are usually operated continuously.

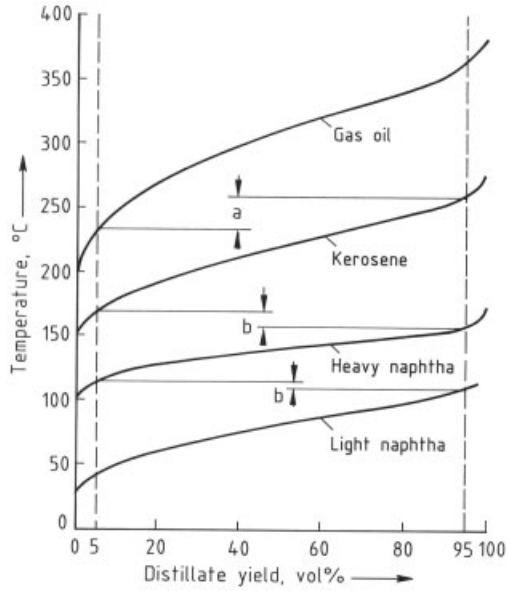

Figure 2. Temperature – yield diagram of ASTM boiling ranges of crude oil fractions
a) Boiling overlap; b) Boiling gap

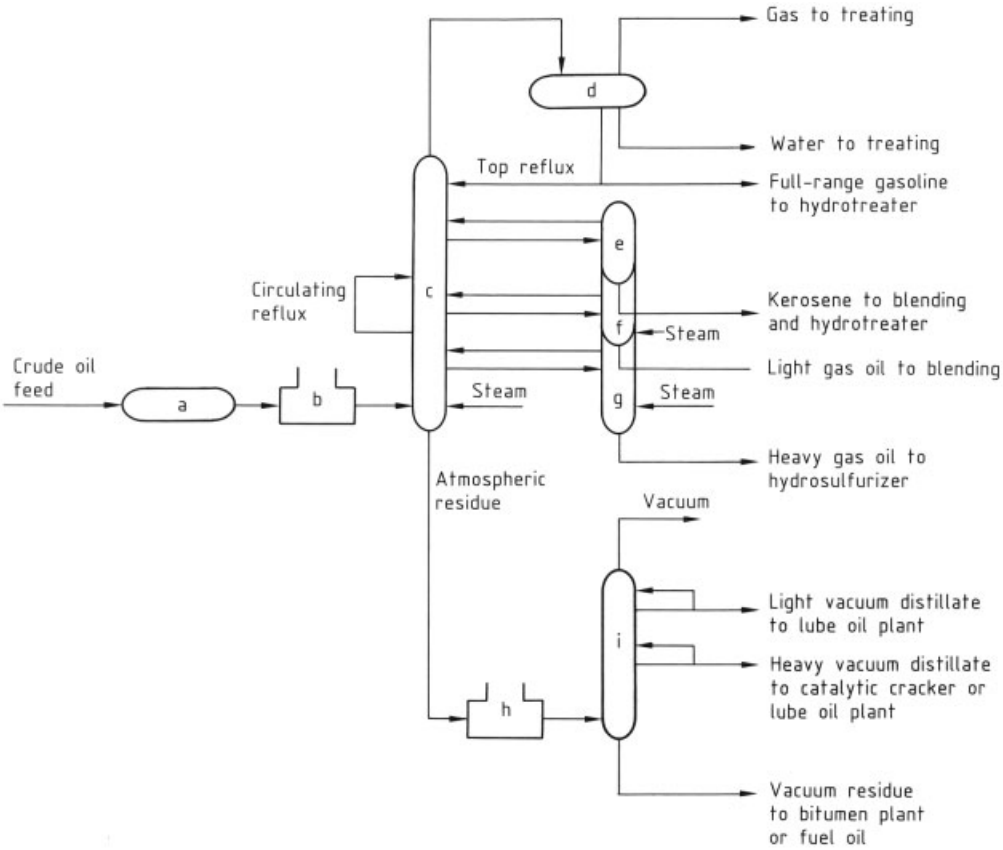

Figure 3. Schematic representation of crude oil distillation including crude desalting and vacuum flashing
a) Crude desalter; b) Crude heater; c) Main fractionator; d) Overhead accumulator; e) Kerosene stripper; f) Light gas oil stripper; g) Heavy gas oil stripper; h) Vacuum heater; i) Vacuum flasher

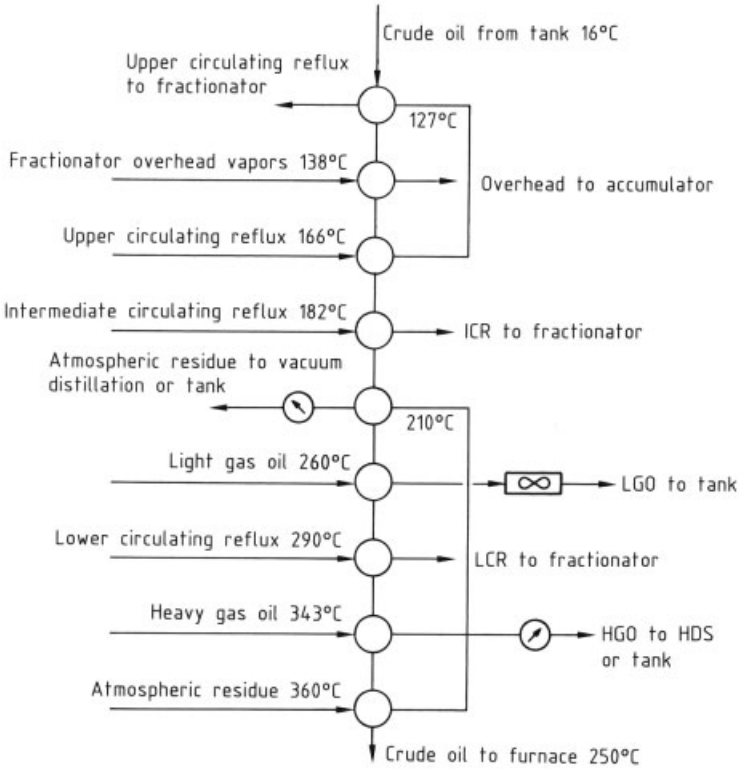

Figure 4. Crude preheat train of a crude oil distilling unit

Some major properties of the final products are established by the fractionation, for example, the boiling ranges of automotive fuels, the flash points and viscosities of lubricating oils, or the penetration (viscosity) of heavy fuel oil.

Modern fractionation columns are designed to yield one or several side streams in addition to a top and a bottom fraction. Figure 3 shows a typical atmospheric crude oil distillation unit, combined with a crude desalter and vacuum distillation. The crude oil is pumped from storage through the desalter (see Section 3.1.3) and is preheated in a series of heat exchangers (see Fig. 4) to ca. 220 – 250 °C before entering the furnace (crude heater). Here it is heated up to the required column inlet temperature which is in the range of 360 – 380 °C. In the main fractionator the partially vaporized crude is separated into a top product, three side products, and a bottom product.

Top Product. The column top vapors are condensed and separated into gas and full-range gasoline in the overhead accumulator. Part of the *liquid gasoline* is pumped back to the column as top reflux; the remainder is fed to the naphtha hydrotreater (see Section 3.3.1). The *gas stream* (uncondensable $C_1 - C_4$ hydrocarbons and H_2S) from the overhead accumulator is processed further in a gas treatment plant.

Process water originating from the stripping steam that is applied to the fractionator and stripper columns, is withdrawn from the sump of the overhead accumulator and treated further in a wastewater treatment plant.

Side Streams. The side streams comprising of the middle distillates, i.e., kerosene, light and heavy gas oil, are transferred to their respective stripper columns where the lighter constituents are stripped off by injection of superheated steam.

The *kerosene stream* from the stripper column is pumped off either to a hydrotreater for the production of jet fuel, or via a heat exchanger and cooler directly to storage if it is to be used as a blending component for diesel fuel.

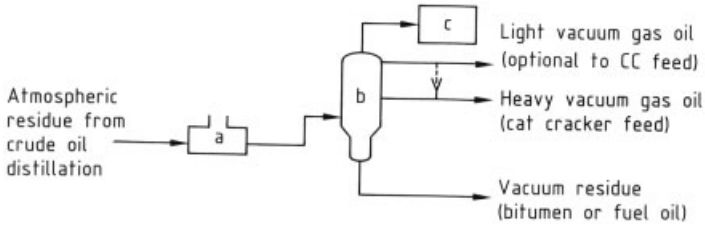

Figure 5. Vacuum flasher for catalytic cracker feedstock
a) Residue heater; b) Vacuum flasher; c) Vacuum system

The *light gas oil* fraction is also pumped to storage via a heat exchanger and cooler and is used for diesel and light heating oil blending.

The *heavy gas oil* fraction is fed to the hydrodesulfurization unit when it is used for the production of diesel fuel and light heating oil, or it is (partially) pumped via a heat exchanger and cooler to storage for direct blending.

One or more liquid side streams of the main fractionator column can be used as circulating reflux (CR). These circulation streams deliver heat to the crude oil via heat exchangers and contribute to the fractionation sharpness by returning the cooled product back to the column.

Bottom Product. The bottom product, namely atmospheric residue, can be supplied to a vacuum fractionator for the production of lubricating oil fractions (see Fig. 3) or to a vacuum flasher for the preparation of catalytic cracker feedstock (see Section 3.1.2). Alternatively, the residue can serve as feedstock for a variety of other cracking processes, e.g., thermal and catalytic cracking, hydrocracking, or hydroconversion.

3.1.2. Vacuum Distillation

Generally two versions of vacuum distillation are applied – the vacuum flasher and the vacuum fractionator.

The *vacuum flasher* has a relatively simple design with a limited number of trays, because it serves only for the separation of atmospheric residue into a wide-boiling vacuum gas oil that is used as catalytic cracker (CC) feedstock and a vacuum residue for bitumen manufacture or visbreaker feed. No specific separation sharpness is required for preparing the CC feed. Figure 5 shows a typical vacuum flasher. The vacuum is produced by a set of steam ejectors as shown in Figure 6.

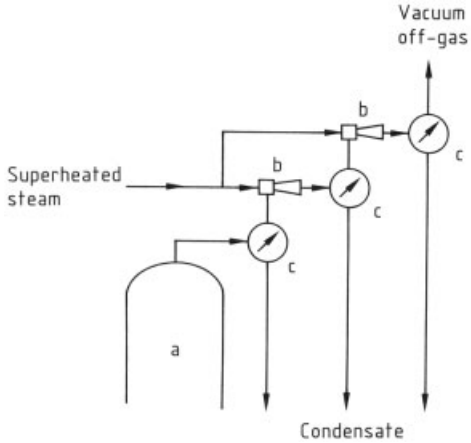

Figure 6. Vacuum ejector system
a) Vacuum column; b) Steam ejectors; c) Condensers

The *vacuum fractionator* is essential for the preparation of lubricating oil fractions. The vacuum fractionator separates the vacuum gas oil portion of the atmospheric residue into light and heavy spindle oils, light and heavy machine oils, and cylinder oil (→ Lubricants and Related Products). To achieve this high degree of fractionation the vacuum distilling section is often splitinto two fractionator columns.

3.1.3. Crude Oil Desalting

Although crude oils delivered from production fields are desalted, a small portion of water and salt always remains in the crudes which reach the refiners. Salt contents (sodium, calcium, and magnesium chlorides and sulfates) can have detrimental and costly effects on oil processing equipment. Examples are corrosion in the fractionator overhead system due to the formation of hydrochloric acid; erosion or sedimentation of salt in lines, furnace tubes, pumps,

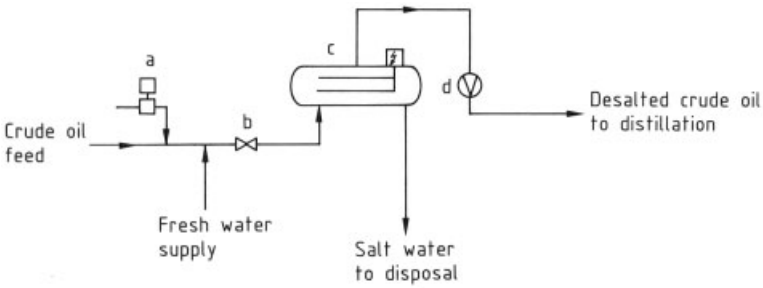

Figure 7. Electrical crude oil desalter
a) Chemical injection; b) Mixing valve; c) Crude desalter; d) Crude booster pump

and valves; and impaired heat transfer due to salt deposits in heat exchangers. It is therefore advisable to install a desalting unit in the crude feed line to the distillation unit.

The most common desalting unit is the *electrical desalter.* An electrical field is applied in the desalting vessel which breaks up saltwater – oil emulsions by coalescence of the water droplets and segregation of the water from the oil phase. A typical electrical desalter is shown in Figure 7.

3.2. Catalytic Cracking (Cat Cracking)

The commissioning of the first catalytic cracking plant in 1936 (Houdry process, United States) initiated a new era in petroleum refining. Until then, thermal cracking of distillates had been the governing conversion process, the role of which was taken over very soon by catalytic cracking. This development was particularly influenced by the increased demand for high-quality transportation fuels during World War II.

Four major advantages of catalytic over thermal cracking supported this development:

1) Lower production of C_1 and C_2 hydrocarbons to the credit of higher yields of C_3 and C_4 hydrocarbons (Fig. 8), which can be used for the manufacture of isoparaffins for motor gasoline via alkylation and polymerization processes.
2) Lower cracking temperatures which result in a lower production of diolefins, thus improving the oxidation stability of the gasoline fractions.
3) Higher octane ratings of gasoline by increased formation of branched alkanes, naphthenes (cycloalkanes) and aromatics.
4) Higher cracking reaction rates, allowing smaller dimensioning of plant equipment.

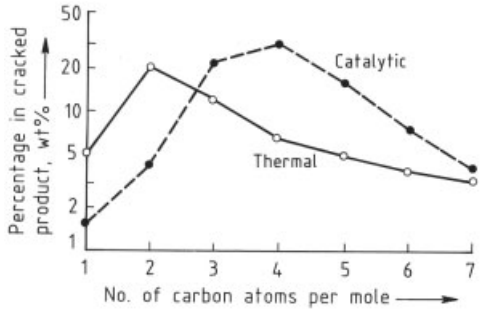

Figure 8. Product distribution of $C_1 - C_7$ hydrocarbons from thermally and catalytically cracked C_{16} key component [3]

The conversion reactions of oil distillates in the catalytic cracking process occur mainly in the vapor phase at elevated temperatures in the presence of a cracking catalyst (originally of the montmorillonite type, currently synthetic crystalline zeolites). The cracking reactions occuring at the active sites of the catalysts proceed via a carbenium ion ion mechanism that predominantly effects the formation of olefins, isomeric components, and aromatics (the latter via intermediate formation of cycloolefins. The formation of low-boiling olefins, branched alkanes, and aromatics favors the production of gasolines with high octane levels.

A considerable portion of the cracker feedstock is converted to gaseous components which are separated in a gas plant into fuel gas (containing hydrogen, methane, ethane, ethylene, and hydrogen sulfide) and LPG fractions, i.e., propane – propene and butane – butene. After treatment, these LPG streams are predominantly

processed in alkylation and polymerization units for the production of high-octane gasoline components (see Sections 3.7.1, 3.7.2). Part of the propane and butanes is also used for LPG sales or as chemical feedstock.

The *cracked gasoline* (research octane number > 90) must be treated to remove undesirable impurities, such as hydrogen sulfide, mercaptans, phenols, and nitrogen compounds, because they are responsible for unacceptable odor and corrosion.

The *middle distillates*, i.e., light and heavy gas oils, serve as blending stocks for light heating and diesel oils and heavy fuel oil, respectively. The cat cracked gas oils are traditionally referred to as light and heavy "cycle oils" since they have been widely used as recycle products to maximize gasoline production in the process.

An important byproduct of catalytic cracking is coke which originates from heavy carbonaceous material. Coke forms deposits on the catalyst particles and thus impairs catalyst activity. Therefore, the process must include a regeneration step where the carbon is burnt off the catalyst for reactivation before entering the reaction step.

Distillate fractions have been used for about four decades as catalytic cracking feedstocks – chiefly *heavy gas oils* (vacuum gas oils) and to some extent *deasphalted oils*. Residual materials could not be processed because of their heavy metals content (nickel and vanadium) since this caused irreversible poisoning of the cracking catalysts. In addition, the metals and the asphaltene contents of the residues led to increased coke deposition on the catalysts. However, the introduction of metal-resistant catalysts and installation of feed pretreatment facilities since the mid 1970s have also made atmospheric residues accessible for catalytic cracking processes (see Section 3.6.4.2).

The development of catalytic cracking units gave rise to many improvements to the process within a relatively short time span. The first Houdry design was a fixed-bed process which required cyclic operation (switching of a series of vessels between reaction and regeneration steps). A new continuous type of process, the Thermofor Catalytic Cracking (TCC), then emerged. This process version used a furnace (Thermofor kiln) as regenerator for coke burning, and mechanical catalyst transport by elevators between reactor and furnace and vice versa. This moving-bed system was later replaced by gas-flow, and even later by gas-lift transport.

The introduction of powdered catalyst types gave way for the development of fluid-bed catalytic cracking (FCC) in 1942 (United States). Currently, FCC represents the most commonly applied catalytic cracking process. In a lifting gas stream the catalyst powder behaves like a fluid and can be transported in fluidized form through the system. Figure 9 shows the principle of an FCC process.

The catalyst flowing down from the regenerator (c) is combined with feed and steam, lifted up through the riser (d) into the reactor (a), where it is fluidized by hydrocarbon vapors. The catalyst then flows down via the stripper (b) to the regenerator, where it is fluidized by combustion gases.

Significant improvements in the quality of the cracking catalyst were also achieved. The change from the natural clay materials (montmorillonite), which were originally used, to synthetic amorphous silica – alumina catalysts in the 1940s and the introduction of synthetic crystalline zeolites in the 1960s produced remarkable gains in activity, stability, and selectivity of the catalysts. The addition of highly active zeolitic material to the silica – alumina catalysts resulted in a strong selectivity gain, i.e., increased yields of gasoline and less gas production.

Process Description. The flow diagram of a typical FCC process is shown in Figure 9. Hot catalytic cracker feed and dispersion steam enter the process at the riser foot (d) where they are mixed with the hot reactivated catalyst flowing down the regenerator standpipe (e_1). The feed is vaporized by the hot catalyst, and the mixture is lifted up the riser into the reactor (a), where the catalyst is segregated from the vapors. The cracking reactions usually start in the riser and are completed in the reactor. The vapors (cracked products) pass through cyclone separators to eliminate any remaining catalyst particles before they enter the fractionator section.

The catalyst in the reactor bed, now loaded with coke, drops into the stripper (b) where remaining hydrocarbons are stripped off by the addition of steam. From there the catalyst flows down via the stripper standpipe (e_2) into the regenerator bed (c), where the coke is burnt off

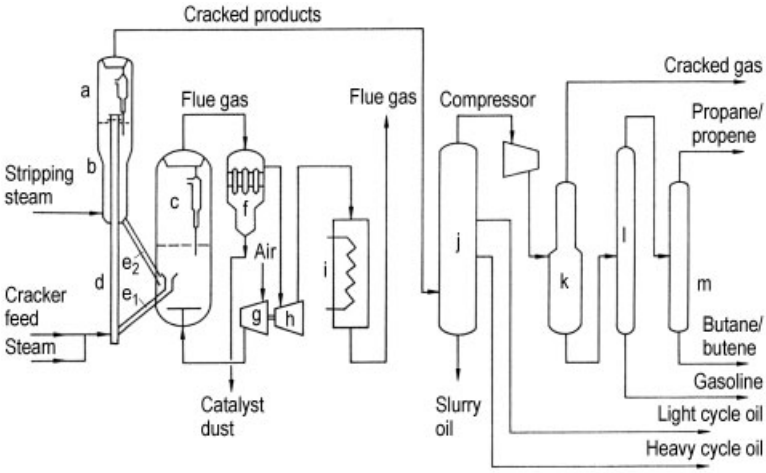

Figure 9. Typical FCC unit
a) Reactor; b) Stripper; c) Regenerator; d) Riser; e_1) Regenerator standpipe; e_2) Stripper standpipe; f) Cyclone vessel; g) Air blower; h) Flue gas expander; i) Waste-heat boiler; j) Fractionator; k) Absorber; l) Debutanizer; m) Depropanizer

by blowing combustion air into the vessel. The regenerated catalyst flows down through the regenerator standpipe to the riser foot. The flue gas leaves the regenerator via cyclone separators (f) to remove catalyst fines to avoid erosion of the internals of the downstream expander by catalyst particles. The gas then enters the heat recovery section (waste-heat boiler i).

The air blower (g) is driven by an expansion turbine (h) which utilizes the pressure difference of the flue gas. The heat of the flue gas is utilized in a waste-heat boiler for steam generation.

The fractionation and gas separation section consists of four columns – the fractionator (j), absorber (k), debutanizer (l), and depropanizer (m). The cracked products from the reactor are separated in these columns into fuel gas (C_2 and lighter), propane – propene, butane – butene, gasoline, light and heavy cycle oils (middle distillates), and slurry oil. The slurry oil contains some catalyst dust (carried over with the reactor product) and is normally recycled to the process.

The cracking sections of four different catalytic cracking processes are depicted in Figures 10, 11, 12, and 13 in more detail. The downstream fractionation and gas separation sections are not shown.

Fluid-Bed Catalytic Cracking Unit (Fig. 10) [4–7]. The partition of the regenerator vessel into a combustor and a disengaging section provides particularly efficient air – catalyst contact in the combustor section and low catalyst loadings in the regenerator cyclones.

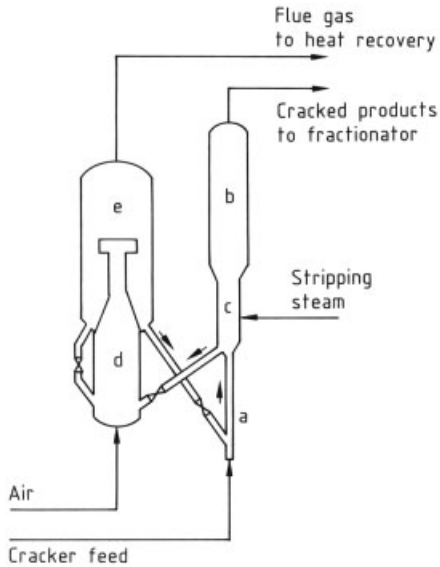

Figure 10. Fluid-bed catalytic cracking
a) Riser; b) Reactor; c) Stripper; d) Combustor; e) Disengaging vessel

Riser Reactor FCC (Fig. 11) [5–9]. The cracking reactions proceed to completion in the riser, whereas the reactor serves as a catalyst – product disengaging vessel. Special flexibility is given by the two feed inlets at different eleva-

tions on the riser, and by the possibility of rapid conversion changes between maximum gasoline and maximum gas oil operation.

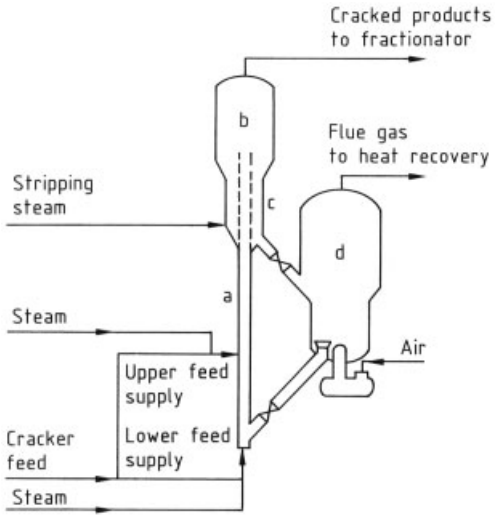

Figure 11. Fluid-bed catalytic cracking with riser reactor
a) Riser reactor; b) Cyclone vessel; c) Stripper; d) Regenerator

Figures 12 and 13 show the *riser-bed* and *transfer line flexicrackers*, respectively [6–8]. The transfer line configuration is applied for operations that require a short reaction time. The cracking reactions are interrupted at the top of the transfer line where fast separation of catalyst and cracked products occurs.

3.3. Hydrotreating

Until the 1940s there was little incentive for the petroleum industry to improve product quality by means of hydrogen treatment. Since the early 1950s, however, several influences have triggered the development of various hydrotreating processes. The increased production of high-sulfur crude oils and consequently the need to remove sulfur compounds from oil fractions, more stringent product specifications because of environmental requirements, and the production of increased quantities of cracked material from conversion processes boosted the installation of adequate hydrotreating capacities in refineries. Hydrotreating plants represent the highest capacities of all secondary refining processes worldwide.

The most important function of hydrotreating is the removal of sulfur components (hydrodesulfurization, HDS) by reaction with hydrogen in the presence of a suitable catalyst, to form hydrogen sulfide. Hydrogen sulfide is removed from the process gas stream using a solvent (e.g., amine) wash and is then converted into elemental sulfur in a Claus plant.

Hydrotreating processes are applied to achieve the following objectives:

1) Removal of sulfur, nitrogen, oxygen, and arsenic compounds from *reformer feedstocks* (naphthas) to prevent poisoning of the plat-

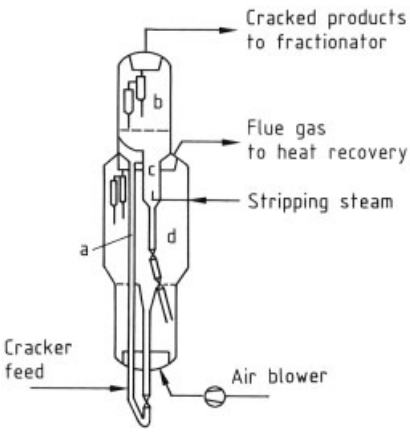

Figure 12. Flexicracking with riser-bed
a) Riser; b) Reactor; c) Stripper; d) Regenerator

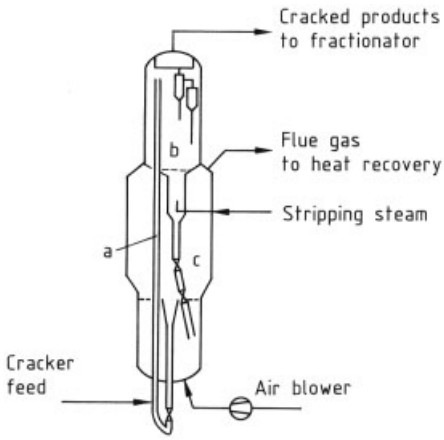

Figure 13. Flexicracking with transfer line reactor
a) Transfer line reactor; b) Cyclone vessel – stripper; c) Regenerator

inum catalysts in the reformers. Hydrogenation of the unsaturated hydrocarbons of cracked naphthas to minimize coking of the reformer catalysts.

2) Increase in the thermal stability of *jet fuels*. Additionally, an improvement of the smoke point can be aimed at by (partial) hydrogenation of aromatics.
3) Removal of sulfur compounds from *middle distillates* (gas oils) to achieve blending qualities for diesel fuels and light heating oil. Unsaturated hydrocarbons in cracked gas oils are hydrogenated to improve the color stability and reduce the coking tendency of the products.
4) Removal of sulfur, nitrogen, and metal compounds from *catalytic cracking feedstocks* (vacuum gas oils, atmospheric residues) to limit the catalyst deactivation by nitrogen and metals, and to yield lower sulfur contents in the cracked products.
5) Reduction of the sulfur content of *heavy fuel oils*.
6) Hydrogenation of diolefins in *pyrolysis gasolines* (originating from ethylene production plants) to avoid the formation of gum which deteriorates the stability of these gasolines.
7) Improvement of odor, color, and oxidation stability of *lubricating oils* by hydrofinishing (mild hydrotreating). Dearomatization and removal of sulfur, nitrogen, and oxygen compounds by deep hydrogenation (high-pressure hydrotreating) of lubricating oils.

3.3.1. Distillate Hydrotreating

Two process types are used – the vapor-phase process for *hydrotreating of gasoline* (where the feed is vaporized in the reaction stage) and the trickle phase process for *hydrotreating and desulfurizing of straight-run middle distillates* or vacuum gas oils (where the major portion of the feed is liquid in the reaction stage).

The process scheme of both hydrotreating process types is identical. The feed stream is mixed with hydrogen-rich gas (usually fresh gas from catalytic reforming plus recycle gas from the hydrotreating process) and heated before it is passed through the catalyst bed in the reactor. The reactor product is cooled and separated into a gas and a liquid phase in two consecutive separation stages (high-pressure and low-pressure separators). The hydrogen-rich recycle gas from the high-pressure separator is recombined with the feed. The low-pressure off-gas which contains hydrogen sulfide is processed further in a gas treatment plant for H_2S removal. The liquid reactor product is stabilized in a debutanizer (gasoline hydrotreater) or stripper column (gas oil hydrodesulfurizer).

The two process types are shown schematically in Figure 14 (gasoline hydrotreating) and Figure 15 (gas oil hydrodesulfurizing), respectively. The catalyst types employed and the main operating criteria of distillate hydrotreatment processes are given in Table 4.

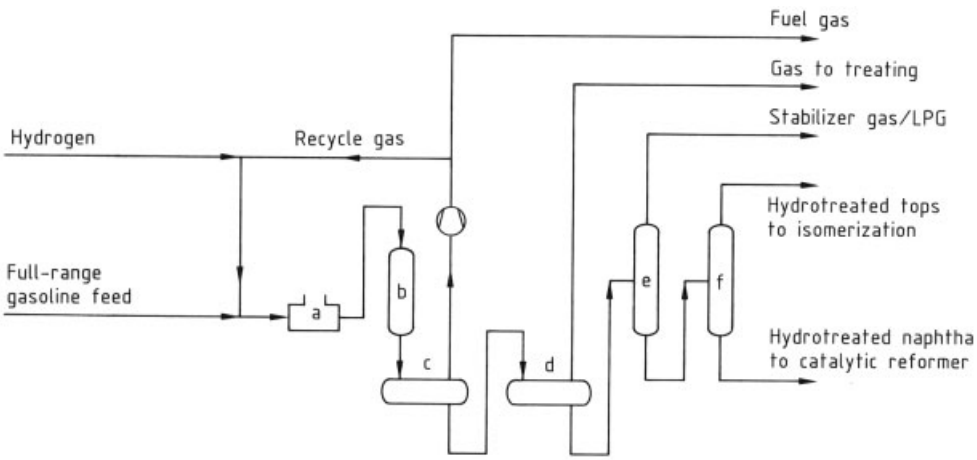

Figure 14. Schematic representation of gasoline hydrotreater
a) Process heater; b) Reactor; c) High-pressure separator; d) Low-pressure separator; e) Stabilizer; f) Gasoline splitter

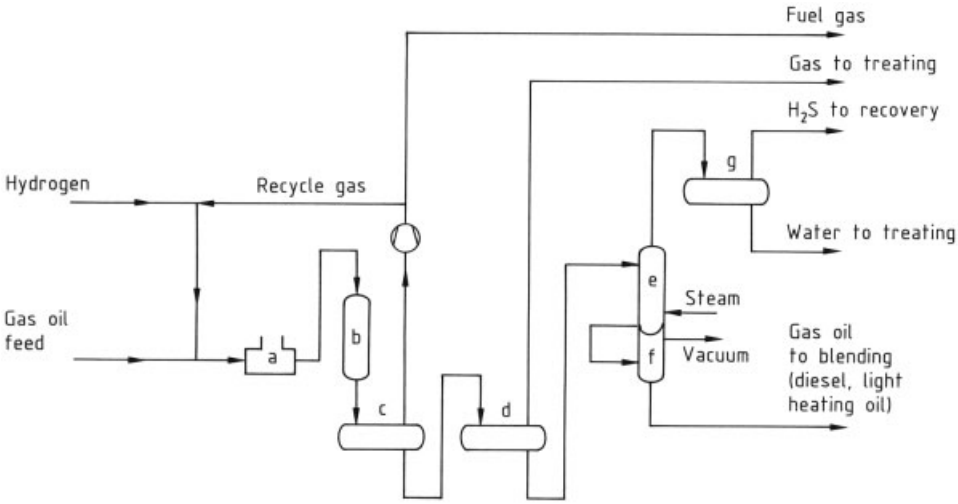

Figure 15. Gas oil hydrodesulfurizer
a) Process heater; b) Reactor; c) High-pressure separator; d) Low-pressure separator; e) Gas oil stripper; f) Gas oil dryer; g) Stripper overhead drum

Table 4. Catalyst types and main operating criteria of distillate hydrotreating processes [18]

Process	Catalyst type	H_2 partial pressure, MPa	LHSV,[a] vol/vol
HT[b] of naphthas	Co – Mo	0.5 – 0.8	5 – 8
HT[b] of kerosene	Co – Mo	0.8 – 1.2	4 – 6
HDS[c] of gas oils	Co – Mo	1.0 – 2.0	4 – 6
HDS[c] of VGO[d]	Co – Mo	2.5 – 4.0	1 – 3
HDS[c]–HDN[e] of naphthas	Ni – Mo	2.0 – 3.0	2 – 3
HDS[c] – HDN[e] of gas oils	Ni – Mo	2.0 – 3.0	1.5 – 2.5
Smoke point improvement of kerosene			
Moderate	Ni – W	2.0 – 3.0	1 – 3
High	Ni – W	3.0 – 4.0	1 – 1.5
HDS[c] of unsaturated naphthas and gas oils	Ni – Mo	2.0 – 3.0	1.5 – 3
(coker, visbreaker, FCC)	Ni – W		

[a] LHSV = liquid hourly space velocity (volume of product per volume of catalyst).
[b] HT = hydrotreating.
[c] HDS = hydrodesulfurization.
[d] VGO = vacuum gas oil.
[e] HDN = hydrodenitrification.

A large variety of commercial distillate hydrotreating processes are described in [10, 11], e.g., *for straight-run distillates*

Distillate HDS & SPI (IFP)
Distillate HT/Gulfining (Air Products – Gulf) [12]
Hydrofining (BP) [13]
Hydrofining (Exxon) [14]
Smoke point improvement (Shell)
Thermal naphtha HT (ARCO)

for vacuum distillates (FCC feedstocks)

HGO – HDS (ARCO)
GO-Fining (Exxon) [15]

for highly unsaturated distillates

DPG – HT (Lummus Crest)
Pyrolysis distillate hydrogenation (IFP) [16]
Pyrolysis gasoline hydrogenation (BASF – VEBA)
Hydroisomerization/HPN IVB (Engelhard) [17].

Hydrotreating (hydrodesulfurization, hydrodemetallization) processes for residues are dealt with in Section 3.6.3.1.

3.3.2. Pyrolysis Gasoline Hydrotreating

This widely applied process can be used in a single-stage and a two-stage version, depending on the destination of the products. The feedstock is usually a pyrolysis gasoline from olefin plants (byproduct of naphtha steam cracking).

In the *first stage* the unsaturated and aromatics-rich material is selectively hydrotreated so that the conjugated double bonds of the diolefins and styrene are hydrogenated to produce a stable gasoline blending component. The main stability criteria are a reduction

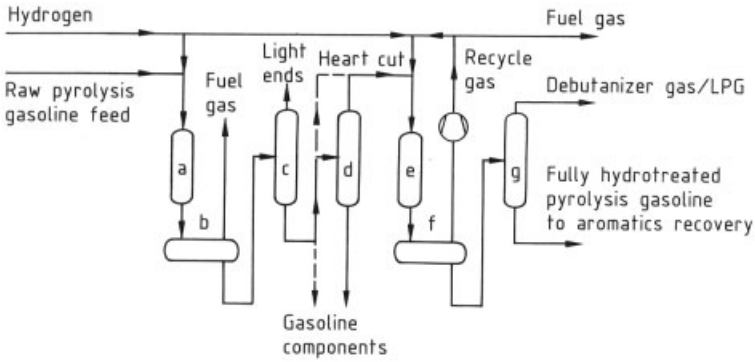

Figure 16. Two-stage pyrolysis gasoline hydrotreater
a) First stage reactor; b) First stage separator; c) Depentanizer; d) Gasoline (heart cut) column; e) Second stage reactor; f) Second stage separator; g) Debutanizer

of the diene value (maleic anhydride value) from a range of 30 – 70 to 1 – 2, accompanied by a reduction of the bromine number from a range of 50 – 90 to 30 – 60. The induction period and potential gum content are also greatly reduced.

In the *second stage*, the first stage product or a fraction thereof ("heart cut") is hydrotreated further to achieve complete hydrogenation of the olefins and to accomplish desulfurization. Thus, an aromatics-rich product is yielded that is suitable for BTX recovery.

The main quality characteristics of the second stage are a reduction of the bromine number to < 1 and a sulfur removal from 1000 – 5000 ppm to < 1 ppm.

Figure 16 shows a typical two-stage unit.

3.4. Catalytic Reforming

3.4.1. Introduction

In the reforming processes, gasoline fractions (naphthas) with a low octane number are converted into a high-octane reformate, which is a major blending product for motor gasolines (→ Automotive Fuels, Chap. 4.). Valuable byproducts are hydrogen and liquefied petroleum gas.

The naphtha feedstocks for reformers can be of straight-run or cracked origin and usually have octane numbers of only 35 – 65. Therefore these products are unsuitable for direct gasoline blending. In addition, the reformer feedstocks must be pretreated to adjust them to the required specifications (e.g., sulfur content, see Section 3.3).

With the development and introduction of noble metal catalysts at the beginning of the 1950s, the catalytic process displaced other (mainly thermal) reforming processes, because the catalytic conversion produces higher liquid yields – together with higher octane numbers – and hydrogen. Catalytic reforming is chiefly based on the catalytic conversion of normal paraffins and cycloparaffins into aromatics and isoparaffins.

The catalysts applied are either of the single metal platinum-on-alumina type, or bimetallic species where platinum is used in combination with a second metal, e.g., rhenium. The bimetallic catalysts have a better operational stability, but are more sensitive toward poisoning (e.g., by sulfur) or deficient regeneration. The catalyst metal is dispersed on the porous carrier material (alumina). The various commercial catalyst types contain 0.25 – 0.8 wt % platinum and up to 1 wt % of a halogen, usually chlorine. In these "bifunctional" catalysts the metal promotes hydrogenation and dehydrogenation reactions, while the chlorine catalyzes isomerization and cracking reactions.

The chemical reaction mechanisms of the catalytic reforming process can be characterized as follows:

Dehydrogenation of C_6-Naphthenes to Aromatics. Example:

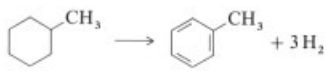

(1)

Conversion of methylcyclohexane into methylbenzene (toluene) with formation of 3 mol hydrogen per mole of converted hydrocarbon. The research octane numbers (RON) of methylcyclohexane and toluene are 75 and >100, respectively.

Dehydroisomerization of Alkyl-C_5-Naphthenes to Aromatics. Example:

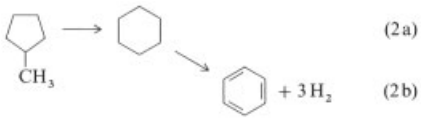

(2a)

(2b)

In the first step, methylcyclopentane is isomerized to cyclohexane (2a).

According to Equation (1), cyclohexane is then dehydrogenated in a second step to form benzene with formation of 3 mol hydrogen (2b).

The RONs of methylcyclopentane, cyclohexane, and benzene are 91, 83, and >100, respectively.

Dehydrocyclization of Paraffins to Aromatics. Example:

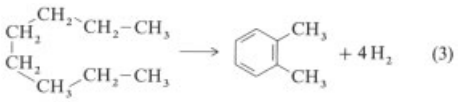

(3)

n-octane is converted to 1,2-dimethylbenzene (*o*-xylene) under formation of 4 mol hydrogen. The RONs of these components are 0 and >100, respectively.

Hydrocracking of Paraffins and Naphthenes to Smaller Paraffin Molecules. Example:

$$CH_3CH_2CH_2CH_2CH_2CH_2CH_3 + H_2 \longrightarrow CH_3CH_2CH_3 + CH_3CH_2CH_2CH_3 \quad (4)$$

n-heptane is split into propane and normal butane under consumption of 1 mol hydrogen per mole of paraffin (2 mol hydrogen per mole of naphthene). This reaction does not affect the RON.

Isomerization of n–Paraffins to Isoparaffins. Example:

$$CH_3CH_2CH_2CH_2CH_3 \longrightarrow CH_3CH_2CH(CH_3)CH_3 \quad (5)$$

Conversion of n–pentane into isopentane. The respective RONs are 62 and 92.

Hydrogenation of Unsaturated Hydrocarbons. This reaction is also important because unsaturated hydrocarbons can act as coke precursors. Coke deposits would deactivate the catalyst.

The dehydrogenation reactions (Eq. 1) and (Eq. 3) are both endothermic and require appropriate heat input (i.e., 32 and 38 kJ/mol, respectively, at reforming temperature level). However, because the reaction in Equation (3) is considerably slower than that for Equation (1), higher temperatures must be applied to achieve the desired reaction rate. The isomerization reactions (Eq. 2a) and (Eq. 5) are thermally almost neutral, whereas the hydrocracking reaction (Eq. 4) is exothermic (ca. −16 kJ/mol at reforming temperature).

Dehydrogenating cyclization of paraffins (Eq. 3) and dehydrogenation of naphthenes (Eqs. 1 and 2b) are the largest contributers to octane quality improvement. The thermodynamic equilibria of these reactions are shifted to the side of the reaction products at low hydrogen partial pressures and a reforming temperature level of ca. 500 °C. The ring-enlarging isomerization of alkyl-C_5-naphthenes to C_6-naphthenes (Eq. 2a) involves a decrease in the octane number, but this effect is more than compensated by the subsequent aromatization step (Eq. 2b). The theoretically attainable equilibrium compositions of the isomerization of *n*- to isoparaffins according to Equation (5) are not be reached at the usual reforming conditions.

Hydrocracking (Eq. 4) is favored at high temperatures and high hydrogen partial pressures. These reactions are usually undesirable in catalytic reforming, because they consume hydrogen and produce gaseous hydrocarbons from liquid ones, thereby decreasing the yield of liquid reformate. This influence is illustrated in Figure 17, where reformate yields are related to reactor pressure at various octane levels.

At a given pressure, the octane number can be increased by applying higher reactor temperatures; this, however, results in yield loss by increased hydrocracking.

As a result, catalytic reformers should generally be operated at low pressure to achieve high liquid yields. However, the hydrogen par-

tial pressure must be high enough to avoid the formation of unsaturated compounds which may polymerize and cause increased coke deposition on the catalyst and hence its deactivation.

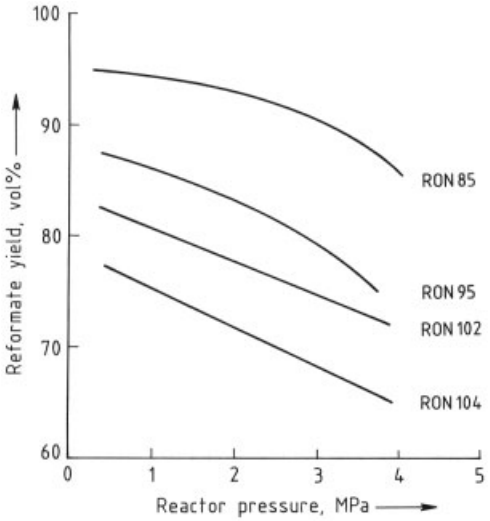

Figure 17. Reformate yield and reactor pressure at various octane levels

Catalytic reformers are usually available in three different versions, namely the semiregenerative, the fully regenerative, and the continuously regenerative reformer. Common to each process type is a reactor section and a product workup section. The fully and continuously regenerative process versions are also equipped with a catalyst regeneration section.

Various licensed reforming processes are described in [19] and [20]. Examples are Platforming (UOP), Magnaforming (Engelhard), Catalytic Reforming (IFP), Powerforming (Exxon), and Rheniforming (Chevron).

3.4.2. Semiregenerative Reformer

This conventional catalytic reforming process is shown in Figure 18. The plant has three or four reactors that are arranged side by side at ground level. After addition of hydrogen-rich recycle gas to the naphtha feed, the mixture is heated up in the charge – product heat exchanger (a) and brought to reaction temperature in the charge heater (b). The vaporized feed is then successively passed through the catalyst beds of the reactors (c, e, g). Between the reactors, makeup heat is provided in the intermediate heaters (d, f) to compensate for heat losses after each (endothermic) reaction step. The effluent from the last reactor is cooled in the heat exchanger and aftercooler (i). It is then separated in the product separator (h) into hydrogen-rich gas and liquid reaction product.

The H_2 gas is recompressed in the recycle gas compressor (k) and part is returned to the feed stream; the net H_2 "make gas" can be supplied to hydrogen-consuming units, such as hydrotreaters and hydrodesulfurizers.

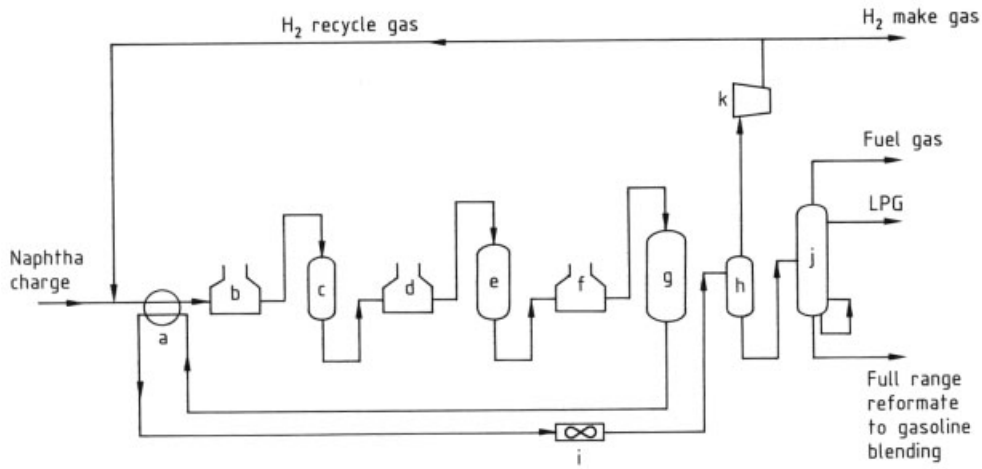

Figure 18. Semiregenerative catalytic reformer
a) Charge – product heat exchanger; b) Charge heater; c) First reactor; d) First intermediate heater; e) Second reactor; f) Second intermediate heater; g) Third reactor; h) Product separator; i) Product cooler; j) Stabilizer; k) Recycle gas compressor

The liquid reactor product is fractionated in a stabilizer column (j) to separate the C_4 and lighter hydrocarbons from the reformate. The LPG stream (C_3 and C_4) is usually split in the LPG recovery plant for propane and butane production; (→ Liquefied Petroleum Gas, Chap. 3.); the stabilized reformate can be used directly for gasoline blending or is fractionated further into light and heavy gasoline components before blending.

The catalyst is gradually deactivated during the operation period; this can be counteracted by stepwise increase in the reactor temperatures to keep the product quality at the desired level. Once the recommended ultimate temperature is reached, the plant has to be shut down and the catalyst is regenerated. This procedure is usually performed in three successive steps, i.e., carbon burn-off, metal reduction, and metal redispersion. The plant is then restarted for its next operation period. When the operation cycles between regenerations become unacceptably short, the catalyst inventory is replaced by a new batch. The spent catalyst batch is worked up for metal recovery (platinum, rhenium).

3.4.3. Fully Regenerative Reformer

The fully-regenerative, or cyclic, reforming processes have been devised to increase the on-stream time of the reforming plants by introducing the "swing reactor" principle. A typical example is the Power forming process [21]. In a system of four reforming reactors, three reactors are on-stream, in the fourth the catalyst is regenerated. After completion of the regeneration, this reactor is brought on-stream, while another reactor is switched to regeneration. The cyclic process requires a complex switch system with automatically operated valves. However, with this method a continuity of operation at high reformation severities can be maintained over a long period, with catalyst lifetimes of up to five years.

3.4.4. Continuously Regenerative Reformer

Special moving-bed techniques were developed for reforming units with continuous catalyst regeneration. The *IFP catalytic reforming process* which employs a conventional side-by-side arrangement of reactors uses a system of lift pots to transport the moving catalyst in portions from the bottom of a reactor to the top section of the next reactor [22, 23]. The catalyst portions obtained from the last reactor are lifted to the regenerator, where carbon burn-off and reconditioning of the catalyst take place. The catalyst is then reinjected into the first reactor bed.

A sophisticated catalyst moving-bed system was realized in the UOP Platforming process as illustrated in Figure 19 [24, 25]. The reactors are

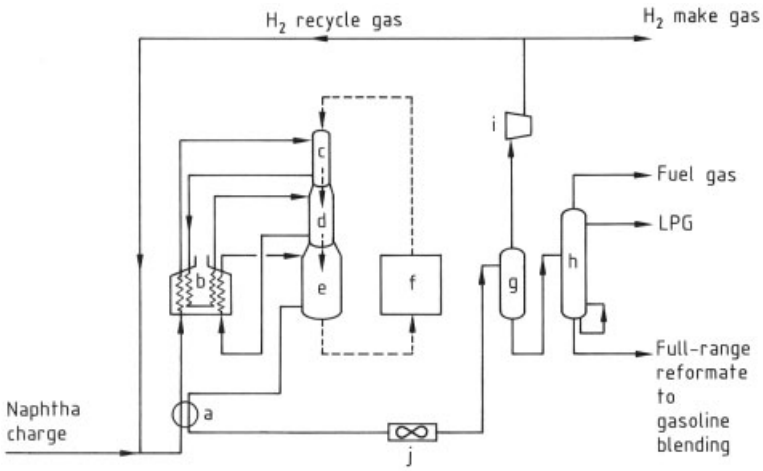

Figure 19. Continuously regenerative catalytic reformer (UOP Platforming)
a) Charge – product heat exchanger; b) Process furnace (charge heating cell, first intermediate heating cell, and second intermediate heating cell); c), d), e) Reforming reactors; f) Catalyst regeneration section; g) Reactor product separator; h) Stabilizer; i) Recycle gas compressor; j) Product cooler

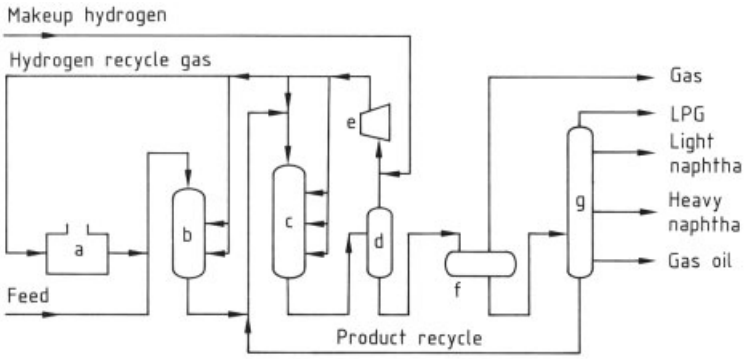

Figure 20. Two-stage hydrocracking [26, 27]
a) Hydrogen heater; b) First-stage reactor (hydrotreating); c) Second-stage reactor (hydrocracking); d) High-pressure separator; e) Hydrogen compressor; f) Low-pressure separator; g) Fractionator

arranged in a "stacked" construction. The catalyst trickles through the system from the top of the first reactor to the bottom of the third reactor, where it is collected and lifted in portions to the regenerator section.

"Sluices" are incorporated in the catalyst transport system, i.e., the spent catalyst from the third reactor is depressurized, freed from hydrogen, and then lifted in an inert gas stream to the regenerator; in turn, the regenerated and reconditioned catalyst has to be flushed free of combustion gases and pressurized with hydrogen before it is reinjected to the first reactor.

With continuous catalyst reactivation severe reforming conditions can be applied to achieve RON values of well above 100 in the reformate product. Moreover, operation at low reaction pressures of 7 – 10 bar results in favorable yield structures, i.e., high reformate and hydrogen quantities at low gas production.

3.5. Hydrocracking

The hydrocracking (HC) process was developed to produce high yields of distillates with better qualities than can be obtained by fluid-bed catalytic cracking. Hydrocracking feedstocks are typically flashed distillates from vacuum distillation, from cat and thermal cracking, or deasphalted oils. Operating conditions in the reactor section of hydrocrackers are usually about 400 °C and 8 – 15 MPa. The reactions take place on metal sulfide catalysts in the presence of hydrogen. Major advantages of the process include the full-scale production of high-quality products, i.e., no low-grade fuel oil remains as residue, and the high flexibility towards product yields, i.e., possible gasoline maximization, or gasoline – kerosene, or kerosene – gas oil maximization.

Modern hydrocrackers are units with two reaction stages – a hydrogenation step for desulfurization, denitrogenation, and desoxygenation using cobalt – molybdenum catalysts, and a hydrocracking step using nickel – tungsten catalysts. After fractionation of the liquid reactor product into gas, naphthas, and middle distillate, the bottom product of the fractionator is fed to the hydrocracking reactor. A typical example of a two-stage hydrocracker is shown in Figure 20. Hydrocracking processes for treatment of residues ("hydroconversion" processes) are described in Section 3.6.3.2.

3.6. Residue Conversion Processes

3.6.1. Introduction

In the previous sections, the standard processes of refinery technology for refining of distillates were described. These are used to improve the product yields recovered from the crude as well as the quality of the individual products.

These steps include catalytic cracking and distillate hydroprocessing, i.e., processes whose main purpose is to increase the yield of light products from heavier crude fractions. These processes were developed and applied mainly in the United States, where the goal had been to achieve a high yield of motor gasoline as the

main product of crude oil processing. In Europe and most other consumer areas, however, the objective was a high yield of diesel fuels, and light and heavy fuel oils.

In addition to these distillate conversion processes, the conversion of residues and heavy crudes in the modern refinery is growing in importance as a comprehensive process step for several reasons [28, 29]:

1) Since the drastic oil price increase following the first oil crisis in 1973 (from ca. $ 2/bbl to more than $ 40/bbl with fluctuating but in the long term rising trends) there is a growing economic need to recover more light, high-value products from the crude oil barrel at the expense of the residue.
2) In many countries, heavy fuel oil has been substituted by other fuels, above all natural gas, because of high sulfur and heavy metal content of residual oils.
3) The oversupply of residues also increases when the crude to be processed becomes heavier, i.e., richer in residue over the medium and long term (African and North Sea oils contain ca. 10 wt % vacuum residue, oils from the Persian Gulf ca. 30 wt %, and Venezuelan heavy oils still higher proportions of vacuum residues), see Table 5.
4) Finally, even the heavy crudes can no longer be processed in a normal refinery without additional treatment because of the high proportions of impurities.

Table 5. Residue contents of typical crudes (in wt %)

	Atmospheric residue	Vacuum residue
Nigeria	34.0	8.2
Algeria	30.9	8.3
North Sea	39.7	13.1
Mexican Isthmus	47.6	19.2
Arabian light	47.6	19.2
Russia	40.9	17.5
Kuwait	53.5	27.2
Arabian heavy	56.6	31.9
Venezuelan heavy	75.6	43.0

Important criteria to be considered when selecting suitable heavy oil upgrading technologies are the range of crudes (i.e., the amount and quality of the residue to be processed), the product requirements (motor gasoline, diesel fuel, heating oil), and the flexibility of the refinery (existing structure, changing feed supply, and product demand).

The feedstocks for residue conversion plants are atmospheric and vacuum residues, heavy oils, bitumen, tar sands, and shale oils. The products or objectives of the conversion vary according to the starting material and the method of conversion. They range from processing a low-sulfur heavy fuel oil where only low yields of lighter products are aimed at, via maximization of the distillate yield to complete conversion of the residue. In the longer term the production of "synthetic crudes" (syncrudes) from tar sands and shale oils will gain in importance (Fig. 21).

3.6.2. Fundamentals of Residue Conversion and Process Options [30, 31]

The C:H ratio in the crude oil fractions increases with increasing molecular weight, i.e., the hydrogen deficiency of the residue (in comparison with lighter hydrocarbons) increases in the sequence atmospheric residue < vacuum residue < visbreaker residue < coker residue < coke.

To obtain lighter products that are richer in hydrogen, two options exist for further processing which lead to a change in molecular structure: (1) hydrogen addition ("H-in") processes (Section 3.6.3); (2) carbon rejection ("C-out") processes (Section 3.6.4).

Additional processes for particular cases are the extraction of asphaltenes and the oxidation to synthesis gas or hydrogen (see Section 3.6.5).

A combination of different process steps is often necessary because of special conditions in the plant and varying process solutions (see Section 3.6.6).

Ca. 570×10^6 t/a of throughput capacity is installed in residue conversion plants worldwide (status 1997/1998). Nearly 60 % of this capacity consist of thermal processes, the most important being delayed coking (→ Petroleum Coke) and visbreaking. The thermal processes still predominate but the proportion of hydrogenation processes (hydrorefining and hydrocracking) is increasing, particularly the processes with high conversion rates, whereas the share of catalytic cracking capacities for treating of residual oils remains relatively constant.

Table 6 shows the approximate residue conversion capacities of the various world regions

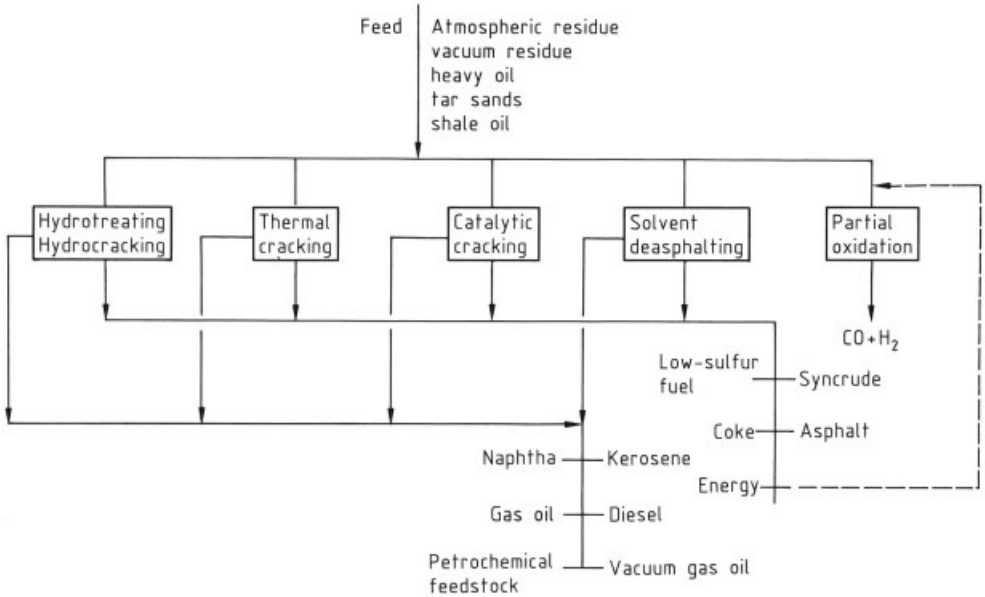

Figure 21. Residue conversion processes

(1997/1998). The resulting worldwide capacity for conversion of residual oils amounted to 14–15 % of total refinery capacity, However, the regional ratios vary widely between 4 % (Former Soviet Union, Africa, and Australia/Oceania) and 22 % (North America). In view of heavier crudes to be processed and consequently higher residue yields anticipated for the near future, a large expansion of heavy oil upgrading technologies will be necessary.

Residue conversion continues to be in a state of rapid development, thus the present review deals mainly with the problems of residue conversion and the currently best-known processes. The details of process engineering will be discussed only insofar as they differ from those of established processes. The best-known commercial processes may be summarized as follows:

Hydrorefining	Resid/HDS (Chevron)
	Resid/HDS (Gulf)
	Residfining (Exxon)
	Unicrackig (Union Oil)
	HDS (Shell)
	RCD Unibon (UOP)

Table 6. Estimated capacities (in 10^6 t/a) of residue conversion in 1997/1998 (derived from [2])

	North America	Latin America	Europe (excl. FSU[f])	Former Soviet Union	Asia (excl. FSU[f])	Africa, AUS/Oc[g].	Total
Refinery capacity	888	396	809	509	1126	190	3918
Thermal cracking[a]	105	49	92	20	64	5	335
Catalytic cracking[b]	36	7	18	3	20	2	86
Hydrorefining[c]	45	1	26		58	1	131
Hydrocracking[d]	9	2	1		9		21
Total[e]	195	59	137	23	151	8	573
% Residue conversion capacity of refinery capacity	21.9	14.9	16.9	4.5	13.4	4.2	14.6

[a] Thermal cracking of residues (visbreaking, coking).
[b] Catalytic cracking of residues (the capacity range for coprocessing of residual oils is estimated at 10–15 % of total cat. cracking capacity).
[c] Catalytic hydrofining of residues and heavy cat. cracking feedstocks.
[d] Catalytic hydrocracking of residual oils.
[e] Total residue conversion capacity.
[f] FSU = Former Soviet Union.
[g] AUS/Oc = Australia/Oceania.

Hydrocracking	H-Oil (HRI)
	LC-fining (Lummus)
	BOC Unibon (UOP)
	Combi-Cracking (VEBA)
	Canmet (Petro-Canada)
	Hycon (Shell)
Thermal cracking	Visbreaking
	Delayed coking
	Fluid coking (Exxon)
	Flexicoking (Exxon)
Catalytic cracking	Heavy oil cracking (Kellogg)
	Resid oil cracking (Total)
	Asphalt residue treatment (Kellogg)
	Reduced crude conversion (UOP)
Solvent deasphalting	ROSE (Kerr McGee)
	DEMEX (UOP)
Partial Oxidation	Texaco
	Shell

3.6.3. Hydrogen Addition ("H-in") Processes

As for hydrotreating of distillates, the term *hydrotreating* is also used for the catalytic hydrogen processing of residues if the main effect is an extensive removal of sulfur and nitrogen compounds (> 90 %) as well as of heavy metals, at a relatively low conversion (5 – 35 %). *Hydrocracking* means the operation under more severe conditions with the objective of high conversion rates of ca. 60 – 90 %.

The hydrogenation processes use high pressures (5 – 20 MPa) and are expensive; the principal cost factors are the technical equipment of the plants, the expensive hydrogen supply, and the catalysts. Estimated investment costs for hydrotreating are ca. $\$ 60 - 90 \times 10^6$ per 1×10^6 t residue per year; for hydrocracking ca. $\$ 150 \times 10^6$ per 1×10^6 t residue per year.

An advantage of H-in processes – especially of hydrocracking – is the recovery of pure, high-grade products in high yields.

3.6.3.1. Residue Hydrotreating (Demetallization, Desulfurization, Denitrification)

Residue hydrotreating processes have been developed by many oil companies and by well-known engineering firms on the basis of experience in the desulfurization of gasoline and middle distillates. They are also operated predominantly in fixed-bed reactors, but with lower specific catalyst loadings (0.5 – 2 m^3 feed material per cubic meter of catalyst). So-called guard chambers are frequently installed upstream as traps for metal impurities and solids to extend the on-stream-time of the actual HDS and HDN reactors. Depending on the residue quality the on-stream times would be only up to 6 – 12 months without preliminary purification. The individual processes differ mainly in the type and number of reactors (1 – 4 in series) and in catalyst types required for the special purpose. In most cases Co-, Ni-, and Mo-containing catalysts are applied, similar to those used for distillate hydrogenation. A typical residue HDS process is shown in Figure 22.

Some processes, such as the Shell HDS process, operate with a fluidized bed ("moving bed" or "bunker flow" reactors) in order to achieve continuous catalyst exchange and avoid fairly frequent plant shutdowns.

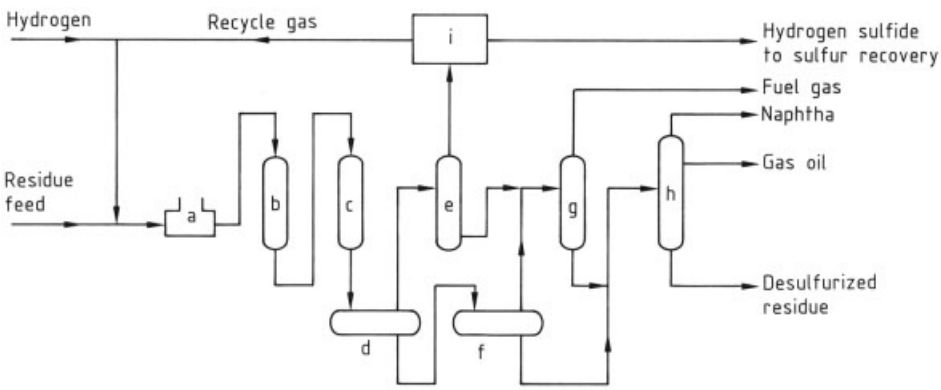

Figure 22. Residue hydrodesulfurizer
a) Process heater; b) Demetallization reactor; c) Desulfurization reactor; d) Hot high-pressure separator; e) Cold high-pressure separator; f) Hot low-pressure separator; g) Cold low-pressure separator; h) Product splitter (fractionator); i) Recycle gas treatment section

The objective of "mild" hydrotreating processes is to obtain low-sulfur residues which can be further processed catalytically (by FCC) or used as nonpolluting fuel oils. The main market for these processes has been in Japan, where "low-sulfur fuels" are mandatory.

3.6.3.2. Residue Hydrocracking (Hydroconversion)

Despite the advantage of a high yield of clean light distillates, residue hydrocracking is used worldwide in only a few plants, because of the high plant cost and the high hydrogen consumption of 300 – 400 m^3 per cubic meter of residue.

Fluidized-bed processes are especially suitable for hydrocracking because of the prevailing severe operating conditions in combination with rapid catalyst deactivation. An additional advantage of these processes is the low pressure drop in the reactor. The catalysts used are either expensive metal catalysts, which are continuously withdrawn and regenerated, or cheap, so-called "throwaway" catalysts, which are removed from the system in a once-through operation.

Processes Using Metal Catalysts. The current commercial processes are the *H-Oil process* developed by Hydrocarbon Research and Texaco and the Lummus *LC-fining process*. In these processes ca. 0.5 – 1 kg Co – Mo or Ni – Mo catalyst are used per ton of raw material in fluidized- or ebullated-bed reactors. The two processes are very similar in technology and process outline. The fuel oil – gas mixture passes upwards through the fluidized- or ebullated-bed reactors, of which up to 4 are connected in series according to the desired cracking severity. Suspended catalyst is added at the top of each reactor to maintain the desired catalyst activity and an equivalent amount of catalyst is withdrawn at the bottom. To maintain a flow rate which holds the catalyst in suspension in the fluidized or ebullated bed and which ensures uniform (isothermal) temperature control, so-called "ebullating pumps" circulate the reactor contents from top to bottom. In the first plants of the H-Oil process these pumps were arranged outside the reactor, where they could easily be cut off and were easier accessible for repairs. In LC-fining, more effective internal pumps are used to achieve an internal recycle. A schematic representation of an LC-fining reactor is shown in Figure 23.

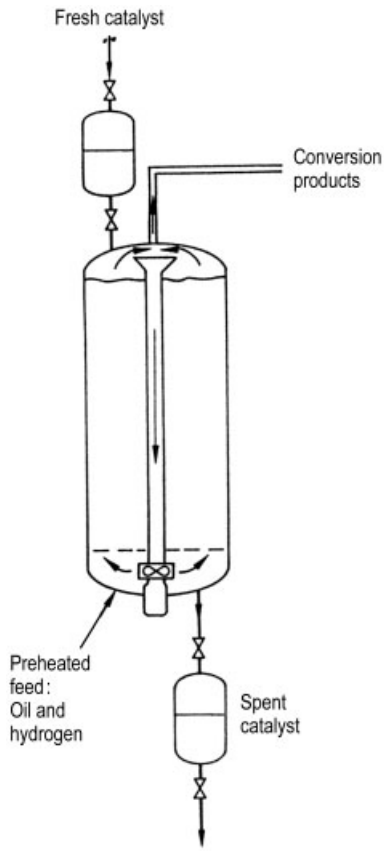

Figure 23. Schematic representation of an LC-fining reactor

Hydrogen is recirculated in the usual way, and the light conversion products escape rapidly overhead from the reactors, whereas the heavy material remains in the lower part so that the desired conversion can be obtained, if the residence time is sufficiently long.

Information on conversion, product yields, and degrees of desulfurization according to data from Lummus are given in Table 7 (see next page).

A modified hydroconversion process, the *Shell Hycon process*, went on stream in 1989 in a large plant with a throughput of 4000 t/d. This process involves the combination of two already proven process stages. In the first step, the metal content of the feed is reduced by 80 – 90 % with a SiO_2 catalyst in three bunker-flow

reactors. In the second step, the desulfurization and conversion is carried out in several fixed-bed reactors operated in series with a Ni – Mo catalyst on an Al_2O_3 support. The operating conditions (15 – 20 MPa and 400 °C) are similar to those for other hydrocracking processes. As the demetallization catalyst in the front-end reactors is deactivated within a short period of time, it is continuously replaced. The catalyst mass flow through the reactor is adjusted such that a shutdown cycle of ca. 2 years is obtained, which is normal for the tail-end fixed-bed reactors. The conversion is reported to be ca. 65 – 70 %, depending on residue quality. With improved catalysts at elevated temperature and with further developments in the reactor area (e.g., in material and catalyst replacement), higher conversion rates up to 90 % are possible even for extremely heavy residues [32].

Table 7. Product yields from hydrocracking at two conversion severities for an Arabian blend vacuum residue (process: LC-fining)

Product yield	Conversion, vol %	
	65	90
Gas	4.5	10.0
Gasoline	10.4	23.1
Gas oil	16.1	35.0
Vacuum gas oil	31.7	19.2
Residue	34.9	10.9
% HDS	83	89
% HDN	47	53
% Demetallization	93	96

BOC (Oil Conversion) is a fixed-bed process, licensed by UOP. The conversion is 60 – 70 % and desulfurization reaches 70 – 80 %. The reactor contains two catalyst beds separated by a specially developed "liquid – gas distributor". In the first bed the catalyst poisons and solids are removed; in the second bed the actual conversion step occurs.

Processes with Coal-Based Catalysts. A group of processes exist in which the residue is converted with pulverized coal as additive or doped with metals (preferably iron compounds) as catalyst. These processes are a further development of the coal liquefaction processes used before and during World War II, mainly in Germany. The amount of additive is between 0.5 and 5 wt %, according to the nature of the residue treated.

In *Veba-Combi-Cracking* (*VCC*), developed by Veba-Oel, liquid-phase cracking is combined with gas-phase hydrogenation of the conversion products formed. Both processes are performed at the same high-pressure level. In the first (liquid) stage (3 – 4 reactors), thermal cracking and hydrocracking reactions occur simultaneously in the fluidized bed at ca. 15 – 25 MPa and 430 – 470 °C. Oil residue, hydrogen, and catalyst are introduced at the bottom of the reactor, and a fluidized bed forms, the height of which depends on flow rate and desired conversion. The reactors have no internals and the temperature is controlled by quenching with recirculated gas. The second (gas) stage consists of a fixed-bed reactor of conventional construction with commercial HDS catalysts (Fig. 24).

The advantage of the direct combination of the two reaction stages in series is that the freshly cracked products are immediately hydrogenated before polymerization can occur to a significant extent. Furthermore, the specific energy con-

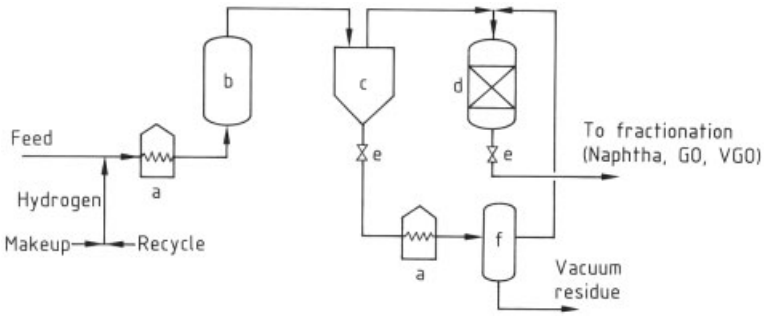

Figure 24. Schematic representation of Veba-Combi-Cracking
a) Heater; b) Liquid-phase reactor; c) Hot separator; d) Gas-phase reactor; e) Depressurizer; f) Vacuum tower

sumptions are lower because of the omission of an intermediate condensation and depressurization step followed in turn by energy-consuming pressurization.

The *Canmet process* of Petro-Canada and Partec Levalin is another process based on a coal catalyst. The process has been tested for the heavy Canadian oils and residues in several pilot and demonstration plants. Plant sizes are such as to perform up to ca. 1×10^6 t/a throughput in a single reactor. The hydrocracking reactions occur again in an up-flow reactor. A certain proportion of solid is withdrawn continuously from the bottom of the reactor. In contrast to the processes in which several reactors are connected in series, in this process the removal of sulfur- and nitrogen-containing compounds is less complete for a given conversion rate, so that larger reprocessing units are required [33].

3.6.4. Carbon Rejection ("C-out") Processes

The carbon rejection processes are designed as purely thermal or – increasingly – as catalytic processes. In these "C-out" processes, excess carbon is deposited as coke and the distillates, which still contain unsaturated hydrocarbons, must undergo further processing. Disadvantages of these coking processes include the unwanted amount of low-priced coke (→ Petroleum Coke) formed and the high gasification rate.

3.6.4.1. Thermal Processes (Visbreaking, Coking)

Visbreaking and coking (with delayed and fluid coking and Flexicoking as variants) are process types which have been known and proven in crude oil processing for many decades. The pioneering work in fluid and Flexicoking was performed by Exxon.

Because of the relatively low investment and operating costs of these processes, the oil companies are working on their further development. A particular aim is the minimization and quality improvement of the coke produced by (1) suitable pretreatment of the raw materials (optimization of the vacuum distillation, desulfurization), (2) minimization of recycle, and (3) optimization of operating parameters (pressure, temperature). *Visbreakers* are available from, e.g., Foster Wheeler, IFP, Kellogg, and Lummus Crest [34].

A typical visbreaker is shown in Figure 25. The vacuum residue feedstock is heated and cracked in a tube furnace (a) at ca. 460 – 480 °C. The furnace effluent is quenched with a cold oil stream, e.g., gas oil, which stops cracking reactions before the product enters the fractionator (b). The cracked product streams from the fractionator are routed to further treatment (e.g., gasoline hydrotreating, gas oil HDS). The visbroken residue usually has a lower viscosity and lower pour point than the feed stream.

Two-stage visbreaking is shown in Figure 26 (see next page). This process is applied for thermal cracking of atmospheric residues. The residue feed is cracked in the first stage residue furnace (a). The combined flashed distillates from the fractionator (b) and vacuum flasher (c) are subjected to a second cracking step in the distillate furnace (d). The temperature in this stage may be as high as 500 °C. This process maximizes the middle distillate (gas oil) yields, but a considerable portion of the gas oil is used as a blendstock to improve the viscosity of the cracked residue to match fuel oil specifications.

In the *coking* processes, the cracked residue is converted further to coke by allowing complete cracking within sufficient residence time [35].

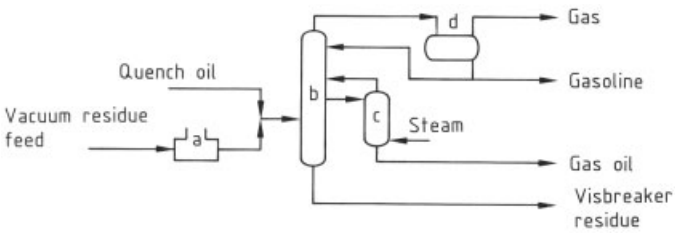

Figure 25. Visbreaker [34–36]
a) Cracking furnace; b) Fractionator; c) Gas oil stripper; d) Overhead accumulator

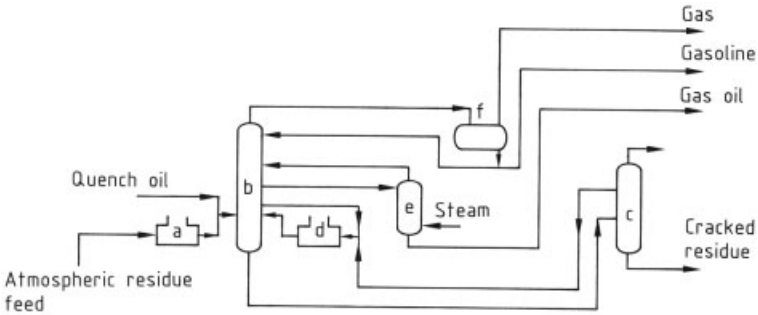

Figure 26. Two-stage visbreaking
a) Residue cracking furnace; b) Fractionator; c) Vacuum flasher; d) Distillate cracking furnace; e) Gas oil stripper; f) Overhead accumulator

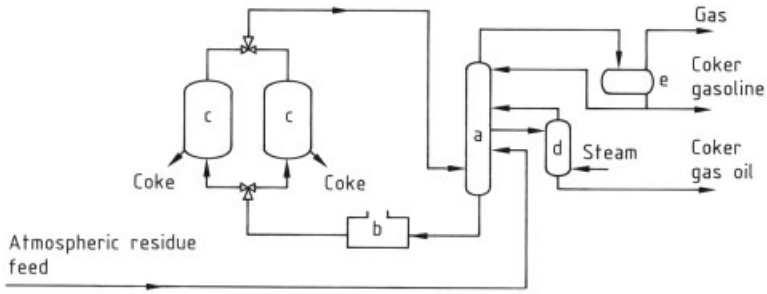

Figure 27. Delayed coking
a) Fractionator; b) Furnace; c) Coke drums; d) Gas oil stripper; e) Overhead accumulator

The principle of the *delayed coking process* is shown in Fig. 27. The atmospheric residue feed is introduced to the fractionator (a) where it condenses some of the cracked vapors. The fractionator bottom product is heated in a tube furnace (b) to ca. 490 °C, and the cracked furnace effluent flows through one of the coke drums (c) in which coke is being formed and deposited. The cracked vapors from the coke drum are separated further in the fractionator. In a 24 h cycle, one of the coke drums is in use while the other is emptied by means of a hydraulic coke removal procedure.

The introduction of the *fluid coking process* brought the advantage of continuous operation, thus avoiding alternate use of the coke drums [37]. The cracking reactions occur at 500 – 550 °C in the reactor in a fluid bed of coke particles into which the residue feed is injected. Coke fines are removed from the cracked vapors in cyclone separators before fractionation. The coke formed in the reactor flows continuously to the heater, where it is heated up to 600 – 650 °C by partial combustion in a fluid bed. The heated coke particles are returned to the reactor, from where the net coke production is withdrawn.

Table 8 shows a yield comparison of delayed versus fluid coking; delayed coking has higher gasoline and coke yields, whereas fluid coking maximizes coker gas oil. Delayed coking is the preferred process for production of high-quality petroleum coke grades.

Table 8. Product yields of delayed and fluid coking

Product	Delayed coking, wt %	Fluid coking, wt %
Gases (C_4 and lighter)	7.5	6.5
Coker gasoline	17	10
Coker gas oil	53.5	72
Coke, total	22	11.5

The *Flexicoking process*, licensed by Exxon, combines fluid coking with gasification [38]. This process offers virtually full conversion of the residues to be cracked. A flow scheme is shown in Figure 28. The coke from the heater (c) of the fluid coking section is fed to the gasifier (d) and converted with steam and air to coke gas. The hot coke gas is recycled to the heater

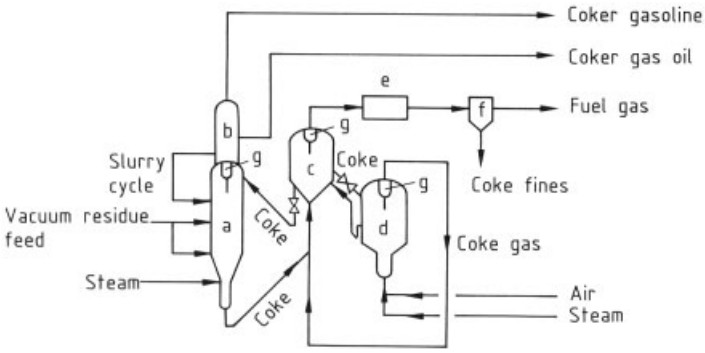

Figure 28. Flexicoking
a) Reactor; b) Fractionator; c) Heater; d) Gasifier; e) Heat recovery – gas cooling; f) Coke fines removal; g) Cyclone separators

(c) and its heat is transferred to the fluid coke bed. The gas mixture leaving the heater is then cooled and cleaned for use as a fuel gas.

3.6.4.2. Catalytic Processes (Residue Cat Cracking)

The residue cat cracking processes represent a further development of the well-known FCC technology which has been widely used for the cracking of vacuum gas oil (see Section 3.2). The primary objective is maximum gasoline production; the $C_3 - C_4$ gases which are formed in increased yield are also converted to high-grade alkylate gasoline. For processing residues in fluid cat crackers itis essential that the contents of asphaltenes and metals, which deactivate the catalyst, are kept under a certain level. The metals content in the feed should not exceed 10 – 30 ppm, otherwise pretreatment is necessary. Since the coke deposits on the catalyst in residue crackers are considerably greater than in vacuum gas oil crackers, research efforts are being undertaken mainly to develop new catalysts (use of zeolites, regeneration techniques) and to achieve a reduction of the residence time (riser know-how). At present only feedstocks with Conradson Carbon Residue < 15 wt % can be processed which means that only 40 – 50 % of the available crudes can be used [39].

Most large-scale plants process mainly atmospheric residues with pretreatment of the feedstock. The best-known commercially operated processes are (1) Heavy Oil Cracking (HOC, Kellogg, Phillips Petroleum); (2) Reduced Crude Conversion (RCC, UOP) with regeneration in two stages; (3) Asphalt Resid Treatment (ART, Kellogg); and (4) Residual Oil Cracking (ROC, Total). In accordance with the development of the oil market, the trend in all these processes is towards conversion of heavier (vacuum) residues [40].

3.6.5. Other Processes

3.6.5.1. Extraction of Asphaltenes

Deasphalting (→ Asphalt and Bitumen, Chap. 4.2.) is a process which was developed in the 1930s for the recovery of high-viscosity lubricating oil components. In this process the residue is separated with the aid of light hydrocarbons into two fractions: a deasphalted oil (DAO) and an asphalt-rich phase. In the original process, propane was mainly used as extractant and a DAO yield of 25 – 40 wt % was obtained. The process was not operated supercritically, i.e., the solvent and extract had to be separated by evaporation and condensation of the solvent with high energy consumption. The development of the supercritical extraction process began in the 1960s with the objectives to improve energy conservation, to achieve an increased DAO yield and to simultaneously reduce the amount of residue by employing a wider range of solvents.

The deasphalting processes are characterized by low investment cost, simple technical operation, and high flexibility with regard to extract yield and quality by choice of solvent.

Interest in deasphalting processes as an alternative to residue processing is increasing because the DAO can be used as a hydrocracker

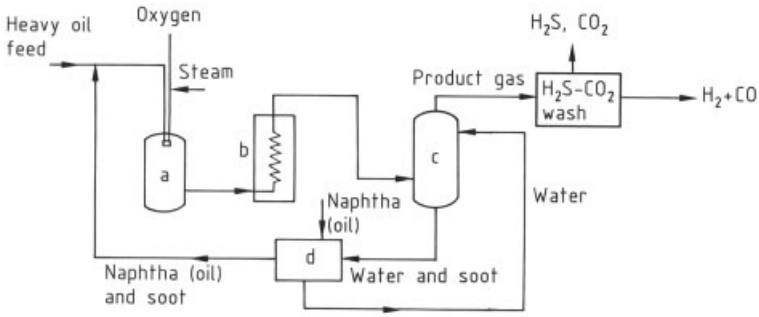

Figure 29. Schematic representation of partial oxidation
a) Reactor; b) Waste-heat boiler; c) Scrubber; d) Soot extraction

or FCC feed and because a large reduction of the unwanted residual products is achieved with DAO yields of 75 – 80 % currently feasible (see Table 9).

Table 9. Deasphalted oil yields obtained with various extractants

DAO	Feed	Extractant			
		Pentane	Butane	Isobutane	Propane
Yield, wt %		80.0	65.0	50.0	35.0
Sulfur content, wt %	5.4	4.7	4.2	3.7	3.2
Nitrogen content, wt %	0.41	0.3	0.2	0.2	0.1
Conradson Carbon Residue, wt %	25	17.0	9.7	5.5	3.1
V+Ni content, ppm	250	60	25	11	5

The extract yield which can be obtained is, however, limited by its metal content, which rises disproportionately in DAO yields of > ca. 50 %. Depending on refinery type and product requirement, the optimum must be sought between DAO yield and acceptable metal content or the costs of a demetallization stage.

Commercial plants are operated mainly in the United States and Central America for processing the heavy crude oils rich in metals and asphalt which are produced there. The best-known industrial processes are the Demex process of UOP and the ROSE (Residuum Oil Supercritical Extraction) process of Kerr – McGee.

3.6.5.2. Partial Oxidation

The partial oxidation of hydrocarbons to a mixture of $CO+H_2$ (synthesis gas) is a well-known process (→ Gas Production, Chap. 3.2.) and can be applied in residue processing within very wide limits. The noncatalytic reaction between hydrocarbons and oxygen in the presence of steam takes place at 1300 – 1500 °C in empty reactors lined with high-temperature-resistant brickwork or corundum refractory. Virtually all residues which are pumpable at the appropriate temperatures can be processed, and almost complete decomposition to $CO + H_2$ with a small (< 5 vol %) CO_2 fraction is achieved. The unburnt soot and ash fraction (≤ 5 wt %) is entrained in the synthesis gas and washed out with water in a downstream scrubber. Various possibilities exist for reprocessing this residual carbon, including complete soot recycle to the gasification reactor (Fig. 29).

The disadvantages of this universally applicable process are the relatively high oxygen costs and the production of a gas mixture of low calorific value which cannot be utilized in a standard refinery. If in the future methanol or higher alcohols become important as nonpolluting motor fuel components, the appropriate syntheses (from $CO+H_2$) could offer an interesting option. Similarly, the refineries' growing hydrogen requirement can be covered by this route even in the medium term, because pure H_2 can be produced by a simple CO-shift conversion step.

The partial oxidation process can be used in particular as a final treatment step in combination with the aforementioned processes with incomplete conversion. This also applies to the coking processes, since even the gasification of solid substances is technically possible after appropriate slurrying with oil or water.

Commercial partial oxidation processes are offered by Texaco and Shell. On a large scale, besides atmospheric and vacuum residues, also heavy thermal cracking residues (with high den-

sity and viscosity and C:H ratios of > 9.5) can be processed.

3.6.6. Process Combinations

Of the different process types of hydrogen addition and carbon rejection, which have been described in Sections 3.6.3 and 3.6.4, each has its advantages and disadvantages when used alone. A comparison of the residue conversion rates that can be achieved with the different processes is given in Figure 30.

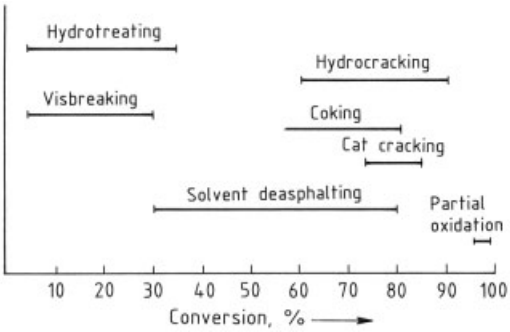

Figure 30. Conversion rates of various residue conversion processes

The "C-out" processes of thermal and catalytic cracking are proven processes with relatively low investment and operating costs, but they give a smaller distillate yield of poorer quality and a considerable amount of unwanted byproducts.

The"H-in" processes give much higher distillate yields, but they are more expensive. Since H-in processes require catalysts, additional costs occur with increasing concentrations of catalyst poisons (reduced on-stream times, higher catalyst consumption).

Therefore, it is usually necessary to combine different conversion processing steps. A great variety of possible combinations exist with regard to conversion, amount, and quality of the desired products, as well as to the existing refinery infrastructure. They include both combinations of different residue conversion processes and further conversion units for the higher-boiling vacuum gas oils fraction, which is formed in considerable amount during the residue conversion.

The medium- and long-term trend towards the use of heavy crudes which are rich in sulfur, nitrogen, and metals, in combination with environmental problems makes it increasingly necessary to add upstream hydroprocessing stages (principally hydrotreating, and in special cases hydrocracking), followed by deeper conversion [41].

The following combined steps in the residue line are used or discussed for the planning of conversion refineries:

hydroprocessing + coking
hydroprocessing + catalytic cracking
hydroprocessing + solvent deasphalting

and as a more far-reaching step:

hydroprocessing + coking / deasphalting / hydrocracking + partial oxidation

In most cases the hydrotreater is installed in the atmospheric residue line, so that the overhead and the bottom products from the downstream vacuum tower can be fed to the subsequent conversion units in sufficiently good quality.

In delayed coking, up to now the most frequently used residue conversion process, an upstream hydrotreater results in a lower yield of coke with improved quality and increased yields of lighter products (Table 10).

Table 10. Delayed coking (feedstock: Arabian heavy vacuum residue), yield improvement by upstream hydrotreating

	Feed	
	Untreated	Pretreated
C_4, wt %	9.5	12.0
C_5 < 200 °C ∗, wt %	19.0	22.1
200 – 360 °C ∗, wt %	31.3	29.9
360 – 500 °C ∗, wt %	19.7	40.0
Coke, wt %	32.5 (fuel)	12.0 (anode)
S content, wt %	6.5	2.5
V content, ppm	300	75

∗ Boiling range of hydrocarbon fraction.

Hydrotreating of the vacuum residue alone can be considered for crudes which give a vacuum gas oil of sufficient quality to be used without pretreatment in an FCC plant or a hydrocracking process. In this case, the hydrotreater can be smaller and the yields of liquid products from the downstream coker can be increased further.

The hydrotreating – deasphalting combination likewise leads to an increased yield of liquid products, because the proportion of a DAO with reduced sulfur and metal contents, which is suitable for further processing, can be increased by up to 30 % after pretreatment of the residue in the downstream extraction step.

Partial oxidation as the possible final step in a complete residue conversion is described in Section 3.6.5.2 (Fig. 29). An important advantage of gasification is that it solves the environmental problem of sulfur, which concentrates in all residues and in coke to such an extent that these residual materials are unsuitable for combustion. In partial oxidation the sulfur compounds in the feed materials are completely converted to H_2S which can be processed to sulfur in a Claus plant.

The combinations of residue conversion processes given as examples here represent only some of the possible and useful variations. The anticipated increasing complexity of refinery processing and the more difficult technical problems to be solved, will result in a steep rise in investment and operating costs. Increasing energy prices, an increasing shortage of crude oil and its poorer quality, as well as growing environmental problems worldwide, mean that the residue problem will in the future be one of the most important tasks for crude oil processing.

3.7. Gasoline Upgrading Processes

3.7.1. Alkylation

The alkylation process is based on the reaction of isoparaffins (e.g., isobutane) with olefins (e.g., propene, butenes, pentenes) to form higher molecular mass branched paraffins in the presence of a strong acid, e.g., sulfuric acid. The alkylation product, or alkylate, is obtained as a mixture of isoparaffins with high research octane numbers, making the product an excellent blending component for motor gasolines.

A typical example of alkylation is the reaction of isobutane (1-methylpropane) with 1-butene to form isooctane (2,2,4-trimethylpentane):

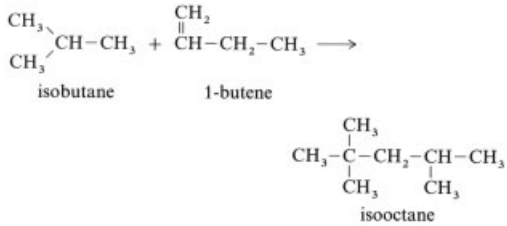

Alkylation reactions are catalyzed by strong acids. The first generation of commercial units applied sulfuric acid as catalyst. During World War II alkylation plants with large capacities were installed to satisfy the drastic increase in the demand for high-octane gasoline. Alkylation units could easily be installed in refineries where olefins were readily available from catalytic cracking units. Isobutane was supplied by LPG streams, e.g., from reformers, and by C_4-isomerization. The newer generation of alkylation units used hydrogen fluoride (HF) as catalyst, which has a better regenerability than sulfuric acid.

The most commonly applied HF alkylation processes were developed by Phillips Petroleum [42] and UOP [43, 63]. A scheme of the latter process is shown in Figure 31. The olefin feed and recycle isobutane are contacted with HF in the reactor (a). The reaction product is separated in the settler (b), from where HF is recycled

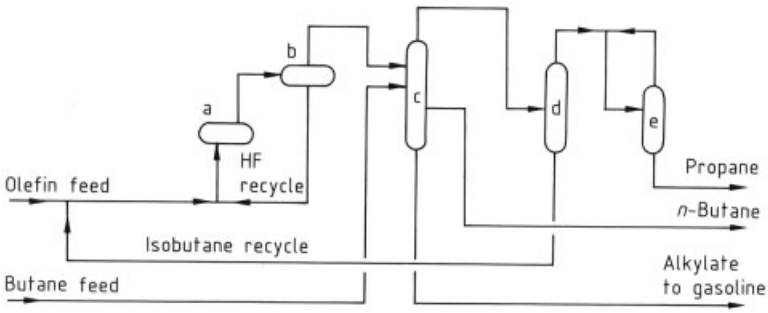

Figure 31. HF alkylation
a) Reactor; b) Settler; c) Isostripper; d) Depropanizer; e) HF stripper

to the reactor, while the alkylation product is fractionated in the isostripper (c) together with mixed butane feed. The isostripper overhead product is further separated in the depropanizer (d), from where the isobutane is recycled to the feed. The propane product is freed from entrained catalyst in the HF stripper (e). *n*-Butane is drawn off as a byproduct from the isostripper. The bottom product of the isostripper is alkylate which is supplied to gasoline blending.

3.7.2. Polymerization

The polymerization process to produce additional gasoline is used mostly in refineries where excess amounts of olefins, e.g., propene and butenes from catalytic crackers, are available. The most widely applied process is the UOP polymerization, using phosphoric acid as catalyst. The C_3/C_4-olefin feed is contacted with the acid in a reactor. After separation and stabilization, the polymerate is obtained and used as a gasoline component. Another polymerization development is the IFP Selectopol process [44, 45].

The more recent IFP Dimersol process [46, 47] is a dimerization process for olefins. Three types of Dimersol processes are now available for producing gasoline components from olefins [64]:

Type G	for high-octane gasoline from propene
Type X	for octenes from butenes, or for hexenes – heptenes – octenes mixtures by co-dimerization of propene and butenes.
Type E	for gasoline from lean ethylene streams

A simplified scheme of the G-Type Dimersol version is shown in Figure 32. The feed is contacted with the liquid catalyst in the reactor. After separation from the catalyst, the reaction product is debutanized in a stabilizer column.

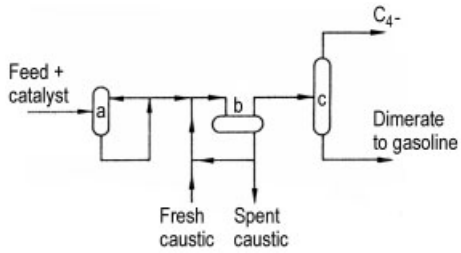

Figure 32. Dimersol process
a) Dimerization reactor; b) Catalyst recovery; c) Stabilizer column

3.7.3. Isomerization

Isomerization of low molecular mass *n*-paraffins has been used for a long time, particularly for isomerizing *n*-butane to isobutane in order to produce makeup feedstock for alkylation units. Isomerization gained importance after the introduction of lead-free gasolines, to compensate for the reduction in octane number. This was achieved by isomerizing C_5 and C_6 *n*-paraffin fractions to the corresponding isoparaffins and led to the development of various catalytic isomerization processes that used hydrogen and operated at moderate reaction conditions.

A typical example of this type of hydroisomerization is the Shell Hysomer process [48, 49] shown in Figure 33. The C_5/C_6-rich light gasoline feed is heated up together with hydrogen in the furnace (a) and isomerized in the reactor (b) containing a noble metal catalyst. The reaction

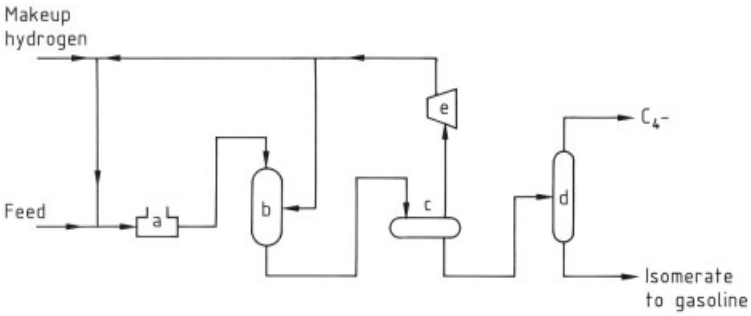

Figure 33. Hysomer process
a) Process heater; b) Isomerization reactor; c) Reactor product separator; d) Stabilizer column; e) Recycle gas compressor

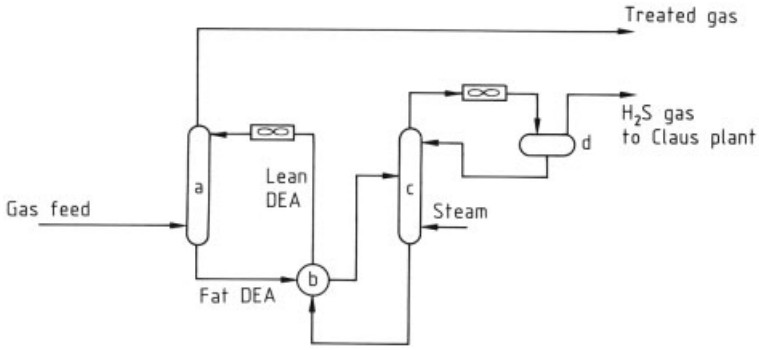

Figure 34. Alkanolamine (DEA) gas treatment
a) Absorber column; b) Lean – fat solvent heat exchanger; c) Regenerator; d) Overhead drum

product is separated, and the stabilized isomerate run down as blending component for motor gasoline.

To maximize production of C_5/C_6 isomers, the total isomerization process (TIP) can be applied. This is a combination of the IsoSiv (Union Carbide) and the Hysomer processes, providing an iso/*n*- paraffin separation in a molecular sieve unit and subsequent isomerization of the fraction containing the *n*-paraffins.

3.8. Other Processes

3.8.1. Gas Treating

Most refinery gas and LPG streams contain poisonous and/or corrosive contaminants, e.g., H_2S, mercaptans, COS, and CO_2, which have to be removed by suitable physical or chemical methods. The gases should be moisture-free, since water in hydrocarbon mixtures may form hydrates which can plug lines and equipment.

Commonly used gas treatment processes for H_2S and other contaminants are based on chemical absorption, using regenerable solvents such as alkanolamines, e.g., monoethanolamine (MEA), diethanolamine (DEA) or diisopropanolamine (DIPA). The Sulfinol process, using a mixed sulfolane – alkanolamine solvent, is also widely applied.

As a typical example the purification of a gas stream in a DEA treater is shown in Figure 34. The gas is contacted with the "lean" (pure) solvent in countercurrent flow in an absorber column (a), where H_2S and CO_2 are absorbed by the solvent.

The "fat" (loaded) solvent enters the regenerator (c), where the contaminants are stripped off by heating and steam injection. The H_2S-concentrated gas is then routed to the sulfur recovery plant.

Caustic treatment is the preferred process for the removal of mercaptans from LPG, using alkali hydroxide solutions (e.g., caustic soda) as solvent. Dryers containing caustic soda or other desiccants are applied to remove traces of water from the LPG.

3.8.2. Sulfur Recovery

The Claus process is the most widely used sulfur recovery method for sulfur recovery, and is indispensable for most refineries with gas streams containing H_2S. The process is based on burning part of the H_2S to form SO_2, which subsequently reacts with H_2S to elemental sulfur:

$$H_2S + \frac{3}{2}\,O_2 \longrightarrow SO_2 + H_2O \tag{6}$$

$$2\,H_2S + SO_2 \rightleftarrows 3\,S + 2\,H_2O \tag{7}$$

Owing to the reaction equilibrium, the recovery of sulfur is limited to 94 – 95 %.

The Claus off-gas (containing sulfur gases) is usually burnt in an incinerator to form SO_2. The sulfur recovery can be increased to above 99.5 % by applying a separate treatment process, e.g., the Shell Claus Off-gas Treating (SCOT), in which the SO_2 is catalytically converted to H_2S, and the H_2S is absorbed by means of solvent (DEA) circulation to and from the Claus unit.

4. Environmental Protection in Oil Refining

4.1. Introduction [50–52]

The natural product crude oil and the products obtained from it are generally easy to handle and with careful treatment – which is necessary because of their flammability – they cause little pollution to the environment. (The CO_2 emissions associated with all fossil fuels and the corresponding effects on the earth's climate cannot be considered in this context.)

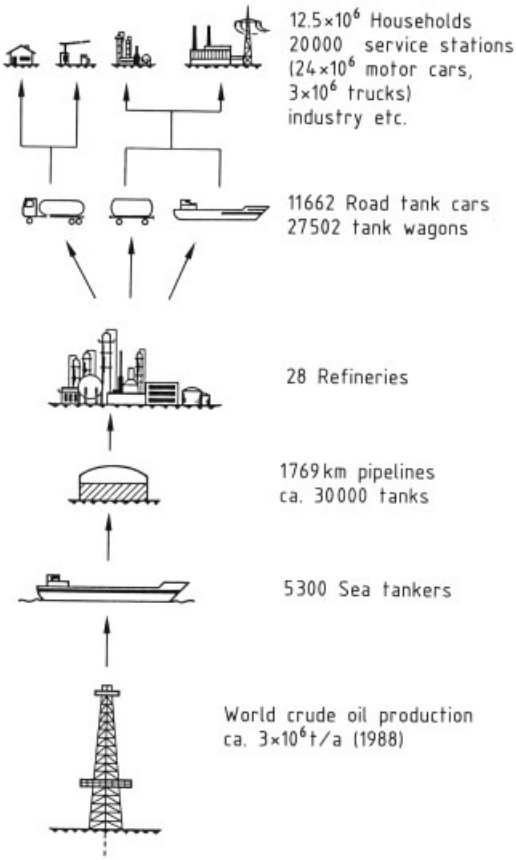

Figure 35. Petroleum's route from the well to the consumer shown as an example for Germany

Environmental problems arise, however, in dealing with crude oil and its products because the petroleum industry – like hardly any other branch of industry – maintains exploration, transport, and refining installations for crude oil that are scattered over the entire globe. The distribution and consumption of burning and transportation fuels are even more scattered (Fig. 35). Since the geographical location of crude oil fields does not coincide with that of the large consuming and refining regions, enormous transportation distances exist. The large amounts handled – the annual refining capacity in 1988 was ca. 3×10^9 t in more than 600 refineries worldwide – means that the industry needs to take precautions in all areas with regard to air emissions, water and soil pollution and, to a lesser degree, noise. The problems of introducing uniform standards are caused by legal standards which differ from country to country depending on the local concentrations of industrial areas with their corresponding environmental pollution and population density.

Because of the high cost of technical measures for environmental conservation, harmonization of legislation is absolutely necessary. As these costs amount to more than 25 % of the total processing costs in the refineries equal competition between different countries must be guaranteed (see Section 4.4).

In the following chapters, the environmental problems in crude oil refining, storage and loading, and in the application of the products are discussed separately, because the problems in the various fields differ.

4.2. Manufacturing Emissions

Although closed, gas-tight systems are generally used in refinery units, emissions into air and water cannot be completely avoided even with careful handling during refining and storage of the crude oil and its products. This is due to the management and control of the process and the properties of the products concerned.

Hydrocarbons are discharged into the air because of their high vapor pressure and they appear in refinery wastewater effluents because of their water solubility, which is, however, small. The carcinogenic aromatic hydrocarbons are particularly dangerous.

Further attention must be paid to the sulfur and nitrogen compounds originating from the heteroatomic compounds in the crude oil, both because of their smell and toxicity and because of the air pollution which arises in the form of

SO_2 and NO_x emissions during firing in process plants.

4.2.1. Hydrocarbons (HC) [53]

4.2.1.1. Hydrocarbons in Air

Hydrocarbon emissions can arise in *production plants* during normal operation (1) from leaking flanges in the pipework system; (2) at the seals of valves, pumps, and compressors; and (3) in the course of sampling. In the case of an accident, gases are led in closed systems to flares, collected as far as possible in gas recovery systems, compressed and returned to the process. The remainder of the gases is burnt in elevated or ground flares at efficiencies which can exceed 99 %. Liquid products are collected in closed slop systems that are equipped with pressure reservoirs and tanks and later returned to the production circuit.

In new plants measures to reduce emissions are taken during their construction, whereas continuous retrofitting is necessary in existing plants. Examples are flangeless piping, low-emission stuffing boxes, and seals, such as duplicated slide ring packings. For intensely odorous, poisonous, and carcinogenic substances more far-reaching measures are necessary (canned-motor pumps, special extraction devices etc.).

Most hydrocarbon emissions in processing occur in the *storage areas*, i.e., tank farms for crude, feedstocks, intermediate, and final products. Pressure – vacuum relief valves and floating cover tanks are generally used to reduce emissions. More recent developments are emission-free tank farms, where several fixed roof tanks fitted with internal floating roofs breathe into a closed system at one common gasometer, which normally absorbs all changes in the tank level. In the rare event of unusually large changes in the system (large amounts flowing into or out of storage, solar irradiation, heavy rain), the surplus quantity is burnt harmlessly in an associated flare or, in the case of a pressure drop, the system is topped up with inert gas.

Considerable amounts of hydrocarbons are also emitted in the *loading facilities*, especially in the loading of gasolines. Here low-emission or emission-free loading for transport by ship, rail, and road has become largely accepted. Various methods are used (Fig. 36):

1) Vapor recovery, where in a closed system the displaced gasoline vapors are either returned to the product tank or collected in a gasometer and used for the firing of process plants
2) Regenerative adsorption of the vapors on suitable adsorbents
3) Recovery of the products in liquid form after cooling or washing out the vapors

Care must be taken to prevent the formation of explosive gasoline – air mixtures which can occur in the road tanker to be loaded and in the adjacent piping and equipment.

This can be achieved by (1) keeping the concentrations outside the explosive range, (2) short transportation paths and exclusion of ignition sources, and (3) extremely strict control of the oxygen contents.

Dispensing of gasoline from the road tanker to the service station is also increasingly carried out with vapor recovery between the road tanker

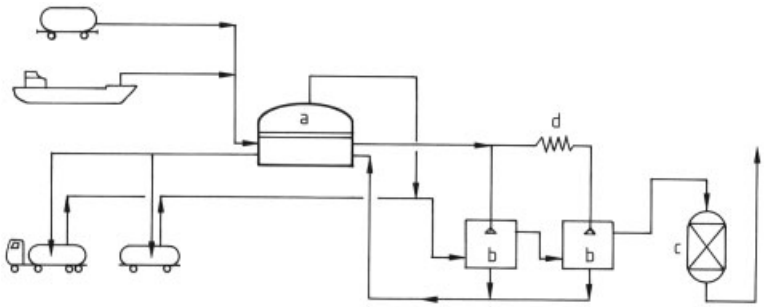

Figure 36. Emission-free loading and unloading of gasoline hydrocarbons with off-gas purification (refinery, intermediate storage)
a) Storage tank with floating roof; b) Exhaust gas washes (gasoline); c) Fine purification (adsorption); d) Low-temperature cooling (to − 40 °C)

and the installed storage tank of the service station. The possibilities of reducing emissions in the final step (filling the customer's motor vehicle) are described in Section 4.3.1.

The measures described reduced the atmospheric emissions from processing, storage, and distribution of refinery products in the mid 1980s in Western Europe to $< 8\,\%$ of the total man-made HC emissions. The portion from the refineries themselves was in turn only a quarter of this value. More recent investigations by CONCAWE show further improvements. Thus, relative to the crude oil throughput, refinery emission losses of ca. 0.05 wt % are expected, depending on the structure and size of the plant. For inland refineries in Germany, for example, values of 0.02 – 0.03 wt % have been published.

4.2.1.2. Hydrocarbons in Wastewater

Hydrocarbon-containing wastewater is unavoidably obtained at various points in the refinery because of the water content of the crude oil itself and because steam is employed in various processing steps. The total amount of hydrocarbon-containing wastewater in a normal refinery is of the order of 60 – 100 m^3/h.

This wastewater must be removed from the process units after separation of the oil phase, and led via a closed system to wastewater purification. Rainwater from exposed plant areas and from tank yards, and possibly contaminated cooling water from leaks or accidents must be treated in the same way. Considerable buffer volumes must be made available for the latter amounts of water which are formed discontinuously and sometimes in large quantities.

Treatment in the wastewater purification system is carried out stepwise by:

1) mechanical separation (sieves, filters, oil – water separators)
2) physicochemical purification (stripping, flocculation, flotation)
3) biological treatment

Biological wastewater treatment of hydrocarbons in refinery wastewaters is normally problem-free. However, the incoming streams and corresponding buffer volume must be continuously monitored to detect pollution by sulfur or nitrogen compounds and by oxygen-containing components such as phenols.

In many countries there is an increasing legal requirement for covered water treatment plants to avoid odor and for total nitrogen removal for further protection of surface waters (rivers, lakes etc.). The latter usually requires an additional purification stage with increased residence time.

After the biological stage, the water is clean and can be discharged into the receiving water. The average analyses of the wastewater from a modern refinery (in mg/L) are:

COD	60 – 100
BOD	5 – 15
Oil	0.5 – 2
Settleable solids	0.1 – 0.3
Phenols	0.1 – 0.2

As a result of increased wastewater purification, the amount of hydrocarbons discharged by the refineries in Western Europe was reduced by ca. 90 % in the period 1970 – 1985 for about the same amount of crude oil processed (Fig. 37) [54].

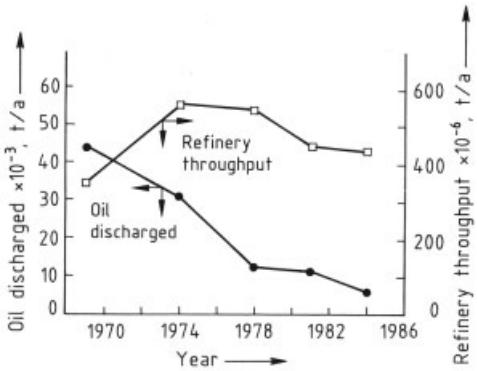

Figure 37. Development of the discharge of hydrocarbons with wastewater in Western Europe (reports from 87 refineries, ca. 80 %)

4.2.1.3. Hydrocarbons in Soil and Groundwater

Since hydrocarbons are water-soluble (even if only to a small degree), their penetration into the soil with possible contamination of the groundwater must be carefully avoided, whenever crude oil or any of its products are handled.

Transport from the oil terminal to the refinery is carried out almost exclusively in underground pipelines, which are also the safest means of

transport. The choice of high-grade steels as construction material, good insulation, cathodic corrosion protection, and continuous monitoring for leaks, including visual monitoring from aircraft or ground inspection, ensure a high level of safety. On difficult terrain and in areas of extreme temperature fluctuations, additional measures must be taken (pipeline compensation, elevated piles, intermediate tanks etc.).

The location of the refinery must be carefully selected with regard to possible dangers to drinking water. According to new legislation, all HC-handling units must be erected so as to prevent the discharge of spilled product to the underground, to adjacent streets, or canals; the storage tanks are placed in collection spaces which are made impermeable to oil using clay layers, plastic tilts, or concrete lining; in case of a leak they must be capable of receiving the entire tank contents.

If hydrocarbon contamination occurs in the soil, the affected portion of soil must be removed to prevent subsequent pollution of groundwater. If small amounts have escaped, the contaminated soil is usually combusted in incinerating plants. With larger amounts – and particularly if large areas are polluted with chemical residues and dangerous refuse – the damage must be treated in situ. Depending on the nature of the soil and the corresponding migration of the oil, this can be done by pumping off and purifying the contaminated water, possibly by additional injection of fresh water in adjacent wells.

Degradation of the oil by microorganisms is becoming increasingly important. This can also be done in situ or by excavating the soil and treating it externally [55].

Loading and storage of the products outside the refinery are subject to similar regulations. However, these are still very different in individual countries. As a result of the large number of external distribution depots for transport fuels and heating oils and of service stations, and because of the enormous number of oil-heated households, special care must be taken against overfilling and escape of products due to leaks. Corrosion-resistant and nonageing steels, plastic-lined steel tanks, and novel, glass-fiber-reinforced, plastic tanks are widely used. These tanks must also be constructed such that the whole volume can be collected if leakage occurs (doublewalled tanks).

In some countries with a population density and dense housing and industrial areas, the regulations which apply to pure hydrocarbons (i.e., products) are also applied to the handling of process water with much lower HC-concentrations.

4.2.2. Sulfur and Nitrogen Compounds

As a natural product, crude oil also contains heteroatomic compounds containing sulfur, nitrogen, and oxygen in addition to hydrocarbons. Whereas the nitrogen and oxygen contents are in the ppm range and play only a secondary role in atmospheric emissions, the sulfur content of the crudes can be as high as several percent. The distribution over the individual refinery fractions varies, but the content increases with increasing molecular size.

4.2.2.1. Sulfur Compounds

Sulfur and its compounds are catalyst poisons and adversely affect atmospheric emissions. Hydrogen sulfide, mercaptans, and disulfides are odor nuisances, and sulfur dioxide is formed during the combustion of crude oil products. Therefore, sulfur and its compounds must be removed or their contents reduced.

The light refinery products, liquefied petroleum gas (→ Liquefied Petroleum Gas, Chap. 2.) and gasoline, must be almost completely sulfur-free (→ Automotive Fuels, Chap. 8.); for diesel fuels (→ Automotive Fuels) and light heating oils (→ Heating Oil), a substantial sulfur reduction to 0.1 – 0.5 % is required by recent legislation.

Serious problems exist with heavy fuel oil (→ Heating Oil, Chap. 1.7., → Heating Oil), which is used almost exclusively as fuel in large industrial furnaces and power stations and leads to considerable SO_2 emissions. Many countries have established a maximum sulfur content in fuels of 1 – 2 wt %. This value can be reached without additional treatment only with a few, low-sulfur crude oils, whose supplies are limited (→ Heating Oil).

Sulfur is generally removed from distillates by hydrodesulfurization, whereby the chemically bound sulfur is converted to hydrogen sulfide. The H_2S is then removed in a gas scrubber,

converted to elemental sulfur in the downstream Claus process, and supplied to the chemical industry as a raw material.

The intense smell of the H_2S-containing gases to be processed and the high toxicity of H_2S, even at high dilution, means that precautions must be taken when handling. All sulfur-processing plants must be completely gastight: in the hazard zones, instruments and alarm devices are installed which automatically shut down the plants in case of danger.

Sulfur recovery in refineries has increased greatly as the result of the lower permitted sulfur content of the products. In conventional Claus plants with two reactors connected in series the conversion of the H_2S feed is ca. 95 %, a third reactor increases the conversion to 96 %. For a further reduction of the H_2S content an aftertreatment step (Scot process, Sulfren, see → Natural Gas) is required. Almost complete removal that is required in various countries (e.g., Germany 10 ppm H_2S in off-gas) can only be achieved with an additional high-temperature combustion step. Figure 38 compares the sulfur removal from the processed crude oil with the crude oil capacity in West German refineries for the period 1970 – 1988. The increase in sulfur removal resulted in a dramatic reduction of overall SO_2 emissions since most of the desulfurized hydrocarbons are used as fuels.

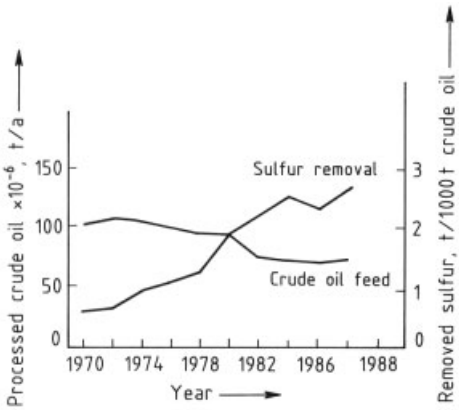

Figure 38. Sulfur removal during crude oil processing

Sulfur dioxide emissions in the refineries themselves are relatively small. Depending on the refinery complexity, only 4 – 6 % of the crude oil used is required for fueling in the process plants. About half of this amount is covered by the low-sulfur or sulfur-free gases obtained during processing, and the remainder is heavy fuel oil. With more severe requirements to lower SO_2 emissions, the refiner might be forced to use as fuel all of the LPG that is produced in the refinery.

Legislation differs greatly in the individual countries: the stipulated limitvalues can be related to a single process unit, a stack, or the refinery as a whole and can refer to the total sulfur discharged or the sulfur concentration in the flue gas. Combinations of several measures are also possible, and so-called "bubble concepts" are used, particularly in the United States. (Bubble concept means that for a given amount of emission from the total complex, different limit values or amounts discharged are allowed for each individual plant.)

The expensive flue gas desulfurization is normally uneconomic for firing installations of the size of those employed in a normal refinery. Therefore it is used only in special cases, e.g., if the refinery is connected with a large power station or if the sulfur content of the fuel is extremely high.

4.2.2.2. Nitrogen Compounds

The nitrogen content of most crude oils is relatively low, and in the distillates it is reduced during hydroprocessing to a few ppm. The residual nitrogen content causes no further difficulty in the use of the products.

Ammonia, which is formed in the hydrogenation steps and added in many refineries for process control in various process stages (pH adjustment), enters the wastewater and can be removed in the biological stage of wastewater purification (see Section 4.2.1.2).

For the NO_x problems in connection with the combustion of fuel oils, see → Heating Oils, Chap. 2.6.

4.2.3. Noise

In the past, noise pollution in the neighborhood of petroleum refineries played only a secondary role. Worldwide, pumps and compressors were installed not in closed buildings but in the open air, because of the flammability of petroleum

and its products. However, because the plants were smaller at that time, the noise emission of the units was relatively small. In addition, the sound radiation from furnaces and their burners produced relatively low sound levels in the refinery surroundings, because of the compact construction of the main process plants that were simultaneously screened by lownoise auxiliary installations (tank farms) at the periphery.

Problems arose, though, in densely populated areas with insufficient distances between industrial and residential areas. Most noise problems were caused by flare noise in the event of process disturbances.

Owing to growing environmental awareness in the public and a correspondingly greater strictness of legislation, further measures will have to be taken in the future to reduce sound emissions. This is also necessary for technical reasons, because the complex conversion refinery maintains more and larger units with greater furnaces and higher-powered gas compressors with high-speed steam turbines etc.

Important measures for noise reduction could cover the following items:

1) Low-noise burners and additional noise insulation on process heaters and piping.
2) Sound hoods on the drive motors and turbines of pumps and compressors. A complete "in-housing," however, may also give rise to safety problems, because fire fighting is more difficult and the danger of explosion is greater in closed rooms.
3) Sound insulation on control valves.
4) Low-noise flare stack tips for elevated flares or additional ground flares.

A noise reduction by 10 dB(A), i.e., a reduction of the observable noise level by about one-half, is possible according to the states of the art, though only with considerable expenditure.

4.3. Consumer Related Emissions

The emissions of SO_2 and NO_x from the combustion of refinery products outside the refinery are much greater than the emissions during refining itself. In addition, the HC emissions in the transportation field, both in refueling and in running motor vehicles, are also considerably greater than in the steps from production to the filling station (see Table 11).

Table 11. Hydrocarbon emissions in the petroleum field (Western Europe 1986)

Source of emission	Proportion	
	10^3 t	%
Crude oil production	20	0.4
Refinery, storage	170	4.1
Distribution	310	7.4
Vehicles		
Refueling	180	4.3
Exhaust gas	2500	59.7
Evaporation	1010	24.1
Total	4190	100.0

4.3.1. Transportation Fuels

When gasoline is combusted in the spark ignition (Otto) engine, varying amounts of CO, NO_x, unburnt hydrocarbons, and lead compounds are emitted with the exhaust gas in addition to the combustion products water and CO_2. The amounts differ according to the driving style, type of fuel, and engine construction. With the diesel engine, SO_2 and soot particles are also emitted, → Automotive Fuels, Chap. 3.2.9.

In addition to exhaust emissions, the emission of hydrocarbon vapor during refueling and running of motor vehicles is an important environmental pollution factor.

4.3.1.1. Motor Gasoline [56]

Different factors cause emissions of hydrocarbons from motor vehicles: firstly, the unburnt hydrocarbons in the motor vehicle exhaust, and secondly, the hydrocarbons emitted during refueling and running, because of their high vapor pressure. In Western Europe in 1986, the emissions (in 10^3 t/a) were distributed as follows:

Exhaust emissions	$2500 \triangleq 68\,\%$
Evaporative emissions	$1010 \triangleq 27\,\%$
Refueling losses	$180 \triangleq 5\,\%$

Specific measures are provided to reduce these different types of emission either in the motor vehicle or at the filling station.

For exhaust emissions, the use of catalysts in the exhaust gas line has generally become accepted (→ Automobile Exhaust Control).

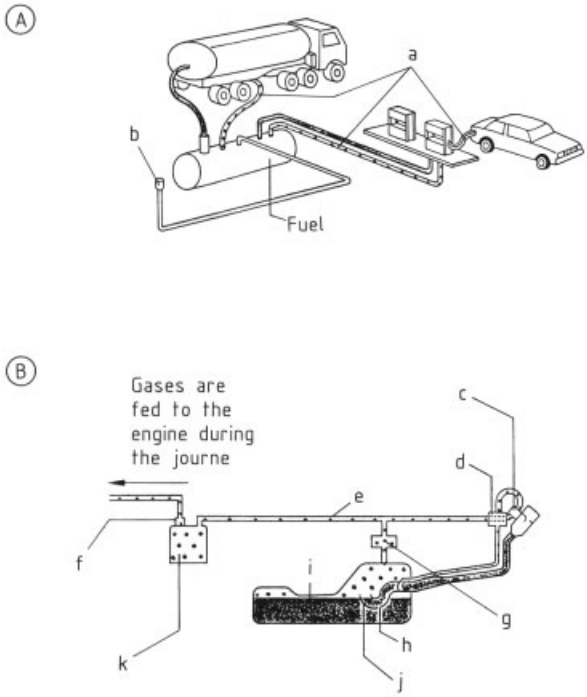

Figure 39. Reduction of hydrocarbon emissions
A) Vapor recovery at the service station; B) Large carbon filter in the motor vehicle
a) Gas displacement pipe; b) Vent; c) Gas venting valve actuated by filling nozzle; d) Gas – liquid separator; e) Gas line; f) Magnetic valve and regeneration control orifice; g) Standard gas vent and overturn protection; h) Outlet; i) Fuel tank; j) Liquid seal in filling tube (reduces escape of gases); k) Activated carbon filter with 4.5 L capacity (traps gases)

Efforts to reduce evaporative emissions and refueling losses are less clear-cut. Their proportion in the emissions from the motor vehicle sector is ca. one-third of the total output, corresponding to ca. 12 % of the total man-made emissions of organic substances.

Two solutions are under discussion (Fig. 39) [57]:

1) Vapor recovery at the service station, possibly supplemented by a smaller activated carbon canister in the motor vehicle.
2) A larger activated carbon canister fitted into the car, whose charge can adsorb all of the displaced hydrocarbons.

Vapor recovery at the service station is only a partial solution, because most of the emission is produced when the car is running and is not influenced by this measure. The advantage of vapor recovery is that it can be carried out at the service station within a rather short period of time. The system is already in use in the United States. The large activated carbon canister ("on-board system") is considered to be the better solution in the long term because of its greater effectiveness and lower cost. Technical trials have been carried out with various types of vehicles. Installation in new cars presents no problems; retrofitting to old vehicles will, however, only be possible in a few cases.

Owing to the wide use of motor fuels, particular attention must be paid to benzene emissions. The limit for the benzene content of gasoline (1 – 5 vol % is under discussion) presents the oil industry with problems, particularly in the production of unleaded gasoline, because the benzene content of various blending stocks can be considerably higher, depending on the quality or origin of the crude, up to 8 vol % for high severity reformate and 18 – 40 vol % for pyrolysis gasoline. Benzene extraction from HC-stocks does not solve this problem because of the cost and the question of reuse of the benzene recovered. Therefore, the evaporative losses must be minimized.

4.3.1.2. Diesel Fuel

The main exhaust gas problems of diesel vehicles concern the emissions of soot particles and SO_2. The NO_x and CO components are less important than in the spark ignition (Otto) engine. Demands are increasingly being made for a reduction of these emissions, because of the smell of the exhaust gases, and because the soot particles may have a carcinogenic effect. Emission reduction of diesel fuels is discussed in more detail in → Automotive Fuels, Chap. 3.2.9.

4.3.2. Burning Fuels

The largest proportion of refinery products worldwide is used as a fuel for domestic or industrial heating as well as for energy production. These are the fractions of LPG (→ Liquefied Petroleum Gas) the middle distillates, and heavy fuel oil (→ Heating Oil). Emissions of SO_2, NO_x, and CO that are associated with combustion, as well as emission of ash and unburnt carbon (in the form of soot) can be considerably influenced by selection and pretreatment of the products used and by the burner design. This is particularly true for the large heating installations operated with sulfur-rich heavy fuel oils.

Liquefied Petroleum Gas (→ Liquefied Petroleum Gas). The liquefied petroleum gases used as fuels are almost sulfur-free as a result of pretreatment in the refineries, and they burn without formation of soot. Their use presents no problem for the user.

Light Heating Oil (No. 2 Fuel) (→ Heating Oil, Chap. 1.). Middle distillates contain up to 1.5 wt % sulfur, depending on the origin of the crude. Most of the sulfur can be removed in the refinery by hydrosulfurization (see Section 4.2.2).

The permissible sulfur limits for light heating oil in most countries are < 0.5 wt %, because of its use in domestic heating. In the EU, the sulfur content of domestic heating oil is now limited to 0.2 wt % (→ Heating Oil, Chap. 1.7.). If lower sulfur contents are to be achieved, desulfurization costs will rise steeply with the current range of crudes, according to studies by Concawe [50] (Fig. 40).

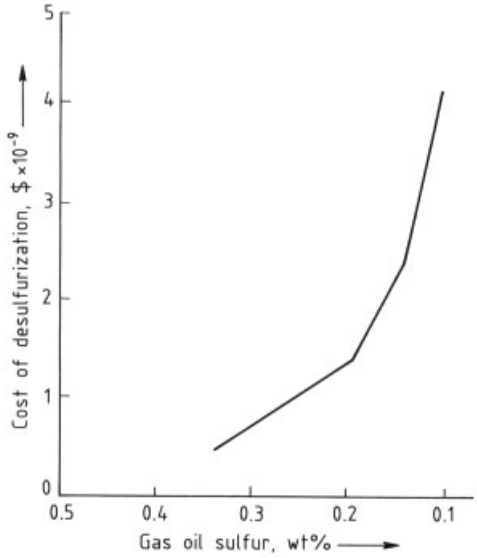

Figure 40. Investment to reduce gas oil sulfur in Western Europe

Heavy Fuel Oil (No. 6 Fuel). The residues from crude oil refining are used as heavy fuel oil for industrial heating and energy production. Sulfur dioxide, nitrogen oxides, CO, and particulate emissions must be considered for heavy fuel oils; for environmental protection measures in this field, see → Heating Oil, Chap. 2.

4.4. Cost of Environmental Conservation

Because the understanding of environmental problems, technical solutions, and legislation are still in rapid development, a view of the expenditure on environmental conservation can only be predicted. This estimation is necessary for an overall view of the problem, because the costs of environmental conservation have reached such a high level that they markedly influence the total processing costs. The great variety of possible solutions worldwide and the different legislation in the various countries and regions also have a corresponding impact on costs.

Besides the investment costs required for new plants (15 – 20 % of the total cost is used for environmental conservation) and the conversion of old plants, there are continuous operating costs

for energy, maintenance, personnel, etc. It is difficult to give current, generally applicable figures due to the increasing debate on the environment and the rapid development in this sector. Table 12 shows the growth of environmental conservation costs in Germany from 1977–1987.

Table 12. Investment and operating costs for environmental conservation in the Federal Republic of Germany in comparison with crude oil processing

	1977	1979	1981	1983	1985	1987
Crude oil processed, 10^6 t/a	99.5	109.6	98.9	90.2	87.3	81.4
Investment costs, 10^6 DM	294	127	130	305	103	221
Operating costs, 10^6 DM	826	894	1129	1212	1097	900

A study was set up in 1986 by the Commission of the European Communities to determine the environmental conservation costs for refineries in the EC member countries [59]. This considered the current (1985) and future (1993) operation of typical refineries and determined the costs of keeping to environmental regulations – at the refinery sites and with regard to product specifications. The current standards and regulations and those expected in future were taken into account. The most important measures in the processing field are:

1) Gas washing systems and Claus plants for sulfur recovery from H_2S followed by fine purification.
2) Closed systems for the discharge of gaseous (via a flare) and liquid hydrocarbons.
3) Floating roofs and floating covers in storage tanks for crude oil and products.
4) Hydrocarbon vapor recovery systems for the storage and loading of volatile products.
5) Collection spaces for escaping hydrocarbons and precipitation water in production plants, tank farms and loading plants. Drainage to wastewater purification plants.
6) Reduction of emissions (SO_2, NO_x) from the plant's firing installations.
7) Closed sampling systems, laboratory analysis, and on-line instruments for pollutant measurement in air and wastewater.

The following measures are for the production of less polluting products:

1) Hydrodesulfurization of gasoline and middle distillates.
2) Reforming and isomerization for the production of high-octane components as a basis for unleaded gasoline.
3) Synthesis of suitable components for unleaded gasoline (MTBE).
4) Conversion of heavy residues into light, clean products.
5) Separate storage and loading of unleaded types of transportation fuels.

In determining the measures required and the resulting costs, the study considered the environmental conservation laws applicable in each EC member country at the time concerned or scheduled up to 1993. The results are given in Table 13.

Table 13. Environmental costs of petroleum processing in the EC member states (in ECU/t)

Country	1985	1993 EC directives only	1993 EC+national regulations
Belgium	1.01	6.25	8.57
France	1.05	3.11	4.54
Germany	3.38	5.87	19.28
Italy	0.34	2.15	2.52
Netherlands	1.01	6.12	17.07
Spain	0.31	3.21	3.58
United Kingdom	0.28	6.64	7.01

They show that: (1) the environmental costs rose in all EC countries; (2) there are large differences in cost between the individual countries; (3) for comparable costs arising from the EC directives, the additional national legislations lead to different costs; (4) the burden of environmental costs is in some countries of the same order of magnitude as that of the other processing costs.

The figures given are more likely to be near the lower limit: the study did not, e.g., consider noise protection measures, soil clean-up or special regional regulations, because these can differ greatly according to the location of the refinery. Moreover, the conservation of groundwater and consideration of climatic changes are becoming increasingly important, and are continually leading to a tightening of the limits of the (current) standards.

The investigations carried out for the EC can be applied to all other regions of the world. Taking into account the high expenditure for minimizing potential dangers (flammability, explo-

sion risk, transport risks) and for protection of the plant workers, crude oil refining is burdened with extremely high additional costs related to product value and technical efforts.

5. Integrated Refinery Models

5.1. Trends of Refinery Structures

The drastic increase in crude oil prices in 1973 – 1974 and 1979 – 1980 as well as changes in the product demand pattern in major consumer regions, i.e., North America, Europe, and Japan, have led to new considerations in integrated refinery processing to achieve better utilization of residual materials, a maximization of distillate production, an improvement in process economics, and to meet current environmental standards.

These trends resulted in a worldwide restructuring of oil refining with the aim of increasing the yields of the high-value products – chiefly transportation fuels and industrial feedstocks – at the expense of low-value heavy fuel oils. Great efforts were made to change from the simple "hydroskimming" type refining scheme to more complex refinery structures with integrated residue conversion processes. A comparison of three conversion concepts versus hydroskimming is given by the following schemes which are based on Arabian light crude processing [60]. Examples of developments in thermal and catalytic cracking processes for heavy fuel oils and of developments in residue hydroprocessing, both in integrated refinery structures, are given in [61] and [62], respectively.

5.2. Hydroskimming Refinery

The process configuration, known as hydroskimming or "fuel oil refinery," is shown in Figure 41. The plant lineup consists of a crude distillation unit, hydrotreaters for gasoline and gas oil, a catalytic reformer, and a gas and sulfur workup section. It does not include a residue conversion unit. Atmospheric residue with a yield of 43.5 % of the crude represents by far the largest product portion, which has to be marketed as heavy fuel oil.

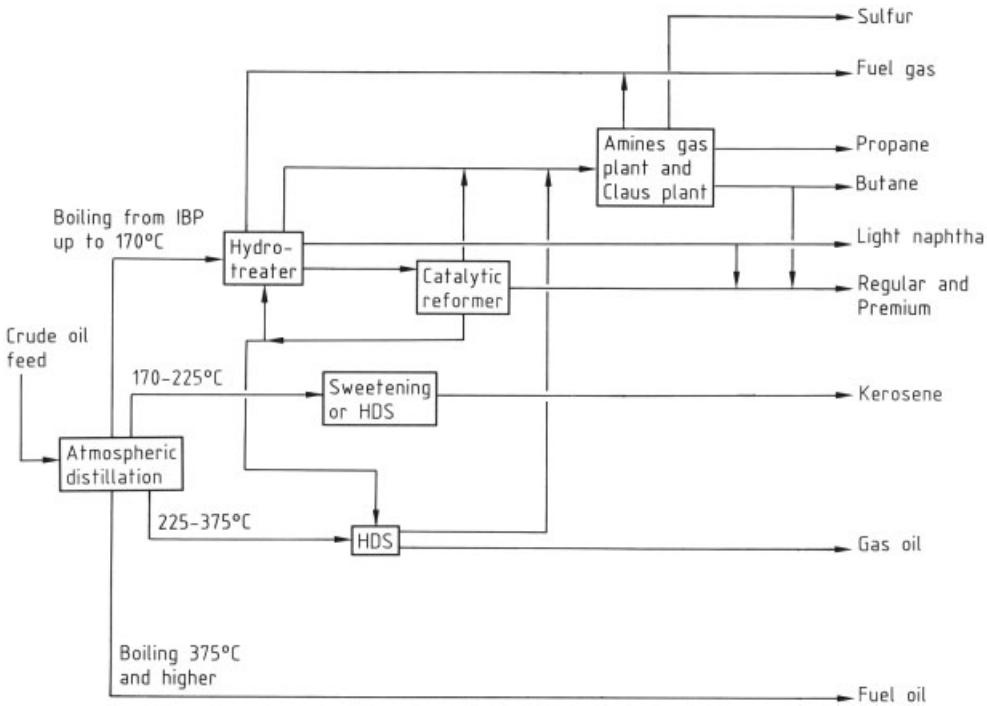

Figure 41. Hydroskimming

Table 14. Yield structures of refinery conversion schemes for Arabian light crude processing [60]

Product	Hydroskimming [a]	FCC – VB [b]	HC – SDA – FCC [c]	HC – Coking [d]
Intake, wt %				
Crude oil	100.0	100.0	100.0	100.0
Hydrogen			0.4	0.5
Output, wt %				
Fuel gas	1.8	3.4	3.1	4.7
LPG	1.6	3.0	2.9	2.9
Chemical feed	2.0			1.7
Gasoline	15.9	33.0	32.9	24.4
Kerosene	9.6	9.6	9.6	9.6
Gas oil	25.3	26.3	40.2	50.9
Fuel oil	43.5	22.7	9.2	
Coke		1.0 [e]	1.1 [e]	5.0 [f]
Sulfur	0.3	1.0	1.4	1.3

[a] Hydroskimming = refinery configuration without residue conversion processes.
[b] FCC – VB = conversion scheme with cat cracker and visbreaker.
[c] HC – SDA – FCC = conversion scheme with hydrocracker, solvent deasphalting, and cat cracker.
[d] HC – coking = conversion scheme with hydrocracker and coker.
[e] Coke burnt in FCC unit.
[f] Coke produced in coker unit.

The naphtha yield is 17.9 % of the crude (15.9 % gasoline and 2 % chemical feedstock), and the middle distillates amount to 34.9 % (9.6 % kerosene, 25.3 % gas oil). A comparison of the product yields of the hydroskimming refinery with the yield structures attainable with different conversion concepts is given in Table 14.

5.3. Conversion Concepts

5.3.1. Cat Cracking – Visbreaking Refinery

A typical refinery for maximum gasoline production using an FCC unit is shown in Figure 42. The FCC feedstock, vacuum gas oil, is produced by vacuum distillation and pretreated in a hydrodesulfurizer for sulfur removal and FCC yield optimization. In a visbreaker unit, the vacuum residue is thermally cracked to produce additional distillates. As can be seen from Table 14, the residue output from this refinery type is 22.7 % of the crude, i.e., only about half of that of the hydroskimming type. The gasoline yield amounts to 33 % and middle distillates make up 35.9 % of the crude.

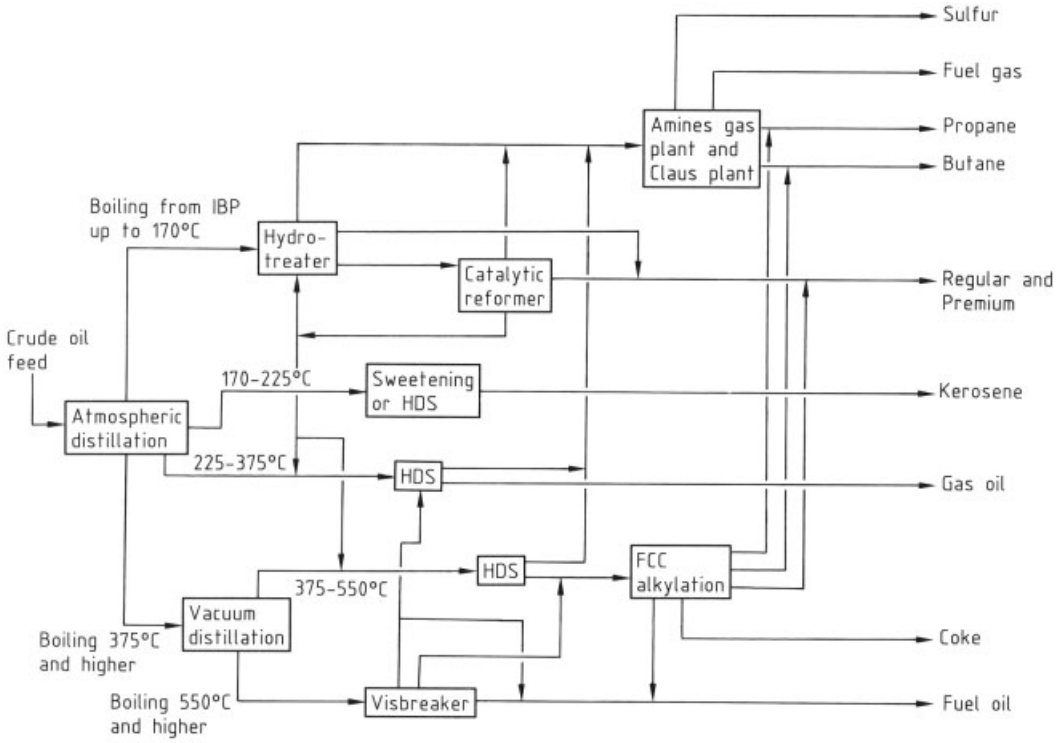

Figure 42. Catalytic cracking – visbreaking

5.3.2. Hydrocracking – Cat Cracking Refinery

A high-conversion refinery which combines fluid catalytic cracking and hydrocracking is shown in Figure 43. The HC feed is vacuum gas oil, whereas the FCC feed (deasphalted oil) is produced in a solvent deasphalting unit and desulfurized in a HDS unit. The asphalt product of the SDA unit can be used as heavy fuel oil; however, it must be blended with a "fuel thinner" (middle distillate) to meet fuel specifications. The hydrogen produced in the catalytic reformer is not sufficient for supplying the HC unit, therefore makeup hydrogen is produced in a steam reformer. As shown in Table 14, the net fuel oil portion is only 9.2 % of the crude. Gasoline and middle distillate yields are 32.9 % and 49.8 % respectively, i.e., an improvement in comparison with the two schemes previously described.

5.3.3. Hydrocracking – Coking Refinery

The hydrocracking – coking refinery is illustrated in Figure 44. In this configuration the production of middle distillates is maximized by using a hydrocracker and a coker. Vacuum residue is used as coker feed while the HC feed is a combination of vacuum gas oils and heavy coker distillates. As in the previous scheme, extra hydrogen has to be supplied from a steam reformer.

The only residual product is 5 % coke (Table 14) which can be sold to the chemical industry or for electrode manufacturing (→ Petroleum Coke). Since no heavy fuel oil is produced, this process configuration is often referred to as a "zero-fuel" refinery. The yield of middle distillates (kerosene and gas oil) is > 60 % and the naphtha yield (gasoline and chemical feedstock) is 26 %.

6. Corrosion and Materials

Prevention of corrosion in refining equipment is determined by material selection in plant construction and day-to-day operational control. As various corrosion criteria are dealt with elsewhere (→ Corrosion), only the main refining-related corrosion problems and the selection of corrosion-resistant materials are summarized as follows:

Corrosion problems are mainly caused by organic sulfur compounds and inorganic chlorides in crude oils. The danger of steel embrittlement by hydrogen in all processes involving hydrogen must also be taken into account.

Corrosion in distillation processes is caused by the following:

1) Formation of iron sulfide and subsequent corrosion by H_2S and mercaptans at elevated temperatures. Suitable materials for corrosion prevention are Cr and austenitic Cr – Ni alloys, particularly for transfer lines and reboiler tubes, and Cr – Mo alloys for furnace tubes. The lower parts of the crude and vacuum columns should be lined with Cr steels.
2) Attack by naphthenic acid to form iron naphthenates and subsequent corrosion. Austenitic Cr – Ni – Mo alloys can be used and naphthenic acid can be removed upstream of the columns.
3) Hydrogen chloride corrosion in overhead condensation systems. Monel or other Ni – Cu alloys can be used. Measures which can be taken to prevent corrosion include desalting of crude oils, neutralization of overhead vapors with ammonia, and pH-control, and use of corrosion inhibitors in endangered equipment sections.

Corrosion in cracking processes is caused by the following:

1) *High-Temperature Sulfur Corrosion.* Cr – Mo alloys in the furnace and heat-exchanger tubes and Cr steel lining in the columns are suitable materials for corrosion prevention.
2) *Erosion by Catalysts.* Ceramic and concrete linings in vessels and cyclones and regenerator internals made of Cr – Mo and austenitic Cr – Ni alloys are suitable materials.

Corrosion in hydrogen-carrying processes is caused by hydrogen attack at high temperatures (blistering and embrittlement of steels). Cr or Mo steels, at higher temperatures austenitic Cr – Ni alloys, can be used to prevent corrosion. In the presence of hydrogen and H_2S, i.e., in hydrotreating and hydrocracking processes, steels with high chrome contents must be used.

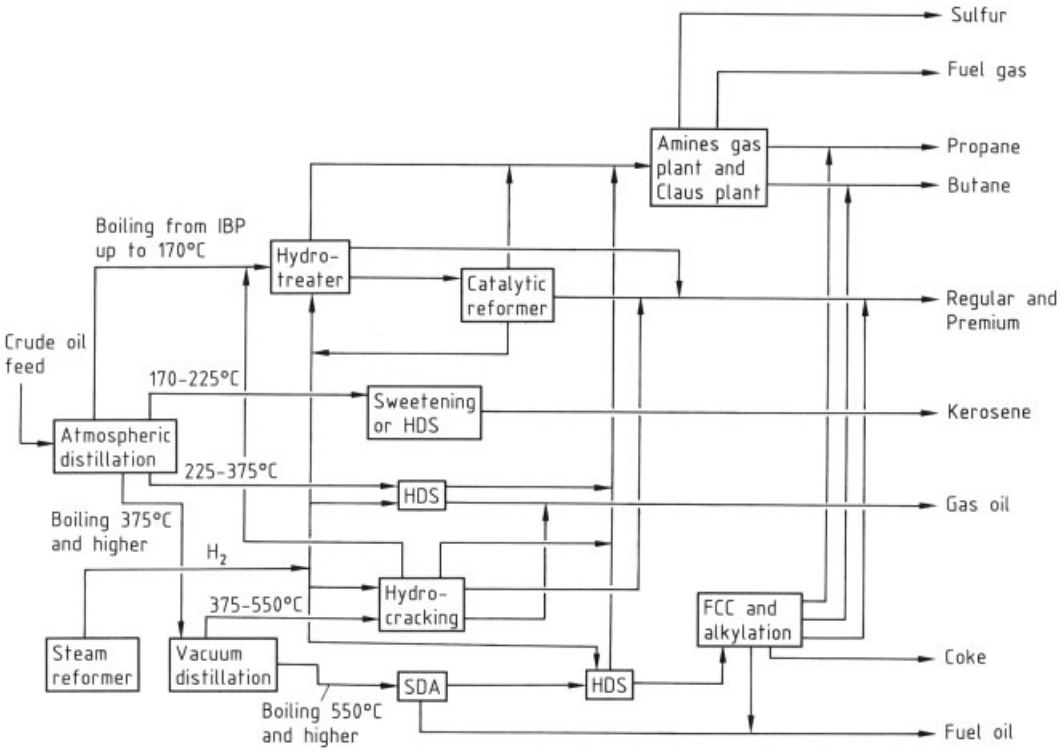

Figure 43. Hydrocracking – catalytic cracking

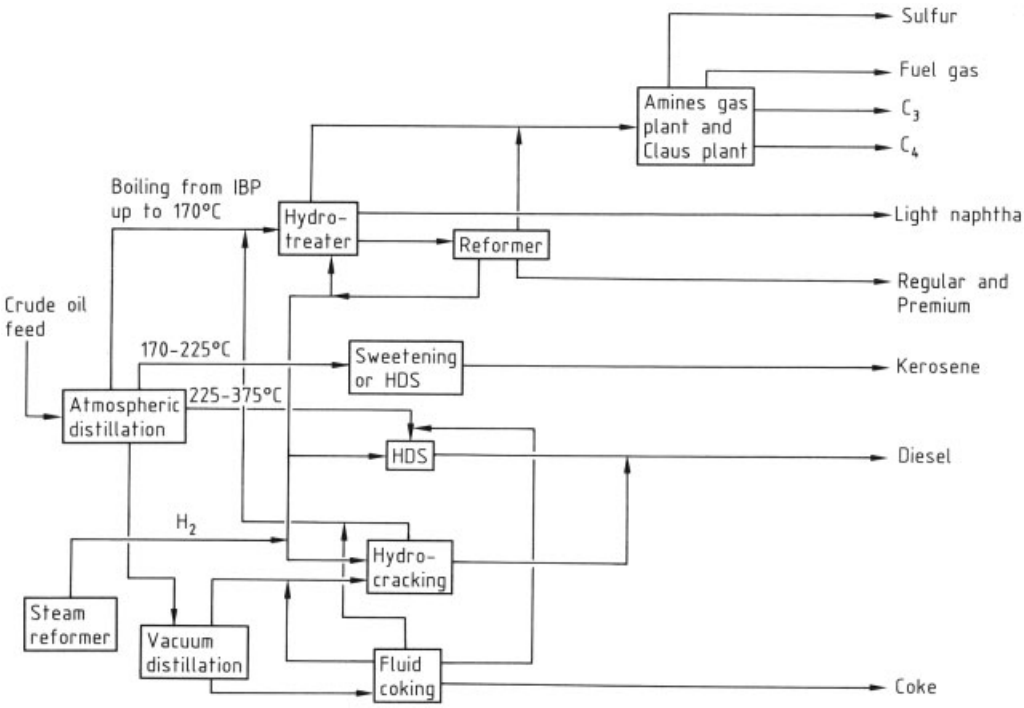

Figure 44. Hydrocracking – coking

7. Testing and Analysis

The chemical and physical properties of crude oils and their fractions depend on the types and concentrations of their consituents. The typical properties of crude oils and their products must be defined and determined by analytical methods, to meet quality requirements for production, transport, storage and handling of crude oils, and to assess specific quality criteria for further processing of products.

This quality testing is performed by regular laboratory analysis or by automatic instruments.

7.1. Crude Oil and Product Properties

Density and API Gravity. The density ranges of various crudes are as follows:

Paraffin-based crudes	815 – 835 kg/m^3
Mixed-based crudes	835 – 860 kg/m^3
Naphthalene-based crudes	860 – 955 kg/m^3

The API gravity (introduced by the American Petroleum Institute) is the internationally used density measure for crude oils. The relation between degrees API and density at 15 °C is:

$$^{\circ}\mathrm{API} = \frac{141.5}{\mathrm{Density\ [kg/L]}} - 131.5$$

Distillation. The *boiling range* of crude oils is one of the most important quality criteria for determining the yield distribution of the different product fractions. Boiling analyses under atmospheric conditions can be supplemented by distillation under reduced pressure for oil fractions that boil above 360 °C.

More detailed and precise results are obtained by applying fractional or *true boiling point* (TBP) distillation.

Viscosity. The viscosity of crude oils is determined by the ratio of low and high molecular mass constituents in the oil. The kinematic viscosity of crude oils (at 38 °C = 100 °F) can vary from 4 mm^2/s (0.04 St) for light (North Sea) crudes to 700 mm^2/s (7 St) for heavy (Venezuelan) crudes.

Flash point. The flash point depends on the concentration of light (low-boiling) constituents. For crude oils, flash points between − 20 °C and + 80 °C are observed.

Pour Point. The pour point is the temperature at which an oil or oil product ceases to flow. It is an important measure of its cold flow properties. (→ Automotive Fuels, Chap. 5.2.3.).

Other properties that are analytically tested include specific heat; calorific value; contents of sulfur, nitrogen, oxygen, and inorganic compounds; the C/H ratio; the hydrocarbon composition; and product-specific criteria (see Section 2.3).

7.2. Testing Methods and Standards

In the oil industry a framework of standard analytical procedures has been established, e.g., by ASTM, DIN and the British Institute of Petroleum, (IP) standards. In addition, various house standards have been introduced by most international oil companies.

A selection of frequently applied DIN methods is listed below:

Density	DIN 51757
Distillation (boiling range)	DIN 51751
Vacuum distillation	DIN 51356
Fractional distillation	DIN 51567
Flash point	DIN 51755, 51758, 51376, 51584, 53213
Viscosity	DIN 53012, 51550
Cold properties (cloud point, pour point, freezing point)	DIN 51583, 51597, 51421
Water content	DIN 51582
Water and sediment content	DIN 51793
Salt content	DIN 51576
Sulfur content	DIN 51400, 51768, 51409, 51418, 51450

Analytical testing by *automatic continous analyzers* has come into use worldwide. The continuous analysis of process streams by automatic "onstream analyzers" has proven to be an indispensable tool for the quality control of products in integrated refining systems.

8. Storage and Transport

Surface Storage. *Crude oils* are stored in cylindrical fixed-roof or floating roof tanks; *LPG* is stored in spherical or cylindrical pressure vessels (→ Liquefied Petroleum Gas, Chap. 5.). Suitable storage tanks for *gasolines, kerosenes,* and *gas oils* are cylindrial floating-roof or fixed-roof tanks (optionally with vapor removal) (→ Automotive Fuels, Chap. 7.). *Heavy fuel oils* and *bitumen* are stored in cylindrical fixed-roof tanks (optionally with gas blanketing).

Underground Storage. Crude oil and liquid (e.g., light heating oil), and gaseous products (LNG) are stored in underground caverns (→ Liquefied Petroleum Gas, Chap. 5.). The underground storage of crudes and liquid products in caverns is usually intended as long term (emergency) storage; caverns for gas are used as buffer storage (→ Natural Gas, Chap. 4.2.2.).

Subsea Storage. Crude oils (from offshore production) and hydrocarbon condensates (from offshore separation) are stored in undersea tanks.

Transport of Crude Oils and Products. Crude oils and bulk (mass) products are normally transported by large tankers and pipelines. Finished products are usually transported by smaller tankers and river barges, product pipelines, rail tank cars, and road tank trucks.

9. References

1. H. S. Bell: *American Petroleum Refining,* D. van Nostrand, Princeton, NJ, 1959.
2. *Oil Gas J.* **95** (1997)Dec. 22, "OGC" Special,"41 – 44.
3. *Hydrocarbon Process.* **48** (1969) no. 2, 77 – 88.
4. *Oil Gas J.* **71** (1973) no. 45, 88 – 92.
5. *Hydrocarbon Process.* **51** (1972) no. 5, 85 – 88.
6. *Hydrocarbon Process* **51** (1972) no. 5, 92 – 97.
7. *Hydrocarbon Process.* **65** (1986) no. 9, 93, 95.
8. In [7], p. 94.
9. *Hydrocarbon Process.* **63** (1984), no. 9, 106.
10. 1984 Refining Process Handbook, *Hydrocarbon Process.* **63** (1984) no. 9, 70 – 84, 89 – 90.
11. 1986 Refining Process Handbook, *Hydrocarbon Process.* **65** (1986) no. 9, 84 – 87.
12. *Hydrocarbon Process.* **52** (1973) no. 9, 131 – 133.
13. *Pet. Eng. Manage.* **28** (1956) no. 3, C 37 – C 44.
14. *Hydrocarbon Process.* **58** (1979) no. 9, 136 – 142.
15. *Hydrocarbon Process.* **58** (1979) no. 5, 103 – 106.
16. *Hydrocarbon Process.* **64** (1985) no. 3, 51 – 59.
17. *Oil Gas J.,* **81** (1983) no. 3, 103 – 113.
18. *Hydrocarbon Process.* **63** (1984) no. 9, 70.
19. *Hydrocarbon Process.* **65** (1986) no. 9, 98 – 100.
20. *Hydrocarbon Process.* **63** (1984) no. 9, 115 – 119.
21. In [20], p. 118.
22. In [20], p. 115.
23. *Oil Gas J.,* **74** (1976) no. 3, 48 – 52.
24. *Oil Gas J.,* **70** (1972) no. 21, 52.
25. *Hydrocarbon Process.* **65** (1986) no. 9, 99.
26. *Oil Gas J.,* **79** (1981) no. 22, 102 – 105.
27. *Hydrocarbon Process.* **63** (1984) no. 9, 104.
28. E. Naber et al.: "Refining in the 1990's," *Erdoel Erdgas* **107** (1991) no. 3, 124 – 132.
29. A. K. Logwinuk et al.: *Factors Affecting Choice of Heavy Hydrocarbon Processing,* NPRA Ann. Meet., San Antonio 1981.
30. H. N. Olson: "Cost and Cost Aspects in Residue Conversion," *Proc. World Pet. Congr. 11th, 1983,* 1 – 16.
31. H. R. Siewert et al.: "Optimize Design for Heavy Crude," *Hydrocarbon Process.* **64** (1985) no. 3, 61 – 66.
32. Shell Internationale Petroleum Maatschappij, EP 264 158, 1987 (R. H. van Dongen et al.).
33. M. A. Menzies, T. F. Scot, J. Denis: "The Canmet Residuum Hydrocracking Process," *Proc. Intsoc. Energy Convers. Eng. Conf., 15th* 1980, 26 – 32.
34. *Hydrocarbon Process.* **65** (1986) no. 9, 97, 98.
35. *Hydrocarbon Process.* **65** (1986) no. 9, 95, 96.
36. V. D. Singh, *Erdoel Kohle* **39** (1986) no. 1, 19.
37. S. W. Martin, L. E. Wills: "Coking Petroleum Residues," in K. A. Kobe and J. J. McKetta (eds): *Advances in Petroleum Chemistry and Refining,* vol. 2, Interscience Publishers, New York – London 1959.
38. *Oil Gas J.* **80** (1982) no. 20, 93 – 102.
39. H. Predel: Delayed Coking, *Erdoel Kohle Erdgas Petrochem.* **43** (1990) no. 4, 145 – 146.

40. M. J. Humbach et al.: "Residue Conversion, RCC in the 1990's," *Erdoel Kohle Erdgas Petrochem.* **43** (1990) no. 12, 477 – 482.
41. A. Rhoe et al.: *Meeting the Refiners Upgrading Needs,* NPRA Annu. Meet., San Francisco 1983.
42. *Hydrocarbon Process.* **63** (1984) no. 9, 123.
43. *Hydrocarbon Process.* **64** (1986) no. 9, 124.
44. *Hydrocarbon Process.* **64** (1986) no. 9, 126.
45. *Hydrocarbon Process.* **60** (1981) no. 9, 134 – 138.
46. *Hydrocarbon Process.* **61** (1982) no. 5, 110 – 112.
47. *Hydrocarbon Process.* **63** (1984) no. 9, 125.
48. *Hydrocarbon Process.* **63** (1984) no. 9, 121.
49. *Hydrocarbon Process.* **59** (1980) no. 5, 110 – 114.
50. *The Oil Refining Industry and the Environment,* CONCAWE, The Hague, Sept. 1987.
51. O. Neuwirth: "Umweltfragen bei der Nutzung von Mineralöl," *Erlanger Forsch. Reihe B; Naturwiss. Med.* **14** (1984) 119 – 134.
52. *Mineralöl und Umweltschutz,* 3rd ed., Publ. Mineralölwirtsch. Verband, Hamburg, 1989.
53. A. H. Edwards et al.: "Volatile Organic Compound Emissions in Western Europe," *CONCAWE Rep.* 1986, no. 2.
54. A. de Roocker et al.: "Trends in Oil Discharged with Aqueous Effluents from Oil Refineries in Western Europe," *CONCAWE Rep.* 1985, no. 61.
55. C. L. F. Webb: "Biodegradation of Contaminated Soil," *Erdoel Erdgas* **106** (1990) no. 11, 453 – 455.
56. J. S. McArragher et al.: "Motor Vehicle Emission Regulations and Fuel Specifications, *CONCAWE Rep.* **1990,** no. 2.
57. International Workshop Berlin: "Emissionsarme Betankung von Kraftfahrzeugen," *Erdoel Kohle Erdgas Petrochem.* **42** (1989) 468.
58. R. J. Ellis: "Cost Estimating Aspects of Environmental Control Technology – HDS as a Case Example," *CONCAWE Rep.* 1988, no. 51.
59. Commission of the European Communities: *Study on the Costs Borne by the Refining Industry in the Member States in Order to Comply with Environmental Legislation,* prepared by Chem. Systems Int., London 1986.
60. G. Heinrich, Y. Jacquin, A. Vidal, *Proc. World Pet. Congr., 10th, 1979,* vol. 4, paper PD 18-1, Heyden & Sons, London 1979.
61. F. G. Turpin, C. R. Lipuma, A. M. Edelman, in [60], paper PD 18-2.
62. C. T. Douwes, J. van Klinken, J. B. Wyffels, W. C. van Zijll Langhout, in [60], paper 18-3.
63. *Erdoel Erdgas Kohle* **105** (1989) no. 9, 369 – 372.
64. *Erdoel Erdgas Kohle* **106** (1990) no. 7/8, 309 – 315.

Oil Sand → Tar Sand

Oil Shale

Hans-Jürgen Weiss, Lurgi AG, Frankfurt a. M., Federal Republic of Germany

1. Introduction . . . 1
2. Origin and Properties of Oil Shale . 2
2.1. Origin, Stratification . . . 2
2.2. Physical Properties . . . 2
2.3. Chemical Composition . . . 2
3. World Resources and Deposits . . . 4
3.1. Europe . . . 4
3.1.1. Western Europe . . . 4
3.1.2. Eastern Europe . . . 5
3.2. America . . . 5
3.3. Asia . . . 6
3.4. Africa . . . 6
3.5. Australia and New Zealand . . . 6
4. Present Exploitation . . . 6
5. Mining Methods . . . 7
5.1. Underground Mining . . . 7
5.2. Above Ground Mining . . . 8
6. Shale Preparation . . . 8
7. Pyrolysis . . . 8
7.1. Definition and Kinetics . . . 8
7.2. Above Ground Retorting Processes 10
7.2.1. Class 1 (Indirectly Gas-Heated) Retorts . . . 11
7.2.2. Class 2 (Directly Gas-Heated) Retorts 14
7.2.3. Class 3 (Directly Solids-Heated) Retorts . . . 19
7.3. In Situ Processing of Oil Shale . . . 24
8. Combustion . . . 24
9. Gasification . . . 25
10. Properties and Uses of the Products 26
11. Outlook . . . 26
12. References . . . 27

1. Introduction

Oil shale is a sedimentary rock that contains a considerable proportion of solid organic material embedded in a matrix of fine inorganic minerals. Other sedimentary rocks, too, contain organic substances, but the name oil shale suggests that it releases a considerable quantity of oil when processed appropriately. Many rocks, although they satisfy the definition of oil shale, are not true shales geologically.

To recover the oil from the shale, its organic portion – termed kerogen – must be decomposed thermally. This thermal decomposition, known as retorting, converts the solid organic material of the shale into liquid and gaseous fractions and a solid carbonaceous residue. The *liquid fraction*, the so-called shale oil, consists of condensable hydrocarbons (C_{5+}) and small quantities of decomposition water. The *gaseous product* is a mixture of carbon monoxide, carbon dioxide, hydrogen, nitrogen, hydrogen sulfide, methane, and other hydrocarbons (C_{4-}). The *carbonaceous residue* – a cokelike material – is obtained in a mixture with the inorganic minerals of the original oil shale.

Oil shale normally contains fewer organic and more inorganic substances than various types of coal. The organic contents of oil shale often account for $<10\,\%$ of the total mass, whereas coal usually contains ≥ 75 wt % organic substances in terms of water-free material.

A major difference between oil shale and coal lies in the fact that the organic compounds of shale are essentially insoluble in organic solvents (e.g., toluene). Bituminous rocks, oil chalk, tar sands, etc., are therefore not included among oil shales.

The molar ratio of hydrogen to carbon in the organic substances in oil shale is higher than that of the pure organic component of various types of coal. The retorting of oil shale normally yields a higher amount of oil than the pyrolysis of a coal sample containing an equivalent amount of organic material. Unlike coal, oil shales have up to now been less investigated, and no generally accepted classification system exists. For this reason, the names given to oil shales of different origins rarely provide a clear indication of their qualities or types. In addition to general descriptions such as bituminous shale or kerogenic rock, the names may relate the material to a certain deposit or area (e.g., algal shale, paper coal, coorongite, ichthyol shale, kukersite, marahunite, posidonian shale, fetid shale, tasmanite, or torbanite).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_101

The organic component of the world's oil shale deposits is a fossil source of energy that is second in abundance only to coal. Despite the large oil shale deposits, fuel production from shale has so far been uneconomical and hence insignificant. Whenever oil shale retorting plants were built, it was due to abnormal conditions such as a scarcity of conventional fuels because of war or heavy government subsidies. However, as the reserves of conventional crude oil gradually decrease, oil production from shale may become necessary and justifiable for economic reasons.

2. Origin and Properties of Oil Shale

2.1. Origin, Stratification

Oil shale originates from a syngenetic accumulation of bituminous plant components which settled as gyttja (sapropel) in marine, paralic, or limnic basins. Unlike coal, it does not contain any humic matter. Oil shale is found in all geologic strata from the Cambrian to the Tertiary. Many deposits exhibit an annual stratification of sediments that gives them a shalelike appearance. In some rocks, these annual strata are very thin. Some are rich in organic substances; others contain few. In most oil shale, the individual particles are very small; their diameter often does not exceed 200 µm.

The most frequently found inorganic minerals in oil shale deposits are quartz, feldspar, various clays, and carbonates such as calcium carbonate and magnesium carbonate. Other minerals, which have apparently been formed in situ within the sediments, consist of the sulfides of iron, copper, and other metals and of certain sodium salts. Table 1 shows the mineral composition of Green River oil shale.

Table 1. Mineral composition of the inorganic component of North American oil shale (Green River)

Mineral	Chemical formula	Content, wt %
Dolomite	$CaMg(CO_3)_2$	32 – 33
Limestone	$CaCO_3$	16 – 20
Quartz	SiO_2	10 – 15
Clays (illite)	$K_2O \cdot 3\,Al_2O_3 \cdot 6\,SiO_2 \cdot 2\,H_2O$	11 – 19
Orthoclase (adular)	$KAlSi_3O_8$	4 – 6
Plagioclase (albite)	$NaAlSi_3O_8 \cdot CaAl_2Si_2O_8$	10 – 12
Zeolites (analcime)	$NaAlSi_2O_6 \cdot H_2O$	1 – 7
Pyrite, marcasite	FeS_2	1 – 3

Traces of numerous metals including uranium, vanadium, arsenic, and zinc are also common. Phosphate-bearing oil shale of marine origin contains fluorine and uranium; the latter is often also found in nonphosphate oil shales. Sodium salts occur as interspersed crystals and also, in a few cases, in the form of salt deposits embedded in the oil shale; these constitute valuable salt reserves.

2.2. Physical Properties

The physical properties of oil shales differ widely. Discrete layers are frequently found in the shale, and many of these rocks are brittle and cleave easily along the laminate boundaries. Other oil shales are very hard and compact, although a stratification pattern is clearly recognizable. The density of the organic matter in the shale is generally on the order of 1.0 g/cm^3, whereas the inorganic minerals have characteristic densities of, for instance, 2.6 g/cm^3 in the case of feldspar and quartz, and 5.0 g/cm^3 in the case of pyrite. Owing to the differences in the specific masses of organic and inorganic components, high-grade (i.e., high organic content) oil shales have a lower density than low-grade ones. The physical properties of low-grade oil shales resemble those of the inorganic rock components. High-grade oil shales are dark colored because of their relatively high organic content. Important physical properties of oil shales are listed in Table 2 (see next page).

At room temperature, the solubility of kerogen is low in most inorganic and organic solvents. Kerogen solubility increases with increasing temperature, but a noticeable effect is not achieved until the temperature has reached a level at which the chemical links in the kerogen begin to break up. Because of the extremely low tendency of kerogen to dissolve or to swell noticeably – except during thermal decomposition – a rigid structure is assumed [3,4].

2.3. Chemical Composition

To be of economic interest today, oil shale should contain at least 10 % organic material. In most cases, economic interest focuses on the kerogen.

Table 2. Physical properties of oil shale (dry)

	Moisture, %	Gross calorific value, kJ/kg	Grinding index, Hardgrove	Bulk density, g/cm³	Density, g/cm³	Ash melting point, °C
Schandelah (Germany)	3–4	4800	76	1.0–1.2		
Dotternhausen (Germany)	4–5	2900–3800	58	1.2–1.3	2.15–2.20	1150
Kvarntorp (Sweden)	1–2	9000		2.1		
Estonia (Soviet Union)	14	11 000–15 000		1.3–2.5	1.5–2.7	
Timahdit (Morocco)	6–7	5700	69	1.0–1.2		1180
Green River (United States)	0.5–1.5	6500	43	1.2–1.4	2.1–2.3	1300
Irati (Brazil)	2–5	5200	68		2.3	
Fushun (China)	5	3350–6100			2.3	
Maoming (China)	15–18	5600	62	0.9–1.0		
El Lajjun (Jordan)	4	6900	73	1.0–1.1		1260
Rundle (Australia)	20	5000–6600	60	0.85–0.95		

However, in individual cases the shale may also be processed for all or part of its mineral content; thus low-grade oil shales may also be exploited.

The character of the kerogens in various oil-bearing strata is so complex that it is difficult to describe, and no exact classification has been developed so far. When the kerogen was formed from spores, algae, and other organic matter, conditions strongly favored decomposition. Thus, the properties of kerogen predecessors are largely unknown. Kerogen may generally be described as a mineraloid with a high molecular mass whose composition varies greatly from one deposit to the next. Its major elements are carbon, hydrogen, nitrogen, sulfur, and oxygen, which form heterogeneous macromolecules consisting of aromatic and polycyclic condensed rings with aliphatic side chains of varying lengths linked weakly to them. These side chains themselves are cross-linked to other aliphatic chains via oxygen or sulfur atoms.

Table 3 shows the elemental analysis of typical fossil and recent organic substances along with their C/H mass ratios. In comparison to most other basic organic substances, the C/H ratio of the shale kerogens is relatively low. The same is also true also for shale oil (see Table 8). Consequently, less hydrogen must be added to these hydrocarbons at the refining stage to obtain products such as fuel oil or gasoline. This

Table 3. Elemental analyses and C/H ratios of typical samples of kerogens, crude oil, bitumen from tar sands, wood, peat, and coal

	Analysis, wt %					
	C	H	N	S	O	C/H ratio
Kerogen						
Schandelah (Germany)	65.8	7.7	2.1	6.9	17.5	8.5
Dotternhausen (Germany)	81.8	8.9	1.5	0.6	7.2	9.2
Kvarntorp (Sweden)	67.7	7.2	0.8	6.8	17.5	9.4
Estonia (Soviet Union)	77.1	9.4	0.4	2.0	11.1	8.2
Alexinac (Yugoslavia)	76.6	10.7	3.9	3.9	4.9	7.2
Timahdit (Morocco)	66.7	10.9	2.4	10.9	9.1	6.1
Green River (United States)	80.5	10.3	2.4	1.0	5.8	7.8
Crude oil						
Oklahoma (United States)	85.4	13.0	0.1	0.6	0.9	6.6
Bitumen						
Athabasca (Canada)	83.1	10.3	0.4	4.6	1.6	8.1
Wood						
Oak	51.2	5.9	0.2	0.0	42.7	8.7
Peat						
Northern Germany	58.8	5.7	1.7	0.4	33.4	10.3
Coals						
Lignite (Rhineland, Germany)	66.6	4.9	1.1	0.4	27.0	13.6
High-volatile bituminous coal (Saar, Germany)	81.7	5.3	1.2	0.8	11.0	15.4
Low-volatile bituminous coal (Ruhr, Germany)	88.6	5.0	1.6	1.0	3.8	19.9
Anthracite (Aachen, Germany)	90.6	4.0	1.3	1.1	3.0	22.6

Table 4. Fischer assay of various oil shales

	Moisture* content, wt %	Assay, wt %			
		Water	Oil	Residue	Gas + losses
Schandelah (Germany)**	3.7	1.7	5.7	84.7	4.2
Dotternhausen (Germany)	3.9	0.4	4.3	88.6	2.8
Kvarntorp (Sweden)	2.0		5.7	87.2	5.1
Estonia (Soviet Union)	dry	1.9	22.0	70.5	5.6
Alexinac (Yugoslavia)	5.5	2.1	12.6	75.3	4.5
Timahdit (Morocco)	6.4	2.0	7.6	80.0	4.0
Green River (United States)	dry	1.4	10.4	85.7	2.5
Irati (Brazil)	4.6	1.2	6.9	83.6	3.7
Fushun (China)	dry	4.9	3.0	90.3	1.8
Maoming (China)	5.0	3.2	7.3	80.6	3.9
El Lajjun (Jordan)	4.0	1.4	10.1	80.8	3.7
Rundle (Australia)	6.1	3.0	9.4	76.6	4.9

* Moisture, as analyzed, does not represent in situ moisture.
** Ash analysis yielded the following figures in wt %: 24.0 CaO, 38.4 SiO_2, 13.8 Al_2O_3, 6.6 Fe_2O_3, 9.7 SO_3, 2.9 MgO, 2.1 K_2O, 0.7 Na_2O, 0.8 TiO_2, 0.4 P_2O_5, 0.1 BaO.

fact, along with the relative ease with which oil can be produced from oil shale, gives it an edge over coal as a raw material for the production of liquid fuels.

In the Fischer assay (common analysis method for oil shales and coals), oil shale is slowly heated to 520 °C and pyrolyzed, and the masses of the resulting water, oil, gas, and residue are measured. Some results of Fischer assays from oil shales are summarized in Table 4.

3. World Resources and Deposits

Oil shale deposits are found on all continents, but no quantitative figures of total deposits worldwide have been compiled as yet. Even in areas with highly promising deposits, resources have generally been investigated very little. Consequently, the estimates are inaccurate and differ widely in some cases. The data on oil shale deposits in Table 5 were compiled by BURGER from various sources [5] and refer also to estimates made by DUNCAN [6]. They make no pretense of completeness.

Table 5. Estimate of known shale oil reserves from oil shale yielding more than 42 L of oil per tonne [5] (only deposits containing $>50\times10^6$ m^3 of shale oil are listed)

Location of reserves	Amount, 10^6 m^3
United States	
Rocky Mountains	318 000
Eastern areas	32 000
Brazil	
Irati	127 000
Paraiba	320
Australia	
Queensland	4 300
Soviet Union	
Estonia/Leningrad	3 500
Rest of Russia	2 000
Siberia	12 400
Zaire	
Kinshasa	16 000
Canada	
New Brunswick	7 000*
Italy	
Sicily	5 600
China	
Fushun	330
Others	4 100
Morocco	440
Thailand [6]	440
Sweden	400
Germany	320
Burma	320
United Kingdom	160
Luxemburg	110
France	70
Argentina	60
Other known deposits in various countries	209

* Rough estimate for high-grade oil shale (>100 L/t).

3.1. Europe

3.1.1. Western Europe

Western Europe has a large number of oil shale deposits dating back to different geological periods. Most of them are small. Of more than regional importance is the marine Liassic formation in Spain, France, Belgium, Germany, and the United Kingdom, which contains oil shale

deposits. Information about these deposits is still fragmentary [7, 8]. As a rule, *marine shales* cover large areas, but their layer thickness is relatively small and they contain only moderate amounts of oil. *Paralic shales* consist mostly of thin layers with higher oil content. They are frequently found in the immediate neighborhood of humic formations. *Limnic deposits* are limited in area but may be of considerable thickness with high oil content.

Germany. Germany has relatively small deposits; posidonian shale (Liassic) is commercially exploited near Balingen. The largest deposit, Schandelah (Lower Saxony), contains 2×10^9 t of oil shale with an oil content of about 6 wt %. The Nördlinger Ries is presumed to contain 2×10^9 t of oil shale. Major quantities of posidonian shale occur in northern and southern Germany at greater depths [9]. A tertiary limnic "paper coal" deposit (Messel, Hesse) has been largely exhausted but is still of great interest to paleontologists [10].

Sweden. Sweden has Cambrian–Silurian deposits of 10-m thickness near Kvarntorp. Their appreciable uranium content may become commercially interesting.

United Kingdom. Seams of oil shale occur in 1800-m Carbonian strata west of Edinburgh. The deposit is mined down to 300-m depth. Kerogen-holding Liassic strata with possibly higher oil contents than those in Scotland are widespread in England [11].

France. France has major Permian deposits: Autun, Aumance (mined for 100 years), Ceverac-le-Chateau, and Creveney.

Spain. A Carbonian oil shale near Puertollano was exploited industrially for about 40 years.

Italy. In Sicily, shale deposits exist that contain ca. 6×10^9 t of oil.

3.1.2. Eastern Europe

In Eastern Europe, oil shale deposits occur mainly in Yugoslavia, Bulgaria, and the Soviet Union. Smaller deposits are also found in Czechoslovakia, Albania, and Greece.

Yugoslavia. Oligocene limnic oil shale is found near Alexinac, Serbia. Its thickness is 80 – 120 m, containing 10 – 15 wt % oil (max. 25 wt %); the total shale deposits amount to ca. 2×10^9 t.

Bulgaria. Oil shale deposits are found in Krassawa (some 50×10^6 t recoverable by opencast mining) and Stara Sagora (thickness 50 – 70 m; 4×10^9 t of oil shale recoverable by opencast mining, containing 4 – 6 wt % oil).

Former Soviet Union. Some 50 oil shale areas are known [12], the largest in the Baltic region and the Volga basin. The Baltic deposits dating from the Ordovician period (kukersite) have oil contents of ca. 20 wt %. They have been commercially exploited since 1916. The recoverable reserves are ca. 2×10^9 t.

3.2. America

Canada. Shale with low oil content occurs in all Canadian provinces; its exploitation is presumably not cost effective.

United States. Deposits with very high oil contents are found in the Green River formation in the Rocky Mountain states of Colorado, Utah, and Wyoming. Their oil content is ca. 10 wt %. The deposits cover ca. 20 000 km^2 and are ca. 30 m thick. Colorado also has deposits with a thickness up to 300 m. This is a low-grade oil shale covering some 50 000 km^2; Utah has deposits with ca. 9 wt % oil content but measure only 10 – 20 m in thickness. Thin Devonian oil shale layers at great depths are found in the eastern United States.

Alaska has little-known deposits dating from various periods [13].

Brazil. Brazil's largest and best-known deposit is the Permian Irati formation. Other low-grade oil shale deposits from the Cretaceous and Tertiary periods are found in various provinces [14].

Oil shale deposits also occur in other South American countries.

3.3. Asia

Oil shale is found in Turkey, Syria [15], Israel [16–18], Lebanon (very high oil contents), and Iraq.

Jordan. High-grade oil shales have been found at El Lajjun (20 km^2 with 130×10^6 m^3 of oil and 10 wt % oil content). Other deposits have been reported (e.g., at Sultani).

Thailand. Particularly high-grade oil shales containing ca. 20 – 25 wt % oil in four seams of low thickness are found in an area of ca. 28 km^2. Some 400×10^6 t of oil are confirmed in Mae Sod. However, their average grade is low (5 wt %).

China. Oil shale deposits occur near Maoming and Fushun; the latter is 100 – 200 -m thick and situated above a very thick coal deposit. Only about 3.5 wt % of oil is obtained by processing the shale.

3.4. Africa

Morocco. Deposits exist near Timahdit and Tarfaya. According to reliable estimates they contain 2×10^9 t and 4×10^9 t, with 7.5 and 5 wt % of oil, respectively.

Zaire. Several layers are spread over ca. 14 000 km^2, with a total thickness of ca. 10 m, that contain oil at a concentration of more than 100 L/t. Estimated reserves are 15×10^9 t of oil.

South Africa. Transvaal has oil shale (torbanite) of very high quality (20 wt % oil) and a seam thickness of ca. 2 m, e.g., near Ermelo and Krill [the deposit at Ermelo was mined and retorted by the South African Torbanite Mining and Refining Co. (SATMAR) for 25 years]. More recent studies have concentrated around the deposit near Krill (15×10^6 m^3 of oil).

3.5. Australia and New Zealand

Economically important oil shale deposits are found in various parts of *Australia*. Recoverable oil reserves in Queensland are estimated at 4.3×10^9 m^3. Major reserves in Queensland are Condor and Rundle.

Little-known deposits from the Tertiary period exist in *New Zealand*.

4. Present Exploitation

Oil shale industries were founded over the last 100 years in Germany, Scotland, Sweden, France, Spain, South Africa, and Australia, but all of them closed after a while. Only two countries – the former Soviet Union and the People's Republic of China – succeeded in setting up oil shale industries that are still processing, respectively, ca. 24×10^6 and 40×10^6 t of oil shale annually. The United States and Brazil, where oil shale reserves are most abundant and richest in oil, have developed oil shale retorting technology to the prototype stage (Union Oil, Petrobras).

Germany produced small quantities of high-sulfur shale oil for medical purposes centuries ago. Industrial oil shale retorting can be traced back to the middle of the last century. Its origins lay in Swabia. Only the Dotternhausen cement works in Germany still uses some 700 t/d of oil shale on a commercial basis. At that factory, fine-grained shale is burned in a fluidized bed (Rohrbach – Lurgi process) to generate energy and produce a residue that is used as an additive in cement production.

A description of the Swedish oil shale industry is given in [19]. Operations were discontinued in 1963. Considerable quantities of shale oil were also produced in Scotland, Spain, and France.

In the lands that later constituted the former Soviet Union, oil shale was initially used as a *fuel*. Retorts were employed for the first time in Estonia in 1925; they were followed by indirectly heated rotary kiln retorts and tunnel kilns with rinse gas heating systems [20]. In 1940, ca. 1.9×10^6 t/a of oil shale was processed in Estonia, 0.4×10^6 t/a in the Leningrad area, and 0.3×10^6 t/a in the Volga basin. Most of these oil shale plants were destroyed during World War II. Reconstruction began shortly after the war and the quantities of oil shale processed in the former Soviet Union grew steadily [21] (quantities in 10^6 t/a).

1945	1950	1955	1960	1965
1.384	4.716	10.793	14.148	21.259

Among these totals, the Estonian oil shale industry accounted for ca. 70 %. Growth trends in the oil shale industry in the Soviet Union are shown in Table 6 [22].

Table 6. Growth of oil shale industries in the Soviet Union (percentage of annual production in 1958 or in the first year of production)

Product	1958	1960	1965	1970	1975
Town gas	100	109	125	135	140
Shale oil	100	119	147	160	180
Aromatic hydrocarbons			419∗	691	1156
Tanning agent		100	178	472	722
Synthetic resins and plastics			500∗	1067	1104
Electrode coke			100	861	1491
Lacquers	100	100	700	2750	3130
Kerogen-70				1102∗∗	3000

∗ Production started in 1962.
∗∗ Production started in 1966.

In *China*, oil shale plants first went on stream in 1914 in Fushun, Manchuria, along with coal mining operations. The oil shale there constituted the overburden that had to be removed to gain access to a coal deposit of considerable thickness. Currently, two retorting plants with an annual consumption of ca. 9×10^6 t of oil shale are operating in Fushun. The Fushun shale is of poor quality. The richest shale found so far contains about 4.7 wt % oil. Normally, shale with an oil content <3.5 wt % is rejected. The strip mine covers an area of 40 km^2 and belongs to the largest open-pitmines of the world.

The Maoming plant in the Guandong Province started up in 1961 and currently produces ca. 100 000 t of shale oil per year. Production was reduced in recent years from 200 000 t of shale oil annually due to the discovery of crude oil [23].

World oil demand is currently met by a sufficient supply of crude oil. Whenever conventional oil supplies are sufficient, shale oil has not been competitive. However, decreasing reserves and an increasing demand for oil will necessarily lead to an escalation of its price in the future. With this perspective in mind, shale oil produced by modern cost-effective technologies may be expected to be competitive in the medium run. Great efforts have been made for years, especially in the United States, Germany, Brazil, and the Soviet Union, to develop more effective shale oil production processes (see Chap. 7).

5. Mining Methods

Before oil shale can be processed in surface retorting plants, it must be mined and transported to these plants. Mining costs constitute a major share of the total shale oil production costs. Modern commercial oil shale processing plants can be cost effective only if large and highly efficient equipment is used, which keeps the cost per tonne of mined material to a minimum. Prospects are good for the introduction of new processes and equipment to mine oil shale deposits of great thickness that often cover considerable areas.

Either underground or open-pit methods can be used to mine oil shale. The method selected in each particular case depends on the depth and size of the deposit; on topographical, geological, and hydrological considerations; and on environmental questions.

5.1. Underground Mining

Room-and-Pillar System. The room-and-pillar system is appropriate wherever conveniently accessible deposits of great thickness are exploited. With this method, part of the shale is taken out in such a way that underground rooms of regular shape are produced. The remaining shale is left in place as pillars between these rooms to support the overburden.

The pillar and room sizes are determined by the physical properties of the shale and the overburden. This mining method is widely used for ore and coal mining, and has also been employed in oil shale mining operations (e.g., in the United States, South Africa, Australia, and Estonia).

In the *Green River formation* the room-and-pillar system was used to mine oil shale deposits 10 – 30 m thick under an overburden of 50 – 150 m. The rooms had a width of ca. 18 m, leaving square pillars of 18-m side length and 18-m spacing between them in offset rows. In this way, 75 % of the total shale can be mined, whereas 25 % must be left in place as supporting pillars. The advantage of this system is that large power shovels, loaders, and hauling equipment can be used underground.

A major portion of the *Estonian oil shale* is currently mined by the room-and-pillar system, which requires a great deal less manual work than former longwall methods.

Longwall Method. Most underground mines in the Estonian oil shale area used the longwall method before the room-and-pillar system was introduced in 1958. Typical face lengths were 85 – 90 m. To facilitate mechanical cutting along the face, holes had to be drilled and explosives used. A lot more manual work is required to load the material, advance the conveying system, and support the roof. Longwall mining was also used in a few oil shale mines in Scotland.

New Underground Mining Concepts. In extremely thick oil shale deposits, low-cost large-capacity mining methods such as block caving, stoping and gobbing, space stoping, or a modified room-and-pillar method can be used.

5.2. Above Ground Mining

Above Ground mining is highly efficient, and mining losses are kept to a minimum. The most important methods are open-pit and strip mining. These methods are employed at present to work oil shale deposits in Estonia, Manchuria, and Brazil.

Open-Pit Mining. In an open-pit mine, the overburden is stripped off and hauled away to the waste dump in the voided part of the mine by trucks, pit railways, or belt conveyors. The exposed shale is then mechanically excavated and hauled off. The type of equipment used depends crucially on the geology and mechanical properties of the overburden and shale. Soft materials are conveniently removed by bucket-wheel excavators, whereas hard rocks require power shovels and scrapers and possible blasting.

Strip Mining. Strip mining employs dragline excavators that are shifted only rarely from one location to another. They discharge the excavated overburden promptly by swiveling the full bucket from the trenchlike digging area to the dumping area. No overburden hauling facilities are thus required. Shale is excavated from a strip mine in the same way as from an open pit.

Newly designed mining systems that are more cost effective or permit opencast mining of deeper shale deposits are currently being tested.

6. Shale Preparation

Unless the oil shale is used in situ, it must be upgraded before it can be retorted, gasified, or burned. This means at least that the shale must be classified and crushed. Raw shale with high moisture content may also need to be dried. Suitable dryers for high throughputs are well known (e.g., rotary kilns and entrained-flow dryers).

High-grade oil shales have a lower density than low-grade shales, so that they can be enriched in settlers, sink and float equipment, etc. Selective comminution may be another method to be considered, because rich and lean shales often differ in their grinding properties. There is doubt whether enrichment processes can ever become important, because oil shale is a low-grade raw material and processing may be expected to be cost effective only for extremely high throughputs. Oil shales that are so lean they must be enriched are of no economic interest in the short or medium run.

The *Brazilian* Petrosix retort – a shaft furnace – requires a feed with a particle size of 6 – 80 mm. To prepare it for this retort, the oil shale is gently comminuted and screened.

At Fushun (People's Republic of China), oil shale containing less than 3.5 wt % oil is rejected in the mine. Fine-grained shale is screened off and kept out of the retorts as well.

In Estonia, shale enrichment is relatively common because of the special quality of the kukersite shale. Settlers, slurry separators, air classifiers, flotators, and selective comminution equipment are used for this purpose.

7. Pyrolysis

7.1. Definition and Kinetics

In pyrolysis an organic substance is thermally converted by heating in the absence of oxygen. The pyrolysis of kerogen yields shale oil, reaction water, noncondensable gases, and a carbonaceous residue. The inorganic mineral matrix of the shale is retained in the form of ash or spent shale.

Oil shale pyrolysis begins at relatively low temperature (200 °C), but proceeds more rapidly and more completely at higher temperature (500 – 600 °C). Various gases and vapors are produced from the organic materials during the process.

Once gas generation ceases, an unburned carbonaceous residue remains that sticks to the solid inorganic material of the original oil shale. The gases and vapors leaving the retort are cooled to condense the oils and reaction water.

The kinetics of oil shale pyrolysis has been investigated, and the following reaction sequence for Colorado oil shale has been described [24, 25]. First, kerogen is broken down to produce bitumen, gases (condensable hydrocarbons and other organic compounds), and a carbonaceous residue. The bitumen is then decomposed further in a second step, yielding other gaseous and liquid hydrocarbons.

The kinetics of shale oil formation can be described by two sequential first-order reactions [26]:

$$K \xrightarrow{k_1} B + G_1 + C_1$$

$$B \xrightarrow{k_2} A + G_2 + C_2$$

where K is kerogen, B is bitumen, A denotes oil, G_1 and G_2 are gaseous products, C_1 and C_2 carbonaceous residues, k_1 is the rate constant of kerogen decomposition, and k_2 is the rate constant of bitumen decomposition.

Kinetic equations were derived for the decomposition of kerogen and for the formation of bitumen and oil. The corresponding rate constants were determined by integrating and combining the equations: above 400 °C the decomposition of kerogen is the rate-determining step; below 400 °C the decomposition of bitumen is rate determining. The low activation energy of 44.56 kJ/mol for the first reaction proves that kerogen decomposition involves the cleavage of relatively weak chemical bonds, whereas the activation energy of 177.6 kJ/mol required for the second reaction indicates that much stronger chemical bonds must be broken for bitumen decomposition.

Many studies have focused on the amount of gaseous and solid products resulting from each of the two reactions. One interpretation is that most of the carbonaceous residue is formed during the kerogen-to-bitumen conversion, whereas most of the gas is evolved during the bitumen-to-oil conversion, although small amounts of hydrogen sulfide and carbon dioxide precede hydrocarbon evolution (e.g., with Australian kerogen).

Since liquid hydrocarbons historically have had a higher economic value than gaseous or solid hydrocarbons, special attention has been directed toward maximizing oil yields. Oil yield from shale is decreased by coking, cracking, and combustion.

Coking is a liquid-phase condensation or polymerization reaction that results in the fusion of two or more molecular species, with ultimate formation of a high-polymer carbonaceous product and minor amounts of lower molecular mass gases and liquids.

Cracking involves vapor-phase bond-fission reactions that eventually lead to a distribution of smaller molecular units plus, possibly, a certain amount of carbonaceous residue. An important difference between coking and cracking is the molecular mass distribution of the products. Cracking reactions lead to a higher yield of hydrocarbons with a lower molecular mass (CH_4, C_2H_4, etc.), whereas coking reactions produce mainly a high-polymer carbonaceous residue.

Combustion is for the most part a mass-transfer-limited reaction that occurs if both oxygen and organic material are present at high temperature. Under fuel-rich conditions, combustion and cracking become closely related. Oil that is not combusted at these temperatures will undergo thermal cracking. In addition, oxygen is a good initiator of free-radical cracking, because it is a free radical.

The physical description of oil production from bitumen is as follows. Bitumen decomposes to oil, gas, and a carbonaceous residue. The oil exists in two physical states: liquid oil and oil vapor.

In the liquid phase, oil is subjected to two competing processes: vaporization, which leads to the production of oil, and coking, which leads mainly to a carbonaceous residue. To obtain high oil yields, the residence time of the oil in the liquid phase must be kept short to minimize oil coking. This can be accomplished either by high retorting temperatures and high gas sweep rates or by high heating rates. The reason for the in-

crease in oil yield with increasing *heating rate* is twofold. As a result of the kinetics of the reaction the average temperature of the oil evolved depends on the heating rate. Higher heating rates result in the oil being evolved at a higher average temperature; hence a smaller fraction of the oil remains in the liquid phase. Furthermore, higher heating rates increase the rate of oil vaporization and thus reduce liquid-phase coking. These two effects increase oil yield by reducing coke formation.

High retorting temperatures increase liquid yield by decreasing oil coking. However, too high a temperature will decrease oil yield because of secondary vapor-phase oil cracking.

Retorting conditions for oil shale have a significant effect on the properties of the oil. For Colorado oil shale, a decrease in the yield of oil due to *coking* causes an increase in the H/C ratio and a decrease in the nitrogen content of the remaining oil. Associated with the increase in hydrogen content is a lower density. In contrast to oil coking, oil *cracking* results in a decrease in the H/C ratio and an increase in the nitrogen content of the remaining oil. Also, the pour point of the oil (which depends on the content of long-chain alkanes) drops sharply with increased degree of cracking.

These differing trends with coking and cracking can be easily understood. Oil coking is most likely caused by condensation and polymerization reactions involving nitrogen-containing compounds. Since nitrogen is contained mostly in aromatic rings, removal of these hydrogen-deficient compounds causes an increase in the H/C ratio. In addition, the lower temperatures for oil evolution associated with *coking* cause a decrease in the alkene-to-alkane ratios in the oil.

In contrast, in *oil cracking*, alkanes are converted to alkenes and aromatics. Thus, the H/C ratio in the remaining oil decreases, a result that is most likely a function of both temperature and conversion. In addition, the nitrogen content of the oil increases with cracking, because the aromatic nitrogen compounds in shale oil are resistant to cracking. As the alkanes and alkenes are partially converted to gases, the nitrogen compounds are selectively concentrated. A 50 % conversion, for example, results in roughly a doubling of the nitrogen content.

7.2. Above Ground Retorting Processes

Commercial oil shale retorting plants have been operated in various countries for more than a hundred years, but their design, low thermal efficiency, and insufficient capacity make them noncompetitive under current economic conditions.

A simple design and reliable operation, along with low capital and operating costs, are prerequisites for a modern oil shale retorting process to become an economic success. Yields reaching 100 % of the Fischer assay are desirable; the process should not depend on additional external energy supplies, should be continuous, and should be environmentally benign.

Classification of Retort Types. Retorting of oil shale entails the heating of a large mass of nonproductive rock to pyrolyze the kerogen. The principal aim in retort design is to accomplish this heating at low cost by using the least amount of energy. Thus, the method of heat transfer to the raw shale provides a convenient way to classify retorts. Three different methods can be distinguished.

Class 1 (Indirectly Gas-Heated Retorts). Heat is transferred to the shale through a wall. A simple example of this type of retort is the Fischer assay apparatus used to measure shale quality. The Pumpherston retort, the Salermo retort, and the Davidson retort are typical examples of class 1 oil shale retorts. This retort type is not being developed further because of the small unit capacity, expensive heat transfer, and low thermal efficiency of the method.

Class 2 (Directly Gas-Heated Retorts). Heat is transferred by passing hot gases directly through the shale, mostly in a vertical shaft kiln. This class can be subdivided into the following:

In class 2 a (internal combustion mode), hot gases are generated by the combustion of residual carbon or a portion of the hydrocarbon product within the retort. Examples of this type are the Fushun retort, the Paraho retort, the Lurgi Spuelgas kiln, the Union A retort, the Superior retort, the Kiviter retort, and the Lurgi–Schweitzer kiln. This method of heat transfer is inexpensive and allows relatively low-cost retort construction. However, the yield of hydrocarbons is lower than with other methods, and the retort gas is largely diluted by combustion products and nitrogen from the combustion air.

In class 2 b (external combustion mode), the gases are externally heated in a gas heater (e.g., by combustion of a portion of the hydrocarbon product). Examples are the Union B retort, the Petrosix retort, and the Lurgi tunnel oven. These retorts require an external gas heater, which may be expensive and lead to operating problems. Hydrocarbon recovery is higher than in the case of class 2 a retorts, and the gas from the retort has a high calorific value because it does not contain combustion products. In contrast, the thermal efficiency is lower because the residual carbon in the spent shale is not utilized in most cases.

Common to most class 2 retorts are (1) their inability to process shale fines and (2) increasing problems with uniform shale-size and gas distribution through the retort shaft area with increasing retort size (i.e., capacity).

Class 3 (Directly Solids-Heated Retort). Heat is transferred by mixing hot solid heat carriers with the fresh shale. This method involves a somewhat more complex heat carrier circulation system but has the advantages of:

- High oil yields due to high heating rates (flash retorting),
- Undiluted product gas and vapors,
- Direct use of spent shale carbon as fuel if spent shale is used as the heat carrier,
- Use of fine granular shale with no minimal size restriction, and
- Easy scaleup to very large unit capacities.

Examples of this retort type are the Lurgi – Ruhrgas retort, the Tosco II retort, the Galother retort, and the Chevron STB retort.

7.2.1. Class 1 (Indirectly Gas-Heated) Retorts

Class 1 retorts were used mainly before 1960. They generally have the advantage of producing an undiluted gas and vapor product stream with few restrictions on the grain size of the feed oil shale. However, a low unit throughput, high specific investment cost, and a low thermal efficiency, have hindered further development of retorting technologies based on this heat-transfer principle.

Pumpherston Retort. The Pumpherston retort emerged as the most successful retort of the Scottish oil shale industry in the second half of the 19th century. It was also widely applied, for example, in France (more than 200 retorts in operation by 1905) and Spain (Puertollano). Pumpherston retorts were mounted in huge banks because of their limited throughput. In 1947 the Westwood plant in Scotland consisted of two benches with 52 retorts each, backed up to one another in pairs. Total capacity of the plant was 1040 t/d (i.e., 10 t/d per retort). Shale processed in this plant in 1947 had an average oil yield assay of 21.4 gallons per ton (ca. 90 L/t) of feed. Daily oil production was about 530 barrels.

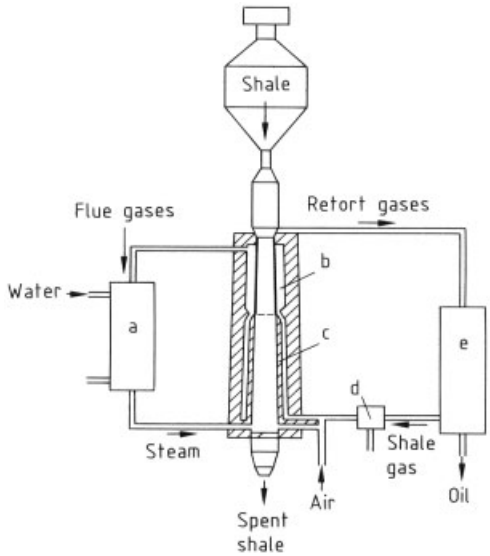

Figure 1. Single Pumpherston retort (1947)
a) Waste-heat boiler; b) Cast-iron retort; c) Brick retort; d) Scrubber; e) Condenser

Figure 1 is a schematic of a single Pumpherston retort in 1947. Above each pair of retorts was a large hopper carrying a 30-h oil shale supply for the two retorts. This fed into the small shale hopper indicated at the top of the retort in Figure 1. Each retort consisted of a cast-iron section 4.3 m long and an oval-shaped and tapered brick section 6.1 m long. The cast-iron part of the retort was surrounded by masonry enclosing a free space for heating the retorts externally by burning noncondensable shale gas. The crushed shale entered the retort at the top, flowed downward by gravity, and emerged at the bottom through a discharge hopper with water-sealed doors. As indicated in Figure 1, shale gas

and air were supplied, and the gas was burned at the bottom of the retort. The combustion products passed upward around the cast-iron section to preheat the entering shale. Residual heat in the flue gas was recovered in the waste-heat boiler. Steam from this waste-heat boiler together with a regulated amount of air was injected into the interior of the retort at the bottom. The steam generated ammonia and swept the oil vapors upward. In addition, at the high temperature maintained in the lower part of the retort (1000 °C), air and steam reacted with the residual carbon to supply much of the heat required in the retorting process. Virtually all of the residual carbon was consumed.

This injection of air at the bottom of the retort changed the process from an indirectly gas-heated (class 1) to a partially directly gas-heated (class 2 a) mode and increased each retort's throughput capacity from ca. 4 to 10 t/d, permitting the use of lower-grade oil shale. Flow of shale through the retort was controlled by the rate of removal of spent shale at the bottom of the retort. Heat was recovered from the exiting spent shale. Spent shale was discarded in large piles called "bings," still in evidence throughout the area of oil shale processing in Scotland. Although some spent shale was made into brick by adding lime and steam-curing the resulting bricks, no major use for the spent oil shale was ever found.

Gases and oil vapors leaving the top of the Pumpherston retort passed through preliminary water sprays and then through water-cooled, multiple condensers. Most of the ammonia remained in the condensed water, but the gases were further scrubbed first with water to absorb the rest of the ammonia and then with gas oil to remove the remaining naphtha. The condensed oil and ammonia-containing water separated easily. The water was processed in an ammonia recovery plant on site. The uncondensed gas used in retorting amounted to about 290 m^3/t (STP) of shale and had a heating value of only about 5600 kJ/m^3, primarily because of dilution with inert gases and carbon dioxide. However, the Westwood Pumpherston plant was energy self-sufficient because the residual carbon was burned in the retorts. Each tonne of shale processed produced about 0.5 barrel of oil and 12 kg of ammonium sulfate. For processing, the oil was shipped to oil refineries serving several retort plants. In [27] the following product distribution is given (in wt %):

Motor spirit (octane number 57)	19.00
Diesel fuel	52.00
Paraffin wax	9.00
Tar (including acids and bases)	11.60
Coke (1 % ash)	2.75
Resin	1.00
Loss (by difference)	4.65

All products found a ready market in Scotland. Acid was recovered from the acid tars and used in ammonium sulfate production. The tars produced were used as refinery fuel. The Scottish oil shale industry closed down in 1963 due to the competition of cheap, imported crude oil, and the depletion of high-grade oil shale reserves.

Davidson Retort. The Davidson retort (see Fig. 2) consisted of a cylinder 1.2 m in diameter and 23 m long, sloped slightly from shale inlet to shale outlet. This cylinder rotated slowly in a brick-lined heating chamber. Heat was generated in a combustion chamber at the outlet end of the retort. The gaseous combustion products

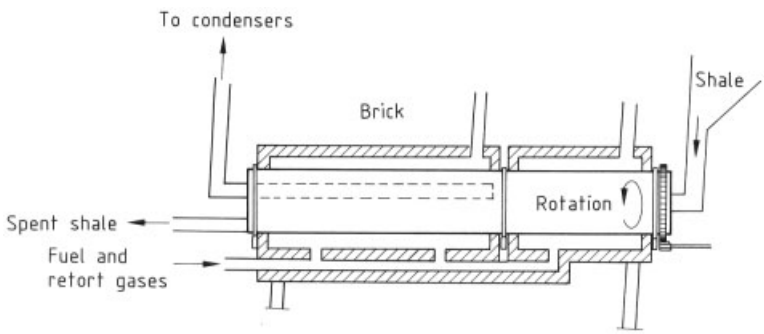

Figure 2. Davidson retort

were passed along a flue to a row of ports into the heating chambers around the rotating retort. Heat admitted to the heating chambers was controlled at the ports. The Davidson retorts normally operated with two temperature zones: a drying and preheating section kept at ca. 400 °C in the first third of the retort, and a retorting zone kept at ca. 630 °C. The capacity of each Davidson retort was about 27 t/d. This feed rate was not very flexible. If too little feed was admitted, the charge would slide instead of tumble, and the shale would tend to stick together in cakes. If too much was fed in, the evolution of oil would not be complete. Spent shale was discharged through a double-door arrangement to prevent air entry into the retort. Each Davidson retort produced about 38 barrels of oil per day at a recovery of about 90 % of Fischer assay. The electrical energy needed to turn the huge retort tube and to meet other requirements was 8 kW · h/t. Gas energy input to the Davidson retort was 2600 kJ/t. In 1947 the carbon on the Davidson spent shale was discarded because of the cheap coal available.

Davidson retorts had been operated in Estonia as early as 1932, and Davidson had several patents on variations.

Salermo Retort. In 1935, SATMAR started retort operation near Ermelo, Transvaal. In the Salermo retort, small oil shale particles were progressively shoveled through a series of semicircular troughs heated from the underside. This retort, called Salermo in South Africa, probably originated from the retort patented by P. M. Salerni in 1928.

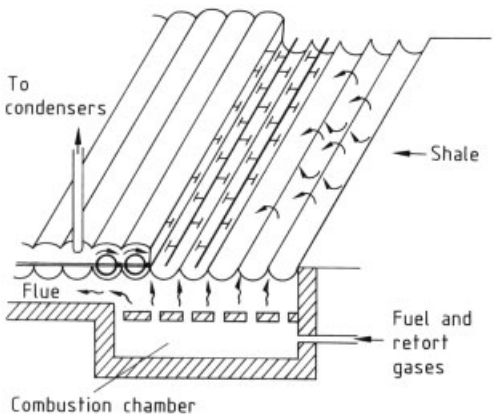

Figure 3. Trough and shovel operation in Salermo retort

Figure 3 is a schematic view of the troughs to illustrate shale travel and operation of the paddles. The SATMAR Salermo retort consisted of 36 semicircular mild-steel troughs 0.4 m in diameter and 2.3 m long. These troughs formed the bottom of the retort. Side, end, and top cover plates isolated the troughs from the atmosphere. A rod with 16 T-shaped paddles attached (Fig. 3) revolved in each trough.

Each rod was rotated from the outside at 5 rpm. Shale entered the retort, and was agitated and moved from trough to trough by paddle rotation. Below the first nine troughs was a combustion chamber leading to flues that distributed heat along the length of the retort. Noncondensable shale gas was the primary fuel, supplemented by producer gas. As the oil shale was shoveled from trough to trough, it was gradually heated. The first troughs were dryers and preheaters. By trough 14 the shale temperature had reached 300 °C, and at the discharge it reached 450 – 500 °C. Small amounts of gas were added and burned along the flues to help raise and maintain the higher temperature toward the discharge end. Evolving gases were led off to condensers for separation.

Each Salermo retort processed 70 t/d on average. During an operating day, each retort produced about 90 barrels (13 t) of oil. The electric power required was about 3.5 kW · h/t. The total heat input to the retort, primarily from 37-MJ/m^3 retort gas but supplemented by producer gas, was about 2.1×10^6 MJ/t of shale. A retort operated continuously for 140 d and was then cooled, overhauled, and reheated, a process that took about 20 d. Recovery was about 90 % of Fischer assay, or about 50 gallons per ton (210 L/t) of oil. Similar to the spent shale from the Davidson retort, Salermo spent shale contained large quantities (up to 40 %) of carbon. Because waste coal was mined with the shale, no effort was made to recover the energy content of this carbon.

There is still one Salermo retort in operation in Clerbagnoix (France) that produces high-sulfur shale oil. This oil is used for the production of ichthyol pharmaceuticals.

In 1947 the SATMAR retorting plant consisted of four Davidson retorts, ten Salermo retorts, an oil shale crushing and screening plant, and seven gas producers gasifying coal to supply fuel gas. The Davidson retorts received ≥25-mm material, whereas all smaller material

was fed to the Salermo retorts. No fines were discarded. Both the Davidson and the Salermo retorts were externally heated by using noncondensable gases from the oil shale supplemented by producer gas generated from coal. The company persisted and even paid dividends until it-mined itself out of existence in about 1962, by which time it had exhausted its oil shale (torbanite) resources.

7.2.2. Class 2 (Directly Gas-Heated) Retorts

Class 2 retorts have found worldwide application because of their simple shaft kiln design.

Class 2 a Retorts. Class 2 a retorts are heated by hot flue gas produced either by combustion of spent shale carbon through injection of air into the lower retort section or by combustion of some product gas and oil in a combustion chamber adjacent to the retort. They operate with *internal combustion*.

Paraho Retort. The Paraho retort (see Fig. 4) was developed in the 1970s on the basis of available lime kiln technology and the earlier gas combustion retort by the Paraho Development Corp. (United States). It is a typical vertical shaft kiln.

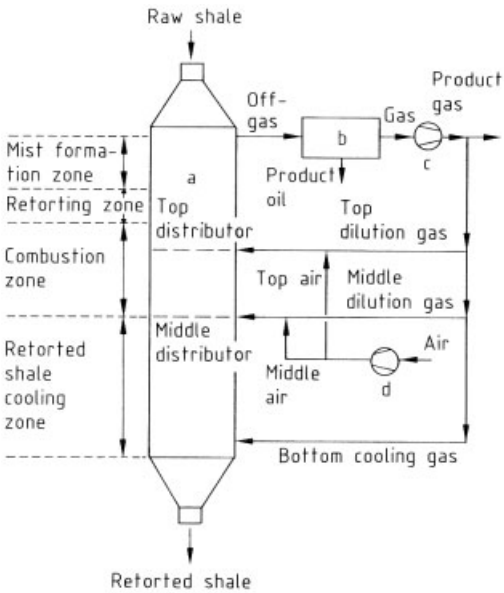

Figure 4. Paraho retort
a) Retort; b) Oil – gas separator; c) Recycle gas blower; d) Air blower

Raw shale, crushed and screened to a feed size of 0.6 – 8 cm, is conveyed to the feed system on top of the retort. In a continuous moving bed, the shale flows downward consecutively through the mist formation, retorting, combustion, and cooling zones. As the shale descends, heat is efficiently exchanged with a countercurrent flow of recycle gas, which is introduced into the retort at different levels by three specific-purpose gas – air and gas distributors. Near the top of the retort (in the *mist formation zone*), the ambient-temperature shale is warmed by rising hot oil vapors and gas which, in turn, are cooled to form an oil mist that is entrained in the gas.

As the shale continues downward into the *retorting zone*, it is heated by rising hot gases until it reaches about 482 °C, where kerogen is pyrolyzed into oil vapor, gas, and carbonaceous residue. This residue, which remains on the shale, is utilized as a fuel when the shale bed descends into the *combustion zone*, where combustion is achieved by the injection of air (diluted with recycle gas) through the middle and top distributors. This combustion provides the heat for the retort process.

Below the combustion zone the descending shale is cooled to about 149 °C by passage through a stream of cool recycle gas, which is injected into the shale bed by the bottom gas distributor. The cooled retorted shale is discharged at the bottom through the hydraulically operated Paraho grate mechanism, which regulates the rate and uniformity of descent of shale through the retort. Displaced shale drops into a collection hopper and passes through a rotary seal and out of the retort, where it is conveyed to the retorted shale disposal site.

Yield and properties of Colorado oil shale processed by the Paraho retort may be summarized as follows:

Yield	
Shale assay	27.4 gallons per ton (114 L/t)
Retort production	24.4 gallons per ton (102 L/t)
Percent of assay	89 – 94 %
Physical properties	
Specific gravity	21.4 °API (→ Oil Refining)
Density	925 kg/m^3
Viscosity (54 °C)	83 SUS (Saybolt units · second)
Viscosity (99 °C)	48 SUS
Pour point	29 °C
Water	0.3 wt %
Sediment	0.1 mL/100g
Elemental analysis	
Carbon	84.8 wt %

Hydrogen	11.4 wt %
Nitrogen	2.0 wt %
Sulfur	0.6 wt %
Oxygen	1.2 wt %

Lurgi Spuelgas Retort. The Lurgi Spuelgas retort, which had been successfully applied to the carbonization of brown coal briquettes and bituminous lumpy coal in more than 250 units, was modified slightly to make it suitable for retorting oil shale. The kiln (Fig. 5) is of the double rectangular-shaft type and incorporates a drying, retorting, and residual carbon gasification zone through which shale travels from top to bottom. The shale is dried and retorted by recycling product gas (Spuelgas) in counterflow to the shale. The Spuelgas is internally preheated by cooling spent shale in the lower shaft section and heated further by mixing with hot gas from the combustion chamber. This kiln yields a highly diluted gas.

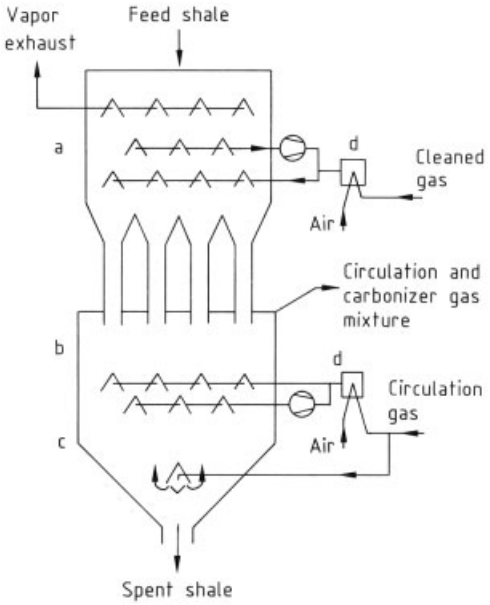

Figure 5. Lurgi Spuelgas retort
a) Drying zone; b) Carbonization zone; c) Cooling zone; d) Burner

In 1944 a 3000-t/d oil shale retorting plant was engineered for Spanish Puertollano oil shale, but Pumpherston retorts were finally installed. In 1958 a 400-t/d kiln including a residual carbon combustion zone was installed near Kisangani for Zaire oil shale. Besides shale oil, spent shale was used by the cement industry. The capacity of a kiln is ca. 500 t/d. Higher capacities involved increasing problems with the uniform distribution of gas and solids through the shaft area of the kiln. Oil yields up to 93 % of the Fischer assay can be obtained by this retort.

Union A Retort. The Union A retort was developed by the Union Oil Company of California in the 1950s. It is also a vertical shaft kiln involving a novel rock pump designed to force raw oil shale up into and through the body of the retort.

The retort (Fig. 6) is shaped like an inverted cone. it admits a crushed shale feed of 6 – 100-mm particle size. In this system, shale moves through the reactor from the bottom upward. A rugged, reliable reciprocating rock pump forces cold feed into the bottom and then further up the retort. Spent shale is discharged over the top of the reactor shell.

Gases flow through the retort downward in countercurrent to the shale. The hottest part is the combustion zone at the top of the retort. There,

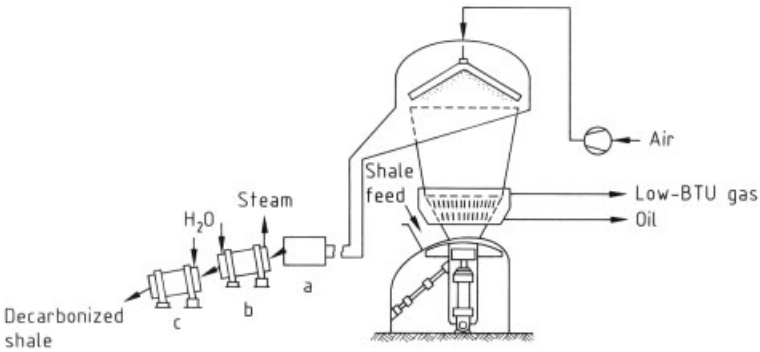

Figure 6. Union A retort
a) Crusher; b) Cooler; c) Wetting

near the solids overflow, air is blown into the reaction bed to burn the carbonaceous residue on the spent shale particles. As it moves up the retort, the cold shale is first preheated and then retorted, and finally the carbonaceous material is burned to provide all the heat required for the process. The oil and gases are withdrawn from the bottom of the conical retort.

To date, only one retort of this type with a 1.7-m feed piston and a maximum cone diameter of 5.2 m has been operated. It reached a throughput of ca. 1000 t of shale per day and a yield of about 85 % of the Fischer assay. The retort gases have a low calorific value of 4000 kJ/m^3 (STP).

Dravo/Superior Circular Grate Retort. The Dravo/Superior Circular Grate Retort was initially developed by Superior Oil Co. in the United States during the 1970s. The retort was based on available circular grate technology developed for sintering iron ore pellets. In the 1980s, Dravo took over this development and tested the process in its 220-t/d pilot plant in Cleveland, Ohio.

In the circular grate retort (Fig. 7), hot recycle gas is combusted with air inside the retort. Both the recycle gas and the combustion air are preheated by passage through the traveling spent shale bed, and the spent shale is cooled at the same time. The recycle gas has a low calorific value because of significant dilution by nitrogen from the combustion air. This low-energy gas is burned in special burners above the entering shale bed. The combustion products at 725 – 775 °C carry heat into the shale bed to generate and drive out oil and gas. The oil and gas pass with the gas flow out of the shale bed to the oil recovery system. The traveling grate then moves the shale bed into the *carbon recovery area*. Preheated air (≈200 °C) that directly enters the top of the spent shale bed is used to burn residual carbon. This combustion heats the shale residue. The combustion products join the recycle gas stream. After oil removal, recycle gases at about 100 °C pass through the heated shale residue on the moving track and are heated to 275 – 375 °C. Final cooling of the spent shale is performed by passing ambient air through the traveling bed. Cooled spent shale leaves the retort through a liquid seal. Although this retort is flexible with regard to feed shale size and offers no major scaleup problems, a large-scale unit has not yet been built.

Fushun Retort. The Fushun retort is still being used in the Chinese oil shale industry. Because it has been in operation for 50 years, it has been modified gradually, but the Fushun retort

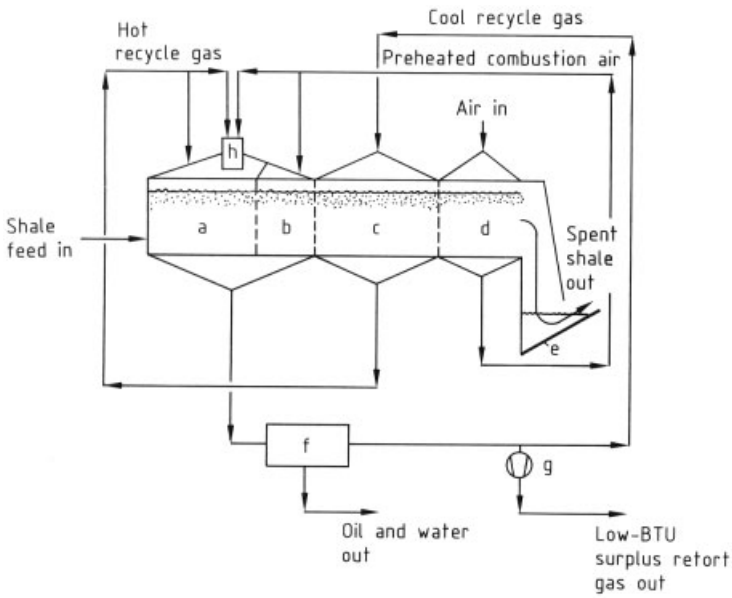

Figure 7. Dravo/Superior retort (direct heating mode)
a) Heating; b) Carbon recovery; c) Cooling by recycle gas; d) Cooling by air; e) Water seal; f) Oil removal; g) Gas compressor; h) Direct-heating low-BTU burner

was initally modeled after an internally heated retort tested in Estonia. In operation, the Fushun retort strongly resembles the gas-combustion retort, a predecessor of the Paraho retort. The retort used at Fushun and at Maoming is a vertical cylinder about 3 m in diameter and 15 m high. Average oil yield of the shale processed in 1979 was about 6.7 wt % .

Crushed oil shale screened to discard particles <8 mm, which incurs a 15 % loss of fines, is fed into the top of the retort. Operation of the retort is continuous because the shale is progressively dried and heated as it moves downward, reaching retorting temperature in about the middle of the retort. Spent shale moves into the "gasification section," where the residual carbon is burned and gasified by reaction with hot air and steam. The temperature in this section is 700 – 800 °C. The hot gases travel upward to heat the incoming shale, while the remaining ash travels downward to heat the incoming air and water. It then drops into the "ash dish," is cooled to 70 °C, and is removed from the system.

The oil vapors and noncondensable gases leave the top of the retort at 90 – 110 °C and pass to the condenser system. The cooling and recovery system strongly resembles that described for the Scottish Pumpherston plant. In addition to oil, ammonium sulfate is also recovered here. A block of 20 to 25 retorts shares a single condenser system. The evolved gases supply heat to the retort and are also used as an energy source. The plant requires no external energy supply. The oil yield of the entire system is about 60 % of the Fischer assay. Each retort processes about 20 t/d of oil shale.

To obtain 1 t of shale oil at Fushun, the following inputs to the system are required:

Oil shale	40 t
Water	5 – 6 t
Electricity	150 – 170 kW · h
Steam	2 – 3 t

The oil shale industry in the People's Republic of China is the largest in the world.

Kiviter Retort. The Kiviter retort (Fig. 8) was designed in the Estonia to process kukersite shale, which tends to soften and cake during retorting. The size of the feed shale ranges from 25 – 100 mm. The Kiviter retort was apparently developed from the Pintsch retort, which is termed a gas-generation retort in the Soviet Union. The Kiviter retort is a vertical system internally heated by combustion of coke residue and noncondensable shale gas. In many respects it resembles the gas-combustion and thus the Paraho retort. Kiviter units with capacities of 180 – 1000 t/d have been put into commercial operation. The unit has also been tested successfully on low-yield oil shale, which permits faster throughput.

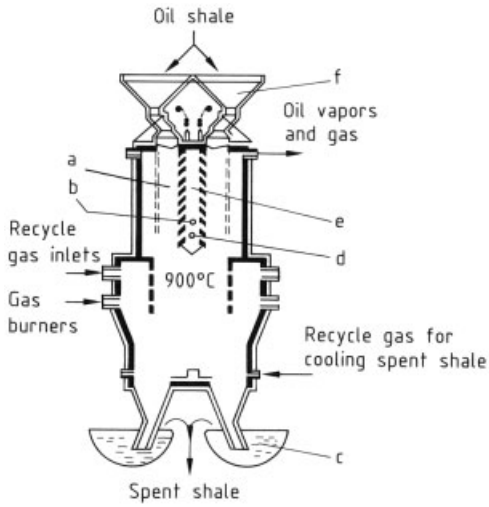

Figure 8. Kiviter retort
a) Oil shale retorting and evacuation chamber; b) Recycle gas inlet; c) Spent-shale discharge device; d) Gas burner; e) Heat carrier, distribution chamber; f) Charging device

Operation of the Kiviter retort is continuous. Raw oil shale enters from the top (Fig. 8) and is heated to the temperatures required to decompose the organic matter. Heat is provided by the rising gases supplemented by recycle gas burned in the heat carrier preparation chamber. The combustion products pass sidewards through the descending shale and carry the oil vapors and evolving gas with them into the collection chamber. From there they are delivered to the condensing system. Additional recycle gas and air are admitted to the chambers near the 900 °C point and heat the shale residue to burn off the coke. Recycle gas entering the bottom of the retort cools the spent shale, which then leaves the retort through a water-sealed discharge system. The Kiviter retort operates on Baltic shales enriched at the mine to an oil yield of 21.3 wt % (233 L/t) and produces the following materials (in kilograms per tonne of wet feed):

Input	
Dry shale	910
Associated water	90
Steam	50
Air	430
Total input	1480
Output	
Shale oil	180
Gas	560
Process water	160
Spent shale	580
Total output	1480

The Kiviter process recovers about 75 % of Fischer assay oil. The electrical energy required by the process is 18 kW · h per tonne of wet feed shale. The gross calorific value of the nitrogen-diluted product gas is 3534 kJ/kg. In the Soviet Union special burners were designed to use this low-energy gas. A large-scale 1000-t/d Kiviter retort was built in Estonia in the late 1970s.

Lurgi – Schweitzer Process. The Lurgi – Schweitzer process was developed in Germany as a low-cost retort during World War II. The principle of this process is the retorting of lumpy low-grade oil shale in a descending gas stream. The retort consists of a cylindrical shaft with no refractory lining and is operated batchwise. After charging, the upper shale layer is ignited like a pipe, and air is sucked through the shale bed from top to bottom, whereby the combustion zone moves slowly downward. The hot gases retort the shale below the combustion zone so that the heat required for the process is produced only by combustion of the residual carbon. On completion of the retorting process, the retort is tilted, discharged, and subsequently recharged. A commercial plant comprising 38 units was installed in 1942 in Württemberg, Germany, and operated until 1949. This process gives a maximum oil yield of 95 % of the Fischer assay.

Class 2 b Retorts. Class 2 b retorts are heated by hot recycled product gas that is reheated in an external (indirectly heated) heat exchanger. These retorts are classified as operating with *external combustion*.

Union B Retort. The Union B retort is very similar to the directly heated Union A retort (Fig. 6). However, it is heated by circulating gases deriving their heat from a recuperative system, whereas the heat required in the Union A retort is generated by internal combustion.

Screened shale particles are fed to two reciprocating rock pumps via a shale oil siphon, and delivered to the bottom of the retort. The shale then moves evenly up the conical part of the retort and is extracted by a special rotating rake from the top. Thereafter, the residue flows into a collecting and cooling system. The top of the retort is covered by a gastight hood that permits an operating pressure of about 1 bar above atmospheric pressure. As in the Union A retort, the condensed oils and vapors are withdrawn from the bed through many vertical slots arranged along the perimeter at the bottom of the cone.

The Union B retort reaches oil yields of about 100 % of the Fischer value and the gas produced has a calorific value of ca. 33 500 kJ/m^3 (STP), because it is not diluted by combustion products. However, the overall thermal efficiency of the process is relatively low (ca. 69 %) because the energy of the carbon in the spent shale (about 4 wt % in the case of Colorado oil shale) is not used. Therefore, Union Oil investigated the addition of a circulating fluidized-bed (CFB) combustor for the spent shale, to increase the thermal efficiency of the overall system.

In 1983, Union Oil commissioned a 10 000 barrel/d Union B retort in Colorado. Because of problems resulting from excessive decrepitation of spent shale in the upper part of the retort, the retort is operated at a lower than design capacity (60 – 80 %). In 1989, however, a high onstream efficiency of ca. 90 % seems to have been achieved. Operation was stopped in 1991 due to currently low crude oil prices.

Petrosix Retort. The Petrosix retort, which was developed in Brazil, is shown in Figure 9. The retort is externally heated to maximize sulfur recovery from the pyritic Irati shale and to avoid agglomeration and sintering problems experienced with directly gas-heated retorts.

Shale sized $\leq$15 cm and blended to uniform grade enters the top of the retort. Hot recycle gas enters the middle of the retort to heat and drive off oil and gas. Hydrogen sulfide is also generated. Heated shale travels downward and is cooled by recycle gas entering the bottom of the retort. The flow rate of the shale through the retort is controlled by the grate discharge mechanism. Spent shale leaves the retort at about 150 °C. It passes through a sampling system that measures discharge flow rate and selects a sample for Fischer assay. The results of the Fischer assay are used as input to the process control sys-

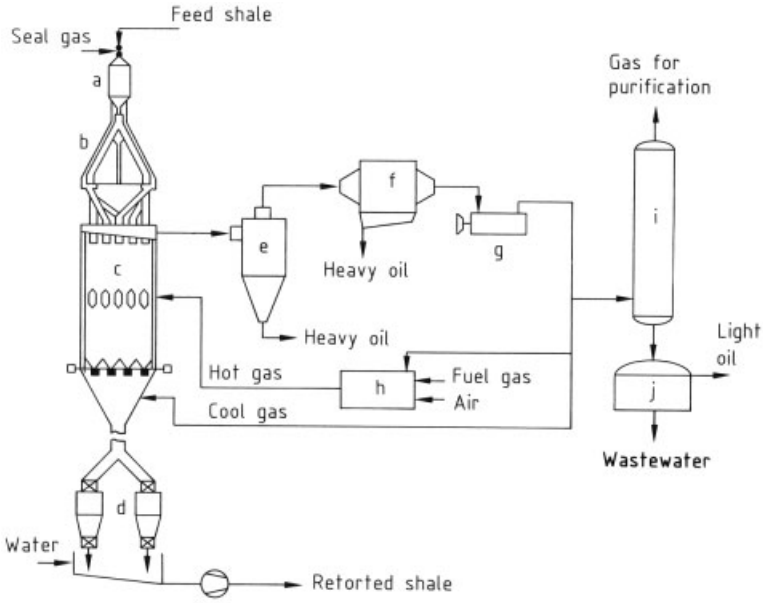

Figure 9. Petrosix retort
a) Feed hopper; b) Shale distributor; c) Retort; d) Seal system; e) Cyclone; f) Precipitator; g) Compressor; h) Heater; i) Condenser; j) Separator

tem. Processed shale enters one of two depressurization vessels operated in parallel. While one is being filled, the other is discharging its load into a mechanism that wets and cools the shale. The shale leaves the retort and is discarded in a water slurry. The residual carbon is not used at present.

Recycle gases carrying the evolved products travel through a *recovery system* that starts with two cyclones to remove dust particles and heavy-oil mist. This operation is followed by electrostatic precipitation. The gases are then compressed, and the recycle gas streams required by the retort are separated. The balance of the gases passes into a refrigerated condenser that separates light oils. The gases are then scrubbed with sodium carbonate in an absorption tower to remove mercaptans and sulfur dioxide. The gases subsequently enter a second tower, where hydrogen sulfide is absorbed by a stream of diethanolamine. The absorbed sulfur is recovered in a Claus unit. The cleaned gases have a high heating value and are suitable for pipelining.

The plant was designed to process 2200 t/d of mined shale and to produce 1000 barrels of oil per day, 17 t/d of sulfur, and 36 500 m^3/d of high-BTU fuel gas. It reached those production levels in 1975 during the demonstration phase, but ithas been operating at somewhat lower throughput in commercial production.

Lurgi Tunnel Kiln. In the Lurgi Tunnel kiln (Fig. 10), the oil shale is retorted in wagonettes moved forward periodically through a tunnel. The heat carrier used is retort gas, which is passed through tubular heaters and then through the shale bed in the wagonettes. The kiln is particularly suitable for very high-grade shales that temporarily become plastic during retorting (e.g., Estonian kukersite shale). Two kilns, each with a throughput capacity of 450 t/d, were installed in the late 1930s for the retorting of oil chalk in Germany, and another four kilns of the same size were erected later for the treatment of Estonian kukersite shale. The oil yield from this class 2 b retorts reaches 95 % of the Fischer assay.

7.2.3. Class 3 (Directly Solids-Heated) Retorts

Class 3 retorts appear to have a high potential for large-scale unit capacities because they can easily be scaled up, produce high oil yields, and

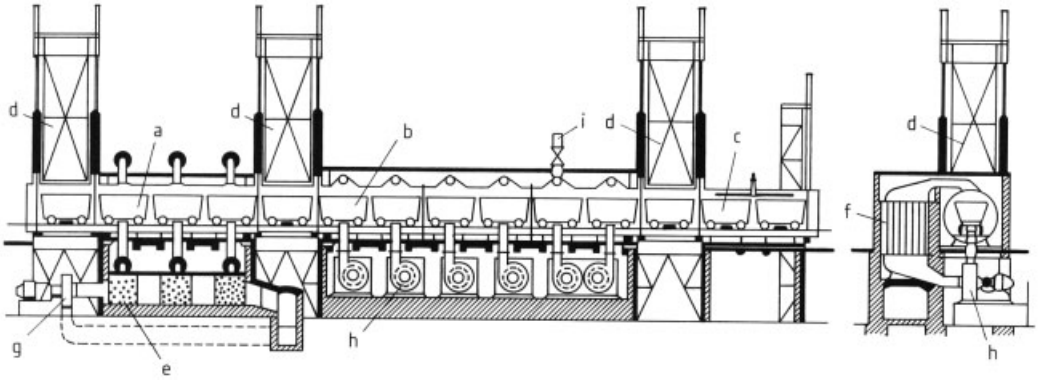

Figure 10. Lurgi tunnel kiln
a) Drying zone; b) Distillation zone; c) Cooling zone; d) Lock gates; e) Tube heater of drying zone; f) Tube heater of distillation zone; g) Circulation blower of drying zone; h) Circulation blower of distillation zone; i) Distillation gas outlet

use fine granular shale that must be rejected in most class 2 retorts.

Lurgi – Ruhrgas Retort. The Lurgi – Ruhrgas (LR) process was used on a commercial scale in the 1960s to carbonize fine-grained coal and to crack naphtha and crude oil to produce olefins. Its great flexibility also makes it suitable for retorting oil shale, tar sands, bituminous rocks, etc. A characteristic feature of the process is that the residue itself is used as a circulating heat carrier, which is reheated by combustion of the remaining organic substance during pneumatic conveying.

As shown in Figure 11, the fine-grained feed is mixed with the hot heat carrier from the collecting bin (b) and instantly heated in the screw mixer (c). The resulting gases and vapors are dedusted in the cyclone (e) and then fed to the cooling and condensing system (f). The mixture of fine-grained residue and heat carrier drops into the surge bin (d) and then flows in a moving bed to the bottom leg of the lift pipe (a). From there, it is lifted by preheated air into the collecting bin (b). On its way, the remaining carbon is burned off the spent shale and the material is reheated. Subsequently, the fine-grained material is separated from the gas and recycled to the mixer (c) to serve as a heat carrier. Surplus material is withdrawn from the system and cooled in a waste-heat recovery system. Flue gas flows through the cyclone (g) and finally through an aftertreatment system (h), (i) to recover its sensible heat and remove the remaining dust before it is discharged to the atmosphere.

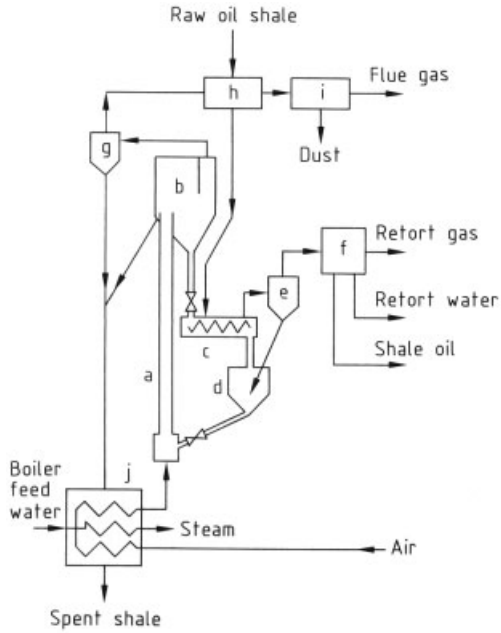

Figure 11. Lurgi – Ruhrgas retort
a) Lift pipe; b) Collecting bin; c) Screw mixer; d) Surge bin; e) Cyclone; f) Cooling and condensation system; g) Cyclone; h) Shale preheaters; i) Electrostatic precipitator; j) Fluidized-bed cooler

This process reaches oil yields of ca. 100 % of the Fischer assay, uses the carbon remaining on the retorting residue to heat the system, produces undiluted retort gas with a calorific value of ca. 38 000 kJ/m^3 (STP), and produces little NO_x because of the low combustion temperature in the lift pipe. The magnesium and calcium carbonates found in numerous oil shales are to a certain

Table 7. Oil shale retorting by the LR process (material balance from a pilot plant in kg/t)

	Shale origin		
	Schandelah (Germany)	Timahdit (Morocco)	Green River (United States)
Quantity, kg	1000.0	1000.0	1000.0
Moisture, wt %	3.5	5.9	0.4
Fischer oil content, wt %	5.6	8.0	9.4
Production, kg			
Fine-grained material from circulation system	464.3	510.5	266.7
Dust from flue gas	330.9	186.2	490.5
Dust from retort gas	13.5	55.8	9.7
Heavy oil, dust-free	28.8	50.6	62.9
Middle oil	18.8	23.5	26.5
Gasoline	5.4	5.5	10.7
Gas liquor	32.9	85.9	24.5
Dry distilled gas (without $\geq C_4$)	27.1	32.8	29.4
Carbonate CO_2 in flue gas	42.9	11.1	42.7
Burnt carbon	35.4	38.1	36.4
Total	1000.0	1000.0	1000.0
Yield in terms of Fischer assay, %	94.6	99.5	106.5

extent converted into oxides during the process. Under the operating conditions prevailing in the lift pipe, the oxides rapidly absorb sulfur dioxide and thus have a self-desulfurizing effect on the process flue gas.

Table 7 summarizes the material balances obtained in a continuous LR pilot plant from different oil shales.

The LR process was developed jointly by Lurgi and Ruhrgas at the beginning of the 1950s to flash-carbonize fine-grained coals. The first technical tests were run in Herten (Germany) in a pilot plant designed for a throughput of 12 t/d. The test program allowed for flash-carbonization of various coals to produce town gas, and for cracking liquid hydrocarbons to obtain olefins, by using sand as the circulating heat carrier. This application of the LR process was thereafter termed *sand cracker*.

Industrial-scale LR plants using oil and/or naphtha as feedstocks were built in Leverkusen, Leuna, Rosario, Chiba, and Lanchow. Commercial scale coal-based LR plants were constructed in Lukavac, Bottrop, and Scunthorpe.

Use of the LR retort for oil shale has been tested in a large number of pilot trials that confirmed the process for this feedstock.

Tosco II Retort. The Tosco II process (Fig. 12, see next page) uses circulating ceramic balls to provide the heat required to heat up and retort fine-grained oil shale. The shale and the hot ceramic balls are mixed in a closed rotary retort. The resulting vapors, the ceramic balls, and the spent shale are discharged from the retort. The heat carrier is then separated from the fine-grained spent shale by screening and is reheated in a furnace fired by retort gases. Because no air is admitted to the rotary retort, the resulting gases are not diluted by nitrogen and have a calorific value of about 38 000 kJ/m^3 (STP). The oil yield reaches ca. 100 % of the Fischer value. A demonstration plant operating between 1970 and 1975 was capable of retorting up to 1000 t/d oil shale.

Galoter Retort. The Galoter process (Fig. 13, see next page), developed in Estonia, combines the characteristics of the Tosco II and Lurgi – Ruhrgas processes. Dried oil shale of < 25-mm particle size is mixed with a heat carrier and fed to a rotary reactor. As in the LR process, hot residue is used as a heat carrier. The oil shale is heated to 520 °C in a rotary kiln retort to pyrolyze the organic substances. Thereafter, the retort gases are cleaned and cooled, and the condensed shale oil is withdrawn. Part of the hot spent shale leaving the combustion zone at ca. 800 °C is mixed with fresh shale and recycled to the retort. The retort gas produced from Baltic shale has a high calorific value of ca. 38 000 kJ/m^3 (STP), and the oil yield reaches 100 % of the Fischer value.

The Galoter process was developed in parallel with the Kiviter process because the fields of application of the two processes complement each other. Oil shale with particles sizes of 25 – 125 mm is retorted by the Kiviter process, whereas the Galoter process is tailored to fine-grained shale up to 25-mm diameter.

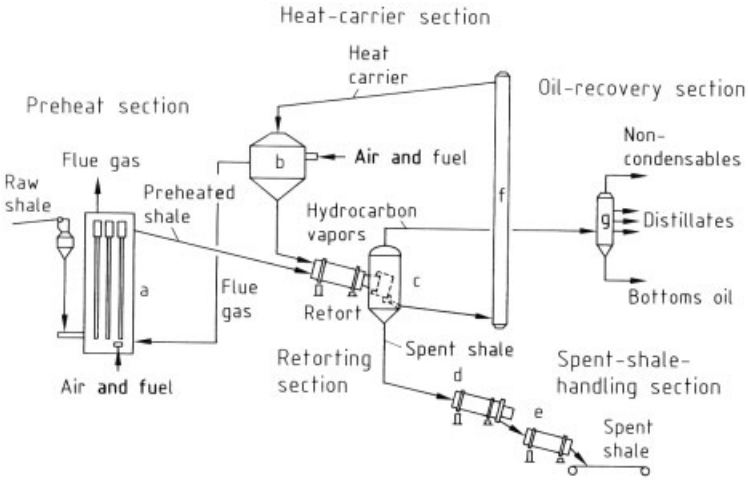

Figure 12. Tosco II retort
a) Lift-pipe system; b) Heater; c) Accumulator; d) Steam generator; e) Moisturizer; f) Skip hoist; g) Fractionator

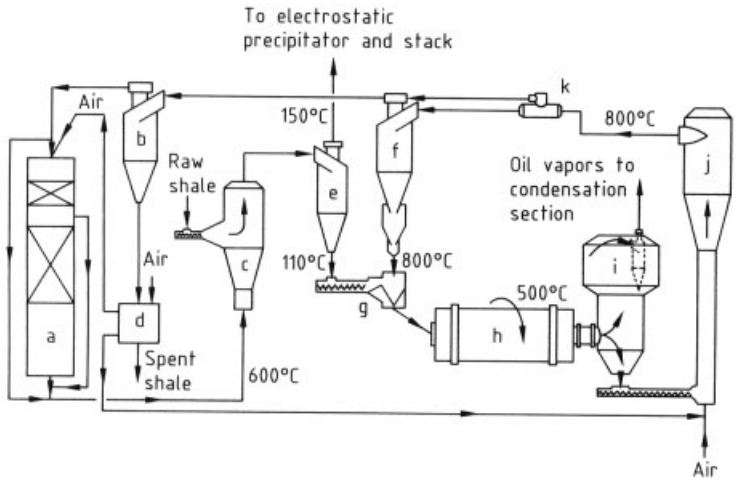

Figure 13. Galoter retort
a) Waste-heat boiler; b) Ash separator; c) Dryer; d) Ash heat exchanger; e) Dry oil shale separator; f) Heat carrier – separator; g) Mixer; h) Reactor; i) Dust removal chamber; j) Air-blown fire box; k) Bypass

A Galoter pilot plant with a capacity of about 500 t/d has been in operation in Estonia since 1964. This pilot plant had operated for 4100 d (11.2 a) by 1979. Over this period 1.76×10^6 t of shale was treated, yielding 235 000 t of commercial oil products and 102×10^6 m^3 of high-energy gas having a combustion heat of ca. 42 000 kJ/m^3 (STP). In 1979 two commercial-size units with a capacity of 3300 t/d each were commissioned in Estonia. The plants were reported to have startup problems in 1981 but appear to have been operational since then.

An Estonian report gives the following process materials balance on 1 t of run-of-mine Baltic shale (in kg/t) yielding 16.6 wt % oil (43 gallons per ton; 180 L/t) by Fischer assay: oil recovery by the Galoter system is 85 – 90 % of Fischer assay. Electrical energy used is 24 kW · h per tonne of wet feed shale.

Input	
Dry shale	876
Contained water	124
Flue gas	2
Steam	3

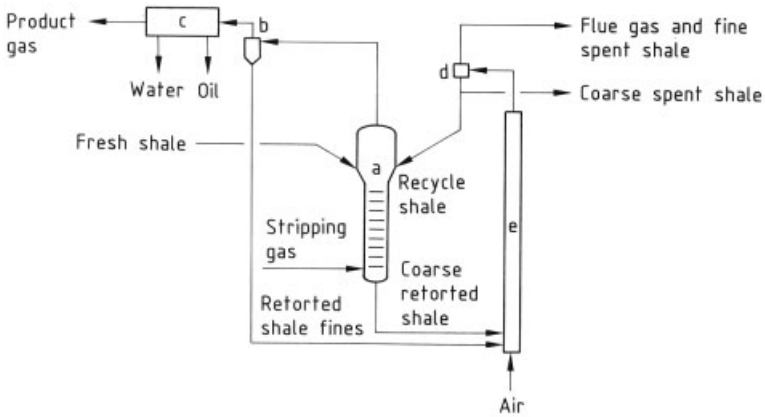

Figure 14. Chevron STB retort
a) Retort; b) Solids separator (hot); c) Liquid product recovery; d) Separator; e) Combustor

Water	174
Air	676
Total	1855
Output	
Shale oil	130
Gasoline	8
Gas	46
Process water	23
Flue gas	900
Steam	174
Spent shale	574
Total	1855

Chevron STB Retort. The Chevron STB retort is a small-particle retorting process in which retorted shale is burned in a separate combustor and recycled to the retort to supply the process heat requirement. The flow scheme for the core of the process is shown schematically in Figure 14. Fresh-shale preheating facilities and heat-recovery facilities for the flue gas and spent shale are not shown.

The environment in the retort section can best be described as a staged, moving bed of particles, in which a portion of the particles are fluidized. Solids move downward through the retort with an average velocity of 0.6 – 1.5 m/min. Proprietary internals within the retort restrict mixing in the vertical direction and thereby produce "staging" of the solids flow. The "staging" gives an approach to plug flow for the solids and thereby improves the stripping action of the countercurrent gas flow and minimizes the residence time required to assure complete retorting. Locally, conditions within the bed have all the appearances of a fluidized state, although the superficial gas velocity is far below the minimum fluidization velocity of many of the particles.

The retort can be operated with a stripping gas consisting of either recycled product gas, steam, or an inert gas. The superficial gas velocity at the bottom of the retort is 0.3 – 1.5 m/s. The volumetric flow of gas in the retort increases significantly in going from the bottom to the top because of the vapor added by the pyrolysis products.

The combination of thermal shock experienced by fresh shale on entering the retort, the pyrolysis of binding kerogen, and turbulent conditions within the retort cause a certain breakup of the shale. The finest particles are swept overhead from the retort with the product vapors. Most of the elutriated particles are recovered from the vapor stream before condensation of the oil. The recovered particles, collected by primary and secondary devices, contain carbonaceous residue and are therefore burned in the combustor section for recovery of their fuel value.

In the *product recovery section,* product vapors are condensed in stages, which yields several oil fractions and foul water. The initial condensation stage must be capable of handling a certain amount of solids, because the fines removal equipment is never 100 % efficient. Either the noncondensable gases are produced directly or a portion is recycled to the retort for stripping. With very lean shale, some or all of the product gas can be used as a supplementary fuel in the combustor.

The *combustion section* receives coarse retorted shale withdrawn from the bottom of the retort and retorted shale fines removed from the product vapors. Air injected at the bottom of the combustor both transports the shale pneumatically and combusts the carbonaceous residue. The temperature of the solids rises rapidly. At the outlet of the combustor, a solids separator splits the stream once again into fine and coarse fractions. Most of the hot, coarse fraction is recycled to the retort to provide the necessary process heat. Excess coarse shale is sent on to the heat recovery section, from which it is eventually discharged as coarse spent shale. The fine fraction exits the separator with the flue gas. This stream is also sent to the heat recovery section, where it is cooled and separated into a fine spent shale stream and a flue gas stream.

Chevron built and operated a 300-t/d pilot unit in its Salt Lake City refinery (Utah) in the 1980s. The plant has since been dismantled.

7.3. In Situ Processing of Oil Shale

An attractive alternative to the technically proven aboveground retorting processes may be available if attempts at developing economically viable in situ processes turn out to be successful. Large quantities of spent shale would not have to be dumped, nor would areas that had been destroyed by shale mining or dumping have to be reclaimed.

In situ processing of oil shale was practiced on an industrial scale in Sweden during and after World War II. Oil yields reached 60 % of the Fischer value. However, the thermal efficiency, including the electricity required to heat the shale, did not exceed 16 %. On a much smaller scale, oil shale was also processed in situ during World War II in Württemberg (Germany).

In 1965, the U.S. Bureau of Mines began an extensive testing program for the in situ retorting of oil shale in the western United States; the tests had little success. Several in situ projects that were test-operated by private oil companies in the western United States a number of years ago were not successful either.

However, from 1975 onward a *vertical, modified in situ process* tested in the oil shale deposits of the western United States gained more attention, and a number of successful results have already been reported. Such projects were operated by the Occidental Oil Company and Rio Blanco in the Green River formation.

These modified in situ processes require that the oil shale be mined underground to a certain degree. A cavity of $40 \times 40 \times 40\ m^3$, for example, is produced by underground mining, and the oil shale above this cavity is then blasted by explosives to provide a chimney or retort, which may be as much as 150 m high, filled with fractured shale. This loose shale bed can now be retorted in situ. Actually, this is not an in situ process in the narrow sense of the word, because some of the oil shale must be mined and retorted aboveground, and the blasted shale filling the chimney has also been transported, although not a long way.

Horizontal, modified in situ retorting is a variation of the vertical process. Geokinetics has operated a plant in the Green River formation in Utah to exploit a deposit ca. 25-m thick by the horizontal process. This also involves fracturing the shale layer (which must be close to the surface) with explosives and subsequently retorting the deposit horizontally.

In the *Bronco project* proposed in the United States, a 90 kilotonne nuclear explosion was planned in an oil shale formation of great thickness, which should break up the rock and make it suitable for subsequent in situ processing. According to this scheme, a burning wavefront would travel through the fractured rock downward from the top. However, social and environmental factors have prevented this project from being realized.

8. Combustion

Combustion of Pulverized Oil Shale. For ca. 60 years, oil shale has been combusted directly in power stations in Estonia. The Narva power station in the Baltic region, one of the largest thermal power stations in the Soviet Union, burns pulverized oil shale with a relatively high calorific value of ca. 9200 kJ/kg. However, the fact that the oil shale deposits are the only fuel resources in the entire region must not be overlooked.

Direct combustion of pulverized oil shale involves a number of problems resulting from its high ash content. Examples are corrosion and

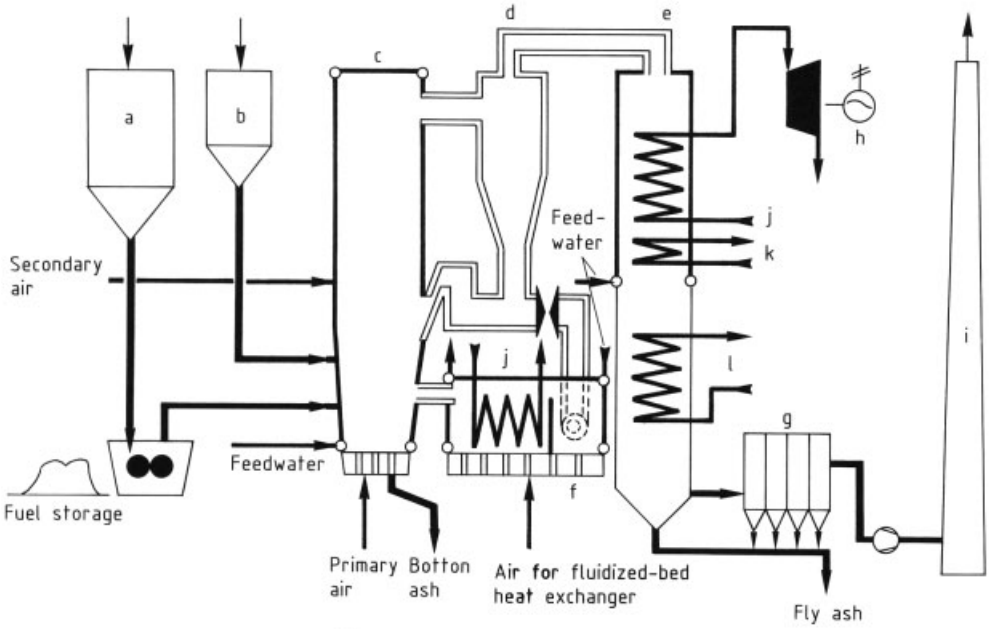

Figure 15. The CFB boiler
a) Fuel hopper; b) Limestone hopper; c) CFB reactor; d) Recycling cyclone; e) Convective pass, air preheater; f) Fluidized-bed heat exchanger; g) Dust filter; h) Turbine; i) Stack; j) Steam generator; k) Feedwater preheater; l) Air preheater
■ Walls of pipes cooled by circulated water; □ Fire resistant pipe walls

fouling of the heat-transfer surfaces. For this reason, oil shale – although in principle valuable as a fuel – is not of general interest today.

Tests performed by German firms proved that Moroccan oil shale can be used for power generation in pulverized shale-fired boilers if the boiler plant and the auxiliary systems are specifically designed for this fuel. Capital investment would then be ca. 14 % higher than for an oil-fired power station with the same output.

Combustion of Fine Granular Oil Shale. A commercial fluidized-bed combustor for 700-t/d fine granular oil shale is operated by the Rohrbach cement works in Germany [28]. Its main product is shale ash that is used in cement production. Heat released during combustion is used to calcine the high-carbonate shale and to generate steam for electric power production.

The combustion of fine granular (0 – 7 mm) oil shale in circulating fluidized bed (CFB) boilers has attracted attention over the past few years because most of the disadvantages associated with pulverized shale boilers can be avoided with this system. Because the combustion temperature is substantially lower (typically 850 °C), sintering or melting of the ash is avoided; sulfur capture by carbonates contained in the ash or added as limestone is higher; and NO_x formation is lower.

Figure 15 shows the CFB boiler system as developed by Lurgi. The CFB boiler has been widely used for various types of coal, but until now only two small oil shale CFB boilers have been built in Canada (70 MW_{th}) and Israel (41 MW_{th}).

9. Gasification

Industrial-scale gasification of oil shale began in Estonia in 1925, with the use of rotating-grate generators (Pintsch type) to produce not only gas but also ca. 170 kg of oil per tonne of shale. This process is still used in Estonia today.

In the United States, work on oil shale gasification has been continuing with certain interruptions since the 1960s. Laboratory tests at the Institute of Gas Technology (IGT) in Chicago have shown that between 87 and 97 % of the carbon contained in the organic material of Colorado oil shale can be converted to useful fuel gases by a hydrogenating gasification process at

hydrogen pressure up to 3.5 MPa. The laboratory reactor was a traveling-bed device with a diameter of 10 cm and a height of 7.5 m. The test results were encouraging. However, the oil shale had to be ground to very fine particles sizes, and difficulty was encountered in maintaining the specified temperature profiles. The reaction products differed as a function of operating temperature, retention time, and hydrogen partial pressure. The carbon was either completely gasified or reacted to a liquid – gas mixture. If carbon in the organic shale material is to be converted completely into gaseous hydrocarbons, severe reaction conditions, such as a very high pressure and a large hydrogen surplus, are required. At lower temperature, more and more liquid products result. This effect has also been observed during tests to upgrade bitumen from tar sands. The tendency therefore is to put up with a larger percentage of liquid products for economic reasons.

Since 1975, IGT has operated a process plant with throughputs up to 900 kg/h. This plant can handle oil shale with particle sizes up to 25 mm in a batchwise operation [29]. Rather than aiming at complete gasification of the oil shale in a single step, IGT's more recent research has concentrated on separate gasification of the shale oil.

Supported by public funds, Dow Chemical Company has been working since 1977 on projects to produce mainly gas by retorting, gasifying, and burning in situ the black Devonian shale found in Michigan.

10. Properties and Uses of the Products

Shale oil, distillation gas, and shale residue are the main products of an oil shale retorting plant.

Shale Oil. Shale oil may be refined to obtain conventional fuel oils, jet propulsion fuels, diesel oil, naphtha, and gasoline. Once the distillation gases have been removed, shale oil is obtained as a dark, viscous liquid with a high content of sulfur, nitrogen, and unsaturated compounds. Shale oil solidifies in the temperature range of 15 – 25 °C because of its high pour point.

Most important among the various *shale oil upgrading processes* are

1) *Coking* This method cannot be recommended, because considerable losses occur due to the production of low-grade coke.
2) *Hydrogenation or Hydrovisbreaking* By catalytic hydrogenation under pressure, shale oils can be converted to motor fuels with low sulfur and nitrogen content, yielding elemental sulfur and ammonia as byproducts.

Typical properties of various crude shale oils are listed in Table 8.

Distillation Gas. After elimination of condensable oils and vapors, the distillation gas from an oil shale retort consists mainly of carbon monoxide, carbon dioxide, hydrogen, methane, hydrogen sulfide, and nitrogen. It also contains liquefied petroleum gas and some naphtha vapors, which may be recovered partly by means of wash oil, activated carbon, or cryogenic separation. Depending on the retorting process used, the calorific value of the distillation gas (without gas naphtha) varies between 5.4 and 39.8 MJ/m^3 (STP). It may serve as fuel gas for industrial purposes, or if itcontains little nitrogen, it may be used in a reformer to produce hydrogen for refining. Alternatively, the product may be shift-converted, cleaned, and methanated catalytically to upgrade it to substitute natural gas.

Shale Residue. The shale residue is normally an undesirable byproduct and must be returned to the pit or dumped in a landfill area. Depending on the retorting process and type of oil shale, the residue still contains 1 – 15 % combustible matter. Aftercombustion of shale residue may therefore be of interest. Residues containing high percentages of calcium oxide may have the properties of hydraulic binders and could be used as additives in the cement industry.

11. Outlook

Although numerous oil shale deposits exist on all continents, the oil shale industry can be expected to develop further only if certain economic conditions are fulfilled. Special events such as an oil embargo or problems of national security may speed up developments, especially if they lead governments to subsidize the oil shale industry.

Table 8. Shale oil analyses

Shale origin		Green River (United States)		Schandelah (Germany)	Timahdit (Morocco)	Alexinac (Yugoslavia)
Retorting method	Fischer assay	Tosco II	Lurgi – Ruhrgas ∗∗	Lurgi – Ruhrgas ∗∗	Lurgi – Ruhrgas ∗∗	Gas combustion
Density (20 °C), g/cm^3	0.930∗	0.927∗	0.908	0.959	0.971	0.893
Pour point, °C	24	27	18	< −30	−23	38
Flash point, °C			46	55		142
Viscosity (°C), mm^2/s	23.7 (38)	22 (38)	8.3 (40)	6.8 (40)	6.1 (40)	15.0 (50)
Conradson carbon residue, wt %			3.9	6.6	5.6	1.5
Bromine number			63	61		
Elemental analysis, wt %						
C	85.2	84.9	85.4	83.4	78.0	82.3
H	11.4	11.5	11.1	9.6	9.7	12.3
N	1.8	1.9	1.7	1.0	1.5	1.0
S	1.0	0.9	0.7	3.8	6.5	0.8
O	0.6	0.8	1.1	2.2	4.3	3.6
C/H ratio	7.5	7.4	7.7	8.7	8.0	6.7
Calorific value, MJ/kg			39.77	41.16	40.19	
Distillation, °C						
Initial boiling point	89	< 93	70	69	78	214
10 vol %	169	135	160	152	167	274
20 vol %	221	210	198	190	195	304
30 vol %	270	261	236	225	227	330
40 vol %		326	280	262	268	349
50 vol %	346		326	302	308	
60 vol %	363	413	371	342	341	
70 vol %	374	455	414	395	377	
80 vol %		493	450		423	

∗ Density at 15.5 °C.
∗∗ Crude oil without gas naphtha.

12. References

General References

1. O. P. Strausz, E. M. Lown (eds.): *Oil Sand and Oil Shale Chemistry,* Verlag Chemie, New York 1978.
2. R. A. Meyers: *Handbook of Synfuels Technology,* McGraw-Hill, New York 1984.

Specific References

3. H. M. Thorne et al.: "Oil Shale Technology: A Review," *Inf. Circ. U.S. Bur. Mines* **8216** (1964).
4. R. F. Cane: "Shale Oil: What's Its Future?," *Hydrocarbon Process.* **44** (1965) no. 9, 185 – 190.
5. J. Burger: "L'Exploration Des Pyroschistes Ou Schistes Bitumeux, Donnes Générales Et Perspectives D'Avenir," *Rev. Inst. Fr. Pet.* **28** (1973) no. 3, 315 –372.
6. D. C. Duncan, V. E. Swanson: "Organic-Rich Shale of the United States and World Land Areas," *Geol. Surv. Circ. (U.S.)* **523** (1965).
7. United Nations Department of Economic and Social Affairs: *Utilization of Oil Shale-Progress and Prospects,* United Nations Publication Sales No. 67.II.B.20 (1967).
8. L. E. Schlatter: "Oil Shale Deposits of Western Europe," *United Nations Symposium on the Development and Utilization of Oil Shale Resources,* Tallinn 1968.
9. H. Steinert: *Ölschieferpotential von beachtlicher Größe,* Handelsblatt, Düsseldorf, Dec. 1, 1977.
10. H. Schade, *Glückauf* **111** (1975) 1172 – 1176.
11. "No Great Prospects for U.K. Shale Oil Industry," *Pet. Times,* March 1975.
12. V. A. Kotlukov: "Geology and Prospects of Studies of Oil Shales in the USSR," *United Nations Symposium on the Development and Utilization of Oil Shale Resources,* Tallinn 1968.
13. J. R. Donnell et al., *Q. Colo. Sch. Mines* **59** (1964) no. 3, 605 – 609.
14. A. J. Kraemer, *Rep. Invest. U.S. Bur. Mines* **4655** (1950).
15. S. A. Hussein, *Mod. Geol.* **6** (1978) 163 – 170.
16. E. Gil-Av, *Bull. Res. Counc. Isr.* **4** (1954) 136 – 143.
17. Y. Schachter, *Proc. Fourth World Pet. Congr.,* Section III H, Paper 4 (1955).

18. Y. Schachter: *Eleventh Israel Conference on Mechanical Engineering,* Technion, Haifa, Israel, 1977.
19. *Ullmann,* 3rd ed., **13,** 40.
20. H. Keltser, *Oel Kohle* **13** (1937) 1139 – 1149.
21. A. P. Petrov, *United Nations Oil Shale Symposium,* Tallinn 1968.
22. A. M. Kotov, *United Nations Oil Shale Symposium,* Tallinn 1968.
23. H. H. Schmitz, A. Tolle, *Erdöl Kohle Erdgas Petrochem.* **40** (1987) no. 6, 245 – 252.
24. A. B. Hubbard, W. E. Robinson, *Rep. Invest. U.S. Bur. Mines* **4744** (1950).
25. V. D. Allred, *Chem. Eng. Prog.* **62** (1966) no. 8, 55 –59.
26. R. L. Braun, A. J. Rothman, *Fuel* **54** (1975) 129 – 131.
27. G. Guthrie, S. Klosky: "The Oil Shale Industries of Europe," *Rep. Invest. U.S. Bur. Mines* **4776** (1951) 73 ff.
28. G. Friese, *Erdöl Kohle Erdgas Petrochem.* **14** (1961) 702 – 703.
29. F. C. Schora et al., *Hydrocarbon Process.* **56** (1977) 107 – 110.

Oils, Essential → Flavors and Fragrances
Oils, Mineral → Natural Gas
Oils, Mineral → Oil Refining
Olefin Polymers → Polyolefins
Olefins → Butadiene
Olefins → Butenes
Olefins → Ethylene
Olefins → Hydrocarbons
Olefins → Isoprene
Olefins → Propene
Olefins → Styrene
Olefins → Terpenes
Oleic Acid → Fatty Acids
Oleoresins → Resins, Natural
Olivine → Silicates

My Profile L

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Ophthalmological Preparations

Standard Article

Terence M. Dolak[1], O. William Lever Jr.[2], David Marsh[3], Irene Moran[4], Scott Sutton[5]

[1]Bausch & Lomb, Rochester, New York, United States
[2]Bausch & Lomb, Rochester, New York, United States
[3]Bausch & Lomb, Rochester, New York, United States
[4]Bausch & Lomb, Rochester, New York, United States
[5]Bausch & Lomb, Rochester, New York, United States

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a18_127
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Anesthetics
2.1. Topical Anesthetics
2.2. Regional Anesthetics
3. Antimicrobial Agents
3.1. Antibacterial Agents
3.1.1. Sulfonamides
3.1.2. β-Lactam Antibiotics
3.1.2.1. Penicillins
3.1.2.2. Cephalosporins
3.1.3. Tetracycline Antibiotics
3.1.4. Aminoglycoside Antibiotics
3.1.5. Polypeptide Antibiotics
3.1.6. Miscellaneous Antimicrobial Agents
3.2. Antifungal Agents
3.3. Antiprotozoal Agents
3.4. Antiviral Agents
4. Anti-inflammatory Agents
4.1. Corticosteroids
4.2. Miscellaneous Anti-inflammatory Agents
5. Antiglaucomatous Agents
5.1. Sympathomimetic Drugs
5.2. β-Adrenergic Blocking Agents

5.3. Carbonic Anhydrase Inhibitors
5.4. Hyperosmotic Agents
5.5. Miotics (Parasympathomimetic Agents)
5.5.1. Cholinergic Agonists
5.5.2. Anticholinesterases
6. Mydriatics and Cycloplegics
6.1. Sympathomimetics
6.2. Parasympatholytics
7. Vasoactive (Adrenergic) Agents
8. Diagnostic Agents
9. Dry Eye Medications
10. Miscellaneous Ophthalmic and Contact Lens Preparations
10.1. Disinfectants and Preservatives
10.1.1. Biguanides
10.1.2. Mercurial Preservatives
10.1.3. Quaternary Ammonium Compounds
10.1.4. Miscellaneous Disinfectants and Preservatives
10.2. Contact Lens Cleaners and Rewetting Agents
10.2.1. Surfactant Cleaners
10.2.2. Enzymatic Cleaners

[Top of Page]

1. Introduction

The eye is an external organ, and is susceptible to injury, infection, and irritation, as well as to organic and functional disorders. The products and preparations for ophthalmic use therefore comprise a broad category and include medications, diagnostic agents, and medical device formulations.

Over-the-counter and prescription medications are used to treat chronic ailments (e.g., glaucoma, blepharitis, and dry eye), as well as temporary conditions (e.g., infections, allergic reactions, and injuries).

Ocular diagnostic agents are a specialized group of drugs and dyes utilized primarily for ocular examination and the identification of potential disease conditions.

Medical device products are used for the hydration, wetting, cleaning, and disinfection of contact lenses.

[Top of Page]

2. Anesthetics

Topical anesthetics diminish sensory nerve impulse generation and conduction locally, and are used to anesthetize the ocular surface during short corneal or conjunctival procedures (e.g., tonometry, removal of sutures or foreign bodies) [1-4]. Proxymetacaine (0.5 %), oxybuprocaine (0.4 %), or tetracaine (0.25 – 1.0 %) solutions are commonly used [1]; proxymetacaine is particularly useful [4]. These agents provide rapid onset of local anesthesia of relatively short duration (e.g., one or two drops of a 0.5 % proxymetacaine solution take effect within 30 s with anesthesia lasting about 15 min). Cocaine exhibits significant corneal epithelial toxicity and is now rarely used; however, it may still be employed for debridement of the corneal epithelium, as in herpetic epithelial keratitis [3], [4]. Topical application of combinations of local anesthetics does not provide additive effects but increases the risk of side effects [2].

Regional anesthetics are employed in ophthalmic surgical procedures. They are administered by injection for infiltration anesthesia or for regional nerve block (facial or retrobulbar) anesthesia [1-3]. Onset of action is typically 3 – 11 min, and anesthesia may last from 30 – 45 min (procaine) up to 300 – 600 min (etidocaine) or longer (480 – 720 min for bupivacaine with epinephrine) [2]. Combination of injectable anesthetics with vasoconstrictors such as epinephrine decreases the rate of systemic absorption and prolongs the duration of anesthesia.

Most of the anesthetics listed in Sections Antibacterial Agents and Antifungal Agents are described in detail elsewhere, see ⟶ .

2.1. Topical Anesthetics

Cocaine hydrochloride [53-21-4], $C_{17}H_{21}NO_4 \cdot HCl$, M_r 339.8. Cocaine is not commercially available as a solution for clinical use.

Proxymetacaine hydrochloride [5875-06-9], 2-diethylaminoethyl 3-amino-4-propoxybenzoate monohydrochloride, proparacaine hydrochloride, $C_{16}H_{26}N_2O_3 \cdot HCl$, M_r 330.9.

Trade names: Ak-taine (Akorn); Alcaine (Alcon); Kainair (Pharmafair); Keracaine, Kerakain (Chibret); Ocu-caine (Ocumed); Ophthaine (Squibb); Ophthetic (Allergan).

Tetracaine hydrochloride [136-47-0], 2-dimethylaminoethyl 4-butylaminobenzoate monohydrochloride, $C_{15}H_{24}N_2O_2 \cdot HCl$, M_r 300.8.

Trade names: Minims Amethocaine Hydrochloride (Smith & Nephew), Pontocaine (Winthrop).

Oxybuprocaine hydrochloride [5987-82-6], 2-diethylaminoethyl 4-amino-3-butoxybenzoate monohydrochloride, benoxinate hydrochloride, $C_{17}H_{28}N_2O_3 \cdot HCl$, M_r 344.9.

Trade names: Alcon Opulets Benoxinate (Alcon); Cebesine (Chauvin-Blanche); Conjuncain (Mann); Minims Benoxinate Hydrochloride, Minims Oxybuprocaine Hydrochloride (Smith & Nephew); Novesin (Dispersa); Novesina (Sandoz); Novesine (Chibret, Wander); Oftalmocaina, Oxibuprokain Minims (Smith & Nephew); Poen Caina, Prescaina (Llorens).

Oxybuprocaine hydrochloride (0.4 %) is available in combination with sodium fluorescein (0.25 %) under the trade name Fluress (Sola/Barnes – Hind) for applanation tonometry.

2.2. Regional Anesthetics

Bupivacaine hydrochloride [18010-40-7] (monohydrate [14252-80-3]), 1-butyl-*N*-(2,6-dimethylphenyl)-2-piperidinecarboxamide monohydrochloride monohydrate, $C_{18}H_{28}N_2O \cdot HCl \cdot H_2O$, M_r 342.9.

Etidocaine hydrochloride [36637-19-1], *N*-(2,6-dimethylphenyl)-2-(ethylpropylamino)butanamide monohydrochloride, $C_{17}H_{28}N_2O \cdot HCl$, M_r 312.9.

Hexylcaine hydrochloride [532-76-3], 2-cyclohexylamino-1-methylethyl benzoate monohydrochloride, $C_{16}H_{23}NO_2 \cdot HCl$, M_r 297.8.

Lidocaine hydrochloride [73-78-9] (anhydrous), monohydrate [6108-05-0], 2-diethyl-aminoaceto-*N*-(2,6-dimethylphenyl)acetamide monohydrochloride monohydrate, $C_{14}H_{22}N_2O \cdot HCl \cdot H_2O$, M_r 288.8.

Mepivacaine hydrochloride [1722-62-9], *N*-(2,6-dimethylphenyl)-1-methyl-2-piperidinecar-boxamide hydrochloride, $C_{15}H_{22}N_2O \cdot HCl$, M_r 282.8.

Prilocaine hydrochloride [1786-81-8], *N*-(2-methylphenyl)-2-(propylamino)propanamide hydrochloride, $C_{13}H_{20}N_2O \cdot HCl$, M_r 256.8.

Procaine hydrochloride [51-05-8], 2-diethylaminoethyl 4-aminobenzoate hydrochloride, $C_{13}H_{20}N_2O_2 \cdot HCl$, M_r 272.8.

Tetracaine hydrochloride (see Section Topical Anesthetics).

[Top of Page]

3. Antimicrobial Agents

3.1. Antibacterial Agents

Successful therapy for bacterial infection with antibiotics relies upon five considerations:

1. Appropriate indications for use
2. Isolation and identification of the infective agent
3. Efficacy of the antibiotic for the infective agent
4. Adequate levels of the antibiotic at the site of infection
5. Low toxicity of the antibiotic for the host [5]

Antibiotics act upon bacteria through a variety of mechanisms. The penicillins, cephalosporins, vancomycin, and bacitracin inhibit development of the bacterial cell wall. Tetracycline, erythromycin, clindamycin, chloramphenicol, and the aminoglycosides inhibit protein synthesis.

3.1.1. *Sulfonamides*

See also ⟶ Chemotherapeutics – Sulfa Drugs.

Although the use of sulfonamides for the treatment of major bacterial infections has been largely supplanted by antibiotics, they are still prescribed for the topical treatment of minor ocular infections. They exhibit broad-spectrum bacteriostatic activity against many gram-positive and gram-negative organisms, and are useful against the protozoa *Toxoplasma gondii* and *Plasmodium falciparum*. Some strains of *Pseudomonas, Chlamydia*, and *Actinomyces* are also sensitive to these drugs. A major disadvantage is the frequent occurrence of cross-resistant microorganisms. Many side effects have been reported following the topical administration of sulfonamides. These drugs are contraindicated for the treatment of keratoconjunctivis sicca or penetrating corneal wounds. Sulfacetamide sodium is the most widely used sulfonamide at a concentration of 10 %, or in combination with the steroid prednisolone acetate (0.2 – 0.5 %).

The sulfonamides are antimetabolites. Their structural similarities to 4-aminobenzoic acid allow them to act as competitive inhibitors in the folic acid biosynthetic pathway.

Sulfacetamide sodium [144-80-9], *N*-[(4-aminophenyl)sulfonyl]acetamide, $C_8H_{11}N_2NaO_3S$, M_r 238, *mp* 182 – 184 °C, is incorrectly called sulfacetimide. For synthesis, see [6].

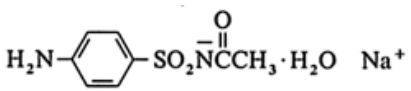

Trade names: Ak-Cide, Ak-Sulf (Akorn); Bleph-10, Blephamide Liquifilm/S.O.P. (Allergan); Cetamide Ointment, Cetrapred Ointment (Alcon); FML-S (Allergan); Isopto Cetmide Solution (Alcon); Metimyd Ophthalmic Solution/Ointment — Sterile (Schering); Ocu-Lone-C Sterile Ophthalmic Ointment/Suspension (Ocumed); Predsulfair Ointment/Suspension (Pharmafair); Sodium Sulamyd Ophthalmic Ointment 10 %, Sodium Sulamyd Ophthalmic Suspension 10 % (30 %) — Sterile (Schering); Sulf-10 Dropperettes 10 %/Ophthalmic Solution 10 % (IOLAB); Sulfair (Pharmafair); Sulphrin (Bausch & Lomb); Sulpred Suspension (Pharmafair); Sulten-10 (Bausch & Lomb); Vasocidin Ointment/Solution, Vasosulf (IOLAB).

Sulfadiazine [68-35-9] (sodium [547-32-0], silver [22199-08-2]), 4-amino-*N*-2-pyrimidinylbenzenesulfonamide, $C_{10}H_{10}N_4O_2S$, M_r 250.3, *mp* 252 – 256 °C. For synthesis, see [7].

Sulfamerazine sodium [127-58-2], 4-amino-*N*-(4-methyl-2-pyrimidinyl)benzenesulfonamide sodium salt, $C_{11}H_{11}N_4NaO_2S$, M_r 286.3. For synthesis, see [7]. Sulfamerazine sodium is a short-acting sulfonamide with action similar to that of sulfamethoxazole. It is usually administered with other sulfonamides.

Sulfamethizole [144-82-1], 4-amino-*N*-(5-methyl-1,3,4-thiadiazol-2-yl)benzenesulfonamide,$C_9H_{10}N_4O_2S_2$, M_r 270.3, *mp* 208 °C. For synthesis, see [8].

Sulfisoxazole diolamine [4299-60-9], 4-amino-*N*-(3,4-dimethyl-5-isoxazolyl)benzenesulfonamide diethanolamine, M_r 372.4. For synthesis, see [9].

Trade names: Ganstrin Ophthalmic Ointment/Solution, Suladrin, Sulfium (Roche).

3.1.2. β-*Lactam Antibiotics*

3.1.2.1. *Penicillins*

Penicillins are effective antimicrobials, although they have two drawbacks: (1) microbial resistance is relatively common, and (2) penicillins may be antigenic; anaphylactic shock is therefore a concern when they are administered systemically. See also ⟶ Antibiotics – Semisynthetic Penicillins.

Ampicillin [69-53-4], (sodium derivative [69-52-3], trihydrate [7177-48-2]), $C_{16}H_{19}N_3O_4S$, M_r 349.42, *mp* 202 °C (decomp.). Ampicillin is a semisynthetic derivative of penicillin. For synthesis, see [10]. See also ⟶ Antibiotics.

Methicillin sodium [132-92-3], $C_{17}H_{19}N_2NaO_6S$, M_r 402.42, *mp* 196 – 197 °C. Methicillin sodium is a semisynthetic antibiotic related to penicillin. For synthesis, see [11]. See also ⟶ Antibiotics.

3.1.2.2. *Cephalosporins*

The cephalosporins originated from a mold of the genus *Cephalosporium*. Although cephalosporins do not usually cause an allergic reaction in patients sensitized to penicillin, allergy to cephalosporins can also develop. The cephalosporins are not inactivated by penicillin-resistant bacteria and have a broader spectrum of activity than penicillin. See also ⟶ Antibiotics – Natural Cephalosporins ⟶ Antibiotics – Semisynthetic Cephalosporins ⟶ Antibiotics – Cephamycins.

Carbenicillin [4697-36-3] (disodium derivative [4800-94-6], indanyl sodium derivative [26605-69-6], phenyl sodium derivative [21649-57-0]), $C_{17}H_{18}N_2O_6S$, M_r 378.42. For synthesis, see [12]. See also ⟶ Antibiotics ⟶ Antibiotics ⟶ Antibiotics.

Cefazolin [25953-19-9] (sodium salt [27164-46-1]), $C_{14}H_{14}N_8O_4S_3$, M_r 454.50, *mp* 198 – 200 °C. For synthesis, see [13]. For structure, see ⟶ Antibiotics – Semisynthetic Cephalosporins.

Cephalothin [153-61-7] (cephalothin sodium [58-71-9]), $C_{16}H_{15}N_2NaO_6S_2$, M_r 418, *mp* 160 °C. Cephalothin is a semisynthetic cephalosporin antibiotic. For synthesis, see [14]. For structure, see ⟶ Antibiotics – Semisynthetic Cephalosporins.

3.1.3. *Tetracycline Antibiotics*

Tetracyclines reversibly block protein synthesis by specifically binding to the bacterial ribosome. Tetracycline and chloramphenicol were the first broad-spectrum antimicrobials discovered. The tetracycline antibiotics listed below are described in detail elsewhere; see ⟶ Antibiotics – Tetracyclines.

Chlortetracycline hydrochloride [64-72-2] (chlortetracycline [57-62-5]), $C_{22}H_{24}Cl_2N_2O_8$, M_r 493.9. Chlortetracycline is isolated from *Streptomyces aureofaciens* [15]. See also ⟶ Antibiotics

Trade name: Aureomycin Ophthalmic (IOLAB).

Oxytetracycline [79-57-2] (hydrochloride [2058-46-0]), $C_{22}H_{24}N_2O_9$, M_r 460.44, is an antibiotic isolated from the fermentation of *Streptomyces rimosus* [16]. See also ⟶ Antibiotics

Trade name: Terramycin (5 mg/mL) with Polymyxin B Ointment (10 000 U/mL) (Roerig).

Tetracycline [60-54-8] (hydrochloride [64-75-5], phosphate complex [1336-20-5]), $C_{22}H_{24}N_2O_8$, M_r 444.43, is an antibiotic produced by *Streptomyces* spp. [17], [18]. See also ⟶ Antibiotics

Trade names: Achromycin and Achromycin Ophthalmic Ointment 1 % (IOLAB).

3.1.4. *Aminoglycoside Antibiotics*

The aminoglycoside antibiotics exhibit broad-spectrum antimicrobial activity, particularly against gram-negative organisms (⟶ Antibiotics – Aminoglycosides). They inhibit protein synthesis and decrease translational efficiency through interaction with the 30 S subunit of bacterial ribosomes. Cross resistance exists among the aminoglycosides.

Aminoglycosides are generally administered topically because systemic adsorption may produce oto- and nephrotoxicity. Aminoglycosides and β-lactam antibiotics are incompatible in the same formulation. Aminoglycosides are not effective under acid conditions, in high salt concentrations, or against intracellular bacteria.

Framycetin sulfate [4146-30-9], [28002-70-2], the sulfate of neomycin B, $C_{23}H_{46}N_6O_{13} \cdot 3\ H_2SO_4$, M_r 908.9. Neomycin B is isolated from the mixture of neomycins A, B, and C produced by *Streptomyces fradiae*; see also ⟶ Antibiotics [19]. Framycetin sulfate has broad-spectrum activity similar to that of neomycin but is more efficacious against *Pseudomonas* spp. Unlike most other aminoglycosides, framycetin sulfate can be administered by subconjunctival injection; however, because of its high degree of systemic absorption, extreme caution must be employed to avoid toxicity.

Trade names: Framygen (Fisons), Soframycin (Roussel), with 0.5 % hydrocortisone acetate in Framycort (Fisons), with 0.05 % dexamethasone and 0.005 % gramicidin in Sofradex (Roussel).

Gentamicin sulfate [1405-41-0] (gentamycin [1403-66-3]) is a complex mixture of the sulfates of gentamicin C_1, C_{1a}, and C_2. This complex is produced by fermentation of *Micromonospora purpurea* [20]; see also ⟶ Antibiotics. Gentamicin sulfate is used as a drop, ointment, or buffered preserved solution, generally at a concentration of 0.3 %.

Trade names: Alcomicin (Alcon); Cidomycin (Roussel); Garamycin (Kirby-Warrick); Garamycin Ophthalmic Ointment/Solution — Sterile (Schering); Genopyic (Allergan); Gentacidin (CooperVision); Gentafair (Pharmafair); Genticina (Antibioticos); Genoptic Liquifilm Sterile Ophthalmic Solution (Allergan); Genoptic S.O.P. Sterile Ophthalmic Ointment (Allergan); Gentafair Cream and Ointment/Solution (Pharmafair); Gentak Ointment and Solution (Akorn); Genticidin Ointment/Solution (IOLAB); Gentrasul Ointment/Solution (Bausch & Lomb); Minims Centamicin Sulfate (Smith & Nephew); Ocu-Mycin Sterile Ophthalmic Ointment/Solution (Ocumed); Pred-G Liquifilm Sterile Ophthalmic Suspension (Allergan); Pred-G S.O.P. Sterile Ophthalmic Ointment (Allergan).

Micronomicin sulfate (micronomicin [52093-21-7]), $(C_{20}H_{41}N_5O_7)_2 \cdot 5\ H_2SO_4$, M_r 1417.5, *mp* 260 °C. This gentamicin C complex antibiotic is produced by *Micromonospora sagamiensis* var. *nonreducans* [21].

Trade names: Sagamicin (Kyowa), Santemycin (Santen).

Neomycin [1404-04-2] is an antibiotic complex made up of neomycins A, B, and C from *Streptomyces fradiae* [22-24]. See also ⟶ Antibiotics.

Trade names: Ak-Spore (H.C.) Ointment/Solution, Ak-Trol Ointment & Suspension (Akorn); Cortisporin Ophthalmic Ointment/Suspension (Burroughs Wellcome); Dexacidin Ointment/Suspension (IOLAB); Dexasporin Ointment/Suspension (Pharmafair); Infectrol Ointment/Suspension (Bausch & Lomb); Maxitrol Ophthalmic Ointment/Suspension (Alcon); NeoDecadron Sterile Ophthalmic Ointment (Merck Sharp & Dohme); Neo Dexair Ophthalmic Solution (Pharmafair); Neosporin Ophthalmic Ointment/Solution (Burroughs Wellcome); Neotricin Ointment/Solution (Bausch & Lomb); Ocu-Cort Sterile Ophthalmic Ointment, Ocu-Spor-B Sterile Ophthalmic Ointment, Ocu-Spor-G Sterile Ophthalmic Solution (Ocumed); Ocutricin HC Ointment/Suspension, Ocutricin Ointment/Solution (Pharmafair); Ocu-Trol Sterile Ophthalmic Ointment/Suspension (Ocumed); Poly-Pred Liquifilm (Allergan).

Sisomicin [32385-11-8] (sulfate [53179-09-2]), $C_{19}H_{37}N_5O_7$, M_r 447.6, monohydrate from ethanol, *mp* 198 – 201 °C. Sisomicin is an antibiotic produced by *Micromonospora inyoesis* [25]. It is usually administered systemically and not commonly used as an ophthalmic antibiotic. See also ⟶ Antibiotics – Sisomicin.

Streptomycin sulfate [3810-74-0] (streptomycin [57-92-1]), $C_{42}H_{84}N_{14}O_{36}S_3$, M_r 1457.40. Streptomycin is produced by *Streptomyces griseus* [26]; see also ⟶ Antibiotics. Corneal penetration of streptomycin is poor but can be greatly enhanced by iontophoresis.

Tobramycin [32986-56-4] (sulfate [79645-27-5]), $C_{18}H_{37}N_5O_9$, M_r 467.5. Tobramycin is produced by *Streptomyces tenebrarius* [27]. See also ⟶ Antibiotics. Tobramycin has a broad spectrum of activity similar to that of gentamicin. Cross resistance of tobramycin and gentamicin has been observed. Tobramycin is four times more effective against *Pseudomonas aeruginosa*, and less effective against *Serratia*, than gentamicin. Unlike other aminoglycosides, tobramycin is ineffective against mycobacteria.

Trade names: TobraDex Ophthalmic Suspension and Ointment, TobraLex, Tobral, Tobramaxin, Tobrex (Alcon); Tobra-Gobens (Normon); Tobra-Laf (Andalucia).

3.1.5. *Polypeptide Antibiotics*

Bacitracin [1405-87-4], bacitracin A (major component): $C_{66}H_{103}N_{17}O_{16}S$, M_r 1422.71. Commercial bacitracin is a mixture of at least nine bacitracins produced by *Bacillus subtilis* and *B. licheniformis* [28], [29]; see ⟶ Antibiotics. It is available singly only as an ointment (500 mg/g) and in combination with other antimicrobials (see Table 1).

Table 1. Combination antibiotic solutions and ointments *

Product	Polymyxin B sulfate, units/g	Neomycin, mg/g	Bacitracin, units/g	Oxytetracycline, mg/g	Gramicidin, mg/mL	Company
Neotal Ophthalmic Ointment	5000	5	400			Hauck
Terramycin with Polymyxin B Ointment	10000			5		Roerig
Ak-Poly-Bac Ointment	10000		500			Akorn
Polysporin Ointment	10000		500			Burroughs Wellcome
Neomycin Sulfate–Polymycin B Sulfate–Gramicidin Solution	10000*	1.75**			0.025	Rugby
Ak-Spore Solution	10000*	1.75**			0.025	Akorn
Neocidin Solution	10000*	1.75**			0.025	Major
Neosporin Ophthalmic Solution	10000*	1.75**			0.025	Burroughs Wellcome
Neotricin Opthalmic Solution	10000*	1.75**			0.025	Bausch & Lomb
Statrol Ointment	10000	3.5				Alcon
Ak-Spore Ointment	10000	3.5	400			Akorn
Neosporin Ophthalmic Ointment	10000	3.5	400			Burroughs Wellcome
Neotricin Ointment	10000	3.5	400			Bausch & Lomb
Mycitracin Ophthalmic Ointment	10000	3.5	500			Upjohn
Statrol Solution	16250*	3.5**				Alcon

* Units/mL. ** mg/mL.

Trade names: Ak-Poly-Bac Ointment, Ak-Spore (H.C.) Ointment (Akorn); Bacitracin Ophthalmic Ointment — Sterile (Pharmafair); Cortisporin Ophthalmic Ointment Sterile, Neosporin Ophthalmic Ointment Sterile (Burroughs Wellcome); Neotricin Ointment/Solution (Bausch & Lomb); Ocumycin (HC) Ointment (Pharmafair); Polysporin Ophthalmic Ointment Sterile (Burroughs Wellcome).

Colistimethate sodium [8068-28-8], $C_{58}H_{105}N_{16}Na_5O_{28}S_5$, M_r 1749.8. Colistimethate sodium is an injectable form of colistin, a cyclopolypeptide produced by *Bacillus colistinus* [30]. See also ⟶ Antibiotics.

Trade name: Coly-Mycin M (Parke-Davis).

Polymyxin B sulfate [1405-20-5] is a mixture of polymyxins B_1 and B_2 produced by *Bacillus polymyxa* [31], [32]. See also ⟶ Antibiotics. Polymyxins are cyclic polypeptides that are similar to detergents in having basic groups with a fatty acid side chain. They exert their antimicrobial effect by solubilizing the cell membrane. Polymyxin B sulfate is available as a sterile powder for reconstitution or in combination with other antimicrobials (see Table 1).

Trade names: Ak-Poly-Bac Ointment, Ak-Spore (H.C.) Ointment/Suspension, Ak-Trol Ointment & Suspension (Akorn); Cortisporin Ophthalmic Ointment/Suspension (Burroughs Wellcome); Dexacidin Ointment/Suspension (IOLAB); Dexasporin Ointment/Suspension (Pharmafair); Infectrol Ointment/Suspension (Bausch & Lomb); Maxitrol Ophthalmic Ointment/Suspension (Alcon); Neosporin Opthalmic Ointment/Ophthalmic Solution Sterile (Burroughs Wellcome); Neotricin Ointment/Solution (Bausch & Lomb); Ocu-Cort Sterile Ophthalmic Ointment (Ocumed); Ocumycin Ointment (Pharmafair); Ocu-Spor-B Sterile Ophthalmic Ointment, Ocu-Spor-G Sterile Ophthalmic Solution (Ocumed); Ocutricin HC Ointment/Suspension, Ocutricin Ointment/Solution (Pharmafair); Ocu-Trol Sterile Ophthalmic Ointment/Suspension (Ocumed); Opthocort (Parke-Davis); Polymyxin B Sulfate Sterile Ophthalmic (Pfizer); Poly-Pred Liquifilm (Allergan); Polysporin Ophthalmic Ointment Sterile (Burroughs Wellcome).

3.1.6. *Miscellaneous Antimicrobial Agents*

Chloramphenicol [56-75-7], $C_{11}H_{12}Cl_2N_2O_5$, M_r 323.1, *mp* 150.5 – 151.5 °C, is a broadspectrum antibiotic produced by *Streptomyces venezuelae* [33-35]. Chloramphenicol reversibly blocks the bacterial ribosome, inhibiting protein synthesis. See also ⟶ Antibiotics.

Trade names: Ak-Chlor Sterile Ophthalmic Ointment/Solution (Akorn); Chlorofair Ointment/Solution (Pharmafair); Chloromycetin Ophthalmic Ointment/Solution/Powder, Hydrocortisone Ophthalmic Solution (Parke-Davis); Chloroptic S.O.P., Chloroptic Sterile Ophthalmic Solution (Allergan); Ocu-Chlor Sterile Opthalmic Ointment/Solution (Ocumed); Opthochlor Ophthalmic Solution (Parke-Davis).

Clindamycin [18323-44-9] (palmitate hydrochloride [25507-04-4], hydrochloride monohydrate [58207-19-5]), $C_{18}H_{33}ClN_2O_5S$, M_r 424.98, is a semisynthetic antibiotic prepared from lincomycin [36]. See also ⟶ Antibiotics.

Trade name: Cleocin Phosphate (Upjohn).

Erythromycin [114-07-8], $C_{37}H_{67}NO_{13}$, M_r 733.9, *mp* 135 – 140 °C, resolidifies with second *mp* 190 – 193 °C. Erythromycin is produced by *Streptomyces erythreus* [37] and is a mixture of erythromycin A, B, and C, with A predominant. It is the most widely used of the macrolide antibiotics and exerts its effect at the level of protein synthesis by reversibly inhibiting the bacterial ribosome. See also ⟶ Antibiotics – 14-Membered Ring Macrolides ⟶ Antibiotics.

Trade names: Ak-Mycin Ointment (Akorn), Erythromycin Ophthalmic Ointment (Pharmafair), Ilotycin (Dista).

Lincomycin [154-21-2], $C_{18}H_{34}N_2O_6S$, M_r 406.6, is an antibiotic produced by *Streptomyces lincolnensis* var. *lincolnensis* [38], [39]. Lincomycin reversibly inhibits the bacterial ribosome. See also ⟶ Antibiotics.

Trade name: Lincocin (Upjohn).

Ofloxacin [82419-36-1], $C_{18}H_{20}FN_3O_4$, M_r 361.4, *mp* 250 – 257 °C, is a broad-spectrum, fluorinated quinoline antibacterial. See also ⟶ . For synthesis, see [40].

Silver nitrate [7761-88-8], $AgNO_3$, M_r 169.9, *mp* 212 °C, is used to treat the eyes of newborns. Caution should be used with repeated applications because cauterization of the cornea and blindness may result. Silver nitrate is incompatible with sulfacetamide preparations.

Trade name: Silver Nitrate (Lilly).

Vancomycin [1404-90-6] (hydrochloride [1404-93-9]), $C_{66}H_{75}Cl_2N_9O_{24}$, M_r 1449.2. Vancomycin is an amphoteric glycopeptide produced by *Streptomyces orientalis* [41]. It inhibits cell wall synthesis. See also ⟶ Antibiotics.

Trade names: Vancocin (Lilly), Vancoled (Lederle), Vancor (Adria).

3.2. Antifungal Agents

For general information see ⟶ Antimycotics.

Ocular fungal infection is a rare disease but is increasing in occurrence [42-44]. Commensal fungi commonly present in the eye include *Aspergillus*, *Penicillium*, *Candida*, *Fusarium*, and *Rhodotorula*. These may become opportunistic pathogens if the eye's defenses are compromised by physical trauma or an underlying disease such as diabetes.

Although natamycin is the only antifungal agent currently available commercially in the United States as a topical ophthalmic formulation, other antifungal drugs (including polyenes and imidazoles) have been used topically in dilute solution or suspension [45].

The *polyene antimycotics* (amphotericin B, nystatin, natamycin) are insoluble in water and unstable in oxygen, light, water, heat, and at extreme pH. Their biological specificity for yeast relies on preferential binding to yeast membranes via ergosterol over their affinity for cholesterol, the primary sterol in mammalian cell membranes. Cell death is induced by increasing the permeability of the fungal membrane, allowing depletion of intracellular components.

The *imidazoles* (miconazole and ketoconazoles) have a broad antifungal and antimicrobial spectrum, with significantly less toxicity than the polyenes. Antifungal effects may result from the inhibition of estergol synthesis, leading to cell membrane permeability. Miconazole physically disrupts the membrane, leading to cell lysis. High concentrations of miconazole also increase intracellular concentrations of hydrogen peroxide, presumably through inhibition of cytochrome C peroxidase.

Flucytosine is a widely used *pyrimidine antimycotic*.

Amphotericin B [1397-89-3], $C_{47}H_{73}NO_{17}$, M_r 924.1, is a polyene antibiotic produced by *Streptomyces nodosus* [46]. Amphotericin can be administered parenterally for systemic mycosis. Topical administeration of 0.1 and 0.25 % has been effective for the treatment of keratomycoses. See also ⟶ Antimycotics – Amphotericin B.

Trade name: Fungizone (Squibb).

Nystatin [1400-61-9] is a polyene antifungal antibiotic complex containing three active components (A_1, A_2, and A_3) produced by several *Streptomyces* spp.[47]. See also ⟶ Antimycotics – Nystatin A1. Nystatin is effective in topical application for ocular infection.

Trade names: Mycostatin (Squibb), Nilstat (Lederle), Nystex (Savage), O-V Statin (Squibb).

Natamycin [7681-93-8], $C_{33}H_{47}NO_{13}$, M_r 665.75, *mp* 280 – 300 °C (decomp.), is a polyene antibiotic produced by *Streptomyces natalensis* and *S. chattanoogensis* [48]. See also ⟶ Antimycotics – Natamycin (Pimaricin). Natamycin is the least toxic, least irritating, and most stable polyene. It is only useful, however, for the treatment of superficial keratomycosis because it penetrates tissue poorly.

Trade name: Natacyn (Alcon).

Miconazole [22916-47-8] (nitrate [22832-87-7]), $C_{18}H_{14}Cl_4N_2O$, M_r 416.12. See also ⟶ Antimycotics – Miconazole. For synthesis, see [49].

Trade names: Micatin, Monistat-Derm (Ortho).

Ketoconazole [65277-42-1], $C_{26}H_{28}Cl_2N_4O_4$, M_r 531.44, *mp* 146 °C, is an orally active, broad-spectrum antimycotic. See also ⟶ Antimycotics – Ketoconazole. For synthesis, see [50].

Trade name: Nizoral (Janssen).

Flucytosine [2022-85-7], 5-fluorocytosine, $C_4H_4FN_3O$, M_r 129.1, *mp* 295 – 297 °C (decomp.). For synthesis, see [51]. Flucytosine, a fluorinated pyrimidine, is metabolized to either 5-fluorouracil, an inhibitor of RNA synthesis, or 5-fluoro-2′-deoxyuridylic acid, a potent inhibitor of DNA synthesis. Flucytosine has a limited spectrum of activity but is effective in topical application against external infections. It exhibits poor ocular penetration but can be administered orally. See also ⟶ Antimycotics – Flucytosine.

Trade name : Ancobon (Hoffmann – La Roche).

3.3. Antiprotozoal Agents

Broline [496-00-4], 4,4′-(trimethylenedioxy)bis(3-bromobenzamidine), $C_{21}H_{30}Br_2N_4O_{10}S_2$, M_r 722.4, *mp* 226 °C, is used to treat *Acanthamoeba* keratitis infections [52]. For synthesis, see [53].

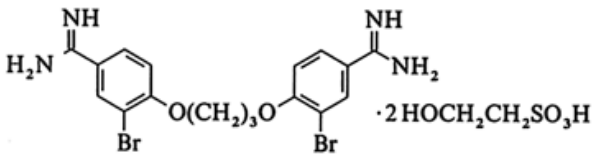

Trade names: Brolene Eye Drops (Bausch & Lomb); Broline Ointment, Brulidene (May & Baker).

Sulfadiazine [68-35-9] (silver salt [22199-08-2]), 4-amino-*N*-2-pyrimidinylbenzenesulfonamide, $C_{10}H_{10}N_4O_2S$, M_r 250.3, *mp* 252 – 256 °C. For synthesis, see [7]. See also ⟶ . This compound is used to treat toxoplasmic retinochoroiditis in triple sulfonamide mixtures with sulfamerazine and sulfamethazine.

Trade names: Coco-Diazine (Lilly), Eskadiazine (SKF).

Antibiotic – Corticosteroid Preparations. A variety of ophthalmic antibiotic – corticosteroid combination preparations are available for treating protozoal infections. Ophthalmic applications of corticosteroids and corticosteroid – antibiotic combination preparations are discussed in Section Corticosteroids.

3.4. Antiviral Agents

Currently marketed antivirals are available only as topical applications for the treatment of infection with the herpes simplex virus (HSV). These agents inhibit viral replication at the level of DNA synthesis. The prophylactic use of antivirals may reduce the frequency of viral latency [54]. The use of high dosages does not rid the host of latent virus but may reduce the frequency of recurrences during therapy [55].

Acyclovir [59277-89-3], $C_8H_{11}N_5O_3$, M_r 225.2, *mp* 256.5 – 257 °C, is an orally active acyclic nucleoside used against HSV. It is also used as a 3 % ointment in HSV infections of the epithelium. For synthesis, see [56]. See also ⟶ .

Trade name: Zovirax (Burroughs Wellcome).

Idoxuridine [54-42-2], 2′-deoxy-5-iodouridine, $C_9H_{11}IN_2O_5$, M_r 354.1. Idoxuridine is a functional analogue of the nucleoside thymidine; see ⟶ . Incorporation into DNA results in death of the viral particle through DNA base-pair mismatches and mutation. The toxicity of the substance is attributable to its incorporation into the host cell DNA. This agent is best used topically on a superficial lesion and is available as a 0.1 % solution or a 0.5 % ointment. For synthesis, see [57].

Trade names: Herpex (Allergan); Stoxil Ointment, Stoxil Solution (SKF).

Trifluridine [70-00-8], 2′-deoxy-5-(trifluoromethyl)uridine, $C_{10}H_{11}F_3N_2O_5$, M_r 296.2, *mp* 186 – 189 °C. Trifluridine is a thymidine analogue but unlike idoxuridine is rapidly metabolized and thus less toxic. Trifluridine is effective against HSV [58] and, in addition, is compatible with corticosteroids [59]. See also ⟶ . Trifluridine is generally available as a 1 % solution or a 3 % ointment. For synthesis, see [60].

Trade names: TFT-Ophtiole (Mann), TFT Eye Ointment (Mann), Viroptic (Burroughs Wellcome).

Vidarabine (anhydrous [5536-17-4], monohydrate [24356-66-9]), $C_{10}H_{13}N_5O_4 \cdot H_2O$, M_r 285.3, *mp* 257 – 257.5 °C, is an adenosine analogue produced by *Streptomyces antibioticus*. Its mechanism of action is similar to that of idoxuridine. The number and severity of adverse reactions of the two drugs are not significantly different, although vidarabine has been used to treat cases unresponsive to idoxuridine [61]. Vidarabine is available as a 3 % ointment. See also ⟶ .

Trade name: Vira-A (Parke-Davis).

[Top of Page]

4. Anti-inflammatory Agents

A variety of agents are available for the treatment of ocular inflammatory conditions, including corticosteroids, nonsteroidal anti-inflammatory agents that block mediator synthesis or release, and, more rarely, immunosuppressive agents. For general information, see ⟶ Anti-inflammatory–Antirheumatic Drugs.

4.1. Corticosteroids

For general information, see ⟶ Hormones – Adrenal Steroid Hormones .

Corticosteroids are widely used for the treatment of ocular inflammation [62-64]. Prednisolone, hydrocortisone, dexamethasone, medrysone, and fluorometholone are available in topical ocular preparations (suspension, ointment, or solution). In certain cases, topical application may be supplemented or replaced by systemic or periocular administration (e.g., severe anterior uveitis and inflammation of the posterior segment, optic nerve, or orbit) [62], [63].

The use of corticosteroids is not without potential ocular and systemic side effects (e.g., elevated intraocular pressure, posterior subcapsular cataracts, and inhibition of epithelial wound healing [60], [62], [65-67]). Because corticosteroids have immunosuppressive properties, resistance to infection is lowered. Accordingly, a number of ophthalmic preparations containing combinations of corticosteroids and antimicrobial agents are available for anti-inflammatory/anti-infective therapy (Tables 2 and 3).

Table 2. Steroid and antibiotic drops

Product	Steroid	Antibiotic	Company
Chlormycetin Hydrocortisone Powder	0.5% hydrocortisone acetate	0.25% chloramphenicol	Parke-Davis
Neo-Cortef Suspension	0.5% hydrocortisone acetate	neomycin sulfate equivalent to 0.35% base	Upjohn
Bacticort Suspension	1% hydrocortisone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Rugby
Triple-Gen Suspension	1% hydrocortisone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Goldline
Cortisporin Suspension	1% hydrocortisone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Burroughs Wellcome
Terra-Cortil Suspension	1.5% hydrocortisone acetate	0.5% oxytetracycline	Roerig
Ak-Neo-Cort Suspension	1.5% hydrocortisone acetate	neomycin sulfate equivalent to 0.35% base	Akorn
Cor-Oticin Suspension	1.5% hydrocortisone acetate	neomycin sulfate equivalent to 0.35% base	Americal
Ortho Drops Suspension	1.5% hydrocortisone acetate	neomycin sulfate equivalent to 0.35% base	Vortech
Poly-Pred Suspension	0.5% prednisolone acetate	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Allergan
Pred-G Suspension	1% prednisolone acetate	gentamycin sulfate equivalent to 0.3% gentamycin base	Allergan
NeoDecadron Solution	0.1% dexamethasone phosphate	neomycin sulfate equivalent to 0.35% base	MSD
TobraDex Suspension	0.1% dexamethasone	0.3% tobramycin	Alcon
Dexasporin Suspension	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Pharmafair, Rugby, and others
Ak-Trol Suspension	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Akorn
Dexacindin Suspension	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	IOLAB
Maxitrol Suspension	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/mL	Alcon

Table 3. Steroid and antibiotic ointments

Product	Steroid	Antibiotic	Company
Orthocort	0.5% hydrocortisone acetate	1% chloramphenicol and 10000 units polymyxin B/g	Parke-Davis
Cortisporin	1% hydrocortisone	neomycin sulfate equivalent to 0.35% base, 400 units bacitracin zinc, and 10000 units polymyxin B sulfate/g	Burroughs Wellcome
Coracin	1% hydrocortisone acetate	0.5% neomycin sulfate, 400 units bacitracin zinc and 10000 units polymyxin B sulfate/g	Coracin
NeoDecadron	0.05% dexamethasone phosphate	neomycin sulfate equivalent to 0.35% neomycin base	MSD
AK-Trol	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/g	Akorn
Dexacidin	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/g	IOLAB
Dexasporin	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/g	Pharmafair, Rugby, and others
Maxitrol	0.1% dexamethasone	neomycin sulfate equivalent to 0.35% base and 10000 units polymyxin B sulfate/g	Alcon

Hydrocortisone [50-23-7], 17-hydroxycorticosterone, cortisol, $C_{21}H_{30}O_5$, M_r 362.5. See also ⟶ Hormones – Naturally Occurring Glucocorticoids.

Hydrocortisone (1 %) is available in antimicrobial – anti-inflammatory ointment or suspension combination products (Tables 2 and 3).

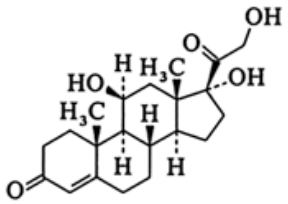

Trade names: Ak-Spore ointment (Akorn; hydrocortisone, neomycin sulfate and polymyxin B sulfate); Cortisporin ointment (Burroughs Wellcome: hydrocortisone, polymyxin B sulfate, neomycin sulfate, and bacitracin zinc); Cortisporin suspension (Burroughs Wellcome: hydrocortisone, neomycin sulfate, and polymyxin B sulfate); Ocu-Cort ointment (Ocumed: hydrocortisone, polymyxin B sulfate, neomycin sulfate, and bacitracin zinc); Ocutricin HC ointment (Pharmafair: hydrocortisone, bacitracin zinc, polymyxin B sulfate, and neomycin sulfate); Ocutricin HC suspension (Pharmafair: hydrocortisone, neomycin sulfate, and polymyxin B sulfate).

Hydrocortisone acetate [50-03-3], hydrocortisone 21-acetate, cortisol acetate, $C_{23}H_{32}O_6$, M_r 404.5.

Trade names: Hydrocortone (Merck Sharp & Dohme), Optef (Upjohn); also see Tables 2 and 3.

Prednisolone [50-24-8], 1,2-dehydrohydrocortisone, $C_{21}H_{28}O_5$, M_r 360.5. See also ⟶ Hormones.

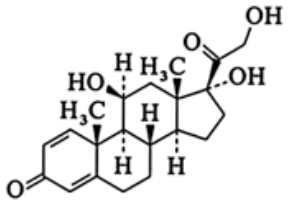

Prednisolone acetate [52-21-1], prednisolone 21-acetate, $C_{23}H_{30}O_6$, M_r 402.5.

Trade names: Ak-Tate (Akorn); Econopred, Econopred Plus (Alcon); Ocu-pred A (Ocumed); Pred Forte, Pred Mild (Allergan); see also Tables 2 and 3.

Prednisolone sodium phosphate [125-02-0], prednisolone 21-(disodium orthophosphate), $C_{21}H_{27}Na_2O_8P$, M_r 484.4, is water soluble and is available as solutions.

Trade names: Ak-Pred — 0.125 % and 1 % (Akorn); I-Pred and I-Pred 1 % (Americal); Inflamase Forte 1 %, Inflamase Mild (IOLAB); Ocu-Pred, Ocu-Pred Forte (Ocumed); Predair, Predair Forte (Pharmafair); see also Tables 2 and 3.

Dexamethasone [50-02-2], 9 α-fluoro-16 α-methylprednisolone, $C_{22}H_{29}FO_5$, M_r 392.5. See also ⟶ Hormones.

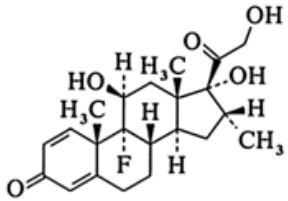

Trade names: Ak-Dex, Ak-Trol (Akorn); Baldex (Bausch & Lomb); Dexacidin (IOLAB); Decadron Phosphate (Merck Sharp & Dohme); Dexair (Pharmafair); Dexasporin (Pharmafair); Infectrol (Bausch & Lomb); Maxidex, Maxitrol (Alcon); I-Methasone (Americal); Ocu-Trol (Ocumed); Tobradex (Alcon); see also Tables 2 and 3.

Medrysone [2668-66-8], 11 β-hydroxy-6 α-methylprogesterone, $C_{22}H_{32}O_3$, M_r 344.5.

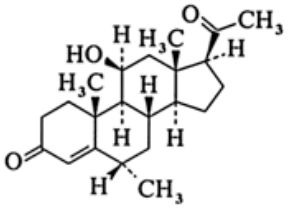

Trade names: HMS Liquifilm, Ipoflogin (Allergan); Medrifar (Farmila); Medriusar (Difa); Ophtocortin (Winzer); Sedesterol (Poen); Spectramedryn (Allergan); Visudrisone (LOA).

Fluorometholone [426-13-1], $C_{22}H_{29}FO_4$, M_r 376.5, is used ophthalmically as a 0.1 or 0.25 % suspension or as a 0.1 % ointment. See also ⟶ Hormones.

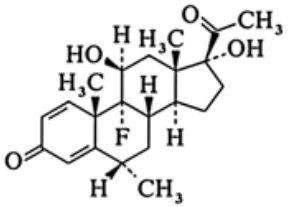

Trade names: Flucon, Isopto-Flucon (Alcon); FML Liquifilm, FML Forte Liquifilm, FML S.O.P. (Allergan); Fluor-Op (IOLAB).

4.2. Miscellaneous Anti-inflammatory Agents

Disodium cromoglycate [15826-37-6], DSCG, cromolyn sodium, sodium cromoglycate, $C_{23}H_{14}Na_2O_{11}$, M_r 5123, appears to achieve its antiallergic and anti-inflammatory effects by stabilizing mast cells and inhibiting antigen-induced mast cell degranulation and the associated release of a variety of inflammatory mediators [62], [68]. See also ⟶ Antiallergic Agents – Mast Cell Stabilizing Agents. Topical ophthalmic DSCG (Opticrom, Fisons) is available for prophylactic treatment of allergic disorders such as vernal or allergic keratoconjunctivitis. Maintaining DSCG therapy throughout the allergy season may be more advantageous than sporadic treatment of flare-ups [68]. Therapy with DSCG may permit the reduction or elimination of topical corticosteroid treatment [62].

Trade names: Opticrom (Fisons), Vividrin (Mann).

Nonsteroidal anti-inflammatory drugs (NSAIDs) inhibit the enzyme cyclooxygenase, which is responsible for the first step in the synthesis of prostaglandins from arachidonic acid. The prostaglandins mediate certain steps of the inflammatory process. The role of arachidonic acid metabolites in ocular inflammation and pressure elevation continues to emerge [62]. Oral administration of NSAIDs has been approached as an alternative to corticosteroid anti-inflammatory therapy [62], [65]. Oral administration of classical cyclooxygenase inhibitors (e.g., acetylsalicyclic acid and indomethacin) has met with mixed success in the treatment of ocular inflammation, partially due to systemic intolerance (e.g., gastrointestinal side effects). A number of more recently developed NSAIDs such as diflusinal (Dolobid) and naproxen (Naprosyn) may be better tolerated orally.

Flurbiprofen [5104-49-4], 2-fluoro- α-methyl-(1,1′-biphenyl)-4-acetic acid, $C_{15}H_{13}FO_2$, M_r 244.3. The sodium salt dihydrate of flurbiprofen is available as an ophthalmic solution (0.03 %) for topical application (Ocufen Liquifilm, Allergan) for inhibition of intraoperative miosis.

Tetracycline (see Section Tetracycline Antibiotics) is used to treat ocular rosacea [69].

Immunosuppressive agents, including alkylating agents such as cyclophosphamide, the folic acid antimetabolite methotrexate, and the antibiotic cyclosporin A, are systemically cytotoxic and dangerous drugs. Although their use in ocular therapy is rare, some of them have been used to treat intractable and progressively destructive ocular conditions [65].

[Top of Page]

5. Antiglaucomatous Agents

Glaucoma is a leading cause of blindness and may be defined as a loss of visual function and optic nerve damage associated with an elevated intraocular pressure (IOP). Ocular hypertension alone is not predictive of glaucoma or visual impairment [70]. In a healthy eye, the IOP (ca. 2 – 2.6 kPa, 15 – 20 mm Hg) [71] is maintained by a balance of aqueous humor formation and outflow. Glaucoma therapy often involves a reduction in IOP by increasing outflow or decreasing production. A variety of therapeutic agents have ocular hypotensive effects [68-74] and include adrenergic agents, miotics, carbonic anhydrase inhibitors, and hyperosmotic agents. The applications and adverse effects of ocular hypotensives are discussed in [70-80].

Glaucoma associated with congenital ocular abnormalities is treated surgically or with a combination of surgical and medical therapy. Glaucomas may be classified as primary (direct disturbance of the

aqueous circulation) or secondary (arising from other disease states). They are also classified as open angle or narrow angle, depending on the anterior chamber angle (the anatomy of the structures within the angle of the eye as visualized by gonioscopy). These classifications are important in determining whether surgery or drug therapy should be employed and which ocular hypotensive drug therapies may be appropriate.

5.1. Sympathomimetic Drugs

Epinephrine [51-43-4], adrenaline, $C_9H_{13}NO_3$, M_r 183.2, is used widely in primary open-angle glaucoma but is contraindicated in narrow-angle glaucoma [70], [71] because of its mydriatic effects (dictation of the pupil). Topically applied epinephrine has mydriatic, vasoconstricting, and ocular hypotensive effects. Although its mechanism of action is not fully understood, it may lower the IOP by reducing aqueous humor production and also increasing outflow [70-74]. Epinephrine is often used in combination with other ocular hypotensives (e.g., pilocarpine).

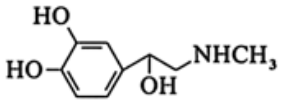

Trade names: Simplene (Smith & Nephew), 1 and 2 % solutions (Pharmafair); U.S.P. Ophthalmic Solution: a solution of epinephrine prepared with hydrochloric acid.

Epinephrine hydrochloride [55-31-2], Epifrin (0.25, 0.5, 1, or 2 %; Allergan), Glaucon (1 or 2 %; Alcon).

Epinephrine bitartrate [51-42-3], adrenaline bitartrate, $C_9H_{13}NO_3 \cdot C_4H_6O_6$, M_r 333.3, Epitrate (Wyeth-Ayerst). Epinephrine bitartrate is also available in combination with pilocarpine hydrochloride (E-Pilo-1 – 6, IOLAB)

Epinephryl borate [5579-16-8] : Eppy/N (0.5, 1.0, and 2 %, Sola/Barnes Hind).

Dipivefrin hydrochloride [52365-63-6], dipivalylepinephrine hydrochloride, dipivalyladrenaline hydrochloride, $C_{19}H_{29}O_5N \cdot HCl$, M_r 387.9. For synthesis, see [81]. Dipivefrin hydrochloride is a dipivalyl prodrug that is converted to epinephrine by esterase action. It has enhanced corneal penetration and hypotensive potency relative to epinephrine [71], [73]. Dipivefrin 0.1 % is about as potent as 2 % epinephrine.

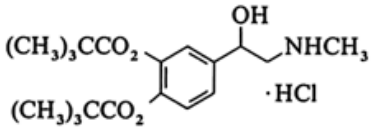

Trade names: Epifrin (Allergan), Glaucothil (Thilo), Glaudrops (Cusi), Propine (Allergan).

5.2. β-Adrenergic Blocking Agents

Although systemic blockers of the β-adrenoceptors are not approved for ocular hypotensive indications, some reduce IOP in addition to having cardiovascular activity against arrhythmias, angina, and hypertension. The β-blockers lower the IOP by decreasing aqueous humor production [72-74]. Although topical preparations are generally well tolerated, systemic (CNS, cardiovascular, and respiratory) side effects have been observed [72-74], [76-78].

Timolol maleate [26921-17-5] (timolol [26839-75-8]), (*S*)-1-[(1,1-dimethylethyl)amino]-3-[(4-morpholinyl-1,2,5-thiadiazol-3-yl)oxy]-2-propanol maleate, $C_{13}H_{24}N_4O_3S \cdot C_4H_4O_4$, M_r 432.5, is available as 0.25 and 0.5 % solutions. For synthesis, see [82]. See also ⟶ Antihypertensives – Timolol Timolol can provide an additive effect when used concurrently with other ocular hypotensive agents such as miotic agents or carbonic anhydrase inhibitors, although the effects with epinephrine are more controversial [72-74].

Trade names: Timoptic, Timoptol (Merck Sharp & Dohme).

Levobunolol hydrochloride [27912-14-7] (levobunolol [47141-42-4]), (–)-bunolol hydrochloride, (–)-5-[3-(*tert*-butylamino)-2-hydroxypropoxy]-3,4-dihydro-1(2*H*)-naphthalenone hydrochloride, $C_{17}H_{25}NO_3 \cdot HCl$, M_r 327.9. For synthesis, see [83].

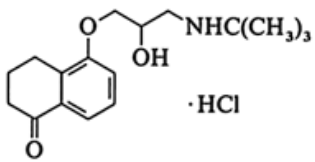

Trade name: Betagan (0.5 %, Allergan).

Betaxolol hydrochloride [63659-19-8] (betaxolol [63659-18-7]), 1-[4-[2-(cyclopropylmethoxy)ethyl]phenoxy]-3-isopropylamino-2-propanol hydrochloride, $C_{18}H_{29}NO_3 \cdot HCl$, M_r 343.9. For synthesis, see [84].

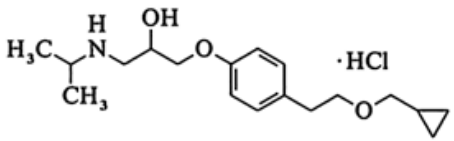

Trade name: Betoptic (0.5 % Alcon).

Metipranolol [22664-55-7], 4-(2-hydroxy-3-isopropylaminopropoxy)-2,3,6-trimethylphenyl acetate, $C_{17}H_{27}NO_4$, M_r 309.4. For synthesis, see [85], [86].

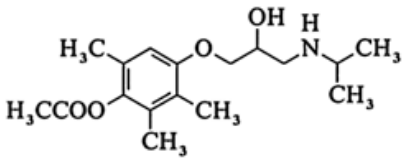

Trade names: Betamann, Beta-Ophtiole (Mann); Glauline, Minims Metipranolol (Smith & Nephew); Optipranolol (Bausch & Lomb); Normoglaucon (Mann).

5.3. Carbonic Anhydrase Inhibitors

Carbonic anhydrase inhibitors decrease the IOP by blocking bicarbonate formation in the ciliary process required for aqueous humor production [74], [79]. These agents are administered systemically and are useful in treating glaucoma cases that do not respond to topical therapy [72], [74]. Virtually all (90 – 99 %) of the carbonic anhydrase activity must be blocked before the IOP is lowered [72], [79], and oral doses are usually administered several times a day. The sulfonamide carbonic anhydrase inhibitors are diuretics that are readily absorbed from the gastrointestinal tract. Effective oral doses of these agents for glaucoma therapy depend on their pharmacokinetic properties [72], [74], [79]. Systemic side effects [72], [74], [76], [79] may limit the use of oral carbonic anhydrase inhibitors, especially in the elderly. Unfortunately, no topically effective preparations of these agents are available, possibly due to poor corneal penetration. Orally administered carbonic anhydrase inhibitors may provide additive effects when used in combination with topically applied antiglaucoma agents (e.g., pilocarpine and timolol) [74], [79].

Acetazolamide [59-66-5], 2-acetylamino-1,3,4-thiadiazole-5-sulfonamide, $C_4H_6N_4O_3S_2$, M_r 222.2. See also ⟶ Diuretics – Carbonic Anhydrase Blockers.

Trade names: Acetamide (Nessa); Diamox, Diamox Sequels (Lederle); Diuriwas (IFI); Edemox (Wasserman); Glaupax (Erco, Dispersa); Theraplix (Cycanamide, Cyanamide-Novalis).

Dichlorphenamide [120-97-8], 4,5-dichloro-1,3-benzenedisulfonamide, $C_6H_6Cl_2N_2O_4S_2$, M_r 305.2. For synthesis, see [87].

Trade names: Antidrasi (ISF, Merck Sharp & Dohme); Daranide (Merck Sharp & Dohme); Fenamide (Farmigea); Glauconide (Llorens); Glaumid (SIFI); Hipotensor Oftalmico, Oralcon, Oratrol (Alcon).

Methazolamide [554-57-4], *N*-[5-(aminosulfonyl)-3-methyl-1,3,4-thiadiazol-2(3*H*)-ylidene]-acetamide, $C_5H_8N_4O_3S_2$, M_r 236.3. For synthesis, see [88].

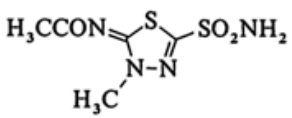

Trade names: Neptazane, Theraplix (Lederle).

5.4. Hyperosmotic Agents

Orally (glycerol, isosorbide, urea) or intravenously (mannitol) administered hyperosmotic agents produce a rapid reduction of IOP as a result of migration of water from the eye to ocular blood vessels [72], [75]. These drugs may have significant side effects [72], [75], [76]. They are used primarily to treat acute IOP elevation and in ocular surgical procedures [72].

Glycerol [56-81-5], propane-1,2,3-triol, $C_3H_8O_3$, M_r 92.1.

Trade names: Ophthalgan (Wyeth-Ayerst), Osmoglyn (Alcon).

Isosorbide [652-67-5], 1,4:3,6-dianhydrosorbitol, $C_6H_{10}O_4$, M_r 146.1.

Trade name: Ismotic (Alcon).

Mannitol [69-65-8], cordycepic acid, mannite, $C_6H_{14}O_6$, M_r 182.2.

Trade name: Mannitol Injection (Astra).

Urea [57-13-6], carbamide, ureum, CH_4N_2O, M_r 60.1.

Trade name: Ureaphil (Abbott).

5.5. Miotics (Parasympathomimetic Agents)

Miotics are agents that cause constriction of the pupil. Two classes of miotics, cholinergics and anticholinesterases (parasympathomimetic agents), have therapeutic roles in glaucoma. Acetylcholine itself is not generally useful because of its short biological half-life and topical ineffectiveness, although it can induce miosis when applied directly to the iris during surgery [73]. Cholinergic agonists (pilocarpine, carbachol) act directly on cholinergic effector cells to mimic the effects of acetylcholine; anticholinesterases (physostigmine, demecarium bromide, echothiophate iodide, isoflurophate) prolong the action of endogenous acetylcholine by blocking its hydrolysis by cholinesterases. In the eye, miotics appear to reduce IOP through activation of muscarinic acetylcholine receptors, which contract the ciliary muscle and enhance flow through the trabecular meshwork. Muscarinic activation also constricts the pupillary sphincter and produces miosis, which is sometimes an undesirable side effect of antiglaucoma therapy. Other ocular and systemic side effects may be associated with the use of miotics for antiglaucoma therapy [72], [73], [76], [77], [80].

5.5.1. *Cholinergic Agonists*

Carbachol [51-83-2], *O*-carbamoylcholine chloride, $C_6H_{15}ClN_2O_2$, M_r 182.7. Carbachol is inherently more potent and longer acting than pilocarpine but has poorer corneal penetration and more severe ocular side effects. Enhancement of corneal penetration may be achieved with 0.03 % benzalkonium chloride (BAK) in the formulation, or with lower levels of BAK along with methyl cellulose or hydroxypropyl methyl cellulose to prolong contact with the ocular surface [73]. Carbachol is used when the response to pilocarpine is inadequate or when a pilocarpine allergy develops [73], [76].

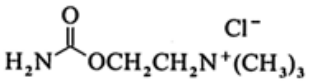

Trade names: Isopto Carbachol (Alcon), also available as a 0.01 % solution; Miostat (Alcon) for intraocular injection for miosis during surgery.

Pilocarpine [92-13-7], (3*S*,4*R*)-3-ethyldihydro-4-[(1-methyl-1*H*-imidazol-5-yl)methyl]furan-2(3*H*)-one, $C_{11}H_{16}N_2O_2$, M_r 208.3. Pilocarpine is one of the most useful drugs for management of glaucomas, including primary open-angle and acute angle-closure glaucomas [72], [73]. Principal ocular side effects of pilocarpine solutions (0.25 – 10 %) include miosis, ciliary spasm, and visual blurring [72], [73], [76]. A sustained-release drug delivery system (Ocusert) for pilocarpine base is also available [72], [73], [76].

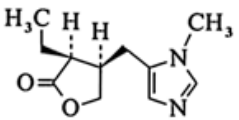

Trade names: Ocusert Pilo-20 and Ocusert Pilo-40 Ocular Therapeutic System (20 or 40-µg pilocarpine delivery per hour for one week, Alza).

Pilocarpine hydrochloride [54-71-7], $C_{11}H_{16}N_2O_2 \cdot HCl$, M_r 244.7.

Trade names: Akarpine (Akorn); I-Pilocarpine (Americal); Isopto Carpine (Alcon); Ocu-Carpine (Ocumed); Pilocar (IOLAB); Pilokair (Pharmafair); Pilopine (Alcon), and in various combinations with epinephrine bitartrate: E-Pilo-1 – 6 (IOLAB).

Pilocarpine nitrate [148-72-1], $C_{11}H_{16}N_2O_2 \cdot HNO_3$, M_r 271.3.

Trade name: Pilagan Liquifilm (Allergan).

5.5.2. *Anticholinesterases*

Anticholinesterases may be divided into two classes. The short-acting “reversible” inhibitors (e.g., physostigmine and demecarium) transfer a carbamoyl group to the enzyme, which is removed within a few

hours to regenerate the enzyme. The long-acting irreversible inhibitors (e.g., echothiophate and isoflurophate) phosphorylate the enzyme, and cholinesterase activity is restored only through synthesis of new enzyme.

The anticholinesterases are contraindicated for angle-closure glaucoma [72], and their ocular use may produce a variety of systemic and ocular side effects [72], [73], [76]. Anticholinesterases are used primarily when other antiglaucoma medications have failed and are the least commonly used antiglaucoma agents [72].

Physostigmine [57-47-6], eserine, $C_{15}H_{21}N_3O_2$, M_r 275.4.

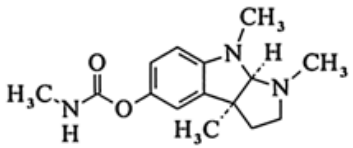

Trade names: Eserine sulfate (0.25 %) Sterile Ophthalmic Ointment (IOLAB).

Demecarium bromide [56-94-0], $C_{32}H_{52}Br_2N_4O_4$, M_r 716.6.

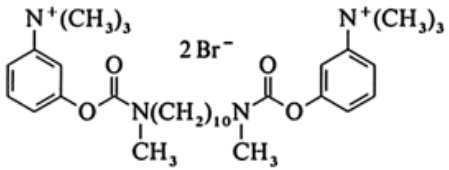

Trade names: Humorsol (Merck Sharp & Dohme), Tosmilen (Chem Linz, Sinclair).

Echothiophate iodide [513-10-0] (echothiophate [6736-03-4]), mercaptoethyltrimethylammonium iodide *O*,*O*-diethyl phosphorothioate, $C_9H_{23}INO_3PS$, M_r 383.2.

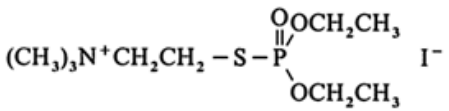

Trade names: Echofilina, Iodeto de fosfolina, Phospholine (Sodip); Phospholine Iodide (Wyeth-Ayerst, Promedica, Chinoin); Phospholinjodid (Winzer).

Isoflurophate [55-91-4], diisopropyl flurophosphate, $C_6H_{14}FO_3P$, M_r 184.1.

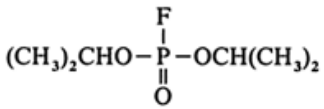

Trade names: Diflupyl (Labaz), DFP-Oel (Winzer), Floropryl (Merck Sharp & Dohme).

[Top of Page]

6. Mydriatics and Cycloplegics

Mydriatic and cycloplegic agents are used routinely in ophthalmic practice for dilating the pupil to facilitate examination of the retina [89-98]. In addition to dilation, cycloplegics cause paralysis of accommodation for near vision and are used primarily as an aid in refraction and in the treatment of uveitis [97]. They are also used as postoperative agents in cataract and retinal detachment surgery.

Mydriatics and cycloplegics are divided into sympathomimetics and parasympatholytics. They can be used alone or in combination. Sympathomimetic agents imitate (direct acting) or potentiate (indirect acting) the action of norepinephrine at sympathetic nerve terminals and produce mydriasis by stimulating the iris dilator muscle fibers. Parasympatholytic agents block the action of acetylcholine at the neuromuscular junction and produce pupil dilation with loss of accommodation by immobilizing the iris sphincter and ciliary muscle.

Although serious side effects from mydriatics and cycloplegics are rare, individual response to these drugs varies [98]. They should be used with caution in patients with closed-angle glaucoma or in patients with a narrow angle between the iris and the cornea because they may increase the IOP and precipitate an acute attack [98].

6.1. Sympathomimetics

Phenylephrine hydrochloride [61-76-7], (*S*)-1-(3-hydroxyphenyl)-2-methylaminoethanol hydrochloride, $C_9H_{13}NO_2 \cdot HCl$, M_r 203.7. Phenylephrine hydrochloride is used topically (2.5 and 10 % solutions) to dilate the pupil both for ophthalmoscopy in the treatment of uveitis and for cataract surgery [97]. Maximum dilation occurs within 45 – 60 min and recovery in about 6 h. A 1 % solution can be used in the diagnosis of Horner's syndrome [97].

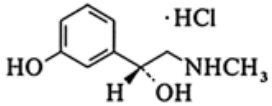

Trade names: Ak-Dilate, Ak-Vernacon (Akorn); Blephamide (Allergan); I Phrine (Americal); Isonefrine (Allergan); Isopto Frin (Alcon); Minims Phenylephrine Hydrochloride (Smith & Nephew); Mydfrin (Alcon); Neosynephrine (Winthrop); Opstin (Allergan); Phenylephrine Hydrochloride (IOLAB); Prefin (Allergan); Vasosulf (Cooper Vision); Vistosan (Allergan); Zincfrin (Alcon).

Hydroxyamphetamine hydrobromide [306-21-8], (±)-4-(2-aminopropyl)phenol hydrobromide, $C_9H_{13}NO \cdot HBr$, M_r 232.1. For synthesis, see [99]. Hydroxyamphetamine hydrobromide is an indirect-acting sympathomimetic agent. It is used in a 1 % solution as a mydriatic (effect comparable to 2.5 % phenylephrine) and in the diagnosis of Horner's syndrome [98]. Hydroxyamphetamine hydrobromide may be used in patients allergic to phenylephrine.

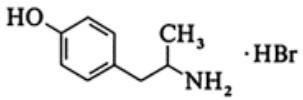

Trade name: Paredrine (SKF).

Cocaine [53-21-4], $C_{17}H_{21}NO_4 \cdot HCl$, M_r 339.8, is an indirect-acting sympathomimetic agent. In addition to its mydriatic action, it is also a local anesthetic and vasoconstrictor. Cocaine is not available commercially as a solution for clinical use.

6.2. Parasympatholytics

Atropine sulfate [5908-99-6] (monohydrate [55-48-1]), $(C_{17}H_{23}NO_3)_2 \cdot H_2SO_4 \cdot H_2O$, M_r 694.8. See also ⟶ Alkaloids – Pyridine Alkaloids; ⟶ Antiasthmatic Agents. Atropine is the most effective cycloplegic agent and has the longest duration of action. It is used to maintain a dilated pupil after intraocular surgery. Pupil dilation occurs within 30 – 40 min and lasts up to 2 weeks.

Trade names: Alcon Opulets Atropine 1 % (Alcon); Atropine-Care Ophthalmic (Akorn); Atropine Minims (Smith & Nephew); Atropine Sulfate Ophthalmic (Alcon, Allergan, Moore, Rugby, Schein, Steris, Balon, Fougera, Lilly); Atropine Sulfate S.O.P. (Allergan); Atropisol (IOLAB); Cicloplegyl (Winzer); Isopto Atropine (Alcon); I-Tropine (Americal); Minims Atropine Sulfate (Smith & Nephew); SMP Atropine (Cooper Vision).

Cyclopentolate hydrochloride [5870-29-1], $C_{17}H_{25}NO_3 \cdot HCl$, M_r 327.9. For synthesis, see [100]. Cyclopentolate hydrochloride has a rapid onset (30 – 60 min) but shorter duration of action (< 24 h) than atropine [97].

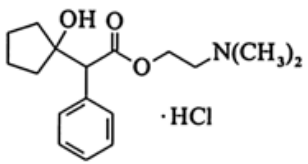

Trade names: Alcon Opulets Cyclopentolate (Alcon); Ak-Pentolate (Akorn); Ciclolux (Allergan); Ciclopegic, Colircusi ciclopejico, Colirio Oculos Cicloplegic (Llorens), Cyclopen, Cyclopentol Colircusi, Cyclomydril, Cyclogyl (Alcon); Cyclopentolat Minims, Cyplegin, Mydplegin, Minims Cyclopentalate Hydrochloride (Smith & Nephew); Mydrilate (Boehringer, Ingelheim); Skiacol (Pos); Zyklolat (Mann).

Homatropine hydrobromide [51-56-9], $C_{16}H_{21}NO \cdot HBr$, M_r 356.3. Homatropine hydrobromide is weaker and less toxic than atropine. The onset of action occurs in about 15 – 20 min and lasts ca. 3 h. Complete recovery time is about 36 – 48 h. Homatropine is used for refraction, in the treatment of inflammatory conditions of the uveal tract, and as a pre- and postoperative agent. It is also used as an optical aid for axial lens opacities [97].

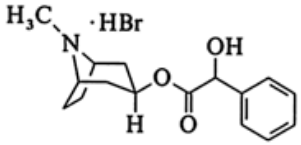

Trade names: Ak-Homatropine (Akorn), Allergan Homatropine (Allergan), Ancatropine Infant Drops (Anca), Homatrine (Americal), Homatropine HBr Ophthalmic (IOLAB), Isopto Homatropine (Alcon), Minims Homatropine Hydrobromide (Smith & Nephew).

Scopolamine hydrobromide [114-49-8], (trihydrate [6533-68-2]), $C_{17}H_{21}NO_4 \cdot HBr \cdot 3\ H_2O$, M_r 438.3. See also ⟶ Alkaloids – Chemical and Physical Properties. Scopolamine hydrobromide is an effective cycloplegic and is used in the treatment of uveitis, in refraction of children, postoperatively, and in patients sensitive to atropine. Mydriasis and cycloplegia occur within 20 – 60 min and last 3 – 7 d. The duration of action is much shorter in eyes with inflammation.

Trade names: Contac (Allergan), Isopto Hyoscine (Alcon), Minims Hyoscine Hydrobromide (Smith & Nephew), Murocoll-2 (Bausch & Lomb).

Tropicamide [1508-75-4], *N*-ethyl-*N*-(4-pyridylmethyl)tropamide, $C_{17}H_{20}N_2O_2$, M_r 284.4. For synthesis, see [101]. Tropicamide is used as a shorter-acting substitute for atropine or scopolamine when prolonged mydriasis and cycloplegia are not required. It is an effective mydriatic with weak cycloplegic activity and is, therefore, useful for ophthalmoscopy and some preoperative and postoperative states. Onset of mydriasis takes place within 15 – 20 min and lasts ca. 7 h [98]. Mydriasis may be counteracted by local application of pilocarpine. Maximum cycloplegia occurs within 20 – 25 min; the duration of this peak effect is 15 – 20 min and has passed in 6 – 8 h.

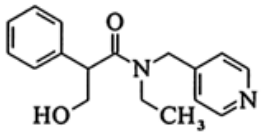

Trade names: I-Picamide (Americal), Minims Tropicamide (Smith & Nephew), Mydriacyl (Alcon), Mydrian (Dispersa), Mydriaticum (Merck Sharp & Dohme), Phenyltrope (Akorn), Tropicacyl (Akorn), Tropikamid Minims (Smith & Nephew), Tropimil (Farmigea), Visumidriatic (Merck Sharp & Dohme).

[Top of Page]

7. Vasoactive (Adrenergic) Agents

In addition to their mydriatic effect, adrenergic agents constrict the vascular system of the conjunctiva within minutes (α-adrenergic effect) and are used to treat congestion and relieve minor allergy irritation and itching of the conjunctiva. Commercially available ophthalmic vasoconstrictive agents usually contain ephedrine, naphazoline, oxymetazoline, phenylephrine, or tetrahydrozoline. Side effects are not typically observed because of the relatively low concentrations used but rebound congestion can occur with extended use [97].

Naphazoline hydrochloride [550-99-2], 2-(1-naphthylmethyl)-2-imidazoline hydrochloride, $C_{14}H_{14}N_2 \cdot HCl$, M_r 246.7. For synthesis, see [102]. One or two drops of naphazoline solution (0.012 – 0.1 %) are administered in the eye every 3 – 4 h.

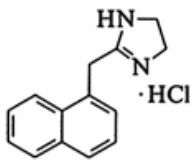

Trade names: Ak-Con (Akorn); Albalon (Allergan); Allerest Eye Drops (Pharmacraft); Allergy Drops (Bausch & Lomb); Clear Eyes (Ross); Clera (Person and Covey); Comfort Eye Drops, Degest 2 (Barnes-Hind); Imidazyl (Allergan); Naphcon (Alcon); I-Naphline (Americal); Opcon (Bausch & Lomb); Optazine (Lederle); Otrivin, Privin, Privina, Privine (Ciba); Vasocon Regular (IOLAB).

Naphazoline formulations that also contain an antihistamine are also available: Ak-Con-A (Akorn); Opcon-A (Bausch & Lomb); Naphcon-A, Vasocon-A (IOLAB); Vasoclear (IOLAB); Vistalbalon, Albalon A Liquifilm (Allergan); Zincfrin-A also contains zinc sulfate (Alcon).

Oxymetazoline hydrochloride [2315-02-8], 2-(4-*tert*-butyl-3-hydroxy-2,6-dimethylbenzyl)-2-imidazoline hydrochloride, $C_{16}H_{24}N_2O \cdot HCl$, M_r 296.8.

Trade names: Ocuclear Eye Drops (Schering), Visine L.R. Eye Drops (Pfizer).

Phenylephrine hydrochloride [61-76-7], (*S*)-1-(3-hydroxyphenyl)-2-methylaminoethanol hydrochloride, $C_9H_{13}NO_2 \cdot HCl$, M_r 203.7. Phenylephrine is used at a concentration of 0.12 %. This relatively low concentration causes vasoconstriction with little or no pupillary dilation. One or two drops are instilled in the eye(s) three or four times daily to relieve burning and itching in mild cases of noninfectious conjunctivitis.

Trade names: Ak-Dilate, Ak-Vernacon, Ak-Nephrin, (Akorn); Isonefrine (Allergan); Isopto Frin, Isopto Phenylephrine (Alcon); Metaoxedrin Minims, Minims Phenylephrine Hydrochloride (Smith & Nephew); Mydfrin (Alcon); Ocu-Phrin (Ocumed); Optistin, Prefrin, Prefrin Liquifilm, Prefrin-A, Prefrin-Z (Allergan); Vasosulf (CooperVision); Relief (Allergan); Zincfrin (Alcon).

Tetrahydrozoline hydrochloride [522-48-5], 2-(1,2,3,4-tetrahydro-1-naphthyl)-2-imidazoline hydrochloride, $C_{13}H_{16}N_2 \cdot HCl$, M_r 236.7. For synthesis, see [103]. Tetrahydrozoline hydrochloride solutions (0.05 %) are used to decrease swelling of the mucosa in conjunctivitis. One or two drops are instilled in the eye up to four times daily. The onset of vasoconstriction becomes apparent within minutes.

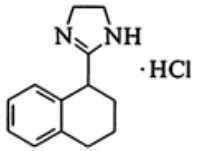

Trade names: Berberin Ophtiole (Mann); Collyrium (Wyeth); Optigene 3 (Pfeiffer); Mallazine Drops (Hauck); Murine (Abbott); Murine Plus (Ross); Soothe Eye Drops (Alcon); Visine A.C., Visine Eye Drops (Leeming); Visine Plus (G.P. Laboratories).

[Top of Page]

8. Diagnostic Agents

Diagnostic agents are used to examine the eye for signs of systemic disease and to diagnose ocular abnormalities. Dyes used for this purpose include fluorescein, rose bengal, indocyanine green [3599-32-4], trypan blue [72-57-1], alcian blue [12040-44-7], methylene blue [61-73-4], and fluorexon. Of these, only fluorescein, rose bengal, and fluorexon are commercially available [104], [105]. Several drugs are also used as ophthalmic diagnostic agents, including edrophonium chloride, methacholine, and pilocarpine.

Fluorescein C.I. 45 350 [2321-07-5] (fluorescein sodium [518-47-8]), $C_{20}H_{12}O_5$, M_r 332.3, *mp* 314 – 316 °C (decomp.). The synthesis, properties, and histological uses of fluorescein have been reviewed [106], [107]. Fluorescein sodium is used typically as a 1 or 2 % solution or as impregnated filter paper strips for detecting foreign bodies and examination of the corneal surface. During corneal surgery, topical administration of fluorescein can be used to detect leakage from the anterior chamber. In cataract surgery, leaks are detected as bright green rivulets [104], [105].

Fluorescein is used for the fitting and management of hard contact lenses. In the presence of fluorescein the tear layer fluoresces green under a cobalt blue light and contrasts with a blue fluorescence or the absence of fluorescence where the lens comes in contact with the cornea. Fluorescein is not useful for fitting soft contact lenses because it penetrates the lens matrix, providing insufficient contrast between the lens and the tissue [104], [105].

Fluorescein is also used intravenously in concentrations of 5 – 25 % for examination of the ophthalmic vasculature and the presence of ocular lesions.

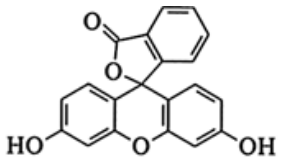

Trade names: Ak-Fluor (Akorn), Fluorescite (Alcon), Fluorets (Akorn), Fluor-I-Strip (Wyeth-Ayerst), Ful-Glo (Barnes-Hind), Funduscein (CooperVision); with proparacaine hydrochloride: Fluoracaine (Akorn).

Rose bengal, C.I. 45 440 [11121-48-5], (disodium salt [632-69-9]), $C_{20}H_2Cl_4I_4Na_2O_5$, M_r 1017.6. The dipotassium or disodium salt of rose bengal dissolves in water, giving a bluish red color [106]. Unlike fluorescein, rose bengal stains degenerated corneal and conjunctival epithelial tissue red. The stain intensity correlates with the state of degeneration, with dead cells staining intensely [108]. The dye is useful in highlighting abnormal epithelial cells in “dry eye” conditions. It is also a useful adjunct in determining the area of epibulbar squamous neoplasms [109].

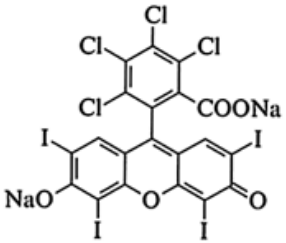

Trade names: EV Rose Bengal (Eagle Vision), Rose Bengal (Akorn).

Fluorexon [1461-15-0], $C_{30}H_{26}N_2O_{13}$, M_r 618.54; tetrasodium salt $C_{30}H_{22}N_2Na_4O_{13}$, M_r 710.5. Unlike fluorescein, fluorexon is slow to penetrate the matrix of soft contact lens materials and may be used as an adjunct for fitting soft contact lenses. (However, it is not recommended for lenses with >60 % water content [110].)

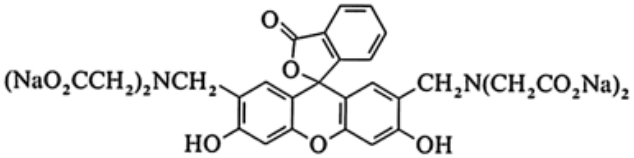

Trade name: Fluoresoft (Holles).

Edrophonium chloride [116-38-1], ethyl-(2-hydoxyphenyl)dimethylammonium chloride, $C_{10}H_{16}ClNO$, M_r 201.7, *mp* 162 – 163 °C (decomp.). For synthesis, see [111]. Edrophonium chloride (an anticholinesterase) may be injected intravenously to determine whether the cause of ptosis (drooping of the upper eyelid) is myasthenia gravis (a neuromuscular disease).

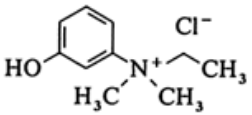

Trade name: Enion (Anaquest), Tensilon (Roche).

Methacholine chloride [62-51-1], acetyl- β-methylcholine chloride, $C_8H_{18}ClNO_2$, M_r 195.7, *mp* 172 – 173 °C. A 2.5 % methacholine (a cholinergic drug) solution is applied topically to the conjunctival sac to diagnose Adie's pupil (a disruption of cholinergic innervation to the iris). Normal pupils do not respond significantly to methacholine, whereas the Adie's pupil responds with intense miosis.

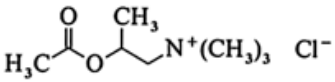

Trade name: Provocholine (Hoffmann-La Roche).

Pilocarpine, see Section Cholinergic Agonists. A 0.1 – 0.125 % solution of pilocarpine can be used to diagnose Adie's pupil. A 0.5 – 1 % solution can be used in the differential diagnosis of the fixed, dilated pupil. The dilated pupil responds to the drug if the dilation is of neurologic origin, where it does not respond if the etiology is anticholinergic in nature [104], [112].

[Top of Page]

9. Dry Eye Medications

“Dry eye” denotes a number of conditions including insufficient lacrimation, mucin deficiency, lipid abnormality, impaired blinking, or primary ocular disorder [113]. The seriousness of dry eye ranges from mild “dry and scratchy” eyes to serious degeneration and keratinization of corneal tissue, which may lead to blindness. The latter condition is often found in developing countries in areas of malnutrition (deficiency of protein and vitamin A). However, in highly developed nations, a milder form of dry eye occurs, which is characterized by chronic eye irritation and decreased visual acuity [114]. Symptoms are normally subclinical but become symptomatic in a dry, windy, or dusty atmosphere, in air conditioning; or when contact lenses are worn.

If dietary deficiency is the underlying cause of the dry eye syndrome, a proper diet should be implemented. *trans*-Retinoic acid ointment may be applied topically to counter vitamin A deficiency [115]. In dry eye conditions induced by infection, treatment with an antimicrobial agent should be initiated.

The most common treatment of nonspecific, incipient dry eye is with tear substitutes [113]. Commercial dry eye preparations typically contain at least one *demulcent*, which lubricates and wets the eye, and a *preservative* to prevent microbial contamination of the product. Nonpreserved preparations are also available in single-dose containers. Some demulcents serve as thickening agents (e.g., substituted cellulose ethers) based on the concept that highly viscous tear substitutes should have a longer retention time in the eye. Because contact lens wearers generally experience improved comfort from the lubricant and wetting properties of demulcents, many lens preparations include one or more of these agents (see Section Disinfectants and Preservatives).

Other components of dry eye medications include *emollients* (such as petrolatum, mineral oil, and lanolin), which lubricate and protect the eye from drying, and *lipids*, which are intended to supplement a deficient superficial lipid layer of tear fluid. Both emollients and lipids may, however, interfere with visual acuity.

The most common demulcents are listed below.

Carboxymethyl cellulose sodium [9004-32-4], exists as white granules, with water solubility dependent on the degree of substitution. It is available in various viscosities.

Trade names: Bro-Lac (Riker), Celluvisc (Allergan).

Dextran [9004-54-0], a polysaccharide with a backbone of D-glucose units is produced from bacteria cultured with a sucrose substrate.

Trade names: With hydroxymethyl cellulose: Muro Tears TM (Muro), Tears Naturale R (Alcon); with hydroxypropyl methyl cellulose and dextran 70: Moisture Drops (Bausch & Lomb).

Gelatin is a mixture of water-soluble proteins obtained after collagen-containing animal tissues (e.g., skin, tendons, ligaments, bones) have been boiled with water.

Trade names: Lacril R (Allergan).

Gycerol [56-81-5], propane-1,2,3-triol, $C_3H_8O_3$, M_r 92.1.

Trade names: Dry Eye Therapy (Bausch & Lomb); with dextran 70 and hydroxypropyl methyl cellulose: Moisture Drops (Bausch & Lomb).

Ophthalmic Preparation : Glycerin Ophthalmic Solution, U.S.P.

Hydroxyethyl cellulose [9004-62-0], nearly odorless, yellowish-white, white, or grayish-white, hygroscopic granules or powder.

Trade names: Adsorbo Tears (Alcon); Lyteers (Barnes-Hind); with polyvinyl alcohol: Comfort Tears, Neo Tears (Barnes-Hind); with lipid: TearGard (Bioproducts).

Contact lens products: Soft Mate ps (Barnes-Hind), Clerz (CooperVision)

Hydroxypropyl cellulose [9004-64-2] is an off-white powder that softens at 130 °C, produces aqueous solutions with a wide range of viscosities, and precipitates from solution at 40 – 45 °C.

Trade name: Lacrisert (Merck Sharp & Dohme), a sterile (prescription only) ophthalmic insert that slowly releases hydroxypropyl cellulose.

Hydroxypropyl methyl cellulose [9004-65-3] is a powder that slowly dissolves in cold water (insoluble in hot water) to give solutions with a wide range of viscosities.

Trade names: Isopto Alkaline, Isopto Plain, Isopto Tears (Alcon); Tearisol (CooperVision); Ultratears (Alcon); with gelatin and polysorbate 80: Lacril (Allergan); with dextran: Tears Naturale (Alcon); with dextran 70 and glycerol: Moisture Drops (Bausch & Lomb).

Lens care products: With poly(vinyl alcohol): Liquifilm.

Methyl cellulose [9004-67-5], consists of white granules that are soluble in cold water (insoluble in hot water), with solubility and viscosity dependent on the degree of substitution.

Trade names: Methopto Forte (Bioproducts), Methulose (Warner-Lambert), Milroy Artificial Tears (Milroy), Murocel Ophthalmic Solution (Muro), Visculose (Warner-Lambert).

Polyethylene glycol [25322-68-3], is a clear liquid or white solid that dissolves in water to form transparent solutions of various viscosities.

Trade names: With poly(vinyl alcohol): HypoTears (CooperVision).

Contact lens products: With polyoxyl 40 stearate: Blink-N-Clean (Allergan).

Poly(vinyl alcohol) [9002-89-5] is a white to cream colored powder that softens at ca. 200 °C; it is available in a variety of viscosities and solubilities in water.

Trade names: Liquifilm, Pre-Sert (Allergan); with poly(ethylene glycol); HypoTears (CooperVision); with hydroxyethyl cellulose: NeoTears (Barnes-Hind); with polyvinylpyrrolidone: Refresh, Tears Plus (Allergan).

Contact lens products: Lens-Wet, Pre-Sert, Total, Wet-N-Soak (Allergan), Hy-Flow (Cooper-Vision); with hydroxypropyl methyl cellulose: Liquifil (Allergan), Barnes-Hind Wetting Solution, Visalens Wetting Solution; with povidone and hydroxyethyl cellulose: Barnes-Hind Wetting and Soaking Solution (Barnes-Hind); with poly(ethylene glycol) and hydroxyethyl cellulose: Lensine 5 (CooperVision).

Povidone [9003-39-8], polyvinylpyrrolidone, a pale yellow solid, M_r 10 000 – 700 000, gives colloidal solutions in water.

Trade names: With other demulcents: AdsorboTear (Alcon); Refresh, Tears Plus (Allergan).

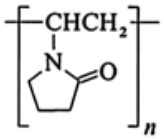

Contact lens products: Soft Care (Barnes-Hind).

Propylene glycol [57-55-6], 1,2-propane-diol, $C_3H_8O_2$, M_r 76.09, *mp* –59 °C.

Contact lens products: Allergan Hydrocare (Allergan).

[Top of Page]

10. Miscellaneous Ophthalmic and Contact Lens Preparations

Contact lenses are broadly classified as “hard” or “soft”. *Hard lenses* are typically composed of poly(methyl methacrylate) (PMMA); they are inflexible and hydrophobic, and retain their shape when removed from the eye. *Soft (“hydrophilic”) lenses* are flexible, hydrophilic (typically with a water content of 30 – 79 %), and conform to the contour of the supporting surface. They are composed of any one of a wide variety of cross-linked polymers that form a hydrophilic gel network (e.g., polymacon, a copolymer of 2-hydroxyethyl methacrylate with 2-hydroxyethyl-2-methyl-2-propenoate) [116-118]. In addition, *“semirigid” or gas-permeable lenses* are available. These lenses are commonly made from PMMA and silicone, and are a hybrid of soft and hard lenses.

The advantages and disadvantages of each type of lens are reviewed elsewhere [116-118]. The wide range of polymers used for contact lens materials results in a broad spectrum of physicochemical interactions between the lenses, tear fluid, eye tissue, and contact lens products. For example, 20 – 30 µg of tear-fluid protein is typically deposited on a polymacon soft lens, whereas >500 µg of protein may be deposited on an etafilcon A soft lens during the same period of wear [119]. (Etafilcon A is a copolymer of 2-hydroxyethyl methacrylate, sodium methacrylate, and the trimethacrylate ester of 2-ethyl-2-hydroxymethyl-1,3-propanediol). Moreover, the compositions of the tear-fluid proteins deposited on each lens type differ: the negatively charged etafilcon A lens attracts more positively charged lysozyme than the nonionic polymacon material.

The effectiveness and cytotoxicity of the agents used for contact lens disinfection, cleaning, lubrication, wetting, etc., vary markedly with the properties of the lens material. Many factors may affect product efficacy, safety, and utility. Heat disinfection is unsatisfactory for some hydrophilic lens materials because it reduces lens life and makes deposits more difficult to remove after thermal disinfection. Chemical *disinfectants and preservatives* are described in Section Disinfectants and Preservatives. Hydrogen peroxide (3 %) is a good antimicrobial agent, but its use requires a neutralization step and has patient compliance problems. Thimerosal is also a good antimicrobial agent, but a substantial proportion of the population has become sensitized to it. Benzyl alcohol is an effective preservative but tends to irritate the eye [120]. Chlorobutanol is only a fair preservative because it has a slow kill rate, is unstable in solutions above pH 6, and is adsorbed by containers. Ethylenediaminetetraacetic acid is not an effective antimicrobial agent when used alone; it disrupts the integrity of bacterial cell walls and is used in combination with other preservatives.

Boric acid (often used as a buffering agent in ophthalmic and contact lens care solutions) and sorbic acid are preservatives with relatively low cytotoxicity, but they are not useful as lens disinfectants. Quaternary amines such as benzalkonium chloride are good disinfectants and preservatives but accumulate on negatively charged hydrophilic lenses with a high water content, causing cytotoxic responses [121-126]. Polyquad is the only quaternary amine currently in use for soft lens disinfection.

Biguanides are both potent disinfectants and good preservatives. Unfortunately, chlorhexidine has been associated with sensitivity reactions. In contrast, poly(hexamethylenebiguanide) (PHMB) has a very low frequency of adverse patient response and is safe for use with soft lens materials. Lenses disinfected with PHMB solution (0.5 ppm) can be placed directly on the eye, without the saline rinse required by most other disinfectants.

Considerable variation exists among individuals in both tear chemistry and tear deposits on contact lenses [127-132].

Most *lens cleaning agents* are used daily or weekly (see Section Contact Lens Cleaners and Rewetting Agents). Daily cleaners (e.g., macrogol esters and ethers, sodium lauryl sulfate, and block

copolymers) remove proteinaceous and lipoidal lens deposits by their surfactant action. However, daily cleaners do not completely remove protein deposits, and proteins tend to accumulate with time, serving as a possible site for microbial growth and causing eye irritation and lens clouding. Weekly cleaners are used to remove accumulated protein deposits. They are generally formulations of proteolytic enzymes (e.g., pancreatin, papain, or subtilisin-A), which cleave the protein into smaller fragments that are more readily rinsed off the lens surface.

In addition to disinfectants, preservatives, and cleaning agents, contact lens preparations may contain *lubricants*, *wetting agents*, or *viscosity adjusters* to improve lens comfort. Formulations containing such materials are listed in Chapter Dry Eye Medications (contact lens products). Contact lens solutions also contain *buffers* (e.g., borate, citrate, and phosphate) and *osmolality adjusting agents* (usually sodium chloride).

10.1. Disinfectants and Preservatives

10.1.1. *Biguanides*

See ⟶ Disinfectants.

Chlorhexidine gluconate [18472-51-0], chlorhexidine digluconate, $C_{22}H_{30}Cl_2N_{10} \cdot 2\ C_6H_{12}O_7$, M_r 897.8.

Contact lens products: Bausch & Lomb Sterile Disinfecting Solution (Bausch & Lomb); Flexcare, Flexsol (Alcon); Soft-Mate (Barnes-Hind).

Poly(hexamethylenebiguanide) hydrochloride [27083-27-8], polyaminopropyl biguanide hydrochloride $[C_8H_{17}N_5]_n \cdot HCl$. For synthesis, see [120].

Contact lens products: ReNu Multi-Purpose Solution; ReNu Saline Solution (Bausch & Lomb).

10.1.2. *Mercurial Preservatives*

Phenylmercury(II) acetate [62-38-4], $C_8H_8HgO_2$, M_r 336.8, *mp* 149 °C.

Trade name: Blinx (Barnes-Hind).

Phenylmercury(II) nitrate [8003-05-2], $C_{12}H_{11}Hg_2NO_4$, M_r 634.4, *mp* 187 – 190 °C (decomp.).

Contact lens preparation: Clean-N-Soak (Allergan).

Thimerosal [54-64-8], $C_9H_9HgNaO_2S$, M_r 404.84.

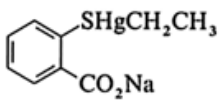

Trade names: Absorbonac, Absorbotear (Alcon); Collyrium (Wyeth); Eye Cool, 20/20 Eye Drops (Milroy); Liquifilm Forte, Prefin Z Liquifilm (Allergan); M/Rinse (Milroy); Neo Tears (Barnes-Hind); Soothe (Alcon).

Contact lens products: Adapettes (Alcon); Cleaning & Disinfecting (Allergan); Clerz (Cooper Vision); Dual-Clean (Sherman); Flex-Care, Flexsol (Alcon); Gel-Clea (Barnes-Hind); LC-65, Lens-Wet (Allergan); Pena-Vel II (Serman Labs.); Soaclens (Alcon); Soft Care, Soft Mate (Barnes-Hind); Stay Brite (Sherman); Sterile Disinfecting Solution, Sterile Lens Lubricant (Bausch & Lomb).

10.1.3. *Quaternary Ammonium Compounds*

See ⟶ Disinfectants.

See ⟶ Biocides – Quaternary Ammonium Compounds.

Benzalkonium chloride [8001-54-5] is a mixture of alkyldimethylbenzylammonium chlorides.

Trade names and contact lens products are too numerous to list.

Benzethonium chloride [121-54-0], $C_{27}H_{42}ClNO_2$, M_r 448.1, *mp* 164 – 166 °C.

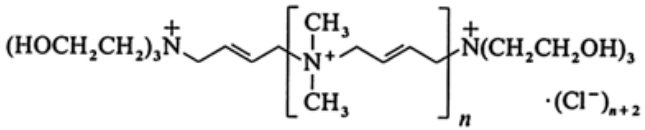

Cetylpyridinium chloride [123-03-5], 1-hexadecylpyridinium chloride monohydrate, $C_{21}H_{38}ClN \cdot H_2O$, M_r 358.0, *mp* 77 – 83 °C.

Cetyltrimethylammonium bromide [505-86-2], cetrimide, is chiefly trimethyltetradecylammonium bromide together with dodecyl- and hexadecyltrimethylammonium bromides.

Trade names: Pilomann, Dexamytrex Ophtiole (Mann).

Polyquad [75345-27-6], poly[(dimethyliminio)-2-butene-1,4-diyl chloride], polyquarternium-1, polyquat.

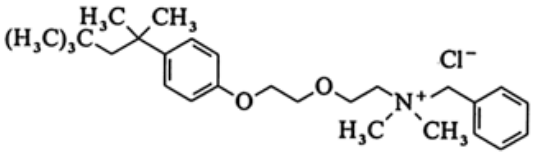

Contact lens products: Opti-Free (Alcon).

10.1.4. *Miscellaneous Disinfectants and Preservatives*

Benzyl alcohol [100-51-6], C_7H_8O, M_r 108.1, *mp* –15.2 °C. See also ⟶ Benzyl Alcohol.

Boric acid [10043-35-3], orthoboric acid, BH_3O_3, M_r 61.84, *mp* ca. 171 °C.

Commercial products: Boric acid is used in a variety of ophthalmic and contact lens care products.

Chlorobutanol [57-15-8], 1,1,1-trichloro-2-methyl-2-propanol, $C_4H_7Cl_3O$, M_r 177.5, *mp* 97 °C.

Trade names: Gentamytrex Ophtiole (Mann); Lacril, Liquifilm Tears, Tears Plus (Allergan).

Ethylenediaminetetracetic acid [60-00-4], EDTA, $C_{10}H_{16}N_2O_8$, M_r 292.2, *mp* >300 °C. See also ⟶ Ethylenediaminetetraacetic Acid and Related Chelating Agents.

Commercial products: EDTA has widespread use in a variety of ophthalmic and contact lens care products.

Hydrogen peroxide [7722-84-1], H_2O_2, M_r 34.0. See also ⟶ Hydrogen Peroxide.

Contact lens products: OxySept (Allergan); Consept (Barnes-Hind); Quik-Sept (Bausch & Lomb); AOSept (Ciba); MiraSept (CooperVision); Murine Pure Sept (Ross).

Sorbic acid [110-44-1], 2,4-hexadienoic acid, M_r 112.12, *mp* 134.5 °C.

Ophthalmic and contact lens products: Sorbic acid is used widely as a preservative in contact lens care products.

10.2. Contact Lens Cleaners and Rewetting Agents

Most lens cleaning agents are used daily or weekly (see Section Disinfectants and Preservatives). Daily cleaners remove proteinaceous and lipoidal deposits by their surfactant action. Weekly cleaners are used to remove accumulated protein deposits.

10.2.1. *Surfactant Cleaners*

Octylphenoxypolyethoxy ethanol [9002-93-1], poly(ethylene glycol) 4-isooctylphenyl ether, $C_{34}H_{62}O_{12}$, average M_r 647.

Contact lens products: Barnes-Hind Wetting & Soaking, Soft-Mate (Barnes-Hind).

Polyoxyl 40 stearate [9004-99-3], poly(ethylene glycol) monostearate.

Contact lens product: Blink-N-Clean (Allergan).

Poly(ethylene glycol) [25322-68-3], typical lens product M_r range 190 – 420; hygroscopic, viscous liquid.

Contact lens products: Blink-N-Clean (Allergan), ReNu Effervescent Enzymatic Contact Lens Cleaner (Bausch & Lomb).

Tyloxapol [25301-02-4], 4-(1,1,3,3-tetramethylbutyl)phenol polymer with formaldehyde and ethylene oxide, tyloxypal.

Trade name: Enuclene (Alcon).

Sodium lauryl sulfate [151-21-3], sodium dodecyl sulfate, $C_{12}H_{25}NaO_4S$, M_r 288.4

Contact lens products: Lens Plus Daily Cleaner, Resolve/GP Daily Cleaner (Allergan).

Poloxamer 407 [9003-11-6], ethylene – propylene block copolymer, average M_r 4000 with 70 wt % polyethylene.

Contact lens products: Ciba Vision Lens Drops, MiraFlow Extra-Strength Cleaner (Ciba).

Poloxamine [110617-70-4], Tetronic, copolymer of ethylene oxide, propylene oxide, and tetrakis ether of ethylenediamine with propanol.

Contact lens products: ReNu Multi-Purpose Solution (Bausch & Lomb).

10.2.2. *Enzymatic Cleaners*

Pancreatin [8049-47-6] is obtained from hog pancreas and has protease, amylase, and lipase activity.

Contact lens products: Optizyme (Alcon).

Papain [9001-73-4] is derived from papaya.

Contact lens products: Allergan Enzymatic, Extenzyme (Allergan).

Subtilisin A [9014-01-1] is extracted from *Bacillus licheniformis*.

Contact lens products: ReNu Effervescent Enzymatic Contact Lens Cleaner, ReNu Thermal Enzymatic Contact Lens Cleaner (Bausch & Lomb); Ultrazyme (Allergan).

[Top of Page]

References

1. J. E. F. Reynolds (ed.): *Martindale, The Extra Pharmacoepia*, 29th ed., The Pharmaceutical Press, London 1989, pp. 1205 – 1227.
2. F. W. Chang: “Local Anesthetics,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 4, Butterworths, Boston 1984.
3. S. S. Gandhi: “Local Anesthetics,” in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology,* Chap. 11, Little, Brown and Co., Boston 1987.
4. J. D. Bartlett: “Topical Anesthesia,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 14, Butterworth, Boston 1984.
5. D. B. Glasser, R. A. Hyndiuk in K. F. Tabbara, R. A. Hynduik (eds.): *Infections of the Eye*, Little, Brown and Co., Boston 1986, p. 211.
6. M. L. Crossley, E. H. Northey, M. E. Hultquist, *J. Am. Chem. Soc.* **61** (1939) 2950. Links
7. R. O. Roblin, Jr., J. H. Williams, P. S. Winnek, J. P. English, *J. Am. Chem. Soc.* **62** (1940) 2002. Links
8. H. Lundbeck, US 2 447 702, 1948 (O. Huber).
9. Hoffman-La Roche, US 2 430 094, 1947 (H. M. Wuest, M. Hoffer).
10. F. P. Doyle, G. R. Fossker, J. H. C. Nayler, *J. Chem. Soc.* 1962, 1440.
11. F. P. Doyle *et al.*, *J. Chem. Soc.* 1962, 1453.
12. E. G. Brain, J. H. C. Nayler, US 3 282 926, 1966.
13. N. Koniwhi et al., US 3 516 997, 1970.

14. R. R. Chauvette *et al.*, *J. Am. Chem. Soc.* **84** (1962) 3401. Links
15. B. M. Duggar, *Ann. NY Acad. Sci.* **51** (1948) 177.
16. A. C. Finlay *et al.*, *Science* **111** (1950) 85. Links
17. J. H. Boothe *et al.*, *J. Am. Chem. Soc.* **75** (1953) 4621. Links
18. L. H. Conover *et al.*, *J. Am. Chem. Soc.* **75** (1953) 4622. Links
19. K. L. Rinehart, Jr. *et al.*, *J. Am. Chem. Soc.* **82** (1960) 3938.
20. M. J. Weinstein, G. M. Luedemann, E. M. Oden, G. H. Wagman, *Antimicrob. Agents Chemother.* 1963, 1. Links
21. R. Okachi *et al.*, *J. Antibiot.* **27** (1974) 793. Links
22. S. A. Waksman, H. A. Lechevalier, *Science* **109** (1949) 305.
23. S. A. Waksman, H. A. Lechevalier, D. A. Harris, *J. Clin. Invest.* **28** (1949) 934. Links
24. E. A. Swart, S. A. Waksman, D. Hutchison, *J. Clin. Invest.* **28** (1949) 1045.
25. M. J. Weinstein *et al.*, *J. Antibiot.* **23** (1970) 551. Links
26. A. Schatz, E. Bugie, S. A. Waksman, *Proc. Soc. Exp. Biol. Med.* **55** (1944) 66.
27. W. M. Stark, N. G. Knox, R. M. Wilgus, *Folia Micobiol. (Prague)* **16** (1971) 205 – 217.
28. B. A. Johnson, H. Anker, F. Meleney, *Science* **102** (1945) 376.
29. H. S. Anker, B. A. Johnson, J. Goldberg, F. L. Meleney, *J. Bacteriol.* **55** (1948) 249. Links
30. T. Suzuki, H. Inouye, K. Fujikawa, Y. Suketa, *J. Biochem. (Tokyo)* **54** (1963) 25. Links
31. K. Vogler, R. O. Studer, *Experientia* **22** (1966) 345 –354. Links
32. H. Paulus: “Polymyxins,” in D. Gottlieb, P. D. Shaw (eds.): *Antibiotics,* vol 2, Springer-Verlag, New York 1967, 254 – 267.
33. Q. R. Bartz, *J. Biol. Chem.* **172** (1948) 445. Links
34. D. Gottlieb, P. K. Bhattacharyya, H. W. Anderson, H. E. Carter, *J. Bacteriol.* **55** (1948) 409. Links
35. J. Ehrlich *et al.*, *J. Bacteriol.* **56** (1948) 467.
36. R. D. Birkenmeyer, F. Kagan, *J. Med. Chem.* **13** (1970) 616. Links
37. J. M. McGuire *et al.*, *Antibiot. Chemother.* **2** (1952) 281.
38. D. J. Mason, A. Dietz, C. Deboer, *Antimicrob. Agents Chemother.* 1962, 555. Links
39. R. R. Herr, M. E. Bergy, *Antimicrob. Agents Chemother.* 1962, 560. Links
40. I. Hayakawa, T. Hiramitsu, US 4 382 892, 1983. H. Egawa, T. Miamoto, J. Matsumoto, *Chem. Pharm. Bull.* **34** (1986) 4098.
41. M. H. McCormick *et al.*, *Antibiotic Annu.* **1957–1958** (1958) 906.
42. B. R. Jones, *Tr. Am. Acad. Ophthalmol. Otolaryngol.* **79** (1975) 15.
43. T. H. Petit, R. J. Olson, R. Y. Foos, W. J. Martin, *Arch. Ophthalmol.* **98** (1980) 1025. Links
44. J. Francois, M. Rijsselaere, *Ann. Ophthalmol.* **6** (1974) 207. Links
45. D. P. Yolton: “Anti-Infective Drugs,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* 2nd ed., chap. 6, Butterworths, Boston 1989.
46. W. Gold, H. A. Stout, J. F. Pagano, R. Donovick, *Antibiotic. Annu.* 1955 – 1956 (1956) 579.
47. R. Brown, E. L. Hazen, *Trans NY Acad Sci* **Ser. II 19** (1956 – 1957) 447 – 456.
48. A. P. Struyk *et al.*, *Antibiotic Annu.* 1957 – 1958 (1958) 878 – 885.
49. E. F. Godefroi, J. Heeres, J. Van Cutsem, P. A. J. Janssen, *J. Med. Chem.* **12** (1969) 784. Links
50. J. Heeres, L. J. J. Backx, J. H. Mostmanns, J. Van Custem, *J. Med. Chem.* **22** (1979) 1003. Links
51. R. Duschinsky, E. Pleven, *J. Am. Chem. Soc.* **79** (1957) 4559. Links
52. J. J. Wiens, W. B. Jackson, *Can. J. Ophthalmol.* **23** (1988) 107; Links
53. S. S. Berg, G. Newbery, GB 598 911, 1948. S. S. Berg, G. Newbery, *J. Chem. Soc.* **1949,** 642.
54. D. Pavan-Lanston, N. H. Park, J. H. Laas, *Arch. Ophthalmol.* **97** (1979) 1964.
55. A. B. Nesburn, D. E. Willey, M. D. Trousdale, *Proc. Soc. Exp. Biol. Med.* **172** (1983) 316. Links
56. H. Matsumoto *et al.*, *Chem. Pharm. Bull.* **36** (1988) 1153. Links
57. F. Prusof, *Biochim. Biophys. Acta* **32** (1959) 295.
58. D. Pavan-Lanston, J. Lass, R. Campbell, *Arch. Ophthalmol.* **97** (1979) 1132. Links
59. F. M. Polak, M. G. Burdette, L. A. Sidry, *Cornea* **1** (1982) 29.
60. C. Heidelberger, D. Parsons, D. C. Remy, *J. Am. Chem. Soc.* **84** (1962) 3597. Links

61. D. Pavan-Langston, R. A. Buchanan, *Trans. Am. Acad. Ophthalmol. Otolaryngol* **81** (1976) 813. Links
62. S. D. Jaanus: “Anti-inflammatory Drugs,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 5, Butterworths, Boston 1989.
63. B. J. Mondino, D. H. Aizuss, M. K. Farley: “Steroids,” in D. W. Lamberts, D. E. Potter (eds): *Clinical Ophthalmic Pharmacology,* chap. 5, Little, Brown and Co., Boston 1987.
64. J. E. F. Reynolds (ed.): *Martindale, The Extra Pharmacoepia*, 29th ed., The Pharmaceutical Press, London 1989, pp. 872 – 902.
65. C. S. Foster: “Nonsteroidal Anti-Inflammatory and Immunosuppressive Agents,” in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology,* chap. 6, Little, Brown and Co., Boston 1987.
66. F. M. Wilson, II, *Survey Ophthalmology* **24** (1979) 57.
67. F. T. Fraunfelder, S. M. Meyer, *Cornea* **5** (1986) 55.
68. S. I. Butrus, J. H. Weston, M. B. Abelson: “Ocular Mast Cell Stabilizing Agents,” in D. W. Lamberts, D. E. Potter (eds): *Clinical Ophthalmic Pharmacology,* chap. 18, Little, Brown and Co., Boston 1987.
69. J. B. Walsh, A. Gold: *Physician's Desk Reference for Ophthalmology*, 17th ed., Medical Economics Company, Inc., Oradell, NJ 1989, p. 9.
70. J. B. Eskridge, J. D. Bartlett: “The Glaucomas,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 29, Butterworths, Boston 1989.
71. A. J. Flach, *J. Toxicol. Cutaneous Ocul. Toxicol.* **3** (1984) 31 – 51.
72. Z. M. Shihab: “Antiglaucoma Therapy,” in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology,* chap. 7, Little, Brown and Co., Boston 1987.
73. S. D. Jaanus, V. T. Pagano, J. D. Bartlett: “Drugs Affecting the Autonomic Nervous System,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 3, Butterworths, Boston 1989.
74. J. A. Hiett: “Inhibitors of Aqueous Formation,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 7, Butterworths, Boston 1989.
75. S. D. Jaanus: “Hyperosmotic Drugs,” in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* chap. 8, Butterworths, Boston 1989.
76. K. W. Benjamin: “Toxicity of Ocular Medications” in *Int. Ophthalmol. Clin.* **19** (1979) 199.
77. B. L. Selvin, *South. Med. J.* **76** (1983) 349 – 358. Links
78. F. T. Fraunfelder, S. M. Meyer, *Medical Toxicology* **2** (1987) 287 – 293; Links *Cornea* **5** (1986) 55 – 59.
79. R. D. Newcomb, M. L. Priest: “Systemic Effects of Ocular Drugs” in J. O. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology,* Butterworths, Boston 1989, chap. 31.
80. B. L. Selvin: “Systemic Effects of Topical Ophthalmic Medications,” in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology,* chap.17, Little, Brown and Co., Boston 1987.
81. A. Hussain, J. E. Truelove, US 3 809 714, 1974;3 839 584, 1974. D. Henschler, J. Wagner, H. Hampel, US 4 085 270, 1978.
82. L. M. Weinstock, R. J. Tull, M. D. Mulvey, US 3 619 370, 1971; B. K. Wasson, US 3 655 663, 1972; L. M. Weinstock, R. J. Tull, M. D. Mulvey, US 3 657 237, 1972; B. K. Wasson et al., *J. Med. Chem.* **15** (1972) 651.
83. C. F. Schwender, J. Shavel, Jr., US 3 649 691, 1972.
84. P. M. J. Manoury, I. A. G. Cavero, H. Mayer, D. P. R. L. Guidicelli, US 4 252 984, 1981.
85. J. W. Clapp, R. O. Robin, Jr., US 2 554 816, 1951.
86. P. Gianfranco, US 2 980 679, 1961.
87. E. M. Schultz, US 2 835 702, 1958.
88. R. W. Young, K. H. Wood, J. R. Vaughn, Jr., US 2 783 241, 1957; R. W. Young et al., *J. Am. Chem. Soc.* **78** (1956) 4649.
89. A. D. Charap: “Ocular Pharmacology” in S. J. Rhode, S. P. Ginsberg (eds.): *Ophthalmic Technology,* Raven Press, New York 1987, pp. 469 – 478.
90. N. Weiner in A. G. Gilman, L. S. Goodman, T. W. Rall, F. Murad (eds.): *The Pharmacological Basis of Therapeutics*, Macmillan Publishing Co., New York 1985, pp. 130 – 144.
91. B. L. Selvin, *South. Med. J.* **76** (1983) 349. Links
92. J. H. Calhoun, *1979 Annual Meeting of the Pennsylvania Academy of Ophthalmology and Otolaryngology*, Bedford, Pennsylvania, May 16 – 19, 1979, pp. 10 – 13.
93. F. T. Fraunfelder, S. M. Meyer, *Medical Toxicology* **2** (1987) 287. Links
94. G. A. Hopkins, W. M. Lyle, *J. Am. Optom. Assoc.* **48** (1977) 1241. Links
95. T. O. Soine: “Autonomic Blocking Agents and Related Drugs,” in C. O. Wilson, O. Gisvold, R. F. Doerge (eds.): *Textbook of Organic Medicinal and Pharmaceutical Chemistry,* 7th ed. J. P. Lippincott Co., Philadelphia, Toronto 1977, p. 481.
96. P. P. Ellis in D. Vaughan, T. Asbury, K. F. Tabbara (eds.): *General Ophthalmology, Commonly Used Eye Medications*, Appleton and Lange, Norwalk, Connecticut; San Mateo, California 1989, pp. 399–407.
97. J. D. Bartlett, S. D. Jaanus (eds.): *Ophthalmic Drug Facts*, J. B. Lippincott Company, St. Louis, Missouri 1989, p. 13 – 27.
98. J. E. F. Reynolds (ed.): *Martindale, The Extra Pharmacoepia*, 29th ed., The Pharmaceutical Press, London 1989, pp. 522 – 545, pp. 1453 – 1486.
99. L. L. Abell, W. R. Bruce, J. Seifter, US 2 590 079, 1952.
100. G. R. Treves, US 2 554 511, 1951.
101. G. Rey-Bellet, H. Spiegelberg, US 2 726 245, 1955.
102. A. Sonn, US 2 161 938, 1939.
103. M. E. Syderholm, L. H. Jules, M. Sahyun, US 2 731 471, 1956; J. F. Gardocki, D. E. Hutcheon, G. D. Lanbach, S. Y. P'an, US 2 842 478, 1958.
104. S. D. Jaanus in J. D. Bartlett, S. D. Jaanus (eds.): *Clinical Ocular Pharmacology*, Butterworths, Boston 1984, pp. 311 – 326.

105. W. A. Boothe in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology*, Little, Brown and Company, Boston 1987, pp. 361 – 415.
106. S. Budavari (ed.): *The Merck Index*, 11th ed., Merck & Co., Inc., Rahway, N.J. 1989.
107. R. F. Steiner, H. Edelhoch, *Chem. Rev.* **62** (1962) 457. Links
108. M. S. Norn, *Acta Ophthalmol.* **48** (1970) 227.
109. F. M. Wilson, *Ophthalmic Surg.* **7** (1976) 21. Links
110. F. J. Holly, US 4 518 579, 1985.
111. J. A. Aeschlimann, A. Stempel, US 2 647 924, 1953.
112. J. B. Walsh, A. Gold, H. Charles in *Physicians' Desk Reference for Ophthalmology*, 18th ed., Medical Economics Company Inc., Oradell, N.J. 1990, pp. 15 – 16.
113. F. J. Holly, M. A. Lemp, *Surv. Ophthalmol.* **22** (1977) 69. Links
114. F. J. Holly in D. W. Lamberts, D. E. Potter (eds.): *Clinical Ophthalmic Pharmacology*, Little, Brown and Co., Boston 1987, pp. 497 – 518.
115. S. C. G. Tseng, *Trans. Ophthalmol. Soc. U.K.* **104** (1985) 489.
116. J. R. Boyd in *Handbook of Non-Prescription Drugs*, 8th ed., American Pharmaceutical Association, Washington D.C., pp. 453 – 476.
117. J. B. Walsh, A. Gold, H. Charles in *Physicians' Desk Reference for Ophthalmology*, 18th ed., Medical Economics Company Inc., Oradell, N.J. 1990, pp. 209 –256.
118. *Contact Lens Forum*, August (1988) 11.
119. W. Reindel *et al.*, *International Contact Lens Clinic* **16** (1989) 232.
120. G. L. Feldman, *Contact Lens Spectrum*, May (1989) 41.
121. J. L. Sterling, A. S. Hecht, *Contact Lens Spectrum*, March (1988) 62.
122. W. R. Baily, Jr., *Contact Lens Soc. Am. J.* **6** (1972) 33.
123. S. S. Davis, M. A. Watson, *J. Pharm. Pharmacol.* **33** (1981) 109 P.
124. S. Lerman, G. Sapp, *Can. J. Opthalmol.* **6** (1971) 1.
125. A. R. Gasset, *Am. J. Ophthalmol.* **84, Series 3** (1977) 169.
126. A. R. Gasset, H. E. Kaufman, *Am. J. Ophthalmol.* **69 Series 3** (1970) 252.
127. N. J. Van Haeringen, *Survey Ophthalmol.* **26** (1981) 84.
128. F. J. Holly, *Contact Intraocul. Lens Med. J.* **4** (1978) no. 2, 14.
129. F. J. Holly, *Contact Intraocul. Lens Med. J.* **4** (1978) no. 3, 52.
130. W. G. Bachman, G. Wilson, *Investigative Ophthalmology Visual Sci.* **26** (1985) 1484.
131. M. A. Lemp, *Clinical Ophthalmology* **4** (1987) 1.
132. R. L. Farris, *Trans. Am. Ophthalmol.* **83** (1985) 501.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Optical Brighteners

Adolf Emil Siegrist, Basel, Switzerland

Claude Eckhardt, Ciba-Geigy AG, Basel, Switzerland

Jürgen Kaschig, Ciba Specialty Chemicals Inc., Basel, Switzerland

Erika Schmidt, Ciba-Geigy AG, Basel, Switzerland

1. Introduction 1
1.1. Physical Principles 2
1.2. Molecular Structure 3
1.3. History of Whitening 4
2. Chemistry of Technical Products . . 4
2.1. Carbocycles 4
2.1.1. Distyrylbenzenes 4
2.1.2. Distyrylbiphenyls 5
2.1.3. Divinylstilbenes 6
2.2. Triazinylaminostilbenes 6
2.3. Stilbenyl-2*H*-triazoles 7
2.3.1. Stilbenyl-2*H*-naphtho[1,2-*d*]triazoles 7
2.3.2. Bis(1,2,3-triazol-2-yl)stilbenes 7
2.4. Benzoxazoles 7
2.4.1. Stilbenylbenzoxazoles 7
2.4.2. Bis(benzoxazoles) 9
2.5. Furans, Benzo[*b*]furans, and Benzimidazoles 10
2.5.1. Bis(benzo[*b*]furan-2-yl)biphenyls . . 10
2.5.2. Cationic Benzimidazoles 11
2.6. 1,3-Diphenyl-2-pyrazolines 11
2.7. Coumarins 12
2.8. Naphthalimides 13
2.9. 1,3,5-Triazin-2-yl Derivatives 14
3. Commercial Forms and Brands . . 14
4. Uses 15
4.1. General Requirements 15
4.2. Textile Industry 15
4.3. Detergent Industry 16
4.4. Paper Industry 17
4.5. Plastics and Synthetic Fibers 17
4.6. Other Uses 17
5. Analytical Methods and Whiteness Assessment 18
5.1. Analytical Methods 18
5.2. Assessment of Whitening Effect . . 18
6. Environmental Aspects 18
7. References 19

1. Introduction

Optical brighteners or, more adequately, fluorescent whitening agents (FWA) are colorless to weakly colored organic compounds that, in solution or applied to a substrate, absorb ultraviolet light (e.g., from daylight at ca. 300 – 430 nm) and reemit most of the absorbed energy as blue fluorescent light between ca. 400 and 500 nm.

In daylight optical brighteners can thus compensate for the aesthetically undesirable yellowish cast found in white industrial substrates, such as textiles, papers, or plastics. Furthermore, since a portion of the daylight spectrum not perceived by the eye is converted to visible light, the brightness of the material is enhanced to give a dazzling white.

Figure 1 illustrates the action of optical brighteners [1–3]. White substrates show a yellowish cast because their reflectance for incident visible light is lower at short wavelengths than at long ones (curve a), i.e., they absorb short-wavelength light.

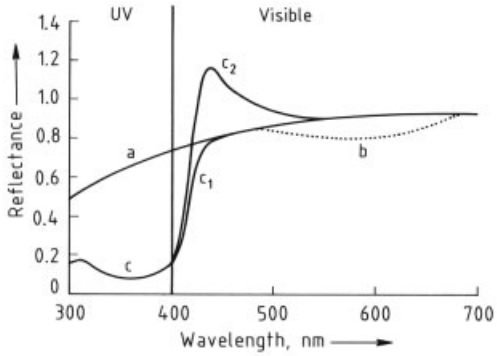

Figure 1. Reflectance spectra for a textile material
a) Bleached; b) Blued; c) Treated with optical brightener [25] (c_1 absorption, c_2 fluorescence)

The first method of compensating for the yellowish cast was the use of a blue dye such

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_153

as ultramarine or indigo. Bluing lowers the reflectance of the sample in the long-wavelength part of the visible spectrum (curve b). As a result, the sample takes on a neutral white appearance, but at the same time it loses brightness so that it looks grayer.

Optical brighteners lower the reflectance mainly in the UV and near-visible by absorption (curve c_1); at visible wavelengths (mostly with a maximum at 435 – 440 nm), they greatly increase the reflectance through fluorescence (curve c_2). The brightener acts as a supplementary emission source.

Optical brighteners are more effective the cleaner and whiter the substrate is. In no case can they replace cleaning or bleaching.

1.1. Physical Principles

Figure 2 illustrates the processes involved in light absorption and fluorescence by optical brighteners [4–8].

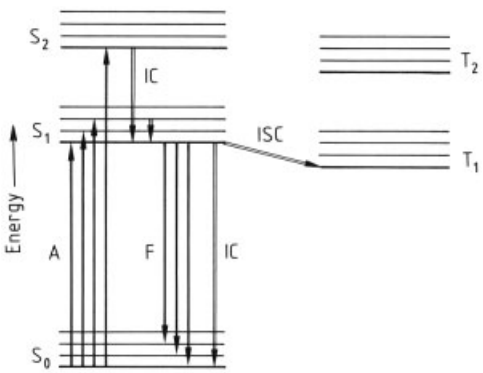

Figure 2. Energy diagram of optical brighteners and transitions
A = absorption; F = fluorescence; IC = internal conversion; ISC = intersystem crossing; S = singlet state; T = triplet state

Absorption (A) of light quanta by the brightener molecules induces transitions from the singlet ground state S_0 to vibrational levels of the electronically excited singlet states S_1 and S_2. The electronic state and vibrational level reached depend on the wavelength of the absorbed light. Absorption occurs within ca. 10^{-15} s and results in a gain in energy E and a change in the electron distribution within the brightener molecule.

Brightener molecules that have been excited to a higher electronic state such as S_2 or into a higher vibrational level of S_1 relax by a nonradiative process within ca. 10^{-12} s to the vibrational ground state of S_1 (internal conversion, IC), which has a lifetime of ca. 10^{-9} s. This time is sufficient for the brightener geometry to adapt to the electron distribution in the S_1 state.

Brighteners in the S_1 state are deactivated by several routes. Fluorescence results from radiative transitions to vibrational levels of the ground state. The fluorescence emission has a lower average energy than the absorbed light (Fig. 2), i.e., fluorescence occurs at longer wavelengths than absorption (Stokes shift).

Deactivation processes competing with fluorescence are mainly nonradiative deactivation to the S_0 state (IC) and nonradiative transition to a triplet state

(intersystem crossing, ISC). Photochemical products are often formed from this triplet state. Important photochemical reactions are the $E \rightarrow Z$ isomerization of ethylene, the oxidation of pyrazoline to pyrazole, and the dimerization of coumarins (see also → Photochemistry).

The efficiency of fluorescence is measured by the quantum yield Φ:

$$\Phi = \frac{\text{Number of quanta emitted}}{\text{Number of quanta absorbed}}$$

It is determined by the relative rates of fluorescence emission and the competing processes.

In many brighteners deactivation from the S_1 state depends on the local environment. When fixed in (highly viscous) solid substrates, brighteners fluoresce with high quantum yields (Φ ca. 0.9). In (low-viscosity) solvents, such as are used before or during application of the brightener, the quantum yield is often low and the probability of formation of photoproducts is higher.

Requirements for optimally effective brighteners are as follows:

1) Brighteners should absorb as much UV light as possible and should have little inherent color. This means that the absorption maximum should lie in the 350 – 375 nm range, the extinction coefficient should be high, and the absorption band should decrease steeply near 400 nm.
2) Fluorescence should be as intense as possible and produce maximum whiteness. This

means that the quantum yield should be near 1.0. The fluorescence maximum should be between 415 and 445 nm depending on the preferred shade (violet-blue to greenish blue), the fluorescence band should decrease as rapidly as possible on the long-wavelength side.

3) The brightener distribution in the substrate should be monomolecular, even at high concentrations because aggregates of brightener molecules show dramatically lowered quantum yields and often different spectra.

Figure 3 shows typical absorption (A) and fluorescence (F) spectra for optical brighteners. Fluorescence spectra F_1 are measured in dilute solutions. On brightened substrates at the usual (i.e., higher) brightener concentrations, short-wavelength fluorescence is reabsorbed by the brightener with the result that the fluorescence band becomes narrower (F_2). A further increase in brightener concentration shifts the center of the band toward longer wavelengths corresponding to more greenish effects, up to unwanted shades.

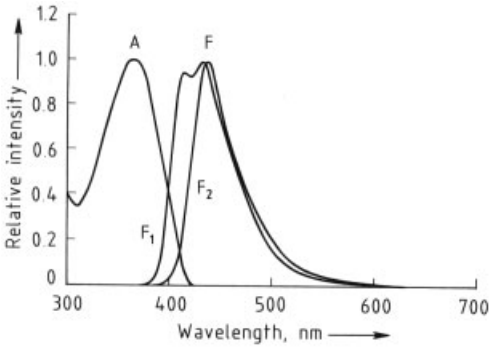

Figure 3. Typical absorption (A) and fluorescence (F) spectra of optical brighteners [25]
F_1 = dilute solution; F_2 = applied to substrate

1.2. Molecular Structure

The chromophores of optical brighteners are π-electron systems in which $\pi - \pi*$ transitions occur. The chromophores must be rigid [5] and their conformations should differ only slightly in the electronic ground state and in the first excited state.

Brightener chromophores can be regarded as combinations of building blocks with the following numbers of π electrons:

2 π electrons:	vinylene, imino
4 π electrons:	2-pyrazolin-1,3-diyl
l6 π electrons:	1,4-phenylene, 2,5-furanediyl, 2,5-thiophenediyl, phenyl, fur-2-yl, pyrazol-4-yl, pyrazol-1-yl, 1,2,3-triazol-2-yl, 1,2,4-triazol-1-yl, 1,3,5-triazin-2-yl, 1,2,4-oxadiazol-5-yl
10 π electrons:	1,4-naphthylene, benzofuran-2-yl, benzoxazol-2-yl, benzimidazol-2-yl
12 π electrons:	4,7- coumarindiyl
14 π electrons:	naphtho[1,2-*d*]triazol-2-yl
16 π electrons:	naphthalimido, pyren-1-yl

The individual building blocks absorb at wavelengths too short for brighteners. Linking the building blocks into a system of conjugated double bonds yields chromophores that absorb at longer wavelengths.

By themselves, the building blocks with two or four π electrons (chiefly vinylene and 2-pyrazolin-1,3-diyl) absorb at very short wavelengths, but they make a significant contribution to shifting the absorption of other building blocks to longer wavelengths. They are usually combined with phenyl and phenylene groups. Examples are compounds **4**, **15**, and **53**.

In contrast, the cyclic building blocks with six π electrons give a very slight bathochromic shift. If several of these building blocks are combined, brighteners of great length and high molecular mass as well as a wide range of structures can be obtained. Examples are compounds **10** and **37**.

The absorption and fluorescence of coumarin with an exocyclic π center, naphthalimide, and pyrene are close to the favorable region for brighteners, and so only slight lengthening of these chromophore systems is necessary (**67**) or possible (**74–79**).

The optical properties of chromophores can be varied by means of substituents. Electron donors (e.g., alkyl, alkoxy, or substituted amino groups) as well as electron acceptors (e.g., cyano, alkylsulfonyl, or carbalkoxy groups) can intensify fluorescence depending on their position on the chromophore.

Substituents that have little influence on the chromophore system affect the application prop-

erties and are important for substantivity (see Chap. 4).

For reviews of optical brightener systems, see [1,3,6–21].

1.3. History of Whitening

Optical brightening was discovered in 1929 by P. Krais [22]. The whiteness of viscose rayon and semibleached flax yarn was increased by treatment with an aqueous solution of esculin (**1**) [*531-75-9*] and drying. Krais recognized that this effect was due to the strong blue fluorescence of esculin, a β-D-glucoside of 6,7-dihydroxycoumarin that is obtained by extracting horse chestnut bark. (For historical references see [23].)

The first industrial application in 1935 involved the use of sodium umbelliferone acetate [*19491-88-4*] and starch [24] and, later, chiefly β-methylumbelliferone (**2**) [*90-33-5*] [6]. As with esculin, however, these treatments lacked light- and wetfastness.

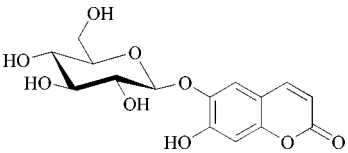

1

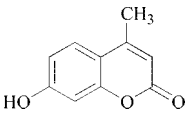

2

With the 1,3,5-triazin-2-yl derivatives of 4,4′-diaminostilbene-2,2′-disulfonic acid proposed by B. Wendt in 1940 [25], optical brighteners came into full-scale industrial use. In 1941, 4,4′-bis[(4-anilino-6-hydroxy-1,3,5-triazin-2-yl)amino]stilbene-2,2′-disulfonic acid disodium salt was introduced commercially under the trade name Blankophor B (**3**) [*1264-32-0*].

Subsequent development in all industrialized countries was rapid, especially during the period 1945 – 1975, leading to more than 1000 commercial products by 1989. Almost 400 compounds of distinct chemical structure or composition are registered in the Colour Index. Many of these compounds, including **3** and heterocycles such as imidazolone, dibenzothiophene-5,5′-dioxides, methine derivatives, dihydrocollidine, pyrazine, carbostyrils, pyrazolo [3,4-*b*] quinolines, and 5,6-benzocoumarins are no longer used and have been replaced by others.

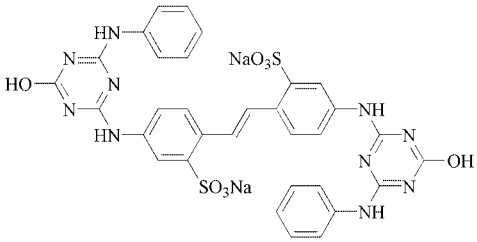

3

2. Chemistry of Technical Products

This chapter deals with the most important commercial products as well as some new developments of industrial interest.

2.1. Carbocycles

The first commercial carbocyclic products were derivatives of 4,4′-diaminostilbene-2,2′-disulfonic acid, e.g., 4,4′-bis(phenylureido)stilbene-2,2′-disulfonic acid disodium salt [*2606-93-1*] (Blankophor R, commercialized in 1942) [26] and acylamino- or methoxy-substituted benzoylamino derivatives [6]. Because of their poor efficiency (i.e., yield of whiteness) and hue, these products have disappeared from the market.

Carbocyclic brighteners gained new impetus after 1959, when the trialkylphosphite-activated ethylenation process developed by Horner became available [27]. This process, which features a high yield and high selectivity for the *E* isomers needed for optical brighteners, is extremely interesting from an industrial point of view. The principal products are distyrylarenes and divinylstilbenes, which have retained their importance to the present day.

2.1.1. Distyrylbenzenes

Cyano-substituted 1,4-distyrylbenzenes (Table 1) are carbocyclic brighteners with a very

Table 1. 1,4-Distyrylbenzenes

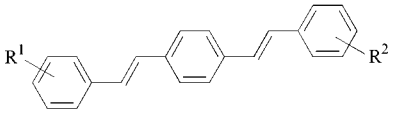

FWA	R^1 (position)	R^2 (position)	CAS no.	Ref.
5	CN (2)	CN (3)	[*79026-03-2*]	
6	CN (2)	CN (4)	[*13001-38-2*]	[29]
7	CN (3)	CN (3)	[*36755-00-7*]	[28]
8	CN (3)	CN (4)	[*79026-02-1*]	[32]
9	CN (4)	CN (4)	[*13001-40–6*]	[28]

high whiteness and good efficiency for plastics and synthetic fibers, especially polyesters [28]. Asymmetrical compounds, mixtures of two or three different di(cyanostyryl)benzenes [29–32], or combinations with polyester brighteners differing in structure (e.g., **15**, **39**, **40**, and **42**) improve uptake and diffusion into the fiber [33] and allow the white hue to be varied over a wide range. On polyester substrates, however, the di(cyanostyryl)benzenes have only moderately good lightfastness.

For the synthesis of **4**, 2-cyanobenzyl chloride is first treated with triethyl phosphite to give 2-cyano-(1-diethylphosphonomethyl)benzene [*23973-65-1*], which reacts with terephthaldehyde in the presence of sodium methoxide to yield 1,4-bis(2-cyanostyryl)benzene (**4**) [*13001-39-3*] (Palanil Brilliant White R, 1966) [28]:

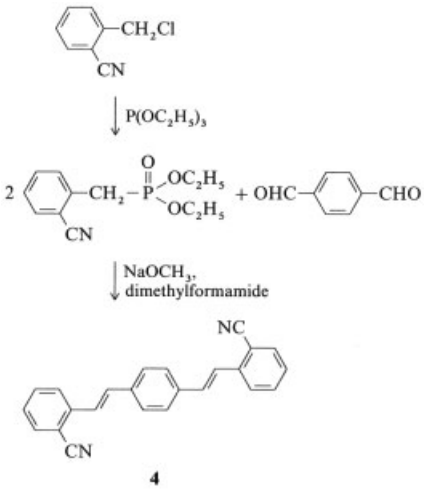

2.1.2. Distyrylbiphenyls

4,4′-Distyrylbiphenyls, discovered in 1967, can be produced in a similar way to distyrylbenzenes [34]. The reaction of 4,4′-bis(chloromethyl)biphenyl with trimethyl phosphite yields 4,4′-bis(dimethoxyphosphonomethyl)biphenyl [*27344-43-0*]. Reaction with, for example, benzaldehyde-2-sulfonic acid gives 4,4′-di(2-sulfostyryl)biphenyl disodium salt (**10**) [*27344-41-8*] (Tinopal CBS, 1972):

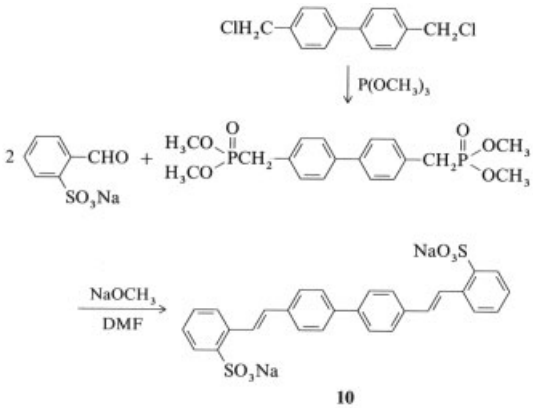

Anionic distyrylbiphenyls have very high efficiencies and are fairly chlorine-stable brighteners that are used for cellulosic fibers and especially for detergents (Table 2). Nonionic [35] and cationic compounds [36] have also been developed, for example [*72796-88-4*]:

$R^1 = -O-(CH_2)_2-{}^{+}N(CH_3)(C_2H_5)_2 \cdot CH_3OSO_3^-$

at the 2-position, $R^2 = H$ (see formula in Table 2).

Table 2. 4,4′-Distyrylbiphenyls

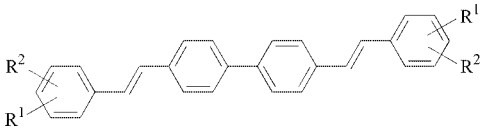

FWA	R^1 (position)	R^2 (position)	CAS no.	Ref.
11	SO_3Na (3)	Cl (4)	[*42380-62-1*]	[34]
12	OCH_3 (2)	H	[*40470-68-6*]	[35]

2.1.3. Divinylstilbenes

Polyester divinylstilbene brighteners with a very high efficiency [37] include 4,4′-di(ethoxycarbonylvinyl)stilbene (**13**) [*60683-03-6*] (Leukophor EHB, 1979), which can be obtained by the reaction of stilbene-4,4′-dicarbaldehyde with triethyl phosphonoacetate in the presence of sodium methoxide.

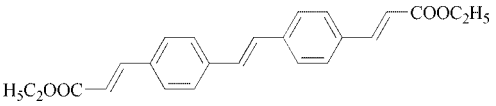

13

Another divinylstilbene brightener with an even higher efficacy is 4,4′-di(cyanovinyl)stilbene [*60682-87-3*] [37,38].

2.2. Triazinylaminostilbenes

1,3,5-Triazinyl derivatives of 4,4′-diaminostilbene-2,2′-disulfonic acid have been available since 1941 and remain the most important class of optical brighteners in terms of quantity. Some 75 derivatives have been put on the market [1]. In the textile industry, they find use in brightening cotton, regenerated cellulosic fibers, and polyamides. They can also be included in detergent formulations and used to brighten paper either in the pulp or in the coating.

The starting compound for all triazinylaminostilbenes is 4,4′-dinitrostilbene-2,2′-disulfonic acid [*128-42-7*], which is obtained by oxidizing 4-nitrotoluene-2-sulfonic acid with aqueous sodium hypochlorite in the presence of sodium hydroxide or, more recently, by atmospheric oxidation in an aqueous ammoniacal medium [39]. The Béchamps reduction using iron filings etched with hydrochloric acid yields 4,4′-diaminostilbene-2,2′-disulfonic acid [*81-11-8*] (DAS).

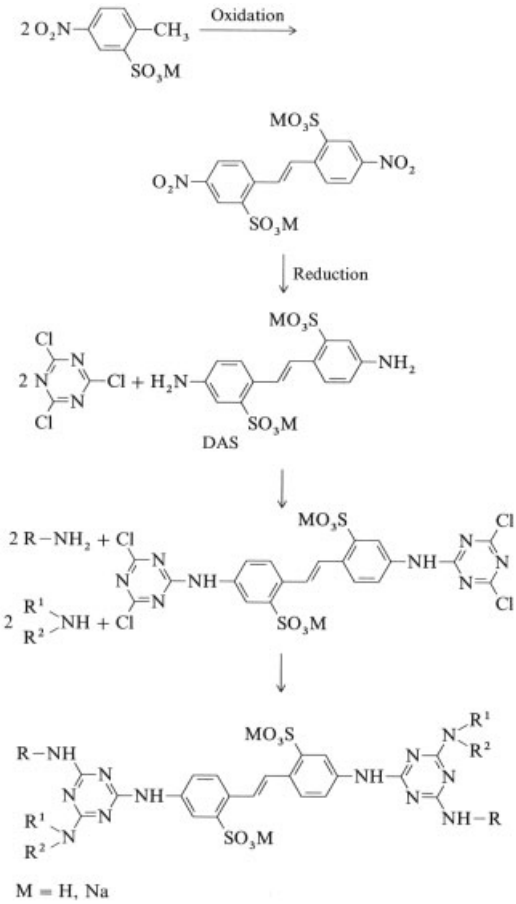

In subsequent steps, DAS is treated in acidic aqueous solution at 0 – 5 °C with cyanuric chloride dissolved in ethyl methyl ketone. The remaining chlorine atoms are then replaced by aliphatic, cycloaliphatic, or aromatic amines at 15 – 35 °C and then at ca. 60 °C. The addition of electrolytes (e.g., sodium carbonate, sodium

hydroxide) or water-soluble aprotic solvents at 90 to > 100 °C leads to the desired β crystal modification, which is nearly colorless.

Another sequence of reactions starting with DAS is needed for weakly basic amines, such as anilinesulfonic acids, and also for alcohols. These are initially reacted with cyanuric chloride, then with DAS in an aqueous medium at 40 – 50 °C. In this way the number of side reactions is substantially reduced.

Some 40 triazinylaminostilbene brighteners are currently marketed; Tables 3 and 4 respectively list the important anilino and anilinosulfonic acid representatives of bis(4,4′-triazinylamino)stilbene-2,2′-disulfonic acid. The latter can be employed over a wide pH range. All of the listed compounds are distinguished by high whitening effects, good efficiency, and adequate lightfastness, but they have only partial fastness to bleaches and none to chlorite solutions.

2.3. Stilbenyl-2*H*-triazoles

Stilbenyl-2*H*-triazoles are used as highly chlorine- and chlorite-stable brighteners for cotton and polyamides.

2.3.1. Stilbenyl-2*H*-naphtho[1,2-*d*]triazoles

The first important member of this group was 4,4′-bis(6-sulfonaphtho[1,2-*d*]triazol-2-yl)stilbene-2,2′-disulfonic acid. As the tetrasodium salt [*7426-67-7*]

(Blankophor G, 1950) [57], however, it produced a greenish-white brightening effect on cotton. Removal of one of the naphthotriazole groups from Blankophor G resulted in asymmetric compounds giving a neutral white, but at the cost of efficiency. The most important representative is the sodium salt of 4-(2*H*-naphtho[1,2-*d*]triazol-2-yl)stilbene-2-sulfonic acid(**32**) [*6416-68-8*] (Tinopal RBS, 1953) [58]. Both products are no longer marketed.

2.3.2. Bis(1,2,3-triazol-2-yl)stilbenes

Bis(1,2,3-triazol-2-yl)stilbene brighteners have high efficiencies and good lightfastness and are very resistant to chemical bleaches. They are employed as brighteners for cotton and polyamides. An example is 4,4′-bis(4-phenyl-1,2,3-triazol-2-yl)stilbene-2,2′-disulfonic acid dipotassium salt (**33**) [*52237-03-3*] (Blankophor BHC, 1970) [59].

Compound **34** is obtained by bis-hydrazone formation from hydroxyiminoacetophenone and 4,4′-dihydrazinostilbene-2,2′-disulfonic acid followed by cyclization with urea [60].

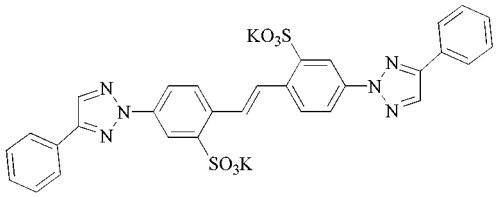

33 R = H, M = K [*52237-03-3*] [60]
34 R = SO_3Na, M = Na [*61968-72-7*] [61]

2.4. Benzoxazoles

The benzoxazole ring is especially suitable as an end group for lengthening conjugated systems that have the ability to resonate. The ring is easily synthesized and imparts good to excellent lightfastness, especially on polyester substrates. Brighteners are applied in the textile industry or added to the spinning mass. In addition, synthetic fibers such as polyamides and cellulose acetates, as well as plastics such as polyolefins, polystyrene, or poly(vinyl chloride) can be brightened with these compounds.

2.4.1. Stilbenylbenzoxazoles

Stilbenylbenzoxazoles are produced by cyclocondensation of stilbene-4-carboxylic acids with 2-aminophenols or by olefination with aromatic anils (anil synthesis); the latter requires activated tolyl compounds as reactants [62].

For example, 5,7-dimethyl-2-(4-tolyl)benzoxazole and the Schiff′s base from diphenyl-4-carbaldehyde and aniline react at 60 – 65 °C to yield 5,7-dimethyl-2-(4′-phenylstilben-4-yl)benzoxazole (**35**) [*40704-04-9*] [63] (see top of page 371).

Table 3. Anilino derivatives of bis(4,4′-triazinylamino)stilbene-2,2′-disulfonic acid

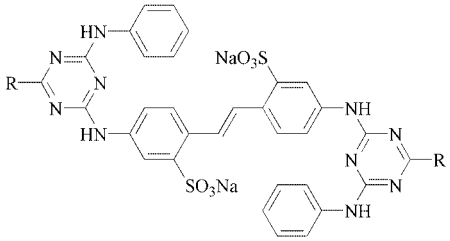

FWA	R	CAS no.	Ref.
14	OCH_3	[*3426-43-5*]	[40]
15	$NH–CH_3$	[*35632-99-6*]	[41–43]
16	$NH–C_2H_5$	[*24565-13-7*]	[42]
17	$NH–CH_2CH_2OH$	[*12224-16-7*]	[44]
18	$-N(CH_3)CH_2CH_2OH$	[*13863-31-5*]	[45]
19	$N(CH_2CH_2OH)_2$	[*4193-55-9*]	[25,46]
20		[*16090-02-1*]	[41]
21	$NH–C_6H_5$	[*133-64-4*]	[25,46]
22	$-N(CH_2CH_2OH)CH_2CH_2CONH_2$	[*68444-86-0*]	[47]

Table 4. Anilinosulfonic acid derivatives of bis(4,4′-triazinylamino)stilbene-2,2′-disulfonic acid

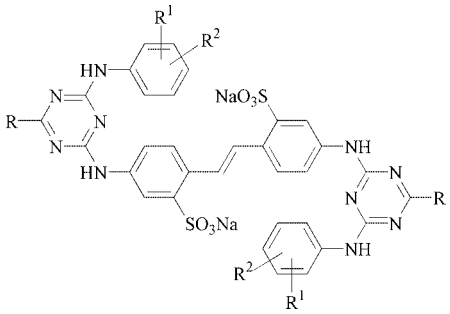

FWA	R	R^1 (position)	R^2 (position)	CAS no.	Ref.
23	$NH–CH_2CH_2OH$	SO_3Na (3)	H	[*61968-74-9*]	[48]
24	$N(CH_2CH_2OH)_2$	SO_3Na (3)	H	[*12224-02-1*]	[49,50]
25	$-NH(CH_2CH(OH)CH_3)_2$	SO_3Na (4)	H	[*99549-42-5*]	[51]
26	$N(CH_2CH_2OH)_2$	SO_3Na (4)	H	[*16470-24-9*]	[50,52]
27	$-N(CH_3)CH_2CH_2OH$	SO_3Na (4)	H	[*74228-28-7*]	[53]
28	$N(C_2H_5)_2$	SO_3Na (2)	SO_3Na (5)	[*83512-97-4*]	[54,55]
29	$N(CH_2CH_2OH)_2$	SO_3Na (2)	SO_3Na (5)	[*76482-78-5*]	[55]
30		SO_3Na (2)	SO_3Na (5)	[*55585-28-9*]	[55]
31	$N(CH_2CO_2Na)_2$	SO_3Na (4)	H	[*17405-36-3*]	[56]

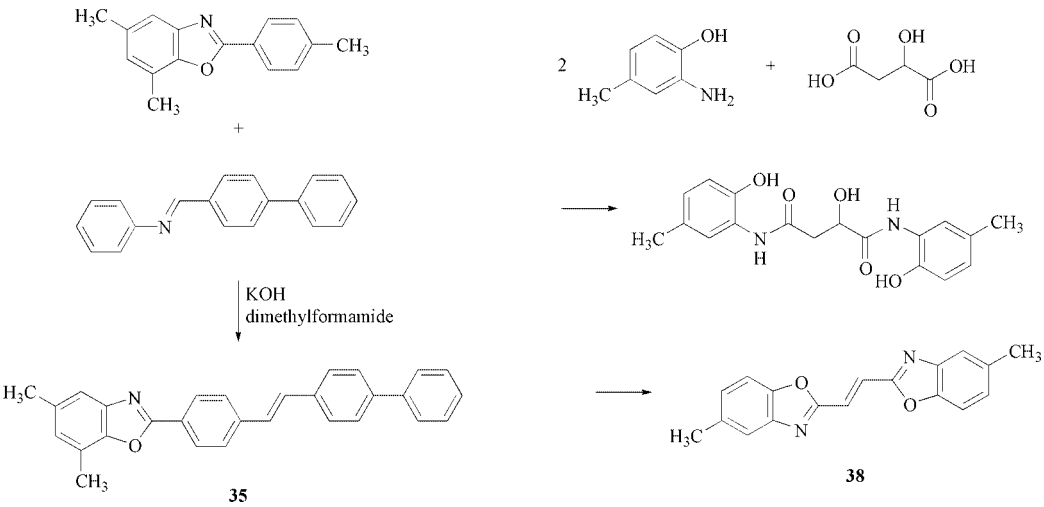

Table 5. Stilbenylbenzoxazoles

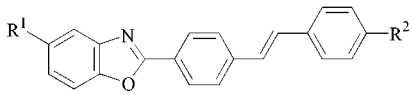

FWA	R^1	R^2	CAS no.	Ref.
36	CH_3	$COOCH_3$	[*18039-18-4*]	[64,65]
37	H	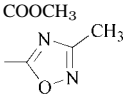	[*64893-28-3*]	[66]

Other stilbenylbenzoxazoles are given in Table 5. Mixtures of **35**–**37** with polyester brighteners of different structure [e.g., **4**, **6**, and **45**] are commercially available.

2.4.2. Bis(benzoxazoles)

Table 6 lists the most important bis(benzoxazole) brighteners.

In addition to the most common synthesis (the reaction of dicarboxylic acids or their chlorides with 2-aminophenols) other industrially interesting processes have been developed.

Compound **38** (Uvitex ERN, 1956) is obtained, for example, by condensing 2-amino-4-methylphenol with maleic acid to yield the diamide. Ring closure and the formation of the ethylene bridge occur on heating in a mixture of diphenyl ether and diphenyl at 135 – 155 °C in the presence of boric acid [68] (see above).

To obtain 2,5-bis(benzoxazol-2-yl)thiophene (**39**) (Uvitex EBF, 1961), two equivalents of 2-chloromethylbenzoxazole are treated with sodium sulfide, and the thiophene ring is formed with glyoxal hydrate [70]:

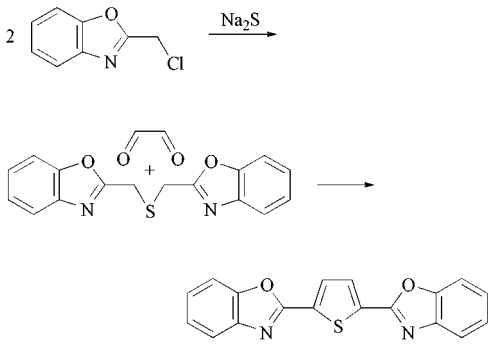

39

For **43** (Eastman OB 1, 1962), a new olefination reaction was discovered. 2-[4-(Chloromethyl)phenyl]benzoxazole is reacted in polar aprotic solvents with strong bases [76]:

Table 6. Bis(benzoxazoles)

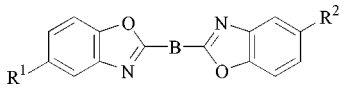

FWA	R^1	R^2	B	CAS no.	Ref.
38	CH_3	CH_3		[*1041-00-5*]	[67,68]
39	H	H	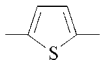	[*2866-43-5*]	[69,70]
40	$C(CH_3)_3$	$C(CH_3)_3$	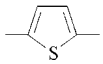	[*7128-64-5*]	[71]
41	H	H	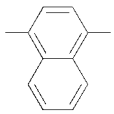	[*5089-22-5*]	[72]
42	CH_3	CH_3	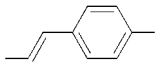	[*1552-46-1*]	[73]
43	H	H	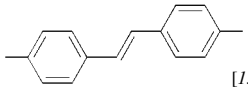	[*1533-45-5*]	[74,75]
44	CH_3	H	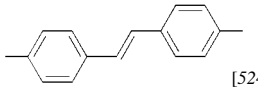	[*5242-49-9*]	[76]

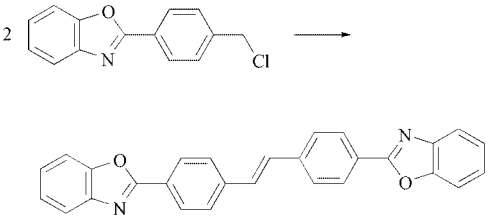

43

2.5. Furans, Benzo[*b*]furans, and Benzimidazoles

Furans and benzo[*b*]furans are further building blocks for optical brighteners. They are used, for example, in combination with benzimidazoles and benzo[*b*]furans as biphenyl end groups.

2.5.1. Bis(benzo[*b*]furan-2-yl)biphenyls

The 4,4′-bis(benzo[*b*]furan-2-yl)biphenyls can be regarded as distyrylbiphenyls that have had their rings closed and are thus fixed in the *E* configuration. Sulfonated derivatives are described as as lightfast brighteners for polyamides and cellulosic fibers [77].

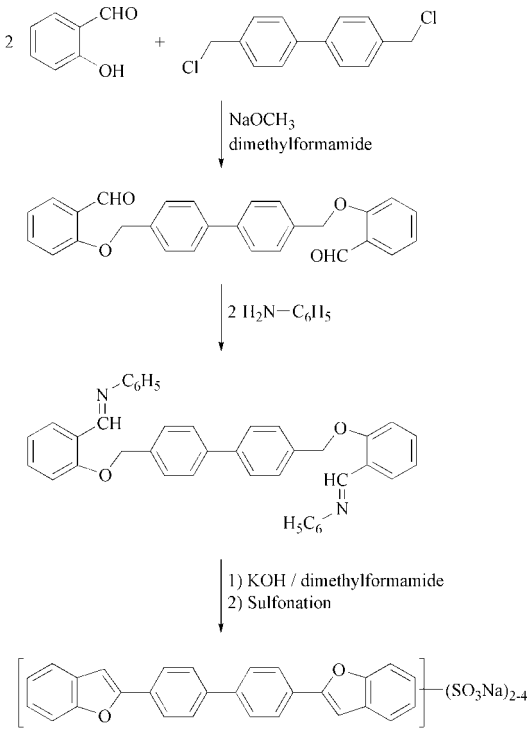

45

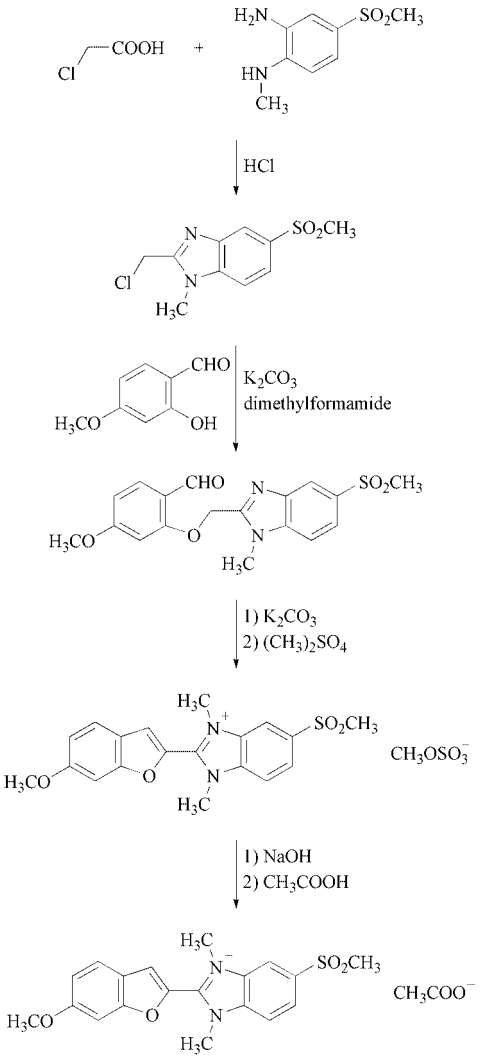

47

Compound **45** is produced by first treating 4,4′-bis(chloromethyl)biphenyl with salicylaldehyde to give 4,4′-biphenyldiyl-bis(methylenoxy-2-benzaldehyde). The formyl groups of this compound are converted to phenyliminomethyl groups with aniline. The double ring closure in dimethylformamide with potassium hydroxide is almost quantitative [78, 79]. The final step is sulfonation. For other syntheses, see [80].

2.5.2. Cationic Benzimidazoles

2,5-Di(1-methylbenzimidazol-2-yl)furan [*4751-43-3*] (Uvitex AT, 1967) [81], which has the ability to form salts, was one of the first cationic benzimidazole brighteners.

More recently, cationic benzimidazoles have been developed as brilliant, chlorine-fast, light-fast brighteners for polyacrylonitrile and cellulose acetate (Table 7). They are marketed in the form of cold- and heat-resistant concentrated aqueous solutions with long shelf lives.

In the synthesis of **47**, 4-methoxysalicylaldehyde is condensed with 2-chloromethyl-1-methyl-5-methylsulfonylbenzimidazole, ring closure is effected, and the product is quaternized with dimethyl sulfate [83]. To enhance water solubility, the methyl sulfate anion is converted to the acetate anion; the benzimidazole ring is first opened with sodium hydroxide and then reclosed with acetic acid because direct ion exchange is unfavorable for solubility reasons [84].

Compound **47** can also be synthesized from 6-methoxybenzo[*b*]furan-2-carboxylic acid and 2-methylamino-5-methylsulfonylaniline by cyclocondensation [83].

2.6. 1,3-Diphenyl-2-pyrazolines

Although the intense blue fluorescence of 1,3-diphenyl-2-pyrazolines had long been known [86], the usefulness of these compounds as optical brighteners was not recognized until 1949 [87]. The most valuable brighteners are the 3-(4-chlorophenyl)-1-(4-phenylsulfonic acid) derivatives because of their ready manufacturing ability, excellent optical properties, and versatility in application.

Industrial synthesis involves the reaction of β-chloro-4-chloropropiophenones with phenylhydrazines containing substituents at the 4-position, such as 4-hydrazinobenzenesulfonamide. The β-chloro-4-chloropropiophenone needed as starting material can be obtained from the Friedel–Crafts reaction of chlorobenzene with 2-chloropropionyl chloride.

1-(4-Amidosulfonylphenyl)-3-(4-chlorophenyl)-2-pyrazoline (**49**) [*2744-49-2*] (Blankophor DCB, 1961) [88] is used for the brightening of polyamides in the laundry.

Table 7. Cationic benzimidazoles

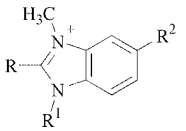

A⁻ (anion)

FWA	R	R^1	R^2	A^-	CAS no.	Ref.
46	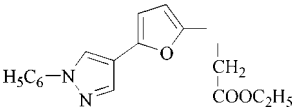		H	Br^-/Cl^-	[*72829-17-5*][*74878-56-1*]	[82]
47	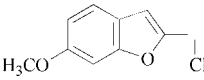		SO_2CH_3	CH_3COO^-	[*74878-48-1*]	[83,84]
48	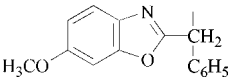		H	$CH_3OSO_3^-$	[*66371-25-3*]	[85]

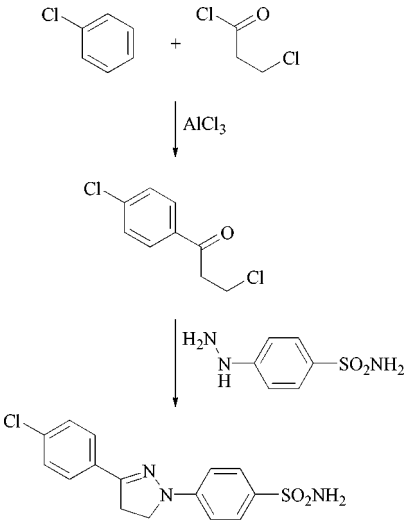

49

The compounds listed in Table 8 and **55** are industrially important nonionic and anionic 1,3-diphenyl-2-pyrazoline derivatives. The anionic compounds are employed to brighten polyamides but do not, however, always have adequate lightfastness. The poor wet lightfastness can be improved by the insertion of substituents in the 3-phenyl and pyrazoline rings, as in **54** and **55** [*81209-71-4*] [94]. The pyrazoline ring system is not stable against oxidative bleaches, which dehydrogenate it to the corresponding pyrazole.

The tertiary and quaternary amine salts of 1,3-diphenyl-2-pyrazoline derivatives (Table 9) are brilliant, moderately lightfast brighteners for (modified) polyacrylonitriles. Some can also be used to brighten cellulose acetate. Depending on the structure and nature of the anion, concentrated aqueous solutions with long shelf lives can be obtained; such products are needed, for example, in gel brightening.

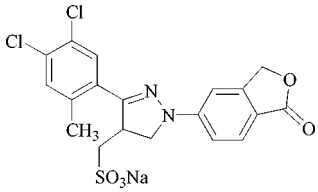

55

The side chains R in Table 8 and $-SO_2R$ in Table 9 can be added before or after cyclocondensation, for example, by the attachment of $NaHSO_3$, amines, or alcohols to vinyl or allylsulfonyl groups.

2.7. Coumarins

Further modification of the coumarin ring system in esculin (**1**) led to the development of 7-amino-4-methylcoumarins with an *N*-alkyl group at the 7-position. The sole representative of this series still on the market is 7-diethylamino-4-methylcoumarin (**63**) [*91-44-1*] (Uvitex WGS, 1954) [101]. Despite its poor lightfastness, the good optical properties of this compound have kept it in use for the brightening of wool, cellulose acetate rayon, and polyamides. Compound **63** is produced by cy-

Table 8. Nonionic and anionic 1,3-diphenyl-2-pyrazolines

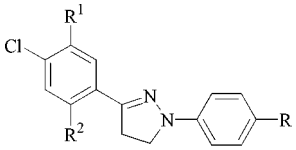

FWA	R^1	R^2	R	CAS no.	Ref.
50	H	H	SO_2CH_3	[*60650-443*]	[89]
51	H	H	$SO_2CH_2CH_2OH$	[*3656-22-2*]	[90]
52	H	H	$SO_2CH_2CH_2SO_3Na$	[*27441–70-9*]	[91]
53	H	H	COONa	[*32020-25-0*]	[92]
54	Cl	CH_3	$SO_2CH_2CH_2SO_3Na$	[*61931-42-8*]	[93]

Table 9. Cationic 1,3-diphenyl-2-pyrazolines

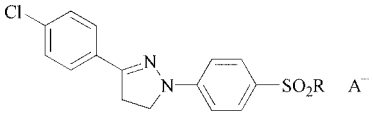

FWA	R	A^-	CAS no.	Ref.
56	$-(CH_2)_2-NH(CH_3)_2$	$H-P(=O)(OH)-O$	[*106359-93-7*]	[95]
57	$-(CH_2)_2-NH(CH_3)_2$	$HCOO^-$		[95]
58	$-CH_2-CH(CH_3)-\overset{+}{N}H(CH_3)_2$	$CH_3CH(OH)COO$	[*85154-08-1*]	[96]
59	$-(CH_2)_2-CONH-(CH_2)_3-\overset{+}{N}H(CH_3)_2$	Cl^-	[*42952-22-7*]	[97]
60	$-(CH_2)_2-O-CH(CH_3)-CH_2-\overset{+}{N}H(CH_3)_2$	Cl^-	[*63310-12-3*]	[98]
61	$-NH-(CH_2)_3-\overset{+}{N}(CH_3)_3$	$CH_3OSO_3^-$	[*12270-54-1*]	[99]
62	$-NH-(CH_2)_3-\overset{+}{N}(CH_3)_2-CH_2CH_2OH$	COO^-	[*36086-26-7*]	[100]

clocondensation of 3-diethylaminophenol with ethyl acetoacetate in the presence of zinc chloride [102].

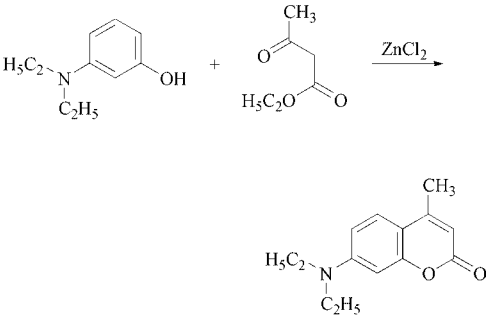

63

Substituted 3-phenylcoumarins with methoxy [103] and triazinylamino groups [104] at the 7-position were developed later. These have been superseded in industry by the 3,7-bis(hetero)arylcoumarins [105–110].

2.8. Naphthalimides

The 4-aminonaphthalimides and their *N*-alkylated derivatives are brilliant greenish yellow fluorescent dyes (→ Fluorescent Dyes, Chap. 2.). Acylation of the amino group at the 4-position of the naphthalimide ring shifts the fluorescence toward blue [111], yielding compounds suitable for use as optical brighteners, such as 4-acetylamino-*N*-(*n*-butyl)naphthalimide [*3353-99-9*]

(Ultraphor APL, 1954) [112].

Later workers, chiefly in Japan, used naphthalimides with alkoxy substituents at the 4- or 4,5-positions and obtained brighteners with good lightfastness for polyester substrates and good chlorite fastness for polyacrylonitriles. The first commercial product was 4-methoxy-*N*-methylnaphthalimide (**64**) [*3271-05-4*]

(Mikawhite AT, 1962) [113]. Table 10 lists the most important compounds.

The naphthalimide ring system has the drawback of low effectiveness, which is mainly attributable to the low molar extinction coefficient.

The industrial synthesis of alkoxynaphthalimides begins with acenaphthene. Chlorination and subsequent oxidation with dichromate give the corresponding naphthalic acids [116], which are converted to the anhydrides on drying.

Mild reaction with methylamine, followed by reaction with sodium methoxide or sodium ethoxide, gives, e.g., **64** or **65**.

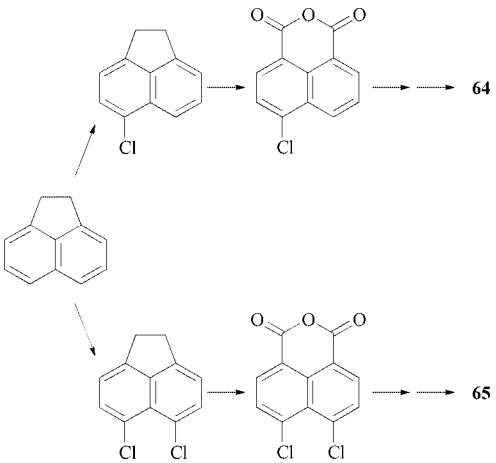

2.9. 1,3,5-Triazin-2-yl Derivatives

The sole representative of this class is 1-(4,6-dimethoxy-1,3,5-triazin-2-yl)pyrene (**67**) [*3271-22-5*]

(Fluolite XMF, 1963), which gives a bluish white brightening effect on polyester substrates; it is employed chiefly in combination with other polyester brighteners such as **38**. Compound **79** is obtained by the Friedel – Crafts reaction of cyanuric chloride with pyrene with subsequent replacement of the remaining chloro substituents with methoxy groups [117].

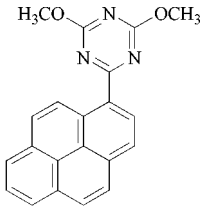

67

3. Commercial Forms and Brands

Commercial Forms and Formulations. Optical brighteners are formulated specifically according to structure, substrate, and mode of application. Nondusting, easily meterable forms are preferred: free-flowing granules with high crushing strength, solutions, or slurries (concentrated dispersions).

Nonionogenic optical brighteners must be as finely divided as possible so that they can be quickly and optimally distributed in apolar substrates. They are marketed as powders (e.g., for plastics or spinning masses) or as low-viscosity dispersions (e.g., for the textile industry).

Formulations are optimized with additives to control granule strength, dissolution, chemical stability, viscosity, stability on storage, or fungal growth. They can also contain toning dyes and auxiliaries to improve their efficiency in a given application.

Commercial Brands. Several thousand products with more than 100 trade names are on the market worldwide. Only a few of the most important are listed here:

Blankophor (Bayer), Hakkol (Showa Kayaku), Hostalux (Clariant), Kayaphor (Nippon Kayaku), Leukophor (Clariant), Optiblanc (3V), Photine (Hickson & Welch), Tinopal (Ciba Specialty Chemicals), Ultraphor (BASF), Uvitex (Ciba Specialty Chemicals), Whitex (Sumitomo), Belofor, Calcofluor White, Eastobright, Enkantine, Heliophor, Mikephor, Ranipal, Rylux, Viophos, and Wobital.

Table 10. Naphthalimides

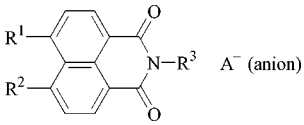

FWA	R^1	R^2	R^3	CAS no.	Ref.
64	OCH_3	H	CH_3	[*3271-05-4*]	[113]
65	OC_2H_5	OC_2H_5	CH_3	[*22330-48-9*]	[114]
66	OCH_3	H	$O(CH_2)_3CH_3$	[*258-31-7*]	[115]

4. Uses

4.1. General Requirements

Optical brighteners (fluorescent whitening agents, FWAs) are used primarily in the textile, detergent, and paper industries and are also added to plastics [118]. The mode of application and performance requirements differ greatly from one substrate to another, but there are five common basic principles:

1) FWAs can only improve whiteness effectively if the substrate does not absorb too strongly in their excitation and fluorescence range. Objects with a poor "basis white", such as unbleached textiles or poorly cleaned articles, cannot be brightened to a high degree of whiteness even with high concentrations of FWAs [119].
2) FWAs only act effectively if they are monomolecularly dispersed in the substrate (see Section 1.1). This depends on the substrate, the FWA type, and the mode of application [4, 5, 120].
3) Whitening effects increase with FWA concentration up to a saturation limit. For this reason, the applied concentrations of FWA must be kept low: ca. 0.002 – 0.2 wt % relative to the substrate, or roughly one tenth of the usual concentration of dyes [4].
4) Since the use of FWAs in industrial applications should not create any problems, specific products have been developed for each type of application.
5) The required shades and fastnesses needed are also selection criteria for a suitable FWA.

4.2. Textile Industry

Brighteners can be applied to any substrate by batch or continuous processes [9, 17, 121–125]. The processes are the same as those used for dyeing (→ Textile Dyeing).

Cellulosic Fibers. Substantivity for cellulosic fibers is offered mainly by anionic FWAs, preferably with sulfonic acid groups. Brighteners such as **14**, **19**, **24**, and **26** with moderate to high affinities are suitable for batch processes. If simultaneous bleaching takes place in the same bath, the brightener must be stable with respect to bleaching chemicals such as peroxide or chlorite [126].

Continuous processes allow high throughputs of woven and knitted fabrics [127]. To achieve a level brightening effect from the beginning to the end of the process and between the center and the edges of the material, FWAs with low to moderate affinity for the fiber are used; examples of principal types are **23**, **25**, **30**, **31**, and **34**.

Brightening is often combined with a high-grade finishing process. The goods are impregnated with a synthetic resin in the padding machine, usually under strongly acidic conditions (Lewis acids can also be employed as catalysts). Brighteners must therefore meet additional stability requirements [128]; types such as **10** and **33** are especially suitable for such a process.

Polyamides. Batch brightening usually takes place in combination with reductive bleaching (dithionite). Continuous processes (padding) are followed by steam, heat, or acid treatment (acid shock) [129]. Anionic FWAs are absorbed similarly to acid dyes, or insoluble dispersion brighteners are suitable. Examples of products listed in order of increasing fastness are diaminostilbenes (**14**), (**19**); pyrazolines (**50**), (**54**); distyrylbiphenyls (**10**); and bis(triazolyl)stilbenes (**33**).

Polyacrylonitriles. An increasing proportion of the polyacrylonitrile fibers produced have a good basis white and do not require further bleaching [130, 131]. The use of FWAs that can be applied in the chlorite bleaching bath is therefore steadily declining. Suitable types include benzimidazoles such as **46**, **47**, and **48**.

If chlorite stability is not required, other less expensive, highly effective products can be used: pyrazoline derivatives such as **49** and **50** and cationic products such as **58–60** and **62**. The optical brighteners are being increasingly applied immediately after the polyacrylonitrile leaves the spinning nozzle while it is still in the gel state. Gel application has already become the dominant process in some countries (e.g., UK). Increasingly stringent fastness requirements relating to, e.g., perborate washing and exposure to light are satisfied by types **58** and **60**.

Polyesters. Dispersions of nonionogenic FWAs are suitable for hydrophobic fibers such as polyesters. Penetration into the fibers and optimal distribution are achieved by batchwise treatment at 110 – 130 °C under pressure, or at the boiling point with a "carrier" (a substance that improves diffusion of the brightener by swelling the fiber); the continuous process in the padding machine involves subsequent heat treatment at 180 – 210 °C (simultaneously resulting in good dimensional stability of the textile) [132]. The FWA molecules become oriented during thermal fixation [133]. Batch processes can be combined with bleaching, which is necessary for many polyester qualities and polyester – cellulosic mixtures.

The brighteners should be adapted to the heat treatment; they should represent an optimal compromise between diffusion into the polyester material and sublimation fastness [32, 33]. High to very high lightfastness and chlorite stability are offered by naphthalimides such as **64** and **65**; and benzoxazoles such as **35-38**, **41**, **42**, and especially **39**. Compounds such as **4 – 6**, **13**, and **67** are very effective but less stable to chlorite and light. " Boosters" (e.g., alkyl esters of stilbene dicarboxylic acids) have been suggested as a way of intensifying the action of FWAs [134].

Brighteners that sublime can be applied by transfer printing [135, 136].

Other Fibers. The main products used for cellulose acetate or triacetate fibers are pyrazoline derivatives such as **49–51**, and for wool typical polyamide brighteners.

4.3. Detergent Industry

Regular replacement of brighteners degraded during use improves the acceptance of textile articles and thus helps extend their service lives [137, 138]. Optical brighteners are therefore used in laundry detergents (→ Laundry Detergents, Chap. 3.4.5.) and are required above all for cotton fabrics [139].

Heavy-Duty Powders. Appropriate brighteners for regions with lower washing temperatures include **17** and especially **10** with its better lightfastness. FWAs such as **18** and especially **20** are used in countries where typical washing temperatures are moderate to high [140]

The preferred FWAs lend a bright, fresh, clean appearance to the detergent powder. Bleaching systems in modern detergents are effective even at low temperature, so that stringent demands are placed on FWA stability [141]. For example, triazinylaminostilbenes must be as free as possible from nonfluorescing triazine byproducts, which lead to discoloration and an unpleasant odor during storage of the detergent powder [142]. The active ingredient may also be degraded; preference is therefore given to **10**.

Heavy-Duty Liquid Detergents. The ease of incorporation and affinity of FWAs are crucial in such formulations. Heavy-duty liquid detergents can also be employed as "prespotters" that are applied in concentrated form to heavily soiled areas before the normal washing process. Bright spots can be formed, especially on colored fabrics, due to washing out of the color; they are intensified by the fluorescence of the locally high FWA concentration. For this reason, FWAs have been developed which cause less pronounced spotting that cannot be seen. The main products in use are **10** and **21** [143].

The addition of bleaches to aqueous or nonaqueous heavy-duty liquid detergents will impose even more stringent stability requirements on FWAs [144–146].

Other Laundry Products. Laundry "boosters" are bleaching systems that are added separately to the wash (→ Laundry Detergents, Chap. 4.3.2.). They contain **10** as a brightener; this compound is easy to incorporate and shows good compatibility and stability during storage [147].

Light-duty detergents contain coumarins (**63**) as brighteners for polyamide and wool. At the washing temperatures acceptable for polyesters, the FWAs cannot penetrate into the fibers. Special products containing water-soluble polymers and a lightfast brightener such as **10** have therefore been developed for use in the rinse bath after the washing of, for example, curtains [148].

4.4. Paper Industry

White paper must satisfy increasingly high standards of whiteness, which can be obtained by means of the quality and degree of bleaching of the raw materials as well as optical brightening. Brighteners can be added at three stages of papermaking: in the pulp, to the paper surface, and during coating [2, 149]. Application in the pulp has leveled off in recent years, while brightening of the paper surface and especially coating have made strong gains.

Brightening in Pulp. The brightener is added in an aqueous solution to the paper pulp during preparation in the hollander or in mixing chests. Inexpensive, highly soluble triazinylaminostilbenes such as **19**, **22**, **24**, **26**, and **30** are used. For acid-sized paper qualities, good stability toward acid and alum are necessary. There is, however, a trend toward neutral sizing.

Impregnation of Paper Web. The paper web is impregnated from one or both sides with an FWA solution that often contains other additives. In the size press, the paper is therefore brightened only on the surface, thus economizing on brightener material. Examples of products employed in this way are **24**, **26**, and **28**–**31**.

Addition to Coating. The paper is often coated with a material that contains binder and/or pigment, which usually reduces the whiteness [150]. To make up for this loss of whiteness or to provide additional brightness, FWAs such as **15**, **19**, **22**, **23**, **28**, and **29** are added to the coating.

4.5. Plastics and Synthetic Fibers

Brighteners for plastics and synthetic fibers must be "dissolved" in the material. They are added dry to the polymer powder, granulate, or chips, either before or during polymerization or else later before high-temperature forming (e.g., immediately prior to spinning in the case of synthetic fibers). Very good chemical and thermal stabilities are prerequisites for such FWAs. The brightener can also be added to the polymer solution if this is appropriate [132, 151, 152]. Low FWA concentrations are often used (e.g., 0.005 wt %) just enough to offset the yellowish cast of the material. In poly(vinyl chloride) articles, higher degrees of whiteness are achieved mainly with compounds such as **12**, **44**, and especially **40**; in polyester fibers, with products such as **43**, and especially **44**.

4.6. Other Uses

Optical brighteners are also used to enhance the whiteness of various inorganic and organic substrates, such as tripolyphosphates, carbonates, and photographic papers [152–154].

Markings visible under UV light can be made with FWAs, for example, to identify petroleum products such as fuel oil or to verify the uniform distribution of sprayed products.

Compounds such as **40** are valuable as scintillators that convert the radiation emitted by radioactive isotopes into visible light which can easily be measured [155]. Substances such as **26** are preferentially absorbed by fungi, making them easily detectable in the skin or other types of cells allowing accurate clinical diagnosis and control [156].

Some FWAs, especially **10**, are of interest as laser dyes [5, 157].

5. Analytical Methods and Whiteness Assessment

5.1. Analytical Methods

Identification. Chromatographic methods [interpretation of fluorescing spots in thin layer chromatography or of UV absorption peaks in (high performance) liquid chromatography] allow fairly rapid and certain identification of FWAs [158–166]. The same methods can be used for qualitative and quantitative characterization of the FWA purity in terms of fluorescent or nonfluorescent byproducts.

Determination in Solutions. The content of active substance is determined by photometric measurement of the extinction (absorbance) in solution (E_1^1) [158]. Both the wavelength and the height of the absorption maximum depend on the solvent. Concentrations of FWAs in application baths can be monitored quantitatively by comparative fluorescence measurements against calibrated standards [167].

When dilute FWA solutions are handled, their photosensitivity must be kept in mind. Lighting that includes UV wavelengths can lead to rearrangement (e.g., $E \rightarrow Z$) and/or degradation of the FWAs to give molecules with different spectra and chromatographic properties [18, 170].

Determination in Substrates. The FWA on a substrate can be directly determined by reflectance measurements. Alternatively, FWAs may be extracted with appropriate solvents and determined in solution; in the case of plastics, this involves dissolving the substrate [168, 169, 171, 172].

Determination in the Aquatic Environment. Identification and quantification of FWAs in sewage water and sludges as well as in surface waters and sediments require special methods of extraction and sensitive determination of photoisomers [173, 174].

5.2. Assessment of Whitening Effect

Whitening can be assessed either visually or instrumentally (by colorimetry). The quantities evaluated are the *degree of whiteness* and the *tint*. Special attention must be paid to the spectral distribution of the incident light in the UV and visible ranges, since this controls the fluorescent contribution of the brightener to the whiteness [175].

Visual assessment is performed by comparison with references in diffuse daylight [176]. White scales consisting of stages with well-defined whiteness values (e.g., the Ciba-Geigy cotton or plastic white scale) can be used [177].

For instrumental evaluation spectrophotometers are preferred; three-filter instruments can, however, also be employed. The colorimetric values obtained depend on the instrument and its instantaneous state (chiefly the sample illumination conditions) and must be controlled by suitable methods [178]. The influence of the instrument can be eliminated by conversion to the standard illumination D 65, representing daylight with a correlated color temperature of 6504 K [179]. The degree of whiteness and the tint values can then be calculated from these colorimetric data with appropriate formulas. A selection of currently used whiteness formulas can be found in [180] and [181]. For attempts at standardizing the assessment methods for white objects, see [9, 182–189].

6. Environmental Aspects

Both the toxicology and the ecotoxicology of commercial optical brighteners have been extensively studied [7, 19, 190, 191].

Acute toxicity by the oral, dermal, and inhalatory routes place these substances outside any poison class. The oral median lethal doses (LD_{50}) are generally above 2000 mg/kg. There are no indications of adverse effects as regards irritation, sensitization, or subacute toxicity. No mutagenic or teratogenic effects have been reported for the many optical brighteners so far tested. Phototoxicity and photocarcinogenicity studies on five FWAs as well as chronic rat studies with four optical brighteners produced a favorable no observable adverse effect level (NOAEL) [7]. FWAs for cotton/paper do not penetrate the skin in measurable amounts, and when fed to animals they are eliminated unchanged. Extensive work has been done to characterize aquatic toxicity and photo- and bio-

degradation. Human and environmental risk assessments of FWAs for cotton/paper show no adverse effects to human and environment [190, 191].

As industrial chemicals, fluorescent whitening agents are regulated by national laws such as the Toxic Substance Control Act in the United States and the relevant classification, packaging, and labeling directives in the European Community. For specific applications such as food packaging, the national regulatory requirements must be complied with.

7. References

General References

1. *Ullmann,* 5th ed., **A18,** 153 – 176.
2. A. E. Siegrist: "Die Anwendung optischer Aufheller in der Papierindustrie", *Papier (Darmstadt)* **8** (1954) 109 – 120.
3. R. Zweidler: "Einführung in die Chemie der optischen Aufheller", *Textilveredlung* **4** (1969) 75 – 87.
4. T. Förster: *Fluoreszenz organischer Verbindungen,* Vandenhoek und Ruprecht, Göttingen 1951.
5. F. P. Schäfer: *Dye Lasers,* Springer Verlag, Berlin 1973.
6. H. Gold: "Fluorescent Brightening Agents" in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. V, Academic Press, New York – London 1971, pp. 535 – 679.
7. R. Anliker, G. Müller: "Fluorescent Whitening Agents" in F. Coulston, F. Korte (eds.): *Environmental Quality and Safety,* vol. IV, Thieme Verlag, Stuttgart 1975.
8. M. Zahradnik: *The Production and Application of Fluorescent Brightening Agents,* John Wiley & Sons, Chichester 1982.
9. R. Levene, M. Lewin: "The Fluorescent Whitening of Textiles" in M. Lewin, S. B. Sello (eds.): *Handbook of Fiber Science and Technology,* vol. 1, part B, Marcel Dekker, New York – Basel, 1984, pp. 257 – 304.
10. A. K. Sarkar: "Fluorescent Whitening Agents" in J. G. Cook (ed.): *Textile Technology,* Merrow Publishing Co., Watford, England, 1971.
11. D. Barton, H. Davidson: "Fluorescent Brighteners, 1967 – 1973", *Rev. Prog. Color. Relat. Top.* **5** (1974) 3 – 11.
12. A. Dorlars, C. W. Schellhammer, J. Schroeder: "Heterocyclen als Bausteine neuer optischer Aufheller", *Angew. Chem.* **87** (1975) 693 – 707.
13. E. Siegel: "Organic Dyes and Optical Brightening Agents", *Int. Rev. Sci. Org. Chem. Ser. Two,* 1976 vol. 2, 259 – 297.
14. E. A. Kleinheidt, H. Theidel: "Optische Aufheller" in A. Chwala, W. Anger, C. Chwala (eds.): *Handbuch Textilhilfsmittel,* Verlag Chemie, Weinheim 1977, pp. 645 – 666.
15. *Kirk-Othmer,* 3rd. ed., **4,** pp. 213 – 226.
16. Y. Hashida, K. Matsui: "Recent Development in Fluorescent Whitening Agents", *Senryo to Yakuhin* **25** (1980) 128 – 137.
17. R. Williamson: "Fluorescent Brightening Agents" in *Textile Science and Technology,* vol. 4, Elsevier, Amsterdam, 1980.
18. I. H. Leaver, B. Milligan: "Fluorescent Whitening Agents, A Survey (1974 – 1982)", *Dyes Pigm.* **5** (1984) 109 – 144.
19. A. E. Siegrist, H. Hefti, H. R. Meyer, E. Schmidt: "Fluorescent Whitening Agents 1973 – 1985", *Rev. Prog. Color. Relat. Top.* **17** (1987) 39 – 55.
20. H. Zollinger: "Fluorescent Brighteners" in *Color Chemistry,* VCH Verlagsges., Weinheim 1987, pp. 203 – 213.
21. B. M. Krasovitskii, B. M. Bolotin: *Organic Luminescent Materials,* VCH Verlagsges., Weinheim 1988.

Specific References

22. P. Krais: "Ueber ein neues Schwarz und ein neues Weiss", *Melliand Textilber.* **10** (1929) 468 – 469.
23. R. Anliker in [7]pp. 12 – 18.
24. Ultrazell, GB 472 473, 1935 (L. Mellersch – Jackson).
25. IG Farbenind., DE 752 677, 1940 (B. Wendt).
26. IG Farbenind., DE 746 569, 1940 (S. Petersen, O. Bayer, B. Wendt).
27. Hoechst, GB 981 178, 1958 (L. Horner, H. Hoffmann, H. G. Wippel); US 3 156 739, 1960 (L. Horner, H. Hoffmann, W. Klink). L. Horner, H. Hoffmann, H. G. Wippel, G. Klahre, *Chem. Ber.* **92** (1959) 2499.
28. BASF, GB 920 988, 1959 (W. Stilz, H. Pommer, K. H. Koenig); US 3 177 208, 1959 (W. Stilz, H. Pommer).
29. Hoechst, US 4 336 155, 1979 (T. Martini et al.).
30. Ciba-Geigy, US 4 778 623, 1979 (L. Guglielmetti).
31. BASF, US 4 380 514, 4 464 284, 1980 (G. Seybold).

32. Ciba-Geigy, EP-A 238 446, 1986; GB 2 200 660, 1987 (L. Guglielmetti, H. R. Meyer, D. Reinehr, K. Weber).
33. T. Martini, H. Probst: "Synergismen bei optischen Aufhellern", *Melliand Textilber.* **65** (1984) 327 – 329.
34. Ciba-Geigy, US 3 984 399, 1967 (K. Weber, P. Liechti, H. R. Meyer, A. E. Siegrist).
35. Ciba-Geigy, GB 1 360 142, 1971 (K. Weber, C. Lüthi).
36. Ciba-Geigy, US 4 339 393, 4 468 352, 4 602 087, 1979 (C. Lüthi, H. R. Meyer, K. Weber).
37. Sandoz, US 4 108 887, 4 196 229, 1975 (F. Fleck, J. Heller).
38. Ciba-Geigy, US 4 708 558, 1985 (K. Burdeska, K. Weber, D. Reinehr); US 4 533 505, 1981 (A. Spencer).
39. Ciba-Geigy, EP-A 305 646, 1987 (L. Guglielmetti).
40. Ciba, US 2 846 496, 1949 (F. Ackermann). General Aniline & Film Corp., US 2 713 046, 1951 (W. W. Wilson, H. B. Freyermuth).
41. ICI, US 2 612 501, 1947 (R. H. Wilson).
42. Ciba, US 2 763 650, 1951 (F. Ackermann).
43. Nippon, JP-Kokai 59 5 8 80, 1956 (T. Noguchi, M. Sumitami), *Chem. Abstr.* **54** (1960) 14707g.
44. Nisso, JP-Kokai 6 325 833, 1961 (G. Hayakawa, M. Orita, M. Hayashi, T. Obizu), *Chem. Abstr.* **60** (1964) 5524h.
45. Geigy, US 2 762 801, 1953 (H. Häusermann); GB 1 116 619, 1966 (H. Häusermann).
46. IG Farbenind., FR 874 939, 1940 (B. Wendt).
47. Sandoz, GB 1 243 276, 1968 (F. Fleck).
48. Nippon, JP-Kokai 6 3 21 636, 1961 (S. Ono, T. Noguchi).
49. Bayer, DE 1 044 758, 1956 (B. Dach); GB 841 189, 1957 (J. Hegemann, A. Mitrowski, H. Roos).
50. Bayer, GB 1 114 021, 1965 (E. A. Kleinheidt, H. Gold). Nippon, JP-Kokai 58 7 6 38, 1958 (S. Ono, T. Noguchi), *Chem. Abstr.* **54** (1960) 2381i.
51. Sandoz, GB 896 533, 1958 (F. Fleck).
52. Du Pont, US 3 025 242 , 1961 (R. C. Seyler); Amer. Cyanamid, US 3 211 665, 1962 (W. Allen, R. F. Gerard).
53. Geigy, GB 930 393, 1959 (H. Häusermann).
54. Wolfen, DD 55 668, 1966 (B. Noll, G. Chmell, G. Knöchel, L. Zölch).
55. Geigy, US 3 479 349, 1967 (R. C. Allison, F. Fischer, H. Häusermann).
56. Sandoz, WO 9 600 221, 1996(J. Cowmann, J. Farrar, J. Martin, M. Graham, N. Mackinnon).
57. Bayer, US 2 668 777, 1948 (H. Gold, S. Petersen).
58. Geigy, US 2 784 183, 1951 (E. Keller, R. Zweidler, H. Häusermann); Ciba-Geigy, US 4 791 211, 1982 (R. B. Lund, L. W. Bass).
59. Bayer, US 3 485 831, 1965 (A. Dorlars, O. Neuner, R. Pütter).
60. Bayer, US 3 666 758, 1967 (A. Dorlars, O. Neuner).
61. Bayer, GB 1 172 657, 1967 (A. Dorlars, O. Neuner).
62. A. E. Siegrist, *Helv. Chim. Acta* **50** (1967) 906 – 957. I. J. Fletcher, A. E. Siegrist, *Adv. Heterocycl. Chem.* **23** (1978) 171 – 261.
63. Ciba-Geigy, US 3 850 914, 1974 (C. Lüthi).
64. Nippon, JP-Kokai 7 6 40 090, 1968 (S. Mitsukuni, S. Kiyoshi), *Chem. Abstr.* **86** (1977) 122956n.
65. Hoechst, US 4 129 412, 1976 (D. Günther, E. Schinzel, R. Erckel, G. Rösch).
66. Hoechst, US 4 142 044, 1979 (D. Günther, R. Erckel, E. Schinzel, G. Rösch); US 4 231 741, 4 363 744, 1977 (D. Günther, R. Erckel, G. Rösch, H. Probst).
67. Ciba, US 2 483 428, 1945 (J. Meyer, F. Ackermann); US 2 488 289, 1945 (J. Meyer, C. Gränacher, F. Ackermann); US 2 875 089, 1955 (F. Ackermann, A. E. Siegrist). Intracolor, US 3 649 623, 1955 (F. Ackermann, M. Dünnenberger, A. E. Siegrist).
68. Ciba-Geigy, EP 31 296, 1979 (W. Schreiber).
69. Ciba, US 2 995 564, 1958 (M. Dünnenberger, A. E. Siegrist, E. Maeder); US 3 095 421, 1960 (P. Liechti, A. E. Siegrist, M. Dünnenberger, E. Maeder); US 3 127 416, 1960 (P. Liechti, A. E. Siegrist, E. Maeder).
70. Ciba-Geigy, GB 1 561 031, 1976 (L. Guglielmetti).
71. Ciba, US 3 135 762, 1961 (E. Maeder, P. Liechti, A. E. Siegrist, M. Dünnenberger).
72. Hoechst, US 3 709 896, 3 743 625, 1966 (H. Frischkorn, U. Pintschovius, H. Behrenbruch).
73. Nippon, US 3 629 246, 1962 (T. Tanaka).
74. Kodak, US 3 260 715, 1962 (D. G. Saunders); US 3 322 680, 1962 (D. G. Hedberg, M. S. Bloom, M. V. Otis).
75. Hoechst, US 4 585 875, 1983 (L. Heiss).
76. Nippon, JP-Kokai 7 2 36 861, 1963 (K. Ohkawa, T. Kinoshita, T. Tanaka), *Chem. Abstr.* **78** (1973) 85946w; JP-Kokai 6 6 20 225, 1964 (T. Tanaka), *Chem. Abstr.* **67** (1967) 55173x; JP-Kokai 6 9 06 978, 1964 (T. Tanaka), *Chem. Abstr.* **71** (1969) 4505g; JP-Kokai 6 9 23 028, 1966 (T. Tanaka), *Chem.*

Abstr. **72** (1970) 12705g. Kodak, US 3 586 673, 1968 (M. S. Bloom, J. A. Hill). Hoechst, US 4 282 355, 1978 (R. Erckel, P. Jürges).
77. Hoechst, US 3 559 350, 1971 (W. Sahm, E. Schinzel, G. Rösch). Nippon, JP-Kokai 7 5 40 627, 1973 (S. Mitsukuni, S. Hajime); *Chem. Abstr.* **83** (1975) 116988w.
78. Hoechst, US 3 864 333, 1971 (W. Sahm, A. Horn); US 3 892 807, 3 994 879, 1972 (W. Sahm).
79. W. Sahm, E. Schinzel, P. Juerges, *Justus Liebigs Ann. Chem.* 1974, 523 – 538.
80. Hoechst, US 4 133 953, 1977, US 4 309 536, 1979 (E. Schinzel).
81. Ciba, US 2 488 094, 1942 (C. Gränacher, F. Ackermann); US 3 103 518, 1963 (M. Dünnenberger, A. E. Siegrist).
82. Sandoz, US 4 018 789, 1974 (P. S. Littlewood, A. V. Mercer).
83. Ciba-Geigy, GB 1 331 670, 1970 (H. Schläpfer); EP 10 063, 1978 (L. Guglielmetti, C. Lüthi).
84. Ciba-Geigy, US 4 313 846, 4 455 436, 1978 (H. R. Meyer, M. Siegrist).
85. Sandoz, GB 1 577 127, 1976 (P. S. Littlewood).
86. L. Knorr, P. Duden, *Ber. Dtsch. Chem. Ges.* **26** (1893) 115. K. v. Auwers, H. Voss, *Ber. Dtsch. Chem. Ges.* **42** (1909) 4411 – 4427.
87. Ilford, US 2 639 990, 1949 (J. D. Kendall, G. F. Duffin).
88. Bayer, US 3 135 742, 1956 (A. Wagner, A. Schlachter, H. Marzolph).
89. Bayer, US 3 378 389, 1962 (C. W. Schellhammer, A. Wagner).
90. Hoechst, US 3 255 203, 1962 (E. Schinzel, K. H. Lebkücher).
91. Hoechst, US 3 619 234, 1968 (J. Weihsbach, O. Smerz, G. Rösch).
92. Bayer, US 2 879 174, 1955 (A. Wagner, S. Petersen).
93. Hoechst, GB 1 360 490, 1970 (H. Mengler, E. Schinzel, G. Rösch); US 3 865 816, 1971 (H. Mengler); DE-OS 2 702 635, 1977 (O. Smerz, E. Schinzel, G. Rösch, J. Weihsbach).
94. Sandoz, US 3 997 556, 1974 (I. J. Bolton, A. V. Mercer).
95. Ciba-Geigy, US 4 816 590, 1985 (H. R. Meyer).
96. Bayer, EP 73 996, 1983 (H. Theidel).
97. Sandoz, US 3 849 406, 1971 (H. Aebli, F. Fleck, P. S. Littlewood, A. V. Mercer).
98. Hoechst, US 3 560 485, 1964 (E. Schinzel, S. Bildstein, K. H. Lebkücher); US 3 690 947, 1969 (G. Rösch, W. Linke, O. Smerz, E. Schinzel).
99. Bayer, US 3 131 079, 1960 (A. Wagner, S. Petersen).
100. Bayer, US 3 189 772, 3 925 367, 1970 (G. Boehmke, H. Theidel).
101. Ciba, US 2 610 152, 1946 (F. Ackermann). Sandoz, US 2 654 713, 1951, US 2 791 564, 1952 (F. Fleck).
102. H. v. Pechmann, M. Schaal, *Ber. Dtsch. Chem. Ges.* **32** (1899) 3690.
103. Bayer, US 3 351 482, 1960 (R. Raue).
104. Geigy, US 2 945 033, 1957; US 3 429 880, 1965 (H. Häusermann). Bayer, US 3 242 177, 1963 (C. W. Schellhammer, E. Degener, H. G. Schmelzer, A. Wagner); US 3 244 711, 1964 (O. Berendes, H. Gold, C. W. Schellhammer).
105. Geigy, US 3 123 617, 1961 (H. Häusermann).
106. Bayer, US 3 646 052, 1966 (O. Neuner, A. Dorlars).
107. Sandoz, US 3 288 801, 1962 (F. Fleck, H. Balzer, H. Aebli).
108. Bayer, GB 1 313 253, 1970 (A. Dorlars, C. W. Schellhammer).
109. Bayer, US 4 005 098, 1968 (A. Dorlars, C. W. Schellhammer, W. D. Wirth); US 3 271 412, 1961 (R. Raue, H. Gold).
110. Bayer, EP 74 590, 1981 (J. Koll, H. Theidel, H. H. Mölls).
111. General Aniline & Film Corp., US 2 600 080, 1946 (M. O. Shrader).
112. BASF, GB 741 798, 1952 (E. Tolksdorf, F. Schubert, E. Kern).
113. Mitsubishi, US 3 310 564, 1961 (T. Kasai).
114. Mitsubishi, JP-Kokai 7 1 05 596, 1968 (T. Kasai), *Chem. Abstr.* **75** (1971) 7435c; JP-Kokai 7 1 13 953, 1968 (S. Kasai), *Chem. Abstr.* **76** (1972) 114764b.
115. Mitsubishi, JP-Kokai 7 0 03 671, 1967 (H. Okada, M. Kaneko), *Chem. Abstr.* **73** (1970) 16310 u.
116. Nippon, JP-Kokai 7 4 05 333, 1969 (T. Noguchi, S. Matsunaga).
117. ICI, US 3 157 651, 1961 (J. R. Atkinson, S. Hartley). Bayer, US 4 278 795, 1979 (H. Harnisch). Hoechst, US 4 675 402, 1984 (E. Schinzel).
118. M. Lüscher: "Psychological Aspects of White", in [7] pp. 1 – 11.
119. I. Soljacic, K. Weber: "Die Wirkung organischer Fremdstoffe auf die Fluoreszenz optischer Aufheller", *Textilveredlung* **14** (1979) no. 3, 97 – 101.

120. I. Soljacic, K. Weber: "Ueber die Konzentrationslöschung der Fluoreszenz optischer Aufheller", *Textilveredlung* **6** (1971) no. 12, 796 – 802. "Washing Powders containing FWA", *Res. Discl.* **151** (1976) no. 11, 18 – 20.
121. H. Hefti: "Whitening of Textiles", in [7] pp. 51 – 58.
122. H. K. Banerjee, K. G. Pai: "Textile Optical Whiteners", *Colourage* **22** (1975) no. 7, 25 – 35.
123. R. Williamson: "Fluorescent Brightening Agents", *Int. Dyer Text. Printer Bleacher Finish.* **157** (1977) no. 8, 359 – 364; no. 9, 408 – 416.
124. R. Levene, M. Lewin: "Some Aspects of the Fluorescent Whitening of Textiles", *Colour. Annu.* 1982 – 83, A 44 – A 49.
125. J. J. Donze: "L'azurage optique", *Ind. Text.* (Paris) **1177** (1987) no. 5, 519 – 524.
126. W. Schuerings: "Weissausrüsten von Polyester-Cellulosetextilien", *Text.-Prax. Int.* **29** (1974) no. 11, 1570 – 1584.
127. W. Schuerings: "Optische Aufheller für Baumwolle in Semi- und Vollkontinue-Verfahren", *Textilveredlung* **15** (1980) no. 7, 236 – 241.
128. W. Schoelermann, H. Mantz: "Der Einfluss von Reaktant und Katalysator auf den Weissgrad in der Hochveredlung", *Textilveredlung* **8** (1973) no. 10, 506 – 514.
129. G. Rösch: "L'azurage optique des fibres synthétiques", *Text. Chim.* **28** (1972) no. 4, 43 – 50.
130. A. Decorte: "La blancheur des fibres acryliques", *Text. Chim.* **27** (1971) no. 4, 4 – 10.
131. R. Anliker, H. Hefti, A. Räuchle, H. Schläpfer: "Das Aufhellen von Polyacrylnitril-Fasern", *Textilveredlung* **11** (1976) no. 9, 369 – 375.
132. C. Eckhardt, H. Hefti: "A Comparison of Fluorescent Brightening in the Production of Synthetic Polymer Fibers and in the Textile Industry", *J. Soc. Dyers Colour.* **87** (1971) no. 11, 365 – 370.
133. H. Hefti: "Optische Aufheller für Polyesterfaserstoffe", *Textilveredlung* **4** (1969) no. 2, 94 – 101.
134. Ciba-Geigy, EP-A 0 280 652, 1988 (H. Hefti, U. Lehmann, D. Reinehr, J. Kaschig).
135. Storey Br, GB 38 491/73, 1973.
136. R. Toex: "Optisches Aufhellen von Polyesterfasern im Transferdruck", *Chemiefasern Textilind.* **30** (1980) no. 8, 642 – 643.
137. P. Carter: "FWA for the Soap and Detergents Industry", *MVC Rep. Miljoevardscentrum, Stockholm,* **2** (1973) 51 – 56; *Chem. Abstr.* **82** (1975) 32767 q.
138. C. Eckhardt, R. von Rütte: "FWA in Detergents", in [7] pp. 59 – 64.
139. A. Siegrist: "The Requirements of Present-Day Detergent FWA", *J. Am. Oil Chem. Soc.* **55** (1978) 114. P. Stensby: "FWA" in W. G. Cutler (ed.): Detergency, Part III, Marcel Dekker, New York 1981, pp. 730 – 806. U. Schüssler: "FWA for Detergents, 2nd World Conf. on Detergents", *J. Am. Oil Chem. Soc.* **63** (1986) 187. W. Findley: "FWA for Modern Detergents", *J. Am. Oil Chem. Soc.* **65** (1988) no. 4, 679 – 683.
140. Ciba-Geigy, US 4 311 605, 1982 (C. Eckhardt, R. von Rütte).
141. G. Becker: "Das Waschen bei niedrigen Waschtemperaturen", Dok. Chemiefasertagung Dornbirn, 1977, 1 – 29; 1980, 1 – 13.
142. Unilever, GB 1 286 459, 1972 (J. Evans).
143. Procter & Gamble, EP-A 0 237 119, 1987 (J. Wevers, J. Walker). Bayer, DE 3 726 266 A1, 1987 (U. Schüssler, F. Seng). Sandoz, EP-A 0 314 630, 1988 (J. Farrar, M. Graham).
144. Ciba-Geigy, DE-OS 2 730 246, 1978 (F. Günter, C. Eckhardt).
145. Clorox, US 4 764 302, 1988 (G. Baker, J. Farr, D. Carty), Albright & Wilson, EP-A 0 273 775, 1988 (E. Messenger, A. Martin). Procter & Gamble, EP-A 0 293 040, 1988 (F. de Buzzaccarini, J. P. Boutique).
146. Colgate-Palmolive, GB 2 158 838 A, 1985 (M. Julemont, L. Laitem, M. Marchal); US 4 781 856, 1988 (T. Ouhadi, L. Dehan); DE 3 704 876 A1, 1987 (T. Ouhadi, L. Dehan); US 4 772 413, 1988 (J. Massaux, N. Mineo, L. Laitem).
147. Procter & Gamble, US 4 179 390, 1979 (G. Spadini, I. Tolliday).
148. Ciba-Geigy, US 4 008 172, 1977, Re. 30 227, 1980 (R. von Rütte, W. Gasser). Bayer, DE-OS 3 046 482 A1, 1982 (U. Schimmel, U. Schüssler, A. Brüggemann).
149. W. Mischler: "FWA in the Paper Industry", *Ciba-Geigy Rev.* 1973, no. 1, 35 – 38.
150. H. G. Oesterlin: "Optische Aufhellung gestrichener Papiere", *Papier (Darmstadt)* **36** (1982) no. 2, 66 – 72.
151. A. Wieber, C. Eckhardt: "Mass Whitening of Synthetic Fibers and Plastics", *Ciba-Geigy Rev.* 1973, no. 1, 26 – 29. A. Wieber: "Mass

Whitening of Synthetic Fibers and Plastics", in [7] pp. 65 – 82.
152. Weifang S. D. P., Chin. FZSGS 87 108 249, 1988 (Y. Ding, X. Cao, Z. Zhang).
153. Weifang S. D. P., Chin. FZSGS 87 108 248, 1988 (Y. Ding, X. Cao, M. Zhang).
154. Kodak, GB 47 527/62, 1962 (D. Saunders, E. Knott, M. Mijovic, H. Wealdstone).
155. E. Kowalski, R. Anliker, K. Schmid: "Criteria for the Selection of Solutes in Liquid Scintillation Counting", *Int. J. Appl. Radiat. Isot.* **18** (1967) 307 – 323. C. T. Peng: "Sample Preparation in Liquid Scintillation Counting", *Rev. Amersham Corp.* **17** (1977) 9. A. Dyer: *Liquid Scintillation Counting Practice,* Heyden & Son, East Kilbride, Scotland, 1980.
156. H. H. Koch, M. Pimsler: "A Nonspecific Fluorescent Stain for Detecting and Identifying Fungi and Algae in Tissue", *Lab. Med.* **18** (1987) no. 9, 603 – 606. E. D. Wachsmuth: "A Comparison of the Highly Selective Fluorescence Staining of Fungi in Tissue Sections", *Histochem. J.* **20** (1988) 215 – 221.
157. H. Telle, U. Brinkmann, R. Raue: "Laser Properties of Triazinylstilbene Compounds", *Opt. Commun.* **24** (1978) no. 1, 33 – 37; "Laser Properties of Bis-styryle Compounds", *Opt. Commun.* **24** (1978) no. 3, 248 –250. M. Ringe, H. Güsten, *Ber. Bunsen-Ges. Phys. Chem.* **90** (1986) 439 – 444.
158. J. Lanter: "Properties and Evaluation of FWA", *J. Soc. Dyers Colour.* **82** (1966) 125 – 132.
159. G. Anders: "Limits of Accuracy Obtainable in the Direct Determination by Fluorimetry of FWA on TLC", in [7] pp. 104 – 110.
160. H. Theidel: "Qualitative TLC of FWA", in [7]pp. 94 – 103.
161. H. Theidel: "Direct Determination of FWA by Absorption Measurement in situ on TLC", in [7]pp. 111 –114.
162. G. R. Weis: "Säulenchromatographische Auftrennung von Aufhellergemischen", *Tenside Deterg.* **12** (1975) no. 1, 43 – 44.
163. D. Kirkpatrick: "Separation of Optical Brighteners by Solid-Liquid Chromatography", *J. Chromatogr.* **139** (1977) 168 – 173.
164. B. McPherson, N. Omelczenko: "The Determination of Optical Brighteners in Laundry Detergents by Reverse Phase and Ion Pair High Performance Liquid Chromatography", *J. Am. Oil Chem. Soc.* **57** (1980) no. 11, 388 – 391. A. Nakae: "Separation and Determination of Fluorescent Whitening Agents in Detergents by High Speed Liquid Chromatography", *Bunseki Kagaku* **29** (1980) no. 1, 69 – 73. G. Micali, P. Curro, G. Calabro: "High Performance Liquid Chromatographic Separation and Determination of FWA in Detergents", *Analyst (London)* **109** (1984) no. 2, 155 – 158.
165. K. Tsuji, S. Setsuda, S. Naito, A. Abe: "Determination of Fluorescent Whitening Agents in Detergents by High Performance Liquid Chromatography", *Bull. Kanagawa P. H. Lab.* **1981** , 65 – 67.
166. J. L. Jasperse, P. H. Steiger: "A System of Determining Optical Brighteners in Laundry Detergents by TLC and HPLC", *J. Am. Oil Chem. Soc.* **109** (1992) 621 – 625.
167. C. Eckhardt: "Examen del blanqueo optico de los textiles tenidos o preblanqueados", *Invest. Inf. Text. Tensioactivos* **16** (1973) no. 4, 465 – 485.
168. B. Werthmann, R. Borowki: "Möglichkeiten zur quantitativen Bestimmung optischer Aufheller für Papier", *Papier (Darmstadt)* **30** (1976) no. 6, 243 –249.
169. W. Archwal, A. Narkar: "Die Bestimmung optischer Aufheller auf Baumwolle mittels Cadoxen als Lösungsmittel", *Textilveredlung* **7** (1972) no. 1, 19 –24.
170. H. Theidel: "Zur Stereochemie optischer Aufheller vom Stilben-Typ", *Melliand Textilber.* **45** (1964) no. 5, 514 – 519. W. Weller: "The Preparation of the Pure cis-Isomers of Four FWA of the Sulfonated Stilbene Type", *J. Soc. Dyers Colour.* **95** (1979) no. 5, 187 – 190. K. Smit, K. Ghiggino: "Fluorescence and Photoisomerisation of Two Stilbene-Based Dyes", *Dyes Pigm.* **8** (1987) no. 2, 83 – 97.
171. J. Schulze, T. Polcaro, P. Stensby: "Analysis of FWA in US Home Laundry Detergents", *Soap Cosmet. Chem. Spec.* **50** (1974) no. 11, 46 – 52.
172. H. Bloching, W. Holtmann, M. Otten: "Beitrag zur Analytik von Weisstönern in Waschmitteln", *Seifen Oele Fette Wachse* **105** (1979) no. 2, 33 – 38.
173. T. Poiger, J. A. Field, T. M. Field, W. Gieger: "Determination of Detergent Derived Fluorescent Whitening Agents in Sewage Sludges by Liquid Chromatography", *Analytical Methods and Instrumentation* **1** (1993) 104 – 113.
174. T. Poiger, J. A. Field, T. M. Field, W. Gieger: "Occurrence of Fluorescent Whitening Agents

in Sewage and River Water Determined by Solid-phase Extraction and High-performance Liquid Chromatography", *Environmental Science and Technology* **30** (1996) 2220 – 2226.
175. G. Anders, E. Ganz: "Metameric White Samples for Testing the Relative UV Content of Light Sources and of Daylight", *Appl. Opt.* **18** (1979) 1067 – 1072.
176. C. Eckhardt: "Visual Assessment of Whiteness", *Ciba-Geigy Rev.* 1973, no. 1, 10 – 13.
177. G. Anders: "Control of Whites: The Cibanoid White Scale", *J. Soc. Dyers Colour.* **84** (1968) no. 2, 125 – 132. J. Rieker, D. Gerlinger: "Neuerungen auf dem Gebiet der farbmetrischen Weissbewertung", *Melliand* **65** (1984) 631 – 632.
178. F. Gärtner, R. Griesser: "Eine Vorrichtung zur Messung von optischen Aufhellern mit konstanter UV-Anregung", *Farbe* **24** (1975) 199 – 207.
179. D. Eitle, E. Ganz: "Eine Methode zur Bestimmung von Normfarbwerten für fluoreszierende Proben", *Textilveredlung* **3** (1968) 389 – 392.
180. G. Wyszecki, W. S. Stiles: *Color Science, Concepts and Methods, Quantitative Data and Formulas,* J. Wiley & Sons, New York – London – Sydney 1967.
181. R. Hunter, *J. Opt. Soc. Am.* **48** (1958) 597. A. Berger, *Farbe* **8** (1959) 187. P. Stensby, *Soap & Chem. Spec.* **43** (1967) 80.
182. A. S. Stenius et al., *Farbe* **26** (1977) 3 – 104.
183. S. V. Vaeck: "Some New Experiments on the Colorimetric Evaluation of Whiteness", *J. Soc. Dyers Colour.* **95** (1979) 262 – 269.
184. F. Grum: "Whiteness and Fluorescence", *Text. Chem. Color.* **13** (1981) 187 – 189.
185. E. Ganz: "Whiteness: Photometric Specification and Colorimetric Evaluation", *Appl. Opt.* **15** (1976) 2039 – 2058. E. Ganz: "Whiteness Formulas: A Selection", *Appl. Opt.* **18** (1979) 1073 – 1078. E. Ganz, R. Griesser: "Whiteness: Assessment of Tint", *Appl. Opt.* **20** (1981) 1395 – 1396. R. Griesser: "Stand der instrumentellen Weissbewertung unter besonderer Berücksichtigung der Beleuchtung", *Textilveredlung* **18** (1983) 157 – 162.
186. A. Brockes: "The Evaluation of Whiteness", *CIE-J.* **1** (1982) no. 2, 38 – 39.
187. B. Mensak: "Exakte Weißmetrik mit Spektralphotometern", *Maschen Industrie* **3** (2000) 40 – 42.
188. R. Griesser: "Assessment of Whiteness and Tint of Fluorescent Substrates with Good Inter-Instrument Correlation", *Color Res. Appl.* **19** (1994) 446 – 460.
189. R. Hayhurst, K. Smith: "Instrumental Evaluation of Whiteness", *J. Soc. Dyers Colourists* **111** (1995) 263 – 266.
190. E. J. van de Plassche, P. F. H. Bont, J. M. Hesse: ″Exploratory Report Fluorescent Whitening Agents (FWAs)″, Dutch National Institute of Public Health and Environment, Report 601503013 (1999).
191. J. Plautz, P. Richner: ″Human & Environmental Risk Assessment on ingredients of European household cleaning products: Fluorescent Brightener FWA-5″, HERA, International Association for Soaps, Detergents and Maintenance Products, A.I.S.E., (2001); http://www.heraproject.com/RiskAssessment.-cfm.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Optically Active Compounds

Standard Article

David N. A. Fox[1] and Timothy Gallagher[2]
[1]Bath University, Bath, United Kingdom
[2]Bath University, Bath, United Kingdom

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a18_177
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Analysis of Optically Active Compounds, Determination of Optical Purity, and Assignment of Absolute Configuration
2.1. Chromatographic Analysis
2.2. Spectroscopic Analysis
3. Occurrence and Production of Optically Active Compounds
3.1. Resolution
3.2. Synthesis of Asymmetric Compounds
3.2.1. Chiral Pool Approach
3.2.2. Use of Chiral Reagents and Auxiliaries
3.2.3. Asymmetric Chemical Catalysis
3.2.4. Enzymatic Transformations

[Top of Page]

1. Introduction

The ability of certain compounds to rotate the plane of polarized light is termed optical activity. This phenomenon was first observed in 1815 by the French scientist BIOT while studying the interaction of plane polarized light with organic materials such as sugar and camphor. Thirty years later, PASTEUR demonstrated that two distinct types of crystals could be separated manually from an optically inactive salt of tartaric acid. These two forms were shown to have bulk structures that were nonsuperimposable mirror images of one another, and each type of crystal exhibited optical activity to a similar extent but in an opposite direction. The crystals that rotated the light plane to the right (clockwise) were designated (+) and termed dextrorotatory (d). Those rotating the plane of polarized light to the left (counterclockwise) were assigned the configuration (–) or levorotatory (l).

Although PASTEUR was able to separate these two optical isomers (enantiomers), the link with molecular structure did not become apparent until 1874. Following KEKULÉ's structural theory that carbon was invariably tetravalent, VAN'T HOFF reasoned that a carbon atom that is attached to four groups would have a tetrahedral geometry. Hence, if the four groups were different, the molecule would be asymmetric and nonsuperimposable on its mirror image. Molecules possessing this property of "handedness" are said to be chiral (Greek: *cheir*, hand), and molecular chirality results in the phenomenon of optical activity that is observed with a chemical compound. This may be illustrated by lactic acid.

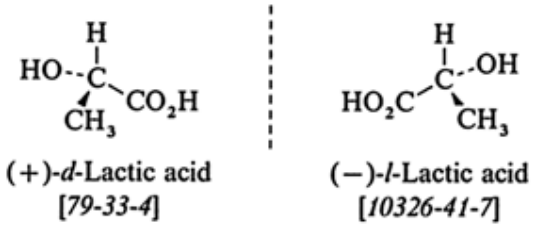

(+)-*d*-Lactic acid [*79-33-4*] | (–)-*l*-Lactic acid [*10326-41-7*]

If a molecule has a plane or center of symmetry, it is achiral.

The two enantiomers of lactic acid, like all pairs of enantiomers in an achiral environment, have identical physical and chemical properties, except for the direction in which they rotate the plane of polarized light and the way in which they interact with other enantiomers.

A consequence of the fact that optical isomers rotate polarized light equal amounts but in opposite directions is that a 1 : 1 mixture of enantiomers is optically inactive. Such a mixture is called a racemate or a racemic mixture in the case of distinct crystalline enantiomers. Racemates comprise eutectic mixtures and behave as though they are pure compounds; thus, they exhibit physical and chemical properties that differ from those of a single enantiomer.

Optical activity is most commonly observed as a property resulting from molecular structure (in the case of solutions, gases, or liquids). However, optical activity can also be the result of a bulk phenomenon in solids such as quartz where there is a dependence on crystal structure. For the latter, the optical activity is lost on melting or dissolution. In this article, only the former type of optical activity is considered.

Specific Rotation. The optical rotation of a compound is described in quantitative terms by the specific rotation [α]. This value is obtained by using a polarimeter in which polarized light is first passed through a solution of the optically active sample and then analyzed to determine the angle through which the plane of polarization has been rotated.

The extent to which light is rotated depends on a number of variables that are discussed in more detail later. Specific rotation attempts to take account of all of these factors and provide a standardized figure for optical activity that can be used for comparison purposes. It is defined for a particular concentration of sample (*c*, expressed as grams per 100 ml) at a temperature *t* (°C) with light of wavelength λ by the equation

$$[\alpha]_{\lambda}^{t} = \frac{100\alpha}{l\,c}$$

where α is the observed rotation in degrees and *l* the sample path length in decimeters. Specific rotation is often measured by using light corresponding to the D line of sodium (λ = 589 nm) and expressed as [α] $_D$.

Representation and Nomenclature of Optically Active Compounds Many naturally occurring compounds are composed of chiral molecules and a large proportion of these compounds exist as single optical isomers. The origin of this optical activity in nature is a highly debated question, with biotic or evolutionary arguments being put forward on one hand and abiotic mechanisms such as chance or parity violations being offered as an alternative [3]. Theories to explain how a slight enantiomeric enrichment of a mixture is amplified to the exclusion of the minor component have also been propounded.

As a result of the abundance and importance of optically active compounds, various systems of nomenclature have been developed to describe absolute configuration. These are based either on the direction in which the material rotates the plane of light or on the molecular stereochemistry of the particular enantiomer being considered.

Fischer Projections. Before such systems were introduced, a convention for representing chiral molecules in two dimensions was developed by FISCHER. The Fischer projection is drawn such that the two horizontal lines represent bonds projecting out of the plane of the paper and those drawn vertically are deemed to be directed into the plane.

(+)-Lactic acid may be represented in a Fischer projection as

(+)-Lactic acid
$[\alpha]_D^{15} = +\ 3.82°(c = 10.5,\ H_2O)$

Some manipulations of these projections (e.g., 180 ° rotation) are permissible (i.e., the configuration is retained) and allow easy comparison with other molecules. However, certain transformations (e.g., interchange of any two groups) result in inversion of the absolute configuration.

D and L Notation. Lactic acid has already been described in terms of the (+)- or *d*- and (–)- or *l*-isomers, by using a system that is based solely on the direction in which each isomer rotates the plane of polarized light. Early in this century, ROSANOFF proposed a system that would relate all molecules containing asymmetric carbon atoms to either (+)- or (–)-glyceraldehyde. Molecules that could be derived from (+)-glyceraldehyde [453-17-8] were designated D and those that were correlated with (–)-glyceraldehyde [497-09-6]were denoted L. The D-(+)-isomer of glyceraldehyde was assigned, arbitrarily, the configuration

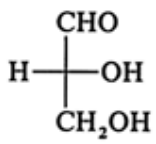

D-(+)-Glyceraldehyde
[*453-17-8*]
$[\alpha]_D^{25} = +\ 8.7°\ (c = 2.0,\ H_2O)$

Thus, by correlating an unknown asymmetric center with either (+)- or (–)-glyceraldehyde, the absolute configuration (either D or L) of a molecule was defined. It was not until 1951, when BIJVOET, by applying X-ray crystallographic techniques to sodium rubidium tartrate, showed that the stereochemical assignment of (+)-glyceraldehyde proposed by ROSANOFF was, in fact, correct [4].

The DL system is not a simple one to apply. First, molecules with identical configurations do not necessarily rotate the plane of polarized light in the same direction, e.g., (–)-glyceric acid [6000-40-4] and (+)-glyceraldehyde both have D absolute configuration. Second, some enantiomers can, by a number of steps, be related to either the D- or L-enantiomer of a known molecule, and such a situation requires an arbitrary assignment. For these reasons, this somewhat cumbersome method of assigning absolute configuration has been superseded by the Cahn – Ingold – Prelog system of nomenclature.

The Cahn – Ingold – Prelog system of nomenclature [5] assigns the absolute configuration of an asymmetric center according to the stereochemical arrangement of the four groups about that center, by application of a set of sequence rules. These rules are independent of both the magnitude and the sign of rotation. The summary of the rules given below is sufficient for assigning the vast majority of chiral compounds.

1. Substituents are ranked in order of decreasing atomic number of the atom attached directly to the central carbon.
2. If rule 1) is insufficient to assign priorities (i.e., two or more of the atoms connected to the asymmetric carbon are identical), the atomic number of the atom at the first point of difference along the substituent skeleton determines the order.
3. Multiply bonded atoms are considered as if they were an equivalent number of singly bonded atoms. Thus, a –CHO substituent is equivalent to $–CH(O)_2$.

For lactic acid, therefore, the four groups in order of decreasing priority are: –OH, $–CO_2H$, $–CH_3$, and –H.

Once the order has been determined, the molecule is oriented so that the substituent with the lowest priority is pointing away from the observer. If the direction of an arrow starting at the highest priority substituent, passing through the second, and ending at the third priority group is clockwise, the molecule is assigned *R* configuration (Latin: *rectus*, right). Conversely, the molecule has the *S* configuration (Latin: *sinister*, left) if the direction indicated by the arrow is counterclockwise.

By using (+)-lactic acid once again as the example, application of the sequence rules allows an assignment of *S* configuration to be made.

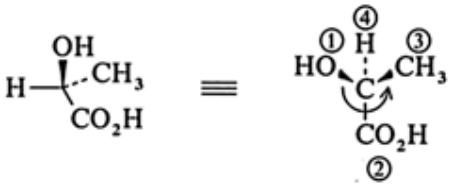

Prochirality. A carbon atom attached to four groups, two of which are the same (CX_2YZ) is achiral because it is superimposable upon its mirror image. Such a molecule is termed prochiral because replacement of one of the two identical groups X by a fourth group W results in a chiral species whose configuration depends on which enantiotopic X group is replaced. The *pro-R* substituent, when replaced by a substituent W of higher priority, affords a product with *R* configuration. Similarly, the *pro-S* substituent results in a chiral molecule with *S* configuration when replaced by W. Such considerations become important when reactions are carried out in an asymmetric environment (e.g., enzyme-mediated transformations), where often only one of the two X groups in a prochiral molecule undergoes reaction [6].

Molecules with More than One Asymmetric Center. In a molecule containing a series of *n* asymmetric centers, each center may be assigned either *R* or *S* configuration according to the sequence rules. Thus, a maximum of 2^n stereoisomers exists that form 2^{n-1} pairs of enantiomers. Stereoisomers that are not enantiomers of one another are related to each other by having some, but not all, asymmetric centers of common absolute configuration. Such isomers are termed *diastereomers*; they differ in physical and chemical properties as well as in their specific rotations. A special class of diastereomers includes those that have all but one asymmetric center with the same absolute configuration. These are termed *epimers*.

A *meso compound* is a diastereomer that is achiral by virtue of a plane of symmetry present in its molecular structure. For a molecule (e.g., tartaric acid) with two asymmetric centers, in which the three different substituents attached to these centers are common to both, only three of the maximum four (2^n, $n = 2$) stereoisomers are possible, i.e., one diastereomer comprising two enantiomers and a *meso*-diastereomer.

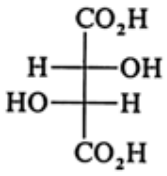

[87-69-4]
(+)-*d*-Tartaric acid
$[\alpha]_D^{20} = +12.0°$ (*c* 20, H_2O)

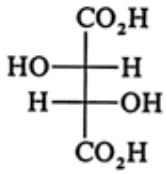

[147-71-7]
(−)-*l*-Tartaric acid
$[\alpha]_D^{20} = -12.0°$ (*c* 20, H_2O)

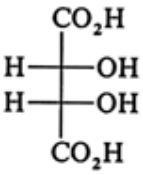

[147-73-9]
meso-Tartaric acid

Corresponding to enantiotopic groups in prochiral molecules, diastereotopic groups are two similar groups attached to the same carbon atom that result in diastereomers when one group is replaced. Diastereotopic groups differ in both chiral and achiral environments.

The two faces of some *trigonal species*, such as alkenes and carbonyl groups, may also be distinguished and are termed enantiotopic or diastereotopic depending on whether reaction at each face results in the formation of enantiomeric or diastereomeric products. Enantiotopic and diastereotopic faces are named by an extension of the Cahn – Ingold – Prelog system [7]. For the trigonal system XYC = Z, the double bond is treated as a single bond to a Z-C substituent. The three substituents (X,Y,Z) are then assigned priorities according to the sequence rules and if, from the direction observed, the groups are arranged in a clockwise manner, this face is termed *Re*. Conversely, if they are arranged counterclockwise, the *Si* face is observed.

Diastereoselectivity and Enantioselectivity. A reaction in which diastereomers (or enantiomers) are obtained as products is termed diastereoselective (enantioselective) if one diastereomer (enantiomer) is formed preferentially. If the starting materials and reaction environment are achiral or racemic, then the products must be optically inactive (i.e., no enantioselectivity is observed).

However, in a chiral environment or with a reagent that is chiral and nonracemic, the reaction is potentially enantioselective.

Enantiospecificity and *diastereospecificity* are terms that relate to stereospecific reactions within the context of enantiomers or diastereomers.

Other Types of Chirality. Thus far, the discussion has dealt solely with asymmetric carbon atoms as a source of chirality. However, any molecule containing an atom that has bonds to four different groups arranged tetrahedrally is chiral and can exhibit optical activity. Examples are silicon [8], germanium, zinc [9], certain transition metals, and nitrogen in quaternary salts or *N*-oxides. Pyramidal nitrogen atoms do not, in general, exhibit optical activity because racemization via inversion is rapid. However, certain aziridines, in which inversion is slow, have been resolved [10], as well as molecules incorporating a conformationally rigid asymmetric bridgehead nitrogen [11]. Phosphorus, arsenic, and antimony compounds undergo inversion more slowly and more commonly exhibit optical activity [12].

Other classes of chirality leading to optical activity are listed below.

Restricted Rotation About Single Bonds – Atropisomerism [13]. Biphenyls, with four large groups in the ortho positions of each ring, cannot freely rotate about the central bond. In such compounds the two rings are in perpendicular planes, as shown below. If W X and Y Z, the molecule lacks a plane of symmetry and is chiral, and the two enantiomers may be resolved.

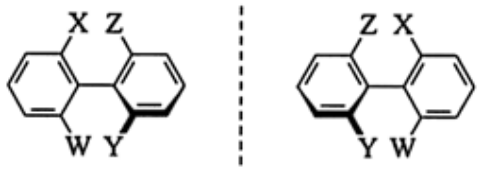

Allenes. Owing to the *sp* hybridization of the central carbon atom, allenes may possess perpendicular dissymmetric planes (i.e., be chiral) if both sp^2-hybridized carbons are unsymmetrically substitued [14].

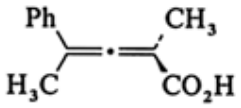

(*S*)-2-Methyl-4-phenyl-2,3-pentadienoic acid
[*36854-46-3*]
$[\alpha]_D^{25} = +149°$ ($c = 0.4$, EtOH)

Chirality Due to Helical Shape. Geometrical constraints in *trans*- cyclooctene result in two possible helical forms, which confers chirality on the molecular structure (the absolute values of the specific rotation of the following enantiomers are not identical, for an explanation see Chap. Analysis of Optically Active Compounds, Determination of Optical Purity, and Assignment of Absolute Configuration) [15].

R-(−)-*trans*-Cyclooctene
$[\alpha]_D^{25} = -458°$ (neat)

S-(+)-*trans*-Cyclooctene
$[\alpha]_D^{25} = +440°$ (neat)

Chirality Caused by Restricted Rotation of Other Types. Metallocenes (e.g., ferrocenes) with at least two different groups on one ring are chiral and are readily resolved [16].

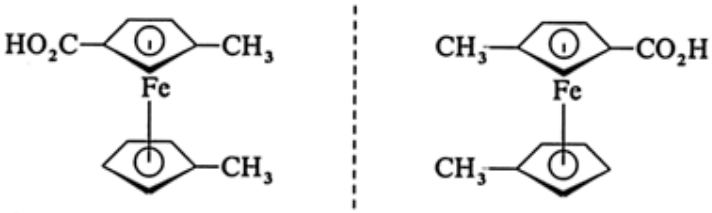

(1*R*)-(+)-1′,3-Dimethyl-ferrocenecarboxylic acid
[*12295-37-3*]
$[\alpha]_D^{22} = +37.5°$ ($c = 0.8$, $CHCl_3$)

(1*S*)-(−)-1′,3-Dimethyl-ferrocenecarboxylic acid
[*12126-75-9*]
$[\alpha]_D^{22} = -36.0°$ ($c = 0.8$, $CHCl_3$)

[Top of Page]

2. Analysis of Optically Active Compounds, Determination of Optical Purity, and Assignment of Absolute Configuration

The optical purity of a compound consisting of two enantiomers (+)-A and (–)-A is commonly expressed as a percentage in terms of enantiomeric excess (% e.e.), where

$$\%\ \text{e.e.} = \frac{\%(-)-\text{A} - \%(+)-\text{A}}{\%(-)-\text{A} + \%(+)-\text{A}}$$

An indirect method for obtaining % e.e. relies on optical rotation, where [α_{max}] is the maximum rotation of the pure enantiomer.

$$\%\ \text{e.e.} = \frac{[\alpha_{\text{observed}}] \times 100}{[\alpha_{\text{max}}]}$$

Although *polarimetry* (i.e., methods of analysis based on specific rotation) and the related techniques of *optical rotatory dispersion* (ORD) and *circular dichroism* (CD) have been used extensively to determine % e.e., the limitations of these methods should be recognized. Specific rotation is concentration, temperature, and wavelength dependent, and care should be taken when comparisons are made to earlier literature. In addition, the presence in a sample of even small amounts of contaminants with high specific rotation gives misleading results.

For these reasons and because of an ever increasing demand for more reliable and precise analytical techniques, other methods for the determination of % e.e. have been developed. These are based on chromatographic and spectroscopic procedures relying on the generation of a diastereomeric interaction that is either covalent or noncovalent in nature.

2.1. Chromatographic Analysis

Reaction of a mixture of enantiomers with an optically pure reagent leads to a pair of diastereomers. For example, acylation of amines or alcohols with an optically pure carboxylic acid leads to diastereomeric amides and alcohols respectively that may be analyzed by e.g., GLC or HPLC. Optically pure amines and alcohols have also been used to derivatize a range of other types of compounds. Derivatives of this type may be used to establish absolute configuration (see Section Use of Chiral Reagents and Auxiliaries).

Alternatively, enantiomers can be analyzed via a chiral recognition mechanism by using an independent *chiral stationary*

phase. A variety of naturally occurring and synthetic chiral stationary phases are commercially available, and these materials allow the advantages of GLC and HPLC, such as speed and operational simplicity, to be employed on both analytical and preparative levels. Reverse phase supports for both HPLC and TLC are also available and are particularly important in the separation of amino acid enantiomers.

2.2. Spectroscopic Analysis

NMR spectroscopy is a flexible and reliable method for the determination of enantiomeric excess. A mixture of enantiomers may be derivatized, as described above, to give spectroscopically distinguishable diastereomers, with ^{1}H NMR being the most widely used technique.

To make the analysis straightforward, a derivative is usually chosen that contains a sharp, readily observable signal such as a methoxy moiety. An example is (*R*)- or (*S*)- α-methoxyphenylacetic acid [3966-32-3], [26164-26-1], respectively, which has been used extensively to analyze enantiomeric mixtures of both amines and alcohols. The Mosher reagent, (*R*)- or (*S*)- α-methoxy- α-(trifluoromethyl)phenylacetic acid (MTPA) [20445-31-2], [17257-71-5], respectively, is a useful alternative because it is stable toward acid- or base- catalyzed racemization, which is sometimes a problem with α-methoxyphenylacetic acid derivatives. In addition, ^{19}F NMR as well as ^{1}H NMR may be used for analysis of MTPA derivatives, and these are also suitable for analysis by GLC and HPLC.

Two other NMR methods have been used to determine the enantiomeric excess of optically active compounds. Enantiomers can often be distinguished by NMR when dissolved in a *chiral solvent*. A range of these chiral solvating agents, such as the commercially available aryltrifluoromethylmethanols and 1-arylethylamines, have found extensive application.

Optically pure variants of the well-known *lanthanide shifts reagents* provide another method of analysis. In general, the lanthanideinduced shifts are larger than those observed with chiral solvating agents, although line-broadening effects may, on occasion, present difficulties. The following are examples of the commercially available reagents, together with their commonly used abbreviations:

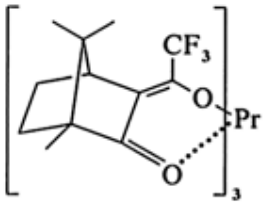

Tris(3-trifluoromethylhydroxy-methylene-*l*-camphorato) praseodymium(III) $Pr(tfc)_3$

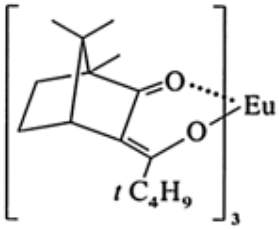

Tris(3-*tert*-butylhydroxy-methylene-*d*-camphorato) europium(III) $Eu(bhmc)_3$

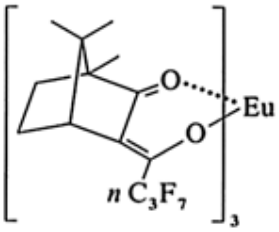

Tris(3-heptafluoropropyl-hydroxymethylene-*d*-camphorato)europium(III) $Eu(hfbc)_3$ or $Eu(hfc)_3$

Assignment of Absolute Configuration. A range of *spectroscopic* procedures are available for the assignment of absolute configuration to a molecule. These methods have been successfully applied on numerous occasions, usually within a related series of compounds and, although of great value, should all be employed with care. *Chemical correlation* with a structure of known configuration is another useful tool, but once again this method can be dangerously misleading for the inexperienced or unwary.

X-ray crystallographic analysis can be used to assign a configuration directly to an enantiomer, but the analysis is more commonly carried out on a diastereomeric derivative containing a stereocenter of known absolute stereochemistry; e.g., MTPA can be used to prepare diastereomeric esters and amides with chiral alcohols and amines, respectively. The problem then reduces simply to one of assigning a relative stereochemistry to the known and the unknown stereocenters, and this technique has found widespread application [17].

[Top of Page]

3. Occurrence and Production of Optically Active Compounds

Although nature frequently exhibits a high degree of symmetry in terms of general morphology, at the molecular level the natural world is highly asymmetric. Chirality is based primarily on the asymmetric carbon center, and enzymes, proteins, polysaccharides, nucleic acids, and many other basic components of plants and animals are all present in optically active

form. The implications of this on the properties of biological systems are profound. The interaction of a chiral molecule with, for example, an enzyme or receptor protein is a diasteromeric one, and a corollary of this is that two enantiomers can exhibit quite different properties. For example, (*R*)-asparagine [2058-58-4] has a bitter taste, whereas the (*S*)-isomer [70-47-3] is sweet. Clearly, if a chiral molecule is directed toward a biological target, the two enantiomers should be viewed as distinct compounds that are capable of acting in different ways. For example, (*R*)-propranolol [5051-22-9] acts as a contraceptive, whereas the (*S*)-isomer [4199-09-1] is a β-blocker. Because the potency, absorption, transport, degradation, and excretion of the two enantiomers may be quite different within the body, a racemic drug may only be "50 % pure." Although this may not be a problem, in the worst scenario the unwanted enantiomer can be highly toxic. As a consequence, the requirement for optically pure compounds is increasing, and legislation is forcing this issue in the pharmaceutical and agrochemical industries. In 1988 the FDA announced a set of guidelines addressing these stereochemical topics in relation to the submission of new drug applications. As a consequence, six of the world's top fifteen drugs are currently marketed as single enantiomers. Another factor involved is that manufacturing an unnecessary enantiomer creates waste that may be financially, as well as socially or environmentally, unacceptable.

Two basic methods exist for the synthesis of optically pure enantiomers:

1. Resolution
2. Asymmetric synthesis

3.1. Resolution

The resolution of racemates is an important source of optically pure materials. Classical resolution of a racemate (±)-A requires the use of an optically pure resolving agent, e.g., (+)-R. Interaction (which may be ionic or covalent in nature) leads to the diastereomers [(+)-A/(+)-R] and [(–)-A/(+)-R], which are separated by crystallization or chromatography. Cleavage of [(+)-A/(+)-R] gives optically pure (+)-A, which is then separated from (+)-R; the latter may be recovered. The enantiomer (–)-A is obtained analogously from diastereomer [(–)-A/(+)-R]. Optically pure carboxylic acids and amines are commonly used as resolving agents because they are readily available and their acidic – basic properties allow easy separation of the resolving agent from the enantiomer of interest. Commonly used *carboxylic acid* resolving agents include tartaric acid (and related analogues); camphorsulfonic acid, (1*R*)- [35963-20-3] and (1*S*)- [3144-16-9]; mandelic acid, (*R*)- [611-71-2] and (*S*)- [17199-29-0]; and pyroglutamic acid, (*S*)- [98-79-3] and (*R*)- [4042-36-8]. Optically pure *amines* used as resolving agents include alkaloids such as brucine [357-57-3] and cinchonidine [485-71-2], as well as ephedrine [299-42-3] and pseudoephedrine [90-82-4]. Both (*R*)- and (*S*)- α-methylbenzylamine [3886-69-9], [2627-86-3], respectively, are commercially available and are extensively employed in resolution processes.

A good example of the use of classical resolution is found in the synthesis of naproxen [22204-53-1], which is the only α-arylpropionic acid anti-inflammatory agent that is actually marketed in the optically pure *S* form. The chiral amine 1-deoxy-1-(octylamino)-D-glucitol [23323-37-3] is used as the resolving agent, and the efficiency of the process is high because the corresponding (*R*)-isomer, can be easily racemized and recycled.

Although resolution is an important source of optically pure compounds, the trial-and-error approach connected with it can be very time consuming. Considerable effort has been made in recent years to understand the factors that control the efficiency of the resolution process, and chemists are beginning to identify rational approaches to a technique that is often considered more art than science.

3.2. Synthesis of Asymmetric Compounds

The asymmetric synthesis of optically active compounds can be achieved in several ways:

1. Use of the "chiral pool"
2. Stoichiometric and catalytic methods for enantioselective synthesis
3. Enzyme-mediated processes

3.2.1. *Chiral Pool Approach*

Nature provides a wide range of optically active, although not always optically pure, materials that can be used as a starting point for asymmetric synthesis. These materials — alkaloids, terpenes, sugars, α-amino- and α-hydroxyacids — represent the so-called chiral pool.

Numerous examples exist of the use of the chiral pool approach in synthesis, many of which have important industrial applications. (*R*)-Limonene [5989-27-5], isolated from orange peel as a byproduct of the orange juice industry, provides the starting material for the commercial production of (*R*)- carvone [6485-40-1], the main constituent of spearmint flavor. The clinically important β-lactam antibiotic thienamycin [59995-64-1] is synthesized from (*S*)-aspartic acid [56-84-8], an inexpensive and readily available member of the chiral pool. Unlike many related antibiotics, thienamycin is not yet available in useful amounts by fermentation methods. However, the classical penicillins and cephalosporins that are produced by fermentation processes may themselves be regarded as components of the chiral pool, because they are used as the basis for the synthesis of a host of commercially successful drugs.

3.2.2. *Use of Chiral Reagents and Auxiliaries*

Since 1979, chiral, nonracemic reagents have been used increasingly to induce asymmetry in reaction products [18]. *Asymmetric induction*, as distinct from asymmetric catalysis (see Section Asymmetric Chemical Catalysis), involves the use of stoichiometric quantities of a chiral moiety, either a reagent or a residue attached covalently to the substrate. Removal of the reagent or auxiliary after reaction yields optically active products, and successful systems lead to chiral compounds of near 100 % optical purity. The chiral environment provided by the chiral moiety in the transition state allows the reaction to proceed solely via one of two diastereomeric pathways. At ambient temperature, a free activation energy difference $\Delta G^{\neq}$ = 12 kJ/mol between the two paths is equivalent to an enantiomeric excess in the final product of 99 %.

Thus relatively small energy differences lead to substantial stereoselectivity, and maximization of this discrimination forms the basis of asymmetric synthesis.

Some fundamental guidelines are available for the design of an efficient asymmetric synthesis:

1. The desired enantiomer should be obtained in high optical and high chemical yield.
2. The chiral auxiliary should be close to the reacting center but should not adversely affect the rate and should ideally be available in both enantiomer forms.
3. The chiral product must be readily separable from any chiral auxiliary or reagent required for synthesis.
4. The incorporation and subsequent removal of any chiral auxiliary should be chemically efficient and achieved without loss of optical activity from either the product of the reaction or the auxiliary itself.
5. If a catalytic process is used, the reagent may not have to be recovered, provided the turnover number is high.

Examples abound of the application of chiral reagents in asymmetric synthesis. Reagents based on hydridoaluminate and hydridoborate complexes have been increasingly successful in the field of asymmetric reduction since 1979. This is illustrated by the work of NOYORI [19] with (*R*)-Binal-H [70945-91-4] and MIDLAND [20] with (1*R*)-B- β-pinanyl-9-borabicyclo [3.3.1]nonane [64106-79-2].

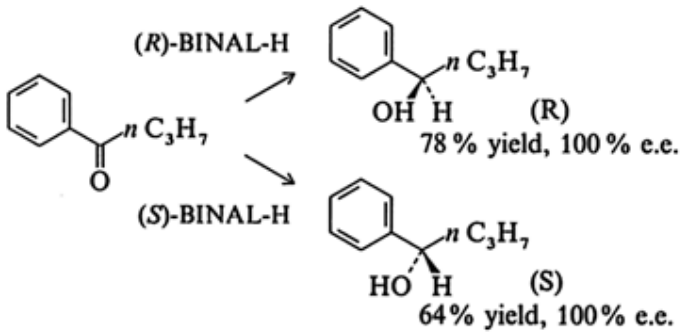

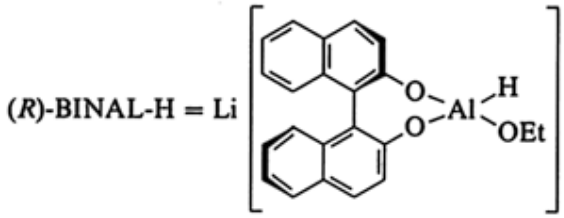

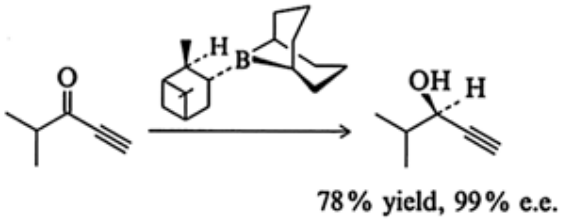

(1*R*)-B-β-pinanyl-9-borabicyclo[3.3.1]nonane (derived from (+)-α-pinene)

Other boron-containing reagents, developed by BROWN [21], have been shown to be versatile and highly selective. Diisopinocampheylborane (Ipc_2BH) [21947-87-5] and monoisopinocampheylborane ($IpcBH_2$) [64234-27-1], both readily prepared in > 99 % e.e., are used extensively in the asymmetric reduction of a variety of functional groups, as well as the asymmetric hydroboration of a range of substituted alkenes.

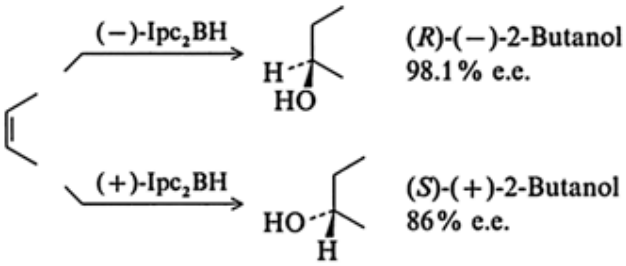

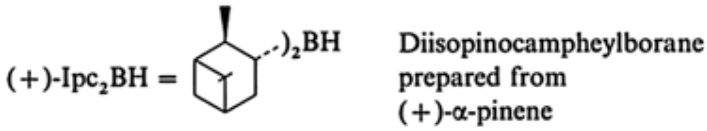

Chiral amide bases that are related to the classic lithium amides (see examples below) have been used more recently for

the desymmetrization of prochiral ketones or the deracemization of racemic ketones [22], [23].

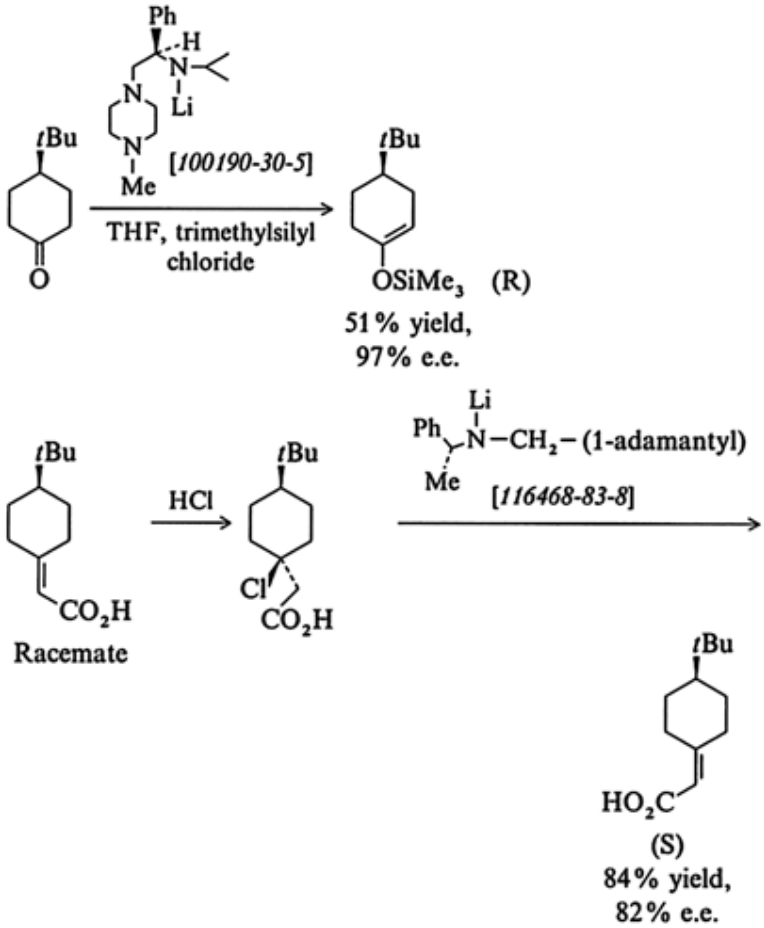

The base selects between enantiotopic protons in kinetically controlled deprotonations, and this area will develop significantly over the next few years as the mechanism by which selectivity is achieved becomes clear.

The chiral auxiliary approach to asymmetric synthesis has, over the last decade, been refined to such an extent that it now provides a versatile and often highly predictable route to a wide range of optically pure compounds. The auxiliary serves both as a chiral-inducing agent promoting the preferential formation of one configuration at the reactive center and as an internal resolving agent. This latter function allows separation of the major diastereomer and thus, on auxiliary removal, formation of optically pure products.

Because of its chemical reactivity and versatility , the carbonyl group is an especially suitable starting group, and a wide range of chiral auxiliaries (Xc) have been developed for use in this area.

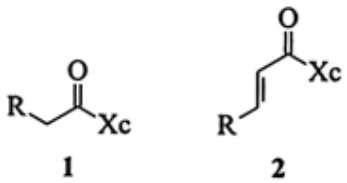

The incorporation of Xc into general structures **1** and **2** allows a wide range of transformations to be carried out in a diastereoselective fashion. These include enolate-based reactions, [24-26] (alkylations and aldol condensations), 1,4-additions (organocuprates) [27], cyclopropanation [28], and Diels – Alder cycloadditions [29], [30]. In all cases the role of Xc is to confer a preferred conformation on an enolate or enone in the substrate, often associated with the formation of a metal chelate, which is accompanied by selective attack at one of the two diastereotopic faces of the reactive prochiral center or centers. The reactivity of the carbonyl group is then exploited to facilitate the removal of Xc.

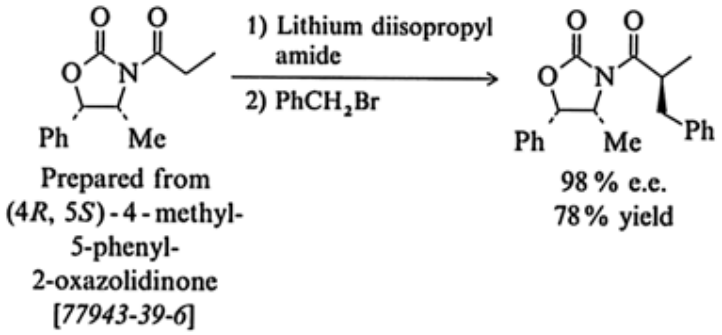

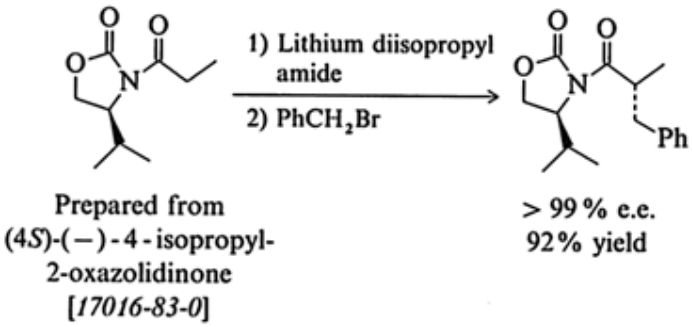

The oxazolidinone chiral auxiliaries introduced by EVANS [24] serve to illustrate these principles.

Systems based on (1*S*,2*R*)-norephedrine [37577-28-9] and (*S*)-valinol [2026-48-4] provide complementary enantioselectivity in a wide range of reactions. Alkylations adjacent to the amide carbonyl proceed with good diastereoselectivity, and the chiral auxiliary may be cleaved with $LiOCH_2Ph$, for example, to yield the corresponding benzyl ester in optically pure form; the chiral auxiliary can then be recovered without loss of optical purity. Oxazolidinones of this type are also used in aldol condensations [31] and, with type **2** substrates, in Diels – Alder cycloadditions [29].

A range of other carbonyl-derived chiral auxiliaries are available for both type **1** and type **2** substrates, and selected examples are shown below [32]:

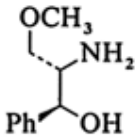

(2*S*, 3*S*)-1-Methoxy-2-amino-3-phenyl-3-hydroxypropane [*51594-34-4*]

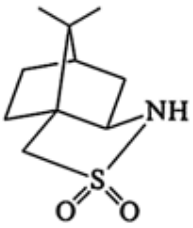

D-2,10-Camphorsultam [*94594-90-8*]

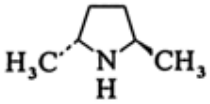

(2*R*, 5*R*)-2,5-Dimethylpyrolidine [*62617-70-3*]

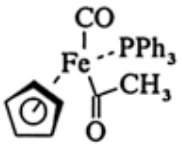

(*R*)-Acetylcarbonyl-(η^5-2,4-cyclopentadien-1-yl) triphenylphosphineiron [*36548-60-4*]

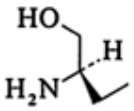

(*R*)-2-Amino-1-butanol [*5856-63-3*]

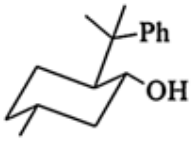

(1*R*)-8-Phenylmenthol [*65253-04-5*]

Constraining a chiral auxiliary with its associated reacting partner into a ring is often an effective means of achieving efficient asymmetric induction. This may be illustrated by the work of SCHÖLLKOPF, who has developed the use of chiral dihydropyrazines for the asymmetric synthesis of amino acids [33]:

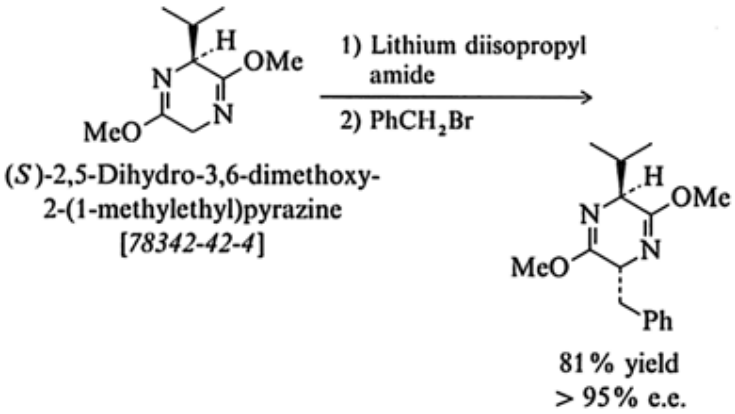

The chiral auxiliary is an amino acid derivative (e.g., (*S*)-valine [72-18-4]), and the presence of the isopropyl residue provides the facial discrimination required for selectivity in alkylation reactions.

Chiral auxiliaries have also found application in the asymmetric synthesis of sulfoxides [34], the alkylation of both cyclic and acyclic amines [35], and 1,3-dipolar cycloaddition reactions [36].

The ideal chiral auxiliary has a low molecular mass, so that unnecessarily large amounts of material are not carried through a sequence only to be eventually removed. An obvious corollary of this is that the cost of an auxiliary should always be calculated on a mole, rather than a gram, basis. However, stoichiometric use of the auxiliary means that for large-scale industrial processes, the approach frequently becomes cumbersome. Further refinements of these methods aim to a achieve similar results under catalytic conditions (see Section Asymmetric Chemical Catalysis).

A different strategy has been developed by SEEBACH [37]. This approach does not strictly involve the use of a chiral auxiliary (which is usually recovered in its original form at the end of a sequence) but is nevertheless a very attractive

means of preparing optically pure compounds. The concept is aptly illustrated using (*R*)-3-hydroxybutanoic acid [625-72-9] as the basis of a general asymmetric synthesis of the corresponding 3-alkyl derivatives.

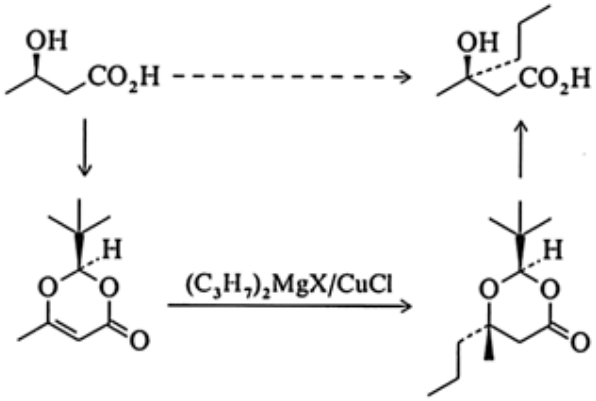

(*R*)-4*H*-2-(1,1-Dimethylethyl)-6-methyl-1,3-dioxin-4-one [*107289-20-3*]

91% yield
Single diastereomer

This process of “self-reproduction of chirality centers” has also been applied to the asymmetric synthesis of substituted amino acids.

3.2.3. *Asymmetric Chemical Catalysis*

A successful catalytic asymmetric transformation must satisfy a number of requirements. The reaction must be mediated or at least accelerated by the presence of the chiral catalyst to ensure that the chiral product is formed in asymmetric environment. The reactive complex must be regenerated and the product released so that further cycles can occur. Finally, the demand for high stereoselectivities is greater than in the case of stoichiometric chiral auxiliaries because the products formed are enantiomers and, therefore, not readily separable.

Arguably, the most important contribution to catalytic asymmetric synthesis in terms of versatility, predictability, and usefulness of products is the asymmetric epoxidation of allylic alcohols developed by SHARPLESS [38]. The transformation is shown in general terms below. The enantiomer of the product epoxide is determined by whether D- or L-tartrate is used and the mechanistic model developed by SHARPLESS has been found to be very reliable. To date no exception to this enantioface selection rule has been found among prochiral substrates.

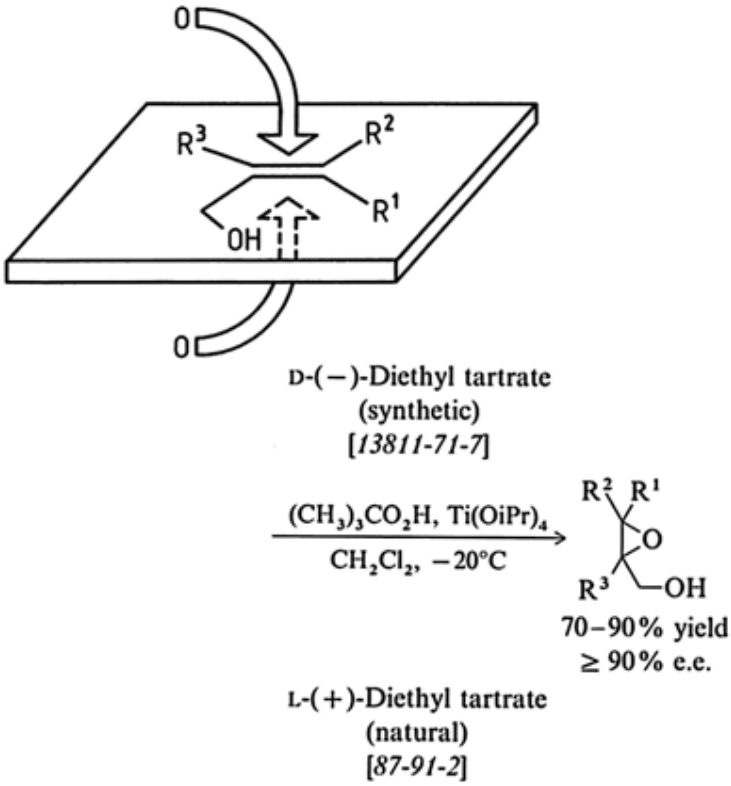

Asymmetric epoxidation provides 2,3-epoxy alcohols in very high optical purity, which are themselves readily manipulated to give a wide range of functionally flexible derivatives. This catalytic process has already been used extensively by synthetic chemists and is also amenable to large-scale work.

Asymmetric hydrogenation is another area that has seen extensive development since the spectacular work by the Monsanto group in 1975. They used a reducing step, that was catalyzed by rhodium complexed to a chiral biphosphine ligand as a key step in the synthesis of L-dopa [59-92-7] [39]. Some impressive results in this area have been achieved by NOYORI [40] using ruthenium complexed to the bulky chiral ligand BINAP, (*R*)- [76189-55-4] and (*S*)- [76189-56-5] (see

below). A range of alkenes with an adjacent polar group (e.g., allylic alcohols) can be hydrogenated with very high stereoselectivity (often > 99 % e.e.) (The concentration ratio of substrate to catalyst is 10 000.)

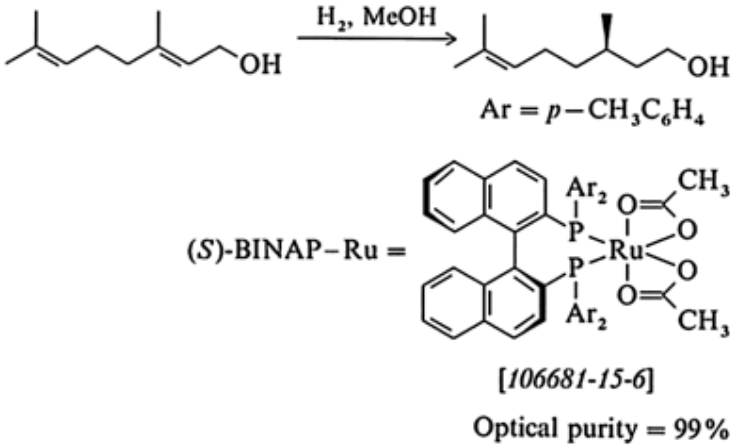

Although a detailed mechanistic interpretation has not emerged, the scope and versatility of this catalytic system is already having a major impacts on organic synthesis.

Rhodium complexes catalyze the isomerization of *allylic amines* to *enamines*, and this reaction has found applications in asymmetric synthesis [41]. A commercial synthesis of (–)-menthol [2216-51-5] is based on the enantioselective isomerization of *N,N*-diethylgeranylamine to citronellal diethylenamine by using rhodium – BINAP.

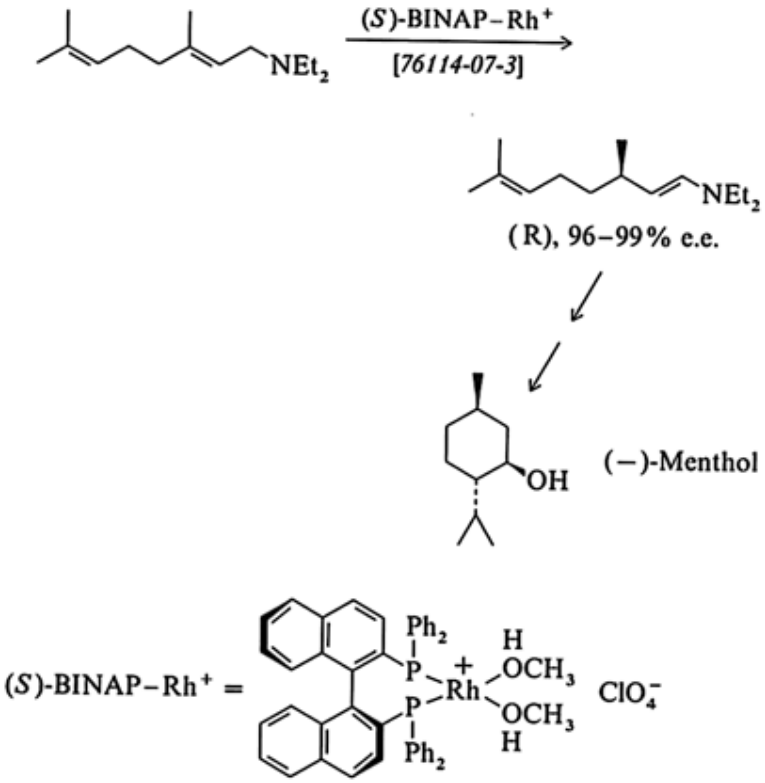

This key step proceeds in 96 – 99 % e.e. and is responsible for a substantial quantity of the world's output of synthetic (–)-menthol.

Another application of asymmetric epoxidation or of reduction using the BINAP catalyst is in the kinetic resolution of a racemic mixture of secondary allylic alcohols [38], [42]. If one enantiomer reacts faster than the other, an optically active product is isolated. The less reactive enantiomer of the allylic alcohol is also recovered in optically active form. The success of a kinetic resolution (the optical enrichment achieved) depends on the relative rates of reaction of a pair of enantiomers. This concept is a powerful one that has also found applications in enzyme-mediated processes.

A number of other transformations have been sucessful in the field of catalytic asymmetric induction. These include the use of proline-derived catalysts in reduction and annulation processes [43], asymmetric aldol reactions mediated by a chiral gold(I) – ferrocenyl complex [44], and asymmetric catalysis in the addition of organozinc reagents to aldehydes and ketones [45]. The Diels – Alder reaction is extremely important in synthesis, and effective asymmetric catalysis for this reaction has been reported [46]. Many of the asymmetric catalyst techniques are still in their infancy and are likely to undergo major refinements in the next few years, particularly in the field of carbon – carbon bond formation.

3.2.4. *Enzymatic Transformations* (⟶ Enzymes, ⟶ Enzymes – Enzymes in Organic Synthesis; ⟶)

Enzymes are particularly valuable chiral catalysts because they can exert complete control over several stereochemical aspects in single-step transformations under mild conditions, while tolerating a wide variety of functionality. As a result of

the high specificity of enzymes, however, the success of an enzyme-mediated reaction often depends on the substrate used. Mammalian enzymes best fit the criteria of stereospecificity coupled with general applicability, whereas microbial enzymes, although more varied in type, usually have narrower structural specificities. For asymmetric synthesis the most important enzymes are oxidoreductases, hydrolases, lyases (catalyzing additions to double bonds), and less generally, ligases (e.g., aldolases). Examples of asymmetric synthesis employing enzymes are given below.

Aspartame [22839-47-0], manufactured by Monsanto, is a highly intensive sweetener (Nutrasweet). The synthesis of aspartame is based on the coupling of L-aspartic acid and L-phenylalanine derivatives, both of which are produced in cultures of microorganisms.

The specificity of enzymes, particularly hydrolases, has been exploited in the large-scale kinetic resolution of *N*-acylamino acids [47]. These enzymes are specific for the hydrolysis of the “natural” L-enantiomer, and the unreactive D-*N*-acylamino acids can then be isolated in optically active form or recycled after chemically induced racemization. Similarly, readily available hydantoinases have been used to resolve racemic hydantoins, hydrolyzing the D-isomers to yield D-amino acids, which are valuable intermediates in the production of semisynthetic penicillins and cephalosporins. The unreactive L-hydantoin can be racemized enzymatically or chemically, and recycled. Other industrial uses of enzymatic resolution include the manufacture of pure D-valine [640-68-6] in the synthesis of fluvalinate [69409-94-5], a broad-spectrum pyrethroid insecticide and (*S*)-homophenylalanine [943-73-7], a useful starting material for angiotensin converting enzyme (ACE) inhibitors.

Alcohol dehydrogenases have been widely used to catalyze the stereospecific reduction of acyclic ketones and β-ketoesters [48]. Yeast reductions are particularly convenient, and models have been suggested that successfully predict the stereoselectivity of these reactions.

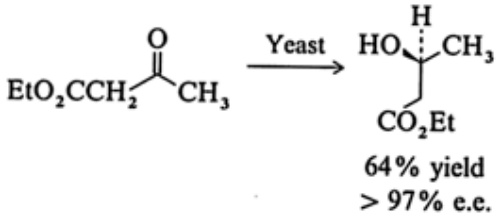

Efforts are currently underway, in the field of genetic tailoring of enzymes and microorganisms, to adapt them for specific transformations. Such developments ensure that enzymatic catalysis will play an increasingly important role in the synthesis of optically active compounds.

[Top of Page]

References

General References

1.

A series of volumes under the title “Asymmetric Synthesis” provides an excellent and comprehensive discussion of modern asymmetric synthesis:

2. J. D. Morrison (ed.): *Analytical Methods*, vol. **1**, Academic Press, Orlando 1983. J. D. Morrison (ed.): *Stereodifferentiating Addition Reactions*, vol. **2, Part A**, Academic Press, Orlando 1983. J. D. Morrison (ed.): *Stereodifferentiating Addition Reactions*, vol. **3, Part B**, Academic Press, Orlando 1984. J. D. Morrison, J. W. Scott (eds.): *The Chiral Carbon Pool and Chiral Sulfur*, Nitrogen, Phosphorus, and Silicon Centers, vol. **4**, Academic Press, Orlando 1984. J. D. Morrison (ed.): *Chiral Catalysis*, vol. **5**, Academic Press, Orlando 1985.

Specific Reference

3. W. A. Bonner, *Top. Stereochem.* **18** (1988) 1. Links
4. J. M. Bijvoet, A. F. Peerdeman, A. J. van Bomnel, *Nature* **168** (1951) 270.
5. R. S. Cahn, C. Ingold, V. Prelog, *Angew. Chem. Int. Ed. Engl.* **5** (1966) 385. Links
6. H. R. Levy, P. Talalay, B. Vennesland in P. B. D. de la Mare, W. Klyne (eds.): *Progress in Stereochemistry*, vol. **II**, Butterworths, London 1962, Chap. 8.
7. K. R. Hanson, *J. Am. Chem. Soc.* **88** (1966) 2731. Links
8. R. J. P. Corriu, C. Guérin, J. J. E. Moreau, *Top. Stereochem.* **15** (1984) 43. Links
9. M. Gielen, *Top. Stereochem.* **12** (1981) 217; *Top. Curr. Chem.* **104** (1982) 57. Links
10. S. J. Brois, *J. Am. Chem. Soc.* **90** (1968) 506, 508. Links
11. V. Prelog, P. Wieland, *Helv. Chim. Acta.* **27** (1944) 1127. Links
12. M. J. Gallagher, I. D. Jenkins, *Top. Stereochem.* **3** (1968) 1. Links
13. M. Oki, *Top. Stereochem.* **14** (1983) 1. Links
14. W. Runge in S. R. Landor (ed.): *The Chemistry of the Allenes*, vol. **3**, Academic Press, New York 1982, pp. 579 – 678.
15. A. C. Cope *et al.*, *J. Am. Chem. Soc.* **85** (1963) 3276. Links
16. K. Schlögl, *Pure Appl. Chem.* **23** (1970) 413.
17. G. R. Sullivan, J. A. Dale, H. S. Mosher, *J. Org. Chem.* **38** (1973) 2143. Links
18. J. W. ApSimon, T. L. Collier, *Tetrahedron* **42** (1986) 5157. Links
19. R. Noyori, *Pure Appl. Chem.* **53** (1981) 2315. Links

20. M. M. Midland, D. C. McDowell, R. L. Hatch, A. Tramontano, *J. Am. Chem. Soc.* **102** (1980) 867. Links
21. H. C. Brown, M. C. Desai, P. K. Jadhav, *J. Org. Chem.* **47** (1982) 5065. Links
22. R. Shirai, M. Tanaka, K. Koga, *J. Am. Chem. Soc.* **108** (1986) 543. Links
23. L. Duhamel, A. Ravard, J-C. Plaquevent, D. Davoust, *Tetrahedron Lett.* **28** (1987) 5517. Links
24. D. A. Evans, M. D. Ennis, D. J. Mathre, *J. Am. Chem. Soc.* **104** (1982) 1737. Links
25. A. I. Meyers *et al.*, *J. Am. Chem. Soc.* **103** (1981) 3081. Links
26. J. d'Angelo *et al.*, *Tetrahedron Lett.* **24** (1983) 5869. Links
27. T. Mukaiyama, N. Iwasawa, *Chem. Lett.* 1981, 913. Links T. Imamoto, T. Mukaiyama, *Chem. Lett.* 1980, 45. Links G. H. Posner, J. P. Mallamo, M. Huke, T. L. Frye, *J. Am. Chem. Soc.* **104** (1982) 4180. Links W. Oppolzer, H. J. Löher, *Helv. Chim. Acta* **64** (1981) 2808. Links
28. P. W. Ambler, S. G. Davies, *Tetrahedron Lett.* **29** (1988) 6979, 6983.
29. D. A. Evans, K. T. Chapman, J. Bisaha, *J. Am. Chem. Soc.* **110** (1988) 1238. Links
30. W. Oppolzer, *Angew. Chem. Int. Ed. Engl.* **23** (1984) 876. Links
31. D. A. Evans, *Aldrichimica Acta* **15** (1982) 23. Links
32. A. I. Meyers, K. A. Lutomski, *J. Am. Chem. Soc.* **104** (1982) 879. Links W. Oppolzer, C. Chapuis, G. Bernardinelli, *Helv. Chim. Acta* **67** (1984) 1397. Links R. P. Short, R. M. Kennedy, S. Masamune, *J. Org. Chem.* **54** (1989) 1755. Links S. G. Davies, *Pure Appl. Chem.* **60** (1988) 13. Links J. K. Whitesell, M. A. Whitesell, *J. Org. Chem.* **42** (1977) 377. Links E. J. Corey, H. E. Ensley, *J. Am. Chem. Soc.* **93** (1975) 6908.
33. U. Schöllkopf, U. Groth, C. Deng, *Angew. Chem. Int. Ed. Engl.* **20** (1981) 798.
34. K. Hiroi, R. Kitayama, S. Sato, *Heterocycles* **15** (1981) 879. Links
35. A. I. Meyers, M. Boes, D. A. Dickman, *Angew. Chem. Int. Ed. Engl.* **23** (1984) 458.
36. A. P. Kozikowski, A. K. Ghosh, *J. Am. Chem Soc.* **104** (1982) 5788. Links T. Koizumi, H. Hirai, E. Yoshii, *J. Org. Chem.* **47** (1982) 4004. Links D. P. Curran *et al.*, *J. Org. Chem.* **52** (1987) 2137. Links
37. D. Seebach *et al.*, *J. Am. Chem. Soc.* **110** (1988) 4763. Links D. Seebach, U. Misslitz, P. Ullmann, *Angew. Chem. Int. Ed. Engl.* **28** (1989) 472.
38. M. G. Finn, K. B. Sharples in J. D. Morrison (ed.): *Asymmetric Synthesis*, vol. **5**, Academic Press, Orlando 1985, Chap. 8.
39. W. S. Knowles, M. J. Sabacky, B. D. Vineyard, D. J. Weinkauff, *J. Am. Chem. Soc.* **97** (1975) 2567. Links
40. H. Takaya *et al.*, *J. Am. Chem. Soc.* **109** (1987) 1596. Links
41. R. Noyori, *Chem. Soc. Rev.* **18** (1989) 187. Links
42. M. Kitamura *et al.*, *J. Org. Chem.* **53** (1988) 708. Links
43. E. J. Corey *et al.*, *J. Am. Chem. Soc.* **109** (1987) 7925. Links E. J. Corey, R. K. Bakshi, S. Shibata, *J. Am. Chem. Soc.* **109** (1987) 5551. Links
44. Y. Ito, M. Sawamura, T. Hayashi, *J. Am. Chem. Soc.* **108** (1986) 6405. Links
45. K. Soai, A. Ookawa, T. Kaba, K. Ogawa, *J. Am. Chem. Soc.* **109** (1987) 7111. Links
46. K. Maruoka, T. Itoh, T. Shirasaka, H. Yamamoto, *J. Am. Chem. Soc.* **110** (1988) 310. Links
47. I. Chibata, T. Tosa, T. Sato, T. Mori, *Methods Enzymol.* **44** (1976) 746.
48. A. I. Meyers, R. A. Amos, *J. Am. Chem. Soc.* **102** (1980) 870.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Optical Materials

Fiber Optics, Glass, Holography, Information Storage Materials, Lasers, Mirrors, Nonlinear Optics, and Photography are separate keywords.

CHRISTIAN HOFMANN, Carl Zeiss Jena GmbH, Jena, Federal Republic of Germany

1. Introduction 1
2. Optical Properties of Optical Materials 1
2.1. Refraction and Dispersion 1
2.2. Interfacial Properties 3
2.3. Birefringence 4
2.4. Absorption and Scattering 6
3. Effect of Optical Material on Imaging 7
3.1. Effect of Refractive Index on Optical Imaging 8
3.2. Effect of Dispersion on Optical Imaging 9
3.3. Effect of Partial Dispersion on Optical Imaging 11
3.4. Effect of Transmittance on Optical Imaging 11
4. Optical Glasses 12
5. Colored or Filter Glasses 14
6. Optical Crystals 14
7. Optical Plastics 14
8. Manufacturing Techniques 16
8.1. Processing of Optical Glasses 16
8.2. Casting of Optical Thermosets 17
8.3. Processing of Thermoplastic Optical Components 18
9. References 18

1. Introduction

In this article optical materials are considered to be transparent solids (e.g., glasses, crystals, and polymers) from which transparent optical components such as lenses, optical flats, and prisms are fabricated. Immersion fluids and optical cements are considered to be special optical materials.

Optical materials influence the light passing through them resulting in refraction, dispersion, and absorption. Light incident on an interface of an optical material may be reflected and/or transmitted, depending on wavelength and on the directions of propagation and polarization. The direction of light passing through an interface changes as a result of refraction, which is a function of the angle of incidence. In the extreme case, light experiences total reflection. Optically anisotropic materials influence the polarization of light through birefringence.

2. Optical Properties of Optical Materials

2.1. Refraction and Dispersion

Dispersion is the result of interaction between light and the atoms in optical materials. The light excites vibrations of valence electrons (dispersion electrons). At resonance, the dispersion electrons reach a metastable excited state. Light is absorbed and its energy is transferred to the electrons. If the frequency ω of the light does not correspond to the frequency ω_{e} of the electron transition, the electrons are excited and undergo phase-shifted polarization vibrations. These dipole oscillators emit secondary waves having the same frequency as the exciting radiation but shifted in phase. The phase velocity $c_{\mathrm{n}}(\omega)$ of the light thus differs from the velocity of light in vacuum c, it depends on the difference $\omega_{\mathrm{e}}^2 - \omega^2$ and the density of the optical material. The ratio of these velocities is called the *refractive index* of the optical medium [1], [2]:

$$n(\omega) = \frac{c}{c_{\mathrm{n}}(\omega)} \tag{1}$$

If the refractive index of a material is independent of position and direction, the material is said to be *homogeneous* and *isotropic*. Homogeneous, isotropic optical materials usually have to be employed for optical components, in particular those in imaging systems. In *anisotropic* optical materials (e.g., birefringent crystals, glasses and plastics under stress), the refractive index depends on the directions of vibration and propagation of the light. The

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_191

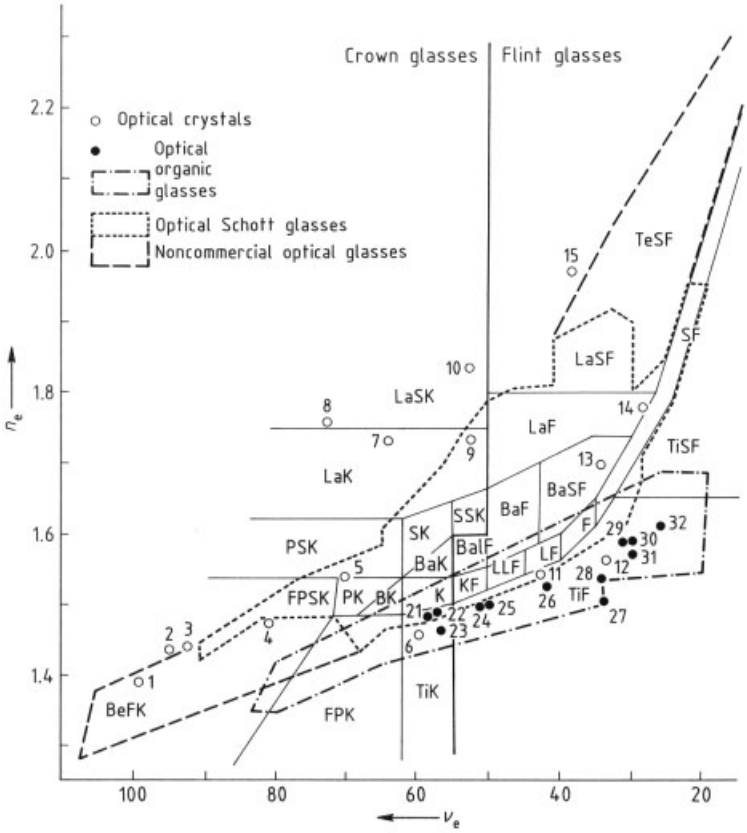

Figure 1. Locations of optical materials in the $n-\nu$ diagram (for explanation of symbols, see Chap. 4)
Crystals: 1) LiF; 2) CaF_2; 3) SrF_2; 4) BaF_2; 5) Quartz; 6) Potassium alum; 7) Spinel; 8) Sapphire; 9) MgO; 10) Yttrium aluminum garnet (YAG); 11) NaCl; 12) KBr; 13) CsBr; 14) PbF_2; 15) Gadolinium gallium garnet (GGG)
Optical polymers: 21) CR 39 (copolymer of diethylene glycol and bisallyl carbonate); 22) Poly(methyl methacrylate) (PMMA); 23) Polymethylpentene (TPX); 24) Cyclohexyl methacrylate – methyl methacrylate copolymer (CMMA); 25) Cellulose acetate; 26) Cyclohexylmethyl methacrylate – styrene copolymer (CMMAS); 27) Polytrifluoromethylstyrene (FMS); 28) Methyl methacrylate – styrene copolymer (NAS); 29) Styrene – maleic acid copolymer (SMA); 30) Polystyrene; 31) Bisphenol A – phosgene polycarbonate; 32) DPSC (copolymer of bisphenol and dihydroxyphenyl sulfide)

birefringence of anisotropic materials is utilized in polarization optics. In *inhomogeneous* optical materials the refractive index varies with position.

The refractive index depends on the frequency and thus the wavelength as is approximately described by Cauchy's dispersion formula

$$n^2 = 1 + A\left(1 + \frac{B}{\lambda^2}\right) \tag{2}$$

where A and B are the refraction and dispersion constants and λ is the wavelength. Because B is small in comparison with A, the wavelength dependence of the refractive index can be neglected for estimation purposes. The phase shifts associated with the wavelength-independent part of the refractive index and the resulting effects (particularly the change in propagation direction when light passes through interfaces of optical materials) are termed *refraction*.

The frequency dependence of the refractive index is called *dispersion*. It is generally defined as the difference between the refractive indices $n(\lambda_1)$ and $n(\lambda_2)$ at two wavelengths:

$$\Delta_\lambda n = n(\lambda_1) - n(\lambda_2) \tag{4}$$

Optical materials are usually characterized by the *principal refractive index* n_e [the refractive index for the Fraunhofer line of mercury, $\lambda_e = 546.1$ nm (green)], and the *mean dispersion* $n_{F'} - n_{C'}$ [the subscripts F′ and C′ refer to

the blue ($\lambda_{F'}$= 480 nm) and red ($\lambda_{C'}$= 643.8 nm) lines of cadmium, respectively].

Another important parameter for characterizing optical materials is the *Abbe number:*

$$\nu_e = \frac{n_e - 1}{n_{F'} - n_{C'}} \tag{5}$$

Along with the principal refractive index, the Abbe number determines the location of the material in the $n-\nu$ diagram of Figure 1.

Optical materials are classified as being analogous tocrown glasses (ν_e>55) or analogous to flint glasses (ν_e<50). Optical materials with 55>ν_e>50 and n_e>1.6 are placed in the crown group (Fig. 1).

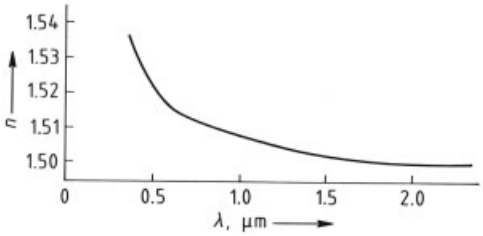

Figure 2. Variation of refractive index with wavelength for Schott BK7 glass

Most optical materials analogous to crown glasses have low refractive indices and dispersions; materials analogous to flint glasses have high refractive indices and dispersions. In conventional optical materials, the mean dispersion is approximately a linear function of the principal refractive index, as described by the law of the iron-clad line:

$$\Delta_\lambda n = Pn + Q \tag{6}$$

This formula applies to most optical glasses. In the dispersion – refractive index diagram, all optical glasses known up to 1884 lie on the iron line [3], [4].

Because the refractive index is a nonlinear function of wavelength (Eqs. 2 and 3, Fig. 2) the *partial dispersions* for stated pairs of wavelengths must be known in order to characterize dispersion in defined spectral regions. The relative partial dispersion

$$\theta = \frac{\Delta_{\lambda_p} n}{n_{F'} - n_{C'}} \tag{7}$$

is a linear function of the Abbe number for most optical glasses:

$$\theta_n = a + b\nu \tag{8}$$

The correction of the secondary spectrum calls for optical materials with relative partial dispersions that differ greatly from the values predicted by the normal straight line (Eq. 8, see Section 3.3). Materials with large anomalous partial dispersions include fluorspar, barium fluoride, potassium alum, and the dense fluorophosphate crown glasses.

2.2. Interfacial Properties

When light travels from one medium to another it is split into a reflected part I″ and a transmitted part I′. In accordance with Fresnel's formulas [2], reflectance and transmittance depend on the wavelength, the polarization and incident propagation directions, and the refractive indices of the optical material and the surroundings. The intensity ratios are called the *spectral reflectance*

$$\varrho(\lambda) = \frac{I''}{I_o} \tag{10}$$

and the *spectral transmittance*

$$\tau_F(\lambda) = \frac{I'}{I_o} \tag{11}$$

where I_o is the intensity of the incident light (the subscript F stands for Fresnel). In the case of normal (i.e., 90°) incidence, the spectral reflectance and transmittance are independent of polarization:

$$\varrho\left(\lambda, \frac{\pi}{2}\right) = \left(\frac{n' - n}{n' + n}\right)^2 \tag{10a}$$

$$\tau_F\left(\lambda, \frac{\pi}{2}\right) = \frac{4n'n}{(n' + n)^2} \tag{11a}$$

where n and n' are the refractive indices in front of and behind the interface.

The reflected light is described by the *law of reflection*

$$\varepsilon'' = -\varepsilon \tag{12}$$

which states that the angle of incidence ε equals the angle of reflection ε'', both angles being measured from the normal to the interface

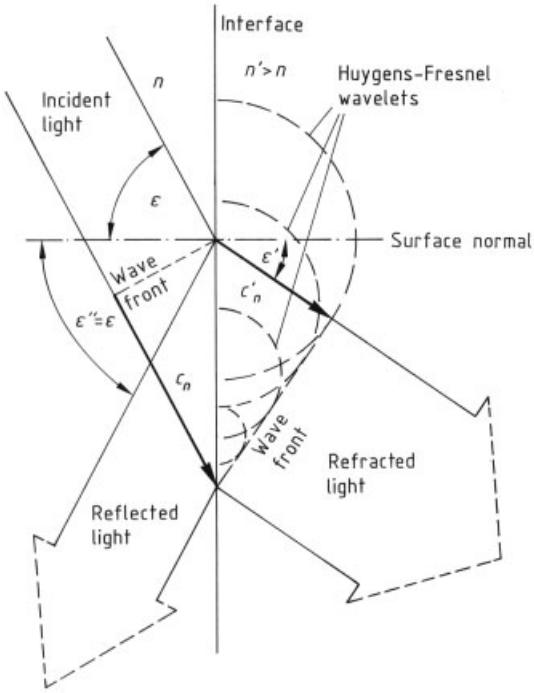

Figure 3. Change in direction of propagation due to change in phase velocity for light obliquely incident on an interface (Snell's law)
c_n and c'_n phase velocities in the two media; ε = angle of incidence; ε' = angle of refraction; ε'' = angle of reflection

(Fig. 3). The incident ray, the reflected ray, and the normal lie in the same plane.

The light passing through the interface is refracted. The difference in phase velocity on either side of the interface leads to a change in the direction of propagation for oblique incidence, as described by *Snell's law*

$$n' \sin\varepsilon' = n \sin\varepsilon \tag{13}$$

where ε' is the angle of refraction. The refracted ray lies in the incidence plane (the plane containing the incident and reflected rays and the surface normal). In optically denser media ($n' > n$), the ray is refracted toward the surface normal ($\varepsilon' < \varepsilon$); in optically less dense media ($n' < n$), $\varepsilon' > \varepsilon$ (Fig. 4 A). At the critical angle ε_g the angle of refraction is 90° ($\varepsilon' = 90°$), i.e., it propagates at a grazing angle in the interface (Fig. 4 B). If the incidence angle exceeds the critical angle (Fig. 4 C)

$$\varepsilon_g = \arcsin \frac{n'}{n} \tag{14}$$

then total reflection occurs (Fig. 4 C).

Total reflection is utilized for light deflection in prisms, and mirrors → Mirrors; refraction forms the basis for optical imaging.

2.3. Birefringence

In anisotropic optical media the lattice structure may lack central symmetry (e.g., in noncubic crystals) or there may be internal and external stresses. As a result the phase velocity and thus the refractive index depend on the polarization and propagation directions of the light. A light wave incident on an optically anisotropic material is generally split into two orthogonally polarized partial waves, which propagate at different phase velocities. In the case of oblique incidence, the two partial waves have different angles of refraction, and thus different refractive indices. This phenomenon is called *birefringence* or *double refraction.*

Birefringence is only absent in defined directions termed the optical (crystal) axes. Tetragonal, hexagonal, and rhombohedral crystals (e.g.,

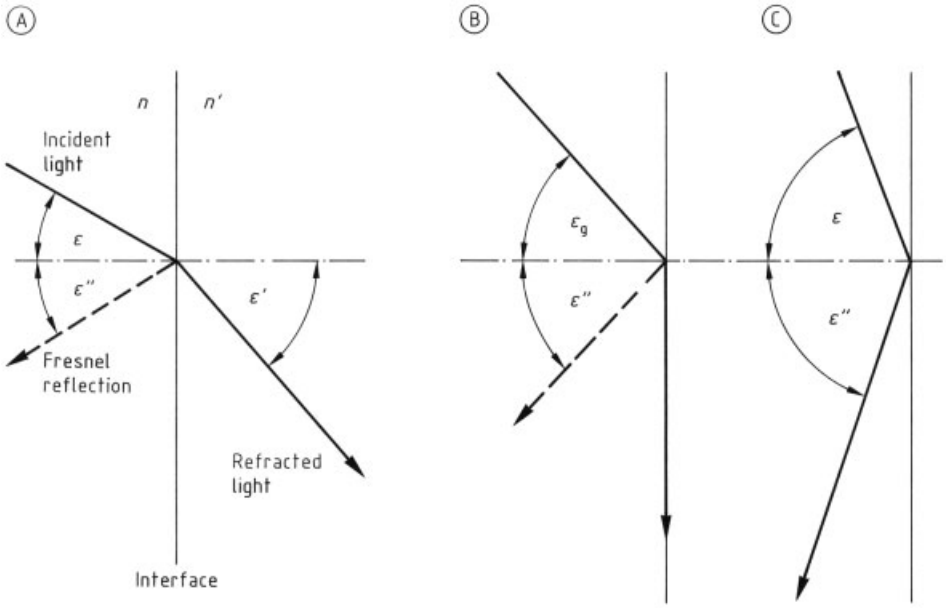

Figure 4. Reflection of obliquely incident light at varying angles of incidence (ε) for two media with refractive indices n and n' ($n > n'$)
A) $\varepsilon < \varepsilon_g$; B) $\varepsilon = \varepsilon_g$; C) Total reflection $\varepsilon > \varepsilon_g$
ε_g = critical angle; ε' = angle of refraction; ε'' = angle of reflection

quartz and calcite) are optically uniaxial; monoclinic, triclinic, and rhombic crystals are optically biaxial. Optically uniaxial crystals are generally of importance in polarization optics.

Any plane that corresponds to an optical axis of a crystal is termed the *principal section (plane)* of the crystal. In optically uniaxial crystals, the electric field of the extraordinary ray vibrates in the principal section. The extraordinary ray has a direction-dependent phase velocity that depends on the direction-dependent extraordinary refractive index $n_{eo}(\varphi)$. The electric field vibration of the ordinary ray is orthogonal to the principal plane section. The phase velocity and the ordinary refractive index n_o of the ordinary ray are independent of direction.

The extraordinary refractive index $n_{eo}(\varphi)$ has an extreme value $n_{eo}(\pi/2)$ perpendicular to the optical axis of the crystal. In the case of negative birefringence (as in calcite with n_o=1.662 and $n_{eo}(\pi/2)$=1.488), this extreme value is a minimum. In the case of positive birefringence (e.g., quartz, n_o=1.546, $n_{eo}(\pi/2)$=1.555), it is a maximum. Birefringent materials are used in double-image polarizing prisms and in delay or phase plates. In prism-type polarizers, such as the Nicol prism (Fig. 5 A), linearly polarized light is obtained from unpolarized light by using total reflection to trap the ordinary ray. In polarization-type beam splitters, such as the Wollaston prism (Fig. 5 B), the two orthogonally polarized beams leave the device in different directions. Delay or phase plates are used to produce elliptically and circularly polarized light. They consist of birefringent plates made of uniaxial quartz or of biaxial mica or gypsum in which the parallel faces lie parallel to a principal section of the crystal. If the light is normally incident, the extraordinary and ordinary rays differ only in the distance traveled. The path difference $d\,[n_{eo}(\pi/2) - n_o]$ depends on the thickness d of the plate and the maximum birefringence $n_{eo}(\pi/2) - n_o$. In a $\lambda/4$ (quarter-wave) plate, the path difference is $(2m+1)\,\lambda/4$ and the phase difference is $(2m+1)\,\pi/2$. The plate thickness is given by

$$d = \frac{(2m+1)}{n_{eo}\left(\frac{\pi}{2}\right) - n_o} \cdot \frac{\lambda}{4} \tag{15}$$

with $m = 0, 1, 2, \ldots$

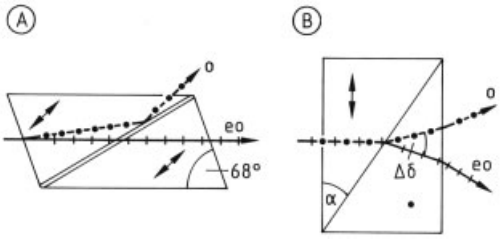

Figure 5. Polarizing crystal prisms
A) Nicol calcite prism used to obtain linearly polarized light;
B) Wollaston prism of calcite or quartz used as a beam splitter
⟷ Optical axis of crystal in drawing plane
• Optical axis perpendicular to drawing plane
———→ Polarization in drawing plane
–•–• → Polarization perpendicular to drawing plane

2.4. Absorption and Scattering

An optical material only transmits light in certain portions of the spectrum. Even in these "spectral windows," some of the light entering the material is lost. Absorption of light depends on the nature of the atoms in the material, their bonding, the wavelength of the light, and the path length d traveled by the light.

The ratio of absorbed intensity $\Delta I'_{\mathrm{A}}$ to incident intensity I_0 is called the *spectral absorptance:*

$$\alpha(\lambda) = \frac{\Delta I'_{\mathrm{A}}}{I_0} \tag{16}$$

The absorbed intensity $\Delta I'_{\mathrm{A}}$ divided by the intensity I' passing through the boundary of the material (adjusted for Fresnel reflection losses I'') is called the *internal spectral absorptance*:

$$\alpha_{\mathrm{i}}(\lambda) = \frac{\Delta I'_{\mathrm{A}}}{I'} = \frac{\alpha(\lambda)}{1-\varrho} \tag{17}$$

The *internal spectral transmittance* is given by

$$\tau_{\mathrm{i}}(\lambda) = 1-\alpha_{\mathrm{i}}(\lambda) = \frac{I'-\Delta I'_{\mathrm{A}}}{I'} \tag{18}$$

In homogeneous media, it is described by the *Bouguer–Lambert law*

$$\tau_{\mathrm{i}}(\lambda) = \mathrm{e}^{-Kd} \tag{18a}$$

which states that the absorption intensity increases with the thickness d of the material. The absorption coefficient

$$K = \frac{4\pi}{\lambda} n\kappa \tag{19}$$

is a function of the absorption constant $n\,\kappa$ and thus of the absorption index κ. The absorption constant is the imaginary part of the complex refractive index

$$\tilde{n} = n(1-i\kappa) \tag{20}$$

of an absorbing medium, where i is the imaginary unit. The reciprocal of the absorption coefficient gives the mean range, which is the path length after which the intensity is reduced by a factor of (1−1/e).

A reflection factor can be defined which describes transmission through both bounding surfaces taking into consideration weak multiple reflection and absorption:

$$\frac{2n}{n^2+1}$$

The formula is only strictly valid for normal incidence. The *spectral transmission factor* is defined as

$$\tau(\lambda) = \frac{2n}{n^2+1} \cdot \tau_{\mathrm{i}}(\lambda) \tag{21}$$

It describes the ratio of light intensity leaving an optical material to that entering the material.

If *scattering losses* also occur in the material, absorption and scattering are added together and referred to as *extinction* or *optical density*. The sum of the absorption coefficient and the analogous scattering coefficient K_{S} is the extinction coefficient:

$$K_{\mathrm{E}} = K + K_{\mathrm{S}} \tag{22}$$

Similarly to Equations (18 a) and (19), the extinction, extinction constant, and extinction index can be defined in terms of K_{E}.

Nonselective internal transmittance is generally desired in optical materials, but no material has this property [5]. Most of the energetically more favorable crystals (e.g., LiF, CaF_2, BaF_2) have their ultraviolet absorption limit at a shorter wavelength than the metastable glasses (Fig. 6) [5].

The dependence of absorption on the propagation and polarization directions, observed in colored birefringent crystals, is called *pleochroism.* Optically uniaxial crystals (e.g., tourmaline

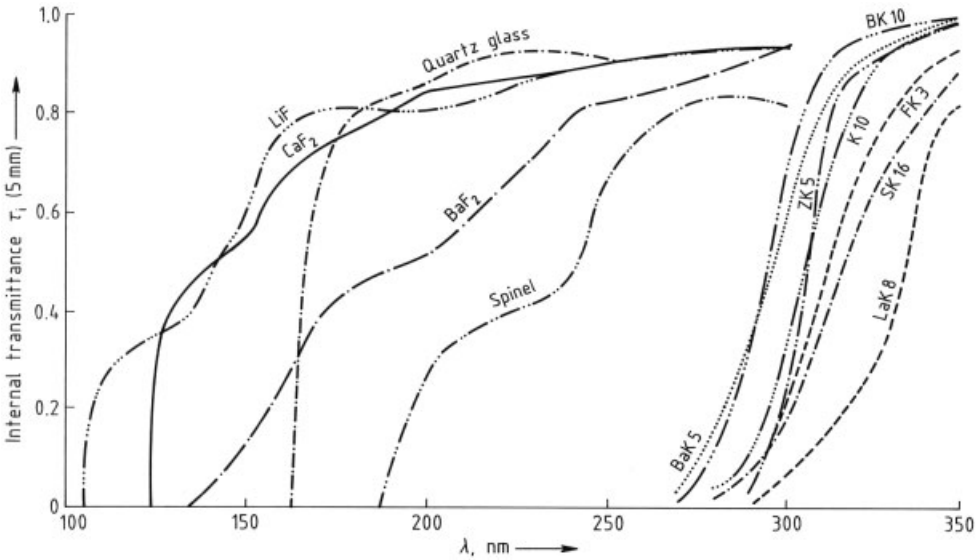

Figure 6. Ultraviolet internal spectral transmission τ_i of optical glasses (Schott, Mainz) and crystals for a 5 mm path (λ = wavelength)

and penninite) exhibit *dichroism* with different absorption constants for two orthogonal polarization directions. By choosing an appropriate material thickness one component can be attenuated to such a degree that the crystalline plate can be used as a dichroic polarizer. Polarizers can also be fabricated by dyeing polyvinyl films.

3. Effect of Optical Material on Imaging

Refraction of light at appropriately shaped boundaries of optical materials is the most important principle employed in conventional optics. It is exploited in lenses and lens systems which produce an optical image by focusing the light from the individual points of an object at the respective conjugate image points.

Lenses are generally cylindrical bodies with spherical refracting surfaces at which light is deflected in accordance with Snell's law (Eq. 13). A lens is characterized in terms of the refractive index n, its central thickness d, and the radii r_1 and r_2 of the refracting surfaces. A radius is positive if the center of curvature C is to the right of the vertex of the corresponding surface, or negative if the center is to the left of the vertex. The line joining the centers of curvature C_1 and C_2 is the optical axis of the lens, which should also be the axis of the cylinder bounding the lens. In drawings, only the meridional section containing the optical axis is shown (Fig. 7).

The action of a lens can be described in a zero-order approximation by a model of two principal planes orthogonal to the optical axis. The lens images these planes on one another at a scale of 1 : 1. Ray deflection is assumed to take place at these planes. The principal points H and H′ lie at the intersection of the optical axis with the principal planes. The distances between the principal points and the points S_1 and S_2 at which the lens surfaces intersect the optical axis are given by

$$s_{H'} = S_2H' = \frac{r_2 d}{n(r_1 - r_2) - d(n-1)} \tag{23a}$$

$$s_H = S_1H = \frac{r_1 d}{n(r_1 - r_2) - d(n-1)} \tag{23b}$$

The light-deflecting action of a lens is determined by the focal length f'. Parallel bundles of rays are approximately focused on the back (image) focal plane, at a distance f' from the back principal plane. The focal point F′ is the point at which the optical axis meets the focal plane. The reciprocal of the focal length is the focal power or strength of the lens:

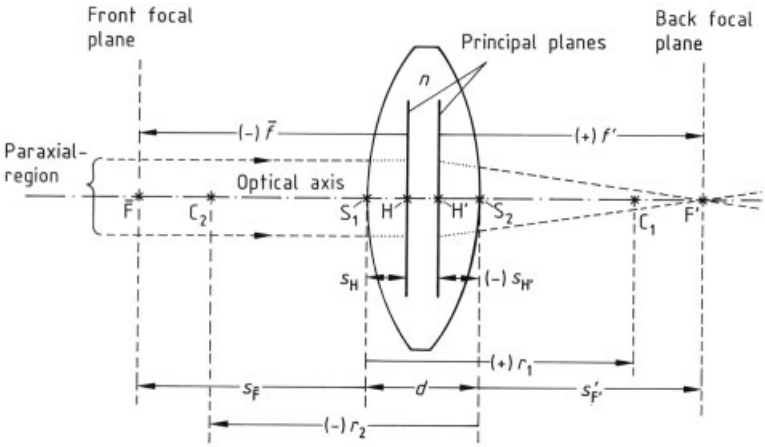

Figure 7. Design parameters, principal points, and focal points of a spherical lens
C = center of curvature; d = central thickness; $\overline{f}$ = front focal length; f' = back focal length; $\overline{\mathrm{F}}$ and F′ = front and back focal points; H and H′ = first and second principle points; r = radius of curvature; $s_{\overline{\mathrm{F}}}$ = front focal distance (from the front lens surface to the first principle focus); $s'_{\mathrm{F}'}$ = back focal distance (from the back lens surface to the back principle focus)

$$D' = \frac{1}{f'} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2} + \frac{n-1}{n}\cdot\frac{d}{r_1 r_2}\right) \tag{24}$$

which serves as a measure of the deflection of a ray. For thin lenses, the third term in the second set of parentheses in Equation (24) can be neglected [2]:

$$D' = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{24a}$$

The refractive power can be expressed in terms of two distances a and a'; a is the distance from the object to the front or object principal plane; a' is the distance from the image plane to the back principal plane. The Gaussian lens formula

$$\frac{1}{a'} - \frac{1}{a} = D' \tag{25}$$

is only strictly valid for small incidence angles at the refracting surfaces: $\varepsilon \approx \sin\varepsilon \approx \tan\varepsilon$ (i.e., for rays in the immediate vicinity of the optical axis). Outside this paraxial region, the paths of rays through the lens do not match the paths predicted by Equation (25) (Fig. 8 A). Such aberrations prevent the ideal convergence of ray bundles at the image points [2]. They can be minimized by making the refracting surfaces flatter to decrease the angle of incidence.

3.1. Effect of Refractive Index on Optical Imaging

The decrease in focal power that occurs in less strongly curved lenses (Eq. 24 a) can be compensated by the use of highly refractive materials. Such highly refractive lenses are thinner and have smaller aberrations than lenses made of normally refractive material (Fig. 8 B) [5]. To achieve a given correction without a highly refractive material, several less strongly curved lenses with the same total focal power must be combined (Fig. 8 C). The use of highly refractive material allows the number of lens elements in an optical system to be reduced; this is particularly important for photographic objectives.

Lenses with a high focal power and a large diameter can be made only from high-refraction material, since otherwise the strongly curved surfaces required by Equation (24 a) can lead to total reflection [5].

Combinations of high- and low-refraction materials used for correction purposes are especially effective in cemented and wringing components because they permit the use of less strongly curved mating surfaces. Low-index optical materials are needed for weak diverging (concave) lenses that are used in optical systems because their strongly overcorrecting action (in relation to the focal power) can offset the aberrations produced by converging (convex) lenses.

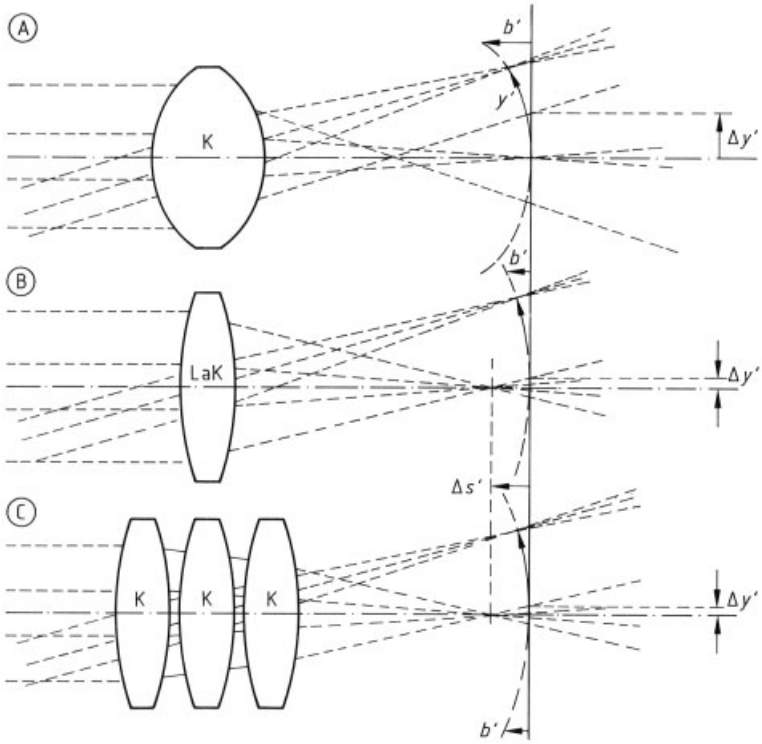

Figure 8. Aberrations of (A) a strongly curved lens made of a normally refractive material ($n = 1.5$), (B) a less strongly curved lens made of a highly refractive material ($n \approx 10$), and (C) a system consisting of several less strongly curved lenses made of a normally refractive material
b' = curvature of field; $\Delta y'$ = transverse aberration

3.2. Effect of Dispersion on Optical Imaging

Since the focal power of a lens depends on the refractive index (Eq. 24) and the refractive index is wavelength dependent, the focal lengths and the position of the focal plane or image plane also depend on wavelength. These variations with wavelength are referred to as *chromatic aberrations* [2].

The chromatic aberration of the focal length $\Delta_\lambda f'$ produces a chromatic variation of the image size $\Delta_\lambda y'$. This is called *transverse chromatic aberration* or color magnification error (Fig. 9).

The locations of the principal planes also depend on the refractive index and thus the dispersion (Eqs. 23 a, 23 b). With thick lenses, a distinction must be made between the chromatic aberration of the focal length $\Delta_\lambda f'$ and the chromatic variation of the paraxial image position $\Delta_\lambda s'$ (Fig. 9 A), also called *longitudinal chromatic aberration.* The latter term refers to the wavelength dependence of the image-plane position; it represents the distance between the image planes for the two extreme wavelengths of a useful spectral region $\Delta\lambda$. The longitudinal chromatic aberration of a thin simple lens is given by Equation (26).

$$\Delta_\lambda s' = \frac{-f'}{\nu} \tag{26}$$

It can be corrected for two wavelengths by cementing together a converging lens with focal length $f'_C > 0$, made of a crown-glass analogue (Abbe number $\nu_C > 55$) and a diverging lens with focal length

$$f'_F = -\frac{\nu_C}{\nu_F} f' = \left(1 - \frac{\nu_C}{\nu_F}\right) f' \tag{27}$$

made of a flint-glass analogue (Abbe number $\nu_F < 50$). Here f' is the overall focal length of the resulting achromat (Fig. 9 B).

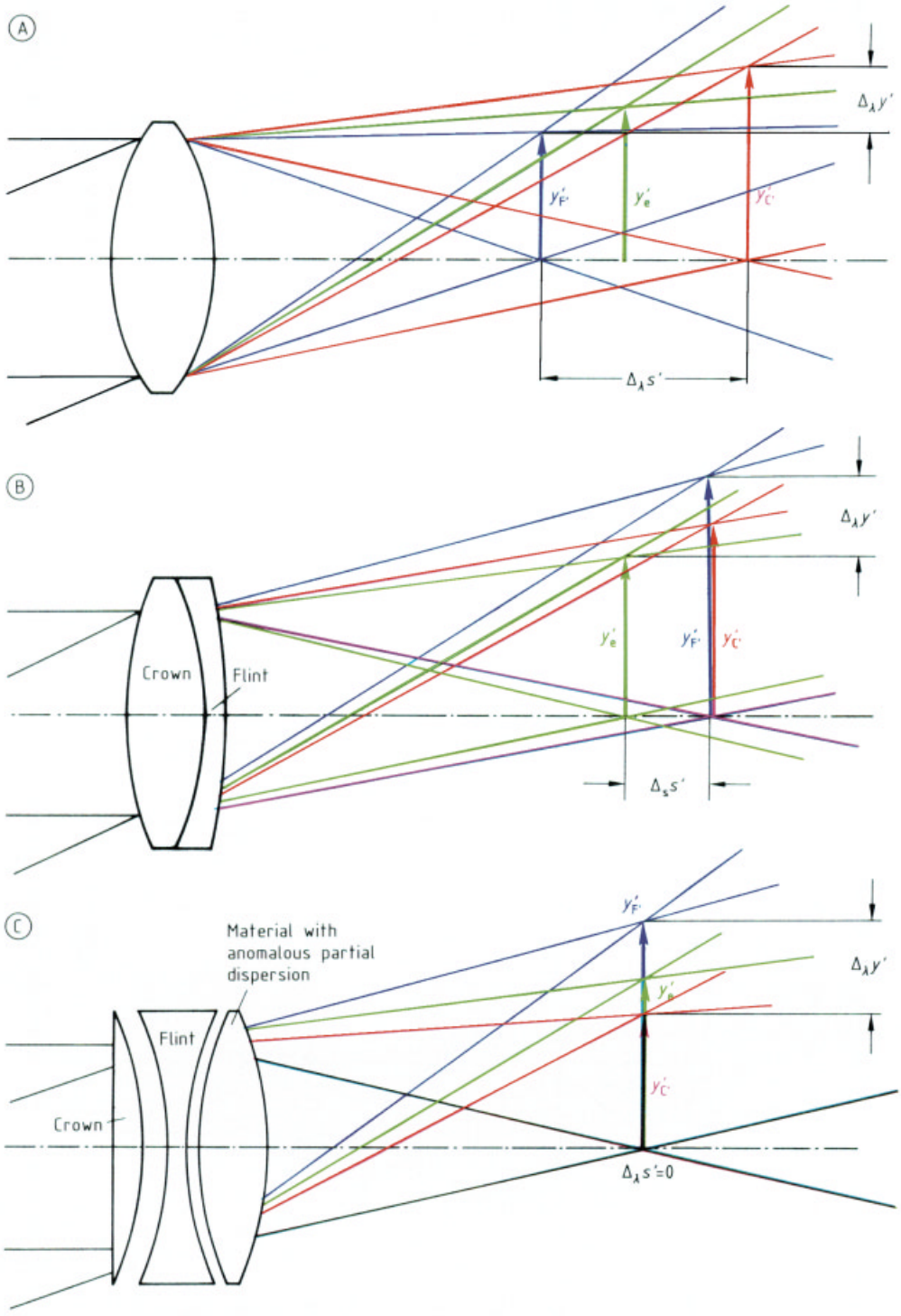

Figure 9. Chromatic aberrations of (A) a single lens, (B) an achromat, and (C) a trichromat, shown as ray diagrams
$\Delta_\lambda y'$ = transverse chromatic aberration; $\Delta_\lambda s'$ = longitudinal chromatic aberration; $\Delta_s s'$ = secondary spectrum; λ = wavelength

In order to correct the monochromatic aberrations of the achromat, in particular the spherical aberration (Fig. 10), f'_C must be much greater than $-f'_F$. This means that the Abbe numbers must differ by at least 25. The *spherical aberration* (also called aperture aberration) denotes the difference between the point where the ray actually passes through the image plane and the ideal image point. This difference increases with the incidence height of the ray on the lens. A diffuse circle is produced instead of the ideal image point. Because the spherical aberrations of a converging lens and a diverging lens are opposite in sense, they can largely compensate one another when the two lenses are cemented together. Such a lens unit, called an *achromat* (or achromatic lens) consists of a low-index crown-glass analogue with low dispersion and a high-index flint-glass analogue with high dispersion; it corrects both spherical aberration and longitudinal chromatic aberration. The lens still has a residual error called *spherochromatism*, which is the chromatic variation of the spherical aberration (Fig. 11). This can be corrected with a new achromat consisting of a high-index material with low dispersion and a low-index material with high dispersion. Since, however, the new achromat also has a spherical aberration, the old achromat is combined with the new achromat to correct both the spherical aberration and spherochromatism. Combinations of this type are called *anastigmats* [2].

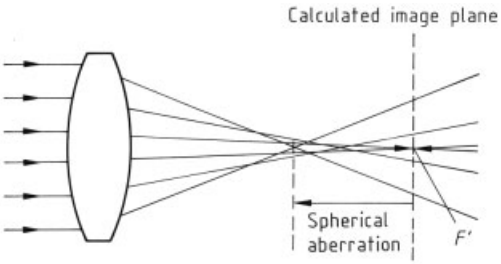

Figure 10. Spherical aberration

3.3. Effect of Partial Dispersion on Optical Imaging

The longitudinal chromatic aberration of an achromat can be corrected only for two wavelengths (Fig. 9 B). This type of cemented lens does not offer achromatic, but rather dichromatic correction.

The residual longitudinal chromatic aberration of such a dichromat is called the *secondary spectrum*

$$\Delta_S s' = -\frac{\theta_C - \theta_F}{\nu_C - \nu_F} f' \tag{28}$$

which depends on the difference between the relative partial dispersions (Eq. 7) of the two materials. As a result of the linear relationship given in Equation (8), equal relative partial dispersions (as required for correction of the secondary spectrum) cannot generally be obtained at the largest possible difference in Abbe numbers ($\nu_C - \nu_F > 25$). For achromats made of optical media that lie on the normal straight line defined in Equation (8), the secondary spectrum is therefore roughly constant:

$$\Delta_S s' = -b f' \tag{28a}$$

The secondary spectrum can be corrected for by supplementing the achromat shown in Figure 9 B with a lens made of an optical medium that exhibits anomalous partial dispersion (i.e., materials such as CaF_2, BaF_2, dense fluorophosphate crowns, or extralight flint glasses that do not obey Eq. 8). Systems of this type are called *trichromats* (Fig. 9 C). If they also correct for spherochromatism, they are called *apochromats* [2].

3.4. Effect of Transmittance on Optical Imaging

Since the resolving power of an optical system is inversely proportional to the wavelength, high resolution dictates the use of short wavelengths. In other words, it is necessary to use optical materials having high transmittance in the UV and DUV (deep ultraviolet). Optical storage, astronomy, microscopy, microlithography, and spectroscopy require internal transmittance of at least 50 % through a 10 mm thickness of material [5].

The UV absorption of optical glasses is roughly proportional to the refractive index [5]. Optical glasses used in UV microlithography therefore have a refractive index below 1.65. For wavelengths shorter than 400 nm, only silicate

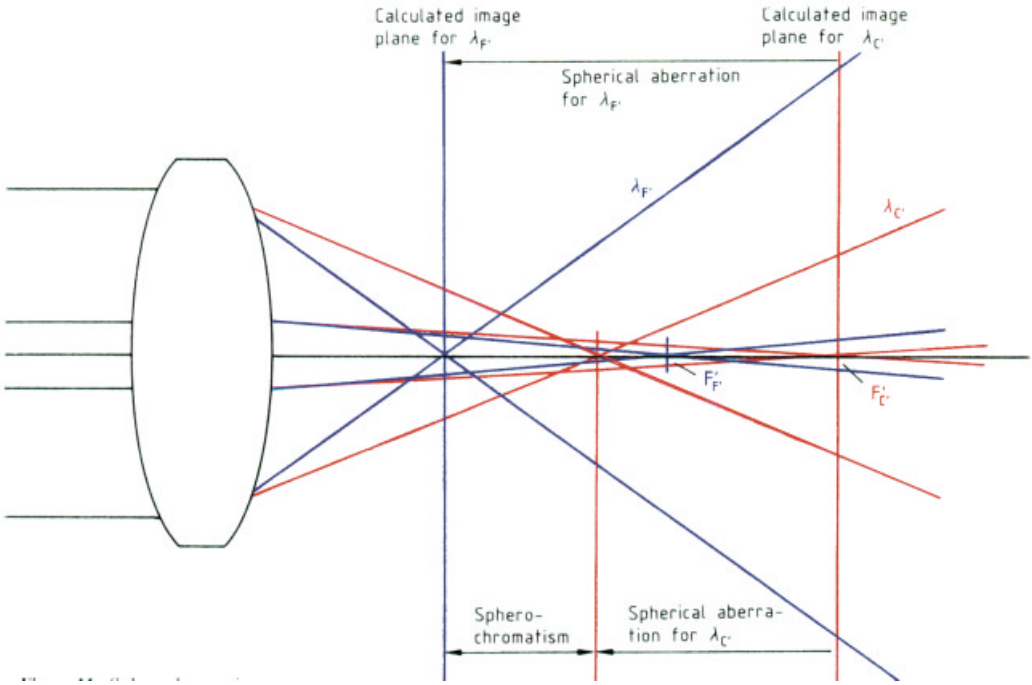

Figure 11. Spherochromatism
The subscripts F′ and C′ refer to the blue ($\lambda_{F'}$ = 480 nm) and red ($\lambda_{C'}$ = 643.8 nm) lines of cadmium, respectively.

glass and isotropic crystals such as CaF_2, BaF_2, spinel, and MgO can be used; weakly birefringent sapphire can also be employed under certain conditions.

4. Optical Glasses

For a detailed description of glass, see →Glass.

Glasses are the most important optical materials because their optical properties vary over wide ranges (Fig. 1). Refractive indices lie between about 1.3 and 2.2. The homogeneity and isotropy of the refractive index are of vital importance. Inhomogeneity and anisotropy give rise to phase shifts that cause undesired noise and distort the image structure.

The principal defect in small-diameter lenses is the occurrence of cords or striae which deform the wavefronts in large pupil areas. Suppression of cords is particularly important for the titanium flints and fluor crowns used in microscopic and photographic objectives. The permissible wave aberrations due to cords range between $\lambda/12$ and $\lambda/100$ [6].

For photolithographic objectives with their large lens diameter, interference due to large area inhomogeneities with low refractive-index gradients predominate. The variations in the refractive index in these inhomogeneities must be kept below 2×10^{-7} to 3×10^{-7} in lens blanks depending on lens diameter and thickness.

In the case of stress birefringence, tolerances are between 15 and 0.1 nm/cm (for polarization microscopy) [5].

Rational production methods demand that the refractive index should be constant to $\pm10^{-4}$ for photographic and microscope optics and to $\pm10^{-5}$ for photolithographic objectives. Microheterogeneities and bubbles act as scattering centers and interfere above all in coherent optical imaging, integrated optics, and fiber-optic communications.

Because the amorphous glassy state is metastable due to its high free energy, lower-energy photons (i.e., longer wavelengths) are needed for electron excitation than in the ordered, lower-energy, more stable crystalline state. Accordingly, the UV absorption edge of glasses occurs at a longer wavelength than that

of crystals (Fig. 6). Silicate and fluorophosphate glasses have a high UV transparency (Fig. 6).

In contrast to pure silicate glass, multicomponent glasses with their higher refractive indices (e.g., BK10, BaK5) do not have internal transmittance values of unity, even though there is no interaction between light and network components in the visible region. This is because even traces of iron, lead, manganese, cobalt, nickel, chromium, titanium, and cerium lead to selective transmission losses and a further shift of the UV absorption edge toward longer wavelengths [5].

The transmittance of high-index dense flint (SF) and dense barium flint (BaSF) glasses is unfavorable because the UV absorption edge is shifted further toward long wavelengths. Such glasses tend to take on a yellow cast with increasing refractive index, leading to shifts in color locus.

Good chemical stability is required for both the use and the processing of optical glasses. Acid attack (due to sweat from the hands, atmospheric CO_2 and SO_2, etc.) leads to exchange of hydrogen ions with network modifiers and the formation of low-index silica gel layers. These layers may result in formation of spots or even disintegration of the glass.

Flowing water has a less drastic action than acids, because the water-soluble alkaline products formed in the primary reactions are removed before the secondary reaction products can attack the glass surface. When a glass is exposed to atmospheric humidity, the primary reaction products remain on the surface and dissolve the network. The glass weathers and layers of material are lost.

Optical glasses are characterized by the principal refractive index n_e, the Abbe number (Eq. 5), and the type designation (e.g., BK 7 for a borosilicate crown). Types of glass are classified by their position in the n–ν diagram. The nomenclature is based on that devised by Otto Schott. Glasses are divided into the less dispersive crown glasses ($\nu_e > 55$) and the more dispersive flint glasses ($\nu_e < 50$). The range $55 > \nu_e > 50$ includes the barium light flints, crown flints, some extralight flints, and crown glasses.

Extremely low refractive indices, down to about 1.3, are found in the beryllium fluoride crowns (BeFK) [7], but these are not approved for use in most European countries because of the high toxicity of beryllium. The dense (heavy) fluorophosphate crowns (FPSK) [8], [9] have low refractive indices, high anomalous partial dispersions, and good UV transmittance. They are made by melting the phosphates and fluorides of the alkaline earth and rare-earth metals, but their strong crystallization tendency and their poor weathering resistance make them difficult to work with. The fluorocrowns (FK) have low refractive index and low density; they contain fluorine in place of oxygen in a glass matrix consisting of the oxides of light elements. The low-index, high-dispersion titanium flints (TiF) and dense titanium flints (TiSF) are obtained by incorporating fluorine in lead flint glasses. Borosilicate crowns (BK) and zinc crowns (ZK) contain up to 20 % B_2O_3 and ZnO, respectively, in place of SiO_2. The barium crowns (BaK), which contain nearly 30 % BaO, have higher refractive indices and dispersions than the light barium crowns (BaLK), which contain a maximum of 10 % BaO. If the BaO content is brought close to 50 %, dense crowns (SK) are obtained; if PbO and TiO_2 are added, the more highly refractive and dispersive extra-dense (very heavy) crowns (SSK) result. The highest refractive indices in the crown-glass region belong to the lanthanum crowns (LaK) and dense lanthanum crowns (LaSK). The high-index lanthanum flints (LaF) and extra-dense lanthanum flints (LaSF) contain oxides of lanthanum, lead and titanium. Flint glasses of moderate dispersion and index follow (listed in order of decreasing BaO and PbO contents): dense barium flints (BaSF), barium flints (BaF), flints (F), light flints (LF), and extra-light (very light) flints (LLF). The highest refractive indices and dispersions belong to the dense flints (SF), containing up to 70 % lead oxide and titanium oxide. They are exceeded only by the not-yet-commercialized telluride glasses, some of which have a strong yellow cast. Short flints, contain ≤ 20 % Sb_2O_3, lower dispersion in the blue, and thus show anomalous partial dispersion.

Optical glasses are supplied in several forms: as slabs with polished inspection surfaces, uncut or cut blocks with two polished surfaces, profile rods with uncut lateral surfaces, round blanks made by gobbing, and preformed molded or cut blanks. The maximum within-melt variation in refractive index for standard quality is $\pm 10^{-4}$, for precision quality glass, $\pm 5 \times 10^{-6}$.

High-volume optical glasses are produced in continuous melters. Molded or gobbed blanks are taken or the molten glass is continuously cast into rods.

5. Colored or Filter Glasses

Colored or filter glasses are employed for selective absorption. For example, a yellow filter intercepty the complementary blue portion of the spectrum. Glass can be colored with ions of heavy metals (e.g., Cu, Co, Ni, Fe, Cr, and Mn) or rare earths (e.g., Nd and Pr) or with colloidal particles of sulfides (e.g., CdS), selenides (e.g., CdSe), or metals (e.g., Cu, Ag, Au, or Se). The composition of the base glass is also important. Copper ions, for example, produce a blue color in conventional potassium – calcium glasses but a blue-green in borosilicate glasses.

The transition from transmission to absorption, especially in ion-colored glasses (Fig. 12 A), generally covers a wider range of wavelengths then in interference filters. Tempered glasses, in which the colored colloids listed above are precipitated after heat treatment, show a steep curve between the short-wavelength absorption region and the long-wavelength transmission region (Fig. 12 B). A steep flank cannot be obtained in the long-wavelength absorption region of filter glasses, neither can narrow spectral bands be filtered out (Fig. 12 C). The gray, compensating or neutral glass filter attenuates light fairly uniformly throughout the visible region (Fig. 12 D).

6. Optical Crystals

Optical crystals have several advantages over optical glasses: better transmittance in the UV and IR, extreme refractive indices and dispersions, and high anomalous partial dispersions [10]. The more stable crystal structure with its lower-energy configuration shifts the UV absorption edge to shorter wavelengths than those found for the metastable, amorphous glassy state (Fig. 6) [5].

Many crystals offer excellent wideband transmission, from UV to IR; this is important chiefly in spectroscopy (e.g., for filters and windows) [11], [12]. Pure oxide crystals such as quartz, magnesium oxide, corundum, spinel ($MgAl_2O_4$), GGG ($Gd_3Ga_5O_{12}$), and undoped YAG (yttrium aluminum garnet) have outstanding transparency into the deep ultraviolet, as do pure fluoride crystals such as LiF, NaF, CaF_2, and MgF_2. This is significant for the use of (especially) BaF_2, CaF_2, and undoped YAG for deep ultraviolet microprojection lithography between 300 and 200 nm.

Typical IR crystals include Si, Ge, NaCl, NaF, KBr, KSR5 (TlBrI), CdTe, ZnSe, ZnS, and GaAs; CsI and Ge transmit into the far infrared beyond a wavelength of 50 µm.

Due to their inherent lattice structure, crystals exhibit birefringence, polarization behavior, electro-optical properties, and nonlinear optical properties which are exploited in crystal polarizers, modulators, beam-deflection systems, and frequency converters. Birefringent crystals include sapphire, quartz, ammonium dihydrogen phosphate, potassium dideuterium phosphate, and calcite.

Drawbacks of optical crystals in comparison with optical glass include higher price, the solubility of many halide crystals in water (only fluorspar, with a minimum water solubility of 1.5×10^{-3} g/100 cm^3, is moisture-resistant enough for practical purposes), and their cleavage anisotropy, which makes fabrication difficult. Optical anisotropy limits the use of birefringent crystals in imaging optics.

Virtually all optical crystals are grown as single crystals either from solution or from a melt. Exceptions are the natural minerals calcite (for polarizers) and mica (for phase plates).

7. Optical Plastics

When compared with glasses and crystals, organic polymers have the crucial advantage that they can be shaped by injection molding [13] or casting [14], techniques that do not involve the removal of material. As a result, complicated optical components such as Fresnel lenses, aspheric lenses, lens – prism combinations, and lenses with integrally molded mounting parts (Fig. 13) can be manufactured by simple economical methods. Because of their poor quality, however, such components are mainly used in

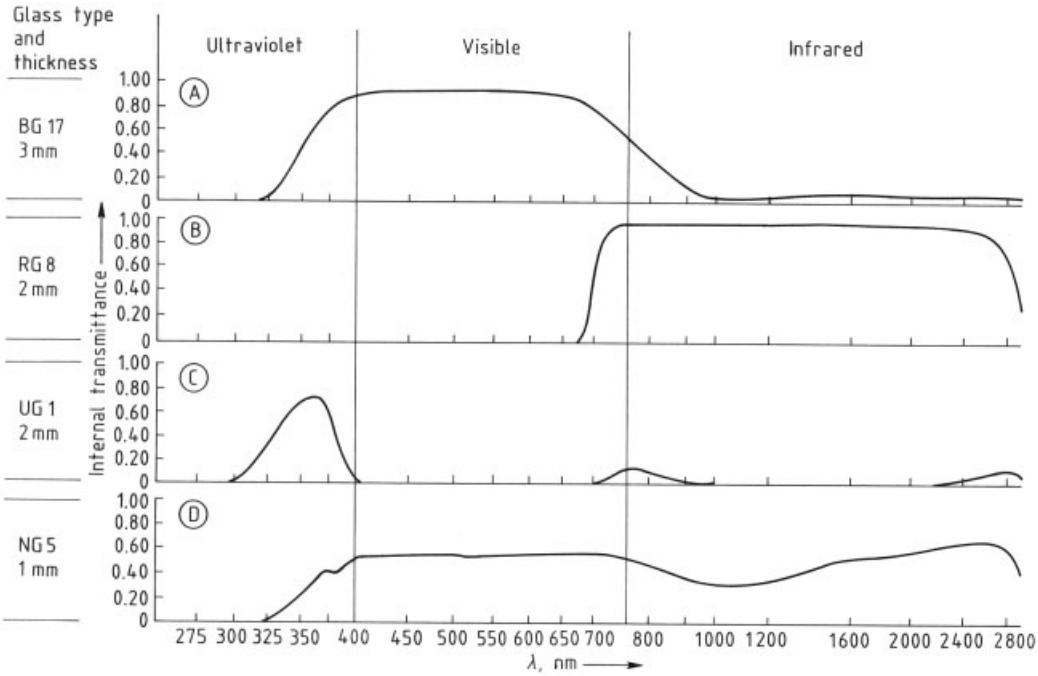

Figure 12. Transmission curves of optical filter glasses
A) Ion-colored heat filter with gradual absorption edges; B) Dark red tempered glass filter with steep edge; C) Ultraviolet filter with narrow passband; D) Light neutral glass filter

consumer goods. The use of plastic optics represents a cost saving of 5 – 15 % in comparison with glass optics [13].

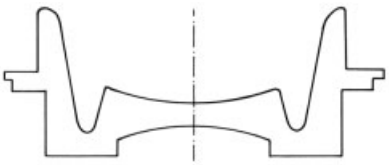

Figure 13. Plastic lens with integrally cast mounting part

Other advantages of optical plastics over glass are a more than 50 % lower density and a 10-fold lower shock sensitivity and resistance to breakage. Their surfaces are resistant to solvents and can easily be dyed. They can be subjected to seamless welding (polymerization). Refractive index and dispersion can be adjusted by the use of copolymers [15]. Replication techniques can be employed to bond aspheric plastic skins on spherical glass lenses (Fig. 14) [16]. Polarizing or colored films can be polymerized onto the base material.

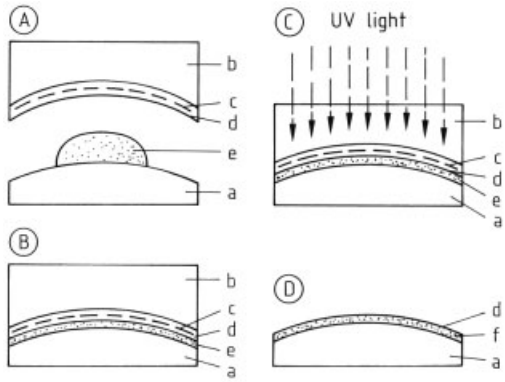

Figure 14. Replication method for making an aspheric precision plastic lens from a spherical glass lens
A) UV-transparent quartz glass mold with release agent and protective coating, base lens with drop of cast resin monomer; B) Molding of a layer of monomer; C) Curing of resin on base lens accelerated by UV light; D) Aspheric replica lens with resin layer after removal from mold
a) Lens; b) Mold; c) Release agent; d) Protective coating; e) Monomer; f) Replica layer

The principal disadvantage of optical plastics as compared against glasses is that their op-

tical and mechanical properties are much more (ca. 10-fold) dependent on temperature. This, together with their up to ninefold lower thermal conductivity and up to threefold higher specific heat, leads to different cooling rates at the surface and in interior regions, causing deformations, inhomogeneities in refractive index, and index shifts of up to $\pm 10^{-3}$ [13].

Another important disadvantage of optical plastics is the strong dependence of the optical and mechanical properties on atmospheric humidity. Swelling after water absorption of 0.1 – 0.8 % can lead to dimensional changes as great as 4 %. Absorbed water, even at low concentrations, also alters the refractive index.

Plastics are also less transpartent than optical glass, internal transmittance of a 2 mm thick flat is 0.6 – 1.9 % lower. Plastics have a poorer thermal stability.

Mold shrinkage amounts to 0.4 – 0.7 % for thermoplastics and 4 – 14 % for thermosets, and causes deformation and residual stress in components with large thickness differences.

Other problems include haze due to surface roughness and Tyndall scattering; contamination as a result of the high surface resistivity and buildup of electric charge; aging with yellowing, haze, cold creep, stress corrosion cracking, and shrinkage; poor resistance to scratching and wiping; and poor adhesion of antireflection and reflective coatings.

More than 100 optically practicable polymers are now known, having refractive indices between 1.35 and 1.72 and Abbe numbers between 86 and 18.8 (Fig. 1) [17]. Most are, however, noncommercial laboratory products that are made from expensive monomers or are difficult to process or color. High indices are achieved by adding S, Cl, I, As, or metal ions. Aromatic carbon is especially effective in raising the refractive index, but often yields expensive colored products that are hard to process. An increase in the index, and mostly without intrinsic color, is obtained by bromination or chlorination. Fluoridation or perfluoridation, on the other hand, lowers the index. The most commonly used commercial optical plastic is the crown-glass analogue poly(methyl methacrylate) (PMMA) [18], but this material absorbs a lot of water (0.8 – 4 %) and swells with dimensional changes [19]. It is likely to be replaced by cyclohexylmethyl methacrylate, which is not yet on the market. Another plastic, polymethylpentene (TPX), is a close optical analogue of PMMA; it is lighter than water and has a relatively high thermal stability (180 °C), but shrinks more than PMMA and has twice as high a coefficient of thermal expansion. Teflon, with an extremely low refractive index, can be used only in very thin sections because of its strong scattering.

Polystyrene, a flint-glass analogue, is inferior to PMMA as far as service qualities are concerned. It is increasingly being displaced by the optically analogous but mechanically and thermally superior polycarbonate (bisphenol A – phosgene polycarbonate) [20]. A still higher refractive index of 1.612 is obtained with DPSC (a copolymer of bisphenol and dihydroxyphenyl sulfide).

The gap between PMMA and polystyrene in the $n-\nu$ diagram (Fig. 1) is being filled by styrene copolymers, which, however, often have a slight yellow cast. These include styrene–acrylonitrile (SAN), styrene–maleic anhydride (SMA), polytrifluoromethylstyrene (FMS), and cyclohexylmethyl methacrylate–styrene (CMMAS) copolymers [18]. They have properties markedly better than those of polystyrene. Despite its high dispersion ($\nu \approx 34$), FMS has an extremely low refractive index (1.508).

8. Manufacturing Techniques

8.1. Processing of Optical Glasses

The classical method of producing optical components from optical glasses [21] involves sawing, cutting, nibbling (shanking), and rounding slabs of glass so that as little material as possible has to be removed in the subsequent grinding process. The preparation of optical crystals is similar.

The more economical technique involving molded blanks, which require little grinding, can only be employed with moldable glasses that do not tend (like many high-index dense lanthanum crowns) to recrystallize when under pressure [5]. On economic grounds, blanks of mass-produced glasses are conventiently gobbed from a continuous melter and then molded. There is a trend toward the use of precision-molded lens blanks that are ready for polishing. Condensing lenses

are made in smooth-surfaced molds so that no polishing is needed.

The classical grinding process, involving the use of a loose abrasive and a grinding tool, is being supplanted by grinding (milling) with diamond-ring tools. Multispindle grinding machines with a central drive are being replaced by numerically controlled single-spindle machines [22], [23]. To obtain short polishing times, fine-milled (fine-ground) lens surfaces should have maximum roughnesses of the order of 1 μm.

Grinding is a purely mechanical process. The subsequent polishing step, however, involves chemical reactions between the polishing medium [Fe_2O_3 (rouge) or cerium oxide], the polishing pitch or felt, and the water in which the polishing medium is slurried [24].

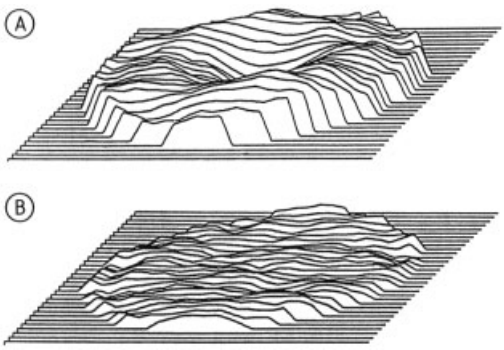

Figure 15. Deformation (departure from sphericity) of a spherical surface (A) before and (B) after computer-controlled correction polishing
Diameter = 29 cm; Radius of curvature = −45 cm; λ = 633 nm; Surface deformation: root mean square value 0.012 λ before and 0.004 λ after correction, peak to valley value 0.075 λ before and 0.028 λ after correction

In classical low-precision polishing (e.g., for condenser and spectacle lenses), the tools are equipped with felt, which allows faster polishing; in precision optics, pitch is used. For high-performance optics, composite aids (polishing films and pellets) and synthetic polishing-medium supports are employed to give a surface accuracy of ca. ± 50 nm. In precision optics, however, the lens surfaces may only have a maximum deformation of ± 20 nm or, under some circumstances, ± 10 nm. High precision can be achieved in large-diameter lenses through local computer-controlled correction polishing [26]. Control is based on the use of interferometry to detect deformations in the polished surface to be corrected. The polishing tool is small in comparison with the diameter of the surface being treated, and remains in the relevant surface regions until the surface deformations have been corrected. Figure 15 shows the smoothing of deformations in a spherical surface in such a correction process. Correction polishing can also be done by ion-beam methods [25], the density of the ion beam being modified with perforated masks or aperture control. The ion-beam density corresponds to the thickness of material to be removed.

After polishing, the lenses are centered and edged so that they are bounded by a cylinder whose axis coincides with the optical axis. The optical axis must be aligned with the rotation axis of a spindle accommodating the lens. Alignment can be carried out optically, with the reflected images of luminous points formed at the centers of curvature of the lens surfaces, or mechanically, with a pair of annular cutting edges arranged in tong fashion. The edge of the lens is then milled to a cylindrical form.

Aspheric lenses made from glass by fine milling and polishing have a maximum accuracy of ca. 0.1 μm. With computer-controlled correction polishing or ion-beam etching, spherical surfaces can be deformed in the range of a few tens of nanometers to give nanometer aspherics (Fig. 16) [26].

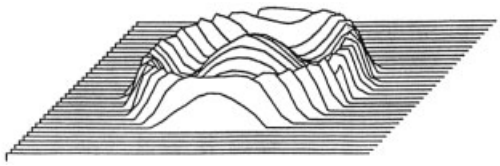

Figure 16. Deformation (departure from sphericity) profile of a nanometer aspheric made by ion-beam etching with perforated mask control
Diameter = 70 mm; Radius of curvature = 225 mm; λ = 633 nm; Surface deformation: root mean square value 0.068 λ

8.2. Casting of Optical Thermosets

Optical components made of thermosets (e.g., CR 39, epoxy resins) can be fabricated by casting and setting [14]. Thermosets harden by polymerization of a liquid monomer (e.g., diethylene glycol bisallyl carbonate) in a mold. The process is relatively slow and is accompanied

by 4 – 14 % shrinkage. Shrinkage can, however, be reduced by the addition of set resin from the same batch, the use of prepolymerized gels which have already undergone most of their shrinking, or by casting around prefabricated component cores. The rather long setting time can be greatly shortened and the shrinkage minimized by accelerating polymerization with UV light or X-rays.

Diffraction-limited aspheric elements can be produced by polymerizing thin aspheric thermoset layers onto spherical glass lenses (Fig. 14) [27]. This method, called replication, employs aspheric quartz glass molds coated with a release agent (silicone oil). The resin layer is pressed against the substrate by the mold (Fig. 14 B) and sets in under 1 min on exposure to UV irradiation (Fig. 14 C). Because thinner layers shrink less on polymerization, the resin must be no thicker than 0.01 mm to a maximum of 1 mm.

8.3. Processing of Thermoplastic Optical Components

The most rational and economical technology for making plastic optical elements is injection molding of granulated thermoplasts. One die can feed resin to several molds simultaneously so that a number of lenses can be produced in a short cycle time. This technique has the advantage that mounting parts can be molded integrally with the optics. The maximum lens diameter is, however, limited to 30 mm.

Because the properties of injection-molded plastics cannot be adequately controlled, they cannot be employed in high-quality optics. Longer cooling times (ca. 1.6 min) are needed for the injection molding of plastic lenses which may alter material properties. Orientation of the molecular chains in the die during injection into the mold can also result in birefringence [11]. By optimal die temperature control, the shrinkage can be kept to 0.3 – 0.6 % (depending on material and process conditions). This shrinkage can be offset by correction of the mold.

Because of cooling variations, solidification occurs unevenly within the plastic lens. The hardening plastic may also stick to the cooler mold. These problems can be overcome by using ceramic molds, which are poor heat conductors that insulate the entire lens volume and hold it above the solidification temperature for some time. In contrast to steel molds, solidification in ceramic molds is virtually stress-free. However, even with ceramic molds, lenses with large diameters and large edge-to-center differences in thickness cool unevenly resulting in differential shrinkage, deformations, stress birefringence, and inhomogeneities.

9. References

1. A. Sommerfeld: "Optik," *Vorlesungen über theoretische Physik,* vol. 4, Akademische Verlagsgesellschaft Geest & Portig KG, Leipzig 1959.
2. C. Hofmann: *Die optische Abbildung,* Akademische Verlagsgesellschaft Geest & Portig KG, Leipzig 1980.
3. C. Hofmann: "Wechselwirkung zwischen Glasindustrie und optischem Instrumentenbau zwischen 1800 und 1884," *Augenoptik (Berlin)* **101** (1984) 131 – 137.
4. C. Hofmann: "Die Auswirkungen der Zusammenarbeit von E. Abbe und O. Schott für die Entwicklung optischer Gläser und für den Jenaer optischen Prä-zisionsgerätebau am Ende des 19. Jahrhunderts," *Augenoptik (Berlin)* **101** (1984) 162 – 168.
5. C. Hofmann, W. Nordwig: "Zu den Anforderungen der Optik an die Entwicklung optischer Medien," *Wiss. Z. PH Erfurt Mühlhausen (Erfurt)* **18** (1982) 5 – 18.
6. V. Giggel, A. Leitel: "Rechnergesteuerte Polierkorrektur sphärischer optischer Funktionsflächen," *Feingerätetechnik* **38** (1989) 3 – 4.
7. C. F. Cline, M. J. Weber: "Beryllium Fluoride Optical Glasses," *Wiss. Z. Friedrich-Schiller-Univ. Jena, Math. Naturwiss. Reihe* **28** (1979) 351 – 364.
8. W. Vogel: "Möglichkeiten und Grenzen der Entwicklung optischer Gläser von Otto Schott bis zur Gegenwart," *Wiss. Z. Friedrich-Schiller-Univ. Jena, Math. Naturwiss. Reihe* **32** (1983) 495 – 508.
9. D. Ehrt, R. Atzrodt, W. Vogel: "Struktur-Eigenschaftsbeziehungen optischer Gläser mit anomaler Teildispersion," *Wiss. Z. Friedrich-Schiller-Univ. Jena, Math. Naturwiss. Reihe* **32** (1983) 509 – 526.
10. U. Steinbruch, J. Kötitz, J. Pudenz: "Kristalle als wichtige Ergänzung des Sortiments der Jenaer optischen Gläser," *Jenaer Rundsch.* **29** (1984) 71 – 73, 82.

11. W. Horn: "Optische Werkstoffe für die unsichtbaren Spektralgebiete Infrarot und Ultraviolett," *Jahrbuch für Optik und Feinmechanik (Wetzlar)* **12** (1965) 85–121.
12. H. Schröder, N. Neuroth: "Optische Materialien für den ultravioletten und infraroten Spektralbereich," *Optik (Stuttgart)* **26** (1967/69) 381–401.
13. G. Klepek: "Herstellung optischer Linsen im Spritzprägeverfahren," *Kunststoffe* **77** (1987) 1147–1151.
14. C. Hofmann: "Möglichkeiten und Grenzen des Einsatzes von Plastoptik," *Bild und Ton* **43** (1990) 197–202.
15. U. Herrmann, C. Hofmann, F. Rudolph: "über die Entwicklung optischer Plastmedien mit gezielter Brechung und Dispersion auf der Basis von Mischpolymerisation von Diäthylenglykol-bis-allylkarbonat und Isophthalsäurediallylester mit unterschiedlicher Konzentration," *Feingerätetechnik* **21** (1972) 30–35.
16. H. M. Weissmann: "Epoxy Replication of Optics," *Opt. Eng.* **15** (1976) 435–441.
17. G. Klepek: "Optische Linsen aus Kunststoffen," *Kunststoffe* **78** (1988) 340–344.
18. R. H. Waxler, D. Horowitz, A. Feldman: "Optical and Physical Parameters of Plexiglas 55 and Lexan," *Appl. Opt.* **18** (1979) 101–104.
19. G. Klepek: "Kunststofflinsen, Herstellung – Vergü-tung – Anwendung in Objektiven," *F & M Feinwerktech. Meßtech.* **96** (1988) 43–48.
20. W. Siegbourg: "Polycarbonat – Ein Werkstoff für optische Speichermedien," *Kunststoffe* **76** (1986) 917–919.
21. H. Schade: *Arbeitsverfahren der Feinoptik,* Deutscher Ingenieur-Verlag, Düsseldorf 1955.
22. W. Warziniak, U. Jungstand: "Entwicklungstendenzen der Werkzeuge und Maschinen für die Feinbearbeitung optischer Funktionsflächen," *Feingerätetechnik* **34** (1985) no. 2, 60–61.
23. W. Emmerich, C. Hofmann: "Trends der Optikentwicklung," *Feingerätetechnik* **36** (1987) no. 11, 488–492.
24. T. Izumitani, Sh. Harada: "Polishing Mechanism of Optical Glasses," *Wiss. Z. Friedrich-Schiller-Univ. Jena, Math. Naturwiss. Reihe* **28** (1979) 389–413.
25. T. Hänsel, F. Bigl, B. Retschke: "Ionenstrahlbearbeitung optischer Oberflächen," *Beiträge zur Optik und Quantenelektronik (Berlin)* **15** (1990) 146–147.
26. K. Merkel, C. Hofmann: "The Development of Microlithographic High-performance Optics," *Int. J. Optoelectronics (London)* **4** (1989) 545–562.
27. M. Braat, A. Schmidt, M. M. Wijnakker: "Design and Production Technology of Replicated Aspheric Objective Lenses for Optical Disk Systems," *Appl. Opt.* **24** (1988) 1853–1855.

Oral Antidiabetics → Antidiabetic Drugs

Oral Hygiene Products

Wolfgang Weinert, Bad Schwalbach, Federal Republic of Germany

1. Introduction 1
2. Dentifrices 1
2.1. Toothpaste 2
2.1.1. Composition 2
2.1.2. Production 4
2.1.3. Quality Specifications, Analysis, and Testing 5
2.2. Tooth Powder 6
3. Mouthwashes 6
4. Denture Cleansers 6
5. References 7

1. Introduction

Dental caries (formation of cavities) and periodontal diseases, the two most prevalent human diseases, are both caused by bacteria present in deposits on the teeth. These deposits, also known as plaque, contain a host of microorganisms. Acid-forming microorganisms are responsible for cavities, whereas anaerobes are primarily responsible for the development of gingivitis, are the first step of periodontal diseases. Therefore, bacterial plaque constitutes the basis for the destruction of teeth and their supporting structures [1–3]. Caries and periodontal diseases are estimated to cost insurance companies in the Federal Republic of Germany as much as €8.6×10^9 (ca. \$ 10^{10}) annually, and the cost is rising. Thus oral hygiene has a considerable economic impact on society, quite apart from its obvious advantages to the individual.

Removal of plaque is the key to preventing caries and gum diseases, but effective plaque removal requires a conscientious, daily effort. Most of the actual work is performed by a toothbrush, which constitutes the most important prophylactic agent. The action of the toothbrush is supported by toothpaste that contains cleansing agents to facilitate the removal of firmly bound deposits. Plaque removal from the relatively inaccessible spaces between the teeth requires the use of dental floss. Special interdental brushes serve a similar purpose, but they can be used only when the gaps between the teeth have already been widened by gum resorption. Plaque dyeing tablets make plaque deposits visible, so that the efficiency of their removal can be checked easily. Electric toothbrushes have also been found to be useful in plaque removal, even exceeding the effectiveness of ordinary toothbrushes when properly used [4,5]. Oral irrigators, on the other hand, are not intended for plaque removal, although they can help to eliminate previously dislodged plaque and food residues from inaccessible interdental openings.

All the tools so far described are based on a mechanical plaque-removal strategy. Chemical plaque control is currently possible only with chlorhexidine digluconate [*18472-51-0*], an antimicrobial agent that suppresses plaque deposition because of its high affinity for teeth and the oral mucous membranes [6]. Continuous use of the substance is recommended only in special cases (e.g., in patients with jaw fractures [7]) because long-term application leads to discoloration of dental surfaces. Although discoloration is reversible, it still represents a cosmetic problem.

2. Dentifrices

The agents used to assist a toothbrush in removing plaque are known as dentifrices, they include toothpaste and toothpowder. *Toothpastes* are regarded as cosmetics rather than drugs and are thus subject only to legal restrictions applicable to cosmetics, including regulations governing the amount and type of dyes, preservatives, and fluorine compounds they may contain. Most commercially distributed toothpaste falls into one of four major categories: *medicinal, anticaries, antitartar,* and special *pediatric* toothpaste. The precise formulation of a toothpaste is a function of the purpose for which it is intended. Medicinal toothpaste contains ingredients designed to combat inflammation of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_209

the gums, whereas fluoride is the most important additive in toothpaste aimed at prevention of caries. Many toothpastes offer significant protection against the formation of tartar, but they cannot remove existing tartar deposits [8]. Because children perceive tastes differently from adults, special pediatric toothpastes have been developed [9]. These usually have fruit flavors and a low fluoride level. The latter is dictated by a concern for possible fluorosis reactions, particularly in children receiving supplemental doses of fluoride in tablet form. In clear gels hydrated silica is used as the cleansing agent; all other abrasive agents produce opaque (white) toothpastes. *Tooth powder* plays a much more limited role in oral hygiene than toothpaste, perhaps because its application by toothbrush is less convenient. Furthermore, toothpaste produces the sensation of being more evenly distributed throughout the mouth. Nevertheless, tooth powder can include active ingredients that are unstable in a medium containing water.

2.1. Toothpaste

2.1.1. Composition

Toothpaste contains active ingredients and auxiliaries. The active ingredients include:

1) cleansing agents,
2) fluorine compounds,
3) surfactants, and
4) other ingredients.

Auxiliaries include:

1) moisturizers,
2) binders,
3) preservatives,
4) flavoring agents,
5) sweeteners, and
6) dyes.

Cleansing Agents. Cleansing agents are usually insoluble, inorganic substances that constitute 15 – 60 wt % of a typical toothpaste. These materials, which might also be described as polishing agents or abrasives, enhance the cleansing action of the toothbrush and reduce the time required for cleaning. Their effectiveness depends on removing plaque deposits with a minimum of abrasion. The particle size of cleansing agents must be such that individual grains are not perceived as foreign bodies in the mouth; the preferred mean particle size is < 15 µm. The cleansing effect of such particles depends on their size, shape, and hardness.

The most commonly employed cleansing agents are aluminum hydroxide [$Al(OH)_3$], calcium carbonate ($CaCO_3$), calcium hydrogen phosphate dihydrate ($CaHPO_4 \cdot 2H_2O$), calcium hydrogen phosphate ($CaHPO_4$), hydrated silica ($SiO_2 \cdot H_2O$), sodium aluminum silicate [e.g., with the zeolite structure $Na_{12}(AlO_2)_{12}(SiO_2)_{12} \cdot 27\ H_2O$], insoluble sodium metaphosphate [$(NaPO_3)_n$], and hydroxyapatite [$Ca_5(PO_4)_3OH$].

Some toothpastes also contain the soluble cleansing agent sodium hydrogen carbonate, usually mixed with calcium carbonate. The soft sodium hydrogen carbonate often consists of relatively large particles and is referred to as a "temporary cleansing agent" because it slowly dissolves in saliva during brushing. A very finely divided form of the water-insoluble, organic cleansing agent poly(methyl methacrylate) is also sometimes employed.

Fluoride in toothpastes increases the resistance of dental enamel against attack by the acid released by bacterial plaque; it also facilitates remineralization of the enamel. Fluoride is thus essential in toothpaste for caries prophylaxis. The most commonly employed fluorine compounds are sodium fluoride, sodium monofluorophosphate, and certain organic amine fluorides [e.g., bis-(hydroxyethyl)-aminopropyl-*N*-hydroxyethyloctadecylamine dihydrofluoride].

Surfactants (foaming agents) lower surface tension and contribute to the uniform distribution of toothpaste during brushing. They also increase cleansing power, especially at sites that are difficult or impossible to reach with a toothbrush. Obviously, the only surfactants that can be considered for toothpaste are those that are physiologically safe, they should also have as neutral a taste as possible. The most commonly used surfactants include sodium lauryl sulfate, sodium lauryl sarcosinate, sodium cocoate, sodium lauryl sulfoacetate, sodium cocomonoglyceride sulfonate, and betaines.

Other Active Ingredients. This category includes substances that either strengthen enamel (e.g., remineralizers) or have an effect on plaque (e.g., antimicrobial agents, plaque suppressants, tartar control agents, or enzymes). Additional possibilities include components with a beneficial effect on the oral mucous membranes, such as anti-inflammatory agents, astringents, or vasoactive agents.

Moisture control agents primarily minimize the tendency of a toothpaste to dry out but also affect its consistency and enhance its stability in the cold. The most common moisture control agents for toothpaste are glycerol, sorbitol, xylitol, 1,2-propylene glycol, and poly(ethylene glycol).

Binders and thickeners determine the consistency of toothpaste and prevent the separation of liquids and solids. The principal binders are natural and synthetic hydrophilic colloids such as alginates, carrageenan, tragacanth, sodium carboxymethyl cellulose, hydroxyethyl cellulose, methyl cellulose, methyl hydroxypropyl cellulose, xanthan gum, and guar gum. Typical thickeners are highly dispersed silica, magnesium aluminum silicate, magnesium lithium silicate, and bentonite.

Sweeteners improve the taste of a toothpaste. The use of sodium saccharine is especially widespread, often in combination with sodium cyclamate. Acesulfame-K and aspartame have also been utilized.

Preservatives are required to protect toothpaste from microbial decomposition. The most important preservatives for this purpose are esters of 4-hydroxybenzoic acid, or benzoic acid and its sodium salt, as well as sorbic acid and potassium sorbate. Compounds such as 2-hydroxybiphenyl, bromochlorophene, dehydroacetic acid, and chlorhexidine have also been employed.

Dyes. Both water-soluble and water-insoluble dyes are used in toothpaste. Water-insoluble lakes and pigments are particularly common for dyeing multicolored, "striped" products, where migration of color into the white toothpaste base must be prevented. The most frequently used dyes are cochineal red A (C.I. 16 255), patent blue V (C.I. 42 051), indigotine (C.I. 73 015), chlorophyll (C.I. 75 810), quinoline yellow (C.I. 47 005), titanium dioxide (C.I. 77 891), indanthrene blue RS (C.I. 69 800), and madder lake (C.I. 58 000). All of these dyes have high resistance to acid and base, as well as excellent temperature stability.

Flavoring Agents. A wide range of flavoring agents can be found in toothpaste, in most cases introduced as mixtures. The essential oils listed in Table 1 are described in more detail elsewhere (→ Flavors and Fragrances).

Table 1. Flavoring agents used in toothpaste

Substance	Origin
American peppermint oil	*Mentha piperita*
Brazilian peppermint oil	*Mentha arvensis*
Spearmint oil	*Mentha crispa*
Anise oil	*Pimpinella anisum*
Eucalyptus oil	*Eucalyptus globulus*
Fennel oil	*Foeniculum vulgare*
Lemon oil	*Citrus limon*
Wintergreen oil	*Gaultheria procumbeus*
Clove oil	*Eugenia caryophyllata*

Flavor specialists have devoted considerable attention to improving the taste of toothpaste with additives. The aim is to increase overall acceptance of the product and thus raise the conscientiousness of the public with respect to hygiene. Cleansing agents such as calcium carbonate, calcium phosphate, hydrated silica, aluminum hydroxide, and sodium metaphosphate can bind fairly large amounts of flavoring components.

Peppermint-flavored toothpastes represent 35 % of the world market. They are distributed throughout the world and are particularly common in Europe (especially Scandinavia, UK, and FRG), playing a smaller role in the United States and Asia. The generic formulation (in weight percent) is

Peppermint oil	ca. 60
Menthol ex peppermint oil	10 – 15
Anise oil, fennel oil	10 – 15
Eucalyptus oil	5 – 10
Spice oils	2 – 5
Nuance substances (floral oils, lemon oil)	0.5 – 10

Spearmint-flavored products represent 40 % of the market. These are also found worldwide, especially in the United States and Asia (less

in Europe). The generic formulation (in weight percent) is

Spearmint oil	ca. 30 – 40
Peppermint oil	20 – 30
Menthol ex peppermint oil	10 – 15
Anise oil, fennel oil	10 – 15
Spice oils	0.5 – 3
Nuance substances	0.1 – 5

Wintergreen-flavored toothpastes account for 20 % of the world market, primarily in the United States and the Philippines. The generic formulation (in weight percent) is

Peppermint oil	ca. 20 – 30
Spearmint oil	5 – 20
Menthol ex peppermint oil	5 – 10
Anise oil, fennel oil	10 – 30
Eucalyptus oil	10 – 20
Wintergreen oil	10 – 30
Spice oils	1 – 10
Nuance substances	1 – 2

Peppermint-anise toothpastes are particularly common around the Mediterranean, accounting for 5 % of the world market. The generic formulation (in weight percent) is

Peppermint oil	ca. 10 – 30
Spearmint oil	10 – 15
Menthol ex peppermint oil	5 – 10
Anise oil, fennel oil	30 – 40
Eucalyptus oil	5 – 20
Spice oils	1 – 2
Nuance substances	2 – 3

Examples of complete toothpaste formulations follow (values are in weight percent):

Toothpaste for gingival prophylaxis	
Methyl cellulose	1.10
Sorbitol, 70 %	15.00
Water	31.00
Sodium saccharin	0.05
Carbamide	5.00
Sodium benzoate	0.20
α-Bisabolol	0.10
Sage extract	0.50
Flavoring agent	1.10
Methyl 4-hydroxybenzoate	0.15
Calcium carbonate	45.00
Sodium cocoate	0.80
Anticaries toothpaste	
Sodium carboxymethyl cellulose	1.40
Sorbitol, 70 %	28.00
Xylitol	5.00
Water, deionized	43.37
Acesulfame-K	0.25
Sodium fluoride	0.24
Methyl 4-hydroxybenzoate	0.18
Propyl 4-hydroxybenzoate	0.06
Flavoring agent	1.00
Silicic acid abrasive	11.00
Silicic acid thickener	8.00
Sodium lauryl sulfate	1.00
Fatty acid tauride	0.50
Antitartar toothpaste	
Sorbitol, 70 %	15.00
Carrageenan	0.80
Glycerol	27.00
Poly(ethylene glycol), PEG-12	2.00
Sodium lauryl sulfate	1.40
Silicic acid abrasive	20.00
Tetrapotassium pyrophosphate	4.50
Tetrasodium pyrophosphate	1.50
Titanium dioxide	0.50
Sodium saccharin	0.25
Flavoring agent	1.00
Sodium fluoride	0.24
Methyl 4-hydroxybenzoate	0.10
Water	25.71
Toothpaste for sensitive teeth	
Hydroxyethyl cellulose	1.40
Glycerol	18.00
Sorbitol, 70 %	12.00
Water	38.68
Sodium saccharin	0.35
Dye (C.I. 45 430)	0.01
Methyl 4-hydroxybenzoate	0.10
Propyl 4-hydroxybenzoate	0.06
Strontium chloride hexahydrate	10.50
Flavoring agent	1.00
Hydrated silica	15.00
Silica, highly dispersed	1.60
Sodium lauryl sulfate	1.30

2.1.2. Production

Toothpaste can be prepared in a batchwise or continuous process at any desired temperature. Cold processing (ca. 25 °C) is generally employed.

Process. The charge anticaries formulation is generally prepared in a 200 – 2500-L vacuum mixer equipped with a planetary stirrer and homogenizers. Sorbitol is introduced first, in which sodium carboxymethyl cellulose is then suspended with stirring. Deionized water is added with continued stirring followed by acesulfame-K, sodium fluoride, xylitol, and the two hydroxy benzoate esters. Stirring is maintained for ca. 20 min until the binder has finished swelling and all water-soluble components are dissolved. The mixer is then evacuated. As soon as the vacuum has reached 2 – 3 kPa, silicic acid abrasive, fatty

acid tauride, and sodium lauryl sulfate are drawn in and the mass is intensively stirred and homogenized for ca. 15 min. The flavoring agents are drawn in last, stirring and homogenization are maintained at full intensity under vacuum for ca. 10 min. until all bubbles have been removed to produce the desired smoothness and the mixture shows a homogeneous structure.

Continuous Process. Most continuous processes for toothpaste manufacture involve six production steps: gel preparation, introduction of liquid components, addition of solids, mixing, homogenization, and degassing.

Gel Preparation. The gel in a typical commercial toothpaste consists mainly of a mixture of glycerol, sorbitol, and binding agent. Problems are most likely to arise during the blending step because the formation of lumps is difficult to avoid with binding agents such as sodium carboxymethyl cellulose. The powder – liquid mixture is passed through a three-stage colloid mill equipped with a specially designed intake screw. Entrapment of finely divided air cannot be avoided and because this could lead to subsequent problems, the output of the colloid mill is passed through a dynamic degasser capable of continuously degassing the gel at its standard input density.

Introduction of Liquid Components. Liquid components are added through volumetric metering pumps. A separate metering head is best used for each liquid component, although premixing is also employed.

Mixing. The biggest problem in continuous processing involves the metering of solid components. The required quantity of mixture must be worked uniformly into the liquid stream. A homogeneous powder mixture is therefore needed. The densities of the individual raw materials can vary from 0.2 to 1.3 g/mL, their contents ranging from 0.5 – 95 % of the amount of powder. Metering may be accompanied by relatively rapid demixing because the powders in the metering vessel flow at different rates depending upon their densities. Screw metering is therefore preferred because it provides good metering accuracy over a variable performance range. Reproducible bulk density is achieved with the aid of preliminary loosening and homogenization of the charge.

The powder and liquid phases are mixed in a conical mixer containing an anchor screw device and a dissolver to ensure prehomogenization of the product. The process is conducted at atmospheric pressure, and entrapped air is subsequently removed with a dynamic degasser.

Homogenization is achieved with a rotary piston pump controlled by a level regulator installed on the mixer. The rotary piston pump transports the toothpaste into a three-stage inline colloid mill.

Degassing is performed in a dynamic degasser. The residence time and thus degassing is controlled with an exit pump subject to microwave adjustment.

2.1.3. Quality Specifications, Analysis, and Testing

Toothpaste is usually stable for at least 30 months even under variable temperature conditions. Physical, chemical, or biological changes can affect its quality. Oral hygiene products are therefore tested during product development for their storage behavior after 1 – 3 months at + 40, + 5, and – 10 °C, as well as after at least 6 months at 20 °C. Results are normally judged on the basis of consistency, color, taste, pH, fluoride content, content of other active ingredients, tube corrosion, gas generation, liquid separation, and microbiological stability [10, 11].

As noted earlier, oral hygiene products may contain many diverse ingredients. The variety of products and starting materials, as well as the multitude of potential test methods in a modern analytical laboratory, makes it difficult to provide a brief, product-oriented summary of oral hygiene product analysis. For detailed information see [12].

Toothpastes are subjected to both in vitro and in vivo testing with respect to their effectiveness. In vitro tests consist mainly of microbiological investigations and studies dealing with remineralization, whereas in vivo tests are designed to check product effectiveness in preventing caries and gingivitis [13–17].

2.2. Tooth Powder

Tooth powder differs from toothpaste primarily because of the absence of moistening agents such as glycerol, sorbitol, and water. As a result, the cleansing agent content can be as high as 90 %. The following formulation (in weight percent) is typical of a tooth powder designed for gingival prophylaxis:

Sodium carboxymethyl cellulose	1.50
Poly(ethylene glycol) 1500	4.00
Sodium saccharin	0.30
Aluminum lactate	8.00
Flavoring agent	1.40
Calcium carbonate	73.90
Sodium hexametaphosphate	6.00
Silicon dioxide, highly disperse	3.50
Sodium lauryl sulfate	1.40

3. Mouthwashes

Mouthwashes are generally not suitable for removing plaque and therefore can not replace toothbrushes and toothpaste. They mainly serve to give the user a certain feeling of freshness and to mask bad breath. To achieve this, mouthwashes contain alcohol and high concentrations of essential oils. Mouthwashes are sold as ready-to-use solutions or as concentrates to be diluted with water. A relatively simple percentage composition for a refreshing mouthwash is given in the following example::

Ethanol (94.7 %)	80.00
Sodium cyclamate	0.15
Flavoring	3.50
Red dye (C.I. 16 255)	0.01
Water, demineralized	16.34

Apart from mouthwashes designed to give a purely refreshing effect, there are also mouthwashes that provide a certain degree of prophylaxis against caries and periodontitis. These mouthwashes, available exclusively as ready-to-use solutions, contain active ingredients such as sodium fluoride or antibacterial substances like cetylpyridinium chloride. However, these mouthwashes alone do not provide adequate oral prophylaxis, and they should be used in combination with the usual oral hygiene measures, either to intensify the remineralizing and enamel-strengthening action of sodium fluoride or to provide antibacterial activity against residual plaque. A typical example of the percentage composition of a sodium fluoride-containing mouthwash is:

Ethanol (94.7 %)	7.000
Glycerol (86 %)	12.000
Sodium fluoride	0.050
Pluronic F-127 (emulsifier)	1.400
Sodium phosphate (buffer, pH 7)	1.100
Sorbic acid	0.200
Sodium saccharide	0.150
Red dye (C.I. 15 850)	0.002
Water, demineralized	77.998

4. Denture Cleansers

Removable dentures must also receive daily cleansing because they are just as subject to plaque deposition as natural teeth, and the associated bacteria cause infections of the oral mucous membranes. Special tablets and powders with antibacterial properties have been developed to facilitate mechanical plaque removal from dentures; these function much like detergents. All raw materials used in denture cleansers must meet regulatory standards for cosmetic agents.

Most denture cleansers contain 1 – 4 % active oxygen as their bactericidal component, which is released in aqueous solution as peroxo compounds. The effervescence that accompanies oxygen release supports the chemical cleansing action of the accompanying surfactants and helps in the mechanical removal of adhering particles. Generic formulations for denture cleansers are given in Table 2.

Table 2. Formulations of dental cleansers

Component	Content, wt %	
	Powder	Tablet
Sodium peroxoborate	20 – 25	20 – 25
Potassium peroxosulfate		20 – 30
Sodium tripolyphosphate	20 – 25	10 – 25
Sodium hydrogen carbonate	20 – 25	10 – 15
Citric acid		15 – 20
Sodium dodecylbenzenesulfonate	0.5 – 1	0.3 – 1
Sodium sulfate	25 – 30	20 – 25
Indigotine	0.005 – 0.02	0.005 – 0.02
Flavoring agent	0.5 – 1	0.5 – 1
Tableting agents		5

5. References

Acknowledgement. *This article is based on the author's contribution in W. Umbach (ed.): Kosmetik: Entwicklung, Herstellung und Anwendung kosmetischer Mittel,* Thieme Verlag, Stuttgart 1988, pp. 181 – 223. Permission to utilize this material is duly acknowledged.

1. L. Bailey, *Clinical Preventive Dentistry* **11** (1989) no. 3, 21 – 27.
2. K. Bößmann, *Oral-Prophylaxe* **10** (1988) 18 – 27.
3. S. Clergeau-Guerithault, *Quintessenz* **6** (1987) 1047 –1051.
4. H.-K. Albers, B. G. Schüßeler, K. Bößmann, *Zahn-ärztl. Mitteilungen* **16** (1988) 1777 – 1780.
5. R. Buchmann, *Quintessenz* **3** (1987) 1 – 8.
6. P. Riethe, B. Stoll, *Dtsch. Zahnärztl. Z.* **32** (1977) 753 – 754.
7. G. Krankel, G. Röthler, *Österr. Z. Stomatol.* **11** (1979) 408 – 413.
8. A. J. Rugg-Gunn, *Br. Dent. J.* **165** (1988) no. 4, 133 – 136.
9. W. Weinert, *Sozialpädiatrie in Klinik und Praxis* **6** (1984) 451 – 452.
10. A. K. Reng, F. J. Dany, *Parfümerie und Kosmetik* **2** (1978) 37 – 45.
11. S. B. Singh-Verma, *Parfümerie und Kosmetik* **6** (1985) 371 – 386.
12. H. König, E. Walldorf, *Fresenius Z. Anal. Chem.* **335** (1989) 216 – 223.
13. P. M. Cahen, R. M. Frank, J. C. Turlot, M. T. Jung, *Com. Dent. Oral Epidemiol.* **10** (1982) 238 – 241.
14. J. M. ten Cate, W. L. Jongebloed, J. Arends, *Quintessenz* **6** (1981) 1109 – 1111.
15. H. R. Wagner, W. Weinert, *Kariesprophylaxe* **3** (1981) 67 – 74.
16. W. Weinert, *Quintessenz International* **11** (1982) 1229 –1235.
17. D. J. White, J. D. B. Featherstone, *Caries Res.* **21** (1987) 502 – 512.

Organometallic Compounds and Homogeneous Catalysis

ARNO BEHR, Universität Dortmund, Dortmund, Federal Republic of Germany

1. Introduction 1
2. Organometallic Compounds of the Main Group Elements 2
2.1. Group 1 2
2.2. Group 2 2
2.3. Group 12 3
2.4. Group 13 4
2.5. Group 14 6
3. Transition Metals 7
3.1. Transition-Metal Complexes 7
3.1.1. Alkyls and Aryls 7
3.1.2. Carbenes and Carbynes 8
3.1.3. Carbonyls and Clusters 8
3.1.4. π-Complexes 9
3.2. Transition-Metal Catalysis 11
3.2.1. Catalytic Elementary Reactions 11
3.2.2. Isomerization 13
3.2.3. Hydrogenation 15
3.2.4. Carbon – Carbon Bond Formation . . 17
3.2.5. Metathesis 21
3.2.6. Oxidation 23
3.2.7. Reactions with Carbon Monoxide . . 24
3.2.8. Functionalization 27
3.2.9. Alkane Activation 30
4. References 30

1. Introduction [1–12]

Organometallic compounds are substances in which a metal or metalloid is involved in at least one direct bond to a carbon atom, whereby the bond can have either covalent or ionic character. Table 1 shows a section of the periodic table, listing those elements whose organic derivatives are regarded as organometallic compounds.

Table 1. Range of organometallic compounds in the periodic table

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Li	Be											B			
Na	Mg											Al	Si	P	
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te
Cs	Ba	La$^+$	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	(Po)
Fr	Ra	Ac∗													

La$^+$ = Lanthanum and lanthanides; Ac∗ = Actinium and actinides.

In this article, only some important examples of organometallic compounds are described: the organic derivatives of lithium, magnesium, zinc, mercury, and boron are important reagents in organic synthesis, while organometallic compounds of aluminum, silicon, tin, and lead are produced industrially on a large scale. With few exceptions (e.g., iron pentacarbonyl), organotransition metal complexes have found only limited application in industrial stoichiometric reactions. On the other hand, they act as intermediates in transition-metal catalysis, in which they largely control the course of the reaction.

Although the first organometallic compound, $[(CH_3)_2As]_2O$ [*503-80-0*], was prepared as early as 1760, it was only in the late 1800s that the first major developments began, with the seminal work of LUDWIG MOND on the carbonyls of nickel and iron and the discovery of the Grignard reagents. The systematic study of organo transition metal complexes and their potential in catalysis began in the mid-1900s with the work of W. REPPE, E. O. FISCHER, K. ZIEGLER, and G. WILKINSON. Some important discoveries of organometallic compounds are listed in Table 2.

Table 2. Historical development of organometallic chemistry

Year	Discoverer	Compound
1760	CADET DE GASSICOURT	$[(CH_3)_2As]_2O$
1827	W. C. ZEISE	$K[PtCl_3(C_2H_4)]$
1849	E. FRANKLAND	$(C_2H_5)_2Zn$
1868	M. P. SCHÜTZENBERGER	$[Pt(CO)Cl_2]_2$
1890	L. MOND	$Ni(CO)_4$
1891	L. MOND	$Fe(CO)_5$
1900	V. GRIGNARD	CH_3MgI
1908	L. MOND	$Co_2(CO)_8$
1919	F. HEIN	$(\eta^6\text{-Arene})_2Cr$
1922	T. MIDGLEY, T. A. BOYD	$Pb(C_2H_5)_4$
1927	A. JOB	$Cr(CO)_6$, $W(CO)_6$
1928	W. HIEBER	Metal carbonyl compounds
1943	E. G. ROCHOW	Organochlorosilanes
1951	P. PAUSON, S. A. MILLER	$Fe(C_5H_5)_2$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_215

Table 2. Continued.

Year	Discoverer	Compound
1955	E. O. Fischer	$Cr(C_6H_6)_2$
1961	G. Wilke	$Ni(\eta^3\text{-}C_3H_5)_2$
1964	E. O. Fischer	$(CO)_5WC(OMe)Me$
1968	A. Streitwieser	$U(C_8H_8)_2$
1972	H. Werner	$[(C_5H_5)_3Ni_2]^+$
1973	G. Wilke	$Ni(\eta^2\text{-}C_2H_4)_3$
1973	E. O. Fischer	$I(CO)_4Cr(CR)$
1979	H. C. Brown	Organoborane
	G. Wittig	Methylenphosphorane
1981	R. West	$R_2Si{=}SiR_2$

Closely linked to the development of synthetic organometallic chemistry is the development of heterogeneous and homogeneous catalysis. Some noteworthy discoveries in this field are listed in Table 3.

2. Organometallic Compounds of the Main Group Elements [13–34]

2.1. Group 1

Organolithium compounds (→ Lithium and Lithium Compounds, Chap. 6.2.) are usually prepared from lithium metal or from commercially available *n*-butyllithium. In the first case, lithium metal is treated with an organic halide to give an organolithium compound and a lithium halide. In the second case, a metallation reaction, *n*-butyllithium is treated with an organic compound containing an acidic C – H group, forming an organolithium compound with liberation of butane. In their chemical behavior, the organolithium compounds resemble the Grignard reagents; they are, however, more reactive. Organolithium compounds have gained importance industrially as initiators for the polymerization of isoprene, in which a polyisoprene that is closely related to natural rubber is formed.

The organic derivatives of the heavier alkali metals are of only minor importance. The ionic character of the metal – carbon bond increases from lithium to cesium. Methylsodium [*18356-02-0*] and methylpotassium [*17814-73-2*] are so reactive that they can even metallate alkanes.

2.2. Group 2

Because of their high air and moisture sensitivity and extreme toxicity, the organoberyllium compounds play only a minor role in synthetic chemistry. In contrast, the organomagnesium halides, or Grignard reagents, are of major importance (see also → Magnesium Compounds, Chap. 6.). They are synthesized from iodine-activated magnesium turnings and an or-

Table 3. Historical development of transition-metal catalysis

Year	Discoverer	Catalytic reaction	Metal or metal complex
1821	J. W. Döbereiner	H_2 oxidation	Pt
1897	P. Sabatier, J. P. Senderens	Hydrogenation of ethene	Ni
1902	P. Sabatier, J. P. Senderens	Synthesis of methane from CO/H_2	Ni
1922	F. Fischer, H. Tropsch	Synthesis of synthol from CO/H_2	Fe
1923	A. Mittasch, M. Pier	Synthesis of methanol from CO/H_2	Zn/Cr
1938	M. Calvin	Homogeneous hydrogenation of benzoquinone	CuOAc
1938	O. Roelen	Hydroformylation of ethene	$Co_2(CO)_8$
1939	W. Reppe	Alkyne oligomerization	$Ni(CN)_2$
1939	W. Reppe	Alkyne hydrocarboxylation	$Ni(CO)_4$
1949	I. Wender	Methanol homologation	Co complexes
1951	P. Arthur, B. C. Pratt	Butadiene hydrocyanation	Ni complexes
1955	K. Ziegler, G. Natta	Alkene polymerization	$TiCl_4/AlEt_3$
1959	F. Smidt	Wacker oxidation	$PdCl_2/CuCl$
1964	R. L. Banks, G. C. Bailey	Alkene metathesis	W/Mo
1965	G. Wilkinson	Selective homogeneous hydrogenation	$(PPh_3)_3RhCl$
1970	Monsanto	Acetic acid synthesis	$[Rh(CO)_2I_2]^-$
1972	Shell	α-Olefin synthesis (SHOP)	Ni chelates
1976	Y. Inoue	C – C bond formation with CO_2	PdL_n
1976	W. N. Lipscomb	Binding properties of borans	
1981	R. Hoffmann K. Fukui	Isolobalanalogs	
1983	R. G. Bergmann W. A. G. Graham	″C-H activation″	

ganic halide RX. In ethereal solution the Grignard reagents exist in the Schlenk equilibrium:

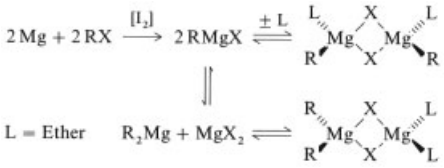

The reactivity of the Grignard reagents is demonstrated by their behavior towards compounds having active protons. For example, the action of water leads to liberation of an alkane; the same is true for reactions with alcohols, carboxylic acids, and amines. In the case of methylmagnesium iodide [*917-64-6*], this reaction can be used to determine the number of active protons in an organic compound by measuring the amount of methane released gas volumetrically (Zerewitinow reaction).

The same reaction is used for selective deuterium labelling of alkanes with D_2O; for example:

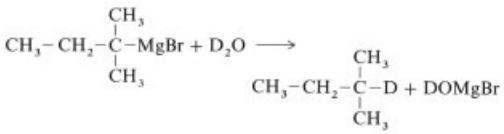

Also of importance is the reaction of Grignard reagents with compounds that contain polar multiple bonds, such as C=O, C=N, C≡N, or C=S. Thus, reaction with carbon dioxide followed by hydrolysis affords carboxylic acids, long-chain aldehydes give secondary alcohols, formaldehyde affords primary alcohols, and ketones afford tertiary alcohols. Some important reactions of Grignard reagents, together with their hydrolysis products, are as follows:

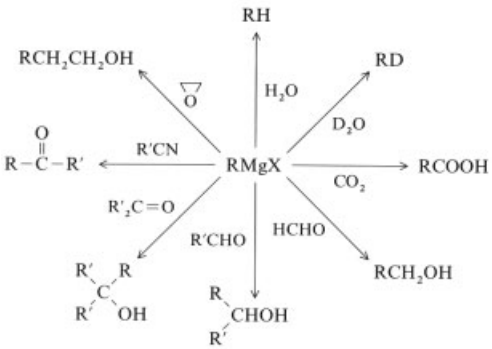

Apart from the Grignard reagents, numerous other organomagnesium compounds are known. Examples are magnesium butadiene [*70809-00-6*] and magnesium anthracene [*50594-04-2*], which are formed by reaction of magnesium with C=C double bond systems.

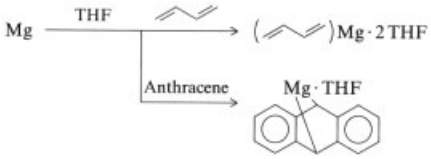

2.3. Group 12

The first organozinc compounds were prepared by E. Frankland in 1849. For example, diethylzinc [*557-20-0*] is formed by reaction of zinc with ethyl iodide, after distillation of the initially formed ethylzinc iodide [*999-75-7*]:

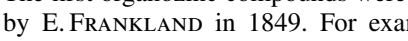

Organozinc compounds are also of importance in the Reformatsky reaction, in which α-halo carboxylic acid esters are converted to organozinc derivatives, which then add to the C=O double bond of aldehydes or ketones. The organozinc intermediate exists as a dimer:

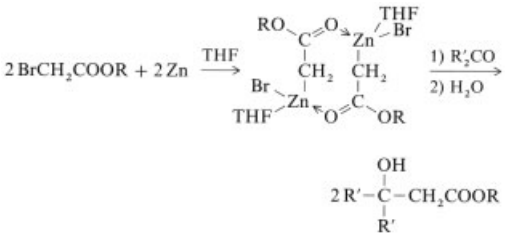

A further synthetic application of organozinc compounds is the cyclopropanation of alkanes (Simmons – Smith reaction). The key starting material for this reaction is (iodomethyl)zinc iodide [*4109-94-8*], which can be obtained by reaction of zinc with diiodomethane. The Simmons – Smith reaction is stereospecific, so that, for example, *cis*-alkenes give exclusively *cis*-cyclopropane derivatives. The attack of the carbenoid on the double bond does not involve the

free carbene, but instead occurs by a concerted reaction:

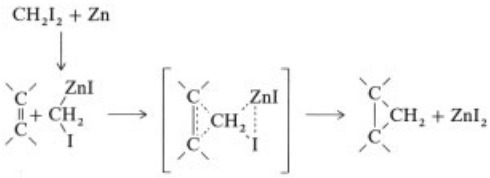

Organocadmium compounds are very similar to the organozinc compounds, but somewhat less reactive. Organomercury compounds are unusual in that the inert, covalent mercury–carbon bond is very stable to water and air. For this reason, organomercury compounds were used in the past as pharmaceuticals, fungicides, and bactericides. These uses have, however, declined drastically due to the toxicity of these compounds. Organomercury compounds have some importance in organic synthesis. In mercuration reactions, aromatic, heteroaromatic, or aliphatic compounds which contain acidic hydrogen atoms are reacted with mercury(II) acetate to form organomercury derivatives that can undergo a range of subsequent reactions. The hydroxymercuration of alkenes with mercury(II) acetate in the presence of water leads to regioselective formation of (hydroxyalkyl)-mercury compounds, which can be converted to alcohols by reaction with sodium borohydride in situ. The overall reaction corresponds to markovnikov addition of water across the C=C double bond:

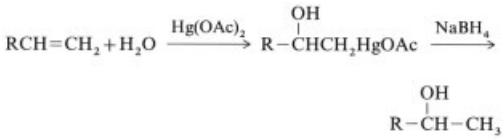

In the Seyferth dihalocyclopropanation reaction, dihalocarbenes are formed from organomercury compounds and then add to unsaturated compounds, for example, cyclohexene:

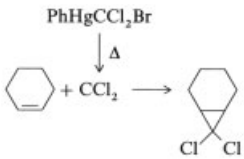

2.4. Group 13

Organoboron compounds can be synthesized by reaction of boron trifluoride with Grignard reagents or through addition of diborane [*14044-65-6*], B_2H_6, to alkenes, the hydroboration reaction (cf. Section 3.2.8):

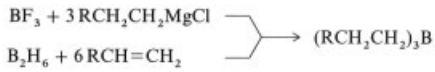

Trialkylboranes are very sensitive to oxidation and may undergo self ignition on contact with atmospheric oxygen. Therefore, trialkylboranes are usually reacted further in situ, for example, by protolysis to give alkanes or by oxidation to alcohols with hydrogen peroxide.

Since the hydroboration of unsymmetric alkenes always affords the anti-markovnikov product, the subsequent oxidation with hydrogen peroxide leads to primary alcohols. Overall, the reaction can be regarded as anti-markovnikov addition of water to an alkene:

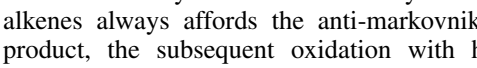

In addition to high regioselectivity, hydroboration also exhibits high stereoselectivity for *cis* addition. Particularly high selectivity is exhibited by organoboron compounds with sterically demanding substituents, such as 9-borabicyclo[3.3.1]nonane (9-BBN) [*280-64-8*], which is synthesized from diborane and 1,5-cyclooctadiene [*1552-12-1*]. Since hydroboration is a reversible reaction, alkylboron compounds can undergo elimination of alkenes at higher temperatures. This reaction can be used for the formally contra-thermodynamic isomerization of alkenes:

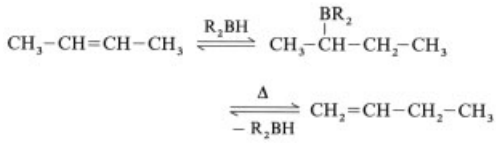

In addition to acyclic boranes, numerous cyclic boranes are also known. Some important examples are the aforementioned 9-BBN, borolane [*1692-31-5*], borole [*287-87-6*], borinane [*289-53-2*], and borobenzene [*31029-61-5*], which generally require stabilization by substitution.

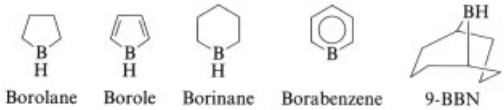

Aluminum alkyls are prepared on an industrial scale. An important method of synthesis is the Ziegler direct synthesis from aluminum, hydrogen, and α-olefins, a cyclic process involving "growth" and "addition" reactions. The two-step process can be summarized as shown here for triethylaluminum [*97-93-8*]:

Growth reaction:

$$Al + 1.5\,H_2 + 2(C_2H_5)_3Al \longrightarrow 3(C_2H_5)_2AlH$$

Addition:

$$3(C_2H_5)_2AlH + 3\,C_2H_4 \longrightarrow 3(C_2H_5)_3Al$$

Overall reaction:

$$Al + 1.5\,H_2 + 3\,C_2H_4 \longrightarrow (C_2H_5)_3Al$$

Aluminum alkyls are also produced industrially by reaction of aluminum or aluminum–magnesium alloys with alkyl halides.

The binary aluminum alkyls are very air and moisture sensitive liquids. The short-chain alkyls undergo self-ignition and react explosively with water. Aluminum alkyls form dimers, which can be considered to contain sp^3 hybridized aluminum and carbon atoms. For example, trimethylaluminum [*75-24-1*] has two bridging methyl groups:

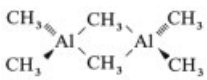

Organoaluminum compounds have various uses in industrial chemistry. An important application is in the Ziegler growth reaction (see figure below). In this reaction, ethene inserts into Al–C bonds, after which either α-olefins are formed in a subsequent displacement reaction (Alfen process, see → Hydrocarbons) or the chain-growth products are first oxidized and then converted to long-chain primary alcohols by hydrolysis (Alfol process, → Fatty Alcohols, Chap. 2.4.1.).

The second important application is the use of aluminum alkyls in Ziegler–Natta catalysis (cf. Section 3.2.4). Ziegler–Natta catalysts are composed of a transition-metal compound (e.g., $TiCl_4$, VCl_3, or Cp_2TiCl_2) and an organoaluminum compound, which may also contain halide substituents (e.g., $AlEt_3$, $AlEt_2Cl$, or $AlEtCl_2$).

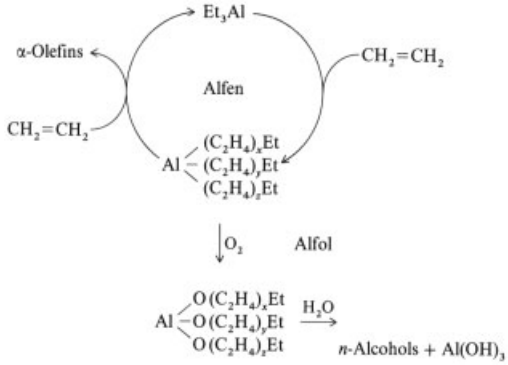

Aluminum alkyls are also used for the dimerization of alkenes. For example, in the presence of AlR_3, propene undergoes dimerization to give 2-methyl-1-pentene, which can be converted to isoprene by isomerization and elimination of methane.

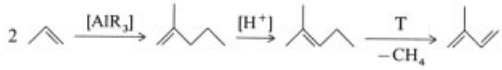

The Dimersol process of the Institut Français du Pétrol (IFP) (→ Hydrocarbons) involves the dimerization of propene or the codimerization of propene and butene in the presence of catalysts consisting of aluminum alkyls and nickel compounds. After hydrogenation, the products are effective octane enhancers for motor fuels.

A further important use of aluminum trialkyls is as alkylating agents, for example, in the synthesis of organozinc and organotin compounds. The addition of alkyl aluminum hydrides to double bonds (hydroalumination) is discussed in Section 3.2.8.

Organometallic compounds of gallium and indium are used for doping semiconductors. Coatings of gallium arsenide can be prepared by thermal decomposition of a mixture of trimethylgallium [*1445-79-0*] and arsine.

$$GaMe_3 + AsH_3 \xrightarrow{\Delta} GaAs + 3\,CH_4$$

2.5. Group 14

The organosilicon compounds (→ Silicon Compounds, Organic) are of major industrial importance. They can be obtained by reaction of silicon tetrachloride [*10026-04-7*] with Grignard reagents. The Müller–Rochow synthesis was especially developed for the production of methylchloro-silanes.

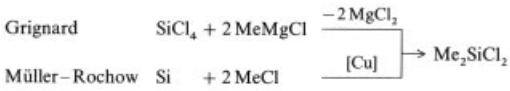

The direct Müller–Rochow process always gives a mixture of all possible methylchlorosilanes. These can, however, be separated by distillation. The reaction of the chloromethane takes place on the surface of a copper–silicon alloy, which is prepared from nine parts silicon and one part copper in a sinter furnace.

The methylchlorosilanes are used for the production of silicones (→ Silicones). Hydrolysis of the isomeric mixture of methylchlorosilanes leads to intermediate silanols, which undergo polycondensation to form linear chains and networks of Si–O bonds, whereby the difunctional dimethyldichlorosilane [*75-78-5*] acts as a chain-extender, monofunctional trimethylchlorosilane [*75-77-4*] as a chain-terminator, and trifunctional methyltrichlorosilane [*75-79-6*] as a network-former. The silicones are used as oils, resins, and elastomers.

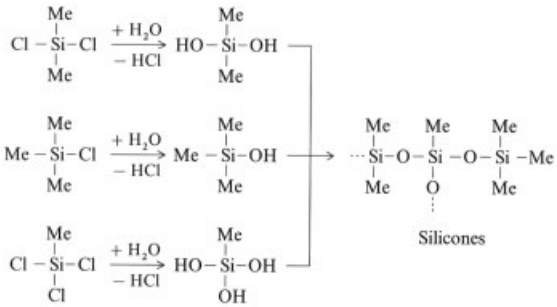

More recently, organosilicon compounds have been used for the production of ceramics. For example, alkyl- and arylchlorosilanes can be converted by dechlorination to polycarbosilanes, which can be melt-spun into ceramic fibers. The final product is a β silicon carbide with exceptionally high tensile strength.

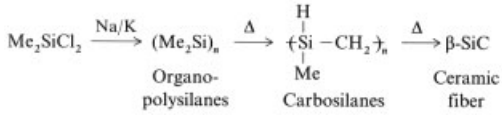

The organometallic compounds of germanium have no industrial uses as yet. Organotin compounds, in contrast, are readily accessible and have versatile applications. They can be prepared by reaction of $SnCl_4$ with Grignard reagents, with aluminum alkyls, or with alkyl halides in the presence of sodium.

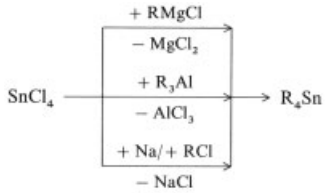

Other important preparation methods are the direct synthesis from tin and alkyl halides and the addition of organotin hydrides to double bonds (hydrostannation). The most important uses of organotin compounds are as stabilizers for PVC, as biocides in industry and agriculture, and as catalysts for the synthesis of polyurethanes. In PVC, organotin compounds suppress the elimination of HCl that occurs during heat treatment. Furthermore, they prevent the photochemical decomposition of the polymer by terminating UV-induced radical chain reactions (→ Plastics, Additives). Organotin compounds such as bis(tributyltin) oxide [*56-35-9*] are used as industrial biocides. Owing to their antibacterial and fungicidal properties, they are used as antifouling agents for wood and for disinfecting circulating cooling water in industrial plants. Triphenyltin acetate [*900-95-8*] and tris(cyclohexyl)tin hydroxide [*13121-70-5*] are used in agriculture as acaricides and for the treatment of algal infestations in rice and potato rot. Important for this application is that the toxicity of the organotin compounds decreases with decreasing extent of alkylation. The organotin compounds undergo degradation in nature as follows:

$$R_4Sn \longrightarrow R_3SnX \longrightarrow R_2SnX_2 \longrightarrow RSnX_3 \longrightarrow SnX_4$$

The final product of this sequence is nontoxic tin dioxide.

Industrially, the most important organolead compound is tetraethyllead [*78-00-2*], which is

produced by reaction of sodium – lead alloy with ethyl chloride under pressure.

$$4\,PbNa + 4\,EtCl \longrightarrow PbEt_4 + 4\,NaCl + 3\,Pb$$

The lead formed in the reaction is reconverted with sodium to the alloy NaPb. An alternative method is the electrochemical Nalco process, in which an ethyl Grignard is converted to tetraethyllead at a lead anode. Tetraethyllead is used as an antiknock for gasoline. Its effectiveness is due to the deactivation of hydroperoxides by lead oxide. However, because of its toxicity its use as an octane enhancer has greatly declined. In the long term, a total ban on lead in gasoline can be expected.

3. Transition Metals

3.1. Transition-Metal Complexes [35–94]

In addition to *s*- and *p*-orbitals, transition metals also possess *d*-orbitals for the formation of metal-carbon bonds. The partial occupation of the *d*-orbitals means that transition metals can act as electron donors and as electron acceptors. A ligand can therefore be bound through both a σ-bond and a π-back-bond. Moreover, transition metals can form direct or ligand-bridged metal – metal bonds, allowing the formation of clusters (see Section 3.1.3). Table 4 gives a survey of transition metal complexes according to ligands.

Table 4. Survey or organometallic transition-metal complexes

Type	Bonding modes (examples)
Alkyls	$M-CR_3$
Aryls	M–
Carbenes	$M=CR_2$
Carbynes	$M\equiv CR$
Carbonyls	$M-C\equiv O$ 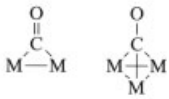
π-Complexes:	
Alkene complexes	$M-\parallel^{CR_2}_{CR_2}$
Alkyne complexes	$M-\equiv^{CR}_{CR}$ $M\langle^{CR}_{CR}$
Allyl complexes	
Cyclopentadienyl complexes	
Arene complexes	

3.1.1. Alkyls and Aryls

The transition-metal alkyls contain only σ-bonds between the metal and carbon. Only a few of the various methods of synthesis are described here. Transmetallation is the reaction of a metal halide with a main-group metal alkyl, for example, a lithium alkyl or a Grignard reagent. The lithium reagents react more quickly than the Grignards and give the higher alkyls. Metallacycles are accessible by using dilithium alkyls.

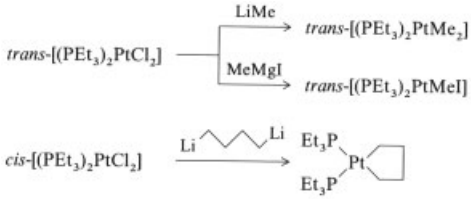

Another synthetic method for the preparation of metal alkyls is the insertion of alkenes into metal-hydride bonds (hydrometallation).

$$(CO)_4CoH + CH_2{=}CHR \longrightarrow (CO)_4Co-CH_2CH_2R$$

This reaction plays a role in many important homogeneous catalytic processes. Some transition-metal hydrides can be converted to alkyl complexes by reaction with diazomethane.

$$CpMo(CO)_3H + CH_2N_2 \longrightarrow CpMo(CO)_3CH_3$$

A further important method is the oxidative addition of alkyl halides to 16-electron complexes, a reaction in which the oxidation state of the metal increases by two. Aryl complexes can be obtained by the same method.

$$(PPh_3)_2Ir(CO)Cl + CH_3I \longrightarrow (PPh_3)_2Ir(CO)(CH_3)ICl$$

$$Ni(PPh_3)_4 + PhCl \longrightarrow (PPh_3)_2Ni(Ph)Cl + 2\,PPh_3$$

Many transition-metal alkyls are, however, unstable and can decompose by β-elimination, giving a metal hydride and an alkene (i.e., the reverse of hydrometallation).

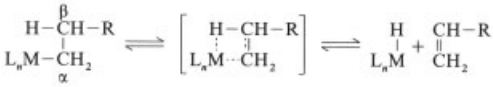

Alkyl complexes can be stabilized by using alkyl groups that have no β-hydrogen atoms, for instance, neopentyl or benzyl.

3.1.2. Carbenes and Carbynes

Metal carbene complexes are compounds having metal – carbon double bonds. The first carbene complex was prepared by E. O. FISCHER in 1964 by addition of a lithium alkyl to tungsten hexacarbonyl, giving a complex with a heteroatom substituted alkylidene ligand (Fischer carbene complex).

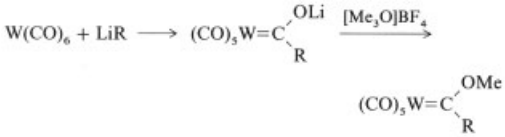

In 1975, SCHROCK found a route to C – H substituted carbene complexes (Schrock carbene complexes). The synthesis is based on the α-elimination of a hydrogen atom from an alkyl group.

Transition-metal carbene complexes have received much attention because they are intermediates in homogeneous catalytic metathesis reactions (see Section 3.2.5). Carbene complexes are also used in organic synthesis. Thus, Tebbe's reagent can be used to convert a C=O group to C=CH_2, analogously to the Wittig reaction.

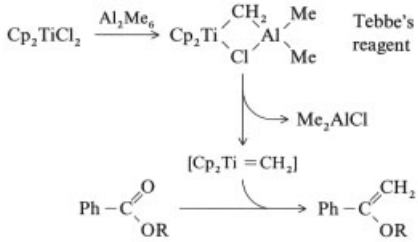

The reactive intermediate is $Cp_2Ti{=}CH_2$ [*83876-46-4*].

The first transition-metal carbyne complexes were also synthesized by Fischer.

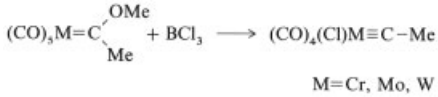

3.1.3. Carbonyls and Clusters

Carbon monoxide bonds to transition metals through both σ-donor and π-acceptor bonds. Since the metals of groups 3 to 5 have either an insufficient number of d-electrons or none at all, they do not form stable carbonyl complexes. Vanadium hexacarbonyl [*14024-00-1*], for example, decomposes at $-70\,°C$. A survey of the neutral binary carbonyls of groups 5 to 10 is given in Table 5.

As shown in Table 4, carbon monoxide exhibits various bonding modes, acting for example as a terminal ligand and as a doubly or triply bridging ligand. This leads to a wide variety of structures, particularly for the polynuclear clusters. Some examples of the structures are given here for iron and cobalt carbonyls.

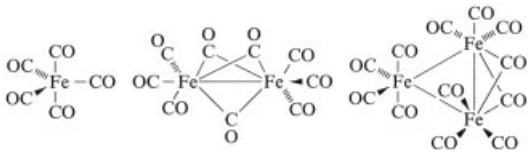

Table 5. Neutral metal carbonyl complexes

5	6	7	8	9	10
$V(CO)_6$	$Cr(CO)_6$	$Mn_2(CO)_{10}$	$Fe(CO)_5$	$Co_2(CO)_8$	$Ni(CO)_4$
			$Fe_2(CO)_9$	$Co_4(CO)_{12}$	
			$Fe_3(CO)_{12}$	$Co_6(CO)_{16}$	
	$Mo(CO)_6$	$Tc_2(CO)_{10}$	$Ru(CO)_5$	$Rh_2(CO)_8$	
		$Tc_3(CO)_{12}$	$Ru_3(CO)_{12}$	$Rh_4(CO)_{12}$	
			$Ru_6(CO)_{18}$	$Rh_6(CO)_{16}$	
	$W(CO)_6$	$Re_2(CO)_{10}$	$Os(CO)_5$	$Ir_4(CO)_{12}$	
			$Os_3(CO)_{12}$	$Ir_6(CO)_{16}$	

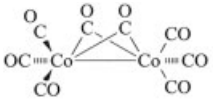

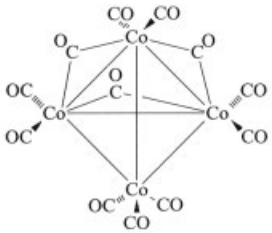

There are several possibilities for synthesizing metal carbonyls. The direct reaction of a metal with carbon monoxide is only possible for nickel and iron. Whereas nickel reacts with carbon monoxide at room temperature to give $Ni(CO)_4$ (→ Nickel Compounds) [*13463-39-3*], the formation of $Fe(CO)_5$ [*13463-40-6*] from iron requires elevated temperature and pressure (→ Iron Compounds, Chap. 3.2.). In most other cases, the synthesis involves the reduction of metal compounds (e.g., halides or oxides) in the presence of carbon monoxide. Typical synthetic methods include:

$$Fe + 5\,CO \longrightarrow Fe(CO)_5$$

$$2\,Fe(CO)_5 \xrightarrow[-CO]{h\nu} Fe_2(CO)_9$$

$$Fe(CO)_5 \xrightarrow{OH^-} [HFe(CO)_4]^- \xrightarrow{MnO_2} Fe_3(CO)_{12}$$

$$2\,CoCO_3 \xrightarrow{CO/H_2} Co_2(CO)_8$$

The metal carbonyls are highly reactive complexes. The CO group can be substituted, either thermally or photochemically, by other ligands such as phosphines and olefins:

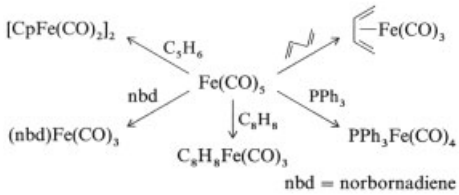

Metal carbonyls can be converted to anionic carbonyl complexes (carbonyl metallates) by means of the Hieber base reaction. In many cases, the carbonyl metallates are converted to metal carbonyl hydrides by protonation.

$$Fe(CO)_5 \xrightarrow{NaOH} Na[HFe(CO)_4] \xrightarrow{H^+} H_2Fe(CO)_4$$

A further important group of compounds are the carbonyl metal halides, which are formed by reaction of carbonyl complexes with halogens.

$$Fe(CO)_5 + I_2 \longrightarrow Fe(CO)_4I_2 + CO$$

$$Mn_2(CO)_{10} + Cl_2 \longrightarrow 2\,Mn(CO)_5Cl$$

3.1.4. π-Complexes

In addition to a ligand to metal donor bond, π-complexes also exhibit a strong metal to ligand acceptor interaction. The alkene complexes of the transition metals are of major importance because of their roles as intermediates in homogeneous catalysis (cf. Section 3.2). The first alkene complex, the ionic platinum complex $K[(C_2H_4)PtCl_3]$ [*12012-50-9*], was prepared by Zeise in 1827. Alkene complexes can be synthesized by addition of alkenes to coordinatively unsaturated complexes.

$$IrCl(CO)(PPh_3)_2 + R_2C{=}CR_2 \longrightarrow (PPh_3)_2Ir(CO)Cl(R_2C{=}CR_2)$$

Coordinatively saturated complexes can be converted to alkene complexes by ligand displacement.

An important method is the reduction of metal salts or complexes in the presence of alkenes. This method is used to prepare the nickel(0) complexes (cyclododecatriene)-nickel [*39330-67-1*], bis(1,5-cyclooctadiene)-nickel [*1295-35-8*], and tris(ethene)nickel [*50696-82-7*].

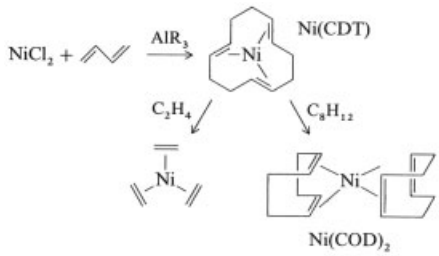

In alkyne complexes, the alkynes can act either as monodentate or as bidentate ligands (see

Table 4). The bonding mode is indicated by the C–C bond length of the coordinated alkyne. Monodentate alkynes have C≡C bond lengths of ca. 120 pm, while bidentate alkynes have a C=C bond length of ca. 134 pm. Therefore, bidentate alkynes, for example in $[(PPh_3)_2Pt(Ph_2C_2)]$, are better described as metallacyclopropenes. Alkyne complexes are prepared similarly to alkene complexes, for example, by substitution of carbonyl ligands. Alkyne complexes play an important role as intermediates in homogeneous catalytic oligomerization (see Section 3.2.4) and in the reaction of alkynes with carbon monoxide (Section 3.2.7) or carbon dioxide (Section 3.2.8). Other reactions involving alkyne complexes are the cyclization of alkynes and nitriles to give pyridines, developed by Yamazaki and Bönnemann:

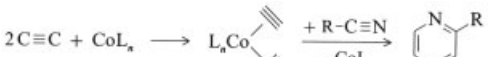

and the Pauson–Khand reaction, in which an alkyne, an alkene, and carbon monoxide react to give a substituted cyclopentenone.

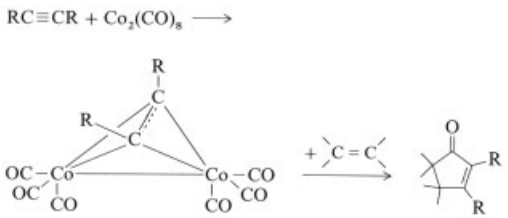

Allyl complexes contain the hydrocarbon ligand C_3H_5, in which the protons may also be replaced by organic substituents. In η^1-allyl complexes, a metal atom is bound to one carbon atom through a σ-bond, while in η^3-allyl complexes three electrons are distributed over three carbon atoms, all of which are bound to the metal center (see Table 4). Allyl complexes are important intermediates in homogeneous catalytic isomerization reactions (see Section 3.2.2) and in the oligomerization and polymerization of 1,3-dienes (see Section 3.2.4). Binary allyl complexes, which can be prepared from transition-metal salts and allyl Grignard reagents, are usually unstable and must therefore be handled at low temperature.

$$NiBr_2 + 2\,C_3H_5MgBr \longrightarrow Ni(C_3H_5)_2 + 2\,MgBr_2$$

Halogen-containing allyl complexes can often be prepared simply from reactions of allyl halides with metal compounds.

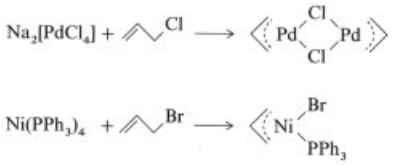

Another route to allyl complexes is the addition of metal hydride complexes to 1,3-dienes. An example is the reaction of hydridocobalt tetracarbonyl with butadiene to give a crotyl complex, which can exist in *syn* and *anti* forms.

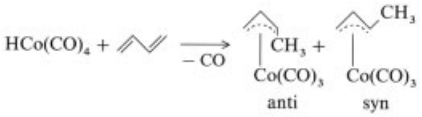

Although η^5-cyclopentadiene complexes have found no major practical application, they are of interest because of the special electronic properties of the η^5-cyclopentadienyl (Cp) ligand, the wide variety of structures, and their versatile reactivity. Structurally, the simplest Cp complexes are the sandwich complexes, for example, ferrocene [*102-54-5*], $FeCp_2$, which was discovered in 1951. Other examples of the structural variety of Cp complexes are the half-sandwich complexes, triple-decker complexes, MCp_4 complexes, and complexes with bent sandwich structures.

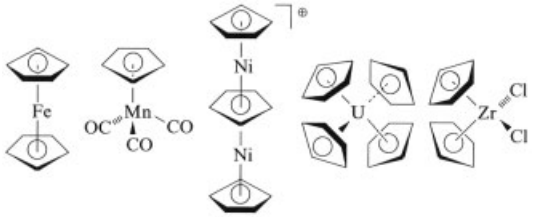

Binary cyclopentadienyl complexes, or metallocenes, are synthesized by reaction of metal salts with a cyclopentadienyl compound (e.g., NaCp or CpMgBr) or with cyclopentadiene, sometimes in the presence of a base.

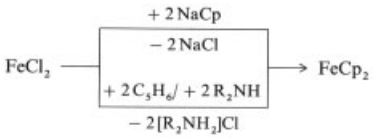

The Cp ligands of metallocenes readily undergo electrophilic substitution reactions. The chemistry of ferrocene, in particular, has been extensively investigated. Arene complexes are metal complexes of benzene or its derivatives. The most important synthetic route is the reduction of a metal salt by Al and $AlCl_3$ in the presence of an aromatic hydrocarbon. This method was first used by FISCHER and HAFNER for the synthesis of bis(η^6-benzene)chromium [*1271-54-1*]. The intermediate in the reaction is a cationic chromium complex, which is converted to the neutral benzene complex by reduction with sodium dithionite or through disproportionation.

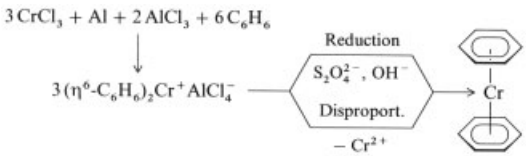

Numerous complexes of benzene, mesitylene, and hexamethylbenzene have been prepared by this method. The synthesis of complexes of polynuclear aromatic hydrocarbons such as naphthalene, anthracene, and phenanthrene requires the use of metal – ligand cocondensation techniques, in which a metal vapor reacts with an arene in the vapor phase, and the resulting complex is condensed on a cooled surface.

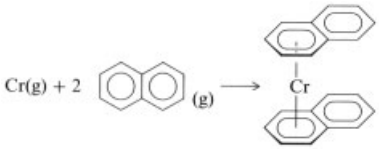

Bis(arene) complexes undergo a wide range of reactions. The aromatic ligands can be substituted via lithiation reactions and displaced by other arene ligands. For example, the naphthalene ligands in bis(naphthalene)chromium [*33085-81-3*] undergo partial exchange with benzene. The arene complexes are, however, of no industrial importance.

3.2. Transition-Metal Catalysis

Homogeneous transition-metal catalysis exploits the versatile reactivity of transition-metal complexes for the synthesis of organic molecules. The critical intermediates in the catalytic cycles are mostly the organo transition metal complexes described in Section 3.1.

The advantage of homogeneous catalysis over heterogeneous catalysis is that the selectivity of the reaction can be controlled by appropriate choice of the ligands on the catalytic metal center. Organometallic catalysis often allows the synthesis of organic compounds that are otherwise only accessible through complex multistep syntheses. Also, homogeneously catalyzed enantioselective reactions with high optical yields are carried out industrially. Noteworthy is that homogeneous catalysis can often be achieved under mild conditions of temperature and pressure. A drawback to these processes is the expense of the required noble metals, for example, rhodium (see Table 6). In spite of the high cost of these metals, many bulk chemicals, such as butyraldehyde and acetic acid, are produced by means of homogeneous transition-metal catalysis at costs of a few cents per mole. This is possible when a large quantity of product is formed for each mole of catalyst; that is, when the reaction has a high turn-over rate. Furthermore, there are many methods by which the catalyst can be recovered almost quantitatively after the reaction, including precipitation, reverse osmosis, immobilization on carrier materials, and reaction in two-phase liquid – liquid systems [95–185].

Table 6. Price of transition metals, $/mol ∗

4	5	6	7	8	9	10	11
Ti	V	Cr	Mn	Fe	Co	Ni	Cu
1.44	1.01	0.78	0.08	0.07	2.83	0.39	0.16
Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag
9.87	17	2.88		173	897	508	19
Hf	Ta	W	Re	Os	Ir	Pt	Au
28	52	2.76		3161	1554	2369	2398

∗ In February 1997.

3.2.1. Catalytic Elementary Reactions [186–233]

Fundamentally speaking, a homogeneous catalytic cycle can be divided into the following steps: formation of the active species (I), coor-

dination of the reactants at the metal center (II), formation of the product (III), and elimination of the product (IV) with reformation of the initial catalytic species.

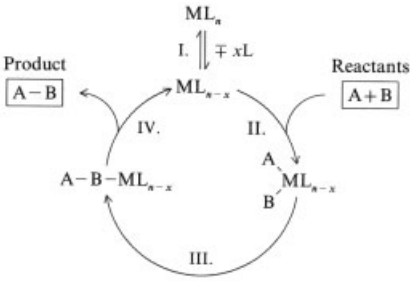

The first step involves in situ formation of the catalytically active species ML_{n-x} from the starting complex ML_n. The active species possesses vacant coordination sites, which is a prerequisite for coordination of the reactants A and B. The active species is usually formed by ligand dissociation; that is, by elimination of Lewis acids (e.g., protons) or Lewis bases (e.g., phosphines).

In the second step, one or more reactants bind to the metal center. This can occur either by ligand association, in which, for example, alkenes, alkynes, or carbon monoxide add to the metal center, or by oxidative addition, in which molecules such as H – H, H – X, or R – X are cleaved and the two fragments bind to the metal by σ-bonds. In the latter case, both the coordination number and the oxidation state of the metal increase by two. An example of oxidative addition is the addition of dihydrogen to the square-planar complex $[Ir(CO)(PPh_3)_2Cl]$ (Vaska's complex). The addition of H_2 to the Ir(I) complex occurs in a synchronous bimolecular reaction, giving an octahedral Ir(III) complex with *cis* hydride ligands.

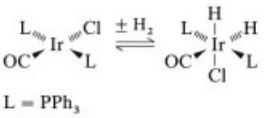

The third step of the catalytic cycle involves bond-forming reactions between the reactants bound to the metal center. Insertion reactions, in which for example ligand B inserts into the metal – A bond are often observed. Examples include the insertion of carbon monoxide into metal – alkyl bonds to give acyl complexes, the insertion of ethylene into metal hydrogen bonds to give ethyl complexes, and the synthesis of crotyl complexes of cobalt through insertion of butadiene into cobalt – hydrogen bonds.

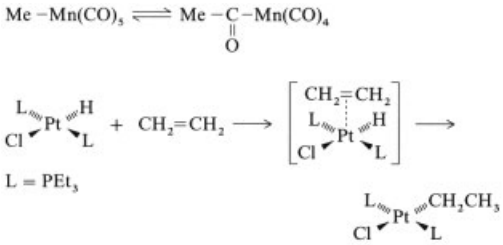

In some cases insertion reactions are reversible; the reverse reaction is termed extrusion. Also noteworthy are multiple insertion reactions. For example, two CO molecules can insert into a metal – alkyl or metal – aryl bond, offering a catalytic route to α-keto carboxylic acids (see Section 3.2.7).

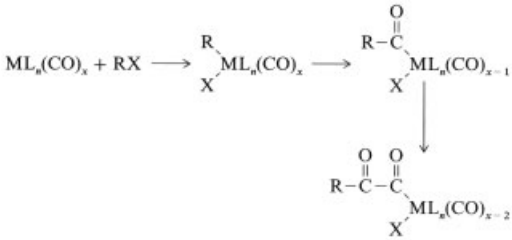

Another important bond-forming reaction is oxidative coupling, which is of particular importance in the dimerization of alkynes, alkenes, and 1,3-dienes (see Section 3.2.4).

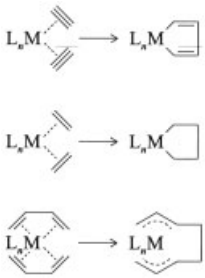

A special case of oxidative coupling is the coupling of metal carbenes with alkenes to give a metallacyclobutane. This has been discussed as the crucial step in the metathesis reaction (Section 3.2.5).

Table 7. Elementary catalytic steps

Elementary catalytic step	General equation	Importance in the catalytic cycle
Association → Dissociation ←	$L_nM + X \rightleftharpoons L_nMX$	Addition and cleavage (Steps I, II, IV)
Oxidative addition → Reductive elimination ←	$L_nM + X-Y \rightleftharpoons L_nM{<}^{X}_{Y}$	
Insertion → Extrusion ←	$L_nMX + Y \rightleftharpoons L_nMYX$	Bond formation (step III)
Oxidative coupling → Reductive cleavage ←	$L_nM{<}^{X}_{Y} \rightleftharpoons L_nM{<}^{X}_{Y}$	

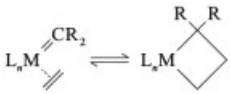

The reverse of oxidative coupling is known as reductive cleavage.

The fourth step of the cycle involves elimination of the product. If the product is bound to the metal through a η^2-bond, this can occur by ligand dissociation.

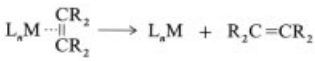

Often, the product is released through reductive elimination, the reverse of oxidative addition. Some examples are as follows:

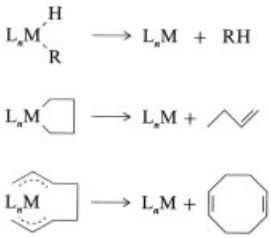

The β-elimination of hydrogen from alkyl complexes, which can be regarded as the reverse of alkene insertion into M–H bonds, leads to formation of alkenes. This reaction is an important chain-terminating step in oligomerization

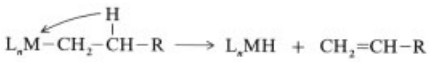

reactions. Table 7 gives a survey of the various catalytic elementary reactions.

The following Sections introduce those homogeneous catalytic reactions which have gained industrial importance. These include isomerization, hydrogenation, oligomerization and polymerization, metathesis, oxidation, and functionalization with molecules such as carbon monoxide. The most important substrates for these reactions are those which can associate with metal complexes and those which can undergo oxidative addition to a metal center; in particular alkenes, alkynes, carbon monoxide, and hydrogen. The wide variety of reactions and the possibility of controlling them by appropriate choice of catalyst allow the synthesis of a wide range of low and high molecular mass hydrocarbons and mono- and polyfunctional compounds. Homogeneous catalysis is used for the synthesis of bulk chemicals such as acetaldehyde, acetic acid, oxoalcohols, α-olefins, and high-density polyethylene and, increasingly, for fine chemicals, including pharmaceuticals, flavors, fragrances, and special agrochemicals. Although organometallic catalysis has a relatively short history, it is already of considerable importance in industrial chemistry. Industrial processes based on homogeneous catalysis reached a turnover of $\$23 \times 10^9$ in the United States in 1990, and new applications can be expected in electronics, metal-containing pharmaceuticals, and specialty polymers.

3.2.2. Isomerization [234–256]

Transition-metal complexes can catalyze the isomerization of double-bond systems such as alkenes, dienes, and allyl compounds and skeletal rearrangements of cyclic aliphatic compounds. Transition-metal hydride complexes of general formula L_nMH, which may also be formed in situ from ML_n and hydrogen, are often used as homogeneous isomerization catalysts.

The driving force for isomerization is the formation of thermodynamically more stable iso-

mers. Isomerization is an equilibrium reaction that is strongly influenced by the reaction temperature. At higher temperature, the proportion of the thermodynamically less stable isomer increases. In unsaturated compounds, the stability of the individual structural units generally follows the following rules:

1) Inner alkenes are more stable than terminal alkenes

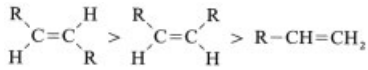

2) Highly branched alkenes are more stable than less branched alkenes

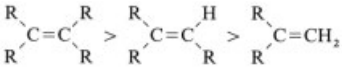

3) Conjugated dienes are more stable than unconjugated dienes

$$R-CH=CH-CH=CH-R > R-CH=CH-(CH_2)_n-CH=CH-R$$

4) In olefins with functional groups, those having conjugated multiple bonds are particularly stable:

$$R-CH_2-CH=CH-X > R-CH=CH-CH_2-X$$
$$(X = COOH, COOR, CN, Ph, \ldots)$$

Two mechanisms are discussed in alkene isomerization. In the hydride mechanism, the alkene adds to the transition-metal hydride to give an η^2-alkene complex, which rearranges to give a σ-alkyl complex. Elimination of the product olefin then leads to reformation of the initial hydride complex.

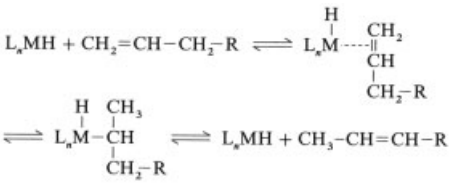

In the allyl mechanism, the transition metal removes a proton adjacent to the double bond to form an η^3-allyl complex as isomerization intermediate.

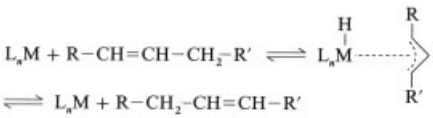

Numerous complexes of metals of groups vier, 6, 8, 9, and 10 can be used as isomerization catalysts. Rhodium chloride, rhodium chloro complexes, iron carbonyls, and palladium and titanium complexes are most widely used in practice.

Alkene Isomerization. 1-Hexene can be converted to 2- and 3-hexene in the presence of iron carbonyls:

Rhodium trichloride [*10049-07-7*] is an efficient catalyst for the isomerization of exocyclic enones into endocyclic enones. This allows the synthesis of 2-substituted cyclopentenones:

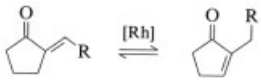

Sesquiterpene-α-methylene-γ-butyrolactones such as damsin [*1216-42-8*] have received much attention because of their antitumor activity. A selective isomerization of damsin [*1216-42-8*] to isodamsin [*17081-89-9*] can be achieved with the rhodium hydride complex [$(PPh_3)_2RhH_2Cl$] [*12119-41-4*]. Since Wilkinson's catalyst [$(PPh_3)_3RhCl$] [*14694-95-2*] is inactive in this isomerization, the reaction probably occurs through addition of Rh – H to the exocyclic double bond:

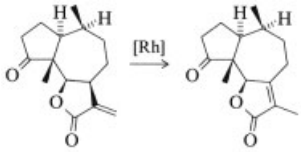

Damsin Isodamsin

A new synthesis of L-menthol [*2216-51-5*] is based on the selective isomerization of nerylamine to give a specific enamine isomer. The catalyst is a complex of rhodium with the (+)-isomer of the ligand BINAP:

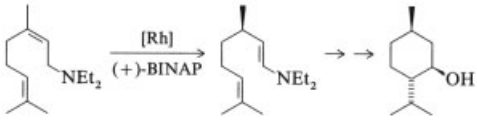

Nerylamine Enamine L-Menthol

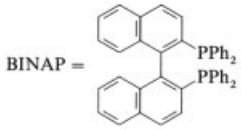

Alkyne Isomerization. Alkynes can be converted to the corresponding allenes by means of manganese complexes:

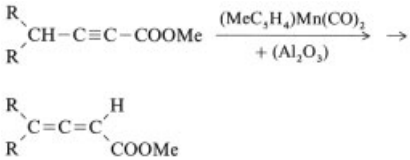

A process that has found practical application is the vanadium-catalyzed isomerization of dehydrolinalool to give citral. Whereas alkyl vanadate catalysts require temperatures above 160 °C, oxovanadium(IV) acetylacetonate [*3153-26-2*] is active at room temperature:

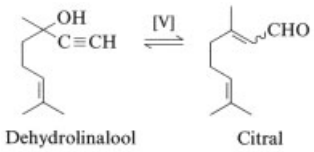

Diene Isomerization. 1,5-Cyclooctadiene [*1552-12-1*] is converted to the thermodynamically more stable 1,3-isomer by iron pentacarbonyl [*13463-40-6*]:

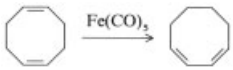

The same reaction can be achieved with hydrido iridium catalysts.

Skeletal Isomerization of Ring Molecules. Molecules with small rings (e.g., cyclopropanes, cyclobutanes, epoxides) can undergo isomerization of their molecular frameworks in the presence of transition-metal catalysts. Typical examples are the isomerization of bicyclo[2.1.0]-pentane [*185-94-4*] to cyclopentene with $[Rh(CO)_2Cl]_2$ [*14523-22-9*], and the isomerization of cyclohexene epoxide [*286-20-4*] to cyclohexanone [*108-94-1*]:

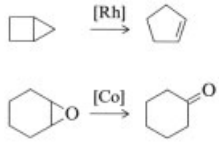

Also of interest is the rearrangement of cubane [*277-10-1*] to give cyclooctatetraene [*629-20-9*] (COT) and of quadricyclane [*278-06-8*] to form norbornadiene [*121-46-0*]. Although both of these isomerizations are forbidden by the Woodward–Hoffmann rules, they can be carried out in the presence of transition metals.

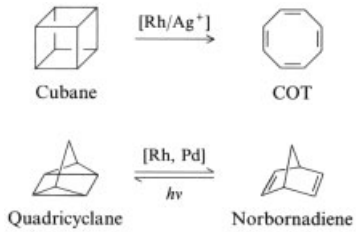

Homogeneous isomerization has found only limited use in industry. However, it does play an important role as a side reaction in hydroformylation, hydrocarboxylation, and addition reactions of alkenes.

3.2.3. Hydrogenation [257–287] (see also → Hydrogenation and Dehydrogenation)

The hydrogenation of alkenes, dienes, and alkynes can be carried out with homogeneous transition-metal catalysts. Typical hydrogenation catalysts are H_2IrCl_6 [*10025-83-9*], $[(PPh_3)_2IrCl(CO)]$ [*14871-41-1*], $[HCo(CO)_4]$ [*16842-03-8*], and $[(PPh_3)_2(CO)_2RuCl_2]$ [*14564-35-3*]. One of the most widely used catalysts is Wilkinson's catalyst $[(PPh_3)_3RhCl]$, which in many cases allows the reaction to be carried out under mild conditions. A simplified reaction scheme for a hydrogenation reaction with Wilkinson's catalyst is as follows:

Oxidative addition of hydrogen gives a dihydrido rhodium (III) complex, which through dissociation of a phosphine ligand provides a vacant coordination site for bonding of an alkene. An insertion step then gives a hydrido alkyl rhodium complex, which upon elimination of an alkane and recomplexation of the phosphine gives the starting complex.

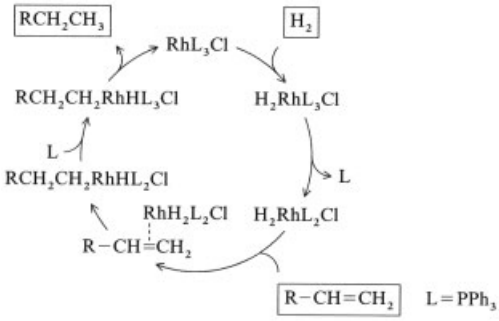

Whereas $[(PPh_3)_3RhCl]$ catalyzes the hydrogenation of terminal and inner alkenes, the hydride complex $[(PPh_3)_3Rh(CO)H]$ [*22172-51-6*] hydrogenates only terminal double bonds; it is also somewhat less reactive. The carbonyl chloro complex *trans*-$[(PPh_3)_3Rh(CO)Cl]$ [*13938-94-8*] catalyzes the hydrogenation of ethene and propene, albeit with only slow rates of reaction. In contrast, the corresponding iridium complex *trans*-$[(PPh_3)_3Ir(CO)Cl]$ is considerably more reactive. Also noteworthy is the high activity of many cationic catalysts; for example, complexes of the type $[L_2Rh(THF)_2]^+$.

Of particular interest is asymmetric hydrogenation, in which prochiral unsaturated substrates are converted to chiral products by means of transition-metal catalysts with optically active diphosphine ligands such as DIOP, CHIRAPHOS, NORPHOS, and DIPAMP:

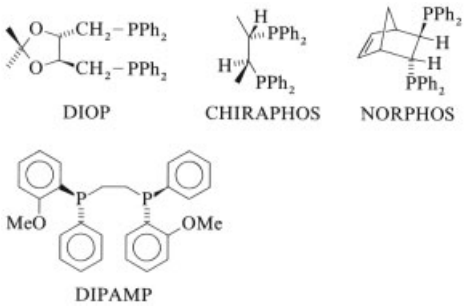

Some commercial applications of asymmetric hydrogenation already exist, the best known being the Monsanto process for the production of the chiral amino acid 3,4-dihydroxyphenylalanine [*63-84-3*], which in its L-form (L-Dopa) [*59-92-7*] is used for the treatment of Parkinson's disease.

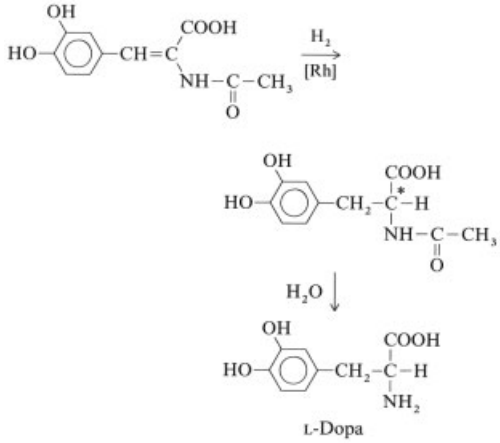

BINAP complexes of ruthenium have proved to be particularly effective catalysts for asymmetric hydrogenation of β,γ-unsaturated carboxylic acids, allyl alcohols, enamides, α-aminoketones, and substituted acrylic acids. The synthesis of the sweetener Aspartame [*22839-47-0*] (the methyl ester of L-aspartyl-L-phenylalanine) also involves asymmetric hydrogenation.

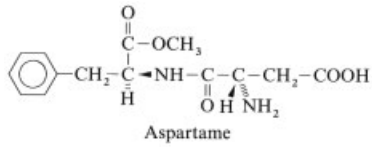

The high selectivity of homogeneous hydrogenation catalysts compared to classical heterogeneous catalysts is demonstrated by two other industrial processes. The fragrance linalool [*22564-99-4*] is produced by BASF in a multistep synthesis, in the last step of which the triple bond of dehydrolinalool is selectively hydrogenated.

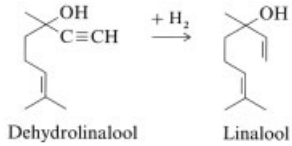

Merck has developed a synthesis of Ivermectin [*70288-86-7*], which is used in the third world for the treatment of worm infestations. Ivermectin is prepared from Avermectin B1 by selective hydrogenation of the double bond in the C22 position (arrow). Although four other

C=C double bonds are present in the molecule, Wilkinson's catalyst leads to hydrogenation exclusively in the desired position.

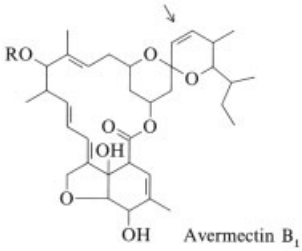

As shown by the above-mentioned examples, homogeneous hydrogenation is gaining increasing importance for the synthesis of pharmaceuticals and fine chemicals.

3.2.4. Carbon – Carbon Bond Formation [288–373]

This section deals with linear and cyclic oligomerization, telomerization, polymerization, and vinylogation. Carbon – carbon bond formation by metathesis, reactions of CO, or addition of HCN or CO_2 to alkenes are described in Sections 3.2.5, 3.2.7, and 3.2.8, respectively.

The oligomerization of alkenes leads to linear or branched high molecular mass alkenes, which are important for the production of detergents, plasticizers, and fine chemicals (see also → Hydrocarbons). Two mechanisms have been discussed: the hydride mechanism and the metallacycle mechanism. In the former, the catalytically active species is a metal hydride complex, into whose M – H bond an alkene inserts. Further insertions of alkene molecules into the resulting metal – carbon bond lead to long-chain metal alkyls. The reaction is terminted by β-hydrogen elimination with liberation of the oligomer and reformation of the metal hydride (see figure at bottom).

An example of the metallacycle mechanism is the dimerization of ethene to give butene. A bis(ethene) complex forms a metallacyclopentane, which rearranges to a butene complex by a hydrogen shift reaction.

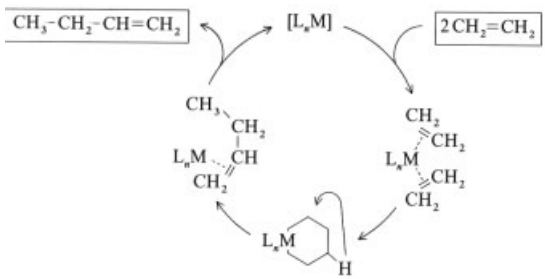

Numerous catalysts for ethene oligomerization have been investigated. An important group are the Ziegler – Natta catalysts, which are prepared from a transition-metal component and an aluminum alkyl. The most important transition metals for these catalysts are titanium, zirconium, cobalt, and nickel. Titanium and zirconium are used mostly in the form of their alkoxides or halides, and cobalt and nickel as acetylacetonates or allyl complexes. The catalytically active species are adducts of the transition-metal compound and the aluminum component. An example is the reaction of titanium tetrachloride with ethylaluminum dichloride, which gives a complex with a free coordination site for the binding of an olefin.

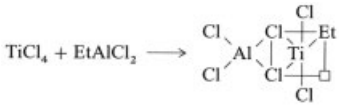

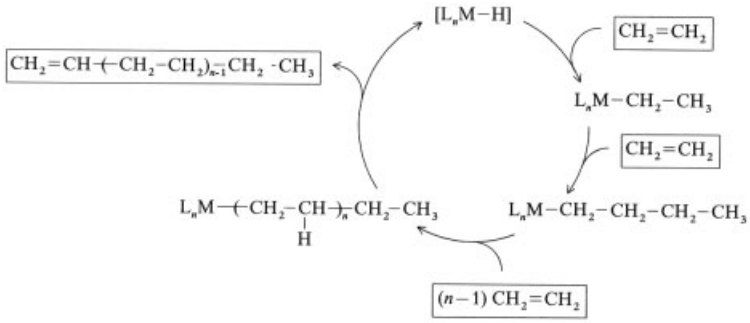

Apart from Ziegler–Natta catalysts, there are other metal compounds that catalyze the oligomerization of alkenes without activation by aluminum alkyls. These include rhodium trichloride [*10049-07-7*], palladium dichloride [*7647-10-1*], ruthenium trichloride [*10049-08-8*], and the cobalt dinitrogen complex $[CoH(N_2)(PPh_3)_3]$ [*16920-54-0*]. Of greatest industrial importance, however, are the nickel complexes of phosphorus/oxygen chelates that are used in the SHOP process (Shell) for the oligomerization of ethylene to α-olefins (→ Hydrocarbons). The catalytically active species is a nickel hydride, which is formed as shown for the following nickel phosphorus ylide complex:

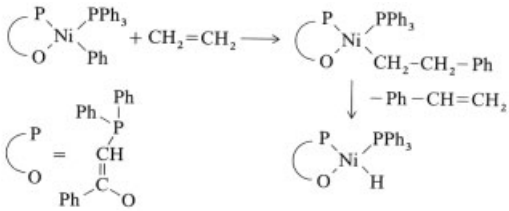

Although higher olefins such as propene or butene can also be oligomerized, the reactions often lead to formation of branched products.

Also of industrial interest is the oligomerization of alkenes with functional groups, such as acrylonitrile, styrene, and acrylates. Linear dimerization of acrylonitrile gives 1,4-dicyanobutenes, which can be used as precursors for nylon 66. The hydrodimerization of acrylonitrile leads directly to adiponitrile [*111-69-3*].

$$2\,CH_2{=}CH{-}CN + H_2 \longrightarrow NC{-}(CH_2)_4{-}CN$$

Also of importance is the linear coupling of methyl acrylate [*96-33-3*], for which $RhCl_3$, $RuCl_3$, and $[PdCl_2(PhCN)_2]$ [*14220-64-5*] were originally used as catalysts. Since then, considerably more selective catalysts have been developed, for example, methallyl cyclooctadiene palladium tetrafluoroborate [*46135-01-7*] and $[Pd(\eta^3\text{-}C_3H_5)(PMe_3)]BF_4$.

Alkynes also undergo oligomerization reactions. Linear dimerization of acetylene gives vinyl acetylene [*689-97-4*], which was formerly an important precursor for chloroprene. A typical catalyst is CuCl in ammonium chloride solution.

$$2\,HC{\equiv}CH \xrightarrow{[Cu]} HC{\equiv}C{-}CH{=}CH_2$$

Cyclic oligomerization of alkynes is, however, more common than linear oligomerization.

Cyclotrimerization of alkynes affords benzene derivatives, while cyclotetramerization yields derivatives of cyclooctatetraene.

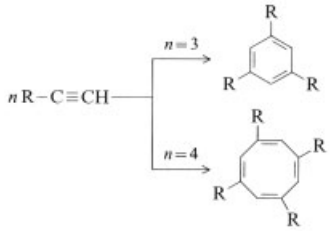

Cyclic alkyne coupling reactions have been used by VOLLHARDT for the synthesis of polycyclic systems, a typical reaction being the coupling of an α,ω-diyne with an alkyne.

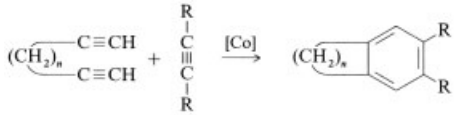

An important catalyst for this reaction is $[CpCo(CO)_2]$ [*12078-23-8*]. This synthetic method has been used for elegant syntheses of steroids, such as the female sex hormone estrogen [*53-16-7*].

1,3-Dienes can also be oligomerized, 1,3-butadiene [*106-99-0*] being particularly reactive. Cyclic oligomerization of butadiene gives isomers having various ring sizes.

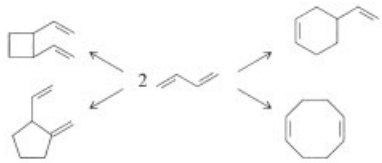

The dimer 1,5-cyclooctadiene [*1552-12-1*] is industrially important. It is synthesized by means of nickel catalysts that contain a sterically demanding ligand. The reduction of nickel bis(acetylacetonate) with $AlEt_2(OEt)$ in the presence of the ligand gives a complex that catalyzes the formation of 1,5-cyclooctadiene in almost quantitative yield under mild conditions. Mechanistically, the reaction is assumed to occur by coupling of two butadiene molecules to give a bis(allyl)nickel complex, whereby the ligand L hinders the coordination of a third butadiene molecule.

1,5-Cyclooctadiene is a versatile product. Complete hydrogenation and subsequent oxidation cleaves the molecule to give suberic acid [*505-48-6*]. Metathesis with ethene gives 1,5-hexadiene [*592-42-7*], and selective hydrogenation affords cyclooctene [*931-87-3*], which can be converted to a polyoctenamer by homometathesis, or to 1,9-decadiene [*1647-16-1*] by ethenolysis. Cyclotrimerization of butadiene is carried out industrially in the presence of a Ziegler – Natta catalyst, prepared from $TiCl_4$ and $Al_2Cl_3Et_3$, to give *cis,trans,trans*-1,5,9-cyclododecatriene [*706-31-0*]. Cyclotrimerization with nickel catalysts requires the use of naked nickel complexes (i.e., without phosphorus ligands). The mechanism can be summarized schematically as follows:

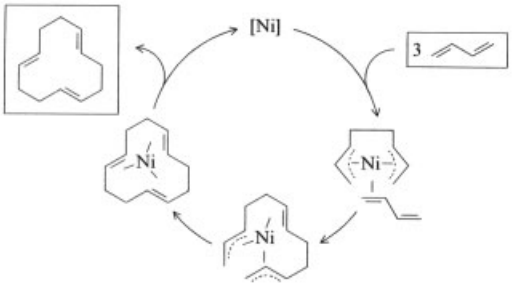

Nickel catalysts give the all-trans isomer. Du Pont (United States) and Hüls (now Degussa Hüls) (Federal Republic of Germany) produce cyclododecatriene as a precursor for nylon fibers. Hydrogenation gives cyclododecane, which can be oxidized to dodecane dicarboxylic acid [*693-23-2*]. Cyclododecane can also be converted to a C_{12} lactam, which is used for the synthesis of nylon 12.

Also of industrial importance are cooligomerization reactions, for example, of dienes and alkenes. An important example is the synthesis of *trans*-1,4-hexadiene [*7319-00-8*], which has been produced by Du Pont since 1963 and is used for production of EPDM elastomers (Nordel). In the copolymerization of the diene with ethene and propene to give EPDM, the terminal double bond of the 1,4-hexadiene reacts, while the less reactive inner double bond can be used for further polymerization reactions.

The synthesis of *trans*-1,4-hexadiene occurs by rhodium-catalyzed cooligomerization of ethene and butadiene. The starting material is rhodium trichloride, which is converted to a rhodium hydride complex. Addition of butadiene then gives an η^3-crotyl complex, which undergoes coordination and insertion of an ethene molecule. Subsequent coordination of butadiene then leads to elimination of *trans*-1,4-hexadiene.

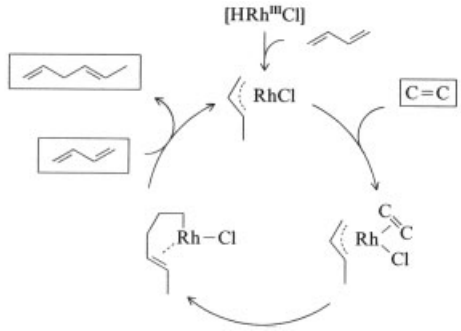

The reaction can also be carried out with Ziegler – Natta catalysts such as $NiCl_2(PPh_3)_2$/*i*-$BuAlCl_2$. The use of cobalt or iron catalysts results in formation of *cis*-1,4-hexadiene [*7318-67-4*].

Another commercially important oligomer is ethylidenenorbornene [*16219-75-3*], which can be prepared from cyclopentadiene and butadiene. An uncatalyzed Diels – Alder reaction gives 5-vinylnorbornene [*3048-64-4*], which can be converted to ethylidenenorbornene by, for example, titanium Ziegler catalysts. An alternative synthesis is the nickel-catalyzed cooligomerization of norbornadiene and ethene. Like 1,4-hexadiene, ethylidenenorbornene is used in EPDM elastomers.

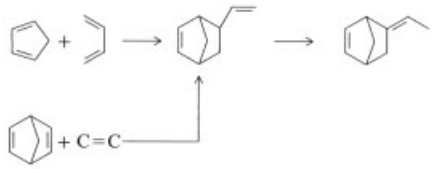

An interesting possibility is the ability to perform asymmetric cooligomerization reactions. For example, the cooligomerization of norbornene [*498-66-8*] with ethene in the presence of metal complexes containing optically active phosphine ligands leads to selective formation of a single enantiomer. Dimenthyl alkyl phos-

phines are effective optically active ligands for this reaction:

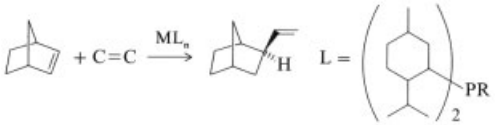

Telomerization is a C – C bond-forming reaction that is closely related to oligomerization. Telomerization is the dimerization of a 1,3-diene with simultaneous addition of a nucleophile HX. Although in principle different isomers can result, their formation is controlled by the catalyst:

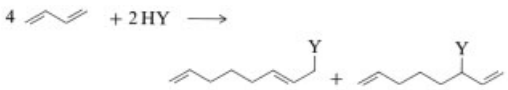

The most important telomerization catalysts are palladium complexes, such as $[Pd(PPh_3)_2(MA)]$ [*17830-50-1*] (MA = maleic anhydride), $[Pd(OAc)(C_3H_5)_2]_2$ [*12084-71-8*], and combinations of palladium acetate or palladium acetylacetonate with phosphines or phosphites. Whereas nickel complexes also give high yields of telomers, only in a few cases are platinum and rhodium active in telomerization.

A wide range of compounds can serve as nucleophiles in these reactions, including water, alcohols, silanols, carboxylic acids, primary and secondary amines, β-diketones, malonates, β-ketoesters, and nitroalkanes. Telomerization has yet to gain major industrial importance. Nevertheless, the multiple functionality of the telomers makes them excellent starting materials for the synthesis of complex natural products. Butadiene and acetic acid undergo telomerization with high selectivity to give an octadienyl acetate, which can be converted in five steps to dihydrojasmone [*1128-08-1*].

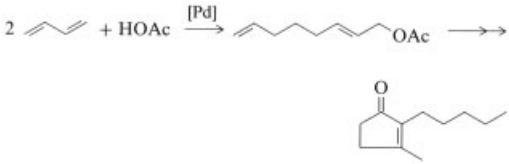

The telomer 3-hydroxy-1,7-octadiene can be dehydrated to give the corresponding ketone, which is an important building block for the synthesis of steroid systems.

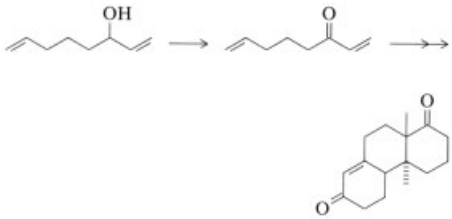

The telomerization of nitromethane with butadiene leads to binding of two octadienyl chains to the carbon atom of the nitroalkane. The resulting molecule can be used for the synthesis of *cis*-civetone [*542-46-1*].

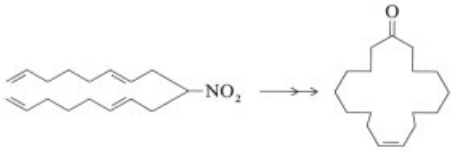

Polymerization of alkenes (→ Polyolefins) is a major field of application for transition-metal catalysis. For example, isotactic polypropylene is produced with titanium catalysts. Up to now heterogeneous catalysts have mainly been used for this purpose. These include γ-$TiCl_3$, prepared from $TiCl_4$ and $AlEt_3$; titanium chloride on a magnesium chloride carrier; and δ-$TiCl_3$.

However, homogeneous catalysts are used preferentially in some areas of polymer technology. For example, high-density polyethylene is produced in the presence of a homogeneous catalyst prepared from $TiCl_4$ and aluminum alkyls. A newly developed, highly active catalyst for the synthesis of linear polyethylene is a combination of $[Cp_2ZrCl_2]$ [*1291-32-3*] and methyl aluminoxane [*90330-46-4*]; the yield is 10^6 kg polyethylene per hour per mole of zirconium complex. Isotactic polypropylene can also be produced by homogeneous catalysis, for example, with ethylene bis(indenyl) zirconium dichloride [*112243-78-4*] and methylaluminoxane.

EDPM polymers are also produced by homogeneous catalysis. The catalysts are presumably alkyl vanadium (III) complexes, which are formed in situ from vanadium compounds such as $VOCl_3$ [*7727-18-6*] or VCl_4 [*7632-51-1*] and aluminum alkyls. The resulting terpolymers are used as highly ozone resistant elastomers. *Cis*-1,4-polybutadiene [*9003-17-2*] is another polymer that is readily accessible by transition-metal catalysis. In the Phillips pro-

cess, Ziegler catalysts prepared from titanium halides and aluminum alkyls are used, while the Goodrich process employs catalysts that are prepared from cobalt salts and alkylaluminum halides. Nickel-based catalysts exhibit similar reactivity and give *cis*-1,4-polybutadiene with selectivities greater than 90 %. A mechanistic explanation for the high cis selectivity is as follows:

The growing polymer chain is bound to the catalytic metal center M by an η^3-allyl bond. Butadiene coordinates to the metal as a chelate and undergoes C – C coupling with the terminal carbon atom of the allyl group, simultaneously forming a new η^3-allyl bond. The η^3-allyl bonding means that each chain extension step forms only a cis double bond.

Further use of organometallic catalysts in polymer chemistry can be expected in the future. The synthesis of polyacetylene [*25067-58-7*], which has applications in electronics and energy storage, and of specialized methacrylate polymers for use in fiber optics have been extensively investigated.

Finally, a few examples of the use of C – C bond-forming reactions in the synthesis of fine chemicals are given below. An important class of reactions are metal-catalyzed cyclopropanations, in which alkyl diazonium acetates react with particular unsaturated compounds to give novel pyrethroid insecticides.

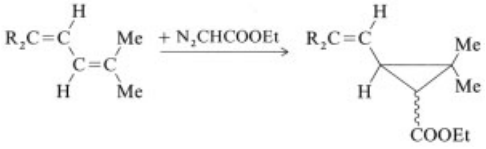

The regio- and enantioselectivity of the reaction can be controlled by choice of the metal catalyst, usually a cobalt, rhodium, or copper complex. Mechanistically, it is assumed that the metal and the diazo compound first form a carbene complex, which reacts with the alkene to give a metallacyclobutane. The cyclopropane derivative is formed by reductive elimination, reforming the catalyst.

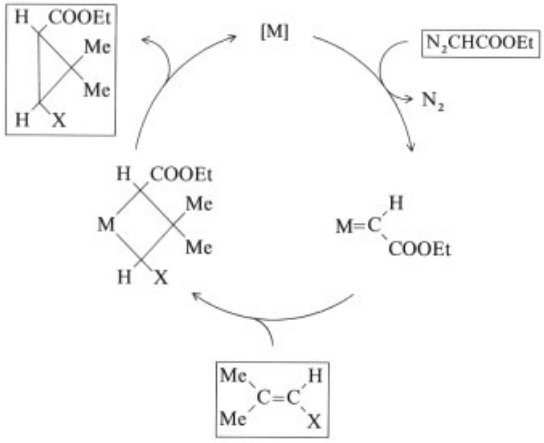

Also of importance is the vinylogation of alkenes, for which vinyl halides are used as vinylogating agents, and palladium complexes, prepared in situ from $Pd(OAc)_2$, PPh_3, and NEt_3, as catalysts.

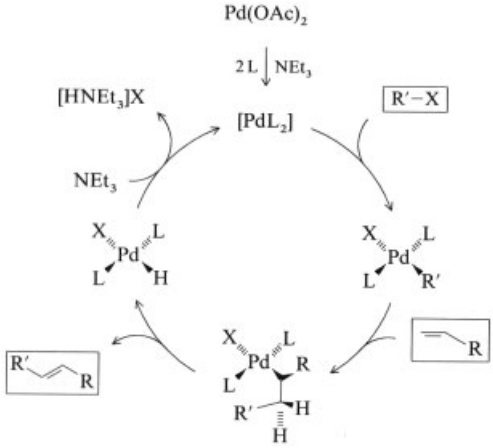

3.2.5. Metathesis [374–404]

Metathesis is formally the exchange of alkylidene groups between two alkenes.

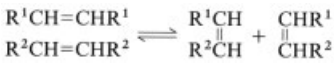

Typical heterogeneous metathesis catalysts are WO_3 [*1314-35-8*] on SiO_2, and Re_2O_7 [*1314-68-7*] on Al_2O_3, while an important homogeneous catalytic system is WCl_6 – EtOH –

R_2AlCl. The reactive intermediates are assumed to be carbene complexes.

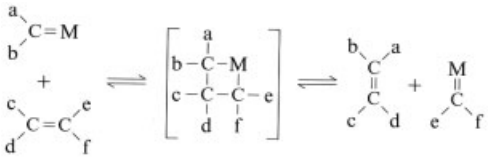

The formation of the metal carbene in the catalytic system $WCl_6 - Me_2AlCl$ is believed to occur as follows:

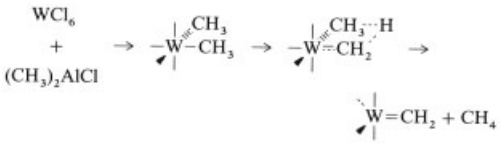

The oldest industrial metathesis process, the Phillips triolefin process, in which 2 mol of propene is converted to ethene and 2-butene, was brought on-line in 1966.

However, the 50×10^3 t/a Phillips plant in Canada ceased production in 1972 because of the changed situation in the olefin market. In the meantime, the reverse reaction, the synthesis of propene from ethene and 2-butene, was taken up by the Atlantic Richfield Corp. In the first step ethene is dimerized to give 2-butene, which is then metathesized with an excess of ethene in the presence of $Re_2O_7 - Al_2O_3$ catalysts. This reaction allows the mixture of lower alkenes formed in the naphtha pyrolysis to be altered to correspond to prevailing market requirements.

The largest use of metathesis is currently the SHOP process (Shell), in which ethylene is oligomerized to give long-chain α-olefins. Undesired low molecular mass olefins ($<C_{10}$) and high molecular mass alkenes ($>C_{20}$) can be converted by metathesis to $C_{10} - C_{14}$ inner alkenes after an isomerization step. The terminal and inner alkenes produced in the SHOP process are used mainly for the production of detergents, plasticizer alcohols, synthetic lubricants, and LLDPE. Total world SHOP plant capacity is currently being increased to over 800×10^3 t/a. Another major application of metathesis is the synthesis of polyalkenamers by metathesis of cyclic alkenes. The initial cyclic molecule undergoes ring opening and is converted stepwise to larger and larger rings and, finally, polymers. This reaction is also called ring-opening metathesis polymerization (Romp).

The first industrial application of ring-opening metathesis dates from 1976. In a 5000 t/a plant of CdF-Chimie (France), norbornene is converted to polynorbornene [*25038-76-0*] (Norsorex), presumably by means of tungsten catalysts.

The metathesis of cyclooctene to give a polyoctenamer has been carried out by Hüls since 1980. The product is sold under the trade name Vestenamer 8012.

The metathesis of dicyclopentadiene to give the polymer Metton has been operated in Canada by Hercules since 1985.

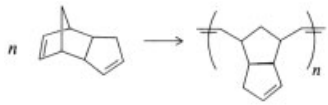

Less important industrial applications of metathesis include the Phillips neohexene process and the Shell Feast process (→ Hydrocarbons, Chap. 2.2.3.). The neohexene is formed by metathesis of 2,4,4-trimethyl-2-pentene [*558-37-2*] and ethene.

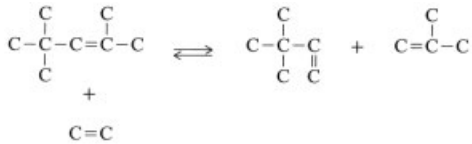

In the Feast process, cycloolefins and ethene are converted to α,ω-dienes, which are otherwise industrially inaccessible. Starting material for the Feast process is 1,5-cyclooctadiene, which can be produced by cyclodimerization of butadiene.

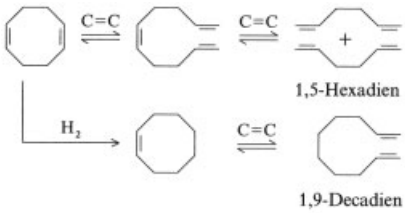

3.2.6. Oxidation [405–435]

In the Wacker–Hoechst process (→ Acetaldehyde, Chap. 4.3.1.), ethene chloro complexes of palladium (II) are decomposed by water to give acetaldehyde and palladium(0). The oxidation of Pd(0) back to Pd(II) can be carried out with air, whereby the use of a Cu(I)/Cu(II) couple allows the reaction to be carried out catalytically. The sum of all the reactions involved is the reaction of ethene and oxygen to give acetaldehyde.

$$
\begin{array}{lll}
1)\ C_2H_4 + PdCl_2 + H_2O & \longrightarrow & CH_3CHO + Pd(0) + 2\,HCl \\
2)\ Pd(0) + 2\,CuCl_2 & \longrightarrow & PdCl_2 + 2\,CuCl \\
3)\ 2\,CuCl + 2\,HCl + 1/2\,O_2 & \longrightarrow & 2\,CuCl_2 + H_2O \\
\hline
\Sigma\quad C_2H_4 + 1/2\,O_2 & \longrightarrow & CH_3CHO
\end{array}
$$

When the reaction is carried out with terminal or inner alkenes, ketones are obtained. For example, propene can be converted to acetone. In the mechanism for acetaldehyde formation, it is assumed that first ethene and then water adds to the metal center. A hydroxyethyl complex is then formed by oxypalladation, which, after rearranging to a hydrido vinylalcohol complex, releases acetaldehyde.

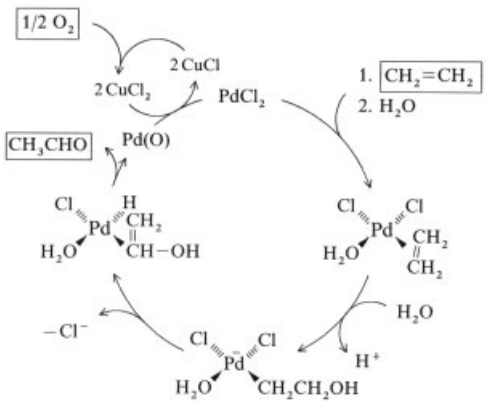

The reaction is versatile in that solvents other than water can be used. Oxidation in alcohols yields the corresponding vinyl ethers, while reaction in acetic acid affords vinyl acetate. A large proportion of the acetaldehyde produced by the Wacker–Hoechst process is oxidized to acetic acid or acetic anhydride by means of manganese, cobalt, or copper salts.

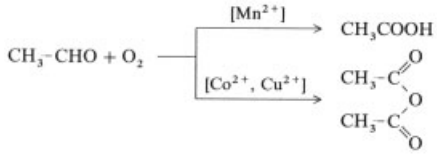

Another oxidation reaction of olefins is epoxidation, which is carried out on an industrial scale by reaction of ethene with oxygen on heterogeneous silver catalysts (→ Ethylene Oxide). Homogeneous epoxidation is also used industrially. For example, selective epoxidation of olefins with hydroperoxides can be catalyzed by molybdenum or vanadium compounds.

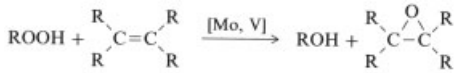

This reaction was developed further by Halcon for the production of propene oxide, in which *tert*-butyl hydroperoxide or ethylbenzene hydroperoxide is used as the oxidant for propene. A simplified scheme for the reaction with a molybdenum catalyst is as follows:

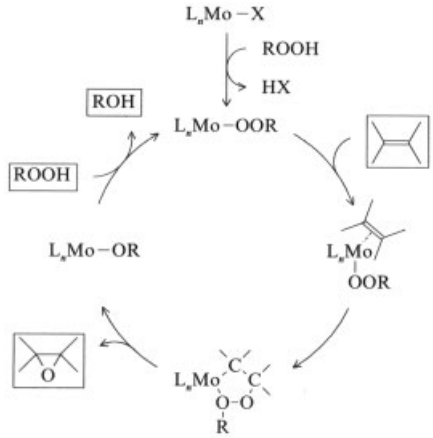

Stereoselective epoxidation is used industrially for the synthesis of fine chemicals. An example is the insect pheromone (+)-disparlure

[*54910-51-9*], which is used against the gypsy moth, a leaf-feeding insect that causes extensive damage to trees in the United States. The synthesis starts from the allyl alcohol *Z*-2-tridecenol, which is converted to the desired epoxide in 95 % enantiomeric purity by titanium alkoxides, D-(−)-diethyl tartrate, and *tert*-butyl hydroperoxide.

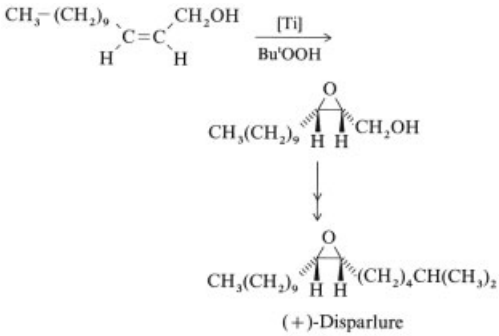

Propranolol [*13071-11-9*] is used for the treatment of cardiac disease and high blood pressure. In the first step of its synthesis, allyl alcohol is epoxidized with *tert*-butyl hydroperoxide, affording the corresponding glycidol. A titanium complex of diethyl tartrate is used as catalyst, whereby the use of (+)- or (−)-tartrate leads to formation of the corresponding glycidol isomer.

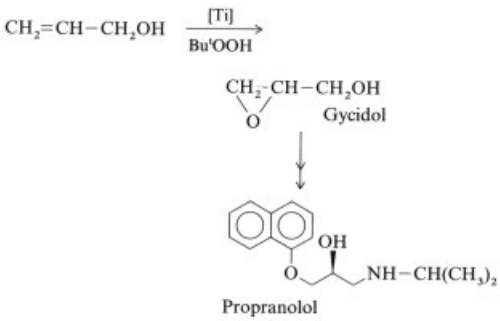

Also of major importance is the homogeneous catalytic oxidation of alkanes and alkyl aromatics. The radical chain reactions are initiated by cobalt or manganese ions. Therefore, unlike olefin oxidation or epoxidation, the oxidation of alkanes and alkyl aromatics does not directly involve the transition-metal complex.

Linear short-chain alkanes, in particular butane, can be converted to acetic acid by liquid-phase oxidation with cobalt acetate as catalyst. The reaction is, however, unselective, and has a maximum yield of ca. 50 %. A similar process is used to oxidize cyclohexane to give a mixture of cyclohexanone [*108-94-1*] and cyclohexanol [*108-93-0*], which can be converted in a second step to adipic acid [*124-04-9*].

p-Xylene is oxidized industrially to give terephthalic acid [*100-21-0*] or dimethyl terephthalate [*120-61-6*]. In order to convert both methyl groups of *p*-xylene to carboxyl groups, either bromide ions are added to the metal catalyst (Amoco process) or an indirect route via methyl toluate is used. It is assumed that the Co(II) or Mn(II) catalysts used are oxidized by the hydroperoxides to give M(III) ions, which can then initiate the oxidation by extraction of a benzylic hydrogen atom. The oxidation of alkyl aromatics is also used for the synthesis of fine chemicals. For example, oxidation of *p*-*tert*-butyltoluene with a cobalt bromide catalyst affords the corresponding aldehyde, which can be converted to the fragrance lilial [*80-54-6*] by aldol condensation with propionaldehyde followed by selective hydrogenation.

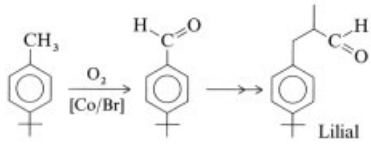

3.2.7. Reactions with Carbon Monoxide [436–468]

Of the many reactions of carbon monoxide, only the important transition-metal catalyzed reactions with hydrogen, alkynes, alkenes, and alcohols are described here.

The synthesis of methanol is the most important of the reactions with hydrogen; it is carried out heterogeneously with zinc, chromium, or copper oxide catalysts. The mixture of carbon monoxide and hydrogen (synthesis gas) is also the starting material for the Fischer – Tropsch reactions. In the presence of heterogeneous cobalt or iron catalysts, these reactions can be used for the production of alkane mixtures that can be used as fuels. Currently, this process is carried out on an industrial scale in South Africa, where cheap coal is used as the source of the synthesis gas. More recent research on Fischer – Tropsch chemistry has the aim of controlling the reac-

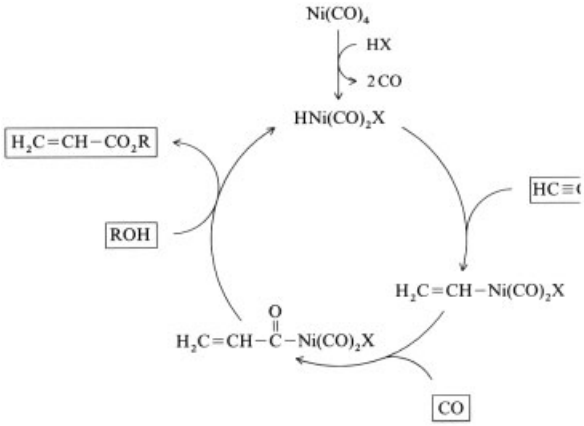

tions by means of the catalyst to give other products such as short-chain olefins, long-chain α-olefins, and alcohols. Homogeneous reactions of hydrogen and carbon monoxide have also been extensively investigated. An interesting result is the rhodium-catalyzed synthesis of ethylene glycol, albeit at pressures above 100 MPa.

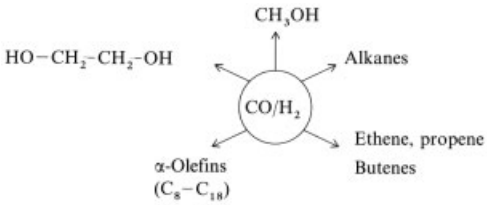

The reactions of carbon monoxide with alkynes are largely based on the work of Reppe. For example, acetylene, carbon monoxide, and water can be converted to acrylic acid with the catalyst $Ni(CO)_4$ [*13463-39-3*]. The reaction is regarded as the hydrocarboxylation of an alkyne. Instead of water, other molecules of general formula HX can be used. In the reactions with alcohols (hydroesterification of alkynes), acrylates are formed, which are used for production of polyacrylates. The reaction involves the formation of a nickel hydride species, which undergoes insertion of acetylene. Subsequent insertion of carbon monoxide and reaction with the alcohol gives the acrylate with reformation of the nickel hydride complex (see figure at top of page).

The reactions of alkenes with carbon monoxide also include hydrocarboxylation and hydroesterification reactions, in which saturated carboxylic acids or their esters are formed; for example, propanoic acid from ethylene. In the presence of iodine-modified nickel carbonyl catalysts, the reactions occur at 190 °C and 10 MPa CO pressure. Longer-chain alkenes always give mixtures of carboxylic acids because the carbon monoxide can attack at various carbon atoms of the alkene due to possible isomerization of the double bond.

BASF has further developed the hydroesterification of butadiene to give dimethyl adipate [*627-93-0*]. The intermediate in the reaction is methyl 3-pentenoate. The two reaction steps, which occur at different CO pressures, require a catalyst prepared from $Co_2(CO)_8$ [*10210-68-1*] and a nitrogen base (e.g., pyridine).

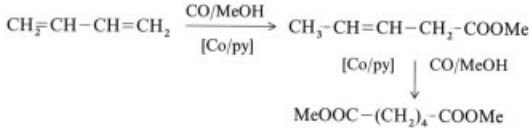

By far the most important reaction of carbon monoxide with alkenes is hydroformylation (→ Oxo Synthesis), discovered by Roelen at Ruhrchemie, in which, for instance, ethene and synthesis gas in the presence of a cobalt carbonyl catalyst give propionaldehyde. Currently, ca. 5×10^6 t/a of C_3-C_{19} oxo products are produced worldwide. The primary products of the reactions are aldehydes, which are mostly converted to other products, principally alcohols. The mechanism of cobalt-catalyzed hydroformylation is believed to involve a hydridocobalt carbonyl as catalytically active species. Insertion of olefin and carbon monoxide gives an acyl complex that is hydrogenated, forming the aldehyde.

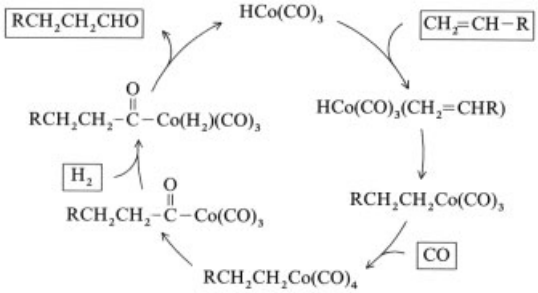

The oxo synthesis catalysts have been developed over a time-span of 50 years. Classical hydroformylation with unmodified cobalt catalysts is performed at 150 °C and 20 – 30 MPa. The aldehyde yield is 90 %, and the ratio of linear to branched (*n*/*iso*) aldehydes is 3 : 1. Modification of the cobalt catalysts, in particular by using phosphine ligands, allowed the reaction pressure to be reduced to 10 MPa and improved the *n*/*iso* ratio to 7 : 1. A major development was the use of rhodium catalysts, which are ca. 1000 times more active. Since the reaction can be carried out at 1 – 2 MPa, it is termed low-pressure oxo synthesis (LPO process). It gives a yield of aldehyde of 98 % and good selectivity for formation of the desired *n*-aldehydes. Use of the catalyst $[HRh(CO)(PPh_3)_3]$ gives an *n*/*iso* ratio of >10 : 1. A further development is the use of rhodium catalysts with water-soluble phosphine ligands such as tris(*m*-sulfonato)triphenylphosphine. In this process, jointly developed by Rhône-Poulenc and Ruhrchemie, the reaction is carried out in a two-phase liquid – liquid system, which simplifies separation of the organic product from the aqueous catalyst phase and thus minimizes catalyst losses. Hydroformylation can be applied to a wide range of unsaturated compounds and is hence suitable for the synthesis of fine chemicals.

Conjugated dienes can be hydroformylated with rhodium carbonyl phosphine catalysts to give dialdehydes. The same catalyst allows allyl alcohols to be hydroformylated without isomerization. Unsaturated carboxylic acids and their esters are also suitable starting materials for rhodium- or cobalt-catalyzed oxo synthesis.

A further important group of CO reactions is the reaction of carbon monoxide with functionalized compounds such as alcohols, esters, and nitro and halogeno compounds. An industrially important reaction is the Monsanto acetic acid process (→ Acetic acid, Chap. 4.1.), which involves the carbonylation of methanol with an iodine-modified rhodium catalyst. The methanol is converted to methyl iodide, which oxidatively adds to the rhodium complex. After formation of an acyl complex, acetyl iodide is eliminated and then reacts with water to give acetic acid. The Monsanto process can be carried out under mild conditions (175 °C, 5 MPa) and gives high yields of acetic acid (90 %) with excellent selectivity (>99 %).

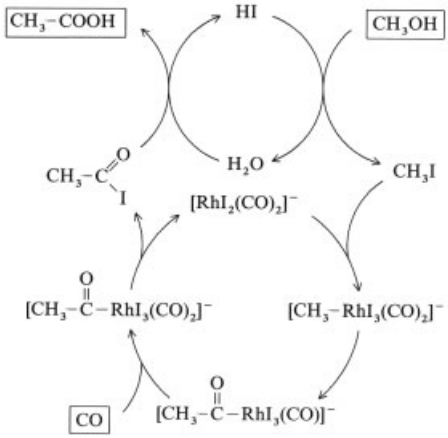

Homologation is another process in which methanol is converted to C_2 species by reaction with synthesis gas. Use of cobalt – iodine catalysts gives high yields of acetaldehyde. The addition of ruthenium increases the hydrogenation activity of the catalysts and leads to preferential formation of ethanol.

In the acetic anhydride process, jointly developed by Tennessee Eastman and Halcon SD, methyl acetate is carbonylated in the presence of rhodium – iodine catalysts (→ Acetic Anhydride and Mixed Fatty Acid Anhydrides, Chap. 1.3.3.). As in the Monsanto process, the catalytically active species appears to be the complex $[RhI_2(CO)_2]^-$. The process is noteworthy in that it for the first time allowed the production of acetic anhydride from coal-based synthesis gas, independent of petrochemical feedstocks.

$$CO + 2H_2 \longrightarrow CH_3OH \xrightarrow{CH_3COOH} CH_3COOCH_3 \xrightarrow{CO} (CH_3CO)_2O$$

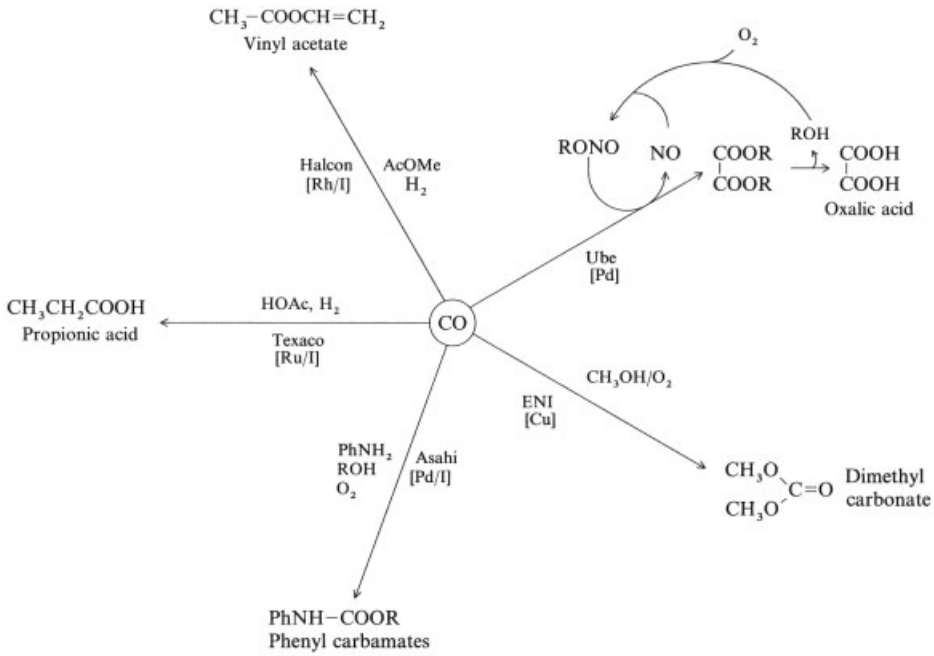

Benzyl bromide undergoes a double carbonylation in the presence of cobalt catalysts and lithium hydroxide to give phenyl pyruvic acid [*156-06-9*]. The palladium-catalyzed reaction of carbon monoxide with aromatic nitro compounds affords industrially important isocyanates.

Finally, some processes are listed that are either under development or are aldready used to a limited extent industrially (see figure above). Noteworthy are the Ube oxalic acid synthesis by oxidative coupling of carbon monoxide (→ Oxalic Acid); the Enichem dimethylcarbonate process involving oxidative carbonylation of methanol; the Asahi carbamate synthesis; the homologation of acetic acid to give propionic acid (Texaco); and the Halcon process for the synthesis of vinyl acetate from methyl acetate and synthesis gas.

3.2.8. Functionalization [469–505]

Functionalization is the addition of a molecule of type HY or YX_2 to C–C multiple bonds. The former reaction type includes hydroboration, hydroalumination, hydrocyanation, hydrosilylation, and hydroamination, while the latter includes cyclopropanation (see Section 3.2.4) and carboxylation with CO_2. Hydroboration ($HY=HBX_2$) normally occurs without catalysis. In some cases, however, titanium salts or complexes have proved to be effective catalysts. For example, the hydroboration of 1-octene with lithium borohydride is catalyzed by titanocene dichloride. Work up with sodium methoxide and hydrogen peroxide gives the 1- and 2-octanols in almost 100 % yield.

$$R{-}CH{=}CH_2 \xrightarrow[{[TiCp_2Cl_2]}]{LiBH_4} Li\left[(RCH_2CH_2)_nBH_{4-n}\right]$$

$$\xrightarrow[H_2O_2]{CH_3ONa} R{-}CH_2{-}CH_2{-}OH$$

The low solubility of $NaBH_4$ in organic solvents can be overcome by use of phase-transfer catalysts such as 18-crown-6. The system $TiCl_3/NaBH_4$/18-crown-6 allows many hydroborations to be carried out at room temperature. Dienes with isolated double bonds can be selectively hydroborated at one double bond. Thus, 1,7-octadiene [*3710-30-3*] can be converted to 8-hydroxy-1-octene [*13175-44-5*] in 80 % yield. Hydroalumination ($HY = HAlX_2$) of olefins is an important method for synthesizing organoaluminum compounds. Important catalysts are $TiCl_4$, $TiCl_3$, $TiCp_2Cl_2$, and $Ti(OR)_4$; zirconium

compounds ($ZrCl_4$, $ZrCp_2Cl_2$) and nickel phosphine complexes can also be used. The most reactive substrates are terminal olefins, and the reactivity of the aluminum hydrides decreases in the order:

$$AlH_3 > H_2Al-Cl > [(R_3Si)_2N]_2Al-H > (R_2N)_2Al-H > HAlCl_2$$

An advantage of hydroalumination over hydroboration is that oxidation of the Al – C bond does not require peroxide, but can be carried out with oxygen. However, attack occurs at functional groups other than the double bond.

Catalytic hydrocyanation (HY = HCN) is a good method for the synthesis of alkyl nitriles and dinitriles, which are starting materials for amides, amines, and carboxylic acids. Although the most important catalysts are nickel complexes, complexes of cobalt, palladium, and copper are also active. An example of the hydrocyanation of a terminal olefin is the addition of HCN to 1-octene. With the catalyst $[Ni\{P(OPh)_3\}_4]$ and cocatalyst $AlEtCl_2$, the anti-markovnikov adduct, 1-nonanenitrile, is obtained with 80 % selectivity.

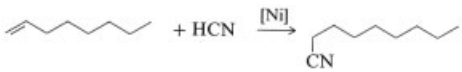

Octenes with internal double bonds give 1-nonanenitrile with the same selectivity, suggesting that a nickel-catalyzed isomerization of the double bond occurs before addition. The mechanism of hydrocyanation involves a hydrido cyano species, which reacts with the olefin to give an alkylcyano intermediate. After elimination of the nitrile, the hydrido cyano species is reformed by addition of HCN.

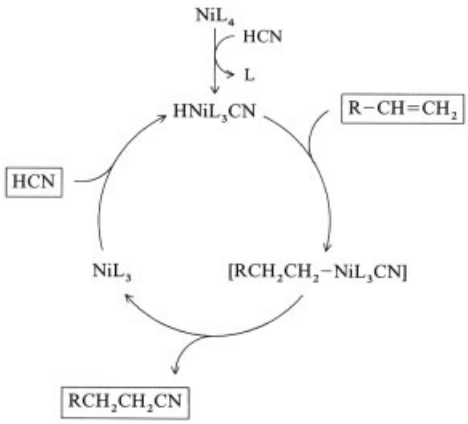

Stereoselective hydrocyanation can be carried out by using catalysts with chiral ligands. The complex $[Pd(DIOP)_2]$ leads exclusively to formation of threo isomers.

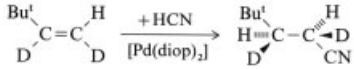

Hydrocyanation can also be carried out with substituted olefins such as styrene or acrylonitrile and cyclic olefins such as cyclopentene or norbornene. Of particular industrial interest is the double hydrocyanation of butadiene to form adiponitrile, which is hydrogenated to hexamethylenediamine, an important precursor for nylon 66. The hydrocyanation of butadiene has been carried out by DuPont since 1971 with nickel(0) phosphite complexes as catalysts.

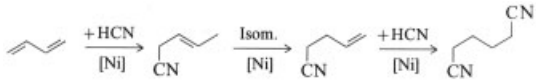

The reaction occurs stepwise, involving hydrocyanation to give 3-pentenenitrile, isomerization to form 4-pentenenitrile, and a further hydrocyanation step in which adiponitrile is formed. The mechanism is believed to involve an allyl nickel complex as the reactive intermediate. Because addition of the cyano group to the 4-carbon atom is strongly favored, the first hydrocyanation step gives pentenenitrile and not the isomeric methyl branched product.

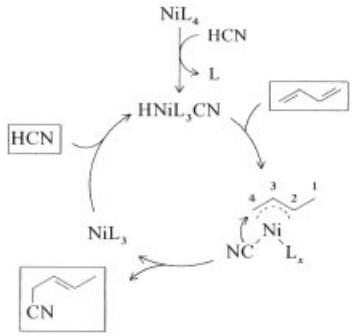

Hydrosilylation (HY = $HSiR_3$) of olefins leads to a range of functionalized silanes that are valuable starting materials in organic synthesis. The most important catalytic metals are platinum, rhodium, and nickel, but reactions with vanadium, palladium, ruthenium, iron, and

cobalt catalysts have also been reported. The reactivity of the silanes decreases in the order:

$$HSiCl_3 > HSiCl_2Et > HSiClEt_2 > HSiEt_3 > HSiPr_3$$

That is, the reactivity decreases with increasing electron density on the silane hydrogen atom. These electronic effects are, however, strongly influenced by additional steric effects.

The reactivity also varies between terminal and inner alkenes. For example, the rhodium-catalyzed hydrosilylation of 1-heptene with trichlorosilane proceeds in 54 % yield, whereas the analogous reactions with *cis*- and *trans*-2-heptene give yields of only 10 % and 1 %, respectively. Therefore, dienes that have double bonds of differing reactivity can often be selectively monosilylated. For example, in the presence of nickel or platinum catalysts, 4-vinylcyclohexene reacts only at the exocyclic double bond. Conjugated dienes such as butadiene can also be hydrosilylated. The proportions of the 1,2 and 1,4 addition products and the *cis/trans* ratio of the product depends on the catalyst and silane used. Hydroamination ($HY = HNR_2$) of mono- and dialkenes with primary or secondary amines leads to the corresponding higher substituted amines. Important catalysts are salts or complexes of palladium, rhodium, nickel, platinum, and iridium. An example is the reaction between 1-butene and dimethylamine, which affords a mixture of two isomeric butyl dimethyl amines.

$$CH_3-CH_2-CH=CH_2 + HNMe_2 \xrightarrow{[Pd]} C_4H_9-NMe_2 + C_2H_5-CH(CH_3)-NMe_2$$

A special functionalization reaction is the carboxylation of unsaturated compounds with carbon dioxide. The reaction of alkynes with CO_2 in the presence of nickel catalysts gives doubly substituted pyrones. The mechanism involves an oxanickelcyclopentene complex, which undergoes ring enlargement by insertion of an alkyne to give a 7-membered ring. Decomposition of the 7-membered ring liberates the pyrone and reforms the nickel catalyst.

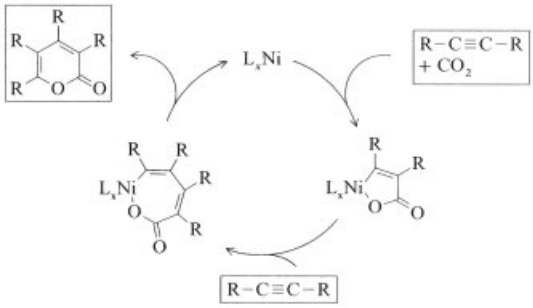

Methylene cyclopropanes can be converted to γ-lactones by palladium catalysis, while allene gives esters.

The reaction between butadiene and carbon dioxide can be controlled to give δ- or γ-lactones by judicious choice of the catalyst and reaction conditions. Two butadiene molecules couple at the palladium center to give a bis(η^3-allyl) complex. Carbon dioxide then inserts in one allyl bond, giving an allylcarboxylate complex from which the lactone can be liberated (see figure below).

Mixtures of dienes (e.e., butadiene and isoprene) also react with CO_2 in the presence of palladium catalysts.

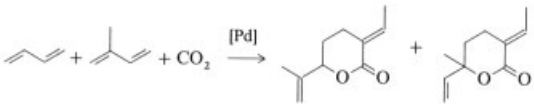

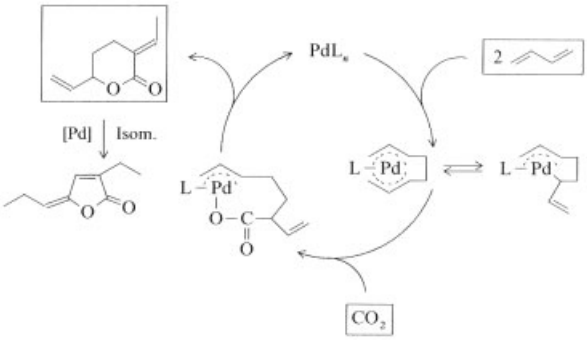

3.2.9. Alkane Activation [506–531]

In theory, it should be possible to carry out alkane activation – the functionalization of unsaturated hydrocarbons – in homogeneously catalyzed reactions.

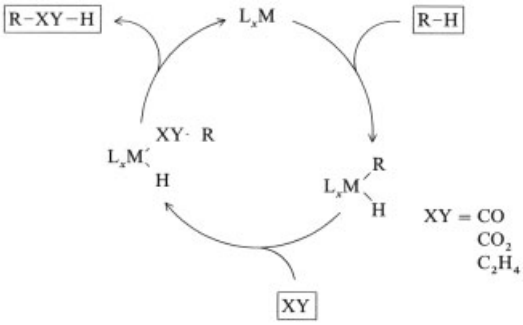

The hydrido alkyl intermediate is, however, highly unstable and generally undergoes spontaneous reductive elimination of the alkane. Nevertheless, it has been shown in stoichiometric reactions that alkane activation is possible. Some examples include:

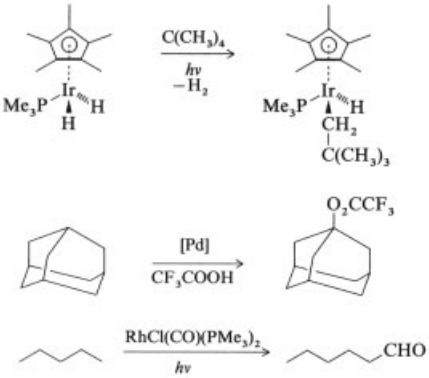

Future research in this area will show whether catalytic alkane activation can achieve practical importance.

4. References

1. G. Wilkinson, F. G. A. Stone, E. W. Abel: *Comprehensive Organometallic Chemistry,* vols. 1 – 9, Pergamon Press, Oxford 1982. *Comprehensive Organometallic Chemistry II: A Review of Literature 1982 – 1994,* vols. 1 – 14, Pergamon Press, Oxford 1982 – 94.
2. J. Buckingham (ed.): *Dictionary of Organometallic Chemistry,* vols. 1 – 3, Chapman and Hall, London 1984.
3. I. Haiduc, J. J. Zuckerman: *Basic Organometallic Chemistry,* Walter de Gruyter, Berlin 1985.
4. A. J. Pearson: *Metallo-Organic Chemistry,* Wiley Interscience, New York 1985.
5. A. W. Perkins, R. C. Poller: *An Introduction to Organometallic Chemistry,* MacMillan, London 1986.
6. P. Powell: *Principles of Organometallic Chemistry,* 2nd ed., Methuen, London 1987.
7. C. Elschenbroich, A. Salzer: *Organometallchemie,* 2nd ed., Teubner, Stuttgart 1988; *Organometallics,* VCH Verlagsgesellschaft, Weinheim 1989.
8. A. de Meijere, H. tom Dieck (eds.): *Organometallics in Organic Syntheses,* Springer Verlag, Berlin 1988.
9. H. Werner, G. Erker (eds.): *Organometallics in Organic Synthesis 2,* Springer Verlag, Berlin 1989.
10. Degussa AG (ed.): *Edelmetall-Taschenbuch,* Hüthig-Verlag, Heidelberg 1995.
11. P. R. Jenkins: *Organometallic Reagents in Synthesis,* Oxford University Press, Oxford 1992.
12. A. E. Martell, R. D. Hancock: *Metall Complexes in Aqueous Solutions,* Plenum Press, New York 1996.
13. R. C. Poller: *The Chemistry of Organotin Compounds,* Lagos Press, London 1970.
14. B. J. Wakefield: *The Chemistry of Organolithium Compounds,* Pergamon Press, Oxford 1974.
15. P. Paetzold: "Neues vom Bor und seinen Verbindungen," *Chem. Unserer Zeit* **9** (1975) 67.
16. T. P. Onak: *Organoborane Chemistry,* Academic Press, New York 1975.
17. R. Okawara, Y. Matsumura: "Recent Advances in Organoantimony Chemistry," *Adv. Organomet. Chem.* **14** (1976) 187.
18. J. J. Eisch: "Rearrangements of Unsaturated Organoboron and Organoaluminum Compounds," *Adv. Organomet. Chem.* **16** (1977) 67.
19. H. Yamamoto, H. Nozaki: "Selective Reactions with Organoaluminum Compounds," *Angew. Chem. Int. Ed. Engl.* **17** (1978) 169.
20. R. C. Larock: "Organoquecksilberverbindungen in der organischen Synthese," *Angew. Chem.* **90** (1978) 28.

21. G. J. M. van der Kerk: "The organic chemistry of tin," *CHEMTECH* 1978, June, 356.
22. E. W. Colvin: *Silicon in Organic Synthesis,* Butterworths, London 1981.
23. W. P. Neumann: "Die organischen Verbindungen und Komplexe von Ge, Sn und Pb," *Naturwissenschaften* **68** (1981) 354.
24. J. M. Connolly, C. Hoff: "Organic Compounds of Divalent Tin and Lead," *Adv. Organomet. Chem.* **19** (1981) 123.
25. J. J. Eisch: *Organometallic Synthesis,* vol. 2, "Non-transition Metal Compounds," Academic Press, New York 1981.
26. H. Reich: "Recent Developments in the Use of Silicon in Organic Synthesis," *Tetrahedron* **39** (1983) 839.
27. H. U. Reissig: "Siliciumverbindungen in der organischen Synthese," *Chem. Unserer Zeit* **18** (1984) 46.
28. B. Bogdanovic: "Katalytische Synthese von Organolithium- und -magnesium Verbindungen," *Angew. Chem.* **97** (1985) 253.
29. M. T. Reetz: *Organotitanium Reagents in Organic Synthesis,* Springer Verlag, Heidelberg 1986.
30. C. Paulmier: *Selenium Reagents and Intermediates in Organic Synthesis,* Pergamon Press, Oxford 1986.
31. M. Pereyre, J. P. Quintard, A. Rahm: *Tin in Organic Synthesis,* Butterworths, London 1987.
32. G. H. Robinson: *Coordination Chemistry of Aluminum,* VCH Verlagsgesellschaft mbH, Weinheim 1993.
33. H. K. Shama, K. H. Panell: "Activation of the Si-Si Bond by Transition Metal Complexes", *Chem. Rev.* **95** (1995) 1351 – 1374.
34. M. S. Holt, W. L. Wilson, J. H. Nelson: "Transition Metal-Tin Chemistry", *Chem. Rev.* **89** (1989) 11 – 49.
35. G. Wilke et al.: "Allyl-Übergangsmetall-Systeme," *Angew. Chem.* **78** (1966) 157.
36. E. O. Fischer: "Auf dem Weg zu Carben- und Carbinkomplexen," (Nobelvortrag), *Angew. Chem.* **86** (1974) 651.
37. E. L. Muetterties: "Molecular Metal Clusters," *Science* **196** (1977) 839.
38. R. Mason, D. W. Meek: "Die Vielseitigkeit tertiärer Phosphane als Liganden in der Koordinations- und Organometallchemie," *Angew. Chem.* **90** (1978) 195 – 206.
39. P. Chini: "Large Metal Carbonyl Clusters," *J. Organomet. Chem.* **200** (1980) 37 – 61.
40. R. Scheffold (ed.): "Transition Metals in Organic Synthesis," *Mod. Synth. Methods* **3** (1980).
41. B. F. G. Johnson, R. E. Benfield: "Stereochemistry of Transition Metal Carbonyl Clusters," *Top. Stereochem.* **12** (1981) 253.
42. R. Hoffmann: "Brücken zwischen Anorganischer und Organischer Chemie," (Nobelvortrag), *Angew. Chem.* **94** (1982) 725.
43. S. G. Davies: *Organotransition Metal Chemistry: Applications to Organic Synthesis,* Pergamon Press, Oxford 1982.
44. H. Vahrenkamp: "Basic Metal Cluster Reactions," *Adv. Organomet. Chem.* **22** (1983) 169.
45. C. M. Lukehart: *Fundamental Transition Metal Organometallic Chemistry,* Brooks/Cole, Monterey 1985.
46. A. Yamamoto: *Organotransition Metal Chemistry,* Wiley-Interscience, New York 1986.
47. B. Walther: "Metallclusterverbindungen," *Z. Chem.* **26** (1986) 421.
48. J. P. Collman, L. S. Hegedus, J. R. Norton, R. G. Finke: *Principles and Applications of Organotransition Metal Chemistry,* 2nd ed., University Science Books, Mill Valley, CA, 1987.
49. G. W. Parshall: "Trends and Opportunities for Organometallic Chemistry in Industry," *Organometallics* **6** (1987) 687 – 696.
50. G. Wilkinson, R. D. Gillard, J. E. McCleverty: *Comprehensive Coordination Chemistry,* vols. 1 – 7, Pergamon Press, Oxford 1987.
51. K. H. Dötz, R. W. Hoffman (eds.): *Organic Synthesis via Organometallics 3,* Vieweg, Braunschweig 1991.
52. D. Enders, H. -J. Gais, W. Keim (eds.): *Organic Synthesis via Organometallics 4,* Vieweg, Braunschweig 1993.
53. M. Gerloch, E. C. Constable: *Transition Metal Chemistry,* Wiley-Weinheim 1994.
54. E. C. Constable: *Metals and Ligand Reactivity,* Wiley-Weinheim 1996.
55. E. T. Edelmann: "Cyclopentadienylfreie Organolanthanoidchemie", *Chem. Rev.* **107** (1995) 2647 – 2669.
56. F. R. Hartley: *Chemistry of the Platinum Group Metals,* Elsevier, Amsterdam 1991.
57. J. D. Atwood: *Inorganic and Organometallic Reaction Mechanisms,* Brooks/Cole. Publ. Com., Monterey 1985.
58. A. S. Abu-Surrah, B. Rieger: *Angew. Chem.* **108** (1996) 2627 – 2629.
59. G. B. Deacon, Q. Shen: *J. Organomet. Chem.* **511** (1996) 1 – 17.

60. M. Hidai, Y Mizobe: "Recent Advances in the Chemistry of Dinitrogen Complexes", *Chem. Rev.* **95** (1995) 1115 – 1133.
61. F. Zaera: *Chem. Rev.* **95** (1995) 2651 – 2693.
62. R. J. Angelici: *Acc. Chem. Res.* **28** (1995) 1.
63. R. H. Crabtree: *Angew. Chem.* **105** (1993) 828 – 845.
64. G. A. Molander: "Application of Lanthanide Reagents in Organic Synthesis", *Chem. Rev.* **92** (1992) 29 – 68.
65. K. Eller, H. Schwarz: "Organometallic Chemistry in the Gas Phase", *Chem. Rev.* **92** (1991) 1121 – 1177.
66. M. I. Bruce: "Organometallic Chemistry of Vinylidene and Related Unsaturated Carbenes", *Chem. Rev.* **91** (1991) 197.
67. J. Chaiken: "Laser Chemistry of Organomettalics", *Chem. & Ind. (London)*1993, 4. Oct. 751 – 157. **107**
68. R. Noyori, M. Kitamura: "Enantioselective Addition von Organmetallreagenzien an Carbonylverbindungen", *Angew. Chem.* **103** (1991) 34 – 35.
69. J. A. Martinho Simoes, J. L. Beauchamp: "Transition Metal-Hydrogen and Metal Carbon Bond Strength: The Keys to Catalysis", *Chem. Rev.* **90** (1990) 629 – 688.
70. R. H. Crabtree: "Dihydrogen Complexes: Some Structural and Chemical Studies", *Acc. Chem. Res.* **23** (1990) 1.
71. H. Werner: "Komplexe von CO und seinen Verwandten", *Angew. Chem.* **102** (1990) 1109 – 1121.
72. G. Süss-Fink: *Angew. Chem.* **103** (1991) 73 – 74.
73. G. Gonzáles-Moraga: *Cluster Chemistry,* Springer, Heidelberg 1993.
74. G. Schmid (ed.): *Cluster and Colloids. From Theory to Applications,* Wiley-VCH, Weinheim 1994.
75. W. Kläui, *Angew. Chem.* **102** (1990) 661 – 670.
76. A.-M. Caminade, J.-P. Majoral, R. Mathieu, *Chem. Rev.* **91** (1991) 575 – 612.
77. N. M. Doherty, N. W. Hoffman, *Chem. Rev.* **91** (1991) 553 – 573.
78. D. H. Busch, *Chem. Rev.* **90** (1993) 847 – 860.
79. L. A. Belfiore, M. P. McCurdie, E. Ueda, *Macromolecules* **26** (1993) 6908 – 6917.
80. V. C. Gibson, *Angew. Chem.* **106** (1993) 1640 – 1648.
81. M. W. Göbel, *Angew. Chem.* **106** (1994) 1201 – 1203.
82. W. Keim: "Nickel: Ein Element mit vielfältigen Eigenschaften in der technisch-homogenen Katalyse", *Angew. Chem.* **102** (1990) 251 – 260.
83. G. Schmid et al., *J. Mol Catal. A: Chemical* **107** (1996) 95 – 104.
84. M. T. Reetz, S. A. Quaiser: *Angew. Chem.* **107** (1995) 2461 – 2463.
85. A. Fürstner, B. Bogdanovic: "Neue Entwicklungen in der Chemie von niedervalentem Titan", *Angew. Chem.* **108** (1996) 2583 – 2609.
86. R. O. Duthaler, A. Hafner, *Angew. Chem.* **109** (1997) 43 – 45.
87. Herrmann/Brauer: *Synthetics Methods of Organometallic and Inorganic Chemistry,* **1 – 8,** Georg Thieme Verlag, Stuttgart 1996.
88. H. Plenio, *Angew. Chem.* **109** (1997) 358 – 360.
89. "Directory of Catalysis", *Catalysis Today* **34** (1997) 1 – 2, p. 1.
90. B. Cetinkaya, J. Özdemir, P. H. Dixneuf, *J. Organomet. Chem.* **534** (1997) 153 – 158.
91. C. Köcher, W. A. Herrmann, *J. Organomet. Chem.* **532** (1997) 261 – 265.
92. M. Shibasaki, H. Sasai, T. Arai: "Asymmetrische Katalyse mit Hetero-Bimetall-Verbindungen", *Angew. Chem.* **109** (1997) 1291 – 1310.
93. M. Bochmann: *Mtetallorganische Chemie der Übergangsmetalle,* Wiley-CVH, Weinheim 1997.
94. A. Behr: (eds. C. -I. Bränden and G. Schneider): "Stoichiometric and catalytic activation of carbon dioxide on transition metal complexes", in :*"CO_2-fixation and CO_2-reduction in biological and model systems",* Oxford University Press, Oxford 1994, 93 – 109.
95. P. Kalck, F. Monteil: "Use of Water-Soluble Ligands in Homogeneous Catalysis", *Adv. Organomet. Chem.* **34** (1992) 219 – 284.
96. J. R. Anderson, E. M. Campi, W. R. Jackson, *Catalysis Letters* **9** (1991) 55 – 58.
97. A. Andriollo et al., *J. Mol Catal. A: Chemical* **116** (1997) 157 – 165.
98. J. P. Arhancet, M. E. Davis, B. E. Hanson, *Catalysis Letters* **9** (1991) 129 – 136.
99. J. P. Arhancet, M. E. Davis, B. E. Hanson, *J. Catal.* **129** (1991) 94 – 99.
100. J. P. Arhancet, M. E. Davis, B. E. Hanson, *J. Catal.* **129** (1991) 100 – 105.
101. J. P. Arhancet, M. E. Davis, J. S. Merola, B. E. Hanson, *J. Catal.* **121**(1990) 327 – 339.
102. H. Bahrmann et al., *J. Mol. Catal. A: Chemical* **116** (1997) 49 – 53.

103. J. Bakos, A. Orosz, B. Heil, M. Laghmari, P. Lhoste, D. Sinou, *J. C. S., Chem. Commun.* (1991) 1684 – 1685.
104. T. Bartik, B. Bartik, I. Guo, B. E. Hanson, *J. Organomet. Chem.* **480** (1994) 15 – 21.
105. T. Bartik, B. Bartik, B. E. Hanson, I. Guo, I. Tóth, *Organometallics* **12** (1993) 164 – 170.
106. M. Barton, J. D. Atwood, *J. Coord. Chem.* **24** (1991) 43 – 67.
107. W. Baidossé, N. Goren, J. Blum, H. Schumann, H. Hemling, *J. Mol. Catal.* **85** (1993) 153 – 162.
108. A. C. Bényei, S. Lehel, F. Jóo, *J. Mol. Catal. A: Chemical* **116** (1997) 349 – 354.
109. A. C. Bényei, F. Jóo, *J. Mol. Catal.* **58** (1990) 151 – 163.
110. R. Breslow: "Hydrophobic Effects on Simple Organic Reactions in Water", *Acc. Chem. Res.* **24** (1991) 159 – 164.
111. A. L. Casalnuovo, J. C. Calabrese, *J.A.C.S.* **112** (1990) 4324 – 4330.
112. T. H. Chan, C. J. Li, M. C. Lee, Z. Y. Wei, *Can. J. Chem.* **72** (1994) 1181.
113. W.-C. Chan, C.-P. Lau, L. Cheng Y. -S. Leung, *J. Organomet. Chem.* **464** (1994) 103 – 106.
114. R. V. Chaudhari, A. Bhattacharya, B. M. Bhanage, *Catalysis Today* **24** (1995) 123 – 133.
115. Y. Chauvin, S. Einloft, H. Olivier, *Ind. Eng. Chem. Res.* **34** (1995) 1149 – 1155.
116. Y. Chauvin, H. Olivier-Bourbigou, *CHEMTECH* 1995 Sept., 26 – 30.
117. Y. Chauvin, L. Mußmann, H. Olivier, *Angew. Chem.* **107** (1995) 2941 – 2943.
118. B. Cornils, *Nachr. Chem. Tech. Lab.* **42** (1994) 1136 – 1138.
119. B. Cornils: "Neue Wege der homogenen Zweiphasenkatalyse", *Angew. Chem.* **107** (1995) 1709 – 1711.
120. B. Cornils, E. Wiebus: "Aqueous catalysts for organic reactions", *CHEMTECH* (1995) Jan., 33 – 38.
121. B. Cornils, E. G. Kuntz, *J. Organomet. Chem.* **502** (1995) 177 – 186.
122. B. Cornils, W. A. Herrmann, R. W. Eckl, *J. Mol. Catal. A: Chemical* **116**(1997) 27 – 33.
123. D. J. Darensbourg, N. W. Stafford, F. Jóo, J. H. Reibenspies, *J. Organomet. Chem.* **488** (1995) 99 – 108.
124. D. J. Darenbourg et al., *Organometalics* **11** (1992) 1990 – 1993.
125. M. E. Davis, "Supported aqueous-phase catalysis", *CHEMTECH* (1992) Aug., 498 – 502.
126. R. M. Deshpande, P. Purwanto, H. Delmas, R. V. Chaudhari, *Ind. Eng. Chem. Res.* **35** (1996) 3927 – 3933.
127. E. V. Dehmlow, S. S. Dehmlow: *Phase Transfer Catalysis,* Wiley-VCH, Weinheim 1993.
128. H. Ding, J. Kang, B. E. Hanson, C. W. Kohlpaintner, *J. Mol. Catal. A: Chemical* **124** (1997) 21 – 28.
129. H. Ding, B. E. Hanson, T. E. Glass, *Inorg. Chim. Acta* **229** (1995) 329 – 333.
130. R. Dupree, F. Delanoe, J. Barbier, Jr., P. Isnard, G. Blanchard, *Catalysis Today* **29** (1996) 317 – 322.
131. J. Hagin: *C & EN,* 1994 10. Oct., 28 – 36.
132. O. Herd, et al., *J. Organmet. Chem.,* **522** (1996) 69 – 76.
133. M. Hernandez, P. Kalck, *J. Mol. Catal. A: Chemical* **116** (1997) 117 – 130.
134. W. A. Herrmann, C. W. Kohlpeitner: "Wasserlösliche Liganden, Metallkomplexe und Komplexkatalysatoren: Synergismen aus Homogen- und Heterogenkatalyse", *Angew. Chem.* **105** (1993) 1588 – 1609.
135. W. A. Herrmann, B. Cornils: "Metallorganische Homogenkatalyse – Quo Vadis?", *Angew. Chem.* **109** (1997) 1074 – 1095.
136. W. A. Herrmann, et al., *J. Mol. Catal. A:Chemical* **97** (1995) 65 – 72.
137. O. Herd, et al., *Angew. Chem.* **105** (1993) 1097 – 1099.
138. I. T. Horváth, *Catalysis Letter* **6** (1990) 43 – 48.
139. I. T. Horváth, J. Rábai, *Science* **266** (1994) 72 – 75.
140. Z. Jiang, A. Sen, *Macromolecules* **27** (1994) 7215 – 7216.
141. F. Jóo, L. Nádasdi, A. C. Benyei, D. J. Darenbourg, *J. Organomet. Chem.* **512** (1996) 45 – 50.
142. Z. Jin, X. Zheng, B. Fell, *J. Mol. Catal. A: Chemical* **116** (1997) 55 – 58.
143. Z. Jin, Y. Yan, H. Zuo, B. Fell, *J. Prakt. Chem.* **338** (1996) 124 – 128.
144. F. Jóo, A. Kathó, *J. Mol. Catal. A: Chemical* **116** (1997) 3 – 26.
145. S. Kanagasabapathy, Z. Xia, G. Papadogianakis, B. Fell, *J. Prakt. Chem.* **337** (1995) 446 – 450.
146. U. Koelle, *Coord. Chem. Rev.* **135/136** (1994) 623 – 650.
147. S. Kolaric, V. Sunjic, *J. Mol. Catal. A: Chemical* **110** (1996) 181 – 188.
148. C. W. Kohlpaintner, M. Beller, *J. Mol. Catal. A: Chemical* **116** (1997) 259 – 267.

149. C. Larpent, E. Bernard, *J.C.S., Chem. Commun.,* 1992 535 – 536.
150. C. Larpent, E. Bernard, F. Brisse-le-Menn, H. Patin, *J. Mol. Catal. A: Chemical* **116** (1997) 277 – 288.
151. S. Lemaire-Audoire, et al., *J. Mol. Catal. A: Chemical* **116** (1997) 247 – 258.
152. M. Lin, A. Sen, *Nature* **368** (1994) 613 – 615.
153. C. -J. Li: "Organic Reactions in Aqueous Media – with a Focus on Carbon-Carbon Bond Formation", *Chem. Rev.* **93** (1993) 2023 – 2035.
154. A. Lubineau, J. Augé, Y Queneau, *Synthesis* 1994 August, 741 – 760.
155. F. Mercier, F. Mathey, *J. Organomet. Chem.* **462** (1993) 103 – 106.
156. F. Monteil, P. Kalck, *J. Organomet. Chem.* **482** (1994) 45 – 51.
157. E. Monflier, et al., *J. Mol. Catal. A: Chemical* **97** (1995) 29 – 33.
158. E. Monflier, G. Fremy, Y. Castanet, A. Mortreux, *Angew. Chem.* **107** (1995) 2450 – 2452.
159. T. Müller, H. Bahrmann, *J. Mol. Catal. A: Chemical* **116** (1997) 39 – 42.
160. D. C. Mudalige, G. L. Rempel, *J. Mol. Catal. A: Chemical* **123** (1997) 15 – 20.
161. A. Oasmaa, A. Johansson, *Energie & Fuels* **7** (1993) 426 – 429.
162. G. Papadogianakis, R. A. Sheldon, *New J. Chem.* **20** (1996) 175 – 185.
163. E. Paetzold, G. Oehme, *J. Prakt. Chem.* **335** (1993) 181 – 184.
164. E. Paetzold, M. Michalik, G. Oehme, *J. Prakt. Chem.* **339** (1997) 38 – 43.
165. G. Papadogianakis, L. Maat, R. A. Sheldon, *J. Mol. Catal. A: Chemical* **116** (1997) 179 – 190.
166. I. Pennequin, J. Meyer, I. Suisse, A. Mortreux, *J. Mol. Catal. A: Chemical* **120** (1997) 139 – 142.
167. L. Plasserau, G. Süss-Fink, *J. Organomet. Chem.* **539** (1997) 163 – 170.
168. G. Pozzi, et al., *Tetrahedron* **52** (1996) 11879 – 11888.
169. T. Prinz, W. Keim, B. Drießen-Hölscher, *Angew. Chem.* **108** (1996) 1835 – 1836.
170. P. Purwanto, H. Delmas, *Catalysis Today* **24** (1996) 135 – 140.
171. D. M. Roundhill, *Adv. Organomet. Chem.* **38** (1995) 155 – 188.
172. S. Sabata, J. Vcelak, J. Hetflejs, *Collect. Czech. Chem. Commun.* **60** (1995) 127 – 136.
173. R. A. Sánchez-Delgado, M. Medina, F. López-Linares, A. Fuentes, *J. Mol. Catal. A: Chemical* **116** (1997) 167 – 177.
174. M. C. Schuster, K. H. Mortell, A. D. Hegemann, L. L. Kiessling, *J. Mol. Catal. A: Chemical* **116** (1997) 209 – 216.
175. T. G. Southern, *Polyhedron* **8** (1989) 407 – 413.
176. J. Struijk, M. d'Angremond, W. J. M. Lucas-de Regt, J. J. F. Scholten, *Appl. Catal. A: General* **83** (1992) 263 – 195.
177. G. Süss-Fink, A. Meister, G. Meister, *Coord. Chem. Rev.* **143** (1995) 97 – 111.
178. I. Tóth, B. E. Hanson, M. E. Davis, *Catalysis Letters* **5** (1990) 183 – 188.
179. I. Tóth, I. Guo, B. E. Hanson, *J. Mol. Catal. A: Chemical* **116** (1997) 217 – 229.
180. T. Viswanathan, J. Jethmalani, *J. Chem. Educ.* **70** (1993) 165 – 167.
181. S. Wache, *J. Organomet. Chem.* **494** (1995) 235 – 240.
182. K. T. Wan, M. E. Davis, *Nature* **370** (1994) 449 – 450.
183. K. T. Wan, M. E. Davis, *J. Catal.* **148** (1994) 1 – 8.
184. E. Wiebus, B. Cornils, *Chem-Ing.-Tech* **66** (1994) 916 – 923.
185. Z. Xia, B. Fell, *J. Prakt. Chem.* **339** (1997) 140 – 144.
186. R. Ugo (ed.): *Aspects of Homogeneous Catalysis,* **vol. 2 (1974), vol. 3 (1977), vol. 4 (1981), vol. 5 (1984), vol. 6 (1988),** Reidel Publ., Dordrecht.
187. H. Pracejus: *Koordinationschemische Katalyse organischer Reaktionen,* Steinkopff, Dresden 1977.
188. G. Henrici-Olivé, S. Olivé: *Coordination and Catalysis,* Verlag Chemie, Weinheim 1977.
189. J. K. Kochi: *Organometallic Mechanisms and Catalysis,* Academic Press, New York 1978.
190. G. W. Parshall: "Industrial Applications of Homogeneous Catalysis," *J. Mol. Catal.* **4** (1978) 243 – 270.
191. Y. Ishii, M. Tsutsui (eds.): *Fundamental Research in Homogeneous Catalysis,* vol. 2, Plenum Press, New York 1978.
192. M. Tsutsui (ed.): *Fundamental Research in Homogeneous Catalysis,* vol. 3, Plenum Press, New York 1979.
193. P. Heimbach, H. Schenkluhn: "Controlling Factors in Homogeneous Transition Metal Catalysis," *Top. Curr. Chem.* **92** (1980) 45.
194. G. W. Parshall, S. D. Ittel: *Homogeneous Catalysis,* Wiley-Interscience, New York 1992.

195. A. Nakamura, M. Tsutsui: *Principles and Applications of Homogeneous Catalysis,* Wiley-Interscience, New York 1980.
196. C. Masters: *Homogeneous Transition-Metal Catalysis – a Gentle Art,* Chapman and Hall, London 1981.
197. J. Falbe, H. Bahrmann: "Homogene Katalyse in der Technik," *Chem. Unserer Zeit* **15** (1981) 37.
198. E. R. F. Gesing: "Übergangsmetallkomplexe in der Synthese komplexer Naturstoffe," *Kontakte (Darmstadt)* 1982, no. 3, 13.
199. W. Keim: "Die homogene Übergangsmetallkatalyse und ihre industrielle Anwendung," *Chem. Ind. (Düsseldorf)* **107** (1984) 397.
200. H. Bönnemann: "Homogene Übergangsmetallkatalyse und ihr Potential in Forschung und Technik," *Österr. Chem. Z.* **85** (1984) 29.
201. J. F. Waller: "Recent Achievements, Trends and Prospects in Homogeneous Catalysis," *J. Mol. Catal.* **31** (1985) 123.
202. J. P. Casey (ed.): "Symposium: Industrial Applications of Organometallic Chemistry and Catalysis," *J. Chem. Educ.* **63** (1986) 188 – 225.
203. P. A. Chaloner: *Handbook of Coordination Catalysis in Organic Chemistry,* Butterworths, London 1986.
204. A. Mortreux, F. Petit (eds.): *Industrial Applications of Homogeneous Catalysis,* Reidel Publ., Dordrecht 1988.
205. B. Cornils, W. A. Herrmann, M. Rasch: "Otto Roelen als Wegbereiter der industriellen homogenen Katalyse", *Angew. Chem.* **106** (1994) 2219 – 2238.
206. W. A. Herrmann: "Metallorganische Chemie in der industriellen Katalyse", *Kontakte (Darmstadt)* **1** (1991) 22.
207. B. Cornils, W. A. Herrmann: *Applied Homogeneous Catalysis with Organomettalic Compounds,* **Vols. 1 and 2,** Wiley-VCH Verlag, Weinheim, 1996.
208. B. Cornils, W. A. Herrmann: *Aqueous-Phase Organometallic Catalysis – Concepts and Applications,* Wiley-VCH Verlag, Weinheim, 1998 (in print).
209. W. A. Nugent, T. V. Rajanbabu, M. J. Burk: "Enantioselective Catalysis in Industry", *Science* **259** (1993) 479 – 483.
210. W. R. Moser, D. W. Slocum: "Homogeneous Transition Metal Catalyzed Reactions", in *Advances in Chemistry Series No. 230,* American Chemical Society, Washington, DC (1992)
211. A. Togni, L. M. Venanzi: "Stickstoffdonoren in der Organometallchemie und in der Homogenkatalyse", *Angew. Chem.* **106** (1994) 517 – 547.
212. J. Hagen: *Technische Katalyse – Eine Einführung,* Wiley-VCH Verlag, Weinheim 1996.
213. L. S. Hegedus: *Transition Metals in the Synthesis of Complex Organic Molecules,* University Science Books, Mill Valley 1994, *Organische Synthese mit Übergangsmetallen,* VCH Verlagsgesellschaft mbH, Weinheim 1995.
214. B. M. Trost: "Atomökonomische Synthesen – eine Herausforderung in der Organischen Chemie: die Homogenkatalyse als wegweisende Methode", *Angew. Chem.* **107** (1995) 285 – 307.
215. R. Whyman: "Let's watch homogeneous catalysis", *CHEMTECH* (1991) July 414 – 419.
216. E. S. Gore: "Platinum Group Metals in Phase Transfer Catalysis", *Platinum Metals Rev.* **34** (1990) 2 – 9.
217. S. L. Blystone: "Synthetic Application of Enantioselective Organotransition-Metal-Mediated Reactions", *Chem. Rev.* **89** (1989) 1663 – 1679.
218. G. Consiglio, R. M. Waymouth: "Enantioselective Homogeneous Catalysis Involving Transition-Metal-Allyl Intermediates", *Chem. Rev.* **89** (1989) 257 – 276.
219. N. Koga, K. Morokuma: "Ab initio Molecular Orbital Studies of Catalytic Elementary Reactions and Catalytic Cycles of Transition-Metal Complexes", *Chem. Rev.* **91** (1991) 823 – 842.
220. R. E. Malz, Jr. (ed.): *Catalytics of Organic Reactions,* Marcel Dekker, Inc., New York 1996.
221. W. A. Herrmann (ed.): *Organolanthanoid Chemistry: Synthesis, Structure, Catalysis,* Springer-Verlag, Berlin 1996.
222. C. L. Hill (ed.): "Polyoxometalates in Catalysis", *J. Mol. Catal. Special issue* **114** (1996) No. 1 – 3.
223. H. Steinhagen, G. Helmchen: "Asymmetrische Zweizentren-Katalyse – von der Natur lernen", *Angew. Chem.* **108** (1996) 2489 – 2492.
224. O. Reiser: "Kombinatorische Methoden – Chancen in der Katalyse?", *Nachr. Chem. Tech. Lab.* **44**(1996) 1182 – 1188.

225. C. B. Dartt, M. E. Davis: "Catalysis for Environmentally Benign Processing", *Ind. Eng. Chem. Res.* **33** (1994) 2887 – 2899.
226. P. G. Jessop, T. Ikariya, R. Noyori: "Homogeneous Catalysis in Supercritical Fluids", *Science* **269** (1995) 1065 – 1069.
227. D. J. Berrisford, C. Bolm, K. B. Sharpless: "Ligandenbeschleunigte Katalyse", *Angew. Chem.* **107** (1995) 1159 – 1171.
228. R. Breslow: "The chelate effect in binding, catalysis and chemotherapy", *Recl. Trav. Chim. Pays-Bas* **113** (1994) 493 – 498.
229. R. A. Sández-Delgado, M. Rosales, M. A. Esteruelas, L. A. Oro: "Homogeneous catalysis by osmium complexes. A review", *J. Mol. Catal. A: Chemical* **96** (1995) 231 – 243.
230. D. A. Tomalia, P. R. Dvornic: "What promise for dendrimers?", *Nature* **372** (1994) 617 – 626.
231. R. Breslow: "Biomimetic Chemistry and Artificial Enzymes: Catalysis by Design", *Acc. Chem. Res.* **28** (1995) 146 – 153.
232. G. Süss-Fink, G. Meister: "Transition Metal Clusters in Homogeneous Catalysis", *Adv. Organomet. Chem.* **35** (1993) 41 – 134.
233. A. Behr: "Anwendungsmöglichkeiten der homogenen Übergangsmetallkatalyse in der Fettchemie", *Fat Sci. Technol.* **92** (1990) 375 – 388.
234. J. E. Arnet, R. Pettit, *J. Am. Chem. Soc.* **83** (1961) 2954.
235. J. L. Eisenmann, *J. Org. Chem.* **27** (1962) 2706.
236. J. F. Biellmann, M. J. Jung, *J. Am. Chem. Soc.* **90** (1968) 1673.
237. P. Z. Bedoukian, *Am. Perfum. Cosm.* **86** (1971) 25 –36.
238. C. P. Casey, C. R. Cyr, *J. Am. Chem. Soc.* **95** (1973) 2248.
239. E. J. Corey, J. W. Suggs, *J. Org. Chem.* **38** (1973) 3224.
240. D. Bingham, B. Hudson, D. E. Webster, P. B. Wells, *J. Chem. Soc. Dalton Trans.* 1974, 1521.
241. R. G. Miller et al., *J. Am. Chem. Soc.* **96** (1974) 4211, 4229.
242. M. Gargano, P. Giannoccaro, H. Rossi, *J. Organomet. Chem.* **84** (1975) 389.
243. K. C. Bishop, *Chem. Rev.* **76** (1976) 461.
244. C. F. Lochow, R. G. Miller, *J. Org. Chem.* **41** (1976) 3020.
245. J. Andrieux, D. H. R. Barton, H. Patin, *J. Chem. Soc. Perkin Trans. I,* 1977, 359.
246. J. Andrieux, D. H. R. Barton, H. Patin, *J. Chem. Soc. Perkin Trans. I,* 1977, 359.
247. M. Tuner, J. v. Jouanne, H.-D. Brauer, H. Kelm, *J. Mol. Catal.* **5** (1979) 425, 433, 447.
248. M. Franck-Neumann, F. Brion, *Angew. Chem. Int. Ed. Engl.* **18** (1979) 688.
249. H. C. Brown, U. S. Racherla, H. Taniguchi, *J. Org. Chem.* **46** (1981) 4313.
250. H. C. Brown, U. S. Racherla, *J. Organomet. Chem.* **241** (1983) C 37 – C 40.
251. H. Taketoni et al., *J. Am. Chem. Soc.* **106** (1984) 5208 – 5217.
252. *Houben Weyl,* **E 18,** 1146.
253. A. Basu, T. G. Kasar, *J. Am. Oil. Chem. Soc.* **63** (1986) 1444 – 1447.
254. N. Satyanarayana, M. Periasamy, *J. Organomet. Chem.* **319** (1987) 113 – 117.
255. A. Rosenfeld, J. Blum, D. Avnir, *J. Catal.* **164** (1996) 363 – 368.
256. S. Bellemin-Laponnaz, H. Gisie, J. P. Le Ny, J. A. Osborn, *Angew. Chem.* **109** (1997) 1011.
257. L. Vaska, R. E. Rhodes, *J. Am. Chem. Soc.* **87** (1965) 4970.
258. B. R. James: *Homogeneous Hydrogenation,* Wiley-Interscience, New York 1973.
259. F. J. McQuillin: *Homogeneous Hydrogenation in Organic Chemistry,* Reidel Publ., Dordrecht 1976.
260. P. N. Rylander: *Catalytic Hydrogenation in Organic Syntheses,* Academic Press, New York 1979.
261. M. Fiorini, G. M. Giongo, *J. Mol. Catal.* **5** (1979) 303.
262. M. Fiorini, G. M. Giongo, *J. Mol. Catal.* **7** (1980) 411.
263. J. C. Chabala, M. H. Fischer, US 4 199 569, 1980.
264. J. Halpern, *J. Organomet. Chem.* **200** (1980) 133.
265. P. J. Brothers, *Prog. Inorg. Chem.* **28** (1981) 1.
266. F. H. Jardine, *Prog. Inorg. Chem.* **28** (1981) 63.
267. B. R. James in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, p. 285.
268. J. W. Faller in L. H. Pignolet (ed.): *Homogeneous Catalysis with Metal Phosphine Complexes,* Chap. 2, Plenum Publishing, New York 1983.
269. W. S. Knowles: "Asymmetric Hydrogenation," *Acc. Chem. Res.* **16** (1983) 106.
270. W. S. Knowles, *J. Chem. Educ.* **63** (1986) 222.
271. H. Takaya et al., *J. Am. Chem. Soc.* **109** (1987) 1596.
272. K. -J. Haack et al., *Angew. Chem.* **109** (1997) 297 – 300.

273. H. S. Hilal, W. Jondi, S. Khalaf, R. Abu-Halawa, *J. Organomet. Chem.* **452** (1992) 161 – 165.
274. A. Thomas, M. Haake, F. -W. Grevels, J. Bargon, *Angew. Chem.* **106** (1994) 820 – 822.
275. P. A. Chaloner, M. Esteruelas, F. Jóo, L. A. Oro: *Homogeneous Hydrogenation,* Kluwer Academic Publ., Dordrecht 1994.
276. W. R. Cullen, S. J. Rettig, E. A. Wickenheiser, *J. Mol. Catal.* **66** (1991) 251 – 269.
277. I. T. Horváth, *Angew. Chem.* **103** (1991) 1042 – 1044.
278. G. Allmang, F. Grass, J. M. Grosselin, C. Mercier, *J. Mol. Catal.* **66** (1991) L27 – L31.
279. J. M. Grosselin, C. Mercier, G. Allmang, F. Grass, *Organometallics* **10** (1991) 2126 – 2133.
280. J. Struijk, R. Moene, T. van der Kamp, J. J. F. Scholten, *Appl. Catal. A.: General* **89** (1992) 77 – 102.
281. E. Bayer, L. Tang, D. Waidelich, LM. Kutubuddin, *Fat Sci. Technol.* **94** (1992) 79 – 82.
282. B. Fell, G. Gurke, *Chemiker-Zeitung* **115** (1991) 83 – 87.
283. Y. Zhang, S. Liao, Y. Xu, S. Chen, *J. Organomet. Chem.* **382** (1990) 69 – 76.
284. B. R. James, *Catalysis Today* **37** (1997) 209 – 221.
285. M. Boutonnet Kizling, C. Bigey, R. Touroude. *Appl. Catal. A.: General* **135** (1996) L13 – L17.
286. A. Behr, et al.: "Selektive Härtung mehrfach ungesättigter Fettsäuren in der Flüssigphase", *Fat Sci. Technol.* **95** (1993) 2 – 12.
287. A. Behr, H. Schmidke: "Selektive Hydrierung ungesättigter Modellverbindungen mit solvensstabilisierten Palladium-Kolloidkatalysatoren", *Chem.-Ing.-Tech.* **65** (1993) 568 – 569.
288. T. Alderson, E. L. Jenner, R. V. Lindsey, *J. Am. Chem. Soc.* **87** (1965) 5698.
289. R. Cramer, *J. Am. Chem. Soc.* **89** (1967) 4621.
290. W. Brenner et al., *Justus Liebigs Ann. Chem.* **727** (1969) 161 – 182.
291. G. Lefebvre, Y. Chauvin in R. Ugo (ed.): *Aspects of Homogeneous Catalysis,* **vol. 1,** Manfredi, Milan 1970, p. 107.
292. M. G. Barlow, M. J. Bryant, R. N. Hazeldine, A. G. Mackie, *J. Organomet. Chem.* **21** (1970) 215.
293. A. Yamamoto, S. Kitazume, L. S. Pu, S. Ikeda, *J. Am. Chem. Soc.* **93** (1970) 371.
294. G. Henrici-Olivé, S. Olivé, *Adv. Polym. Sci.* **15** (1974) 1.
295. M. Hidai, A. Misono in R. Ugo (ed.): *Aspects of Homogeneous Catalysis,* **vol. 2,** Reidel Publ., Dordrecht 1974, p. 39.
296. W. Keim, F. H. Kowaldt, R. Goddard, C. Krüger, *Angew. Chem. Int. Ed. Engl.* **17** (1978) 466.
297. R. F. Heck, *Acc. Chem. Res.* **12** (1979) 146.
298. B. Bogdanovic, *Adv. Organomet. Chem.* **17** (1979) 105 – 140.
299. A. C. L. Su, *Adv. Organomet. Chem.* **17** (1979) 269.
300. K. P. C. Vollhardt, *J. Am. Chem. Soc.* **103** (1981) 6259.
301. W. Keim, A. Behr, M. Röper in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, p. 371.
302. P. Grenouillet, D. Neibecker, I. Tkatchenko, *Organometallics* **3** (1984) 1130.
303. A. Behr in R. Ugo (ed.): *Aspects of Homogeneous Catalysis,* **vol. 5,** Reidel Publ., Dordrecht 1984, pp. 3 – 73.
304. K. P. C. Vollhardt: "Cobalt-vermittelte 2 + 2 + 2-Cycloadditionen," *Angew. Chem.* **96** (1984) 525.
305. W. Keim, *Chem. Ing. Tech.* **56** (1984) 850.
306. H. Bönnemann: "Organocobaltverbindungen in der Pyridinsynthese," *Angew. Chem.* **97** (1985) 264.
307. *Houben-Weyl,* **E 18.**
308. G. P. Chiusoli, *J. Organomet. Chem.* **300** (1986) 57 –80.
309. B. M. Trost, *J. Organomet. Chem.* **300** (1986) 263 –280.
310. A. Behr in A. Mortreux, F. Petit (eds.): *Industrial Applications of Homogeneous Catalysis,* Reidel Publ., Dordrecht 1988, pp. 141 – 175.
311. S. J. Brown, A. F. Masters, *J. Organomet. Chem.* **367** (1989) 371 – 374.
312. V. Rautenstrauch, P. Mégard, J. Conesa, W. Küster: "2–Pentylcyclopent-2–en-1–on durch katalytische Pauson-Khand-Reaktion", *Angew. Chem.* **102** (1990) 1441 – 1443.
313. W. Schulz, H. Pracejus, G. Oehme: "New three-step synthesis of 2–vinylpyridine starting from acrylonitrile", *J. Mol. Catal.* **66** (1991) 29 – 36.
314. J. Strupinska: "Oligomerization of α-Olefins to Higher Oligomers", *Chem. Rev.* **91** (1991) 613 – 648.
315. R. M. Waymouth, K. S. Knight: "Constructing from simple olefins", *CHEMTECH* (1995) April, 15 – 21.

316. C. Siv, G. Peiffer, A. Bendayan, *J. Organomet. Chem.* **525** (1996) 151 – 153.
317. P. W. Jolly, N. Kokel, *Synethesis* (1990) Sept., 771 – 773.
318. H. Anderson, H. Heldt, K. Heldt, *Chem.-Ing.-Tech.* **66** (1994) 69 – 72.
319. M. Camargo, P. Dani, J. Dupont, R. F. de Souza, M. Pfeffer, J. Tkatchenko, *J. Mol. Catal. A: Chemical* **109** (1996) 127 – 131.
320. K. Hill, B. Gruber, K. J. Weese, *Tetrahedron Lett.* **26** (1994) 4541 – 4542.
321. G. Peiffer, S. Chhan, A. Bendayan, B. Waegell, J.-P. Zahra, *J. Mol. Catal.* **59** (1990) 1 – 9.
322. D. J. Berrisford, C. Bolm, *Angew. Chem.* **107** (1995) 1862 – 1864.
323. K. Ohe, M. Kojima, K. Yonehara, S. Uemura, *Angew. Chem.* **108** (1996) 1959 – 1962.
324. P. Wagler, B. Heller, J. Ortner, K.-H. Funken, G. Oehme, *Chem.-Ing.-Tech.* **68** (1996) 823 – 826.
325. H. B. Kagan, O. Riant: "Catalytic Asymmetric Diels-Alder-Reactions", *Chem. Rev.* **92** (1992) 1007 – 1019.
326. Y. Ohgomori, S. Ichikawa, N. Sumitani: "Tail-to-Tail Dimerization Reaction of Acrolein", *Organometallics* **13** (1994) 3758 – 3760.
327. H.-G. Schmalz: "Katalytische Ringschluß-Metathese", *Angew. Chem.* **107** (1995) 1981.
328. G. Roso-Levi, I. Amer, *J. Mol. Catal. A: Chemical* **106** (1996) 51 – 56.
329. M. Schäfer, N. Mahr, J. Wolf, H. Werner, *Angew. Chem.* **105** (1993) 1377 – 1379.
330. G. J. P. Britovsek, W. Keim, S. Mecking, D. Sainz, T. Wagner, *J.C.S. Chem. Commun.* (1993) 1632 – 1634.
331. J. H. Rigby, *Acc. Chem. Res.* **26** (1993) 579 – 585.
332. H. Butenschön, *Angew. Chem.* **106** (1994) 664 – 666.
333. S. Murai et al., *Nature* **366** (1993) 529 – 531.
334. H. Werner, M. Schäfer, J. Wolf, K. Peters, H. G. von Schnering, *Angew. Chem.* **107** (1995) 213 – 215.
335. J. Iqbal, B. Bhatia, N. K. Nayyar, *Chem. Rev.* **94** (1994) 519 – 564.
336. E. P. Kündig, B. Bourdin, G. Bernardinelli, *Angew. Chem.* **106** (1994) 1931 – 1934.
337. J. E. Hill, et al., *Organometallics* **10** (1991) 3428 – 3430.
338. B. M. Trost, *Acc. Chem. Res.* **23** (1990) 34 – 42.
339. E. J. Vandenberg, J. C. Salamone (eds.): *Catalysis in Polymer Synthesis,* ACS Symposium Series 496, Washington, DC 1992.
340. W. Kaminsky: "Zirconocene catalysts for olefin polymerization", *Catalysis Today* **20** (1994) 257 – 271.
341. K. B. Sinclair, R. B. Wilson: "Metallocene Catalysts: A revolution in Olefin Polymerisation", *Chem. & Ind. (London)* (1994) Nov., 857 – 862.
342. P. C. Möhring, N. J. Coville: "Homogeneous Group 4 metallocene Ziegler-Natta catalysts", *J. Organomet. Chem.* **479** (1994) 1 – 29.
343. A.. W. Al-Sa'doun, *Appl. Catal. A.: General* **105** (1993) 1 – 40.
344. R. D. Leaversuch: "Metallocene resins", *Mod. Plastics Int.,* (1994) June, 38 – 40.
345. G. Ricci, S. Italia, *Macromol. Chem. Phys.* **195** (1994) 1389 – 1397.
346. W. Kaminsky, S. Lenk, *Macromol. Chem. Phys.* **195** (1994) 2093 – 2105.
347. H. Cherdron, M. -J. Brekner, F. Osan: "Cycloolefin – Copolymere", *Angew. Makromol. Chem.* **223** (1994) 121 – 133.
348. W. Kaminsky: "Olefinpolymerisation mittels Metallocenkatalysatoren", *Angew. Makromol. Chem.* **223** (1994) 101 – 120.
349. W. Keim, R. Nücker, *Chem.-Ing.-Tech.* **66** (1994) 950 – 951.
350. M. van Halteren, W. Keim, *Chem.-Ing.-Tech.* **64** (1992) 1117 – 1118.
351. R. Ceder, G. Muller, J. Sales, J. Vidal, D. Neibecker, I. Tkatchenko, *J. Mol. Catal.* **68** (1991) 23 – 31.
352. W. Keim, Z. Chen, Z. Shen, *Catalysis Lett.* **10** (1991) 233 – 238.
353. P. Brun, A. Tenaglia, *J. Mol. Catal.* **70** (1991) L1 – L3.
354. P. Denis, J. -F. Croizy, A. Mortreux, F. Petit, *J. Mol. Catal.* **68** (1991) 159 – 175.
355. K. Nomura, M. Ishino, *J. Mol. Catal.* **68** (1991) L5 – L7.
356. G. L. Tembe, M. Ravindranathan, *Ind. Eng. Chem. Res.* **30** (1991) 2247 – 2252.
357. W. Keim, W. Meltzow, Z. Chen, Z. Huang, Z. Shen, *J. Mol. Catal.* **137** (1992) 423 – 428.
358. S. M. Pillai, G. L. Tembe, M. Ravindranathan, *Appl. Catal. A.: General* **81** (1992) 273 – 278.
359. E. Drent, J. A. M. van Broekhoven, , M. J. Doyle, *J. Organomet. Chem.* **417** (1991) 235 – 251.
360. W. Keim, H. Maas: "Copolymerization of ethylene and carbon monoxide", *J. Organomet. Chem.* **514** (1996) 271 – 276.
361. P. S. Subramanian, K. J. Chou, *Trends in Polym. Science* **3** (1995) 324 – 329.

362. C. Bergemann, R. Cropp, G. Luft, *Chem.-Ing.-Tech.* **68** (1996) 417 – 419.
363. B. Heller, D. Heller, G. Oehme, *J. Mol. Catal. A: Chemical* **100** (1996) 211 – 219.
364. M. Bianchi, G. Menchi, U. Matteoli, F. Piacenti, *J. Organomet. Chem.* **526** (1996) 33 – 41.
365. Y. Chauvin et al., *J. Catal.* **165** (1997) 275 – 278.
366. H. K. Hall, Jr.: "Atom economy and monomer synthesis", *CHEMTECH* (1997) June, 20 – 25.
367. R. G. da Rosa, M. O. de Souza, R. F. de Souza: "Oligomerization and co-oligomerization of α-olefins", *J. Mol. Catal. A: Chemical* **120** (1997) 55 – 62.
368. G. J. Meuzelaar, L. Maat, R. A. Sheldon, I. V. Kochernikow: "Heteropoly acid-catalyzed Diels-Alder reactions", *Catal. Lett.* **45** (1997) 249 – 251.
369. A. Behr, A. Laufenberg: "Synthese neuer verzweigter Fettsäureester durch homogene Rhodiumkatalyse", *Fat Sci. Technol.* **93** (1991) 20 – 24.
370. A. Behr, S. Bellmann, H. -P. Handwerk: "Katalytische Oligomerisierung von Fettsäuren", *Fat Sci. Technol.* **93** (1991) 340 – 345.
371. A. Behr, H. -P. Handwerk: "Übergangsmetallkatalysierte Addition von Maleinsäureanhydrid an Fettstoffe", *Fat Sci. Technol.* **94** (1992) 204 – 208.
372. A. Behr, H. -P. Handwerk: "Katalytische Verknüpfung ungesättigter Fettstoffe mit Formaldehyd", *Fat Sci. Technol.* **94** (1992) 443 – 447.
373. A. Behr, H. -P. Handwerk: "Katalytische Addition von Acryl- und Vinylverbindungen an ungesättigte Fettstoffe", *Chem.-Ing.-Tech.* **65** (1993) 1377 – 1379.
374. T. J. Katz: "The Olefin Metathesis Reaction," *Adv. Organomet. Chem.* **16** (1977) 283.
375. N. Calderon, J. P. Lawrence, E. A. Ofsteadt: "Olefin Metathesis," *Adv. Organomet. Chem.* **17** (1979) 449.
376. R. Streck: "Some Applications of the Olefin Metathesis Reaction to Polymer Synthesis," *J. Mol. Catal.* **15** (1982) 3 – 19.
377. R. H. Grubbs in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, p. 499.
378. J. C. Mol: "Metathesis of Functionalized Olefins," *CHEMTECH* 1983, April, 250 – 255.
379. K. J. Ivin: *Olefin Metathesis,* Academic Press, London 1983.
380. S. Warwel, H. Ridder, G. Hachen: "Olefinsynthesen durch Metathese," *Chem. Ztg.* **107** (1983) 115 – 120.
381. V. Dragutan, A. T. Balaban, M. Dimonie: *Olefin Metathesis and Ring Opening Polymerization of Cyclo-Olefins,* Wiley-Interscience, Chichester 1985.
382. *Houben-Weyl,* **E 18,** 1163 – 1223.
383. X. Xiaoding, J. I. Benecke, C. Boelhouwer, J. C. Mol, *Appl. Catal.* **28** (1986) 271 – 283.
384. R. L. Banks: "Discovery and Development of Olefin Disproportionation," *CHEMTECH* 1986, February, 112 – 117.
385. S. Warwel: "Industrielle Prozesse der Olefin-Metathese," *Erdöl-Erdgas-Kohle* **103** (1987) 238 – 245.
386. A. Ellison (ed.): "Results of the 7. Int. Symp. on Olefin Metathesis," *J. Mol. Catal.* **46** (1988).
387. C. M. Huwe, J. Velder, S. Blechert, *Angew. Chem.* **108** (1996) 2542 – 2544.
388. M. Schuster, J. Pernerstorfer, S. Blechert, *Angew. Chem.* **108** (1996) 2111 – 2112.
389. P. A. von der Schaaf, A. Hafner, A. Mühlebach, *Angew. Chem.* **108** (1996) 1974 – 1977.
390. K. Weiss, U. Neugebauer, S. Blau, H. Lang, *J. Organomet. Chem.* **520** (1996) 171 – 179.
391. W. J. Feast, E. Khosravi (eds.), *J. Mol. Catal. A.: Chemical* **115** (1997), Special issue: Olefin Metathesis and Related Chemistry.
392. W. J. Zuercher, M. Hashimoto, R. H. Grubbs, *J. Am. Chem. Soc.* **118** (1996) 6634 – 6640.
393. P. Schwab, M. B. France, J. W. Ziller, R. H. Grubbs, *Angew. Chem.* **107** (1995) 2179 – 2181.
394. W. A. Herrmann, W. C. Schattenmann, O. Nuyken, S. C. Glander, *Angew. Chem.* **108** (1996) 1169 – 1170.
395. P. Dounis, W. J. Feast, A. M. Kenwright: "Ring-opening metathesis polymerization", *Polymer* **36** (1995) 2787 – 2796.
396. U. Koert: "Ringschluß-Olefin-Metathese", *Nachr. Chem. Tech. Lab.* **43** (1995) 809 – 814.
397. Z. Wu, A. D. Benedicto, R. H. Grubbs, *Macromolecules* **26** (1993) 4975 – 4977.
398. M. B. France, R. A. Paciello, R. H. Grubbs. *Macromolecules* **26** (1993) 4739 – 4741.
399. J.-L. Couturier, K. Tanaka, M. Leconte, J.-M. Basset, J. Ollivier, *Angew. Chem.* **105** (1993) 99 – 102.
400. P. H. Wagner: "Olefin Metathesis-Applications for the Nineties", *Chem. & Ind. (London),* (1992) May, 330 – 333.

401. W. A. Herrmann, W. Wagner, U. N. Flessner, U. Volkhardt, H. Kombler: "Methyltrioxorhenium als Katalysator für die Olefin-Metathese", *Angew. Chem.* **109** (1991) 1704 – 1706.
402. H. L. Kraus (ed.), *J. Mol. Catal.* **65** (1991) No 1 – 2, Special issue: Olefin Metathesis.
403. P. Buchacher, W. Fischer, K. D. Aichholzer, F. Stelzer, *J. Mol. Catal. A.: Chemical* **115** (1997) 163 – 171.
404. M. J. Marsella, H. D. Maynard, R. H. Grubbs, *Angew. Chem.* **109** (1997) 1147 – 1150.
405. J. E. Bäckvall, B. Akermark, S. O. Ljunggren: "The Wacker-Process," *J. Am. Chem. Soc.* **101** (1979) 2411.
406. E. S. Gore: "Ruthenium Catalysed Oxidations of Organic Compounds," *Platinum Met. Rev.* **27** (1983) 111 – 125.
407. J.-E. Bäckvall: "Palladium in some Selective Oxidation Reactions," *Acc. Chem. Res.* **16** (1983) 335 – 342.
408. D. Schnurpfeil: "Über die katalysierte Flüssigphasenoxidation von Olefinkohlenwasserstoffen," *Wiss. Z. Tech. Hochsch. Carl Schorlemmer Leuna Merseburg* **26** (1984) 411 – 425.
409. A. H. Haines (ed.): *Methods for Oxidation of Organic Compounds,* Academic Press, London, New York 1985.
410. J. F. Perez-Benito, D. G. Lee: "Oxidation of Hydrocarbons," *Can. J. Chem.* **63** (1985) no. 12, 3545 –3550.
411. P. A. Chaloner in: *Handbook of Coordination Catalysis in Organic Chemistry,* Butterworth, London 1986, pp. 451 – 679.
412. *Houben-Weyl,* **E 18,** 1077 – 1136.
413. W. J. Mijs, C. R. H. I. De Jonge (eds.): *Organic Syntheses by Oxidation with Metal Compounds,* Plenum Press, New York 1986, pp. 1 – 908.
414. J. E. Lyons, G. Suld, C.-Y. Hsu: "Multiple Roles of Palladium in Liquid Phase Oxidation," in: Y. Yermakov, V. Likholobov (eds.): *Proc. Int. Symp. Relat. Heterog. Homogeneous Catal. Phenom.,* 5th, Novosibirsk 1986, VNU Science Press, Utrecht, pp. 117 –138.
415. M. Wenkin, R. Touillaux, P. Ruiz, B. Delmon, M. Devillers, *Appl. Catalysis A.: General* **148** (1996) 181 – 199.
416. S.-J. Murahashi, *Angew. Chem.* **107** (1995) 2670 – 2693.
417. R. A. Sheldon: "Fine chemicals by catalytic oxidation", *CHEMTECH* (1991) Sept., 566 – 576.
418. S.-J. Murahashi et al., *J. Org. Chem.* **55** (1990) 1744 – 1749.
419. J.-E. Bäckvall, R. L. Chowdhury, U. Karlsson, *J.C.S., Chem. Commun.,* 1991, 473 – 475.
420. W. A. Herrmann, R. W. Fischer, D. W. Marz, *Angew. Chem.* **103** (1991) 1706 – 1709.
421. G. Speier: "Organometallic oxidation catalysts", in *The Chemistry of the Metal-Carbon Bond,* **Vol. 5** , chap. 5, Wiley, New York 1989, pp. 147 – 198.
422. W. P. Griffith, A. M. Z. Slawin, K. M. Thompson, D. J. Williams, *J.C.S. Chem. Commun.* (1994) 569 – 570.
423. R. A. T. M. van Beuthem et al., *Angew. Chem.* **107** (1995) 500 – 503.
424. W. Adam, W. A. Herrmann, J. Lin, C. R. Saha-Möller, R. W. Fischer, J. D. G. Correia, *Angew. Chem.* **106** (1994) 2545 – 2546.
425. E. Monflier, E. Blouet, Y. Barbaux, A. Mortreux, *Angew. Chem.* **106** (1994) 2183 – 2185.
426. T. Mallat, A. Baiker, *Catalysis Today* **19** (1994) 247 – 284.
427. S. Wilson, *Chem. &. Ind.,* (1994), April, 255 – 258.
428. B. Meunier, *Chem. Rev.* **92** (1992) 1411 – 1456.
429. J. Sundermeyer, *Angew. Chem.* **105** (1993) 1195 – 1197.
430. T. T. Wenzel, *J.C.S. Chem. Commun.* 1993, 862 – 864.
431. R. A. Sheldon (ed.), *J. Mol. Catal. A.: Chemical* **117** (1997) No 1 – 3, special issue: Activation of Dioxygen and Homogeneous Catalytic Oxidation.
432. E. Gusevskaya, J. A. Gonsalves, *J. Mol. Catal. A.: Chemical* **121** (1997) 131 – 137.
433. A. M. Romano, M. Ricci: "New efficient catalysts for the aerobic oxidation of ethers", *J. Mol. Catal. A.: Chemical* **120** (1997) 71 – 74.
434. A. Behr, K. Eusterwiemann: "Controlling factors in the Ruthenium-catalyzed oxidation of primary alcohols with sodium bromate", *J. Organomet. Chem.* **403** (1991) 209 – 214.
435. A. Behr, K. Eusterwiemann: "Ruthenium-catalyzed oxidation of primary alcohols with the new oxidation system N-methylmorpholine/hydrogen peroxide", *J. Organomet. Chem.* **403** (1991) 215 – 219.
436. J. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls,* **vol. 1 (1968), vol. 2 (1977),** Wiley-Interscience, New York.
437. B. Fell, M. Barl, *J. Mol. Catal.* **2** (1977) 301.
438. E. L. Muetterties, J. Stein, *Chem. Rev.* **79** (1979) 479.

439. F. G. A. Stone, R. West (eds.): *Advances in Organometallic Chemistry,* **vol. 17,** Academic Press, New York 1979.
440. J. Falbe: *New Syntheses with Carbon Monoxide,* Springer-Verlag, Berlin 1980.
441. J. F. Knifton: "Syngas Reactions," *J. Organomet. Chem.* **188** (1980) 223 – 236.
442. J. F. Knifton: "The Catalysis of Diene Carbonylation," *Ann. N.Y. Acad. Sci.* **333** (1980) 264 – 272.
443. P. Pino: "Hydroformylation of Olefinic Hydrocarbons with Rh and Co Catalysts," *J. Organomet. Chem.* **200** (1980) 223 – 242.
444. M. Orchin: "$HCo(CO)_4$, the Quintessential Catalyst," *Acc. Chem. Res.* **14** (1981) 259 – 266.
445. D. Forster, A. Hershman, D. E. Morris, *Catal. Rev. Sci. Eng.* **23** (1981) 89 – 105.
446. W. Cornely, B. Fell, *J. Mol. Catal.* **16** (1982) 89.
447. R. P. A. Sneeden: "Organic Syntheses where Carbon Monoxide is the Unique Source of Carbon," in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* vol. 8, Pergamon Press, Oxford 1982, p. 19.
448. I. Tkatchenko: "Synthesis with Carbon Monoxide and a Petroleum Product," in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* vol. 8, Pergamon Press, Oxford 1982, p. 101.
449. W. Keim (ed.): *Catalysis in C_1* Chemistry, Reidel Publ., Dordrecht 1983.
450. R. A. Sheldon: *Chemicals from Synthesis gas,* Reidel Publ., Dordrecht 1983.
451. F. R. Hartley, S. Patai: *Chemistry of Metal-Carbon Bond,* vol. 3, Wiley-Interscience, Chichester 1985.
452. M. Röper: "Zur direkten und indirekten Synthese sauerstoffhaltiger Verbindungen aus CO/H2 ", *Habilitationsschrift,* RWTH, Aachen 1985.
453. R. P. Chiusoli: "Catalytic C – C Bond Formation by Carbonylation of Organic Substrates," *Transition Met. Chem. (Weinheim, Ger.)* **12** (1987) 89 – 96.
454. W. Keim: "Synthesis Gas – Feedstock for Chemicals," *ACS Symp. Series* **328** (1987) 1 – 16.
455. B. Fell, Z. Jin, *Chem. Z.* **111** (1987) 317 – 323.
456. M. Röper: "Chemicals from Methanol and Carbon Monoxide," in A. Mortreux, F. Petit (eds.): *Industrial Applications of Homogeneous Catalysis,* Reidel Publ., Dordrecht 1988.
457. J. Gauthier-Lafaye, R. Perron: "Carbon Monoxide and Fine Chemicals Synthesis," in A. Mortreux, F. Petit (eds.): *Industrial Applications of Homogeneous Catalysis,* Reidel Publ., Dordrecht 1988.
458. M. Kilner, N. J. Winter, *J. Motol. Catal. A.: Chemical* **121** (1996) 327 – 345.
459. O. Reiser: "Metall-katalysierte Hydroformylierungen", *Nachr. Chem. Tech. Lab.* **44** (1996) 996 – 1001.
460. B. Lee, H. Alper, *J. Mol. Catal. A.: Chemical* **111** (1996) 17 – 23.
461. A. M. Trzeciak, J. J. Ziolkowski, R. Choukroun, *J. Mol. Catal. A.: Chemical* **110** (1996) 135 – 139.
462. F. Agbossou, J.-F. Carpentier, A. Mortreux: "Asymmetric Hydroformylation", *Chem. Rev.* **95** (1995) 2485 – 2506.
463. C. Abu-Gnim, I. Amer, *J. Organomet. Chem.* **516** (1996) 235 – 243.
464. J. Freiberg, A. Weigt, H. Dilcher, *J. Prakt. Chem.* **335** (1993) 337 – 344.
465. J. Beisser, B. Fell, *J. Prakt. Chem.* **334** (1992) 355 – 357.
466. A. M. Trzeciak, J. J. Ziolkowsky, R. Choukroun, *J. Organomet. Chem.* **525** (1996) 145 – 149.
467. S. Oi, N. Nomura, T. Aiko, Y. Inoue, *J. Mol. Catal. A.: Chemical* **115** (1997) 289 – 295.
468. T. Miyawaki, K. Nomura, M. Hazama, G. Suzukamo, *J. Mol. Catal. A.: Chemical* **120** (1997) L9 – L11.
469. G. P. Pez, S. T. Butany: "Functionalization of Olefins," in M. Tsutsui, R. Ugo (eds.): *Fundam. Research in Homogeneous Catalysis,* Plenum Press, New York 1977, pp. 53 – 67.
470. I. Ojima, K. Yamamoto, M. Kumada: "Asym. Hydrosilylation," in R. Ugo (ed.): *Asp. Homog. Catal.,* vol. 3, Reidel Publ., Dordrecht 1977, p. 186.
471. J. F. Harrod, A. J. Chalk: "Hydrosilylation," in I. Wender, P. Pino (eds.): *Org. Synth. Met. Carbonyls* **2** (1977) 673.
472. E. S. Brown in I. Wender, P. Pino (eds.): *Organic Synthesis by Metal Carbonyls,* **vol. 2,** Wiley-Interscience, New York 1977, pp. 655 – 672.
473. F. Sato: "Hydroalumination," in Y. Ishii, M. Tsutsui (eds.): *Fundam. Research in Homogeneous Catalysis,* vol. 2, Plenum Press, New York 1978, pp. 81 – 92.
474. J. L. Speier: "Hydrosilation," *Adv. Organomet. Chem.* **17** (1979) 407 – 447.
475. S. E. Diamond, A. Szalkiewwicz, F. Mares: "Hydroamination," *J. Am. Chem. Soc.* **101** (1979) 490.

476. W. Keim, A. Behr, H.-O. Lühr, J. Weisser: "Catalytic Hydrocyanation of Dienes and Trienes," *J. Catal.* **78** (1982) 209 – 216.
477. B. R. James in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, pp. 285 – 369.
478. M. B. Gasc, A. Latters, J. J. Perie: "Hydroamination," *Tetrahedron* **39** (1983) 703.
479. A. Behr: "Activation of Carbon Dioxide via Coordination to Transition Metal Complexes," in W. Keim (ed.): *Catalysis in C1-Chemistry,* Reidel Publ., Dordrecht 1983, p. 169.
480. W. C. Seidel, C. A. Tolman, *Ann. N.Y. Acad. Sci.* **415** (1983) 201 – 221.
481. H. S. Lee, K. Isagawa, Y. Otsuji: "Hydroboration," *Chem. Lett.* 1984, 363.
482. H. S. Lee, K. Isagawa, H. Toyoda, Y. Otsuji: "Hydroboration," *Chem. Lett.* 1984, 673.
483. F. Sato: "Hydroalumination," *Fundament. Chem. Res.* **4** (1984) 81.
484. C. A. Tolman et al.: "Homogeneous Nickel-Catalyzed Olefin Hydrocyanation," *Adv. Catal.* **33** (1985) 1 – 46.
485. I. Mamalis et al., *J. Catal.* **102** (1986) 357 – 364.
486. *Houben-Weyl,* **E 18,** 668, 802.
487. M. Hodgson, D. Parker, *J. Organomet. Chem.* **325** (1987) C27 – C30.
488. A. Behr: "Kohlendioxid als alternativer C_1-Baustein: Aktivierung durch Übergangsmetallkomplexe," *Angew. Chem.* **100** (1988) 681; *Angew. Chem. Int. Ed. Engl.* **27** (1988) 661 – 678.
489. A. Behr: *Carbon Dioxide Activation by Metal Complexes,* VCH Verlagsgesellschaft, Weinheim 1988.
490. D. Y. Waddan et al., *J. Catal.* **116** (1989) 415 – 421.
491. J.-J. Brunet, D. Neibecker, F. Niedercorn: "Catalytic Amination of Monoolefins," *J. Mol. Catal.* **49** (1989) 235 – 259.
492. M. Beller: "Die palladiumkatalysierte Aminierung von Arylhalogeniden", *Angew. Chem.* **107** (1995) 1436 – 1437.
493. *Houben-Weyl,* **vol. E21a – E21f.**
494. A. H. Hoveyda, D. A. Evans, G. C. Fu, *Chem. Rev.* **93** (1993) 1307 – 1370.
495. K. Burgess, M. J. Ohlmeyer: "Transition-Metal-Promoted Hydroborations of Alkenes", *Chem. Rev.* **91** (1991) 1179 – 1191.
496. L. N. Lewis J. Stein R. E. Colborn, Y. Gao, J. Dong, *J. Organomet. Chem.* **521** (1996) 221 – 227.
497. M. Rivera-Claudio, J. Rozell, E. Ramirez-Oliva, J. Cervantes, K. H. Pannell, *J. Organomet. Chem.* **521** (1996) 267 – 270.
498. M. R. Kesti, R. M. Waymouth, *Organometallics* **11** (1992) 1095 – 1103.
499. I. Ojima, R. J. Donovan, N. Clos, *Organometallics* **10** (1991) 2606 – 2610.
500. W. Wolfsberger: "Hydrogermylierungsreaktionen", *J. Prakt. Chem.* **334** (1992) 453 – 464.
501. J. Y. Corey. X. -H. Zhu, *Organometallics* **11** (1992) 672 – 683.
502. M. Bruncko, G. Schlingloff, K. B. Sharpless, *Angew. Chem.* **109** (1997) 1580 – 1583.
503. S. Sakaki, M. Ogawa, Y. Musashi, *J. Organomet. Chem.* **535** (1997) 25 – 28.
504. S. Ützkan, V. M. Akhmedov, C. Kayrau, *J. Organomet. Chem.* **533** (1997) 103 – 108.
505. A. Tillak et al., *J. Organomet. Chem.* **532** (1997) 117 – 123.
506. D. E. Welster: "Activation of Alkanes by Transition Metal Compounds," *Adv. Organomet. Chem.* **15** (1977) 147.
507. A. H. Janowicz, R. G. Bergman, *J. Am. Chem. Soc.* **104** (1982) 352.
508. J. K. Hoyano, W. A. G. Graham, *J. Am. Chem. Soc.* **104** (1982) 3723.
509. G. W. Parshall, D. L. Thorn, T. H. Tulip: "Activate C – H bonds," *CHEMTECH* 1982, Sept., 571 – 576.
510. R. G. Bergman, *Science* **223** (1984) 902.
511. A. E. Shilov: *Activation of Saturated Hydrocarbons by Transition Metal Complexes,* Reidel Publ., Dordrecht 1984.
512. R. H. Crabtree, *Chem. Rev.* **85** (1985) 245.
513. J. Halpern, *Inorg. Chim. Acta* **100** (1985) 41.
514. M. Hackett, J. A. Ibers, P. Jernakoff, G. M. Whitesides, *J. Am. Chem. Soc.* **108** (1986) 8094 – 8095.
515. J. M. Cogen, W. F. Maier, *J. Am. Chem. Soc.* **108** (1986) 7752 – 7762.
516. J. W. Park, P. B. Mackenzie, W. P. Schaefer, R. H. Grubbs, *J. Am. Chem. Soc.* **108** (1986) 6402 – 6404.
517. R. A. Periana, R. G. Bergman, *J. Am. Chem. Soc.* **108** (1986) 7332 – 7346.
518. M. Ephritikhine, *Nouv. J. Chim.* **10** (1986) 9.
519. B. Meunier, B. Chaudret (eds.): *Perspectives in the Selective Activation of C – H and C – C Bonds in Saturated Hydrocarbons,* Final Report of the International Workshop in Bonas, France, 30 June – 3 July 1987.
520. M. V. Baker, L. D. Field, *J. Am. Chem. Soc.* **109** (1987) 2825 – 2856.

521. M. Ephritikhine in A. Mortreux, F. Petit (eds.): *Industrial Applications of Homogeneous Catalysis,* Reidel Publ., Dordrecht 1988, p. 257.
522. W. D. Jones, F. J. Feher, *Acc. Chem. Res.* **22** (1989) 91 – 100.
523. R. H. Crabtree: "Aspects of Methane Chemistry", *Chem. Rev.* **95** (1995) 987 – 1007.
524. K. A. Gardner, J. M. Mayer, *Science* **269** (1995) 1849 – 1851.
525. D. Schröder, H. Schwarz, *Angew. Chem.* **107** (1995) 2126 – 2150.
526. A. L. Tonkovich, R. W. Carr, R. Aris, *Science* **262** (1993) 221 – 223.
527. J. A. Davies, P. L. Watson, A. Greenberg, J. F. Liebman (eds.): *Selective Hydrocarbon Activation – Principles and Progress,* Wiley-VCH, Weinheim 1990.
528. D. A. Hickman, L. D. Schmidt, *Science* **259** (1993) 343 – 346.
529. J. J. Schneider, *Angew. Chem.* **108** (1996) 1132 – 1139.
530. G. J. Stiegel, R. D. Srivastava, *Chem. & Ind. (London),* 1994 7. Nov., 854 – 856.
531. D. H. R. Barton, D. Doller, *Acc. Chem. Res.* **25** (1992) 504 – 512.

Orthoformates → Esters
Orthophosphoric Acid → Phosphoric Acid and Phosphates
Osmium → Platinum Group Metals and Compounds
Osmosis → Membranes and Membrane Separation Processes

Oxalic Acid

WILHELM RIEMENSCHNEIDER, Hoechst AG, Frankfurt, Federal Republic of Germany (Chaps. 1 – 3, 5 – 6, Section 4.1, 4.2, 4.3, 4.5, Chap. 7 in part)

MINORU TANIFUJI, Ube Industries, Tokyo, Japan (Section 4.4, Chap. 7 in part)

1. Introduction 1
2. Physical Properties 1
2.1. Anhydrous Oxalic Acid 2
2.2. Oxalic Acid Dihydrate 2
3. Chemical Properties 2
4. Production Processes and Raw Materials 3
4.1. Oxidation of Carbohydrates 4
4.2. Oxidation of Ethylene Glycol 5
4.3. Oxidation of Propene 6
4.4. Production from Carbon Monoxide . 7
4.5. Production of Anhydrous Oxalic Acid 8
5. Chemical Analysis 8
6. Uses 9
6.1. Metal Treatment 9
6.2. Textile Treatment 9
6.3. Bleaching Agents 9
6.4. Chemical Uses 9
7. Economic Aspects 10
8. Storage, Handling, Transportation, Waste Disposal 10
9. Derivatives 11
9.1. Salts 11
9.2. Organic Derivatives 11
10. Toxicology 12
11. References 13

1. Introduction

Oxalic acid, ethanedioic acid, acidum oxalicum, has the structural formula

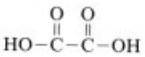

and is the simplest saturated dicarboxylic acid [1]. The compound exists in anhydrous form [*144-62-7*] or as a dihydrate [*6153-56-6*]. The anhydrous acid is not found in nature and must be prepared from the dihydrate even when produced industrially. Oxalic acid is widely distributed in the plant and animal kingdom (nearly always in the form of its salts) and has various industrial applications.

The acidic potassium salt of oxalic acid is found in common sorrel (Latin: oxalis acetosella) and the name oxalic acid is derived from that plant. Examples of plants in which oxalic acid occurs (in the form of potassium, sodium, calcium, magnesium salts, or iron complex salts) are given below (oxalic acid content in milligrams per 100 g dry weight):

Spinach	460 – 3200
Rhubarb	500 – 2400
Chard	690
Parsley	190
Beets	340
Cocoa	4500
Tea	3700
Beet leaves	up to 12 000

Oxalic acid is formed in plants through incomplete oxidation of carbohydrates, e.g., by fungi (*Aspergillus niger*) or bacteria (*acetobacter*) and in the animal kingdom through carbohydrate metabolism via the tricarboxylic acid cycle. The urine of humans and of most mammals also contains a small amount of calcium oxalate. In pathological cases, an increased calcium oxalate content in urine leads to the formation of kidney stones [2]. Calcium and iron(II) oxalates are also found as minerals.

Oxalic acid is one of the oldest known acids and was already identified in the potassium salt of sorrel by WIEGLEB in 1769. The compound also has historical significance in chemistry because WOEHLER prepared oxalic acid by hydrolysis of cyanogen [*460-19-5*] in 1824, thereby synthesizing the first natural product, even before his famous urea synthesis in 1828.

2. Physical Properties

Both the anhydrous and dihydrated forms of oxalic acid form colorless and odorless crystals.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_247

2.1. Anhydrous Oxalic Acid

Anhydrous oxalic acid [*144-62-7*] exists as rhombic crystals in the α-form and as monoclinic crystals in the β-form [3]. These forms differ mainly in their melting points. The slightly stable β-form changes into the α-form at 97 °C and 0.2 bar. Anhydrous oxalic acid is prepared by dehydration of the dihydrate through careful heating to 100 °C. It is then sublimated in a dry air stream. The sublimation is fast at 125 °C and can be carried out at temperatures up to 157 °C without decomposition. The dehydration can also be accomplished by azeotropic distillation with benzene or toluene [4]. Anhydrous oxalic acid is slightly hygroscopic, it absorbs water from moist air ("weathers") to form the dihydrate again. The hydration occurs very slowly because of surface caking. The physical and thermal chemical properties of anhydrous oxalic acid may be summarized as follows:

M_r	90.04
mp	
α	189.5 °C
β	182 °C
Decomposition temperature	187 °C
Density d_4^{20}	
α	1.900 g/cm^3
β	1.895 g/cm^3
Refractive index β, n_4^{20}	1.540
Vapor pressure (solid, 57 – 107 °C)	$\log P = -(4726.95/T) + 11.3478$ kPa
Specific heat (solid, −200 to 50 °C)	$c_p = 1.084 + 0.0318\,t$ J/g
Heat of combustion ΔH_c° (25 °C)	−245.61 kJ/mol
Standard enthalpy of formation $\Delta H^\circ{}_f$ (25 °C)	−826.78 kJ/mol
Standard free energy of formation $\Delta G^\circ{}_f$ (25 °C)	−697.91 kJ/mol
Heat of solution (in water)	−9.58 kJ/mol
Heat of sublimation	90.58 kJ/mol
Heat of decomposition	826.78 kJ/mol
Entropy, S° (25 °C)	120.08 J mol^{-1} K^{-1}
Thermal conductivity (0 °C)	0.9 W m^{-1} K^{-1}
Ionization constant	
K_1	6.5×10^{-2}
K_2	6.0×10^{-5}

2.2. Oxalic Acid Dihydrate

Oxalic acid dihydrate [*6153-56-6*], HOOC−COOH · 2 H_2O is the industrially produced and usual commercial form of oxalic acid. The compound forms colorless and odorless prisms or granules that contain 71.42 wt % oxalic acid and 28.58 wt % water. Oxalic acid dihydrate is stable at room temperature and under normal storage conditions. The most important physical properties are as follows:

M_r	126.07
mp	101.5 °C
Density d_4^{20}	1.653 g/cm^3
Refractive index n_4^{20}	1.475
Standard enthalpy of formation ΔH_f° (18 °C)	−1422 kJ/mol
Enthalpy of solution (in water)	−35.5 kJ/mol
pH (0.1 M solution)	1.3
Water vapor partial pressure	
at 20 °C	2.5 kPa
at 50 °C	20 kPa
Bulk density, regular product	0.977 g/cm^3
Bulk density, coarse product	0.881 g/cm^3

The solubility in water and the density of these solutions are presented in Table 1 [5]. Oxalic acid is readily soluble in polar solvents such as alcohols (although partial esterification occurs), acetone, dioxane, tetrahydrofuran, and furfural. Oxalic acid is sparingly soluble in diethyl ether (1.5 g oxalic acid dihydrate in 100 g ether at 25 °C), and insoluble in benzene, chloroform, and petroleum ether. The ionization constants show that oxalic acid is a strong acid. The value of K_1 is comparable to that of mineral acids and the value of K_2 corresponds to ionization constants of strong organic acids, for example, benzoic acid.

Table 1. Solubility of oxalic acid in water at different temperatures and corresponding densities

Temperature, °C	Solubility, g/100 g ∗	Relative density $d_4^{17.5}$
0	3.5	1.016
10	5.5	1.025
17.5	8.5	1.038
20	9.5	
30	14.5	
40	22	
50	32	
60	46	
80	85	
90	120	

∗ Anhydrous basis.

3. Chemical Properties

In the homologous series of dicarboxylic acids, oxalic acid, the first member, shows unique be-

havior because of the interaction of the neighboring carboxylate groups. This results in an increase in the value of the dissociation constant and in the ease of decarboxylation: Upon rapid heating to 100 °C oxalic acid decomposes into carbon monoxide, carbon dioxide, and water with formic acid as an isolable intermediate. In aqueous solution decomposition is induced by light, and to a much greater extent by γ- or X-rays (to carbon monoxide, carbon dioxide, formic acid, and occasionally hydrogen). This decomposition is catalyzed by the salts of heavy metals, for example, by uranyl salts. Oxalic acid cannot form an intramolecular anhydride. Upon heating to over 190 °C or warming in concentrated sulfuric or phosphoric acid, oxalic acid decomposes to carbon monoxide, carbon dioxide, and water: this decomposition is not exothermic.

Ease of reaction with oxygen is also a typical characteristic of oxalic acid:

$$HOOC{-}COOH + 1/2\ O_2 \longrightarrow 2\ CO_2 + H_2O$$

This reaction can be exploited in quantitative analysis (Manganometry, see Chap. 5).

The reducing properties of oxalic acid (which itself is oxidized to the harmless end products carbon dioxide and water) form the basis for the variety of practical applications (see Chap. 6). Oxalic acid is also oxidized relatively easily to carbon dioxide by many other oxidizing agents in addition to air, especially in the presence of the salts of heavy metals. Oxalic acid is easily esterified, whereby two types, the acidic mono or neutral diesters can result. These esters are applied as intermediates in chemical syntheses. They react relatively easily with water, ammonia, or amines to afford the corresponding acyl derivatives.

Important chemical characteristics are also demonstrated by the metal salts of oxalic acid. These exist in two types–the acidic and neutral salts. The alkali metal and iron (III) salts are readily soluble in water. All other salts are sparingly soluble in water. The near complete insolubility of the alkaline-earth salts of oxalic acid, especially of calcium oxalate, finds some applications in quantitative analysis. When heated all these metal salts lose carbon monoxide. Other salts which are easier decomposable lose carbon dioxide in addition. The alkali and alkaline-earth salts form carbonates under these conditions. Manganese, zinc, and tin salts form oxides; iron, cadmium, mercury, and copper salts form mixtures of oxides and metals. Nickel, cobalt, and silver salts afford pure metals. Anhydrous fusion of oxalates with alkali yield carbonates and hydrogen. For review see [6].

4. Production Processes and Raw Materials

There are many industrial processes for the production of oxalic acid, which have been carried out by a large number of companies and are in part still used today. In recent years both the number of firms producing oxalic acid and the number of starting materials used for its production have been reduced. Generally, only three classes of compounds are employed as raw materials for the production of oxalic acid. These are:

carbohydrates (including ethylene glycol and molasses)
olefins
carbon monoxide

Nevertheless, three outdated methods, which have been abandoned in important industrial nations for economic reasons, are described briefly [7], because in the third world small firms might exist that still employ these production methods:

1) The sodium formate method, which is also based on carbon monoxide as the raw material, but involves very cumbersome procedures of questionable technical safety that are no longer economically feasible (heating to 400 °C, elimination of hydrogen, calcium precipitation, and precipitation with sulfuric acid). Nevertheless, this process is probably still being used in China.
2) The alkali fusion of cellulose, which is the oldest production method [8]. Although cellulose-containing material from every source can be used in this method, this process has been discontinued because of the low yields, large amounts of required alkali and sulfuric acid, and the large amounts of wastewater.
3) The isolation of oxalic acid as a byproduct of carbohydrate fermentation, for example, in the synthesis of citric acid [9]. The process involves the biotechnical exploitation

of the tricarboxylic acid cycle through enzymatic cleavage of sugars. The intermediary formed oxalacetic acid esters and oxalosuccinic acid (1-oxopropane-1,2,3-tricarboxylic acid) esters are cleaved by hydrolases. Especially because of the handling of large amounts of water, this process is no longer economically feasible.

In all industrial processes, oxalic acid is produced as the crystalline dihydrate. In the following sections four methods are described that are currently used industrially. Three of them are based on a nitric acid oxidation and one is a synthesis using carbon monoxide. A new review of oxalic acid production from various sources is given in [10].

4.1. Oxidation of Carbohydrates

This method, the oldest of chemical productions, was in principal discovered by SCHEELE more than 200 years ago (1776). He oxidized sugar with concentrated nitric acid to oxalic acid. However, it was not until ca. 1940 that this process gained industrial importance when the nitrogen oxides produced in the reaction could be recovered and recycled. In Germany, I.G. Farben produced 2000 t of oxalic acid per year by this method until 1944 [11]. This process is employed in Brazil, China, and in several Eastern European countries. Raw materials employed in this process include sugar, glucose, fructose, corn starch, wheat starch, reclaimed starch, potato starch, corncobs, tapioca, and molasses. Choice of starting material depends on availability and price and on optimization of the reaction process involved. For example, depending on the starting materials foam, greases, slime, and varying yields can be expected and must be dealt with accordingly [1]. Molasses and other agricultural waste have gained increasing importance as raw materials. These materials are not only inexpensive, but must be further processed in any case due to environmental reasons. Molasses contains, as a production residue of sugar refinement, many nitrogen-containing compounds, which are, however, for the most part removed during the nitric acid oxidation. This explains the difficulties encountered with excessive foam formation during production that must be overcome. If starch products are used as raw materials they must be hydrolyzed to glucose in an extra batchwise step prior to the actual oxidation [12, 13]. This is carried out by boiling the starch suspension in the presence of oxalic acid or sulfuric acid for about six hours. The oxidation of the resulting glucose and the following steps are then essentially continuous operations. The oxidation occurs according to the following idealized equation:

$$C_6H_{12}O_6 + 6\,HNO_3 \xrightarrow{V_2O_5,\ Fe^{3+}} 3\,(COOH)_2 + 6\,NO + 6\,H_2O$$

In fact besides NO other nitrogen oxides such as NO_2, N_2O, N_2O_5 are formed.

Allied Chemical Process. The Allied Chemical process is schematically presented in Figure 1. The concentration of the glucose solution obtained by hydrolysis of starch is approximately 50 – 60 wt %. This material, along with the recycled supernatent from the crude centrifugation, is placed in a reactor with ca. 50 % sulfuric acid, vanadium pentoxide (0.001 – 0.05 wt %) and iron (III) sulfate (0.39 – 0.8 wt %), based on the total mass of the mixture. Nitric acid (65 %) is then added slowly under vigorous stirring in a tightly controlled temperature range around 70 °C– the temperature range differs according to the raw material being used. The strongly exothermic reaction must be well cooled. Air is simultaneously blown into the reactor to support the reaction and to remove NO and other nitrogen oxides of low oxidation number. The nitrogen oxides expelled are collected in an adsorption system and are recycled after oxidation. Crude oxalic acid is obtained after cooling and centrifugation of the reaction mixture. The crude acid is again dissolved in hot water, put through a grease separator and recrystallized. After a second centrifugation and drying, oxalic acid dihydrate is obtained with a purity >99 %. Precise control of the reaction parameters is critical for a good yield [14]. For good quality raw materials ca. 65 % of the theoretical yield is obtained. The absent ca. 35 % of the yield mainly results from losses through CO_2 formation. This process is employed today in Spain, Brazil, The People's Republic of China, Taiwan, Korea, and India.

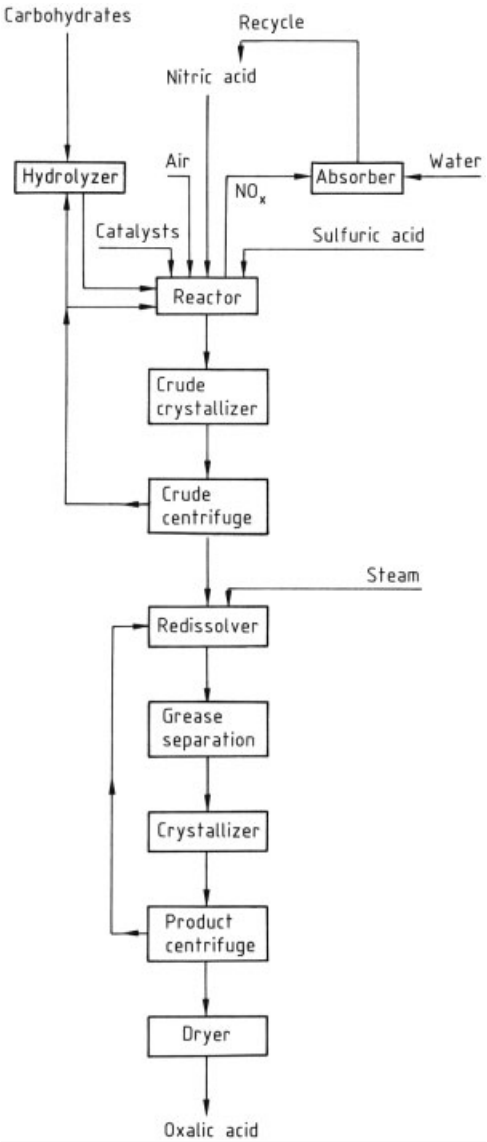

Figure 1. Production of oxalic acid by oxidation of carbohydrates (Allied Chemical process)

4.2. Oxidation of Ethylene Glycol

The oxidation of ethylene glycol with nitric acid is a one-step process. A flow chart is shown in Figure 2. An oxidizing mixture of 30 – 40 % sulfuric acid and 20 – 25 % nitric acid is used. The oxidation is carried out in the presence of vanadium pentoxide and iron (III) salts, at 50 – 70 °C and atmospheric pressure. Only CO_2 is formed as a byproduct. This original process has been improved by the Mitsubishi Gas Company and in Japan most of the oxalic acid is produced this way [15]. The improved process operates at about 10 bar in a pressure reactor, in which ca. 60 % nitric acid and possibly also sulfuric acid at 80 °C, is pressurized with oxygen at 3 – 10 bar. Ethylene glycol is then continually pumped in and the temperature and pressure are maintained by addition of oxygen. The exothermic reaction must be held at a temperature of 50 – 70 °C by effective cooling. The yield of oxalic acid, based on ethylene glycol, is 90 – 94 % of the theoretical value. The nitric acid used can be recovered almost completely. This indicates that the nitrogen oxides and nitric acid only act as catalysts and the actual oxidation has been effected by oxygen. The idealized overall equation for this reaction is:

$$\begin{array}{c}CH_2OH\\|\\CH_2OH\end{array} + 2O_2 \xrightarrow[V_2O_5,\ Fe^{3+}]{NO_2} \begin{array}{c}COOH\\|\\COOH\end{array} + 2H_2O$$

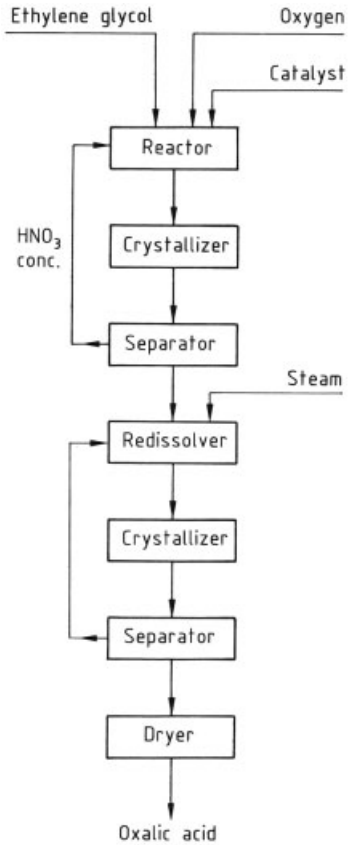

Figure 2. Production of oxalic acid from ethylene glycol

Other suitable starting materials are propylene glycol, acetaldehyde, or glycolic acid. Vanadium compounds (0.001 – 0.1 wt %) are recommended as promoters. Sodium nitrite, formic acid, or formaldehyde are recommended as initiators [16]. Ube Industries in Japan have further developed the process so that it operates only with nitric acid (no sulfuric acid), however, pressure is required. In an oxalic acid process developed in the former Soviet Union a mixture of nitric and sulfuric acids is used. The reaction is performed at 1.3 – 4.0 kPa (the pressure is less than that employed in the Mitsubishi process) to remove NO_x more effectively.

Although the ethylene glycol process clearly affords the best yields by a wide margin, the corn starch and molasses oxidations are still more advantageous from an economic point of view. Ethylene glycol oxidation is carried out only by Mitsubishi in Japan.

4.3. Oxidation of Propene

The second most important industrial process for the production of oxalic acid (after carbohydrate oxidation) is the reaction of propene with nitric acid. The worldwide production of oxalic acid by this method in 1978 was 65 000 t/a. Pure propene or propene rich fractions from petrochemical refineries can be used as feedstocks. This earlier developed reaction was significantly improved by Rhône-Poulenc in France by technical engineering and construction improvements and strict observance of the reaction conditions. These refinements increased the yield and solved the serious safety problems [17]. Rhône-Poulenc is reputed to be the largest oxalic acid producer in the world. The two-step process can be represented by the following equations approximate stoichiometry:

Initial stage:

$$\underset{\text{Propene}}{CH_3CH{=}CH_2} + 3\,HNO_3 \longrightarrow \underset{\alpha\text{-Nitratolactic acid}}{CH_3\underset{\underset{ONO_2}{|}}{C}HCOOH} + 2\,NO + 2\,H_2O$$

Final stage:

$$CH_3\underset{\underset{ONO_2}{|}}{C}HCOOH + 5/2\,O_2 \longrightarrow \underset{\text{Oxalic acid}}{HOOC-COOH} + CO_2 + HNO_3 + H_2O$$

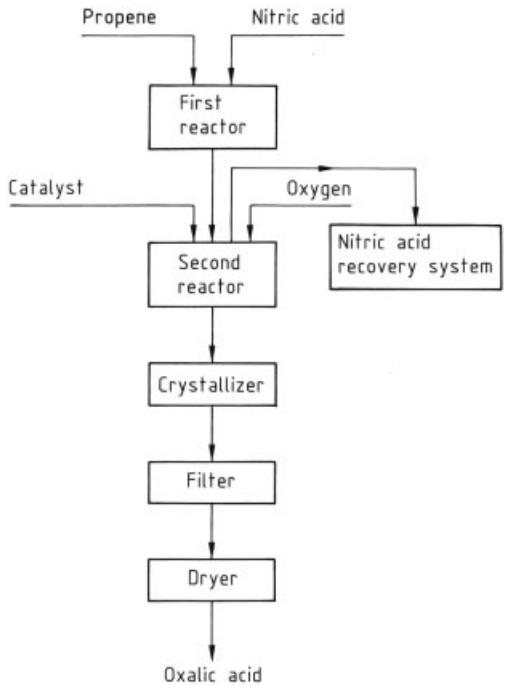

Figure 3. Production of oxalic acid from propene

Process Description (see Fig. 3). Propene is introduced into a solution of nitric acid (50 – 75 wt %) in the first reactor at 10 – 40 °C, while keeping the molar ratio of propene to nitric acid at 0.01/1.0 – 0.5/1.0 to produce watersoluble intermediates of α-nitratolactic acid and lactic acid. In the second step the solution of these partially oxidized products is treated with oxygen at 45 – 100 °C in the presence of a catalyst. Oxalic acid is formed, crystallized, and filtered. The overall yield based on propene is claimed to be ca. 90 % and the conversion of propene is ca. 100 %. The nitrogen oxides are removed from the second reactor and recovered. The consumption of nitric acid is 1.2 kg per kg of oxalic acid dihydrate produced. The process can also be employed for the production of lactic acid instead of oxalic acid. An improvement in the yield is achieved by addition of catalysts (salts or compounds containing Fe, Al, Cr, Sn, Bi, or I). In the chemical industry chromium (III) nitrate, iron (III) nitrate or tin (II) chloride are employed in quantities of 0.4 – 0.5 wt % (based on the metal) and dissolved in 65 % nitric acid. The scope and limitations of this process are restricted by safety

problems because the nitrates of α-hydroxycarboxylic acids that are formed as intermediates from α-olefins and NO_2 are unstable and lead to uncontrollable decomposition and violent explosions [18]. As a result many technical difficulties need to be overcome to carry out these processes. Today this process is only operated by Rhône-Poulenc. Analogous to propene, ethylene can also be oxidized to oxalic acid.

4.4. Production from Carbon Monoxide [19,20]

The newest process for the industrial production of oxalic acid was developed by Ube Industries. Their synthesis is performed in two steps. In the first step CO and a lower alcohol are reacted under pressure and in the presence of a catalyst to form the corresponding diester of oxalic acid. This diester is hydrolyzed in the second step to oxalic acid and the original alcohol which is then recycled. Palladium on charcoal and alkyl nitrites are employed as catalysts and the reaction is carried out at 10 – 11 MPa. *n*-Butanol is the preferred alcohol and is simultaneously used as the solvent. This process has the advantage of inexpensive starting materials, however, it is disadvantageous with regard to the much higher capital investment (high pressure) and utility costs.

$$2\,CO + 2\,ROH + 1/2\,O_2 \xrightarrow{\text{Pd/C, RONO}} \begin{matrix}COOR\\ |\\ COOR\end{matrix} + H_2O \quad (1)$$

$\Delta H = -52.8$ kJ/mol

Oxalic acid

$$\begin{matrix}COOR\\ |\\ COOR\end{matrix} + 2\,H_2O \longrightarrow \begin{matrix}COOH\\ |\\ COOH\end{matrix} + 2\,ROH \quad (2)$$

New perspectives and developments are given in [21]. The use of palladium catalysts is discussed in [22].

The catalytic action of the alkyl nitrite on the first reaction is believed to be the following:

$$2\,RONO + 2\,CO \longrightarrow (COOR)_2 + 2\,NO \quad (3)$$

$$2\,NO + 1/2\,O_2 + 2\,ROH \longrightarrow RONO + H_2O \quad (4)$$

Butyl nitrite and methyl nitrite are the preferred alkyl nitrites for industrial purposes. The first plant using butyl nitrite in the production of oxalic acid was constructed in 1978 (6000 t/a oxalic acid). The advantages of using butyl nitrite are as follows:

1) It functions not only as reaction component but also as a dehydrating agent. Butyl nitrite has a higher boiling point than water and butyl nitrite and water form an azeotropic mixture with a lower boiling point than butyl nitrite. Thus, the water formed can be removed as an azeotropic mixture and most of the butyl nitrite remains in the system. Butyl nitrite is recovered from the azeotropic mixture and reused.
2) Butyl nitrite is finally hydrolyzed to butanol which can easily be separated from oxalic acid solution by phase separation.

A flow chart of the process using butyl nitrite is given in Figure 4. A circulation liquid containing butyl nitrite, butanol, and the palladium catalyst and a circulation gas containing carbon monoxide and oxygen are pressurized and fed to a reaction tower. Here dibutyl oxalate is formed and butyl nitrite is consumed (Eq. 3). Butyl nitrite is regenerated simultaneously (Eq. 4). Both reactions are exothermic. The selectivity for dibutyl oxalate increases with increasing reaction pressure and decreasing reaction temperature; however, the velocity of the reaction increases with increasing temperature. Thus, the reaction is conducted industrially at 10 – 11 MPa and 90 – 100 °C.

Gas circulation is carried out in order to hold the oxygen concentration below the explosion limit during the reaction and to improve the carbon monoxide utilization and gas–liquid contact rate.

The solution released from the tower is flashed to evaporate the carbon dioxide gas formed as a byproduct. The reaction water is removed by azeotropic distillation with butyl nitrite and most of the solution is recycled to the reaction tower. Part of the solution is further processed: it is first filtered to separate the catalysts; then butyl nitrite, and butanol are removed to yield pure dibutyl oxalate. The palladium catalyst, butyl nitrite, and butanol are recovered and returned to the circulation liquid, which is also supplied with make-up butanol and nitric acid. The circulation liquid is then pressurized and fed back to the reaction tower.

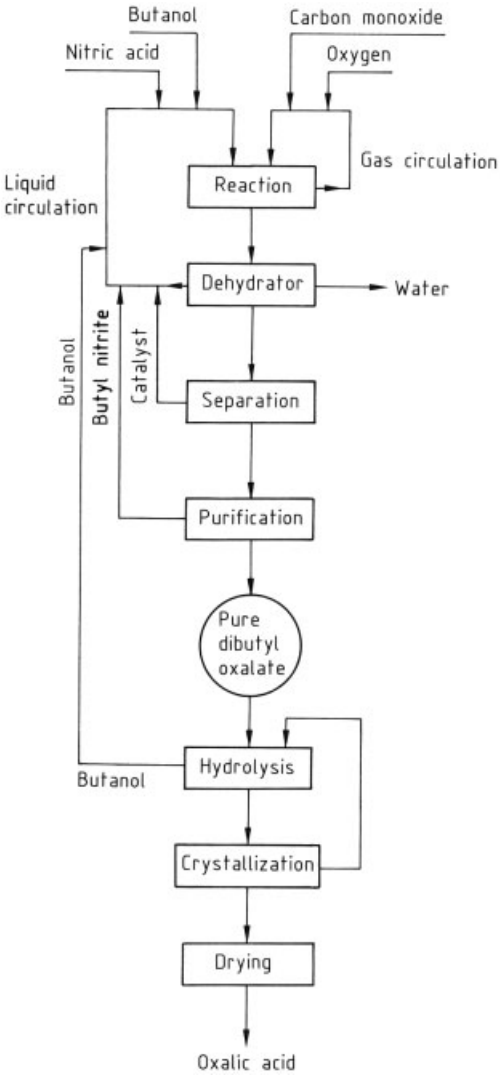

Figure 4. Production of oxalic acid from carbon monoxide an butanol using butyl nitrite as catalyst

The refined dibutyl oxalate is mixed with the mother liquor from the oxalic acid crystallization and hydrolyzed at about 80 °C into oxalic acid and butanol. The oxalic acid solution and butanol are separated by phase separation. Then oxalic acid dihydrate is crystallized from the oxalic acid solution, and butanol is recycled to the process after refining. A gas-phase process using methyl nitrite has been established [21, 22]. This process employs a lower reaction pressure (390 kPa) and is more economic than the liquid-phase method because of the following reasons:

1) The consumption of electric power is greatly reduced owing to the low pressure operation (about 60 % saving).
2) Dimethyl oxalate formation and methyl nitrite regeneration reaction are separated; hence, side-reactions are decreased and the yield of dimethyl oxalate is increased (98 %).
3) Inexpensive methanol is used instead of butanol.

4.5. Production of Anhydrous Oxalic Acid

In all processes oxalic acid is obtained as the dihydrate. Anhydrous oxalic acid can be obtained by heating oxalic acid dihydrate to ca. 3 °C below its melting point or by azeotropic distillation with benzene or toluene. The demand for anhydrous oxalic acid is, however, very low.

5. Chemical Analysis

Qualitative Analysis. Upon heating oxalic acid with diphenylamine a blue color results, the so-called "aniline blue". Analysis limit: 5 µg [23].

Quantitative Analysis. In addition to simple acid–base titration, which of course can only be conducted in the absence of other acids, four important analytical methods exist.

Manganometry. As the reaction of oxalic acid and $KMnO_4$ is quantitative, oxalic acid can be easily titrated with a standard solution of $KMnO_4$.

$$5\,(COOH)_2 + 2\,KMnO_4 + 3\,H_2SO_4 \longrightarrow K_2SO_4 + 2\,MnSO_4 + 10\,CO_2 + 8\,H_2O$$

This method can only be used in the absence of other reducing agents.

Gravimetry. Because of the insolubility of calcium oxalate in water (100 g of water at 18 °C dissolves only 0.0006 g of calcium oxalate), oxalic acid can be very easily and quantitatively precipitated out of aqueous solutions, e.g., with calcium chloride. The oxalic acid content of the precipitate is determined either gravimetrically (after ashing as CaO) or manganometrically.

Determination by HPLC. In this determination the oxalic acid must be esterified, separated by HPLC and analyzed by gas chromatography subsequently [24].

Enzymatic Determination. This method is usually used for analysis of oxalic acid in biological material, such as in fruit or vegetables and in body fluids. Either oxalate oxidase [25, 26], oxalate decarboxylase, or formate dehydrogenase are employed [27].

6. Uses

Apart from the application of oxalic acid as a synthetic intermediate, the properties of oxalic acid or its salts have generally been exploited as reducing agents (in which case the harmless final products of carbon dioxide and water are formed), as precipitants for calcium ions, or as complexing agents for the salts of heavy metals. The principal applications worldwide can be divided into the following groups in percentages of total use:

Metal treatment	ca. 28 %
Textile treatment	ca. 25 %
Bleaching agents	ca. 20 %
Chemical uses	ca. 27 %

6.1. Metal Treatment

Sodium, potassium, or ammonium oxalates form soluble double salts with heavy metals, especially iron. Iron (III) oxalate is the only readily soluble heavy metal salt. Oxalic acid salts are therefore used (1) for rust removal, e.g., from cooling systems (automobile radiators), boilers, or from steel plates before phosphating, (2) for removal of iron veins in marble, (3) as a constituent of metal cleaners with rust protection, especially for use on copper, silver, or aluminum, (4) in anodizing of aluminum, for the formation of very hard, abrasion and corrosion protective coatings, where through the use of additives different surface colors can be imparted to aluminum [28, 29]. Oxalic acid is also used for the cleaning of materials for electronic devices, and as coatings for protection of stainless steel, nickel alloys, chromium, alloy steels, and titanium [30, 31].

6.2. Textile Treatment

Oxalic acid is used as mordant for the printing and dyeing of wool and cotton. It serves as a stripping agent for wool colors for special pattern affects, as auxiliary in indigo caustic discharge printing, and is used in vat dyeing as a reducing agent for potassium dichromate. In laundries oxalic acid is used for the removal of rust stains and the neutralization of alkalinity. It also has an antibacterial effect [32]. Oxalic acid can be employed as a catalyst for cross linking of textile finishing agents in the permanent pressing of cellulose fabrics and for flameproofing [33].

6.3. Bleaching Agents

Oxalic acid is used in the tanning and bleaching of leather, and in bleaching of cork, wood (especially veneered wood), straw, cane, feathers as well as natural and synthetic waxes. The majority of oxalic acid is used for the bleaching of pulpwood. In leather processing oxalic acid is used in whitening instead of sodium dithionite.

6.4. Chemical Uses

Preparation of Esters and Salts. Because of the two carboxyl groups oxalic acid forms acid esters and salts as well as normal esters and salts.

Reagent in Chemical Synthesis. Since oxalic acid is a strong dehydrating agent it can be used in condensation reactions, in the esterification of secondary alcohols to esters, or in cyclization reactions [34].

Oxalic acid	$1.10/kg
Ammonium oxalate	$3.80/kg
Ammonium iron (III) oxalate	$4.40/kg
Potassium oxalate	$4.60/kg
Sodium oxalate	$5.50/kg

Concentration of Rare Earth Elements. Oxalic acid is used for the concentration and isolation of rare earth elements from aqueous solutions that are formed during the digestion of lanthanide ores (for example, monazite) with sulfuric, hydrochloric, or phosphoric acids. The

precipitation of oxalates is almost quantitative even from strongly acidic solutions. After precipitation the oxalates are converted to their oxides by heating and processes further in this form [35,36].

Other Uses. Small particles or powders, which are prepared from oxalates, are used for magnetic recording tapes. Oxalic acid is used for the preparation of polyester oxalates and copolyoxamides [37,38]; for the production of inks, pigments, and paints [39]; and for the preparation of catalysts especially those which contain precious metals and as redox initiators.

7. Economic Aspects

Specification for commercial-grade oxalic acid dihydrate is a purity of >99.5 %. Supply and demand in the world market is presented in Table 2. No significant increase in demand is anticipated since no new application areas are indicated. The average prices per kilogram in the United States (truckload quantities) in 1988 were [40]:

The most important producers of oxalic acid are:

Western Europe:	Rhône-Poulenc (Chalampé, France)
	Soc. Franc. Hoechst (Trosly–Breuil, France)
	EniChem Sintesi (Pieve-Vergonte, Spain)
	DAVSA (Vimbodi, Spain)
Japan:	Mitsubishi Gas (Yokkaichi, Mie)
	Ube Industries (Ube, Yamaguchi)
Taiwan:	Uranus Chemicals (Hsinchu, Hsien)

8. Storage, Handling, Transportation, Waste Disposal

Dry oxalic acid is packed and sold in polyethylene lined multilayered bags or in polyethylene or poly(vinyl chloride) vessels or drums. Reinforced fiberglass plastics are also used commonly and are recommended for storage.

Oxalic acid is not flammable and also not self inflammatory up to 400 °C. It is not shock sensitive and does not cause dust explosions. Oxalic acid is a weakly dangerous substance for water (hazard class 1) and exhibits low toxicity for fish and bacteria. Labelling according to Gef Stoff V: X_n (less poisonous). R-phrase: 21/22 (dangerous to health through contact with the skin and through ingestion); S-phrase: 2, 24/25. For inhalation protection masks with fine particle filters (filter class P 2) and protective gloves made from PVC or rubber are recommended [41].

Table 2. World supply – demand balance of oxalic acid ∗

Country	Production capacity	Production	Consumption	Export	Import
North America					
United States	0.0	0.0	9.0	0.0	9.0
Canada, Mexico	0.0	0.0	0.9	0.0	0.9
Total	0.0	0.0	9.9	0.0	9.9
South America					
Brazil	7.0	5.0	1.0	4.0	0.0
Venezuela	0.0	0.0	0.1	0.0	0.1
Total	7.0	5.0	1.1	4.0	0.1
Western Europe					
Spain	10.0	8.5	7.0	1.5	0.0
France	8.0	8.0	5.0	4.0	1.0
Others	4.0	3.0	15.0	1.0	13.0
Total	22.0	19.5	27.0	6.5	14.0
Eastern Europe	13.0	8.0	10.0	3.0	5.0
Asia					
China	100.0	64.0	50.0	14.0	0.0
Japan	18.0	12.4	9.9	4.8	2.3
Korea	4.8	4.5	2.5	2.0	0.0
Taiwan	4.9	4.8	2.2	2.5	0.0
India	6.5	5.5	9.5	0.0	4.0
Others	0.0	0.0	1.6	0.0	1.6
Total	134.2	91.2	75.7	23.4	7.9
World total	176.2	123.7	123.7	36.9	36.9

∗ Data are from an Ube survey, 1. Feb. 1990; data are given in 1000 t.

Neither oxalic acid nor its solutions should be discarded in the environment before proper treatment. Common treatments include incineration or heating with sulfuric acid to convert it to carbon monoxide, carbon dioxide, and water. Small amounts of oxalic acid can be treated with potassium permanganate to give carbon dioxide. Neutralization with lime results in precipitation of insoluble calcium oxalate. The latter can be safely deposited.

9. Derivatives

9.1. Salts

In addition to neutral and acid salts, oxalic acid also forms many complex compounds that easily crystallize. Simple oxalates of nearly all organic and inorganic bases are known. The three most important salts of oxalic acid are potassium hydrogen oxalate (salt of sorrel), ammonium oxalate, and ammonium iron (III) oxalate. All other salts of oxalic acid are less important industrially.

Potassium oxalate [*583-52-8*], $K_2C_2O_4 \cdot H_2O$, M_r 184.24, has the highest water solubility of any neutral oxalate (at 20 °C, 25 wt %). It was formerly used for the preparation of photographic developers and is employed today in the production of uranium in the extraction of the so-called yellow cake and in metal and textile treatment.

Potassium hydrogen oxalate [*127-95-7*], $KHC_2O_4 \cdot H_2O$, M_r 146.15, is found in many plants and was the first oxalate discovered in nature. LDLo, oral, human: 660 mg/kg.

Ammonium oxalate [*1113-38-8*], $(NH_4)_2C_2O_4 \cdot H_2O$, M_r 142.08. The anhydrous salt dissolves in water at 0 °C up to 2.27 wt % and at 100 °C up to 25.8 wt %. Ammonium oxalate is used in textile and leather processing as a replacement for oxalic acid, for the precipitation of rare earth elements, and in ore beneficiation (precipitation of heavy metals from acidic extracts).

Ammonium iron (III) oxalate [*29696-35-3*], $(NH_4)_3[Fe(C_2O_4)_3] \cdot 3H_2O$, M_r 428.08, forms emerald green crystals and is readily soluble in water and sensitive to light. It is used in the anodizing of aluminum and imparts, dependant on concentration, either yellow, gold, or brown colorations to aluminum. It is also employed in the treatment of other metals.

Sodium oxalate [*62-76-0*], $Na_2C_2O_4$, M_r 134.01, is easily prepared in high purity from oxalic acid and sodium hydroxide. Therefore, it is used as a titrimetric standard substance. The solubility in water at 0 °C is 2.6 wt % and at 100 °C is 6.3 wt %. Sodium oxalate is used for the production of super white pulp.

9.2. Organic Derivatives

(Cyanogen → Cyano Compounds, Inorganic, Chap. 5.)

Esters. Oxalic acid can be easily esterified by the usual methods. It forms both neutral (ROOC−COOR) and acidic esters (ROOC−COOH). The inherent acidity of oxalic acid is sufficient to catalyze the esterification in most instances. In some cases the esterification stops at the monoester stage. In this case anhydrous oxalic acid is used in preparation of the diester. Newer methods include the preparation from carbon monoxide and alcohols according to the Ube process (see Section 4.4) and the reaction of the easily accessible oxamide (see below) with alcohols in the presence of concentrated sulfuric acid [42].

Of industrial interest are the neutral esters, especially dimethyl, diethyl, and di-*n*-butyl oxalates. Except for the dimethyl ester which is a solid at room temperature they are all oily liquids with a faint odor, and thermally stable. Some esters can be distilled at atmospheric pressure without decomposition. The esters are readily soluble in alcohols, benzene, diethyl ether, acetone, and dioxane. Selected physical properties are presented in Table 3 (see next page). The neutral esters are good solvents for ethers and esters of cellulose (e.g., nitrocellulose) and for resins. They are therefore used for the production of special lacquers. Of greatest significance is the diethyl ester. The di-*n*-butyl ester is also used as a plasticizer. A further application results

Table 3. Physical properties of selected oxalates

Ester	M_r	*mp*, °C	*bp*, °C	d_4^{20}	CAS registry no.
Dimethyl	118.09	54	163.5	1.148 (54 °C)	[*553-90-2*]
Diethyl	146.15	−40.6	184.4	1.0785	[*95-92-1*]
Di-*n*-butyl	202.15	−29.6	245.5	0.9873	[*2050-60-4*]

from the ability of oxalic acid esters to easily undergo condensation reactions. Especially diethyl oxalate plays an important role in organic synthesis.

Oxamide [*471-46-5*], H_2N-OC-CO-NH_2, M_r 88.07, *mp* ca. 350 °C (decomp.), is sparingly soluble is water (0.3 g/L at 20 °C; 6 g/L at 100 °C).

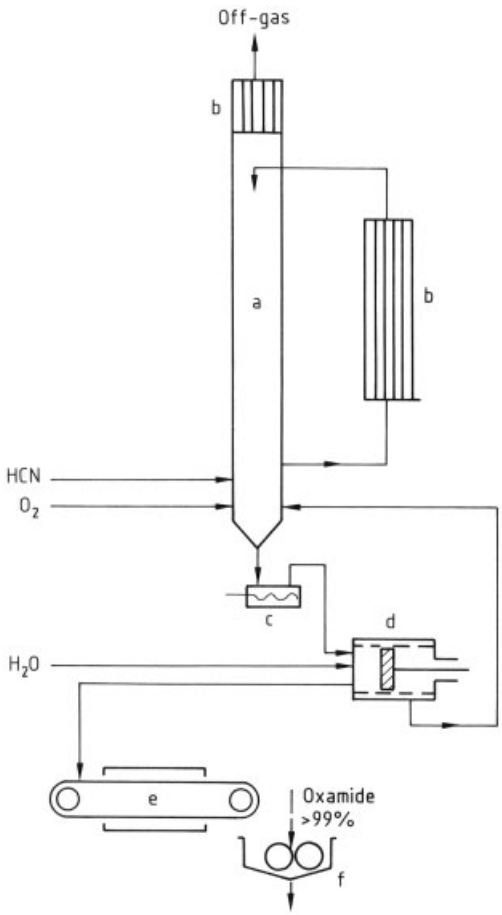

Figure 5. Continuous production of oxamide
a) Reactor; b) Cooler; c) Sludge pump; d) Centrifuge; e) Dryer; f) Granulator

The most economic method for the production of oxamide is a one-step process by Hoechst that is based on a catalytic oxidation of hydrogen cyanide [43]:

$$2\,HCN + 1/2\,O_2 + H_2O \longrightarrow \begin{matrix} CONH_2 \\ | \\ CONH_2 \end{matrix}$$

$$\Delta H = -440\ \text{kJ/mol}$$

Oxamide

The flow chart is shown in Figure 5. A solution of copper nitrate in aqueous acetic acid is employed as the catalyst. The reaction is run at 65 °C and the yield is nearly quantitative. The oxamide that is insoluble in the catalytic solution precipitates at the bottom of the reactor and is pumped from there to a centrifuge, washed, dryed and, if desired, granulated. Without further treatment it has a purity of >99 % [44]. Using this method Enimont in Gela/Sicily, began a production plant in 1990 with a capacity of 10 000 t/a [45].

Oxamide can be converted to ammonium oxalate and oxalic acid by heating it for long periods with water; best results are obtained by the application of pressure. Oxamide has proved to be of great interest as a long-term fertilizer (→ Fertilizers, Chap. 4.4.). A single application is sufficient because the nitrogen is released gradually throughout the vegetation period in a form that the plant can easily use. A granular application form of oxamide is required. The poor aqueous solubility of oxamide guarantees that the fertilizer will not be washed away by rainfall and that it can be used in rice plantations. Oxamide can also be used as a starting material for chemical syntheses.

10. Toxicology [46,47]

Like other strong acids, oxalic acid has a local caustic effect. The compound is absorbed relatively well through the skin and very well through mucous membranes, so that contact with larger amounts can quickly lead to absorptive poisoning. Oral ingestion causes acute irritation and pain in the mouth, esophagus, and stom-

ach. Nausea, vomiting of blood, collapse, convulsions, irritability, and cramps (calcium precipitation) are known to occur. Premature coronary and/or circulatory arrest, and/or a kidney malfunction which results in anuria and uremia are possible. Eventual transition to polyuremia can occur. Death occurs in the coma or through kidney failure. Workers who had contact with a 5 % solution of oxalic acid without wearing protective gloves, complained of pains in their hands within a few days. Blood clots and damaged blood vessels and finally a very painful gangrene made amputation of the fingers necessary. With solutions of higher concentrations the superficial caustic effects become increasingly significant. Inhaling the vapors and/or dust evokes responses that range from irritation of the mucous membranes to rashes, headaches, overexcitability, and albuminuria. Persistent coughing, vomiting, and general weakness are signs of chronic effects. Skin rashes and brittle yellow fingernails are also known.

Intake of food containing higher quantities of oxalic acid should be avoided because of an impaired calcium absorption through the formation of insoluble calcium oxalate: 4 – 5 g oxalic acid can be deadly for humans because the precipitated calcium oxalate plugs the kidney channels. The normal excretion of oxalic acid from humans is 20 mg per day. Two-thirds of all kidney stones are composed of calcium oxalate. The toxicological levels are [48]:

LD_{50}, oral, rat	375 mg/kg
LDLo, oral, human	71 mg/kg
RTECS-no.	Ro 245000
EC-no.	607-006-00-8

11. References

1. *Kirk-Othmer,* 3rd ed., **16,** 618 – 636.
2. R. Schröder, *DMW Dtsch. Med. Wochenschr.* **105** (1980) 997.
3. R. C. West (ed.): *CRC Handbook of Chemistry and Physics,* 61st ed., CRC Press, Boca Raton, FL, 1980.
4. H. T. Clarke, A. W. Davis: *Organic Syntheses,* 2nd ed., **coll. vol. I,** J. Wiley and Sons, New York 1946, pp. 421 – 425.
5. W. Riemenschneider: unpublished results.
6. D. Dollimore, *Thermochim. Acta* **117** (1987) 331 – 363.
7. *Ullmann,* 4th ed., **17,** 476, 477.
8. *Ullmann,* 4th ed., **17,** 478, 479.
9. *Ullmann,* 4th ed., **17,** 479.
10. S. K. Dube et al., *J. Chem. Technol. Biotechnol.* **32** (1982) no. 10, 909 – 919.
11. BIOS final report no. 663, 1946.
12. Allied Chem., US 2 057 119, 1936 (G. S. Simpson).
13. Allied Chem., US 2 322 915, 1943 (M. J. Brooks).
14. Allied Chem., US 3 536 754, 1970.
15. Mitsubishi Gas, US 3 691 232, 1972 (E. Yonemitsu et al.).
16. Mitsubishi Gas, DE 2 011 998, 1969.
17. Allied Chem., US 3 081 345, 1963 (E. J. Carlson, E. E. Gilbert).
18. H. Saito, *Shokubai* **23** (1981) no. 2, 139 – 142; *Chem. Abstr.* **95** (1981):79382u.
19. S. Uchiumi, Y. Masayosi, *J. Japan Petrol. Inst.* **25** (1982) no. 4, 197 – 204.
20. D. M. Fenton, P. J. Steinwand, *J. Org. Chem.* **39** (1974) 701.
21. Rhône-Poulenc, DE 1 618 824, 1967.
22. A. M. Gaffney et al., *Chem. Ind. (Dekker)* **33** (1988) 87 – 100.
23. F. Feigl et al., *Mikrochemie* **18** (1935) 272.
24. S. M. Steinberg et al., *Int. J. Environ. Anal. Chem.* **19** (1985) 251 – 260.
25. U. Rehmert et al., *Lab. Med.* **7** (1983) 29 – 32.
26. J. E. Buttery et al., *Clin. Chem. (Winston-Salem N.C.)* **29** (1983) 700 – 702.
27. H.-O. Beutler et al., *Z. Anal. Chem.* **301** (1980) 186.
28. S. Werneck, R. Pinner: *The Surface Treatment and Finishing of Aluminum and its Alloys,* R. Draper Ltd., Teddington 1972.
29. S. John, B. A. Shenoi, *Met. Finish* **74** (1976) no. 9, 48, 57.
30. P. E. Morris, *Corrosion (Houston)* **33** (1977) 17.
31. W. Widerholt: *The Chemical Surface Treatment of Metals,* R. Draper Ltd., Teddington 1965.
32. H. Cohen, G. E. Linton: *Chemistry and Textiles for the Laundry Industry,* Textile Book Publishers Inc., New York 1961.
33. N. A. Cashen et al., US 3 811 210, 1974.
34. L. Re, B. Maurer, T. Ohloff, *Helv. Chim. Acta* **56** (1973) 1882.
35. L. A. Sarrer et al., *J. Am. Chem. Soc.* **49** (1972) 943.
36. G. M. Varshal, *Metody Khim. Anal. Khim. Sostav Miner.*, 1971, 18 – 33; *Chem. Abstr.* **75** (1971): 133488k.

37. A. Alsnis et al., *Latv. PSR Zinat. Akad. Vestis* **5** (1976) 81 – 90; *Chem. Abstr.* **85** (1976): 78382a.
38. O. Vogl et al., *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **19** (1978) no. 2, 75 – 80.
39. N. H. Andersen et al., *Synth. Commun.* **3** (1973) 125.
40. SRI International: *Chem. Econ. Handbook,* June 1988, 682 1000 A-P.
41. GefStoffV of Aug. 28th, 1986.
42. Röhm GmbH, DE 1 668 639, 1967 (W. Gruber, J. Dehler).
43. Hoechst AG, DE 2 308 941, 1973 (W. Riemenschneider).
44. W. Riemenschneider, *Chem. Ing. Tech.* **50** (1978) 55.
45. *Chem. Ind.* **7** (1989) 18.
46. N. I. Sax: *Dangerous Properties of Industrial Material,* 6th ed., Van Nostrand Reinhold, New York 1984, p. 2098.
47. Kühn-Berett: *Merkblätter Gefährliche Arbeitsstoffe* **12/1982,** Verlag Moderne Industrie, München 0-03-1/2.
48. *Registry of Toxic Effects of Chemical Substances,* US Department of Health and Human Services, Cincinnati, Ohio 1983/1984.

Oxazine → Azine Dyes

Oxidation

Gerhard Franz, Consultant, Marl, Federal Republic of Germany (Chaps. 1 – 5)

Roger A. Sheldon, Technische Universiteit Delft, Delft, The Netherlands (Section 5.1 in part)

1. Introduction 1
2. Thermodynamics; Stability of Hydrocarbons 2
3. Reaction Types 2
4. Homolytic Oxidation 13
4.1. Historical Development; Reaction Mechanism; Radical Reactivity . . . 13
4.2. Kinetics of Autoxidation 19
4.3. Differences between Liquid- and Gas-Phase Autoxidations 23
4.4. Homolytic Oxidation in the Liquid Phase 23
4.4.1. Secondary Reactions of Radicals, Peroxides, and other Intermediates 23
4.4.2. Metal Catalysis in Homolytic Liquid-Phase Oxidation 29
4.4.3. Inhibitors of Homolytic Liquid-Phase Oxidation 33
4.4.4. Special Cases of Homolytic Oxidation 34
4.4.5. Process Technology of Homolytic Liquid-Phase Oxidation 35
4.5. Homolytic Gas-Phase Oxidation . . 38
4.6. Gas Explosions and Safety Data . . 41
5. Heterolytic Oxidation 43
5.1. Heterolytic Oxidation in the Liquid Phase 44
5.2. Heterogeneously Catalyzed Gas-Phase Oxidation 49
5.2.1. Oxidation Catalysts 50
5.2.2. Process Technology of Heterogeneously Catalyzed Oxidations 54
6. References 57

1. Introduction

The chemical term oxidation, used here only in the sense of the take-up of oxygen by organic and inorganic materials, includes such differing processes as the combustion of organic material to obtain energy or synthesis gas, biological oxidation, industrial oxidations for the production of desired substances, the corrosion of metals, and the ageing of organic polymers.

Although each of these areas is very important industrially, in this article only an overview of the chemical and industrial fundamentals of the oxidation of hydrocarbons with atmospheric oxygen or with H_2O_2 or organic hydroperoxides, and of gas explosions will be given.

Atmospheric oxygen is by far the most important and also the cheapest oxidizing agent, since all others, for example, HNO_3, H_2SO_4 and SO_3/H_2SO_4 (oleum), Cl_2, MnO_2, CrO_3, as well as H_2O_2 and the organic hydroperoxides, must first be produced by using oxygen. However oxidations with molecular oxygen are usually considerably less selective than, for example, those with Cl_2 or HNO_3 and they generally require much higher investment, which becomes economically worthwhile only on exceeding a minimum capacity. This currently lies between 10^4 and 10^5 t/a, depending on the product.

Before World War II, organic materials were mostly oxidized with inorganic oxidizing agents and only rarely directly with O_2, since the most important products of industrial organic chemistry at that time were pharmaceuticals and dyes, for which only relatively small quantities of organic intermediates were required.

Exceptions included phthalic anhydride, for which heterogeneously catalyzed gas-phase oxidation of naphthalene was developed to an industrial level in Germany in 1916 (BASF) and in the United States (Bureau of Chemistry), and maleic anhydride, which has been obtained since 1928 by the Barret Co. in an analogous process involving the oxidation of benzene.

The age of plastics, with the mass production of polyamides, polyesters, polyurethanes etc., and the precursors necessary for them, only began after World War II.

Since about the mid 1970s ecological reasons became increasingly important in decreasing the use of inorganic oxidizing agents, although they are still used to some extent. An example is hydroquinone, which was previously produced by oxidation of aniline to *p*-benzoquinone with

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_261

MnO_2 and subsequent hydrogenation, but is now obtained from diisopropylbenzene by a process analogous to the Hock phenol process.

Table 1 gives an overview of the oxidation processes used today, whereby the data are predominantly obtained from [1]. Although the table is not intended to be comprehensive, it does give an impression of the importance which oxidation with O_2 has achieved in the meantime [2].

2. Thermodynamics; Stability of Hydrocarbons

Under normal conditions in the presence of O_2, CO_2 is the only thermodynamically stable carbon compound; at higher temperatures CO also becomes increasingly stable. Thus, in the presence of free O_2, all organic materials (living organisms, wood, coal, oil, synthetic polymers, solvents, etc.) are only metastable intermediates on the way to CO_2. The stability of all organic compounds in the presence of O_2 – and hence the yield of desired products in oxidations – therefore depends only on the kinetics of the reaction involved, i.e., on the prevailing conditions (temperature, effect of high-energy radiation, nature and concentration of catalysts or stabilizers (inhibitors) present, presence of other reaction partners, impurities, etc.).

Table 2 lists roughly estimated rates of typical reactions, and the linearly extrapolated time required for complete conversion.

3. Reaction Types

The oxidation of organic materials can be roughly divided into two groups:

Homolytic oxidation reactions are chain reactions of organic and inorganic compounds with oxygen involving radicals formed by homolytic cleavage of interatomic bonds. Explosions and combustion reactions, ageing of polymers and oils, and many industrial oxidation reactions in the homogeneous liquid and gas phases belong to this group. All types of organic compounds can be subjected to radical chain oxidation. Because many of these reactions proceed spontaneously even at low temperatures they are also called autoxidations.

Radicals recombine readily on surfaces. Therefore surface enlargement is a proven tool to stop radical reactions like explosions and flames.

Metal ions that can transfer only one electron, such as Co^{2+}/Co^{3+}, Mn^{2+}/Mn^{3+}, Fe^{2+}/Fe^{3+}, and Cu^{+}/Cu^{2+}, act as catalyst in the liquid or in the solid phase. The main function of the metal ions is splitting of the hydroperoxides. Substances which readily form radicals, such as phenols, many halogen compounds etc., can act either as catalysts or inhibitors, depending on their nature, concentration, and temperature.

Because radicals isomerize easily, and the resulting intermediates are usually more readily oxidizable than the starting hydrocarbons, homolytic oxidations generally have low selectivity. Only in several special cases homolytic oxidation can be performed with high selectivity. Examples include oxidation of cumene to the corresponding hydroperoxide, of alkylhydroanthraquinone to alkylantraquinone and H_2O_2, of *p*-xylene to dimethylterephthalate (DMT) or terephthalic acid (TPA), of aldehydes to acids and some other processes.

As an introduction [3–6] are recommended.

Heterolytic Oxidation Reactions. In heterolytic oxidation reactions an active oxygen compound (e.g., a peroxyacid, hydroperoxide, or O_2) or a metal ion in its high valence state (e.g., Pd^{2+} or V^{5+}) oxidizes the starting material in a two-electron transfer reaction. Normally, a stoichiometric amount of the oxidizing compound is used and a catalyst, for instance a complex of Mo, V, Ti has to be used.

The catalyst can be dissolved homogeneously in a liquid or be present in solid form. The reaction can take place in the liquid phase or in the heterogeneous gas phase.

Examples of heterolytic oxidations in the homogeneous liquid phase are the epoxidation of olefins with hydroperoxides, catalyzed by Mo complexes, and the oxidation of ethylene to acetaldehyde by Pd^{2+} complexes.

Examples of heterogeneous gas-phase reactions are the oxidation of benzene or butane/butene to maleic anhydride; and the oxidation of propene with mixed-oxide catalysts containing Mo to give acrylonitrile (or acrolein and

Table 1. Overview of industrial oxidation processes [1]

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
I. Basic chemicals				
1: $S + O_2 \longrightarrow SO_2$ $H_2S + 1\,1/2\,O_2 \longrightarrow SO_2 + H_2O$ Metal sulfide $+ O_2 \longrightarrow SO_2$ + metal oxide	1: many processes	sulfuric acid		140×10^6
2: $SO_2 + 1/2\,O_2 \longrightarrow SO_3$	2a: gas and liquid phase, 70 – 90 °C, 1 bar, cat.: NO/NO_2, 2b: gas phase, cat.: $V_2O_5/K_2S_2O_7$, 440 – 550 °C, $\leq$ 5 bar, C: $\geq$ 99.5 %		2a: lead chamber process 2b: contact process (BASF)	2a: no longer used
$(CH_2)_x + x/2\,O_2 \longrightarrow x(CO + H_2)(+H_2O)$	oil, gas: flame reaction, coal, coke: gas – solid reaction, thermal efficiency: 70 – 95 %	synthesis gas, e.g., for methanol, NH_3, oxo process, H_2 production	many names	$> 10^8$
$2\,NH_3 + 7/2\,O_2 \longrightarrow 2\,NO_2 + 3\,H_2O$	gas phase, 900 °C, $\leq$ 5 bar, cat.: Pt/Rh gauze, Y: 93 – 95 %	nitric acid	Ostwald process	ca. 30×10^6
$2\,HCl + 1/2\,O_2 \longrightarrow Cl_2 + H_2O$	gas phase, cat.: NO_x	chlorine	Kel-Chlor process (Kellogg)	ca. 2×10^5
II. Organic intermediates **a: Olefins**				
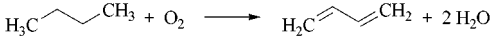	gas phase, 480 – 600 °C, 1 bar cat.: Fe_2O_3 (ferrite) + Zn, Mn, Mg, or Bi/Mo oxide, C: 65 – 80 %, S: 88 – 93 % (butadiene)	butadiene	OXD process (Phillips), Oxo-D process (Petro-Tex)	nowadays of only minor importance
b: Epoxides				
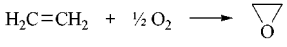	gas phase, 245 – 275 °C, $\leq$ 20 bar, cat.: Ag/carrier, ethylene excess, S: 70 – 80 %	ethylene oxide	(UCC, IG-Farben, Shell)	ca. 6×10^6
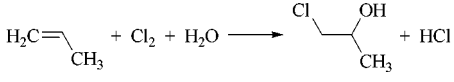	liquid phase, 35 – 50 °C, 1 bar, no cat., S: $\geq$ 95 %	propylene oxide	chlorhydrin process	ca. 10^6
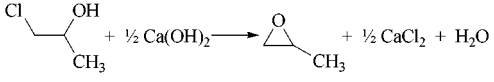				

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
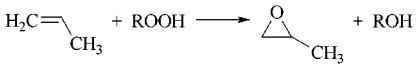	liquid phase, 90 – 120 °C, ≤ 60 bar, cat.: Mo complexes, olefin excess, C (olefin): ca. 10 % S: ca. 90 %	propylene oxide	(Halcon & Arco, Oxirane Division)	ca. 2×10^6
c: Hydroperoxides				
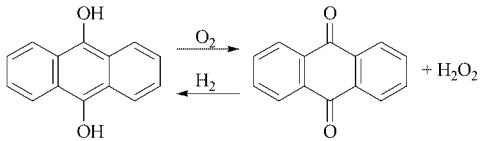	liquid phase, cyclic process, oxidation: no cat., 30 – 80 °C hydrogenation: Pd cat., 40 °C, ≤ 5 bar	H_2O_2 (in H_2O)	anthraquinone process (BASF, IG-Farben)	ca. 5×10^5
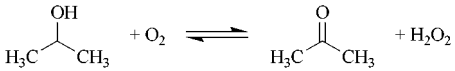	liquid phase, 90 – 130 °C, ≤ 4 bar, no cat., C: ca. 15 %, S: ca. 90 %	H_2O_2 and acetone	(Shell)	last plant shut down in 1980
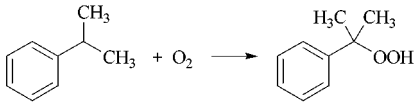	liquid phase, 80 – 120 °C, 4 – 10 bar, no cat., pH 8.5 – 10.5 C: < 30 %, S: > 90 %	cumene hydroperoxide (for phenol + acetone)	Hock phenol process (Hercules, Gulf)	ca. 4×10^6
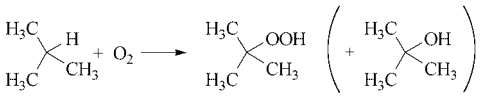	liquid phase, 120 – 140 °C, ≤ 40 bar, no cat., C: < 30 %, S: ca. 60 %	*tert*-butyl hydroperoxide (for propene epoxidation)	(Halcon + Arco, Oxirane Division)	ca. 4×10^6
d: Alcohols				
$Al[(C_2H_4)_n—C_2H_5]_3$ $+ 1½\ O_2$ ↓ $Al[O(C_2H_4)_n—C_2H_5]_3$ $+ 3\ H_2O$ ↓ $Al(OH)_3 + 3\ C_2H_5(C_2H_4)_n—OH$	3-stage process (polymeriz. + oxidat. + hydrolysis), oxidation: liquid phase, 50 – 100 °C, 10 bar, cat., very dry air, hydrolysis: only H_2O or dil. H_2SO_4, S: 85 – 91 % (alcohols)	primary fatty alcohols (+ $Al(OH)_3$ or $Al_2(SO_4)_3$)	Alfol process	ca. 2×10^5

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
Paraffins + O_2 → OH, R^1, R^2 + H_2O	liquid phase, 150 – 180 °C, ≤ 20 bar, cat.: HBO_2, C: < 25 %, S: > 70 % alcohol (rest: 20 % ketones, 10 % acids)	secondary alcohols (for surfactants, etc.)	Bashkirov process (Riebeck'sche Montanwerke)	
O_2, $-H_2O$; H, OH + O; 7 : 1	liquid phase, 150 – 180 °C, ≤ 20 bar, cat.: HBO_2, C: < 25 %, S: > 80 % alcohol (< 20 % ketones)	cyclododecanol/-one (for dodecanedicarboxylic acid)	(Hüls, Shell)	$< 10^5$
O_2, $-H_2O$; HO H + O; 9 : 1	liquid phase, 150 – 180 °C, ≤ 20 bar, cat.: HBO_2, C: < 15 %, S: > 90 %	cyclohexanol/-one (for adipic acid)		ca. 10^6
O_2, $-H_2O$; HO H + O; 1 : 1	liquid phase, 130 – 160 °C, ≤ 15 bar, cat.: Co, Mn, C: < 15 %, S: 80 – 85 %	cyclohexanol/-one (for adipic acid)		
H_2C OH; 1. H_2O_2 (or AcOOH) 2. H_2O → OH, HO OH	liquid phase, 50 – 70 °C, cat.: WO_3, MoO_3, Y: ca. 90 %	glycerol		ca. 10^5
e: Aldehydes + ketones				
$CH_3OH + ½ O_2 \longrightarrow CH_2O + H_2O$	gas phase a: cat.: Ag, 600 – 720 °C, 1 bar, O_2 deficiency, (+ H_2O), C: ≈ 95 %, Y: > 98 %, b: cat.: MoO_3/Fe_2O_3, 350 – 450 °C, O_2 excess, C: 95 – 99 %, Y: 91 – 94 %	formaldehyde	a: (BASF) b: Formox process (Perstorp)	$\leq 4 \times 10^6$

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
$CH_3CH_2OH + ½ O_2 \longrightarrow CH_3CHO + H_2O$	gas phase, 450 – 550 °C, 1 – 2 bar, cat.: Ag, O_2 deficiency, Y: ca. 95 % (+ 2 % AcOH)	acetaldehyde	(VEBA-Chemie, now Hüls)	ca. 10^5, shut down
$H_2C{=}CH_2 + ½ O_2 \longrightarrow CH_3CHO$	1- or 2-stage process, liquid phase, 100 – 120 °C, $\leq$ 10 bar, cat.: $Pd/Cl_2/CuCl_2$, Y: $\leq$ 95 %	acetaldehyde	(Wacker–Hoechst)	ca. 2×10^6
$H_2C{=}CH{-}CH_3 + O_2 \longrightarrow H_2C{=}CH{-}CHO + H_2O$	gas phase, 320 – 400 °C, < 2 bar, cat.: $MoO_3/Bi_2O_3/WO_3$, Fe, Co, K, C: $\geq$ 95 %, Y: 80 – 90 % acrolein, 15 – 5 % acrylic acid	acrolein, acrylic acid	(Sohio)	ca. 10^5
f: Phenols + precursors				
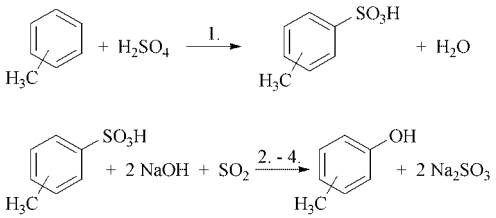	4-stage process, liquid phase 1: sulfonation, ca 100 °C 2: salt formation (with Na_2SO_3) 3: NaOH melt ($\leq$ 350 °C) 4: liberation with SO_2 with toluene with xylene	phenol resorcinol [a] (cresols) (xylenols) + ca. 1.4 t Na_2SO_3 + ca. 2 t Na_2SO_4 per tonne phenol cresoles xylenols		ca. 4×10^4 [a]
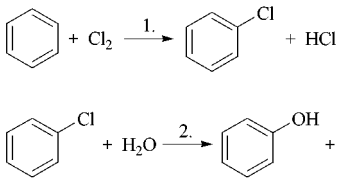	1a: Cl_2 addition, liquid phase, 20 – 50 °C, cat.: $FeCl_3$ 1b: oxychlorination (HCl + air), gas phase, 240 °C, cat.: $CuCl_2$, $FeCl_3/Al_2O_3$, C (benzene): ca. 10 %,	chlorobenzene (byproduct: $\leq$ 10 % dichlorobenzenes)		ca. 5×10^5
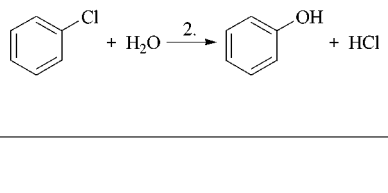	2a: liquid phase, $\leq$ 390 °C, 300 bar, + NaOH 2b: gas phase, 450 °C, + H_2O, cat.: $Ca_3(PO_4)_2/SiO_2$, S (both stages): 70 – 85 % phenol	phenol		$< 10^5$

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
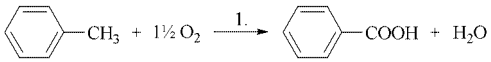	1: liquid phase, 110 – 120 °C, ≤ 3 bar, cat: Co, S: ca. 90 %	benzoic acid		1.5×10^5
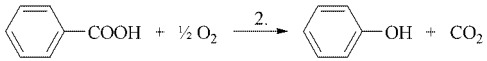	2: liquid phase, 220 – 250 °C, cat.: Cu salts, S (toluene): 70 – 80 %	phenol	(California Research)	ca. 2×10^5
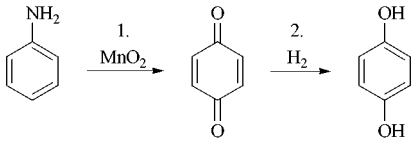	1: liquid phase, 3 – 10 °C, oxidant: MnO_2 or $CrO_3 + H_2SO_4$ 2: liquid phase, 50 – 80 °C, reductant: Fe^0 + HCl	hydroquinone		no longer used in Japan and United States
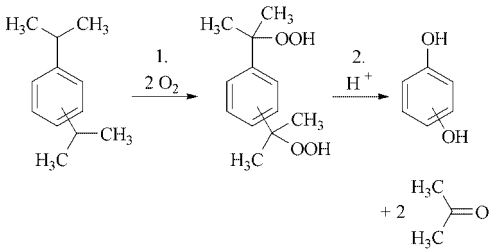	1: liquid phase, 110 – 120 °C, ≤ 8 bar, no cat. 2: liquid phase, acid cleavage	resorcinol (+ hydroquinone) hydroquinone (+ resorcinol)	Hock phenol process (Mitsui Petrochem., Sumitomo Chem. Co.)	ca. 10^4 ca. 10^4
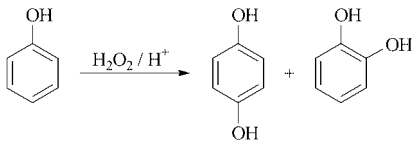	liquid phase, 90 °C a: cat.: carboxylic or mineral acid C: < 10 % b: cat.: Fe, Co salts, 40 °C, C: 20 %, S: > 90 %	hydroquinone (+ pyrocatechol) a: ratio: 60/40 to 40/60 b: ratio: 40/60 to 20/80		ca. 3×10^4 (+ ca. 2×10^4)
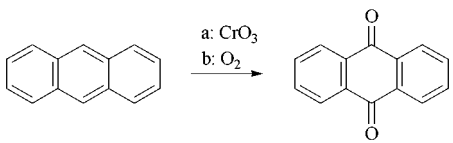	a: liquid phase, 50 – 100 °C, $CrO_3 + H_2SO_4$ b: gas phase, 340 – 390 °C, cat.: $FeVO_4$, air C: 100 %, S: ca. 90 %	anthraquinone	(BASF)	4×10^4

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
g: Carboxylic acids + anhydrides				
Butane (naphtha) $\xrightarrow{O_2}$ CH_3COOH (+ byproducts)	various variants, all liquid phase, a: no cat., 180 °C, O_2/20 bar (UCC), air/50 bar (Brit. Dist.) b: cat.: Co/Mn, 175 °C, 60 bar, C: 10–20 %, S: ca. 40 % (AcOH)	acetic acid, many side products: propionic acid, formic acid, methyl ethyl ketone, acetone, methyl acetate, ethyl acetate	a: (Brit. Distillers) b: LPO process (Celanese)	ca. 10^5 many plants shut down
Paraffins $\xrightarrow{O_2}$ fatty acids (+ byproducts)	liquid phase, 110–130 °C, cat.: Co/Mn salts, C: $\leq$ 30 %, S: ca. 80 %	fatty acids	(Imhausen, Henkel)	ca. 10^5
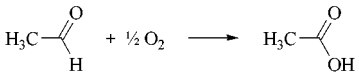	liquid phase, 50–90 °C, $\leq$ 10 bar, cat.: Mn acetate, O_2, S: 95–97 %	acetic acid		ca. 10^6
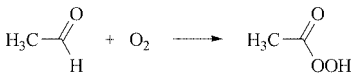	liquid phase, − 15 °C, no cat.	peracetic acid		$< 10^5$
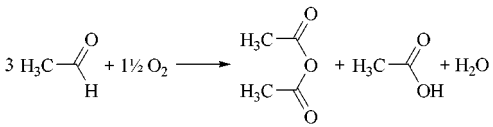	liquid phase, 50–60 °C, $\leq$ 5 bar, cat.: Cu/Co acetate, C: 80 %, S: $\leq$ 60 % (rest AcOH) (+ CO + CO_2)	acetic anhydride	(Hoechst-Knapsack, Shawinigan; now Gulf Oil)	ca. 6×10^5
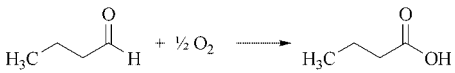	liquid phase, 50–90 °C, $\leq$ 5 bar, no cat., C: 98 %, S: > 95 %	*n*-butyric acid		$< 10^5$
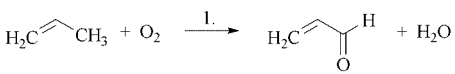	two-stage process 1: see above (acrolein) 2: gas phase, 250–300 °C, < 2 bar, cat.: e.g., $MoO_3/V_2O_5/WO_3$/Cu/Cr, H_2O addition, C: 100 %, S: > 95 %, Y: ca. 85 % (propene)	acrylic acid	(Sohio, Nippon Kayaku)	ca. 10^6
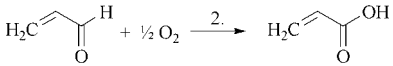				

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
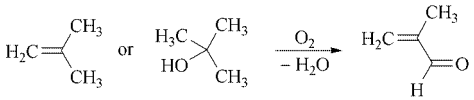	two-stage process gas phase, 300 – 350 °C, < 2 bar, cat.: mixed oxide (with Mo, Sb, W, P)	methacrylic acid	(Asahi Chem., Mitsubishi Rayon, Nippon Shokubai, Sumitomo)	ca. 10^5
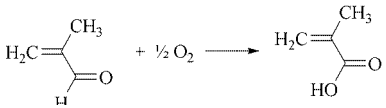				
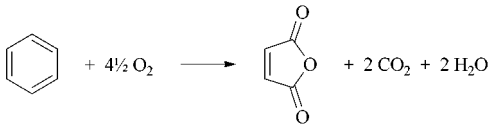	gas phase, 380 – 450 °C, < 3 bar, cat.: V_2O_5(+ MoO_3) on Al_2O_3, C: 85 – 95 %, S: 60 – 75 % (new cat. S: ≤ 95 %)	maleic anhydride	(BASF, Scientific Design)	ca. 2×10^5
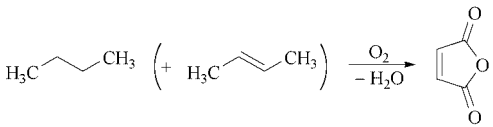	gas phase, 400 – 450 °C, < 3 bar, cat.: $V_2O_5/P_2O_5/TiO_2$ on Al_2O_3, C: 80 – 95 %, S: 40 – 55 %	maleic anhydride	(Petrotex, now Denka) ALMA process (Alusuisse/Lummus)	ca. 3×10^5
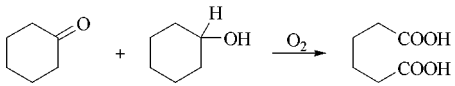	various variants, all liquid phase, usually AcOH as solvent, 80 – 85 °C, 6 bar, cat.: $Co^{2+/3+}/Mn^{2+/3+}$	adipic acid		only minor importance
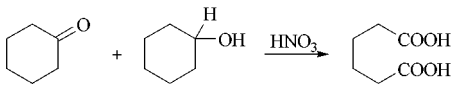	liquid phase, 60 – 80 °C, < 2 bar, cat.: $(VO)^{2+}$, Cu^{2+} C: 100 %, S: 92 – 96 %	adipic acid analogue: dodecanedicarboxylic acid	(DuPont, BASF, Hüls)	$> 10^6$ $< 10^5$

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
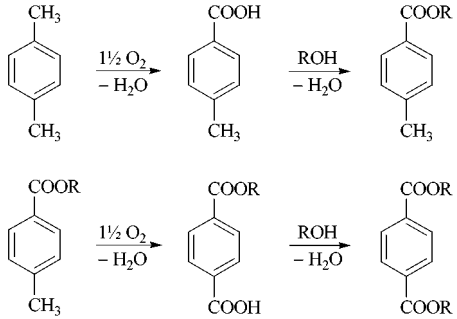	1- or 2-stage process 1: oxidation, liquid phase, 150 – 180 °C, < 8 bar, cat.: $Co^{2+/3+}$, $Mn^{2+/3+}$, C: 100 %, S: 85 – 90 % (*p*-xylol), 80 % (methanol) 2: esterification, liquid phase, 250 – 280 °C, ≤ 100 bar, no cat. [a] oxidation + esterification simultaneously	dimethyl terephthalate (DMT)	Witten–Imhausen–Katzschmann process, now Hüls (Kretzschmann/Witten, BASF[a], DuPont[a], Montecatini[a])	ca. 3.5×10^6
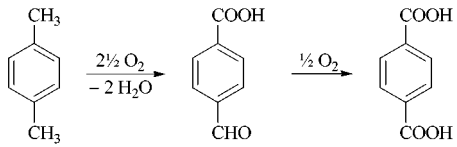	1-stage, liquid phase, solvent: AcOH, a: 160 – 190 °C, < 30 bar, cat.: $Co^{2+/3+}/Mn^{2+/3+}$ + Br, C: ≥ 95 %, S: ≥ 90 % b: 100 – 140 °C, MEK or paraldehyde as cooxidant cat.: $Co^{2+/3+}Mn^{2+/3+}$ C: ≥ 95 %, S: ≥ 95 %	terephthalic acid (TPA)	a: (Amoco/Mid Century; now Standard Oil of Indiana, IFP) b: (Toray, Eastman Kodak, Mobil)	ca. 3×10^6
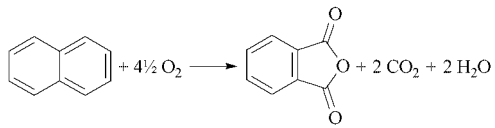	gas phase, fixed or fluidized bed, 350 – 400 °C or 400 – 550 °C, cat.: V_2O_5/K_2SO_4 on SiO_2, C: ca. 90 %, S: ca. 90 %	phthalic anhydride	(BASF, Chem. Fabrik von Heyden)	$< 2 \times 10^5$
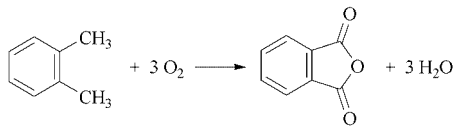	gas phase, 370 – 420 °C, cat.: $V_2O_5/P_2O_5/TiO_2$ on Al_2O_3, C: ≥ 90 %, S: ca. 80 %	phthalic anhydride	(BASF, Chem. Fabrik von Heyden, Alusuisse)	2.5×10^6

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
h: Esters + lactones				
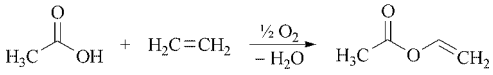	a: gas phase, 150 – 250 °C, ≤ 10 bar, cat.: Pd, Bi_2O_3/Na acetate, O_2 deficiency, C (C_2H_4): ≤ 10 %, C (AcOH): ≤ 25 %, S (allyl acetate): ≥ 90 % b: liquid phase, 110 – 130 °C, ≤ 40 bar, cat.: $PdCl_2$/$CuCl_2$/Na acetate (analogue Wacker process for acetaldehyde)	vinyl acetate	a: (Bayer, Hoechst, USI) b: (ICI, Celanese)	2×10^6 shut down
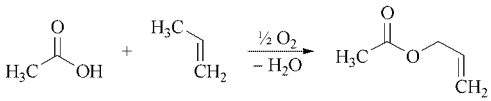	analogue vinyl acetate variant a	allyl acetate	(Bayer, Hoechst)	
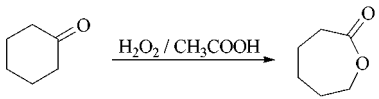	liquid phase, 50 – 70 °C, S: ca. 90 %	caprolactone (for ε-caprolactam)	Baeyer–Villiger reaction (UCC)	2.5×10^4
i: Nitriles				
$CH_4 + 1½\,O_2 + NH_3 \longrightarrow HCN + 3\,H_2O$	various variants, gas phase, 1000 – 1200 °C, cat.: Pt/Rh-gauze, S: ≤ 90 % (CH_4), 90 % (NH_3)	HCN (for cyanuric chloride)	Andrussow process (BASF)	ca. 10^5
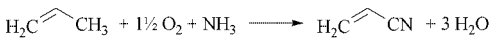	gas phase, 320 – 450 °C, < 2 bar, fixed or fluidized bed, cat.: mixed oxide, e.g., $Bi_2O_3/MoO_3/Fe^{2+/3+}$, C: 95 %, S: ≤ 80 %	acrylonitrile, side products: acetonitrile, HCN	(Sohio)	2.7×10^6

Table 1. (Continued)

Oxidation reaction	Reactions conditions*	Target product	Process name (developed by)	Installed capacity, t/a
j: Oximes				
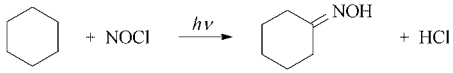	photonitrosation, liquid phase, $< 20\,^{\circ}C$, S: 86 % (cyclohexane)	cyclohexanone oxime (for ε-caprolactam)	PNC process (Toray)	1.5×10^5
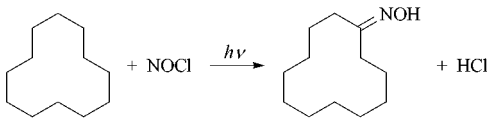	photonitrosation, liquid phase, 70 °C, C: ≤ 70 %, S: ≥ 85 %	cyclododecanone oxime (for laurolactam)	(Aquitaine Total Organico, Toray)	$< 10^4$
k: Oxychlorination				
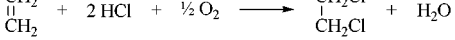	gas phase, 220–240 °C, ≤ 4 bar, cat.: $CuCl_2$ + alkali/carrier, C: ≤ 100 %, S: ≤ 96 %	dichloroethane (for vinyl chloride)	(Goodrich, Dow)	$> 2 \times 10^6$
l: Sulfoxidation				
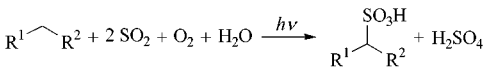	liquid phase, $\leq 40\,^{\circ}C$, no cat., UV, H_2O added, C: ≤ 5 %, S: ≥ 90 %	paraffin sulfonic acids (H_2SO_4, dilute)	light-water process (Hoechst)	$\geq 10^5$

*C = conversion; S = selectivity; Y = yield.

Table 2. Rates u of various oxidation reactions and linearly extrapolated time t required for complete formula conversion

u^a, molecules $cm^{-3} \cdot s^{-1}$	u^b, mol L^{-1} s^{-1}	t^c	Examples of reactions
$10^3 - 10^{10}$	$10^{-18} - 10^{-11}$	$10^{10} - 10^3$ yr	initial rate of uncatalyzed hydrocarbon oxidations
$10^{13} - 10^{14}$	$10^{-8} - 10^{-7}$	2 yr – 10 weeks	degradation of unstabilized plastics under irradiation
10^{16}	10^{-5}	ca. 20 h	industrial radical reactions (e.g., liquid-phase oxidation and polymerization)
10^{18}	10^{-3}	ca. 10 min	heterolytic reactions (e.g., epoxidation with ROOH/Mo complex)
10^{20}	10^{-1}	ca. 6 s	heterolytic reactions heterogeneously catalyzed hydrocarbon oxidations
10^{22}	10	ca. 10^{-2} s	heterogeneously catalyzed hydrocarbon oxidations
10^{24}	10^3	ca. 10^{-4} s	explosions; ion reactions

[a] Pure hydrocarbons contain ca. 6×10^{21} molecules per cubic centimeter.
[b] 1 mol/L corresponds to ca. 6×10^{20} molecules per cubic centimeter.
[c] 1 year $\approx$ 8760 h $\approx 3.15 \times 10^7$ s.

acrylic acid) and with Pd catalysts to give allyl acetate.

Unlike radical autoxidations, heterolytic oxidations proceed via defined complexes and are therefore, in principle, very selective.

For an introduction and review see [4] and [16]; further literature: [12], [17–22].

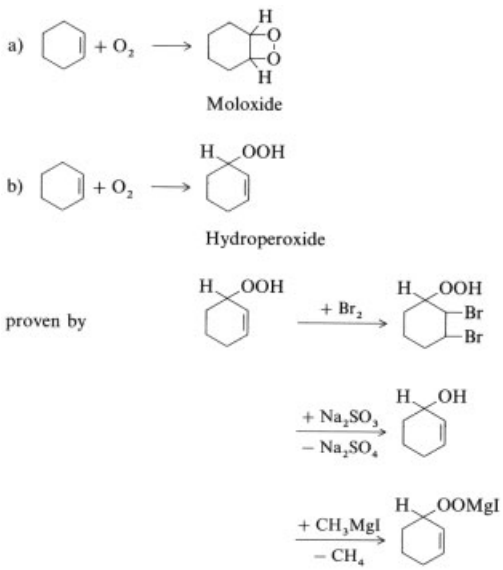

Figure 1. Concepts of the structure of peroxides
a) Moloxide (ENGLER, MOUREU et al.); b) Hydroperoxide [3]

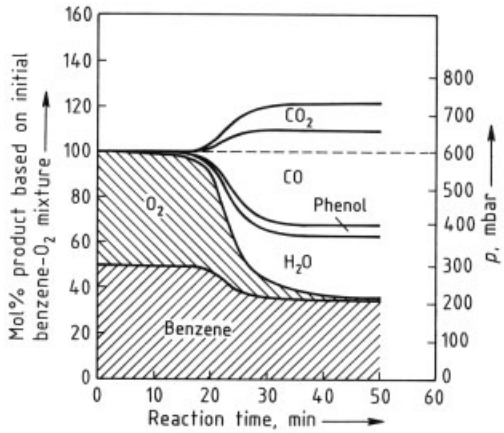

Figure 2. Typical course of a slow hydrocarbon oxidation (oxidation of a 1:1 benzene/O_2 mixture in a glass flask; $T = 470\,°C$, initial pressure $p = 600$ mbar)

4. Homolytic Oxidation

4.1. Historical Development; Reaction Mechanism; Radical Reactivity

An extensive study of the history of homolytic oxidation is still missing. Short reviews of the historical development of knowledge of homolytic oxidation can be found, for example, in [3, pp. 1 – 14], [5], and, with emphasis on the development of homogeneous gas-phase oxidation, in a publication by SHTERN [10].

The development of concepts of oxidation began with the introduction of the phlogiston the-

ory in the 17th century by BECHER and STAHL and its disproving in 1773 by LAVOISIER. The development of theories for the oxidation of organic materials started in 1897 with ENGLER's proposal that the first step in the oxidation of an olefin consists in the addition of one O_2 molecule to a C=C double bond forming a four membered ring (see Fig. 1 a). Later on this theory was called "moloxide"theory by STAUDINGER. Despite the fact, that a growing number of arguments could be brought forward against the moloxide theory, it was accepted until 1939. In this year CRIEGEE and coworkers could unambiguously show, that in the first stage of the oxidation of an olefin a hydroperoxide is formed (see Fig. 1 b). Already in 1900 BAEYER and VILLIGER had isolated perbenzoic acid as an intermediate in the oxidation of benzaldehyde to benzoic acid.

Another problem that was difficult to understand for a long time was the S-shaped slope of the velocity curve of the oxidation of hydrocarbons (see Fig. 2). Since the few radicals known to the late 1920s were very reactive and therefore short lived, nearly nobody could imagine that radicals could significantly participate in the slow oxidation of hydrocarbons. CHRISTIANSEN had therefore developed in 1924 the concept of an "energy chain reaction" with the inclusion of a "hot" molecule.

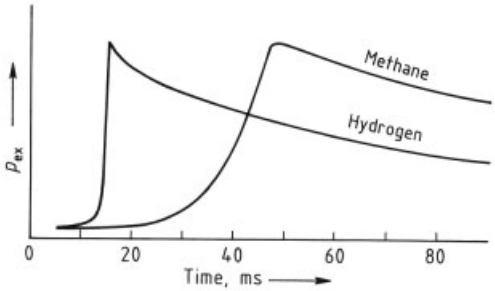

Figure 3. Course of a methane explosion and a hydrogen explosion in a sealed vessel (stoichiometric mixture with O_2) [23, p. 10]

In 1927 BÄCKSTRÖM opposed this idea with the concept of a radical chain reaction, on the grounds of the high quantum yield in photochemical oxidations.

A mathematical solution to a similar problem – the reaction of H_2/O_2 mixtures – was found by SEMENOV in 1926 who based his theory on the work of BODENSTEIN and NERNST. The latter developed the theory of the unbranched chain reaction taking the experimental results of BODENSTEIN, who studied the photochemically induced Cl_2/H_2 reaction in 1913 (Scheme I). In 1926 SEMENOV first proposed the theory of the branched chain reaction describing mathematically the pressure rise in the H_2/O_2 explosion (Scheme II, see [7–10]).

Scheme I. Unbranched radical chain reaction

Initiation

$Cl_2 + h\nu \longrightarrow 2\,Cl\cdot$ (1)

Propagation

$Cl\cdot + H_2 \longrightarrow HCl + H\cdot$ (2)

$H\cdot + Cl_2 \longrightarrow HCl + Cl\cdot$ (3)

Termination

$2\,H\cdot + (wall) \longrightarrow H_2 + (wall)$ (4)

$2\,Cl\cdot + (wall) \longrightarrow Cl_2 + (wall)$ (5)

$H\cdot + Cl\cdot + (wall) \longrightarrow HCl + (wall)$ (6)

Scheme II. Branched radical chain reaction

Inititiation

$H_2 + O_2 \longrightarrow 2\,HO\cdot$ (7^r)

Propagation

$HO\cdot + H_2 \longrightarrow H_2O + H\cdot$ (8^r)

$HO\cdot + \cdot OH \longrightarrow HOO\cdot + H\cdot$ (9^r)

$\cdot O\cdot + H_2 \longrightarrow \cdot OH + H\cdot$ (10^r)

Branching

$H\cdot + O_2 \longrightarrow HO\cdot + \cdot O\cdot$ (11^r)

$HOO\cdot \longrightarrow HO\cdot + \cdot O\cdot$ (12^r)

Termination

$H\cdot + \cdot OH + (wall) \longrightarrow H_2O + (wall)$ (13^r)

$H\cdot + \cdot H + (wall) \longrightarrow H_2 + (wall)$ (14^r)

$\cdot O\cdot + \cdot O\cdot + (wall) \longrightarrow O_2 + (wall)$ (15^r)

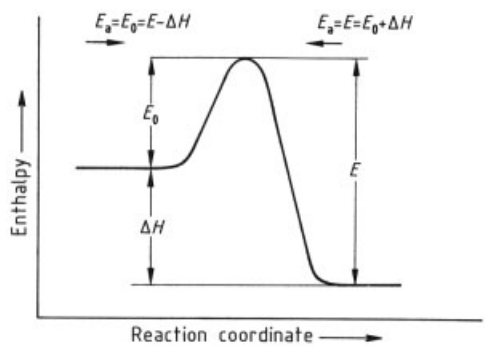

Figure 4. Relationship between E_a, E, E_o, and ΔH for exothermic ($E_{a\rightarrow}$) and endothermic ($E_{a\leftarrow}$) reactions

In the early 1930s SEMENOV was able to prove mathematically that the pressure rise in the autoxidation of hydrocarbons could also be explained using the model of a branched chain reaction. Because of the long induction period and the slow rise of the pressure, compared with the

very fast H_2/O_2-reaction, he called this reaction "degenerate" branched chain reaction (see Fig. 3 and [9, pp. 317ff and pp. 438ff]).

Confirming the findings of CRIEGEE, FARMER in 1941 – 1942 was the first to propose the chemical radical chain mechanism of the homolytic oxidation of hydrocarbons (Scheme III).

Scheme III. Hydrocarbon oxidation (degenerate branched radical chain reaction)

Initiation

$R\text{–}H + O_2 \longrightarrow R\cdot + \cdot OOH$ (16^r)

$R\text{–}H + O_2 + H\text{–}R \longrightarrow R\cdot + HOOH + \cdot R$ (17^r)

$R\text{–}R + h\nu \longrightarrow R\cdot + \cdot R$ (18^r)

Propagation

$R\cdot + O_2 \longrightarrow ROO\cdot$ (19^r)

$R\text{–}H + ROO\cdot \longrightarrow R\cdot + ROOH$ (20^r)

$ROO\cdot + C{=}C \longrightarrow ROO\text{–}CR_2\text{–}C(R_2)\cdot$ (21^r)

Branching

$ROOH\,(ROOR) \longrightarrow RO\cdot + \cdot OH\,(+\cdot OR)$ (22^r)

$ROOH + ROOH \longrightarrow ROO\cdot + H_2O + RO\cdot$ (23^r)

$ROOH + ROH \longrightarrow RO\cdot + H_2O + RO\cdot$ (24^r)

Termination

$R\cdot + R\cdot \longrightarrow R\text{–}R$ (25^r)

$RO\cdot + R\cdot \longrightarrow ROR$ (26^r)

$ROO\cdot + ROO\cdot \longrightarrow ROOR + O_2$ (27^r)

$ROO\cdot + R\cdot \longrightarrow ROOR$ (28^r)

The essential differences between the relatively slow hydrocarbon oxidation and the very fast oxygen – hydrogen reaction (see Fig. 3) and similar branched radical chain reactions are:

1) In the hydrocarbon oxidation, hydroperoxide formation occurs prior to the branched hain reaction. Hydroperoxide formation takes place by an unbranched radical chain reaction according to Reactions (19^r) and (20^r) (propagation in Scheme III).
2) The chains branch by decomposition of the alkylhydroperoxide molecule. However, molecules react orders of magnitude more slowly than radicals (therefore "degenerate" chain branching). Only at high temperatures branching of chains occurs via peroxy radicals in hydrocarbon oxidations. But even then these oxidations are not as fast as the H_2/O_2-reaction (see Fig. 3 and Section 4.5).
3) In large hydrocarbon radicals the energy absorbed on dissociation is dissipated rapidly over the entire radical by atomic vibrations. The radical thus loses a considerable portion of reactivity. This behavior is particularly marked for radicals from aromatic compounds and olefins in which the unpaired electron (odd electron) is additionally stabilized by the π-electron system.
4) Because of their lower reactivity these radicals have a longer lifetime than the more reactive radicals.
5) Because of the longer lifetime and the ability to rapidly dissipate energy, the large radicals can readily recombine in binary collisions (chain termination). The recombination of small radicals (e.g., $H\cdot + \cdot H$ or $H\cdot + \cdot OH$) occurs only in the much rarer three-body collisions or at the wall because of the need to remove the bond energy liberated.

Bond Strength and Radical Reactivity. Radical chain reactions essentially occur through the attack of a radical at a covalent bond such that an existing bond is broken and a new one formed:

$R_3C\text{–}H + X\cdot \longrightarrow R_3C\cdot + H\text{–}X \; -\Delta H$ (29^r)

In such homolytic substitution reactions, a radical is more reactive the greater the value of $-\Delta H$, i.e., the greater the difference between the bond energies according to:

$$\Delta H_{(H-X)} - \Delta H_{(C-H)} = -\Delta H \text{ kJ/mol} \quad (1)$$

According to the rule of POLANYI and EVANS, more precisely formulated by SEMENOV [9, p. 26 ff], from the equation:

$$E_a = 48 + 0.25\,(-\Delta H) \text{ kJ/mol} \quad (2)$$

the activation energy E_a of an exothermic reaction is lower and the reaction faster the greater the value of $-\Delta H$ (see Figs. 4 and 5).

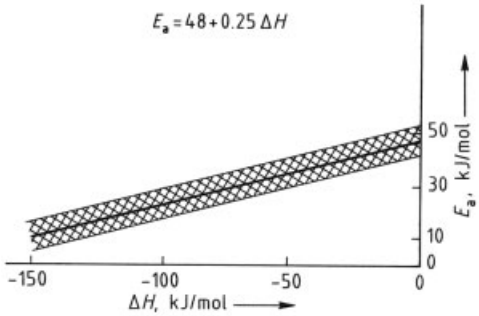

Figure 5. Dependence of the activation energy E_a on the enthalpy of reaction ΔH for exothermic radical exchange reactions of type $A - B + C\cdot \rightarrow A\cdot + B - C$ [7–9]
The reactions investigated by POLANYI and EVANS lie in the shaded area

In Figure 6 dissociation energies for various types of bond are shown. The following is noted:

1) The attack of a radical at a C – H bond takes place more easily and faster the lower the energy of the bond and the greater the energy of the attacking radical. The reactivity of C – H bonds increases with decreasing bond energy:
   Vinyl ≈ phenyl < primary < secondary < tertiary < allyl < cumyl < diallyl
2) The first members of homologous series (e.g., methane, ethylene, benzene) have particularly high C – H bond energies and are correspondingly stable towards attack by radicals. In alkanes the energy of the primary C – H bonds decreases with increasing chain length. Alkenes are either attacked at the allylic C – H bond or the radicals add to the C=C double bond.
3) The hydroxy, alkoxy, and acyloxy radicals can abstract a H atom from all C – H bonds. The peroxy radicals, in contrast, exhibit a similarly high reactivity only towards C – H bonds of hydrocarbons, alcohols, and aldehydes that lie below the tertiary C – H bond of isobutane in the energy scale. With C – H bonds of compounds which lie above isobutane, these radicals react much more slowly, but much more selectively (see Section 4.4.1, Tables 9, 10).
4) Through the cleavage of the comparatively weak O – O bond in peroxides the highly active HO · and RO · radicals are formed.

According to Figure 6 the C – H bond of the aldehyde group is less reactive than the allylic C – H bond in olefins or the tertiary C – H bond of cumene. However, in reality aldehydes are much more reactive than these hydrocarbons because they can be oxidized at much lower temperatures than e.g., cumene or butene (see Table 3, last line). There are three reasons for this confusing phenomenon:

1) The data in Figure 6 originate from measurements in the gas phase. In the case of hydrocarbons these data can be applied to the reactivity in the liquid phase without complications because in the liquid phase hydrocarbon molecules interact with one another only weakly and are not altered chemically until they are oxidized. But in the case of aldehydes these data are misleading, because aldehydes form addition products with polar compounds like water, alcohols, and acids in the liquid phase. In the addition product the remaining C – H bond of the aldehyde group is strongly activated by either two OH groups or one OH and one O–R group at the same C-atom.
2) Peroxyacyl radicals are much more reactive than peroxyalkyl radicals in cleaving a C – H bond to gain a H-atom in order to stabilize to the corresponding peroxy acids.
3) Peroxy acids decompose much faster to radicals than alkyl hydroperoxides (see Section 4.4.1).

Initiation; Induction Period; Peroxide Formation. Small molecules such as H_2, CH_4, and NH_3 normally exist as gases. Because of the stable bonds, which necessitate high dissociation energies, and because of the weak interactions between the gas molecules, these types of molecule only react with oxygen at high temperatures, if no catalysts or initiators are present. At low temperatures their mixtures with oxygen are metastable in the absence of catalysts.

In the autoxidation of higher hydrocarbons the situation is quite different, especially in the liquid and solid phase. Because of intermolecular interactions and the relatively low bond energies of many C – C and C – H bonds, radicals are formed relatively rarely, but continually, even at low temperatures. This may occur by reaction with O_2 (Reactions 16^r or 17^r) or by concentration of the thermal vibration energy into one bond (cf., the lowering of the ignition temperature for gas explosions from ca. 600 °C for methane to ca. 200 °C for decane, see Table 15). The rate of radical formation is increased drastically by high-energy radiation (sunlight) and heat. In most cases these radicals recombine – above all in the liquid or solid phase, where the mobility of the molecules and their fragments is severely limited as a result of the forces of interaction (solvent cage). However, if oxygen is present, it competes with the simple radical recombination (Reaction 25^r) because of its diradical structure (Reaction 19^r). These types of radical recombination take place without activation energy and are therefore very fast. This mechanism is the

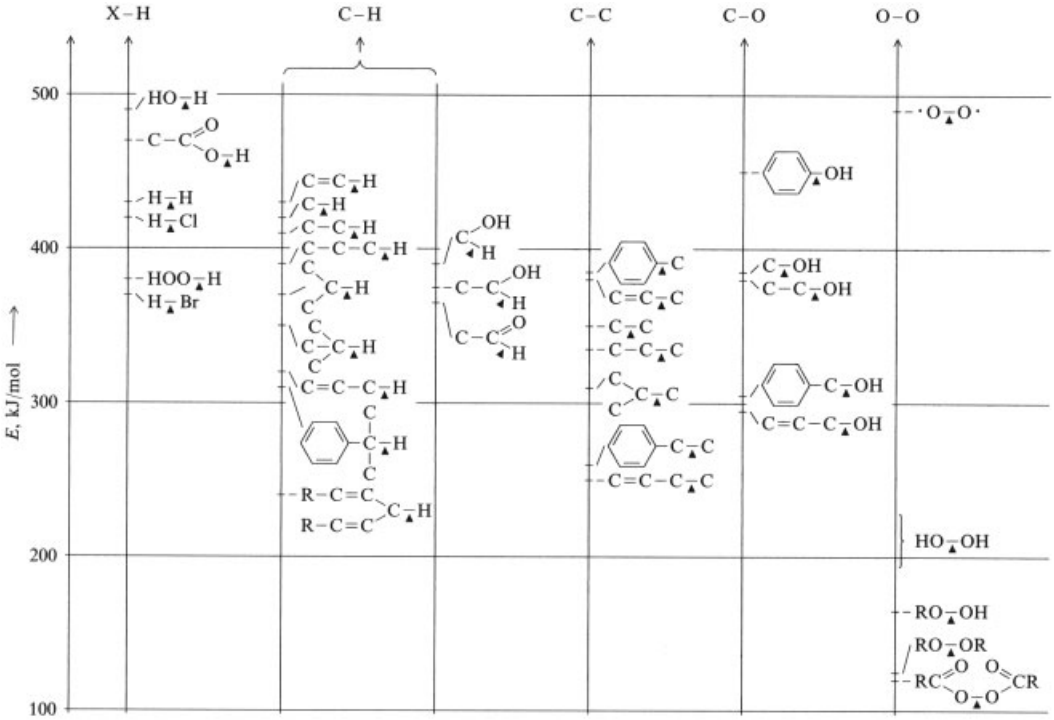

Figure 6. Dissociation energies of various bonds [9, p. 16 ff.]

cause of the ageing of organic materials in the presence of oxygen even at low temperatures. It is also the mechanism of initiation of autoxidation.

Although radical chains are continuously initiated even at low temperatures, the reaction does not exceed a uniform low rate as long as more radicals are destroyed than are formed. This is the cause of the induction period, which is always exhibited by uncatalyzed autoxidation reactions. In particular hydroperoxides accumulate during this induction period as active intermediates, which are formed in an unbranched chain reaction according to Reactions (19^r) and (20^r).

Dangerous peroxide build-up can occur in olefin-containing liquids, since peroxy radicals also add easily to double bonds according to Reaction (21^r). Because oxygen generally reacts much faster with the alkylperoxyalkyl radicals formed than the olefin, copolymeric peroxides are formed in this way. Because of the missing OH group, these are considerably more stable than hydroperoxides and can therefore accumulate in large quantities. At higher temperatures they can decompose explosively. Because the common reagents for peroxide detection are mostly sensitive to hydroperoxide, these polyperoxides often remain undetected. For more details, see [24–27].

Hydroperoxides also accumulate in paraffins, glycol ethers, and other nonvolatile organic compounds on exposure to air. These can initiate rapidly accelerating oxidation reactions even at moderate temperatures. Cleaning cloths soaked with linseed oil can even ignite at room temperature. The self-ignition temperature of paraffins, surfactants, glycol ethers, and similar substances drops below 100 °C if, for example, insulating material with a large surface area, such as rock wool which has been moistened with them has been lying for a few days in the air.

The induction period can be shortened or completely eliminated by the addition of peroxides or other radical-forming reagents, or through irradiation.

Table 3. Oxidizability, absolute chain propagation rate k_p, and chain propagation rate relative to cumene $(k_p)_{rel}$ for various organic compounds at 30 °C , [4, p. 19], [5, p. 35]

Compound	$(k_p/\sqrt{2k_t})\cdot 10^3$, $mol^{-0.5}\,s^{-0.5}$	k_p, $L\,mol^{-1}\,s^{-1}$	$(k_p)_{rel}$
n-Nonane			0.19
n-Dodecane			0.32
Cyclooctane			0.39
Cyclododecane			0.55
1-Octene	0.062		
		1.0	
Cyclohexene	2.3		10
		5.4	
2,3-Dimethyl-2-butene	3.2		20
		2.6	
Toluene	0.014		0.28
		0.24	
p-Xylene	0.049		0.73
		0.84	
Cumene	1.50		1.00
		0.18	
Tetralin	2.30		4.0
		6.4	
Benzyl alcohol	0.89		
		2.4	
Benzaldehyde	290	33 000	

Termination; Inhibition. Radical chain reactions can be slowed down or stopped by reducing the concentration of active intermediates, for instance by enhancing radical recombination reactions.

In homolytic gas- and liquid-phase oxidations, solid surfaces are the most effective radical scavengers. Fillings or packings made from thin tubes (capillaries) or Raschig rings are used overall in industry as flame and explosion barriers, and certain inorganic dusts (e.g., ammonium phosphate) are among the most effective fireextinguishing agents.

Organic antioxidants (e.g., phenols, thiols, amines) act in different ways. Phenols, for instance, are transformed by active radicals into low energy radicals, which subsequently add other active radicals forming stable products. Thiols are oxidized by hydroperoxides to sulfoxides whereby the hydroperoxides are reduced to alcohols. In these reactions the organic antioxidants are, however, gradually consumed. Once they have been consumed, the autoxidation continues at the rate it would have in the unstabilized substance (see Fig. 7); for more details, see Section 4.4.3. An introduction to organic stabilizers and antioxidants can be found in [3]; fire and explosion protection are described in [23], [28–36].

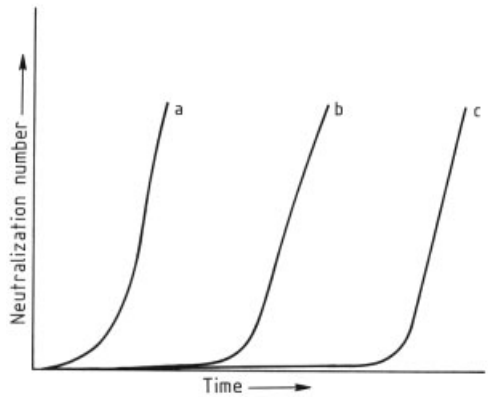

Figure 7. Typical autoxidation curves in the absence and presence of antioxidants [3, p. 4]
a) Absence of antioxidant; b) *x* parts antioxidant; c) 2 *x* parts antioxidant

Rise in the Reaction Rate; Final Conversion. As soon as the peroxide concentration has exceeded a critical value – which depends on the nature of the compound to be oxidized, the state of aggregation, the presence or absence of catalysts, and the temperature – the overall rate of radical formation by peroxide decomposition (Reactions 22^r to 24^r) begins to exceed that of radical destruction (Reactions 25^r to 28^r). From this point, the overall reaction rate increases ex-

ponentially to a maximum, which is reached at an early stage of the overall reaction.

Afterwards the reaction rate slows down to zero. In the oxidation of hydrocarbons the reaction normally ceases long before the starting material is completely consumed. One reason for this behavior is the dilution of the reaction mixture by the end products formed, mainly water and lower carboxylic acids such as acetic acid, the latter being very resistant against oxidation. In the presence of metallic ions as catalysts, water and carboxylic acids act as catalyst poisons because they are stronger complexing agents than hydroperoxides (see Section 4.4.2). Raising the temperature can revive the reaction for a more or less short time.

High starting material conversions are, however, only rarely desired in homolytic oxidations because the product selectivity decreases rapidly with increasing conversion due to secondary reactions. Frequently, starting material conversions of 10 – 30 %, sometimes even only 1 %, are aimed for. (For more details see Section 4.2.)

Hydroperoxide Decomposition. The decomposition of hydroperoxides is the main source of radicals during the entire reaction (Scheme III, Reactions 22^r to 24^r).

The thermal decomposition of peroxides has activation energies of between 120 and 164 kJ/mol. These are the highest activation energies of the reactions shown in Scheme III.

The hydroperoxide decomposition therefore determines the temperature range for the oxidation of organic compounds. Temperature ranges for liquid-phase oxidations are as follows: aldehydes: up to 80 °C; olefins and cumene: up to 130 °C; aliphatics, alkylaromatics, alcohols, ketones, esters: 80 to 200 °C.

4.2. Kinetics of Autoxidation

Kinetic measurements, in particular with respect to individual elementary reactions, are associated with particular difficulties and uncertainties in radical chain reactions because of the great number of parallel and secondary reactions and the normally very low concentration of the reacting compounds. Nevertheless, many insights into the course of the reaction could be obtained, for example, by

1) Following the pressure rise in gas-phase oxidations
2) Investigation of the peroxide decomposition in hydrocarbons
3) Simultaneous oxidation (cooxidation) of different compounds
4) Oxidation in the presence of a tertiary hydroperoxide
5) By using specific analytical methods: ESR, chemiluminescence, and UV absorption for radicals; liquid chromatography for hydroperoxides and nonvolatile compounds; gas chromatography for the stable, volatile products etc., and many other methods.

The main problem is still, however, the quantitative determination of all components present in a reaction mixture. For example, active intermediates react further after the sample has been taken, so that the composition of the sample changes rapidly. For low-conversion reactions, which are required for the precise determination of the kinetics, the sensitivity limits of all analytical methods are quickly reached. For a review of experimental methods for the determination of the kinetics of radical chain oxidations see [6], and for analytical methods for peroxide determination see [37].

Under mild conditions (temperature < 100 °C and sufficient oxygen, i.e., for a liquid-phase reaction: O_2 partial pressure > 10 kPa, for a gas-phase oxidation: O_2 excess) the rate of oxygen conversion is only determined by Reaction (20^r) in the case of the oxidation of hydrocarbons, since Reaction (19^r) is very fast and the concentration of R · is very small. The following equation is therefore valid:

$$-\frac{\mathrm{d}\,[\mathrm{O_2}]}{\mathrm{d}t} = k_\mathrm{p}\,[\mathrm{ROO^{\cdot}}]\;[\mathrm{RH}] \tag{3}$$

Since a kinetic equilibrium (steady state) is established very rapidly after initiation of the radical formation and radical scavenging reactions, and because only Reaction (27^r) plays a role as a chain termination reaction at sufficient O_2 supply, the following is valid:

$$u_\mathrm{i} = u_\mathrm{t} = 2k_\mathrm{t}\,[\mathrm{ROO^{\cdot}}]^2 \tag{4}$$

where u_i = rate of chain initiation and u_t = rate of chain termination. Inserting Equation (4) into (3) gives:

$$-\frac{\mathrm{d}\,[\mathrm{O_2}]}{\mathrm{d}t} = \frac{k_\mathrm{p}}{\sqrt{2k_\mathrm{t}}}\,[\mathrm{RH}]\cdot\sqrt{u_\mathrm{t}} \tag{5}$$

Equation (5) is only valid for low conversions as the secondary reaction products formed at higher conversions are generally more readily oxidized than the starting hydrocarbon. The expression $k_\mathrm{p}\,/\,\sqrt{2k_\mathrm{t}}$ is referred to as the "oxidizability", since it indicates how much faster one organic compound is oxidized than another. Table 3 lists values for the oxidizability as well as k_p and relative k_p (based on cumene) values for various hydrocarbons. Under mild oxidation conditions, the value of k_p increases in the following order: *n*-alkanes < isoalkanes < alkylaromatics < alkenes < alkynes ≈ styrene derivatives ≪ aldehydes.

If, however, the oxidation rates of alkanes are compared under more severe conditions (temperature > 120 °C) where the hydroperoxide decomposition determines the reaction rate, then the order of overall oxidation rate is as given in Table 4 and Figure 8.

Table 4. Relative oxidation rate of various *n*-alkanes , [15, p. 94]

n-Alkane	Relative oxidation rate [mol O_2 (mol alkane)$^{-1}$ time^{-1}]
Ethane	0.001
Propane	0.1
Butane	0.5
Pentane	1.0
Hexane	7.5
Octane	200
Decane	1380

It can be seen that the overall oxidation rate increases sharply both with increasing chain length and with decreasing branching. The reasons for this are:

1) In the series primary → secondary → tertiary C – H bonds, the rate of formation of the hydroperoxides increases dramatically, but their rate of decomposition decreases equally drastically. The secondary hydroperoxides are formed more slowly than tertiary hydroperoxides but decompose considerably faster, so that the lower rate of formation is more than compensated for. Therefore, under these conditions, hydrocarbons with secondary C – H bonds are oxidized faster than hydrocarbons having tertiary C – H bonds.
2) The number of secondary H atoms per molecule increases with decreasing branching. All secondary C – H bonds in an alkane are attacked equally rapidly.

Tables 5, 6, 7, 8 list some literature values for the collision factor A, activation energy E_a and rate constant k for the initiation of the elementary reaction, peroxide decomposition (branching), propagation, and termination [5–9], [24] (Because autoxidation is generally initiated by the decomposition of added peroxides, peroxide decomposition frequently is designated "initiation" in the literature, which is incorrect and may lead to confusion.)

The following points may be made:

1) According to Equation (5) the rate of O_2 take-up in the initial phase of an oxidation is only proportional to the hydrocarbon concentration and is independent of the O_2 concentration under mild reaction conditions.
2) The k value of Reaction (19$^\mathrm{r}$) is estimated to be 10^7 to $> 10^9$ L mol^{-1} · s^{-1} (corresponding to 10^{-14} to $> 10^{-12}$ cm^3 molecule^{-1} · s^{-1}) at $E_\mathrm{a} \approx 0$ kJ/mol [4]. It is orders of magnitude faster than Reaction (20$^\mathrm{r}$), which therefore determines the propagation rate. Because of its high rate, Reaction (19$^\mathrm{r}$) is generally controlled in the liquid phase under industrial conditions by O_2 diffusion.
3) The number of repetitions of Reactions (19$^\mathrm{r}$) and (20$^\mathrm{r}$) is referred to as the kinetic chain length ν. It is given by the ratio of the propagation rate to the termination rate. Because of the stationary state $u_\mathrm{i} = u_\mathrm{t}$, it follows that:

$$\nu = \frac{u_\mathrm{p}}{u_\mathrm{t}} = \frac{k_\mathrm{p}}{\sqrt{2kt}\,\sqrt{u_\mathrm{i}}}\cdot[\mathrm{RH}] \tag{6}$$

With increasing rate of radical formation and radical concentration, i.e., with intensification of the reaction conditions (e.g., by raising the temperature and/or initiator concentration), the chain length becomes shorter. Under mild reaction conditions ν can exceed 1000. Under industrial conditions it is usually ca. 10.
4) The hydroperoxide decomposition has the highest activation energy. The values given are extrapolated to infinite dilution and are

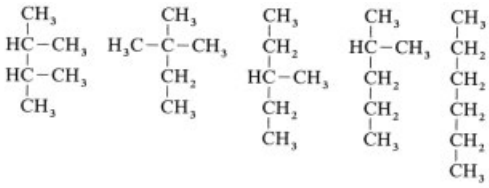

Relative oxidation velocity	1	12	60	560	1580

Figure 8. Oxidation rates of hexane isomers [mol O_2 (mol alkane)$^{-1}$ time^{-1}] relative to 2,3-dimethylbutane [15, p. 95]

Table 5. Kinetic data for the thermal decomposition of some initiators at 80 °C, [5, p. 23], [6, p. 21 ff], [24, p. 95] ∗

Initiator	Solvent	T, °C	k_d, s^{-1}	A_d, s^{-1}	$E_{a,d}$, kJ/mol
2,2′-Azodiisobutyronitrile	benzene		7.59×10^{-3}	1.6×10^{15}	129
Cyclohexyl hydroperoxide	cyclohexane	150	3.1×10^{-5}	1.2×10^{13}	142
tert-Butyl hydroperoxide	*n*-octane	150	0.90×10^{-5}	1.2×10^{15}	163
	water			4.0×10^{12}	142
Di-*tert*-butyl peroxide	benzene		9.55×10^{-6}	1.2×10^{14}	142
Cumene hydroperoxide	cumene	150	1.24×10^{-5}	7.2×10^{8}	111
	mineral oil	150	1.35×10^{-4}	1.3×10^{11}	121
tert-Butyl peroxybenzoate	ethyl benzene		9.55×10^{-5}	3.8×10^{14}	139
	n-undecane		2.34×10^{-6}	7.1×10^{16}	157
Benzoyl peroxide	benzene	79.8	3.3×10^{-5}		139
Dibenzoyl peroxide	benzene		5.75×10^{-4}	1.2×10^{13}	120
	n-octane		2.63×10^{-3}	3.5×10^{14}	121
Tetralyl hydroperoxide	mineral oil	150	1.35×10^{-4}	1.3×10^{11}	121
9-Decalyl hydroperoxide	decalin	150	9.9×10^{-5}	8.5×10^{13}	134

∗ The values are only valid for highly dilute solutions. Only under these conditions is the decomposition unimolecular.

Table 6. Preexponential factors and activation energies for the initiation reactions (16^r) and (17^r)^a [5–9]

Compound	Reaction mechanism^b	T, °C	$A_{i,1}$, L mol$^{-1}\cdot$ s^{-1}	$E_{a,i,1}$, kJ/mol	$A_{i,2}$, L^{2}mol$^{-2}\cdot$ s^{-1}	$E_{a,i,2}$, kJ/mol
Tetralin	17^r	130			3.4×10^{3}	86.7
Cyclohexane	16^r		8×10^{12}	167.5		
Cyclohexanol	17^r				8.3	67.0
Cyclohexanone	17^r				2.1×10^{2}	73.3
Cumene	16^r	120	3.5×10^{9}	113.0		
o-Xylene	16^r		2×10^{8}	129.8		
Ethanol	16^r		6×10^{7}	125.6		
Benzaldehyde	17^r	5			2.5×10^{3}	48.5

^{a} (16^r): $R{-}H + O_2 \longrightarrow R\cdot + \cdot OOH$ | $u_{i,16}$, $k_{i,16}$, etc.
(17^r): $R{-}H + O_2 + H{-}R \longrightarrow R\cdot + HOOH + \cdot R$ | $u_{i,17}$, $k_{i,17}$, etc.
$u_{i,16} = k_{i,16}\cdot[RH][O_2]$; $u_{i,17} = k_{i,17}\cdot[RH]^2\cdot[O_2]$;
$k_{i,16} = A_{i,16}\cdot\exp(-E_{a,i,16}/RT)$; $k_{i,17} = A_{i,17}\cdot\exp(-E_{a,i,17}/RT)$.
Equation (17^r) is probably an oversimplification because there is experimental evidence that this reaction proceeds via two or more steps involving a complex of the type [RH...O_2] [6, p. 20 ff.].
^{b} If RH consumption is proportional to the reaction rate u_i, the reaction mechanism (16^r) is valid. With a quadratic relationship between RH consumption and u_i, the reaction mechanism (17^r) is valid. For more details, see [6, p. 18 ff.].

Table 7. Rate constants, preexponential factors, and activation energies for the chain-propagation reaction (20^r)[a]

Compound	T, °C	k_p, L mol^{-1} s^{-1}	A_p, L mol^{-1} s^{-1}	$E_{a,p}$, kJ/mol
Tetralin	25	13.3[b]	2.5×10^4	18.8
Tetralin	30	1.6[b]		
Cyclohexane	60	0.53		
Cyclohexane			1.5×10^6	35.6
Cyclohexane	75	0.34		57.8
Cyclohexanol	40	0.70		
Cumene	30	0.18		
Cumene	65	0.50	1.1×10^4	28.1
Benzaldehyde	5	1900	4.8×10^4	7.5
Benzaldehyde	30	33 000		

[a] (20^r): R−H + ROO· ⟶ R· + ROOH; $u_p = k_p \cdot [ROO\cdot] \cdot [RH]$; $k_p = A_p \cdot \exp(-E_{a,p}/RT)$.
[b] Data from different authors sometimes differ by more than one order of magnitude.

Table 8. Rate constants, preexponential factors and activation energies for the chain termination reaction (27^r)[a]

Compound	Solvent	T, °C	$2k_t$, L mol^{-1} s^{-1}	$2A_t$, L mol^{-1} s^{-1}	$E_{a,t}$, kJ/mol
HOO·		30	7.60×10^5		
Primary $RCH_2OO\cdot$		30	10^7		
Tetralin[b]	Tetralin	60	2.14×10^7	4×10^7	1.7
Cyclohexane[b]	Cyclohexane	60	5.25×10^6	5.9×10^7	6.7
Ethylbenzene[b]	Ethylbenzene	60	1.90×10^7	1.9×10^7	0.0
Cumene[b]	Cumene	60	1.12×10^5	1.3×10^{10}	32.2
Isobutane[b]	Isobutane	60	1.23×10^4	1.6×10^9	32.6

[a] (27^r): ROO· + ROO· ⟶ ROOR + O_2; $u_t = k_t \cdot [ROO\cdot]^2$; $k_t = A_t \cdot \exp(-E_{a,t}/RT)$.
[b] The compound name stands for the secondary or tertiary peroxy radical, respectively.

therefore valid for the uncatalyzed, unimolecular, thermal decomposition. Under normal oxidation conditions (no metal catalyst, but typical concentrations of radicals, hydroperoxides, alcohols, acids, and other substances with an accelerating effect), the activation energies for the hydroperoxide decomposition are considerably lower, especially in the liquid phase. Nevertheless, they are still higher than those of the other reactions. The hydroperoxide decomposition therefore determines the temperature range of the oxidation.

5) With progressing oxidation an increasing number of intermediates are formed (aldehydes, alcohols, ketones) that are more readily oxidized than the starting hydrocarbon. These products form, on the one hand, other hydroperoxides and peroxy radicals, some more labile, some more stable, and on the other hand some of these products and their radicals catalyze the decomposition of the hydroperoxides.

This has several consequences:

a) The reaction rate increases until it reaches its maximum with the appearance of aldehydes.
b) The O_2 uptake is accelerated to a greater extent than the consumption of the starting material owing to formation of the more readily oxidizable intermediates.
c) The products formed from these intermediates – principally lower carboxylic acids – are more stable to oxidation than the starting hydrocarbons and, in the case of metal-catalyzed autoxidations inhibit the peroxide decomposition with increasing concentration. The rate of radical formation thus decreases gradually to zero and the reaction stops before the starting material is completely consumed.

6) Aldehydes are particularly readily oxidized (see Sections 4.1 and 4.4.1).

4.3. Differences between Liquid- and Gas-Phase Autoxidations

In principle, autoxidations proceed with the same mechanism in the liquid and gas phases. However, to achieve comparable reaction rates, the temperature in the gas phase must be 100 – 200 °C higher than in the liquid phase because of the weaker interactions between the molecules (in the gas phase nearly no hydrogen bonds are formed, no metallic catalysts are present, etc.).

Above 250 °C paraffins increasingly react with oxygen in a first step to olefins and hydrogen peroxide. Hydrogen peroxide is far more stable than alkylhydroperoxides against splitting of the O – O bond. Additionally, above 250 °C peroxy radicals increasingly do not stabilize to form hydroperoxides, but isomerize and/or split directly to products such as aldehydes plus RO•. But these reactions require higher activation energies than decomposition of hydroperoxides. Therefore, in the region 350 – 450 °C, depending on the hydrocarbon, the reaction velocity decreases with increasing temperature (region of negative temperature coefficient). At higher temperatures the reaction rate rises again with temperature.

The effect of material and heat transport is of considerable importance for the different pathways followed by homolytic oxidations in the liquid or gas phase.

In the liquid phase, heat transfer is good and the oxidation rate is mostly limited by oxygen transport from the gas phase. In the gas phase mass transfer does not limit the oxidation rate but heat transfer is poor due to the low mass of the gas. The oxidation reaction can therefore readily turn into a thermal explosion due to insufficient heat removal.

Table 9. Relative rate of Reaction (20^r)∗ as a function of the nature of the attacked C – H bond

R–H	C–H	prim.	sec.	tert.
2-Methylpentane	u_{rel}	1	30	300

More details on gas phase oxidation are given in Section 4.5. A detailed account of the differences between gas- and liquid-phase oxidations can be found in, for example [6, pp. 442 – 495].

4.4. Homolytic Oxidation in the Liquid Phase

4.4.1. Secondary Reactions of Radicals, Peroxides, and other Intermediates

It is only possible here to go briefly into the secondary reactions of the intermediates and radicals that are important for the understanding of radical chain oxidations by means of selected examples. For further information see [37], [54].

Reactions of Alkyl and Aralkyl Radicals. The most important reactions of these radicals are:

1) Addition of O_2 (Reaction 19^r).
2) Recombination with other radicals (termination Reactions 25^r, 26^r, 28^r).
3) Addition to C=C double bonds.
4) Isomerizations (Reactions 30^r to 33^r).

Because Reaction (19^r) is so rapid, even a low O_2 concentration (O_2 partial pressure in the gas phase 25 kPa) is sufficient to make the concentration of hydrocarbon radicals so small that they can practically no longer participate in recombination reactions.

In industry, however, oxidation reactions are mostly run with O_2 off-gas concentrations near to zero. The upper part of the bubble column is thus deficient in O_2, so that the concentration of hydrocarbon radicals increases, and radical recombination reactions can occur in this region. Thus in practically all industrial autoxidations radical trapping products formed according to Reactions (25^r) to (28^r) are found.

Table 10. Relative rate of Reaction (20^r)∗ as a function of the substituent X at the same carbon atom bearing the attacked C – H bond

X–R–H	X:	H	OOH	OH
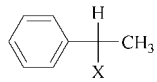	u_{rel}:	1	9.5	12
Ethylbenzene-X [(1-X,1-phenyl)ethane]				
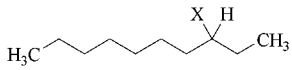	u_{rel}:	1	13	6.3
n-decyl-X				

∗ Reaction (20^r) R – H + R'OO• ⟶ R• + R'OOH

The addition of alkyl radicals to C=C double bonds occurs in olefin oxidations and acts as an

initiation reaction for radical polymerization in the oxygen-deficient regions of the bubble column.

Radicals that are stabilized by π-electron systems isomerize very readily; examples include Reactions (30^r) to (33^r):

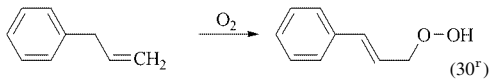

(30^r)

Yield 100 %;
no polymeric peroxide

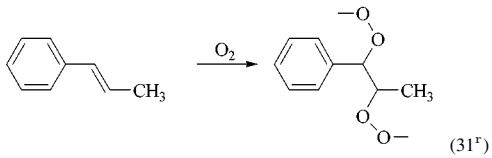

(31^r)

Yield 100 % polyperoxide;
no hydroperoxide

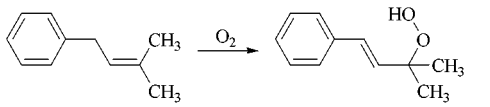

(32^r)

Yield 100 % hydroperoxide

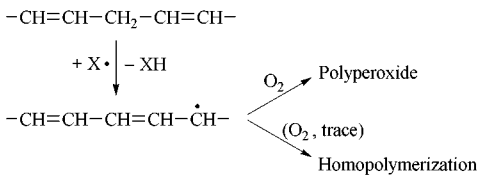

(33^r)

Reaction (33^r) involves a doubly activated CH_2 group as is present in, for example, linolenic acid derivatives. Double bond systems of this type isomerize almost completely during oxidation to give conjugated double bonds. With an adequate O_2 supply they react – like all olefins – to form copolymeric peroxides. With a deficiency of O_2 they react to give homopolymers (principle of self-drying oil paints).

Reactions of Peroxy Radicals. The most important reaction of the peroxy radical is H abstraction (Reaction 20^r). The relative rates of attack of R–OO · on primary, secondary, and tertiary C–H bonds and on C–H bonds which have another substituent (OH, OOH) on the same C atom are given in Tables 9 and 10. For example, alkylperoxy radicals react with the primary, secondary, and tertiary C–H bonds in 2-methylpentane in the ratio 1/30/300.

Because of their comparatively low reactivity the ROO · radicals can accumulate in detectable concentrations (unlike, for example, HO ·, R ·, RO ·). In the oxidation of cumene, radical concentrations of up to 8×10^{-5} mol/L, corresponding to 5×10^{16} molecules/cm^3, were detected by ESR. Peroxy radicals therefore frequently recombine, for example by Reactions (34^r, Russel reaction) and (35^r).

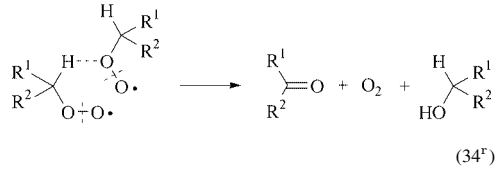

(34^r)

Russel mechanism (for primary and secondary ROO ·)

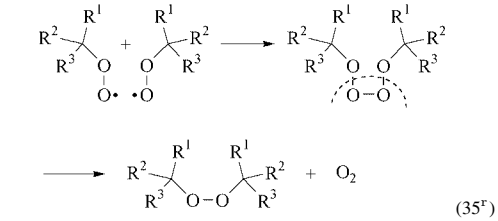

(35^r)

Lacking H atoms at the C-atom that binds the peroxy group, tertiary ROO · radicals cannot be stabilized in the same simple way as primary and secondary peroxy radicals (see Reaction 34^r). They generally recombine to give tetroxides, which subsequently decompose forming very stable di-*tert*-alkyl peroxides (Reaction 35^r) or *tert*-alkoxy radicals. The latter may either form tertiary alcohols or decompose further with cleavage of the carbon framework to form ketones, aldehydes, alkanes, or other decomposition products:

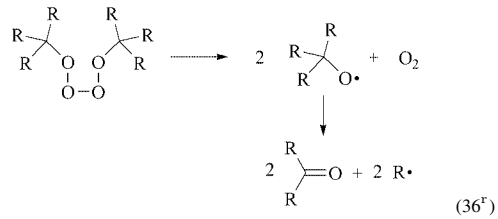

(36^r)

Because reactions of type (36^r) have high activation energies, the overall reaction of the recombination of tertiary ROO · radicals proceeds in general with high activation energy (see Table 8).

Rate constants and activation energies for the recombination of primary, secondary, tertiary, and allylic ROO · radicals are listed in Table 8. The oxidation rate is inversely proportional to these chain termination rate constants (see Eq. 5). Therefore, under mild conditions (ca. 100 °C) autoxidations of primary and secondary C – H bonds are slower than those of tertiary C – H bonds, not only because of their higher bond energies but also because of the higher recombination rates of primary and secondary ROO · radicals.

In contrast, hydrocarbons with allylic C – H bonds such as ethylbenzene and tetralin are oxidized faster than those with tertiary C – H bonds. Nevertheless, because of the higher recombination rates of allylic ROO · radicals, addition of the corresponding hydrocarbons does not enhance, but strongly hinders the autoxidation of tertiary C – H bonds under mild conditions. For example, the rate of oxidation of cumene is reduced by two thirds by the addition of 3 % tetralin because the rapidly formed tetralyl peroxy radicals capture the cumyl peroxy radicals, which are formed more slowly. Ethylbenzene has an analogous effect.

Peroxy radicals also add readily to C=C double bonds, forming either epoxides and alkoxy radicals (Twigg mechanism), or polyperoxides and their decomposition products (epoxides, aldehydes, alcohols, and ketones) (Mayo mechanism).

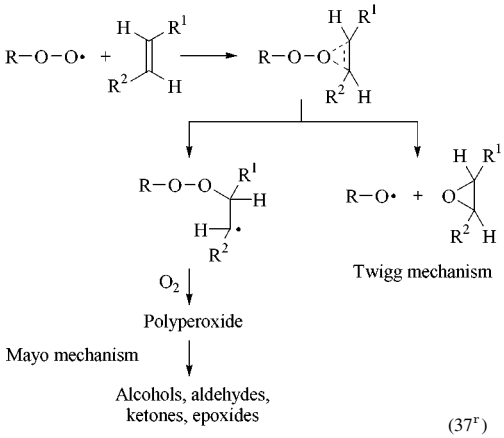

(37r)

They can also add to ketones (radical variant of the Baeyer–Villiger reaction) to give esters or lactones.

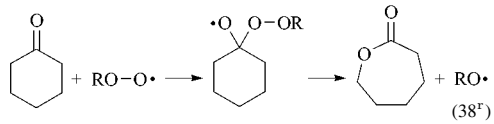

(38r)

More often, attack at the C – H bond α to the keto group occurs, followed by chain cleavage (or ring-opening) and oxidation to carboxylic acids:

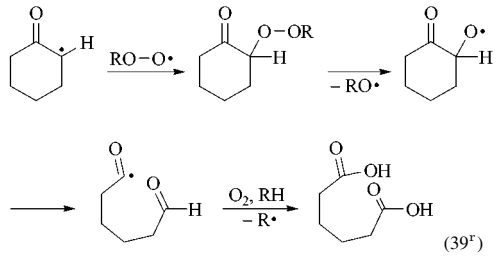

(39r)

Reactions of Hydroperoxides, Dialkylperoxides, and Peroxyacids. The stability of hydroperoxides depends on (1) the structure, (2) the surrounding medium, and (3) the active components in this medium.

For aliphatic hydroperoxides, the stability decreases in the order: tertiary > secondary > primary. Electron-withdrawing groups (–C≡N, –COOR) have a destabilizing effect, and methyl groups a stabilizing one. Hydroperoxides in which the HOO-group is in a position α to a double bond or an aromatic ring (resonance stabilization) are fairly stable. Because C – H bonds in such α-positions are weakened due to the resonance stabilization of the incipient radical, these types of hydroperoxide are formed at lower temperatures than others, and can be obtained in high yields (e.g., cumene hydroperoxide).

The homolytic decomposition of hydroperoxides is catalyzed by compounds that form hydrogen bonds, i.e., by the hydroperoxides themselves and by alcohols and carboxylic acids (Reactions 23r, 24r, 40r). The reason is that the leaving group H_2O is preformed by these O· · ·H bridges. Table 5 lists kinetic data for the monomolecular decomposition of peroxide in highly dilute solution (see Reaction 22r). In hydrocarbons the decomposition rate depends strongly on the C–H dissociation energy of the

solvent: the lower the C–H dissociation energy, the faster the decomposition (catalysis by metals, see Section 4.4.2).

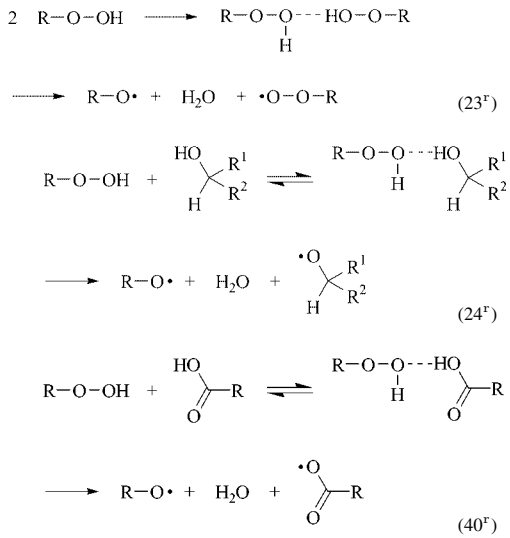

Radicals also catalyze hydroperoxide decomposition. Accordingly, hydroperoxides can be stabilized by the addition of radical scavengers.

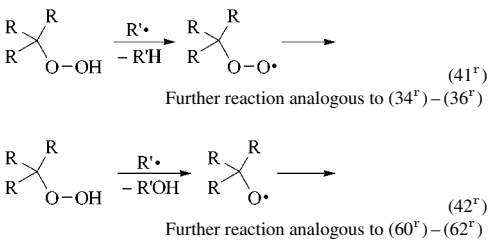

Because of their acidic character, peroxyacids catalyze their own decomposition much more strongly than the alkyl hydroperoxides:

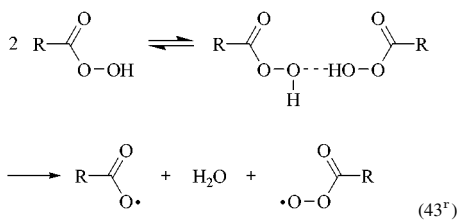

The acyloxy radicals formed are more reactive than the alkyl peroxy radicals. They can react in different ways:

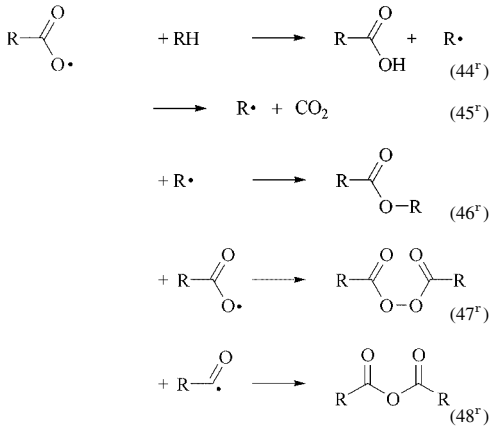

Normally, Reaction (44ʳ) is the preferred one. But if there are not sufficient molecules with active C–H bonds available, that can be attacked by the acyloxy radical in order to stabilize the radical as a carboxylic acid, Reaction (45ʳ) becomes important (formation of CO_2). Reaction (45ʳ) seems to be important also in the case of aromatic acyloxy radicals. There the odd electron is attracted by the π-electron system of the aromatic ring which leads to radical isomerization, and subsequently splitting into a phenyl radical and CO_2. Reactions (46ʳ) to (48ʳ) are of minor importance. Only in the presence of copper ions, a modified version of Reaction (48ʳ) gains importance (see Reaction 87ʳ).

Cleavage by Mineral Acids. Mineral acids protonate peroxides and thus induce their decomposition. The best known example is the Hock rearrangement of cumene hydroperoxide to give phenol and acetone:

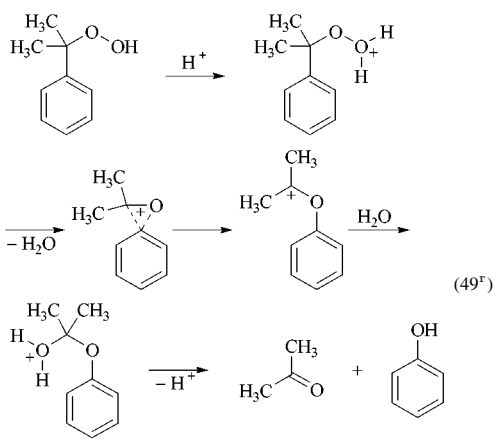

Another industrially important example is the oxidation of *p*-xylene to terephthalic acid, catalyzed by $CoBr_2$ in acetic acid (see Reaction 92[r]). Oxidations of alkylaromatics in which mineral acids such as HBr can be formed only proceed at a satisfactory rate if it is ensured that the concentration of the mineral acid remains low. Otherwise, phenol, which acts as an oxidation inhibitor, may be formed by acidic hydroperoxide cleavage.

Reactions with Bases. Bases react with primary and secondary hydroperoxides as follows (PRITZKOW believes that these findings are artefacts, caused by very small impurities of metal salts in the bases):

$$ROOH + B^- \rightarrow ROO^- + B\text{–}H \qquad (50^r)$$

Chain reaction:

$$ROO^- + ROOH \rightarrow RO^- + O_2 + ROH \qquad (51^r)$$

$$RO^- + ROOH \rightarrow ROH + ROO^- \qquad (52^r)$$

In secondary dialkyl peroxides, the C–H bond at the same atom as the peroxy group is attacked:

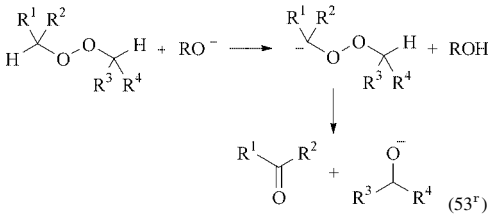

(53[r])

Bases have practically no effect on tertiary dialkylperoxides.

Reactions with Olefins. In the absence of epoxidation catalysts (e.g., Mo complexes) hydroperoxides predominantly react with olefins according to:

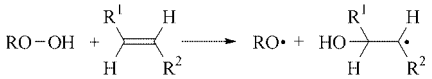

(54[r])

whereas peroxyacids mostly give epoxides very selectively in a nonradical reaction (Prilezhaev reaction; see Section 5.1)

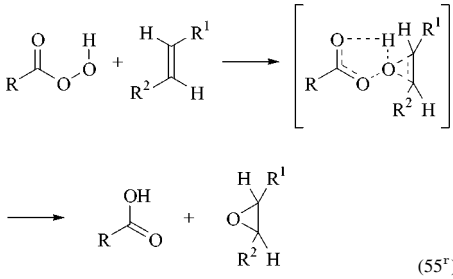

(55[r])

Reactions with Ketones. Hydroperoxides readily add to ketones. The resulting perketals are labile and decompose readily, forming a variety of products:

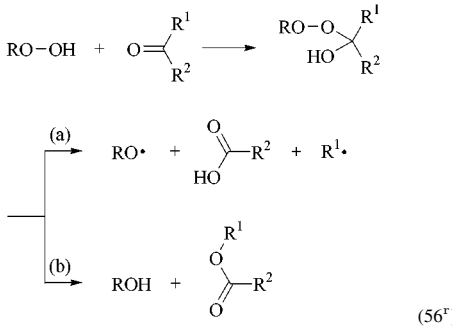

(56[r])

Peroxyacids react with ketones with C–C bond cleavage to give esters or lactones (Baeyer–Villiger reaction).

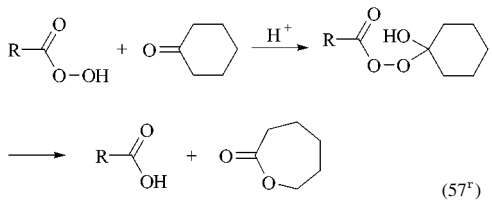

(57[r])

Reactions with Aldehydes. Peroxy acids add to aldehydes more readily than to ketones, forming the corresponding aldehyde monoperesters. Cleavage of these compounds gives the corresponding acids. This cleavage is selectively catalyzed by Mn^{2+}/Mn^{3+} ions (see Section 4.4.2) In the oxidation of aldehydes to acids almost the entire acid is formed via this pathway:

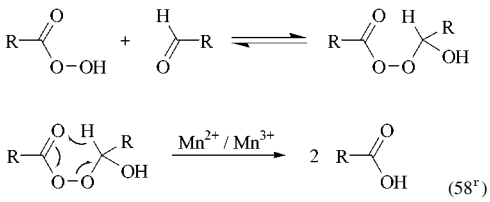

(58r)

Reactions with Radicals. Radicals preferably attack C–H bonds at the same C-atoms that are already bearing other substituents like OOH or OH groups (see Table 10):

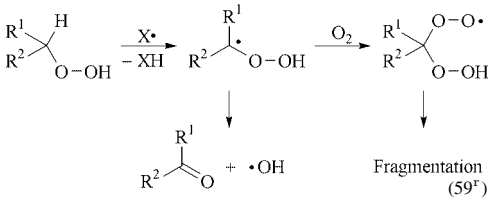

(59r)

Reactions of Alkoxy Radicals. The alkoxy radicals formed in the decomposition of hydroperoxides are very reactive (see Chap. 4.1). However, only primary alkoxy radicals appear to react predominantly by abstraction of hydrogen atoms to form alcohols. Secondary and, above all, tertiary alkoxy radicals react predominantly by C–C cleavage to give aldehydes or ketones, together with small amounts of alcohols.

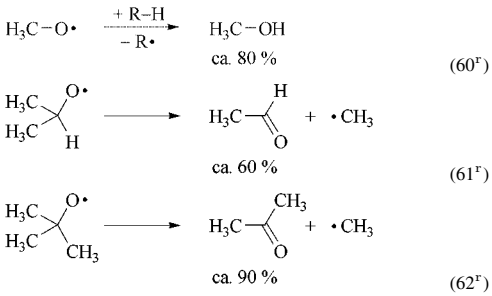

(60r)

(61r)

(62r)

The C–C cleavage reactions (Reactions 61r and 62r) are very important in autoxidation because the carbon chain degradation proceeds mostly by these mechanisms. However, in the following examples, abstraction of H atoms or addition is preferred to C–C cleavage:

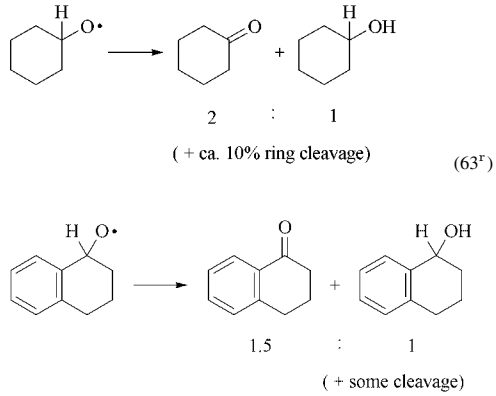

(63r)

(64r)

In olefin oxidations, alkoxy radicals also add to C=C double bonds:

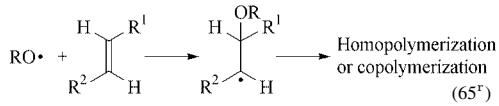

(65r)

Oxidation of Alcohols, Aldehydes, and Ketones. The oxidation of primary alcohols proceeds via intermediate aldehydes and acetals to give the corresponding acids and esters.

In the autoxidation of secondary alcohols ketones and H_2O_2 are formed:

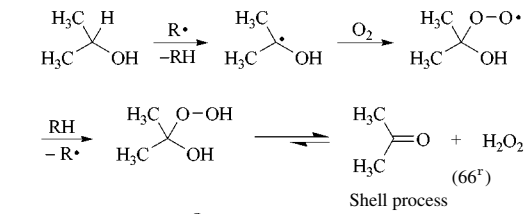

(66r)

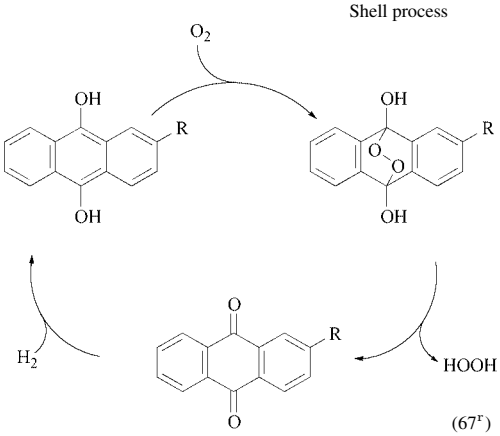

(67r)

H_2O_2 process from BASF

The oxidation of cyclohexanol gives cyclohexanone and H_2O_2, which are in equilibrium with the intermediate cyclohexylperoxyacetal:

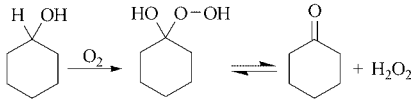

(68r)

The ketones formed are further oxidized via the α-hydroperoxyketones:

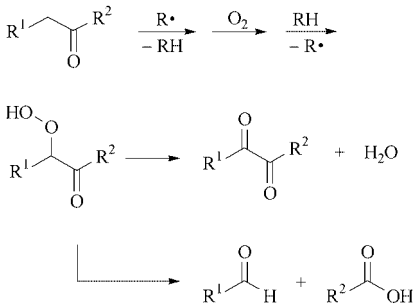

(69r)

The addition of peroxy radicals and hydroperoxides to ketones has already been mentioned (Reactions 38r, 56r, and 57r).

Aldehydes form addition products with polar compounds in the liquid phase. In the addition product the C–H bond of the aldehyde group becomes strongly activated by two polar groups (see Section 4.1, and Table 10, where the activation of the C–H bond by one polar substituent is shown):

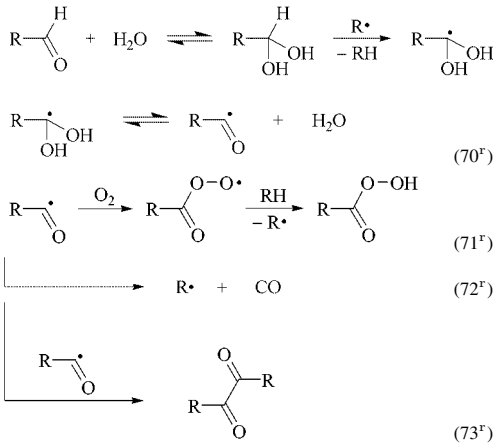

(70r)

(71r)

(72r)

(73r)

The acyl radicals formed initially can decompose forming an alkyl radical and CO if there is an oxygen deficiency (Reaction 72r) or can also recombine to give diketones (Reaction 73r: synthesis of diacetyl from acetaldehyde).

Carbon monoxide and carbon dioxide are formed in autoxidation almost exclusively by decomposition of acyl and acyloxy radicals (Reactions 72r and 45r).

The presence of CO in the reaction waste gas (at an order of magnitude of ≤ 0.1 vol %) is thus a sensitive indicator of oxygen deficiency in autoxidations.

Cooxidation is the simultaneous oxidation of several substances. The cooxidation of aldehydes and olefins to produce epoxides (and acids) is particularly well known:

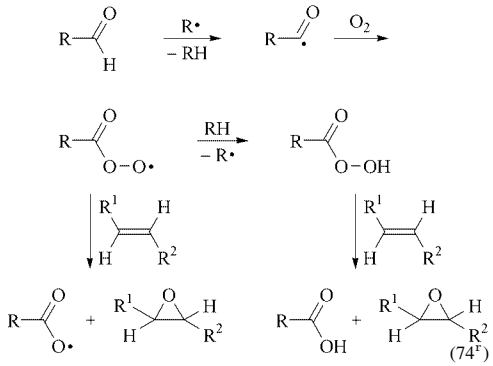

(74r)

as is the cooxidation of *p*-xylene and paraldehyde or methyl ethyl ketone in acetic acid to give terephthalic acid and acetic acid (Toray, Olin-Matheson).

As is apparent from the preceding sections, practically every autoxidation becomes a cooxidation soon after initiation. Planned cooxidation has achieved great importance, e.g., for elucidating reaction mechanisms and determining the relative reactivity of homologous species.

4.4.2. Metal Catalysis in Homolytic Liquid-Phase Oxidation

Homolytic liquid-phase oxidation is influenced by many metal ions, but only metals that can occur in two valence states of approximately equal stability in the oxidation medium can act catalytically. These are above all Co^{2+}/Co^{3+} and

Mn^{2+}/Mn^{3+}. Ionic pairs such as Cu^{+}/Cu^{2+} and Fe^{2+}/Fe^{3+} are important only in special cases.

The catalytic activity of these metals is based on the acceleration of peroxide decomposition by lowering the activation energy:

Thermal activation ($E_a = 90 - 170$ kJ/mol):

$$ROOH + ROOH \longrightarrow ROO\bullet + H_2O + RO\bullet \qquad (23^r)$$

Metal catalysis ($E_a = 40 - 50$ kJ/mol):

$$ROOH + Co^{2+} \longrightarrow RO\bullet + Co^{3+} + OH^- \qquad (75^r)$$

$$ROOH + Co^{3+} \rightleftharpoons ROO\bullet + Co^{2+} + H^+ \qquad (76^r)$$

$$2\ ROOH \xrightarrow{(Co^{2+}/Co^{3+})} ROO\bullet + H_2O + RO\bullet \qquad (77^r) = (75^r) + (76^r)$$

Because alkyl hydroperoxides are strong oxidants, Reaction (76^r) is only important in the decomposition of pure hydroperoxides, i.e., during the initiation step. After initiation the metal ions in the higher valency state are preferably reduced by aldehydes, because aldehydes are strongly reducing compounds:

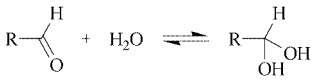

(78^r)

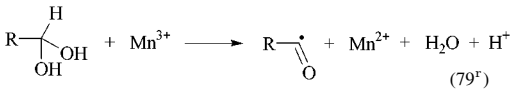

(79^r)

Cobalt is the most important catalyst for homolytic oxidations, because only Co^{3+} has an oxidation potential high enough to oxidize hydrocarbons to cation radicals which decompose instantaneously to give radicals and protons:

$$RH + Co^{3+} \rightarrow RH\cdot^{+} + Co^{2+} \qquad (80^r)$$

$$RH\cdot^{+} \rightarrow R\cdot + H^{+} \qquad (81^r)$$

Therefore, in the decomposition of hydroperoxides (Reaction 75^r) two active species ($RO\cdot$ and Co^{3+}) are formed only with cobalt as catalyst. Formation of these species results in branching of the reaction chain and an increase in the reaction rate.

Mn^{2+}/Mn^{3+} catalyze the decomposition of hydroperoxides in the same way as cobalt ions. But since Mn^{3+} is unable to oxidize hydrocarbons, its catalytic activity is considerably lower than that of Co^{3+}. However, manganese is of great importance in the homolytic oxidation of hydrocarbons to acids, because it selectively catalyzes the splitting of the addition product of peracids and aldehydes (the aldehyde monoperester) into two molecules of acid (see Reaction 58^r).

Prior to decomposition, the hydroperoxides have to form complex compounds with the metal ions. Therefore, the rates of these redox reactions depend strongly on the medium. In polar media, such as acetic acid, the redox reactions can take place via extremely fast ligand transfer.

In nonpolar media, such as hydrocarbons, only electron transfer is possible, which is orders of magnitude slower than ligand transfer but still considerably faster than thermal decomposition.

In the series: alkyl hydroperoxide $<$ alcohol $<$ water $\approx$ acid, the hydroperoxides are the weakest complexing agents. Therefore alcohols, but in particular water and acids, act as catalyst poisons by competing with the hydroperoxides for the metal ion complex sphere. If these compounds are present in appreciable amounts – that is the case in nearly every normal oxidation – $Co^{2+/3+}$ has to be added in fairly high concentrations, e.g., in polar media 0.05 – 1 wt % and in nonpolar media 1 – 500 ppm (ca. $10^{-4} - 10^{-5}$ mol/L).

Because water and acids act as catalyst poisons, water and acids formed during the reaction should be stripped off to as low concentrations as possible. Acids, which cannot be stripped off, will slow down the oxidation rate long before the hydrocarbon starting material is consumed. In the special case of oxidation of *p*-xylene to DMT, the acid intermediate is esterified in order to achieve conversion of the second methyl group as well (see below: Witten – Imhausen – Katzschmann process, now Hüls process).

Peroxy radicals are strong oxidants. Therefore they oxidize Co^{2+} according to:

$$ROO\cdot + Co^{2+} \rightleftharpoons ROO^{-}\ Co^{3+} \qquad (82^r)$$

In the oxidation of *p*-xylene to TPA the peroxy anion formed further reacts to form the corresponding aldehyde:

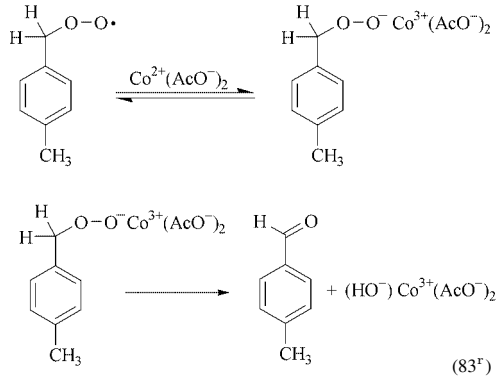

(83[r])

(AcO^- = acetate anion)

In nonpolar media the initiation of an oxidation by hydroperoxides can come to a halt if the concentration of the Co ions becomes greater than the hydroperoxide concentration (catalyst-inhibitor conversion). Because all hydroperoxide is tied up in the complex sphere of the Co ions, excessive Co^{2+} competes effectively with the substrate RH for the alkylperoxy radicals according to Reaction (82[r]) and hinders chain propagation by:

$$ROO\cdot + RH \longrightarrow ROOH + R\cdot \qquad (20^r)$$

In polar media such as acetic acid these phenomenon does not appear because sufficient complex ligands are available.

The decomposition of dialkyl peroxides is accelerated by Fe^{3+}, Cu^{+}, and complex compounds of Mo^{6+}, W^{6+}, V^{6+} and Ti^{4+} (see Section 5.1). $Co^{2+/3+}$ and $Mn^{2+/3+}$ decompose dialkyl peroxides only under conditions which are not relevant to autoxidation.

Peroxyacids are decomposed analogously to alkyl hydroperoxides:

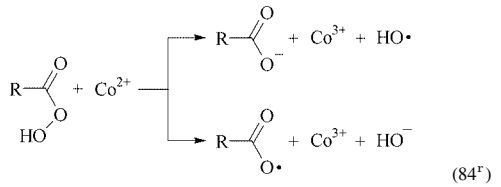

(84[r])

Because peroxyacids are stronger oxidizing agents and much weaker reducing agents than hydroperoxides, reduction of Co^{3+} ions only occurs with aldehydes according to Reactions (78[r] and 79[r]).

Effect of the Catalyst and the Medium. The metal ions have a major influence on the course of the reaction, as the following examples show.

Oxidation of cyclohexane:

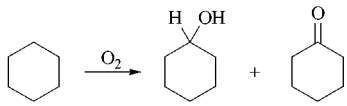

Without catalyst	1	:	2
With Co or Mn catalyst	3	:	2

(85[r])

The oxidation of acetaldehyde can be controlled to give considerable amounts of diacetyl or acetic anhydride besides acetic acid, depending on the nature and concentration of the catalyst. In both cases the production of CO, CO_2, and CH_4 is enhanced, too.

Oxidation in acetic acid with catalytic quantities of Mn gives acetic acid according to Reactions (58[r]), (70[r]), (71[r]), (78[r]), and (79[r]).

Addition of catalytic quantities of copper acetate to the $Co^{2+/3+}$ catalyst leads to formation of acetic anhydride in addition to acetic acid:

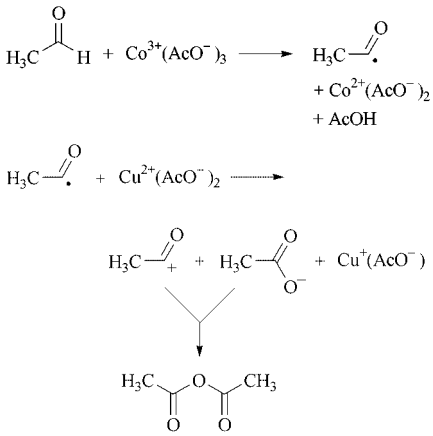

(86[r])

(87[r])

The Cu cations seem to stabilize the acyl cations and anions and the peroxyacyl radicals.

Oxidation in the presence of large quantities of manganese(III) acetate gives diacetyl, acetic acid (and CO, CO_2, CH_4):

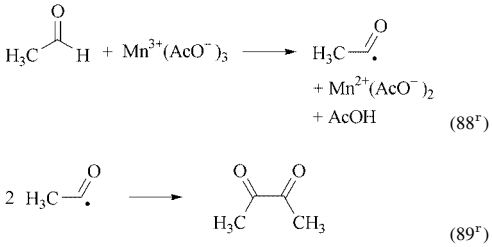

Oxidation of p-xylene: pure *p*-xylene is mainly oxidized to *p*-methylbenzoic acid in the presence of 100 – 200 ppm Co and up to 100 ppm Mn acetate in a normal radical chain reaction at 140 – 180 °C and 10 bar. The greater part of the second methyl group can only be oxidized after esterification of the *p*-methylbenzoic acid (hindrance by the acid formed) (Witten – Imhausen – Katzschmann process, now Hüls process):

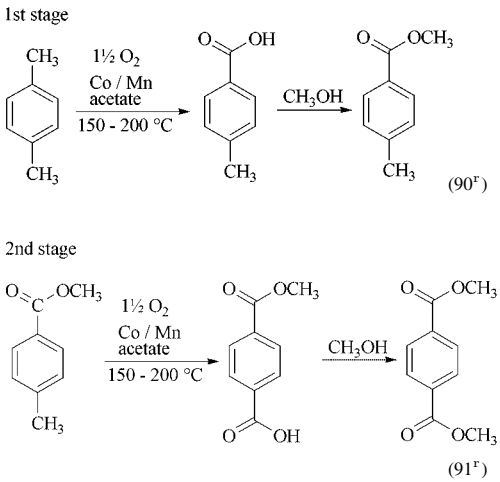

Single-stage oxidation of both methyl groups succeeds in acetic acid in the presence of 500 – 1000 ppm Co and Mn acetate and either a peroxyacid-forming cooxidizing agent such as paraldehyde or acetaldehyde (Toray, Eastman-Kodak), or with Br ions as radical sources (Mid-Century/Amoco).

Whereas the peroxyacid formed in the Toray or Eastman-Kodak process overcomes the low oxidizability of the second methyl group, which is due to the carboxyl group, in the Mid-Century/Amoco process Co^{3+} and the Br • radical formed act as abstractors for protons or H atoms, respectively:

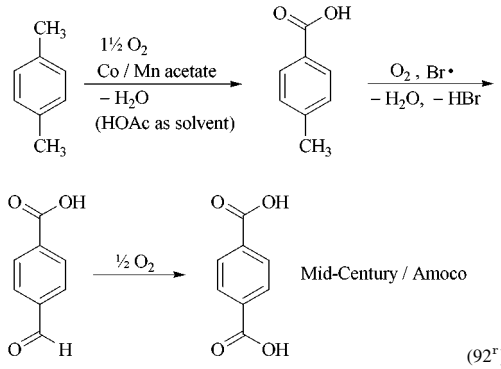

The intermediate 4-carboxybenzaldehyde is rapidly oxidized further to give the acid. Nevertheless, the removal of residual aldehyde is the major problem in the purification of the crude terephthalic acid (TPA), which crystallizes out rapidly due to its insolubility in the reaction medium and occludes small amounts of aldehyde, which are thus protected from further oxidation. Amoco solved this problem by hydrogenating the carboxybenzaldehyde with palladium as catalyst.

In the cooxidation process, only ca. 0.2 t acetic acid per tonne TPA are formed although the cooxidant agent must be present in at least a stoichiometric ratio.

In the Mid-Century/Amoco process 0.05 t acetic acid are consumed per tonne TPA.

One reason for the low yield and high loss of acetic acid is its metal-ion-catalyzed decarboxylation (Reaction 93r).

The decarboxylating activities of some metal ions are listed in Table 11. The reaction (93r) is fast above 140 °C.

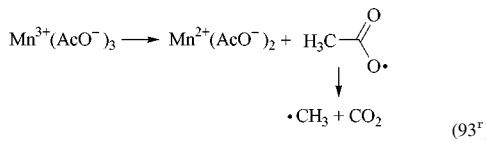

Table 11. Relative activity of various metal ions in the decarboxylation of carboxylic acids

Metal ion	Relative activity
Ce^{4+}	1
Ag^{2+}	2
Pb^{4+}	20
Co^{3+}	100
Mn^{3+}	500

4.4.3. Inhibitors of Homolytic Liquid-Phase Oxidation

Inhibitors are substances that hinder the formation of the active intermediates (radicals, hydroperoxides, catalysts) or lower their concentration by complexing, decomposing, or trapping them.

The effect of phenols and aromatic amines is particularly well known. Because of their weak O–H and N–H bonds, they readily transfer H atoms to radicals. The resulting phenoxy or arylamino radicals are extremely unreactive due to resonance stabilization and, in the case of 2,6-disubstituted phenols, because of the strong steric hindrance. The radical chain reaction is thus interrupted, as shown in Reaction (94^r):

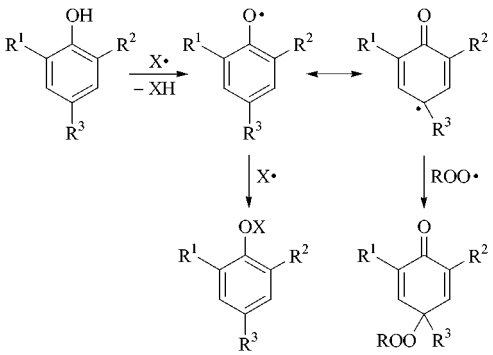

(94^r)

X = Alkyl; not alkoxy

An inhibitor molecule can thus capture two radicals. The relative efficiency of the inhibitors depends strongly on the substituents (Table 12).

Table 12. Relative inhibitor effect of trialkylphenols∗ on the oxidation of crude oil , p. 50[3]

R^1	R^2	R^3	Relative inhibitor effect ∗∗
H	H	Me	10
tert-Bu	*tert*-Bu	*tert*-Bu	36
Me	Me	*tert*-Bu	46
tert-Bu	*tert*-Bu	*sec*-Bu	80
tert-Bu	*tert*-Bu	Me	100
Me	Me	Me	118
tert-Bu	*tert*-Bu	Ethyl	125
tert-Bu	*tert*-Bu	*n*-Bu	140
Me	*tert*-Bu	Me	170

∗ 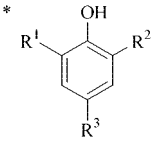

∗∗ 2,6-Di-*tert*-butyl-4-methylphenol = 100

Whereas phenols and amines mainly capture radicals, organic sulfides and thiols usually react with hydroperoxides as follows:

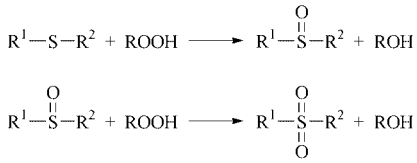

(95^r)

For this reason mixtures of organic sulfides and phenols or amines have a stronger inhibiting effect than the individual components (inhibitor synergism).

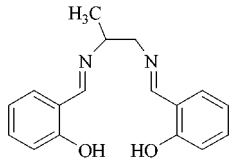

Disalicylidine-1,2-propylenediamine

The inhibiting effect of an excessively high metal concentration has already been mentioned (see Section 4.4.2). Compounds that form metal complexes can also act as inhibitors if the number of donor centers corresponds to the coordination number of the metal. For example, disalicylidene-1,2-propylenediamine is added to gasoline in concentrations of 10^{-8} mol/L because it forms stable complexes with metals,

thus preventing them from catalyzing autoxidation.

It is noteworthy that phosphates, especially di- and triphosphates, appear to inhibit radical chain oxidation. They are not only the most effective fire-extinguishing powders (in the form of ammonium hydrogen phosphate), and, as pyrophosphate ($P_2O_7^{2-}$), a component of many heterogeneous gas-phase oxidation catalysts, but are also components of important building blocks of all living cells in the nucleotides and nucleic acids. In what way they interfere with the radical chain oxidation has apparently not yet been investigated.

For an introduction to autoxidation and the effect of inhibitors and stabilizers, in particular with reference to the protection of polymers from oxidative aging, see [3].

4.4.4. Special Cases of Homolytic Oxidation

Boric-Acid-Modified Alkane Oxidation (→ Alcohols, Aliphatic, Chap. 2.3.6.). The oxidation of hydrocarbons in the presence of suspended metaboric acid, known since 1928, appears to be a borderline case between homolytic and heterolytic oxidation. Secondary alcohols are preferably formed with unchanged carbon frameworks (selectivity > 80 %), together with ketones and diols. In the presence of olefins, epoxides are also formed in considerable yields.

The way of action of the boric acid is not clear yet. Until the 1970s it was believed that boric acid reacts only with the alcohol to form boric ester, which are very resistant to further oxidation. Later on epoxides were found, that are formed during the oxidation of olefins in the presence of boric acid. This led to the assumption that the boric acid already catalyzes the decomposition of the hydroperoxides. But, because in the oxidation of paraffins the ratio of secondary alcohols to ketones is very high (nearly no Russell reaction), it seems very likely that the boric acid reacts directly with the peroxy radicals to the corresponding boric esters and hydroperoxy radicals.

After oxidation, the alcohols must be liberated from the boric esters by hydrolysis. Only one OH group of the boric acid reacts with the alcohol to give the ester.

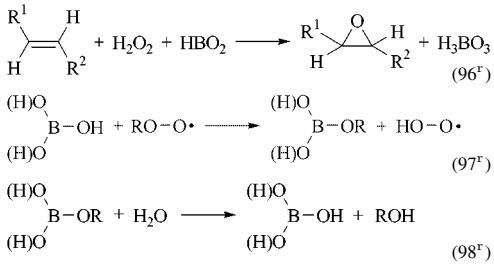

In Reactions (97r) and (98r) the protons of the boric acid depicted in parentheses do not behave like normal protons. They do not take part in esterification.

Alkyl aromatics cannot be oxidized in this way because boric acid also catalyzes the Hock cleavage to give phenols, which inhibit further oxidation shortly after initiation.

Paraffin Sulfonation (→ Surfactants, Chap. 6.2.3.). A further special case is the paraffin oxidation process in the presence of SO_2. Of several variations, two are carried out industrially.

Sulfochlorination (Reed Process). Sulfur dioxide and chlorine are introduced into the paraffin at 25 – 35 °C in the presence of UV radiation. The reaction proceeds by a radical chain mechanism with alkylsulfonyl chloride (i.e., chlorosulfonic acid) as the final product.

The kinetic chain length is 4000 – 5000. The conversion is limited to 30 % to minimize further oxidation to di- and polysulfonic acids.

Sulfoxidation (Light – Water Process). If SO_2 and O_2 (in a molar ratio of 2: 1 to 10: 1) are passed through an *n*-alkane in the presence of UV radiation at 20 – 40 °C, alkyl persulfonic acid is initially formed in a radical chain reaction. In pure paraffin, the persulfonic acid reacts further to give dark colored products. If water is added to the alkane, the reaction is slower and for each mole alkyl sulfonic acid an equivalent quantity of sulfuric acid is formed. However, the product remains colorless because the water extracts the alkylpersulfonic acid and catalyzes its reaction with SO_2 to give alkylsulfonic acid and sulfuric acid. The alkane conversion is limited to ca. 1 % to keep the undesired polysulfonation low.

Initiation

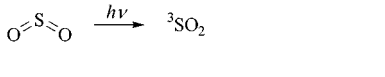

(99r)

Chain propagation

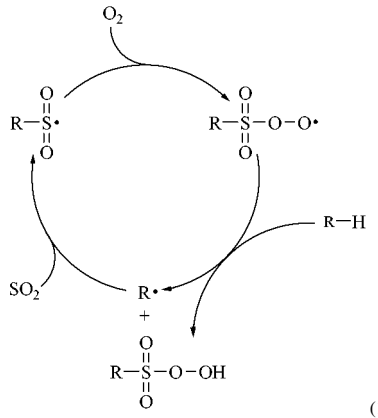

(100r)

Decomposition of peracid

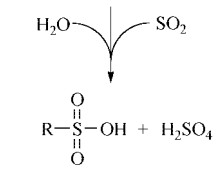

Alkane sulfonic acid
(R = C_{10} - C_{15})

(101r)

Remarkably, SO_2 adds faster to the alkyl radical than O_2. Excess of SO_2 is therefore important. The reactivity of the alkylpersulfonyl radical formed by subsequent addition of oxygen is similar to that of the alkylperoxy radical. The sulfonic acid group is thus situated predominantly at carbon atoms in the middle of the chain.

Aromatics, olefins, and branched-chain paraffins inhibit the reaction.

4.4.5. Process Technology of Homolytic Liquid-Phase Oxidation

The various starting materials (liquefied petroleum gas, low-boiling aldehydes to high-boiling, long-chain paraffins, alkylhydroanthraquinone etc.) and the widely differing products (from H_2O_2 to terephthalic acid) also require differing conditions (temperature, pressure, solvents, reactor material). Some general rules may be given, which are valid for most radical chain oxidations.

Reaction Parameters, Reactor Geometry, Operation. The aim of industrial reactions is to produce as much product as possible per unit time and reactor volume. Because the mass flow of the educts, reaction temperature, and catalyst concentration can be chosen freely within limits, these parameters are normally maximized to give high throughput and production numbers. Consequently, O_2 transport from the gas to the liquid phase becomes the rate-determining step, often at the expense of lower product selectivity. This is even worsened by the fact that for safety reasons the O_2 content of the off-gas is usually kept near to 0 vol %, resulting in O_2 deficiency in the upper part of the bubble column.

Table 13. Solubility of oxygen in organic compounds at 1 bar [38, p. 1 ff]

Solvent	T, °C	O_2 concentration, 10^{-3} mol/L	α*
Benzene	20	9.1	0.204
Benzene	60	9.6	0.216
Toluene	20	5.3	0.119
p-Xylene	26	8.3	0.118
p-Xylene, tech.	23	7.2	0.162
p-Xylene, tech.	80	8.0	0.179
p-Xylene, tech.	100	8.6	0.192
Isooctane	20	13.9	0.312
Isooctane	50	14.6	0.327
Dodecane	25	8.25	0.185
Methanol	20	10.5	0.235
Ethanol	20	6.4	0.143
Acetone	20	9.6	0.215
Cyclohexanol	116	9.5	0.213
Cyclohexanone	100	10	0.224
Benzaldehyde	5	4**	0.090
Decanal	5	4	0.090

* α=cm^3 O_2 (STP) cm^{-3} bar^{-1}.
** The determination of O_2 solubility requires a long time for equilibrium to be reached. Hence, the values for readily oxidizable substances such as aldehydes are probably too low.

To test whether an oxidation is controlled by O_2 transport, it is sufficient to increase the O_2 pressure and to see if the reaction rate subsequently rises. According to Equation (5) if there is an adequate O_2 supply the reaction rate is independent of the O_2 concentration and thus of the O_2 transport. The solubility of O_2 in organic solvents is only ca. 10^{-2} mol/L (see Table 13).

Apart from the O_2 partial pressure, the rate of O_2 transport from the gas to the liquid phase

depends only on the area of the phase boundary surface, i.e., the bubble size. Therefore, the gas must be divided into the smallest bubbles possible on entry into the reactor and these must be prevented from coalescing into larger bubbles while they rise up the liquid column.

If turbulence is too severe and there is consequent coalescence, large gas bubbles are formed and the oxygen cannot pass into the liquid phase quickly enough. Oxygen then appears in the waste gas, although the oxidation suffers from an O_2 deficiency.

The gas velocity based on the reactor cross-section should thus clearly lie below 0.02 m/s. At higher velocities turbulence arises in the bubble column which promotes coalescence of the gas bubbles.

The height of the liquid phase in the reactor above the air inlet must be adapted to the rate of O_2 take-up. It can be determined in pilot experiments (see below). Generally 4 – 6 m should be sufficient. The volume required for the necessary throughput should be obtained by increasing the cross-section of the reactor (tanks instead of columns). In industry, for historical reasons, most oxidation reactors in operation are much higher (up to 30 m and more with liquid heights of 8 to more than 20 m). The reaction pressure should be chosen in a range where boiling of liquids is avoided because the vapor formed would enlarge the gas bubbles and lower the O_2 concentration in the bubbles, thus hinder oxygen transport.

Reactors for radical chain reactions must have the smallest possible surface to volume ratio because of the undesired radical recombination that always occurs at the walls. Apart from necessary built-in devices, like sieve plates for gas redispersion and cooling tubes, packings and devices with large surface areas should not be used in the reaction section. However, in the gas space above the bubble column packings with large surface areas are recommended to suppress gas explosions.

Oxidation can be operated as a batch or continuous process. In a continuous process the residence time spectrum is broadened, as part of the fresh starting material is carried straight out again because of the generally strong back-mixing. Accordingly, part of the product remains longer in the reactor and is overoxidized. The selectivity is thus lower than in batch processes. For this reason many large-scale oxidations are still performed batchwise even today. The selectivity in continuous oxidations can generally be improved by using baffles (sieve plates, etc.), which convert the reactor into a bubble column cascade and lower back-mixing.

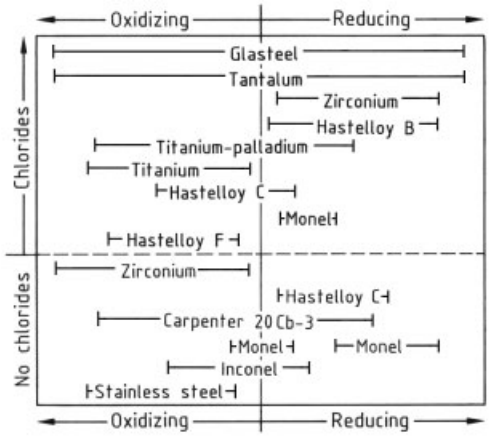

Figure 9. Corrosion resistance of various materials
The effective corrosion resistance of each material in an acid environment is that portion of the chart below each bar (with kind permission of Gulf Publishing Comp., taken from *Hydrocarbon Processing,* June 1979, p. 214).

Reactor Materials; Corrosion. The choice of reactor materials is determined by the reaction medium and conditions (temperature, pressure, etc.). Most base metals and alloys are only protected from chemical attack by a dense, tightly adhering layer of oxide (passivation). These materials (Al alloys, Ti, stainless steel) are generally stable towards oxidizing media, but are readily attacked by reducing media (see Fig. 9).

Because many heavy metal ions also have a catalytic effect, aluminum and its alloys are recommended for the oxidation part under moderate conditions of temperature and pressure. For higher temperatures and pressures, stainless steel is recommended. Stainless steel is, however, rapidly attacked by halide ions even in an oxidizing medium. Therefore, for example, for the Mid-Century process reactors coated with titanium or Hastelloy C must be used.

Copper can inhibit many oxidations even in extremely low concentrations. Because copper is contained in fine dispersions in some lubricating agents and slip additives, particular attention

should be paid to such substances in the case of initiation problems.

In all parts of a plant in which O_2 is absent, i.e., where reducing conditions prevail, special care must be taken to ensure that the reactor material and the temperature are mutually compatible. For example Material no. 4571 (V4A E; US specification: SS 316 Ti) is rapidly attacked by acetic acid in the absence of O_2 above 140 °C. In the presence of O_2 it is stable towards the acid up to and above 160 °C.

The corrosivity of organic materials increases with the acidity, reducing capacity, and capacity for complex formation. Besides formic acid, hydroxyacids such as lactic and glycolic acids are the most corrosive.

Oxidation on Laboratory and Pilot Plant Scale. An apparatus is shown in Figure 10 which is suitable for most laboratory scale homolytic oxidations. Some aspects to which attention should be paid are as follows:

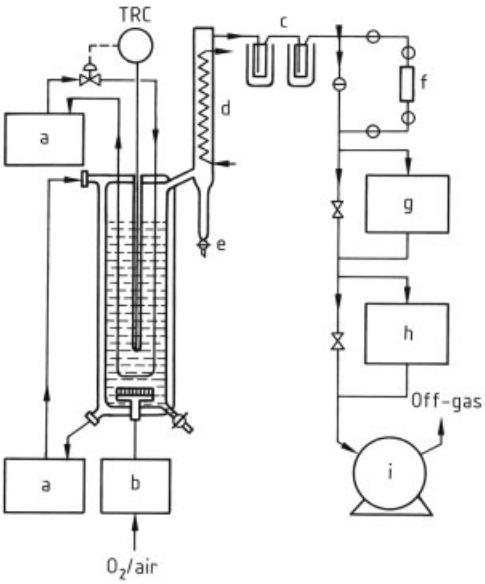

Figure 10. Laboratory apparatus for liquid-phase oxidation of high-boiling compounds
a) Thermostats; b) Gas metering; c) Cold traps; d) Off-gas cooler; e) Water separator; f) Gas sampling; g) O_2 analysis; h) CO/CO_2 analysis; i) Gasmeter

Reactor dimensions:

	Laboratory	Pilot plant
Diameter, cm	> 5	> 30
Height, cm	> 30	up to 800

In larger pilot reactors back-mixing can be limited by the incorporation of, for example, sieve plates.

Reactor Material. In the laboratory, glass should be used if possible (so that the reaction can be observed). With metals, oxygen-free zones should be avoided because of possible corrosion, possibly by O_2 purging.

Reaction temperature is the most important parameter; constant temperature can best be achieved by using two heating circuits: one at the desired reaction temperature for heating the jacket and one for inner cooling which is kept 10 – 50 °C below the reaction temperature.

The size and shape of the *cooling circuit* depend on the reaction and dimensions of the reactor. Cooling tubes should be arranged vertically and sufficiently far from one another and from the wall to minimize disturbances to the gas distribution.

Gas distribution is the second most important parameter. Tubes with small holes or frits made of glass, Teflon, or metal are the most suitable. For metal reactors, the frits should be tested periodically in glass vessels under conditions similar to those of the reaction.

Oxygen Content of the Off-gas. At 0 % oxygen content the experimental results are uncertain because the size of the O_2 deficiency zone is unknown. If higher O_2 contents are permitted in the off-gas, the necessary safety measures must be adhered to very precisely (see Section 4.6).

A water separator is important in the oxidation of nonpolar substances and is recommended for other substances if water is formed as a separate phase.

Gas Sampling. The O_2, CO, and CO_2 content in the off-gas should be continually monitored if possible; attention should be paid to volatile substances (e.g., H_2, CH_4, CH_3CHO). The gas sampling vessel should be purged with at least ten times the volume of the vessel to obtain correct values.

Course of a Liquid-Phase Homolytic Oxidation. Batch and continuous oxidation reac-

tions can be initiated as follows. The carefully cleaned reactor is charged with the starting material and the catalyst under nitrogen purging, the pressure is adjusted to the required value, and the mixture heated to the initiation temperature. Then air or O_2 is introduced and, if necessary, an initiator added. The start of the reaction can be detected by the decrease in the O_2 content of the off-gas and a rise in the reaction temperature. From this time onwards the O_2 supply is increased in such a way that the O_2 content in the off-gas does not exceed a critical limit, for instance 5 %. The acceleration of the reaction to maximum velocity usually occurs rapidly (within 10 to 30 min). The heat of reaction must be removed and the desired reaction temperature regulated. With proper layout, the temperature of the cooling medium and the control function of the valve in the cooling circuit can be adjusted on installation of the system such that the oscillations of the reaction temperature are maintained below 0.1 °C. For batch oxidations, the temperature of the cooling circuit must gradually approximate the reaction temperature towards the end of the reaction, to avoid excessive temperature fluctuations or breakdown of the oxidation process. In continuous oxidations, charging of starting material and removal of the product begin after the required conditions have been established.

Applicability of the Results from Laboratory and Pilot Plant to Full Plant Scale. Scale-up problems are mainly caused by the following factors:

1) The laboratory and pilot plant reactors often have too small a diameter relative to their height. Because the rising gas bubbles entrain liquid, a downward streaming in the rest of the liquid is always produced, predominantly at the wall of the reactor. The resulting turbulence leads to rapid coalescence of the gas bubbles in reactors that are too narrow. Thus the exchange area becomes much smaller and the reaction appears to be slower than it is.
2) The O_2 content of the off-gas is kept near 0 % (for further explanation, see above).

Results that are representative for full plant scale are only obtained with reactors which are of the same height as large-scale reactors and have a diameter such that the gas velocity remains well below 0.02 m/s (diameter 30 cm minimum).

4.5. Homolytic Gas-Phase Oxidation

Homolytic gas-phase oxidations are all combustion processes and gas explosions that take place in free gas space without heterogeneous catalysts. Examples include the combustion of heating oil and natural gas for the production of heat and energy and the rapid combustion of gasoline and diesel in internal combustion engines. Thus the homolytic gas-phase oxidation is one of the most important chemical reactions with respect to the amounts of materials consumed.

Mechanism. In principle up to 250 °C homolytic gas-phase oxidations proceed by the same mechanism as liquid-phase oxidations (see Section 4.3).

However, whereas oxygen transport from the gas to the liquid phase generally limits the rate of the overall reaction in a controlled liquid-phase oxidation, no comparable restriction exists in the gas phase. In the liquid phase the reaction temperature can be held constant at low levels by simple means. The rise in temperature is inversely proportional to the mass of the reaction mixture, and the heat of reaction is quickly transported to the walls of the cooling tubes by convection of the liquid. For reactions in the homogeneous gas phase, the temperature difference for adequate heat removal must be much greater because of the small mass. If the reaction is not limited by the quantity of fuel or oxygen, and the reaction space is not kept very small and well cooled, temperatures of over 1000 °C can occur in gas-phase oxidations, whereby heat transfer to the surroundings occurs mostly by radiation.

Because the intermolecular interactions, in particular through hydrogen bond formation, are considerably weaker in the gas phase than in the liquid phase due to the greater intermolecular distances, in the absence of catalysts, hydroperoxide decomposition can only occur thermally. Because the activation energy for thermal hydroperoxide decomposition is high (see Section 4.4.2), gas-phase oxidations only proceed at the same rate as a comparable liquid-phase

Table 14. Limiting values for the inerting of flammable gases and vapors with N_2 or CO_2 at 20 °C and 1 bar [28]

Compound	Minimum volume ratio				Maximum O_2 concentration in the F+A+inert gas mixture for inerting with N_2 or CO_2	
	of inert gas to fuel (F) for inerting towards addition of air		of inert gas to air (A) for inerting towards addition of fuel		N_2	CO_2
	N_2	CO_2	N_2	CO_2	O_2 (max), vol %	O_2 (max), vol %
Hydrogen	17	10	3.00	1.56	5.0	5.1
Carbon monoxide	4	2.2	2.13	1.13	5.4	5.4
Methane	6	3.3	0.61	0.34	12.1	14.6
Ethane	13	7.5	0.82	0.49	11.0	13.3
Ethylene	16	9	1.00	0.67	10.0	11.7
Ethylene oxide	17.2	15.5	*	*	*	*
Propane	15	8	0.75	0.43	11.8	14.2
Propene	14.5	8	0.75	0.43	11.5	14.1
Cyclopropane	15.5	8	0.75	0.45	11.7	13.9
Butane	17	9.5	0.70	0.39	12.1	14.5
Butadiene	19.5	13	0.89	0.51	10.4	13.0
Hexane	25	14	0.72	0.41	12.1	14.5
Benzene	21	13	0.79	0.45	11.2	13.9

* Values do not exist due to ready decomposition of ethylene oxide.

oxidation at considerably higher temperatures. However, because heat removal is poor, an initiated gas-phase oxidation accelerates considerably faster than a corresponding liquid-phase oxidation and can readily become a thermal explosion.

Up to ca. 250 °C, hydroperoxides predominate as the first neutral intermediates. Between ca. 300 and 450 °C, isomerization and decomposition of the peroxy radicals to give aldehydes and other fragments becomes faster than hydroperoxide formation. Thus above ca. 420 – 450 °C practically no more alkylhydroperoxides are formed, and aldehydes and alcohols together with olefins and H_2O_2 are the primary neutral intermediates.

The decomposition of peroxy radicals, usually with preceding isomerization, has a higher activation energy than the hydroperoxide decomposition. Therefore, in the transition range between 320 and 450 °C, the overall reaction rate of gas-phase oxidation decreases with increasing temperature (region of negative rate – temperature coefficients), and subsequently rises more sharply than before.

Catalysis and Inhibition. Gas-phase oxidations are catalyzed by radical-forming substances (halogens and their compounds, NO/NO_2, nitro compounds, volatile peroxides, phenols, and amines) and also by high-energy radiation and cracking or dehydrogenation catalysts on the reactor walls.

In the gas phase, radical-forming substances generally only act catalytically if they are added in very low concentrations. In higher concentrations (sometimes even in the ppm range), the radicals formed act as radical scavengers for the more reactive hydrocarbon radicals. The mechanism of action of halogen compounds as fire-extinguishing agents and as flameproofing agents in plastics, and that of tetraethyl lead – used as an antiknock agent in gasoline up to the 1980s – is based on this radical trapping.

Chain termination by recombination of smaller, and therefore highly energetic, active radicals (such as $H\cdot$, $OH\cdot$, $CH_3\cdot$) can occur in the gas phase only if a third collision partner absorbs the energy of recombination (bond energy) liberated. Apart from the very rare three-body collisions, this occurs mostly through collisions with surfaces. Thus gas-phase oxidations, like all radical reactions occurring in the gas phase (e.g., decomposition of acetylene or ethylene oxide), are sensitive towards an increase in surface area. This is the mechanism of action of packings as flame arrestors, powder fire extinguishers, and dusts to control firedamp in mining. Degenerate radical chain explosions that proceed via hydroperoxides (e.g., in hy-

Table 15. Explosion limits and ignition temperatures of some compounds (reproduced from [31] by kind permission of Deutscher Eichverlag)

Compound	Explosion limits in air (20 °C, 1033 mbar)				Self-ignition temperature, °C
	Lower, vol %	Upper, vol %	Lower, g/m^3	Upper, g/m^3	
H_2	4.0	75.6	3.3	64	560
CO	12.5	74	145	870	605
NH_3	15	28	105	200	630
H_2N-NH_2	4.7	100	60	1265	(270)
CH_4	5.0	15	33	100	595
C_2H_6	3.0	12.5	37	155	515
C_2H_4	2.7	28.5	31	330	425
C_2H_2	1.5	100	16	1080	305
C_3H_8	2.1	9.5	39	180	470
C_3H_6	2.0	11.7	35	210	455
n-C_4H_{10}	1.5 – 2.0	8.5	37 – 49	210	365
$C(CH_3)_4$	1.3	7.5	40	230	(450)
n-C_5H_{12}	1.4	7.8	41	240	285
n-$C_{10}H_{22}$	0.7	5.4	41	320	205
C_6H_5-C_6H_5	0.7	3.4	45	220	570
C_6H_6	1.2	8.0	39	270	555
$C_6H_5CH_3$	1.2	7.0	46	270	535
$CH_3C_6H_4CH_3$	1.1	7.0	48	310	525
CH_3OH	5.5	31 – 44	73	410 – 590	455
$(CH_3)_2CHOH$	2.0	12	50	300	425
n-C_3H_7OH	2.1	13.5	50	340	405
$(C_2H_5)_2CHOH$	1.2	8	53	300	360
n-$C_5H_{11}OH$	1.3	10.5	47	380	300
CH_2O	7.0	73	87	910	(420)
CH_3CHO	4.0	57	73	1040	140
CH_2CHCHO	2.8	31	65	380	
$(CH_3)_2CHCHO$	1.6	11	47	320	165
CH_3Cl	7.1	18.5	150	400	625
CH_3Br	8.6	20	335	790	535
CH_2CHCH_2Cl	3.2	11.2	105	360	390
CH_2CHCH_2Br	4.3	7.3	215	370	(295)
n-$C_5H_{11}Cl$	1.4 – 1.6	8.6	60 – 70	380	255

drocarbon – oxygen mixtures) can be stopped if fire-extinguishing powder or other radical scavengers are added early enough. This is because that due to indirect chain branching, the pressure rise during the initiation period is so slow, that sufficient time remains for releasing the fire-extinguishing device. But after the initiation period temperature and pressure can rise very sharply, changing the slow oxidation in a thermal explosion. These types of thermal explosion can rapidly become detonations, especially in enclosed spaces such as pipelines.

Nondegenerate branched radical chain reactions (such as the H_2/O_2 reaction), on the other hand, accelerate so rapidly due to direct chain-branching that no time remains for extinguishing measures once the reaction has started (see Fig. 3). They can only be hindered by preventive measures such as filling empty spaces with packings or capillary bundles.

Syntheses of Intermediates by Homolytic Gas-Phase Oxidation. Because of the extreme conditions, only a few simple products are produced by oxidation in the homogeneous gas phase. These include:

1) Synthesis gas (or hydrogen) by combustion of hydrocarbons or coal with a substoichiometric amount of oxygen and in the pres-

ence of water (→ Gas Production, Chap. 3.2.; → Gas Production, Chap. 4.1.).
2) Carbon black (rubber filler and strengthener) by combustion of oil or natural gas with a substoichiometric amount of oxygen (see → Carbon, Chap. 6.4.).
3) Acetylene by partial combustion of hydrocarbons (→ Acetylene, Chap. 4.2.).
4) CO_2 by complete combustion of natural gas.

The degradation of hydrocarbons and chlorohydrocarbons in the atmosphere takes place by radical chain reactions.

4.6. Gas Explosions and Safety Data

Modern civilization depends to a great extent on the use of hydrocarbons. However, great dangers can arise from these hydrocarbons in the presence of O_2 (see Fig. 11). Consequently, an extensive body of safety data has been generated to keep these dangers within acceptable limits and make the handling of hydrocarbons less hazardous. Extensive literature, standards and regulations exist in all industrial countries [23] , [28–36].

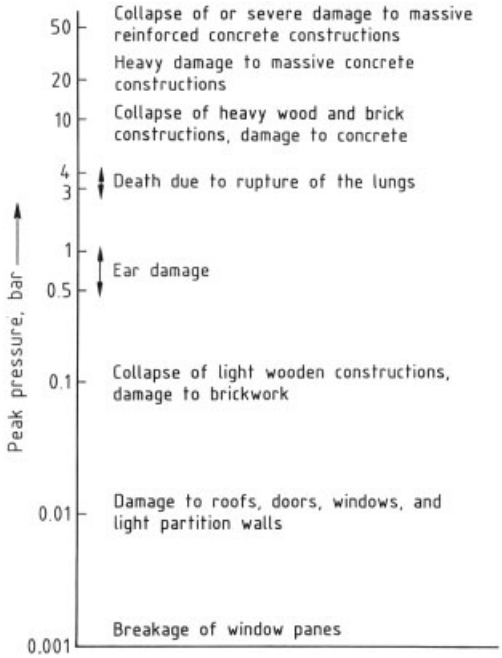

Figure 11. Characteristic damage caused by an atmospheric shock wave reflected from an object as a function of the peak pressure [28, p. 60]

The various safety terms are explained in [28], which also gives the corresponding standards for France, Germany, Russia, the United Kingdom, and the United States.

Explosion Limits, Explosion Groups, Ignition Temperatures, Temperature Classes, Flash Points, and Hazard Classes. All hydrocarbon vapors, and also many other substances, form explosive gas mixtures with oxygen within certain concentration limits, and these mixtures can be made to ignite if a certain, usually quite small amount of energy (electrical sparks, hot surfaces, etc.) is supplied. The lower and upper limiting concentrations of combustible gas at which an explosion can no longer be set off are called the lower and upper explosion limits.

In [28, p. 65] the opposite and therefore incorrect definition of the explosion limits is given.

Figure 12 shows the explosion limits of propane in air as a function of the pressure of the gas mixture before ignition and the ignition energy. Whereas the upper explosion limit is extended significantly by increasing either the pressure or the ignition energy, the lower explosion limit is lowered only to a small extent.

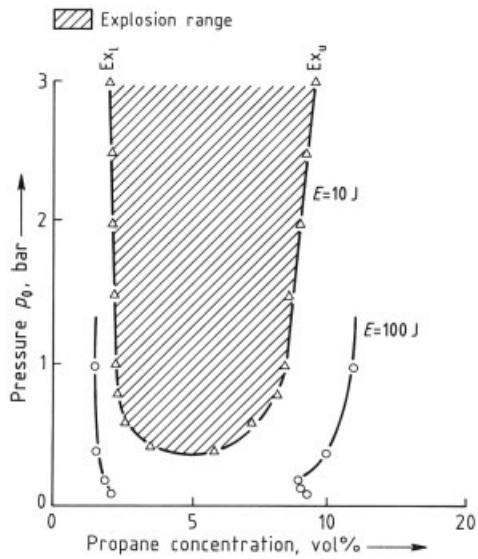

Figure 12. Explosion limits of propane in air as a function of the pressure p prior to ignition and the ignition energy E in a 7-L vessel
Ex_l = lower explosion limit; Ex_u = upper explosion limit
Reproduced from [23] with permission of the publisher

Combustible dusts behave in a manner similar to vapors [29].

In enclosed spaces (tanks, silos) the O_2 concentration can be lowered by dilution with inert gases (e.g., N_2, CO_2) to such an extent that an explosion can no longer be set off (see Table 14).

Above a certain temperature, the mixture ignites without an external ignition source due to autoxidation. This ignition temperature decreases with increasing chain length (see Table 15). It also decreases, in particular in the case of hydrocarbons, the longer the gas mixture has been kept at an elevated temperature below the ignition temperature (see Section 4.5). In pure oxygen the ignition temperatures can be up to 300 °C lower than in air. Like the explosion limits, the ignition temperatures are not substance-dependent constants but depend to a great extent on the measuring conditions. Table 15 lists explosion limits and ignition temperatures for some compounds in air at atmospheric pressure and 20 °C.

For safe handling of hydrocarbons being processed or stored in a plant the surfaces of hot apparatus, tubing, machines, etc. must be so well insulated that no danger of ignition can arise. For this reason combustible substances are classified according to ignition temperature into temperature classes T1 to T6 (see Table 16).

Table 16. Criteria for temperature classes [32]

Ignition temperature of flammable material, °C	Temperature class	Maximum surface temperature, °C	Former ignition class
>450	T1	450	G1
>300	T2	300	G2
>200	T3	200	G3
>135	T4	135	G4
>100	T5	100	G5
> 85	T6	85	

Also, areas where there is a danger of explosion are classified into areas according to the probability of the occurrence of explosive gas mixtures [28]:

Zone 0	Areas in which an explosive atmosphere is always present or present for an extended period of time.
Zone 1	Areas in which an explosive atmosphere is occasionally present.
Zone 2	Areas in which an explosive atmosphere is rarely present or present for only a short time.

Whereas hydrogen is diluted very quickly in air because of its low density and high rate of diffusion, hydrocarbons from propane onwards sink to the ground because of their relatively high density and low rate of diffusion. It has been shown experimentally that 10 m^3 of liquid propane evaporated within 10 min and distributed itself in a layer of 0.7 m thickness over an area of 700 m diameter. In addition the lower explosion limit of hydrogen is 4 vol %, while that of hydrocarbons lies between 1 and 2 vol %. Hydrocarbons are at least as dangerous as hydrogen.

Explosions cannot occur in spaces occupied by packing material. Explosions which have already started are quenched in vessels or pipes filled with Raschig rings, tube bundles, or similar packing materials (explosion barriers). The critical diameter of the capillary tubes at which no flame can be propagated is defined as the quenching distance. The quenching distance depends above all on the combustible gas, but also on the material, temperature, and mass of the explosion barrier. Capillaries and bead fillings are more effective flame barriers than parallel plates (see Table 17).

Table 17. Quenching distances of various flame barriers, [23, p. 58]

Fuel	Flame barrier		
	Parallel plate w_{max}, mm	Capillary, d_{max}, mm	Bead filling $d*_{max}$, mm
H_2	0.3	1	2
CH_4	1.3	4	7
C_3H_8	0.9	3	6

∗ Bead diameter.

Closely associated with the quenching distance is the maximum experimental safe gap (MESG), which principally applies to electrical equipment in surroundings where there is a danger of explosions. The safe gap is defined as that through which an ignition occurring inside the equipment is just unable to penetrate to the outside.

Substances are divided into three explosion groups corresponding to their hazard and assigned particular MESG values (see Table 18). The MESG is smaller than the quenching distance because, unlike the explosion barrier, the gap has a depth of only 25 mm.

Table 18. Classification of explosion groups according to maximum experimental safe gap (MESG) and minimum ignition current ratio (MIC) [32]

Classification	Condition	Example
II A	0.9 mm < MESG 0.8 < MIC	propane
II B	0.5 mm ≤ MESG ≤ 0.9 mm 0.45 ≤ MIC ≤ 0.8	ethylene
II C	MESG ≤ 0.5 mm MIC ≤ 0.45	H_2, C_2H_2, CS_2

The vapor pressure of liquids increases with temperature. The temperature at which the vapor pressure of organic liquids exceeds the lower explosion limit so that the vapor/air mixture above the liquid can be ignited, is defined as the flash point. Because dangers can arise, particularly from low-boiling substances, they have been divided into hazard classes according to flash point (Table 19). Because of the complexity and great number of regulations, the hazard classes according to the UN Recommendation which have arisen from the IATA and RID regulations are becoming increasingly important.

Self-Ignition of Solids, Fires from Liquids. Similar to gases, heat removal from solids is poor. In particular, with solids having large surface areas (coal dust, flour, hay, straw, paper waste, wood shavings etc.) oxidation can result in a significant temperature rise, whereby ignition occurs (see Table 20). The ignited solid only glows (reaction between O_2 and solids). Flames are formed by gas-phase oxidation of the gases and vapors formed by cracking and evaporation of the volatile compounds.

In the case of liquids it is also only the vapors formed which burn. For this the liquids must be heated to temperatures above the combustion point. At the flash point only the gas mixture above the liquid burns and the flame is then extinguished; however, when the combustion point is reached so much liquid evaporates that the gas mixture continues to burn. The combustion point is usually a few degrees above the flash point. Halogenated hydrocarbons, for which the combustion point can be up to 100 °C above the flash point, are an exception. This striking behavior and the suitability of halogenated hydrocarbons to be used as fire extinguishers are due to the same cause: These hydrocarbons form too many radicals. Therefore, at low temperatures, the proportion of chain termination reactions is too high.

Particularly critical situations can arise from fires involving containers of low-boiling liquids, if the container is heated. This can lead to sudden evaporation of a large quantity of liquid and a subsequent explosion. This process is known as BLEVE (burning liquid evaporating vapor explosion). A decision on the particular procedure can only be taken on the spot. But in all cases where the liquid can neither be pumped off nor quenched by dilution with water, a BLEVE should be hindered by intensive cooling of the container. The liquid which could otherwise not be disposed of should, however, be allowed to burn away, since clouds of gas forming after complete quenching by subsequently evaporating liquid could ignite at places which were still hot.

5. Heterolytic Oxidation

Both liquid- and gas-phase heterolytic oxidations have long been known and are used both in the laboratory and in industry.

For example, the oxidation of naphthalene to phthalic anhydride was carried out by BASF with sulfuric acid and $K_2Cr_2O_7$ since 1872, and since 1896 by the Sapper process with sulfuric acid and mercury as the catalyst. From 1916 this process was replaced by air oxidation on V_2O_5 catalysts.

Oxidations with HNO_3, SO_3/H_2SO_4, peroxomonosulfuric acid, MnO_2, KIO_4, etc., still play a role in industry today [17–20].

A more systematic investigation and determination of these reactions began in the 1960s, principally as a result of two developments:

1) The oxidation of ethylene to acetaldehyde with $PdCl_2$, developed by Wacker [39] and Hoechst.
2) Molybdenum-catalyzed epoxidation of propene with *tert*-butyl hydroperoxide, introduced by Halcon and Arco, and their joint venture, Oxirane.

Oxidation in the presence of heterogeneous catalysts is still based almost entirely on empirical principles (both in the liquid and gas phases). It profits considerably from developments in homogeneous liquid-phase oxidation

Table 19. Flash point limits of various transportation and storage regulations [28]

Regulation	Applies to	Class	Flash point limit, °C	Notes
VbF	storage	A I A II A III	< 21 21 – 55 55 – 100	incompletely miscible with water
		B	< 21	completely miscible with water
IATA (RAR)	air transport		< 37.8 37.8 – 93.3	measured in closed cup
IMDG Code	sea transport	3.1 3.2 3.3	< −18 −18 to 23 23 to 69	
RID, EVO, ADR, ADNR, etc.	transport by rail, road, and canal	Materials of UN class 3	21 55 100	

Table 20. Smoldering temperature∗ of deposited dust (5 mm thick layer) [28]

Material	Particle size		Bulk density, kg/L	Smoldering temperature, °C
	Majority, µm	Max., µm		
Coal dust	5 – 10	50	0.41	225
Charcoal	1 – 2	20	0.36	340
Flour	30 – 50	200	0.31	325
Phosphorus, red	30 – 50	150	0.99	305
Polystyrene	40 – 60	150	0.23	melts
Poly(vinyl chloride)	4 – 5	10	0.55	chars
Zinc	10 – 15	90	4.9	430

∗ Temperature of the surface on which the dust lies. It decreases with increasing thickness of the dust layer. The maximum permissible surface temperature must be ca. 100 °C lower than the smoldering temperature.

since the course of reactions on solid surfaces cannot yet be directly investigated. It may be assumed, however, that many reactions on heterogeneous catalysts proceed in the same or at least similar ways with corresponding catalysts in the homogeneous liquid phase.

Depending on the substrate, oxidizing agent, type of catalyst, participating complexing agents, reaction medium, temperature, and the other reaction conditions, heterolytic oxidations proceed by various mechanisms. From the combination of different substrates with the different catalysts, oxygen donors, and reaction media, there is an extraordinarily large number of possible reactions. For this reason only a short overview can be given here, based on [16]. Further reviews can be found in [4], [12], [18–21].

5.1. Heterolytic Oxidation in the Liquid Phase

In heterolytic oxidations in the liquid phase other oxygen donors are frequently used instead of molecular oxygen. Some of these are listed in Table 21, together with the byproducts formed from them.

Table 21. Oxygen donors, [16, p. 5]

Donor	Active oxygen, %	Byproduct
H_2O_2	47.0	H_2O
tert-BuO_2H	17.8	*tert*-BuOH
NaClO	21.6	NaCl
$NaClO_2$	35.6	NaCl
NaBrO	13.4	NaBr
$C_5H_{11}NO_2$ [a]	13.7	$C_5H_{11}NO$
$KHSO_5$	10.5	$KHSO_4$
$NaIO_4$	29.9 [b]	NaI

[a] *N*-Methylmorpholine-*N*-oxide.
[b] Assuming all four O atoms are utilized.

Example 1: Uncatalyzed Epoxidation of Olefins by Peroxyacids (Prilezhaev Reaction). Due to the polarization of the neighboring C=O bond, the hydroperoxy group of peroxyacids can epoxidize C=C double bonds. Alkyl hydroperoxides are less electrophilic and require a catalyst (see Example 3 b)

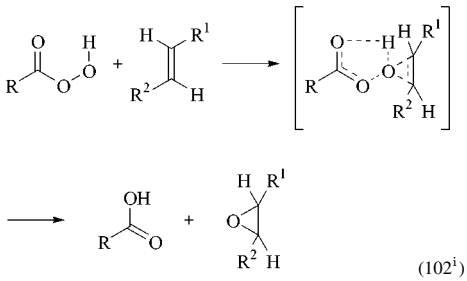

(102[i])

Example 2: Stoichiometric Oxidation with Metal Oxides. Oxides such as SeO_2, CrO_2Cl_2, MnO_4^-, OsO_4, etc. have long been used as oxidizing agents for organic substrates. The following mechanism is assumed for the oxidation of propene by SeO_2:

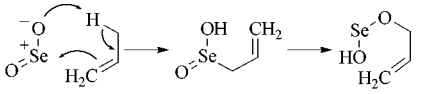

(103[i])

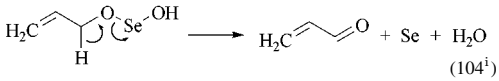

(104[i])

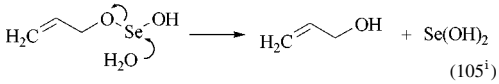

(105[i])

In the presence of *tert*-butyl hydroperoxide (TBHP) and small quantities of SeO_2 the reaction proceeds catalytically (see also Example 3).

The active oxidizing agent is an oxometal species in the case of metal ions that are strong oxidizing agents in their highest oxidation state (e.g., Se^{4+}, Cr^{6+}, V^{5+}, Ce^{4+}, Ru^{8+}), while weakly oxidizing metal ions (e.g., Mo^{6+}, Zr^{4+}, Ti^{4+}) react as peroxometal species. V^{5+} can react by both mechanisms.

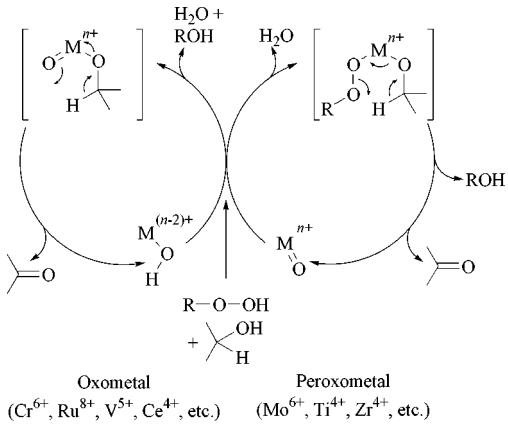

It should be noticed that metal ions reacting via "oxometal" mechanism are reduced by the substrate and reoxidized by the active oxidizing agent during one cycle, whereas the metal ions in the "peroxometal" complexes do not change their oxidation state but add the active oxygen donor to their complex sphere. In this way, primary alcohols can be oxidized (with Zr complexes) to aldehydes, secondary alcohols to ketones, and olefins to epoxides or glycols (see Example 3).

Example 3: Transition Metal Catalyzed Oxidation of Olefins.

a) Some transition metal oxides form inorganic peroxoacids with H_2O_2, analogous to the reactions shown in the above scheme. These inorganic peroxoacids, known as Milas reagents, oxidize olefins to epoxides. The epoxide formed is hydrolyzed in water (H_2O_2 is usually added as a 30 % aqueous solution):

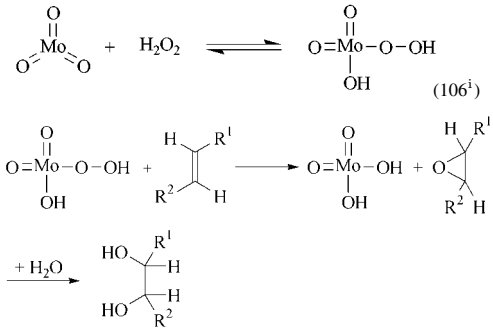

(106[i])

(107[i])

The best known Milas reagent is OsO_4. However, it does not oxidize the olefins via the epoxides but via the osmates:

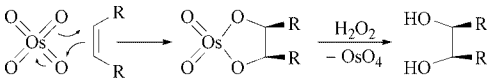

(108[i])

b) Halcon – Oxirane Process (→ Propylene Oxide). If olefins are treated with organic hydroperoxides in nonaqueous solvents in the presence of Milas reagents as catalysts, epoxides are formed in high yields:

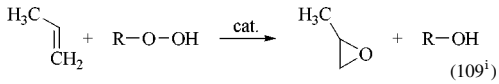

(109[i])

Homogeneous catalysis: Mo^{6+}, W^{6+}, V^{5+}, Ti^{4+} (Arco)
Heterogeneous catalysis: Ti^{4+}/SiO_2 (Shell)

The following mechanism is assumed:

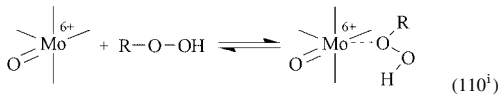

(110[i])

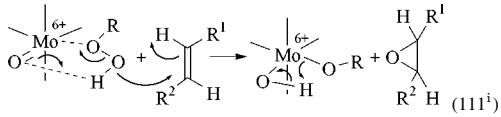

(111[i])

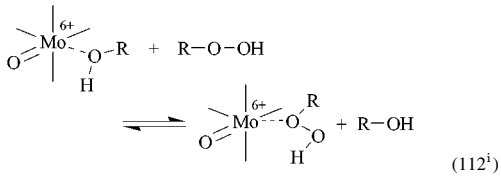

(112[i])

Whereas the alcohols formed from the hydroperoxides have a strongly inhibiting effect on the activity of vanadium, they inhibit the action of molybdenum only in high concentrations (e.g., when used as the solvent). For homogeneous catalysis the metals are added in the form of complexes that are readily soluble in organic materials in concentrations of 1 – 500 ppm. The reaction temperatures are between 80 and 120 °C. The problem of catalyst recycling was solved by Shell by rendering the catalyst heterogeneous.

The major problem of this process is that ca. 2 t *tert*-butanol (TBA) are formed per tonne of propylene oxide. Up to the 1980s Arco recycled the TBA by dehydration to isobutene and subsequent hydrogenation to isobutane. Today, isobutene is converted to methyl-*tert*-butyl ether (MTBE) for which a large market as gasoline additive has developed (→ Methyl-*Tert*-Butyl Ether).

Example 4: Palladium-Catalyzed Oxidation of Olefins. Wacker/Hoechst process (→ Acetaldehyde). SMIDT et al. [39] attempted to oxidize ethylene to ethylene oxide in the gas phase over Pd/C. In contrast to their expectations acetaldehyde was formed. In this way a reaction was rediscovered, which was known since 1894 and used analytically for a long time for determination of palladium. Since the catalysts rapidly became inactive in the gas phase because of formation of palladium aggregates, it was decided to carry out the reaction in the liquid phase (water). By combination with a redox system ($CuCl/CuCl_2$), a catalytic process for acetaldehyde synthesis was developed. Considerable corrosion problems arose due to the use of chloride ions in an acidic medium (pH 0.8 – 3) at 100 – 130 °C.

The reaction scheme is as follows:

$$C_2H_4 + PdCl_2 + H_2O \rightarrow CH_3CHO + Pd^0 + 2\,HCl \qquad (113^i)$$

$$Pd^0 + 2\,CuCl_2 \rightarrow PdCl_2 + Cu_2Cl_2 \qquad (114^i)$$

$$Cu_2Cl_2 + 2\,HCl + 1/2\,O_2 \rightarrow 2\,CuCl_2 + H_2O \qquad (115^i)$$

From experiments with C_2D_4 (very small isotope effect), D_2O ($k_H/k_D = 4.5$, no deuterium in the acetaldehyde), with varying H^+ and Cl^- concentrations (both ions inhibit the reaction), with various nucleophiles (acetic acid or ethanol instead of water as the solvent), and from the finding that *trans* OX substitution occurs in ethylene (OX thus comes directly from the solvent and not from the ligand sphere of the complex) it was concluded that the mechanism is as follows:

1) Substitution of a chloride ligand by ethylene, with formation of a π-complex, which is in equilibrium with the corresponding solvent complex by rapid ligand exchange:

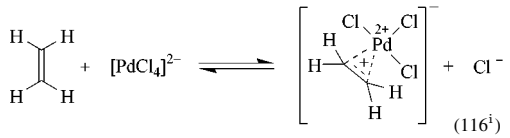

(116^i)

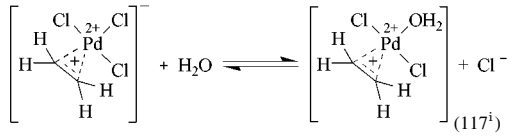

(117^i)

2) Formation of the π-complex lowers the electron density on the ethylene ligand, rendering it susceptible to nucleophilic attack. The π-complex is in equilibrium with the σ-complex. The latter loses a Cl^- ion in a slow reaction and decomposes after hydride migration from C_1 to C_2 via the Pd, and two-electron transfer from C_1 to Pd, to give the final products. The Pd^0 is then reoxidized. Alternatives to this mechanism are discussed in [40], [41]

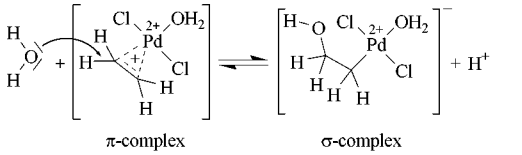

(118^i)

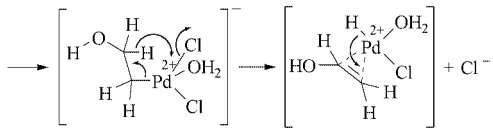

(119^i)

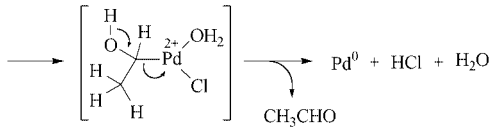

(120^i)

The two major drawbacks of the Wacker process – chlorinated byproducts and severe corrosion – can be overcome by using heteropolyacids (HPA) as the reoxidizing agents [42] instead of the $CuCl/CuCl_2$ system.

Higher olefins are oxidized to ketones. Yields of over 95 % are only obtained with HPA as the reoxidizing agent for palladium.

As shown above, vinyl acetate is formed in acetic acid (and in the presence of Na acetate). Isopropenyl acetate is formed from propene since Pd^{2+} is electrophilic and therefore reacts with the double bond in the manner shown in Reactions (116^i – 120^i) and (121^i). In contrast, Pd^0 abstracts the allylic H atom and gives a π-allyl complex. Thus propene is oxidized to allyl compounds on Pd^0 (on activated carbon or another carrier), whereby a cocatalyst (e.g., $CuCl_2$) is not required, unlike with Pd^{2+}. With water as solvent acrylic acid is formed (124^i).

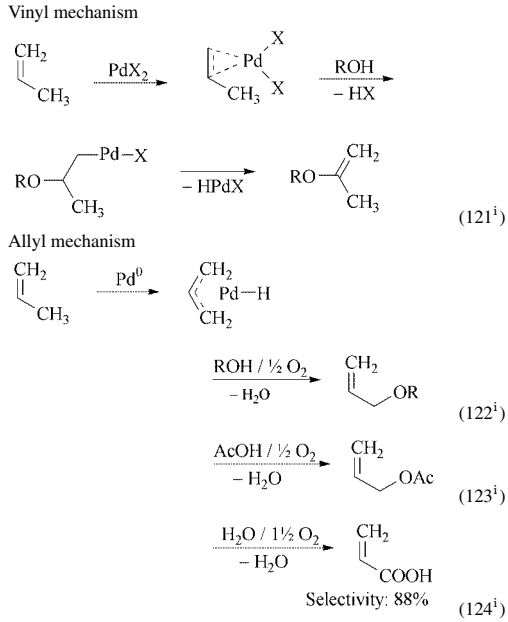

(121^i)

(122^i)

(123^i)

(124^i)

Palladium(0) on activated carbon also catalyzes the oxidation of the methyl group of toluene under mild conditions:

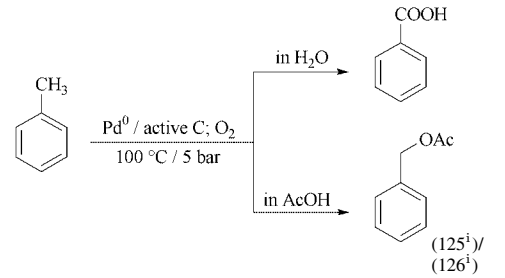

(125^i)/ (126^i)

Further examples of the heterolytic oxidation of olefins are:

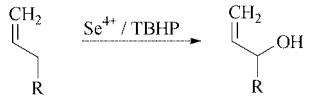

(127[i])

TBHP = *tert*-butyl hydroperoxide

Oxidative C=C cleavage:

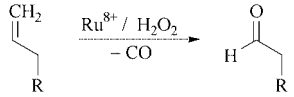

(128[i])

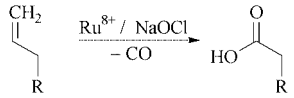

(129[i])

Alcohols can be oxidized in a similar manner, as the following examples show:

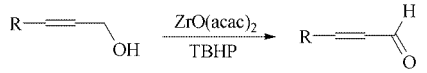

Yield: 85 - 95 % (130[i])

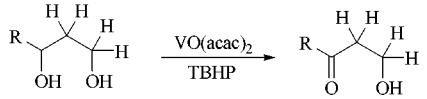

Yield: 96 % (131[i])

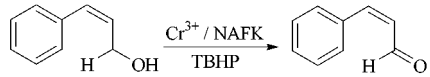

Yield: 81 % (132[i])

acac = acetylacetonate
NAFK = Nafion R 511 perfluorinated ion-exchange resin

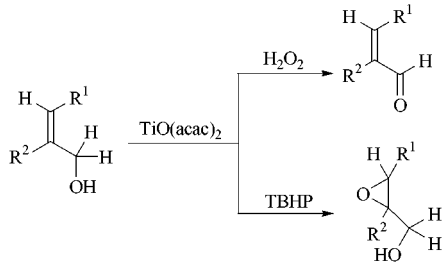

(133[i]), (134[i])

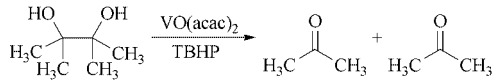

(135[i])

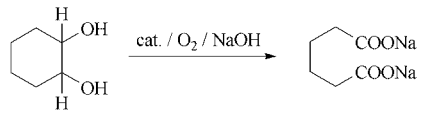

(136[i])

cat. : $Pb_{2.63}Ru_{1.37}O_{6.5}$ or $Bi_{2.39}Ru_{1.61}O_{7-y}$

Heteropolyacids (HPA) are gaining increasing importance because of their unique combination of properties:

1) Strong Brønsted acids
2) Multi-electron oxidizing agents
3) Solubility in water and oxygenated organic solvents (soluble oxides)
4) HPA anions are stable to oxidation and function as multi-electron ligands that can stabilize reactive, higher-valent oxometal species

For example, the oxidative cleavage of diols was achieved with H_2O_2 in *tert*-butanol in the presence of the cetylpyridinium salt of the heteropolyacid $H_3PW_{12}O_{40}$ as catalyst.

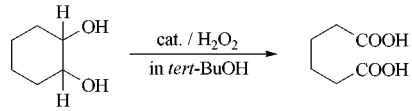

(137[i])

cat.: (cetylpyridinium)$_3$ $PW_{12}O_{40}$

To avoid the problems associated with catalyst recycling, Anic/Enichem has incorporated Ti^{4+} into a zeolite lattice and obtained titanium-silicalite (TS-1). TS-1 has the same elemental composition as the Shell Ti–SiO_2 catalyst, but a considerably wider range of applications. It is assumed that active, isolated titanyl species are responsible for the catalytic activity:

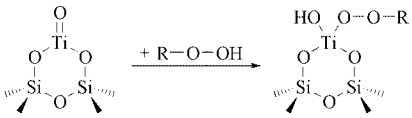

TS-1 (138[i])

With TS-1 as catalyst, the following reactions have, among others, been carried out with 30 % H_2O_2:

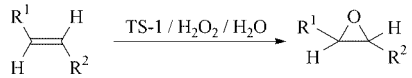

Selectivity: 96 % (139[i])

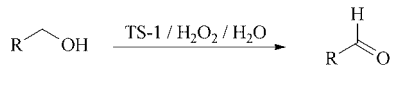

(140[i])

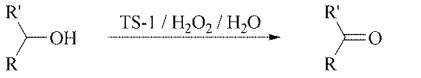

(141^i)

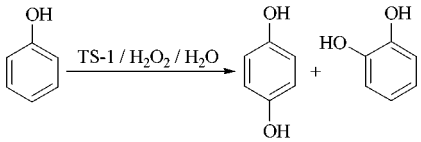

(142^i)

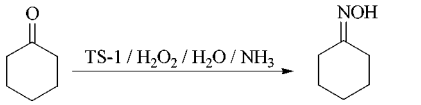

(143^i)

For hydrogen peroxide and the inorganic water-soluble oxygen donors, there is often a miscibility problem with the organic compound. This problem can be solved in many cases by using a two-phase aqueous – organic solvent system in conjunction with a phase transfer catalyst.

Process Technology of Heterolytic Liquid-Phase Oxidation. Because of the great number of possible reactions, general rules cannot be drawn up. In the cases where only liquids are reacted with each other, stirred tanks are suitable reactors for small batches. For continuous production, reactors with turbulent plug flow are generally the most suitable.

If a catalyst precursor is used a formation phase is necessary for initiation. After this formation phase, and with the use of preformed catalysts, the reaction proceeds immediately with its maximum possible rate.

The reactions are generally very selective. If the catalyst permits, they can be run until complete conversion of the starting material is reached.

5.2. Heterogeneously Catalyzed Gas-Phase Oxidation

The area of heterogeneously catalyzed gas phase oxidations is still largely empirical, with regard to the composition and preparation of the catalysts, and the control and influencing of the processes occurring on the catalyst surface during the reaction.

Considerable progress has been made in the investigation and characterization of surfaces. Surface analysis methods, such as Auger electron spectroscopy (AES), secondary ion mass spectroscopy (SIMS), and low energy electron diffraction (LEED), have been applied to catalysts. By using such techniques, considerable advances in the characterization of and the differentiation between good and poor catalysts have been made. This is important both for reproducible catalyst production and for the explanation of poisoning and aging processes.

Table 22. Types of catalyst for heterogeneous gas-phase oxidation

Catalyst	Reactions and examples
Noble metals(Pt, Pd, Ag)	
Pt(+ Rh,Au) gauze	NH_3 oxidation to NO/NO_2
	Andrussow process:
	$CH_4 + NH_3 + \frac{3}{2}\,O_2 \longrightarrow HCN + 3\,H_2O$
Pt(+ Rh)/TiO_2/carrier	off-gas combustion/purification
Pt(+ Bi,Au)/carrier	oxyacetylation, e.g.:
	ethylene ⟶ vinyl acetate
	propene ⟶ allyl acetate
Ag(+Cs)/carrier	ethylene oxide direct synthesis
Ag (pure)	oxidative alcohol dehydrogenation:
	methanol ⟶ formaldehyde
	ethanol ⟶ acetaldehyde
Cu - Fe catalysts	
$CuCl_2$(Alkali chloride)/carrier	Deacon process:
	$2\,HCl + \frac{1}{2}\,O_2 \longrightarrow Cl_2 + H_2O$
	Oxychlorination:
	$C_2H_4 + 2\,HCl + \frac{1}{2}\,O_2 \longrightarrow ClCH_2{-}CH_2Cl + H_2O$
Cu - Mn - Cr oxide	off-gas combustion/purification
Mg - Cr ferrite	oxidative dehydrogenation:
	butene ⟶ butadiene ∗
CuO - ZnO	oxidative decarbonylation:
	α-branched aldehydes ⟶ ketones ∗
	(e.g., isobutyraldehyde ⟶ acetone)
Cu_2O/carrier	propene ⟶ acrolein
Catalysts based on V_2O_5 + MoO_3	
V_2O_5(+ $K_2S_2O_7$)/carrier	Contact process: $SO_2 + \frac{1}{2}\,O_2 \longrightarrow SO_3$
V_2O_5(+ P_2O_5)/carrier	o-xylene ⟶ phthalic anhydride (→ phthalic acid)
V_2O_5(+ P_2O_5 + TiO_2)/carrier	butane/butene ⟶ maleic anhydride (→ maleic acid)
V_2O_5(+ MoO_3 + additives)/carrier	benzene ⟶ maleic anhydride (→ maleic acid)
V_2O_5(+ SnO_2 + additives)/carrier	ammonoxidation: alkyl aromatics ⟶ aromatic nitriles
MoO_3 - Bi_2O_3(+ various additives)	oxidative dehydrogenation:
	butene ⟶ butadiene
	ammonoxidation:
	propene ⟶ acrylonitrile
	propene ⟶ acrolein
MoO_3 - V_2O_5(+ various additives)	acrolein ⟶ acrylic acid
MoO_3 - Fe_2O_3	Formox process:
	methanol ⟶ formaldehyde

∗ Of minor industrial importance.

Most of these spectroscopic methods can, however, only be performed under high vacuum. Thus for all industrially interesting cat-

alysts one is still far away from investigating the processes occurring on the catalyst surface during the reaction. Nevertheless, a large number of observations and indirect methods of investigation, (e.g., sorption measurements, use of isotopically labeled compounds, alternating oxidation and reduction) in combination with the above-mentioned methods of physical analysis allows a rough picture to be built up of what happens on the surface of the catalyst.

As long as the processes on the surface of the catalyst are practically inaccessible to direct investigation, all proposed mechanisms can only be speculative.

For a summary of the literature up to 1972, see [43]; up to 1979, [12], [44], [45]; (see also → Heterogeneous Catalysis and Solid Catalysts).

5.2.1. Oxidation Catalysts

Table 22 lists the most important oxidation catalysts and the reactions catalyzed by them [46]. The catalysts are divided into three groups:

1) The *noble metals Pt and Ag* are the longest known catalysts. Their large-scale use as oxidation catalysts was limited for a long time to the combustion of ammonia (Pt/Rh, Ostwald process) and the synthesis of formaldehyde (Ag; BASF process). After World War II, the direct oxidation of ethylene to ethylene oxide on Ag, the oxydehydrogenation of ethanol on Ag to give acetaldehyde, and the Andrussow process for the synthesis of HCN on Pt catalysts became important. In the late 1960s, the combustion of off-gases on Pt/Rh and the acetoxylation of olefins on Pd were introduced.
2) In the 1940s *copper-based catalysts* became important for the oxidation of propene to acrolein. In the 1950s various mixed-oxide catalysts based on Fe/Cr oxide were developed for the oxidative dehydrogenation of butenes to give butadiene. In the 1960s the $CuCl_2$/carrier catalysts, known since 1868, achieved renewed industrial importance for the oxyhydrochlorination of ethylene/HCl mixtures to give dichloroethane (precursor for vinyl chloride).
3) After Pt and Ag, the *vanadium oxide catalysts* are the longest used industrial catalysts for oxidation processes; V_2O_5 was first used in the oxidation of SO_2 to SO_3, and later in the oxidation of naphthalene to phthalic anhydride, and benzene to maleic anhydride.

The discovery of the $Bi_2O_3 \cdot MoO_3$ catalysts by Sohio had a similarly revolutionary effect on the development of mixed-oxide catalysts for gas-phase oxidation, as the discovery of the palladium catalysts for ethylene oxidation on the development of the heterolytic liquid-phase oxidation.

Properties and Mechanism of Action of Heterogeneous Oxidation Catalysts. Most heterogeneously catalyzed oxidations probably proceed by heterolytic mechanisms. As shown in Section 5.1 these mechanisms differ, depending on the catalyst used.

Noble Metal Catalysts. A characteristic of noble metal catalysts is that reducing conditions must generally be used for selective oxidations; i.e., the hydrocarbon must be in excess and oxygen in deficiency. With excess oxygen the pure platinum metals in particular catalyze only complete combustion, which begins at 200 – 250 °C on active catalysts.

The most active catalysts for total combustion (off-gas purification) are based on Pt/Rh; they catalyze the complete oxidation of hydrocarbon compounds to CO_2 and H_2O at a stoichiometric O_2 availability.

For the acetoxylation of olefins over a Pd^0 catalyst, a mechanism was proposed in Reaction (123^i). It is assumed that this mechanism is also valid for the corresponding reactions in the gas phase. The palladium must be modified with Bi and Au, and sodium acetate must be continuously added to the starting material as a moderator, so that the reaction proceeds with the desired selectivity.

Oxidation of Ethylene to Ethylene Oxide on Silver Catalysts (see also → Ethylene Oxide, Chap. 4.1., → Ethylene Oxide, Chap. 4.2.). Up to now ethylene is the only olefin which can be selectively oxidized to the epoxide in the gas phase, and silver is the only catalyst on which the reaction proceeds with high selectivity.

Ethylene can be oxidized to ethylene oxide with an excess of oxygen on silver catalysts.

Because the selectivity decreases with increasing conversion, small ethylene conversions are used industrially. Because of the need to recycle the ethylene, pure oxygen is generally used. The oxygen concentration is determined by the lower explosion limit with respect to oxygen.

Figure 13 shows the development of the ethylene oxide selectivity. The first catalysts gave selectivities of ca. 50 %, which were gradually increased to 65 – 70 % by addition of a few ppm of a chlorohydrocarbon to the make-up gas to lower the activity, and by improvements in catalyst preparation. In the 1970s Shell raised the ethylene oxide selectivity from below 70 % to ca. 75 % by doping the catalyst with 100 – 500 mg/kg cesium [47]. Later, ethylene oxide selectivities of over 86 % were achieved by additional doping of the Ag/Cs/carrier catalysts with rhenium and sulfate ions.

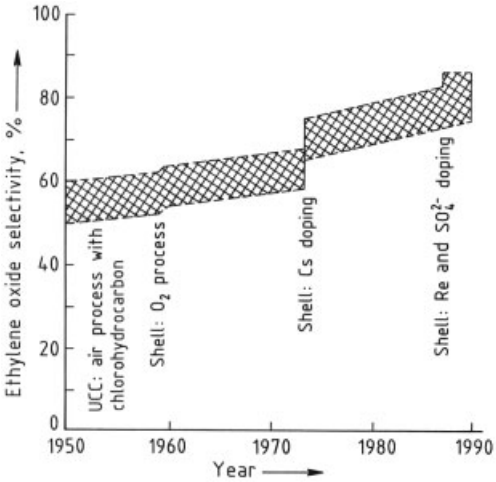

Figure 13. Development of the selectivity in the oxidation of ethylene oxide since 1950
The selectivities attained in industrial plants lie in the shaded region

Although the oxidation of ethylene on silver has been known for 60 years and has been investigated intensively, there is no clarity over the mechanism either with regard to the activation of the oxygen or the ethylene. Two mechanisms are principally discussed for the oxygen activation (see also → Ethylene Oxide, Chap. 4.2.):

According to *mechanism 1* oxygen molecules adsorbed end-on on the silver surface at certain interstitial sites are believed to oxidize the ethylene diffusing to the site from the gas phase to give ethylene oxide selectively. The oxygen atom bound to the silver remains on the surface.

Ethylene oxide and carbon dioxide formation proceed in parallel in the industrially used temperature range and also over wide ranges of ethylene and oxygen concentration, and appear to be coupled with one another. Moreover, several studies showed that chemisorbed oxygen atoms only totally oxidize ethylene. It was therefore assumed that the chemisorbed oxygen atom remaining after transfer of one oxygen atom to ethylene, is subsequently removed by the other ethylene molecule with total oxidation. Subsequently, the cycle can start again.

In this mechanism one ethylene molecule can remove a maximum of six oxygen atoms from the interstitial sites at which ethylene oxide has previously been formed. Therefore, the ethylene oxide selectivity should not exceed the limit of $100 \times 6/7 = 85.7$ %.

In *mechanism 2* oxygen atoms adsorbed on the silver surface bring about both selective and total oxidation of ethylene. Is is assumed that promoter and moderator ions in the neighborhood of the active oxygen atom (e.g., Cs^+, Cl^-, and also O^{2-}) influence its electronegativity and thus determine whether it acts as a selective or total oxidant (lowering the electronegativity favors selective oxidation).

Given the available experimental evidence, arguments can be made in support of either mechanism [48].

If the ethylene oxide selectivity of more than 86 % reported in [49] is confirmed, the second part of mechanism 1 would be refuted.

There have been practically no investigations of the nature of the ethylene activation. It only seems certain that it does not react directly from the gas phase – as postulated in mechanism 1 – but is first adsorbed on the catalyst surface.

The reaction is zero, first, or fractional order, depending on the ethylene/oxygen ratio in the gas phase. Under industrial conditions (concentration < 9 % O_2, 30 – 40 % ethylene; per pass, 40 % of the O_2 and only ca. 10 % of the ethylene are converted), the reaction is zero order with respect to ethylene and first order with respect to oxygen.

The oxidative dehydrogenation of ethanol to give acetaldehyde (see → Acetaldehyde, Chap. 4.1.) on silver or copper catalysts has long been known and was formerly the method

of choice in the laboratory. Formaldehyde was also previously produced by this method, but is now mainly produced in the Formox process by oxidative dehydrogenation of methanol on iron molybdate catalysts (→ Formaldehyde, Chap. 4.2.).

In the Hüls process (originally developed by BASF), ethanol is mixed with air and passed over coarse-grained crystals of pure silver at 500 – 650 °C. There is only 50 – 70 % conversion of ethanol, but complete consumption of oxygen. The off-gas contains up to 11 % unoxidized hydrogen.

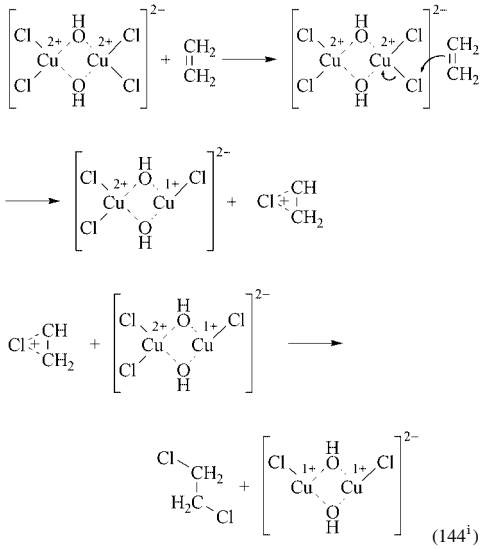

(144^{i})

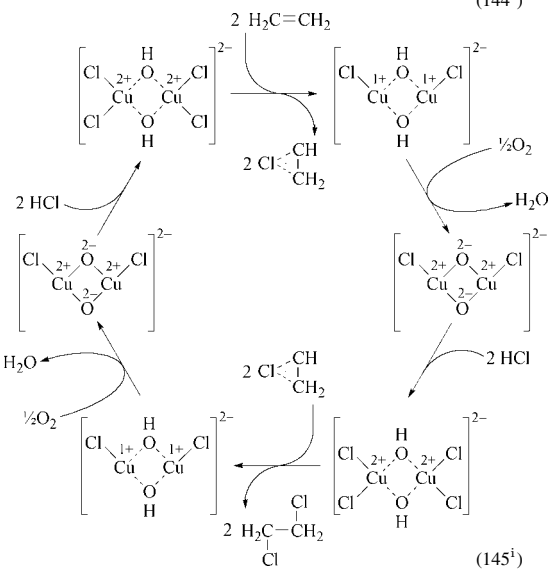

(145^{i})

Copper Catalysts. Whereas the oxidation of HCl and the oxychlorination of ethane require temperatures over 400 °C, the oxychlorination of ethylene proceeds at ca. 230 °C. An electrophilic attack of the ethylene double bond on a copper hydroxy chloro complex with formation of an ethylene-chloronium complex and reduction of Cu^{2+} to Cu^{+} with subsequent transaddition of a second chloride ion to this chloronium complex is assumed. This is followed by reoxidation of the copper (Reaction 144^{i} + 145^{i}).

Mixed Oxide Catalysts. There are essentially two groups of mixed oxide catalysts:

1) Molybdenum-based catalysts that contain bismuth or vanadium and/or tungsten as the second main component and which are used, for instance for the oxidation of propene to acrolein and acrylic acid or to acrylonitrile, and for oxidative dehydrogenation of butene to give butadiene.
2) Catalysts based on vanadium, with, e.g., phosphate, sulfate, TiO_2, or MoO_3 as the second main component, are principally used for the oxidation of aromatics and butane/butene to give acid anhydrides (benzene or butane/butene to maleic anhydride, naphthalene or *o*-xylene to phthalic anhydride).

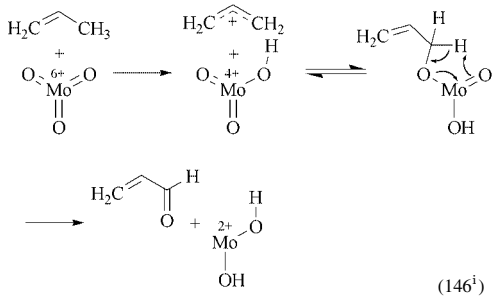

(146^{i})

While V_2O_5 and MoO_3 belong to the oxides which are themselves catalytically active, oxides such as Bi_2O_3, Sb_2O_3, SnO_2, and TiO_2 possess virtually no catalytic activity of their own. However, in combination with V_2O_5 and MoO_3 the catalytic activity and selectivity of the latter are greatly increased.

Hydrocarbon Activation on Mixed-Oxide Catalysts. Experiments with ^{14}C-labeled propene indicate that it is chemisorbed as a

π-allyl complex on the Mo/Bi catalysts with hydrogen abstraction. It is assumed that this π-complex reacts further with the metal oxo complex in a manner analogous to the reaction with SeO_2 with oxygen transfer and metal reduction (see Reactions 103^i to 105^i).

On a catalyst with the composition $Mo_{10}BiFeCo_4W_2K_{0.06}Si_{1.35}$ a selectivity for acrolein of 90 % was achieved at a propene conversion of 98 %. Only ca. 6 % of the acrolein was further oxidized to acrylic acid on this catalyst [total selectivity (acrolein + acrylic acid): 96 %].

The selective further oxidation of acrolein can be achieved on Mo – V catalysts (see → Acrylic Acid and Derivatives, Chap. 1.3.1.). Obviously the addition of vanadium leads to stronger adsorption and activation of the acrolein than on the Mo – Bi catalyst.

The oxidative dehydrogenation of butene to give butadiene proceeds similarly. On Mo – Bi catalysts the first step of the oxidation proceeds as with propene. Like acrolein, the butadiene formed is practically completely desorbed from the catalyst (selectivity > 90 %) and is not further oxidized. Two mechanisms can be discussed:

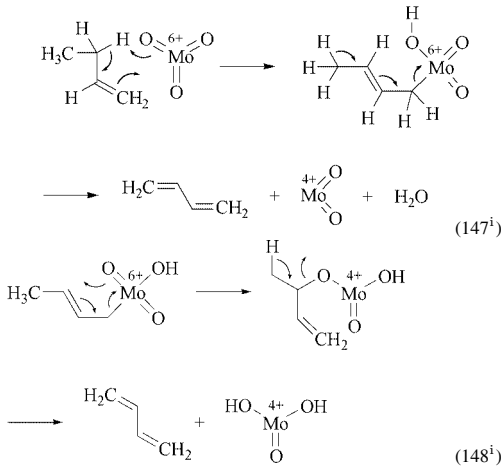

On vanadium-containing catalysts (see Table 22), however, the butadiene is rapidly oxidized further to maleic anhydride.

At the high temperatures necessary for the oxidation of butane (400 – 450 °C), vanadyl phosphate catalysts, $(VO)_2P_2O_7$, have proven particularly selective. The P : V ratio in the bulk catalyst particle is 1 : 1, corresponding to the stoichiometry, but on the catalyst surface it is 2 : 1. The phosphate apparently has two functions: it stabilizes the vanadium in high valence states by complexation and it inhibits homogeneous gas-phase oxidation of butane and its intermediates.

Methyl aromatics (toluene, xylene) and naphthalene are oxidized with the greatest selectivity on vanadium catalysts via the aldehydes to the acids or acid anhydrides. High yields are obtained only with those hydrocarbons that form moderately polar, readily desorbable, volatile products, such as *o*-xylene, from which phthalic anhydride is formed. *p*-Xylene initially affords terephthalic acid, which is too polar and is therefore hardly desorbed. The anhydride of terephthalic acid is oligomeric and therefore insufficiently volatile. For this reason *p*-xylene can be oxidized selectively in the gas phase only by ammoxidation to give terephthalonitrile. Normal air oxidation, in contrast, results predominantly in total oxidation.

Mixed oxide catalysts are only catalytically active and selective in their higher oxidation states. Unlike the noble metal catalysts they must be used in an excess of oxygen, i.e., in an oxidizing atmosphere.

It is generally assumed that the oxygen originates from the oxide lattice in selective oxidations on mixed oxide catalysts and that the metal ions are reduced and oxidized again alternately (redox mechanism of Mars and Van Krevelen). Up to now there are only vague ideas about the way in which atmospheric oxygen is activated and incorporated into the lattice.

Whereas reoxidation proceeds very rapidly on vanadium catalysts (the reaction temperature is mainly determined by the temperature necessary to activate the hydrocarbon), the reoxidation of the catalyst is the activity-determining process for molybdenum-based catalysts, which also determines the reaction temperature. This is demonstrated, for example, by the fact that for catalysts based on V_2O_5 the active mass can be applied as a thin layer on a carrier (egg-shell catalyst), while catalysts based on MoO_3 only exhibit sufficient activity as bulk catalysts.

The importance of the reoxidation rate in mixed oxide catalysts, is exemplified by the development of the Bi–Mo catalysts. When they were introduced by Sohio in the early

1960s, the temperature required for oxidation of propene was 400 – 450 °C, and the selectivity for acrolein, between 60 and 70 %. By the addition of metal ions such as Fe(II/III), Cu(II), Co(II) and Ni(II), which only catalyze one-electron transfers, the reaction temperature could be lowered to 300 – 330 °C and the selectivity increased to 90 – 95 %. It is assumed that the very high activation energy of the oxygen dissociation necessary for the reoxidation of molybdenum is strongly lowered by these additives.

Today the reduction/oxidation-mechanism – developed by MARS and VAN KREVELEN in the 1960s – is generally accepted as the only valid one. However, taking the later developed "oxometal-peroxometal" scheme (see Section 5.1) into account, with oxidation catalysts containing Mo and even V, a peroxometal-mechanism without reduction and reoxidation of the central metal ion should be considered too.

The oxidation of hydrocarbons on mixed oxide catalysts generally follows Langmuir – Hinshelwood kinetics, i.e., the reaction rate depends on both the concentration of the hydrocarbon and the concentration of oxygen in the gas phase.

Because of the large number of parallel and consecutive processes occurring simultaneously on the catalyst surface (adsorption, chemical reactions, and desorption) kinetic measurements only give apparent values for heterogeneously catalyzed oxidations. CREMER was the first to point out the "compensation effect" occurring in these types of kinetic measurements [50]. In the Arrhenius equation

$$\mathrm{k} = A_{\mathrm{o}} \cdot \mathrm{e}^{-E_a/RT} \tag{7}$$

the collision factor A decreases with decreasing activation energy and vice versa.

5.2.2. Process Technology of Heterogeneously Catalyzed Oxidations

Here only the oxidation processes on carrier or mixed oxide full-contact catalysts are described, whereas processes involving gauzes or thin layers of noble metals and waste gas combustion are not considered.

Catalyst Preparation and Characterization [22] (see also → Heterogeneous Catalysis and Solid Catalysts). The surface area of oxidation catalysts should be relatively small (0.1 – 10 m^2/g), i.e., the catalyst should have no pores smaller than 50 nm, since these favor total combustion due to the poor material transport.

Reactors. Because of the large quantities of heat liberated in oxidation reactions and the poor heat transport through the gas, either multitube reactors (with up to 40 000 tubes of 10 – 40 mm diameter and 1 – 12 m length) or fluidized-bed reactors are used.

In multitube reactors the heat is removed either by boiling liquids (water or organic heat transfer media) or by salt melts (e.g., 53 wt % KNO_3, 40 wt % $NaNO_2$, 7 wt % $NaNO_3$, *mp* 150 °C, usable up to 550 °C; reactor material preferably untempered mild steel). In fluidized-bed reactors the heat is removed by cooling tubes containing water (high-pressure steam production).

Because of the unavoidable back-mixing in fluidized-bed reactors the selectivity is always lower than in multitube reactors with their plug flow. For this reason fluidized-bed reactors are only used for catalysts with particularly short lifetimes (< 1 year). In multitube reactors, catalyst exchange always means a shutdown of several weeks. In particular, the pressure loss compensation of the freshly filled tubes is still a tedious manual process. In fluidized-bed reactors the catalyst can be continuously exchanged, even during operation. Typical examples for the use of fluidized beds are in the older acrylonitrile plants, and the oxidation of butane to maleic anhydride.

Temperature Profiles. In fluidized-bed reactors, heat transport by the turbulent mixed solid catalyst is very good. Therefore, uniform temperature is established throughout the entire bed.

In fixed-bed reactors, heat transport by the gas is poor and characteristic temperature profiles develop along the axis of each tube. Because the temperature differences between catalyst bed and bath in each spot along the tubes are proportional to the heat evolved in the catalyst bed at this spot, these temperature profiles give a

fairly accurate indication of the overall reaction taking place at each spot along the catalyst bed. If capillaries (thermowells) are mounted in the axis of some tubes, these temperature profiles can be measured by moveable thermocouples. Figure 14 shows three typical temperature profiles.

With highly active catalysts (Fig. 14 A) the main reaction occurs in a very limited region with a sharp temperature peak (hot spot), whose height is generally proportional to the hydrocarbon loading (large reserve of catalytically active centers in the region of the main reaction zone). Because of the large temperature difference between the cooling medium and the catalyst in the region of the temperature maximum (high temperature in the main reaction zone and relatively low temperature in the post-reaction zone), the hydrocarbon is generally not completely converted, but often only to the extent of 80 or 90 %. The reaction is first order with respect to the hydrocarbon.

With a moderately active catalyst (Fig. 14 B), there is only a small temperature rise with increasing hydrocarbon load, and the temperature maximum is broadened (most active centers of the catalyst in the region of the temperature maximum participate in the reaction). The reaction is first order with respect to the hydrocarbon provided that sufficient catalytically active sites are free in the contact bed region below the reaction zone, which is indicated by the decrease of the reaction temperature below the temperature maximum to the level of the cooling medium. The hydrocarbon conversion is more complete the smaller the difference between the temperature maximum and the cooling medium (mode of operation typical of an integral reactor).

With a low-activity catalyst (Fig. 14 C) the reaction is zero order with respect to the hydrocarbon because all active centers are occupied. Because of the hydrocarbon excess which is usually necessary, the reaction is run according to the permitted oxygen conversion (mode of operation typical of a differential reactor with hydrocarbon recirculation).

Regular measurements of the temperature profile reveal changes in catalyst behavior (e.g., due to aging or poisoning). In addition, the true catalyst performance (kilogram product per kilogram catalyst per second) in sections of the bed can be evaluated from the temperature profile.

Addition of Steam. Water is adsorbed more strongly on the catalyst surface than most organic products because of its high polarity. By addition of steam to the make-up gas, polar substances can thus be desorbed from the catalyst surface by competing adsorption and their yields increased. This method is used above all in the oxidation of propene to acrolein and acrylic acid and in oxidative dehydrogenation. For acid anhydrides the addition of steam clearly leads to a lower yield, although the dehydration equilibrium lies completely on the side of the anhydride at the reaction temperatures.

Addition of steam also improves heat removal. Because the diffusion rate in gases increases with decreasing molecular mass, the heat conducting capacity of a gas mixture increases with the number of small molecules. Like methane and ammonia, water is one of the smallest stable gas molecules.

Laboratory and Pilot Plant Experiments; Applicability of the Results. The choice of reactor is orientated according to the planned investigations:

For kinetic measurements gradientless reactors must be used; details of construction are given in [22], [51], [52].

For the development of catalysts, reactors similar to those used on industrial scale should be used.

In *tube reactors* catalyst testing is straightforward. A suitable apparatus is described in [53]. Smaller reactors (e.g., made of glass with an inner diameter of 20 mm and a bed height of 100 – 200 mm) gave identical results to pilot plant reactors with bed heights of 1 – 3 m, provided that the reactor diameter, catalyst diameter and height, and the residence time based on the free reactor volume were in agreement with the pilot plant-reactor; that the residence time was not greater than 5 s (at longer residence times it is not certain whether the flow is still turbulent); and that the catalyst was not too active.

With very active catalysts, whose temperature profiles are similar to Figure 14 A, scaling up is often associated with loss in selectivity due to the higher temperature in the hot spot or a lower conversion as a result of the lower bath

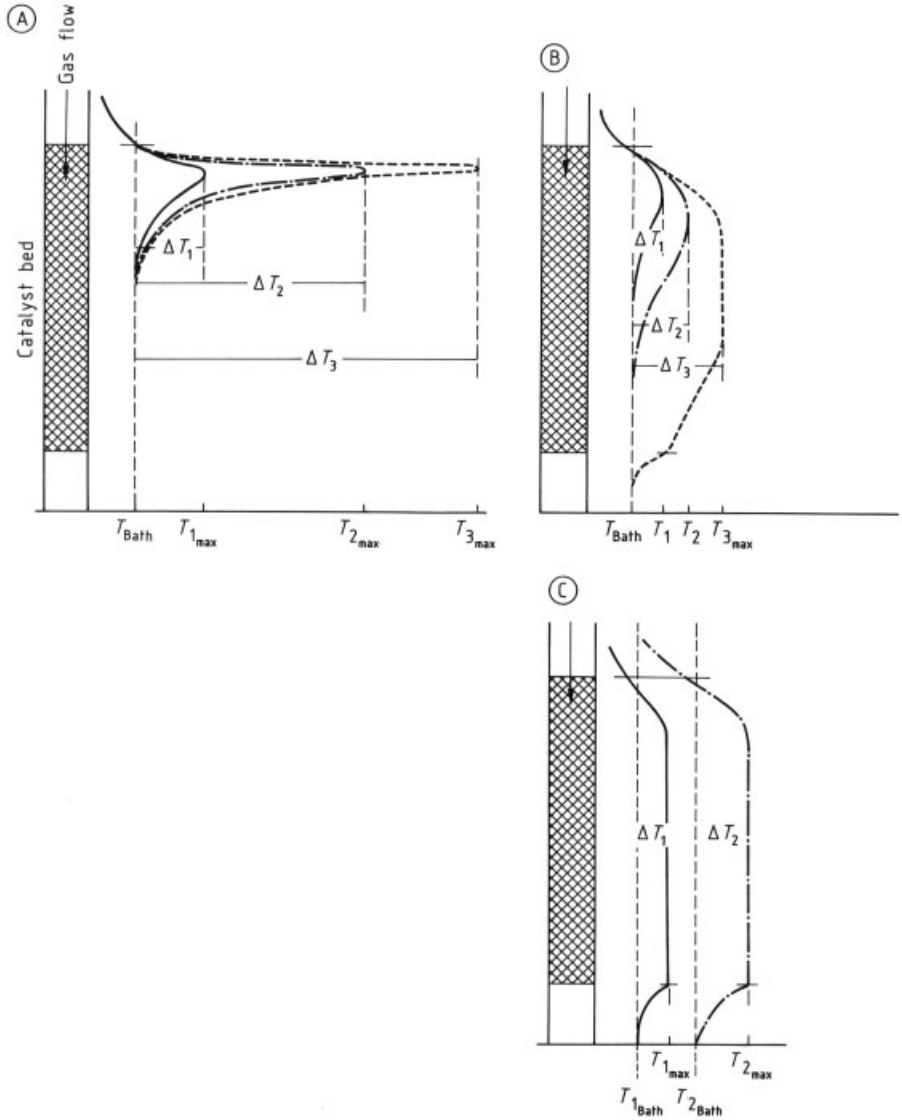

Figure 14. Temperature profiles of various catalysts
A) Temperature profile of a highly active catalyst (V_2O_5 type) as a function of hydrocarbon loading (ΔT_1, low loading; ΔT_3, high loading).
B) Temperature profile of a moderately active catalyst ($BiO_3 \cdot MoO_3$ type) as a function of hydrocarbon loading (ΔT_1, low loading; ΔT_3, high loading).
C) Temperature profile of a low-activity catalyst (Pd^0/Al_2O_3 + moderator). The activity can only be increased by raising the temperature.

temperature to compensate for the increasing heat production with increasing starting material load.

For catalysts with temperature profiles corresponding to those in Figures 14 B and 14 C, laboratory and pilot plant results are largely consistent, i.e., they are also valid for the industrial multitube reactor, as the pilot plant reactor behaves like a single tube from the multitube reactor.

Fluidized beds behave similarly to liquid bubble columns; i.e., results that are valid for a large industrial scale are only obtained in beds with a diameter exceeding 300 mm. The expen-

diture associated with this requirement in the pilot plant is considerable. Due to lower back mixing the selectivity data obtained in the narrower laboratory fluidized beds are generally considerably higher than on the industrial scale.

Influence of Homogeneous Gas-Phase Oxidations. As a rule, homogeneous gas-phase oxidations generally play only a minor role, if any, in heterogeneously catalyzed gas-phase oxidation, because for safety reasons the temperature of the gas mixture above the catalyst bed must be kept well below the (sometimes very low) ignition temperature. Owing to the short residence times, a significant oxidation reaction cannot normally occur in this region. In the contact bed no significant oxidation reaction should be able to take place in the homogeneous gas phase because of the very small free gas space. Below the catalyst bed, the gas mixture is normally cooled so rapidly that hardly any oxidation in the homogeneous gas phase can occur.

Safety. Although under normal operating conditions homogeneous gas-phase oxidation does not play an important role, critical states can quickly arise during start-up and shutdown and in the case of faults. Particular attention must be paid to the make-up gas compressors, which drive the make-up gas through the reactor. If the compressor stops, the flow rate of the gas about to enter the reactor decreases so much that the gas mixture in the voluminous head of the reactor, can be ignited by the hot catalyst. Equally critical states can occur if the catalyst falls out of one reactor tube during the reaction or if the flow resistance in the reaction tubes becomes very different. Particularly in processes with high oxygen conversion, ignitible gas mixtures can enter the space downstream from the catalyst bed and can ignite either spontaneously or through the action of catalyst dust. Most plants should have been equipped in the meantime with safety devices which render the plant inert in critical areas – independent of service personnel – or bring it to a safe state in another way.

6. References

1. K. Weissermel, H.-J. Arpe: *Industrial Organic Chemistry,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1997.
2. M. Sittig: *Combining Oxygen and Hydrocarbons for Profit,* Gulf Publishing Company, Houston 1962.
3. G. Scott: *Atmospheric Oxidation and Antioxidants,* Elsevier Publishing Comp., Amsterdam 1965.
4. R. A. Sheldon, J. K. Kochi: *Metal-Catalyzed Oxidations of Organic Compounds,* Academic Press, New York 1981.
5. *Autoxidation von Kohlenwasserstoffen. Technische Fortschrittsberichte,* vol. 65, Deutscher Verlag für Grundstoffindustrie, Leipzig 1981.
6. N. M. Emanuel, G. E. Zaikov, Z. K. Maizus: *Oxidation of Organic Compounds (Medium Effects in Radical Reactions),* Pergamon Press, Oxford 1984.
7. N. N. Semenov: *O nekoterikh problemakh khimicheskei kinetiki i reaktsionnoi sposobnosti,* Publ. Akad. Nauk SSSR, Moscow 1958.
8. N. N. Semenov: *Some Problems of Chemical Kinetics and Reactivity,* Pergamon Press, London 1959.
9. N. N. Semenov: *Einige Probleme der chemischen Kinetik und Reaktionsfähigkeit, (Freie Radikale und Kettenreaktionen),* Akademie-Verlag, Berlin 1961.
10. V. Ya. Shtern: *The Gas-Phase Oxidation of Hydrocarbons,* Pergamon Press, Oxford 1964.
11. N. M. Emanuel (ed.): *The Oxidation of Hydrocarbons in the Liquid Phase,* Pergamon Press, Oxford 1965.
12. Catalytica Associates, Selective Catalytic Oxidation of Hydrocarbons: A Critical Analysis, Study no. 1077, Santa Clara, Calif., Oct. 1977.
13. F. R. Mayo (ed.): "Oxidation of Organic Compounds", vol. I – III, *Adv. Chem. Ser.* 1968 no. 75, 76, 77.
14. E. T. Denisov, N. I. Mitskevich, V. E. Agabekov: *Liquid-Phase Oxidation of Oxygen-Containing Compounds,* Consultants Bureau, New York 1977.
15. R. Schoellner: *Die Oxidation organischer Verbindungen mit Sauerstoff,* Akademie-Verlag, Berlin 1964.
16. R. A. Sheldon: "Catalytic Oxidations in the Manufacture of Fine Chemicals," in G. Centi, F. Trifiro (eds.): *New Developments in Selective Oxidation,* Elsevier Science Publ., Amsterdam 1990.
17. *Ullmann,* 3rd ed., **13,** 95 – 109.

18. *Houben-Weyl,* 4th ed., **4,** part 1 A and 1 B.
19. *Organic Chemistry, A Series of Monographs,* Academic Press, New York, K. B. Wiberg (ed.): "Oxidation in Organic Chemistry," **vol. 5 a,** 1965; W. S. Trahanovsky (ed.): "Oxidation in Organic Chemistry," **vol. 5 b,** 1973.
20. R. L. Augustine (ed.): *Technics and Application in Orgnanic Chemistry,* Marcel Dekker, New York, "Oxidation," **vol. I,** 1969; "Oxidation,", **vol. II,** 1971.
21. Catalytica Associates, Advances in Multimetallic Catalysis, Study no. 4155, Santa Clara, Calif., May 1982.
22. Catalytica Associates, A Practical Guide to Catalyst Testing, Study no. 4186 TE, Mountain View, Calif., 1987.
23. W. Bartknecht: *Explosionen–Ablauf und Schutzmaßnahmen,* Springer-Verlag, Berlin 1980.
24. O. L. Magelli, Ch. S. Sheppard in [37], vol. I, pp. 1 –104.
25. R. D. Mair, R. T. Hall, in [37], vol. II, pp. 535 – 636.
26. R. Hiatt in [37], vol. II, pp. 1 – 151.
27. *Houben-Weyl,* 4th ed., **8,** 3 – 74.
28. W. Berthold, U. Löffler: *Lexikon sicherheitstechnischer Begriffe in der Chemie,* Verlag Chemie, Weinheim 1981.
29. W. Bartknecht: *Staubexplosion–Ablauf und Schutzmaßnahmen,* Springer-Verlag, Berlin 1987.
30. H. Freytag: *Handbuch der Raumexplosionen,* Verlag Chemie, Weinheim 1965.
31. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* 2nd ed., Deutscher Eichverlag, Braunschweig 1978.
32. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* **5th suppl.,** Deutscher Eichverlag, Braunschweig 1980.
33. D. R. Stull: "Fundamentals of Fire and Explosion," *AICHE Monogr. Ser.* **73** (1977) no. 10.
34. L. Bretherick: *Handbook of Reactive Chemical Hazards,* Butterworth, London 1975.
35. H. F. Coward, G. W. Jones: "Limits of Flammability of Gases and Vapors," *Bull U.S. Bur. Mines* 1952, no. 503.
36. K. N. Palmer: *Dust Explosions and Fires,* Chapman and Hall, London 1973.
37. D. Swern (ed.): *Organic Peroxides,* Wiley-Interscience, New York, vol. I, 1970; vol. II, 1971; vol. III, 1972.
38. *Landolt-Börnstein,* 6th ed., **2,** part 2 b, 1 ff.
39. J. Smidt et al., *Angew. Chem.* **71** (1959) 176.
40. R. Jira in B. Cornils, W. A. Hermann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* VCH Verlagsgesellschaft, Weinheim 1996, p. 374 – 393.
41. P. A. Jacobs in G. Ertl, H. Knözinger (eds.): *Handbook of Heterogeneous Catalysis,* VCH Verlagsgesellschaft, Weinheim 1997, p. 2284 – 2295.
42. Catalysis Institute Nowosibirsk, DE – OS 2 516 244, 1975.
43. D. J. Hucknall: *Selective Oxidation of Hydrocarbons,* Academic Press, London 1974.
44. R. Higgins, P. Hayden in: *Catalysis,* vol. I, Chemical Society, London 1977.
45. W. D. Mross, J. Jira in J. Falbe, U. Hasserodt (eds.): *Katalysatoren, Tenside und Mineralöladditive,* G. Thieme-Verlag, Stuttgart 1978.
46. M. V. Twigg (ed.): *Catalyst Handbook,* Wolfe Publ. Co., 1989, pp.
47. Shell, US 4 012 425, 1972 (R. P. Nielsen, J. H. LaRochelle).
48. W. M. H. Sachtler, C. Backx, R. A. van Santen, *Catal. Rev. Sci. Eng.* **23** (1981) 127 – 149.
49. Shell, EP 0 266 015, 1988 (A. M. Lauritzen).
50. E. Cremer: *Advances in Catalysis,* vol. 7, Academic Press, New York 1955.
51. L. K. Doraiswamy, D. G. Tajbl, *Catal Rev. Sci. Eng.* **10** (1974) 177 – 219.
52. J. Nelles et al., *Chem. Techn. (Leipzig)* **30** (1978) 441 – 446, 555 – 559.
53. H. Pichler, F. Obenaus, *Brennst. Chem.* **45** (1964) 97 –103; **46** (1965) 258 – 264.
54. Wataru Ando (ed.): *Organic Peroxides,* J. Wiley & Sons, New York1992.

Oxide Ceramics → Ceramics, Advanced Structural Products
Oxide Ceramics → Superconductors
Oxirane → Ethylene Oxide

Oxocarboxylic Acids

Glyoxylic Acid is a separate keyword.

Franz Dietrich Klingler, Boehringer Ingelheim, Federal Republic of Germany (Chaps. 1, 2 , 4 – 6)

Wolfgang Ebertz, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 3)

1. Introduction 1
2. Pyruvic Acid 1
3. Acetoacetic Acid and Derivatives . . . 2
3.1. Properties 2
3.2. Production 4
3.3. Uses and Economic Aspects 5
3.4. Toxicology and Environmental Protection 5
4. Levulinic Acid 5
5. Acetonedicarboxylic Acid 6
6. Other Oxocarboxylic Acids 7
7. References 7

1. Introduction

Oxocarboxylic acids are classified according to the relative positions of the oxo and carboxyl groups as α-, β-, γ-, and δ-oxocarboxylic acids. Glyoxylic acid [*298-12-4*] is the simplest α-oxocarboxylic acid (an aldehyde carboxylic acid) (→ Glyoxylic Acid).

Oxocarboxylic acids, particularly α-oxocarboxylic acids [1], are important metabolites in biochemical cycles. They form the link between α-amino acid, carbohydrate, and fat metabolism [2], [3]. Whereas the free α-oxocarboxylic acids, the most important example of which is pyruvic acid, are relatively stable, the β-oxocarboxylic acids can be isolated in the free state only by taking special precautions. They decompose by decarboxylation to give the corresponding ketones. The γ-oxocarboxylic acids, with levulinic acid as the simplest example, are again stable and do not decarboxylate.

Very specific processes are required for the industrial production of oxocarboxylic acids. A general method for the production of oxocarboxylic acids is the hydrolysis of fatty acids with concentrated sulfuric acid or other hydrolytic reagents. α-Oxocarboxylic acids can be produced by treatment of α-amino acids with trifluoro acetic anhydride [*407-25-0*]. 2-Trifluorooxazolones are formed as intermediates which hydrolyze to the corresponding α-oxocarboxylic acids [4]. The carbonylation of alkyl halides with cobalt carbonyls (typically carried out at ca. 5 MPa and 100 °C) leads to α-oxocarboxylic acids [5]. General methods for the sythesis of oxocarboxylic acids have been described by Robinson [6] and Stetter [7].

2. Pyruvic Acid

Pyruvic acid [*127-17-3*], 2-oxopropanoic acid, pyroracemic acid, α-ketopropionic acid, $H_3C{-}CO{-}COOH$, M_r 88.06, is the most important α-oxocarboxylic acid. It plays a central role in energy metabolism in living organisms [8]. During exertion, pyruvic acid is formed from glycogen in the muscle and reduced to lactic acid [*79-33-4*]. In the liver, pyruvic acid can be converted into alanine [*56-41-7*] by reductive amination. Pyruvic acid was discovered and first described in 1835 by Berzelius [9].

Physical Properties. Pyruvic acid is a colorless liquid with an odor similar to that of acetic acid. Its most important physical data may be summarized as follows:

Density (20 °C)	1.268 g/cm^3
mp	13.6 °C
bp	
101.31 kPa	165 °C (decomp.)
1.33 kPa	57.9 °C
0.13 kPa	21.4 °C
Refractive index (25 °C)	1.4259
Flash point (DIN 51 758)	91 °C
Ignition temperature (DIN 51 794)	305 °C

Pyruvic acid is totally miscible with water, ethanol, and ether. Pyruvic acid exists in the keto form; the enol form has not been detected [10].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_313

Chemical Properties. Pyruvic acid reacts as both an acid and a ketone. It forms, for example, oximes, hydrazones, and salts. 4,5-Dioxo-2-methyltetrahydrofuran-2-carboxylic acid [*24891-71-2*] (**1**) is formed from pyruvic acid either slowly on standing or more quickly under acid catalysis [11].

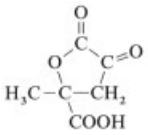

1

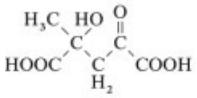

2

On standing in aqueous solution, pyruvic acid polymerizes to higher molecular mass products via the dimeric ketoglutaric acid (**2**) [*19071-44-4*] and the trimeric aldol product [12], [13].

Like all 2-oxo acids, pyruvic acid eliminates carbon monoxide on treatment with concentrated sulfuric acid [14].

Oxidation of pyruvic acid gives acetic or oxalic acid [*144-62-7*] and carbon dioxide, depending on the conditions [15]. Lactic acid is obtained by reaction with reducing agents [1].

Reaction of α-amino acids with pyruvic acid gives, besides carbon dioxide, alanine [*56-41-7*] (transamination reaction) and the corresponding aldehyde with one carbon atom less [16]. Alanine is also obtained by reductive amination of pyruvic acid [1]. Phenylethylamines react with pyruvic acid to form the corresponding tetrahydroisoquinolines via the Bischler – Napieralski reaction [17]. Reaction with *o*-phenylenediamines gives quinoxalinols [18]. In a similar reaction the corresponding hydroxypteridines are obtained from 4,5-diaminopyrimidines and pyruvic acid [19]. Pyruvic acid reacts with aldehydes to form the corresponding α-keto-γ-hydroxy acids, which then cyclize to butyrolactone derivatives [1]. Friedel – Crafts type reactions of aromatic compounds with pyruvic acid yield diarylpropionic acids. These compounds have achieved a certain degree of importance because they provide a good route to 1,1-diarylethylenes by dehydration and decarbonylation [15], [20].

Production. On an industrial scale, pyruvic acid is produced by *dehydration and decarboxylation of tartaric acid* [*87-69-4*] [21]. In this process, pyruvic acid is distilled from a mixture of tartaric acid and potassium and sodium hydrogen sulfates at 220 °C. The crude acid obtained (ca. 60 %) is then distilled in vacuum. The reaction temperature can be lowered to 160 °C by adding ethylene glycol [*107-21-1*] [22]. Pyruvic acid can also be obtained by the *gas-phase oxidation of lactic acid* [23], but this process has not been successful industrially. In contrast, *microbial oxidation of* D-*lactic acid* by a new process results in high yields [24]. Microbial oxidation of 1,2-propanediol [*57-55-6*] to pyruvic acid has also been described [25]. Another process describes the hydrolysis of 2,2-dihalopropionic acids to pyruvic acid [26]. A process for the oxidation of methylglyoxal [*78-98-8*] with halogens has been published [27].

Uses. Pyruvic acid is used mainly as an intermediate in the synthesis of pharmaceuticals. It is also employed in the production of crop protection agents, polymers, cosmetics, and foods.

Storage and Quality Specifications. Pyruvic acid is stored and transported in tightly closed polyethylene containers. It can be kept for only a limited period and must therefore be stored in refrigerated areas at a maximum of 10 °C. At higher temperature, explosion can occur through spontaneous self-condensation [28]. The concentration of the commercial product is determined acidimetrically and decreases by ca. 1 % per month during storage.

Toxicology. Pyruvic acid has a corrosive effect and irritates the eyes, skin, and respiratory passages.

3. Acetoacetic Acid and Derivatives

3.1. Properties

The physical properties of the most important acetoacetic esters and aromatic amides are listed in Table 1.

Chemical Properties. Acetoacetic acid and its derivatives exist in the tautomeric keto – enol forms.

Table 1. Physical properties of important acetoacetic esters and amides*

	R	CAS registry number	M_r	*mp*, °C	*bp*, °C	d^{20}, kg/m³	Flash point, °C
1	$-OCH_3$	[*105-45-3*]	116.1	−35	170	1075	67
2	$-OC_2H_5$	[*141-97-9*]	130.1	−43	180	1030	71
3	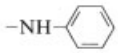	[*102-01-2*]	177.2	82		1260	185
4	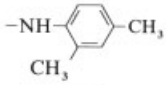	[*97-36-9*]	205.0	88		1240	171
5	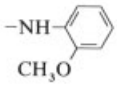	[*92-15-9*]	207.2	86		1130	162
6	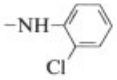	[*93-70-9*]	211.6	104		1440	176
7**	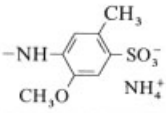	[*72705-22-7*]	318.3				

* $CH_3-\overset{O}{\overset{\|}{C}}-CH_2-\overset{O}{\overset{\|}{C}}-R$

** Hygroscopic.

The enol content depends on the type of derivative and the polarity of the solvent. In nonpolar solvents, stabilization of the enol form occurs by formation of a hydrogen bond [29].

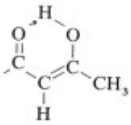

Acetoacetic acid is a strong organic acid that decomposes rapidly to acetone and carbon dioxide when heated above its melting point of 35–37 °C.

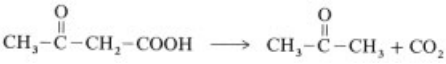

Because of this instability, acetoacetic acid as such is not used in organic synthesis. Its derivatives, however, in particular esters and amides, are important intermediates because of their higher stability in comparison to the free acid.

β-Ketoacids and their derivatives have three reactive functional groups, which allow the synthesis of many important compounds. The carbonyl, methylene, and carboxylic groups take different reaction paths.

The tendency of β-dicarbonyl compounds to form the enol results in high reactivity of the C–H bond of the *methylene group*. Hydrogen can be abstracted easily from the methylene group by strong alkali. This reaction is usually carried out with acetoacetic esters and alkoxides. The resulting salt of the ester can be alkylated or acylated.

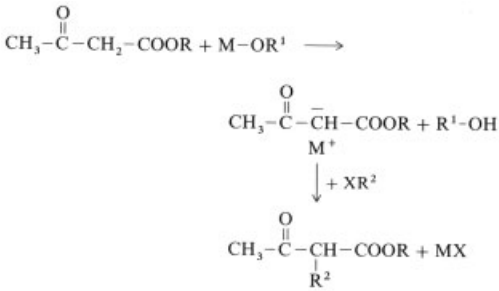

The reactivity of the remaining C–H bond is still high enough to undergo a second substitution to form disubstituted acetoacetic esters. When treated with diluted alkali these substituted esters are decarboxylated to form substituted acetones [30].

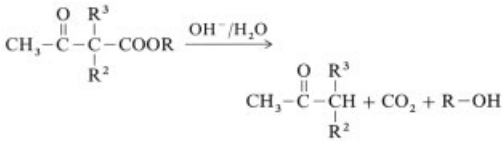

The cleavage of substituted acetoacetic esters under strongly basic conditions leads to carboxylic acids (retro-Claisen condensation) [30].

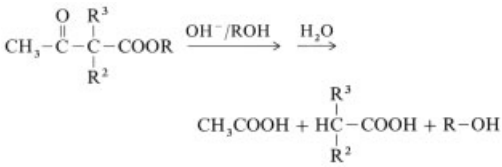

The *carbonyl group* is the second reactive group in acetoacetic compounds. Under conditions that stabilize the keto form, its reactivity is comparable to that of other ketones. Of particular interest are hydrazines and other difunctional compounds, which react with acetoacetic esters to yield heterocyclic compounds (e.g., 3-methyl-5-pyrazolones):

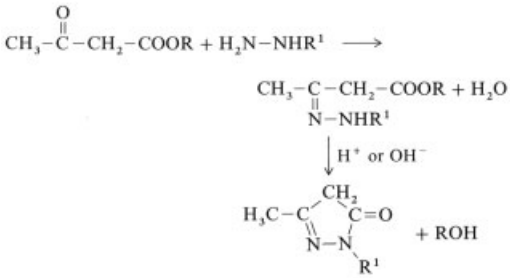

Reaction with ureas leads to uracil derivatives and with hydroxylamines leads to isoxazoles. The reactivity of the *ester group* is comparable to that of other carboxylic esters. Acetoacetic esters are used as starting materials in the synthesis of acetoacetamides.

3.2. Production

Diketene is used as starting material in the production of acetoacetic esters and amides. The use of diketene makes strict safety regulations necessary. Enrichment of diketene in the reaction mixture must be avoided, therefore reaction temperature, flow, and pressure must be carefully controlled. A temporary decrease in catalytic activity (temperature drop, dosage inaccuracies) or inaccuracies in the addition of reactants can result in the strong rapid exothermic decomposition of enriched diketene. The reaction mixture or vapor must also be kept from contact with stored diketene.

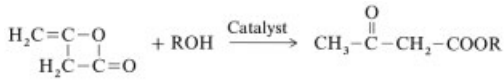

Production of Acetoacetic Esters. The methyl and ethyl esters are the preferred acetoacetic esters for industrial use. They are produced in continuous processes, usually from *crude diketene* and methanol or ethanol. Basic catalysts such as amines or their acetates are commonly used, although strong acids also have catalytic action. The exothermic reaction is carried out at 100 – 140 °C and 1.5 – 2 bar. The residence time in the reactor must be short (0.5 – 2 h) to avoid formation of condensation products of acetoacetic esters (dipyrones).

Since crude diketene always contains a certain amount of acetic anhydride, acetic acid and acetic esters are formed as byproducts. Together with excess alcohol and acetone (the latter formed in side reactions from diketene) acetic acid and esters constitute a low-boiling fraction in the crude acetoacetic ester, which must be removed by distillation. Crude diketene contains oligomers and polymers of ketene. These substances partially react with the alcohol and form the high-boiling fraction in the crude product. After purification of the ester by distillation, they remain as a partially solid residue, which is combusted.

Production of Acetoacetamides. *Distilled diketene* is used for the production of acetoacetamides. The reaction is carried out in aqueous solution or suspension or in dilute acetic acid.

Because of the basic reactant amine, no catalyst is needed.

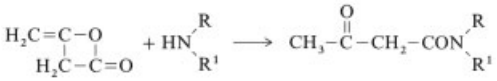

Acetoacetamide is produced by reacting a dilute aqueous ammonia solution with diketene and is used in further reactions (e.g., to pyrazolones) as an aqueous solution. Aromatic amides of acetoacetic acid, the so-called arylides, crystallize from the cooled reaction mixture, and are separated and dried. When no diketene is available, acetoacetic esters are used as starting materials for the production of acetoacetamides. The alcohol released from the amidation reaction must be recycled.

3.3. Uses and Economic Aspects

Acetoacetic esters are important starting materials in the synthesis of intermediates for drugs, dyes, and agrochemicals. The worldwide installed production capacity of acetoacetic esters is more than 80 000 t/a. More than 80 % of all esters produced are methyl and ethyl acetoacetates (**1** and **2**, respectively, in Table 1).

The main application of *acetoacetamides* is in the production of dyes and organic pigments. The most important amides in pigment production are acetoacetanilide (acetoacetaminobenzene, **3**) and acetoacet-*m*-xylidide (4-acetoacetamino-1,3-dimethylbenzene, **4**), followed by acetoacet-*o*-anisidide (2-acetoacetamino-1-methoxybenzene, **5**) and acetoacet-*o*-chloroanilide (2-chloro-1-acetoacetaminobenzene, **6**). The so-called cresidine (3-acetoacetamino-4-methoxytoluene-6-sulfonic acid, ammonium salt, **7**) is an intermediate in the production of reactive dyes.

Producers. The main producers of acetoacetic compounds are BASF (Germany), BP Chemicals (UK), Colour Chem. (India), DAICEL (Japan), Eastman Kodak (United States), Hoechst (Germany), Lonza (United States, Switzerland, Germany), Nippon Gosai (Japan), and Wacker Chemie (Germany).

3.4. Toxicology and Environmental Protection

Acetoacetic esters have a low acute toxicity. The LD_{50} values are 3230 mg/kg (rat) [31] for methylacetoacetate and 3980 mg/kg (rat) [32] for ethylacetoacetate. Chronic diseases caused by acetoacetic esters have not been reported. The methyl ester causes reversible irritation of skin and mucous membranes [33]. Acetoacetic esters are completely degradable in biological wastewater treatment and have low toxicity for fish and bacteria.

The toxicity and degradability of *acetoacetamides* are similar to those of the esters.

To avoid negative environmental effects, the effluents of acetoacetic compound production must be purified in a biological wastewater treatment plant. In handling and shipping acetoacetic esters the rules for products with low toxicity must be obeyed:

RID/ADR	3/32 c
ADNR	3/4/K 3
ICAO/IATA	not classified

4. Levulinic Acid

Levulinic acid [*123-76-2*], β-acetylpropionic acid, γ-ketovaleric acid, 4-oxopentanoic acid, $H_3C-CO-CH_2-CH_2-COOH$, M_r 116.11, is the simplest and most important γ-oxocarboxylic acid. It was first described and named levulinic acid by Grote and Tollens in 1875 [34]. Only since 1960 has levulinic acid gained commercial importance [35], [36].

Physical Properties. Levulinic acid forms colorless crystals that are readily soluble in water, ethanol, and diethyl ether. Important physical properties are given below:

mp	37 °C
bp	
(101.31 kPa)	245 – 246 °C
(1.33 kPa)	137 – 139 °C
Specific density (20 °C)	1.14
Refractive index (25 °C)	1.4396
Flash point	138 °C

Chemical Properties. Levulinic acid reacts both as a ketone and as a carboxylic acid. Angelica lactone [*591-12-8*] is formed by elimination of water on prolonged heating [37], [38]. γ-Valerolactone [*108-29-2*] is formed by catalytic hydrogenation of levulinic acid [39]. Condensation reactions with aldehydes can occur at the α-, β-, or δ-carbon atoms, depending on conditions [35]. Levulinic acid is the starting compound in the synthesis of many heterocycles. 5-Methyl-2-pyrrolidone [*108-27-0*] is formed by reductive amination [40], and 6-keto-3-methyl-1,4,5,6-tetrahydropyridazone by reaction of levulinic acid with hydrazine [41].

Production. Levulinic acid is produced industrially from sugar polymers such as cellulose or starch via the monomeric hexoses [36] according to many patented processes. The reaction is usually acid catalyzed: D-glucose [*50-99-7*] is formed first and is then isomerized enzymatically to D-fructose [*57-48-7*]. D-Fructose is subsequently converted to hydroxymethylfurfural [*67-47-0*], an intermediate that reacts further to form levulinic acid. The classical levulinic acid synthesis, the treatment of D-fructose with hydrochloric acid, is also being used [36].

A newer patented process is based on the use of petrochemical raw materials. Levulinic acid is obtained by ozonolysis of unsaturated hydrocarbons in a relatively sophisticated process [42].

Uses. Levulinic acid is used as a raw material for the synthesis of pharmaceuticals [43], [44], for the production of plasticizers [45], and as an auxiliary in electroplating [45].

The calcium salt is used in calcium therapy [46], and levulinic acid esters are used in the production of cosmetics [47].

Storage and Quality Specifications. Levulinic acid is stored and transported in tightly closed polyethylene containers, protected from light. It is stable for long periods at room temperature.

Toxicology. The LD_{50} (oral, rat) is 1850 mg/kg. Levulinic acid can cause acid burns, and concentrated solutions can cause irritation of the skin and mucous membranes [36].

5. Acetonedicarboxylic Acid

Acetonedicarboxylic acid [*542-05-2*], 3-oxoglutaric acid, β-ketoglutaric acid, 3-oxopentane dicarboxylic acid, $HOOC-CH_2-CO-CH_2-COOH$, M_r 146.10, was first isolated and described by VAN. PECHMANN [48], [49].

Physical Properties. Acetonedicarboxylic acid forms colorless crystals that are readily soluble in water and ethanol, and sparingly soluble in trichloromethane and diethyl ether. It melts with decomposition at 138 °C. The dissociation constants (in 0.01 M solution) are $K_1^{25} = 4.68 \times 10^{-4}$ and $K_2^{25} = 5.37 \times 10^{-5}$ [50].

Chemical Properties. On heating above its melting point, acetonedicarboxylic acid decomposes into carbon dioxide and acetone [49]. When acetonedicarboxylic acid is warmed in water, a two-step decomposition occurs with acetoacetic acid [*541-50-4*] as an intermediate that decomposes further to acetone and carbon dioxide [50]. The decomposition is catalyzed by metal ions and protons [51]. Phenols react with acetonedicarboxylic acid in the presence of dehydrating agents to form coumarin derivatives [52–54].

Production. Acetonedicarboxylic acid is produced from citric acid [*77-92-9*] by many industrial processes that differ only slightly [55], [56]. Citric acid is treated with oleum and reacts to yield acetonedicarboxylic acid via decarbonylation and dehydration.

Other possible production methods include reaction of acetone with carbon dioxide [57], [58], oxidation of citric acid with chlorosulfuric acid [59], or reaction of ketene with phosgene [60].

Uses. Acetonedicarboxylic acid is an important starting material in the production of pharmaceutically active alkaloids [61–63]. It is also suitable as a stabilizer for natural fats and oils [64], [65].

Storage and Quality Specifications. Large quantities of acetonedicarboxylic acid can be stored for short periods in tightly closed polyethylene containers in refrigerated areas. The acid must not be stored in metal containers. Laboratory-scale quantities of the pure substance can be kept for longer periods over phosphorus pentoxide in a desiccator [66]. The commercial product has a concentration of ca. 98 %; the purity is determined by titration.

Toxicology. Acetonedicarboxylic acid can cause irritation and acid burns to the eyes, respiratory passages, and skin.

6. Other Oxocarboxylic Acids

Oxaloacetic acid [*328-42-7*] (2-oxosuccinic acid), an important intermediate in the tricarboxylic acid cycle, is known only in the form of its *cis* and *trans* enol tautomers, hydroxymaleic acid and hydroxyfumaric acid, respectively [67], [68]. Its diethyl ester, obtained by Claisen condensation of diethyl oxalate [*95-92-1*] and ethyl acetate [69], is used in the production of higher α-oxocarboxylic acids (by decarboxylation of the condensation product) and higher carboxylic acids (retro-Claisen condensation).

α-Oxoglutaric acid [*328-50-7*], which plays an important role in amino acid metabolism [70], is obtained by condensation of diethyl succinate [*123-25-1*] with diethyl oxalate [*95-92-1*] [71], [72]. It was used as the starting material for the production of L-glutamic acid [*56-86-0*] in a microbial process [73].

7. References

1. A. J. L. Cooper, J. Z. Ginos, A. Meister, *Chem. Rev.* **83** (1983) 321.
2. A. E. Martell, *Acc. Chem. Res.* **22** (1989) 115.
3. M. Walser, J. R. Williamson: *Metabolism and Clinical Implications of Branched Chain Amino- and Ketoacids,* Elsevier, Amsterdam 1981.
4. F. Weygand, W. Steglich, H. Tanner, *Justus Liebigs Ann. Chem.* **658** (1962) 128.
5. B. Bell, H. Chrobaczek, W. Kohl, *Chem. Ztg.* **109** (1985) 167.
6. R. Robinson, *J. Chem. Soc.* 1930, 745; 1934, 1543.
7. H. Stetter, *Angew. Chem.* **88** (1976) 695.
8. K. Schreiber: *Die Brenztraubensäure und ihr Stoffwechsel,* Editio Cantor, Aulendorf 1956.
9. J. J. Berzelius, *Ann. Phys.* **36** (1835) 1.
10. A. Schellenberger, K. Winter, *Chem. Ber.* **92** (1959) 793.
11. L. Wolff, *Justus Liebigs Ann. Chem.* **317** (1901) 1.
12. H. Goldfine, *Biochim. Biophys. Acta* **40** (1960) 557.
13. A. Schellenberger, E. Podany, *Chem. Ber.* **91** (1958) 1781.
14. A. Bistrzycki, B. v. Siemiradzki, *Ber. Dtsch. Chem. Ges.* **39** (1906) 58.
15. S. Patai, S. Dayagi, *J. Chem. Soc.* 1958, 3058.
16. R. M. Herbst, L. L. Engel, *J. Biol. Chem.* **107** (1934) 505.
17. G. Hahn, A. Hansel, *Ber. Dtsch. Chem. Ges.* **71** (1938) 2163.
18. O. Hinsberg, *Justus Liebigs Ann. Chem.* **237** (1887) 327.
19. G. B. Elion, G. H. Hitchings, P. B. Russel, *J. Am. Chem. Soc.* **72** (1950) 78.
20. Bayer, DE 2 830 953, 1978 (W. Meyer, H. Rudolf, E. Cleur, E. Schoenhals).
21. J. W. Howard, W. A. Fraser, *Org. Synth. Coll. Vol.* **1** (1932) 475.
22. J. D. Riedel, DE 281 902, 1913.
23. C. H. Boehringer Sohn, DE 523 190, 1931 (F. Zumstein).
24. BASF, EP-A 0 313 850, 1988 (B. Cooper).
25. Y. Izumi, Y. Matsumura, Y. Tani, H. Yamada, *Agric. Biol. Chem.* **46** (1982) 2673.
26. Dow Chemical, US 3 524 880, 1966 (L. H. Lee, D. E. Ranck).
27. BASF, DE-OS 3 219 355, 1982 (U. R. Samel, L. Hupfer).
28. *Sichere Chemiearbeit* **29** (1977) 87.
29. S. Gould: *Mechanisms and Structure in Organic Chemistry,* 2nd ed., Henry Holt and Company, New York 1959.
30. L. F. Fieser, M. Fieser: *Organische Chemie,* 2nd ed., Verlag Chemie, Weinheim 1972, p. 535.
31. H. F. Smyth et al.: *J. Ind. Hyg. Toxicol.* **30** (1948) 63.
32. H. F. Smyth, C. P. Carpenter, C. S. Weil, *J. Ind. Hyg. Toxicol.* **31** (1949) 60.
33. *Food Cosmet. Toxicol.* **16** (1978) 815.
34. A. v. Grote, B. Tollens, *Justus Liebigs Ann. Chem.* **175** (1875) 181.
35. M. Kitano, F. Tanimoto, M. Okabayashi, *CEER Chem. Econ. Eng. Rev.* **7** (1975) 25.

36. V. Sunjic, J. Horvat, B. Klaic, S. Horvat, *Kem. Ind.* **33** (1984) 593.
37. J. Thiele, R. Tischbein, E. Lossow, *Justus Liebigs Ann. Chem.* **319** (1901) 180.
38. Newport Ind., US 2 761 869, 1956; US 2 809 203, 1959 (R. H. Leonard).
39. Quaker Oats, US 2 786 852, 1957 (A. P. Dunlop, J. W. Maden).
40. L. Frank, W. R. Schmitz, B. Zeidman, *Org. Synth. Coll. Vol.* **3** (1955) 328.
41. W. G. Overend, L. F. Wiggins, *J. Chem. Soc.* 1947, 239.
42. BP Chem., GB 1 596 651, 1981 (J. Russel, C. G. Gasser).
43. Edelfettwerke W. Schlueter, Glyco Iberica, DE-OS 2 360 795, 1975.
44. Henkel, DE-OS 2 533 101, 1975 (H. Moeller, H. Schnegelberger, R. Osberghaas).
45. Leningrad Technological Institute, SU 722 943, 1980 (Y. M. Postolov et al.).
46. A. R. Carnes, D. D. Mann, DE-OS 2 830 636, 1979.
47. Y. Hikotaro, JP 73 43 178, 1973.
48. H. v. Pechmann, *Ber. Dtsch. Chem. Ges.* **17** (1884) 929, 2542.
49. H. v. Pechmann, *Justus Liebigs Ann.* **261**(1891) 155.
50. R. W. Hay, K. N. Leong, *J. Chem. Soc. A* 1971, 3639.
51. J. E. Prue, *J. Chem. Soc.* 1952, 2331.
52. H. P. Kansara, N. H. Shah, *J. Univ. Bombay Sci.* **17 A** (1948) 47.
53. K. A. Thakar, *J. Indian Chem. Soc.* **40** (1963) 397.
54. V. M. Dixit, A. N. Kanakudati, *J. Indian Chem. Soc.* **28** (1951) 323.
55. *Beilstein,* **3,** 789; **3 (1),** 275; **3 (2),** 482; **3 (3),** 1369.
56. R. Adams, H. M. Chiles, C. F. Rassweiler, *Org. Synth. Coll. Vol.* **1** (1932) 9.
57. Mitsui Toatsu Chem Inc., JP-Kokai 75 71 625, 1975 (M. Kawamata, H. Tanabe).
58. Montedison, DE-OS 2 429 627, 1974 (E. Alneri, G. Bottaccio, V. Carletti, G. Lana).
59. C. H. Boehringer Sohn, DE 1 160 841, 1961 (F. Gerner).
60. Akzo, DE-OS 2 409 342, 1974 (N. Heyboer).
61. R. C. Menzies, R. Robinson, *J. Chem. Soc.* **125** (1924) 2163.
62. Sadolin + Halmblad, GB 791 770, 1958 (N. Elming, P. Nedenskov).
63. C. Schöpf, G. Lehmann, *Justus Liebigs Ann. Chem.* **518** (1935) 1.
64. Secretary of Agriculture, US 2 605 186, 1951 (A. W. Schwab, H. A. Moser, C. D. Evans).
65. A. Mosca, IT 494 153, 1951.
66. E. O. Wilg, *J. Phys. Chem.* **32** (1928) 961.
67. Merck Index 10th ed., 6782, Rahway, N.J. 1983.
68. *Beilstein,* **3,** 777; **3 (1),** 272; **3 (2),** 478; **3 (3),** 1359; **3 (4),** 1808.
69. Höchster Farbenwerke, DE 43 847, 1887.
70. K. Heyns, *Angew. Chem.* **61** (1949) 474.
71. L. Friedman, E. Kosower, *Org. Synth. Coll. Vol.* **3** (1955) 510.
72. E. M. Bottorff, L. L. Moore, *Org. Synth. Coll. Vol.* **5** (1973) 687.
73. Int. Minerals and Chem. Corp., US 2 933 434, 1960 (R. C. Good).

Oxo Synthesis

HELMUT BAHRMANN, Celanese GmbH Ruhrchemie, Oberhausen, Federal Republic of Germany

HANSWILHELM BACH, Celanese GmbH Ruhrchemie, Oberhausen, Federal Republic of Germany

1. Introduction 1
2. Theory 2
3. Industrial Applications 3
3.1. Rhodium Processes 4
3.2. Cobalt Processes 7
4. References 7

1. Introduction

Olefins react with synthesis gas (carbon monoxide and hydrogen) in the presence of homogeneous catalysts to form aldehydes containing an additional carbon atom. This hydroformylation of olefins, known as oxo synthesis, was discovered in 1938 by ROELEN of Ruhrchemie [1], [2–8, 10].

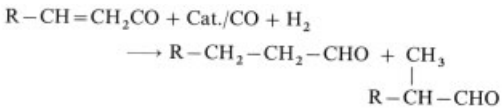

The most important oxo products are in the range $C_3 - C_{19}$, [2], [8]; with a share of roughly75 % butanal is by far the most significant. In 1997 the total worldwide oxo production capacity for aldehydes and alcohols was 6.5×10^6 t/a. Table 1 gives an overview of the known facilities and capacities.

Up until the mid-1970s only the cobalt-based catalysts were important industrially. The situation changed in 1974 – 1975 when Union Carbide and Celanese, independently of one another, introduced rhodium-based catalysts on an industrial scale. Since then, cobalt catalysts for the hydroformylation of propene have been replaced in nearly all major plants by the more advantageous rhodium catalysts despite the higher price of the noble metal. This substitution was made possible because the high price of rhodium was offset by cheaper equipment, increased activity, and a generally higher selectivity and efficiency. Efficiency became extremely important following the oil crisis in 1973, which shifted the the cost structure in aldehyde production towards the raw materials. For example, propene and synthesis gas account for about 75 % of the product value [9]. With the introduction of more expensive rhodium catalysts, catalyst reprocessing, which can be carried out batchwise or continuously, became increasingly important [12–22]. The catalyst lifetime can be severely diminished by "extrinsic poisons" such as strong acids, hydrogen cyanide, sulfur, hydrogen sulfide, COS, oxygen, and dienes. The feedstocks must therefore be very pure. Since 1978 [23] increased attention has been paid to "intrinsic deactivation" (i.e., internal decomposition) of the relatively expensive catalyst ligands [12] , [18], [24]. Next to the efficiency, activity, and selectivity of the catalyst, the catalyst lifetime is the most industrially important reaction parameter.

Table 1. Worldwide oxo capacities (1989)

Country	Capacity, 10^3 t/a	Products	Process type ∗
Brazil	120	C_4	1
Bulgaria	20	C_4	5
China	300	C_4	1, 5
Czech Republic	40	C_4	5
Former Soviet Union	580	C_4	5
France	355	$C_4, C_8 - C_{10}, C_{13}$	1, 5
Germany	1695	$C_3 - C_{13}$	1 – 5
India	40	$C_4, C_8 - C_{10}$	1, 4, 5
Italy	85	$C_{12} - C_{15}$	5
Japan	755	$C_4, C_7 - C_{15}$	1, 4, 5
Poland	150	C_4	1
Romania	135	C_4	5
Saudi Arabia	190	C_4	1
South Korea	420	C_4	1, 2
Spain	70	C_4	1
Sweden	220	C_4	1
The Netherlands	200	$C_8 - C_{10}$	5
United Kingdom	150	$C_7 - C_{15}$	4, 5
United States	2500	$C_3 - C_{15}$	1, 4, 5

∗ See Table 2.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_321

2. Theory

The mechanism developed by HECK and BRESLOW [25–27] for the reaction with cobalt catalysts (see → Organometallic Compounds and Homogeneous Catalysis, Chap. 3.2.7.) can be largely applied to hydroformylation with rhodium. However, the introduction of the triphenylphosphine (PPh_3) ligands in the catalyst system brings about critical changes which already affect the course of the reaction by formation of various Rh complexes, which leads to new effects that cannot be explained by the older cobalt model. For example, with the cobalt system the fraction of unbranched aldehyde product (or the *n*/*i* ratio) increases with increasing partial pressure of carbon monoxide, while with rhodium no effect or a lower *n*/*i* ratio is observed [1].

Fundamental mechanistic studies of the industrially important triphenylphosphine rhodium catalyst system are based on the work of G. WILKINSON's group and workers at Union Carbide in the late 1960s [25], [28], [29].

Model studies with the hydride complex $[RhHCO(PPh_3)_3]$, the key compound in the hydroformylation reaction, were the first to provide satisfactory explanations for the different reaction behavior compared to cobalt catalysts.

For example, studies on the molecular mass of this compound in various organic solvents (e.g., benzene) indicated extensive dissociation of this complex. This shows that an equilibrium exists between the various substituted rhodium complexes before the catalytic cycle occurs [28]:

$$\underset{\mathbf{1}}{[RhH(CO)(PPh_3)_3]} \underset{+PPh_3}{\overset{-PPh_3}{\rightleftharpoons}} \underset{\mathbf{2}}{[RhH(CO)(PPh_3)_2]} \underset{+PPh_3}{\overset{-PPh_3}{\rightleftharpoons}} \underset{\mathbf{3}}{[RhH(CO)(PPh_3)]}$$

The predominant catalytic species is thus determined by this equilibrium, which has important consequences when the reaction is carried out on an industrial scale.

According to R. L. PRUETT and J. A. SMITH, under industrial reaction conditions (i.e., at higher temperatures and under CO/H_2 pressure) the above-mentioned equilibrium is also accompanied by the competition of CO and PPh_3 for free coordination sites on the rhodium atom [29]:

$$\underset{\mathbf{1}}{[RhH(CO)(PPh_3)_3]} \underset{-PPh_3}{\overset{+CO}{\rightleftharpoons}} \underset{\mathbf{4}}{[RhH(CO)_2(PPh_3)_2]} \underset{-PPh_3}{\overset{+CO}{\rightleftharpoons}} \underset{\mathbf{5}}{[RhH(CO)_3(PPh_3)]}$$

Each catalytic species is assigned an individual reaction rate and a characteristic product distribution. High phosphorus/rhodium ratios and a lower partial pressure of carbon monoxide favor the type **1** complex, which is assumed to give a high *n*/*i* ratio as a result of steric effects, since it is more highly substituted with phosphine ligands.

Rhodium complexes with three phosphine ligands were also postulated in 1981 by J. D. UNRUH and O. R. HUGHES of Celanese as the basis for a high *n*/*i* selectivity in the presence of particular chelating ligands in combination with certain monophosphine ligands [30]. In contrast, the two basic dissociative and associative mechanisms, proposed by WILKINSON in 1968, are based on rhodium complexes with either one or two phosphines [28]. Both mechanisms start with the key complex **4** but differ as regards the primary reaction step, i.e., coordination of the olefin to the rhodium center.

In the *dissociative mechanism* dissociation of a phosphine ligand from the bis(phosphine) complex **4** is followed by addition of the olefin to the coordinatively unsaturated square-planar complex **6** (Fig. 1). There is very little steric hindrance because the complex only has a single phosphine ligand. In contrast, in the *associative mechanism* the olefin attaches directly to the bis(phosphine) complex **4**. The associative mechanism accounts for the fact that a higher selectivity for the *n*-aldehyde is obtained in the presence of excess phosphine. However, a high *n*/*i* ratio can also be explained by the dissociative mechanism if the coordinatively unsaturated *trans* bis(phosphine) complex **2**, $[RhH(CO)(PPh_3)_2]$ is assumed to be the active catalytic species. WILKINSON [28] had already taken this possibility into consideration, but rejected it (Fig. 2).

As the olefin approaches the trans bis(phosphine) complex (Fig. 2 a) and forms an alkene complex (Fig. 2 b) the substituents on the phosphine ligands impinge on the alkene due to rotation about the Rh – P bonds, whereby

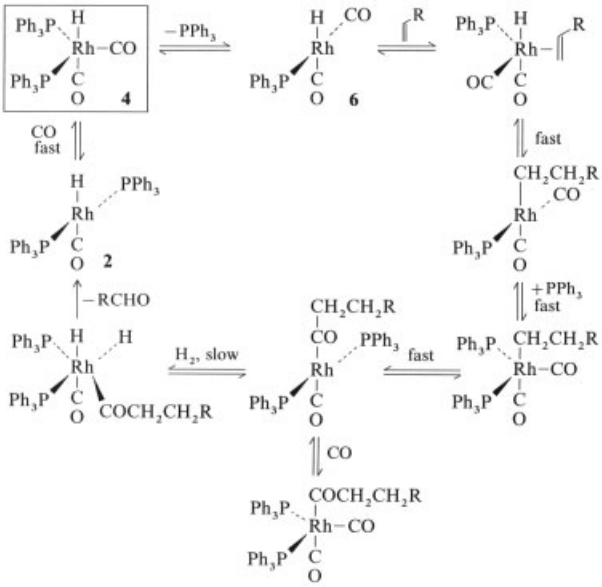

Figure 1. Dissociative mechanism after G. WILKINSON [7]

the alkene groups are bent away from the direction of approach. Additionally, the rotation and sweeping of the phenyl groups of the PPh_3 ligands in the transfer step from the π complex (Fig. 2 b) may favor formation of the *n*-alkyl complex (Fig. 2 c) rather than the *i*-alkyl complex due to minimization of steric hindrance, the *i*-alkyl complex being more crowded than the *n*-alkyl complex.

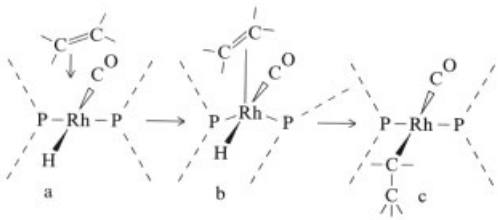

Figure 2. Control of the *n*/*i* ratio by a trans bis(phosphine) rhodium complex [28]

Investigations by Exxon [31], [32], Johnson Matthey [33], Montedison [24], [34], and J. M. BROWN and A. G. KENT [35] clearly support the dissociative mechanism. The active catalytic species are probably unsaturated square-planar rhodium complexes containing one or two coordinated phosphine ligands (Fig. 3). These complexes are in equilibrium with one another, whereby complex **2** produces predominantly the *n*-aldehyde and complex **6** the *i*-aldehyde.

Since 1990 an increasing number of investigations [46–49] have been performed on the water-soluble tris(trisodium-3-sulfophenyl)phosphine-modified [$P(3\text{-}C_6H_4SO_3Na)_3$] rhodium hydroformylation catalyst system used in the Ruhrchemie – Rhône-Poulenc (RCH – RP) process (Fig. 4). A higher *n*/*i* ratio is obtained in the RCH – RP process than in the UCC process, which uses triphenylphosphine as the ligand. This can be explained by a relative shift of the equilibrium between complexes **6** and **4** (see Fig. 1) toward complex **2** because of the higher dissociation energy of the tris(trisodium-3-sulfophenyl)phosphine ligand compared to the triphenylphosphine ligand.

3. Industrial Applications

The hydroformylation reaction is used on an industrial scale to produce aldehydes and alcohols (→ Alcohols, Aliphatic; → Aldehydes, Aliphatic and Araliphatic; → Butanals). These include bulk products as well as high-value specialty products.

The industrially important processes can be divided into five main types. Table 2 gives an

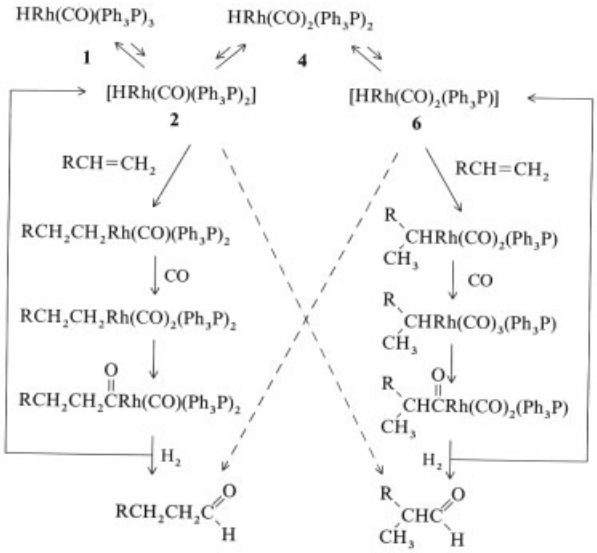

Figure 3. Modified dissociative mechanism [32]

overview of their catalyst systems, reaction conditions, and expected products.

The cobalt catalysts are extensively described in [1–3], [8]. In the meantime their decline in industrial importance has come to an end. Steady improvements in the processing of branched medium- to long-chain olefins (Table 2, Type 5 [50–52]) and inner long-chain olefins (Table 2, Type 4) have strengthened the position of these older cobalt-based processes.

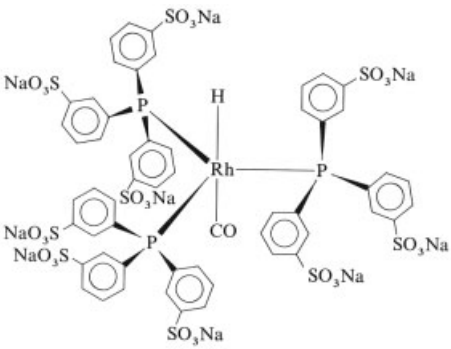

Figure 4. Structure of the water-soluble rhodium carbonyl hydride complex

3.1. Rhodium Processes

As shown in Table 2, rhodium-based processes can be classified into three types. The most important of these on an industrial scale uses the so-called phosphine-modified catalyst system. The unmodified rhodium carbonyl complex is used for the reaction of special olefins.

Low-Pressure Oxo (LPO) Process. Table 2, Type 1. In the mid-1970s Union Carbide and Celanese succeeded in using rhodium – triphenylphosphine catalysts for the hydroformylation of olefins on an industrial scale. Since then some other companies have developed modifications of this process. The most important of these, however, is the low-pressure oxo (LPO) process jointly developed by Union Carbide, Davy McKee, and Johnson Matthey. A detailed description of this specific low-pressure route follows with the conversion of propene to butyraldehyde as an example.

Figure 5 shows a schematic of hydroformylation with the LPO gas-recycle process [8], [36].

The reaction takes place in a stirred-tank reactor (a) made of stainless steel. The reactants and supplementary catalyst to make up for catalyst lost in production are fed in from below. Due to the sensitivity of the rhodium catalyst system toward catalyst poisons, the olefin and synthesis gas or hydrogen used as makeup gas must first be carefully purified (b). The reactor contains the catalyst dissolved in high-boiling reaction byproducts. In order to maintain catalyst activ-

Table 2. Comparison of the various oxo processes

Catalyst	$[RhH(CO)(PR_3)_3]$		$[RhH(CO)_4]$	$[CoH(CO)_3PR_3]$	$[CoH(CO)_4]$
	$R = C_6H_5$	$R = 3\text{-}C_6H_4SO_3Na$			
Process type no.	**1**	**2**	**3**	**4**	**5**
Hydroformylation conditions					
Pressure, bar	15 – 20	10 – 100	200 – 300	50 – 100	200 – 350
Temperature, °C	85 – 115	50 – 130	100 – 140	160 – 200	110 – 180
Results					
Selectivity for aldehydes	high	high	high	low	medium
n/i Ratio	92 : 8	95 : 5	50 : 50	88 : 12	80 : 20
Hydrogenation	low	low	low	high	medium

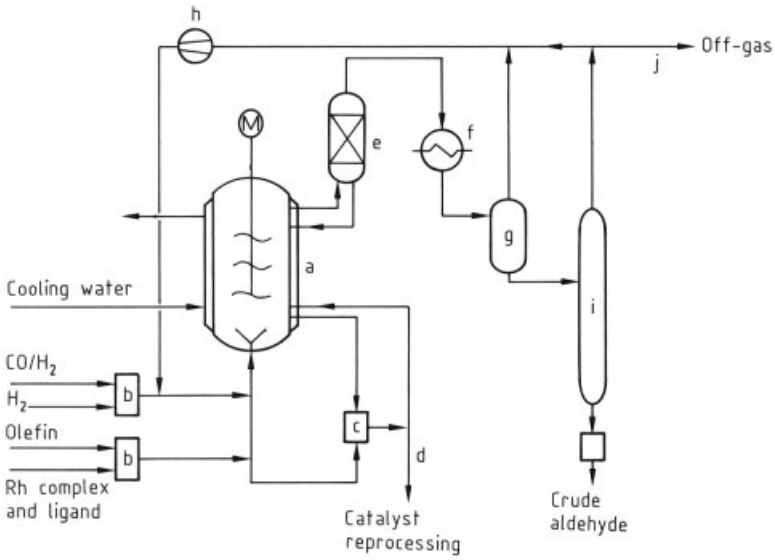

Figure 5. LPO process (schematic)
a) Stirred-tank reactor; b) Reactant purification system; c) Reprocessing; d) To catalyst reprocessing; e) Demister; f) Condenser; g) Separator; h) Compressor; i) Stripping column; j) Outlet

ity a portion of the solution must be continuously removed and reprocessed separately (c), and the noble metal and the phosphine returned to the process. From time to time, all of the catalyst must be removed via line (d) and reprocessed externally [12], [16]. The hydroformylation reaction takes place at < 20 bar and 85 – 115 °C. The reactor jacket is cooled to remove the heat of reaction. The reaction products and unreacted gaseous reactants (conversion of about 30 % per pass) are forced out of the reactor by the recycled gas and pass through the demister (e) and the condenser (f) into the separator (g).

Unreacted starting materials and part of the propane formed as byproduct are recycled to the reactor by means of the compressor (h). The level of propane in the circulating gas is adjusted by means of the outlet (j). The liquid reaction products are freed from residual olefin in a stripping column (i), and are worked up by multistage distillation. Residual olefin from the stripping column is recycled. To limit the buildup of inerts in the recycled gas stream and to reduce losses by venting, other treatment steps may be applied. These include extraction of propene with the aldehyde products, stripping the olefin from the aldehydes with synthesis gas and recycling both to the reactor [55], and washing the off-gas with stripped catalyst solution [56].

The LPO gas-recycle process has been partly replaced by a liquid-recycle variant, in which the catalyst solution and the aldehyde products

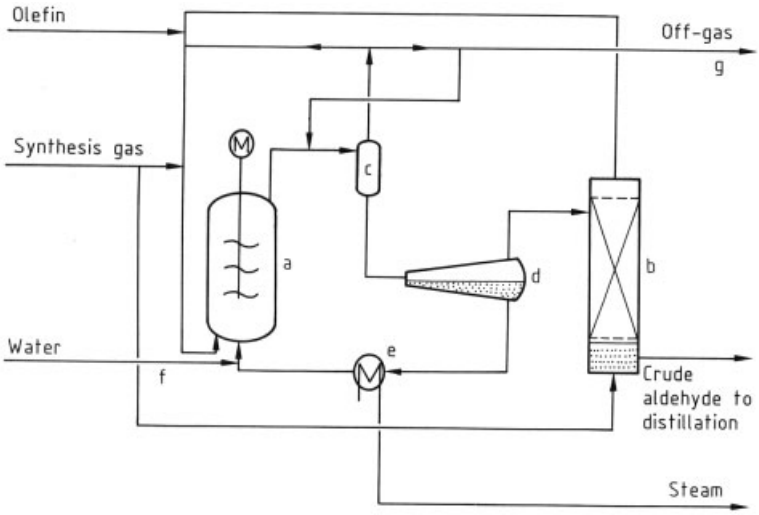

Figure 6. Schematic of the RCH – RP process
a) Stirred-tank reactor; b) Stripping column; c) Separator; d) Phase separator; e) Heat exchanger; f) Water inlet; g) Off-gas outlet

leave the reactor as a liquid. The catalyst solution is separated from the aldehydes in several distillation steps and recycled. Combinations of gas and liquid recycle have also been described and are claimed to give increased propene conversion [53], [54].

Ruhrchemie – Rhône-Poulenc (RCH – RP) Process. Another industrial low-pressure process for olefin hydroformylation (Table 2, Type 2) is based on a water-soluble rhodium catalyst (see Chap. 2) [37–43]. As with the LPO process, the RCH – RP process has found its greatest importance in the hydroformylation of propene.

The use of a water-soluble catalyst system is associated with substantial advantages for industrial practice, because the catalyst can be considered to be heterogeneous. Since the catalyst is insoluble in the organic phase formed, separation of the aqueous catalyst phase and the butanal is greatly simplified by phase separation, and losses of the noble metal in the crude aldehyde stream are negligible. High-boiling byproducts do not dissolve in the aqueous catalyst phase, dispensing with the need for continuous catalyst regeneration. In the LPO process, however, these byproducts are retained in the catalyst phase and lead to catalyst problems, unless the catalyst system is continuously regenerated.

The process is explained by means of the simplified flow sheet (Fig. 6). The reaction takes place in a stirred-tank reactor (a), which contains the catalyst solution; the reactants are introduced from below. Before entering the reactor, the synthesis gas is first passed through a stripping column (b) in countercurrent to the crude aldehyde stream in order to recover the unreacted propene. Furthermore, purification of the reactants can be avoided by this procedure [42]. The crude aldehyde product leaves the top of the reactor and passes through the trap (c) into the phase separator (d) where it is separated from the entrained catalyst solution. The catalyst solution is returned to the reactor via the heat exchanger (e).

Because of the higher temperature compared to the LPO process (see Table 2), the heat of reaction can be used for steam generation in the heat exchanger (e). The crude aldehyde from the phase separator (d) is distilled. Part of the water is retained in the crude aldehyde in homogeneous solution. This loss of water is compensated for via inlet (f). Part of the off-gas which escapes via the separator (c) can be recirculated; part must be drawn off to maintain a constant propane level in the gas. Depending on the purity of the olefin used, the off-gas (g) may contain considerable amounts of olefin. The RCH – RP process may then, for example, be combined

with a cobalt high-pressure process to convert the residual olefin [44].

Rhodium High-Pressure Process. If unmodified rhodium carbonyl hydride is used as a hydroformylation catalyst, the reaction product consists of roughly equal amounts of branched and straight-chain aldehydes (see Table 2, Type 3). For this reason this catalyst is only applicable if the *n/i* ratio is not important (i.e., both aldehydes are valuable products) or if the formation of a branched aldehyde is impossible (e.g., in the hydroformylation of ethylene to give propanal) [45]. Anhydrous propanal can be obtained by this process.

3.2. Cobalt Processes

Modified Catalysts (Shell Process). If the corresponding alcohol rather than the aldehyde is the desired product, the Shell process can be applied (Table 2, Type 4). A phosphine-modified cobalt catalyst with a strong hydrogenation activity is used and gives a favorable *n/i* ratio (88 : 12) in the initially formed aldehyde product mixture. The catalyst readily isomerizes feed olefins and can therefore be used to hydroformylate inner olefins. For a detailed description of this process see [2], [8].

Unmodified Catalysts. The first and, for a long time, the most important, industrial oxo process was based on an unmodified cobalt carbonyl hydride catalyst (Table 2, Type 5) which required high pressures and temperatures. However, since the successful introduction of rhodium catalysts its importance has declined but since stabilized at a lower level.

4. References

General References

1. B. Cornils in J. Falbe (ed.): *New Syntheses with Carbon-Monoxide,* Springer-Verlag, Berlin 1980.
2. *Ullmann,* 4th ed., **7,** 118.
3. *Kirk-Othmer,* 3rd ed., **16,** 637.
4. R. L. Pruett, *Adv. Organomet. Chem.* **17** (1979) 1.
5. *Houben-Weyl,* 4th ed., **E 3,** 180.
6. F. H. Jardine, *Polyhedron* **1** (1982) 569.
7. I. Wender, P. Pino (eds.): *Organic Syntheses via Metal Carbonyls,* **vol. 2,** John Wiley & Sons, New York 1977, p. 43.
8. *Winnacker-Küchler* 4th ed., **5,** 537.
9. C. D. Frohning, C. W. Kohlpaintner in B. Cornils, W. A. Herrmann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* **vol. 2,** VCH, Weinheim 1996, p. 64.

Specific References

10. Ruhrchemie AG, DE 849 548, 1938 (O. Roelen).
11. B. Cornils, R. Payer, K. C. Traenckner, *Hydrocarbon Process.* **54** (1975) 89.
12. P. E. Garrou, *Chem. Rev.* **85** (1985) no. 3, 171.
13. Ruhrchemie AG, DE 3 235 030 A 1, 1982 (R. Gärtner, B. Cornils, H. Springer, P. Lappe).
14. Ruhrchemie AG, EP 0 163 233 A 1, 1985 (H. Bahrmann, et al.).
15. Ruhrchemie AG, EP 0 175 919 A 1, 1985 (L. Bexten, B. Cornils, D. Kupies).
16. P. E. Garrou, R. A. Dubois, W. J. Chu, *CHEMTECH* 1985, no. 2, 123.
17. Ruhrchemie AG, EP 0 103 845 A 2, 1983 (R. Gärtner, B. Cornils, L. Bexten, D. Kupies).
18. A. G. Abatjoglou, E. Billig, D. R. Bryant, *Organometallics* **3** (1984) no. 6, 923.
19. UCC, EP 0 357 997, 1988 (D. J Miller, D. R. Bryant, E. Billig, L. B. Shaw).
20. UCC, EP 0 552 797, 1993 (J. E. Babin, D. R. Bryant, A. M. Harrison, D. J. Miller).
21. Hoechst, EP 544 091, 1992 (W. Konkol, H. Bahrmann, W. A. Herrmann, C. Kohlpaintner).
22. UOP, EP 504 814, 1992 (D. R. Bryant, J. E. Babin, J. C. Nicolson, D. J. Weintritt).
23. G. Gregorio et al., *Symp. Rhodium Homogeneous Catal. [Proc.]* 1978, 121.
24. G. Gregorio et al., *Chim. Ind. (Milan)* **62** (1980) 389.
25. D. Evans, J. A. Osborn, G. Wilkinson, *J. Chem. Soc. A* 1968, 3133.
26. D. S. Breslow, R. F. Heck, *Chem. Ind. (London)* 1960, 467.
27. R. F. Heck, D. S. Breslow, *J. Am. Chem. Soc.* **83** (1961) 4023.
28. D. Evans, G. Yagupsky, G. Wilkinson, *J. Chem. Soc. A* 1968, 2660.
29. R. L. Pruett, J. A. Smith, *J. Org. Chem.* **34** (1969) 327.
30. O. R. Hughes, J. D. Unruh, *J. Mol. Catal.* **12** (1981) 71.
31. Exxon Research and Engineering Co., DE 3 034 352, 1980, (A. A. Oswald, T. G. Jermansen, A. A. Westner, I-Der. Huang).

32. I. T. Horvath, R. V. Kastrup, A. A. Oswald, E. J. Mozeleski, *Catal. Lett.* **2** (1989) 85.
33. M. J. H. Russell, *Platinum Met. Rev.* **32** (1988) no. 4, 185.
34. P. Cavalieri d'Oro et al., *Chim. Ind. (Milan)* **62** (1980) 572.
35. John M. Brown, Alexander G. Kent, *J. Chem. Soc. Perkin Trans. 2* 1987, 1597.
36. F. Heinrich, M. Bernard, *27th DGMK-Haupttagung,* Aachen Oct. 6 – 8, 1982, Compendium 82/83, p. 189.
37. Rhône-Poulence Industries, DE 2 627 354, 1976 (E. Kuntz).
38. Rhône-Poulenc Chimie de Base, EP 0 104 967 B 1, 1982 (J. L. Sabot).
39. Rhône-Poulenc Recherches, EP 0 133 410 A 1, 1984 (J. Jenck, D. Morel).
40. Rhône-Poulenc Recherches, EP 0 158 572 A 1, 1985 (C. Barre, M. Desbois, J. Nouvel).
41. Rhône-Poulenc Industries, FR 8 005 488, 1980 (J. Jenck).
42. Ruhrchemie AG, EP 0 103 810, 1983 (B. Cornils et al.).
43. Ruhrchemie AG, EP 0 158 246 A2, 1985 (B. Cornils et al.).
44. Ruhrchemie AG, EP 0 111 257 B 1, 1983 (B. Cornils et al.).
45. *Ullmann,* 4th ed., **19,** 443.
46. B. Cornils, *Org. Process Res. Dev.* **2** (1998) no. 2, 121.
47. W. A. Herrmann, C. W. Kohlpaintner, *Angew, Chem. Int. Ed. Engl.* **32** (1993) 1524.
48. M. Beller, B. Cornils, C. D. Frohning, C. W. Kohlpaintner, *J. Mol. Cat. A* **104** (1995) 17 – 85.
49. F. Joó, A. Kathó, , *J. Mol. Cat. A* **116** (1997) 3 – 26.
50. Exxon Chemical Patents, EP 0 357 405, 1989 (N. A. De Munck).
51. Exxon Chemical Patents, EP 0 391 650, 1989 (D. M. Olijve, N. A. De Munck).
52. Exxon Chemical Patents, EP 0 372 925, 1988 (E. van Driessche et al.).
53. Union Carbide, EP 0 188 246, 1986 (D. L. Bunning, M. A. Blessing).
54. Davy Powergas, GB 1 387 657, 1973 (R. Fowler).
55. Union Carbide Chemicals and Plastics Company, EP 048 976, 1991 (K. D. Sorensen).
56. Union Carbide Chemicals and Plastics Company, EP 0 404 193, 1991 (D. L. Bunning).

Oxygen

MARK J. KIRSCHNER, The BOC Group, Inc., Murray Hill, New Jersey 07974, United States

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 4
4.1. Oxygen from Cryogenic Air Separation 4
4.2. Oxygen from Adsorptive Air Separation 5
4.3. Oxygen from Membrane Air Separation 6
5. Quality Specifications and Analysis . 7
6. Storage and Transportation 9
6.1. Storage 9
6.2. Material Selection for Oxygen Compatibility 9
6.3. Transportation 9
7. Uses 10
7.1. Combustion and Metallurgy 11
7.2. Production of Chemicals 13
7.3. Oxygen Bleaching of Wood Pulp . . . 15
7.4. Wastewater Treatment 15
7.5. Medicine 16
8. Economic Aspects 17
9. Toxicology and Occupational Health . 18
10. References 18

1. Introduction

Oxygen [*7782-44-7*] forms acids by combining with many different substances and accordingly derives its name from the Greek *oxys* "sharp acid", and -*genes* "born" meaning acid former. It is a nonmetallic element of group 16 of the periodic table. Under standard conditions it is a colorless, odorless, and tasteless diatomic gas, having the molecular formula O_2. Oxygen is essential to life, as we know it, and the most abundant element on the Earth's surface. Oxygen accounts for 23.2 wt % of the atmosphere, 85.8 wt % of seawater, and 46.6 wt % of the Earth's crust.

The discovery of oxygen is credited to the independent work of the Swedish chemist SCHEELE and the English chemist PRIESTLEY. SCHEELE (1772) obtained oxygen by heating potassium nitrate and mercury(II) oxide to decomposition and collecting the evolved gas. PRIESTLEY (1774) also thermally decomposed mercury(II) oxide to obtain oxygen and further demonstrated that at a lower temperature, mercury and oxygen could be quantitatively recombined in the reverse reaction to reproduce the starting oxide. Neither researcher, however, recognized oxygen for what it was. The French chemist LAVOISIER (1777), who first understood the gas to be elemental, gave the new substance its present name and, in opposition to the prevailing phlogiston theory, explained combustion as the union of oxygen with a burning fuel.

Oxygen is nonflammable but readily supports combustion. All materials that are flammable in air burn much more vigorously in oxygen, some with an explosive intensity.

Free oxygen constitutes about one-fifth of the atmosphere (20.99 vol %) and is the result of photosynthesis. In green plants, incident solar energy is captured and used to convert water, carbon dioxide, and minerals into oxygen, which is released into the atmosphere, and energy-rich organic compounds (carbohydrates, amino acids, proteins, lipids, and pigments) required by the plants for growth and reproduction. Isotope studies with ^{18}O have confirmed that the source of the released oxygen is water. Oxygen combined as carbon dioxide is retained in the organic products.

2. Physical Properties

Oxygen has three known allotropic forms: diatomic (O_2) ordinarily found in air; triatomic (O_3), known as ozone; and a rare, unstable tetratomic form (O_4) that readily dissociates into two diatomic oxygen molecules. In the diatomic form, two unpaired electrons lie in antibonding orbitals. The paramagnetic behavior of oxygen confirms the presence of these electrons.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_329

The seven isotopes of oxygen are listed in Table 1. Some physical properties are as follows [1]:

A_r	15.9994
M_r	31.9988
Specific volume (21.1 °C, 101 kPa)	0.7541 m^3/kg
bp (101 kPa)	−182.96 °C
mp (101 kPa)	−218.78 °C
Triple point	
Temperature	−218.79 °C
Pressure	0.1480 kPa
Density, gas (21.1 °C, 101 kPa)	1.326 kg/m^3
Relative density, gas (21.1 °C, 101 kPa, air=1)	1.105
Density, liquid (− 182.96 °C)	1141 kg/m^3
Density, gas (− 182.96 °C)	4.483 kg/m^3
Critical temperature	−118.57 °C
Critical pressure	5043 kPa
Critical density	436.1 kg/m^3
Latent heat of vaporization (− 182.96 °C)	213 kJ/kg
Latent heat of fusion (− 218.78 °C)	13.86 kJ/kg
Specific heat, gas (21.1 °C, 101 kPa)	
C_p	0.9191 kJ kg^{-1} K^{-1}
C_v	0.6578 kJ kg^{-1} K^{-1}
Viscosity, gas (25 °C, 101 kPa)	20.8×10^{-3} Pa · s
Viscosity, liquid (− 173.3 °C)	156×10^{-3} Pa · s
Thermal conductivity, gas (25 °C, 101 kPa)	26.6×10^{-3} W m^{-1} K^{-1}
Thermal conductivity, liquid (− 173.3 °C)	192.9×10^{-3} W m^{-1} K^{-1}
Surface tension (− 173.3 °C)	13.2×10^{-3} N/m

Table 1. Oxygen isotopes

Mass number	Natural abundance, %	Half-life, s
Stable isotopes		
16	99.759	
17	0.037	
18	0.204	
Unstable isotopes		
14		71
15		124
19		29
20		14

Many applications of oxygen (e.g., wastewater treatment, homogeneous and heterogeneous liquid phase reactions where oxygen must flow through the liquid to the reaction site) are dependent on its behavior according to Henry's law, which states that the concentration of a gas in a liquid is directly proportional to the partial pressure of the gas in equilibrium with the liquid. At 0 °C, the absorption coefficient (Henry's law constant) for oxygen in water is 70.5 mg L^{-1} atm^{-1}. Since the concentration of oxygen in air is 21 % (p = 0.21 atm), the equilibrium saturation concentration of oxygen in water at 0 °C is the product of these, or 14.8 mg/L. As the proportion of oxygen in the gas increases from 21 to 100 %, the equilibrium saturation concentration increases approximately fivefold, raising its value from 14.8 to ca. 70.5 mg/L.

Additionally, the saturation concentration is proportional to absolute pressure. Doubling the absolute pressure of a system will double the saturation concentration of oxygen in that system.

3. Chemical Properties

Oxygen, atomic number 8, has the electron configuration $1s^2 2s^2 2p^4$. Thus six electrons occupy the outermost shell. With the exception of fluorine, oxygen has the highest electronegativity and electron affinity of any element; the Pauling electronegativity is 3.44.

The oxygen molecule is unusual in that it is paramagnetic in the ground state. This can be explained by the molecular orbital theory, whereby the two least strongly bound electrons occupy degenerate orbitals of π symmetry and have parallel spins. Because there are four more electrons in bonding MOs than in antibonding MOs, diatomic oxygen contains a double bond [2].

The molecular species O_2 is not especially reactive at ambient temperature and pressure. The atomic species O is far more reactive, with its energy of formation ($O_2 \rightarrow 2\,O$) being quite large (490.7 kJ/mol). In nonaqueous reactions, the reduction of O_2 is a single-electron process (Eq. 1), which yields the superoxide anion, but in aqueous systems, it is a two-electron process (Eq. 2), which results in hydroperoxide and hydroxide anions:

$$O_2 + e^- \rightleftharpoons O_2^- \quad (1)$$

$$O_2 + H_2O + 2\,e^- \rightleftharpoons HO_2^- + OH^- \quad (2)$$

Oxygen forms oxides, peroxides, and superoxides, according to the periodic position of the corresponding reactant. Alkali and alkaline-earth metals, which have relatively small ionization energies and large oxidation potentials,

form ionic oxides (M_xO_y). The most reactive of these metals give peroxides (e.g., Na_2O_2) and superoxides (e.g., KO_2).

Less active metals, typified by those of groups 13 – 16, are far less ionic, and their oxides are often variable in their stoichiometry because of the possibility of more than one valence state. If more than one metal oxide species is possible for a given metal, the oxide of the lower oxidation state is more ionic, more soluble in water, and a stronger base.

The metalloids boron, silicon, phosphorus, and germanium form polymeric oxide networks (e.g., SiO_2) and acidic oxides (B_2O_3). The bonds formed between oxygen and metalloids have a largely covalent character.

Because atoms of nonmetals have very little tendency to give up electrons to the chemical bond, the resultant bonds between oxygen and nonmetals are covalent. Nitrogen (group 15), sulfur and selenium (group 16), and the halogens (group 17) form oxides that are gaseous at ambient temperature and pressure (e.g., SO_2) and react with water to form acids.

Noncatalytic oxidation of organic materials is part of a general reaction class involving the incorporation of elements such as nitrogen, phosphorus, oxygen, sulfur, fluorine, chlorine and bromine into the reacting molecule [3]. With oxygen, oxidation typically proceeds via breaking of bonds in the reactant molecule along with the incorporation of oxygen. Carried to the extreme, the result is called combustion.

Table 2. Bond dissociation energies

Bond type	Dissociation energy, kJ/mol
O=O	498
H−H	435
CH_3 −H	427
C_2H_5 −H	406
CH_3 −CH_3	351
H_2CCHCH_2 −H	322

Table 2 [4] compares the bond dissociation energy of oxygen (498 kJ/mol) to that of other chemical bonds that are important in these reactions. Because the oxygen – oxygen double-bond dissociation energy is significantly higher than the dissociation energy of most other bonds in organic reactants, oxidation of organic molecules is commonly initiated with the breaking of a lower energy bond in the reactant molecule [5]. For example, in ethane, the first bond to be broken is expected to be the carbon – carbon σ-bond, which has a dissociation energy of 351 kJ/mol:

$$H_3C{-}CH_3 \longrightarrow H_3C^{\bullet} + {}^{\bullet}CH_3 \tag{3}$$

A methyl radical then attacks an oxygen molecule at one of the bonding electron pairs. As the methyl – oxygen bond is established, one electron from the oxygen double bond must be promoted to a π antibonding level, whereby the oxygen –oxygen bond is significantly weakened:

$$H_3C^{\bullet} + O_2 \longrightarrow H_3COO^{\bullet} \tag{4}$$

At higher temperature the peroxy radical undergoes bond rupture, but at lower temperature it may be sufficiently stable to undergo a second attack such as

$$H_3COO^{\bullet} + H_3CCH_3 \longrightarrow H_3COOH + H_2\dot{C}CH_3 \tag{5}$$

Consequently, the oxygen–oxygen linkage is further weakened by the addition of an electrophilic group, and it eventually ruptures:

$$H_3COOH \longrightarrow H_3CO^{\bullet} + {}^{\bullet}OH \tag{6}$$

The result is a degenerate branched chain reaction, which can proceed at an increasing rate, possibly until it becomes explosive. At this stage, more extensive oxygen additions can occur by reactions such as

$$RO^{\bullet} + O_2 \longrightarrow \underset{\underset{O}{\|}}{R} - O - O^{\bullet} \tag{7}$$

A second class of reaction is called partial oxidation (see also→Oxidation). These reactions are catalyzed by transition-metal oxides and are some of the most widely applied processes for converting hydrocarbons into valuable chemical intermediates, such as those incorporating aldehyde, carboxylic acid, epoxide, and nitrile. Similarly, partial oxidation reactions catalyzed by metal-containing enzymes (e.g., cytochromes) are involved in the sequences of reactions responsible for metabolism in living organisms.

Industrial catalysts are usually complex, solid transition-metal oxides, whose surfaces donate oxygen to adsorbed hydrocarbon reactants. Catalyst selectivity, measured by the rate of formation of desired products relative to carbon

monoxide, carbon dioxide, water, and other undesired coproducts, is often sensitive to reaction conditions, especially temperature. Reactor design for partial oxidation is difficult because the reactions are strongly exothermic and reactor temperature profiles are difficult to predict. Some industrially significant partial oxidations with oxygen are described in Section 7.2.

In this class of oxidation reactions, a different mechanism prevails in the breaking of the oxygen – oxygen bond. The transition-metal catalyst transfers electrons via intermediates such as free radicals, O_2^-, H_2O_2, and transition-metal ions [6]. Such reactions are nonbranched chain reactions; occasionally though, the free radicals may start thermal (noncatalytic) chain reactions, including branched chain reactions. It is sometimes advantageous to use two different metal ions, one to attack the organic molecule and the other to interact with oxygen and facilitate electron transfer.

4. Production

4.1. Oxygen from Cryogenic Air Separation

In the late 1800s, various attempts were made to separate oxygen from air on a commercial scale. These all relied on chemical processes employing thermal cycling [7]. The most successful processes involved heating barium oxide in air at 590 °C under pressure, so that it accumulated oxygen to form barium peroxide. The peroxide, in turn, was heated to ca. 870 °C, at which point it released half of its oxygen, which was collected at ca. 95 % purity. This was subsequently improved by maintaining the barium oxide at 600 °C, while oxygen was absorbed from air supplied at 170 kPa and subsequently desorbed under vacuum.

At this time, successful experiments in air liquefaction were being conducted, which eventually led to the separation of oxygen via distillation. Because the percentage of oxygen in the gas phase that is in equilibrium with liquid air (containing 21 % oxygen) is only 6 %, separation of oyygen by distillation is convenient. In 1901, Carl von Linde (Germany) demonstrated the first practical method for controlling the distillation of liquid air: rectification. Early columns were inefficient, but by 1910, Linde had introduced a double column, one above the other, with a condenser – evaporator in between. This remains the concept employed for producing oxygen by cryogenic air separation.

Over the years, oxygen production plants have continued evolving into larger, more efficient process systems. To date, the largest modular plant is capable of producing 2250 t/d [8].

Cryogenic air separation involves three steps:

1) Purification of the incoming air to remove particles, carbon dioxide, and water
2) Refrigeration and economization of refrigeration values contained in the product and waste streams
3) Separation by distillation

Figure 1 shows a schematic of a process embodying these steps to produce oxygen. Filtered air compressed to 600 – 700 kPa is cooled and scrubbed in a direct-contact water wash tower (a) and then directed into a plate-fin reversing heat exchanger (b), where it is cooled further in counterflow to product oxygen and waste nitrogen. Carbon dioxide and water are removed from the air stream by condensation and solidification in this heat exchanger. Every few minutes, the paths of the incoming air and the waste nitrogen streams are reversed (hence the name reversing heat exchanger) so that deposited impurities can be purged from the exchanger. In this way, the process is able to operate continuously on a cyclic basis. State-of-the-art plants are, however, being constructed with molecular sieve adsorption units, in place of the reversing heat exchangers. After carbon dioxide and water have been removed by the sieve, the incoming air is then cooled by cold waste gas in counterflow heat exchangers.

Refrigeration for this process is achieved primarily by the Claude principle of conversion of energy into work by the expansion of a compressed gas in a turbine. Some of the cooled air is returned through the cold end of the heat exchanger before expansion to ca. 120 kPa in a turbine (c), and then introduced at an appropriate point in the upper, low-pressure column of the double column rectifier (d). The remainder of the air passes into the lower, high-pressure column

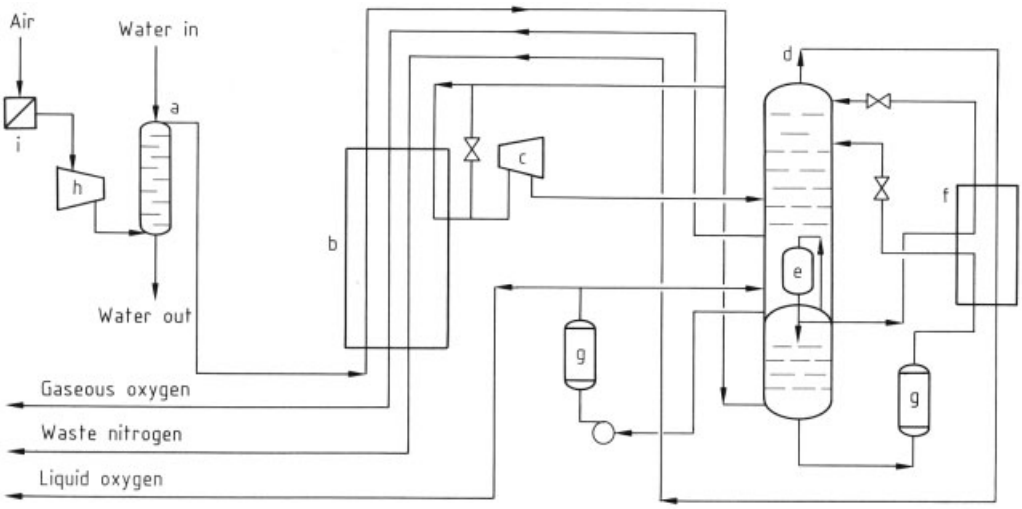

Figure 1. Cryogenic oxygen plant (schematic)
a) Water wash cooler; b) Reversing heat exchanger; c) Expansion turbine; d) Double column rectifier; e) Condenser; f) Subcooler; g) Adsorber; h) Compressor; i) Filter

at 500 – 600 kPa where it is separated into nitrogen gas at the top and oxygen-enriched liquid (38 % O_2) at the bottom.

Gaseous nitrogen is condensed by liquid oxygen (which is vaporized) in the condenser – reboiler (e). Part of this liquid nitrogen is returned as reflux to the lower column, and part is expanded into the top of the upper column as reflux liquid. The oxygen-enriched liquid (rich liquid) from the sump of the lower column is expanded as feed into an appropriate point in the upper column. The reflux liquids used are subcooled by returning nitrogen in the heat exchanger (f), to reduce the amount of liquid vaporized on expansion.

In the upper column, the feed is separated into pure oxygen (99.6 %) at the bottom and waste nitrogen with 1 – 2 % oxygen at the top. The adsorber (g), packed with silica gel, removes trace hydrocarbon matter from liquid oxygen in the condenser – reboiler and from the rich liquid, ensuring that unsafe hydrocarbon concentrations do not accumulate in the process.

4.2. Oxygen from Adsorptive Air Separation

Various forms of molecular sieves exhibit a greater capacity for nitrogen than for oxygen, in a ratio of ca. 2 : 1. When used in an isothermal fractionator (pressure-swing adsorption) with compressed air feed, the nitrogen is preferentially adsorbed, leaving a product gas rich in oxygen and argon. Because the actual nitrogen capacity is relatively small, the phenomenon is basically a transient one, with nitrogen breaking through after one to two volume changes with feed.

The principle of oxygen enrichment from air by pressure-swing adsorption (PSA) is relatively simple. Compressed air enters the adsorption bed at high pressure. Typical adsorbents are usually based on zeolites. Because zeolite has a higher selectivity for nitrogen, and oxygen is at a lower partial pressure, nitrogen is adsorbed to an appreciably greater extent and oxygen-enriched air is obtained at the outlet. The bed is then blown down to a lower pressure to desorb nitrogen. A purge step may sometimes be used to sweep nitrogen out of the bed before a new cycle begins [9].

Oxygen PSA had its beginnings in the late 1960s as an extention of hydrogen production by synthetic zeolites. The driving force for this development was linear capacity turndown and low operating costs for wastewater treatment. The high power consumption and relatively high capital cost, compared to cryogenic systems, were characteristic of a design that was ad-

vantageous to the federal government-funded wastewater treatment market. The government supported capital but not operating expenses for these wastewater treatment facilities. These same factors, however, initially limited success in the traditional industrial sector, but this has now changed. Advances in adsorptive materials and process design in recent years have improved the process economics of oxygen PSA.

A simple PSA plant, as illustrated in Figure 2, consists of two vessels containing beds of zeolite, a feed of clean compressed air, waste gas discharge, and a control system that switches valves according to the adsorption – desorption sequence of operation. Bed 1 is shown in the adsorption mode: its feed air and product valves are open, and the waste gas valve is closed. Nitrogen in the feed air is adsorbed onto the zeolite, permitting the product gas, which is enriched in oxygen, to pass through to whatever process is being served. Bed 2 is in the regeneration or desorption mode: its feed air and product valves are closed, and the waste gas valve is open, thereby allowing the vessel to depressurize and venting desorbed nitrogen to the atmosphere. After some time, normally several minutes, valve positions are switched. Bed 2 goes into the adsorption mode while bed 1 is regenerated. Thus, by periodic bed switching, a continuous supply of oxygen is generated.

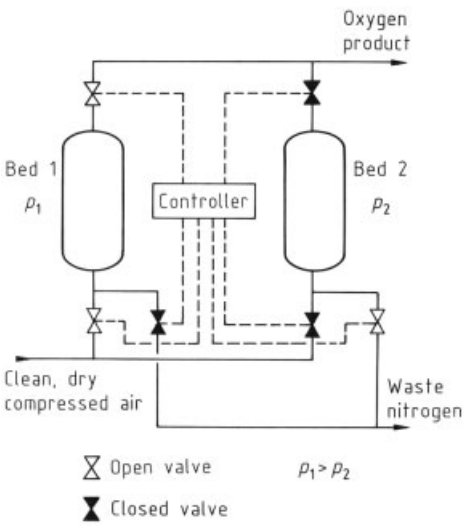

Figure 2. Oxygen pressure swing adsorption (PSA) system

Sometimes more than two vessels are used, with a more complicated valve switching sequence. In this way, the batch PSA process can better approximate a continuous process. Another product smoothing technique is to insert a surge vessel in the product gas line, downstream of the adsorption beds, to dampen pressure pulses caused by bed switching.

Other variations have been engineered, including vacuum-swing adsorption and a combined vacuum – pressure-swing adsorption process.

By careful selection of the molecular sieve and creative process engineering, oxygen concentrations up to 93.5 % can be achieved. Because most oxygen applications do not require the highest purity available, lower purities are generally provided, with an attendant decrease in specific capital and power charges.

Air enrichment makes a product of ca. 75 % oxygen, which is subsequently diluted with air to the required enrichment level; this procedure is more economical than direct production of a product containing a lower level of oxygen.

4.3. Oxygen from Membrane Air Separation

Membrane separation of oxygen is accomplished by a thin polymeric film that allows selective transport of oxygen through it as a result of a driving force, such as differential partial pressure. Permeation of oxygen through a membrane can be expressed in terms of the permeability coefficient of the membrane, P defined as

$$P = \frac{Jl}{\Delta\phi}$$

where J is the flux of oxygen, l the membrane thickness, and $\Delta\phi$ the driving force between the upstream and downstream gas phases.

Oxygen dissolves into the membrane at its upstream surface and diffuses across the membrane to the downstream surface, where it volatilizes into the adjacent gas phase. Other gases in air also dissolve into the membrane and diffuse along with oxygen; the most important of these is nitrogen. Thus, selectivity of a membrane material for oxygen becomes an important design parameter. The preferred membrane material must have both a high permeability

and a high selectivity for oxygen. Unfortunately, achieving a high permeability rate concurrent with a high selectivity for oxygen – nitrogen separation in a single material has been elusive. Polymers derived from silicone rubber offer the best performance (Table 3), but even these are not competitive with alternative oxygen supply means, except for the smallest amounts, where high purity is not required. Small membrane units, used to supply a single patient's medical oxygen needs, meet these criteria and have been commercially successful. These systems deliver 30 – 40 % oxygen at flow rates of 1 – 10 L/min.

5. Quality Specifications and Analysis

Table 4 lists quality specifications for commercially available grades of oxygen [11]. Oxygen of U.S.P. grade must be manufactured, filled, and handled in accordance with regulations directed by the Food and Drug Administration in the United States [12], [13]. A manufacturer or filler of U.S.P. oxygen must be registered with the Food and Drug Administration as a drug manufacturer.

The quantitative determination of oxygen is important for monitoring many processes such as respiration and combustion. Physical methods rely on oxygen's paramagnetism or solid-state diffusion. Chemical methods include potentiometry and catalytic combustion.

Oxygen is one of three gases that exhibit paramagnetism, but it is several times more paramagnetic than either nitric oxide or nitrogen dioxide. This property is exploited in gaseous oxygen analyzers based on magnetic susceptibility.

One type of paramagnetic oxygen analyzer incorporates a small glass dumbbell suspended on a taut quartz fiber in the magnetic field. When a sample containing oxygen is drawn into the test chamber surrounding the dumbbell, the magnetic force is altered, causing the dumbbell to rotate away from the region of maximum magnetic flux density. The degree of rotation is proportional to the partial pressure of oxygen in the sample. A mirror attached to the dumbbell reflects an incident beam of light onto a translucent scale calibrated for oxygen concentration. With a full-scale span of 5 % oxygen, an accuracy of 0.05 % can be achieved. Response time is about 10 s, and either static or continuous samples may be measured [14].

Another paramagnetic technique employs a variation of hot-wire anemometry. The gas sample is directed across a test cell containing an electrically heated hot wire; a strong magnetic flux from a permanent magnet is applied normal to the length of the wire. Oxygen from the sample is pulled into the region around the hot wire by the magnetic flux and is heated. The warmed oxygen loses its magnetic susceptibility in inverse proportion to the square of its absolute temperature. In this way, heated, relatively nonparamagnetic oxygen is continually displaced by cooler, more paramagnetic oxygen being attracted by the magnetic flux. A flow of gas proportional to the amount of oxygen in the sample is established across the hot wire, which is consequently cooled, decreasing its resistance. The resistance of the wire in the test cell is compared with the resistance of a similar wire in a reference cell that also contains a sample of the gas being analyzed, but does not have a magnetic flux around the wire. Thus all variables except cooling due to the "magnetic wind" are canceled. The overall accuracy of this technique is claimed to be ± 0.25 % oxygen up to 20 % and ± 2.5 % up to 100 % [15].

The Hersch galvanic cell provides an electrolytic method for the determination of low concentrations of oxygen in gas streams. The gas stream passes through an electrolytic cell, which consists of a silver cathode and an anode of active lead or cadmium. The electrodes are separated by a porous tube saturated with an electrolyte of potassium hydroxide. Oxygen in the sample is reduced to hydroxyl ions at the silver cathode. The metallic lead or cadmium in turn is oxidized to lead(II) ions or cadmium hydroxide. The magnitude of the cell current is a measure of the oxygen in the sample. Sensitivities less than 1 ppm are attainable by this method.

Oxygen in flue gases may be monitored by probe-type analyzers that are inserted directly into high-temperature exhaust streams. These probes operate on oxygen diffusion, controlled by the defect chemistry of zirconium oxide (ZrO_2), doped with small amounts of calcium oxide and ytterbium(III) oxide. The ceramic contains cation vacancies and anion holes whose populations increase with increasing temperature. At ca. 1500 °C, oxygen anions tend to mi-

Table 3. Permeability and selectivity of polymer membranes [10]

Polymer	Permeability $P \times 10^8$, cm^3 (STP) cm/(cm^2 s Pa)		Selectivity, P_{O_2}/P_{N_2}
	Oxygen	Nitrogen	
Silicone rubber	37.5	9.75	3.85
Polyisoprene (natural rubber)	1.73	0.63	2.75
Ethyl cellulose	0.203	0.063	3.22
Polyethylene (ϱ=0.92)	0.0443	0.015	2.95
Polychloroprene	0.03	0.009	3.33
Butyl rubber	0.0098	0.0024	4.06
Cellulose acetate	0.006	0.0021	2.86

Table 4. Commercial grades of oxygen[a]

Limiting characteristics	Maxima for type I (gaseous)							Maxima for type II (liquid)				
	A	B	C	G	D	E	F	A	B	G	C	D
Oxygen	99.0	99.5	99.5	99.5	99.5	99.6	99.995	99.0	99.5	99.5	99.5	99.5
Nitrogen				100						100		
Odor	none				none			none				none
Water, ppm (v/v)			50	2	6.6	8	1.0		6.6	2	26.3	6.6
Dew point, °C			− 48.1	− 71.7	− 63.3	− 62.2	− 76.1		− 63.3	− 71.7	− 82.3	−63.3
Water (condensed)		5 ml per container										
Total hydrocarbon content (as methane)				25		50	1.0			25	67.7	
Methane					50							25
Ethane and other hydrocarbons (as ethane)					6							3
Ethylene					0.4							0.2
Acetylene					0.1		0.05				0.5	0.05
Carbon dioxide	300[b]				10		1.0	300[b]				5
Carbon monoxide	10[b]						1.0	10[b]				
Carbon dioxide and carbon monoxide				5						5		
Nitrous oxide				2	4		0.1			2		
Halogenated refrigerants					2							1
Solvents					0.2							0.1
Other components					0.2							0.1
U.S.P.	yes							yes				
Permanent particles											1 mg/L, 1 mm	

[a] Units are in ppm, mol/mol, unless otherwise stated. [b] Test is not required when oxygen is produced by air liquefaction.

grate into the holes, causing the vacancies to be displaced. When a differential partial pressure of oxygen across the ceramic exists, an electric potential is established by the redox reactions (Eqs. 8 and 9):

Cathode reaction $O_2 + 4e^- \longrightarrow 2O^{2-}$ (8)

Anode reaction $2O^{2-} \longrightarrow O_2 + 4e^-$ (9)

A voltmeter in series with platinum electrodes on opposite sides of the ceramic measures the potential difference which, if the oxygen partial pressure is controlled on the reference side, is proportional to the difference between the oxygen concentrations [16].

Dissolved oxygen has been measured by the Winkler titration method for many years. This is based on the quantitative oxidation of alkaline manganese(II) hydroxide by oxygen. After acidification in the presence of excess iodide, an amount of iodine is produced, equivalent to the amount of originally dissolved oxygen. Titration

with sodium thiosulfate determines the amount of iodine in the sample.

Real-time dissolved oxygen measurements can be made with oxygen-specific electrodes, utilizing either potentiometric or galvanic analysis. The electrodes are protected by an oxygen-permeable polymeric membrane, which provides a diffusion barrier against other solution impurities [17].

6. Storage and Transportation

6.1. Storage

Compressed oxygen is frequently stored in high-pressure cylinders or tube trailers, depending on the frequency and rate of consumption. For larger needs, liquid oxygen is more economically stored at the user's site and vaporized as required. This is replenished by cryogenic tank trailers.

At consumer sites, liquid-oxygen stations must comply with NFP 50, Standard for Bulk Oxygen Systems at Consumer Sites [18]. Although oxygen is nonflammable, it vigorously accelerates combustion. The distances between an oxygen station and various exposures, illustrated in Figure 3, have been established to minimize the possibility of liquid oxygen, or an oxygen-rich air mixture, interacting with persons or flammable substances.

6.2. Material Selection for Oxygen Compatibility

At liquid-oxygen temperature, ordinary carbon steels and many alloy steels lose their ductility and are therefore considered unsuitable for liquid-oxygen service. Satisfactory metals and alloys include austenitic stainless steels (i.e., 304 and 316), nickel – chromium alloys, nickel, Monel 400, copper, brasses, bronzes, and aluminum alloys.

Many materials have been evaluated for oxygen compatibility with respect to their ease of ignition and their combustion intensity with oxygen [19]. Careful selection of nonmetallic components for oxygen service is especially important, because, if they are ignited, the metallic portion of the system may also be ignited.

Oxygen equipment and systems must be cleaned adequately to remove harmful contamination prior to the introduction of oxygen. Harmful contamination includes both organic and inorganic materials such as oil, grease, fibers, sawdust, solvents, welding slag, rust, and dirt. Organic materials are combustible, and inorganic particles impacting at high speed, particularly at elbows in the distribution piping, can be an ignition source. The cleaning of oxygen systems is described in [20].

6.3. Transportation

Under controlling regulations, oxygen may be shipped as either a compressed gas (13 790 kPa or above) or a cryogenic liquid (below 1379 kPa and − 146.5 °C) as follows [21]:

By Rail: In cylinders and tube tank cars, for both gaseous and liquid oxygen. Special insulated cylinders and tank cars for liquid oxygen are not subject to U.S. Department of Transportation (DOT) regulations below 172 kPa, except in Canada where oxygen is regulated regardless of pressure.

By Road: Compressed in cylinders on trucks and in tube trailers. Liquid oxygen in insulated cylinders and cargo tanks.

By Water: Compressed in cylinders on cargo and passenger vessels and on ferry and railroad car ferry vessels, both passenger and cargo. In authorized tank cars on cargo vessels only. Liquefied oxygen in insulated containers on cargo and passenger vessels and on ferry and railroad car ferry vessels, both passenger and cargo.

By Air: A maximum net weight of compressed oxygen per cylinder of 68 kg aboard passenger aircraft and 136 kg aboard cargo aircraft. Liquid oxygen is not accepted for air shipment under present regulations.

The shipment of compressed or liquid oxygen is under the jurisdiction of DOT and must be in containers that meet regulatory specifications [22]. Cylinders meeting DOT specifications 3A, 3AA, or 3AL are the types usually employed to ship gaseous oxygen, but oxygen is authorized for shipment in any cylinders designated for nonliquefied compressed gases. These include cylinders that comply with DOT specifications 3A, 3AA, 3AX, 3AAX, 3AL, 3B, 3E,

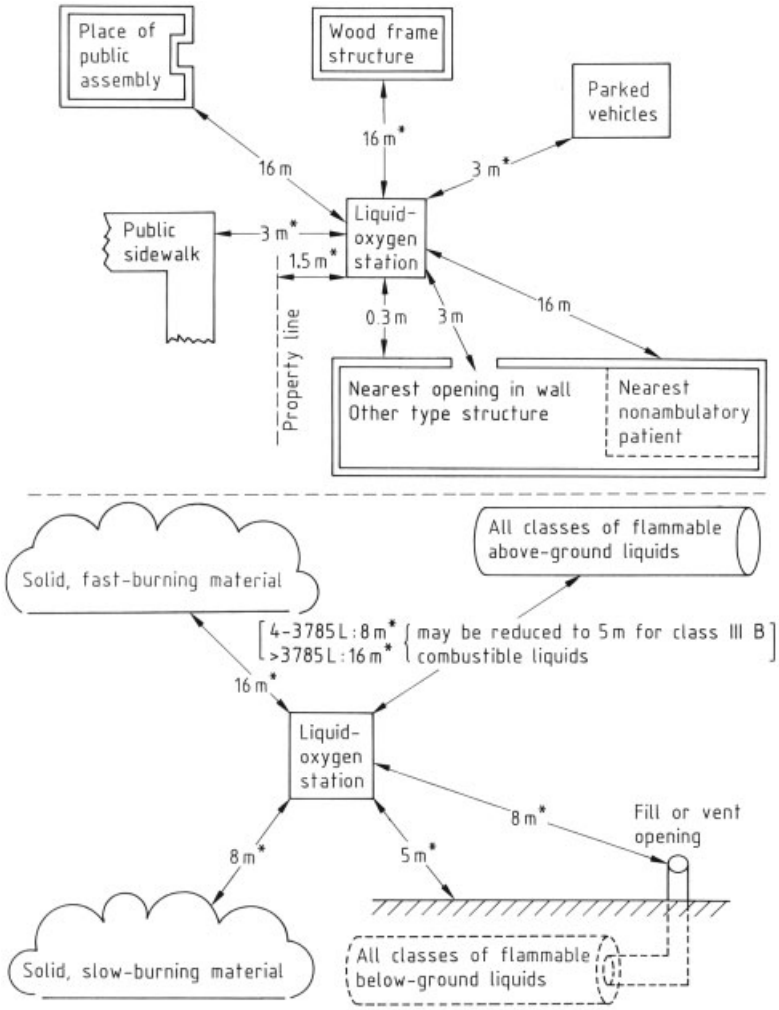

Figure 3. Minimum distances between liquid-oxygen storage and exposures
Note: These distances do not apply where protective structures, having a minimum fire rating of 2 h, interrupt the line of sight between uninsulated portions of the bulk oxygen storage installation and the exposure. The protective structures protect uninsulated oxygen storage containers or supports, control equipment, and system piping (or parts thereof) from external fire exposure. Liquid-oxygen storage containers are insulated; such containers may provide line-of-sight protection for uninsulated system components [18].

3T, 4B, 4BA, 4BW, and 39. Some new, higher strength steel cylinders and fiber-wrapped aluminum cylinders are authorized under DOT exemptions [23].

Liquid oxygen is authorized for shipment in cylinders that meet DOT specification 4L. However, when it is packaged in dewars – small vacuum-isolated containers holding from 4 to 90 L at atmospheric pressure – DOT cylinder regulations do not apply.

Most of these cylinders must be requalified for use by hydrostatic testing every 5 or 10 years. Specifications for filling limits [22], pressure relief devices [24], and valve outlet connections [25] are also regulated.

7. Uses

In many applications, the choice is to operate with ambient air as the source of oxygen or to ob-

tain the oxygen in concentrated form, otherwise known as commercial oxygen. The choice is always dictated by the relative process economics, and experience shows that most of the oxygen consumed in industrial and biological processes is introduced in the form of air. Sometimes oxygen is used to assist an air-based process to increase production without significant capital expenditures, or to overcome a mass-transfer limitation when dissolved oxygen is required. At other times, pure oxygen-based systems are employed that fully exploit the chemical and physical properties of oxygen by eliminating nitrogen, the attendant diluent that is always present in air-based systems.

The first significant industrial use of oxygen began in the early 1900s, when gaseous oxygen became available relatively cheaply and in quantity from air liquefaction. The oxyacetylene flame was improved, modified, and developed into a cutting tool and welding torch. During this time, hospitals began using oxygen to treat patients suffering from respiratory ailments.

7.1. Combustion and Metallurgy

The source of oxygen for the combustion of carbonaceous fuel is usually air.

The kinetic theory of gases [26] predicts that the number of binary collisions in a unit volume per unit time is proportional to the square of the pressure of the gas. Thus, increasing the partial pressure of oxygen in a combustion process exponentially increases the number of oxygen – fuel interactions per unit volume per unit time. The result of increasing the oxygen concentration in an air – fuel flame is a higher rate of flame propagation, or flame velocity, due to more frequent molecular collisions, and a higher flame temperature due to the higher rate of heat release within a given combustion volume. Figures 4 and 5 demonstrate the effect of oxygen enrichment on the velocity and temperature of a methane flame.

A radiative flame transfers heat more readily at a higher temperature, and a higher velocity flame allows more heat to be released from a given fuel within a limited combustion space. Both of these attributes are exploited in various industrial operations that employ oxygen enrichment to increase production capacity.

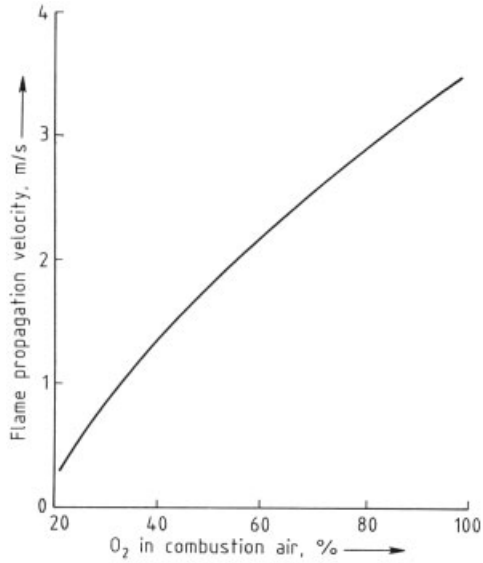

Figure 4. Effect of oxygen enrichment on flame velocity of methane

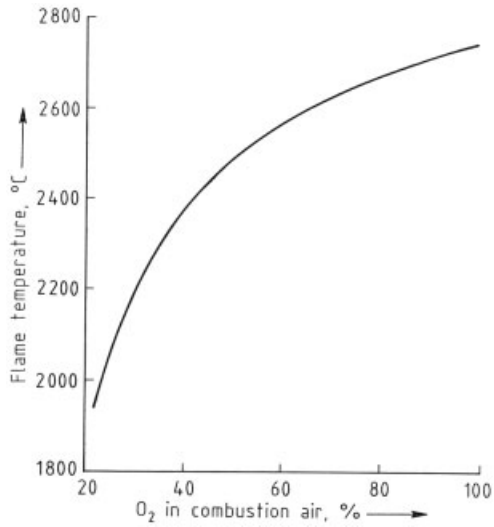

Figure 5. Effect of oxygen enrichment on flame temperature of methane

Oxygen enrichment also displaces nitrogen from the system. The presence of nitrogen in combustion has a number of effects, one of which is to chill the flame. Nitrogen is heated by the flame, thereby decreasing the overall thermal efficiency of the process.

Because the release of 1 GJ requires the addition of ca. 280 m^3 of air and only ca. 55 m^3

of oxygen, the combustion space required for pure oxygen is much less. Because nitrogen does not have to be expanded, the space requirement for a given thermal demand becomes minimal, adding to the effect of increased flame velocity. Although the combustion products, water and carbon dioxide are heated initially to a higher temperature, their rate of thermal decay over the cold charge is much more rapid than that of nitrogen, and they leave the furnace at a temperature that is the same or lower than with air combustion.

Pure oxygen combustion on an exclusive basis, however, is not an industrial practice because of economic and material properties limitations. To overcome these restrictions, the industrial gas industry has developed methods for oxygen enrichment and supplementary oxygen – fuel burners to aid many of the pyroprocess industries.

The most common approach is the use of oxygen to enrich the air – fuel burner combustion air. A diffuser in the air main introduces commercial oxygen so that the oxygen content of air reaching the burner is elevated above its normal 21 %. As oxygen is added to the air, the flow of air is reduced to maintain the desired combustion stoichiometry. In practice, the final oxygen concentration is held to 23 – 25 % because of refractory temperature limitations. With oxygen enrichment of air, the radiative heat-transfer mode is most affected.

The advantage of this method, besides simplicity, is that it affects the entire combustion system and helps overcome many deficiencies of existing air – fuel burners. Postmix burners (diffusion flames) usually lend themselves to these moderate enrichment levels without modification, but premix burners are severely limited because the increase in flame speed, imparted by additional oxygen, can result in flashback into the mixer.

Some of the constraints imposed on general oxygen enrichment by furnace design can be overcome by more selective techniques. Depending on the burner type used, oxygen may be introduced into the primary or atomizing air where liquid fuels are used. This more closely approaches the characteristics of pure oxygen combustion. Oxygen – fuel mixing is improved, and the oxygen is subject to less dilution by the secondary air and combustion products.

In some cases, pure oxygen is lanced into an air – fuel flame to obtain selective enrichment of the flame in the lower section adjacent to the furnace charge. The unaffected upper portion of the flame protects the furnace refractory by providing a shielding area against the intense radiation of the lower section. This procedure is frequently used in steel mill open hearths, glass melting furnaces, and rotary kilns.

Another variation in oxygen use is the impingement of an oxygen – fuel flame against the air – fuel flame. This is especially effective in stabilizing flames fired with hard-to-burn fuels such as coal or hogged wood. The air – fuel stream is selectively preheated and a portion of the fuel is volatilized. Selective preheating in this case means that the air – fuel stream receives an enthalpy increase from the highly directional oxygen – fuel burner, and since the oxygen – fuel flame is not radiative, the refractory and furnace contents are not directly heated by this flame. This technique is particularly suitable for rotary kilns [27].

A very effective method of oxygen use is to add supplementary oxygen – fuel burners to obtain all the advantages of pure oxygen combustion. These include accelerated melting, reduced fuel consumption, and reduced oxidation losses by virtue of shorter melt cycles. By adding 25 % additional fuel through oxygen – fuel burners, melt rates can be doubled while fuel consumption is reduced by about 30 %. The rapid heat transfer of the highly directional convective oxygen – fuel flame allows the placement of burners to add heat directly to the charge without heating the refractory. In most cases, the refractory temperature is lower than normal, and the flue gas temperature is unchanged despite the higher thermal input. This technique has found widespread acceptance in steel, copper, lead, and aluminum melting operations.

In 1856, Sir Henry Bessemer invented his process for the conversion of low sulfur and phosphorus hematite iron into steel. He produced malleable steel by blowing air through liquid metal; this resulted in the exothermic oxidation of carbon, silicon, and manganese. The later development of the basic Bessemer or Thomas Converter allowed the phosphoric limonite ore fields of France and Germany to be exploited. This process was the major steel-producing method employed in Europe until it

was superseded in the 1960s by top-blown oxygen converter processes [28].

An over-simplification of the chemistry, but a nevertheless useful way to understand the role of oxygen in steel production, is illustrated by the following equation:

$$Fe\,(l) + C\,(s) + 1/2\,O_2\,(g) \longrightarrow Fe\,(l) + CO\,(g) + Heat \quad (10)$$

Carbon is removed from solution in the iron by conversion to carbon monoxide.

Increased demand for steel during the 1940s and 1950s stimulated research into rapid steelmaking processes, which utilized tonnage oxygen being made available for the first time in economic quantities during the immediate postwar period. A number of novel oxygen-blown converter processes were developed, the most successful of which was the Austrian LD process (Linz and Donawitz). This process accounts for more than 80 % of the world's steelmaking capacity, but outside Europe it is usually referred to as BOS (basic oxygen steelmaking) or BOP (basic oxygen process).

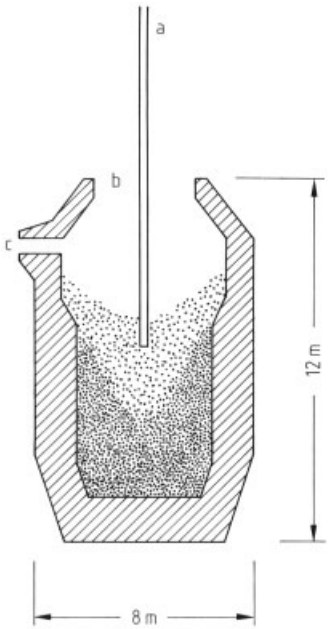

Figure 6. Oxygen-blown LD converter
a) Oxygen lance; b) Charging mouth; c) Tap hole

The LD converter (Fig. 6) is mounted on trunnions so that it can be tilted for loading and pouring. The process begins when the vessel is tilted downward for loading, then charged with hot iron from a blast furnace and up to 30 wt % scrap metal. After the vessel is turned upright, the oxygen lance is lowered and delivers 50–55 m^3 of oxygen per tonne of steel into the vessel, which contains about 250 t of charge. Initially, soft blowing, from 2–4 m above the metal level, is required to produce foaming slags for phosphorus removal. Later, the lance is lowered to within 1 m of the molten metal surface, and hard blowing follows, which results in greater penetration of oxygen into the bath [29]. The oxygen lance is water cooled with three or four exit nozzles at the tip. With oxygen pressure up to 1.5 MPa, supersonic exit velocities are achieved that provide the necessary penetration into the bath. At the end of the blowing period, which may be as short as 15 min, the lance is withdrawn and the vessel tilted again to tap off first the slag and then the molten steel.

7.2. Production of Chemicals

The trade-offs between air and commercial oxygen for the synthesis of chemicals via partial oxidation must consider capital investment, operating cost, reaction selectivity, environmental impact, safety, and the vent losses of product or raw materials associated with the discharge of large quantities of nitrogen after passage through the chemical reaction process. Where large volumes of air must be compressed and contained in process equipment, oxygen may be the oxidizing medium of choice. On an energy basis, an operating pressure of more than 690 kPa favors the use of oxygen. This choice is favored further if conversion rates per reactor pass are low, necessitating heavy recycle of the feed material.

Oxygen and feed material are never charged to a reactor without a diluent (ballast), which is required to adsorb and desorb the reactant onto and from the catalyst surface. It also discharges free radicals, whose unchecked propagation produces nonselective reactions. Its main function, however, is to act as a heat-transfer medium, cooling the catalyst by rejecting reaction heat to the surface of the externally cooled reaction tubes.

With air, the ballast must be nitrogen, but nitrogen is a poor thermal transfer fluid as well as a poor quencher of free radicals. With oxygen, a wider choice of ballast is available. Methane,

ethane, and carbon dioxide have all been used successfully, as have mixtures.

From an environmental point of view, the significant difference between using air and commercial oxygen is the volume of process effluent. In the production of vinyl chloride monomer, the volume of the vent stream from an air-based process is about 100 times larger than that from an oxygen-based process [30] because of the high nitrogen content (> 90 %). With oxygen, nitrogen diluent in the vent stream is eliminated, thereby making pollution abatement more manageable and economical, such as when the waste stream is either scrubbed or incinerated to remove or destroy contaminants.

Acetaldehyde is useful primarily as an intermediate for other chemicals, such as acetic acid, acetic anhydride, and butanol. Of the various precursors available, ethylene appears to be the least expensive [31]:

$$H_2C{=}CH_2 + 1/2\,O_2 \xrightarrow[130\,^\circ C]{PdCl_2,\ CuCl_2,\ H_2O} CH_3CHO \qquad (11)$$

This is known as the Wacker process. Commercial oxygen allows it to be conducted at lower pressure and in a single step, as opposed to a two-step process when air is used to regenerate the catalyst [32]. Acetaldehyde may be oxidized further to form acetic anhydride by contact with oxygen in the presence of cobalt or manganese acetate.

The oxidation of *o*-xylene to phthalic anhydride is performed in the gas phase, with a large excess of air:

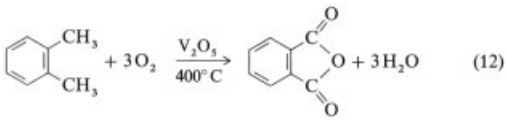

In contrast, the oxidation of *p*-xylene to 1,4-benzenedicarboxylic acid (terephthalic acid) is carried out with air oxidation in a solvent (acetic acid) in the presence of a cobalt catalyst. Oxygen enrichment of the air feed has been used successfully on a commercial scale to expand production capacity in this process.

In the production of maleic anhydride from benzene (Eq. 13) and formaldehyde from methanol (Eq. 14), air is favored as the source of oxygen. High-purity oxygen does not offer any particular process benefit that can offset its higher cost, and the vapor-phase oxidations are conducted with a mixture that is above the upper flammability limit in air [33]. Switching to oxygen would require an even greater excess of reactant for a given production per unit volume of catalyst in order to operate safely.

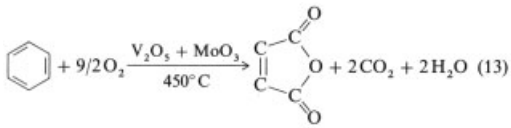

$$CH_3OH + 1/2\,O_2 \xrightarrow[575\,^\circ C]{Cu} HCHO + H_2O \qquad (14)$$

Ethylene oxide is an important petrochemical intermediate made from the direct oxidation of ethylene [34–36] over a silver-based catalyst in either fixed or fluid beds (Eq. 15).

$$H_2C{=}CH_2 + 1/2\,O_2 \xrightarrow[550\,^\circ C]{Ag_2O} H_2C{-}\overset{O}{\diagup\!\diagdown}{-}CH_2 \qquad (15)$$

About 60 % of the ethylene feedstock is converted to the desired product. Approximately 40 % of the feed is converted to carbon dioxide [37] and must be removed. Although feed losses to carbon dioxide are undesirable, with a pure oxygen process the carbon dioxide evolves in concentrated form and may be liquefied and sold as a coproduct, rather than vented to the atmosphere as in air-based processes.

Acrylonitrile from propene ammoxidation is the primary raw material for nylon 66. Commercial oxygen is used in this process.

$$H_2C{=}CHCH_3 + NH_3 + 3/2\,O_2 \xrightarrow[425\,^\circ C]{Bi_2Mo_3O_{12}} H_2C{=}CHCN + 3H_2O \qquad (16)$$

Poly(vinyl chloride) plastics have created enormous demand for the monomer vinyl chloride. The usual route [38] is by oxychlorination

of ethylene via the 1,2-dichloroethane intermediate.

$$2H_2C{=}CH_2 + Cl_2 + 1/2O_2 \xrightarrow[180\,^\circ C]{CuCl_2} 2H_2C{=}CHCl + H_2O \quad (17)$$

By operating with pure oxygen instead of air, vent volumes from this process have been greatly reduced [29]. Moreover, the stripping loss of ethylene from the process is greatly diminished, which further improves process economics.

The production of acetone and phenol from isopropylbenzene (cumene) involves the formation of cumene hydroperoxide via liquid-phase oxidation (commercial oxygen is used) and subsequent decomposition with acid to the two coproducts (Eq. 18). This process not only is the main source of phenol but also has become one of the most competitive processes for acetone production [39].

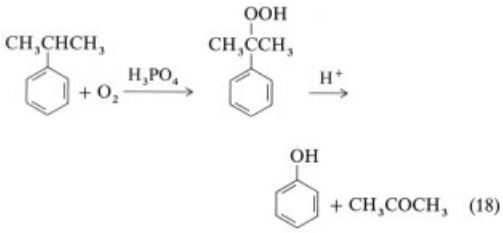

(18)

7.3. Oxygen Bleaching of Wood Pulp

(see also → Paper and Pulp)

The entry of oxygen into the pulp mill was prompted by efforts to find the best practical answer to water pollution problems associated with conventional bleaching methods: color and chlorinated organic compounds. In the face of ever-increasing restraints on effluent loading, oxygen bleaching (and oxygen-derived compounds such as ozone and hydrogen peroxide) appears to present the most satisfactory solution. If "zero discharge" ever becomes law, bleach plant effluents will have to be recycled within the mill. This is not practicable with chlorine or chloride-containing effluent streams. The feasibility of absorbing the oxygen stage effluent in a conventional kraft chemical recovery system was originally demonstrated in the Republic of South Africa [40] and has since been implemented in every major pulp-producing country.

Lignin comprises about 25 % of the wood that is converted into papermaking pulp and, for high-grade paper, must be removed from the cellulosic portion of the fibers. Oxygen delignification (removal of lignin with oxygen) proceeds with the hydroperoxide anion as the most active species. It may be formed initially by oxygen abstracting hydrogen atoms and producing organic radicals (Eq. 19 and 20)

$$8RH + O_2 \longrightarrow R^\bullet + HOO^\bullet \quad (19)$$

$$RH + HOO^\bullet \longrightarrow R^\bullet + H_2O_2 \quad (20)$$

that may subsequently form organic peroxides or terminate by any of the classical methods. The result is that side chains containing α-carbonyl and α-carbinol structures are eliminated by conversion into free acids and aldehydes, respectively, and ring opening with attendant formation of dicarboxylic acid groups [41] renders the larger lignin fragments soluble in the aqueous alkaline medium. These organic reaction products are subsequently washed from the cellulose fibers and, because they are chloride free, are ultimately reclaimed as fuel in the pulp mill's chemical recovery system.

Oxygen delignification of wood pulp is carried out at either high (30 %) or medium (12 %) pulp consistency. High-consistency pulp is relatively dry in appearance and presents a large fibrous network surface area to the oxygen atmosphere inside a continuous, fluffed-bed, downflow reactor. At medium consistency, the reaction rate is controlled by oxygen diffusion through the alkaline solution surrounding the pulp. Prior to entering the upflow reactor, oxygen is mixed with the pulp slurry under high shear, to fully disperse the oxygen within the fibrous network and aid the diffusion process by increasing the interfacial surface area. In either case, ca. 25 kg of oxygen and 25 kg of sodium hydroxide per tonne of pulp are applied.

7.4. Wastewater Treatment

The use of relatively high purity oxygen gas in the activated sludge process was established

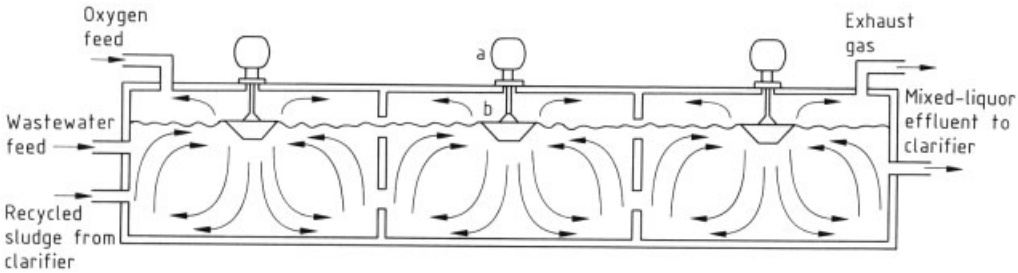

Figure 7. Three-stage UNOX system (Union Carbide)
a) Motor; b) Surface aerator

commercially in the mid-1960s with the successful introduction of the UNOX system (Union Carbide). Today, this process and subsequent variations are used worldwide to treat municipal sewage and industrial wastewater from diverse manufacturing operations [42].

The principle involved is not the increased level of dissolved oxygen that can be obtained with the use of oxygen; rather, it is the rate at which oxygen can be transferred into the mixed liquor. With higher oxygenation rates, the organic loading on the system can be increased substantially. By simultaneously increasing the organic loading and the mixed-liquor volatile suspended solid levels, a high degree of treatment efficiency can be obtained at the higher loading and higher oxygen transfer rate conditions.

The UNOX system is illustrated schematically in Figure 7. The system uses a covered, multistage oxygenation and reaction system in which the wastewater and activated sludge are contacted in a series of cocurrent, gas – liquid mixing stages. The system is operated at essentially atmospheric pressure, with relatively high purity oxygen gas, wastewater, and recycle sludge being introduced in the first stage and flowing cocurrently through successive contacting stages. Exhaust gas is vented from the last stage, and the mixed-liquor is settled in a conventional gravity clarifier. The oxygen feed-gas rate can be controlled automatically in direct proportion to the actual oxygen uptake rate in the system by means of a simple pressure controller in the first stage. As the oxygen uptake rate diminishes, the pressure in the first stage tends to increase and the oxygen feed rate is reduced. Alternatively, as the oxygen demand increases, the pressure in the first stage tends to decrease and the oxygen feed rate is increased. The number of stages employed, overall reactor geometrical configuration, tank depth, and specific type of gas – liquid contacting devices or aerators used in the stages can be economically optimized for each specific installation and process requirement. Standard surface aerators are typically used in most installations.

Oxygen is often used to augment air-based wastewater systems that have become mass-transfer limited. This condition often results years after a facility is commissioned because of increasing load on the plant.

7.5. Medicine

Oxygen is used extensively in medicine for therapeutic purposes, for resuscitation in asphyxia, and with other gases in anesthesia.

Adequate respiration consists of the uptake of sufficient amounts of oxygen and the elimination of adequate amounts of carbon dioxide to maintain their partial pressures, p_{O_2} and p_{CO_2}, in arterial blood at normal values. Respiratory failure is associated with disturbances in the exchange of oxygen and carbon dioxide between gas in alveoli and blood in pulmonary capillaries. Respiratory failure is not a disease but a disorder of function that can be caused by a variety of conditions that affect the lungs; in some instances the lungs are completely normal, as with an overdose of sedative drugs.

When respiratory failure is severe, death results from the central nervous system or cardiovascular consequences of hypoxia (oxygen deficiency). During emergencies, supplementary oxygen is given without worry about the concentration being used; in general, the higher the

concentration, the better. After the patient's condition has stabilized, attention is directed to administering oxygen in the lowest possible concentration required to correct the hypoxia. Any more oxygen than that required to increase the arterial p_{O_2} to a safe level exposes the patient to the direct toxicity of oxygen on the lung lining [43].

Supplementary oxygen can be administered to a patient in several ways. An oxygen tent provides the lowest inspired oxygen concentration per liter of oxygen flow. Its inefficiency is only in part related to the difficulties encountered in proper draping of the tent about the upper part of a bed, because rapid loss of oxygen tends to occur by diffusion through the canopy. Poor economy and bulk, together with increased nursing problems and fire hazard, limit its use to certain postoperative treatments.

Plastic, dome-shaped hoods enclose the head and neck of the patient, and are used with a recirculating carbon dioxide-absorbing and air conditioning system, such as those employed with oxygen tents. The smaller volume, better seal, and nonpermeability of the materials used combine to provide the hood with a higher ratio of inspired oxygen concentration to oxygen flow.

The concentration of oxygen in inspired air can be enriched by nasal cannulas, catheters, or prongs. These devices work well even when patients breath through their mouths. If more than 3 – 5 L/min is needed to achieve satisfactory arterial oxygenation, other methods of administration are advisable. At this rate, unhumidified oxygen has a drying effect on the nasal mucous membranes.

Venturi masks work on the principle of entrainment of a fixed proportion of air that mixes with the oxygen supplied and results in a constant inspired concentration (24, 28, 35, or 40 % O_2). The mask is somewhat uncomfortable and must be removed for eating and drinking.

When high concentrations of oxygen (40 – 80 %) are necessary for patients who are not intubated, reservoir masks are used. To ensure optimal efficiency of operation, the masks must be tight fitting to avoid leaks; because this often causes discomfort, delivering high concentrations of oxygen for long periods is difficult.

Mechanical ventilators have regulators that can be set to deliver an inspired oxygen concentration that ranges from 21 to 100 %. One of the best ways of ensuring that patients actually receive high concentrations of oxygen (60 – 100 %) is to use a mechanical ventilator connected to an endotracheal or tracheostomy tube.

For home therapy, three types of oxygen delivery are available. Large tanks of compressed oxygen (H or K cylinders) are easily obtainable but bulky. Smaller D and E cylinders are used for portability. The large tanks may be used with long nasal cannulas (20 m) to allow limited ambulation in the home. Liquid-oxygen reservoirs are more expensive but allow filling of relatively lightweight portable units that may be carried with a shoulder strap. This is the most convenient system for ambulation. Oxygen concentrators are probably the least expensive system in the long run. They require electric power and utilize either a molecular sieve or a membrane to produce oxygen-enriched air. The former systems are capable of delivering 85 – 90 % oxygen at 0.3 – 1 m^3/min; the latter, 30 – 40 % oxygen at 0.3 – 3 m^3/min. A backup cylinder is recommended in case of power or mechanical failure, and a small portable tank is useful for ambulation [44].

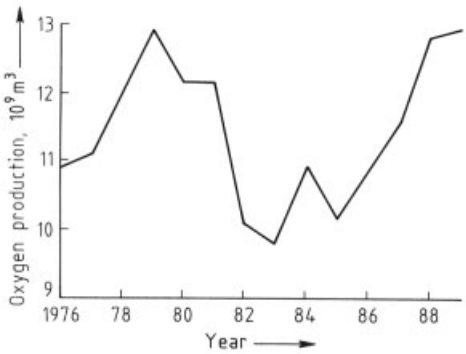

Figure 8. Oxygen production in the United States (U.S. Bureau of the Census, Current Industrial Reports M-28-C)

8. Economic Aspects

Commercial oxygen is one of the world's most significant industrial chemicals, although it is not traded on the international market. Because no fundamental limitations exist as to where oxygen can be generated, it is manufactured where needed, and its production rate is closely

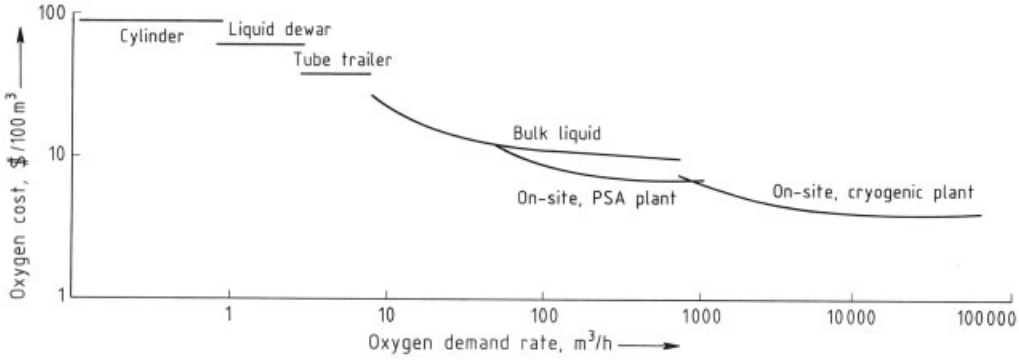

Figure 9. Cost of oxygen as a function of rate of use and supply mode

linked with the health of the local economy. Figure 8 illustrates the ca. 30 % decrease in U.S. production of oxygen during the recessionary period of the early 1980s.

Oxygen is supplied in several delivery modes, depending on the user's requirements. Unit pricing is directly related to rate of use and supply mode, diminishing over two orders of magnitude with increasing volume (Fig. 9).

9. Toxicology and Occupational Health

The inhalation of oxygen in concentrations greater than that found in air (21 %) should be commenced only under the direction of a physician, except in emergencies.

At atmospheric pressure, 100 % oxygen can cause pulmonary irritation and edema after 24 h of exposure. The earliest symptoms are pleuritic substernal pain and dry cough, occurring after only 6 h. Adult respiratory distress syndrome, which involves interstitial and intraalveolar fluid extravasation in the lung tissue, follows after 24 – 48 h. Other known toxic effects include retrolental fibroplasia, which has occurred in premature infants exposed to high concentrations of oxygen at birth; retinal circulatory injury; and erythrocyte hemolysis in adults [45].

When pure oxygen is inhaled at 2 atm or more, central nervous system (CNS) toxicity supervenes. Symptoms include nausea and vomiting; dizziness or vertigo; muscle twitching; vision changes; loss of consciousness; and generalized seizures. At 3 atm, CNS toxicity occurs in less than 2 h, and at 6 atm in only a few minutes. Physical exertion decreases the period before toxic symptoms and signs appear [45].

The toxic effects of oxygen are attributed to its free radical activity. Biological oxidation and autoxidation can convert molecular oxygen to the free-radical form, the superoxide anion (O_2^-). Other reactive intermediates include hydrogen peroxide (H_2O_2) and the hydroxyl radical ($OH^\bullet$). When these react with lipids, potent lipoperoxides form, which damage cell membranes and other vital cellular and subcellular structures [45].

10. References

1. W. Braker, A. L. Mossman: *The Matheson Unabridged Gas Data Book,* Matheson, East Rutherford, N.J., 1975.
2. M. C. Day, Jr., J. Selbin: *Theoretical Inorganic Chemistry,* 2nd ed., Van Nostrand Reinhold, New York 1969, p. 189.
3. B. C. Gates, J. R. Katzer, G. C. A. Schuit: *Chemistry of Catalytic Processes,* McGraw-Hill, New York 1979, p. 329.
4. R. T. Morrison, R. N. Boyd: *Organic Chemistry,* 2nd ed., Allyn and Bacon, Boston 1966, p. 46.
5. N. N. Semenov: *Chemical Kinetics and Chain Reactions,* Clarendon Press, Oxford 1935.
6. B. C. Gates, J. R. Katzer, G. C. A. Schuit: *Chemistry of Catalytic Processes,* McGraw-Hill, New York 1979, p. 335.
7. C. H. Simpson: *Chemicals from the Atmosphere,* Doubleday & Company, Garden City, New York 1969, p. 96.

8. F. Wolf, D. V. Eyer, M. Grenier: "Oxygen Plants: Ten Years of Development and Operation," *Chem. Eng. Prog.* **75** (1979) no. 7, 83 – 87.
9. C. W. Sharstrom in N. N. Li (ed.): *Recent Developments in Separation Science,* **vol. 2,** CRC, Boca Raton, Fla., 1980, p. 95.
10. C. E. Rodgers, M. Fels, N. Li in N. N. Li (ed.): *Recent Developments in Separation Science,* **vol. 2,** CRC, Boca Raton, Fla., 1980, p. 116.
11. Commodity Specification for Oxygen, CGA G-4.3, Compressed Gas Association, Arlington, Va., 1980.
12. Code of Federal Regulations, Food and Drugs, Title 21 CFR Parts 200, 201, 207, and 211, Superintendent of Documents, U.S. Government Printing Office, Washington, DC, 1990.
13. Compressed Medical Gases Guideline, Division of Drug Quality Compliance (HFN-320), U.S. Food and Drug Administration, Rockville, Md., 1990.
14. H. H. Willard, L. L. Merritt, Jr., J. A. Dean, F. A. Settle, Jr.,: *Instrumental Methods of Analysis,* 6th ed., D. Van Nostrand, New York 1981, p. 910.
15. H. H. Willard, L. L. Merritt, Jr., J. A. Dean, F. A. Settle, Jr.: *Instrumental Methods of Analysis,* 6th ed., D. Van Nostrand, New York 1981, p. 911.
16. H. H. Willard, L. L. Merritt, Jr., J. A. Dean, F. A. Settle, Jr.,: *Instrumental Methods of Analysis,* 6th ed., D. Van Nostrand, New York 1981, p. 912.
17. M. L. Hitchman: *Measurement of Dissolved Oxygen,* John Wiley & Sons, New York 1978.
18. NFP 50, Standard for Bulk Oxygen Systems at Consumer Sites, National Fire Protection Association, Quincy, Mass., 1985.
19. R. Lowrie: "Materials for Oxygen Service," *Chem. Eng.* **94** (1987) no. 6, 75 – 80.
20. Cleaning Equipment for Oxygen Service, CGA G-4.1, Compressed Gas Association, Arlington, Va., 1985.
21. *Handbook of Compressed Gases,* 3rd ed., Compressed Gas Association, Van Nostrand Reinhold, New York 1990, p. 532.
22. Code of Federal Regulations, Transportation, Title 49 CFR parts 100 – 199, Superintendent of Documents, U.S. Government Printing Office, Washington, DC, 1990.
23. *Handbook of Compressed Gases,* 3rd ed., Compressed Gas Association, Van Nostrand Reinhold, New York 1990, p. 533.
24. Pressure Relief Device Standards, CGA S-1.1, Part 1, Cylinders for Compressed Gases, Compressed Gas Association, Arlington, Va., 1989.
25. American National, Canadian and Compressed Gas Association Standard for Compressed Gas Cylinder Valve Outlet and Inlet Connections, ANSI/CSA/CGA V-1, Compressed Gas Association, Arlington, Va., 1987.
26. G. M. Barrow: *Physical Chemistry,* 2nd ed., McGraw-Hill, New York 1966, p. 44.
27. M. J. Kirschner: "Use of Oxygen in Kraft Lime Kilns Increases Capacity and Saves Fuel," *Pulp & Paper* **3** (1979) 181 – 185.
28. C. Moore, R. I. Marshall: *Modern Steelmaking Methods,* The Institute of Metallurgists, London 1980, p. 13.
29. C. Moore, R. I. Marshall: *Modern Steelmaking Methods,* The Institute of Metallurgists, London 1980, p. 39.
30. W. E. Wimer, R. E. Feathers: "Oxygen Gives Low Cost VCM," *Hydrocarbon Process.* **55** (1976) 81 – 84.
31. G. T. Austin: "The Industrially Significant Organic Chemicals – Part 1," *Chem. Eng.* **81** (1974) no. 2, 127 – 132.
32. R. Jira, W. Blau, D. Grimm: "Acetaldehyde Via Air or Oxygen," *Hydrocarbon Process.* **55** (1976) no. 3, 97 – 100.
33. R. Maux: "Air Best for Formaldehyde and Maleic Anhydride," *Hydrocarbon Process.* **55** (1976) no. 3, 90.
34. I. Kiguchi, T. Kumazawa, R. Nakai: "For EO: Air and Oxygen Equal," *Hydrocarbon Process.* **55** (1976) no. 3, 69 – 72.
35. B. DeMaglie: "Oxygen Best for EO," *Hydrocarbon Process.* **55** (1976) no. 3, 78 – 80.
36. M. Gans, B. J. Ozero: "For EO: Air or Oxygen?" *Hydrocarbon Process.* **55** (1976) no. 3, 73 – 77.
37. G. T. Austin: "The Industrially Significant Organic Chemicals – Part 5," *Chem. Eng.* **81** (1974) no. 9, 143 – 150.
38. G. T. Austin: "The Industrially Significant Organic Chemicals – Part 3," *Chem. Eng.* **81** (1974) no. 6, 87 –92.
39. P. R. Pujado, J. R. Salazar, C. V. Berger: "Cheapest Route to Phenol," *Hydrocarbon Process.* **55** (1976) no. 3, 91 – 96.
40. B. Coetzee: "Continuous SAPOXAL Bleaching – Operation, Technical Experience," *Pulp Pap. Mag. Can.* **75** (1974) June.

41. R. C. Eckert, H.-M. Chang, W. P. Tucker: "Oxidative Degradation of Phenolic Lignin Model Compounds with Oxygen and Alkali," *Tappi* **56** (1973) no. 6, 134–138.
42. J. R. McWhirter (ed.) *The Use of High-Purity Oxygen in the Activated Sludge Process,* **vol. 1,** CCR Press, West Palm Beach, Fla., 1978, p. 5.
43. J. B. Wyngaarden, L. H. Smith: *Cecil Textbook of Medicine,* 18th ed., **vol. 1,** W. B. Saunders, Philadelphia, Pa., 1988, p. 478.
44. R. E. Rakel: *Conn's Current Therapy,* W. B. Sauders, Philadelphia, Pa., 1988, p. 126.
45. *Effects of Exposure to Toxic Gases – First Aid and Medical Treatment,* 3rd ed., Matheson Gas Products, Secaucus, N.J., 1988.

Ozone

MARK J. KIRSCHNER, The BOC Group, Inc., Murray Hill, New Jersey, 07974 United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 3
5. **Environmental Protection** 5
6. **Analysis** 5
7. **Uses** 6
7.1. **Potable Water Filtration and Disinfection** 6
7.2. **Wastewater Treatment** 6
7.3. **Other Uses** 7
8. **Toxicology and Occupational Health** . 7
9. **References** 8

1. Introduction

Ozone [*10028-15-6*] was first acknowledged in 1840 by SCHONBEIN, who determined that the odor produced during sparking experiments was caused by an unknown compound. He also observed this odor as being the same as that produced after a lightening flash and named the substance ozone, after the Greek "ozein": to smell. But it was not until twenty years later that the new substance was revealed to be a triatomic allotrope of oxygen.

Formation of ozone is endothermic:

$$3\,O_2 \longrightarrow 2\,O_3 \quad \Delta H_f^0 = 144.8\,\text{kJ/mol}$$

Ozone is thermodynamically unstable and spontaneously reverts back into diatomic oxygen. This process is promoted by the presence of transition metals or their oxides.

An irritating pale blue gas, ozone is explosive and toxic, even at very low concentrations. At −111.9 °C it condenses to form a dark violet liquid, which freezes at −192.7 °C. In the Earth's stratosphere, it occurs naturally (5–10 ppm), protecting the planet and its inhabitants by absorbing ultraviolet radiation of wavelength 290–320 nm.

Ozone is a powerful oxidizing agent, capable of participating in many chemical reactions with inorganic and organic substances. Commercially, ozone has been applied as a chemical reagent in synthesis, used for potable water purification, as a disinfectant in sewage treatment, and for the bleaching of natural fibers.

2. Physical Properties

The average value of the nonlinear O–O–O bond angle is given as 116.5° from microwave investigations [1], [2], and 127° from electron diffraction studies [3], [4]. The discrepancy is no doubt due to the difficulty of preparing ozone samples free of oxygen, and there is always the chance that trace impurities of sensitizing substances are present. In addition, spectral studies on ozone are complicated by the possibility that the incident radiant energy excites the molecule. Calculations [5], [6] of the O–O bond length, based on spectral data, have resulted in 126 pm.

By analysis of the electronic structure [2], the molecule is considered to resonate between **1** and **2**.

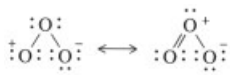

1 **2**

This resonance accounts for the electrophilic nature of the ozone molecule. The resonance does not permit unshared electrons, so unlike oxygen, ozone does not exhibit paramagnetism.

Ozone is soluble in many substances, forming either stable or metastable solutions. An example of a stable solution is that of ozone in liquid oxygen. Liquid ozone–oxygen mixtures are completely miscible in all proportions above −180.5 °C. When the mixture is cooled below the critical solution temperature (−180.5 °C), it separates into a dilute phase and a denser, concentrated phase. The composition of each phase

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_349

depends on the degree of cooling below the critical solution temperature.

Under practicable conditions in water, ozone is about 14 times more soluble than oxygen but forms a metastable solution. The stability is influenced by the presence of sensitizing impurities, such as heavy-metal cations and metal oxides, and by temperature and pressure. Because of this, literature values for the solubility of ozone vary.

Most of the solubility determinations have been performed with dilute ozone, and the values extrapolated to 100 % ozone. Table 1 lists the solubility of 100 % ozone in pure water, for the range of 0 – 60 °C [7].

Table 1. Solubility of ozone in water [7]

Temperature, °C	Bunsen coefficient ∗	O_3 solubility, kg/m^3
0	0.51	1.09
10	0.38	0.78
20	0.29	0.57
30	0.21	0.40
40	0.15	0.27
50	0.11	0.19
60	0.08	0.14

∗ m^3 O_3/m^3 H_2O (reduced to 0 °C and 101 kPa).

Some other physical properties of ozone are as follows:

M_r	48.0
bp (101 kPa)	− 111.9 °C
mp	− 192.7 °C
Critical temperature	− 12.1 °C
Critical pressure	5.53 MPa
Critical density	437 kg/m^3
Critical volume	1.471×10^{-4} m^3/mol
Density, gas (0 °C, 101 kPa)	2.144 kg/m^3
Density, liquid	
− 112 °C	1358 kg/m^3
− 183 °C	1571 kg/m^3
− 195.4 °C	1614 kg/m^3
Surface tension	
− 183 °C	3.84×10^{-2} N/m
− 195 °C	4.38×10^{-2} N/m
Heat capacity, liquid	
(− 183 to − 145 °C)	1884 J kg^{-1} K^{-1}
Heat capacity, gas	
0 °C	794 J kg^{-1} K^{-1}
25 °C	818 J kg^{-1} K^{-1}
Viscosity, liquid (−183 °C)	1.57×10^{-3} Pa · s
Heat of vaporization	15.2 kJ/mol
Heat of formation	144.8 kJ/mol

3. Chemical Properties

The chemistry of ozone is largely governed by its strongly electrophilic nature. Table 2 compares the oxidation potential of ozone with other strong oxidizing agents.

Table 2. Relative oxidation potentials

Species	Oxidation potential, eV
Fluorine	3.06
Hydroxyl radical	2.80
Nascent oxygen	2.42
Ozone	2.07
Hydrogen peroxide	1.77
Perhydroxyl radical	1.70
Hypochlorous acid	1.49
Chlorine	1.36

Much of ozone's organic chemistry is based on its attack on carbon double bonds (ozonolysis). The mechanism for the ozonolysis reaction proceeds through electrophilic addition to a double bond (**3**), forming an intermediate carbonium ion (**4**), which reacts further to form an ozonide (**5**). If water is present during this reaction or is later added, the ozonide undergoes hydrolysis to yield cleavage products. Ozonides are not isolated and purified commercially because they are explosive.

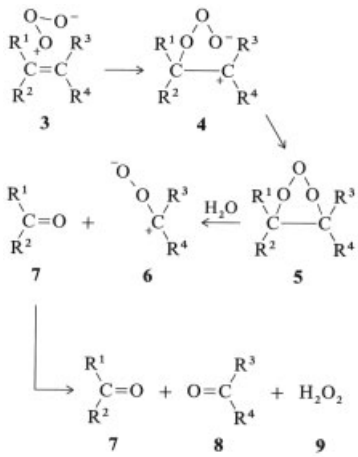

The mechanism proceeds through a zwitterion (**6**), a dipolar ionic species [8]. The cleavage products consist of two molecules each containing one carbonyl group (**7**,**8**), plus hydrogen

peroxide (**9**). Each carbonyl group is located at the sight of one of the original olefinic carbon atoms.

Ozone has proven useful as an oxidant in both laboratory and industrial scale organic synthesis. In many instances, ozone acts as a catalyst for another oxidant, usually oxygen. Chief among these are the preparation of aliphatic acids from the corresponding alcohols and olefins. Saturated hydrocarbons can be oxidized to yield a variety of products, depending on reaction conditions.

Ozonation can be used to convert trimethylhydroquinone to Vitamin E [9]. Fatty acids and their derivatives when treated with ozone yield a variety of oxidation products. Examples include azelaic and pelargonic acids from oleic acid [10], reductive ozonolysis of methyl oleate in methanol to produce aldehydes [11], and pelargonic acid from ozonation of methyl esters and amides of oleic and erucic acids [12].

Many other organic compounds are oxidized by ozone. Representative examples are the oxidation of diazo compounds [13], azides [14], phosphites [15–17], dinitrophenylhydrazones [18], nitriles [19], amines [20], [21], acetylenes [22–25], ethers [26], [27], organomercurials [28], silanes [29], oximes and Schiff bases [30], nitrones [31], [32], amides [33], organotin compounds [34], acetals [35], chloromethanes [36], imines [37], and alcohols [38], [39].

Ozonation reactions can be categorized into four groups, depending on how they are conducted: (1) conditions which permit isolation of the ozonide or other peroxidic product, (2) oxidative conditions yielding acids, (3) reductive ozonolysis to produce aldehydes or alcohols, and (4) ozonation in a reactive solvent (H_2O, RNH_2, ROH).

Oxidations with molecular oxygen are often catalyzed by ozone. Ethers can be prepared by treating petroleum distillates with ozone and oxygen in the presence of a metal salt [40]. By selecting the appropriate primary or secondary alcohols, a wide range of carboxylic acids may be synthesized [41]. Ozone induces the oxidative attack of oxygen on polytetrafluoroethylene [42]. Isobutene is converted to methacrolein [43], illustrating how unsaturated aldehydes may be formed when olefins are treated with oxygen containing a small amount of ozone.

4. Production

Commercially useful volumes of ozone are nearly always generated from atmospheric oxygen in a corona discharge, although small quantities are sometimes generated by ultraviolet irradiation of oxygen, either in air or as a solution in water. Ozone is always generated on a demand basis, at its point of use.

The dominant mechanism for the formation of ozone in a corona discharge is believed to proceed by the interaction of high-energy free electrons with oxygen molecules to form transient oxygen atoms [44]:

$$e^- \text{ (high energy)} + O_2 \longrightarrow 2\,O + e^- \text{ (low energy)} \quad (1)$$

Ozone then results from either a two- (2) or three-body (3) collision reaction, involving oxygen atoms and oxygen molecules:

$$O + O_2 \longrightarrow O_3 \quad (2)$$

$$O + O_2 + Z \longrightarrow O_3 + Z \quad (3)$$

where Z is another gaseous species such as nitrogen or carbon dioxide. The rate of ozone production may be increased or decreased by a given species Z, depending on its concentration in the feed gas. For example, as the nitrogen concentration in oxygen is increased from 0 to 7 vol % there is a slight increase in ozone production, due to nitrogen oxides produced in situ. Above 7 %, however, the rate decreases dramatically [45], due to the higher concentration of nitrogen oxides. In addition, the ozone production is further influenced by the effectiveness of the generator cooling system, since the competing ozone decomposition reactions (4) and (5) are thermally activated:

$$O + O_3 \longrightarrow 2\,O_2 \quad (4)$$

$$e^- + O_3 \longrightarrow O_2 + O + e^- \quad (5)$$

Also, the ozone yield is dependent on the oxygen content of the feed gas and the corona power density.

A corona is a low-current electrical discharge between two electrodes separated by a gas-filled

gap. In an ozone generator, there is also a dielectric material present, which, together with the gap between the electrodes, presents a capacitive load to the power supply. The high-voltage applied is on the order of the voltage breakdown of the dielectric and gap, which is a function of the dielectric material and its thickness, the width of the gap, and the gas which fills the gap and its pressure. Figure 1 shows a schematic of a corona cell.

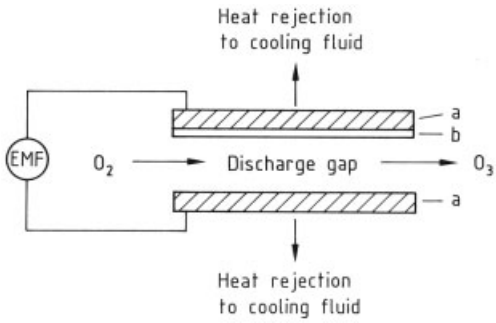

Figure 1. Schematic of corona discharge ozone apparatus
a) Electrode; b) Dielectric

High-energy electrons (6 – 7 eV) in the corona resonate between the electrodes at the frequency of the applied alternating current. To a point, higher frequencies increase the ozone yield of the generator, but there are practical limitations, such as heat removal, dielectric breakdown, and current saturation, which differ according to the design.

Figure 2 graphically represents the operation of a corona cell. During the initial voltage increase from t_0 to t_1, the voltage gradient in the gas space is less than that required to ionize the gas. Because the dielectric constant of the gas is much smaller than that of the dielectric material, the capacitance of the dielectric material is considerably greater than that of the gas gap. Therefore, at any time from t_0 to t_1, the voltage drop across the gap is considerably larger than that across the dielectric material.

At point t_1, the driving voltage reaches the corona start potential (V_{cs}), and the voltage across the gas space reaches its sparking potential (V_s). The voltage breakdown at this point establishes the corona. Because the gas in the gap becomes conductive with an impressed voltage of V_{cs}, the voltage drop across the gap itself remains essentially constant at V_s.

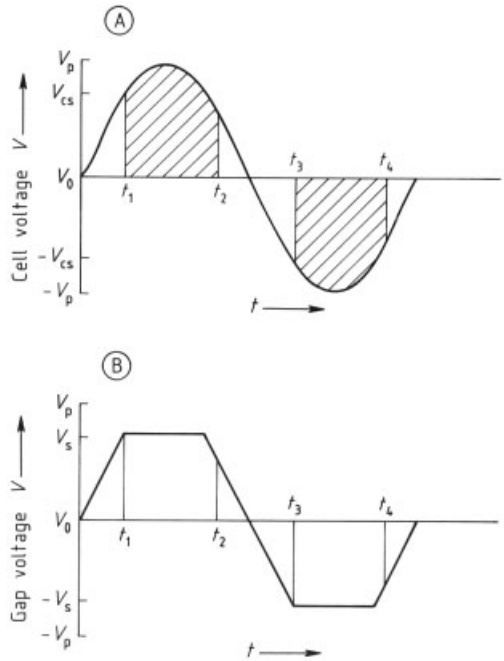

Figure 2. Driving voltage (A) and discharge gap voltage (B) waveforms

The ozone power equation (6) is the theoretical basis for the design of corona cells. For a given geometry, gap thickness, and gas pressure, the power drawn by the corona can be increased: (1) by operating at higher frequencies; (2) by employing thinner dielectrics with higher dielectric constants; or (3) by operating at higher peak driving voltages. Most generators are powered at 60 Hz, although medium-frequency (400 – 600 Hz) and high-frequency (nominally 2 kHz) equipment are available. Dielectric materials are commonly glass or ceramic, with thicknesses of 0.5 – 3 mm. Driving peak voltages are typically 8 – 30 kV [46].

$$P = 4C_d V_s f \left[V_p - \left(\frac{C_d + C_g}{C_d} \right) V_s \right] \tag{6}$$

P = corona power, W
C_d = dielectric capacitance, F
C_g = gap capacitance, F
V_s = gap sparking potential, V
V_p = peak driving potential, V
f = driving voltage frequency, Hz

Three types of ozone generators have been developed: the plate type, the shell and tube type, and the double-fluid-cooled tube type. The plate

type is air cooled and is only used in the smallest capacity models. The shell and tube type illustrated in Figure 3 is used in large ozone generators, employing multiple corona discharge tubes that are externally cooled with water. A variation, the double-fluid-cooled type, employs concentric gas and liquid conduits which results in both electrodes having surfaces that are cooled by water or by another coolant.

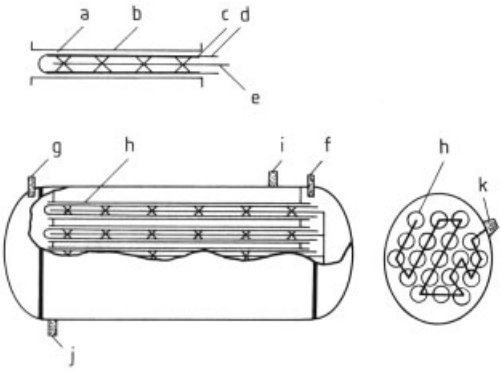

Figure 3. Horizontal shell and tube type ozone generator
a) Corona discharge gap; b) Electrode; c) Electrode; d) Dielectric (glass tube); e) Electric contact to electrode; f) Oxygen carrier gas inlet; g) Ozone outlet; h) Ozone generation tube; i) Coolant outlet; j) Coolant inlet; k) High-voltage terminal

Figure 4 shows the specific energy required for ozone generation from pure oxygen feed in large-scale units. For a feed of dry air, the specific energy, for a given weight ozone concentration, is about twice that of oxygen for the same concentration, production rates being equal.

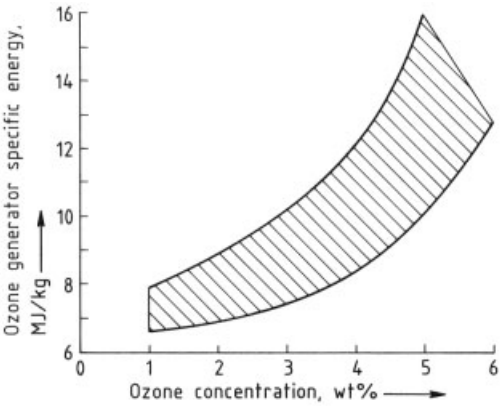

Figure 4. Specific energy as a function of ozone product concentration for representative ozone generators with dry oxygen feed

5. Environmental Protection

Ozone contacting systems are not 100 % efficient in mass transfer of ozone from the carrier gas stream. Therefore, vent gas from the contactor will contain some amount of waste ozone.

The simplest and least expensive method of treatment is dilution with a large volume of fresh air, venting through an elevated stack. Dilution, however, should not be used in populated areas or where atmospheric inversions occur.

Because ozone is a very powerful oxidant, it is readily reduced to oxygen by scrubbing with a solution of sodium bisulfite or other reductant. Large installations may find the cost of chemicals and their subsequent disposal cost to be relatively expensive, compared to other treatment methods.

Decomposition of ozone by granular activated carbon is similar to bisulfite reduction as both constitute chemical reactions in which the reagents are oxidized by the ozone, and are thereby depleted. Besides the cost of replacing consumed activated carbon, the ozone-containing stream may have to be compressed to overcome the pressure drop through the carbon bed.

Transition metals and their oxides catalyze the decomposition of ozone to oxygen. A small amount of catalyst is dispersed on high-surface-area media such as porous alumina. The ozone-containing waste gas is preheated to 50 – 150 °C, depending on the catalyst. Capital cost is relatively high and catalyst replacement is necessary every two years or so.

Ozone is thermally destroyed in 1 – 2 s at temperatures above 300 °C. Heat may be supplied by electric gas heaters or by direct fired natural gas. More than 50 % of the thermal input may be reclaimed by heat exchanging the final waste gas against the incoming ozone stream.

6. Analysis

In the application of ozone for various treatment processes, process control is normally based on the measurement of the initial ozone dose, or the dissolved residual ozone after some detention time has elapsed. Absorbed ozone is taken as the difference between the concentrations of

ozone in the contactor exhaust gas and the contactor feed gas.

The methods employed most often to analyze gaseous samples are the direct ultraviolet (UV) absorbance method [47], and chemiluminescence of ethylene [48]. In the past, it was common to employ a modification of the iodometric procedure [49], [50], though this is largely reserved for analyzing liquid samples.

Ozone measurement by UV radiation is based on its strong absorption in the vicinity of the 254 nm line of mercury lamps. On entering the analyzer, the sample is diverted into a catalytic converter which converts any ozone present to oxygen. This ozone-free sample now enters the detection chamber, and the UV absorbance is measured. This measurement serves as the reference.

The catalytic converter is then switched out of the gas stream and the detection chamber is filled with the ozone- containing sample, and the absorbance is measured again. The difference between the two measurements is then converted to the ozone concentration.

Analysis of ozone by ethylene chemiluminescence is based on the gas-phase reaction between ozone and ethylene to produce formaldehyde and oxygen with emission of light. The intensity of this photoemission is proportional to the concentration of ozone, and is measured by a photomultiplier tube.

When using the iodometric titration [51] for analyses of aqueous samples, care is required to account for interfering factors such as solution pH, temperature, and the presence of other interfering oxidizing substances. The procedure is based on the oxidation of iodide to iodine by ozone (7):

$$O_3 + 2\,I^- + H_2O \longrightarrow O_2 + I_2 + 2\,(OH)^- \qquad (7)$$

The liberated iodine can be determined by thiosulfate titration in the presence of a starch indicator (or by colorimetry, or amperometric titration).

7. Uses

7.1. Potable Water Filtration and Disinfection

Ozone has been used for the treatment of drinking water since 1906, when it was used in Nice, France for disinfection. Although many water treatment plants throughout the world still utilize ozone for disinfection, most modern plants use ozone as an oxidant to aid the filtration process. The benefits of ozonation include oxidation of inorganic and organic materials, and microfloculation for removal of turbidity or suspended solids [52].

In potable water supply systems in the United States, suspended matter is usually removed by coagulation, precipitation, and filtration. Typically, an inorganic salt or polyelectrolyte (or both) is added to aid floc formation during the coagulation step. In these operations, predisinfection with chlorine is normally practiced to prevent aquatic growths in the water, on the walls of the treatment units, and in the filter beds, and to kill bacteria and viruses.

Public concern about the quality of drinking water has risen sharply since the early 1970s. In part, due to the discovery that naturally occurring organic compounds were being converted to chloroorganics, principally trichlorome-thane, in chlorinated potable water. Eventually, the EPA specified a 0.1 g/m^3 upper limit for trihalomethanes (THM) in finished drinking water. Until this problem with chlorination was discovered, ozonation was virtually unused for treating municipal water systems in the United States.

Ozone's chief advantage over chlorine is that it not only kills bacteria, but it also destroys viruses and waterborne parasites. Furthermore, ozone removes odors and color from water, and leaves no residue. The applied dosage is ca. 1.5 g/m^3 of water. By the time the water reaches the user, the ozone is depleted since its half-life in a distribution system is about 25 min at ambient temperature. However, this also means that purified water in a distribution network could become contaminated, e.g., if ground water incursion occurs. For this reason, some chlorine is added to the ozone-treated water after filtration to assure its continued purity. By this stage, the organic materials have been removed, so formation of THM is not a concern.

7.2. Wastewater Treatment

Although chlorine is the most widely used reagent for treating wastewater prior to its dis-

charge, the advantages of ozone treatment are becoming increasingly important, creating demand for ozone in this application. In collecting waters, the presence of several carcinogenic organic compounds has been attributed to chlorinated discharges from industrial and municipal wastewater treatment plants. Ozone-treated wastewater is safe to the receiving water body and carries the maximum dissolved oxygen concentration. Along with disinfection, ozone also lowers the COD, color, and turbidity of the discharged water. The typical dosage rate is about 10 g/m^3 of effluent [53].

7.3. Other Uses

To lessen the environmental impact of wood pulp bleaching with chlorine and chlorine derivatives, ozone has been investigated [54] as a replacement for chlorine-based treatments on chemical pulps, chiefly kraft softwoods and hardwoods (→ Paper and Pulp). The bleaching response is comparatively high with respect to the amount of ozone applied, about 10 kg/t pulp, and the reaction is very fast giving practical bleaching times of several minutes. Although there are no commercial ozone bleaching units presently operating, the first start-up is anticipated by 1995.

Ozone has been used to replace traditional cooling water treatment chemicals in closed recycling systems [55] for large air conditioning units. Over several years operation, scaling did not occur, as evidenced by maintaining a good heat-transfer rate. In the electric power industry, ozone has been successfully applied in single pass cooling systems to prevent condenser biofouling [55]. The use of ozone was prompted by impending limitations on chlorine discharge from such operations.

High-purity water is important to certain taste-sensitive products such as beer and bottled water. Ozone is used by the beverage industry to remove residual organic compounds in feed water, which could contribute to taste and odor problems. In the microelectronics industry, highly purified water is required for wafer washing between the various process steps. Ozone protects these systems from biological fouling without causing ionic or microparticle contamination.

Odor control by ozone has been demonstrated for many diverse sources [56]. In gas-phase reactions, ozone oxidizes the odor-causing functional groups of many substances.

8. Toxicology and Occupational Health

Ozone, being an extremely powerful oxidizing agent, readily oxidizes a variety of functional groups in biochemicals. Studies indicate that free radical formation, lipid peroxidation, carbonyl and aldehyde formation, and oxidation of SH groups, are some of the major sites of attack [57]. Ozone readily reacts with olefinic compounds, particularly polyunsaturated lipids, forming unstable ozonides. Their decomposition results in the formation of toxic free radicals, which can in turn amplify the primary cytotoxic or tissue damage. The most commonly proposed cytotoxic mechanism is that ozone reacts with phospholipid micelles of plasma membranes and intracellular organelles causing alterations in membrane related functions such as intracellular activity of enzymatic systems [58].

While ozone is considered to be a toxic gas, there are factors which mitigate the immediate danger to individuals working with it. Toxicity is dependent on concentration and length of exposure. OSHA has set an 8-h TWA-PEL of 0.2 mg/m^3 (0.1 ppm) for ozone [59].

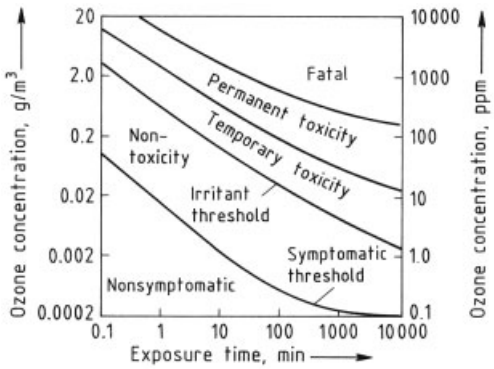

Figure 5. Human toxicity limits for ozone exposure

Figure 5 illustrates the relationship between various exposure levels and exposure time for humans [60]. An exposure of less than 0.2 mg/m^3 can be tolerated indefinitely, 2 mg/m^3

(1 ppm) can be tolerated for 8 min, and up to 8 mg/m^3 (4 ppm) can be tolerated for one minute without producing the symptoms of coughing, eye watering, and irritation of the nasal passages.

There are other factors which lessen the risk to personnel working with ozone. The odor threshold concentration for ozone is ca. 0.02 – 0.04 mg/m^3 (0.01 – 0.02 ppm). Thus ozone is generally detected by personnel before dangerous concentrations are reached. Moreover, once a critical concentration level is reached, the results are not immediately toxic but merely symptomatic. And since ozone is always consumed as it is generated, significant quantities are never on hand to cause a catastrophic event.

Where ozone generating equipment is installed indoors, the general practice is to install an ambient air ozone monitor set to sound alarm if the ambient ozone concentration rises to 0.2 mg/m^3. At that point, ozone generation ceases and exhaust fans vent the building. Personnel entering such contaminated areas must wear a self- contained breathing apparatus.

9. References

1. R. H. Hughes: "The Microwave Spectrum and Structure of Ozone," *J. Chem. Phys.* **21** (1953) 959 – 960.
2. R. Trambarlulo, S. Ghosh, C. Burrus, W. Gordy: "The Molecular Structure, Dipole Moment, and G Factor of Ozone from its Microwave Spectrum," *J. Chem. Phys.* **21** (1953) 851 – 855.
3. W. G. Penney, G. Sutherland, "The Relations Between the Form Force Constants and Vibration Frequencies of Triatomic Systems," *Proc. of the Royal Soc. (London)* **156 A** (1936) 654 – 678.
4. W. Shand, Jr., R. A. Spurr: "The Molecular Structure of Ozone," *J. Am. Chem. Soc.* **65** (1943) 179 – 181.
5. W. H. Eberhardt: "The Structure of Ozone," *J. Chem. Phys.* **14** (1946) 525 – 530.
6. M. J. Klein, F. F. Cleveland, A. G. Meister: "Force Constants Frequency Assignments, and Thermodynamic Properties of Ozone," *J. Chem. Phys.* **19** (1951) 1068 – 1069.
7. C. E. Thorp: "Physical and Pharmacological Properties," *Bibliography of Ozone Technology,* vol. 2, Armour Research Foundation, Illinois Institute of Technology, Chicago 1955, p. 43.
8. R. Criegee: "The Course of Ozonization of Unsaturated Compounds," *Record of Chemical Progress* **18** (1957) 111 – 120.
9. H. Mayer: "A New Synthesis of Vitamin E," *Chimia* **16** (1962) no. 11, 367 – 368.
10. G. M. M. Houben: "The Manufacture of Azelaic Acid and Pelargonic Acid in Technical Scale," *Chem. Weekbl.* **58** (1962) no. 22, 258 – 259.
11. E. H. Pryde, J. C. Cowan: "Aldehydic Materials by the Ozonation of Vegetable Oils," *J. Am. Oil Chem. Soc.* **39** (1962) no. 11, 496 – 500.
12. J. Pasero, M. Naudet: "Preparation of Functional Acids by the Oxidative Ozonolysis of Derivatives of Monounsaturated Fatty Acids," *Rev. Fr. Corps Gras* **10** (1963) 453 – 462.
13. A. M. Reader, P. S. Bailey, "Ozonation of Diphenyl-Diazomethane," *Chem. Ind. (London)* **40** (1961) 1620 – 1621.
14. D. H. R. Barton, L. R. Morgan, Jr.: "Photochemical Transformations, XII: The Photolysis of Azides," *J. Chem. Soc.* **2** (1962) 622 – 631.
15. F. J. Dillemuth: "The Reaction of Ozone with Acetylene Hydrocarbons," *Combust. Flame* **6** (1962) no. 3, 211 – 212.
16. W. H. Chang: "The Synthesis of Bicyclic Esters of Phosphorous Acids," *J. Org. Chem.* **29** (1964) no. 12, 3711 – 3712.
17. Q. E. Thompson: "Ozone Oxidation of Nucleophilic Substances. I: Tertiary Phosphite Esters," *J. Am. Chem. Soc.* **83** (1961) 846 – 851.
18. R. E. Erickson, A. M. Riebel, P. S. Bailey: " Ozonation of 2,4-Dinitrophenylhydrazones," *Justus Liebigs Ann. Chem.* **635** (1962) 129 – 138.
19. J. Klinot, A. Vystreil: "Beckmann Rearrangement of Triterpene-3-Ketoximers," *Collect. Czech. Chem. Commun.* **27** (1962) 377 – 386.
20. N. L. Bauld, J. A. Thompson, C. F. Hudson, P. S. Bailey: "Stereospecificity in Ozonide and Cross-Ozonide Formation," *J. Am. Chem. Soc.* **90** (1968) 1822 – 1830.
21. A. Maggiolo: "Ozone Oxidation and Peroxidation of Fatty Acids and Oils," *Annu. Symp. Am. Oil Chem. Soc.,* Newark, N. J., Feb. 1962.
22. P. S. Bailey, Y. Chang: "Rearrangements of Certain Alpha-Alkoxy-Alpha-Hydroperoxyacetophenones. Alcoholysis of

Acetic Benzoic Anhydride," *J. Org. Chem.* **27** (1962) 1192 – 1197.
23. E. Bernatek, T. Ledaal, S. Asen, "Ozonolysis of Acetylenedicarboxylic Acid," *Acta Chem. Scand.* **18** (1964) 1317 – 1319.
24. J. G. Cannon, L. L. Darko: "Ozonolysis of Certain Acetylinic Alcohols," *J. Org. Chem.* **29** (1964) 3419 –3420.
25. W. B. Demore, C.-L. Lin: "Intermediates in the Ozonation of Simple Alkynes," *J. Org. Chem.* **38** (1973) 985 – 989.
26. R. E. Erickson, R. T. Hansen, J. Harkins: "Mechanism of Ozonation Reactions. III: Ethers," *J. Am. Chem. Soc.* **90** (1968) 6777 – 6783.
27. C. C. Price, A. L. Tumolo: "The Course of Ozonation of Ethers," *J. Am. Chem. Soc.* **86** (1964) 4691 – 4694.
28. P. E. Pike, P. G. Marsh, R. E. Erickson, W. L. Waters: "The Ozonolysis of Organomercurials," *Tetrahedron Lett.* (1970) 2679 – 2682.
29. J. D. Austin, L. Spialter: "Substitution Effects in the Oxidative Cleavage of Organosilanes by Ozone," *Oxidation of Organic Compounds,* **vol. III,** The American Chemical Society, Washington, D. C., 1968, pp. 26 –31.
30. R. E. Erickson et al.: "Mechanism of Ozonation Reactions. IV: Carbon-Double Bonds," *J. Org. Chem.* **34** (1969) 2961 – 2965.
31. R. E. Erickson, T. M. Myszkiewicz: "Mechanism of Ozonation Reactions. Nitrones," *J. Org. Chem.* **30** (1965) 4326 – 4327.
32. J. Hamer, A. Macaluso: "Nitrones," *Chem. Rev.* **64** (1964) 473 – 495.
33. P. Kolsaker, O. Meth-Cohn: "Ozonation of Tertiary Amines and Amides. An Ozone-Induced Polonovski Reaction," *Chem. Commun.* **18** (1965) 423 – 425.
34. Yu. A. Aleksandrov, N. G. Sheyanov, "First Organotin Ozonide-Product of Ozonolysis of Trietylstannic Chloride," *J. Gen. Chem. USSR (Engl. Trans.)* **37** (1967) 2026.
35. P. Deslongchamps, C. Moreau: "Ozonolysis of Acetyls. (1) Ester Synthesis, (2) THP Ether Cleavage, (3) Selective Oxidation of B-Glycoside, (4) Oxidative Removal of Benzylidene and Ethylidene Protecting Groups," *Can. J. Chem.* **49** (1971) 2465 – 2467.
36. L. A. Hull, I. C. Hisatsune, J. Heicklen: "The Reaction of O_3 with CCl_2CH_2," *Can. J. Chem.* **51** (1973) 1504 – 1510.
37. J. K. Crandall, L. C. Crawly: "Reactions of Ketenimines with Peracids, Ozone, and Methylene-Transfer Reagents," *J. Org. Chem.* **39** (1974) 489 – 492.
38. A. Ya. Gerchikov, E. P. Kuznetsova, E. T. Denisov: "Kinetics and Rate Constants of the Reactions of Ozone with Alcohols in Carbon Tetrachloride," *Kinetika i Kataliz* **12** (1974) 509 – 511.
39. M. C. Whiting, A. J. N. Boltz, J. H. Parish: "Reaction between Ozone and Saturated Compounds," *Adv. Chem. Ser.* **77** (1968) 4 – 14.
40. H. Huber, US 3 052 727, 1974.
41. S. Miyazaki, Y. Suhara: "Preparation of Carboxylic Acids from Hydroxy Compounds by Oxidation with Ozone," *J. Am. Oil Chem. Soc.* **55** (1978) 536 –538.
42. F. Gozzo, A. Bergomi, G. Camaggi, V. Oprandi: "The Ozone-Induced Oxidation of Tetrafluorethylene," *Eur. Polym. J.* **6** (1970) 219 – 232.
43. A. C. Schotts, US 3 095 450, 1963.
44. S. W. Bensen: "Kinetic Considerations of Efficiency of Ozone Production in Gas Discharges," in "Ozone Chemistry and Technology," *Adv. Chem. Ser.,* 1959, no. 21, 405.
45. W. E. Cromwell, T. C. Manley: "Effect of Gaseous Diluents on Energy Yield of Ozone Generation from Oxygen," in "Ozone Chemistry and Technology," *Adv. Chem. Ser.,* 1959, no. 21, 304.
46. J. J. Carlins, R. G. Clark: "Ozone Generation by Corona Discharge," in R. G. Rice, A. Netzer (eds.): *Handbook of Ozone Technology and Applications,* vol. 1, chap. 2, Ann Arbor Science Publishers, Ann Arbor, Mich. 1982, p. 47.
47. K. A. Rehme, J. C. Puzak, M. E. Beard, G. F. Smith, R. J. Paur: "Evaluation of Ozone Calibration Procedures," U.S. EPA Report no. EPA-600/S 4 – 80 – 050. NTIS Report No. PB-81 – 118, 911 (1981).
48. S. Toby: " Chemiluminescence in the Reactions of Ozone," *Chem. Rev.* **84** (1984) 277 – 285.
49. B. E. Saltzman, N. Gilbert: "Iodometric Microdetermination of Organic Oxidants and Ozone," *Anal. Chem.* **31** (1959) 1914 – 1920.
50. *Standard Methods of the Examination of Water and Wastewater,* 13th ed., Public Health Association, Washington, D. C., 1971, pp. 107 – 143, 271 – 275.

51. P. R. Stroy, J. A. Alford, W. C. Ray, J. R. Burgess: "Mechanisms of Ozonolysis. A New Unifying Concept," *J. Am. Chem. Soc.* **93** (1971) 3044.
52. R. G. Rice, C. M. Robson, G. W. Miller, A. G. Hill: "Uses of Ozone in Drinking Water Treatment," *J. Am. Water Works Assoc.* **73** (1981) 44 – 57.
53. R. Junkins: "Ozone Disinfection: An O & M Survey of Selected Water and Wastewater Treatment Plants," *Sixth Ozone World Congress Proceedings,* The International Ozone Association, Vienna 1983.
54. R. P. Singh (ed.): *The Bleaching of Pulp,* Technical Association of the Pulp and Paper Industry, Atlanta, GA., 1979.
55. R. G. Rice (ed.): *Ozone Treatment of Water for Cooling Applications,* The International Ozone Association, Vienna 1981.
56. C. Nebel: *Industrial Odor Technology Assessment,* Ann Arbor Science Publishers, Ann Arbor, Mich., 1975 chap. 23.
57. A. J. DeLucia, M. G. Mustafa, M. Z. Hussain, C. E. Cross: "Ozone Interaction with Rodent Lung. III: Oxidation of Reduced Glutathione and Formation of Mixed Disulfides between Protein and Nonprotein Sulfhydryls," *J. Clin. Invest.* **55** (1975) 794.
58. F. G. Hueter, M. Fritzhand: "Oxidants and Lung Biochemistry. A Brief Review," *Arch. Intern. Med.* **128** (1971) 48.
59. F. J. Kelly, W. E. Gill: "Ozone Poisoning," *Arch. Environ. Health* **10** (1965) 517 – 519.
60. C. E. Thorp: "The Toxicity of Ozone. A Report and Bibliography," *Ind. Med. Surg.* **19** (1950) 45 – 57.

Packaging Materials → Films
Packaging Materials → Foods, 4.Packaging
Packaging Materials → Paper and Pulp
Paint and Varnish Removers → Paints and Coatings

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Paints and Coatings

Standard Article

Dieter Stoye[1], Werner Funke[2], Lutz Hoppe[3], Jürgen Hasselkus[4], Larry G. Curtis[5], Klaus Hoehne[6], Hans-Joachim Zech[1], Peter Heiling[7], Masaaki Yamabe[8], Klaus Dören[1], Hans Schupp[9], Rolf Küchenmeister[6], Martin Schmitthenner[10], Wolfgang Kremer[11], Wolfhart Wieczorrek[6], Hans Gempeler[12], Wolfgang Schneider[12], James W. White[13], Anthony G. Short[13], Werner J. Blank[14], Leonard J. Calbo[14], Dieter Plath[15], Friedrich Wagner[16], Werner Haller[17], Karl-Martin Rödder[18], Hans-Joachim Streitberger[19], Edmund Urbano[20], Richard Laible[21], Bernd D. Meyer[22], Engin Bagda[23], Frederick A. Waite[24], Michel Philips[25], Klaus Köhler[11], Peter Simmendinger[12], Wolfgang Roelle[26], Wilfried Scholz[27], Andreas Valet[12], Mario Slongo[12], Thomas Molz[18], Rainer Hiller[28], Klaus Werner Thomer[29], Klaus Vogel[30], Ulrich Schernau[31], Bernhard Hüser[31], Alfred Brandt[32], Alex Milne[33], Helmut Weyers[32], Wolfgang Plehn[34], Hanns-Adolf Lentze[35]

[1]Hüls AG, Marl, Federal Republic of Germany
[2]Institut für Technische Chemie, Stuttgart, Federal Republic of Germany
[3]Wolff Walsrode AG, Walsrode, Federal Republic of Germany
[4]Krahn Chemie GmbH, Hamburg, Federal Republic of Germany
[5]Eastman Chemical Products, Kingsport, Tennessee, United States
[6]Bayer AG, Leverkusen, Federal Republic of Germany
[7]Wacker-Chemie GmbH, Burghausen, Federal Republic of Germany
[8]Asahi Glass Co. Ltd., Hazawa-cho, Kanagawa-ku, Yokohama, Japan
[9]BASF AG, Ludwigshafen, Federal Republic of Germany
[10]Hüls AG, Witten, Federal Republic of Germany
[11]Bayer AG, Krefeld-Uerdingen, Federal Republic of Germany
[12]Ciba-Geigy AG, Basel, Switzerland
[13]Dow Corning Ltd., Barry, South Glamorgan, CF6 7YL, United Kingdom
[14]King Industries, Norwalk, Connecticut, United States
[15]Hoechst AG, Wiesbaden, Federal Republic of Germany
[16]Sika-Chemie, Stuttgart, Federal Republic of Germany
[17]Henkel KGaA, Düsseldorf, Federal Republic of Germany
[18]Hüls AG, Troisdorf, Federal Republic of Germany
[19]BASF Corporation, Southfield, Michigan, United States
[20]Vianova Kunstharz AG, Graz, Austria
[21]Akzo Coatings GmbH, Stuttgart, Federal Republic of Germany
[22]VP Resicoat GmbH, Reutlingen, Federal Republic of Germany
[23]Deutsche Amphibolin-Werke, Ober-Ramstadt, Federal Republic of Germany
[24]ICI Paints, Slough, Berkshire SL2 5DS, United Kingdom
[25]Radcure Specialties, Drogenbos, Belgium
[26]Bassermann & Co., Mannheim, Federal Republic of Germany
[27]BYK-Chemie GmbH, Wesel, Federal Republic of Germany
[28]BASF Lacke + Farben AG, Münster-Hiltrup, Federal Republic of Germany
[29]Opel, Bochum, Federal Republic of Germany
[30]Herberts GmbH, Wuppertal, Federal Republic of Germany
[31]BASF Lacke + Farben AG, Münster, Federal Republic of Germany
[32]ICI Lacke Farben GmbH, Hilden, Federal Republic of Germany
[33]Courtaulds Coatings Ltd., Felling, Gateshead NE10 0JY, United Kingdom
[34]Umweltbundesamt, Berlin, Federal Republic of Germany
[35]CEPE, Brussels, Belgium

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a18_359.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Fundamental Concepts
1.2. Historical Development
1.3. Composition of Paints
1.3.1. Binders and Resins
1.3.2. Plasticizers
1.3.3. Pigments and Extenders
1.3.4. Paint Additives
1.3.5. Solvents
1.4. Paint Application
1.5. Drying and Film Formation
1.6. Multicoat Systems
1.7. Economic Aspects
1.8. Future Outlook
2. Types of Paints and Coatings (Binders)
2.1. Oil-Based Coatings
2.2. Cellulose-Based Coatings
2.2.1. Nitrocellulose Lacquers
2.2.1.1. Raw Materials
2.2.1.2. Application and Uses
2.2.2. Organic Cellulose Ester Coatings
2.2.2.1. Cellulose Acetate Butyrate

2.2.2.2. Cellulose Acetate Propionate
2.3. Chlorinated Rubber Coatings
2.3.1. Starting Products
2.3.2. Chlorinated Rubber Paints
2.3.3. Chlorinated Rubber Combination Paints
2.4. Vinyl Coatings
2.4.1. General Properties
2.4.2. Coatings Based on Polyolefins and Polyolefin Derivatives
2.4.3. Poly(Vinyl Halides) and Vinyl Halide Copolymers
2.4.3.1. Poly(Vinyl Chloride) and Vinyl Chloride Copolymers
2.4.3.2. Vinylidene Chloride Copolymers
2.4.3.3. Fluoropolymer Coatings
2.4.4. Poly(Vinyl Esters)
2.4.4.1. Solid Resins
2.4.4.2. Dispersions
2.4.5. Poly(Vinyl Alcohol)
2.4.6. Poly(Vinyl Acetals)
2.4.7. Poly(Vinyl Ethers)
2.4.8. Polystyrene and Styrene Copolymers
2.5. Acrylic Coatings
2.6. Alkyd Coatings
2.6.1. Alkyd Resin Binders and Uses
2.6.2. Additional Raw Materials
2.6.3. Production
2.6.4. Environmental and Health Protection Measures
2.7. Saturated Polyester Coatings
2.7.1. Properties
2.7.2. Production of Polyester Resins and Paints
2.7.3. Cross-Linking of Polyester Resins
2.7.4. Uses
2.8. Unsaturated Polyester Coatings
2.8.1. Unsaturated Polyester Binders
2.8.2. Other Raw Materials
2.8.3. Production, Properties, and Uses
2.8.4. Transportation and Handling
2.9. Polyurethane Coatings
2.9.1. Raw Materials
2.9.2. Polyurethane Systems
2.9.2.1. One-Pack Systems
2.9.2.2. Two-Pack Systems
2.9.3. Properties and Uses
2.10. Epoxy Coatings
2.10.1. Epoxy Resin Types
2.10.2. Curing Agents
2.10.3. Chemically Modified Epoxy Resins
2.10.4. Uses
2.10.4.1. Curing at Ambient Temperature
2.10.4.2. Curing at Elevated Temperature
2.10.4.3. Radiation Curing
2.10.4.4. Occupational Health
2.11. Silicone Coatings
2.12. Urea, Benzoguanamine, and Melamine Resins for Coatings
2.13. Phenolic Resins for Coatings
2.13.1. Resols
2.13.2. Novolacs
2.13.3. Modified Phenolic Resins
2.14. Asphalt, Bitumen, and Pitch Coatings
2.14.1. Asphalt and Asphalt Combination Coatings
2.14.2. Bitumen Coatings
2.14.3. Bitumen Combination Coatings
2.14.4. Pitch Coatings
2.14.5. Pitch Combination Coatings
2.15. Silicate Coatings
2.15.1. Water Glass Coatings
2.15.2. Alkyl Silicates
3. Paint Systems
3.1. Solventborne Paints
3.1.1. General Information
3.1.2. Properties and Raw Materials
3.1.3. Environmental Protection and Application Technology
3.2. Solvent-Free and Low-Solvent (High-Solids) Paints
3.2.1. Principles
3.2.2. Production and Uses
3.3. Waterborne Paints
3.3.1. Properties
3.3.2. Production and Application

3.3.3. Uses and Environmental Aspects
3.4. Coating Powders
3.4.1. Introduction and Economic Importance
3.4.2. Production
3.4.3. Properties
3.4.4. Testing
3.4.5. Storage and Transportation
3.4.6. Environmental Aspects and Safety
3.4.7. Uses
3.5. Waterborne Dispersion Paints (Emulsion Paints)
3.6. Nonaqueous Dispersion Paints
3.7. Radiation-Curing Systems
3.7.1. Introduction
3.7.2. Radiation-Curable Systems Based on Acrylates
3.7.3. Equipment
3.7.4. Fields of Application
3.8. Electrodeposition Paints
4. Pigments and Extenders
4.1. Inorganic Pigments
4.2. Organic Pigments
4.3. Extenders
4.3.1. Introduction
4.3.2. Properties
4.3.3. Modification of Extenders
5. Paint Additives
5.1. Defoamers
5.2. Wetting and Dispersing Additives
5.3. Surface Additives
5.4. Driers and Catalysts
5.5. Preservatives
5.6. Rheology Additives
5.7. Light Stabilizers
5.8. Corrosion Inhibitors
5.9. Use and Testing of Additives
6. Paint Removal
6.1. Paint Removal from Metals
6.1.1. Chemical Paint Removal
6.1.2. Thermal Paint Removal
6.1.3. Mechanical and Low-Temperature Paint Removal
6.2. Paint Removal from Wood and Mineral Substrates
7. Production Technology
7.1. Principles
7.2. Paint-Making Processes
7.2.1. Varnishes
7.2.2. Paints
7.2.3. Powder Coatings
7.3. Apparatus
7.3.1. Mixers
7.3.2. Dissolvers
7.3.3. Kneaders and Kneader Mixers
7.3.4. Media Mills
7.3.5. Roller Mills
7.3.6. Auxiliary Equipment
8. Paint Application
8.1. Types of Substrate
8.2. Pretreatment of Substrate Surfaces
8.2.1. Pretreatment of Metallic Substrates
8.2.1.1. Cleaning
8.2.1.2. Degreasing
8.2.1.3. Formation of Conversion Layers
8.2.2. Pretreatment of Plastics
8.2.3. Pretreatment of Wood
8.3. Application Methods
8.3.1. Spraying (Atomization)
8.3.2. Electrostatic Atomization
8.3.3. Dipping
8.3.4. Miscellaneous Wet Paint Coating Methods
8.3.5. Powder Coating
8.3.6. Coating of Plastics and Wood
8.4. Paint Curing Methods
9. Properties and Testing
9.1. Properties of Coating Materials
9.2. Properties of Coatings
9.2.1. Films for Testing
9.2.2. Optical Properties
9.2.3. Mechanical Properties

9.2.4. Chemical Properties
9.2.5. Weathering Tests
10. Analysis
10.1. Analysis of Coating Materials
10.1.1. Separation of the Coating Material into Individual Components
10.1.2. Analysis of Binders
10.1.3. Analysis of Pigments and Extenders
10.1.4. Analysis of Solvents
10.1.5. Analysis of Additives
10.2. Analysis of Coatings
11. Uses
11.1. Coating Systems for Corrosion Protection of Large Steel Constructions (Heavy-Duty Coatings)
11.2. Automotive Paints
11.2.1. Car Body Paints
11.2.2. Other Automotive Coatings
11.3. Paints Used for Commercial Transport Vehicles
11.3.1. Railroad Rolling Stock
11.3.2. Freight Containers
11.3.3. Road Transport Vehicles
11.3.4. Aircraft Coatings
11.4. Marine Coatings
11.4.1. Substrate, Surface Preparation, and Priming
11.4.2. Ship Paint Systems
11.4.3. Fouling and Antifouling
11.5. Coil Coating
11.6. Coatings for Domestic Appliances
11.7. Coatings for Packaging (Can Coatings)
11.8. Furniture Coatings
11.9. Coatings for Buildings
11.9.1. Exterior-Use Coatings
11.9.2. Interior-Use Coatings
12. Environmental Protection and Toxicology
12.1. Clean Air Measures
12.2. Wastewater
12.3. Solid Residues and Waste
12.4. Toxicology
13. Economic Aspects

[Top of Page]

1. Introduction

Dieter Stoye

1.1. Fundamental Concepts

Paints or coatings are liquid, paste, or powder products which are applied to surfaces by various methods and equipment in layers of given thickness. These form adherent films on the surface of the substrate.

Film formation can occur physically or chemically. Physical film formation from liquid coatings is known as drying, whereas for powder coatings, it is a melting process. Drying is always associated with evaporation of organic solvents or water. Physical film formation is only possible if the coating components remaining on the substrate are solid and nontacky. Chemical film formation is necessary if the coating components are liquid, tacky, or pasty; conversion to a solid nontacky film takes place by chemical reaction between the components. The reactive components can be constituents of the coating, and the reaction can be initiated by energy (heat or radiation) after application of the coating. However, it is also possible to add a reaction partner while applying the coating (multipack paints). A special case of chemical film formation is the oxidation of coating component(s) by atmospheric oxygen (air drying). Physical and chemical film formation are often combined, e.g., in solvent-containing stoving paints, where the first stage is solvent evaporation, after which the film is cured by stoving. The properties of a paint are determined by its qualitative and quantitative composition, suitable choice of which enables the viscosity, electrical conductivity, and drying behavior to be matched to the application conditions. Also, the properties of the coating film (luster, elasticity, scratch resistance, hardness, adhesion, and surface structure) are determined by the paint properties. However, the condition of the substrate surface (cleanliness and freedom from dust and grease) is also important.

Coatings must fulfill many requirements. They protect the substrate against corrosion, weathering, and mechanical damage; have a decorative function (automotive coatings, household appliances, furniture); provide information (traffic signs, information signs, advertising); or have other specific properties.

“Coating” is a general term denoting a material that is applied to a surface. “Paint” indicates a pigmented material, while “varnish” refers to a clear lacquer (ISO 4618/1, DIN 55 945).

1.2. Historical Development

The earliest evidence of well-preserved prehistoric paintings, dating from the 16th millenium B.C., can be found in caves in Southern France (Font-de-Gaume, Niaux, Lascaux), Spain (Altamira), and South Africa. The colors used were pure oil paints prepared from animal fat mixed with mineral pigments such as ocher, manganese ore (manganese dioxide), iron oxide, and chalk. The oldest rock paintings from North Africa (Sahara, Tassili n'Ajjer) date from between the 5th and the 7th millennium B.C. Many examples of paintings from Babylon, Egypt, Greece, and Italy dating from the 1st and 2nd millenium B.C. are also known.

The first painted objects come from China. Furniture and utensils were covered with a layer of paint in an artistic design. The oldest tradition work dates from around 200 B.C. The lacquer used was the milky juice from the bark of the lacquer tree (*Rhus vernicifera*). This was colored black or red with minerals, and later also with gold dust or gold leaf.

The oldest recipe for a lacquer, from linseed oil and the natural resin sandarac, dates from 1100 A.D. and was due to the monk ROGERUS VON HELMERSHAUSEN. Natural products such as vegetable oils and wood resins remained the most important raw materials for paint production, into the early 1900s. Only the introduction of faster production equipment such as belt conveyors made the development of new paints necessary. Initially, the rapid-drying binder used was nitrocellulose, which after World War I could be manufactured on a large scale in existing guncotton plants. Phenolic resins were the first synthetic binders (ca. 1920), followed by the alkyd resins (1930). The large number of synthetic binders and resins now available are tailored for each application method and area of use. These paint raw materials are based on petrochemical primary products. Vegetable and animal oils and resins are now seldom used in their natural form, but only after chemical modification. The tendency to use such “renewable” raw materials is increasing. Consumer demand has led to a marked renaissance of natural products (“biopaints”).

The use of organic solvents in paint technology was linked to the development of modern rapid-drying binders. Whereas the liquid components previously used in coatings were vegetable oils or water and possibly ethanol, it now became necessary to use solvent mixtures to give accelerated drying and optimized paint-application properties. Production of a wide range of solvents began worldwide in the chemical industry in the 1920s.

Methods of applying paints also underwent major changes in the 1900s. Whereas up to this time coatings were applied manually with a brush, even in industry, this technique is today only used in the handicraft and DIY areas. Modern mechanized and automated application methods are used today for industrial-scale application because of greater efficiency, low material losses, qualitatively better results, and lower labor costs. They include high-pressure spraying using compressed air or electrostatic charging, modern automatic and environmentally friendly dipping and electrophoretic processes, and application by rollers.

Problems of environmental pollution also followed from the introduction of solvents. These were recognized by the late 1960s and became the subject of development work. Waterborne coatings, low-solvent coatings, solvent-free powder coatings, and new radiation-curing coating systems with reactive solvents that are bound chemically during the hardening process were developed. These environmentally friendly coating systems have gained a considerable market share. However, in some areas solvent-containing coatings are difficult to replace without affecting quality. For this reason, solvent-recycling and solvent-combustion plants have been developed to recover or incinerate the solvents in the waste air.

1.3. Composition of Paints

Paints are made of numerous components, depending on the method of application, the desired properties, the substrate to be coated, and ecological and economic constraints. Paint components can be classified as volatile or nonvolatile.

Volatile paint components include organic solvents, water, and coalescing agents. Nonvolatile components include binders, resins, plasticizers, paint additives, dyes, pigments, and extenders. In some types of binder, chemical hardening can lead to condensation products such as water, alcohols, and aldehydes or their acetals, which are released into the atmosphere, thus being regarded as volatile components.

All components fulfill special functions in the liquid paint and in the solid coating film. Solvents, binders, and pigments account for most of the material, the proportion of additives being small. Low concentrations of additives produce marked effects such as improved flow behavior, better wetting of the substrate or pigment, and catalytic acceleration of hardening.

Solvents and pigments need not always be present in a coating formulation. Solvent-free paints and pigment-free varnishes are also available.

The most important component of a paint formulation is the binder. Binders essentially determine the application method, drying and hardening behavior, adhesion to the substrate, mechanical properties, chemical resistance, and resistance to weathering.

1.3.1. *Binders and Resins*

Binders are macromolecular products with a molecular mass between 500 and ca. 30 000. The higher molcular mass products include cellulose nitrate and polyacrylate and vinyl chloride copolymers, which are suitable for physical film formation. The low molecular mass products include alkyd resins, phenolic resins, polyisocyanates, and epoxy resins. To produce acceptable films, these binders must be chemically hardened after application to the substrate to produce high molecular mass cross-linked macromolecules.

Increasing relative molecular mass of the binder in the polymer film improves properties such as elasticity, hardness, and impact deformation, but also leads to higher solution viscosity of the binder. While the usefulness of a coating is enhanced by good mechanical film properties, low viscosity combined with low solvent content are also desirable for ease of application and for environmental reasons. Therefore, a compromise is necessary.

The low molecular mass binders have low solution viscosity and allow low-emission paints with high solids contents or even solvent-free paints to be produced. Here, the binder consists of a mixture of several reactive components, and film formation takes place by chemical drying after application of the paint. If chemical hardening occurs even at room temperature, the binder components must be mixed together shortly before or even during application (two- and multicomponent systems).

Today, most binders are synthetic resins such as alkyd or epoxy resins.

The natural resin most commonly used as a binder today is rosin (colophony; ⟶ Resins, Natural), which is often tailored by chemical modification to suit specific applications. Also, many synthetic hard resins, mainly based on cyclohexanone, acetophenone, or aldehydes, are used in the paints industry. Hard resin binders increase the solids content, accelerate drying, and improve surface hardness, luster, and adhesion.

Most synthetic binders are softer and more flexible than hard resins. Consequently, they impart good elasticity, impact resistance, and improved adhesion, even to critical undercoats, as well as offering adequate resistance to weathering and chemicals. These binders are produced with a property profile tailored to suit particular application methods and to comply with a range of technical requirements, including environmental protection, low toxicity, and suitability for recycling and disposal.

1.3.2. *Plasticizers*

Plasticizers (⟶ Plasticizers) are organic liquids of high viscosity and low volatility. The esters of dicarboxylic acids (e.g., dioctyl phthalate) are well-known examples. Plasticizers lower the softening and film-forming temperatures of the binders, They also improve flow, flexibility, and adhesion properties. Chemically, plasticizers are largely inert and do not react with the binder components. Most binders used today are inherently flexible and can be regarded as “internally plasticized” resins. For this reason, use of plasticizers has declined.

1.3.3. *Pigments and Extenders*

Pigments and extenders in coatings are responsible for their color and covering power, and in some cases give the coating film improved anticorrosion properties. Pigments and extenders are finely ground crystalline solids that are dispersed in the paint. They are divided into inorganic, organic, organometallic, and metallic pigments. By far the most commonly used pigment is titanium dioxide. As a rule, mixtures of pigments are used for technical and economic reasons. The hiding power and tinting strength of a paint depend on the particle size of the pigment. The usual size range aimed at is 0.1 – 2.0 µm, which means that the pigment has a high surface area that must be wetted as effectively as possible by the binder components to give the coating film good stability, weathering resistance, and luster. This is achieved by bringing the pigment and binder into intimate contact under the influence of high shear forces. The high hiding power of some pigments enables them to be partially replaced by the cheaper extenders such as barium sulfate, calcium carbonate, or kaolin. Extenders have a particle size distribution similar to that of the pigments and are incorporated into the coating in the same way. The concentration of pigment in coating films is expressed by the pigment volume concentration (PVC). This is the ratio of the volume of pigments and extenders to the total volume of the nonvolatile components. Each coating system has a critical pigment volume concentration (CPVC) at which the binder just fills the free space between the close-packed pigment particles. At higher pigment concentrations, the pigment particles in the coating film are no longer fully wetted by the binder, leading to a marked deterioration in coating film properties such as luster, stability, strength, and anticorrosion properties.

1.3.4. *Paint Additives*

Paint additives are auxiliary products that are added to coatings, usually in small amounts, to improve particular technical properties of the paints or coating films. Paint additives are named in accordance with their mode of action.

Leveling agents promote formation of a smooth, uniform surface on drying of the paint. Suitable materials include certain high-boiling solvents such as butyl ethers of ethylene glycol, propylene glycol and diglycols, as well as cyclohexanone and alkylated cyclohexanones, and in some cases aromatic and aliphatic hydrocarbons. Low molecular mass resins (e.g., some polyacrylates and silicones) are also used. Solid leveling agents, such as special low molecular mass resins, are also useful for improving the surface properties of films produced from powder coatings. Flow agents act by reducing the paint viscosity during drying. The effectiveness of a particular flow agent depends on the type of binder and the drying or hardening temperature.

Film-formation promoters, which are closely related to flow agents, reduce the film-forming temperature for film formation from dispersions, leading to a surface that is as pore-free and uniform as possible. Certain high-boiling glycol ethers and glycol ether esters are used, often in combination with hydrocarbons.

Wetting Agents, Dispersants, and Antisettling Agents. Wetting agents form one of the largest groups of coating additives. These are surfactants which aid wetting of the pigments by the binders and prevent flocculation of the pigment particles. This leads to the formation of a uniform, haze-free color and a uniformly high luster of the coating film. This group also includes the dispersants, which give good pigment wetting and hence optimum dispersion of the pigments in the

paint, thereby preventing sedimentation particularly of high-density pigments. As well as good wetting properties, some pseudoplasticity is also necessary. Antisettling agents have similar characteristics to dispersants.

Antifoaming agents are used to prevent foaming during paint manufacture and application and to promote release of air from the coating film during drying. Various products are used, including fatty acid esters, metallic soaps, mineral oils, waxes, silicon oils, and siloxanes, sometimes combined with emulsifiers and hydrophobic silicas.

Catalysts are added to paints to accelerate drying and hardening. They include drying agents (driers, siccatives), which, in the case of the air-drying binders (including some alkyd resins or unsaturated oils), accelerate decomposition of the peroxides and hydroperoxides that form during the drying process, thereby enabling radical polymerization of the binders to take place. The driers used are mainly metallic soaps such as cobalt naphthenate; manganese, calcium, zinc, and barium salts; and zirconium compounds.

Various products are used to catalyze the cross-linking of binder systems at room temperature. For acid-catalyzed systems such as polyester – melamine resin systems, free acids, their ammonium salts, or labile esters are suitable, while for base-catalyzed systems such as polyester – isocyanate, tertiary amines or dibutyltin dilaurate are used. The amount of catalyst used must be such that the pot life is not impaired.

Antifloating and antiflooding agents prevent horizontal and vertical segregation of pigments with different densities and surface properties. This prevents differences in the color and luster of the surface of the film, which can lead to a blotchy appearance.

Antiskinning agents are added to air-drying paints to prevent surface skin formation caused by contact with atmospheric oxygen. In the film, they produce uniform drying and prevent shrinkage (wrinkling). Chemically, these materials are antioxidants such as oximes, which evaporate with the solvents during the drying process.

Matting agents are used to produce coatings with a matt, semi-matt, or silk finish. They include natural mineral products such as talc or diatomites and synthetic materials such as pyrogenic silicas or polyolefin waxes. Matting can also be obtained by special formulations that exploit the incompatibility between binder components and their cross-linked structures.

Neutralizing agents are used in waterborne paints to neutralize binders and stabilize the product. Ammonia and various alkylated aminoalcohols are used, depending on the type of binder and method of application. On hardening, the amines mainly evaporate along with the water.

Thickening agents control the rheological properties of paints of various types. They include inorganic (mainly silicates), organometallic (titanium and zirconium chelates), naturally occurring organic (mainly cellulose ethers) and synthetic organic products (polyacrylates, polyvinylpyrrolidone, polyurethanes).

Preservatives (biocides, fungicides) prevent the attack of paint systems, principally water-based, by microorganisms.

Corrosion inhibitors are used to prevent the formation of corrosion products when waterborne paints are applied to metallic substrates (flash rust). They include oxidizing salts such as chromates, metaborates, nitrites, and nitrates; organic amines or sulfur-containing products; and organic salts (benzoates, naphthenates, octoates).

1.3.5. *Solvents*

Solvents (⟶ Solvents) are compounds that are normally liquid at room temperature. Those most commonly used in coatings technology are aromatic and aliphatic hydrocarbons, esters of acetic acid, glycol ethers, alcohols, and some ketones. Solvents dissolve solid and highly viscous binder components. They enable incompatibility between paint components to be overcome, improve pigment wetting and dispersion, and control storage stability and viscosity of the coating. They promote the release of included air from the liquid coating film, control the drying behavior of the coating, and optimize flow properties and luster. Organic solvents are used in most liquid coatings systems, including waterborne coatings, in which they perform important functions.

After paint application, the solvents should evaporate as quickly as possible, leaving the film. If no special precautions are taken, the solvents enter the atmosphere as pollutants. To protect operating personnel from the toxic effects of evaporating solvents, safety measures such as ventilation and air exhaust are necessary. To protect the environment, incineration and sometimes solvent-recovery plant is installed to prevent solvents entering the atmosphere. Other measures for the protection of the workplace and the environment from solvent vapors include the development and use of new low-solvent or solvent-free coatings, e.g., high-solids paints, waterborne coatings, and powder coatings.

1.4. Paint Application

Paint application can be performed manually, for example with brushes or rollers, or by mechanical methods such as spraying, atomization by rotating disks or cones, dipping, pouring, rotating drums and tumbling equipment, and automated application by rollers. Powder coatings are applied by electrostatic spraying or by dipping components into the powders. Multicomponent coatings are applied with multicomponent spraying equipment.

1.5. Drying and Film Formation

As the paint dries on the substrate, a firmly bonded film is formed. The properties of this film are determined both by the substrate and its pretreatment (cleaning, degreasing) and by the composition of the coating and the application method used.

Drying of the paint on the substrate takes place physically (1 – 3) or chemically (4):

1. Evaporation of the organic solvents from solvent-containing paints
2. Evaporation of water from waterborne paints
3. Cooling of the polymer melts (powder coatings)
4. Reaction of low molecular mass products with other low or medium molecular mass binder components (polymerization or cross-linking) to form macromolecules

Physical Drying Physical drying takes place mainly for paints with high molecular mass polymer binders such as cellulose nitrate, cellulose esters, chlorinated rubber, vinyl resins, polyacrylates, styrene copolymers, thermoplastic polyesters, and polyamide and polyolefin copolymers. These materials give good flexibility and stability because of their high molecular mass. Their glass transition temperature should be above room temperature to ensure adequate hardness and scratch resistance. With these polymers, film formation can also take place from solutions or dispersions in organic solvents or water, from which the solvent or water evaporates, leaving behind the chemically unchanged polymer film.

Film formation can be accelerated by drying at elevated temperatures (forced drying). Physically drying solvent-containing paints have a low solids content because the molecular mass of the binder is relatively high. Higher solids contents are obtained by dispersing the binder in water (dispersions, emulsions) or in organic solvents (nonaqueous dispersion or NAD systems). Films formed from physically drying paints, especially those formed from solutions, are sensitive to solvents (dissolution or swelling). The physically drying coatings also include many powder coatings that contain thermoplastic binders. Film formation takes place by heating the powder that has been applied to the substrate above its melting point. This ensures that a sealed film of polymer is formed.

Plastisols and organosols are a special case of physically drying coatings systems in which the binders consist of finely dispersed poly(vinyl chloride) or thermoplastic poly(meth)acrylates suspended in plasticizers. Organosols also contain some solvent. On drying at elevated temperatures, the polymer particles are swollen by the plasticizer, a process known as gelation.

Chemical Drying. Chemically drying paints contain binder components that react together on drying to form cross-linked macromolecules. These binder components have a relatively low molecular mass, so that their solutions can have a high solids content and a low viscosity. In some cases, solvent-free liquid paints are posible. Chemical drying can occur by polymerization, polyaddition, or polycondensation.

When *polymerization* is used as the hardening principle, reactive components combine to form the binder, e.g., unsaturated polyesters with styrene or acrylate monomers. Here, one component often behaves as a reactive solvent for the other, and low-emission coating systems are the result. Cross-linking can be carried out at room temperature (cold curing) or by radiation curing.

In drying by *polyaddition*, low molecular mass reactive polymers such as alkyd resins, saturated polyesters, or polyacrylates react with polyisocyanates or epoxy

resins to form cross-linked macromolecules. Because this reaction can take place at room temperature, the binder components must be mixed shortly before application. The period of time during which a coating of this type remains usable after mixing of the components is known as the pot life. These are known as two-pack coatings, differing from the one-pack systems, which can be stored for months or even years.

Chemically blocking one of the polyaddition binder components (e.g., the polyisocyanate) gives a coating system stable at room temperature. Heat is required to deblock the component and enable cross-linking to occur. Stoving paints of this type are used in industry and in powder coatings.

Polycondensation drying requires the addition of catalysts or the use of higher temperatures. Acid-catalyzed coatings are well-known cold-curing paint systems used in the furniture industry, while heat-curing and stoving paints are used as industrial and automotive coatings. The binding agents used are functional alkyd resins, saturated poylesters, or polyacrylates in combination with urea resins, melamine resins, or phenolic resins. On cross-linking, water, low molecular mass alcohols, aldehydes, acetals, and other volatile compounds are released.

In practice drying of coatings and paints does not take place by one method alone. With solvent- containing and waterborne heat-curing coatings, physical drying by solvent evaporation always precedes chemical drying. Depending on the composition of the binder system, physical and chemical drying can take place simultaneously, and the various mechanisms of chemical drying can proceed concurrently or consecutively, depending on the nature of the binder. A knowledge of binder composition is important in order to assess the drying of a coating and able to accelerate it by heat, radiation, and addition of catalysts.

1.6. Multicoat Systems

Because dried coating films are not always pore-free, optimal protection of the substrate is not always ensured by one coat. A single coat can seldom fulfill all requirements such as good adhesion, corrosion protection, elasticity, hardness, decorative effect, coloration, and resistance to weathering and chemicals. Coatings with different compositions and functions are therefore often applied in succession. For example, primers provide good adhesion to the substrate and maximum corrosion protection, whereas color stability, gloss, and resistance to weathering are better provided by a top coat which is specially designed for this purpose but may not have particularly good corrosion resistance.

Intermediate coatings between the top coat and the primer are also applied if the highest quality is required, e.g., in the automobile industry. These have the task of providing adhesion between the primer and the top coat, and they also smooth out irregularities on the substrate, thereby indirectly helping to ensure good flow of the top coat and a high gloss with no defects.

1.7. Economic Aspects

In the industrialized countries of Europe and in North America, annual paint consumption of coatings per capita is >20 kg, and high growth rates can be expected in the less industrialized countries of Eastern Europe, Asia, and South East Asia. The volume of coatings produced in Western Europe in 1991 was ca. 5×10^6 t, and, in the United States, 4.7×10^9 L. The annual growth rates of solvent- containing conventional coatings are estimated to be 1 – 2 %, and, of environmentally friendly coatings (high solids, water-based, powder coatings, etc.), ca. 5 %.

1.8. Future Outlook

In the past, the development of coatings was mainly based on technical, quality, and economic considerations. These factors are just as important today from a business point of view and will continue to be so in the future. However, other considerations are now very much in the foreground, i.e., environmental protection, toxicology, environmentally friendly disposal of paint residues and coated articles at the end of their life cycle, the recycling of coated articles, and the conservation of raw materials and energy.

Thus, numerous low-emission paints have been developed, including high-solids paints, waterborne paints, aqueous dispersions for industrial use, powder coatings, and radiation-curing coatings. At the forefront in adopting these environmentally friendly products is heavy industry, in particular the automobile and household appliance industries. Medium-sized and smaller businesses will profit from this experience, adapting it for their own needs.

To conserve raw materials based on mineral oil, renewable raw materials derived from natural oils and resins will be investigated and assessed for potential use in paints.

Nevertheless, the principal development goal is still to secure further improvements in the quality of paint systems and to prolong the durability of coating films. The longer the renewal of a coated surface can be delayed, the less the environment is polluted, and the smaller are the amounts of waste produced and of raw materials and energy consumed.

The continuous increase in automation and electronic control of paint production and application are equally relevant, enabling products to be manufactured that are consistently of the highest quality.

[Top of Page]

2. Types of Paints and Coatings (Binders)

Werner Funke, Lutz Hoppe, Jürgen Hasselkus, Larry G. Curtis, Klaus Hoehne, Hans-Joachim Zech, Peter Heiling, Masaaki Yamabe, Klaus Dören, Hans Schupp, Rolf Küchenmeister, Martin Schmitthenner, Wolfgang Kremer, Wolfhart Wieczorrek, Hans Gempeler, Wolfgang Schneider, James W. White, Anthony G. Short, Werner J. Blank, Leonard J. Calbo, Dieter Plath, Friedrich Wagner, Werner Haller, Karl-Martin Rödder

In this chapter paints and coatings will be discussed according to their binders.

2.1. Oil-Based Coatings [11]

Composition. Oil-based paints (oil paints) are among the oldest organic coating materials; in China, they have been known for more than 2000 years. Oil paints consist of natural drying oils (e.g., linseed oil, China wood or tung oil, and soybean oil) which undergo autoxidative polymerization in the presence of catalytic driers and atmospheric oxygen (⟶ Drying Oils and Related Products). Further constituents may include hard resins (e.g., alkylphenolic resins) that generally react with the drying oils at elevated temperature (230 – 280 °C) to form oleoresinous binders. On account of the air sensitivity of the oils, heating mainly takes place under an inert gas atmosphere.

Auxiliaries may be added to oil paints to improve their wetting and flow properties. The desired handling consistency is generally adjusted with aliphatic hydrocarbon solvents such as mineral spirits and in certain cases with toluene or xylenes.

With clear varnishes 5 – 10 wt % of solvent is sufficient, with paints 10 – 20 wt % is sufficient. There are very few restrictions in the choice of pigment; basic pigments (e.g., zinc oxide) can be used.

Conventional dispersion equipment (e.g., ball, roller, or sand mills) are suitable for producing oil paints.

Oil paints are relatively environmentally friendly as long as harzardous solvents and toxic pigments (e.g., red lead or zinc chromate) are not used. The oils used in such paints have a low viscosity. They are therefore particularly suitable for priming coats on manually derusted steel surfaces since they wet and penetrate the residual layers of rust well, resulting in thorough coverage. Oil paints are easily applied by conventional methods (e.g., brushing, roller coating, spraying and dipping).

During film formation (curing), atmospheric oxygen reacts with the oil to form hydroperoxides which decompose into radicals and then initiate polymerization of the binder. Driers (metallic soaps such as cobalt, lead, and manganese naphthenates or octoates) catalyze formation and decomposition of the hydroperoxides and thereby accelerate film formation. A combination of several driers is normally used to control the curing reaction at the surface and in the interior of the coating.

The thickness of an oil-paint coating is restricted on account of the atmospheric oxygen required for curing. With thick layers (25 – 30 µm on vertical surfaces and 40 – 50 µm on horizontal surfaces), the oxygen penetrates too slowly and the lower region of the paint layer remains soft. Since the shrinkage of the coating differs in various layer regions during oxidative drying, wrinkles may form if the layer is too thick. The drying time is highly temperature dependent and may increase substantially in the absence of light. At room temperature, oil paint films dry in ca. 12 – 24 h depending on the amount of drier added, whereas several weeks are required in the vicinity of the freezing point of water.

During drying the films take up ca. 10 – 20 wt % of oxygen (relative to the pure oil). The smell detected during drying is partly due to decomposition products of the binder that are formed during autoxidative polymerization.

Coatings derived from oil paints are tough but not excessively hard, and exhibit limited weather resistance. They lose their high gloss relatively quickly (ca. two years) and yellow much more than other binders, both in the light and dark as well as at elevated temperature. The coatings are readily hydrolyzed and are therefore unsuitable, at least as a topcoat, in applications involving exposure to strong chemical influences.

On account of these disadvantages and the relatively long drying time, oil paints have almost completely lost their former importance since the 1960s in favor of oxidatively drying alkyd resins. Being “naturally-based paints”, renewed interest has, however, recently been shown in oil paints owing to ecological reasons.

Chemically modified, oxidatively drying oils (e.g., polyurethane oils, Section One-Pack Systems) are being increasingly used as binders for high-solids coatings (e.g., for wood protection). Binders for oil paints include low molecular mass 1,4-*cis*-polybutadienes, known as polyoils and produced by Hüls (⟶ Hydrocarbons – Synthesis of Dienes and Polyenes by Oligomerization). Relatively short drying times (8 – 12 h) are achieved as a result of the polybutadiene component. The polyoils are heated with oxidatively drying oils or modified with maleic anhydride; the maleic anhydride units being converted into imide groups with amines. Such oil paints are suitable as priming coats for corrosion protection of manually derusted steel surfaces because they stabilize residual rust layers.

2.2. Cellulose-Based Coatings

2.2.1. *Nitrocellulose Lacquers*

Nitrocellulose (cellulose nitrate) lacquers are a mixture of binders (nitrocellulose and resins), plasticizers, and (optionally) pigments dissolved in organic solvents. The nonvolatile components are:

1. Nitrocellulose
2. Resin
3. Plasticizer
4. Pigment (extender, dye)

The volatile components are

1. Active solvents
2. Latent solvents
3. Nonsolvents (diluents such as benzene or toluene)

Physical evaporation of the solvents results in formation of the desired solid film on the substrate surface. Films may also be obtained from aqueous, low-solvent, or solvent-free nitrocellulose emulsions or dispersions [12].

2.2.1.1. *Raw Materials*

The compositions of cellulose nitrate lacquers are summarized in Table 1.

Table 1. Formulation of nitrocellulose and lacquers

Lacquer type	Ingredients	Quantitative (weight) ratio
Primer sealer	DBP, blown castor oil, maleate resin	NC:PL:resin = 1:0.3:0.5–1
Film lacquer	Dicyclohexyl phthalate, dammar resin (dewaxed), maleate resin, paraffin	NC:PL:resin = 1:0.5:0.5 + 8% paraffin (*mp* 57–62 °C)
Gel dipping lacquer	DOP, amino resin/alkyd resin (1:1)	NC:PL:resin = 1:0.2:2 alcohol-soluble NC:solvent 2-propanol:toluene = 60:40 to 30:70
Incandescent lamp lacquer	tricresyl phosphate, alkyd resin	NC:PL:resin = 1:0.5:0.6 colored with ceres dyes
Primer	DBP, unrefined castor oil	NC:PL = 1:0.4; ratio of DBP to castor oil (plasticizer) = 1:1
Wood lacquer	DBP/unrefined castor oil (1:1), alkyd resin/ maleate resin (2:1)	NC:PL:resin = 1:0.5:3
Leather lacquer	DBP, blown castor oil	NC:PL = 1:0.9 high-viscosity NC (e.g., E 950, E 840, Wolff Walsrode)
Light-metal lacquer	DOP, vinyl chloride copolymer	NC:PL:resin = 1:0.5:1
Open-pore wood varnish	DBP/blown castor oil (0.5:1), melamine resin, dammar or ketone resin	NC:PL:resin = 1:0.5:1
Furniture lacquer	DBP, polyacrylate resin, resin ester, alkyd resin	NC:PL:resin = 1:0.25–0.5:3–4
Nail polish	plasticizer	NC:PL = 1:1.2
Paper lacquer	DBP/castor oil (1:2)	NC:PL = 1:0.8
Polishing lacquer	DBP/DOP (1:1.7), groundnut oil alkyd resin	NC:PL:resin = 1:0.3:0.7
Polish	DBP, shellac	NC:PL:resin = 1:0.2:1
Sanding primer	DBP/castor oil (2:1)	NC:PL:resin = 1:0.3:1 + ca. 0.1 zinc stearate
Buffing lacquer	DBP/DOP (1:2), groundnut oil alkyd resin	NC:PL:resin = 1:0.3:0.5
Zapon lacquer	plasticizer	NC:PL = 1:0.1–0.3 high-viscosity NC (e.g., E 1160, Wolff Walsrode)

DBP = dibutyl phthalate; DOP = dioctyl phthalate; NC = nitrocellulose; PL = plasticizer.

Nitrocellulose. For the preparation and properties of nitrocellulose [9004-70-0] (cellulose nitrate) see ⟶ Cellulose Esters – Cellulose Nitrate. Nitrocellulose is an outstanding film-forming substance which displays rapid solvent evaporation (short drying time). It is compatible with most coating raw materials.

Nitrocellulose is characterized by its nitrogen content and solubility. The nitrogen contents are:

Ester-soluble nitrocellulose 11.88 – 12.2 wt %
Alcohol-soluble nitrocellulose 10.9 – 11.3 wt %

High-viscosity, medium-viscosity, and low-viscosity formulations of each type are available [13]. Important producers of nitrocellulose used in lacquers include Hercules (USA), ICI (UK), SNPE (France), Wolff Walsrode (FRG), and NQB (Brazil).

Films formed from high-viscosity nitrocellulose have good flexibility combined with a high crack resistance. They are therefore employed where high mechanical stress is to be expected (e.g., in leather coatings, putty, adhesives). Only lacquers with low solids contents can be obtained from high-viscosity nitrocellulose.

Low-viscosity nitrocellulose is used to prepare high-solids lacquers. Since low-viscosity nitrocellulose produces hard to brittle coating films, plasticizers and elastifying resins must be added to the lacquer formulation. They are used in putty, dipping paints, and printing inks.

The medium-viscosity nitrocelluloses have the broadest application range, a major field being furniture lacquers. They are also employed in paper and metal coatings as well as in reaction lacquers (e.g., acid-catalyzed lacquers and polyurethane paints).

Ester-soluble nitrocelluloses are mainly used in the lacquers described above. Alcohol-soluble nitrocellulose (which is also soluble in esters and ketones) is used for odorless lacquers, particularly for printing inks and sealing waxes.

According to international agreement, industrial nitrocelluloses must have a maximum nitrogen content of 12.6 wt % and are stabilized (phlegmatized) for commercial use. Wetted nitrocellulose cotton (with water, ethanol, 2-propanol, or butanol) contains 65 (or 70) wt % nitrocellulose and 35 (or 30) wt % wetting agent.

Nitrocellulose is also available in the form of chips containing ≤ 82 wt % nitrocellulose and ≥ 18 wt % plasticizers (e.g., dibutyl phthalate); pigments may also be incorporated if desired.

Wetted nitrocellulose must not be allowed to dry out because of the risk of explosion.

Plasticizers. Plasticizers that are compatible with nitrocellulose and resins are used in coatings for the following purposes:

1. To improve adhesive strength and gloss
2. To improve mechanical properties such as elongation, pliability, buckling strength, crease resistance, and deep-drawing ability
3. To increase resistance to light, heat, cold, and sudden temperature changes (cold-check test)

Plasticizers may be solvents or nonsolvents for nitrocellulose. The type used depends on the application. Nitrocellulose is for example soluble in dibutyl phthalate, dioctyl phthalate, dicyclohexyl phthalate, tricresyl phosphate, and triphenyl phosphate. Plasticizers in which nitrocellulose is insoluble include crude and blown vegetable oils, stearates, and oleates.

Resins. A large number of synthetic coating resins (e.g., alkyd, ketone, urea, maleate, and acrylic resins) are available for formulating nitrocellulose combination lacquers. Selection criteria include price, color, influence on solvent release, gloss, hardness, sandability, yellowing, and durability of the final coating.

Nitrocellulose (generally in the form of chips) is used in polyurethane coatings to improve drying behavior, to increase bulk, and to obtain good flow.

Solvents. The solvent mixture has a large influence on the quality of the coated film. The solvent that evaporates last should be a solvent for all raw materials in the lacquer formulation. The most important active (true) solvents are acetate esters (e.g., ethyl, butyl, or propyl acetate) and ketones (e.g., acetone, methyl ethyl ketone, methyl isobutyl ketone).

Latent solvents, which become effective only in the presence of active or true solvents, include alcohols (e.g., methanol, ethanol, and propanol). Like the nonsolvents these are used to reduce costs. The lower alcohols (e.g., methanol or ethanol) are, of course, true solvents for alcohol-soluble nitrocellulose.

2.2.1.2. *Application and Uses*

The preparation of nitrocellulose lacquers is simple and involves dissolution and mixing procedures. The viscosity should be compatible with the equipment used.

Nitrocellulose lacquers can be sprayed efficiently with compressed air or by an “airless” technique. Electrostatic spraying is employed to reduce the overspray and for good coverage (e.g., when coating chairs). Flat articles, thin sheets (foils), or paper can be coated inexpensively on casting machines. High-viscosity lacquers are frequently applied by roller coating. Smaller objects are often coated by the dipping method. The pushing-through process is used for coating pencils.

An important use of nitrocellulose lacquers is in printing inks employed in flexographic, gravure, or silk-screen printing.

The most important areas of use of nitrocellulose lacquers are for coating wood, metal (e.g., automotive repair), paper, foil (cellophane, aluminum), leather, and textiles and in nail polish.

Aqueous nitrocellulose lacquers that contain small amounts of solvent are used in the form of emulsions or dispersions to coat leather [14] and decorative foils. Solvent-free dispersions are cured by UV radiation after evaporation of the water and are used to coat furniture, profiled boards, and paper.

2.2.2. ***Organic Cellulose Ester Coatings [15-17]***

Cellulose acetate [9004-35-7], the simplest organic cellulose ester, offers excellent properties in coating films (e.g., flame resistance, high melting point, toughness, and clarity). These esters have limited solubility and compatibility with other resins; this is, however, necessary for widespread use.

Cellulose butyrate contains the bulkier butyryl group; these esters are more compatible and soluble than acetates, but are too soft for most coating applications. Cellulose esterified with blends of alkyl groups can provide many intermediate properties needed in coatings. Selection of the appropriate cellulose acetate butyrate [9004-36-8] (CAB) and cellulose acetate propionate [9004-39-1] (CAP) content must be based on specific application requirements.

A detailed description of organic cellulose esters is given elsewhere, ⟶ Cellulose Esters – Organic Esters. Production starts by mixing the appropriate organic acids and anhydrides, sulfuric acid catalyst, and purified cellulose. Esterification proceeds rapidly until all three anhydroglucose hydroxyls are esterified with acyl groups. Anhydride mixtures produce mixed esters (e.g., CAB and CAP). Fully acylated cellulose is of limited value in the coatings and plastics industries. Some free hydroxyl groups along the cellulose chain are necessary to provide solubility, flexibility, compatibility, and toughness. Since termination of the esterification reaction is not feasible, the fully acylated triester is slowly hydrolyzed to give the desired hydroxyl content.

Following esterification and hydrolysis, the product undergoes additional manufacturing steps that include filtration, precipitation, washing, and drying. The final product is usually a dry, free-flowing powder.

2.2.2.1. *Cellulose Acetate Butyrate*

Tennessee Eastman is presently the world's only manufacturer of CAB and CAP. Table 2 lists the properties of the commercially available CAB and CAP products.

Table 2. Properties of cellulose acetate butyrates (CAB) and cellulose acetate propionates (CAP) (Tennessee Eastman)

Cellulose ester	Viscosity*, Pa · s	Acetyl** content, wt %	Butyryl content, wt %	Propionyl content, wt %	Hydroxyl content, wt %	Melting range, °C	M_r	T_g, °C	Density, g/cm^3
CAB-171–15S	5.7	29.5	17		1.5	230 – 240	65 000	161	1.26
CAB-321–0.1	0.038	17.5	32.5		1.3	165 – 175	22 000	127	1.20
CAB-381–0.1	0.038	13.5	38		1.3	155 – 165	20 000	123	1.20

CAB-381–0.5	0.19	13.5	38		1.3	155 – 165	30 000 130 1.20
CAB-381–2	0.76	13.5	38		1.3	170 – 185	40 000 133 1.20
CAB-381–2BP	0.84	14.5	35.5		1.8	175 – 185	40 000 135 1.20
CAB-381–20	7.6	13.5	37		1.8	195 – 205	70 000 141 1.20
CAB-381–20BP	6.08	15.5	35.5		0.8	185 – 195	65 000 128 1.20
CAB-500–5	1.9	4.0	51		1.0	165 – 175	57 000 96 1.18
CAB-531–1	0.722	3.0	50		1.7	135 – 150	40 000 115 1.17
CAB-551–0.01	0.0038	2.0	52		1.5	127 – 142	16 000 85 1.16
CAB-551–0.2	0.076	2.0	51		1.8	130 – 140	30 000 101 1.16
CAB-553–0.4	0.114	2.0	46		4.8	150 – 160	20 000 136 1.20
CAP-482–0.5	0.152	2.5		45	2.6	188 – 210	25 000 142 1.22
CAP-482–20	7.6	2.5		46	1.8	188 – 210	75 000 147 1.22
CAP-504–0.2	0.076	0.6		42.5	5.0	188 – 210	15 000 159 1.26

* ASTM D 817 (Formula A) and D 1343.
** ASTM D 817.

Properties. The large size and low polarity of the butyryl groups separate the cellulose chains and lowers the attraction between them. As butyryl content increases, properties are affected as follows:

1. Solubility increases
2. Tolerance for diluents increases (tolerance signifies the ability to withstand inclusion of a nonsolvent in the solvent system before haze or precipitation occurs)
3. Water tolerance decreases
4. Compatibility increases
5. Flexibility increases
6. Moisture resistance increases
7. Grease resistance decreases
8. Tensile strength decreases
9. Hardness decreases
10. Melting range decreases

The *hydroxyl content* of CAB is perhaps the most important chemical variable on the cellulose chain. It affects properties as follows:

1. *Solubility*. At hydroxyl levels of <1 %, solubility is limited. Solubility increases as the hydroxyl content increases. At hydroxyl levels of ca. 5 %, CAB is soluble in lower molecular mass alcohols.
2. *Moisture Resistance*. The greater the hydroxyl content, the more hydrophilic are films formed from it.
3. *Toughness.* Fully substituted esters are not as tough and flexible as those with a low hydroxyl content.
4. *Reactivity*. The degree of reactivity increases as hydroxyl content increases. When cross-linked with other resins, the cross-link density of resultant films increases correspondingly.

The higher viscosity form of a particular ester type is associated with a higher molecular mass and longer chain length. Increasing molecular mass of a particular ester (and coatings formulated with it) slightly lowers its solubility and compatibility but does not affect hardness or density. Melting ranges and toughness increase with molecular mass.

Uses. Protective and decorative *coatings for metals* can be formulated as converting or curing systems or as air-drying lacquer systems. Cellulose acetate butyrate is included in many such coatings as a modifying resin to impart specific properties. It can also be used as the primary film-forming resin. Cellulose acetate butyrate is usually included in coatings for metal to accelerate solvent release from the film. This significantly reduces the dry-to-touch time and consequently reduces dirt pickup. The cellulose resin may be added at levels of 1 – 5 wt % to improve film leveling and reduce cratering.

Clear-on-base (basecoat – clearcoat) automotive coatings are stoving enamels that are used worldwide. Proper application of the basecoat is critical for obtaining the desired appearance of the coating. Normally, aluminum flake is used for pigmentation and must be oriented parallel to the substrate. This can be achieved by inducing maximum shrinkage of the basecoat during solvent evaporation. Cellulose acetate butyrate (20 – 30 wt % of total resin solids) greatly assists film shrinkage and metallic flake orientation by increasing coating viscosity following atomization during painting operations (Fig. 1). The viscosity increase permits the application of a relatively thick, wet coating without sagging and running. Subsequent solvent evaporation results in film shrinkage causing the aluminum to assume a position relatively parallel to the substrate. The CAB prevents redissolution of the basecoat when the clear topcoat is subsequently applied.

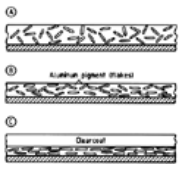

Figure 1. The function of CAB in automotive clear-on-base finishes

A) Viscosity of CAB permits application of heavy basecoat; B) Solvent evaporates, film shrinks, and flake orientation begins; C) The CAB prevents solvents in the clear topcoat from redissolving the basecoat

Cellulose acetate butyrate is used in a wide variety of *coatings for wood* because they provide many desirable properties (fast solvent release, flowout, and leveling; excellent spray characteristics, nonyellowing, and cold crack resistance).

The surface of molded plastic parts is often coated to obtain properties that the plastic does not have (e.g., mar resistance, solvent resistance, reduced dirt pickup in barrier coatings). (Mar resistance is the ability to withstand scratching and scuffing caused by sliding a rough object or cracking upon impact of a hard object.) Cellulose acetate butyrate is used in *coatings for plastic* because of its toughness, low color, color stability, good abrasion resistance, and generally good bonding characteristics.

The development of urethane elastomers has allowed the formulation of tough, abrasion-resistant coatings for many flexible substrates including textiles. One widely used application of *CAB-modified urethane elastomeric systems* is for coating lightweight outdoor backpacking and camping equipment (e.g., portable tents).

Cellulose acetate butyrate is beneficial in *radiation-curing systems*. CAB 551–0.01 with its high butyryl content and low viscosity is soluble in many vinyl

monomers used in this area. Levels as low as 1 – 5 wt % provide good flowout and leveling of the coating which often tends to form craters and pull back at edges.

Another very important application is the *dispersion of pigments* that are difficult to disperse (e.g., carbon black, transparent iron oxides, phthalocyananine blue and green, and perylene red). The use of CAB and two-roll milling is the most efficient method of dispersion.

Application. Cellulose acetate butyrate lacquers are usually applied by spraying (air atomization, airless, spinning disk). Application by brush or dip is possible but less commonly used.

2.2.2.2. *Cellulose Acetate Propionate*

Cellulose acetate propionates (CAP) have the same characteristics as CAB, including high solubility and compatibility with other resins. They also have a very low odor; this is important in printing applications and in reprographic processes. Commercially available products and their typical properties are listed in Table 2.

Cellulose acetate propionate is used mainly in printing inks where a low odor is required (e.g., in food packaging). It is also used for coating leather clothing and for printing gift wrapping paper.

2.3. Chlorinated Rubber Coatings

2.3.1. *Starting Products [18-20]*

Chlorinated rubber was originally produced by chlorination of natural rubber that had previously been degraded to low molecular mass compounds by mastication or addition of radical formers. Chlorine contents are typically 64 – 68 wt %. Modern polymer technology now provides other synthetic substances that make production of chlorinated rubber independent of natural rubber. For example, synthetic polyisoprene, polyethylene, polypropylene, and other polyolefins are now used exclusively or in part as starting materials. For a detailed description of chlorinated rubber, see ⟶ Rubber, 4. Chemicals and Additives.

Chlorination products containing >64 wt % chlorine and with mean molecular masses of 50 000 – 350 000 exhibit similar properties (DIN 53 243). Chlorinated polyisoprene is assumed to have the following structure [20]:

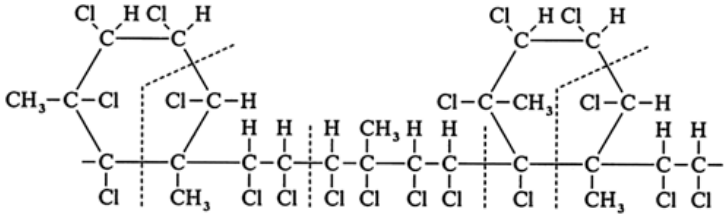

Commercial Products. Chlorinated rubber from natural rubber or synthetic polyisoprene is only produced by a few manufacturers. Trade names include Adeka (Asahi Denka), Alloprene (ICI), Amolac (VEB Chemische Werke Buna), Chlortex (Caffaro), Pergut (Bayer) [18], Superchlor (Sanyo-Kokusaku Pulp). These products are available in various viscosity grades, whose ranges largely coincide for the aforementioned commercial products (Table 3).

Table 3. Viscosity grades of Pergut, an example of a commercial chlorinated rubber product

Designation	Viscosity*, mPa · s	Mean molecular mass** (M_w), g/mol
Pergut S 5	4.5 – 7.5	60 000
Pergut S 10	10.0 – 12.0	124 000
Pergut S 20	16.0 – 24.0	160 000
Pergut S 40	33.0 – 51.0	213 000
Pergut S 90	75.0 – 110.0	302 000
Pergut S 130	120.0 – 150.0	327 000
Pergut S 170	150.0 – 190.0	359 000

* Measured in a 20 % solution in toluene at 23 °C in a Höppler viscometer (DIN 53 015).
** Measured by a combination of gel permeation chromatography and viscometry.

Depending on production technology, chlorinated rubbers supplied by the manufacturers may contain 5 – 8 % carbon tetrachloride. Some manufacturers have reduced the carbon tetrachloride content to 1 %, which means that paints (formulations) can be produced that do not have to be labeled according to the GefStoffV (December 16, 1987, Appendix I.2.1., solvents) with the warning “contains carbon tetrachloride”. Carbon tetrachloride is not only important for occupational health and safety reasons. It is one of the chlorine-containing compounds that are responsible for the formation of the “ozone hole”. A worldwide ban on these substances is now imminent. Since 1991 supply forms that contain <0.1 % carbon tetrachloride (Bayer, ICI) were commercially available. It is expected that an exemption order will be granted for the use of these products.

Properties. The high degree of chlorination substantially alters the properties of the starting polymers. A hard, granular, white powder with the following properties is obtained: high resistance to oxidizing agents (e.g., ozone or peroxide), water, inorganic salts, acids, alkalis, and gases; good solubility in almost all conventional solvents except water, aliphatic hydrocarbons, and alcohols; good compatibility with a wide range of paint resins and plasticizers; low flammability; fungistatic and bacteriostatic behavior; pigmentability with almost all inorganic pigments and extenders, as well as many organic pigments.

Disadvantages resulting from the high chlorine content include low temperature resistance (60 °C wet, 90 °C dry) on account of elimination of hydrochloric acid. Chlorinated rubber also tends to undergo yellowing and chalking more readily when exposed to atmospheric influences than do lightfast binders such as aliphatic polyurethanes or polyacrylates. (Chalking refers to the degradation of the binder on the surface of the coating. The pigment is set free, forming a white deposit that can be wiped off.)

Other Areas of Application. In addition to its principal use as a paint binder, chlorinated rubber is added to rotary gravure inks; it facilitates transport of the pigments from the inking device to the print carrier and good setting. This results in good printability, fast drying, and a shiny, smudgeproof print. These properties are important for printing illustrated magazines, prospectuses, and catalogs with large circulation numbers.

In polychloroprene-based adhesives, addition of ca. 10 % chlorinated rubber (based on the polychloroprene content) increases the initial adhesive strength (contact adhesives) and improves heat stability.

2.3.2. *Chlorinated Rubber Paints*

Chlorinated rubber and related chlorinated polymers form coating films by physical drying. Plasticizers or resins have to be added since otherwise brittle films are formed.

Composition. The *binder* consists of ca. 65 % chlorinated rubber (usually low-viscosity grades) and ca. 35 % plasticizer. Chlorinated paraffins (Hordaflex, Hoechst; Chlorparaffine and Witaclor, Hüls; Cereclor, ICI) are used as *plasticizers*. Special nonhydrolyzable plasticizers may be added if necessary, e.g., bisphenoxyethylformal (Desavin, Bayer) or resin-modified phenyl alkylsulfonates (Leromoll, Bayer). This composition ensures the “nonhydrolyzability” of the binder (resistance to water, acid, and alkali), which is not the case if hydrolyzable phthalate or adipate plasticizers are used. Nonhydrolyzable resins (e.g., coumarone – indene resins or other hydrocarbon resins) are often added as “extenders”.

Red lead has proved outstandingly suitable as a *pigment* for priming coats on steel, and is fully effective in chlorinated rubber coatings. For reasons of environmental protection and occupational health, the use of toxic lead compounds is diminishing. Zinc phosphate is used instead, although it does not have the same corrosion protection effect. Conventional metal pigments (e.g., lead dust, aluminum bronze, and zinc dust) produce diffusionproof coatings with good mechanical properties. In the case of aluminum bronze and zinc dust, stabilization of the paint is required to prevent gelatinization. Iron oxide, chromium oxide, and titanium dioxide pigments, commonly used in the paint industry, are suitable for finishing and topcoats. Zinc oxide, white lead, and lithopone are, however, unsuitable.

All inert minerals are suitable as *extenders*. Carbonate-containing extenders may only be used if no stringent requirements have to be satisfied as regards resistance to water and chemicals.

The choice of *solvent* is practically unlimited. Xylene or other alkylbenzenes are generally recommended. Mixtures of esters and mineral spirit can be used to avoid compulsory warning labels.

Hydrogenated and modified castor oil is used as an *additive* to adjust the viscosity and facilitate application with a brush or spray gun (compressed air or airless); layer thicknesses of ≥ 100 µm are thereby achieved [21-23].

Production. Chlorinated rubber paints are produced by conventional means. The plasticizer, resins, and in some cases a proportion of the chlorinated rubber are first dissolved in the solvent. The high-boiling solvent contained in the formulation is preferred for this step. The hydrogenated castor oil is then added and the resultant mixture is dispersed in a dissolver. In order to obtain optimum “digestion”, the instructions of the castor oil supplier should be strictly observed; the temperature should not be allowed to exceed ca. 60 °C. Dispersion is followed by the formation of a paste with the pigments and extenders, and grinding. Conventional apparatus including dissolvers is suitable as grinding equipment; grinding with steel balls should be avoided since the iron dust that is formed can cause the final paint to gelatinize after prolonged storage. The ground material is then combined with the separately prepared chlorinated rubber solution.

Application. Chlorinated rubber paints can be applied with all conventional coating equipment. The suppliers' (manufacturers') instructions must, however, be observed since the coating material (chlorinated rubber paint) is specifically formulated for the recommended application equipment.

Uses. On account of their high water resistance, chlorinated rubber paints are used for underwater coatings on steel and concrete (e.g., water storage vessels, swimming baths, sewage systems, harbor installations, and docks). The chemical resistance is exploited in vessels, tanks, and constructional parts used in mines, chemical plants, etc., in which aqueous solutions of inorganic chemicals are handled. Coatings for concrete require chlorinated rubber as a binder due to the alkalinity of the concrete surface.

The main area of use of chlorinated rubber paints is for underwater coatings on ships (see also Section Marine Coatings). Favorable properties for this application are high water resistance, rapid drying (which is independent of the external temperature in the shipyard), good mutual adhesion of the individual layers, and the fact that old coats of paint can easily be renewed.

2.3.3. *Chlorinated Rubber Combination Paints*

Composition. Chlorinated rubber combination paints contain a second resin as the property-determining binder. The chlorinated rubber is added to an alkyd resin, acrylic resin, or bituminous substances to improve properties such as drying rate, water resistance, or chemical resistance. This application only accounts for a small proportion of the total chlorinated rubber consumption.

The proportion of chlorinated rubber in the binder varies from 10 to 50 wt % depending on the intended application; plasticizers and/or alkyd resins and/or acrylic resins account for the remainder.

In combinations with bituminous substances the proportion of chlorinated rubber ranges from 1:10 to 10:1. The ratio depends on whether the goal is to improve the bitumen-based coating without any substantial increase in cost, or to reduce the cost of the chlorinated rubber coating. Adhesion is improved but with the disadvantage of darker shades caused by the black bitumen.

Production corresponds to that of pure chlorinated rubber paints (see Section Chlorinated Rubber Paints).

Chlorinated Rubber – Alkyd Resin Combinations. In these combinations chlorinated rubber accounts for 25 – 50 % of the binder. Chlorinated rubber is used to increase the drying rate and/or improve the chemical resistance. These paints also exploit the benefits of the alkyd resin, e.g., good brushability and nonsolubilization. They are used for corrosion protection in industrial plants or marine environments to protect steel, galvanized steel, and aluminum; air-drying or forced-dried industrial paints (e.g., for agricultural machinery); and road marking paints.

Chlorinated Rubber – Acrylic Resin Combinations. Physically drying acrylic resins are used for these combinations. These combinations have the same drying rates as normal chlorinated rubber paints (see Section Chlorinated Rubber Paints). They have improved flow properties (particularly when applied by pouring techniques), improved weather resistance (chalking and yellowing), and favorable mechanical properties (adhesion and extensibility). Applications include topcoats for ship superstructures and priming coats on galvanized surfaces.

Combinations with Bituminous Substances. Chlorinated rubber can be combined with bitumen and tars but compatibility has to be checked. Addition of chlorinated rubber reduces thermoplasticity, accelerates drying, and prevents cracking of the final coating in adverse weather conditions, without, however, adversely affecting the good adhesion, water resistance, and chemical resistance of the bituminous substance.

Bituminous coatings reinforced with chlorinated rubber are used in silos, tanning pits, drinking water containers, and on ships' hulls (on the underwater part). Bituminous substances for coatings are supplied as special products that are free from carcinogenic constituents.

2.4. Vinyl Coatings

This section deals with paints based on vinyl resins (including vinyl copolymers) which are synthesized by polymerization of monomers containing terminal CH_2=CH groups. Polyolefins, poly(vinyl halides) and vinyl halide copolymers, poly(vinyl esters), poly(vinyl alcohol), poly(vinyl acetals), poly(vinyl ethers), and polystyrene are discussed. Polyacrylates (acrylic resins) are treated in Section Acrylic Coatings.

2.4.1. *General Properties*

Paints and coating materials based on vinyl resins are generally physically drying. Only in a few cases can vinyl resins be chemically cross-linked with other reactants via incorporated reactive groups. The properties of the paints are therefore primarily determined by the chemical and physical nature of the vinyl resin. Despite the large number of available vinyl resins this class of binders has some common features.

All vinyl resins have a linear carbon chain with lateral substituents and exhibit a range of molecular masses. Increasing molecular mass is accompanied by improved mechanical properties, a decrease in solubility, and an increase in the viscosity of their solutions. Vinyl resins of high molecular mass can therefore only be used in the form of dispersions or powders for paint applications. Solvent-containing paints require vinyl resins of considerably lower molecular mass than plastics, since only then can a sufficient binder content be achieved in the viscosity range required for paint application.

The properties of vinyl resins, paints, and coatings are chiefly determined by the nature and number of substituents. The substituents influence the crystallization behavior and thus the properties of interest in paint technology such as the softening range, mechanical properties (film flexibility, cold embrittlement tendency, film hardness), the film-forming temperature in dispersions, solubility, and compatibility with other binders. Chemical behavior also depends on the substituents: ester groups can be hydrolyzed, free carboxyl groups improve adhesion to metals, and hydroxyl groups permit cross-linking with reactants such as isocyanates. Pigment wetting, pigment loading, water swelling capacity, diffusion of water vapor and other gases, solvent retention, and many other phenomena are also largely determined by the substituents.

Copolymerization and the associated introduction of further substituents allows individual properties to be modified. For example, internal plasticization can be achieved by copolymerizing “rigid” monomers with “soft” monomers.

Vinyl paints are produced by conventional techniques (see Chap. Production Technology) and can be applied by all conventional methods (e.g., spraying, brushing, roller coating, and dipping). They are dried at room temperature; heating can be used to shorten the drying process.

2.4.2. *Coatings Based on Polyolefins and Polyolefin Derivatives*

Polyolefins are described in detail elsewhere, ⟶ Polyolefins.

Polyethylene can only dissolve in hydrocarbons above its melting point. On account of its low solubility it is used in coatings solely in the form of powders and dispersions.

On account of their paraffinic nature, polyethylene coatings are highly resistant to chemicals. They are, however, attacked by strong oxidizing agents. Polyethylene coatings are elastically tough, flexible (even at low temperature), resistant to hot water, and nontoxic.

Polyethylenes of lower molecular mass are added as slip and matting agents to paints and printing inks and can also produce dirt-repellent and abrasion-resistant effects. The wax dispersions are produced by hot dissolution and precipitation in aromatic hydrocarbons. This is not necessary with commercial microcrystalline grades. Aqueous dispersions of polyethylene are important as polishing agents.

Polyisobutenes do not crystallize. Polyisobutenes of low molecular mass are flexible resins, and those of high molecular mass are elastomers. Their use in the paint sector is restricted on account of their aliphatic nature and limited compatibility with other binders. They are employed in combination with hydrocarbons such as paraffins, rubber, and bitumen. They are used as plasticizing components.

Ethylene Copolymers. *Ethylene – vinyl acetate copolymers* differ according to their vinyl acetate content and molecular mass. With increasing vinyl acetate content, compatibility with paraffin waxes decreases, but that with other binders increases. Low molecular mass types containing 25 – 40 % vinyl acetate are readily or sufficiently soluble in solvents. With a 40 % vinyl acetate content, they can be combined with polar resins and nitrocellulose. Terpolymers with free carboxyl groups exhibit improved adhesion. Ethylene – vinyl acetate copolymers are primarily added to waxes to improve their properties, but are also used to increase flexibility and adhesion in paints, printing inks and adhesives, and for hot melt coatings. Ethylene – vinyl acetate copolymers have low water vapor and gas permeabilities (barrier effect).

Powder coatings are also formed from ethylene – vinyl acetate copolymers (see Section Coating Powders). *Copolymers of ethylene with maleic acid* (*anhydride*) of low molecular mass are water-soluble, form salts, and undergo cross-linking reactions.

Chlorinated Polyethylene and Polypropylene. Totally chlorinated polyethylene and polypropylene have a chlorine content of 64 – 68 %. Their properties largely correspond to those of chlorinated rubber (see Section Chlorinated Rubber Coatings). Chlorinated polypropylenes can be used for chemical-resistant and weather-resistant coatings. These binders are important in adhesion priming coats and heat-sealing lacquers for polypropylene sheets.

Commercial products include chlorinated polyolefins CP (Eastman) and Hardlen (Toyo Kasei Kogyo).

Chlorosulfonated polyethylene is obtained from polyethylene by simultaneous treatment with chlorine and chlorosulfonic acid. The binder is cross-linked via its sulfonyl group with metal oxides, preferably lead oxide, in the presence of organic acids and sulfur-containing organic accelerators. The paints may be formulated as air- and oven-drying, one- and two-pack systems with a pot life of one to two weeks. Cross-linked films of chlorosulfonated polyethylene have an extremely high chemical resistance, particularly against oxidizing agents (it is used in internal coatings of chromic acid baths).

2.4.3. *Poly(Vinyl Halides) and Vinyl Halide Copolymers*

2.4.3.1. *Poly(Vinyl Chloride) and Vinyl Chloride Copolymers*

For a detailed description of poly(vinyl chloride), see ⟶ Poly(Vinyl Chloride)).

Poly(vinyl chloride) [9002-86-2] (PVC) is sparingly soluble in the solvents used in the paint industry, and so is rarely used as a paint binder. It is used to a significant extent in the form of paste-forming PVC powders in plastisols and organosols. Plastisols are PVC dispersions in plasticizers that also contain stabilizers, extenders, pigments, and processing agents. Organosols additionally contain solvent-soluble binders. Since plastisols and organosols adhere poorly to metals, adhesion promoters are necessary; organosols can also be combined with adhesive resins when applied to metals.

Film formation takes place at gelation temperatures from 160 to 200 °C. The properties of the coating depend on the type and amount of PVC, plasticizer, and extender. A foaming effect can be achieved during gelation by adding blowing agents. Poly(vinyl chloride) is also used in powder coatings, which are applied by fluidizedbed coating and electrostatic spraying (corrosion protection for metal furniture, wire, aluminum front elements for buildings, and road-marking posts). PVC coatings have favorable mechanical properties, a high abrasion resistance, and high chemical resistance.

Commercial products include Ekavyl, Lucovyl (Atochem); Geon (Goodrich); Hostalit (Hoechst); Solvic (Solvay); Vestolit (Hüls); and Vinnol (Wacker).

Chlorinated Poly(Vinyl Chloride). Post-chlorinated PVC combines the advantageous properties of PVC e.g., good chemical and weather resistance with good solubility in most conventional solvents. Its importance has, however, continually decreased in the paints sector.

Vinyl Chloride Copolymers. Vinyl chloride copolymers can be used in a wide variety of paint technology applications. The solubility of vinyl chloride copolymers is considerably higher than that of the PVC homopolymer. Important examples are copolymers without additional functional groups formed with vinyl acetate, dibutyl maleate, or isobutyl vinyl ether; terpolymers with carboxyl groups formed with dibutyl maleate or vinyl acetate and a dicarboxylic acid; and copolymers and terpolymers with hydroxyl groups formed with hydroxyacrylates or with vinyl acetate and vinyl alcohol.

Commercially available PVC copolymers differ in composition and molecular mass. They are physically drying binders that undergo film formation by solvent evaporation. (The types with hydroxyl groups can, however, also be used as combination binders in reactive systems. They can be cross-linked, for example, with melamine resins or polyisocyanate resins.) The films are tough, abrasion resistant, thermoplastic, of low flammability, colorless, odorless, tasteless, and physiologically harmless. Flexibility and abrasion resistance improve with increasing molecular mass. Vinyl chloride copolymers exhibit a good water resistance and outstanding resistance to alkalis, dilute mineral acids, salt solutions, oils, fats, greases, alcohols, and gasoline. Paints can readily be formulated with this group of binders and can be applied by conventional methods.

Preferred solvents include ketones, esters, and chlorinated hydrocarbons. Aromatic hydrocarbons have a swelling effect on most vinyl chloride copolymers, but are widely used as diluents. Alcohols and aliphatic hydrocarbons do not dissolve vinyl chloride copolymers.

Both monomeric and polymeric plasticizers are suitable for plasticization. Practically all monomeric plasticizers for PVC can be used. Suitable polymeric plasticizers include polyadipates, chlorinated paraffins, carbamide resins, and epoxides. Vinyl chloride copolymers are compatible with most conventional pigments and extenders. Despite their high intrinsic stability, paints based on vinyl chloride copolymers have to be stabilized against dehydrochlorination in the presence of heat and/or UV radiation for some applications. Epoxy compounds are often sufficient for thermal stabilization.

The composition of vinyl chloride copolymers without functional groups influences their solubility behavior and compatibility with other paint binders. For example, copolymers with isobutyl vinyl ether or maleate esters dissolve in aromatic hydrocarbons, whereas copolymers with vinyl acetate merely swell in these solvents. Paint films formed from vinyl chloride copolymers without functional groups are heat-sealable on account of their thermoplastic character. Since the films adhere poorly to nonabsorbing substrates such as metals, they are suitable as binders for strippable coatings. On account of their good chemical resistance, vinyl chloride copolymers are also extremely suitable as binders for exterior-use paints, traffic paints, and paper and foil lacquers; their lack of taste and odor means that they can be used as pasteurization-resistant coatings for can interiors.

Vinyl chloride terpolymers containing carboxyl groups adhere extremely well to metals. Due to their special properties (outstanding adhesion to aluminum, good chemical resistance, heat-sealable from ca. 140 °C—the sealing temperatures can be lowered by adding plasticizers) these copolymers are ideal binders for heat-sealable finishes of aluminum foils used in the packaging sector.

Vinyl chloride copolymers containing hydroxyl groups can be used in combination with other binders in reactive systems. The non-cross-linked paint films are thermoplastic and therefore heat-sealable. Cross-linking (e.g., with melamine resins or polyisocyanate resins) lowers the thermoplasticity and improves adhesion and resistance. The hydroxyl groups are also responsible for very good adhesion to organic substrates; these copolymers are therefore used as intermediate layers in marine coatings. In two-component polyurethane lacquers, hydroxyl-containing vinyl chloride copolymers can be used alone or in combination with polyols. In the latter case, lacquer viscosity can be adjusted and initial physical drying can be accelerated. As a result of their outstanding pigment wetting and stabilization properties, their high pigment loading, and compatibility with polyester and polyurethane resins, hydroxyl-containing PVC copolymers are used as binders for magnetic storage media.

Commercial products include Corvic (ICI); Hostaflex (Hoechst); Laroflex, Lutofan (BASF); S-Lec (Sekisui); Ucar, Vinylite (Union Carbide); Vilit (Hüls); and Vinnol (Wacker).

2.4.3.2. *Vinylidene Chloride Copolymers*

On account of its low thermal stability, poly(vinylidene chloride) is seldom used in paints. Vinylidene chloride copolymers with vinyl chloride, acrylonitrile, or acrylates are mainly employed. These heat-sealable copolymers are efficient gas barriers and have an outstanding resistance to chemicals. They are marketed as solid resins and dispersions. Vinylidene chloride copolymers are mainly used for coating food-packaging foils. They are also important in paint coatings where good chemical resistance is required.

Commercial products include Diofan (BASF); Haloflex (ICI); Ixan (Solvay); and Saran (Dow Chemical).

2.4.3.3. *Fluoropolymer Coatings*

Organic fluoropolymers have been used in many fields because they have special properties that no other polymers can provide (⟶ Fluoropolymers, Organic).

Coating is one of the important uses of fluoropolymers, since it enables them to exhibit their characteristics on the surface of a substrate. Some of the conventional fluoropolymers such as polytetrafluoroethylene [9002-84-0] (PTFE), tetrafluoroethylene – hexafluoropropylene copolymer [25067-11-2] (FEP), and ethylene – tetrafluoroethylene copolymer [25038-71-5] (ETFE) have been used as antistick or anticorrosive coatings. Only poly(vinylidene fluoride) [24937-79-9] (PVDF) has so far been used in paints. The major difficulties in employing thermoplastic fluoropolymers in paints and coatings result from their poor solubility in organic solvents and also from the necessity to bake them at a rather high temperature (> 200 °C). However, novel fluoropolymers with curable characteristics have also been developed (one of them was successfully commercialized), mainly as highly weather-resistant paints [24], [25].

Coatings with Thermoplastic Fluoropolymers. *Poly(vinylidene fluoride)*, PVDF, is the only conventional thermoplastic fluoropolymer that is used as a commercial product for weather-resistant paints. This crystalline polymer is composed of $-CH_2CF_2-$ repeating units; it is soluble in highly polar solvents such as dimethylformamide or dimethylacetamide (⟶ Fluoropolymers, Organic – Poly(Vinylidene Fluoride)). Poly(vinylidene fluoride) is usually blended with 20 – 30 wt % of an acrylic resin such as poly(methyl methacrylate) to improve melt flow behavior at the baking temperature and substrate adhesion. The blended polymer is dispersed in a latent solvent (e.g., isophorone, propylene carbonate, dimethyl phthalate). The dispersion is applied to a substrate and baked at ca. 300 °C for ca. 40 – 70 s. The weather resistance of the paints exceeds 20 years [26-28].

Copolymers of vinylidene fluoride with tetrafluoroethylene or hexafluoropropylene have been developed as an air-drying paint mainly for repair coating of the PVDF finish [29-31]. Poly(vinylidene fluoride) is now widely used for coil coating of galvanized iron and aluminum sheets, and as a maintenance-free coating for walls of skyscraper buildings and roofing of industrial constructions.

Polytetrafluoroethylene. See also ⟶ Fluoropolymers, Organic – Polytetrafluoroethylene. An aqueous dispersion of PTFE is produced by emulsion polymerization of tetrafluoroethylene followed by thermal concentration of the latex up to ca. 60 wt %. Since the polymer has a very high melt viscosity and does not adhere to many substrates, the properties of the finish are greatly influenced by pretreatment of the substrate. On physically or chemically pretreated iron or aluminum, the PTFE coating is rather vulnerable to scratching (soft coat). If, however, the treated substrate is first coated with a ceramic-powder primer to form a coarse, hard surface, the subsequently applied PTFE finish becomes tough and scratch-resistant (hard coat). These coatings are used in hot cookware (e.g., frying pans) with the advantage of being nonstick and easily lubricated [32, 33]. Recently, mixtures of PTFE dispersions and heat-resistant hydrocarbon polymers (e.g., polyimide, polyether sulfone, or polyphenylene sulfide) have been developed to improve the poor adhesion of fluoropolymer to a substrate and applied as a primer or one-coat enamel [34].

Tetrafluoroethylene Copolymers. Tetrafluoroethylene – hexafluoropropylene copolymer (FEP) and tetrafluoroethylene – perfluoroalkoxyethylene (PFA) are used as dispersion coatings in the same way as PTFE, taking advantage of their low melt viscosity and low viscosity at baking temperature. (See also ⟶ Fluoropolymers, Organic – Tetrafluoroethylene–Hexafluoropropylene Copolymers.)

Powder coating is another popular technique in fluoropolymer coating. Tetrafluoroethylene – ethylene copolymer (ETFE) has a highly alternating sequence with a low melting point (280 °C) and low melt viscosity (⟶ Fluoropolymers, Organic – Tetrafluoroethylene–Ethylene Copolymers). Tetrafluoroethylene – ethylene copolymer powder has a better melt processability than PTFE; electrostatic coating gives thick finishes without pinholes that have excellent anticorrosive and antistick characteristics. Tetrafluoroethylene – perfluoroalkoxyethylene is also applied as a powder coating, exploiting its low melt viscosity in the same way as ETFE.

Coatings with Curable Fluoropolymers. In order to facilitate the application of fluoropolymers, extensive studies have been performed on curable fluoropolymers. A block copolymer containing 65 % vinylidene fluoride, 25 % tetrafluoroethylene, and 10 % vinyl ester forms a highly weatherable, strongly adhering, solventborne coating on metals or cellulosic materials after photoinitiated cross-linking (UV curing) [35, 36].

Hydroxyl-containing fluoropolymers made of fluoroolefin and hydroxyalkyl vinyl ether, hexamethoxymethylmelamine, and silica form highly cross-linkable liquid mixtures and can be applied as an excellent scratch-resistant coating on plastic objects [37, 38].

Although these fluoropolymers are soluble in organic solvents and can be applied at a rather low temperature (< 200 °C), characteristic properties of the coating seem to be too specialized to be widely used in the coating area.

A fluoroolefin – vinyl ether terpolymer (Lumiflon) has recently been developed [24]. This polymer is an amorphous, alternating copolymer of a fluoroolefin with several vinyl monomers (Fig. 2). The alternating sequence is responsible for the high performance of the resultant finish. The combination of vinyl monomers provides the polymer with various properties necessary for a coating material.

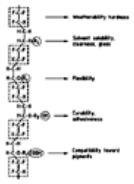

Figure 2. Structure of Lumiflon polymer

The outstanding characteristics of this polymer as a coating material are its excellent weatherability and its ease of handling and processing [33], [39-41]. The hydroxyalkyl groups in the polymer react with polyisocyanates at room temperature and with melamine resin or blocked isocyanates at higher temperature. Lumiflon can therefore be formulated for both on-site coatings that are cured at ambient temperature and for thermoset coatings in the factory. Practical application to plastics, buildings, various architectural structures, bridges, and automotives is now proceeding and the market is expanding annually.

Several fluoroolefin–vinyl monomer terpolymers have been developed based on the Lumiflon polymer [42-45]:

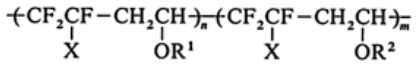
[2.14]
[2.29]

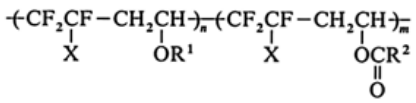
[2.32]

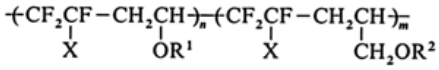
[2.33]

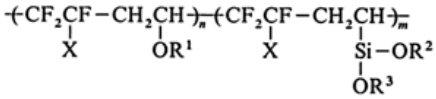
[2.34]

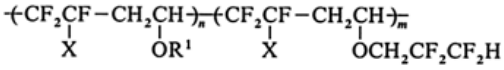
[2.35]

Recently waterborne coatings based on fluoroolefin – vinyl ether terpolymers have been reported, in which a macromonomer with a hydrophilic side chain was copolymerized with fluoroolefin [46].

A unique fluoroepoxy compound is liquid at ambient temperature and can be cured by incorporation of suitable agents (e.g., amino silicone compounds) [47]. It forms finishes with low friction, oil and water repellency, and antifouling properties [48, 49].

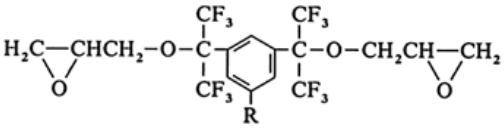

2.4.4. ***Poly(Vinyl Esters)***

Poly(vinyl esters) used in paints and adhesives are available as homopolymers and copolymers in the form of solid resins, solutions, and dispersions. For a detailed description of poly(vinyl esters), see ⟶ Poly(Vinyl Esters).

2.4.4.1. *Solid Resins*

Poly(vinyl acetate) and vinyl acetate copolymers with crotonic acid, vinyl laurate, and dibutyl maleate are important solid resins; some are available in solution.

Poly(Vinyl Acetate). Commercial grades of poly(vinyl acetate) [9003-20-7] differ in molecular mass and therefore in their viscosity when dissolved. Poly(vinyl acetate) is a physically drying binder that forms transparent, lightfast films with good hardness, gloss, and adhesive strength. It dissolves in lower alcohols, glycols, esters, ketones, and toluene. Thanks to their elastic properties, paints and adhesives based on poly(vinyl acetate) generally require little plasticizer. Plasticizers not only increase elasticity but also lower the glass transition temperature, which adversely affects blocking stability and can lead to sticky surfaces.

The neutral behavior of poly(vinyl acetate) allows the use of all conventional pigments. Poly(vinyl acetate) is highly compatible with ester-soluble nitrocellulose and improves the adhesion and lightfastness of the latter. Poly(vinyl acetate) can also be readily combined with phenolic resins, ketone resins, and colophony resins.

On account of their excellent lightfastness, high gloss, and physiological harmlessness, poly(vinyl acetates) are used as binders in nitrocellulose combination paints for paper, labels, cardboard, wood, leather, and certain plastics. Low molecular mass grades are used in impregnation coatings that are resistant to oil, grease, and moisture, and in priming coats on cardboard or masonry. On account of the thermoplastic properties, heat-sealable lacquers can also be formulated. Poly(vinyl acetate) is an extremely important raw material in the adhesives industry.

Commercial products include Gelva (Monsanto); Mowilith (Hoechst); Rhodopas (Rhône – Poulenc), Vinac (Air Products); Vinavil (Montedison); and Vinnapas (Wacker).

Vinyl Acetate Copolymers. Copolymerization of vinyl acetate with other monomers allows specific improvement of certain properties. Copolymers generally exhibit a broader compatibility than the homopolymer. For example, softer, permanently flexible polymers with a lower water uptake and higher alkali resistance are obtained by polymerizing vinyl acetate with vinyl laurate. On account of their thermoplastic properties these copolymers are used in heat-sealable finishes on paper, cardboard, and aluminum foil. In cellulose nitrate lacquers they increase adhesion, lightfastness, and the body fullness of the paint film. They can also be used for priming coats and for stabilizing porous or absorbent substrates.

Copolymers of vinyl acetate and dibutyl maleate are used in adhesives and as binders in deep sealers. Copolymers of vinyl acetate and crotonic acid dissolve in aqueous alkalis with salt formation. The carboxyl groups confer better metal adhesion.

Vinyl acetate copolymers are also used as a raw material for wash-off adhesives, textile finishing agents, and marking inks with extremely good adhesion on a wide variety of substrates.

Commercial products include Mowilith (Hoechst) and Vinnapas (Wacker).

2.4.4.2. *Dispersions*

Poly(vinyl ester) dispersions are quantitatively more important than solid resins. Homopolymer and copolymer dispersions are used for binders in emulsion (dispersion) paints, plastic-bonded plasters, and water-thinnable adhesives. Poly(vinyl acetate) dispersions are less important than vinyl acetate copolymer dispersions. The most important comonomers of vinyl acetate are vinyl laurate, dibutyl maleate, Versatic Acid esters (VeoVa, Shell), ethylene, vinyl chloride, and butyl acrylate. Poly(vinyl propionate) and copolymers of vinyl propionate with butyl acrylate, styrene, or vinyl chloride are also marketed and used as dispersions.

The properties of the films and coatings obtained from dispersions depend primarily on the polymer composition, stabilization system, and particle size. The type of polymer determines film-forming properties, resistance to hydrolysis, and to some extent the water resistance, flammability, and mechanical properties such as flexibility, elongation at break, and tensile strength. The stabilization system [protective colloids such as poly(vinyl alcohol) and cellulose derivatives or emulsifiers] influences behavior under mechanical stress, pigment and extender compatibility, pigment loading, water resistance, and rheology.

Film formation in polymer dispersions occurs as a result of the agglomeration and fusion of the polymer particles (diameter 100 – 1000 nm) after evaporation of the water. The minimum temperature at which the particles fuse to form a film (minimum film-forming temperature) is related to the glass transition temperature. It can be lowered to facilitate paint application by adding plasticizers or solvents. Poly(vinyl acetate) dispersions form lightfast, dry, hard, brittle films. Plasticizers therefore have to be used (external plasticization), which are, however, volatile and lead to embrittlement of the films after a relatively short time. Internally plasticized dispersions of copolymers of vinyl acetate with vinyl laurate, butyl maleate, Versatic Acid esters, or ethylene form permanently flexible, nonaging films that are not, however, always sufficiently resistant to hydrolysis. Terpolymer (vinyl acetate – ethylene – vinyl chloride) dispersions form films that are more resistant to hydrolysis than homopolymer and copolymer dispersions. The films also have a higher mechanical strength and lower flammability. The glass transition temperature of the terpolymer can be varied within wide limits and properties can be matched to requirements by using a suitable choice of

comonomers. The same is true of vinyl propionate copolymer dispersions.

Poly(vinyl ester) dispersions are important binders for indoor (conventional, solvent-free) and outdoor paints, special coatings, and textured finishes. Special types are used for wood paints and for coating paper and cardboard. Poly(vinyl ester) dispersions are also important in the adhesives and textile finishing industries.

Commercial products include Airflex (Air Products); Dilexo (Texaco); Elotex (Ebnöther); Emultex (Revertex); Ertimul (ERT); Mowilith (Hoechst); Propiofan (BASF); Ravemul (ANIC); Rhodopas (Rhône – Poulenc); Ubatol (Cray Valley Kunstharze); Ucar (Union Carbide); Vinamul (Vinyl Products); Viking (Kirkless Chemicals); Vinnapas (Wacker); and Walpol (Reichhold).

2.4.5. *Poly(Vinyl Alcohol)*

Poly(vinyl alcohol) [9002-89-5] is obtained by hydrolysis of poly(vinyl acetate) (⟶ Polyvinyl Compounds, Others). Commercial grades of poly(vinyl alcohol) differ in the degree of polymerization (molecular mass) and degree of hydrolysis [residual poly(vinyl acetate) content].

The water solubility of these polymers declines with decreasing hydrolysis and increasing molecular mass. Poly(vinyl alcohols) containing up to 20 wt % vinyl acetate are insoluble in organic solvents. Poly(vinyl alcohol) is resistant to oils, fats, greases, and waxes. Films obtained from aqueous solutions are clear and colorless, have a high crack resistance, exhibit good lightfastness, and are impermeable to water vapor.

Water-soluble organic compounds with highly polar groups and a high boiling point may be used as plasticizers. Poly(vinyl alcohol) has a good pigment binding capacity and is compatible with the pigments and extenders conventionally used in the paint industry. Resistance to water can be improved by reacting the hydroxyl groups with aldehydes or by cross-linking with urea resins or melamine resins in the presence of acid catalysts.

Poly(vinyl alcohol) is used in the coating sector (e.g., as a thickening agent for aqueous systems), in adhesives, for finishing paper and cardboard, for coating paper, as a binder for strippable coatings, and as a protective colloid for dispersions.

Commercial products include Airvol (Air Products); Elvanol (Du Pont); Ertivinol (ERT); Gohsenol (Nippon Gohsei); Mowiol (Hoechst); Polyviol (Wacker); Poval (Kuraray, Denka, Shinet-Su); and Rhodoviol (Rhône – Poulenc).

2.4.6. *Poly(Vinyl Acetals)*

Poly(vinyl acetals) are produced by reacting poly(vinyl alcohol) with aldehydes (⟶ Polyvinyl Compounds, Others). Since acetalation does not proceed quantitatively and a proportion of the acetyl groups remains after hydrolysis of poly-(vinyl acetate) to poly(vinyl alcohol), poly(vinyl acetals) may be regarded as terpolymers of vinyl alcohol, vinyl acetal, and vinyl acetate.

Poly(vinyl formals) and poly(vinyl butyrals) are of importance in the paint industry. Commercial products differ in the degree of polymerization and acetalation, but especially in the residual poly(vinyl alcohol) content.

Poly(Vinyl Formals). Poly(vinyl formals) [63148-64-1] are physically drying binders. The films exhibit high resistance to chemicals and favorable mechanical properties. Relatively powerful organic solvents are required for dissolution (e.g., chlorinated hydrocarbons, cyclic ethers, and mixtures of alcohols and aromatic hydrocarbons). The free hydroxyl groups of the polymers confer good pigment wetting and adhesion to various substrates. Poly(vinyl formals) can be cross-linked via the hydroxyl groups; this improves chemical resistance.

Poly(vinyl formals) are compatible with a range of plasticizers (that are used, for example, to improve the low-temperature flexibility) and binders, in particular polyisocyanate, phenolic, epoxy, and melamine resins.

The most important use of poly(vinyl formals) is for coating wire. They are also used for coating magnetic recording media and as adhesives for metal – metal composites used in construction.

Commercial products include Formvar (Monsanto); Pioloform FN (Wacker); Polyvinylformal (Siva); and Vinylec F (Chisso).

Poly(Vinyl Butyrals). Poly(vinyl butyrals) [63148-65-2] are physically drying binders. On account of their free hydroxyl groups they can also be used in reactive systems. Non-cross-linked films are thermoplastic and therefore heat-sealable. As well as having a high heat resistance and lightfastness, poly(vinyl butyrals) also exhibit good resistance to fats, greases, oils, bitumen, and gasoline. Outstanding properties are their good adhesion to metals, glass, absorbent substrates, and plastics sheets, as well as their excellent pigment wetting properties. Alcohols are preferred as solvents; esters and ketones have a somewhat lower solvent power.

Poly(vinyl butyrals) form very flexible films. Plasticizers can be used to improve cold flexibility and reduce solution viscosity. Poly(vinyl butyrals) are also compatible with a wide range of resins (e.g., epoxy, urea, melamine, phenolic, polyisocyanate, nitrocellulose, polyethyleneimines, and ketone resins). The resistance of poly(vinyl butyral) to chemicals can be improved by cross-linking. As an additive poly(vinyl butyral) reduces the brittleness of highly cross-linked coatings and improves the adhesion, particularly to metals. Polymer compatibility, the solubility, and plasticizer compatibility depend on molecular mass and polarity (degree of acetalation). Poly(vinyl butyrals) are compatible with the pigments and extenders normally used in the paint industry. The free hydroxyl groups confer outstanding pigment wetting.

Poly(vinyl butyrals) are an important class of binders. One of their main uses is in priming coats for corrosion protection. Poly(vinyl butyrals) are employed together with corrosion protection pigments and phosphoric acid as adhesion priming coats (wash primers); combination with a suitable finishing coat confers protection against corrosion and subsurface rusting in metals. Reactive primers or reinforced primers are obtained by combining poly(vinyl butyrals) with phenolic resins that contain corrosion protection pigments and phosphoric acid. On account of their very high solubility in solvents such as ethanol, their good flow properties, and excellent pigment wetting, poly(vinyl butyrals) are used for formulating flexographic and gravure printing inks for food packaging. High molecular mass grades are used to produce glass composites (laminated safety glass) on account of their lightfastness and good adhesive power. Poly(vinyl butyrals) are also important as binders for primer sealers (e.g., to prevent the migration of bitumen) and as a temporary binder in ceramics production.

Commercial products include Butvar (Monsanto); Denka Butyral (Denki Kagaku); Mowital (Hoechst); Pioloform B (Wacker); and S-Lec-B (Sekisui).

2.4.7. *Poly(Vinyl Ethers)*

Poly(vinyl ethers) [⟶ Poly(Vinyl Ethers)] formed from methyl ethyl ether or isobutyl ether are used as soft plasticizing resins. Solubility and compatibility depend on the alkyl group. In the paint sector, poly(vinyl ethers) are used mainly as plasticizing and in some cases as adhesion-improving resins for chlorine-containing binders, styrene polymers, nitrocellulose, and brittle resins. An example of a commercial product is Lutonal (BASF).

2.4.8. *Polystyrene and Styrene Copolymers*

For a detailed description of polystyrene and styrene copolymers, see ⟶ Polystyrene and Styrene Copolymers.

Polystyrene Dispersions. On account of their glass transition temperature (T_g) of ca. 100 °C, polystyrene dispersions do not form films at room temperature. These rigid polymers can only be applied with means of heat drying (e.g., to stiffen fabrics and nonwovens). Film formation is not required in agents used to protect floor coverings and paper coatings (plastic pigments); in this case polystyrene is therefore applied in the form of a dispersion at room temperature.

Styrene Copolymer Dispersions. The T_g and hardness of polystyrene can be adjusted over a wide temperature range by copolymerization of styrene with soft monomers such as butadiene and acrylate esters. Styrene – butadiene (SB) dispersions are quantitatively the most important. With a styrene – butadiene weight ratio of 85:15 the T_g is ca. 80 °C, at a ratio of 45 : 55 the T_g is ca. –25 °C. On account of the cross-linking capability of butadiene, SB copolymers are not thermoplastics, but elastomers. Elasticity can be modified by controlling the molecular mass and degree of cross-linking.

Styrene – butadiene dispersions are generally stabilized with anionic, or anionic and nonionic emulsifiers. Carboxylation (incorporation of a small proportion of unsaturated carboxylic acids) of SB dispersions increases their stability and improves adhesion to various substrates. Almost all SB dispersions used in the coating sector are carboxylated.

Styrene – butadiene dispersions may undergo oxidative post-cross-linking via the double bond in the butadiene unit. Uncontrolled oxidation leads to embrittlement and, finally, to breakdown of the binder (chalking). The dispersions are thus unable to satisfy stringent requirements regarding color stability and UV resistance.

In conventional exterior-use paints SB dispersions have largely been replaced by styrene – acrylate dispersions, and their use is now restricted to special applications (corrosion protection primers, wood primers, mortar modification) where low film permeability to gases, water vapor etc. and complete resistance of the polymer to hydrolysis are necessary. In order to achieve a uniform surface and thus improve printability, paper and card are coated with paper-coating colors. Carboxylated SB dispersions are used in these paints as binders.

The many uses of SB copolymer dispersions in the textile sector include tufting, textile, and carpet coatings; needle felt and card web reinforcement; textile finishes; and impregnation.

In addition to butadiene, other monomers can be used for internal plasticization of polystyrene. In the paper sector, styrene – butadiene – acrylate terpolymers have advantages in special applications.

Soluble Polystyrene. Styrene homopolymers (generally in admixture with other binders) are used in special applications such as zinc dust primers and bronze lacquers, in paints for mineral substrates and paper, as well as in adhesives. Polystyrene and poly(α-methylstyrene) have good water resistance, low permeability, resistance to chemicals and light, and also dry quickly. Suitable solvents include aliphatic and aromatic hydrocarbons, in some cases containing ketones or esters.

Soluble Styrene Copolymers. Copolymers internally plasticized by acrylate esters, butadiene, maleate esters, or acrylonitrile are of greater importance than the homopolymers. The proportion, nature, and molecular mass of the comonomer determine the softening point, solubility, flexibility, drying rate, and resistance to water, UV light, and chemicals. Incorporation of butadiene decreases lightfastness, while acrylic acid confers solubility in alcohol. The range of potential uses extends from soft resins for adhesives and paper, to rigid resins for paints and hot-melt adhesives. Styrene – and vinyltoluene – acrylate polymers are used particularly in exterior-use paints, paints for concrete and road-marking, metal lacquers, and printing inks. Styrene – allyl alcohol copolymers may also be used as cross-linkable polyol components.

Commercial Products.
Soluble systems include Amoco Resins (Amoco); Hercoflex, Kristaflex, Piccolastic, Piccotex (Hercules); Pliolite, Pliowag (Goodyear); RJ-Products (Montesano); andSupraval (BASF).

Dispersions. The main suppliers of SB dispersions are: BASF, Bayer, Doverstrand, Dow Chemical, Enichem, Hüls, Rhône – Poulenc, Synthomer. Commercial coating products include Butonal, Styrofan (BASF); Dow-Latices (Dow Chemical); Lipaton (Hüls); Rhodopas (Rhône – Poulenc); andSynthomer-Latices (Synthomer).

2.5. Acrylic Coatings

General Properties. Paints containing acrylic resins as binders have been known since the 1930s. They are now one of the largest product classes in the paint and coatings sector. For a detailed description of acrylate polymers, see ⟶ Polyacrylates, ⟶ Polymethacrylates.

Polyacrylates as binders consist of copolymers of acrylate and methacrylate esters. Other unsaturated monomers (e.g., styrene and vinyltoluene) may also be incorporated, but usually to a lesser extent. Copolymers formed exclusively from acrylates and/or methacrylates are termed pure acrylics. The comonomers differ as regards the alcohol residues of the ester group, which also allow incorporation of additional functional groups. Choice of suitable monomers allows wide variation of the physical and chemical properties of the resulting polymer. Hydrophilicity, hydrophobicity, and acid – base properties can be adjusted; resins containing hydroxyl, amine, epoxy, or isocyanate groups can also be produced.

The resin products may be solids, solutions in organic solvents or water, emulsions, or dispersions.

Acrylate resins have several advantages over other paint binders [50-54]:

1. Polyacrylates are only slightly attacked by chemicals, and confer a high degree of resistance to paints.
2. Polyacrylates are colorless, transparent, and do not yellow, even after prolonged thermal stress.
3. Polyacrylates do not absorb above 300 nm and are therefore not degraded by UV radiation (as long as they do not contain styrene or similar compounds).
4. Polyacrylates do not have unstable double bonds.
5. Polyacrylates have outstanding gloss and gloss retention.
6. Acrylates, and especially methacrylates, are stable to hydrolysis.

The following properties of the coating can be ascribed to individual monomers [51]. Methyl methacrylate promotes weather resistance, lightfastness, hardness, gloss, and gloss retention. Styrene increases hardness and resistance to water, chemicals, and salt spray, but reduces lightfastness and gloss retention. Alkyl acrylates and alkyl methacrylates impart flexibility and hydrophobicity, while acrylic acid and methacrylic acid improve adhesion to metals.

The increasing importance of environmental considerations places new requirements on paint resins and has broadened the range of paint systems. Paints are now required that have a low solvent content (medium-solids, high-solids coatings) or are solvent-free (powder coatings), that can be adjusted by dilution with water (waterborne paints), and that are thermoplastic or capable of undergoing cross-linking. All of these properties must be obtained via the polymer structure of the binders. Important parameters are described below.

The *glass transition temperature* (T_g) affects adhesion, flaking and peeling from the substrate, crack formation, and resistance to impact shock [54]. In acrylates adjustment of T_g is achieved relatively easily via the ratio of methyl methacrylate (T_g of the homopolymer +105 °C) to *n*-butyl acrylate (T_g of the homopolymer – 54 °C) [50, 51, 55]. The T_g also influences properties of dispersions [55] and the viscosity of solutions [56-58]. A high T_g value is associated with a faster drying rate [59]. In the low molecular mass range (< ca. 6000), which is of interest particularly for high-solids paints, the T_g depends on the molecular mass. Subsequent cross-linking leads to an increase of T_g which is dependent on the cross-linking density.

The *styrene content* in the binder reduces resistance to yellowing and weathering [59], but improves resistance to chemicals, hydrophobic properties, adhesion, and pigment wetting. Styrene is therefore largely avoided in topcoat paints for exterior use and in clearcoats.

The development of low-solvent (high-solids) paints requires resins with a very low *viscosity*. The principal viscosity-determining parameters for such binders are the molecular mass and molecular mass distribution [55-57], [60-63]. Oligomers with a molecular mass of ca. 1000 – 3000 are required for high-solids paints [58], [64]. An acrylate binder with a molecular mass of 100 000 can be processed to form a paint with 12.5 % solids content at the application viscosity; a molecular mass of ca. 6000 results in a paint with 50 % solids content [56]. A narrow molecular mass distribution is beneficial in achieving low viscosity [55], [60]. However, the mechanical properties of a paint are favored by a high molecular mass [54]. Low molecular mass binders that cross-link after application are therefore used exclusively for high-solids paints. When it is applied, the paint contains low-viscosity oligomers; a highly polymeric system is formed after cross-linking and curing [58], [60]. Further possibilities of reducing the viscosity include specific interactions between the binder molecules [65] and choice of a low-viscosity solvent [57] that does not interact significantly with the resin [58]. The melt viscosity is an especially important criterion in powder coatings; acrylic resins have disadvantages in this respect when compared with polyesters [55].

Incorporation of functional groups in the polymer skeleton is necessary for the *production of dispersions*. Most water-dispersable systems are polymers with free carboxyl groups. Water thinnability is achieved by neutralizing the acid groups with aqueous alkali or amines. Binders may also contain basic nitrogen-containing groups; dispersion can then occur after neutralization (e.g., with acetic or lactic acid). Since the viscosity of dispersions is very low irrespective of the molecular mass, polymers of very high molecular mass are generally used; dispersions are therefore ideal for physically drying coatings.

Solvent emission from paints can be reduced without lowering the molecular mass by using *nonaqueous dispersions* (NAD) (see Section Nonaqueous Dispersion Paints). Acrylates have been described as binders for NADs, but apart from a low viscosity they offer few advantages over conventional coatings [60] and moreover must compete with high-solids paints and powder coatings [66].

Cross-Linking of Acrylate Resins. In contrast to linear polymers, cross-linked polymers are insoluble and are also harder and more resistant to chemicals; these properties are extremely important for high-grade coatings. Cross-linking reactions became important in the 1950s with the introduction of acrylic resins in the automotive sector.

A further impetus was given by increasingly stringent environmental legislation; lower solvent contents and the replacement of conventional solventborne paints by medium-solids and high-solids paints mean that the molecular mass of the binders had to be lowered to a range where the required paint properties (e.g., film formation, hardness, and flexibility) no longer exist. These properties must be obtained by increasing the molecular mass by cross-linking after application. Chemical reaction after application also provides advantages for high molecular mass dispersions; the glass transition temperature and film hardness are increased.

A widely used method for cross-linking paint films consists of reacting hydroxyl-containing acrylates with melamine – formaldehyde resins or urea – formaldehyde resins [50], [63], [67]. The hydroxyl-containing acrylates are prepared using comonomers such as hydroxyethyl (meth)acrylate or butanediol monoacrylate. Amino resins are to some extent self-cross-linking, but also form cross-links with acrylates via the hydroxyl groups [62]. Cross-linking can be effected by curing at ca. 130 °C or by acid catalysis. The paints exhibit outstanding gloss and weather resistance [68].

A second important cross-linking method is the combination of hydroxyl-containing acrylates with polyisocyanates as hardener. This mixture forms cross-links at room temperature and must therefore be produced and stored as a two-component (two-pack) system [50]. Reaction between aromatic isocyanates and the hydroxyl-containing acrylates proceeds extremely rapidly; as aliphatic isocyanates react slowly, the reaction has to be catalyzed (e.g., with dibutyl tin dilaurate, amines, or acids) [69]. The properties of such polyurethane paints (see also Section Polyurethane Coatings) are superior to those of most other coating materials, and they are used in an increasing range of applications. One-pack polyurethane paints are also available and are based on hydroxyl-containing acrylates; blocked isocyanates are used as hardeners. Such systems generally require relatively high stoving temperature (> 150 °C).

A third group of cross-linking reactions involves acrylic resins containing free carboxylic acid groups. Polyepoxides are mainly used as cross-linking agents in paints dissolved in organic solvents or powder coatings. This combination is superior to other combinations (e.g., isocyanate cross-linking or melamine resins) as regards hardness and resistance to detergents, lyes, and solvents. Extremely high stoving temperatures (≥ 200 °C) are required; the temperature can be reduced to ca. 120 – 150 °C if tetrabutylammonium iodide or tertiary amines are used as catalysts [67-70]. Use of catalysts is, however, offset by a low storage stability (only a few weeks).

If less stringent requirements with regard to chemical resistance, abrasion, and hardness are placed on the cross-linking density, carboxyl-containing acrylates can also be cross-linked by salt formation (e.g., with diamines [54] or metal complexes [71]). This procedure is widely employed, particularly for aqueous dispersions. Cross-linking with bisoxazolines has also been reported [55]. Epoxy groups can be incorporated into the binder via glycidyl (meth)acrylate and cross-linked with dicarboxylic acids [50].

Aqueous acrylic dispersions are being increasingly used for corrosion protection coatings. These paints are often not baked; instead, mechanical properties are improved by cross-linking at room temperature. Aziridines or bishydrazides are generally used as the cross-linking agent and are mixed with the dispersion after the production process.

Many other cross-linking processes exist that are not widely adopted or have only recently been developed. Cross-linking of epoxy-containing acrylates with amino resins and reactions with polysulfonazides have been reported [67]. Reactions of unsaturated acrylic groups with ketimines and reaction of acetoacetate-containing polymers with ketimines should also be mentioned [62], [72].

An alternative way of curing paints is to produce self-cross-linking acrylic resins that react internally at elevated temperature without the addition of an external cross-linking agent. Such systems have advantages as regards resistance to chemicals, hardness, and elasticity, but are less variable in their formulation and can present problems as regards storage stability. Furthermore, in order to achieve a sufficiently high degree of cross-linking a minimum molecular mass is required that is higher than that of resins which do not self-cross-link; high-solids paints accordingly cannot be produced with such systems.

Applications. Acrylic paints are used in many different areas and applied by all commonly used methods. Low-solvent paints and aqueous dispersions require special formulation.

Acrylic paints with a low-solvent content (medium- and high-solids systems) present some difficulties in application. Only a very small amount of solvent evaporates from the wet film, the increase in viscosity during drying is slight, and the paint remains liquid with a low viscosity for a long time. If an excessively thick coating is applied, this can easily lead to sagging and evaporation is also hindered; control of the layer thickness therefore becomes difficult. During stoving, viscosity initially falls; viscosity and thus paint stability only increase after initiation of the cross-linking reaction and the associated increase in molecular mass. The rheology of high-solids paints must therefore be finely controlled with appropriate additives during application and stoving. These additives act as thixotropic agents and ensure good leveling during application without adversely affecting the stability of thick layers [57], [62]. Sagging can also be avoided by applying the paint at elevated temperature (e.g., 50 – 60 °C). The low molecular mass of high-solids paints also affects the stoving process. The cure window (i.e., the temperature and time ranges used for stoving) is therefore smaller the lower the molecular mass of the binder [63]. Paints produced from carboxyl-containing acrylic resins and diepoxides as cross-linking agents exhibit better rheological behavior during application and stoving, presumably on account of the reversible formation of intermolecular hydrogen bonds [70].

Problems are also associated with *waterborne acrylic paints*. Water has an evaporation enthalpy of 2.26 kJ/g and thus requires substantially more energy for drying than organic solvents (evaporation enthalpies ca. 500 J/g). Pitting and blistering often occur at stoving temperatures above 100 °C. In physically drying paints the drying behavior depends strongly on the atmospheric humidity and the application zone often has to be air-conditioned. The stability of acrylate dispersions depends on a delicate equilibrium between factors such as pH, emulsifier content, and sedimentation stability [72]. The formulation of a paint dispersion (emulsion paint) therefore requires additional components (e.g., amines as solubility promoters, dispersants, defoamers, thickening agents) and is much more complicated than that of a conventional solvent-based paint [56]. Aqueous dispersions generally contain a few per cent of high-boiling solvents which act as temporary plasticizers and lower the minimum film-forming temperature, thus allowing film formation to occur. A film of sufficient hardness is obtained only after complete evaporation of the solvent. The film may, however, be somewhat hydrophilic due to the presence of carboxyl groups and emulsifier residues; this reduces its water resistance, gloss, and gloss retention.

Uses. The largest application sector for acrylate binders are emulsion paints for ceilings, walls, and building fronts. These emulsion paints are generally physically drying and only contain a small amount of binder; the main constituents are pigments and extenders. Acrylate dispersions for paints have a good water vapor permeability and good water resistance.

Acrylic resins have been used in the automotive sector since 1957. They are now important binders in automotive finishes and topcoats, and have replaced alkyd resins in many cases. Advantages for automotive finishes and coatings are high transparency, weather resistance, gloss retention, and yellowing stability [50]. Automotive finishes are always cross-linked; melamine resins are generally used as hardeners [54], polyisocyanates are being increasingly used for clearcoats [62]. Automotive fillers and metallic basecoats may be formulated with acrylate dispersions to reduce solvent emission. Solvent-containing systems are, however, still indispensable in the topcoat sector.

In the industrial coatings sector, acrylates are used to satisfy special requirements. Self-cross-linking acrylates are used for domestic appliances on account of their good resistance to detergents and lyes as well as their temperature and yellowing stability. Metal substrates are stoved with combinations of acrylate and melamine or urea resins; recently, stoving dispersions have also become available. Wood surfaces subjected to a high degree of wear (e.g., in kitchens) are being increasingly coated with polyurethane paints based on acrylic resins and isocyanates.

2.6. Alkyd Coatings

The binders used in the production of alkyd paints, namely alkyd resins, are oil-modified or fatty-acid-modified condensation polymers of polybasic acids and polyhydric alcohols. For further details, see ⟶ Alkyd Resins. Alkyd resins account for ca. 45 % of the total world production of paint raw materials, excluding plastics latexes and polyvinyl dispersions.

Production of alkyd resins began in 1930. They owe their importance to their universal applicability. In the architectural coatings sector, alkyd resins replaced stand oils on account of their faster drying and curing rates as well as their better film hardness and gloss retention.

Alkyd paints are used for protection and decoration in virtually all sectors including coatings in the steel, sheet, and metal processing sector, house and decorative paints, do-it-yourself paints, wood varnishes, wood protection agents, and automotive finishes. Alkyd resins are extremely important for corrosion

protection coatings.

2.6.1. *Alkyd Resin Binders and Uses*

DIN 53 183 defines an alkyd resin as follows: “Alkyd resins are synthetic polyester resins produced by esterifying polyhydric alcohols with polybasic carboxylic acids. At least one of the alcohols must be trihydric or higher. Alkyd resins are always modified with natural fatty acids or oils and/or synthetic fatty acids. In order to obtain particular application technology properties, alkyd resins may be additionally modified with compounds such as resin acids, benzoic acid, styrene, vinyltoluene, isocyanates, acrylic, epoxy, or silicone compounds.”

Worldwide alkyd resin production in 1988 (including captive-use amounts) was estimated to be ca. 2.4×10^6 t. Detailed information on commercially available alkyd resins can be found in [73].

The properties of alkyd resins can be widely modified by selecting appropriate vegetable oils and fats/fatty acids, varying their content in the resin, and by using special synthetic acids and a variety of modifying agents.

Classification. Normal alkyd resins containing oil/fatty acid, dicarboxylic acid (mainly phthalic acid or phthalic anhydride), and polyhydric alcohol are classified as short oil (≤ 40 % oil content), medium oil (41 – 60 % oil content), and long oil (61 – 70 % oil content) resins. Very long oil alkyd resins with 71 – 85 % oil content are mainly synthesized using isophthalic acid. The term oil content denotes the triglyceride (fatty acid) content relative to the solvent-free resin. Alkyd resins are subdivided into drying and nondrying resins according to the nature of the oils/fatty acids used. See ⟶ Alkyd Resins.

Long oil drying alkyd resins are used as binders for architectural and do-it-yourself paints, and are mainly modified with soybean oil. In the absence of direct light, they do not yellow as much as linseed oil alkyd resins and are less expensive than safflower oil types. Long oil linseed oil alkyd resins are principally used in corrosion protection coatings and colored architectural coatings.

Medium oil alkyd resins (oil content ca. 50 %) are based on drying oils or fatty acid mixtures, and are used as binders for air-drying and forced-drying machinery coatings and industrial coatings. They are also used in car repair finishes (refinishes), for original equipment manufacturing (OEM) finishes, and refinishes of trucks and large-capacity vehicles.

Short oil alkyd resins are used in combination with amino resins for industrial stoving finishes (e.g., for metal furniture, radiators, bicycles, garage doors, and small steel items). Dehydrated castor oil types mixed with alkyd resins based on synthetic acids are combined with melamine resin for OEM automotive finishes. Short oil resins are also combined with urea resins to produce acid-curing paints for wood and mixed with nitrocellulose for inexpensive, easily applied furniture lacquers and automotive refinishing systems.

Combination of Alkyd Resins with Other Binders. Alkyd resins can be combined in many different ways with other paint raw materials. In one method alkyd resins are mixed with preformulated products; this is normally carried out in the paint factory. Alternatively, the reactivity of the initial products can be employed to modify the alkyd resin during production.

Alkyd resins can be mixed with nitrocellulose, chlorinated rubber, highly chlorinated poly(vinyl chloride), amino resins, and colophony-modified phenolic and maleic resins. Chlorinated rubber and highly chlorinated poly(vinyl chloride) are combined with alkyd resins with the addition of plasticizers for corrosion protection coatings. These combinations have a higher resistance to chemicals than pure alkyd resin coatings. Combinations with amino resins are used in oven-drying industrial coatings and automotive finishes. The gloss, body, adhesion, and the resistance to preserving wax are determined by the alkyd resin. Amino resins allow rapid drying at elevated temperature and improve the mechanical resistance of the coatings.

Chemical reactions with alkyd resins can take place via their hydroxyl or carboxyl groups as well as via the double bonds of the unsaturated fatty acids. Isocyanates, epoxy resins, or colophony, for example, may be reacted with the hydroxyl groups. The carboxyl groups can be reacted with polyamidoamines (reaction products formed from dimerized linoleic acid and ethylenediamine) to form thixotropic resins, or can react with hydroxy-functional silicone precondensates. The double bonds of the unsaturated fatty acids permit copolymerization with vinyl compounds [e.g., styrene or (meth)acrylic acid derivatives].

Styrene-Modified Alkyd Resins. Styrene was first copolymerized with drying oils in the early 1940s. These binders had a substantially improved drying behavior and resistance to water and chemicals compared with untreated oils. The first patents for the production of styrenated alkyd resins were granted in England in 1942 [74]. Mixtures of styrene and α-methylstyrene are also used for a more reliable reaction; vinyltoluene yields copolymers with improved thinnability in mineral spirit and pigment absorption.

Advantages of styrenated alkyd resins compared with nonstyrenated resins are faster drying, better resistance to water and chemicals, and a lower yellowing tendency of their films. Disadvantages include the poor drying and low solvent resistance of the paint films, which frequently lead to “lifting” during recoating. A defect of styrene – alkyd resin topcoats is their high surface sensitivity, especially to scratching; they are therefore mainly used in the primer sector where the danger of lifting can be avoided by appropriate pigmentation.

Modification with Other Vinyl Compounds. Low molecular mass acrylate and methacrylate esters are used to modify alkyd resins containing oil/fatty acid [75], [76]. Methacrylates give products that are more interesting from the paint technology aspect. Acrylated alkyd resins have a good pigment absorption capacity and yield paints with fast drying properties, a high permanent elasticity, and very good adhesion (e.g., to aluminum, brass, and some plastics). Special advantages include better weather resistance and gloss retention; through drying and surface (scratch) resistance are superior to those of styrenated alkyd resins.

Silicone-Modified Alkyd Resins. Hydroxy-functional silicone precondensates can be used to produce air-drying alkyd resins. Methoxy-functional silicone precondensates are used for oven-drying silicone-modified alkyd resins [77]. Air-drying coatings have outstanding chalking resistance and gloss retention, and thus substantially improved weather resistance. They are therefore often used on inaccessible objects that can only be repainted at long intervals. The U.S. Navy Specification TTE 049 prescribes silicone alkyd resins for marine coatings. In the oven-drying coatings sector, the improved gloss retention and the substantially higher thermostability/yellowing resistance are important.

Thixotropic Alkyd Resins. Thixotropic flow is always associated with structural viscosity. Thixotropic paints can be obtained by various methods. Use of a thixotropic alkyd resin is the best way of ensuring reproducible flow properties. The alkyd resins become thixotropic after reaction with 10 – 20 % polyamidoamines at elevated temperature (ca. 200 °C) [78], [79].

Thixotropic resins are mainly used for decorative paints and primers, and in corrosion protection coatings. Thixotropic alkyd resins prevent the formation of precipitates, improve brushability, and reduce the sagging of paints on vertical surfaces and from edges and corners. Thixotropic coatings (including those based on thixotropic urethane alkyd resins) are frequently used in corrosion protection because the required layer thickness and good edge covering can be achieved with a small number of layers.

Thixotropic air-drying and oven-drying alkyd resins with low oil contents are used as binders for one-coat industrial paints with a high dry-film thickness of 50 – 75 µm. In special-effect paints (e.g., hammer finishes, soft-feel coatings, textured finishes) they facilitate fixation of the desired effect and prevent smudging.

Urethane Alkyd Resins. Attempts to modify drying oils with isocyanates date from the early 1940s [80] and led to the development of urethane alkyd resins. Urethane alkyd resins are generally produced by reacting long oil alkyd resins containing excess hydroxyl groups with diisocyanates. Since conversion is quantitative, the properties of the long oil alkyd resins are largely retained in the final products. For example, urethane alkyd resins are readily soluble in mineral spirit, are compatible with many other paint raw materials, and have a good storage stability. Their advantages in paint technology are rapid drying, high hardness, very good film elasticity, and above-average abrasion resistance. Paint films are highly waterproof and show improved resistance to chemicals, even in the alkaline range.

Toluene diisocyanate and isophorone diisocyanate are mainly used to modify alkyd resins. Isophorone diisocyanate yields products with a lower yellowing tendency, higher pigment absorption, and considerably improved chalking resistance under weathering conditions; however, its use is limited by its considerably higher price.

Urethane alkyd resins are used as varnish binders for the interior and exterior coating of wood. Further applications include parquet floor seals, marine paints, and varnishes for the joinery and do-it-yourself sectors. They are used in pigmented paints to produce high-gloss and matt decorative coatings with extremely good mechanical resistance as well as resistance to cleaning agents and many disinfectants. They allow sanding to be carried out at an early stage in rapid-

drying primers. They open up interesting possibilities in the corrosion protection sector because their drying is less sensitive to temperature and atmospheric humidity than that of standard resins.

Alkyd Resins with Highly Branched Carboxylic Acids. Highly branched carboxylic acids are produced industrially from olefins, carbon monoxide, and water [81]. They are marketed by Shell under the name Versatic Acids (⟶ Carboxylic Acids, Aliphatic – Versatic Acids and Neo Acids). Their esters have an outstanding resistance to hydrolysis. Highly branched fatty acids are, however, difficult to esterify; their sodium salts are therefore reacted with epichlorohydrin to form glycidyl esters which can then react as the epoxides, for example, with carboxylic acids [82]. These products are marketed by Shell under the name Cardura resins. The glycidyl esters are esterified with phthalic anhydride or adipic acid and glycerol. Cardura resins are nondrying alkyd resins modified with Versatic Acid. Their fatty acid content is generally ca. 40 %. They are soluble in aromatic solvents, ketones, and esters, and are compatible with many amino resins, short oil alkyd resins, oven-drying acrylic resins, epoxy resins, and nitrocellulose. Their main areas of use are stoving finishes, combining extremely good hardness with high elasticity and resistance to chemicals. The coatings are resistant to overstoving and have a good gloss retention and chalking resistance under weathering conditions. Their very good adhesion (even on untreated metal) is an important advantage. They are used in enamels and coatings for domestic appliances, washing machines, and refrigerators, as well as for OEM automotive finishes.

Waterborne Alkyd Resins [83]. Alkyd resins can be converted into a waterborne form in two ways:

1. Alkyd resins with high acid numbers (> 50) are neutralized with amines (normal air-drying resins have an acid number < 10, oven-drying resins 20 – 35); solubility in water is due to salt formation.
2. Alkyd resins are emulsified in water after addition of emulsifying agents and stabilizers.

Neutralization with amines is mainly employed with medium and short oil resins that are modified with drying and semidrying oils/fatty acids and supplied as 70 – 80 % solutions in butyl glycol, or mixed with 1-methoxy-2-propanol or *N*-methylpyrrolidone. Neutralization is generally performed with triethylamine to obtain air-drying alkyd resins; dimethylethanolamine is used for oven-drying alkyd resins [84]. The amines used for neutralization represent an ecotoxicological problem, and attempts are therefore being made to find more environmentally compatible agents. In contrast to alkyd resin emulsions, waterborne alkyd resins permit the production of high-gloss paints. Sensitivity of the resins to hydrolysis during storage of the paints may cause problems and generally results in a decrease in viscosity. With air-drying systems the drying ability can also be adversely affected. These disadvantages are largely overcome by incorporating acids such as trimellitic acid which is less susceptible to hydrolysis for steric reasons; paints based on this system are therefore more stable. Coating materials based on amine-neutralized alkyd resins are used for forced-drying and oven-drying industrial paints (e.g., in combination with waterborne amino resins), for wood protection, and for corrosion protection.

Alkyd resin emulsions or “self-emulsifying” alkyd resins can be produced relatively easily (resin content 40 – 60 %). However, optimization of their properties requires a great deal of experience and careful selection of additives [85]. In order to obtain satisfactory water resistance and drying behavior, the content of emulsifiers should be kept as low as possible. The size of the emulsified particles is in the range 5 – 15 µm. A small particle size ensures good storage stability of the coating materials and improved gloss of the dried films; it also improves miscibility with polymer dispersions. The more homogeneous the system, the better the weather resistance of the resultant coating films. Alkyd resin emulsions are mainly used in architectural paints, corrosion protection coatings, wood protection agents, and house paints. They are particularly suitable for combination with polymeric, especially acrylate, dispersions. The use of alkyd resin emulsions in industrial coatings is still insignificant, an exception being furniture coatings and varnishes.

Alkyd Resins for High-Solids Paints. See also Section Solvent-Free and Low-Solvent (High-Solids) Paints. Alkyd resins have been developed for producing high-solids paints to reduce solvent emission. One-pack, oven-drying high-solids paints have solids contents of 65 – 70 wt % in the ready-for-use formulation. Air-drying high-solids paints have solids contents of 85 – 90 wt %. A high solids content can be obtained in three ways:

1. Use of solvents with stronger solvent power, so that resin solutions with lower viscosities are obtained with less solvent. This method is used for stoving paints based on conventional alkyd – melamine resin combinations; it can also be applied with certain restrictions to air-drying systems.
2. Use of reactive diluents that are incorporated into the film by condensation when the stoving paints harden.
3. Use of low-viscosity, reactive alkyd resins that have a lower solvent requirement than conventional products.

With high-solids stoving paints, sagging generally has to be prevented when coating vertical surfaces. Rheology-controlling additives can provide a remedy, but generally lead to a reduction in the solids content.

The solids contents of air-drying paints (i.e., architectural and do-it-yourself paints) can be adjusted relatively easily to 75 – 80 %. However, for a solids content of 90 %, very low viscosity binders with a low degree of condensation are necessary and often do not satisfy requirements concerning drying, hardness, and resistance of the coatings. Developments have been aimed at overcoming the deterioration in drying and mechanical properties by incorporating 10 – 20 % of an allyl ether [86-88].

Commercial products include Alftalat (Hoechst); Alkydal (Bayer); Beckosol (Reichhold Chemie); Beetle (BIB Chemical); Dynotal (Dyno Industries); Heso-Alkyd (Hendricks + Sommer); Rhenalyd (RWE – DEA); Setal (Synthese); Synolac (Cray Valley Products); Synthalat (Synthopol Chemie); Uralac (DSM Resins); and Vialkyd (Vianova Kunstharz).

Testing. See also Chapters Properties and Testing and Analysis. DIN 53 183 (Alkyd Resin Testing) describes the most important tests for alkyd resins in the Federal Republic of Germany. A comparable U. S. standard is ASTM D 2689–88 (Standard Practices for Testing Alkyd Resins). ISO 6744 (Binders for Paints and Varnishes – Alkyd Resins – General Methods of Tests) defines and classifies the term alkyd resins and contains a summary of international standards for testing alkyd resins. ASTM D 4712–87 (Standard Guide for Testing Industrial Water-Reducible Coatings) may also be mentioned; it includes standards covering the testing of wet paints (properties, storage, application, film formation) and paint films.

2.6.2. *Additional Raw Materials*

See also Chapter Paint Additives.

Solvents. Alkyd resins are generally supplied as concentrated (≥60 %) solutions (except in the case of aqueous dispersions and emulsions). Further solvent must be added to bring the paints into a ready-for-use state. The choice of solvent is determined by the solubility of the alkyd resin and ecological compatibility.

Mineral spirit is the most commonly used solvent for *long oil alkyd resins*. Additions of other solvents during paint production modify specific properties. The brushability and wet edge time can be improved by addition of higher boiling mineral spirit fractions (*bp* 180 – 210 °C), aromatics-free petroleum fractions, or isoparaffins [89]. (The wet edge time denotes the period that the paint can be brushed side by side without leaving overlap marks.) Additions of small amounts of glycol ethers (2.5 – 5 % of total solvent content) also improve the leveling and application of the paints. Lower alcohols and glycol ethers reduce and stabilize the viscosity. Additions of aromatic hydrocarbons of the solvent naphtha or xylene type improve adhesion of corrosion protection coatings to existing coats. Polar solvents should not generally be used in thixotropic paints because they destroy the thixotropy.

Mineral spirit is the most important solvent for *medium oil alkyd resins* in industrial air-drying and forced-drying paints. On account of the high viscosity of these resins, mineral spirit must be combined with solvents of stronger solvent power or with viscosity-lowering solvents (e.g., xylene, solvent naphtha, lower alcohols, and glycol ethers). These solvents promote leveling and stabilize the paints during storage.

Short oil alkyd resins require special solvent mixtures for each area of use. Principal solvents include aromatic hydrocarbons (e.g., xylene), solvent naphtha fractions, and glycol ethers. Lower alcohols, even in small amounts, have a powerful viscosity-lowering effect. The choice of solvent depends on the solvent power for the resins and the application method. The last solvent to evaporate during drying must be a true solvent for the binders because otherwise the paint film becomes opaque or inhomogeneous.

For their largest use sector (oven-drying industrial coatings) the short oil alkyd resins are combined with amino resins (melamine, urea, or benzoguanamine resins). The paints are cured at 80 – 250 °C. The amino resins are generally soluble in alcohols and glycol ethers; aromatic hydrocarbons are used as diluents for amino resins. Effective solvent mixtures often consist of aromatic hydrocarbons, glycol ethers, and alcohols in a volume ratio of 7 : 1 : 2 in the ready-for-use paint. The glycol ethers and alcohols improve the leveling and surface of the coat as well as the storage stability of the paints. With spraying and roller paints, relatively high-boiling solvents should also be used to achieve good leveling and avoid boiling phenomena (bubble formation). In dipping paints, small additions of low-boiling solvents can prevent sagging.

Short oil alkyd resins can also be combined with nitrocellulose. They are readily soluble in typical nitrocellulose solvents (esters and ketones). Special short oil

alkyd resins that are soluble in lower alcohols are often processed into furniture lacquers and acid-curing paints, and are also used in paper (spray) coatings and in flexoprinting inks.

The choice of solvents for *waterborne alkyd resins* is restricted. Resins neutralized with amines are normally supplied in glycol ethers or glycol ether – alcohol mixtures [90]. After neutralization they are generally diluted with water. Small amounts of propylene glycol are added to alkyd resin emulsions to prolong the wet edge time. Coalescing agents are used in dispersions with high latex contents. These specially formulated high-boiling solvents lower the film-forming temperature.

Driers are added to drying oils and drying alkyd resins to accelerate oxidative drying. Driers include oil- and solvent-soluble and solvent-soluble metallic soaps. The naphthenate and octoate salts of cobalt, lead, and manganese with the auxiliary metals calcium, zinc, and cerium are generally used. Calcium and zinc salts are added to the comminuted material during paint production to exploit their action as wetting agents.

For environmental reasons modern alkyd paints are dried without lead driers and with minimum amounts of cobalt driers [91]. Manganese driers are only used in colored paints and corrosion protection coatings because they affect the shade of white and pastel coatings. Special nondiscoloring driers (e.g., Dapro 5005, Daniel Products, USA) are used to obtain particularly pale white paints or honey-yellow varnishes.

Antiskinning agents suppress skin formation on air-drying alkyd paints in the can, particularly after the can has been opened. Antiskinning agents may be phenol derivatives (e.g., 4-*tert*-butylphenol) or oximes (e.g., methyl ethyl ketone oxime or butanaloxime). Oximes are used almost exclusively, at a concentration of 0.5 – 1.5 % relative to the solvent-free alkyd resin. They are colorless, do not affect paint color, and have a negligible effect on drying time.

Antisettling Agents. High-density pigments and extenders tend to settle, especially in low-viscosity paints. They form deposits that are difficult to dislodge. Antisettling agents are used to overcome this problem. They may be minerals (e.g., pyrogenic silica), organically treated minerals (e.g., montmorillonite treated with stearic acid), or organic compounds (e.g., hydrogenated castor oil). Excessive amounts of antisettling agents frequently cause thickening and have an adverse effect on rheological behavior. They should therefore be used with care.

Wetting Agents. A stable mixture of binder solution, pigments, and extenders is essential for the production of storable pigmented paints and primers. Pigments and extenders are usually covered by a layer of water molecules that makes wetting difficult or even impossible. Carbon black and some organic pigments can only be wetted with difficulty by weakly polar or nonpolar binder solutions due to their polar structure.

Wetting agents are surfactants and reduce the surface tension between the binder solution and the pigment particles. Wetting agents used in paint technology include calcium or zinc octoate or naphthenate, fatty alcohol sulfonates, castor oil acid derivatives and conversion products, poly(glycol ethers), and low-viscosity silicone oils.

Wetting agents are also often used as antifloating agents. Floating of pigments may occur in combinations of pigments of widely differing densities, particle sizes, differing polarities, and with unsatisfactory wetting.

2.6.3. *Production*

See also Chapter Production Technology.

In liquid coating materials, pigments and extenders have to be uniformly distributed and fully wetted in the binder solution or dispersion. The pigment agglomerates are therefore comminuted (grinding, dispersion) into primary particles and simultaneously incorporated into the binder solution. Grinding is usually carried out with a dissolver and agitator ball mill, but in some cases only with a dissolver. Grinding can be carried out in roller mills after predispersion of the pigments and extenders with kneaders or stirrers. Three-roll mills are used to produce highly viscous materials (e.g., fillers or printing inks).

Agitator ball mills (i.e., sand or bead mills) permit continuous operation and shorter dispersion times. Particle sizes of < 10 µm can be achieved with relatively fast throughputs. These mills are therefore also used for high-quality paints (e.g., for automotive finishes or car refinishing).

Production with a dissolver and a sand mill proceeds as follows:

1. Dilution of the required amount of binder solution
2. Addition of wetting and antisettling agents required for incorporation of pigments and extenders
3. Addition of the pigments and extenders in a relatively slowly running dissolver
4. Dispersion of the pigments and extenders with the sand mill
5. Monitoring of particle fineness
6. Addition of the remaining auxiliaries (e.g., driers and antiskinning agents)
7. Adjustment of the viscosity
8. Sieving
9. Testing, approval, and packaging

2.6.4. *Environmental and Health Protection Measures*

The solvents in the as-supplied alkyd resin solutions and in the paint composition represent a potential hazard due to their (eco)toxicological effects and flammability. Manufacturers and suppliers of resin solutions and paints must indicate the potential dangers involved in the handling, transportation, and storage of solvents by labeling drums and containers according to national and international regulations.

Health risks depend on the nature of the solvents and their concentration in the paint. Risks arise from solvent vapors during paint application. Working areas must be adequately ventilated to prevent vapor accumulation and respirators must be worn during spraying to prevent inhalation of paint aerosols.

Paint residues require special handling and disposal (special tips). Paints, solvents, and resins must under no circumstances be allowed to pollute the ground or effluent. Legal provisions concerning emission limits must be observed when applying paints.

Solvents differ as regards their miscibility with water and flash point and thus their potential fire hazards (flammability, combustibility). Explosive solvent vapor – air mixtures may be formed when applying paints. Spark formation (e.g., by static charging) must be avoided and hot surfaces or naked flames must be removed.

Combustible materials contaminated with air-drying and oven-drying alkyd paints, and spray dusts, may tend to exhibit spontaneous ignition due to autoxidation. Storage under exclusion of air and regular cleaning of the spray cabins improves safety.

2.7. Saturated Polyester Coatings

Polyester resins are condensation products of di- or polyfunctional monomers containing hydroxyl groups and carboxyl groups. The development of saturated polyesters began in 1901 with “Glyptal resins”, formed from glycerol and phthalic anhydride (Smith, United States). For a detailed description, see ⟶ Polyesters. Soluble resins obtained with fatty acids were first employed in 1925. Alkyd resins formed from unsaturated fatty acids can be cured by atmospheric oxidation (see Section Alkyd Coatings). Other unsaturated polyester resins are discussed in Section Unsaturated Polyester Coatings.

The subsequent development of saturated polyester resins for paints and other uses (e.g., adhesives, foams, fibers) was largely dependent on the introduction of new raw materials. A broad range of raw materials are now available, examples follow:

Polycarboxylic acids	
Terephthalic acid (dimethyl ester)	hard
Isophthalic acid	hard
Phthalic anhydride	hard

Hexahydrophthalic anhydride	hard
Trimellitic anhydride	hard
Adipic acid	soft
Sebacic acid	soft
Dodecanoic acid	soft
Dimerized fatty acid	soft
Polyalcohols	
Ethylene glycol	hard
1,2-Propanediol	hard
Neopentyl glycol	hard
1,4-Cyclohexanedimethanol	hard
Trimethylolpropane	hard
Diethylene glycol	soft
1,4-Butanediol	soft
1,6-Hexanediol	soft
Glycerol	soft

Paint polyesters are generally copolyesters composed of several polycarboxylic acids and polyalcohols. The glass transition temperature of copolyester resins can be varied within wide limits by choosing appropriate raw materials. Different polymer units (e.g., polyethers, polysiloxanes) can be incorporated by condensation into the molecular chain. Instead of terminal hydroxyl or carboxyl groups, the molecules may also contain other functional groups (e.g., acrylic or epoxide groups) that can be cross-linked with suitable reactants during paint hardening (curing) to form thermosetting paints. Saturated polyester resins for paints are available in various formulations [92]: low to high molecular mass, linear or branched, as solid or liquid resins, soft/elastic or rigid/brittle, and amorphous or crystalline. This versatility, together with their good resistance to light, moisture, heat, oxygen, and many other substances, accounts for the importance of saturated polyester resins as paint binders.

2.7.1. *Properties*

Molecular Mass. High molecular mass copolyesters (M_r 10 000 – 30 000) are generally linear. They are formed from terephthalic and isophthalic acids, aliphatic dicarboxylic acids, and various diols. Good solubility in conventional paint solvents is achieved by suitable formulation. In some cases (heat-sealable foil lacquers, printing inks, etc.), high molecular mass copolyesters are used as physically drying binders. Optimum paint film properties are, however, only obtained in combination with cross-linking resins. Special, crystalline, high molecular mass copolyesters are ground and used in thermoplastic powder coatings.

Linear, low molecular mass polyester resins are available with molecular mass up to 7000; branched grades have a molecular mass of up to 5000. These resins are not suitable for physically drying paints; they should be regarded as prepolymers for reaction systems with cross-linking resins. Classes of polyester prepolymers, their preferred cross-linking agents, and uses are summarized in Table 4.

Table 4. Classification of saturated polyester resins for paints

Structure	Class	M_r (weight average)	Cross-linking agent	Uses
HO∼∼∼∼OH	linear, high molecular mass	10 000 – 30 000	melamine resins, benzoguanamine resins	coil/can coating, flexible packaging, printing inks
HO∼∼∼OH	linear, low molecular mass	1000 – 7000	melamine resins, blocked polyisocyanate resins	coil/can coating, automotive/industrial paints
OH HO∼∧∼OH	branched, low molecular mass, hydroxy-functional	1000 – 5000	melamine resins, free/blocked polyisocyanate resins	automotive/industrial paints, polyurethane/powder coatings
COOH HOOC∼∧∼COOH	branched, low molecular mass, carboxy-functional	1000 – 5000	triglycidylisocyanurate, epoxy resins, melamine resins	powder coatings, waterborne paints
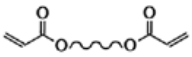	low molecular mass, acrylate group content	1000 – 5000	UV and electron-beam curing	paper/plastic coatings, printing inks

Terminal Groups. The polyester – amine system, consisting of hydroxy-functional polyester resins and, for example, melamine resins, is extremely important for the production of heat-curable industrial stoving enamels [93].

Polyurethane systems based on hydroxypolyester resins and free polyisocyanate resins (two-pack polyurethane paints, Section Polyurethane Coatings) are used particularly for finishes of road vehicles, railroad vehicles, aircraft, and plastics. Blocked polyisocyanate resins are available for formulating one-pack stoving enamels which are particularly suitable for coil and can coatings [94].

The reaction system consisting of “acid” (i.e., carboxy-functional) polyester resins and epoxy resins is important for powder coatings [95].

Acid polyester resins can be neutralized (e.g., with amines), diluted with water, and processed into aqueous stoving enamels [96].

Acrylate-modified polyester resins have proved suitable as binders for radiation-curable coatings [97].

Glass Transition Temperature. The T_g of saturated polyester resins can be varied by a suitable choice of aromatic or aliphatic raw materials. The T_g of nonplasticized aromatic copolyesters is ca. 70 °C, and that of copolyesters formed from cycloaliphatic glycols exceeds 100 °C. Aliphatic polyesters with long methylene chains between the ester groups have T_g values as low as –100 °C (limiting value, polyethylene –125 °C).

Solubility, Crystallinity, and Compatibility. Polyester solubility is largely determined by the nature and proportions of the constituent monomers. Polyesters with a regular structure are crystalline. Examples of highly crystalline polyesters are poly(ethylene terephthalate) and poly(butylene terephthalate). Although medium to highly crystalline copolyesters are insoluble in paint solvents, they can be processed into adhesives and paints in the hot-melt process [98]. Slightly crystalline copolyesters are soluble, for example, in ketones and are mainly used for laminate adhesives.

A low molecular mass and a low T_g generally favor the compatibility of polyester resins with other binders (e.g., acrylic resins, epoxy resins, precondensed amino resins, cellulose esters). Not all polyester resins are mutually compatible; for example, polyesters based on phthalic acid are not always compatible with other polyester resins.

2.7.2. *Production of Polyester Resins and Paints*

The polycondensation of polycarboxylic acids and polyalcohols proceeds via esterification with the elimination of water [92]. Transesterification reactions should be taken into account in copolyester resins derived from more than two monomers. Esterification and transesterification can be accelerated with metal-based catalysts.

Low molecular mass linear and branched polyester resins are produced in a one-stage process at 125 – 240 °C. The volatile condensation products are removed in vacuo (melt condensation process) or by passing a stream of inert gas through the resin melt (gas stream condensation process). Polycondensation in solution with azeotropic removal of water by solvent distillation (azeotropic process) is of lesser importance. High molecular mass copolyesters are produced in two stages as is used for poly(ethylene terephthalate). A precondensate is first obtained by transesterification of dimethyl terephthalate with an excess of diols. In the second stage, the molecular mass of the precondensate is adjusted to the desired value by polycondensation in special reactors with the maximum possible elimination of water and excess diols in vacuo at ca. 250 °C.

Saturated paint polyester resins are soluble in esters, ketones, and in some cases aromatic hydrocarbons, but are insoluble in aliphatic hydrocarbons and alcohols. Solvent-containing paints based on high molecular mass resins have solids contents of 35 – 50 wt % at application viscosity. Paint solids contents for low molecular mass polyesters are 50 – 70 wt %.

Polyester resins for waterborne paints mostly contain anionic groups (e.g., carboxyl groups) and become water soluble after amine neutralization. Waterborne paints based on polyesters also contain cosolvents, preferably glycol ethers. Powder coatings based on low molecular mass resins are solvent-free, like radiation-curable polyester paints.

Conventional grinding methods are used to incorporate pigments into solventborne and waterborne polyester resins as well as into radiation-curable polyester paints containing reactive diluents. Pigmented powders are produced by coextrusion of polyester granulate with pigments, followed by grinding.

2.7.3. *Cross-Linking of Polyester Resins*

Amino Resins. Formaldehyde-modified amino resins (i.e., melamine, benzoguanamine, and urea resins, ⟶ Amino Resins) are the most important resins for the heat curing of hydroxyfunctional polyester resins. These resins are readily available with a low molecular mass (very good polyester compatibility, but less reactive) or in precondensed form (limited compatibility, but very reactive). In order to prevent premature reaction in the wet paint, the amino resins are blocked by etherification with, for example, methanol or butanol. Sulfonic acids (e.g., *p*-toluenesulfonic acid, dodecylbenzenesulfonic acid) have proved suitable for accelerating the deblocking of amino resins during heat curing. These acids must also be used in blocked form (ammonium salts, thermolabile adducts).

The polyester – amino resin ratio is generally in the range 90 : 10 to 70 : 30. High molecular mass polyester resins can be cross-linked with hexamethoxymethylmelamine resins (HMMM) above ca. 190 °C. Amino resins undergo selfcondensation above this temperature [93]. Low molecular mass linear polyesters react above 160 °C, and branched types above ca. 100 °C. The latter generally contain acid groups (acid number 5 – 25 mg KOH/g) in addition to the hydroxyl groups to enhance reactivity with amino resins. Stoving enamels based on high molecular mass polyester resins and HMMM or benzoguanamine resins have excellent flexibility and surface hardness as well as outstanding adhesion to metal substrates [99], [100]. Low molecular mass linear polyesters combined with HMMM resins have lower flexibility, but have the advantages of higher reactivity, better pigmentability, and higher paint solids content [101], [102]. Low molecular mass branched resins form particularly reactive stoving enamels, but the resulting films are less flexible and do not adhere as well to metal substrates.

Polyisocyanate Resins. Hydroxy-functional polyester resins play a significant role in two-pack polyurethane paints. Combinations with nonblocked polyisocyanate resins are discussed in Section Polyurethane Coatings .

Storage-stable one-pack polyester stoving enamels can be formulated with polyisocyanate resins that are thermoreversibly blocked. Tin catalysts (e.g., dibutyl tin dilaurate) are particularly suitable for accelerating hardening with hydroxypolyester resins.

Although linear polyester resins do not readily undergo heat curing, they form extremely flexible paint films with excellent adhesion to metal substrates. Polyisocyanate resins are more resistant to hydrolysis than amino resins. Polyester – polyurethane paints therefore exhibit improved resistance to moisture and weathering but are more expensive.

Epoxy Hardeners. The heat curing of carboxy-functional polyester resins with epoxy resins is mainly used in powder coatings. Catalysts (e.g., benzyltrimethylammonium chloride or 2-methylimidazole) are used to accelerate the reaction during heat curing. Triglycidyl isocyanurate plays a special role as cross-linking agent. Powder coatings with outstanding weather resistance and very good mechanical properties are obtained [95]. Hybrid powder coatings based on epoxy resins of the bisphenol A type should also be mentioned.

Radiation Curing. The cross-linking of acrylate-modified polyester resins with UV radiation or an electron beam is an energy-saving alternative to the heat-curing systems. Reactive diluents (e.g., polyfunctional acrylates) are required to adjust the viscosity for application; although they increase reactivity, they reduce the flexibility and substrate adhesion of the paint film.

2.7.4. *Uses*

Coil Coating. On account of their outstanding hardness – elasticity balance, good adhesion to metals, good application properties, and favorable cost, paints based on saturated polyester resins represent the most important system for coil coating.

On account of their insensitivity to solar radiation, some polyester resins are particularly suitable for weather-resistant topcoats (e.g., facade claddings, blades of shutters and venetian blinds, vehicle claddings, metal signs). Melamine-formulated paints are economic and have a well-balanced property profile. Polyurethane-formulated paints have an even better processability and weather resistance (colorfastness, chalking resistance) [103].

Specially adapted polyester resins are available for topcoats for interior architectural use (e.g., underfloor ceilings, light fittings), for appliance claddings (e.g., domestic appliances, hifi equipment), or for automobile fittings. Hard, high molecular mass copolyesters are suitable for primers containing corrosion protection pigments on account of their low oxygen permeability and very good adhesion to metals [104].

Can Coating. Polyester stoving enamels are widely used for coating can exteriors (metal decorating) but also for protecting can interiors. Polyester resins are available that satisfy the FDA regulations for internal coatings of cans used for food packaging [105]. Polyester coatings do not affect the taste of the food.

High molecular mass polyester resins, cross-linked with benzoguanamine and melamine resins, are particularly suitable for highly drawable, sterilization-resistant decorative stamping enamels [99], [100]. Low molecular mass polyester resins also exhibit good processability and sterilization resistance in polyurethane formulations.

Automotive Finishes. Important industrial automotive topcoats are produced from alkyd resins. Two-coat metallic finishes consist of an acrylic clearcoat and a metallic basecoat. Solventborne basecoats are generally composed of cellulose acetobutyrate and a low molecular mass, branched polyester resin which exhibits high reactivity with precondensed melamine cross-linking resins.

Surfacers are mainly produced from low molecular mass polyester resins cross-linked with amino resins. Blocked polyisocyanate resins may also be used for high stone chip resistance. Water-soluble polyester resins are increasingly used for water-thinnable stoving surfacers. Spray and dip paints based on polyester resins are used to coat automobile accessories (e.g., windscreen wipers, axle parts).

Industrial Paints. Stoving enamels used to spray or dip-coat machinery, domestic appliances, vehicles, and office furniture are termed industrial paints. On account of their good adhesion and resistance to hydrolysis, polyester – melamine stoving enamels provide very good protection of metal surfaces against undercoat corrosion creep. Polyester resin grades are available for special mechanical requirements (high impact resistance, hardness, abrasion resistance) or for exterior use (weather resistance).

Polyester paints with high solids contents (65 – 75 wt % at application viscosity) can be produced from very low molecular mass resins [106]. Low-pollution paints can also be produced from water-soluble polyester resins [96]. On account of their good water solubility, HMMM resins are particularly suitable for cross-linking these resins. Organic cosolvents (mostly glycol ethers) must be added to waterborne polyester paints to control their viscosity and applicability (leveling, substrate wetting).

Two-Pack Polyurethane Paints. Polyester – polyurethane paints are described in detail in Section Polyurethane Coatings . Polyester polyols for polyurethane

systems are generally adjusted to be “softer” than those used for stoving enamels. Their hydroxyl number and degree of branching are adapted to the recommended polyisocyanate cross-linking agent.

Polyester resins combined with aliphatic polyisocyanate resins have outstanding weather resistance, good substrate adhesion, and high flexibility even under very large temperature changes (aircraft finishes).

Polyester – polyurethane paints are particularly suitable for coating plastics (maximum drying temperature generally 80 °C) on account of their very high flexibility and good adhesion.

Powder Coatings. Polyester resins are available for two different reactive powder coating systems:

1. Hydroxy-functional resins for combination with blocked polyisocyanate resins (based on isophorone diisocyanate) [107]
2. Carboxy-functional (acid) resins for combination with triglycidyl isocyanurate or epoxy resins

The resulting powder coatings have good weather resistance as well as excellent impact strength and adhesion to metals (even under humid conditions) and are therefore suitable for many uses (e.g., in exterior and interior architectural applications, for coating machinery, domestic appliances, steel furniture, garden tools).

Radiation-Curable Polyester Coatings. Radiation-curable polyester paints (see also Section Radiation-Curing Systems) are particularly suitable for coating flat surfaces (strips, boards, sheets) and for temperature-sensitive substrates (paper, plastics). Acrylate-modified polyester resins are used for UV printing inks, UV varnishes for paper and plastics, as well as for pigmentable, electron beam-curable gloss and matt paints [97]. Radiation-curable polyester paints are also suitable for metal coatings with good elasticity.

Special Uses. High molecular mass copolyester resins are used in the manufacture of flexible packaging. Terephthalate resins are particularly suitable as adhesion promoters for printing inks, lacquers, and adhesives on poly(ethylene terephthalate) films. Some polyester printing inks adhere directly to these sheets. Lacquers that can be heat sealed at relatively low temperatures can be produced from high molecular mass, soft copolyester resins. Special linear copolyester resins are used for magnetic tape coatings [108].

2.8. Unsaturated Polyester Coatings

Unsaturated polyester coatings are composed of unsaturated polyester resins (UP resins, ⟶ Polyester Resins, Unsaturated), hardening (curing) agents, accelerators, promoters, photoinitiators, stabilizers, covering agents (paraffin), air release and leveling agents, pigments, and extenders. Most of the coatings also contain copolymerizable vinyl monomers.

Although styrene is the most commonly used reactive monomer, esters of acrylic acid have also been introduced, particularly for UV curing applications. Monomer-free UP resins that are either dissolved in solvents (e.g., butyl acetate) or are water dispersible have also become important [109].

Unsaturated polyester systems are cured by free-radical polymerization. Free radicals are generated by use of peroxide (conventional curing) or addition of a photoinitiator and exposure to UV light (UV curing). Conventional curing can be accelerated by use of heat or IR radiation.

Due to the incorporation of reactive monomers, the coatings harden to form layers of practically unlimited thickness. Styrene initially acts as a solvent but then copolymerizes with the double bonds of the UP resin and is firmly incorporated into the film. Since only a small proportion of the styrene evaporates, excellent filling properties are obtained with these virtually “solvent-free” coating materials [110].

2.8.1. *Unsaturated Polyester Binders*

Unsaturated polyester resins are linear, soluble polycondensation products of maleic or fumaric acid and dihydric alcohols, such as ethylene glycol and 1,2-propylene glycol. In special applications, a portion of the α, β-unsaturated dicarboxylic acids can be replaced by phthalic and/or adipic acid. Unsaturated polycondensation products of maleic acid or maleic anhydride and glycols as well as their copolymerization with styrene are described in [111]. The testing of UP resins is regulated by DIN 53 184.

Worldwide UP resin production in 1988 including captive use amounts was estimated to be ca. 200 000 t. Western Europe accounts for 65 000 t, of which ca. two thirds of this is used for furniture coatings and the rest for filler and putty production.

Commercial products of UP resins are marketed under the following trade names [112]: Alpolit (Hoechst); Crystic (Scott Bader); Distitron (Alusuisse); Estratil (Rio Rodano); Gohselac (Nippon Gohsei); Ludopal (BASF); Poloral; Verton (Galstaff); Polylite (Reichhold); Roskydal (Bayer); Silmar (Silmar); Synolite (DSM); Unidic (Dainippon Ink); Vestopal (Hüls); Viapal (Vianova).

Paraffin-Containing UP Resins. Since atmospheric oxygen inhibits the copolymerization of many types of UP resins with monomers (chain termination), the film surface remains tacky. This can be prevented by adding a small amount of paraffin to the coating formulation [113]. The paraffin is initially soluble in the UP coating but its solubility decreases as polymerization progresses. It accumulates on the film surface to form a layer that prevents access of oxygen. The paraffin layer also reduces evaporation loss of styrene to < 5 % [114]. This is particularly important in view of the low maximum workplace concentrations allowed for styrene.

Unsaturated Polyester Resins for Fillers and Putties. Nontacky film surfaces are also obtained if phthalic anhydride is replaced by tetrahydrophthalic anhydride [115]. Use of *endo*-methyltetrahydrophthalic anhydride (cycloaddition product of maleic acid and cyclopentadiene) also improves air drying [116]. Both classes of products have become important, especially for the production of fillers and putties for autobody repair.

Air-Drying Unsaturated Polyester Resins. Use of UP resins modified with α, β-unsaturated ethers give coatings that surface-harden completely in the presence of atmospheric oxygen without the incorporation of paraffin wax [117-119]. The resulting films therefore have a very high gloss. Examples of ether reagents are diallyl ethers of glycerol or trimethylolpropane, as well as di- and triallyl ethers of pentaerythritol. Since they are cured to give a high gloss in the presence of oxygen, these materials are referred to as air-drying or gloss UP resins.

2.8.2. *Other Raw Materials*

The composition of the coating material should be adjusted in accordance with the intended application as regards choice of resin, curing system, and additives that modify the color, leveling, mechanical properties, and chemical resistance.

Hardeners. Polymerization is initiated by peroxide radicals. Typical examples are cyclohexanone and methyl ethyl ketone peroxides (hydroperoxides), benzoyl peroxide, perbenzoates, and peroctoates (acyl peroxides). About 4 % (relative to amount of resin and styrene) of a desensitized peroxide (50 % peroxide in plasticizer) is normally used. To avoid potential hazards manufacturers' instructions should be carefully observed when handling peroxides.

Accelerators are used to promote peroxide decomposition at room temperature. Hydroperoxides are split by heavy metal salts. Normally 0.02 – 0.05 % of cobalt as cobalt naphthenate or cobalt octoate dissolved in aromatic compounds is added to the system shortly before application.

Acyl peroxides are split by tertiary aromatic amines; manufacturers add up to 2 % amine to filler and putty resins (amine-preaccelerated UP resins).

Promoters. Substances such as acetylacetone, ethyl acetoacetate, or tertiary aromatic amines accelerate the curing reaction initiated by hydroperoxide and cobalt octoate. The promoter is generally at a level of 1 – 5 %, relative to the as-supplied resin form.

Photoinitiators. In UV curing, benzoin and benzil derivatives are added as photoinitiators in amounts of 1 – 3 %. These compounds are split into radicals by UV light, which in turn initiate polymerization [120-122].

Stabilizers. If stabilization is necessary (e.g., to prolong pot life or increase storage stability) 0.01 – 0.03 % of solid hydroquinone or *tert*-butylcatechol as a 1 – 5 % solution may be added.

Paraffin. The addition of paraffin prevents inhibition of polymerization by atmospheric oxygen [123] (see Section Unsaturated Polyester Binders). With higher

application temperatures (max. 40 °C) higher melting point paraffins are required (*mp* 40 – 70 °C). They are usually added at a level of ca. 0.1 % with respect to the coating, if possible as a 10 % solution in toluene. On cooling, the paraffin contained in coating materials and fillers may crystallize out. Paraffin grades that have been approved for UP resins should be used. The paraffin solution is normally added with the accelerator (cobalt compound) to the coatings material and fillers shortly before application.

Air-Release and Leveling Agents. On account of the many possible formulations and applications, the type and amount must be matched to the relevant system. The additive Baysilone Paint Additive PL (Bayer) is effective in many instances.

Solvents. Coating compositions based on UP resins are generally solvent-free. When inert or other coreacting binders are to be used, low-boiling solvents (e.g., ethyl acetate) are employed. With matt coatings containing nitrocellulose or cellulose acetobutyrate that are applied as thin layers, the solvent content may be as high as 75 %. In coatings applied as a thick layer the content should not exceed 4 %. Unsaturated polyester resins may also be combined with polyisocyanates to improve adhesion to the substrate and resistance to chemicals.

Pigments and Extenders. The *pigments* and *extenders* must neither accelerate nor inhibit the hardening process. The main function of the extenders (talc, barytes, calcite, dolomite) is to reduce the 7 – 10 % shrinkage that occurs on hardening and thus improve adhesion. Ready-formulated filler compositions should have an adequate storage stability.

In order to apply paints based on UP resins in thicker layers to vertical surfaces, the coating formulation has to be adjusted so as to be thixotropic. Thixotropic or standard resins are used for this purpose in combination with *thickeners* (e.g., highly dispersed silicic acid, bentonites, and/or hydrogenated castor oil derivatives). For *matt coatings*, waxes (polyethylene and polypropylene waxes) and/or silicic acid are added in amounts of up to 10 %.

2.8.3. *Production, Properties, and Uses*

Fillers and Putties for Wood. *Putties* are highly filled (resin to extender ratio 1: 2.5 to 1: 3), highly viscous compositions that are used to level out unevenness in the substrate (e.g., chipboard). *Fillers* contain less extender (resin to extender ratio 1: 1.5) and are generally used as an intermediate layer between the putty and the topcoat. Fillers and putties are produced with dissolvers, butterfly mixers, or stirrer-type ball mills.

Putties are applied with special roller coaters. Fillers are applied with reverse rollers (direction of roller rotation is opposite to that of the material to be coated), spray guns, or curtain coaters. The amount applied is 50 – 200 g/m^2 depending on the substrate. The high filler content of putties prevents inhibition of polymerization by atmospheric oxygen. The main binders for putties are paraffin-type UP resins that are UV cured in a paraffin-free formulation. Air-drying UP resins in combination with elastifying resins are mainly used for hydroperoxide – cobalt-hardening fillers.

UV curing takes place at belt speeds of ca. 3 m/min with a mercury vapor high-pressure lamp (80 W/cm). Conventional curing of sprayed fillers in circulating air ovens at 80 °C requires a hardening time of 10 min. These systems produce ideal primers for a wide range of topcoats.

Roller, Curtain, and Spray Coatings for Veneered Boards and Panels (Thin Layers). Veneered boards and panels that are to be coated in the open-pored form are first primed and sanded. The UV-curable UP *primers* are based on air-drying UP resins. UV-curable primers can be applied with rollers in amounts of 20 – 40 g/m^2. UV curing is performed at a belt speed of 3 m/min. UV-curable solvent-containing primers applied by spraying (60 – 70 g/m^2) have a solids content < 30 %; they are then subjected to aeration for 4 – 5 min at 50 °C to remove any remaining solvent, followed by UV curing at a belt speed of 3 m/min.

Cobalt – hydroperoxide-containing matt coatings for *topcoats and primers* can be left to stand for several hours due to their high solvent content of up to 75 %. This is an important prerequisite for their application. Solids settling is achieved by adding highly viscous nitrocellulose or cellulose acetobutyrate; it is desirable for economic reasons to obtain special surface effects (open pore effect in wood furniture). The coating materials are applied in amounts of 20 – 100 g/m^2 with a drying time of 6 h at room temperature. After evaporation of the solvent (4 – 5 min at 50 °C), the coating can be cured in 10 min at 80 °C in circulating air, in 1 – 2 min under IR radiation, or at a belt speed of 3 m/min under UV radiation. The coatings have a high scratch resistance, good adhesion, and high resistance to household chemicals.

Roller, Curtain, and Spray Coatings for Wood (Thick Layers). Coatings based on paraffin-type UP resins are mainly used as clear or pigmented thick coatings for wood, with or without matting agents. They are applied by roller, curtain, or spray coating. In special cases gloss polyesters are also used. After application and standing overnight, the coating is sanded and polished to give a stable high-gloss surface with excellent holdout and depth. Such systems are commonly used as substrates for topcoats based on polyurethane or other binders.

UV-curable UP coatings for both gloss and matt thick-layer varnishes have been used increasingly on account of their rapid drying properties.

Application. The following application methods are used industrially:

Roller Coating. Up to ca. 150 g/m^2 of UV-curable coating materials can be applied with reverse roller coating. These materials are preferred on account of their long application time (no pot life). A long application time is advantageous because it allows adjustment of viscosity.

Curtain Coating. The main problem in the application of UP materials is the fact that only a few minutes elapse between mixing and the onset of polymerization. This short pot life is circumvented with curtain coating by separating the cobalt and hydroperoxide used to initiate cross-linking. For example, in the active priming method [124] peroxide-containing nitrocellulose lacquers (which may also contain styrene-free UP resins) are first applied with rollers or a curtain coater. After evaporation of the solvent, a layer of cobalt-containing UP material is applied. Other, less widely used methods include the double head, sandwich, and reverse methods [125]. Since UV coating materials do not have a pot-life time (one-component systems) they are particularly suitable for curtain coating application.

Spraying. Pot-life-free UV coating materials can be applied using all spraying equipment. Two-component units are used for cobalt – hydroperoxide-containing paints with a short pot life.

Curing. Paraffin-containing coatings gel after ca. 10 min at room temperature, are touch-free after ca. 30 min, and are stackable after ca. 2 h. Air-drying UP coatings can be cured directly by heating without pregelling. UV coatings are cured after deaeration for 30 s (curtain coating) or 5 min (spraying) at a belt speed of 1.25 m/min. Coatings of up to 500 g/m^2 can be hardened by conventional curing (cobalt salt and hydroperoxide at ambient or elevated temperature) and up to 250 g/m^2 by UV curing.

Putties and Fillers for Metals and Marble. Knifing and spot-repair putties are particularly suitable when a high degree of unevenness has to be leveled in a few, rapidly successive operations. The putty resins (hard to elastic) are adjusted to a pasty consistency with extenders. The curing system consists of benzoyl peroxide and a tertiary aromatic amine.

The putty composition is thoroughly mixed with the benzoyl peroxide paste required for the hardening (2 – 3 %) and applied with a putty knife. The pot life is generally ca. 5 min; the hardening time is ca. 15 min. The dry putty can be sanded. Shrinkage of the hardened putty composition is 1 – 2 %.

In order to ensure that all pores are filled, a spraying filler and a nonsanding sealer must be applied to the putty. The spraying filler (based on air-drying UP resins with hydroperoxide and cobalt octoate) fills in the ghost lines. The nonsanding sealer (based on polyurethane) prevents bubble formation after topcoat application, particularly in a warm, damp atmosphere. The spraying filler and nonsanding sealer can be coated. Knifing putties and spraying sealers are used in automotive repair finishes and machinery coatings (on sanded or sandblasted iron).

The hardened putties exhibit very good adhesion, complete and even hardening in both thin and thick layers, and rapid hardening even at temperatures as low as 0 °C.

Special UP resins are used in stone putties and fillers and coatings for marble. The fillers are pasty to pourable, they are hardened with benzoyl peroxide and a tertiary aromatic amine. The fillers can be sanded and polished after 15 – 20 min. These hard, reactive coating compositions harden without discoloration and exhibit good mechanical properties: spraying or curtain-coating topcoats are produced from air-drying UP resins. Glossy surfaces are obtained after curing with cobalt and hydroperoxide.

2.8.4. *Transportation and Handling*

Unsaturated polyester resins cannot be stored for an unlimited period. If the resins are protected against light and heat, manufacturers generally guarantee a shelf life of 6 months (23 °C). Unsaturated polyester resins are delivered in sealed containers, drums, canisters, and road tankers made of stainless steel, aluminum, tin plate, or sheet iron that are sometimes internally coated. The flash point of UP resins is 28 – 35 °C, depending on the styrene content. Appropriate precautions should be observed to prevent electrostatic charging.

The transportation of UP resins is governed by the following provisions: GGVSee/IMDG code: class 3.3; UN no.: 1866; Medical First Aid Guide (MFAG): 310; Emergency Procedures (EmS): 3 05; GGVE/RID: Rn 301 D; GGVS/ADR: Rn 2301 D; ADNR: unclassified; ICAO/IATA-DGR: 3., 1866 III.

The toxicology of styrene (⟶ Styrene) should be considered while producing, handling, and applying coating compositions containing styrene-dissolved UP resins. Respirators should be worn to prevent any inhalation of paint aerosols during spraying application.

2.9. Polyurethane Coatings [126-129]

The term polyurethane paints (coatings) originally referred to paint systems that utilized the high reactivity of isocyanates for chemical hardening (curing). However, this term now includes a large variety of binders. The amount of polyurethane raw materials processed into coatings is steadily increasing, and in 1990 was estimated to be ca. 400 000 t worldwide.

Polyurethane paint films all have a polymeric structure with urethane, urea, biuret, or allophanate coupling groups. Coupling can occur during paint hardening (curing) as the result of polyaddition of relatively low molecular mass starting products. Alternatively the paints may already contain high molecular mass polymers synthesized by the coupling of appropriate monomers. High molecular mass adducts with excess isocyanate groups or adducts in which curing occurs via oxidation of conjugated double bonds are also common.

Other variants of technical interest include blocked reactive groups, that can be activated by heat or atmospheric moisture. Microencapsulation of polyisocyanates is also becoming an important deactivation method.

Depending on their chemical composition, polyurethane paints are formulated as twocomponent (two-pack) or one-component (one-pack) mixtures. They are applied from the liquid phase, which may be solvent-containing, aqueous, or solvent-free, or from the solid phase as a powder coating. They may be cured under a wide range of conditions: drying above 0 °C to stoving at ca. 200 °C.

2.9.1. *Raw Materials*

See also ⟶ Polyurethanes.

Polyisocyanates. Numerous diisocyanates may be considered as a basis for paint polyisocyanates. However, only toluene diisocyanate (TDI), hexamethylene diisocyanate (HDI), isophorone diisocyanate (IPDI), methylenediphenyl diisocyanate (MDI), and 1,1-methylenebis(4-isocyanato)cyclohexane (HMDI) are of commercial importance. (For a more detailed description of diisocyanates, see ⟶ Isocyanates, Organic.) High molecular mass polyisocyanates or prepolymers derived from such products are the basis of most polyurethane paint formulations. They contain only very small amounts (<0.5 %) of monomers that are volatile at room temperature and thus permit safe application from the point of view of industrial hygiene.

Polyisocyanates differ as regards their chemical structure, reactivity, functionality, and isocyanate content. They are the principal curing constituent of two-pack polyurethane paints. Curing agent solutions have an isocyanate content of 5 – 16 wt %. Solvent-free, liquid curing agents may contain up to 30 wt % isocyanate.

Aliphatic and cycloaliphatic polyisocyanates can be used to produce lightfast, weather-resistant coatings for extremely severe conditions. Aromatic polyisocyanates are of limited suitability for exterior applications and react more quickly than aliphatic polyisocyanates. The different reactivities of aromatic and aliphatic polyisocyanates influence their use because they affect drying and curing.

Polyisocyanate types with free, excess isocyanate groups include adducts of diisocyanates with polyols, isocyanurates formed by trimerization of diisocyanates, and high molecular mass products containing biuret or allophanate groups. Monomeric MDI and its mixtures with oligomers have a low vapor pressure and can be safely handled at room temperature.

Blocked polyisocyanates are obtained by chemical addition of compounds that contain acidic or potentially acidic hydrogen atoms (e.g., alcohols). These isocyanates are not reactive at room temperature since they do not contain free isocyanate groups. Such polyisocyanates can, however, be regenerated by heating to eliminate the blocking agent.

Commercial products include Coronate (Nippoly); Desmodur (Bayer/Mobay); Isocyanate (Hüls); Takenate (Takeda); and Tolonate (Rhône – Poulenc).

Polyols. In conventional two-pack polyurethane paints, high molecular mass compounds containing hydroxyl groups form the “base” that cross-links and hardens with polyisocyanates. Polyester polyols, hydroxyl-containing acrylic copolymers, and polyether polyols are particularly important. They may be linear or branched and their hydroxyl group content ranges from 0.5 to 12 %.

The structure and content of reactive groups determine the property profile of the polyurethane formed with the polyisocyanate. The higher the hydroxyl group content and the degree of branching, the better the hardness and solvent and chemical resistance of the paint films.

Commercial products include Desmophen (Bayer/Mobay); Lumitol (BASF); Macrynal (Hoechst); and Phtalon (Galstaff).

2.9.2. *Polyurethane Systems*

Polyurethane systems for paints and coatings are summarized in Table 5.

Table 5. Polyurethane systems for paints and coatings

System designation	Hardening principle	Application form	Film-forming temperature
One-Pack Systems			
1	Oxidation with atmospheric oxygen	solvent-containing	up to 80 °C
2	Reaction of NCO groups with atmospheric moisture	solvent-containing	ambient temperature
3	After unblocking reaction of NCO groups with OH or NH_2 groups	solvent-containing, solvent-free, or powder	110–200 °C
4	After activation of microencapsulated polyisocyanate with OH or NH_2 groups	solvent-free	100–160 °C
5	Physical evaporation	solvent-containing	up to 80 °C
6	Physical evaporation (if necessary additional cross-linking with melamine resin)	aqueous (waterborne)	ambient temperature (with melamine resin cross-linking at 140 °C)
7	Reaction of oxazolidine with atmospheric moisture and further reaction with NCO groups	solvent-containing	ambient temperature
Two-Pack Systems			
8	Reaction of NCO groups with OH groups	solvent-containing, solvent-free, or aqueous	up to 130 °C
9	Reaction of ketimines with atmospheric moisture and further reaction with NCO groups	solvent-containing	up to 130 °C

2.9.2.1. *One-Pack Systems*

Curing with Atmospheric Oxygen. Type 1 in Table 5. "Urethane oils" are produced by reacting diisocyanates (e.g., TDI or IPDI) with polyol-modified drying or semidrying oils. "Urethane alkyds" correspond in structure to alkyd resins, in which a proportion of the dicarboxylic acid is replaced by a diisocyanate. Urethane oils and alkyds do not contain free isocyanate groups. Drying and film formation occur oxidatively, as with alkyd resins (⟶ Alkyd Resins). Compared to the latter, however, they generally dry more quickly and the resultant films have better mechanical properties and a higher solvent resistance.

Curing with Atmospheric Moisture. Type 2 in Table 5. High molecular mass polyaddition products of polyols with excess diisocyanate contain reactive isocyanate groups. They are used to formulate one-pack polyurethane paints that cross-link with the formation of urea groups under the influence of atmospheric moisture to produce paint films with excellent resistance to chemical and mechanical attack. Solvent-free products and dissolved products with isocyanate contents of 5 – 15 % (based on the solid resin) are commercially available.

Since products that contain free isocyanate groups react with water, moisture may affect storage stability when producing pigmented coatings. This has to be prevented, and use of water-binding additives has proved effective (e.g., special monoisocyanates such as Additive TI (Bayer).

Blocked Polyisocyanate Systems. Type 3 in Table 5. The isocyanate groups of polyisocyanates can be "blocked" with agents such as phenol, butanone oxime, ε-caprolactam, or diethyl malonate that are easily eliminated or that rearrange under the action of heat. Mixtures of blocked polyisocyanates with polyols or polyamines are stable, and thus permit the production of storable one-pack stoving finishes. Stoving urethane resins containing both reaction components are commercially available. The stoving temperatures vary from 120 to 220 °C depending on the blocking agent, but can be lowered in certain cases by 10 – 20 °C using catalysts (e.g., dibutyl tin dilaurate).

The resulting films have a very high mechanical resistance. They are particularly important for stone impact protection in automobile finishes, for coating high-quality industrial goods, for electrical insulation, as well as for coil coating applications. Solvent-free products are used in liquid form as thick-layer systems or in solid form for powder coatings.

Microencapsulated Polyisocyanate Systems. Type 4 in Table 5. Some powdered diisocyanates or polyisocyanates can be deactivated when dispersed in a liquid solvent-free medium and protected by a suitable coating. Diamines or polyamines can, for example, be added to produce a stable polyurea layer that requires ca. 1 % of the isocyanate groups. Since microencapsulated polyisocyanates do not cross-link with reactive groups under normal conditions, they form stable one-pack systems at room temperature. At higher temperature (> 80 °C) the stabilizing coating melts and dissolves, allowing curing to occur. Such systems (e.g., those based on dimerized TDI) are used as underbody coatings in automobile construction.

Physically Drying Systems. Type 5 in Table 5. These systems are based on linear or slightly branched polyurethanes without reactive groups that are soluble in organic solvents; MDI and IPDI are usually used as starting isocyanates. On account of their high molecular mass these polyurethane resins generally yield highly viscous solutions with a relatively low solids content. The rapidly drying paints are extremely flexible and elastic and have a high resistance to mild solvents. Important uses include the coating of flexible substrates such as leather imitates and magnetic tapes.

Aqueous Systems. Type 6 in Table 5. Aqueous dispersions of polyurethanes that are slightly supported by ionic groups also undergo physical film formation. The binder consists of polymer chains that are coupled via urethane and urea groups, and contain basic or acid groups. Neutralization by salt formation provides the necessary hydrophilicity if the "self-emulsifying" properties resulting from incorporation of hydrophilic polyether radicals are insufficient.

Anionic dispersions based on IPDI, HMDI, and HDI are particularly important and are largely used for industrial applications (e.g., as glass fiber sizing or as a finish for leather and leather imitates). Chemical cross-linking is achieved by the use of water-soluble melamine resins. Stoving temperatures of ca. 140 °C are necessary.

Systems with Polyisocyanates and Blocked Reactants. Type 7 in Table 5. Instead of blocking the isocyanate groups, they can be combined with latent reactants (e.g., oxazolidines) to obtain storage-stable one-pack systems. After application, curing is achieved not by heat, but by atmospheric moisture which induces ring opening and the formation of reactive hydroxyl and amino groups.

2.9.2.2. *Two-Pack Systems*

Polyisocyanate and Polyhydroxyl Systems. Type 8 in Table 5. This is by far the most important system and comprises conventional polyurethane paint. The principal binder constituents (polyisocyanates and polyhydroxyl compounds) must be kept separate before application. In addition to polyols, many other hydroxyl-containing components can be used for the formulation (e.g., alkyd resins, epoxy resins, castor oil, and cellulose nitrate; see Section Raw Materials). The mixing ratio of base (hydroxyl component) to curing agent (polyisocyanate) is generally between 1: 1 and 10: 1. The two components are mixed homogeneously immediately before application. Industrially, mixing can be performed automatically in two-component spray equipment.

Two-pack polyurethane paints can be applied by all conventional methods apart from dipping. Curing takes place at ambient temperature, but can be accelerated if necessary by heat. Stoving is not necessary (energy saving).

The reaction between the isocyanate and the hydroxyl groups results in 100 % cross-linking if the reactants are converted in stoichiometric proportions. Optimum paint films are generally formed when NCO : OH equals 1. However, it may be advantageous to deviate from this ratio (under- or over-cross-linking). With aliphatic polyisocyanates, a catalyst (accelerator) is generally necessary. Small amounts (0.05 – 0.5 %) of tertiary amines or metal-containing compounds are normally added to the binder. Amine accelerators can also be added to the air stream used for spraying. Alternatively, the paint films may be cured in an amine-containing atmosphere.

Esters, ketones, and ether esters of polyurethane grade (i.e., absence of reactive constituents and a maximum water content of 0.05 %) are required for solvent-containing paints. Aromatic hydrocarbons are suitable as diluents. High-solids two-pack paints containing 30 – 40 wt % of volatile components are

becoming more popular on account of their low solvent emission.

Solvent-free two-pack systems require liquid reaction components and are normally based on MDI and its homologues. In contrast to solvent-containing paints, pigmented solvent-free two-pack systems require the use of water-binding agents (e.g., zeolites) to obtain thick, bubble-free coatings.

Aqueous two-pack polyurethane paints are obtained by combining, for example, hydroxylcontaining polyacrylate dispersions with water-emulsifiable, low-reactive aliphatic polyisocyanates.

Systems with Polyisocyanates and Blocked Reactants. Type 9 in Table 5. In special cases, compounds with blocked amine groups (e.g., ketimines) can be used as reaction partners for polyisocyanates. The chemical curing of these two-pack systems is initiated by humidity which causes regeneration of the reactive amine groups.

2.9.3. *Properties and Uses*

Polyurethane coatings have three main advantages: high mechanical resistance, outstanding chemical resistance, and (in the case of aliphatic polyisocyanates) excellent lightfastness and weather resistance. The properties of the paint systems which harden at ambient temperature are unsurpassed. They can also be combined and formulated in a large number of ways. Specific properties can be optimally adjusted even for extreme conditions.

On account of their broad range of properties, polyurethane paints are used industrially and on a trade scale in virtually all sectors. Practical applications range from paper coating to the protection of equipment in industrial plants. Major areas of use include transportation (large motor vehicles, rail vehicles, aircraft, automobile finishes and refinishes), the building sector (parquet floor coatings, outdoor coatings, floor coatings, sealing membranes), the industrial paints sector (equipment, machinery, furniture), as well as steel structural engineering and hydraulic steel structures. Polyurethane paints are particularly important in the surface treatment of plastics.

2.10. Epoxy Coatings

Epoxy resins are generally not used alone but require a reaction partner in order to be cured (hardened). A large number of reaction partners may be used for curing at elevated or at ambient temperature. The cured films have high adhesion, flexibility, hardness, abrasion resistance, resistance to chemicals, and corrosion protection. In 1995 ca. 700 000 t of epoxy resins was used worldwide by the coating industry (including civil-engineering applications).

2.10.1. *Epoxy Resin Types*

Bisphenol A Resins. Most epoxy resins are condensation products of bisphenol A and epichlorohydrin (⟶ Epoxy Resins). Depending on the ratio of bisphenol A to epichlorohydrin, resins of varying chain length are obtained, which differ in molecular mass, melting point, viscosity, solubility, and content of epoxy and hydroxyl groups. The low molecular mass types (M_r ca. 360 – 500) are liquid and the medium molecular mass types (M_r ca. 500 – 7000) are solid at room temperature. Low and medium molecular mass resins require reaction partners to give useful coatings. High molecular mass resins give useful, physically drying binders, mainly primers. These resins contain very few epoxy groups but many hydroxyl groups. The hydroxyl groups can be reacted at ambient temperature with polyisocyanates or at elevated temperature (180 – 300 °C) with amino resins or phenolic resins.

Bisphenol F Resins. Bisphenol F is produced by condensing phenol with formaldehyde in an acid medium (⟶ Epoxy Resins – Bisphenol F Epoxy Resin). Bisphenol F is a mixture of isomeric and oligomeric products (novolacs). Resins produced by reaction of bisphenol F with epichlorohydrin have lower viscosities and a somewhat higher functionality (more epoxy groups) than the corresponding bisphenol A resins. Like the liquid bisphenol A resins, they tend to crystallize. Mixing of bisphenol A and F resins prevents or reduces crystallization. On account of their slightly higher epoxy content, coatings based on bisphenol F resins have a somewhat higher solvent resistance than similarly formulated bisphenol A resins. Coatings based on bisphenol F yellow more than those based on bisphenol A.

Epoxy Novolacs. Epoxy novolacs are made by condensing formaldehyde with phenolic substances in an acid medium, followed by epoxidation (⟶ Epoxy Resins – Cresol Epoxy Novolacs). Only the epoxy phenol novolac and cresol novolac resins have gained importance in the market. Although bisphenol A novolac resins are also available, they have not succeeded in making a commercial breakthrough. All of these epoxy novolacs have a substantially higher functionality than bisphenol A resins, ranging from ca. 2.2 (phenol novolacs, e.g., D.E.N. 431, Dow Chemical) to ca. 5.8 (cresol novolacs, e.g., Araldite ECN 1299, Ciba-Geigy). The high functionality of these resins allows formulation of coating systems with a high solvent resistance (high network density). The aromatic structure of the resins is responsible for the high glass transition temperature and good resistance to aqueous and acid solutions of the coatings. The chemical structure and high functionality lead to coatings with limited flexibility and a somewhat lower adhesion to metallic substrates than bisphenol A epoxy resins.

Aliphatic Epoxy Compounds. Epoxidized alcohols are used as reactive diluents for epoxy resins. The reaction products of diols with epichlorohydrin (diglycidyl ethers of butanediol, polypropylene glycol etc., ⟶ Epoxy Resins – Monofunctional Glycidyl Ethers and Aliphatic Glycidyl Ethers) are also used as reactive plasticizers. Polyglycidyl ethers (e.g., of sorbitol and pentaerythritol) are commercially available. They have a higher viscosity than the aliphatic mono- and diglycidyl ethers. They show good solvent resistance when cured with cycloaliphatic amines. Such resins are also cured at elevated temperature with carboxy-functional polyesters, polyacrylates, or polyanhydrides to give coatings with good weather resistance.

Aliphatic epoxy resins generally have higher color stability and reactivity than aromatic epoxy resins, but their resistance to aqueous acid solutions is much lower.

Cycloaliphatic Epoxy Resins. Cycloaliphatic epoxy resins are produced by oxidizing olefins with peracids (⟶ Epoxy Resins – Cycloaliphatic Epoxy Resins and Epoxidized Vegetable Oils). They are of only minor importance in surface coatings. They are cured with anhydrides, carboxy-functional substances, or Lewis acids at 150 – 200 °C. Ultraviolet curing with triarylsulfonium salts and ferrocenes has become technically important. Coatings based on cycloaliphatic epoxy resins have a high gloss and good weather resistance. They adhere to metallic substrates better than the acrylate esters normally used for UV curing but are considerably more expensive.

Glycidyl Esters. Glycidyl ester resins were originally developed for electrical applications. Glycidyl esters of phthalic acid, hexahydrophthalic acid, terephthalic acid or trimellitic acid (e.g., Araldite PY 284, PT 910) cured with carboxy-functional polyesters or polyacrylates at elevated temperatures give coatings with both excellent color stability and outdoor resistance.

Heterocyclic Epoxy Compounds. The technically most important heterocyclic epoxy resin is triglycidyl isocyanurate, obtained by reacting cyanuric acid with epichlorohydrin (Araldite PT 810, Ciba-Geigy; Tepic, Nissan). This trifunctional epoxy resin is combined with carboxy-terminated polyesters to give weather-resistant powder coatings [133], [134], [136].

2.10.2. *Curing Agents*

The most important combinations of epoxy resins and curing agents are summarized in Table 6. Depending on their molecular mass, bisphenol A epoxy resins can be cured by polyaddition via their epoxy or hydroxyl groups. Polyamines, polythiols, and polyisocyanates are suitable for room temperature cure. Polyanhydrides, polyphenols, acids, and carboxy-functional polyesters are suitable for hot cure. Epoxy resins can also be cured by polycondensation with amino resins or phenolic resins. Epoxy resins can be polymerized with catalysts such as tertiary amines, boron trifluoride complexes, ferrocenes, and triarylsulfonium salts.

Table 6. The most important epoxy resin – curing agent combinations and their uses

Curing agent	Curing mechanism		Resin type*	M_r	Typical use
Amines and polyamidoamines	via epoxy group	curing at room temperature	liquid	400	two-component systems for heavy-duty corrosion protection and floorings building blocks for chemical modification
Polyamidoamines and amine adducts			solid, type 1	1 000	industrial maintenance and marine coatings
			solid, type 2		
Latent amines			solid, type 3		powder coatings
Polyesters					
Phenolic hardeners		hot curing	solid, type 4	2 000	powder coatings
Amino resins			solid, type 4	2 000	epoxy esters for industrial finishes and can coating
Phenolic resins			solid, type 6		can coatings and finishes
Amino resins			solid, type 7		
Polyanhydrides			solid, type 8	4 000	coil coatings
Phenolic resins	via hydroxyl group		solid, type 9		can coatings and finishes
Amino resins					
Isocyanates				5 000	two-pack polyurethane paints
Isocyanates			high molecular mass epoxy resins	10 000	primers (cold curing)
Phenolic resins					can and coil coatings
Amino resins					

* Types 1–9 indicate increasing molecular mass of the respective solid resins.

Polyamines. Aliphatic polyamines cure epoxy resins via their epoxy groups at ambient temperature. Excess amine is generally prereacted with the epoxy resin to form a polyamine adduct. Amines react with monomeric or dimeric fatty acids to form polyamidoamines. Polyamine adduct-cured epoxy resins have a high chemical resistance. Polyamidoamine-cured epoxy resins exhibit good adhesion and flexibility.

Cycloaliphatic polyamines are less reactive than aliphatic amines, and an accelerator must therefore be used for curing at room temperature (e.g., salicylic acid). Aromatic polyamines are even less reactive than cycloaliphatic amines. Therefore, accelerators must be used to cure aromatic amines with epoxy resins at room temperature [137].

Thiols. Thiols have an extremely unpleasant smell, and are only of industrial importance for adhesives and joint groutings. They give a high, permanent flexibility to epoxy resins.

Isocyanates. Whereas polyamines and thiols cure epoxy compounds via their epoxy groups, isocyanates cross-link high molecular mass epoxy resins via their hydroxyl groups to form polyurethanes. The reaction takes place at ambient temperature. These combinations cure more rapidly and at lower temperature than epoxy resins cured with polyamines.

Anhydrides. Polyanhydrides and not monoanhydrides must be used to cure epoxy surface coatings. They are used in powder form for powder coatings and in solution for can coatings; both forms are hot curing. The films have a good acid resistance, and do not impart an undesirable taste to foods.

Acids and Carboxy-Functional Polyesters. Curing of epoxy resins with acids and carboxy-functional polyesters requires heat. Industrially, these systems are most important in the formulation of powder coatings: more than 70 % of powder coatings are based on epoxy resins and carboxy-functional polyesters (see also Sections Production and Properties).

Polyphenols. Polyphenols react with epoxy resins on heating, but require an accelerator (e.g., tertiary amines, imidazoles). On account of their low color stability, powder coatings based on this combination are not used for decorative purposes, but exclusively for functional purposes where high thermal, mechanical, and chemical resistance is required (e.g., pipe coatings).

Amino Resins. Urea, melamine, and benzoguanamine resins react with high molecular mass epoxy resins on heating. Their hydroxymethyl groups react with one another and with the hydroxyl groups of the epoxy resins to form ether bonds. Such systems are used for coating domestic appliances as well as for packaging.

Phenolic Resins. Phenolic resins react with high molecular mass epoxy resins on heating. Such systems are used as “gold lacquers” to line food containers. The similarly synthesized bisphenol A resols produce colorless coatings with a higher chemical resistance, less odor during cure, and less alteration of the taste of food in contact with the coating. Flexibility is, however, lower than with standard phenolic resins.

Catalytically Curing Compounds. Although tertiary amines can polymerize epoxides even at room temperature, the degree of polymerization is too low to obtain useful coatings. Boron trifluoride complexes polymerize epoxides on heating. Such combinations are sometimes used in powder coatings if high solvent resistance is required.

Ultraviolet curing of epoxy resins has become important. Cycloaliphatic epoxy resins are combined with substances that induce polymerization under the influence of UV radiation or an electron beam. Industrially important products include triarylsulfonium salts (UVE 1014, 1016, General Electric) and arene – ferrocenium compounds (CG 24–061, Ciba-Geigy). To ensure good film flexibility and chemical resistance, polyols should be added. The films should be postcured for 30 – 120 s at ca. 100 °C [138].

2.10.3. *Chemically Modified Epoxy Resins*

Epoxy Resin Esters. Air- or oven-drying esters can be produced by esterification of epoxy resins (M_r 1000 – 2000) with fatty or oleoresinous acids (fatty acids of linseed oil, soybean oil, tall oil, coconut oil, or dehydrated castor oil).

Plasticized epoxy resins can be produced by reacting low molecular mass epoxy resins with dimeric fatty acids. These resins (e.g., Epikote 872, Shell) generally exhibit better substrate wetting than the unmodified epoxy resins. This reaction is, however, associated with the formation of ester bonds, and the resins therefore have a poor resistance to alkaline solutions. Instead of fatty acids, carboxy-functional polybutadieneacrylonitrile elastomers can also be used to flexibilize epoxy resins.

Esters of Epoxy Resins and Acrylic Acid. If low molecular mass liquid to semisolid epoxy resins are heated with acrylic acid at ca. 120 °C, epoxy acrylates are produced. These resins can be dissolved in low molecular mass acrylate esters and used as raw materials for UV-curable finishes. They are mainly used for paper and cardboard (record sleeves), but also for metals.

Esters of Epoxy Resin and Methacrylic Acid. Epoxy resins can also be reacted with methacrylic acid on heating. The resulting methacrylate ester resins are dissolved in polymerizable solvents and catalyzed with peroxides (e.g., Derakane, Dow Chemical). These catalyzed resins are applied as thick-layer laminates in combination with glass fibers, fabrics, or mats. The laminates are generally cured at ca. 60 °C and have extremely good chemical resistance; they are therefore used in the chemical industry for the internal lining of reactors and storage tanks [139-142].

2.10.4. *Uses*

2.10.4.1. *Curing at Ambient Temperature*

Solvent-Free Coatings. Solvent-free applicable binder systems which cure at ambient temperature are not only used for coating applications. By adding quartz sand to low-viscosity binder systems, floorings can be formulated. Depending on the chosen filler/binder mixing ratio, these systems are applied as self-levelling floorings or as mortar screeds. Adhesives and joint fillers for concrete applications are formulated in a similar way. For all these uses, only low molecular mass

liquid bisphenol A and bisphenol F epoxy resins (M_r 340 – 400) or their mixtures are suitable. While solvent-free coatings are applied in film thicknesses between 300 and 1000 µm, floorings are applied in 2 – 10 mm layers, depending on the quartz sand content and the application method used. Reactive diluents (cresyl and isooctyl glycidyl ethers, diglycidyl ethers of 1,4-butanediol or 1,6-hexanediol) are often added in amounts of 10 – 20 % to these resins. They react with the curing agent and lower the viscosity of the epoxy resin (1000 – 3000 mPa · s at 25 °C) to improve processing properties. Polyamines and modified polyamines are used as curing agents. Aliphatic polyamines are industrially very important because they are highly reactive and can cure epoxy resins at ambient temperature. Epoxy resin coatings that are cross-linked with aliphatic amines have the highest solvent resistance. Cycloaliphatic amines (Laromin C 252, C 260, BASF; Vestamin IPD, Hüls) have to be modified and accelerated to permit curing at room temperature. These amines give a very attractive appearance to the cured film and are mainly used for decorative wall and floor finishes. Aromatic amines (e.g., diaminodiphenylmethane) are solid, but can be used for cold curing by adduct formation and acceleration. Films based on such curing agents (Ancamin LT, Anchor; HY 830, HY 850, Ciba-Geigy) have a glasslike appearance and extremely good resistance to organic and inorganic acids.

Solvent-free epoxy systems must be applied within an extremely short time (30 – 60 min), which presents problems, particularly in hot countries. Such systems are also frequently applied with heatable two-component spraying equipment. Ketimines give a longer pot life. They are produced by condensing ketones and polyamines. When applied as a thin layer, they react with atmospheric moisture and are converted back into the ketone and polyamine. The ketone evaporates from the film and the amine cross-links the epoxy resin. Ketimines can only be used if a sufficiently high ambient temperature (> 20 °C) and high atmospheric moisture (> 75 % R. H.) can be guaranteed.

Solvent-Containing Coatings. Semisolid to solid bisphenol A resins (M_r 500 – 1500) are used to formulate solvent-containing paints. They are sometimes mixed with epoxy novolacs to increase the network density and thus the solvent resistance of the coating. Aromatic hydrocarbons, often mixed with alcohols, are generally used as solvents. The epoxy resins are cross-linked with modified aliphatic polyamines, in the form of in situ amine adducts or as isolated amine adducts dissolved in xylene and butanol. Polyamine adduct – epoxy resin combinations have an extremely good hardness and good resistance to aqueous alkaline solutions and organic solvents.

In practice, polyamidoamines are the most important curing agents for these types of epoxy resins. Polyamidoamines and polyamidoamine adducts with epoxy resins confer maximum flexibility and adhesion to epoxy coatings. Such systems are frequently used as primers on metallic and mineral substrates. As topcoats, epoxy paints are only used for interior applications since they tend to yellow, become matt, and exhibit chalking under the influence of light and weather. The cured films exhibit good resistance to aqueous, alkaline, and neutral solutions; their resistance to organic solvents is limited.

To reduce the use of organic solvents in paints, low-VOC (high-solids) coatings are becoming more and more important in both Europe and the USA. The solids content of these ambient-cured coatings can be increased by the partial replacement of the solid epoxy resins by liquid bisphenol A or bisphenol F epoxy resins. Some raw material suppliers offer modified bisphenol A resins (e.g., Epikote 874 X 90, Shell; Araldite XB 290, Ciba) and curing agents (Witco, various modified polyamidoamine-based hardeners), which allow the formulation of coatings with a low VOC content (< 300 g/L).

High molecular mass epoxy resins (M_r 3000 – 20 000) may be regarded as polyols, and are cross-linked with isocyanates to form extremely chemically resistant coatings (Desmodur N, L, Bayer; IPDI, Hüls). Epoxy resin – isocyanate combinations are also used to formulate rapidly curing shop primers.

Solvent-Containing, Air-Drying Epoxy Esters. In order to produce epoxy resin esters that can be cured at room temperature, solid epoxy resins (M_r 1500 – 2000; e.g., Araldite GT 6084, Ciba-Geigy; Epikote 1004, Shell) are reacted with air-drying fatty acids (e.g., linseed oil or soybean oil fatty acid). The area of application of these resins is very similar to that of air-drying alkyd resins, but they have a higher chemical resistance and improved flexibility. Normal driers (e.g., cobalt octoate) are used to accelerate drying.

Waterborne Coatings. Low molecular mass liquid bisphenol A and bisphenol F epoxy resins (M_r 340 – 400) and their mixtures are made water emulsifiable by either adding nonionic emulsifiers or epoxy-group-containing emulsifiers which take part in the curing reaction (Araldite PY 340–2, PY 3960, Ciba; Beckopox EP 147w, Vianova). Water-soluble polyamidoamines or modified polyimidazolines (Casamid 360, 362, Anchor former Akzo; Hardener HZ 340, Ciba) are used as curing agents. Such two-component systems are environmentally friendly and are mainly used for coating concrete and other mineral substrates as well as pretreated plastics. Up till now more than 80 % of waterborne binder systems are applied to these substrates, because waterborne epoxy systems adhere well to wet concrete surfaces and do not attack plastics like solvent-containg coatings do. Aqueous emulsions based on solid resins (M_r ca. 1000) containing ca. 10 % organic solvents are commercially available (Beckopox EP 384, Vianova; Araldite PZ 3961, Ciba). These emulsions are less reactive than those based on liquid bisphenol A epoxy resins. Therefore, they can be combined with water-soluble polyamine adducts (e.g., Beckopox Hardener EH 623, EH 613, Vianova; Hardener HZ 3980, HZ 3981, HZ 3982, Ciba; Hardener XE 36, Witco) so that even white coatings with good chemical resistance properties can be formulated. Due to the higher film-forming temperature compared to the liquid-resins-based systems, such binder systems should not be used for civil-engineering applications. They are mainly intended for the surface protection of metallic substrates. Coatings with improved flexibility and adhesion to steel can be obtained by the use of modified bisphenol A epoxy resin emulsions (e.g., Beckopox EP 385, Vianova; Araldite PZ 3962, Ciba) which allow the formulation of highly filled primers. Their corrosion protection properties are similar to primers formulated with solvent-containing epoxy resins.

2.10.4.2. *Curing at Elevated Temperature*

Powder Coatings. Powder coatings mainly contain solid bisphenol A resins with M_r between 1000 and 2000 (Araldite GT 6063, 6064, 7004, Ciba-Geigy; DER 662, 663, 664, Dow Chemical; Epikote 1055, Shell). Carboxy-functional polyesters are the most important curing agents (Grilesta, Emser Werke; Grylcoat, UCB; Uralac, DSM). Modified and accelerated dicyanodiamides and polyphenols, and in some cases polyanhydrides may also be used. For a detailed description of powder coatings, see Section Coating Powders.

Powders coatings based on aromatic epoxy resins have high chemical and mechanical resistance, but are not resistant to weathering. To obtain weather-resistant powders, triglycidyl isocyanurate (Araldite PT 810, Ciba-Geigy; Tepic, Nissan) must be used as the epoxy component with carboxy-functional polyesters.

Animal studies have shown that triglycidyl isocyanurate (TGIC) is a severe eye irritant, a mild skin and nasal irritant, and can cause genetic damage. This has led to concern over potential reproductive or carcinogenic effects and therefore alternative, less-toxic epoxy resins have been introduced. Most similar to the TGIC/carboxy-functional polyester powder coatings in their film properties are those based on glycidyl esters (e.g., Araldite PT 910, Ciba), which are also cured with carboxy-functional polyesters. These two systems dominate the architectural and exterior coatings sector since they combine good weather stability with good flexibility and excellent adhesion to steel and aluminum [134]. Latest developments by DSM combine epoxidized linseed oil with carboxy-functional polyesters, so that binder systems are available which do not need to be labelled [135].

Solvent-Containing Paints. Butylated urea – formaldehyde resins, butylated or methylated melamine – formaldehyde resins, and butylated benzoguanamine – formaldehyde resins react with the hydroxyl groups of epoxy resins (M_r 3000 – 10 000) at elevated temperature (160 – 200 °C) under elimination of alcohols. Ether bonds are formed that are largely responsible for the good chemical resistance of such paints. The coatings have outstanding adhesion, impact resistance, good flexibility, and high surface hardness. Epoxy resin – urea resin combinations are mainly used as primers or one-coat systems in the packaging sector, whereas combinations with melamine and benzoguanamine resins are mostly used as topcoats [143].

Phenol – formalydehyde resins (in some cases butylated) and to a small extent bisphenol A – formaldehyde resins have become important as curing agents for epoxy resins in can coatings. Curing occurs primarily via the hydroxyl group of the epoxy resin and the etherified methylol groups of the phenolic resin. The epoxy groups also react with the phenolic hydroxyl groups of the phenolic resin. The coatings are cured at ca. 200 °C and are extremely resistant to foods, cosmetics, and pharmaceuticals.

The use of carboxy-functional or anhydride-functional low molecular mass polyesters as curing agents is more recent. They are used to formulate white-pigmented can coatings. Blocked isocyanates (Desmodur AP stable, Bayer) cure high molecular mass epoxy resins on heating. The films have an extremely high hardness, good flexibility, and high resistance to chemicals.

Hot-Curing Epoxy Resin Esters. Solid epoxy resins with a molecular mass of ca. 2000 can be esterified with semidrying and/or nondrying fatty acids. These resins are used in combination with amino resins for stoving primers, can coatings, and industrial finishes. The importance of epoxy resin esters has, however, decreased in recent years.

Waterborne Paints and Coatings [144], [145]. For ecological reasons waterborne industrial paints and coatings have become extremely important. The epoxy

or hydroxyl groups of epoxy resins must be chemically modified to make them waterborne. Water-soluble epoxy paints can be produced by esterifying the resins (M_r 1000 – 2000) with a monocarboxylic fatty acid, followed by reaction with an anhydride. The resins are then dissolved in a water-soluble solvent (e.g., glycol ether or an alcohol) and made waterborne by neutralization with a monoamine. Such resins are cured with amino resins (e.g., hexamethoxymethylmelamine) at ca. 160 °C. The waterborne epoxy paints that are applied by anodic electrodeposition have a similar structure. Cathodically deposited coatings (as employed in the automotive industry) are generally produced by reacting low molecular mass epoxy resins (M_r ca. 1000) with a secondary amine, followed by neutralization with an organic acid. Blocked isocyanates are usually used for curing.

Waterborne coatings are increasingly used for can coatings. The binders are mainly based on epoxy resins. High molecular mass epoxy resins (M_r 3000 – 4000) can be made waterborne by graft polymerization with acrylates, methacrylates, and/or acrylic or methacrylic acid in the presence of styrene, and subsequent neutralization with amines. These waterborne polymers are normally cured with amino resins. All of the coatings exhibit good adhesion and flexibility and are mainly used for beverage cans.

2.10.4.3. *Radiation Curing*

Cycloaliphatic epoxy resins are normally used for cationic radiation curing (see also Section Electrodeposition Paints). A strong Lewis or Brønsted acid is formed after photolysis of the initiator. This acid opens the epoxide ring, thus initiating the polymerization reaction. The chain length is thereby increased. Aryldiazonium salts are no longer used as initiators due to their poor dark storage stability and unsatisfactory film quality. Diarylhalonium salts are only employed to a small extent. Triarylsulfonium salts are the most important initiators. The special advantages of cationic cure are the possibility of coreaction with substances containing reactive hydrogen atoms (e.g., for flexibilization), insensitivity to oxygen during cure, and low shrinkage.

The systems are color-stable, weather resistant, flexible, and also adhere well to tin-free and low-tin sheet metals, as used in the manufacture of beverage and aerosol cans. Although arene – ferrocenium complexes even allow the cure of bisphenol A epoxy resins, provided postcuring takes place, these systems have not yet become important in surface protection applications. The reasons for this are inadequate color stability and weather resistance as well as the intense color of these films.

2.10.4.4. *Occupational Health*

Liquid unmodified bisphenol A, A/F, and F resins produce little or no irritation of the skin or mucous membranes; they can, however, act as sensitizers and cause dermatological allergies.

Solid epoxy resins based on bisphenol A can be categorized as effectively nontoxic; they are nonirritating and scarcely sensitizing. The same is true for the novolac epoxy resins, although their liquid forms can have a sensitizing effect.

Reactive diluents and low molecular mass aliphatic epoxy resins of low viscosity show a detectable vapor pressure, and have to be handled with care. These substances produce medium to strong irritation of the skin and mucous membranes; they are also sensitizers and can cause skin disorders.

Epichlorohydrin, which is used for the synthesis of epoxy resins, is a suspected carcinogen in animals. Thanks to the well-developed manufacturing processes, practically all currently commercially available epoxy resins contain only the smallest traces of epichlorohydrin [146].

2.11. Silicone Coatings

Polysiloxanes or silicones have been used in surface coatings since the earliest days of the silicone industry [147-149]. They are versatile materials with an enormous range of properties and physical forms [150]. For a detailed description of silicones, see ⟶ Silicones. They have a very low surface energy, high thermal stability, are UV transparent, and have a low glass transition temperature (T_g). These properties are utilized by the paint industry to provide excellent water repellency and resistance to weathering, UV radiation, and thermal cycling. Silicones used in paints can be conventiently classified into four types: silicone resins, blending resins, silicone organic copolymers, and paint additives.

Silicone Resins. Silicone resins used in paint applications are usually produced from mixtures of di- and trifunctional organosilane monomers (e.g., phenyl- and methyltrichlorosilane, methylphenyldichlorosilane, and dimethyldichlorosilane) by controlled hydrolysis and condensation. Typically, mixed chlorosilanes in xylene are added to water; hydrolysis produces a layer of concentrated hydrochloric acid which is then removed. The remaining solution of siloxane oligomers (M_r 2000 – 4000) is neutralized with sodium hydrogencarbonate and washed with water. At this stage the low-viscosity resin contains 3 – 6 % silanol. Further condensation (bodying) and azeotropic distillation of water in the presence of catalysts such as zinc octanoate reduces the silanol level to 0.5 – 2 % and yields high-viscosity network resins (M_r 200 000 – 500 000). Alternatively, alkoxysilanes (most commonly methoxysilanes) may be used as starting materials in which case alcohol, usually methanol, is liberated during hydrolysis, bodying, and curing. The final properties of the resin can vary greatly depending on the nature of the substituents, the degree of substitution, and the extent of the condensation reaction and hence the content of residual reactive groups [148]. Products vary from hard, brittle monomethyl flake resins to soft, flexible phenyl-containing resins dissolved in a solvent. Resins with a higher phenyl content are more thermoplastic, have greater heat and oxidation resistance [151], longer shelf life, and tend to be more compatible. A high methyl content confers faster cure rates, flexibility, hardness, and better retention of properties at low temperature. The level and type of substituents selected depend on the application requirements. For example, silicone resins formulated with white or colored pigments can withstand continuous service temperatures of 260 – 370 °C, whereas aluminum-filled paints made with phenyl-containing silicone resins resist temperatures up to 700 °C by forming protective ceramic frits [152]. Silicone-resin-based paints are used in extreme environments where organic paints are ineffective [153], e.g., boiler stacks, exhausts, ovens, furnaces, heat exchangers, hot plates, grills, air-conditioning units, and jet engines.

Silicone Organic Blends. Pure silicone resins are too soft and thermoplastic for general paint application but are frequently blended with traditional organics such as alkyd, acrylic, epoxy, and phenolic resins. The higher levels of reactive groups in unbodied silicones provide sites for copolymerization on curing; the low molecular mass (2000 – 4000) ensures good compatibility which is essential for optimum film-forming properties. The organic component can improve the hardness and drying properties of the silicone while the silicone adds gloss and color retention at elevated temperature. These properties are employed in exterior maintenance paints. Improvement in durability is generally proportional to the silicone content of the blend. This in turn is determined by performance – cost considerations and usually lies in the range 30 – 50 %.

Silicone Organic Copolymer Resins. Simple blends of organic and silicone resins are unsuitable for many high-performance applications where mechanical properties and good weather resistance are of primary importance. They may become incompatible at the required levels and copolymerization is therefore preferred. At levels of 20 – 60 % the silicone copolymers impart considerable improvements over unmodified organics, particularly in heat stability, chalking resistance, and gloss retention [154]. The silicone intermediates are usually low molecular mass oligomers that are rich in reactive hydroxyl or alkoxy groups and have a high level of organic substitution. They are copolymerized with polyester, alkyd, and acrylic resins containing 80 – 250 % excess hydroxyl groups, usually in the presence of titanium catalysts [148]. If silanol-functional intermediates are used, the titanium catalysts promote copolymerization via – COH and – SiOH groups, thus eliminating the tendency for homopolymerization via SiOH groups. Choice of solvent and catalyst, concentration of reactants, and extent of reaction play an important role in the final composition and performance of the copolymer. Methoxy-functional silicone intermediates react with organic hydroxyl groups by elimination of methanol. Use of special catalysts, e.g., $Al(CH_3COCH\ COCH_3)_3$ and $Ti[OCH(CH_3)_2]_4$, allows better control of the condensation reaction and lessens the yellowing caused by many catalyst systems at elevated temperature [155]. Materials made in this way (e.g., silicone polyesters) can contain as much as 80 % silicone resin. Almost all types of organic coating have at some time been modified with silicone resins [156] although commercial success has been realized mainly with polyesters. A major user of silicone polyester paints is the coil coating industry where coatings containing > 50 % silicone have retained their gloss and color for more than ten years [153]. Other applications include appliance finishes, exterior cookware coatings, and colored maintenance and architectural paints.

Curing. Curing cycles are primarily determined by the silicone content of the resin [152]. A typical cure schedule for a 100 % silicone resin is 30 min at 260 °C. For 50 – 80 % silicones, satisfactory cures are obtained in 15 min at 220 °C. For silicone-modified organic coatings, complete cure can be achieved using ambient or low-bake (< 180 °C) cure conditions.

Silicone Paint Additives. See also Section Surface Additives. Because of their very high surface activity, modified silicone fluids and resins are widely used in paint formulations to achieve special desirable effects [157], [158]. The only silicone additive likely to cause surface finish defects is an overdose of polydimethylsiloxane. New siloxane organic copolymers are compatible with organics and overcome problems with fish eyes (oval surface defects), cratering, and recoatability. They perform at very low concentrations (generally 0.01 – 0.5 %) and are used to improve flowout, mar and scuff resistance, adhesion, leveling, and gloss. They can also control foaming [159], aid pigment dispersion, and impart special textured effects (e.g., hammer finishes). Silicone additives can be used in most paints including solvent-based, waterborne, high-solids, and powder coatings.

2.12. Urea, Benzoguanamine, and Melamine Resins for Coatings

Alkylated urea –, benzoguanamine –, and melamine – formaldehyde resins (⟶ Amino Resins) represent a versatile group of cross-linking agents for hydroxy-, amide-, and carboxy-functional polymers. They are used in both waterborne and solventborne coatings, including high-solids industrial coating systems. These amino resins are almost always employed in combination with other flexibilizing functional polymers (backbone polymers) to form highly cross-linked networks.

The major application for amino resins is in baking systems; there is a small market for acid-catalyzed wood coatings that are cured at room temperature. Excellent properties can be obtained in both the formulated coating and the cured film. Advantages include excellent exterior durability, adhesion to metal, gloss and color retention, relatively low cost, and excellent stability. Amino resins have found widespread use in automotive primers and topcoats, appliance coatings, metal decorating, and most general industrial applications. They are usually modified by etherification with alkyl ether groups; in the paint industry the term alkylated (methylated, butylated etc.) is commonly used for these modified resins.

The first amino resins used in coatings were partially butylated, polymeric urea – formaldehyde (UF) resins introduced in the late 1920s. Melamine – formaldehyde (MF) resins followed in 1935. In 1960 a range of fully alkylated, predominantly methylated MF resins were introduced. These products are more monomeric in nature and supplied as 100 % solids; hexakis(methoxymethyl)melamine (HMMM) is a typical example. Fully alkylated MF resins have gained a major market share in the United States accompanied by increasing penetration in Europe and Japan. Replacing a conventional partially alkylated amino resin with a fully methylated resin increases the solids content of a paint by 3 – 8 %. Reasons for their success include higher application solids, improved hardness – flexibility relationship, better exterior durability, improved reaction with functional groups on the backbone polymer, stability in waterborne coatings, reduced emission of formaldehyde, and high reactivity under controlled conditions. The incorporation of cross-linking agents of this type provides the formulator with the capability to utilize low molecular mass backbone polymers.

Properties. Alkylated amino resins can be classified into two general classes: (1) polymeric, partially alkylated resins which have a lower solids content and (2) the more monomeric, fully and partially alkylated products which have a higher solids content.

Typical *polymeric, partially alkylated amino resins* are butylated or isobutylated condensates with an average degree of polymerization between 3 and 8 and a combined formaldehyde content of 1.4 – 1.8 per amino group [160-162]. The hydroxymethyl (methylol) groups are partially alkylated (degree of alkylation 40 – 80 %).

The final resin has a solids content of ca. 50 – 65 %, and a combination of functional sites of different reactivity [161]. The solvents most commonly used are aromatic hydrocarbons and alcohols.

The higher reactivity of partially alkylated amino resins was long attributed to the presence of free hydroxymethyl groups. Under weakly acidic cure conditions, these hydroxymethyl groups demethylolate to form alkoxymethyl groups adjacent to secondary amide groups which respond more favorably to general acid catalysis [162-165].

Since 1970 *fully alkylated monomeric melamine resins* based on hexakis(methoxymethyl)melamine (HMMM) and higher-solids methylated melamine resins have replaced conventional partially butylated amino resins in many applications. They have also made significant inroads into the UF and the benzoguanamine market. Although these resins are designated monomeric, they are of oligomeric nature, with a degree of polymerization between 1.4 and 3.0. The amount of combined formaldehyde is about 1.7 – 1.9 per amino group. The degree of alkylation for fully alkylated products is 90 – 95 % of all available hydroxymethyl groups.

Fully alkylated amino resins require strong acid catalysis for fast and/or low-temperature cross-linking. Their catalysis mechanism is different from that of partially alkylated resins which respond to weak acid catalysts or general acid catalysis. A fully alkylated melamine resin catalyzed by a strong acid catalyst is a faster curing (cross-linking) agent than a partially butylated amino resin.

The catalysts which are predominately used are *p*-toluenesulfonic acid, dodecylbenzenesulfonic acid, and dinonylnaphthalenedisulfonic acid, and their amine salts [165]. Compared to partially alkylated amino resins, fully alkylated resins have a lower tendency to undergo self-condensation and produce films which are hard and still more flexible.

Production. Although the patent literature cites many other amino and amide compounds, only urea, melamine, benzoguanamine, acrylamide, and glycoluril have found a market position as starting materials for the production of amino resins. Formaldehyde is the only aldehyde used on a commercial scale. Methanol, butanol, and isobutanol are mainly employed as alkylating alcohols.

The manufacturing process for partially alkylated amino formaldehyde resins consist of two consecutive reactions. The first reaction is the methylolation (hydroxymethylation) of an amino compound such as urea, melamine, or benzoguanamine with formaldehyde and can be carried out under basic or acidic conditions. Hydroxymethylation under acidic conditions leads to simultaneous alkylation and polymerization of the amino resin. The second reaction, alkylation (or etherification), is carried out under acidic conditions.

Conventional low-solids butylated and isobutylated amino resins are normally prepared in this fashion [166]. The reaction water formed during the hydroxymethylation or alkylation step can be removed by azeotropic distillation. The acids used for catalysis may be organic (e.g., formic, oxalic, or phthalic acid) or weak inorganic acids (e.g., phosphoric acid). The end point of the alkylation reaction can be controlled by compatibility testing of the resin with low-polarity solvents. The hydroxymethylation catalyst remains in the resin either in free form or as the salt and may influence the reactivity of the resin.

Monomeric partially and fully alkylated amino resins are prepared in two separate reaction steps. Hydroxymethylation is carried out under basic conditions to minimize self-condensation of the amino resin. The pH is then lowered by addition of a mineral acid and alkylation with methanol is carried out. To obtain monomeric amino resins, an excess of formaldehyde has to be used to assure a high level of hydroxymethylation and a low level of residual amide groups that would otherwise lead to polymer formation during the alkylation step. Since water removal by azeotropic distillation is not possible with methanol, a large excess of alcohol is required to achieve complete alkylation. After completion of alkylation the resin solution is neutralized; water, alcohol, and residual formaldehyde are removed by vacuum distillation. The salt formed during neutralization is removed by filtration [167].

The fully alkylated amino resin is characterized by its degree of alkylation and polymerization, and its residual hydroxymethyl and amino content, all of which can significantly affect physical and chemical properties.

Environmental Protection. The potential release of formaldehyde during application and cure is a serious problem in the handling of amino resins. Depending on the structure and the conditions of manufacture a resin can contain 0.25 – 3 % free formaldehyde. The formaldehyde liberated from an amino resin may originate from three sources: unreacted formaldehyde, formaldehyde bound to the resin in the form of hydroxymethyl groups, or as a result of hydrolysis of the alkoxymethyl groups during cure and subsequent dehydroxymethylation [162].

Fully alkylated MF resins may contain < 0.25 % free formaldehyde. Partially alkylated resins usually have a higher content (≤ 3 %). The free formaldehyde content of an amino resin is largely responsible for the formaldehyde released during the application process. This is only a small fraction of the total formaldehyde released. Hydrolysis of the amino resin and subsequent dehydroxymethylation during curing account for > 80 % of the total formaldehyde release (see also Toxicology and Occupational Health.).

Although resins with a high amino content have a higher content of free formaldehyde than fully alkylated resins, they release lower amounts of formaldehyde during cure. The lowest amount of formaldehyde during application and cure is released by alkylated glycoluril – formaldehyde resins [168].

Quality and Specifications. A typical specification for an amino resin contains information on solids, viscosity, color, density, and free formaldehyde content.

A combination of analytical techniques has to be used to fully characterize an amino resin: molar ratios of amino compound to aldehyde and alcohol, molecular mass, molecular mass distribution, and functional group content are important data.

Amino resins are offered at a solids content of 50 – 100 %. For solutions a reliable nondestructive drying method is required before analysis [169]. Drying at < 50 °C in a thin film or under vacuum gives reproducible results without loss of resin components or degradation of the resin. This is especially critical for highly reactive resins. Solids methods which dry the resin at 100 – 110 °C result in dehydroxymethylation of the resin and reduced solids content particularly when applied to polymeric butylated types.

Total amino resin content can be determined in the solid residue by the Kjeldahl nitrogen method or by CHN analysis. In melamine resins titration of melamine with perchloric acid is possible. The total level of free and combined formaldehyde can be determined by phosphoric acid decomposition and by distillation.

Free formaldehyde is readily determined by titration in ice water with sodium sulfite solution. Titration at higher temperature results in dehydroxymethylation. An indirect method for determining the structure of an amino resin consists of CHN analysis and ^{1}H-NMR to determine the CH_2 : alkoxy ratio [170-172]. The hydroxymethyl content of amino resins can also be determined by reaction with iodine under basic conditions or by ^{13}C-NMR [173], [174].

Storage and Transportation. Amino resins should be stored in stainless steel or in lined carbon steel containers. The lining should be resistant to degradation or swelling by alcohols. Unlined steel vessels can be used under certain circumstances. The unlined vessel should have a desiccator on the gas-inlet line to remove water vapor from the gas that is used to replace the resin as it is pumped out of the tank. Methylated amino resins are preferably stored in lined drums or in fiber pack containers.

Amino resins should be stored at 20 – 27 °C. Lower temperatures can cause crystallization of hexakis(methoxymethyl)melamine and partially methylated melamine resins. Careful reheating is then required. Maximum storage temperature for fully alkylated melamine resins is 50 – 65 °C for 14 – 30 days. Amino resins with a lower degree of alkylation should not be stored at these temperatures. Fully alkylated amino resins are stable at 20 – 27 °C for more than 24 months. Butylated and partially methylated amino resins should not be stored for more than 6 – 12 months.

Since many amino resins contain solvents, local safety codes regarding flammable liquids should be observed. For bulk storage an inert gas atmosphere above the resin should be used [175].

Uses. The use of amino resins in packaging is regulated by the FDA under 21 CFR 175.105, adhesives; 21 CFR 175.300, resinous polymeric coatings; 21 CFR 176.170, paper and paperboard in contact with aqueous, fatty, and dry food.

Alkylated melamine –, urea –, and benzoguanamine – formaldehyde resins are the principal cross-linking agents in many industrially applied baked coatings. They are combined with acrylic, alkyd, epoxy, and polyester resins. The amide, hydroxyl, or carboxyl groups of these backbone polymers are used as functional sites for reaction with the amino resin.

The choice of amino component depends on the end use. Butylated UF resins are principally used in low-temperature curing applications, such as wood and general industrial coatings. Small amounts of these resins are also employed in container coatings in combination with epoxy resins. With the development of high-solids coatings, urea resins have been replaced in many applications with HMMM-type resins.

The market for benzoguanamine resins is small, and is mainly limited to container and appliance coatings because of its superior anticorrosion properties.

The largest market for amino resins is in the automotive topcoat area. Most heat-cured automotive primers, primer surfacers, basecoats, and clearcoats utilize melamine cross-linking agents. In the United States most of the amino resins used in high-solids automotive topcoats are monomeric, mixed ether, fully alkylated (methylated/butylated) MF resins. Outside the United States lower-solids coatings are predominantly used, which are cross-linked with conventional low-solids, partially butylated MF resins. The level of cross-linker is ca. 25 % in low-solids coatings and 35 – 40 % in high-solids coatings.

The cure temperature for most amino-resin-based automotive coatings is 120 – 150 °C. In end-of-line repair applications the same coatings are cured with addition of acid catalyst at 80 – 90 °C.

Amino resins are also employed in waterborne systems used in industrial coatings (e.g., automotive spray primers and basecoats, can and coil coatings). Stable coatings can be formulated with hexakis(methoxymethyl)melamine or partially methylated melamine resins.

Commercial products and suppliers of melamine resins are listed below:

United States		
American Cyanamid		Cymel/Beetle
Monsanto		Resimene
Europe		
Vianova	Austria	Viamine
Krems	Austria	Sakopal
Hoechst/Casella	FRG	Maprenal, Resamine
BASF	FRG	Luwipal, Plastopal
Hendricks & Sommer	FRG	Heso Amin
Synthese	Netherlands	Setamine
DSM	Netherlands	Uramex
Dyno Cyanamid	Norway	Dyanomin-Cymel, UFR
Ciba-Geigy	Switzerland	Cibamine
Reichhold	Switzerland	Super Beckamine, Beckamine
BIP	UK	Beetle
Monsanto	UK	Resimene
Japan		
Mitsui Cyanamid		Cymel
Mitsui Toatsu Chemical		U-Van
Dainippon Ink & Chemicals		Super Beckamine, Beckamine
Hitachi		Melan
Sanwa Chemical		Nickalac
Sumitomo Chemical		Suminal
Korea		
Monsanto – Kumbo		Resimene

Economic Aspects. The growth of the amino resin market is closely related to the overall growth of industrial coatings. The development of new coating technologies (e.g., powder coatings and radiation curing) may depress the future expansion of this market. However, this is counteracted to some extent by increased use of melamine resins in high-solids coatings which is greatly influenced by the global trend to less polluting coatings with a lower volatile organic content.

The world market for amino resins used in coatings in 1990 was estimated to be 200 000 – 225 000 t [176]. The U.S. market is about 50 000 t, with the European market estimated to be of similar size. The Japanese market is approximately 33 000 t. A breakdown of the U.S. market for amino resins used in coating applications is as follows (10^3 t):

Automotive	13.5
General industrial	8.5
Wood	4.5
Paper, film, and foil	5.8
Appliances	1.8
Coil	4.8
Container	4.8
Other	4.0
Total	47.7

In the United States high-solids amino resins (predominantly methylated resins with a solids content of > 80 %) enjoy a major market position. Both in Europe and Japan methylated resins are used only in applications such as waterborne coil and metal decorating, where superior performance is required. Conventional low-solids butylated amino resins are still the most important products of the industry in these countries.

Toxicology and Occupational Health. Most of the toxicologial and occupational problems in handling alkylated MF, UF, and BF resins are caused by formaldehyde release during application and cure [177]. Additionally, some release of formaldehyde can take place from cured objects.

The amount of formaldehyde released during the application process is related to the free formaldehyde content in the coating formulation. Loss of formaldehyde from hydroxymethyl groups during storage can raise the free formaldehyde content of a coating formulation.

The release of formaldehyde during the curing process is predominantly the result of dehydroxymethylation reactions. Some of the hydroxymethyl groups are formed during the curing process by hydrolysis of alkoxy groups from moisture in the air. For example, HMMM releases about 0.5 – 0.8 mol of formaldehyde during the cure process at normal humidity levels. High-imino melamine resins exhibit lower emissions of formaldehyde during cure. Formaldehyde release during cure can be almost eliminated with glycoluril – formaldehyde resin [168].

2.13. Phenolic Resins for Coatings [178-180]

Phenolic resins (i.e., condensation products of phenols and formaldehyde, ⟶ Phenolic Resins) are among the oldest synthetic binders, and their first use in paint technology dates back to the early 1920s. Their primary uses have constantly changed since then, and new classes of synthetic binders have become increasingly important.

Initially, phenolic resins attracted a great deal of interest because they appeared to be the first synthetic products that could be used as a substitute for natural resins (rosin, copal resins, shellac). Nowadays attention is mainly focussed on performance and technical-economic competition between the widely differing groups of resins.

The main disadvantage of phenolic resins is their intrinsic yellow to brown color. As a result they cannot be used for colored and white paints. They can only be used for decorative coatings in a few cases (e.g., gold lacquers). Phenolic resins have favorable mechanical properties and a high chemical resistance. Paint systems that are optimally adjusted to the requirement profile can be developed by suitable formulations.

2.13.1. *Resols*

Resols are phenolic aldehyde resins that undergo self-cross-linking catalyzed with bases or basic salts. Their structure depends on the choice and molar ratios of the raw materials (e.g., phenols and cresols), the solvents, and the type and amount of catalyst used. Resols have free hydroxymethyl groups which can react on heating to form homocondensates. Heterocondensation with other reaction partners is also possible.

Phenolic resins are generally pale yellow to dark brown in color, occasionally of high intensity. In special cases intensely colored resols are used as coloring resins to produce defined shades in coatings (gold lacquers). The products are available in solid (solvent-free) and dissolved form. The choice of solvents is governed by the intended application; resols are generally dissolved in alcohols, glycol ethers, aromatic compounds, or mixtures thereof.

Resols used for coatings generally have a relatively low melting point (ca. 50 °C), with the result that solid resins may sinter at high ambient temperatures (ca. 30 °C).

Resols suitable for use in paints generally cannot be used as the sole binder because they produce very hard, brittle paint films after stoving. They must therefore be combined with other paint raw materials [e.g., epoxy resins, poly(vinyl butyrals)].

Resols as Sole Binders. Resols are not normally used as the sole binder because of their brittleness. There are cases, however, where flexibility is not important. Uses of phenolic resins as sole binders are restricted to coating rigid constructions (e.g., pipelines and reaction vessels). Several coats are often applied to obtain the desired layer thicknesses. The first coats are cross-linked at relatively low temperature (ca. 170 °C) to prevent stress formation within the film. The overall film structure only undergoes complete cross-linking when stoving the last layer. Relatively high stoving temperatures up to 220 °C are required for optimum cross-linking.

Resols used alone as thermohardening; binders produce coatings with outstanding resistance to chemicals and solvents. Cross-linking takes place primarily via the free hydroxymethyl groups of the resols. The disadvantage of the low flexibility of these coatings has to be borne in mind.

Coatings for Sheet-Metal Cans and Containers. By far the most important area of use is the coating of sheet-metal containers and cans for the storage and preservation of food and other products. Thin sheets of galvanized steel (tin plate), nickel-plated steel (tin-free steel), or aluminum alloys are generally used as substrate. Untreated steel sheet may, however, also be used (e.g., to produce containers and drums for chemicals) and are usually coated with nonpigmented paint systems (varnishes).

In many cases the metal sheets must be coated before they are formed into containers. This demands extremely high flexibility of the cross-linked paint system. However, the coating material must also be extremely resistant to the container contents, must not discolor, and must, at least in the food-packaging sector, satisfy relevant legal provisions (Bundesgesundheitsamt in, FDA in the United States). Depending on the raw material, most as-supplied resols contain relatively large amounts of toxic substances (phenols, formaldehyde, etc.). They must therefore be properly labeled, stored, and handled. As a result of improvements in production methods and the choice of suitable raw materials, usable resols with greatly reduced levels of toxic substance have been developed and made available to the paint industry.

Resols are of limited use as sole binders on account of their brittleness; they are generally used in combination with plasticizing co-resins.

Combinations with Epoxy Resins. When resols are cross-linked with epoxy resins, the latter serve as plasticizing components. The hydroxymethyl groups of the resols are assumed to react with the hydroxyl groups of the epoxy resins to produce flexible coatings.

Epoxy resins with an epoxy equivalent weight of > 1500, a relatively high molecular mass (ca. 8000), and a high glass transition temperature (ca. 75 °C) are generally employed. Depending on the required flexibility and resistance properties, the resol – epoxy resin mixing ratio is generally between 15 : 85 and 40 : 60 (relative to weight of solid resin). The viscosity of the mixture is adjusted to the required value (governed by the method of application) with suitable solvents. The paints are generally provided with flow-improving additives (e.g., melamine resins) and lubricants (e.g., waxes). Typical application methods for this area of use include roller coating (normal coating, reverse roller coating), coil coating, and spray coating. In order to obtain the required cross-linking densities, the resol – epoxy resin paints must be stoved at relatively high temperature, normally > 180 °C. Production demands very short stoving times, a typical time being 12 min at 200 °C.

The general trend toward shorter production times and rationalization of the production process is reflected in the “shock drying process”, in which the paints are cured within a few seconds at up to 300 °C. The thickness of the stoved film is normally only 2 – 10 µm; optimum flow behavior is therefore a prerequisite for a satisfactory pinhole-free film.

Catalysts are used to accelerate hardening phosphoric acid and organic phosphates are generally used in practice. These compounds also improve adhesion of the resol – epoxy resin to the metallic substrate.

Resol – epoxy resin mixtures can also be precondensed by refluxing the prediluted components under an inert atmosphere for several hours at ca. 100 °C. Although a substantial increase in the mean molecular mass is not normally observed, mixtures pretreated in this way exhibit considerably better flow properties; the susceptibility to external flow interference (craters) is also reduced.

Waterborne systems based on resol – epoxy resin precondensates are already at an advanced development stage. Carboxyl groups are introduced into the preformed resols and made water soluble by salt formation with amines. These systems offer a significant saving in solvent compared with that of conventional high-solvent, high-viscosity products.

Combinations with Poly(Vinyl Butyrals). Poly(vinyl butyral) resins with a high acetal content are used, like epoxy resins, to plasticize resols. The weight ratio of resols to poly(vinyl butyrals) is generally between 95:5 and 70: 30. The employed application methods and curing conditions are similar to those used in the resol – epoxy resin systems. The resistance of such paint films to organic solvents, amines, and other chemicals are better than that of the corresponding combinations with epoxy resins. In contrast to the latter, resol – poly(vinyl butyral) combinations are often used as pigmented paints, but heat-resistant inert pigments have to be used. Certain shades (e.g., gold lacquers) can often be obtained more satisfactorily with such combinations on account of the darker intrinsic color produced on stoving.

Priming Coats. Resols in combination with poly(vinyl butyral) are used for acid-catalyzed priming coats that are cured at room temperature. These systems produce rapidly drying corrosion-resistant priming coats for metallic substrates. The mixing ratio of resol to poly(vinyl butyral) is generally 25:75 to 50:50. These priming coats (shop primers or wash primers) serve as a temporary corrosion protection in steel construction work. They also form an excellent adhesive layer for subsequent topcoats on substrates that are otherwise difficult to coat (e.g., aluminum or galvanized steel). A major use is therefore as a priming coat for rust protection in ships. Phosphoric acid is the preferred catalyst and, depending on the amount used, this coating is formulated as one-pack (acid and paint are premixed) or two-pack (acid and paint are supplied separately) systems. Corrosion-protection pigments are also employed; for ecological reasons zinc phosphates have replaced the previously used chromate-containing pigments. Highly volatile alcohols with small amounts of aromatic hydrocarbons as diluents are used as solvents.

Electrical Insulation Paints. Resols combined with poly(vinyl butyral), oil-free polyester resins, or poly(vinyl formal) as plasticizing components are used for this special application. The presence of small amounts of free phenols can increase the reactivity of these systems. These paints are used to coat wires, sheet metal for transformers and dynamos, as well as to impregnate electric motor windings. The paints are cross-linked at 100 – 300 °C.

Commercial products include Bakelite (Bakelite); Phenodur (Hoechst); Schenectady (Schenectady Chemicals); Uravar (DSM Resins); Varcum (Reichhold Chemicals).

2.13.2. *Novolacs*

If phenols and formaldehyde are reacted in the presence of acid catalysts, condensation products are obtained that do not contain reactive groups. These products can therefore neither react with themselves nor with other compounds and are termed novolacs.

Novolacs are solid at normal temperature and generally have a well-defined melting point. The largely inert behavior of novolacs means that they can only be used for physically drying paints and coatings. They are extremely soluble in polar solvents such as alcohols, ketones, and esters, but are insoluble in aromatic and aliphatic hydrocarbons (e.g., xylene, benzene). “Spirit varnishes” are novolacs dissolved in alcohol that are primarily used as unpigmented varnishes; they have an intrinsic yellow color. Their main uses include furniture polishes and rapidly drying coatings (e.g., for wood and toys) in industrial and consumer sectors. Novolacs have a high dilectric constant which means that they can also be used as electrical insulation paints.

The paint solutions dry rapidly and produce hard coatings that are largely resistant to water and hydrocarbons. Novolacs are readily compatible with many other binder classes. Addition of small amounts of alkyd resins or poly(vinyl butyral) resins improves the flexibility of the coatings.

Compared with phenolic resol resins, novolacs are of only minor importance as paint binders. Novolacs are, however, still very important as binders for basic dyes in printing inks.

Commercial products include Alnovol (Hoechst); Bakelite (Bakelite); Schenectady (Schenectady Chemicals).

2.13.3. *Modified Phenolic Resins*

Modified phenolic resins are condensation products of the resol type which contain other starting materials besides phenol and formaldehyde (e.g., acrylic monomers). The phenolic component itself is often modified (alkyl- or arylphenols). Modification with rosin is the most important. The compatibility of phenolic resins with other binders can be substantially improved by modification. Rosin-modified phenolic resins may be combined with linseed oils and alkyd resins. Examples of use include putties, priming coats, rust protection paints, and colored topcoats.

Only a limited amount of resin can be added to white paints due to its intrinsic color. Addition of these hard resins increases the hardness and the gloss of the paint films; accelerates and improves the drying of oxidatively cross-linking alkyd resins; and optimizes sanding properties and corrosion protection in putties. Modified phenolic resins have lost much of their importance because they have been replaced by more efficient binder systems (e.g., thermoplastic cross-linkable acrylic resins, polyurethane systems). Rosin-modified phenolic resins are, however, still extremely important in the production of printing inks.

Commercial products include Alresen, Durophen (Hoechst); Bakelite (Bakelite); Schenectady (Schenectady Chemicals); and Super Beckacite (Reichhold Chemicals).

2.14. Asphalt, Bitumen, and Pitch Coatings [181-184]

For a detailed description of asphalt and bitumen, see ⟶ Asphalt and Bitumen. Pitch is described in ⟶ Tar and Pitch.

2.14.1. *Asphalt and Asphalt Combination Coatings*

Composition. Only natural asphalts with small amounts of mineral constituents (asphaltites) can be used to produce paints. Of the asphaltites, only gilsonite with its moderately high softening point, particularly low mineral content (< 1 %), and high solubility in hydrocarbons is suitable. This natural product is a black, resinous, thermoplastic material with a relatively high molecular mass; it does not contain any carcinogenic constituents. Since gilsonite is very hard and brittle, it cannot be used as the sole film-forming substance and must be combined with resins and/or plasticizers. It can also be plasticized with softer bitumen or mineral oils.

Combination of gilsonite with drying vegetable oils (e.g., linseed or tung oil) produces high-grade asphalt combination paints, especially if the two binder components are heated (cooked) together.

Asphalt and asphalt combination paints are produced and applied only in dissolved form. Aromatic and aliphatic hydrocarbons of the normal boiling point range as common in the paint industry are suitable as solvents [e.g., mineral spirit (*bp* 140 – 200 °C), light solvent naphtha (*bp* 155 – 180 °C)]. Unfilled (unextended) asphalt or asphalt combination paints are rarely used nowadays.

Finely ground minerals such as talc (10 –20 wt %) and mica, or aluminum powder (1 –5 wt %) are usually incorporated into the asphalt to improve its film properties, in particular its mechanical, weather, and aging resistance. Antisettling agents such as montmorillonite – bentone (0.5 – 1.5 wt %) also have to be used.

On account of the relatively low inherent color of gilsonite, asphalt-based paints can be colored with inorganic pigments (e.g., iron oxide, chromium oxide, or aluminum).

Properties and Uses. Asphalt paints have good resistance to weathering, temperature fluctuation, condensation, aging, and also have good anticorrosion properties. They adhere well to metallic substrates, asphalt, and roofing felt.

Asphalt paints, particularly those based on combinations of gilsonite with drying oils, are used to protect exposed steel structures (e.g., bridges, lattice towers,

penstocks) and are recommended for these purposes in DIN 55 928. Since they are also heat resistant, paints of this type are used for protection against corrosion in coking plants and blast furnace plants, as well as to protect the reinforcement bars in aerated concrete. Soft-formulated asphalt paints can be obtained by blending with mineral oils, soft bitumens, or drying (fatty) oils. These are used to protect and seal roofs.

2.14.2. *Bitumen Coatings*

Composition. Bitumen paints contain bituminous substances as binders. Bitumen is defined in DIN 55 946 as “a low-volatile, darkly-colored mixture of organic substances obtained in the processing of mineral oils and petroleum, whose viscoelastic behavior changes with temperature. The typical properties of bitumen can be attributed to a colloidal system in which the disperse phase (asphaltenes) is stably distributed in a coherent phase of high-boiling oils (maltenes).”

Bitumens are thermoplastic, weather resistant, water resistant, absorb very little water, do not release any toxic substances in water, and are soluble in hydrocarbons.

Depending on the application, relatively hard distillation bitumen, oxidation bitumen, or high-vacuum bitumen (hard bitumen) are used for physically drying bitumen paints. Waterborne bitumen paints (bitumen emulsions) are produced from moderately hard distillation bitumen.

The properties of various types of bitumen should be taken into account when selecting the most suitable bitumen for a particular application. Oxidation bitumens have the widest range between the temperature at which they become brittle (fracture point, << 0 °C) and the softening point. Their plasticity range is between 90 and 130 °C. The relatively hard distillation bitumens have plasticity ranges of ca. 60 – 70 °C with fracture points around 0 °C. The high-vacuum bitumens have plasticity ranges of ca. 80 °C with fracture points above 0 °C. Distillation bitumens and high-vacuum bitumens (hard bitumens) are substantially more extensible in the plasticity range than oxidation bitumens.

The distillation and high-vacuum bitumens have a comparatively limited plasticity range and higher fracture point in comparison with the oxidation bitumens but this is compensated for by the following advantages. Distillation bitumens and high-vacuum bitumens require less solvent (less environmental contamination) to obtain the same paint viscosity and equal hardness. Their solutions are stable (long shelf life), whereas solutions of oxidation bitumens tend to thicken (gel). Solutions of distillation bitumen and high-vacuum bitumen penetrate porous substrates more readily. After drying, they adhere better to a wide range of substrates. Their weather resistance is better and they absorb less water. Since they are more stable to UV radiation than oxidation bitumens, they have a better aging resistance.

Petroleum fractions and aromatic hydrocarbons with a boiling point range between ca. 130 and 200 °C are suitable solvents for bitumen solutions. Benzines (mineral spirit) are preferred on account of their less obtrusive smell and the fact that they are more environmentally friendly. Use of chlorinated hydrocarbons should be avoided on account of their toxic properties.

Waterborne bitumens may be anionic, cationic, or nonionic emulsions, and are produced with the aid of emulsifiers and, if necessary, stabilizers conventionally used in the paint industry. The properties of bitumen paint (solutions and emulsions) (e.g., stability; mechanical, weather, water, and chemical resistance; corrosion protection properties) can be substantially improved by incorporating extenders. Suitable extenders include ground slate, ground limestone, chalk, talc, and mica (25 – 45 wt %, depending on extender type and required properties). The use of extenders requires the addition of antisettling agents. Long-term weather resistance can be considerably improved by incorporating soot.

Bitumen paints can only be colored to a slight extent (e.g., with iron oxide or aluminum powder) on account of their intrinsic color.

Properties and Uses. Thoroughly dried films of bitumen paints are resistant to water, salt solutions, dilute acids, and alkalis. Extended paints are weather resistant. On account of the thermoplasticity of bitumen, resistance decreases with increasing temperature. Bitumen paints are not resistant to organic solvents, fuels, oils, and greases. Optimally formulated bitumen paints are nontoxic and may therefore be used to protect surfaces that come into contact with drinking water.

On account of their favorable properties and comparatively low price, bitumen paints have been used to protect concrete structures (foundations, bridge abutments), felt roofs, sheet-metal roofs, drinking water reservoirs, silos, pipes made from fiber-reinforced cement, concrete, steel and cast iron, and in vehicle construction (wagon underframes, car underfloor protection). Unextended as well as extended bitumen paints are listed in DIN 55 928.

2.14.3. *Bitumen Combination Coatings*

The properties of bitumen paints (Section Bitumen Coatings) can be favorably modified and adjusted to suit practical requirements by combination with other film-forming substances. For example, the thermoplasticity can be reduced and/or mechanical properties (e.g., hardness, extensibility) can be improved by adding polymers such as polyethylene, polypropylene, polyisobutene, and styrene – butadiene copolymers. The chemical resistance can also be improved; high-quality corrosion protection coatings can be obtained by combination with alkyd resins.

Bitumen is permanently compatible with only a few polymers and only when the latter are added in small amounts. Compatibility depends largely on the nature and origin of the bitumen.

Combinations of bitumen with reaction-curing paint binders (e.g., epoxy resins and polyurethanes) are not usually employed in solvent-based paints due to their unsatisfactory compatibility. Only with high epoxy resin contents can a certain compatibility be achieved by addition of solubilizing phenols and aromatic oils.

2.14.4. *Pitch Coatings*

Composition. Special coal tar pitches are mainly used to produce one-pack, physically drying pitch paints. According to DIN 55 946, these are coal tar pitches whose “physical and chemical properties have been modified by special processes (e.g., polymerization)”. Nonmodified coal tar pitches are no longer used. Other pitches obtained from tar residues occurring in various processes (e.g., thermal treatment of pyrolytic oils, distillation of peat or lignite) are also used in pitch paints.

Both types of pitch consist mainly of polynuclear aromatic hydrocarbons. The pitches and pitch products used in the production of pitch paints are classified as carcinogenic materials (especially because of their benzo[*a*]pyrene content). Relevant regulations should therefore be observed in their production and use. It has, however, become possible to reduce the benzo[*a*]pyrene content of some pitches to such an extent (<0.005 wt %) that the “carcinogenic” warning label is no longer needed.

Nevertheless, pitches and pitch-based paints should not come into contact, even in the completely dried state, with drinking water, foods, or fodder. Pitches with softening points of 50 –100 °C are normally used for producing solvent-containing pitch paints. The content of sparingly soluble, high molecular mass constituents (free carbon) should be as low as possible to obtain solutions that are free of precipitates.

For waterborne coating materials based on pitch (DIN 55 946, Part 2, pitch emulsions or suspensions), the softening point of the pitch must be below 60 °C. High-boiling coal tar distillates, mineral oil extracts rich in aromatics, or plasticizing compounds normally used in the paint industry (e.g., benzyl phthalate) may be used as plasticizers for hard pitches. Aromatics boiling below 200 °C (e.g., xylene, Shellsol) are suitable as solvents. Petroleum spirits as well as most conventional paint solvents are unsuitable as solvents because they cause precipitation of the relatively sparingly soluble pitches and thus result in segregation. Waterborne pitch-based paints can be produced using special emulsifiers (e.g., bentonite).

Properties such as stability, mechanical and aging resistance of pitch paints can be improved by incorporating extenders as are used for bitumen paints (see Section Bitumen Coatings).

The tar crazing (alligatoring) that occurs during weathering of nonextended or unsatisfactorily extended pitch paints can be prevented by the optimum use of fillers. On account of the very dark color of pitch, it is virtually impossible to color pitch paints.

Properties and Uses. Nonextended and extended solventborne pitch paints can penetrate porous substrates very well and have excellent weather resistance. These paints and the waterborne pitch emulsions provide good corrosion protection properties and are resistant to water, salt solutions, dilute acids, alkalis, and microbiological degradation. They have only limited resistance to mineral oils, fuels, and greases. On account of the pronounced thermoplasticity of pitch (plasticity range 40 – 50 °C), the resistance of pitch paints decreases significantly with increasing temperature.

On account of their special properties and low cost, pitch paints have been used to protect concrete structures against chemically aggressive water (including effluent and wastewater) and for corrosion protection of steel constructions in industry, hydraulic steel structures, and underground pipelines made of steel, cast

iron, concrete, and fiber-reinforced cement. Pitch paints must not be used on objects that come into contact with drinking water, food, and fodder. Both nonextended and extended pitch paints are listed in DIN 55 928.

2.14.5. *Pitch Combination Coatings*

Composition. Pitch paints can be combined with thermoplastic polymers as well as with reaction-curing resins in order to improve their properties. Pitch paints are mixed with thermoplastic polymers such as poly(vinyl chloride), chlorinated rubber, polychloroprene, polyacrylonitrile, or polystyrene. The polymer content is generally low (ca. 4 – 8 wt %) on account of the limited compatibility and the impairment of paint application properties. Polymer addition reduces the marked thermoplasticity of pitches in one-pack, physically drying pitch combination paints. Mechanical properties (e.g., hardness, extensibility, resistance to mineral oils, and aging resistance) can also be improved.

Extended combinations of pitch paints (pitch + solvent), “tars” (pitch + high-boiling tar oils), and anthracene oil (high-boiling tar oils) with reaction-curing resins (e.g., epoxy resins and polyurethanes) have become extremely important. Combinations not classified as carcinogenic can be obtained by using pitches, “tars”, or anthracene oils with very low benzo[*a*]pyrene contents.

Combinations with reaction-curing resins are mainly supplied as two-pack paints, whose components are mixed together shortly before application. In combinations with polyurethanes, one-pack paints that cure by reaction with atmospheric moisture are also feasible.

Low-solvent and solvent-free paints can be formulated by using liquid film-forming substances (liquid “tars” or anthracene oils, liquid resins and curing agents). Very thick coatings [ca. 1 mm dft (dry film thickness) per coat] can be produced with epoxy resin combinations. Others can also be applied to moist substrates and cure under water.

In the cured state, solid epoxy resins (e.g., bisphenol A resins and polyaminoamides) are incompatible with the pitch or “tar”, and thus produce inhomogeneous films. Mixtures of pitch-, “tar”-, and anthracene-oil-containing film-forming substances with epoxy resins and polyisocyanates suffer from limited shelf life because their constituents react with one another. This increases viscosity and adversely affects film curing.

In combinations with epoxy resins, two-pack systems are used: one component contains the pitch-, “tar”-, or anthracene-oil-containing fraction and the curing agent (e.g., polyamine), and the other component contains the epoxy resin. In combinations with polyurethanes, one component contains the tar-derived fraction and the polyol, while the other component contains the polyisocyanate.

Properties and Uses. Pitch paints modified with thermoplastic polymers have better thermal properties in the dry state (cold and hot behavior) than conventional pitch paints (Section Pitch Coatings). They also have better cohesion, greater extensibility, and increased chemical resistance. On account of these properties, their good water resistance, and comparatively low cost these pitch paint combinations are used to protect construction works especially in the sewerage and effluent sector.

Optimally formulated combinations with reaction-curing resins have excellent water and weather resistance as well as a very good permanent resistance to acids, alkalis, mineral oils, fuels, and detergents. When cured they are no longer thermoplastic and are therefore resistant to heat and hot water, as well as being very abrasion resistant.

On account of their outstanding properties and relatively low cost, such combinations are widely used in the following areas: in hydraulic steel structures (hydro dams, lock gates, sheet pile boards); shipbuilding and harbor construction (ships' hulls, quay installations); sewerage sector (sewers, wastewater treatment plants); and pipeline construction (penstocks). They are listed in DIN 55 928.

Due to their smell and taste, all tar combination paints are unsuitable for residential buildings, stables, and surfaces that come into contact with drinking water, food, or fodder.

Commercial products include Eurolan (Deitermann); Inertol (Sika Chemie); and Prodorit (T.I.B.)

2.15. Silicate Coatings

2.15.1. *Water Glass Coatings*

The expression water glass paints is understood to mean coating systems based on the binder water glass (potassium, sodium, or occasionally, lithium water glass). The general expression silicate coatings has been adopted for these systems and is therefore used throughout Section Water Glass Coatings. For a detailed description of water glass, see ⟶ Silicates.

History. The history of silicate paints began with the rediscovery of potassium and sodium water glass by J. N. VON FUCHS around 1818 [185]. The use of these compounds as a flame retardant coating was first described in 1823 [186]. FUCHS, J. SCHLOTTHAUER, and W. V. KAULBACH subsequently developed the “stereochromy” process, a new painting technique in which water glass replaced the previously used binders lime and casein for mural painting. A further significant development was made by A. W. KEIM in 1878 [187], who created the two-component technique, still used nowadays, involving the binder (fixative) and pigments. The modern one-component dispersion silicate paints were developed and introduced to the market in 1967 (Silin silicate paint, van Baerle & Co.; Kieselit, Henkel).

Properties. The special properties of silicate paints are imparted by the binder, generally potassium water glass. The paint coats harden physically by evaporation of water, as well as chemically by several reactions. These stabilization and consolidation processes are termed “silicification” [189]. The water glass reacts as follows with atmospheric carbon dioxide:

$$\underset{\text{Water glass}}{m\,K_2O\cdot n\,SiO_2\cdot x\,H_2O} + m\,CO_2 \longrightarrow \underset{\text{Silica gel}}{n\,SiO_2\cdot x\,H_2O} + m\,K_2CO_3$$

Reaction also takes place with the pigments and extenders of the paint. If the paint is applied to a mineral substrate, a third reaction occurs, e.g., with calcium hydroxide:

$$\underset{\text{Water glass}}{m\,K_2O\cdot n\,SiO_2\cdot x\,H_2O} + Ca(OH)_2 + m\,CO_2 \longrightarrow \underset{\text{Metasilicate gel}}{CaSi_nO_{2n+1}\cdot(x+1)H_2O} + m\,K_2CO_3$$

The silica gel and metasilicate gel formed in these complex reactions contain a large amount of water, and are then converted into solid, amorphous silicate structures with only a low residual water content by evaporation of water [188]. Extremely finely divided potassium carbonate is formed as a hygroscopic byproduct which occurs as a transparent layer that is slowly dissolved by moisture. The sodium carbonate produced with sodium water glass leads to the formation of saltlike efflorescences. Potassium water glass is therefore mainly used for coatings.

Coatings formed from silicate paints differ from organic coating systems. Chemical bonds are formed by silicification in the paint layer and with the mineral substrate, and adhesion is therefore not purely physical. The paint layer has a silicate structure and is thus relatively permeable with a high degree of hardness. Evaporation of moisture is not hindered, peeling and flaking are therefore avoided. Since silicate paint layers have a high gas permeability they do not disturb the equilibrium with atmospheric carbon dioxide when applied to a lime plaster or mortar. Since the paint layer and the mineral substrate have a chemically similar structure they both exhibit the same expansion behavior under the action of heat and cold; thermal stress between the paint and substrate is therefore minimized.

The good resistance of silicate paints to weathering, environmental influences, and temperature is a result of their mineral, chemically resistant structure and the almost exclusive use of lightfast inorganic pigments. Silicate paints dry to give a matt surface.

Composition, Application, and Uses. Architectural and exterior coating systems may be either one- or two-component (DIN 18 363, Section General Properties). Two-component silicate paints may consist only of potassium water glass, pigments, and extenders that are resistant to potassium water glass.

One-component dispersion silicate paints may consist of potassium water glass, pigments that are resistant to potassium water glass, hydrophobing (waterproofing) additives, and no more than 5 wt % of organic constituents, relative to the total amount of the coating material. This total amount is measured by the ignition loss (Technischer Arbeitskreis Dispersionsfarben, TAKD, resolution, Feb. 1984) [190].

In two-component silicate paints, the potassium water glass binder (fixative) and the powdered pigments, extenders, and auxiliaries are mixed (kneaded) in a 1: 1 ratio by the user and homogenized after a few hours. The end product and reproducibility depend on the kneading time, degree of homogenization, temperature, and metering accuracy. On account of the high reactivity the paint should be applied within a few days of preparation. In contrast, one-component silicate paints are formulated ready for application in the paint factory [191], [192]. Depending on the paint manufacturer, water or a water – primer (fixative) mixture may be added by the user to optimize adaptation to the substrate. The permissible proportion of up to 5 wt % of organic constituents does not affect the advantageous properties of silicate paints. Indeed, they stabilize the system and improve water resistance. Storage stabilities of more than one year can be obtained and requirements with regard to the water uptake coefficient (DIN 18 558 E) are met [193].

On account of their properties, silicate paints are used as weather-resistant and wash-resistant coatings on mineral substrates in both exterior and interior applications. Substrates include lime, lime cement, cement renders (mortar groups I, II, and III), concrete, sandy limestone, as well as old silicate or mineral coatings. Old lime coats should be removed completely, and gypsum substrates should be treated only after special priming. A special water glass priming coat and one or two topcoats are required, depending on the substrate. On account of their silicate nature and the fact that they have the same color effect as historical lime paints, silicate paints are also used for restoring and maintaining historical monuments. Their high lightfastness, weather resistance, and water vapor permeability reduce the danger of coating damage even in structures and buildings that do not have adequate transverse insulation. Further special uses of silicate coating systems include corrosion protection (see Section Coating Systems for Corrosion Protection of Large Steel Constructions (Heavy-Duty Coatings)) and flame retardant coatings on wood [194], [195].

Commercial Products. Commercial one-component products include Calsilit (Kabe Karl Bubenhofer); Caparol – Sylitol (Deutsche Amphibolin Werke); Color-Silikatfarbe (Sikkens); Granital/Biosil (Keim Farben); Kieselit (Henkel); and Silin (van Baerle). Two-component products include Purkristallat (Keim Farben) and Silin (van Baerle).

2.15.2. *Alkyl Silicates [196]*

Alkyl silicates may be regarded as esters of the hypothetical silicic acid, which may exist in monomeric or oligomeric form. They are also referred to as silicate esters or alkoxysilanes. See also ⟶ Silicon Compounds, Organic.

Hydrolysis of alkyl silicates produces silicic acids with varying degrees of condensation and ultimately silicon dioxide. The reaction time increases with increasing size of the alkyl group. Elevated temperature and high moisture content accelerate hydrolysis and condensation.

The resultant silica gel can be used as a binder for solid particles, i.e., for pigments utilized in the paint industry. Inorganic coatings with special properties can be formulated. The silanol groups formed during hydrolysis make the alkyl silicate binders highly reactive in contrast to silica sols formed from colloidal silicic acid. They differ from water glass paints (Section Water Glass Coatings) in that alkali is not released during hydrolysis.

The properties of the hardened film (high surface hardness, temperature resistance, and electrical conductivity) account for their main use as binders for corrosion protection coatings containing zinc dust (see Section Coating Systems for Corrosion Protection of Large Steel Constructions (Heavy-Duty Coatings)). Depending on the temperature sensitivity of the pigment used, they can withstand temperatures up to 1000 °C. The absence of organic polymers makes the paints extremely resistant to organic solvents, and the high degree of hardness confers excellent abrasion resistance.

Composition and Properties. The binders are produced from orthosilicate esters $Si(OR)_4$ or industrial mixtures of polysilicate esters. Although the methyl esters are readily hydrolyzed and have a very high silicon content (generally given as SiO_2 content), the ethyl esters are normally used for physiological and toxicological reasons. The less readily hydrolyzable esters of longer-chain alcohols are employed in special applications.

Various polymeric silica distributions are obtained depending on the production process. Although this distribution may possibly affect hardening behavior, investigations have shown that this is insignificant in practical applications.

The use of alkyl silicates in coatings is based on the following reaction:

$$Si(OR)_4 + 2\,H_2O \rightarrow SiO_2 + 4\,ROH \text{ where } R = \text{alkyl}$$

This overall equation can be subdivided into two steps: hydrolysis

$$Si(OR)_4 + 4\,H_2O \rightarrow Si(OH)_4 + 4\,ROH$$

and condensation

$$Si(OH)_4 \rightarrow SiO_2 + 2\,H_2O$$

Both reactions occur in parallel, but their rates can be independently influenced. Condensation can be suppressed by lowering the pH value to obtain hydrolyzates that remain stable for relatively long periods (6 – 12 months) without gelling.

It has been suggested that this occurs because protons associate with the silanols which then repel one another on account of their positive charge and thereby oppose condensation. If the pH value is raised, these charges are dissipated and the silanols can condense [197].

The gelling time depends on a range of parameters [198], the most important being the SiO_2 concentration, the degree of hydrolysis, the type of solvent, and storage temperature. (The degree of hydrolysis is the degree of conversion, complete conversion to SiO_2 and ethanol corresponding to 100 % hydrolysis.)

In practice the ester is mixed with the required amount of water and a suitable solvent (e.g., alcohols, esters, ketones). To obtain 100 % hydrolysis of 1 mol of an orthosilicate, 2 mol of water must be added. By maintaining a constant pH (for ethyl silicate ca. 2) the condensation reaction can be suppressed to such an extent that hydrolyzates of ethyl silicate containing 20 % SiO_2 and with a degree of hydrolysis of 80 – 90 % remain stable for about one year. If this equilibrium is disturbed (e.g., by shifting the pH value or evaporating the solvent), condensation commences and the solution solidifies after a short time. This mechanism is used in two-pack zinc dust paints. In the simplest case the binder is a hydrolyzate in alcoholic solution. Zinc dust and, if necessary, extenders, antisettling agents, colorants, etc. are then added. Addition of zinc consumes acid and the pH value increases.

$$Zn + 2\,HCL \rightarrow ZnCl_2 + H_2$$

Condensation therefore commences and the mixture becomes solid.

On application, the paint layer hardens within a short time due to evaporation of the solvent, progressive hydrolysis of ester groups after uptake of atmospheric moisture, reaction of silanol groups with the zinc and substrate, and condensation of silanol groups to form sparingly soluble, cross-linked silica gel [199]. Conditions are generally chosen such that the film is hard enough to allow transportation after 4 h and a topcoat after 24 h. Alkyl silicate zinc dust paints are particularly suitable for applications involving high atmospheric humidity and high temperature because these conditions promote hardening. This explains their widespread use for corrosion protection in maritime tropical regions.

Historical Development. The importance and advantages of ethyl silicate zinc dust paints were recognized and intensively investigated in the late 1940s [200], [201]. The fundamental work of LOPATA and KEITHLER [202] on the partial acid hydrolysis of tetraethyl orthosilicate and the use of the product as a zinc dust paint binder was followed by attempts to improve the application properties and quality of the coatings (e.g., by adding poly(vinyl butyral) [203], alkyl titanates [204], borate esters [205], or trialkyl phosphates [206].

Ethyl silicate was transesterified with higher alcohols to produce binders with a higher flash point [207], [208].

In addition to the general disadvantages associated with a two-pack system, ethyl silicate hydrolyzate also ages. The condensation reaction is merely slowed

down and not stopped. As a result application properties such as viscosity and reactivity (and thus the hardening rate), alter during storage. The use of binders of varying ages can lead to considerable practical difficulties.

Continuous efforts have therefore been made to develop binders for one-pack paints which must be completely compatible with active pigments such as zinc dust. Mixtures of ethyl silicate and zinc can be stored for unlimited periods in the absence of water. If the hydrolysis and condensation reactions caused by atmospheric moisture can be sufficiently accelerated during application of the coating, a one-pack system can be produced.

Use of basic catalysts allows formulation of such systems because the reaction mechanism of hydrolysis and condensation proceeds differently in the alkaline range [197]; condensation occurs more rapidly.

The pronounced settling behavior due to the high density of zinc and the low viscosity of the alkyl silicate frequently resulted in solid precipitates and caused problems. Nowadays, however, one-pack zinc dust paints can be used even after several years because they contain special rheological additives (e.g., bentonite, pyrogenic silica). The application properties of these paints are similar to those of paint systems based on organic polymer binders. In addition to the general advantages associated with a one-pack system, these alkyl silicate paints do not exhibit age-dependent changes in their application properties (e.g., viscosity, brushability, and hardening behavior) because the degree of condensation does not alter.

All binders formed from ethyl silicate have a relatively low flash point due to the ethanol released during hydrolysis (flash point of ethanol 12 °C). The paint may have a somewhat higher flash point (15 °C) because its ethanol content is low. The use of higher alkyl groups yields systems with higher flash points.

Production. On account of the many possible variations in batch formulation and production procedure, only a general outline and instructions for producing one- and two-pack binders from ethyl silicate can be given here. The following batch formulation (in parts by weight) is typical for a two-pack binder containing 20 wt % SiO_2: ethyl polysilicate 500, solvents 440, and 1 % hydrochloric acid 60. Alcohols such as ethanol and 2-propanol, often mixed with one another and with ketones, ethers, and esters, are used as solvents.

The solvents, dilute hydrochloric acid, and ethyl silicate are mixed and stirred and the temperature rises; an upper limit of 25 – 30 °C is often recommended. The second component of the two-pack paint is the pigment – extender mixture. The binder can be used after 1 – 3 d; the storage stability is limited and is generally specified by the manufacturer. In the above example it is ca. one year.

Binders with a fairly high flash point require a suitable solvent (e.g., propanol, butanol) and a silicate base that cannot release any alcohols of low flash point during partial hydrolysis (e.g., tetrapropyl or tetrabutyl silicate). The slower evaporation of such alcohols delays the initial hardening of these paints.

The amount of water present determines the possible extent of hydrolysis of the Si–OR groups (in the above example ca. 82 %) and thus storage stability and reactivity. For example, the addition of ten additional parts by weight of water would reduce the storage time to < 25 % of the original value. Conversely, a lower amount of water would give a product with a longer storage stability which, however, would react more slowly possibly resulting in uneconomically long drying times. Suitable catalysts include hydrochloric acid, other acids, and salts (e.g., zinc chloride or amine hydrochlorides). Their concentration determines the rate of the hydrolysis, but also affects properties such as storage stability of the binder, pot life, and paint drying.

The production of binders for one-pack zinc dust paints is simple: alkyl silicate, catalyst, and solvent are mixed and diluted to the desired concentration. The resulting mixtures are extremely sensitive to moisture; even small amounts of water in the solvents can lead to gelling. Appropriate equipment and trained personnel are required to avoid mishaps. Interactions between the constituents can greatly influence application properties and lead to failure. The choice of an effective antisettling agent is difficult due to the large difference between the densities of zinc powder and alkyl silicate. Most rheological additives are ineffective in highly polar alkyl silicate solutions.

In order to produce ready-for-use zinc dust paint, zinc dust is slowly added to the binder in a ratio of 4 – 4.5: 1 while stirring. Binder additives (e.g., antisettling agents such as bentonite, pyrogenic silica) and zinc dust additives (e.g., extenders such as mica, talc, kaolin, or colored components such as chromium(III) oxide, titanium dioxide) can be used to adjust the application behavior and properties of the paints.

Commercial products containing ethyl silicate binders include Dynasil (Hüls); Ethylsilicate (Union Carbide); Ethylsilikat (Wacker); Silbond (Akzo); and Silester (Monsanto).

Important manufacturers of ethyl silicate zinc dust paints include Ameron, Carboline, Hempel's Marine Paints, International Paint, Jotun, Mobil Chemical, and Sigma Coatings.

Transportation and Storage. Containers and drums with a polyethylene insert or resistant lining are suitable. The binders should be stored at as low a temperature as possible and under absolute exclusion of moisture. The storage time of two-pack systems varies and is specified by the manufacturer. Commercial products are stable for about 9 months when stored at 20 °C.

Properties of the Coatings and Uses. Impurities in steel (e.g., carbon) can act as local cells and cause corrosion. The iron is first oxidized to divalent ions. A metal with a lower electrode potential (e.g., zinc) is, however, preferentially oxidized. The electrons that are thereby released counteract oxidation of the iron:

$$Zn^0 \rightarrow Zn^{2+} + 2e^-$$

$$2e^- + Fe^{2+} \rightarrow Fe^0$$

Theoretically, oxidation of iron can only begin when all the metallic zinc has been oxidized. This process can occur only if there is sufficient electrical conductivity (i.e., contact) between zinc and iron. Thorough cleaning of the metallic substrate is therefore important to ensure the effectiveness of corrosion protection coatings.

On account of their higher electrical conductivity, silicate esters have clear advantages over organic polymers as binders. The resistivity of a one-pack silicate ester zinc dust paint is ca. $1\times10^6\ \Omega \cdot cm$, i.e., two to three orders of magnitude lower than that of commercially available purely organic systems with a comparable zinc content.

The zinc oxidation products form a hard cementlike layer in the silicic acid film. Resistance to diffusion is increased and attack of atmospheric oxygen on the substrate is thus slowed down (barrier protection).

Ethyl silicate binders are mainly used in applications involving high temperatures and corrosive influences. The principal application of ethyl silicate zinc dust coatings is the protection of iron and steel against corrosion; it is effective even at temperatures of up to 400 °C. Examples include ships, harbor construction work, bridges, steel structures exposed to atmospheres in industrial areas, tanks, containers, pipelines, and chimneys.

Ethyl silicate zinc dust paints are applied by air or airless spraying, often directly after jet cleaning of the substrate. They are seldom applied with a brush or roller. Layer thicknesses range from 15 to 20 µm (shop primers) through 50 to 70 µm (primers) to > 100 µm (single thick-layer application).

The contents of two-pack paints must be constantly stirred after mixing to prevent settling. The pot lives are ca. 8 h. With one-pack paints, layer thicknesses of up to 200 µm can be obtained in a single application. Pot life is unlimited provided moisture is excluded.

Silicic acid is not attacked by radiation, and silicate esters are therefore particularly suitable binders for coatings in radiation-hazard areas such as nuclear power stations.

Ethyl silicate zinc dust paints can be covered with many conventional topcoat systems [209]. This is necessary in an acid or alkaline environment to prevent the zinc from being attacked. Ethyl silicate is also suitable as a binder for inorganic pigments on account of its UV resistance.

[Top of Page]

3. Paint Systems

Hans-Joachim Streitberger, Edmund Urbano, Richard Laible, Bernd D. Meyer, Engin Bagda, Frederick A. Waite, Michel Philips

3.1. Solventborne Paints

3.1.1. *General Information*

Paints and coating materials normally consist of a physical mixture of binders, pigments, extenders, additives, and solvents. Depending on the method of application and area of use, the solids content may reach 80 wt %, the proportion of pigment may reach 60 % of the solids content. The technologically most important component is the binder (or binder mixture). Binders may be classified as physically or chemically drying according to their film-forming mechanism.

Physically drying paints are solutions of thermoplastic polymers with molecular masses exceeding 20 000; on account of their low solubility they have a high solvent content (> 60 %) and a low solids content. Chemically drying paints have a fairly low solvent content (30 – 60 %) and a high solids content because the polymer network is formed by cross-linking of the binder (M_r ca. 800 – 10 000) to form thermosetting coatings. Oxidatively drying paints contain allyl groups and reactive double bonds, and cross-link by absorbing oxygen and forming ether bridges.

Normally paints are also classified according to the nature of the principal binder and its associated film properties; e.g., alkyd, acrylic, polyester, nitrocellulose, epoxy, and oil-based paints (see Chap. Types of Paints and Coatings (Binders)). The method of application, surface properties, and intended use are also utilized for classification [210]. Since the beginning of the 1980s environmental requirements have become increasingly important for two main reasons, especially in the case of paints with low material transfer (application) efficiencies (see Section Environmental Protection and Application Technology):

1. Avoidance of the use of toxic, carcinogenic, mutagenic, or teratogenic organic solvents
2. Drastic and in some cases legally imposed reduction of solvent contents (see also Section Environmental Protection and Application Technology)

Legal regulations have resulted in a sharp rise in the solids content and a reduction in the solvent content of the most important paint systems. Among solventborne paints, high-solids paints have enjoyed the largest growth rate; in the Federal Republic of Germany production of this type of paint rose by 32 % between 1988 and 1989, whereas the production of all solventborne paints increased by only 3 % [211]. A sharp drop in high-solvent paints is to be expected in the United States and Europe [212], [213]. The definitions of “low-” and “high-solids” paints differ. For example, the Verband der Deutschen Lackindustrie (Association of the German Paint Industry) defines “high-solids” paints as paints with a nonaqueous (i.e., organic) solvent content of less than 30 wt %, whereas the Environmental Protection Agency (EPA) in the United States defines “high-solids” paints as having volatile organic contents of < 2.8 pounds/gallon (336 g/L) [214]. High-solids paints are described in Section Solvent-Free and Low-Solvent (High-Solids) Paints.

Typical examples of low-solids paints (solvent content usually > 60 wt %) are:

1. Metallic(-effect) base paints for mass production of automobiles and touch-up finishes
2. Thermoplastic coatings
3. Coatings for the electronics and optoelectronics industries

Of these three groups metallic basecoats are quantitatively the most important. Technical reasons for the high solvent content of these paint systems include the need for thin layer thicknesses (e.g., when applying paint with a spin coater). Secondly, binders such as polyimides for electrical insulation coatings and thermoplastic acrylic resins have low solubilities. Finally, in some cases (e.g., in metallic basecoats) a high solvent content is needed to optimize rheological behavior [215]. Rheological factors limit the reduction of the solvent content, particularly when coating vertical surfaces. Significant improvements can often only be achieved by replacing organic solvents with water. For example, the solventborne metal effect paints used in the automobile industry are increasingly being replaced by waterborne paint systems with a reduction in the solvent content in the paint to ca. 10 % based on 25 % solids [216] (see also Section Solvent-Free and Low-Solvent (High-Solids) Paints).

3.1.2. *Properties and Raw Materials*

Solvents. Organic solvents are used to dissolve binders (resins) and additives and to disperse pigments and extenders in order to produce ready-for-use paints and coatings. For a detailed description, see ⟶ Solvents. The dissolution behavior (solvency) and interaction of organic solvents (e.g., ketones, glycol ethers, glycol ether esters, alcohols, esters, hydrocarbons, and terpenes) with binders have been investigated in detail [210]. The objectives of such studies are the optimum adjustment of processability to obtain the appropriate rheological behavior before and during application, and the maximization of the binder content in the paint formulation. The legally imposed requirement to employ toxicologically harmless solvents for the formulation of paints has stimulated the development of computer programs for optimizing solvent mixtures based on less harmful or innocuous solvents. This required a detailed knowledge of the interaction between binders and solvents [217]. Initial programs for calculating optimum solvent mixtures were developed by taking into account dispersive interactions (C_d), polar interactions (C_p), and interactions resulting from hydrogen bond formation (C_h) (solubility parameters) [218], [219]. Solvents were chosen in regard to toxicological factors [220] (see also Section Environmental Protection and Application Technology). Conventional tests for checking solubility are still used.

As well as the solubility, the solvent release during and after application is another important criterion for solvent selection. Evaporation of the solvent mixture from the film is important not only for a smooth surface, but also for obtaining optimal, reproducible mechanical properties of the film. The boiling points or vapor pressures of the solvents do not provide sufficient information for formulating a paint system [221]. A model for calculating solvent evaporation also includes transport data of the solvents in the film in the form of diffusion parameters [222]. Evaporation behavior is also determined experimentally (e.g., according to ASTM D 3539.76).

In solventborne paints, small amounts of the organic solvents may remain in the resultant film after cross-linking. Solvent retention can reduce internal tension, linear shrinkage behavior, and the glass transition temperature [223].

Film Properties. Physically drying solventborne paints contain cellulose esters, vinyl resins, thermoplastic acrylic resins, polyurethanes, rubber derivatives, and hydrocarbon resins as principal binders. Chemically drying solventborne paints that undergo oxidative drying are based on alkyd resins, epoxy esters, modified phenolic resins, and urethane oils. Cross-linkable paints play the most important role. They contain binders with unsaturated components (e.g., unsaturated polyester resins), ethoxy groups (e.g., epoxy resins), and hydroxyl groups (e.g., alkyd resins, polyesters, vinyl resins, and acrylic resins). A wide range of cross-linking agents are available for a variety of applications [220]. They include phenolic resins and amino resins. Isocyanates also play an important role as cross-linking agents. Hydroxyl-containing binders (e.g., polyester resins, acrylic resins) that can be cross-linked with polyfunctional isocyanates have been introduced as two-pack paints, particularly in the automobile and plastics industries [224], [225]. Advantages lie in the good weather resistance of the urethane groups, the low curing temperature, and the high gloss level of the coating.

Clear varnishes used in the automobile industry have to satisfy particularly stringent requirements: high gloss, color durability, and maintenance-free behavior for more than 5 years [226]. This is expressed in terms of permanently high water contact angles (> 90°) and a high gloss after several years' exposure of the paint in Florida. New acrylic resins have been developed that are copolymerized to about 50 % with fluorine-containing monomers such as chlorotrifluoroethylene [227]. Acrylic resins produced by group transfer polymerization represent a new class of resins with a narrow molecular mass distribution which is advantageous for dissolution [228].

New cross-linking reactions for curing paints and coatings have been developed, but their technical implementation has not been very successful. The Michael addition of C–H-azide compounds to activated unsaturated double bonds is used for automotive repair and touch-up finishes [229]. Esters activated with alkoxy and amide groups have also been described [230]. Epoxy–amine, epoxy – carboxy, and epoxy – anhydride reactions are gaining importance in aliphatic epoxy resin cross-linking agents [231], [232]. However, films produced in this way do not exhibit any major advantages over those produced from polyurethane resins. Much research and development has been devoted to solventborne paints that produce a multilayer system after a single application; this leads to considerable savings in material and time. The phase behavior of the binder components and solvents plays an important role in these paint systems [233], [234].

3.1.3. *Environmental Protection and Application Technology*

The paint transfer efficiency is defined as the amount of paint on the object expressed as a percentage of the amount of paint used for application. A low theoretical paint transfer efficiency means high amounts of solvent emission and waste generation. Special attention must therefore be given to environmental aspects when using conventional spraying methods for producing smooth, high-quality surfaces (see Table 7).

Table 7. Comparison of theoretical solvent emission data of various application methods*

Application method	A Paint solids content, wt %	B Organic solvent content, % of solids content	C Transfer efficiency, %	D Solvent emission, kg/kg film*
Pouring, roller application	70	43	100	0.43
Pneumatic spraying				
Metallic basecoat	13	670	60	11.20
Topcoat	50	80	60	1.33
High-speed rotating atomizers and electrostatic spraying				
Metallic basecoat	13	670	75	8.93
Two-pack varnish	60	67	75	0.85

*D = B/C.

The development of electrostatic spraying (see Section Electrostatic Atomization) has led to a sharp rise in paint yield and material transfer efficiency [235]. The use of high-speed rotating atomizers has also contributed to this and allows a very fine droplet spectrum to be produced. The droplet mist is directed onto the object by a guiding air jet; with a favorable charge to mass ratio, a high material transfer efficiency (> 75 %) is obtained. This high efficiency together with the evaporation losses of the solvent results in a source of emission that has led to the development of solvent and paint recovery units in plants with a particularly high throughput (e.g., in the automobile industry). Successful solvent recovery is of paramount importance in automobile factories [236]. Resins, solvents, additives, cross-linking agents, and other paint components are subjected to toxicological tests to determine any possible health-hazard effects before they are used in commercial formulations.

3.2. Solvent-Free and Low-Solvent (High-Solids) Paints

Solvent-free and low-solvent paints are a development of conventional high-solvent paints and are formulated to comply with increasingly stringent environmental requirements. In this section only those liquid coating materials are discussed that can be applied by conventional methods (coating, spraying, rolling, dipping, pouring). Solvent-free paints that are applied with new technologies and methods are described in Sections Coating Powders (powder coatings) and Radiation-Curing Systems (radiation-curing systems). Waterborne paints are discussed in Section Waterborne Paints.

Low-solvent paints are commonly referred to as *high-solids paints* and are defined as systems with a solids content exceeding 85 wt %. In practice, however, paints with a solids content of 60 – 80 wt % are also termed high-solids paints, particularly if the corresponding conventional (high-solvent) system normally contains < 50 wt % nonvolatile components [237].

Low-solvent paints have existed for a long time in the form of oil-based paints and varnishes (Section Oil-Based Coatings). The more recent development of high-solvent paints satisfying strict quality requirements has largely displaced this type of coating materials. Extensive development of high-solids paints occurred only after regulations governing solvent emissions and air pollution came into force (“Rule 66” in California, and TA Luft in the Federal Republic of Germany). There are several reasons for developing solvent-free and low-solvent paints:

1. Reduction of atmospheric pollution
2. Savings in materials and transportation costs
3. Savings in energy costs involved in paint production and use
4. Time savings due to higher layer thicknesses per application cycle
5. Improvement in safety due to low-solvent content

The advantage of high-solids paints compared with more recent technologies lies in the fact that they can be produced and applied with conventional equipment and by well-known methods. Consequently potential users are primarily small and medium-sized paint shops, the maintenance and architectural sector, as well as the do-it-yourself sector.

3.2.1. *Principles*

One way of lowering the solvent content of a paint is to reduce the viscosity during application by physical or chemical methods. For example, dilution of solvent-containing paints can be minimized by hot spraying. The solids content of solvent-containing paints can be increased by using solvents with solubility parameters that largely correspond to those of the binder, by adding viscosity-reducing cosolvents, or by using additives. For example, the formation of hydrogen bonds between hydroxyl and carboxyl groups increases the viscosity, which however can be overcome by adding small amounts of alcohols.

Solids contents exceeding 70 wt % cannot be achieved in this way—modification of the binder is necessary. Thus, the binders used in high-solids systems have a much lower intrinsic viscosity than conventional high-solvent paints. This is usually achieved by reducing the mean number-average molecular mass or by using a narrow molecular mass distribution (reduction of the weight-average molecular mass). The use of esterification and transesterification catalysts in polycondensation reactions can contribute to a saving in solvents in this manner. Lead and tin compounds have proved particularly suitable for this purpose, but are not appropriate for all applications on account of their toxicity. Production of the resin in solution followed by concentration to a higher solids content is a further possibility. On account of the low solvent content in the formulation further measures may, however, be required because the binder is also responsible for flowability, antisagging on vertical surfaces, prevention of wrinkle formation, and controlling the drying behavior. In high-solvent paints these properties are partially determined by the solvent. Evaporation of the solvent after application produces a sharp rise in viscosity which prevents the paint from sagging on vertical surfaces and accelerates the initial drying. Solvents also permit the use of high-viscosity binders with high glass transition temperatures which undergo physical drying. These advantages are largely lost in low-solvent paints, and development is therefore concentrated on the synthesis of special binders with the following features:

1. Reduction of the number-average molecular mass
2. Narrow molecular mass distribution (reduction of the weight-average molecular mass)
3. Chemical modification of the binders
4. Use of nonvolatile reactive diluents
5. Use of metal compounds as cross-linking agents

In addition to the absence of physical drying, low-solvent paints require a longer time for cross-linking due to the lower velocity of polymerization of the binder. This has to be compensated by higher unsaturation or a higher proportion of functional groups in thermosetting paints. When larger amounts of curing agents are used, however, a denser network than in conventional systems may be produced, resulting in better chemical and weather resistance.

As well as the mean molecular mass and molecular mass distribution, the number and nature of the functional groups as well as the degree of branching of the polymer molecules are important. Starlike branched polymers exhibit weaker intermolecular interactions on account of steric factors. The rigidity of polymers containing olefinic double bonds interferes with the association of the chains. Both effects lower the viscosity.

The term reactive diluents denotes low molecular mass reactants that act as solvents with an extremely low volatility and participate in the cross-linking reaction. Reactive diluents have become especially important in one-pack systems because they can partially or wholly replace the solvent in high-solvent

paints. Reactive diluents participate in the cross-linking reaction because their functional groups are identical or similar to those of the principal binder. They are completely incorporated into the polymer network during curing. Good compatibility with the principal binder in all cross-linking stages is necessary for continuous curing of the paint. If this is not the case, component segregation can occur, resulting in hazy films with reduced gloss. Resistance to chemicals and weathering as well as mechanical strength are also adversely affected. The most commonly used reactive diluents for oxidatively drying systems are unsaturated compounds. Polyfunctional methacrylate esters have replaced the more volatile monomers. Diacrylates or triacrylates of polyalcohols [238], dicyclopentenyloxyethyl methacrylate [239], [240], or their (co)polymers are recommended. Reaction products of dicyclopentadiene and butadiene also act as reactive diluents [241].

Low-viscosity metal compounds are also used in cross-linking reactions. For example, alcoholates of aluminum, titanium, and zirconium as well as their complexes with volatile ligands are primarily employed. The reaction involves substitution of the alcohol and formation of a coordination compound with the binder [242].

During evaporation of the solvent, sagging on vertical surfaces must be prevented by using thixotropic paints. These paints can be formulated with extenders and additives (e.g., smectite, montmorillonite, highly dispersed silica, hydrogenated castor oil, polyethylene waxes, and metallic soaps). These paint additives also act as antibleeding and antisettling agents. Modern urethane additives achieve the same effect without, however, affecting the gloss of the paint films [243]. A widely used alternative is thixotropic modification of the binder or combination of the binder with highly thixotropic resins (e.g., alkyd resins that have been reacted with polyamides).

3.2.2. *Production and Uses*

Low-solvent and solvent-free paints are not associated with specific classes of binders, and may be either one- or two-pack systems. In one-pack systems, hardening is effected by oxidative cross-linking with the aid of driers (see Section Alkyd Coatings) or by thermoreactive curing agents such as etherified amino resins or “blocked” polyisocyanates. Two-pack systems are mixed immediately prior to application, and curing occurs as the result of a polyaddition reaction between the binder and a curing agent. Such paints can be classified according to the nature of the raw materials and binders.

Alkyd Resins. The high versatility and broad range of applications of alkyd resins (Section Alkyd Coatings) have benefitted the development of high-solids paints. The commonly used one-pack paints contain high-oil, oxidatively drying alkyd resins, and have a solids content of 85 – 90 wt %. They are used as maintenance and architectural coatings. Unmodified alkyd resins require long drying times, tend to form wrinkles when applied in thick layers, and become yellow under the action of heat and in the dark, and they have therefore been superseded by modified alkyd resins. Low molecular mass alkyd resins with a narrow molecular mass distribution are reacted with modifiers such as trimethylolpropane diallyl ether to yield low-viscosity alkyd resins with improved drying properties [244]. Incorporation of 1,4,5,6,7,7-hexachlorobicyclo[2,2,1]-5-heptene-2,3-dicarboxylic anhydride [115-27-5] yields low-viscosity, satisfactorily drying binders that are recommended as flame-retardant marine and maintenance coatings.

Nondrying alkyd resins cross-linked with melamine resins at elevated temperature are used for industrial coatings. Resins containing synthetic fatty acids and resins based on glycidyl esters of Versatic Acid (Cardura E 10, Shell) are used in combination with low-viscosity melamine resins such as hexamethoxymethylmelamine. *p*-Toluenesulfonic acid and its amine salts are used as catalysts. On account of their high storage life, hydrophobic oxime esters of sulfonic acids are also employed.

Polyester Resins. Saturated polyesters (Section Saturated Polyester Coatings) of low molecular mass are formulated with hydroxymethylated, highly etherified melamine resins or blocked isocyanates to form one-pack paints, or with polyisocyanates to form two-pack paints. 2-Butanone oxime, caprolactam, and glycol ethers are used as blocking agents for isocyanates. Aliphatic polyisocyanates are preferred on account of their high resistance to yellowing. Linear or slightly branched oligoesters with terminal hydroxyl groups are usually used as resins. On account of their low molecular mass they require more curing agent than high-solvent binders and more powerful catalysis for sufficiently rapid cross-linking.

Given the possibilities for postcombustion of the solvent in industrial coating, the trend towards high-solids paints is, however, not so pronounced in this class of binders as it is, for example, in the case of maintenance paints and architectural coatings. High-quality paints with a solids content > 70 wt % are not widely available. Powder coatings are preferred as an alternative in the industrial sector.

Polyurethanes. The comments made concerning polyester paints also apply to polyurethanes (see also Section Polyurethane Coatings). Aliphatic isocyanates with a high yellowing resistance are increasingly used despite their high cost. Polyurethanes have already made large inroads into all application sectors [245]. Low-viscosity polyurethane oligomers can be combined with polyester, acrylic, and alkyd resins. They serve as modifiers for waterborne and low-solvent binders and improve the hardness and the flexibility of the paint film [246]. Two-pack polyurethane systems may contain as polyhydroxyl component both low-solvent or solvent-free oligoester diols and oligourethane diols. Although the latter are more expensive, they are more stable against hydrolysis.

Acrylic Resins. Low molecular mass acrylic resins (see also Section Acrylic Coatings) may be obtained by using fairly large amounts of chain-transfer reagents (e.g., thiols). The resultant problems of smell have, however, led to the implementation of new polymerization techniques, such as group-transfer polymerization or dead-end polymerization. Bifunctional (telechelic) polymers [247] are obtained and can be used as binders for low-solvent paints.

Epoxy paints (see also Section Epoxy Coatings) are usually applied as two-pack systems. Curing with amines is preferred in high-solids systems. Epoxy resins may also be used as coreactants or reactive diluents in low-solvent paints based on acrylic, polyester, or alkyd resins.

PVC plastisols, i.e., poly(vinyl chloride) dispersed in plasticizers, are used in special applications such as underseals for automobiles or as sealing compositions.

Commercial products include Alftalat (Hoechst); Alkydal (Bayer); Beckosol (Reichhold); Bevibond (Bergvik); Synthalat (Synthopol); Vialkyd (Vianova); and Worlekyd (Worlee).

3.3. Waterborne Paints

Waterborne (water-thinnable) paints were developed in the 1950s with the aim of replacing the common organic paint solvents by water, which has the obvious advantages of being noncombustible and nontoxic [248].

Waterborne paints provided the technological basis for electrodeposition paints (Section Electrodeposition Paints), in which negatively charged paint particles (anaphoresis, industrial introduction at the beginning of the 1960s), or positively charged paint particles (cataphoresis, industrial introduction at the end of the 1970s) are deposited from aqueous solution onto metallic substrates by application of an electrical field [249].

With the scarcity of raw material resources (e.g., petroleum) and the introduction of more stringent environmental legislation, waterborne paints were developed on a broader basis during the 1970s (e.g., as spraying and dipping paints). Among those coating materials that can be classed as environmentally friendly, waterborne paints have the widest potential as regards application, drying methods, and industrial uses.

3.3.1. *Properties*

In the production and application of waterborne paints, water is used as solvent or diluent. The physical properties of water and organic solvents differ (Table 8); some of these properties have to be taken into account when water is used as a paint solvent. The water molecules have a high dipole moment and associate with one another. This means that water has a high boiling point and high latent heat of evaporation despite its low molecular mass. This in turn results in fairly long evaporation times or in the need to supply energy in the form of heat to evaporate the water and dry the paint film. The high dipole moment of water is also responsible for its high surface tension. With substrates having a low critical surface tension (e.g., plastics or unsatisfactorily degreased metals) this leads to inadequate wetting, unsatisfactory edge covering, and crater formation. Critical surface tensions (at 20 °C and in millinewtons per meter) of water and other substances follow:

Aliphatic hydrocarbons	18.0 – 28.0
Aromatic hydrocarbons	28.0 – 30.0
Ketones	22.5 – 26.6
Esters	21.2 – 28.5
Alcohols	21.4 – 35.1

Water	72.5
Poly(vinyl fluoride)	18.5
Oil	28.9
Polypropylene	29.0
Stainless steel	30.0
Steel	55.0

The high polarity of water is responsible not only for differences between the application behavior of waterborne paints and solvent-containing paints. It also means that the organic polymers used as binders for waterborne paints must have a different structure from those used in solventborne paints.

Table 8. Physical properties of water, butanol, and xylene

Property	Water	Butanol	Xylene
Dipole moment, 10^{-30}C·m	6.17	5.54	2.07
Density, 10^{-3} g/cm^3	1000	811	874
bp, °C	100	117.7	140
Specific heat, J g^{-1} K^{-1}	4.187	2.998	1.666
Latent heat of vaporization (100 °C), J/g	2261	590	523
Evaporation number (diethyl ether = 1)	80	33	13.5
Latent heat of vaporization (liquid at 20 °C), J/g	2596	829	657
Surface tension (20 °C), mN/m	72.5	25.5	29.5

Waterborne paints differ according to the nature of their stabilization in water; the polymer molecules are dissolved in water or dispersed in water in the form of polymer dispersions or emulsion polymers. Further developments include polymer particles formed in organic solvents and then emulsified in water with low or, more commonly, high molecular mass emulsifiers being used for internal or external emulsification. Internal emulsification denotes that part of the binder molecule functions as an emulsifying moiety, whereas for external emulsification separate emulsifiers are required.

Water-Soluble Binders. Water-soluble binders consist of relatively low molecular mass polymers (M_r < 10 000) (e.g., alkyds, polyesters, polyacrylates, epoxides, and epoxy esters) whose individual molecules dissolve in water due to salt formation involving functional anionic or cationic groups.

Most water-soluble binders are anionic. They are made water soluble by neutralizing carboxyl groups with ammonia or highly volatile (generally secondary or tertiary) amines that evaporate during film formation. In cationic binders, salt formation usually occurs between the amine groups of the resin and organic acids (e.g., acetic or lactic acid). Salt formation with inorganic acids (e.g., phosphoric acid) is, however, also used. Typical applications include cationic electrodeposition paints.

The binder structure must be resistant to hydrolysis in aqueous, alkaline, or acid media. For alkyd or polyester resins the introduction of the carboxyl groups to be neutralized via full-ester bonds is advantageous; semi-ester bonds tend to undergo hydrolysis. Organic solvents and organic cross-linking agents (e.g., melamine resins) also influence viscosity stability on storage [250].

Water-soluble binders generally contain organic solvent (≤ 10 – 15 wt %) that originates from their production via polycondensation or polymerization reactions in an organic medium. They can, however, still be dissolved or diluted with water after neutralization. Anomalous viscosity behavior may, however, arise; it is characterized by a viscosity increase on dilution with water due to association of the binder molecules in water (Fig. 3) [251].

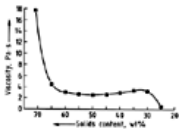

Figure 3. Viscosity behavior of a short-oil alkyd resin on dilution with water

Starting conditions: solids content 71.0 wt %; organic cosolvents 24.3 wt %; dimethylethanolamine 4.7 wt %

The organic cosolvents are mainly alcohols, glycol ethers, and other oxygen-containing solvents that are soluble or miscible with water.

Waterborne paints formed from water-soluble binders can undergo different forms of film formation and hardening, including drying at room temperature (e.g., polyacrylates, epoxy esters, and alkyds). The rates of initial drying and through drying are controlled by evaporation of water from the paint film (physical drying). Oxidatively drying, water-soluble binders (e.g., alkyds) also undergo chemical cross-linking that can be accelerated with driers.

Water-soluble binders (e.g., polyesters, polyacrylates) can also be chemically cured with cross-linking agents at elevated temperature (oven drying). These binders can be reacted with water-soluble or water-miscible amino resins (especially fully or partially methylated methoxymethylmelamine resins) or with water-soluble or water-miscible blocked polyisocyanates, which only react after the isocyanate groups have been unblocked.

On account of the viscosity anomaly, waterborne paints based on water-soluble binders have a relatively low solids content (ca. 30 – 40 wt %) and require relatively large amounts of organic cosolvents (up to ca. 15 %) to ensure water solubility and film formation. They also have the advantage of a broad drying spectrum (physical, oxidative, oven drying) and a wide range of application methods (dip coating, flow coating, spray coating, electrodeposition coating).

Paint films formed from water-soluble binders tend to be water-sensitive due to their hydrophilic, solubilizing groups. They can be formulated to have a high gloss due to good pigment wetting and stabilization. They can also have a high level of corrosion protection which depends on the corrosion-inhibiting pigments used and the chemical nature of the binder. The latter determines adhesion to the substrate and diffusion of water and oxygen through the paint film.

Polymer Dispersions (Emulsion Polymers). Waterborne paints based on polymer dispersions (usually referred to as emulsion paints) are not water soluble. They are water-thinnable systems composed of dispersions of polymer particles in water (see Section Waterborne Dispersion Paints (Emulsion Paints)). The particles consist of high molecular mass polymers (e.g., of styrene, butadiene, acrylate, or vinyl monomers) and are produced by emulsion polymerization (⟶ Polymerization Processes). These waterborne paints also contain small amounts of organic solvents (< 5 wt %) that serve as film-forming (coalescing) agents that evaporate on drying.

The required application viscosity of waterborne emulsion paints is generally obtained by adding a small volume of water. The evaporation behavior of polymer dispersions is similar to that of conventional, solvent-based paints.

Film formation occurs via evaporation of water from the paint film and coalescence of the high molecular mass binders; the release of water takes place relatively quickly (physical drying). The minimum film-forming temperature depends on the chemical structure of the polymers. Film formation can be facilitated by the addition of organic solvents (e.g., alcohols, butylglycol ethers) or by the action of heat.

Alternatively, chemical cross-linking reactions can also be used to harden polymer dispersions (oxidative drying or oven drying). An example is the cross-linking of hydroxyl groups of the polymer dispersions with hexamethoxymethylmelamine resins.

Characteristic properties of waterborne emulsion paints are relatively high solids contents (50 – 60 wt %), rapid dilution with water without any viscosity anomaly, and a low content of organic solvents (< 5 %).

As a result of the particulate nature of the binders only a low gloss and, in some cases, only limited corrosion protection can be obtained. Waterborne paints based on dispersions can be applied by spraying, however they are of only limited use for electrostatic coating and dipping applications due to their rapid drying properties.

Hybrid Systems. Combinations of water-soluble and water-dispersed binders may be used to achieve synergistic effects (e.g., to control the amount of organic solvents or the application behavior). Polymer dispersions or powdered binders may be used as water-dispersible binders [252]. The use of polymer dispersions allows the solids contents of water-soluble binders to be increased, the level of organic solvents to be reduced, and the physical drying time to be shortened. The use of water-soluble binders provides a broad application spectrum and yields paint films with well-balanced properties.

Internally and Externally Emulsified Binders. In recent developments polyurethane, polyacrylate, or polyester binders are synthesized in an organic medium and partially dissolved in water by salt-forming groups to form colloidal solutions. This involves internal emulsification. Alternatively, the polymers are dispersed and stabilized in the aqueous phase by adding low or high molecular mass emulsifiers (external emulsification). The disadvantages of water-soluble binders as regards low solids and high organic solvent contents, and polymer dispersions as regards film formation can be compensated in this way.

These binders form the basis for new industrial waterborne paints, an example being aqueous metallic base paints for automotive coatings and finishes. The binders of these paints are acrylic core – shell polymers [253], [254]. The core and shell have a different monomer composition, the shell has anionic groups which interact with water and stabilize the dispersion. The anionic shell allows pseudoplastic flow behavior which ensures parallel orientation of the aluminum pigments in the wet paint film. This orientation and the low solids content are responsible for the metallic gloss and high color flop (change in color observed on varying the viewing angle) of the base paints.

Generally, the binders have a low content of organic solvents (ca. 10 %) and can be applied over a wide temperature and humidity range. Water and organic cosolvents used in clearcoats applied by wet-on-wet coating evaporate rapidly after predrying with convection or IR radiation.

3.3.2. *Production and Application*

Most of the pigments used in conventional paints and coatings are also suitable for waterborne systems. Some exceptions are lead chromates and molybdates, manganese lakes (Pigment Red 48 and 52), some perylenes (Pigment Red 223), benzimidazolone yellow (Pigment Yellow 151), and isoindolines (Pigment Yellow 13a) which all have limited alkali resistance.

Another limitation results from pigments that contain a high amount of water-soluble components and electrolytes which may reduce storage stability, cause coagulation of the binder emulsion or dissolved resins, and result in loss of adhesion or blistering during exposure to humidity and water. Examples are surface-treated organic and inorganic pigments (some phthalocyanine blues, carbon black, transparent iron oxides).

As a rule, pigments can easily be dispersed in water-soluble binders because the water-soluble binder molecules promote wetting and stabilization of the pigments. The dispersion and stabilization of pigments in polymer dispersions is, however, more difficult because the dispersions form a continuous phase in water and uniform distribution of the dispersed pigment particles in the paint film is hindered by coalescence of the polymer particles.

Waterborne paints can be applied by a wide range of methods (e.g., by dipping electrodeposition, or spraying) followed by air, forced-air, or oven drying. The influence of climatic conditions (e.g., temperature and atmospheric humidity) on sprayability, substrate wetting, pouring and leveling behavior, as well as the rate of water evaporation from the wet paint film should, however, be taken into account. For example, the drying of waterborne paints proceeds more slowly at low temperature and/or high relative atmospheric humidity. These disadvantages can be remedied by suitable formulation of the paints, by heating the intake air to the spray cabins and the ventilation zone, and also by maintaining the atmospheric humidity within specific limits (e.g., 50 – 80 %).

In order to reduce the tendency to “pinholing” that occurs with higher layer thicknesses, control of application parameters is advantageous. Examples include using suitable spraying equipment, feeding preheated air to the ventilation zone, using IR drying in the ventilation zone, and avoiding steep stoving curves.

Corrosion-resistant and wear-resistant materials (e.g., stainless steel and in special cases suitable plastics) have to be used for the equipment required to produce and apply waterborne paints. This applies to production vessels, storage and transportation vessels, to the feed system used for application (e.g., closed circuit tanks, pipelines, and pumps) all of which must be able to withstand chemical mechanical stress. The use of corrosion-resistant materials for the spraying equipment is also advantageous.

Waterborne paints should be protected against frost during storage. In general, storage areas do not have to satisfy special fire prevention regulations.

3.3.3. *Uses and Environmental Aspects*

Waterborne paints for industrial use only account for a small proportion of the total production of coatings, paints, and thinners in the Federal Republic of Germany (Table 9) (i.e., for the serial production of industrial and consumer goods such as machinery, tools, electrical equipment). Whereas the proportion of waterborne emulsion paints and plaster for decorative use (architectural paints, see Section Waterborne Dispersion Paints (Emulsion Paints)) accounted for ca. 45 % of the total production between 1979 and 1989, the proportion of waterborne paints for industrial use merely doubled from ca. 2 % to almost 4 %. However, industrial waterborne paints now have a broad application spectrum based on the development of new paint technologies. The most important use sectors include electrodeposition coating of automobile bodies and accessories; spraying and dipping primers, as well as one-layer coatings for engines, chassis, axles, and automobile accessories; sprayed fillers for automobile bodies (first industrial use in Europe 1981); waterborne paints for automobile bodies (1987); aqueous protective waxes for preservation of automobile body shells (1989); aqueous primers and one-layer coatings for metal furniture, electrical and electronics equipment, machines, machine parts, agricultural equipment and machinery, construction machinery; and aqueous primers and one-layer coatings for plastic parts (e.g., radio and television cabinets and office equipment).

Table 9. Production of paints in the Federal Republic of Germany (Statistisches Bundesamt, BRD)

Percentage of total production	1979	1989
Dispersions and plaster (decorative)	46.0	41.6
Waterborne coatings	1.9	3.7
Powder coatings	0.6	2.4
Others	51.5	52.3
Total production, t	1.33×10^6	1.42×10^6

The prospects for the increasing use of industrial waterborne paints lie in their economic advantages and in the possibility of reducing solvent emission during application to comply with legal requirements. Savings in organic solvents as diluents, savings in insurance premiums, lower energy consumption in spray cabins, ventilation zones, and drying ovens all contribute to the overall economy of waterborne coating materials.

Thanks to their low organic solvent content, the use of waterborne paints substantially reduces solvent emission from paint or spraying shops (spray cabin, ventilation zone, drying oven).

Waterborne paints can generally be classed as less toxicologically harmful than corresponding solvent-based paints. Nevertheless, lung-penetrating paint mists (aerosols) of water-based paints present a health hazard and appropriate protective measures (e.g., use of respirators) must be taken depending on the workplace concentration. Water-based paints are not, however, subject to such stringent regulations concerning hazardous substances as conventional, solvent-based paints.

3.4. Coating Powders

3.4.1. *Introduction and Economic Importance*

Coating of substrates (particularly metallic substrates) with particulate plastics is a well-established technology. The fluidized-bed process was developed in 1952. The electrostatic powder coating technique introduced in 1965 has, however, made a substantially greater impact [255]. These two processes are discussed in Section Powder Coating .

Other powder coating processes include blowing thermosetting plastics powders onto hot parts for electrical insulation, flocking polyethylene onto hot pipes to

produce an external coating, and coating small, cold items in electrostatic fluidized beds, followed by thermal curing. Lances are also used for the internal coating of preheated vessels and pipes, vibration and rotation sintering, as well as flame spraying for repair work and large articles, and the minicoat and maxicoat processes for coating small items.

Thermosetting coating powders are applied mainly by electrostatic spraying and dominate the market. The annual growth rates worldwide are ca. 10 – 20 %. Total world production in 1989, including "functional materials", was 250 000 t with a value of ca. $ 1.2×10^9. This was subdivided according to geographical regions and countries as follows (10^3 t):

Europe	140
North America	55
Far East	37
Rest of the world	18
United States	47
Italy	33
Federal Republic of Germany	33
United Kingdom	19
France	18
Japan	18
Spain	11

In contrast to wet materials, almost no material is lost during application of coating powders (no solvents, no split-off products, powder overspray is recovered). These tonnages roughly correspond to three times the amount of ready-for-spraying, conventional industrial paint.

Table 10 summarizes the consumption of thermosetting binders used in various regions. The epoxy – polyester systems predominate over polyester – trisglycidyl isocyanurate and epoxy systems. The consumption of thermoplastic coating powders, particularly those used in the fluidized-bed and flocking processes, is substantially lower than that of thermosetting powders (Table 11). Their production does not show a comparably significant growth rate.

Table 10. Production of thermosetting coating powders in 1988 classified according to binder types (as percentages)

Binder types	World	Europe	United States *	Japan *
Epoxy	20–25	14–16	29	23
Epoxy–polyester	40–45	55–60	17	20
Polyester–trisglycidyl isocyanurate	20	20–25	21	5
Polyester–polyurethane	10–15	4– 5	32	35
Acrylic	2	0	1	17

* [256]

Table 11. Estimated production (in 10^3 t) of thermoplastic coating powders in 1989 classified according to binder types

Binder types	Europe	World
Polyethylene (PE)	6–10	18–20
Polyamide (PA)	3– 4	5– 7
Poly(vinyl chloride) (PVC)	2– 4	4– 8
Ethylene–vinyl alcohol copolymer (EVOH)	1– 2	2– 3
Thermoplastic polyester (SP)	1	1– 2
Total	14–17	30–36

3.4.2. *Production*

See also Section Powder Coatings.

Thermosetting coating powders are produced exclusively by the extruder process. This method is employed because it is the cheapest and the most widely applicable and flexible; it allows large plants to be supplied with material of consistent quality. Production with extruders includes the following steps:

1. Testing of raw materials
2. Weighing of the premixture
3. Mixing in the premixer
4. Homogenization and dispersion in the extruder (80 – 140 °C)
5. Cooling the extrudate to room temperature
6. Intermediate testing (monitoring the extrudate)
7. Flaking
8. Fine grinding in an air-classifying mill
9. Screening in a sieve
10. Checking specific properties of the coating powder

Since modification of the coating powder (e.g., tinting) after production) is nearly impossible, all relevant process parameters (e.g., raw materials, machine settings) must be kept constant. This has the advantage for the user that changes such as thickening, settling, or floating cannot occur during application and storage. Large batch sizes are necessary because cleaning of the plant and equipment is time-consuming and expensive. An extremely high level of cleanliness must be maintained during production because different coating powders may be incompatible.

Thermoplastic coating powders can also be produced with extruders. Grinding is particularly important here and, depending on the binder and desired particle size distribution, can be carried out at ambient temperature [e.g., some polyethylene (PE) powders for the fluidized-bed process] or at low temperature using liquid nitrogen, etc. [e.g., polyester (SP) electrostatic powders]. Some poly(vinyl chloride) (PVC) powders are also produced in this way. The most efficient process for producing PVC powders is to mix the components in a high-speed mixer at 80 – 110 °C, followed by cooling in a second mixer, and screening.

Polyamide (PA) powders are generally precipitated from solution. Ethylene – vinyl alcohol copolymer (EVOH) powders are precipitated by hydrolyzing ethylene – vinyl acetate copolymer solutions.

3.4.3. *Properties*

The typical constituents of a coating powder are binders (resins, if necessary hardeners and accelerators), colorants (mainly pigments), additives, and in some cases extenders.

Binders. Thermosetting and thermoplastic binders are listed, together with their abbreviations, in Table 12. A broad range of raw materials is now available and it has become difficult to systematically classify binder properties. Outstanding properties are obtained with suitably selected systems and have contributed to increased use of coating powders.

Table 12. Summary of coating powder binders

Use	Thermosets		Thermoplasts	
Mainly interior	EP–SP	epoxy resin – COOH polyester resin	PE	polyethylene LDPE, LLDPE, HDPE
	EP–DCD	epoxy resin – modified dicyanodiamide		
	EP	epoxy resin – phenolic hardener	PA	polyamide 11, 12
	EP	epoxy resin – imidazoline derivative	SP	polyester
	EP	expoy resin – anhydride adducts		
Interior and exterior	SP–TGIC	COOH polyester resin – trisglycidyl isocyanurate	EVOH	ethylene – vinyl alcohol copolymer
	SP–PUR	OH polyester resin – isocyanate adduct	PVC	poly(vinyl chloride)
	AC–PUR	OH acrylic resin – isocyanate adduct		
	AC–DDA	glycidylacrylic resin – dodecanedicarboxylic acid		

The *thermosetting binders* are the most important. Their properties are summarized in Table 13. The quantitatively predominant EP, EP – SP, and SP – TGIC types cure by polyaddition without release of split-off products. Only SP – PUR and AC – PUR systems are cured by polycondensation and release some caprolactam.

Table 13. Typical properties of thermosetting powders [a]

Compound [b]	Stoving temperature, °C	Weight loss during stoving, wt %	60° gloss, %	Other properties
EP (cold electrostatic spraying) [d]	120–240	<0.1 organics	0–100 [c]	depending on the hardener, excellent mechanical, chemical, electrical, anticorrosive properties, yellows if overstoved, marked loss of gloss and chalking under weathering; with coating of preheated parts and sufficient heat content of the parts, hardening takes place through the residual heat, organic pigmentation often restricted due to chemical reaction
EP (hot electrostatic spraying, fluidized bed, blowing) [d]	130–240	≤1 cracking products	70–100	
EP–SP	140–220	<0.1 organics	20–100 [c]	excellent mechanical properties, almost no yellowing on overstoving, loss of gloss and chalking under weathering
SP–TGIC	160–220	<0.1 organics	10–100 [c]	excellent mechanical properties, no yellowing on overstoving, outstanding weather resistance
SP–PUR	180–220	2–4 caprolactam	0–100	good mechanical properties, some yellowing on overstoving, good weather resistance, layer thickness limited due to caprolactam release
AC	180–220	<1 organics or 2–4 caprolactam	70–100	moderate mechanical properties, very good hardness and weather resistance, in the case of PUR hardening layer thickness limited due to caprolactam release

[a] The coating powders are applied by cold electrostatic spraying; EP resins can also be applied by hot electrostatic spraying, fluidized bed, and blowing. [b] For definition of abbreviations, see Table 12 [c] Smooth flow and low gloss are not obtained at low stoving temperatures. [d] Hot and cold denotes the temperature of the workpiece.

[c] [257].
[d] [258].

Thermosetting powders are mainly applied to cold workpieces by electrostatic spraying. Materials for interior use are generally produced from epoxy resins based on bisphenol A – epichlorohydrin and are cured with acidic polyester resins or modified (i.e., substituted or accelerated) dicyanodiamides. Materials for highly reactive systems are cured with phenolic hardeners (i.e., oligomers with phenolic terminal groups). Other curing agents include imidazoline derivatives (for matt coatings), and anhydrides and their adducts. The application of thermosetting powders to hot substrates by electrostatic spray or fluidized-bed processes is restricted to EP coating powders.

Coating powders for exterior use are mainly based on the SP – TGIC system. Of some importance is the SP – PUR system, i.e., the combination of hydroxy-functional polyester resins with caprolactam-blocked isocyanate adducts (primarily based on isophorone diisocyanate). Acrylic resins are of importance mainly in the United States (AC – PUR) and in Japan (AC – DDA). The AC – PUR system is used for automotive components and in domestic appliances. The PUR systems based on uretdione (a dimerization product of isocyanates) do not release reaction products and may be mentioned as a specialty.

Typical properties of *thermoplastic binders* are summarized in Table 14. Application is mainly by fluidized bed but includes application by flocking. Electrostatic spraying is also used for finely ground powders. The main uses are in the interior sector. The PA and EVOH powders are the most widely used thermoplastic binders.

Table 14. Typical properties of thermoplastic powders

Binder [a]	Application method [b]	Preheating and stoving temperature [c], °C	Adhesive primer required [c, d]	Weight loss during stoving, wt %	Other properties
PE	hot flocking, and hot fluidized bed	280–400	in some cases	≤1 paraffinic hydrocarbons	without primer (hot-melt adhesive) poor adhesion, soft, high viscosity, complicated shapes are difficult to coat, poor weather resistance
PA	hot fluidized bed	280–400	in some cases	0.2 laurolactam	primer required for good wet adhesion, hard, tough, good resistance to chemicals and detergents, good anticorrosive properties, average weather resistance, undergoes yellowing
PVC	cold electrostatic spray, hot fluidized bed	160–350	yes	1–6 plasticizer	extremely flexible, good weather resistance, undergoes yellowing, inexpensive
EVOH	hot fluidized bed	200–400	no	<0.1 organics	good adhesion, elastic, tough, good anticorrosive properties, good weather resistance
SP	hot electrostatic spray	240–400	no	≤1 organics	good adhesion, film formation possible in a few seconds, good weather resistance, very good elasticity, resistant to sterilization

[a] For definition of abbreviations, see Table 12 [b] Hot and cold denote that the workpiece is preheated or at ambient temperature.

Colorants. Coating powder colorants are almost exclusively pigments and must satisfy the following requirements:

1. Thermal stability at the curing temperature
2. Only a slight increase in binder viscosity
3. No reaction with other constituents of the formulation
4. Stability towards shearing forces during extrusion and grinding

Inorganic pigments (e.g., titanium dioxide, iron oxides, and chromium oxides) satisfy these criteria the best. To obtain maximum durability in exterior-grade powder coatings, lead pigments are sometimes still required. In Europe the use of cadmium pigments is declining. Replacement by organic pigments generally restricts color selection in the red, orange, and yellow ranges. Metallic and pearlescent effects are obtained by using flakes of aluminum or coated mica.

Other Constituents. The remaining constituents of coating powders impart specific properties (e.g., additives to adjust flow) and reduce cost (extenders such as calcium carbonate or barium sulfate).

3.4.4. *Testing*

DIN 55 690 classifies and identifies coating powders according to the binder. Standards for testing coating powders in powder form are summarized in Table 15. Relevant European standards (CEN) have not yet been drafted. Quality aspects are described in detail in [259].

Table 15. Standards for testing coating powders in powder form

Parameter	ISO	DIN	AFNOR *	BS	ASTM
Sampling	DIS 8130-10	53225/53226	T 30 – 048	3900-J 1	D 1898
Characterization					
Density (gas comparison method)	8130-2	55990-3	T 30 – 505		
Density (liquid method)	8130-3	55990-3		3483-B 8	D 3451
Gel time	DIS 8130-6	55990-8	T 30 – 506	3900-J 3	D 3451
Sintering range			T 30 – 502		D 3451
Resistance to blocking		55990-7	T 30 – 507	3900-J 4	D 3451
Chemical storage stability		55990-8	T 30 – 507	3900-J 4	D 3451
Acid extract	6713	DIN-ISO 6713	T 30 – 210	3900-B 5	
Application					
Particle size distribution	8130-1	55990-2	T 51-701	3900-J 2	D 3451
Lower ignition limit	8130-4	55990-6			
Free flow					D 3451
Fluidization	DIS 8130-5		T 30-501		
Flowability (fluidized)	DIS 8130-5		T 30-500		
Deposition efficiency					D 3451
Stoving conditions		55990-4			
Loss on stoving	8130-7	55990-5	T 30-503	3900-J 5	D 3451

* AFNOR = Association Française de Normalisation.

To check the film quality of a powder coating, the same standard methods as for liquid coatings are used. Minimum requirements for powder coatings have been laid down for certain areas of use in relevant standards (e.g., DIN 30 671 for pipes, DIN 30 677 for mountings and fittings, BS 6496 for individually coated structural elements) or by quality associations (e.g., Qualicoat or GSB, “Quality Association for Individually Coated Building Elements” relating to coating of aluminum and “Quality Association for Heavy Corrosion Protection” relating to the coating of valves and fittings).

3.4.5. *Storage and Transportation*

Coating powders are shipped in containers of varying size. They may, for example, be packaged in cardboard boxes each containing one or more polyethylene bags holding 20 – 25 kg. For larger customer off-takes, reusable metal (e.g., aluminum) containers with capacities for 400 – 1000 kg are used. Containers may be equipped with built-in fixtures such as a porous layer at the base for aerating and fluidizing the contents. However, containers without such fixtures are often preferred because they are easier to clean. Thermoplastic powders are often supplied in plastic drums or sacks.

Coating powders should be stored in a cool dry place. They should not be subjected to excessive pressure or temperature.

3.4.6. *Environmental Aspects and Safety*

Powder coating is the coating method that produces the least environmental pollution because

1. the occurrence of paint slurry is avoided,
2. the separation of dust – air mixtures is simple, and
3. the coating powders are solvent-free and generally cure by polyaddition.

Emissions from coating powder spray cabins are generally << 5 mg coating powder per cubic meter of exhaust air. Emissions from the driers are generally also extremely low. Other substances, apart from very small amounts of water (≤ 1 %), that are released in the stoving of coating powders are listed in Tables 13 and 14. On account of these very low emission levels, the use of coating powders is not limited by environmental protection legislation.

Coating powders that do not contain any hazardous components (e.g., lead pigments) should generally be regarded as inert dusts and do not have toxic or fibrogenic effects. As a rule, the fraction of respirable dust that can penetrate the lungs is << 10 wt %. Disposal of coating powder waste in landfills does not in general pollute groundwaters or surface waters.

Coating powders are combustible dusts and they can form explosive powder – air mixtures in certain concentration ranges. However, the danger risk is substantially lower than that for solvent-based paints for the following reasons:

1. When stored in bulk, coating powders are difficult to ignite and when heated generally only fuse together
2. Higher energy sparks are required to ignite powder – air mixtures than solvent – air mixtures
3. The glow temperature and the minimum ignition temperature of a powder – air mixture are very high (450 – 600 °C)
4. Powders without fines for fluidized-bed application are extremely difficult to ignite

The safe operation of coating powder plants and equipment is covered by national provisions and regulations. In the case of Europe, the ten-language brochures “Safe Powder Coating” published by CEPE (the European Committee of Paint Associations, Brussels) should be mentioned.

3.4.7. *Uses*

Thermosetting coating powders are widely used for industrial metal coating and are distributed among the various markets as indicated in Table 16. Uses range from manual application up to fully automatic, large-scale plants with an annual coating powder consumption of several hundred tonnes. Job coaters as a service industry account for a significant proportion of the coating powder consumption.

Table 16. Uses of thermosetting coating powders (1987) expressed as percentage of production

Use	Europe, %	USA, % *
Appliances	10	30
Automotive	8	24
Outdoor building elements	13	4
Job coaters	25	42
Metal fixtures, indoor building elements	12	
Metal furniture	12	
Machinery	6	
Heating elements	6	
Others (e.g., electrical insulation, corrosion protection, coating of glass and ceramics	8	

* [260]

The main reasons for the steadily increasing use of thermosetting coating powders (see Section Introduction and Economic Importance) are as follows:

1. Economy
2. Simple, easily automatable application
3. Ecological advantages
4. Good film properties, wide range of layer thicknesses
5. Improved coating powders and plants
6. New areas of application
7. Many years' experience with all sizes of coating plants

The building and construction industry is the main user of coating powders, for both exterior and interior use. Coating powders have proved popular for coloring light metal façades, panels, and profiles on account of their outstanding weather resistance and robust surface. They also allow architects complete freedom of expression and use.

In the interior-use sector, a diverse range of objects and parts can be powder coated. Examples include partition walls, ceilings, door and window frames, and radiators. The coating of radiators is particularly interesting because it involves penetration of the coating powder into deep Faraday cages. The excellent impact resistance of most powder coating films is important in the building sector because the articles are roughly handled.

A further important use is the coating of steel and garden furniture, machine parts, and electrical components such as switch cabinets. Applications in the furniture sector include the coating of metal rails and fittings.

The coating of steel chairs and garden furniture can be more easily automated with electrostatic powder coating than with conventional wet coating. Chairs have an unfavorable shape and coating with wet paint can only be partially automated; generally a final coat is applied manually. With coating powders, however, coating can be performed fully automatically. The paint film has a high scratch resistance which considerably simplifies packaging.

The coating of cast iron is an important use in the engineering sector. Specially formulated coating powders are available for porous substrates to prevent bubble formation in the film.

In the electrical industry, the coating of switchboards with textured powders should be mentioned. Important requirements include satisfactory punching quality, recoatability with almost all paints, and constant visual properties.

Further important users are the household appliance industry and the automotive parts and manufacturing industry. The first large-scale plant for powder coating (for coating refrigerators and deep freezers) came into operation in 1973. Powder coating has since become widespread in the household appliances sector.

An extremely wide range of automotive parts and components on, in, or underneath automobiles are powder coated (e.g., trim, headlights, springs, transmission shafts, horns). By coating reflectors, smooth leveling is achieved on the rough, deep-drawn substrate, which eliminates the time-consuming, expensive polishing that is necessary with a liquid paint. Interesting multilayer coatings that combine powder and/or paint layers can be applied to wheels to obtain decorative effects. In Japan and the United States coating powders are used as primers or surfacers on automotive bodies.

Thermosetting coating powders are also used for storage furniture shelves, shopfitting, and wire goods. The coating of nonmetallic substrates (e.g., glass or ceramics) and highly decorative articles (e.g., spectacles and sanitary fittings) should also be mentioned.

Important areas of use of epoxy resin powders include the interior and exterior coating of pipes, fittings, fixtures, filter plates, and steel bars for reinforced concrete, as well as the electrical insulation and the encapsulation of electronic components.

Thermoplastic Coating Powders.
Polyethylene is mainly used for the exterior coating of pipes. It is also employed for coating wire goods, especially in the domestic appliances sector.

Polyamide is extremely resistant to washing lyes and chemicals, provides good corrosion protection, and has a high impact strength and low degree of abrasion. Areas of use are therefore very varied, and include wire articles (e.g., dishwasher baskets) and metal furniture (e.g., for hospitals).

Poly(vinyl chloride) is mainly used for wire articles (e.g., fences, domestic and household appliances). The release of plasticizer during coating is a disadvantage.

Ethylene – vinyl alcohol copolymer is used particularly for heavy corrosion protection (e.g., in water treatment and chemical plant construction) but also for wire goods (e.g., household appliances).

The main use of *thermoplastic polyesters* is the coating of weld seams in cans.

3.5. Waterborne Dispersion Paints (Emulsion Paints) [261], [262]

The paints and coatings sector underwent a major revolution after World War II. In Germany the changeover to synthetic polymers [poly(vinyl acetate) and polyacrylates] took place during the war due to the lack of natural raw materials (vegetable oils for alkyd resins). Since the postwar demand for synthetic styrene–butadiene rubber fell in the United States, part of the production capacity was converted to styrene – butadiene latex for paints (latex paints). Dispersion paints, commonly referred to as emulsion paints are now the largest product group worldwide in the paint and coating industry. In the Federal Republic of Germany, for example, the total production of paints and coating materials in 1988 was 1.2×10^6 t, of which interior-use emulsion paints accounted for 226 000 t, exterior-use emulsion paints for 120 000 t, emulsion lacquers for 20 000 t, and dispersion-bound plasters for 124 000 t.

Binders consisting of plastics (polymer) dispersions can be used for matt to gloss coats, smooth paint-like products, and plasters (synthetic resin-bound plasters, DIN 18 550) containing coarse, structure-imparting aggregates.

Substrates such as plaster, concrete, cementbonded slabs, gypsum plaster, plaster board, wallpaper (wood-chip, glass textile, vinyl wallpaper, etc.), brickwork and sand–lime bricks, cement or concrete roofs, floors, sidewalks, wood and wood materials (e.g., chipboard and hardboard), and many other substrates can be coated with coating materials based on polymer dispersions. Many coating materials containing polymer dispersions that form films and dry at elevated temperature (>40 °C) are used for the industrial coating of plastic articles (e.g., audio equipment housings, dashboards and instrument panels in vehicles), metals, window frames, etc.

The advantage of waterborne coatings containing polymer dispersions is that they can be applied without stringent safety measures because water is their principal liquid component. Most of them dry physically, by evaporation of water, under ambient conditions and generally from + 5 °C (minimum film-forming temperature, MFT) without producing any permanent odor. The MFT depends on the auxiliaries and residual monomers in the plastics dispersions. Emulsion paints are noncombustible in liquid form (easy storage and transportation); equipment, hands, and spillages can be washed and cleaned with water.

According to DIN 55 947, a plastics dispersion consists of polymer particles distributed in a liquid phase (water in the case of emulsion paints). Polymer dispersions may be fine (particle diameter 0.1 – 0.3 µm), medium (particle diameter 0.3 – 2 µm), or coarse (particle diameter 2 – 5 µm). Fine and medium dispersions are mainly used as binders for emulsion paints. If a particularly good penetration ability is required, very fine dispersions with particle sizes of 0.04 – 0.1 µm are also used. Polymer dispersions of relatively uniform particle size are termed monodisperse, while those with a nonuniform or broad particle size distribution are termed heterodisperse or polydisperse. Fine dispersions (0.1 – 0.3 µm) generally have a fairly high pigment binding capacity (i.e., at equal pigment volume concentration (*PVC*) the pigments are bound better than with coarser dispersions). Paints with improved application properties can be formulated with dispersions in the particle size range > 0.3 µm. Satisfactorily leveling and glossy coatings can be produced with fine, heterodisperse dispersions.

Emulsion paints consist of polymer dispersions as binders, pigments, extenders, and small amounts of auxiliaries (in some cases < 1 %). Waterborne polymer dispersions are produced by emulsion polymerization; monomer droplets are polymerized in water-containing surfactants and protective colloids. The size and size distribution of the dispersed polymer particles can be controlled by adjusting the stirring rate in the polymerization reactor and by selecting appropriate protective colloids. For a detailed description of emulsion polymerization, see ⟶ Polymerization Processes.

Polymers. The nature and amount of the employed dispersion greatly influences the quality of an emulsion paint. The nature of the plastics dispersion is governed by the monomers used for emulsion polymerization. The most important polymers for emulsion paints are listed below in order of increasing resistance to alkali (which may be of importance for use on exterior mineral substrates):

Vinyl acetate copolymers are used worldwide in polymer dispersions. Copolymers with dibutyl maleate are particularly elastic, while copolymers with vinyl esters of Versatic Acid have particularly good weather resistance for shade formulations. Copolymers with acrylate esters and vinyl acetate – ethylene – vinyl chloride terpolymers with a good cost – performance ratio are also important.

Vinyl propionate copolymers (e.g., with acrylate esters) are used in the same way as vinyl acetate copolymers.

Acrylate – methacrylate copolymers (pure acrylates) are used in particular for house paints and other emulsion paints with high weather resistance and for waterborne industrial paints.

Styrene – acrylate copolymers are widely used as binders for interior-use paints, plasters, and exterior-use paints with a relatively high *PVC*. Water absorption and elasticity decrease with increasing styrene content.

Styrene – butadiene polymers have a very low water uptake but tend to undergo chalking and are therefore used for formulating interior-use (latex) paints that can withstand repeated cleaning.

Commercial Products. Examples of commercial products are Acronal, Propiofane (BASF); Lipaton (Hüls); Mowilith (Hoechst); Plextol (Röhm); Primal (Rohm & Haas); Rhodopas (Rhône-Poulenc); and Vinnapas (Wacker).

Pigments. Many colored pigments are used in emulsion paints. Titanium dioxide is responsible for the hiding power, whiteness and, depending on the type of TiO_2 used (rutile or anatase), the chalking resistance of paints. Surface-treated rutile should be used for chalking-resistant paints because anatase reacts photocatalytically with oxygen and moisture resulting in radical reactions, decomposition of the binder at the TiO_2 interface, and chalking.

Extenders are used to adjust the properties of a coating (e.g., layer thickness and structure). They influence water uptake and permeability to water vapor, carbon dioxide, sulfur dioxide, and nitrogen oxides. Inorganic extenders are obtained from natural minerals by crushing, grinding, elutriation, and drying. The most important extenders are calcite, chalk, dolomite, kaolin, quartz, talc, mica, diatomaceous earth, and barytes. Synthetic inorganic extenders are produced from inorganic products by digestion, precipitation, or annealing. The most important synthetic extenders are precipitated calcium carbonate, barytes (blanc fixe), and silicates (Aerosil). Synthetic organic extenders include polymer fibers, as well as the more recently introduced hollow polymer spheres which can be used to formulate paints with a particularly low density.

Auxiliaries.
Dispersants ensure that the individual particles in the emulsion paint do not combine to form agglomerates. *Protective colloids and emulsifiers* are used during emulsion polymerization to ensure that small polymer spheres are formed in the aqueous phase but do not fuse together. They influence film formation of the emulsion paint and can cause foaming. Protective colloids include poly(vinyl alcohols) and cellulose ethers. Emulsifiers include anionic and nonionic surfactants.

Wetting agents are used to disperse the extenders and pigments. They include sodium polyphosphates, and sodium and ammonium salts of low molecular mass poly(acrylic acids). Polyphosphates tend to undergo secondary reactions; polyacrylates are therefore preferred from the point of view of storage stability, gloss, etc.

An excess of dispersion auxiliaries increases the water sensitivity of the coating films and thus reduces washfastness and weather resistance. The optimum amount and composition of dispersants is of decisive importance as regards the quality of each formulation.

During the production and transportation of emulsion paints the dispersants and emulsifiers cause foaming which leads to the undesirable formation of craters and pores in the coating film. *Defoamers* reduce the surface tension of the liquid to such an extent that the air bubbles in the foam collapse. They include silicone oils, waxes, fatty acids, if necessary on a solid carrier, e.g., on silicates. The defoamers are formulated in an auxiliary liquid (e.g., mineral oil, silicone oil, or glycol).

Preservatives prevent microbial growth in emulsion paints stored in drums or cans. Fungi and bacteria are present in the raw materials (especially water) and also pass during the production process into the emulsion paints. The most commonly used preservatives include formaldehyde-releasing agents, chloroacetamide, isothiazolinone, and heterocyclic nitrogen – sulfur compounds. For health reasons mercury compounds and organic tin compounds should not be used. At higher concentrations, preservatives have a fungicidal action and prevent fungal infestation of the coating and coated substrate. Coatings may also contain algicides.

Thickeners are used to modify the rheological properties of emulsion paints. Rheological properties include viscosity, consistency, application behavior, coating resistance, splashing during roller application, leveling, covering of butt surfaces and edges, pouring resistance, settling behavior, and storage stability (postthickening).

Similar to all other auxiliaries, thickeners influence the water retention and open time, water resistance, washfastness, abrasion resistance, weather resistance, and gloss. The most important thickeners are methyl cellulose and hydroxyethyl cellulose. The viscosity rises with increasing polymerization of the cellulose ether. Synthetic thickeners based on poly(meth)acrylates, polyvinylpyrrolidone, and polyurethanes are also important. Inorganic thickeners include layer silicates of the montmorillonite type (e.g., aluminum silicate and magnesium silicate). They confer a better washfastness and abrasion resistance than the cellulose ethers and polyacrylates. However, they develop a lower water retention capacity (open time).

Film-forming auxiliaries are high-boiling solvents that promote film formation and reduce the minimum film-forming temperature. They act as temporary plasticizers until the end of film formation and then evaporate during the drying phase. Film-forming auxiliaries include butyl diglycol acetate, butyl diglycol, glycol ether, and glycol ether esters, which in some cases are also termed plasticizers. An example of one of the most important types is dibutyl phthalate.

Film formation is the physical “solidification” of an emulsion paint. Water evaporates from the wet coating film and the polymer particles approach one another until they come into contact. On further evaporation of water, capillary pressure is generated in the water-filled interstices between the polymer spheres, as well as between the polymer spheres and other solids (extenders, pigments). The polymer particles thus coalesce to produce a coating film. A precondition for this film formation is that the minimum film-forming temperature (MFT) is reached or exceeded during drying. Below the MFT an emulsion paint forms a noncoherent, fissured, mechanically unstable film. The MFT depends on the structure of the polymer binder, its glass transition temperature, and on the nature and amount of film-forming auxiliaries and other additives in the formulation. The MFT range may extend from –5 to +50 °C; in the case of house paints and wall paints the range is + 5 to + 10 °C.

Production. Emulsion paints are produced in paddle mixers, of which high-speed dispersers are by far the most widely used and effective. High shear forces are produced in these dispersers which account for their good dispersing action. With some viscous, coarsely filled emulsion paints, however, the shear forces can destroy the dispersion; beam agitators are more suitable in this case.

Properties. The most important properties of emulsion paints are their rheology (which affects the application properties); hiding power (extinction); pigment volume concentration (influences stress, water uptake, water vapor permeability, weather resistance); water vapor and water permeability (influences the moisture balance of the substrate and thus the durability and protection of the substrate against weathering); adhesion to the substrate (particularly with alternating moist – dry, moist – hot, and moist – frosty conditions); and the elasticity (in the case of crack-covering coatings and wood coatings).

Application. Emulsion paints can be applied by brush, roller, dipping, inundation, and spraying. A wide range of qualities of emulsion paints with tailor-made properties can be produced from the available binders, thickeners, and auxiliaries.

Types of Emulsion Paints. The most important types of emulsion paints are the following:

Emulsion lacquers have a pigment volume concentration of 15 – 20 % and contain almost exclusively TiO_2 and finely particulate inorganic and organic colored pigments with which satisfactorily leveling paints with a uniformly high surface gloss can be formulated. Emulsion lacquers are produced mainly with fine, pure acrylate dispersions. Lacquers used for industrial coatings are also emulsion lacquers.

Exterior-use paints have a pigment volume concentration of 30 – 55 %. They may be formulated with all of the aforementioned dispersions. Properties are adapted as appropriate to the substrates, climatic conditions, market circumstances, etc. The most important properties of exterior-use paints are their water and water vapor permeability (moisture balance), chalking resistance, and adhesion to the substrate. Special requirements apply in certain areas of application (e.g., CO_2 protection on reinforced concrete, fungal and algal resistance in moist surroundings, crack bridging on cracked substrates, and alkali resistance on new mineral substrates).

Interior-use paints have a pigment volume concentration of up to 80 % depending on the required surface effects. The gloss, cleaning ability, mechanical strength, washfastness, and abrasion resistance decrease with increasing pigment volume concentration.

Tinting paints are used to tint and shade emulsion paints. They contain colored pigments and wetting agents to ensure good color development in white emulsion paints.

Synthetic resin-bound plasters contain, in addition to the components of an emulsion paint, coarse aggregates (sand, marble chippings, colored stones) to produce decorative effects.

3.6. Nonaqueous Dispersion Paints

The term nonaqueous dispersions can be applied to all particulate dispersions in any liquid medium other than water. It has, however, come to have a more restricted meaning, namely, particulate dispersions of organic polymers in organic liquids.

The upsurge of interest in this type of colloidal system followed the development in the 1950s of the technique known as dispersion polymerization [263]. This process provides a means of preparing nonaqueous polymer dispersions in a controlled manner. A wide range of such dispersions have been made, mainly by free-radical addition polymerization.

As with many technological advances, however, the development of stable, nonaqueous colloidal dispersions posed theoretical problems including the mechanism responsible for their stabilization. Electrostatic stabilization, operative in many aqueous systems, could not account for the stability of dispersions in low dielectric media and the concept of steric stabilization was proposed [264].

Steric Stabilization. Steric stabilization involves interactions between polymer chains located on the surface of particles [265-268]. Stabilizing repulsive forces are generated by the interaction of polymer chains located on approaching particles. These forces are related to the free energy changes accompanying the increase in the concentration of polymer segments during interpenetration or compression of the polymer chains. For practical purposes, sterically stabilized dispersions are achieved if:

1. Every particle in the dispersion carries a complete sheath of solvated polymer chains, each chain being anchored at one or a few points to the particle surface. Solvated denotes that the liquid phase of the dispersion would be a good solvent for the polymer chains if they were not anchored to the particle surface.
2. The chains are so strongly anchored to the particle surface that the sheath remains intact during encounters with other particles.
3. The sheath of polymer chains is thick enough to render the van der Waals attraction between the particles insignificant at their point of closest approach. In general, all polymer chains are long enough to provide such a sheath [264].

Simple homopolymers or random copolymers that are solvated by the dispersion medium are only weakly adsorbed to the surface of the polymer particles and cannot therefore provide the strongly anchored sheath of solvated polymer chains necessary for steric stabilization [269]. Block and graft copolymers play a key role in the steric stabilization of nonaqueous dispersions. Block and graft copolymers contain long runs of identical monomers in the polymer chain (block) or in side chains (graft):

```
–AAABBBAAA–        Block copolymer
–AAAAAAAAA–        Graft copolymer
        B
        B
        B
        B
        B
        |
```

To provide strong adsorption (i.e., anchoring of a solvated polymer sheath) to the surface of the particle, at least one component of the block or graft copolymer must be solvated by the medium and at least one component must be insoluble in the medium and preferably compatible with, or identical to, the polymer particle to be dispersed.

Block and Graft Copolymer Stabilizers in Dispersion Polymerization. A sterically-stabilized, nonaqueous, polymer dispersion is made simply by heating a solution of a free radical initiator (e.g., azobisisobutyronitrile), an appropriate monomer, and a suitable block or graft copolymer in an organic liquid which is a nonsolvent for the polymer product and acts as a diluent for the dispersion. The block or graft copolymer acts as a stabilizer for the dispersed polymer obtained by free radical polymerization of the monomer. Although both block and graft polymers can act as stabilizers, graft polymers are normally used because they are more easily synthesized.

Although dispersion polymerization is similar to conventional aqueous emulsion polymerization, there are some crucial differences. In dispersion polymerization, the monomer is usually soluble in the organic liquid in which the polymer dispersion is to be made; a surfactant is therefore not necessary to maintain the monomer in emulsion form. Since the monomer is in solution, agitation of the polymerizing system is of secondary importance. Furthermore, in contrast to emulsion polymerization, the initiator fragments do not stabilize the resulting dispersion. The initiator to monomer ratio can therefore be used to control the molecular mass of the polymer without affecting the particle size of the dispersion [270].

In dispersion polymerization, sufficient graft copolymer must be present to satisfy the increase in particle surface area as polymerization proceeds. The graft copolymer can be either made in situ as a side reaction during dispersion polymerization, or it may be premade and added as required. The side reaction for the formation of the graft copolymer occurs in two ways:

1. Hydrogen abstraction grafting to a simple polymer dissolved in the diluent for the dispersion [271]
2. Copolymerization with a dissolved polymer carrying a copolymerizable residue (stabilizer precursor) [272]

In *hydrogen abstraction grafting*, the initiator abstracts a hydrogen atom from the main dissolved polymer. The graft copolymer is then formed by free-radical copolymerization of an appropriate monomer:

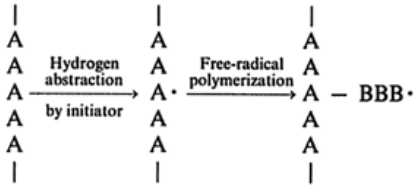

For optimization of the hydrogen abstraction grafting process, the soluble polymer must be carefully chosen and should be used with an initiator that generates very reactive free radicals (e.g., benzoyl peroxide). Adventitious hydrogen abstraction and grafting can also occur. Many dispersions, in which steric stabilization is claimed to arise from the adsorption of simple soluble polymers, may thus actually be stabilized by graft copolymers formed in situ during the preparation of the dispersions [273].

Although the hydrogen abstraction process is a useful technique for many practical purposes, it is not easy to control. Control is considerably improved when a diluent-soluble polymer carrying one or a few copolymerizable groups, often known as a *stabilizer precursor*, is used under conditions that minimize hydrogen abstraction grafting (e.g., use of initiators such as azobisisobutyronitrile that produce moderately reactive free radicals).

Stabilizer precursors can be made by copolymerization of a monomer that contains a condensable group. The copolymer is then reacted with a monomer containing a complementary group. For example, lauryl methacrylate may be copolymerized with hydroxyethyl methacrylate. The resulting copolymer contains hydroxyl groups that can be esterified with methacrylic acid, thereby introducing copolymerizable vinyl groups:

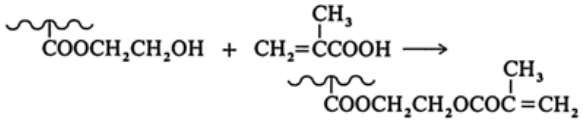

A more convenient way is to use the addition reaction of a carboxylic acid and a polymer containing an epoxide group [272]:

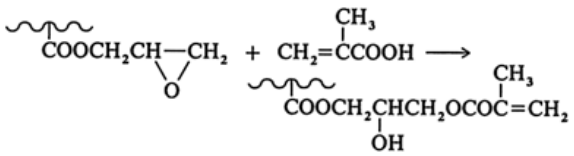

An alternative type of stabilizer precursor can be made by polymerizing a monomer with an initiator (e.g., azobiscyanovaleric acid) in the presence of a transfer agent (e.g., thioglycolic acid); both initiator and transfer agent have the same condensable group (in the above case a carboxylic acid group) [274], [275]. Under suitable conditions, > 95 % of the resulting polymer chains carry a single terminal condensable group. When the polymer is reacted with a monomer containing an appropriate group (e.g., a carboxyl-terminated polymer with glycidyl methacrylate), a monofunctional stabilizer precursor is produced (i.e., it has a single terminal copolymerizable group).

The development of monofunctional precursors led to the production of graft copolymers that were fairly pure, in that they were relatively free from ungrafted polymer. Monofunctional precursors can also be made by condensation [275]. Indeed, graft copolymers made by the copolymerization of monomers with the reaction product of glycidyl methacrylate and the self-condensate of 12-hydroxystearic acid [276] have been used extensively in dispersion polymerizations, not only for technological applications, but also to study the mechanisms of dispersion polymerization and steric stabilization.

The emphasis so far has been placed on processes which, by their random nature, lead to polydispersity. For some fundamental studies and technological applications, block copolymers ionic initiation and “living polymer” techniques have been used [277], [278]. The growing polymer chains remain active until they are deliberately quenched. These processes yield polymers that are monodisperse with respect to molecular mass. Measurements made on dispersions, sterically-stabilized with almost monodisperse block copolymers, yield results which are more readily interpreted.

Stabilizer precursors can be used in conventional dispersion polymerization only because the starting reaction mixture is homogeneous. Some nonaqueous dispersions, however, are made by processes in which one or more of the reactants is/are insoluble in the liquid diluent (e.g., polycondensation) [279]. In these cases, preformed graft copolymers must be used because they function not only as stabilizers for the final dispersion, but also as dispersants or emulsifiers for

the starting materials.

In order to maintain stability, the solvated sheath must remain anchored to the particle surface. In some circumstances (e.g., exposure to strong solvents which tends to swell the particles), anchoring can fail during preparation and use of the resulting dispersion. This can be avoided by including in the anchor component of the block or graft copolymer a group which is capable of coreacting with the polymer being formed in the polymerization or condensation process as the particles are being made [280]. The solvated sheath on the particles can then only be removed by totally disrupting the particles.

Characterization. The solvated sheath which provides the steric stabilization of nonaqueous dispersions has been extensively studied. Hydrodynamic techniques have been used to measure the thickness of the sheath [281], [282], low angle neutron scattering to measure the polymer segment distribution within the sheath [283], and compression experiments to measure the strength of the sheath – sheath interaction [266], [284]. The importance of the solvation power of the medium for steric stabilization has been demonstrated by flocculation experiments [285]. Flocculation is observed as the solvent approaches the "theta" condition for the unattached polymer chains, both by using solvent mixtures and by change of temperature. The theta condition is the state at which the configurational entropy losses caused by constraints on the polymer molecule are balanced by an increase in energy associated with polymer segment association.

Flocculation has also been observed when the diluent of the dispersion is a good solvent for the chains constituting the steric barrier. This is observed when free polymer, even when it is identical to that constituting the steric barrier, is dissolved in the diluent of the dispersion [286]. The magnitude of flocculation is a function of both the molecular mass and the concentration of the added polymer. The depletion flocculation theory suggests that this phenomenon is related primarily to the perturbation of the free polymer chains in solution rather than those in the solvated sheath [287].

The anchoring of the solvated sheath to the surface of the polymer particles has received relatively little attention. At least one component of a graft or block copolymer must be insoluble in the diluent for the dispersion. Preferably, the anchor component should be identical to or compatible with the polymer constituting the particles. There is some evidence that anchor polymer segments which are incompatible with particle polymers tend to clump at the surface of the particles [288], [289].

Similarly, the structure of polymer particles produced by dispersion polymerization has not been extensively studied. The kinetics of classic dispersion polymerization imply that particle nucleation is followed by gel-phase polymerization within the particles. This suggests that the resulting particles should have a fairly homogeneous structure [290], [291]. There is evidence, however, that in some circumstances (e.g., when the monomer is not a good solvent for the resulting polymer), polymerization occurs predominantly in solution. The polymer then precipitates out and is swept onto the surface of existing particles leading to a more heterogeneous structure [292]. Fluorescence quenching studies on dispersions made from labeled reactants have confirmed the heterogenicity of particles made by some dispersion polymerization procedures [293], [294].

Properties of Nonaqueous Dispersions. The stabilizing solvated sheath of nonaqueous dispersions is of limited dimensions. It therefore only has a significant effect on the effective volume fraction, and hence the rheology of the dispersions, when the dispersed particle size is very small and/or the volume fraction of the dispersion is very high. Dispersions of high molecular mass polymers have a much lower viscosity than their solutions at the same temperature.

Dispersions with uniform particle size have been used to model the rheological behavior of dispersions of noninteracting hard spheres [295]. Liquid dispersions with the appropriate particle size distribution have been made with a very high volume fraction [296].

Uses. Nonaqueous polymer dispersions have been used as the main film former in thermoplastic and, more sucessfully, in thermosetting automotive coating compositions [297]. The systems have the advantage of a higher solids content at application and hence lower emission of volatile organic solvents. Cross-linked particles when swollen with solvent have been used as rheological control agents to improve the application of spray coatings in which the main film former is solventborne [298-300]. Discrete polymer particles in a continuous matrix have also been used to improve the properties of the final paint films [301], [302]. Nonaqueous dispersions have proved convenient for the production of polymer in a form suitable for powder coating applications [303].

More generally, nonaqueous dispersions have been used as components of polishes [304], adhesives [305], [306], fiber coatings [307], textile impregnating agents [308], printing inks [309], [310], and electrophotographic toners [311], [312].

Finally, the preparation of nonaqueous polymer dispersions has been used as a technique for polymer production. Process advantages such as the use of lower temperatures for polyamide and polyester synthesis can be obtained [279]. Products such as synthetic elastomers can be obtained in powder form [313]. The fine particulate form of dispersions of water-soluble polymers (e.g., polyacrylamide) offers a convenient route to aqueous solutions of polymeric flocculants and thickeners [314].

3.7. Radiation-Curing Systems

3.7.1. *Introduction*

Radiation curing is a relatively new technology which uses electromagnetic (mainly UV) or ionizing (mainly accelerated electrons) radiation to initiate a chain reaction in which mixtures of polyfunctional compounds are transformed into a cross-linked polymer network.

Resin Systems. Radiation-curable resin systems can be classified as follows:

1. Polyester – styrene mixtures
2. Products with methacrylic groups
3. Thiol – thiene mixtures
4. Prepolymers and monomers with acrylic functionalities
5. Epoxy resins

The polymerization of all these systems can be divided into initiation, propagation, and termination steps. The initiation step involves generation of a reactive species (free radical or acid). During irradiation with UV light, the reactive species are formed by chemical decomposition of a photoinitiator. In electron-beam (EB) curing, reactive species (radicals) are generated by interaction of accelerated electrons with organic compounds.

In systems 1 – 4 radical polymerization takes place. Decomposition of the photoinitiator (e.g., benzoin and its derivatives) leads to formation of free radicals which react with the carbon – carbon double bonds. In system 5 acidic reactive species are formed which react with cycloaliphatic epoxy compounds and vinyl ethers to form a cross-linked network. Diazonium, sulfonium, and iodonium salts are used to initiate this cationic polymerization, a typical example is $Ph_3S^+PF_6^-$.

Polyester – styrene resins have a low raw material price and are used in wood finishing. Disadvantages are the presence of volatile styrene and the low cure speed. *Products with methacrylic groups* are used where line (cure) speed is less important, mainly in combination with acrylates in special end uses (e.g., in the electronics industry) to improve adhesion or to obtain specific physical properties. *Thiol – thiene mixtures* can lead to very flexible cured films but their odor seriously limits their use. *Epoxy resins* are used in metal finishing and can coatings mainly because of their high adhesion.

Prepolymers and monomers bearing acrylic functionalities still dominate the radiation curing market and are discussed in more detail here.

3.7.2. *Radiation-Curable Systems Based on Acrylates*

A UV-curable acrylate system usually contains four main components: prepolymer(s), monomer(s), photoinitiator(s), and additives. A fully formulated EB-curable system is not merely a UV system without photoinitiator but the same principles apply and the same raw materials are used.

Prepolymers. A broad range of acrylated resins (oligomers) are commercially available. The film-forming properties depend on the oligomer system. One of the most common is the *acrylated epoxy* system. In *acrylated urethanes*, an isocyanate-functional prepolymer with a polyol backbone can be reacted with a hydroxy-functional monomer (e.g., hydroxyethyl or hydroxypropyl acrylate). Many different resins can be synthesized by varying the polyol backbone, the isocyanate type, and the hydroxy-functional monomer. *Polyester acrylates* are another example of commercially important prepolymers. *Acrylated acrylics* have an acrylic backbone with pendant acrylate functionality.

Selection of the prepolymer is guided by the desired film-forming properties. For example, acrylated epoxy resins typically provide hard, solvent-resistant films with fast cure at relatively low cost. Acrylated urethanes may be chosen for their excellent film flexibility and toughness. Aliphatic urethanes are used in preference to aromatic urethanes for stringent exterior weatherability requirements. The acrylated acrylics offer excellent nonyellowing properties when a soft

film can be used.

Polyester acrylates are used in a wide range of applications from overprinting varnishes to lithographic (offset) inks.

Monomers. In radiation-curable systems, the solvent is replaced by reactive diluents (monomers) which are incorporated in the network during cross-linking. These monomers have two important functions: they reduce the viscosity of the mixture and they strongly influence the physical and chemical properties of the final coating. Monomers can be divided into three groups:

1. *Monofunctional monomers:* improve the flexibility of the end product (e.g., isobornyl acrylate, *N*-vinylpyrrolidone).
2. *Difunctional monomers:* e.g., hexanediol diacrylate (HDDA), tripropylene glycol diacrylate (TPGDA).
3. *Tri- or tetrafunctional monomers:* increase the cross-linking density of the final coating. Examples include trimethylolpropane triacrylate (TMPTA), pentaerythritol tri- or tetraacrylate (PETA).

Additives can be extremely important in high-performance radiation-curable formulations. Pigments for inks or coatings, fillers for cost control or viscosity control, defoamers and wetting agents, flatting agents and slip aids can all play essential roles in UV/EB inks and coatings. Reactive additives such as surfactants or slip aids to avoid migration in the final coating.

Photoinitiators are essential for UV-curable systems and can be classified into two groups:

1. Photoinitiators that produce radicals by intermolecular abstraction of hydrogen (e.g., benzophenone, thioxanthone). These products require a co-initiator (e.g., tertiary amines, acrylated amines).
2. Photoinitiators that produce radicals by intramolecular photocleavage (e.g., benzoin ethers).

Radical polymerization is inhibited by oxygen, leading to tacky surfaces. This can be overcome in UV-curable formulations through the use of amines (co-initiator) and in EB formulations by working under an inert atmosphere.

Advantages. An important ecological advantage is that only reactive products are used; this means that there is no solvent recovery problem and no pollution. Energy consumption is low. Curing is performed at room temperature so heat-sensitive substrates can be used.

Perhaps the most important advantage is the economic aspect. Curing is done at high speed, manipulations and installation are easy, and only minimal floor space and manpower are needed. A further advantage is that the finished products are of high quality.

Disadvantages. Raw material costs are high but taking into account solvent recovery costs for solvent-based systems this is changing in favor of radiation curing. The products used are potentially hazardous (irritant) and special precautions must be taken for safe handling. Because of the extremely high reactivity and fast cure, there is a tendency for shrinkage leading to adhesion problems on certain substrates.

3.7.3. *Equipment*

Radiation curing requires only a fraction of a second exposure to the UV or EB source.

UV Curing. The UV lamps are usually medium-pressure mercury lamps or electrodeless gas-filled lamps. Cooling is necessary to avoid overheating of the lamp which would lead to decreased lifetime and changes in the spectral output.

Semielliptical reflectors are commonly used to concentrate the light and cure coatings applied to a moving flat substrate. Parabolic reflectors provide a parallel beam of light and are mainly used to cure coatings on uneven substrates.

EB Curing. Electron-beam accelerators (150 – 300 kV) used in finishing plants are of two types:

1. *Scanner type:* electrons are generated from a point cathode in high vacuum. They are then focused, accelerated, and scanned along the scan horn whose length is adjusted to the coating width.
2. *Linear cathode type:* electrons are emitted from suspended wire cathode(s). Contrary to the scanner-type accelerators, electrons are not focussed but are emitted over the total length of the wire(s).

3.7.4. *Fields of Application*

Radiation curing is used for protective coatings, inks, adhesives, and in other minor areas.

Coatings are either clear or pigmented. Coatings for paper include clearcoats for laminated paper on pressboard (wood-grain papers to simulate natural wood). Inks and overprint varnishes for the graphic art industry are a large, rapidly growing segment. High-gloss overprint varnishes for magazine covers, record jackets, and other consumer items are also often radiation-cured.

Plastic coatings for interior and exterior applications are also an important use segment. Important factors are cost, adhesion, weatherability, and availability of raw materials.

The coating of wood is an important use of radiation curing. Three-dimensional curing is very difficult and requires spray application, good control of film thickness, and well-engineered curing lines to assure complete cure of coated objects (e.g., chairs). UV-curable coatings are also used for cork in Spain and Portugal.

Adhesion to metal is sometimes difficult to achieve. However, due to the development of adhesion promoters industrial applications on metal exist. Electronics applications are important; photoresists (both wet and dry film), solder masks, potting compounds, and conformal coatings are products based on UV-curable materials.

Optical fibers have been coated with a protective layer of UV-curable materials for a long time.

Inks. Radiation-curable inks are applied to metal, paper, wood, and plastics (⟶ Printing Inks). Lithographic (offset) and screen printing inks are the most important printing inks. Flexographic inks are known but it is difficult to achieve the low viscosity required for this printing technique while maintaining the good cured ink requirements. Intaglio inks (special inks used to avoid counterfeiting of items such as bank notes) are also employed and ink-jet printing with UV-curable materials is starting.

Adhesives are well established, mostly in lamination. Some UV-curable pressure-sensitive adhesives are used in the automotive industry.

Other uses include the production of flexographic printing plates from UV-curable materials. An interesting application is three-dimensional modeling (stereolithography) in which a solid part is made from a vat of UV-curable liquor by the use of UV lasers controlled by computers using CAD/CAM software.

Release coatings (casting paper, caul sheet, labels) are manufactured by EB curing. Other new applications of radiation curing include binders for magnetic media, leather coatings, and paper upgrading.

3.8. Electrodeposition Paints

Electrodeposition (electrophoretic) coating was developed in the 1960s as a method of priming electrically conducting bulk articles for industrial use [315]. A high application efficiency (> 95 %) can be achieved by using ultrafiltration techniques [316] and has promoted the use of electrodeposition. In 1989 ca. 90 % of all automobiles and many industrial goods were primed by electrodeposition. The market volume for electrodeposition paints in Europe in 1988 was estimated to be ca. 90 000 t/a, corresponding to a turnover of ca. 470×10^6 DM. The increase in sales in the Federal Republic of Germany in 1989 was ca. 8 % [317].

Properties. Electrodeposition paints are aqueous dispersions of organic polymers which are used to apply organic coatings to electrically conducting substrates by passage of direct current [318-320]. When direct current flows between two electrodes, charged dispersion particles are deposited on one of the electrodes

in a diffusion-controlled boundary layer after an induction time t_i. Coagulation of the particles occurs as a result of the change in pH value caused by the electrolysis of water [321]. If the surface charge of the dispersion particles is positive, coagulation occurs at the cathode [322]; if the charge is negative, coagulation occurs at the anode [323]. The deposited polymer particles consolidate due to electroosmosis and rapidly produce a high electrical resistance that slows down deposition [324]. This behavior results in a uniform coating, even of parts with complicated shapes (e.g., automobile bodies). The surfaces to be coated are pretreated with inorganic conversion layers (usually zinc phosphate) to improve adhesion of the electrocoat and corrosion protection [325].

The compositions of cathodic and anodic electrodeposition baths are given in Table 17. The main components are resins, pigments, extenders, some solvent, and neutralization agents. The latter are used to neutralize the charge on the resins so that they become water dispersible. The application data affect the plant and equipment technology (e.g., choice of construction materials, energy consumption, and rectifier design) [318], [326].

Table 17. Typical bath and application data of electrodeposition paints

Parameter	Cathodic	Anodic
Solid (2 h at 130 °C), wt %	15 – 25	8 – 15
Ash (450 °C), wt %	1 – 5	2 – 6
Solvent, wt %	1 – 4	4 – 10
pH	4 – 7	7 – 9
Specific bath conductivity (20 °C), mS/cm	0.8 – 1.5	0.8 – 1.5
Voltage, V	250 – 450	150 – 350
Maximum current density, mA/cm^2	4 – 6	4 – 6
Bath temperature, °C	25 – 35	20 – 28
Coating time, s	120 – 240	120 – 240
Deposition equivalent, mg/C	30 – 40	60 – 80

Electrodeposition tanks range in size from 0.2 to 400 m^3 depending on the objects being coated and the throughput rate. Tanks may be operated batchwise or continuously. The former are mainly used for truck bodies and small mass-produced parts, the latter for high throughputs as in car body priming. In addition to the coating tank, the most important parts of the unit are the counter electrodes, rectifier, circulating pumps, heat exchanger, filters, rinsing zones, and ultrafiltration units (see Section Dipping) [318].

The non-water-soluble resins are converted into a soluble (i.e., ionized) state by using neutralizing agents (e.g., acetic acid). In order to maintain constant bath composition and thus constant coating quality, two compensation methods are used for feeding the baths. In the first method the neutralizing agent released during the deposition period is used to disperse underneutralized, often one-pack feed material in the bath. The solids content of the bath is, however, usually kept constant by using fully neutralized, uniformly predispersed material. Excess neutralizing agent must then be removed through a circulation system at the counter electrodes with the help of ion-exchange membranes [327].

Binders. The first important *anodic electrodeposition paints* were formulated with unsaturated oils that were functionalized with maleic anhydride. Reaction with maleic anhydride was followed by semiesterification of the succinic anhydride derivatives [328]. These resins are cross-linked by oxidative radical polymerization via their residual double bonds. Specially developed polybutadiene oils subsequently replaced natural oils as raw materials. Unsaturated polyesters, saturated polyesters, and alkyd resins can be dispersed in water at an acid index of 40 – 80 mg KOH/100 g solids after neutralization with amines, and can be used as anodic electrodeposition paints. Acrylic copolymers form an important class of binders for one-coat, weather-resistant coatings. These electrodeposition paint binders can easily be synthesized by copolymerizing acrylic acid with unsaturated carboxylate esters and/or vinylogous compounds.

Cathodic electrodeposition paints important for corrosion-resistant priming coats are based on modified epoxy resins that contain amino groups. Dispersion is obtained by neutralizing the amino groups with organic acids [329], [330]. Acrylic resins modified with amino groups can also be used for cathodic one-coat paints. Polybutadienes can be modified for cathodic deposition (e.g., by amidation with diamines after reaction with maleic anhydride). Blocked isocyanates are mainly used as cross-linking agents [331], [332]. The properties of cathodic electrodeposited coatings are listed in Table 18.

Table 18. Properties of cathodic electrodeposited coatings [333]

Property or test	Value
Stoving conditions	20 min, 165 – 185 °C
Layer thickness, mm	0.015 – 0.035
Corrosion protection on pretreated metal sheets	
Salt spray mist test (DIN 50021, 480 h), mm	0.3 (subsurface migration)
Alternating climatic test (VDA 621-415, 12 cycles), mm	1.8 (at the crack)
Draw behavior (DIN 53/56), mm	4 – 8

Conventional *paint additives* (e.g., cross-linking catalysts, corrosion inhibitors, and surfactants) are used to complete the formulation of commercially employed electrodeposition paints. Dispersion mixtures for modifying film properties have also been described [333].

Environmental Protection. On account of the high paint transfer efficiency and low solvent content, environmental pollution in electrodeposition coating is much lower than with spray coating. The solvent contents of electrodeposition paints relative to the solids content are 5 – 10 % in comparison with solventborne spray paints with a solvent content of 50 – 100 %. The most important sources of emission are the electrodeposition tank, the rinsing zone, the stoving oven, and the drain of the ultrafiltrate [334]. Developments in new electrodeposition paints aim to reduce the organic volatile solvent content, to have fewer condensation products in the stoving ovens, and to have lead-free formulations [335].

Uses. Electrodeposition coatings are used primarily for corrosion protection of steel (e.g., in automobiles, agricultural machinery, steel furniture, and appliances). New areas of application include coil coating primers [336], beverage can coatings [337], and photocurable electrodeposition paints as photoresists [338].

[Top of Page]

4. Pigments and Extenders

Klaus Köhler, Peter Simmendinger, Wolfgang Roelle

Pigments, dyes, and extenders are defined in DIN 55 943 under the heading “colorants” (coloring materials). Further standards are DIN 55 944 and 55 949. Similar definitions are given in ASTM D 16 – 84, ISO 4617, and ISO 4618. A *pigment* is defined as a substance that is insoluble in the application medium (e.g., a paint) and is used as a colorant or on account of its corrosion-inhibiting or magnetic properties. A *dye* is defined as a colorant that is soluble in the application medium.

Extenders are sometimes difficult to distinguish from pigments; they are defined as substances that are insoluble in the application medium and are used to increase the volume, obtain certain technical properties, and/or improve optical properties (see also Section Introduction).

Pigments, dyes, and extenders are incorporated in binders to obtain decorative effects (e.g., color, gloss, dulling) or functional effects (e.g., corrosion protection) on a surface coating. For optimization of these effects the colorants must be mixed in a suitable manner and in appropriate amounts with the relevant binder. The distribution of the primary pigment particles in the polymer matrix should be as uniform as possible.

4.1. Inorganic Pigments [339-347]

In this section the emphasis is on inorganic pigments for topcoats and finishing coats. Functional pigments (e.g., for corrosion protection) are discussed to some extent in more detail in Section Coating Systems for Corrosion Protection of Large Steel Constructions (Heavy-Duty Coatings). The particle size of inorganic pigments used in paints is 0.05 – 10 µm, that of transparent pigments 0.01 – 0.05 µm. The pigment volume concentration in gloss paints is 10 – 35 % but may reach >80 % in matt emulsion paints. For a detailed discussion of inorganic pigments, see ⟶ Pigments, Inorganic.

White Pigments. The most important white pigment for surface coatings is titanium dioxide [13463-67-7] with a worldwide consumption in 1989 of ca. 1.8×10^6 t (60 % of total consumption). World wide consumption in 1995 is ca. 3.1 x 10^6t [588]. Of the three known crystal modifications, only rutile and anatase are produced synthetically and utilized on a commercial scale. The brookite modification has no economic importance. The light scattering capacity of titanium dioxide is superior to all other white pigments due to its high refractive index; pigmented coatings therefore have an extremely good hiding power. When combined with colored pigments, titanium dioxide exhibits the highest brightening power of all industrially produced white pigments. On account of these unique optical properties, its toxicological innocuousness, and its chemical resistance, titanium dioxide has largely replaced all other white pigments.

Doping the crystal lattice and coating the surface with metal oxides or hydroxides has reduced the photocatalytic polymer decomposition caused by titanium dioxide to such an extent that all stability requirements can be met. Since anatase is generally more photoactive than rutile, aftertreated and/or lattice-stabilized rutile pigments are preferred for exterior applications (exposure to sunlight).

Commercial products include Bayertitan (Bayer); Finntitan (Kemira Oy, Vuorikemia); Hombitan (Sachtleben); Kronos (Kronos Titan); Tibras (Titanio do Brasil); Tiofine (TDF Tiofine BV); Tiona (SCM Chem.); Tioxide (Tioxide Group); Tipaque (Ishihara S.K.); Tripure (Du Port); Titafrance (Thann et Mulhouse); and Tronox (Kerr McGee).

White pigments of historic importance or less economic importance than titanium dioxide include zinc sulfide, lithopone (mixtures of ZnS and $BaSO_4$), zinc oxide/zinc white (ZnO produced from metallic zinc by oxidation of metal vapors), and white lead (basic lead carbonate, $2\ PbCO_3 \cdot Pb(OH)_2$). On account of their specail properties these pigments have been able to compete against titanium dioxide in some applications. Special properties of zinc sulfide and lithopone are their low absorption in the near UV range; they can therefore be used to color UV-curable paints. Lithopone-pigmented putties can easily be sanded and have favorable rheological properties. The luster of lithopone is used to advantage in wallpaper coatings. Zinc sulfide and white lead are sometimes used in coating materials because of their fungicidal properties and to neutralize acid to protect against corrosion. Also, the lower abrasivity in comparison to TiO_2 pigments can be of importance.

Iron Oxide Pigments. Iron oxide pigments may be natural or synthetic, the synthetic oxides have a wider color range and more uniform quality. Worldwide consumption of synthetic iron oxides in 1989 was 500 000 t, coating materials accounted for about 185 000 t. The color range includes the basic colors red (α-Fe_2O_3, hematite), black (Fe_3O_4, magnetite), yellow (α-FeOOH, goethite), and orange (γ-FeOOH, lepidocrocite). Mixtures of these colors are used as brown pigments; the γ-Fe_2O_3 (maghemite) normally used for magnetic recording purposes can, however, also be used to impart brown coloration. Manganese- and chromium-containing iron oxide phases are important as heat-stable brown pigments (e.g., for stoving enamels).

Iron oxide pigments are extremely important in coating materials due to their outstanding hiding power, excellent lightfastness and weather resistance, insolubility in water and organic solvents, resistance to alkalis, and their toxicological inertness. Only micronized pigments should be used in paints because they have better dispersion properties. Pigments surface-coated with metal oxides (e.g., Al_2O_3) frequently exhibit good flocculation stability.

Commercial products include Bayferrox (Bayer) and Deanox (Deanshanger Oxide Works). Other manufacturers include S.I.L.O., Columbian Chemical Company, and Mobay.

Chromium Oxide Pigments are composed of chromium(III) oxide [1308-38-9], Cr_2O_3, crystallized in a corundum lattice. This crystal structure impact has a certain degree of hardness and abrasivensss. On account of their excellent colorfastness and insolubility in solvents, acid, and alkali, chromium oxide pigments have successfully competed against more brilliant organic green pigments. Hydrated chromium oxide pigments are less important. Pure chromium oxide pigments are not toxic. About half of the consumption of chromium oxide pigments in the Western World (total production, 48 000 t in 1995) was used for coating materials.

Manufacturers include Bayer (green chromium oxide) and British Chrome and Chemicals and Nikon Kagagu (chromium oxide).

Mixed Metal Oxide Pigments. This term usually refers to metal oxide pigments that are derived from spinel ($MgAl_2O_4$), rutile (TiO_2), hematite (α-Fe_2O_3), or bixbyite (α-Mn_2O_3) structures. Substitution of metal ions in the host lattice by other chromophoric metal ions opens up a broad color spectrum. The oxidic mixed-phase pigments generally exhibit outstanding lightfastness and resistance to weathering, heat, and chemicals. Heavy-metal ions can only exect their toxic effects in dissolved form. Since these pigments are sparingly soluble, they may be classified as toxicologically innocuous. Around half the production capacity of ca. 11 000 t (1988) was processed into coating materials.

Commercial products include Heucodur (Dr. H. Heubach); Irgacolor (Ciba-Geigy); Kerafast (Blythe Colours); lightfast pigments (Bayer); mixed-phase pigments (Degussa, Ferro B. V., Ferro Corporation, Shephard Chemical Corporation); and Sicopal; Sicotan (BASF Lacke und Farben).

Cadmium pigments may be regarded as mixed-phase pigments, and are derived from the wurtzite lattice of cadmium sulfide (cadmium yellow). Partial substitution of cadmium by zinc shifts the color of cadmium yellow to greenish shades. Replacement of the sulfide ions with selenium yields orange to dark red pigments (cadmium red). Of all inorganic pigments, cadmium pigments exhibit the most brilliant shades; they have excellent hiding power and also a much higher heat stability than organic pigments and are completely stable against bleeding. Public discussion concerning environmental problems has lead to a sharp fall in the consumption of cadmium pigments in paints. In the Federal Republic of Germany, only ca. 22 t of cadmium pigments were used in paints in 1986.

Commercial products include Languedomer (Société Languedocienne de Micron Couleur); and Sicotherm (BASF Lacke und Farben). Other manufacturers include Blythe Colours, Brown, Degussa, Ferro Corporation, Harshaw Chemical Company, and Reckitts Colours. In 1987 Bayer stopped production and marketing of their Cadmopur pigments.

Bismuth pigments are based on bismuth orthovanadate [14059-33-7], $BiVO_4$ or produced as the two-phase system $BiVO_4$ – Bi_2MoO_6 to improve their colorfastness. The colors of this group of pigments are similar to that of cadmium yellow.

Commercial products include Sicopal Yellow (BASF Lacke und Farben).

Chromate and molybdate pigments are generally mixed-phase pigments from the system $Pb(Cr, S, Mo)O_4$. They include chromium yellow (molybdenum-free) and molybdate orange and red. Variations in shades are possible by adjusting the crystal modification (orthorhombic or monoclinic) or proportion of free lead sulfate. Pigments from the system $PbCrO_4 \cdot PbO$ should also be classified in this group (chromium orange and chromium red). The chromate and molybdate pigments have a high tinctorial strength combined with good hiding power and brilliance. Depending on the nature and degree of stabilization, products are classified as standard, lightfast, or SO_2-resistant. Chromate and molybdate pigments are classed as harmful to health and must be identified. A considerable proportion (>60 %) of the total world production of chromate and molybdate pigments (130 000 t in 1988) is used in coating materials.

Commercial products include Heucotron (Dr. H. Heubach) and Sicomin (BASF Lacke und Farben). Other manufacturers include Burrell Colours, Ciba-Geigy, Ten Horn B.V., ETS, and Capelle Frères).

Ultramarine pigments are sodium aluminum silicates of the composition $Na_8Al_6Si_6O_{24} \cdot S_x$ (Na-rich) or $Na_{8-y}Al_{6-y}Si_{6+y}O \cdot S_x$ (Si-rich). Small sulfur-containing anions (e.g., S_3^- and/or S_2^-) are bound as chromophores in the interstices of the crystal lattice to equalize the charge. Depending on composition, blue, red, green, or violet pigments can be obtained. Ultramarine pigments have a high heat resistance but their universal use in paint systems is limited on account of their poor hiding power and limited weather resistance. Special types coated with silica must be used when resistance to acids is required. Ultramarine pigments are regarded as being physiologically innocuous. The worldwide ultramarine production capacity is ca. 30 000 t.

Manufacturers in Western Europe include Reckitts Colours, UK.

Blue iron pigments have the general composition M(I)Fe(III) $(CN)_6 \cdot H_2O$, where M(I) is Na, K, or NH_4. These pigments have a very high tinctorial strength, but are difficult to disperse on account of their fine state of division and tendency to agglomerate. Micronized types behave more favorably on dispersion. Thermal resistance (up to ca. 180 °C) is generally sufficient for stoving lacquers. Stability to dilute acid is good, although there are problems in the alkaline pH range. These pigments can be used in coatings for exterior use due to their lightfastness and weather resistance. Their main area of application is, however, the printing ink sector. Blue iron pigments are toxicologically innocuous. Total world production of blue iron pigments was ca. 50 000 t in 1988, a considerable proportion of which was used in paints and, in particular, printing inks.

Commercial products include Miloriblau (ETS, Capelle Frères) and Vossenblau (L. Vossen und Co., Degussa).

Nacreous (pearlescent) pigments and interference pigments generally consist of platelets with a high refractive index (at least in some regions in pigments with a layer structure) and a high transparency. The first such pigments were based on natural products such as fish silver, as well as synthetic products such as bismuth oxychloride, BiOCl, and basic lead carbonate, $Pb(OH)_2 \cdot 2\ PbCO_3$. Layered pigment structures (especially coated mica particles) now dominate, however. The mica surface is coated with a titanium dioxide layer of defined thickness. The photoactivity of the titanium dioxide for exterior applications can be lowered by suitable aftertreatment (see White Pigments.). Variations in the layer thickness produce a range of interference colors. Lowering the titanium dioxide content or additional coating with Fe_2O_3 or Cr_2O_3 (if necessary without a TiO_2 layer) results in further interesting color effects based on combination of interference and selective light absorption. These pigments can also exhibit metallic effects. Other layer structures (e.g., alternating coatings of mica with TiO_2 and Fe_2O_3) also produce metallic effects. Inorganic pigments based on mixed phases of platelets with the composition $Al_xMn_yFe_{2-(x+y)}O_3$ ($x = 0 – 0.2$, $y = 0 – 0.06$) also create metallic or interference effects in coatings.

Special effect pigments based on mica are regarded as toxicologically harmless. Their economic importance has arisen steadily; production capacities amounted to several thousand tonnes in 1989.

Commercial products include Iriodin (E. Merck); Merlin (Mearl Corporation); and Paliochrome (BASF Lacke und Farben).

Metallic effect pigments consist of highly lustrous flakes of nonferrous metals (e.g., aluminum, copper) or alloys (e.g., brass). *Leafing types* undergo special organic surface treatment with fatty acids and collect at the surface of the coating film. *Nonleafing types* are coated with alkylamines and are uniformly distributed in the solid paint film. When the pigments are dispersed in aqueous media, care must be taken to prevent reaction with water on freshly created metal surfaces because this leads to graying; for the same reason care should also be taken when using chlorine-containing solvents. The importance of metallic effect pigments in the decorative sector has risen sharply in recent years (e.g., in the automobile topcoat and finish sector).

Manufacturers in Western Europe include Eckart-Werke (Fürth, FRG) and Silberline Ltd. (UK).

Transparent Pigments. Transparent coatings are obtained when there are slight differences between the refractive indices of the pigment and binder. Pigment particle sizes that are substantially smaller than the wavelength of light must be used. Light scattering under these conditions is very low and light absorption predominates. Many transparent shades can be produced from colored inorganic pigments; examples include α- and γ-FeOOH (yellow), α-Fe_2O_3 (red), α-CrOOH (green), and $CoAl_2O_4$ (blue).

On account of their fine state of division, transparent pigments may cause dispersion and flocculation problems. They are used not only for their decorative effects — their high UV absorption and lightfastness are also exploited to protect the binder matrix and substrate (e.g., wood).

Manufacturers in Western Europe include BASF Lacke und Farben (Stuttgart, FRG).

4.2. Organic Pigments

Organic pigments are used to color and/or cover a substrate. They have gained importance because the use of some inorganic pigments containing heavy metals has been legally restricted.

Organic pigments have a high light absorption and a low scattering power, whereas inorganic pigments have a low light absorption and high scattering power. Combinations of organic and inorganic pigments are therefore often advantageous. Organic pigments are therefore often advantageous. Organic pigments also have a lower density and higher surface area than inorganic pigments; their color purity and tinting strength are often higher. Organic pigments tend to dissolve at high temperature and in binders and solvents (migration).

Organic pigments may be transparent or opaque. Transparent pigments are used in glazes and in combination with inorganic pigments including special-effect pigments.

The most important class of organic pigments in terms of production are the azo pigments. Other important pigment classes include metal-complex pigments (e.g., copper phthalocyanine) and higher polycyclic compounds (e.g., anthraquinone, quinacridone, isoindolinone, and perylene). For a detailed description of these pigments, see ⟶ Pigments, Organic.

Organic pigments are synthesized and then dried to form powders. The powder consists of agglomerates of pigment particles. Breaking up and incorporation of the agglomerate in the paint (dispersion) are very important because they determine the appearance and film properties of the pigmented paint layer. These steps are described in detail in Chapter Production Technology. Pigment dispersions for paint formulation are also commercially available.

Properties. *Lightfastness* and *weather resistance* are the most important properties of paint pigments. The organic pigments should be colorfast and stable towards radiation, heat, and atmospheric substances. Pigments may also be partially responsible for photochemical or thermal degradation of the binder which leads to deterioration of optical film properties (e.g., gloss).

Migration should not occur when organic pigments are used in paints that are applied in differently colored layers on top of each other. Resistance to migration depends on the pigment, binder, and solvent.

The pigment should be *heat stable* because it may be subject to increasing temperature during dispersion, stoving, or use.

Chemical resistance is required for various areas of use (e.g., in automotive finsihes to resist car washes and battery acids and in washing machine coatings to resist detergent solutions). Chemical resistance is also of importance in paint formulation if acids or bases are employed as catalysts or if an aqueous basic or acidic binder is used. Since coatings are permeable to industrial gases (e.g., SO_2 and NO_x), the pigments must be resistant to these gases that can affect the colorfastness of the pigment.

The pigment must also be *compatible with the paint binder*. In chemically drying paints, the pigment may react with the reactive components of the paint system. This can lead to changes in the pigment or in the cross-linking stoichiometry. A change in the cross-linking density can affect the properties of the coating.

Uses in Paints. Most organic pigments on the market are suitable for use in paint systems. Apart from the properties mentioned above, the rheological character of the pigment is important because it affects both dispersion and the paint. Rheological behavior during dispersion determines how much pigment can be dispersed per batch. Favorable rheology (i.e., low viscosity) allows higher pigment and binder concentrations at equal application viscosity. “High-solids” coatings have a low solvent content and are therefore more environmentally acceptable.

Pigments can also influence other coating properties (e.g., film gloss) as a result of complex interactions between the pigment, binder, and solvent, and the method of application.

The choice of pigment depends on the required color, binder, and fastness. Pigment manufacturers have therefore produced pattern cards and information describing their properties and areas of application. The pigment should, however, be tested in the binder to be used and for the particular application.

Commercial products include Cinquasia, Cromophtal, Irgalith, Irgazin, Monastral (USA only) (Ciba-Geigy); Cromofine (Dainichi Seika); Graphtol, Sandorin (Sandoz); Hansa, Hostaperm, Novoperm, Permanent (Hoechst); Helioecht, Indofast, Palomar, Perrindo, Quindo (Bayer, Mobay); Heliogen, Paliogen, Paliotol,

Sico (BASF); Monastral, Monolite (ICI); and Rubicron (Toyo Soda).

Commercially available pigment preparations containing dispersion resins or a dispersant are used when pigments are difficult to disperse in a given binder or if the dispersion stage of production is to be omitted. The pigment preparation is then simply stirred into the paint. The compatability of the dispersion resin or dispersant with the other paint components determines the area of application. Some preparations can be used for a wide variety of uses such as solventborne or waterborne systems. Others have very restricted application areas and can only be used with a few binders and solvents. Trade names and manufacturers include Colanyl, Flexonyl, Hostatint (Hoechst); Encelac, Luconyl, Sicoflush, Sicomix (BASF); and Mikrolith, Unispers (Ciba-Geigy).

Solventborne Systems. The trend towards high solids in solventborne paint systems has made rheology an important factor. Viscosity is affected by the binder, solvent, and pigment. The effect of the pigment on viscosity is reduced by optimization of the particle size and by suitable surface treatment. This is particularly important for organic pigments which tend to have a higher viscosity than inorganic pigments because of their smaller particle size.

The pigment is dispersed in the binder and solvent which wet and stabilize the pigment. Destabilization of the pigment (e.g., reagglomeration, flocculation, and flotation) must not occur when the other paint components are added.

Aqueous Systems. Dispersion of organic pigments in an aqueous phase is difficult because of their hydrophobic character. Surfactants are therefore often used to reduce the surface tension of the water and thereby improve wetting. Suitable treatment of the surface of the pigments can also facilitate wetting. In addition to classical steric stabilization of the pigment, electrostatic stabilization is also possible due to the high dielectric constant of water. Special dispersion resins or surfactants are used to aid dispersion; they must be compatible with the binders and must not alter the paint properties. Commercial pigment preparations are also marketed which contain appropriate dispersion resins and surfactants.

Powder Coatings. Powder coatings do not contain solvents or water, they consist solely of binder and pigment. The pigment is incorporated into the reactive resin (e.g., with an extruder). The particle size required for the application is then obtained by grinding and sieving. Cross-linking of the paints is carried out at high temperature (> 150 °C), hence the heat stability of the organic pigments is very important.

Coil Coatings. Only a few high-grade organic pigments are suitable for use in coil coatings since they must be stable to storing temperatures > 250 °C.

4.3. Extenders

4.3.1. *Introduction*

Definitions. According to DIN 55 943 extenders are powdered substances that are practically insoluble in the application medium and are used to modify the volume, achieve or improve technical properties, and/or change optical properties. DIN 55 945 states furthermore that in white extenders, the refractive index is generally below 1.7; in certain media an extender may also be a colorant (pigment).

Since the refractive indices of most binders are also < 1.7, extenders do not contribute significantly to the hiding power when they are used in an amount below the critical pigment volume concentration (*CPVC*). The hiding power of a dry film depends mainly on the amount of titanium dioxide present, a white pigment which has a high hiding power due to its high refractive index.

Above the *CPVC*, however, dry hiding becomes apparent. Air inclusions are formed in the micropores of the dry film, and consequently the hiding power depends on both the amount of titanium dioxide and the amount of trapped air.

If the refractive index of an extender is greater than 1.7 (e.g., zinc oxide), it can also be termed a pigment, though the transition is not clear-cut. The refractive indices of some important extenders and white pigments are as follows:

Water	1.33		Mica	1.58
Chalk	1.55	⟶ Polymer	Barytes	1.64
		1.5 – 1.6	Zinc oxide	2.06
Kaolin	1.56		Zinc sulfide	2.37
Talc	1.57		Anatase (TiO_2)	2.55
Siliceous earth	1.55		Rutile (TiO_2)	2.75

Classification. Extenders are mainly natural minerals that are converted into a usable form by working natural deposits, separating secondary constituents, and comminution (micronization). Synthetic products such as precipitated carbonates [e.g., calcium carbonicum praecipitatum (ccp), precipitated sulfates (e.g., blanc fixe), precipitated and pyrogenic silica, and silicates are used for optical brightening. Synthetic fibers (generally organic) are used for reinforcement. The classification of extenders used in surface coatings is summarized in Table 19. They are discussed in detail elsewhere (⟶ Aluminum Oxide, ⟶ Calcium Sulfate, ⟶ Magnesium Compounds, ⟶ Mica, ⟶ Silica, ⟶ Silicates, ⟶ Talc).

Table 19. Classification of extenders used in coatings

Types of compound	Natural extenders (ISO 3262)*	Synthetic extenders
Oxides and hydroxides	aluminum oxide (DIN 55628) magnesium oxide	aluminum hydroxide (DIN 55628) magnesium hydroxide
Silica and silicates	quartz (DIN 55926) cristobalite (DIN 55927) kieselguhr (DIN 55920) talc (DIN 55924) kaolin (DIN 55922) siliceous earth (DIN 55920) mica (DIN 55929) wollastonite feldspar plastorite slate	pyrogenic silica (DIN 55921) precipitated silica (DIN 55921) aluminum silicates (DIN 55921) calcined aluminum silicates (DIN 55631)
Carbonates	calcite (DIN 55918) chalk (DIN 55918) dolomite (DIN 55919) magnesite	precipitated calcium carbonate (DIN 55918)
Sulfates	barytes (DIN 55911) gypsum	precipitated barium sulfate (DIN 55911)
Special extenders	cork flour, etc. rock wool	glass beads glass fibers (DIN 60001) polymer fibers

* The DIN and ISO standards define supply specifications.

Economic Aspects. The calcium carbonates (calcites, chalks, limestone powders, dolomites, precipitated calcium carbonates, and surface-treated carbonates) account for 80 – 90 % of the extenders used in Western Europe. In the Federal Republic of Germany calcium carbonates are more predominant than in the remainder of Western Europe (ca. 90 %); calcites with a high degree of whiteness and a mean particle size of 2 – 5 µm are most widely used. Next follow talc, kaolins (with calcined kaolins), silica, barytes, and mica.

Occupational Health. Some mineral extenders with extremely fine particles (dusts) are governed by special legal provisions. The most important regulations in Germany, for example, are [348]:

1. Safety precautions (Unfallverhütungsvorschriften, UVV 26), “Protection against Noxious Mineral Dust” (VBG 119)
2. Technical rules for dangerous substances (Technische Regeln für Gefahrstoffe, TRGS), TRGA 508, (siliceous dust)
3. MAK values, TRGS 900, maximum workplace concentrations and biological tolerance values

Reference should also be made to the use of paints and coating materials for food containers and packaging [349].

4.3.2. *Properties*

Extenders were originally employed for increasing volume (i.e., filling) to reduce the cost of the end product. Sophisticated and modified processing techniques (e.g., micronization, surface treatment, and precipitation) have been used to modify and improve the extenders and have led to the discovery of other important properties (see Table 20).

Table 20. Properties of extenders

Filler	Chemical composition	Occurrence	Particle shape	Density, g/cm^3	Oil number, g linseed oil/100 g extender	Brightness	Specific surface area, m^2/g	pH	Particle size distribution, µm [c]
Calcium carbonate	$CaCO_3$	natural	rhombohedral (calcitic)	2.7	14 – 21	75 – 95	0.5 – 3	9.0 ± 0.5	0.5 – 100 (300)
		natural	microcrystalline	2.7	18 – 22	70 – 90	0.5 – 3	9.0 ± 0.5	0.5 – 100 (300)
		synthetic	aragonite	2.9	25 – 35	>95	7 – 10	9.5 ± 0.5	fiber length 1 – 3 fiber diameter 0.2 – 0.5
		synthetic	rhombohedral (calcitic)	2.7	35	>95	7 – 10	9.5 ± 0.5	fiber length 0.3 – 0.8 fiber diameter 0.2
Dolomite	$CaCO_3$ /$MgCO_3$	natural	rhombohedral	2.8	16 – 25	85 – 92	0.5 – 3	9.8	0.5 – 100 (300)
Talc	Mg hydrosilicate	natural	platelet-like	2.7 – 2.8	30 – 55	75–>90	3 – 12	9.5 ± 0.2	0.2 – 80
China clay	Al silicate	natural	platelet-like	2.6	35 – 65	79 – 92	8 – 20	5 (natural) 7.5 – 9 (primary dispersion)	0.2 – 40
China clay	Al silicate	synthetic calcined	amorphous platelets	2.6	ca. 50	88 – 90	8 – 10	6.5 ± 0.5	0.2 – 40
Siliceous earth	Al silicate/quartz	natural	platelet-like corpuscular	2.6	ca. 45	81 – 90	10 – 15	7 – 8	0.1 – 20
Wollastonite	$CaSiO_3$	natural	fibrous	2.8 – 2.9	25 – 30	85 – 88	2 – 5	9 – 11	3 – 100
Pyrogenic silica	SiO_2	synthetic	amorphous	2.2	280 [a]	>90	100 – 400	4 ± 0.5	0.008 – 0.04 (average particle size)
Quartz	SiO_2	natural	trigonal	2.6	13 – 27	75 – 88	1 – 18 (90)	7 – 8	2.5 – 100 (250)
Cristobalite	SiO_2	synthetic calcined	tetragonal	2.4	23 – 25	85 – 90	ca. 2	8 – 9	2.5 – 100 (250)
Mica	Al silicate	natural	platelet-like	2.7 – 2.9	35 – 75	60 – 85	3 – 12	7.5 – 9.5	1 – 100
Barytes	$BaSO_4$	natural	rhombic	4.0	8 – 10	60 – 90	1 – 3	8 – 9	0.8 – 50
Barytes	$BaSO_4$	synthetic	rhombic	4.0 – 4.2	13 – 20	92 – 95	1 – 4	6 – 7	0.3 – 10
Aluminum hydroxide	$Al(OH)_3$	synthetic	platelet-like	2.4	20 – 40	95 – 97	2 – 10	9 ± 1	0.05 – 100 (250)
Gypsum	$CaSO_4$	natural	platelet-like	2.4 – 2.6	17 – 20	83 – 87	5000 – 6000 [b]	7 ± 0.5	2 – 40
Slate	$AlH(SiO_3)_2$	natural	platelet-like	2.7	30 – 35	28	8000 [b]	9 ± 0.5	30 – 200
Plastorite	Mg, Al, K silicate	natural	platelet-like	2.75	23 – 33	81 – 85	6000 – 14 000 [b]	7.5 ± 0.5	5 – 250
Glass fibers	C-, E-Glass	synthetic	fibrous	2.5					fiber length 0.5 – 10 mm fiber thickness 9 – 12
Glass beads	A-Glass	synthetic	solid, spherical	2.5	16		0.4 – 1.5		0 – 50
Glass beads	A-Glass	synthetic	hollow, spherical	0.2 – 0.4	16		0.4 – 1.5		0 – 50

[a] Oil adsorption (lubrication point according to ASTM D 281).
[b] According to BLAINE, cm^2/g.
[c] This particle size is mostly used in paints, other particle sizes are also possible.

Extenders included in Table 20 as well as their physical characteristic data are described in [350]. Measurement of these properties allows definition and comparison of the extenders. Such properties include:

chemical composition
particle shape and size distribution
density
tamped density and tamped volume
specific surface area
brightness (whiteness)
abrasion
pH
wettability, dispersibility
oil absorption

Chemical Composition (DIN 51 070 and ISO 787, Parts 3, 8, 13, 14, 19). The extenders should be chemically inert in the paint. Secondary constituents (e.g., iron ions) may produce color changes. Sulfates, sulfides, alkali, alkaline-earth, and heavy-metal ions may cause problems due to their chemical reactivity and toxicity, and also affect abrasion and weather resistance.

Since these components mainly occur as water-soluble salts, they lead to problems in the paint film; for example, the paint film effloresces under the influence of humidity. This is important in exterior-use paints because it affects weather and corrosion resistance.

The extender should be acid-resistant (e.g., barytes, talc, mica, china clay, silica, silicates). Minerals with a platelet (lamellar) structure (e.g., talc, mica, and china clay) are particularly suitable for corrosion protection systems since they cover the surface better than particles of other shapes. Better covering makes the paint film less permeable to water vapor, thereby reducing bubble formation and improving film adhesion. New developments have shown that special sorts of calcites result in a remarkable corrosion resistance; this can be explained by the buffering effect when the paint film is exposed to harmful acid atmospheres [351]. Platelet-shaped extenders should not, however, be “contaminated” with differently shaped particles (e.g., talc with carbonates).

Particle Shape and Size Distribution (DIN 66 111, 66 115 – 66 120, 66 141, 66 143 – 66 145, 66 165, ISO 25 591). Particles may have the following shapes (figures denote the aspect ratio) [352]:

Spheres	1/1
Cubes	1/1
Parallelepiped	(1 – 4)/(1 – < 1)
Platelets (lamellae)	1/(1/4 – 1/100)
Fibers	1/< 1/10

Extender producers try to retain the original particle shape during processing because the paint producer does not want to have the platelet shape of kaolin, talc, etc. destroyed during micronization.

Particle size and size distribution are important factors in determining the usability of the extender. The particle size spectrum produced by micronization corresponds to a Gaussian distribution. The size spectrum is numerically characterized by the mean particle size (d_{50}), the fines, an upper limit, and by the screen oversize (⟶ Particle Size Analysis and Characterization of a Classification Process).

A spectrum of particle sizes ensures optimum packing density. Interstices between individual particles can be filled with finer particles that prevent the penetration of binder. Both the manufacturer and user try to ensure that a specified product always has the same particle size distribution.

If, however, a certain extender grade does not have optimum packing density, it can be combined with fine extenders or extenders with different particle shapes.

Lamellar extenders improve coating properties, e.g., rheology (in particular thixotropy by talc and kaolin), matting (talc), dry hiding (talc, kaolin), and adhesion (talc, kaolin).

Density (ISO R 787, Part 10). The densities of all mineral extenders lie in a relatively narrow range (ca. 2.5 – 2.8 g/cm^3). The exception is barytes with a density of ca. 4.0 g/cm^3.

Extenders with a lower density (< 2 g/cm^3) aroused interest for volume packaging (i.e., packaging based on volume rather than weight) in the emulsion paint sector. However, paints produced with low-density extenders had completely different application properties. Further developments have therefore been aimed at modifying emulsion paints to allow the use of more platelet-like extenders (e.g., talc or kaolin) in combination with the commonly used carbonates.

Tamped Density and Tamped Volume (ISO 787, Part II). The *tamped density* is the ratio of mass to volume, while the *tamped volume* is the ratio of volume to mass after tamping under specified conditions. These quantities provide information about the volume requirements of the products.

Specific Surface Area (DIN 66 131, BET method). The activity of extenders (i.e., their influence on the properties of the end product) depends on their surface area and their interaction with, for example, the polymer matrix.

The specific surface area is measured by absorption methods. The BET (Brunauer, Emmet, and Teller) method is widely used for light-colored extenders.

Extenders can be classified as follows (BET surface area in m^2/g):

Inactive	< 10
Semiactive	10 – 50
Active	50 – 100
Highly active	> 100

Brightness and Yellowness (DIN 5033, 53 163). Extenders should be chromatically neutral and should not cause discoloration or fading in the paint. Color, brightness, and yellowness are determined quantitatively (see Section Optical Properties).

Measurement is carried out by directing light on the surface of the test sample. The proportion of light reflected by the sample is determined with a spectrophotometer and compared with that reflected by a standard surface (usually barium sulfate). The brightness is the reflectance obtained with a R 457 filter (λ = 457 nm). The yellowness is obtained by subtracting the brightness from the reflectance measured with a R 57 filter (λ = 570 nm). Generally, measurement methods must be specified by the manufacturer to allow comparison.

Abrasion. The abrasion of extenders depends on their hardness and particle size. A high degree of abrasion is undesirable because it causes wear in machinery and contamination of the blended material with abrasion products. Contamination produces surface defects in the film. Einlehner's method is often used to investigate the influence of the extender on abrasion in machinery. An aqueous slurry of the extender is stirred for 1 h in a special apparatus. A metal sieve is placed in this slurry throughout this period and the weight loss of the sieve is then measured and converted to a specific surface area.

The abrasion of a mineral decreases with the particle size. Abrasion is low if a mineral does not contain any secondary abrasive constituents. Thus, for example, the abrasion of talc increases with the content of secondary constituents such as quartz and calcium carbonate. Abrasion can also be reduced by

subjecting the material to surface treatment, normally with stearic acid or stearates.

pH Value (DIN 53 200; DIN ISO 787, Part 9). The pH value of an extender refers to the pH value of an aqueous suspension of the relevant extender. It is very important for paint manufacturers to know whether an extender contains acidic or alkaline constituents to predict its behavior in an existing formulation.

Wettability, Dispersibility (DIN 53 328; ISO 8780, 8781). The *dispersibility* of a powder is defined as its capacity for being dispersed. Dispersibility depends on the wettability of the substance and on the number and strength of the binding sites between the agglomerates.

The rate of color development and the decrease in granularity may serve as a measure of the dispersibility.

Dispersion is defined as the formation of smaller particles from powder agglomerates in a medium and the simultaneous wetting of the particles by the medium. In a fully dispersed material there is a statistically uniform distribution of the particles throughout the entire medium.

Grinding produces crystal fracture surfaces and increases the surface activity of extenders. Since the surface also bears hydroxyl groups, the ground extenders are generally hydrophilic.

Such extenders are consequently not wetted, or if so only poorly, by aliphatic hydrocarbon solvents. However, the binder takes over this function to some extent because it contains surfactants and wetting agents added by the formulator [353].

In nonaqueous systems the disadvantages of extender hydrophilicity can be overcome by surface treatment (stearic acid and stearates for carbonates, silanes for silicates). Such treatment is not necessary for aqueous systems and may even be damaging.

Wetting and micronization of the agglomerates play important roles in the dispersion process. The resultant dispersion should remain stable over a prolonged period, flocculation is undesirable. For a more detailed discussion of dispersion, see ⟶ Disperse Systems and Dispersants.

Oil Absorption (DIN 53 199, DIN ISO 787, Part 5). The oil number is the amount of linseed oil that is absorbed by a pigment or extender sample under defined conditions. The oil number depends on the nature of the extender, its particle size distribution, packing density, and specific surface area.

Heterodisperse extenders have a smaller binder requirement than homodisperse extenders because their interstices can be filled with fine particles. In homodisperse systems the binder penetrates the interstices.

The wet point in waterborne paints is analogous to the oil number in solventborne paints. It represents the number of milliliters of water (containing 2 % wetting agent) that are required to convert 100 g of extender to a compact mass.

4.3.3. *Modification of Extenders*

Precipitated Extenders. Precipitated extenders such as calcium carbonate (calcium carbonicum praecipitatum), barium sulfate (blanc fixe), and silica have been known for a fairly long time. When natural minerals are micronized, a minimum particle size limit is encountered. Lower particle sizes (< 1 µm) can only be obtained with considerable technical effort and expense. Extenders with finer particle sizes are produced synthetically by precipitation. The structure of one and the same extender can be modified by controlling the precipitation conditions. Particularly pure and thus bright extenders can be produced by a suitable choice of starting substances [350].

Surface-Treated Extenders. Some extenders tend to form agglomerates during storage, or absorb moisture or additives from the paint mixture. This can be prevented by surface treatment. Stearic acid and its salts are used in large amounts to coat calcium carbonates.

In subsequent developments the aim was to chemically bond the extender and substrate, with the surface-treatment agent acting as a bridging element. Silanes have proved particularly outstanding in this respect and a large number of products with widely varying organic groups (e.g., amino, mercapto, vinyl, and methacrylic groups) are available for this purpose. Such products are usually used to treat kaolins and other silicates.

[Top of Page]

5. Paint Additives

Wilfried Scholz, Andreas Valet, Mario Slongo

In addition to resins, solvents, and pigments, paints also contain additives. The additive content is typically between 0.01 and 1 %. Paint additives are used to prevent defects in the coating (e.g., foam bubbles, poor leveling, flocculation, sedimentation) or to impart specific properties to the paint (e.g., better surface slip, flame retardance, UV stability) that are otherwise difficult to achieve.

Such products were formerly termed “auxiliaries” and were often only used to correct a paint batch that did not comply with the required specifications and showed some defects. Nowadays additives are already taken into consideration when a new paint formulation is created; they form an essential constituent of the coating. In view of the increasingly stringent quality and environmental requirements for the production and use of coatings, high-quality paint systems are almost always formulated with additives.

Many effects can be achieved with additives, and not all of their, in some cases, very specific uses can be discussed here. Information on paint additives and their mode of performance is given in the literature and company brochures. There is very little comprehensive literature on additives [354], [355], [356], [357]. The physical and chemical principles of paint production and application are helpful in understanding their mode of action [358-360]. A large number of paint additives are listed according to areas of use in special tables [361, 362]. Today the Internet is a very helpful tool for collecting information and this holds also for additives. All major additive suppliers maintain websites where detailed information about their products can be found. Additionally, these websites often contain background information and guidelines for additive selection.

Additives may be classified in the following groups:

1. Defoamers
2. Wetting and dispersing additives
3. Surface additives
4. Rheology additives
5. Driers and catalysts
6. Preservatives
7. Light stabilizers

This subdivision is by no means exhaustive but includes the major additive groups. Individual products cannot always be assigned to one of these groups because one additive may simultaneously influence a combination of paint properties. Additionally it is sometimes difficult to draw the line between additives and other paint ingredients (e.g., resins, pigments, extenders, and solvents) when these materials are used in an additive-like manner.

5.1. Defoamers

Although foam may occur as an interfering factor already during paint production, most problems arise when it causes surface defects during the application process.

Liquid foams are a fine distribution of a gas (normally air) in a liquid. Thin films of liquid (the lamellae) separate the gas bubbles from one another and the gas – liquid interfacial area is quite high. For a detailed description, see ⟶ Foams and Foam Control.

Defoamers (antifoaming additives) are liquids with a low surface tension which have to satisfy three conditions [363-365]:

1. They must be virtually insoluble in the medium to be defoamed
2. They must have a positive penetration coefficient *E*

3. They must have a positive spreading coefficient *S*

$$E = \sigma_L - \sigma_D + \sigma_{L/D} > 0$$

$$S = \sigma_L - \sigma_D - \sigma_{L/D} > 0$$

σ_L = surface tension of the liquid phase

σ_D = surface tension of the defoamer

$\sigma_{L/D}$ = interfacial tension between the liquid and the defoamer

If both *E* and *S* are positive, the defoamer penetrates into the foam lamella and spreads across the surface. This creates interfacial tension differences that destabilize the lamellae and cause the foam to collapse. In simple terms it can be said that defoamers act because of their controlled incompatibility with the paint system. If a defoamer is too compatible its defoaming effect is not sufficient, if it is too incompatible film defects occur (e.g., gloss reduction, formation of craters). Defoamer action is often very system specific and slight changes in a coating composition may already require a different defoamer. As foam mostly occurs during the application process, also application parameters and substrate properties (e.g., porosity) can influence foam behavior.

For *waterborne paint systems* (especially emulsions used for decorative purposes) defoamers based on mineral oils are often used (⟶ Foams and Foam Control – Paint and Coating Industry). In addition to the mineral oil as carrier, these products contain finely dispersed hydrophobic particles (e.g., silica, metal stearates, polyureas) as defoaming components. A small amount of a polysiloxane is sometimes included to intensify the defoaming action. For high-quality waterborne coatings in industrial applications, defoamers are used that contain hydrophobic silicone oils (modified polysiloxanes) as the principal defoaming component instead of mineral oils (in combination with hydrophobic particles or without). They have a better defoaming effect, but are more expensive. In most cases silicone defoamers do not cause the gloss reduction that is often observed with mineral oil products. More recently silicone-free defoamers that make use of other foam-destroying polymeric structures have also come to the market.

Silicones are also the predominant defoamer components in *solventborne coatings.* Products with a correct balance of compatibility and incompatibility can be synthesized by selectively modifying the silicone backbone with polyether and/or alkyl chains.

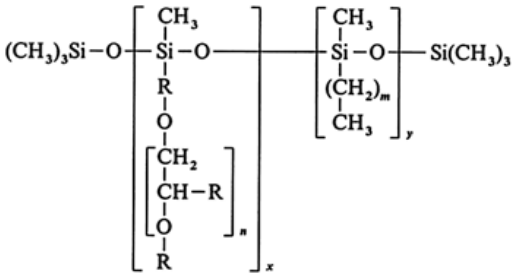

Silicone-free defoamers based on other incompatible polymers (e.g., acrylates and acrylic copolymers) are also commercially available.

5.2. Wetting and Dispersing Additives

In the production of pigmented paints, the pigment particles must be distributed as uniformly and as finely as possible in the liquid phase (see Section Paints). The pigment agglomerates must first be wetted by the binder solution. This fairly complex process (described in detail in [366]) mainly depends on the chemical nature of the pigments, solvents, and binders and loosely packed agglomerates and low viscosity of the penetrating binder solution are helpful for fast agglomerate wetting. Surface tension of the liquid phase and the contact angle between the liquid and the pigment surface also strongly control the wetting speed, and this is where the *wetting additives* exert their influence: Wetting additives are materials of low molecular mass with a typical polar – nonpolar surfactant structure. They reduce the surface tension of the binder solution to such an extent that the contact angle becomes zero, which is the requirement for good wetting. Further reduction of the surface tension should be avoided as it can reduce wetting speed.

After the agglomerates have been broken down into smaller particles by impact and shear forces (grinding, milling), the pigment dispersion must be stabilized to avoid reformation of larger pigment clusters by flocculation. Flocculation is due to the van der Waals forces between the particles, which are attractive in nature. They are effective only over very short distances, but the Brownian motion of the pigment particles in the liquid causes particle collisions and thus flocculation. To avoid flocculation it is necessary to build up an energy barrier between the particles that keeps them well separated. If this barrier is higher than the thermal energy (*kT*) of the particles, flocculation can be effectively minimized. *Dispersing additives* are stabilizing substances that are adsorbed onto the pigment surface via pigment-affinic groups (anchor groups with a high affinity for the pigment surface) and establish repulsive forces between individual pigment particles. Stabilization is achieved either via electrostatic charge repulsion (Fig. 4 A) or via steric hindrance due to molecular structures that project from the pigment surface into the binder solution (Fig. 4 B and C) [359], [367], [368]. The first mechanism is prevalent in waterborne emulsion systems and makes use of polyelectrolyte structures to create electrical charges on the pigment particles. The latter predominates in solventborne paints, and additive molecules can have relatively simple fatty acid based chemistry or more complex polymeric structures. In coatings with water-soluble resins both mechanisms can be employed, but in practice steric stabilization is often preferred, especially in high-quality coating formulations.

Figure 4. Stabilization of pigment dispersions

A) Electrostatic charge repulsion induced by polyelectrolytes; B) Steric hindrance through low molecular mass dispersing additives; C) Steric hindrance through polymeric dispersing additives

a) Pigment particle; b) Polyelectrolyte; c) Molecular structure causing steric hindrance; d) Pigment-affinic group

Good adsorption of the additive to the pigment surface is necessary for efficient stabilization. Problems may arise in this respect with many organic pigments because of their highly nonpolar surface. This is where polymeric dispersing additives have an advantage in comparison to the low molecular mass structures. These polymeric wetting and dispersing additives can better stabilize such difficult pigments by virtue of their macromolecular structure and the large number of pigment-affinic groups (Fig. 4 C).

Wetting and dispersing additives can also solve flooding and floating problems. Since most paints contain more than one pigment, the pigments often segregate in the paint film during drying. Nonuniform pigment distribution within the film surface is termed *floating* [formation of Bénard cells (Fig. 5 A) and streaks]. In *flooding* the surface is uniformly colored, concentration and thus shade differences occur only perpendicular to the surface; this phenomenon only becomes evident in the rub-out test (Fig. 5 B). In this test, after a short drying period part of the wet paint film is rubbed with the finger until almost dry (i.e., until it starts to become tacky). This treatment distributes the pigments evenly in the paint film and segregation is not possible. A color difference detected between the rubbed section and the untouched area indicates flooding.

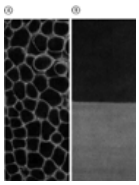

Figure 5. Uneven pigment distribution (color differences) due to flooding and floating

A) Bénard cells; B) Rub-out test

Flooding and floating are caused by local eddies in the drying paint film. The pigment particles undergo eddy motion and if they differ in mobility, they can become separated from one another. The mobility of the pigments depends on density, size, and the strength of their interactions with the binder molecules. Additives can minimize mobility differences between different pigments by controlling these pigment – binder interactions and thereby prevent flooding and floating.

Another way of avoiding flooding and floating is to prevent the separation of the pigments by coflocculation. Additives that work in this way are known as controlled flocculating additives (Fig. 6). They form bridges between pigment particles and thus build up flocculates. Size and stability of the flocculates are controlled by the additive. This method is, however, not ideal for high-quality topcoats because flocculation may reduce gloss and impair other paint properties (e.g., hiding power, color strength, transparency). Controlled flocculation also changes the rheology of the paint system (see Section Rheology Additives). Wetting and dispersing additives with such properties are often used in combination with other rheological additives. They enhance the action of the rheological additives, often synergistically, and problems such as sagging and settling can be overcome. In the case of settling, the presence of an additive layer on the pigment surface prevents the formation of hard sediment which would be difficult to stir in again. Instead any settled material formed is soft and easy to incorporate again. *Antisettling additives* generally increase the low shear viscosity to improve suspension of the pigment particles and avoid the formation of hard sediments.

Figure 6. Controlled flocculation

5.3. Surface Additives

During and directly after application of a coating several defects can develop in the coating surface. Many of these defects can be explained by differences in interfacial tension. Poor substrate wetting, for example, must be expected if the paint has a higher surface tension than the substrate to be coated. When spray dust or solid dust particles fall onto a freshly coated surface, *craters* are formed if the deposited droplets or particles have a lower surface tension than the surrounding paint material. Craters are also formed if the surface to be coated is locally contaminated with substances having a very low surface tension (e.g., oils) and the surface tension of the paint is too high to wet these contaminated areas.

Surface tension differences may also develop within the paint itself: during drying the solvent evaporates and this change in composition also alters the surface tension. Even slight surface tension differences lead to the formation of Bénard cells which may result in visible surface defects such as *orange peel* and *air draught sensitivity*.

In general, surface tension differences lead to material transport in the liquid paint film from the region of lower surface tension to that of higher surface tension. This movement is responsible for the above-mentioned defects. Other phenomena such as fat edges, picture framing, and ghosting can be explained in a similar way.

Silicone additives (mainly organically modified methylalkyl polysiloxanes) lower the surface tension of coatings and minimize surface tension differences.

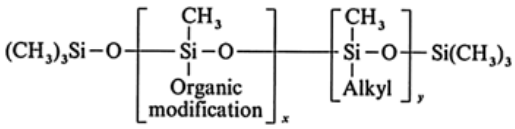

They are therefore ideal for solving the problems described above. Organic modification of the silicone (polyether and polyester chains, aromatic groups) serves to adjust the compatibility with the paint system. The alkyl groups have a strong influence on the surface tension: methyl groups give very low surface tension, longer alkyl chains give higher values.

Silicone additives also improve the *slip properties* of the dried coating (reduced coefficient of friction), which then exhibits improved blocking and scratch resistances. Similar effects can be achieved with wax additives (wax emulsions and dispersions in water or organic solvents and micronized waxes), which can have an additional flattening effect and also improve surface protection.

Poor *leveling* is also considered a surface defect. The leveling properties of a coating depend on many factors. Silicones influence the surface structure by suppressing eddy motion during drying. Acrylate copolymers are also used for the same purpose. They are incompatible with the paint system and accumulate at the surface. They also have a stabilizing effect on the surface but do not lower the surface tension as strongly as silicones.

Leveling also depends highly on paint rheology which can be modified by using special solvent blends. Finally, wetting and dispersing additives can also alter the rheology and thus influence leveling.

Besides modified polysiloxanes and acrylate copolymers also fluorinated structures and additives based on acetylene diols are used in practice to counteract surface defects.

5.4. Driers and Catalysts

Driers (siccatives) are used in paint systems that dry at ambient temperature by oxidation processes. They accelerate the drying process by catalyzing the autoxidation of the resin. Driers are in general organometallic compounds (metallic soaps of monocarboxylic acids with 8 – 11 carbon atoms, mainly naphthenates and octanoates), the metal being the active part. Cobalt, manganese, and lead (primary or active driers), calcium, zinc, zirconium, and barium (secondary or auxiliary driers) are mainly used (⟶ Metallic Soaps). In practive, mixtures of metallic soaps are commonly used to obtain the optimum ratio of auxiliary to active drying. Co – Pb – Ca is a typical combination that has been widely used in the past. Due to increasing environmental awareness, lead as a drier metal was banned for most applications, and lead-free drier systems use zirconium instead. More recently the use of cobalt, which is the most important drier metal today, has come under discussion because of its toxicological potential. Replacing it by precomplexed manganese driers in which organic compounds such as 2,2′-bipyridyl are used as ligands may be the solution in future.

Auxiliary driers cannot be used on their own, they always have to be combined with active driers.

The driers mentioned above can also be used in waterborne formulations, but it was found that they show reduced effectivity in such systems. One reason is that the metal ions are easily complexed with water molecules. A possible way around this problem is to use precomplexed driers.

Driers can cause skin formation during paint storage, particularly if the can or container has been opened. Oximes or alkylphenols are added as *antiskinning additives*. They block the action of the driers in the can, but at the correct dosage do not prolong the drying time of the applied paint film due to their volatility.

The curing of coatings that are cross-linked by other chemical reactions can be accelerated with *catalysts*. Acid catalysts are the most important and are used for a large number of stoving enamels and force-dried, acid-curing wood paints. They are mostly sulfonic acids of widely varying structure, often blocked with amines to allow formulation of storage-stable paints. The use of a catalyst can lower the stoving time and/or stoving temperature to save energy or to permit the coating of temperature-sensitive substrates.

Catalysts also include accelerators for two-pack polyurethane paints (e.g., tin and zinc compounds, tertiary amines), initiators for unsaturated polyester resins that act as radical-forming agents, and photoinitiators for UV curing systems.

5.5. Preservatives

Waterborne coatings systems are easily attacked by microorganisms. The biodegradable paint material detoriates and, as a result, bad smell, discoloration, and growth of mildew can be observed. Solventborne coatings are not prone to this kind of attack because of their content of organic solvents, but the dry paint film can also be attacked by microorganisms. In this case, it does not matter whether the coating was solventborne or waterborne, and under humid conditions algae and fungi can grow on the coating surface. In both cases, preservatives (biocides) are used to protect the coating. *In-can preservatives* protect

waterborne paint systems against contamination by microorganisms during production, transportation, and storage. *Dry-film preservation* is aimed at preventing the growth of bacteria, fungi, and algae on the applied paint film. A special area of use is the protection of wood against biodegradation by putrefactive fungi and insects. Antifouling additives for underwater coatings that are intended to prevent marine growth are also included in this category (see also Section Marine Coatings).

Preservative measures are governed by the intended use of the coating. There are no universal additives on account of the large number of possible types of damage; combination products containing several active ingredients are available and often used. Organomercury compounds, chlorinated phenols, and organotin compounds were often used, but these environmentally harmful products are now being replaced more and more by metal-free organic substances, mainly nitrogen-containing heterocycles.

In-can preservatives, besides being effective against a wide range of microorganisms and having low toxicity for humans and animals, must have good water solubility and good compatibility with the coating system. Dry-film preservatives must have low water solubility and volatility so that they stay in the coatings film as long as possible.

5.6. Rheology Additives

The flow properties of paints can be modified with rheology additives. This is important for obtaining optimum application behavior of the paint material, i.e., good leveling and no sagging at the desired film thickness. Paint rheology also plays a role in storage stability (settling, see Section Wetting and Dispersing Additives).

Thickeners, mainly cellulose derivatives (e.g., methyl cellulose, ethylhydroxypropyl cellulose) or polyacrylates, are generally used in emulsion paints. Polyurethane thickeners (associative thickeners) with more favorable leveling properties are also increasingly used.

A large number of rheological additives for solventborne systems are commercially available. Hydrogenated castor oils, pyrogenic silica, modified ureas, and modified montmorillonite clays (organoclays, e.g., bentonite) are preferred.

The rheological action of the above additives is based on the fact that they form three-dimensional network structures via hydrogen bonds in the paint. These lattice structures are destroyed by shear forces but are restored when the forces are removed. This recovery is not, however, immediate. The rising viscosity initially allows leveling of the surface but subsequently prevents sagging. This time-dependent change in viscosity is known as thixotropy (Fig. 7). In paints thixotropy is more desirable than purely pseudoplastic flow behavior because it allows a compromise between sagging and leveling.

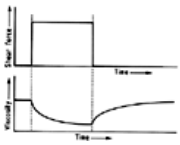

Figure 7. Thixotropy: viscosity is shear-force-dependent and time-dependent

5.7. Light Stabilizers

High-quality industrial coatings, especially automotive finishes, are subjected to severe weathering in exterior applications. In two-coat metallic coatings, exposure to UV light, oxygen, moisture, and atmospheric pollution causes decomposition of the polymer material in the automotive finishes. This decomposition results in loss of gloss, crack formation, color changes, and delamination phenomena [369].

High-energy UV light is particularly detrimental because each polymer material can be damaged particularly easily at one or more wavelengths in the UV range. Light stabilization is therefore essential [370].

Methods of Stabilization. Two stabilization methods have been adopted industrially [371], [372]:

1. Competitive UV absorption by UV absorbers in the wavelength range 290 – 350 nm.
2. Trapping of the radicals formed during polymer degradation by radical scavengers (hindered amine light stabilizers, HALS).

In two-coat metallic paints, the basecoat is protected against color change and photochemical decomposition (which leads to delamination) by the filter effect of the UV absorber that is added to the clearcoat; UV absorbers cannot trap radicals. Hindered amine light stabilizers do not absorb in the UV region, but trap radicals already formed, and are mainly responsible for the gloss retention and prevention of crack formation in paints. Optimum protection against decomposition phenomena in the coating is achieved by using a combination of both stabilization methods.

UV Absorbers. Four different classes of UV absorbers are shown below:

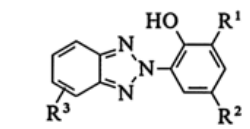

Hydroxyphenylbenzotriazoles

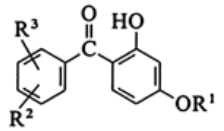

Hydroxybenzophenones

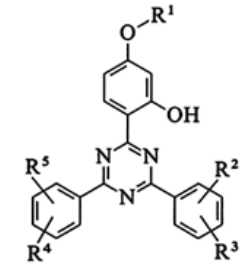

Hydroxyphenyl-*S*-triazines

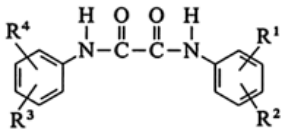

Oxalic anilides

The hydroxyphenylbenzotriazoles are the most important. They absorb the damaging UV light and rapidly convert it into harmless heat (keto – enol tautomerism) [372].

The action of all UV absorbers depends on the Lambert – Beer law, and the absorption properties of the UV absorber. The further the absorption edge extends into the near UV region, the more UV light can be filtered out. Of the four UV absorber classes shown above, the hydroxyphenylbenzotriazoles have the broadest absorption band [369], [373]. In addition to thermal stability [373] and stability to extraction with water or organic solvents, photochemical stability is important [374-376]. Ultraviolet reflection spectroscopy can be used to establish whether the employed UV absorber is still effective, even after several years' external weathering [377]. The hydroxyphenylbenzotriazoles have a much higher photochemical resistance than oxalic anilides and hydroxybenzophenones [369], [373], [378].

Radical Scavengers (Sterically Hindered Amines). Two typical sterically hindered amines (HALS = Hindered Amine Light Stabilizer) follow:

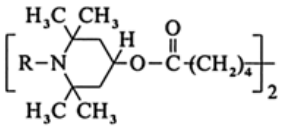

HALS I, $R = CH_3$ $pK_b \approx 5.5$
HALS II, $R = O\text{-}i\text{-}C_6H_{17}$ $pK_b \approx 10$

The tetramethylpiperidine group is responsible for the stabilizing action. Different substituents on the nitrogen atom result in different pK_b values, which are important in the area of use of the products. HALS I, bis(1,2,2,6,6-pentamethyl-4-piperidinyl) ester of decanedioic acid [41556-26-7], is used in systems that are not catalyzed by strong acids (interaction of the acid with the basic nitrogen atom). HALS II, bis(2,2,6,6-tetramethyl-1-isooctyloxy-4-piperidinyl) ester of decanedioic acid [122586-52-1], was developed for acid-catalyzed systems because it does not undergo undesirable interactions with acid catalysts.

The mode of action (Densiov cycle) of HALS (as deduced from investigations on polyolefins) follows (P = polymer) [372], [379]:

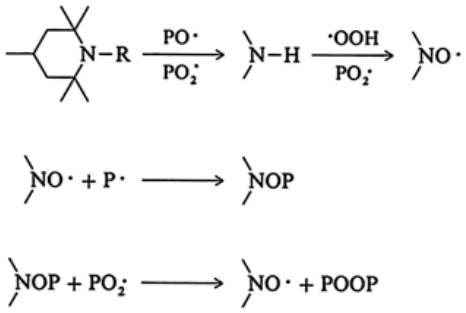

The formation of nitroxyl radicals ⟩NO· is essential for stabilization since the concentration of harmful peroxy radicals falls sharply in their presence [380].

Light stabilizers are tested under artificial weathering conditions (accelerated weathering) and under outdoor weathering (Florida, 5° South, black box, not heated) [378]. Figure 8 and Table 21 illustrate the influence of light protection agents on gloss retention and crack formation in two-coat metallic coatings.

Table 21. Outdoor exposure of a two-coat metallic coating [a]

Light stabilizer	20° gloss after *n* years Florida				
	n=0	*n*=2	*n*=4	*n*=6	*n*=8
Unstabilized [b]	93	45			
1 % Benzotriazole I [c]	94	71	58		
1 % Benzotriazole I and 1 % HALS I	94	70	67	56	50

[a] Clearcoat, acrylic – melamine; basecoat, polyester – cellulose, acetobutyrate – melamine, silver metallic; bake: 130 °C, 30 min; exposure: Florida, 5° South, black box unheated; percentage of light stabilizer relative to binder solids; benzotriazole I, 2-(2 *H*-benzotriazole-2-yl)-4,6-bis(1-methyl-1-phenylethyl)phenol [70321-86-7].
[b] Cracking after 2.25 years.
[c] Cracking after 5.5 years.

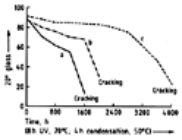

Figure 8. Accelerated weathering of an acid-catalyzed two-coat metallic coating in an UVCON apparatus (Atlas Corp.)

Clearcoat: high-solids acrylic – melamine; Basecoat: high-solids acrylic – melamine, silver metallic; Bake: 120 °C, 30 min; Benzotriazole II: 3-(2 *H*-benzotriazole-2-yl)-5-(1,1-dimethylethyl)-4-hydroxyoctyl benzenepropanoate [84268-23-5]

a) Unstabilized; b) 2.5 % Benzotriazole II; c) 1.5 % Benzotriazole II and 1 % HALS II (percentage relative to binder solids)

5.8. Commercial Products

The number of additives available on the market is quite large and picking some of these and mentioning them as examples is not very helpful, because their use is often system specific. In addition, new additives are constantly being developed to meet market requirements. Therefore, it seems more useful to give a list of the main additive suppliers, and further up-to-date information can be obtained from these sources.

This following list contains the main companies worldwide that manufacture additives for coatings. However, such a list can never be exhaustive and there are many other companies active in the same market. The information given here may change without notice due to mergers and acquisitions and other circumstances.

Company	Products
Air Products www.airproducts.com/	Wetting and dispersing additives, defoamers, surface additives
BYK-Chemie www.byk-chemie.com/	Wetting and dispersing additives, defoamers, surface additives, waxes, rheology additives
Ciba www.cibasc.com/	Wetting and dispersing additives, defoamers, surface additives, waxes, preservatives, UV stabilizers, photoinitiators
Cognis www.cognis.com/	Wetting and dispersing additives, defoamers, surface additives, rheology additives, corrosion inhibitors, plasticizers, preservatives,
Cytec www.surfacespecialties.com/	Wetting and dispersing additives, defoamers, surface additives, antiskinning additives, catalysts, driers, preservatives

Degussa	Wetting and dispersing additives, defoamers, surface additives, rheology additives, adhesion promoters
www.degussa.com/	
www.tego.de/	
Deuchem	Wetting and dispersing additives, defoamers, surface additives, adhesion promoters, catalysts
www.deuchem.com/	
Elementis	Wetting and dispersing additives, defoamers, waxes, rheology additives, driers, corrosion inhibitors, UV stabilizers
www.elementis-specialties.com/	
King Industries	Wetting and dispersing additives, defoamers, rheology additives, catalysts, corrosion inhibitors
www.kingindustries.com/	
Kusumoto	Wetting and dispersing additives, defoamers, surface additives, waxes, rheology additives, preservatives
www.kusumoto.co.jp/	
Lanxess	Wetting and dispersing additives, defoamers, surface additives, catalysts, driers, antiskinning additives
www.lanxess.com/	
www.borchers.com/	
Micro Powders	Waxes
www.micropowders.com/	
Münzing Chemie	Wetting and dispersing additives, defoamers, surface additives, rheology additives
www.munzing-chemie.com/	
Noveon	Wetting and dispersing additives, defoamers, surface additives, waxes, rheology additives
www.noveoncoatings.com/	
Troy	Wetting and dispersing additives, defoamers, surface additives, rheology additives, driers, preservatives
www.troycorp.com/	
Worlée-Chemie	Wetting and dispersing additives, defoamers, surface additives, rheology additives
www.worlee.com/	

5.9. Use and Testing of Additives

Additives are generally used in very small amounts (≤1 % of total formulation) and correct dosage is extremely important for optimum effectiveness and also to avoid undesirable side effects. The dosage has to be determined in test series.

Many additives can be incorporated relatively easily into the paint system during or after the last phase of production (let down). In some cases, however (e.g., with many rheological additives), certain incorporation conditions have to be observed. Wetting and dispersing additives, for example, always have to be present in the millbase to achieve the desired results.

Ideally, the effectiveness of an additive should be checked in the complete formulation and under conditions as close as possible to those prevailing in practice. Checking a defoamer just in the binder system, for example, can only be regarded as a preliminary test, because the behavior in the complete formulation may differ substantially.

Many defects that are to be eliminated by the use of additives are also influenced by the substrate to be coated and the application method. Differences between additives can be established in simple laboratory tests. However the final composition of a specific formulation must take into account as many application parameters as possible (e.g., the state of substrate, application method, and drying conditions).

In most cases additives influence not just one property of the coating. They may also have undesirable or beneficial side effects. Additives are not “magical” products but need to be used rationally and carefully to provide the desired satisfactory results. As detailed and complete a knowledge as possible of the mechanism of action of the products, their possible effects and side effects, their limitations, and the underlying causes of paint defects are certainly helpful, but due to the complexity of paints and coatings empirical knowledge is indispensible.

[Top of Page]

6. Paint Removal

Thomas Molz

The nature, condition, and quality of the paint and substrate are important in paint removal. The paint binder plays a decisive role in paint dissolution; the substrate influences the choice of paint removal method. Various chemical and physical methods exist for removing paint from different substrates (metals, wood, and mineral substrates) [381], [382].

6.1. Paint Removal from Metals

6.1.1. *Chemical Paint Removal*

Paint layers can be stripped (dissolved) or degraded with chemicals [383]. Paint dissolution is performed with organic solvents, whose action is assisted by surfactants. Paint binders can be degraded with strong alkali or acid. Depending on their use, paint removers may also contain cosolvents, activators (acid or alkali), wetting agents, emulsifiers, evaporation retarders, corrosion inhibitors, and thickening agents. The efficiency of paint removers is improved by increasing the temperature; the paint removal time can be substantially shortened by raising the temperature of the bath (2 – 6 h at 20 – 95 °C).

The *advantages* of chemical paint removal are that it can be used for almost all types of paints, geometries, and heat-sensitive items. Investment costs are low and waste air does not cause serious environmental problems. The *disadvantages* are the relatively long removal time and formation of a paint slurry which leads to higher waste disposal costs.

Hot Alkaline Paint Removal Baths. Paint removal takes place in hot, aqueous, alkaline baths (> pH 13) at 50 – 95 °C. The alkalinity is adjusted by adding large amounts of alkali-metal hydroxides or organic amine or hydroxy compounds. The chemical bonds in the paint binders are hydrolyzed by the high alkalinity. Penetration into the paint layer and migration underneath the film is promoted by adding surfactants. Hot paint removers may, for example, consist of:

50 – 70 %	alkali-metal hydroxides (KOH/NaOH)
0 – 20 %	strongly basic organic hydroxy or amine compounds (e.g., alkanolamines)
5 – 20 %	high-boiling solvents (e.g., glycol ethers)
5 – 20 %	organic acids (e.g., gluconic acid) and their alkali-metal salts
0 – 5 %	surfactants

Strongly hydrolyzing, aggressive alkali lyes can be used for removing paint from steel whereas strong organic bases are used for light and nonferrous metals. The hot paint removal bath is overlaid with a layer of an organic compound of low volatility (e.g., paraffin oil) to prevent evaporation of water and active constituents. Alkaline hot paint removal agents are extremely economical and can be used for most paints.

Neutral paint strippers include halogen-free organic solvents (e.g., glycols, glycol ethers, 1-methyl-2-pyrrolidinone) which are generally used at 20 – 40 °C. In contrast to the alkaline paint removers, neutral paint strippers result in purely physical dissolution of the paint from the substrate. Their use is therefore restricted to removing physically drying paints.

Acid Paint Removers. Like the alkaline products, the acid paint removal agents (pH < 1) based on mineral acids result in chemical degradation of the paint binders. They are usually based on sulfuric acid and are covered with a protective layer of paraffin wax. Alkaline agents have a pronounced swelling action, whereas the acid agents result in disintegration of the binder. The acid products are used only in special cases (e.g., to remove epoxy and polyamide powder coatings from heat-sensitive substrates) because of their aggressive, corrosive properties; bath temperature is 20 – 50 °C.

Cold Paint Removal. Up to a few years ago, cold paint removal based on the use of dichloromethane [75-09-2] (methylene chloride) was the most widely used paint removal method. Use of this method is declining in favor of chemical or thermal methods due to concern about the environmental and occupational safety of halogenated hydrocarbons.

6.1.2. *Thermal Paint Removal*

Thermal paint removal methods exploit the fact that organic paint constituents can be readily pyrolyzed or combusted, as well as the high heat capacity of the organic material. They can be used to remove paint from steel and aluminum but not from heat-sensitive substrates (e.g., zinc and zinc alloys). The short removal times and small amount of waste ash formed are advantages, although this is offset by the high capital investment resulting from stringent safety regulations and the need for equipment to combust the resultant pyrolysis gases.

Low-Temperature Carbonization Method. Pyrolysis is performed for 3 – 9 h in closed furnaces at 380 – 420 °C in a controlled, low-oxygen atmosphere.

Fluidized-Bed Method. The fluidized bed consists of inert, finely granular, inorganic material (usually alumina) that is fluidized by injecting compressed air through a special fluidization floor. Organic paint material is oxidatively degraded by the entrained atmospheric oxygen at 400 – 480 °C in 20 – 60 min. This method is not suitable for some geometries (e.g., hollow items).

Salt Bath Method. The object from which the paint is to be removed is immersed in a salt melt heated to 300 – 500 °C. The short time required for paint removal (30 – 120 s) is the result of the spontaneous heat transfer and high oxidation potential of the salt melt (e.g., alkali nitrates).

6.1.3. *Mechanical and Low-Temperature Paint Removal*

High-Pressure Removal with Water. Paint can be removed from a metallic substrate (e.g., grids and skids) by the high kinetic energy of a high-pressure (70 – 100 MPa) water jet.

Blasting. In analogy with sandblasting of metal surfaces, paint can be removed from various substrates by blasting with air or water containing abrasive blasting media (e.g., sand, plastic granules, metal particles) and/or other additives (e.g., alkali salts). Paint is, for example, removed from aircraft (aluminum) with a water jet containing special additives.

Low-temperature paint removal exploits the shrinkage and embrittlement of paint layers that occurs after cooling in liquid nitrogen (–196 °C) for 1 – 3 min.

The mechanical methods and low-temperature paint removal are restricted to a few special areas of application. Chemical and thermal methods have their specific advantages and disadvantages, not only as regards paint removal but also with respect to environmental pollution by paint slurry, rinse water, effluent, and waste air.

6.2. Paint Removal from Wood and Mineral Substrates [384]

Paint strippers used to remove coatings from wood and mineral substrates may be either alkaline or based on organic solvents.

Alkaline paint strippers (e.g., alkali-metal hydroxides, sodium carbonate, sodium metasilicate, trisodium phosphate) can be manufactured in liquid, paste, or powder form. They are only suitable for removing hydrolyzable coating materials (e.g., alkyd resins and oil paints). Alkaline paint strippers may also be used to remove coatings on facades. Tests should be carried out to check that the coating is hydrolyzable.

The *solvent-containing paint strippers* may contain chlorinated hydrocarbons (dichloromethane) and cosolvents (e.g., alcohols and aromatic hydrocarbons). Systems that do not contain chlorinated hydrocarbons (CHC-free paint strippers) are, however, becoming increasingly important because environmental considerations demand reduction in the use of chlorinated hydrocarbons. The dichloromethane-containing formulations evaporate relatively quickly and subsequently enter the atmosphere. The CHC-free paint strippers are formulated with a combination of various slowly evaporating but more effective solvents. Typical solvents are high-boiling glycol ethers, dicarboxylic acid esters, *N*-methylpyrrolidone, alcohols, and ketones in combination with auxiliary substances. CHC-free paint strippers are suitable for removing facade coatings whereby many paint layers can be removed simultaneously. Removal takes a longer time than with CHC-containing paint strippers but this can be compensated for by using different application techniques. It is important that the paint stripper wastes are readily biodegradable.

[Top of Page]

7. Production Technology

Rainer Hiller

7.1. Principles

A large number of paint formulations are produced in a limited number of production steps. The special know-how for paint production therefore involves skilful adaptation of the processing steps to the relevant production plant (scale-up, quality planning). Until recently, paint production was a traditional craft in which formulations were adapted in a large number of production stages. Quality testing was carried out at the end of the production process. Fluctuations in quality were rectified in time-consuming, and thus expensive, correction steps. With the transition to industrial production, quality planning and assurance systems were increasingly implemented in conjunction with the scale-up of unit operations to achieve the necessary quality specifications without additional correction stages.

Many other factors besides the choice of suitable raw materials have to be taken into account if a paint formulation is to serve as a basis for commercial production of high-quality paints. This is because the quality and stability of the final product is determined not only by the choice and quantity of ingredients (pigments, film formers, solvents, additives, etc.) but also by how and in which order they are combined.

Source of Added Value. The main physical process involved in the production of coating materials is the homogeneous, irreversible mixing of the liquid components. In pigmented systems, complete wetting and a uniform, stabilized distribution of pigment particles in the liquid medium (resin solution or dispersion) are also important. The main reason why these apparently simple processes result in a relatively high added value is the crucial but difficult to achieve requirement for microhomogeneity with a particle size of ≤ 1 µm. Homogeneity or dispersity at this level is very important because of the interaction of the paints with the absorbed light with wavelengths in the visible region. High energy inputs are necessary to obtain microhomogeneity (or dispersity) due to the interaction of strong adhesive forces in this region.

Production Strategy. An all-embracing description of paint production processes cannot be given, mainly because paint manufacturers range from very small operations producing only a few hundred tonnes of paint annually up to large companies producing several hundred thousand tonnes annually. Large companies account for only about 20 % of the worldwide market (ca. 30×10^6 t). A further difficulty is the multiplicity of formulations: large paint factories must hold a range of up to 20 000 formulations to be able to serve their markets. In addition, national differences in environmental legislation make a uniform approach difficult.

Since customer orders differ not only as regards paint formulation but also in the amount ordered, manufacturers must have a corresponding “mix” of equipment (different sizes of mixers, tanks, production equipment, filling and packaging lines, etc.). Production according to customer orders (including just-in-time

delivery) results in widely varying batch sizes, ranging from a few kilograms up to 20 or 50 t, but of the same, constant quality. A large paint factory therefore has a complex structure with regard to material flow and the provision of materials, equipment, and staff [394].

Table 22 shows the compositions of a filler and a one-pack topcoat (white and colored) for application in automotive mass production. The paint-making process can be subdivided into four main stages:

1. Preparation of the millbase (premixing)
2. Continuous dispersion
3. Completion of the formulation
4. Correction and adjustment of the final product

Two strategies are adopted for producing the millbase:

1. Ab initio production
2. Mixed production, using pastes

Ab initio production generally starts from raw materials, and pigment mixtures are dispersed directly in the batch. Mixed production starts from ready-prepared pigment pastes: only one kind of pigment is usually dispersed in the binders, and is ready-for-use. Paste series differ according to the nature of the film formers used.

Table 22. Examples of compositions of coating materials for automotive mass production (wt %)

	Filler	One-pack topcoat, white	One-pack topcoat, colored
Component	**Ab initio**	**Mixed Ab initio**	**Mixed Ab initio**
Dispenser formulation (comminuted material)			
Paste 1		99.5	60
Paste 2		0.5	30
Paste 3			7
Paste 4			2
Paste 5			1
Solvent	20	18	25
Film former (resin) (100 %)	15	20	23
Pigment	25	60	50
Extender	35		
Additives (volatile)	5	2	2
Total liquid components	75	80	73
Pigment and extender	60	60	50
Film former (100 %)	15	20	23
Proportion of dispersed millbase in total formulation	60	40	20
Total formulation			
Solvent	31	40	45
Film former (100 %)	30	35	45
Pigment	15	24	10
Extender	21		
Additives	4	1	1
Total liquid components	65	60	55
Liquid components, pigment and extender	35	25	10

Smaller factories generally buy in their film-forming materials and if necessary convert them into solutions. The major paint companies, on the other hand, produce considerable amounts of the film formers (resins) themselves and are therefore able to develop “tailor-made” film-forming agents that are optimized to the formulations, production conditions, application technique, and intended use of the resulting coating composition. Such companies can supply special, high-quality products.

Material Flow. Figure 9 illustrates the material flow in paint and coating production, starting with the reception and storage of the raw materials and finishing with the end product ready for dispatch. The type of raw material, its consistency, the amount consumed, and the packaging in the as-supplied state are decisive for material flow and metering. In large factories, large storage tanks for resin solutions and solvents have to be installed and equipped with the latest safety devices. The raw materials are usually delivered in tank trains. These bulk raw materials can be pumped and conveyed to the sites of use via pipes. A considerable number of other liquid, pasty, solid, or powder raw materials are supplied, palletized, and depalletized in containers, barrels, drums, or sacks. Open-air storage facilities and warehouses for free-standing stacking are provided for this purpose. The raw materials for 1, 2, 4, and in some cases 3, are delivered to the appropriate stores by road. The material flow then continues to the production facilities (9 and 10).

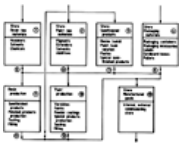

Figure 9. Material flow in paint production (for explanation of the numerals see text)

Four types of product are made (see Section Paint-Making Processes):

1. Clear or colored varnishes (do not contain pigment)
2. Paints (contain pigment)

3. Powder coatings
4. Special products (e.g., putties, fillers, plasters)

The paint maker also produces semifinished products (e.g., pastes) for internal use. The semifinished products from 12 are stored in 3. Packaging materials flow via 4 to the filling areas at the end of resin (9) and paint (10) production. Many different packaging types, sizes, and designs are used. The end products may be dispensed into containers the size of a road tanker, into drums and barrels, or into cans with a capacity of as little as 50 mL. Finally, the ready-for-use material is sent to the storage area 13 before it is dispatched.

Production based on vertical material and process flow is most commonly used and will now be considered in a conventional factory (Fig. 10). Smaller factories operate on a horizontal concept, but the principles are the same.

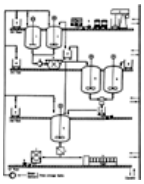

Figure 10. Process flow in conventional paint making

a) Disperser; b) Scales; c) Premixer, dissolver; d) Movable container; e) Pump; f) Agitator mill; g) Paste mixer; h) End product mixer; i) Filter, sieve; j) Container; k) Packing machine

The material and process flow proceeds downward through four floors as follows:

Fourth Floor. The fourth floor is used for intermediate storage and metering of the raw materials required to produce the millbase. Solid raw materials (e.g., pigments and fillers) are stored on pallets in sacks, drums, barrels, or cartons. Liquid or pasty raw materials are kept in drums or cans. These materials are weighed (b) and charged into the premixers or dispersers (a, c). Solvents and solutions of film-forming materials are fed from the storage tanks into premixers or dissolvers (c) via pipelines and filling stations with oval disk meters.

Third Floor. The mixture is dispersed using agitator mills (f). With mixed production the dispersed millbase is then fed as a single-pigment paste to paste mixers (g) and then to the end product mixers (h). In ab initio production the multipigment paste is sent directly from the agitator mills to the end product mixers (h).

Second Floor. A large number of end product mixers (h) of different sizes are used. With mixed production the mixed pastes are metered from the paste mixers (g) into mobile containers and then added to the end product mixers (h). In ab initio production the dispersed material is pumped directly from the agitator mills (f) into the end product mixers (h). Formulation of the paste mixture or dispersion is completed in the end product mixer by adding film-former solutions, solvents (via oval disk meters), and additives. The product is sampled, tested, and adjusted (e.g., shade, viscosity, other technically important properties).

First Floor. The end product is sieved, filtered, and dispensed from the end product mixers via filling lines into containers, drums, or small packages.

The material flow system is complex because the mixers and machines have to be changed in accordance with the product type and batch size. Equipment also has to be reserved for specific product groups with special shades or chemical properties. Only in this way can the effective cleaning of equipment and pipelines during batch changes be guaranteed.

In large paint factories there is an increasing tendency to perform the precharging (i.e., the accurate weighing and preparation of the solid paint constituents) centrally in the storage area (Fig. 11). This has economic and organizational advantages; it also reduces the risk of contamination. Filling is also centralized for the same reasons. Materials are conveyed in mobile containers or via pipelines. Cleaning inside pipelines can be optimized by using scraper (go-devil) techniques.

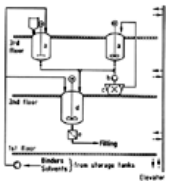

Figure 11. Process flow in modern paint making

a) Premixer, dissolver; b) Pump; c) Agitator mill; d) End product mixer; e) Filter, sieve

Very few product lines in the paint industry are mass-produced; white emulsion paints, plasters, rust protection paints, and underbody protection materials are, however, exceptions. Paint factories producing these bulk products are highly automated but extremely inflexible as regards the use of different raw materials and semifinished products and their proper scheduling. Enterprises that manufacture “tailor-made” products must have a wide range of equipment, and simple and flexible production methods.

Quality Assurance. Almost every aspect of the paint-making process is subjected to quality assurance measures. Depending on the type of product, this involves a number of correction stages and delay periods, particularly in the end product mixers. In the most commonly employed correction strategy, the desired quality is approached and achieved in small steps. Correction in this context does not mean the elimination of defects, but the addition of small amounts of materials. Quality assurance corresponding to the ISO Standards 9000 – 9004 (e.g., failure mode-and-effect analysis (FMEA) and statistical process control (SPC) [395]) has been introduced in modern paint factories.

Environmental Protection. The paint industry has adopted a two-pronged strategy to ensure environmental protection:

1. Apparatus and processes are enclosed as far as possible to minimize emissions
2. If enclosure is not possible, the sources of emission are selectively aspirated. Gaseous pollutants and vapors are collected and passed to central or decentralized disposal equipment [e.g., combustion plants (thermal or catalytic) or adsorption units], particulate impurities being retained in filters.

7.2. Paint-Making Processes

The wide variety of paint formulations are reflected in the variety of paint-making processes. Figure 12 summarizes the essential features of these processes. Only a few special products (e.g., putties, plasters, or underbody protection materials) are not covered by this scheme; they are produced by special proprietary procedures (generally of the one-pot type).

Figure 12. General flow diagram of the paint-making process

7.2.1. *Varnishes*

Varnishes are usually colorless but may be slightly colored due to the intrinsic color of the binders, driers, or other additives. Colored transparent varnishes

contain soluble dyes or a small amount of finely dispersed pigments. Matt varnishes may be opaque (milky), colorless, or transparent; after drying they have a matt to semigloss surface due to the presence of dulling (mainly inorganic) additives.

Liquid raw materials are mixed in a specific sequence in slow- to high-speed stirrers or mixers to give a homogeneous, streak-free mixture (see Section Kneaders and Kneader Mixers). In larger mixers and tanks (up to 50 000 L capacity) the mixing device is permanently installed in the vessels. Precondensates and wax solutions are prepared in heatable and coolable vessels equipped with stirrers.

Solids are usually dissolved in liquids using high-speed stirring equipment. If the formulation contains dissolving and nondissolving solvents (diluents), then the nondissolving solvents are often added first, followed by the solid raw materials (e.g., powders, granules, fibers), and then finally the “true” solvent. With this order of addition the surfaces of the solid particles are wetted more effectively, resulting in uniform penetration of the “dissolvers” into the solids. Agglomeration is thereby avoided and dissolution is speeded up. In Figure 12 the sequence of operations is 1, 2, 3, or 4, 7, 8, 9.

7.2.2. *Paints*

Dispersion. The liquid semifinished products (intermediates) used in paint making are prepared in a similar manner to varnishes (see Section Varnishes). The most important stage in the production of pigment-containing paints is complete wetting and uniform distribution (i.e., dispersion) of the solid pigment particles in the liquid, film-forming material. To obtain optimal paint properties the pigment must be optimally dispersed in the liquid. Many pigments are surface-finished. An extremely fine grain size in the finished paint results in undesirable effects (e.g., gloss streaks).

Historically, to obtain homogeneous pigment wetting and dispersion, kneaders, rotor and stator mills, and roller mills came first. These were followed by ball mills, tank mills, attritors, and (open) sand (bead) mills. Enclosed ball mills are now widely used to comply with increasingly stringent quality and environmental protection requirements. Fitted with an enclosure, a roller mill is still acceptable for some product lines (e.g., pasty printing inks). Paint production proceeds according to steps 1, 2, 4, 5, 6, 7, 8, and 9 in Figure 12.

Commercial coating materials seldom have the same composition as the ready-dispersed millbase. The proportion of film former in the final product is generally higher. Highly concentrated (high-solids) paint resins must be added to increase the proportion of film former to the required value. This is generally performed in formulation vessels equipped with appropriate stirrers. This stage can present serious problems because the widely differing solvent concentrations of the two substance streams may cause irreversible diffusion.

Since the solvent molecules are more mobile than the film-former molecules, they migrate down their concentration gradient from the millbase into the paint solution, resulting in a decrease in volume of the pigmented phase. The distance between the pigment particles is reduced to such an extent that, even in stabilized pigment dispersions, repulsive forces between individual pigment particles are overcome and flocculation occurs. This process (pigment shock) can be avoided by modifying the compositions of the millbase and the final formulation, or by suitable process technology (e.g., rapid homogenization during final formulation). Flocculation may also occur when solvents are added because they remove the stabilizing film-forming layers from the pigment surface.

Flushing was developed industrially in the early 1930s. Aqueous pigment press cakes or doughs are kneaded with binders, wetting agents, and additives until phase inversion occurs, i.e., until the binder has displaced the water from the pigment surface and the hydrophilic pigment – water mixture has been converted into a hydrophobic pigment – binder system. The bulk of the water is decanted, any residual water being removed by applying a vacuum or in heated roller mills.

Flushing has the advantage that the original primary form of the finely divided pigment particles found in the aqueous dough is retained. Efficiently flushed pigments exhibit a purer shade, better transparency, and possibly higher yield in paints.

7.2.3. *Powder Coatings*

In order to produce powder coatings (see Section Coating Powders) solid binders (e.g., epoxy, polyurethane, acrylic – polyester resins) are comminuted and thoroughly mixed with pigments, catalysts, hardeners, and additives (Fig. 13, a). The mixture then passes through a metering device into an extruder (b) where it is melted and homogenized at temperatures above the softening point of the binder and below the activation temperature of the catalyst. The extruded paste is rolled out to a thickness of 2 – 3 mm and transported on a cooling belt (c) or drum to a roll crusher (d) which comminutes the cold mass into chips. The chips are finely ground in an impact pulverizer or air separation mill (k). The powder is then screened (m) and classified to the desired particle size fineness. Oversize particles are reground [396-398]. In addition to the vertical plant design illustrated in Figure 13 production may also be performed horizontally.

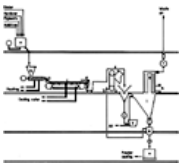

Figure 13. Powder coating production

a) Premixer; b) Extruder; c) Cooling belt; d) Roll crusher; e) Tipping device; f) Ventilator; g) Bag filter; h) Explosion valve; i) Hopper; j) Cellular wheel sluice; k) Air separation mill; l) Cellular wheel sluice; m) Screening device; n) Packaging

7.3. Apparatus

7.3.1. *Mixers*

See also ⟶ Stirring; ⟶ Mixing of Highly Viscous Media.

A wide variety of containers and tanks are used as mixers in the paint industry. Their materials of construction must be appropriate to the contents being mixed. Paints containing organic solvents are produced in steel or aluminum vessels. Waterborne paints generally require stainless steel because of their high corrosion potential. The tanks may also be lined with enamel or plastics.

The stirrers are mainly of the flat-blade, disk, propeller, or turbine type [399-402]. Rotor – stator systems are also used as continuous mixers (Fig. 14). These machines are employed particularly if an emulsion or an emulsion-like product has to be produced (e.g., waterborne paints) [403-405].

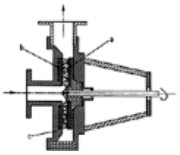

Figure 14. Continuous mixer for paint making (rotor–stator system)

a) Rotor; b) Stator; c) Housing

7.3.2. *Dissolvers*

Dissolvers are specially modified mixers (Fig. 15), and are mainly used in paint making for mixing solids (pigments) and resin solutions [406]. They have a good mixing effect and are even capable of breaking down coarse pigment agglomerates. The high shear forces required are produced by appropriate design and high speeds (up to 30 m/s at the stirrer tip). Power consumption may reach 50 kW/m^3. Dissolvers exist in single-shaft and multishaft versions. Vacuum dissolvers are used for product degassing.

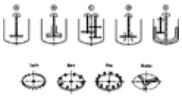

Figure 15. Types of dissolvers and dissolver disks

A) Simple dissolver; B) Dissolver with stator system; C) Double-shaft dissolver with three disks; D) Dissolver with disk for pumping and shearing; E) Dissolver with two drives for separate functions

7.3.3. *Kneaders and Kneader Mixers*

Various kneader designs (e.g., ram, internal, screw, paddle, Z-, arm, and planetary kneaders, Fig. 16) are used to produce highly viscous (pasty) paint concentrates (master batches), uncured rubber compositions, and putties (⟶ Mixing of Highly Viscous Media). Kneaders have a high energy consumption because they are generally used to produce systems that have a low binder content or are difficult to wet [406].

Figure 16. Planetary (A) and vat (B) kneaders

a) Drive; b) Vat; c) Kneader hock

Heatable and coolable kneaders are available and both types can be fitted with a vacuum device. They may operate batchwise or continuously and are equipped with screw conveyors.

7.3.4. *Media Mills*

See also ⟶ Size Reduction.

On account of the widely varying production requirements (see Section Principles), a wide range of mills are used for dispersion. *Ball mills,* rotating vessels filled with balls (diameter > 10 mm), were developed first (⟶ Size Reduction – Grinding Media Mills). Ball mills are inefficient compared with modern units because they operate batchwise and large batches cannot be produced. In addition to safety problems, their high noise level is a disadvantage.

To overcome the safety problems associated with rotating ball mills, *tank mills* were developed which consist of a stationary vessel equipped with a slowly rotating stirrer. Use of grinding media less than 10 mm in size increased product quality, grinding efficiency, and thus output.

Increasing the stirring power, optimizing the geometry, and introducing an external forced circulation led to further advantages and the development of the *attritor* (attrition mill).

The transition from batch production to continuous production required the development of *sand mills*. Sand mills can be regarded as slender attritors with a forced product circulation. They are charged with product vertically from the bottom upward, gravity being mainly responsible for the separation of the grinding media. A sieve is installed to improve separation. The name of the mill is derived from the fact that Ottawa sand (particle size ca. 0.5 – 3 mm) is usually used as the grinding medium. The development of sand mills increased both product quality and productivity. Throughput was, however, restricted by the force of gravity. The viscosity of the paint pastes therefore had to be relatively low, which did not promote optimum dispersion. Another disadvantage is the open construction, which led to additional emission problems.

A further advance was the development of *enclosed agitator mills.* The horizontal design was intended to eliminate the influence of gravitational forces. Numerous design features have been optimized [401], [402], [405-415]. The most important developments are illustrated in Fig. 17.

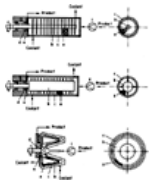

Figure 17. Closed agitator mills [387]

a) Grinding media; b) Rotor; c) Grinding chamber; d) Sealing system; e) Separator; f) Pump

Disadvantages associated with these mills are the high wear of the stirrer elements, the expensive axial face seals, and the limited throughput due to the often extremely narrow devices used to separate the grinding media (friction gaps, sieve). Since the free cross-sectional area of the separation devices can only be decreased within certain limits, small grinding media (< 1 mm) cannot be used to increase product quality [416].

The breakthrough in wet mill technology came with the development of the *centrifugal vortex mill* (Fig. 18) [416], [417]. The major innovation in this mill is the separation of the active grinding zone from the passive separation and sealing zone. This results in two advantages: (1) very low wear in the passive separation and sealing zone, and (2) the incorporation of large surface areas for separation (sieves). The second advantage is essential for any substantial increase in output: the large sieve areas allow throughput performance to be considerably increased (by a factor of 10 – 20). The dwell time is considerably shortened by a recirculation or multipass system. A fine, narrow particle size distribution can then be obtained more quickly. A device that combines the advantages of a large product throughput and the use of small grinding media, is the *double cylinder bead mill.* This mill has an internal grinding medium circulation to avoid compression of the grinding media [418].

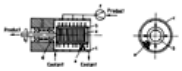

Figure 18. Centrifugal vortex mill

a) Grinding media; b) Rotor; c) Grinding chamber; d) Sealing system; e) Separator; f) Pump

Dispersion Methods. The dispersion methods most commonly used in the paint industry are illustrated in Figure 19. The *single-pass batch procedure* (Fig. 19 A) is the most common method. Its disadvantage is the broad dwell time distribution and the resulting wide particle size distribution. It is also extremely difficult to control, i.e., productivity falls sharply if the batch has to be treated again to improve its quality.

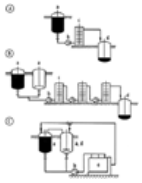

Figure 19. Dispersion methods

A) Passage; B) Cascade; C) Swing-batch, circulation

a) Mixer; b) Pump; c) Conventional mill; d) Let-down tank; e) Centrifugal vortex mill

A *cascade of agitator mills* (Fig. 19 B) produces a considerably narrower particle size distribution. Control is, however, more complex than with the single-pass procedure.

The *swing-batch and circulation procedures* (Fig. 19 C) are suitable for machines with high product throughputs (e.g., the centrifugal vortex mill and double cylinder bead mill). Control is simple.

The swing-batch procedure involves two vessels whose contents are alternately passed through the agitator mill. The circulation procedure is even simpler to control because the product can be recirculated until the desired quality and properties have been obtained. Only one vessel is required per machine, which is an advantage with frequent changes of product. A disadvantage is the backmixing which occurs in the vessel, making the dwell time distribution less favorable than in the pendulum procedure. Above about 20 cycles or passes the difference is negligible and the particle size distributions in the two procedures are

similar.

7.3.5. *Roller Mills*

Like agitator mills, roller mills are also used to disperse and homogenize highly viscous pigment pastes (Fig. 20). Roller mills are often also used for finishing (postdispersion or fine dispersion). The fact that dirt and coarse particles float to the top in the feed nip is a beneficial side effect: the roller mill acts as a filter.

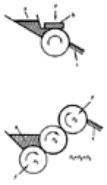

Figure 20. Roller mills (one- and three-roller)

a) Filling; b) Compression bar; c) Scraper

Roller mills may have one to five smooth rollers (rolls). Triple-roll systems are most often used, the rolls being arranged in various ways (e.g., horizontal, inclined, or chair arrangement). In modern high-speed systems the axis of the third roll is generally fixed in position, while the first two rolls are hydraulically actuated. The rotational velocities are 40 – 100 min^{-1} for the front roll, 140 – 200 min^{-1} for the middle roll, and 300 – 600 min^{-1} for the third roll. Shear energy for dispersion depends on rotational speed differences, and the roll nip clearance (controlled by the compression force). In order to compensate for viscosity fluctuations the roller mills are generally designed so that they can be heated or cooled.

7.3.6. *Auxiliary Equipment*

Sieves. Paints containing undesirably coarse particles must be screened. The simplest method is to sieve them through metal (steel) or plastic fabrics of specific mesh size. Polyamide screen bags are employed to facilitate handling. Metal fabric is fastened or soldered onto cylindrical metal frames.

Vibrating screens are used for products (e.g., metallic effect paints) that are difficult to screen with fabrics. These screens have a driven cam which causes the oscillating frame to roll. The material is vibrated horizontally and flows rapidly through the sieve.

Filters. In the production of intermediates, varnishes, and synthetic resin solutions, turbidity sometimes occurs which has to be eliminated by filtration. The simplest method is filtration through *bag filters* made of needled nonwovens. The bag filters can be installed in metal filter housings with closable, sealable covers. A perforated sheet-metal basket protects the filter bag against overloading.

If impurities cannot be removed with bag filters, *filter cartridges or candles* are used. The filter cartridges are inserted in individual housings or, in large-scale equipment, up to 24 cartridges are inserted side by side or one above the other and held in place with quick-release mechanisms. Resin-bonded, needled nonwovens or, in special cases, sintered metals are used as filter material. The fineness of the filter fabric (5 – 100 µm) determines the filtration effect. Water is first pumped through the filtration equipment to remove loose fabric constituents, lint, etc. from the filter. If serious problems arise (gel particles or turbidity), layered filters such as plate and disk filters with or without a diatomaceous earth bed are used.

An extremely time-consuming and expensive method of removing unwanted coarse material ("oversize") from paints and paint pastes, is *centrifugation.* This technique should only be used if filters and screens prove unsuccessful because it is often accompanied by a high product loss.

With the increasing intensification of quality assurance measures before (incoming inspection, FMEA) and during manufacture (cleanliness tests, on-line particle size measurement) the use of auxiliary equipment is declining or limited to a protective function ("policing filtration").

[Top of Page]

8. Paint Application

Bernd D. Meyer and Klaus Werner Thomer

8.1. Types of Substrate [419]

All solid substrates can be coated, including metals, plastics, wood, minerals, paper, and leather.

The surface structure of *metals*, their melting point, and susceptibility to the effects of weathering and chemicals (corrosion) can substantially influence the coating process and quality.

The coatability of *plastics* depends to a large extent on the material, design of the article, forming process, and the processing agents used for this purpose.

The coating of *wood* is strongly influenced by its structure and properties (elasticity, hardness, strength, shrinking tendency). Some wood constituents (e.g., pigments, resins, or tanning acids) can interfere with the coating process. See also ⟶ Wood, Surface Treatment.

Coatings are used for different purposes:

1. Protection of an object against its environment (e.g., corrosion protection)
2. Improvement of appearance
3. Surface modification to meet special physical and functional requirements

The choice of coating technology and the quality of the coat applied to a given article primarily depends on the substrate, its geometry, design, and surface state.

8.2. Pretreatment of Substrate Surfaces

The surface of workpieces may be covered or contaminated with coherent or discontinuous layers of the following materials:

Lubricants	mineral oil products
	metallic soaps
	graphite
	molybdenum disulfide
Corrosion-inhibiting oils	petroleum sulfonates
	inhibitors
	mineral oil products
Fluxes	inorganic salts
Polishing pastes	mineral oils
	waxes
Scale, corrosion products	metal oxides
	metal salts
Finely divided, foreign substances	dust
	ash

Miscellaneous	slag (from welding work) dermal deposition from fingers/hands absorbed gas films old coatings adhering moisture

These foreign materials all have adverse effects on the wetting behavior, adhesion, and corrosion resistance of coating materials. The purpose of treatment prior to coating is therefore to remove foreign material from the workpiece surface. Another important function is to create a reproducible surface state that ensures long-term protection against corrosion subsurface migration and paint flaking at weak points or locally damaged areas of the paint film.

8.2.1. *Pretreatment of Metallic Substrates*

Many processes and techniques are used for pretreating metallic substrates (Table 23), [419-422].

Table 23. Summary of pretreatment methods

Method	Material			Impurities						Temporary corrosion	Improved corrosion
	Fe	Al	Zn	Lubricants, corrosion protection oils	Scale, rolling skin	Dirt, shavings	Drawing agents	Corrosion products	Old paint layers	protection after treatment	protection with coatings
Mechanical methods											
Steam and high-pressure water blasting	+	+	+	+	–	+	O	–	*	–	–
Jet blasting (e.g., wet jets)	+	+	+	–	++	+	O	++	+	—	–
Manual abrasive methods (e.g., brushing, grinding)	+	+	+	–	+	+	O	+	+	—	–
Wet chemical methods											
Solvent degreasing											
Steam	+	+	+	+	–	O	–	–	O	–	–
Dipping	+	+	+	+	–	–	–	–	O	–	–
Ultrasonic dipping	+	+	+	+	–	+	–	–	O	–	–
Alkaline degreasing											
Spraying	+	+	+	+	–	+	–	–	–	–	–
Dipping	+	+	+	+	–	O	–	–	–	–	–
Pickling											
Phosphoric acid	+	+	+	*	*	+	–	+	–	O	*
Sulfuric acid	+	–	–	*	+	+	O	+	O	—	–
Hydrochloric acid	+	–	–	*	+	+	O	+	O	—	–
Caustic soda	–	+	–	*	*	*	*	+	+	–	–
Phosphating											
Alkali phosphate	+	+	+	*	–	*	–	–	–	+	+
Zinc phosphate	+	+	+	–	–	–	–	–	–	+	++
Chromating	–	+	+	–	–	–	–	–	–	+	++
Adhesion promoters (wash primer)	+	+	+	–	–	–	–	–	–	++	++

The symbols have the following meanings: ++ Extremely satisfactory. + Satisfactory. O Moderately satisfactory. – Unsatisfactory. — Particularly critical.* Satisfactory only with special methods.

8.2.1.1. *Cleaning*

Mechanical cleaning is mainly used to remove scale, rolling skin, rust, burr, and paint residues. Hammering, brushing, and lapping jelly are used in smaller workshops and on building and construction sites. Grinding is also widely employed for finishing weld seams and soldering points, smoothing and leveling file or tool marks, as well as for roughening material surfaces to improve paint adhesion. Blasting methods (sandblasting, compressed air, or water jet) remove impurities as a result of the kinetic energy of small hard particles. Abrasives may be metallic (e.g., spherical cast iron shot) or nonmetallic products (e.g., corundum). Blasted and abraded surfaces are extremely dusty and have to be cleaned by blowing, wiping, or washing.

Chemical Methods. See also ⟶ Metals, Surface Treatment – Cleaning. Chemical methods include pickling, which can be used to remove scale and metal oxides. Sulfuric acid and hydrochloric acid are usually used to treat steel and zinc, while nitric acid, nitric acid – hydrofluoric acid, or sulfuric acid – hydrofluoric acid are used for aluminum. Although pickling in phosphoric acid is somewhat slower, it has the advantage that a thin conversion layer of the substrate phosphates is produced, which provides temporary corrosion protection and is a good base for the subsequent coating. Previously, sulfuric acid was preferred

for pickling because of its low vapor pressure; acid losses are therefore slight and environmental pollution is low. Nowadays the tendency is to use hydrochloric acid because it allows better cleaning of the metal surface (even slightly alloyed steels). Evaporation of hydrochloric acid from the pickling baths is limited by using self-contained or completely enclosed units. Organic inhibitors are generally added to pickling acids to limit the pickling action to the oxidic impurities and to minimize the dissolution of the base material. Addition of surfactants has a limited degreasing effect.

Pickling in aqueous caustic soda solution is widely used for cleaning aluminum workpieces on account of their amphoteric behavior. However, after alkaline pickling the article generally has to be treated with acid to remove loosely adhering layers of pickling sludge and to brighten the surface.

8.2.1.2. *Degreasing*

Processes using aqueous solutions or organic solvents have become extremely important for removing organic impurities. In special cases, salt melts and treatments at elevated temperature in the gas phase are also employed. See also ⟶ Metals, Surface Treatment – Cleaning.

Organic Solvents. Chlorinated hydrocarbons are widely used. They remove oils and greases extremely effectively (generally in the vapor phase). Cleaning in immersion baths can be significantly improved by using ultrasound. Increasingly stringent environmental protection legislation will, however, greatly restrict the future use of chlorohydrocarbon solvents.

Aqueous Media. Degreasing may be performed with alkaline, neutral, or acidic aqueous media. Approximately neutral (mild alkaline) degreasing agents with fairly high concentrations of wetting agents and surfactants are mainly used. An advantage of these agents over alkaline or acidic media is a simpler and more economical effluent treatment. The degreasing agents hydrolyze animal and plant oils and grease. Non-hydrolyzable components (mineral oils and grease) are dissolved and dispersed by adding colloidal emulsifiers and wetting agents. These baths are operated at 60 – 80 °C and a pH of 8 – 9. The concentration of cleansing agents is between 1 g/L (spraying methods, pressure 0.15 – 0.25 MPa) and 50 g/L (dipping methods). Both stationary and flow-through baths are used.

8.2.1.3. *Formation of Conversion Layers [419-423]*

Conversion layers generally consist of inorganic compounds formed on the metal surface. They are used to increase the corrosion resistance and to improve the paint adhesion of the metal surface.

Industrially, phosphate layers are the most important and phosphating is used to treat steel, aluminum, and zinc. Chromating produces layers containing trivalent or hexavalent chromium compounds and is mainly used with aluminum and zinc. Special oxide layers and inorganic – organic coatings are used for special purposes in strip treatment.

The surface weight of conversion layers is 0.05 – 5 g/m^2. With higher surface weights the flexibility of the layers decreases, which has an adverse effect on the flexural adhesion of the organic coating.

Phosphating Processes. See also ⟶ Metals, Surface Treatment – Phosphating. The most important phosphating processes are alkali, zinc, and zinc – calcium phosphating. In alkali phosphating the layer-forming cation originates from the substrate, in zinc phosphating processes it originates from the phosphating solution.

Alkali phosphating (iron phosphating) is mainly used when corrosion protection does not have to satisfy stringent requirements. The solutions (pH 4 – 6) consist of acid alkali phosphates, free phosphoric acid, and small amounts of additives; oxidizing agents (e.g., chlorates, chromates, or nitrites), condensed phosphates (e.g., pyrophosphate or tripolyphosphate), and special activators (e.g., fluorides or molybdates). The first reaction is the pickling reaction which produces Fe^{2+} ions from the substrate (steel). These ions react with phosphate ions from the solution to form sparingly soluble iron phosphate that precipitates and adheres strongly to the metal surface. Zinc phosphate layers are formed in an analogous reaction sequence on zinc surfaces. Aluminum is usually treated with fluoride-containing solutions; thin, complex coatings are formed that contain aluminum, phosphate, and fluoride. The baths are adjusted to a concentration of 2 – 15 g/L. Contact with the surface may take place via spraying, flooding, or dipping. The bath temperature is normally 40 – 70 °C, but can be lowered to 25 – 35 °C with special bath compositions. Treatment times are 5 – 10 s (spraying of strip material) and 1 – 3 min (spraying or dipping of individual parts). Iron phosphating includes both thin-coating (0.2 – 0.4 g/m^2) and thick-coating methods (0.6 – 1.0 g/m^2). The color of the layers is blue-green, and in some cases also reddish iridescent. The surfaces become matter and grayer with increasing coating weight.

Zinc phosphating is primarily used for the surface treatment of steel and zinc as well as composites of these metals with aluminum. Aqueous phosphoric acid solutions (pH 2.0 – 3.6) containing dissolved acidic zinc phosphate, $Zn(H_2PO_4)_2$, are used. For further details, see ⟶ Metals, Surface Treatment – Chemistry.

The phosphate layers are gray in color (weight 1.2 – 6.0 g/m^2) and consist of $Zn_3(PO_4)_2 \cdot 4\,H_2O$ (hopeite), $Zn_2Fe(PO_4)_2 \cdot 4\,H_2O$ (phosphophyllite), and $Zn_2Ca(PO_4)_2 \cdot 4\,H_2O$ (scholzite). Layer formation is complete when the metal is completely covered with a phosphate layer, and the pickling action initiating layer formation has stopped. The treatment baths contain 0.4 – 5 g/L of zinc and 6 – 25 g/L of phosphate, calculated as P_2O_5.

The phosphating baths are usually used in automatic or semiautomatic dipping, spraying, or flooding plants at 45 – 70 °C; low-temperature processes operating at 25 – 35 °C also exist. Treatment times are 60 – 120 s (spraying process) and 3 – 5 min (dipping process). The iron phosphate sludge must be removed periodically or continuously from the bath.

Low-zinc processes were developed with the introduction of cathodic electrodeposition coating. In the normal zinc processes flat, sheetlike crystallites (mainly hopeite) are formed which may project from the surface. In the low-zinc process the layers mainly consist of phosphophyllite. They have a parallel orientation relative to the metal substrate and are more finely crystalline and compact than the hopeite layers. Very thin layers with a higher iron content are produced with nitrite-free low-zinc processes.

Chromating. See also ⟶ Electrochemical and Chemical Deposition – Chromating In chromating, metal surfaces (mainly aluminum and magnesium) are brought into contact with aqueous acid solutions of chromium(VI) compounds and additives that activate and accelerate pickling. The pickling reaction converts acidic Cr(VI) into basic Cr(III); cations of the treated metal simultaneously accumulate in the liquid film on the metal surface leading to precipitation of a gel-like layer containing chromium (III), chromium(VI), cations of the treated metal, and other components. The final conversion layer is formed after aging and drying. Treatment can be carried out by spraying or dipping (6 – 120 s) at 25 – 60 °C. Effluent waste must be treated to remove Cr(VI); the most common method being reduction with sulfite to form Cr(III) followed by precipitation of chromium (III) hydroxide with milk of lime.

For the *yellow chromating* of aluminum, solutions containing chromium(VI) compounds as well as simple or complex fluorides and activators are used to accelerate layer formation. The pH value is 1.5 – 2.5 at total bath concentrations of 5 – 20 g/L. The conversion layers consist of oxides or hydrated oxides of trivalent and hexavalent chromium and aluminum. The color of the layer may range from colorless through yellowish iridescent to yellowish brown, corresponding to an increase in the surface weight from 0.1 to 3 g/m^2.

The essential constituents of the aqueous solutions for *green chromating* are chromic acid, fluorides, and phosphates. The pH value of the baths is slightly less than in the case of yellow chromating, and the bath concentration is normally 20 – 60 g/L. The conversion layers consist largely of chromium(III) phosphate and aluminum(III) phosphate, with small amounts of fluorides and hydrated oxides. The surface weight is 0.1 – 4.5 g/m^2, and the color ranges from iridescent green to deep green.

Aqueous, chromium-free acidic solutions have also been developed for aluminum materials that may contain complex fluorides of titanium and zirconium, phosphate, and special organic compounds. These solutions are applied by spraying or dipping (up to 60 °C) and produce thin, almost colorless conversion layers with a surface weight < 0.1 g/m^2.

Aqueous chromic acid solutions containing chlorides, simple and complex fluorides, sulfates and formates as activators, are used for *chromating zinc.* The total bath concentration is 5 – 30 g/L, the pH value 1.2 – 3.0. The bath temperature is in the range 25 – 50 °C, the process times are 5 – 120 s, the achievable surface weight is 0.1 –3 g/m^2. The layer color changes with increasing surface weight from iridescent through yellow to brown or olive green.

Rinsing. A passivating rinse is necessary to exploit the quality-improving properties of conversion layers. The most important rinsing agents are dilute aqueous solutions of chromic acid, optionally with additional amounts of chromium(III). Equally good or only slightly worse results can be obtained with solutions free

from chromic acid and containing polyvalent metal cations (e.g., chromium(III) or also organic components). The concentration of active substances in the rinsing baths is 100 – 250 mg/L. Rinsing is performed at 20 – 50 °C, treatment time ranges from a few seconds to about a minute. The surfaces should be sprayed with demineralized water as a last rinse to prevent crystallization of water-soluble salt residues.

8.2.2. *Pretreatment of Plastics [419-421], [424-426]*

Pretreatment of plastic surfaces is necessary for the following reasons:

1. To increase adhesion strength
2. To reduce the concentration of interfering constituents and mold release agents
3. To eliminate surface defects (e.g., bubbles)
4. To remove interfering foreign substances
5. To increase electrical conductivity

Many pretreatment techniques are used in practice (Table 24). The normal physical method used to improve the adhesive strength of the coating to the substrate is to slightly roughen the surface by solvent treatment, abrasion, or blasting. Some plastics (e.g., polyolefins) require special pretreatment methods; processes that modify the surface molecular layers of the plastic to increase their polarity have proved suitable (e.g., flaming, immersion in an oxidizing acid, immersion in a benzophenone solution with UV irradiation, corona treatment, plasma treatment).

Table 24. Pretreatment methods for plastics

Pretreatment methods		Use
Physical and mechanical methods		
Abrasion	dry, or wet dust cloth	removal of contamination (cleaning)
Blowing-off	oil- and water-free compressed air	
	ionized compressed air	+ reduction of electrostatic charge
Sanding		
Blasting		increase in adhesive strength, elimination of surface defects and foreign substances
Steam degreasing		increase in adhesive strength, elimination or reduction of interfering constituents and adhering process auxiliaries, as well as foreign substances
Washing	solvent, aqueous surfactant solutions	
Spraying	antistatically adjusted solutions	+ reduction of electrostatic charge
Dipping	conducting solutions	+ increase in electrical conductivity
Chemical methods		
Oxidative	flame treatment	increase in adhesive strength in special plastics, particularly polyolefins
	corona discharge, plasma treatment	
	oxidizing acids	
Cross-linking	benzophenone solution with UV irradiation	
Miscellaneous methods		
Tempering		elimination or reduction of surface defects, constituents, or process auxiliaries
Storage, aging		
Application of a dielectric layer		reduction of electrostatic charge

Corona discharge is performed in a high-frequency alternating field (14 – 40 kHz) at 10 – 20 kV between two electrodes. The plastic surface is oxidized in a very short period (milliseconds). Plasma treatment is carried out under a moderate vacuum down to ca. 10 Pa. The advantage of this technique is the better penetration depth and the fact that it is also possible to treat shaped parts more easily. The plasma can also burn in gases (e.g., argon), whereby special effects (e.g., plasma polymerization) can be achieved.

Adhering processing additives (lubricants, release agents) can be removed by cleaning with solvents or aqueous surfactants. The solvent stability and solvent and water absorption of the plastic material should be taken into account. In order to reduce migration of constituents (shaping agents, plasticizers, dyes, organic pigments, stabilizers) during and after coating, preliminary tempering is often recommended (at the same time surface defects can also be detected).

Surface defects resulting from production (e.g., pores, bubbles, flow seams and projecting fibers) are rectified by surface appearance enhancement. Deeper-lying defects are filled in and smoothed with putties.

Elimination of foreign substances (e.g., dirt particles and fibers) is very difficult due to the electrostatic charge of the plastics material. Alternative methods are wiping with a lint-free cloth wetted with water or a water – alcohol mixture or blowing with ionized compressed air. Plastics can also be made permanently antistatic by applying a dielectric coating.

8.2.3. *Pretreatment of Wood [419], [420]*

Properly graded sanding with appropriate sandpapers is a prerequisite for a satisfactory wood surface. Industrially, sanding is performed on cylindrical abrasive-belt machines or automatic grinders, followed by brushing and suction to remove abrasive dust.

Surface pretreatment includes the following steps:

1. Removal of resins, e.g., by hydrolysis (wood soaps or soda solution) or dissolution (e.g., alcohol, acetone)
2. Removal of adhesive residues
3. Rectification (filling, patching) of processing and growth defects in the wood
4. Structuring by brushing, burning, sandblasting, embossing, or leaching
5. Staining with dyes dissolved in water or solvents, or with pigment dispersions

8.3. Application Methods

Many techniques have been developed for the industrial application of coatings. The individual industrial coating methods can, however, only be employed in limited areas if design and production sequence are not matched to the requirements of the coating technique. Adoption of more environmentally friendly coating methods is therefore often less a problem of investment, than a problem of the application limits of the relevant processes. Analysis of the criteria for choosing a coating method is a complex task. Of particular importance are the workpiece (design, material), the coating material, the number of workpieces and batch sizes, range of workpieces, requirements demanded of the coating (e.g., decorative appearance, corrosion protection), economic factors, legal provisions, and available facilities, premises, and equipment. Economic factors generally have top priority for choosing an application method on a commercial basis. Coating systems and processes are usually preferred that best satisfy the demands and requirements for thin coatings, high degree of material utilization, low energy costs, and good automation. Modern coating methods that best comply with these requirements are electrodeposition coating, electrostatic atomization, and electrostatic powder spraying.

8.3.1. *Spraying (Atomization) [419-421], [427]*

In conventional spraying, atomization is the result of external mechanical forces, i.e., the exchange of momentum between two free jets (air and paint). See also ⟶ Spraying and Atomizing of Liquids. Atomization may be classified as compressed air atomization (air 0.02 – 0.7 MPa, paint 0.02 – 0.3 MPa), airless atomization (paint 8 – 40 MPa), and air-assisted airless atomization, also termed airmix process (air 0.02 – 0.25 MPa, paint 2 – 8 MPa), and special technologies (Table 25).

Table 25. Nonelectrostatic atomization methods

Advantages	Disadvantages	Examples of areas of use
Compressed air (pneumatic) atomization		
Universally employable	spraying experience required	large-scale series coating applications (automobile industry)
Simple to use	paint mists constitute a health hazard	repair and touch-up finishes
Uniform layer thicknesses	ventilation required in enclosed spaces	industrial coatings (furniture, domestic appliances, etc.)
Applicable to complicated work-piece geometries	compressed air supply required	
Small amounts can be applied	unsatisfactory material utilization	
Rapid change of colors	danger of film defects due to spray mist	
Very good optical paint film quality		
Suitable for special-effect paints		
Airless (hydraulic) atomization		
Very high operating speed	expensive apparatus	coating of large objects (shipbuilding, steel construction work, machines, lorries, etc.)
Low spray mist formation (low losses)	equipment parameters must be matched to the coating material	
Large film thicknesses in one application	limited number of spray jets due to overlapping	
Uniform surface and film thickness	amount can be regulated during application	
Suitable for high-viscosity paints	nozzle subject to high degree of wear	
Direct application from as-supplied drums and containers	danger of sagging with sensitive paints	
Substrate with deep pores can be wetted		
Airmix atomization		
Combines advantages of pneumatic and hydraulic atomization	paint pressure generator and compressed air supply required	industrial coating
Hot spraying		
High-viscosity paints can be applied	trained workforce required	industrial coating
Large film thicknesses	additional heaters necessary	
Low sagging tendency		
Quicker drying of the paint film		
Two-pack coating		
Advantageous for temperature-sensitive workpieces	expensive and complicated equipment	industrial coating, coating of wood and plastics
Hardening at room temperature (energy saving)	exact metering and mixing required (automatic monitoring)	protection of buildings and structures
Better paint film quality	safety measures required with isocyanate-containing systems	corrosion protection
High resistance to mechanical, chemical, and climatic influences	trained workforce	large equipment and apparatus (machines, aircraft, ships, commercial vehicles, etc.)
Low content of organic solvents		

In *compressed air* (*pneumatic*) *atomization*, compressed air flows through an annular gap in the head of the spray gun that is formed between a bore in the air cap and the concentric paint nozzle. Further air jets from air-cap bores regulate the jet shape and assist atomization. The expanding compressed air leaves the paint nozzle at high velocity. A low-pressure area is formed in the nozzle aperture which exerts a suction effect and assists outflow of the paint. The difference between the velocities of the compressed air and the exiting paint atomizes the paint into particles that are conveyed as spherical droplets in the free jet. In the high-pressure process (0.2 – 0.7 MPa) the exiting air jets can atomize the paint material extremely finely. The size of the liquid droplets varies from ca. 10 µm to 100 µm (depends on the liquid viscosity, amount of delivered paint, and air pressure). In the low-pressure process (0.02 – 0.2 MPa) atomization is correspondingly coarser (20 – 300 µm). Depending on the viscosity and throughput, the paint can be fed to the nozzle via a suction cup, a pressure cup, a flow cup, or pressure tank.

In *airless* (*hydraulic*) *atomization* the paint is forced through a slit nozzle of hard metal under high pressure (8 – 40 MPa). On account of the high degree of turbulence, the paint stream disintegrates immediately after leaving the fluid tip. A similar atomization process occurs in spray cans where the paint pressure is produced by the propellant gas.

The combined *airmix process* operates at a lower paint pressure (2 – 8 MPa). Additional low-pressure air jets (0.02 – 0.25 MPa) from the air-cap bores impinge on the spray jet to mix and homogenize it. In addition to the atomizer and a compressed air generator (airless pump), the airmix unit therefore also requires compressed air for postatomization. Advantages over the airless method are the less sharply defined spray jet and the smaller droplet size. Compared with compressed air atomization, a low-mist coating is possible.

Hot spraying can be combined with all of the spraying methods described above and is used for large film thicknesses or highly viscous, high-solids paints (lower solvent consumption). The paint is heated to 50 – 80 °C in a heat exchanger. Immediately after atomization the heat content of paint droplets is transferred to the air and the workpiece. The droplets therefore cool and their viscosity rapidly increases; the risk of sagging at larger layer thicknesses is thus reduced.

In *two-pack paints* both the binder and the hardener have to be mixed before application. In paints with a short pot life the paint must be metered and mixed in the atomization equipment immediately prior to use. The two reactive components are normally mixed in static mixers after metering.

8.3.2. *Electrostatic Atomization [419-421], [427]*

In *purely electrostatic spraying*, the paint is atomized solely by electrostatic forces. In *electrostatically assisted spraying*, atomization takes place by the methods described in Section Spraying (Atomization) , with simultaneous or subsequent electrical charging.

In all electrostatic coating methods an electric field is applied between the atomization equipment and the workpiece. The advantages and disadvantages of this technology are listed in Table 26. The paint is electrically charged by a concentrated electric field at a high-voltage electrode. In “lead charging” the nonatomized paint is charged by direct contact with the electrode. In “ionization charging” mechanically produced paint droplets are charged by attachment of ions from the air; the electrode serves as a corona tip that generates these ions.

Table 26. Electrostatic atomization methods

Advantages	Disadvantages	Examples of areas of use
Purely electrostatic spraying methods		
Very high application efficiency Relatively low investment and operating costs Low wear of structural parts	unsuitable for parts with complex shapes low paint throughput limited choice of coating materials high consistency of paint data low flexibility	housings and flat parts in the electrical industry (washing machines, refrigerators, switching cabinets)
Electrostatically assisted rotation atomization		
High application efficiency Fine atomization Practically all paints can be applied	expensive plant and safety technology unsuitable for parts with complex shapes excessive edge coating may occur	large-scale series coating (automobile industry), industrial coating
Electrostatically assisted compressed air, airless, and airmix atomization		
As for nonelectrostatic methods but with higher application efficiency	coarser atomization otherwise as for rotation atomization	industrial mass-produced articles, handicrafts

In *purely electrostatic spraying* the paint is atomized solely by electric field forces. The paint flows as a thin film over a high-voltage sharp edge, where it is subjected to high field forces. The paint film breaks up into threads and then into charged droplets that follow the electric force lines to the workpiece. Only paints with a moderate viscosity and an electrical conductivity in the range 5×10^{-6} to 5×10^{-9} S/cm can be applied in this way. The best known designs are the electrostatic spray gap (AEG method), the electrostatic spray cone (diameter 70 – 250 mm, max. rotational speed 1500 min^{-1}), and the electrostatic spray disk (diameter 400 – 700 mm, max. rotational speed 3000 min^{-1}). Purely electrostatic spraying methods are only used in special cases on account of their limitations (workpiece geometries, type and throughput of the paint).

Electrostatically assisted atomization methods are more versatile than purely electrostatic methods because atomization takes place mechanically. The electric field serves only to charge the paint material and to transport the charged droplets to the workpiece. The following systems are used:

1. Electrostatic high-speed turbo bells (diameter 30 – 80 mm, rotational speed 15 000 – 40 000 min^{-1}, voltage 70 – 120 kV)
2. Electrostatic high-speed disks (diameter 150 – 250 mm, rotational speed up to 20 000 min^{-1}, voltage 70 – 120 kV)
3. Electrostatically assisted atomization guns (charging by means of needle-shaped electrodes arranged directly in or on the paint nozzle, voltage 50 – 100 kV).

The extremely high electrical conductivity and high dielectric constant of waterborne paints should be taken into account. The paint supply system must be electrically insulated to prevent short circuiting. Ionization charging or, in the case of automatic equipment, the installation of an insulated paint supply system have proved of value.

8.3.3. *Dipping [419][420][421], [428]*

Dip coating is one of the simplest and oldest coating methods. In addition to dipping in solvent- or waterborne paints, electrodeposition has become important for large-scale series production (Table 27).

Table 27. Dipping methods

Advantages	Disadvantages	Examples of areas of use
Conventional dipping		
Simple process Can easily be automated High application efficiency High economy Low wage costs High article throughputs	special plant and equipment technology unsuitable for many shades ventilation and fire prevention measures required with solvent-borne paints paint analysis control edge coating often unsatisfactory danger of sagging, dripping, paint splashing foam and bubble formation	priming or one-layer coating of mass-produced articles
Electrodeposition coating		
Complete and uniform paint film, also in cavities (wrap-around) Very high material utilization	complex equipment unsuitable for more than one shade	priming in the automobile industry priming of mass-produced articles
No sagging or droplet formation Fully automatic operation High parts throughput Environmentally friendly (waterborne paints)	complex bath monitoring highly trained workforce high material costs	

Conventional Dipping. In conventional dipping the workpieces are immersed in the paint and then removed (Fig. 21). The liquid paint adheres to the surface and is then dried or stoved. Care should be taken to ensure that the workpieces do not float during dipping and that air bubbles do not become trapped. The speed at which the workpiece is removed from the bath must be selected so that excess paint adhering to the surface can run off. The draining and evaporation time must be sufficiently long to ensure satisfactory evaporation of the solvents (if necessary a hot air zone should be included for waterborne paints).

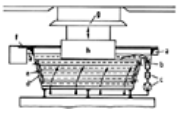

Figure 21. A conventional dipping unit

a) Edge suction; b) Overflow; c) Circulation system with pump, filter, piping, and nozzles; d) Dipping tank; e) Heating and cooling device; f) Cover (in the form of roller-type covers); g) Raising and lowering system; h) Workpiece

Electrodeposition. Electrodeposition paints are suspensions of binders and pigments in fully demineralized water with low concentrations (ca. 3 %) of organic solvents (see Section Electrodeposition Paints). Electrodeposition coating may be either anodic or cathodic.

In *anodic electrodeposition* the workpiece acts as the anode. This method is only used to a small extent. Disadvantages compared with cathodic electrodeposition are its poorer handling and corrosion protection. Advantages include the lower paint price and lower expenditure on plant technology.

In *cathodic electrodeposition* the workpiece acts as the cathode; this method is more important than anodic electrodeposition. The binders consist largely of non-water-soluble epoxy resins, and to a lesser extent of acrylic resins (one-layer coatings). These resins are converted into a water-soluble (i.e., ionized) state by neutralization with organic acids (e.g., acetic acid):

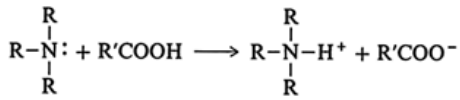

Water is decomposed by electrolysis at the electrodes

$$2\,H_2O \rightarrow O_2 + 4H^+ + 4\,e^-$$

and at the workpiece

$$2\,H_2O + 2e^- \rightarrow H_2 + 2\,OH^-$$

Iron from the high-grade steel anodes is also oxidized at the anode and passes into solution. Hydroxyl ions are formed at the cathode and react with the solubilized resin causing reversal of the neutralization:

$$H_2O + 3\,H^+ + 4\,e^- \longrightarrow 2\,H_2 + OH^-$$

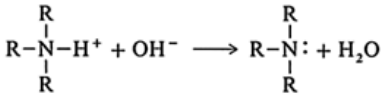

The binder then coagulates and is deposited as an irregular, porous layer on the workpiece. It is converted into a uniform, sealed paint film by stoving.

Electrodeposition equipment and technology is expensive and is therefore practicable only for large-scale series coating (Fig. 22). In addition to the dipping tank and a storage tank, circulation systems for the paint and auxiliary materials, rinsing systems (ultrafiltrate, fully demineralized water), a regulated d.c. supply (200 – 400 V), and the workpiece transporting system (curent supply) also have to be installed. Since electrodeposition systems require temperature control, production tanks are equipped with heaters and chillers.

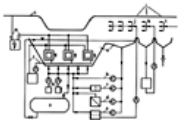

Figure 22. Cathodic electrodeposition coating unit

a) Dipping tank with overflow; b) Recycle circulation; c) Paint filter circulation; d) Paint cooling circulation; e) Ultrafiltration; f) Anolyte circulation; g) Anodes; h) Metering of paint; i) Rinsing system; j) Ultrafiltration rinsing; k) Recirculated material rinsing; l) Water rinsing; m) Power supply; n) Transportation system; o) Storage tank

8.3.4. *Miscellaneous Wet Paint Coating Methods [419-421], [428]*

Other wet paint coating methods are summarized in Table 28. Application with a *brush or roller* is now only used to a great extent in the handicrafts sector, do-it-yourself sector, or on building sites.

Table 28. Miscellaneous wet coating methods

Advantages	Disadvantages	Examples of areas of use
Brushing		
Simple equipment High paint yield No specially trained workforce required Universally applicable Good wetting of the substrate	highly labor-intensive (high wage costs) nonuniform film thicknesses danger of brush marks	steel superstructures lattice constructions handicrafts do-it-yourself
Roller application		
Fast and easy to master High paint yield Uniform film thicknesses	only suitable for smooth surfaces worse wetting of the substrate labor-intensive	steel superstructures handicrafts do-it-yourself
Wiping		
Fast application Uniform film thickness High paint yield	unsuitable for workpieces with complex shapes	exterior coating of pipes application of bitumen to pipelines wood coating
Rolling, printing, strip (coil) coating		
High degree of automation High paint yield High economy Very uniform coating	only suitable for flat surfaces (strips) high investment in plant and equipment limited potential uses	strip and panel coating (sheet metal, wood, films, paper, paperboard)
Pouring		
Very good material utilization Applicable for slightly curved parts Easily automated Uniform coating	special equipment not universally applicable	chipboard, paper, and cardboard coating
Troweling		
Easily automated Very low material losses	specially prepared materials only for flat, strip, or panel-type parts	wood panel coating
Flow coating		
Good material yield Easily automated	nonuniform film thicknesses danger of paint slurry formation	large, bulky articles (radiators, frames for commercial vehicles, etc.)
Centrifugation, drum application		
High paint yield Good economy	no special surface quality (nonuniform leveling, pressure points, etc.)	small mass-produced articles (hooks, eyelets, screws)

Flat workpieces (e.g., paper sheets, films, wooden boards, and panels) can be coated economically and quickly by rolling, pouring, or knife coating. These methods have become important because they are easily automated, have a high material yield, and are environmentally friendly. With *roller coating* the paint material is transferred from rotating rubber rollers to one or both surfaces of the workpiece (Fig. 23). In forward roller coating (layer thickness ≤ ca. 12 µm) the workpiece and paint application roller run in the same direction. With reverse roller coating (layer thickness 3 – 100 µm) they run in opposite directions.

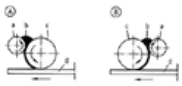

Figure 23. Roller coating

A) Reverse roller coating; B) Forward roller coating

a) Metering roller; b) Coating material; c) Paint roller; d) Workpiece

The *pouring method* is commonly used in the wood and timber trade. Here the paint is pumped into a pouring head that has a paint outflow slit with adjustable lips (Fig. 24). Paint that does not come into contact with the workpiece is returned to the tank via a collecting channel. Since the paint is constantly circulated and is heated by the circulation pump, it must be cooled and must contain high-boiling solvents.

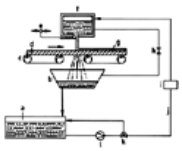

Figure 24. Pouring method

a) Paint tank; b) Drip pan; c) Transporting belt; d) Article being coated; e) Aperture regulation; f) Pouring head; g) Paint film; h) Pressure release value; i) Paint filter; j) Paint line; k) Valve for quantitative adjustment; l) Pump

Knife coating is used in the paper and textile industry for coating continuous material. Knife coating can also be used to print and coat flat, two-dimensional workpieces. The coating is applied to the substrate which then passes under a doctor knife. The coating material is pressed onto the workpiece with the doctor knife (Fig. 25). The knife also removes excess coating material and smooths the surface.

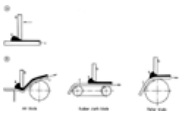

Figure 25. Knife coating

A) Coating of unit items (panels, plates, disks, etc.); B) Coating of semifinished articles (strips, sheets)

a) Coating material; b) Doctor knife; c) Substrate

Highly pigmented, pasty coatings and viscous putties and fillers are applied by *troweling*. The materials are applied with a pair of counterrotating rollers. Excess material is smoothed and “pressed” into the workpiece surface. This method can only be used for flat, striplike, or panel-shaped workpieces (e.g., wooden panels).

With *flow coating* gentle streams of paint are pumped over the workpieces via nozzles. Excess paint flows into a collecting trough and can be recirculated (Fig. 26).

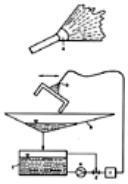

Figure 26. Flow-coating method

a) Paint nozzle; b) Drip pan; c) Filter; d) Paint adjustment valve; e) Pump; f) Paint tank

Small items (e.g., hooks, eyelets, clasps, and buckles) can be coated by centrifugation or drum coating. The workpieces must not adhere to one another, nor must they become entangled. In *centrifugation* the workpieces are placed in a wire basket, and then immersed in a paint bath and centrifuged at a rotational speed of about 500 min^{-1}. With *drum coating* paint is fed into a rotating drum containing the workpieces. The paint feed can be achieved by adding special, low-viscosity paint or by spraying with spray guns.

Coil coating denotes the continuous coating of cold-rolled steel strip (including galvanized strip) or aluminum with organic polymers. In automated plants the metal strip is first cleaned and chemically pretreated. The strip is then roller-coated on one or both sides, with one or more coats of liquid, thermosetting or thermoplastic coating materials. The paint coat is dried in an oven after each application. Throughput rates are of the order of 1 – 3 m/s. Constant production conditions using appropriate control devices ensure high quality.

8.3.5. *Powder Coating [420], [421]*

In powder coating the coating material is applied to the workpiece in the form of dry (i.e., solvent-free) thermoplastic or thermosetting powder. The powder particles are heated and melt to form a sealed film. The thermoplastic powders melt and fuse on heating whereas the thermosetting powders also become chemically cross-linked. Two main application processes are used: electrostatic spraying and fluidized-bed coating.

Electrostatic Spraying. The principle of electrostatic spraying is simple. A coating powder is a dust with a particle size in the range 10 – 80 µm. When dispersed (fluidized) in air the powder flows in the same way as a liquid and is applied in spray cabins using special spray guns onto the cold or hot workpiece. The coating powder adheres to cold workpieces electrostatically, whereas on hot workpieces it adheres by fusion. The workpiece generally passes through a dryer. The coating powder overspray is suctioned off, separated from the air, screened, and reused.

Care should be taken to minimize the length of pipes for powder transport and to maximize the deposition efficiency in order to avoid problems caused by a shift in the particle size distribution. A sufficiently large number of guns, a coating powder cloud size matching the size of the workpiece, adjustment of the oscillating movement of the gun supports, gun triggering, and a continuous powder metering system with maximum accuracy are advantageous. Minimizing the powder content of a plant facilitates cleaning.

Coating powders are mainly sprayed with guns with negative corona charging. Guns with friction charging (positive charging by charge separation on polytetrafluoroethylene), in some cases with additional corona charging are, however, also used. A new generation of corona-charged guns with ventilated deflector plates or without deflectors need to be cleaned far less frequently.

Further developments include systems with more uniform metering, low-voltage guns producing fewer free ions and therefore better penetration into corners and cavities (Faraday cages), and higher deposition efficiencies obtained by using slit-shaped nozzles and a triggered powder output that may be related to the shape of the object. Increasing the number of guns or reducing the gun output often improves coating application. Parts with complicated shapes, particularly those with Faraday cages, are extremely difficult to coat, at least with a constant layer thickness.

Good grounding (earthing) is necessary for satisfactory results. For safety reasons, earth leakage resistances of less than 1 MΩ (at 5 kV) are required. Good grounding can also be obtained via contactless, electrostatic processes.

A large proportion of coating powder plants operate partly or fully automatically. Many types of equipment are commercially available [429], [430]. The most important cabin type is the compact unit with replaceable filters in which the air that entrains and removes nonadherent powder is fed to a filter housing, flange-mounted on the rear wall. In the filter the powder is separated from the air and concentrated. The air is then vented into the atmosphere or directly recycled to the working area. If the color has to be changed the cabin is cleaned and a new filter housing is mounted. If the unit operates with one filter housing only, the filter housing is cleaned.

Criteria for spray cabins, powder recovery systems, and powder preparation systems are mainly determined by the time required to change colors. A compact

unit with one replaceable filter is used for one or a few colors. A compact unit with one or more replaceable filters or a filter carpet unit is used for several colors. For many colors a unit equipped with cyclones (or a filter carpet unit) or a compact unit with several replaceable filters should be used. If necessary, secondary colors can be applied without recovery or with liquid paints.

After the coating powder has been applied the workpiece can be passed directly to the dryer. An evaporation zone is not required as is the case for solvent-based paints. Since only the carrier medium air, and not a solvent, has to escape from the film, the coating powder can be heated much more quickly at the oven inlet than is the case with conventional solvent-based paints. Heating can be performed with IR zones or, after the coating powder has fused, with special blowing zones. Rapid heating results in short oven lengths and usually improves leveling and wetting. Since the stoving loss of coating powders is very low the ovens can be operated in a highly energy-efficient manner with a relatively low rate of exhaust air.

Fluidized-bed coating leads to substantially thicker layers than electrostatic spraying. The workpieces are preheated to 200 – 400 °C in an oven, dipped briefly (1 – 10 s) in vessels containing fluidized coating powder (particle size 40 –200 µm). The excess powder is removed by vibration, shaking, or blowing and the workpiece is optionally postheated and then cooled with air or water. The capacity of the fluidization basins ranges from a few grams to several tonnes.

The layer thickness is determined by the preheating temperature, the heat content of the parts, the dipping time, and the coating powder. Parts with complicated shapes or with different substrate thicknesses are difficult to coat. The layer thicknesses obtained by the various application methods and the resultant use profiles are summarized in Table 29.

Table 29. Layer thicknesses and typical uses in powder coating

Application method	Temperature of workpiece	Typical layer thickness, µm	Uses
Electrostatic	cold	25 – 100	interior: decoration, corrosion protection
Electrostatic	cold	50 – 120	exterior: weather resistance, decoration, corrosion protection
Electrostatic	cold	50 – 250	resistance to chemicals, electrical insulation, corrosion protection
Electrostatic	hot	150 – 600	
Fluidized bed, flocking	hot	200 – >1000	

8.3.6. *Coating of Plastics and Wood [420]*

Most coating methods described in Sections Spraying (Atomization) , Electrostatic Atomization , Dipping , Miscellaneous Wet Paint Coating Methods , and Powder Coating are also suitable for plastics and wood. The resistance to solvents and heat as well as the electrical conductivity of the plastics or wood must, however, be borne in mind.

Inmold coating (IMC) is an application technique used for plastics in which the coating material is applied during production of the plastics. The coating material is first injected (generally electrostatically) into the mold that is previously coated with a release agent. After the mold has been closed the plastic is then injected, foamed, and hardened. This method can be used for glass-fiber-reinforced plastics and polyurethane foams. A similar method is used with sheet molding compounds (SMC resin mats). After the resin mat has been inserted in the mold, the SMC is compressed and hardened. The mold tool is then opened and coating material is injected through the slit opening, compressed, and hardened.

8.4. Paint Curing Methods [419-426]

The performance of a coating depends on the chosen curing conditions. For example, mechanical and other properties associated with a given thermally cross-linking paint can only be guaranteed if curing is adequate. In undercured paint systems, adhesion to a subsequent layer may be adversely affected. If, however, the maximum curing temperature is exceeded the paint layer may become brittle or yellow. The curing time is also important.

Curing and hardening may be either physical or chemical. With a suitable combination of binders, both types of curing methods may proceed in parallel or overlap.

Physical Curing [419]. Physical curing occurs when polymers dissolved in organic solvents gradually cohere to form a solid film and then a network. Cohesion occurs solely as a result of solvent evaporation without chemical cross-linking. Such polymer films are generally reversible, i.e., they dissolve in the original solvent. Physically drying binders (e.g., nitrocellulose and its esters, vinyl resins, polystyrene, acrylate esters, chlorinated rubber, bitumen) are mostly chainlike or threadlike molecules with short side chains.

Chemical Curing [419]. During chemical curing film formation occurs as a result of formation of chemical bonds between the binder molecules. The binders become increasingly insoluble as cross-linking proceeds, and ultimately form irreversible (i.e., insoluble) thermosetting films. In solvent-containing systems physical curing also occurs simultaneously.

Curing methods may be divided into three groups:

1. Curing with a heat carrier (air)
2. Curing with radiation (IR, UV, electron beams, laser beams, plasma arc)
3. Curing by means of electrical processes (inductive curing, resistance, high-frequency and microwave curing)

Curing with Heat Carriers. Curing with a circulating hot air stream is the most important curing method. The heat causes the solvent vapors to evaporate from the coating film and they are removed by the air current. Heat transfer and thus film formation occur from the exterior to the interior. Paint curing is also possible with parts having a complicated shape. Since the whole workpiece has to be heated, long curing times and thus also large ovens are required. Heat consumption is relatively high because the workpiece, paint film, transporting device, parts of the conveying system, as well as the fresh air (maximum permissible concentration of combustible solvents in the oven is 0.8 vol %) all have to be heated.

Curing with Radiation. See also Section Radiation-Curing Systems. Radiation curing methods have become increasingly important in the last few years. Paint curing proceeds more rapidly than in circulating air curing since the whole workpiece does not have to be heated. However, only large flat parts can be satisfactorily treated.

Microwaves (frequencies 3 – 600 GHz, wavelength 0.5 mm – 10 cm) are generated in magnetrons and transmitted in hollow conductors. They exhibit wave effects such as interference which lead to localized concentrations of energy (wave peaks). Interaction between the polar material in the paint film and the electromagnetic alternating field is manifested macroscopically as a heating effect. Since microwaves are reflected by electrically conducting surfaces, this method can be used only for nonconducting substrates (plastics, wood, or paper).

Infrared (IR) radiation (wavelength 0.76 µm – 1 mm) is absorbed, reflected, or transmitted by an object. In the paint film absorbed radiation is converted into heat. The paint film cures from underneath and the paint surface does not harden initially (solvents can, however, still escape without any problem). The wavelength and intensity of the IR radiation must be matched to the paint being hardened. Absorption behavior is determined by the pigment, pigment volume concentration, and binder. Longwave IR radiation (4.0 µm – 1 mm) is absorbed by the pigments at the surface, while shortwave radiation (0.76 – 2 µm) can penetrate the paint. Normal thermal outputs are 5 – 25 kW/m^2, and may be up to 100 kW/m^2 if appropriate regulating equipment is used.

UV radiation with a wavelength range of 0.32 – 0.4 µm is used for curing paints. The UV radiation initiates photochemical reactions which lead to cross-linking. An added photoinitiator (sensitizer) decomposes in the paint into free radicals that initiate polymerization of the binder. Curing with UV radiation is of practical importance for hardening colorless polyester putties or primers, and offset and printing inks. The hardening times are of the order of a few seconds.

In pulsed radiation curing (PRC) exothermically reacting paints and printing inks are cured by UV pulses at a wavelength of ca. 197 nm. The pulses break the carbon double bonds of the binder because this wavelength range corresponds to their resonance frequency. A chain reaction starts throughout the whole layer, and curing takes place within a few seconds.

Electron beams are generated by applying an accelerating voltage (150 kV) to a thermionic cathode. An electron beam (ca. 6 mm diameter) is spread out into a curtain beam by a beam splitter. The electrons leave the beam distribution housing through a very thin metal sheet. When these electron beams strike binder monomers, they initiate polymerization in the paint film. Polymerization occurs in a fraction of a second and must be performed in a vacuum or in an inert gas atmosphere. The equipment must be screened to protect the operators.

Sharply localized curing can be achieved with *laser beams.* The carbon dioxide laser, which has a total beam output of 100 W/cm^2 at a wavelength of 10.6 µm, can be used for this purpose.

A *plasma arc* may also initiate cross-linking in a paint film. The high temperatures produced in the interior of the plasma arc are transmitted only to a small extent to the paint film.

Drying by Electrical Methods. In electrical methods electric current is directly converted into heat (resistance drying) in the workpiece or in the paint film.

In *inductive curing* an induction coil is located close to a metallic workpiece and generates eddy currents in the latter. The workpiece therefore becomes hot and the paint is heated from beneath the film; solvent loss or curing occurs.

In *high-frequency curing* the workpieces are arranged between two capacitor plates in a radiofrequency field (10^8 – 10^9 Hz). Molecules (dipoles) align themselves and are polarized in the alternating electric field. They therefore oscillate about their equilibrium position, resulting in heating.

[Top of Page]

9. Properties and Testing

Klaus Vogel

Anyone wishing to test the quality of a paint or coating quickly realizes that only a few properties can be accurately scientifically defined. In many cases there is a good correlation between defined physical properties and the behavior of interest to the scientist or practitioner. In some cases, however, it is impossible to obtain such a correlation. A large number of laboratory testing methods have therefore been developed for paints and coatings that are intended to simulate in-use conditions. These testing methods are often similar but their results are not fully comparable. Standard manuals provide a good overview of available test methods [431-435]. In this chapter attention is focused on methods that are widely known and internationally standardized, or whose international standardization is in progress.

9.1. Properties of Coating Materials

Many properties of liquid paints can be measured with considerable accuracy. Samples must be homogeneous. They must also be sufficiently large to be representative for a given batch. Impurities and permanent material defects (e.g., skin formation, a hard sediment, or gelling of the paint) can be detected. The most common investigations to which a sample is subjected before starting the tests are described in ISO 1513. Only those tests that are important for paint storage, transportation, and application are described here.

Viscosity. Although paint viscosity can be accurately measured with viscometers, paint consistency is normally assessed with *flow cups*. The time in seconds required for a known volume of paint to flow out of the cup through a jet is measured. Paints of higher or lower viscosity can be matched by using cups with different jet diameters.

Measurement of the run-out time from flow cups has been adopted worldwide since this test can be performed anywhere (e.g., in the laboratory, during production, or on a building site). Nationally standardized sets of flow cups are normally used in major industrial countries and give similar but not identical results. An internationally standardized system of flow cups has been introduced (ISO 2431) to overcome this problem. Since temperature fluctuations greatly affect the viscosity measurement, the flow cups should only be used in conjunction with thermostated jackets.

Measurements should be made at 23 ± 0.5 °C. According to ISO 2431 flow cups should only be used for substances exhibiting Newtonian flow. However, they are also often employed for near-Newtonian paints where the flow behavior at the desired viscosity deviates only slightly from Newtonian behavior (e.g., when adjusting the application consistency by dilution).

Viscometers are being increasingly used for accurate measurements on modern industrial paints, especially waterborne paints. Rotating viscometers with a concentric cylindrical geometry (Searle system, Fig. 27) are advantageous for paints. Precise thermostatic control is easily achieved because the outer cylinder does not rotate. The drive and torque sensor are combined to form a single unit with the rotating inner cylinder.

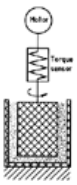

Figure 27. Rotating viscometer: Searle system

The Couette system (Fig. 28) is another concentric cylinder system with the advantage that the drive and sensor are separate. The motor drives the outer cylinder; the inner, stationary cylinder is connected to the sensor.

Figure 28. Rotating viscometer: Couette system

Viscometers with cylinder and cone/plate geometries can also be employed. The cylinder viscometers are easier to use and provide more reproducible results. Cone and plate systems can be used to investigate the hardening behavior of paints. The system can easily be cleaned and only a small amount of sample is required. High velocity gradients can be achieved with small cone angles. The potential uses of cone and plate systems are limited for several reasons and they cannot be used with dispersions [436], [437].

Falling ball viscometers and capillary viscometers are not generally used for testing paints. Flow cups are, however, special capillary viscometers with a short capillary in which the force of gravity acts on the paint.

Other Rheological Properties. Paint brushability, sagging, and leveling are highly dependent on viscosity and are usually evaluated subjectively in an application test.

Brushability (the ease with which a paint can be brushed) is evaluated by actual brushout of the paint by a painter. To evaluate the force required for brushing,

the viscosity should be measured at a high shear rate of about 10 000 s^{-1}.

Sagging is the downward movement of a paint film that occurs between the time of application and setting. The viscosity has to be considered at a shear rate of at most 1 s^{-1} (low-shear viscometers) in order to assess sagging on vertical walls. The run-out rate depends on the square of the layer thickness. Irregularities in the layer thickness can cause undesirable sagging (curtaining or crawling).

Leveling is the measure of the ability of a paint to flow out after application (e.g., to obliterate brush marks). It is measured by comb tests. Theoretical correlations have been established between viscosity measurements and practical results. Surface irregularities can be sensed mechanically and visualized with modern methods of image analysis [438].

Pot Life. The pot life is the length of time that a paint can be used after necessary preparations for application have been made. Twice the flow time measured with a flow cup is often used to assess the pot-life time.

Flash Point. The flash point of a liquid is a measure of the flammability of its vapors on application of an external flame. It is used to assess fire hazards. The flash point of paints may have to be measured to comply with legal requirements relating to the storage, transportation, and use of flammable products. According to ISO 1523, the flash point (closed cup) is the minimum temperature to which a product, confined in a closed cup, must be heated for the vapors emitted to ignite momentarily in the presence of a flame, when operated under standard conditions.

- ISO 1523 states that the Abel, Abel – Pensky, and Pensky – Martens cups satisfy the necessary requirements.
- ISO 3679 specifies another apparatus that provides similar results using a more rapid procedure and with a smaller test portion (2 mL).
- ISO 9038 aims to test the combustibility of paints, and involves the use of a Cleveland cup (ISO 2592) or a similar device.

Content of Nonvolatile Matter. Nonvolatile matter is defined in ISO 3251 as the residue left when the product is heated at an elevated temperature for a definite period under prescribed test conditions. A 1 g sample is evenly distributed in a flat-bottomed metal or glass dish (diameter 75 mm) and weighed before and after heating. Six different heating conditions are recommended in ISO 3251. The nonvolatile content of a paint is not an absolute quantity, but depends upon the temperature and period of heating used for the test as well as upon the test portion. If the test method specified by ISO 3251 is used, only relative values are obtained due to solvent retention, thermal decomposition, and evaporation of low molecular mass constituents. The method is therefore primarily intended for testing subsequent deliveries of a given product.

Density. Paint density is usually expressed in grams per milliliter at a reference temperature of 20 °C.

Pyknometer Method (ISO 2811). A pyknometer is filled with paint of known mass. The density is calculated from the mass of the paint and the volume of the pyknometer. Cylindrical metal pyknometers and glass Gay – Lussac or Hubbard pyknometers are suitable. The Hubbard pyknometer is especially useful for highly viscous paints.

Immersed Body Method (DIN 53 217 T 3). The density is calculated from the buoyancy of spherical bodies immersed in the paint. This method is used for low- and medium-viscosity paints and is particularly suitable for production control.

Vibration Method (DIN 53 217 T 5). A U-shaped tube is clamped at both ends and filled with paint. When the tube is subjected to vibration its resonance frequency depends on the mass in a given tube volume and thus on the density of the paint. The tube is calibrated with two media of known density and the density of the sample is then calculated from the resonance frequency. This method is used for liquid and pasty coating materials [439].

Fineness of Grind (ISO 1524). The fineness of grind of paints is determined with a gauge graduated in micrometers. The reading obtained on the standard gauge under specific test conditions, indicates the depth of the groove of the gauge at which discrete solid particles in the paint are readily discernible.

9.2. Properties of Coatings

9.2.1. *Films for Testing*

Preparation of Test Samples. Almost all properties of coatings depend on the layer thickness of the film. To obtain reproducible comparable test results, measurements must be made on carefully prepared films with a defined layer thickness. Films may be either free or applied to a substrate. In some cases film properties are also measured on films with different thicknesses.

Many properties of coatings also depend on the substrate and the method of application. The substrate and method of application must therefore be described in detail when citing results.

Paint-making conditions should be reproduced in the laboratory when necessary. Often it is also useful to take into account the production boundary conditions (e.g., maximum and minimum temperature during stoving). If particularly good correlation of test results between a paint manufacturer and paint user is necessary, identical equipment (panels, spray units, etc.) must be used for sample preparation.

The use of standard panels according to ISO 1514 also has advantages. Standard panels are marketed by several companies. It is advantageous to use flat, inert substrates for optical and some mechanical tests if the reproduction of identical substrate surfaces presents serious difficulties (e.g., surfaces of wood or mineral substrates). For example, glass plates are used to test gloss and color; polished steel sheets are used to test flexibility.

In the laboratory, large-scale industrial plants for paint application are often simulated with simple coating equipment. The most commonly used devices for producing films on test panels are the film applicator, film caster, film spreader, and drawdown bar (Fig. 29). When operated at a constant rate these devices produce a uniform layer thickness on the test panels. Small machines are often used to coat test panels, in which either the film applicator or the test panel is uniformly moved by a motor-driven unit.

Figure 29. Devices for producing films on test panels (Courtesy of Erichsen GmbH+Co KG, Hemer, FRG)

A) Film applicator with adjustable blade; B) Film applicator with four fixed blades of different clearance; C) Grooved-rod film applicator

Combination of the drawdown technique with standard panels is quicker, universally applicable, and cheaper than simulating large-scale industrial application conditions, and is satisfactory for many tests. Drawdowns give more reproducible results than spray application.

Sample Conditioning. The physical and mechanical properties of paints and coatings generally depend on the environmental conditions used during testing; the most important variables are temperature and humidity. The degree to which each of these variables need to be controlled is determined by the effect the variable has on the property being measured. Thus for the measurement of viscosity and density the temperature must be held constant within limits of ± 0.5 °C. Standard conditions recommended in ISO 3270 are 23 ± 2 °C and 50 ± 5 % relative humidity. Paints that dry at ambient temperature should be dried under these standard conditions for the specified period of time. Paints that are dried at high temperature should be stored for two days under these conditions before being tested. Samples subjected to such conditions have the properties of a new coating. A coating that has been in use for a year exhibits different properties. Surprisingly, however, the properties of coatings are rarely tested after artificial aging.

Measurement of Film Thickness. An accurate definition of film thickness can only be given if the coating has even upper and lower surfaces and a defined density. In practice, neither the surface of the coating nor that of the substrate is even. Surface irregularities and density variations influence the results of each test method in a specific way. Results of different tests performed on the same sample may therefore differ substantially. Results of film thickness measurements therefore always have to be quoted together with details of the measurement method and instrument used [440], [441]. A survey of methods used to measure paint thickness is given in ISO 2808.

Two instruments are generally used for determining the *wet film thickness*: the Interchemical Wet Film Thickness Gauge (ASTM D 1212, Wheel Gauge) and the Comb Gauge. With the wheel gauge (Fig. 30) contact of the wet film and eccentric rim of the wheel is influenced by the surface tension of the paint. It is therefore necessary to observe the first and last contacts and calculate the mean value. The influence of surface tension is excluded if the comb gauge (Fig. 31) is used.

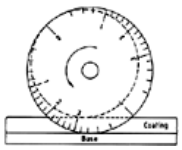

Figure 30. Wheel gauge

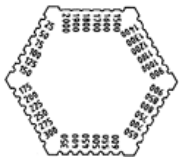

Figure 31. Comb gauge

Many methods are used for measuring the *dry film thickness*. Thickness can be determined by mechanical contact instruments such as hand-held or stationary micrometers (ASTM D 1005). The apparatus comprises a dial comparator, dial gauge, or suitable indicator for linear measurement, equipped with a mechanical contact foot and a mechanical, optical, or electronic indicator.

Two reference methods are recommended for painted surfaces: A method in which the surface profile is measured (ISO 4518) and a microscopy method (ASTM D 2691) for coatings on wood and other substrates. Other optical methods include the wedge-cut method and the profile-measuring microscope method.

Magnetic instruments based on the magnetic induction principle or on the permanent magnet pull-off principle are often used on magnetic metallic substrates (ISO 2178, ASTM D 1186). Instruments that operate on the eddy current principle are used for nonmagnetic metallic substrates (ISO 2360, ASTM E 376).

Two contact-free methods employ instruments based on the back scattering of β-particles (ISO 3543) or X-ray fluorescence (ISO 3497).

The uncertainty of film thickness measurement is usually 10 % of the thickness or ca. 1.5 µm, whichever is the greater. The measurement uncertainty of reference methods is normally < 10 % of the film thickness; microscopy methods have a minimum error of 0.8 µm.

Determination of Hiding Power. The hiding power is the ability of a paint to hide a surface over which it has been uniformly applied. The hiding power of light-colored paints can be determined to an accuracy of ca. ± 10 % by a simple method. Films are applied in different thicknesses to charts that are printed and varnished so as to produce an array of adjacent black and white areas. After drying, the surface is illuminated (1000 – 4000 lx) and visually examined. The layer thickness (in micrometers) at which contrast between the black and white areas is no longer visible is determined and is taken as a measure of the hiding power.

ISO 6504/1 describes a modification of this method in which the hiding power (V in m^2/L) is measured with a reflectometer; reproducibility is 2 %.

The contrast ratio (opacity, hiding power) for a defined area covered by a given volume of paint is determined according to ISO 3905 using black and white charts, and according to ISO 3906 using a polyester film substrate. The contrast ratios of paints of the same type and color can be compared according to ISO 2814 [442].

Drying Time. Evaluation of the drying time is simple in the case of stoving materials, but is a major problem with all paints and varnishes that dry physically or chemically at ambient temperature.

Many expressions are used in connection with drying (e.g., dust-free, tack-free, surface-dry, dry-to-touch, dry-to-handle, dry-through, dry-to-recoat) but these terms do not have accurate definitions. However, equipment is available for measuring drying times in the laboratory under defined temperature and humidity conditions (Fig. 32). Internationally standardized test methods also exist:

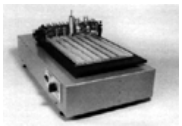

Figure 32. Electromotive drying equipment (Courtesy of Erichsen GmbH+Co KG, Hemer, FRG)

Ballotini Method (ISO 1517). Small glass beads are poured onto a horizontal film surface. After 10 s the panel is inclined and brushed lightly. The coating is “surface-dry” if all the spheres can be brushed away.

Print-Free Test (ISO 3678). A square of polyamide gauze is pressed onto the surface. The surface is “print-free” if no imprint can be seen.

Pressure Test for Stackability (ISO 4622). Two panels are coated and dried; one panel is placed on top of the other and set under pressure. No damage to the coating must occur.

9.2.2. *Optical Properties*

The appearance of a coating involves not only color but also gloss, luster, and translucency.

Gloss. The term gloss is commonly used but is difficult to define [443]. The Commission Internationale de l'Éclairage (CIE) defines the gloss of a surface as “the mode of appearance by which reflected highlights of objects are perceived as superimposed on the surface due to the directionally selective properties of that surface.”

Specular gloss denotes the degree of mirror reflection and is the primary visual gloss impression. Flat coated panels can be visually classified according to their specular gloss. The ISO 4628 system classifies specular gloss as high gloss, gloss, semigloss, semimatt, matt, and dull matt.

Arbitrary gloss scales were used initially. Little effort was made to establish perceptual gloss scales. “Psychromatic scaling of gloss” has been proposed and describes a method devised for the multidimensional assessment of gloss [444].

Although specimen panels can be classified according to the specular gloss criterion, the human eye recognizes additional gloss phenomena. These phenomena are generally described according to R. S. HUNTER who, in 1939, proposed six types of gloss: specular gloss, sheen, contrast gloss/luster, absence of bloom/reflection haze, distinctness of image/image clarity, and directional/surface uniformity.

Specular gloss and the distinctness of image are of practical importance. In many paint uses maximum gloss is required. Humans experience gloss as particularly brilliant if images are reflected with clear definition from a plane surface. Various forms of black–white contrasts are formed on the paint surface for the visual evaluation of the distinctness of image (DOI).

Specular gloss and distinctness are also most accessible to objective measurement. In 1939 HUNTER and JUDD found that specular gloss measured at 60° in a reflectometer with specified light source and receptor apertures provided a useful classification of paint finishes according to glossy appearance [445]. The instrument was standardized as ASTM D 523, improved, and subsequently standardized as ISO 2813. The 60° reflectometer is now used worldwide as the standard instrument for measuring specular gloss because visual evaluation correlates well with the reflectometer values. The directionality can also be recorded with this instrument as the difference between the maximum and minimum readings when the instrument is rotated in its own plane [449].

The 60° reflectometer has been supplemented by the 20° and 85° reflectometers. The 20° instrument provides better differentiation in the high-gloss range, while the 85° instrument provides a better differentiation in the low-gloss range. However, values measured with these two reflectometer geometries do not always correlate satisfactorily with visual gloss evaluation.

In order to measure the distinctness of image, attempts were made to obtain a second reflectometer value at a larger receiver-side aperture angle and to combine the two reflectometer values in such a way that the measured values could be assigned to the relevant visual evaluations. Sophisticated instruments are now available, in which the reflected image of a grating projected onto the surface is analyzed by means of Fourier transformations.

Goniophotometers are used for more detailed laboratory investigations of the gloss indicatrix. A plane section is laid through the gloss indicatrix and can be measured (ASTM E 167) [446-452].

Color. See ⟶ Dyes, General Survey – Color and Structure and ⟶ Dyes, General Survey – Color Measurement and Colorant Formulation for a more detailed description of color. The impression of color is produced by the absorbance of light by colorants (pigments) in the paint film. To ensure that the color of the coating is not altered by the color of the substrate, the coating must “hide” the latter. Visual color evaluation involves comparison of a sample with a standard (ISO 3668). Visual comparison of the color of paints requires an illumination of at least 2000 lx. The comparison should be made against a color-neutral (e.g., matt gray) background and the samples should be illuminated by diffuse daylight. This may be achieved with light falling onto the sample through a north-facing window (in the northern hemisphere), or in a color-matching booth with a light source giving a power distribution approximating to that of CIE Standard Illuminant D 65. If the standard and test panels contain different pigment mixtures, their colors may match under the standard light source but not under another source (metamerism).

Color comparison is most successful if the two paint films have the same or only slightly different gloss, or if the panels are viewed at an angle which minimizes gloss differences. Color differences are conveniently described according to the criteria of lightness, chroma, and hue in conjunction with the rating schemes given in ISO 4628/1. In the testing of coatings with special optical effects (e.g., metallic finishes) the flop effect also has to be included. Flop is the change in color observed when the angle of observance is varied. The tester should stand with his back to a north-facing window, hold the test panels in his outstretched arms at eye level, and tilt them slowly to give observation angles of 15°, 45 ° , and 75° [453], [454].

Colors can be defined with the CIELAB color space. All colors are arranged in a circle around a central vertical lightness axis. The center of the hue circle is considered to be neutral gray and saturation is quantified by the distance from the axis. All colors can be defined in the color space by color coordinates [455]. For further details, see ⟶ Dyes, General Survey – Color Measurement and Colorant Formulation.

The three different color-response mechanisms in the human eye are the scientific basis for the measurement of color that is standardized in the CIE Standard Observer (1931). This system specifies color according to *X, Y, Z* tristimulus values and *x, y, z* chromaticity coordinates. Several alternative scales were subsequently developed to provide a better adaptation to the color differentiation ability of the human eye [456].

ISO 7724 describes methods for the instrumental determination of the color coordinates and color differences. The standard is based on the CIE 1976 (*L*, a*, b**) color space. ISO 7724 specifications are satisfied by many color measuring devices: tristimulus colorimeters, spectrophotometers, and abridged spectrophotometers. Spectrophotometric data are now preferred and have replaced tristimulus colorimeters. Measuring conditions 8/d or d/8 (with or without gloss trap), or 45/0 or 0/45 are used [456], [457]. Metallic automotive finishes are measured with a sensor having a variable angle geometry or with a two-angle instrument [45/0 for face tone (color observed when paints are observed at an angle close to the perpendicular) and a second viewing angle to evaluate the flop tone (color observed near the grazing angle)].

Spectrophotometers can measure the spectral reflection of a sample and represent it graphically by means of computer peripherals. Spectrophotometers are a useful aid in the analysis and synthesis of color samples. All colorimeters enable calculation of the total color difference with the CIELAB color difference formula.

Color measuring instruments are superior to the human eye as regards the ability to differentiate between chroma and hue. The trained eye is, however, superior to color measuring instruments in the evaluation of differences between full shade colors and dark colors. If paint has to satisfy stringent color limits, final pass – fail decisions are made on the basis of visual inspection; instrumental color measurement serves only as an aid [458].

Color values are generally only measured at one point on each sample panel. Colorimetric evaluation can, however, be improved if several measurements are made at different points on the surface and the results are statistically evaluated [459].

9.2.3. *Mechanical Properties*

Adhesion. The cross-cut test is the most commonly used procedure for assessing the resistance of a coating to separation from a substrate. A right-angle lattice pattern is cut into the coating and penetrates through to the substrate. Cutting was previously performed with a single cutting tool or, on site, with a multiple cutting tool (DIN 53 151, ISO 2409). Sharp razor blades were also used. These tools are now being replaced by steel cutting blades which can be broken into sections. The tool may be mounted on a motor-driven apparatus to ensure more uniform cutting. A piece of adhesive tape is placed over the lattice pattern, firmly rubbed with a fingertip, and removed. A six-step classification is used to evaluate the results. This procedure should not, however, be regarded as a means of measuring adhesion [460].

The pull-off test for measuring adhesion is described in ISO 4624. Adhesion is assessed by measuring the minimum tensile stress necessary to detach or to rupture the coating in a direction perpendicular to the substrate. The result gives the minimum tensile stress required to break the weakest interface (adhesive failure) or the weakest compound (cohesive failure) of the test assembly. Mixed adhesive – cohesive failure may also occur.

Hardness. The hardness of a coating is examined by three types of test: the scratch test, the damping test, and the indentation test with a penetrating body. The results of one of these tests do not allow conclusions to be drawn as regards the behavior in another test [461].

Scratch Test. Many methods have been developed for testing scratch resistance. The method standardized in ISO 1518 employs a scratching needle with a hardened steel hemispherical tip of 1 mm diameter. The needle is moved along the surface of the coating under a specified load either in a hand-operated or in a motor-driven apparatus.

The coil coater method (ECCA-T 4, ASTM-D 3363) uses pencils of different hardnesses which have a “squared tip” and a circular cross section. The pencil is held at an angle of 45° to the coating surface. The hardness of the coating is equated with the hardness at which the pencil can press into and penetrate the paint film. Pencils that are too soft slip over the film without scratching it.

Damping Test. Pendulums according to KÖNIG [462] and PERSOZ (ISO 1522) or the Sward Rocker (ASTM D 2134) are used in the damping test. The damping time of a pendulum oscillating on the surface is taken as a measure of the hardness of the coating. With air-drying oil paints the hardness increases during drying. The pendulum oscillates for a longer time as the paint hardens and the internal damping decreases. The behavior is not found in modern stoving finishes. Surface hardness may be high but the internal damping may be large. A relationship between surface hardness and internal damping is no longer given in some synthetic resin binders.

Indentation Test. Indentation hardness is measured by two methods. A spherical or pyramidal indentor is applied to the paint surface for a specified period. The penetration depth under load is measured (ISO 6441, ASTM D 1474, Pfund hardness number). In the other method an indentor with a specified load is applied for a specified period to the paint surface and then removed. The depth of the impression is then measured.

According to ASTM D 1474 the Knoop hardness number is measured with the pyramidal Knoop diamond. According to ISO 2815 the Buchholz instrument is used. This method cannot detect elastic deformation of the coating that disappears immediately after the load is removed.

The Wallace microindentation tester employs a Vickers diamond pyramid indentor and electrical measurement of capacitance to determine the depth of

indentation under load.

In modern instruments the load on the indentor is applied in very small steps and at varying rates, with the simultaneous measurement of the depth of penetration. The results are plotted as a curve showing the depth of penetration. The measurement can be continued with stepwise removal of the load [463].

Flexibility. Viscoelastic properties are sometimes measured on free paint films [464], [465]. However, the degree to which the coating allows bending of a substrate without cracking or peeling is more important. Flexibility depends on ambient temperature and humidity, as well as on the bending rate of the substrate. Both cylindrical and conical mandrels are quite often used to evaluate the flexibility of coatings in bending tests (ISO 1519, ISO 6860). Slow deformation tests such as the Erichsen cupping test (ISO 1520) also exist.

Rapid deformations are produced in impact tests (falling weight tests). A metal sheet coated on one side is deformed by a falling weight that has a hemispherical indentor. Deformation occurs directly or indirectly, i.e., the coated or the uncoated side of the sample is struck by the falling body and deformed (ISO 6272).

The viscoelastic properties of coatings depend on structural factors, particularly on the degree of chemical cross-linking, and can be accurately measured by monitoring the bending vibrations of thin, coated steel bars. The bars are subjected to forced oscillations whose resonance curves are evaluated with a computerized measuring system. The device can provide data on the flexibility of coatings after different curing processes. The method can be used to determine the optimum conditions for curing with heat [466], [467].

Abrasion Resistance. Abrasion (wear) resistance is a basic factor in the durability of a coating, and is determined by the interaction of the abrasive medium and the coating. Many types of abrasive media and treatment are used for testing e.g., fast and slow movement, small or large load, high and low temperature, constant or intermittent contact. The falling sand (abrasive) method, abrasive blast method, rotating disk method, rotating wheel method (Taber abraser, ISO 7784), and Gardner wet abrasion method (used with emulsion paints in DIN 53 778 T 2) are commonly used.

The stone impact resistance is an important property of automotive finishes. Several instruments (gravelometers) have been developed for testing this parameter. One or more projectiles are hurled or shot in the direction of the painted surface at a defined velocity [Verband der Automobilindustrie (VDA) Guideline 621–427] [468], [469].

9.2.4. *Chemical Properties*

Since coatings are generally used to protect substrates against corrosion, two factors must always be considered:

1. Are the coatings themselves resistant to aggressive media?
2. Can the coatings protect the substrate?

It should always be borne in mind that chemicals may have different effects depending on whether they are in liquid or vapor form.

Water Vapor and Water Resistance. The rate of water vapor transmission of free paint films is measured according to ISO 7783. The resistance of coated substrates to water condensation is, however, generally measured (e.g., according to DIN 50 017). Water vapor condenses on coated substrates in airtight cabinets (capacity ca. 300 L) or in walk-in chambers. The coated substrates are subjected to 100 % relative humidity at 40 °C for 8 h, followed by a 16-h period at 23 °C and 50 % relative humidity.

The resistance to humid atmospheres containing sulfur dioxide is measured according to ISO 3231 in a 300-L cabinet, to which 0.2 or 1.0 L of sulfur dioxide is added before the cabinet is heated to 40 °C.

Resistance to humidity (continuous condensation) is measured according to ISO 6270 in the Cleveland condensation cabinet, in which the specimens form the roof of the cabinet. The cabinet contains a water bath set at 40 °C. The water vapor condenses on the coatings that are backed by porous or nonporous substrates which are maintained at 23 °C and 50 % relative humidity. The water vapor diffuses through the porous substrate [470].

Corrosion Resistance. The determination of the resistance to neutral salt spray fog according to ISO 7253 has been adopted from ASTM B 117. This widely used method is regarded as the classic corrosion test. The specimen panels are artificially damaged with a scratch that exposes the substrate. The salt mist acts continuously on the specimen panels during the test. In filiform corrosion on steel and aluminum, thread-like infiltrations of the coatings start from edges or damaged sites. This phenomenon can be tested according to ISO 4623 on steel sheets by damaging a specimen with a scratch and immersing the specimen in salt solution or exposing it to a neutral salt spray fog in a test cabinet maintained at 40 ± 2 °C and 80 ± 5 % relative humidity.

Comparisons of the results of laboratory corrosion tests with damage sustained in use have revealed differences in corrosion and corrosion products. For example, the constant salt spray fog test leads to a different type of rust formation than that produced by natural weathering, with its alternating damp and dry periods. Test cycles have therefore been developed in which the individual stresses are applied in sequence. A good correlation with natural corrosion in automobiles was demonstrated in the VDA alternating test. The sample is subjected to salt spray fog for 24 h, followed by four 24-h cycles each consisting of 8 h at 40 °C and 100 % relative humidity, followed by 16 h at 23 °C and 50 % relative humidity, and two days of ambient conditions (VDA Guideline 621–415).

Since atmospheric corrosion is an electrochemical process, attempts have been made to develop electrochemical test methods that can predict the corrosion resistance of the coating – substrate system [471], [472].

The specimens are damaged with a scratch that exposes the metallic substrate, and are cathodically polarized. After several hours a loss of adhesion at the scratch can be observed, which can be correlated with natural corrosion.

Resistance to Liquids. Various methods are used to test the resistance of coatings to the action of liquids and pastes (e.g., mustard, soap solutions, ketchup). Standard methods are defined in ISO 2812. Coated specimen panels are partially immersed in the test liquid so that the changes in gloss, color, and swelling can also be evaluated at the liquid – air interface. Disks of absorbent material are immersed in the test liquid and placed on the coating. The test liquid or test paste is also dropped directly onto the coating. If the test medium is to be prevented from evaporating, the test surface is covered with a glass dish. Elevated temperature generally increases the aggressiveness of the test medium.

ISO 4628 describes a general system for evaluating the decomposition of coatings. The intensity, quantity, and size of common defects are classified on a numerical scale ranging from 0 to 5.

9.2.5. *Weathering Tests*

Coatings that are exposed to weathering undergo aging. Aging is defined as the sum of all irreversible chemical and physical processes that occur in the coating over the course of time. Aging is mainly caused by radiation, temperature, and moisture (rain and atmospheric humidity).

Solar radiation (λ = 290 – 3000 nm) is the primary cause of aging. Solar radiation causes heating of the coating that depends on the color of the surface. UV radiation initiates photochemical aging. Weathering produces changes in gloss and color, cracking, blistering, loss of adhesion, and loss of flexibility; these parameters are used to evaluate aging [473-478]. ISO 2810 gives guidelines on how to perform natural weathering tests. The manner and location in which the coating is exposed must be appropriate for the end use of the product under test. Flat specimen panels or structures are fastened to exposure racks at a defined angle. The exposure period should be one or more years. The intensity, quantity, and magnitude of defects (e.g., blistering, chalking, cracking) caused by natural weathering can be classified according to ISO 4628 (evaluation of degradation of paint coatings). Measurement of degradation of gloss and color is described in Section Pretreatment of Plastics .

Commercial test institutes in Florida are often commissioned to carry out natural weathering tests. The weather conditions in Florida are preferred because the same aging phenomena occur as in other locations but more quickly. Exposure time can be reduced by a factor of two to four when compared with other locations. The test institutes in Florida also offer accelerated aging by means of exposure on black boxes and on heated black boxes.

In laboratories, coatings are artificially weathered in specially designed apparatus to simulate or measure the aging processes that occur during natural weathering. Artificial weathering involves a smaller number of parameters than natural weathering but can be controlled more uniformly and allows accelerated test conditions [479-481].

Generally valid correlations between aging processes during artificial and natural weathering cannot be expected because they are influenced by many factors. Clearly defined relationships can only be expected if the most important parameters are the same or their influence on the coatings is known [482], [483]. ISO 4892 and 11 341 specify filtered xenon arc radiation and other conditions used for the artificial weathering of coatings. The optical radiation source and its filter system is specified so as to produce a spectral distribution of the irradiance sufficiently similar to the global solar radiation defined in CIE Publication No. 85. The irradiance is measured inside the apparatus with a radiation meter. The temperature is measured with a black standard thermometer. The test panels are wetted by spraying or flooding with water. ISO 2809 and 787/15 also provide information on how to simulate aging processes occurring during natural weathering under a glass cover.

ASTM D 822 describes a standard procedure for operating light- and water-exposure apparatus (carbon-arc type) for testing coatings. ASTM G 53 and DIN 53 384 describe similar standard procedures for the light exposure or light and water exposure (fluorescent UV type) of nonmetallic materials [484].

[Top of Page]

10. Analysis

Ulrich Schernau and Bernhard Hüser

Modern coating materials are complex mixtures of binders, solvents, pigments, extenders, and additives. Most of these components consist of several constituents. Complete analysis of coating materials therefore requires comprehensive knowledge and the use of various analytical methods.

Surveys of the analysis of coating materials are given in [485-487]. The proceedings of conferences devoted specifically to paints (e.g., FATIPEC, International Conference on Organic Coatings, Science and Technology in Athens, Waterborne and Higher-Solids Coatings Symposium, the conference of the ACS Polymeric Materials Division, and ASTM meetings) are particularly important because they describe the use of analytical methods from the point of view of their suitability for investigating coating materials. Periodic literature reviews on the analysis of coating materials are also published in the journal *Analytical Chemistry*.

The analysis of coating materials is often employed in the investigation of complaints and substandard batches, or to evaluate competing products. Analysis also plays an important role in the assessment of raw materials, occupational safety and hygiene, and the emission of solvents and decomposition products during paint curing.

In the case of complaints, the binder and pigment (extender) composition of the individual layers have to be established to determine the origin of the paint material in doubtful cases. Establishing the cause of coating defects (e.g., inclusions, delamination, peeling) is also important. These and other analytical investigations are time-consuming and expensive. Optimal utilization of analytical resources therefore requires a clear definition of the problem in hand.

10.1. Analysis of Coating Materials

10.1.1. *Separation of the Coating Material into Individual Components*

For a complete analysis it is advantageous to separate the coating material into its components. The binder and pigment (extender) fractions can only be investigated in detail after separation and isolation. The solvent composition can usually be determined by gas chromatographic analysis of the complete coating material. In conventional, solventborne systems, the pigment (extender) fraction is centrifuged off after diluting the coating material with a suitable, volatile solvent (ethyl acetate, tetrahydrofuran, methyl ethyl ketone) [486]. Centrifuge speeds of 5000 min^{-1} are generally sufficient, however if finely divided pigments (extenders) or soot are present, a low-density, low-viscosity solvent (e.g., acetone) should be used and the rotational speed should be increased to 20 000 min^{-1} [488]. To achieve quantitative separation the pigment sediment is repeatedly shaken with solvent and recentrifuged. Prior to analysis the combined binder containing supernatant fractions are dried in a drying cabinet. The centrifuged pigment is also dried in a drying cabinet before being analyzed.

Problems are often encountered if centrifugation is applied to waterborne systems, particularly if the binder is insoluble or only sparingly soluble in organic solvents. In these cases the coating material should be carefully dried (vacuum drying cabinet, freeze drying). A representative binder fraction is then obtained by exhaustive extraction with a suitable solvent (e.g., 1,2-dichlorobenzene, dimethylformamide, and/or tetrahydrofuran) [489], [490].

10.1.2. *Analysis of Binders*

Modern coatings are expected to provide permanent protective action, outstanding mechanical – technological properties, and an attractive surface appearance. This requires binders consisting of special resin combinations (Chap. Types of Paints and Coatings (Binders)) and selected additives (Chap. Paint Additives).

Preliminary tests, color reactions, and spot tests [491] were formerly used to identify individual resins but are no longer important because they are not sufficiently specific and they do not provide quantitative results. They have been largely replaced by modern spectroscopic and chromatographic methods, which often require preliminary chemical workup of the sample.

Infrared and nuclear magnetic resonance spectroscopy are the most important spectroscopic methods for analyzing coating materials. Near infrared Fourier transform (NIRFT) Raman spectroscopy [492] also has great potential, particularly for aqueous systems. UV/VIS spectroscopy is used only in exceptional cases, e.g., to determine light protection agents (UV absorbers).

Infrared spectroscopy has the advantage of simple sample preparation and measurement; practically all types of samples (both as regards the state of aggregation and solubility) can be investigated with the aid of special measuring techniques. Infrared spectroscopy is frequently employed to obtain an overview of the binders and binder classes. A common procedure is to pour a few drops of the binder-containing supernatant from the separation (centrifugation) onto a NaCl or KBr crystal. A thin binder film is obtained after drying in the drying cabinet and IR absorption is then measured. The IR spectra provide qualitative and semiquantitative information about the binder composition. Comprehensive spectra catalogs of organic polymers [493] and commercially available coating materials [494] are available for comparison purposes. Modern IR spectrometers, particularly the Fourier transform infrared (FTIR) spectrometers, offer further possibilities. These instruments have short measurement times and high wavelength reproducibility; they can therefore accumulate data and provide a considerably better signal to noise ratio than conventional grating or prism instruments. IR spectroscopy can then be used to analyze extremely small sample amounts. Examples are diffuse reflectance FTIR spectroscopy (DRIFTS), investigations of sample surfaces by measurement with attenuated total reflection (ATR), or the analysis of local defects using an IR microscope [495]. A further advantage of FT spectrometers is the fact that the data are available in digital form. Spectroscopic data banks can therefore be compiled either on the spectrometer computer or on an external computer. Commercial data banks (e.g., Sadtler) and in-house, laboratory-specific data collections can be installed, the latter are often better suited to the particular interests of the laboratory. Data banks facilitate archiving, and also allow quick comparisons of measured and library spectra. Quantitative evaluations are also facilitated after appropriate calibration. The theoretical aspects of FTIR spectroscopy and analysis are described in [496]. The use of FT and conventional IR spectroscopy to investigate coatings and coating materials is described in [497-499].

Nuclear Magnetic Resonance Spectroscopy. Like IR spectroscopy, NMR spectroscopy requires little sample preparation, and provides extremely detailed information on the composition of many resins. The only limitation is that the sample must be soluble in a deuterated solvent (e.g., deuterated chloroform, tetrahydrofuran, dimethylformamide). Commercial pulse Fourier transform NMR spectrometers with superconducting magnets (field strength 4 – 14 Tesla) allow routine measurement of high-resolution ^{1}H- and ^{13}C-NMR spectra. Two-dimensional NMR techniques and other multipulse techniques (e.g., distortionless enhancement of polarization transfer, DEPT) can also be used [500]. These methods are employed to analyze complicated structures. ^{13}C-NMR spectroscopy is particularly suitable for the qualitative analysis of individual resins in binders, quantiative evaluations are more readily obtained by ^{1}H-NMR spectroscopy. Comprehensive information on NMR measurements and the assignment of the resonance lines are given in the literature, e.g., for branched polyesters [501], alkyd resins [502], polyacrylates [503], polyurethane elastomers [504], fatty acids [505], cycloaliphatic diisocyanates [506], and epoxy resins [507].

Chromatography. Liquid chromatography is the most important chromatographic method for the investigation of binders and resins. Special applications are also opening up for supercritical fluid chromatography (SFC) [508]. Methods of particular importance are size exclusion chromatrography (SEC) (also termed gel permeation chromatography, GPC) and reverse-phase high performance liquid chromatography (HPLC).

Gel permeation chromatography separates molecules according to their size. This technique is used to determine the molecular mass distribution of resins. A review of modern GPC methods for analyzing coating materials is given in [509]. Combination of GPC with modern spectroscopic techniques (in particular FTIR spectroscopy) is a valuable aid for identifying individual resins in binder systems; in some cases additives can also be identified, because their molecular mass is often substantially lower than that of the resin constituents. The coupling of liquid chromatography and FTIR spectroscopy is described in [510]. Off-line investigations of fractions from an analytical GPC run of a binder by FTIR measurements under diffuse reflection provide detailed information on the

composition of the binder [511].

Reverse-phase HPLC (mainly octyl- or octadecyl-modified silica gels) separates molecules according to their partition coefficients. It is particularly suitable for characterizing relatively low molecular mass resins (e.g., epoxy, phenolic, and melamine resins).

The high resolution of modern columns and the high sensitivity of UV detection allow individual oligomers (including positional isomers and secondary compounds) to be separated and identified. The peak distribution pattern often permits identification of commercial products. Detailed HPLC investigations have been carried out on epoxy resins [512]. HPLC can also be applied to resins without UV-active groups by using mass detectors [513].

On-line coupling of pyrolysis, gas chromatography, and mass spectrometry is a quick and elegant method for the qualitative detection of monomer units in many resins (e.g., polyesters, polyurethanes, phenolic resins, and polyacrylates). Identification of comonomers of polyacrylates, including hydroxy-functional and carboxy-functional monomers, is facilitated if the sample is silylated before pyrolysis [514].

Chemical Workup. Chemical decomposition of resins followed by qualitative and quantitative analysis is still an important technique because, apart from NMR spectroscopy, none of the instrumental analytical methods provides reliable quantitative values. Chemical workup is also essential for low concentrations of resin building blocks that are often unknown; it simplifies and enables the desired substances to be concentrated. Qualitative and quantitative determination is carried out by instrumental methods.

Well-established methods of chemical workup are available for alkyd resins based on *o*-phthalic acid. Alkaline hydrolysis can be used for the quantitative determination of the dicarboxylic acid, fatty acid, and polyol fractions [515]. In the IUPAC method the carboxylic acids are determined by gas chromatography after transesterification with lithium methoxide; the polyols are determined by gas chromatography after aminolysis and acetylation [516]. Chemical workup methods (e.g., hydrolysis with alcoholic alkali, alkali fusion, aminolysis with hydrazine, and transesterification with sodium methoxide) for various resins are described in [517]. Alkaline hydrolysis has the disadvantage that it results in low fractions for polyunsaturated fatty acids and lower polyols. Transesterification with methanol or trimethylsulfonium hydroxide provides substantially better results for unsaturated fatty acids. This transesterification process was originally proposed for determining fatty acids in triglycerides [518] but can also be applied to alkyd resins and polyesters. After evaporation of methanol and silylation, the polyols from the reaction mixtures can then be determined by gas chromatography. This procedure yields more accurate values than with the previously mentioned methods, particularly for the lower polyols.

10.1.3. *Analysis of Pigments and Extenders*

The separated pigment (extender) is usually used as starting material for analysis (see Section Separation of the Coating Material into Individual Components). If quantitative separation is not possible (e.g., in emulsion paints) the inorganic pigment (extender) fraction can be obtained by ashing the nonvolatile fraction; for further details see [489].

The isolated pigment (extender) fraction is analyzed by various chemical and instrumental methods. Methods of elemental analysis are used for inorganic pigments (extenders). These include traditional chemical methods involving separation and gravimetric, tritrimetric, or polarographic determination of the elements. These methods are being replaced by instrumental methods such as atomic absorption spectroscopy (AAS), atomic emission spectroscopy (AES), and X-ray fluorescence analysis (XFA). A further valuable tool is IR spectroscopy, which provides characteristic spectra for many inorganic extenders and pigments (e.g., chalk, dolomite, kaolin, talc, and barium sulfate). The most elegant and informative method, but the most expensive as regards equipment, is X-ray diffraction [519], [520]. Sample preparation is simple and the method can be used on hardened coating materials; problems associated with isolating the extender – pigment fraction can therefore be avoided. The principal advantage, however, is that the substances (e.g., $BaSO_4$) can be identified directly and not indirectly via their elements (e.g., Ba and S). This is particularly advantageous if both silica and silicates (kaolin, talc) are present. Minerals such as dolomite, chalk, talc, and kaolin originating from different geographical locations have specific elemental compositions. Their spectra can therefore be used to identify the source of supply when investigating coating materials [521].

For many inorganic and organic substances, X-ray diffraction spectra recorded on powdered materials are commercially available on a data carrier (CD-ROM) [522].

The pigmentation of monochrome coating materials usually comprises several constituents: organic and/or inorganic pigments, titanium dioxide to improve the hiding power, and other inorganic extenders. The inorganic fraction is generally dissolved with acid and can then be analyzed by conventional chemical methods. The organic pigments remain undissolved and are investigated by IR spectroscopy, photometry, X-ray diffraction, and separation methods such as thin layer chromatography (TLC) or HPLC. Modern FTIR spectroscopic methods including library searches; further developments in TLC and HPLC are discussed in [523]. A separation process for organic pigments has been developed which exploits solubility differences in various solvents (ranging from *n*-hexane to sulfuric acid) [524]. The extracted pigments are identified by their VIS spectra in the range 400 – 900 nm. The advantages and limitations of X-ray diffractometry in the analysis of organic pigments are discussed in [525].

The increasing use of special-effect paints, particularly in automobile finishes, places new requirements on analytical techniques. The previously described methods are largely unsuitable for identifying substances used to produce special effects. Aluminum pigments in metallic paints can be roughly classified by light microscopy. Since these pigments are often subjected to special pretreatment, the determination of foreign elements with X-ray microprobes or the determination of an organic agent used for surface treatment may be necessary for more accurate characterization. Light microscopy can also be used for initial characterization of nacreous pigments, more detailed information can be obtained with a microscope spectral photometer (e.g., UMSP 80, Zeiss) or transmission electron microscopy [526].

10.1.4. *Analysis of Solvents*

The solvent composition of coating materials can be determined in two ways. In the first method, the coating material is subjected directly to gas chromatography, if necessary after dilution with a suitable solvent. In the second method, the solvent fraction is separated from the coating material by vacuum distillation and then analyzed by gas chromatography or, in special cases, by IR spectroscopy. ASTM methods (D 3271 and D 3272) exist for both processes. The advantages and disadvantages of the two methods in the analysis of solventborne and waterborne systems are discussed in [486].

Direct gas chromatographic analysis with modern capillary column technology and reproducibly operating injection systems (e.g., modern autosamplers) gives highly reproducible, accurate results. This method is therefore used most widely. With complex solvent mixtures, unambiguous identification of the peaks is facilitated by simultaneously analyzing the sample on two columns with different polarities or by using GC – MS. For further information on gas chromatography, see [527].

10.1.5. *Analysis of Additives*

Qualitative and, in particular, quantitative analysis of coating additives is difficult due to their low concentrations and chemical diversity. A general outline for their investigation cannot be given, isolation and analysis depend on chemical structure. The liquid or hardened coating material is often extracted with a suitable solvent, followed by spectroscopic or chromatographic analysis. For example, plasticizers can be extracted very efficiently with pentane and detected by IR spectroscopy or gas chromatography. Supercritical fluid chromatography (SFC) is being increasingly used for the analysis of polymer additives [528]. Especially in combination with supercritical fluid extraction (SFE) as a relatively fast and efficient sample preparation method [529], it may prove to be of interest for the analysis of many additives used in coating materials.

10.2. Analysis of Coatings

Coatings are practically free of solvents and the binder is generally cross-linked (i.e., insoluble). These factors require special sample preparation and analytical methods, for a detailed discussion see [530]. The separation of binder and pigment (extender) fractions for further investigation is only possible with non-cross-linked (physically drying) binders. The coating film is soaked in a suitable solvent and the pigment is centrifuged off after dissolving the binder. Provided sufficient material is available, the isolated components can be analyzed by the methods described in Section Analysis of Coating Materials. The above method cannot be used with cross-linked binders (e.g., two-pack systems, stoving finishes). The binder and pigment (extender) fractions in coating layers are most simply analyzed by means of IR spectroscopy. Different measurement techniques are available for this purpose, which require various degrees of preparative effort. The most important techniques are measurements on KBr pellets prepared from scratched off paint material, measurement of the coated surface with the method of attenuated total reflection (ATR), and measurements on cross sections of the coating with FTIR microspectroscopy [531].

Further methods for determining the binder structure of cross-linked systems include the use of pyrolysis gas chromatography or alkaline hydrolysis followed by analysis of the degradation products by gas chromatography. In multilayer coatings this may prove difficult because the materials have to be prepared from individual layers.

The inorganic pigment (extender) fraction can be obtained by ashing the isolated paint material and analyzing it by the methods described in Section Analysis of Pigments and Extenders.

The question of the cause of coating defects often arises in the investigation of coatings. Defects may be found at localized sites (specks, craters) and impair the appearance of the coating. They may also occur as planar defects that for example reduce adhesion to the substrate or between the individual layers. Defects are investigated with light microscopy, FTIR microspectroscopy, X-ray methods [532], and modern surface analysis methods such as time-of-flight secondary ion mass spectrometry (TOF-SIMS) [533], laser-microprobe mass analysis (LAMMA) [534], and X-ray photoelectron spectroscopy (XPS) [535].

[Top of Page]

11. Uses

Alfred Brandt, Alex Milne, Helmut Weyers

The application of paints to various substrates (e.g., metals, wood, plastics, and concrete) is the most widely used method of protecting materials against corrosion and degradation. It is also used to obtain properties that include gloss, color, completely smooth or textured surfaces, abrasion resistance, mar resistance, chemical resistance, and weather resistance. Normally, a combination of properties is required. Paint systems are therefore applied that generally consist of a primer, an intermediate coat, and a topcoat. These coats of paint together with the substrate surface and surface layers resulting from substrate preparation and pretreatment form the coating system. Only this complete coating system can provide the combination of properties required for the wide range of uses of organic coatings.

Most paints are supplied as liquids that are applied by different methods, using various types of equipment (see Chap. Paint Application). The properties and uses of powder coatings are described in Section Coating Powders. Once applied, the wet paint film must dry to a hard solid film. Paints that dry at ambient temperature (air-drying paints) may be force-dried at temperatures up to 100 °C. Other types of paints require higher temperatures (120 – 220 °C) for film formation that involves reaction of two and more binder components. Thus, the paint formulator has to consider both the properties of the liquid film and those of the final dry film. Liquid film properties have to be considered during storage, application, and curing.

To obtain a properly formulated paint, testing has to be carried out in different stages (Chaps. Properties and Testing and Analysis): weathering and corrosion tests, application tests, field trials that test in-use behavior, and durability. The paints can only be used commercially when they have passed these tests.

11.1. Coating Systems for Corrosion Protection of Large Steel Constructions (Heavy-Duty Coatings)

[536], [537]

Large steel constructions have vast metal surfaces which must be protected against corrosion to maintain their proper function. Such constructions include road and railroad bridges, electric pylon lines, radio and radar antennae, gas tanks, storage tanks (e.g., for oils, chemicals, cement, and grain), loading equipment (e.g., cranes, conveyors), mining and drilling constructions, as well as steelworks and chemical plants. Although these constructions are sometimes protected by inorganic coatings, about 90 % are coated with paints based on organic binders.

Appropriate surface preparation is of utmost importance for a long service life. Best results are achieved by blasting, using grade 2 1/2 or 3 according to the Swedish Standard SIS 055 900 (equivalent to DIN 55 928, part 4, FRG; BS 7079, part A 1, 1989, UK; SSPC-SP 5-SP 6, ASTM, USA; ISO 8501 – 8503). Although mill scale and old paint are sometimes removed by flame descaling, this method is less effective. Residues of rust and corrosion-promoting chemicals (e.g., salts) on the prepared surface lead to early rusting under the new paint and must therefore be removed completely.

In galvanized steel constructions, blasting is also necessary to provide a coatable surface. The sweep-blasting method is used in which the zinc surface is roughened without removing a significant amount of the zinc layer. When constructions made of aluminum alloys are coated, their surfaces must be blasted with iron-free blasting materials.

Heavy-duty coating systems generally comprise two primer coats and two topcoats, modern system sometimes consist of one primer coat and two topcoats. The total dry film thickness (DFT) of such anticorrosive paint systems is 150 – 200 µm, each layer has a minimum DFT of 40 – 50 µm. Due to their excellent adhesion, the first and second primers prevent corrosion of the metal surface. The pigments and extenders allow the primers to react with ions (Cl^- and SO_4^{2-}) that diffuse into the film from the atmosphere. The pigmented organic film also forms a barrier against humidity that may otherwise initiate a corrosive process.

Heavy-metal pigments (mainly lead pigments) and zinc chromates were used succesfully in earlier decades. These pigments are now being replaced by nontoxic pigments (see Section Railroad Rolling Stock, and ⟶ Pigments, Inorganic).

The first and second topcoats build up the necessary dry film thickness and protect the entire coated construction against the adverse influence of the atmosphere.

Binders based on linseed oil and other oils have been used for many years in anticorrosive primers. Alkyd binders, especially those with high fatty acid contents, perform similarly. The main disadvantages of these binders is their limited chemical resistance and their slow drying.

Chlorinated rubber and poly(vinyl chloride) (PVC) resins allow the formulation of coatings with good chemical resistance. They are therefore used for steel constructions in chemical plants. Since they are not resistant to many organic solvents, they should not be used in oil refineries or plants handling solvents. The undesirable fact that these binders contain halogens in high amounts is responsible for their decreasing use. Overspray of chlorinated rubber and PVC paints and contaminated blasting materials produced after removing old paint cause severe problems in waste incineration plants (generation of hydrochloric acid), as well as in waste disposal areas (pollution of soil and water).

Epoxy resins cured with aminoamide resins or amine adducts are often used for large metal constructions. Paints based on these resins are normally applied in four layers. Epoxy coatings form films that are resistant to organic solvents and a wide range of chemicals. Epoxy coatings are used for the majority of steel and aluminum constructions, but are also suitable for use on other construction materials (e.g., concrete). They can protect buildings in chemical plants and nuclear power plants. Epoxy coatings are less susceptible to deterioration by radiation than other organic films, and are also resistant to decontaminating chemicals (usually aqueous detergent solutions) used to remove radioactive dust from walls and other surfaces in nuclear power plants.

Heat-resistant coatings have silicone-resin binders. Pigments for such paints are zinc dust, flakes of aluminum or stainless steel, titanium dioxide, or silicon carbide. Such paints can withstand temperatures up to 600 °C.

Paints with inorganic binders are also used for corrosion protection of steel constructions. These paints are based on organic silicates which are soluble in mixtures of alcohols or other water-miscible solvents (see Section Alkyl Silicates). Ethyl silicate is often used and mostly pigmented with zinc dust. Zinc-rich primers and single coats are available as one- or two-pack products. Zinc-rich ethyl silicate paints dry to form inorganic films that are very durable even under adverse atmospheric conditions, (e.g., onshore and at sea). These coatings have excellent resistance to oil, solvents, and mechanical impact, and are therefore used on drilling stations, oil rigs, and ships. Since zinc-rich silicate coatings are heat resistant, they are also used in hot areas of iron works, coal mines, and coking plants.

Heavy-duty coatings are often still applied manually with brushes or rollers that completely wet the metal surface; holes and pores are filled with paint. This is especially important when old, partially rusted constructions are repainted after sanding. Brushing and rolling, however, only allow a slow working speed. Larger surface areas must be painted with airless spraying equipment.

11.2. Automotive Paints

11.2.1. *Car Body Paints*

Cars are coated to achieve maximum, long-lasting corrosion resistance. Cars must also be given an optimum appearance that lasts for many years. Long-lasting color and gloss retention as well as resistance against cracking (especially in clearcoats of two-coat metallics) are therefore necessary. Topcoats of automobiles must withstand solar radiation and atmospheric pollution (e.g., acid rain and soot from oil combustion). Aggressive chemicals (e.g., road salts and cleaning agents containing detergents) can damage the coating if they come into contact with the car surface. Furthermore, small stones cause heavy impact on automobile surfaces and corrosion via chipping.

Large numbers of cars are manufactured on fast-running assembly lines. The paints must therefore be applied with highly efficient equipment, and must dry very quickly. The paint products are classified as primers, intermediate coats (also called fillers or surfacers), and topcoats (or finish). The primers and fillers are designated as the undercoating system.

Car paints are cured with heat in special oven lines. Electrodeposition coatings (used as anticorrosive primers) contain only small amounts of volatile organic compounds (VOC), whereas intermediate and topcoats release considerable amounts of VOCs. Intermediate coats based on waterborne resins have been developed to decrease VOC emission and are already being used in some automotive plants. Basecoats, as part of base – clear topcoat systems, contain very high amounts of volatile organic solvents. Waterborne basecoats were developed more recently to lower this source of solvent emission. Some car manufacturers are operating pilot lines with the aim of introducing waterborne basecoats into their production processes. Many car producers in the United States and Europe have already switched their topcoat lines over to waterborne basecoats [538].

Pretreatment. Various metals are used for manufacturing car body shells: steel, galvanized steel, aluminum alloys, and zinc-rich precoated steel. The surfaces of these metals are routinely contaminated with oils, drawing lubricants, dirt, and assembly residues (e.g., welding fumes). The body shells are pretreated to remove these contaminants and to obtain a well-defined, homogeneous surface that has the necessary properties for adhesion of primers. Pretreatment includes surface cleaning and formation of a phosphate conversion coat on the shell surface (see Section Pretreatment of Metallic Substrates); six to nine discrete steps are involved using either spraying devices or baths. Continuous control of phosphating solutions ensures good results [536], [539].

Anticorrosive Primers. Anticorrosive primers are applied in dip tanks so that they reach all parts of the car body; dipping is a fast method of application. The standard method for application of primers is electrodeposition. Anodic electrodeposition paints were used when the electrocoating technique was first applied, but cathodic electrodeposition is now predominant because it provides better corrosion protection.

The binders for cathodic electrodeposition are epoxy resin combinations dispersed in water (see Section Electrodeposition Paints). Advantages of anticorrosive electrocoatings include excellent corrosion resistance at a dry film thickness of ca. 20 – 30 µm. Electrocoats are stoved at 165 – 185 °C to obtain films with the desired properties. The paint industry is now developing electrocoats that can be cured at lower temperatures (140 – 150 °C). Electrocoating produces a homogeneous film that covers the entire car body surface, including recesses and cavities.

Although the dry film thickness on the metal edges is somewhat lower, these areas are still efficiently protected against corrosion. The ultrafiltration technique results in a very high transfer effect and a uniform coating: paint solids from the bath are deposited on the metal surface without loss. Since electrodeposition paints have a low organic solvent content, air pollution is low. The dip tank contents are not flammable, which reduces insurance costs [540].

Intermediate Coats. Intermediate coats (fillers) are applied between the anticorrosive primers and the topcoat systems. They provide good filling and flowing layers which are normally smoothed by sanding. Oil-free polyesters are used as binders for fillers. They react with blocked isocyanates in 20 min at 165 °C. Their high flexibility gives the whole coating system a highly effective mechanical (stone chip) resistance.

Fillers are applied with electrostatic spraying devices (fast-rotating bells) to give dry film thicknesses of about 40 µm. Waterborne fillers with polyester – melamine binders (primer surfacers) have been developed to reduce the volatile organic content. They yield a film thickness of 30 µm after a prereaction time of 10 min at 100 °C and a reaction time of 20 min at 165 °C. The properties of the films are similar to those formed by solventborne paints. More recently, waterborne fillers based on blocked isocyanates have been developed. Field trials have shown that their mechanical resistance is very good.

Topcoat Systems. Topcoats form an important part of the protection system of the car body surface, but are much more important for decoration. The basic requirements for a car topcoat are:

1. Full, deep gloss (wet-look)
2. Highly brilliant metallic effects
3. Long-lasting resistance against weather and chemical influences
4. Easy to polish and repair

Topcoats based on nitrocellulose combinations with plasticizers and alkyd resins were used in the first decades of industrial car manufacturing. These were followed by thermosetting alkyd – melamine combinations, and later by thermosetting acrylics. The use of stoving enamels as thermosetting paints also accelerated production significantly. Although the properties of these coatings during application and in use were very good, their high content of volatile organic solvents had to be lowered to comply with legal restrictions.

The basecoat – clearcoat system is presently the most commonly used type of topcoat for cars because it is the standard application system for metallic colors. Today, about 70 % of all cars have metallic topcoats. The basecoat – clearcoat system consists of a colored layer (basecoat) which is overcoated after a short flash-off time with a protective layer of clearcoat. Both coats are cured together at 120 – 140 °C. The basecoat contains pigments which provide two types of finish: solid (straight) colors or metallic.

Solventborne metallic basecoats contain ca. 15 % solids and ca. 85 % volatile organic solvents. These solvents are not released into the atmosphere, but are converted to combustion gases in afterburners. To reduce emission of organic solvents from this source, waterborne basecoats have been developed.

Waterborne basecoats with higher solids contents are now available: metallic basecoats contain about 18 wt % solids and solid (straight) color basecoats 25 – 40 wt %. The solvent in waterborne paints is not pure water; about 15 % of organic solvents is still needed as a cosolvent for proper film formation. Metallic basecoats are applied at a DFT of 15 µm, solid color basecoats at a DFT of 20 – 25 µm.

Basecoats are sprayed in two layers. The first layer is sprayed electrostatically with high-speed rotation bells, the second layer is sprayed with compressed air to achieve proper orientation of the aluminum particles in metallic paints. The basecoat is then dried for 3 – 5 min in a warm air zone at 40 – 60 °C.

A final layer of clearcoat is applied with electrostatic high-speed rotation bells [538], [542] to protect the system against atmospheric influences, including wear and tear during use.

Alkyd – melamine clearcoats with an approximate solids content of 50 % contain UVabsorbing agents to prevent deterioration in extreme climates.

Some car manufacturers use clearcoats with acrylic binders that are cured with aliphatic isocyanates. Their chemical and mechanical properties are better than those of alkyd – melamine clearcoats. Solid contents are as high as 58 %.

Car Repair Paints [536]. Repair paints are used in considerable amounts for refinishing cars. Since repair shops cannot provide the same facilities as those of car manufacturers, repair paints are dried at ambient temperature or elevated temperature up to 80 °C (metal temperature). Alkyd repair paints and nitrocellulose paints were standard materials, but two-pack acrylate – isocyanate refinish paints are now more common. Their properties are similar to those of the original car coatings (long-lasting gloss and color, mechanical and fuel resistance). Car refinish paints are available in a wide range of colors, solids as well as metallics. They are often supplied to shops and retailers as mixing schemes.

Paint systems for car repair comprise anticorrosive primers, putties, intermediate coats, and topcoats; repair coatings applied to refinished cars have similar durabilities to those of the originally manufactured coating systems.

11.2.2. *Other Automotive Coatings*

The properties of coating systems used for car components differ considerably from those of systems used for exterior car surfaces. Color is not important (and is mainly black or gray), but anticorrosive properties similar to those of car body coatings are required. Since car components are produced in large numbers, coatings are commonly baked at high temperature to ensure a high reaction rate and rapid film formation.

Wheels are electrocoated; engine blocks are coated with heat-resistant, usually waterborne materials. Other parts (e.g., steering equipment and shock absorbers) are painted with two-pack, one-coat epoxy systems that are usually solventborne; use of waterborne systems is, however, increasing.

11.3. Paints Used for Commercial Transport Vehicles

11.3.1. *Railroad Rolling Stock*

Large numbers of railroad electric and diesel engines, passenger cars, and freight cars are in service. They are designed to function for at least 30 years with minimum maintenance of their coatings.

Coating systems for railroad rolling stock are used under severe working conditions. Engines and passenger cars are run at high speeds in many types of climates. They are frequently cleaned with strong chemical solutions to remove the heavy dirt adhering to the coating. Freight cars are also exposed to additional stress as a result of impact from mechanical handling during loading and unloading. Many transported goods (chemicals, fuels) attack the coating of freight cars. Coating systems for railroad rolling stock must be highly resistant in all respects. This is achieved by using two-pack coating systems (mainly epoxy but also polyurethane) and, more recently, coatings of acrylic resins which are applied as one-pack, waterborne dispersions.

Before applying the coatings, surfaces are pretreated by blasting. Longlife coatings require surface pretreatment according to Swedish Standard SIS 055 900, grade 2 1/2 (equivalent to DIN 55 928, part 4, FRG; BS 7079, part A 1, 1989, UK; SSPC-SP 5-SP 6, ASTM, USA; ISO 8501 – 8503).

Engines and Passenger Cars. The exteriors of engines, passenger cars, and similar vehicles (subway coaches) are normally painted in a three-coat system. An anticorrosive, two-pack epoxy primer is applied at a DFT of 80 µm. The curing agent is either an amine adduct or an aminoamide resin. Zinc chromate was used for many years as an effective anticorrosive pigment but has now been replaced by chromate-free pigments (e.g., zinc phosphates, barium metaborate, calcium borosilicates, and zinc phosphomolybdates) to avoid the risk of carcinogenicity.

Anticorrosive primers are followed by a 40 – 50 µm intermediate coat, based on two-pack polyurethane resins. This coat is sanded to yield a smooth surface.

Sometimes, the intermediate coat is overcoated with the topcoat in a wet-in-wet system. Normally, however, two-pack polyurethane topcoats are applied to the dried filler. These topcoats with aliphatic isocyanate hardeners have excellent gloss and color retention. The DFT of topcoats is 40 – 80 µm, depending on the hiding power of the topcoat.

Red, orange, and yellow pigments used in topcoats are now free of toxic lead and heavy metals to protect workers and the environment. Railroad companies expect a service life of 15 years before the topcoat has to be replaced (provided that spot repair is carried out when necessary). The whole coating system should not need to be renewed before a minimum service life of 30 years.

Although two-pack epoxy primers and polyurethane intermediate coats have high solids contents, they still contain significant amounts (20 – 30 wt %) of organic solvents. In polyurethane topcoats, the VOC is even higher. Anticorrosive, waterborne primers based on aqueous dispersions of two-pack epoxy resins and one-pack acrylic resins have been developed to decrease solvent emission. Waterborne, one-pack acrylic topcoats are also used. All of these waterborne paints contain 2 – 5 % organic cosolvents that are required for film formation.

Field trials have shown that the expected durability of the new waterborne coating systems is equal to that of conventional solventborne paints. As a further advantage, the number of single coats can be reduced from three to two. Standard freight cars are painted with a three-coat alkyd system consisting of an anticorrosive primer, an intermediate coat, and a topcoat, total DFT is ca. 150 µm. Such systems are being replaced by single-coat systems of waterborne acrylic resins and acrylic copolymers to reduce solvent emission.

Freight Cars. Special freight cars transport dry and liquid goods that can contain aggressive chemicals which would destroy alkyd coatings. An *exterior coating system* based on two-pack epoxy resins is therefore used. It consists of the same anticorrosive primer employed for engines and passenger cars. This two-pack epoxy primer is applied to blasted steel, sometimes also stainless steel. The next layer is a two-pack intermediate epoxy coating with chemically inert pigmentation, DFT 40 – 60 µm.

Topcoats are also based on epoxy resins, their DFT is 40 – 50 µm. Chemical resistance requires pigments that are not affected by the transported goods; the available color range is therefore limited.

Although epoxy topcoats show chalking and loss of gloss after a short period of outdoor use, this does not affect their chemical resistance. Chalking does not cause a significant reduction of film thickness.

When goods such as salts or fertilizers are transported, the interior coatings must be chemically resistant. Chemical resistance is obtained by using the same system that is used for exterior coating. In the case of food transport (e.g., flour, sugar, or grain) the coating must also comply with relevant legal regulations; thresholds are stipulated to limit migration of coating ingredients into the transported goods. The same resin system as for the exterior coating may be used, but there are strong limitations for the use of pigments, plasticizers, and additives.

In the case of abrasive freight goods, the interior of the car must be lined with a thick coat system consisting of solvent-free two-pack polyurethane or epoxy material. Toughness combined with flexibility results in a film that is highly resistant to abrasion.

Application. The standard application method for most paints used on railroad vehicles is airless spraying. A combination of airless and compressed air spraying has been recently introduced, it is mainly used for applying waterborne topcoats.

11.3.2. *Freight Containers*

Hundreds of thousands of freight containers have been built worldwide since the early 1960s. They are exposed to many types of climate and are heavily stressed by wear and tear. Freight containers consist of steel or aluminum alloy frames, clad with sidewalls and roofs made of steel, stainless steel, or aluminum alloys. The walls and roofs may be coil-coated. Container frames are usually blasted to grade 2 1/2 of the Swedish Standard SIS 055 900.

Prime coating is either a two-pack, zinc-rich epoxy primer with a DFT of ca. 30 – 40 µm, or a chromate-free, two-pack epoxy primer. Topcoats based on PVC copolymerisates, chlorinated rubber, silicone-reinforced epoxy esters, or alkyds were used for many years. Now, the standard European topcoat is based on acrylic resins with terminal epoxy groups that are cross-linked with amine adducts or aminoamide resins.

11.3.3. *Road Transport Vehicles*

Coatings on road transport vehicles are expected to have a minimum service life of ten years. Paint systems must therefore be of high quality.

Trucks. Drivers' cabins are coated in the same way as cars (see Section Car Body Paints). Truck chassis are especially prone to corrosion caused by stone impact and road salts. Chassis parts are supplied by special manufacturers who normally provide an anticorrosive prime coating.

Side and cross members are pretreated by blasting and phosphating. An electrodeposited primer coating follows and is sometimes immediately overcoated with a topcoat based on an air-drying alkyd or epoxy ester, or on an oven-drying alkyd – melamine resin combination. After assembly of the truck, a third layer (i.e., a second topcoat) is applied that serves as a supply finish. These finishes are mostly two-pack acrylic –aliphatic isocyanate topcoats.

Trailers. Trailer chassis are generally coated in the same way as truck chassis but with a two-coat system. The primer is either a one-pack alkyd paint or a two-pack epoxy paint that are cured at ambient temperature with amine adducts or isocyanates.

Truck bodies are often constructed of steel frames with aluminum side walls. The aluminum alloys are pretreated by application of 5 – 8 µm of a chromate-containing wash primer. The topcoat is a two-pack acrylic resin bearing hydroxyl groups and is cross-linked with aliphatic isocyanates.

Buses have a profiled steel framework that is clad on the sidewalls and roof with steel, galvanized steel, and aluminum alloys. The anticorrosive primer must therefore adhere well to all of these metals, and to glass-fiber-reinforced plastics.

Two-pack epoxy primers with chromate-free pigments show the best performance. They are applied with a DFT of ca. 80 µm, and simultaneously form the anticorrosive primer and a sandable filler. They can protect the metal substrates against corrosion.

The metals are joined by spot welds and rivets that may cause electric currents in the joining areas. Specially designed epoxy primers prevent corrosion in these areas. At ambient temperature epoxy primer fillers dry thoroughly in about 16 h. Since this is too long for assembly lines, they are normally cured in ovens (80 °C metal temperature).

Bus topcoats are mainly two-pack acrylic –isocyanate paints which dry very rapidly at ambient temperature or under low baking conditions. This is important

because many buses, especially luxury coaches, are painted in multistripe color designs. The resulting coatings have excellent color and gloss retention under service conditions. This is not only important for long-lasting corrosion protection, but also for the image desired by the owners.

11.3.4. *Aircraft Coatings*

Aircraft coating systems have an extremely high quality. Corrosion must be prevented to guarantee safe functioning of the aircraft. Aircraft manufacturers issue very comprehensive specifications to ensure the required coating properties. This demands extensive development work and careful paint formulation, because the coating properties must be obtained with a minimum dry film thickness to minimize weight. Aircraft coating systems must also be resistant to hot hydraulic fluids.

The aluminum alloys used for aircraft construction are pretreated by anodizing, chromating, application of wash primer (6 – 8 µm), or pickling.

Anticorrosive primer formulations are based on epoxy resins that react at ambient temperature with amine adducts, aminoamide resins, or isocyanates (two-pack primers). Zinc or strontium chromates are used as anticorrosive pigments. These electrolytically active pigments prevent dangerous filiform corrosion. Extensive development work is being carried out on chromate-free aircraft primers [543], [544].

Aircraft coatings normally consist of two coats. The topcoat used for civil aircraft is a high-gloss two-pack polyurethane product, cured with aliphatic isocyanates. Single-coat systems based on thermoplastic acrylic resins have been used for military aircraft, but are being replaced by polyurethane topcoats [545].

11.4. Marine Coatings [546-548]

The requirements of any coating system are determined by its in-use environment (service, construction, or maintenance). The sea is wet, salty, usually oxygenated, sometimes anaerobic, and full of living organisms which may colonize the surface of a vessel or platform [546].

During construction and maintenance cycles, coatings have to adapt to the environment without disrupting time schedules and productivity. They must also conform to increasingly stringent considerations of occupational safety and health, and to environmental constraints. While construction to the “block-stage” (stage at which the prefabricated block is transferred to the building dock) frequently takes place under cover, final assembly and painting take place under ambient conditions (–10 to + 40 °C, and 40 – 100 % relative humidity). Marine paints must be able to tolerate this range of climate as well as application methods and skills. For maintenance, the advent of larger vessels, smaller and more specialized crews, and increased vessel activity, means fewer or no opportunities for in-service maintenance. Improved plant reliability means longer intervals between drydocking.

The requirements of marine coatings may therefore be considered under the following headings:

1. *Substrate*: usually hot-rolled mild steel, high-tensile steel, or aluminum
2. *Surface preparation*: usually centrifugal or pneumatic blasting with recycled shot or expendable grit blasting media
3. *Priming*: the dominant consideration is the effect on the efficiency and quality of new construction
4. *Design requirements*: examples are anticorrosive, antifouling, chemical resistance, impact and abrasion resistance, cosmetic qualities, friction and camouflage at various acoustic and electromagnetic wavelengths

11.4.1. *Substrate, Surface Preparation, and Priming*

Substrate. In marine construction the substrate is generally hot-rolled mild steel, with some high-tensile steel in highly stressed areas (apertures in container ships), and aluminum where there are weight penalties (e.g., topsides of passenger liners, ferries, and naval vessels).

Surface Preparation. Steel hot rolled at 800 – 900 °C acquires a tenacious oxide layer (mill scale) that is cathodic with respect to the steel to the extent of about 300 mV. In the presence of an electrolyte (seawater containing 3.5 % salts, mainly sodium chloride) the steel would corrode and pit and roughen severely. The first process in new construction and refurbishment is therefore the complete removal of mill scale. A small amount of very light-gauge steel is prepared by acid pickling, but most steel for ship and off-shore construction is centrifugally or pneumatically blasted with steel shot that can be recycled or with expendable abrasive grit. Freedom from scale and soluble salt contamination are the main requirements, texture and profile are less important [548].

Priming. In new construction the freshly blasted steel surface is highly reactive and requires corrosion protection during construction. It immediately passes through an automatic shop or blast priming plant. The design criteria for shop primers are set principally by the ship-builder, with the consequence that their “fitness for purpose” is compromised. The main criteria are:

1. *Speed of Drying.* The primer is applied to the substrate on a horizontal moving conveyor; the plate must dry quickly to allow handling and stacking within a few minutes.
2. *Speed and Quality of Welding.* The primed plates are cut and welded by a number of techniques (plasma, submerged arc, inert gas, manual metal arc, laser). The primers must not interfere with the speed and quality of welding. The principal defects of the previous generation of materials was their high organic content, which yielded porous welds.
3. *Occupational Safety and Health.* Emission of toxic vapors must be minimized; hence, shop primers that form inorganic films (zinc silicate) are now mainly used.
4. *Lifetime.* Ship construction times have dropped and much prefabrication is done undercover, so that anticorrosive protection (weathering) during construction may only be required for 3 – 9 months.
5. *Burning and Weld Spatter.* Limiting degradation of the primer in the heated zone and nonadhesion of weld spatter are important considerations for subsequent coating operations.

The above constraints mean that only a limited range of coating materials is suitable for marine construction:

1. Wash or etch primers based on poly(vinyl butyral) and phenolic resins pigmented with zinc tetroxychromate
2. Epoxy primers cured with amine adducts or polyamides, pigmented with zinc, zinc chromate and potassium chromate, strontium chromate, or, more recently zinc phosphate and calcium phosphate
3. Epoxy primers cured with amine adducts, containing inert pigments (usually iron oxide)
4. Epoxy primers, zinc-rich, cured with polyamines or polyamides
5. Epoxy primers cured with amine adducts, containing leafing aluminum flake pigmentation

These organic primers have, however, largely been discarded in favor of the “organic zinc silicate” shop primers (see also Section Alkyl Silicates). These are low-zinc versions of zinc silicate anticorrosive primers. Low zinc concentrations are used to minimize volatilization of zinc metal and formation of zinc oxide aerosols at welding temperatures.

Dilution of the zinc with other pigments (extenders) requires careful compromise: the protective qualities of a conductive zinc film with continuous contact should be retained without increasing the emission of toxic vapors. Diluent pigments are therefore chosen for their heat resistance and contribution to the impermeability of the films; they include titanium dioxide, iron(III) oxide, talc, mica, and china clay.

Limitations in performance are several. Zinc is volatilized at weld lines and cut edges, reducing anticorrosive protection. In areas of marine and industrial pollution, thorough washing and/or grit sweeping is necessary to remove water-soluble corrosion products of zinc (sulfate, chloride, and ammonium salts) before overcoating. Performance is also reduced if the complete paint system requires chemical or impact and abrasion resistance. In such cases reblasting is necessary (particularly in chemical tanks, but also in ballast spaces and on the external hull) to completely remove the shop primer. Grit sweeping may also be used but is not as efficient. Maximum primer thickness must be strictly controlled to about 25 µm.

11.4.2. *Ship Paint Systems*

The painting areas of a ship are shown in Figure 33. The main requirement of marine coating systems is corrosion prevention. Detailed requirements vary with the particular internal or external area (e.g., chemical resistance in cargo tanks; resistance to seawater in ballast spaces; heat resistance in engine rooms; impact and abrasion resistance on boottoppings, external hulls, and decks; cosmetic qualities on superstructure and topsides).

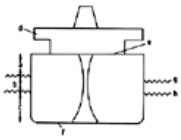

Figure 33. End-on view of a ship's hull showing painting areas

a) Topsides; b) Boottopping or "wind and water line" (may be exposed or immersed); c) Lower sides (always immersed); d) Superstructure; e) Decks; f) Bottom; g) Water level (heavy load); h) Water level (light load)

Although still referred to as anticorrosives, marine coatings are technically barrier coatings that depend on ohmic resistance for their protective qualities. The ohmic resistance is maintained by impermeability to water and hydroxyl ions. The use of toxic anticorrosive pigments such as red lead, white lead, lead silicochromate, and zinc chromate has declined for safety and health reasons. Zinc (or calcium) phosphate is one of the few anticorrosive pigments still in common use. There has also been a move away from medium oil alkyds and tung oil/phenolic drying oils for primers and undercoats, and long oil alkyds for finishing. They are convenient maintenance materials, but reduced maintenance requirements have meant a move towards nonconvertible (chemically drying) epoxy and polyurethane coatings, and a decline in the use of convertible (physically drying) acrylic, vinyl, or chlorinated rubber coatings for topsides and boottoppings. Retention of color and gloss have been improved with rust-hiding, antistaining additives (e.g., calcium etidronate or other iron(III) sequestering agents).

For similar reasons in the boottopping area and on decks, where the most severe mechanical and corrosive problems occur, traditional drying oil – red lead compositions are being replaced by solvent-free or high-solids, impact- and abrasion-resistant epoxies with outstanding mechanical properties. External underwater hull coatings have changed more slowly. They originally consisted of mixtures of coal tar pitch, modified with natural bituminous materials (e.g., asphaltums and gilsonite) by cooking with lead and drying oils, pigmented with leafing aluminum flake. They had a high water and oxygen impermeability but deficient mechanical properties. Furthermore they are not resistant to cathodic protection, either by impressed current or sacrificial zinc or aluminum anodes.

Coal tar pitches are less readily avialable than they used to be; their content of polynuclear aromatic hydrocarbons makes their safe use suspect. They have been replaced to some extent by the less effective petroleum bitumens whose mechanical properties are improved by combination with poly(vinyl chloride) and other chlorinated polymers as one-pack, nonconvertible coatings, or by combination with epoxy or polyurethanes in high-performance two-pack convertible coatings. Similar materials are preferred for coating ballast spaces and double bottoms in vessels, where economical systems are required that do not have to be attended throughout the lifetime of the vessel (15 – 25 years). Multiple coats with a total thickness of 250 – 400 µm are usually employed.

The environment in chemical tanks is amongst the severest to which marine coatings are subjected. Each tank may have to transport some 1500 bulk, liquid cargoes that include crude oil, refined gasoline, aviation spirit, diesel oil, solvents, vegetable oils, or wine. Inorganic cargoes are also carried in solution (e.g., alkalis and acids). New demands also appear such as methanol as a fuel and feedstock, and methyl *tert*-butyl ether as an additive for lead-free petrol.

No single type of coating is universally applicable, and nonpaint alternatives may also be used (e.g., rubber linings and stainless steel). Vessels that are not dedicated to a single type of cargo (parcel tankers) usually have a number of coating systems.

The tank coating has three purposes:

1. To prevent corrosion of the steel
2. To prevent contamination of the cargo
3. To facilitate cleaning and avoid cross-contamination of one cargo by another

Variants of six main paint systems are used (Table 30).

Table 30. The most important paint systems for marine tank coatings

Paint system	Advantages	Limitations
Zinc silicate	outstanding neutral solvent resistance	poor acid and alkali resistance poor film formation at low humidity
Epoxy cured with polyamide	tolerant to application conditions	poor resistance to polar solvents
Epoxy cured with amine adduct	wide range of solvent resistance	minimum application temperature 5 °C
Epoxy cured with isocyanate	low temperature, elastic and adhesive	poor alkali resistance dangerous to human health
Epoxy–phenolic resin	outstanding overall resistance, including methanol	requires heat for ultimate cure
Polyurethane	excellent low-temperature curing	poor methanol alkali resistance and dangerous to human health

Zinc silicates are outstanding for nonreactive hydrocarbon cargoes. The remaining systems are all highly cross-linked and their advantages are generally a function of the cross-linking density and their specific chemical stability. Heat is of great advantage in achieving adequate cross-linking density. Some systems require the carriage of heated cargo (e.g., palm oil) for up to six months after application. Heating, ventilation, and film thickness all need to be closely controlled during application. The steel surface temperature has to be maintained at a constant level. In Northern latitudes the outside of the vessel may be lagged with polyurethane foam and the tanks are ventilated with hot dehumidified air. Certain sequences of cargo have to be avoided since the coating systems swell and are stressed; for example, permeability to water (from damp cargo or steam cleaning) may increase following carriage of methanol or contact with tank-cleanning chemicals. Chemical tank coating is thus one of the most critical coating processes requiring careful coordination of coating formulation, testing, manufacture, surface preparation, and control and inspection of application. The paints are generally applied by airless spraying. Difficult geometrically complex areas and welds are usually cleaned, smoothed by mechanical abrasion, and coated with a brush [549-552].

11.4.3. *Fouling and Antifouling*

Fouling and antifouling are unique to the marine industry. In contrast to other coatings whose primary function is structural, preservation, the outer hull surface of a ship also has to be smooth to maximize ship speed and reduce fuel consumption [553]. Fouling with marine organisms, considered as biological roughness, can be immediate and drastically increase fuel consumption or reduce speed. Some 630×10^6 deadweight (tonnes) of marine transport burns 184×10^6 t of fuel at a cost of \$ 18.4×10^9/a. Physical roughness would increase this by 10 %, fouling by 30 – 40 %. The improvement in fouling control and roughness of ships made since 1974 has been calculated to have led to worldwide annual fuel savings of \$ 720×10^6; the extension of drydocking from 21 to 28 months has been calculated to have saved \$ 800×10^6.

Much effort has been directed to the prevention of fouling [547], [554]. The major concern on wooden ships was the attack of wood-boring mollusks (*Teredo* and *Martesia*) or crustacea. Almost any sessile marine organism may be found as a fouling organism. Seriousness depends on size and frequency of occurrence; severity of settlement depends on geography and vessel itinerary. Because of the wide range of organisms (e.g., bacteria, diatoms, algae, barnacles, hydroids, mollusks, ascidians, and sponges) methods of prevention have to be extremely general. All antifoulings have so far depended on a small range of organic, metallic, or organometallic biocides. They include copper, water-soluble or water-reactive compounds, organoarsenical, organomercurial, organolead, and organotin compounds. The biocides function by slowly dissolving (leaching) in the seawater adjacent to the hull and killing the organism in question. There is a critical leaching rate for each biocide and for each organism. The objective of antifouling design is to select an optimum biocide or mixture of biocides and to control the release rate at just above the critical leaching rate for macrofouling (i.e., fouling that affects ship performance). Ideally the leaching rate should be linear or zero order. Because leaching has to continue for 3 – 5 years under adverse environments only the most potent biocides have been found useful (LD_{50} against target organisms may be 3 ng/mL). Control of release was at first poor because the choice of matrix or vehicle was limited to a small number of natural oils and resins with the appropriate permeability. Controlled-release media have, however, since been developed. Stable film-forming resins composed of trialkyl- and triaryltin esters of acrylic copolymers are used [555-557]. When exposed to water the film dissolves with zero-order kinetics and the surface becomes smoother in turbulent flow [558]. Self-polishing or ablating antifoulings based on tributyltin and copper(I) oxide can have antifouling lives of five years or more. Trade names include Intersmooth (International Paint, UK); Nautic (Hempel, Denmark); Seaflo (Chugoku, Japan); and Seamate (Jotun, Norway).

For reasons of safety, health, and environmental protection, organolead, -arsenic, and -mercury compounds are no longer used [559]. Use of organotin compounds is restricted to commercial vessels. Although nontoxic, nonpolluting, and nonhazardous solutions to the fouling problem are needed, alternative compounds to the tributyltin acrylates have proved extremely difficult to find. However, apart from their susceptibility to biocides one of the common features of all fouling organisms is adhesion. Prevention of adhesion with antiadhesive coatings has been successful [560-562]. This approach requires an understanding of the physics and chemistry of bioadhesion. Hydrophobic fluoropolymers have proved ineffective as marine antiadhesive coatings. Silicone elastomers are adequate for most release purposes and physiologically benign, they have shown the most promise. Claims have also been made for nonadhesion with hydrophilic agents, but much work remains to be done in this area [563].

11.5. Coil Coating [536], [537]

Coil coating is a special application method in which coiled metal strips are unwound and then passed through pretreating, coating, and drying equipment before being finally rewound (see Section Miscellaneous Wet Paint Coating Methods). The coated metal strips (0.2 –2 mm thick) are supplied as coils, and are used to manufacture a wide variety of metal constructions, ranging from flat panels to complex geometrical forms. The organic coating film must have a high elasticity to allow the metal to be shaped without damaging the film at the edges.

Coil coating allows very efficient coating of large surface areas in a short time, the coated surfaces have a very high quality. The metal sheets are shaped to the required form after coating.

Coating coils of strips in a large coating facility instead of in many small ones results in optimization of the coating process and reduces pollution of the environment (lower emission of volatile organic solvents) [536].

Prior to coating, the metal surfaces must be cleaned and pretreated with aqueous solutions to form conversion coatings. The following metal surfaces can be coil-coated:

1. Steel strips with a width up to 1850 mm
2. Electrolytically galvanized metal bands
3. Hot-dipped galvanized steel bands (using the Sendzimir method)
4. Strips from aluminum alloys with a width up to 1650 mm

Coil coatings are normally applied by roller coating machines, but are sometimes sprayed. The topside of the metal band is normally painted with a primer (DFT ca. 5 µm) and a topcoat (DFT 20 – 22 µm). Zinc-rich coil coatings are applied with a DFT of 5 – 20 µm, plastisol films have a DFT of 80 – 400 µm. The reverse side of the coils is coated with backing coats based on binders, such as alkyd, polyester, and epoxy resins (DFT 8 – 10 µm).

After a very short flash-off time (3 – 10 s), the coated metal passes through a drying oven, which can be up to 50 m long. Drying time is 20 – 60 s at 180 – 260 °C, depending on the coating type.

Epoxy primers are used as general-purpose primers on steel, galvanized steel, and aluminum, and under all types of finishes. These primers give excellent corrosion resistance on steel. Acrylic primers have a corrosion resistance similar to that of epoxy primers, and are used primarily under fluorocarbon finishes because of their superior intercoat adhesion.

Polyester primers are widely used in Western Europe. Although their corrosion resistance (salt spray test) is not as good as that of the other primer types, their mechanical properties are superior.

The choice of binders for the topcoat depends on the end use of the coil-coated metals. Acrylic topcoats were developed earliest, and are widely used indoors for surfaces that are not exposed to water, chemicals, or mechanical stress. Polyester coil coatings provide a good general-purpose finish, and can be used for a wide range of applications (e.g., car interiors and accessories, caravan exteriors, domestic appliances). Tough, abrasion-resistant, and durable general-purpose finishes can be formulated with polyurethane or polyester resins.

Topcoats based on fluorocarbon resins [poly(vinylidene fluoride)] produce extremely durable coatings (life to first maintenance > 20 years under most climatic conditions) with excellent color and gloss retention. Like silicone-modified resins they are used for applications where weather resistance is required (e.g., exterior panels on buildings).

11.6. Coatings for Domestic Appliances [536]

Domestic appliances (e.g., refrigerators, deepfreezers, dishwashers, washing machines, and laundry dryers) are mainly made from blank sheet iron, or electrolytically or hot dip galvanized steel. Coatings must protect these construction materials against corrosion caused by food ingredients, household chemicals, and the humid atmosphere in a kitchen or laundry.

Domestic appliances are manufactured in large numbers necessitating adequate application and drying facilities and fast production lines. Stoving materials are used that cure in a short time at comparatively high temperatures (10 min at 160 °C or 30 min at 120 °C). Binders for these stoving paints are mainly self-cross-linking acrylic resins that cure by polycondensation to give a smooth homogeneous film with excellent mechanical and chemical resistance.

Stoved acrylic resin films are generally applied as single coats which requires careful pretreatment of the metal surfaces by cleaning and phosphating. The paints are applied by electrostatic spraying using rotating bells or disks.

Certain components of domestic appliances are also coated with electrodeposition paints. Two-pack polyurethanes are applied if the dimensions of the parts are too large for the electrodeposition bath. These paints dry at ambient temperature to give films with similar properties to stoving acrylic enamels. Solvent-free powder coatings are also used to coat domestic appliances.

11.7. Coatings for Packaging (Can Coatings) [536]

About 100×10^9 cans are produced annually worldwide for packing perishable food (⟶ Foods, 2. Food Technology – Food Canning Technology). Cans can be considered as a single material, consisting of a metal substrate with an organic lacquer.

The interior coating of cans is very important because it prevents the metal from reacting with the ingredients of the filling goods. The can exterior is painted to prevent corrosion, but also for decorative reasons. Coatings based on combinations of polyester or acrylic resins with melamine resins are used for can exteriors. Lacquers based on phenolic resins are especially resistant against aggressive can contents. Curing at 200 °C produces densely cross-linked films with high chemical resistance, but poor flexibility.

Combining phenolic resins with epoxy resins results in films with good chemical resistance and flexibility. Interior coatings based on these resin combinations are applied with a solids content of 33 – 43 % and cured at 200 °C in 80 s. Epoxy–phenolic resin lacquers are used in high amounts for can coating.

Interior can coatings based on PVC organosols are highly flexible and resistant. They are used mainly for cans which are heavily shaped during the manufacturing process.

Three-piece (body and two lids) cans are now welded instead of soldered. The welding seam is coated separately with epoxy – phenolic resins or PVC organosols that are cured at 260 °C for 10 – 20 s. The PVC organosols result in higher dry film thicknesses, whereas the epoxy–phenolic welding seam coatings have better chemical resistance.

Special filling goods and a long shelf life sometimes require an extremely high dry film thickness for the welding seam coating. Powder paints are then applied electrostatically to give a dry film thickness of ca. 50 µm.

Beverage cans are two-piece (body and lid) cans produced by the drawn and wall ironing (DWI) process. They require highly flexible coatings based on epoxy resins cross-linked with aminoplast resins.

For many years, paints used to coat cans contained considerable amounts of volatile organic solvents. Waterborne can coatings were developed to reduce solvent emissions and are used worldwide. Binders used in waterborne can coatings are modified epoxy resins (see Section Epoxy Coatings). Acidic acrylate

chains are grafted onto an epoxy molecule. After partially neutralizing with amines, the resins can be dispersed in water.

Both waterborne and solventborne can coatings must not affect the can contents, especially their taste. Components of coating films are not allowed to migrate into food, beverages, or other filling goods. In most countries, food packaging is subjected to legal regulations. The raw materials used to produce can coatings and the coatings themselves are strictly limited.

11.8. Furniture Coatings [536]

See also ⟶ Wood, ⟶ Wood, Surface Treatment, and ⟶ Wood, Preservation

The long-term value of furniture depends to a high degree on its surface characteristics. Untreated or uncoated surfaces very quickly lose their good appearance and deteriorate under conditions of use. Different forms of wood are used for furniture (e.g., solid wood, veneer, plywood, particle board, and chipboard). Solid wood and veneer may originate from different types of trees with a wide range of properties partly due to their contents of resins and essential oils.

The natural humidity content of fresh wood has to be lowered to a maximum level of 10 – 15 % before it is coated. Continuously varying atmospheric humidity in air leads to changes in the volume of the furniture wood. Furniture coatings must therefore have excellent film flexibility as well as film hardness, and resistance to abrasion and fluids (e.g., alcohols). Prior to coating, wood surfaces usually have to be smoothed by sanding, using putty, patinating, staining, or pore filling.

Resins and essential oils are often extracted with organic solvents to prevent the coating from cracking, discoloring, and developing other faults.

Automatic spraying, machine roller coating, and curtain coating are used for industrial application of furniture coatings. Drying is accelerated by air circulation and/or increased temperature up to 80 °C. IR drying ovens and tunnels are also used.

Furniture coatings can be clear or pigmented with many different color shades. The surface can be high gloss, semigloss, matt, or textured. Some coatings leave the wood pores open to give the wood surface a more natural appearance; the “bottom” of the pores must, however, be coated evenly to protect the wood surface completely. Application of a film that fills the wood pores results in a completely smooth surface coating.

For many years, nitrocellulose coating systems were preferred for indoor furniture and other wood parts. These systems are still used on low-price furniture, but are gradually being replaced by coatings based on polyurethanes and unsaturated polyesters. Most modern wood and furniture coatings are based on special acrylic resins and unsaturated polyester resins, cured by UV or electron-beam radiation [565], [566].

The properties of furniture coatings can be summarized as follows:

1. *Nitrocellulose Paints*. Cheap, high content of organic solvents, low solids content, dry at ambient temperature, not very resistant to chemicals or mechanical stress, dried film is soluble in organic solvents.
2. *Polyester Paints*. Binder based on unsaturated polyester resins (copolymerized with styrene), low emission of organic solvents, catalytic curing by organic peroxides or UV radiation, highly resistant to abrasion, alcohols, and other chemicals (cleaning agents).
3. *Polyurethane Paints*. Polyester, polyether, or acrylic resins containing hydroxyl groups are cured with isocyanate hardeners, medium to high solids contents, best overall properties in terms of mechanical and chemical resistance [567].

11.9. Coatings for Buildings

Coatings for buildings are solvent- or waterborne. They include coatings that can protect all materials used in building and construction work (e.g., wood, steel, light metals and alloys, plastics, concrete, plaster) against corrosion and decomposition; they can also give a decorative appearance.

Solventborne paints are mainly alkyd paints with mineral spirit as solvent. Paints based on synthetic resins and two-pack paints with hydrocarbon, ester, or ketone solvents are, however, also used. Gloss ranges from silk-matt to high gloss.

Waterborne systems are formulated with pure acrylates and have a solvent content below 10 %, preferably glycols. The gloss is between silk-matt and full gloss.

Emulsion paints (see also Section Waterborne Dispersion Paints (Emulsion Paints)) with a solvent content below 4 % are based on vinyl or acrylic polymers and copolymers (e.g., with styrene). Gloss is between matt and silk-glossy.

Architectural paints are applied in situ with brushes or rollers in Europe, but spray application is widely used in the USA on concrete and wooden buildings. Paints on windows and doors are brush applied. Structural components (e.g., doors, windows, radiators) that receive primers or complete coating systems during production or in the workshop are dipped or sprayed.

Since the decorative effect is a very important factor in architectural paints a large number of shades in various gloss gradations are available. Manufacturers' color cards, the German RAL-register (Deutsches Institut für Gütesicherung und Kennzeichnung), and the Swedish NCS (Natural Color System) shade charts provide a survey of available shades.

11.9.1. *Exterior-Use Coatings*

Exterior-use coatings must be weather-resistant: they must adhere to a wide range of substrates and retain their gloss and shade fastness for a number of years. Suitable binders and pigments must therefore be chosen [568].

Coatings for Mineral Substrates. The binders must be particularly resistant to alkali because otherwise hydrolysis can occur in the presence of moisture due to the high pH of the substrate. Aqueous systems such as silicate paints and vinyl or acrylic emulsion paints with silicone resin additives are suitable for opaque and semitransparent coatings. Solventborne systems are based on synthetic resins (e.g., styrene – acrylate polymers). Two-pack polyurethane coating materials with an aliphatic isocyanate as hardener can also be used. Colorless, water-repellent impregnations are obtained with silicone resins or siloxanes in solvent and silicates in water [569].

Coatings for Wood. Wooden structures can receive colorless, semitransparent, or opaque coatings. The binder and pigment of the systems must be formulated such that a sufficient film thickness reduces exchange of moisture and thus swelling and shrinkage of the wood.

The UV transparency of the coating has to be low to prevent depolymerization of the wood lignin. This is achieved by using UV absorbers in varnishes and transparent iron oxides in stains [570].

Long oil alkyd resins are used as binders for solventborne products; acrylic resin dispersions alone or mixed with alkyd resin emulsions are used for waterborne products. Stains may be of the impregnating or coating type. Low-build stains (solids content < 30 %) are used on non-dimensionally stable wooden surfaces and on dimensionally stable structural parts as priming coats only. High-build stains (solids content > 40 %) are used as intermediate- and topcoats to protect dimensionally stable structural parts against direct weathering [571].

Coating for Metals. Steel, zinc, and light metals are largely topcoated with alkyd resin systems. Choice of an appropriate surface treatment and a suitable primer are important, particularly with galvanized surfaces and light metals, because adhesion to these substrates presents difficulties [572].

Primers based on modified alkyd resins or two-pack epoxy resins for derusted ferrous metals mainly contain zinc phosphate and zinc oxide as a corrosion protection pigment. Nonferrous metals are first washed with an ammoniacal wetting agent before applying the primer that contains a binder based on synthetic resins (e.g., PVC copolymers, chlorinated rubber) which ensure good adhesion to the substrate. The same primers must be used on zinc or galvanized surfaces because the use of alkyd resins causes embrittlement [573].

Coatings for Plastics. The use of plastics in building and construction (e.g., glass-fiber-reinforced resins, rigid PVC) is constantly increasing, as is the trend towards renovating such surfaces. The surface is first cleaned and, if necessary, roughened. A one-pack primer based on chlorinated rubber or chlorosulfonated polyethylene (as is conventional with galvanized surfaces) or a two-pack epoxy primer is then applied. The primer can then be coated with a topcoat system that is usually formulated with medium or long oil alkyd resins [574].

11.9.2. *Interior-Use Coatings*

Wall coatings for mineral substrates are largely formulated as dispersions. Cleanability (wipefastness, washfastness, abrasion resistance) rather than weather resistance is, however, the prime concern (DIN 53 778). Fungal contamination can easily occur in damp, moist areas and is prevented by adding fungicides (e.g., carbamates or imidazoles).

Various surface effects can be produced by varying the viscosity and adding coarse, possibly colored extenders or fibers. Two-pack systems based on polyurethane resins or epoxy resins are used for wall coatings that require a good resistance to agents used for chemical cleaning and decontamination.

Floor Coatings. Concrete floors are coated with low-solvent or solvent-free epoxy or acrylic resin materials that may be applied in any desired thickness. They are extremely resistant to abrasion, can be made slip resistant with sand, silicon carbide, or high-grade steel granulate, and are also resistant to mineral or vegetable oils and gasoline (used for warehouses and factory halls).

Pigmented, two-pack, waterborne epoxy resin coatings (garages) or one-pack waterborne acrylic resin emulsion paints are used for areas that receive less wear (e.g., cellars). Wooden parquet floors are coated with one- or two-pack polyurethane varnishes that can be applied by spraying or brushing. Acrylate-based waterborne parquet varnishes are also used because they are environmentally friendly.

Radiator coatings are intended to protect radiators against corrosion without, however, affecting their heating effects (DIN 55 900). Primers based on special alkyd resins are generally applied industrially. They have to satisfy the usual requirements for preventing radiator corrosion during transportation and at the building site. The topcoats applied on site by rolling or inundation are based on medium oil alkyd resins.

Heating oil storage premises must be equipped with a collection trap so that any heating oil leaking from the tank cannot contaminate the soil. The interior of these premises must be painted with an officially approved coating material that is not dissolved or penetrated by heating oil. The coating must also cover cracks in the substrate. Multilayer systems based on waterborne acrylic resin dispersions are suitable for this purpose.

Fire Retardant Coatings. The flammability of combustible wood structures can be reduced in accordance with DIN 4102 by applying a fire retarding paint that forms an insulating layer [572]. Dispersion paints based on poly(vinyl acetate) with addition of ammonium phosphate, a nitrogen compound (e.g., melamine), and a carbon-forming agent (e.g., pentaerythritol) are suitable for this purpose. The thermal insulation is so good that ignition can be delayed by at least 10 min.

[Top of Page]

12. Environmental Protection and Toxicology

Wolfgang Plehn

Paints and coating materials frequently contain substances that may be a hazard both to human health and to the environment. This applies particularly to organic solvents, to certain reactive binder constituents, to pigments containing heavy metals, and to some additives. Evaluation of the environmental properties of paints must take into consideration their effects on the atmosphere, water, and the soil, the potential danger to the user, the use of low-residue application techniques, and the suitability for use. The primary concern is to minimize adverse effects in all sectors.

12.1. Clean Air Measures

Organic solvents in paints constitute significant sources of atmospheric pollution. The direct effects of these substances and their mixtures (see Section Toxicology), particularly the odor nuisance, should be taken into account in the vicinity of sources of solvent emissions. In the atmosphere the solvents gradually decompose or participate in chemical reactions under the influence of sunlight or traces of other substances present in the air. Photochemical decomposition in the presence of nitrogen oxides leads to formation of intermediates which are termed photooxidants on account of their oxidizing action. Ozone is regarded as the tracer for photooxidants. Even low concentrations of photooxidants harm plant life and may damage the human respiratory tract. Atmospheric pollution caused by photooxidants can occur particularly in the summer (summer smog), and was first observed in Los Angeles, where measures were adopted at an early stage to reduce pollution.

Solvent emissions can be reduced in three ways:

1. Use of low-solvent or solvent-free products
2. Application methods with low solvent emissions
3. Implementation of waste air treatment as a secondary measure in paint application facilities

In painting work performed by tradesmen, emissions can only be lowered by using low-solvent products.

Low-Solvent Coating Materials. A great deal of progress has been made in the development of low-solvent paints [575]. The use of water as a replacement for organic solvents is particularly important in this respect, although alternatives such as powder coatings or one-pack and two-pack high-solids paints have also led to significant improvements. Solvent emissions can be drastically reduced by using low-solvent coating materials. This is particularly apparent in automobile finishes: in a metallic effect finish based on conventional coating materials and solvent contents of up to 85 %, the specific emission is ca. 180 g of solvent per square meter of car body. With an average car body surface of 90 m^2, ca. 16 kg of solvents are released per automobile. By using low-solvent coating materials (electrodeposition paints, waterborne primer surfacers, waterborne metallic basecoats, and a conventional solventborne clearcoat) solvent emissions in modern plants are now only 30 – 45 g/m^2. Waterborne clearcoats are under development, but the use of powder clearcoats is also under discussion. Use of these systems should allow solvent emissions to be reduced further to ca. 20 g/m^2.

In other industrial application sectors there is an increasing tendency to employ low-solvent paints. In furniture production, for example, waterborne, UV-curing paints are superceding high-solvent nitrocellulose lacquers.

Stringent coating stability requirements on metal surfaces can be met by using powder coatings. Use of powder coatings is therefore increasing in new areas of application (e.g., for the colored coating of sanitary ware [576] and corrosion protection of truck tanks [577]).

The need for waste air treatment can generally be avoided by replacing conventional, high-solvent paints by low-solvent coatings. In some cases smaller waste air treatment plants may be necessary for specific demarcated areas where, for example, a solventborne layer is applied.

In the architectural paints sector, conventional, high-solvent paints based on alkyd and synthetic resins can largely be replaced by waterborne dispersions of acrylate polymers and copolymers, aqueous alkyd resin emulsion paints, and oil-based high-solids paints. Modern low-solvent coating materials not only decrease environmental pollution, but have also improved paint quality (e.g., resistance to weathering). In some countries low-solvent architectural paints are clearly labeled as such to assist workers and users. In the Federal Republic of Germany waterborne emulsion gloss paints containing ≤ 10 wt % solvent and high-solids paints containing ≤ 15 wt % solvent are identified by an “environment label”, a symbol created for products having a favorable environmental compatibility. In Switzerland paints containing ≤ 5 wt % of solvents may be labeled as “solvent-free”, and with ≤ 30 wt % as “low-solvent”. Some states of the United States (e.g., California) stipulate maximum solvent limits for individual product groups.

Application Methods with Low Solvent Emission. The method used to apply the coating material also influences the level of solvent emissions as well as the amount of waste and water pollution. Application efficiency depends on the application method and shape of the article being coated. Application methods are described in Chapter Paint Application. Spraying and dipping are most commonly employed for industrial coating. Spraying methods have the highest solvent emissions because overspray losses are unavoidable, and low-viscosity paints generally have to be applied. With conventional high-solvent paints the solvent content at the ready-to-spray viscosity value may be as high as 90 wt % but is usually 50 – 70 wt %. In conventional spraying with compressed air the application efficiency is particularly low (20 – 65 %, depending on the shape of the part being coated). Higher application efficiencies are achieved with airless and electrostatic spraying methods, and may reach 90 % in the case of electrostatic spraying. The overspray of coating powders can be largely recovered and reused directly. Problems arise with changes in color which can, however, largely be solved. The highest application efficiencies for wet paints (nearly 100 wt %) are obtained with brushing, rolling, and pouring methods, which, however, can only be used on flat surfaces. Parts with hidden areas can be effectively coated by dipping methods.

Waste Air Treatment. Two main methods are used for treating waste air in coating and paint shops: afterburning with heat recovery and adsorption with solvent recovery. In order to reduce expenditure on waste air treatment, the amount of waste gas should be minimized by enclosing the paint application area

and recycling circulating air. Water washers (scrubbers), fabric filters, and electrical separators are used to remove paint aerosols from the atmosphere. Recycling of circulating air is already common practice in modern automated paint shops (e.g., in the automobile industry). If the workforce is protected against relatively high solvent concentrations by respirators equipped with a fresh air supply, manual spraying zones can also be operated with recycled air. If such a concentration procedure is not possible, large-volume waste air streams with a low solvent content can be concentrated with a continuously operating adsorption wheel (e.g., with special activated carbon) and then treated [578]. The solvent from the waste air is adsorbed on one side of the rotating wheel and desorbed with a small air stream on the other side. The concentrated waste air (10 % of the original waste air stream) can then be purified either by an adsorption unit with solvent recovery, or by afterburning. The heat from the afterburning plant should be utilized; if this is not possible, newly developed thermal methods with internal heat utilization may be used [579].

Biological waste gas treatment methods are also suitable for purifying solvent-containing waste gases, especially slightly contaminated, large-volume waste gas streams; they have already been tested [580]. In biological methods organic substances are degraded by microorganisms on the surface of a wet filter layer or in a scrubber.

In various countries (e.g., the Federal Republic of Germany, the United States, Scandinavia, and Switzerland) regulations exist concerning the treatment of waste air from paint shops. Large paint shops (e.g., in the automobile industry) are covered by these regulations. In the Federal Republic of Germany the maximum permissible emissions from automobile paint shops are limited by the amount of solvent used per square meter of car body (see Low-Solvent Coating Materials.). For automated spraying zones in other paint shops the emission of organic substances in the waste gas is restricted to a maximum of 150 mg/m^3.

12.2. Wastewater

In industrial paint application the principal sources of wastewater are spraying cabins, wet filters, and scrubbers. Further sources of wastewater are the cleaning of apparatus, equipment, vessels, tanks, and working areas, as well as the retentate produced in the ultrafiltration of electrodeposition paints. To reduce environmental pollution, attempts should be made to minimize the amount of wastewater. Wastewater from spraying cabins may be treated by coagulating the overspray and continuously extracting the paint slurry, as well as by reducing the amount of water used. Continuous methods for cleaning the circulating water are used both for solventborne and waterborne paints. Products based on alumina, metal hydroxides, and organic fatty acid derivatives are used as coagulating agents. These auxiliaries coat and envelop the paint particles.

With waterborne paints, stable dispersions or emulsions are sometimes formed in the circulating water. The coagulating agent also has to “break” these disperse systems. A very fine flocculant coagulate is often formed which has to be separated with special filters or a centrifuge, or converted into larger, more easily removable flakes by using a further coagulating agent. If the circulating water from the spray cabin has to be drained off due to high levels of contamination, it must be treated before being discharged into the sewage system or wastewater treatment plant. Treatment usually comprises flocculation, neutralization, and filtration. With certain water-soluble toxic substances (e.g., heavy-metal compounds), organic solvents, and additives, further purification steps may be necessary. Heavy metals can be precipitated. Methods used for organic solvents depend on their nature and concentration in the water; they include ultrafiltration, reverse osmosis, adsorption (e.g., on activated charcoal), biological purification, and, with high solvent concentrations, distillation [581]. The concentration of organic substances in the wastewater is described by the chemical oxygen demand (COD) and the biological oxygen demand within 5 days (BOD_5). Statutory requirements governing the preliminary purification of the wastewater can vary widely. They depend on purification facilities in the existing wastewater treatment plants. The effort and expense involved in wastewater pretreatment can be considerably reduced by avoiding the use of toxic substances (e.g., heavy metals). With waterborne paints particular attention should be paid to adequate removal of water-soluble organic substances (e.g., solvents) from the wastewater.

12.3. Solid Residues and Waste

Considerable amounts of solid residues and other waste are produced when paints are applied, particularly by spraying. The overspray is collected as a coagulated residue from the spray cabin water. Articles such as contaminated filters, paint residues, and empty containers also have to be disposed of. For ecological reasons minimization of waste production and reutilization should take precedence over disposal methods (incineration, landfill). In the Federal Republic of Germany, for example, this principle is laid down in waste control and emission legislation (Abfallgesetz, Bundes-Immissionsschutzgesetz). Up to now paint slurries were mainly disposed of in special landfills. On account of increasingly stringent requirements to prevent pollution of soil and groundwater, paint slurries will have to be disposed of in special refuse incinerators. This will inevitably lead to higher disposal costs; avoidance of waste and recycling will therefore be of economic advantage.

The production of paint residues and waste can be prevented or reduced by using coating methods with high application efficiencies (dipping, brushing, rolling, and pouring). Compared with the spraying technique, these methods generally also result in lower solvent emissions (see Application Methods with Low Solvent Emission.). The amount of overspray produced in spraying methods can be lowered by using electrostatic application procedures. In powder coating the overspray is trapped and separated in a dust-removal filter; it can then be directly recycled, if necessary after purification.

In the spray application of wet paints the overspray can also be recovered and recycled by various methods which have not, however, all been tested industrially [582]. The overspray can be recovered with rotating disks or circulating belts. Stable paints can be reused directly after conditioning (e.g., viscosity adjustment). In solventborne paints the disk or belt often has to be wetted with solvents, the waste air should therefore be treated to prevent high solvent emission. Some waterborne paints can also be recovered by this method. In paint recycling the overspray should not be entrained with the waste air from the spray cabin; certain preconditions should therefore be observed: airless spraying guns are most suitable and the parts to be coated should not be too large or have a complex three-dimensional structure (e.g., car bodies). The overspray can be recovered from the cabin circulation water if it can be coagulated without destroying its chemical structure. After mechanical dewatering with kneaders or mixers, purification, and work-up, small amounts of this material can be added to the new paint. If direct recycling is not possible, the material can be separated by centrifugation into a binder solution and pigment concentrate that can be used as raw materials for paint production. Physically drying paints and stoving finishes are particularly suitable for recovery.

Paint coagulates can also be used as a binder constituent in the production of molded plastics and as a filler replacement in plastics dispersions. If used as a binder for molded plastics, the material must not be cross-linked; the coagulate is worked up into an aqueous dispersion which is used to wet or impregnate fiber mats that are then compressed. In order to produce fillers the paint coagulate first has to be dehydrated and dried, the material becomes completely cross-linked and can be ground into a powder. The powder is used as a filler for plastics dispersions (e.g., for underbody protection in automobiles and in sealing materials).

12.4. Toxicology

Many different substances are used in paints and coating materials as binders, pigments, solvents, and additives. Workers involved in painting and coating work are regularly exposed to volatile organic compounds, especially solvents. In spraying application methods the inhalation of all paint constituents in the form of aerosols should be borne in mind even if they are nonvolatile or of low volatility. Contact with the skin represents a further source of exposure to paint constituents, many of which can be absorbed through the skin.

With manual application by brushing or rolling the health hazards due to solvent exposure (aliphatic and aromatic hydrocarbons, esters, ketones, alcohols, and glycol ethers) are a major factor. Solvents are predominantly absorbed via the respiratory tract. Their toxic effects depend on the nature of the solvent, its concentration, and the length of exposure. Depending on the concentration, symptoms after acute exposure include irradiation of the mucous membranes (eyes and respiratory tract), vertigo, nausea, and vomiting; narcosis symptoms are also observed which are attributed to disturbances of the central nervous system. Chronic poisoning is initially undetectable, but may subsequently produce damage to organs that is specific for the solvent concerned.

The neurotoxic effects found in painters and coaters exposed to solvents are the subject of controversy. Some studies describe subjective symptoms such as fatigue, difficulty in concentrating, and short-term memory problems in workers employed in industrial paint and coatings application. These symptoms have not, however, been observed in painters employed in the architectural and exterior-use paints sectors who mainly use waterborne paints [583].

During surface treatment prior to paint application, abrasive dust and pyrolysis products produced during the removal of paints and solvents may also be inhaled. The dust produced from corrosion protection agents and some older colored paints is often contaminated with heavy metals. Chlorinated hydrocarbons are still used in paint strippers.

Frequent skin contact with paints and coating materials can cause skin disorders, particularly on the hands, in painters and coaters. The lipid-solubilizing properties of the organic solvents may cause or at least promote contact eczema. In particular, paints based on reactive resins (e.g., epoxy and polyester resins) may cause allergic skin disorders. Skin-sensitizing substances include residual monomers and reactive diluents (e.g., acrylates and epoxides) and paint additives (e.g., acid anhydrides, peroxides, amines, as well as cobalt and zirconium in driers, and formaldehyde and isothiazolinone in biocides). Some of these paint constituents are also skin irritants.

In spraying methods often employed for industrial paint application, workers are not only exposed to solvents, they may also inhale paint constituents in the form of aerosols. On account of their very small size, some aerosol components can reach and penetrate the lung virtually unhindered. Substances that have a particularly sensitizing and irritant action on the skin can thus also affect the respiratory tract. Isocyanates can have a sensitizing effect even at very low concentrations (1 µL/m^3) and can cause chronic bronchial asthma in particularly susceptible persons. A literature study carried out by a working group of the International Agency for Research on Cancer came to the conclusion that there is sufficient evidence for carcinogenicity due to the occupational exposure of painters [584]. Occupational exposure in paint manufacture cannot be assessed however.

Depending on the application method (brushing, spraying) and the paint used, technical and personal work safety measures should be adopted when applying paints and coatings. Technical measures include adequate supply of fresh air and removal of waste air (e.g., in special hoods), as well as the replacement of "hazardous" paints with less dangerous ones [585]. Many hazardous dangerous substances have maximum workplace concentrations (threshold limit values) which should be strictly observed and monitored. Adequate facial and skin protection must also be ensured (e.g., with masks and gloves).

[Top of Page]

13. Economic Aspects

Hanns-Adolf Lentze

[586], [587]

Paints and varnishes (coatings) have two primary functions: protection and decoration. Other objectives include information, identification, safety, insulation, vapor barrier, nonskid surface, and control of temperature, light, and dust. A range of product categories with a wide variety of application is therefore available:

1. *Architectural (decorative) coatings* include exterior and interior house paints which are normally distributed through wholesale-retail channels and purchased by the general public, painters, building contractors, government agencies, etc.
2. *Product finishes* are coatings formulated specifically for original equipment manufacture (OEM) to satisfy application conditions and manufacturing requirements for a wide variety of industrial and consumer products, e.g., wood and metal furniture and fixtures; automotive and nonautomotive transportation, aircraft, machinery and equipment, appliances, electrical insulation, film, paper, foil, toys, and sports goods.
3. *Special-purpose coatings* are formulated for special applications or extreme environments and include automotive and machinery refinishing, high-performance maintenance, road markings, marine (bridge) maintenance, crafts, metallic and multicolored coatings.

The number of coatings producers worldwide was estimated at about 10 000 in 1988. The total world coatings market was estimated to be ca. $ 36×10^9; the product market sectors were as follows:

Decorative	52 %
Auto OEM	6 %
Refinishes	5 %
Can coatings	3 %
Others including industrial paints	34 %
Total production	12×10^9 L

World paint markets by region were

North America	32 %
Western Europe	30 %
Japan	13 %
Rest of Asia-Pacific	10.5 %
Rest of world	14.5 %
Total production	12×10^9 L

In 1989 the top ten paint companies accounted for about a third of the total world market, by the year 2000 they could well account for up to half. This development is expected because of permanent streamlining of activities by larger companies through selective acquisitions and/or divestments.

Single sourcing as in the car industry (one supplier for one model), marine sector (direct availability at each ship yard), or canning industry (worldwide health and safety standards) is a key factor in this globalization. Recouping in international markets for expenditure in research and development for technically sophisticated, high added value products is the other reason for this evolution.

Internationalization and the generally high standard of technical products shifts the economic importance from countries to companies. Therefore simple national per capita consumption figures are no longer indicative of productivity and standard of living. This is independent of some standard parameters such as climatic, cultural, or other impacts.

The paint and coatings industry as a whole is considered as a mature industry. An overall growth of 2.5 – 3.0 % is estimated for the 1990s assuming overall growth of the corresponding gross national product of + 3.5 %. This is remarkable in view of the fact that improved techniques such as high-solids coatings and coating powders, reduction in overspray, recovery, and recycling have considerably increased the surface area covered by a given amount of paint.

Raw materials account for roughly half of the production costs; prices of many of them are linked directly or indirectly to the price of crude oil.

In the decorative market products are mainly waterborne and consumption is dominated by new construction work and maintenance. Higher growth rates are therefore expected in newly industrialized and developing economies. In the automotive sector paint supply (not necessarily production) follows the requirements of car producers. Major growth potential for packaging (food and drink cans) lies in developing countries. Industrial paint markets are characterized by replacement of solventborne paints by high-solids, waterborne, and powder coatings to reduce environmental pollution. Therefore, coil coating can also avoid the classical painting process.

[Top of Page]

References

General References

1. S. LeSoto *et al.*: *Paint/Coatings Dictionary*, Fed. of Soc. for Coatings Technology, Philadelphia, Pennsylvania 1978.
2. H. Clausen: *Wörterbuch der Lacktechnologie*, Edition Lack und Chemie, Elvira Moeller GmbH, Filderstadt, 1980.
3. D. H. Solomon: *The Chemistry of Organic Film Formers*, J. Wiley and Sons, New York 1967.
4. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **I – VIII**, Verlag W. A. Colomb, Stuttgart – Berlin 1971 – 1980.
5. S. T. Harris: *The Technology of Powder Coatings*, Portcullis Press Ltd., London 1976.
6. A. V. Patsis (ed.): *Advances in Organic Coatings Science and Technology – Proceedings of the Intern. Conf. in Org. Coatings Sci. and Techn.*, vol**s. 1 – 10**, Technomic Publishing Co., Westport, Connecticut 1979 – Lancaster, Pennsylvania 1988.
7. G. Odian: *Principles of Polymerization*, J. Wiley and Sons, New York 1981.

8. R. A. Dickie, F. L. Floyd: *Polymeric Materials for Corrosion Control*, Am. Chem Soc., Washington, DC 1986.
9. D. R. Randell: *Radiation Curing of Polymers*, The Royal Society of Chemistry, London 1987.
10. E. V. Schmid: *Exterior Durability of Organic Coatings*, FMJ International Publications Ltd., Redhill 1988.
11. J. Bentley in R. Lambourne (ed.): *Paint and Surface Coatings: Theory and Practice*, Ellis Horwood, Chichester 1988, pp. 46 – 52. A. E. Rheineck: "Drying Oil – Modification and Use," in R. R. Myers, J. S. Long (eds.): *Treatise on Coatings,* vol. 1, part II, Marcel Dekker, New York 1969.
12. Wolff Walsrode, EP 0 076 443, 1986 (E. Lühmann, L. Hoppe, K. Szablikowski); EP 0 154 241, 1989 (E. Lühmann, L. Hoppe, K. Szablikowski, K.-F. Lampert).
13. DIN 53 179.
14. A. Kraus: *Handbuch der Nitrocelluloselacke*, vol. **2**, Westliche Berliner Verlagsgesellschaft Hunemann KG, Berlin 1963, p. 248 ff.
15. L. G. Curtis, J. D. Crowley: *Applied Polymer Science*, 2nd ed., American Chemical Society, Washington, DC p. 1053.
16. C. J. Malm, G. H. Hiatt: *Cellulose and Cellulose Derivatives*, 2nd ed., Wiley Interscience, New York 1955; vol. 5, part 2, p. 798.
17. M. Salo, *Official Digest*, 31 Sept. 1959.
18. Bayer AG: Bayer Produkte für die Lackindustrie, 1967.
19. K. Hoehne, Lt. Nordt: "Neues über chlorierte Polymere" *Farbe + Lack* **76** (1970) no. 2, 123 – 126.
20. B. Dodson, I. C. McNeill: "Thermal Degradation of Chlorinated Rubber, Evidence for an Alternative Cyclic Structure for Chlorinated Rubber," *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2305 – 2315.
21. R. Lapasin, G. Torriano, S. Volpe: *Ind. Vernice* **29** (1975) no. 4, 2 – 19. (Thick-film paints based on chlorinated rubber and inert pigments – special aspects of anticorrosive properties and the film-forming process)
22. G. C. Reid: "Chlorinated Rubber, the Solution to Many Corrosion Control Problems," *Anti-Corros. Methods Mater.* **22** (1975) no. 12, 8 – 13.
23. K. Hoehne: "Zusammensetzung und Eigenschaften von Chlorkautschuk-Dickschichtanstrichen," *Farbe + Lack* **77** (1971) no. 9, 866 – 869.
24. M. Yamabe, H. Higaki, G. Kojima in G. D. Parfitt, A. V. Patsis (eds.): *Organic Coatings Science and Technology*, vol. **7**, Marcel Dekker, New York – Basel 1984, p. 25.
25. V. Handforth, *J. Oil. Colour Chem. Assoc.* **73** (1990) no. 4, 145 – 148.
26. Pennwalt Chemical Co., Technical Data, M. Kevin: *Building Design & Construction*, March 1983, p. 2.
27. L. Stonberg, *Conf. Proc. Alum. Finish* **87** (1986) 26.1 – 26.17.
28. J. E. McCann, *Surf. Coat. Aust.* **27** (1990) no. 1 – 2, 8 – 12.
29. K. V. Summer, *Conf. Proc. Alum. Finish* **87** (1986) 36.1 – 36.19.
30. Nippon Oil and Fats, JP-Kokai 86 238 863 A2, 1986 (A. Hikita *et al.*).
31. Pennwalt Chemical Co., Technical Data, Philadelphia, June 1984.
32. Du Pont, US 4 011 361, 1975 (E. Vassiliou *et al.*).
33. G. Kojima *et al.* in A. V. Patsis (ed.): *11th International Conference in Organic Science and Technology*, vol. **9**, Technomic Publishing Company, Lancaster, Pa. 1987, pp. 120 – 128.
34. Du Pont, GB 1 064 840, 1967 (J. C. S. Fang).
35. Du Pont, US 3 318 850, 1967 (F. B. Stilmar).
36. L. Harold *et al.*, *157th ACS Meeting, Minneapolis* **29** (1969) April, 221.
37. Du Pont, US 2 468 664, 1949 (E. William *et al.*).
38. Du Pont, US 3 651 003, 1972 (M. F. Bechtold *et al.*).
39. S. Munekata *et al.*, *Report of Research Laboratory, Asahi Glass Co.* **34** (1984) no. 2, 205.
40. N. Miyazaki, T. Takayanagi, *Report of Research Laboratory, Asahi Glass Co* **36** (1986) no. 1, 155.
41. S. Munekata, *Prog. Org. Coat.* **16** (1988) 113 – 134.
42. Dainippon Ink and Chemicals, JP-Kokai 84 102 962, 1984 (M. Ooka *et al.*).
43. Central Glass Co., JP-Kokai 86 57 609, 1986 (T. Koishi *et al.*).
44. Mitsui Petrochemical Industries, JP-Kokai 86 141 713, 1986 (S. Honma *et al.*).
45. Daikin Industries, JP-Kokai 84 189 108, 1984 (A. Oomori *et al.*).
46. N. Miyazaki, M. Kamba, *First Pacific Polymer Conference*, Hawaii, Dec. 1989, p. 557.
47. J. R. Griffith, *CHEMTECH* **12** (1982) 290.
48. S. J. Shaw *et al.*, *Polym. Mater. Sci. Eng.* **56** (1987) 209.
49. J. R. Griffith *et al.*, *Ind. Eng. Chem. Prod. Res. Dev.* **25** (1986) no. 4, 572.
50. W. S. Zimmt, *CHEMTECH* **11** (1981) 681.
51. M. K. Yousuf, *Mod. Paint Coat.* **79** (1989) 48.
52. J. L. Gerlock, D. R. Bauer, L. M. Briggs, *Prog. Org. Coat.* **15** (1987) 197.
53. A. J. Tortorello, M. A. Kinsella, *J. Coat. Technol.* **55** (1983) 99.
54. W. Brushwell, *Farbe + Lack* **86** (1980) 706.
55. R. Zimmermann, *Farbe + Lack* **82** (1976) 383.
56. A. Mercurio, S. N. Lewis, *J. Paint Technol.* **47** (1975) 37.
57. C. K. Schoff, *Prog. Org. Coat.* **4** (1976) 189. Links
58. M. Takahashi, *Polym. Plast. Technol. Eng.* **15** (1980) 1.
59. K. O'Hara, *J. Oil. Colour Chem. Assoc.* **71** (1988) 413.
60. H. Kittel, *Adhäsion* **21** (1977) 162.
61. D. W. Aldinger: *Congr. FATIPEC I*, 1984, p. 69.
62. J. P. H. Juffermans, *Prog. Org. Coat.* **17** (1989) 15.
63. D. R. Bauer, R. A. Dickie, *J. Coat. Technol.* **58** (1986) 41.
64. S. Günther, *Ind. Lackierbetr.* **57** (1989) 167.
65. J. R. Erickson, *J. Coat. Technol.* **48** (1976) 58.
66. R. Dowbenko, D. P. Hart, *Ind. Eng. Chem. Prod. Res. Dev.* **12** (1973) 14.
67. G. Y. Tilak, *Prog. Org. Coat.* **13** (1985) 333.
68. L. W. Hill, A. Kaul, K. Kozlowski, J. O. Sauter, *Polym. Mater. Sci. Eng.* **59** (1988) 283.
69. T. Nakamichi, M. Ishidoya, *J. Coat. Technol.* **60** (1988) 33.
70. H. Sander, R. Kroker, *Farbe + Lack* **82** (1976) 1105.
71. R. D. Athey, *Farbe + Lack* **95** (1989) 475.
72. A. Noomen, *Prog. Org. Coat.* **17** (1989) 27.
73. E. Karsten, O. Lückert: *Lackrohstoff-Tabellen*, Curt R. Vincentz Verlag, Hannover 1987. *Resin Index*, British Resin Manufacturer's Association, Redhill, Surrey 1986. *Raw Materials Index-Resins*, National Paint and Coatings Association, Washington, DC 1988.
74. Lewis Berger & Sons, EP 573 809, 1942; 573 835, 1943; 580 912, 1942; DE 975 352, 1949.
75. Röhm und Haas, DE 1 022 381, 1953.
76. Röhm und Haas, DE 912 752, 1951; F. Schlenker, *Farbe + Lack* **58** (1952) 174.
77. W. Krauß, *Farbe + Lack* **64** (1958) 39; **64** (1958) 209; **70** (1964) 876.
78. W. Götze, *Fette Seifen Anstrichm.* **65** (1963) 493.

79. W. B. Winkler, US 2 663 649, 1953.
80. I. G. Farben, DE 738 254, 1940.
81. Shell International Research, EP 942 465, 1963 (Dutch prior. 1959 and 1960).
82. Shell International Research, EP 1 269 628, 1960.
83. K. Hamann, *Farbe + Lack* **81** (1975) 907.
84. R. Küchenmeister, *Farbe + Lack* **78** (1972) 550; E. Knappe, *Fette Seifen Anstrichm.* **77** (1975) 149; J. Schoeps, *J. Oberflächentechnik* **11** (1975) 32.
85. J. E. Glass (ed.): *Water Soluble Polymers – Beauty with Performance*, American Chemical Society, Washington, D.C. 1986.
86. Akzo, DE-OS 3803 141 A 1, 1989.
87. Coates Brothers, EP-A 0 253 474 A 2, 1988 (Brit. prior. 1986, GB 86 12 193).
88. K. Holmberg: *High Solids Alkyd Resins*, Marcel Dekker, New York NY.
89. S. J. Storfer, J. T. Di Piazza, R. E. Moran, *J. Coat. Technol.* **60** (1988) no. 761, 37.
90. J. Spauwen, *J. Oil Colour Chem. Assoc.* **71** (1988) no. 2, 47.
91. S. J. Bellettiere, D. M. Mahoney, *J. Coat. Technol.* **59** (1987) no. 752, 101.
92. *Ullmann*, 4th ed., **19**, 61.
93. R. Seidler, H. G. Stolzenbach, *Farbe + Lack* **81** (1975) 281.
94. V. Mirgel, K. Nachtkamp, *Farbe + Lack* **89** (1983) 928.
95. W. Marquardt, H. Gempeler, *Farbe + Lack* **84** (1978) 301.
96. J. Dörffel, *Farbe + Lack* **81** (1975) 10.
97. E. Häring, O. Lückert, in: *Emissionsarm lackieren*, Curt. R. Vincentz Verlag, Hannover 1987, p. 51.
98. K. Brüning, K. G. Sturm, *Adhäsion* **3** (1973) 83.
99. Dynamit Nobel AG, DE 2 126 048, 1971 (G. Schade, H. Hermsdorf, H. Vltavsky).
100. Hüls AG, DE 2 521 791, 1975 (K. Burzin, K.-H. Haneklaus, D. Stoye).
101. K. H. Hornung, U. Biethan, *Farbe + Lack* **76** (1970) 461.
102. U. Biethan, J. Dörffel, D. Stoye, *Farbe + Lack* **77** (1971) 988.
103. Hüls AG, DE 2 346 818, 1973 (K. Schmitt, J. Disteldorf, F. Schmitt); DE 2 938 855, 1979 (R. Gras, E. Wolf).
104. M. Schmitthenner: *XIV. Internat. Conference in Org. Coatings Sc. and Techn.*, Athens, 1988.
105. Code of Federal Regulations (annual publication), § 175.300, Chap. (b) (3) (vii), “Polyester resins”.
106. D. Stoye, W. Andrejewski, J. Dörffel: *Congr. FATIPEC XIII*, 1976, p. 605.
107. Hüls AG, EP 45 994, 1981 (J. Disteldorf, R. Gras, H. Schnurbusch).
108. S. Tochihara, *Prog. Org. Coat.* **10** (1982) 195.
109. Bayer, DE 2 804 216, 1978 (F. Frank, O. Bendszus, H. J. Freier, H. J. Traenckner).
110. H. Wagner, H. F. Sarx: *Lackkunstharze*, 5th ed., Carl Hanser Verlag, München 1971, p. 129.
111. I. G. Farben Ind., DE 540 101, 1930; DE 544 326, 1930; DE 571 665, 1930; DE 598 732, 1932.
112. E. Karsten, O. Lückert: *Lackrohstoff-Tabellen*, 8th ed., C. R. Vincentz Verlag, Hannover 1987, pp. 466 – 479 (“Resin Index,” British Resin Manufacturers Assoc.; “Raw Materials Index-Resins,” National Paint and Coatings Assoc., USA).
113. BASF, DE 948 816, 1951. Scott-Bader, EP 713 312, 1951.
114. W. Gebhardt, W. Herrmann, K. Hamann, *Farbe + Lack* **64** (1958) 303.
115. Chem. Werke Hüls, EP 842 958, 1960.
116. ICI, EP 578 867, 1944.
117. Bayer, DE-AS 1 024 654, 1955 (W. Biedermann, K. Raichle).
118. Bayer, DE 1 494 437, 1956 (K. Raichle).
119. Reichhold Chemie, DE 1 028 333, 1956.
120. H. G. Heine, H.-J. Traenckner, *Prog. Org. Coat.* **3** (1975) 115 – 139. Links
121. J. Ohngemach, K. H. Neisius, J. Eichler, C. P. Herz, *Kontakte (Darmstadt)* **3** (1980) 15 – 20.
122. M. Jakobi, A. Henne, A. Böttcher, *Adhäsion* **4** (1986) no. 4, 6 – 10.
123. K. Demmler, E. Müller, M. Schwarz, *DEFAZET Dtsch. Farben Z.* **31** (1977) 115 – 128.
124. Herberts GmbH, DE-AS 1 025 302, 1954 (K. H. Hauck, F. Hecker-Over).
125. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **5**, Verlag W. A. Colomb, Berlin 1977, pp. 311 – 315.
126. G. Oertel: *Polyurethane Handbook*, Hanser Publishers, New York 1985.
127. M. Bock, R. Halpaap, *Farbe + Lack* **93** (1987) 264.
128. W. Wieczorrek, *Polym. Paint Colour J.* **178** (1988) no. 4225, 827.
129. B. Taub, G. H. Patschke, *Modern Paints Coat.* **79** (1989) 41.
130. H. Lee, K. Neville: *Handbook of Epoxy Resins*, McGraw Hill, New York 1982.
131. A. M. Paquin: *Epoxidverbindungen und Epoxidharze*, Springer Verlag, Berlin-Göttingen-Heidelberg 1958.
132. H. Wagner, H. F. Sarx: *Lackkunstharze*, Hanser Verlag, München 1971.
133. H. Batzer, F. Lohse, *Kunststoffe* **66** (1976) no. 10, 637.
134. W. Marquardt, H. Gempeler, *Farbe + Lack* **84** (1978) 301 – 306.
135. DSM, DE 4 340 974 C2.
136. W. Marquardt, H. Gempeler, *Farbe + Lack* **86** (1980) 696 – 698.
137. W. Schneider, *Farbe + Lack* **85** (1979) 925 – 929.
138. F. Lohse, H. Zweifel, *Adv. Polym. Sci.* **78** (1986) 61 – 81. Links
139. M. Gaschke, B. Dreher, *J. Coat. Technol.* **48** (1976) no. 617, 46 – 51.
140. K. Brugger, R. Schmid: *Congr. FATIPEC XI*, 1972, p. 307.
141. J. Bersier, P. M. Bersier, B. Dreher, A. Heer: *Congr. FATIPEC X*, 1970, p. 255.
142. K. Buser, *Plaste Kautsch.* **3** (1963) 189.
143. A. Manz, *Beckacite-Nachr.* **25** (1966) no. 3, 70.
144. W. J. van Westrenen *et al.*: *Congr. FATIPEC VIII*, 1966, p. 126.
145. W. Weger: *Congr. FATIPEC XVII*, 1986, p. 43.
146. Ciba: “Occupational Hygiene Advices for Manufaturing and Processing Plastics,” company brochure.
147. W. A. Finzel, E. P. Plueddemann: “The History of Silicone Coatings,” *ACS Meeting*, Miami Beach, Fla., September 13, 1989.
148. L. H. Brown in R. R. Myers, J. S. Long (eds.): *Treatise on Coatings,* vol. 1, part III, chap. 3, “Silicones in Protective Coatings,” Marcel Dekker, New York, 1972.
149. S. Paul: *Surface Coatings*, "chap. 2.3.", J. Wiley & Sons, Chichester 1985.
150. T. C. Kendrick, B. Parbhoo, J. W. White in S. Patai, Z. Rappoport (eds.): *The Chemistry of Organic Silicon Compounds,* part 2, chap. 21, “Siloxane Polymers and Copolymers,” J. Wiley & Sons, Chichester 1989.
151. N. Grassie, I. G. Macfarlane, *Eur. Polym. J.* **14** (1978) no. 11, 875. Links
152. W. A. Finzel: “Properties of High Temperature Silicone Coatings,” *J. Protective Coatings Linings* **4** (1987) no. 8, 38 – 43.

153. Dow Corning Corp.: *Silicones for Protective Coatings*, 48686–0994, Midland, Mich., 1985.
154. W. A. Finzel: “Silicone Coatings: Popular Choice for the Tough Jobs,” *Am. Paint Coat. J.* **65** (1981) 18.
155. Dow Corning, EP 279 513, 1988 (A. G. Short).
156. R. W. Cope, M. A. Glaser: *Silicone Resins for Organic Coatings*, Federation of Society for Paint Technology, Philadelphia 1970.
157. L. H. Brown: *Handbook of Coating Additives*, Marcel Dekker, New York 1987.
158. Dow Corning Europe: *Silicone Additives: That Little Bit Extra*, **48686–0994**, 1988.
159. G. C. Sawicki, J. W. White: “Speciality Silicones for Foam Control,” *Chemspec Europe 89 BACS Symposium*, Manchester 1989.
160. N. J. Albrecht, W. J. Blank: “The Use of Triazine Resins in High Solids Coatings,” *Adv. Org. Coat. Sci. Technol. Ser.* **4** (1982).
161. W. J. Blank: “Amino Resins in High Solids Coatings,” *J. Coat. Technol.* **54** (1982) no. 687, 26 – 41.
162. W. J. Blank: “Reaction Mechanism of Melamine Resins,” *J. Coat. Technol.* **51** (1979) no. 656, 61 – 70.
163. A. Berge, S. Gudmunden, J. Ugelstad, *Eur. Polym. J.* **5** (1969) 171. Links
164. A. Berge, B. Kvaeven, J. Ugelstad, *Eur. Polym. J.* **6** (1970) 981. Links
165. L. J. Calbo: “Effect of Catalyst Structure on the Properties of Coatings Crosslinked with Hexa(methoxymethyl)melamine,” *J. Coat. Technol.* **52** (1980) no. 660, 75 – 83.
166. W. Schedlbauer: “Melamin-Formaldehyd-Lackharze,” *Farbe + Lack* **79** (1973) no. 9, 846 – 852.
167. W. F. Herbes, S. J. O'Brien, R. G. Weyker: “Melamine-Formaldehyde,” *Chemical Aftertreatment of Textiles*, Wiley Interscience, New York 1971.
168. G. G. Parekh: “Chemistry of Glycoluril-Formaldehyde Resins and their Performance in Coatings,” *J. Coat. Technol.* **51** (1979) no. 658, 101 – 110.
169. G. Christensen: “Analysis of Functional Groups in Amino Resins,” *Prog. Org. Coat.* **5** (1977) 255 –276.
170. H. Schindlbauer, J. Anderer: “Eine einfache Charakterisierung von Melamin-Formaldehyd-Kondensaten mit Hilfe von NMR-Messungen,” *Fresenius Z. Anal. Chem.* **301** (1980) 210 – 214.
171. M. Chiavarini, N. Del Fanti, R. Bigatto: “Compositive Characterization of Urea-Formaldehyde Adhesives by NMR Spectroscopy,” *Angew. Makromol. Chem.* **46** (1975) no. 695, 151 – 162. Links
172. M. Chiavarini, N. Del Fanti, R. Bigatto: “Compositive Characterization of Melamine-Formaldehyde Condensates by NMR Spectroscopy,” *Angew. Makromol. Chem.* **56** (1976) no. 818, 15 –25. Links
173. B. Tomita, H. Ono: “Melamine-Formaldehyde Resins: Constitutional Characterization by Fourier Transform 13C-NMR,” *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 3205 – 3215.
174. A. J. J. de Breet, W. Dankelman, W. G. B. Huysmans, J. De Wit: “^{13}C-NMR Analysis of Formaldehyde Resins,” *Angew. Makromol. Chem.* **62** (1977) no. 877, 7 – 21. Links
175. Technical literature of Monsanto Chemical Company.
176. Estimates of market size were obtained from private industry sources. Published information on amino resin usage bundles other end uses and does not permit any reasonable estimates of market size.
177. S. E. Feinman: *Formaldehyde Sensitivity and Toxicity*, CRC Press, Boca Raton, Fla. 1988.
178. K. Hultzsch: *Chemie der Phenolharze*, Springer Verlag, Berlin 1950.
179. Robert W. Martin: *The Chemistry of Phenolic Resins*, J. Wiley & Sons, New York 1956.
180. P. Nylén, E. Sunderland: *Modern Surface Coatings*, J. Wiley & Sons, London 1965.
181. H. Abraham: *Asphalts and Allied Substances*, D. van Nostrand Comp., New Jersey 1960.
182. F. M. Depke: “Lacktechnische Schriften,” in vol. 8: *Bitumen- und Teerlacke,* Verlag W. A. Colomb, Stuttgart – Berlin 1970.
183. H. Walther: *Bituminöse Stoffe im Bauwesen, Straßenbau*, Chemie und Technik Verlag, Heidelberg 1962.
184. H.-G. Franck, U. G. Collin: *Steinkohlenteer, Chemie, Technologie und Verwendung*, Springer Verlag, Berlin – Heidelberg – New York 1968.
185. L. Bernhard: *Das Wasserglas, seine Darstellung und Anwendung*, H. Bechold, Frankfurt 1893.
186. H. Trillich: *Die Wasserglas-Anstrich- und Mal-Verfahren*, B. Heller, München 1929.
187. A. W. Keim, DE 4315, 1878.
188. R. Wenda, W. Lukas, *Farbe + Lack* **93** (1987) 186 – 189.
189. H. Wulf: *Große Farbwarenkunde*, R. Müller, Köln 1967.
190. E. Bagda, *Farbe + Lack* **90** (1984) 443.
191. Tiedemann (ed.): *Grundlagen zur Formulierung von Dispersions-Silikat-Systemen*, Viernheim 1988.
192. Technical Service Report on Trasol: *Production of Trasol Paints, Plasters and Primers*, Henkel KGaA, Düsseldorf 1986.
193. W. Haller, *Adv. in Org. Coatings*, in preparation.
194. J. W. Lyons: *The Chemistry and Uses of Fire Retardants*, Wiley-Interscience, New York 1970.
195. A. Williams: *Flame Retardant Coatings and Building Materials*, Chemical Technology Review 25, Noyes Data, Park Ridge 1974.
196. T. Ginsberg, *J. Coat. Technol.* **53** (1981) no. 677, 23 – 32.
197. T. Ginsberg, L. G. Kaufmann, *Mod*, Paint Coat. (1981) October, 138 – 143.
198. W. C. La Course *et al.*, *J. Can. Ceram. Soc.* **52** (1983) 18 – 23.
199. T. Ginsberg, C. N. Merriam, L. M. Robeson, *J. Oil Colour Chem. Assoc.* **59** (1976) 315 – 321.
200. J. Loomans, K. van Leberghe, *Chim. Peint.* **12** (1949) 184 – 190.
201. H. Anders, *DEFAZET Dtsch. Farben Z.* **8** (1954) 452 – 453.
202. Carboline, US 3 056 684, 1959 (S. L. Lopata, W. R. Keithler).
203. Zinc Lock Corp., US 3 392 130, 1962 (R. A. Rucker, J. B. Heymes).
204. Hempel's Marine Paints, US 3 442 824, US 3 546 155, 1967 (J. W. Chandler).
205. Stauffer Chem. Corp., US 3 392 036, 1965 (G. D. McLeod).
206. Dynamit Nobel, DE 2 000 199, 1970 (W. Dittrich).
207. Stauffer Chem. Corp., US 3 730 743, 1970 (G. D. McLeod).
208. The Anderson Development Co., DE-OS 2 147 865, 1971 (A. R. Anderson).
209. D. H. Gelfer, P. Vandorsten, *Mater Perform.* **20** (1981) no. 5, 23 – 28.
210. *Ullmann*, 4th ed., **15**, 652.
211. Verband der deutschen Lackindustrie (VdL), information 9/90, 17.04.1990.
212. R. B. Seymour, *J. Coat. Technol.* **61** (1989) no. 776, 73.
213. J. Hambrecht, *Coat. Mag.* **12** (1990) 44.
214. ASTM D 3960–87, Determining Volatile Organic Content (VOC) of Paints and Related Coatings.
215. K. Tachi, C. Okuda, S. Suzuki, *J. Coat. Technol.* **62** (1990) no. 782, 43.
216. H. J. Drexler, B. Biallas, *I-Lack* **58** (1990) 208.
217. J. Ramsbotham, *Prog. Org. Coat.* **8** (1980) 113.
218. R. C. Nelson, V. T. Figurelli, J. G. Walsham, G. D. Edwards, *J. Paint. Technol.* **42** (1970) 644.
219. C. M. Hansen, R. W. Tess (eds.): “Solvents Theory and Practice,” *Adv. Chem. Ser.* **124** (1973) 1 ff.
220. D. W. Richards, *Am. Paint Coat. J.* **73** (1988) 126.
221. J. Ramsbotham, *J. Oil Colour Chem. Assoc.* **26** (1979) 359.
222. R. A. Waggoner, F. D. Blum, *J. Coat. Technol.* **61** (1989) no. 768, 51.
223. M. Shimbo, M. Ochi, K. Arai, *J. Coat. Technol.* **57** (1985) no. 728, 93.

224. R. Henning, *Proc. Polyurethane World Congr.* 1987, 626.
225. B. Henshaw, *Surf. Coat. Aust.* **25** (1988) 12.
226. M. Kashimoto, *JSAE Rev.* **11** (1990) 88.
227. V. Handforth, *J. Oil Colour Chem. Assoc.* **73** (1990) 145.
228. J. A. Simms, H. J. Spinelli, *J. Coat. Technol.* **59**, (1987) no. 752, 125.
229. A. Noomen, *Proc. Int. Conf. Org. Coat. Sci. Technol. 13th* 1989, 251.
230. H. R. Lucas, *J. Coat. Technol.* **57** (1985) no. 731, 49.
231. PPG, US 4 650 718, 1987 (D. Simpson *et al.*).
232. ICI, EP 0 134 691, 1984 (S. M. Horley *et al.*).
233. U. Zorll, *Farbe + Lack* **95** (1989) 801.
234. M. Schmitthenner, W. Funke, *Proc. 2nd Solvents Symp. "Solvents: The Neglected Parameter*," University of Manchester 1977, 101.
235. M. J. Diana, *SAE Tech. Pap. Ser.* (1991) 910089.
236. H. Scheene, T. Heere, *Chem. Ing. Tech.* **61** (1989) 330.
237. K. Holmberg: *High Solids Alkyd Resins*, Marcel Dekker, New York 1987.
238. Du Pont, US 4 224 202, 1980 (P. Heiberger).
239. Rohm and Haas, US 4 261 872, 1981 (W. D. Emmons).
240. Bayer, EP 057 834, 1982 (E. Eimers, R. Dhein, K. Fuhr).
241. Bayer, EP 032 657, 1981 (E. Eimers, R. Dhein).
242. D. J. Love, *Paint and Resin* **59** (1989) 19 – 22.
243. J. P. H. Juffermans, *Prog. Org. Coat.* **17** (1989) 15 –26.
244. Akzo N.V., EP-A 234 641, 1987 (D. J. F. Geerdes, P. J. C. Nelen, K. H. Zabel).
245. T. A. Potter, *Polym. Mater. Sci. Eng.* **55** (1986) 425 –433.
246. W. J. Blank, *J. Coat. Techn.* **60** (1988) no. 764, 43 –50.
247. W. Heitz, *Angew. Chem.* **145/146** (1986) 37 – 68. Links
248. K. Hönel, *Farbe + Lack* **59** (1953) 174.
249. K. Weigel: *Elektrophorese-Lacke*, Wissenschaftl. Verlagsgesellschaft, Stuttgart 1967.
250. H. J. Luthardt, *Congr. FATIPEC XIV*, 1978, p. 413.
251. W. Burckhardt, H. J. Luthardt, *J. Oil Colour Chem. Assoc.* **62** (1979) 375.
252. EP-B 0 015 035, 1979 (G. Günter, H. Haeufler).
253. EP-B 0 038 127, 1980 (A. J. Backhouse).
254. EP-A 0 238 108, 1986; EP-A 0 273 530, 1986; EP-A 0 287 144, 1987 (R. Buter).
255. B. D. Meyer in: *Umweltfreundliche Lackiersysteme für die industrielle Lackierung,* Expert-Verlag, Kontakt und Studium, vol. 271, Ehningen 1989, pp. 146 – 244.
256. *Farbe + Lack* **95** (1989) no. 9, 671.
257. *VDI-Handbuch Kunststofftechnik*, VDI 2538, Sept. 1979, VDI Düsseldorf.
258. *VDI-Handbuch Reinhaltung der Luft*, vol. **3**, Entwurf VDI 2588, May, 1986, VDI Düsseldorf.
259. B. D. Meyer, *Farbe + Lack* **89** (1983) no. 10, 763 –771.
260. The Powder Coating Institute; *Powder Coating 88,* Alexandria, Virginia 1988.
261. H. Dörr, F. Holzinger: *Kronos Titandioxid in Dispersionsfarben*, Kronos Titan, Leverkusen 1989.
262. E. Nägele: *Dispersionsbaustoffe*, Rudolf Müller Verlag, 1989.
263. K. E. J. Barrett (ed.): *Dispersion Polymerization in Organic Media*, J. Wiley & Sons, London 1975.
264. Chap. 2 in [263]
265. R. H. Ottewill, T. Walker, *Kolloid Z. Z. Polym.* **227** (1968) 108.
266. A. Doroszkowski, R. Lambourne, *J. Poly. Sci. Part C* **34** (1971) 253.
267. F. Th. Hesselink, *J. Phys. Chem.* **73** (1969) 3488.
268. D. H. Napper, *J. Colloid Interface Sci.* **29** (1969), 168. Links
269. Chap. 3 in [263]
270. M. W. Thompson in R. Buscall, T. Corner and J. Stagemen (eds.): *Polymer Colloids*, J. F. Elsevier, London 1985.
271. ICI, GB 893 429, 1957; GB 990 144, 1960; GB 1 017 931, 1966 (D. W. J. Osmond).
272. ICI, GB 1 052 241, 1962 (D. W. J. Osmond).
273. F. Bueche, *J. Colloid. Interface Sci.* **41** (1972) 374. Links
274. F. A. Waite, *J. Oil Colour Chem. Assoc.* **54** (1971) 342.
275. ICI, GB 1 096 912, 1963 (M. W. Thompson, F. A. Waite).
276. ICI, GB 1 174 391, 1966 (D. W. J. Osmond, F. A. Waite, D. J. Walbridge).
277. J. V. Dawkins, S. A. Shakir, T. G. Croucher, *Eur. Polym. J.* **23** (1987) no. 2, 173. Links
278. J. V. Dawkins, G. Taylor in R. M. Fitch (ed.): *Polymer Colloids II*, Plenum Press, New York 1980, p. 447.
279. ICI, GB 1 373 531, 1971 (P. F. Nicks, P. G. Osborne).
280. BALM Paints, GB 1 231 614, 1967.
281. S. J. Barsted, L. J. Novakowska, I. Wagstaff, D. J Walbridge, *Trans. Faraday Soc.* **67** (1971) 3598.
282. J. V. Dawkins, G. Taylor, *J. Chem. Soc. Faraday Trans. 1* **76** (1986) 1263.
283. D. J. Cebula *et al.*, *Colloid Polym. Sci.* **261** (1983) no. 7, 555. Links
284. R. J. R. Cairns, R. H. Ottewill, D. W. J. Osmond, I. Wagstaff, *J. Colloid Interface Sci.* **54** (1976) 45. Links
285. D. H. Napper, *Trans. Faraday Soc.* **64** (1968) 1701.
286. B. Vincent, J. Clarke, K. G. Barnett, *Colloids Surf.* **17** (1986) no. 1, 51.
287. G. J. Fleer, J. H. M. H. Scheutjens, B. Vincent, *ACS Symp. Ser.* **240** (1984) 245.
288. J. V. Dawkins *et al.*, *ACS Symp. Ser.* **165** (1981) 189.
289. J. S. Higgins *et al.*, *Polym. Commun.* **29** (1988) no. 5, 122.
290. K. E. J. Barrett, H. R. Thomas, *J. Polym. Sci. Part A* **7** (1969) 2621.
291. K. E. J. Barrett, H. R. Thomas, *Br. Polym. J.* **2** (1970) 45.
292. K. E. J. Barrett, H. R. Thomas, *Kinet. Mech. Polyreactions Int. Symp. Macromol. Chem. Prep.* **3** (1969) 375.
293. G. Chin, M. A. Winnik, M. D. Croucher, *Colloid Polym. Sci.* **264** (1986) 25.
294. M. D. Croucher, L. S. Egan, M. A. Winnik, *Polym. Eng. Sci.* **26** (1986) 15.
295. T. A. Strivens, *Colloids Surf.* **18** (1986) no. 2 – 4, 395.
296. ICI, GB 1 157 630, 1965 (M. W. Thompson, D. J. Walbridge).
297. R. Dowbenko, D. P. Hart, *Ind. Eng. Chem. Prod. Res. Dev.* **12** (1973) 14.
298. A. J. Backhouse, *J. Coat. Technol.* **54 (693)** (1982) 83.
299. ICI, US 4 180 489, 1979 (M. S. Andrew, A. J. Backhouse).

300. ICI, US 4 242 384, 1980 (M. S. Andrew, A. J. Backhouse).
301. ICI, GB 1 242 054, 1967 (N. D. P. Smith, M. W. Thompson, E. J. West).
302. C. B. Arends, D. K. Hoffman, G. C. Kolb, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **26** (1985) no. 1, 232.
303. Rohm & Haas, GB 956 454, 1959.
304. Rohm & Haas, GB 993 794, 1961.
305. Union Carbide, GB 1 199 651, 1966.
306. Du Pont, GB 1 272 890, 1969.
307. ICI, DE 2 139 315, 1970.
308. BASF, GB 1 203 427, 1966.
309. Hercules Inc., BE 877 607, 1980.
310. Nippon Oils and Fats, GB 2 071 116, 1981.
311. Xerox Corp., US 4 617 249, 1985.
312. C. K. Ober, K. P. Lok, *Macromolecules* **20** (1987) no. 2, 268. Links
313. Mobil Oil Corp., EP-A 290 195 A1, 1988.
314. Allied Colloids Ltd., EP-A 150 933 A2, 1985.
315. G. C. Burnside, G. E. P. Brewer, *J. Paint Technol.* **38** (1960) 96.
316. L. R. LeBras, R. M. Christensen, *J. Paint Technol.* **44** (1972) 63.
317. Verband der Lackindustrie e. V., Mitteilung 9/90, April 17th 1990.
318. A. D. Wilson, J. W. Nicholson, H. J. Posser (eds.): *Surface Coatings – 2*, Elsevier Applied Science Publishers, Barling 1988, p. 39.
319. M. S. ElAasser *et al.*, *J. Coat. Technol.* **56** (1984) 37.
320. V. I. Eliseeva, V. N. Cherny, *Prog. Org. Coat.* **17** (1989) 251.
321. F. Beck, *Prog. Org. Coat.* **4** (1976) 1. Links
322. H. Spoor, H. U. Schenck, *Farbe + Lack* **88** (1982) 94.
323. P. E. Pierce, J. Kovacs, C. Higgenbothan, *Ind. Eng. Chem.* **17** (1978) 317.
324. P. E. Pierce, *Ind. Coat. Techn.* **53** (1981) 52.
325. W. Rausch, *Ind. Lackierbetrieb* **49** (1981) 413.
326. W. Woebcken: *Kunststoffhandbuch, Duroplaste*, Hanser Verlag, München – Wien 1988, p. 1023.
327. ICI, GB 1 109 979, 1964 (B. A. Cooke).
328. PPG Ind., FR 1 394 007, 1962 (D. P. Hart, R. H. Christensen).
329. H. Verdin, *J. Oil Colour Chem. Assoc.* **59** (1986) 81.
330. P. I. Kordomenos, J. D. Nordstrom, *J. Coat. Technol.* **54** (1982) 33.
331. BASF, EP 125 438, 1983 (E. Schupp *et al.*).
332. Z. W. Wicks, *Prog. Org. Coat.* **3** (1985) 73.
333. PPG, US 4 423 166, 1982 (T. Moriarity, W. Geiger).
334. H. Kraus, *Metalloberfläche* **36** (1982) 531.
335. H. J. Streitberger, K. Arlt, F. J. Pötter, *Fahrzeug und Karosserie* **7** (1984) 34.
336. B. A. Baker, *Light Metal Age* **41** (1983) 20.
337. Th. Brücken, *Canmaker* **2** (1989) 31.
338. Mutsutseshi Denhi, Kansai Paint, DE 3 628 340 (M. Hoshino *et al.*).
339. W. Büchner, R. Schiebs, G. Winter, K. H. Büchel: *Industrielle anorganische Chemie*, Verlag Chemie, Weinheim, Germany 1984.
340. G. Buxbaum *et al.*: "Anorganische Pigmente," *Winnacker-Küchler, Chemische Technologie*, 4th. ed., vol. **3**, Carl Hanser Verlag, München – Wien 1983.
341. P. A. Lewis: *Pigment Handbook*, J. Wiley & Sons, New York 1988.
342. G. Benzing *et al.*: *Pigmente für Anstrichmittel*, expert verlag, Grafenau 1988.
343. Bayer AG: "Anorganische Pigmente für Anstrich- und Beschichtungsstoffe," company bulletin, Leverkusen 1989.
344. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* vol. 2, Verlag W. A. Colomb, Berlin 1974.
345. A. Goldschmidt et al: "Lacke und Farben," *GlasuritHandbuch*, C. R. Vincentz Verlag, Hannover 1984.
346. G. D. Parfitt: *Dispersions of Powders in Liquids*, Applied Science Publishers, London 1981.
347. T. C. Patton: *Paint Flow and Pigment Dispersion*, J. Wiley & Sons, New York 1978.
348. *TRGS, UVV, MAK-Werte (TRGS 900)*, C. Heymanns Verlag KG, Köln 1989.
349. Kunststoff LMBG XL: Gesundheitliche Beurteilung von Kunststoffen im Rahmen des Lebensmittel- und Bedarfsgegenständegesetzes: XL Lacke und Anstrichstoffe für Lebensmittelbehälter und -verpackungen, Bundesgesundheitsblatt, no. 13, 1968, pp. 192 – 194.
350. O. Lückert: *Pigment+Füllstoff-Tabellen*, 4th ed., Laatzen 1989.
351. H. Haagen, D. Martinovic, *Farbe+Lack* **95** (1989) 892 – 895.
352. H. S. Katz, J. V. Milewski: *Handbook of Fillers and Reinforcements for Plastics*, Van Nostrand Reinhold Co., New York 1978.
353. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **II**, Verlag W. A. Colomb in der H. Heenemann GmbH, Berlin–Oberschwandorf 1974, pp. 457.
354. J. Bieleman (ed.): *Coatings Additives*, Wiley-VCH, Weinheim, Germany 2002.
355. H. Kittel: *Lehrbuch der Lacke und Beschichtungen, Band 4*, Hirzel-Verlag, Stuttgart 2006.
356. L. J. Calbo (ed.): *Handbook of Coatings Additives*, Marcel Dekker, New York 1987.
357. D. R. Karsa (ed.): *Additives for Waterbased Coatings*, The Royal Society of Chemistry, Cambridge 1991.
358. R. Lambourne (ed.): *Paint and Surface Coatings: Theory and Practice*, Ellis Horwood, Chichester 1987.
359. T. C. Patton: *Paint Flow and Pigment Dispersion*, J. Wiley & Sons, New York 1979.
360. A. Goldschmidt, H.-J. Streitberger: *Basics of coating technology*, Vincentz Network, Hannover 2003.
361. E. Karsten, O. Lückert: *Lackrohstoff-Tabellen*, Vincentz Network, Hannover 2000.
362. National Paint and Coatings Association (ed.): *Raw Materials Index*, NPCA, Washington 1989.
363. P. R. Garrett (ed.): *Defoaming, Surfactant Science Series*, Vol. 45, Marcel Dekker, New York 1993.
364. R. J. Pugh: "Foam Breaking in Aqueous Systems" in K. Holmerg (ed.): *Handbook of Applied Surface and Colloid Chemistry*, John Wiley & Sons, New York 2001.
365. R. J. Pugh: "Foams and Foaming" in K. Holmerg (ed.): *Handbook of Applied Surface and Colloid Chemistry*, John Wiley & Sons, New York 2001.
366. G. D. Parfitt (ed.): *Dispersion of Powders in Liquids (with Special Reference to Pigments)*, Applied Science Publ., London 1981.
367. J. Lyklema: *Fundamentals of Interface and Colloid Science*, Vols. 1 – 3, Academic Press, London 1991–2000.
368. D. H. Napper: *Polymeric Stabilisation of Colloidal Dispersions*, Academic Press, London 1983.
369. G. Berner, M. Rembold, *Farbe+Lack* **89** (1983) 840.
370. A. Jurgetz, H. Rothbächer, *Farbe+Lack* **91** (1985) 921.
371. G. Berner, G. Deslex, K. Schoohf: *Congr. FATIPEC*, vol. **II**, Amsterdam 1980, p. 61.
372. H. J. Heller, H. R. Blattmann, *Pure Appl. Chem.* **30** (1972) 145; **36** (1973) 141.
373. A. Valet, F. Sitek, G. Berner, *Farbe+Lack* **94** (1988) 734.
374. J. Otterstedt, *J. Chem. Phys.* **58** (1973) 5716. Links

375. W. Klöppfer, *J. Polym. Sci. Polym. Symp.* **57** (1978) 205.
376. F. Werner, H. E. A. Kramer, *Eur. Polym. J.* **13** (1977) 501. Links
377. M. Rembold, G. Berner, *Org. Coat. Appl. Polym. Sci. Proc.* **47** (1982) 624.
378. A. Valet, *Farbe+Lack* **96** (1990) 185.
379. D. M. Wiles, D. J. Carlson: *New Trends in the Photochemistry of Polymers*, Elsevier, London–New York 1985.
380. J. L. Gerlock, D. R. Bauer, D. F. Mielewski, *11th Int. Conf. on Advances in the Stabilization and Controlled Degradation of Polymers*, Luzern 1989, p. 25.
381. Kirk-Othmer, 3rd. ed., **16,** pp. 762 – 768.
382. V. Wagenseil, *Oberfläche JOT* **12** (1988) 26 – 28.
383. W. Antony, *Oberfläche JOT* **8** (1988) 16 – 18.
384. G. P. Wahl, *Das deutsche Malerblatt* **7** (1990) 41 –44.

General References

385. G. D. Parfitt: *Dispersion of Powders in Liquids*, Elsevier, Amsterdam–London–New York 1969.
386. H. Rumpf: *Mechanische Verfahrenstechnik*, C. Hanser Verlag, München–Wien 1975.
387. W. Jänicke, *Comp. Chem. Eng.* **3** (1987) 303 – 309.
388. ISO 9000 – 9004.
389. N. Stehr, J. Schwedes, *Aufbereit. Tech.* **10** (1983) 597 – 604.
390. N. Stehr, *Keram. Z.* **42** (1990) no. 3, 162 – 167.
391. H. Weit, J. Schwedes, *Chem.-Ing.-Tech.* **58** (1986) no. 10, 818 – 819.
392. J. Winkler, L. Dulog, *Farbe+Lack* **90** (1984) no. 4, 244 – 250 (1), no. 5, 355 – 359 (2).
393. O. Schmitz, J. Luo: *Congr. FATIPEC XVIII*, vol.**I b**, Venice 1986, pp. 623 – 655.

Specific References

394. *Galasurit-Handbuch "Lacke+Farben"*, BASF Lacke+Farben AG, C. R. Vincentz Verlag, Hannover 1984.
395. M. Bruhm in W. Masing (ed.): *Handbuch der Qualitätssicherung*, C. Hanser, München–Wien 1988.
396. M. H. Pahl, *Chem.-Ing.-Tech.* **57** (1985) no. 5, 421 – 430 (part 1), *Chem.-Ing.-Tech.* **57** (1985) no. 6, 506 – 510 (part 2).
397. J. McErlane, *Pigm. Resin Technol.* **8** (1987) 4 – 7.
398. O. Wälty, *Farbe+Lack* **92** (1986) no. 8, 734 – 773.
399. H.-P. Wilke *et al.*: *Rührtechnik*, Hüthig-Verlag, Heidelberg 1988.
400. R. Pörtner, U. Werner, *Chem.-Ing.-Tech.* **61** (1989) no. 3, 250 – 251.
401. R. Geisler, A. Mersmann, H. Voit, *Chem.-Ing.-Tech.* **60** (1988) no. 12, 947 – 955.
402. J. Raasch, K. Sommer, *Chem.-Ing.-Tech.* **62** (1990) no. 1, 17 – 22.
403. L. Pedrocchi, *Swiss Chem.* **11** (1989) no. 5, 23 – 32.
404. L. Pedrocchi, F. Widmer, *Chem.-Ing.-Tech.* **61** (1989) no. 1, 82 – 83.
405. W. Müller, *Plaste Kautsch.* **33** (1986) no. 1, 32 – 34.
406. G. Nöltner, *Chem.-Ing.-Tech.* **60** (1988) no. 12, 986 – 994.
407. H. Reichert, *Chem.-Ing.-Tech.* **47** (1975) no. 7, 313.
408. H. Reichert, *Chem.-Ing.-Tech.* **45** (1973) no. 6, 391 – 395.
409. W. Herbst, *Farbe + Lack* **77** (1971) no. 11, 1072 – 1080 (I);no. 12, 1197 – 1203 (II).
410. B. Heinrich, L. Kreitner, *Aufbereit. Tech.* **10** (1981) 556 – 624.
411. J. Johnson, A. Szegvari, M. Li, *Polym. Paint Colour J.* (1982) no. 7, 459 – 463.
412. H. Mölls, R. Hörnle: *Wirkungsmechanismus der Naßzerkleinerung in der Rührwerkskugelmühle*, Dechema Monographie 69, 1972.
413. K. Engels, *Farbe + Lack* **87** (1981) no. 1, 18 – 25.
414. K. Engels, *Farbe + Lack* **94** (1988) no. 5, 350 – 364.
415. I. Berg, *Polym. Paint Colour J. Suppl.* **23** (1983) 1 –28.
416. R. Stadler, R. Polke, J. Schwedes, F. Vock, *Chem.-Ing.-Tech.* **62** (1990) no. 11, 907 – 915.
417. F. Vock, K. Warnke, K. Sollik, *Farbe+Lack* **11/12** (1987) 905 – 910, 995 – 1000.
418. N. Stehr, *Chem.-Ing.-Tech.* **61** (1989) no. 5, 422 –423.
419. A. Goldschmidt, B. Hantschke, E. Knappe, G.-F. Vock: *Glasurit-Handbuch*, 11th ed., Curt R. Vincentz Verlag, Hannover 1984.
420. D. Ondratschek, K. Ortlieb: *Taschenbuch für Lackierbetriebe*, 47th ed., Curt R. Vincentz Verlag, Hannover 1989 (published annually).
421. K. W. Thomer, D. Ondratschek: "Lackieren," in: *Handbuch der Fertigungstechnik,* vol. 4/1, Carl Hanser Verlag, München 1987.
422. D. Oppen: "Die Vorbehandlung des metallischen Anstrichuntergrundes," Seminar: Lackierungen in der Metallindustrie, Technische Akademie Esslingen, Esslingen 1983.
423. W. Rausch: *Die Phosphatierung von Metallen*, 2nd ed., Eugen G. Leuze Verlag, Saulgau 1988.
424. U. Hoffmann: "Grundsätzliche verfahrenstechnische Gesichtspunkte bei der Kunststofflackierung," Seminar: Lackieren von Kunststoffoberflächen, Württembergischer Ingenieurverein Stuttgart, Stuttgart 1984.
425. K. Wittel, *Ind. Lackierbetr.* **54** (1986) no. 3, 121 – 127.
426. R. Hartmann: "Physikalische Vorbehandlungsmethoden von Polymeren," Seminar: 4th Deutsche Kunststoff-Finish, Bad Nauheim 1987, Praxis-Forum, Berlin 1987.
427. W. Burckhardt, *VDI-Ber.* **449** (1982) 89 – 98.
428. W. Burckhardt, *Ind. Lackierbetr.* **51** (1983) no. 5, 162 – 164.
429. B. D. Meyer in: *Umweltfreundliche Lackiersysteme für die industrielle Lackierung,* Expert-Verlag,Kontakt und Studium, vol. 271, Ehningen 1989, pp. 146 – 244.
430. *Fachtagung Pulverlack,* Curt R. Vincentz Verlag, Hannover 1985.
431. F. Wilborn *et al.*: *Physikalische und technologische Prüfverfahren für Lacke und ihre Rohstoffe*, Berliner Union, Stuttgart 1953.
432. H. A. Gardner, G. S. Sward: *Physical and Chemical Examination of Paints*, Varnishes, Lacquers and Colors, Bethesda, Md., 1962.
433. G. Zeidler, G. Bleisch: *Laboratoriumsbuch für die Lack- und Anstrichmittel-Industrie*, W. Knapp Verlag, Düsseldorf 1967.
434. G. G. Sward (ed.): *Paint Testing Manual*, 13th ed., **ASTM STP 500**, ASTM, Philadelphia, Pa. 1972.
435. H. Knittel *et al.*: "Untersuchung und Prüfung," *Lehrbuch der Lacke und Beschichtungen*, vol. VII, part 1, Verlag W. A. Colomb, Berlin–Oberschwandorf 1980.
436. H. Schmidbauer: *Congr. FATIPEC XX.*, Nice 1990.
437. M. Klarskov, A. Saarnack, J. Jakobsen: *Congr. FATIPEC XX*, Nice 1990.
438. K. Armbruster, M. Breucker, *Farbe+Lack* **95** (1989) 896 – 899.
439. D. Baumgarten, *Farbe + Lack* **87** (1981) 375 – 381.
440. H. Plog: *Schichtdicken-Messung, Verfahren und Geräte*, 2nd ed., Lenze Verlag, Saulgau 1968.
441. U. Zorll, *Dtsch. Farben Z.* **30** (1976) 334 – 339.
442. M. Cremer, *Progr. Org. Coat.* **9** (1981) 241 – 279.
443. H. Schene: *Congr. FATIPEC XX*, Nice 1990.
444. F. X. D. O'Donnell, F. W. Billmeyer Jr: "Psychrometric Scaling of Gloss," in I. J. Rennilson, W. N. Hale (eds.): *Review and Evaluation of Appearance,* ASTM, Philadelphia, Pa., 1986.

445. R. S. Hunter, D. B. Judd, *ASTM Bulletin* **97** (1939) 11 – 14.
446. H. K. Hammond III, J. Nimeroff, *J. Res. Nat. Bur. Stand. (U.S.)* **44** (1950) 585 – 598.
447. U. Zorll, *Prog. Org. Coat.* **1** (1972) 113 – 155. Links
448. U. Zorll, *Farbe+Lack* **79** (1973) 191 – 201.
449. R. S. Hunter, R. W. Harold: *The Measurement of Appearance*, Wiley-Interscience, New York – Chichester – Brisbane – Toronto – Singapore 1987.
450. H. K. Hammond III, *ASTM Standardization News* 1987, 36 – 40. Links
451. W. Czepluch, *Farbe+Lack* **93** (1987) 119 – 123.
452. W. Czepluch: *Vergleiche von Tastschnitten, optischen und rechnerischen Verfahren zur Kennzeichnung von Oberflächenprofilen*, VDI-Verlag, Düsseldorf 1987.
453. F. Finus: *Congr. FATIPEC XI*, Garmisch-Partenkirchen 1974.
454. F. Sadowski, *Farbe+Lack* **91** (1985) 702 – 705.
455. W. N. Hale, Jr. *ASTM Standardization News* 1987 46 – 49.
456. W. Schultze: *Farbenlehre und Farbenmessung*, 3. ed., Springer Verlag, Berlin–Heidelberg–New York 1975.
457. H. G. Völz: *Industrielle Farbprüfung*, VCH Verlagsges., Weinheim, Germany 1990.
458. H. Schmelzer, *J. Coat. Technol.* **58** (1968) 53 – 59.
459. H. Cloppenburg, D. Schmittmann, *Farbe+Lack* **95** (1989) 631 – 636.
460. J. Sickfeld, *Industrie-Lackierbetrieb* **58** (1990) 53 –60.
461. H. Kraus, J. Sickfeld, *Industrie-Lackierbetrieb* **58** (1990) 12 – 20.
462. W. König, *Farbe+Lack* **59** (1953) 435 – 443.
463. W. Weiler, H. Fischer, *Materialprüfung* **29** (1987) 308 – 311.
464. A. Zosel, *Farbe+Lack* **94** (1988) 809 – 815.
465. M. Osterbrock, O. T. de Vries, *J. Oil Colour Chem. Assoc.* **73** (1990) 398 – 402.
466. G. W. Becker, K. Frankenfeld, *Farbe+Lack* **66** (1960) 433 – 448.
467. U. Zorll, *Materialprüfung* **31** (1989) 381 – 384.
468. E. Ladstädter, *Farbe+Lack* **90** (1984) 646 – 653.
469. R. Harlfinger, *Farbe+Lack* **94** (1988) 179 – 183.
470. E. Bagda, *Farbe+Lack* **94** (1988) 606 – 609; **95** (1989) 176 – 178.
471. R. R. Wiggle, A. G. Smith, J. V. Petrocelli, *J. Paint Technol.* **40** (1968) 40 – 42.
472. W. Göring, *Industrie-Lackierbetrieb* **57** (1989) 397 – 402.
473. J. G. Calvert, J. N. Pitts: *Photochemistry*, J. Wiley & Sons, New York 1967.
474. P. A. Mullen, N. D. Searle, *J. Appl. Polym. Sci.* **14** (1970) 765 – 776. Links
475. B. Ranby, F. J. Rabek: *Photodegradation, Photooxidation and Photostabilisation of Polymers*, J. Wiley & Sons, London 1975.
476. N. D. Searle in A. V. Patsis (ed.): *Advances in the Stabilization and Controlled Degradation of Polymers*, vol. **I**, Technomic Publishing Co., Lancaster 1989, pp. 62 – 74.
477. P. Trubiroha in A. V. Patsis (ed.): *Advances in the Stabilization and Controlled Degradation of Polymers*, vol. **I**, Technomic Publishing Co., Lancaster 1989, pp. 236 – 241.
478. G. Kämpf, K. Sommer, E. Zirngiebl, *Farbe+Lack* **95** (1989) 883 – 886.
479. W. Papenroth, *Farbe+Lack* **76** (1970) 467 – 471.
480. E. Preininger, *Farbe+Lack* **87** (1981) 100 – 104.
481. J. A. Simms, *J. Coat. Technol.* **59** (1987) 45 – 53.
482. Association of Automobile Industries, *J. Coat. Technol.* **58** (1986) 57 – 65.
483. D. R. Bauer, M. C. Paputa Peck, R. O. Carter III, *J. Coat. Technol.* **59** (1987) 103 – 109.
484. P. Trubiroha, *Industrie-Lackierbetrieb* **58** (1990) 180 – 184.
485. J. Boxall, *Pigm. Resin. Technol.* **14** (1985) no. 9, 4 – 10.
486. J. B. Lear, *J. Coat. Technol.* **53**, (1981) 51 – 57.
487. Oil and Colour Chemist's Association, Australia: *Surface Coatings*, vol. **2**, Paints and their Applications, Tafe Educational Books, Randwick, Australia, 1984, pp. 759 – 790.
488. A. Goldschmidt, B. Hantschke, E. Knappe, G.-F. Vock: *Glasurit-Handbuch*, Curt R. Vincentz Verlag, Hannover 1984, p. 323.
489. S. Hvilsted, *Prog. Org. Coat.* **13** (1985) 253 – 271.
490. Standard Recommended Practice for Qualitative Identification of Polymers in Emulsion Paints, ASTM D 3168–73 (reapproved 1979).
491. H. Wagner, H. F. Sarx: *Lackkunstharze*, Carl Hanser Verlag, München 1971.
492. G. Ellis, M. Clayborn, S. E. Richards, *Spectrochimica Acta* **46 A** (1990) 227 – 241.
493. D. O. Hummel, F. Scholl: *Atlas of Polymer and Plastics Analysis,* vol. 2, Carl Hanser Verlag, München 1984.
494. Chicago Society for Coatings Technology: *An Infrared Spectroscopy Atlas for the Coatings Industry*, Federation of Societies of Coatings Technology, Philadelphia, Pa. 1980.
495. R. G. Messerschmidt, M. A. Harthcock: *Infrared Microspectroscopy*, Marcel Dekker, New York 1988.
496. P. R. Griffiths, J. A. de Haseth: *Fourier Transform Infrared Spectrometry*, J. Wiley & Sons, New York 1986.
497. L. J. Mathias, *Mod. Paint Coat.* **75** (1985) no. 11, 38 – 42, 44.
498. L. J. Mathias, *Proc. Water-Borne Higher-Solids Coat. Symp. 12th* (1985) 25 – 43.
499. I. L. Davis, *J. Oil Colour Chem. Assoc.* **68** (1985) no. 5, 109 – 115.
500. D. M. Doddrell, D. T. Pegg, M. B. Bendall, *J. Magn. Reson.* **48** (1982) 323 – 327.
501. S. Hvilsted in O. Kramer (ed.): *Biological Synthetic Polymer Networks*, Elsevier, London 1988, pp. 243 – 254.
502. G. L. Marshall, J. A. Lander, *Eur. Polym. J.* **21** (1985) 949 – 958. Links
503. K. J. Ivin, S. Pitchumani, C. R. Reddy, S. Rajadurai, *Eur. Polym. J.* **17** (1981) 341 – 346. Links
504. K. H. Matthews, A. McLennaghan, R. A. Pethrick, *Br. Polym. J.* **19** (1987) 165 – 179.
505. J. T. K. Woo, J. E. Evans, *J. Coat. Technol.* **49** (1977) 42 – 50.
506. D. Wendisch, H. Reiff, D. Dietrich, *Angew. Makromol. Chem.* **141** (1986) 173 – 183. Links
507. C. F. Poranski, Jr., W. B. Moniz, *J. Coat. Technol.* **49** (1977) 57 – 61.
508. R. D. Smith, B. W. Wright, C. R. Yonker, *Anal. Chem.* **60** (1988) 1323 A–1336 A. Links
509. C. Y. Kuo, T. Provder, *ACS Symp. Ser.* **352** (1987) 2 – 28.
510. J. W. Hellgeth, L. T. Taylor, *J. Chromatogr. Sci* **24** (1986) 519 – 527. Links
511. U. Schernau, *Eur. Coat. J.* **4** (1989) no. 6, 510 – 520.
512. C. Schenk, A. H. M. Kayen: *Congr. FATIPEC XV*, vol. **III**, Amsterdam 1980, pp. 345 – 360.
513. J. H. Charlesworth, *Anal. Chem.* **50** (1978) 1414 –1420. Links
514. U. Schernau, M. Freitag: *Congr. FATIPEC XIX*, vol. **I**, Aachen 1988, pp. 419 – 430.
515. C. P. A. Kappelmeier: *Chemical Analysis of Resin-Based Coating Materials*, Interscience Publishers Inc., New York 1959.
516. IUPAC, Recommended Methods for the Analysis of Alkyd Resins, *Pure Appl. Chem.* **33** (1973) 413 –435.

517. J. K. Haken, *Prog. Org. Coat.* **14** (1986) 247 – 295.
518. W. Butte, *J. Chromatogr.* **261** (1983) 142 – 145. Links
519. R. W. Scott, *J. Paint Technol.* **41** (1969) 422 – 430.
520. P. Kamarchik, Jr., *J. Coat. Technol.* **52** (1980) 79 – 82.
521. W. Kugler, Landeskriminalamt Baden-Württemberg, Stuttgart, personal communication.
522. R. Jenkins, M. Holomany, *Powder Diffr.* **2** (1987) 215 – 219.
523. E. Baier, *Farbe+Lack* **96** (1990) 173 – 179.
524. M. Saltzmann, R. Kumar, F. W. Billmeyer, *Color. Res. Appl.* **7** (1982) 327 – 337.
525. A. Whitaker, *J. Soc. Dyers Colour.* **102** (1986) 66 –76.
526. R. Goebel, W. Stoecklein, *Scanning Microsc.* **1** (1987) 1007 – 1015.
527. K. J. Hyver, P. Sandra: *High Resolution Gas Chromatography*, 3rd ed., Hewlett-Packard Company, Avondale 1989.
528. R. Moulder *et al.*, *HRC CC, J. High Resolut. Chromatogr. Chromatogr. Commun.* **12** (1989) 688 – 691.
529. S. B. Hawthorne, *Anal. Chem.* **62** (1990) 633 A–642 A.
530. J. B. Lear, *J. Coat. Technol.* **53** (1981) 63 – 65.
531. W. Stoecklein, M. Gloger, *Nicolet FT-IR, Spectral Lines* **9** (1988) no. 1, 2 – 6.
532. R. Blaschke, N. Dingerdissen, R. Göcke, K. Franke: *Congr. FATIPEC XVII*, vol. **II**, Lugano 1984, pp. 363 – 387.
533. A. Benninghoven, *J. Vac. Sci. Technol.* **A 3** (1985) 451 – 460.
534. N. Basener, *Farbe+Lack* **94** (1988) 821 – 825.
535. J. A. Gardella, Jr., J. J. Pireaux, *Anal. Chem.* **62** (1990) 645 A–661 A.
536. *Glasurit Handbuch*, 11th ed., Curt R. Vincentz Verlag, Hannover 1984.
537. K. A. van Oeteren: *Korrosionsschutz durch Beschichtungsstoffe,* vols. 1 and 2, Carl Hanser Verlag, München 1980.
538. W. Göldner, "Wasserverdünnbare Lacke für die Autoserienlackierung," *Ind. Lackierbetr.* **58** (1990) no. 6, 205 – 207.
539. K. Wittel: "Vorbehandlungsverfahren beim Einsatz verschiedener Metalle," *Ind. Lackierbetr.* **57** (1989) no. 3, 97 ff.
540. R. D. Athey: "Electrocoating Review," *Met. Finish.* **87** (1989) no. 4, 23 – 27.
541. G. Lovell: "Electrocoat Basics," *Prod. Finish.* **54** (1990) no. 7, 58 ff.
542. J. Goebbels: "Bedingungen für die Anwendung von Wasserbasislack," *Oberfläche±JOT* **29** (1989) no. 10, 48 ff.
543. R. W. Blackford: "Corrosion Protective Coatings in Aerospace," *Schriftenreihe Praxisforum*, vol. **1**, Berlin, 18 – 21 March 1991, pp. 242 – 249.
544. L. W. Jedermann: "Chromatfreie Primer für die Luftfahrt–Luftfahrtspezifischen Anforderungen an das Korrosionsschutzvermögen und an die Mediumbeständigkeit," *Schriftenreihe Praxisforum*, vol. **1**, Berlin, 18 – 21 March 1991, pp. 251 ff.
545. J. Hirst, C. R. Hegedus: "Synergistic Soiling, Cleaning and Weathering Effects on Aircraft Polyurethane Topcoats," *Met. Finish.* **87** (1989) no. 1.
546. D. M. James: "Marine Paints," in R. R. Myers, J. S. Long (eds.): *Treatise on Coatings,* vol. 4, part 1, Marcel Dekker, New York 1975, pp. 415 –475.
547. *Marine Fouling and Its Prevention*, Woods Hole Oceanographic Institute & United States Naval Academy, Annapolis, Md. 1952.
548. R. Lambourne: "The Painting of Ships," chapter 13 in R. Lambourne (ed.): *Paint and Surface Coatings,* Ellis Horwood, Chichester 1987, pp. 528 –546.
549. A. V. Robinson: "Tank Coatings," *Proc. MariChem.* **80** (1980) 169.
550. D. Banks: "The Effects of Low Molecular Weight Cargoes on Tank Coating Systems," *Proc. MariChem.* **85** (1985) 169 – 172.
551. C. E. M. van der Kolk: "OCL Epoxy Tanklining," *Proc. MariChem.* **89** (1989) session 8, paper 8.6.
552. O. B. Sorensen: "The Factors Influencing Ventilation and Drying of Epoxy Coated Cargo Tanks," *Proc. MariChem.* **89** (1989) session 8, paper 8.7.
553. R. L. Townsin, D. Byrne, T. Svensen, A. Milne: "Speed, Power and Roughness; The Economics of Outer Bottom Maintenance," *Trans RINA* **122** (1980) 459 – 474.
554. J. D. Castlow, R. C. Tipper: *Marine Biodeterioration, and Interdisciplinary Study*, Naval Institute Press, Annapolis 1984.
555. J. A. Montemarano, J. Dyckman: "Performance of Organo-Metallic Polymers," *J. Paint Technol.* **47** (1975) 59.
556. A. Deeks, D. Hudson, D. James, B. W. Sparrow: "Tributyltin Methacrylate Copolymers in Antifouling Paints," *2nd International Congress of Marine Corrosion and Fouling*, Athens 1968.
557. D. James, GB 1 124 297, 1968.
558. A. Milne, G. Hails, GB 1 457 590, 1976.
559. A. Milne: "Self-Polishing Coatings in Marine Antifouling Paints," Royal Inst. Chem. Annual Chemical Congress, Durham, April 1980.
560. A. Milne, M. E. Callow, R. Pitchers: "Control of Fouling by Non-Biocidal Systems," in Evans & Hoagland (eds.): *Algal Biofouling,* Elsevier, Amsterdam 1986, pp. 145 – 158.
561. A. Milne, GB 1 470 465, 1989.
562. R. R. Brooks, EP 329 375, 1989.
563. A. Milne, I. S. Millichamp, GB 2 218 708 A, 1989.
564. B. Meuthen *et al.*: "25 Years of Wide Strip Coil Coating," *ECCA Congress Brussels*, Nov. 1985.
565. S. Fabriz: "Neue Lacksysteme im systematischen Vergleich der Einsatzchancen unter heutiger Gesetzgebung," *Ind. Lackierbetr.* **57** (1989) no. 4, 177 ff.
566. W. Baulmann: "Durchbruch für die Elektronenstrahlhärtung von Möbeloberflächen," *Ind. Lackierbetr.* **57** (1989) no. 6, 208 ff.
567. B. Riberi: "Polyurethanlacke für die Möbelindustrie,", *Ind. Lackierbetr.* **55** (1987) no. 10, 341 ff.
568. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* vol. 1, Verlag W. A. Colomb, Berlin 1977.
569. A. Brasholz: *Handbuch der Anstrich- und Beschichtungstechnik*, Bauverlag, Wiesbaden 1978.
570. Institut für Fenstertechnik, Report 1983.
571. Bundesausschuß Farbe und Sachwertschutz, " Beschichtungen auf nichtmaßhaltigen Bauteilen aus Holz," Data sheet no. 3, 1991; "Beschichtungen auf Fenstern und Außentüren," Data sheet no. 18, Frankfurt/Main 1989.
572. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* vol. 5, Verlag W. A. Colomb, Berlin 1977.
573. Bundesausschuß Farbe und Sachwertschutz, "*Beschichtungen auf Zink und verzinktem Stahl,*" Data sheet no. 5, Frankfurt/Main 1982.
574. Bundesausschuß Farbe und Sachwertschutz, "Beschichtungen auf Kunststoff im Hochbau," Data sheet no. 22, Frankfurt/Main 1984.
575. F. Müller, *Ind. Lackierbetr.* **58** (1990) 48 – 52.
576. "Sanitär-Armaturen farbig pulverbeschichten," *Ind. Lackierbetr.* **57** (1989) 455 – 458.
577. G. Senn, *Oberfläche±JOT* **29** (1989) no. 9, 38 – 42.
578. S. Schmid, *Ind. Lackierbetr.* **57** (1989) 311 – 316.
579. F.-J. Schanze: "Betriebserfahrungen mit dem Thermoreaktor," *VDI Ber.* **730** (1989) 61 – 77.
580. W. Koch, M. Angrick: "Neue Entwicklungen im Bereich der Biofiltertechnik," *VDI Ber.* **735**, 349 –355. A. Windsperger, R. Bucher, K. Stefan, *Staub Reinh. d. Luft* **50** (1990) 465 – 470.
581. D. Schneider, *Oberfläche ± JOT* **29** (1989) no. 9, 61 – 64.
582. T. R. May, *Farbe+Lack* **95** (1989) 816 – 822.
583. A. Askergren *et al.*, *Arb. Hälsa* **4** (1988) 1 – 64.
584. "Some Organic Solvents, Resin Monomers and Related Compounds, Pigments and Occupational Exposures in Paint Manufacture and Painting," *IARC Monogr. Carcinog. Risks Chem. Humans* **47** (1989) 385 – 425.
585. K. Ringen: "Health Hazards among Painters," in A. Englund, K. Ringen, M. A. Mehlman (eds.): *Occupational Health Hazard of Solvents, Advances in Modern Environmental Toxicology,* vol. II, Princeton Sci. Publ., Princeton 1982, pp. 111 – 131.
586. *Panorama of EC Industry 1989*, EEC Brussels-Luxembourg 1988, pp. 7 – 10.

587. *Paint & Ink International* **2** (1989) , no. 3, 14 – 46.
588. K. H. Büchel, H.-H. Moretto, P. Woditsch, *Industrial Inorganic Chemistry*, Wiley-VCH, Weinheim (2000), p 551.

[Top of Page]

Related Articles
⟶ Artists' Colors; ⟶ Metals, Surface Treatment; ⟶ Printing Inks; ⟶ Solvents; ⟶ Wood, Surface Treatment are separate keywords.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Paper and Pulp

Rudolf Patt, University of Hamburg, Hamburg, Federal Republic of Germany (Chap. 1)

Othar Kordsachia, University of Hamburg, Hamburg, Federal Republic of Germany (Chap. 1)

Richard Süttinger, Schongau, Federal Republic of Germany (Chap. 2)

Yoshito Ohtani, Japan Pulp & Paper Research Institute, Tsubuka, Japan (Chap. 3, Section 6.3)

Jochen F. Hoesch, Gaggenau-Sulzbach, Federal Republic of Germany (Sections 4.1, 4.2)

Peter Ehrler, Institut für Textil- und Verfahrenstechnik , Denkendorf, Federal Republic of Germany (Section 4.3)

Rudolf Eichinger, Institut für Papier, Zellstoff und Fasertechnik, Graz, Austria (Chap. 5)

Herbert Holik, Sulzer Escher Wyss GmbH, Ravensburg, Federal Republic of Germany (Sections 6.1 and 6.2, Chap. 7)

Udo Hamm, Institut für Papierfabrikation, Darmstadt, Federal Republic of Germany (Chap. 8)

Michael E. Rohmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 9)

Peter Mummenhoff, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 10)

Erich Petermann, Sankt Augustin, Federal Republic of Germany (Chap. 11)

Richard F. Miller, Akzo Chemicals Ltd., Fairfield, Connecticut 06430, United States (Chap. 12)

Dieter Frank, Akzo Chemicals Ltd., Fairfield, Connecticut 06430, United States (Chap. 12)

Renke Wilken, Papiertechnische Stiftung, München (PTS), Federal Republic of Germany (Chap. 13)

Heinrich L. Baumgarten, Papiertechnische Stiftung , München (PTS), Federal Republic of Germany (Chap. 13)

Gert-Heinz Rentrop, Verband Deutscher Papierfabriken, Bonn, Federal Republic of Germany (Chap. 14)

1. Pulp 5
1.1. Raw Materials for Pulp 6
1.2. Processing of Raw Materials . . . 9
1.2.1. Processing of Wood 9
1.2.2. Processing of Annual Plants 13
1.3. Pulping Processes 14
1.3.1. Alkaline Pulping Processes 15
1.3.1.1. Kraft (Sulfate) Pulping 15
1.3.1.2. Soda Pulping of Annual Plants . . . 20
1.3.2. Sulfite Process 20
1.3.2.1. Reactions in Sulfite Pulping 22
1.3.2.2. The Various Sulfite Processes . . . 23
1.3.2.3. Multistage Sulfite Processes 26
1.3.2.4. Anthraquinone-Catalyzed Sulfite Pulping 27
1.3.3. Other Pulping Processes 27
1.3.3.1. Nitric Acid Pulping 27
1.3.3.2. Pulping with Organic Solvents . . . 28
1.3.3.3. Pulping with Organic Acids 29
1.4. Technology of Pulp Production . 29
1.4.1. Pulping 29
1.4.2. Washing and Screening of Pulp . . 31
1.4.3. Pulp Drying 35
1.5. Waste Liquor Treatment 36
1.5.1. Recovery of Useful Products from Waste Liquor 36
1.5.2. Energy Generation from Waste Liquor 38
1.5.3. Chemicals Recovery 40
1.6. Pulp Bleaching 43
1.6.1. Conventional Bleaching 43
1.6.2. Chlorine-Reduced and Chlorine-Free Pulp Bleaching 47
1.6.3. Technology of Bleaching 50
1.7. Environmental Aspects of Pulp Production 52
1.8. Pulp Properties and Applications 54
2. Mechanical Pulp 54
2.1. Introduction 54
2.1.1. The Raw Material, Wood 54
2.1.2. Characteristics Required of Mechanical Pulps 55
2.2. Groundwood 56
2.2.1. The Grinder Stone 57

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_545

2.2.2. Groundwood Production 58
2.2.3. Newer Methods of Groundwood Production 59
2.3. Refiner Mechanical Pulps 59
2.3.1. Historical Survey 59
2.3.2. Theory of the Refiner Process . . . 59
2.3.3. TMP Production 60
2.3.4. New Methods of Refiner Pulp Production 60
2.4. Screening of Mechanical Pulps . 61
2.4.1. Purpose of Screening 61
2.4.2. Mechanical Pulp Screening 62
2.5. Bleaching of Mechanical Pulps . 62
2.6. Comparison of Mechanical Pulps 63
3. Wastepaper 63
3.1. Types of Wastepaper 63
3.2. Origin, Collection, Handling, and Economic Aspects 64
4. Nonwood Plant Fibers 65
4.1. Rags 65
4.2. Other Nonwood Plant Fibers . . . 66
4.3. Synthetic Fibers 67
4.3.1. Survey 67
4.3.2. Marketing and Packages of Synthetic Fibers 68
4.3.3. Processing Properties and Characteristics of Synthetic Fibers 68
4.3.4. Influence of Synthetic-Fiber Properties on Product Properties . . 68
4.3.4.1. Polymer-Dependent (Uninfluenceable) Fiber Properties 68
4.3.4.2. Type-Dependent (Influenceable) Fiber Properties . . 69
4.3.5. Special Fibers: Bonding Fibers . . 70
5. Raw Materials for Stock Preparation 70
5.1. Fibrous Raw Materials 70
5.1.1. Pulp 71
5.1.2. Mechanical Pulp 72
5.1.3. Wastepaper 72
5.2. Stock Additives 72
5.2.1. Additives with Qualitative Effects . 72
5.2.1.1. Fillers 72
5.2.1.2. Strength Additives 74
5.2.1.3. Sizing Agents 74
5.2.1.4. Dyes 75
5.2.1.5. Optical Brighteners 76
5.2.1.6. Chemicals for Specialty Papers . . 76
5.2.2. Processing Aids 76
5.2.2.1. Retention Aids 76
5.2.2.2. Defoamers 76
5.2.2.3. Biocides 77
5.2.2.4. Dispersing Agents 77
5.2.2.5. Complexing Agents 77
5.2.2.6. Precipitating and Fixing Agents . . 77
5.2.2.7. Drainage Aids 77
5.2.2.8. Additives for Wastepaper Processing and Deinking 78
6. Stock Preparation 78
6.1. Process Steps 78
6.1.1. Fiber Disintegration 79
6.1.2. Separation 80
6.1.3. Deforming/Disperging of Impurities 83
6.1.4. Refining 84
6.1.5. Storage and Mixing 84
6.2. Preparation of Primary Raw Stocks 85
6.2.1. Preparation of Primary Pulp in Bales 85
6.2.2. Special Processing Features for Various Primary Raw Stocks 86
6.3. Wastepaper Systems 86
6.3.1. Repulping 86
6.3.2. Deinking 87
6.3.3. Screening and Cleaning 88
6.3.4. Bleaching 89
6.3.5. Examples of Wastepaper Processing Systems 89
7. Paper and Board Manufacturing 89
7.1. Introduction 89
7.2. Sheet Forming 91
7.2.1. The Stock Approach Flow System 91
7.2.2. The Headbox 92
7.2.3. The Wire Section 93
7.2.4. Systems for the Production of Multilayer Products 97
7.3. Press Section 97
7.4. Dryer Section 99
7.5. Size Press 99
7.6. Calendering 100
7.7. Supercalendering 101
7.8. Reel-Slitting Machine 101
8. Water Circuits and Effluent Treatment 103
8.1. Introduction 103
8.2. Fresh Water 103
8.3. Water Circuits 104
8.4. Closed Water Circuits 104
8.5. Save-Alls 105
8.6. Wastewater Pollution 107
8.7. Effluent Treatment Plants 108
9. Coating of Paper and Board . . . 109
9.1. Requirements of Coated Papers . 111
9.2. Properties and Components of Coating Colors 111
9.2.1. Pigments 112
9.2.2. Dispersing Agents 112

9.2.3. Binders 113
9.2.4. Other Additives 115
9.2.5. Coating Color Formulations 116
9.3. Coating Color Preparation 117
9.4. Coating Machines 118
10. Paper Dyeing 120
10.1. Dyeing Processes 120
10.2. Fastness, Dyes, Pigments 121
11. Types of Paper, Cardboard, and Board 123
11.1. Graphic Papers 123
11.2. Office and Administration Papers 124
11.3. Paper, Board, and Cardboard for Packaging 126
11.4. Hygienic Papers 128
11.5. Papers and Boards for Technical and Specialty Uses 128
12. Book and Paper Preservation . . 131
12.1. Introduction 131
12.2. Paper and Deacidification Chemistry 131
12.3. Development of Mass Deacidification Processes 132
12.4. Commercial Mass Deacidification Processes 133
12.5. Commercial Prospects 137
13. Properties and Testing 137
13.1. The Material Paper: A Survey . . 137
13.1.1. Introduction 137
13.1.2. Material Properties 138
13.1.3. Summary 140
13.2. Paper Testing 140
13.2.1. Preliminary Measures 140
13.2.2. Basic Properties 140
13.2.3. Composition and Chemical Paper Testing 141
13.2.4. Strength Properties 141
13.2.5. Load – Deformation Properties . . . 142
13.2.6. Surface Properties 143
13.2.7. Optical Properties 143
13.2.8. Printing Properties 143
13.2.9. Behavior Towards Liquids 143
13.2.10. Exclusion of Gases and Vapors . . 144
14. Economic Aspects 144
15. References 147

1. Pulp

Wood and other lignocelluloses, such as straw or other annual plants consist of many different cell types in which cellulose, hemicelluloses, and lignin are the main chemical components. In the chemical digestion of lignocelluloses, lignin, which is embedded predominantly between the wood fibers as a binder and in the outer cell wall layers, and the hemicelluloses are largely dissolved out of the fiber matrix. The cohesion of the structural elements is lost in this process. The fibrous material obtained after digestion, consisting principally of cellulose, is called pulp.

The chemicals used in the digestion cannot quantitatively eliminate lignin from the wood matrix. In addition, unwanted, sometimes colored accompanying substances, called wood extractives, may remain in the cell walls. For most uses, the residual lignin and the extractives must be eliminated in a subsequent bleaching treatment. Moreover, partial or complete removal of hemicelluloses by bleaching and purification is required if the cellulose in the pulp is to be employed as raw material for chemical products.

Pulps that are used to make paper and board are called paper or chemical pulps, and those processed in the chemical industry are known as dissolving pulps. Pulps are also classified according to the type of raw material used (softwood or long-fiber pulps, hardwood or short-fiber pulps, straw pulps, etc.) and the pulping process employed (e.g., kraft, sulfite, and soda pulps).

The pulping and bleaching processes must be adapted to the final use of the pulp. The digestion of the raw material and the bleaching of the pulp can be carried out with varying intensity, or the latter step can be completely omitted. Depending on the intensity of pulping, it is possible to differentiate between semichemical and high-yield pulps and between chemical pulps and dissolving pulps. Depending on the residual lignin content of chemical and dissolving pulps, a difference is also made between harder and softer pulp qualities. Semichemical and high-yield pulps can be used with or without bleaching, but chemical and dissolving pulps must usually be bleached. There are slightly bleached, partially bleached, and fully bleached qualities of pulp. Paper and board production is by far the dominant use; only about 5 % is processed in the chemical industry.

1.1. Raw Materials for Pulp

Apart from good availability and processibility, the distribution and morphology of the cell types and the chemical composition are important criteria for the suitability of lignocelluloses for pulp production. The proportion of fibrous elements with a favorable length-to-width ratio is crucial for the technological properties of the resultant pulps. The cellulose content as well as the amount of lignin and its structure determine the pulp yield and the conditions to be used for digestion. In particular, acid sulfite digestion, and also the further processing of pulp, can be adversely affected by accompanying substances, such as fats, waxes, phenolic components, or mineral inclusions. The average composition of important raw materials for pulp production is presented in Table 1. Note that considerable deviations in individual analytical data can be observed, depending on the method of analysis and especially on the sample of raw material used. However, the values given do exemplify the essential differences in the chemical composition of conifers (softwoods), deciduous woods (hardwoods), and annual plants.

The most important raw material for the production of pulp is wood, softwoods being preferred. However, hardwoods have gained increasing importance because, in comparison with softwoods, they have some interesting advantages for the production of pulp and paper (Table 2). Only 7 – 8 % of pulp production is based on annual plants, which are mainly employed when wood is not available [16].

The morphology of the lignocelluloses varies considerably (see also → Wood). Conifers have the most homogeneous structure. They consist to more than 90 % of tracheids with an average length of 3 mm and a width of 30 – 50 µm. These fiber cells have a length-to-width ratio > 100:1 and are ideally suited to papermaking. On the other hand, the parenchyma cells, which occur only in small amounts in conifers, have unfavorable technological properties. They are very small and have an unfavorable length-to-width ratio of ca. 3 : 1.

The structure of hardwoods is much more heterogeneous. The most important cell type are the libriform (or scleremchyma) fibers, which account for 40 – 60 % of the tissue of most deciduous trees. The average length of hardwood fibers is usually less than 1 mm. The fiber width varies greatly, in the range 10 – 40 µm, depending on the type of wood. In addition, hardwoods contain vessels and, in comparison with conifers, a higher content of parenchyma cells. The vessels are very voluminous, with relatively thin, partially perforated walls and a low length-to-width ratio.

The cellular dimensions of annual plants are extremely variable. In comparison with wood, the cells are, in most cases, shorter and have thinner walls, and the parenchyma content is very high. A simplified survey of the cellular composition and morphology of various raw materials for pulp is presented in Figure 1 [17].

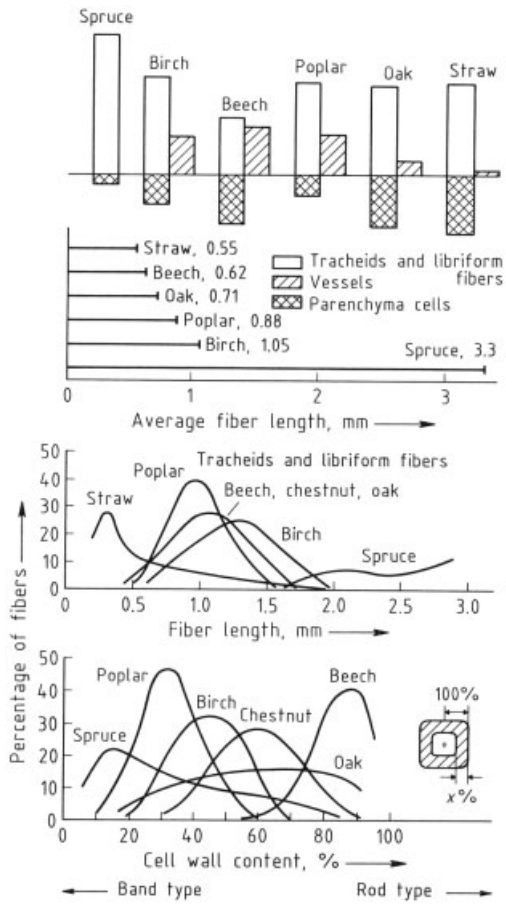

Figure 1. Composition and morphology of various pulp raw materials

In the pulp and paper industry, a morphologically correct differentiation of cell types is not

Table 1. Chemical composition of important raw materials for pulp production

Common name	Scientific name	Cellulose, %	Hemi-cellulose, %	Pentosan, %	Lignin, %	Ethanol/ benzene extract, %	Ash, %	Silica, %
Scots pine	*Pinus sylvestris*	44	26	9	28	4	0.4	
European spruce	*Picea abies*	43	27	9	29	2	0.4	
Douglas fir	*Pseudosuga menziesii*	47	22	7	30	5	0.3	
Silver fir	*Abies alba*	43	27	11	29		0.5	
Silver birch	*Betula verrucosa*	46	36	25	19	2	0.3	
European birch	*Fagus sylvatica*	45	35	22	22	2	0.3	
Trembling aspen	*Populus tremuloides*	50	31	17	18	4	0.3	
Blue gum	*Eucalyptus globulus*	47	27	17	26	1.5	0.3	
Wheat straw	*Tritium vulgare*	38	36	28	19	4	5	3.5
Rice straw	*Oryza sativa*	32	36	25	12	4	16	12
Sugarcane bagasse	*Saccharinum officinarum*	38	36	27	21	3	2	1.5
Bamboo	*Dendrocalamus strictus*	35	32	18	26	5	3.5	2.5

Table 2. Advantages of softwoods and hardwoods for pulping and papermaking

	Softwoods	Hardwoods
Wood production	usually higher productivity	in some cases extremely high productivity (e.g., eucalyptus, poplar), higher stock stability
Wood	more uniform raw material, especially better stem formation, easier to debark, longer fibers, higher yields in acidic cooking	higher specific weight (e.g., eucalyptus, beech, birch) → lower transport cost → higher digester capacity → higher pulp output lower lignin content → easier to pulp → less energy and chemicals consumption higher yields in alkaline cooking
Pulp	better wet web strength, better drainage properties	requires less beating energy, rapid strength development during beating, better bleachability → fewer bleaching stages → lower chemical demand → less pollution (chlorine-free bleaching possible)
Paper	high strength properties, especially tear strength, better runability	better sheet formation (higher content of hemicelluloses), usually smoother surface, higher bulk, higher opacity → better printability

made. All cells are designated as fibers. The suitability of cells for papermaking depends essentially on the length-to-width ratio and the ratio of twice the radial cell-wall thickness to the lumen diameter. The latter is important for the flexibility or stiffness of a structural element. In this respect, considerable differences can be observed within the same cell type due to the seasonal growth rhythms and to the formation of specific cell structures because of special stresses in certain tissue regions. At the start of the growth period, cells having a large diameter and thin walls (springwood) are formed, while at the end of the vegetation period, thick-walled cells are produced (summerwood). Compressive and tensile stresses can also result in the formation of especially thick-walled cells that are capable of absorbing these stresses. The increase in the wall thickness of spruce tracheids during the growth period is shown in Figure 2.

The structure of the tissue elements influences the technological properties of the resulting paper in the following manner. In papermaking, the fibers suspended in water deposit on the wire of the paper machine to form a fiber web. Hydrogen bonds are responsible for bonding the fibers together. Apart from the presence of water, formation of hydrogen bonds requires hy-

drophilic groups for contact between the fiber surfaces, and small interfiber distances. The removal of the hydrophobic lignin from the surface of the fiber exposes the cellulose and hemicelluloses with their hydrophilic OH groups. The tensile strength and burst strength of a paper sheet depend primarily on the number of hydrogen bonds per unit volume between the fiber surfaces. Therefore, during paper sheet formation it must be ensured that the interfiber distance required for the formation of hydrogen bonds is attained. Springwood fibers with thin walls and large diameters (i.e., with large volumes) collapse during paper sheet formation because the lignin responsible for the stiffness of the cell wall was dissolved during digestion. Hence, the fiber becomes much more flexible, and its surface can touch neighboring fibers in large areas of the paper web, leading to increased hydrogen bond formation. This effect can be increased still further by mechanical treatment of the pulp in a beating step.

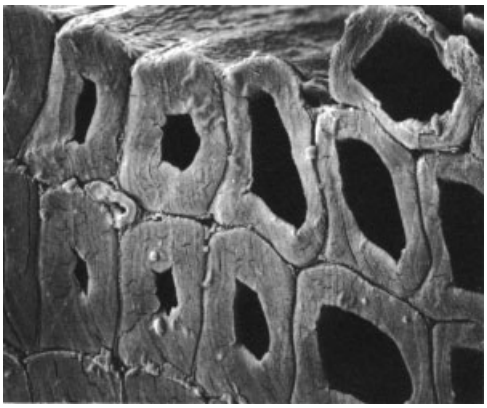

Figure 2. Increase of the cell wall thickness of spruce tracheids during the growth period

Summerwood fibers, having thick cell walls and being less flexible, hardly collapse and therefore form fewer hydrogen bonds per unit volume. However, they contribute greatly to the tear strength of the fiber web because of their inherent high fiber strength. In addition, they increase the volume of the paper and improve the optical properties by increasing the opacity of the paper. Thus, long, thick-walled fibers are required for high tear resistance of the pulp, while high fiber flexibility and a large surface-to-volume ratio are required for good interfiber bonding and high tensile strength. Therefore, fibers with a good length-to-width ratio have optimal properties for papermaking. This ratio should be at least 100 : 1. Figures 3 and 4 show the embedding of unbeaten and slightly beaten fibers in a paper web.

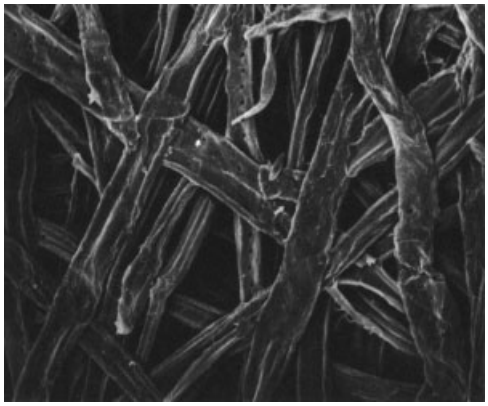

Figure 3. Electron micrograph of a test sheet made of unbeaten oak pulp

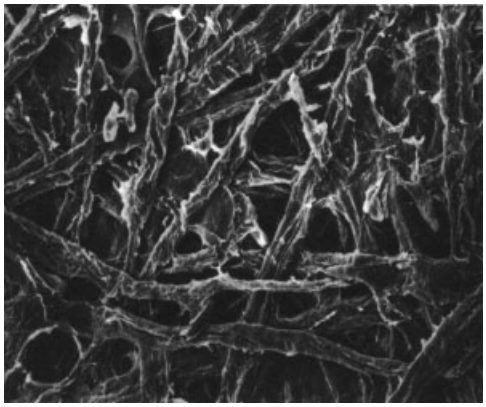

Figure 4. Electron micrograph of a test sheet made of oak pulp after 20 min beating

In Europe, sawmill residues and wood from thinning of forests are used in the production of wood pulp. In countries with good growth conditions, wood for pulp production is predominantly cultivated on plantations. Conifers, particularly spruce and pine, are preferentially used because of their favorable morphological structure. However, pulps of good quality can also be made from hardwoods, such as birch, beech, eucalyptus, and poplar (see Tables 1 and 2). In

particular, the importance of eucalyptus for the fiber industry should increase steadily because of its high growth rate and the excellent properties of pulps made from this wood [18–20]. Each type of wood is normally processed separately because, to achieve optimal pulp properties, the production process must be adjusted to the specific requirements of that particular type of wood. Mixtures of woods of greatly varying morphology and chemical composition, e.g., mixtures of hardwoods and softwoods or mixtures of hardwoods of widely varying wood density, lead to unsatisfactory digestion results and pulp properties.

1.2. Processing of Raw Materials

1.2.1. Processing of Wood

Wood is the most important raw material for the pulp industry and is obtained as logs, saw-mill residues, and chips. Logs normally enter the pulp mill with bark, but industrial residues should be bark-free.

Storage of Wood. Weeks and months often pass between wood harvesting and its processing in the pulp mill. In this time, biological respiration of living cells, chemical reactions, and attack by microorganisms can degrade the wood. The extent of wood degradation depends on many factors, the most important being the duration of storage, the storage conditions, the climate, and the type of wood.

In the storage of very fresh wood, especially from summer felling, the respiration of nutrients still present in the wood continues, generating heat, especially in the center of wood chip piles. Temperatures of 60 °C can be reached in only one to two weeks and in extreme cases temperatures as high as 90 °C have been observed. Spontaneous ignition can even occur under these conditions. After this increase in temperature, fungi and bacteria attack the wood, reducing the temperature again. Normally a temperature gradient arises, with the highest temperatures in the upper part of the pile.

Apart from microbial oxidation, chemical autooxidative and hydrolytic degradations also occur. In autooxidation, the wood components react with atmospheric oxygen; the final products are frequently organic acids. Substances that primarily undergo autooxidation are wood extractives, especially unsaturated fatty acids, resin acids, and terpenes. Although hydrolysis, particularly of carbohydrates, contributes only to a small extent to the generation of heat, it releases acetic acid from the hemicelluloses, which causes further hydrolysis.

Fungi and bacteria have the greatest destructive effect on wood. Most of the microorganisms that destroy wood have an optimum temperature between 25 and 50 °C; the most damaging microorganisms have an optimum of 30 °C. The optimal moisture content for microbial attack is 24 – 32 %. Degradation is accompanied by a reduction of strength of the wood and of the individual tissue components. Frequently, discoloration of the wood also occurs.

No general statements can be made about the losses of wood incurred during storage because of a great variety of interconnected factors. However, it can be assumed that wood losses of at least 1 % occur for each month of storage. Considerably higher values are observed for hardwoods than for softwoods, and the degradation of sapwood is greater than that of heartwood. The loss in digestion yield resulting from the decay of wood during storage is substantially greater. Microbial attack results not only in the loss of wood substance, but also in the reduction of the degree of polymerization of cellulose. This results in decreased pulp strengths. In addition, condensation reactions occur in lignin which lead to darkening of the wood and reduction of lignin reactivity. Hence, stronger digestion and bleaching conditions are required, which result in an increased degradation of carbohydrates.

The duration of storage of wood should be limited to two to three months to prevent storage damage. Since fungi do not attack highly moist wood, it is of advantage to store wood in a water saturated state. This applies in particular to the storage of chips. In principle, chemical treatment with wood preservatives is possible, but is now hardly considered for economic and environmental reasons.

Wood Debarking. The bark of a tree accounts for 10 – 20 % of the trunk, depending on the type of tree, its age, and the growth conditions. The bark must be removed before the

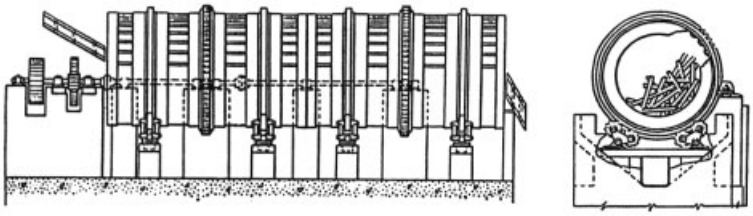

Figure 5. Drum barker

wood is cooked because it interferes with the digestion process and reduces the quality of the pulp. Bark has a higher content of lignin and extractives and a much lower content of cellulose than wood. Up to 25 % of the tissue components of hardwood barks is fiber. However, softwood barks have a fiber content of at most 5 %. The presence of bark leads to an increased consumption of digestion chemicals, and reduces not only pulp yield, but also the technological properties and the brightness of the pulp. In addition, acidic digestion processes can only degrade and dissolve bark to a limited extent. Hence, the resulting pulp contains more impurities and requires stronger bleaching conditions. Bark, especially outer bark, contains considerable amounts of colored substances which usually have a phenolic structure. Bark phenols can cause serious digestion problems in the sulfite process. They react with lignin, preventing its sulfonation and dissolution. The kraft process can cope with the largest proportions of bark because most of the bark components are quite readily soluble in alkali. The higher the quality of pulp required, the lower the content of bark that can be tolerated. The use of bark-containing wood also has the disadvantage that dirt such as sand and stones enters the pulping process, and must subsequently be removed.

The efficiency of debarking depends on the structure of the bark, the strength of bonding between the bark and the wood, and the shape of the trunk. The most important influencing factor is the bonding between the bark and the wood, which depends on the time of felling and storage influences. In most cases, wood felled during the vegetation period can be debarked in the fresh state rather easily. Wood which is felled in summer and allowed to dry in its bark can be debarked only with great difficulty. However, considerable differences are observed between different types of wood. After exceptionally long storage, the bark can be removed from the wood without additional mechanical effort because of microbial degradation. However, overstored wood should not be used for the production of pulp. The debarking of wood during periods of frost causes great problems as frozen bark can hardly be removed from the wood. Therefore, wood of this type must be subjected to steam treatment before debarking. Wood which is suitable for debarking should be as straight as possible and contain only a few depressions and knots. Before debarking, the wood is cut into segments.

In the pulp industry, drum barking is the most frequently used debarking process (Fig. 5). In this process, small short logs are continuously fed into a slightly-inclined, horizontally rotating drum. The bark is sheared off by friction between the logs and between the logs and the inner wall of the drum. The bark falls through slots in the drum wall onto a conveyor belt below the drum. Although the spraying of water improves debarking, this process is usually not employed today because the resulting wastewater is polluted and the heating value of the bark is reduced because of the increased moisture content. Drum barkers have diameters of 3 – 5 m, lengths of up to 30 m, and rotate at 5 – 10 rpm. Since the wood is subjected to intense mechanical stress in the drum, losses occur, but these do not normally exceed 3 % for normal wood. However, losses are much higher if the wood has been damaged by fungi or is very thin.

Ring barkers are normally used only for barking large-diameter logs (Fig. 6). In this process, logs are passed through a rotating ring and scrapers are hydraulically pressed against the surface of the trunk. Wood losses of 3 – 5 % can occur and, in addition, the barking efficiency for strongly adhering bark is frequently inadequate.

Knife barking machines and hydraulic barkers are no longer used in the debarking of wood. The output of knife barkers is too low and wood losses too high, while hydraulic barkers require too much energy and cause wastewater problems.

The debarking efficiency should be in the range of 85 to 95 %, depending on the digestion process used.

The bark obtained from the debarking process is usually utilized, either directly or after press dewatering, for the production of power in a bark combustion furnace.

The debarking of chips has been the subject of intensive studies, but the results were unsatisfactory. The debarking of chips is still an unsolved problem.

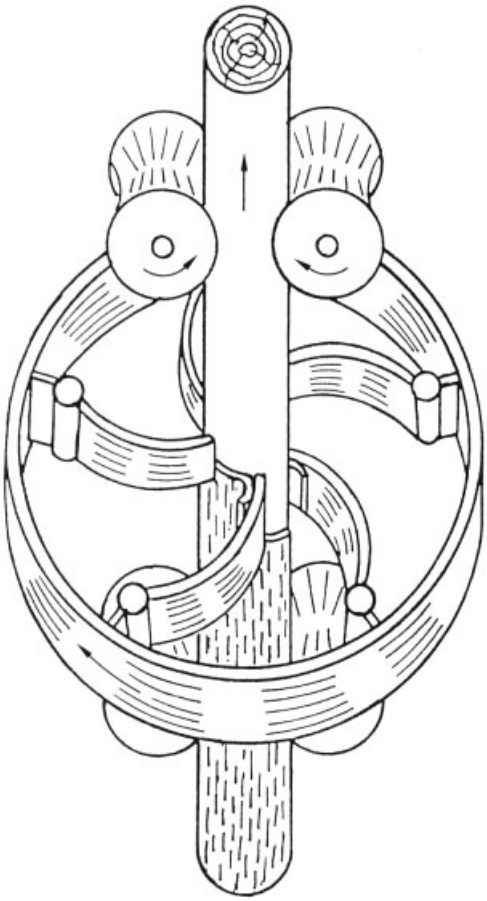

Figure 6. Ring barker

Production of Chips. In pulp production, the digestion of wood requires homogeneous impregnation with digestion chemicals. Therefore, chipping of wood is necessary. However, in the mechanical production of chips the wood can be damaged in various ways, and the resulting fiber damage influences the pulp quality. The smaller the chips, the more often the fibers are cut. The extent to which chipping reduces the average fiber length also depends on the average native length of the wood fiber. The smaller the average fiber length of the wood, the smaller the chips can be cut without reducing the average fiber length. Therefore, the chip size is always a compromise between the demand for chips that are mechanically as undamaged as possible and the necessity of achieving the quickest and most extensive impregnation with the digestion liquor. Moreover, the chips produced should have as homogeneous a size distribution as possible. Optimally dimensioned chips are 25 – 35 mm long and 3 – 7 mm thick.

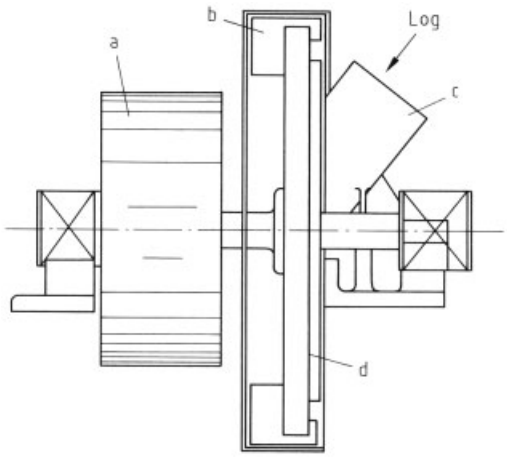

Figure 7. Schematic of a disk chipper
a) Flywheel; b) Crushers; c) Spout; d) Disk with knives

The machines used most frequently are disk chippers (Fig. 7). They have a disk diameter of 1.2 – 4.5 m, with chopping knives fastened in holes in the disk. The number of knives per disk varies between 4 and 14 and the speed between 300 and 900 rpm. The wood is transported towards the rotating disk at an angle of about 45 ° to the longitudinal direction of the wood, and chopped into slices. Due to the wedge-shaped form of the knife and the position of the knife

with respect to the cutting surface, the knife blade applies a shearing force in the longitudinal direction of the wood which results in a predetermined chip thickness. The length of the chips depends on how far the knife projects out of the chopping disk. The cut slices are flung outwards, and are then disintegrated into chips by a crusher fastened at the periphery of the chipping disk. Disk chippers have a capacity of more than 100 m^3 of wood per hour.

Drum chippers (Fig. 8) are used for chopping round wood and sawmill scraps in particular. The knives of a drum chipper are arranged on a rotating drum. They penetrate the wood from above at an angle of about 40 °; during chopping, the cutting angle changes constantly because of the rotary motion. The wood is transported into the chopper at right angles to the drum axis. Depending on the cutting angle, the chips have surfaces that are compressed to a greater or lesser extent. The proportion of fines is high.

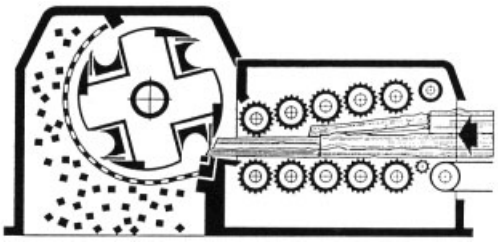

Figure 8. Schematic of a drum chipper (Courtesy of Pallmann Maschinenfabrik, Zweibrücken, Germany)

A development that has considerably influenced sawmill techniques in the last years is the headrige cutter (Fig. 9). By using two parallel cutter heads of greatly varying shape, chips can be cut directly out of the trunk surface, resulting in a beam with two smooth cut surfaces. The knives can be arranged on a conical cutter head in spirals or rings. The space between the cutter heads can be varied, depending on the dimensions of the log to be cut. The knives penetrate the wood at an angle of 35 – 40 °. As in the drum chipper, the cutting angle changes constantly due to the rotation of the cutting heads. In the final phase of cutting, the chips are removed from the wood parallel to the fiber direction, resulting in little fiber damage. The design of the knives determines both the length of the chips and their thickness, so that the production of chips with a homogeneous size distribution is possible.

The CCL chipper functions in a similar manner to the cutting heads of the headrige cutter. The two cutting heads are fitted together to form a V-shaped opening on the surface. The trunk is transported horizontally towards the rotating disk, and the knives cut in the same way as in the headrige cutter, except that the entire log is cut. With well-dimensioned wood and a constant cutting geometry, very homogeneously dimensioned chips having little fiber damage can be produced.

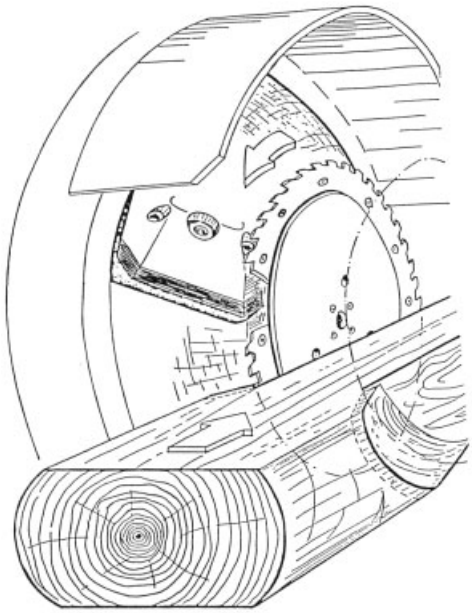

Figure 9. Schematic of a headrige cutter (Courtesy of Linck Holzbearbeitungstechnik, Oberkirch, Germany)

In chip production, not only are fibers shortened due to cutting, but damage also occurs within the chips. This damage includes primarily cracks in the middle lamella, and cracks and breaks in the inner, cellulose-rich cell wall layers due to the slipping of the parallel fibril layers. The latter, in particular, represents irreversible damage and allows the digestion chemicals to attack the carbohydrates directly. This results in poorer yields and strength losses. In thin-walled springwood fibers, damage is more pronounced than in the more stable summerwood fibers. This mechanical damage is particularly evident in

acid sulfite digestion. It can reduce the strength of sulfite pulps by 10 – 20 %. Fewer negative effects are observed in alkaline digestion.

Chips must be screened after production. Oversized and fine components are normally separated by rotating or vibrating screens (Figs. 10 and 11). Oversized chips are rechipped and the sawdust fraction is used for another purpose, e.g., production of chipboard and fiberboard. The entire course of wood processing is shown in Figure 12.

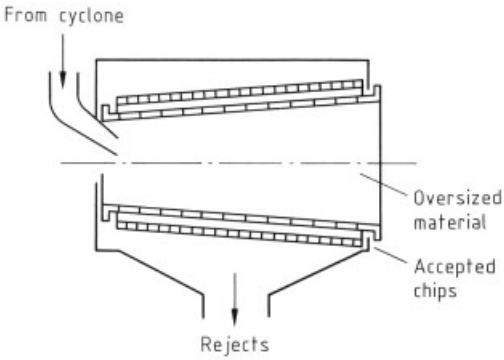

Figure 10. Rotary screen

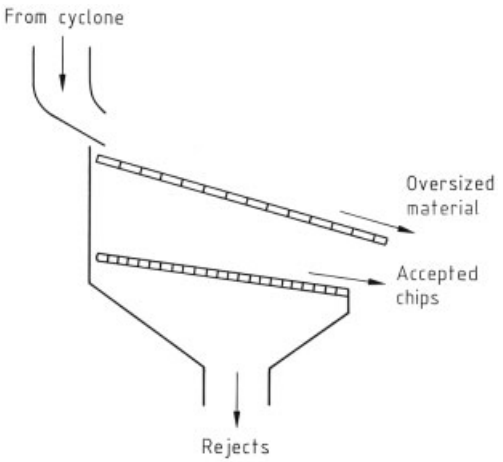

Figure 11. Vibratory screen

1.2.2. Processing of Annual Plants

The processing of annual plants for pulp production poses specific problems [21]. Harvesting is limited to a few weeks a year; hence, a pulp mill requires an extensive store capacity to ensure an all-year supply. As a result of their morphological and chemical structure, most annual plants are attacked rapidly and intensively by microorganisms. To minimize degradation, these plants should be stored as dry as possible or in a water-saturated state. Straw and strawlike annual plants are usually stored as bales.

It is essential that the plants be thoroughly washed before processing to remove adhering soil and other impurities. Chopping to lengths of about 4 cm is usually carried out in a chipper.

Plant material of high pith content, especially sugarcane, must be subjected to pith removal because the parenchyma pith cells are unsuitable for papermaking. Pith removal can be achieved either in the moist or in the dry state. Wet pith removal in hammer mills and screens is more intensive, but pollutes the wastewater considerably. In the dry process, the starting material is crushed in disk mills and the pith is removed with screens or by air separation.

1.3. Pulping Processes

At elevated temperatures lignin, a large proportion of the hemicelluloses, and some of the cellulose are dissolved by the action of pulping chemicals on wood. The fiber yield is 45 – 90 %, depending on the intensity of the chemical action. The dissolution of lignin weakens the fiber matrix. The fibers can be separated easily after about 40 – 45 % of the wood substance has been dissolved. This stage is depicted in Figure 13 for the pulping of spruce chips.

The corresponding pulps are called chemical pulps. If the pulping process is continued and if in addition to lignin, the hemicelluloses are also dissolved extensively, pulp yields of less than 45 % are obtained, but the pure cellulose content is high. Pulps of this type are processed further into chemical products. Very mild pulping conditions result in high yield pulps or semichemical pulps. The yields are in the range of 60 – 90 %, and the defibration point is not reached during digestion, i.e., the bonding substances in the wood are not rendered sufficiently soluble to disrupt the fiber binding and separate the cells from one another. Therefore, mechanical defibration must follow digestion. A survey

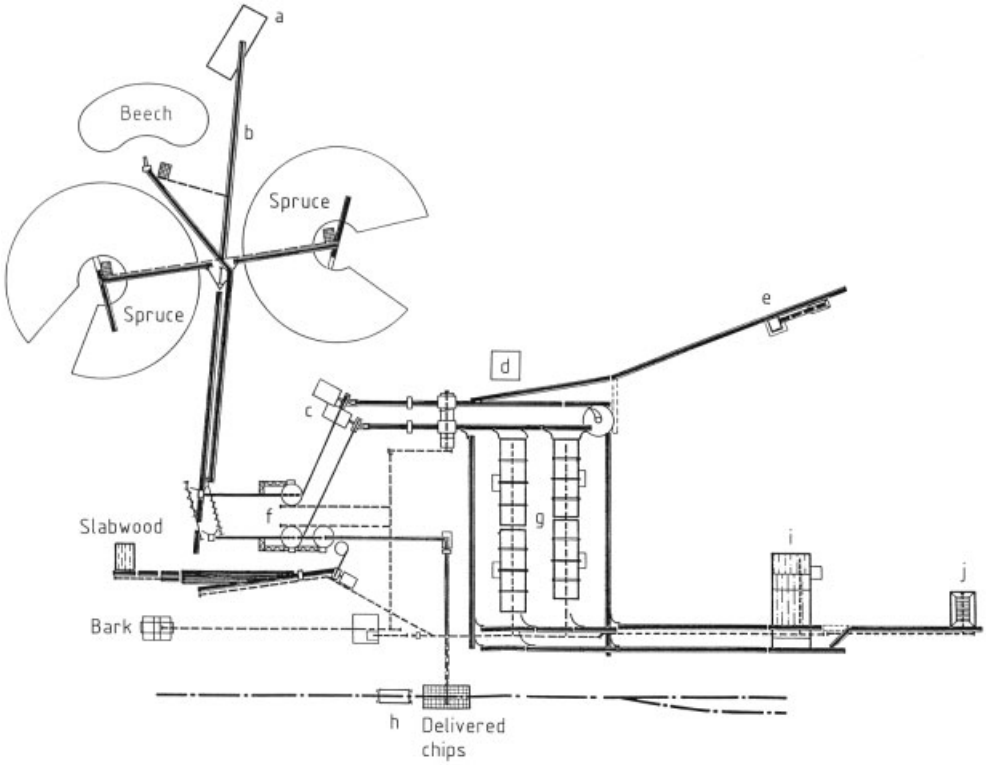

Figure 12. Schematic of wood processing (LEYKAM-Mürztaler)
a) Digester house; b) Chip storage; c) Chipper; d) Control room; e) Intermediate storage for debarked beech wood; f) Screening; g) Debarking; h) Tipping device; i) Saw; j) Log feeder

of the various processes involved in pulp production is given in Table 3.

The most important pulping processes are the sulfate or kraft process, the acid sulfite process, and the soda process. The advantages and disadvantages of the kraft and sulfite processes are listed in Tables 4 and 5, respectively.

The sulfate or kraft process has become increasingly important and is now the principal pulping process, accounting for ca. 80 % of world pulp production. In comparison, the importance of the sulfite process has decreased steadily. Today, only 10 % of the world pulp production is obtained by this method. The soda process is used primarily for the pulping of annual plants and is responsible for 5 % of the total pulp production. The remaining pulp production is based on other processes.

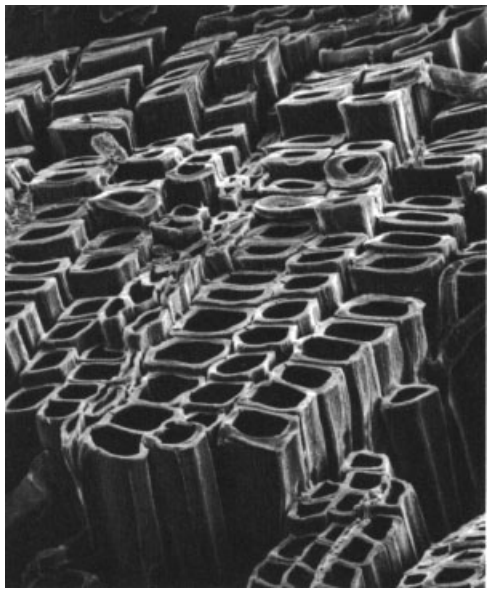

Figure 13. Electron micrograph of a spruce chip after the defibration point has been reached

Table 3. Pulp production processes

Process	Raw material	Yield, %	Brightness, % ISO	Breaking length[a], m	Tear strength[a], cN	Uses
High-yield neutral sulfite	softwood	85–93	60–80	6000–7000	80–100	wood-containing printing paper tissue, fluff
Neutral alkaline sulfite		85–90	60–80	4000	60	writing and printing paper tissue
Neutral sulfite, semi-chemical	hardwood	70–80	60–80	7000	80	fluting for corrugated board
Sulfate, semichemical	softwood	55–70	30	8000	150	liner for corrugated cardboard
Sulfite, chemical	softwood	48	90	7000	70	wood-containing papers and boards
	hardwood	43	90	5500	60	wood-free papers
Kraft	softwood	43	88	9000	130	wood-free and wood-containing papers and boards
	hardwood	50	90	8000	90	wood-containing and wood-free papers and boards
Soda	straw	43	90	6000	45	wood-free fine papers

[a] Strength values based on a degree of beating of 30 °SR.

Table 4. Advantages and disadvantages of kraft pulping

Advantages	Disadvantages
Universal raw material basis	Small process flexibility
High insensitivity to bark	Low yields for softwoods
Fast pulping process, short cover-to-cover times	High residual lignin content and poor bleachability of the pulps
High yields for hardwoods	High bleaching chemicals demand
Good pulp strength	Indispensable use of chlorine-containing bleaching agents; high water pollution
Low extract content of the pulps	Offensive smell due to volatile reduced sulfur compounds

Table 5. Advantages and disadvantages of sulfite pulping

Advantages	Disadvantages
High process flexibility	Limited raw material basis
High yields for softwoods	High requirements on raw material quality, especially high sensitivity to bark
Good bleachability	Poor pulp strength
Low water pollution	Low pulping yields for hardwoods
Possible utilization of dissolved carbohydrates	Air pollution by SO_2 emission

1.3.1. Alkaline Pulping Processes

1.3.1.1. Kraft (Sulfate) Pulping

The sulfate or kraft process exhibits a higher selectivity of delignification and, for this reason, has almost completely replaced the older soda or sodium hydroxide pulping process. Only in the case of annual plants that are easily cooked has the soda process been able to assert itself because the sulfate process offers no distinct advantages [22].

The predominant position of the kraft process is, above all, due to its applicability to almost all lignocellulose-containing raw materials, its low sensitivity to variations in the quality of raw materials, and the excellent technological properties of the resulting pulps (see also Table 4). Furthermore, the original disadvantages of the kraft process, such as poor bleachability of the pulps and complicated recovery of chemicals, have now largely been eliminated.

The terms "sulfate" and "soda" process are derived from the make-up chemicals, sodium

sulfate and sodium carbonate, which are added in the recovery cycle to compensate for chemical losses.

Pulping Conditions. The digestion liquor (white liquor) used in the kraft process consists of sodium hydroxide and sodium sulfide as active pulping chemicals, as well as sodium carbonate as residue from the causticizing step. Furthermore, small amounts of Na_2SO_4, Na_2SO_3, and $Na_2S_2O_3$ from side reactions are present in the pulping liquor. The following definitions are used to characterize white liquor, sodium equivalents expressed as Na_2O or NaOH being used to calculate the chemicals employed.

$$\text{Titratable alkali} = NaOH + Na_2S + Na_2CO_3$$
$$\text{Active alkali} = NaOH + Na_2S$$
$$\text{Effective alkali} = NaOH + 1/2\, Na_2S$$
$$\text{Sulfidity, } \% = 100 \times \frac{Na_2S}{NaOH + Na_2S}$$
$$\text{Causticizing efficiency, } \% = 100 \times \frac{NaOH}{NaOH + Na_2CO_3}$$
$$\text{Degree of reduction} = 100 \times \frac{Na_2S}{Na_2S + Na_2SO_4}$$

The sodium sulfide added is hydrolyzed almost completely to sodium hydroxide and sodium hydrogen sulfide

$$Na_2S + H_2O \rightarrow NaOH + NaSH$$

and, therefore, only half of the Na_2S is effective available alkali.

In comparison with the sulfite process (see Table 5), the kraft process is a very uniform method that exhibits little flexibility and is well suited to continuous production of a given quality of pulp. In particular, pH adjustments over a wide range are not possible. As a result of the large amounts of sodium hydroxide used, the pH value at the start of the digestion is between 13 and 14. It decreases continuously during the course of digestion because organic acids are liberated from lignin and carbohydrates during the pulping reaction.

The amount of chemicals required for pulping, their composition and the pulping parameters to be applied depend on the type of raw material used, the quality of pulp desired, and especially on the extent of delignification required. The production of semichemical or high-yield pulps requires between 10 – 15 % of active alkali, chemical pulp based on hardwoods requires 18 – 22 %, and the corresponding softwood pulp 20 – 25 %, calculated in each case as NaOH. The amount of sodium solution used should be calculated such that the pH does not fall below 10, because otherwise the dissolved lignin would reprecipitate on the pulp fibers. Nevertheless, a drop in pH towards the end of the digestion process is, as a rule, desired because then dissolved xylan fragments reprecipitate on the cellulose fibers [23–26].

A very important pulping parameter is the sulfidity of the white liquor. The optimal sulfidity depends on the raw material, the amount of chemicals used, and the pulping temperature. Since the Na_2S used in the kraft process improves the selectivity of pulping by accelerating the delignification reactions without simultaneously enhancing the dissolution of carbohydrates, pulping time can be shortened by maintaining a high sulfidity. This raises the pulp yield, and the higher hemicellulose content increases the tensile strength of the pulp. However, using more sodium sulfide leads to more odors due to the formation of reduced sulfur-containing gases, and increases the corrosiveness of the digestion liquor. A sulfidity of 25 – 40 % is normally applied in softwood pulping, while a lower sulfidity in the range 20 – 30 % is chosen for hardwoods.

Kraft pulping is usually carried out at 165 – 175 °C. The digestion time at the maximum temperature is in the range of 1 – 2 h, depending on the type of wood, the extent of delignification desired, and the digestion temperature. The heating of the digester can be carried out relatively quickly because, as a rule, the impregnation of the chips poses no problem due to the strong swelling of the fiber walls under the strongly alkaline conditions. The liquor to wood ratio influences the intensity of pulping through the chemical concentration. Depending on the chip size and the packing density of the chips in the digester, the pulping process (continuous or batch) etc., the liquor to wood ratio is adjusted to 2.5 – 4.5. A higher ratio favors uniform impregnation of the chips but reduces the delignification rate. As a result of the poor bleachability of kraft pulps and the pollution caused by pulp bleaching, efforts are being made to continue digestion to give the lowest possible kappa number so as to achieve an easier bleachability of the pulp. The kappa number is a measure of the residual

lignin content of the pulp. The amount of lignin that remains in the pulp can be determined approximately by multiplying this number by the factor 0.15 for sulfate pulp and 0.165 for sulfite pulp.

Chemical pulps made from conifers (spruce and pine) are cooked to kappa numbers of 25 – 30, giving yields between 45 and 47 %. In comparison, hardwoods give yields of >50 % when delignified to kappa numbers of 15 – 20.

Pulping Reactions. Lignin is a macromolecule based on phenylpropane units, three-dimensionally cross-linked through various ether and C–C bonds (→ Lignin). The dominant bonds are α- and, above all, β-arylether bonds. A large proportion of these bonds must be cleaved in the pulping process to achieve the fragmentation and subsequent dissolution of lignin. Compared with the soda process, delignification is accelerated in the kraft process by the presence of strongly nucleophilic hydrogen sulfide ions, and the extent of lignin breakdown is increased, especially in the case of softwoods. Phenylpropane units with free phenolic OH groups are very quickly attacked. The initial step is the formation of a quinone methide structure, which already results in lignin degradation by cleavage of α-arylether bonds. The HS^- ion (OH^- ion) can subsequently add to the quinone methide intermediate and, as a strong nucleophile, can initiate the cleavage of an arylether bound to a neighboring β-C atom (Fig. 14). Phenolate anions formed by this cleavage reaction favor the continued breakdown of lignin.

Even β-arylether bonds in nonphenolic lignin units are cleaved, albeit more slowly, if an adjacent ionized hydroxyl group was originally present (OH groups in α- or γ-position) or were formed during the pulping reaction. During kraft pulping, some of the methoxyl groups of the phenolic ring are split off to give methyl mercaptan (CH_3SH), which can be converted subsequently to dimethyl sulfide (CH_3SCH_3) or dimethyl disulfide (CH_3SSCH_3). These compounds are responsible for the unpleasant odor surrounding kraft pulp mills. Apart from the cleavage of arylether bonds, carbon – carbon bonds are also cleaved to some extent. The degradation of lignin by the pulping chemicals competes with condensation reactions caused by internal nucleophiles, especially in the final phase of digestion. These reactions counteract further delignification by forming C–C bonds and greatly reduce the bleachability of the pulp produced [27–36].

The polysaccharides in the raw material are also attacked and degraded under the conditions of alkaline pulping. The carbohydrate losses incurred and the reduction of the degree of polymerization are caused essentially by two reactions: the depolymerization reaction and alkaline hydrolysis. The former is primarily responsible for yield losses, while the latter mainly causes a reduction of the chain length. In particular, the hemicelluloses are subject to degradation reactions because of their amorphous structure and their lower degree of polymerization. Galactoglucomannans and glucomannans are dissolved much faster than the fairly alkali resistant xylans [37], [38]. For this reason, the xylan-rich hardwoods give higher yields in alkaline pulping than the low-xylan conifers.

The acetyl side groups of the xylans of hardwoods and the galactoglucomannans of softwoods are hydrolyzed at the very start of digestion. The primary depolymerization reaction begins above ca. 100 °C. In this process, sugar units are successively cleaved in a β-alkoxy elimination reaction, commencing at the reducing end of the carbohydrate chains [39–42]. An average of 50 – 60 units are cleaved from the cellulose chain molecules until depolymerization is stopped by competing stopping reactions and formation of an alkali-stable carboxyl group at the end of the chain [43–45]. Chain scission reactions occur at temperatures above ca. 150 °C due to alkaline hydrolysis, which lowers the degree of polymerization of cellulose [46–48]. The ends of the chains, exposed by the hydrolytic cleavage of glycosidic bonds, are the starting point for the secondary depolymerization reaction. With increasing digestion time and more extensive lignin dissolution, the carbohydrate degradation reactions dominate and, because of the reduced selectivity of the pulping reaction, digestion must be terminated to avoid excessive losses of yield and pulp strength.

The carbohydrate losses caused by the depolymerization reactions can be limited by reduction or oxidation of the aldehydic end groups of the cellulose to give alkali-stable alcohol or carboxyl groups [49–52]. Although reduction by the addition of 1 – 2 % sodium borohydride,

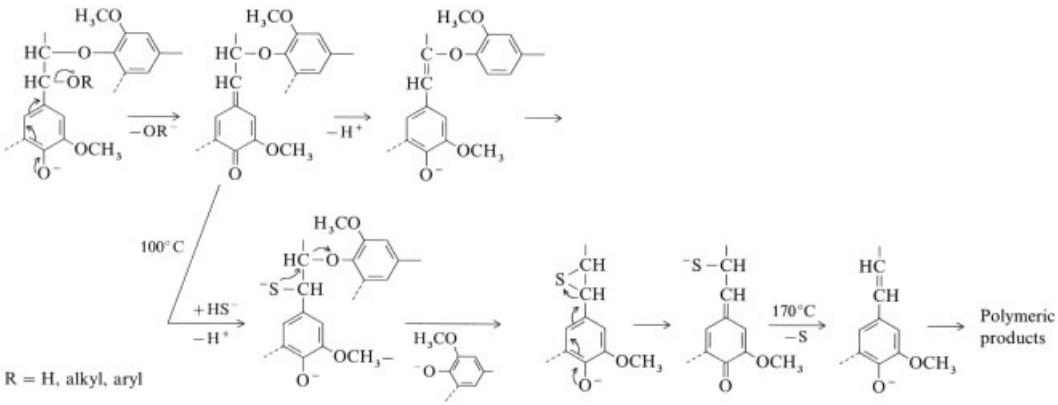

Figure 14. Main reactions of lignin during soda and kraft pulping

based on o.d. (oven dry) raw material, produces up to 6 % higher pulp yields [53], [54], the high cost make this process uneconomical. In comparison, oxidation by the addition of 3 – 4 % polysulfide is used on an industrial scale in some cases. A 6 % increase in yield can be obtained by polysulfide digestion, which, however, also leads to an increased emission of odorous sulfur compounds [55–59]. The use of anthraquinone (AQ-catalyzed pulping) stabilizes carbohydrates against degradation by oxidizing the end groups.

Anthraquinone-Catalyzed Pulping. Anthraquinone (AQ) which, when used in catalytic amounts of 0.05 – 0.2 % based on o.d. wood in alkaline pulping, causes efficient stabilization of carbohydrates and simultaneous acceleration of the delignification reaction [60], [61]. The addition of 0.05 % AQ to softwood pulping increases pulp yield at a given kappa number by 2 – 3 % [62], [63]. Although various AQ derivatives are also effective catalysts for alkaline pulping, they are less cost effective than anthraquinone [64].

The effect of anthraquinone is based on the AQ/AHQ (AHQ = anthrahydroquinone) redox system. Already in the initial phase of pulping, anthraquinone oxidizes the reducing end groups of the carbohydrates to give aldonic acid end groups and thus stabilizes them towards degradation by the depolymerization reaction. The monoanion of the anthrahydroquinone produced in this reaction can add to the α-C atom of the lignin quinone methide intermediate. AHQ monoanion can initiate the cleavage of β-arylether bonds in the same way as HS^- ions. AHQ is oxidized in this redox reaction, regenerating AQ for further redox cycles [65–70].

Anthraquinone was first used on an industrial scale in 1977. In the meantime it has been employed in more than 70 pulp mills on a temporary or permanent basis [71], [72]. Since anthraquinone is relatively expensive and cannot be recovered because it is decomposed extensively in the pulping process, the decision to use AQ depends on the wood costs, the pulp prices attainable, and on the specific production conditions. In principle, many different objectives can be pursued with the use of AQ in the sulfate process:

1) Increase in pulp yield and therefore enhanced production in the case of limited capacity for evaporation or combustion of waste liquor
2) Reduced wood consumption per tonne of pulp
3) Energy savings by use of lower pulping temperatures or shorter digestion times
4) Reduced pulping chemicals demand per tonne of pulp
5) Reduction of sulfidity in order to reduce the formation of volatile, odorous sulfur compounds
6) Increase in the extent of delignification in order to improve the bleachability and reduce water pollution caused by bleaching effluents

In comparison to kraft pulping, a far greater effect can be achieved by the use of anthraquinone

in the soda process. During lignin degradation, AHQ acts in a similar way to hydrogen sulfide ions. Therefore, the rate and extent of delignification is increased to a much greater amount by the addition of AQ to the soda process. In fact, the application of the soda AQ process permits the production of bleachable chemical pulp from softwoods. The same applies to alkaline and semialkaline sulfite pulping (see Section 1.3.2.4) [73–78].

Prehydrolysis Kraft Process. The chemical industry requires dissolving pulps with a high α-cellulose content for processing into cellulose products. If the kraft process is used, the wood hemicelluloses must be degraded by acid hydrolysis prior to digestion because the alkaline pulping conditions stabilize the hemicelluloses remaining in the pulp towards further alkaline attack. In contrast, the hemicelluloses in sulfite dissolving pulps can be removed from the fiber matrix by an alkaline purification step. Prehydrolysis can be carried out by using mineral acids (e.g. 0.25 – 0.5 % H_2SO_4 based on o.d. wood) at 120 – 140 °C, or by steam or water treatment at 160 – 170 °C. In the latter case, acetyl groups are split off from the xylans of hardwoods or the galactoglucomannans of softwoods, resulting in a pH of 3 – 4. Whereas cellulose is very stable to such acid treatment, hemicelluloses are broken down into smaller, water-soluble chains. Bonds between hemicelluloses and cellulose and lignin are hydrolyzed as well. The average degree of polymerization of the hemicelluloses is decreased by about 30 % during prehydrolysis. If softwoods are used, ca. 10 – 15 % of the organic matter is dissolved in this step, and with deciduous woods the losses are 15 – 20 %. In particular when hemicellulose-rich hardwoods are used, the large-scale utilization of the hydrolysate obtained is of interest. Dissolved sugars and their degradation products, such as furfural, hydroxy-methylfurfural, levulinic acid, acetic acid, formic acid, and methanol, are present in the hydrolysate.

The subsequent kraft pulping is carried out analogously to the process used in the production of paper pulps. The digestion conditions depend primarily on the required α-cellulose content of the pulp. The hemicelluloses still present in the wood after prehydrolysis are subject to intensive degradation, especially through the depolymerization reaction that can commence at the numerous carbonyl groups formed during prehydrolysis.

Softwoods, and hardwoods, and low-ash annual plants are suitable raw materials for prehydrolysis kraft pulping. Hardwoods have a higher hemicellulose content than softwoods and usually give lower yields, but they can be much more easily and extensively delignified. Also, the risk of lignin condensations occurring during prehydrolysis is far lower. These condensations impede the subsequent pulping and bleaching steps. The conditions for the prehydrolysis of hardwoods and softwoods are summarized in Table 6 [79–85].

1.3.1.2. Soda Pulping of Annual Plants

The most important annual plants for the pulp industry are agricultural waste materials (bagasse and straw) and naturally occurring or cultivated bagasse and reeds. Cultivated fiber plants, such as kenaf, sisal, jute, hemp, and cotton, are valuable raw materials for the production of specialty pulps [16], [21], [86–88].

For the pulping of annual plants, the soda process is usually employed. The kraft and neutral sulfite processes are used less frequently. The acid sulfite process is not applied because it produces brittle pulps with a high ash content and inadequate strength properties.

Since annual plants are impregnated easily and have a low content of reactive lignin, the amount of chemicals required is lower than in wood pulping. In soda pulping, 10 – 15 % NaOH, based on the raw material, is normally employed at a pulping temperature of 160 – 170 °C. The yields attainable are in the range 40 – 55 % and are influenced greatly by the type and quality of the raw material, especially by the lignin content and the proportion of parenchyma cells. A high parenchyma content of the raw material not only reduces the yield, but also increases alkali consumption and diminishes the dewatering properties and strength of the pulp by increasing the formation of fines. For this reason, bagasse is subjected to a depithing process to eliminate parenchyma pith cells before pulping.

While various fiber plants such as hemp, kenaf, jute, and sisal give high strength pulps,

Table 6. Prehydrolysis conditions for hard- and softwood

	Charge, %/o.d. wood	Temperature, °C	Time, h
Prehydrolysis with mineral acids			
Hardwood, H_2SO_4	10 – 30	70 – 80	0.5 – 3
Hardwood, H_2SO_4	0.3 – 0.5	120 – 140	0.5 – 3
Hardwood, HCl	0.5 – 0.7	95	0.5 – 3
Softwood, H_2SO_4	0.25 – 0.5	130 – 140	0.5 – 1
Prehydrolysis with liberated organic acids			
Hardwood		160 – 170	0.5 – 3
Softwood		170 – 175	0.5 – 1

sugar-cane bagasse and straw pulps have lower strengths than wood pulps because of their short, thin-walled fibers. However, these pulps are commonly added in the production of fine papers to give the paper certain characteristics, e.g., a smooth, closed surface [89].

1.3.2. Sulfite Process

The sulfite process is characterized by its high flexibility. In principle, the entire pH range can be used for pulping by changing the dosage and composition of the chemicals. In comparison, kraft pulping can be carried out only with highly alkaline digestion liquor. Thus, the use of sulfite pulping permits the production of many different types and qualities of pulps for a broad range of applications (Table 7).

Today, the term sulfite process also refers to the many different variations of this process. Previously, this term was reserved for the acid (calcium) bisulfite process, which was used almost universally. The sulfite process can be subdivided into the following types of pulping:

1) acid bisulfite pulping,
2) bisulfite pulping,
3) neutral sulfite pulping,
4) alkaline sulfite pulping,
5) multistage sulfite pulping, and
6) anthraquinone-catalyzed sulfite pulping.

All these processes are based on the use of aqueous sulfur dioxide and a so-called base, i.e., calcium, magnesium, sodium, or ammonium. In the pulping system, pH-dependent equilibrium reactions occur that depend on the ratio of SO_2 to base. If sodium hydroxide is added to the aqueous sulfur dioxide, first bisulfite and then sulfite are formed:

$$H_2O + SO_2 \rightarrow SO_2 \cdot H_2O$$

$$SO_2 \cdot H_2O + NaOH \rightarrow NaHSO_3 + H_2O$$

$$NaHSO_3 + NaOH \rightarrow Na_2SO_3 + H_2O$$

A pure bisulfite solution is present at pH 4.5 and a pure unbuffered sulfite solution at approximately pH 9 (see Fig. 15). At these inflection points, a slight change in the chemical composition leads to a large pH shift [90]. As shown below, practically the entire pH range can be used for sulfite pulping with Na as "soluble" base (the same applies to NH_4).

Chemical composition	pH
$SO_2 \cdot H_2O/NaHSO_3$	1.5 – 4.5
$NaHSO_3$	4.5
$NaHSO_3/Na_2SO_3$	4.5 – 9.0
Na_2SO_3	9.0
Na_2SO_3/Na_2CO_3	9 – 12.5
$Na_2SO_3/Na_2CO_3/NaOH$	10 – 13
$Na_2SO_3/NaOH$	13 – 14

Since calcium sulfite is only slightly soluble, this base can only be used in acid bisulfite pulping, i.e., in the presence of excess SO_2; $Ca(HSO_3)_2$ precipitates at pH values above 2.0 – 2.3. In comparison, magnesium bisulfite solutions are stable up to pH 5 – 6, so that acid bisulfite as well as the bisulfite (magnefite) and multistage pulpings can be carried out. Due to the rise in the pH value with increasing pulping temperatures [91], the pH of the cold pulping liquor must be in the range of 1.5 (acid bisulfite) and 4.0 (bisulfite) to avoid precipitation in the digestion phase.

Table 7. Sulfite pulping processes

Process	pH range	Base	Active agent	Digestion temperature, °C	Digestion time, min	Pulp yield, %	Special applications
Acid (bi)sulfite	1–2	Ca^{2+}, Mg^{2+}, Na^+, NH_4^+	$SO_2 \cdot H_2O$, H^+, HSO_3^-	125–143	180–420	40–50	dissolving pulp, tissue, printing paper
Bisulfite	3–5	Mg^{2+}, Na^+, NH_4^+	HSO_3^-, H^+	150–170	60–180	50–65	printing paper
Neutral sulfite (NSSC)	5–7	Na^+, NH_4^+	HSO_3^-, SO_3^{2-}	160–180	25–180	75–90	corrugated medium
Alkaline sufite	9–13.5	Na^+	SO_3^{2-}, OH^-	160–180	180–300	45–60	kraft-type pulp
Two-stage sulfite							
Stora							
1st stage	6–8	Na^+	HSO_3^-, SO_3^{2-}	135–145	126–360		
2nd stage	1–2		$SO_2 \cdot H_2O$, H^+, HSO_3^-	125–140	120–240	50–60	greaseproof paper
Sivola							
1st stage	3–4	Na^+	HSO_3^-, H^+	140–150			
2nd stage	7–10		SO_3^{2-}	160–180	60–180	35–45	dissolving pulp
AQ–sulfite (NS – AQ, AS – AQ)	9–13	Na^+	SO_3^{2-}, OH^-, AHQ^-	170–175	150–240	55–65	kraft-type pulp

In practice, apart from the chemical composition and the pH values, the terms free, combined, and total SO_2 are used to characterize sulfite solutions. The content of theoretically base bound bisulfite and sulfite ions is referred to as combined SO_2. The difference between the total SO_2 and the combined SO_2 is then free SO_2. According to this definition, half of a pure bisulfite solution is free and the other half is bound SO_2.

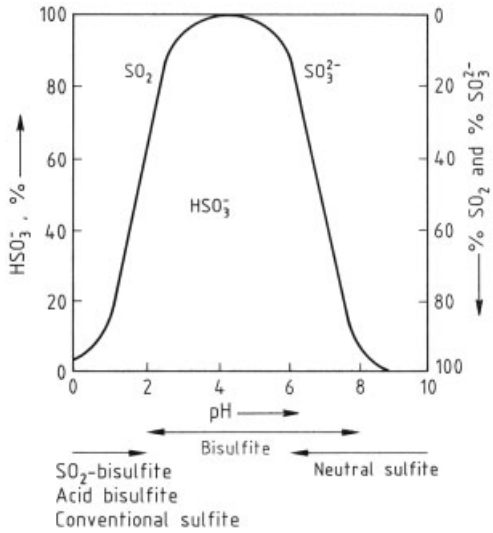

Figure 15. Composition of a dilute SO_2-H_2O solution versus pH at 20 °C

1.3.2.1. Reactions in Sulfite Pulping

In sulfite pulping, the reactivity of lignin is essentially determined by the pH of the pulping liquor. Fundamental differences exist in the mechanism, rate, and extent of delignification under acidic, neutral, and alkaline conditions. Regardless of the pH, the α-position of the quinone methide intermediate formed from the lignin phenylpropane units is always the most reactive site for attack.

Under acidic pulping conditions, aqueous sulfur dioxide is the active pulping chemical that causes extensive sulfonation, especially by addition to the α-C of the propane side chain. Bisulfite ions can also take part in the sulfonation reactions. The initial step involves the protonation and subsequent cleavage of α-

hydroxy, α-alkoxy, and α-aroxy groups leading to the formation of resonance stabilized carbonium – oxonium ions. Since these ions are very reactive, the weakly nucleophilic pulping reagents can readily add to the α-C atom (Fig. 16).

Other sulfonation reactions can occur on the γ-C aldehyde groups or on the α-C carbonyl groups of the propane side chain. The solubility of the lignin molecule is increased by the hydrophilic sulfo groups. However, the fragmentation of lignin is restricted to the cleavage of non-cyclic α-arylether bonds between the lignin units. In acid sulfite pulping, the β-arylether bonds are essentially stable due to the absence of strong nucleophiles, and methoxy groups attached to the phenolic ring are also not subject to cleavage [36], [92–94].

Condensation reactions of nucleophilic lignin fragments compete with the sulfonation reaction. Both reactions are favored by low pH values. The danger of lignin fragments undergoing recondensation to give higher molecular, partially degradable units exists, especially if the amount of combined SO_2 available at the end of digestion is not sufficient. The resulting " black digest" can be avoided by raising the bisulfite ion concentration, i.e., by increased addition of base. Phenolic extractives from the bark or heartwood constituents such as pinosylvin (3,5-dihydroxystilbene) and its methyl ether from pine heartwood, or taxifolin from the heartwood of larch and Douglas fir, readily condense with lignin and can greatly impede acid bisulfite pulping or bring it to a complete standstill.

In bisulfite pulping (pH ca. 4), these condensation reactions can be largely suppressed, and a more selective, but slower sulfonation is achieved by the bisulfite ions. The main disadvantage of bisulfite pulping is the lower degree of delignification that can be attained.

In the neutral sulfite process, the reactions of lignin are restricted to the phenolic structural units, which are converted to quinone methide structures, as in alkaline pulping. After sulfonation at the α-C atom, the neighboring β-aroxy group can be slowly cleaved sulfolytically [95]. The nonphenolic β-arylether units, however, are resistant to neutral sulfite pulping [29]. Since practically only phenolic lignin structures react, only semichemical pulps with a high residual lignin content can be produced.

In the alkaline sulfite process, which is carried out in the presence of NaOH (NH_4OH), the cleavage of α-ether bonds is also restricted to the phenolic lignin structures. Apart from the sulfolytic β-arylether cleavage described above for the neutral sulfite process, β-aryl ether (β-O-4) bonds of nonphenolic units are also cleaved to a certain extent by the action of alkali, analogous to the reaction in the soda process. Consequently, chemical pulps similar to kraft pulps can be produced by alkaline sulfite pulping.

The reactions of carbohydrates in sulfite pulping are also determined by the pH value. At high pH values, rapid deacetylation of hemicelluloses takes place and alkali-induced degradation reactions (see Section 1.3.1) occur. In the neutral pH range, carbohydrates are attacked to the lowest extent, but their susceptibility increases considerably with increasing acidity. The glucosidic bonds are readily cleaved by acid, especially under the conditions of acid bisulfite pulping. Hemicelluloses in particular are subject to hydrolytic degradation because their glycosidic bonds are more easily accessible and less resistant than the bonds of cellulose. With increasing degradation, highly depolymerized hemicelluloses dissolve and are hydrolyzed step by step until monomeric sugar building blocks are obtained. The depolymerization of cellulose leads essentially to a reduction in the degree of polymerization (DP value); only under drastic conditions are soluble cellulose fragments formed. In general, acid pulping of softwoods gives higher yields because galactoglucomannans, the dominant hemicelluloses, are relatively acid resistant, whereas the xylans occurring predominantly in hardwoods are degraded much faster [96–98].

The monosaccharides formed in pulping are not completely stable and are partially dehydrated and oxidized. Aldonic acids in particular are produced by bisulfite oxidation. The thiosulfate formed by the reduction of bisulfite can, in turn, react with reactive lignin structures and contribute to lignin condensation [99–101].

1.3.2.2. The Various Sulfite Processes

Acid Bisulfite Process. In the acid bisulfite process, pulping is performed with a large excess of free SO_2 at pH 1.2 – 1.8. All of the above-

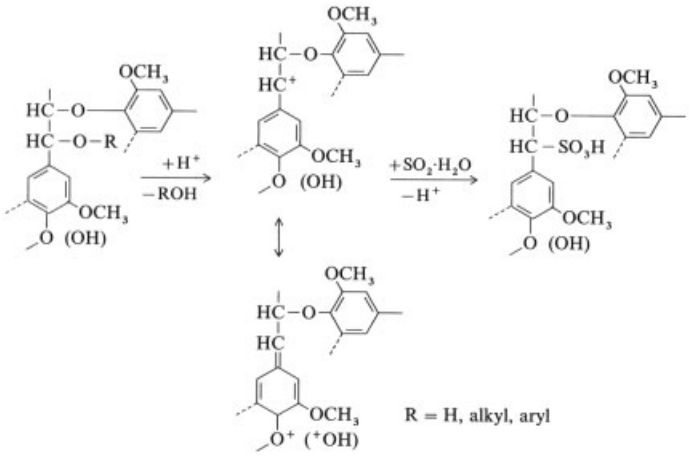

Figure 16. Reactions of lignin during acid sulfite pulping

mentioned bases can be employed. No basic differences in their mode of action exist, but the pulping conditions considered to be optimal vary slightly. The course and the result of pulping are influenced by the base selected [102], [103]. For instance, magnesium bisulfite pulping results in the best yields and the highest pulp brightness. Use of the monovalent bases Na^+ and NH_4^+ results in stronger pulps and a lower shive (fiber bundle; see section 2.4.1) content. If ammonium is used as base, the pulping process is very fast, but the pulps are dark in color and cannot be used in the unbleached state. Pulps of the lowest strength are obtained with calcium as base, and this pulping process must be controlled more carefully because calcium precipitates readily.

The total SO_2 used in acid bisulfite pulping is in the range 18 – 25 %, depending on the type of wood used, degree of digestion desired etc., and the SO_2 to base ratio is adjusted to 4.5 – 6.0. During the digestion phase, the liquor to wood ratio is usually 3 – 4 : 1. The chips are first impregnated with a higher liquor ratio to achieve a homogeneous distribution of the pulping liquor. A part of the pulping liquor (side relief) and the SO_2 gas (top relief) is then removed after impregnation and returned to the accumulator to prepare fresh digestion liquor. Pulping temperatures of 125 –140 °C are used for the production of paper pulps. Unlike bisulfite pulping, the acid bisulfite process is also well suited to the production of dissolving pulps because the hemicelluloses can be dissolved to a large extent in the highly acidic pulping process. However, higher maximum temperatures of 145 – 150 °C are employed for this process. The digestion time is usually 2 – 5 h at this pulping temperature, but can be extended to more than 7 h at lower pulping temperatures. A long heating time is also necessary because the SO_2 penetrates into the wood much faster than the base due to the high vapor pressure of the excess free SO_2. Rapid heating can result in lignin condensations due to the high acidity within the chip centers. Since digester relief must also be effected slowly so that the excess SO_2 can be reused in the subsequent digestion, the result is a longer cover-to-cover cycle of 8 – 12 h.

From its first industrial application in 1874 until the end of World War II, the calcium bisulfite process was the principal process for wood pulping used worldwide because of the low costs and good availability of limestone and sulfur dioxide as basic materials for the production of the digestion acid. Since the pulping chemicals were inexpensive, it was not necessary to recover them before discharging the sulfite waste liquor into rivers or lakes. At the beginning of the 1950s, factories in North America started using ammonium as the base and it was possible to use the resultant waste liquor as fertilizer. Today, calcium bisulfite pulping, ammonium bisulfite pulping, and other sulfite pulping processes based on ammonium are of very little importance. If there is no possibility of utilizing the waste liquor, it is burnt for reasons of environ-

mental protection and process energy production. However, base recovery is pursued neither in the calcium bisulfite process nor in the ammonium sulfite process.

The recovery of calcium fails because of the high temperatures of more than 1150 °C required to completely decompose the gypsum, formed during waste liquor combustion, to calcium oxide and sulfur dioxide. For ammonium sulfite waste liquor, semitechnical recovery processes based on ion exchange have been developed. However, these processes have not gained acceptance in industry because of high costs. Thus, recovery in both processes is limited to the portion of SO_2 that is not bound to the base, i.e., a maximum of 80 % of the sulfur dioxide charge is recoverable.

Therefore, the use of ammonium as base was only a temporary alternative. With the development of recovery processes for magnesium and sodium sulfite liquors the corresponding pulping processes have increasingly gained acceptance for economic and environmental reasons. Moreover, compared with the calcium bisulfite process, the possibility exists that improved paper pulp qualities and higher yields can be achieved and the utilizable range of raw materials can be extended by increasing the pH value of the pulping liquor.

Bisulfite Pulping. The pulping liquor used in the bisulfite process consists of approximately equal parts of free and combined SO_2, i.e., it contains no excess free SO_2. The pH value is between 3.5 and 5.5, depending on the temperature and the chemical concentration. The use of calcium as base in this pH range is not possible. Magnesium is usually employed as base (magnefite process) [104] because the recovery of chemicals is easy and inexpensive. Ammonium bisulfite [105] and sodium bisulfite pulping [106] (Arbiso process) have also found industrial applications in individual cases.

Bisulfite pulping is normally carried out with 4 – 5 % total SO_2 in the digestion liquor and a liquor to wood ratio of 4 : 1, i.e., the total SO_2 used, based on wood, is 16 – 20 %. The pulping temperature is ca. 155 °C for hardwoods and 165 °C for softwoods. The higher temperatures required, compared with the acid bisulfite process, compensate for the lower rate of delignification caused by the lower acidity of the pulping liquor. However, carbohydrates are also dissolved to a smaller extent during bisulfite pulping. Hence, these pulps have a high hemicellulose content, and yields can reach 60 %. In comparison with acid bisulfite pulping, shorter digestion times of 1 – 3 h are required because of the higher digestion temperatures. The heating period is also considerably shorter since the risk of lignin condensation is lower and, therefore, the digestion liquor can be pumped at higher temperature into the digester and heated more quickly to the maximum temperature. On the whole, the cover-to-cover cycle is much shorter (approx. 6 – 8 h), which increases the production capacity.

Since phenolic components of bark and heartwood interfere with bisulfite pulping only to a small extent, the raw material base is considerably wider than for the acid bisulfite process. In addition to spruce and fir, pine, Douglas fir, and larch can be used in the production of long-fiber pulps. Bisulfite pulps are brighter, stronger, and better suited to papermaking than the corresponding acid bisulfite pulps. However, these pulps are not suitable for further processing in the chemical industry because of their high hemicellulose contents. A serious disadvantage of the bisulfite process is that the extent of delignification attainable is lower than that obtained under highly acidic pulping conditions, since otherwise the losses of yield and strength that occur are excessively high. The higher residual lignin content of the pulp requires the use of more bleaching chemicals and hence produces a greater effluent charge. As a result of stricter environmental controls, many mills that had already switched from the acid bisulfite to the bisulfite process reversed their decision.

Neutral Sulfite Process. The NSSC (neutral sulfite semichemical) process is not suited for making chemical pulp because the degree of delignification attainable is too low. This process uses sodium as base and is widely applied in the production of semichemical pulps, especially in North America. These pulps are employed mainly for the middle layers of corrugated cardboard and in newsprint [107], [108]. Neutral sulfite pulp mills are sometimes used in combination with kraft mills, which allows the combined recovery of chemicals (cross recovery). The NSSC waste liquor can be used as

make-up chemical to compensate for losses of chemicals in the kraft process.

The neutral sulfite process is a two-stage process with a chemical delignification step followed by mechanical defibration. In the chemical treatment, neutral to slightly alkaline sulfite solutions are used to dissolve hemicelluloses and lignins and, thus, weaken the bonds between the fibers. In this way, the mechanical effort required to expose fiber bundles and individual fibers is reduced considerably.

Hardwood is the raw material of choice; it can be digested easily and gives bright pulps with high stiffness and crush strength. Annual plants, such as straw and bagasse, and softwoods can also be used. Conifers produce pulps of good quality, but the chemicals requirement is higher and more energy is needed for defibration and subsequent beating.

In the neutral sulfite pulping of hardwood, between 8 and 18 % of sodium sulfite or ammonium sulfite is used, based on o.d. wood. To maintain the pH in the range 7 – 9, the digestion liquor must be buffered by the addition of 2.5 – 5 % of alkali ($NaHCO_3$, Na_2CO_3, or NaOH) because acids are released during digestion, particularly acetic acid by the cleavage of acetyl groups from hemicelluloses. The digestion time is 30 – 180 min at the maximum temperature (ca. 170 °C). The reaction proceeds very slowly and only 25 – 50 % of the lignin and about 30 – 45 % of the hemicelluloses are dissolved. Accordingly, high yields of about 65 – 75 % are obtained for deciduous woods and up to 85 % for conifers.

Alkaline Sulfite Pulping. Alkaline sulfite pulping is carried out with liquors containing sodium sulfite and sodium hydroxide (sodium carbonate) at initial pH values of 9 – 13. When, as in NSSC pulping, a low chemical charge, low alkalinity, and low temperature are applied, semichemical pulps can be prepared, especially from annual plants. The boundaries between the two processes are fluid. If the pulping liquor consists of approximately equal amounts of sodium sulfite and NaOH, and if strongly alkaline conditions (pH >13) are chosen, chemical pulps can be produced as well. This pulping exhibits characteristics typical of both the sulfite process (no odors due to formation of reduced sulfur compounds; good bleachability of pulps) and the kraft process (no raw material restrictions, high pulp strengths, low yields, and low brightness) [109]. However, the higher chemicals requirement, longer pulping times, higher residual lignin content, and the more complicated chemicals recovery, compared to the kraft process, prevented establishment of this process in industrial practice. However, the especially pronounced effect of anthraquinone in alkaline sulfate pulping may lead to renewed interest in this process.

1.3.2.3. Multistage Sulfite Processes

The various modifications that can be achieved in the sulfite pulping system by adjustment of the pH can be advantageously combined by changing the pH during pulping. In this way, it is possible to produce special pulps that are ideally suited to certain fields of application. Sodium, as a soluble base, is particularly suitable for multistage digestions of this type. Magnesium can be used as base within limits. In the two-stage systems, pulping is either carried out first at a higher pH and the pH is subsequently decreased or vice versa. The Stora process, applied for the first time in the Skutskär factory of Stora-Kopparsberg, has achieved commercial importance [110], [111]. In this process, a sodium bisulfite – sulfite solution at pH 6 – 8 is used in the impregnation and first digestion stages. Subsequently, excess pulping liquor is removed and the pH is adjusted to ca. 1.5 by introducing SO_2. This is followed by normal acid bisulfite pulping as the main delignification step. The reactive centers of lignin are sulfonated practically completely in the first stage, and thus condensation reactions with phenolic extractives under strongly acidic conditions are blocked, so that even pine heartwood, for instance, can be digested satisfactorily by this method. Furthermore, glucomannan stabilization takes place in the bisulfite – sulfite step such that the acetyl groups are cleaved first, allowing the glucomannans to attach themselves firmly to the cellulose fibrils by increased formation of hydrogen bonds [112–114]. The resistance to acid hydrolysis is thereby increased. However, this stabilization effect is restricted essentially to softwoods, with their high galactoglucomannan content. An increase in yield of over 5 % can be achieved in this manner. The very bright, easily bleachable pulps

are specially suited to the production of greaseproof paper because of their high hemicellulose contents and their resultant fast beatability.

A corresponding process can be carried out with magnesium as base [115]. The term FB process results from the very fast beatability, a characteristic property of these pulps. Analogous to the process based on sodium, pretreatment in a slightly acidic to neutral medium is followed by pulping at a highly acidic pH. As a result of the limited solubility of magnesium sulfite, the pulping liquor is pumped into the digester at pH 5 for the first step. The required pH value of 5.5 – 6 is then adjusted in the digester by the direct addition of magnesium hydroxide. In the second pulping stage, the pH is lowered by addition of SO_2.

The second type of process involving an increase in the pH in the second pulping stage is the Sivola (Rauma) process [116–118]. In a two-stage procedure, a pH 3 – 4 sodium bisulfite solution is applied in the first stage, and the pH value is subsequently adjusted to 7 – 10 with sodium hydroxide. A three-stage variant of the Sivola process, with a somewhat higher pH in the first stage, an intermediate acid bisulfite stage, and a final neutral or slightly alkaline stage is especially suited to extractive-rich wood species, such as pine. This method can be used to produce dissolving pulps with a particularly high α-cellulose content [119], [120]. Dissolving pulps can also be produced by the two-step process, and it is also possible to make many different paper pulps by altering the pulping conditions. Hemicelluloses can be dissolved to a great extent by using a high pH in the second stage. The tear strength of the pulps thus obtained rivals that of kraft pulps. On the other hand, the hemicelluloses are largely retained in the pulp by second-stage digestion under neutral conditions, so that higher yields are achieved.

A corresponding two-stage process based on magnesium is called the HO process (high opacity). Impregnation with a bisulfite solution (pH 3.8 – 4) is carried out in the first stage. The pH is then adjusted to 6.5 – 7 by the injection of magnesium oxide milk and digestion is completed at a high pulping temperature (170 °C) [121]. Since dissolved organic acids apparently stabilize the pulping liquor, no precipitation occurs in spite of the high pH.

1.3.2.4. Anthraquinone-Catalyzed Sulfite Pulping

An exceptional feature of the sulfite pulping system is that the catalytic activity of anthraquinone is also exerted at slightly alkaline to neutral pH values, whereas effective acceleration of delignification in soda pulping requires a pH above 13 [122], [123]. This explains how sulfite pulping in a moderately alkaline pH range, which would otherwise permit only inadequate delignification, can be used for the production of chemical pulps if small amounts of AQ (0.1 – 0.2 %) are added [124–127], [73–78]. The pulping liquor used in this process consists predominantly of sodium sulfite. Sodium hydroxide solution and/or sodium carbonate is added as additional alkali. The alkali ratio (sodium sulfite/total alkali) is preferably 0.8 – 0.9. In comparison with the AS – AQ process with NaOH (pH 10 – 13.5), the term NS – (neutral sulfite) AQ process is frequently used when sodium carbonate is employed as the buffer below pH 10. The transition is, however, fluid and both process variants require alkaline conditions (final pH 7 – 10.5).

This pulping procedure is faster and more selective than alkaline sulfite pulping with approximately equal sodium sulfite and sodium hydroxide charges and without AQ. Better delignification is achieved in spite of the lower chemicals demand (20 – 25 % calculated as NaOH/o.d. wood) and shortened digestion times (150 – 240 minutes).

AQ-catalyzed sulfite pulping combines the advantages of the sulfite and kraft processes. First, there is no raw material restriction and high pulp strengths are achieved. Second, high yields are obtained, the pulps are easily bleachable and the emission of malodorous sulfur compounds is avoided. The main disadvantage of this process is the relatively high residual lignin content in the pulps, which have a high chemicals demand for bleaching and require the use of chlorine-containing bleaching agents.

1.3.3. Other Pulping Processes

Environmental problems and the high investment costs for pulp mills (€1500 – 2000 t^{-1} a^{-1})

have led to the development of new pulping processes. The aim is to develop a process that can produce high strength pulps with good yields from all lignocellulose-containing raw materials that are suited to fiber production. The lignin content of these pulps must be low and the residual lignin must be highly reactive to permit easy, high-brightness bleaching. An important criterion is also the bleachability with chlorine-free chemicals. This makes possible the treatment of the bleach effluent together with the waste digest liquor in the recovery stage and reduces the fresh water demand of a pulp mill to a minimum. Other aims are to reduce the cost of pulp production by reducing investment costs, e.g., in the fields of chemical recovery and utilization of useful byproducts from the dissolved wood components present in the black liquor.

1.3.3.1. Nitric Acid Pulping

The first investigations made by Payen (1838) showed that wood can be delignified with nitric acid. The breakdown of lignin proceeds via electrophilic substitution on the aromatic rings and subsequent oxidation reactions. Apart from the partial cleavage of methoxy groups and propane side chains, the ether bonds between the lignin units are also partially cleaved. As a result of the oxidation reaction, the nitrated lignin fragments possess mainly a quinoid structure and are highly colored. Alkali treatment is required to dissolve the partially water soluble nitrolignin.

Nitric acid pulping has achieved importance only to a modest extent in small pulp mills for the pulping of hardwoods and annual plants. The process was used for the first time on an industrial scale in the 1930s in Wolfen, Germany. Dissolving pulps with a high cellulose content (98 – 99 %) were produced from beech wood with a yield of ca. 40 %. The chips were first impregnated for 3 h under pressure at 40 – 45 °C with 15 % nitric acid. After the removal of excess acid and reduction of the acid concentration to 4 %, the chips were digested for 7 h at 70 – 95 °C. To dissolve the nitrated lignin, a two-step alkali extraction was then performed with 1 % NaOH for 60 min at 70 °C, followed by 3 % NaOH for 180 min at 80 °C.

Apart from this process, numerous other nitric acid pulping processes and modifications have been developed which differ primarily in the acid concentration, pulping temperature and time, and the extraction conditions [128–132].

Nitric acid pulping can be accelerated greatly by the addition of sulfuric acid or aluminum sulfate so that the digestion time for pressured digestion at 110 °C can be reduced to a few minutes. The alkali extraction step can be enhanced by the addition of oxygen, and ammonium hydroxide can be used instead of sodium hydroxide. The black liquor obtained in this way has a high nitrogen content and can be utilized as long-term fertilizer. Although the amount of nitric acid required has been reduced drastically from 550 kg/t, e.g., to 130 kg/t of pulp, by changing the original pulping procedure, this method may be considered economical only if cheap nitric acid is available and special pulps with a high market value can be made.

1.3.3.2. Pulping with Organic Solvents

Alcell Process. The use of low-boiling alcohols in wood pulping was proposed in the 1930s by Kleinert, who applied mixtures of ethanol and water to dissolve lignin [133]. In the 1970s, this method [134–136], in a modified form, was adopted in many new pulping processes [137–159]. The Alcell process [137–141] comes closest to Kleinert's original idea: hardwoods are cooked with a 50:50 ethanol – water mixture. Three extractors in series are used in which the chips are delignified by successive exposure to different pulping liquors. Since the solution applied first has already been used previously in two extraction stages, it has the highest content of dissolved wood components. Fresh solution is employed in the third delignification stage. The exchange of pulping liquor is effected by displacement. The alcohol is recovered by condensation of the gases during relief of the digester after the third extraction stage. The alcohol remaining in the pulped material is recovered by introducing steam into the extractor.

The waste liquor from the first extraction step is processed, the alcohol is distilled off, and the lignin that then precipitates is separated from the carbohydrates remaining in the aqueous solution by centrifugation. In this process, lignin and low-molecular carbohydrates and their reaction products can be recovered in addition to

pulp. Depending on the type of wood, 15 – 20 % lignin, 7 – 10 % xylose, and 1 – 3 % of other sugars based on o.d. wood are recovered from the black liquor.

At still acceptable pulp strength, the kappa number of the pulps is in the range 25 – 30. The yield for beechwood pulping is 50 % (based on o.d. wood) and thus, less than that given by kraft pulping. The relatively high lignin contents make it difficult to bleach the pulp to high brightness when exclusively chlorine-free bleaching agents are used. Alcell pulps subjected to conventional bleaching with chlorine-containing chemicals have a 10 % lower breaking length and at least 20 % lower tear strength than comparable kraft pulps. Until now, use of the Alcell process has been restricted to the pulping of hardwoods, which are easy to delignify. Due to lower costs for recovery of chemicals, it can be assumed that plant costs are 10 – 15 % lower than those of kraft pulp mills. The economy of this process can be further improved by alternative use of lignin and carbohydrates. However, this process cannot be regarded as a generally applicable alternative to the kraft process because the strength properties of the pulps are markedly lower than those of kraft pulps. Also, the restriction to hardwoods limits the applicability of the process.

Organocell Process. The limited capacity of aqueous alcohol solutions to dissolve lignin led to the development of organosolv processes, which involve the addition of conventional pulping chemicals. The Organocell process is a two-stage process [143–145]. In the first pulping stage, the chips are treated with a mixture of methanol and water (50 : 50) at 190 °C for 20 – 40 min. In this stage, equal amounts of lignin and carbohydrates, accounting for 10 – 20 % of the wood substance, are dissolved.

In the main pulping stage, the methanol content of the liquor is lowered to 35 vol % and ca. 20 % sodium hydroxide and 0.1 – 0.2 % anthraquinone based on o.d. wood are added. In this stage, the pulping time for softwoods is 40 – 60 min at ca. 170 °C. The yield and the residual lignin content of the pulp are comparable to the kraft process. Apart from pulp production, the utilization of substances in the black liquor is an objective of this process. Distillation is used to separate the alcohol from the pulping liquor from the first stage. Lignin precipitates in this process and can be removed by centrifugation. The carbohydrates remain in the aqueous solution. In this manner, about 6 % lignin and 5 % carbohydrates based on o.d. wood can be recovered in softwood pulping. Since the second stage requires sodium hydroxide, the carbohydrates dissolved in this stage are degraded extensively, and only lignin can be recovered from the black liquor. After methanol distillation, lignin precipitates after acid treatment or in the electrolytic recovery of sodium hydroxide. Since the recovery of sodium hydroxide by electrolysis consumes a lot of energy and the lignin recovered cannot be sold at a cost-covering price, the simplest solution is conventional black liquor treatment with evaporation, combustion, and recovery of chemicals. The recovery process, being free of sulfur, is unproblematic.

Organocell pulping can be used to process the same range of raw materials as the kraft process. However, it has disadvantages with regard to the pulp yields and the extent of delignification. The tear strengths of organocell pulps are comparable to those of kraft pulps, but the tensile strengths are much lower.

ASAM Process. Like the AS – AQ (NS – AQ) process, sodium sulfite is used as the main delignification agent in the alkaline sulfite process with the addition of anthraquinone and methanol (ASAM). Sodium hydroxide, sodium carbonate, or a mixture of the two can be added to increase the alkalinity of the pulping liquor. The optimal ratio of sodium sulfite to added alkali varies between 70 : 30 and 80 : 20, depending on the type of wood to be pulped and the desired pulping result. In addition to anthraquinone, used in amounts of 0.1 % based on o.d. wood, 10 – 25 vol % of methanol is added to the pulping liquor to intensify the delignification process.

The same range of raw materials as in the kraft process can be digested at 175 – 180 °C in 2 – 3 h. Both the yield and the pulp strength, especially the breaking length, are superior to those obtained from the kraft process. However, the essential advantages of the ASAM pulps are the lower residual lignin content (kappa numbers between 10 and 20), the relatively high brightness and the easy bleachability. Even completely

chlorine-free bleaching sequences can produce high brightness levels [152–155].

1.3.3.3. Pulping with Organic Acids

Both formic acid and acetic acid, with or without the addition of peroxides, can extensively remove lignin from wood. In all cases, concentrated acids are required, and temperatures range from 100 to 190 °C. The use of lower temperatures has the advantage that pressureless systems can be employed. In this case, however, an acid catalyst (e.g., HCl, aluminum chloride) or the addition of peroxide is required. Very low kappa numbers can be achieved in acetic acid pulping. The pulps as well as the dissolved lignin and the hemicelluloses are acetylated. Dissolved carbohydrates and lignin components can be separated and processed into byproducts. Chlorine-free bleaching in the acetic acid system can follow directly if peroxides are used. Another bleaching chemical that has proved effective under these conditions is ozone. The pulps have fairly good tensile strengths; the tear strengths are similar to those of sulfite pulps [160–167].

1.4. Technology of Pulp Production

1.4.1. Pulping

Wood pulping can be carried out in a continuous or batch process. The steel digesters are protected internally against the corrosive digestion liquor. In older batch digesters, acid resistant tiles or carbon bricks were used for the digester lining. A lining of this type reduces the digester volume by 7 – 15 %. In addition, the bricks and tiles must be inspected regularly because they can crack easily, especially at the joints, due to the high temperature differences during heating up and cooling down. Today, digesters are usually plated on the inside with stainless steel. Batch digesters (Fig. 17) have a volume in the range of 100 – 400 m^3. Rotary digesters are older types with smaller volumes. They are rotated head over in order to thoroughly mix the chips with the digestion liquor, and are heated directly by the introduction of steam into the digester. In this way, the liquid volume increases and the concentration of chemicals is reduced.

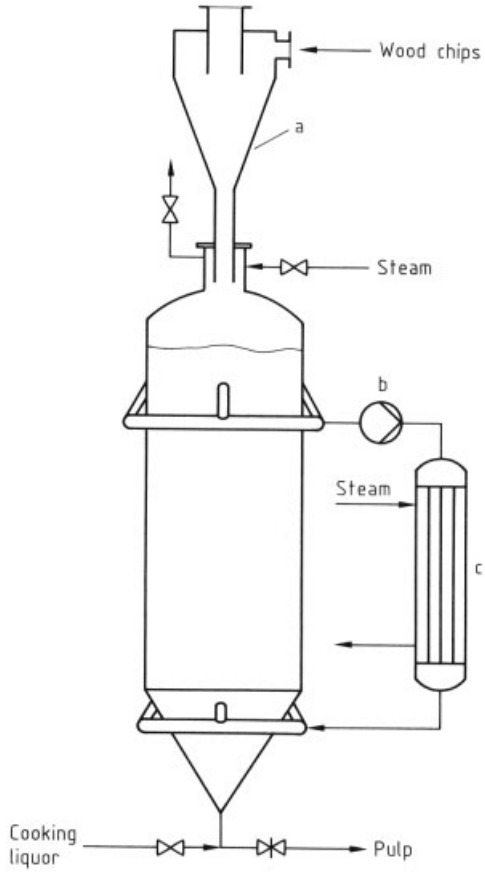

Figure 17. Batch digester
a) Wood chip collector; b) Circulation pump for digestion liquor; c) Heat exchanger

In more modern pulp mills that use batch pulping systems, the digesters are arranged upright in batteries and each is equipped with a liquor circulation system. The liquor is removed at the bottom of the digester and pumped into the top of the digester. A heat exchanger in the circulation cycle is used for indirect heating. Digesters with liquor circulation systems and direct heating are very rare.

The investment costs for conventional batch pulping systems are lower than for continuous systems. However, it is more difficult to maintain the constant process conditions that are required for uniform pulping results. Moreover, the heat economy of batch pulping systems is unfavorable. To reduce or completely eliminate these disadvantages, energy saving systems have been developed for batch digestion systems, which

use the heat content of the waste liquor to heat up the next digester. This can be achieved either by heat exchangers or by collecting the black liquor in thermally insulated collecting tanks and using it for heating chips. In this way, the residual pulping chemicals in the black liquor are also made available for pulping. Apart from improved heat economy, impregnation of the pulping material is also increased. The chemical dosing during the entire cook can be better adjusted to the pulping requirements, reducing side reactions, especially with the carbohydrates. Pulps made by such liquor displacement can be delignified to a greater extent. They have better yields and very good strengths.

Modern pulp technology is dominated by continuous digesters. The leading system worldwide was developed by Kamyr (Fig. 18). In this process, chips are continuously fed into upright digestion tubes from the top while the pulp is removed at the bottom of the digester. digestion systems of this type have a very wide capacity range of 200 – 2000 t/d. All conventional pulping processes can be conducted in this system, and many different types of pulp, ranging from semichemical pulp to dissolving pulp, can be produced. Accordingly, both the design and the process mode of this system can be modified.

The standard design functions as follows. Chips are introduced continuously via a low-pressure valve into a presteaming vessel, in which they are heated with low-pressure steam (150 kPa) to prepare them for subsequent impregnation. The chips are transferred to an impregnation tower through the chip shute together with the pulping liquor via a high-pressure valve. Complete impregnation of the chips takes place here. The chips are discharged from the bottom of the system and are conveyed to the digester together with the liquor. They are then heated to the maximum digestion temperature. The chips are transported from the top of the digester downwards and pulped in the process. The pulping occurs in the interchanging action of liquid and vapor phase digestion. Recirculation of the pulping liquor is achieved by means of a system of concentric tubes in the center of the digester and outlet screens set in the wall. Varying numbers of these screens are arranged at different levels in the digester. In each case, the inlet tubes end slightly above the outlet screens, and the pulping liquor is fed centrally into the digester. As a result of pressure differences, it flows from the center to the outlet screens in the wall. This liquid cycle causes the reaction products to be washed out of the chips and replaced by new pulping chemicals, which react in the vapor phase between the screens.

Wash liquor is added in the lower part of the digester and moves upwards countercurrently to the stream of chips. Thus the spent liquor present in this region is displaced, and the pulp is subjected to preliminary washing. In addition, this technique serves to reduce the fiber temperature, and only little fiber damage occurs during pulp discharge through a valve by rapid pressure relief. The wash liquor used comes from washers located behind the digester. The black liquor is withdrawn through the last screen ring and fed into the liquor evaporation system via a heat exchanger.

Apart from the Kamyr system, various companies have developed continuous pulping systems for the production of semichemical and chemical pulps based on both wood and annual plants. In addition to upright digesters, inclined and horizontal digesters are also employed, in which a conveyor system such as a screw, is used to transport chips and chemicals. All these systems are only partially suited to the production of chemical pulps from wood because liquid-phase pulping cannot be carried out due to the lack of liquor recycling systems. All the pulping chemicals must be homogeneously distributed throughout the raw material before pulping starts. Subsequently, pulping occurs in the vapor phase without exchange of chemicals.

1.4.2. Washing and Screening of Pulp

Brown-Stock Washing. In the production of chemical pulps, 45 – 55 % of the wood substance fed into the digester is dissolved. The remainder is present as fibers, largely still in the form of the original chips. In addition, ca. 200 – 500 kg of pulping chemicals per tonne of pulp are present, depending on the process and the extent of delignification. It is essential for economic and environmental reasons that the dissolved wood substances and the pulping chemicals are separated from the pulp as completely as possible with a minimum amount of wash water. The black liquor contains organic components

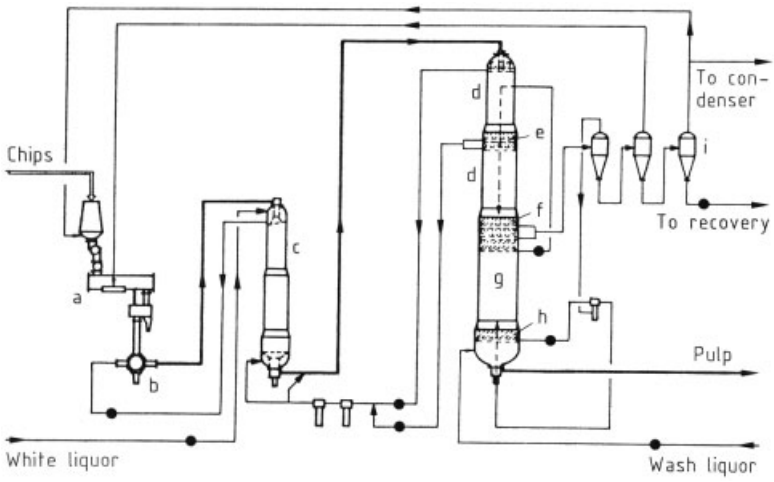

Figure 18. Continuous digester (Kamyr system)
a) Presteaming vessel; b) High pressure valve; c) Impregnation tower; d) Digester; e) Strainers for recirculation of digestion liquor; f) Strainers for waste black liquor; g) Washing zone; h) digestion liquor outlet; i) Heat exchangers

of high heating value, whose combustion can cover all or most of the energy requirements of a pulp mill. For economic operation of the mill, the digestion chemical make-up demand should be as low as possible. The washing efficiency must be as high as possible to minimize the pollution load of the mill effluent stream. However, this objective must be achieved with a minimum of wash water because its evaporation before the subsequent black liquor combustion is an energy consuming process.

The washing process is influenced by many factors, e.g., the type of wood, the chip size, the packing density of the chips in the digester, the extent of delignification, the concentration of the black liquor, amount of wash water, flow rate, temperature, washing time, diffusion gradient, and the pH.

Pulp washing can be divided into two steps. The first stage involves the displacement of the free spent digest liquor. This occurs either with the wash water or by pressing the pulped material. The second step involves the removal of the black liquor from the chips or fibers. The washing process occurs largely by diffusion since the free flow of the wash water is impeded by the tissue structure of the wood. The washing process must be controlled such that both steps are carried out under optimal conditions. In other words, the free black liquor must first be displaced from the brown stock to establish a concentration gradient between the liquor inside the chips and the liquor surrounding the chips. Then optimal diffusion conditions must exist to ensure the fastest possible removal of the dissolved wood substances and the pulping chemicals from inside the chips and fibers. In practice, these two washing steps, displacement of free waste liquor and subsequent diffusion, cannot be clearly separated from the viewpoint of time and space. In the displacement of free black liquor, it is unavoidable that the black liquor mixes with the wash liquor. A region of low black liquor concentration is established where diffusion occurs again on account of the concentration gradient. The washing is influenced considerably by the accessibility of the chips and the individual fibers after pulping. The higher the delignification, the easier it is to wash the black liquor solids and the chemicals out of the chips. A very important factor for the washing efficiency is whether the process is dominated more by displacement washing or by diffusion. If the wood substance is still present as chips, the dissolved substances require a correspondingly longer time to find their way out of the chips and into the wash water. Diffusion is of importance primarily in the removal of dissolved substances from the cells. The rate of diffusion depends on the concentration gradient between the black liquor in the cell wall and in the wash water. Moreover, the washing temperature is of significance. At elevated

temperatures, the solubility of the substances to be washed out increases and the viscosity of the wash water decreases. The diffusion rate can be doubled by increasing the temperature by 30 %, which favors high-temperature washing. However, this can be carried out only in closed systems. If the pulped material is still present as chips during washing, diffusion plays the dominant role and the time required to achieve the effective removal of black liquor is relatively high. In the case of defibrated material, displacement of black liquor by the wash water dominates.

Washing Systems. In the early days of pulping, washing was carried out in the digester or in the blow pit into which the digesters were emptied. The amount of wash water used did not have to be restricted because the waste liquor was not monitored; the dissolved wood substance and the pulping chemicals were simply discharged into the receiving stream. Today, a washing efficiency of 99 % must be attained with dilution factors of 2 (wash water/pulp) or less. This is possible only in closed systems in which the repeated use of the wash water takes place in a countercurrent washing process.

In the continuous pulping system described above, pulp is prewashed in the lower part of the digester with wash filtrate from a subsequent washing stage. The washing efficiency attainable at this stage is already above 80 %. The pulp is then fed into a single- or multistage diffusion washer (Fig. 19). Since this is a closed system, effective, high-temperature washing can be performed above 100 °C. The pulp enters the diffuser at the bottom with a consistency of 5 – 10 %. In the interior, concentric perforated outlet screens are to be found over the entire area. The wash water is injected into the system by rotating jets between the outlet screens and diffuses through the pulp to the outlet screens, where the waste liquor is withdrawn. Several of these washing units can be placed at different levels in one diffuser and operated in series, with fresh water being used only in the last step. The system is operated countercurrently and the pulp leaves the diffuser from the top. The digester washing and a subsequent two-step diffuser give a washing efficiency of more than 99 %.

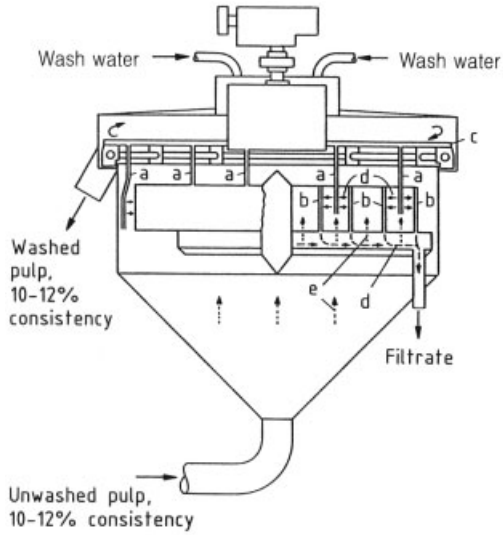

Figure 19. Diffusion washer
a) Rotating nozzle for wash water addition (attached to (c)); b) Concentric screen rings for removal of wash water; c) Rotating pulp scraper; d) Wash water outlet; e) Upflowing pulp

The standard washing in a pulp mill is carried out in a washing line which operates with rotary filters (Fig. 20). In this process, drums are used that are closed at the sides and have a screen spanned over their shell surfaces. The drum dips into a chest containing the pulp suspension at a consistency of ca. 4 %. During rotation through the suspension, the fibers become attached to the filter surface, while the waste liquor enters the filter and is pumped off. The fiber mat on the filter surface is removed by a scraper. Modern filters are closed and can be operated above 100 °C at an overpressure. The washing effect is increased by spraying wash water onto the pulp web on the filter surface. In modern filter designs, several washing steps are installed in a single unit. A washing line consists of several filters operating in series. Fresh water is used only in the last filter, the preceding filters each operate with the filtrate from the following washing step. This countercurrent operation minimizes water consumption. The filtrate from the first washing stage, which is polluted to the greatest extent, is used to dilute the pulp in the digester.

Double-sieve presses and double-roll presses have also proved effective in pulp washing. In the former, the pulp is dewatered between two flat sieves which run on pressure rollers. The second system involves double presses with perfo-

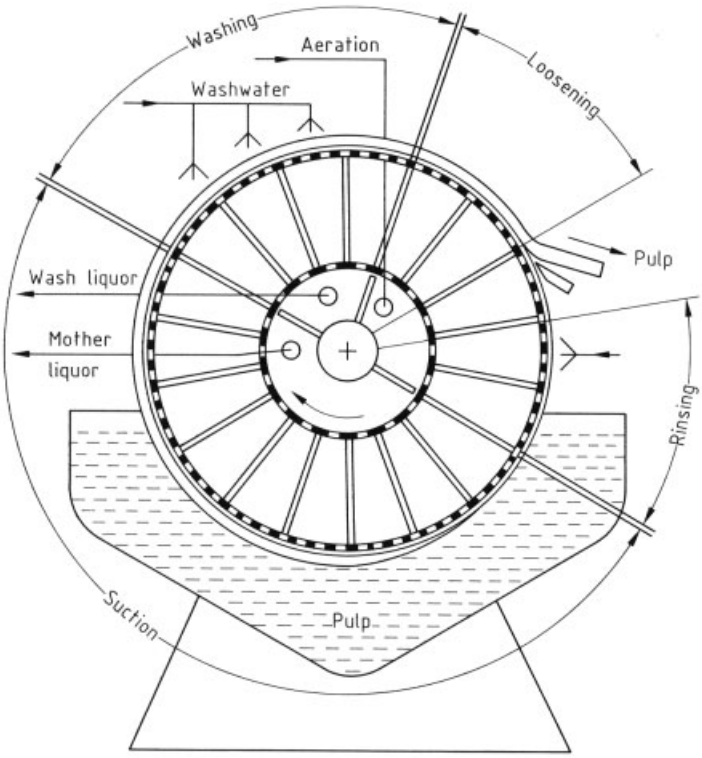

Figure 20. Rotary filter for washing pulp and recovering waste liquor

rated surfaces. They dip into a pulp suspension and their surfaces are pressed together. While ca. 15 % solids content can be achieved in a pulp suspension by dewatering filters, the pulp consistency attainable with double sieves reaches 40 % and with double-roll presses >40 %.

Fiber Screening. After the removal of dissolved substances from the pulp, all solid impurities that are not part of the fibrous material must be removed. These include bark residues and undercooked wood particles such as knots and shives, as well as external impurities such as sand, stones, and even metal objects.

As a result of the varying size ranges and other sorting criteria, e.g., specific weight, screening and cleaning must be carried out with different separating aggregates in a multistage process.

Coarse screens or knotters, commonly of vibratory design, are used to separate large rejects (Fig. 21). The flat screens have perforations in the form of slots or round holes. The screen vibrates elliptically, and the pulp flows onto the screen with a consistency of 1 – 2 %. The accepts pass through the screen, while the rejects, owing to the vibration direction of the screen, are transported over its surface to the outlet. The spraying of water on the surface of the screen intensifies the screening process.

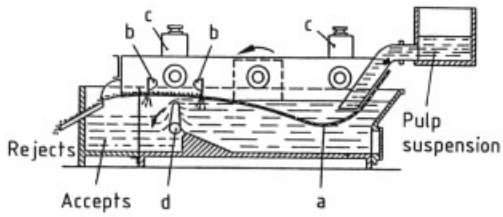

Figure 21. Jönsson shaker
a) Vibrating screen; b) Canned sprayer; c) Casing for spring; d) Overflow control

Vibrating screens can also be designed in cylindrical form. They consist of perforated, vibrating cylinders that have a longitudinal axis of rotation. Depending on the design of the screen

surface, they can be used in both coarse and fine screening.

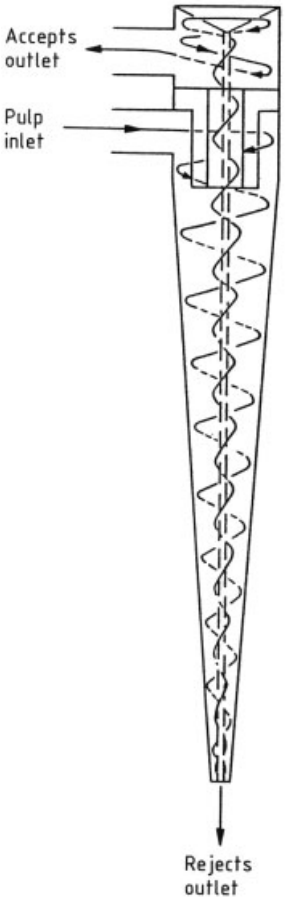

Figure 22. Schematic of a centrifugal cleaner (hydrocyclone)

Centrifugal strainers have a stationary horizontal cylindrical screen. The pulp suspension enters this body with a stock consistency of less than 1 %. The rotating suspension is flung against the screen, causing the fibers to pass through it. To improve the screening process, water can be injected into the drum. Multistage screening in one aggregate is possible by using screen perforations of different sizes at different levels in the cylinder.

Today, pressure screens are most often used for pulp screening. They have a high output and produce excellent separation. These are vertical centrifugal screens with a vertical shaft and a fixed screen basket. The stock is introduced tangentially from the top. The centrifugal motion of the stock is intensified by the propeller blades rotating at the bottom of the screen. Very effective screening is possible because of the pressure pulsation that occurs due to the scraper closely passing the screen. Moreover, the blades keep the screen surface free. Since screening is accelerated by applying overpressure, the rejects must be removed batchwise. On the other hand, the accepts go into a chest or directly to the paper machine, thus maintaining the pressure in the system. Pressure screens are used almost entirely for fine screening.

The screens described above remove impurities primarily on the basis of particle size. Unwanted accompanying substances with the same dimensions as fibers, e.g., sand or bark residues, cannot be separated in this way. Centrifugal cleaners (hydrocyclones) are suitable for the separation of these contaminants. A cleaner of this type is shown in Figure 22. The stock enters tangentially so that a vortex flow is produced in the tube. The tangential, axial, and radial velocity components of the swirling motion of the suspension are dependent on the location. The suspension near the cyclone wall moves predominantly towards the rejects outlet, while that in the interior moves towards the accepts outlet.

It can be assumed that in the cyclone, fibers and impurities move approximately with the tangential and axial velocity of the carrier liquid. In the radial direction, however, considerable differences in velocity occur which are responsible for the separation process. If a particle moves towards the cyclone wall as a result of the centrifugal force and the radial pressure gradient, this motion counteracts the resistance to fluid flow because the particle must move through the carrier liquid. The larger the projection area of a particle, the higher the resistance. In addition, the migration velocity of the particle depends on its specific weight. The higher the mass per unit projection area, the quicker it migrates to and along the wall to the rejects outlet. Hence, specifically heavy particles such as sand are removed from the stock flow relatively early. In the case of particles of the same specific gravity, e.g., fibers and some bark components, separation is based on the different ratio of projection area to volume. The larger the projection area of a particle, the higher the power of resistance of the liquid. As a result, even in an early

phase of the separation process, large fibers with correspondingly large projection surfaces reach the fiber outflow in the center of the cleaner. In comparison, smaller particles are carried further down and then pass into the rising accepted stock flow. By using a particular entry velocity and dimensions of the centricleaner, particles with defined length to thickness ratios can be eliminated. Since fibers or fiber fragments and bark components can have the same dimensions, separation can be achieved only incompletely in a single step. For this reason, cleaners are arranged in several stages and the rejects are after-screened in each case. Cleaners normally operate with dilute suspensions with pulp consistencies of less than 1 %.

The arrangement and operation of the different screening systems in a pulp mill can vary considerably, and are determined largely by the type of pulp being made. Thus, in the prescreening of unbleached pulp, swinging perforated sieves and centrifugal sorting machines are used to remove knots and other coarse impurities. In final screening, however, swinging cylindrical sieves and centrifugal cleaners are used.

1.4.3. Pulp Drying

After being cleaned in the screening process, pulp is usually dried either in the bleached or unbleached state. This drying to a solids content of ca. 90 % is required in all cases for pulp that is to be marketed. Also, in the case of integrated pulp and paper mills, the pulp is sometimes dried because the fiber properties of interest for papermaking are changed by drying. Pulp that has been subjected to intermediate drying swells to a lower extent and can be drained more easily on the paper machine. The opacity of paper made from this pulp is higher. Intermediate storage in the dried state poses no problems.

Conventional dewatering and pulp drying is carried out on a Fourdrinier. As in the production of paper and board, the fiber suspension flows through a stock inlet onto the wire and is then drained. The single-layer mat of pulp thus formed is removed at the end of the wire, passed through presses onto conventional steam-heated drying cylinders where the final drying to a predetermined moisture content takes place. To increase drying efficiency, the final drying of the pulp, after web formation and pressing, is often completed in flakt or jet dryers. These are usually arranged in several levels one above the other. The pulp passes through them and is dried in a stream of hot air. After drying, the mat of pulp is cut into sheets, stacked, pressed, and packed in bales.

Fiber Flash Drying. In fiber flash drying, the moisture content of the pulp is first reduced to about 50 % in double-screen, disc, or double-roll presses. The predrained pulp is then flaked in a shredder. The pulp flakes are subsequently blown into a hot-gas channel and dried in two steps. The drying stages involve ventilators, high-performance suspension towers, and cyclone separators. The dried flakes are fed continuously into a bale press, which produces bales with a specific weight of about 0.5 g/cm^3. The bulk density is increased to at least 0.8 g/cm^3 by repressing the bales in a hydraulic bale press. As a result of the relatively severe drying conditions, the fibers can become horny, especially if the hemicellulose content of the pulps is high.

1.5. Waste Liquor Treatment

The waste liquor obtained from the production of chemical pulps contains at least 50 % of the wood substance and pulping chemicals. For environmental and economic reasons, this waste liquor must be utilized as far as possible. Indeed, its utilization to cover the energy requirement of pulp production is of primary importance, and in most processes, this is combined with the recovery of the pulping chemicals. In addition, part or all of the waste liquor can be utilized directly or after chemical or biological treatment.

1.5.1. Recovery of Useful Products from Waste Liquor

The composition of waste liquor depends on the raw material and pulping process used. For instance, utilizable carbohydrates are found only in waste liquor from acid pulping processes, while resin acids and fatty acids occur only in black liquor from alkaline pulping processes that use resin-containing softwoods. The lignin structure is also influenced by the

pulping process. Thus, water soluble lignosulfonates are found in the sulfite process, whereas alkali lignins are formed in the kraft or soda process. The composition of typical sulfite and kraft waste liquors from softwoods is shown in Tables 8 and 9.

Table 8. Composition of softwood sulfite spent liquor

Product	Quantity, kg/t pulp
Methanol	7–10
Acetic acid	30–90
Formic acid	0.5–1
Formaldehyde	2–6
Furfural	5–6
Sugar sulfonic acids and aldonic acids	150–250
Sugars	200–400
Cymenes	0.3–1
Lignosulfonates	600–800

Table 9. Composition of softwood kraft black liquor

Product	Quantity, kg/t pulp
Methanol	5
Acetates	100–200
Carbohydrate degradation products (lactones, sugar acids, etc.)	350–400
Aliphatic sulfur compounds (methyl mercaptan, dimethyl sulfide, dimethyl disulfide)	1
Terpentine	8–10
Tall oil	20–100
Alkali lignin	400–600

Waste liquor from the sulfite process is better suited to the nonenergetical waste liquor utilization than kraft black liquor. Concentrated sulfite waste liquor is used both as a pelletizing agent in the feed industry as well as in the pelletizing of coal and ore. The carbohydrates obtained in the sulfite process can be utilized as shown in Figure 23. Hexoses are fermented to ethanol, while the pentoses are converted to yeast, or reduced to xylitol, or converted to furfural [168], [169].

The lignosulfonates in sulfite waste liquor can be used in many ways, e.g., they are employed as concrete additives to improve the viscosity. The dispersing properties of lignosulfonates are exploited in the dispersion of pigments in dyes, as dispersing agents for water insoluble compounds in pesticides, as drilling agents, and as additives in the production of adhesives for the wood industry. The residual waste liquor components are present in amounts that are usually too small for cost-effective utilization. However, the elimination of these substances from waste liquor may be necessary for pollution control.

In the recovery of byproducts from kraft black liquor, only two groups of substances are of economic interest: alkali lignins and the resins and terpenes. The terpene fraction is steam volatile, so when the digester is relieved, this fraction is found in the gas phase and can be recovered by condensing the digester gases. The wood resins are dissolved in the black liquor. During concentration in black liquor evaporation plants, the resins can be skimmed off the waste liquor as sodium soaps with a solids content of ca. 28 %. They are salted out and separated by distillation.

Alkali lignins as such are of little interest economically. They are, however, subjected to fractional precipitation and sulfonation to give high-quality lignosulfonates. They can also be converted to polymers. The following reactions are of potential interest: lignin and formaldehyde to give phenolic plastics; lignin and epichlorohydrin to give epoxide resins; and lignin and epoxides to give polyalcohols, followed by reaction with diisocyanates to polyurethanes [170], [171]. However, these have not been used until now because products recovered from fossil raw materials can be produced in higher quality and often more cheaply. Lignin can be converted to a nitrogen-yielding fertilizer by reaction with ammonia and oxygen under pressure. A series of low molecular mass products based on lignin can also be produced. Vanillin and dimethyl sulfoxide are recovered on an industrial scale. Due to low demand, however, only a very small part of the waste liquor lignin obtained can be used for this purpose. Potential alternative uses include hydrogenolysis of lignin to give phenols, and high-temperature pyrolysis for the recovery of gaseous or liquid products such as carbon monoxide or ethane and methane. However, these processes are not used industrially because they are not cost effective.

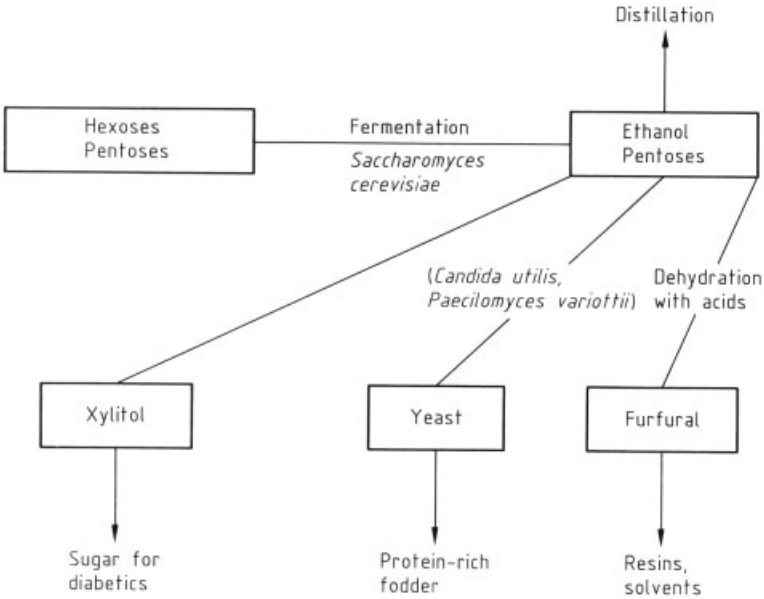

Figure 23. Processing of pentoses and hexoses

1.5.2. Energy Generation from Waste Liquor

After pulp washing the black liquor has a solids concentration of 10 – 18 %. The solids content must be increased to 55 – 70 % before utilization for energy. Counterflow multistage vacuum systems are the most frequently used liquor evaporating systems. They consist of 4 to 7 evaporators arranged in series. The individual evaporators usually comprise a lower evaporating part and an upper vapor space. The dilute black liquor flows between the vapor space and the evaporating part into the evaporating body. The evaporating part consists of either a steam-heated tube bundle or falling-film evaporators. The waste liquor is heated in these systems, which operate as heat exchangers, and the steam produced is withdrawn through the vapor space and later removed from the system as vapor condensate. The waste liquor concentrated in this manner passes into the next evaporating body and is concentrated further with steam. It flows through the entire evaporating system in this way and leaves with the required final concentration. Because the steam is reused repeatedly, its temperature decreases. To maintain the evaporation capacity, this loss of temperature is compensated for by operating the evaporators under vacuum; the lower the steam temperature, the higher the vacuum employed. The higher the rise in concentration of waste liquor in the individual evaporating steps, the higher the evaporation temperature used. In this way a sufficiently high evaporating efficiency is ensured and, at high solids concentrations, the viscosity of the waste liquor is maintained at such a level that precipitation and deposition on the heat exchangers is avoided. A flow sheet of an evaporation plant is shown in Figure 24.

The risk of precipitation in the liquor evaporating system is especially high for more concentrated liquor. This problem can be combated by rinsing the evaporating system from time to time with vapor condensate. During this time, the evaporating body being rinsed cannot be used for evaporation. The evaporating system can be designed so that half of it is used for evaporation while vapor condensate flows through the other half. The two systems can be used alternatingly by changing from one to the other. Moreover, systems exist that operate without vacuum. In these systems, evaporation is also carried out in counterflow but at correspondingly higher temperatures. The lowest steam temperature of a system of this type used to concentrate the dilute black liquor flowing into the first stage must

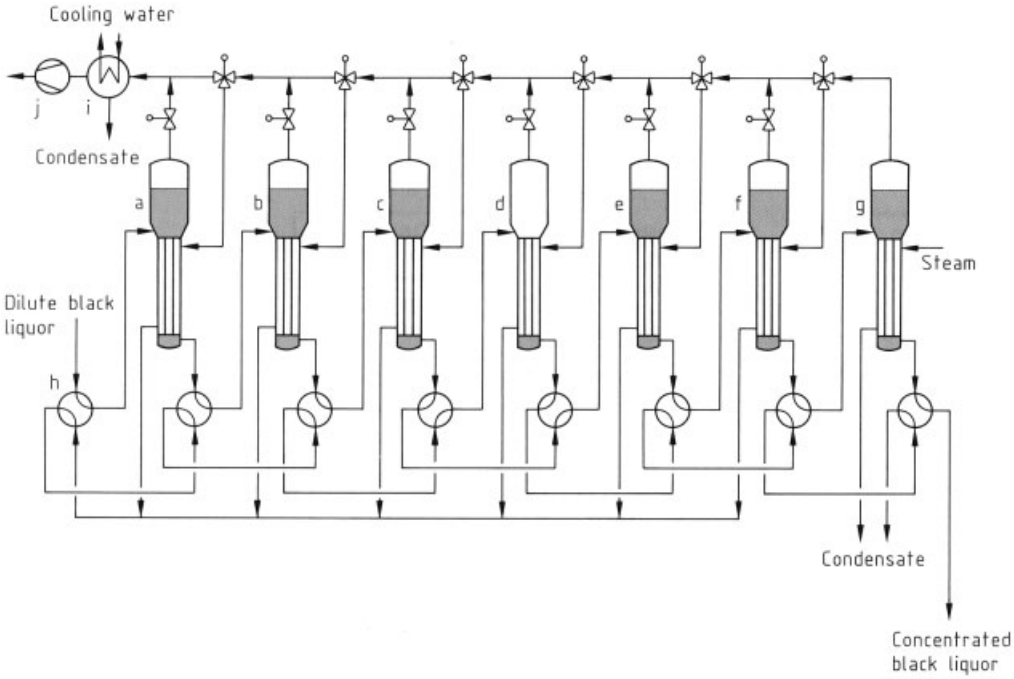

Figure 24. Schematic of a multistage vacuum evaporating plant
a) – g) Evaporators (Evaporator d is switched to washing); h) Reversing valve; i) Condenser; j) Compressor

be well above 100 °C. Although the energy requirement in the individual evaporating stages is quite favorable in a system of this type, utilization of the waste steam is necessary if the total energy requirement is to be minimized. This often causes problems in a pulp mill, because most mill operations require higher steam temperatures.

In modern evaporating plants, liquor concentrations of up to ca. 80 % are attainable. However, the composition of the liquor plays a considerable part in concentration. The viscosity is very important as is the concentration-dependent solubility. Here, the molecular structure of lignin and its solubility in water and alkali are of significance. The concentration and composition of inorganic components in the waste liquor are also important. Silicates, which are present in large amounts particularly in annual plants, are extremely critical because they deposit as sodium silicates on heat exchanger surfaces during alkaline pulping. The concentration of problematic waste liquor of this type to a high solids content is possible only by direct evaporation. In this process, the liquor is dried directly in a hot-air drier, which is fed with waste gas from the black liquor combustion stage. This can lead to air pollution problems because reduced or oxidized gaseous sulfur compounds are liberated from the waste liquor. These must then be introduced into the liquor combustion system or the off-gas purification system.

Combustion of Concentrated Liquor. Liquor combustion combines several objectives. It is a means of disposal; serves the recovery of energy, making pulp production self-sufficient; and allows the recovery of the majority of the pulping chemicals charged. Combustion is carried out in a special combustion furnace which functions as a thermal power station and is equipped with a chemical recovery system. Both the combustion furnace and the following chemical regenerating system are adapted to a particular type of waste liquor. A furnace suited to the combustion of waste liquor based on sodium is shown in Figure 25.

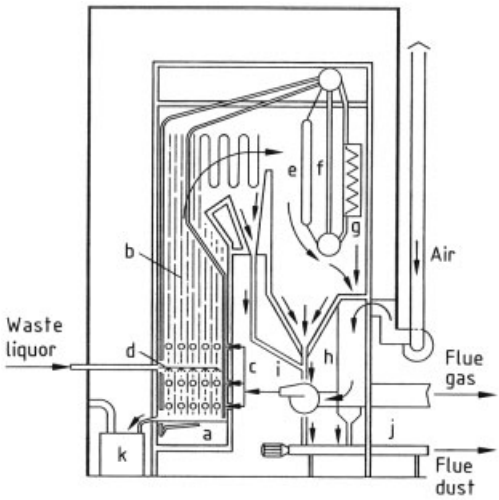

Figure 25. Black liquor combustion furnace
a) Melting zone; b) Boiling tubes; c) Nozzle for combustion air; d) Nozzle for black liquor; e) Superheater; f) Preevaporator; g) Fresh water preheater; h) Air preheater; i) Pipes for flue gas transport; j) Transport system; k) Smelt dissolver

The concentrated waste liquor is injected into the lower part of the furnace. In the combustion of waste liquor with a low content of organic components (e.g., waste liquor from the production of semichemical pulp), the simultaneous combustion of heavy fuel oil or coal dust is required. Injection occurs through oscillating nozzles to achieve uniform distribution on the furnace walls where the remaining water evaporates. The organic components of the liquor burn at ca. 1000 °C. The dosing of oxygen during the combustion process is of considerable importance. Sodium-containing waste liquor is normally burnt under reducing conditions in the furnace bed. This means that enough oxygen must be present in the combustion air, so that a complete conversion of organic substance to CO_2 and water occurs. However, the sulfur compounds in the waste liquor should be present in the reduced form, i.e., bound to sodium as sodium sulfide in the smelt. In contrast, in sulfite waste liquor based on magnesium, combustion conditions must be regulated so that magnesium oxide is present as ash, and SO_2 in the gas phase. Both compounds must be separated from the combustion gas. In the combustion of kraft black liquor, most of the pulping chemicals are present in the form of sodium sulfide and sodium carbonate in the smelt at the bottom of the furnace and are then transferred to a dissolving tank. The flue gases pass through heat exchangers in which high-pressure steam is produced. The gases contain a considerable amount of entrained dust and must therefore be purified in mechanical dust collectors and electrostatic filters. An additional gas purification step is required to meet environmental regulations. Sulfur-containing gaseous compounds, in particular, must be eliminated. This is achieved by wet cleaning in scrubbers. The solid and gaseous combustion products which contain pulping chemicals can either be reused directly in the production of pulping liquor or they must be first subjected to chemical conversion.

1.5.3. Chemicals Recovery

Kraft Process. A series of reactions take place in the combustion furnace during the combustion of kraft black liquor. The carbon in the organic material combines with oxygen to form carbon dioxide, which can react with free NaOH to give sodium carbonate. Since combustion partly takes place under reducing conditions, sodium compounds react with the sulfur to form sodium sulfide. The chemical losses that have occurred in the entire pulping process are compensated by addition of sodium sulfate as make-up before combustion. This sodium sulfate is reduced to sodium sulfide in the combustion furnace. For economic reasons, other sodium and sulfur-containing materials are often used to compensate for losses. In modern mills, the chemical losses, calculated as sodium sulfate, are 10 – 20 kg per tonne of pulp. This corresponds to >90 % recovery of the pulping chemicals used. The reactions occurring in the combustion furnace can be simplified as follows:

$$C + O_2 \rightarrow CO_2$$

$$2\,NaOH + CO_2 \rightarrow Na_2CO_3 + H_2O$$

$$Na_2O + CO_2 \rightarrow Na_2CO_3$$

$$Na_2SO_4 + 2\,C \rightarrow Na_2S + 2\,CO_2$$

Therefore, the smelt present at the bottom of the combustion furnace consists primarily of sodium carbonate and sodium sulfide. A solution of this smelt in water is referred to as green

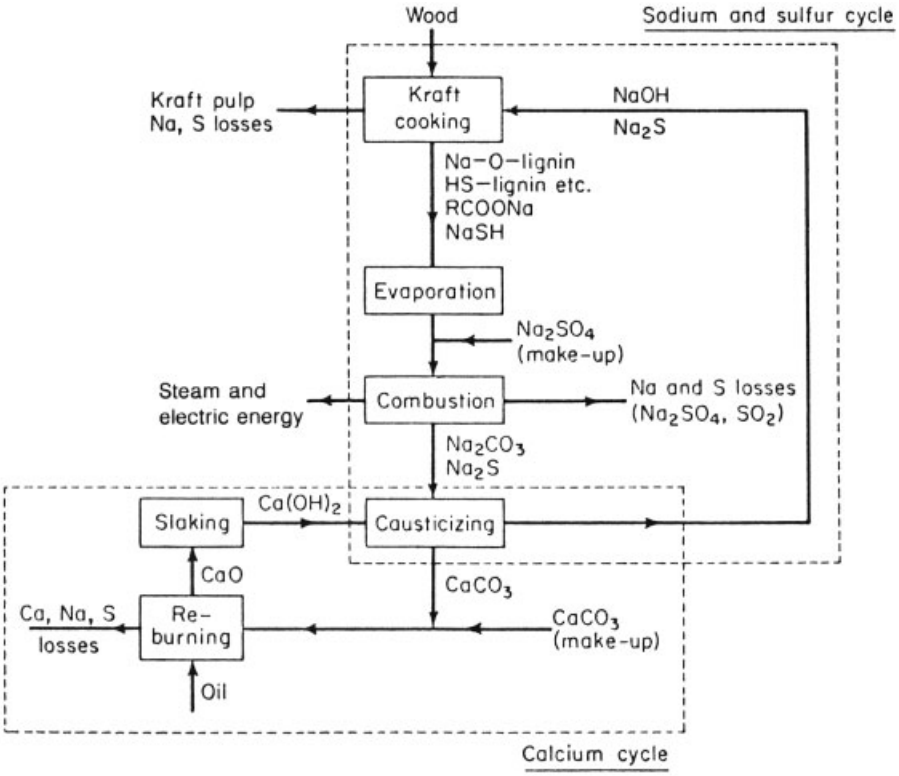

Figure 26. Recovery cycles for a kraft mill

liquor. The greenish color is due to iron sulfides contained in the solution.

To recover the pulping chemicals used, sodium carbonate must be converted to sodium hydroxide. This takes place in the causticization step, in which sodium carbonate reacts with slaked lime to give sodium hydroxide and calcium carbonate:

$$Na_2CO_3 + CaO + H_2O \rightarrow 2\,NaOH + CaCO_3$$

This is an equilibrium reaction which, under normal conditions, produces degrees of causticization of 85 to 90 %. The solubility of calcium carbonate increases with increasing OH^- content, and the reaction equilibrium is shifted to the right. Normally 2 – 3 % of calcium oxide remains in the sludge produced by causticization. Although the addition of more quicklime improves the degree of causticization slightly, a multistage causticization process is more effective. The calcium carbonate formed is separated from the white liquor and rewashed to reduce chemical losses. About 250 kg quicklime per tonne of pulp are required for causticization. Amounts of this magnitude are very expensive and the large capacity of kraft pulp mills makes this not only a procurement problem, but also a waste problem. Therefore, the lime sludge is reburnt in a rotary kiln in the plant, converting the calcium carbonate in an endothermic reaction to calcium oxide:

$$CaCO_3 \rightarrow CaO + CO_2$$

The lime kilns are a major source of air pollution because mercaptans are released from liquor residues in the sludge. The outlet air from the kilns can be returned to the liquor combustion furnace to reduce odors. A chemial recovery system for the kraft process is shown in Figure 26.

Calcium Bisulfite Process. Apart from calcium oxide and sulfur dioxide, calcium sulfate is formed predominantly in the combustion of calcium bisulfite liquors. The complete

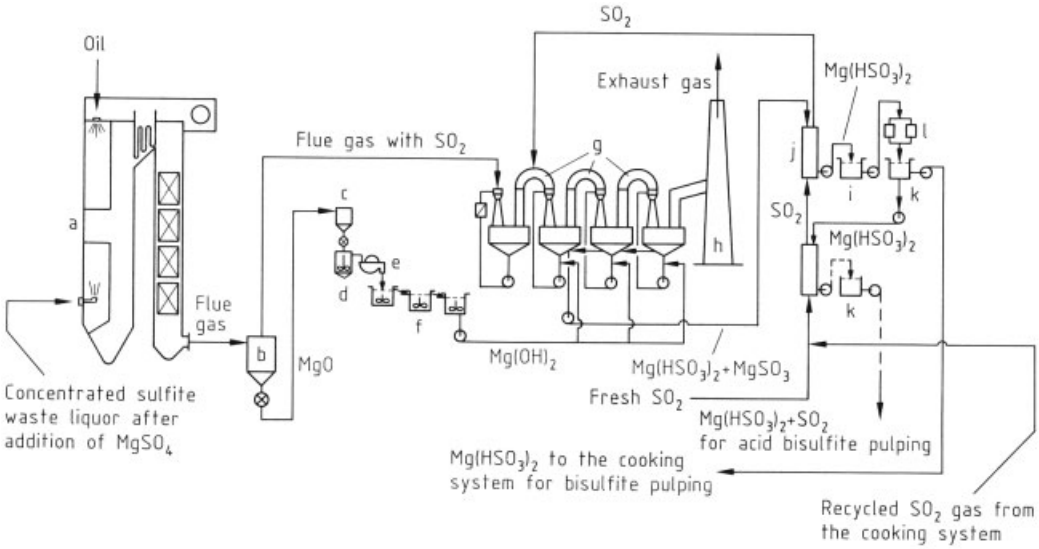

Figure 27. Recovery of the chemicals from the spent liquor from the magnesium bisulfite process
a) Waste liquor combustion furnace; b) Dust collector; c) MgO silo; d) Slurry tank; e) Wash filter; f) Hydration vessels; g) Flue gas scrubber; h) Stack; i) Sedimentation tank; j) Fortification tower; k) Storage tank; l) Filter

thermal decomposition of calcium sulfate occurs only above 1150 °C; temperatures above 1200 °C must be used to achieve an essentially quantitative decomposition. This causes problems because dead burning of lime occurs at these high temperatures, changing its crystal structure and reducing its reactivity.

Magnesium Bisulfite Process. In comparison with calcium sulfate, the magnesium sulfate obtained from the combustion of magnesium sulfite waste liquor can be decomposed thermally at normal combustion temperatures to give gaseous SO_2 and magnesium oxide. The latter can be separated from the combustion gases in powder form. In slaking tanks, the magnesium oxide is converted to magnesium hydroxide, which reacts with SO_2 to form a mixture of magnesium sulfite and magnesium bisulfite. In the fortification tower, this solution is treated with SO_2 so that either a pure magnesium bisulfite solution is formed which can then be used in the magnefite process, or a magnesium bisulfite solution with excess SO_2 for acid magnesium bisulfite pulping. A schematic of the process is shown in Figure 27.

Flue gas purification using electrostatic filters is not widely employed because chlorides, for example, are not removed in this process. Instead, the pulp particles are discharged from the gas flow in mechanical purification units or multiple cyclones. In the subsequent wet cleaning, water-soluble solid components of the flue gases are removed. The flue gases then again pass through wet dust separators, which precipitate further chlorides. Chloride discharging occurs in a retention tank for magnesium oxide. The chloride level of the chemical cycle can be controlled by regulating the overflow.

In sulfite mills, the spent liquor contains considerable amounts of acetic acid, furfural, and methanol, especially when hardwoods are used for pulping. For this reason, the vapor condensate obtained on evaporation is highly polluted. Therefore, the digestion spent liquor is usually neutralized before evaporation to convert the low-boiling, acidic components to salts and thus prevent their transfer to the vapor condensate. For waste liquor neutralization in magnesium bisulfite mills, the amounts of MgO required are of the same order of magnitude as the amounts of magnesium oxide used in digestion liquor preparation. The volatile compounds can also be separated by stripping the vapor condensates. Acetic acid can also be recovered by liquid – liquid extraction.

The degree of chemicals recovery in the magnesium bisulfite process is 80 – 90 % for both

magnesium and sulfur. Therefore, the make-up requirement is 10 – 15 kg magnesium oxide and SO_2 per tonne of pulp.

Sodium-Based Sulfite Process. The recovery of chemicals in sodium-based sulfite processes is relatively complicated. For this reason, many different methods have been developed. Until now, only a few processes have proved useful in industrial practice. These are based on the principle of carbonation or pyrolysis. Direct sulfitation is a very interesting process, but it has never been used in a large-scale operation.

Carbonation. The concentrated liquor is burnt in a kraft-type furnace under reducing conditions. Sodium sulfide and sodium carbonate are present in the smelt. The sodium sulfide reacts with CO_2 from the flue gases or pure CO_2 or bicarbonate to yield sodium carbonate and hydrogen sulfide. The gaseous hydrogen sulfide is withdrawn and burnt to give SO_2. In an absorption tower, SO_2 reacts with sodium carbonate to give sodium sulfite or sodium bisulfite, depending on the stoichiometry. Carbonation processes in various modifications are still the only large-scale recovery processes for sodium sulfite liquor [172], [174].

Pyrolysis. Sodium and sulfur can be separated in a single step by pyrolysis of sodium sulfite waste liquor to produce sodium carbonate and solid carbon, while sulfur is obtained as gaseous hydrogen sulfide, which gives SO_2 on combustion. In comparison with other combustion processes, the pyrolysis temperatures are relatively low because no smelt should be formed in this process. Sodium carbonate is then dissolved in water and the carbon is filtered off and burnt in a separate step. Sodium carbonate and SO_2 are recombined to give the pulping liquor. In the industrial application of this process, however, substantial oxidation of sodium sulfur compounds occurs during pyrolysis, forming sodium sulfate. This must be carried through the process cycle as a chemical inert for pulping. Temperature control is also a problem in this process. If the temperature is too low, incomplete chemical conversion occurs. On the other hand, if the temperature is too high, a smelt is formed which contains sodium sulfides [175].

Direct Sulfitation. In the direct sulfitation, the smelt containing sodium sulfide and sodium carbonate, formed after spent liquor combustion under reducing conditions, is reacted directly with SO_2 to yield sodium sulfite. Sodium sulfide reacts to give sodium sulfite and H_2S, which is in turn converted to SO_2 by combustion. The industrial application of this process failed because thiosulfate is formed in considerable amounts from sodium sulfide, SO_2, and sodium sulfite in side reactions. Thiosulfate catalytically decomposes bisulfite during the digestion process.

Ammonium Sulfite Process. Today, ammonium sulfite solutions are seldom used for pulp production. The main reasons for this are the relatively high price of ammonia and the fact that base recovery is not cost effective. However, after combustion of the waste liquor the SO_2 can be recovered from the gas flow.

1.6. Pulp Bleaching

Lignin cannot be completely dissolved out of the wood in a pulping process because the rate of carbohydrate degradation in the final phase of delignification exceeds the rate of lignin dissolution. This leads to high yield losses and reduced pulp strengths due to the depolymerizing attack on the carbohydrates. For this reason, pulping must be stopped and the lignin that remains in the pulp must be removed by selective lignin dissolving chemicals. The residual lignin content of pulps depends on the type of raw material, the pulping process, and the required pulp properties. The residual lignin content of softwood pulps is normally higher than that of hardwood pulps, and sulfite processes can delignify more extensively than the kraft process. The normal residual lignin content of unbleached pulps is in the range 2 – 5 %. The elimination of this residual lignin is required primarily to brighten the pulps. However, mainly the structure and not the content of residual lignin is responsible for the color of unbleached pulp. Depending on the pulping reactions, more or less chromophoric groups are formed. Sulfite pulps always have a higher brightness than kraft pulps. Table 10 lists residual lignin contents and pulp brightnesses for various pulps.

Pulp bleaching is carried out in several stages, with the elimination of residual lignin as far as possible being the primary objective of the first

stages. The subsequent bleaching stages are responsible for brightening the pulp. Accordingly, bleaching is divided into prebleaching and final bleaching. The following abbreviations have been introduced for the bleaching stages:

C– chlorination
E– alkaline extraction
H– sodium or calcium hypochlorite bleaching
D– chlorine dioxide bleaching
O– oxygen bleaching in alkaline medium
P– peroxide bleaching
Z– ozone bleaching
Y– dithionite (hydrosulfite) bleaching

Table 10. Residual lignin content and brightness of various pulps

Type of pulp	Residual lignin content (kappa number)	Brightness, % ISO
Softwood kraft	27–33	25–30
Hardwood kraft	17–20	25–35
Softwood sulfite	15–25	55–60
Hardwood sulfite	12–22	55–65
Semichemical pulp/neutral sulfite	>80	50–60
Semibleached pulps	1–5	70–85
Fully bleached pulps	<1	85–92

1.6.1. Conventional Bleaching

In conventional pulp bleaching, chlorine, hypochlorite, and chlorine dioxide are used with or without intermediate alkaline extraction steps. For purposes of comparison, the amounts of chlorine chemicals employed are expressed as active chlorine, the varying molecular weight and oxidation potential of the chlorine-containing bleaching agents being taken into account. Thus, the oxidation potential of chlorine in hypochlorite is twice that of chlorine and the oxidation potential of chlorine in chlorine dioxide is five times that of chlorine. Taking into consideration the different molecular masses of the various chlorine compounds, the chemical equivalence compared to molecular chlorine is 1.05 for sodium hypochlorite and 2.63 for chlorine dioxide.

The use of chlorine and chlorine-containing compounds in pulp bleaching is ecologically the most problematic step in the entire process of pulp production. The chlorine compounds react in various ways with the lignin of pulp. Substitution reactions on the phenol rings of lignin, which occur to a large extent especially in chlorination, are particularly critical with regard to environmental pollution. The effluents from the bleaching process contain chlorinated organic compounds and cannot be disposed of by combustion with the spent digestion liquor. It must be released into the receiving stream, possibly after being subjected to biological treatment. It pollutes this stream considerably with its oxygen demand and its toxic and genotoxic effects [176–179]. Among others, various dibenzofurans and dioxins and even the extremely toxic 3,4,7,8-tetrachlorodibenzo-*p*-dioxin have been detected in small concentrations [179–183]. The combustion of bleaching waste liquor of this type would cause corrosion problems because of the formation of hydrochloric acid and would lead to the production of other highly toxic chlorinated products.

Chlorination. The first step in conventional lignin-dissolving bleaching is usually chlorination [184]. Chlorine is an effective, selective, and inexpensive bleaching agent that reacts exceptionally quickly with lignin. However, instead of brightening the pulp it usually causes a decrease in the brightness. Chlorination is normally carried out at a low temperature (ca. 30 °C). A short reaction time is sufficient, usually 30 min for sulfite pulp and up to 60 min for kaft pulp. The stock density for chlorination is usually ca. 3 %. Due to the fast reaction of chlorine with pulp, this low stock consistency is required to guarantee quick chlorine mixing and, hence, a uniform reaction. In particular, overchlorination of parts of the pulp and the resultant fiber damage should be avoided. However, the development of high performance mixers has made chlorine bleaching possible even at higher pulp consistencies of up to 10 % [185].

Chlorine water is normally used for chlorination. As a result of the more favorable reaction conditions and the possibilities now available for the optimal mixing of chlorine and pulp, chlorine gas has also been used recently on an industrial scale.

The most important function of chlorination is the conversion of the residual lignin in the pulp

to products that are soluble in water and alkali. Chlorine reacts primarily with the benzene rings of lignin in a nonradical reaction, in which both substitution and oxidation reactions take place. At the same time, addition reactions occur on the propane side chains of the lignin monomers if double bonds are present. However, since this is very seldom the case, addition reactions play only a minor role.

Electrophilic substitution with the chloronium ions formed by heterolytic cleavage of the Cl−Cl bond results in the replacement of the hydrogen atoms of the benzene ring by chlorine atoms [36], [186–189]. The number of chlorine atoms and their position on the ring depend on the structure of the lignin molecule. The use of excess chlorine can result in the maximum replacement of the three ring hydrogens at positions 5, 6, and 2 of the aromatic nucleus by chlorine atoms; 50 – 70 % of the total chlorine requirement of a bleaching sequence is consumed in chlorination.

The amount of chlorine used for different pulps is normally in the range of 1.0 – 1.7 g active chlorine per gram residual lignin. Lower amounts are required for sulfite pulps, which contain more reactive residual lignin, than for kraft pulps. Independent of the pulping process, softwood pulps require more chlorine than hardwood pulps. Under industrial chlorination conditions, ca. 1 – 2 chlorine atoms per lignin unit are attached to the phenol ring in substitution reactions. If the phenolic OH group in *para*-position to the side chain is not etherified, the substitution reaction occurs preferentially at the 5-position of the aromatic ring. If there is an alkyl, aryl, or methoxy group at position 5 of the ring, substitution occurs only at positions 6 or 2. The substitution of chlorine on the guaiacyl ring has the following priority: $6 > 5 > 2$.

Electrophilic substitution can also occur at position 1 of the phenol ring if the α-C atom of the side chain bears an OH or OR group. In this reaction, the aliphatic moiety is cleaved from the aromatic ring of the lignin phenylpropane unit (side-chain elimination). In contrast to the substitution of hydrogen atoms on the aromatic ring, the lignin molecule is fragmented in this reaction. Apart from the substitution reaction, oxidation also occurs during chlorination [190]. The initial step in the oxidation reaction is again the electrophilic addition of chlorine. Subsequent demethylation gives rise to quinoid intermediates, which can be oxidized further to dicarboxylic acids upon ring opening. The reactions occurring on the aromatic rings during chlorination of lignin are shown in Figure 28 [36].

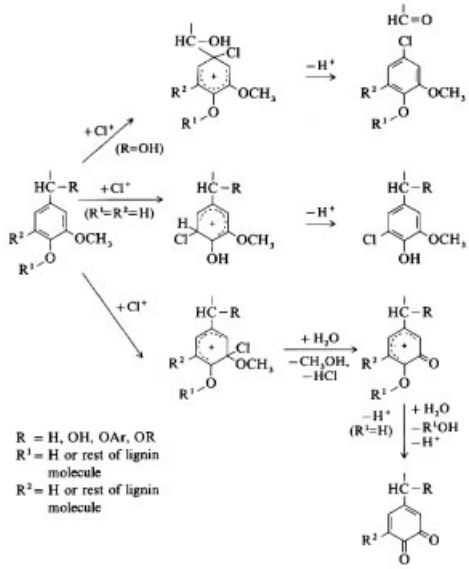

Figure 28. Substitution and dealkylation of phenolic and nonphenolic units by chloronium ions

Apart from the reactions described above, a large number of side reactions occur which give rise to many different chlorinated compounds, some of which are toxic, resistant to biological degradation, or potentially mutagenic or carcinogenic.

In industrial chlorination processes, the ratio of substitution to oxidation reactions is about 50 : 50. The chlorination conditions chosen previously favored the substitution reaction because carbohydrates are attacked thereby only to a minor extent. However, the chlorination process is now operated increasingly such that the oxidation reactions dominate, thus avoiding the aromatic chlorine compounds produced in the substitution reactions. This can be achieved by increasing the reaction temperature and pH.

Chlorine is a selective bleaching agent that reacts faster with lignin than with cellulose. Nevertheless, carbohydrates are also subject to attack by the chlorine radicals formed [184]. Carbohydrate degradation occurs via oxidation as well

as hydrolysis. The positions 1, 2, 3 and 6 of anhydroglucose units are very sensitive. If oxidation occurs at position 1, depolymerization takes place by cleavage of the β-1,4-glycosidic bond.

In other oxidation reactions, primary and secondary alcohol groups can be converted to aldehyde and keto groups, which facilitates depolymerization of the cellulose chains in subsequent bleaching steps.

Alkaline Extraction. Chlorination makes lignin partially soluble in water and alkali. After chlorination, in fact, sulfite lignin is for the most part already soluble in water, whereas the majority of kraft lignin can be dissolved only in alkali [191]. Due to the action of alkali, chlorine atoms bound to the aromatic nuclei are partially replaced by hydroxyl groups, increasing lignin solubility. Apart from the dissolving of chlorolignin, the function of alkali extraction in the production of dissolving pulps is to remove the hemicelluloses from the pulp as extensively as possible. For this purpose, higher amounts of alkali and severer conditions are required. In the production of paper pulps, however, hemicelluloses should remain in the pulp as far as possible and the DP value of the cellulose should be maintained at the highest possible level. Therefore, alkali extraction should be carried out gently.

In the extraction step, 1 – 1.5 % NaOH based on o.d. pulp is used for sulfite pulps, and up to 3 % for kraft pulps. The extraction time is 60 – 90 min at 40 – 60 °C. For dissolving pulps, temperatures up to 90 °C and up to 15 % sodium hydroxide based on o.d. pulp are used if pulps with extremely high α-cellulose contents are to be made. The consistency in the extraction stage is generally ca. 10 %.

As a result of the intensive dissolution of lignins and carbohydrates in the chlorination and extraction stages of pulp bleaching, highly polluted wastewater is produced which accounts for up to 90 % of the COD of the total bleach and ca. 100 % of the color load. The chlorinated lignin degradation products have, in part, a potential toxic effect on living organisms as well as a mutagenic effect. In particular, chlorine compounds with molecular masses of less than 1000 are problematic because they are small enough to diffuse through the membranes of living cells and can accumulate in fat tissues. The effluents from the chlorination and extraction steps can be degraded only to a limited extent by biological treatment. In single-stage plants, rates of degradation of approximately 50 % can be achieved, carbohydrates being broken down preferentially.

Hypochlorite Bleaching. After chlorination and alkaline extraction, the delignification of pulp is almost complete. The residual lignin content of the pulp is ca. 0.5 – 0.8 %. In conventional bleaching, this lignin is dissolved with oxidizing bleaching agents. In the past, hypochlorite was the agent of choice, but today chlorine dioxide is used almost exclusively, especially for kraft pulps. If both bleaching agents are used, hypochlorite is usually used first, followed by chlorine dioxide. Whereas calcium hypochlorite was used predominantly in the past, especially in calcium bisulfite pulp mills, sodium hypochlorite is used almost exclusively today in pulp bleaching [192].

In the hypochlorite bleaching step, the pH is the most important influencing factor. It can be varied within a wide range by varying the ratio of alkali to chlorine. At pH values less than 2, only elemental chlorine is present in the bleaching liquor. However, HOCl is formed at higher pH values, and the dissociation of this acid produces OCl^- ions at still higher pH values. Above pH 9.5, only OCl^- ions are present. Chlorination reactions take place predominantly at acidic pH values, whereas oxidation reactions occur preferentially at higher pH values. Since the latter are preferred, the initial pH of the hypochlorite solution should be adjusted to 11 – 12. The pH drops during bleaching owing to the formation of acidic products such as carbonic acid and oxalic acid. The pH should not fall below 8 at the end of the reaction because cellulose is also attacked extensively at lower pH values.

Normally, 1 – 2 % hypochlorite based on o.d. pulp is used at 30 – 50 °C. Temperatures up to 70 °C are also acceptable if the pH does not fall below 11 during bleaching. The consistency in the hypochlorite stage is 10 %, and the bleaching time is 2 – 4 h, depending on the raw material used, the pulping process, and the pulp properties required.

The oxidation reactions occurring at high pH in hypochlorite bleaching attack the free phenolic hydroxyl groups or the phenolic ethers at positions α and β of the phenylpropane side chain. The products of lignin degradation from

hypochlorite bleaching are methanol, organic acids, and carbon dioxide.

Hypochlorite is, however, not a selective oxidizing agent; it also reacts with carbohydrates. The following reactions can occur [192], [193]:

1) Oxidation of primary hydroxyl groups at C_6 to carbonyl and further oxidation to carboxyl groups
2) Formation of keto groups at atoms C_2 and C_3; ring cleavage to give dicarboxylic acids
3) Cleavage of the β-1,4-glycosidic bond and oxidation of the reducing end groups to carboxyl groups
4) Ring fission between C_1 and C_2 with formation of an aldehyde group at position 2 and a carboxylic ester at position 1; further oxidation to a carboxyl group at position 2 is possible

Depolymerization, which leads to poorer technological pulp properties, can be limited by using high pH values and lower reaction temperatures.

Chlorine Dioxide. Chlorine dioxide is an extremely effective and selective, but also expensive bleaching chemical. In the gas phase, chlorine dioxide is explosive at concentrations above ca. 12 %. It is, however, uncritical in dilute aqueous solution. The production of chlorine dioxide is carried out at the pulp mill by reduction of sodium chlorate. For a long time, it was not possible to use chlorine dioxide in pulp bleaching because of its toxicity, corrosivity, and the risk of explosion. It was the industrial application of chlorine dioxide shortly after World War II that made it possible to bleach even kraft pulps to high brightness, and led to the dominance of kraft pulping.

Chlorine dioxide has five oxidation equivalents. It can develop its full oxidation potential at a pH of 4 because all five oxidation equivalents can be used. In practice, the highest brightness is obtained when the pH at the beginning of bleaching is adjusted to approximately 6 and is around 4 at the end of the process [194].

The bleaching effect of chlorine dioxide is based on the oxidation of lignin. The initial step in lignin breakdown involves radical attack on structures with phenolic OH groups, forming phenoxy radicals, which undergo further reactions to give quinoid structures or, after ring cleavage, to form muconic acid derivatives [36], [195]. Ring chlorination reactions are caused by elemental chlorine, which is either contained in technical chlorine dioxide or formed by its decomposition.

Under normal bleaching conditions, chlorine dioxide hardly degrades carbohydrates. However, carbohydrate degradation can occur at temperatures above 80 °C, pH values above 7, and in the presence of excess chlorine dioxide, especially if the carbohydrates have already been attacked in the preceding bleaching stages.

The application of chlorine dioxide is a first response to the necessity of using environmentally harmless bleaching agents. The amount of bleaching agent used is small. Extensive degradation of lignin to aliphatic compounds occurs and chlorinated compounds are formed only to a very small extent.

Chlorine dioxide has an exceptionally positive effect on pulp brightness. It can degrade even largely unreactive lignin structures and is indispensible for the elimination of dirt specks in pulp.

The temperature during chlorine dioxide bleaching is between 70 and 80 °C and the bleaching time is 3 – 4 h. The pulp consistency is ca. 10 – 12 %. The amount of chemicals used can vary considerably depending on the position of the chlorine dioxide stage in the bleaching sequence and the type of pulp to be bleached. Normally, the amounts used are between 0.5 and 1.5 % active chlorine based on o.d. pulp.

Conventional Bleaching Sequences for Various Pulps. Depending on the type of wood, the pulping process used, the residual lignin content of the pulp, and the target brightness, different bleaching sequences must be used in the production of the various pulps. In pulps of all types, conventional bleaching starts with chlorination and extraction stages. In the case of readily bleachable sulfite pulps having low residual lignin contents, a hypochlorite stage follows CE prebleaching, and chlorine dioxide is used in the final bleaching. The bleaching of sulfite pulps to brightnesses of 90 % ISO and higher is possible with this CEHD standard sequence. Semi-bleached pulps having brightnesses of 80 % can be produced by using the sequence CEH.

In the conventional bleaching of kraft pulps, the sequence CEDED has become established.

Brightnesses of 90 % can be achieved while maintaining good pulp quality. Bleaching yields of ca. 93 % are obtained for softwood kraft pulps with an initial kappa number of about 30 (residual lignin content ca. 4.5 %). Bleaching yields of 96 – 97 % are achieved for sulfite pulps with low unbleached kappa numbers in the range 12 – 15.

The properties of pulp change during bleaching in a characteristic manner. As a result of the removal of hemicelluloses, the bonding potential of the fibers is reduced, which, in turn, reduces the breaking length in particular. On the other hand, the tear strength can increase.

1.6.2. Chlorine-Reduced and Chlorine-Free Pulp Bleaching

In response to pollution linked to the use of chlorine in pulp bleaching, considerable efforts have been made to replace chlorine by an environmentally less harmful chemical. Moreover, efforts are also being made to replace as far as possible the other chlorine containing bleaching agents and to develop bleaching sequences that require no chlorine-containing compounds. Then, the requirements would be met not only for the combined evaporation and combustion of waste liquor from bleaching and digestion, but also for the recovery of the alkali used in the various bleaching stages and for carrying out pulp production with a greatly reduced water consumption.

At the beginning of this development, attempts were made to replace the highly polluting CE prebleaching by chlorine-reduced or even chlorine-free alternatives. The first steps in this direction were the partial or complete replacement of chlorine by chlorine dioxide and the introduction of oxygen delignification. Ozone has also been employed on a laboratory and pilot scale. Peroxide has proved useful in enhancing oxygen delignification and is also being used increasingly in final bleaching.

Substitution of Chlorine Dioxide for Chlorine in Prebleaching. Owing to the different modes of reaction of chlorine and chlorine dioxide, the formation of chloroorganic compounds decreases with increasing use of chlorine dioxide [196–201]. At the same time, the selectivity of the bleaching reaction is improved [202–208]. Even small amounts of chlorine dioxide added in a chlorination stage reduce carbohydrate degradation considerably because chlorine dioxide functions as a radical scavenger and suppresses the oxidation of alcoholic OH groups to carbonyl groups by chlorine radicals [209], [210].

In industrial prebleaching, up to 70 % of the chlorine is replaced by chlorine dioxide. Especially in the case of kraft pulps, the complete replacement of chlorine is impracticable because the delignifying effect of chlorine dioxide is much lower than that of chlorine. The use of mixtures of chlorine dioxide with chlorine has the following advantages:

1) The process can be carried out at higher temperatures (60 – 70 °C) and higher consistencies (10 – 15 %). This corresponds to the conditions used in the following bleaching stages. Thus, a reduction of the energy and water requirements is possible.
2) Since chlorine dioxide has five oxidation equivalents, the use of this agent, compared to chlorine, corresponds to a reduction of the chloride ion content to 20 %.
3) The oxidative attack of chlorine dioxide on lignin leads to water-soluble degradation products, so that the amount of alkali used in the subsequent extraction stage can be reduced.
4) The use of chlorine dioxide produces effluents that are less chlorinated, less colored, and contain smaller amounts of toxic and genotoxic substances.
5) Pulps bleached with chlorine dioxide are of a better quality. Their extractives content is lower, and the brightness, viscosity, brightness stability, and technological properties are improved. Yields increase since chlorine dioxide attacks the carbohydrates in pulp to a lesser extent.

Alkali/Oxygen Bleaching. The increased use of chlorine dioxide also has disadvantages in that the investment and chemical costs are higher. Because of the high oxidizing power of chlorine dioxide, bleaching equipment must be made of corrosion-resistant material.

The treatment of pulp with oxygen under alkaline conditions is a chlorine-free alternative to prebleaching [211–219]. Sodium hydroxide is

usually used as alkali; magnesium hydroxide can be used in magnesium-based sulfite mills. The bleaching agent employed is gaseous oxygen, which is a very inexpensive chemical. However, oxygen is only very slightly soluble in water, especially at the high bleaching temperatures that must be used in the oxygen stage to achieve an adequate reaction rate. Therefore, the process must be carried out under pressure for a sufficient amount of oxygen to be available in the bleaching liquor. Oxygen in alkaline medium is, under these conditions, a relatively strong oxidizing agent. As a biradical, it can remove an electron from the phenolate ions present in the alkaline medium. The phenoxy radical thus formed is subject to further reactions. Hydroperoxides are produced which are further degraded by intramolecular nucleophilic attack of the peroxide anions [36], [218–221].

Oxygen does not, however, react selectively with lignin, but also attacks carbohydrates to a considerable extent [122], [123]. Above all, oxidations at atoms C 2 and C 3 of the anhydroglucose units occur. The unstable carbonyl compounds facilitate further degradation, which can result in the cleavage of the glycosidic bonds. To reduce carbohydrate degradation, the extent of delignification occurring in this stage must be limited. Inhibitors such as magnesium sulfate can be added to protect the carbohydrates. In industrial operation, delignification in the oxygen stage is restricted to about 50 % in spite of these additives. Consequently, oxygen bleaching is less effective than conventional prebleaching with chlorine.

Oxygen bleaching was originally developed as a high consistency bleaching process for pulp consistencies of 30 – 35 %. However, newly installed large-scale plants operate almost exclusively in the MC range (medium consistency), i.e., at consistencies of 10 – 15 % [224–227]. While high consistency bleaching requires temperatures above 100 °C and pressures up to 0.8 MPa, MC oxygen bleaching is carried out at temperatures below 100 °C and pressures of 0.2 – 0.4 MPa. The temperature can be reduced further to 70 °C in the bleaching of sulfite pulps. Modern MC oxygen stages usually operate under hydrostatic pressure. The oxygen is mixed in at the bottom of the bleaching tower under the hydrostatic pressure of the reactor. Thus, a reactor height of 20 m, for instance, results in a pressure of 0.2 MPa.

The reaction time in the oxygen stage is 30 – 90 min, and 20 – 40 kg alkali per tonne o.d. pulp and about 0.5 – 2.5 kg magnesium sulfate are charged. Depending on the delignification efficiency, 10 – 25 kg oxygen per tonne of pulp are consumed.

In order to improve delignification and to apply milder conditions, especially lower temperatures, hydrogen peroxide is added to the oxygen stage, particularly in sulfite pulp bleaching [228–231]. Sulfite pulps can be delignified by 50 – 60 % at 70 – 80 °C in the oxygen stage by adding 1 % peroxide. Moreover, the pulp properties profit from the milder reaction conditions.

Pulp Bleaching with Ozone. Since oxygen delignification is limited, other delignifying agents must additionally be used before final bleaching is performed. In practice, especially in kraft pulp bleaching, chlorine and chlorine dioxide are used today in varying proportions. Ozone has been tested for this purpose both on a laboratory and pilot-plant scale [232–240]. However, the use of ozone in pulp bleaching requires new technology. Like oxygen, ozone is only slightly soluble in water. Ozone is also very unstable and must be prepared at the mill by electrical discharge in air or oxygen. The ozone concentration in the air flow reaches 3 – 5 %, while concentrations of 6 – 10 % can be achieved with oxygen. In order that sufficient amounts of ozone reach the fibers, extremely low pulp consistencies of less than 2 % must be used; otherwise this process must be carried out in the gas phase at consistencies above 35 %. Pressurized medium consistency ozone bleaching is also possible. The water should be as free as possible of heavy metals to minimize catalytic decomposition of ozone.

The most important reaction of ozone with lignin is the cleavage of the β-O-4 bonds between the lignin units. Ozone can attack both the aryl and the alkyl moieties of the molecule. The attack on the aromatic rings can lead to ring cleavage between the C 3 and C 4 atoms. Double bonds in the aliphatic side chain are also attacked, forming carbonyl and peroxide structures [36], [241–244].

The intense attack on the carbohydrates is a disadvantage. The dominant reactions are cleavage of the glycosidic bonds and oxidations at

the C 2 and C 3 atoms, which facilitate further chain breakdown in subsequent alkaline treatment [222], [245], [246].

The formation of hydroxyl and perhydroxyl radicals is of considerable importance in these reactions. The hydroxy radicals in particular react very unspecifically and lead to severe carbohydrate degradation. Since the formation of these radicals is favored by hydroxyl ions, the ozone treatment should be carried out under acidic conditions. For this reason, the pulp must be acidified, and SO_2 and sulfuric acid have proved effective for this purpose. Since ozone is extremely reactive, a very short reaction time can be used, provided that thorough mixing of ozone and pulp takes place. If this is the case, the entire duration of treatment including an appropriate after-reaction time can be less than 10 min. The temperature in the ozone step should be as low as possible to avoid harmful side reactions. This applies in particular to low-consistency bleaching. In high-consistency ozone bleaching with very thorough mixing, temperatures up to 50 °C can be used.

Low-consistency ozone bleaching is of advantage with regard to the selectivity of lignin degradation [247]. However, high-quality water and low temperatures are required. If the water cycle of the mill is largely closed, both requirements can be met only with difficulty. The stronger attack on carbohydrates in high consistency ozone bleaching must be restricted by limiting the amount of ozone charge. Amounts exceeding 1 % ozone based on o.d. pulp should not be applied. The pulp can be delignified by about 10 kappa numbers with this amount of ozone. Under these conditions, ozone can be used for pulps with a high lignin content as the second bleaching stage after oxygen prebleaching. Alternatively, it can be used as the sole delignifying agent for sulfite pulps with low initial kappa numbers. As a result of its higher delignifying capacity, ozone, unlike oxygen, is also well suited to the bleaching of low-residual-lignin pulps. The kappa number after the ozone stage should be in the range of 1.5 – 3.

An emerging technology is ozone bleaching at medium consistency. The MC ozone bleaching must be carried out under pressure to reduce the gas volume and to have a good mixing effect. On the other hand, the disadvantages of ozone bleaching at low or high consistency are greatly eliminated. First pilot plant results have been satisfactory; in particular, carbohydrate degradation was substantially reduced.

Until now, ozone has not been used for large-scale bleaching of pulp. However, the existing laboratory and pilot plant results and the necessity for chlorine-free bleaching suggest that the industrial application of this bleaching agent is in the offing.

Peroxide Bleaching. Hydrogen peroxide in alkaline medium is well suited to the final bleaching of chemical pulps [248]. The bleaching of mechanical pulp has shown that it destroys the chromophoric groups present in lignin by cleaving conjugated double bonds. At 70 – 80 °C, the highly nucleophilic perhydroxyl ion formed is capable of further degrading quinoid lignin structures, produced by the attack of electrophilic bleaching chemicals. This is generally accompanied by ring fission [36], [221], [249]. As mentioned above, peroxide is used in prebleaching as an intensifying agent in the oxygen stage. In final bleaching, it is often used as the concluding treatment. The amounts used are 1 – 2 % at consistencies of 10 % and temperatures of 70 – 80 °C. An advantage of the use of peroxide in final bleaching is the high brightness stability of peroxide-bleached pulps.

Chlorine-Free Bleaching Sequences. The choice of the bleaching sequence to be used, if chlorine-free bleaching agents are applied exclusively, depends on the raw materials used in pulping, the pulping process, the residual lignin content of the pulp, and the demands made on the quality of the bleached pulp. The sequence OZEP is suitable for sulfite pulps with high residual lignin contents. After the oxygen stage, the kappa number of the pulp should be 10. The kappa number can then be reduced to ca. 3 by the use of 0.8 % ozone based on o.d. stock. These pulps can be brightened to 90 % ISO by subsequent alkaline extraction and the addition of 1 % hydrogen peroxide in the final bleaching step.

At present, kraft pulps still cannot be bleached economically to high brightness levels in a chlorine-free sequence. In the case of softwood kraft pulps, strength losses of 20 – 25 % must be accepted in chlorine-free bleaching on reaching brightness levels of 85 % ISO.

On the other hand, several sulfite mills have switched to completely chlorine-free bleaching sequences; oxygen and hydrogen peroxide are employed. Prebleaching is achieved with alkali/oxygen, partially intensified by hydrogen peroxide. Peroxide is used in final bleaching, usually in a two-step process. Bleaching sequences of this type give brightnesses up to 85 % ISO. Compared with conventional bleaching and ozone bleaching, however, strength losses must be accepted.

1.6.3. Technology of Bleaching

Today, bleaching is always carried out in a closed system. The individual bleaching stages take place in bleach towers in which the material flows either from top to bottom or vice versa (Fig. 29). Washing units (filters or sieve presses) located between the bleach towers remove the used chemicals and the reaction products from the pulp. The washing efficiency is about 80 %. Countercurrent operation is employed as far as possible to minimize water consumption. In modern plants, the tendency is to use similar reaction conditions in all bleaching stages. The bleaching stages are preferentially performed at a medium consistency of 10 – 15 %. Bleaching towers with an upward flow can be used for this purpose without problems. The addition of the chemicals occurs before the pulp enters the tower. An exception is the chlorination step, which is performed at consistencies of 3 – 5 % and at the lowest possible temperatures. However, the temperature and the consistency can be adapted to those of the other bleaching stages by partially replacing chlorine by chlorine dioxide. Chlorination towers are usually rubber lined, while chlorine dioxide towers are made of high-alloy stainless steel.

High-consistency oxygen bleaching requires a technology which differs from conventional bleaching stages. Alkali and the degradation inhibitor (magnesium sulfate) are charged before the pulp enters the reactor. After mixing, the stock is adjusted to a consistency of at least 30 % and then fluffed at the top of the reactor. The tower is operated from the top downwards at a high oxygen pressure. The tower can be equipped with movable trays over which the pulp moves and is gassed with oxygen. Before being discharged, the pulp on the reactor floor is diluted to give a pumpable suspension and it then enters a washing stage.

On the other hand, the medium consistency oxygen stage used more frequently today uses an up-flow. Alkali and magnesium sulfate are added before the pulp enters the tower. Oxygen is added in an additional mixer. Bleaching occurs either at hydrostatic pressure or with oxygen overpressure. The stock is discharged continuously from the top of the tower and is transported to a washer. The MC oxygen bleaching technology is not problematic because no explosive mixtures of carbon monoxide/oxygen can form. Increasing the consistency in the bleaching stages saves energy and reduces water consumption, but problems with the mixing of bleaching chemicals and the stock can occur at higher consistencies. The water and energy requirements in the medium-consistency range can be reduced by adapting the reaction conditions in the different bleaching steps and by using countercurrent washing. A system of this type is shown in Figure 30.

A high-consistency ozone stage can be operated with the technology of the high-consistency oxygen stage. However, a pressure reactor is not required. Ozone can be added in the fluffer. On the other hand, up-flow towers can be used for low and medium consistency ozone bleaching.

1.7. Environmental Aspects of Pulp Production

The production of pulp is associated with considerable pollution. The type and extent of pollution depend on the pulping and bleaching processes used and on the internal and external measures taken to reduce pollutants. If all the possible and economically justifiable measures are taken to reduce the environmental burden in the production of kraft pulp, pulpmaking with a wastewater COD below 50 kg/t of pulp appears to be possible. The specific water requirements of a mill of this type is ca. 50 m^3 per tonne of pulp. In the production of sulfite pulp, the COD burden is even lower. Values of 30 kg COD per tonne of pulp and water consumption corresponding to that of kraft pulp production are possible. Base losses of 1 – 1.5 % based on o.d. pulp and sulfur losses, measured as SO_2, that are

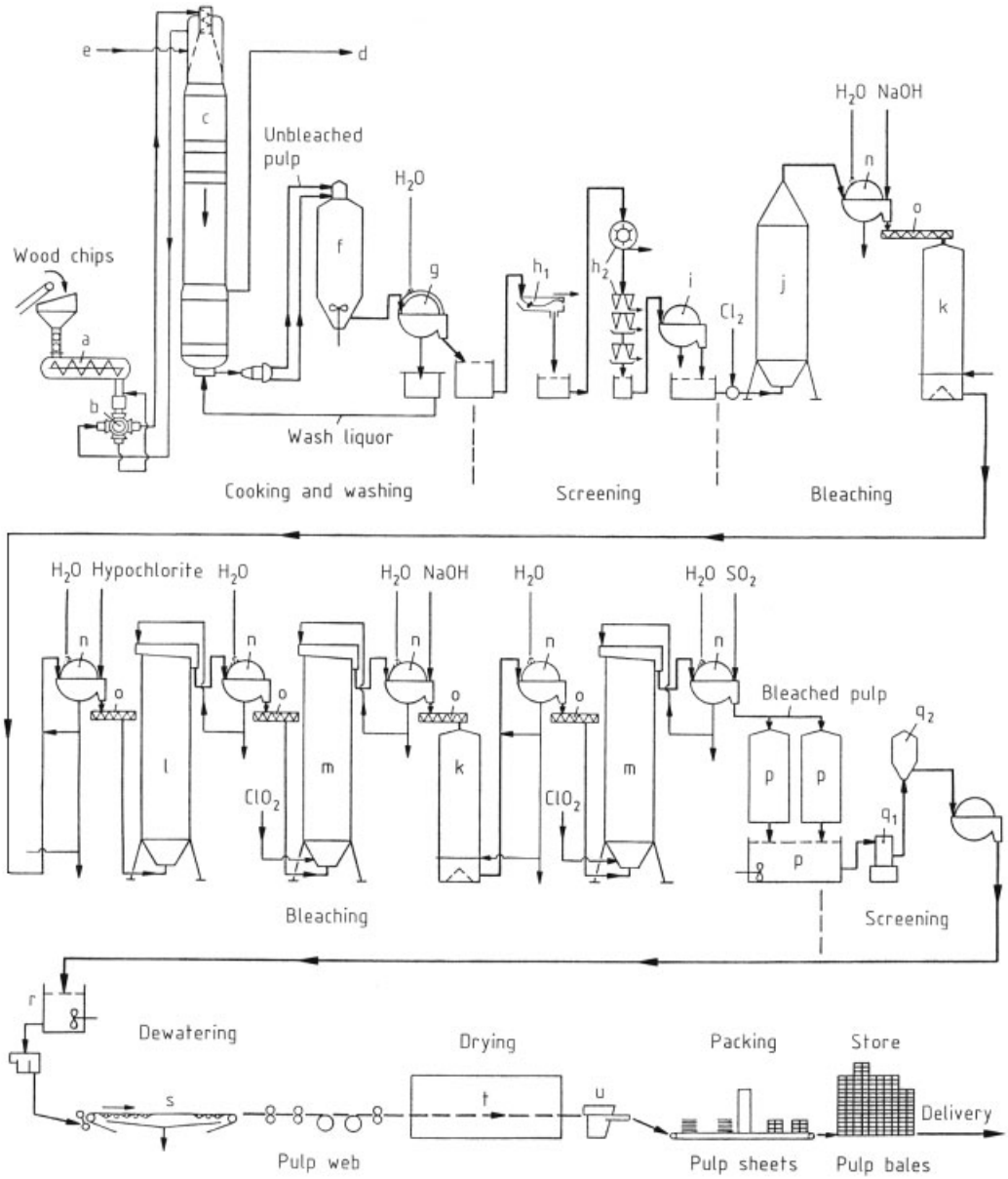

Figure 29. Schematic of a pulp mill
a) Steaming drum; b) High-pressure valve to the continuous digester; c) Continuous digester for kraft and bisulfite pulping; d) Waste liquor outlet; e) digestion liquor inlet; f) Blow tank; g) Filter for pulp washing; h_1) Coarse screening; h_2) Fine screening; i) Filter thickener; j) Chlorine bleaching tower; k) Bleaching tower for alkali treatment; l) Bleaching tower for hypochlorite treatment; m) Bleaching tower for chlorine dioxide treatment; n) Wash filters; o) Mixers; p) Storage tanks; q_1) Coarse postscreening; q_2) Fine postscreening; r) Machine chest; s) Dewatering machine; t) Tunnel dryer; u) Sheet cutter

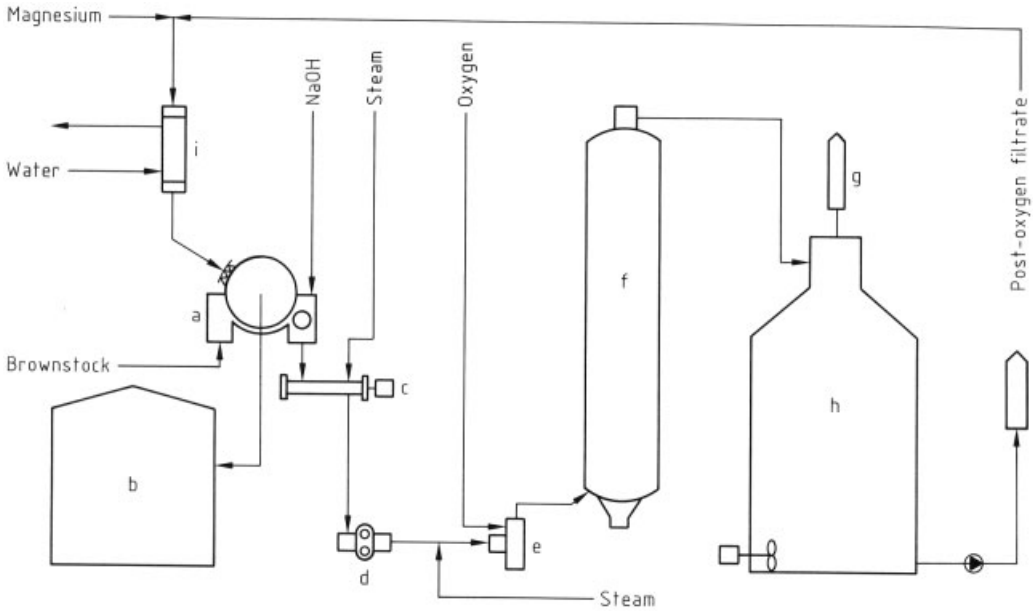

Figure 30. A medium consistency oxygen bleaching system
a) Brownstock washer; b) Seal tank; c) Steam mixer; d) Medium consistency pump; e) Oxygen mixer; f) Oxygen reactor; g) Vent; h) Oxygen blow tank; i) Post-oxygen washer

only insignificantly higher are incurred by modern kraft and sulfite pulp mills. Today, combustion furnaces operate with dust emissions below 50 mg/m^3, NO_x emissions below 300 mg/m^3, and SO_2 emissions below 1.5 g/m^3.

Sources of Emission. In the pulping process, not only gaseous emissions, but also wastewater is produced. Today, the washing efficiency after the digestion process exceeds 99 %. The original load of the entire wood substance dissolved by pulping amounts to ca. 1500 kg COD per tonne of pulp and a final charge of ca. 15 kg COD is produced. Screening and liquor evaporation cause further pollution. In addition, pulp bleaching is a significant source of pollution. A rough assessment shows that about half of the COD of a pulp mill is caused by pulp production and the other half by pulp bleaching. In modern sulfite and kraft pulp mills, the gaseous sulfur compounds formed are withdrawn at the point of emission and charged into the chemical recovery system. The separation of gaseous sulfur compounds from the flue gases takes place in wet cleaners. The major sources of SO_2 in the sulfite process are the digester, the blow tank, the pulp wash, and the chemical recovery system. In the kraft process, lime kilns are a further substantial source of emissions. Considerable amounts of chlorinated organic compounds are formed particularly in the chlorination and extraction stages of conventional bleaching. These compounds can be determined as organic halogens (AOX values) adsorbable on active charcoal. Ranges of AOX values for various hard- and softwood kraft pulps are listed in Table 11.

Table 11. AOX values from the bleaching of various pulp

Bleaching sequence	AOX, kg/t	
	Hardwood pulp	Softwood pulp
CEHDED	4–5	8–10
O($C_{85}D_{15}$) EDED	2.5–5	4–5
O(D_{70}/C_{30}) (EO) DED	1–2	1.5–2.5

See Section 1.6 for abbreviations.

The chlorinated organic material can be divided into high and low molecular compounds, the dividing line being at a molecular mass of 1000. The high molecular mass fraction is not directly toxic. However, biological degradation

of these compounds can give rise to low molecular toxic products. The low molecular portion of the chlorinated organic compounds can, in turn, be divided into a water soluble, an ether soluble, and a highly volatile fraction. The ether soluble substances, particularly the chlorinated phenolic compounds, are especially hazardous, some of them being highly toxic, chemically very stable, and able to accumulate in the fatty tissue of water organisms via the food cycle. This fraction can also contain polychlorinated dibenzofurans and dibenzodioxins.

Prevention Measures. Pollution caused by pulping can be reduced by both internal and external measures. Internal measures refer to changes in the production process that lead to a reduction of pollution. These include extended digestion to reduce the residual lignin content of unbleached pulp, increasing the black liquor recovery rate by recycling the wastewater into the process, and switching to chlorine-free bleaching.

External basic prevention measures include the elimination of pollutants formed. Wastewater treatment for the reduction of pollutants can be carried out by filtration, sedimentation, flotation, and removal of solids. These processes can be supplemented by physicochemical procedures such as precipitation of dissolved compounds or flocculation.

Dissolved organic compounds can be degraded biologically in subsequent process steps. Both aerobic as well as anaerobic wastewater treatment are used for this purpose. Biological treatment is indispensible in modern pulp mills. To achieve the highest possible rate of degradation, multistage processes are frequently used; the sequence anaerobic/aerobic has proved to be especially effective. Depending on the wastewater composition, the COD of pulp mill effluents can be reduced by 50 % in single-stage plants. In two- or three-stage treatment plants, a decrease in the COD of up to 80 % and a AOX reduction of 50 % can be achieved.

In pulp production, the atmospheric emissions can be greatly reduced by rigorous detection of all flows of exhaust air and their return to the chemical recovery system, and by the purification of the flue gases.

1.8. Pulp Properties and Applications

Pulps have a wide range of properties and uses. The technological properties of pulps depend largely on the raw material, the pulping process, the extent of delignification, and the bleaching sequences applied. Basic differences exist between paper and dissolving pulps. The latter must not only be absolutely free of lignin, but the content of extractives, inorganic impurities, and hemicelluloses must not exceed relatively narrow limits. In addition, the degree of polymerization of cellulose must fall within a definite range, depending on the intended use. Dissolving pulps are produced by using the prehydrolysis–kraft process or the acid sulfite process. In the prehydrolysis–kraft process, the hemicelluloses are prehydrolyzed to such an extent that pulp with a high α-cellulose content is formed in subsequent digestion. The bleaching conditions, especially in the extraction stage, are also characterized by relatively stringent conditions. This refers to both the alkali dosing as well as the bleaching temperatures. A survey of the requirements to be met by dissolving pulps for the production of various products is given in Table 12.

For pulps used in the production of paper and board, the strength properties are of special importance. Kraft pulps are clearly superior to sulfite pulps in this respect, especially with regard to the tear strength. The pulp fibers form the supporting structure of paper and board, while mechanical pulp fibers improve the opacity and the volume of paper. Mineral fillers also increase the opacity and the surface smoothness of writing and printing papers. On the other hand, recycled fibers from wastepaper contributes little to paper strength. They are, however, inexpensive and their reutilization reduces waste problems. Indeed, wastepaper is used as a substitute for mechanical pulp in many cases. Today, the sulfite processes are of secondary importance in the making of paper pulp because their strength properties are inadequate for many applications. For instance, the tear strength of acid bisulfite pulps is only about two thirds that of kraft pulps. However, they have the advantage of easy bleachability and beatability. Semichemical pulps made by the bisulfite, neutral sulfite, or alkaline sulfite processes have good strength properties and, compared to competing

Table 12. Specifications of dissolving pulps for various products

Specifications	Viscose					Cellulose derivatives				
	Cordrayon	Modal fibers	Cellulose guts	Cellulose film	Acetyl cellulose	Cellulose [c] ethers	Cellulose [d] ethers	Cellulose nitrates	Methyl-cellulose	Carboxy-cellulose
Brightness, % ISO			>90	>90		89	93	>90	>88	>88
R_{18} [a]	>99	>96	>95	>93	96	93 – 95	92	>92	>92	>90
R_{10} [a]	>98	>93	>93	>91	94	91 – 94	89	>90	>90	89
Extract in CH_2Cl_2, %	0.02 [b]	0.02 [b]	<0.2	<0.2	0.1	0.1	0.4	<0.2	0.2	0.2
Ash	0.15	0.18		<0.1	0.1	<0.2	0.1	0.1	<0.2	0.2
DP_W [e]	760 – 860	460 – 520	900 – 1700	900 – 1200				900 – 4500	500 – 4500	700 – 4500
Specific density, kg/m^3					600	650 – 800	800	500 – 700	ca. 700	ca. 700

[a] Residues in 10 or 18 % sodium hydroxide at 20 °C.
[b] In ethanol.
[c] Softwoods.
[d] Hardwoods.
[e] Degree of polymerization.

pulps made by the kraft process, better brightness levels. A survey of pulps made by various processes, their properties, and possible uses is presented in Table 3.

2. Mechanical Pulp

2.1. Introduction

Mechanical pulp denotes the semifinished product, used for the production of paper and board, which is made solely by the use of mechanical energy to separate the wood into individual fibers and fiber fragments. According to present findings, the internal processes occurring in such nonchemical production processes are predominantly of a physical nature.

There are two main processes used for the production of mechanical pulp. In the stone groundwood (SGW) process, logs are pressed against a rotating grinder stone with simultaneous addition of water. Refiner mechanical pulps (RMP, TMP) are produced by defiberizing wood chips between refiner disks to give fibers and fiber fragments.

High yields are characteristic for these two processes, i.e., the losses occurring during the processes, mainly in the form of dissolved ligneous substances, are low.

The most serious disadvantages associated with all types of mechanical pulp are their low resistance to ageing, which results in loss of strength and the tendency to discolor. For this reason, the large scale use of this pulp is limited to "short-lived" paper such as newsprint.

In the Federal Republic of Germany mechanical pulp accounted for about 15 % of the total production of paper and board from 1986 to 1988 [250]. This percentage is also more or less valid for world production.

2.1.1. The Raw Material, Wood

For the production of mechanical pulps, which to a large extent are responsible for the structural stability of the paper produced, softwood is used worldwide. The most common woods used in Europe are spruce and fir, while in climatically more favorable countries, woods such as the fast growing *Pinus insignis, Pinus maritima,* hemlock, etc. are used. Softwoods have two predominant advantages over hardwoods. The fiber content of softwoods is ca. 90 % as compared to 50 % for hardwoods and, in addition, the fiber length of softwoods of 3 – 4 mm is considerably larger than the 1 – 2 mm of hardwoods. Under the same manufacturing conditions these two characteristics are always responsible for the higher strength of softwood mechanical pulps.

The logs required for the groundwood process are obtained by thinning out forests and are usually 10 – 20 cm in diameter. The bark has to be removed before the logs are processed. The

wood chips used for the production of refiner mechanical pulps are predominantly obtained from sawmill residues. It is of great importance for the quality of mechanical pulp that the moisture content of the wood is as high as possible. This results in a higher flexibility of the fibers, which improves their fibrillation.

2.1.2. Characteristics Required of Mechanical Pulps

About 90 % of the mechanical pulp produced is used to make graphic papers, such as newsprint; uncoated and coated magazine papers, e.g., SC (super calendered), and LWC (light weight coated) papers [251]. The desired properties of mechanical pulps, used for graphic papers, are as follows:

1) The mechanical pulp should contribute substantially to the structural strength of the paper. Newsprint, for example, may contain more than 80 % mechanical pulp.
2) The printability of the paper is mainly determined by the quality of the mechanical pulp.
3) Since graphic paper is usually printed on both sides, a very low show-through and strike-through of the printing ink is very important. In offset printing, the surface of the paper should have a high picking resistance and be free of dust.
4) The drainage ability (freeness) of mechanical pulp has to be high to allow the paper machines to be operated at high speeds (ca. 1400 m/min). Furthermore, the paper has to be able to cope with the stresses within the printing machines.

The technological data of mechanical pulp is determined by means of standardized test methods. The United States and its neighbors follow the TAPPI regulations [252], the Scandinavian countries use the Scan methods [253], and Germany and its neighbors, the Zellcheming regulations [254].

Although the methods used vary and the results are thus not directly comparable, all standards require determination of the following mechanical pulp properties:

1) Drainage characteristics
2) Stability properties such as tensile strength tearing strength, and picking resistance
3) Optical properties such as brightness and opacity
4) The quantitative composition with respect to the portions of long fibers, fiber fragments, and fines as well as the size of the shives

Furthermore, proof printing and other tests provide information about the suitability of a mechanical pulp.

All in all, the test methods in question permit adequate assessment of the mechanical pulp and are essential for mechanical pulp production.

2.2. Groundwood

Until the 19th century, used textiles were virtually the only raw material used for papermaking. With the constantly increasing demand for paper, the supply of new raw materials was an urgent and widely known problem. In 1843 a weaver from Saxony, F. G. Keller, shredded wood by pressing it against a hand-driven grinder stone adding water to the grinding process. From the resulting pulp, he then made a sheet of paper. By using wood, Keller found a raw material which is still the most important in the paper industry. After this discovery was adopted, the first mechanical grinder was put into operation in 1852.

It was not until 1930 that the internal processes involved in grinding were systematically studied with the aim of obtaining a higher quality groundwood [255]. The investigations carried out in Canada in the early 1960s on an experimental grinder operating under pressure led to the pressure grinding process for the production of pressure groundwood (PGW) [256]. The Tampella company of Finland then developed this process and introduced it in 1978 [257]. The thermogrinding process for the production of thermogroundwood (TGW) was introduced by the Voith company in 1984 [258]. This grinding process is carried out at a normal pressure and with process temperatures that reach the feasible upper limit.

Theory of the Grinding Process. The purpose of the grinding process is to disintegrate the wood fiber bond system. The logs are pressed against the rotating grinder stone with the wood

fibers parallel to the axis of the stone (Fig. 31) [259]. Based on the usual operating parameters, a stone particle requires about 0.02 s to pass the grinding zone, removing 0.03 mm of the wood substance. The total or partial separation of the fibers from the wood and the properties of the groundwood achieved with this process depend on a series of factors. These are, above all, the surface structure of the grinder stone, the nature of the abrasive particles (round or sharp-edged) and their depth of penetration into the wood. Fiber removal occurs as soon as the stone particles are in contact with the wood, and the availability of sufficient water is of basic importance for heat dissipation and friction reduction. When sliding along the wood surface, a stone particle not only causes deformation of the fibers, but simultaneously changes the direction of already separated fiber parts (Fig. 32). These deformations lead to the fracture of the fiber-to-fiber bonds and of the bonds within the fiber walls. The results of this effect, comparable to "brushing" of the deflected fiber parts, are the fibrillation of the fiber walls and the creation of fines [260].

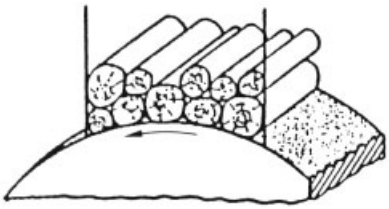

Figure 31. Arrangement of the wood on the grinding stone

It can be assumed that most of the energy put into the grinding process is transformed into heat; values of 90 to 99 % have been reported. The resulting increase in temperature is effective in the wood and in the grinding zone [261]. Due to the poor thermal conductivity of wood, temperatures of 180 °C or higher may develop within the wood directly above the grinding zone [262]. In the conventional grinding process at normal pressure, the average temperature in the grinding zone is below 100 °C. Evaporation may occur only in those areas where the amount of water present is insufficient to dissipate the heat [261].

During the development of grinding processes it was confirmed that the temperature is a decisive factor in the production of groundwood. The explanation for this was given by Goring, who found that lignin with a water content of more than 30 % softens at temperatures as low as 90 °C (Fig. 33) [263]. Since the lignin-rich middle lamellae of the wood fibers contribute substantially to the mutual bonding of the fibers and to the structural stability of the wood, the softening of lignin must inevitably facilitate the exposure of the fibers.

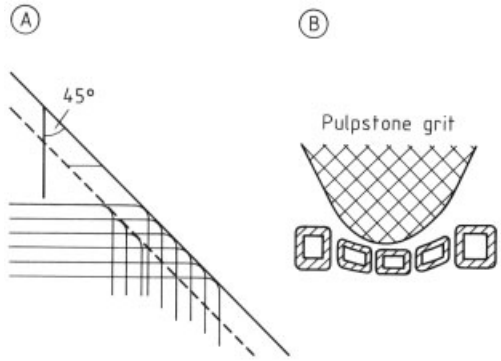

Figure 32. Processes occurring in the grinding zone
A) Angle of removal; B) Deformation of fibers by pulpstone grit

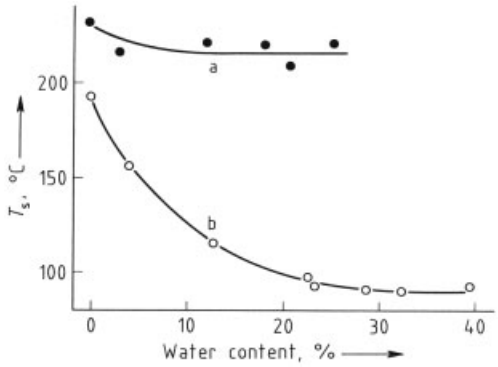

Figure 33. Variation of softening temperature T_s with water content
a) Sulfite pulp; b) Periodate lignin

A sharp-edged abrasive particle has an energy-saving cutting action on the wood structure, whereas round particles exert a predominantly fibrillating effect. The result is a groundwood with superior strength properties, but with the disadvantage of a lower production rate. The ratio of the motor energy to the amount

of groundwood produced is called the specific energy requirement (kWh/t), and is an important process parameter. An increase in the specific energy requirement is always linked with better strength properties, but also with a lower drainage capacity. The magnitude of the specific energy requirement depends on the interaction of all grinding conditions and the surface characteristics of the grinder stone. The effect of the grinding conditions – such as the grinding surface area, the peripheral speed, the specific grinding pressure, etc. – on the grinding process and on the groundwood quality has been the subject of detailed studies, which have in turn contributed to the optimization of the process [255], [264–266].

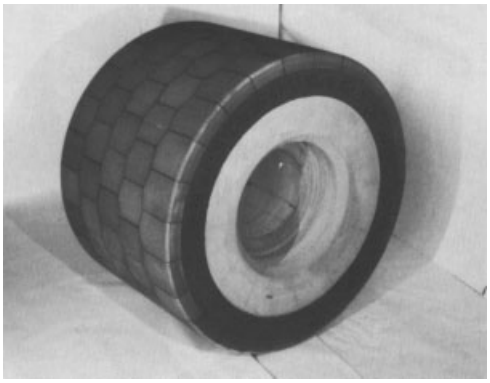

Figure 34. Ceramic-bound grinder stone

2.2.1. The Grinder Stone

Apart from the cement stones, now of declining importance, the grinder stones used are exclusively of the ceramic-bound type (Fig. 34) with a diameter 1.5 – 1.9 m. The stone surface consists of segments fastened to the concrete core of the stone [267]. The abrasive particles used consist either of sharp-edged carborundum (SiC) or, increasingly, of the softer and less sharp-edged corundum (Al_2O_3). A pattern is impressed on the grinder stone surface (Fig. 35) by traversing sharpening rollers, a procedure which has to be repeated periodically depending on the wear of the pattern. The intervals at which the stone is resharpened vary greatly between 1 and 100 d. This sharpening process must be carried out with extreme care because at present there is no practically useful technique available for measuring the pattern [268], [269]. This sharpening pattern has two functions. It meets the requirements for the separations of the fibers from the wood substances, and it transports the water needed in the grinding zone within the grooves of the sharpening pattern.

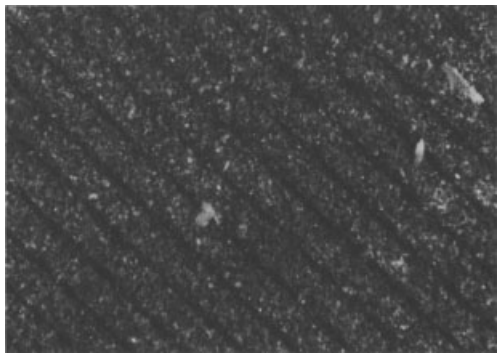

Figure 35. Pattern on a grinder stone surface

2.2.2. Groundwood Production

For the production of groundwood two grinder systems are predominantly used. In press grinders the logs are transported automatically into chambers, where they are pressed against a grinder stone by means of hydraulic presses (Fig. 36 A). In continuous grinders, logs are pressed against the stone by means of specifically shaped chains (Fig. 36 B) [270]. The advantage of the latter system is that the continuous grinding process better utilizes the motor load. The two-press grinder, on the other hand, has a higher specific production due to its larger grinding surface area. A comparison of the two grinding systems is given in Table 13.

Table 13. Comparison of the two types of grinder for wood of 1 m diameter

	Two-press	Continuous
Grinding surface, m^2	2.0	1.2
Motor output, kW	5000	3000
Motor output utilized, kW	4500	2700
Production (1250 kWh/t), t/d	78	50
Area required/unit, m^2	26	13
Space required/unit, m^3	100	80

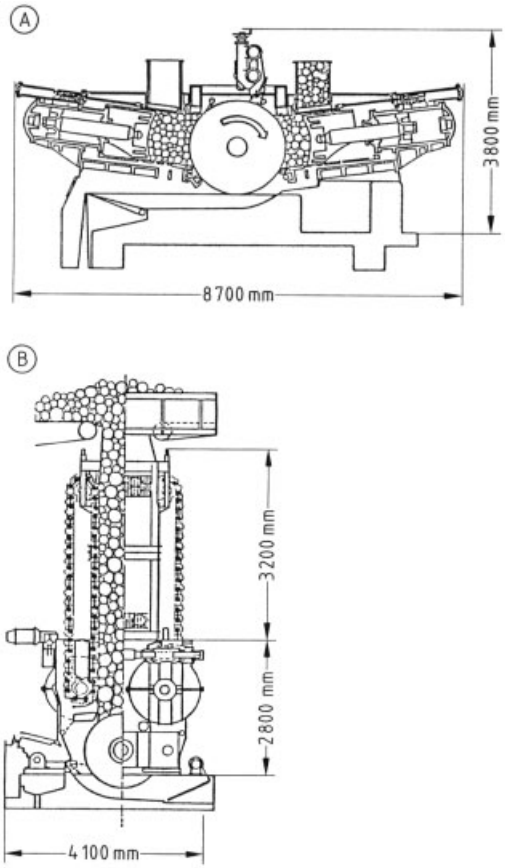

Figure 36. A) Tampella press grinder; B) Voith chain grinder

The grinders are equipped with control systems that keep the power input, the production rate, or the specific energy consumption constant. Each of these control systems has its advantages and disadvantages. In fact, combinations of two of the above control systems are usually applied.

2.2.3. Newer Methods of Groundwood Production

The superior strength properties of refiner mechanical pulps, especially of TMP, forced further development of the conventional grinding process if it was to remain competitive. The findings on the influence of process temperature [256] led to the development of the pressure grinding process (PGW process). In this process, the Tampella pocket grinder is sealed so that the grinding process takes place at overpressure. Overpressures of up to 0.3 MPa, produced with compressed air, allow the process to be operated with white water temperatures of 95 °C and grinder pit temperatures of 125 °C. The more intensive softening of lignin associated with these temperatures results in improved groundwood qualities [271–275]. The strength of pressure groundwood is 30 – 35 % higher than that of conventional groundwoods produced with the same specific energy consumption. Although this process represents a tremendous improvement, the technical and financial requirements are considerably higher.

The manufacturers of continuous grinders were also compelled into improving the quality of their groundwood. This was achieved by reducing the heat losses [276] which result from evaporation in the grinding zone and by maximizing the process temperature, without using overpressure, by means of stability controllers [258]. The qualities of groundwood obtainable from the thermogrinding process (TGW process) are about 15 – 20 % higher than that of conventional groundwood and are comparable to lower-grade pressure groundwoods. A certain advantage is the fact that the conversion of existing grinders to the TGW process is relatively inexpensive.

2.3. Refiner Mechanical Pulps

2.3.1. Historical Survey

Refiners have been available since the thirties and were used particularly in the fiberboard industry. After 1955 they also found application in the paper industry for the production of mechanical pulps [277]. Sawmill residues are used as a cheap raw material by being transformed into wood chips. The refiner mechanical pulp (RMP) process, which is hardly used today, was the first stage within this development. After 1960, the advantages of higher process temperatures were exploited in the thermomechanical pulp (TMP) process. This process is still being improved and is used for the production of the majority of refiner mechanical pulps.

Since 1980, the chemothermomechanical pulp (CTMP) process is becoming increasingly

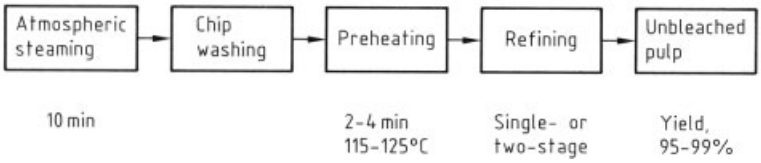

Figure 37. Schematic of the TMP process

important. Because wood chips are readily impregnated, small additions of chemicals give further improvements in the pulp properties. A further increase in the amount of chemicals used, as in the more recently developed CMP process, results in mechanical pulps comparable to semi-chemical pulps.

2.3.2. Theory of the Refiner Process

The objective of the refiner process is the same as that of groundwood production, i.e., the breaking loose and fibrillation of the wood fibers [278]. As a result of the large surface area of wood chips, the advantages of higher temperatures can be exploited to a much greater extent than in the grinding process. In the TMP process, the chips are subjected to thermal pretreatment, and the disintegration and defibration process is carried out in the refiner at overpressure. To obtain a large proportion of intact fibers, fiber separation should preferentially be a fiber-to-fiber action rather than an action between the refiner plates and the wood [279], [280]. Pulp consistencies in the refining zone of at least 40 % are suitable for this purpose. Nevertheless, the design of the refiner plate pattern exerts a marked influence on the quality of the wood pulp.

TMP with good strength properties is obtained as a result of the relatively small number of cutting actions. However, the specific energy consumption is 40 – 50 % higher than that of the grinding process. The temperatures resulting from the excess pressures of up to 5 bar in the refiners permit the recovery of the extra energy demand by means of heat exchangers [281], [282].

2.3.3. TMP Production

A schematic of the TMP process is shown in Figure 37 [283]. Since the pulping of wood chips to the desired pulp quality is usually not possible in a single refiner stage, two stages are used. The pulping of screen rejects is often carried out in the second refiner stage.

The capacity of a two-stage refiner system is much higher than that of a grinder; 300 o.d. t/d of TMP can be produced at an installed motor power of 15 000 kW per refiner. The specific energy consumption, including that of the auxiliary machines is ca. 2100 kWh/o.d. t. The arrangement of the refiner disks can be quite different (Fig. 38) and each of these models has its advantages [284].

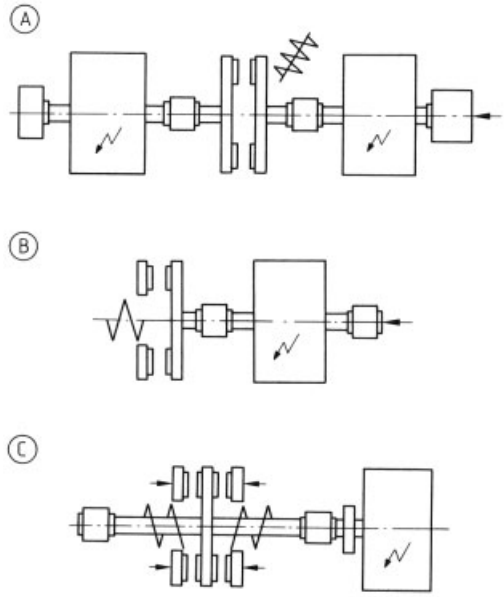

Figure 38. Refiner configurations
A) Double-disk refiner; B) Single-disk refiner; C) Twin refiner

The refiner disks consist of 6 – 12 segments and have a diameter of more than 1.5 m. The profile of the plates is designed such that the direction of disk rotation can be reversed, extending the plate life-time, which ranges between 1000 and 2000 h.

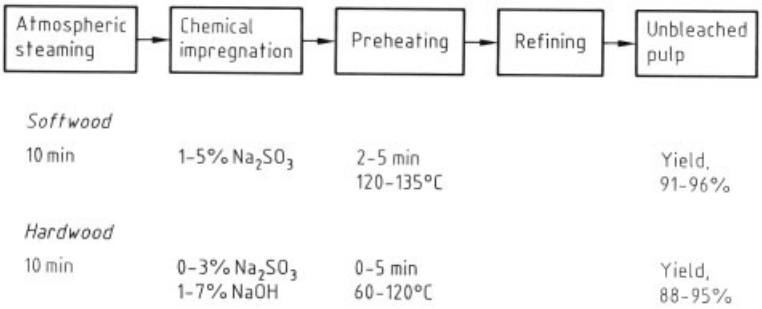

Figure 39. Schematic of the CTMP process

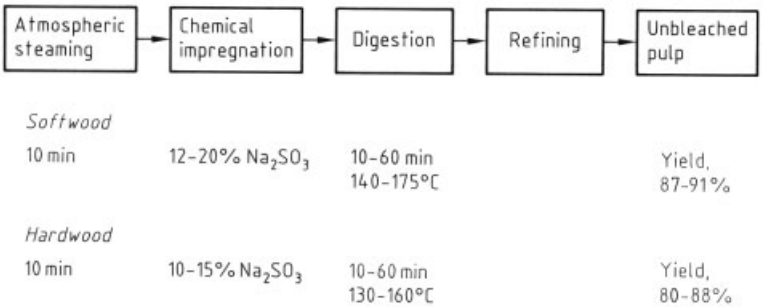

Figure 40. Schematic of the CMP process

The design of the refiner plates is also of great importance. This applies both to the optimum profile and to the alloy composition of the plate material [285]. The possibility of using elsewhere the heat generated in the process by means of heat exchangers has already been mentioned. In this way, 50 – 70 % of the motor energy can be saved [281], [282].

2.3.4. New Methods of Refiner Pulp Production

Although the addition of chemicals in the groundwood process did not prove worthwhile, their use in the refiner process has become increasingly important because wood chips can be impregnated very easily.

Sulfonation of the wood chips in combination with the TMP process has proved most effective. The properties of the CTMP can be varied to a great extent by changing the amount and the nature of the chemicals. CTMP has a wide field of application, ranging from tissue to LWC. A schematic of the process is given in Fig. 39 [286], [287].

Sulfonation results in permanent lignin softening which permits easy separation of the fibers. CTM pulps have a favorable dewaterability and are characterized by high strength, a large long fiber fraction, and a very low shive content [288].

A further development of the CTMP process is the CMP process. In this method, the process conditions are intensified, giving products comparable to semichemical pulps (Fig. 40) [286].

CTMP and CMP have the advantages of good strength properties and low shive content, but they also have the disadvantages of higher solvent losses and, therefore, environmental problems. Their tendency to yellowing is typical of all lignin-containing pulps [289]. For this reason, use of CTMP and CMP is limited to paper that is used for only a relatively short period of time.

2.4. Screening of Mechanical Pulps

2.4.1. Purpose of Screening

All mechanical pulps contain undesired components such as large, insufficiently pulped frag-

ments (i.e., coarse rejects) and the shives, which consist of many fiber bundles (Fig. 41).

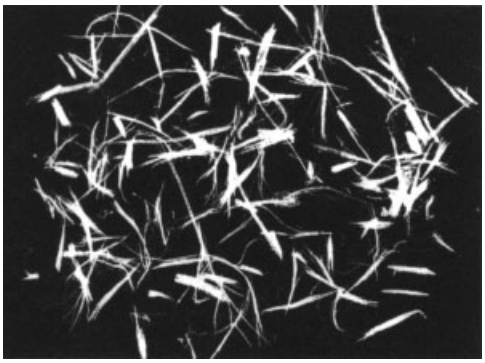

Figure 41. Fiber bundle or shives

The shives have to be removed because they reduce the strength and the printing quality of paper [251]. The screening out of the coarse rejects is easy to perform, but the removal of the shives requires a more elaborate technique. The shive content of unscreened mechanical pulp may be as high as 5 %, depending on the process used. The purpose of screening is to remove the shives to a large extent and, at the same time, to keep the simultaneous screening out of valuable long fibers as low as possible. The extent of the required shive removal depends on the type of paper to be produced. With increasing printing quality, priority is given to shive removal over the prevention of fiber losses [290].

For the evaluation of the action of a screen or a screening system, it is not sufficient to evaluate the accepts alone; the quality of the inflow and of the rejects also has to be taken into consideration. The higher the efficiency of a screen, the higher the relative shive retention at the lowest possible overflow [291], [292].

2.4.2. Mechanical Pulp Screening

Coarse rejects are removed by vibrating flat screens (bull screens) equipped with perforated plates (3 – 10 mm). Hammer mills or pulpers are used for the reduction of the coarse rejects, also called "sauerkraut". The bull screen accepts-mainly contain coarse shives and are usually added to the rejects of the fine screening system for further processing.

The screens used for the separation of shives usually have vertically arranged screen cylinders. The screening process is carried out at overpressure (Fig. 42). The size of the holes in the screen plates is chosen in such a manner that fibers and accepted stock pass through the screen, whereas shives above a certain size are retained. Depending on the type of paper required, screens with holes of 1.0 to 1.8 mm in diameter are employed, and the screening process is carried out at high dilutions (pulp consistencies below 1 %) to promote the screening action.

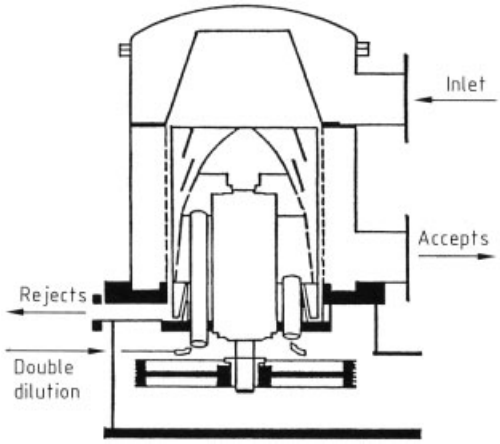

Figure 42. Hooper screen

Although older models are occasionally used, the new perforated plates equipped with slots (0.2 – 0.3 mm) instead of holes give an excellent screening effect [293], [294]. They relieve the cleaners, which are mainly used for the separation of cubelike particles, and are still an essential component of any multistage screening installation. The screening reject rate may be as high as 30 % of the inflow. A screening installation consists of various stages in order to reduce the strain on subsequent installations and to save most of the acceptable components contained in the rejects.

The rejects of the last stage, after thickening to ca. 20 % consistency, are processed in refiners and recycled to the screening process.

In the case of groundwood, the energy requirement for screening, for refining the screen rejects, and for subsequent thickening accounts

for ca. 25 % of the total energy. The value for refiner mechanical pulp is ca. 29 % [295].

2.5. Bleaching of Mechanical Pulps

Mechanical pulps need to be brightened to suit their intended use. Since higher process temperatures are always accompanied by a loss of brightness, the amount of wood pulp that requires bleaching is constantly increasing. Two bleaching processes or a combination of the two are used to achieve the desired final brightness [296–298].

Reductive bleaching uses sodium dithionite ($Na_2S_2O_4$) and can lead to an increase in brightness of up to 12 % ISO. Depending on the required final brightness, the amount of $Na_2S_2O_4$ used is 0.1 to 1.2 %, based on o.d. pulp. A suitable pH value is 5.6 to 6.5, and a temperature up to 70 °C accelerates the bleaching process. However, a further increase in temperature is detrimental. The optimum pulp consistency is 3 – 5 % and the bleaching time is ca. 1 h. The bleaching tower operates with an overflow and is charged from the bottom upstream, due to the detrimental effect of the presence of free oxygen on the process [296], [297].

Oxidative bleaching is carried out in the pH range 10 – 10.5. Sodium peroxide (Na_2O_2) or hydrogen peroxide (H_2O_2) is used in amounts of 0.4 – 4 % H_2O_2 (50 %), based on o.d. pulp. The bleaching time is 1 – 2 h at 50 – 60 °C. High pulp consistencies of 15 – 20 % are favorable and the maximum increase in brightness is 14 % ISO. However, an increase in brightness of 16 – 20 points can be achieved by combining the two bleaching processes in the order peroxide–dithionite (Fig. 43) [297], [299]. Both bleaching processes result in lower brightness in the presence of heavy metal ions. For this reason chelating agents (0.3 – 0.5 %) are added before bleaching [296], [297], [300], [301].

The technological properties of mechanical pulps are not changed appreciably by bleaching. However, solvent losses vary substantially, being distinctly lower with reductive bleaching. Finally, it should be mentioned that the tendency of mechanical pulps to yellowing is not changed by bleaching.

2.6. Comparison of Mechanical Pulps

Each mechanical pulping method results in very specific properties of the pulps produced. Furthermore, the pulp properties can be influenced to a certain extent by varying the pulping conditions or by additional treatments. A decision in favor of one or the other process is always a compromise because of the great variety of often contrasting properties.

It is generally valid that increasing process temperatures result in increased solvent losses and reduced brightnesses. Although the use of chemicals (CTMP, CMP) is of a certain advantage (for example, allowing the pulping of hardwoods) it causes a further increase in solvent losses.

Table 14 lists some technological characteristics of the various processes after thickening and before bleaching. These values are intended only to indicate the trends at about the same freeness level (excluding CMP and halfbleached sulfate pulp), and it should be mentioned that the values for the specific energy requirement, the strength, and fiber classification have a margin of error of ca. ± 10 %.

3. Wastepaper

Paper and board are indispensable resources for an information and distribution society. However, their production consumes large amounts of forest resources and energy. Furthermore, uncollected wastepaper which has been rapidly increasing is causing a large problem in the processing of urban waste materials. Although various proposals (e.g., specially designed incineration, biodegration, and efficient landfill) will be performed for solving these problems in the future, a possible method may be to promote the utilization of all kinds of wastepaper immediately.

In 1987, worldwide production of paper and board was 25.6×10^6 t, of which 6.73×10^6 t was collected as wastepaper (collection rate = 30.9 %). 6.69×10^6 t of wastepaper was utilized (utilization rates = 30.7 %) as raw material for paper and board production, accounting for about one-third of the total raw materials. Wastepaper has thus become an indispensable raw material for the paper manufactur-

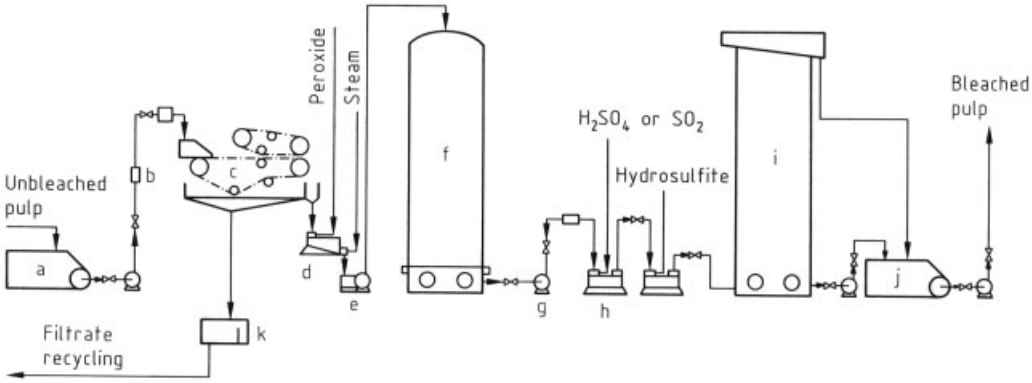

Figure 43. Schematic of a two-stage, continuous mechanical pulp bleaching process
a) Unbleached pulp chest; b) Pulp density regulator; c) Double wire press; d) High consistency mixer; e) High consistency pump; f) High consistency peroxide tower; g) Pump; h) Low consistency mixer; i) Low consistency hydrosulfite tower; j) Bleached pulp chest; k) Filtrate vessel

Table 14. Comparison of unbleached mechanical pulps (TAPPI standard)

	SGW	TGW	PGW	RMP	TMP	CTMP	CMP	Semibleached kraft pulp
Specific energy, kWh/t	1500	1530	1550	2100	2200	2100	1550	100
Process temperature, °C	75–80	93	115	90	120–140	120–140	150–160	
Yield, %	97–98	97–98	96–97	97–98	94–96	91–94	82–90	45
BOD, kg/t	7–15	9–16	10–17	ca. 15	18–25	30–40	50–85	200
Freeness (CSF), mL	100	100	100	100	110	110	300–400	350
Tensile index, Nm/g	30	34	37	36	42	47	54	65
Tear index, $mN \cdot m^2/g$	3.4	3.9	4.7	3.9	7.3	7.3	9.0	14.0
Quality value (QV)*	98	112	131	114	188	193	234	345
Long fiber content (R 30), %	21	24	30	29	40	40		
Fines content (D 100), %	45	41	37	40	33	32		
Brightness, % ISO	63	62	60	58	56	60	52	

* QV = tensile index + 20 x tear index.

ing industry. The collection rates and utilization rates of wastepaper in some countries [302] are listed in Table 15. Both the collection and utilization rate are very high in EC countries. Although the collection rate is high in Scandinavian countries, the utilization rate is extremely low, because most of the paper and board produced is exported. In Asian countries wastepaper recovery and utilization are very high.

3.1. Types of Wastepaper

Wastepaper is generally classified according to origin, quality, and usage. The classifications vary according to country. For example, they are classified according to the 1983 NARI standards [303] in North America, the European wastepaper standards (CEPAC) [304] in Europe, and classification guides for wastepaper [305] in Japan. In CEPAC the wastepaper is roughly classified into four groups, corresponding to ordinary, medium, high, and kraft qualities; these are further subclassified. However, these official classifications are rarely utilized as market nomenclature.

3.2. Origin, Collection, Handling, and Economic Aspects

Wastepaper contains contaminants such as inks, resins, and other substances. Since uniformity is important for pulp furnishes, the quantitative and qualitative variations of impurities should be as small as possible. Thus, the collecting methods and the segregation of the wastepaper at an early stage may be more important than the technical efforts to develop new processes for secondary fiber production.

Table 15. Wastepaper recovery and utilization in some countries in 1987

Country or region	Recovery rate, %	Utilization rate, %
EC	35	47
United Kingdom	30	55
Germany	41	44
Italy	23	41
Spain	41	58
Asia	37	47
Japan	50	51
Taiwan	40	87
China	21	25
Latin America	32	42
Mexico	40	67
North America	28	23
USA	29	26
Canada	20	11
Scandinavia	36	7
Sweden	40	10
Oceania	26	29
East Europe	26	25
Africa	17	27

The wastepaper collected from houses and markets etc., such as corrugated containers, newspapers, and magazines is usually called "collected wastepaper", while the shavings (cuttings), brokes (waste materials), and overissued papers generated in printing, paper-container manufacture, and book binding mills are termed "industrial wastepaper". Wastepaper occurring in supermarkets, department stores, and large office buildings are called "semi-industrial wastepaper" and are classified as industrial wastepaper.

Every country uses various collecting methods because the different climates influence wastepaper recycling. Most wastepapers are collected by dealers. However, in some countries collection boxes are provided without charge, and collection is carried out voluntarily by groups such as schools.

The collected wastepaper is brought to the collection yards, then weighed and sorted. They are compacted by baling machines (balers) which are installed in most collection yards. Small amounts of wastepaper are sometimes shipped in unbaled form. Industrial wastepapers from large generators are usually delivered to and processed in wastepaper yards integrated in the pulp mill. Detrimental substances (e.g., plastics; pressure sensitive and hotmelt adhesives; and carbonless copy, laminated, and thermosensitive papers) are removed before baling. The baled wastepaper is stored in the yards until shipment.

Fluctuating moisture content may cause economic problems when the wastepapers are accepted by the mills, because wastepaper is bought and sold on the basis of weight. The property of secondary fiber furnishes also varies because chemicals are added on the basis of dry weight. Various measuring tools are being experimented with in all countries, for example, core–drillers and " aquaboy" which work by measuring electric resistance.

There are seasonal deviations in the occurrence of wastepapers; therefore storage sometimes causes problems. Unsuitable storage conditions result in an increased short fiber content. The processing of deteriorated wastepaper requires large amounts of drainage aids, the reduction of paper machine speed, and leads to decreased sheet strength. In this case the mill operators should optimize the turnover of wastepapers, and establish a maximum storage period.

Because the secondary fibers have already passed through stock-preparation equipment when the original paper was made, they require comparatively little energy for processing. Therefore, the capital investment for the mills and the running costs are lower than those of kraft or sulfite pulping mills. This means that secondary fibers are usually cheaper than other pulps.

Recent incentives for increased wastepaper utilization are a result of the favorable price of wastepaper in comparison with the corresponding grades of market pulp. However, in some countries wastepaper recycling is promoted by government assistance programs and policies for forest preservation and environmental protection. This trend will be accelerated more and more in the future.

4. Nonwood Plant Fibers

4.1. Rags

The textile fibers of cotton linters, flax, hemp, Manila hemp (abaca), sisal, and jute are used in papermaking. The fibers of cotton linters are the original fibers on cotton seeds. These fibers are

processed as original fibers or as cotton waste from the spinning factory.

The term rags refers to old or new fabrics produced from textile fibers, e.g., waste cloth, mailbags, foot cloth, ropes, binding twine, etc. Blended fabrics that contain plastic fibers or plastic coatings cannot be used. The amount of rags and cotton linters used by the pulp and paper industry in the Federal Republic of Germany was 92 000 t/a in 1962 and 38 000 t/a in 1988 (data from the Verband deutscher Papierfabriken).

The reasons for the replacement of rag and nonwood plant fibers are:

1) Wood fibers are substantially cheaper
2) Long-fiber softwood pulps are now of sufficiently high quality to compete with rag fibers in many applications
3) The very low availability of rag and nonwood plant fibers compared with wood fibers

Range of Use of Nonwood Plant Fibers in the Paper Industry. Rag fibers are used for thin and lightweight paper which must meet exacting demands with respect to strength, toughness, opacity, aging resistance (printing paper), absorbing capacity, chemical purity (linters), and runability. The following papers contain nonwood plant fibers:

cigarette paper (flax, hemp, and sisal)
tea-bag paper (sisal, Manila hemp)
casing paper (Manila hemp)
thin carbon raw silk paper (flax, hemp)
brush matrix (abaca, sisal)
special filter papers (cotton linters)
banknotes and security papers (hemp, cotton linters)
paper for dust bags in vacuum cleaners (various rag fibers)
special drawing, watercolor, and handmade paper (cotton linters)
voluminous electrolyte paper (Manila, sisal)
special thin Bible paper (flax, hemp)

Depending on the rag-fiber content these papers are designated as:

1) All-rag paper (100 % rags)
2) Rag paper (at least 50 % rags)
3) Rag-containing paper (at least 10 % rags)

The fiber additives are wood pulps and, in some cases, synthetic fibers.

The individual fibers are broken free from the fibrous material mainly by alkaline or neutral sulfite digestion. Stronger alkaline pulping is necessary to eliminate the lignified parts of fibers. In the case of rags, however, mild digestion is sufficient. The half-stock is subsequently bleached. The fiber yield is 40 – 50 %. Fiber yields exceeding 70 % can be achieved, depending on the quality of the rags.

Prospects. Efforts are being made to develop kenaf-hemp (*Hibiscus cannabinus*) and *Hibiscus sabdariffa* (Malvaceae family), also known as Mesta, Bimli, or Deccan hemp, Siam jute or Thai kenaf, into an industrially usable fiber plant for the paper and pulp industry. These plants occur in tropical and subtropical regions. *Kenaf rosella* is more robust, grows on poor soil more slowly, and gives lower fiber yields than kenaf (*Hibiscus cannabinus*).

4.2. Other Nonwood Plant Fibers

Other nonwood plant fibers used in the paper and pulp industry are:

bagasse
bamboo
esparto
reeds
rice straw
wheat straw
ramie
papyrus (upper Sudan)

In 1987 the estimated worldwide production of pulp from the following nonwood plant fibers was:

straw	5.4×10^6 t
bagasse	2.1×10^6 t
bamboo	1.4×10^6 t
reeds	1.6×10^6 t

It is estimated that nonwood pulp products accounted for 8.1 % of total world pulp production in 1985 and showed a tendency to rise. The reasons for this are:

1) Saving of foreign exchange by the replacement of imported pulp by native fibrous raw materials

2) Creation of jobs in one's own country, both in agriculture as well as in the following industry
3) The consumption of wood-free fibers will increase on account of the progressive death of forests throughout the world and the growing demand for fiber raw materials
4) Efforts are being made to promote the cultivation of industrial plants to reduce agricultural surpluses
5) Recently developed, sulfur-free pulping methods could have a positive effect on the increased use of wood-free fibers in the pulp and paper industry

4.3. Synthetic Fibers

4.3.1. Survey

According to DIN 60 000 T 1, synthetic fibers belong to the group of man-made fibers that also includes fibers from natural polymers, e.g., viscose fibers. The term synthetic fibers covers a very wide spectrum of polymers, comparable to the range of plastics. The most important representatives are: PA 6 and 6.6, PETP, PAC, PP, PE, PVA, PTFE, PUE (elasthane), PVC, aramides, and PPS.

A comprehensive survey of the many different synthetic fibers available, their production, and properties is given in [306–309] (see also → Fibers, 4. Synthetic Organic).

The main application of synthetic fibers is textile production. Up to now, only a small quantity of synthetic fibers have been used in papermaking. The reasons for their use are as follows:

PA (6 & 6.6)	very high resistance against alternating bending, high strength, favorable dyeability
PETP	high strength (at a lower price compared to PA), favorable chemical stability, excellent electrical insulating capacity, very large selection of suppliers and types
PP, PE	very favorable chemical stability, low specific weight, relatively low melting point (ca. 160 °C or ca. 130 °C) which permits heat-sealing
PVA	used in the "unhardened" form due to the hot-water solubility, which can be used for pre-bonding of a web [310]

There are, in addition, some copolymer and mixed polymer fibers with low melting points that can be used as binding fibers [309], [311], [312].

Papers containing synthetic fibers are:

1) Document papers with especially high resistance to alternating bending stresses. These papers are used for identity cards and maps. Types of fibers preferred: PA and PETP.
2) Papers that are subjected to high mechanical stress, e.g., for envelopes. Type of fiber preferred: PETP.
3) Filter papers (dry and wet filtration) with high chemical stability. Type of fiber preferred: PETP, PP.
4) Papers with an especially high dimensional stability under varying environmental conditions, used for stencils and cartridges. Type of fiber preferred: PETP.
5) Papers that are suitable for clean rooms because of low particle emission. Type of fiber preferred: PETP and PE.
6) Wrapping paper in food technology: for tea bags and sausage skins. Type of fiber preferred: PETP, PP, PE [313], [314].
7) Electrical insulating papers. Type of fiber preferred: PETP and PE.
8) Overlay papers for composites in plastic technology. Type of fiber preferred: PETP.
9) Decorative papers with colored fibers as part of the pattern. Type of fiber preferred: PA and PETP.
10) Roller coating papers with PA as additive.

This list is incomplete and the designation of these products as paper is also disputed.

According to DIN 60 000 T 1, glass fibers do not belong to the group of synthetic fibers. However, these fibers are of great importance for wet-laid nonwovens (cf. DIN 61 210).

A close relationship exists in part between paper and wet-laid nonwovens (→ Nonwoven Fabrics). According to DIN 61 210, two criteria serve to distinguish them: the pulp content and the bulk density. Thus, products that contain pulp and textile fibers are regarded as paper if:

1) the pulp content is at least 50 % and
2) the bulk density is at least 0.4 g/cm^3.

If the pulp content is higher than 70 %, the product is generally considered to be paper and if it is lower than 50 %, the product is generally a non-

woven fabric. This definition has not yet been accepted internationally.

4.3.2. Marketing and Packages of Synthetic Fibers

In the case of several synthetic fibers, especially PETP and PA, many different types of each fiber are available. The properties of these various types can vary considerably even though they are made of the same polymer.

The main application of synthetic fibers is textile production. Accordingly, the package forms are adapted to textile process technology. They are characterized by:

1) crimped cut fibers, with fiber lengths ≥ca. 30 mm; or flat or textured multifilament yarns with unlimited length; and
2) spin finishes, which determine the properties of the fiber surface.

As a rule, special package forms are required for paper production: lengths of 2 – 30 mm, uncrimped fibers, and a hydrophilic spin finish for quick wetting of the fibers. If necessary, further shortening is carried out by beating (Hollander).

The low demand for fibers of this type (less than one-thousandth of the total amount produced) by manufacturers of paper and wet-laid nonwovens limits the choice of fiber types drastically because efficient production in the usual large-scale plants is possible only in amounts larger than ca. 10 t.

4.3.3. Processing Properties and Characteristics of Synthetic Fibers

The demands made on synthetic fibers by paper process technology do not play an important part in textile technology. They include:

Fiber Length [309], [313], [315]. Fibers that are too long result in "interspinning" (fibers form bundles and knots) during the suspending process. Therefore, the length is restricted and, at the same time, depends on the degree of fiber fineness. A favorable degree of fineness (calculated as fiber length (mm)/fiber fineness (dtex)) is 4 – 8, which corresponds to a length of 6 – 10 mm for thin fibers (ca. 1.5 dtex) and a length of up to 30 mm for coarse fibers (ca. 11 dtex). In view of this variation range, the varying polymer-dependent flexural rigidity does not have to be considered separately. Coarser fibers extend the range of processable fiber lengths.

Overlengths and Deformation of Cut Ends [309], [313], [315] arise through production defects. Overlengths are caused, in particular, by irregular fiber positioning and deformation is due to incorrectly adjusted or blunt blades. In the most unfavorable case, crushing of the cut ends can result in the welding together of neighboring fibers or entire fiber bundles, which cannot be opened during the suspension process.

Spin Finishes. Fibers free of spin finish are not produced. Spin finish – a multicomponent system of anionic or nonionic surfactants – cannot be quantitatively analyzed. It can only be described through its effects. The spin finish determines the wettability and suspendibility of the fibers [315], [316]. In fact, even polyolefin fibers (density <1.0) can be processed without problems if the formulation is optimal. If the spin finish is unsuitable, air bubbles become attached to the fiber surface in the vat and result in the flotation of the synthetic fibers in the vat or in the headbox. Spin finishes are partially water soluble and can therefore accumulate in closed process water cycles.

4.3.4. Influence of Synthetic-Fiber Properties on Product Properties

4.3.4.1. Polymer-Dependent (Uninfluenceable) Fiber Properties

Density. The density of common synthetic fibers extends over a wide range from 0.9 g/cm^3 for polyolefin fibers to 2.1 g/cm^3 for PTFE fibers. This has an effect on the number of fibers in the product. The lower the density, the more snythetic fibers are contained in the product, fiber size being the same. Thus, fiber distribution in the paper web becomes increasingly uniform.

Chemical and Physical Stability. In principle, fibers and the corresponding plastics behave in the same way. In borderline cases, however, the synthetic fiber is always less stable because

the lower cross-sectional area of the fiber and the larger surface give rise to a higher rate of damage.

Adhesiveness. In contrast to pulp, synthetic fibers are not capable, to any appreciable extent, of forming hydrogen bonds. Auxiliary agents, e.g., binders (latices), are required to bond synthetic fibers with each other or with other fibers [314]. Alternatively, thermal treatment can be carried out, i.e., sealing or pressing in the softened state (warm calendering). The adhesion of the synthetic fiber surface to a binder film depends on the polymer. Moderate adhesion is observed with PTFE, PP, and PE fibers and good adhesion with PA and PAC fibers. PETP fibers occupy the mid-position.

Fibrillation. Treatment of some types of PAC, especially wet-spun types, in a pulper or Hollander causes fibrillation. A similar effect also occurs in the case of highly oriented (highly stretched) PP fibers. Fibrillation increases the surface areas of the fibers and the fiber flexibility. These are two desirable effects which increase paper strength and reduce flexural rigidity.

If fibrillation is carried out during fiber production, "fibrids" are produced. These are very thin, short fibers which cohere in large numbers due to stemlike linkages and give rise to net-shaped formations. Polyolefin and aramide fibers are especially suitable for this purpose.

Biocompatibility. In particular, PETP fibers have been intensively studied because they are used as implants (substitute for blood vessels). This material can be employed without reservations. It is suitable for use in the hygiene and food sectors provided that the spin finish has been carefully removed. PTFE and PP fibers can also be used for this purpose if they are free of spin finish. A general statement of this type cannot be made for any of the other synthetic fibers.

4.3.4.2. Type-Dependent (Influenceable) Fiber Properties

Fiber Fineness. As a rule, the fineness of fibers is adjustable within wide limits: between ca. 1 and 20 dtex in standard types. Melt-blown fibers extend this range to about 0.05 dtex and coarse fibers to about 200 dtex. The fineness used depends on the type of paper, the most common range being 1.5 dtex to about 7 dtex.

Fiber Cross Section. Melt-spun fibers, such as PETP, PA, PP, and PE always have circular cross sections. However, other shapes can be produced by changing the shape of the nozzle: ring-shaped (i.e., hollow fibers), triangular, square, and star-shaped. These forms influence flexural rigidity and fiber gloss in particular. The circular cross section gives rise to the relatively lowest flexural rigidity and the most pronounced, sometimes "greasy" gloss. Multicornered or star-shaped cross sections stiffen the fibers and lower the gloss. For some synthetic fibers, e.g., PAC or PVA, the cross section depends on the spinning conditions. Consequently, a wide assortment of various noncircular cross sections can be formed, depending on the type and on the producer.

Strength. Strength can be adjusted within wide limits during fiber production by changing the spinning and aftertreatment conditions. As a rule, finer fibers have a higher strength per unit fineness than coarser fibers. The extent to which fiber strength influences paper strength depends primarily on the size of the fiber. Long, thin synthetic fibers increase paper strength to the greatest extent. The degree of fiber strength in itself is of minor importance.

Thermoshrinkage. Synthetic fibers shrink upon application of heat. The degree of shrinkage extends from $<2\,\%$ for thermofixed types to ca. 40 % for high-shrinking types. The degree of shrinkage in boiling water corresponds to the degree of hot-air shrinkage at about 160 °C. Creases are formed in the product web if uncontrolled high shrinkage takes place during the drying process.

Crimp. In textile applications, fiber crimp helps to increase the volume of yarns and webs and also functions as a processing aid. In the paper and wet-laid nonwoven sector, it is the volume increasing effect, produced by a sawtooth shaped fiber, which is of interest [317]. (Note: three-dimensional screw-shaped crimp usually leads to disturbing "interspinning"). Only the

careful adjustment of crimping geometry (number of crimp curves per unit length, curve height) and spin finish guarantees problem-free processing in paper technology.

Delustering. Melt spun fibers (PETP, PA, PE, and PP) are usually glossy. This interferes with many textile applications. Therefore, the fibers are delustered by the addition of titanium dioxide. The degree of dulling extends from "micromatt" (<0.1 % TiO_2) to "deep matt" (>2.5 % TiO_2). Glossy fibers are usually supplied if the papermaker does not have any special requests when purchasing fibers.

Optical Brightening. "Pure white" fiber types are produced by the use of optical brighteners during spinning ("raw white" is the trade name without brighteners). Fibers of this type fluoresce on exposure to UV light, an effect that can be exploited, e.g., in document papers or used for quality control.

Color. Apart from a few exceptions (aramides, polyimides, PTFE), synthetic fibers are white. However, they can be dyed almost any color during production by means of pigments (spin dyeing). The availability of dyed fibers is still very restricted. This limitation is due to the producer requirement that minimum amounts in the range of 10 t per color have to be purchased. Colored synthetic fibers produce patterns in decorative paper.

Flammability. If high demands are made on the nonflammability, it is necessary to use special types of fibers or flame retardants, which are also widely used in paper technology. The available selection of these fiber types, especially those based on PETP and aramide, is constantly increasing because of worldwide growing demands for industrial and garment textiles.

Electrostatic Properties. Synthetic fibers that are free of spin finish become highly charged by friction and discharge very slowly. In the case of textiles, this unfavorable electrostatic property is reduced to an uncritical level by using antistatic agents, which are an essential component of all spin finishes, or by adding conductive fibers. The same solutions would also be appropriate for papers and wet-laid nonwovens if, as a result of a high proportion of synthetic fibers, static charges are generated during use or if these papers are used in areas protected against explosion.

4.3.5. Special Fibers: Bonding Fibers [309], [310], [312], [313]

The extensive assortment of synthetic fibers includes special "bonding fibers". For instance: hot-water soluble PVA fibers and low melting copolymer or mixed polymer fibers.

PVA fibers dissolve at water temperatures of 60 °C – 80 °C. This effect is utilized in the drying of a wet paper or nonwoven web. The PVA "solution" is distributed in the web during the drying process and then solidifies as a result of further heating after the water has evaporated.

Low-melting copolymer and mixed polymer fibers include copolyamide and copolyester fibers and mixed polymer fibers containing 85 % poly(vinyl chloride) and 15 % poly(vinyl acetate). The melting or softening point can be adjusted in each case by altering the mixing ratio. The typical temperature lies in the range 120 – 200 °C for copolyamide and copolyester fibers, and at ca. 160 °C for the mixed copolymer fibers. Softening can already occur in water at 100 °C.

At least in the case of copolyamide fibers, the melting temperature in aqueous media is considerably lower, so that bonding can occur already during the drying of the wet webs due to the melt–water system formed.

PP and PE fibers are preferentially used as bonding fibers for dry-laid nonwovens. They are softened/melted under the influence of pressure and heat and then bond to other fibers such as PETP. Hot water or saturated steam are inadequate for the softening process. This fact distinguishes the PP and PE fibers from the special "bonding fibers".

5. Raw Materials for Stock Preparation

5.1. Fibrous Raw Materials

Mulberry fibers, bamboo bark, China grass, cotton, hemp, and old rags were the first plant fibers

that were used for papermaking by the Chinese before the birth of Christ. Later, rice straw was employed, and the Japanese also used mulberry bark, Mitsumata, and Gampi. The Arabs adopted the technology and used linen and hemp as raw materials. Rags also played an important role in the early days of papermaking in Europe. In fact, old rags remained the only raw material till well into the nineteenth century. However, it became increasingly scarce with the invention of printing by GUTENBERG (1440), the publications of the reformation, the introduction and development of the postal service (after 1500), and the invention of the paper machine (patented in 1799 by ROBERT). The search for other fibrous raw materials resulted in a breakthrough in 1844, when KELLER succeeded in producing groundwood mechanically. However, the disadvantages of groundwood soon became apparent, and in the late 1800s methods were developed for the chemical production of pulp suitable for papermaking from annual and perennial fibrous raw material. Today, fibrous wood-based materials (residues from forestry and wood processing) are used almost exclusively for the production of bulk papers (printing papers, packaging papers) in North and South America, Europe, and Japan. Other raw materials are annual plants, such as wheat or maize straw and bagasse.

Today, the composition of the raw material used for paper is influenced more than ever before by the cost of the individual components. Since the cost of the raw materials for papermaking frequently accounts for more than 50 % of the total costs, changes in composition can have a major effect on the total production costs. Quality permitting, it is necessary to economize on fibrous materials not only because they are much more expensive than, e.g., mineral fillers, but also for reasons of environmental protection and better utilization of existing plant raw materials. This is particularly important in view of the fact that in the future the demand for paper can be expected to increase due to rising living standards and the growing world population.

Specialty papers still require fibrous materials that are otherwise too expensive to be used. For instance, document paper, banknotes and identification papers contain not only synthetic fibers, but also cotton and jute pulp from rags.

5.1.1. Pulp

The term chemical pulp for paper refers to fibrous material, usually obtained from wood, that only has a low lignin content as a result of chemical digestion. Paper containing only chemical pulp as the fiber component is called wood-free paper. In other words, this paper is free or largely free of the woody substance lignin. Pulps represent the most valuable constituent of the fiber component and are responsible for the highest attainable strengths. Owing to the elimination of lignin and the partial removal of hemicelluloses, hydrogen bonds can form between the fibers in the papermaking process. These bonds are responsible for the strength of paper, which results predominantly from fiber-to-fiber bonds. Fiber matting is practically of no importance for the very short fibers found in modern pulps. Furthermore, the quality of paper depends on whether the pulp was produced from softwood or hardwood. Softwood pulps have longer fibers (4 – 6 mm) and therefore give higher strengths, whereas hardwood pulps (fiber length ca. 1 mm and less) produce a more uniform sheet. In addition, the fibers are shortened and fibrillated by the mechanical treatment during beating. The shortening increases the uniformity of the resulting paper but can also lead to a slight loss of strength. On the other hand, fibrillation roughens the surface of the fiber, and the projecting fibrils can form more bonding points. The tensile strength, in particular, can be increased by this process. The type of digestion also plays a very important role here. Sulfite pulps can be beaten more easily, i.e., the strength potential can be increased more quickly by beating. In the case of kraft pulps, the increase in strength is slower but higher strengths can be achieved than with sulfite pulps, although this requires a correspondingly higher energy expenditure. For unbeaten or slightly beaten pulps, there is an increased danger of flocculation occurring in the liquid stock prior to sheet formation. This results in poor formation and nonuniform paper. Since beating also increases the flexibility of the fibers, a higher packing density can be achieved during sheet formation. This increases the strength, but the paper becomes more transparent because of the presence of small amounts of enclosed air. Although this is required for some specialty papers, for the majority of writing and printing pa-

pers it is a disadvantage because writing or pictures on the one side of the paper would interfere with those on the other side.

5.1.2. Mechanical Pulp

Papers that contain pulp produced by mechanical or chemomechanical processes are referred to as wood-containing papers. In the production of mechanical pulp, fibers are separated from the wood as gently as possible by mechanical processing. To produce useable fibrous material, the lignin binding the wood fibers together must be softened. This process is promoted by heating (friction heat) and is achieved more efficiently the higher the moisture content of the wood. The finished stock contains fibers as well as fiber fragments. For this reason, papers produced from pulps of this type have lower strengths than wood-free papers. Another reason for the lower strength is the high lignin content, which blocks hydrogen bonding sites on the fibers and stiffens the fibers themselves. Consequently, in the course of papermaking, fewer bonds can form during dewatering. However, the resulting paper has a higher specific volume. As a result of the higher volume and the presence of amorphous lignin, paper made from groundwood has a good opacity, which in turn allows even papers of low basis weight to be printed on both sides.

The residual lignin in mechanical pulp results in a poorer and less stable brightness level. Mechanical pulps cannot be brightened by oxidative and reductive bleaching to the same extent as chemical pulps because, in contrast to the bleaching of chemical pulp, no delignification takes place. The aging, or yellowing, is accompanied by a decrease in strength by which the paper becomes highly brittle. Therefore, fibrous materials of this type are not suitable for the production of paper for archives.

5.1.3. Wastepaper

In the paper industry, it is a very old tradition to make new paper from wastepaper. Today, wastepaper is also regarded as fibrous raw material because many types of paper either contain a large proportion of wastepaper or are produced entirely from it. There are many different types of wastepaper, and these require exact classification and separation. This secondary fibrous material has various uses, depending on its history and degree of soiling. After being subjected to a dissolving process, pure grade, unprinted wastepaper can be used for the production of paper either directly or after a purification step. Mixed, unsorted paper, e.g., from household paper collections, can be used only in the production of paper whose quality is not reduced by the presence of many different fibers or accompanying waste paper components. These include, for instance, the middle layer of board or the components of corrugated cardboards. White papers such as newsprint can be made from wastepaper by using the deinking process. In this process, most of the printing ink is removed from printed old white paper (newspapers, magazines). Consequently, deinked stock is considerably whiter than stocks that have only been dissolved. Today, hygienic paper and writing paper are also made from wastepaper. It must be taken into account in the recycling of wastepaper that fibers of organic origin cannot be reused indefinitely because their strength is reduced by the stresses in production processes, such as dissolving, beating, sheet formation, drying and further processing (e.g., printing), and in use. For this reason, the addition of primary fibers (chemical or mechanical pulp) is necessary to maintain the strength of the paper. The stability of wastepaper-containing papers can also be limited because they can contain an uncontrolled, large proportion of mechanical pulp.

5.2. Stock Additives

5.2.1. Additives with Qualitative Effects

The demands made on the quality of many types of paper are very high because of the increased efficiency of modern processing machines. These requirements can no longer be met by natural raw materials alone. A large number of additives are available for altering and improving the properties of paper.

5.2.1.1. Fillers

Mineral fillers have long been used in papermaking. Previously, they were used to obtain a

more closed surface for writing papers and to improve the writing properties. Today, the main reason for the use of fillers in printing papers is to increase printability. They produce not only a higher smoothness and a higher gloss after calendering but also a higher brightness and opacity. In addition to improving quality and printability, mineral fillers are considerably cheaper than fibrous materials and, therefore, efforts are made to include as much filler as possible in the paper in order to reduce material costs. The limits are set by the properties of the sheet because the filler cannot form hydrogen bonds and so reduces the strength by blocking fiber-to-fiber bonds. This applies to both the tensile strength (breaking length, breaking load) and the surface strength (picking resistance), which is especially important in offset printing. If excessive amounts of filler are used, they can no longer be adequately fixed in the paper, resulting in dusting.

The demands made on fillers include the highest possible brightness level for the production of high-quality writing and printing paper, high refractive index so that the opacity is high enough even in thin papers to allow printing on both sides, chemical inertness so that no reactions occur between the filler and other paper components or dyes, and a defined particle size distribution to give a good surface finish. The particle size is also of importance for the production technology because, together with hardness and particle shape, particle size determines the abrasive wear of the paper machine wires. A survey of the properties of fillers is given in Table 16.

Kaolin, an aluminum silicate, is still the most frequently used filler. It is supplied as a slurry or dried in powder form. The hexagonal crystalline platelets of kaolin produce a high gloss in the finished paper after calendering. This gloss depends on the degree of delamination, i.e., the extent to which the platelet agglomerates are broken into individual platelets. The brightness of simple filler kaolins is not especially high. As a result of its chemical composition, kaolin can also be used as filler for papermaking under acid conditions. The retention of kaolin is better than that of granular fillers because of its plate structure.

Calcium carbonate is obtained by grinding limestone and has a granular structure. The greatest advantages here are the high brightness and the somewhat lower price compared to kaolin. This filler must be used in papermaking in the absence of alum because of its solubility in acid media. This still causes problems in the production of wood-containing papers. Calcium carbonate is used predominantly to produce papers with a matt surface. Precipitated grades are only employed for specialty papers, primarily because of its high cost.

Talc is hydrophobic and is therefore used when dealing with resin problems. Resin, which is problematic in paper production, is carried into the process with fresh fibers or on the sticky components of wastepaper. Talc particles are adsorbed on the resin and render these interfering substances harmless by covering them. As a result of its pronounced platelet structure, the use of talc as filler gives good retention and paper surfaces that have a high smoothness and gloss.

Titanium dioxide has the highest brightness and, because of its high refractive index, gives paper of the highest opacity. It is because of its excellent opacity that titanium dioxide, despite high cost and poor retention, is still indispensible as a filler especially in the production of thin printing papers. However, due to its high cost and for reasons of environmental protection, titanium dioxide is used in combination with extenders. These extenders are other minerals which are capable of replacing titanium dioxide to a limited extent while producing similar results.

Gypsum is currently not used on a large scale because of its high solubility in water and the resulting side reactions that occur in the circulation system of the paper machine. The high availability of gypsum, especially from flue-gas desulfurization, has stimulated intensive research in ways of using gypsum without these side effects.

Various other mineral fillers are used only for specialty papers because they are expensive or have special properties. *Barite* (rhombic barium sulfate) has high whiteness and is of importance for high-quality writing and printing papers, and for photographic papers. *Magnesite* (magnesium carbonate) is of some importance in the production of cigarette papers. *Diatomaceous*

Table 16. Properties of fillers

Filler	Chemical composition	Crystal form	Brightness, % ISO	Refractive index	Density, g/cm^3	Specific surface, m^2/g	Mohs hardness	Surface energy, 10^{-7} J/cm^2
Kaolin	$Al_4(OH)_8$ $[Si_4O_{10}]$	hexagonal platelets	70 – 90	1.56	2.8	4 – 7	1 – 2.5	508 – 668
Calcium carbonate	$CaCO_3$	orthorhombic	82 – 92	1.58	2.7	3 – 4	3	78
Ppted.∗ calcium carbonate	$CaCO_3$		95 – 98	1.56	2.7	3 – 10	3	
Talc	$Mg_3(OH)_2$ $[Si_4O_{10}]$	monoclinic lamellas	70 – 88	1.57	2.7	5 – 20	1	68
Titanium dioxide	TiO_2	tetragonal	97 – 99	2.60	4.2	8 – 11	6 – 6.5	640
Gypsum	$CaSO_4$	orthorhombic	92 – 98	1.58	2.8	0.7 – 2	2	39
Barite	$BaSO_4$	rhombic	93 – 95	1.64	4.5	0.7 – 2	3 – 3.5	
Magnesite	$MgCO_3$				2.9 – 3.1		4 – 4.5	
Diatomaceous earth	SiO_2		60 – 90	1.33	2.3			
Synthetic ppted.∗ silicates	mostly Al or Ca silicates	mostly amorphous	94 – 96	1.54	2.1	30 – 150		
Aluminum hydroxide	$Al(OH)_3$	platelets pseudohexagonal	96 – 98	1.58	2.4	5 – 9	2.5	

∗ Ppted. = precipitated.

earth, an amorphous silicium dioxide with a very low specific weight, can be used for very highly voluminous papers. *Synthetic precipitated silicates* and *aluminum hydroxide* also give a high specific paper volume and especially good printing properties.

Synthetic organic fillers can give a printability comparable to that attained with mineral fillers. In particular, the opacity is increased in combination with high volume. However, these fillers are seldom used partly because of problems with the uniform distribution in the paper and, above all, because of the high cost.

5.2.1.2. Strength Additives

It is possible to distinguish between three basic types of paper strength. The initial wet strength is the strength of the paper in the wet state before drying. It greatly influences the operational characteristics of the paper machine. The dry strength which the paper possesses in the finished state is responsible for a trouble-free running in further processing steps, e.g., in a printing press or packaging machine. Finally, the wet strength describes the strength of the paper after it has been wet.

Substances that improve dry strength are of the greatest importance in papermaking because they allow the use of cheaper raw materials, such as fillers. Dry-strength additives form additional hydrogen bonds, which lead to stronger fiber-to-fiber bonding. In this way, it is possible to compensate for the absence of hydrogen bonds due to the presence of mechanical pulps or fillers. Starches and their derivatives, plant gums, cellulose derivatives, and synthetic polymers are used for this purpose. These substances are usually added to the pulp slurry and, therefore, their retention in the stock often poses a problem. In general, the use of these additives is determined by economic considerations, i.e., whether it is more economical to use expensive raw materials or even more expensive binders, which, however, need only be used in small amounts to produce the desired strengths.

The wet strength of paper is essential for some of its applications. For instance, some packaging papers must be resistant to rainwater, while it should be possible to remove labels from bottles without turning the wash solution into a fiber slurry. Wet-strength additives protect the natural hydrogen bonds from attack by moisture and form additional water-resistant bonds. Wet-strength additives include poly(ethylene imines), polyamides, and urea – or melamine – formaldehyde condensates.

5.2.1.3. Sizing Agents

Paper usually has a high absorbing capacity for liquids because of its hygroscopic fibers and the many fine pores of the web. However, this has

a negative effect on the writing properties and printability of the paper. The absorptivity can be reduced by the addition of sizing agents to the pulp or by their application to the surface of the paper. Sizing makes the paper partially hydrophobic and prevents the spreading of ink or color, but not the receptivity for these fluids. In the production of hand-made paper, individual sheets were coated with starch or animal glue which covered the fine pores in the paper, resulting in a certain degree of sizing. This process corresponds to modern surface sizing, which is used either independently or to supplement pulp sizing, and is usually carried out with a size press in the paper machine. The most widely used sizing agent in pulp sizing is still rosin size, which is a partially saponified colloid or dispersion of colophonium. Sizing is achieved by distributing the size as finely as possible in the pulp slurry and subsequently precipitating it on the fibers as an aluminum resinate by adding aluminum sulfate. However, this process served as the basis for the paper disintegration (see Chap. 12) feared by libraries and archives because the acid residues remaining in the paper reduce the elasticity of the fibers and make the paper brittle and unusable. Today, printing paper is usually produced in neutral to slightly alkaline systems without the use of aluminum sulfate. This requires either synthetic sizing agents, primarily dimeric alkylketenes, or new precipitants for rosin size, which are in part still being developed.

Furthermore, aluminum sulfate is also an excellent precipitant for interfering substances and is therefore still very difficult to dispense with when producing wood-containing printing paper. The new precipitants such as polyDMDAAC [poly(dimethyl diallyl ammonium chloride)] must also take over this role.

5.2.1.4. Dyes

The coloring of paper can be achieved by pulp dyeing, dip dyeing, or surface dyeing (see Chap. 10). A large number of physical and chemical forces are involved in each of these dyeing processes. The most important processes are:

1) Penetration of the dye molecules into the capillaries and cavities of the pulp fibers
2) Adsorption of the dye molecules onto the inner and outer fiber surface
3) Formation of a sparingly soluble complex in the presence of salts or fixing agents

The natural inorganic pigments possess a low dyeing power, but are very lightfast. The synthetic inorganic dyes have improved dyeing power. Natural organic dyes are no longer of importance. The synthetic organic dyes make up the most important group; they include basic, acid, and direct dyes, as well as pigments.

Basic dyes are salts of organic bases which cannot bond with pure cellulose. For this reason, they are usually used only for wood-containing papers because they form sparingly soluble colored lakes with the acidic groups of lignin.

Acid dyes are usually the alkali-metal salts of dye sulfonic acids. They have no affinity for plant fibers and must be precipitated as colored lakes in the paper stock. This is achieved with aluminum sulfate, a substance used frequently in the paper industry; however, the use of rosin size gives better binding.

Direct dyes are usually the sodium salts of azo dyes containing sulfo groups. They are absorbed by bleached pulp fibers without the addition of a fixing agent. However, they are not very suitable for the dyeing of wood-containing stocks because the incrusting components (binding substances) of wood prevent uniform dyeing. The lightfastness is usually good, but the colors attainable are generally not as brilliant as those produced by acid and basic dyes.

In pigments the colorant is present in a water insoluble, finely dispersed form. The binding of the pigment to the fibers is improved by rosin sizing and the use of aluminum sulfate. Fixing agents must be employed during papermaking in neutral or slightly alkaline systems. The lightfastness of pigments is excellent and mottling usually does not occur. Since these pigments are expensive, they are used only for specialty light-fast papers. The naturally occurring coloring pigments such as ocher, terra sienna, or umber are of little significance at present. However, the synthetic inorganic (iron oxide, cadmium, chromium oxide), the organic (azo and polycyclic), and the metal complex pigments are very important.

In combination dyeing, it should be taken into account that various dyes, e.g., acid and basic or direct and basic, must be dissolved separately

because they would otherwise precipitate each other during the dissolving process.

Dyeing is also of some importance for white papers, which are treated with small amounts of a blue dye to achieve a color shift and remove the yellowish tinge. This process is called shading.

5.2.1.5. Optical Brighteners

Optical brighteners are substances which transform ultraviolet light into visible blue light. Paper is treated with optical brighteners to achieve a whiter visual appearance. Although this effect can also be obtained with shading dyes, optical brighteners increase the luminosity as well. These substances are mainly derivatives of diaminostilbenedisulfonic acid and can form hydrogen bonds, allowing them to absorb onto cellulose fibers in the same way as direct dyes.

Optical brighteners act only when the light incident on the paper has an UV component, i.e., they do not act in artificial light. They lose their effect with increasing application of printing colors. Furthermore, they are not suitable for long-life papers because they are not resistant to sunlight. For this reason and because of pollution problems, the use of optical brighteners in paper is still controversial.

5.2.1.6. Chemicals for Specialty Papers

A large number of specialty papers exist for various applications. However, only a few specialty papers and the additives required in their production will be mentioned here.

Fire-resistant papers are used for wallpaper base materials, Chinese lanterns, and paper streamers. These papers are produced by the addition of flame retardants to the pulp or to the coating. The flame retardants either release incombustible gases on heating which prevent the entry of atmospheric oxygen, or produce a nonflammable melt when heated that surrounds the paper. Examples include calcium chloride, magnesium chloride, diammonium ethyl phosphate, and mixtures of zinc borates, antimony oxides, and organic halide salts [324] as well as inorganic bromides and oxybromides.

Anticorrosion paper is impregnated or coated with chemicals that inhibit corrosion, e.g., sodium nitrite or sodium benzoate. The paper is coated by deposition of the chemicals from the vapor phase. These papers prevent the rusting of iron parts and the tarnishing of silver, aluminum, and copper.

5.2.2. Processing Aids

5.2.2.1. Retention Aids

The term retention refers to the holding back of the papermaking stock on the wire during dewatering. The fibers are retained on the wire fractions and fillers better than the fines, which may be washed through the wire and even through the fiber mat formed. Retention aids increase adsorption of fine particles onto the fibers so that they are retained with the fibers. This adsorption must be able to resist the high shear forces which arise in modern high-speed paper machines. Both cationic and anionic retention aids are used. Although nonionic products are rarely employed, they also produce good effects if their molecular length is sufficiently large. The effect of retention aids is not only dependent on the charge, but is also closely related to the molecular length. However, the particularly effective high molecular products tend to flocculate. The fibers themselves form larger flocks by adsorption resulting in a nonuniform paper. Such a cloudy formation of the paper sheet decreases both the strengths and the printability.

The retention aids usually employed are synthetic polymers, such as poly(ethylene imine) or polyacrylamide, cationic starches, and carboxymethylcelluloses.

5.2.2.2. Defoamers

Foam is a dispersion of air in water and has physical and chemical causes. The violent motion of water during the drainage procedure and in the entire cycle, especially at high paper machine speeds, produces a very stable foam in the presence of surface active agents and colloidal surface active film formers, such as starch and hemicelluloses. The most effective way to prevent and control foam is to determine the cause in each case so that the appropriate defoamer can be employed. Defoamers include mixtures

of higher alcohols, salts of fatty acids, and water-emulsible phosphate esters.

5.2.2.3. Biocides

The large number of organic substances (e.g., starch), which are not only employed for various purposes, but can also originate from the use of coated wastepaper, serve as nutrients for microorganisms, and typical paper-machine temperatures are usually ideal for the growth of these organisms. The nutrients that are important for bacterial growth become enriched due to their low retention in the paper web and the closed paper-machine cycles. To control slime formation, an exact examination of the raw materials and the strains of microorganisms introduced into the cycle is required in order to find a solution which is both economically and environmentally acceptable. Slime control agents include bromine, sulfur, nitrogen, or chlorine compounds, and acetic acid derivatives. Microorganisms can develop resistance to particular biocides, and it is therefore advisable to alternate between different products. With regard to biological wastewater purification, it is important that completely degradable products are used.

5.2.2.4. Dispersing Agents

Dispersing agents are required for both fibers and fillers. They increase the toxicity of slime control agents and, due to their (usually) anionic charge, they disperse resins and other sticky impurities originating from the pulp or wastepaper.

In the dispersion of fillers and coating pigments, dispersing agents prevent secondary agglomeration which would lead to disturbances in the paper sheet. Dispersing agents are also necessary for controlling the viscosity of filler and coating pigment suspensions because without these agents, the solids content of the suspensions required for economic and technological reasons would not be attainable. Dispersants are either added during redissolving after the drying process, where they increase the effect of shear forces of the dispersing machines, or they are added during wet grinding of the filler material. Dispersing agents are used in the transportation and storage of fillers and coating pigments to maintain them in a pumpable slurry form.

5.2.2.5. Complexing Agents

The presence of heavy metals negatively influences many production processes and paper properties. In particular, they interfere with the bleaching of wood pulp by reducing the effect of bleaching chemicals. Complexing agents that contain amino and carboxyl groups mask metal ions effectively. Chemical compounds of this type are nitrilotriacetic acid (NTA), ethylenediaminetetraacetic acid (EDTA), diethylenetriaminepentaacetic acid (DTPA), and hydroxyethylethylenediaminetriacetic acid (HEEDTA), which mask metal ions by chelation. Other complexing agents include the soluble salts of oxalic acid, citric acid, tartaric acid, gluconic acid, amines, and ammonia. From the point of view of environmental protection, the use of these chemicals is controversial because they are only poorly biodegradable.

5.2.2.6. Precipitating and Fixing Agents

Precipitating and fixing agents are used primarily for the precipitation of soluble substances onto the fibers. Acid and direct dyes require cationic fixing agents, and basic dyes require anionic agents. Fixing agents form sparingly soluble complexes with the dyes, and coarse flocculates when used in too high a concentration. Therefore a compromise must be made between the need to minimize the turbidity of the coloring and the requirement to maintain the wastewater as clear as possible. Increasing the amount of fixing agent leads to increased dye fixation and clearer wastewater, but also lowers the brilliance of the coloring. Fixing agents are used for tissue coloring to prevent bleeding.

Aluminum sulfate is the most commonly used precipitating agent. Other agents include condensation products of formaldehyde, urea, dicyandiamide, melamine, and condensation products of aromatic sulfonic acids with formaldehyde.

5.2.2.7. Drainage Aids

In the drainage of paper, board, and cardboard made from materials that are difficult to dewater (e.g., highly beaten pulps or wastepaper) the

use of drainage aids is often necessary. Drainage aids are usually ionogenic polymers similar to those used as retention aids. Drainage aids function by binding the fines to the fibers. This prevents their forming flocks which would block the pores between the fibers during formation of the paper web, thus hindering drainage. In order that this process can function smoothly under the high shear forces in modern paper machines, the order of addition of the raw materials and additives must be optimized. The position at which the drainage aids are added to the stock flow is equally important, since this determines the magnitude of the stresses to which the bonds formed by the drainage aids are subjected at various points in the paper machine.

5.2.2.8. Additives for Wastepaper Processing and Deinking

The removal of ink from wastepaper is determined by the type of ink binder and chemicals used. Two separating processes are employed: flotation and washing. The most widely used method in Europe is flotation. The most important chemical is sodium hydroxide, which is used to adjust the pH to 10 – 11. Saponifiable binders in the ink are saponified. In addition, the fibers swell substantially in this environment, making it easier to detach the dye particles from the fibers. However, sodium hydroxide solution also causes a strong yellowing of wood-containing stocks. In order to prevent this, hydrogen peroxide is used because, apart from its saponifying effect, it also acts as a bleach. In addition, water glass is added to stabilize the hydrogen peroxide. Since hydrogen peroxide is more effective at higher concentrations, the dissolving process is carried out at stock consistencies of 12 – 15 % (thick stock dissolver or dissolving drums). Soaps and fatty acids are used as dirt collectors and flotation agents. They form calcium soaps with hard water or the calcium carbonate from the coating or filler of the wastepaper. Calcium chloride must be added if the water is not sufficiently hard. For higher quality papers, further bleaching with sodium dithionite and/or hydrogen peroxide may follow. Coated papers can be deinked more easily because the inks are fixed only with the coating. The wastewater situation can be controlled very well; however, the COD value increases with increasing pH.

The wash deinking process is used in North America. After saponification, the pigments are removed from the stock by washing. Since large amounts of wastewater are produced and the solids losses are high, it is not surprising that the flotation deinking process is also gaining a foothold in North America.

Only pure-grade papers can be used in the deinking process. Newspapers and magazines are used in the largest amounts. Because of the printing inks still present in the finished stock, deinked stocks can be employed only for newsprint (up to approx. 60 %), hygienic papers and improved recycling writing paper (100 %), and in board production.

6. Stock Preparation

Stock preparation denotes the operations that must be undertaken to convert raw stock into finished stock (furnish) for the paper machine. The raw stocks used are the various types of chemical pulp, mechanical pulp, and wastepaper. Raw stock is available in the form of bales, loose material, or, in the case of integrated mills, as suspensions. The finished stock is a suspension of a defined quality as far as the mixture and condition of the fibers, additives, and impurities are concerned. This quality essentially determines the properties of the paper produced [325].

A stock preparation system consists of several process steps that are adapted to one another. These systems differ considerably depending on the raw stock used and on the quality of furnish required. Thus, refining is of central importance in the preparation of virgin pulp, whereas the separation and neutralization of interfering components are the main purposes of wastepaper preparation [326].

6.1. Process Steps

The important steps in stock preparation, their location in the course of the process, and the corresponding machines used are presented in Figure 44. It can be seen that some process steps are required almost throughout the entire process, whereas others are used only at the beginning

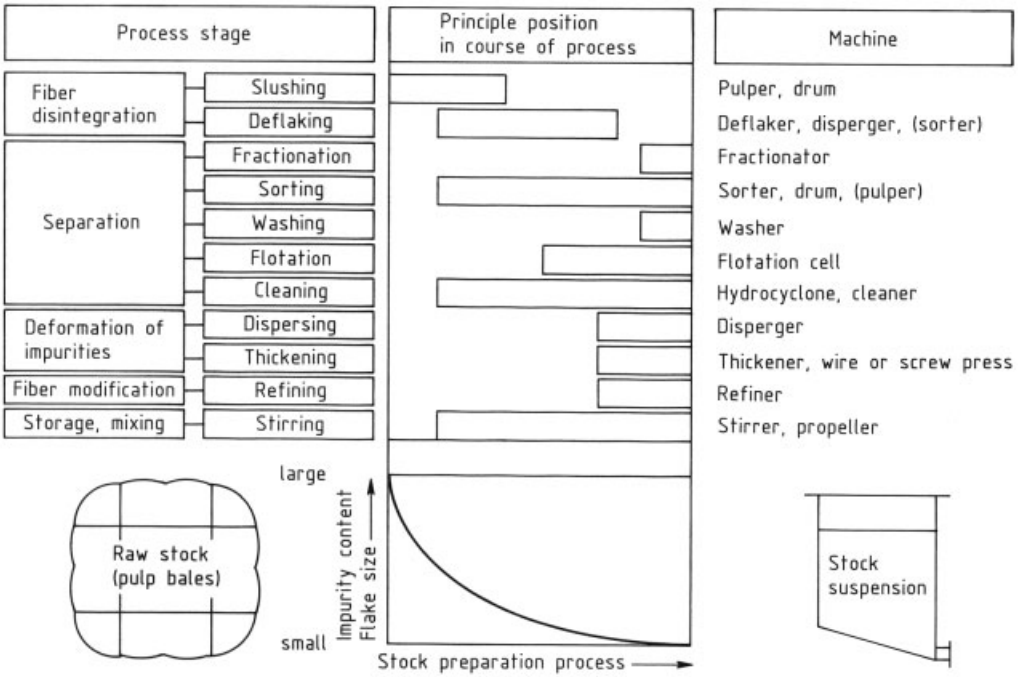

Figure 44. Process stages in stock preparation and the corresponding machines (courtesy of Sulzer Escher Wyss)

and at the end of the process [327]. It is possible to use different machines for a particular process step. In addition, it should be taken into account that most of the machines not only have one effect, but can also exert side effects that may or may not be desirable.

6.1.1. Fiber Disintegration

The fiber networks in raw stock should be broken up as quickly as possible with a minimum of mechanical work and chemicals so as not to alter the fiber properties. In this process, the matrix strength, which results primarily from hydrogen bonds, is reduced greatly by wetting with water. In the case of wet-strength wastepaper, this process is promoted by elevated process temperatures (e.g., >70 °C) and by the addition of chemicals (acidic or alkaline). The fiber matrix is broken up by the application of tensile stresses. Accelerative forces, shear forces, and holding forces due to mechanical clinging are decisive [328].

Slushing. The first rough disintegration to give a pumpable suspension is carried out in a pulper or dissolving drum. The pulper (Fig. 45) usually consists of a vessel with a rotor and a perforated plate that is kept free with the help of clearing vanes. The axis of the rotor is arranged vertically, horizontally, or inclined. The fiber consistency in the pulper is 3 – 15 % [329], and the peripheral velocity of the rotors is correspondingly 12 – 24 m/s.

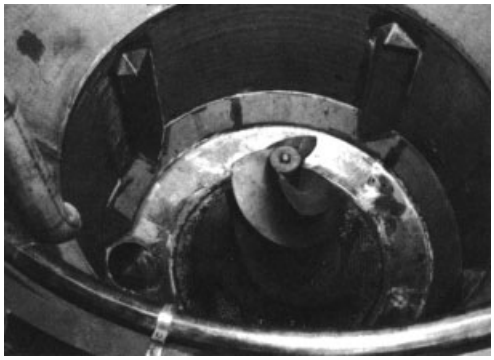

Figure 45. View of an empty pulper drum (courtesy of Sulzer Escher Wyss)

Depending on the raw stock and the type of pulper, the specific energy consumption for

slushing is ca. 10 – 40 kW · h/t and higher for wet-strength papers.

In many cases, the first separation of impurities from the fibers takes place in the pulper. This can occur by means of:

1) A ragger, which continuously removes wires and spinnable parts from the pulper tank;
2) Periodic removal of the residues retained by the pulper sieve. For this purpose, operation of the pulper is interrupted and the residues are washed out; and/or
3) Withdrawal of a partial stream from the pulper which is fed forward into the storage chest after coarse cleaning by a drum screen or coarse pressure screen. Other, older systems return this cleaned partial stream diluted back into the pulper.

High-consistency pulpers are designed increasingly without a perforated plate so that the stock is discharged from the pulper and then coarse screened. Drums are used only for dissolving wastepapers that have very low wet strengths (see Section 6.3).

Formerly, slushing, deflaking, and refining were performed in a single stage in a hollander. However, hollanders do not meet modern production and capacity requirements and are therefore no longer used [330].

Deflaking. The flakes (fiber bundles and pieces of paper) that pass through the pulper screen are separated into individual fibers in a deflaker. A disperging stage may be required in the case of wastepaper with high wet strength. It is more economical to carry out the process of fine disintegration outside the pulper in specially designed machines than in the pulper itself.

The deflaker shown in Figure 46 has a rotor and a stator disk, each of which bears three interlocking slotted rings. As the suspension passes through the slots, accelerative forces, shear forces, and holding forces due to the clinging of the edges break the flakes into individual fibers.

The deflaker operates at stock consistencies of 3 – 6 %. The peripheral velocity of the disks is up to 30 m/s. The specific energy consumption is 20 – 60 kW · h/t for easily deflakable stocks. Stocks that are more difficult to deflake must be subjected to two or more runs through the deflaker [331].

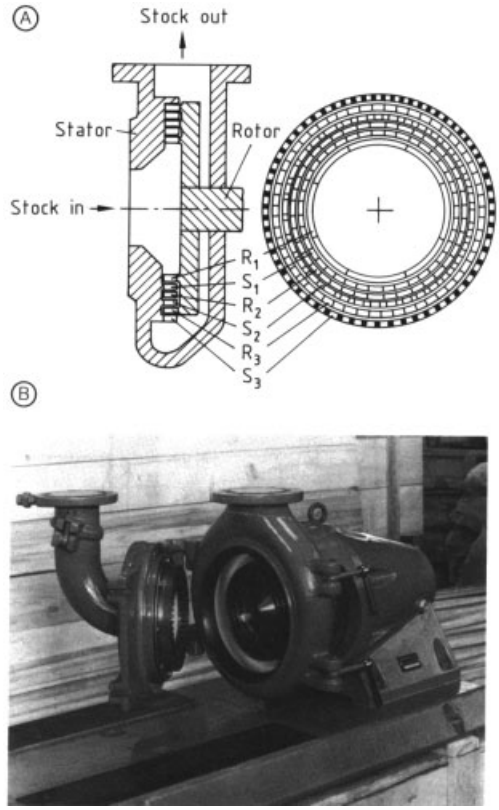

Figure 46. Schematic (A) and photo (B) of a deflaker (courtesy of Sulzer Escher Wyss)

Flakes of high-wet-strength materials can usually be broken down only with a disperger (see Section 6.1.3).

6.1.2. Separation

Separation refers to the processes that are used to separate substances of the same type (fractionation of fibers) or foreign substances (screening, washing, flotation, and cleaning) from fibers in a suspension.

Fractionation. By means of screening, the fibers present in the suspension are divided according to length into a fraction enriched in short fibers and a fraction enriched in long fibers and flakes [332]. These partial streams are then subjected to further treatment separately (deflaking, refining, and disperging) and are used, e.g., for the individual layers of multilayer products such

as board. The fractionating machine usually resembles a pressure screen, but the sieve holes are smaller than those of a pressure screen. Until now, fractionating machines based on other concepts have not been used in the paper industry.

Screening. The objective of screening is the removal of interfering substances from the fiber suspension that differ from the fibers in shape and deformability. The suspension is passed through a screen with apertures in the form of slots or round holes, and the impurities to be separated are rejected by the screen. The screen can be flat or cylindrical and the suspension can be pressurized or pressureless. Clearing devices produce pressure pulsations and thus prevent blocking of the screen. They also aid the transport of the impurities to the overflow.

Rough Screening. The screen apertures of pulpers, secondary pulpers, and drum screens used for rough screening are about 4 – 25 mm, i.e., they are usually capable of only rough screening.

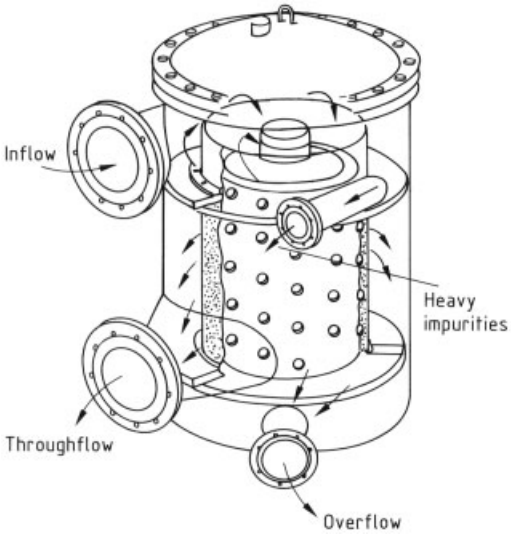

Figure 47. Schematic of a pressure screen with a cylindrical screen basket (Omniscreen) (courtesy of Sulzer Escher Wyss)

Fine Screening. For fine screening with a high screening quality, screens with aperture diameters in the range of ca. 0.8 – 2.4 mm and slot openings in the range of 0.2 – 0.8 mm are used. The screening efficiency depends on the geometry of the screen and the rotor, the direction of flow of the suspension towards the surface of the screen, the volumetric ratio of the overflow to the inflow, and on the specific throughput [333].

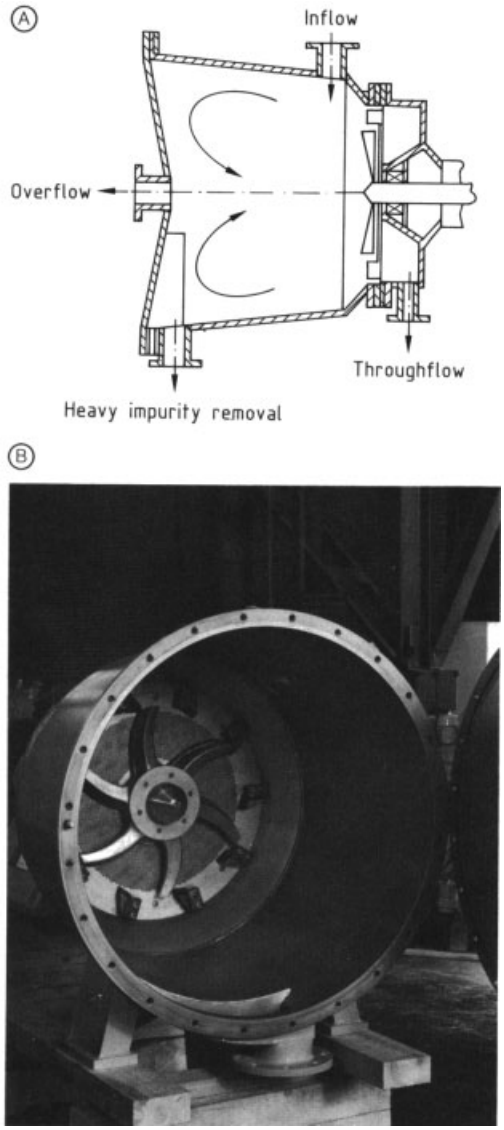

Figure 48. A) Schematic of a flat pressure screen; B) View of flat pressure screen after opening (courtesy of Sulzer Escher Wyss)

A pressure screen with a cylindrical sieve is shown in Figure 47. These screens operate with stock consistencies up to ca. 4 %, and the specific

energy consumption is 5 – 20 kW · h/t. Figure 48 shows a flat screen [334], [335]. Flat pressure screens require stock consistencies of 3 – 4.5 % and have a specific energy consumption of 20 – 40 kW · h/t. Pressure screens in the approach flow system operate with stock consistencies of ca. 1 %.

In the case of pressure screens, the overflow is highly laden with interfering substances and also contains a large number of fibers. These fibers should be recovered in the tail screen without impairing the total screening result. The vibrating screens used in the past as tail screens (cf. Fig. 21) are suitable only for the separation of larger impurities, while the smaller impurities are returned to the cycle in front of the first-stage screening device. Hence, the feeding back of this throughflow with all these impurities leads to a reduction of quality.

A highly effective tail screen is shown in Figure 49 [336]. Similar to the vibrating screen, it has a pressureless mode of operation, and rotors fitted with vanes are used to transport the stock and keep the screen clear. This sorter requires a stock consistency of 1 – 4 %, and it has a specific energy requirement of 20 – 40 kW · h/t. Screens with aperture diameters in the range of 1.0 – 4.0 mm or slot openings of 0.6 – 1 mm are used. The rejects accumulate in a form that poses no disposal problems. As a result of the good screening effect, the throughflow contains only a small amount of impurities and can therefore be fed forward into the mainstream production.

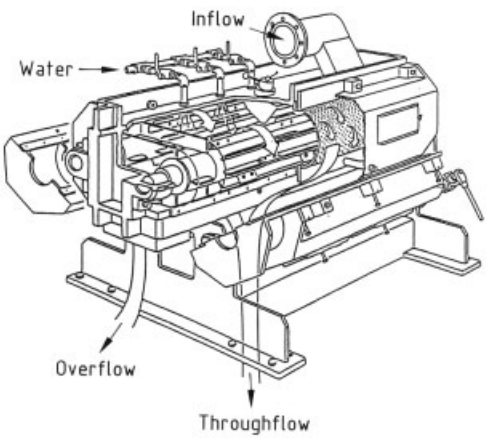

Figure 49. Schematic of a tail screen (reject sorter) (courtesy of Sulzer Escher Wyss)

Washing [337–340] is, in principle, also a screening and fractionating process, in which particles smaller than 30 – 40 µm, such as ash, printing inks, and inactive fines, are removed from the suspension. In the washing stage, the suspension is dewatered, and the solids that pass through the screen represent the washing effect. Upright inclined screens, curved screens, round screens, inclined screws, or horizontal presses are used for washing, but they all produce a limited washing effect. An effective washer is represented by the machine shown in Figure 50, which functions in a manner similar to the web-forming part of a paper machine. The suspension emerging from the headbox with a consistency of 0.7 – 1.4 % is dewatered to a stock consistency of 5 – 12 % by a wire running over a central roll. The thickness of the resulting fiber layer, i.e., the dry basis weight, essentially determines the washing effect. The machine is usually operated at a speed of 400 – 1000 m/min and 30 – 40 g/m^2 dry basis weight. The specific energy consumption is 5 – 10 kW · h/t.

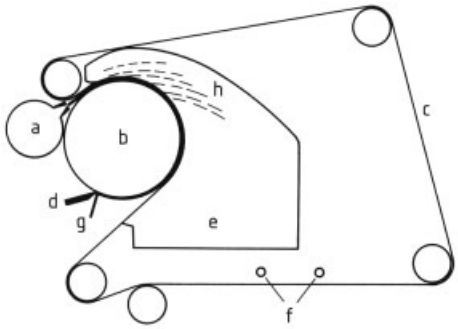

Figure 50. Schematic of a washer for the removal of ash and fines from a fiber suspension (Variosplit) (courtesy of Sulzer Escher Wyss)
a) Distributor; b) Center roll; c) Wire; d) Take-off doctor; e) Filtrate tray; f) Wire cleaning; g) Accepts; h) Filtrate

The washing effect can be further improved by disperging the printing inks and coating particles contained in the stock in a preceding stage.

Flotation [341–345]. A distinction is made between total flotation, which should remove all the solids from a suspension (e.g., purification of the circulation water) in an expansion flotation step, and selective flotation, in which impurities made hydrophobic are separated from the

fiber suspension by aeration (deinking flotation) [341].

In *total flotation* the solids are flocculated by addition of cationic and anionic chemicals [e.g., poly(aluminum chloride), polyacrylamide, poly(ethylene oxide)]. About a third of the clarified water is saturated with air at a pressure of ca. 0.9 MPa, and is then mixed with the remaining suspension after being expanded. The fine air bubbles formed during expansion adhere to the flocs and cause them to float. The clarified water and the floating sludge are withdrawn separately. A system for the purification of suspensions (e.g., the filtrate from the washing stage) is described in [343], [344] (see also Chap. 8).

Selective flotation is the specific removal of dirt from the suspension. For this purpose, air is mixed thoroughly with the suspension. The air bubbles then become attached to the dirt particles, which have previously been rendered hydrophobic, and carry them to the surface of the suspension. The resulting foam containing the dirt is withdrawn separately from the suspension. Selective flotation is used in wastepaper preparation to remove printing inks from the fiber suspension (deinking, see Section 6.3) [341].

Cleaning is the separation of impurities from a fiber suspension in a centrifugal field. The sinking and floating properties of the impurities to be removed must differ from those of the fibers. Cleaning is carried out in hydrocyclones [346]. A distinction is made between heavy-particle and light-particle cleaners, depending on the purpose of separation, and between cocurrent and countercurrent hydrocyclones, depending on the direction of the main flow. The outward (heavy particles) or inward (light particles) radial velocity of particle motion is determined by the equilibrium between the centrifugal force and the resistance to motion of the impurities. The required rotation of the suspension is usually produced by the entering velocity of the suspension into the hydrocyclone (i.e., by the pump energy), but sometimes by a rotor in the head of the hydrocyclone. The centrifugal acceleration of cleaners operating at low stock consistencies ($\leq$ ca. 1.2 %) is up to 1000 g ($g = 9.81$ m/s^2) and that of high consistency cleaners ($\leq$ ca. 6 % stock consistency), up to 40 g.

Three-, four-, and five-way cleaners are used, but the efficiency of the four-way and, above all, the five-way types is much lower than that of the three-way types [347].

Most cleaner plants are multistage systems (up to 5 stages). The overflow from the first stage is fed into the second stage. The material that flows through the second stage is returned to the inflow of the first stage and the overflow is led to the next stage, and so on in a cascade arrangement. Each stage consists of several cleaners operating in parallel because of the high flow rates to be processed [348–350].

6.1.3. Deforming/Disperging of Impurities

In the disperging system, impurities should be deformed so that they are separable in a subsequent screening stage and, above all, reduced in size to such an extent that they no longer interfere in the papermaking process or in the paper. In addition, residual fiber flakes from high-wet-strength papers should be disintegrated. Disperging is carried out with high shear forces at high stock consistencies. A disperging system has the functions of thickening, heating, and generating shear forces (Fig. 51). In order to apply high shear forces, the viscosity of the suspension is increased by dewatering it to 22 – 30 % stock consistency. Wire presses, double wire presses, or screw presses, frequently combined with thickeners, are used for this purpose. The stock is heated with steam to ca. 95 °C (pressureless) or up to 130 °C (under steam pressure) in a heating screw to reduce the shear strength of temperature-sensitive impurities such as wax, bitumen, and adhesives [351].

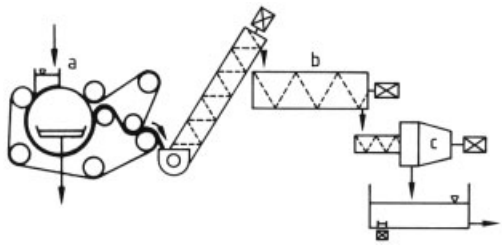

Figure 51. Schematic of a disperging system (courtesy of Sulzer Escher Wyss)
a) Wire press; b) Heating screw; c) Disperger

The disperger has fittings that consist of a series of toothed rings. The stator and rotor rings are interlocking and leave a very narrow gap. The peripheral velocity is ca. 50 – 60 m/s. The stock is driven through the gaps between the teeth and is repeatedly subjected to an intense shearing stress due to the high relative velocity between stator and rotor (Fig. 52). The specific energy transferred to the stock depends on the stock, stock consistency, stock throughput, and on the width of the gap, and is in the range 30 – 120 kW · h/t. The magnitude of the specific energy input determines the quality and strength characteristics of the stock. Dispergers can also be used as mixers, e.g., for mixing bleaching agent with stock in high-consistency secondary fiber bleaching.

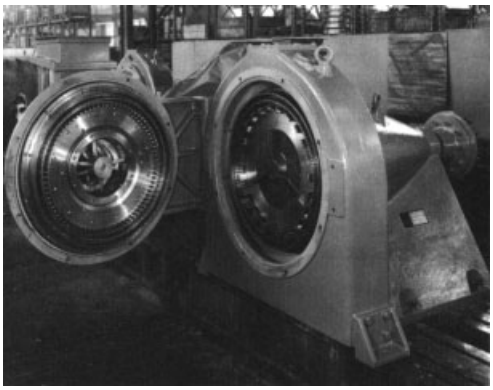

Figure 52. Disperger, with open casing, showing the fillings (courtesy of Sulzer Escher Wyss)

In other systems a slower running kneader (10 – 20 m/s peripheral velocity) is used instead of the fast running disperger [352], [353].

6.1.4. Refining

Refining is used to improve the bonding ability of the individual fibers of the finished paper as well as the fiber properties with regard to formation quality in paper web forming. The refining process determines the properties of the finished product, e.g., whether an opaque printing paper or a transparent drawing paper has to be formed [354]. Refining is carried out in low and wide-angle conical refiners or disk or double disk refiners. The refiner fillings are equipped with bars and are designed as rotor and stator. The angle formed by the bars of the rotor and the stator is called the bar or cutting angle. The bars are subjected to wear during operation. The edges of the bars, which run against each other, develop a shape that depends on the bar material and the load applied and has a great influence on the refining result [355].

During operation, the rotor and the stator are pressed together electrically or pneumatically. This influences the specific refining energy, i.e., the amount of refining energy transferred to a specific amount of stock. Besides the specific refining energy, the specific edge load is also used to assess refining. This is the ratio of the effective refining energy to the edge length per second. The effective refining energy is the difference between the power input and the pumping work of the refiner fillings produced at a rotor peripheral speed of up to 25 m/s [356]. The edge length per second refers to the length formed per second by the edges of the bars as they move past each other.

A distinction is made between shortening refining, which mainly involves cutting of fibers, and fibrillating refining, which involves the splitting of the outer layers of the fiber walls, a process that greatly increases the surface area of the fibers and their bonding ability.

Low cutting angles, sharp bar edges, and high specific edge loads result in more cutting refining. Large cutting angles, rounded bar edges, and lower edge loads have a more fibrillating effect. The beating action of a refiner is assessed by the beating energy required to increase the degree of beating ($kW \cdot h\,t^{-1}\,{}^{\circ}SR^{-1}$) [357]. This value is influenced by the type of stock processed, the degree of beating, and the type of beating (shortening/fibrillating), and is ca. $0.5 – 2\,kW \cdot h\,t^{-1}\,{}^{\circ}SR^{-1}$.

6.1.5. Storage and Mixing [358]

The storage and mixing of suspensions require different types of agitators. In the former case, the solids should remain suspended, i.e., settling out or floating on the surface should be prevented. In comparison, mixing involves the blending of the different components of the suspension. Agitators of the type shown in Figure 53 are used for storage. This rotating agitator moves automatically in a circular path through

the stock. In this way, a definite part of the contents of the chest is stirred about every 20 seconds. The rate of rotation of the agitator is 60 rpm and the specific power is 0.02 – 0.1 kW/m^3.

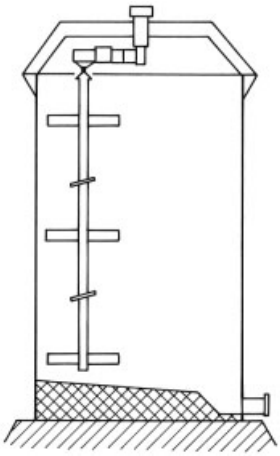

Figure 53. Schematic of a storage chest with rotating agitator to prevent settling out of fibers (Economix; courtesy of Sulzer Escher Wyss)

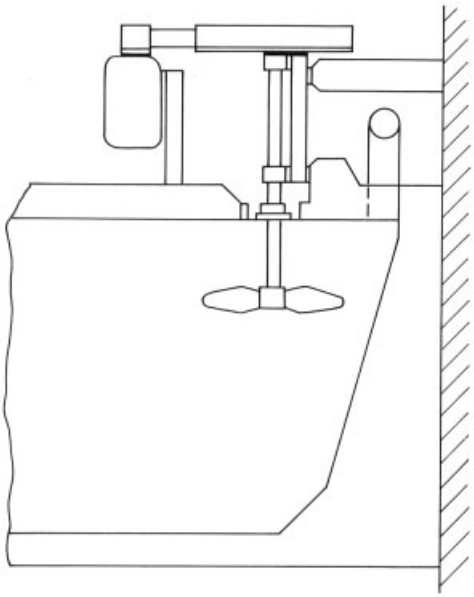

Figure 54. Schematic of a chest propeller for stock mixing (courtesy of Sulzer Escher Wyss)

The chest propeller shown in Figure 54 is used for intensive mixing. This type of agitator requires more energy because the entire contents of the chest are circulated continuously. The chest propeller rotates at ca. 130 rpm, and the specific power is ca. 0.5 kW/m^3.

6.2. Preparation of Primary Raw Stocks

Complete stock preparation for a paper machine usually consists of several lines, in each of which different raw stocks are prepared. These are then mixed in the mixing chest in proportions appropriate for producing a particular grade of paper. The processing of broke from the paper machine is also a part of this process. The broke is collected in the paper machine in the couch broke chest after the wire part, and in broke chests after the pressing and drying sections. Broke may also come from the reel-slitting machine and the coating machine. Modern stock preparation systems are largely automated, and the preparation varies according to the type of stock. As an example, stock preparation of primary raw stock in bales will be described below.

6.2.1. Preparation of Primary Pulp in Bales

An example of a stock preparation line for primary pulp in bales is shown in Figure 55. Pulp is usually delivered as bales, which must be placed individually on the conveyer belt. The wires holding the bales together are subsequently cut and removed by hand or by machine. The bales reach the pulper via a weighing device, and slushing is started at a calculated stock consistency of 4 – 6 %. The high consistency cleaner that follows is a hydrocyclone operating at a stock consistency of 4 – 6 %; it removes heavy particles to protect the deflaker from damage. The stock is then collected in the dump chest at stock consistencies of 3.5 – 4.5 % and kept in suspension by agitators.

Fiber treatment in the subsequent refining stage depends on the requirements to be met by the final paper. The refined stock is collected in the beating chest and then transferred to the mixing chest in appropriate amounts. The various components are mixed thoroughly with a chest propeller. The stock consistency is then adjusted before transfer to the machine chest. The finished stock is collected in the machine chest, and any slight residual nonuniformity in the stock consistency is smoothened out at this

stage. The mixing and machine chests are the transition to the stock approach flow system (see Chap. 7).

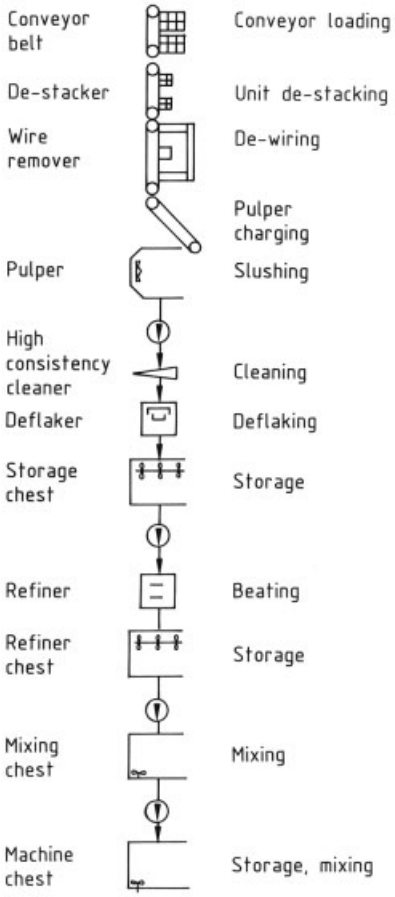

Figure 55. Schematic of a stock preparation line for primary raw stock in bales (courtesy of Sulzer Escher Wyss)

6.2.2. Special Processing Features for Various Primary Raw Stocks

Depending on the raw stock to be prepared, some process steps are of special importance or some of the process steps described above are omitted (Fig. 56). For instance, floc-dried pulp must be subjected to intensive deflaking to prevent the occurrence of "fish eyes" in paper, which are due to non-disintegrated fiber bundles. In the case of pulp being pumped directly from the pulp mill, the slushing, cleaning, and deflaking stages are omitted. Refining is not required for the various types of mechanical pulp.

6.3. Wastepaper Systems

The wastepaper processing system varies according to the pulp grade to be produced, e.g., pulps for printing, writing, household paper, or linerboard. The systems for the former three consist of a combination of repulping, deinking, screening, cleaning, and bleaching processes. Relatively mild digestion is sometimes carried out as the first stage of the process for household paper production.

6.3.1. Repulping

The wastepapers are put into a pulper together with hot water or white water, and pulped with mechanical agitation. Some chemicals, such as deinking agent, NaOH, etc., are often added as pulping additives. There are three types of pulper: low consistency, high consistency, and drum pulpers.

The *low consistency pulper* (Fig. 57 A) operates at pulp consistencies of 4 – 6 % and is classified into two types: vertical pulpers, which have a repulping rotor at the bottom of the pulper, and horizontal pulpers, with the rotor on the side wall. Also, there are batch and continuous types. Contaminants are removed during operation by a ragger and dirt trap, in order to avoid the contaminants breaking into small pieces or accumulating in the pulper.

The *high consistency pulper* (Fig. 57 B) operates at pulp consistencies of ca. 15 % by means of a special helical rotor shaped like a screw. The wastepaper is mostly disintegrated by interfiber friction, which is very effective and selective for detachment of ink. Therefore, contaminants rarely break into small pieces in spite of excellent pulping and ink detachment abilities.

The *drum pulper* (Fig. 58), also known as a fiber-flow pulper, consists of a high consistency pulping zone and a low consistency screen zone. The wastepapers fall repeatedly after being carried up by lifters on the inside wall of the rotating drum, resulting in their disintegration into flocks. The contaminants and pulp flocks are discharged continuously from the side of

Raw stock \ Process step		Slushing	Cleaning	Deflaking	Refining
Chemical pulp	sheet	●	●	●	●
	flocks	●	●	●	●
	suspended				●
Mechanical pulp		●	●	●	
Broke		●	●	●	

Figure 56. Stock preparation process stages required for various raw stocks (courtesy of Sulzer Escher Wyss)

the drum opposite to the inlet and enter the screen zone. Because the pulping is mild, removal of contaminants is successful. Pulping is sometimes preceded by sorting to remove relatively large contaminants, such as polyvinyl films, rags, polypropylene bags, wires, metals, and stones, which cause operational problems. Strainers, raggers, dirt traps, bar screens, drum screens, etc., are usually used.

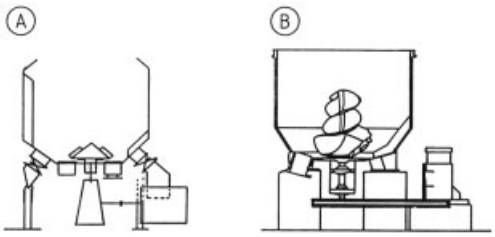

Figure 57. A) Low consistency pulper; B) High consistency pulper

6.3.2. Deinking

The deinking process is composed of detachment and separation stages.

Detachment Stage. The detachment of ink already begins in the repulping stage. The high consistency pulper in particular has a relatively high ability for detachment. In many cases, however, detachment is insufficient, and soaking and kneading stages are added in order to increase the detachment of inks.

In the soaking process, the repulped and sorted pulps are dewatered to consistencies of ca. 20 %. After the addition of deinking agents, NaOH, and H_2O_2, they are retained in the soaking tower at 40 – 60 °C. Soaking results in extensive penetration of the chemicals, increases the detachment of inks and contaminants, and decreases the number of undisintegrated pulp flocks. The resulting pulps are deflaked by high speed deflakers which further promote disintegration and detachment of inks.

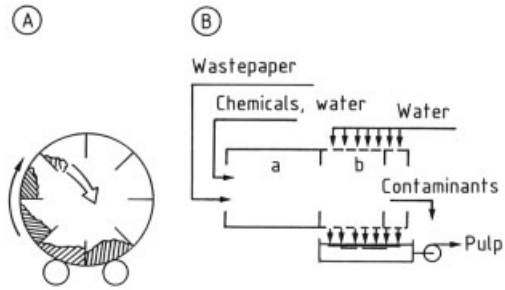

Figure 58. Drum pulper
A) End-view, showing motion of wastepaper;
B) Side view: a) Pulping zone; b) Screen zone

Inks containing large amounts of binders, such as linseed oil and resin, are transformed to nondetachable substances during the aging of

paper. They are difficult to detach from fibers by repulping and soaking alone. In this case, *kneading* is carried out with 30 % pulp consistency at normal or elevated temperature to increase the detachment and dispersion of the inks. Kneading is carried out in, e.g., single- and double-shaft mixers or screw-press pulpers. It is used in the production of high-quality deinked pulp.

Separation Stage. The ink particles liberated from the fiber surfaces and dispersed in the pulp slurry by the detachment process are removed in the separation process by flotation or washing.

In *flotation* the ink particles are rendered hydrophobic, become attached to air bubbles, are floated, and then removed. The flotation mechanism is illustrated in Figure 59. The reject ratio, deinking agent, type of flotator, pulp consistency, and residence time are the factors that influence the ink-removal efficiency. Generally, flotation conditions are: pulp consistencies of 0.8 – 1.2 %, 7 – 12 min residence time, and reject ratios of 2 – 3 %. Effective flotators have a high air-mixing rate and generate small air bubbles.

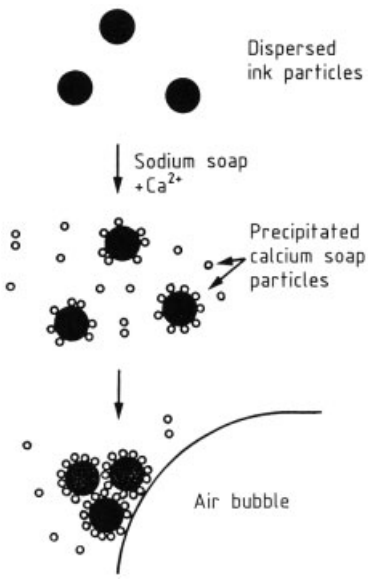

Figure 59. Mechanism for flotation of ink particles

The *washing process*, in which the ink is dispersed and then removed with the liquor in a subsequent dewatering stage, is used mainly in the United States. In other countries it is utilized as a subprocess of flotation because it is effective for the removal of fine dirt or fillers which are hardly removed by flotation. The washing process is carried out in, for example, a thickener, disk filter, screw press, extractor, or twin wire press.

Deinking agents are formulated with either dispersion- or coagulation-type surfactants. The latter type, which has high aggregative and captive forces for inks, is used in flotation; it consists mainly of fatty acids. Dispersion-type deinking agents contain, anionic and nonionic surfactants and are mostly used in washing.

6.3.3. Screening and Cleaning

Residual contaminants are removed by fine screens and cleaners. Pressurized centrifugal screens are usually used for fine screening. The screen plates have either holes or slotted orifices; the latter are effective for removal of sticky contaminants. Since the stickies are soft and flexible, they readily pass through the screen together with the pulp fibers. Therefore, the slit width of the screen plate must be as small as possible, and may be as narrow as 0.15 – 0.2 mm. In addition, special wave and profile screens, which have an uneven plate surface, have been developed. Their screening effect and specific throughput differ from those of screens with smooth surfaces.

Conventional cleaners (e.g., forward-flow types) are used to remove large and heavy contaminants such as sand and metal. Since deinked pulp contains large amounts of lightweight contraries, special cleaners have been developed for their removal. These include the reverse cleaner, in which the accept and reject outlets are reversed as compared with the conventional cleaner, and the three-outlet cleaner, which has an extra outlet for lightweight contraries collected at the center of the vortex. They tend to be applied only in the treatment of screen rejects because of their high energy consumption.

Washing, flotation, cleaning, and screening are each effective for the removal of certain sizes of contaminants, as shown in Figure 60. Therefore, a combination of all these units is necessary for the removal of all sizes of contaminants.

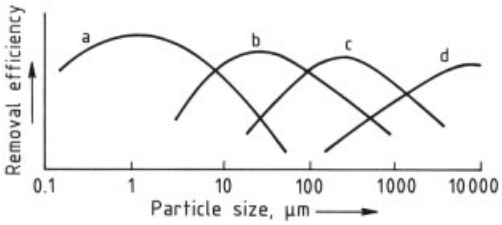

Figure 60. Particle size and contaminant removal efficiency a) Washing; b) Flotation; c) Cleaning; d) Screening

6.3.4. Bleaching

Since the brightness of wastepaper is often low and decreases further during storage or through the action of alkaline chemicals during processing, bleaching chemicals are added in the pulper, soaking tower, or bleaching tower to maintain or increase the brightness. Hydrogen peroxide and dithionite are usually used. More recently, formamidine sulfinic acid has been introduced as a reducing bleaching agent. Hydrogen peroxide bleaching is carried out in the presence of NaOH, sodium silicate, and chelating agents. Wood-free wastepapers are sometimes bleached with hypochlorite.

6.3.5. Examples of Wastepaper Processing Systems

Since various combinations of the process steps may be important for effective elimination of the contaminants, many processing systems have been proposed. Recent systems pay more attention to the removal of contaminants earlier in the process. Figure 61 shows a deinking system for the production of newspaper-grade pulps, and Figure 62 shows a system for processing corrugated containers into linerboard-grade pulp. The systems for linerboard pulps differ slightly according to the raw materials (for example, old corrugated containers, magazines, and newspapers). Although most of the systems are similar, repulping and deflaking are more important in systems using old corrugated containers because of the difficulty in disintegrating these materials.

7. Paper and Board Manufacturing

7.1. Introduction

The essential steps in the historical, manual papermaking process were sheet forming, pressing, drying, sizing, and calendering. This fundamental process has not changed since the first sheet of paper was produced more than 2000 years ago (Fig. 63).

In sheet forming, the water from a fiber suspension is filtered off on a rotating continuous wire to produce a wet web. The mass distribution in the web should be as uniform as possible. In the press section, the wet web is compressed between rolls and rotating felts by using mechanical pressure, and dewatered mechanically as far as possible. Thus, the strength of the web is increased and energy is saved in the subsequent drying section, in which the residual water, with the exception of a few percent, is evaporated by bringing the web into contact with steam-heated cylinders. In the size press, starch, size, or pigments are transferred onto or into the fiber mat. The calender finally smooths the surface of the paper.

Varying quality requirements have to be met by various types of paper, such as writing and printing paper, tissue paper, packaging paper, board and cardboard. The machines and processes used for the production of the different types of paper and board vary correspondingly.

Modern paper machines are up to 10 m wide, and the fastest machines operate at speeds exceeding 2000 m/min. A single machine for the production of packaging paper can produce over 2000 t/d (Fig. 64).

Paper consists essentially of fibers, either primary or secondary. The quality characteristics of the finished paper sheet are determined by the fiber properties, which in turn depend on the type of fibers and the treatment during stock preparation, the auxiliary agents and additives (e.g., retention agents and defoamers as auxiliary agents; fillers and dyes as additives), and the process conditions on the paper machine. The demands made on the production process have increased considerably, especially because of the use of secondary fibers from wastepaper.

Important quality characteristics of the finished paper sheet are:

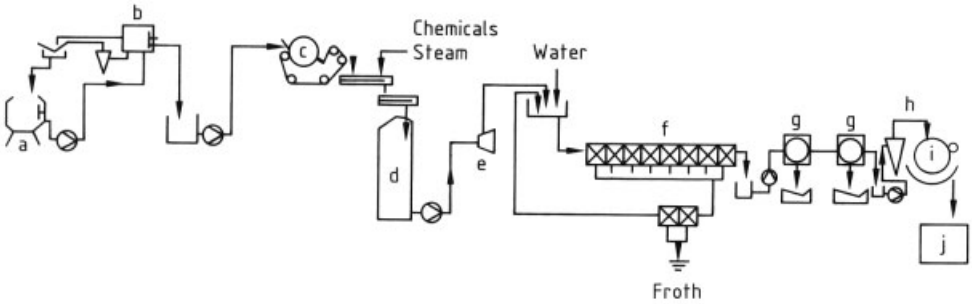

Figure 61. Deinking system for production of newspaper grade pulps
a) Pulper; b) Sorter; c) Thickener; d) Soaking tower; e) Deflaker; f) Flotation; g) Screens; h) Cleaner; i) Washer; j) Chest

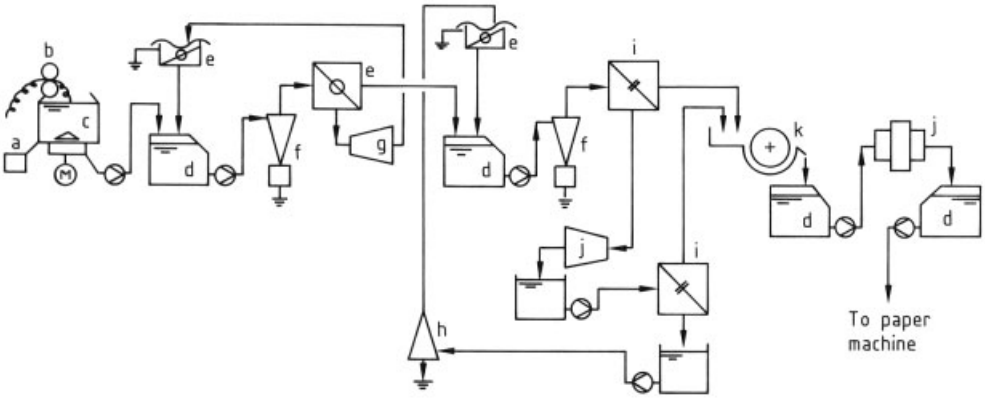

Figure 62. System for processing old corrugated containers
a) Junk box; b) Ragger; c) Pulper; d) Chests; e) Screens (hole); f) Hydrocyclones; g) Deflaker; h) Cleaner (reverse); i) Screens (slotted); j) Refiners; k) Filter

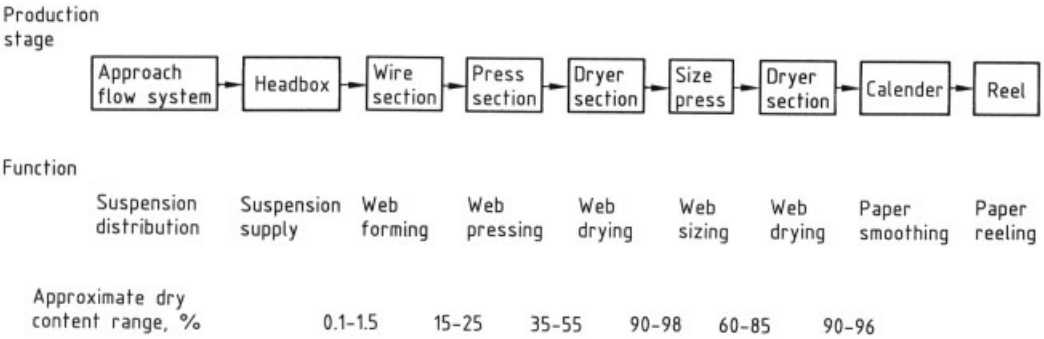

Figure 63. Flow sheet for paper and board manufacturing (courtesy of Sulzer Escher Wyss)

1) The mass distribution in the paper, which is discernible as cloudiness when the paper is held up to the light
2) The strength properties in the machine direction (MD) and cross machine direction (CD); machine direction/cross machine direction ratio, MD/CD
3) The optical properties (brightness, opacity)
4) The surface properties (smoothness or roughness)
5) The absorptive behavior, which is important for writing and printing
6) Behavior in the converting process

These properties must be constant in the longitudinal and cross directions in a given product.

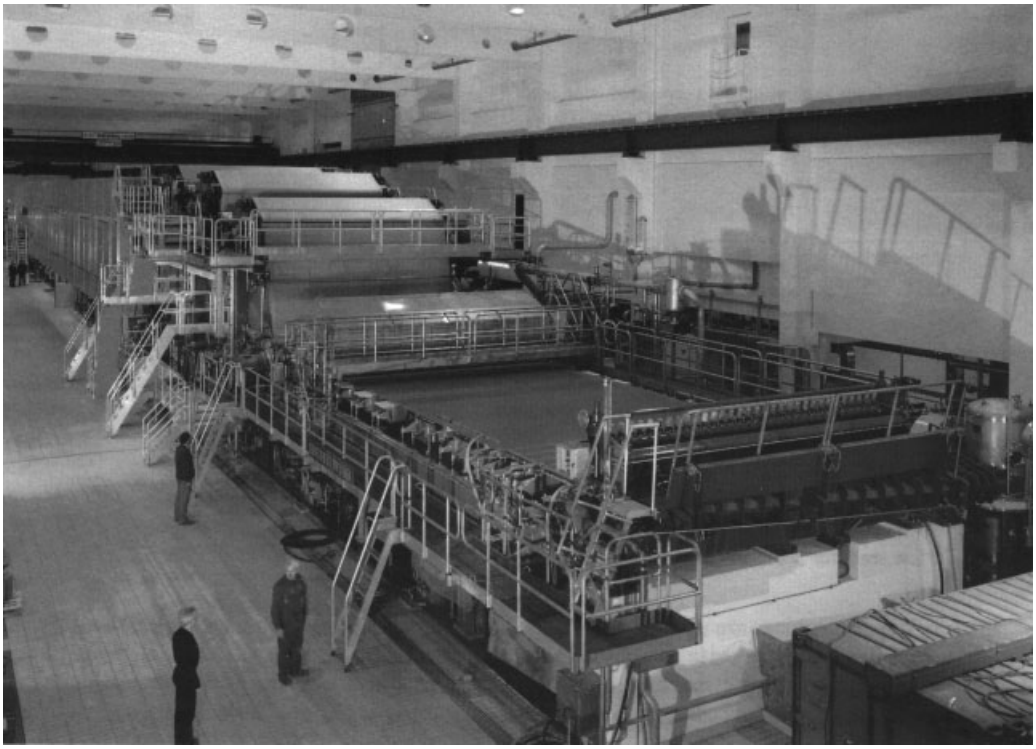

Figure 64. Paper machine, showing (from right to left) the headbox, fourdrinier, press section, and drying section (courtesy of Sulzer Escher Wyss)

The most important working agent in papermaking is water. It is required to provide the individual fibers in the suspension with as much free space as possible for undisturbed movement, slowing down reflocculation, i.e., the agglomeration of fibers to form flocs. Futhermore, water is required to produce strength in paper which is, in turn, the result of the formation of hydrogen bonds between the fibers during the drying process.

7.2. Sheet Forming [363–378]

In the sheet forming stage a uniform sheet with definite properties should be formed continuously. The sheet forming system consists of the stock approach flow system, the headbox, and the wire part.

7.2.1. The Stock Approach Flow System

The stock approach flow system (Fig. 65) is responsible for metering the amount and type of stock components used, mixing with white water, and for delivering the suspension in the required amount and with the appropriate pressure to the headbox. The suspension delivered must be of constant pressure, quantity, and composition in order to ensure constant, uniform paper quality.

After stock preparation, the various types of stocks are metered and mixed in the mixing chest at a stock consistency of ca. 3 – 4 %. Variations in the stock consistency are then adjusted by addition of filtrate water; the remaining small variations in consistency are rectified in the machine chest. The stock consistency is now largely constant, and the amount of thick stock is usually dosed via a stock level box and diluted with white water to a stock consistency of 0.1 – 1.5 %. In simple systems, the suspension is then fed to the headbox through a screen. The screen is intended to protect the headbox and the wire from mechanical damage.

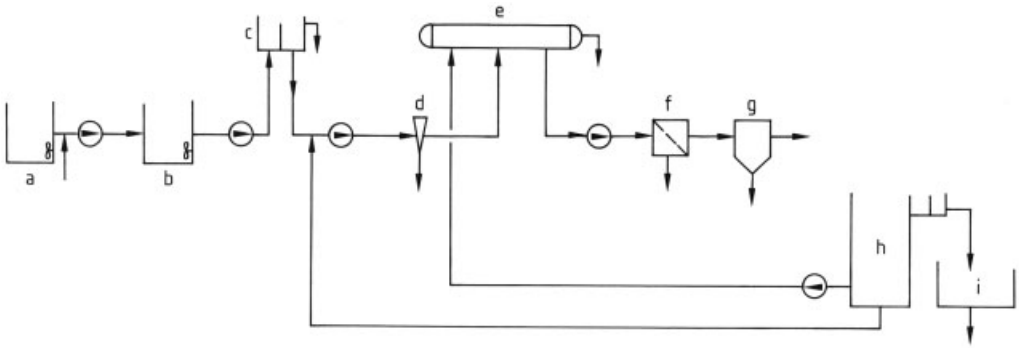

Figure 65. Schematic of the stock approach flow system (courtesy of Sulzer Escher Wyss)
a) Mixing chest; b) Machine chest; c) Level box; d) Cleaner; e) Vacuum tank; f) Pressure screen; g) Pulsation damper; h) White water I tank; i) White water II tank

In more elaborate systems, the quality of the final product is improved still further by removing sand and other heavy or light impurities from the suspension by means of conical hydrocyclones. Other impurities are separated in pressure screens with fine holes or slots. In addition, the suspension is often deaerated under vacuum to improve quality and dewaterability. Sometimes the suspension is fed through a pulsation damper.

To prevent unwanted sedimentation of fillers or the build up of bacterial slime and fiber stringing, the flow rates in the piping of the stock approach flow system must be sufficiently high (at least 1 m/s to more than 2 m/s suspension velocity, depending on the kind of stock) and the surfaces coming into contact with the suspension must be smooth.

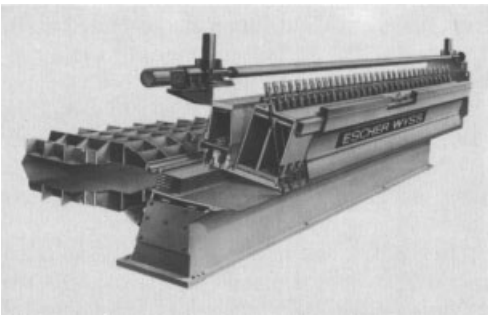

Figure 66. Example of a high-turbulence headbox (courtesy of Sulzer Escher Wyss)

7.2.2. The Headbox

The headbox (Fig. 66) produces a jet of suspension which should have a largely constant velocity, thickness, and direction across the width of the paper machine and be independent of time.

The pumping of the suspension into the headbox and cleaning with pressure screens generates pressure fluctuations, which must be dampened to keep the velocity of the jet constant. For this purpose, pulsation dampers are either used separately or integrated into the headbox.

The suspension, which is delivered through piping that has a round cross section, must be turned and distributed uniformly and with great accuracy across the full width of the headbox. It must then flow out of a narrow slit (normally called a slice, 6 – 20 mm in height) that is up to 10 m wide. In modern headboxes, this uniform distribution across the width is achieved by passing the suspension through a distributor on the backside of the headbox to a perforated plate. The many individual streams thus produced are reunited to form a single uniform stream in the nozzle, which then provides a constant jet of suspension. The breaking of fiber flocs that have already formed and the prevention of flocculation, at least for a short time until the suspension is transferred to the wire, are essential requirements for good sheet formation. For this purpose, turbulent shear forces are generated in the suspension. The turbulence should have an intensity that can be adjusted to the stock type, but the wavelength (i.e., effective length) should be short. In modern headboxes (high-turbulence headboxes) turbulence is produced by friction

in tubes and channels, or by changes in cross section as carried out in step diffusors.

In the headbox nozzle, the suspension is accelerated to approximately machine speed. The thickness of the jet and, therefore, the amount of suspension is usually adjusted by swinging the upper wall of the nozzle. The slice at the nozzle outlet is limited by a bar, which can be adjusted to an accuracy of ca. 1/1000 mm by means of spindles. This bar is adjusted locally across the width to compensate for deviations in the cross machine weight profile of the web. The direction of the jet with respect to the horizontal is influenced by shifting the upper wall of the nozzle. In this way, the point and angle of impingement of the jet on the wire can be adjusted.

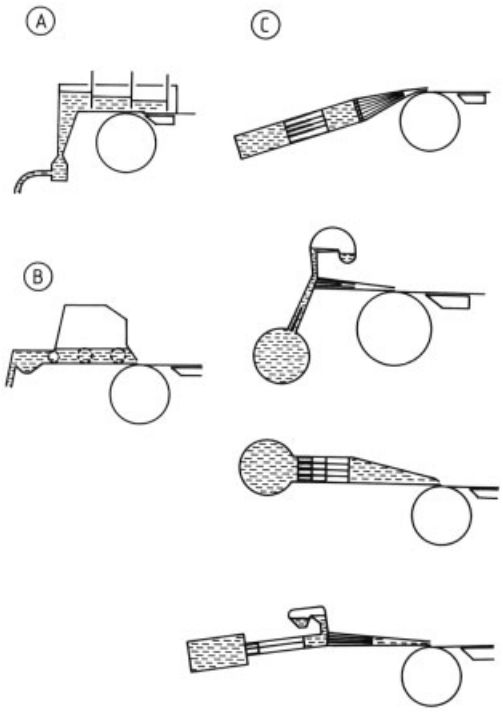

Figure 67. Schematic representation of various headbox types (courtesy of Sulzer Escher Wyss)
A) Pond slice headbox; B) Evener roll headbox; C) Various high-turbulence headboxes

In addition to modern high turbulence headboxes, there are also evener roll headboxes and antiquated pond slice headboxes (Fig. 67). The high-turbulence headboxes (also called hydraulic headboxes) have, as described above, a closed suspension guidance without a free surface. The essential functional element for distribution and dispersion in evener roll headboxes are the perforated rolls that rotate in the suspension. The suspension passes through these rolls, thereby generating shear forces. These headboxes possess a large free surface and are more suited to lower throughputs and machine speeds. The air cushion above the suspension is usually under pressure (air cushion headboxes). As an open headbox, the pond slice headbox no longer meets the most modern demands on throughput and speed.

7.2.3. The Wire Section

In the wire section, a fiber web is formed from the suspension. Three objectives are of primary importance:

1) Extensive separation of fibers from water
2) Prevention of fiber flocculation
3) Well-ordered deposition of the fibers on the wire

The separation of the fibers from water is a filtration process. The driving force for dewatering can be:

1) The height of the suspension above the wire
2) A vacuum behind the wire, produced by the action of dewatering elements
3) The pressure generated by tension in the wire over a curved surface

The forming and dewatering elements used in the wire section are (Fig. 68):

1) The forming board, which is a dewatering element at the beginning of the fourdrinier wire where the stock jet impinges. It consists of several blades or bars arranged closely together. Thus, it performs gentle, initial drainage of water from the suspension.
2) Table rolls, used in the fourdrinier section for drainage and to generate turbulence. Pressure is developed in the upstream wedge and in the region of contact between the wire and roll, and a vacuum is induced in the downstream nip.

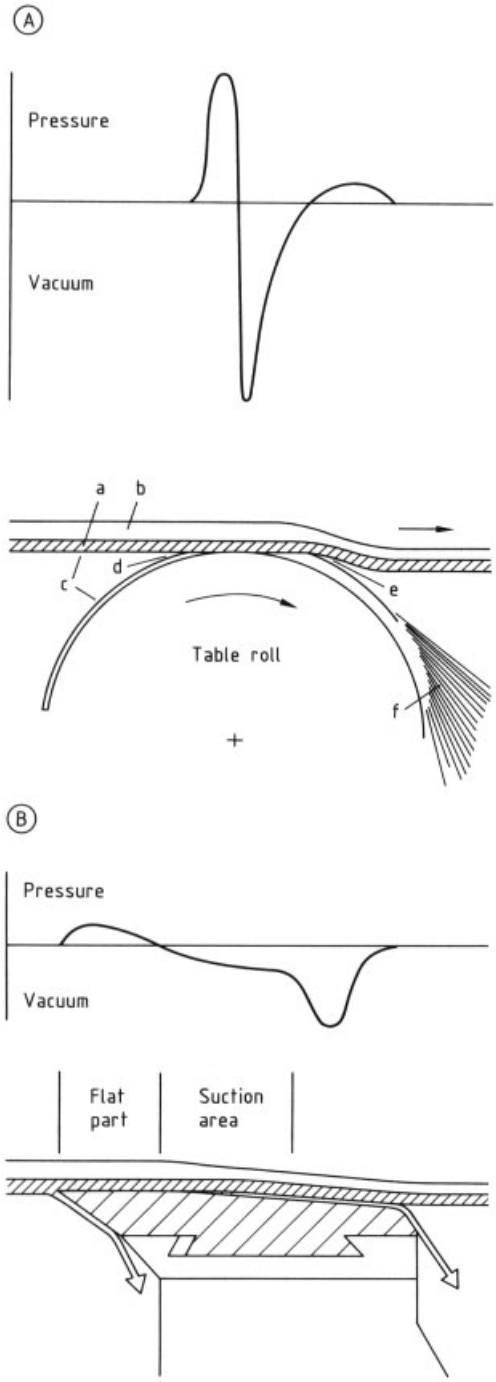

Figure 68. Schematic and mode of operation of dewatering elements (courtesy of Sulzer Escher Wyss)
A) Table roll; B) Foil
a) Wire; b) Fiber suspension; c) Water film; d) Overpressure; e) Suction; f) Discharged water

3) Dewatering foils, used with both the fourdrinier wire and twin wire formers. They have an acute-angled leading edge to doctor off the water hanging under the wire and a slope on the downstream side (foil angle of 0 – 3°) which induces a vacuum for drainage. Apart from wedge shaped foils, step foils can also be used.
4) Foil boxes which combine several foils in a unit. In addition, the foil box can operate under vacuum (vacuum foil box).
5) Wet suction boxes, which are dewatering elements that are located in front of the water line. They operate under vacuum and, in contrast to suction boxes, they drain mainly white water. The water line is a line on the fourdrinier wire that is discernible by light reflection and beyond which no free water is present on the surface of the freshly formed web.
6) Suction boxes, which are dewatering elements that are generally located behind the water line. They operate under vacuum and, in contrast to wet suction boxes, also suck air through the paper web.
7) Suction rolls, which have an open shell; sometimes vacuum is applied through the interior. They are used in sheet formation as suction breast rolls, suction formers, suction forming rolls, or suction rolls. Suction rolls accelerate drainage and increase the dry content in sheet formation.

Drainage is opposed by the resistance to filtration, which depends on the degree of beating, chemical treatment, and type of stock, and on the amount of fines and fillers present. The drainage resistance is increased by drainage that is too intensive at the beginning of the filtration process. The dry content after the wire section is 15 – 25 %.

The water removed in the filtration process (white water) carries away fibers, fines and fillers. The solids content retained, also called retention, can be increased by the addition of retention aids. The white water is reused to dilute the thick stock in the stock approach flow system.

The short-wave turbulence (microturbulence), generated in the suspension in the headbox to maintain fiber deflocculation, dissipates rapidly. For this reason, good formation

requires that either the fiber web be fixed very quickly or additional turbulence be generated in the suspension to be dewatered. This can be achieved by means of pressure and vacuum impulses from table rolls and foils. However, impulses that are too strong are harmful. Hence, table rolls are no longer used at higher machine speeds (above ca. 500 m/min) because their pressure pulses strongly increase with increasing machine speed.

Formation on fourdriniers is sometimes improved by agitating the wire. A shaker vibrates the fourdrinier wire with a frequency of 2 – 7 Hz and an amplitude of up to 10 mm in the horizontal direction.

The difference between the velocity of the jet and the wire is decisive for the controlled deposition of the fibers on the wire. If the jet and the wire have the same velocity, the fibers are deposited with random orientation or according to a possible preorientation in the headbox nozzle. If the jet is slower or faster than the wire, more fibers are aligned in the machine direction. The highest value for the tensile strength of paper is observed in the direction of the main fiber orientation. The relationship between the properties in the longitudinal and cross directions is often important in the processing and use of paper. If the jet is not directed exactly in the machine direction, this angular deflection is magnified on the wire many times over. The main fiber orientation is then no longer in the machine direction, which can lead to problems (diagonal sheet stress) in certain types of paper (e.g., copying paper).

The nature of the wire, which acts as a filtering auxiliary layer, influences the surface properties of the web (wire mark), fiber orientation, retention, dewatering velocity, and machine operation. Important parameters are the topography of the wire surface, resistance to fluid flow, free volume, cross stability (so that the wire remains level), and the wear properties of the wire. Today, single layer or multilayer woven wires are used. The material employed is predominantly plastic, but bronze is still used occasionally.

During the forming process, the suspension is guided between the wire and air (fourdrinier wire), between the wire and a solid wall (former), or between two wires (twin wire formers). Drainage can occur on one or both sides of the web. Two-sided drainage produces a more symmetrical paper. The different principles involved in sheet formation are shown schematically in Figure 69.

The Fourdrinier Wire. The fourdrinier wire (see Fig. 69 D) is the classical method of sheet formation. Today, speeds exceeding 1200 m/min are achieved with the fourdrinier. Normally, the fourdrinier is fitted with a wire table, foils and/or table rolls, suction boxes, wet suction boxes, and suction rolls. Drainage proceeds in the direction of gravity. A dandy roll is often used just in front of the water line to improve formation. This is a wire-covered open roll, with a honeycomb structure, which dips into the suspension and distributes the fibers in the web more uniformly. The dandy roll is also used to produce watermarks in the web through displacement and compression of fibers. These marks are visible in the finished paper in transmitted light. They are generated by a raised pattern soldered onto the roll.

Cylinder Formers. In the cylinder former (see Fig. 69 A), web formation occurs on a wire-covered, water-permeable cylinder. The uniflow and contra-flow former with an immersed mold represents the oldest design. These were later replaced by the nonimmersed mold former and finally by the suction former. In suction formers, drainage is further increased by reduced pressure in the interior of the forming cylinder. The suspension is led between the wire cylinder and a solid wall (lip, former box). Suction formers are employed in the making of the single layers of board at lower speeds. Higher speeds (up to 1500 m/min) are only achieved with the suction breast roll former in the production of tissue.

Twin Wire Formers. In twin wire formers, (Fig. 69 F), the suspension is led between two wires operating at the same speed, and is drained through one or both sides. One of the driving forces in twin wire dewatering is the drainage force due to wire tension, which counteracts the centrifugal force of the suspension. Open rolls, suction rolls, forming blades, and vacuum shoes are used to increase drainage capacity and improve formation. In the twin wire former, the jet is injected directly into the gap between the two wires (gap former). The first section of the hy-

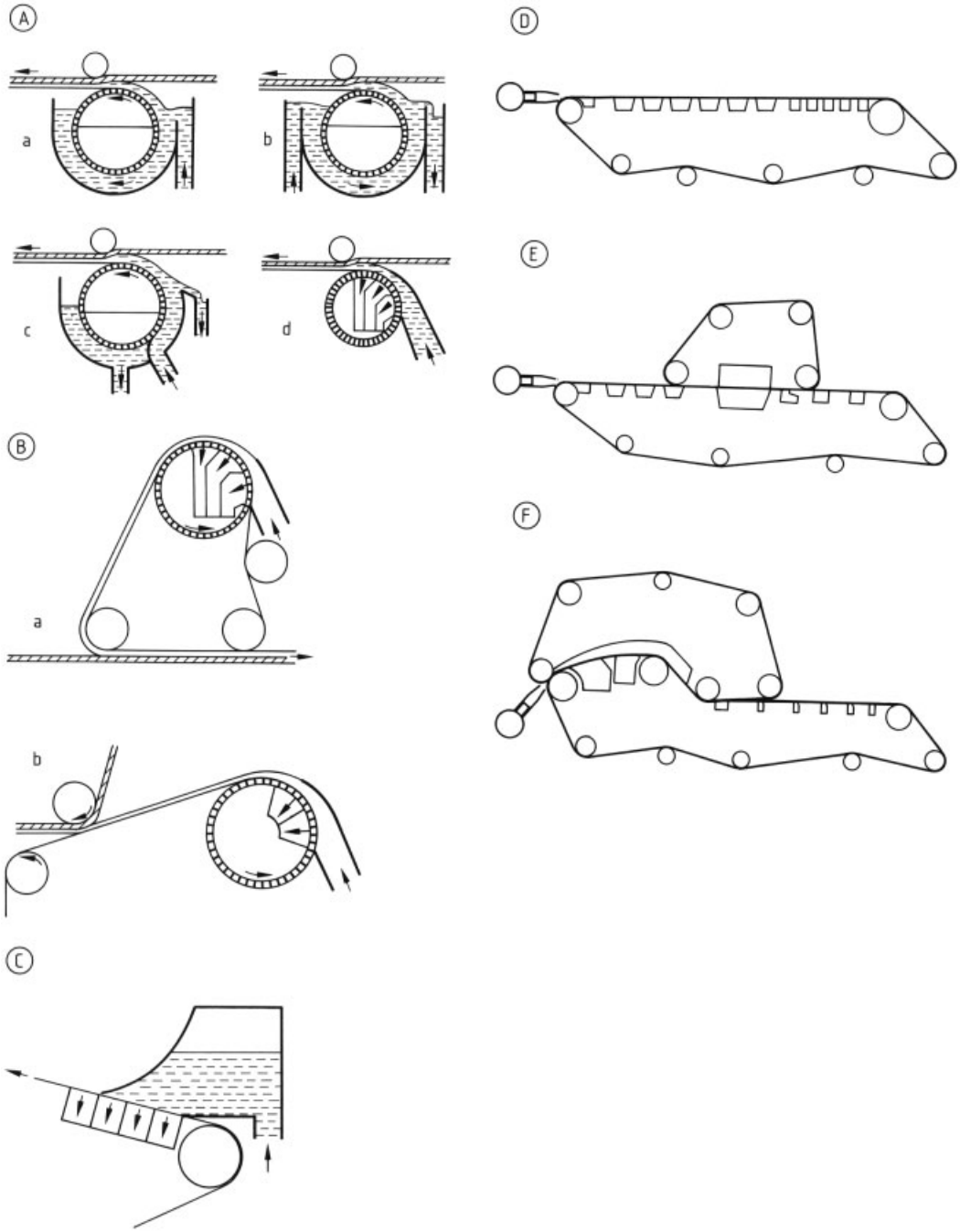

Figure 69. The various principles of web formation (courtesy of Sulzer Escher Wyss)
A) Cylinder formers: a) Contra-flow former; b) Uni-flow former; c) Nonimmersed mold former; d) Suction former
B) Suction former with rotating wire: a) Former isolated on fourdrinier; b) Suction breast roll
C) Inclined wire; D) Fourdrinier; E) Hybrid former; F) Twin wire (gap) former

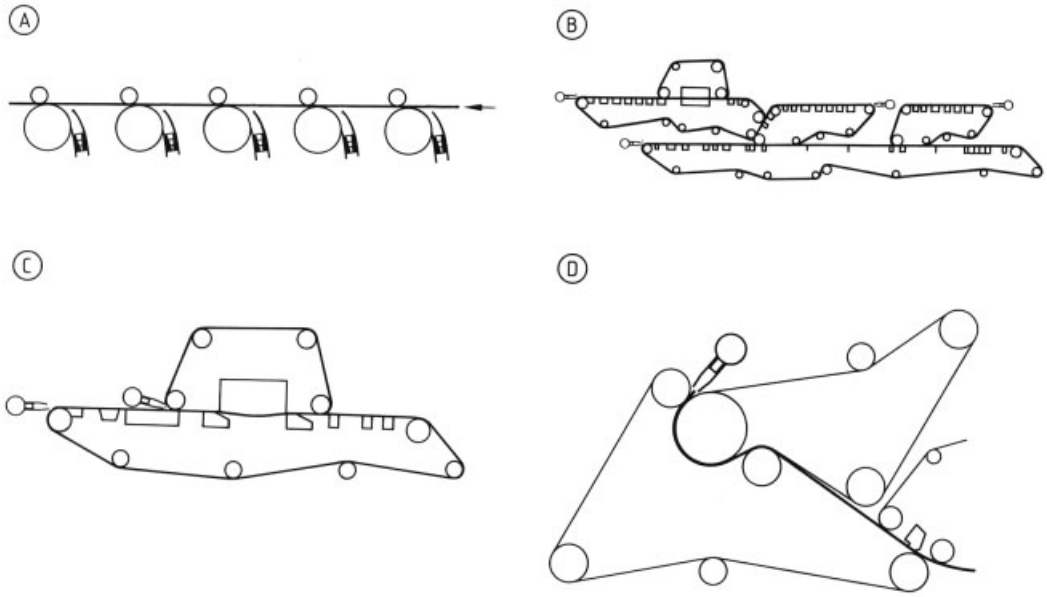

Figure 70. Various methods for the production of multilayer products (courtesy of Sulzer Escher Wyss)
A) Suction forming section; B) Fourdrinier coupled with fourdriniers and a hybrid former; C) Hybrid former with secondary headbox; D) Tissue former with double-layer headbox

brid former consists of a fourdrinier, which is followed by a twin wire section in which dewatering also occurs through the top wire.

7.2.4. Systems for the Production of Multilayer Products

Multilayer products such as board are produced by forming the individual layers separately in several web-forming units and then couching them together in the wet state (multi-former). This process gives the entire web improved uniformity compared with a single layer web. Since the cross section of the web can be tailored to suit its intended purpose, and different materials can be used to form the individual layers, increased functional value and improved economy can be achieved. The latter applies also to products that are formed on a single web-forming unit in which the headbox delivers a jet having two or three layers of varying stock composition (multi-layer headbox). Here, dewatering is usually carried out in a twin wire former. The use of a secondary headbox allows the application of a further layer of suspension to a web that has already been formed on a fourdrinier wire. Dewatering then occurs through the web downwards or in a hybrid former to the top (Fig. 70).

7.3. Press Section [379–388]

The purpose of the press section is to increase the dry content of the paper web by dewatering and to increase the strength of the web by compression before the web enters the subsequent dryer section.

The press section usually consists of several successive press nips, formed between two rolls (Fig. 71). The rolls are pressed with linear forces of 20 – 150 N/mm and up to 1000 N/mm in some special designs. Today, the dry content after the press section reaches 55 % and more, depending on the product and the raw material used. Because of these high dry contents, less thermal energy is required in the dryer section, and the increased web strength results in fewer breaks.

In modern press sections, the web is transported between one or two felts through the nip. The water removed from the web is stored and carried away in the felt and, in the case of high water extraction, also in grooves or perforations in the surfaces of the rolls. With suction press

rolls, the storage and transportation of water are improved and rewetting of the paper web is reduced. The most important properties of the felt are the storage capacity under pressure, the water retention behavior on leaving the nip, which, in turn, prevents rewetting of the web, and the smoothest possible surface in the nip for uniform pressure transfer to the web. To meet these requirements, the felts are equipped with a coarser, frequently multilayered, base fabric and a fine, surface layer made of synthetic and natural fibers needled into the base layer. The dewatering of the felts is performed with suction pipes. In order to maximize the working life of the felt, it is constantly or intermittently treated with a fine, high-pressure water jet and/or wetting facilities (felt conditioning) to keep the felt clean.

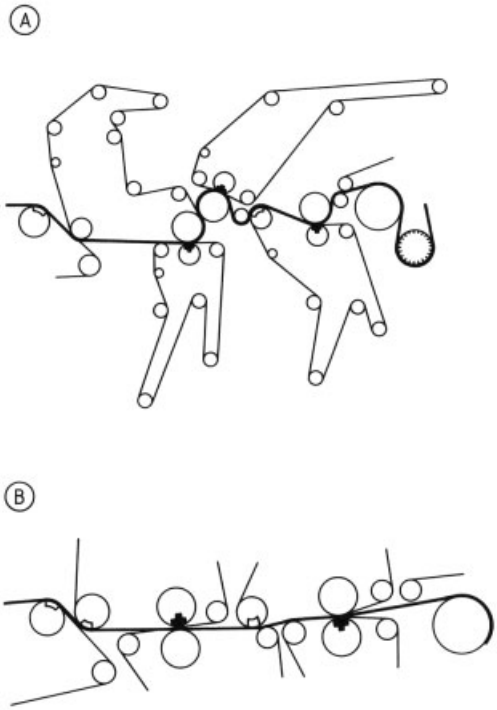

Figure 71. Schematic of two press sections (courtesy of Sulzer Escher Wyss)
A) Press section of a newsprint machine with rolls for line force control across the machine width; B) Press section of a packaging paper machine with two wide-nip rolls

The magnitude of the pressure applied and the time in which it acts depend mainly on the diameter of the roll, the elastic properties of the surface of the roll or the roll coatings (rubber, plastic, steel, or bronze; grooved/perforated) and on the felts, and less on the elastic properties of the web. The applied pressure must first overcome the structural resistance of the web before it can generate the hydraulic pressure required for web dewatering. The higher the hydraulic pressure and the time of action, the higher the dewatering performance. The pressure is increased from nip to nip, corresponding to the decreasing water content. If the hydraulic pressure is too high, however, the paper web can be damaged (crushing).

The dewatering pressure is counteracted by the structural strength of the web and the flow resistance of the water in the web, which is determined by the type of stock, the fiber treatment, and the amount of fines and filler. Increasing the web temperature by steam or infrared heating facilitates dewatering by lowering the viscosity of the water and the structural strength of the web.

The time during which the pressure acts on the web can be prolonged almost tenfold by using wide nip pressing. In this process, a concave hydrostatically or hydrodynamically supported shoe presses a flexible revolving plastic sleeve or a plastic belt against the other press roll.

A closed web guidance system is important to prevent web breaks, especially at high machine speeds, low basis weights, and low dry contents of the web. Indeed, in modern paper machines the web is not conveyed freely, but it is nearly always supported by a felt or the surface of a roll.

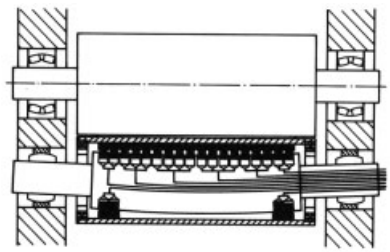

Figure 72. Example of a press roll with which the line force can be varied locally across the width (courtesy of Sulzer Escher Wyss)

A uniform moisture profile in the cross machine direction on the reel is very important for reasons of quality and economy. Press rolls which can vary the line force selectively across

the width (Fig. 72) and sectioned steam boxes which heat the web to varying extents across the width are suitable for cross machine moisture profile correction.

7.4. Dryer Section [389–393]

In the dryer section, water is removed by heating, whereby fiber – fiber bonding via hydrogen bonds is effected and the web strength is increased. The dry content of the web after the dryer section is 90 – 98 %.

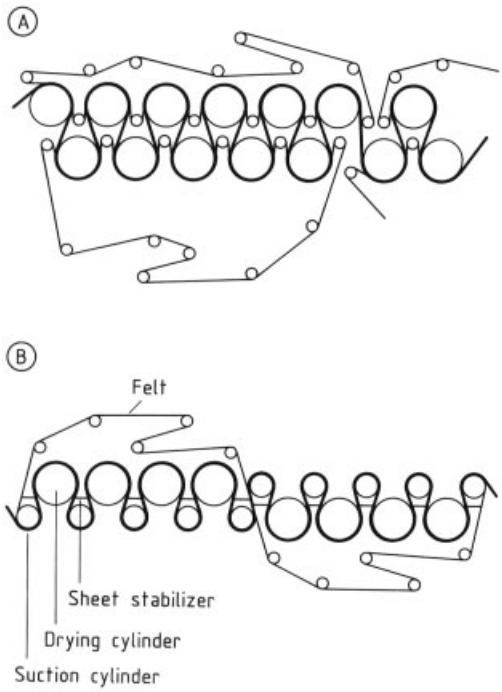

Figure 73. A) Schematic of a multiple cylinder dryer section; B) Schematic of a multiple cylinder dryer section for high speeds and for minimizing paper shrinkage in the cross machine direction (courtesy of Sulzer Escher Wyss)

In the multiple-cylinder dryer section, the web is transported in a slalom course over a number of steam-heated cyliders (contact drying). Dryer felts or dryer fabrics press the web against the drying cylinder for better heat transfer (Fig. 73). The dryer felts are, in turn, freed of water by means of felt dryers (steam-heated cylinders, blowpipes, blow boxes). The dryer section is aerated with hot air (pocket ventilation) to carry off the water vapor generated and to increase the drying efficiency. The entire dryer section is enclosed in a drying hood.

Apart from contact drying, which is the most commonly used process, forced convection drying and infrared drying are also used in the paper industry, usually in combination with contact drying.

The paper web shrinks during drying. The extent of shrinkage depends on the type of stock, degree of beating, the fiber orientation, and on forces applied that counteract shrinkage. The shrinkage in the longitudinal direction can be controlled by stretching the paper between the individual drying groups. Shrinkage across the width is nonuniform and is higher at the sides than in the center of the web. Attempts have been made to influence the shrinkage across the width by using a special web guidance system. In this system, the web in one row of cylinders is pressed against the drying cylinders by the drying fabric, and in the other row, it is held by reduced pressure in suction rolls and partly by web stabilizing elements on the drying fabric (see Fig. 73).

7.5. Size Press [394], [395]

In the size press, starch, size, or pigments are applied to the fiber matrix to increase the strength of the paper web and to modify the surface properties with respect to liquid uptake during writing, printing, or coating.

In sizing, the web is passed through the sizing liquor pond, which is located above a roll nip. As a result of capillary action in the pond and hydraulic pressure in the roll nip, the paper web absorbs the sizing liquor. Some modern size presses for high machine speeds no longer have a pond. A thin layer of size is applied to the rolls and is subsequently transferred to the paper web in the nip (Fig. 74).

The amount of size taken up depends on the dry content of the web. To guarantee uniform size uptake across the width, the web must be pressed uniformly across the width in the nip and must exhibit a constant moisture content. The dry contents before the size press are high and can reach 98 %.

The water applied in the size press is evaporated in the after-dryer section. The first cylinder

in particular must be protected against the formation of coatings.

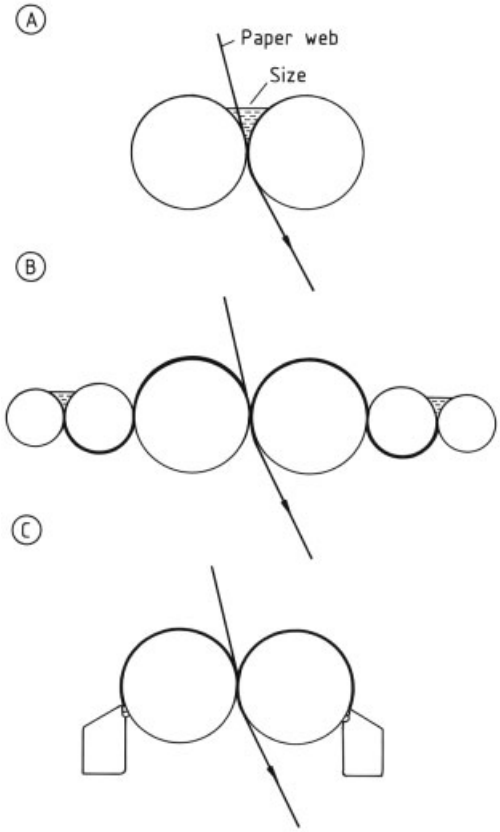

Figure 74. Size presses (courtesy of Sulzer Escher Wyss) A) Size press; B) Gate roll; C) Speed sizer

7.6. Calendering [396–402]

The objective of calendering is to produce a smooth paper surface which meets the printing and writing requirements for the intended use. Not only must smoothness and gloss be produced, but thickness reduction is also important in some cases. However, the increase in smoothness and gloss must frequently be as high as possible, but thickness reduction of the web must be as low as possible. Apart from thickness, the opacity, tear strength, and absorptivity of the paper are reduced to a certain extent by calendering.

In calendering, the web is fed through counteracting press rolls and in this process the surface roughness of the web is leveled by the action of pressure and sometimes temperature. Calenders consist of two or four to eight cast-iron rolls with very smooth surfaces that are arranged one on top of the other. The web is passed through the press nips of these rolls (hard nips). Calenders that are being increasingly used today have a heated cast iron roll and a mating roll with a "soft" plastic cover (soft nip). Two or more of these are arranged one behind the other. The rolls are heated externally by induction or internally with hot water, steam, or heating oil (Fig. 75).

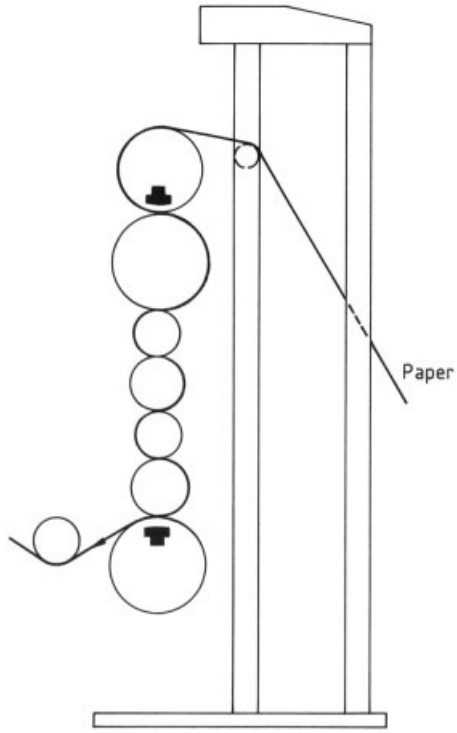

Figure 75. Example of a seven-roll calender (courtesy of Sulzer Escher Wyss)

The level of smoothness that can be attained is greatly influenced by the type of stock, the moisture content and temperature of the paper web, the pressure applied, the retention time in the nip, and the surface temperature of the roll.

On entering the nip, irregularities in the web with regard to thickness, web weight, moisture and temperature, and local differences in the pressure applied have a negative effect on calendering and on the product. This can be cor-

rected by using rolls which can specifically adjust the linear force locally across the width (cf. Fig. 72). Moreover, the rolls can be heated or cooled locally across the width with blow jets, by which the diameter of the rolls is changed locally. The roll diameter is also changed locally by induction heating that is controllable across the machine width. In some cases, the web is specifically steamed before calendering.

7.7. Supercalendering [403–406]

The paper or board web that has been presmoothed in the calender is often subjected to a further finishing treatment called supercalendering. A uniform, high surface smoothness is produced in this process, which is carried out independently of the paper machine, usually in supercalenders and sometimes with gloss and glassyne calenders. This treatment also improves the gloss. Supercalendering increases printability, e.g., in the case of picture printing. Before supercalendering, the web is frequently checked in rewinders for possible defects. Defective parts are eliminated and the ends of the web are spliced together again. Supercalenders usually process the web in its full width as it comes off the paper machine or the coating machine. The system consists of the reel-off stand, the actual supercalender, and the reel-up stand (Fig. 76). The supercalender is more than 10 m high, the working width up to 10 m, and the maximum working speed ca. 1000 m/min. The roller stack usually consists of 12 (or more) rolls, half of which are cast-iron rolls of a high surface quality (the diameter of the top or bottom roll is up to 1000 mm and that of the other rolls is up to 560 mm) that are sometimes internally heated, while the other half are paper rolls (called filled rolls, diameter up to 800 mm), which have shells composed of paper rings that consist of pure cotton or a mixture of cotton and wool or soda paper, depending on the desired property (e.g., hardness). These rings are threaded onto a steel core and are pressed together axially under a high pressure. This stack is then turned and ground. The hardness of the cover is ca. 82 – 90 Shore D. The cast-iron rolls and the softer paper rolls are installed alternatingly. At about the middle of the calender, the sequence of the rolls is reversed so that both sides of the web come in contact with the paper rolls. The hard roll presses into the soft roll on loading. The resultant relative movements between the two rolls confer smoothness and gloss on the paper web as it passes through the roll nip. After each nip, the web is led over paper guide rolls to compensate for width increases of the web produced by the supercalendering process in the nip, and to prevent curling.

In modern supercalenders, the top and the bottom roll are deflection control rolls, which can be used to adjust nip forces locally across the width. This system permits a uniform loading of the web across the width in all the nips over a wide range of linear forces. Sometimes the entire roll stack floats, in which case very uniform loads across the width are achieved.

In the case of web breaks or other disturbances, the roll stack is opened quickly so that the sensitive paper rolls are not damaged. To save the threading time for each new roll of paper, the beginning of the new roll is sometimes stuck (flying splice) to the end of web of the old roll at low speeds (max. 40 m/min). Subsequently, the machine is quickly switched to normal operating speeds.

The result of supercalendering depends not only on the linear forces, speed, and surface temperature of the roll, but also on the paper properties, especially on the type of fibers, coatings, and on the moisture content and temperature of the paper web.

Other calender types which are less important are as follows:

1) Calibrating calenders that calibrate the thickness of the web (usually with cardboard)
2) Embossing calenders that transfer a pattern onto the surface of the web (usually with cardboard)
3) Friction calenders that utilize the friction between separately driven rolls of a roll pair to smooth the web
4) Sheet calenders that are used for heavy grades (above ca. 300 g/m^2) and are especially gentle in bending the thick product.

7.8. Reel-Slitting Machine

In the reel-slitting machine, the paper web is cut to the width required for further processing.

(A)

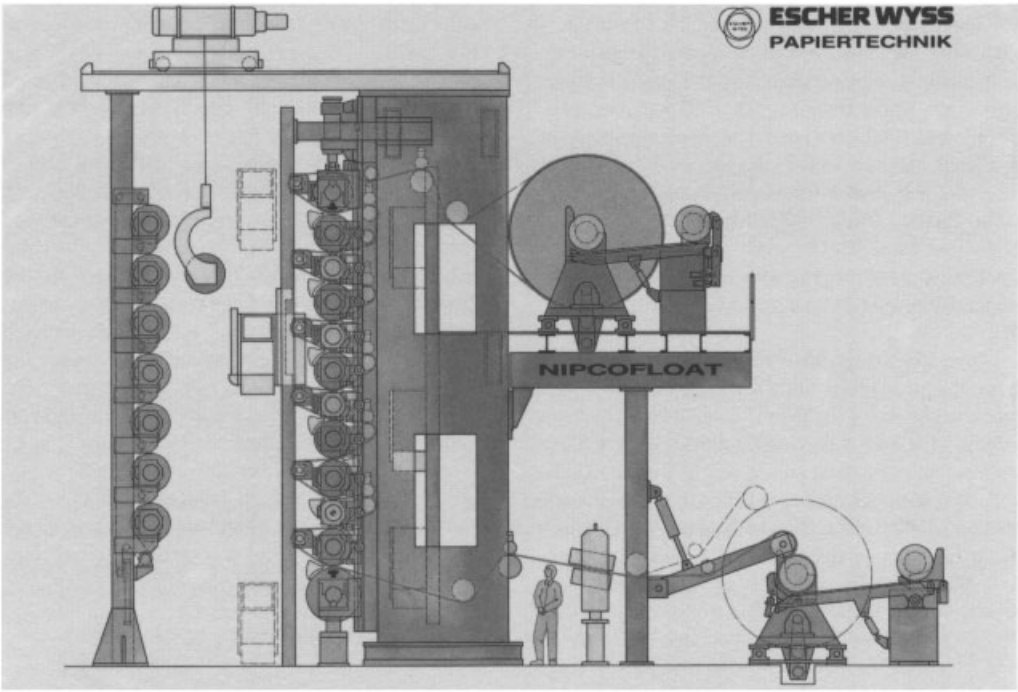

(B)

Figure 76. Schematic (A) and photo (B) of a supercalender (courtesy of Sulzer Escher Wyss)

Reel-slitting machines process web widths of up to 10 m, working speeds exceed 2500 m/min, and diameters of individual rolls reach 1500 mm and sometimes 2400 mm. The functions of a reel-slitting machine (Fig. 77) include

1) Unreeling the roll present in the full paper machine width
2) Cutting to the desired web width
3) Reeling the individual rolls on to separate cores

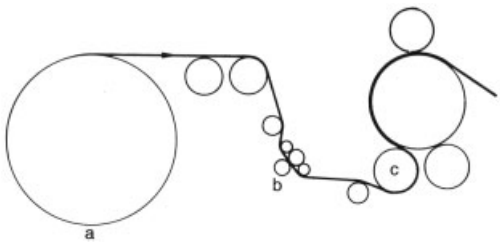

Figure 77. Schematic of a reel-slitting machine (courtesy of Sulzer Escher Wyss)
a) Machine-width roll; b) Rotating circular knives; c) Wind-up drum

The web is cut with rotating circular knives of high-alloy tool steel, hard metal, or ceramic. At a defined contact pressure, the disk-shaped upper knives penetrate into the groove of the lower knives. The positioning of the knives and, consequently, the adjustment of roll width is performed manually or automatically.

Roll changing is automated, and changing times of less than 40 s have been achieved. In the reeling up of the individual rolls, the reel hardness is important for further use. Apart from paper tension, the reel hardness is essentially determined by the linear force of the load roll, which can be controlled for this reason.

The most important requirements to be met by the cut rolls with regard to further processing (printing, cross cutting) are uniform reel hardness across the width of the roll and in the diameter, straight front surfaces, and absence of roll defects and roll damage.

8. Water Circuits and Effluent Treatment

8.1. Introduction

Papermaking is a branch of industry that requires large amounts of water. However, the fresh water demand and, consequently, the wastewater emission can be reduced substantially by the realization of almost closed water circuits and the reuse of process water. The water is recycled either directly or via clarifiers (save-alls). Fibers and fillers are removed from process water by sedimentation, flotation, or filtration. The resulting clarified water is reused in production and replaces fresh water at many points in the production process. The separated solids are also reused. Hence, the closure of water circuits is favored from the viewpoint of environmental protection, the saving of resources, and economy of operation.

8.2. Fresh Water

Primarily surface water and to some extent groundwater are used to meet the water demand of the paper industry. The essential criteria for the choice of sources of water are sufficient availability and the local conditions. A combined utilization of water sources is common. Irrespective of its source, fresh water should neither reduce the quality of the paper nor cause disturbances in the papermaking process (e.g., score breaks, lime precipitation, corrosion). Surface water frequently does not meet these requirements and must be treated before use.

Sedimentation processes (with chemical flocculation) and/or gravel filter processes are principally used for conditioning fresh-water treatment. In the production of specialty papers (e.g., photographic base paper, filter paper, cigarette tissue, capacitor tissue), fresh water is also deferrized, demanganized, softened, and desalinated (ion exchanger). Boiler feed water for power plants within the mill is always softened and/or desalinated.

Fresh water is chiefly used:

1) As cleaning water for the paper machine (wire and felt cleaning)
2) As sealing water and confining water in stuffing boxes, suction boxes, and cleaners

3) As a solvent and dispersant for fillers and additives

Fresh water used as cooling water is sometimes introduced into the process in a separate pipeline and removed after use. However, it is usually also employed as process water. In calculating the specific fresh water consumption, cooling water is not taken into account provided that it had no contact with fibrous materials and additives.

8.3. Water Circuits

There are three process water circuits in a paper mill [407], [408] (Fig. 78): the primary circuit, the secondary circuit, and the tertiary circuit. In the primary circuit, the fiber-rich water obtained in the sheet-forming zone of the wire section (white water I) is recycled for stock dilution in the stock approach flow system. The primary circuit is maintained as closed as possible. It is the most important circuit in terms of the amounts of water and stock.

Excess water from the sheet-forming section, suction and press water, as well as cleaning water is called white water II and is circulated in the secondary circuit. A save-all is generally installed in this system. Fibers and filler materials are largely removed from the circulating water by sedimentation, flotation, and filtration. The resulting clarified water is returned to the process circuit, where it replaces fresh water at several points. The separated solids, a concentrate containing fibers and fillers (3 – 10 % consistency), is used for suspension and dilution in stock preparation.

The tertiary circuit, also called the outer cycle, contains excess water from the secondary circuit and, in addition, all other process water which is not, or is not directly, reusable because of its degree of contamination. The water in this circuit represents the wastewater of a paper mill; it is treated in chemomechanical and/or biological wastewater purification plants. Purified water can be either completely (closed water circuit) or partially (tertiary circuit) returned to the production process.

Residual substances are obtained as sludges in wastewater purification. These can either be dewatered and disposed of or reused in production. In particular, fiber-containing sludges from chemomechanical clarification plants are returned to the process. In some cases (e.g., in board production [409]), sludges from biological wastewater treatment plants are also reused. However, these sludges are usually disposed of in dumps or by incineration or used as raw materials in the production of other products, especially in the brick industry.

The specific amounts of wastewater listed in Table 17 can be used to classify the closure of water circuits within the mill [410]. The specific wastewater is the ratio of the amount of wastewater to the gross production of the mill. It is 1 – 2 L/kg less than the specific fresh water consumption because evaporation losses occur in the dryer section of the paper machine. Almost closed water circuits are used in the production of packaging papers from wastepaper (corrugating medium, screeenings). The reusability of process water and therefore the extent to which the water circuits can be closed are limited by, for example, buildup of salt loading, enrichment of dissolved organic substances, increase in temperature, slime formation, corrosion phenomena, and odors. Causes and remedial measures are described in [411–414].

Table 17. Classification of paper mill water circuits [410]

Water circuit	Specific effluent volume, L effluent/kg paper
Open	> 100
Less closed	10 – 100
Almost closed	< 10
Completely closed	0

8.4. Closed Water Circuits

A closed water circuit is achieved when no wastewater leaves the paper mill. The process water is then utilized to almost 100 %. Only evaporation and vaporization losses, and the water content in the paper and in the residual matter must be replaced by fresh water (1 – 2 L/kg).

A closed water circuit is shown schematically in Figure 79. The essential requirement for reusing circulating water in the tertiary circuit is adequate save-all clarification. The quality of the clarified water must be suitable for the operation of sensitive systems (e.g., showers,

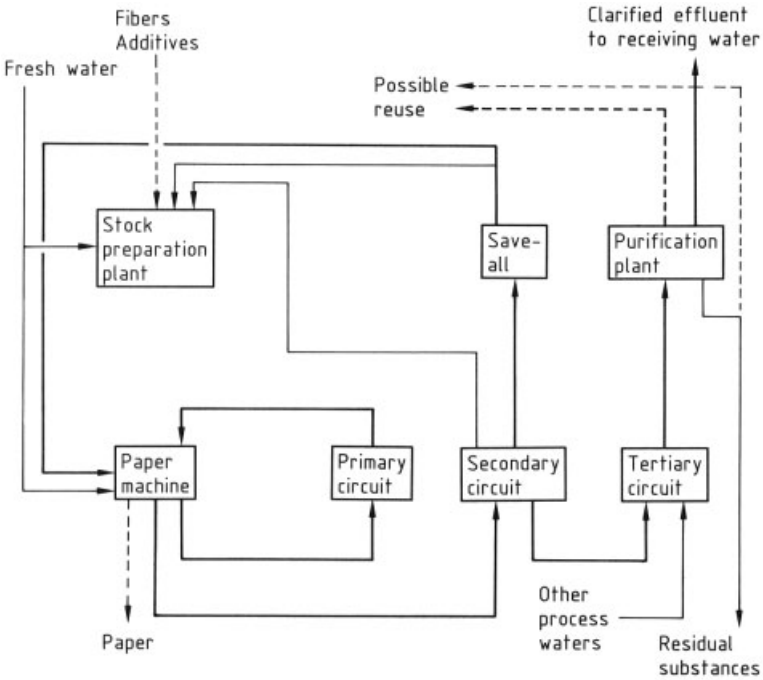

Figure 78. Scheme of the water circuits in a paper mill [407]

trim showers). There must be enough storage capacity to compensate for variations in quantity caused by disturbances and interruptions in production. Until now, closed circuits have been realized only in paper mills processing wastepaper. Measures and operating experiences are described in [415–417].

8.5. Save-Alls

Save-alls have a dual function: stock recovery and water clarification. The save-alls used almost exclusively in the paper industry are based on sedimentation, flotation, and filtration. The requirement for optimal solids recovery efficiency is the flocculation of the fibers and fillers to form larger agglomerates. High-molecular, water-soluble polymers (e.g., polyacrylamides, polyamines, polyethyleneimines) or inorganic electrolytes (iron(III) chloride, aluminum sulfate, calcium hydroxide) are used as flocculants.

Apart from economic factors, the choice of the save-all system depends on the solid composition (fiber and filler components), space requirement, reliability of operation, energy consumption, and chemicals consumption. The advantages and disadvantages of the individual systems are presented in Table 18.

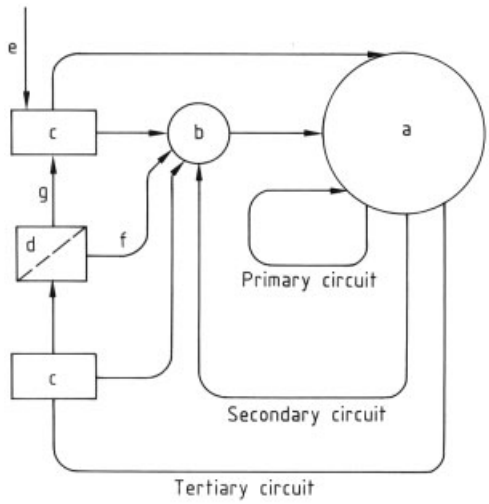

Figure 79. Scheme of a closed water circuit
a) Paper machine; b) Stock preparation plant; c) Storage; d) Save-all; e) Fresh water; f) High density stock; g) Clarified water

Table 18. Advantages and disadvantages of various save-all systems [408]

	Sedimentation	Filtration	Flotation
Space requirement	large	very low	low
Holding time	long	very short	short
Buffer efficiency	moderate	none	slight
Sensitivity to load variations	low	low	somewhat higher
Control and maintenance requirement	low	low	higher
Sensitivity to disturbances	low	low	somewhat higher
Solids removal efficiency			
with long fibers	good	very good	very good
with short fibers	good	moderate	good
with fillers	very good	moderate	moderate
Undissolved substances in clarified water, mg/L	10 – 100	10 – 300	10 – 50
Thickening efficiency	low	very high	high
High density stock concentration, %	1 – 2.5	8 – 30	1 – 15
Investment costs, specific	low	high	moderate
Operating costs	low	moderate	high
Energy consumption	very low	low	moderate
Chemicals consumption	low	low	higher

Sedimentation plants (rectangular scraper save-alls, round clarifiers, and cone save-alls) are suitable for the clarification of filler-loaded process water; the solids removal efficiency is greater than 90 %. However, the consistency of the stock is relatively low (1 – 2.5 %). A further disadvantage of sedimentation plants is the long retention time of the water to be clarified in the clarifying basins. This necessitates large volumes and therefore a large space requirement.

The flotation plants employed in the paper industry are either open or closed. Both full stream and partial stream aeration processes are used. In the former, the entire wastewater is aerated under pressure (up to 800 kPa), while in the latter, only a partial stream of either wastewater or circulated clarified water is aerated. Figure 80 shows an open flotation plant with partial stream clarified water aeration. The plant consists of a clarifying basin with sludge removal, aeration equipment for a partial stream of clarified water, and a dosing plant for the flotation chemicals.

Flocculation chemicals are added to the process water to be purified before it is fed into the clarifying basin. At the same time, compressed air is dissolved under pressure in a partial stream of clarified water. This partial stream is introduced into the stream of process water via an expansion valve, whereby a part of the dissolved air is degassed into very fine bubbles. The water, containing air bubbles and flocculation chemicals, enters into the clarifying basin, where the flocculation and flotation occur. The floated solids form a floating layer of sludge, which is removed by reamers. The consistency of the floatate is between 7 and 15 % and the efficiency of solids removal is often almost 100 %. Computers and flotation simulators have been used to obtain a low flocculant consumption and a stable operation with process water of varying solids content. Computerized flotation plants are described in [418], [419].

Filtration save-alls (disk filters and drum filters) have the lowest space requirement compared with other save-all systems. The retention time of the water to be clarified in the aggregates is very small. Filtration save-alls are mainly installed in paper mills where the grade of paper produced is frequently changed. They are also often used in mills with large, high-speed paper machines. Solids removal efficiencies of up to 90 – 95 % are obtained for fiber-containing process water. The efficiency decreases with increasing concentration of filler. Better efficiencies are attained by using fibrous filter aids (chemical pulp, groundwood, and paper stock), which form a layer of fibers on the screen that can retain the smaller filler particles and fines. Of all save-all systems, filtration produces stocks of the highest consistency (10 – 30 %). A disk filter that uses fibrous filter aids is shown in Figure 81.

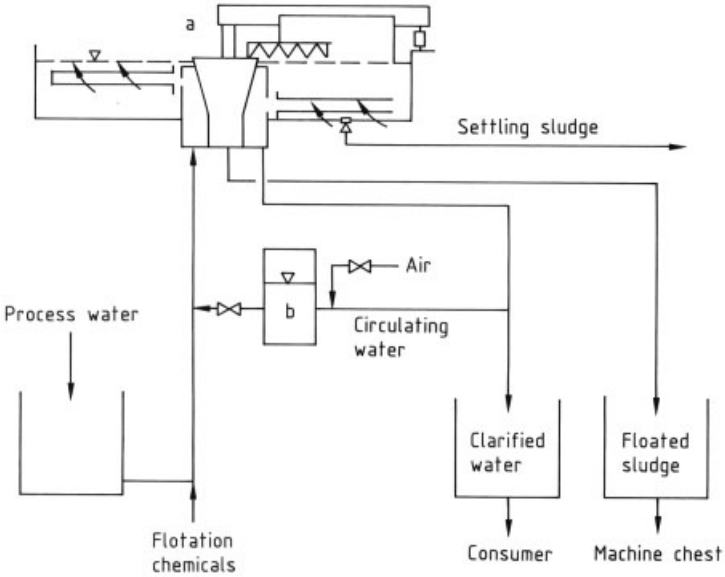

Figure 80. Scheme of a flotation plant with partial current clarified water aeration
a) Clarifying basin; b) Air saturation tank

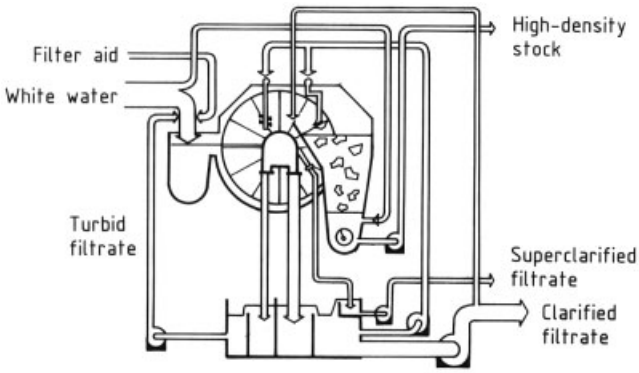

Figure 81. Scheme of a disk filter

8.6. Wastewater Pollution

The pollution in the wastewater of a paper mill depends on the type of raw material, the type and amount of fillers and additives applied, and on the degree of circuit closure. Effluents from mills that largely or exclusively process chemical pulps have a lower degree of pollution than those from mills in which mechanical pulp or wastepaper is employed. The use of starch or other organic additives results in a marked increase of the oxidizable effluent components, measured as BOD_5 or COD. Dyes and fillers can lead to discoloration and/or turbidity of effluents. Effluents from paper mills have a low content of nitrogen and phosphorus compounds. For this reason, these elements must be added as nutrient salts (e.g., urea and phosphoric acid) to feed microorganisms during treatment in biological purification plants. The closure of the water circuit within the mill results in an increase in the concentrations of the effluent components. On the other hand, the inorganic and organic load is reduced owing to its partial elimination via the paper produced. For this reason, the specific load in kg per tonne of paper (specific load = concentration × specific amount of effluent) is a more suitable parameter for quantifying the

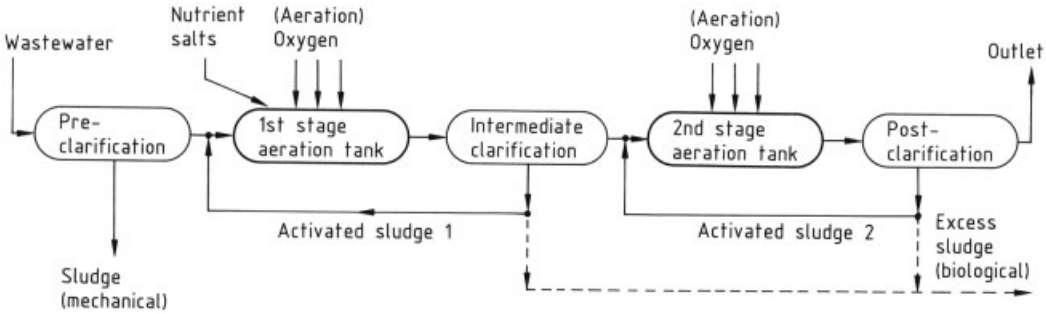

Figure 82. Scheme of biological effluent treatment with a two-stage activated sludge process [408]

effluent pollution. Information on the specific COD and BOD_5 loads, classified according to product groups, is given in [420].

8.7. Effluent Treatment Plants

Effluents from paper mills contain solids and dissolved substances. Solids (fibers, fillers) are mostly removed from the effluent in a chemomechanical clarification process by the use of flocculants. The degradation of dissolved organic substances is performed in aerobic and anaerobic biological treatment plants.

Chemomechanical clarification of effluents is carried out almost exclusively in sedimentation plants (round and rectangular basin with bottom sludge removal). Only in a few cases, is it necessary to neutralize the effluents. Rakes for the separation of coarse material and sand traps are seldom used. The clarifying efficiency of sedimentation plants is increased considerably by the use of flocculants. Undissolved substances are removed with an efficiency exceeding 90 %. Primarily activated sludge processes and, less often, trickling filter processes are employed for aerobic biological treatment. In North America and Northern Europe, effluent purification is frequently carried out in aerated oxidation ponds [421], [422]. Anaerobic treatment has also become established, especially in paper mills processing wastepaper.

The activated sludge processes applied are single-stage processes, systems with recycled sludge aeration, cascade systems, and two-stage processes. Atmospheric oxygen transfer is employed, using surface aerators (roll aerator, gyratory aerator) or pressure aerators. Pure-oxygen processes with pressure aeration are also used occasionally [423]. It is possible to improve the purification efficiency of activated sludge plants by using pulverized brown coal [424] or foamed plastic carriers [425] in the activated sludge tank. Biological treatment with a two-stage activated sludge process is shown schematically in Figure 82. The trickling filter system is usually operated in combination with an activated sludge process. The trickling filter can be either the first stage (in high-load operation) or the second stage (in low-load operation). Combination processes also include activated sludge processes with fixed-bed internals and submerged disk filters. Activated sludge plants and trickling filter plants are almost exclusively fed with mechanically preclarified wastewater. The addition of nitrogen and phosphorus compounds is required.

Anaerobic processes have been employed recently for the treatment of more highly polluted effluents (COD > 2000 mg/L). Above all, effluents from paper mills processing wastepaper are treated in anaerobic contact systems with sludge recycling and in USAB (upflow anaerobic sludge blanket) reactors. Information on plants installed up to now and on operational experience is given in [426], [427]. Fixed-bed reactors, fluidized-bed reactors, and other reactor concepts are still being tested. Anaerobic effluent purification is almost exclusively operated as a preliminary step to the aerobic treatment. The BOD_5 efficiency attainable with activated sludge processes is usually in the range 90 – 98 %, and the COD efficiency between 80 and 95 % [428].

For final clarification after the aeration tank, horizontal (rectangular and round basin) or vertical-flow (hopper basin) sedimentation

basins are preferred. Biofilters are occasionally employed as a further purification stage [429]. Third-stage purification processes for the elimination of nitrogen and phosphorus compounds are not required.

The sludges generated in effluent treatment must be dewatered and disposed of if they cannot be returned to the papermaking process. Screen belt presses, chamber presses, vacuum filters, or centrifuges are used for dewatering. The dewatering behavior can be improved by the addition of chemicals (polymers, trivalent metal salts, lime). Machines and plants for sludge treatment and their dewatering efficiencies are described in [430].

Most dewatered sludges (solid content 20 – 60 %) are disposed of in landfills. Thermal utilization (combustion) is becoming increasingly important. The sludges also can be composted together with bark, applied to agricultural land, and used as a porousing agent in the production of bricks, as additives for fiberboard, and in making cat litter.

9. Coating of Paper and Board

Paper, composed of individual fibers, has a rough surface and pores whose size corresponds approximately to the fiber dimensions. Since this roughness has a negative effect on printing, a large amount of paper and board is coated in order to impart a smooth surface for printing. Coating is a process by which a mixture of water, white pigments, binder, and various additives are applied to one or both sides of a paper sheet. The resulting covering and smoothening of the surface of the paper improves its glaze, gloss, brightness, and above all, its printability in all conventional printing processes.

Economic Importance. In spite of the higher costs of coated papers, the proportion of the total production of paper and board that is coated has increased steadily in the last 25 years because of the quality advantages described above. Current (1988) prognoses assume that this development will continue up to the year 2000 (see Table 19). Of the 57×10^6 t of paper produced in Europe in 1988, 12.7×10^6 t (22.3 %) were coated. However, as shown in Table 20, this percentage varies greatly from country to country depending on the structure of the paper industry in each case.

Table 19. Estimated development of the world consumption of paper and of coated grades between 1988 and 2000

	Consumption, 10^6 t		
Paper grade	1988	2000*	Increase, %/a
All grades	227	289	2
Coated grades			
Paper	20	32	
Board	12	15	
Total	32	47	3.2
% Coated	14.1	16.3	

*estimated

The largest producer of coated paper and board in Europe is Germany, followed by Finland and Italy. About 39 % of the 5.2×10^6 t of graphic paper produced in Germany is coated, ca. 50 % of which is wood-containing, and the remainder wood-free. About 45 % (1.4×10^6 t) of the board employed for packaging (folding boxboard, liquid packaging) is coated.

Table 21 shows that between 1986 and 1988 the proportion of coated papers increased continuously both in Germany and in the United States. It should be taken into account that in the same period of time, the export of coated papers from Europe into the United States also increased and has now reached ca. $600 – 800\times10^3$ t. The paper manufactures in Europe are concerned about the fact that many producers have already made or are planning considerable investments because of the favorable prognosis. The production capacity for wood-containing coated papers will increase by about 800×10^3 t and that for wood-free coated papers by 2.4×10^6 t from 1990 to 1992. It is questionable whether this drastically increased supply will meet with an equally high demand. The enormous capital required for these new investments – a new paper machine and coating machine cost about 255×10^6 € – has previously led to the closing down of companies, even in Europe.

Coated Paper and Board Grades. The classification of coated grades has not yet been standardized. However, certain terms are recognized throughout Europe.

Table 20. Coated production as a percentage of the total production in some Western European countries in 1988

Country	Total production, 10^6 t	Coated production, 10^3 t			
		Board	Paper	Total	% of total production
Western Europe	57.0	3 870	8 840	12 710	22.3
Germany	10.6	683	2 044	2 727	25.7
Sweden	8.2	592	476	1 068	13.0
France	6.3	281	1 193	1 474	23.4
Finland	8.7	602	1 621	2 223	25.6
Italy	5.4	589	1 279	1 868	34.6
United Kingdom	4.3	236	353	589	13.7

Art Paper. This term refers to wood-free or slightly wood-containing paper that is coated on both sides. The weight of the coating exceeds 20 g/m^2 per side, resulting in the complete disappearance of the fiber structure of the raw paper. These papers can be either dull or glossy. Up to three coats are applied to each side to obtain the high quality required. Art paper is used for high quality printed products.

Table 21. Coated papers as a percentage of the total graphic papers in Germany and the United States, 1986 – 1988 (in 10^3 t)

	1986	1987	1988
Germany			
Graphic papers	4565	4828	5185
Coated	1657	1786	2044
%	36.3	37.0	39.4
United States			
Graphic papers	22 666	23 897	24 950
Coated	5710	6230	6700
%	25.2	26.1	26.9

Chromopapers. Chromopaper is coated only on one side. However, the weight and quality of the coating correspond to that of art paper. Much of this paper is employed for packaging purposes.

Machine-Coated Papers. This term was previously used only for papers coated on-machine. Today, it is also applied to papers which are coated off-machine and have a similar coating weight and appropriate quality. It includes both wood-free and wood-containing papers with a coat weight of up to 18 g/m^2 per side. Higher weights are often obtained by applying a double coat. The coating formulations are normally simpler than those for art paper. The covering of the fiber structure depends on the particular coating weight and the roughness of the base paper. Machine-coated papers are used chiefly for printing.

Lightweight Coated (LWC) Papers. The English term is used internationally for this grade of paper. These papers are highly wood-containing and generally have coat weights in the range 5 – 12 g/m^2 per side. They are mass papers and are used for the production of catalogs, magazines, advertising pamphlets etc. The fiber structure can still be discerned in most cases. Either web offset or rotogravure printing is used for LWC paper. The total weight of LWC papers varies between 51 and about 71 g/m^2. From about 1982 the weight range has been extended below (ULWC or LLWC) and increasingly also above (MWC = medium-weight coated and HWC = high-weight coated) these limits. Today, the heavy grades are often provided with a double coat and are, consequently, grouped with machine-coated grades.

Cast coated papers are specialty paper and board grades. The mirror-like gloss is obtained by a process called cast coating, in which the still moist or remoistened coating surface is brought in contact with a highly polished cylinder and dried. Only one side is coated in most cases.

Folding Boxboard and Chromoboard. In these grades of board, the white topliner, which is usually made of pulp, is coated on the board machine. The remaining layers of board can consist of bleached or unbleached pulp or of wastepaper. However, the use of wastepaper for the topliner is also rapidly gaining acceptance. The coating weight is in the range 12 – 25 g/m^2 and the coat is normally comprised of a precoat and a topcoat. Three coating layers have also been applied. The board is rarely coated on the reverse and if it is, it is mainly to improve the flatness and to reduce

dust deposition. High-gloss board is produced by using a brush finishing system or a gloss calender.

9.1. Requirements of Coated Papers

The composition of the coating color is largely determined by the demands made on the surface of the paper by the particular printing process. Offset and letterpress printing require an especially well bound coat because considerable tensile forces are exerted by the highly viscous inks at the high printing speeds used today. At the same time, the coating must be sufficiently porous to permit fast, controlled removal of the printing ink vehicle without reducing the gloss of the ink. For offset printing, the surface of the coating must have not only an especially high water resistance, but also the ability to absorb sufficiently quickly the fountain solution applied in the printing machine. In general, coatings must have adequate stiffness, good brightness, low yellowing tendency, and good aging resistance.

In rotogravure printing, the tensile forces between the printing cylinder and the surface of the paper are very small because of the low viscosity, solvent-containing inks that are used. For this reason, the amount of binder required should only be sufficient to bind the pigment particles to each other and the paper surface so that picking of the coating is prevented during the further processing of the coated rotogravure paper. Important criteria for coated papers that are to be used for rotogravure printing are good compressibility, a high degree of smoothness, and adequate receptivity for the ink solvent.

Apart from the optical and printing properties, there are other criteria that are important for paper and board used for packing purposes. These include low odor, flexibility, glueability, and crease resistance. Coated specialty papers require specific properties, such as alkali resistance (labels), washability (wallpaper), and solvent resistance (silicone release paper).

9.2. Properties and Components of Coating Colors

Apart from water, the main components of a coating color are coating pigments and binders. In addition, a variety of auxiliary agents, such as dispersing agents for the pigments, additives to reduce wet abrasion, products for the control of viscosity and water retention, foam control agents, shading dyes, optical brighteners, and lubricants to improve paper calendering, are employed.

Coating colors, which are almost without exception all water-based, are fairly complicated multicomponent systems. Nowadays, paper is coated on high speed coating machines that are either a part of the paper machine or separate from it (on- or off-machine coaters). Coating colors are classified according to their solids content:

1) Low-concentration coating colors, solids content 35 – 45 %; Newtonian flow behavior; viscosity (Brookfield method) 100 – 500 m Pa · s (at 100 rpm, 20 °C); coating process or equipment: a) air knife, b) bar coater (for low coating weights)
2) High-concentration coating colors, solids content 55 – 70 %; occasionally Newtonian, mostly pseudoplastic-thixotropic flow behavior; viscosity (Brookfield) 600 – 2500 mPa · s (at 100 rpm, 20 °C); coating process or equipment: a) roll coater, b) blade coater, e.g., inverted blade, billblade, c) roller blade
3) Coating colors of moderate concentration, solids content 45 – 55 %, Newtonian to slightly pseudoplastic flow behavior at greatly varying viscosity; coating process or equipment: a) roller blade (for intermediate coating weights), b) roll coater (gate roll coater, size press), c) blade coater

The properties of the coating colors must be adapted to the coating process and the intended use of the coated paper. The *solids content* greatly influences the runnability of the coating color in the coating machine. It basically influences the viscosity of the color and its flow behavior at varying shear stresses in the coating machine. The various components of the coating color determine the solids content attainable at a given viscosity. The highest possible solids content is aimed for to save drying energy. Another important property of the coating color which greatly influences its workability is the pH. Most coatings are processed in the pH

range 7.5 – 11.5. The alkaline range is preferred because most of the color components have an anionic character and, from a colloidal chemical viewpoint, form considerably more stable systems at higher pH values than at low pH values. In addition, the drying time of the mostly oxidatively drying offset inks is greatly reduced with increasing pH.

For most coating processes, the water retention capacity of the coating color is also important; i.e., the ability of the coating color to retain water in spite of the sucking action of the coating base paper. If the water retention capacity of a coating color is too low, it may be too strongly dewatered between application and leveling.

9.2.1. Pigments

Pigments form the actual "body" of paper coatings. They usually account for at least 80 % of the total dry coating weight. Thus, pigments largely determine the quality and cost of the coating. White pigments are used almost exclusively in the coating plant, the most important being clay, ground and precipitated calcium carbonate, calcium sulfoaluminate (satin white), titanium dioxide, talc, and occasionally barium sulfate. Calcium sulfate (gypsum) has also been used in exceptional cases. For special applications, calcined and delaminated clays and synthetic pigments based on aluminum silicate are employed (cf. Table 16). The pigment most often used is coating clay. It is especially well suited to the production of a smooth coating surface because of its platelike structure. Furthermore, the large number of types makes it possible to take into account all the requirements of the coating and of the processing machines. Most of the coating clay used in Europe comes from England. However, North and South American clays are being used in European coating plants to an increasing extent. American clays are normally a little yellower than the English varieties and also have a different area-to-thickness ratio (shape factor). This, in turn, results in a lower viscosity and a somewhat lower water retention capacity of the coating color.

Apart from clay, finely ground grades of calcium carbonate have become particularly important, and the amounts used are now comparable to clay. Typical grades of calcium carbonate have a particle fineness of 90 – 95 % $< 2\,\mu$ and a brightness of 93 – 95 %. They permit the production of highly concentrated coating colors which can still give a very smooth and glossy coating. Calcium carbonate normally increases the microporosity and brightness of the coating, but the ground form can give lower opacity.

Satin white is still of some importance in paper coating. This pigment has very fine particles; it is extra white, and has a low density. It increases ink absorption and gloss. The disadvantages of satin white are its sensitivity to increases in temperature and decreases in pH, and its high adhesive demand.

In spite of the high price and limited availability of titanium dioxide, it is employed especially in the United States to compensate for the yellowish U.S. clay. As a result of its high refractive index, this pigment greatly increases the opacity of coated paper, even in the wet state. For this reason, it is used in Europe mainly for specialty papers (e.g., labels).

9.2.2. Dispersing Agents

The first stage in the production of coating colors is the conversion of the coating pigments, which are usually supplied in the dry state, into a flowable and pumpable aqueous slurry. This dispersion process consists of two steps: the mechanical dispersion of the coarse pigment agglomerates in an aqueous medium and the chemical stabilization of the slurry towards reagglomeration. The second step requires dispersing agents.

Suitable dispersing agents are polyanions, e.g., salts of polyphosphoric acids (pyrophosphates, hexametaphosphates) or poly(acrylic acids). These products permit the production of stable, highly concentrated pigment slurries that are still flowable. Coating colors with concentrations of 70 % and more can be produced. Today, ready-for-use pigment slurries are often available from pigment suppliers. The specific requirement of dispersing agents for the individual pigments varies greatly. English clays, for instance, require ca. 0.3 % and American clays ca. 0.1 % of dispersing agents. Calcium carbonates need 0.4 – 0.8 %, depending on the fineness and the pretreatment, and satin white requires 2 – 3 %. The protective colloid action

of organic dispersing agents stabilizes the rheological properties of the coating color, especially in the case of satin white.

9.2.3. Binders

The function of the binder or adhesive is to bind the pigment particles to each other and to fix the coat to the base paper. In spite of their low proportion by weight in the coating color (generally 5 – 25 % based on the pigment), binders have a large influence on the coating and processing properties of the coating color.

Binders from Natural Raw Materials. The natural binders all have hydrophilic properties. The most widely employed binders of this type are starch and its derivatives, cellulose ethers such as carboxymethyl cellulose, and in the United States, hydroxyethyl cellulose, soybean protein, and to a very small extent casein and alginates.

Natural binders act as protective colloids that prevent the flocculation of the pigments; they increase the viscosity and water retention of coating colors and give the coat a higher stiffness. In general, natural binders are used as additives for synthetic polymer binders.

Starch derivatives are intended to lend the coating color good rheological properties at high shearing forces. They should have a good pigment-binding power and ensure good adhesion of the coating to the paper. In order to meet these and other requirements, mixtures with other binders, and variation in type of derivatives are used. For example, oxidized starches are usually employed together with polymer dispersions. Hydrolyzed starches exhibit high stability in solution, good binding power, and good flow behavior. Hydrolyzed, esterified starches exhibit good stability in solution, high binding power, and increased reactivity towards wet-strength additives such as urea – formaldehyde resins. Hydrolyzed, etherified starches exhibit the same properties as the esterified derivatives. However, in contrast to the esterified derivatives, they can be used at pH values above 8 without the risk of saponification. Starch derivatives that contain phosphate and amino groups are compatible with cationogenic substances such as satin white. The phosphate groups react with multivalent metal ions, such as aluminum ions, which in turn leads to a certain degree of water resistance. The amino groups react with aldehydes, which in turn enhances the activity of wet-strength additives.

Cellulose Ethers. The sodium salt of carboxymethyl cellulose (CMC) is the principal cellulose ether used as a binder in papermaking. Pure CMC coats of 0.5 – 3 g/m^2 increase the grease resistance and printability of paper. Depending on the use, various mixtures of low- and high-viscosity CMC are employed. Pigment-containing CMC coats, which can be applied on the size press, contain up to 10 % semitechnical, low-viscosity CMC or salt-free, purified CMC. CMC is usually processed together with other natural or synthetic binders. Above all, CMC increases the water retention of the coating mixture. However, the water retention is so high that the addition of other binders that promote water retention in the coat to a lesser extent is necessary. In general, the amounts of CMC employed are 0.3 –1.5 % based on the pigment, and very low viscosity types are preferred. This gives an adequate coating color viscosity, even if the solids content of the coating color is high. Types of CMC that are soluble in cold water are rapidly becoming established because they do not require dissolving at elevated temperature. The presence of satin white pigment can cause strongly interfering coagulations in the coating color.

Casein must be present in the dissolved state. It is dissolved by the addition of alkali (e.g., ammonia, sodium hydroxide, borax, or sodium carbonate) either separately in a cooker (up to 70 °C) or with the pigment in a kneader. The casein concentration is limited to ca. 20 % by its high viscosity. The limit of processability is shifted to ca. 33 % by the addition of urea or dicyanodiamide, which reduces the viscosity and increases the storage stability of casein solutions. The mixing of casein solutions with pigments, especially China Clay, can cause a "shock" phenomenon (a rapid increase of viscosity).

Soybean protein has properties very similar to those of casein. Isolated soybean proteins are hydrolyzed, isoelectric proteins. They are used in the form of alpha and delta proteins with four different viscosities (extra low, low, medium, and high). The viscosity refers to the

dissolved soybean protein, the solvent of choice being aqueous ammonia (26 Bé). Proteins dissolved in this way exhibit very low sensitivity to water after drying. Like casein, soybean protein is mostly used as a mixture with polymer dispersions. This combination permits the preparation of coating colors with high solids content and a relatively low viscosity. Solids contents of ca. 60 %, suitable for blade coaters, can be achieved.

Synthetic Binders. In the initial phases of paper coating, natural binders were used exclusively. However, in the 1940s in the United States and in 1950s in Germany, the demand for synthetic binders in the form of polymer dispersions started increasing. These products made it possible for the first time to attain a high solids content at low viscosities, a prerequisite for modern high-speed coating machines. In addition, polymer dispersions give the coat a higher water resistance, better flexibility, higher gloss, and better printability. They are aqueous dispersions of polymers, usually stabilized with anionic emulsifiers. The solids content of the dispersions is generally 45 – 50 %. In most cases, the polymers are copolymers of several monomers, e.g., styrene, butadiene, acrylic ester, vinyl acetate, and acrylonitrile. Apart from these components – the so-called main monomers – small amounts of auxiliary monomers, such as acrylamide, acrylic acid, maleic acid, and methacrylamine are added to modify the dispersion properties.

The styrene – butadiene dispersions are most widely used worldwide. Products of varying film hardness are obtained, depending on the proportion of styrene used. The compatability of these highly hydrophobic polymers with the other coating components can be improved considerably by carboxylation, i.e., by incorporating small amounts of unsaturated carboxylic acids such as acrylic acid. Approximately equal proportions of styrene and butadiene give binders which provide a relatively soft film and a very good pigment binding capacity. However, the odor of the dispersion and the tendency of the the films to yellow when exposed to light are disadvantages.

Acrylate dispersions are of comparable importance in Europe. These are copolymers of butyl acrylate with styrene or vinyl acetate. The ratio of the soft component (butyl acrylate) to the hard one (styrene or vinyl acetate) determines the application characteristics of the dispersion. In general, acrylate dispersions have an excellent ageing resistance and are less odorous. Apart from these two most important groups of polymer dispersions, vinyl acetate homo- or copolymers have gained entry into paper coating plants, especially in the United States. These products generally have a lower binding power, but provide very hard and porous coats, and have an excellent ageing resistance. Polymer dispersions based on methacrylates and copolymers of vinyl acetate and ethylene are less important in paper coating.

The above-mentioned products are low-viscosity aqueous dispersions which usually do not influence properties such as viscosity and water retention that are important for the flow behavior of the coating. For this reason, only in a few cases are these "multipurpose dispersions" used alone. They are usually mixed with cobinders, which are responsible for adjusting the flow properties of the coating color. In general, these cobinders are naturally occurring products, such as starch, casein, CMC, and alginate.

Products based on acrylate esters have also been used as cobinders. Due to their high degree of carboxylation, these dispersions turn into colloidal dispersions upon addition of alkali. Thus the laborious and energy consuming dissolution and preperation processes involved in the use of natural binders are avoided. Another synthetic cobinder that has long been known is polyvinyl alcohol. Although its binding power exceeds that of all other binders used in paper coating, it has gained only limited acceptance. This is mainly due to the fact that the application of large amounts of polyvinyl alcohol leads to rheological problems on the coating machine during processing.

Apart from these two-component systems, synthetic sole binders which do not require a cobinder have been available since the early 1960s. On addition of alkali, these products develop the required viscosity and water retention, but retain their dispersion form. Special binders of this type have become popular, especially for illustration paper that is produced in large amounts and printed by the rotogravure or web offset process. Table 22 lists the various binder systems, their influence on the produc-

Table 22. Comparison of various binder systems

Binder system	Production	Rheological behavior	Coating properties
Polymer dispersion alone (rarely used)	simple mixing process	usually unfavorable	variable via type of dispersion; waterproof, glossy coat attainable; no activation of optical brighteners
Polymer dispersion + watersoluble cobinder, e.g., casein, starch, CMC (very frequent)	mixing and dissolving processes, sometimes separately with heating	largely controllable via cobinder	variable via single components; limited water resistance and gloss; full activation of optical brighteners
Polymer dispersion + synthetic cobinder (frequent)	integrated mixing/dissolving processes, no heating	controllable via cobinder	limited variability; waterproof and glossy coat attainable; full activation of optical brighteners
Polymer dispersion + viscosity regulating additives (in special cases)	integrated mixing/dissolving processes, no heating	limited controllability	variable only via type of dispersion; good water resistance and gloss, no activation of optical brighteners
Special dispersion (for publication papers)	integrated dissolving/mixing processes, no heating	slight controllability	limited variability; improved water resistance with additives; good gloss attainable; optical brighteners can be activated only via auxiliary components

tion of coating colors, the rheological properties of the coating color, and the coating properties.

9.2.4. Other Additives

In many coated papers and boards, the coating must be resistant to water. For instance, offset papers are exposed to the fountain solution in the printing machine, labels should not disintegrate when wet, and some wallpapers must be resistant to cleaning with detergents.

Good waterproofing of coatings can be achieved by using water-insoluble polymer dispersions. However, additional auxiliary agents are required to harden the coating since a water-soluble cobinder is used in most cases, which renders the coat sensitive to water to some extent. Three main classes of products are employed as hardening agents: melamine – formaldehyde resins, urea – formaldehyde resins and bifunctional aldehydes such as glyoxal, which is added only to coatings containing starch. The other two resins are used in various binder combinations. High proportions of natural binders require the addition of 10 – 15 % of hardening agents, based on the natural binder. For sole-binder coatings or coatings with only very small amounts of cobinder, 0.8 – 1 % is used, based on the pigment.

The waterproofing of the coating is due to a cross-linking reaction between the resin and the binder, with elimination of water. The rate and degree of cross-linking depend on the pH and the temperature. The desired effect increases with decreasing pH of the coating color and intensity of drying. The worldwide discussion on formaldehyde being a possible carcinogen has led to an intensified search for other cross-linking systems. For instance, products based on ammonium zirconium carbonate or epoxides are being used in some cases.

Foam-control agents (→ Foams and Foam Control) are another important class of additives. Most coating colors contain considerable amounts of products that act as foam producers or stabilizers (e.g., emulsifiers, protective colloids). Due to the constant motion of the coating color, air is easily introduced into it. The resulting foam causes an increase in viscosity and processing problems. However, the more serious problems are flaws in the finished coat such as craters and longitudinal streaks caused by the bursting of foam bubbles. The foaming tendency

decreases with increasing solids content and is, therefore of minor importance for modern high-solids coatings. The use of appropriate systems that minimize the ingress of air during preparation and circulation of coating color is the principal measure taken to prevent foam. However, in many cases this is not sufficient and chemical defoamers must be added.

A distinction is made between agents that prevent foam formation (foam inhibitors) and those that destroy foams (defoaming agents). Coating color defoamers must not only be easily emulsifiable, but should also be worked into the coating color under an appropriately high shear gradient. Otherwise, or if too much is used, agglomerates of defoamers are formed in the coating and these in turn lead to spots of low surface energy where printing ink will not adhere (fish eyes). In practice, the amounts of foam inhibitors used vary from 0.02 to 0.05 %, based on the pigment. Defoamers are usually diluted with water and added to the coating mixture dropwise. Amounts ranging from 0.01 to 0.05 vol % are usually sufficient.

In many cases, the yellowish color of the coating clay does not satisfy market requirements. Apart from the addition of pigments such as titanium dioxide which increase brightness, blue-violet dyes (*shading dyes*) are often used at a concentration of about 2 g per 100 kg of pigment to attain a bluish white coating surface which appears brighter to the eye. Not only basic dyes, but also the considerably more lightfast pigments may be used. If the brightness required is very high, *optical brighteners* are added in addition (→ Optical Brighteners). In order to be effective, most optical brighteners need a substance to which they can be attached. Substances of this type are polyvinyl alcohol, carboxymethyl cellulose, casein, starch, and some synthetic cobinders. Polymer dispersions do not have this property.

So-called *lubricants* such as calcium or ammonium stearate are used in many illustration papers to reduce dusting on the calender and produce a higher paper gloss.

9.2.5. Coating Color Formulations

The coating color formulations presented below are examples of formulations commonly used in Europe at present. However, they must usually be adapted to suit the particular conditions in the individual mills.

LCW Rotogravure Paper. A coating color based on a special sole binder is usually used for this paper, which is produced on very large machines ($150-250\times10^3$ t/a). Composition (in parts):

Clay	100
Dispersing agent	0.1 – 0.3
Sole binder	4.8 – 5.5
Sodium hydroxide solution added until pH 8.5 is reached	
Calcium (ammonium) stearate	0.7
Solids content	45 – 60 %
Viscosity (Broofield; for blade coater)	400 – 2500 mPa · s at 100 rpm

LWC web offset paper (in parts):

Pigment (clay + $CaCO_3$)	100
Dispersing agent (as required)	
Multipurpose dispersion	9 – 12
Natural or synthetic cobinder	0.5 – 4
Hardening agent	0.5 – 1
Calcium (ammonium) stearate	up to 1
Solids content (for blade application)	58 – 70 %

This is an example of a large number of different compositions. Today, great importance is attached in many cases to the stiffness of the paper. Consequently, the binders should form very stiff films.

Sheet offset paper (in parts):

Pigment (clay + $CaCO_3$)	100
Dispersing agent (as required)	
Multipurpose dispersion (see above)	10 – 14
Natural or synthetic cobinder	0.5 – 4
Hardening agent	0.5 – 1
Solids content (for blade application)	57 – 70 %

A wide variety of coating formulations are available.

Board (in parts):

Pigment (clay + $CaCO_3$)	100
Dispersing agent (as required)	
Multipurpose dispersion, e.g., based on acrylate	12 – 20
Natural or synthetic cobinder	1 – 6
Hardening agent	1

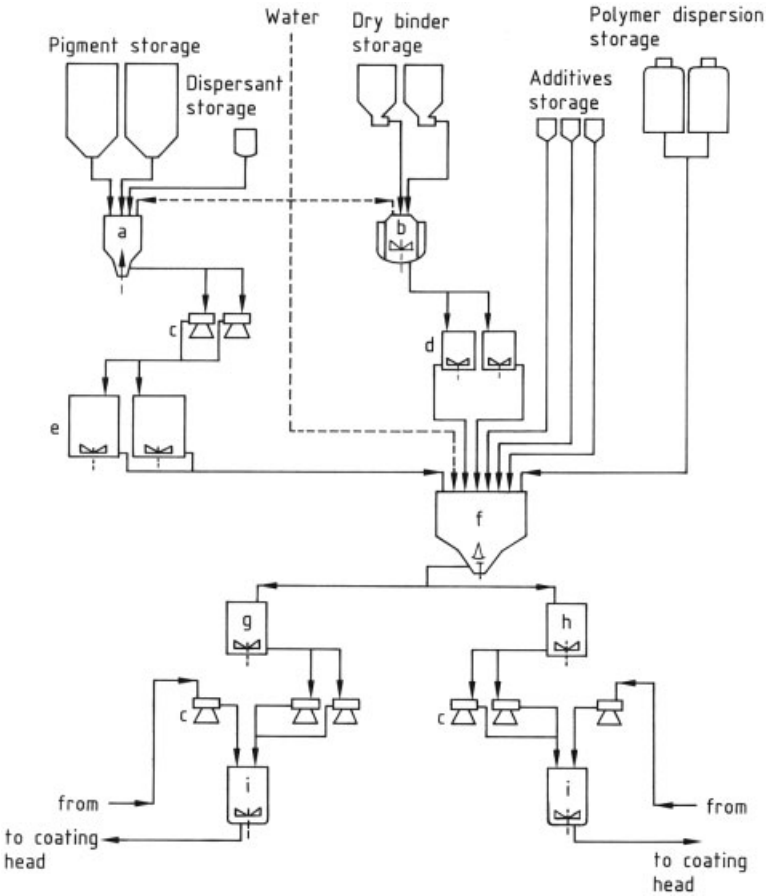

Figure 83. Production of coating colors
a) Disperger; b) Dissolving vessel; c) Filters; d) Vessels for dissolved binder; e) Containers for pigment slurry; f) Coating mixer; g) Reservoir for coating 1; h) Reservoir for coating 2; i) Working vessels

Starch is used only rarely in the topcoat. CMC or a synthetic acrylate based cobinder is preferred for board.

9.3. Coating Color Preparation

A flow sheet for the batch production of coating colors is shown in Figure 83. These, or similar plants that permit adaptation to a very variable coating program, are used in practice in by far the most cases.

Water, dispersing agent and, when required, alkali are fed to the dispersing machine and the dry pigments are then added. While kneaders were used previously for dispersion, high-speed mixers with rotor – stator systems have now gained acceptance. These turbomixers permit the production of a completely dispersed pigment slurry in 5 – 60 min, depending on the amount of pigment. The next step is either the intermediate storage of the slurry or the immediate preparation of the coating color in the turbomixer, depending on the size of the plant. Dispersion requires the highest rate of rotation, but the speed is reduced when binding and auxiliary agents are added. Natural binders such as protein or starch are dissolved in advance by heating in a separate vessel.

When adding the binding and auxiliary agents, the amount of water required to give the desired final solids content is added to the pigment slurry first. The natural cobinders are then added with constant stirring, followed by

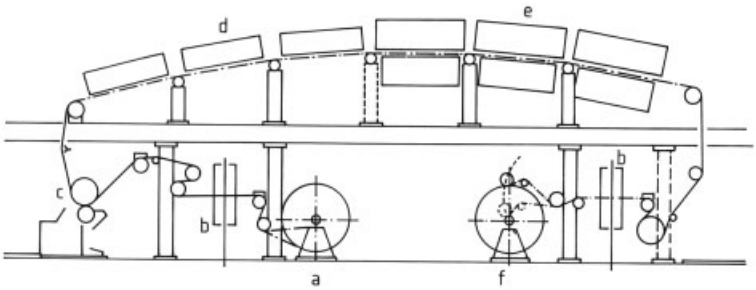

Figure 84. Schematic of coating plant
a) Unwind station; b) Moisture and coating weight measurement; c) Coating station; d) Infrared dryer; e) Hot-air dryer; f) Rewind station

the polymer dispersion, and finally the auxiliary agents, such as hardeners, shading dyes, optical brighteners, and lubricants. This is followed by the final adjustment of the pH, viscosity, and solids content, and pumping the mixture into the storage vessel.

The pigment slurry and also the finished coating color are screened at least once before they enter the coating machine. The use of a screen in the coating color return flow is also recommended to remove particles torn off the paper. Pressure or disk filters or vibrating screens are used for the filtration of the coating color. The mesh size of the screens is 5000 – 10 000 apertures/cm^2. Coarser screens with mesh sizes of 2500 – 3500 apertures/cm^2 are required for clay. Apart from batch plants, continuous plants also exist and they have the advantage of requiring less space, and the amount and composition of coating color can be more easily changed. The precondition for a continuous plant is that all the components of the coating color are present in the liquid state. They are then transferred to the dispersion apparatus by dosing pumps. After passing through the shearing zone the finished coating color leaves the closed system.

9.4. Coating Machines

The first coating machines imitated the process of painting in that the coating was leveled with oscillating brushes. The brushes were then replaced by distributor rollers, and finally by systems such as the air knife or trailing blade. Modern coaters (see Fig. 84) consist of the unwind station, coating head, a dryer, and the rewind station. Paper may be coated either on equipment that is an integral part of the paper machine (on-machine) or on separate coating equipment (off-machine). A compromise is frequently made between the two types, the coater being installed directly after the paper machine (in-line). In this case, the coater can be used when required.

The most important part of the coating machine is the coating application unit. It has three functions: transferring the coating color onto the web, metering of the excess, and leveling of the remaining coating. The most important coating equipment is listed in Table 23.

The air knife is widely used in Europe. Metering and leveling are achieved with a thin jet of air. This contact-free doctoring off permits the use of almost all coating raw materials and an almost unsurpassed range of coat weights. The air brush is also used as an intermediate coating aggregate for triple-coated board.

The coating equipment most often used for on-machine coating is the size press. It is used almost exclusively for the precoat, and the application weight is relatively limited.

The blade coating processes have rapidly gained tremendous importance due to their simple operating principle, which permits the attainment of speeds exceeding 1400 m/min. The blade is usually made of flexible steel of 0.3 – 0.4 mm thickness. It is used to remove excess coating color from the paper. Especially in the case of the inverted blade, coating weights of 4 – 26 g/m^2 can be achieved. Blade coatings have a particularly high smoothness. This system has the advantage of requiring a lower drying energy because coating colors of high solids

Table 23. Survey of the most widely used coating equipment

	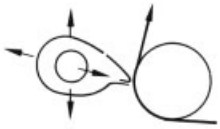 Air knife	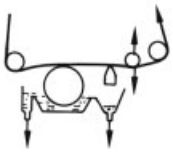 Bar coater	Size press	Film press
Usual mode of operation	separate and board machine	board machine	paper and board machine	paper machine
Application	one-sided	one-sided	two-sided	two-sided
Use for multilayer coatings	topcoat, intermediate coat	precoat, topcoat	precoat	
Working speed (max.), m/min	500	300	800	400
Coating weight (per side), g/m^2	10–30	4–10	4–8	8–15
Solids content of coating, %	35–50	40–50	25–45	50–60
Viscosity (Brookfield, 100 rpm), mPa·s	100–300	100–300	100–300	1000–3000
Types of paper usually coated	Art paper, chromopaper, specialty paper, wallpaper, packaging paper, board	board	graphic	graphic

	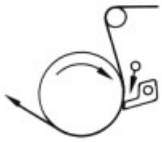 Trailing blade	Bill blade	Inverted blade	SDTA	Roller blade
Usual mode of operation	separate	paper machine	separate and paper machine	separate and paper machine	board machine
Application	one-sided	two-sided	one-sided	one-sided	one-sided
Use for multilayer coatings	topcoat	precoat, topcoat	precoat, topcoat	precoat, topcoat	topcoat
Working speed (max.), m/min	1200	1000	1500		1000
Coating weight (per side), g/m^2	6–15	6–12	4–25	4–25	6–12
Solids content of coating, %	55–62	50–65	45–70	45–70	50–65
Viscosity (Brookfield, 100 rpm), mPa·s	500–1500	300–1000	500–2000	400–1000	500–1500
Types of paper usually coated	graphic, LWC	graphic	graphic, LWC, art paper, specialty paper, board	LWC, board	board

content (70 %) can be processed in blade coating plants. Apart from the usual application with a roll, the short dwell or short dwell time applicator (SDTA) system has been used increasingly in the past years. In the SDTA, the coating color is in a chamber under pressure directly in front of the blade. The distance between coating and blade, which is about 80 cm in normal roll applicator coaters, is reducd to 3 – 30 cm. Thus, the time in which the coating color is in contact with the base paper before blade metering is reduced substantially. A relatively new system used in

paper machines is the bill blade, which simultaneously coats both sides of the paper, and the roller blade, which has a rotating bar instead of the rigid blade placed on the running paper web.

Drying. The coat must be dried after leveling in the coater. The drying facilities vary greatly. The festoon drying system, in which the web was dried on slowly moving rods in a long tunnel, is now obsolete. Hot-air hoods and drying cylinders are used on the paper machine. The latter are often coated with teflon or a similar substance to prevent the sticking of the moist coating. In off-machine coaters hot air tunnels are often used, through which the paper web is transported. The paper is supported by rolls when only one side is coated and by a hot air cushion when both sides are coated. Air foils are a modern version of this process, involving air-cushion drying with the paper web being supported and dried on both sides by hot air rapidly streaming out of narrow jets. In addition, infrared dryers are usually used in combination with the above mentioned drying facilities. These IR dryers immediately fix the coating after it leaves the coater, preventing migration of binder in the paper.

The drying facilities not only determine to a large extent the coating speed and the maximum application weight, but they also greatly influence the quality of the coat.

10. Paper Dyeing [431–440]

Colored papers are obtained by dyeing the paper stock or the paper surface (size press, paper coating, printing). Optically brightened papers can be produced in the same manner. Dip dyeing is used only for special purposes. Anionic direct, cationic direct, and basic dyes are the most important dyes employed for paper coloring, and are now frequently available as liquid formulations. Acid dyes are used in individual cases and pigments only in special grades.

10.1. Dyeing Processes

Stock dyeing is the most widely used type of paper dyeing. Dyes, pigments, and optical brighteners are added either batchwise in the pulper or mixing chest or introduced continuously into the stock flow.

The choice of dye and the fixing and dyeing conditions largely depend on the raw material used in papermaking (recycled fibers, groundwood, unbleached pulp, bleached pulp, CTMP, TMP) and on the pH of the dyeing system. In continuous dyeing, the point of addition is also determined by these two factors (low consistency dyeing at a stock consistency of ca. 0.8 % and high consistency dyeing at a stock consistency of 3 – 4 %). In general, there is a trend towards dyeing processes in the neutral pH range. The higher the degree of beating of the pulp, the deeper the coloring.

Fillers increase the dye requirement because they absorb dyes and, at the same time, reduce the coloration owing to their covering power and brightness. The addition of aluminum sulfate usually promotes the absorption of dyes and yields less colored effluent. The two stock dyeing processes have the following advantages and disadvantages.

Batch addition has the advantage of thorough mixing of all additives with the paper stock, optimal fixation due to longer contact time between fiber and dye, as well as easy handling. The disadvantage is that the time required for color correction and color change is long.

Continuous addition has the advantage of a shorter zone in the stock line that must be cleaned when the color is changed, and of less broke because the desired shade is attained more quickly. However, a lower color yield (low contact time) is obtained for intensely colored papers, and the more complex equipment required for this dyeing process must also be taken into consideration. For example, the shade of a production is controlled by continuous color measurements in the paper machine, and the feeding pumps can be adjusted automatically to carry out the required corrections.

Surface Coloring in the Size Press. In this process, dissolved dyes are added to the size press liquor. Optical brighteners and other additives have long been applied in this way. An essential prerequisite for uniform dyeing is adequate and, above all, uniform absorbency of the paper.

The advantages of this process are: quick changeover of shades, absence of dyes in the wa-

ter circuit, and, in the case of papers with higher basis weights ($> 80\,g/m^2$), the saving of dye. Nevertheless, compared with stock dyeing – the classical dyeing process – surface dyeing has gained acceptance only in individual cases because uniform dyeing of the paper is difficult to achieve. It is occasionally advisable to combine the stock and surface dyeing processes, e.g., to correct two-sidedness. The bleed fastness of surface dyed papers is generally lower, compared to internal addition.

Dip Dyeing. A small group of specialty papers, called effect papers (flower crepe paper, tissue paper), is noted for its intensive and brilliant shades. The paper is passed through a dipping bath containing an aqueous solution of the dye. The excess dye liquor is pressed off between two rolls and the wet paper is creped, if required, before drying. Acid dyes are usually used because they have high solubility and bright shades. The low affinity of these dyes for plant fibers results in uniform dyeing, even in the case of papers with greatly varying fiber composition. The bleed fastness of dip dyed paper is poor, corresponding to that of surface-colored paper.

Surface Coloring by Coating. In normal coating, the surface of the paper or board is covered with a pigment coating (see Chap. 9). In the case of colored coatings, the starting material is the white coating mixture, and the desired shade is attained by adding a dispersion of an organic or inorganic pigment. Water-soluble dyes can be used only to a limited extent because, even with the use of fixatives, they cannot be rendered bleed fast in the coat (danger of migration into the base paper). Also, the inadequate lightfastness compared to pigments limits the use of these dyes.

10.2. Fastness, Dyes, Pigments

Depending on the intended purpose of the paper, the following types of fastness are required of paper dyeing [432]:

Lightfastness is required for all kinds of fine paper, especially document paper, photographic paper, laminate base paper, and wood-free writing and printing paper. Organic and inorganic pigments and direct dyes are suitable for this purpose.

Bleed fastness is required for paper and board that are used for food-packaging and tissue papers. Organic and inorganic pigments and direct dyes are suitable for this purpose.

Rub fastness is required for album and wrapping paper and for cover board. Direct dyes are suitable for this purpose.

Acid fastness is required for parchment and vulcanized fiber base paper, writing, and printing paper. Organic and inorganic pigments, selected representatives of all colorant groups, are suitable for this purpose.

Solvent resistance is required for labeling paper (packaging of perfumes, medicines, and spirits). Special pigments are suitable for this purpose.

Heat resistance is required for cable and core paper and laminate paper. Organic and inorganic pigments are suitable for this purpose.

Basic Dyes (Cationic Dyes). The affinity of basic dyes is extremely good on lignin-containing fibrous materials (e.g., groundwood) and unbleached pulps, which are negatively charged. Salt formation gives rise to very stable color lakes that are insoluble in water.

Anionic fixatives, e.g., Mesitol (Bayer) or Tamol (BASF), also form color lakes of this type with cationic dyes, and can therefore be used to improve the fixation of basic dyes on bleached pulps that contain only small amounts of lignin. Better dye fixation always results in richer colors, improved bleed fastness, and less effluent staining.

Advantages: brilliant and tinctorially strong dyes, which result in good fastness to water and steam on groundwood-containing furnishes.

Disadvantages: poor substantivity on bleached furnishes, mottling tendency (i.e., parts of the paper stock or stock mixture are dyed more deeply than others), and poor lightfastness.

Applications: low-quality paper, packing paper, kraft paper, cartridge and cheap wrapping paper, dip dyeing, and surface coloring.

Direct dyes, compared with basic dyes, have a high affinity for wood-free, bleached pulps. Cationic fixatives are not required to increase the dye yield of direct dyes, unless deep shades with

good bleed fastness are required (e.g., deeply colored napkins).

Advantages: high affinity for wood-free bleached stocks. The lightfastness is often superior to that obtained with basic dyes.

Disadvantages: direct dyes are less suitable for the dyeing of unbleached furnishes (groundwood) because the shades obtained are not as brilliant as those produced by basic and acid dyes. Mottling tendency is occasionally observed with stock mixtures (cellulose – groundwood).

Applications: text and cover paper, cable paper, quality writing paper, bookkeeping and carbon paper, punch and filing cards, all unsized papers such as blotting paper and hygienic papers (e.g., serviettes, handkerchieves, toilet paper).

Cationic direct dyes are moderately absorbed on bleached and wood-containing stocks, and even in the case of deep shades, alum or a fixative is not required. These dyes produce fairly good bleed fastness. Combination dyeing with anionic direct dyes can be carried out to reduce the contamination of the wastewater and to improve bleed fastness.

Acid dyes have little affinity for natural plant fibers. Although the dye molecules penetrate well into the capillaries of the fibers, they must be fixed with aluminum sulfate (pH ca. 4.5) and cationic fixing agents, e.g., Levogen (Bayer) or Tinofix (Ciba-Geigy), in order to achieve a satisfactory dye yield.

Advantages: good solubility, no tendency to mottle on stock mixtures, and very well suited to dip dyeing.

Disadvantages: no affinity for plant fibers, suitable only for sized papers when aluminum sulfate and fixatives are added, colored effluent, and poor bleed fastness.

Applications: certain writing and printing papers, decorative crepe paper, and carbon paper.

The importance of acid dyes for pulp dyeing has decreased tremendously because they pose wastewater problems.

Pigments. Inorganic pigments (iron oxides and chromium oxides), organic pigments (e.g., azo and phthalocyanine types), and carbon black are used for paper dyeing. The retention of pigments on paper fibers is based on the following phenomena:

1) Filtration, i.e., the fibers on the wire of the paper machine act like a filter. The retaining power of the stock is directly proportional to the degree of beating to which it has been subjected.
2) Absorption (van der Waals forces and electrostatic effects).
3) Occlusion, i.e., the pigment particles are enveloped by other substances (colloidal components of the paper stock, aluminum hydroxide hydrate, resin precipitate).
4) Flocculation, i.e., improvement of retention by the addition of agents such as cationic retention or wet-strength additives, which bind the pigment to the fibers.

Advantages: excellent fastness and no risk of mottling.

Disadvantages: no affinity for fibrous materials, two-sidedness in the case of pulp dyeing (the wire side of the paper is less strongly colored), and low tinctorial strength.

Applications: lightfast writing and printing papers, bookkeeping and label papers, document and laminate base papers.

Blueing. This type of paper dyeing basically consists of counteracting the slight yellow tinge of all paper stocks by adding "violet". Pure blue and brilliant red dyes, which together give violet, or violet dyes are used for blueing.

The highest degree of whiteness is obtained with *fluorescent whitening agents*, which, like direct dyes, are absorbed directly on the fiber. This method is mainly applied to paper stock and to paper surfaces (size press, coating). The presence of optical brighteners in paper can be detected by the characteristic fluorescence under a UV lamp.

Dyeing of Specialty Papers.

Veined paper is made from a mixture of bleed fast dyed fibers (cellulose, cotton fibers, wool, or synthetic fibers) and pulp that is either dyed differently or undyed.

Effect papers (see the Section on page 121)

Laminate base papers for the production of high quality laminate materials, e.g., for the furniture industry, are obtained by pigment dyeing.

Bronze-glazed paper : the color bearing component of the coating color is aluminum or

brass powder, which produces a silver or gold effect or, in combination with dyes, a metallic effect.

Security papers are treated with dyes or chemicals that change color under the influence of reducing or oxidizing agents, alkaly, or acids and thus make forgeries more difficult.

11. Types of Paper, Cardboard, and Board

Paper is a product made from plant fibers that have been mechanically or chemically digested. The fibers are bonded together in an aqueous suspension and processed into sheets with the addition of auxiliary agents.

Accordingly, paper is defined as a planar material made essentially from fibers that are mostly of plant origin. It is formed by draining a fiber suspension through a sieve. The resulting fiber web is subsequently compacted and dried to give a basis weight of < 225 g/m^2 [441], [442].

There are approximately 3000 kinds of paper products [443]. These can be divided into four main groups:

1) Graphic papers
2) Packaging papers
3) Hygienic papers
4) Specialty papers

This classification has also been largely adopted by the Statistisches Bundesamt (Federal Statistical Office) of Germany in Chapter 48 of the "commodity index for foreign trade statistics" [444]. The basis of this systematology is a "harmonized system for the marking and coding of merchandise" of the "council for tariff cooperation".

11.1. Graphic Papers

The term graphic papers refers to a large range of different papers that are suitable for printing.

Printing and Press Papers [445]. Printing and press papers are divided into the following four groups:

1) Newsprint
2) Wood-containing natural printing papers
3) Wood-free natural printing papers
4) Coated printing papers

Newsprint represents an important subgroup. It consists primarily of a mixture of primary and secondary fibers. The primary fibers include groundwood and small amounts of pulp, which are used to increase strength. "Improved" newsprint, which has a high filler content and is calendered (closed surface), also belongs to this group. As a result of the good runnability in the printing machine, the main field of application is newspaper production. Other uses are, e.g., paperbacks, calculating-machine rolls, telex rolls, and interleafing paper.

Newsprint has a basis weight of 40 – 56 g/m^2. It is machine-finished or supercalendered. Standard newsprint should have an ISO brightness of 57 – 60 %.

Improved newsprint is used, e.g., for book and offset printing. The ISO brightness should not be less than 63 %.

Wood-Containing Natural Printing Papers. This is another important group of printing and press papers. The term wood-containing indicates that these printing papers consist mainly of groundwood along with pulp, while "natural" refers to the fact that they are not coated. Fillers, e.g., kaolin or calcium carbonate, are used to produce high smoothness and good opacity. These natural papers include gravure printing paper, also called SC paper (supercalendered).

Natural magazine papers have high lignin content, and are filled and highly supercalendered. Their main use is the production of magazines and other products of the press.

Gravure papers (SC gravure) have the same properties as natural magazine paper, but are specially suited to gravure printing. ISO brightness > 67 %.

Newsprint, improved for gravure printing has the same properties as natural magazine paper; ISO brightness < 67 %.

SC offset paper has the same properties as natural magazine paper. The ISO brightness (> 67 %) is optimized for offset printing.

Wood-containing thin printing papers have a basis weight of ≤ 45 g/m^2. They are used

for telephone, railroad timetable, and address books.

Book papers are used in the production of books. They have a groundwood content of 5–20 %. The filler content depends on the desired volume.

Voluminous book papers are produced with $\geq$1.4-fold volume.

Supercalendered Writing Papers. These woodcontaining natural papers are sized and supercalendered. They can also be made from wastepaper.

Machine-finished offset papers are supplied in sheets or rolls.

Wood-Containing Coating Base Papers and Boards. Coating base papers are papers that are intended to be further processed into coated papers or boards.

Wood-Free Natural Printing Papers. Wood-free natural printing papers consist almost exclusively of pulp. The proportion of fibers from mechanical pulp should not exceed 5 %. Wood-free printing papers do not yellow to any appreciable extent and have a high strength. This is the main difference between these papers and the wood-containing grades described above. The papers are uncoated and normally leave the paper machine in a machine-finished state. These papers are calendered on a supercalender for certain printing processes. This group of papers includes:

Wood-Free, Machine-Finished Natural Printing Papers. Basis weight:

Group 1	$\leq 80\,g/m^2$
Group 2	$81-149\,g/m^2$
Group 3	$> 150\,g/m^2$

Wood-free, Supercalendered Natural Printing Papers. The classification of the basis weights is as described for machine-finished papers.

Letterpress printing papers are voluminous and are particularly suitable for books (volume $\geq$ 1.4 fold).

Wood-free Bible and thin printing papers can be rag-containing or rag-free. They are highly filled to give good opacity. The basis weight is in the range below 40 g/m^2 to above 50 g/m^2.

Coating Base Papers for Printing Papers and Boards. Coating base papers and boards are materials that are intended to be further processed into coated papers or boards.

Coated printing papers represent the largest group within the printing and press papers. In the case of these printing papers, the base paper is finished by the application of a layer of coating color to both sides of the paper web, which imparts a closed surface to the printing papers (see Chap. 9). These papers are supercalendered and supplied with a dull or glossy surface and are described as wood-containing or wood-free, depending on the raw materials used. The wood-containing, coated printing papers represent the largest part and the LWC papers (light weight coated), in turn, make up the main part of this subgroup. Examples include:

Wood-Containing, Two-Sided Coated Papers. The minimum coating application is 5 g/m^2 per side. Main areas of application: magazines, catalogs, and job printing.

Wood-containing LWC paper is coated on both sides; basis weight $\leq$72 g/m^2. It is suitable for gravure printing and roller offset printing.

Wood-containing consumer papers, coated on both sides must have a minimum groundwood content of 50 %.

Wood-Free, Two-Sided Coated Papers. The coating application is 5 g/m^2 per side and the fields of application are the same as for the corresponding wood-containing papers.

Wood-free LWC paper is coated on both sides and has a basis weight of $\leq$72 g/m^2.

Art papers (picture printing papers) have a minimum coat application of 20 g/m^2 per side and are usually wood-free letterpress printing papers with a surface suitable for multicolor printing and printing from blocks with fine screens.

Pigmented papers have a coat application of less than 5 g/m^2 per side. Wood-containing or wood-free coating base paper can be used for their manufacture.

11.2. Office and Administration Papers

These papers offer a wide spectrum for very different areas of use, e.g., writing papers, papers for data printing, papers for forms, copy

and duplicating papers, envelope paper, postcard board, self printing papers, refined papers for offset printing, blueprint papers, work drawing paper, bank notes, documents, and map paper. These printing and writing papers and the corresponding base papers are the papers that are used predominantly in offices, administration, and schools. Examples are:

Data papers and boards are used for the manual, mechanical, electronic, and magnetic recording of data and its input and output.

Bookkeeping transparent paper is highly beaten and is therefore transparent.

Account book paper or paper for bookkeeping machines is wood-free, highly supercalendered, and fully sized.

Punched tape paper is vegetable parchment or other specialty papers for use in tape instruments.

Perforated tape board, magnetic tape board is wood-free kraft board, also semibleached, with high dimensional stability and rigidity.

Optical character recognition must comply with the requirements in DIN 6723 and 6724 [446], [447], which guarantees that the exchange of machine-readable vouchers between companies, e.g., in automated data exchange, proceeds without problems. The papers are uncoated.

Paper for continuous forms is wood-containing or wood-free, usually machine-finished reeled paper, including recycled paper.

Index board is wood-containing or wood-free, supercalendered, single-layered board that can also contain wastepaper.

Writing papers are used almost exclusively for manual writing. They are supercalendered and can also be made with varying amounts of rags. This group also includes handmade papers, moldmade papers, bond, and air mail papers.

School writing paper is a machine-finished, highly sized paper with a basis weight of 60 – 80 g/m^2.

Typewriting, copy, duplicating, and photocopy papers are wood-containing or wood-free papers that may also be made from wastepaper. They are machine finished or supercalendered, sometimes voluminous, and have a basis weight of up to 100 g/m^2. They can also be used as transfer paper or for copying.

Air mail paper is light-weight, wood-free, often rag-containing paper with high opacity.

Carbon copy paper is well sized, has a basis weight between 23 and 30 g/m^2 and is used in the copying of letters and manuscripts.

Carbonless copy papers contain color forming reagents, which produce a contour-true copy when subjected to pressure (exerted on underlying sheets).

Correspondence envelope paper can be made from a wide range of stocks, from wastepaper to pulp. This paper is used exclusively in the production of envelopes and mailing bags.

Kraft paper is used to make envelopes and mailing bags and is made from unbleached or bleached kraft pulp.

Envelope lining tissue is an intensely colored, sometimes dip dyed, wood-free tissue paper with a basis weight of 22 – 30 g/m^2. It is used for the lining of expensive envelopes.

Base paper for diazotype is wood-free, has a basis weight of 55 to $>$ 200 g/m^2 and is used in the making of blueprint papers.

Copy base papers are wood-free. As a result of special properties, they can be finished to give various coated copy papers, e.g., zinc oxide paper, thermocopy, and photocopy papers.

Base papers for offset films are predominantly wood-free papers that are used for making paper offset plates.

Carbonizing base papers are carrier papers for a wax-bearing color mass. They are used for the production of carbon and blue papers.

One-time carbon (OTC) base paper is used in the production of one-time carbon paper.

Multiple carbon base papers are used in the production of multiple carbon or blue papers.

Base papers for nonimpact printing processes are predominantly wood-free and are specially suitable for nonimpact printing processes (thermosensitive, electrosensitive, thermostatic papers).

Blotting papers are wood-containing or wood-free, white or colored, and sometimes have veined fibers. They possess a very high absorbency and wettability.

Security, banknote, and archival papers are wood-free and/or rag-containing or linters-containing. These heavy duty, nonaging papers are suitable for color printing and can be made resistant to forgery, e.g., by means of a watermark.

Map and nautical chart papers are wood-free, dimensionally stable, and are used for the printing of maps and sea charts.

Letter board is wood-containing or wood-free and may contain wastepaper. It has a basis weight of >150 g/m^2 and a supercalendered surface.

Picture postcard board is wood-containing or wood-free and is used in the production of picture postcards.

Letter File Cardboard. This solid fiber board used in the production of files is multilayered, vat-lined or unlined and has a basis weight of 1000 – 2000 g/m^2. Special properties: exactly planar, and precisely maintained moisture level.

Filing board is a special, tough board which is usually colored and supercalendered.

Ticket board is wood-containing or wood-free and often colored. It is used, e.g., for streetcar tickets, entrance and weighing tickets. Ticket board is produced in accordance with regulations issued by the railroad authorities.

Fancy cardboard can be pasted or unpasted and is often wood-free, e.g., ivory, Bristol, and opaline board. It is used in the production of visiting cards, invitations etc.

11.3. Paper, Board, and Cardboard for Packaging

A large number of grades of paper, board, and cardboard are available for packaging and are employed for many different purposes. These include:

Folding Boxboard. This product is ideally suited for the production of folding boxes because of its properties, such as good scorability, creasing ability, groovability, and printability. This board can be coated or uncoated; the former gives a better printed format.

Machine-made board (also solid fiber board) is produced almost exclusively from wastepaper and is processed primarily into mailing boxes. Pressed rolled board is also used for this purpose.

A great variety of papers are used for packaging. These range from simple WP wrapping paper (made from mixed wastepaper) to kraft paper, which is made predominantly in rolls from various pulps (sulfite/kraft pulp) or from mixtures of pulp and wastepaper. These grades are employed in the production of bags, beer mats, carrier bags, gift wrapping paper, etc.

For the foodstuff sector, special types of paper and board are produced which must fulfil the requirements of the food laws (see also → Foods, 4. Food Packaging). These papers are also widely used in the nonfood sector, e.g., as coating papers. The possibilities of combination with each other and with other materials (plastics) are numerous. In this way, special packing materials can be produced for most packaging.

Greaseproof papers are used for the packaging of butter, margarine, meat, and sausage. These papers include vegetable parchment and glassyne, which were provided with barrier polymers.

Depending on their intended purpose (for the smooth surface or the corrugated inner layer of corrugated board), corrugated papers are produced with varying strength, rigidity, wet strength, and printability. The normal designations, which are all used for the production of corrugated board, are kraft liner, test liner, covering stock, and corrugated stock.

For most of the papers, boards, and cardboards described above, the stock is adjusted to meet the requirements in each case. Examples are:

Kraft Papers. In order to exhibit the high strength properties required, these papers consist of primary or primary and secondary fibers of kraft pulp. They are used predominantly for packaging purposes.

Kraft sack paper, bleached and unbleached, includes kraft bag asphalt base paper which is used in the production of asphalt paper.

Extensible kraft sack paper (clupack) corresponds to the kraft sack papers, but it has a higher extension (> 6 %) in the longitudinal direction. It can also be used as base paper for plastic coating.

Secondary kraft sack paper is made from at least 50 % unbleached kraft primary fibers.

Kraft packing papers, slightly bleached, unbleached, or bleached must contain 80 % of primary fibers.

Kraft tissue, bleached or unbleached, has a basis weight of < 30 g/m^2. It is a one-sided smooth or fluted kraft paper that is employed for wrapping glass or metal goods; it is also suitable for packing fruit, e.g., oranges.

Sulfite Wrapping Paper (ZP). The stock composition of this paper is on average 70 % of primary fiber pulp and not more than 30 % of fibers from sulfite or kraft wastepaper. The basis weight is $>30\,g/m^2$.

ZP 1:	made from knotter pulp, 30 % secondary fibers permitted
ZP 2:	inferior sulfite pulp, up to 30 % higher quality secondary fibers permitted
ZP 3:	100 % inferior sulfite pulp, up to 30 % groundwood or higher quality secondary fibers permitted
ZP 4:	65 % sulfite pulp, up to 35 % groundwood permitted
ZP 5:	made from 100 % unbleached or bleached sulfite pulp

WP packing papers consist of mixed wastepapers of varying composition; they may also contain pulp.

Crepe packing material is a highly sized and creped packing paper of any stock composition.

Mixed soda pulp-paper is made predominantly from secondary fibers of kraft pulp.

Glassyne is a highly supercalendered paper made from strongly beaten pulp. It is largely greaseproof, but not water resistant. The high transparency is achieved by intensive supercalendering. It is frequently colored. This paper is used, e.g., as wrapping paper for chocolates, covering paper in photograph albums, for dust jackets, envelope windows, and for packing high-quality preserves.

Supercalendered base paper for waxing is designed to be waxed and frequently contains titanium dioxide as filler. The waxed paper is often used as wrapping paper for candies.

Silicon base paper is a supercalendered paper to be used for siliconization.

Greaseproof paper is made resistant to grease primarily by using a special beating process (high fibrillation) and by the addition of auxiliary agents. In comparison with vegetable parchment, this paper is neither waterproof nor boil-proof.

Parchment base paper is unsized, highly absorptive and is made of 100 % pulp fibers. It contains no chemical auxiliary agents and is employed in the production of parchment paper.

Vegetable parchment [448] is a highly waterproof, greaseproof, boilproof paper product of high purity. It is used in the packing of fats, also in combination with aluminum foil or with a plastic coating. Silicon-coated parchment paper has a wide range of applications, e.g., baking paper, antiadhesive papers, etc.

Coated papers are often coated on one side and are suitable for offset and gravure printing. They can be painted, bronzed, and punched. They are waterproof and alkali resistant to prevent the detachment of labels in washing machines. They are used, e.g., as wrapping paper for chocolates, cigarettes, soap, and as dust jackets for books and records.

Coated Packaging Papers. These bleached kraft papers have a one-sided dull or glossy coat and are suited to the production of bags and carrier bags.

Corrugated Papers. These papers include:

Kraft Liner. This covering paper is made of at least 80 % kraft pulp, bleached or unbleached, and has a bursting strength per basis weight of $3.43\,kPa \cdot m^2\,g^{-1}$.

Kraft-Faced Liner. This covering paper does not have a specified stock composition, but it has guaranteed strength properties.

Two-layer paper is a double-layered covering paper.

Semichemical paper is corrugated paper made of at least 65 % semichemical pulp.

Fluting Medium Paper. The corrugated paper used for the production of corrugated board has no specified stock composition; however, definite strength properties are required.

Board and cardboard are planar materials consisting mainly of fibers of plant origin. In terms of basis weight, these materials can be included in both the paper sector and the board sector (250 to $>600\,g/m^2$).

1) Folding Boxboard. The following board types are included in this group:
   Coated pulp board is made from bleached pulp.
   Cast coated pulp board is used for high-grade packaging and for folding boxes, fine pasteboard boxes, paperback covers, etc.
   Chromo board, chromo duplex board, chromo triplex board have very different stock compositions, depending on the intended purpose.
   Artificial Chromo Board, Triplex Board. A gray core can also be used in this case.
2) Machine-Made Boards. This group includes the following board types:

Mechanical pulp board has a light or gray core. It is used, e.g., in the production of paper plates and round cheese boxes.
Brown mechanical pulp board is made predominantly from wastepaper and has a brown colored cover and reverse. It is used for the production of mailing boxes.
Chipboard consists predominantly of wastepaper and is employed for the production of mailing boxes, box lining, etc.
Pasted board consists of two or more layers of mill board that have been pasted together in the dry state.
Sheet-lined board is made from mill board. It may also be pasted, is lined on one or both sides with paper, and is used for high-grade packaging.
Mill board is made by winding one or several fiber webs on a press roll. Thus, high basis weights and the production of especially stable packaging material are possible.
Cartridge papers and boards are usually made from wastepaper and are used in the production of cartridges, special cartridges, cardboard tubes, and containers such as detergent drums.
Board for cups is a single layered hard board consisting of bleached pulp and is used for the production of drinking cups, etc.

11.4. Hygienic Papers

The major component of all hygienic papers is tissue. Tissue is a high-grade, fine paper with a closed structure. Tissue is usually used in multiple layers because of its low basis weight. Cellulose wadding and crepe paper also belong to this group.

Cellulose wadding is a loosly structured, highly absorbent fiber product. Its uses include the production of sanitary towels, surgical artificial cotton, patients' sheets, and packing cotton.

Soft tissue is made entirely or predominantly of pulp fibers, with fine, soft creping and closed formation. It is especially absorbent and is employed in the production of toilet paper, handkerchiefs, kitchen towels, hand and face towels, serviettes, and for covering diapers and sanitary towels. The basis weight is <27 g/m^2.

Crepe paper is a single-layered, highly absorbent hygienic paper made from groundwood and/or secondary fibers from wastepaper. The basis weight is >27 g/m^2. The most important uses are toilet paper, paper handkerchiefs, and cleaning rolls.

11.5. Papers and Boards for Technical and Specialty Uses

This group of papers and boards covers a wide range of grades. The most important subgroups comprise base papers for decor, priming and barrier foils, body, coating and soakage base papers and base papers for wallpaper, papers for electrotechnical purposes, drawing paper, photographic base paper, cigarette paper, filters and filter layers and special cardboards for the various branches of industry, e.g., presspahn, car body and shoe cardboard. All these papers and boards have a common requirement; they must exactly fulfil defined quality criteria. This group includes:

Chromatography papers are porous, absorbent special papers made from linters without the addition of sizing or filling agents. They are often acid washed and contain silicon compounds to render them hydrophobic.

Reagent and indicator papers are high-grade special papers used for impregnation with test reagents and indicator dyes.

Household filter papers are wood-free special papers that give a defined rate of filtration. They must comply with food laws. They are used, e.g., for coffee filters. The rate of filtration and the separating ability depend on the degree of fiber beating.

Analytical and industrial filter papers are usually made from cotton, dissolving pulps, and glass microfibers. They possess a uniform stock distribution and, thus, a homogeneous pore size. Analytical filters are acid washed and are made without paper making additives. Filter layers with a basis weight of up to ca. 750 g/m^2 are produced to withstand special stress, e.g., in the beverage industry. Filter layers of this type, that are

also used for sterilization filtration, must comply with food laws. For industrial use, the paper and filter layers may also be impregnated with melamine – urea resins or similar compounds.

Filter mass is made from soft-cooked, highly bleached pulps or from rags or linters and is used in the form of flocks for industrial filtration.

Cable and condenser paper (capacitor tissue paper) is stiff, nonporous paper made from pulp and is usually impregnated with artificial resins. These papers contain neither fillers, metal impurities, nor salts. The absence of pores is achieved by fine beating of the pulp fibers and by supercalendering. Cable paper must exhibit high strength in the longitudinal direction. The thickness of condenser paper is 0.006 – 0.012 mm and the basis weight is 6 – 7 g/m^2. These papers must meet IEC regulations [449].

Presspahn is a single-layered, highly rigid product that is made from pure pulp and couched in the wet state. When used in the electrical industry, the pulp must be produced with a high degree of electrochemical purity.

Building Papers and Boards.

Roofing Felt Base. This wool – felt board (reclaimed wool) is impregnated with tar/bitumen and used to produce roofing sheets.

Felt and wool – felt board consist of waste-paper and rags. They are used as the base material for textile floor coverings and as insulating material in various fields, predominantly in the automobile industry. They are also employed as separating layers on flat roofs that are covered with a plastic foil.

Gypsum liner board is used for the production of gypsum plates (building plates).

Covering Papers and Cardboards. This general term describes various papers and cardboards that are used as protective coverings. These single-layered or multilayered materials can be made from various raw materials. They are used, e.g., as covering papers in road construction.

Substrate Papers, Saturating Base Papers, and Coating Base Papers.

Substrate papers are used in many different technical systems as manufacturing aids. They are frequently removed from the finished product at the end of the production process, rewound and reused in the same process.

Coating base papers must be moisture resistant and stable to organic solvents. One or both sides of the paper are treated with various solutions or dispersions, which represent the requirement for the desired coating.

Saturating base papers must be absorbent and have wet strength and often also resistance to organic solvents.

Raw Material for Artificial Leather. This tough kraft paper is not dyed and is usually moisture resistant. It forms the web base for the production of artificial leather.

Base Paper of Vulcanized Fiber. This absorbent, voluminous paper is made from cotton rags, cotton linters, and refined pulps.

Photographic Base Paper. A suitable base paper for the production of photographic paper is a moisture resistant, dimensionally stable, chemically neutral pulp paper that is free from traces of iron or copper. Today, papers coated with a thin polyethylene foil on both sides have replaced barite paper. This paper is used for the uptake of light-sensitive emulsions.

Saturating base paper (overlay paper) is made from special pulp and has basis weights of 15 – 40 g/m^2. It is highly porous, free of dirt, and protects the underlying decor in the production of laminated products.

Decor base paper is a soakage base paper made from bleached or unbleached kraft pulp; it is used for resin absorption in the production of laminated plates.

Base papers for abrasive papers are tough, sized papers made from kraft pulp and are coated with an abrasive.

Calender bowl paper is strong, compressible, heat-resistant paper used for the manufacture of calender bowls.

Technical Drawing Paper. This group includes the following:

Opaque drawing papers are fine look-through papers made from pulp and have good sizing and hardness. Exacting demands are made on the erasability.

Translucent Drawing Paper. A distinction is made between natural transparent paper and paper made transparent by chemical treatment. Natural transparent papers are made from pulp, linters, or mixtures of the two by intensive beating of the fibers. They are well sized and exhibit good surface smoothness and look-through. Opaque papers are rendered transparent by soaking them in a solution of synthetic resin.

Sketching papers are wood-containing or wood-free, machine finished, sized natural papers with a basis weight $\geq 80\,g/m^2$.

Art drawing papers and board are suited to special drawing and painting techniques, e.g., watercolor paper. The latter has a rough surface which accepts water colors but does not allow their penetration. This material also has good erasability.

Base Paper and Board for Wallpaper. This is the general name for papers and boards that are suitable for the production of wallpaper [450]. They can be single- or multilayered, couched or pasted, wood-containing or wood-free, coated or uncoated. They may contain wastepaper or have a woodchip coat. Strip-off wallpapers or preglued wallpapers can also be produced.

Papers for Tobacco Products.

Imitation cork paper is a brownish yellow colored paper used for the production of cigarette tips. It has an imitation cork imprint and joins the filter to the tobacco stick.

Filter encasing paper is used for the production of acetate and/or cellulose filters.

Cigarette paper has a high filler content and a basis weight of $18-28\,g/m^2$. To control the smoking properties, this paper has a porosity that is suited to the type of tobacco and contains additives that regulate burning.

Cigar or cigarillo casing paper holds the chopped tobacco together and serves as the inner casing.

Long-fiber paper is used for tea bags and artificial sausage skin and is made from fibers having an average length >3 mm. The addition of synthetic fibers is permitted.

Binding papers are tough, wood-free papers that are frequently embossed.

Cover paper and board is tough, elastic, durable, wood-free white paper or board that has a linen finish and is specially suited to the production of covers for brochures, paperbacks or similar products.

Bookbinding board is press rolled or mill cardboard that is pasted or unpasted and has high rigidity, dimensional stability, and perfect lining ability.

Car body cardboard is a hard cardboard used predominantly as shaped parts in automobile manufacture; it is usually capable of being cupped.

Leather fiberboard is an especially tough hard cardboard that is flexible and resistant to moisture.

Suitcase board is an embossed hard cardboard used in the production of luggage.

Jacquard cardboard is a tough, dimensionally stable cardboard used in the production of perforated cards for the control of textile machines.

Flong. This coated cardboard (solid board) is made from very high-grade stocks (rags). It is processed with fillers and is completely knot-free and can be subjected to hot embossing. It is used to produce the matrices for high-pressure dry stereotyping.

Cardboard for beer mats is a highly absorbent groundwood cardboard that may be pulp covered. It can be punched, embossed, and printed.

Base papers for chart paper have a basis weight of $30-130\,g/m^2$ and are usually filled and sized. They have a refined surface and good dimensional stability.

Gardener, Flower, and Decoration Crepe. *Gardener's crepe* is a white or colored, highly expandable paper that has longitudinal grooves and serves as decorative flowerpot wrappers.

Flower crepe paper is wood-free and is used in the making of artificial flowers.

Decoration crepe paper is flameproofed by impregnation with chemicals and is used for decorations of various types.

Board for playing cards is made of two or three layers pasted together. It is coated on both sides, has a supercalendered surface, and is sometimes linenized. The basis weight is 280 – 340 g/m^2.

Album Board. Board for photograph, stamp, and other albums is often wood-free and voluminous.

Hot setting paper and hot sealing paper are coated on one side, become sticky on heating and can consequently be fixed to another material.

Insect resistant or insecticide paper is resistant to insects. It can kill insects by means of volatile components or by contact.

Antitarnish Paper. This impregnated or coated paper protects other materials against corrosion by releasing substances in the vapor phase or by direct contact.

Photographic protective wrapping (black photo paper) is lightproof, black-dyed, chemically neutral paper that can be folded without breaking and is suitable for wrapping photographic film.

Sterilization paper is a wood-free paper that, even when coated, can tolerate sterilization processes and is used for packing sterilized materials.

12. Book and Paper Preservation

12.1. Introduction

Most modern paper lacks the durability and permanence that has characterized its use since its earliest application to chronicle the history of man and his civilization. Although fundamental papermaking techniques have changed little in nearly 2000 years, two 19th century innovations, namely: 1) the use of alum as a component of size and 2) the replacement of cotton fiber by wood pulp, are responsible for this predicament. Papermaker's alum readily forms sulfuric acid on exposure to water in air; this acid in turn readily catalyzes the depolymerization of cellulose. Further, cellulose from wood pulp is more susceptible to this attack than linen- or cotton-derived fiber. The inherent acidity of such papers accounts for 85 – 90 % of the destruction in book papers, resulting in a half-life for fold endurance of an estimated 7.5 years [451].

One of the first to detail the problem and correctly identify the cause was William Barrow at the Virginia State Library [452] beginning in the 1930s. Prior to his work, book and paper conservators relied as much on art as science. While highly skilled, these practitioners often selected materials and treatment without adequate consideration of the long-term consequences.

Today mass deacidification of books and archives materials has become a commercial reality. The Wei T'o process, which employs methyl magnesium carbonate in a nonaqueous liquid-phase process, operates at the National Archives of Canada, while a variation on it treats books for the Bibliotheque Nationale in France. The Diethyl Zinc process invented by the Library of Congress now treats library materials in a United States facility owned by Akzo. There are four additional processes in various stages of development, including two that are marketed for their ability to also impart strength, that are of general interest to the preservation community.

12.2. Paper and Deacidification Chemistry

Paper nearly universally involves cellulose, although its source and methods of processing have evolved steadily throughout the centuries. As production costs dropped significantly, the application of paper dramatically expanded so that its traditional applications, particularly its use in books and official records, were dwarfed by other functions. Such widespread use of paper reduced the emphasis on permanence.

The sizing of paper continues to be the predominant cause of paper deterioration. In the early 1800s, about the time of ROBERT's invention of the fourdrinier machine for continuous paper manufacture, gelatin sizing was abandoned in favor of alum or alum-rosin blends. Papermaker's alum, $Al_2(SO_4)_3 \cdot 18\,H_2O$, is readily hydrolyzed by water to form sulfuric acid. Acid-catalyzed hydrolytic attack causes degradation of cellulose by random scission of the hemiacetal links, leading to breakage of chain linkages (Fig. 85). Even a few scissions per molecule cause a substantial loss in physical properties [453].

In 1851, the discovery of the soda process for separating cellulose from wood added a further dimension to the problem since paper derived from wood pulp is more susceptible to acid attack than the cotton cellulose obtained from linen and rags.

The lack of paper permanence is attributable to both internal and external influences [454]. In addition to sizing materials, internal factors include the type of fiber, coatings, and the presence of acidic and metallic compounds. External variables are the conditions during storage and use. Heat and humidity accelerate the deterioration of paper, and atmospheric pollution is frequently the source of external acid attack. A comparison of identical volumes stored in the New York Public Library and the Royal Library in The Hague attest to the importance of proper storage as the New York books were found to be in a far more advanced stage of deterioration [455].

It wasn't until 1926, when GOSTA HALL of Sweden developed a satisfactory method for accelerated aging [456], that researchers had a means for simulating the effects of the natural aging process. Although EDWIN SUTERMEISTER of the S. D. Warren Company recognized as early as 1901 that stable paper should contain an alkaline filler, it was left to WILLIAM BARROW of the Virginia State Library to pull the disparate elements of early researchers together in a scientific analysis of the problem [457].

As acid-catalyzed hydrolysis of cellulose is responsible for up to 90 % of paper's loss of permanence, researchers have concentrated on deacidification as the primary solution to the impermanence of modern papers. Paper deacidification involves introduction of a strong base to convert acid species to their corresponding neutral salts, and the establishment of a neutral buffer to protect against future acid-catalyzed damage. The first recorded attempt to neutralize paper was advanced by ARTHUR CHURCH at the Victoria and Albert Museum in London who in 1891 advocated immersion of paper in a methanolic solution of barium hydroxide [458].

Subsequent attempts to develop effective deacidification focused on aqueous techniques, which offered the additional benefit of modest strengthening through reestablishment of hydrogen bonding. OTTO SCHIERHOLZ, working for the Ontario Research Foundation in Toronto, secured a patent in 1936 for his work which involved dipping paper in a solution of alkaline earth bicarbonates [459]. In 1943 BARROW described a two-step process, involving immersion in a calcium hydroxide bath and subsequent treatment with calcium bicarbonate. Commentators differ as to whether BARROW was aware of the earlier Canadian work.

Further refinement of the technique included the 1957 development by JAMES GEAR et al. at the United States National Archives of a one-step process involving magnesium bicarbonate, and the use of the more soluble magnesium hydroxide by scientists at the Library of Congress in 1978. While effective and proven, these techniques involve painstaking care as they are applied to single sheets.

12.3. Development of Mass Deacidification Processes

Mass preservation is designed to meet the needs of national and major research libraries and archive groups. With a mean collection deterioration rate of 4.7 % [460] and limited resources, a treatment based on painstaking manual work is inconsistent with the need. Further, emphasis has focused on nonaqueous and gas phase deacidification techniques to offset some of the intrinsic limitations of aqueous methods.

Nonaqueous deacidification systems contain a deacidification agent and an organic solvent that serves as the carrier. The solvents offer the advantage of rapid penetration and drying. The latter is particularly important in minimizing the obvious appearance of treatment, most notably cockling. Liquid techniques can function eco-

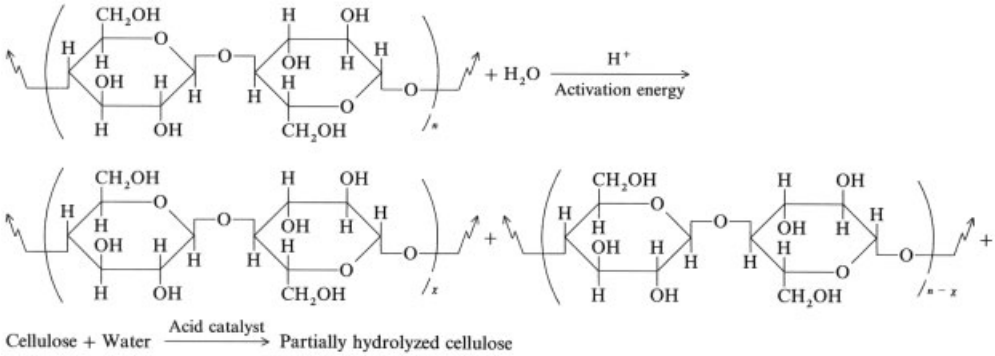

Figure 85. Acid-catalyzed hydrolysis of cellulose

nomically over a broad treatment range. The late 19th century approach of CHURCH was updated by A. D. BAYNES-COPE in the mid-1960s [461]. This followed the 1961 experimentation by BAYNES-COPE with a magnesium methoxide solution which, while quite effective, was impractical in terms of preparation and storage. Independently, RICHARD SMITH at the University of Chicago developed a process that utilized a magnesium methoxide compound solubilized in a mixture of methanol and chlorofluorocarbons (CFC) [462]. This overcame many of the problems associated with earlier approaches regarding stability and ease of use.

In contrast, a vapor-phase approach offers prospects for minimal preselection and maintenance of document integrity because no solvents are involved. At least one researcher likened this approach to a "fountain of youth" among deacidification methods [463]. Gas-phase processes work best at substantial scale because the facilities characteristically have high capital intensity.

Early attempts at gaseous deacidification included work reported in 1962 by Y. P. KATHPALIA at the Nehru Library in New Delhi, involving exposure of books to high concentrations of ammonia [464]. The compound proved to be too weak a base to completely neutralize strong acid and the ammonia buffer was highly impermanent. At about the same time, BARROW working in the United States and BAYNES-COPE in England combined ammonia and short-chain amines, including butylamine, diethylamine and piperidine, but observed similar results – a short term alkalization that was reversed as the amine left the treated paper.

A more promising approach was advanced by WILLIAM LANGWELL who placed cyclohexylamine carbamate impregnated sheets in books [465]. The active compound diffuses into the book paper to neutralize acidity, but no alkaline reserve is established and the neutralization is reversible. Finally, morpholine treatment was promoted by KUSTERER and others working at the Barrow Laboratory in 1977. Besides the lack of permanence, significant discoloration was observed in the paper of treated books.

GEORGE KELLY and JOHN WILLIAMS at the Library of Congress designed a mass deacidification process in 1977 that utilizes diethylzinc (DEZ) [466]. The compound easily diffuses between the pages of closed books, neutralizing the acidity and establishing a uniform, controllable reserve of zinc oxide.

12.4. Commercial Mass Deacidification Processes

Today mass deacidification is a routine undertaking with prospects for substantial use by libraries and archives. The national libraries of France, England, Austria, and the United States have developed preservation processes, and the private sector has advanced additional techniques. Table 24 describes the processes that are under development to treat the broad range of materials that exist in library and archive collections. Three of the processes use chlorofluorocarbons as the solvent or as part of the solvent

Table 24. Mass deacidification processes

	BPA	DEZ	Wei T'o	Bookkeeper	British Library	Lithco
Developer	BPA	Library of Congress	Wei T'o	Koppers	British Library	Lithco
Type	gas	gas	liquid	liquid	liquid	liquid
Neutralizing agent	amine complex	diethylzinc	methoxy Mg methyl carbonate	MgO	amine complex	magnesium butoxytriglycolate
Carrier	none	none	Freon	Freon	none	Freon
Cycle time, h	24	60	24	5	< 15	3 – 5
Preselection ∗	minimal	minimal	some	minimal	minimal	minimal
pH of treated paper ∗	6.5 – 9.0	7.0 – 7.5	8.0 – 9.5	8.0 – 9.0	7.0 – 9.0	8.0 – 9.5
Alkaline reserve, % ∗	up to 1.0	1.0 – 2.0	0.7 – 0.8	up to 2.0	up to 1.0	1.5 – 2.5
Commercial status	yes	yes	yes	pilot	pilot	pilot
Strengthening ∗	some	no	no	no	yes	yes

∗ Most process results have minimal third party verification.

system. Additional development work is underway based on 1) operation in a closed system to preclude loss of CFC's and/or 2) operation with CFC replacements and subsequent paper testing.

BPA Process. Book Preservation Associates (BPA) was formed in 1988 as a wholly owned subsidiary of ETO Sterilization, a processor of spices, medical products, and cosmetics. Books or other library materials are placed in unopened boxes in a vacuum chamber. The pressure is reduced and the books are exposed to ammonia for 6 h. The chamber is then purged and reevacuated. Ethylene oxide is introduced and the materials are held for 12 h at nearly atmospheric pressure. The chamber is alternately purged and evacuated up to 25 times before treated materials are removed. The total cycle time is 24 h [467]. Books are placed in an aeration room for two days to permit off-gasing of the ethylene oxide before being returned to libraries.

The process was designed to create high molecular mass primary, secondary, and tertiary amines within the cellulose matrix of the book paper. The in situ generation of amines, together with free ethylene oxide, neutralizes acidity and establishes an alkaline reserve of ca. 1 %. Ethanolamine generation during treatment is described as follows. (see right column)

Developers of the process report that the mono-, di-, and triethanolamines are formed in approximately equal amounts, with a modest dominance by the monoethanolamine. The reaction is dependent upon the ratio between reactants and is adjusted intentionally.

BPA personnel report that the pH drops approximately 1.5 points in the two months following treatment. As the company has readily available capacity for deacidification, commercial activity has been initiated.

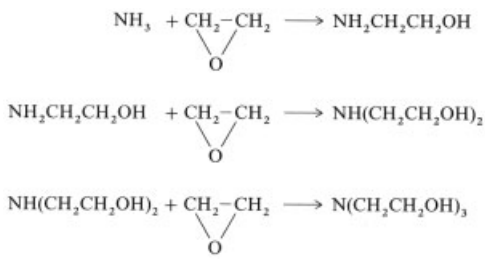

The DEZ Process. The DEZ Process was developed by the Library of Congress over more than 15 years, progressing from simple laboratory experiments to an operating small-scale commercial unit. Much of the latter work was in conjunction with Texas Alkyls, one of three major producers of compounds such as diethylzinc (DEZ). Akzo secured an exclusive license for the DEZ Process and is presently commercializing the technology.

As strong alkylating agents, metal alkyls react readily with acids to form the corresponding neutral salts. DEZ also reacts with residual water in the paper to form a stable, permanent buffer of zinc oxide. Unlike most other processes, the neutralization of existing acidity is independent of the formation of the alkaline reserve. Thus both highly acidic and neutral paper are both

provided with an optimum alkaline reserve. Key reactions are summarized below [468]:

$$H_2SO_4 + (C_2H_5)_2Zn \longrightarrow ZnSO_4 + 2\,C_2H_6$$

$$2\,CH_3CO_2H + (C_2H_5)_2Zn \longrightarrow Zn(CH_3CO_2)_2 + 2\,C_2H_6$$

$$H_2O + (C_2H_5)_2Zn \longrightarrow ZnO + 2\,C_2H_6$$

The buffer established by the DEZ process is more nearly neutral (saturated solution pH of 7.3 – 7.5) than that of magnesium or calcium salts used in other deacidification methods.

The process has three steps: pretreatment, permeation, and postconditioning; the total cycle time is 55 – 60 h. Akzo operates a treatment facility near Houston, Texas with capacity for ca. 40 000 books per year. Owing to its high reactivity, liquid DEZ is dangerous if mishandled. Poor design and operating procedures on the part of an earlier contractor threatened to halt further development of the process because of the inoperative first pilot plant. The issue of safe handling was fully resolved in the second facility.

The Wei T'o Process. Wei T'o Associates, developers of a nonaqueous liquid deacidification process, derives its name from an ancient Chinese god regarded as a protector of books. The National Archives of Canada operates a treatment facility in Ottawa with capacity for ca. 40 000 books per year. The process is based on work conducted by Richard D. Smith, a graduate student at the University of Chicago in the mid-1960s. In 1989, Union Carbide acquired exclusive rights to the process and intends to market the technique as part of a paper preservation services unit.

The deacidification agent is methoxymagnesium methyl carbonate carried in a solvent system composed of methanol, trichlorotrifluoroethane (Freon 113), and dichlorodifluoromethane (Freon 12). Books are placed in wire baskets and dried to a moisture content of 0.5 %. The pretreated materials are immersed in the treatment solution for 1 h under modest pressure. Drying is initiated by evacuating the chamber and recovering the solvents. Final drying is performed under vacuum. It is during this stage that the deposition of the magnesium compounds occurs. Careful control of pressure regulates the evenness of deposition [469].

Following deposition, the active agent reacts with water from the air to form a mixture of intermediate magnesium compounds [458]:

$$H_3COMgOCOOCH_3 + 2\,H_2O \longrightarrow Mg(OH)_2 + 2\,CH_3OH + CO_2$$

$$H_2O + CO_2 \longrightarrow H_2CO_3$$

$$Mg(OH)_2 + H_2CO_3 \longrightarrow MgCO_3 + 2\,H_2O$$

$$MgCO_3 \longrightarrow MgO + CO_2$$

These magnesium compounds react with strong acids to form the corresponding neutral salts. For example, magnesium hydroxide converts sulfuric acid to magnesium sulfate.

$$Mg(OH)_2 + H_2SO_4 \longrightarrow MgSO_4 + 2\,H_2O$$

Magnesium sulfate is also generated by the reaction of magnesium carbonate with acid:

$$MgCO_3 + H_2SO_4 \longrightarrow MgSO_4 + H_2O + CO_2$$

Residual magnesium hydroxide, carbonate, and oxide form the alkaline buffer that protects paper from subsequent acid attack.

Similar chemistry is practiced by the Bibliotheque Nationale at its treatment facility in Sable sur Sarthe. The unit has annual capacity for 60 000 volumes. However, the French process employs ethoxymagnesium carbonate as the deacidification agent.

The Bookkeeper Process is a liquid, nonaqueous process that employs a Freon 113 carrier to deposit a submicron dispersion of magnesium oxide. No alcohol is involved and a surfactant is employed to keep the neutralization agent evenly distributed in the fluorocarbon solvent. The technique was developed by Koppers Chemical Company in 1981, who transferred the rights to the process to Richard Spatz in 1987. Book Preservation Technologies, in which Spatz is a principal, continues the development of the method.

Books are placed open on a supporting grate which is loaded into a treatment chamber. No predrying is required. The chamber is flooded with the treatment solution and the grate is mechanically agitated so that the pages are exposed and completely wetted. Books are subsequently removed from the chamber and placed

in an air dryer, from which the solvent is condensed and recycled [470]. Total process time is 5 h or less. The process can be operated in batch or continuous mode.

The Bookkeeper Process relies on the migration of paper acid to the magnesium oxide for neutralization to occur. To data the method has only been operated on a modest scale and there has been no known commercial application.

The British Library Process. The British Library has achieved promising results in its effort to deacidify and strengthen book paper in a bulk process that has been under development since 1980. The method involves graft copolymerization techniques initiated by low-intensity γ-radiation. Work has been conducted by the Industrial Chemistry Group of the University of Surrey under British Library supervision and sponsorship.

Short-chain monomers, in particular ethyl acrylate and methyl methacrylate in a volume ratio of 5:1, are introduced in liquid form without solvents in a closed system. Low-intensity γ-radiation polymerizes the monomers under ambient conditions. Additionally, the liquid monomer mixture has been modified to include an amine-substituted alkyl methacrylate to neutralize acidity and establish a buffer against future acid-catalyzed hydrolysis [471]. The presence of a small amount of a diacrylate leads to cross-linking and hence strengthening of the paper.

Paper strength, as measured by fold endurance, is enhanced by a factor of 5 to 10 and a weight gain of 15 – 20 % occurs. No dimensional changes in treated books have been observed. Researchers have concluded that the polymer is deposited within the cellulose network of the paper as well as on and between the fiber surfaces. Treated books retain their original appearance and pages do not stick together at the prescribed polymer deposition levels. Future development will include operation on a pilot-plant scale to permit refinement of the monomer mixture and processing details.

The Lithco Process. Lithco began development of a nonaqueous deacidification technique in 1988. Magnesium butoxytriglycolate, the active ingredient, is solubilized in Freon 113 as a ca. 10 % solution. The company markets the process for its ability to both deacidify and strengthen book paper.

The process consists of three steps: preconditioning, impregnation, and postconditioning. Books or other library materials are dried rapidly by dielectric heating to a moisture content of ca. 2 %. The documents are immersed in a pressurized Freon solution for deposition of the neutralization agent. Finally, excess solution is drained, excess chemicals are rinsed out, the solvent is removed under vacuum, and the books are removed. Total cycle time is less than 5 h.

The butoxytriglycolate ligands are intended to interact with cellulose to strengthen paper through hydrogen bonding. Similar chemistry based on zinc compounds has also been explored by Lithco on a more limited basis. Additionally, the company utilizes dielectric heating to speed the drying during both the first and last steps [472].

Early work was conducted in a demonstration unit capable of processing 20 standard bound volumes per batch. Lithco commissioned a larger pilot unit with capacity for 1000 books per day in May 1990.

Other Processes. Additionally, there are several other deacidification or strengthening techniques that have been commercialized and applied selectively to portions of a research library or archives collection. For instance, several such processes are utilized routinely to treat newspapers to neutralize acidity and restore strength. Newpaper is typically made of groundwood pulp to which cellulose fiber is added in processing. The resulting paper exhibits very low permanence, and represents a particular challenge for a research or national library.

The Austrian National Library Process. Work on an effective deacidification process for bound newspapers began in the Austrian Library in 1973. The technique, which has been in regular operation in Vienna since 1987, involves immersion in an aqueous solution of calcium hydroxide and methyl cellulose.

Process steps include: preparation, treatment, drying, and rebinding. Covers are removed from bound newspapers, and the materials are placed in a vacuum chamber. Following immersion in the treatment solution, the paper is shock frozen at $-30\,^{\circ}$C to prevent the formation of large crys-

tals that would damage the paper web. Final drying consists of freeze-drying to avoid pages sticking together. Treated papers experience no significant change in dimensions, permitting use of the original covers [473].

Although testing and preselection is required, the Library has successfully applied the technique to various other archival materials. Capacity is 3000 bound volumes per year.

Leipzig Paper Splitting Technique. Since 1980, the Deutsche Bücherei in Leipzig has successfully operated a process that splits heavily mechanically damaged paper and introduces a new paper carrier composed of cotton linters. Developers intend to follow the mechanical treatment with an aqueous liquid-phase introduction of calcium hydroxide to neutralize the paper. Mechanization is under development.

German State Archives in Bückeburg. Pioneering work conducted in Bückeburg focuses on the fixation by cationic and anionic active agents prior to aqueous liquid phase impregnation with a solution of magnesium hydrogencarbonate for neutralization and carboxymethyl cellulose for strengthening. Successful pretreatment to fix dyes and inks would minimize the preselection characteristically required before the use of liquid-phase processes [474].

Strengthening without Deacidification. The most commonly discussed method to accomplish strengthening alone involves the gas-phase deposition of poly-*p*-xylylene. Paralyne is the generic name for the various members of the polymer series of *p*-xylylenes developed by the Nova Tran Corporation. The technique is capable of establishing a thin, durable coating of the polymer on the paper. No deacidification is accomplished, however, and the high cost of treatment has discouraged widespread use.

12.5. Commercial Prospects

The most compelling evidence for the need for mass deacidification is the substantial involvement of national libraries in the development of treatment methods. Research libraries and archives groups universally seek a low-cost option that will permit preservation of collection materials in their original form. Current deacidification techniques represent a viable and cost effective preservation solution to the problem of acid paper. Although it is presently difficult to define costs in order to make meaningful comparisons among processes, mass deacidification is unquestionably far less costly than reformatting options such as microform. In addition, users of research library and archives collections continue to prefer original format materials.

More than one mass deacidification method will be adopted by the library and archives community. The problem is so great and the needs are so diverse that a single institution may well elect to pursue multiple treatment options.

Table 25. Comparison of selected properties of wood and paper

Property	Wood	Paper
Density, g/cm^3	0.3 – 0.5	0.5 – 1.5
Young's modulus, GPa	1.3	1.8 – 4.3
Tensile strength, MPa	90 – 200	20 – 200
Breaking length, km	7 – 30	1 – 8
Strain to rupture, %	0.6 – 0.8	0.8 – 4
Fiber alignment	ordered	random

13. Properties and Testing

13.1. The Material Paper: A Survey

13.1.1. Introduction

The term paper refers to sheet material that is made essentially from fibers of plant origin. The characteristic feature of this nonhomogeneous material is its fiber network, which is usually arranged in layers containing pores of varying size. The properties of paper differ substantially from those of the fiber resources. These differences are illustrated in Table 25, which shows some selected material properties of wood, as the most important fiber source, in comparison with paper. As a result of the physical and chemical properties and the highly ordered state of wood fibers, wood is unsuitable for direct production of sheet material which could serve as writing material. The construction of paper, i.e., sheet material with properties that are optimal for writing or printing, is possible only after the breakdown of wood into its elemental fiber building blocks, their modification in the pulping process, and their controlled bonding in the papermaking process. Paper can be made with an

Table 26. Selected properties of some grades of paper and board

Property [a]		LWC paper	Newsprint	Wood-free offset printing paper	Bag paper	Creped bag paper	Boxboard
Basis weight, g/m^2 [b]		55.2	53.5	81.4	69.7	91.7	286
Caliper, µm [c]		52	79	110	106	248	453
Density, g/cm^3 [c]		1.05	0.68	0.74	0.66	0.37	0.63
Specific volume, cm^3/g [c]		0.95	1.48	1.35	1.52	2.70	1.59
Tensile strength,	MD	40.9	26.5	64.5	78.5	65.0	245.5
N [d]	CD	14.6	12.6	25.1	53.5	55.0	80.5
Elongation,	MD	1.0	0.8	1.6	2.8	8.4	2.0
% [d]	CD	1.6	1.8	3.0	6.2	6.6	3.2
Breaking length,							
km	MD	4.95	3.30	5.25	7.50	4.70	5.75
	CD	1.75	1.55	2.05	5.15	4.00	1.85
Tear growth							
resistance,	MD	180	120	340	640	910	1200
mNm/m [e]	CD	250	170	410	630	1110	2440
Bursting stength							
(Mullen), kPa [f]		100	50	150	350	450	550
Smoothness	fs	633	71.4	18.2	6.1	<1	22.9
(Bekk) [g]	ws	600	63.4	8.6	4.4	<1	2.9
Roughness	fs	22	90	273	702	>500	113
(Bendtsen),							
mL/min [h]	ws	20	114	469	838	>500	>500
Specific bending							
stiffness,	MD	0.06	0.12	0.37	0.39	0.39	37.8
Nmm [i]	CD	0.03	0.04	0.19	0.18	0.54	13.1
Brightness,	fs	73.1	55.9	84.6			81.5
% [j]	ws	72.3	56.3	84.6			
Opacity, % [k]		92.7	96.5	90.5			

[a] = machine direction; CD = cross machine direction; fs = felt side; ws = wire side. [b] ISO 536. [c] ISO 543. [d] ISO 1924. [e] DIN SCAN/P 11:64. [f] ISO 2758. [g] ISO 5672. [h] ISO 2494. [i] ISO 5629. [j] ISO 2470. [k] ISO 2471.

extremely wide spectrum of properties by varying the construction parameters: type of wood, pulping process, fiber preparation, and papermaking process. This is illustrated in Table 26.

The wide range of product properties that result from the construction parameters mentioned above can be increased still further by the use of additives such as mineral and chemical aids. For instance, the smoothness and brightness of printing paper can be increased by addition of inorganic pigments (fillers) [475].

Many special demands made on paper can be met by using suitable chemical additives, e.g., wet strength by using wet-strength additives or printability by using sizing agents [476]. However, paper produced with all these aids will never neet all conceivable requirements at a time. In each case only a few paper properties that are particularly important for the intended purpose can be optimized. As far as the other properties are concerned, concessions must be made.

13.1.2. Material Properties

The specific characteristic features of paper vary widely, depending on the type of paper. Nevertheless, there are a number of properties that are characteristic of all papers. These are inhomogeneity, hygroscopicity, anisotropy, and viscoelasticity.

Paper is an inhomogeneous material made from the homogeneous elements: fibers, fillers, and air-filled pores. The homogeneous regions extend over a few micrometers (a few millimeters in the longitudinal fiber direction), corresponding to the characteristic dimensions of these elements.

A sheet of paper also exhibits other inhomogeneities, which can have characteristic di-

mensions of a few millimeters or centimeters. These are known as the cloud effect and result from undesired fiber flocculation in the sheet forming process [477]. In this way, variations in caliper and density develop, and this leads to corresponding variations in transparency, the "clouds". Finally, the production process can also give rise to inhomogeneities, e.g., due to stock pulsations in the headbox and in the sheet forming section [478]. The existence of inhomogeneities must be taken into account in paper testing, e.g., by choosing the appropriate size of sample.

The most important characteristic of paper is its hygroscopicity, i.e., its ability to absorb or release moisture, depending on the ambient climate, until an equilibrium is reached. It is significant whether the state of equilibrium is established by absorption or desorption of water [479]. The moisture content of some types of paper at various levels of relative humidity is listed in Table 27.

As a result of the hygroscopicity of paper, physical paper properties, such as sheet dimensions, basis weight, tensile strength, strain to rupture, folding strength, bending stiffness, etc., are dependent on the ambient conditions [480]. With a change in the ambient climate, the attainment of a new equilibrium moisture content of paper is a slow process which takes several hours. However, the change in the physical properties of the paper occurs almost without delay. For this reason, paper should be processed in rooms in which the climatic conditions are favorable for the particular paper property required. The paper must be in equilibrium with the climate in the room and the climatic conditions must be kept largely constant. The testing of paper requires the establishment of testing climatic conditions (standard climate: 23 °C, 50 % relative humidity, cf. Section 13.2.1).

Paper is an anisotropic material with regard to many physical properties. This anisotropy is due to the anisotropic properties of the individual fibers, which result from the fibrillated microstructure of the fiber rather than the fiber shape [481]. As a result of the fibrillated structure, the fiber can accept, for example, high tensile forces in the direction of the fiber axis with low elongation; however, even small tensile forces acting perpendicular to the fiber axis cause high elongations. During drying in the paper machine the fiber shrinks in the axial direction by only a few percent. In comparison, shrinkage in the direction perpendicular to the fiber axis reaches 40 %. If the fiber orientation were completely random, however, the anisotropy of the fibers would cancel out and the paper properties would be isotropic. Thus, the observed anisotropy is due to the production process. First, the fibers are not aligned randomly in the sheet-forming process because the machine direction is preferred (anisotropy of fiber orientation), and second the fiber mat is passed through the paper machine under tension, which prevents free shrinkage of the fibers during drying. This inhibition of shrinkage is only effective in the machine direction and is essentially responsible for the anisotropy of the moisture expansion of finished paper, which is generally far lower in the machine direction than in the cross-machine direction. The inhibition of shrinkage during drying also affects the load – deformation properties of paper, which are therefore also anisotropic (cf. Table 26).

Many papers exhibit a further anisotropy with respect to their composition in the normal direction (z-direction). The process technology of production on the wire is responsible for this phenomenon. On the wire side, substances that can pass through the wire are washed out, while on the other side, they are retained by the fiber mat. This results in different smoothness properties on the two sides of the paper, also known as the two-sidedness of paper.

As a result of the in-plane anisotropy of the sheet, the machine direction must be considered when taking samples for testing anistropic properties, such as strength. Two-sidedness is important in the testing of printability, for example.

Finally, paper has viscoelastic properties [482], [483], i.e., it can be elastic like a solid or viscous like a thick liquid. The viscoelasticity is also a result of the superimposition of the properties of the individual fibers and those of the fiber network. The viscous flow of the individual fibers is caused by the sliding of fibrils at high tensile loads. The individual fibers exhibit elastic behavior at low tensile loads. In the fiber structure of paper, the sliding of fibers at fiber intersections results in a flow effect and the paper undergoes plastic deformation.

A characteristic feature of viscoelasticity is the dependence of the onset of flow on the load-

Table 27. Moisture content of selected types of paper in equilibrium with various levels of humidity at 23 °C

Type of paper	Basis weight, g/m^2	Relative humidity, %		
		20	50	75
Newsprint	50	6.5	8.5	10.0
Gravure printing	65	6.0	8.0	9.2
Offset, mechanical	80	5.2	6.4	7.9
Offset, wood-free	80	5.0	6.3	7.5
Art paper, mechanical	120	4.1	5.7	6.9
Art paper, wood-free	120	3.8	5.3	6.5

ing rate. Small loads acting over long periods of time result in flow, whereas large intermittently acting loads cause only elastic deformation. Therefore, in the determination of the tensile strength of paper, the time is fixed within which the tensile stress applied causes the paper to break.

13.1.3. Summary

The material paper has essentially four characteristic features: inhomogeneity, hygroscopicity, anisotropy, and viscoelasticity. These features must be taken into consideration in the testing of paper (cf. Table 28).

Table 28. Characteristic material properties and their effect on testing conditions

Property	Effect
Inhomogeneity	sample dimensions, number of measurements
Anisotropy	testing direction
Two-sidedness	tested side
Hygroscopicity	relative humidity, temperature
Viscoelasticity	duration of test

The magnitude of these features depends on the type of fiber, the fiber raw material, the pulping process, and on the process techniques used in stock preparation and papermaking. Consequently, there are a large number of variables in papermaking and many degrees of freedom are involved in the adjustment of the desired paper properties. Therefore, several different paths can usually be pursued to obtain papers with comparable technical properties. This is of importance in the standardization of paper properties. Only the technical characteristics may be stipulated in each case, but not the process used to realize these characteristics. Otherwise, there would be the danger of excluding new technologies and impeding technical progress.

13.2. Paper Testing [484–486]

13.2.1. Preliminary Measures

A direct consequence of the paper properties described above is the necessity of having available sufficient sample material that can be tested. Sampling must be carried out correctly because the test results obtained with a sample should be representative for all of the paper from which the sample was taken. A straightforward process for cut-size and reel paper is described in ISO 186 (1988).

As a result of its hygroscopic properties, paper must be tested under standard climatic conditions and the samples must be conditioned in this climate before testing (sample conditioning). In ISO 187 (1977), the preferred climate for paper testing is specified as 23 °C and 50 % relative humidity.

13.2.2. Basic Properties

The basic properties of paper include the dimensions and the mass. In the case of sheet material such as paper, the basis weight m_A is determined in accordance with ISO 536 (1976). The value m_A is the ratio of the mass m to the area A of a sample. If m is measured in grams and A in square centimeters, then

$$m_A = \frac{m}{A} \cdot 10000\,\mathrm{g/m^2}$$

Another basic property of paper is the caliper. The caliper of paper must be defined in terms

of the measuring instrument. ISO 534 (1988) specifies an instrument with two parallel planar measuring surfaces of 200 mm^2 which act on the sample with a surface pressure of 100 kPa. The resulting distance between the surfaces is the caliper D of the paper.

The ratio of the basis weight to the caliper is the density of paper. The reciprocal density is the specific volume.

The moisture expansion of paper, i.e., the change in dimensions with changing ambient climate, is also a basic property. According to DIN 53 130, the moisture expansion (ME) is the ratio of the length L of a sample paper strip, resulting from a change in the ambient climate from 23 °C/45 % relative humidity to 23 °C/83 % relative humidity, to the original length L_0. The temperature can also be 20 °C in each case.

$$\mathrm{ME} = \frac{L}{L_0} \cdot 100\,\%$$

The moisture expansion test is also standardized in ISO 8226-1 (1985).

13.2.3. Composition and Chemical Paper Testing

The fiber composition of paper is determined with a light microscope (fiber microscopy). Due to the varying fiber morphology, both the fiber raw material and the pulping process can be determined. Chemical dyeing methods are used for contrast enhancement.

The term moisture content refers to the amount of water in the paper in equilibrium with a defined ambient climate. It is determined gravimetrically in accordance with ISO 287 (1985). A dry sample is produced by drying the paper to constant weight at 105 °C. The moisture content is expressed in percent, based on the moist sample.

Chemical testing is used to determine the components of paper. Moreover, it is of significance in papermaking for quality control and process control. Another important field of chemical paper testing is the control of papers for the packaging of food.

The entire spectrum of analytical methods used for the chemical testing of paper ranges from gravimetric methods to spectroscopy, electrochemical and enzymatic methods, as well as sensors and microbiological processes.

Only in exceptional cases can the components of paper be determined nondestructively. In general, the components to be determined must be extracted out with water or organic solvents. For example, the resin content of paper is determined gravimetrically in the extraction residue by using an organic solvent as an extracting agent.

The inorganic fillers contained in paper are determined as residue on ignition (ash content). In accordance with ISO 2144 (1987), the sample is ignited at 900 °C until no change in weight is observed. The ignition residue is then determined gravimetrically and expressed as a percentage of the orginal weight of the sample. Other ignition temperatures are used for chemical pulp (575 °C) and filter papers (800 °C). X-ray microanalysis is employed for the quantitative determination of the types and amounts of fillers. Elements such as calcium, magnesium, and metals in trace amounts can be analyzed by atomic absorption spectroscopy. For the determination of organic substances in paper extracts, IR spectrophotometric and chromatographic methods are applied, e.g., gas chromatography for volatile components and high performance liquid chromatography for thermally sensitive constituents.

Other chromatographic methods are used for the determination of chlorides, nitrates, and sulfates, or for the differentiation and characterization of dyes and optical brighteners.

Apart from the analytical methods mentioned above, the sum parameters are of special importance. These include the ash content, pH (ISO 6588, 1981), electrical conductivity of aqueous extracts (ISO 6587, 1983), and the resin content of paper.

Apart from the ISO standards mentioned above, other standards for chemical testing of paper are, e.g., the ZELLCHEMING, SCAN, and TAPPI test methods.

13.2.4. Strength Properties

In the tensile strength test, the tensile strength and the strain at rupture are determined as values characteristic of paper strength (ISO 1924,

1981). The tensile strength together with the sample width gives the tensile strength per unit width, measured in N/m, which together with the sample thickness gives the tensile strength, measured in kPa.

In paper technology, the breaking length is also of some importance as a relative measure of strength. This is the length of a freely suspended paper strip of any constant width and thickness that breaks at the point of suspension due to its own weight.

The breaking length L in kilometers is then calculated from the tensile strength force F_B of the paper measured in N, the basis weight m_A in g/m^2, the width of the strip w in mm, and the gravitational constant $g = 9.81$ m/s^2.

$$L = 10^3 \cdot \frac{F_B}{m_A w g}$$

The bursting strength is important in the characterization of packaging papers. The bursting strength of paper is determined in accordance with ISO 2758 (1983) and that of board in accordance with ISO 2759 (1983). In the testing instrument, the free area of the sample stretched in a circular frame is exposed to increasing pressures until it ruptures.

Packing materials are also characterized by their bursting strength. For instance, set values are stipulated in DIN 55 468 Part 1 for the strength of corrugated boards, and are used for classification. The same standard specifies set values for the puncture resistance. This strength property is determined, mainly for cardboard and corrugated board, by measuring the resistance offered by a sample to the penetration of a pyramidal body (ISO 3036, 1975).

To determine the tear resistance (Elmendorf), a sample is torn with the help of a pendulum device starting from a predetermined cut, and the work required for a tear of a given length is measured. The test is described in ISO 1974 (1990).

The plybond strength (DIN 54 516) is a measure of the matrix strength of multi-ply papers and boards. It is determined by measuring the maximum force per unit sample width required to split the sample.

The crush resistance of paper under compressive stress is measured as the short span compression strength. To prevent the sample kinking, the free span length must be very small (typically 0.7 mm). The resistance measured according to DIN 54 518 gives the ultimate crushing load per unit width.

The flat crush resistance (ISO 3035, 1982) and the edge crush resistance (ISO 3037, 1982) are used to characterize the strength of corrugated board. In the former test, a circular sample of corrugated board is exposed to increasing pressures between two parallel planar plates until the corrugation collapses, and the maximum load is measured. In the second test, the load is applied in the direction of corrugation starting from the edge of the sample. The edge crush resistance is used in DIN 55 468 Part 1 for the classification of corrugated boards.

The folding strength is an important property of both packaging and graphic papers. It can be determined either as described in ISO 5626 (1978) as the Schopper double fold value, or as described in SCAN P 17 (1977) as the Köhler – Molin folding endurance value. The sample is subjected to specified tension in both cases. The Brecht – Wesp pressure folder functions without tension. A sharp fold in the sample is produced between two rolls and is repeated up to fifty times in the same position. The ratio of the tensile strength of the folded sample to that of the unfolded sample is the folding strength.

13.2.5. Load – Deformation Properties

Not only can the strength properties, tensile strength and strain to rupture be determined in the tensile strength test (ISO 1924, 1983), but also the tensile energy absorption, TEA, which is especially important for bag paper. The TEA value represents the nonelastic portion of the deformation energy and thus that portion of dissipated energy which should be high in the case of bag paper.

The resistance to flexural stress is measured as the bending stiffness under approximately pure elastic deformation as defined in DIN 53 121. This test is usually performed with a two-point beam method. To guarantee elastic deformation, maximum bending angles, which depend on the span length and the sample thickness, are specified. This may be circumvented by the resonance length method (ISO 5629, 1983) in which the resonance length of a free sample strip is measured. Resonance is generated via a clamp vibrating at 25 Hz.

Deformations in board and cardboard can also be produced by creasing. Testing the properties of creases requires defined production processes, e.g., DIN 55 437 Part 1. In Parts 2 and 3 of this standard, methods are described for the manual folding of the creases and for the visual evaluation of the folding or the technical evaluation of the creases with a folding-moment tester.

13.2.6. Surface Properties

A large number of methods are available for evaluating the topography of paper surfaces. The Bekk smoothness is determined according to ISO 5627 (1984) as follows. At a defined pressure difference, the time in seconds is measured which a specified amount of air requires to flow radially inwards between the paper surface and a ring-shaped glass plate and on into a vacuum chamber. The smoother the surface of the paper, the longer the pressure equalization takes. The determination of roughness according to Bendtsen (ISO 2494, 1974) is related to the Bekk method for the determination of smoothness. In this case, however, the direction of the air flow is reversed and the magnitude of the air flow is measured. The air escapes under defined conditions between the measuring ring of the measuring head and the paper surface under constant overpressure.

Another surface property of paper is the abrasion resistance. The mechanical abrasion resistance of surfaces is determined in the friction wheel process (DIN 53 109). In this process, the amount of abrasion is measured which is obtained by abrading the conditioned or wet sample with an abrasion wheel of defined quality under defined conditions.

13.2.7. Optical Properties

An object, e.g., a paper surface, is termed white when the illumination intensity and the absorption capacity of the surface are independent of the wavelength. Deviations confer a more or less pronounced color shade on the surface.

In the paper industry, a special process is used to characterize the brightness because this is one of the most important optical properties of paper. The determination of the reflectance factor (ISO brightness) is based on ISO 2470 (1977). For this test, a filter is used which has an intensity maximum at a wavelength of 457 nm. The reflectance (blue component) measured in a reflectometer under specified conditions is known as brightness. It is expressed as a percentage of the brightness of a white standard.

Another optical property of paper is its transparency, i.e., a measure of its light transmittance. It is calculated from the reflectance factors R_0, R_w, and $R_{(w)}$, which are determined in accordance with DIN 53 145 Part 1 or Part 2. The reflectance factor of the individual sheet on a completely black background is R_0, R_w is the reflectance factor of the individual sheet on a white background, and $R_{(w)}$ is the reflectance factor of the white base.

The opacity is a measure of light-tightness. It is defined in ISO 2471 (1977) as the ratio of the reflectance factor R_0 to the reflectance factor R_∞. Both reflectance factors are determined in accordance to DIN 53 145. R_0 is measured as the reflectance factor of an individual sheet on a completely black background and R_∞ as the reflectance factor of an "infinitely" thick stack of the same paper.

13.2.8. Printing Properties

The printing properties of paper result from complex interactions between printing ink, printing process, and paper. Practice-oriented printability tests must be performed to evaluate these properties. Test printers are also suitable for this purpose. Instruments for offset printing, gravure printing, and flexographic printing can be used with standard printing inks under laboratory conditions to test the dry pick resistance or the wet pick resistance of papers. Mottling and ink penetration properties can also be tested. Conversely, the behavior of various printing inks towards standard papers can also be evaluated.

Furthermore, defined proof copies produced with the test printers can be used to test full ink coverage, color density, color gloss, shade, abrasion resistance, stacking ability, and contact yellowing.

13.2.9. Behavior Towards Liquids

The behavior of liquids towards paper is characterized by the processes of wetting and pene-

tration. In both cases, the characteristic physical property is the surface tension. This value can be measured directly and tensiometrically in the case of liquids and indirectly, via the contact angle of test liquid droplets, in the case of solids such as paper. A liquid wets the surface of paper only if its surface tension is lower than that of the paper. The same holds for the wetting of the capillary walls upon penetration of liquids into the capillaries of the paper.

If the wetting and the penetrating capacity of liquids are to be changed, the surface tension of the paper must also be changed. This is achieved, for instance, by sizing the paper, a process which must fulfill the requirements regarding printability with aqueous inks. According to DIN 53 126, paper is considered to be printable if an ink line drawn with a normally adjusted drawing pen has neither run nor penetrated into the paper after 24 h.

The water absorption W_A (Cobb) (ISO 535, 1976) refers to the amount of water that is absorbed by a certain area of paper on one-sided contact for a specified exposure time. The time of exposure to water must be chosen such that a sufficient amount of water enters into the fiber matrix, but does not penetrate to the opposite side of the sample.

In the determination of grease permeability (DIN 53 116), red-colored palm kernel oil is used as the testing agent. The passage of fat through the sample under specified conditions is then evaluated.

13.2.10. Exclusion of Gases and Vapors

As a rule, papers have only a limited ability to exclude gases and vapors. Of particular importance are the air permeability (e.g., for filtration properties) and the water-vapor permeability.

There are two standardized test methods available for the determination of the mean air permeability: the Bendsen method (ISO 5636-2, 1984) and the Schopper method (ISO 5636-3, 1984).

A gravimetric method for the determination of water-vapor permeability is described in ISO 535 (1976). This method is not suitable for very dense or coated papers. In these cases, the use of electrolysis described in DIN 53 122 Part 2 is recommended.

14. Economic Aspects

In 1988 world production of paper and board was ca. 228×10^6 t. Although paper and board are used in some form or other in all parts of the world which are inhabited by people, the production is left to a limited number of countries. More than 4000 paper machines are responsible for production in 95 of the approximately 200 countries in the world, whereby the 15 largest producer countries alone account for 85 % of total world production (Table 29).

Table 29. World paper and board production according to countries (1988)

Country	Production, 10^6 t	Share, %
United States	71.0	31
Japan	24.6	11
Canada	16.6	7
China	12.6	6
Soviet Union	10.7	5
FRG	10.5	5
Finland	8.6	4
Sweden	8.1	4
France	6.3	3
Italy	5.3	2
Brazil	4.6	2
UK	4.2	2
South Korea	3.6	2
Spain	3.4	1
Taiwan	2.9	1
80 other countries	34.1	15
World total	227.8	100

The various requirements for the production of paper are not met everywhere. For instance, suitable raw materials must be available, including water, energy, and trained personnel. Furthermore, the construction of paper mills requires a very high investment, and a sufficiently large market and a favorable location are both essential prerequisites for papermaking. For these reasons, the modern paper industry is based mainly in the coniferous forest zone of the Northern Hemisphere, with centers in Northern and Western Europe, North America, as well as in Eastern Europe and Japan.

World paper production has increased more than fivefold since 1950, when production was ca. 44×10^6 t (Table 30). During this time, the regions which did not belong to the classical paper

Table 30. World production of paper and board (10^6 t)

	1950	1960	1970	1980	1988
Europe	13.194	25.826	45.377	59.495	75.934
East	2.840	5.765	10.560	14.103	16.480
West	10.354	20.061	34.817	45.392	59.454
North America	28.286	39.393	56.323	71.179	87.638
United States	22.108	31.255	45.186	57.789	71.000
Canada	6.178	8.138	11.137	13.390	16.638
Japan	0.873	4.513	12.973	18.087	24.624
Others	1.387	5.193	15.014	23.357	39.650
World	43.740	74.925	129.687	172.118	227.846

Table 31. Regional shares of world production of paper and board (%)

Region	1950	1960	1970	1980	1988
Europe	30.1	34.5	35.0	34.6	33.3
East	6.5	7.7	8.1	8.2	7.2
West	23.6	26.8	26.9	26.4	26.1
North America	64.7	52.6	43.4	41.3	38.5
United States	50.5	41.7	34.8	33.6	31.2
Canada	14.2	10.9	8.6	7.7	7.3
Japan	2.0	6.0	10.0	10.5	10.8
Others	3.2	6.9	11.6	13.6	17.4
World	100	100	100	100	100

countries have come to the fore, and their share of world production has grown from ca. 3 % in 1950 to 17 % in 1988. At the same time, the circle of paper producing countries has widened considerably from 61 to 95 and now includes a large number of developing countries. In these countries, the tendency towards self sufficiency is influenced by national economic considerations and the need to save foreign exchange for paper imports.

In 1988, Europe, North America, and Japan still produced 83 % of the total world production (Table 31). However, the development in these regions has varied considerably. In Japan, the paper industry has developed from a low during the postwar years to become the second largest paper producer in the world. In the last three decades, Western Europe has more or less been able to maintain its share of world production. North America has had to accept large losses of its share of production, which has, however, been increasing relatively constantly.

In 1988 graphic papers accounted for more than 40 % of total paper production. This included newsprint (31.7×10^6 t) and other printing and writing papers (62.4×10^6 t). The production of packaging papers was 95.8×10^6 t, and that of the remaining papers and boards was 37.9×10^6 t (Table 32).

Table 32. World production of paper and board, arranged according to main types (1988)

Paper and board type	Production, 10^6 t	Share, %
Newsprint	32	14
Other printing and writing papers	62	27
Packaging papers and boards	96	42
Other papers and boards	38	17
Total	228	100

At present, a little more than one-fifth of the world production of paper and board is sold across borders. In 1988 exports and imports accounted for about 50×10^6 t.

The net imports to the United States, i.e., imports minus exports, amount to 7.1×10^6 t, which corresponds to a quota of 9 % (based on the consumption). In contrast, Canada, with its comparatively low population, has net exports amounting to 10.5×10^6 t, which is 63 % of its own paper production. The predominant part of Canadian exports go to the United States.

In Western Europe, the foreign trade situation within the EC and EFTA also varies greatly. The EC, with 12.4×10^6 t, a quota of 26 %, is a net importer. The EFTA with 15.6×10^6 t or 70 % (based on the internal production) is a net exporter. The EC as a whole has more than 320×10^6 consumers and is the largest market in

the world. As a customer, it plays a paramount role for the paper industries of the EFTA countries.

In Japan the amounts imported and exported balance out at ca. 10^6 t each.

Of all the remaining regions, the countries in Africa and Asia are the largest net importers with 0.8×10^6 t and 4×10^6 t, respectively. The imports and exports of South and Central America are roughly balanced. World consumption of paper and board in 1988 was 226×10^6 t. The United States has by far the highest consumption, followed by Japan and China. Germany is the fourth largest consumer, before the Soviet Union (Table 33).

Table 33. World paper and board consumption according to countries (1988)

Country	Consumption, 10^6 t	Share, %
United States	78.0	35
Japan	25.0	11
China	13.2	6
FRG	12.3	5
Soviet Union	10.0	4
UK	9.3	4
France	7.9	3
Italy	6.2	3
Canada	6.1	3
Spain	3.9	2
Brazil	3.8	2
South Korea	3.4	2
Taiwan	3.0	1
The Netherlands	2.8	1
Australia	2.5	1
188 other countries	38.0	17
World total	226.2	100

Looking at the world paper consumption from a geographical point in view, it is apparent that the industrial countries of Europe and North America, and Japan not only produce, but also consume the bulk of the paper (Table 34).

There is also a relatively rapid increase in the consumption of paper in the remaining regions, especially in Asian countries such as South Korea and Taiwan, which are becoming increasingly important for the international paper market. In the past 30 years, countries in South and Central America as well as in Eastern Europe and Africa have increased their share of world paper consumption annually. Improvements in the living conditions in these regions are accompanied by a corresponding increase in the consumption of paper. The rise in both the standard of living and the individual income is as important a prerequisite as the spread of literacy for the consumption of writing paper and printing products. Another important factor is the building up of export and consumer industries with their demand for packaging materials.

The amount of paper consumed, however, is not an adequate measure of the standard of living of a country. Only the relative per capita consumption can give a very rough indication of the living standard, because other factors such as the average income, way of life, and consumer patterns must also be taken into consideration. The United States has the highest per capita consumption per year (1988) with 311 kg, followed by Sweden (247 kg), Canada (236 kg), Switzerland (207 kg), Japan and Finland (204 kg each), Denmark (202 kg), and Germany (201 kg). In terms of paper consumption, the countries can be grouped according to their per capita consumption. For example, while the average consumption in Western Europe is 152 kg and in North America 304 kg, the per capita consumption in Eastern European countries is 40 kg, in Latin America 25 kg, and in Africa only 5 kg. Paper consumption per inhabitant is less than 1 kg in about 30 developing countries.

A comparison between population and paper consumption in different regions shows that Asia, which represents 60 % of the world population, accounted for only 24 % of world paper consumption. Conversely, North America has 5 % of the world's population but a consumption of 37 % (Table 35). An important indicator for the development of paper consumption is not only the gross national product but also the population growth.

World population has doubled from 2.5×10^9 in 1950 to 5.1×10^9 in 1988. The per capita consumption of paper worldwide was 18 kg in 1950 and 44 kg in 1988, an increase of 144 %. Thus the relative consumption of paper per capita has increased considerably faster than the world population. It is obvious that these average global values do not reflect the substantial regional differences.

In summary, the largest growth percentage potential for paper consumption is in the Third World countries, especially in Asia, while industrialized countries such as the United States, Japan, and the EC countries have relatively low growth rates, but still represent in absolute terms a considerable market potential.

Table 34. Regional shares of world paper and board consumption (%)

Region	1950	1960	1970	1980	1988
Europe	27.3	33.6	34.7	33.9	31.8
East	6.3	7.7	8.3	8.5	7.1
West	21.0	25.9	26.4	25.4	24.7
North America	64.2	50.9	43.4	38.4	37.2
USA	60.6	47.7	40.7	35.8	34.5
Canada	3.6	3.2	2.7	2.6	2.7
Japan	2.0	5.8	9.8	10.5	11.1
Others	6.5	9.7	12.1	17.2	19.9
World	100	100	100	100	100

Table 35. Population and paper consumption (1988)

Region	Population		Paper consumption	
	$\times 10^6$	Share, %	10^6 t	Share, %
Europe	750	14	72	32
Africa	610	12	3	1
North America	273	5	84	37
Central and South America	429	8	11	5
Asia	3 069	60	53	24
Australia/ Oceania	26	1	3	1
World	5 112	100	226	100

15. References

General References

1. J. P. Casey (ed.): *Pulp and Paper, Chemistry and Chemical Technology,* 3rd ed., Wiley Interscience, New York 1980.
2. D. Fengel, G. Wegener: *Wood, Chemistry, Ultrastructure, Reactions,* Walter de Gruyter, Berlin 1984.
3. M. J. Kocurek, O. V. Ingruber, A. Wong (eds.): "Sulfite Science and Technology," *Pulp and Paper Manufacture,* 3rd ed., **vol. 4,** Joint Textbook Committee of the Paper Industry, Montreal – Atlanta 1985.
4. M. J. Kocurek, F. Hamilton, B. Leopold (eds).: "Secondary Fibers and Non-Wood Pulping," *Pulp and Paper Manufacture,* 3rd ed., vol. 3, Joint Textbook Committee of the Paper Industry, Montreal – Atlanta 1987.
5. M. J. Kocurek, T. M. Grace, E. Malcolm (eds.): "Alkaline Pulping," *Pulp and Paper Manufacture,* 3rd ed., **vol. 5,** Joint Textbook Committee of the Pulp and Paper Industry, Montreal – Atlanta 1989.
6. P. Lengyel, S. Morvay: *Chemie und Technologie der Zellstoffherstellung,* Güntter-Staib-Verlag, Biberach/Riß 1973.
7. A. Mimms, M. Kocurek, J. Pyatte, E. Wright (eds.): *Kraft Pulping,* Tappi Press, Atlanta 1990.
8. W. H. Rapson (ed.): "The Bleaching of Pulp," *Tappi Monogr. Ser.* **27** (1963).
9. S. A. Rydholm: *Pulping Processes,* Interscience Publ., New York 1965.
10. R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979.
11. E. Sjöström: *Wood Chemistry, Fundamentals and Applications,* Academic Press, New York 1981.
12. G. A. Smook: *Handbook of Pulp & Paper Technology,* Tappi Press, Atlanta 1990.
13. C. F. B. Stevens, M. J. Kocurek (eds.): "Properties of Fibrous Raw Materials and Their Preparation for Pulping," *Pulp and Paper Manufacture,* 3rd ed., **vol. 1,** Joint Textbook Committee of the Paper Industry, Montreal – Atlanta 1983.
14. H. F. J. Wenzl: *Sulphite Pulping Technology,* Lockwood Trade Journal Co., New York 1965.
15. H. F. J. Wenzl: *Kraft Pulping,* Lockwood Publ. Co., New York 1967.

Specific References

16. J. E. Atchison, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1988, 25 – 45.
17. B. Steenberg, *Sven. Papperstidn.* **64** (1961) 126 – 129.
18. L. Deslandes, *Paper (London)* **207** (1987) no. 5, 19 – 23.
19. B. Zobel, *Tappi J.* **72** (1988) no. 12, 42 – 46.

20. S. Sideaway, *Tappi J.* **72** (1988) no. 12, 47 – 51.
21. J. T. Jeyasingam, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1988, 571 – 579.
22. A. M. Hurter, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1988, 139 – 154.
23. S. Yllmer, B. Enström, *Sven. Papperstidn.* **59** (1956) 229 – 232.
24. S. Yllmer, B. Enström, *Sven. Papperstidn.* **60** (1957) 549 – 554.
25. I. Croon, B. F. Enström, *Tappi* **44** (1961) 870 – 874.
26. S. Axelsson, I. Croon, B. Enström, *Sven. Papperstidn.* **65** (1962) 693 – 697.
27. J. Gierer, I. Norén, *Acta Chem. Scand.* **16** (1962) 1713 – 1729.
28. J. Gierer, L. A. Smedman, *Acta Chem. Scand.* **19** (1965) 1103 – 1112.
29. J. Gierer, *Sven. Papperstidn.* **73** (1970) 571 – 596.
30. J. Marton in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignins, Occurrence, Formation, Structure and Reactions,* Wiley-Interscience, New York 1971, pp. 639 –694.
31. J. Gierer, I. Pettersson, L. A. Smedman, I. Wennberg, *Acta Chem. Scand.* **27** (1973) 2082 – 2094.
32. J. Gierer, S. Ljunggren, *Sven. Papperstidn.* **82** (1979) 503 – 508.
33. J. Gierer, S. Ljunggren, *Sven. Papperstidn.* **82** (1979) 71 – 81.
34. J. Gierer, *Wood Sci. Technol.* **14** (1980) 241 – 266.
35. J. Gierer, I. Norén, *Holzforschung* **34** (1981) 197 –200.
36. J. Gierer, *Holzforschung* **36** (1982) 43 – 51, 55-65.
37. M. H. Johansson, O. Samuelson, *Wood Sci. Technol.* **11** (1977) 251 – 263.
38. M. H. Johansson, O. Samuelson, *Sven. Papperstidn.* **80** (1977) 519 – 524.
39. B. Lindberg, *Sven. Papperstidn.* **59** (1956) 531 – 534.
40. D. B. Mutton, *Pulp Pap. Mag. Can.* **65** (1964) T 41 – T 51.
41. E. Sjöström, *Tappi* **60** (1977) no. 9, 151 – 154.
42. J. W. Green et al., *Tappi* **60** (1977) no. 10, 120 – 125.
43. O. Franzon, O. Samuelson, *Sven. Papperstidn.* **60** (1957) 872 – 877.
44. B. Alfredsson, L. Gedda, O. Samuelson, *Sven. Papperstidn.* **64** (1961) 694 – 698.
45. M. H. Johansson, O. Samuelson, *Carbohydr. Res.* **34** (1974) 33 – 41.
46. R. D. Brooks, N. S. Thompson, *Tappi* **49** (1966) 362 – 366.
47. C. H. Matthews, *Sven. Papperstidn.* **77** (1974) 629 –635.
48. Y. Z. Lai, *Ekman-Days 1981, Int. Symp. Wood Pulping Chem.* 1981, vol. 2, 26 – 33.
49. N. Hartler, *Tappi* **50** (1967) no. 3, 156 – 160.
50. P. J. Kleppe, *Tappi* **53** (1970) no. 1, 35 – 47.
51. A. R. Procter, R. H. Wiedenkamp, *J. Polym. Sci. Part C* **28** (1969) 1 – 13.
52. N. Hartler, L. A. Olsson, *Sven. Papperstidn.* **75** (1972) 559 – 565.
53. N. Hartler, *Sven. Papperstidn.* **62** (1959) 467 – 470.
54. R. Aurell, N. Hartler, *Tappi* **46** (1963) 209 – 215.
55. E. Hägglund, *Sven. Papperstidn.* **49** (1946) 204.
56. B. Alfredsson, O. Samuelson, B. Sandstig, *Sven. Papperstidn.* **66** (1963) 703 – 706.
57. N. Hartler, *Sven. Papperstidn.* **68** (1965) 369 – 377.
58. V. B. Lasmarias, R. C. Peterson, *Cellul. Chem. Technol.* **3** (1969) 479 – 496.
59. A. Teder, *Sven. Papperstidn.* **72** (1969) 294 – 303.
60. B. Bach, G. Fiehn, *Zellst. Pap. (Leipzig)* **21** (1972) 3 – 7.
61. H. H. Holton, *Pulp Pap. Can.* **78** (1977) T 218 – T 223.
62. H. H. Holton, F. L. Chapman, *Tappi* **60** (1977) no. 11, 121 – 125.
63. K. Goel, A. M. Ayroud, B. Branch, *Tappi* **63** (1980) no. 8, 83 – 85.
64. T. J. Blain, H. H. Holton, *Pulp Pap. Can.* **84** (1983) T 124 – T 129.
65. L. Löwendahl, O. Samuelson, *Sven. Papperstidn.* **80** (1977) 549 – 551.
66. L. Löwendahl, O. Samuelson, *Tappi* **61** (1978) no. 2, 19 – 21.
67. B. I. Fleming, G. J. Kubes, J. M. MacLeod, H. I. Bolker, *Tappi* **61** (1978) no. 6, 43 – 46.
68. J. R. Obst, L. L. Landucci, N. Sanyer, *Tappi* **62** (1970) no. 1, 55 – 59.
69. J. Gierer, O. Lindeberg, I. Norén, *Holzforschung* **33** (1979) 213 – 214.
70. G. J. Kubes, B. I. Fleming, J. M. MacLeod, H. I. Bolker, *Wood Sci. Technol.* **14** (1980) 207 – 228.
71. A. Wong, *CPPA Annual Meeting,* Montreal 1987, Prepr. 73 B, 249 – 254.
72. G. W. Kutney, *Pulp Pap.* **61** (1987) 73 – 75.
73. N. E. Virkola, J. Kettunen, I. Yrjälä, *Paperi ja Puu* **61** (1979) 685 – 700.
74. J. Kettunen, N. E. Virkola, J. Yrjälä, *Paperi ja Puu* **61** (1979) 686 – 700.

75. S. Raubenheimer, S. H. Eggers, *Papier (Darmstadt)* **34** (1980) no. 10 A, V 19 – V 23.
76. N. E. Virkola, R. Pusa, J. Kettunen, *Tappi* **64** (1981) no. 5, 103 – 107.
77. A. Wong, *PPI, Pulp Pap. Int.* **23** (1981) 55 – 59.
78. O. V. Ingruber, M. Stradal, J. A. Histed, *Pulp Pap. Can.* **83** (1982) T 342 – T 349.
79. R. E. Dörr, *Sven. Papperstidn.* **46** (1943) 361 – 369.
80. G. A. Richter, *Tappi* **39** (1956) 193 – 210.
81. M. V. Tuominen, *Paperi ja Puu* **50** (1968) 517 – 528.
82. L. A. Hiett, *Tappi J.* **68** (1985) no. 2, 42 – 48.
83. H. Crönert, *Papier (Darmstadt)* **40** (1986) 619 – 626.
84. H. Crönert, *Papier (Darmstadt)* **41** (1987) 63 – 71.
85. N. Hartler, L. A. Lindström, *Int. Dissolving Pulps Conf. (Prepr.)* 1987, 47 – 52.
86. D. K. Misra in J. P. Casey (ed.): *Pulp and Paper, Chemistry and Chemical Technology,* 3rd ed., **vol. 1,** Wiley-Interscience, New York 1980, pp. 504 – 568.
87. Tappi, Non Wood Plant Fiber Pulping Progress Report Nr. 1 – 18.
88. M. F. Judt, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1987, 451 – 459.
89. F. Wultsch, *Papier (Darmstadt)* **24** (1970) 394 – 408.
90. N. Hartler, L. Stockman, O. Sundberg, *Sven. Papperstidn.* **64** (1961) 33 – 49, 67 – 85.
91. N. H. Schön, L. Wannholt, *Sven. Papperstidn.* **72** (1969) 489 – 492.
92. G. Gellerstedt, J. Gierer, *Sven. Papperstidn.* **74** (1971) 117 – 127.
93. G. Gellerstedt, *Sven. Papperstidn.* **79** (1976) 537 –543.
94. D. W. Blennie in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignins – Occurrence, Formation, Structure and Reaction,* Wiley-Interscience, New York 1971, pp. 597 – 637.
95. G. Gellerstedt, J. Gierer, *Acta Chem. Scand. B* **37** (1977) 729 – 730.
96. J. K. Hamilton, *Wood Chemistry Symp. Proc.* (1962) 197 – 217.
97. F. Shafizadeh, *Tappi* **46** (1963) 381 – 383.
98. J. F. Harris, *Appl. Polym. Symp.* **28** (1975) 131 – 144.
99. N. Hartler, L. Lind, L. Stockman, *Sven. Papperstidn.* **64** (1961) 160 – 166.
100. M. Goliath, B. O. Lindgren, *Sven. Papperstidn.* **64** (1961) 109 – 112, 469 – 471.
101. N. H. Schöön, *Sven. Papperstidn.* **61** (1964) 624 – 633.
102. N. E. Virkola, A. A. Alm, M. Mäkinen, R. Soila, *Paperi ja Puu* **42** (1960) 665 – 676.
103. F. M. Ernest, S. M. Harman, *Tappi* **50** (1967) no. 12, 110 A–116 A.
104. G. H. Tomlinson, G. H. Tomlinson II, J. R. G. Bryce, N. G. M. Tuck, *Pulp Pap. Mag. Can.* **59** (1958) 247 – 252.
105. T. W. Meloney, V. Gibbs, D. H. Andrews, *Pulp Pap. Mag. Can.* **74** (1973) T 378 – T 384.
106. R. M. Dorland, R. A. Leask, J. W. McKinney, *Pulp Pap. Mag. Can.* **59** (1958) 236 – 246.
107. W. W. Marteny in J. P. Casey (ed.): *Pulp and Paper. Chemistry and Chemical Technology,* 3rd ed., **vol. 1,** Wiley-Interscience, New York 1980, pp. 252 – 291.
108. H. E. Wörster, *Pulp Pap. Manuf.* **4** (1985) 130 –158.
109. O. V. Ingruber, G. A. Allard, *Pulp Pap. Mag. Can.* **74** (1973) T 354 – T 369.
110. S. Lagergren, B. Lundén, *Pulp Pap. Mag. Can.* **60** (1959) T 338 – T 341.
111. S. Lagergren, *Sven. Papperstidn.* **67** (1964) 238 –243.
112. G. E. Annergren, S. A. Rydholm, *Sven. Papperstidn.* **62** (1959) 737 – 746.
113. G. E. Annergren, L. Croon, B. F. Enström, S. A. Rydholm, *Sven. Papperstidn.* **64** (1961) 386 – 396.
114. G. E. Annergren, S. A. Rydholm, *Sven. Papperstidn.* **63** (1960) 591 – 605.
115. E. L. Bailey, *Tappi* **45** (1962) 689 – 691.
116. T. A. Pascoe, J. S. Buchanan, E. H. Kennedy, G. Sivola, *Tappi* **42** (1959) 264 – 281.
117. N. Sanyer, E. L. Keller, G. H. Chidester, *Tappi* **45** (1962) no. 2, 90 – 104.
118. J. E. Laine, J. Turumen, R. H. Rasanen, *Tappi* **62** (1979) no. 5, 65 – 68.
119. T. Ulmanen, R. Rasanen, M. Rantanen, *Tappi* **46** (1963) 151 A–153 A.
120. I. Hassinen, R. Räsänen, *Pulp Pap. Mag. Can.* **75**(1974) T 13 – T 15.
121. J. R. G. Bryce, G. H. Tomlinson, *Pulp Pap. Mag. Can.* **63** (1962) T 355 – T 361.
122. D. W. Cameron et al., *Appita* **35** (1982) 307 – 315.
123. I. D. Suckling, *Holzforschung* **43** (1989) 111 – 114.
124. A. Wong, *Tappi* **63** (1980) no. 4, 53 – 57.
125. E. Ojanen, J. Tulppala, J. Kettunen, *Paperi ja Puu* **64** (1982) 453 – 464.
126. O. V. Ingruber, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1985, 461 – 469.
127. P. Tikka, N. E. Virkola, A. Wong, *Papier (Darmstadt)* **49** (1989) no. 10 A, V 1 – V 7.

128. R. P. Wither, H. A. Captein, *For. Prod. J.* **10** (1960) 174 – 177.
129. D. L. Brink, *Tappi* **44** (1961) 256 – 263.
130. D. L. Brink, J. Vlamis, M. M. Merriman, *Tappi* **44** (1961) 263 – 270.
131. D. L. Brink, M. M. Merriman, E. J. Schwerdtfeger, *Tappi* **45** (1962) 315 – 326.
132. J. H. Kalisch, *Tappi* **50** (1967) 44 A–51 A.
133. T. N. Kleinert, K. V. Tayenthal, *Angew. Chem.* **44** (1931) 788 – 791.
134. T. N. Kleinert, *Tappi* **57** (1974) no. 8, 99 – 102.
135. T. N. Kleinert, *Tappi* **58** (1975) 170 – 171.
136. T. N. Kleinert, *Papier (Darmstadt)* **30** (1976) no. 10 A, V 18 – V 24.
137. R. Katzen, R. E. Fredrickson, B. F. Brush, *Pulp Pap.* **54** (1980) no. 8, 144 – 149.
138. R. Katzen, R. E. Fredrickson, B. F. Brush, *Chem. Eng. Progr.* **76** (1980) no. 2, 62 – 67.
139. J. H. Lora, S. Aziz, *Tappi J.* **68** (1985) no. 8, 94 – 97.
140. E. K. Pye, *Pima* **69** (1987) no. 11, 21 – 23.
141. E. K. Pye, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1990, 991 – 996.
142. M. Baumeister, E. Edel, *Papier (Darmstadt)* **34** (1980) no. 10 A, V 9 – V 18.
143. E. Edel, *DPW Dtsch. Papierwirtsch.* 1987, no. 1, T 39 – T 45.
144. E. Edel, *Papier (Darmstadt)* **41** (1989) no. 10 A, V 116 – V 123.
145. G. Dahlmann, M. C. Schröter, *Tappi J.* **73**(1990) no. 4, 237 – 240.
146. J. Nakano, H. Daima, S. Hosoya, A. Ishizu, *Ekman-Days 1981, Int. Symp. Wood Pulping Chem. Prepr.* 1981, vol. 2, 72 – 77.
147. R. Marton, S. Granzow, *Tappi* **65** (1982) no. 6, 103 – 106.
148. U. P. Gasche, *Papier (Darmstadt)* **39** (1985) no. 10 A, V 1 – V 7.
149. L. Paszner, N. C. Bahera, *Holzforschung* **39** (1985) 51 – 61.
150. L. Paszner, *Holzforschung* **43** (1989) 159 – 168.
151. L. Paszner, H. J. Cho, *Tappi J.* **72** (1989) no. 2, 135 – 142.
152. R. Patt, O. Kordsachia, *Papier (Darmstadt)* **40** (1986) no. 10 A, V 1 – V 8.
153. O. Kordsachia, R. Patt, *Holzforschung* **42**(1988) 203 – 209.
154. R. Patt, O. Kordsachia, K. Kopfmann, *Papier (Darmstadt)* **41** (1989) no. 10 A, V 108 – V 115.
155. O. Kordsachia, B. Reipschläger, R. Patt, *Paperi ja Puu* **72** (1990) 44 – 50.
156. J. Janson, R. Uvorisalo, *Paperi ja Puu* **68** (1986) 610 – 615.
157. B. Lönnberg, T. Laxén, R. Sjöholm, *Paperi ja Puu* **69** (1987) 757 – 762.
158. B. Lönnberg, T. Laxén, A. Bäcklund: *Paperi ja Puu* **69** (1987) 826 – 830.
159. T. Laxén, J. Aittamaa, B. Lönnberg, *Paperi ja Puu* **70** (1988) 891 – 894.
160. H. H. Nimz, R. Casten, *Holz Roh-Werkst.* **44** (1986) 207 – 212.
161. H. H. Nimz et al., *Papier (Darmstadt)* **41** (1989) no. 10 A, V 102 – V 108.
162. R. A. Young, E. B. Wiesmann, J. Davis, *Holzforschung* **40** (1986) 99 – 108.
163. R. Y. Young, *Tappi J.* **72** (1989) no. 4, 195 – 200.
164. K. Poppius et al., *Paperi ja Puu* **68** (1986) 87 – 92, 622 – 629.
165. J. Sundquist, *Paperi ja Puu* **68** (1986) 616 – 620.
166. J. Sundquist, L. Laamanen, K. Poppius, *Paperi ja Puu* **70** (1988) 143 – 148.
167. K. Poppius-Levlin, R. Mustonen, T. Huovila, J. Sundquist, *Paperi ja Puu* **73** (1991) 154 – 158.
168. S. A. Saponitzky: *Verwertung von Sulfitablauge,* VEB Fachbuchverlag, Leipzig 1963.
169. F. Melms, K. Schwenzon, *Verwertungsgebiete für Sulfitablauge,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1967.
170. S. Y. Lin, *Ekman-Days, Int. Symp. Wood Pulping Chem.* 1983, vol. 3, 108 – 113.
171. R. W. Coughlin, D. W. Sundstrom, H. E. Klei, E. Avni, "Conversion of Lignin to Useful Chemical Products," in D. L. Wise (ed.): *Bioconversion Systems,* C. R. C. Press, Boca Raton, Fla., 1984, 42 – 58.
172. O. V. Ingruber, *AIChE Symp. Ser.* **200** (1980) no. 76, T 196 – T 208.
173. T. Hauki, R. Reilama, *Pulp Pap. Can.* **85** (1984) T 120 – T 123.
174. R. Brännland, R. Gustafsson, B. Hultman, *Sven. Papperstidn.* **18** (1976) 591 – 594.
175. E. A. Horntvedt, *Tappi* **53** (1970) 2147 – 2152.
176. L. Stockman, L. Strömberg, F. de Sousa, *Cellul. Chem. Technol.* **14** (1980) 517 – 526.
177. I. Palenius, *Papier (Darmstadt)* **36** (1982) V 13 –V 19.
178. K. P. Kringstad, K. Lindström, *Environ. Sci. Technol.* **19** (1985) 1219 – 1224.
179. K. P. Kringstad, *Papier (Darmstadt)* **43** (1989) no. 10 A, V 12 – V 19.
180. H. Beck, K. Eckart, W. Mathar, R. Wittkowski, *Chemosphere* **17** (1988) 51 – 57.
181. S. E. Swanson, C. Rappe, J. Malmström, K. P. Kringstad, *Chemosphere* **17** (1988) 681 – 691.

182. J. J. Lindberg, *Paperi ja Puu* **80** (1988) 713 – 718.
183. C. Rappe et al., *Pulp Pap. Can.* **90** (1989) T 273 –T 278.
184. C. W. Dence, G. Annergren: "Chlorination," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 29 – 80.
185. J. Gullichsen, *Tappi* **59** (1976) no. 11, 106 – 109.
186. C. W. Dence, K. V. Sarkanen, *Tappi* **43** (1960) no. 1, 87 – 96.
187. J. B. van Buren, C. W. Dence, *Tappi* **50** (1967) 553 –560.
188. J. B. van Buren, C. W. Dence, *Tappi* **53** (1970) 2246 – 2253.
189. C. W. Dence: "Halogenation and Nitration," in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignins –Occurrence, Formation. Structure and Reactions,* Wiley Interscience, New York 1971, pp. 373 –432.
190. J. Gierer, L. Sundholm, *Sven. Papperstidn.* **74** (1971) 345 – 351.
191. R. P. Singh, E. S. Atkinson: "The Alkaline Extraction," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 81 – 100.
192. L. E. Larsen, H. deV. Partridge: "Bleaching with Hypochlorites," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 101 – 112.
193. D. B. Mutton, *Pulp Pap. Mag. Can.* **63** (1964) T 41 – T 51.
194. W. H. Rapson, G. B. Strumila: "Chlorine Dioxide Bleaching," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 113 – 157.
195. B. O. Lindgren, *Sven. Papperstidn.* **74** (1971) 57 – 63.
196. P. Axegård, *Pulp Pap. Sci.* **12** (1986) no. 3, J 67 – J 72.
197. P. Axegård, *Tappi J.* **69** (1986) no. 10, 54 – 59.
198. P. Axegård, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1987, 105 – 111.
199. P. Axegård, *Proc. Tech. Assoc. Pulp Pap. Ind.* 1988, 69 – 76.
200. D. C. Pryke, *Tappi J.* **72** (1989) no. 10, 147 – 155.
201. D. C. Pryke, *Pulp Pap. Can.* **90** (1989) T 203 – T 207.
202. W. H. Rapson, C. B. Anderson, *Pulp Pap. Mag. Can.* **61** (1960) T 495 – T 504.
203. W. H. Rapson, C. B. Anderson, *Pulp Pap. Mag. Can.* **67** (1966) T 47 – T 55.
204. W. H. Rapson, C. B. Anderson, D. W. Reeve, *Pulp Pap. Can.* **78** (1977) T 137 – T 148.
205. D. W. Reeve, W. H. Rapson, *Tappi* **64** (1981) no. 9, 141 – 144.
206. J. V. Hatton, *Pulp Pap. Mag. Can.* **68** (1967) T 181 – T 190, T 204.
207. B. J. Fergus, *Tappi* **56** (1973) no. 1, 114 – 117.
208. U. Germgård, R.-M. Karlsson, *Paperi ja Puu* **66** (1984) 627 – 634.
209. P. S. Fredricks, B. O. Lindgren, O. Theander, *Tappi* **54** (1971) 87 – 90.
210. P. S. Fredricks, B. O. Lindgren, O. Theander, *Cellul. Chem. Technol.* **4** (1970) 533 – 547.
211. I. Croon, D. H. Andrews, *Tappi* **54** (1971) no. 11, 1813 – 1898.
212. R. P. Singh, B. C. Dillner: "Oxygen Bleaching," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 159 – 209.
213. L. Almberg, I. Croon, A. Jamieson, *Tappi* **62** (1979) no. 6, 33 – 35.
214. G. L. Akim, *Paperi ja Puu* **63** (1981) 291 – 298.
215. P. J. Kleppe, *Papier (Darmstadt)* **39** (1985) no. 10 A, V 8 – V 13.
216. U. Germgård et al., *Sven. Papperstidn.* **88** (1985) R 113 – R 117.
217. T. J. McDonough, *Tappi J.* **69** (1986) no. 6, 46 – 52.
218. N. Liebergott, B. van Lierop, G. Teodorescu, G. J. Kubes, *Pulp Pap. Can.* **87** (1986) no. 5, 64 – 68.
219. L. Tench, S. Harper, *Tappi J.* **70** (1987) no. 11, 55 –61.
220. R. C. Eckert, H.-M. Chang, W. P. Tucker, *Tappi* **56** (1973) no. 6, 134 – 138.
221. J. Gierer, F. Imsgard, *Sven. Papperstidn.* **80** (1977) 510 – 518.
222. J. S. Gratzl, *Papier (Darmstadt)* **41** (1987) 120 –130.
223. O. Theander in J. S. Gratzl, J. Nakano, R. P. Singh (eds.): *Chemistry of Delignification with Oxygen, Ozone, and Peroxides,* Uni Publishers Co., Tokyo 1980, pp. 41 – 57.
224. L. Nasman, G. Annergren, *Tappi* **63** (1980) no. 4, 105 – 109.
225. E. F. Elton, V. L. Magnotta, L. D. Markham, C. E. Courchene, *Tappi* **63** (1980) no. 11, 79 – 82.
226. P. J. Kleppe, P. C. Knutsen, F. Jacobsen, *Tappi* **64** (1981) no. 6, 87 – 90.
227. K. Idner, *Tappi* **71** (1988) no. 2, 47 – 50.
228. H. Krüger, H. U. Süss, *Pulp Pap. Can.* **85** (1984) T 297 – T 299.
229. H. U. Süss, *Papier (Darmstadt)* **40** (1986) 10 – 15.

230. H. U. Süss, *Papier (Darmstadt)* **40** (1986) 150 – 153.
231. D. Helmling, H. U. Süss, J. Meier, M. Berger, *Tappi J.* **72** (1989) no. 7, 55 – 61.
232. N. Soteland, *Pulp Pap. Mag. Can.* **75** (1974) T 153 – T 158.
233. N. Soteland, *Nor. Skogind.* **32** (1978) 199 – 204.
234. R. B. Secrist, R. P. Singh, *Tappi* **54** (1971) no. 4, 581 – 584.
235. R. P. Singh, *Tappi* **65** (1982) no. 2, 45 – 48.
236. N. Liebergott, B. van Lierop: "Oxygen, Ozone and Peroxide Bleaching," Tappi Seminar Notes, New Orleans 1978, pp. 90 – 106.
237. N. Liebergott, B. van Lierop, G. J. Kubes, *Ozone: Sci. Eng.* **4** (1982) 109 – 120.
238. N. Liebergott, B. van Lierop, G. J. Kubes, *Tappi J.* **67** (1984) no. 8, 76 – 80.
239. C.-A. Lindholm, *Paperi ja Puu* **69** (1987) 211 – 218.
240. R. Patt, O. Kordsachia, D. L.-K. Wang, *Papier (Darmstadt)* **42** (1988) no. 10 A, V 14 – V 23.
241. R. C. Eckert, R. P. Singh in J. S. Gratzl, J. Nakano, R. P. Singh (eds.): *Chemistry of Delignification with Oxygen, Ozone and Peroxides,* Uni Publishers Co., Tokyo 1980, pp. 229 – 243.
242. P. J. Balousek, T. J. Mc. Donough, R. D. McKelvey, D. C. Johnson, *Sven. Papperstidn.* **84** (1981) R 49 – R 54.
243. H. Kaneko, S. Hosoya, K. Iiyama, J. Nakano, *J. Wood Chem. Technol.* **3** (1983) 399 – 411.
244. T. Eriksson, J. Gierer, *Ekman-Days, Int. Symp. Wood Pulping Chem.* 1983, vol. 4, 94 – 98.
245. A. A. Katai, C. Schuerch, *J. Polym. Sci. Part A 1,* 1966, 2683 – 2703.
246. G. Pan, C.-L. Chen, H.-M. Chang, J. S. Gratzl, *Ekman-Days 1981, Int. Symp. Wood Pulping Chem.* 1981, vol. 2, 132 – 144.
247. C.-A. Lindholm, *Paperi ja Puu* **68** (1986) 283 – 290.
248. D. H. Andrews, R. P. Singh: "Peroxide Bleaching," in R. P. Singh (ed.): *The Bleaching of Pulp,* Tappi Press, Atlanta 1979, pp. 211 – 253.
249. G. Gellerstedt, I. Pettersson, S. Sundin, *Ekman-Days 1981, Int. Symp. Wood Pulping Chem.* 1981, vol. 2, 120 – 124.
250. VDP Papier 89: *Zahlenreihen zur Entwicklung der Zellstoff- und Papierindustrie in der BRD,* Verband deutscher Papierfabriken, Bonn 1989.
251. "Holzstoff – Halbstoff mit Zukunft," *Papier (Darmstadt)* **37** (1983) no. 10 A, V 11.
252. American National Standards, Tappi, Atlanta 1989.
253. *Scan Norm,* Testing Committee, Stockholm.
254. *Merkblatt VI/1/66,* Verein der Zellstoff- und Papier-Chemiker und -Ingenieure, Darmstadt.
255. W. Brecht: *Der Versuchsschleifer im Dienste der Holzschlifforschung,* Voith Forschung und Konstruktion, Heft 8, 1962.
256. F. G. Powell, F. Luhde, K. C. Logan: "Supergroundwood by Grinding," *Pulp Pap. Mag. Can.* **66** (1965) no. 8, T 399.
257. M. Aario, P. Haikkala, A. Lindahl: "Pressure Grinding – A New Method to Produce Mechanical Pulps," *CPPA Annual Meeting,* Montreal 1979.
258. A. Meinecke, R. Süttinger, K. Schmidt: "Das Thermoschleifen," *Wochenbl. Papierfabr.* **112** (1984) no. 9, 307.
259. *Ullmann,* 4th ed., **17,** 561.
260. A. Meinecke: "Über die Technologie und die Maschinen zur Holzstofferzeugung," *Wochenbl. Papierfabr.* **100** (1972) no. 9, 309.
261. R. Süttinger: "Ein Beitrag zur Kenntnis der Temperaturverhältnisse beim Schleifprozeß," *Wochenbl. Papierfabr.* **113** (1985) no 7, 247.
262. A. Atack, J. T. Pye: "The Measurement of Grinding Zone Temperature," *Pulp Pap. Mag. Can.* **65** (1964) Sept. T 363.
263. D. A. J. Goring: "Thermal Softening of Lignin, Hemicellulose and Cellulose," *Pulp Pap. Mag. Can.* **64** (1963) no. 12, T 517.
264. D. K. Alexander: "The Mechanism of Wood Grinding and Relationship between Energy and Fiber Quality," 7. International Mechanical Pulp Conference, Ottawa, June 1971.
265. D. Atack, L. R. Heffell: "High Speed Grinding of Eastern Black Spruce," *Pulp. Pap. Mag. Can.* **73** (1972) no. 9, T 78.
266. R. Süttinger: *Die Technologie des Schleifprozesses,* Voith Forschung und Konstruktion, Heft 26, 1979.
267. K. H. Klemm: *Neuzeitliche Holzschlifferzeugung,* Sändig Verlag, Wiesbaden 1957, p. 106.
268. M. Loly: *Pulpstone Sharpening Manuel,* Norton International Inc., Worcester 1978.
269. G. Schweizer, N. Pancur, P. Hell: "Die Holzschliffqualität als Funktion des Steinprofils, *Wochenbl. Papierfabr.* **108** (1980), no. 11/12, 361.
270. K. H. Klemm: "Technik und Wirtschaftlichkeit der mechanischen Gewinnung von Faserstoffen aus Nadelholz," *Wochenbl. Papierfab.* **106** (1978) no. 11/12, 405.
271. A. Kärnä: "Pressure Groundwood, Status of Development and Use," *Eucepa Meet,* München Oct. 1980.

272. M. Aario, P. Haikkala: "Pressure Grinding is Proceeding," *Tappi* **63** (1980) no. 2, T 139.
273. D. Atack, J. Fontebasso, M. J. Stationvala: "Die Vorgänge beim Schleifen in einer Druckatmosphäre," *Papier (Darmstadt)* **35** (1981) no. 9, 397.
274. H. P. Sollinger: *Die Temperatur und der Druck als Einflußgrößen auf die Schliffqualität beim Holzschleifverfahren,* Dissertation, TH Aachen 1982.
275. J. Seehofer, G. Lippert: "Einjährige Erfahrungen mit einem Druckschleifer," *Wochenbl. Papierfabr.* **111** (1983) no. 17, 600.
276. R. Süttinger: *Die Weiterentwicklung des Holzschleifverfahrens,* Voith Forschung und Konstruktion, Heft 31, 1985, p. 9.
277. F. Luhde: "Der Aufschluß des Holzes beim Schleifverfahren und beim Hackschnitzel-Refinerverfahren," *Papier (Darmstadt)* **16** (1962) no. 11, 655.
278. D. Atack: "On the Characterization of Pressurized Refiner Mechanical Pulps," *Sven. Papperstidn.* **75** (1972) no. 3, T 89.
279. W. D. May: "A Theory of Chip Refining – The Origin of Fiber Length," *Pulp Pap. Mag. Can.* **74** (1973) no. 1, T 70.
280. D. Atack, W. D. May: *Mechanical Reduction of Chips by Double-Disk Refining,* Paper and Pulp Research Inst. Canada, Dec. 1962.
281. B. Ferrari, O. Danielsson: "Energierückgewinnung in TMP- und CTMP-Systemen," *Wochenbl. Papierfabr.* **114** (1986) no. 18, 739.
282. B. Engstrom, M. Jackson: *Heat Recovery Aspects of Sunds Defibrators TMP/CTMP Systems,* Sunds Defibrator AB, Sweden, 1984.
283. M. Jackson: *The Manufacture, Physical Properties and End-Uses of High-Yield Pulps,* Sunds Defibrator AB, Sweden 1988.
284. H. Münster: "Thermo-Mechanischer HolzstoffFaserstoff der Zukunft," *Wochenbl. Papierfabr.* **105** (1977) no. 5, 155.
285. B. R. Stein: *A Practical Approach to Refiner Plate Optimization,* vol. 1, nos. 1 – 3, J. & L. Plate, Optima, Wis., 1988.
286. G. äkerlund, M. Jackson: *Manufacture and End-Use Application of CTMP and CMP from Softwoods and Hardwoods,* Sunds Defibrator AB, Sweden, 1984.
287. J. V. Hatton, S. S. Johal: "Chemimechanical Pulps from Hardwood/Softwood Chip Mixtures," *Pulp Pap. Can.* **90** (1989) no. 3, T 91.
288. J. Kurdin: "Refinerbased Chemi-Thermo-Mechanical-Pulping," *Pulp Pap. Technol.* 1989.
289. H. U. Süss, W. Eul: "Hochgebleichter CTMP – ein Zellstoffersatz ?" *Wochenbl. Papierfabr.* **114** (1986) no. 9, 320.
290. S. Nikula: "On Development of Mechanical Pulp Screening," *Paperi ja Puu* **71** (1989) no. 5, T 458.
291. E. Böttger, R. Rienecker: "Neue Entwicklungen bei der Herstellung von Holzschliff hoher Qualität," *Wochenbl. Papierfabr.* **108** (1980), no. 11/12, 429.
292. F. Stolze: *Beurteilungskriterien für Sortierschaltungen in Holzstofferzeugungsanlagen,* I. M. Voith, company publ. 1987, p. 2728.
293. K. Schmidt: *Holzschliffsortierung mit Schlitzsiebkörben,* I. M. Voith, company publ. 1987, p. 2743.
294. A. W. Hooper: "The Screening of Mechanical Pulp," Mechanical Pulping Course, CPPA Technical Section, S. W. Hooper Corp., Atlanta, Ga., 1987.
295. K. Süttinger: "Analyse des Energieverbrauchs in der Holzstofferzeugung," *Wochenbl. Papierfabr.* **107** (1979) no. 2, 40.
296. *Ullmann,* 4th ed., **17,** 565.
297. K. Schröter: *Die Holzschliffbleiche,* Güntter Staib Verlag, Biberach 1976.
298. M. J. Ducey: "Mechanical Pulp Bleaching Survey Explores New Equipment Technology," *Pulp Pap.* **63** (1989) no. 66, T 130.
299. R. Bott: "Moderne Verfahrenstechnik der Holzschliffbleiche," *Wochenbl. Papierfabr.* **100** (1972) no. 21, 791.
300. G. Schweizer: "Theorie und Praxis der Holzschliffbleiche," *Wochenbl. Papierfabr.* **100** (1972) no. 11/12, 433.
301. W. Auhorn, J. Melzer: "Untersuchung von Störsubstanzen in geschlossenen Kreislaufsystemen," *Wochenbl. Papierfabr.* **107** (1979) no. 13, 493. Chapter 3
302. *Recycling of Fiber and Filler in Pulp and Paper Industry,* Proc. of EUCEPA Symposium 1989, pp. 5 – 14.
303. Paper Stock Institute of America, *Paper Stock Standards and Practices,* Circular PS-83.
304. CEPAC rue de Crayar, 14 1050 BRUXELLES.
305. Wastepaper Recycling Promotion Center in Japan: Classification Guide of Wastepaper, Tokyo, 1986. Chapter 4
306. B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981, pp. 123 – 131.
307. G. Egbers, P. Ehrler, W. Schauler: *Denkendorfer Fasertafel,* Inst. f. Textiltechnik, Reutlingen 1974. Revised by E. Kleinhansl, J. Mavely, 1984.

308. Z. A. Rogowin, W. Albrecht (eds.): *Chemiefasern,* Thieme-Verlag, Stuttgart 1984.
309. W. Albrecht: "Faserstoffe" in J. Lünenschloß, W. Albrecht (eds.): *Vliesstoffe,* Thieme-Verlag, Stuttgart 1982.
310. P. Ehrler, J. Janitza: "Wasserlösliche Polyvinylalkoholfasern für Naßvliesstoffe," Reutlinger Kolloquium "Bindung von Vliesstoffen," Reutlingen 1973.
311. Institut für Textiltechnik, Reutlingen (ed.): Reutlinger Kolloquium "Bindung von Vliesstoffen," Reutlingen/Eningen 1973.
312. P. Ehrler: "Kohäsive Verfestigung von Vliesen" in J. Lünenschloß, W. Albrecht (eds.): *Vliesstoffe,* Thieme-Verlag, Stuttgart 1982.
313. B. Scholz: "Naßverfahren" in J. Lünenschloß, W. Albrecht (eds.): *Vliesstoffe,* Thieme-Verlag, Stuttgart 1982.
314. P. Ehrler, R. Vialon: "Vliesverfestigung durch Wärme" in J. Lünenschloß, W. Albrecht (eds.): *Vliesstoffe,* Thieme-Verlag, Stuttgart 1982.
315. P. Ehrler, J. Janitza: "Qualitätskontrolle von Chemiefaser-Kurzschnitt," *Wochenbl. Papierfabr.* **102** (1974) no. 19, 731 – 736.
316. S. G. Yue, P. Ehrler: "Faserpräparationen und ihre Charakterisierung durch das Strömungspotential," Rundgespräch der Cellulosechemiker Grenzflächen und Zetapotential, Baden-Baden 1975. *Papier (Darmstadt)* **29** (1975) 547 – 555.
317. P. Ehrler, J. Janitza: "Gekräuselter Faserkurzschnitt und voluminöse naß gelegte Vliese," *Melliand Textilber. Int.* **54** (1973) 466 – 469.

General References

318. J. P. Casey: *Pulp and Paper, Chemistry and Chemical Technology,* J. Wiley & Sons, New York 1980.
319. P. Lengyel, S. Morvay: *Chemie und Technologie der Zellstoffherstellung,* Akademiai Kiado, Budapest 1973.
320. L. Placzek: *Produkte der Papierchemie,* P. Keppler Verlag KG, Heusenstamm 1973.
321. H. Kotte: *Welches Papier ist das ?* P. Keppler Verlag KG, Heusenstamm 1973.
322. G. A. Smook: *Handbook for Pulp and Paper Technologists,* Tappi Press, Atlanta 1983.
323. W. F. Reynolds: *Dry Strength Additives,* Tappi Press, Atlanta 1980.

Specific References

324. S. B. Humphrey, US 3 770 577, 1973.
325. W. Felsch et al.: *Papiermacher Taschenbuch,* 5th ed., Curt Haefner Verlag, Heidelberg, pp. 65 – 73, 87 – 107.
326. "Altpapier – Probleme der Rohstoff-Rückgewinnung bei der Beschaffung, bei der Aufbereitung und beim Einsatz," *APV-Jahrestreffen, 1975,* Güntter-Staib-Verlag, Biberach.
327. L. Göttsching (ed.): *Papier in unserer Welt – ein Handbuch,* Econ Verlag, Düsseldorf 1990, pp. 128 – 145.
328. H. Holik: "Towards a Better Understanding of the Defibring Process," *Tappi Eng. Conf., (Proc.),* Chicago, Sept. 1988, p. 223.
329. W. Siewert: "Die Auflösetechnologie im Stoffdichtebereich um 15 % und ihre Ergebnisse," *Papier (Darmstadt)* **38** (1984) no. 7, 313.
330. *Ullmann,* 4th ed., **17,** 591.
331. Sulzer-Escher Wyss, Entstipper, Prospect no. 23.12.32.
332. H. Selder: "Fibre Fractionation – a New Method of Optimizing Paper Quality and Reducing Raw Material and Energy Costs," Sulzer Escher Wyss, Report no. 23.71.32 e; "Fraserfraktionierung – Ein neuer Weg zur Optimierung der Papierqualität bei gleichzeitiger Senkung der Rohstoff- und Energiekosten," *Papier (Darmstadt),* **38** (1984) no. 9, 435.
333. Ch. McCarthy: "Factors Affecting Pressure Screen Capacity," *Tappi Eng. Conf.,* (Proc.), New Orleans, Sept. 1987, p. 351.
334. Ph. Goldenberg: "Recent Developments in Screening," *Tappi Pulping Conf., (Proc.),* 1989, pp. 73 –75.
335. R. Rienecker, J. Nittke: "Konstruktionsmerkmale von Drucksortierern," *Wochenbl. Papierfabr.* **19** (1988) 831.
336. S. Siewert: "Function and Significance of a New Type of Tailing Screen for Stock Preparation," *Tappi Pulping Conf., (Proc.),* 1984; "Funktion und Bedeutung eines neuartigen Endstufensortierers für die Stoffaufbereitung," *Papier (Darmstadt)* **38** (1984) no. 10, 485.
337. W. H. Siewert, G. R. Horsch: "Operating Results and Experience with the Escher Wyss Vario Split," *Tappi Pulping Conf., (Proc.),* San Francisco 1984, p. 611.
338. E. Linck, H. Holik: "Advanced Technology for Secondary Fibre Treatment in Tissue Systems," *Svenska Tröforskiningsinstituted*

(STFI) Tissue Machine Symposium, Stockholm, June 1983.
339. W. Matzke, H. Selder: "Various Approaches for Understanding and Improving Secondary Fiber Brightness, *Tappi Pulping Conf., (Proc.),* New Orleans 1988, p. 203.
340. E. Linck, H. Mayr, W. H. Siewert: "Der ‚Vario-Split', eine Maschine für die Stoffaufbereitung mit vielfältigen Einsatzmöglichkeiten," *Papier (Darmstadt)* **38** (1984) no. 9, 441.
341. H. E. Ortner: *Recycling of Papermaking Fibers – Flotation Deinking,* Tappi Press, Atlanta 1981.
342. E. Linck: "Ink and Speck Removal Efficiency – A Matter of the Right Flotation Cells and System Layout," *Paper Industry Technical Association (PITA) Annual Conference,* Manchester, March 1990.
343. W. Matzke, W. Siewert, W. Bassler: "The Purgomat – a Gray Water Clarification System of the Future," *Tappi Pulping Conf., (Proc.),* Washington 1987, p. 157.
344. C. Bülow, U. Hamm, L. Göttsching: "Die Flotationsanlage Purgomat," *Wochenblatt Papierfabr.* **6** (1989) 217.
345. K. Eisenschmid, B. Steinbeis: "Kreislaufreinigung in einer Deinking-Anlage," *Wochenbl. Papierfabr.* **14** (1989) 645.
346. J. W. J. Ferguson: "Theoretical Aspects of a Pulp Suspension Flowing in a Conventional Hydrocyclone," *Tappi J.* **71** (1988) no. 1, 125.
347. Ch. W. Bouchillon, A. Franko: "Comparison of Hydrocyclone Designs for Speck Dirt Removal Compared with Shive Removal," *Tappi Papermakers Conf., (Proc.),* Denver, April 1985.
348. J. C. Kerr: "Use of Hydrocyclones for Cleaning Mechanical Pulp," *Tappi Pulping Conf., (Proc.),* Houston 1983.
349. D. A. Hoffmann, C. M. Reid: "Improved Hydrocyclone Cleaners for the Pulp and Paper Industry," *Tappi J.* **56,** (1973) no. 4, 106.
350. B. Ullmann: "Praktische Erfahrungen bei der Projektierung und beim Betrieb von Cleaneranlagen," *Zellst. Pap. (Leipzig)* **38** (1989) no. 2, 63.
351. G. Zachrisson: *Hot Dispersion of Laserprint Particles and the Implication on Waste Paper,* Pira Paper and Board Division, March 1989.
352. E. Linck: "Überlegungen zur Theorie der Dispergierung," *Papier (Darmstadt)* **35** (1981) no. 2, 67.
353. H. Selder: "Erfahrungen mit dem Escher Wyss Disperger," *Wochenbl. Papierfabr.* **7** (1978) 269.
354. The Institute of Paper Chemistry: *Intenational Symposium on Fundamental Concepts of Refining,* Appleton, Wis., Sept. 1980.
355. S. Hietanen, K. Ebeling: "A New Hypothesis for the Mechanics of Refining," *Pap. Puu* **72** (1990) 172.
356. S. Hietanen, K. Ebeling: "Fundamental Aspects of the Refining Process," *Pap. Puu* **72** (1990) 158.
357. R. J. Kerekes: "Characterization of Pulp Refiners by a C-Factor," *Nordic Pulp and Paper Research Journal* **1** (1990) 3.
358. H. Schnell: "Büttenpropeller oder Rührwerk?" *Papier (Darmstadt)* **8** (1984) 393.
359. W. Felsch et al.: *Papiermacher Taschenbuch,* 5th ed., Curt Haefner Verlag, Heidelberg, pp. 120, 173, 184.
360. L. Göttsching (ed.): *Papier in unserer Welt – Ein Handbuch,* Econ Verlag, Düsseldorf 1990, p. 141.
361. "Die Herausforderung – Geschwindigkeitssteigerung einer Papiermaschine," *APV-Jahrestreffen 1984,* Güntter-Staib-Verlag, Biberach.
362. H. Holik: "Improving Paper Quality by Optimum CD Profile Control," *Proceedings Control Symposium,* Stockholm, May 1986, pp. 87.
363. W. S. Miller and G. R. Oleson: "Wet-End Pressure Pulsations: Case Histories, *Tappi J.* **71** (1988) no. 9, 93.
364. B. Norman, A. Tegengren: "Stock Preparation – a Key to Grammage Control," *Pap. Technol.* **31** (1990) no. 1, 42.
365. Y. Reijonen: "Quality Control Starts in the Short Circuit System," *Pap. Technol.* **30** (1989) no. 12, 10.
366. R. Leadbeater: "Wet-End Barring on a Multi-Ply Machine," *Pap. Technol.* **31** (1990) no. 1, 30.
367. H. Holik, H.-J. Schultz: "How to Avoid Basis Weight Distribution Irregularities," *Pulp Pap. Can.* **80** (1979) no. 12, 101.
368. H. Kießling: "Stoffauflauf und konstanter Teil," *Papier (Darmstadt)* **7** (1987) 370.
369. M. H. Waller: "Recent Developments in Headboxes," *Tappi J.* **70** (1987) 33.
370. H. Holik, E. Weisshuhn, H. Dahl: "Influence of Headbox, Flow Conditions on Paper Properties and Their Constancy," *Tappi Eng. Conf., (Proc.),* New Orleans, Sept. 1987, p. 469.

371. R. Kurtz, W. Seebass: *Affects of Headbox Design on Web Formation and Product Quality – an Update on the Escher Wyss Step Diffusor Headbox,* Pira Paper and Board Division, Nov. 1989.
372. J. Hujala, T. Malmberg: "Roll and Blade Forming Proves to be Unbeatable Both in Pilot and Full Scale," *42nd Appita Annual General Conference,* Hobart, Australia, 1988.
373. R. Bluhm, A. Bubik, H. Holik: "Multi-Layer and Multi-Ply Paper Production," *EUCEPA Conference,* Harrogate, May 1988, p. 161.
374. A. Thorp: "Twin-Wire Forming – Rapid Commercialization Ahead," Wet End Operations, Tappi Seminar Notes, 1985.
375. K. Müller, A. Bubik: "Blattbildung mit dem Twinformer," *Wochenbl. Papierfabr.* **11/12** (1989) 470.
376. "Blattbildung – Theoretische Aspekte und praktische Erfahrungen," *APV-Jahrestreffen 1977,* Güntter-Staib-Verlag, Biberach.
377. *Naßpartie-Symposium, München 1981,* Güntter-Staib-Verlag, Biberach.
378. "Rund um das Langsieb," *APV-Jahrestreffen 1987,* Güntter-Staib-Verlag, Biberach.
379. B. Wahlstrom: "Opportunities in Pressing", part 1, *Tappi* **64** (1981) no. 1, 75; Part 2, *Tappi* **64** (1981) no. 2, 57.
380. G. Carlsson, T. Lindström, B. Norman: "Some Basic Aspects on Wet Pressing of Paper," *J. Pulp Pap. Sci.* **9** (1983) no. 7, 101.
381. Szikla, Z., and H. Paulapuro: "Changes in Z-Direction Density Distribution of Paper in Wet Processing" *J. Pulp Pap. Sci.* **15** (1985) no. 1, 11.
382. H. Holik, H.-J. Schultz: "The Nipco Roll System: Automatic Moisture Profile Control," *Pulp Pap. Can.* **80** (1979) no. 4, 88.
383. "Die Bedeutung des Naßpressens für Produkt und Produktivität," *APV-Jahrestreffen 1988,* Güntter-Staib-Verlag, Biberach.
384. T. Gudehus: "Stoffentwässerung im Walzenpreß-spalt," part I, *Papier (Darmstadt)* **42** (1988) no. 4, 174; "Entwässerungsleitung," part II *Papier (Darmstadt)* **42** (1988) no. 5, 221; "Einflußfaktoren," part III *Papier (Darmstadt)* **42** (1988) no. 7, 362.
385. H. Dahl, W. Schuwerk: "Der Entwicklungsstand bei Preßwalzen unter besonderer Berücksichtigung der NIPCO-Walzen," *Papier (Darmstadt)* **36** (1982) no. 7, 327.
386. H. Holik, R. Lehmann: "The Escher Wyss Nipco Roll," *Pulp Pap. Can.* **75** (1974) no. 10, T 355.
387. F. El-Hosseiny: "Compression Behavior of Press Felts and Wet Webs," *Nord. Pulp & Pap. Res. J.* (1990) no. 1, 28.
388. P. Mirsberger, H. Holik: "New Concepts and Trends in Wet Pressing," *World Pulp and Paper Technology* **1** (1989) 195.
389. "Trocknung von Papier," *APV-Jahrestreffen 1973,* Güntter-Staib-Verlag, Biberach.
390. S. T. Han: "Drying of Paper," *Tappi* **53** (1970) no. 6, 1034.
391. W. A. Spraker: "Heat and Mass Transfer from Cylinder to Paper and within the Paper during Drying," *Paper Trade J.* **5** (1970) 28.
392. M. Soininen, M. Karlsson: "The Drying Characteristics of the Paper Machine," *Proceedings of the 3rd International Drying Symposium,* vol. 2, Drying Research, Wolverhampton 1982, p. 351.
393. M. Kobari, Y. Shimizu, M. Endo: "Contact Drying of Fibrous Sheets," *Proceedings of the 3rd International Drying Symposium,* vol. 1, Drying Research, Wolverhampton 1982, p. 483.
394. J. Mardon, J. Acford: "Size Press Operation," *Paper* (1973) part 1, p. 154; part 2, p. 210; part 3, p. 281.
395. H. Wurster: "Leimpresse und Geschwindigkeitssteigerung der Papiermaschine," *Wochenbl. Papierfabr.* **5** (1985) 165.
396. R. H. Crotogino: "Calendering: Making a good impression," *Pulp Pap. Can.* **90** (1989) no. 7, 107 TS.
397. H. Holik, H. Hess: "Modern Tools for High Quality On-Line Calendering," *Pap. Technol.* **30** (1989) no. 3, 34.
398. R. Kurtz, H. Hess: "Soft Nip Calendering – Latest State of the Art for a Wide Range of Grades," *CPPA Annual Meeting,* Feb. 1990, B 361.
399. H. Holik, H. Hess, W. Schuwerk: "CD Caliper Control in Calenders," *Tappi Eng. Conf., (Proc.),* Atlanta, Sept. 1989, p. 95.
400. J. Zwart, W. R. Farrel: "Oxbow Effect and Surface Temperature Profiles of Calender Rolls," *Proc. CPPA Newsprint Conference 1989,* Quebec, Canada.
401. G. Müller-Mederer: "Maschinen für Glätteerzeugung," *Wochenbl. Papierfabr.* **3** (1970) 41.
402. W. Stotz: "Floating-Stack, seine Anwendung und Vorteile bei der Papierherstellung," *Papier (Darmstadt)* **42** (1988) no. 10 A, V 184.
403. W. Stotz, W. Schuwerk: "New Generation of Supercalenders," *Tappi J.* **72** (1989) no. 7, 95.

404. M. V. Tuomisto: "Supercalendering in Europe and North America: Getting Closer, but Differences Still Exist," *Tappi J.* **72** (1989) no. 1, 68.
405. E. P. Laumer: "Troubleshooting at the Reversing Nip of a Supercalender," *Tappi J.* **70** (1987) no. 5, 175.
406. K. H. Hilden, D. Sawley: "Calender Steam Showers, An Effective New Way of Hot Calendering," *Tappi J.* **70** (1987) no. 7, 87.
407. H.-L. Dalpke in L. Göttsching (ed.): *Papier in unserer Welt,* Econ-Verlag, Düsseldorf 1990.
408. W. Brecht, H.-L. Dalpke: *Wasser, Abwasser, Abwasserreinigung in der Papierindustrie,* Güntter-Staib-Verlag, Biberach 1980.
409. K. Mörch, J. Hartmann, *Wochenbl. Papierfabr.* **118** (1990) no. 2, 55 – 61.
410. H. L. Baumgarten, U. Welcker, *Papier (Darmstadt)* **40** (1986) no. 10 A, V 181 – V 192.
411. F. Börner, H.-L. Dalpke, A. Geller, L. Göttsching, *Wochenbl. Papierfabr.* **110** (1982) no. 9, 287 – 294.
412. A. Geller, W. Lüttgen, *Papier (Darmstadt)* **35** (1981) no. 10 A, V 151 – V 159.
413. W. Auhorn, *Wochenbl. Papierfabr.* **112** (1984) no. 2, 37 – 48.
414. A. Geller, *Wochenbl. Papierfabr.* **112** (1984) no. 2, 49 – 58.
415. K. Diedrich, *Wochenbl. Papierfabr.* **112** (1984) no. 4, 116 – 120.
416. A. Geller, *Tappi* **65** (1982) no. 9, 97 – 101.
417. M. B. Jansson, *For. Prod. J.* **32** (1982) no. 6, 39 – 46.
418. W. Baßler, W. Matzke, W. H. Siewert, *Wochenbl. Papierfabr.* **116** (1988) no. 10, 381 – 385.
419. C. Bülow, U. Hamm, L. Göttsching, *Wochenbl. Papierfabr.* **117** (1989) no. 6, 217 – 224.
420. H.-L. Dalpke, L. Göttsching, *Wochenbl. Papierfabr.* **102**(1974) no. 19, 721 – 730.
421. G. Eischen, J. D. Keenan, *Tappi* **68** (1985) no. 9, 104 – 107.
422. R. Saunamäki, *Pap. Puu* **71** (1989) no. 2, 158 – 164.
423. H. Schneeberger, R. Weisz, *Abwassertechn. Abfalltech.* **38** (1987) no. 4, 21 – 23.
424. M. Erken, G. Ritter, *Wochenbl. Papierfabr.* **115** (1987) no. 18, 810 – 814.
425. M. Morper, *APR Allg. Pap. Rundsch.* 1990, no. 20, 584 – 587.
426. J. Pearson, *PPI Pulp Pap. Int.* **30** (1988) no. 3, 28 – 30.
427. D. Z. Maat, L. H. A. Habets, *Pulp Pap. Can.* **88** (1987) no. 11, 410 – 414.
428. C. H. Möbius, *APR Allgem. Pap. Rundsch.* 1984, no. 24 – 25, 718 – 730.
429. G. Papier, R. Wittal, *APR Allgem. Pap. Rundsch.* 1990, no. 20, 590 – 594.
430. D. Holle, *Wochenbl. Papierfabr.* **118** (1990) no. 3, 100 – 106.
431. B. Cornely: *Das Färben von Papier,* Springer-Verlag, Berlin 1951.
432. L. Placzek: *Produkte der Papierchemie,* **vol. "Anwendungssystematik, "** P. Keppler-Vlg., Heusenstamm 1973, p. 176.
433. H. Hentschel: *Chemische Technologie der Zellstoff- und Papierherstellung,* 2nd ed., VEB Fachbuchvlg., Leipzig 1962, pp. 119, 421.
434. J. P. Casey: "Papermaking," *Pulp and Paper,* 2nd ed., vol. 2, Interscience, New York 1961, p. 1190.
435. Bayer Farben Revue, Färben von Papier, Leverkusen 1984.
436. Sandoz Papierratgeber, Einführung in die Papierfabrikation, Basel 1975.
437. BASF Ratgeber für die Verwendung von BASF-Erzeugnissen in der Papierindustrie, Ludwigshafen.
438. G. Martin, E. Arnold: "Technische Fortschritte beim Färben von Tissue," *Wochenbl. Papierfabr.* **116** (1988) no. 15, 617 – 622.
439. J. S. Arnold, W. Günder, A. Moormann-Schmitz: "Verhalten von anionischen und kationischen Direktfarbstoffen bei saurer und neutraler Fahrweise mit Clay und Kreide," *Wochenbl. Papierfabr.* **113** (1985) no. 8, 272 – 278.
440. U. Beck, F. Müller, P. Mummenhoff: "Färben und Aufhellen – Echtheits- und ökologieaspekte im Neutralbereich," *Wochenbl. Papierfabr.* **111** (1983) no. 15, 529 – 534.
441. DIN 6730, Papier und Pappe; Begriffe, Aug. 1985.
442. ISO 4046, Paper, Board, Pulp and Related Terms; Vocabulary, Nov. 1978.
443. *Unser Papier,* 1st ed., Schriftenreihe des Verbandes Deutscher Papierfabriken, 1988.
444. Statistisches Bundesamt, Warenverzeichnis für die Außenhandelsstatistik, Metzler-Poeschel, Stuttgart 1990.
445. L. Göttsching: *Papier in unserer Welt,* Econ Verlag, Düsseldorf, 1990.
446. DIN 6723, part 1, Papiere für die Datenverarbeitung; 90-g/m^2-Papier für Belegsortierleser; ungestrichen, unbeschichtet; Anforderungen, Prüfung, Nov. 1981.

447. DIN 6724, part 1, Papiere für die Datenverarbeitung; 80-g/m^2-Papier für Seitenleser; ungestrichen, unbeschichtet; Anforderungen, Prüfung, Nov. 1981.
448. RAL-RG 473, Echt Pergament; Gütesicherung, Dec. 1982.
449. Bureau Central de la Comission Electrotechnique International, Specification for Cellulose Papers for Electrical Purposes, IEC Publication no. 554-3-3, Genf 1980.
450. DIN-EN 233, Wandbekleidung in Rollen; Festlegungen für fertige Papier-, Vinyl- und Kunststoffwandbekleidungen, Oct. 1989.
451. J. M. Banks: "Mass Deacidification at the National Library of Canada," *Annual Meeting of the Society of American Archivists,* Washington, D. C., Sept. 1, 1984.
452. W. J. Barrow: *Deterioration of Book Stock – Causes and Remedies,* The Virginia State Library, Richmond, Va., 1959.
453. C. J. Shahani, W. K. Wilson, "Preservation of Libraries and Archives," *Am. Sci.* **75** (1987) 240 – 251.
454. B. L. Browning: *Analysis of Paper,* Marcel Dekker, New York, 1977, chap. 25.
455. R. D. Smith, W. O. Wozniak: *Comparison of Paper Condition in Identical Copies of Dutch Imprints from the New York Public Library and the Royal Library of Holland,* study in progress, Aug. 1988.
456. G. Hall: "Permanence of Paper," *Pap. Trade J.* **82** (1926) 185.
457. W. J. Barrow, "Restoration Methods," *Am. Archivist* **6** (1943) 151.
458. D. N.-S. Hon: "Critical Evaluation of Mass Deacidification Processes for Book Preservation," in "Historic Textile and Paper Materials III" *ACS Symp. Ser.* 1989 no. 410, chap. 2.
459. Ontario Research Foundation, US 2 033 452, 1936 (O. J. Schierholtz).
460. R. D. Smith: "Deacidifying Library Collections: Myths and Realities," *Restaurator* **8** (1987) no. 2/3, 69.
461. A. D. Baynes-Cope: "The Non-Aqueous Deacidification of Documents," *Restaurator* **1** (1969) 1.
462. R. D. Smith: "Paper Deacidification: A Preliminary Report," *Lib. Q.* **36** (1966) no. 4, 273 – 292.
463. R. D. Smith, *Can. Lib. J.* **36** (1979) 326.
464. Y. P. Kathpalia: "Deterioration and Conservation of Paper, IV. Neutralization," *Indian Pulp Pap. J.* **17** (1 962) 230.
465. W. H. Langwell: "Vapour Phase Deacidification: A Recent Development," *J. Soc. Achivists* **4** (1973) 597.
466. J. C. Williams, G. B. Kelly, Jr.: "Research on Mass Treatments in Conversation," *Bull. Am. Inst. Cons.* **14** (1974) 69.
467. M. Howe: "The Book Preservation Associates Process," *Mid-Atlantic Regional Archives Conference,* New Brunswick, N. J., May 19, 1990.
468. D. K. Sebera, P. G. Sparks: "The Library of Congress DEZ Gas Diffusion Deacidification Process," *TAPPI Paper Preservation Symposium,* Washington, D. C, Oct. 21, 1988.
469. G. Morrow: "Mass Deacidification: Operational Experience at The National Archives and The National Library of Canada," *Proceedings of the 10th Annual Conference, Institute of Paper Conservation,* Oxford Univesity, April 14 – 18, 1986.
470. U. S. Congress, Office of Technology Assessment, Book Preservation Technologies, OTA-0-375,Government Printing Office, Washington, D. C., 1988.
471. D. W. G. Clements: "Paper Strengthening at the British Library," *National Library of Austria Conference on the Preservation of Library Materials,* Vienna, April 7 – 10, 1986.
472. R. S. Wedinger: "The Lithco Process," *MidAtlantic Regional Archives Conference,* New Brunswick, N. J., May 19, 1990.
473. G. Banik: "Deacidification and Strengthening of Bound Newspapers Through Aqueous Immersion," *TAPPI Paper Preservation Symposium,* Washington, D. C., Oct. 21, 1988.
474. K. Bredereck, A. Siller-Grabenstein: "Fixing of Ink Dyes as a Basis for Restoration and Preservation Techniques in Archives," *Restaurator* **9** (1988) 113 – 135.
475. H. Solbrig: "Mineralische Hilfsstoffe in der Papierindustrie", *Papiertechnische Bibliothek,* vol. 6, Sändig-Verlag, Wiesbaden 1955 – 1960.
476. J. W. Swanson: "The Science of Chemical Additives in Papermaking," *Tappi J.* **44** (1961) no. 1, 142 A–181 A.
477. T. Krause, W. Schempp: "Physikalisch-chemische Faktoren der Blattbildung," *Wochenbl. Papierfabr.* **110** (1978) no. 1, 21 – 27.
478. H. Praast, L. Göttsching: "Markierungen und Formation von Druckpapieren," *Papier (Darmstadt)* **41** (1987) no. 10 A, V 167 – V 189.

479. J. P. Casey: *Pulp and Paper Chemistry and Chemical Technology,* 2nd ed., vol. III, Interscience Publishers, New York 1961.
480. H. L. Baumgarten: "Beziehungen zwischen Druckpapier-Eigenschaften und Umgebungsklima," *Der Druckspiegel* **38** (1983) no. 9, 764 – 767, 832 – 852.
481. O. Kallmes: "Über die Grundlagen der Formation, der Struktur und der Eigenschaften von Papier," *Papier (Darmstadt)* **35** (1981) no. 10 A, V 56 – V 73.
482. B. Steenberg: "Paper as a Viscoelastic Body, part 1: General Overview," *Sven. Papperstidn.* **50** (1947) no. 6, 127 – 140.
483. Brecht, W., E. Führlbeck: "Das rheologische Verhalten von Papier verschiedenen Feuchtigkeitsgehaltes bei kurzzeitiger Zugbeanspruchung," *Papier (Darmstadt)* **13** (1959) no. 13/14, 293 – 301.
484. R. Korn, F. Burgstaller: *Papier- und Zellstoffprüfung,* 2nd ed., Springer-Verlag, Berlin 1953.
485. K. Frank: *Taschenbuch der Papierprüfung,* Eduard Roether Verlag, Darmstadt 1958.
486. R. E. Marks (ed.): *Handbook of Physical and Mechanical Testing of Paper and Paperboard,* Marcel Dekker, New York 1983.
487. Standards cited in Section 13.2
DIN 53 109, Prüfung von Papier und Pappe: Bestimmung des Abriebs nach dem Reibradverfahren, Nov.
1985. DIN 53 116, Prüfung von Papier: Bestimmung der Fettdurchlässigkeit, Sept.
1977. DIN 53 121, Prüfung von Papier und Pappe: Bestimmung der Biegesteifigkeit nach der Balkenmethode, Nov.
1974. DIN 32 122, Prüfung von Kunststoffolien, Elastomerfolien, Papier, Pappe und anderen Flächengebilden: Bestimmung der Wasserdampfdurchlässigkeit, gravimetrisches Verhalten, Nov.
1974. DIN 53 126, Prüfung von Papier: Bestimmung der Beschreibbarkeit mit Tinte, Feb. 1985. DIN 53 130, Prüfung von Papier und Pappe: Bestimmung der Feuchtdehnung, Oct. 1978. DIN 53 140, Prüfung von Papier und Pappe: Meß-grundlagen zur Bestimmung des Reflexionsfaktors, Aug.
1978. DIN 53 145, Prüfung von Papier und Pappe: Meß-grundlagen zur Bestimmung des Reflexionsfaktors, Aug.
1978. DIN 54 516, Prüfung von Papier und Pappe: Bestimmung des Spaltwiderstandes, May 1985. DIN 54 518, Prüfung von Papier und Pappe: Streifenstauchwiderstand, June
1987 DIN-EN 55 437, Prüfung von Pappe: Prüfung von Rillungen, July
1988. DIN 55 468, Packstoffe: Wellpappe, July 1986. ISO 186, Paper and Board: Sampling to Determine Average Quality,
1988. ISO 187, Paper and Board: Conditioning of Samples, 1977. ISO 287, Paper and Board: Determination of Moisture Content – Oven Drying Method, 1985. ISO 534, Paper and Board: Determination of Thickness and Apparent Bulk Density or Apparent Sheet Density, 1988. ISO 535, Paper and Board: Determination of Water Absorption – Cobb Method, 1976. ISO 536, Paper and Board: Determination of Grammage, 1983. ISO 1924, Paper and Board: Determination of Tensile Properties, 1983. ISO 1974, Paper: Determination of Tearing Resistance,
1990. ISO 2144, Paper and Board: Determination of Ash, 1987. ISO 2470, Paper and Board: Measurement of Diffuse Blue Reflectance Factor (ISO Brightness), 1977. ISO 2471, Paper and Board: Determination of Opacity (Paper Backing) – Diffuse Reflectance Method, 1977 ISO 2494, Paper and Board: Recommended Procedure for the Determination of Roughness –Constant Pressure Air-flow Method, 1974. ISO 2528, Sheet Materials: Determination of Water Vapour Transmission Rate – Dish Method, 1974. ISO 2758, Paper: Determination of Bursting Strength, 1983. ISO 2759, Board: Determination of Bursting Strength,
1983. ISO 3035, Single-faced and Single-wall Corrugated Fibreboard: Determination of Flat Crush Resistance, 1982. ISO 3036, Board: Determination of Puncture Resistance,
1975. ISO 3037, Corrugated Fibreboard: Determination of Edgewise Crush Resistance,
1982. ISO 5626, Paper: Detemination of Folding Endurance, 1978. ISO 5627, Paper and Board: Determination of Smoothness (Bekk Method), 1984. ISO 5629, Paper and Board: Determination of Bending Stiffness – Resonance Method, 1983. ISO 5636-1, Paper and Board: Determination of Air Permeance (Medium Range) – Part 1: General Method,
1984. ISO 5636-2, Paper and Board: Determination of Air Permeance (Medium Range) – Part 2: Schopper Method, 1984 ISO 6587, Paper, Board and Pulps: Determination of Conductivity of Aqueous Extracts,
1980. ISO 6588, Paper, Board and Pulps: Determination of pH of Aqueous Extracts,

1981. ISO 8226-1, Paper and Board: Measurement of Hygroexpansivity-Part 1: Hygroexpansivity up to a Maximum Relative Humidity of 68 %, 1985. SCAN P, Folding Endurance of Paper Determined by the Köhler-Molin-Tester, 1977.

Paraformaldehyde → Formaldehyde
Paraldehyde → Acetaldehyde
Parasite Infections → Anthelmintics
Parasite Infections → Chemotherapeutics
Parasite Infections → Veterinary Drugs

Parasympatholytics and Parasympathomimetics

Parasympatholytics and Parasympathomimetics act on the parasympathetic peripheral nervous system (→ Neuropharmacology, Chap. 3.1.). Neurotransmission in this system is cholinergic, i.e., the neurotransmitter is acetylcholine [*51-84-3*].

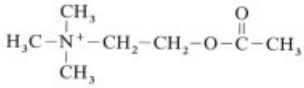

Acetylcholine

Parasympatholytics. Parasympatholytics inhibit the action of acetylcholine by competing for the acetylcholine receptor; therefore, they are also called anticholinergics, cholinolytics, or vagolytics. Atropine [*51-55-8*] is the model substance for all parasympatholytics but it is rarely used in therapy because of its narrow range of tolerance. Other important parasympatholytics are scopolamine [*51-34-3*], homatropine [*87-00-3*], tropicamide [*1508-75-4*], and ipratropium bromide [*22254-24-6*].

Parasympatholytics act as antagonists to parasympathomimetics. Atropine and other alkaloids of this type (→ Alkaloids, Chap. 4.3.) are widely used as mydriatics to dilate the pupil of the eye for diagnostic purposes (→ Ophthalmological Preparations). Anticholinergic drugs block the action of acetylcholine in the bronchi and therefore act as bronchodilators in asthma treatment (→ Antiasthmatic Agents, Chap. 2.3.). Parasympatholytics also decrease gastric secretion, which makes them useful in peptic ulcer treatment (→ Antiulcer Drugs, Chap. 3.4.3.). Atropine and related stubstances, e.g., scopolamine, are used to inhibit vomiting (→ Antiemetics, Chap. 1.). Other important uses of parasympatholytics are as spasmolytics (→ Spasmolytics) and for treatment of Parkinson's disease (→ Parkinsonism Treatment).

Parasympathomimetics. Parasympathomimetics are pharmaceuticals that act on the parasympathetic nerves in the same way as acetylcholine; they are also called cholinergic drugs. They are divided into *directly acting parasympathomimetics* and *indirectly acting parasympathomimetics*. Choline esters (e.g., carbachol [*51-83-2*] and methacholine [*55-92-5*]) and other parasympathomimetic alkaloids (e.g., pilocarpine [*92-13-7*] and muscarine [*300-54-9*]) stimulate the postsynaptic acetylcholine receptor and, therefore, act directly on the nerves. Pharmaceuticals that act indirectly inhibit the enzyme acetylcholinesterase and thus lead to the accumulation of acetylcholine. Examples of this second group of indirectly acting parasympathomimetics are physostigmine [*57-47-6*], neostigmine [*59-99-4*], distigmine bromide [*15876-67-2*], and demecarium bromide [*56-94-0*].

Parasympathomimetics decrease the rate of heart beat, increase sweat secretion (hydrotics), increase the peristaltic action of the intestine (which is used in postoperative atony of the digestive tract), and cause contraction of the pupil of the eye (which is used in treatment of glaucoma, see also → Ophthalmological Preparations).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a18_693

Parkinsonism Treatment

Wolfgang Schatton, KliniPharm, Frankfurt am Main, Federal Republic of Germany

Ralf H. Lyssy, Institut für Pharmazeutische Chemie, J. W. Goethe-Universität, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Therapy of Parkinson's Disease . 4
2.1. Centrally Acting Anticholinergic Drugs 4
2.1.1. Tropane Alkaloids 6
2.1.2. Piperidine Derivatives 7
2.1.3. Phenothiazine Derivatives 8
2.1.4. Antihistamines 9
2.2. Drugs Affecting Brain Dopamine Levels 10
2.2.1. Drugs That Increase Brain Dopamine Levels 10
2.2.1.1. L-Dopa Monotherapy 10
2.2.1.2. Peripheral Decarboxylase Inhibitors 10
2.2.1.3. L-Dopa Derivatives 11
2.2.1.4. Monoamine Oxidase Inhibitors . . . 11
2.2.1.5. Catechol-*O*-Methyl-Transferase Inhibitors 12
2.2.1.6. Monoamine Reuptake Inhibitors . . 13
2.2.2. Dopamine-Releasing Drugs 13
2.2.2.1. Direct Release 13
2.2.2.2. Dopamine Release by Glutamate Antagonists 13
2.2.3. Dopamine Receptor Agonists 14
3. New Developments 17
4. References 19

1. Introduction

Parkinson's disease (PD) is one of the major degenerative neurological diseases afflicting humans, with a worldwide prevalence exceeding 1 per 1000 population. It is, however, primarily a disease of middle-aged and elderly individuals [1]. Thus, 1 % of persons over 50 years of age are sufferers, and more than 5 % of those older than 60 years.

Early diagnosis is important if patients are to be relieved of both early and later symptoms with currently available therapies. Early diagnosis may also postpone development of the more distressing later manifestations of the disease. Although therapy alleviates symptoms, as well as restoring and sustaining a worthwhile quality of life for the patient, it does not appear to retard progression of the disease [2–5].

Clinical Diagnostic Criteria. The major clinical diagnostic criteria for PD are tremor, rigidity, akinesia, and bradyphrenia [6], [7]. Depression is the most frequent psychopathologic disorder. Reports on its frequency range from ca. 30 to 90 %, strongly supporting the hypothesis that depression in idiopathic PD is due to endogenous neurophysiological defects, particularly a reduction in the level of the neurotransmitter serotonin (5-hydroxytryptamine) in the brain.

Tremor is a major symptom; it usually begins unilaterally, often as the typical "pill-rolling" hand and arm activity. Tremor is present at rest, can be decreased or blocked by conscious will, but is aggravated by stress especially when directed motor actions are attempted.

Rigidity is reported as stiffness and cramp-like pains in the limbs. A plastic resistance to passive movement is apparent on examination that, when coexisting with tremor, may appear as a "cogwheel effect".

Akinesia is the inability to translate an intended movement into real activity. The execution of movement is slow and is associated with postural changes in the neck and trunk, as well as a characteristic shuffling gait consisting of small hesitant steps that quicken to prevent falling. Coordinated simultaneous movement of the arms and legs is impossible.

Bradyphrenia is a slowness of cognitive processing associated with apathy, impaired concentration, and finally dementia. It may correlate positively with depression.

Other characteristics of PD are an expressionless masklike face with greasy skin, increased salivation, and impaired swallowing. Handwriting may become small, and speech

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_001

difficulties produce a soft, monotonous voice. Sweating attacks, constipation, urinary frequency, and incontinence are not uncommon.

Pathophysiology of Parkinson's Disease. The central nervous system (CNS) defects associated with the development of PD are centered in the paired basal ganglia, including the striatum (caudate nucleus plus putamen), the globus pallidus (pallidum), the subthalamic nucleus, and the substantia nigra (Fig. 1) [8–13]. The latter is divided into two regions, the pars compacta and pars reticulata.

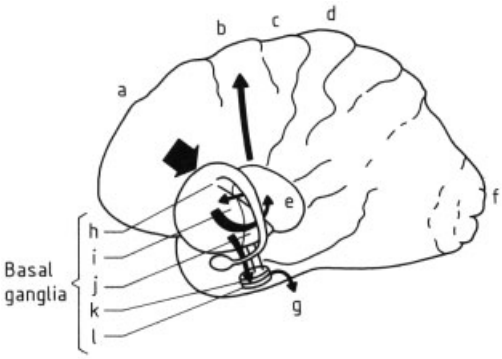

Figure 1. The basal ganglia and some of their connections [14]
a) Prefrontal cortex; b) Premotor cortex; c) Motor cortex; d) Sensory cortex; e) Thalamus; f) Visual cortex; g) Brainstem; h) Striatum (caudate nucleus); i) Striatum (putamen); j) Pallidum; k) Substantia nigra (pars compacta); l) Substantia nigra (pars reticulata)

The basal ganglia belong to the extrapyramidal tract and are responsible for the transmission of cortical motor signals associated with voluntary movement (→ Neuropharmacology, Chap. 4.4.). Parkinson's disease is one of a number of extrapyramidal disorders (i.e., neurological disorders) that arise from lesions specifically affecting the basal ganglia. Other extrapyramidal disorders include Huntington's chorea, Wilson's disease, dystonia, dyskinesia, and torticollis. Clinically, they represent combinations of abnormal voluntary movements, and disturbances in postural stability and muscle tone, and are collectively termed movement disorders.

The substantia nigra is severely damaged in PD. The basal ganglia appear to operate together with the cerebellum as interfaces between the cognitive and motor systems of the brain, converting intentions into motor actions. Destruction of any of the components of the basal ganglia, like that which occurs in the substantia nigra in PD, consequently impairs muscle function and control.

The nerve fibers of the basal ganglia are parts of input, processing, and output circuits. The input pathway from the cortex to the basal ganglia contains the largest number of neuronal fibers in the basal ganglia system. Input connections run from the cortex to the striatum (Fig. 2). The striatum in turn sends nerve fibers to dopamine-secreting nerve cells (neurons) in the substantia nigra (pars compacta). The dopamine-secreting neurons connect back to cholinergic interneurons in the striatum. The output structures of the striatum include the pallidum and the subthalamic nucleus. From here, nerve fibers project to the motor cortex via the ventrolateral thalamus to complete the overall cortex – basal ganglia – cortex loop. This sophisticated feedback control system works through a finely tuned balance of excitatory and inhibitory neurotransmitters. All the nerve fibers leading from the striatum to the pallidum and from there to the ventrolateral thalamus are in effect inhibitory pathways that secrete the inhibitory transmitter γ-aminobutyric acid (GABA). The circular pathway cortex–basal ganglia– cortex is a feedback loop that is needed to stabilize motor control. Neurotransmission from the striatum to the substantia nigra also involves inhibitory GABA secretion at the substantia nigra nerve terminals.

The darkly staining melanophore cells in the substantia nigra send out nerve axons to the striatum to form the nigrostriatal pathway and employ dopamine as an inhibitory transmitter. Both GABA and dopamine normally control their own respective brain structures unless some disease or lesion specifically damages the pathways in which they are involved.

Another significant source of input to the striatum is the corticostriatal pathway derived from the neocortex that uses glutamate as its neurotransmitter. Glutamate is an important excitatory transmitter within the basal ganglia complex and is thought to be responsible for about half of the total input to the striatum. Acetylcholine, another excitatory neurotransmitter, is secreted from interneurons within the striatum. This interconnected system is maintained in balance via neuronal feedback control loops.

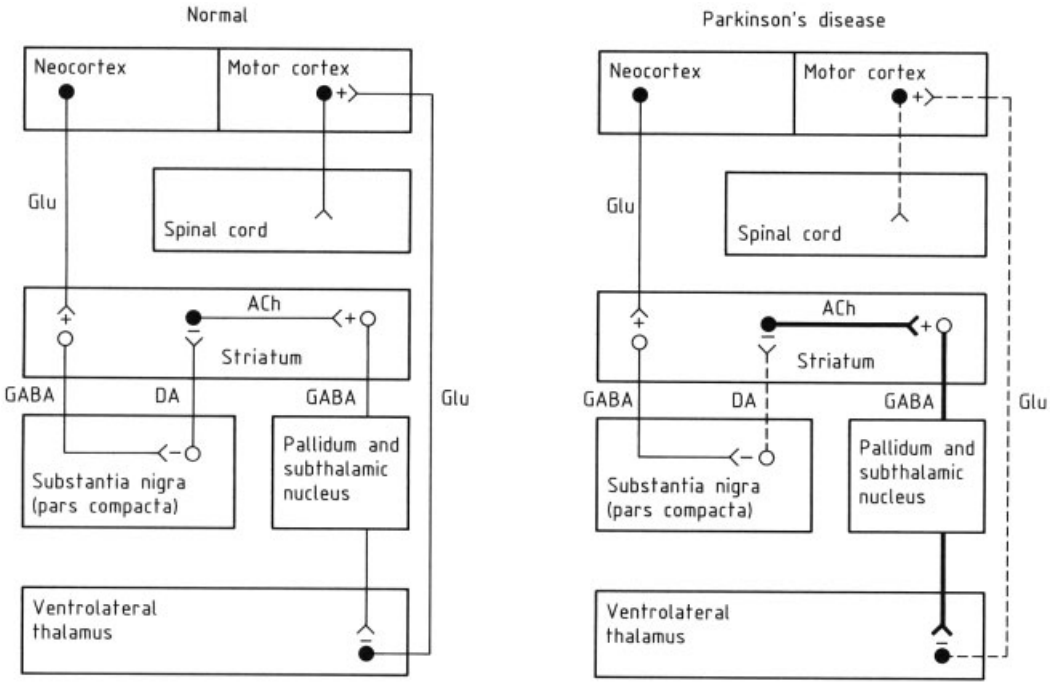

Figure 2. Schematic diagram illustrating the basal ganglia circuitry in normal humans and in patients suffering from Parkinson's disease (PD).
Excitatory (•) and inhibitory (○) neurons are shown with their axons, which form synaptic contacts with other nerve cells (+ excitatory synapse, − inhibitory synapse). Neurotransmitters secreted at the nerve terminals include γ-aminobutyric acid (GABA), dopamine (DA), glutamate (Glu), and acetylcholine (ACh). Broken and thick lines indicate reduced and increased activity, respectively. In PD, dopaminergic neurons of the substantia nigra (pars compacta) are destroyed, resulting in disinhibition (excitation) of cholinergic cells in the striatum. The overactivity of the striatal complex and its increased inhibitory output lead to a marked reduction of the activity of the thalamocortical feedback loop. These mechanisms can explain the symptoms of PD such as tremor, hypokinesia, or akinesia.

The clinical symptoms of parkinsonism are caused by destruction of the dopaminergic fibers of the darkly staining neuromelanin neurons in the substantia nigra (pars compacta) that project to the striatum (see Figs. 1 and 2). This destruction results in a dopamine deficiency. The cholinergic function in the striatum is simultaneously maintained, leading to an imbalance between inhibitory and excitatory neurotransmitters. A reduction in dopamine levels leads to excessive excitation by acetylcholine, which in turn augments the inhibitory (GABA) output of the striatum. As a consequence, the activity of the thalamocortical feedback loop is markedly reduced, leading to impaired motor function. Parkinson's disease can therefore be regarded as an overactivity of the striatal complex resulting from the loss of the fine control normally exercised by the neuromelanin-producing dopaminergic cells. By the time the first clinical symptoms (e.g., tremor or hypokinesia) become apparent, some 80 % of these pigmented substantia nigral cells have already been destroyed.

Etiology of Parkinson's Disease. The cause of parkinsonism is not known with certainty. However, it is probably triggered by some environmental insult in early to middle life. As a result, the pigmented neurons are destroyed much sooner than in normal age-dependent degeneration. Pesticides used in rural areas (e.g., paraquat) may constitute such an environmental insult. However, the disease was known long before it was described by James Parkinson in 1817; thus, other environmental factors must also be responsible.

One hypothesis links PD to lipid oxidation involving excessive formation of oxygen-

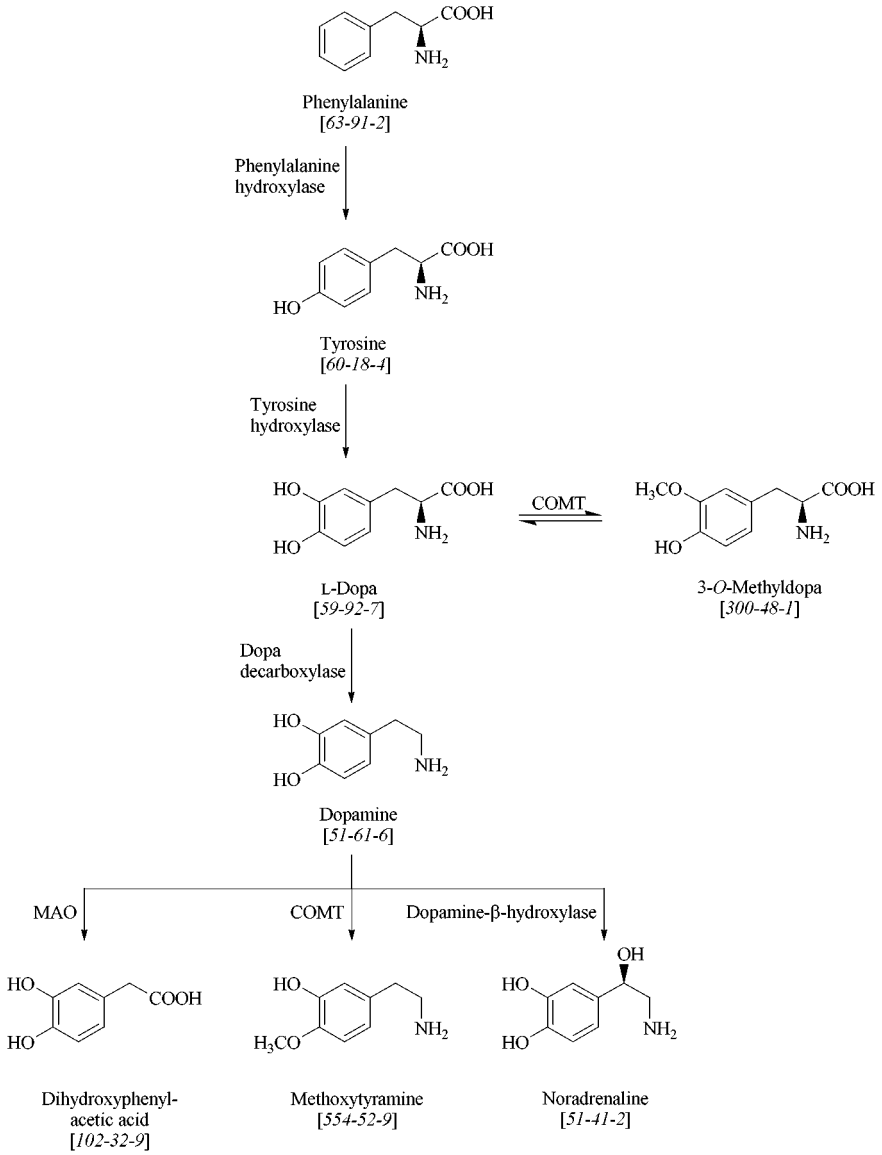

Figure 3. Biosynthesis and metabolism of dopamine
COMT = Catechol-*O*-methyl transferase; MAO = Monoamine oxidase

derived free radicals in the basal ganglia. The free radicals then progressively destroy dopamine-producing neurons under participation of the immune system. Thus, Parkinson's disease is regarded as a failure of scavenger molecules produced by the basal ganglia cells to neutralize the harmful free radical products. Further acceleration may be caused by the enzyme monoamine oxidase-B (MAO-B), which catalyzes the metabolism of catecholamine neurotransmitters and simultaneously produces reactive oxygen species. Therefore, therapeutic inhibition of MAO-B may also inhibit neurodegeneration in PD. Of major interest in this context is parkinsonism induced by 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine [*28289-54-5*] (MPTP) because of its importance for new in vitro and in vivo

pharmacological research models. This comparatively simple molecule selectively destroys dopaminergic cells in the substantia nigra and induces advanced parkinsonian symptoms in humans within a few weeks. Thus, the possibility that an environmental toxin triggers the development of PD appears reasonable.

2. Therapy of Parkinson's Disease

Modern therapy started with the discovery that the brains of Parkinson patients were deficient in dopamine. Dopamine was originally regarded as solely a precursor in the biosynthesis of noradrenaline from phenylalanine [9]. The biosynthesis and metabolism of dopamine are summarized in Figure 3. Subsequent studies led to the first clinical therapeutic use of L-dihydroxyphenylalanine (L-dopa) in 1969. Parkinson's disease was found to be due to an excess of excitatory cholinergic activity and lack of a modulatory balancing effect by inhibitory dopamine. Dopamine production progressively decreases with increasing destruction of dopaminergic neurons caused by the disease process. Therapy aims to restore the balance in several ways (Fig. 4):

1) By damping the cholinergic activity with centrally acting anticholinergic drugs (Section 2.1)
2) By administering L-dopa or its derivatives to increase brain dopamine levels (Section 2.2.1)
3) By increasing dopamine release from the nerve terminals (Section 2.2.2)
4) By using dopamine agonists to stimulate the dopamine receptors (Section 2.2.3)

2.1. Centrally Acting Anticholinergic Drugs

In PD the balance between the transmitters dopamine and acetylcholine in the striatum is disturbed in the direction of cholinergic dominance (Figs. 2 and 4). This imbalance may be corrected by using drugs that block the cholinergic system.

For more than 100 years, extracts of belladonna and, later, atropine formed the basic treatment for alleviating PD symptoms. Tinctures of deadly nightshade (*Atropa belladonna*) and henbane (*Hyoscyamus niger*) were used to reduce excessive sialorrhea and alleviate tremors and muscular stiffness.

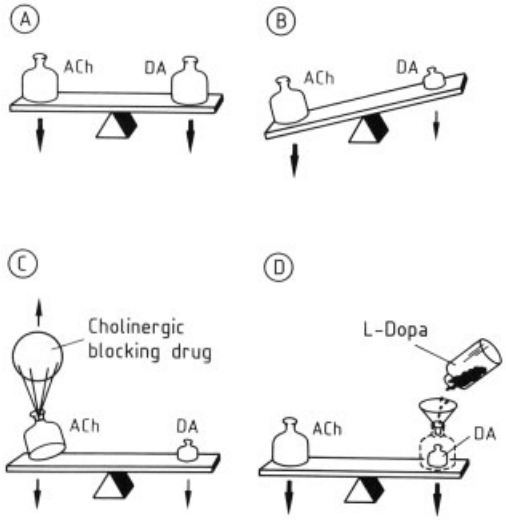

Figure 4. Diagram showing dopamine (DA) and acetylcholine (ACh) acting as antagonist neurotransmitters in the striatum [15]
A) Normal balance. B) In Parkinson's disease, dopaminergic function is reduced. The balance is disturbed in the direction of cholinergic dominance. C) The disturbed balance may be corrected by reducing the effect of acetylcholine with drugs blocking the cholinergic neurotransmitter system. D) Alternatively, the disturbed balance may be corrected by administering the immediate precursors of dopamine such as L-dopa or dopa derivatives.

When the anticholinergic properties of these plant extracts were discovered and their ability to normalize the neurotransmitter imbalance in PD was recognized, semisynthetic derivatives of naturally occurring alkaloids and synthetic anticholinergic drugs were developed. These synthetic drugs include piperidyl compounds such as biperiden, trihexiphenidyl hydrochloride, or procyclidine (Section 2.1.2). Phenothiazine derivatives (e.g., ethopropazine, see Section 2.1.3) and antihistamines (e.g., diphenhydramine or orphenadrine, see Section 2.1.4) also have some antiparkinsonian activity that is attributed to their central anticholinergic properties. They produce fewer side effects and are well tolerated, especially by elderly patients.

Today, such anticholinergics are generally used in early cases to reduce tremor and rigid-

ity. They are coprescribed with dopaminergic drugs such as L-dopa and bromocriptine. Common side effects are dryness of mouth, nausea, and visual disturbances; they may also promote dementia. Anticholinergics are contraindicated in patients with glaucoma or urinary retention and in the aged PD patient [17].

2.1.1. Tropane Alkaloids

See also → Alkaloids, Chap. 4.

Belladonna Extract. Extraction of dried leaves and roots of *Atropa belladonna* L., Solanaceae, growing in Southern and Central Europe, Asia Minor, and Algeria and cultivated in North America, yields atropine, hyoscyamine, and scopolamine. Leaves usually contain 0.3 – 0.5 % total alkaloids. Dried root contains 0.4 – 0.7 % total alkaloids [16].
Trade Name. Bellafolin (Sandoz).

Atropine [*51-55-8*], 3α(1αH,5αH)-tropanyl-DL-tropate, $C_{17}H_{23}NO_3$, M_r 289.38, crystallizes as long, orthorhombic prisms from acetone, *mp* 114 – 116 °C, and sublimes in high vacuum at 93 –110 °C, p*K* 4.35.

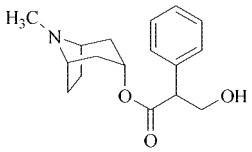

Atropine is a parasympatholytic alkaloid isolated from *Atropa belladonna* L., *Datura stramonium* L., and other Solanaceae. Synthesis is described in [18].
Trade Names. Atropina (Farmigea), Atropisol (Cooper Vision).

Scopolamine [*51-34-3*], [7(*S*)-(1α,2β,4β,5α,7β)]-α-(hydroxymethyl)benzeneacetic acid 9-methyl-3-oxa-9-azatricyclo[3.3.1.0.$_{2,4}$]non-7-yl ester; 6β,7β-epoxy-3α(1αH,5αH)-tropanyl (−)-tropate, $C_{17}H_{21}NO_4$, M_r 303.35, is a viscous liquid that forms a crystalline monohydrate, *mp* 59 °C.

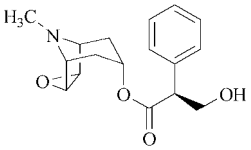

Scopolamine is a tropane alkaloid isolated from *Datura metel* L., *Scopola carniolica* Jacq., and other Solanaceae. Synthesis is described in [19–21].
Trade Names. Scopoderm TTS: Transderm-V (Ciba-Pharma).

Benztropine mesylate [*132-17-2*], 3-diphenylmethoxy-8-methyl-8-azabicyclo[3.2.1]octanemethanesulfonate, $C_{21}H_{25}NO \cdot CH_4O_3S$, M_r 403.53, crystallizes from acetone, *mp* 143 °C; it is soluble in water.

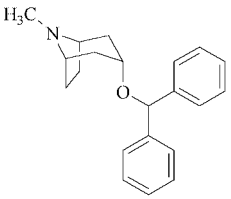

Benztropine mesylate is synthesized by reaction of diphenyldiazomethane with tropine [22].
Trade Names. Cogentin (Merck, Sharp & Dohme), Cogentinol (Astra).

Tigloidine [*495-83-0*], 2-methyl-2-butenoic acid [1α,3α(*E*),5α]-8-methyl-8-azabicyclo[3.2.1]-oct-3-yl ester, $C_{13}H_{21}NO_2$, M_r 223.31; the hydrobromide, *mp* 234 – 235 °C, is soluble in chloroform.

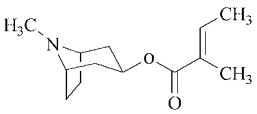

Tigloidine is isolated from *Duboisia myoporoides* R. Br., Solanaceae, or synthesized from tropine and tigloyl chloride [23].
Trade Name. Tiglyssin (Duncan, Flockhart).

2.1.2. Piperidine Derivatives

Biperiden [*514-65-8*], 1-(bicyclo[2.2.1]-hept-5-en-2-yl)-1-phenyl-3-piperidinopropanol, $C_{21}H_{29}NO$, M_r 311.45, crystals *mp* 112 – 116 °C (decomp.), sparingly soluble in water, slightly soluble in ethanol, and readily soluble in methanol. For synthesis see [24].

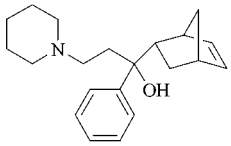

Trade Names. Akineton (hydrochloride: Knoll, Searle, Abbott, Nordmark, Yoshitomi), Tasmolin (Yoshitomi).

Trihexylphenidyl hydrochloride [*52-49-3*], (*RS*)-1-cyclohexyl-1-phenyl-3-piperidinopropanol hydrochloride, α-cyclohexyl-α-phenyl-1-piperidinepropanol hydrochloride, $C_{20}H_{32}ClNO$, M_r 337.92, crystals *mp* 258.5 °C (decomp.), free base *mp* 114.3 – 115.0 °C (decomp.), solubility (g/100 mL): water at 25 °C, 1; ethanol, 6; chloroform, 5. It is more soluble in methanol and very slightly soluble in diethyl ether. The pH of a 1 % aqueous solution is 5.5 – 6.0. For synthesis, see [25].

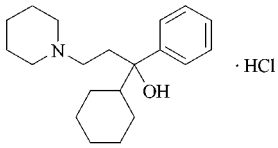

Trade Names. Artane (Lederle), Bentex (Steinhard), Broflex (Bio-Med), Pargitan (KabiVitrum), Parkinane (Lederle), Triphenidone (Toho).

Procyclidin [*77-37-2*], (*RS*)-1-cyclohexyl-1-phenyl-3-(1-pyrrolidonyl)-1-propanol, $C_{19}H_{29}NO$, M_r 287.43, crystallizes from petroleum ether, *mp* 85.5 – 86.5 °C. For synthesis, see [26].

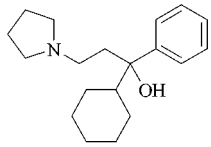

Trade Names. Hydrochloride [*1508-76-5*]: Arpicoline (R.P. Drugs), Kemadrin (Borroughs Wellcome), Osnervan (Wellcopharm).

Prodipine [*31314-38-2*], 1-isopropyl-4,4-diphenylpiperidine, $C_{20}H_{25}N$, M_r 279.43, oil *bp* (1.3 Pa) 117 – 125 °C (decomp.); the hydrochloride crystallizes from 2-propanol, *mp* 267 °C (decomp.). For synthesis, see [27].

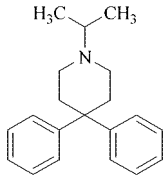

Prodipine is a CNS stimulant with antidepressant and antiparkinsonian properties.
Trade Name. Anthen (Byk Gulden).

Pridinol [*511-45-5*], 1,1-diphenyl-3-piperidino-1-propanol, α,α-diphenyl-1-piperidinepropanol, $C_{20}H_{25}NO$, M_r 295.41, crystals *mp* 120 – 121 °C (decomp.), is soluble in acetone.

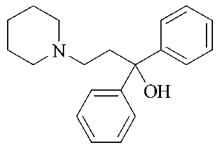

Pridinol can be prepared by Grignard reaction of ethyl 3-piperidinopropionate with phenylmagnesium bromide [28].
Trade Names. Lyseen (Hommel, Therabel, Zyma), Mitanoline (Toyo), Nonpressin (Santen), Tirashizin (Hishiyama).

Budipine [*57982-78-2*], 1-*tert*-butyl-4,4-diphenylpiperidine, $C_{21}H_{27}N$, M_r 293.46. For synthesis, see [29].

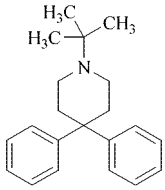

Dexetimide [*21888-98-2*], (+)-(*S*)-3-(1-benzyl-4-piperidyl)-3-phenyl-2,6-piperidinedione, $C_{23}H_{26}N_2O_2$, M_r 362.45, crystals *mp* 181–183 °C (decomp.). The dextroenantiomer is responsible for the pharmacological activity of racemic benzetimide. For synthesis, see [30].

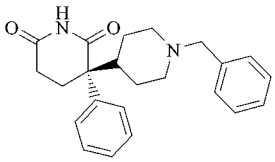

Trade Names. Serenone (Janssen), Tremblex (Janssen).

2.1.3. Phenothiazine Derivatives

Diethazine [*60-91-3*], 10-(2-diethylaminoethyl)phenothiazine, $C_{18}H_{22}N_2S$, M_r 298.44, is an oily liquid, *bp* (0.5–0.65 kPa) 195–208 °C (decomp.).

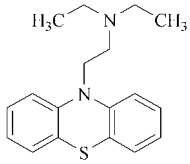

Diethazine is synthesized by reacting 10-(2-chloroethyl)phenothiazine with diethylamine in the presence of copper powder or by reacting diethylaminoethyl chloride with phenothiazine [31].
Trade Name. Deparkin (Levica).

Ethopropazine [*522-00-9*], (profenamine), (*RS*)-10-(2-diethylaminopropyl)phenothiazine, $C_{19}H_{24}N_2S$, M_r 312.46, crystals *mp* 53–55 °C (decomp.).

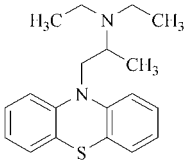

Ethopropazine is synthesized from Grignard complexes of diethylaminopropyl halide and phenothiazine [31].
Trade Names. Parkin (Yoshitomi), Parsidol (Parke Davis).

Piroheptine [*16378-21-5*], 3-(10,11-dihydro-5*H*-dibenzo[*a,d*]cyclohepten-5-ylidene)-1-ethyl-2-methylpyrrolidine, $C_{22}H_{25}N$, M_r 303.45, liquid *bp* (0.5 kPa) 167 °C (decomp.). For synthesis, see [32].

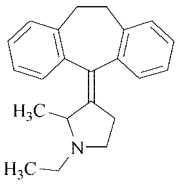

Trade Name. Trimol (Fujisawa).

2.1.4. Antihistamines

Diphenhydramine [*58-73-1*], 2-diphenylmethoxy-*N,N*-dimethylethylamine, $C_{17}H_{21}NO$, M_r 255.35, oil *bp* (0.25 kPa) 150–165 °C; the hydrochloride crystallizes from alcohol and diethyl ether, *mp* 166–170 °C, it is freely soluble in water and soluble in ethanol, chloroform, and acetone. For synthesis, see [33].

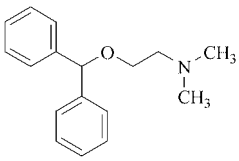

Trade Names. Allergan (Bouty), Benadryl (Parke Davis, Warner Lambert), Dibondrin (Montavit), Sekundal-D (Rorer), Somonex (Cooper), Wehedryl (Hauck).

Bietanautine [*6888-11-5*], 1,2,3,6-tetrahydro-1,3-dimethyl-2,6-dioxo-7*H*-purine-7-acetic acid complex with 2-diphenylmethoxy-*N,N*-dimethylethylamine (2 : 1), $C_{35}H_{41}N_9O_9$, M_r 731.79, crystals *mp* 168 – 170 °C (decomp.), is soluble in ethanol and sparingly soluble in water. For synthesis, see [34].

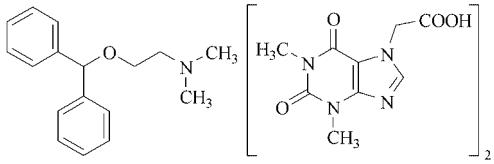

Trade Name. Nautamine (Delagrange).

Orphenadrine [*83-98-7*], (*RS*)-*N,N*-dimethyl-2-(α-2-tolylbenzyloxy)ethylamine, *N,N*-dimethyl-2-[(2-methylphenyl)phenylmethoxy]ethylamine, $C_{18}H_{23}NO$, M_r 269.37, oil *bp* (1.6 kPa) 195 °C.

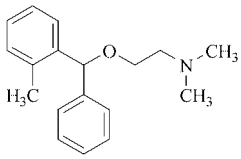

Orphenadrine is synthesized by reacting 2-methylbenzhydryl chloride with dimethylaminoethanol [35].
*Trade Names.*Hydrochloride: Biorphen (Medical), Brocasipal (Gist-Brocades), Disipal (Brocadex, Riker), Marflex (Vortech). Citrate: Banflex (O'Neal), Norflex (Riker), X-Otag (Reid-Provident).

Ethylbenzhydramine [*642-58-0*], *N,N*-diethyl-(2-benzhydryloxy)ethylamine, 2-diphenylmethoxy-*N,N*-diethylethylamine, $C_{19}H_{25}NO$, M_r 283.40, *bp* (20 Pa) 140 – 142 °C (decomp.). The hydrochloride crystallizes in prisms from ethanol and ethyl acetate, *mp* 140 °C; it is freely soluble in water and soluble in ethanol, chloroform, and acetone. For synthesis, see [36].

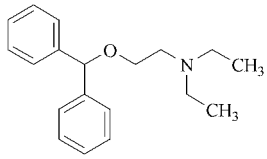

*Trade Names.*Methyl iodide: PKM (Montavit).

2.2. Drugs Affecting Brain Dopamine Levels

Drugs may increase brain dopamine levels, increase dopamine release, or act as dopamine agonists that stimulate the dopamine receptors.

2.2.1. Drugs That Increase Brain Dopamine Levels

2.2.1.1. L-Dopa Monotherapy

Administration of L-dopa to increase brain dopamine levels proved to have shortcomings due to side effects and to long-term therapeutic problems (e.g., decreasing effectiveness and alleviating the symptoms of PD but having no effect on degenerative processes).

Up to 95 % of the drug did not reach the brain, due mainly to the action of decarboxylases in the gut wall, liver, kidney, and cerebral capillaries. Formulation of L-dopa with a decarboxylase inhibitor such as benserazide or carbidopa (see Section 2.2.1.2) overcame this problem. Only one-quarter of the earlier L-dopa dosage was then required to produce comparable therapeutic benefits, at the same time minimizing gastrointestinal, cardiovascular, and other side effects.

Despite circumvention of the L-dopa decarboxylation problem and the associated benefits of lower dosage levels, long-term therapy with the drug also had disadvantages. Patients' earlier disabilities began to return, as early as one or two years after commencement of treatment. Disturbing "on – off" effects were reported, with drug benefits oscillating with periods when it had no effect. A "wearing-off" effect was observed; the initial benefit following oral L-dopa administration, which usually lasted 4 – 6 h, decreased to as little as 30 min after long-term use. More dramatic swings from mobility to immobility are unpredictable, variable, and unrelated

to the timing of oral L-dopa dose. They can, however, be influenced favorably by the infusion of amantadine sulfate.

L-Dopa [*59-92-7*], levodopa, (−)-3-(3,4-dihydroxyphenyl)-L-alanine, $C_9H_{11}NO_4$, M_r 197.19, crystallizes as colorless, odorless, and tasteless needles from water, *mp* 276 – 278 °C (decomp.). It is freely soluble in dilute hydrochloric and formic acids but practically insoluble in ethanol, benzene, chloroform, and ethyl acetate. Solubility in water is 66 mg/40 mL. In the presence of moisture, L-dopa is rapidly oxidized by atmospheric oxygen, with darkening.

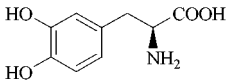

L-Dopa is the biological precursor of the catecholamines (see Fig. 3). It can be prepared from 1-3-dinitrotyrosine, 3-(3,4-methylenedioxyphenyl)-L-alanine, and L-tyrosine, and by fermentation of L-tyrosine [37–39]. A comprehensive description of pharmacology, toxicology, clinical safety, and efficacy is given in [40], [41].
Trade Names. Bendopa (ICN), Brocadopa (Merckle), Dopaflex (Egis), Eldopar (Weifa), Larodopa (Roche).

2.2.1.2. Peripheral Decarboxylase Inhibitors

A comprehensive description of the pharmacology, toxicology, clinical safety, and efficacy of peripheral decarboxylase inhibitors is given in [44–46].

Carbidopa [*28860-95-9*], (2*S*)-3-(3,4-dihydroxyphenyl)-2-hydrazino-2-methylpropionic acid monohydrate, α-hydrazino-3,4-dihydroxy-α-methylbenzhydrocinnamic acid monohydrate, $C_{10}H_{14}N_2O_4 \cdot H_2O$, M_r 244.25, crystallizes from hot water, *mp* 203 – 205 °C (decomp.). For synthesis, see [42].

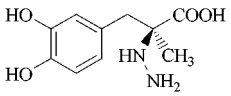

Trade Name. Combination with L-dopa: Nacom (Frosst).

Benserazide [*322-35-0*], *N′*-2,3,4-trihydroxybenzyl-DL-serylhydrazide, DL-serine[2′-(2,3,4-trihydroxybenzyl)hydrazide], $C_{10}H_{15}N_3O_5$, M_r 257.25. For synthesis, see [43].

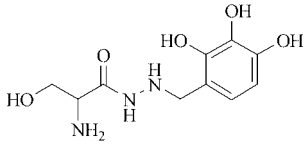

Trade Name. Combination with L-dopa: Madopar (Hoffmann-La Roche).

2.2.1.3. L-Dopa Derivatives

L-Dopa derivatives were synthesized as metabolic precursors of brain dopamine. Examples follow.

Droxidopa [*23651-95-8*], (−)-threo-3-(3,4-dihydroxyphenyl)-L-serine, $C_9H_{11}NO_5$, M_r 213.19, crystallizes from ethanol and diethyl ether, *mp* 232-235 °C (decomp.). For synthesis, see [47]. Droxidopa is a synthetic precursor of noradrenaline and dopamine.

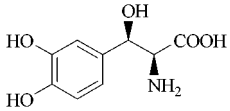

3-*O*-Methyldopa [*300-48-1*], 3-methoxy-L-tyrosine, $C_{10}H_{13}NO_4$, M_r 211.22, is a major metabolite of L-dopa in humans (see Fig. 3) and has a considerably longer biological half-life than L-dopa. It is considered a precursor of dopamine because it is partially demethylated in the organism [48], [49].

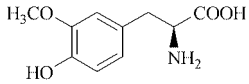

2.2.1.4. Monoamine Oxidase Inhibitors

As the name implies, monoamine oxidases (MAOs) are enzymes that oxidize and deaminate monoamines. In dopa metabolism, their

key role is the deamination (i.e., inactivation) of dopamine to form dihydroxyphenylacetic acid (DOPAC, see Fig. 3). Two types of MAO exist, the A and B types: MAO-A is more specifically oriented to the oxidation of 5-hydroxytryptamine, noradrenaline, and adrenaline in peripheral tissues; MAO-B predominates in the brain, and its main substrates are phenylethylamine, *p*-tyramine, and dopamine.

The first MAO inhibitors were nonselective and blocked both A and B enzymes. A major problem associated with MAO-A inhibitors in humans is that dietary tyramine gains access to the bloodstream; high-protein meals (most notably cheese) provide enough of this pressor amine to produce hypertension crises (the " cheese effect"). L-Deprenyl (selegiline) is a relatively selective inhibitor for MAO-B and has been used successfully as an adjunct to L-dopa therapy, potentiating its pharmacological action. The inhibition of MAO-B by deprenyl results in the accumulation of dopamine in the glial cells associated with synapses. This could have beneficial effects in PD; for example, the increased dopamine levels could block the metabolic pathways leading to the formation of potential toxins such as endogenous reactive oxygen species that are thought to form tissue-toxic free radicals (hydroxyl, hydrogen peroxide, and superoxide). Increased levels of dopamine in the synapses could also diffuse to appropriate target cell receptors.

Deprenyl is used in PD therapy, primarily as an adjunct for patients manifesting the motor fluctuations associated with chronic L-dopa treatment. Deprenyl usually permits a reduction of the L-dopa dosage.

Selegiline [*2323-36-6*], (deprenyl), (*R*)-(−)-*N*-methyl-*N*-phenylisopropyl-2-propynylamine, (*R*)-(−)-*N*-α-dimethyl-*N*-2-propynylphenethylamine, $C_{13}H_{17}N$, M_r 187.29, oil *bp* (0.63 kPa) 103 – 110 °C (decomp.). For synthesis, see [50].

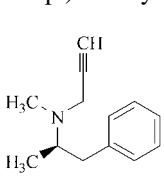

A comprehensive description of the pharmacology, toxicology, clinical safety, and efficacy of deprenyl is given in [51–53].

Trade Names. Eldeprine (Britannia), Jumex (Chinoin), Jumexal (Galenica), Movergan (Astra).

Rasagiline [*136236-51-6*], (*R*)-*N*-2-propynyl-1-indanamine, $C_{12}H_{13}N$, M_r 171.24, *mp* 183 – 185 °C (hydrochloride), 175 – 177 °C (tartrate), $[\alpha]_D$+30.9 ° (2 % ethanol), $[\alpha]_D$+34.3 ° (*c* 1.5, water).

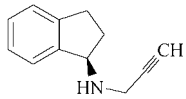

A comprehensive description of the chemistry, pharmacology, toxicology, clinical safety, and efficacy of rasagiline is given in [54].

Rasagiline is being tested in phase III clinical trials as mesylate (TVP-1012).

Producer. Teva Pharmaceutical Industries Ltd. (Israel).

Lazabemide [*103878-84-8*], *N*-(2-aminoethyl)-5-chloropyridine-2-carboxamide, $C_8H_{10}ClN_3O$, M_r 199.64, *mp* 193 – 195 °C (hydrochloride). For synthesis and pharmacological data see [55].

Lazabemide is undergoing phase III clinical studies.

Producer. Roche.

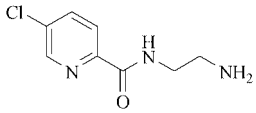

Mofegiline [*119386-96-8*], (*E*)-2-(4-fluorophenetyl)-3-fluoroallylamine, $C_{11}H_{13}F_2N$, M_r 197.23, colorless crystals, *mp* 131 – 133 °C (hydrochloride). A drug monograph is presented in [56].

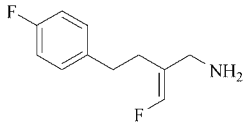

2.2.1.5. Catechol-*O*-Methyl-Transferase Inhibitors

In the beginning of the 1990s, a new class of drugs, COMT inhibitors, have been introduced into clinical practice.

These compounds represent a new strategy for the therapy of Parkinson's disease and may be useful as add-on in patients treated with levodopa by inhibiting peripheral and central metabolism of levodopa to 3-OMD (Fig. 3).

When a peripheral dopa decarboxylase inhibitor is used along with levodopa, *O*-methylation is the major metabolic pathway. 3-OMD has no antiparkinsonian effects and may reduce the efficacy of levodopa. For that reason, COMT-inhibitors could decrease the first-pass metabolism of levodopa, both in the systemic circulation and in the brain. This mechanism enables to enhance the bioavailability of levodopa, prolong its half-time and increase its uptake across the blood – brain – barrier.

Entacapone [*130929-57-6*], (*E*)-2-cyano-*N,N*-diethyl-3-(3,4-dihydroxy-5-nitrophenyl)propenamide, $C_{14}H_{15}N_3O_5$, M_r 305.29, *mp* 153 – 156 °C . A comprehensive description of entecapone is given in [57].

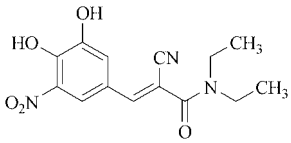

Entacapone is a peripheral COMT inhibitor. *Trade Names.* Comtan (Orion) Comtess (Novartis).

Tolcapone [*134308-13-7*], 3,4-dihydroxy-4'-methyl-5-nitrobenzophenone, $C_{14}H_{11}NO_5$, M_r 273.25. For synthesis, see [58]. An overview of pharmacological, toxicological and clinical trials data is presented in [59].

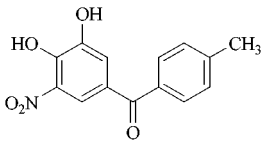

Tolcapone acts as central and peripheral COMT inhibitor and was marketed under the trade name Tasmar (Roche). However, following case reports on serious hepatotoxicity and three mortalities, it was withdrawn from the European market in November 1998.

CGP 28014 [*111757-17-6*], 6-dipropylaminomethylidenimino)pyrid-2(1*H*)-one, $C_{12}H_{19}N_3O$, M_r 221.30, *mp* 102 – 104 °C. For further details see [60].

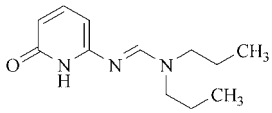

Producer. Ciba-Geigy.

2.2.1.6. Monoamine Reuptake Inhibitors

Monoamine reuptake inhibitors could increase the dopamine levels in the synaptic cleft. However, there is data indicating an association between serotonine reuptake inhibitors (fluoxetine [61], sertraline [62]) and the occurence of movement disorders.

The selective dopamine reuptake inhibitor GBR-12909 [63] has not been evaluated for PD therapy but for treating cocaine addiction.

Dopamine reuptake is strongly dependent on dopamine transport (see page 18).

Brasofensine [*171655-91-7*], (1*R*,2*R*,3*S*,5*S*)-3-(3,4-dichlorophenyl)-8-methyl-8-[3.2.1]octan-2-carbaldehyde-(*E*)-*O*-methyloxime, $C_{16}H_{20}Cl_2N_2O$, M_r 327.26. For further information see [64].

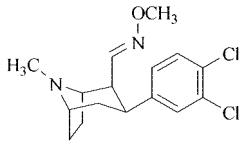

2.2.2. Dopamine-Releasing Drugs

Drugs may release dopamine from the nerve terminals either directly or indirectly via antagonism of glutamate receptors.

2.2.2.1. Direct Release

Amantadine [*768-94-5*], 1-tricyclo-[3.3.1.1^{3,7}]-decylamine, 1-adamantanamine, $C_{10}H_{17}N$, M_r 151.26; the hydrochloride and citrate are freely soluble in water, the sulfate only poorly soluble. For synthesis, see [65].

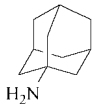

Amantadine was introduced as an antiviral agent for the prophylaxis of A2 influenza and unexpectedly resulted in symptomatic improvement of patients with PD [66–68]. It probably acts by releasing dopamine from intact nerve terminals in the nigrostriatum. Its therapeutic efficacy is enhanced by L-dopa. Compared to L-dopa or anticholinergic drugs, amantadine (especially the sulfate) is relatively free of side effects.
Trade Names. Hydrochloride: Amazolon (Sawai), Mantadine (Du Pont), Mantadix (Du Pont, Rhône-Poulenc), Symmetrel (Ciba-Geigy), Virofral (Ferrosan). Sulfate: Contenton (SKD), PK-Merz (Merz).

2.2.2.2. Dopamine Release by Glutamate Antagonists

The deficit of dopamine can be counteracted by restoring the balance between the cholinergic and dopaminergic function of the striatum (Fig. 4). This can be achieved by directly stimulating the dopaminergic neurons of the nigrostriatal pathways or by reducing the activity of the glutamatergic fibers projecting from the neocortex to the striatum (Fig. 2). Antagonists of the excitatory transmitter glutamate cause disinhibition of nigrostriatal neurons via the GABA-secreting neurons and hence increase the release of dopamine into the striatum.

The action of glutamate antagonists, e.g., NMDA receptor blockers, has been reviewed in [69].

Memantine [*19982-08-2*], 3,5-dimethyltricyclo[3.3.1.1^{3,7}]decylamine, 3,5-dimethyl-1-adamantanamine, $C_{12}H_{21}N$, M_r 179.31, is an oil, n_D^{25} 1.4941; the hydrochloride crystallizes from ethanol and diethyl ether, *mp* 290 – 295 °C (decomp.). For synthesis, see [70].

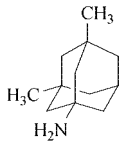

Memantine is a potent antagonist of excitatory glutamate neurotransmission in the CNS. It blocks the ion channel associated with the *N*-methyl-D-aspartate (NMDA) subtype of glutamate receptor [71], [72]. In contrast to other NMDA antagonists being developed, it does not induce psychotropic effects or memory impairment, and is tolerated well [73], [74]. Adamantane derivatives of the memantine type have opened new therapeutic perspectives for PD and may reduce the fatal consequences of CNS traumata, delay the progression of PD, and retard neurodegenerative processes in Alzheimer's or acquired immune deficiency syndrome (AIDS) patients.
Trade Name. Akatinol-Memantine (Merz).

Dizocilpine [*77086-21-6*], 10,11-dehydro-5-methyl-5*H*-dibenzo[*a,d*]cyclohepten-5,10-imine, (+)-5-methyl-10,11-dihydro-5*H*-dibenzo[*a,d*]cyclohepten-5,10-imine, $C_{16}H_{15}N$, M_r 221.30; (−)-form, colorless crystals from cyclohexane *mp* 68.5 – 69 °C; maleate (MK-801), crystals *mp* 208.5 – 210 °C. For synthesis, see [75], [76].

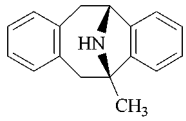

Dizocilpine is a noncompetitive antagonist of NMDA receptors with protective effects against ischemia-induced neuronal degeneration. It shows anticonvulsant activity in animal models and has been clinically evaluated in attention-deficit disorders [77–79].
Trade Name. Neurogard (MSD).

Riluzole [*001744-22-5*], 2-amino-6-(trifluoromethoxy)benzothiazole, $C_8H_5F_3N_2OS$,

M_r 234.20, *mp* 119 °C. For chemical and pharmacological properties see [80].

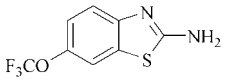

Riluzole is synthesized by condensation of 4-(trifluoromethoxy)aniline with potassium thiocyanate by means of Br_2 in acetic acid. Riluzole acts as glutamate release inhibitor for use in the treatment of amyotrophic lateral sclerosis. The substance is under evaluation for the therapy of Parkinson's disease [81].
Trade Name. Rilutek (Rhône-Poulenc Rorer).

2.2.3. Dopamine Receptor Agonists [82], [83]

The gradual loss of responsiveness during long-term (i.e., one to five years) L-dopa therapy stimulated the search for specific agonists that directly stimulate striatal dopamine receptors. Apomorphine, commonly used as an emetic, was the first dopaminergic agonist reported to have beneficial effects in PD. However, because renal damage occurs on long-term administration, natural and semisynthetic ergolines derived from ergot alkaloids have become the primary sources of such agonists.

Dopamine D2 receptor agonists (e.g., bromocriptine), proved clinically to best benefit patients with decreased response to L-dopa. They bypass the degenerated presynaptic neurons of the substantia nigra and stimulate postsynaptic dopamine receptors directly. They are used in combination with L-dopa. Results may improve if D2 agonist therapy is combined with L-dopa before chronic motor fluctuations occur. The main role of these drugs may thus be in the early stage of PD before the patients develop diurnal fluctuations or dyskinesias. Dopamine agonists reduce the dose of L-dopa required for an anti-parkinson effect; they may thus delay the complications of high-dose, long-term L-dopa treatment.

Bromocriptine [*25614-03-3*], (5′α)-2-bromo-12′-hydroxy-2′-(1-methylethyl)-5′-(2-methylpropyl)ergotaman-3′,6′,18-trione, 2-bromoergocryptine, $C_{32}H_{40}BrN_5O_5$, M_r 654.62, crystallizes from a mixture of methyl ethyl ketone and isopropyl ether, *mp* 215 – 218 °C (decomp.); the mesylate crystallizes from methyl ethyl ketone, *mp* 192 – 196 °C (decomp.).

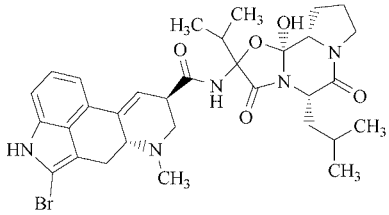

Bromocriptine is a derivative of the ergotoxin group of ergot alkaloids [84]. A comprehensive description of pharmacology, toxicology, clinical safety, and efficacy is given in [85–88].
Trade Names. Mesylate: Parlodel, Pravidel (Sandoz).

Lisuride [*18016-80-3*], *N*′-[(8α)-9,10-didehydro-6-methylergolin-8-yl]-*N,N*-diethylurea, $C_{20}H_{26}N_4O$, M_r 338.46, crystallizes from benzene, *mp* 186 °C; hydrogen maleate crystallizes from ethanol, *mp* 200 °C (decomp.). For synthesis, see [89].

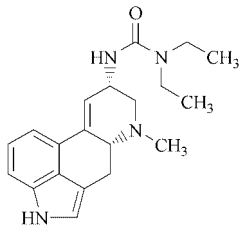

Trade Names. Hydrogen maleate: Cuvalit (Schering), Dopergin (Schering), Eunal (Nihon Schering), Lysenyl (Levica), Prolacam (Schering).

Terguride [*37686-84-3*], *N,N*-diethyl-*N*′-[(8α)-6-methylergolin-8-yl]urea, $C_{20}H_{28}N_4O$, M_r 340.47, crystallizes from ethanol, *mp* 203 – 204 °C (decomp.); the hydrogen maleate crystallizes from ethanol, *mp* 190 – 191 °C. For synthesis, see [90].

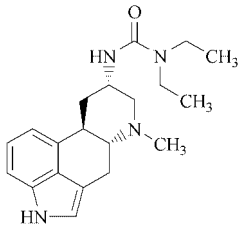

*Trade Name.*Hydrogen maleate: Mysalfon (Levica).

Mesulergine [*64795-35-3*], *N′*-[(8α)-1,6-dimethylergolin-8-yl]*N,N*-dimethylsulfamide, $C_{18}H_{26}N_4O_2S$, M_r 362.49; the hydrochloride crystallizes from ethanol, *mp* 226 – 228 °C (decomp.). For synthesis, see [91].

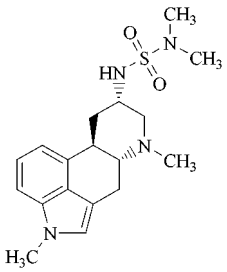

Pergolide [*66104-22-1*], 8β-(methylthiomethyl)-6-propylergoline, $C_{19}H_{26}N_2S$, M_r 314.49, crystals *mp* 206 – 209 °C (decomp.); mesylate crystals *mp* 225 °C (decomp.). For synthesis, see [92].

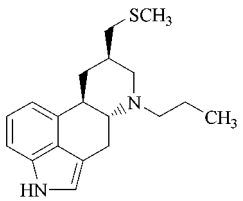

Naxagolide [*88058-88-2*], (4a*R*)-*trans*-3,4-4a,5,6,10b-hexahydro-4-propyl-2*H*-naphth[1,2-b]-1,4-oxazin-9-ol, $C_{15}H_{21}NO_2$, M_r 247.34, crystals *mp* 158 – 160 °C; hydrochloride crystals *mp* 303 – 305 °C. For synthesis, see [93], [94].

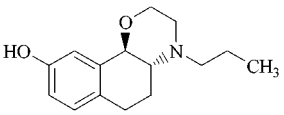

Pramipexole [*104632-26-0*], (*S*)-(−)-2-amino-6-(propylamino)-4,5,6,7-tetrahydrobenzothiazole, $C_{10}H_{17}N_3S$, M_r 211.33, crystals, *mp* (hydrochloride) 296 – 298 °C, $[\alpha]_D^{20}$ −67.2 ° (*c* = 1, methanol). Racemic crystals, *mp* 286 – 288 °C. For synthesis see [95]. An overview of pharmacological, toxicological and clinical trials data is given in [96], [97].

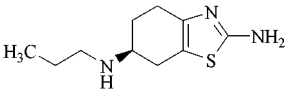

Pramipexole acts as dopamine D2 agonist.
Trade Name. Mirapex (Boehringer Ingelheim).

ABT-431 [*71752-56-0*], 9,10-diacetoxy-2-propyl-4,5,5a(*R*),6,7,11b,(*S*)-hexahydrobenzo[*f*]thieno[2,3-*c*]quinoline, $C_{22}H_{25}NO_4S$, M_r 399.51, *mp* 234 – 237 °C (hydrochloride), $[\alpha]_D$ −228 ° (*c* = 1.125, methanol).

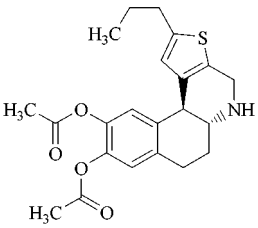

ABT-431 is the diacetyl prodrug of A-86929, a selective dopamine D1 subtype agonist. For synthesis see [98].

The pharmacological actions and clinical trials are reviewed in [99].
*Producer.*Abbott Laboratories (USA).

Ropinirole [*91374-21-9*], 4-[2-(dipropylamino)ethyl]-2-indolinone, $C_{16}H_{24}N_2O$, M_r 260.38, *mp* 241 – 243 °C.

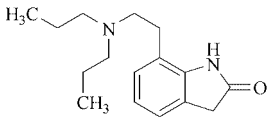

Ropinirole is a selective dopamine D2 subtype receptor antagonist. Pharmacological data of 2-indolinones are presented in [100].
Trade Name. Requip (SmithKline Beecham).

Cabergoline [*81409-90-7*], 1-[(6-allylergolin-8(-yl)carbonyl]-1-[3-(dimethylamino)propyl]-3-ethylurea, $C_{26}H_{37}N_5O_2$, M_r 451.61, white crystals, *mp* 100 – 105 °C. For chemical and pharmacological data, see [101].

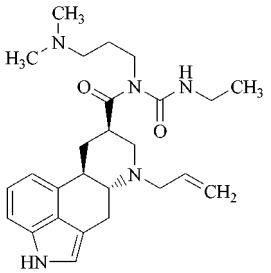

Cabergoline is a long lasting selective dopamine D2 receptor agonist currently licensed as L-dopa add-on.
Trade Names. Cabersase, Cabaseril, Dostinex (Pharmacia & Upjohn).

N-0923 [*125572-93-2*], (−)-(*S*)-2-[*N*-propyl-*N*-[2-(2-thienyl)ethyl]amino]-5-hydroxy-1,2,3,4-tetrahydronaphthalene, $C_{19}H_{25}NOS$, M_r 315.48, white crystalline solid, *mp* (hydrochloride) 186.5 – 187.5 °C. A comprehensive description of the chemistry, pharmacology, toxicology, clinical safety, and efficacy of N-0923 is given in [102].

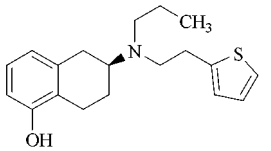

In the form of transdermal patch N-0923 has been licensed to Yoshitomi for further development and marketing in Japan [103].

Quinpirole [*85798-08-9*], (−)-(4a*R*, 8a*R*)-4,4a,5,6,7,8a,9-octahydro-5-propyl-1*H*-pyrazolo[3,4-g]quinoline, $C_{13}H_{21}N_3$, M_r 219.32, *mp* 250 – 257 °C. For synthesis and pharmacological data see [104], [105].

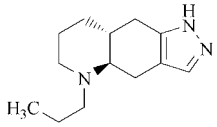

Quinpirole is a selective dopamine D2 subtype receptor antagonist developed for treatment of hypertension by Eli Lilly.

3. New Developments

In addition to the therapeutic approaches described above, there are some new concepts under evaluation for the treatment of Parkinson's disease.

Neurotrophic Factors. The role of neurotrophic growth factors as a means of preventing loss of dopaminergic neurons in the substantia nigra is becoming an accepted treatment strategy for Parkinson's disease. There are a variety of neurotrophic factors [106], [107] [basic fibroblast growth factor (bfGF, FGF-2), transforming growth factor-beta (TGF-beta) [108], brain-derived neurotrophic factor (BNDF) [109], nerve growth factor (NGF) [110], glial cell line-derived neurotrophic factor (GNDF) [111], growth/differentiation factor 5 (GDF5) [112] etc.] that offer new research potential into the cause and treatment of neurodegenerative diseases.

Neurotrophic Immunophilin Ligands. Immunophilins are intracellular receptor proteins for immunosuppressants such as cyclosporin A or FK-506. In addition to their immunosuppressive activity these drugs stimulate nerve growth [113]. In the 1990s several neurotrophic FK-binding protein ligands without immunosuppressive properties have been developed. The lead substance GPI-1046, $C_{21}H_{28}N_2O_4$,

shows comparable potency to NGF or BDNF but is active following oral administration.

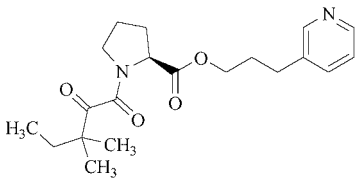

Producer. Guilford/Amgen.

Pre-clinical evaluation of neuroimmophilins revealed promising potential for the treatment of neurodegenerative diseases [114]. The second-generation ligand NIL-A is expected to start clinical trials for Parkinsonism therapy in 1999 [115].

Gene Therapy. The findings of the potential use of neurotrophic factors in the therapy of Parkinson's disease lead to some gene therapy related experiments and models for the treatment of neurodegenerative diseases. A recombinant adenovirus encoding GDNF was used for injection into brain and induced a massive production of GDNF, which protected rats in a lesion model [116]. By retrovirus-mediated transfer of the human tyrosine hydroxylase (TH) gene, astrocytes secreting high levels of dopa were generated. These dopa-producing astrocytes may be a promising option for ex-vivo gene therapy of Parkinson's disease [117]. TH gene introduction was also carried out by lipofectin-mediated gene transfer to rat striatum [118] and a retroviral vector containing the cDNA for rat TH [119]. Principles of gene therapy of Alzheimer and Parkinson's diseases are described in [120].

Adenosine A_{2A} Antagonists. The adenosine A_{2A} receptor plays an important role in brain function. Though its distribution in brain tissue and its action on striatal GABA-containing neurons is not exactly clear [121–123], A_{2A} antagonists such as KF17837 [124] may be useful as anti-Parkinson's disease drugs due to their modulating effects in neurotransmission [121], [125].

KF17837 [*141807-96-7*], (*E*)-8-(3,4-dimethoxystyryl)-1,3-dipropyl-7-methylxanthine, $C_{22}H_{28}N_4O_4$.

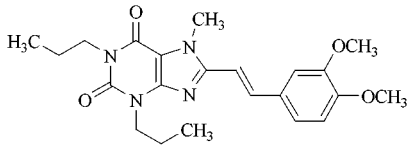

Neuronal Nicotinic Receptor Agonists. Nicotine and selective nicotinic acetylcholine (nACh) receptor agonists seem to have modifying effects on motor disabilities and dementia. There is some evidence that these substances have neuroprotective activity against glutamate, dopamine release enhancing properties, and effects on nerve growth factor-related mechanisms [126–128].

SIB1508Y is a selective brain nictoninic cholinergic subtype agonist which regulates the release of dopamine and acetylcholine [128], [129]. This compound is presently in phase II trials targeted for the motor and cognitive components in Parkinson's disease [130].

Dopamine Transport Inhibitors. Blocking the dopamine transporter increases the endogenous dopamine levels in the synaptic cleft. NS-2214 is a novel cocaine-derived dopamine transport inhibitor under clinical evaluation in phase II [126], [131]. The pharmacology and regulation of the neuronal dopamine transporter is reviewed in [132], [133].

Peptides. *Melanostatin* [*9083-38-9*], melanocyte stimulating hormone-release inhibiting factor, MIF, MRIH, MSH-release inhibiting hormone, L-prolyl-L-leucylglycinamide, $C_{13}H_{24}N_4O_3$. For synthesis, see [134].

Melanostatin inhibits the release of melanocyte stimulating hormone (MSH) from the pituitary (→ Peptide and Protein Hormones). The first factor to be isolated with MSH-release inhibiting activity was the side chain tripeptide of oxytocin. A structurally identical factor was also isolated from bovine pituitary [135–137]. Melanostatin has extrapituitary effects, particularly potentiation of L-dopa-induced behavioral changes [138–140]; it is, therefore, under development for PD therapy.

Various *thyrotropin-releasing hormone (TRH) analogues* are being developed. They

release endogenous dopamine in vitro, potentiate the locomotor activity of dopamine agonists in animals, show activity in different animal models of PD, may protect in electroshock and hypoxia models, and improve learning deficits. Representative derivatives are RGH-2202 [*49760-92-1*], (5-oxo-L-proline-L-leucine-L-prolinamide) under early clinical development by Richter; orotirelin [*62305-86-6*] (CG-3509, orotyl-L-histidine-L-prolinamide TRH) and montirelin [*62305-91-3*] (CG-3703, *N*-[(6-methyl-5-oxo-3-thiomorpholinyl)carbonyl]-L-histidine-L-prolinamide) both developed by Grünenthal; and azetirelin [*95729-65-0*] (YM-14673, $N(\alpha)$-([(*S*)-4-oxo-2-azetidinyl]carbonyl)-L-histidine-L-prolinamide) in early clinical development by Yamanouchi.

Transplants [141], [142]. Grafting tissue capable of supplying dopamine to the damaged striatonigral areas of the brains of patients with moderate to advanced PD may provide a cure [2], [143–145]. The grafting can be a homograft of dopamine-producing chromaffin tissue (e.g., the adrenal medulla). Human fetal brain tissue can also be used. The grafting of fetal dopamine-producing neuroblasts proved successful in animal models and in a severely affected patient [146]. Adult adrenal medullary tissue survives after grafting into host striatum. Administration of nerve growth factor increases both the number of permanently surviving catecholamine-producing cells and the number of nerve fibers formed by such cells. Dopaminergic neurons may, however, require specific growth factors to promote their survival, function, and neurite elongation [147].

The first clinical trials of the effect of grafting chromaffin tissue into brains of severe human PD cases have been performed. Despite earlier euphoria, experience now seems to indicate that neurografting has been beneficial only to certain individuals; many of these may be early onset PD cases [148], [149]. Thus, criteria for the selection of appropriate PD patients and evaluation of a possible therapeutic effect still have to be defined.

The development of NMDA antagonists that can retard the progression of PD appears to be a more promising approach to future PD therapy [150].

4. References

General References

1. J. Jörg, A. Janssen, *Z. Gerontol.* **20** (1987) 38 – 43.
2. S. Fahn, C. D. Marsden (eds.): *Movement Disorders* **4, suppl. 1** (1989). D. B. Calne, *Drugs for the Treatment of Parkinson's Disease,* Springer Verlag, Berlin 1989.
3. P. Sonsalla, L. J. Golbe, *Clin. Neuropharmacol.* **11** (1988) 500 – 511.
4. J. L. Juncos et al., *J. Neurol. Neurosurg. Psychiatry* **50** (1987) 194 – 198.
5. A. E. Lang, *Pharmacol. Ther.* **32** (1987) 51 – 76.
6. R. Duvoisin, *Pharmacol. Ther.* **32** (1987) 1 – 17.
7. D. Rogers et al., *Brain* **110** (1987) 761 – 776.
8. R. Sandyk, R. P. Iacono, *Int. J. Neurosci.* **33** (1987) 257 – 259.
9. O. Hornykiewicz, *Biochem. Pharmacol.* **24** (1975) 1061 – 1065.
10. H. Bernheimer et al., *J. Neurol. Sci.* **20** (1973) 415 – 455.
11. T. Müller, W. Kuhn, H. Przuntek, *Fortschr. Med.* **115** (1997) 31 – 34, 37 – 41.
12. W. Kuhn, T. Müller, *Fortschritte der Neurologie, Psychiatrie* **65** (1997) 361 – 347.
13. R. Pahwa, W. Koller, *Drugs of Today* **34** (1998) 95 – 105.

Specific References

14. J. F. Stein "The Control of Movement" in : C. Coen (ed.), *Functions of the Brain,* Oxford University Press, Oxford 1985, pp. 67 – 97.
15. H. E. Bradford: Chemical Neurobiology: An Introduction to Neurochemistry, Freeman, New York 1986.
16. G. Phokas, E. Steinegger, *Pharmazie* **11** (1956) 652.
17. *American Medical Association Drug Evaluations,* 5th ed., W. B. Saunders, Philadelphia, Pa. 19105, 1983, pp. 329 – 351.
18. G. Schwenker et al., *Chem. Ber.* **99** (1966) 2407 – 2408.
19. G. Fodor, *Chem. Ind.* 1956, 764.
20. P. Dobo et al., *J. Chem. Soc.* 1959, 3461 – 3465.
21. G. Fodor, *Tetrahedron* **1** (1957) 86 – 102.
22. Merck & Co., US 2 595 405, 1952 (R. F. Philips).
23. W. G. Evans, M. Wellendorf, *J. Chem. Soc.* 1959, 1406 – 1409.

24. Knoll, US 2 789 110, 1957 (W. Klavehn).
25. J. J. Denton et al., *J. Am. Chem. Soc.* **71** (1947) 2053 – 2054.
26. D. W. Adamson et al., *J. Chem. Soc.* 1951, 52 – 60.
27. Byk Gulden, DE 1 936 452, 1971 (H. G. Menge, J. Klosa).
28. Wellcome Found., GB 624 118, 1949 (D. W. Adamson).
29. Byk Gulden, DE 1 936 452, 1969 (H. G. Menge, J. Klosa).
30. A. Janssen et al., *Arzneim. Forsch. Drug Res.* **21** (1971) 1365 – 1373. A. Spek et al., *Nature (London)* **232** (1971) 575 –576.
31. Rhône-Poulenc, US 2 607 773, 1952 (S. S. Berg, J. N. Ashley).
32. Fujisawa, NL 6 609 280, 1967 (Y. Deguchi et al.).
33. Parke Davis, US 2 421 714, 1947 (J. Rieveschl).
34. Delagrange, US 2 942 000, 1960 (M. Mizier).
35. U. G. Bijlsma et al., *Arzneim.-Forsch. Drug Res.* **5** (1955) 72 – 73.
36. Geigy, US 2 397 799, 1946 (M. Martin et al.).
37. H. Yamada et al., *Chem. Pharm. Bull.* **10** (1962) 693 – 697.
38. C. J. Sih et al., *J. Am. Chem. Soc.* **91** (1969) 6204.
39. J. W. Becker et al., *Biochem. Biophys. Res. Commun.* **41** (1970) 444 – 449.
40. A. Barbeau, *Can. Med. Assoc. J.* **101** (1969) 59 – 68.
41. R. Gomez et al. in K. Florey (ed.): *Analytical Profiles of Drug Substances,* **vol. 5,** Academic Press, New York 1976, pp. 189 – 223.
42. M. Sletzinger et al., *J. Med. Chem.* **6** (1963) 101 – 103.
43. Hoffmann-La Roche, BE 619 015, 1962 (B. Hegedüs et al.).
44. G. Bartholini, A. Pletscher, *J. Pharmacol. Exp. Ther.* **161** (1968) 14 – 20.
45. I. Kuruma et al., *J. Pharm. Pharmacol.* **24** (1972) 289.
46. R. M. Pinder et al., *Drugs* **11** (1976) 329 – 377.
47. B. Hegedüs et al., *Helv. Chim. Acta* **58** (1975) 147 – 162.
48. J. Pletscher et al., *Brain Res.* **4** (1967) 106 – 109.
49. G. Bartholini et al., *Life Sci.* **14** (1974) 323 – 328.
50. J. S. Flower, *J. Org. Chem.* **42** (1977) 2637 – 2639.
51. G. P. Reynolds et al., *Biochem. Soc. Trans.* **7** (1979) 143 – 145.
52. P. Riederer, G. P. Reynolds, *Br. J. Clin. Pharmacol.* **9** (1980) 521 – 523.
53. P. N. Tariot et al., *Arch. Gen. Psychiatry* **44** (1987) 427 – 433.
54. *Drugs of the Future* **21** (1996) 903 – 605.
55. *Drugs of the Future* **18** (1993) 113 – 115.
56. *Drugs of the Future* **16** (1991) 428 – 431.
57. *Drugs of the Future* **19** (1994) 641 – 645.
58. *Drugs of the Future* **16** (1991) 719.
59. *Drugs of the Future* **21** (1996) 870.
60. *Drugs of the Future* **17** (1992) 275 – 277.
61. P. E. Gerber, L. D. Lynd, *Ann. Pharmacother.* **32** (1998) 692 – 698.
62. R. A. Hauser, T. A. Zesiwicz, *Mov. Disord.* **12** (1997) 756 – 759.
63. R. B. Rothmann, J. R. Glowa, *Mol. Neurobiol.* **11** (1998) 1 – 19.
64. *Chem. Abstr.* **124** (1996) 13 P176602r.
65. H. Stetter et al., *Angew. Chem.* **71** (1959) 429 – 430.
66. R. S. Schwab et al., *J. Am. Med. Assoc.* **222** (1972) 792 – 795.
67. D. Parkes in N. J. Harper, A. B. Simmonds (eds.): *Advances in Drug Research,* Academic Press, London 1974, pp. 11 – 81.
68. V. G. Vernier et al., *Toxicol. Appl. Pharmacol.* **15** (1969) 642 – 665.
69. M. Starr, *J. Neural Transm.* **80** (1995) 141 – 185.
70. K. Gerzon et al., *J. Med. Chem.* **6** (1963) 760 – 763.
71. J. Bormann, *Eur. J. Pharmacol.* **166** (1989) 591 – 592.
72. J. Kornhuber et al., *Eur. J. Pharmacol.* **166** (1989) 589 – 590.
73. P. A. Fischer et al., *Arzneim. Forsch. Drug Res.* **27** (1977) 1487 – 1489.
74. W. Wesemann et al., *Arzneim. Forsch. Drug Res.* **33** (1983) 1122 – 1134.
75. M. E. Christy et al., *J. Org. Chem.* **44** (1979) 3117 – 3127.
76. T. R. Lamanec et al., *J. Org. Chem.* **53** (1988) 1768 – 1774.
77. G. N. Woodruff et al., *Neuropharmacology* **26** (1987) 903 – 909.
78. K. Sato et al., *Brain Res.* **463** (1988) 12 – 20.
79. F. W. Reimherr et al., *Psychopharmacol. Bull.* **22** (1986) 237 – 242.
80. *Drugs of the Future* **19** (1994) 920 – 923.
81. C. Martin, *Chem. Br.* **34** (1998) Sept. 40 – 42.
82. C. Goetz, *Experimental Neurology* **144** (1997) 17 – 20.
83. J. Hagan et al., *Trends Pharmacol. Sci.* **18** (1997) 156 – 163.
84. Sandoz, DE 1 926 045, 1969 (E. Flückiger).

85. P. Weber, *Adv. Biochem. Psychopharmacol.* **23** (1980) 25 – 34.
86. N. Camerman, A. Camerman, *Mol. Pharmacol.* **19** (1981) 517 – 519.
87. D. Parkes, in A. B. Simmonds (ed.): *Advances in Drug Research 12,* Academic Press, New York 1977, pp. 247 – 344.
88. K. Y. Ho, M. O. Thorner, *Drugs* **36** (1988) 67 – 82.
89. V. Zikan, H. Semonsky, *Pharmazie* **23** (1968) 147.
90. Spofa, DE 2 238 540, 1973.
91. Sandoz, DE 2 656 344, 1978 (P. L. Stütz et al.).
92. Lilly, US 4 166 182, 1979 (B. J. Nicholas, E. C. Kornfeld).
93. D. G. Melillo et al., *J. Org. Chem.* **52** (1987) 5143 – 5150.
94. D. Dykstra et al., *Eur. Med. Chem. Chim. Ther.* **20** (1985) 247 – 250.
95. C. Schneider, *J. Med. Chem.* **30** (1987) 494.
96. *Drugs of the Future* **17** (1992) 291 – 292.
97. M. Piercy, *Clin. Neuropharmacol.* **21** (1997) 156 – 151.
98. M. Michaelides, *J. Med. Chem.* **40** (1995) 1585 – 1599.
99. *Drugs of the Future* **22** (1997) 821 – 824.
100. *Drugs of the Future CIPS* **14** (1989) 781 – 797.
101. *Drugs of the Future* **12** (1987) 842 – 844.
102. *Drugs of the Future* **18** (1993) 1005 – 1008.
103. *Drugs of the Future* **22** (1997) 1290.
104. *Drugs of the Future* **12** (1989) 558 – 561.
105. J. K. Rowlett, *Synapse* **20** (1995) 209 – 216.
106. B. Hoffer, A. Olson, *J. Neural Transm. Suppl.* **49** (1997) 1 – 10.
107. M. During, P. Leone, *Experimental Neurology* **144** (1997) 74 – 81.
108. P. Bratt, J. McPherson, *Cytokine and Growth Factor Reviews* **8** (1997) 267 – 292.
109. R. Maggio et al., *J. Neural Transm.* **104** (1997) 1113 – 1123.
110. C. Svendsen et al., *Experimental Neurology* **148** (1997) 135 – 146.
111. D. Choi-Lundberg et al., *Science* **275** (1997) 838 – 41.
112. A. Sullivan, *Neuroscience Letters* **233** (1997) 73 – 76.
113. T. Dawson et al., *Proc. Natl. Acad. Sci.* **94** (1997) 2019 – 2024.
114. G. Hamilton, J. Steiner, *Current Pharmaceutical Design* **3** (1997) 405 – 428.
115. Guilford Pharmaceuticals, Press release Sept. 1998.
116. A. Bilang-Bleuel et al., *Proc. Natl. Acad. Sc.* **94** (1997) 8818 – 8823.
117. C. Lundberg et al., *Experimental Neurology* **139** (1996) 39 – 53.
118. L. Cao et al., *Human Gene Therapy* **6** (1995) 497 – 501.
119. J. Wolff et al., *Proc. Natl. Acad. Sc.* **86** (1989) 9011 – 9014.
120. U. Muller, M. Graeber, *Journal of Molecular Medicine* **74** (1996) 71 – 84.
121. P. Richardson, H. Kase, P. Jenner, *Trends Pharmacol. Sci.* **18** (1997) 338 – 344.
122. B. Fredholm, P. Svenningson, *Trends Pharmacol. Sci.* **19** (1998) 46 – 48.
123. E. Ongini et al., *Annals of the New York Academy of Sciences* **825** (1997) 30 – 48.
124. T. Kanada et al., *Annals of Neurology* **43** (1998) 507 – 513.
125. M. Nomoto, *Nihon Shinkei Seishin Yakurigaku Zasshi* **16** (1996) 113 – 122.
126. M. Williams, S. Wrightm, G. Lloyd, *Trends Pharmacol. Sci.* **18** (1997) 307 – 310.
127. K. Hirata, S. Katayama, *Rinsho Shinkeigaku – Clinical Neurology* **36** (1996) 1318 – 1321.
128. A. Sacaan et al., *J. Pharmacol. Exp. Ther.* **280** (1997) 373 – 383.
129. Sibia Neurosciences Inc. Press release, Jan. 1998.
130. Sibia Neurosciences Inc. Press release, May 1998.
131. R. Pearce, *Exp. Opin. Ther. Pat.* **6** (1996) 949 – 953.
132. M. E. Reith, C. Xu, N. H. Chen, *Eur. J. Pharmacol.* **324** (1997) 1 – 10.
133. A. H. Newman, G. E. Agoston, *Curr. Med. Chem.* **5** (1998) 305 – 319.
134. H. Irie et al., *Chem. Lett.* 1980, 705 – 708.
135. A. J. Kastin, A. V. Schally, *Gen. Comp. Endocrinol.* **7** (1966) 452 – 456.
136. A. J. Kastin et al., *Yale J. Biol. Med.* **46** (1973) 617 – 622.
137. M. E. Celis et al., *Proc. Natl. Acad. Sci. USA* **68** (1971) 1428.
138. A. J. Kastin, A. Barbeau, *Can. Med. Assoc. J.* **107** (1972) 1079.
139. A. V. Schally et al., *Rec. Prog. Horm. Res.* **24** (1968) 497 – 581.
140. A. J. Kastin et al., *Life Sci.* **25** (1979) 401 – 414.
141. N. Diederich, F. Alesch, *Nervenarzt* **68** (1997) 466 – 476.
142. J. Krauss, *Schweizerische Medizinische Wochenschrift* **127** (1997) 1671 – 1675.
143. R. A. E. Bakay et al., *Appl. Neurophysiol.* **48** (1985) 358 – 361.

144. R. J. Wyatt et al. in J. B. Martin, J. D. Borchas (eds.): *Neuropeptides in Neurologic and Psychiatric Disease,* Raven Press, New York 1986, pp. 199 – 208.
145. P. Brundin, *Exp. Brain Res.* **65** (1986) 235 – 240.
146. O. Lindvall et al., *Science (Washington, DC)* **247** (1990) 574 – 577.
147. F. Hefti et al. in R. T. Bartus (ed.): *Neurobiology of Aging,* **vol. 10,** Pergamon Press, New York 1989, pp. 515 – 533.
148. E. O. Backlund et al., *J. Neurosurg.* **62** (1985) 169 – 173.
149. L. Olsen et al. in M. D. Yahr, K. J. Bergmann (eds.): *Advances in Neurology 45,* Raven Press, New York 1987, pp. 85 – 94.
150. J. W. Olney, *Biol. Psychiatry* **26** (1989) 505 – 525.

Particle Board → Wood

Peat

HEINRICH GRUMPELT, C. Deilmann AG, Bad Bentheim, Federal Republic of Germany

1. Introduction 1
1.1. Fundamentals 1
1.2. Formation and Varieties of Mires 3
1.3. Composition and Properties 4
1.4. History of Bog Utilization and Present Importance 7
2. Occurrence 7
3. Peat as a Raw Material, Carbon and Energy Source 9
3.1. Dehydration of Peat 9
3.2. Industrial Peat Production 9
3.2.1. Development of Peat Bogs 10
3.2.2. Mechanical Peat Production Methods 10
3.2.3. Legal Regulations for Peat Production 12
3.3. Uses of Peat and Peat Products . . 12
3.3.1. Plant and Vegetable Production . . . 12
3.3.2. Chemical Raw Material 13
3.3.2.1. Classification and Methods of Peat Processing 14
3.3.2.2. Peat Processing Technologies 15
3.3.3. Peat as an Energy Source 25
3.3.3.1. Peat Power Stations 27
3.3.3.2. Decentralized Energy Generation by Multistage Combustion 29
3.3.3.3. Peat Briquetting 32
3.3.4. Other Uses of Peat 32
3.3.4.1. Peat as a Filter Medium 32
3.3.4.2. Peat for Medical Purposes 32
4. Worldwide Economic Importance 32
4.1. Peat Industry and Availability of Raw Material Resources 34
4.2. Possibilities for the Substitution of Peat and Peat Products 34
5. Raw Material Production in Peat Bogs and Nature Conservation . . 34
5.1. Peat Production and Nature Conservation 34
5.2. Recultivation, Renaturation, and Regeneration 35
6. References 35

1. Introduction

1.1. Fundamentals

Geologically, peat is the youngest and least altered member of the combustible rocks. The position of peat within the usual classification of fossil fuels is shown in Figure 1 and Table 1. Compared to the other fossil fuels, peat is characterized by the lowest content of fixed carbon along with the highest content of volatile constituents.

Table 1. Classification of combustible rocks according to rank (C-enrichment) and the content of elements which vaporize on heating [1]

Fuel sources	C	H	O+N	Volume loss
Wood	50	6	44	
Peat	60	5	35	1/5
Lignite	65 – 78	5 – 6	17 – 29	1/4
Hard coal	79 – 92	4 – 5	4 – 16	1/2
Anthracite	93 – 98	1 – 4	1 – 3	
Graphite	100			

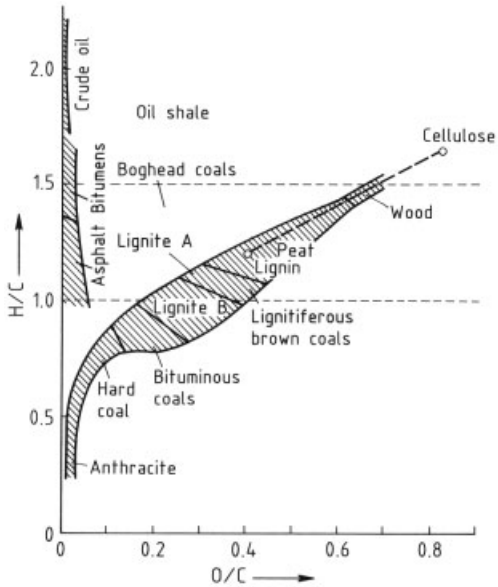

Figure 1. Classification of the fossil fuels

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_015

Deposits of peats are found in mires, where the oxygen content decreases and the nitrogen, hydrogen, and carbon contents increase with depth. The degree of decomposition of peat therefore also increases with depth.

Elemental analysis of the anhydrous organic matter of raised-bog peats depends on the degree of decomposition, as shown in Table 2. The chemical composition of pure peat is almost the same for all varieties of peat with the same decomposition. The water content of fresh peat is up to 96 wt %. Dry distillation of air-dried peat produces a cokelike residue which, like briquettes, coke, coke oven gas, tar oil etc., belongs to the artificial or processed fuels.

Peat is an acidic mixture of dead and decomposed, mainly vegetable, matter which forms in boggy areas [2]. The formation of peat occurred in the last 10 000 to 12 000 years in sedimentary deposits of the postglacial age. Peat is also found in mires formed in the interglacial periods. Peat is formed under exclusion of atmospheric oxygen by the action of water by a reduction process. The speed of formation depends on the climatic conditions.

The humification process of the bog plants, the coalification, starts as a biochemical process to form peat and continues at a much slower rate to ultimately form lignite. Up to this stage, cellulose, lignin, suberin, cutin, and sporenins can survive in the vegetable matter. As a result of the climatic differences during the formation of peats and lignites, the vegetable composition and the chemical and physical properties of the peats of Central Europe are usually very different from those of the Tertiary lignites in this region. Some peats from Brazil and Southern Europe, however, have undergone greater biochemical coalification than the Central European peats. It has therefore been concluded [3] that consolidation is the only difference between peats and lignites, since coalification is already evident in the peat phase. This difference of consolidation could occur as a result of pressure from overlying accumulated sediments and is determined by the moisture content.

The structure of raised-bog peats in Central Europe is strawlike, feltlike, or amorphous. Stratification by parallel orientation of plant residues is rarely found. Colors vary between light yellow and deep black.

Mires are defined as sediments which have a minimum depth of 30 cm in the undrained state. They may consist of one or more layers of peats, sapropel, or muds [2], [4].

Sapropel is formed when substances rich in wax and fat are mixed with these sediments.

Muds are formed by mineral deposits in stagnant waters.

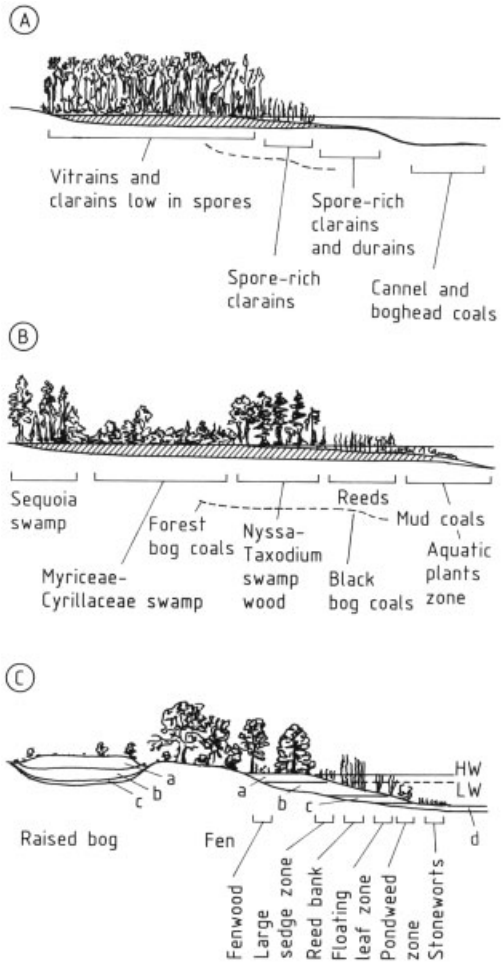

Figure 2. Mire types and coal or peat facies of the main moorland periods of the earth's history: Upper Carboniferous (A), Miocene (B), Holocene (C)
(according to Overbeck and Teichmüller)
Slightly decomposed peat (white peat); b) Highly decomposed peat (black peat); c) Muds; d) Bog lime or limestone gyttja
HW=high water; LW=low water

Table 2. Elemental composition of the dry matter of raised-bog peat as a function of the degree of decomposition

Degree of decomposition∗	Contents, wt %				
	C	O	H	N	S
H 1 – H 3	48 – 53	40 – 46	5.0 – 6.1	0.5 – 1.0	0.1 – 0.2
H 6 – H 7	56 – 58	34 – 39	5.5 – 6.1	0.8 – 1.1	0.1 – 0.3
H 8 – H 10	59 – 63	31 – 34	5.1 – 6.6	0.9 – 1.9	0.2 – 0.5

∗ H 1 – H 10 are the humosity grades (see Table 3).

1.2. Formation and Varieties of Mires

The growth of the postglacial mires in the humid latitudes of the Northern Hemisphere started when the climate changed after the Ice Age, and also occurred in the corresponding regions of the Southern Hemisphere. Carex peat and brown moss peats from the Alleröd period are the first indicators of bog formation and suitable conditions for bog growth have existed since then. They still persist in undisturbed mires to the present day. Figure 2 shows different types of mires which existed during the Upper Carboniferous, Miocene, and Holocene periods.

Mires form where there is an excess of water for most or all of the year. Hence bogs form when sufficient surface or groundwater and ample precipitation is available and water loss by evaporation or runoff is less than the water supply. Because of the continual excess of water, the organic substances cannot be completely degraded and therefore accumulate as sediments to form peats. The plant residues do not merely accumulate but also undergo many chemical transformations, especially in raised bogs, in the course of humification or coalification [3].

The main reason for the incomplete decomposition and steady accumulation of the (peat) sediments in the humid climatic zones is that the vitality of soil organisms is greatly inhibited under the very wet conditions (lack of air), a typical C/N ratio for raised bogs of > 50, and a low pH. Structureless humification products alongside original vegetable matter are therefore typical of a peat mixture. The cell structure of the matter is largely preserved and rich in acidic compounds. With increasing proximity to the permafrost zones, the ratio of the two constituents changes and fully preserved, but dead, vegetable matter is found. This is due to the fact that in the short summer months or weeks of these regions, the solar energy is still sufficient to support the growth of mire plants, even on partly thawed permafrost, and to contribute to the humification of the dead vegetable matter beneath the surface of the mire.

In tropical and subtropical regions the process described is inverted due to the high air and water temperatures which favor microbial activity. The growth rate and amount of biomass formed is greater than the fast microbial decomposition of the dead vegetable matter, enabling incompletely decomposed vegetable sediments to accumulate. These peat sediments are often lower in organic matter, are much less structured and have much higher contents of mineral constituents (ash) than the peats in northern latitudes. Deposits are formed which are often similar to those of coal swamps, lignite swamps, and lignite beds.

The organic matter of peat therefore differs in composition according to the nature of its starting materials. The more decomposed the peat, the higher the carbon content becomes. As regards the composition of the ash, whose content should not exceed 1 wt %, raised-bog peats have a sulfur content of 0.1 – 0.5 wt % and a phosphorus content of 0.008 – 0.025 wt %. Compared with other fossil fuels, these are extremely low values.

Two main types of mires are distinguished according to their formation, development, and structure: the *topogenic fens* as groundwater bogs, and the *ombrogenic raised bogs* as rainwater bogs [2], [5].

Fens. Fens occur near nutrient-rich (eutrophic) soils or are formed when nutrient-rich waters fill with sediment. Their surface is flat or sunken in the middle. Fens are rich in plant nutrients and thus support luxuriant vegetation. Since the ground- and/or seawater, especially

when rippling, contains oxygen (humid climatic zones) and/or when the water temperatures are high (tropics and subtropics), rapid decomposition of the dead plants and thus little accumulation of peat takes place. The typical structure of a fen is characterized by a regular succession of layers of reed peats, carex peats, and wood-rich fenwood peats (see also Fig. 2 C) [1]. Fen peats are therefore very heterogeneous, have pH values of 4.8 – 5.8, and are usually rich in calcium and nitrogen. The ash content of fen peat, relative to peat dry matter, in Central and Northern Europe is usually 3 – 14 wt %, and can be considerably higher in tropical mires. The nitrogen content available to plants in fen peat is between 2.5 and 4.5 wt %. In many parts of the world fens have been used agriculturally for centuries and have thus frequently disappeared from the landscape.

Raised Bogs. Raised bogs are only typical of the moderate to cold climatic regions of the earth and are independent of the terrain. Raised bogs are naturally treeless and are convex or domed in form, several meters above the mineral soil surface. Raised bogs are therefore independent of springs, groundwater, and stagnant water and have an autonomous water regime based on rainwater. Their groundwater level is therefore influenced solely by precipitation. At pH values of 2.5 – 3.5, only undemanding plants can grow. The so- called peat mosses (sphagnums) are the main peat-formers in raised bogs. Since the water has a very low pH value and is very low in oxygen, decomposition of organic matter is inhibited and as a result, considerable amounts of biomass can be produced in raised bogs (2.3 – 9.6 t dry matter per year and hectare). Peat mosses have annual growth rates of ca. 20 cm and more. The annual growth rate of the peat, though, is only ca. 1 mm. Raised-bog formations develop both on fens and directly on mineral soils as so- called own-rooted raised bogs. Figure 3 shows a cross section through a raised bog.

Transition Moors. Mires in which the ecological conditions of the raised bogs and fens overlap are called transition moors. They develop at the groundwater level or at fluctuating water level.

1.3. Composition and Properties

Accurate chemical methods for determining the degree of decomposition of peats are known, but it is usually sufficient to determine the degree of decomposition or humosity of freshly dug peat in H values using von Post's method [6]. A small amount of peat is crushed in the hand. The color of the water which then runs out and the nature of the crushed residue allows the degree of decomposition of the peat to be determined in values of H 1 to H 10 according to the scale shown in Table 3.

Humosity grades H 1 – H 6 are shown by the loose peats deposited uppermost in a raised bog bed, which according to DIN 11 540 are called *white peats* (sphagnum peat) [7]. *Black peats* have humosity grades H 6 – H 10. The bulk of Central and Northern European raised bogs consist of the older black peat. Fens mostly have a black peat character.

White and black peat differ in their structure, chemical, and physical properties. The loose white peat is absorbent, since the original plant structure with its capillaries largely survives, whereas the black peat is almost structureless and is no longer absorbent. The carbon content is ca. 50 wt % for white peat and ca. 60 wt % for black peat.

Raised-Bog Peat. Physical properties, such as high pore volume, water and air capacity, ion exchange capacity, and the high content of humic substance complexes and substances with plant physiological activity make raised-bog peats an irreplaceable raw material for gardening as a soil improver and for the production of substrates for plant cultivation [7]. These properties are described in DIN 11 540 (see also Table 4).

For industrial use as a chemically and thermally processed product, the content of the following elements are important: H, C, N, and O, which primarily form the organic substances of the peat; the primarily biogenic elements such as P and K; and the elements which are present in peat and all other fossil fuels, frequently in considerable amounts [8] (ash content, i.e., Na, Mg, Al, Si, S, Ca, Fe). The composition of this ash depends more on the geological environment than

Table 3. Determination of the degree of decomposition by von Post's humosity grade (H)

Humosity grade according to von Post	Designation of the peat	Color of the water running out between the fingers	Fraction of the peat pulp running out between the fingers	Residue in the hand	
				Form	Plant structure
H 1	completely undecomposed	colorless, clear	no peat pulp runs out between the fingers	not pulpy	clearly identifiable
H 2	almost completely undecomposed	pale yellow-brown, almost clear			
H 3	very slightly decomposed	brown, turbid			
H 4	slightly decomposed	intense brown, very turbid	some one-third	very pulpy	still clearly identifiable
H 5	somewhat decomposed				
H 6					
H 7	highly decomposed	very turbid	one-half		still identifiable
H 8	very highly decomposed		two-thirds	mainly as resistant root fibers, wood etc.	
H 9	almost completely decomposed		almost all		
H 10	completely decomposed	all runs out as a peat pulp between the fingers		no residues in the hand	

Table 4. Properties of raised-bog peats and transition-moor peats according to DIN 11 540, 1989 [a]

Properties	Raised-bog peat and transition-moor peat [b] without additives		
	Slightly decomposed	Moderately decomposed	Frozen highly decomposed
Bulk density (dry), g/L	40 – 80	70 – 150	120 – 250
Pore fraction, vol %	95 – 98	91 – 96	85 – 93
Water capacity, vol %	40 – 82	40 – 85	60 – 87
Air capacity, vol %	16 – 58	11 – 56	6 – 33
Degree of decomposition (r value), %	< 42	42 – 62	> 62
Humosity grade H (von Post)	1 – 3	4 – 6	7 – 10
Organic matter in the dry matter, wt %	94 – 99	94 – 99 (≥ 90)	94 – 99 (≥ 85)
Residue on ignition in the dry matter, wt %	1 – 6	1 – 6 (≤ 10)	1 – 6 (≤ 15)
pH (in $CaCl_2$)	2.5 – 3.5	2.5 – 3.5 (≤ 4.0)	2.5 – 3.5 (≤ 5.0)
pH (in H_2O)	3.0 – 4.0	3.0 – 4.0 (≤ 5.0)	3.0 – 4.0 (≤ 6.0)
Electrical conductivity (volume method, 1 : 3.6), µS/cm	< 175	< 175	< 175
Electrical conductivity (weight method, 1 g DM : 25 g), µS/cm	< 370	< 370	< 370

[a] The physical properties of peats are determined by the botanical composition, the degree of decomposition, the water content, the type of drying and production, the particle size and the granulometric composition. For this reason, slightly, moderately, and highly decomposed peats can overlap in their physical properties.
[b] The data in brackets refer to transition-moor peat. DM=dry matter.

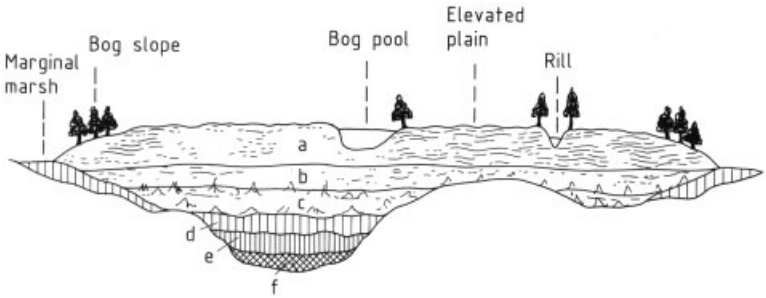

Figure 3. Schematic section through a raised bog that has built up over fen deposits
a) Slightly decomposed sphagnum peat (white peat); b) Highly decomposed sphagnum peat (black peat); c) Alder marsh peat, above it birch – pine – transition fenwood peat; d) Carex peat; e) Reed peat; f) Mud

on the plant species and is virtually unaffected by the mire facies.

Ash from raised-bog peat generally has low potassium and high silicon contents, whereas phosphorus, sulfur, calcium, magnesium, and iron always occur in similar ratios even in different kinds of peat.

Since raised-bog peats could only form above the groundwater, their content of mineral plant nutrients is very low. In the dry matter of raised-bog peats, mineral contents of 1 – 3 wt % are generally found. The following composition (ignition residue at 550 °C, wt %) is typical:

SiO_2	10 – 45
Al_2O_3	1 – 11
Fe_2O_3	1.1 – 5.5
CaO	2 – 45
MgO	1 – 20
MnO	0.1 – 0.3
K_2O	0.1 – 2.5
Na_2O	0.2 – 5
P_2O_5	1 – 3
SO_3	5 – 20

A typical elemental analysis (in wt %) of raised-bog peat is given below; the range depends on the degree of decomposition:

Carbon	50 – 60
Oxygen	30 – 45
Hydrogen	5 – 6.5
Nitrogen	0.5 – 2
Sulfur	0.10 – 0.5
Phosphorus	0.008 – 0.025

The contents increase with the degree of decomposition, except for the oxygen content which decreases.

The individual groups of substances present in raised-bog peat are varied as a result of the many peat-forming plants, and include carbohydrates and lignin. Waxes, resins, colorants, and essential oils are present in very low and variable concentrations. The bitumen content of peats increases with the number of bush plants which took part in forming the peat substance of deposit. A typical substance group analysis of raised-bog peat shows the following average values (wt %):

Waxes, resins, bitumen	4 – 7
Cellulose	3 – 5
Lignin	30 – 40
Humic acids	25 – 38
Insolubles	≤ 22

The content of cellulose decreases with increasing degree of decomposition, whereas all the others increase.

1.4. History of Bog Utilization and Present Importance

In Europe, peat has been used for centuries. Outside Europe, peat was only important intermittently and in a few areas as a source of fuel in the past.

From about 1975 on, though, a changing trend became apparent, and peat appears to be gaining in economic importance as an energy raw material in those regions of the world where it is the only fossil energy source. In Central Europe, in contrast, peat remains important only as a source of raw material (activated carbon and

synthesis gas production) and of humus (gardening uses).

Finland and Ireland, however, continue to rely on peat as an important energy source. The former Soviet Union, in contrast, which has the richest peat resources in the world, has run down its peat power stations and briquette works and gradually replaced them by nuclear power stations [9–12].

In the peat-rich area of the north-west German plain and other European mire areas a new kind of mire use is being developed because of increased environmental awareness. It has become necessary to do equal justice to both peat extraction, which is necessary for economic reasons, and to nature conservation and care for the landscape. Initial attempts or efforts already being made to make moorlands usable for forestry have acquired a completely new importance in the effort to protect the atmosphere by reducing the CO_2 discharged from fire.

Until well into the last century, peat was used as fuel after hand cutting and, up to the present, is still used in large amounts as an energy source. Original extraction of the peat by hand was restricted to the fringes of the mires, since the interior of the deposits could not be walked upon because of the inadequacy or absence of drainage. Only the highly decomposed black peat was extracted at first, owing to its higher calorific value [5].

Today, both black and white peat are economically important. The extraction methods in use today are adapted to the particular structural properties of the peats: the looseness of the white peat is preserved, whereas the black peat is further compacted. Originally, the slightly decomposed raised-bog peat was used to produce soil improvers. Nowadays, the highly decomposed black peat is also increasingly extracted for this purpose and, to a limited extent (ca. 230 000 t/a), in the form of fuel peat sods for e.g., thermal processing to activated carbon [5], [13], [14].

2. Occurrence

Mires are distributed (in various forms, depths, and areas) over the whole earth, apart from very dry and icy regions. Figure 4 shows different types of mires.

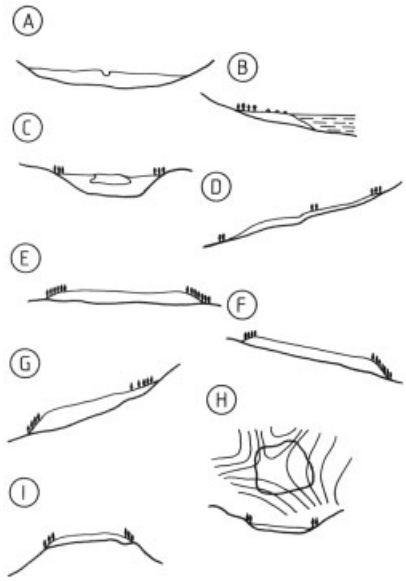

Figure 4. Types of mire according to [1]
A) Paludification fen; B) Bog of lacustine origin; C) Floating vegetation mat; D) Soligenous bog on hill slope; E) Symmetrical raised bog; F) Asymmetrical raised bog; G) Ombro-soligenous raised bog; H) Ombro-soligenous raised saddle bog; I) Crest raised bog

Regions with intensive peat formation and accumulation are found especially in Northern Europe, West Siberia, North America, and the humid and warm tropics (chiefly the Congo basin and the islands of the Malaysian archipelago). The accumulation of peat in the tropics is much less than in the northern latitudes.

Mires in Europe, West Siberia, and North America. The former Soviet Union has the largest bog deposits, with 61.3 % of all known areas (see Table 5). This includes the largest continuous bog region, the Vajguganskojer bog, which is 53 700 km^2 in area between the rivers Ob and Irtysch in West Siberia [4].

A great variety of mire types exist in North America, including Canada, the northern United States, and the Atlantic coastal strip down to Florida, as a result of the varying physiology and the different climatic conditions.

The mires of Europe have been thoroughly investigated. Here mires cover about 5.5 % of the total area (excluding the former Soviet Union). Finland has the greatest share of the resources (32 %) but the mires in Sweden and Ireland are also important.

Table 5. Distribution of mires in various countries [1]

Countries	Mire area	
	km^2	% (of country area)
Africa		
Uganda	14 200	6.0
Zambia	11 060	1.5
Guinea	5 250	2.1
Congo	2 900	0.85
Madagascar	1 970	0.34
America		
Canada	1 113 300	12.0
United States	596 400	6.4
Brazil	15 000	0.2
Chile	10 470	1.4
Venezuela	10 000	1.1
Guiana	8 140	3.8
Panama	7 870	10.4
Cuba	7 670	6.7
Honduras	4 530	4.0
Nicaragua	3 710	2.8
Asia		
Indonesia	263 000	17.7
China	41 590	0.45
Malaysia	23 610	7.2
Japan	2 190	0.6
Europe		
Former Soviet Union [a]	715 000	3.2
Finland	104 000	33.5
Sweden	70 000	15.6
Norway	30 000	9.2
United Kingdom	15 087	6.5
Germany	14 790	5.65
Poland	13 000	4.1
Ireland, Republic of	11 756	17.2
Iceland	10 000	9.7
France	1 200	0.2
Hungary	1 000	1.1

[a] The Soviet Union contains ca. 200×10 mires
[9] t=ca. 60 % of the world peat resources. This includes 49 % raised bog, 13.4 % fens, and 3.1 mixed forms.

In the Federal Republic of Germany the mires are distributed unevenly, according to the climate and topography. Raised bogs are found mainly in the rainy coastal regions, in the Voralpenland (Bavaria), and in the higher regions of the mountains in Baden-Württemberg. Fens exist almost everywhere, but they were largely drained centuries ago to be used for intensive agriculture. A complete inventory of the raised bogs (which are now the only mires of economic interest) and their present condition exists only for Lower Saxony: the figures for the other parts of Germany are partly based on estimates of older statistics [15]. These yield a total mire area in the Federal Republic of Germany of ca. 8715 km^2 (5153 km^2 fens, 3562 km^2 raised bogs).

Lower Saxony contains ca. 70 % of the German mire areas, and thus the German peat industry is largely located there. In addition, about 90 % of Germany's recoverable or workable peat resources are in Lower Saxony [16]. The geological peat resources of Lower Saxony consist of 850×10^6 m^3 white peat and 1.5×10^9 m^3 black peat [16]. In 1989, 60.4×10^6 m^3 white peat and 138.6×10^6 m^3 black peat were extracted.

Mires of the Tropics and Subtropics. Deposits comparable with the European mires are formed even in the fairly high mountainous regions of the tropics above the forest zones and in other regions. These deposits, though, have scarcely any economic interest for the country concerned owing to their small thickness and great remoteness [4].

Other Mire Areas. The mires of Australia, Tasmania, and New Zealand are predominantly raised bogs as a result of high precipitation and cooler climate [4].

Knowledge of the mires in China is limited: fens predominate with an area at present reported as about 4.2×10^4 km^2, i.e., about 0.45 % of the total area of the country. In Japan, small fens also predominate and are used agriculturally [17].

Peat Resources of the Earth. Since knowledge about the worldwide distribution of mires is still very incomplete, it can be assumed that the total known mire area of the earth previously established, of ca. 3.2×10^6 km^2 to date, will increase considerably [18].

The peat resources appear to have increased in the period from 1976 to 1989. New methods of aerial photogrammetry and remote sensing have enabled the earth's surface to be recorded much more accurately and in more detail, and to be scanned and documented cartographically even in remote regions. Also as a result of improved interpretation and analysis of the peat deposits, the new techniques have also given more precise figures on the known and newly discovered peat deposits which have been investigated [18].

Nevertheless it has not so far been possible to carry out a new and more reliable calculation of the world reserves and resources of peat [19].

3. Peat as a Raw Material, Carbon and Energy Source

3.1. Dehydration of Peat

In the crude state, peats still have water contents of 92 wt % (slightly decomposed peats) or 90 wt % (highly decomposed peats) and consequently extremely small energy contents. Therefore, the peat must be air-dried to make processing technically and economically feasible [20], [22].

Figure 5 shows the dependence of drying time on the moisture content of shaped pieces of highly decomposed black peat (fuel peat sods).

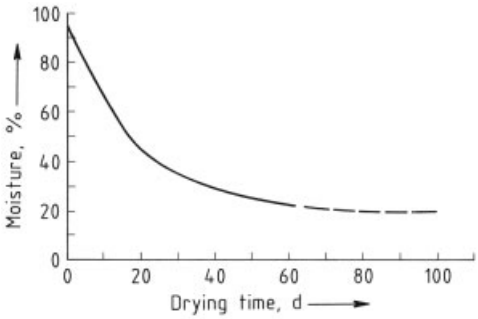

Figure 5. Drying curve of black peat sods (fuel peat)

3.2. Industrial Peat Production

Generally, peat production proceeds in two stages: (1) production of white peat and, when completed, (2) production of black peat. Different excavation methods are used.

In Germany the white and black peat were previously extracted in cuboid-shaped pieces, the so-called sods. Today and throughout the world, more economical, mechanized, peat production methods are used–milling and heap production of loose peat. Lump production and hydropeat production are practiced in isolated cases.

Bogland development and drainage plans, long-term extraction plans, and measures for subsequent recultivation or regeneration works need to be considered.

Figure 6 shows the structure and thickness of the individual layers of a raised bog [23], which determine the methods and machines which can be used for peat extraction.

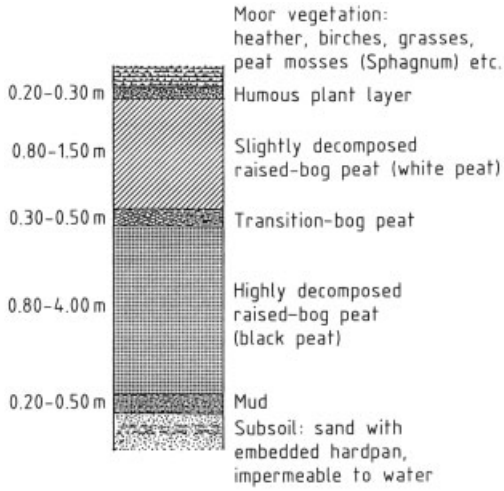

Figure 6. Schematic profile through a raised bog

3.2.1. Development of Peat Bogs

The water content in the bog must be reduced from its original 96 wt % to at least 90 wt % in order that light machines can be driven on the deposit.

A network of main and subsidiary ditches is laid out with draining channels laid perpendicular to the main ditches. In this way up to $0.8\,\mathrm{L\,h^{-1}\,ha^{-1}}$ water can be discharged from virgin mires.

Drainage measures must operate for at least five years before the start of actual peat production. The draining of raised bogs should be carried out during the fall and winter months, since the frozen bog substance can then bear more weight and permits machines to be used.

After draining the surface of the bog must be cleared and/or freed from the overlying cultivated layers before levelling of the bog can begin. The machines used for this purpose must have a particularly low surface pressure or soil pressure [24].

3.2.2. Mechanical Peat Production Methods

The use of sod peat production methods only became possible on the industrial scale with the development of suitable peat machines. Powerful extraction machines have been available for fuel peat production since 1908 (the peat bucket chain excavator developed by STRENGE and WIELANDT) but that has been applied to white peat production only since 1955 [25].

For both types of production, which are still important today, the most suitable block has proved to be a sod with dimensions of $43 \times 15 \times 15$ cm. For the production of such blocks (sods) in the extraction of white and black peat, the peat bog is divided into long extraction strips, the so- called trenches. If peat bog areas are too small or unsuitably cut, this division into trenches can lead to high extraction losses and permits only a limited output per unit area and time depending on the weather. To increase production and profitability of present-day industrial peat production operations, other extraction methods, such as the milling and heap production methods have been developed.

White Peat Production. White peat (sphagnum peat moss) can be produced in the form of sods throughout the year, provided there is no frost. Frost prevents the cutting of the fibrous white peat but not its drying which, in contrast to that of the structureless, pasty, black peat, continues without disruption even below the freezing point. The white peat cutting machines in use today weigh ca. 2.5 t, run on caterpillar mountings and imitate the processes of cutting, lifting, and regular depositing of the sods in manual cutting. They have an output of ca. 50 – 80 m^3/h, corresponding to an annual output of ca. 40 000 – 80 000 m^3 [24], [26].

The subsequent drying processes of the white peat sods are now also mechanized. The way in which the cut sods are piled up by the cutting machine in layers on the bog into long ridges affects the success of the subsequent drying operations. With favorable weather conditions, white peat dries in the course of a year to a residual water content of 50 – 55 %, which can then be processed further. About 800 L water per cubic meter peat must be evaporated. The collection of dried white peat sods also proceeds mechanically with the aid of railroad vehicles on caterpillar tracks, which finally transfer the product to portable railroads, thus taking over transport of the peat to the site of further processing.

Black Peat Production. This begins in raised bogs, usually after completion of white peat extraction (if these layers are present and extractable). Careful preparation of the peat bog is required if the black peat is to be used as fuel peat, including thorough drainage down to the sandy subsoil [27]. If, however, loose production of black peat as milled peat is required, less thorough drainage of the trenches is sufficient.

Sod Peat Production Process. For this purpose, special bucket- chain excavators on caterpillar crawlers are used. These sod peat excavators lift the bog substance out of the bog and convey it to a mixer situated on the excavator. This mixer homogenizes the peat material and transfers it as a compacted strip onto a conveyor, up to 50 m in length, which finally deposits this strip onto the prepared, well drained, dry field. Sod- cutting disks then divide up the strip to give the sods their final form. The minimum trench length should be 2000 m.

Only the period from the end of April to the beginning of August is available for fuel peat production in Central Europe since the sods must be thoroughly dried to yield usable fuel peat. Overwintering results in further compaction, so that in the following year the peat has the residual moisture content of ca. 35 wt % required for industrial processing and can be easily transported [27], [28].

Using modern bucket- chain excavators in a well-drained bog, one dredger can give extraction rates of up to 200 000 m^3/a of crude peat with a water content of 90 wt %, corresponding to 27 000 t/a of air-dried fuel peat with a residual moisture content of ca. 30 – 35 wt %.

Milled Peat Production Process. Whereas in the sod peat process the whole thickness of the black peat layer is extracted in a single operation, in the milled peat production process, 1 – 3- cm thick layers of the carefully drained and levelled bogs are milled off horizontally in each harvesting cycle. The present-day milled peat production processes are fully mechanized. Pin rollers, spoon harrowers, and mechanical and pneumatic collecting devices carry out the

individual operations prior to stacking of the peat in heaps, which are then thoroughly dried to 40 – 65 wt % water [23]. As in the sod peat production process, the milled peat method also requires division of the bog into long trenches which are, however, much narrower and are separated from each other by drainage ditches. These fields must be laid out with convex curvature perpendicular to the longitudinal axis of the trenches and be well levelled so that rainwater runs off as quickly as possible and the peat already milled continues its drying process.

Whereas the sod peat process requires warm winds for drying, the milled peat production process requires primarily intense solar radiation. Up to 25 milled peat harvests are possible per extraction field during the summer peat production season. According to the degree of decomposition and final moisture level, 7 – 16 t milled peat with a bulk density of 150 – 350 kg/m^3 can be produced per hectare and harvesting cycle [23], [29].

Milled peat production is the most widespread production method for peat of all qualities, in Scandinavia, the former Soviet Union, Ireland, Poland, the United States, and Canada. The milled peat production method made it possible to produce electricity in Ireland, Finland, and the Soviet Union using the economical pulverized fuel firing system known from lignite power stations in place of the much less effective sod firing system. The introduction of milled peat production has also enabled peat briquettes to be produced in large amounts.

In the milled peat production process, the risk of self-ignition of the peat on storage is a disadvantage. There is also a risk of fermentation starting in the milled peat heaps, in particular for milled raised-bog white peat.

Heaped Peat Production Process. The heaped peat production process, which is used increasingly in Germany for the production of peat soils from loose black peat predried in the air, is a variant of milled peat production [28]. In contrast to the milled peat process, all layers of the bog are extracted in one operation. Complete mixing of the different bog layers is necessary to ensure a more uniform quality of the peat. The peat is therefore produced first along the usual trench edges, either batchwise with simple machines such as mechanical shovels, clamshell diggers, or dragline bucket excavators, with the aid of a sod bucket elevator – distributor system, or with other machines. The peat is then scattered parallel to the trench edges on well-drained and levelled areas of bog in a thicker layer than in milled peat production. The peat heaped or scattered in this way can usually be re-stored after an initial storage period of only a few days. The heaped peat production process can also be economic on residual areas, so that practically no extraction losses occur, and in principle year-round production would be possible [30].

Hydropeat Production Process. In western Canada, the climate and difficult storage conditions for the peat neither allow thorough predrainage of the bog nor the use of the previously described methods of peat production; the so-called hydropeat production process is used here [23]. Low natural gas prices allow profitable thermal postdrying and further processing of the dried peat. The bogs intended for extraction are neither drained nor dug; a few square meters of bog is simply excavated from the bog edge. On the pond formed, which rapidly fills with water, a pontoon is then set up with a clamshell or multiclamshell bucket dredger for peat production. This dredger removes the whole thickness of the peat including all the roots it contains and transfers the raw material mixture first to a sieve drum in which primarily the tree and bush roots and stones are removed. The peat suspension, containing a maximum of 2 – 3 wt % dry matter, then passes to the pump tank in the float of the dredger pontoon and from there is pumped for further processing to a dewatering unit (screen belt presses and decanters). The peat is then washed (essentially only the fibrous material still survives), thermally dried, and the muddy water is returned to the bog. A short-fibered, fine-granular to powdery peat is obtained as the final product of drying which differs greatly from the European peat products in its fineness and is sold solely as a soil improvement agent.

3.2.3. Legal Regulations for Peat Production

Since the end of World War I, increased peat extraction resulting from the increased need for fuel have made legislative measures necessary.

The objectives are on the one hand the preservation or at least restoration of the mires as areas for agricultural and forestry use, but on the other hand the state also wishes to assure itself a source of income through taxes (royalties) by granting mining rights for exploitation of the bogs. In addition to the laws created especially for peat extraction, laws were enacted for conservation of the nature and countryside, such as the German Federal Nature Conservation Law of 20th Dec. 1976. The new laws provide for the reintegration into the countryside activity of all mires which have more or less lost their peat [31], [32].

3.3. Uses of Peat and Peat Products

Figure 7 shows the various uses of peat, and especially of raised-bog peat.

The main fields of use of raised-bog peat are:

1) the supply of peat substrates to the horticultural trade and to the plant cultivation industry (60 %),
2) production of activated carbons (20 %), and
3) the soil improvement field (20 %).

The German peat industry extracts 12×10^6 m^3 of raw peat from the mires for these uses. In Finland more than 20×10^6 m^3 peat are used for the energy industry. For the year 2000, a demand of 30×10^6 m^3 peat is expected in Finland, and is planned to cover ca. 15 % of the national energy consumption [10].

3.3.1. Plant and Vegetable Production

Peat and peat substrates for plant cultivation must satisfy definite quality criteria which are summarized in DIN 11 540 (see Table 4). The following are important properties of raised-bog peat: low pH value, high water capacity, high air capacity, good structural stability, absence of pollutants and agents of disease, lack of microbial stimulation, and low nutrient content.

The peat cultivation substrates are adjusted to a specific pH value according to plant demands and cultivation method, fertilizer is added, and, after preliminary sieving to adjust the structure as required, they are mixed if necessary with pure additives (e.g., sand, clay, perlite, vermiculite, rock wool or mineral components) or organic additives (bark humus, rice husks, or coconut fibers). Peat without nutrient additives is used as a soil improver (e.g., in tree nurseries). Today peats of all degrees of decomposition are used, i.e., not only slightly or moderately decomposed white peats but also increasingly highly decomposed black peats.

The most important criteria for assessing commercial peats, cultivation composts, and garden soils are assembled in Table 6 (DIN 11 540).

3.3.2. Chemical Raw Material

Peat is a much lower-valued raw material than crude oil or coal because of its high water content. It also contains many organic constituents which are destroyed or converted when peat is fired, pyrolyzed, or gasified. Extraction of the peat constituents lignin, humic acids and humins, pectin, hemicelluloses and cellulose, as well as peat bitumen and waxes therefore might be of interest [21], [29], [34–37]. However, their concentration in peat is in some cases only small and their isolation can be costly and energy-demanding so that their recovery is uneconomical. In the Soviet Union, numerous products continue to be produced, e.g., peat wax, peat dyes, and extracts for the stimulation of plants and microorganisms as well as for the production of medicinal ointments. In addition, the production of methanol from peat and the catalytic liquefaction of peat are being discussed [27].

In Finland, the use of peat as a chemical raw material for the production of ammonia is very advanced, and the first production plant of this kind was started up by Kemira in Oulu in 1988 [38].

3.3.2.1. Classification and Methods of Peat Processing

In all methods of *thermochemical peat processing* the high-molecular mass peat constituents are converted to low-molecular mass organic constituents. These destructive methods of peat processing are similar in their variety of possible processes and products to coal chemistry [39] (see Fig. 8).

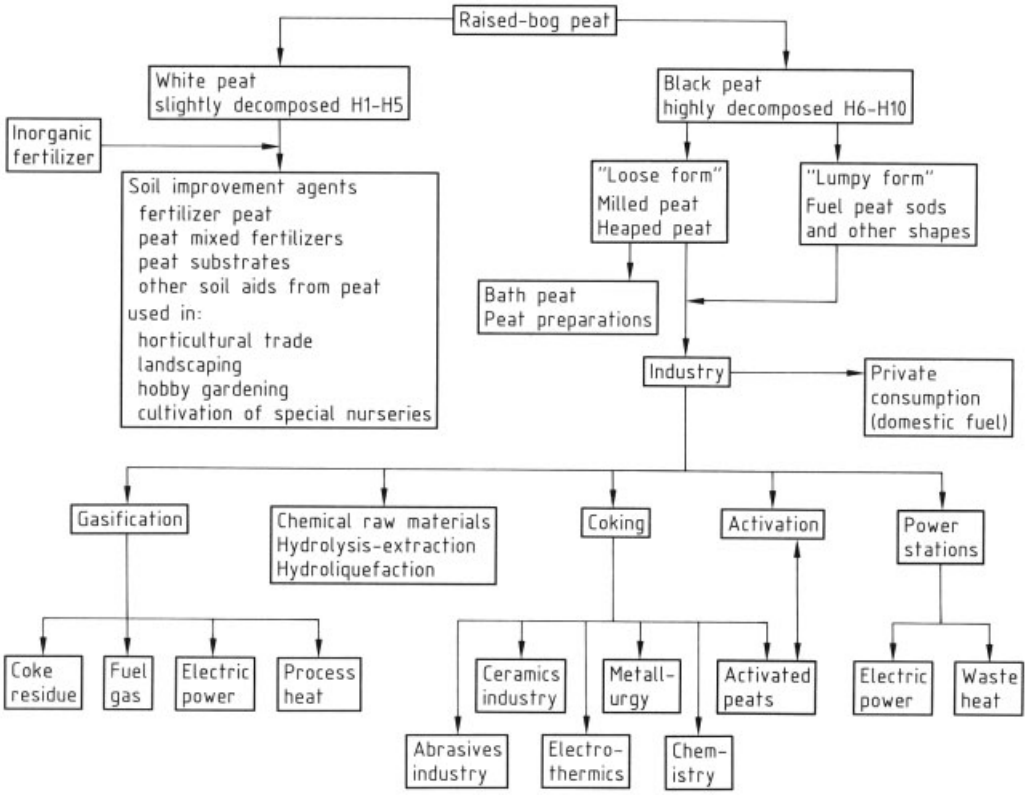

Figure 7. Use of raised-bog peats [33]

Table 6. Criteria for the assessment of commercial peat products (peats, peat substrates, and garden soils)

Product [a]	Bulk density, dry,[b] g/L	pH (in $CaCl_2$)	Salt content,[c] g/L	Available to plants [d] N, mg/L	P_2O_5, mg/L	K_2O, mg/L
Peat for production of substrates	$\leq$250	2.5 – 3.5	$\leq$ 0.4	$\leq$ 50	$\leq$ 30	$\leq$ 40
Neutral substrates [e]	$\leq$ 250	5.0 – 7.0	$\leq$ 0.8	$\leq$ 50	$\leq$ 30	$\leq$ 40
Propagation and transplanting substrates	$\leq$ 250	5.0 – 7.0	0.8 – 1.5	100 – 250	100 – 250	100 – 350
Potting substrates	$\leq$ 250	5.0 – 7.0	1.5 – 3.0	200 – 450	200 – 400	200 – 500
Garden soil	$\leq$ 250	5.0 – 7.0	1.0 – 3.0	100 – 450	100 – 400	100 – 500
Peat substrate for plant cultivation	$\leq$ 250	3.5 – 4.5	0.5 – 1.0	80 – 160	30 – 120	60 – 160

[a] For special peat substrates not listed in the table, the standard values correspond to the requirements of the type of cultivation concerned. [b] Relative to peat. [c] The salt content, expressed as KCl, can be determined from the electrical conductivity of an aqueous extract at a fixed extraction ratio. [d] The content of nutrients available to plants is determined by standardized methods applicable to the analysis of garden soils. [e] These are limed products with trace elements but without the main nutrients (N, P, K) [8].

The following basic processes can be distinguished in peat processing:

1) Anaerobic fermentation to methane and carbon dioxide
2) Acidic or enzymatic hydrolysis to monosaccharides and their subsequent fermentation to alcohols or acids
3) Thermochemical liquefaction with gaseous reductants and solvents together with catalysts
4) Low-temperature carbonization and coking or pyrolysis to peat coke, peat tar, and peat gas

5) Gasification by partial combustion with pure or atmospheric oxygen and steam to synthesis gas and ammonia [38].

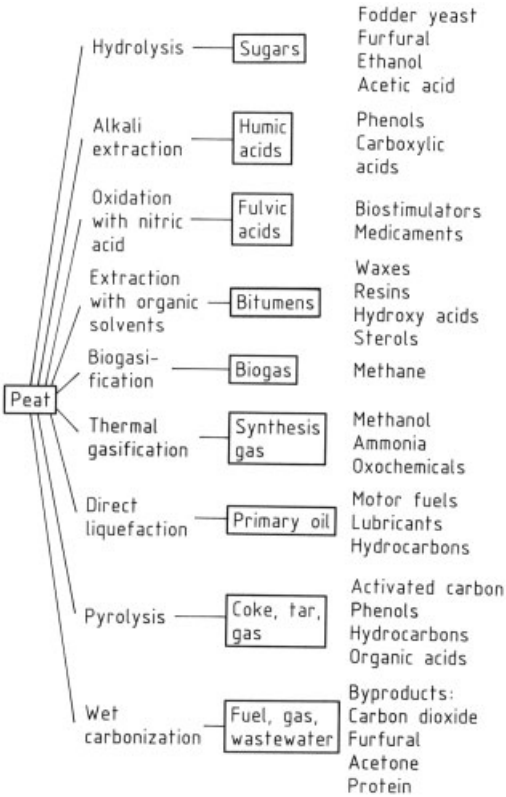

Figure 8. Peat conversion technologies [10]

The technology of purely energetic use of peat in (4) has had the greatest importance. The same applies, to a much lesser degree, for the processes assembled in (5), which have only recently regained outstanding importance for peat.

The processes (1) – (3) are responsible for the degradation of biomass. The processes mentioned in (3) are of purely scientific importance at present, especially since the products may be very heterogeneous or of complex composition, and cannot be isolated and used directly.

Technological, Structural, Stoichiometric, and Energetic Criteria. In contrast to coal formation, the chemistry of peat formation does not go through a coalification process so that the H/C ratio of peat is much more favorable for thermochemical conversion than that of coal, especially if the peat is first gasified [38]. Whereas synthesis gas production from coal can start out from a great variety of technologies or processes successfully practiced on a large scale since the 1920s, for peat the special know-how required for this purpose has been worked out only from the 1970s on [40]. The thermochemical processing of lignites has made a considerable contribution to the high-temperature process.

Feasibility of Economic Realization and Practical Importance. The economic feasibility of peat as an alternative carbon source for thermochemistry is influenced or determined not only by energetic and technological and/or stoichiometric aspects, but also by the cost of competitive raw materials. Nevertheless, the commercial use of peats as a source of raw material, carbon, and energy has a long tradition, and it cannot be ruled out that the production of C_1- and C_2-carbon building blocks from peat (e.g., synthesis gas, methane, methanol) could become competitive if prices for crude oil and natural gas continue to rise. The example of Finland shows this to be possible under certain national economic assumptions [23], [38], [39].

3.3.2.2. Peat Processing Technologies

The various processes for production of thermochemically or physicochemically upgraded peat products (see Fig. 8) have reached different levels of technical maturity. It is probable that peat will be used for the production of synthesis gas and fuel gas and to a smaller extent of liquid fuels, whereas regenerable raw materials, especially wood, will be used more for the production of higher-valued chemicals [41], [42].

Safety Aspects. Numerous explosions and fires in peat processing plants during recent decades have shown how reactive peat can be under certain conditions during its processing. Some peats have twice the oxygen content of the lignites, which are known for their high reactivity (see Table 1). It is therefore important to know the explosive behavior of raised-bog black peat as a function of its botanical composition, water content, particle-size distribution, and possible additives during its processing.

In former years, the explosion and fire behavior of the less decomposed white peat, which was

processed as a soil improver, were studied [43–45]. In the mid 1970, diminishing white peat resources have resulted in increased use of the more highly decomposed black peat. Combustion, coking, and gasification processes of highly decomposed peats have simultaneously become more important. Figure 9 shows the explosion characteristics of peat dusts. The explosion characteristics of peat are especially important in conversion processes under high pressure conditions, but dust explosions and self-ignition can also occur in all other operations of fuel processing.

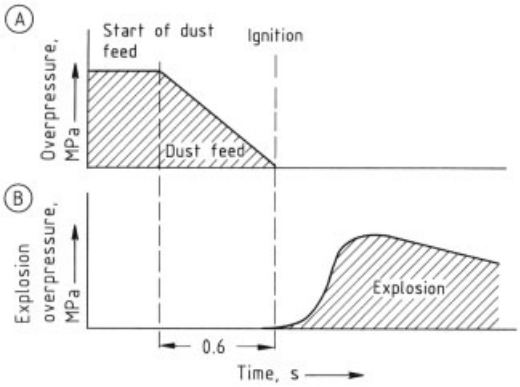

Figure 9. Explosion characteristics of peat dust
A) Dust supply vessel; B) Explosion vessel

In Finland, the explosive properties of native grades of peat were investigated [45].

Even *highly decomposed peat* is extremely susceptible to dust explosion. This sensitivity decreases only at high water contents and large particle sizes; thus milled peats and dust from fuel peat (lump peat) are not explosive because of their high moisture content (≥ 40 wt %).

With *predried peats*, care must be taken that the residual oxygen content of peat dusts does not exceed 12.5 vol %, even in bunkers blanketed with nitrogen [45], [46]. In this case peat dust explosions are no longer possible at normal pressure.

Peat coke dusts, by contrast, have an ignition temperature of about 560 °C, even at particle sizes less than 200 µm. They are therefore effectively nonexplosive, even at moisture contents < 10 wt % [47].

Wastewater. In hydrolysis, thermochemical liquefaction, low-temperature carbonization or coking, wet carbonization, and gasification of peat, contaminated wastewaters are always produced if these processes are not specially designed to be dry technologies free of wastewater. The (low-temperature carbonization) waters usually have a very high content of highly toxic phenols, aliphatic hydrocarbons, acetic acid, acetone, etc. This leads to total organic carbon (TOC) values of ≥ 30 g/L and creates special problems in the treatment of these aqueous solutions.

Therefore, modern peat processing technologies aim at avoiding production of wastewaters, which can be achieved above all by drying the raw material as thoroughly as possible.

Anaerobic Fermentation (Methane Production). The production and further processing of peat from bogs is technically demanding and costly. There has therefore always been an interest in leaving the raw material in the deposit and converting it directly on site to methane.

History reveals a sufficient number of examples of methane production from the very varied peat bog regions of Europe [48], [49]. Anaerobic fermentation processes with methane gas formation are sufficiently well known in nature [50]. In addition, large amounts of atmospheric methane originate from bogs [51], where the decomposition of plant and animal matter takes place in various processes. These processes are generally referred to as decay, rotting, peatification, or putrefaction, but there are no sharp separations, but transitions, between them. The atmospheric oxygen and water (moisture) present, the nature of the material to be decomposed, and the microorganisms taking part in the decomposition, all play an important role.

The actual feasibility of methane gas production from bogs is discussed in [49].

In Sweden, systematic attempts have been made to produce methane gas from bogs by the *Vyrmethane process* [52]. This process exploits the fact that the external appearance of the bogs is scarcely altered by a plant for gas production. The bogs are not seriously harmed, because even the prior drainage of the bog substance, which is indispensable for all other methods of peat production, must be dispensed with.

In the selected bog, a liquid circulation process is set up, by means of which bacteria-inoculated water flows through the peat lay-

ers and absorbs the methane gas formed by the methane bacteria in dissolved form. The methane-containing water is then transported via a drainage pipe system to an overground degassing station and from there to the consumer.

The residence time of the water in the bog depends on the physicochemical conditions in the bog "decomposition reactor" which determines the activity of the methane bacteria, i.e., their capability and rate of gas production and the dissolution rate of the methane in the water transport medium. The injection pressures of the circulating medium, i.e., the pressure difference between injection and extraction or production wells, essentially depend on the flow rate of the carrier liquid in the bog substance, which is determined by the porosity of the peat layers.

Separation of the methane from the circulating liquid occurs in the overground degassing station by pressure release inside large tanks. From there the circulating water again enters the bog, where a proportion of the methane-gas-producing bacteria are again stimulated to gas release and the methane released can be absorbed by the water and transported to the surface.

In pilot tests in a bog area of 25 ha and in a long-term test in a 150 ha bog, methane was formed continuously and at an increasing rate, during which an average methane gas concentration of 20 mg/L was measured. This methane content was three times greater than that originally found in the peat.

The advantages of the Vyrmethane process compared to the conventional peat production methods are:

1) the simple type of equipment and operation of the (methane) production process
2) the effective energy transport system
3) the original character of the bog is scarcely altered
4) year-round production
5) no disadvantages from changing quality characteristics of the peats (e.g., ash content in the deposit)
6) no movement of the peat substance
7) suitability even for small peat deposits
8) no large central installations for distribution, transport, and conversion of the methane into other energy forms
9) usability even of uranium-containing peat deposits
10) no adverse environmental effects
11) applicability of simple process routes for conversion of methane into other energy forms (e.g., electric power, methanol).

In the bogs of Central and Northern Europe, natural methane gas production mainly reached its peak in late summer and early fall (provided that mostly cellulose-rich reed peats had taken part in the fen formation) when the soil warmth stimulated methane bacterial activity, but was scarcely or not at all observable in the other seasons.

In the subtropics and tropics, reed peats and fenwood peats contribute greatly to the structure and have a rich mineral supply. Fens are thus formed instead of the raised bogs which prevail in temperate latitudes. In these fens the bacteria responsible for the decomposition process find more ideal conditions than in raised bogs [9].

Systematic methane production from bogs of temperate geographical latitudes could be initiated in fens by inoculation with specially cultured methane bacteria. Climatic and other reasons favor the application in tropical and subtropical bogs and the process will very probably be unable to satisfy the basic conditions (e.g., for aerobic fermentation) for a noteworthy production of methane.

Thermochemical Liquefaction. When wood and cellulose are heated in water or aqueous alkalis above ca. 175 °C, they are appreciably dissolved. During this process, small amounts of gases are formed, e.g., H_2, CO, CO_2, and hydrocarbons, and also dark oils and carbonized residues, particularly at high process temperatures [53].

There are basically two thermochemical routes for the synthetic production of liquid fuels: direct and indirect liquefaction (see page 21).

For the *direct liquefaction* of peat, as for other fossil energy sources and for biological waste products, a lignin-based aromatics chemistry or the chemistry of the xylose structural units of hemicellulose must first be developed in order to reach total conversion without burdensome residue problems [54]. Peat is usually converted in the wet raw state, using high pressures in the

presence of carbon monoxide and/or hydrogen in a reducing atmosphere [55–58]. This results in thermal cleavage of the polymeric starting material and saturation of the radicals formed by hydrogenation.

Coking. Coking is defined as dry distillation or carbonization in the absence of air. For the production of processed products from peat rich in carbon, i.e., cokes, thermal and physicochemical processes must be used. In carbonization, peat coke remains as a residue of the dry distillation feed material (peat) which contains all mineral constituents (including ash).

Peat coke is usually produced by coking highly decomposed sphagnum peat from raised bogs. This is generally used in the form of air-dried sods but can also be used as granular material. The loose peat itself can also be used as raw material. Byproducts of peat coke formation are carbonization gas, peat tar, and carbonization liquor.

Early trials were made of coking peat in piles and using the peat coke so obtained as a replacement for fat coal and reducing carbon in metallurgical works. True peat production started at the beginning of the 1900s in shaft furnace retorts (see Fig. 10) in the Elisabethfehn peat coking plant and ca. 15 years later in Klazinaveen in the Netherlands in similar furnaces. For many decades these coking plants remained the only plants of their kind worldwide. In 1976 an attempt was made in Finland to produce peat coke in rotary kilns. Although the Finnish peat coke works ceased production after approximately eight years of operation because of its unprofitability, the old coke works in Elisabethfehn worked for almost 58 years without major interruption before it had to cease production for environmental protection reasons. Only the peat coke works in Klazinaveen is still in operation, but has been continuously modernized and now uses rotary kilns up to 36 m in length and 6 m in diameter for carbonization and activation of peat.

Since the 1970s the quality demanded of peat coke and its processed products has continuously risen. As a result the peat production technologies used were improved and, if necessary, replaced, and processes and equipment for coking were adapted to meet new technical, economic, and ecological demands and procedures for the product.

Raw Material Selection. The composition of peat mainly depends on its degree of decomposition. For the production of peat coke, therefore, only black peat can be used, because the more decomposed the peat, the higher is its carbon content. For the elemental composition of black peat see Section 1.3.

Peat is a poor conductor of heat and has no caking power. For this reason the horizontal chamber coking process, for example, which is well established for coking hard coal, cannot be used. Peat coke becomes denser and stronger as the density of the shaped piece of starting material increases. This is of great importance, since many of its applications depend on its density. Accordingly, some of the recent developments are directed to the production of peat coke from fine-grain or granular materials with high friability.

Process Selection. The grain and lump size of the raw material to be coked and the lumpiness and quality of the coke demanded by the market influence the choice of the peat coking process. Techniques for the production of fine and lump coke are different; they can also be distinguished by the type of heat transfer used (direct or indirect).

In principle, most coking processes for the thermal processing of hard coal and lignite are suitable for peat coking [23], [59–61]. The individual coking processes are described in → Coal Pyrolysis; → Gas Production; → Hydrogen; → Oil Shale.

Since peat production and peat coking are interdependent, out of fourteen possible coking processes for peat only the rotary kiln and the fluidized-bed process are economic and flexible enough to meet all the conditions for the production of marketable peat coke products.

If fuel peat is available in sufficient quantity and reasonably priced, with constant moisture contents of 30–35 wt %, a modified *Wielandt process* produces the best possible coke. This coke can no longer be produced economically in a peat coking plant with battery construction and partial automation, even if the capacity is as high as 25 000 t coke per year, and even if the fuel peat sods are thermally predried to prevent formation of tar and coke liquor products [23], [62].

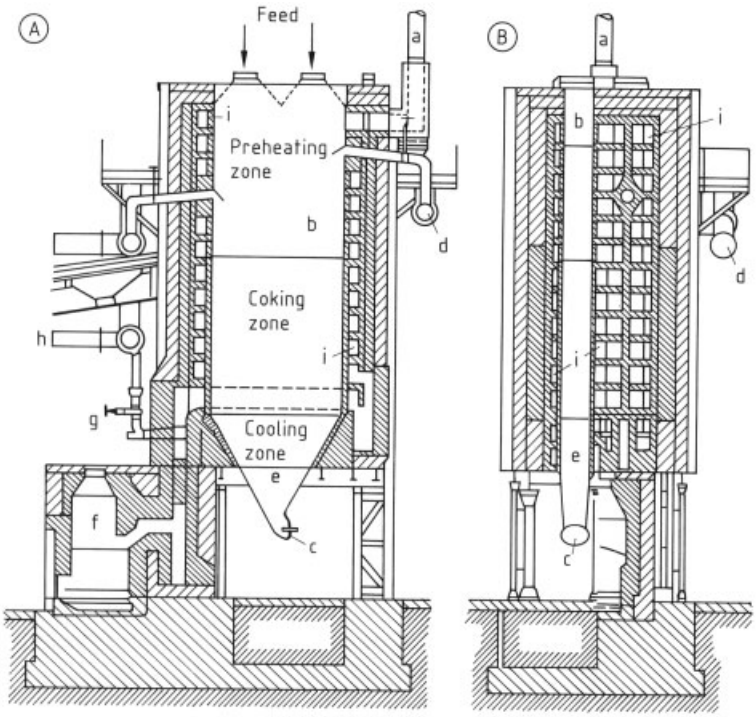

Figure 10. Longitudinal section (A) and cross section (B) of the Wielandt retort
a) Stack; b) Coking chamber; c) Peat coke outlet; d) Raw carbonization gas; e) Cooling zone; f) Peat gasifier for process starting; g) Burner; h) Cleaned carbonization gas; i) Heating channels

If the peat moisture and lumpiness fluctuate within wide limits, *rotary kiln coking with predrying* is more flexible and more economic, especially since it is able to cope with both sod peat and large amounts of peat dust. Its advantages over the Wielandt process include its simpler reactor construction, lower investment, similar harmlessness to the environment, and possibly higher specific throughputs [63].

If the quality of the available lump peats and fuel peats deteriorates, and if the peat coke is subsequently processed further, it may be more logical to use the *fluidized-bed technology* to obtain fine-grain peat coke or granular material as product. Even the cheapest loose peat can then be used, and with little expense on equipment, a semipeat coke or high-temperature coke product which can be of different qualities can be produced. By appropriate preparation of the feed peats the fluidized-bed technology can be operated without gas condensation, and therefore without the formation of tar and sulfur liquor. With less expenditure on equipment than for rotary kiln and shaft furnace coking, the fluidized-bed technology also has the lowest investment costs. Therefore, for specific applications of peat coke (e.g., for activated carbon production), this process is probably the most economic method, irrespective of plant size.

Coking of Sod Peat. A high-grade carbon concentrate with a calorific value > 31.4 MJ/kg and a very low volatiles content is formed from fuel peat, shaped peat pieces, or loose peat with a calorific value of up to 17 MJ/kg by coking at $\geq 800\,°C$ for 10 – 18 h in Wielandt retorts (fuel peat sods) or for a few minutes in the fluidized bed (fine-grain feed materials).

The following characteristic phases can be distinguished during peat coking:

Preheating Phase (20 – 120 °C). Part of the water is removed to form carbon monoxide and carbon dioxide as minor products.
Drying Phase (120 – 200 °C). Here the absorbed water is driven off. Carbon monox-

ide and carbon dioxide formation continues. The evolution of gases increases.

Precarbonization Phase (200 – 450 °C). Up to 280 °C most of the tar is formed along with hydrogen, methane and other hydrocarbons, methanol, phenol, acetic acid, ammonia, acetone, and small amounts of carbon monoxide and carbon dioxide. At 280 °C decomposition of the combustible material becomes increasingly vigorous, i.e., an exothermic reaction starts and continues spontaneously with greater heat liberation, reaching its peak at 350 °C.

Main Carbonization Phase (450 – 520 °C). By reheating the retort, the process temperatures are driven higher, so that gas evolution also increases. Methane, hydrogen, and carbon monoxide are the main products. Hydrocarbon, ammonia, and some tar are also formed, though in smaller amounts than before. The fixed carbon content of the coke being formed increases greatly, the hydrogen and oxygen contents decrease significantly and its nitrogen content remains more or less unchanged.

Postcarbonization Phase (520 – 1000 °C). Here the carbon substance consolidates. At the end of the true coking process (900 °C), the weight loss and shrinkage reach constant values and the finished coke finally contains 95.5 wt % fixed carbon, 1.17 wt % residual moisture, and 3.28 wt % ash (all relative to dry matter).

In continuous processes, e.g., in the fluidized bed and in rotary kilns, the transitions between the individual stages of carbonization are not so clearly defined. The stages overlap or proceed side by side. Nevertheless, if the process is carefully controlled, an exceptionally reactive coke is always formed. The higher the coking temperatures, the more the tars are decomposed to give hydrogen, methane, heavy hydrocarbons, and carbonaceous residues. Rapid coking produces cokes which are richer in fixed carbon if the atmospheres in processes and plants are oxygen-free and heating is indirect. In this way the yields shift and the energy content of the gas increases.

Figure 11 shows the product distribution from peat coking in externally heated Wielandt retorts. The product yield is determined mainly by the type of peat and its moisture content, the coking process, the coking time, and the heating rates and process temperatures.

The heat requirement for dry distillation of peat in externally heated retorts is 0.84 – 1.09 MJ/kg of peat (dry matter). This can be significantly reduced in modern processes, e.g., in a modified Wielandt process or in similarly arranged fluidized-bed and rotary kiln technologies, by various steps relating to the equipment and process engineering if compact designs are used.

If the production of valuable substances is largely abandoned and an integrated heating plant is built instead, the combustion of carbonization gas and tar enables the total process heat and steam requirement to be covered, not only for coking but also for peat predrying coupled upstream and possibly for an added activation of the coke product. Surplus energy (e.g., in the form of electricity) can be delivered to external customers.

Properties and Uses of Peat Coke. Peat coke has a structure with fine pores and thin cell walls and therefore has a much lower density than other cokes. The more marked the shrinkage during coking, the denser and more abrasion-resistant the coke is. High-temperature cokes are more abrasion-resistant than pure low-temperature cokes. Shrinkage values of 67.5 % are readily achieved.

The most important physical properties of peat coke are as follows:

Grain size	≤ 200 mm
Largest grain	$30 \times 60 \times 200$ mm
Apparent density	0.664 g/cm^3
True density	1.776 g/cm^3
Porosity	75 %
Bulk density	
Fine dust	450 kg/m^3
Coarse coke	290 kg/m^3
Specific surface area	≤ 220 m^2/g
Electrical resistance	insulator
Ignition point	230 °C
Compressive strength	390 N/cm^2
Abrasion resistance	87.2 %
Ash melting behavior in weakly reducing atmosphere	
Softening point	1300 °C
Melting point	1350 °C
Flow point	1450 °C
Lower calorific value	29 – 33 MJ/kg

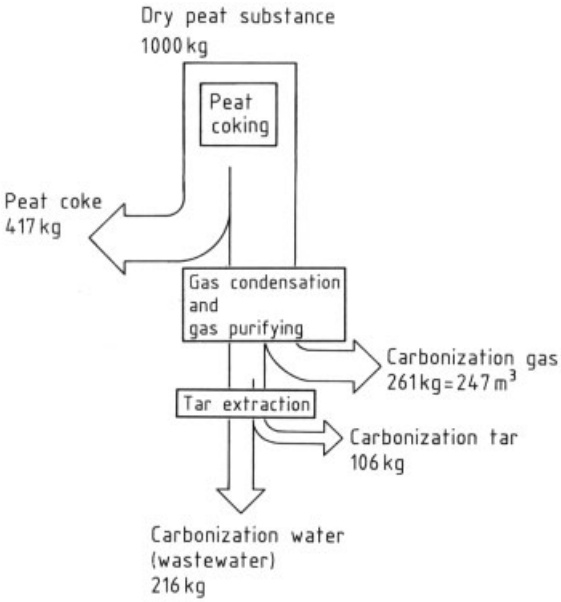

Figure 11. Material balance of peat coking

The chemical composition of peat coke (wt %, dry matter) is as follows:

Ash content	3.5
Fixed carbon	91.06
H_2	1.51
N_2	1.39
S, total	0.21
Iron	0.14
Calcium	0.38
Magnesium	0.48
Aluminum	0.20
Silicon	0.95
Sodium	0.09
Potassium	0.06
Phosphorus	0.03

The specific surface area of peat cokes are generally $< 200\,m^2/g$, but on activation with steam these can be readily adjusted to 800–1600 m^2/g because of the high reactivity of the peat coke to gasification agents.

Peat coke is mainly used in electrothermics, for the production of activated carbons, and in metallurgy, especially in foundry technology and for the carburizing of iron. Figure 12 shows the various applications of peat coke.

With reduced availability of wood charcoals from high-quality deciduous trees, the importance of peat coke as a special coke may increase. Processing technologies are much further advanced than those of wood charcoal and its joint processing with other pure carbons present special advantages for improving the quality of the end product.

Wet Carbonization and Wet Oxidation. Wet carbonization and wet oxidation produce improvements of the feed peat, in the reduction of its water content to $\leq 30-40$ wt % and in the conversion of pasty to friable-dense and fiber-containing (raw) bog substances into granular-powdery, free-flowing, and dried matter resembling fine coal. This is achieved by irreversible destruction of the colloids in the raw peat [63].

Wet Carbonization. In wet carbonization a peat suspension (or milled peat) is reacted without additional oxygen in an autoclave at ca. 2.2 MPa and 220–250 °C. The colloidal structure of the peat is broken down by destruction of its carboxyl and hydroxyl groups. The oxygen content of the peat decreases because carbon dioxide and water are formed, whereas the calorific value of the product increases and its dehydration is considerably facilitated.

A process for the wet carbonization of peat was first patented in 1900. Since then repeated attempts have been made to improve the process technically and economically [22], [63–67]. Up to now, no modern wet carbonization plant for peat exists on the industrial scale together

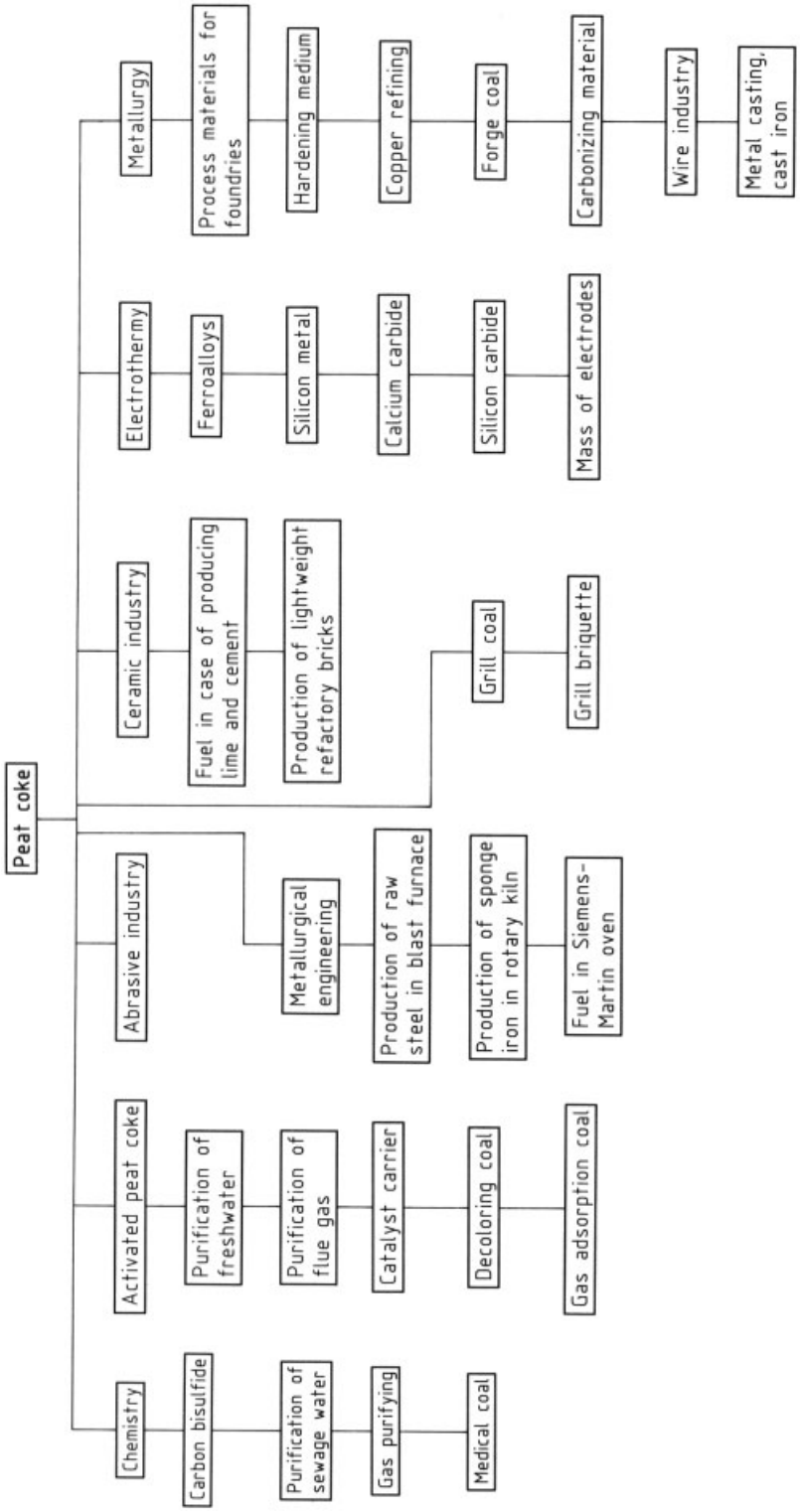

Figure 12. Market areas for peat coke

with operative wastewater purification. There is therefore doubt about the practicability of the process.

Wet Oxidation.

Wet oxidation is now generally used industrially especially for destroying organic pollutants in wastewaters. Process and equipment are in principle the same as for wet carbonization.

The essential difference between the two processes is that in wet oxidation, oxygen at a pressure of up to 2.0 MPa is fed to an autoclave, leading to rapid heating of the peat present in aqueous suspension and to destruction of undesirable constituents (e.g., phenols). The remaining carbonized peat substance, however, has a lower calorific value than that from wet carbonization. It can be mechanically dried and has the same applications as the products of wet carbonization.

The wet oxidation process is still in the exploratory stage.

Gasification. Peat can be converted into gaseous fuels in various ways, and the gas produced is always a mixture of various combustible and incombustible components. Figure 13 shows a survey of the possible process steps in peat gasification, including carbonization.

Depending on the gasification agent used (air, oxygen, steam, CO_2, etc.), the nature of the contact between peat and gasification agent, and the gaseous product formed, various types of gasifier are used (see Fig. 14) [68–71].

Peat gasification processes are usually carried out autothermally. According to the processing sequence and the melting point of the peat's incombustible gasification residue, the residue is formed either as solid ash or as liquid basic slag. To avoid breakdowns as a result of slagging problems in the gasifier, the composition of the peat used must be accurately defined. A reliable, complete, and continuous outward transfer of the ash from the various types of gasification reactor is particularly important.

Synthesis gas or fuel gases can be produced together with process heat and electric power by gasification from peat better than by all other peat conversion processes.

In the past, repeated attempts were made in Germany, Scandinavia, and the United States to use peat for the production of fuel gas or synthesis gas but did not lead to an industrially usable process. The main operating problems which occurred (incompletely dried peat, generator operation giving uneven gasification, slagging problems, large production of unusable aqueous condensates) could not be brought under control.

As a result of the first increase in the oil price in 1973, renewed trials were made to produce industrial gas from peat in the United States [72], Finland [68], [73], and Germany [23]. In the United States high-pressure gasification was preferred for production of *synthesis gas.* In Finland fixed-bed gasification was first investigated unsuccessfully; the high-temperature Winkler gasification (HTW) was then used as the route to synthesis gas production.

In Germany, the Winkler gasification route was also followed at first, though with the objective of generating *process heat and room heating* for decentralized heating systems in small manufacturing plants. However, this fluidized-bed gasification technology was less economic than a novel modified combination of fixed-bed gasification and combustion. Therefore further development was devoted to producing lean gas by atmospheric gasification of peat for so-called multistage combustion systems (see Section 3.3.3.2.). All three methods of peat gasification (high-pressure, HTW, and atmospheric gasification) proved to be technically feasible.

For peat, only the further development of *atmospheric peat gasification* processes and plants shows promise: new routes for utilizing peat as an energy source have opened up with synthesis gas production from milled peat in Finland [68] and with the modern nonpolluting multistage combustion systems in Germany [74].

If the objective of gasification is the production of a fuel gas for raw material and energy conversion purposes, a synthesis gas with the highest possible proportions of CO and H_2 must be aimed at, irrespective of the process. The prerequisites for this are:

1) predrying of the peat to residual water contents < 10 wt %,
2) adaptation of the peat grain size to the process requirements (e.g., for fluidized-bed gasification, size reduction to < 1 mm), and
3) use of pure oxygen and steam as gasification agents (no dilution of product gas by atmo-

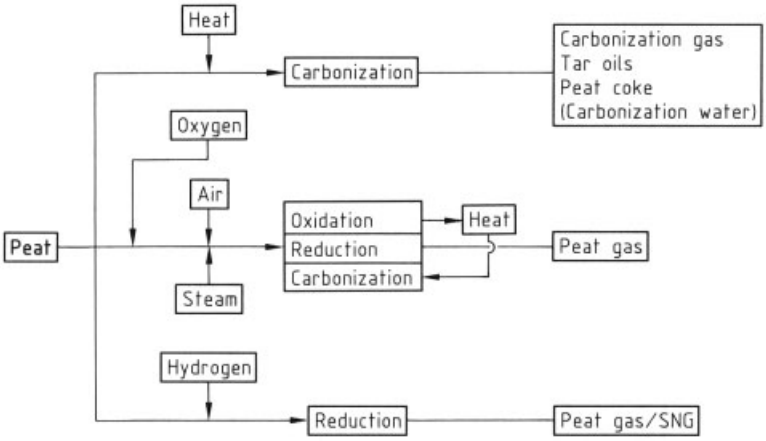

Figure 13. Schematic processing possibilities in peat gasification

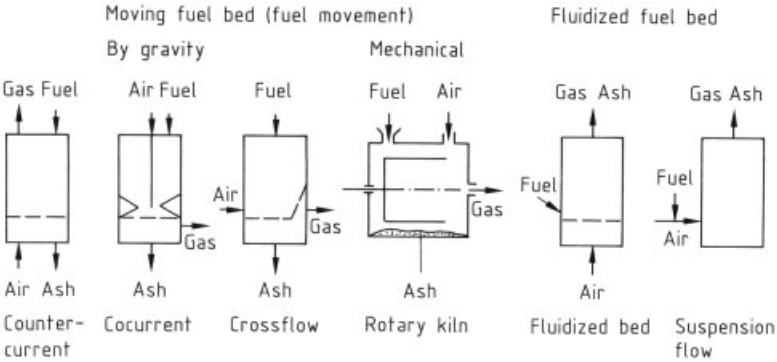

Figure 14. Schematic of gasifier types resulting from different feed materials and different gas supplies and their basic designs

spheric nitrogen) and reduced formation of condensates at $> 900\,^{\circ}C$.

For the gasification of peat the construction of combined power stations could be the most advantageous route (→ Gas Production, Chap. 7.3.), since substantial increase in efficiency and the best possible utilization of the peat (a low-calorific value fuel) would be achieved. A combined power station with integrated gasification stage is shown in Figure 15.

Apart from the costs, the size and lifetime of the required resources are essential factors [75].

Currently, neither conventional fluidized-bed gasifiers nor the incompletely developed Finnish fixed-bed gasifiers are suitable for the production of heat or process heat in decentralized heating plants. New German process developments have shown that gasification technologies and equipment are available which are suitable, economic, and reliable in service, even if only peat with moisture contents up to 45 wt % must be used as starting material (see Section 3.3.3.2). Peat gasification may only gain in importance if energy prices rise again.

Activated Carbons from Peat. Activated peat is a valuable type of activated carbon. Internal surfaces of $\geq 1600\,m^2/g$ can readily be produced.

Whereas until a few years ago activated carbon grains and powders of very widely varying degrees of activation were the main products of activated carbon formation from peat, the importance of shaped-peat activated carbons (granular materials) has currently risen.

For the production of activated carbons, powdered and granulated low-ash raised-bog black peat is the most suitable. Owing to its high volatiles content, this black peat previously had to be carbonized before gas activation (see page

18), but with chemical activation and the appropriate degree of predrying it could be used directly.

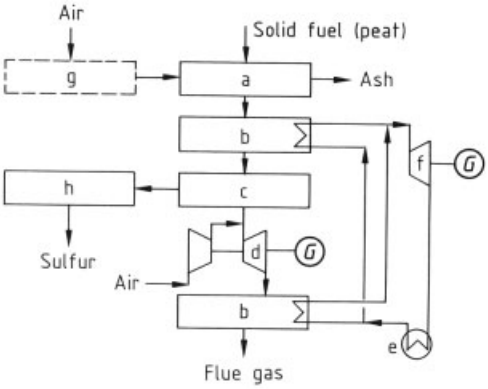

Figure 15. Schematic of peat gasification in a combined coal – peat power station
a) Gasification; b) Heat recovery; c) Gas purification from dust and H_2S; d) Gas turbine; e) Condenser; f) Steam turbine; g) Air separation; h) Sulfur recovery

With the gas and steam activation of peat carbon currently applied it is no longer necessary to use peat sods of a certain grain size as starting material. Activated carbon is produced either directly in a one-stage process at 0 – 1000 °C or in a two-stage process with peat coke as intermediate. For reasons of energy and ecology the peat product concerned should have less than 25 – 30 wt % residual water content before activation. This ensures the process-engineering reliability of an autothermal process without wastewater, increased outputs, and better product qualities.

For the production of activated peats, shaft furnaces, rotary kilns, and multiple hearth and fluidized-bed furnaces are suitable. The shaft furnaces and rotary kilns should preferably be charged with peat sods while multiple hearth and especially fluidized-bed furnaces that are used for the production of specific peat activated carbons require charging with a narrow grain size range.

Because of more stringent ecological constraints, shaft furnaces could no longer be used for the production of activated carbons. Rotary kilns will also be abandoned if this technology cannot be further developed despite the fact that the largest amount of activated carbon from peat are still made in them. In contrast the importance of multiple hearth furnace activation and the fluidized-bed technology will increase, because only these technologies produce new shaped activated carbons of greater hardness which can be thermally reactivated after use.

So far up to 250 000 t raised-bog peat sods with an average residual water content of 30 wt % have been processed annually to activated carbon, corresponding to 2×10^6 m^3 of raw peat bog substance. In future this figure might even increase if shaped activated carbons are increasingly produced from peat – pure carbon mixtures.

3.3.3. Peat as an Energy Source

The potential for the use of peat as fuel depends on the size of a country's peat deposits and especially on competitiveness of peat with other energy sources. In the Scandinavian countries, chiefly Finland and Sweden, interest in the more intensive use of peat as an indigenous energy raw material has grown considerably since the mid-1970s [76], [77].

The suitability of peat for use in electric power generation is discussed taking Finland as an example. Finland is rich in peat deposits. These cover more than 10.4×10^4 km^2, or one-third of the area of the country. Hitherto far less than 1000 km^2 of the 5000 km^2 of economically profitable mire deposits have been developed for peat production.

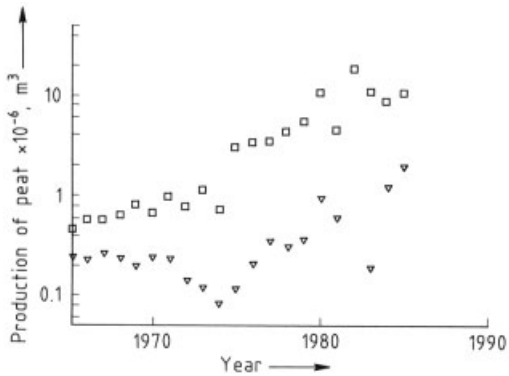

Figure 16. Volume growth in the production of milled peat and sod peat in Finland
▼ =Sod peat; □=Milled peat

Figure 16 illustrates the level reached in the production of milled peat and sod peat; milled peat has a 90 % share of all peat production. Cen-

tral to the exploitation of peat is its use as an energy source (see Fig. 17).

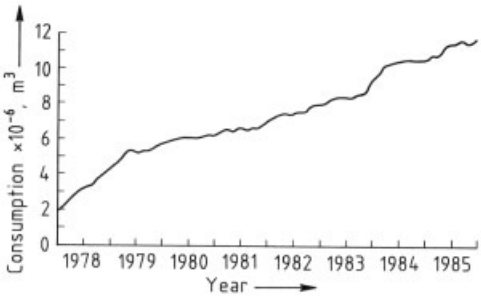

Figure 17. Growth of peat consumption for energy purposes in Finland

The total expenditure required for developing the peat bogs and raising peat production to its present level, was ca. $500-1000\times10^6$ finnmark (ca. $\$120-240\times10^6$). In the financial support given by the Finnish government it had also to be taken into account that the construction of central heating plants, combined heating and power stations, and power stations for electricity generation is more cost-intensive than actual peat production. The investment required per megawatt to install a heat supply for a large appartment block is ca. 85 % of that for the corresponding thermal power station.

There were essentially two reasons why Finland favored peat as an energy raw material. Peat has always been more cost effective in comparison with other fuels. Milled peat in particular was always considered to be the most inexpensive of all fuels available to Finland. Reportedly, the price of sod peat has always been lower than that of petroleum products in the last decade.

Although the fuel cost is only a part of the cost involved in obtaining energy, the production of heat and electric power based on peat in medium-size and larger power stations ($>5\,MW_{th}$) is said to have been profitable in Finland for many years.

However, the security of the energy supply was more important for the Finnish government, because the energy demand per capita in Finland is one of the highest in the world. Finland is highly dependent on energy imports; about two-thirds of the energy consumed in the country must be imported.

In the mid-1970s petroleum constituted about 56 % of the national energy demand. Currently, none of the energy sources supplies more than one-third of the total energy demand. Fuel peat has been able to increase its share, which was negligible at the start of the period under consideration, to more than 3 %.

Nevertheless, peat remains an "exotic fuel" [78]. Yet it has the advantage over other primary energy sources of a much lower content of pollutants (especially sulfur). When modern combustion techniques are applied, peat is advantageous with regard to the unavoidable residual emissions.

Drying of Peat. Any profitable use of peat requires an efficient dewatering and drying of this material. Possibilities for the dewatering and drying of raw peat are shown in Figure 18. Figure 19 shows the increased suitability of peat as a fuel as a function of its residual moisture content [79].

With regard to the economy of the subsequent further processing of the peat as a fuel or raw material, air-drying of the raw peat on the field is the only feasible process. *Air-drying* proceeds in the following steps [12]:

1) In the first step, the water bound mechanically on the surface and in the large pores evaporates from water-saturated peat surfaces. The evaporation depends on the humidity and flow rate of the air.
2) In the second step, the water of the capillary walls evaporates out of the now unsaturated peat surface. The evaporation rate depends essentially on the capillary length, and so on the layer thickness of the lump of peat to be dried; in this step a final water content of 45 – 55 wt % can be obtained.
3) In the third step the drying rate decreases because of the reduced diffusion rates in the narrower pores of the peat, which are constantly shrinking.

The *thermal drying of fuel peat sods* (lump peat) must be the nondestructive continuation of the shrinking process begun on the field. The process used is characterized by multizone drying at $<100\,°C$ with integrated heating-up at high initial air humidity and cooling of the sods downstream [33].

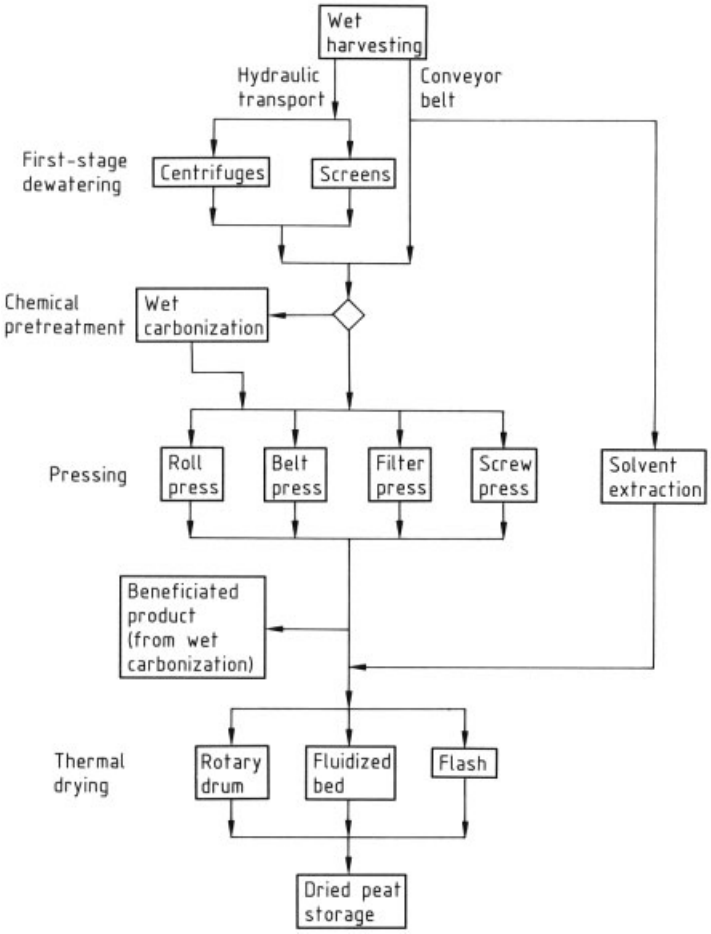

Figure 18. Basically possible and so far technically proven methods for drying peat [79]

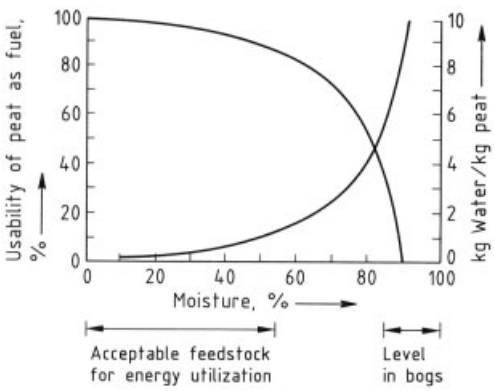

Figure 19. Usability of peat as fuel as a function of its residual moisture content

In the *thermal drying of milled peat*, two different procedures are used, depending on its subsequent use:

1) If milled peat is used as a power station fuel, the flue gases produced are used for drying. Herewith the often highly fluctuating initial moisture content of the milled peat (45 – 60 wt %) can be lowered at least to an average of 45 – 50 wt %.
2) In the drying of milled peat for peat briquette production to maximum residual moisture contents of 12 ± 5 wt % , more thermal energy is necessary than when drying for power station purposes, i.e., advanced equipment is required in combination with higher drying costs. Because of the increasing flammability of the peat with increased drying,

additional precautions must be taken. Any change in the residual moisture content of the dry peat has a much greater effect than all other briquetting parameters on the briquette hot compressive strength [80]. Therefore, it is desirable that thermal drying of milled peat uses starting materials of lower average moisture content than is the case in power station operation. However, only the milled peat favored by the briquette works (maximum average moisture content 45 – 50 wt %) can be supplied in particularly warm and windy summers.

In future, steam fluidized-bed drying will be used instead of the vertical tubular drying used up to now. Steam fluidized-bed drying is very much more effective and more flexible in operation than vertical tubular drying. In steam fluidized-bed dryers (see Fig. 20), the peat's binding potential for water is reduced to a minimum by reduction of the particle size, hence causing the destruction of the capillaries. Steam fluidized-bed drying has the advantage that it can be controlled such that overdrying of the milled peat can be avoided [81], [82], which lowers the risk of flammability.

3.3.3.1. Peat Power Stations

For centuries, black peat in the form of fuel peat sods was used as an industrial and domestic fuel. In Germany the generation of peat-based electrical energy started in 1908. All power stations were dependent on the supply of air-dried sod peat (fuel peat with 30 – 35 % water content). The power station in Wiesmoor with an output of 15 MW_{el}, consumed 100 000 t air-dried fuel peat annually. To generate 1 kWh_{el}, 1.0 – 1.1 kg fuel peat were required. After World War II the remaining fuel-peat power stations in Germany, Denmark, and the Netherlands were shut down. In the former Soviet Union, Ireland, and Finland, however, peat became increasingly important for energy supply. Whereas here also the first peat power stations were still based on sod peat, the current peat power stations use milled peat exclusively to increase the output.

The largest peat power station currently operating in Finland, Haapavesi, has an electrical output of ca. 120 MW_{el} and requires about 2.5×10^6 m^3 of milled peat per year as feedstock [10]. This power station, in contrast to others, runs exclusively on peat. Therefore, the quality of the peat must be maintained as uniform as possible throughout the year. Since ca. 1.3 –1.6 kg milled peat of 50 wt % moisture content are required for the generation of 1 kWh of electrical power, enormous amounts of milled peat must be made available.

The peat power station at Rauhalahti is described as an example to demonstrate the special features of a peat power station [77]. The power station must be continuously supplied with milled peat because large bunkers are deliberately avoided in peat power stations for safety reasons (self-ignition). Instead stockpiling of peat is provided for only 8 – 10 h of power station operation at full capacity (i.e., 2500 m^3 of milled peat daily). The delivered raw milled peat is directly transferred to a bunker belt conveyor which is used for uniformly supplying the belt conveyor flights to the power station and also allows the simultaneous sorting out of pieces of iron and steel with magnetic separators. Peat that has already been self-ignited can also be separated easily. The usable milled peat is fed to a *rotary disk deck screen*, with which oversize material (e.g., lumps of frozen peat, roots, pieces of wood and stones) are sorted out and fed to a crusher. After that the screen underflow and the crushed oversize material are transported via belt systems to the peat intermediate bunker, from which the milled peat is then withdrawn according to the requirement of the power station boiler.

Bituminous coal and heavy fuel oil are used for auxiliary firing and as reserve fuels. In the power station boiler, a dust firing principle which is a modification of that used in conventional lignite power station is used both for milled peat and for coal. 100 % of the power station's full output can be produced with peat, 70 % with coal, and 40 % with heavy oil. The Rauhalahti peat power station is equipped with turbine bypass valves, to enable full use of the steam without power generation, i.e., pure heating-station operation.

There are two peat firing and two coal firing lines. Separate burners are provided for peat and bituminous coal. It is possible in principle to change the proportions of the fuels used during operation.

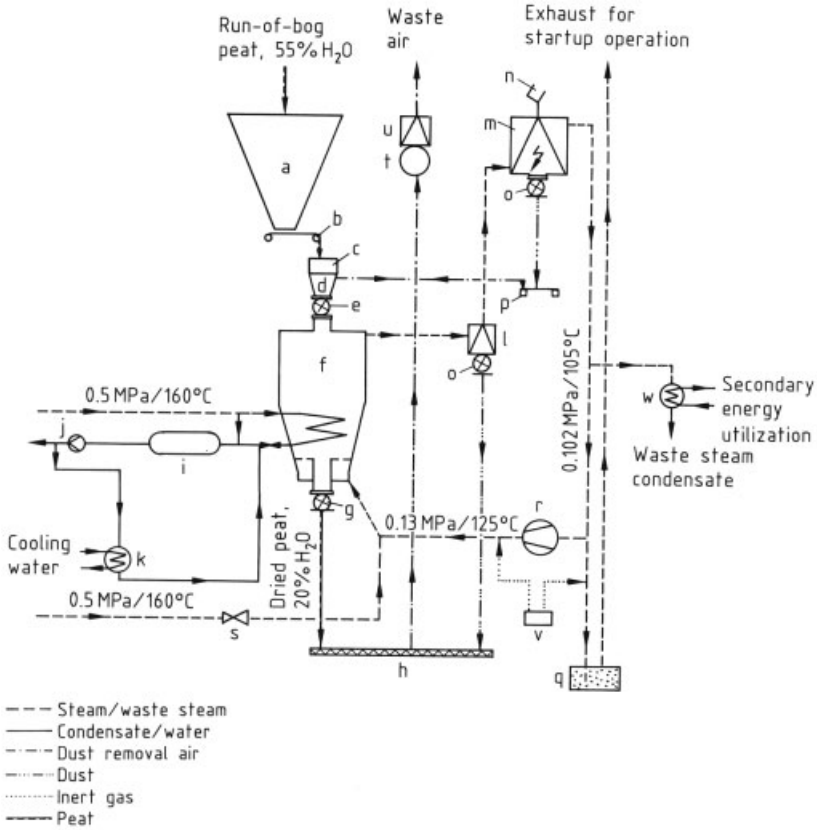

Figure 20. Schematic of process and plant principle of a steam fluidized-bed dryer
a) Run-of-bog peat bunker; b) 2 Controlled feed-regulating conveyors; c) 2 Mills; d) 2 Heated chutes; e) 2 Heated-rotor star-wheel feeders; f) Fluidized-bed steam dryer; g) 4 Star-wheel feeders; h) 2 Screw conveyors; i) High-pressure condensate vessel; j) Condensate pump; k) Condensate cooler; l) 2 Cyclone preliminary separators; m) Electrical flue gas cleaning filter system; n) 5 Pressure relief valves; o) 4 Star-wheel feeders; p) Dust conveyor; q) Water reservoir; r) 2 Centrifugal compressors; s) Steam state converter; t) Internal dust separator fan; u) Wet separator; v) Nitrogen cylinder vehicle; w) Waste steam condenser for waste-heat utilization (heat exchanger)

99 % of the dust (i.e., the ash) is removed from the flue gas stream in a single- chamber electrostatic filter before the flue gas enters the 130 m-high stack. Usually ca. 10 000 – 16 000 t/a fly ash are produced (30 000 – 50 000 t/a in the Haapavesi power station).

The experience in Finland is as follows: peat, as an indigenous primary energy source, can make a contribution to the energy supply of a country. In addition, the technology of the peat power stations can be developed toward a combined power station for the simultaneous production of electricity, process heat, and space heating, as well as of fuel gas or synthesis gas [83]. A further development of the HTW fluidized-bed gasification process (higher process pressure and increased solid circulation rate) is especially suitable for the production of synthesis gas and alternatively fuel gas, e.g., for use in a combined gas- and steam-turbine power station (for a description of combined cycle power stations, see → Gas Production, Chap. 7.3.).

Finally, another modern power station technology should be mentioned, i.e., a power station furnace with a circulating fluidized bed in place of the usual dust-blower furnaces (see Fig. 21). This arrangement offers less pollution in combination with more elegant process engineering. Fluidized-bed technology should advantageously be introduced to future power station construction before the development in the di-

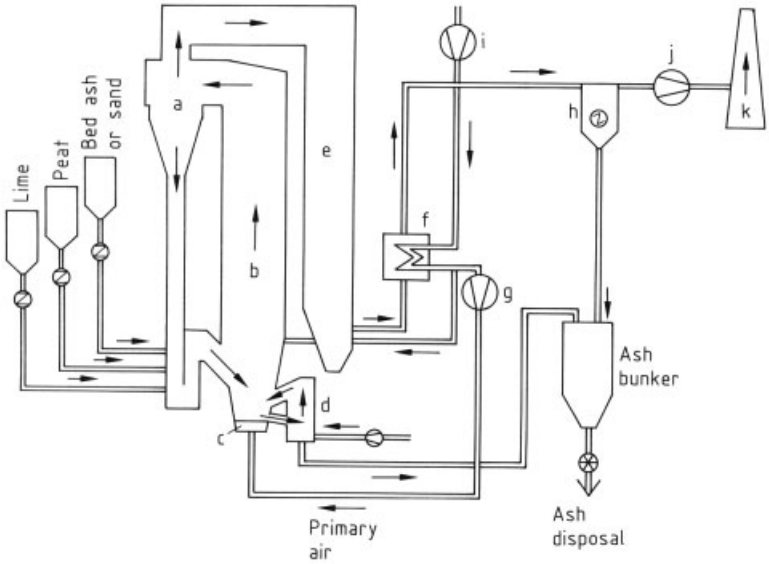

Figure 21. Process and plant layout of a circulating fluidized-bed furnace for peat power stations
a) Cyclone; b) Fluidized-bed combustion chamber; c) Nozzle plate; d) Ash discharge system; e) Second pass; f) Air preheater; g) Primary air blower; h) Dust filter; i) Fresh air blower; j) Flue gas exhauster; k) Stack

rection of combination power stations has taken place.

The Finnish example demonstrates how important an indigenous energy source can be for other areas poorer in fuel resources (e.g., Africa, Brazil, Indonesia).

In Finland ca. 3×10^6 t/a milled peat are consumed in 21 peat power stations, while in Ireland 3.3×10^6 t/a milled and fuel peat (sod peat) supply 8 power stations. In the former Soviet Union, the peat requirement of the remaining 15 peat power and heating stations is much higher (ca. 14.5×10^6 t/a). In 1965, 58 peat power stations were in operation in the former Soviet Union which consumed 28.9×10^6 t/a milled peat. Since then, most of them have been shut down because of difficulties of supply and have been replaced by nuclear, natural gas, and coal power stations [11].

3.3.3.2. Decentralized Energy Generation by Multistage Combustion

In the large peat power stations, full-load operation is mostly aimed at. This is very rarely the case for decentralized systems for generating process heat and space heating (e.g., heating stations in smaller and medium-size industrial plants and in residential districts) which must be adapted to fluctuating energy consumption of the consumers.

Such operational fluctuations increase the emission values of the power station. Accordingly, peat research institutes [84], especially in Finland and Germany, and manufacturers of solid fuel combustion systems [85] have been engaged in developing simple and much less polluting firing systems for peat. Even in partial load and fluctuating load operations, these systems must be guaranteed to be able to adhere to the permitted emission limits.

Process Principle of Multistage Combustion (MSC). Multistage combustion is geared to the full utilization of the energy content of peat and/or of production residues with a minimum calorific value of ca. 9.2 MJ/kg (required for the autothermic mode of operation of the MSC systems) which had previously to be disposed of unused.

The basic process principle of the multistage combustion system used for decentralized generation of process heat and space heating is shown in Figure 22 A. It is characterized by great adaptability to changing composition and

calorific values of the fuels and by simple basic construction (Fig. 22 B).

In a multistage combustion system, the complete combustion of the fuel peat takes place in three mutually coordinated stages:

Stage 1. Complete conversion of the fuels fed, largely conducted under fuel-rich conditions, to hot combustible gases and residual mineral ash.

Stage 2. Stepwise further heating (cracking) and complete conversion of the combustible gases into dry flue gases by stepwise addition of atmospheric oxygen at high temperature.

Stage 3. Use of the energy content of the flue gases for process heat generation.

Each multistage combustion system consists of the following main plant components (see Fig. 22 B):

1) *System Module 1. Precombustion Chamber.* The reactor charging can occur according to the feed material, either via a double-gate air lock (e.g., for sod peat or molded products), or without a lock (e.g., by screw feeder for milled peat feed). Ash comminution, continuous ash discharge ignition burner, and devices for measurement of filling level, temperature, and pressure, and electrically controlled valves are included in the precombustion chamber.
2) *System Module 2. Labyrinth Flame Duct.* The flame duct is constructed from prefabricated ceramic segments with multizone insulation and a steel jacket as well as a starting burner for heating up the flame duct during first cold start of the system, controllable secondary air feed flaps; temperature measuring devices; etc.
3) *System Module 3. Waste Heat Boiler or Heat Exchanger.*
4) *Stack with Flue Gas Blower.*
5) *Control Panel.* The control panel contains units for the measurement and control technique required for automatic operation of the MSC plant which is equipped with continuous exhaust gas monitoring (e.g., for O_2, CO/CO_2, NO_x, SO_2, etc.).

Previous Operating Experience. Multistage combustion plants have previously been used successfully both as single-grate systems for mainly lumpy fuels and corresponding suitable or processed operating residues and as multigrate systems for milled peat or other low-calorific value fuels and suitable fine-grain or pourable operating residues.

Multistage combustion plants can be used for autothermal burning of milled peat and often very heterogeneous production residues such as fuel peat sods of high moisture contents (up to 45 wt % more than is usually the case after air drying) that cannot be charged to conventional combustion plants.

Multistage combustion can also be used for heat recovery from mixtures of peat products and other production residues, if these meet the following preconditions for an autothermic multistage combustion process:

Carbon content (as coke)	min. 6 wt % (relative to dry matter)
Lower calorific value	min. 9 – 20 MJ/kg
Water content	max. 35 – 45 wt %

Ash melting and compaction which often disturb the operating sequence is avoided in the multistage combustion process by precautionary measures with the apparatus. This can be achieved even when the deposits are formed by sintering which occurs in fluidized-bed gasification of peat [68].

The multistage combustion technology is particularly suitable for decentralized energy generation from peat even in medium-sized heating plants.

3.3.3.3. Peat Briquetting

The first successful attempts to briquette peat were made in Germany in the first half of the 19th century [86]. Peat briquette works require large surge bins in order to always guarantee a milled peat mixture with an approximately uniform average moisture content of ca. 52 wt % as starting material (bulk density ca. 250 kg/m^3) [87].

Extrusion presses [88], [89] operating without binders are used for the production of peat briquettes, and briquettes of a wide variety of shapes can be produced. Peat briquettes correspond in appearance and calorific value to the lignite briquettes and in countries such as Ireland and the former Soviet Union must serve essentially the same purpose as lignite briquettes.

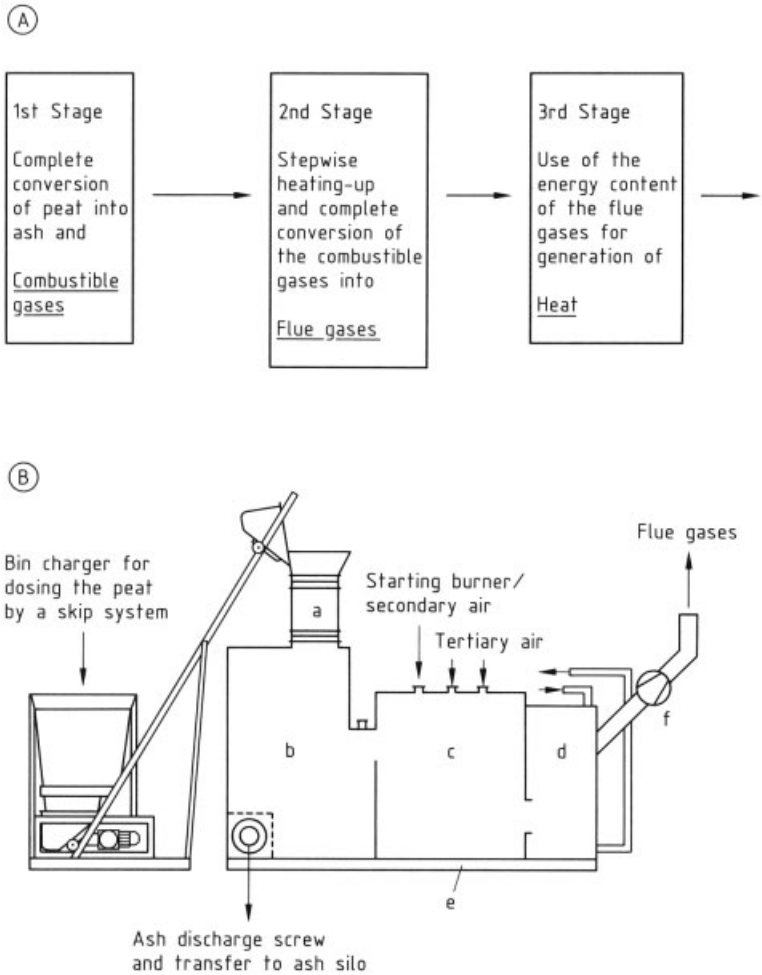

Figure 22. Multistage combustion system
A) Principles; B) Components
a) Air lock; b) Module 1, recombustion chamber; c) Module 2, labyrinth flame duct; d) Module 3, three-pass boiler (heat exchanger); e) Base frame; f) Flue gas chamber

The required high briquette strength necessitates good predrying of the peat. Figure 23 shows the typical construction of an Irish peat briquette works.

At present, peat briquettes are manufactured almost exclusively in Ireland and the former Soviet Union. In Ireland four briquette works are still in operation, with a total production of ca. 500 000 t/a. In the former Soviet Union there are still 120 briquette works, each with an annual production of 25 000 to 60 000 t.

3.3.4. Other Uses of Peat

Aside from its industrial and horticultural uses, peat is used as a filter medium and in medicine.

3.3.4.1. Peat as a Filter Medium

Because of their high specific surface area, white peats (peat mosses) are excellent filter media and adsorbents [90]. Dried fibrous white peats ($< 50-55$ wt % moisture) are preferred to mineral adsorbents because of their hydrophobic

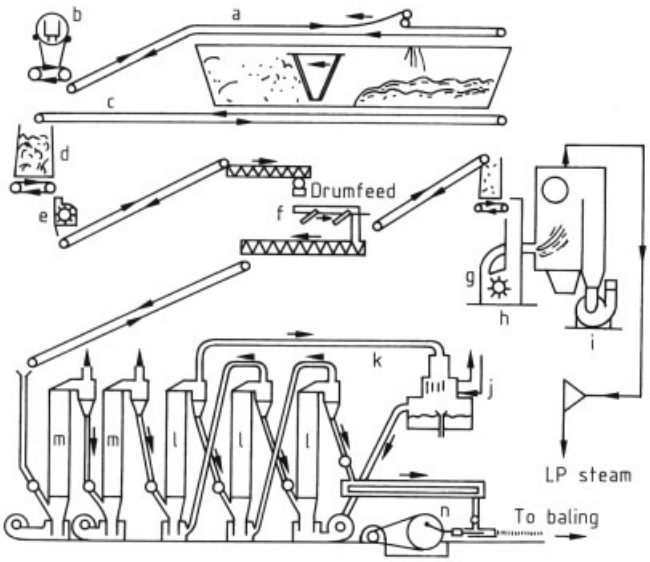

Figure 23. Typical equipment arrangement of an Irish peat briquette works [46]
a) Blending bunker; b) Rotary tippler, c) Magnet; d) Buffer bunker; e) Mill; f) Screen; g) Peat firing mill; h) Boiler; i) Turbo alternator; j) Heat exchanger; k) Scrubber; l) Steam dryer pipes; m) Hot water dryer pipes; n) Press

nature and easy disposal [91]. Tests in wastewater technology show that even nonthermally-processed peats in granular form are suitable for removing heavy metals from wastewaters. For this purpose, however, peat is inferior to activated carbons and other adsorbents (e.g., ion exchange resins) [92].

Peat (e.g., raised-bog peat) is an effective biological exhaust air filter. In particular, filters built up from coarse peat, fibrous peat, heather, and spruce brushwood, which are used for the purification of exhaust air from sewage treatment plants, compost works, animal carcass processing plants, etc., have good purification capacities [93–95].

3.3.4.2. Peat for Medical Purposes

More than 500 000 m^3 of fresh raw peat are used annually as mud baths and mudpacks in balneology in Germany [96], [97]. These mud baths and mudpacks are prescribed, e.g., for illnesses of the rheumatism group. Peat-based bath additives, pastes, and ointments are supplied for external application and drink cures.

4. Worldwide Economic Importance

Since the mid 1960s, white peat produced from raised bog for horticultural purposes has become increasingly important economically. At the same time, the demand for this raw material has shifted increasingly to countries with warmer climates (e.g., Israel and some African countries), where the soil is used ever more intensively for agriculture. Worldwide increases in the price and scarcity of primary energy sources have caused peat to become increasingly important as an energy supplier. This is especially true where no other fossil fuels are available.

In the European countries the peat production costs have developed on the basis of air-drying techniques in the peat bogs: expensive thermal drying can only be used in peat briquetting. For Ireland, for example, the following criteria influence the peat production costs:

1) the average thickness of an undrained bog must not be less than 5.5 m;
2) the average moisture content in the peat of undrained mires must not exceed 93.5 wt %;
3) the ash content must not exceed 2.4 wt % relative to dry matter;
4) for the recultivation, a remaining peat thickness of the bog of 0.5 m must be guaranteed;

5) the production areas must comprise at least 80 % of the deposit;
6) the production losses must not exceed 7.5 % of the available deposit content;
7) the development period of a bog must not exceed 4 to 5 years;
8) the calorific value (relative to dry matter) of a tonne of milled peat produced for power stations must not fall below 20 GJ;
9) the peat bog surface going into production must be free from wood and plants, so that powerful machines can be used;
10) the climatic conditions must be suitable for fully mechanized peat production;
11) the processing installations of the peat business must guarantee profitable production of a product range at constant quality, even if there is a great variety of products; and
12) the infrastructure of the traffic or road network must be suitable for a controlled flow of goods.

In developing countries, the following additional factors can be of importance:

1) Thickness (2 – > 30 m) and condition of the mires can differ considerably from those of European mires.
2) The moisture contents can be 85 – 95 wt %, leading to longer development times and very much higher costs.
3) The ash contents of subtropical and tropical mires are usually very much greater than of the raised bogs of the Northern Hemisphere.
4) The climatic and topographical conditions of peat production are often much less favorable.

4.1. Peat Industry and Availability of Raw Material Resources

The Netherlands and France, as large-scale users of peat for horticultural purposes, have stopped peat extraction for lack of suitable deposits or because of ecological pressure. A similar situation will arise in Germany in the next few decades. Of the formerly untouched ca. 2000 km^2 of raised bogs and fens in Germany, only a few areas have remained close to their natural state (ca. 5000 km^2 raised bogs and ca. 100 km^2 fens). Only 350 km^2 are still available as areas for peat extraction.

Altogether it is assumed that at an annual volume extraction of ca. 12×10^6 m^3, the white peat resources will theoretically last for a further ca. 15 years and the black peat resources for a further ca. 30 years.

In the former Soviet Union, Finland, and Ireland peat resources will last for more than 30 years because of the enormous peat reserves in these countries.

The peat consumption of other highly developed countries (e.g., United States and Canada) is very small and almost insignificant in relation to the great wealth of native peat resources.

From the developing countries as well as a few old peat countries (e.g., Poland and the Baltic states) no reliable data on peat deposits are yet available.

4.2. Possibilities for the Substitution of Peat and Peat Products

As white peat becomes scarcer, suitable peat additives or peat substitutes are required, e.g., from compost production. The substitutes must of course meet the quality standards set by the market [98]. But while valid quality criteria exist for peat products, these are still partly lacking for peat substitutes. This is true especially of the impurity limits, e.g., for heavy metals and other harmful substances. Thus, composts will probably not be used as cultivation substrates [8], [99]. For this reason new fields of application for composts, chiefly in soil improvement, must be developed.

5. Raw Material Production in Peat Bogs and Nature Conservation

Virgin peat bogs have almost disappeared from Central Europe as a result of settlement and intensive exploitation of the soil. The few peat bogs close to nature which still remain in developed economic expanding areas therefore require maintenance and conservation and exploited peat deposits (bogs) must be reintegrated into the landscape.

5.1. Peat Production and Nature Conservation

The current need for the conservation of mires must be justified not only by ecological but also by economic factors [100]. A distinction should always be made between conservation need and the merit of conservation, which must also be contrasted with the conservability of a bog [101].

The *need for conservation* of a bog results from scientific interests, ecological equilibrium, ethical motives, aesthetic standpoints, and economic effects. The bog's *conservability* can be hydrologically, chemically, and biologically based. The *merit of conservation* results from the naturalness, degree of threat, irreplaceability, diversity, rarity, perfection, and uniqueness of a bog.

The Netherlands were the first to develop conversation measures for peat bogs. These regulations formed the basis for an extensive legal framework on mire conservation that was introduced in Germany [102], [103]. The German Arbeitskreis Moornutzung und Landschaftspflege (Working Party for Mire Use and Countryside Care) has been occupied since 1977 with stocktaking in mires and establishing proposals on the questions of peat extraction and the protection and development of raised-bog areas [102–106].

5.2. Recultivation, Renaturation, and Regeneration

According to present knowledge, three important phases can be distinguished in the redevelopment of raised bogs, i.e.: rewaterlogging, renaturation, and regeneration [103].

Rewaterlogging serves as a short term measure. Rewaterlogging is achieved, e.g., by closing the existing drainage ditches of a previous peat extraction works and damming up the precipitation water. The remaining peat layer swells up in the course of 2 – 4 years after rewaterlogging.

Renaturation can be started after rewaterlogging is complete. Renaturation pursues as a priority the repression of untypical plant and animal species in degenerate raised bogs.

Regeneration, finally, corresponds to the successful restoration of conditions typical of raised bogs with regard to vegetation and the animal world. The resumption of peat formation can commence, as experiments in rewaterlogged areas have shown [107].

In Germany more than half of the areas currently used by the peat industry, i.e., ca. 180 km^2, will be renatured in the next few decades. This is a contribution to nature conservation which could then be an example to other peat-extracting countries in Europe. As many of the bogs then protected will be very small, it should be borne in mind at the regeneration stage that edge effects must be minimized by hydrological and immission protection zones as buffer zones.

Wholly different conditions prevail in the other most important peat-producing countries of Europe and North America. In Finland, Ireland, the former Soviet Union, Sweden, and Canada, increased efforts were even made to exploit the existing peat bogs more profitable to the individual national economies, chiefly for energy supply, in place of other imported raw materials.

But peat extraction and nature conservation can be complementary, as examples in Finland and Ireland confirm, if initially both interests are coordinated and respected in practice.

6. References

1. M. Teichmüller, *Geol. Jahrb.* **64** (1950) 429 – 488.
2. F. Overbeck: *Botanisch-geologische Moorkunde,* K. Wachholtz Verlag, Neumünster 1975, p. 719.
3. J. Koch, *Geol. Jahrb.* **89** (1981) 315 – 328.
4. K. Göttlich: *Moor- und Torfkunde,* 3rd. ed., E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1990, p. 529.
5. J. Günther, G. Schmilewski, H. Terkamp: *Der Rohstoff Torf,* 1st ed., Zentrale Informationsstelle Torf und Humus (ZIT), Hannover 1990, p. 16.
6. L. von Post, *Arsb. Sver. Geol. Unders.* **19** (1925) 4.
7. DIN 11 540, April 1989.
8. J. Günther, *Niedersaechs. Akad. Geowiss. Veröffentl.* **5** (1990) 66 – 73.
9. H. Grumpelt, *Telma (Hannover)* **17** (1987) 151 – 177.
10. J. Wendland, *Energie* **42** (1990) no. 10, 26 – 28.

11. B. N. Sokolov (ed.): *Peat in the National Economy,* Nedra Publishers, Moscow 1988, pp. 47 – 48.
12. *Annual Report of the Finish Industry 5,* Helsinki 1989.
13. F. Overbeck: "Moore," in "Das Känozoikum in Niedersachsen (Tertiär, Diluvium, Alluvium und Moore)," in K. Gripp, F. Drewers, F. Overbeck (eds.): *Geologie und Lagerstätten Niedersachsens,* vol. 3, Schriften der Wirtschaftswissenschaftlichen Gesellschaft zum Studium Niedersachsens, Bremen-Horn 1950, p. 122.
14. *Erdöl, Kohle, Erdgas, Petrochem.* **44** (1991) 54.
15. H. Falkenberg, *Niedersaechs. Akad. Geowiss. Veröffentl.* **5** (1990) 62 – 65.
16. P. Steffens, *Niedersaechs. Akad. Geowiss. Veröffentl.* **5** (1990) 23 – 29.
17. K. M. Schallinger (ed.), *Proc. Int. Symp. Peat Agric. Hortic. 2nd.,* Bet Dagan, Israel, 1983, p. 397.
18. G. W. Lüttig, *Gb+Gw* **51** (1988) 2255 – 2257.
19. A. A. Archer, G. W. Lüttich, I. I. Snezhko: *Man's Dependence on the Earth,* UNESCO-Schweizerbart, Paris – Stuttgart 1985.
20. S. Schneider: "Der Rohstoff Torf, seine Gewinnung und Verwendung," Lecture held at the Staatliches Torfinstitut Hannover, Hannover 1956, pp.65 – 84.
21. C. H. Fuchsman: *Peat, Industrial Chemistry and Technology,* Academic Press, New York 1980, p. 279.
22. *Ullmann,* 3rd ed., **17,** 597 – 618.
23. *Ullmann,* 4th ed., **23,** 333 – 356.
24. G. Kampf, *Telma (Hannover)* **3** (1973) 27 – 38.
25. K.-H. Richard: *Rückblick auf die Entwicklung der maschinellen Brenntorfsoden-Gewinnung,* Moormuseum Groß-Hesepe, Groß-Hesepe 1990, p. 78.
26. W. Ehlert: *Die Entwicklung der Torfindustrie in Niedersachsen,* Torfforschung GmbH, Bad Zwischenahn 1978, p. 62.
27. H. Grumpelt, *Unser Betrieb* 1976, no. 3, 10 – 16; (1976), no. 4, 10 – 12.
28. H. Grumpelt, *Unser Betrieb* 1975, no. 4, 14 – 19; (1976), no. 1, 20 – 25.
29. *Proc. VIII. Int. Peat Congr.,* vol. I – IV, Leningrad, Aug. 14 – 21, (1988).
30. Turveruukki Oy, *Energy in Finland* (1987/1988) 40 – 41.
31. G. Eichhorn, *Niedersaechs. Akad. Geowiss. Ver-öffentl.* **5** (1990) 45 – 48.
32. E. Pollak, *Niedersaechs. Akad. Geowiss. Veröffentl.* **5** (1990) 42 – 44.
33. H. Grumpelt, *Unser Betrieb* 1983, no. 2, 3 – 7.
34. P. J. Belkevich, J. I. Lishtavan, G. P. Viryasov, G. B. Naumovea: *World Conference on Future Sources of Organic Raw Materials,* Toronto, July 10 – 13, (1978), pp. 315 – 323.
35. C. H. Fuchsman, *Telma (Hannover)* **9** (1979) 75 – 85.
36. S. W. Souci, G. Sommer, *Mitt. f. Landkultur, Moor- und Torfwirtsch.* **7** (1959) 113.
37. H. H. Oehlert, N. N. Son, T. Dhinh-Van, *Chem. Ing. Tech.* **49** (1977) no. 11, 913.
38. K. Sipila et al., *Turpeen Kaasutuste-Kniikan Kelu Haminen – Oulun turveammoniakkitehdas,* Paper D, Department of Commerce, Helsinki 1989, p. 111.
39. H. Grumpelt, *Telma (Hannover)* **9** (1979) 87 – 108.
40. H. Teggers, J. Lambertz, U. Keller, *VGB Kraftwerkstech.* **68** (1988) no. 11.
41. H. El-Saied, U. H. H. Oehlert, *Cellul. Chem. Techn.* **14** (1980) 507 – 516.
42. K. Sipilä, *Energy in Finland* (1986) 28 – 30.
43. E. W. Scholl, Bulletin of the Torfforschung GmbH **144** (1978) 7 – 21.
44. J. Ewers, H. Mertikat, W. Günster, J. Keller, *Brennst. Waerme Kraft.* **41** (1989) 23 – 31.
45. H. Weckman, Kotimaisten polttoaineiden turvallinen tuotanto ja Kayttö. ösa 4: Turpeen palo-ja rajähdysominaisuudet, Report of the Finnish Technical Research Centre, Report No. 448, Espoo 1986.
46. J. Martin: "Briquetting of Peat Fuel," *Proc. Inst. Briquet. Agglom. Bienn. Conf. 14th,* **14** (1975) 153 – 171.
47. K. Pihlaja, *Suo* **40** (1989) no. 2/3, 125 – 128.
48. H. Grumpelt: *Methangas aus Mooren – Eine kritische Stellungnahme zum Potential der Moore,* Lecture held at the Zwischenahner Torftage, Bad Zwischenahn, Nov. 13 – 14, (1986), pp. 167 – 197.
49. H. Grumpelt, *Telma (Hannover)* **18** (1988) 227 – 250.
50. M. Dambroth, *Nachr. Chem. Tech. Lab.* **29** (1981) no. 1, 12 – 17.
51. Frankfurter Allgemeine Zeitung, 07. Aug. 1985, Frankfurt am Main.
52. R. Martinell, D. Tannenberg, *Proc. Peat Environment '85,* International Peat Society Symposium 17. – 20. Sept. 1985, Jönköping (1985), pp. 287 – 295.
53. H. R. Appell, I. Wendler, R. D. Miller: "Solubilisation of Low Rank Coal with Carbon Monoxide and Water," *Chem. Ind.* (1969), 1703.

54. M. Rutishauser: *Bioenergie,* Gottlieb Duttweiler Institut, Rüschlikon (1978), p. 260.
55. E. Chornet et al., *Cdn. J. Chem. Eng.* **58** (1980) 713–717.
56. B. Björnbom, L. Granath et al., *Fuel* **60** (1980) 7–13.
57. G. R. Youngquist, R. Partch, L. G. Donaruma, *Des. Manage. Resour. Recovery* **1** (1980) 159–180.
58. U. Schuchardt, F. d. A. P. Matos, *Fuel* **61** (1982) 106–110.
59. H. Grumpelt: "Present State and Possibilities of an Economical and Environmentally Beneficial Peat Coke Production," in *A Technical and Scientific Conference on Peat and Peatlands,* Rivière-du-Loup, 16–20. June 1985, pp. 154–173.
60. J. Jüntgen, K. H. van Heek, *Fuel Process. Technol.* **2** (1979) 261.
61. K. G. Beck, *Glückauf* **117** (1981) no. 10, 568–577.
62. P. Rudolf, P. Schmalfeld, D. Sauter: "Gasification and Carbonization of Peat," in *Reports of the International Symposium of DGMT and Section II of IPS,* pp. 149–161; *Recent Technologies in the Use of Peat,* Bad Zwischenahn, Nov. 5–8, 1979, E. Schweizerbart'sche Verlagsbuchhandlung (Nägele & Obermiller), Stuttgart 1983.
63. Voest-Alpine Industrieanlagenbau, *Coal Drying Plants – Fleißner System,* Linz 1986.
64. P. Herber, J. Fohl, *Braunkohle (Düsseldorf)* **42** (1990) no. 5, 16–21.
65. J. W. Rohrer, B. Myreen, *PDT Large-Scale Industrial Peat Refining,* Jaako Pöyry Energie OY, Finland, 1980, p. 13.
66. B. Myreen, US 4 200 600, 1980.
67. J. Fohl, W. Lugscheider, *österreichische Chemie-Zeitschrift* 1986 no. 12, 350.
68. J. Koljonen, E. Kurkela: "Experiences in the Operation of the HTW Process at the Peat Ammonia Plant of Kemira OY," *Symposium 107,* vol. 1, Finnish Technical Research Centre, Espoo 1990, pp. 361–372.
69. F. Lürmann, *Stahl Eisen* **2** (1982) 17.
70. H. Jüntgen, K. H. van Heek: *Kohlevergasung,* Karl Thieme Verlag, Stuttgart 1981, p. 185.
71. Solar Energy Research Institute (SERI): *A Survey of Biomass Gasification,* vol. I–III, SERI TR-33-239, US Department of Energy Contract, No. EG-77-C-01-4042, Golden, Colo., 1979.
72. D. V. Punivani, R. Biljetina: *Peatgas Process Development Status,* Paper presented at IGT Conference, Arlington, Va., 1990.
73. J. Keller: *Kohlevergasung als Basis für die Stromerzeugung: Ein umweltfreundliches Anlagenkonzept,* Lecture held at ACHEMA, Frankfurt am Main, June 9, 1988.
74. H. Grumpelt, *Unser Betrieb* (1991), no. 1, 3–9.
75. H. Grumpelt, *Unser Betrieb* (1979), no. 4, 10–15; (1980), no. 1, 6–12.
76. H. Grumpelt, *Telma (Hannover)* **15** (1985) 269–271.
77. H. Grumpelt, *Unser Betrieb* (1986), no. 3, 3–6;(1987), no. 1, 18–21.
78. K. Mutka, *Energy in Finland,* Dec. 1985/Jan. 1986, 28–30.
79. Montreal Engineering Company: *The Mining of Peat – A Canadian Energy Resource,* EMR-CANMENT Contract 7 – 9047, 1979.
80. M. Schärfke, *Braunkohle (Düsseldorf)* **42** (1990) no. 7, 35–37.
81. H. Zabinski, A. Lange, H. Böse, *Braunkohle (Düsseldorf)* **42** (1990) no. 7, 26–30.
82. Orgreb-Institut für Kraftwerke: *Bereich Dresden: Bericht zur Inbetriebnahme und des Probebetriebes des Prototyp-DWT-Anlage Borna für den Zeitraum 06.10.89 – 31.03.90,* Dresden April 30, 1990.
83. K. H. Heek, *Glückauf* **126** (1990) no. 7/8, 357–362.
84. E. Ekman, D. Aplund: "Pilot Plant Equipment for Research, Experiences for Thermal Processing of Peat at the Technical Research Centre of Finland (VTT)," in *Reports on the International Symposium of DGMT and Section II of IPS,* Bad Zwischenahn, Nov. 5–8, 1979, E. Schweizerbart'sche Verlagsbuchhandlung (Nägele u. Obermiller), Stuttgart 1983, pp. 93–103.
85. H. Grumpelt et al., DE 3 918 508, 1990.
86. H. Kühn: "110 Jahre Braunkohlenbrikettierung," *Braunkohle* **21** (1969) no. 3, 80–84.
87. *Ullmann* 3rd ed., **17,** 612–613.
88. R. Veber: *Peat Briquetting,* State Energy Publishing House, Moscow 1957.
89. E. Scherrer, *Braunkohle (Düsseldorf)* **32** (1980) no. 11, 362–368.
90. R. N. Soniassy, *CIM Bull.* **913** (1974) 95–98.
91. T. Viraragharan, G. N. Mathavan, *Suo* **40** (1989) no. 5, 163–168.
92. Engler-Bunte-Institut der Universität Karlsruhe: 7. Vortragsreihe mit Erfahrungsaustausch über spezielle Fragen der Wassertechnologie (Adsorption), 1975 no. 9, 394.

93. H. D. Zeisig, A. Holzer, J. Kreitemeier: *Anwendung von biologischen Filtern zur Reduzierung von geruchbildenden Emissionen aus Tierkörperverwertungsanstalten,* Papers of the Landtechnik Weihenstephan 1981.
94. W. Koch, H. G. Liebe, *Fleischmehlindustrie* **8** (1982) 165 – 176.
95. Bulletin of the International Peat Society (IPS) **21** (1990) 24 – 31.
96. H. Schneekloth: *Die Moorbäder in Niedersachsen,* Kommissionsverlag Göttinger Tageblatt, Göttingen 1985, p. 68.
97. G. W. Lüttich, *Telma (Hannover)* **19** (1989) 157 –164.
98. H. C. Deilmann, *Niedersaechs. Akad. Geowiss. Veröffentl.* **5** (1990) 59 – 61.
99. J. Günther: "Blickpunkt: Industriesubstrate und Umwelt," Reprint, *ZVG Gartenbaureport* **9** (1990) 4.
100. K. Wieborg: "Zum Verbrennen zu schade," in *Moor im Emsland,* 1st ed., Emsländischer Heimatbund, Sögel 1979, pp. 60 – 81, 92.
101. H. Kuntze, R. Eggelsmann, *Telma (Hannover)* **11** (1981) 197 – 212.
102. Bundesnaturschutzgesetz: BNatSchG from Dec. 20, 1976, Bundesgesetzblatt, Part 1, no. 147, pp. 3573 –3582.
103. Niedersächsisches Landesverwaltungsamt: Fachbehörde Naturschutz, Informationsdienst Naturschutz Niedersachsen (3), Hannover (1990), p. 56.
104. B. Birkholt et al.: Untersuchungen an den niedersächsischen Torflagerstätten zur Beurteilung der abbauwürdigen Torfvorräte und der Schutzwürdigkeit im Hinblick auf deren optimale Nutzung, Naturschutz und Landschaftspflege in Niedersachsen (12), Hannover 1980, p. 402.
105. Niedersächsisches Ministerium für Ernährung, Landwirtschaft und Forsten: Niedersächsisches Moorschutzprogramm – Teil 1: Programm der Niedersächsischen Landesregierung zum Schutze der für den Naturschutz wertvollen Hochmoore mit näheren Festlegungen für rund drei Viertel der noch vorhandenen geologischen Hochmoorfläche in Niedersachsen, Dec. 1, 1981.
106. Niedersächsisches Ministerium für Ernährung, Landwirtschaft und Forsten: Niedersächsisches Moorschutzprogramm – Teil 2: Programm der Niedersächsischen Landesregierung zum Schutz der für den Naturschutz wertvollen Hochmoore und Kleinsthochmoore, Jan. 14, 1986.
107. "Herrichtung von Hochmoorflächen nach Torfabbau," in *Nds. MBL* no. 19, (1988), p. 521.

Pectin → Polysaccharides
Pencils → Drawing and Writing Materials
Penicillin → Antibiotics
2,4-Pentanedione → Ketones
Pentanes → Hydrocarbons

Pentanols

PETER LAPPE, Hoechst Aktiengesellschaft, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany (Chaps. 1 – 7)

THOMAS HOFMANN, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 8)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 1
4. Production 4
5. Uses 6
6. Quality Specifications and Chemical Analysis 7
7. Storage and Transportation 8
8. Toxicology and Occupational Health . . 8
9. References 9

1. Introduction

Pentanols (amyl alcohols), $C_5H_{11}OH$, M_r 88.15, are saturated aliphatic C_5-alcohols. The eight structural pentanol isomers are often named by common names in addition to the systematic IUPAC nomenclature (see Table 1). Four of the pentanol isomers are primary alcohols, three are secondary alcohols, and 2-methyl-2-butanol is a tertiary alcohol.

Pentanols are found mainly in fruit, honey, tea, cheese, and alcoholic beverages. They were first discovered in fusel oils and were side products in the fermentation of starch-containing materials.

Three of the C_5-alcohol isomers–2-pentanol, 2-methyl-1-butanol, and 3-methyl-2-butanol–are optically active.

2. Physical Properties

The most important physical properties of the eight pentanol isomers are listed in Table 2. The pentanols are colorless liquids of characteristic odor, with the exception of 2,2-dimethyl-1-propanol, which is solid at room temperature. The pentanols are only sparingly miscible with water at 20 °C. However, they are almost completely miscible with most organic solvents. Binary liquid – liquid equilibrium constants for pentanols in water at 0 – 90 °C are given in [1].

2-Pentanol, 2-methyl-1-butanol, and 3-methyl-2-butanol can be separated into their optically active isomers by suitable separation processes. Some specific rotations are compiled in Table 3. Many binary and tertiary azeotropic mixtures of various pentanol isomers with other liquids are known. The boiling points and compositions of binary systems with water are listed in Table 4.

3. Chemical Properties

The chemical behavior of pentanols is determined primarily by the hydroxyl group. The most important reactions are dehydration, oxidation, and esterification.

Dehydration. Catalytic dehydration of alcohols for the production of olefins is a well-established reaction [4]. Isomeric mixtures of the C_5-olefins usually are obtained during dehydration of pentanols. *2-Pentanol* yields mainly 2-pentene [5] with $ZnO - Al_2O_3$ as catalyst at 270 – 370 °C; 1-pentene is produced through the use of thorium oxide-containing catalysts [6]. Many other studies of the influence of transition-metal oxides on the mechanism and selectivity of the catalytic dehydration of 2-pentanol have been carried out by B. H. DAVIS [7].

2-Methyl-1-butene is obtained from *2-methyl-1-butanol* on special aluminum oxide contacts with a selectivity of 86 – 90 % [8].

Dehydration of *3-methyl-1-butanol* yields mainly 3-methyl-1-butene [9] or 2-methyl-2-butene [10], depending on reaction conditions. A mixture of these two olefins is produced by the dehydration of 2-methyl-2-butanol [11].

Production of dipentyl ethers during the dehydration of pentanols is a commonly observed

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_049

Table 1. Pentanol isomers

Formula	Systematic name	Common name	CAS registry no.
$CH_3CH_2CH_2CH_2CH_2OH$	1-pentanol	*n*-amyl alcohol	[*71-41-0*]
$CH_3CH_2CH_2CH(OH)CH_3$	2-pentanol	*sec*-amyl alcohol	[*6032-29-7*]
$CH_3CH_2CH(OH)CH_2CH_3$	3-pentanol		[*584-02-1*]
$CH_3CH_2CH(CH_3)CH_2OH$	2-methyl-1-butanol	active primary amyl alcohol	[*137-32-6*]
$CH_3CH(CH_3)CH_2CH_2OH$	3-methyl-1-butanol	isoamyl alcohol	[*123-51-3*]
$CH_3CH_2C(CH_3)(OH)CH_3$	2-methyl-2-butanol	*tert*-amyl alcohol	[*75-85-4*]
$CH_3CH(CH_3)CH(OH)CH_3$	3-methyl-2-butanol	*sec*-isoamyl alcohol	[*598-75-4*]
$CH_3C(CH_3)_2CH_2OH$	2,2-dimethyl-1-propanol	neopentyl alcohol	[*75-84-3*]

Table 2. Physical properties of pentanols

	1-Pentanol	2-Pentanol	3-Pentanol	2-Methyl-1-butanol	3-Methyl-1-butanol	2-Methyl-2-butanol	3-Methyl-2-butanol	2,2-Dimethyl-1-propanol
mp, °C	−78.5	<−60	<−75	<−70	−117.2	−11.9		53
bp, °C	137.8	119.3	115.6	128.0	130.5	101.8	112.0	113–114
Density ϱ, g/cm^3	0.8144	0.8090	0.8218	0.816	0.812	0.809	0.819	
Refractive index n_D^{20}	1.4099	1.4053	1.4098	1.4097	1.4078	1.4052	1.4095	
Flash point, °C	50	33	40	43	44	21	26	36
Ignition temperature, °C	310	340	360	400	365	435		430
Viscosity, mPa · s	3.31 (25 °C)	0.53 (20 °C)	4.12 (25 °C)	5.09 (20 °C)	3.86 (23.8 °C)	3.79 (25 °C)		
Specific heat at 20 °C, Jg^{-1} K^{-1}	2.98	3.40	2.85	2.50	2.28	3.15	2.63	2.69
Heat of vaporization, J/g	505		405	419	441	445		
Vapor pressure at 20 °C, kPa	0.14	0.53	0.93	0.40	0.27	1.60	1.07	
Solubility in H_2O at 20 °C, wt %	2.6	4.7		2.2	2.5			3.0
Expansion coefficient	0.00092	0.00095	0.00149	0.00078	0.00090	0.00133		

side reaction. Dipentyl ethers are normally obtained from pentanols under the influence of catalysts such as synthetic bentonites [12], modified montmorillonites [13], or solid superacids (e.g., Nafion-H) [14].

Dehydration of 2,2- dimethyl-1-propanol in the presence of nickel- or palladium- containing catalysts yields the corresponding dipentyl ether [15]. Syntheses of other dipentyl ethers are described in [16–18].

Oxidation. Primary and secondary pentanols are transformed to aldehydes, carboxylic acids or ketones under usual conditions; catalytic oxidation dehydrogenation is preferred industrially. Thus, 2-methylbutanal is obtained from 2-methyl-1-butanol by

using MnO – NiO – MgO as catalyst [19]. Other suitable oxidation dehydrogenation catalysts for the synthesis of C_5-aldehydes from pentanols include lead(II) acetate [20], MgO – B, Al, Si, P, Sb [21], and copper(I) chloride – phenanthroline [22]. Further suitable oxidants are osmium tetroxide [23], dimethylsulf-oxide – trifluoroacetic anhydride [24], CrO_2Cl_2 [25], and $(Ph_3BiCl)_2O$ [26].

Table 3. Specific rotations $[\alpha]_D$ of optically active pentanols [2]

Compound	Specific rotation	
(*R*)-2-Methyl-1-butanol	+ 3.1°	(20 °C)
(*S*)-2-Methyl-1-butanol	− 5.88°	(18 °C)
(*S*)-3-Methyl-2-butanol	+ 5.34°	(20 °C)
(*R*)-2-Pentanol	−16.8°	(20 °C)
(*S*)-2-Pentanol	+16.6°	(20 °C)

Recommended catalysts for the conversion of secondary pentanols to the corresponding ketones include modified bentonites [27], $(Ph_3P)_2Ru(O_2CCF_3)CO$ [28], $ZnO – Cr_2O_3 – CuO – CdO$ [29], and phosphine-containing ruthenium(IV) – oxobipyridine complexes [30]. Other suitable oxidants are ozone [31], potassium permanganate [32], silver peroxy disulfate [33], nitrogen dioxide [34], and *tert*-butyl hypochlorite – pyridine [35].

The C_5- carboxylic acids can be obtained from primary pentanols under appropriate reaction conditions. 2,2-Dimethylpropionic acid (pivalic acid) and 2-methylbutanoic acid are obtained by oxidizing the corresponding alcohols with $KMnO_4 – CuSO_4$ in the presence of potassium hydroxide [36]. Other suitable oxidizing agents are ozone [37] and nickel oxide – hydroxide [38].

Esterification. Pentanols can react with carboxylic acids or carboxylic anhydrides to yield esters by various methods. Examples are continuous reaction of 2-pentanol with acrylic acid under the catalytic action of acid ion exchangers in the liquid phase [39]; azeotropic esterification of formic acid with 1-pentanol with diisopropyl ether as entraining agent [40]; and reaction of bicyclo[2.2.0]oct-2-en-5,6-dicarboxylic anhydride with 1-pentanol [41].

Another frequently used method for the production of esters is transesterification. Methyl acetate reacts with 1-pentanol in the presence of catalysts such as the oxides of aluminum, titanium, tin, or zirconium at 20 °C to give pentyl acetate in high yield [42]. Diesters such as methyl succinate react with 2-methyl-2-butanol in the presence of lithium to give 2-methylbutyl succinate [43].

The conversion of pentanols to pentyl valerates (part of the pentanol is oxidized to the corresponding aldehyde which reacts with unconverted pentanol via the hemiacetal to the valerate) is catalyzed particularly by ruthenium-containing compounds. Thus, 2-methylbutyl 2-methylbutyrate is obtained from 2-methyl-1-butanol in the presence of $RuH_2(PPh_3)_4$ [44]; 2,2-dimethylpropyl 2,2-dimethylpropionate is formed from 2,2-dimethyl-1-propanol by using $Ru_3(CO)_{12}$ as catalyst [45]; and pentyl valerate is obtained from 1-pentanol and calcium hypochlorite with a yield of 83 % [46].

Pentanols can react with isocyanates and aminonitriles in a similar manner. Phenylisocyanate reacts with 2,2-dimethyl-1-propanol in the presence of lithium alkoxides or di-*n*-butyltin diacetate to yield the corresponding *N*-phenylcarbamate [47]. 3-Methylbutyl 2-aminopropionate is obtained by reaction of 2-aminopropionitrile with 3-methyl-1-butanol [48].

Other Reactions. Pentanols are often used as building blocks in many chemical syntheses. For example, 2-pentanol reacts with methyl acrylate to yield 5-methyl-5-propyltetrahydrofuran-2-one, and 1-pentanol reacts with methyl allylacetate (methyl 4-pentenoate) to give 6-pentyltetrahydropyran-2- one [49]. *N*-Heterocyclic compounds can be produced by reacting pentanols with iminodiacetonitrile [50] or acetic anhydride and ammonia [51]. Substituted thiophenes can be synthesized by gas-phase reaction of 2-pentanol with carbon disulfide under the catalytic action of $Cr_2O_3 – Al_2O_3 – K_2CO_3$ [52].

Cyclization of pentanols such as 2-methyl-2-butanol and 2-methyl-1-butanol in the presence of $V_2O_5 – MoO_3$ at 360 – 410 °C yields citraconic anhydride (3-methyl-2,5-furandione) [53].

Other industrially important reaction products of C_5-alcohols include pinacolone produced from 2-methyl-2-butanol [54]; neopentylamine from 2,2-dimethyl-1-propanol [55]; xanthates from 1-pentanol and other C_5-alcohol

Table 4. Binary azeotropic mixtures with water [3]

Compound	*bp* of compound, °C	*bp* of azeotropic mixture, °C	Water in azeotrope, wt %
1-Pentanol	137.8	95.8	54.4
2-Pentanol	119.3	91.7	36.5
3-Pentanol	115.4	91.7	36.0
3-Methyl-1-butanol	132.05	95.15	49.6
3-Methyl-2-butanol	112.9	91.0	33.0
2-Methyl-2-butanol	102.85	87.35	27.5

isomers (especially 2-methyl-1-butanol), carbon disulfide, and sodium hydroxide [56]; and dipentyl carbonate from 1-pentanol and carbon dioxide [57].

Pentanols with a methyl group in the 2-position as well as the corresponding C_{10}-alcohols can be produced from pentanols under the catalytic action of nickel, palladium, or rhodium and a sodium alkoxide base (Guerbet reaction) [58]. Various ethers and glycols can be obtained from pentanols and epichlorohydrin [59] and from pentanols and formaldehyde [60]. Finally C_5-alcohols can be transformed to pentyl formates by rhodium-catalyzed formylation reaction [61] and to dipentyl tricyclo[5.2.1.0^{2,6}]dec-3-ene-8,9-dicarboxylate by reaction with dicyclopentadiene and carbon monoxide in the presence of $PdCl_2 - CuCl_2$ catalyst [62].

4. Production

Only a few industrial processes are known for the production of pentanols, namely, hydroformylation of butenes, hydrolysis of chloropentanes, and hydration of pentenes. Production of C_5-alcohols (mainly 3-methyl-1-butanol and 2-methyl-1-butanol) from fusel oils by fermentation has been replaced mainly by the oxo synthesis (discovered by O. Roelen) and currently has little industrial importance. Production of pentanols from pentane by the Sharples process (chlorination of pentanes and subsequent hydrolysis) has also been largely replaced by oxo synthesis.

Hydroformylation of butenes (1-butene, 2-butene, and isobutene) and subsequent hydrogenation yield 1-pentanol, 2-methyl-1-butanol, and 3-methyl-1-butanol almost exclusively. The remaining C_5-alcohol isomers are obtained chiefly by hydration of C_5-olefins, hydrogenation of corresponding carbonyl compounds, and other special techniques.

Hydroformylation of Butenes. The most important process for the production of pentanols is hydroformylation of butenes. The C_4-olefins are reacted with a carbon monoxide – hydrogen mixture under the catalytic action of cobalt- or rhodium-containing compounds to yield C_5-aldehydes. Hydrogenation of the latter gives the corresponding C_5-alcohols. The flow sheet for this process is given in Figure 1.

Raffinate I from various cracking processes (after the separation of 1,3-butadiene) is used as raw material and contains ca. 44 – 49 vol % isobutene, 24 – 28 vol % 1-butene, 19 – 21 vol % 2-butene, and 8 – 11 vol % C_4-hydrocarbons [64]. Further processing of this C_4-raffinate converts isobutene by reaction with methanol to methyl *tert*-butyl ether (→ Methyl Tert-Butyl Ether, Chap. 4.). The remaining C_4-olefin mixture (raffinate II) is the most important raw material in the industrial production of 1-pentanol and 2-methyl-1-butanol (→ Butenes, Chap. 5.).

Cobalt-catalyzed hydroformylation of *n-butenes* and subsequent hydrogenation yield a mixture containing ca. 70 % 1-pentanol and 30 % 2-methyl-1-butanol [65]. If rhodium – triphenylphosphane is used as catalyst, a mixture containing up to 90 % 1-pentanol and 10 % 2-methyl-1-butanol is obtained.

2-Butene can be converted to mixtures of C_5-aldehydes of varying composition, depending on the rhodium-containing catalyst system and reaction conditions. Thus, a ratio of 1.13:1 of 1-pentanal to 2-methylbutanal is obtained with a rhodium – phosphite catalyst at 130 °C [66]; a ratio of 0.2:1 is obtained by hydroformylation

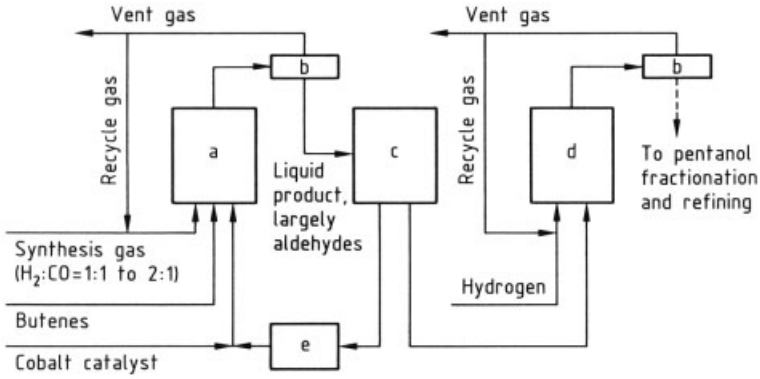

Figure 1. Flow diagram for the production of pentanols [63]a) Oxo reactor (100 – 200 °C, 10 – 45 MPa); b) Separator; c) Catalyst removal (decobalting); d) Hydrogenation unit (50 – 250 °C, 0.3 – 24 MPa); e) Catalyst recovery

with a rhodium – tricyclohexylphosphane catalyst at 110 °C [67]. The dependence of the *n*:*i* ratio on catalyst species and temperature during conversion of 1-butene and 2-butene with a rhodium –triarylphosphane catalyst system is given in [68].

Hydroformylation of *isobutene* with rhodium or cobalt catalysts yields mainly 3-methyl-1-butanol [69]; small amounts of 2,2-dimethyl-1-propanal occur chiefly in the cobalt- catalyzed oxo synthesis [65].

Apart from raffinate II and pure olefins, other C_4- olefin mixtures can be used as raw materials for hydroformylation. Thus, hydroformylation of an olefin mixture containing 1-butene, 2-butene, and isobutene (ratio 1:1:2) with $Co(CO)_4$ as catalyst at 100 °C yields a mixture of 1-pentanol, 2-methyl-1-butanol, and 3-methyl-1-butanol in a ratio of 15:5.2:11 [70]. More detailed information on the hydroformylation of butenes is given in [65].

Catalysts based on CuO – ZnO [71], nickel [72], zirconia [73], and Cu – TiO_2 [74] are used industrially for the subsequent hydrogenation of C_5-aldehydes.

Hydration of Pentenes. 2-Pentanol, 3-pentanol, 2-methyl-2-butanol, and 3-methyl-2-butanol are produced mainly by hydration of pentenes (see Table 5 for details). Apart from classical hydration processes, many modifications of the hydroboration reaction are described in the literature, which are applied especially to stereoselective reactions [83]. The reaction of the monobromoborane – dimethyl sulfide complex with 2-methyl-2-butene yields 3-methyl-2-butanol with high selectivity [84]. 2-Pentene reacts with dimesitylborane at 65 °C to give a mixture containing 76 % 2-pentanol and 24 % 3-pentanol [85]. Several other hydroboration reactions are described in [86].

Other Production Processes. An important process for production of the three secondary pentanols–2-pentanol, 3-pentanol, and 3-methyl-2-butanol – is hydrogenation of the corresponding ketones. Catalysts for these reactions include not only commercially available nickel, copper, and palladium catalysts [87] but also a range of specialized catalyst systems such as cationic rhodium complexes [88], ruthenium – polystyrene [89], and $ClMgAlH_4$ [90].

Various acid-stable catalysts were developed particularly for the production of 2,2-dimethyl-1-propanol from pivalic acid. Pivalic acid is converted to 2,2- dimethyl-1-propanol with a yield of 94.3 % in the presence of rhenium(VII) oxide (Re_2O_7) or osmium tetroxide (OsO_4) in dioxane at 90 °C and 100 bar [91]. Other active catalyst systems include CuO – ZnO – Al_2O_3 [92], zirconia [93], and $HTcO_4$ or $KReO_4$ – $HClO_4$ [94].

In addition to pivalic acid its esters react with complex hydrides [95] or Raney Ni – Sn [96]; pivalic anhydride, (reaction with organoaluminum hydrides) [97] and pivalonitrile [in the presence of $ZrO(OH)_2$] can give C_5-alcohols in high yield [98].

Table 5. Hydration of pentenes

Olefin	Alcohol	Process catalyst	Reference
1-Pentene	2-pentanol	zeolites, aluminosilicates	[75]
2-Methyl-2-butene	2-methyl-2-butanol	acid ion exchangers	[76]
2-Methyl-1-butene/2-Methyl-2-butene	2-methyl-2-butanol	acid ion exchangers	[77]
2-Methyl-2-butene	2-methyl-2-butanol	zeolites	[78]
2-Methyl-2-butene	2-methyl-2-butanol	layer silicates	[79]
2-Methyl-2-butene	2-methyl-2-butanol	formic acid	[80]
2-Methyl-2-butene/2-Methyl-1-butene	2-methyl-2-butanol	HCl, $SbCl_3$, $ZnCl_2$	[81]
2-Methyl-2-butene/2-Methyl-1-butene	2-methyl-2-butanol	acetic acid	[82]

Use of carbon monoxide and hydrogen to produce higher alcohols was investigated intensely in the 1980s. However, despite the development of various catalyst systems [99], this synthesis yields a mixture of many different alcohols and hence has not been used for industrial production of pentanols. The latter also applies to the synthesis of pentanols from the homologization (i.e., chain elongation by one or more methylene groups) of lower alcohols such as methanol and ethanol [100].

An interesting process for the production of 3-methyl-1-butanol is based on the reaction of isobutene with formaldehyde (Prins reaction). The intermediates formed– 3-methyl-3-buten-1-ol or 4,4-dimethyl-1,3- dioxane (depending on conditions) – can be hydrogenated to 3-methyl-1-butanol by a catalyst system such as $L_2Cu(OAc)_2/Cr_2O_3 - ZnO - BaO$ (L=quinoline) [101]. Other suitable catalysts for this reaction are given in [102]. This process is used by several companies for the production of isoprene. Cleavage of the dioxane derivative occurs at 240–400 °C on carbon or H_3PO_4–$Ca_3(PO_4)_2$ catalyst [64].

3-Methyl-2-butanol not only is formed by hydration of 2-methyl-2-butene, but also occurs as a byproduct during the production of isoprene via acetylene – acetone [103].

Finally, different methods for production of C_5-alcohols by oxidation of C_5-hydrocarbons exist [104]: oxidation of isopentane with Co – Mo stearate as catalyst at 110 – 140 °C yields 2-methyl-2-butanol [105].

5. Uses

The range of applications of 1-pentanol, 2-methyl-1-butanol, and 3-methyl-1-butanol (obtainable by the oxo synthesis) are discussed separately because of their individual significance.

1-Pentanol. 1-Pentanol is particularly useful as a solvent, extracting agent [106], and starting material for lubricant additives (zinc dipentyl dithiophosphates) [107] and for auxiliaries in flotation (xanthates). It is used as solvent in the reaction of substituted diamines with diisocyanates [108] and for the production of sulfurized olefins [109], polyacrylates [110], and pyridine-2,3-dicarboxylic acid derivatives [111]. 1-Pentanol is employed as an extracting agent in the purification of phosphoric acid [112] and the separation of strontium chloride from aqueous metal chloride solutions [113].

Other industrially important application of 1-pentanol are the synthesis of starting materials for the production of liquid crystals [114], dyes [115], flavorings [116], and special catalysts [117].

1-Pentanol is used as starting material for several syntheses in the pharmaceutical industry. Examples include the production of isoindolinone derivatives [118], 2-alkylbenzimidazoles [119], and condensed pyridine – pyrazine derivatives [120]. In the plant protection sector it is used as starting material in the synthesis of cyclopropanecarboxylates [121] and phenylurea derivatives [122].

1-Pentanol is employed in the production of many esters that are applied in different fields.

2-Methyl-1-butanol, like 1-pentanol, is used mainly as a solvent and an extracting agent and for the production of various esters; examples are given in [123]. Other important applications of 2-methyl-1-butanol include its use in the production of flotation reagents [124], liquid crystals [125], plant protection agents [126], fragrances [127], and corrosion inhibitors [128]. 2-Methyl-1-butanol is also part of the catalyst system employed in the polymerization of caprolactam [129]; it is used as a fuel additive [130] and in the extraction of crude oil [131]. A mixture of 1-pentanol and 2-methyl-1-butanol is generally used in practical applications instead of either of the pure alcohols.

3 -Methyl-1-Butanol. 3-Methyl-1-butanol is used mainly as a solvent and extracting agent in several practical applications [132], [133]. An important use is in the production of di(3-methylbutyl) ether; the latter is a component of a catalyst system used for the industrial production of polypropylene [134]. Other uses of 3-methyl-1-butanol in catalysis are discussed in [135].

3-Methyl-1-butanol is used in the production of herbicides based on 1,2,4-triazolecarboxamides [136], *N*-methyl- 4-pyridone derivatives [137], and substituted phenylhydrazones [138]. It is also employed for the synthesis of fragrances and flavors: 3-methyl-1-butanol reacts e.g., with ethylidenenorbornene to give the corresponding norbornyl ether [139], with bicyclopentadiene to yield tricyclo[5.2.1.$0^{2,6}$]dec-3-en-8(9)-yl-3-methylbutyl ether [140], and with vanillin alcohol to give the corresponding vanillin ether derivative [141].

3-Methyl-1-butanol is also used for the production of various additives [142], corrosion inhibitors [143], and azo dyes [144]. It is a component of surfactants [145].

Other Pentanols.

2-Methyl-2-butanol is employed as solvent for coating materials based on epoxy resins and polyurethanes [146], in the production of aryl pyruvic acids [147], and in the oxidation of olefins to the corresponding carbonyl compounds [148]. It is also used as a stabilizing agent for 1,1,1-trichloroethane, chlorinated C_3-C_4-hydrocarbons, and *N,N,*- dialkylallylamines [149–153]. The reaction of 2,2,6,6-tetramethyl- 4- cyano-4-hydroxypiperidine with 2-methyl-2-butanol yields highly active light stabilizers for polymers [154].

2-Pentanol is used as a solvent for primers containing phosphoric acid [150] and for the extraction of 2-hydroxy- 4-methylthiobutyric acid [151].

3-Methyl-2-butanol is employed as a stabilizing agent for tetrachloroethane [155].

2-Pentanol, 3-pentanol, and *2-methyl-2-butanol* are used in the production of catalyst systems for dimerization and polymerization, and for the synthesis of methanol from carbon monoxide and hydrogen [156].

Other applications of C_5-alcohols that are inaccessible by the oxo synthesis are in the production of lubricant additives [157], cleaning agents [158], defrosting agents [159], various herbicides [160], and many fragrances [161].

6. Quality Specifications and Chemical Analysis

The eight C_5-alcohol isomers are available commercially as pure substances or as mixtures of isomers. Generally, high purity is required because of the variety of applications.

The quality of C_5-alcohols is controlled by determination of various characteristics such as density, refractive index, color, water content, boiling point, and hydroxyl value. These properties are usually determined according to ASTM, DIN, or BS methods (→ Alcohols, Aliphatic, Chap. 2.6.). Isomeric purity and traces of byproducts are best determined by gas chromatography. Polar or nonpolar packed columns, as well as appropriate capillary colums, can be employed for separation. A thermal conductivity detector system is used for general purity determination; a flame ionization detector is employed for trace analysis (of byproducts); suitable carrier gases are hydrogen, nitrogen, and helium.

The commercially available C_5-alcohol mixture obtained from hydroformylation of 1-butene with subsequent hydrogenation contains ca. 70 % 1-pentanol and 30 % 2-methyl-1-butanol. This product is marketed under the name amyl alcohol. Typical specifications for amyl alcohol may be summarized as follows:

Purity (gas chromatography)	min. 98 wt %
Boiling range at 101.3 kPa, min. 95 vol % (DIN 53 171)	133 – 139 °C
Density at 20 °C (DIN 51 757)	0.813 – 0.817 g/cm^3
Refractive index n_D at 20 °C (DIN 51 423, part 2)	1.410 – 1.411
Hazen color number (DIN 53 409)	max. 10 APHA
Water content (DIN 51 777, modif.)	max. 0.2 wt %
fp	< – 60 °C
Viscosity at 20 °C (DIN 51 562)	ca. 4.3 mPa · s
Solubility in water at 20 °C	2.5 wt %
Water absorption at 20 °C	10 wt %
Vapor pressure at 20 °C	0.3 kPa

7. Storage and Transportation

Pentanols can be stored or dispatched in normal or enameled steel containers, provided that ingress of moisture is prevented. Stainless steel containers can also be used. National and international safety regulations relating to air, sea, inland water, rail and road transport and current safety data for pentanols must be observed.

Flash points of the eight C_5-alcohol isomers lie between 30 and 50 °C; they belong to safety class A II (VbF). Some safety specifications for the pentanols available through the oxo synthesis are compiled in Table 6.

8. Toxicology and Occupational Health

Kinetics and Metabolism. Pentanols are absorbed quickly after oral, dermal, and inhalative administration. *Primary pentanols* are very quickly oxidized (especially in the liver) to form the corresponding aldehydes. The latter quickly oxidize further to acids. Oxidation of *secondary pentanols* to the corresponding ketones proceeds more slowly and takes place only partly in the liver. *Tertiary pentanol* is mainly excreted unchanged or as glucuronide. Elimination takes 3.5 – 9 h for primary pentanols, 13 – 16 h for secondary pentanols, and 50 h for tertiary pentanol [162]. The metabolite valeric acid is eliminated very slowly and is probably responsible for the hangover effects observed after intake of fusel oils [163].

Acute Toxicity. Pentanols have stronger narcotic and toxic effects than the lower alcohols. The LD_{50} values after oral administration are in the range 1.0 – 5.77 g/kg body weight; primary pentanols have the lowest toxicity and tertiary alcohol has the highest [164]. The main symptoms are CNS depression and liver or kidney damage [165], [166]. Pentanols have an irritating effect on the skin and mucous membranes. Humans who inhale pentanol vapors exhibit irritation of the eyes and trachea, headache, dizziness, nausea, vomiting, diarrhea, breathing problems, coughing, and central nervous symptoms [164]. 3-Methyl-1-butanol is not a sensitizer in humans [167]. 2-Methyl-2-butanol was previously used as a sedative.

Subchronic and Chronic Toxicity. Daily oral administration of *1-pentanol* to rats for 13 weeks (dose 1000 mg/kg body weight) did not produce any toxic effects [168]. Daily oral administration of *3-methyl-1-butanol* to rats for 17 weeks (dose 1000 mg/kg body weight) had a slightly adverse effect on body weight gain and caused a slight reduction in food intake [169]. Rats that received 3-methyl-1-butanol for their lifetime, either as an oral dose of 81 mg/kg body weight twice a week or as a subcutaneous dose of 32.5 mg/kg body weight per week, showed higher tumor incidence. Furthermore, haematotoxic and hepatotoxic effects were observed [170]. Repeated dermal application of *tertiary pentanol* to rabbits (dose 3440 mg/kg body weight per day) resulted in serious systemic effects and death. Animals that received a dose of 344 mg/kg body weight per day under similar conditions showed slight irritation at the site of application. Thirteen-week inhalative exposure of 3600 mg/m^3 of 2-methyl-2-butanol to rats, mice, and dogs resulted in impairment of motor coordination, lacrimation, and increased liver weight in rats and dogs. Inhalative exposure of 810 mg/m^3 of 2-methyl-2-butanol caused slight irritation of mucous membranes; exposure to 180 mg/m^3 of *tert*-pentanol had no adverse effects [164].

Mutagenicity. 2-Methyl-2-butanol did not exhibit mutagenic activity in the Ames test [164]. In contrast, higher reversion rates were observed with pentanol on *Escherichia coli* [171]. After incubation of V 79 cells (fibroblasts) with 1-pentanol, an increase in the number

Table 6. Safety data for pentanols

Compound	LD_{50}, [a] mg/kg	Flash point, [b] °C	Ignition temperature [c], °C	Temperature class [d]	Group and hazard class [e]	Class and number of GGVE/GGVS/RID/ADR/ADNR	IMDG code	UN no.
1-Pentanol	3030	50	310	T2	AII	3/3	3.3	1105
2-Methyl-1-butanol	4920	43	400	T2	AII	3/3	3.3	1105
3-Methyl-1-butanol	1300	44	365	T2	AII	3/3	3.3	1105
Amyl alcohol	2690	43	365	T2	AII	3/3	3.3	1105

[a] Oral, rat.
[b] DIN 51 755/51 758.
[c] DIN 51 794.
[d] DIN 57 165/VDE 0165.
[e] According to VbF.

of cells with more than 22 chromosomes was reported [172].

Teratogenicity. Exposure of pregnant rats to 1-pentanol at a concentration of 14 000 mg/m^3 of air resulted in decreased food intake in dams. The juvenile exhibited retarded bone development; no malformations were found [173].

Occupational Health. A TLV – TWA value of 360 mg/m^3 and an MAK value of 360 mg/m^3 have been set for 3-methyl-1-butanol [174], [175].

9. References

1. R. Stephenson, J. Stuart, M. Tabak, *J. Chem. Eng. Data* **29** (1984) 287 – 290. Z. Maczynska, *Solubility Data Ser.* **15** (1984) 159 – 161. A. Maczyski, *Solubility Data Ser.* **15** (1984) 203 – 209.
2. L. Adembri, *Ann. Chim. (Rome)* **46** (1956) 62 – 68. M. Windholz et al. in: *The Merck Index,* 9th ed., Merck, Rahway, NJ, 1976, no. 5905.*Handbook of Chemistry and Physics,* 55th ed., CRC Press, Cleveland, Ohio, 1974/75. *Beilstein* **E IV 1,** 1655 – 1656, 1666.
3. L. H. Horsley: "Azeotropic Data III," *Adv. Chem. Ser.* **116** (1973).
4. J. Falbe (ed.): "Grundgerüste," *Methodicum Chimicum,* **vol. 4,** Thieme Verlag, Stuttgart 1980.
5. Phillips Petroleum Co., US 4 260 845, 1981 (T. K. Shioyama).
6. Institut Français du Pétrole, FR 1 603 748, 1971 (J. M. Bonnier, J. P. Damon, B. Delmon).
7. B. H. Davis, *J. Catal.* **79** (1983) 58 – 69. A. H. Dabbagh, B. H. Davis, *J. Catal.* **110** (1988) 416 – 418. A. H. Dabbagh, B. H. Davis, *J. Mol. Catal.* **47** (1988) 123 – 127.
8. Phillips Petroleum Co., US 4 529 827, 1985 (C. A. Drake); US 4 602 119, 1986 (C. A. Drake).
9. Phillips Petroleum Co., US 4 234 752, 1980 (Y. Wu, S. J. Marwil).
10. V. P. Kasumova, A. M. Ragimova, *Model. Optim. Khim. Protsessov* **3** (1977) 128 – 131; *Chem. Abstr.* **89** (1978) 163 154.
11. Hindustan Lever Ltd., IN 161 104, 1987 (S. U. Kulkarni).
12. British Petroleum Co., GB 2 164 636, 1986 (P. A. Diddams, W. Jones, J. M. Thomas).
13. British Petroleum Co., GB 2 179 563, 1987 (J. A. Ballantine et al.).
14. M. He, X. Fu, G. Zeng, *Xinan Shifan Daxue Xuebao, Ziran Kexueban* **2** (1987) 114 – 116; *Chem. Abstr.* **109** (1988) 189 789.
15. H. Pines, J. Hensel, J. Simonik, *J. Catal.* **24** (1972) 206 – 210. E. Licht, Y. Schaechter, H. Pines, *J. Catal.* **55** (1978) 191 – 197.
16. J. A. Ballantine et al., *J. Chem. Soc., Chem. Commun.* (1981), no. 9, 427 – 428.
17. J. Simonik, H. Pines, *J. Catal.* **24** (1972) 197 – 205.
18. E. Licht, Y. Schaechter, H. Pines, *J. Catal.* **31** (1973) 110 – 118.
19. Phillips Petroleum Co., US 4 304 943, 1981 (G. Bjornson).
20. B. M. Choudary, P. N. Reddy, L. M. Kantam, Z. Jamil, *Tetrahedron Lett.* **26** (1985) 6257 – 6258.
21. Nippon Shokubai Kagaku Kogyo Co., EP-A 183 225, 1986 (Y. Shimasaki, Y. Hino, M. Ueshima).
22. C. Jallabert, H. Riviere, *Tetrahedron* **36** (1980) 1191 – 1194.

23. V. P. Somaiah, B. K. Reddy, B. Sethuram, N. T. Rao, *Indian J. Chem. Sect. A* **27A** (1988) no. 10, 876 – 879; *Chem. Abstr.* **111** (1989) 6716.
24. S. L. Huang, K. Omura, D. Swern, *J. Org. Chem.* **41** (1976) 3329 – 3331.
25. J. San Filippo, C. I. Chern, *J. Org. Chem.* **42** (1977) 2182 – 2183.
26. Agence Nationale de Valorisation de la Recherche, FR 2 441 602, 1980 (D. Barton, W. Motherwell).
27. Mitsui Petrochemical Industries, JP-Kokai 62 142 134, 1987 (T. Fujita, K. Mizuno, K. Saeki).
28. A. Dobson, S. D. Robinson, *Inorg. Chem.* **16** (1977) 137 – 142.
29. Wacker-Chemie, DE-OS 2 358 254, 1975 (O. Sommer, F. Knoerr).
30. M. E. Marmion, K. J. Takeuchi, *J. Chem. Soc. Dalton Trans.* (1988), no. 9, 2385 – 2391.
31. W. L. Waters et al., *J. Org. Chem.* **41** (1976) 889 –891.
32. S. N. Murthy, E. V. Sundaram, *Indian J. Chem. Sect. A* **16 A** (1978) no. 9, 806 – 808; *Chem. Abstr.* **90** (1979) 38 354.
33. C. Walling, D. M. Camaioni, *J. Org. Chem.* **43** (1978) 3266 – 3271.
34. Kogai Boshi Chosa Kenkyusho, DE-OS 2 826 065, 1979 (W. Ando, I. Nakaoka).
35. J. N. Milovanovic, M. Vasojevic, S. Gojkovic, *J. Chem. Soc. Perkin Trans. 2* 1988, no. 4, 533 – 535.
36. C. W. Jefford, Y. Wang, *J. Chem. Soc. Chem. Commun.* (1988), no. 10, 634 – 635.
37. S. Miyazaki, Y. Suhara, *J. Am. Oil Chem. Soc.* **55** (1978) 536 – 538.
38. J. Kaulen, H. J. Schaefer, *Tetrahedron* **38** (1982) 3299 –3308.
39. Knapsack, FR 2 186 457, 1974.
40. W. Werner, *J. Chem. Res. Synop.* (1980), no. 6, 196.
41. A. S. Kyazimov et al., *Dokl. Akad. Nauk. Az. SSR* **39** (1983) 39 – 42; *Chem. Abstr.* **102** (1985) 24 153.
42. Japan Tobacco Inc., EP-A 230 286, 1987 (H. Matsushita, M. Shibagaki, K. Takahashi).
43. Ciba-Geigy, EP-A 278 914, 1988 (U. Frei, R. Kirchmayr).
44. S. Murahashi et al., *J. Org. Chem.* **52** (1987) 4319 –4327. S. Murahashi et al., *Tetrahedron Lett.* **22** (1981) 5327 –5330.
45. Y. Blum, Y. Shvo, *J. Organomet. Chem.* **263** (1984) 93 –107.
46. S. O. Nwaukwa, P. M. Keehn, *Tetrahedron Lett.* **23** (1982) 35 – 38.
47. W. J. Bailey, J. R. Griffith, *J. Org. Chem.* **43** (1978) 2690 – 2692.
48. BASF AG, DE-OS 3 145 736, 1983 (W. Boell, M. Kroener, K. H. Beyer, D. Hertel).
49. Ube Industries, JP 53 036 466, 1978 (S. Yamazaki et al.). Asahi Chemical Ind., JP-Kokai 57 156 481, 1982; JP-Kokai 57 167 980, 1982.
50. BASF AG, DE-OS 3 242 195, 1984 (F. Brunnmueller, M. Kroener).
51. Eastman Kodak Co., EP-A 96 641, 1983 (H. G. Rajoharison, C. M. A. Roussel).
52. Miland-Yorkshire Tar Distillers, DE-OS 2 225 443, 1973 (N. R. Clark, W. E. Webster).
53. Pfizer Inc., DE-OS 2 522 568, 1975 (R. G. Berg, B. E. Tate).
54. Kuraray Co., JP-Kokai 54 148 711, 1979 (S. Kyo, H. Tsuchiya, H. Tanako); JP-Kokai 54 066 613, 1979 (H. Tanaka, H. Tsuchiya, S. Kyo); DE-OS 2 918 521, 1979 (S. Kyo, H. Tsuchiya, H. Tanaka).
55. Bayer AG, EP-A 22 532, 1981 (F. Werner et al.).
56. CIL Inc., ZA 80 4227, 1981 (D. J. Gannon, D. T. F. Fung). J. Sejbl, *Rudy* **31** (1983) 163 – 165; *Chem. Abstr.* **99** (1983) 175 192.
57. Sun Abbott Co., JP-Kokai 57 058 645, 1982.
58. J. Sabadie, G. Descotes, *Bull. Soc. Chim. Fr.* (1983), no. 9 – 10, Part 2, 253 – 256. P. L. Burk, R. L. Pruett, K. S. Campo, *J. Mol. Catal.* **33** (1985) 1 – 21.
59. Kanebo, JP-Kokai 54 115 307, 1979 (K. Honda, Y. Nishijima).
60. British Petroleum Co., EP-A 147 983, 1985 (C. G. Griggs); EP-A 147 133, 1985 (C. G. Griggs, D. A. Pippard).
61. G. Jenner, G. Bitsi, *J. Mol. Catal.* **45** (1988) 235 – 246.
62. Nissan Chemical Ind., JP-Kokai 59 139 376, 1984; JP-Kokai 60 104 080, 1985.
63. H. Bieber, *Encycl. Chem. Process. Des.* **3** (1977) 278 – 289.
64. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1988.
65. W. J. Scheidmeir, *Chem. Ztg.* **96** (1972) 383 – 387.
66. Union Carbide Corp., US 4 717 775, 1988 (E. Billig, et al.).
67. Exxon EP-A 195 656, 1986 (D. A. Young).
68. Exxon Res. and Eng., WO 84 3697, 1984 (A. A. Oswald, J. S. Merola, J. C. Reisch, R. V. Kastrup).

69. UOP Inc., US 4 438 287, 1984 (T. Imai). Inst. de Cercetari Ing. Tehnologica, RO 83 795, 1984 (I. Muja, D. Goidea).
70. Monsanto Co., EP-A 94 456, 1983 (G. E. Barker, D. Forster).
71. Union Carbide Corp., DE-OS 3 737 277, 1988 (J. E. Logsdon, R. H. Loke, J. S. Merriam, R. W. Voight); EP-A 8767, 1980 (C. C. Pai).
72. J. M. Bonnier, J. P. Damon, J. Masson, *Appl. Catal.* **30** (1987) 181 – 184.
73. M. Shibagaki, K. Takahashi, H. Matsushita, *Bull. Chem. Soc. Jpn.* **61** (1988) 3283 – 3288. Japan Tobacco Inc., JP-Kokai 86 204 143, 1986 (S. Mizusaki et al.).
74. F. S. Delk, A. Vavere, *J. Catal.* **85** (1984) 380 – 388.
75. Toa Nenryo Kogyo, EP-A 127 486, 1984 (Y. Okumura, S. Kamiyama, H. Furukawa, K. Kaneko).
76. EC Erdölchemie, EP-A 325 144, 1989 (R. Malessa, B. Schleppinghoff). B. Schleppinghoff, C. Gabel, H. L. Niederberger, *Erdöl Erdgas* **104** (1988) 173 – 177. Dow Chemical Co., US 4 182 920, 1980 (J. H. Giles, J. H. Stultz, S. W. Jones).
77. Z. Prokop, K. Setinek, *Collect. Czech. Chem. Commun.* **52** (1987) 1272 – 1279.
78. Centre National de la Recherche Scientifique, Institut Français du Pétrole, BE 904 797, 1986 (F. F. Francois et al.).
79. J. A. Ballantine et al., *J. Mol. Catal.* **26** (1984) 57 – 77.
80. A. A. Makhnin, A. F. Frolov, K. A. Aronovich, *Kinet. Katal.* **22** (1981) 344 – 348; *Chem. Abstr.* **95** (1981) 79 557. A. A. Makhnin et al., *Khim Tekhnol. Topl. Masel.* 1979, 34 – 35; *Chem. Abstr.* **91** (1979) 6161.
81. Japan Oil Co., JP-Kokai 72 39007, 1972 (A. Nambu); JP-Kokai 72 39013, 1972 (A. Nambu).
82. Gulf Res. and Development Co., US 3 285 977, 1966 (A. M. Henke, R. C. Odioso, B. K. Schmid).
83. Y. Aoyama et al., *J. Org. Chem.* **52** (1987) 2555 –2559. Aldrich-Boranes Inc., EP-A 188 235, 1986 (H. C. Brown). H. C. Brown, U. S. Racherla, *J. Org. Chem.* **51** (1986) 895 – 897. H. C. Brown, N. Ravindran, *J. Org. Chem.* **42** (1977) 2533 – 2534.
84. H. C. Brown, N. Ravindran, *Synthesis* 1977, 695 – 697.
85. A. Pelter, S. Singaram, H. C. Brown, *Tetrahedron Lett.* **24** (1983) 1433 – 1436.
86. H. C. Brown, M. C. Desai, P. K. Jadhav, *J. Org. Chem.* **47** (1982) 5065 – 5069. H. C. Brown, P. K. Jadhav, A. K. Mandal, *J. Org. Chem.* **47** (1982) 5074 – 5083. H. C. Brown, J. A. Sikorski, *Organometallics* **1** (1982) 28 – 37. H. C. Brown, B. Singaram, C. P. Mathew, *J. Org. Chem.* **46** (1981) 2712 – 2717. M. Anez, G. Uribe, L. Mendoza, R. Contreras, *Synthesis* (1981), 214 – 216.
87. Hoechst AG, Katalysatoren Hoechst, Frankfurt 1979.
88. H. Fujitsu et al., *J. Chem. Soc. Perkin Trans. 1* (1981), 2650 – 2655.
89. F. Ciardelli, P. Pertici, *Z. Naturforsch. B Anorg. Chem. Org. Chem.* **40B** (1985) 133 – 140.
90. Studiengesellschaft Kohle, DE-OS 3 536 797, 1987 (B. Bogdanovic, M. Schwickardi).
91. Kao Corp., JP-Kokai 87 210 056, 1987 (Y. Kajiwara, Y. Inamoto, J. Tsuji).
92. Air Products and Chemicals Inc., EP-A 180 210, 1986 (S. A. Butter, I. Stoll).
93. Japan Tobacco Inc., EP-A 285 786, 1988 (H. Matsuhita, M. Shibagaki, K. Takahashi). Mitsubishi Chemical Industries Co., JP-Kokai 87 108 832, 1987 (T. Maki, M. Nakajima, T. Yokoyama, T. Setoyama).
94. Akademie der Wissenschaften, DD 142 050, 1980 (M. Wahren et al.).
95. H. C. Brown, S. Narasimhan, *J. Org. Chem.* **49** (1984) 3891 – 3898. K. Soai, H. Oyamada, M. Takase, A. Ookawa, *Bull. Chem. Soc. Jpn.* **57** (1984) 1948 – 1953.
96. Institut Français du Pétrole, EP-A 172 091, 1986 (J. P. Bournonville, J. P. Candy, G. Mabilor).
97. Anic S.p.A., DE-OS 2 945 348, 1980 (G. Dozzi, S. Cucinella); DE-OS 2 849 767, 1979 (S. Cucinella).
98. Japan Tobacco Inc., EP-A 271 092, 1988 (H. Matsushita, M. Shibagaki, K. Takahashi).
99. Standard Oil Co., US 4 677 091, 1987 (T. J. Mazanec, J. G. Frye). Eastman Kodak Co., US 4 748 144, 1988 (J. R. Monnier, G. R. Apai, M. J. Hanrahan). Belge Secretaire General des Services de Programmation, EP-A 283 635, 1988 (A. Noels et al.). W. X. Pau, R. Cao, G. L. Griffin, *J. Catal.* **114** (1988) 447 – 456. J. G. Nunan et al., *J. Catal.* **116** (1989) 195 – 221. E. Tronconi et al., *Appl. Catal.* **47** (1989) 317 – 333. Metallgesellschaft, DE-OS 3 403 492, 1985 (G. Cornelius et al.). Phillips Petroleum Co., US 4 537 909, 1985 (F. N. Lin, F. Pennella).
100. Metallgesellschaft, DE-OS 3 641 774, 1988 (W. Hilsebein, E. Supp, F. Möller, P. König). Dow Chemical Co., EP-A 180 719, 1986 (G. J.

Quarderer et al.). Union Oil Co. of California, US 4 540 836, 1985 (D. M. Fenton).
101. A., V. Devekki et al., *Zh. Obshch. Khim.* **53** (1983) 2652 – 2653; *Chem. Abstr.* **100** (1984) 156303.
102. N. N. Gonchakova et al., *Neftekhimiya* **21** (1981) 567 – 575; *Chem. Abstr.* **95** (1981) 188 395. Kuraray Co., JP-Kokai 74 11806, 1974 (S. Kyo, T. Yasui). Mitsubishi Gas Chemical Co., JP 73 43084, 1973 (A. Miyamoto et al.).
103. C. Moberg, L. Rakos, *J. Organomet. Chem.* **335** (1987) 125 – 131. G. D. Zakumbaeva, L. A. Beketaeva, S. Z. Aitmagambetova, *Izv. Akad. Nauk. Kaz. SSR Ser. Khim.* (1979), 17 – 21; *Chem. Abstr.* **92** (1980) 11 749. N. A. Zakarina, V. P. Kuidina, A. M. Khisametdinov, *Izv. Akad. Nauk. Kaz. SSR Ser. Khim.* (1982), 30 – 33; *Chem. Abstr.* **96** (1982) 217 175. D. V. Sokolskii et al., *Zh. Khim.* **23** (1987) 1430 – 1432; *Chem. Abstr.* **108** (1988) 166 935.
104. Institut Français du Pétrole, FR 2 175 689, 1973 (H. Mimoun, C. Bocard, I. Seree de Roch). K. S. Suslick, B. R. Cook, M. Fox, *J. Chem. Soc. Chem. Commun.* (1985), 580 – 582. B. R. Cook, T. J. Reinert, K. S. Suslick, *J. Am. Chem. Soc.* **108** (1986) 7281 – 7286. C. A. Dewar, C. J. Suckling, R. Higgins, *J. Chem. Res. Synop.* 1979, 335. R. V. Nazarova et al., *Issled. Obl. Khim. Vysokomol. Soedin. Neftekhim.* (1977), 23; *Chem. Abstr.* **93** (1980) 45 891.
105. S. A. Yurzhenko, V. A. Lobov, *Neftepererab. Neftekhim. (Kiev)* **15** (1977) 92 – 95; *Chem. Abstr.* **88** (1978) 89 034.
106. Hoechst AG, Alkohole Hoechst, Frankfurt 1984.
107. Standard Oil Co., BE 896 226, 1983.
108. Texaco Inc., US 4 761 465, 1988 (G. P. Speranza, J. J. Lin, M. Cuscurida).
109. Lubrizol Corp., US 4 764 297, 1988 (R. L. Sowerby, S. A. Di Biase).
110. Rohm and Haas Co., US 4 594 363, 1986 (R. M. Blankenship, A. Kowalski).
111. Sugai Chemical Industry Co., EP-A 274 379, 1988 (T. Yamashita, M. Kodama, S. Shimada).
112. K. P. Ehlers, *Ber. Bunsen-Ges. Phys. Chem.* **83** (1979) 1113 – 1116.
113. VEB Erdoel-Erdgas Gommern, DD 243 687, 1987 (H. Richter et al.).
114. Kanto Chemical Co., JP-Kokai 01 51538, 1989 (K. Suzuki, A. Sugiura, T. Fujii). Mitsui Toatsu Chemicals, EP-A 305 938, 1989 (M. Nakatsuka et al.). Alps Electric Co., JP-Kokai 62 294 663, 1987 (M. Kano). ICI, EP-A 259 995, 1988 (D. A. Jackson, P. A. Gemmell).
115. Hoechst AG, DE-OS 3 644 661, 1988 (K. Guenther, G. Mau). Hitachi Chemical Co., EP-A 284 370, 1988 (S. Tai et al.).
116. International Flavors and Fragrances Inc., US 4 435 428, 1984 (R. M. Boden, M. R. Hanna, T. J. Tyszkiewicz). BASF AG, DE-OS 3 116 913, 1982 (F. Thoemel, W. Hoffmann).
117. L. Cerveny et al., *React. Kinet. Catal. Lett.* **33** (1987) 471 – 476. Nisso Petrochemical Industries Co., DE-OS 3 025 434, 1981 (K. Kogoma et al.).
118. Rhône-Poulenc, FR 2 607 506, 1988 (J. D. Bourzat et al).
119. Lonza, CH 653 022, 1985 (R. Bergamin, B. Bourgeois, J. Jovanovic).
120. Shionogi and Co., EP-A 157 260, 1985 (I. Adachi, T. Yamamori, M. Ueda). Lion Corp., DE-OS 3 432 983, 1985 (S. Suzuki).
121. Roussel-Uclaf, CA 1 248 455, 1989 (J. Martel, J. Tessier, A. Teche).
122. Takeda Chemical Industries, JP-Kokai 62 205 053, 1987 (J. Yamada, H. Miki, T. Uekado).
123. Celanese Corp., US 4 470 881, 1984 (L. Berg). BASF Wyandotte Corp., US 4 426 291, 1984 (S. G. Sharangpani, F. G. Spence). S. C. Johnson and Son Inc., EP-A 71 116, 1983 (G. O. Schulz, D. M. Wilson).
124. Lubrizol Corp., WO 89 00 457, 1989 (A. C. Clark, E. P. Richards, D. R. Shaw).
125. Centre National de la Recherche Scientifique, FR 2 612 182, 1988 (N. Huu Tinh, C. Salleneuve, C. J. P. Destrade). Idemitsu Kosan Co., EP-A 258 898, 1988 (K. Morita, S. Uchida, S. Hachiya). Merck, WO 87 05 018, 1987 (R. Eidenschink et al.). Canon K.K., DE-OS 3 629 446, 1987 (K. Yoshinaga, K. Katagiri, T. Kobayashi, K. Shinjo).
126. Monsanto Co., FR 2 479 200, 1981 (G. H. Alt); FR-A 2 479 202, 1981 (G. H. Alt); US 4 284 564, 1981 (G. H. Alt, J. P. Chupp).
127. Toray Industries Inc., WO 84 00 161, 1984 (M. Naruto, H. Kondo, G. Hata).
128. Bayer AG, DE-OS 2 535 685, 1977 (H. D. Block).
129. Stamicarbon, EP-A 238 143, 1987 (J. J. M. Bongers, A. A. Van Geenen).
130. Union Carbide Corp., BR 8 306 714, 1984 (E. J. Derderian). Texaco Inc., US 4 539 014, 1985 (W. M. Sweeney); US 4 713 087, 1987 (R. H. Jenkins et al.).

131. M. A. Dahami, W. D. Constant, J. M. Wolcott, *Fuel* **67** (1988) 1242 – 1248. Marathon Oil Co., US 3 613 786, 1971 (S. C. Jones, W. O. Roszelle, M. A. Svaldi).
132. Mitsubishi Gas Chemical Co., EP-A 127 888, 1984 (T. Isshiki, T. Yui, H. Uno, M. Abe). Mobil Oil Corp., US 4 532 084, 1985 (A. O. M. Okorodudu). Soken Chemical Eng. Co., JP 62 045 580, 1987 (S. Morikawa). V. Macho et al., CS 254 225, 1988.
133. Haarmann und Reimer, DE-OS 2 729 075, 1979 (K. Bauer et al.). Negev Phosphates, DE-OS 3 438 655, 1986 (A. Iosef, M. Bar-On, J. Oren, D. Kellerman). M. Muhammed, Y. Zhang, *Hydrometallurgy* **21** (1989) 277 – 292.
134. Solvay, EP-A 288 109, 1988 (P. Fiasse); EP-A 261 727, 1988 (P. Fiasse, A. Bernard). Chisso Corp., EP-A 284 287, 1988 (J. Masuda, T. Shiraishi, N. Komori, T. Uwai). BP Chemicals, EP-A 238 345, 1987 (J. C. A. Bailly, S. Sandis).
135. Nippon Oil Co., EP-A 143 334, 1985 (M. Matsuno, M. Kudoh, H. Imai). A. Sarpudeen, I. Tajuddin, Q. Anwaruddin, *Eur. Polym. J.* **20** (1984) 1019 – 1020. K. Kishore et al., *J. Appl. Polym. Sci.* **31** (1986) 2829 – 2837.
136. Kureha Chemical Industry Co., JP 63 230 676, 1988 (T. Shida et al.). Kureha Chemical Industry Co., JP 63 230 664, 1988 (T. Shida, Y. Kubota, I. Ichinose).
137. Bayer AG, DE-OS 3 430 232, 1986 (H. J. Knops et al.).
138. Kureha Chemical Industry, EP-A 285 294, 1988 (T. Shida et al.).
139. International Flavors and Fragrances, US 4 379 060, 1983 (M. A. Sprecker); US 4 311 861, 1982 (M. A. Sprecker); US 4 369 133, 1983 (M. A. Sprecker).
140. BASF AG, DE-OS 2 642 519, 1978 (W. Hoffmann, L. Schuster).
141. Takasago Perfumery Co., JP 57 009 729, 1982.
142. Standard Oil Co., US 4 155 958, 1979 (E. K. Fields). VEB Petrolchemisches Kombinat Schwedt, DD 201 146, 1983 (H. Brendel et al.). Lubrizol Corp., WO 88 04 685, 1988 (T. F. Steckel).
143. L. Komora et al., CS 229 760, 1986.
144. Ciba-Geigy, DE-OS 2 922 227, 1979 (R. Portmann).
145. Phillips Petroleum, US 4 504 399, 1985 (P. R. Stapp). UOP Inc., US 4 501 675, 1985 (T. P. Malloy, R. J. Swedo).
146. Lechler Chemie, DE-OS 3 033 613, 1982 (W. Oechsner, O. Vogt). Dow Chemical Co., US 4 778 863, 1988 (C. S. Wang, Z. K. Liao).
147. Sagami Chemical Research Center, JP 60 061 550, 1985; JP 62 116 541, 1987 (K. Hirai, T. Fuchigami).
148. Shell Internationale Research Maatschappij, GB 2 167 406, 1986 (B. L. Feringa).
149. Dow Chemical Co., US 4 808 440, 1989 (E. L. Tasset).
150. Monsanto Co., EP-A 143 100, 1985 (D. A. Ruest, M. Takano, L. R. Wolf); EP-A 142 488, 1985 (D. A. Ruest, M. Takano, L. R. Wolf).
151. Rhône-Poulenc, GB 2 024 242, 1980 (Y. Correia, J. Lesparre). Toyo Soda Mfg. Co., JP 61 189 237, 1986 (Y. Oda, J. Mizoguchi, K. Yabuta); JP 61 191 629, 1986 (Y. Oda, J. Mizoguchi, K. Yabuta). Dow Chemical Co., US 4 309 301, 1982 (N. Ishibe, T. G. Metcalf).
152. Deutsche Solvay Werke, EP-A 309 958, 1989 (W. Dilla, G. Jakobson).
153. Wako Pure Chemical Industries, JP 49 125 309, 1974 (K. Yamanishi, Y. Takezoe).
154. Hoechst AG, DE-OS 2 742 582, 1979 (H. Wiezer, N. Mayer, G. Pfahler).
155. Asahi Chemical Ind., JP 69 20 965, 1969 (Y. Hashiguchi, T. Owaki).
156. Lithium Corp. of America, US 4 555 498, 1985 (C. W. Kamienski). E. Alder, H. Fuellbier, W. Gaube, DD 152 778, 1981; DD 152 779, 1981. Hercules Inc., EP-A 41 362, 1981 (W. P. Long, A. S. Matlack). Shell Internationale Research Maatschappij, EP-A 306 114, 1989 (S. T. Sie, E. Drent, W. W. Jager); EP-A 289 067, 1988 (S. T. Sie, E. Drent, W. W. Jager).
157. Asahi Denka Kogyo, JP 86 87 690, 1986 (S. Katsumata et al.).
158. Asahi Glass Co., JP 72 20 232, 1972 (H. Ukibashi, K. Kirimoto, A. Hara).
159. Hoechst AG, GB 2 027 046, 1980 (H. Peters, E. Surma).
160. American Cyanamid Co., US 4 188 487, 1980 (M. Los). Ciba-Geigy, DD 206 525, 1984 (W. Kunz, W. Eckhardt). Bayer, DE-OS 3 643 851, 1988 (K. Sasse et al.).
161. Mitsui Petrochemical Industries, DE-OS 2 427 465, 1974 (H. Miki, H. Hasui). International Flavors and Fragrances Inc., DE-OS 2 330 648, 1974 (J. B. Hall, M. G. Beets, L. E. Lala, W. I. Taylor). Mitsui Petrochemical Ind., DE-OS 2 427 465, 1979 (H. Miki, H. Hasui). Lucta S.A., ES 4 993 976, 1981 (A. Gavalda Sola).
162. E. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier Publ., Amsterdam 1965.
163. E. Rüdell et al., *Blutalkohol* **18** (1981) no. 5,315 – 325.

164. V. K. Rowe, S. B. McCollister in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 C,** Wiley-Interscience, New York 1982, p. 4527.
165. R. A. Scala, E. G. Burtis, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 493 – 499.
166. I. F. H. Purchase, *S. Afr. Med. J.* **43** (1969) 795 – 798.
167. D. L. J. Opdyke, *Food Cosmet. Toxicol.* **16** (1978) Suppl. 1, 785 – 787.
168. K. R. Butterworth et al., *Food Cosmet. Toxicol.* **16** (1978) 203 – 207.
169. F. M. B. Carpanini et al., *Food Cosmet. Toxicol.* **11** (1973) 713 – 724.
170. W. Gibel, K. Lohs, G. P. Wildner, *Arch. Geschwulstforsch.* **45** (1975) no. 1, 19 – 24.
171. H. Hilscher, E. Geissler, H. Lohs, W. Gibel, *Acta Biol. Med. Ger.* **23** (1969) 843 – 852.
172. A. önfelt, *Mutat. Res.* **182** (1987) 135 – 154.
173. B. K. Nelson et al., *J. Am. Coll. Toxicol.* **8** (1989) no. 2, 405 – 410.
174. ACGIH, TLV's–Threshold Limit Values and Biological Indices for 1989 – 1990, Cincinatti, Ohio, 1989.
175. DFG: *Maximale Arbeitsplatzkonzentrationen und biologische Arbeitsstofftoleranzwerte 1989,* VCH Verlagsgesellschaft, Weinheim 1989.

Peptic Ulcer Treatment → Antiulcer Drugs

Peptide and Protein Hormones

Interferons is a separate keyword.

WOLFGANG KÖNIG, Hoechst Aktiengesellschaft, Frankfurt am Main, Federal Republic of Germany

1. Introduction . . . 3
1.1. Synthesis of Peptide and Protein Hormones . . . 3
1.2. Mechanism of Action of Peptides and Protein Hormones . . . 4
1.3. Nomenclature and Abbreviations 4
2. Gonadoliberin, Thyroliberin, Gonadotropins, Thyrotropin, Inhibin, and Related Hormones . . . 7
2.1. Gonadoliberin . . . 8
2.2. Thyroliberin . . . 11
2.3. Gonadotropins . . . 12
2.4. Thyrotropin . . . 14
2.5. Inhibins, Activins, and Related Hormones . . . 14
2.5.1. Inhibins . . . 14
2.5.2. Activins . . . 15
2.5.3. Transforming Growth Factor-β . . . 15
2.5.4. Bone Morphogenetic Proteins and Osteogenic Proteins . . . 16
2.5.5. Muellerian Inhibiting Substance . . 17
2.6. Follistatin . . . 17
3. Parathyroid Hormone and the Calcitonin Family . . . 17
3.1. Parathyroid Hormone and PTH-Related Protein . . . 17
3.2. Calcitonin . . . 19
3.3. Calcitonin Gene Related Peptide 21
3.4. Amylin . . . 21
4. Corticoliberin – Proopiomelanocortin Cascade . . . 22
4.1. Corticoliberin . . . 23
4.2. Corticotropin . . . 25
4.3. Melanotropins . . . 27
4.4. Opioid Peptides . . . 29
4.4.1. β-Endorphin . . . 30
4.4.2. Preproenkephalins A and B . . . 31
4.4.2.1. Peptides Derived from Preproenkephalin A . . . 31
4.4.2.2. Peptides Derived from Preproenkephalin B . . . 33
4.4.3. Dermorphin and Deltorphins . . . 33
4.4.4. Exorphins . . . 34
4.4.5. Femarfarmamide and Related Structures . . . 35
5. Blood Pressure Regulating Peptides . . . 35
5.1. The Angiotensin – Kinin System . 35
5.1.1. Angiotensins . . . 36
5.1.2. Kinins . . . 38
5.2. Substance P, Neurokinins, and Tachykinins . . . 40
5.2.1. Introduction . . . 40
5.2.2. Amyloid A4 Protein . . . 42
5.3. Vasopressin, Oxytocin, and Melanostatin . . . 42
5.3.1. Vasopressin and Oxytocin . . . 42
5.3.2. Melanostatin . . . 46
5.4. Endothelin . . . 46
5.5. Atrial Natriuretic Factor . . . 47
5.6. Pancreatic Peptide Family . . . 49
5.6.1. Pancreatic Peptide . . . 50
5.6.2. Neuropeptide Y . . . 50
5.6.3. Peptide with N-terminal Tyrosine and C-terminal Tyrosinamide . . . 52
5.6.4. Neuromedin U . . . 52
5.6.5. Delta-Sleep-Inducing Peptide . . . 52
6. Cholecystokinin and Gastrin . . . 53
6.1. Trypsin-Sensitive Cholecystokinin Releasing Peptide . . . 53
6.2. Cholecystokinin and Cerulein . . 53
6.3. Gastrin Releasing Peptide and Bombesin . . . 55
6.4. Gastrin . . . 57
7. Secretin Family . . . 58
7.1. Secretin . . . 58
7.2. Vasoactive Intestinal Peptide . . . 60
7.3. Pituitary Adenylate Cyclase Activating Polypeptide (PACAP) . . . 61
7.4. Glucagon and Related Peptides . 62
7.4.1. Glucagon . . . 62
7.4.2. Glicentin and Enteroglucagon . . . 63
7.4.3. Glucagon-Like Peptide-1-(7 – 36)amide . . . 64
7.5. Gastrin Inhibiting Peptide . . . 64
7.6. Growth Hormone Releasing Hormone and the Growth Hormone Cascade . . . 65
7.6.1. Galanine . . . 65

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_061

7.6.2. Growth Hormone Releasing Hormone 66
7.6.3. Growth Hormone, Placental Lactogen, and Prolactin 68
7.6.3.1. Growth Hormone 68
7.6.3.2. Prolactin 70
7.6.3.3. Placental Lactogen 71
7.6.4. Somatomedins, Insulin, and Relaxin 71
7.6.4.1. Somatomedins 71
7.6.4.2. Insulin 73
7.6.4.3. Relaxin 75
7.6.5. Somatostatin 76
7.6.6. Pancreastatin 78
8. Neurotensin 78
9. Motilin 80
10. Pancreatic Spasmolytic Peptide . 80
11. References 81

1. Introduction

In the past, peptide hormones were mainly classified according to their site of synthesis or release, e.g., peptide hormones of the hypothalamus, pituitary gland, gastrointestinal tract, or thyroid gland. This classification is no longer expedient because peptide hormones are not just produced at a single site in the body. Classification according to clinical indications is also unsuitable because these hormones can develop different activities depending on the site of action and on the dose. Peptides that are formed in endocrine cells can have an endocrine or a paracrine function. An *endocrine peptide* is secreted directly into the bloodstream and produces an effect at a distant receptor. A peptide exerts a *paracrine effect* if, after being released, it affects the metabolism of either its own cell or a neighboring cell. Peptides that exert a paracrine effect are also referred to as peptide regulatory factors (PRFs) [1].

An attempt will be made to classify structurally similar peptides into families and, at the same time, to integrate the activity cascades [2].

1.1. Synthesis of Peptide and Protein Hormones [3]

Peptide and protein hormones are synthesized in the cell in the form of large prepropeptides which are then cleaved and modified by enzymes to give active products [4]. Many biologically active peptides have an amide group at the C-terminus. This amide group is formed from the C-terminal glycine by peptidylglycine-α-amidating monooxygenase.

Naturally occurring peptides can be isolated by conventional separation techniques (precipitation, chromatography etc.). However, the demand for increased purity and economic considerations have resulted in the development of methods for synthesizing peptides and proteins. R. GEIGER calculated that peptides up to a chain length of about 20 amino acids could be produced economically using peptide chemistry [3]. For peptides and proteins with a chain length exceeding 20 amino acids, genetic engineering methods are preferred.

Peptides are usually synthesized by condensation of amino acids starting at the C-terminus (→ Peptide Synthesis). Synthesis may be performed in solution or using a solid phase. In solid-phase synthesis the C-terminus of the amino acid is bound to a polymeric support. Various groups are used to protect side-chain functional groups. The hydroxyl groups of tyrosine, serine, and threonine are mostly protected by the formation of ethers. Amino groups are blocked by urethane protecting groups and carboxyl groups are esterified with alcohols. Peptide synthesis using proteases is becoming increasingly important alongside the purely chemical methods.

Foreign peptides and proteins can also be produced in genetically altered microorganisms (e.g., *Escherichia coli*) [5], [6]. Plasmids or cosmids are used to introduce synthetic, semisynthetic, or natural genetic material (DNA) into the microorganisms. The foreign DNA is then transcribed into messenger RNA which is finally translated into peptides and proteins by the protein synthesis "machinery" of the microorganism (→ Genetic Engineering).

DNA recombination using genetic engineering is also often used for cloning DNA sequences whose structure is to be elucidated. The struc-

tures of many proteins and peptides have been deduced from the base sequences of the DNA used to direct their synthesis.

1.2. Mechanism of Action of Peptides and Protein Hormones

Peptide and protein hormones bind to specific receptors where they stimulate the synthesis and activation of enzymes via second messengers such as intracellular calcium, cyclic 3′,5′-adenosine monophosphate (cAMP), cyclic 3′,5′-guanosine monophosphate (cGMP), inositol triphosphate, diacylglycerol, or lysophosphatidylinositol. The hormones also stimulate the release of prostaglandins, steroid hormones, thyroid hormones, peptides, and glycoproteins. Occasionally, internalization of the receptor may occur in which excess peptide hormones causes disappearance of the receptors (receptor down regulation). This results in temporary stimulation and subsequent inhibition of the target cells.

Peptide *agonists* are peptide hormone analogues which stimulate the cell. *Antagonists* are competitive analogues which bind to the peptide receptors, and thus compete with but do not stimulate the cell. They compete with the agonists for the peptide receptor and thus inhibit the activation of the cell stimulated by agonists without receptor down regulation.

1.3. Nomenclature and Abbreviations

According to IUPAC, amino acids can be abbreviated with three-letter [5] or one-letter symbols [6]. The one-letter nomenclature is becoming increasingly popular. Both systems are used here.

The regulatory peptides [7] and their analogues [8] will also be designated in accordance with IUPAC, if possible. Small letters in front of the name of the peptide refer to the origin of the peptide and provide information on the species. The first letters of the latin name are used for this purpose. The position of an amino acid in a peptide chain is denoted by a superscript, after the three-letter symbol (e.g., His^5). Regions within peptide chains are indicated in parentheses, e.g., β-EP-(1 – 27) denotes amino acids in positions 1 to 27 of β-endorphin.

	One-letter symbols for **l**-*amino acids*[a]
A	alanine
B	aspartic acid or asparagine[b]
C	cysteine
D	aspartic acid
E	glutamic acid
F	phenylalanine
G	glycine
H	histidine
I	isoleucine
K	lysine
L	leucine
M	methionine
N	asparagine
Ng	asparagine with *N*-glycosidic carbohydrate residue
P	proline
Ph	4-hydroxyproline
Q	glutamine
*Q	pyroglutamic acid
R	arginine
S	serine
Sg	serine with *O*-glycosidic carbohydrate residue
T	threonine
V	valine
W	tryptophan
X	unknown amino acid
Xa	amino acid amide
X-ol	amino acid with carboxyl group reduced to alcohol function
Y	tyrosine
Ys	*O*-sulfated tyrosine
Z	glutamic acid or glutamine[b]
/	amino acid deletion
[a]	D-amino acids are denoted by the corresponding small letters
[b]	uncertainty in structural elucidation

	Three-letter and other symbols for natural and unnatural amino acids and protecting groups
Abu	α-aminobutyric acid
Ac	acetyl
Aca	ε-aminocaproic acid
Ace	1-aminocyclopropane-1-carboxylic acid
Acm	acetamidomethyl
Ada	adamantylalanine
Agm	agmatin
Ahx	6-aminohexanoic acid
Aib	α-aminoisobutyric acid

Ala alanine
Ams unknown amino acid
Ams-NH_2 amino acid amide
Ams-ol amino acid with carboxyl group reduced to alcohol function
Aoa 8-aminooctanoic acid
Aoc azabicyclo[3.3.0]octane-3-carboxylic acid
Arg arginine
Asn asparagine
Asp aspartic acid
Asu α-aminosuberic acid
Asx aspartic acid or asparagine[b]
Aza denotes amino acid in which the α-CH group is replaced by N
Bip biphenylalanine
Boc *tert*-butoxycarbonyl
Bzl benzyl
Car carboranylalanine
Chg cyclohexylglycine
Cit citrulline
4-ClPhe 4-chlorophenylalanine
Cys cysteine
de denotes a deletion
Dip 4-β-diphenylalanine
Et ethyl
4-FPhe 4-fluorophenylalanine
Gln glutamine
Glu glutamic acid
Glx glutamic acid or glutamine[b]
Gly glycine
Har homoarginine
His histidine
Hmp L-2-hydroxy-3-mercaptopropionic acid
Hyp 4-hydroxyproline
Ile isoleucine
imBzl imidobenzyl
Lac L-lactic acid
Leu leucine
Lys lysine
Me methyl
Met methionine
Met (O) methionine sulfoxide
Met (O_2) methionine sulfone
Mpr 3-mercaptopropionic acid
Nal 3-(2-naphthyl)alanine
Nic nicotinoyl
Nle norleucine
Oic octahydroindolecarboxylic acid
Orn ornithine
Pal 3-(3-pyridyl)alanine
Peg phenethylglycine
Pen penicillinamine
Pgl phenylglycine
pGlu pyroglutamic acid
Phe phenylalanine
Phe(Cl_2) 3,4-dichlorophenylalanine
Phe(NO_2) 4-nitrophenylalanine
Phg phenylglycine
Pic picolyl
Pmp β,β-cyclopentamethylene-3-mercaptopropionic acid
Pro proline
Rha α-L-rhamnopyranosyl
Sar sarcosine (*N*-methylglycine)
Ser serine
Sta statine
Thi β-(2-thienyl)alanine
Thr threonine
Tic tetrahydroisoquinolinecarboxylic acid
Trp tryptophan
Tyr tyrosine
Tyr (Me) *O*-methyltyrosine
Z benzyloxycarbonyl
Δ dehydro
Ψ denotes a substitution of the peptide bond

Prefixes indicating species origin

anq- eel *(Anquilla)*
b- cow *(Bos)*
ca- dog *(Canis)*
cap- goat *(Capra)*
cav- guinea pig *(Caviidae)*
cu- rabbit *(Cuniculus)*
g- chicken *(Gallina)*
go- gudgeon *(Gobius)*
h- man *(Homo)*
m- mouse *(Mus)*
o- sheep *(Ovis)*
p- pig *(Porcus)*
pl- flounder *(Platichtys flesus)*
r- rat *(Rattus)*
ra European common frog *(Rana temporaria)*
sa- salmon *(Salmo)*
xe- xenopus *(Xenopus laevis)*

Abbreviations for peptides and other compounds

A4 amyloid A4 protein
ACE angiotensin converting enzyme
ACTH corticotropin (adenocorticotropic hormone)
AMP adenosine 5′-monophosphate

ANF atrial natriuretic factor
ANP atrial natriuretic peptide
APP amyloid precursor protein
AT angiotensin
ATG angiotensinogen
ATP adenosine 5′-triphosphate
ATV activin
BB bombesin
BK bradykinin
BMP bone morphogenic protein
BNP brain natriuretic peptide
cAMP cyclic 3′,5′-adenosine monophosphate
CCK cholecystokinin
CCKRP cholecystokinin releasing peptide
CG chorionic gonadotropin
cGMP cyclic 3′,5′-guanosine monophosphate
CGRP calcitonin gene related peptide
CLIP corticotropin-like intermediate lobe peptide
CM casomorphin
CNP C-type natriuretic peptide
CPON C flanking peptide of neuropeptide Y
CRH corticoliberin (corticotropin releasing hormone)
CRL cerulein
CT calcitonin
CTP celltropin
DFT dentinal fluid transport stimulating peptide
DM dermorphin
DP dynorphin
DSIP delta-sleep-inducing peptide
DT deltorphin
EDRF epidermis derived relaxing factor (nitrogen monoxide)
EGF epidermal growth factor
EK enkephalin
EP endorphin
ET endothelin
FMRF femarfarmamide
FSH follitropin (follicle stimulating hormone)
GABA γ-aminobutyric acid
GAL galanine
GAP gonadotropin releasing hormone associated peptide
GG glucagon
GH growth hormone (also known as somatotropin)
GHRH growth hormone releasing hormone (also known as somatoliberin)
GIP gastric inhibitory peptide
GLI glicentin
GLP glucagon-like peptide
GMP guanosine 5′-monophosphate
GnRH gonadoliberin (gonadotropin releasing hormone)
GnTH gonadotropin (gonadotropic hormone)
GRP gastrin releasing peptide
GT gastrin
IGF insulin-like growth factor
IGFBP insulin-like growth factor binding protein
IHB inhibin
IL interleukin
KD kallidin
KG kininogen
Lac lactic acid
LH lutropin (luteinizing hormone)
LHRH lutropin releasing hormone (luteinizing hormone releasing hormone)
LPH lipotropin (lipolytic hormone)
MG menopausal gonadotropin (urogonadotropin)
MIF melanostatin (melanotropin inhibiting factor)
MIS muellerian inhibiting substance
MPF melanotropin potentiating factor
MRF melanoliberin (melanotropin releasing factor)
MSH melanotropin (melanocytes stimulating hormone)
MT motilin
NK neurokinin
NMU neuromedin U
NPK neuropeptide K
NPY neuropeptide Y
NT neurotensin
OP osteopoetin
OT oxytocin
PACAP pituitary adenylate cyclase activating polypeptide
PDGF platelet derived growth factor
PG prostaglandin
PH parotid hormone
PHI peptide with N-terminal histidine and C-terminal isoleucinamide
PHM peptide with N-terminal histidine and C-terminal methioninamide
PIF prolactin inhibiting factor
PL placental lactogen
PLA_2 phospholipase A_2
PLC prolactin related cDNA
PLF proliferin

PLP	prolactin-like proteins
PMSG	pregnant mare serum gonadotropin
POMC	proopiomelanocortin
PP	pancreatic peptide
PRC	prolactin related cDNA
PRL	prolactin
PRP	proliferin related protein
PSP	pancreatic spasmolytic peptide
PST	pancreastatin
PTH	parathyroid hormone
PTHrP	parathyroid hormone related protein
PTSI	pancreatic secretory inhibitor
PYY	peptide with N-terminal tyrosine and C-terminal tyrosinamide
RLX	relaxin
SEC	secretin
SL	somatolactin
SM	somatomedin
SP	substance P
SRH	somatoliberin (somatotropin releasing hormone; growth hormone releasing hormone, GHRH)
SRIH	somatostatin (somatotropin release inhibiting hormone)
STH	somatotropin (somatotropic hormone; growth hormone, GH)
T_3	triiodothyronine
T_4	thyroxine
TGF	transforming growth factor
TNF	tumor necrosis factor
TRH	thyroliberin (thyrotropin releasing hormone)
TSH	thyrotropin (thyroid stimulating hormone)
VIP	vasoactive intestinal peptide
VP	vasopressin
VT	vasotocin
vWF	von Willebrand factor

2. Gonadoliberin, Thyroliberin, Gonadotropins, Thyrotropin, Inhibin, and Related Hormones

Thyroliberin (thyrotropin releasing hormone, TRH), and gonadoliberin (gonadotropin releasing hormone, GnRH) are formed in the hypothalamus and release thyrotropin (thyroid stimulating hormone, TSH) and the gonadotropins (follitropin, follicle stimulating hormone, FSH) or lutropin (luteinizing hormone, LH) respectively from the anterior pituitary gland (adenohypophysis).

TSH stimulates the formation of the thyroid hormones, triiodothyronine (T_3) and thyroxine (T_4) in the thyroid gland; T_3 and T_4, in turn inhibit the release of TSH but stimulate the release of TRH (Fig. 1).

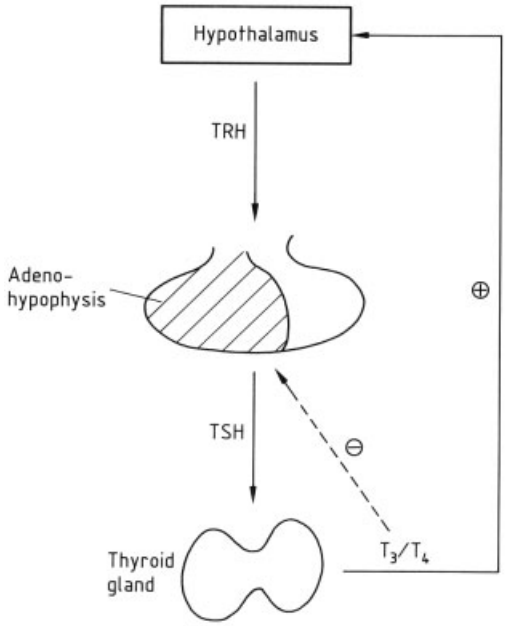

Figure 1. Control of thyroid hormone release

The gonadotropins LH and FSH stimulate the synthesis of the steroid sex hormones in the gonads (Fig. 2). These steroid hormones, in turn, inhibit the release of gonadotropins, but stimulate the release of GnRH. In the gonads, FSH stimulates the release of the inhibins (IHB), which exert a negative feedback to inhibit further secretion of FSH from the hypophysis. The hormone follistatin formed in the gonads also inhibits the release of FSH. The activins (homo- and heterodimers of the IHB β-chains), like the transforming growth factor β (TGF-β), stimulate the formation of FSH.

The tripeptide TRH and the decapeptide GnRH have structural similarities:

GnRH:	∗**Q**	**H**	W	S	Y	G	L	R	P	Ga
TRH:	∗**Q**	**H**	/	/	/	/	/	/	**P**a	

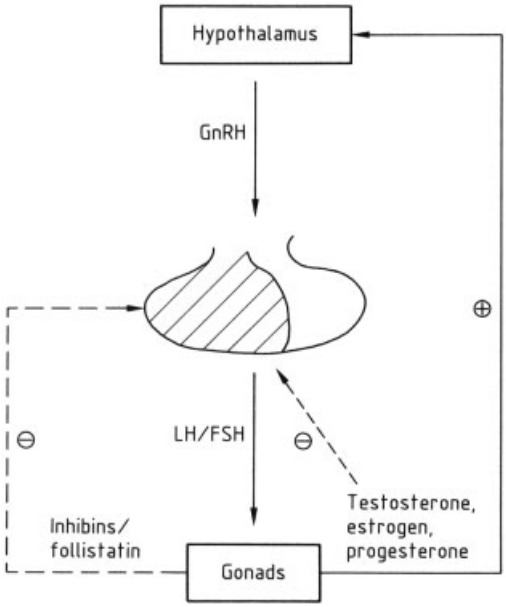

Figure 2. Control of steroid sex hormone release

The gonadotropins (LH and FSH) and TSH also belong to one family. They consist of two noncovalently bound glycoproteins termed the α- and the β-chain; the α-chains of TSH and gonadotropins are identical.

TGF-β, the inhibins, the muellerian inhibiting substance, and the bone morphogenic proteins are structurally related. The inhibins consist of two glycoprotein chains that are linked together by disulfide bridges; their α-chains are identical. The activins are homo- and heterodimers of the inhibin β-chains. TGF-β is a homodimeric compound.

Follistatin is a single-chain glycoprotein which has no structural similarity to the TGF-β/ inhibin family.

2.1. Gonadoliberin

Occurrence. The decapeptide gonadoliberin [*9034-38-2*], [*9034-40-6*], also known as gonadotropin releasing hormone (GnRH), M_r 1182, is formed as a precursor peptide consisting of 92 amino acids (prepro-GnRH) [9]. The C-terminal peptide of prepro-GnRH (gonadotropin releasing hormone associated peptide, GAP) consists of 56 amino acids and inhibits prolactin synthesis [10]. GnRH is found not only in the hypothalamus, but also in the brain, heart, liver, pancreas, kidneys, adrenals, small intestine, gonads, and in the milk of lactating rats [11].

No differences are observed in the GnRHs of different mammalian species. Two GnRHs (g-GnRH I and II) are found in chicken, fish [12], and amphibians [13]. α-Factors that are similar to GnRH have been isolated from microorganisms (e.g., brewer's yeast) and cause cell aggregation and the exchange of chromosomes.

h-,p-,o-GnRH	∗**Q**	**H**	**W**	**S**	Y	**G**	**L**	R	**P**	**Ga**			
g-GnRH I	∗**Q**	**H**	**W**	**S**	Y	**G**	**L**	Q	**P**	**Ga**			
g-GnRH II	∗**Q**	**H**	**W**	**S**	H	**G**	W	Y	**P**	**Ga**			
α-Factor α1	W	**H**	**W**	/	L	Q	**L**	K	**P**	**G**	Q	P	M	Y

Release [14]. The synthesis of GnRH from its precursor requires a peptidylglycine α-amidating monooxygenase that is dependent on Cu^{2+} and ascorbic acid [15]. GnRH is secreted from the hypothalamus in short pulses at intervals of 60 – 120 min and reaches the anterior pituitary via the hypophyseal portal system where it induces the release of the gonadotropins LH and FSH.

Although gonadotropins should be expected to inhibit release of GnRH by a negative feedback mechanism, they stimulate the release of GnRH from GnRH-containing neurons of male rats [16].

The steroid hormones of the gonads (testosterone, estrogen, and progesterone) stimulate the release of GnRH [17]. The post-mortem hypothalamic level of GnRH is lowered in women after the menopause [18]. Vasoactive intestinal peptide (VIP) [19], neuropeptide Y (NPY) [20], thymosin-β_4, and compounds of the arachidonic acid cascade [21] also stimulate the release of GnRH.

Prolactin [14], stress, and the stress hormones (CRH, ACTH, β-endorphin, oxytocin, and glucocorticoids) inhibit the formation of GnRH [14], [22–25]. The effect of steroid hormones depends on the endogenous opioids. Testosterone stimulates the release of GnRH when endogenous opioids are blocked and inhibits GnRH release when the endogenous opioids are stimulated [26].

Receptors [27]. Receptors for GnRH are found not only in the pituitary, but also in the go-

nads, placenta [28], corpus luteum [29], and on the oocytes [30]. In human breast cancer tissue, two classes of [D-Trp^6]GnRH receptors were detected: one with a high affinity and low capacity and the other with a low affinity and a high capacity. These correlate positively with estrogen and progesterone binding [31].

In female rats, physiological doses of GnRH stimulate the formation of GnRH receptors in the hypophysis [32], high doses have the opposite effect [33]. Highly active GnRH agonists reduce the number of GnRH receptors, while the GnRH antagonists have no effect [33]. GnRH agonists are rapidly taken up into the cell by endocytosis, while the antagonists remain on the membrane [34].

Biological Effects. Under physiological conditions (pulsatile application every 60 – 120 min), GnRH stimulates the calcium-dependent hypophyseal synthesis and release [35] of LH and FSH in the pituitary. However, high nonphysiological doses suppress LH and FSH secretion in rats and humans by a down regulation of GnRH receptors, this process does not require calcium [33].

Estrogen stimulates the GnRH-induced secretion of LH when low doses of GnRH are infused, but inhibits secretion at high GnRH doses [36]. Estradiol, insulin [33], adrenaline, noradrenaline [37], and neuropeptide Y (NPY) [38] enhance the LH-releasing effect of GnRH. Progesterone and dihydrotestosterone, as well as stress and glucocorticoids in female rats [39–41] inhibit the GnRH-induced release of LH.

GnRH directly affects steroid synthesis in the gonads of experimental animals. Small doses or short treatment have a stimulating [42] and high doses or long-term treatment an inhibitory effect [43].

In chondrasarcoma and other tumors [44], high doses of GnRH agonists inhibit tumor growth. Similar to physical castration, chemical castration with high doses of GnRH agonists regenerates the thymus in old male rats [45].

Structure – Activity Relationships. [46], [47]. *Gonadoliberin Agonists.* In mammals substitution of His^5 for Tyr^5 and Trp^7 for Leu^7 gives peptides with full GnRH activity: [His^5]-GnRH is twice as effective as GnRH in releasing FSH [48].

The activity of GnRH is increased 100 – 200-fold by substituting an ethylamine residue for Gly^{10}-amide and by replacing Gly^6 by D-amino acids. It is important that the D-amino acid should contain an α-CH and a β-CH_2 group at position 6. Still higher activities can be achieved with synthetic aromatic D-amino acids, e.g., D-3-(2,4,6-(Me)$_3$Phe) [49]. A combination of the hydrophilic AzaGly-amide in position 10 with hydrophobic aromatic amino acids led to [D-Nal^6, AzaGly^{10}]GnRH, one of the most active compounds now available (230 times more active than GnRH) [50]. Hydrophilic D-Ser-*O*-glycosides [51] or β-Asp α-esters [52] in position 6 also give very active compounds.

Gonadoliberin Antagonists. Modification at positions 1, 2, 3, 6, and 10 lead to GnRH antagonists. Highly active compounds are:

1) Ac-[D-Nal^1, D-4-ClPhe^2, D-Trp^3, D-Har(Et)$^6_2$, D-Ala^{10}] GnRH (detirelix) [53], [54]
2) Ac-[D-Nal^1, D-4-ClPhe^2, D-Pal^3, D-Har(Et)$^6_2$, Har(Et)$^8_2$, D-Ala^{10}]GnRH (RS-26306) [47]
3) Ac-[D-Nal^1, D-4-ClPhe^2, D-Trp^3, D-Ser(Rha)^6, AzaGly^{10}] GnRH (HOE 013) [55]
4) Ac-[D-Nal^1, D-4-ClPhe^2, D-Pal^3,Arg^5, D-4-methoxybenzoyl-Abu^6, D-Ala^{10}]GnRH (ORF 21 243 = "Nal-Glu") [56], [57]
5) Ac-[D-Nal^1, D-4-ClPhe^2, D-Pal^3,Lys(Pic)^5, D-Lys(Pic)^6, *N*-isopropyl-Lys^8, D-Ala^{10}]GnRH (ORF 23 541, "Nal-Lys" = antide) [58]
6) Ac-[D-Nal^1, D-4-ClPhe^2, D-Pal^3-D-Cit^6, D-Ala^{10}]GnRH (SB-75, D 20 453, Cetrorelix) [59]

A problem associated with GnRH antagonists is the anaphylactic, histamine-releasing effect, mainly caused by the additional basic charges at position 6 and by the strongly lipophilic N-terminal amino acids [60]. This effect is reduced if the arginine in position 6 is moved to position 5 (e.g., compound 4), D-Trp^3 is replaced by D-Pal^3 (e.g., 2, 4 – 6) and the arginine in position 6 or 8 is modified (e.g., 2) or suitably substituted (e.g., 3, 4, 6) [59], [61–64].

These antagonists protect germ cells from X-ray damage by reversibly blocking the GnRH receptors [65].

Uses [47]. Cryptorchism is treated by the intranasal application of GnRH (gonadorelin; Kryptocur, Hoechst) [66].

Delayed puberty, hypogonadism, anorexia nervosa, Kallmann syndrome, and secondary amenorrhea can be treated with a pulsatile GnRH infusion (Cyclomat, Ferring Arzneimittel). GnRH is also used in the differential diagnosis of fertility disturbances.

In anorexia nervosa, the parenteral application of 5 µg of triptorelin every second day increased the release of LH and initiated the menstrual cycle [67]. A case of acute intermittent porphyria could be controlled by the subcutaneous administration of 5 µg/day of histrelin [68].

Analogues that have a longer and stronger effect are used predominantly in high doses to suppress the gonads. The following GnRH analogues are either on the market or being clinically tested.

1) [*tert*-Butyl-D-Ser^6]GnRH-(1 – 9)nonapeptide-ethylamide (buserelin [*57982-77-1*])
2) [*tert*-Butyl-D-Ser^6,AzaGly^{10}]GnRH (goserelin [*65807-02-5*]) [69]
3) [D-Trp^6]GnRH (triptorelin [*57773-63-4*]) [70]
4) [D-Trp^6]GnRH-(1 – 9)nonapeptide-ethylamide (deslorelin)
5) [D-Leu^6]GnRH-(1 – 9)nonapeptide-ethylamide (leuprolid acetate, leuprorelin [*74381-53-6*]) [71]
6) [D-Trp^6,*N*-MeLeu^7]GnRH-(1 – 9)nonapeptide-ethylamide (lutrelin) [72])
7) [D-Nal^6]GnRH (nafarelin [*76932-56-4*]) [73]
8) [D-His(imBzl)^6]GnRH-(1 – 9)nonapeptide-ethylamide (histrelin) [74], [75])

GnRH agonists can be applied parenterally, nasally, or as a biodegradable implant. They are being clinically tested or applied for the inhibition of testosterone- and estrogen-dependent tumors (carcinoma of the prostate and the breast), in endometriosis, as contraceptives [76], in pubertas precox [77], and for in vitro fertilization. Other indications are regression of benevolent uterus myoma and leiomyoma [78], hirsutism [79], premenstrual intermittent porphyria (disturbance of heme synthesis) [80], and normal prostate hyperplasia [81].

In women, suppressive therapy with GnRH agonists lowers the mineral content of the bones [82].

2.2. Thyroliberin [2]

Occurrence. Thyroliberin [*24305-27-9*], also known as thyrotropin releasing hormone or factor (TRH or TRF), M_r 362, is found primarily in the brain, gastrointestinal tract, pancreas, and prostate. It is formed as prepro-TRH; r-prepro-TRH consists of 225 amino acids and contains 5 copies of the sequence:

- K R Q H P G K R -

which contains the pro-TRH sequence QHPG. Prepro-TRH is cleaved by endopeptidases to release pro-TRH. The N-terminal glutamine of pro-TRH is cyclized to pyroglutamic acid either spontaneously or enzymatically and the C-terminal glycine is converted to the amide function with a peptidylglycine-α-amidating monooxygenase. The peptides located between the TRH sequences are called the prepro-TRH connecting peptides.

The anorexigenic peptide, pGlu-His-Gly-OH, isolated from the urine of women suffering from anorexia nervosa, and a metabolite of TRH, cyclo-(His-Pro), have a central appetite-inhibiting effect [83]. Another structurally similar peptide, pGlu-Pro-amide, was found in rabbit prostate and human sperm [84].

Release. TRH is released centrally by electric shocks [85], 6-hydroxydopamine [86], dopamine [87], cold [88], and thyroid hormones [89]. It is inhibited by opioid peptides [90]. In the periphery, TRH release is increased by histamine [91], testosterone [92], and serotonin [93], and decreased by T_3 [94], carbachol [93], streptozotocin [95], and dopamine [96].

An enzyme (thyroliberinase, pyroglutamyl-peptidase II, E.C.3.4.19) that specifically cleaves the pyroglutamyl – histidine bond of TRH is found in brain homogenates [97]. It deactivates TRH and thus decreases TSH and PRL synthesis.

Receptors. The receptors responsible for TSH secretion are down regulated by chronic pulsatile TRH application, the receptors for PRL secretion are not as sensitive [98].

Thyroidectomy [99] and thyroxine treatment [100] increase the number of TRH binding sites in the rat pituitary. Both κ- and δ-opioids inhibit the binding of radioactively labeled [N^{τ}-MeHis^2]TRH to brain membranes, but μ-opioids are inactive [101]. The blood pressure of spontaneously hypertensive rats correlates positively with the number of TRH receptors in the brain, which could be responsible for the pathophysiology of high blood pressure [102].

Chlorodiazepoxide is a competitive antagonist of the TRH receptor in GH3 cells [103] and the rat amygdala [104]. Other TRH receptor antagonists are diazepam, midazolam, and (−)-4-methylmidazolam [105].

Biological Effects [106].

Biochemical and metabolic functions comprise the stimulation of 1,2-diacylglycerol and inositol-1,4,5-triphosphate levels, the increase of intracellular calcium, and the decrease of serum calcium.

Endocrine functions. TRH secreted in the hypothalamus stimulates the release of thyrotropin (TSH) and prolactin (PRL) in the adenohypophysis of the pituitary. The TRH-stimulated TSH secretion can be inhibited by somatostatin, dopamine [107], and calcitonin [108]. The TRH-stimulated release of PRL is enhanced by estrogen and inhibited by testosterone [109] and dopamine. TRH increases the release of T_3, T_4, and thyroglobulins from the thyroid gland either via TSH [110] or directly [111]. In the pancreas it inhibits the release of glucagon.

Vascular effects are species specific [112]. Vasodilatation occurs via activation of the vagus and vascular contraction occurs via activation of the sympathetic nervous system [113]. Blood pressure is increased [114].

Effects on the nervous system [115] include the potentiation of K^+-stimulated acetylcholine release [116], reversal of natural or narcotic sedation, anticonvulsant activity, inhibition of cortical neuron activity (antinociceptive action) [117], muscle tone improvement, as well as increased blood pressure, heart rate, cerebral blood flow, renal renin release [118], and catecholamine release from the adrenal medulla.

TRH induces hypo- or hyperthermia (depending on species) and reverses induced hypothermia. It also increases the respiratory rate.

Gastrointestinal effects comprise decreased food and water intake, increased gastric secretion (gastric ulcer) and gastric emptying, increased or decreased motility (depending on species), vagus-mediated contraction of the stomach [119], and increased pancreatic exocrine secretion.

Structure – Activity Relationships. Only four synthetic analogues are more effective than TRH in releasing TSH. [2-(L-β-(pyrazolyl-1)Ala]TRH and [1-(3-carboxy-5-oxothiamorpholine)]TRH are 1.5 times more active than TRH. [2-(N^{τ}-MeHis]TRH and the corresponding *N*-amylamide [120] have eight times and approximately ten times the potency of TRH, respectively.

Many TRH analogues exhibit the CNS activities of TRH, without releasing TSH to the same extent [121–130].

1) L-Pyro-2-aminoadipyl-His-thiaPro-amide (MK 771)
2) 6-Methyl-5-oxothiamorpholinyl-3-carbonyl-His-Pro-amide (montirelin, CG 3703) [122]
3) [(*S*)-4-Oxo-2-azetidinyl)-carbonyl-His-Pro-amide (YM-14673, azetirelin dihydrate) [124]
4) γ-Butyrolactone-γ-carbonyl-His-Pro-amide (DN-1417) [125]

MK-771 and CG-3703 are stable to the TRH-specific pyroglutamylpeptidase II, but DN-1417 is degraded faster than TRH. All three are competitive inhibitors of pyroglutamylpeptidase II [130].

Uses. TRH (generic name Protirelin) is used as a diagnostic reagent for thyroid and hypophyseal functions (Relefact TRH, Hoechst; Antepan, Henning; Thyroliberin/TRH, Merck). TRH preparations for parenteral (0.2 – 0.4 mg), oral (40 mg), and intranasal (1 mg, Relefact TRH nasal) application are available.

Possible therapeutic applications are listed below [106]:

1) Motor neuron disease, amyotrophic lateral sclerosis (ALS)
2) Spinal injury
3) Cerebral ischaemia (stroke)
4) Circulatory shock
5) CNS trauma
6) Alzheimer's disease, senile dementia
7) Down's syndrome
8) Schizophrenia
9) Depression
10) Antinociception/analgesia (with morphine)
11) Respiration
12) Reversal of anesthesia and narcotic overdose
13) Spinocerebellar degeneration
14) Epilepsy

2.3. Gonadotropins [131]

Occurrence. The gonadotropins (gonadotropic hormones, GnTH) are glycoproteins and comprise lutropin [*9002-67-9*] (luteinizing hormone, LH), M_r (protein part) 23 113 (h-LH); follitropin [*9002-68-0*] (follicle stimulating hormone, FSH), M_r (protein part) 22 395 (h-FSH); and chorionic gonadotropin [*9002-61-3*] (CG), M_r (protein part) 25 726 (h-CG). LH and FSH are formed in the adenohypophysis and CG is produced in the chorion. They consist of two (α and β) protein chains, each chain contains disulfide bridges. Disulfide bridges between the two chains do not exist. The α-chains of all three hormones are identical but species dependent (89 amino acids in h-LH and h-FSH, 92 amino acids in h-CG) [132]. The β-chains (145 amino acids in h-CG, 121 in h-LH, and 111 in h-FSH) are responsible for activity [133]. The chains are synthesized separately as procompounds.

The structures of GnTHs from humans, horse, sheep, cow, and pig as well as LH sequences from rat, rabbit, salmon, and pike eel have been elucidated [134].

Thyrotropin (TSH) is structurally related to the gonadotropins because it has an identical α-chain.

Release. The action of pulsatile GnRH on the gonadotropic cells of the adenohypophysis controls the release of the gonadotropins, LH, and FSH (see Section 2.1). FSH is preferentially released at a slow pulse frequency and LH is preferentially released at a faster GnRH frequency. Centrally applied GnRH lowers LH secretion [135]. Endothelin is as effective as GnRH in releasing the gonadotropins from anterior pituitary cells but this process does not involve GnRH receptors [136].

Low doses of estrogen and testosterone stimulate the release of gonadotropins by stimulating GnRH formation [137–139]. However, the long-term administration of higher doses directly inhibits gonadotropin release (negative feedback) [137], [140]. Calcitriol also potentiates GnRH-induced gonadotropin release [141]. Phospholipase A_2 [142], prostaglandin E_2 [143], and diacylglycerol [144] stimulate release, whereas cyclooxygenase inhibitors inhibit release [145].

The frequency of LH secretion in women increases from the early follicular phase to the mid and the late follicular phases of the menstrual cycle. The LH amplitude decreases in the midfollicular phase (negative feedback by estrogen). The LH pulse frequency and the amplitude sink during the luteal phase (effect of progesterone). Levels of FSH correlates with those of LH [146]. The inhibitory influence of the steroid hormones is lacking in the menopause and the level of LH and FSH rises. Treatment with progesterone [147] or ethinylestradiol [148] lowers the increased plasma gonadotropin levels. Testosterone is responsible for this effect in men [149]. FSH release is inhibited by the inhibins and stimulated by the activins and TGF-β.

Gonadotropin release is lowered by several factors, usually via the inhibition of GnRH release: deprivation of food [150], stress [151], morphine [152], [153], endogenous opioids [154], marihuana [155], glucocorticoids [156], dopamine agonists [157], alcohol [158], and lithium salts [159]. Central noradrenergic mechanisms inhibit the release of LH via the activation of β-adrenergic receptors [160].

Receptors. The h-LH/CG [161] and r-LH/CG [162] receptors have been characterized. The former consists of 696 amino acids with a signal peptide containing 27 amino acids.

FSH stimulates the formation of LH receptors in rat granulosa cells. Estrogen or testosterone is important for the FSH-stimulated synthesis of LH receptors [163]. Androgens [164] and interleukin-1 (IL-1) [165] inhibit the FSH-stimulated synthesis of LH receptors.

Biological Effects. In the ovaries, LH stimulates the synthesis of cAMP and inositol triphosphate [166] and reduces levels of cGMP by inhibiting guanylate cyclase [167].

The activities of LH and FSH complement each other. In the male organism, spermatogenesis can occur only if both hormones are present. FSH is responsible for sperm development, and LH stimulates the testosterone secretion in the interstitial cells, a process necessary for sperm maturation.

Although FSH activates follicular growth in women, ovulation and formation of the corpus luteum do not occur. Ovulation, corpus luteum formation, and estrogen secretion require the simultaneous presence of LH.

Structure – Activity Relationships. The intercystine loops in positions (38 – 57) of the β-chains of h-CG and h-LH slightly inhibit the binding of ^{125}I-h-CG to rat ovarian membranes and stimulate testosterone synthesis [168].

The tetrapeptides h-FSH-β-(34 – 37) and h-FSH-β-(49 –52) inhibit the receptor binding of ^{125}I-h-FSH. The peptide h-FSH-(33 – 53) binds strongly and stimulates estradiol synthesis [169]. The peptides h-FSH-β-(81 – 95)amide and h-FSH-β-(31 – 45)amide bind tightly to the FSH receptor, the former stimulates estradiol synthesis in Sertoli cell cultures [170]. Partial sequences of the α-chain also displace h-CG from the receptor [171].

Antibodies against h-FSH-β-(76 – 118) or h-FSH-β-(1 – 33) bind to h-FSH, LH, and TSH [172]. A monoclonal antibody directed against h-CG binds in the vicinity of Lys^{45} of the α-chain and Asp^{112} of the β-chain. A peptide in which these two regions are combined stimulates the production of a specific antibody against h-CG and could be used for the development of a synthetic vaccine for castration [173].

The sugar residues on the peptide chains are of great importance for biological activity. Although deglycosylated h-CG binds more strongly to the receptors than intact h-CG, it does not activate cAMP synthesis in vitro and it inhibits the activity of glycosylated h-CG. However, deglycosylated or desialylated h-CH stimulates the synthesis of testosterone in vivo (monkeys) [174].

Uses [175]. h-CG or h-MG (menopausal gonadotropin, urogonadotropin, a mixture of h-CG and h-FSH), and h-FSH (Metrodin, Serono) are used predominantly in human medicine. h-CG (Choragon, Ferring; Predalon, Organon; Pregnesin, Serono; Primogonyl, Schering) and h-MG (Humegon, Organon; Pergonal, Serono) are used to treat gonadal disorders, e.g., primary and secondary amenorrhea, sterility, spermatogenesis, delayed puberty, cryptorchism, oligospermia, and menorrhagia. The main indication is polycystic ovary disease and in vitro fertilization [175]. Pulsatile therapy with Metrodin [176] or Pergonal [177] has advantages over daily injections.

2.4. Thyrotropin [178]

Occurrence. Thyrotropin [*9002-71-5*], also known as thyroid stimulating hormone (TSH), M_r (protein part) 22 813 (h-TSH), is formed in the adenohypophysis.

Release [179]. The release of TSH from the hypophysis is stimulated by TRH. High doses of the thyroid hormones, triiodothyronine (T_3) and thyroxine (T_4) inhibit TSH synthesis in the hypophysis by a negative feedback mechanism. Small doses of T_3 and T_4 stimulate the release of TSH via the activation of TRH. TSH release is also stimulated by calcitriol [180], prostaglandins [181], and Arg-vasopressin [182].

TSH release is inhibited by neuromedin B [183], glucocorticoids [184], estrogen, testosterone, α-endorphin [185], and somatostatin [186].

Receptors. The structures of the TSH receptor in the dog [187] and humans [188] have been established via the elucidation of the cDNA. The human receptor has 744 amino acids. There are sequence similarities between these receptors and the LH/h-CG receptors.

Biological Effects. TSH stimulates the synthesis of thyroglobulin and the synthesis and secretion of T_3 and T_4 in the thyroid gland through cAMP [189]. It enhances DNA synthesis and the proliferation of epithelial follicular cells from human thyroid tissue [190], a process that can be inhibited by somatostatin [191]. Lithium salts

inhibit the TSH-stimulated synthesis of T_3 and T_4 and produce hypothyreosis in women [192].

Structure – Activity Relationships. The sequences (26 – 46), (31 – 45), and (21 – 35) of the α-chain [193] and (101 – 112), (71 – 85), (31 – 45), (41 – 55), and (1 – 5) of the h-TSH-β-chain [194] inhibit the binding of TSH to human thyroid membrane preparations.

2.5. Inhibins, Activins, and Related Hormones

The inhibins, activins, and the transforming growth factor-β (TGF-β) consist of two protein chains linked by disulfide bridges. The inhibins (IHB) are heterodimers consisting of an α- and a β-chain. The α-chains of IHB-A and IHB-B are identical and there are structural similarities between the α- and β-chains. Activin A and TGF-β are homodimeric compounds consisting of two identical chains of the β-type. Related proteins include muellerian inhibiting substance, bone morphogenetic proteins, and osteogenic proteins. They are high molecular mass glycoproteins consisting of 400 – 500 amino acids. Their C-terminal regions are very similar to those of TGF-β and the IHB β-chains.

2.5.1. Inhibins [195], [196]

Occurence. Inhibins occur mainly in the gonads. Other sites of synthesis are the placenta (in humans) [197] and the brain [198]. The structure of IHB-A from humans [M_r (protein part) 27 667], pigs, cows, and rats and that of IHB-B from humans [M_r (protein part) 27 508], pigs, and rats have been elucidated. The α-chain (134 amino acids in humans) and the β-chain (h-IHB-βA 115 amino acids, h-IHB-βB 116 amino acids) are formed from different messenger RNAs in the form of prepropeptides.

Release. Inhibin is released primarily by FSH, presumably via an increase in cAMP [199]. h-CG, LH, testosterone [200], [201], somatomedin C [202] and, in the case of hypophysectomized female rats, estradiol stimulate in vitro inhibin release. GnRH agonists inhibit both basal as well as PMSG-stimulated IHB release, and a GnRH antagonist potentiates the stimulating effect of PMSG on IHB release [203].

The plasma levels of both estradiol and inhibin are low during the early and mid follicular phase of menstruation and the FSH level is slightly elevated. In the late follicular phase, inhibin increases parallel to estradiol, while the levels of FSH and LH decrease and increase, respectively. Inhibin reaches a first maximum when LH and FSH exhibit a maximum [204]. The activity of IHB is reduced in the early luteal phase of the menstrual cycle (parallels that of estrogen) [204]. In the mid luteal phase, inhibin increases parallel to progesterone, while plasma levels of FSH and LH fall.

The plasma inhibin level in men suffering from idiopathic hypogonadotropic hypogonadism is low [205]. The level falls by 70 % during treatment with h-CG, which is associated with a complete loss of endogenous FSH [206].

Female patients with ovarian granulosa cell tumors [207], polycystic ovarian diseases [208], and duodenal ulcers [209] have high serum inhibin levels.

Biological Effects. IHB-A and IHB-B inhibit the synthesis and release of FSH in the hypophysis by specifically inhibiting messenger RNA synthesis [210]. Inhibin suppresses the GnRH-induced formation of GnRH receptors in cell cultures of rat pituitary [211]. In vitro, it inhibits progesterone synthesis [212] and aromatase activity [213] in the gonads.

2.5.2. Activins [195]

Occurrence. Activins (ATV) are found in the ovarian follicular fluid and in leukemic cells. h-ATV-A [*104625-48-1*], M_r (protein part) 25 934, is a homodimer of the β-chain of IHB-A consisting of 115 amino acids each, [214], [215] and h-ATV-A-B [*114949-23-4*], M_r (protein part) 25 775 is a heterodimer consisting of the β-chains of IHB-A and IHB-B with 115 and 116 amino acids, respectively [216].

Follistatin [*117628-82-7*] is an activin-binding protein that has been isolated from rat ovaries [217].

Biological Effects. The activins stimulate the release of FSH from the hypophysis. Unlike GnRH, which releases FSH and LH very rapidly, the activins require >4 h to bring about the secretion of FSH. Activins also potentiate the FSH-induced aromatase activity which is important for estrogen synthesis [216], but inhibit progesterone synthesis [218].

Activin produced by leukemic cells also acts like erythropoietin (stimulates blood formation) [219].

2.5.3. Transforming Growth Factor-β [219], [220]

Occurrence. Transforming growth factor-β (TGF-β) [*98726-64-8*], M_r (protein part) 25 572 (h-TGF-β), consists of two identical peptide chains of 112 amino acids each, linked by disulfide bridges [221]. It is stored in platelets in the form of a high molecular mass, inactive complex [222]. TGF-β occurs widely; messenger RNA coding for TGF-β occurs in the liver cells of adult animals, but also in all cell lines studied so far [221].

Release. TGF-β is activated and released from its inactive storage complex in platelets by thrombin [223], acid, or urea (i.e., by blood clotting) [222]. Heparin cleaves the complex between TGF-β and α_2-macroglobulin formed in the blood plasma and, thus, potentiates the effect of TGF-β [224]. Bone-degrading factors (e.g., parathyroid hormone and calcitriol), release TGF-β from the skulls of rat fetuses and of newborn rats. Calcitonin decreases TGF-β activity [225]. Activated osteoclasts release TGF-β from the TGF-β complex formed by bone organ cultures [226]. Osteoclasts are cells responsible for bone degradation.

Receptors. Three different types [219] of receptor for TGF-β have been found in practically all cells that have been studied, they appear to be insensitive to down regulation. In osteoblasts (bone-forming cells) the binding of TGF-β to the receptors is reduced by parathyroid hormone [227].

Biological Effects. h-TGF-β stimulates both the secretion of FSH from rat anterior pituitary cells [228] and the aromatase activity in ovarian granulosa cells [213], [218], [229]. These effects are similar to those produced by the activins, and are inhibited by the epidermal growth factor-like tumor growth factor-α (TGF-α) [229].

In adrenal cells TGF-β inhibits cholesterol synthesis and thus basal and ACTH-stimulated steroidogenesis [230]. In rat pancreas cells, TGF-β_1 and TGF-β_2 stimulate insulin release, this process is glucose-dependent [231].

TGF-β is a general regulator of cell growth. Both TGF-α and TGF-β stimulate the growth of mesenchymal cell cultures via the induction of platelet derived growth factor (PDGF) [232]. However, TGF-β inhibits the growth of epithelial cells [233] and many different human cancer cell cultures [221], [234], [235]. It potentiates the PDGF-stimulated growth of fibroblasts, but inhibits growth of these cells stimulated by the epidermal growth factor (EGF) [235]. TGF-β inhibits or stimulates cell differentiation; it inhibits the differentiation of B-lymphocytes and muscle myoblasts [236], but stimulates the conversion of 3T3 fibroblasts to adipocytes.

Differentiation and proliferation of epithelial cells are enhanced by TGF-β and these cells exhibit increased production of fibronectin, collagen, and other cell-adhesive proteins [237]. Formation of receptors for these cell-binding proteins is also enhanced by TGF-β [219]. In smooth muscle cells from the porcine aorta, TGF-β stimulates the production of elastin [238]. In the extracellular matrix, TGF-β increases the expression of chondroitin and dermatan sulfate proteoglycan [239], it also inhibits protease activity induced by interleukin 1 (IL-1) and proteoglycan degradation [240].

TGF-β also acts as a regulator of the immune system. For example, it inhibits the basal and the TNF-α-stimulated (tumor necrosis factor α) adhesion of neutrophil blood cells to endothelial cells [241], inactivates macrophages [242], and inhibits the IL-1-stimulated degradation of cartilage [243].

TGF-β appears to be of importance for early bone development and bone repair. In rat fetal skulls it increases the synthesis of DNA and collagen [244], stimulates the formation of osteopontin in osteosarcoma cells [245], and inhibits the production of osteoclast-like cells in cell cultures [246]. The effect of TGF-β is inhibited by parathyroid hormone in osteoblasts [227].

Uses. TGF-β stimulates collagen matrix contraction by fibroblasts, a process that should accelerate the healing of wounds [247].

2.5.4. Bone Morphogenetic Proteins and Osteogenic Proteins

Occurrence. Decalcified bone matrix stimulates the formation of new bone [248]. Bone-growth-inducing substances isolated from decalcified bones include osteogenin [249], osteopoetin (OP) [250], and bone morphogenetic proteins (BMP) [251]. h-BMP-2A, h-BMP-2B, and h-BMP-3 contain 396, 408, and 427 amino acids, respectively [251]. The structures of h-BMP-2A and h-BMP-2B are very similar. The C-terminal ends of all three glycoproteins and the TGF-β/IHB family are also similar. Sequenced fragments of b-osteogenin have almost the same amino acid sequence as BMP-3 [252]. The glycoprotein, h-OP-1 contains 431 amino acids and a signal sequence of 29 amino acids [253].

Release. Vitamin D stimulates BMP synthesis [254]. BMP production decreases with age [255]. Extracts of rat chondrasarcoma and osteosarcoma show no BMP activity, and inhibit the bone-inducing activity of the BMPs [256].

Biological Effects. The BMPs initiate the formation of cartilage, resulting in the production of new bone. Recombinant h-BMP-2A [257] and isolated b-osteogenin [252] induce the formation of cartilage (chondrogenesis) and bone (osteogenesis) in rats. The activity of demineralized bone matrix is reduced in rats deficient in vitamin D [258]. Aluminum salts [259] and vitamin A [260] have an adverse effect on chondrogenesis and osteogenesis induced by demineralized bone matrix. In older rats, PDGF potentiates the cartilage and bone formation induced by demineralized bone matrix [261].

Uses. Gelatin that contains osteogenin corrects diaphyseal defects in sheep [262]. Bone defects in patients with cholesteatomas could be rectified with osteopoetin implants [250].

2.5.5. Muellerian Inhibiting Substance

Occurrence. The muellerian inhibiting substance (MIS) [*80497-65-0*] is a glycoprotein found in large amounts in the Sertoli cells of the testes of newborn calfs. Native MIS is probably a glycosilated disulfide-linked dimer (M_r ca. 140 000). h-Mis, M_r (protein chain) 58 000, consists of 535 amino acids, its C-terminal sequence shows similarities to the TGF-β/inhibin family [263].

Biological Effects. MIS inhibits the development of the muellerian duct during gonadal differentiation in the male embryo. In the female embryo the muellerian duct develops into fallopian tubes, uterus, and vagina. Complete virilization requires the presence of MIS. Preparations of b-MIS inhibit the in vitro and in vivo growth of tumor cells derived from the ovaries and the endometrium [263]. The growth-inhibiting effect of MIS on the muellerian duct can be reversed in vitro by epidermal growth factor [264].

Structure – Activity Relationships. A C-terminal fragment (M_r 25 000) that is similar to TGF-β forms a complex with the N-terminal region of MIS. This complex can be made by limited digestion of MIS with pepsin and has full biological activity [265].

2.6. Follistatin [266]

Occurrence. Follistatins [*117628-82-7*] are single-chain glycoproteins that were originally isolated from ovarian follicular fluid. Two human follistatin precursors containing 344 and 317 amino acids give rise to two follistatins consisting of 315 (M_r ca. 34 800) and 288 amino acids (M_r ca. 31 600), respectively [267].

Follistatin has three homologous domains: amino acid sequences (66 – 135), (139 – 210), and (216 – 287). These domains have a ca. 50 % sequence similarity both among themselves and with the human pancreatic secretory inhibitor (h-PTSI). The 309 – 314 region (ISSILE) is identical to a tyrosine kinase domain of the EGF receptor.

Biological Effects. In hypophyseal cell cultures, both follistatins inhibit the release of FSH.

Like the inhibins, the follistatins also inhibit the FSH-stimulated synthesis of estrogen. Follistatin is a binding protein for the activins [268].

3. Parathyroid Hormone and the Calcitonin Family

Parathyroid hormone (PTH) and calcitonin (CT) are secreted from the parathyroid and thyroid glands, respectively. They regulate the plasma calcium level and bone formation. The plasma calcium level is increased by PTH and decreased by CT; PTH stimulates the turnover of bone and CT inhibits it. PTH also stimulates synthesis and release of calcitriol (1α,25-dihydroxyvitamin D_3) in the kidneys. Like high serum calcium levels, this substance inhibits release of PTH and stimulates the synthesis of calcitonin.

The reproductive hormones described in Chapter 2 also affect bone formation. TGF-β not only stimulates the release of FSH, but also initiates chondrogenesis and osteogenesis (i.e., synthesis of cartilage and bone) [269]. Lack of steroidal sex hormones (estrogen, testosterone, and progesterone), which are important for the release of gonadoliberin, leads to osteoporosis. The chaotic release pattern of PTH appears to be dependent on estrogen levels in women [270].

PTH-related peptide, PTHrP, is formed primarily in tumors and has properties similar to those of PTH. A CT-gene related peptide (CGRP) has also been found. Amylin is a peptide produced by the β-cells of the pancreas that has a structural similarity to CGRP.

3.1. Parathyroid Hormone and PTH-Related Protein [271–273]

Occurrence. Parathyroid hormone (PTH) [*9002-64-6*] (IUPAC name parathyrin) is formed in the parathyroid gland but is also found in the brain and the pituitary gland. A PTH-related protein (PTHrP) with 141 amino acids (M_r 16 005, h-PTHrP) is formed in tumors that are accompanied by hypercalcemia [274] and in lectin-stimulated lymphocytes [275].

h-, b-, p-, and r-PTH consist of 84 amino acids (M_r 9425, h-PTH); g-PTH has 88 amino acids [276]. The precursor of PTH is prepro-PTH, a peptide containing 31 additional N-terminal amino acids [277].

Release of PTH. Low plasma calcium increases the synthesis and release of PTH, while high plasma calcium, an infusion of calcium, or a calcium diet [278] lowers the level of PTH. A circadian rhythm is observed with two daily PTH maxima followed by calcium peaks with a delay of 2 h [279]. PTH secretion is also stimulated by insulin-induced hypoglycemia [280], a phosphate-rich diet [281], calcium antagonists [282], a NaCl-rich diet [283], a diet low in vitamin D_3 and calcium [284], lithium [285], stress (induced hypocalcemia) [286], noradrenaline [287], histamine, 17-β-estradiol, and progesterone [288]. In osteoporotic women, the frequency and amplitude of the chaotic pulsatile microsecretion of PTH are greatly reduced after the menopause and estrogen therapy partly regenerates this pulsatile secretion [289].

PTH stimulates the synthesis of calcitriol in the kidneys. Calcitriol inhibits the synthesis and release of PTH [290]. See also → Hormones, Chap. 5.

Magnesium slightly inhibits the release of PTH [291]. Dietary magnesium correlates positively with the bone mineral density. The release of PTH is also inhibited by calcium canal activators [282], cholinergic agonists [292], H_2-receptor blockers [293], and β-receptor blockers [294].

Occurrence of high PTH levels increases with age [295]. Raised levels of PTH are also observed in nephrocalcinosis, osteitis fibrosa cystica, chronic kidney failure [296], liver disease [297], myotonic dystrophy [298], obesity [299], and coronary arterial disease [300].

The release of PTHrP is reduced by glucocorticoids, calcitriol [301], and octreotide [302].

Receptors. PTH – receptor complexes are internalized by bone cells, thus lowering receptor density on the cell surface.

Biological Effects. The physiological role of PTH is to maintain the extracellular concentration of calcium (1.1 – 1.3 mmol/L) and to prevent hypocalcemia. It is also the principal regulator of bone turnover. PTH stimulates intracellular formation of cAMP and the phos-

phorylation of intracellular proteins, it also increases intracellular calcium levels.

Low doses of PTH stimulate bone synthesis and high doses result in bone resorption [303]. Hyperparathyroidism causes osteoporosis, while hypoparathyroidism results in increased bone formation [304].

In cultured, osteoblast-like cells, PTH stimulates DNA synthesis [305], alkaline phosphatase activity, amino acid incorporation [306], the release of mediators (e.g., interleukin 6) [307], which stimulate the bone-resorbing activity of the osteoclasts [308]. In osteoclasts, PTH stimulates carbonic anhydrase [309] and acid phosphatase [310].

In vivo, PTH also potentiates bone formation induced by insulin-like growth factor I [311] and the activity of TGF-β [312]. In the kidney, PTH stimulates the hydroxylation of 25-hydroxyvitamin D_3 to calcitriol which increases enteral absorption of calcium. High plasma calcium levels lower the formation of calcitriol [313] and low calcitriol levels are therefore observed in hyperparathyroid patients with high calcium levels [314]. In the kidneys, PTH causes calcium derived from bone to be returned to the blood and phosphate is excreted in the urine. After prolonged infusion of PTH ($>$ 12 d), calcium and magnesium excretion is increased [315].

PTH lowers blood pressure and causes vasodilatation [316]; cAMP plays a role in the lowering of blood pressure [317]. PTH inhibits the entry of calcium into the cell which may also play a part in its vascular action [318]. PTH-(1 – 34) causes a marked reduction of the blood pressure of spontaneously hypertensive rats [319]. However, about 40 % of patients suffering from hyperparathyroidism are hypertensive because PTH potentiates the hypertensive effect of hypercalcemia [320].

At high calcium concentrations PTH (unlike CT) enhances the arginine-induced secretion of glucagon, which can lead to hyperglycemia. Low PTH doses stimulate the glucose- and calcium-dependent release of insulin in isolated rat pancreas cells, high doses have an inhibitory effect.

Parenteral administration of PTH promotes the development of gastric ulcers, which can be prevented by CT [321]. In contrast, intracerebroventricular application of r-PTH-(1 – 34) inhibits the secretion of gastric acid and the development of gastric ulcers in rats [322].

The properties of PTHrP are similar to those of PTH [323]. It increases the level of calcium and stimulates the formation of calcitriol. The calcium-mobilizing activity of PTHrP-(1 – 34) amide is comparable with that of b-PTH-(1 – 34) [324].

Structure – Activity Relationships. The (1 – 34) N-terminal sequence of PTH contains all the structural requirements for full biological activity. Extension and shortening at the C-terminus of PTH-(1 – 34) results in loss of biological activity. The region (25 – 34) is required for receptor binding, while the first two amino acids at the amino end are of great importance for biological activity. g-PTH-(1 –34), mainly differs structurally from b-PTH between positions 15 and 27 and has only one-tenth the biological activity [325].

Oxidation of Met^8 and Met^{18} in PTH-(1 – 34) abolishes its hypotensive and vasodilating effects without affecting its hypercalcemic effect. Oxidation of PTH-(1 – 34) reduces adenylate cyclase activation in renal blood capillaries, but not in the tubuli [326].

The activity of b-PTH-(1 – 34) is increased by the substitution of Tyr or D-Tyr for Phe^{34} and the replacement of the two Met residues by Nle. The blocking of the terminal carboxyl group by an amide function also enhances biological activity. Thus, [$Nle^{8,18}$,D-Tyr^{34}]b-PTH-(1 – 34)amide is about four times more effective than b-PTH-(1 – 34).

PTH Antagonists. The peptides PTH-(3 – 34) and [$Nle^{8,18}$,Tyr^{34}]b-PTH-(3 – 34)amide are weak in vitro inhibitors of PTH. [Tyr^{34}]b-PTH-(7 – 34)amide inhibits the PTH-induced excretion of phosphate in urine, cAMP formation [327], and the increase in plasma calcium levels [328]. Cleavage of PTH-(1 – 34) with cathepsin D yields PTH-(8 – 34), which also acts as an antagonist in vitro [329]. Subsitution of Gly^{12} by D-Trp leads to antagonists that are 10 – 30 times more effective [330].

Uses. PTH-(1 – 34) has been used in the treatment of osteoporosis. Alternating therapy with calcitriol [331] or sa-calcitonin [332] appears to be especially favorable. PTH-(1 – 34) can

also serve as a diagnostic aid to differentiate between pseudohypoparathyroidism and hypoparathyroidism [333].

3.2. Calcitonin [334]

Occurence. Calcitonin (CT) [*9007-12-9*] is mainly synthesized in the C-cells of the thyroid gland, CT-like material is also found in the brain, hypothalamus, pituitary, lungs, thymus, liver, gastrointestinal tract, adrenals, muscle, parathyroid gland, cerebrospinal fluid, seminal fluid [335], and in breast milk [336]. The calcitonins are peptide amides containing 32 amino acids, M_r 3418 (h-CT). Marked structural differences are observed between species. Calcitonin is formed as prepro-CT which contains 109 amino acids. The structures of h-prepro-CT [337] and of g-prepro-CT [338] were elucidated via the cDNA.

Release of CT. CT is released from the C-cells in response to calcium, strontium, barium, cholinergic peptides (e.g., cholecystokinin and cerulein), secretin, and glucagon. Chronic hypercalcemia has an inhibitory effect on the acute calcium-stimulated release of CT [339]. 17-β-Estradiol and progesterone stimulate the in vitro secretion of CT from the C-cells of the thyroid glands of eight day old rats [340]. Somatostatin and low plasma levels of calcium inhibit the release of CT. Increased intracellular calcium and cAMP appear to be the second messengers for CT release [341].

Men have a higher plasma CT level than women [342]. A raised plasma level of CT is found in cancer of the thyroid gland [343], heroin addicts [344], and urticaria pigmentosa [345]. A reduced plasma CT level is found in hypothyroidism [346], Cushing's syndrome [347], Turner's syndrome [348], postmenopausal women [349], and hypogonadal osteoporotic men [350].

Receptors. Receptors for CT in osteoblast-like cells are increased by calcitriol [351] and decreased by CT (via internalization) [352].

Biological Effects. Low doses of CT stimulate cGMP formation. High doses activate adenylate cyclase (cAMP formation) and inhibit the release of cGMP [353], [354]. In mammals, CT lowers plasma calcium and phosphate levels by inhibiting bone resorption and promoting renal excretion [355].

In women, the ovarian steroid hormones potentiate the hypocalcemic effect of CT [356]. CT exerts antidiuretic and natriuretic effects [357], promoting the excretion of uric acid [358]. It also stimulates 25-hydroxyvitamin D_3-1α-hydroxylase (i.e., calcitriol synthesis in the kidneys) [359].

The mobility of osteoclasts (giant cells in the bone marrow that are responsible for bone resorption) is inhibited by CT [360] via increased uptake of calcium [361] and phosphate [362]. CT stimulates the proliferation of osteoblasts [363] and activates cartilage synthesis [364].

In humans, CT lowers the insulin level [365], reduces the suppressive effect of glucose on glucagon release, and thus increases blood sugar levels. CT inhibits the release of glucagon induced by arginine or hypoglycemia, as well as glucagon-stimulated glycogenolysis. This may explain why CT lowers plasma glucagon levels and the blood sugar concentration in insulin-dependent diabetics [366]. Intracerebroventricular administration of sa-CT increases the glucose-stimulated release of insulin in rats [367].

Fatty acid synthesis in the liver is increased by CT via calcium-dependent activation of fatty acid synthetase [368] and ATP citrate lyase [369].

CT has both stimulating and inhibitory effects on the hypophyseal hormones. It stimulates the release of the stress hormones ACTH [365] and β-endorphin [365] (analgesic effect) and thus release of cortisol from the adrenals [365]. CT inhibits the release of growth hormone [370], prolactin [371], lutropin, follitropin [372], and thyrotropin [373].

Subcutaneous application of CT reduces the stress-induced secretion of gastric acid and the formation of ulcers [374]. In the gastrointestinal tract, CT releases somatostatin, inhibits the release of gastrin [375], and increases motility [376]. Secretion of pancreatic enzymes and the contraction of the gallbladder are also reduced [377]. These effects may be due to action on the CNS because they are also obtained after intracerebroventricular application. Furthermore, central application also causes hypocalcemia in

rats [334], delays the emptying of the stomach [378], and reduces absorption of food and water [379]. Intracerebroventricular application of CT increases blood pressure, heart rate [380], and body temperature [381], but intravenously applied CT has no effect [380].

Structure – Activity Relationships. In mammals sa- and anq-CTs produce a 10 – 30 times stronger hypocalcemic effect than that exerted by the mammalian peptides. The C-terminal prolinamide is especially important for biological activity. Substitution of the C-terminal carboxamido group or elimination of the C-terminal proline reduces activity by $> 99\,\%$. Acylation of the amino group of the N-terminal cysteine or its substitution by hydrogen slightly increases activity. The N-terminus has a ring structure due to disulfide bond formation between Cys^1 and Cys^8. This is important for the activity of h-CT, but not for sa-CT [382]. The disulfide structure or the size of the ring is not essential for the activity of anq-CT [383]. The amino acids Ser^{29} and Thr^{31} are partly responsible for the high activity of sa-CT. In sa-CT, Val^8 can be replaced by Gly without loss of activity. However, this does not apply to Met^8 of h-CT [383].

de-Leu^{16}-sa-CT and de-Phe^{16}-h-CT exhibit only one- tenth of the corresponding CT activity in binding tests [384], whereas Tyr^{22} [383], [385], Leu^{19}, Ser^{13} [386], Gln^{20}, and Thr^{21} [385] of sa-CT can be deleted without loss of activity.

Active sa-CT analogues with a stable α-helical structure or with a low α-helicity exhibit high biological activity [387], [388]. The α-helical structure seems to be important for action on the kidney, but not for that on the brain [389].

Uses. CT can be administered by the subcutaneous, intramuscular, rectal, or intranasal route. Intranasal administration has the fewest side effects [390]. Nasal administration of only 50 IU of sa-CT per day inhibits bone resorption without having a hypoglycemic effect or increasing plasma cAMP levels [391].

sa-CT (Calcitonin L and Karil, Sandoz; Calsynar, Rover), h-CT (Cibacalcin, Ciba-Geigy), p-CT (Calcitonin S, Rover) and, in Japan, [1,7-L-aminosuberic acid]anq- CT (Elcatonin, Carbocalcitonin, and Turbocalcitonin, Toyo, Ioza) are used to treat Paget's disease, hypercalcemia, hyperparathyroidism, osteoporosis [392], osteolysis [393], the Sudeck syndrome, acute pancreatitis, chronic polyarthritis [394], and tumor-osteolysis. CT is also used as a centrally acting analgesic [395].

3.3. Calcitonin Gene Related Peptide [396]

Occurrence. The messenger RNA for calcitonin gene related peptide (CGRP) [*83652-28-2*] is found primarily in the central and peripheral nervous systems [397], [398]. Circulating CGRP is formed primarily by the perivascular and cardiac nerves.

CGRP consists of 37 amino acids, M_r 3790 (h-α-CGRP). Two human and rat CGRPs have been observed: α-CGRP is coded in the α-gene and β-CGRP in the β-gene [399]. The α-gene is also responsible for coding CT.

Release. CGRP is released by capsaicin [400] and glucocorticoids [401]. Testosterone [402] and δ- and μ-opioid receptor agonists [403] inhibit the release of CGRP.

Biological Effects. CGRP is involved in sensory, motor, and autonomous nervous functions. It stimulates the formation of cAMP.

Intravenous injections result in a rapid decrease of blood pressure and an increase in the heart rate. CGRP is probably of physiological importance for the prevention of cerebral ischemia caused by excessive vasoconstriction and for the control of the peripheral resistance [404]. CGRP stimulates proliferation of human endothelial cells and may be important for the formation of new blood vessels, e.g., in the healing of wounds [405]. Noradrenaline release in mice is stimulated by intracerebroventricular application resulting in increased arterial blood pressure and tachycardia. In the lungs, CGRP exerts a higher bronchoconstricting effect than substance P [406]. Intradermal administration of CGRP causes a more pronounced reddening of the skin than substance P [407]. CGRP may potentiate the release of substance P from primary afferent terminals and promote the transmission of nociceptive information induced by

mechanical noxious stimuli [408]. Administration of CGRP to rats inhibits secretion of growth hormone stimulated by a variety of agents [409]. Stress-induced release of prolactin (PRL) is reduced by intraperitoneal or subcutaneous application of CGRP [410].

In the adrenals, CGRP increases the secretion of cortisol and aldosterone and inhibits the release of Met-enkephalin [411].

CGRP appears to be a central mediator of ingestive behavior because central application suppresses food intake. Similar to CT, both the intravenous and the central application of CGRP inhibits gastric acid secretion [412]; this may be a direct effect or/and mediated through the release of somatostatin [413]. CGRP relaxes the intestinal muscle [414] and inhibits the emptying of the stomach in rats [415]. Low doses of α- and β-CGRP have an antisecretory effect on the epithelium of the colon, but high doses stimulate secretion [416].

CGRP inhibits insulin secretion [417] and insulin-stimulated glycogen synthesis [418], but promotes glycogenolysis [418]. CGRP (and/or amylin) may be responsible for insulin resistance in type II diabetes mellitus.

CGRP inhibits proliferation of cells from the lymph nodes and the spleen of mice [419]. The effect of plasma calcium level and on bone metabolism is similar, but less potent, than that of CT.

Structure – Activity Relationships. The complete structure is required for full biological activity. Acetylation of Lys^{24}, and Lys^{35}, or the N-terminal amino group significantly reduces biological activity [420]. The N-terminal region appears to be important for activity and the C-terminal region for receptor binding.

The N-terminal fragments (1 – 12), (1 – 15), and (1 – 22) of CGRP reduce the blood pressure of rats anesthetized with urethane in a dose-dependent process [421]. h-α-CGRP-(12 – 37) has a weak antagonistic effect in guinea pig heart preparations [422].

3.4. Amylin [423]

Occurrence. Amylin [*112938-42-8*] (islet amyloid polypeptide), a peptide amide containing 37 amino acids, M_r 3903 (h-amylin), was isolated from amyloid deposits in the pancreatic islet cells of patients suffering from type II diabetes (non-insulin-dependent) [424]. There is a 46 % structural similarity between amylin and CGRP. The disulfide ring and the C-terminal amide are important for full biological activity [425].

Amylin also occurs in normal β-cells of various mammals. Not only patients suffering from type II diabetes, but also healthy persons have amylin in their islet β-cells. However, amyloid deposits are only found in type II diabetics and not in healthy persons, indicating a pathogenic effect of amylin [426].

Release. Amylin, like insulin, is released from the rat pancreas by glucose or arginine stimulation [427]. The human plasma level of amylin is low in insulin-dependent diabetes [428].

Biological Effects. Amylin, like CGRP, inhibits the basal and insulin-stimulated glucose uptake and glycogen synthesis in skeletal muscle [429]. Amylin (and/or CGRP) may be responsible for insulin resistance in type II diabetes mellitus.

Amylin exerts similar effects to CT: it inhibits the osteoclastic resorption of bone at low levels of plasma calcium in rats and rabbits and lowers the plasma calcium level in rats [430].

4. Corticoliberin – Proopiomelanocortin Cascade

Corticoliberin (corticotropin releasing hormone or factor, CRH or CRF) and the peptides derived from proopiomelanocortin (POMC) are called the stress hormones because they are released in response to stress. The hypothalamic factor, CRH, stimulates cAMP synthesis and a cAMP-dependent protein kinase [431], resulting in secretion of POMC from the corticotrophic cells of the adenohypophysis, the endocrine cells of the pars intermedia, and central neurons that originate in the hypothalamus. The prohormone, POMC (h-POMC has 241 amino acids, M_r 26 678) is then enzymatically cleaved to form POMC-(1 – 108)-NH_2 (HP-N-108), corticotropin (ACTH), and β-lipotropin (β-LPH),

see Figure 3. Further degradation of HP-N-108 gives POMC-(1 – 76) and HP-N-30 [POMC-(79 – 108)-amide] [432]. The former is degraded to POMC-(1 – 28) [433], δ-melanotropin (δ-MSH), and γ-MSH [434]. ACTH can be cleaved to form α-MSH and the corticotropin-like intermediate lobe peptide (CLIP). Cleavage of the N-terminal tetrapeptide of CLIP gives β-cell tropin (β-CTP), and β-LPH yields γ-LPH, β-MSH, β-endorphin (β-EP), and the melanotropin potentiating factor (MPF). Acetylation by opiomelanotropin acetyltransferase increases the biological activity of α-MSH, decreases the activity of ACTH, and abolishes the opiate activity of β-EP.

In the adrenal cortex, ACTH stimulates the formation of the glucocorticoids and α-MSH releases aldosterone in the presence of angiotensin II (AT II). The gluco- and mineralocorticoids play an important role in stress situations. The glucocorticoids enhance liver gluconeogenesis and exhibit immunosuppressive and anti-inflammatory activity. For further details, see → Hormones, Chap. 4. The release of ACTH and β-EP is inhibited by the glucocorticoids via a negative feedback mechanism. The mineralocorticoid aldosterone exerts antidiuretic and antinatriuretic effects which are important during blood loss.

Both CRH and POMC-derived stress hormones usually inhibit release of the reproductive hormones [435] and enhance release of the atrial natriuretic factor (ANF). ANF then down regulates the stress hormones and stimulates release of the reproductive hormones; it is presumably important as a regulator in coping with stress. The androgens and gonadotropins stimulate POMC synthesis [436], whereas estrogen and testosterone inhibit the formation of the stress hormones [437]. Expression of the POMC gene in the testes correlates closely with the maturation of the Leydig cells. In contrast to the adenohypophysis, the POMC gene in the testes is not regulated by glucocorticoids [436].

Permanent stress probably desensitizes the immunosuppressive ACTH – glucocorticoid system. This condition is characterized by chronic fatigue and increased formation of antibodies against the body's own tissues, a preliminary stage of autoimmune diseases [438].

The structures of h- [439], r- [440], b- [441], sa-, and xe-POMC [442] have been elucidated via their DNA sequences.

4.1. Corticoliberin

Occurrence. Corticoliberin [*9015-71-8*], also known as corticotropin releasing hormone or factor (CRH or CRF), is found in nerve fibers throughout the brain, in the pancreas [443], adrenal medulla [444], placenta [445], stomach, and testes [446].

Although h-CRH is identical to r-CRH (peptide amide with 41 amino acids, M_r 4757), the precursor proteins of these two species have different sequences [447]. p-CRH [448] is very similar to h- and r-CRH.

Release. There is a circadian rhythm in the secretion of CRH. Similar to ACTH, the plasma level of CRH in humans peaks at 6 a.m. and has its lowest values at 6 and 10 p.m. [449]. CRH is released in response to stress (e.g., hunger [450], loss of blood [444], peripheral increase in noradrenaline [451]). Very small doses of catecholamines stimulate and large doses inhibit the central release of CRH [452].

The pulse frequency of CRH in the rat is greatly increased by the chronic administration of alcohol [453]. The release of CRH from the hypothalamus is also stimulated by morphine [454], acetylcholine [455], and interleukins [456]. Endogenous glucocorticoids appear to be important for the synthesis and release of CRH.

The release of CRH from the hypothalamus is inhibited by ACTH [457] and endogenous opioids [458] via a negative feedback mechanism. Very small doses (0.1 – 1 nmol/L) of β-endorphin stimulate the release of CRH in vitro, but higher doses (> 0.1 µmol/L) inhibit both basal and acetylcholine-stimulated release [459].

CRH is involved in the paracrine formation of ACTH in the placenta. Plasma CRH levels increase continually during pregnancy from the second trimester up to birth and fall rapidly after delivery [460].

In Alzheimer's disease, a reduced CRH level is observed in the brain [461]. Patients suffering from anorexia nervosa [462] and depression [463] exhibit enhanced release of CRH.

Receptors. CRH receptors are found in the adenohypophysis and in regions of the brain that

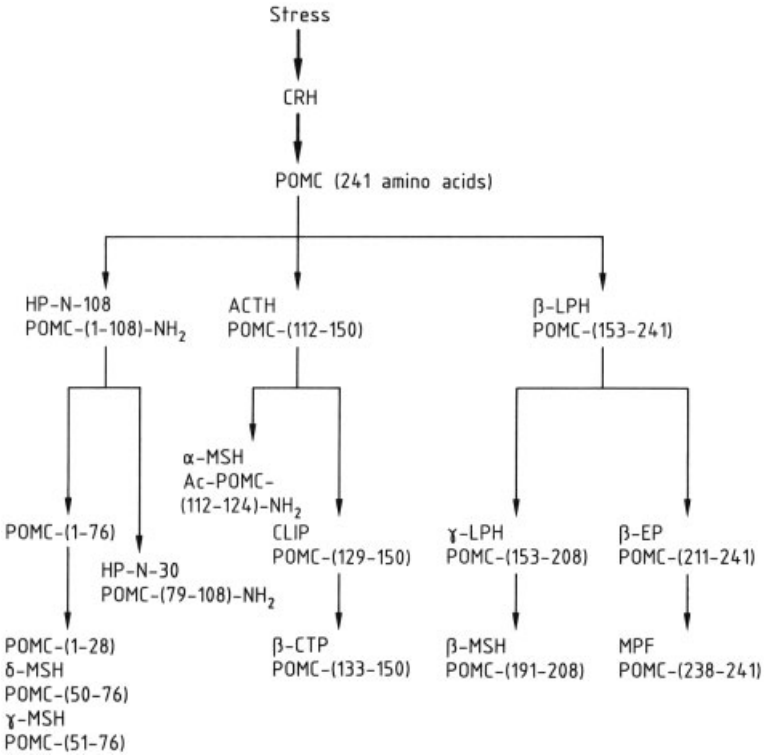

Figure 3. Degradation of proopiomelanocortin (POMC)

are related to the limbic and autonomic nervous systems [464].

In spite of the increased plasma ACTH level in adrenal-ectomized rats, the number of CRH receptors in the hypophysis falls. This down regulation is due to enhanced secretion of vasopressin [465]. In contrast, the CRH receptors on membranes of the liver, spleen, testes, prostate, and pancreas [466] increase in number after adrenalectomy.

Biological Effects. CRH stimulates the synthesis and release of proopiomelanocortin as well as its degradation in the adenohypophysis and placenta [467] to form ACTH, α-melanotropin (α-MSH) [468], γ_3-MSH [469], and β-endorphin (β-EP). It also increases plasma levels of corticosterone, aldosterone [470], and 18-hydroxycorticosterone [471] via release of ACTH. CRH also directly potentiates ACTH-stimulated synthesis of corticosterone [472]. The physiological importance of CRH is emphasized by the fact that CRH antigens cause atrophy of the adrenal glands and lower plasma levels of ACTH and β-EP.

Extracellular Ca^{2+}, the entry of Ca^{2+} into the cell, calmodulin [473], and cAMP-dependent protein kinase [474] all play a role in the CRH-stimulated release of ACTH.

Intravenously applied CRH raises the plasma level of oxytocin and vasopressin, which potentiate the activity of CRH [475]. Angiotensin II [476] and gastrin-releasing peptide-(14 – 27) [477] also potentiate the ACTH-releasing effect of CRH. Somatostatin, melanophore concentrating hormone [478], delta-sleep-inducing peptide [479], glucocorticoids, morphine [480], and endogenous opiates inhibit the ACTH-secreting effect of CRH.

In vitro, CRH stimulates the release of endorphin and dynorphin from the rat pituitary [481]. Intravenous dynorphin A-(1 – 13) and intravenous or intradermal CRH inhibit the inflammatory response to scalding in rats independent of the pituitary and adrenal functions [482]. The

anti-inflammatory effect is not dependent on the hypophysis, but on adrenal function [483].

CRH inhibits the reproductive functions of both sexes [484], [485]. The release of GnRH to the hypophysis from the hypothalamus is prevented by CRH [486]. The prolactin-dependent inhibition of GnRH release is partly due to the activation of CRH-containing neurons [487]. CRH also centrally inhibits the release of growth hormone by stimulating somatostatin release in the hypothalamus [488]. Intravenous CRH raises plasma levels of prolactin [489].

High intravenous doses of CRH selectively dilate the blood vessels of the mesenterium; this is responsible for a decline in peripheral resistance and systemic arterial blood pressure. CRH activates prostaglandin synthesis. The prostaglandins, in turn, inhibit the hypotensive effect of CRH by a negative feedback mechanism [490]. The central application of CRH raises blood pressure and pulse rate by increasing adrenaline and noradrenaline levels. In vivo, CRH potentiates the cholecystokinin-induced pancreatic secretion of protein and the secretin-induced secretion of HCO_3^- [491].

Intravenous CRH results in a strong and persistent lowering of the level of pentagastrin-stimulated gastric acid, while the plasma gastrin level increases [492]. Intraperitoneal CRH prevents formation of stress-induced gastric ulcers in rats [493]. CRH inhibits the emptying of the stomach [494], reduces the transit in the small intestine, increases the transit in the large intestine, and causes diarrhea [495].

In humans, intravenous CRH increases the breath-time volume, breath frequency [496], and attentiveness; it also improves the mood and well-being of depressive patients [497].

CRH exerts a central antipyretic effect [498]. In experimental animals (mainly rats) central application of CRH increases locomotoric activity, emotionalism [499], and cleaning behavior; reduces food uptake [500] and aggressiveness; and enhances defensive behavior [501].

Structure – Activity Relationships. The three amino acids at the N-terminus of CRH are not important for biological activity. However, vasodilating and ACTH-releasing activity fall to below 0.1 % on further deletion of N-terminal amino acids.

The C-terminal region is of extreme importance for CRH activity. Both 41-deamido-o-CRH-(1 – 41) with a free C-terminal carboxyl group and o-CRH-(1 – 39)amide have < 0.1 % of the activity of CRH. Met^{21} is of some significance for activity because $[Met(O)^{21}]$o-CRH exhibits only 10 % of the biological activity. However, norleucine and norvaline can be substituted for this amino acid with practically no loss of activity.

Uses. Subcutaneous injection is advisable and is equally as effective as intravenous administration. Activity after intranasal administration is only about 1 % of that obtained on parenteral administration [502].

A pulsatile injection of CRH normalizes ACTH and cortisol secretion after glucocorticoid therapy (secondary adrenal insufficiency) [503] and apparently does not desensitize the pituitary [504]. Pulsatile application (6.25 µg every 30 min) is preferable to long-term infusion (50 or 100 µg/4 h) [505].

CRH can be used as a diagnostic aid for hypophyseal function [506]. After intravenous administration of 100 µg of CRH, the plasma ACTH and cortisol levels in patients suffering from Cushing's syndrome increase to a higher extent than in healthy persons [507]. In depressive patients with a high level of cortisol, CRH produces a lower ACTH response than in normal persons [508]. An abnormal increase in growth hormone after application of CRH is observed in patients suffering from congenital thyrotropin deficiency [509].

4.2. Corticotropin [510]

Occurrence. Corticotropin [*9002-60-2*], also known as adrenocorticotropic hormone (ACTH), has 39 amino acids, M_r 4541 (h-ACTH), and is formed in the adenohypophysis from the precursor POMC. It is also found in the interstitial fluid of the testes [511] and in the pancreas [512]. Enzymatic degradation and acetylation of ACTH give rise to the N-terminal derivative α-melanotropin (α-MSH) and two C-terminal derivatives: the corticotropin-like intermediate lobe peptide (CLIP) and β-celltropin (β-CTP). ACTH sequences from various species are given in [513].

Release. ACTH is formed from the precursor POMC in response to stress, CRH, and stimuli that release CRH or potentiate its activity. The formation of ACTH is accompanied by an increase in cAMP (see Section 4.1) [510]. ACTH is secreted in micropulses lasting a few minutes. Pulse frequency remains the same throughout the day, but the amplitude is higher in the morning than in the afternoon and night [514], it is also higher in men than in women [515]. Serotonin increases the formation of ACTH and cortisol by stimulating release of CRH. It appears to be partly responsible for the circadian increase in cortisol [516].

In vivo, corticosteroids inhibit the secretion of ACTH [510] in the adenohypophysis and cerebrospinal fluid. Glucocorticoids do not affect the in vitro CRH-stimulated release of ACTH from the pituitary [517]. γ-Aminobutyric acid (GABA) appears to reduce the release of ACTH via inhibition of CRH release [516]. Increased salt intake depresses the formation of ACTH in response to stress or CRH [518].

Inhibitors of angiotensin converting enzyme inhibit the release of ACTH via decreased synthesis of angiotensin II.

ACTH secretion is increased in Cushing's syndrome, but lowered in Sheehan's syndrome. The circadian ACTH rhythm is disturbed in patients suffering from depression; hypersecretion of ACTH and cortisol is observed [519].

Receptors. An ACTH receptor (M_r 225 000) from mouse adrenal cells consists of four different subunits [520]. In humans, ACTH receptors have also been found on splenocytes and peripheral mononuclear leukocytes [521].

IGF-I potentiates the ACTH-stimulated formation of ACTH receptors [522] and TGF-β reduces the number of ACTH receptors on adrenocortical cells [523].

Biological Effects. In the adrenal cortex, ACTH stimulates the synthesis of glucocorticoids (cortisol, cortisone, corticosterone) and mineralocorticoids (cortexone and aldosterone) in a calcium-dependent process [524] that involves activation of adenylate cyclase. ACTH must be administered in the morning for maximum stimulation of glucocorticoid synthesis. The sensitivity of adrenal cells to ACTH is reduced when they are exposed to ACTH for longer periods [525].

The ACTH-stimulated secretion of aldosterone probably proceeds via angiotensin II [526]. ACTH also enhances the release of adrenaline and noradrenaline in the adrenal gland [527]. It exerts a trophic effect on the adrenocorticotropic cells of the adrenal gland. Adrenal hyperplasia is observed as a result of the hypersecretion of ACTH (Cushing's disease). The adrenal glands shrink as a result of ACTH insufficiency of the pituitary (Sheehan's syndrome).

Corticostatin from granulocytes is a peptide containing 34 amino acids that inhibits the ACTH-stimulated formation of corticosterone [528].

ACTH and N-terminal fragments of ACTH, e.g., ACTH-(4 – 10) [529] also have a therapeutic effect in cases of shock [530]. Cholinergic mechanisms in the central nervous system play an important role here [531].

ACTH-(1 – 24) stimulates the pancreatic secretion of $NaHCO_3$ and protein via cholinergic mechanisms [532]. The ACTH molecule also has insulinotropic and hypoglycemic activity [512]. However, it also potentiates the diabetogenic activity of growth hormone induced by hypoglycemia [533].

ACTH binds specifically to splenocytes, inhibiting the production of antibodies and interferon (direct immunosuppressive activity) [521].

ACTH-(4 – 10) is responsible for the influence of ACTH on behavior and on learning processes (see Section 4.3): it increases attentiveness, sexual activity, and the memory capacity of rats, at the same time reducing anxiety. Cholinergic neurons and muscarinic receptors play a role in ACTH-dependent behavior patterns [534].

The central application of ACTH, ACTH-(1 – 24), and α-MSH decreases food uptake [535]. Centrally applied ACTH increases blood pressure and heart rate, it also prevents persons from falling asleep, while de-Ac-α-MSH and CLIP intensify slow-wave sleep and paradoxical sleep [536].

Structure – Activity Relationships. A fragment as small as ACTH-(4 – 7) (10^{-4} mol/L) stimulates in vitro corticosterone synthesis. A dramatic increase in activity occurs with the in-

clusion of Met^3. However, the full activity of the N-terminal sequence is attained with a free amino group at the N-terminal Ser^1. Thus, ACTH-(1 – 10) is 100 times more active than ACTH-(5 – 10), and ACTH-(1 – 18) has the full activity of native ACTH.

ACTH-(1 – 24) (tetracosactid) is more active in releasing corticosteroids than ACTH in vitro. ACTH can be divided into at least four segments. Sequence (11 – 18) is very important for receptor binding. Sequence (4 – 10) is responsible for the corticotropic effect, and the N-terminal tripeptide (1 – 3) is the "amplifier". The C-terminal fragment, ACTH-(25 – 39), is responsible for antigenicity and safe transport. ACTH-(11 – 24) is a competitive antagonist for the corticotropic activity of ACTH. [Phe^9]ACTH-(1 – 24) also inhibits glycolysis and steroidogenesis.

ACTH from guinea pigs in which Pro^{24} is replaced by Ala has a higher aldosterone-releasing activity than hACTH or ACTH-(1 – 24)amide [513]. [Cys(carboxamidomethyl)^{25}]ACTH-(1 – 26) has three to four times the activity of ACTH in stimulating aldosterone secretion, but a lower corticosterone-stimulating activity. The substitution of D-Ser or β-Ala for Ser^1, of Lys for Arg^{17}, and of Lysin-amide or 1,4-diaminobutylamine for Arg^{18} gives long-acting analogues that have about up to eight times the activity of ACTH.

[β-Ala^1, Lys^{17}]ACTH-(1 – 17)-4-amino-*n*-butylamide (ACTH-(1 – 17), alsactide) is stable to aminopeptidases and carboxypeptidases and has a long duration of action.

CLIP exhibits insulin-releasing activity which decreases after shortening or acetylation of the molecule at the N-terminus [537]. CLIP, like α-MSH and de-Ac-α-MSH, lowers the β-endorphin-stimulated secretion of prolactin in rats [538].

Uses. Natural ACTH from the pig is marketed under the name Acethropan (Hoechst). The synthetic ACTH derivatives, tetracosactide (Synacthen, CIBA) and alsactide (Synchrodyn 1 – 17, Hoechst), are also used. ACTH preparations are used as diagnostic aids for the functioning of the adrenal cortex and as a therapeutic agent for insufficient functioning of the adrenal cortex, in multiple sclerosis, inflammatory rheumatic diseases, collagen diseases, acute gout, radicular pain syndrome, severe allergic skin diseases, and collitis ulcerosa. ACTH also has an antiemetic effect and is administered to cancer patients treated with *cis*-platinum [539].

4.3. Melanotropins

Occurrence. α-Melanotropin [*37213-49-3*], also known as α-melanocytes stimulating hormone (α-MSH), M_r 1665, is derived from POMC (POMC-112 – 124). In mammals its amino acid sequence is:

`Ac-S Y S M E H F R W G K P Va`

It is found primarily in the hypophyseal pars intermedia and the hypothalamus.

γ- and δ-MSH are derived from POMC-(1 – 28) and β-MSH from γ-LPH (see Fig. 3). In many species, two different β-MSH peptides are found.

Release. Release of α-MSH is probably regulated through the enzymatic degradation of oxytocin as the prohormone for melanostatin (MIF) and melanoliberin (MRF). Stress and corticoliberin increase brain and plasma α-MSH levels. β-Endorphin can also release α-MSH in the central nervous system. The α-MSH is released in the septum in response to fever induced by interleukin-1, especially when the temperature increases (shivering phase) [540].

Increased levels of γ-MSH are observed in physiological stress, cardiovascular distress, blood loss [541], and cardiac arrest [542].

Biological Effects. Melanin is a black-brown skin pigment that consists mainly of polymerized dihydroxyphenylalanine. In melanocytes, α-MSH stimulates tyrosinase and thus the synthesis of melanin (melanogenesis); it also promotes pigment transport of pigment granules (melanosomes). Dispersion of the melanosomes through the numerous dendrites of the melanocytes causes the skin to darken, the skin is lightened by aggregation of melanosomes. Human skin becomes darker within 24 h after administration of α-MSH. Darkening of the skin is observed in kidney damage and diseases in which the plasma α-MSH and ACTH levels are increased (Cushing's, Addison's, Nelson's disease, and ACTH-secreting tumors). The skin becomes lighter in hypophyseal insufficiency. The C-terminal tetrapeptide

of β-endorphin, the melanotropin potentiating factor (MPF), enhances the melanotropic activity of α-MSH.

The second messenger of α-MSH appears to be cAMP. Inhibitors of α-MSH (e.g., melatonin) inhibit the α-MSH-stimulated formation of cAMP but increase the concentration of cGMP. Melatonin, a hormone found in the pineal body, stimulates the aggregation of melanosomes.

α-MSH has many other activities in mammals [543]:

- Increase in testicular Sertoli cell cAMP
- Secretion of estradiol and plasminogen activator
- Increase in lipogenesis and sebum production in the skin
- Increase in lipolysis in adipose tissue
- Increase in plasma free fatty acid levels
- Increase in adrenal steroidogenesis related to fetal growth and development
- Increase in aldosterone synthesis and secretion by the adrenal zona glomerulosa (angiotensin II seems to be important [544])
- Increase in pineal serotonin levels
- Decrease in pineal melatonin levels
- Increase in growth hormone secretion from the pituitary
- Inhibits stress and β-endorphin-stimulated prolactin release [545]
- Increase in plasma luteinizing hormone [546]
- Increase in plasma glucagon and insulin
- Decrease in blood pressure
- Decrease in plasma calcium levels
- Decrease in bone resorption
- Decrease in immunomodulatory and inflammatory activities of IL-1 [547]
- Reduction of body temperature following intracerebroventricular or parenteral administration [548]
- Behavior (increased arousal, attention, learning, memory retention, sexuality)
- Improved nerve regeneration

h-β-MSH has the same activity as α-MSH in the Anolis skin test (darkening of the skin). β-MSH sequences from other species also show high melanotropic activity, which is usually slightly lower than that of α-MSH. Like α-MSH, β-MSH also raises the plasma levels of glucose, glucagon, insulin, and free fatty acids in the rabbit.

Of the γ-MSH peptides, Ac-γ_1-MSH exerts the highest melanotropic effect but still has $< 0.1\,\%$ of the α-MSH activity. γ_2-MSH inhibits the release of α-MSH from the hypophysis and reduces the β-endorphin-induced analgesic and hypothermic effect. The direct infusion of γ-MSH into the renal artery results in prompt excretion of sodium and potassium [549]. The ACTH-stimulated synthesis of glucocorticoids and aldosterone is potentiated by γ_3-MSH [550]. Intraventricular application of γ-MSH leads to a longterm increase in blood pressure [551]. The functions of POMC-(1 – 108) and its fragments are probably based on its growth-promoting effect on the adrenal cortex and its hypertensive effect.

Structure – Activity Relationships. The N-terminal acetyl group is important for the melanotropic activity of α-MSH. α-MSH can be subdivided into three regions:

1) The classical messenger sequence, His-Phe-Arg-Trp (α-MSH-(6 – 9))
2) The C-terminal tetrapeptide Gly-Lys-Pro-Val-amide which has seven times the melanotropic activity of α-MSH-(6 – 9)
3) The N-terminal sequence Ac-Ser-Tyr-Ser-Met-Glu which acts as a potentiator

Within the messenger sequence, -Phe-Arg- appears to be crucial for melanotropic activity. The minimum effective sequence is Ac-His-Phe-Arg-Trp-amide [552]. The lysine in the C-terminal tetrapeptide is important for activity. Met^4, Gly^{10}, and Pro^{12} are important for MSH activity [552]. If Met^4 and Gly^{10} are replaced by a cysteine disulfide bridge ([Cys^4,Cys^{10}]α-MSH), a cyclopeptide is obtained which contains the two important messenger sequences. This peptide is a superagonist and is 10 000 times more active than α-MSH in the frog skin test. The cyclic analogue Ac-[Cys^4,Cys^{10}]α-MSH-(4 – 10)amide is less active than α-MSH, whereas Ac-[Cys^4,Cys^{10}]α-MSH-(4 – 13)-amide is again superactive and has the same activity as [Cys^4,Cys^{10}]α-MSH in the frog skin test. Thus, the C-terminal sequence is of great significance for MSH activity. The disulfide bridge is important for biological activity. Reduction of the disulfide bridge results

in a 1000 – 10 000-fold decrease in biological activity [553].

Met^4 is sensitive to oxidation, and can be replaced by norleucine without loss of activity. Treatment of α-MSH with hot alkali prolongs the duration of action. [Nle^4,D-Phe^7]-α-MSH is 60 times as active as α-MSH in the frog skin test. In mice, [Nle^4,D-Phe^7](4 – 11) or -(4 – 10)α-MSH has 100 times the activity of α-MSH [554]. Even Ac-[Nle^4,D-Phe^7]α-MSH-(4 – 9)amide is still ten times as active as α-MSH in the melanoma tyrosinase and lizard skin tests, but it is ten times less active in the frog skin test [555]. The central application of [Nle^4,D-Phe^7]α-MSH has an antipyretic effect that is ten times greater than that of α-MSH. The antipyretic effect after intravenous application is, however, not pronounced. The C-terminal tripeptide appears to be very important for antipyretic activity [556].

Cyclic compounds containing D-Phe^7, an acidic amino acid in position 5, and a basic amino acid in position 10 are more active than α-MSH [557].

The activity of [Cys^4,D-Phe^7,Cys^{10}]α-MSH on melanocytes is about as high as that of [Cys^4,Cys^{10}]α-MSH, but it acts considerably longer. [Cys^4,D-Phe^7,Cys^{10}]α-MSH-(1 – 12)amide has the same potency as [Cys^4,D-Phe^7, Cys^{10}]α-MSH, indicating that Val^{13} is not required for melanotropic activity [558].

An analogue of α-MSH-(4 – 9), Met(O_2)-Glu-His-Phe-D-Lys-Phe-OH (ORG-2766) [559], [560] is about 100 –1000 times more active than α-MSH-(4 – 9) in promoting learning. However, high doses have the opposite effect. The subcutaneously or orally applied peptide has an antiamnesic effect, antagonizes pentobarbital anesthesia, and reduces morphine uptake in the brain. Met(O_2)-Glu-His-Phe-D-Lys-Phe-NH-$(CH_2)_8$-NH_2 (HOE 427, Ebiratide [561]) and Met(O)-Glu-His-Phe-D-Lys-Phe-NH-$(CH_2)_8$-NH_2 have 100 times the ORG-2766 activity in this test. In learning tests with rats, HOE 427 is about 500 times more active than ORG-2766 [562]; the sequence Phe-D-Lys-Phe appears to be especially important [563].

Antagonists [543]. Ac-α-MSH-(7 – 10)amide is a weak, selective α-MSH antagonist in the lizard skin bioassay. Ac-[D-Trp^7,D-Phe^{10}]α-MSH-(7 – 10)amide is a competitive inhibitor of α-MSH in the frog and lizard skin assays [564]. A strong antagonist for frog skins is the uncyclized peptide Ac-Nle-Asp-Trp-D-Phe-Nle-Trp-Lys-amide; the cyclic lactam of this compound is a full agonist. Other antagonists are the growth hormone releasing peptide, His-D-Trp-Ala-Trp-D-Phe-Lys-amide and its analogues.

Uses. [Nle^4,D-Phe^7]α-MSH (intermedin alpha) can be applied topically and is absorbed through the skin causing increased pigmentation in the yellow mouse [565]. It is being tested as a suntanning agent.

α-MSH-(6 – 9) exerted a positive transdermal effect (better mood, as well as less anxiety, pain, spasticity, and muscular weakness) in patients suffering from multiple sclerosis [566].

ORG-2766 improved the mood and level of performance of patients without influencing their sleep [554]. Even in elderly, mentally weak patients, ORG-2766 increased attentiveness and induced social behavior [567]. Ebiratide is being tested on patients suffering from Alzheimer's disease.

4.4. Opioid Peptides

Many peptides that have an effect on opioid receptors are formed from prohormones. *β-Endorphin* (β-EP, Section 4.4.1) is derived from proopiomelanocortin (Fig. 3), its activity is based on the N-terminal pentapeptide sequence Tyr-Gly-Gly-Phe-Met. This sequence is also found in *Met-enkephalin* (Met-EK), which is produced from prepro-EK A of the adrenal medulla. A series of *extended Met-EKs* that are structurally related to prepro-EK A have been isolated from the adrenal medulla (Section 4.4.2.1). The *extended Leu-EKs* isolated from the pituitary and hypothalamus (e.g., dynorphins, neoendorphins, and rimorphin) are derived from a common precursor, prepro-EK B (Section 4.4.2.2). Prepro-EK A and B are structurally similar.

The C-terminal tetrapeptide sequence of prepro-EK A is, apart from the missing amide group, identical to that of femarfarmamide (FMRF-amide), a peptide isolated from mollusks. FMRF-amide is not an opiate. It is not derived from prepro-EK A because the C-terminal glycine important for the formation of the amide group is missing. In mollusks, it is formed from

a prepropeptide which contains 21 copies of the precursor FMRFG sequence.

Kyotorpin, a Tyr-Arg dipeptide with analgesic activity, was isolated from the bovine brain. It is probably derived from the propeptide neokyotorpin (Thr-Ser-Lys-Tyr-Arg).

The *dermorphins* and *deltorphins* (Section 4.4.3) are heptapeptide amides from the skin of the South American frogs *Phyllomedusa sauvagii, P. rhodei,* and *P. bicolor*. They contain D-alanine or D-methionine.

The *exorphins* (Section 4.4.4) are opioid peptides which are formed during food digestion. Examples are α- and β-casomorphins.

Modified morphines (e.g., naloxone or certain EK analogues) are antagonists for endogenous opioids and exogenous opiates. In a certain dosage range, cholecystokinin (CCK-8) also acts as an endogenous selective antagonist for the analgesic activity of opioids.

Opioid Receptors. Prior to the discovery of endogenous opioids, three opioid receptors were postulated: the μ-receptor (for morphine), κ-receptor (for ketocyclazocin), and σ-receptor (for SKF 10 047). The discovery of the enkephalins (EK) led to the introduction of the δ-receptors. β-EP prompted the postulation of the ε-receptors, and the naloxone-sensitive effect on thermoregulation gave rise to γ-receptors. Iota-receptors have been postulated in the dog and rabbit intestine, and λ-receptors with a high affinity for 4,5-epoxymorphinane have been described. The μ-receptors have been subdivided into μ_1- and μ_2-receptors. There are three types of κ-receptor [568].

All endogenous opioids are ligands for the μ-, δ-, κ-, and ε-receptors. Met-EK, Leu-EK, and β-EP are the main ligands for the μ- and δ-receptors. The dermorphins are specific for the μ-receptors, the deltorphins for the δ-receptors, and the dynorphins for the κ-receptors. β-EP binds primarily to the ε-receptors. Opiates (e.g., morphine) bind to the μ_2-receptors.

In vitro bioassays for the receptors are described in [569]. Selective ligands for the μ- and δ-receptors are described in [570]. Peptide ligands for the κ- and ε-receptors are given in [571–574].

Biological Activity. The most important effects mediated by the opioid receptors follow [575]:

μ_1-mediated: supraspinal analgesia, inhibition of gastrointestinal transit and motility [576], suppression of experimentally induced diarrhoea [576], and development of anorexia.

δ-mediated: spinal analgesia, inhibition of acetylcholine release from rat corpus striatum [577], impairment of avoidance learning in rats [578], and inhibition of SP-stimulated plasma extravasation and vasodilation [579].

κ-mediated: analgesia, increased food intake [580], ACTH release in rats [581], and diuresis (mediated by the adrenal medulla) [582].

ε-mediated: supraspinal analgesia.

σ-mediated: psychomimetic effects and behavioral changes.

4.4.1. β-Endorphin

Occurrence [583]**.** β-Endorphin (β-EP) [*60617-12-1*], M_r 3465 (h-β-EP), contains 31 amino acids and is derived from proopiomelanocortin (see Fig. 3). It has been isolated primarily from the pituitary, but is also found in human placenta cell cultures, ovaries [584], sperm [585], endometrium [586], gallbladder [587], pancreas, and small intestine.

β-EP derivatives such as β-EP-(1 – 27) (C fragment or δ-EP), Ac-β-EP, Ac-β-EP-(1 – 27) [588], and β-EP-(1 – 18) are found in the rat adenohypophysis. Acetyl-β-EP-(1 – 18) mainly occurs in the rat neurohypophysis [589]. In rats, the degree of acetylation of the hypothalamic endorphins increases with age [590].

The C-terminal tetrapeptide of β-EP (Lys-Lys-Gly-Glu in humans) is the melanotropin potentiating factor (MPF).

Release. β-EP is secreted in the hypophysis in response to stress, corticoliberin, angiotensin II, lipoxygenase or epoxygenase products [591], insulin-induced hypoglycemia (via cholinergic mechanisms [592]), adrenaline, food uptake, and chronic alcohol consumption [593]. β-EP is released from the mucosa of the small intestine by gastric acid or bile acid and from the gallbladder mucosa by CCK-8 [588]. Estrogen and testosterone increase the plasma

level of β-EP [594]. Both h-CG and PMSG stimulate formation of β-EP in the ovaries [595] and in the testicular interstitial fluid [596].

In humans, somatostatin and oxytocin have no effect on the basal plasma β-EP level, but lower the increased levels of β-EP, β-LPH, and cortisol caused by insulin-induced hypoglycemia [597].

In humans, release of β-EP exhibits a circadian rhythm similar to that of cortisol: a high level in the morning and a low level at night [598]. The basal β-EP level is lower in people who bear a high risk of becoming alcoholics compared to those who do not. The plasma level of β-EP increases with alcohol consumption in the high risk group, but not in the low risk group [599]. Plasma levels of β-EP are significantly raised in depressive patients [600] and in obesity [601].

Biological Effects. β-EP has analgesic and lipolytic activity [602]. The intracerebroventricular application of CCK-8 [603] and α-MSH inhibits the analgesic activity of β-EP. β-EP lowers phosphodiesterase activity, somatostatin secretion in the isolated pancreas of the dog, the release of GnRH in the mesencephal central grey substance, plasma LH, the oxytocin release induced by sucking, and, in high doses, the formation of CRH, ACTH, LPH, and cortisol. Very low doses release CRH in rat hypothalami [604].

Intraventricular and intravenous application of β-EP causes release of growth hormone and prolactin. The β-EP-stimulated release of prolactin can be inhibited by α-MSH or CLIP [605]. β-EP increases plasma insulin, glucagon, and glucose, especially in obese humans [606].

Pharmacological doses of β-EP reduce the left ventricular systolic and diastolic pressure in the isolated rat heart [607]. Depending on the dose and conditions, the intracerebroventricular application of β-EP results in hyper- or hypothermia. Other central effects are the release of catecholamines, Met-EK [608], thyrotropin, vasopressin, and α-MSH as well as increased food uptake.

Structure – Activity Relationships. β-EP has a highly specific opiate recognition sequence at the N-terminus (positions 1 – 5) which is linked to an amphiphilic helix (positions 13 – 31) by a hydrophilic region (positions 6 – 12) [609]. Since the N-terminal Tyr-Gly bond of β-EP is more resistant to aminopeptidases than that of the EKs, β-EP has a longer duration of action. Acetylated sa-β-EPs have no activity in an opiate receptor assay.

Incorporation of Gln or Arg at position 8 doubles the analgesic effect, Trp in position 27 quadruples the analgesic effect. [Gln^8,Trp^{27}]h-β-EP has almost eight times the receptor binding ability of h-β-EP, but its analgesic effect corresponds to that of h-β-EP. The replacement of Glu^8 appears to be of great importance for receptor binding and may find application for the design of β-EP antagonists [610]. [Gln^8,Gly^{31}]h-β-EP-Gly-Gly-amide is a strong antagonist for β-EP-induced analgesia, it is 200 times more active than naloxone [611].

The substitutions D-Ala^2 and $MePhe^4$ increase the binding of the β-EPs to the δ-receptors. Increasing hydrophobicity at position 5 correlates with decreasing analgesic activity [612].

Substitution of the dermorphin sequence for the seven amino acids at the N-terminus gives a highly analgesic peptide which is 4.4 times more active than h-β-EP and about as active as dermorphin [612].

Analogues which do not show homology with β-EP in the twelve amino acids at the C-terminus but have the helical structure of the C-terminus are as active or more active than β-EP. Cysteine bridges between positions 14 and 26, 15 and 26, 16 and 26, as well as 17 and 26 are tolerated. These derivatives exhibit stronger receptor binding than h-β-EP [613].

β-EP-(1 – 27) has only 0.2 % of the β-EP activity in the opiate receptor displacement assay. β-EP-(1 – 27) inhibits β-EP-induced analgesia and release of growth hormone [614], and the β-EP-induced hypothermia in the mouse [615], but not the β-EP-induced release of prolactin [614].

The N-terminal tyrosine is very important for the analgesic activity of β-EP: de-Tyr^1-β-EP does not bind to opiate receptors. h-β-EP-(6 – 31) inhibits β-EP-induced analgesia, the β-EP stimulated release of prolactin, but has no effect on the release of TSH [616]. h-β-EP-(6 – 31), h-β-EP-(28 – 31), and h-β-EP-(30 – 31) inhibits the α-MSH-induced grooming, stretching, and yawning syndrome, as well as β-EP-induced grooming and catatonia [617]. The C-

terminus is of prime importance for the lipolytic activity of β-EP. Derivatives in which the two C-terminal amino acids are deleted have no lipolytic activity [618].

4.4.2. Preproenkephalins A and B

There is considerable structural similarity between the two preproenkephalins (prepro-EKs) isolated from endocrine and nerve tissue. The structures of h-, b-, r- [619], xe- [620] prepro-EK A and h- and b-prepro-EK B have been elucidated. Prepro-EK A is mainly cleaved to give Met-EK and extended Met-EK peptides (e.g., peptide E and adrenorphin). Prepro-EK B is cleaved to give Leu-EK, the neoendorphins, and the dynorphins.

Prepro-EKs are widely distributed in the body and are processed differently according to their location [621].

4.4.2.1. Peptides Derived from Preproenkephalin A

Occurrence. Preproenkephalin A (prepro-EK A) [*88895-24-3*], M_r 30 781 (h-prepro-EK A), contains 267 amino acids. Numerous peptides derived from prepro-EK A have been isolated from the adrenal medulla [622]; these peptides are in turn precursors of Met-EK (Tyr-Gly-Gly-Phe-Met, M_r 574) and Leu-EK (Tyr-Gly-Gly-Phe-Leu, M_r 556). Plasma proteins treated with pepsin also generate peptides related to Met-EK [623].

Release. The release of the enkephalins is stimulated by GABAnergic mechanisms, insulin-induced hypoglycemia [624], endotoxin shock [625], electroshock [626], intraventricular application of β-endorphin [627], and glucocorticoids [628]. Dopamine inhibits the formation of Met-EK [629]. Synthesis of messenger RNA for prepro-EK A is stimulated by stress [630], angiotensin II [631], FSH, and cAMP [632], but inhibited by chronic administration of glucocorticoids [630]. The plasma level of Met-EK increases considerably in marathon runners [633].

Met- and Leu-EK are rapidly degraded by aminopeptidases and enkephalinase, which cleaves the C-terminal dipeptide.

Biological Effects [634]**.** For a description of the activity of the enkephalins, see Section 4.4 (μ_1-, δ-, and κ-mediated opioid actions). Enkephalins inhibit transmitter release from nerve terminals in the central and peripheral nervous systems by blocking calcium channels [635].

Met- and Leu-EK mainly bind to δ-receptors and only have an analgesic effect if applied centrally. Analogues that preferentially bind to the μ- and δ-receptors exhibit an analgesic effect, even after peripheral application. The tachyphylaxis and substance dependency produced by morphine are also observed after chronic application of the highly active EK analogues.

Endocrine and exocrine pancreatic and gastrointestinal secretion are modulated by the enkephalins [636]. Enkephalins contract the lower esophagus and the pyloric sphincter [637], delay gastrointestinal transit [638], trigger gallbladder contraction [639], relax the Oddi sphincter [640], cause vasoconstriction in the lungs [641], and inhibit bladder motility [642].

Depending on the dosage and species, enkephalins exert a hypotensive [643] or hypertensive effect [644]. The intracerebroventricular application of μ- and δ-agonists to rats inhibits the release of oxytocin [645] and vasopressin (diuresis).

The enkephalins modulate the release of ACTH and cortisol via extrahypophyseal mechanisms [646]. They potentiate the ACTH-stimulated release of corticosterone [647] and inhibit the CRH-induced increase in the plasma levels of ACTH, β-EP, and cortisol [648].

The enkephalins increase the plasma level of prolactin by inhibiting dopamine release. They stimulate the release of growth hormone via the growth hormone releasing hormone.

The enkephalins inhibit ovulation and the release of lutropin by inhibiting GnRH synthesis. Endogenous opioids inhibit gonadotropin secretion; this is dependent on the gonadal steroids. The FSH-stimulated formation of progesterone is promoted by the enkephalins [649], while the secretion of testosterone from the testes of immature rats is suppressed [650]. In experimental animals, endogenous opioids inhibit the release of TRH and, thus, the formation of TSH [651].

The endogenous opioids can also act as immunostimulants [652] and immunosuppressants [653].

Structure – Activity Relationships. Tyr^1 is essential for opiate activity. However, its N-terminal amino group can be methylated, guanylated, or extended by amino acids without significantly affecting the activity. The *N*-allyl-EKs (particularly *N*-allyl-Met-EK) are, like naloxone, antagonists of morphine and EK.

Gly^2 can be replaced by α-aminoisobutyric acid or D-amino acids to give compounds that are considerably more active and more resistant to enzymatic degradation. The more lipophilic the substitution, the stronger the binding to the μ-receptors and, therefore, the analgesic activity. Substitution with hydrophilic groups increases affinity for the δ-receptors.

Gly^3 is important for the biological activity of the EKs. Replacement of Phe^4 by other amino acids leads to loss of activity. However, N-methylation or substitution by AzaPhe increases analgesic activity.

The more lipophilic the amino acid in position 5, the stronger the analgesic effect after parenteral administration. For example, the intracerebroventricular application of the *O*-galactosyl derivative of [D-Met^2,Hyp^5]EK-amide produces an analgesic effect that is 50 000 times that of morphine [654]. Position 5 does not necessarily have to be occupied by an amino acid: methioninol sulfoxide, the thiolactone of homocysteine, and substituted hydrazides are also suitable.

Shortening of the chain at the C-terminus leads to the tripeptide-*N*-methylphenethylamides or tripeptide-2-amino-4-methylpentane amides, with a still higher analgesic activity. H-Tyr-NH(Me)-$(CH_2)_4$-CO-NH(Me)-CH_2-CH_2-C_6H_5 also shows high analgesic activity when applied intravenously [655]. An EK analogue that is cyclized via a disulfide bridge is [D-Cys^2-D-Cys^5]EK-amide also has a powerful analgesic effect. All of the above highly active compounds are more stable to enzymatic degradation than the rapidly degradable natural EKs.

The hydrophilic EK analogues, [D-Arg^2,Phe$(NO_2)^4$, Pro^5]EK-amide (BW 942C) [656] and [D-Met(O)^2, Phe$(NO_2)^4$, Pro^5]EK-amide (nifaltide) [657] are effective against diarrhoea.

4.4.2.2. Peptides Derived from Preproenkephalin B

Occurrence. Preproenkephalin B (prepro-EK B) [*88895-25-4*], M_r 28 422 (h-prepro-EK B), contains 254 amino acids. The most important peptides derived from prepro-EK B are the pentapeptide Leu-enkephalin (Leu-EK), M_r 556 [658]; the longer dynorphins (DPs); and the neoendorphins (neo-EP). The DPs and neo-EPs were first isolated from the hypothalamus, hypophysis, and duodenum of the pig. Later, DPs were found in the bovine adrenal gland [659], guinea pig heart [660], and rat duodenum [661].

Leu-EK	Y G G F L
DP-A-8	Y G G F L R R I
DP-A-9	Y G G F L R R I R
DP-A-11	Y G G F L R R I R P K
DP-A-13	Y G G F L R R I R P K L K

Release. During dehydration, immunoreactive DP is elevated in the hypothalamus but lowered in the hypophysis.

In ovariectomized rats, the level of immunoreactive DP increases in the adenohypophysis, this increase is prevented by estrogen [662]. Accumulation of immunoreactive DP-8 in the hippocampus and frontal cortex of rats appears to be accompanied by reduced learning ability [663]. After traumatic injuries, immunoreactive dynorphin increases in the spinal cord of the rat. This appears to be responsible for hind limb paralysis, an attendant symptom of traumatic injuries [664]. Immunoreactive DP levels are low in the spinal cord of schizophrenic patients [665].

Biological Effects. See Section 4.4 (κ-receptor-mediated opioid actions). DPs potentiate the glucose- or amino-acid-stimulated release of insulin [666], but inhibit the release of somatostatin [666], TRH [667], oxytocin [668], vasopressin [669], and counteract morphine tolerance [670].

The central application of DPs in rats decreases body temperature, increases food [671] and water intake [672], suppresses motor activity [673], reduces the response to acoustic signals [673], and leads to hind-limb paralysis similar to that observed after spinal cord injury [674].

Structure – Activity Relationships. DPs are specific ligands for the κ-receptors. The affinity for the δ-receptors increases with decreasing chain length [675]. Short-chain DPs are rapidly degraded by peptidases and therefore have a considerably shorter biological activity. DP-A-8 and DP-A-9 are assumed to have a neural transmitter or modulator function at the κ-receptors, and the more stable DP-A-13 and DP-A-17 to have a more hormonal function. As with other opioids, the N-terminal tyrosine is essential for opiate activity.

The basic amino acids, particularly Arg^6 and Arg^7, are important for the biological activity of DP-A-13 [676]; substitution of Ala for Ile^8 gives a more active (2 – 9 times) compound [676]. [D-Cys^2,Cys^5-*N*-$MeArg^7$,D-Leu^8]DP-(1 – 8)ethylamide is highly analgesic, it binds more strongly to the μ- and δ-receptors than to the κ-receptors [677].

The substitution of D-Trp in DP-A-11 gives rise to weak, nonselective antagonists [678]. *N*-diallyl derivatives of DP-A-11 are also opioid antagonists with weak selectivity for κ-receptors [679]. De-Tyr^1-rimorphin inhibits morphine-induced effects [680].

Uses. The intrathecal application of 15 µg of DP-A-13 to cancer patients produces a nociceptive effect that lasts for >4 h [681]. [D-Ala^2]DP-A-6 [*081733-79-1*] (Dalargin) is used to treat duodenal peptic ulcers [682].

4.4.3. Dermorphin and Deltorphins

Occurrence. Dermorphin (DM) [*77614-16-5*], M_r 803, and deltorphin (DT) [*119975-64-3*], M_r 955, each contain a D-amino acid:

Dermorphin	Y	a	F	G	Y	P	Sa
Deltorphin	Y	m	F	H	L	M	Da

They were isolated from the skin of the South American frogs, *Phyllomedusa sauvagii*, *P. rhodei*, and *P. bicolor*. One precursor of DM contains five copies of a sequence of 35 amino acids, the C-terminus of this sequence contains the DM sequence [683]. Another precursor contains four copies of DM and one copy of DT [684]. The precursors contain alanine and methionine in the L form, which are converted to the D conformation in a posttranslational reaction.

Immunoreactive DM has also been found in the brain, adrenal glands, and the gastrointestinal tract of the rat [685].

Biological Effects. DM binds preferentially to the μ-receptors. When given intravenously, it has a powerful analgesic effect (10 times that of morphine).

DM inhibits the secretion of gastric acid, the emptying of the stomach, secretion from the pancreas, and intestinal motility. It increases the plasma levels of prolactin [686], growth hormone [687], thyrotropin [688], somatostatin, gastrin, glucagon [689], blood pressure, and heart rate. It inhibits the release of ACTH, β-LPH, β-EP [690], LH [691], secretin, and pancreatic peptide [692].

DT binds to the δ-receptors. When given intracerebroventricularly it improves the memory of mice [693].

Structure – Activity Relationships. The C-terminal amide group is important for the biological activity of DM [694]. In the guinea pig ileum, [Tyr^7]DM has twice the activity of DM and 1.4 times the analgesic activity [695]. The substitution of Hyp for Pro^6 gives a compound with the same analgesic and gastrointestinal activity, but a higher prolactin-releasing activity [694]. [$Tyr(OMe)^5$]DM has a higher affinity for the μ-receptors and [Phe^5]DM binds more strongly than DM to the δ-receptors [696]. [D-Arg^2]DM has about the same analgesic activity as DM [697]. Shortening the [D-Arg^2]DMs at the C-terminus enhances activity [697]. DM guanylated at the N-terminus has a higher analgesic effect and inhibits the gastrointestinal transit [698].

In the case of the tetrapeptide analogues, amides with a bulky side chain, N-terminal guanylated compounds, and the [Sar^4], [D-Arg^2], and [D-$Met(O)^2$] substitutions all increase activity. Tetrapeptide analogues are described in [699–706].

The selectivity of DM and DT for the opioid receptors depends on charge effects and the hydrophobicity of the C-terminus [707]. DT-(1 – 4)amide binds almost exclusively to the μ-receptors. Met^6, Asp^7, and Leu^5 are important for binding to the δ-receptors.

4.4.4. Exorphins

The term exorphins refers to peptides with opiate activity that are produced from food during digestion.

Occurrence. β-Casomorphin-7 (β-CM-7) [*79805-24-6*], M_r 790, was isolated from casein peptone and corresponds to b- or o-β-casein-(60 – 66) [708]. The more active N-terminal β-CM tetrapeptide amide (morphiceptin) has been synthesized and isolated from enzymatically digested milk proteins [709].

α-Casomorphins were isolated from α-casein after treatment with pepsin (amino acid sequences 90 – 95 and 90 – 96). An exorphin-like sequence was found between positions 43 and 49 of α-gliadin [710].

α-CM-7	R Y L G Y L E
α-CM-6	R Y L G Y L
β-CM-7	Y P F P G P I
β-CM-5	Y P F P G
β-CM-4	Y P F P
Morphiceptin	Y P F P a

Biological Effects. The β-CMs have only slight opiate activity. Intragastric β-CMs cause the release of somatostatin. In a process that can be reversed by naloxone, digested gluten and β-CM (oral) stimulate the release of insulin and glucagon in dogs after a test meal [711].

The β-CMs exert an opioid effect on intestinal electrolyte transport [712]. When applied parenterally, they stimulate the postprandial release of pancreatic peptide and the amino-acid- or glucose-stimulated release of insulin [713], increase the plasma level of prolactin [714], and inhibit the release of somatostatin, thyroliberin, and thyrotropin [715].

Structure – Activity Relationships. The N-terminal tripeptide of β-CM has no activity. Morphiceptin, the N-terminal tetrapeptide amide, has the same analgesic activity as Met-EK in the guinea pig ileum assay and is a specific ligand for the μ-receptors. The peptides, h-β-CM-(1 – 4) amide (valmuceptin) and h-[D-Val^4]β-CM-(1 – 4)amide (devalmuceptin) bind more tightly to the μ-opiate receptors than morphiceptin [716].

Substitution of D-Pro or D-pipecolic acid for Pro^4 gives [D-Pro^4]β-CM-5 (deprolorphin), [D-Pro^4]morphiceptin (deproceptin, Wellcome PL 017) [717], and [D-pipecolic acid^4]β-CM-5 (depilorphin), which are more active than morphine both in vivo and in vitro. These compounds have a high analgesic activity in the rat [718]. Substitution of D-Phe for Phe^3 increases antinociceptive activity [718]. Replacement of D-Pro^2 leads to loss of opioid activity, whereas [D-pipecolic acid^2]β-CM-5 has a higher long-lasting analgesic effect [718]. Substitution of D-Pro for Pro^4 and D-Phe for Phe^3 promotes binding to the μ-receptors.

4.4.5. Femarfarmamide and Related Structures

Occurrence. The tetrapeptide femarfarmamide (FMRF-amide), M_r 599, was isolated from mollusks. It is not derived from prepro-EK A (which also contains the FMRF sequence at its C-terminus) but is coded in a gene that contains 21 copies of the FMRF sequence [719].

Immunocytochemical methods have shown that FMRF-like material occurs in the pancreas of chicken, ileum of the dog, and in brain neurons of the frog and rat. The first FMRF-like peptide from vertebrates, LPLRF-amide, was isolated from chick brain.

Other structurally related compounds are found in many animals (e.g., cockroach [720] and hawk moth [721]).

Biological Effects [722]. In mollusks, FMRF-amide exerts both a stimulating and an inhibitory effect on the heart. It hyperpolarizes neurons in the snail. In rats, both FMRF-amide and LPLRF-amide [723] (intravenously and centrally applied) increase arterial blood pressure.

FMRF-amide inhibits the spontaneous or acetylcholine-induced contraction of the anterior gizzard of *Aplysia california* and the stomach of *Navanax* [724]. It also inhibits colon motility in mammals [722].

FMRF-amide has no opiate activity, but acts as an opiate antagonist, e.g., [725], [726]. However, it only binds weakly to κ- or μ-receptors [727].

On intracerebroventricular application in rats, FMRF-amide exhibits amnesic activity

[728] and increases the plasma level of growth hormone [729].

Structure – Activity Relationships. The C-terminal amide group and the full length of the molecule are important for the biological activity of FMRF-amide. Acetylation and benzoylation increase its contractile effect. A hydrophobic amino acid is required in position 2, while short-chain amino acids decrease activity [730].

5. Blood Pressure Regulating Peptides

This chapter deals with the angiotensin–kinin system, the neurokinins and tachykinins, vasopressin and the structurally related peptides, oxytocin and vasotocin, the endothelins, the atrial natriuretic factor (ANF), and neuropeptide Y (NPY). These peptide hormones influence blood pressure and function as modulators of the reproductive hormones (Chap. 2) and of the stress-induced proopiomelanocortin (POMC) cascade (Chap. 4). AT II and, above all, the kinins appear to be involved in ovulation and the formation of semen. AT II stimulates the POMC cascade, presumably via CRH, and the formation of gluco- and mineralocorticoids; it also raises the level of luteotropin (LH) when applied centrally. Vasopressin has a peripheral stimulating effect and inhibits the POMC cascade when administered centrally. Oxytocin plays a role during mating as well as during and after birth. Endothelin releases LH from the pituitary; ANF, which can be released by endothelin, inhibits the formation of the stress hormone cascade and the synthesis of aldosterone. NPY also exerts an effect on the secretion of LH through the release of LHRH from the hypothalamus.

5.1. The Angiotensin – Kinin System

The kinins and angiotensins are formed in biological fluids by the enzymatic cleavage of protein precursors. Kallikreins cleave kininogens (KG) to yield kinins, which are blood pressure lowering peptides. The species-specific enzyme renin can also be generated by kallikrein from inactive prorenin and is responsible for the formation of the decapeptide angiotensin I (AT I) from the N-terminus of the α-globulin angiotensinogen (ATG). The structures of r-prepro-ATG and h-prepro-ATG have been elucidated [731]. Glucocorticoids stimulate the formation of ATG [732], [733], IL-6 potentiates the glucocorticoid-stimulated synthesis of ATG [733], and AT II or [Sar^1]AT II inhibits the production of ATG in hepatocytes [734]. Although AT I has no effect on blood pressure, it is acted upon by the membrane-bound angiotensin converting enzyme (ACE or kininase II), a zinc-containing carboxydipeptidase, to form the octapeptide angiotensin II (AT II), a vasopressor (i.e., blood pressure increasing agent). The hypotensive kinins are, however, degraded by ACE. Therefore, the inhibition of ACE has a double hypotensive effect because both the formation of the blood pressure increasing AT II as well as the degradation of blood pressure lowering kinins are inhibited. The first inhibitors of ACE were isolated from snake venom and called bradykinin potentiating peptides. Examples of orally active ACE inhibitors are captopril, enalapril, and ramipril. Inhibitors of renin also exert a hypotensive effect. Potent inhibitors of renin are ATG analogues in which the peptide bond between positions 10 and 11 is reduced or replaced by statin, an amino acid occurring in pepstatin, or by similar compounds.

5.1.1. Angiotensins

Occurrence. Angiotensins (ATs) are found both in the periphery and in the brain. Plasma AT II is formed primarily in the lungs. AT II and AT III are also synthesized in the juxtaglomerula cells of the kidney [735], the adrenal cortex [736], the gonadotropic cells of the pituitary [737], the hypothalamus [738], the ovarian follicle [739], and the Leydig cells of the testes. Roughly equimolar amounts of AT II-(1 – 7), AT II, and AT I have been found in the rat hypothalamus [738]. The sequences of human and rat angiotensins are as follows:

D R V Y I H P F H L
← AT I →
← AT II →
← AT III →

Release. See Figure 4. AT I [*9041-90-1*], M_r 1296.5, is formed by the action of renin on α-globulin angiotensinogen (ATG). AT I is in turn cleaved by the angiotensin converting enzyme (ACE), also known as kininase II, or carboxypeptidases to form AT II [*11128-99-7*], M_r 1046.2 [740]. AT III [*12687-51-3*], M_r 931.1, is formed from AT II (aminopeptidase A) or de-Asp^1-AT I (ACE or carboxypeptidases). Large amounts of aminopeptidase A are found in the adrenal cortex.

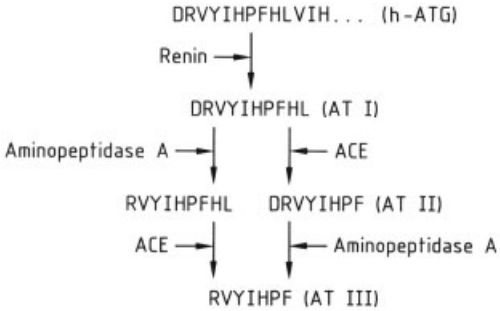

Figure 4. Formation of angiotensins (AT) from angiotensinogen (ATG)

Estradiol or a combination of estradiol and progesterone increases the plasma level of AT I and II in rats, whereas progesterone alone exerts no such effect [741]. A low-NaCl diet raises and a high-NaCl diet lowers the serum level of AT II [742]. The intracerebroventricular application of sodium chloride appears to result in a pressor effect through secretion of AT II [743]. The plasma AT II level is increased in thirst [744], dehydration, loss of blood [745], and endotoxin shock [746].

Receptors. Two AT II receptors occur in the adrenal gland: The *AT 1 receptor* (adrenal cortex) is sensitive to DuP-753 (Du Pont) [747], the *AT 2 receptor* (adrenal medulla) is sensitive to PD 123 319 (Parke Davis), Exp-655 (Merck MSD/Du Pont) [747], and WL-19 (Werner Lambert) [748].

AT III has its own subclass of receptors in the vascular smooth muscles: [Sar^1,Ile^7]-AT III is a selective antagonist for AT III [749].

The density of the AT II receptors can be increased by dehydration [750], stress [751], or IGF-I [752]. Treatment with AT II leads to the down regulation of the receptors in rat hepatocytes [753]. The density of the AT II receptors on the glomeruli in the kidney is reduced by a low intake of sodium chloride [744], diabetes mellitus [754], and by ACTH-(1 – 24) [755]. Chronic treatment with estrogen reduces the AT II receptor density in the hypophysis, adrenal cortex [756], and the rat placenta [741].

Biological Effects. AT II and AT III increase intracellular levels of calcium, phosphatidic acid, phosphoinositides [757], and cGMP [758], and increase the formation of PGE_2 and PGI_2 by activating phospholipase A_2 [759].

In the kidney, AT II inhibits the activity of renin [760] and the synthesis of messenger RNA for renin by a negative feedback effect through a lipoxygenase product [761].

AT II causes contraction of the vascular smooth muscles of the uterus, intestine, aorta, myocardium, and kidney, resulting in a potent vasoconstricting and blood pressor activity. Prostaglandins counteract the activity of AT II in the kidney [762].

AT II has many other effects. It increases plasma vasopressin levels [763] and exerts a positive inotropic effect on the myocardium [764]; in the kidney it reduces the glomerular filtration rate and, in physiological doses, the excretion of sodium ions and water [765], but it exerts a natriuretic and diuretic effect when given in high doses [766]. It reduces water absorption by the jejunum via the synthesis of prostaglandins [767] and stimulates the formation of aldosterone and corticosterone in the adrenal glands after short-term administration [768]. It may play a role in ovulation [769]: it stimulates the secretion of androgen and estrogen from rat ovaries [739] and reduces the LH-stimulated synthesis of progesterone in bovine luteal cells [770]. Like AT II, AT III stimulates the release of aldosterone and prostaglandin.

Structure – Activity Relationships. The C-terminal carboxyl group and the aromatic amino acids are important for the activity of AT II [771]. The N-terminal sequence is responsible for the specificity, intensity, and duration of its biological activity. However AT II-(1 – 7), like AT II, releases vasopressin and probably plays a physiological role in its formation [772].

h-[Sar^1]AT II, h-[Aib^1]AT II, and [D-Ser^1]AT II [773] are more stable to aminopeptidases and thus more potent than AT II.

The activity is lowered if Asn occupies position 1 [774].

Although the N-terminal amino group is not important for the pressor activity of AT II, acylation decreases the myotropic and aldosterone-stimulating effects.

Tyr^4 is very important for biological activity: fragments without Tyr^4 are devoid of activity [771]. The activity depends on the electronegativity of the aromatic ring in position 4. Increasingly electronegative substituents lower activity. [Sar^1,3-NH_2-Tyr^4]AT II [775] and monoiodo-AT II [776] are, however, more active than [Sar^1]AT II and AT II. A β-branched amino acid in position 5 is important for agonistic activity. *O*-Methylthreonine is more effective than Ile in position 5 [777].

If Pro^7 in AT I or AT II is replaced by Ala, cardioselective substances are obtained which have a positive inotropic activity and lower pressor effect [764]. [Cys^3,Cys^5]AT II is a cyclic analogue with about the same activity as AT II [778].

AT II Antagonists. Positions 4 and 8 are responsible for the agonistic or antagonistic action of AT II. Competitive antagonists result when the side chain of Phe^8 is extended or shortened by a methylene group or when aliphatic or branched aromatic [779] amino acids are substituted for Phe^8. One of the most active antagonists of this series is [Sar^1,Bip^8]AT II [779] (Bip = biphenylalanine). The presence of D-amino acids, such as D-phenylglycine (D-Phg) or D-phenethylglycine (D-Peg), in position 8 gives active antagonists, which are not degraded by carboxypeptidases [780]. [Sar^1,Val^5,Ala^8]AT II (saralasin) and [Sar^1,Ile^5,Thr^8]AT II (DE-3489, sarthran) have been examined more closely. [Sar^1,$Thr(Me)^{5,8}$]AT II and [Sar^1,Chg^5,Lac^8] AT II [781] (Chg = L-cyclohexylglycine, Lac = L-lactic acid) are more active than saralasin. Unlike the agonists, the antagonists do not require a β-branched amino acid in position 5. For instance, [Sar^1,Tyr^5,Ile^8]AT II has three times the antagonistic activity of [Sar^1,$Ile^{5,8}$]AT II [777].

Alkylation of the phenolic hydroxyl group of Tyr^4 [782] or the substitution of Phe [783] or 4-ClPhe [784] for Tyr^4 also gives AT II antagonists. However, [Sar^1,$Tyr(Me)^4$]AT II (sarmesin) is less active than saralasin.

Shortening the molecule at the C-terminus also yields AT II antagonists. Saralasin and sarmesin are about six times as active as [Sar^1] AT II-(1 – 7)amide [785]. The latter compound has no agonistic properties and does not inhibit the central dipsogenic activity of AT II [786].

Nonpeptide AT II receptor antagonists are imidazole derivatives and are described in [747], [787], [788].

AT III Analogues. Substitution of D-*N*-methylalanine for Arg^1 in AT III produces a more potent analogue. [D-*N*-$MeAla^1$,Ile^7]AT III is a selective AT III antagonist [789]. [Sar^1, Ile^7]AT III is a selective vascular AT III antagonist. [Ile^7]AT III inhibits both AT III and AT II [749].

Uses. [Asn^1,Val^5]AT II (angiotensin amide [*53-73-6*], Hypertensin CIBA, Ciba), an AT II agonist, is administered in states of shock and collapse when normal blood pressure should be restored as quickly as possible.

The AT II antagonist saralasin [*34273-10-4*] (Sarenin, Röhm Pharma) is used in the diagnosis of AT II-dependent forms of hypertonia and in the preliminary treatment of donor kidneys before transplantation to minimize loss of functions of the transplanted kidney caused by ischemia.

5.1.2. Kinins [790]

Occurrence. Kininogens (KG) are cleaved by kallikrein to give kinins. This system is widely distributed, e.g., in the liver, in acinar cells of the rat submandibular gland, in the human kidney, in platelets, endothelial cells, rat vascular smooth muscle cells (important for the regulation of the local vascular tone) [791], in the brain [792], Leydig cells, ovaries, and in milk [793]. The amino acid sequences of mammalian kinins follow:

Bradykinin (BK) [*58-82-2*], M_r 1060		R	P	P		G	F	S	P	F	R	
[Hyp^3]BK		R	P	Ph		G	F	S	P	F	R	
Kallidin [*342-10-9*], M_r 1188	K	R	P	P		G	F	S	P	F	R	
[Hyp^3]Lys-BK	K	R	P	Ph		G	F	S	P	F	R	
Met-Lys-BK	M K	R	P	P		G	F	S	P	F	R	
T-Kinin [*86030-63-9*], M_r 1260	I S	R	P	P		G	F	S	P	F	R	
T-Kinin-Leu	I S	R	P	P		G	F	S	P	F	R	L

Large amounts of BK, kallidin, [Hyp^3]Lys-BK, and [Hyp^3]BK were found in human urine [794], plasma [795], and in ascites of patients suffering from stomach cancer [796].

T-Kininogens have also been found in the rat [797]. These precursors are cleaved by T-kininogenase [798] (endopeptidase K [799]) or high concentrations of trypsin to give T-kinin (Ile-Ser-BK).

Structural similarities have been observed between kinins and various hormones of insects and crustaceans [800], [801].

Release. Injury lowers tissue respiration and thus the pH [802] which activates the Hagemann factor (the blood clotting factor XII) resulting in the release of bradykinin (BK) [803]. Kallekrein in the plasma cleaves kininogen to produce BK, whereas tissue kallikrein cleaves kininogen to form kallidin (KD = Lys-BK), which can in turn be converted to BK by aminopeptidases.

Kinins are released during muscular work [804], PMSG/h-CG induced ovulation [805], and inflammation.

Prostaglandins inhibit the formation of kinins. Inhibitors of prostaglandin synthesis therefore potentiate BK synthesis [806]. Aspirin, for instance, potentiates the antihypertensive effect of Captopril in spontaneously hypertensive rats [807].

Inhibitors of kallikrein retard the release of kinins. Endogeneous kallikrein inhibitors include α-macroglobulin and aprotinin (Trasylol, Bayer; Antagosan, Behring Werke). Synthetic inhibitors are substrate analogues of the kininogens, an example is D-Chg-L-Chg-L-Arg-4-nitroanilide, a structural analogue of the C-terminal cleavage sequence, Pro-Phe-Arg-Ser [808]. The bradykinin B_2-receptor antagonists also inhibit kallikrein [809].

BK is deactivated primarily in the lungs by the dipeptidase kininase II, which is identical to ACE. Inhibitors of ACE (e.g., Captopril, Squibb; Enalapril, Merck; Ramipril, Hoechst) potentiate BK [810–812]. Other endopeptidases, carboxypeptidases (kininase I activity), and aminopeptidase degrade BK [813–815].

In patients suffering from hypertension, the concentrations of kallikrein, prekallikrein, and kinin in the urine are significantly lower than in persons with normal blood pressure [816]. Captopril decreases blood pressure in hypertensive patients with a normal urine kallikrein content, but not in patients who have a low urine kallikrein content [817]. Kallikrein activity decreases in old age.

An increased level of kinin or kallikrein is found in diabetics with orthostatic hypotension [818], allergic persons [819], and patients suffering from inflammatory stomach diseases, ulcerative colitis, pancreatitis, rheumatic inflammation, or myocardial infarction.

Only in rats, do prepro-T-kininogen messenger RNA [820] and T-kininogen [821] appear to be formed primarily in response to inflammatory stimulants. Expression of the T-kininogen gene increases with age [822].

Receptors. The B_1- and B_2-receptors for BK have been studied intensively. The B_1-receptors bind de-Arg^9-BK more strongly than BK, while the B_2-receptors have a greater affinity for BK. Studies with agonists and antagonists on different tissues have led to the assumption that there are multiple B_2-receptors for kinins [823]. The vasorelaxing effects of BK exerted via B_2-receptors in isolated large vessels require intact vascular endothelial cells. The vasoconstricting effects of BK do not require an intact endothelium. The BK-receptor complex is internalized, causing temporary down regulation of the BK-receptor [824].

Biological Effects [825].

B_1-receptor-mediated effects. The most important B_1-receptor-mediated effects [826] are contraction of rat duodenum, relaxation of rabbit coeliac artery [827], capsaicin-induced inflammation in mice [828], and release of IL-1 and TNF-α in macrophages.

B_2-receptor-mediated effects [826] include release of EDRF and calcium from intracellular stores, leading to activation of PLA_2 and increase in PGE_2 production; the PGE_2-dependent activation of adenylate cyclase in arterial smooth muscle cells [829]; hypotensive activity: reduces blood pressure increased by AT II [830], adrenaline [831], and vasopressin [790]; increase of vascular permeability and pain [832] in rat skin; improvement of heart function [833]; and increase in human sperm motility.

Other effects of the kinins include stimulation of ovulation [805], [834], rhinitis [790], re-

lease of SP- and CGRP-like materials in neurons sensitive to capsaicin [835], and glucose uptake in working skeletal muscle via prostaglandins and phosphofructokinase [836].

Structure – Activity Relationships. De-Arg^9-BK and de-Arg^{10}-Met-kallidin (stronger) are ligands for B_1-receptors. The C-terminal arginine residue is very important for binding to the B_2-receptors. De-Arg^9-[D-Phe^8]BK has four times and Sar-[D-Phe^8]de-Arg^9-BK ten times the effect of de-Arg^9-BK on B_1-receptors [837].

De-Arg^9-[Leu^8]BK, the corresponding KD derivative, de-Arg^9-[Gly^7]BK, and de-Arg^9-[D-Ala^7]BK [838] are powerful B_1-receptor antagonists.

The action of KD on B_2-receptors is greater than that of BK. The guanido groups of Arg^1 and Arg^9 and the proline residue in positions 2 and 3 of BK are important for biological activity at the B_2-receptors. Strong agonists are obtained by the substitution of β-(2-thienyl)-Ala (Thi) or dehydro-Phe (Δ-Phe) for Phe^5 and Phe^8, Hyp for Pro^3, α-aminoisobutyric acid for Pro^7, Tyr(Me) for Phe^8, or by reduction of the peptide bonds between positions 8 and 9 [839] and 6 and 7 [840]. The effect of [Hyp^3,Tyr(OMe)^8]BK on B_2-receptors is about three times that of BK [837]. [Phe^8-ψ-(CH_2-NH)-Arg^9]BK is a selective B_2-receptor agonist, which is five times as active as BK [839]. The effect of [Δ-Phe^5]BK on the uterus and ileum is twice that of BK and the effect on blood pressure is 23 times that of BK.

The replacement of Pro^7 in BK by D-Phe, D-Nal, D-Pal, or D-tetrahydroisoquinolinecarboxylic acid (D-Tic) [841], and the substitution of octahydroindolecarboxylic acid (Oic) [841] for the Phe^8 gives strong B_2-receptor antagonists. Depending on the test, D-Arg-[Hyp^3,Thi^5, D-Tic^7,Oic^8]BK (Hoe 140) [841] is a 100 – 1000 times stronger BK antagonist than the standard peptide, D-Arg-[Hyp^3,Thi^{5,8},D-Phe^7]BK [842].

Modification of the BK antagonists at positions 1, 2, 3, 8, and 9, alters tissue selectivity; extension with D-Arg at the N-terminus improves affinity and inhibits enzymatic degradation [843].

BK-receptor antagonists with Phe^8 or Thi^8 also act as B_1-receptor blockers [844] because carboxypeptidases convert these peptides to biologically active B_1-receptor antagonists by cleaving the C-terminal arginine [845]. The histamine-releasing effect of BK antagonists can be lowered by acetylating the N-terminal amino group [846].

Uses. BK antagonists may find application for treating inflammation, pain, rheumatoid arthritis, osteoarthritis, inflammatory stomach diseases, rhinitis, asthma, and gout [790].

5.2. Substance P, Neurokinins, and Tachykinins [847]

5.2.1. Introduction

Occurrence. The neurokinins include substance P (SP) [*516-47-2*], neurokinin A (NKA) [*86933-74-6*], and neurokinin B (NKB) [*86933-75-7*]. Similar compounds which occur in cold-blooded animals are called tachykinins.

h-,b-Substance P	R P K P Q Q F F G L Ma, M_r 1348
h-,g-,p-Neurokinin A	H K T D S F V G L Ma, M_r 1133
p-Neurokinin B	D M H D F F V G L Ma, M_r 1210
Scyliorhinin II	S P S N S K C P D G P D C F V G L Ma, M_r 1851

Substance P was the first of a large number of peptides that were found in the gastrointestinal tract and the brain. Substance P and the neurokinins are produced in the form of preproneurokinins. Three preproneurokinins (α, β, and γ) have been characterized in the cow and the rat [848]. There is a very great similarity between h-β-prepro neurokinin and b-β-prepro neurokinin [849]. Neuropeptide K (NPK) is neurokinin A that is extended by 26 amino acids at the N-terminus [850]. NPK, NKA [851], NKA-(3 – 10), and NKA-(4 – 10) [852] occur in the plasma and tumor tissue of carcinoid patients and in the bronchi. Apart from NKA, NKB (neuromedin K) has also been isolated from the spinal cord of the pig.

Substance P occurs in the spinal cord, the sensory nuclei of the brain stem, and the sensory nerve terminals in the gastrointestinal tract, urogenital tract, bile duct, bronchi, skin [853], and thymus [854]. Neurokinin A has been isolated from the rabbit iris [855].

Tachykinins are found in the salivary glands of the octopus, the skin of amphibians [856], and the locust [857]. Scyliorhinin I and II are the tachykinins of the mud fish, *Scyliorhinus caniculus* [856]. There are indications that tachykinin-like substances also occur in mammals [852]. There is a structural similarity between the tachykinins and the neurotoxic β-amyloid protein-(25 – 35) [858].

Release. Substance P is released after stimulation of the sensory neurons. It is assumed to be a pain transmitter, especially since the level of SP in the spinal cord is increased by mechanical pressure or hot water applied to the extremities [859], and analgesics inhibit the release of SP from sensory nerve ends [859].

Substance P and NKA are released by capsaicin, bombesin [860], CCK-8 [861], and by cholinergic [862], serotoninrgic [863], and dopaminergic [864] agents.

A high NKA- and SP-like immunoreactivity is found in tumor tissue and plasma NKA immunoreactivity can be used as tumor marker. "Flushing" episodes and diarrhoea correlate positively with plasma NKA [865]. The neurokinin level is also elevated in the joint fluid of patients with rheumatic inflammatory diseases [866].

In the brain, SP is preferentially cleaved by metalloendopeptidase E.C.3.4.24.11 (also called enkephalinase or neutral endopeptidase) to form the N-terminal heptapeptide SP-(1 – 7). The C-terminal heptapeptide SP-(5 – 11) is, however, also formed in the brain after cleavage with the post-proline-cleaving enzyme [867].

Digestion in the plasma occurs preferentially through a post-proline-cleaving enzyme to give dipeptides [868]. SP is also a substrate for the angiotensin converting enzyme (ACE). Levels of SP are increased by smoking (inhibition of the neutral endopeptidase) [869] and by ACE inhibitors [870].

Receptors. Three neurokinin receptors have been found: NK1 for SP, NK2 for NKA, and NK3 for NKB [871]. The r-NK1 receptor, the r-NK2 receptor [872], and the b-NK2 receptor [873] have been characterized via their cDNA. They occur in a wide variety of tissues [874].

Selective ligands of neurokinin receptors are described in the literature [874]:

NK1 receptor agonists	[871], [877–879]
NK2 receptor agonists	[878], [880–882]
NK3 receptor agonists	[871], [876], [883]
NK1 receptor antagonists	[884–886]
NK2 receptor antagonists	[875], [887–889]
NK3 receptor antagonists	[859], [890]

Biological Effects [874].

NK1-receptor-mediated effects. The most important NK1-receptor-mediated effects are hyperalgesia, hypotension, salivary secretion, and increase of capillary permeability.

NK2-receptor-mediated effects include bronchoconstriction [850], [891], [892], activation of guinea pig alveolar macrophages [893], protection against gastric lesions in rats [894], and tachycardia.

NK3-receptor-mediated effects are analgesia [895], bradycardia, and increase in capillary permeability.

Other effects of SP include activation of phospholipase C and D [896], stimulation of the proliferation of synoviocytes and the release of PGE_2 and collagenase (rheumatoid arthritis) [897], modulation of gastrointestinal motility [898], [899], and immunostimulating activity [900], [901].

Structure – Activity Relationships. In the brain, SP-(5 – 11) is a pain neurotransmitter. It causes aggressive, stress-induced behavior, impairs memory, and increases blood pressure when applied centrally. SP-(1 – 7) exerts an analgesic effect, increases learning capacity, improves memory, and lowers blood pressure [867], [902], [903]. Based on the opposing effects of the N- and C-terminal peptides of SP, STEWART postulated the existence of SP-N and SP-C receptors [867]. The regeneration of nerve fibers is stimulated by N-terminal fragments of SP [904].

SP Agonists. The shortest active SP analogues are acylated SP-(7 – 11) derivatives [905], some are more active than SP. The most effective compound in the guinea pig ileum test is SP-(7 – 11) with a 4-hydroxyphenylacetic acid residue at its N-terminus.

The C-terminal amide group, a neutral lipophilic amino acid in position 11 [906], Leu^{10}, and Gly^{9} are important for biological activity. Phe^{8} can be replaced by Tyr(OMe), cyclohexylalanine, or Ile without loss of activity, whereas substitution of these residues for Phe^{7}

produces only slightly active compounds. The remaining N-terminal amino acids are not essential for biological activity.

pGlu-Gln-Phe-MePhe-Sar-Leu-Met-amide is an SP analogue with a long-lasting effect. It has only 10 % of the spasmogenic activity, but all of the aggressive CNS activity of SP and is resistant to the proteolytic enzymes of the hypothalamus.

A derivative that is dimerized with succinic acid via the N-terminal amino groups of SP-(3 – 11) has 2.4 times the receptor affinity and 75 times the saliva-producing effect of SP [907].

NKA-(4 – 10) is equally as active as NKA. Asp^4, Phe^6, and Val^7 are important for the biological activity of NKA and NKB [908].

SP Antagonists. The standard antagonist is spantide I which binds primarily to the NK1 and NK2 receptors.

Spantide I	[D-Arg^1,D-$Trp^{7,9}$,Leu^{11}]SP
Spantide II	[D-$Lys(Nic)^1$,Pal^3,D-$Phe(Cl_2)^5$,Asn^6,D-$Trp^{7,9}$, Nle^{11}]SP

D-Phe(Cl_2) = 3,4-D-dichlorophenylalanine
D-Lys(Nic) = *N*-nicotinoyl-D-lysine

The activity of SP antagonists is enhanced by the replacement of Gln^6 by Asn^6 [909]. As a result of the substitution of D-Lys(Nic) for D-Arg^1, spantide II, unlike spantide I, has no neurotoxic properties [910] and a reduced histamine-releasing activity [911]. Spantide I acts as a bombesin antagonist but spantide II has no effect on bombesin-induced contractions. Spantide II binds to all three neurokinin receptors [911] and exerts a short antinociceptive effect when applied intrathecally [912].

Reduction of peptide bonds also gives SP antagonists which have been evaluated by the displacement of SP from SP receptors in guinea pig acinar cells [913].

SP antagonists with basic N-terminal amino acids release histamine, whereas shortened antagonists (e.g., [D-Pro^4,D-$Trp^{7,910}$]SP-(4 – 11) can inhibit the SP-induced release of histamine in vitro.

[D-Arg^1,D-Phe^5,D-$Trp^{7,9}$,Leu^{11}]SP was the most suitable peptide for inhibiting the bombesin- and vasopressin induced mitogenesis in small cell lung cancer cells. It has the same affinity for the bombesin and SP receptors and is ten times more active than spantide I on the bombesin receptor [914]. Under comparable conditions, spantide I and other N-terminal-shortened SP antagonists did not inhibit the bombesin- or vasopressin-induced growth of Swiss 3T3 cells.

5.2.2. Amyloid A4 Protein [915]

Occurrence. In Alzheimer's disease and in Down's syndrome (Mongolism), fibrillary amyloid is found within the cortical neurons as neurofibrillary tangles and extracellulary in meningeal and intracortical blood vessels as amyloid plaques. These deposits also occur in the skin, subcutaneous tissue, and in the intestine of Alzheimer patients and healthy elderly persons [916].

The plaques mainly consist of an insoluble peptide called the amyloid A4 protein (A4) [*12627-51-9*], M_r 4514 [917], or amyloid-β protein [918]. It contains 42 amino acids [919]. The gene that codes for A4 produces at least three messenger RNAs for amyloid precursor protein, they form the APP_{695}, APP_{751}, and APP_{770} proteins [920].

APP_{695} consists of 695 amino acids and contains A4 in positions 597 – 638. APP_{751} is identical to APP_{695} apart from an insert of 56 amino acids (HL 124i) C-terminal from Arg^{288} and the substitution of Ile for Val^{289}. Apart from an insert of 19 amino acids C-terminal from HL 124i and the substitution of Leu for Ile^{289}, APP_{770} is identical to APP_{751}. The HL 124i insert has the structure of a Kunitz inhibitor, which specifically inhibits serine proteases, such as trypsin, chymotrypsin, elastase, plasmin, and cathepsin G. In Alzheimer's disease, the total increase in the messenger RNA for APP in neurons from the locus ceruleus and nucleus basalis is due solely to an increase in APP_{695} that lacks this inhibitor domain. Therefore, it is assumed that A4 is readily formed from this precursor by the action of cerebral proteases [920]. These A4 precursors are transmembrane proteins and receptors for neuronal adherons [921]. Adherons are extracellular, adhesion-mediating particles (mostly glycoproteins) found in the extracellular matrix.

The gene for APP is expressed in certain neurons, some glia cells, and in brain macrophages. The A4 formed in the glia cells and macrophages

is responsible for the extracellular formation of the amyloid plaques [919]. Human platelets contain APP_{751} in high concentrations [922].

The A4 of patients suffering from Alzheimer's disease and Down's syndrome is identical to the intraneural amyloid of the Parkinson dementia of Guam and the vascular amyloid of sporadic cerebral amyloid angiopathy. The amyloid protein of patients with leptomeningiosis haemorrhagica interna is shorter by three amino acids at the C-terminus [923].

Release of APP and A 4. The APP gene is localized on chromosome 21 and amplified in Down's syndrome but not in Alzheimer's disease. It is assumed that the A4 protein is formed by increased proteolytic activity in Alzheimer's disease [924]. Proteolysis of APP in membranes at the C-terminus of A4 does not change the aggregation properties. However, enzymatic cleavage at the extracellular N-terminus of A4 is of great importance for plaque formation. Proteinase K cleaves a pro-A4 fragment to peptides of the size of A4 [925]. Platelets also release APP during degranulation [926].

Biological Effects. APP is widely distributed in the neurons of the rat brain and appears to play a role in cell – cell contact, which is required for memory [917]. APP released from cells has an autocrine effect on growth regulation [927].

Low concentrations of A4 and A4-(25 – 35), which is structurally similar to the tachykinins, have a neurotrophic and at higher concentrations a neurotoxic effect on mature neurons. This neurotoxic activity is also exhibited by some tachykinin antagonists and can be abolished by tachykinin agonists [928].

5.3. Vasopressin, Oxytocin, and Melanostatin

5.3.1. Vasopressin and Oxytocin

Occurrence. The nonapeptides oxytocin (OT) [*50-56-6*], M_r 1007, and vasopressin (VP) [*11000-17-2*], M_r 1084, are formed as prohormones. OT and VP are located at the N-termini of these prohormones, the residual C-termini are the neurophysins (neurophysin I for OT and neurophysin II for VP). OT and VP are transported together with the neurophysins as transport molecules in neurosecretory vesicles from the pituitary down the axon to the neurohypophysis. OT and VP are also synthesized in the pineal body, the retina of mammals, the thymus [929], corpus luteum [930], and testes [931].

	1	2	3	4	5	6	7	8	9
Oxytocin	C	Y	I	S	N	C	P	Q	Ga
Vasopressin	C	Y	F	Q	N	C	P	R	Ga
Vasotocin	C	Y	I	Q	N	C	P	R	Ga

Vasotocin (VT) [*9034-50-8*], M_r 1050, occurs in the pineal body of mammals and in the cerebrospinal fluid of humans. $N\alpha$-Ac-OT has been found in the pineal body of the cow [932] and rat [933], and, along with $N\alpha$-Ac-VP, also in the pars intermedia and the brain of the rat [933].

Naturally occurring peptides with OT and VP activity are described in [934].

Release. Pneumadin, a decapeptide amide from the lungs, releases VP from the pituitary and has antidiuretic activity [935]:

h-Pneumadin: A G E P K L D A G Va

The plasma VP level increases in response to potassium uptake [936], dehydration and thirst [937], blood loss, intravenous infusion of a hypertonic NaCl solution or hydrochloric acid [938], smoking [939], hypoxia (lack of oxygen), hypercapnia (increase in the CO_2 concentration in arterial blood), and to sexual arousal in men [940]. The plasma level of OT also increases on infusion of sodium chloride [941]. VP release is stimulated by acetylcholine (mediated by the hypothalamus) [942], AT II, AT II-(1 – 7) [943], histamine [944], and dopamine [945]. The activation of central α_1-adrenoceptors increases the release of VP [946], whereas the stimulation of the α_2-adrenoceptors decreases release [947].

A low-NaCl diet [948], water uptake [949], chronic alcohol consumption [950], and GABA [951] decrease the release of VP.

The vasoactive intestinal peptide (VIP) [952], angiotensin II (AT II), corticoliberin [953], fever [954], and endotoxin [955] release both VP and

OT. β-Endorphin and other endogenous opiates inhibit the release of OT and VP.

The plasma level of OT is increased by manual stimulation before milking [956], during labor [957], by CCK, after the intake of food [958], during orgasm in the woman, and ejaculation in the man [959].

Receptors. There are two VP receptors, the V_1-receptor for pressor activity and the V_2-receptor for antidiuretic activity. They are down regulated by VP agonists [960].

The density of the OT receptors in the brain is increased by estrogen or testosterone [961]. Progesterone increases the binding of OT in the posterior part of the ventromedial nuclei of the hypothalamus, the site at which feminine mating behavior is induced by infusion with OT [962].

Biological Effects of Vasopressin.

V_1-receptor-mediated effects. The most important V_1-receptor-mediated effects [963] are stimulation of phosphatidylinositol turnover and aldosterone synthesis in rat adrenal glomerulosa cells [964], vasoconstriction and hypertension [965], stimulation of prostaglandin synthesis in the rabbit kidney [966], promotion of platelet aggregation via serotonin release from platelets [967], and stimulation of the clotting factors vWF and factor VIII [968].

V_2-receptor-mediated effects [963] are stimulation of adenylate cyclase, antidiuretic effect, and induction of hypoosmolality and cerebral edema in rats [969].

Other effects include direct release of ACTH [970], potentiation (VP and OT) of the effect of corticoliberin on ACTH release [970–972], improvement of memory and learning capacity of older persons [973], inhibition of the diuretic [974] and potentiation of the natriuretic effect of ANF [975].

Biological Effects of Oxytocin. Oxytocin stimulates adenylate cyclase and the release of arachidonic acid and $PGF_2\alpha$ from human decidual cells [976]; induces contraction of the uterus and lactation; prevents the fall in blood pressure associated with hemorrhage [977]; releases relaxin; stimulates the release of insulin and glucagon [978]; increases plasma glucose levels [979]; and releases POMC peptides, α-MSH, β-endorphin, and ACTH from human placenta [980].

Biological Effects of Vasotocin [981]. Vasotocin inhibits the stimulating action of pregnant mare's serum on mice uteri and ovaries; lowers emotional arousal in rats [982]; and increases rapid eye movement (REM) sleep in children via serotonin release.

Structure – Activity Relationships. [983]. The ring structure is of special importance for the activity of OT and VP agonists. However, a disulfide bridge is not essential. 1-carba-OT, where the sulfur from Cys^1 is substituted by carbon, is more potent than OT. Extension of the N-terminus (e.g., with triglycine) causes a depot effect. For instance, Gly_3-[Lys^8]VP (terlipressin; Gycylpressin, Ferring) has clear advantages over VP in the treatment of bleeding of esophageal varices. De-9-Gly-amide-VP (Org 5667, Organon) has no cardiovascular or antidiuretic activity, but improves the learning and memory capacity of rats and also has the VP-stimulating effect on penis erection [984]. The memory-enhancing effect of VP is possibly caused by a metabolite.

Substitution of Phe for Ile^3 increases the vasopressor activity at the expense of the OT activity (uterine contraction), the replacement of Arg^8 or Lys^8 by Leu has the opposite effect. Other examples for the alteration of activity as a result of amino acid substitution are given in Table 1. Highly specific analogues can be produced by combining these substitutions. Compounds in which OT activity predominates are deamino^1-dicarba-[Gly^7]OT (cargutocin), deamino^1-1-monocarba-[$Tyr(Me)^2$]OT (carbetocin, weaker but acts longer than OT) [992], [Thr^4,Gly^7]OT, and [Thr^4,Sar^7]OT [993]. Deamino^1-[4-$EtPhe^2$]6-carba-OT (narcartocin) has a higher natriuretic and diuretic activity, but is less effective in initiating uterine contraction than OT [994]. Deamino^1-[D-4-$EtPhe^2$]6-carba-OT is an OT antagonist, which inhibits the uterotonic and galactogogic activity of OT [995].

A powerful antidiuretic preparation with slight pressor activity but with vasodilating activity [996] is deamino^1-[D-Arg^8]VP (dVDAVP, desmopressin, trade name Minirin, Ferring). Deamino^1-[Val^4,D-Arg^8]VP (dKDAVP) lacks pressor activity but has antidiuretic ac-

tivity. Deamino^1-[Tyr(Me)^2,Val^4,D-Arg^8]VP is another antidiuretic compound, but also a strong OT antagonist [997]. Deamino^1-[MeAla^7,D-Arg^8]VP is another antidiuretic with only slight vasopressor activity [986].

Pressor activity is especially pronounced in [Phe^2, Lys^8]VP (felypressin, octapressin) and [Phe^2,Ile^3,Orn^8]VP.

Antagonists [998]. Strong inhibitors with antioxytocin and antivasopressor activity and slight anti-antidiuretic (aquaretic) activity are variants of peptides with β,β-dialkylated or β,β-cyclopentamethylene-3-mercaptopropionic acid (Pmp) in position 1, *O*-alkylated tyrosine in position 2, and ornithine in position 8, e.g., the OT antagonists, [β-mercapto-β,β-diethylpropionic acid^1, Tyr(Me)^2,Orn^8] VT or deamino^1-[Tyr(Me)^2,Orn^8]VT. Relative to [Pen^1]OT and [Pen^1,Phe^2,Thr^4]OT (Pen = penicillinamine), the prototypes for OT antagonists, additional Orn^8 and alkylated Tyr^2 increase duration of action but not potency; D-amino acids in position 2 increase activity but not duration of action. Thus, although [Pen^1,D-Phe(Me)^2,Thr^4,Orn^8]VT is somewhat more potent than [Pen^1,Phe(Me)^2,Thr^4, Orn^8]VT, it has a shorter duration of action [999]. Very strong and selective OT antagonists are [Pmp^1, Tyr(Me)^2,Thr^4,Orn^8,Leu^9]VT with a low antivasopressor and antidiuretic activity and [Pmp^1,Tyr(Me)^2,Thr^4, Orn^8]VT-(1 – 8) with a weak antivasopressor and aquaretic activity [1000].

Other powerful OT antagonists are deamino^1-[D-Tyr(Et)^2,Thr^4,Orn^8]VT (ORF 22 164) [1001] and [MePmp^1]OT [1002]. ORF 22 164 strongly inhibits contraction of the human myometrium, and has a very slight vasopressor and antidiuretic activity [1003].

Deamino^1-[Pen^1,Tyr(Me)^2]VP and [Pmp^1,Tyr(Me)^2] VP (SK&F-100 273) have high antivasopressor activity (V_1 antagonists) [1004]. Highly selective V_1 antagonists with low antidiuretic activity are produced by introducing a 4-methyl or 4-phenyl group in Pmp^1 [1005], [1006].

In these analogues, the aquaretic activity can be increased by substituting D-Phe, D-Tyr, *O*-alkylated D-Tyr, or D-Ile for Tyr^2. The aquaretic–antivasopressor ratio is further improved by Ile^4, Val^4, Abu^4, or Ala^4 [1007], [1008]. Introduction of a *cis*-4-methyl group in Pmp^1 can increase antagonistic activity and greatly reduce the agonistic activity of antagonists of the type [Pmp^1,D-Tyr(Et)^2,Val^4]VP [1009]. The deletion of Gly-amide and Pro^7 does not significantly affect the aquaretic activity of the antagonists [1010]. Replacement of the C-terminal tripeptide by, for example, Arg-NH_2, Arg-D-Arg-amide [1011], or diaminoalkylamides [1012] gives potent V_2 antagonists with aquaretic activity. One of the strongest aquaretic VP antagonists is dicarba-[Pmp^1,D-Tyr(Et)^2,Val^4,Arg^7, D-Arg^8]VP-(1 – 8)amide (SK&F 105 494) [1013]. A very selective V_2 antagonist with an effective dose ratio (V_2:V_1) of 440 is [Pmp^1,D-Ile^2,Ile^4]VP-(1 – 8) [1014]. Nonpeptide structures are also tolerated at the C-terminus [1015].

The cyclic structure is not required for antagonistic activity. Thus, a reduced antagonist with free sulfhydryl groups, [Pmp^1,D-Tyr(Et)^2,Val^4]VP, still has a high aquaretic activity. Adamantylacetyl-[D-Tyr(Et)^2,Val^4,Abu^6,Arg^9] VP-(2 – 9)amide has a higher aquaretic activity than [Pmp^1,D-Tyr(Et)^2,Val^4]VP [1016].

Substitution of sarcosine or *N*-methylalanine for Pro^7 in VP antagonists yields OT antagonists. Thus, [Pmp^1,D-Phe^2,Sar^7]VP or [Pmp^1,D-Phe^2,MeAla^7]VP binds to OT receptors and V_1-receptors, but has only slight antidiuretic activity. The OT and V_1 antagonist, [Pmp^1,D-Phe^2, MeAla^7]VP is a very strong antiglycogenolytic compound with antigalactogenic activity [1017].

Novel OT antagonists are based on a natural product of *Streptomyces silvestris*. L-156,373 [cyclo-(Pro-D-Phe-*N*-OH-Ile-D-piperazyl-L-piperazyl-*N*-Me-D-Phe)] binds to OT receptors with an approximately 20-fold selectivity compared with V_1- and V_2- receptors. Derivatives of this compound are described in [1018–1020].

Uses. Oxytocin (Orasthin, Orasthin "stark", Hoechst; Oxytocin 3 "horm", Oxytocin forte "horm", Hormonchemie; Partocon, Ferring; Pitocin, Parke Davis; Syntocinon, Sandoz) is used to induce birth in primary and secondary uterine inertia, in placenta retention, postpartum hemorrhage, uterine atonia, lactation disturbances, Cesarean section, and mastitis prophylaxis.

Vasopressin (Pitressin, Pitressin Tannat, Parke Davis) is used to treat the bleed-

Table 1. Alteration of the activity of oxytocin (OT) and vasopressin (VP) by amino acid substitution

Amino acid substitution	Effect
Substitution in [Arg^8]VP	
Deamino^1 Tyr(Me)^2, Phe^2, Val^4, Pro^4 [938], Arg^4 [985], Δ^3-Pro^7, 4-Hyp^7, Sar^7, MeAla^7 [986], D-Arg^8, Ala^9, D-Ala^9 [987], Acc^9 [988]	increases ratio of antidiuretic: pressor activity
4-Hyp^7 [981]	increases ratio of contraction of uterus: pressor activity
Ala$^\circ$ [989]	increases ratio of pressor: antidiuretic activity
Substitution in [Lys^8]VP	
Phe^2, Ile^3, Orn^8	increases ratio of pressor: antidiuretic activity
Substitution in OT	
4FPhe^2, Thr^4, Gly^7, thiazolidine carboxylic acid^7, Sar^7,Δ^3-Pro^7 [990]	increases ratio of contraction of uterus: antidiuretic activity
Deamino^1, carba-1, Tyr(Me)^2, carba-6, Gln^8	increases ratio of antidiuretic: contraction of uterus
Asn^4, Val^4, Glu(OMe)^4, β-acetylamino-L-α-aminopropionic acid^4, carba-1, MeAla^7, Gly^7, Acc^7 [990], α-*tert*-butylGly^8	increases ratio of milk ejection: contraction of uterus
Deamino^1, Hmp^1, carba-1 [964], Thr^4 [991]	increases ratio of contraction of uterus: milk ejection activity
Leu^4	diuretic effect, VP antagonist
3-ITyr^2	oxytocin inhibitor

∗ The superscript $^\circ$ denotes N-terminal extension; carba denotes that the sulfur atom of cysteine is replaced by carbon.

ing of esophageal varices and intestinal atonia, to stop bleeding in gynecological operations, and in diabetes insipidus. [Lys^8]VP (lypressin) [*50-57-7*] is available as Postacton (Ferring) or Vasopressin-Sandoz (Sandoz) for pituitary function tests, paralytic ileus, and diabetes insipidus. [Orn^8]VP (ornipressin) [*3397-23-7*] is available as Por 8 Sandoz (Sandoz) and triglycyl-[Lys^8]VP (glypressin, terlipressin [*14636-12-5*]) as Glycylpressin (Ferring). These compounds are used primarily to treat bleeding of esophageal varices, acute uterine bleeding, and in operations. Deamino^1-[D-Arg^8]VP (desmopressin [*16679-58-6*]) is administered as Minirin (Ferring) in diabetes insipidus, after hypophysectomy, for increasing Factor VIII activity in hemophiliacs and in von Willebrand's disease [1021]. Since desmopressin reduces blood loss during operations, it is recommended by Jehovah's Witnesses [1022].

In clinical studies, the V_1 antagonist SK&F-100 273, [Pmp^1,Tyr(Me)^2]VP inhibits the lypressin-induced increase in blood pressure [1023].

The OT antagonist deamino^1-[Tyr(Et)^2,Thr^4, Orn^8]VT can be administered in premature labor [1024].

5.3.2. Melanostatin

Occurrence. Melanostatin (MIF) [*9083-38-9*], M_r 284, has been isolated from extracts of the median eminence and has the structure Pro-Leu-Gly-amide. It is identical to the C-terminal tripeptide of oxytocin (OT).

Biological Effects. MIF inhibits the release of α-MSH from the pituitary. As a weak μ-receptor antagonist, it impedes the antinociceptive activity of morphine, β-endorphin, and morphiceptin [1025]. It reduces aggression [1026], impedes morphine- [1027] or alcohol- [1028] induced tolerance, and has an antiamnesic [1029] and inhibitory effect on tremor induced by oxotremorin in mice. MIF increases the affinity of dopamine agonists for dopamine receptors and reduces dopamine D_2-receptor supersensitivity [1030]. Some of its CNS effects also appear to proceed via dopamine receptors [1031].

Structure – Activity Relationships. Structure–activity studies using the oxotremorin tremor test have shown that Pro can be replaced by a pGlu residue and the amide group can be alkylated. For instance, pGlu-Leu-Gly-*N*(Me)$_2$ has 32 times the potency of MIF. [*N*-Me-D-Leu^2]MIF (pareptide) potentiates L-dopa-induced behavior patterns

and has antiparkinsonian and antidepressive activity. [D-Pip^1]MIF has a stronger antiamnesic activity than MIF [1029]. [3-(+)-thiazolidine-2-carboxamide]MIF, [3-(−)-thiazolidine-2-carboxamide]MIF and [3-L-3,4-dehydroprolinamide]MIF [1032] cause a two- to threefold stimulation of the affinity of dopamine agonists for dopamine receptors and 3(*R*)-(*N*-L-prolylamino)-2-oxo-1-pyrrolidine acetamide is 10^4 times more potent than MIF [1033].

Uses [1034]. MIF and its analogues are reported to exert an antidepressive effect, potentiate the learning capacity, and serve in the treatment of Parkinson's disease (→ Parkinsonism Treatment, Chap. 3.).

5.4. Endothelin [1035]

Occurrence. Endothelin (ET) [*116243-73-3*], M_r 2492 (h-ET 1), is formed in the endothelial cells and occurs in the kidneys [1036], hypothalamus, pituitary [1037], and human milk [1038].

Three different endothelins have been found in humans (h-ET 1, h-ET 2, h-ET 3) and all contain 21 amino acids with two disufide bridges, one between Cys^1 and Cys^{15} and the other between Cys^3 and Cys^{11} [1039]. p-ET [1040] is identical to h-ET 1 and r-ET is identical to h-ET 3. There is great structural similarity between the endothelins and the sarafotoxins from the venom of the snake, *Atractaspis engaddensis* [1041]. The structure of the prepro-compounds of p-ET 1, h-ET 2 [1042], and h-ET 3 [1043] have been elucidated via their cDNA.

Release. Endothelin is released in response to shear stress [1044], thrombin, angiotensin II, vasopressin [1045], TGF-β, TNF-α, and IL-1β [1046]. The release of ET is reduced by EDRF [1047], dehydration [1037], and lack of oxygen (hypoxia) [1048].

The plasma level of ET is low in genetically hypertensive rats [1049].

Plasma ET levels are elevated in patients suffering from essential hypertension [1050], cardiogenic shock, pulmonary hypertension, and in dialysis patients [1051].

Receptors. Two different ET receptors are found in the rat lung with molecular masses of 44 000 and 32 000, respectively. ET 1 and ET 2 bind tightly to the former and ET 3 binds preferentially to the latter [1052].

Treatment with ET [1053] or AT II [1054] lowers the number of ET receptors.

Biological Effects. The endothelins are regulators of blood pressure and circulation [1029]. After a quick drop in blood pressure (via EDRF release [1055]), a strong pressor effect is observed [1056]. They contract vascular and nonvascular smooth muscles [1057]. They stimulate the release of aldosterone and corticosterone in dispersed zona glomerulosa cells [1058], modulate the synthesis of noradrenaline and have a very strong stimulating effect on the release of α-ANP [1059], [1060]. Independent of gonadoliberin receptors, they also stimulate the release of gonadotropins from anterior hypophyseal cells [1061].

Possible pathological conditions arising from an increased formation of endothelin include coronary and cerebral vasospasm, cardiovascular diseases, bronchospasm, atherosclerosis and hypertension, stomach and duodenal ulcers, and acute renal failure [1062].

Structure – Activity Relationships. ET 1 and ET 2 have a higher muscle-contracting activity than ET 3 [1063].

Procompounds of ET 1 that are extended at the C-terminus [p-ET-(1 – 39) and h-ET-(1 – 38)] are 100 times less active than ET in vitro, but almost equally as active in vivo [1064].

The N-terminal amino group, the C-terminal carboxyl group, the carboxyl groups of Asp^8 and Glu^{10}, the aromatic ring in position 14, and Trp^{21} are important for activity [1065]. Loss of Trp^{21} results in a thousand-fold decrease in activity, further shortening of the molecule at the C-terminus reduces the activity still further. The N-terminal 1 – 16 fragment has no vascular activity. ET-(16 – 21) causes contraction of the isolated guinea pig bronchus (ED_{50} = 0.228 µmol/L) but not, of the rat aorta [1066].

Both disulfide bridges (Cys^1-Cys^{15}, Cys^3-Cys^{11}) are important, but not necessary, for activity. Thus, a cyclic derivative with a disulfide bridge between Cys^1 and Cys^{15}, acetamidomethyl protection in positions 3 and 11, and a C-terminal amide have no effect on the arterial blood pressure [1056]. Derivatives with

incorrect disulfide bridges are also less active. [Ala^{1,15}]ET 1 and [Ala^{3,11}]ET 1 have vasoconstrictive activity, but [Ala^{1,3,11,15}]ET 1 is devoid of activity [1067].

5.5. Atrial Natriuretic Factor [1068]

Occurrence. A group of peptides (atrial natriuretic factor, ANF, and atrial natriuretic peptides, ANP) that exert natriuretic (i.e., increased renal sodium excretion), diuretic, and vasodilating effects have been isolated from the atrium of the mammalian heart. They are all derived from a protein precursor with 151 amino acids (prepro-ANF). Prepro-ANF is enzymatically degraded or processed and shortened at its N- or C-terminus. Depending on the species, the first 23 – 25 amino acids at the N-terminus of prepro-ANF belong to the precursor sequence. Asn^1 is the first amino acid in the 126 amino acids of ANF, M_r 13 679 (h-ANF). ANF (also called γ-ANP or cardionatrin IV) was isolated from atrial tissue. In the rat, ANF is stored in the granulocytes of the atrium and not metabolized further [1069]. α-ANP, M_r 3080.5 (h-α-ANP), [ANF-(99 – 126)], and α-ANP-(3 – 28) are found in the circulating blood [1070]. In humans, α-ANP and β-ANP (an antiparallel dimer of h-α-ANP) are found in the plasma [1071]; h-β-ANP can be converted to h-α-ANP [1072]. In the rat, atriopeptin III [r-ANF-(103 – 126)] also occurs as a circulating peptide [1073]. ANF-(1 – 98) and α-ANP are secreted from the cardiocytes of the atrium into the circulation, the former accumulates to a greater extent than α-ANP because of its longer half life [1074]. ANF-(1 – 30) and ANF-(31 – 67) are also found in human serum and show ANF activity [1075]. The structures of ANF from various species are described in [1076].

Immunoreactive ANF has been detected in a wide variety of tissues (e.g., brain [1077], pituitary [1078], and gastrointestinal tract [1079]). ANP-like peptides have also been isolated from the hearts of chicken [1080], eels [1081], and frogs [1082].

An ANF-like peptide, the brain natriuretic peptide (p-BNP) has been found in porcine brain [1083]. The precursors of the BNPs have been characterized through the cDNAs. The BNP structure is located at the C-terminus. b-prepro-BNP is cleaved to b-γ-BNP with 106 amino acids and to smaller forms, such as b-BNP-26 and b-BNP-32 [1084].

Another ANF-like peptide, the C-type natriuretic peptide (CNP), has been found in the porcine brain. The pharmacological spectrum of CNP is similar to that of ANP and BNP .

Release. The calcium-dependent formation of α-ANP appears to proceed via the phosphoinositide and cAMP system. The plasma levels of ANF correlate positively with the pressure of the pulmonary artery and the pressure in the right ventricle of the heart [1087].

The release of α-ANP is stimulated by an increase in volume of the blood, in hypertension, by vasopressin [1088], oxytocin, angiotensin II [1089], adrenaline [1090], by the stimulation of α_1-adrenergic [1091] and α_2-adrenergic receptors [1092], by substance P [1093], endothelin [1094], thyroliberin [1095], thyroid hormones [1096], by the stimulation of muscarinic and cholinergic receptors, by stress [1097], glucocorticoids [1098], mineralocorticoids [1099], NaCl infusion, a high-NaCl diet [1100], and anesthesia [1097].

Plasma ANP levels are especially high in newborn children and premature babies [1101]. In humans the plasma level of α-ANP increases with age [1102], in congestive cardiac defects [1103], attacks of supraventricular tachycardia and atrial bradycardia, Cushing's syndrome [1104], primary aldosteronism [1105], liver cirrhosis [1106], hyperthyroidism [1107], anorexia nervosa [1108], during exercise [1109], in pregnant women, and in chronic kidney failure [1110].

ANF secretion is lowered by hypophysectomy [1111], hypothyroidism [1112], after operations [1113], after a decrease in the plasma volume, after dehydration [1114], and by a low-NaCl diet.

The endopeptidase E.C.3.4.24.11 cleaves h-α-ANP at four sites [1115]. UK 69 578, an inhibitor of endopeptidase E.C.3.4.24.11, increases the plasma level of ANF in healthy persons [1116] and in patients with chronic cardiac defects; it also reduces blood pressure, and increases natriuresis and diuresis [1117].

Receptors. Three different receptors (ANF A, B, and C) have been isolated by mo-

lecular cloning [1118]. ANF A is a membrane-bound guanylate cyclase, it binds ANF and BNP peptides. ANF B is also a membrane-bound guanylate cyclase but binds BNP peptides preferentially. ANF C is coupled to the adenylate cyclase – cAMP system [1119]. The three receptors have a high sequence similarity.

Specific receptors exist in smooth muscle for ANF-(1 –30) and ANF-(31 – 67), these receptors differ from those for ANF-(99 – 126) [1120].

Glucocorticoids [1121] and dehydration increase the number of ANP binding sites [1122]. The receptor density on the blood platelets decreases with a high-NaCl diet and increases with a low-NaCl diet [1123]. The number of ANF receptors in the rat aorta is reduced by pretreatment with α-ANP (receptor down regulation) [1124].

Biological Effects [1125]. Peripheral application of ANF peptides results in vasodilation, reduces the plasma volume [1126], lowers blood pressure, stimulates natriuresis and diuresis, and inhibits the renin – angiotensin system and the stress hormone cascade.

The vasodilating activity of ANF peptides is based on inhibition of the intracellular increase in calcium [1127].

In models that are not dependent on renin, α-ANP reduces the cardiac output and increases the total peripheral resistance. In hypertensive models in which the renin – angiotensin system is stimulated, α-ANP lowers the blood pressure by reducing the total peripheral resistance [1128]. The decreased blood pressure and cardiac output are due to the lower amount of venous blood reaching the heart and the lower blood volume [1129].

In humans, the diuretic activity of α-ANP is highest on infusion with 0.075 $\mu g\, kg^{-1}\, min^{-1}$. The natriuretic activity is highest at 0.015 $\mu g\, kg^{-1}\, min^{-1}$ (2 h). A hypotensive effect is only observed at higher doses. The renal effects disappeared at doses of 0.3 $\mu g\, kg^{-1}\, min^{-1}$ [1130]. In healthy subjects, vasopressin inhibits the diuretic effect of ANF and potentiates its natriuretic effect [1131]. In numerous animal models for kidney failure, ANF peptides improve kidney functions, e.g., [1132].

Other effects of ANF peptides include relaxation of the pulmonary artery [1133] and blood vessels of the lung in patients with chronic obstructive pulmonary disease [1134]; inhibition of adenylate cyclase activity [1135] and stimulation of guanylate cyclase and release of cGMP [1136]; inhibition of the release of CRH and proopiomelanocortin [1137]; inhibition of the release and activity [1138], [1139] of vasopressin (the vasorelaxing [1140] and the diuretic [1141], but not the natriuretic activity of ANF are inhibited by vasopressin); inhibition of the AT II-stimulated formation of aldosterone [1142]; inhibition of the synthesis of pregnenolone, deoxycorticosterone, progesterone, cortisol, and corticosterone; and stimulation of the formation of testosterone in interstitial cells of the mouse [1143].

Structure – Activity Relationships. [1144]. The α-ANP nomenclature is used for small C-terminal molecules and the ANF-(1 – 126) for larger molecules:

h-α-ANP

```
SLRRSSCFGGRMDRIGAQSGLGCNSFRY
1         10        20      28
```

[= h-ANF-(99 – 126)]

α-ANP peptides that are cyclized via a disulfide bridge or by means of α-aminosuberic acid [1145] exhibit good activity. The C-terminal Tyr can be replaced by an amide function. N-terminal extension of α-ANP leads to more active derivatives [1146]: urodilatin, ANF-(95 – 126), has a stronger renal effect than α-ANP [1147]. The aromatic ring of Phe^8, but not that of Phe^{26} is essential for the activity of α-ANP [1148]. Deletion of Gly^9 in h-α-ANP produces an inactive product, whereas full biological activity is retained on substitution of D-Ala for Gly^9 [1145]. Position 12 of α-ANP must be lipophilic (Met or Ile). Asp^{13} is not essential for activity [1145]. α-ANP activity is increased by the phosphorylation of Ser^6 [1149].

The sequence Arg^{11} – Gly^{16} is important for receptor binding and Phe^8 – Gly^{16} for the vasodilating activity [1150].

Superactive analogues are obtained by deleting the N-terminal peptide chain up to Cys^7 and by replacing Cys^7 by 3-mercaptopropionic acid (Mpr), 2-mercaptoacetic acid, or 4-

mercaptobutyric acid [1151]. [Mpr^{7},D-Ala^{9,16}]α-ANP-(7 – 27)amide has about two to three times the biological activity (receptor binding, drop in blood pressure, and natriuresis in rats) of h-α-ANP [1152]. [Mpr^{7},D-Ala^{9}]α-ANP-(7 – 28) has twice the hypotensive effect and five times the cGMP-stimulating effect of α-ANP-(5 – 28) [1153]. r-[D-Tyr^{6}]α-ANP-(6 – 28) (dextronatrin) has the same diuretic activity as α-ANP but a lower hypotensive effect [1154]. [Aoc^{6,28},D-Ala^{9}]α-ANP-(6 – 28) [1155] exhibits strong receptor binding ($IC_{50} = 10^{-11}$–10^{-12} M).

A cyclopeptide from [Phe^{6},Pro^{7},Glu^{13}]r-α-ANP-(6 –22), cyclo-(F P F G G R I E R I G A Q S G L G), still produces the full biological effect, but with a low activity [1156].

Inhibition of the CRH-stimulated release of ACTH requires an intact amino terminal of α-ANP-(1 – 28): α-ANP-(5 – 28) is inactive but α-ANP-(1 – 11) still exerts a weak inhibitory effect [1157].

Surprisingly ANF-(1 – 30), ANF-(31 – 67), and ANF-(79 – 98) have a vasodilating and a cGMP-stimulating activity similar to that of ANF [1158]. ANF-(1 – 30), ANF-(31 – 67), ANF-(79 – 98), and ANF-(99 – 126) increase sodium excretion by 231, 973, 167, and 1405 %, respectively. All these peptides except ANF-(79 – 98) have diuretic activity; ANF-(79 – 98) exhibits kaliuretic activity (i.e., increases renal potassium excretion) [1159].

h-α-ANP-(7 – 23)amide and substituted [Cys(R)^{7,23}]h-α-ANP derivatives inhibit h-α-ANP-stimulated synthesis of cGMP [1160]. In rats, cyclo-(Phe-Gly-Gly-Arg-Leu-Asp) and cyclo-(Phe-Gly-Gly-Arg-Met-Asp-Arg-Ile-Gly) have hypertensive activity [1148].

5.6. Pancreatic Peptide Family

The pancreatic peptide family includes three peptide amides containing 36 amino acids: pancreatic peptide (PP), neuropeptide Y (NPY), and the peptide with N- and C-terminal tyrosine (PYY).

There is a certain structural similarity between these peptides and neuromedin U (NMU) and the delta-sleep-inducing peptide (DSIP).

	1	18
h-PP	A **P** LE **P**V Y **PG** D N **A** TP**E** Q MA-	
h-,r-NPY [1161]	Y **P** S K **P**D N **PG** E **D A** P A**E** D MA-	
h-PYY [1162]	Y **P** I K **P**E A **PG** E **D A S** P**E** E L N-	
cu-DSIP	WA**G** G **D A S** G**E**	
	19	36
h-PP	Q **Y** AA D**L** **R** R **Y** I **N** M**LTR** P **R** **Ya**	
h-,r-NPY	R **Y** YS A**L** **R** HY I **N** L I **TR** Q **R** **Ya**	
h-PYY	R **Y** YA S**L** **R** HY L **N** L V**TR** Q **R** **Ya**	
p-NMU-8	Y F **LFR** P **R** Na	

5.6.1. Pancreatic Peptide

Occurrence. Pancreatic peptide (PP) [*59763-91-6*], M_r 4182 (h-PP), is produced primarily in the pancreas. In squirrels, it was found in the same cells that produce glucagon [1163]. The precursor of h-PP has been characterized [1164]. Sequences of PP from various species are given in [1165].

Release. The pulsatile release of PP in the pancreas is coupled to an insulin pulse [1166]. The plasma level of PP is increased by electrical stimulation of the vagus, food intake [1167], stimulation of β-adrenergic receptors, the μ-receptor ligands morphine and morphiceptin, hypoglycemia induced by insulin or fasting, acetylcholine, cholecystokinin, cerulein, secretin, neurotensin [1168], bombesin [1169], gastrin, GIP, and VIP. The PP level oscillates in a 90-min rhythm in test persons with an empty stomach. The plasma level of PP increases with age [1170]. In general, patients with a low exocrine secretion from the pancreas have an excessively high plasma PP level. In half of patients suffering from pancreatic endocrine tumors, plasma PP levels are raised and cannot be lowered by atropine [1171].

The plasma PP level is lowered by stimulating the α-adrenergic receptors, by δ-receptor agonists ([D-Ala^{2},D-Leu^{5}]enkephalin and [D-Ala^{2}]-enkephalin), by the intravenous administration of glucose or fatty acids, by glucocorticoids, p-PYY, atropine, and somatostatin [1172].

Biological Effects. PP inhibits the exocrine secretion of enzymes from the pancreas which is stimulated by, for example, secretin, cerulein, cholecystokinin, bethanechol, HCl, or sodium

oleate [1173], and cerulein-induced pancreatic hypertrophy [1171].

Other effects of PP include antagonization of VIP-induced vasodilatation of vascular smooth muscle [1174], decrease in plasma insulin levels (rat), relaxation of the Oddi sphincter in small doses and contraction in high doses [1175], contraction of the lower esophageal sphincter [1176], release of $NaHCO_3$ from the gastric and duodenal mucosa [1177], and reduction of food intake.

Structure – Activity Relationships. The C-terminal tyrosinamide residue is very important for biological activity. Since the C-terminal hexapeptide of PP simulates the inhibitory effect of PP on exocrine pancreatic secretion, the active center appears to be located in this region. In contrast to PP, however, both the C-terminal hexa- and decapeptides are unable to lower the insulin level in rats.

5.6.2. Neuropeptide Y [1178]

Occurrence. Neuropeptide Y (NPY) [*82785-45-3*], M_r 4272 (h-NPY), is widely distributed in the nervous system, usually in the vicinity of arteries. In the brain, the highest concentrations are found in the hypothalamus. Immunoreactive NPY is associated with somatostatin in many neurons of the cerebral cortex, it also occurs in some cells of the diencephalon. In neurons of the arcuate nucleus, NPY is associated with growth hormone releasing hormone (GHRH) [1179]. In the adrenal medulla, NPY is found in the majority of noradrenaline-releasing granules; Met- and Leu-enkephalin have also been observed in some of these cells. NPY has been found together with VIP and PHI in the nonadrenergic nerve fibers of the small intestine [1180].

h- [1181] and r-NPY [1182] are synthesized in the form of a preprocompound with 97 amino acids. The 30 amino acid C-terminal peptide of prepro-NPY (C flanking peptide of neuropeptide Y, CPON) is associated with NPY in the brain and the peripheral nervous system [1183]. p-NPY-(12 – 36)amide (p-NPY-25) has been isolated from the porcine brain [1184]. Sequences of NPY from various species are described in [1163].

Release. NPY is released along with noradrenaline on activation of the sympathetic nerves. Release from the perfused guinea pig heart depends on extracellular calcium, calcium influx, and activation of protein kinase C [1185]. NPY is released by tyramine [1186], insulin-stimulated hypoglycemia [1187], loss of blood [1188], stress [1189], and deprivation of food [1190].

An increased level of NPY is found in platelets of spontaneously hypertensive rats [1191]. The peptide is assumed to play a role in the pathophysiology of congestive heart failure [1192]. A high plasma NPY level is observed in patients with pheochromocytoma (tumor of the adrenal medulla) [1193] and septicemia [1194].

The release of NPY is reduced by sympathectomy, reserpine, or CCK-8 [1195]. The NPY concentration in the hypothalamus decreases in old rats [1196].

Receptors. There are two receptors for NPY. The Y_1-receptor is a glycoprotein (M_r 70 000) which binds NPY, but not NPY-(13 – 36). The Y_2-receptor is a glycoprotein (M_r 50 000) which binds both NPY and NPY-(13 – 36). Both receptors also bind PYY [1197], [1198]. The vascular effects of NPY are presumably mediated by the Y_1-receptors and the cardiac effects by the Y_2-receptors [1199]. NPY inhibits the release of noradrenaline in the heart and reduces heart rate [1200]. [Leu^{31},Pro^{34}]NPY is a selective Y_1-receptor agonist [1201].

Biological Effects. NPY increases intracellular calcium in vascular smooth muscle cells [1202], inhibits cAMP formation which is stimulated by, e.g., forskolin [1203], and the calmodulin-stimulated phosphodiesterase [1204]. NPY exerts a strong, calcium-dependent vasoconstricting effect on cerebral arteries in vitro. It inhibits the VIP-induced relaxation of arteries [1205] and causes a persisting increase in blood pressure when applied parenterally.

NPY has many other effects:

- Stimulation of diuresis and natriuresis through release of ANF and inhibition of vasopressin [1206].
- Inhibition of the secretin- and cholecystokinin-stimulated exocrine pancreatic secretion and reduction of the glucose

or β-adrenergic stimulated release of insulin [1207].
- Inhibition of the pentagastrin-stimulated secretion of gastric acid [1208].
- Increase in the serum LH level in male rats [1209] and in ovariectomized, estrogen-treated rats via the release of gonadoliberin [1210] and the potentiation of gonadoliberin activity [1211].

When applied centrally, NPY stimulates the release of ACTH and cortisol via the potentiation of CRH [1212], increases food and water intake [1213] in the rat, delays the emptying of the stomach [1214], and reduces the blood pressure [1215] and body temperature [1216]. NPY may play a role in bulimia nervosa.

Structure – Activity Relationships. Both the N- and C-termini of NPY are important for binding to the Y_1-receptors, while C-terminal fragments suffice for the binding to the Y_2-receptors. The sequence 5 – 24 is not necessary for binding to Y_1-receptors. Thus, short sequences such as

$$[\text{D-Cys}^{7},\text{Aoa}^{8-17},\text{Cys}^{20}]\text{p-NPY}$$
$$[\text{Cys}^{5},\text{Aoa}^{7-20},\text{D-Cys}^{24}]\text{p-NPY}$$

bind to mouse brain membranes [1217] (Aoa = 8-amino-octanoic acid). [Aca^{5-24}]NPY (Aca = ε-aminocaproic acid) still has about one-third the receptor binding capacity of NPY [1218] and [Ahx^{5-24}]NPY (Ahx = 6-aminohexanoic acid) has about one-tenth the activity of NPY [1219].

[Pro^{34}]p-NPY [1220], [His^{34}]p-NPY, and [Leu^{31}, Pro^{34}]NPY are selective agonists for Y_1-receptors [1221]. [Leu^{31},Pro^{34}]NPY has a higher blood pressure increasing activity than NPY [1221].

Arg^{33} and Arg^{35}, but not Arg^{25} are important for biological activity [1219]. The twelve C-terminal amino acids and the amide function are required to activate the Y_2-receptor when tested on the electrically stimulated rat vas deferens. NPY-(26 – 36)amide and PYY-(26 – 36)amide have no activity, NPY-(16 – 36)amide is most active in this test [1222]. NPY-(12 – 36)amide, NPY-(22 – 36)amide and NPY-(22 – 33)amide inhibit calmodulin-stimulated phosphodiesterase more than NPY. N-terminal fragments are devoid of activity [1204].

NPY-(18 – 36) is an NPY antagonist for binding to membranes of the heart [1192].

5.6.3. Peptide with N-terminal Tyrosine and C-terminal Tyrosinamide

Occurrence. Peptide with N-terminal tyrosine and C-terminal tyrosinamide (PYY), [*4985-46-0*], M_r 4310 (h-PYY), occurs primarily in the small intestine. In humans, PYY-forming cells are found primarily in the duodenum; in the dog and the pig, PYY is mainly synthesized in the mucous membranes of the ileum and the colon [1223]. Sequences of PYY in various species are described in [1163].

Release. The plasma level of PYY increases after a meal and after the application of p-gastrin releasing peptide [1224].

A raised basal and postprandial level of PYY is observed in steatorrhea (fat diarrhoea) caused by atrophy of the small intestine and in chronic pancreatitis [1225]. In patients suffering from Dumping syndrome (sweating, weakness and diarrhoea caused by stomach operations), plasma levels of PYY increase 6 – 7-fold on oral glucose administration; somatostatin inhibits this increase [1225].

Biological Effects. PYY reduces the plasma CCK level [1226], the postprandial [1227], or secretin- or CCK-8-stimulated [1228] exocrine secretion from the pancreas and the postprandial secretion of gastric acid [1227]. The vasoconstricting and antisecretory effects on the pancreas are partially abolished by adrenalectomy [1229].

Other effects include increase in blood pressure, reduction of plasma PP and motilin levels [1230], prolongation of transit in the small intestine [1231], and strong emetic activity (PYY > NPY $\gg$ PP) [1232].

5.6.4. Neuromedin U

Occurrence. Neuromedin U (NMU) [*111745-44-9*] was first isolated from the spinal

cord of the pig (p-NMU-25, M_r 3143; and p-NMU-8). r- [1233], ra- [1234], and cav-NMU [1235] were recovered from the gastrointestinal tract.

NMU-like immunoreactivity is widely distributed, however. In the rat, it is found primarily in the pituitary and the gastrointestinal tract, but also occurs in the brain, spinal cord, and genital tract [1236].

Biological Effects. NMU increases blood pressure. It also contracts the rat uterus [1233], the isolated human ileum [1237], and muscle strips from the human bladder [1237].

Structure – Activity Relationships. r-NMU-25 is about twice as active as p-NMU-25 on the rat uterus. The C-terminal heptapeptide amide is essential for activity [1233].

5.6.5. Delta-Sleep-Inducing Peptide [1238]

Occurrence [1239]. Delta-sleep-inducing peptide (DSIP) [*69431-45-4*], M_r 849, was first isolated from a dialysate of venous blood in rabbits which were kept awake by electrically stimulating the thalamic region. Although the nonapeptide is widely distributed in the brain and other tissues, it is most concentrated in the thalamus and pineal gland. DSIP has been found in the pituitary [1240], urine, cerebrospinal fluid, and human milk. The immunoreactive DSIP in plasma, urine, and cerebrospinal fluid is bound to proteins which protect it from proteolysis.

Release. The concentration of DSIP in the brain increases during hibernation. In rats addicted to alcohol, however, the DSIP concentration in the brain falls. Corticoliberin and vasopressin inhibit the release of DSIP from the adenohypophysis [1240].

Biological Effects [1239]. DSIP can pass through the blood – brain barrier. Infusion into the ventricular system of the rabbit brain stimulates the δ waves of the electroencephalogram which are characteristic of natural sleep. Intracerebroventricular injection of DSIP in mice produces an analgesic effect [1241]. Although naloxone abolishes the sleep-inducing [1242] and analgesic effects [1241], DSIP does not bind to opioid receptors [1243]. DSIP increases resistance to acute emotional stress in rats [1244], inhibits the release of prolactin [1245], and the CRH-stimulated increase in ACTH [1246] and glucocorticoids [1247].

DSIP directly stimulates the release of growth hormone [1248] and inhibits the release of growth hormone releasing hormone [1249] via dopaminergic mechanisms. Similar to NPY and PP, intracerebroventricular application of DSIP stimulates the release of LH in estrogen- and progesterone-treated ovariectomized rats. On the other hand, DSIP has no LH-releasing activity in vitro (hypophysis) [1250].

Structure – Activity Relationships. Shortening the chain at the N- or C-terminus or replacement of α-Asp by β-Asp produces inactive peptides. Replacement of Trp^1 by D-Phe or 4-nitrophenylalanine is without effect, but [D-Trp^1]- and [Phe^1]DSIP increase the motor activity. Phosphorylation of the hydroxyl group of Ser^7 leads to a more potent and longer-acting product (DSIP-P) that is more stable to enzymatic degradation than DSIP.

Uses. In patients with sleep disturbances, the duration of sleep has been significantly prolonged by the administration of DSIP. It increases the power of concentration and tolerance to stress in the waking phase. DSIP exerts an analgesic and antidepressive effect in patients with migraine, headaches, psychogenic pain attacks, and tinnitus [1251]. Treatment with DSIP produces a favorable effect in opiate or alcohol withdrawal, where it abolishes increased blood pressure, tachycardia, tremor, and sweating [1238].

6. Cholecystokinin and Gastrin

Cholecystokinin (CCK), which stimulates contraction of the gallbladder, is identical to pancreozymin, which was made responsible for enzyme release from the pancreas. The counterparts of CCK in amphibians are cerulein (CRL) and related peptides, their common C-terminal pentapeptide is identical to that of gastrin.

Gastrin stimulates the release of gastric acid and pepsin in the stomach. Its activity spectrum

is dictated mainly by its C-terminal pentapeptide, whereas that of CCK and CRL depends on the C-terminal heptapeptide, which should contain a sulfated tyrosine residue.

Release of CCK and GT are initiated by special peptides. A trypsin inhibitor with a structure similar to that of the epidermal growth factor has been characterized as the CCK-releasing peptide. The gastrin-releasing peptide is structurally related to bombesin from amphibians.

6.1. Trypsin-Sensitive Cholecystokinin Releasing Peptide

A peptide with 61 amino acids has been isolated from the pancreatic juice of rats and stimulates pancreatic enzyme secretion by inducing the release of CCK. This cholecystokinin releasing peptide (CCKRP) is released in response to a diet rich in protein [1252]. It is structurally similar to the pancreatic trypsin inhibitors of the Kazal type and is classified as such [1252]. Structural similarities are found between CCKRP and the epidermal growth factor (EGF) [1253]. CCKRP also binds to EGF receptors and exerts a growth-promoting effect [1253].

6.2. Cholecystokinin and Cerulein

Occurrence. Cholecystokinin (CKK) [*9011-97-6*] is formed as prepro-CCK. The structures of h- [1254], p- [1255], and r-prepro-CCK [1256] have been elucidated via their cDNA. CCK has been isolated in the form of peptides containing 58 (CCK-58) [1257], 39 (CKK-39), 33 (CCK-33), 22(CCK-22) [1258], 12 (CCK-12), 8 (CCK-8) [1258], 5 (CCK-5) [1259], and 4 (CCK-4 and GT-4) C-terminal amino acids. These peptides are synthesized in the brain, small intestine, and the pancreas.

The r-prepro-(24 – 32)nonapeptide (V-9-M) is found in the rat brain. It has a persisting memory-improving effect when injected into the lateral ventricle [1260].

CRL has a similar sequence to CCK-8 and has been isolated from the Australian frog, *Litoria caerulea.* The prepro-CRL from the skin of *Xenopus laevis* codes for five heterogeneous pro-CRL peptides [1261].

h-CCK-8	M_r 1143	D Ys M G W M D Fa
CRL	M_r 1352	*Q Q D Ys T G W M D Fa

Release. Ingestion of unsaturated fats and fatty acids causes a greater release of CCK than ingestion of the corresponding saturated compounds. CCK is also released in response to the acidification of the duodenum [1262], trypsinsensitive cholecystokinin releasing peptide, bombesin [1263], depolarizing K^+ concentrations, and veratridin (a Na^+ channel activator) [1264].

Bile acids [1265], somatostatin, and atropine [1266] inhibit the release of CCK. Stress modulates the CCK content in the rat hypothalamus [1267].

The postprandial level of CCK in the plasma of patients with pancreatic insufficiency is lower than that of healthy persons [1268]. The plasma CCK level in patients with non-insulin-dependent diabetes [1269], hepatic cirrhosis [1270], and chronic pancreatitis [1271] is elevated.

Receptors [1272]. There are two main types of CCK receptor: CCK-A and CCK-B [1261]. The CCK-A receptors are found on pancreatic acinar cells, in the gallbladder, on adenohypophysis cells, and on inhibitory neurons of the lower esophageal sphincter. Only sulfated CCK derivatives bind to these high-affinity receptors. Gastrin exhibits a 500 to 1000 times weaker binding. CCK-B receptors are found in the cerebral cortex and in the central nervous system. They have a tenfold higher preference for sulfated than for nonsulfated CCK derivatives. A third class of receptors termed the gastrin receptors are found on the parietal cells and the gastrointestinal smooth muscles. They have about the same affinities for sulfated and nonsulfated CCK derivatives [1273].

Binding of CCK to pancreatic receptors leads to an up regulation of the CCK receptors [1274]. In newborn rats, the number of CCK receptors is increased by hydrocortisone [1275]. The CCK receptors on rat pancreatic acini are increased fourfold in diabetic rats [1276].

Selective ligands for the CCK-A receptor are described in [1277]. Selective ligands for the CCK-B receptor are described in [1278–1282].

Biological Effects. The activity spectrum of CCK was mainly studied with synthetic CCK-8 (sincalide) and CRL (ceruletide).

Receptor activation leads to the breakdown of phosphoinositides, mobilization of cellular calcium, activation of protein kinase C, and the phosphorylation of intracellular proteins.

The most important physiological effects of CCK are the CCK-A-receptor-mediated stimulation of bile flow [1283] (contraction of the gallbladder, relaxation of the bile duct and the Oddi sphincter), the exocrine pancreatic secretion [1284], [1285], and potentiation of the secretin-induced exocrine pancreatic secretion. The CCK-8-induced contraction of the gallbladder is increased in patients suffering from constipation and lowered in patients with diarrhoea [1286] and celiac disease [1287]. VIP and secretin potentiate the effect of CCK on the gallbladder and pancreas. Somatostatin [1288] and atropine [1289] inhibit the CCK-induced contraction of the gallbladder.

cAMP is activated in CCK-induced relaxation. In contractions, cGMP is activated and intracellular calcium and acetylcholine are released [1290].

Other CCK-A-receptor-mediated effects include decreased food intake in rats after peripheral injection of CCK-8 [1291], trophic effects on the pancreas [1292], increased insulin and glucagon secretion in mice [1293], inhibition of pentagastrin-stimulated gastric acid secretion [1294], and increased dopamine efflux in rat brain [1295].

CCK-B-receptor-mediated effects are CCK/opiate interactions [1296] and anxiety [1279]. CCK has an analgesic effect through δ-opioid receptors via the release of endogenous opioids [1297] but it reduces morphine- or β-endorphin-induced analgesia [1298] via μ-receptors [1299].

Other effects of CCK include relaxation of the corpus of the stomach and the lower esophageal sphincter, contraction of the pylorus and prolongation of the time required to empty the stomach, increase in motility from the distal part of the duodenum up to the rectum [1300], stimulation of gastric acid secretion, and inhibition of the gastrin-stimulated secretion of gastric acid.

Structure – Activity Relationships. The C-terminal octapeptide of CCK (CCK-8) has the full activity spectrum of CCK. The C-terminal heptapeptide CCK-7 and CRL-(4 – 10) [1301] are required for CCK activity. In vivo, deamino^1-CCK-7 has the same protein secretory activity as CCK-7, but is more effective in contracting the gallbladder than CCK-7. CCK-6 has no activity.

The sulfated tyrosine residue is important for binding to CCK-A receptors and the activity spectrum of CCK and CRL. The activity of desulfated CCK-8 resembles that of gastrin [1302]. The substitution of Phe(CH_2SO_3H) for Tyr(SO_3H) gives more stable, highly active derivatives [1303]. If the N-terminal Tyr(SO_3H) in CCK-7 is replaced by a 3-(4-sulfoxyphenyl)-2-methylpropanoyl group, the analgesic activity is about five times that of CCK-7 [1259].

Met^3 in CCK-8 is not essential for activity and can be replaced by Leu or Thr; Met^6 can only be replaced by Nle without loss of activity. Thus, Arg-[Thr^3,Nle^6]CCK-8 [1304] and Boc-[Nle^{3,6}]CCK-8-(2 – 8) [1305] are stable to oxidation and fully active.

Gly^4 can be replaced by D-amino acids. [D-Ala^4]CCK-8 and [D-Trp^4]CCK-8 are almost as active as CCK-8, but they are stable to brain proteases and act longer [1306].

The effect of Boc-[Nle^3-ψ-($COCH_2$)Gly^4]CCK-8-(2 – 8) on the release of amylase from dispersed rat pancreatic acini is about the same as that of CCK-8, but it also has a central antagonistic effect [1307].

[β-Asp^7]CCK-8 loses its cholecystokinin properties but retains its pancreozymin activity. The substitution of Ser(SO_3H), Thr(SO_3H), or Hyp(SO_3H) for Asp^7 in Ac-CCK-8-(2 – 8) gives preparations that have 2 – 3 times the activity of CCK-8 on isolated strips of gallbladder [1308].

Phe^8 can be replaced by 3-(2-naphthyl)alanine or by 3-cyclohexylalanine without loss of activity [1309].

The C-terminal amide function is very important for the activity of CCK peptides [1310]. CCK receptor antagonists are obtained by deleting C-terminal amino acids. The shortest active fragment is CCK-8-(1 – 5) [1311]. Stronger antagonists are succinyl–CCK-8-(2 – 7)2-phenethylamide [1312] and Boc-[D-Trp^5,Nle^6]CCK-8-(2 – 7)2-phenethyl ester [1313]. Surprisingly strong antagonists are also

obtained by substituting D-Orn(Z) for Met ^{6} in CCK-8 [1282].

Relatively simple derivatives of glutamic acid and tryptophan act as CCK antagonists [1314]. Examples are the inhibitor of acid secretion proglumide (*N*-benzoyl-L-glutamic acid-α-dipropylamide; Milid, Offermann), lorglumide (CR-1409, 3,4-dichlorobenzoyl-Glu-dipentylamide), loxiglumide (CR-1505, 3,4-dichlorobenzoyl-Glu-pentyl-3-methoxypropylamide), and 4-chlorobenzoyltryptophan (benzotript) [1315].

Strong CCK antagonists are asperlicin, a nonpeptide component of *Aspergillus alliaceus*, and its benzodiazepine variants, devazeptides (MK-329 = L-364,718), L-365,260, trifluadom, and L-156,440. L-365,260 preferentially binds to CCK-B receptors, the others bind to CCK-A receptors [1314].

Uses. CCK-8 (sincalide, Squibb – Heyden) [*25126-32-3*], and CRL (ceruletide; Takus, Farmitalia) [*17650-98-5*] [1281] are used in X-ray diagnostics and in the diagnosis of pancreatic function. Their administration halves the time required to pass the small intestine and thus permits quick examination of the intestine with an oral contrast medium. Ceruletide is used therapeutically in postoperative intestinal atonia and paralytic ileus because of its stimulating effect on the small intestine.

Ceruletide can be employed for the expulsion of gallstones [1316] and in biliary colic [1317]. Sincalide and ceruletide (nasal application three times daily, 100 µg) alleviate chronic pancreatitis [1318]. Clinical reports confirm the analgesic properties of CCK-8 [1319].

6.3. Gastrin Releasing Peptide and Bombesin

Occurrence. Gastrin releasing peptide (GRP) [*80043-53-4*], M_r 2859 (h-GRP), is a peptide amide consisting of 27 amino acids and has been isolated from the gastric tissue of pigs, chicken, and dogs and from human lung tumors [1320]. Its C-terminal heptapeptide is identical to that of bombesin. Bombesin (BB) [*31362-50-2*], M_r 1620, 14 amino acids, is a peptide amide from the skin of the frog *Bombina bombina*. Alytesin, litorin, and ranatensin are peptides with related structure and activity that have been isolated from amphibians (frog) [1321].

The neuromedins B, B-32, B-30, and C have been isolated from the brain and spinal cord of the pig; p-neuromedin C corresponds to the C-terminal decapeptide of p-GRP.

Immunoreactive GRP has been found in the central nervous system and peripheral tissue (e.g., gastrointestinal tract). Material similar to GRP-(14 – 27) has been found in endocrine tumors [1322] and cow's milk [1323].

Release. Electrical stimulation of the vagus increases formation of GRP in the pancreas [1324]. The β-adrenergic stimulation of gastrin and somatostatin appears to be mediated by GRP [1325].

Receptors. At least two different types of receptor are assumed for GRP, BB, and the neuromedins. BB and GRP bind to pancreatic acinar cells and murine 3T3 cells with high affinity, while neuromedin B binds with low affinity. The esophageal mucosa possesses receptors which have a high affinity for neuromedin B, but a low affinity for GRP. The GRP antagonist [D-Phe^{6}]BB-(6 – 13)OEt binds preferably to the pancreas, while [Tyr^{4},D-Phe^{12}]BB and the SP antagonists [D-Pro^{4},D-Trp^{7,9,10}]SP-(4 – 11) and [D-Arg^{1},D-Trp^{7,9},Leu^{11}]SP bind better to the esophagus than to the pancreas [1326].

Biological Effects. BB and GRP have similar effects. In healthy persons, a BB infusion increases the plasma levels of gastrin, CCK, motilin, pancreatic peptide, vasoactive intestinal peptide, gastrin inhibiting peptide, glucagon, insulin, and trypsin [1327], as well as the plasma level of gonadoliberin-stimulated lutropin and follitropin [1328]. The blood glucose level is lowered and plasma Ca^{2+} levels are raised. Somatostatin inhibits BB-induced hormone secretion [1329]. In healthy test persons, the food-induced release of insulin [1324] and the thyroliberin-stimulated levels of thyrotropin and prolactin in the plasma are lowered by an infusion with BB [1328]. These effects are due to increased plasma somatostatin levels. Other effects of GRP/BB include stimulation of gastric acid secretion and exocrine pancreatic secretion,

increase in bile flow [1330], trophic action on the gastrin-producing cells of the stomach [1331] and on the pancreas [1332], as well as autocrine growth in cancer cells [1333].

Central application of BB induces hypothermia and adrenaline release which, in turn, inhibits insulin release, and stimulates glucagon release. Other central effects are the reduction of the plasma growth hormone level [1334], stimulation of dopamine synthesis in the hypothalamus [1335], suppression of gastric acid and pancreatic secretion, and inhibition of food and liquid intake [1334].

Structure – Activity Relationships. GRP-(14 – 27) acts like BB [1336]. The C-terminal nonapeptide of BB and Ac-GRP-(20 – 27) have minimum length and maximum effect. GRP-(23 – 27) is the smallest active compound. Shortening the molecule at the C-terminus yields preparations with low affinity and activity. Acetylated peptides shortened at the N-terminus are biologically more active than the corresponding compound with a free amino group. Trp^8 and His^{12} are very important for GRP activity.

Antagonists. [D-Phe^{12},Leu^{14}]BB is a competitive BB antagonist. It inhibits in vitro BB-induced amylase secretion from guinea pig acini cells. [D-Phe^{12}]BB and [Tyr^4,D-Phe^{12}]BB have similar activity [1337]. Introduction of D-Phe^6 enhances the antagonistic activity, while D-Phe^5 has no effect [1338].

BB antagonists are also formed by deleting the C-terminal methionine [1339]. Me_3-C-CO-His-Trp-Ala-Val-D-Ala-His-Leu-OMe exhibits good receptor binding and inhibits mitogenesis of mice Swiss 3T3 cells.

The best (100 %) in vivo inhibition of BB-stimulated amylase secretion was produced by Me_2-CH-CO-His-Trp-Ala-Val-D-Ala-His-Leu-NH-Me (ICI 216 140) [1339].

Other good de-Met-antagonists in the receptor binding assay are [D-Phe^6]BB-(6 – 13)ethylamide, [D-Phe^6]BB-(6 – 13)propylamide [1340], Ac-GRP-(20 – 26)OEt [1341], and [D-Phe^6]BB-(6 – 13)OEt [1342].

Reduction of the peptide bond between positions 13 and 14 yields potent BB antagonists, e.g., [Leu^{13}-ψ-CH_2NH-Leu^{14}]BB [1343]. Shortening of this antagonist at the N-terminus reduces antagonistic activity. Introduction of D-Phe at position 6 and N-terminal shortening produce a compound with ten times the activity [1344].

Uses. BB is used as a diagnostic aid in the gastrin stimulation test. For instance, patients suffering from antral gastritis have an abnormally low gastrin level after BB stimulation.

In severe chronic pancreatitis [1345] and in patients with duodenal ulcers [1346], BB increases the gastrin level to a greater extent than in healthy persons. BB increases the plasma trypsin level in healthy persons, but not in patients with pancreatic insufficiency [1347].

6.4. Gastrin

Occurrence. Gastrin (GT) [*9002-76-0*], M_r 2098 (h-GT-17), was the first gastrointestinal hormone to be structurally characterized. It is synthesized primarily in the antrum and the duodenum as prepro-GT. Immunoreactive GT has also been found in the pituitary, pancreas, and nerves. In humans, tyrosine sulfation depends on the site of synthesis and the degree of development [1348].

GT occurs in several molecular sizes: GT-34 (big gastrin), GT-17 (little gastrin), GT-14 (mini gastrin), GT-6, and GT-4. It is structurally related to the transformation protein of polyoma virus (TPPV) and the chicken antral peptide (g-AP) which stimulates gastric acid production in chickens [1349]. The main physiologically active component is GT-17. Although GT-14 has full biological activity, it only occurs in very small amounts.

A sulfated myotropic neuropeptide, leucosulfakinin, has been isolated from the head of the cockroach [1350]; its intestinal myotropic activities are similar to those of GT. Related peptides are found in other insects.

```
h-GT-34   *Q L G P Q G P P H L V A D P S K K-
           Q G P W L E E E E E A Y G W M D Fa
h-GT-17   *Q G P W L E E E E E A Y G W M D Fa
h-GT-14         W L E E E E E A Y G W M D Fa
  GT-4                            W M D Fa
```

Release. The atropine-resistant secretion of GT-17 and GT-34 occurs postprandially, especially in response to protein or on stimulation

of the vagus. GT is also released by the gastrin releasing peptide, bombesin, carbachol [1351], Ca^{2+}, vitamin D_3 [1352], tolbutamide [1353], PGE_2 [1354], hypoglycemia [1355], amino acids and the corresponding amines [1356], inhibitors of monoamine oxidase [1357], and testosterone [1358].

The release of GT is reduced by Mg^{2+}, acidification of the food in the antrum to pH 1.2 – 1.4 [1359], inhibitors of amino acid decarboxylases, and estrogen [1358].

In humans, the GT level rises during the day and reaches a maximum at about 6 p.m. [1360]. The plasma level of gastrin is increased in the Zollinger – Ellison syndrome (a GT-producing tumor), pernicious anemia (some patients develop autoantibodies against gastrin receptors [1361]), gastric ulcers [1360], and hyperthyroidism [1362].

Biological Effects. GT increases the formation of cAMP and of inositol phosphates [1363] and decreases the plasma level of calcium.

In the stomach, GT increases the formation of gastric acid and pepsin and stimulates the blood circulation in the mucosa [1364]. GT-17 stimulates the secretion of gastric acid in patients with duodenal ulcers more than in healthy persons [1365]. The gastrin inhibitory peptide (GIP), somatostatin, secretin, neurotensin, glucagon, glicentin, thyroliberin, the anorexigenic peptide, cerulein, and cholecystokinin inhibit the GT-stimulated secretion of gastric acid.

The synthesis and marketing of Boc-β-Ala-GT-4 [*5534-95-2*] (pentagastrin) a GT analogue that is fully active in humans, stimulated intensive investigation of the activity of GT. An infusion with pentagastrin, like a protein meal, increases endogenous GT, insulin, GHRH, and calcitonin, but decreases the plasma level of somatostatin (SRIH). The primary effect of pentagastrin appears to be the inhibition of SRIH release; gastrin, insulin, growth hormone, and calcitonin can then be released.

Other effects of gastrin include trophic effects on the oxyntic mucose, enterochromaffin-like cells, A-cells [1366], pancreas (potentiated by secretin [1367]), tumors of the human stomach and colon [1368], [1369], and rat pancreatic adenocarcinoma [1370]; stimulation of histamine secretion; increase in the pressure of the lower esophageal sphincter; and inhibition of the CCK-induced contraction of the pylorus.

The common C-terminal tetrapeptide amide of gastrin and CCK (GT-4, trymafan, tetrin) stimulates the endocrine secretion of insulin, glucagon, somatostatin, and pancreatic peptide. Exocrine secretion from the pancreas is primarily stimulated by the higher molecular forms of gastrin and CCK rather than by GT-4.

In healthy subjects, intravenous GT-4 causes short attacks of panic or symptoms of fear [1371]. In rats, intracerebroventricular application causes deterioration of memory [1372].

Structure – Activity Relationships. GT-34 is more potent and acts longer than GT-17 when tested in humans [1373]. The five glutamic acid residues at positions 6 – 10 in h-GT-17 are of importance for biological activity [1374].

GT-4 has the structure required for gastrin activity. The C-terminal amide is important for biological activity. Substitution of other amino acids for Asp results in substantial loss of activity. replacement of the GT-4 amino acids by the corresponding D-amino acids decreases activity. Agonists and antagonists can be obtained by reducing the peptide bonds. Boc-[trp^1-ψ-(CH_2NH)-Leu^2]GT-4 is an agonist which binds more tightly to gastric mucosa cells than Boc-GT-4 [1375]. Trp^1 is not required for the insulin-releasing activity of GT-4, it can be replaced by L- or D-Orn [1376]. [Pro^2]GT-4 is more effective than GT-4 in releasing insulin from isolated islet cells of the rat pancreas [1377].

Very potent analogues are obtained by N-terminal extension of GT-4; pentagastrin in which GT-4 is extended by Boc-β-alanine is the best known. Glutaroyl-[Leu^5]GT-7 (deglugastrin) is more active by a factor of about two. Peptides having a GT-4 or GT-6 structure with a glycosylated N-terminal amino group have a higher water solubility combined with excellent biological activity.

Oxidation of Met causes loss of gastrin activity. However, both Met residues of GT-7 can be replaced (e.g., by Leu) without loss of activity.

Antagonists. C-terminal GT sequences that lack the C-terminal Phe-amide inhibit GT activity (e.g., bocgT-8-(1 – 7)amide). The smallest antagonistic sequence is Boc-Trp-Met-Asp-amide [1378]. Somewhat stronger antagonists are obtained by substituting β-homoaspartic

acid for asp, e.g., Boc-Trp-Leu-β-homo-Asp-Phe-amide [1379].

Antagonists are also obtained by reducing the peptide bond between leu and Asp [1375].

Uses. Pentagastrin (Gastrodiagnost, Merck) is used as a diagnostic aid to determine the maximal stimulation of gastric acid secretion.

7. Secretin Family

The first peptide hormone of the gastrointestinal tract, secretin (sEC), was discovered in 1902 by BAYLISS and STARLING. The structure of secretin was established in the 1960s by V. MUTT and coworkers. There is considerable structural similarity between SEC and the vasoactive intestinal peptide (viP), p-PHI-27, h-PHM-27, pituitary adenylate cyclase activating polypeptide (PACAP), glucagon (GG), two glucagon-like peptides (gLP-1 and GLP-2), glucose-dependent insulinotropic and gastrin inhibiting peptide (GIP), and growth hormone releasing hormone (GHRH).

GHRH activity triggers the growth hormone cascade: it releases growth hormone in the pituitary which in turn stimulates the synthesis of the somatomedins. Hormones that are structurally related to growth hormone are placental lactogen, prolactin, proliferin, and prolactin-like and prolactin-related proteins. The somato-medins are structurally related to insulin and relaxin. On the basis of their activity, somatostatin, galanine, and pancreastatin also belong to this group. Galanine triggers the growth hormone cascade by releasing GHRH and, like pancreastatin, inhibits the glucose-induced release of insulin. Somatostatin inhibits the release and activity of GHRH, insulin, glucagon, and many other peptide hormones.

7.1. Secretin

Occurrence. Secretin (sEC) [*1393-25-5*], M_r 3039.5 (h-SEC), contains 27 amino acids and is secreted primarily by the mucosa of the duodenum and upper jejunum. It is also found in the antrum, brain [1380], hypothalamus, and pituitary. Propeptides of p-secretin have been isolated and characterized [1381]. The structures of the preprohormones of r- and p-SEC have been established via the corresponding cDNA [1382].

h-SEC

H S D G T F T S E L S R L R E G
A R L Q R L L Q G L va

There is some similarity between secretin and the dentinal fluid transport stimulating peptide (DFT), also called the parotid hormone (ph). DFT has been isolated from the parotid gland [1383] and increases the transport of fluid in the teeth [1384]. The C-terminal regions of SEC show structural similarity with allatostatin (a peptide from the brain of the cockroach) [1385] and the bombilitins (mast cell degranulating peptides from the poison of the bumble bee) [1386].

Release. In the small intestine, hydrochloric acid (pH < 4) stimulates the secretion of SEC into the blood stream. Acid-stimulated release of SEC is mediated by a SEC-releasing peptide [1387]. Stress produces a 3 – 6-fold increase in the plasma levels of secretin. Secretin secretion from the intestinal mucosa is stimulated by high extracellular concentrations of K^+ and inhibited by the absence of extracellular Na^+ and Ca^{2+} [1388].

SRIH inhibits the release of SEC and exocrine pancreatic secretion. Patients with chronic kidney damage and Zollinger – Ellison syndrome (a gastrin-producing tumor) have a significantly raised level of SEC.

Receptors. SEC binds to the perfused pancreas and the gastric mucosa of rats. in contrast to SEC bound to the pancreas, SEC bound to the gastric mucosa can be displaced by tetragastrin [1389].

Biological Effects. The activity of SEC is generally expressed in clinical units (CU) which are determined in the dog (pancreatic secretion of aqueous nahCO_3): 1 mg of SEC has 4000 – 5000 CU. The physiological level of about 20 pg/mL, observed in the human jejunum after H^+ stimulus, should be achieved after an infusion of 0.03 CU kg^{-1} h^{-1}.

SEC stimulates the release of an aqueous $NaHCO_3$-containing secretion from the panreas

which neutralizes any hydrochloric acid entering the small intestine [1390]. Cholecystokinin, neurotensin [1391], and adenosine [1392] potentiate the SEC-stimulated pancreatic secretion. SEC-induced protein secretion from the pancreas is described frequently but is due to impurities, such as cholecystokinin [1393].

Other effects of SEC include:

1) activation of adenylate cyclase and cAMP synthesis
2) proliferative effect in the pancreas which is potentiated by CCK [1394]
3) reduction of the plasma gastrin level and of the formation of gastric acid (via SRIH and prostaglandins [1395], potentiated by neurotensin [1391])
4) reduction of stress-induced hemorrhage from the stomach [1396]
5) vasodilative and motility inhibiting effect (dilation of the pancreatic ducts [1397], reduction of the lower esophageal sphincter pressure [1398], increase in the circulation to the pancreas and duodenum)
6) contraction of the pyloric sphincter (delay in the emptying of the stomach [1399])
7) potentiation of the CCK-stimulated flow of bile [1400]
8) release of insulin in vitro

Structure – Activity Relationships. Analogues of secretin that are more active than the native peptide have not been found. Shortening the molecule at the N- or the C-terminus produces less active products. For instance, SEC-(2 – 27)amide has only 1 % of the activity of secretin. The shortest known active sequence is SEC-(1 – 13): 0.1 mg/kg of this substance still exerts a significant effect on the pancreas. However, 1 mg/kg of SEC-(1 – 11)amide shows no activity in the rat. Similar to His^1, Asp^3, and Phe^6, Arg^{12} is important for secretin activity.

Secretin extended at the C-terminus by Gly-Lys-Arg (natural pro-SEC) still has about 80 % of the SEC activity [1401].

The Asp^3–Gly^4 bond of secretin is responsible for its frequently reported instability. In an acid medium the β-carboxyl group of aspartic acid acylates the adjacent C-terminal peptide bond. This aspartoyl peptide then rearranges to give a peptide in which the peptide bond is formed with the β-carboxyl group of aspartic acid (and not the α-carboxyl group). The Asp-Gly-(3 – 4) and the Asp-Ser-(15 – 16) bonds in p-SEC are cleaved in aqueous solution at pH 5 [1402]. The main decomposition product in neutral phosphate buffer is [β-Asp^3]SEC [1403]. In acidic acetate buffer (pH 4), SEC-(16 – 27), SEC-(4 – 27), [β-Asp^3]SEC, [3-aspartoyl]SEC, and [βAsp^{15}]SEC are formed [1403]. These decomposition products have little or no biological activity.

Uses. Secretin is generally administered parenterally, but the intranasal application of a dry powder has also been described. Secretin (Sekretolin, Hoechst) is applied parenterally in doses of 75 – 100 CU to test pancreatic function and to diagnose the Zollinger – Ellison syndrome. In contrast to healthy subjects, patients with the Zollinger – Ellison syndrome react to a diagnostic intravenous dose of secretin (75 CU) with increased plasma levels of gastrin and GIP and not a decreased level of gastrin [1402]. SEC normally increases the diameter of the pancreatic duct, but this is not the case in patients with chronic pancreatitis [1404]. Infusion of sekretolin (0.25 – 0.5 CU kg^{-1} h^{-1}) stops bleeding of the upper gastrointestinal tract [1405]. SEC improves the condition of patients suffering from peptic ulcers [1406]. The relaxing effect on the lower esophageal sphincter can be exploited in patients with achalasia [1398].

The insulin-releasing effect of SEC (2 CU/kg) can be used to differentiate between single insulinomas or nesidioblastosis and multiple B- cell adenomas or hyperplasia [1407].

7.2. Vasoactive Intestinal Peptide [1408]

Occurrence. Vasoactive intestinal peptide (VIP) [*37221-79-7*], M_r 3326, was first isolated along with secretin from the intestine of the pig. It is distributed throughout the body, primarily in the nervous system.

VIP is a peptide amide containing 28 amino acids. It is formed from the same prepropeptide as PHI-27 (pig) and PHM-27 (humans) that are peptide amides containing 27 amino acids. PHI-27 stands for peptide with N-terminal histidine and C-terminal isoleucinamide, PHM-27 for peptide with N-terminal histidine and C-terminal methioninamide. An extended version

of VIP with Gly-Lys-Arg at the C-terminus has also been found [1409].

p-,b-,r-,ca-,cu-,cap-,o-, and h-VIP are identical [1410]:

```
H S D A V F T D N Y T R L R K
Q M A V K K Y L N S I L Na
```

There is little difference in the sequences of h-PHM-27, p-, b- [1411], and r-PHI-27 [1412]; cu-PHI-27 is identical to p-PHI-27 [1413].

VIP and PHI-27 are found in the neurons of the central nervous system, in the digestive tract, lungs, urogenital tract, gallbladder [1414], exocrine glands, thyroid gland, adenohypophysis [1415], and the adrenal gland. It occurs associated with acetylcholine and galanine [1416].

Release. VIP is released by electrical stimulation of nerve fibers, from the gallbladder wall by stimulation of the vagus, and from the hypothalamus by prostaglandin E_1. Plasma levels are increased by oxytocin, sexual stimulation in women, the sucking stimulus [1417], physical stress [1418], operation stress [1419], and corticosterone [1420]. VIP is also released by Met-enkephalin and β-endorphin [1421] in the hypothalamus, by insulin in the intestine [1422], and by estrogen in the adenohypophysis [1423].

The release of VIP is inhibited by glucose infusions [1418], peripheral application of endogenous opiates [1418], pancreatic peptide, and somatostatin.

Patients with multiple sclerosis have a significantly lower level of VIP in their cerebrospinal fluid than healthy persons. Impotent men have a low plasma VIP level. In patients suffering from cystic fibrosis, fewer VIP-immunoreactive nerves are found in the eccrine sweat glands [1424]. Intravenous glucagon or secretin raises the plasma levels of gastrin and VIP in patients with Zollinger – Ellison syndrome [1425]. The treatment of patients with congestive heart failure with ACE inhibitors increases plasma VIP levels [1426].

Receptors [1427]. At least two high-affinity VIP receptors and one low-affinity VIP receptor may exist. The antagonist [4-Cl-D-Phe^6,Leu^{17}]VIP differentiates between two high-affinity VIP receptors in guinea pig acinar cells [1428]; carbachol binds to the low-affinity VIP receptors [1429]. The number of VIP receptors on human mononuclear leucocytes increases drastically in response to physical activity and an energy deficit [1430].

Biological Effects. In the cell, VIP stimulates the synthesis of cAMP and catecholamines (activation of tyrosine hydroxylase) [1431]. It also potentiates the muscarinic stimulation of the turnover of phosphoinositides [1432]. VIP inhibits the activation of the calmodulin-sensitive phosphodiesterase [1433]. VIP and acetylcholine act synergistically in secretion and circulation. PHI exhibits VIP-like activity. VIP has many biological effects [1408]. The most important are listed below:

1) Cardiovascular system: vasodilation, hypotension, positive chronotropic and inotropic effects [1434].
2) Respiratory system: bronchodilation, augmented ventilation, pulmonary vasodilation, stimulation of bronchial secretion (VIP and asthma [1435]).
3) Gastrointestinal tract: relaxation of lower esophageal sphincter; relaxation of smooth muscle in the stomach and suppression of acid secretion; stimulation of pancreatic water and HCO_3 secretion; increased bile flow; inhibition of intestinal absorption; stimulation of intestinal water and chloride secretion; relaxation of circular muscles and contraction of longitudinal muscles in the colon [1436], [1437].
4) Urogenital tract: increase of blood circulation in the vagina and erection of the penis.
5) Metabolic effects: stimulation of glycogenolysis and lipolysis, bone resorption, hair growth [1438]; regulation of mitosis, differentiation and survival of cultured sympathetic neuroblasts [1439]; stimulation of in vitro meiosis in follicle-enclosed rat oocytes [1440]; and regulation of neuronal survival.
6) Endocrine functions: increased pituitary release of ACTH, prolactin [1441], growth hormone, and LH; suppression of somatostatin release; an increased release of gonadoliberin in the hypothalamus and of iodine-containing hormones in the thyroid; ACTH-like action in the adrenals (stimulation of aldosterone and corticosterone synthesis [1442]); increase of serum testos-

terone and dihydrotestosterone in women [1443]; increased aromatase activity in the neonatal rat ovary [1444]; increase of serum estradiol in women [1443]; stimulation of estradiol and progesterone release from rat ovaries [1445]; and stimulation of testosterone, progesterone, and pregnenolone release in cultured neonatal testicular cells [1446].

Structure – Activity Relationships. The C-terminal amide is not required for the biological activity of VIP; deamido^{28}-VIP, VIP-Gly, and [Leu^{17}]VIP-Gly-Lys have similar activities [1447]. Shortening the molecule at either the C- and N-terminus decreases biological activity. VIP-(10 – 28) exerts an antagonistic effect on the VIP-stimulated synthesis of cAMP and secretion of pepsinogen [1448]. Met^{17} can be replaced by norleucine without loss of activity.

The hydrophobic amino acids of the helical domain (6 – 28) are important for receptor binding, the hydrophilic amino acids in this region appear to be less important. Substitutions of the amino acids in positions 6, 10, 12, 17, 19, 22, 23, and 27 produce less active derivatives [1449]. Iodination leads to at least two VIP derivatives with full biological activity [1450].

Displacement of [^{125}I]VIP bound to receptors on rat liver membranes by various analogues occurs in the following order: VIP > [D-Ala^4]VIP > [D-Asp^3]VIP > [D-Ser^2]VIP > [D-His^1]VIP > [D-Phe^2]VIP > [D-Arg^2]VIP. [D-Phe^2]VIP is a competitive VIP antagonist, the other analogues have agonistic activity [1451]. Still better VIP antagonists are [D-4-ClPhe^6,Leu^{17}]VIP [1452] and Ac-[Tyr^1,D-Phe^2]GHRH-(1 – 29)amide [1453]. [D-Phe^4]PHI-27 is a highly selective VIP agonist [1454].

7.3. Pituitary Adenylate Cyclase Activating Polypeptide (PACAP)

Occurrence. There is considerable homology between VIP and o-PACAP-38 [1455] and o-PACAP-27 [1456], two peptide amides from the sheep hypothalamus. The prepropeptide of h-, o- [1457], and r-PACAP [1458] has been characterized via the cDNA. Highly specific receptors have been found primarily in the hypothalamus, brain stem, cerebellum, and lung. PACAP-27 binds more tightly than PACAP-38.

In the rat hypophysis, o-PACAP-38 and o-PACAP-27 stimulate adenylate cyclase 1000 times more effectively than VIP [1455], [1456]. Like VIP, o-PACAP-27 exhibits vasodepressor activity [1456]. Concentrations of 10^{-10} mol/L release growth hormone, prolactin, and ACTH from rat hypophyseal cells, 10^{-9} mol/L releases LH as well [1459].

7.4. Glucagon and Related Peptides

Glucagon (GG) is formed together with the two glucagon-like peptides, GLP-1 and GLP-2, as prepro-GG (r- [1460], h- [1461], b- [1462], p- [1463]). Glucagon is prepro-GG-(33 – 61), GLP-1 is prepro-GG-(72 – 108), and GLP-2 is prepro-GG-(126 – 159). In the pancreatic A-cells, prepro-GG is cleaved to GG and longer C-terminal peptides that contain GLP-1 and GLP-2. In the small intestine, however, prepro-GG is cleaved to form glicentin [prepro-GG-(1 – 69)], enteroglucagon [prepro-GG-(33 – 69)], GLP-1, and GLP-2 [1464]. In the small intestine the six N-terminal amino acids of GLP-1 are removed to give the C-terminal amide GLP-1-(7 –36)amide [prepro-GG-(78 – 107)amide] [1465].

The sequences of h-prepro-GG [1461] and its cleavage products are shown in Figure 5.

7.4.1. Glucagon

Glucagon (GG) [*642-83-1*], M_r 3483 (h-GG), is formed in the endocrine A- cells of the pancreas. It is, however, also found in the rat pituitary.

Release. Pulsatile GG release is observed in humans at 13 – 20 min intervals [1466]. The Ca^+-dependent secretion of GG is stimulated by amino acids, stress [1467], β-endorphin [1468], morphine [1469], dopamine, carbachol, cholecystokinin [1470], bombesin, xenopsin, substance P, and high doses of oxytocin and vasopressin. Parathyroid hormone potentiates the arginine-stimulated release of GG.

GG secretion is inhibited by glucose [1471], Mg^{2+}, thyroliberin, calcitonin, SRIH, glicentin-(1 – 16), glicentin-(62 – 69) [1472], and sulfonyl

```
1                                                   30
R S L Q D T E E K S R S F S A S Q A D P L S D P D Q M N E D K R–

33                                  53          61
H S Q G T F T S D Y S K Y L D S R R A Q D F V Q W L M N T–
←————————————————————— Glucagon —————————————————————→

62  64        69  71
K R N R N N I A K R
    ←— SpP–I —→

72          78                                                  108
H D E F E R H A E G T F T S D V S S Y L E G Q A A K E F I A W L V K G R G–
←————————————————————————————— GLP I —————————————————————————————→

109 111                 123 125
R R D F P E E V A I V E E L G R R–
    ←———— SpP–II ————→

126                                                         159
H A D G S F S D E M N T I L D N L A A R D F I N W L I Q T K I T D R
←————————————————————— GLP II —————————————————————→
```

Figure 5. Amino acid sequences of h-prepro-GG and its cleavage products

urea derivatives, e.g., glibenclamide [1473]. The plasma level of GG is elevated in overweight persons [1474] and is reduced in type I diabetics [1475].

Biological Effects [1476]. Glucagon plays a role in carbohydrate, fat and protein metabolism. In the liver, its most important target organ, GG stimulates the formation of glucose from glycogen (glycogenolysis) and of ketone bodies from amino acids. Glycogen synthesis is simultaneously inactivated via a phosphorylation reaction [1477] and the binding of insulin in the liver is reduced [1478]. Intracerebroventricular administration of GG increases blood glucose levels via cholinergic and α-adrenergic mechanisms much more than intravenous administration [1479].

In hepatocytes, GG inhibits fatty acid synthesis via activation of a cAMP-dependent protein kinase and subsequent phosphorylation of an acetyl coenzyme A (CoA) carboxylase [1480]. It stimulates lipolysis in brown adipose tissue [1481]. It reduces the activity of β-hydroxy-β-methylglutaryl CoA reductase [1482] and lowers the plasma level of apolipoprotein E, cholesterol, triacyclglycerol, and phospholipids [1483].

Other effects of GG include stimulation of the plasma level of cAMP, reduction of the blood levels of Ca^{2+} and PO_4^{3-}, spasmolytic activity in high doses, relaxation of the esophagus, contraction of the pyloric sphincter [1484], relaxation of the Oddi sphincter [1485], stimulation of bile flow [1486], bronchodilatory effect in asthmatic patients [1487], and inhibition of anaphylactic histamine release.

Structure – Activity Relationships. The shortening of GG and the replacement of its N-terminal histidine reduces glycogenolytic activity [1488]. GG-(1 – 21) is fully active as regards spasmolytic activity but only about 0.1 % as active as regards cAMP-releasing activity. Thus, GG-(1 – 21) exerts a relaxing effect on the Oddi sphincter [1489] and an analgesic effect in biliary colic [1490]. Further shortening of the compound, e.g., GG-(1 – 20) and GG-(2 – 21), reduces the spasmolytic activity to $\leq 2\,\%$ of the GG activity. GG-(19 – 29) does not activate adenylate cyclase, but is 1000 times more active than GG in inhibiting the Ca^{2+}-activated and Mg^{2+}-dependent ATPase [1491] and the liver plasma membrane calcium pump [1492].

Substitutions in the N-terminal region of GG usually lower biological activity. However, [D-Ser^2]GG [1493] and [D-Phe^4]GG had twice and six times the glycogenolytic activity in the rat, respectively [1494].

The amino acid region 10 – 13 is not essential for biological activity, but is important for receptor binding [1495].

S-Methylation of Met^{27} reduces glucagon activity to 1 %; [*S*-carboxymethyl-Met^{27}]GG [1496], [2-thiol-Trp^{25}]GG, [oxindolyl-Ala^{25}] GG, [$Met(O_2)^{27}$]GG, monoiodo-GG and [*N*-ε-alkylamidino-Lys^{12}]GG are fully active.

N-α-Trinitrophenyl-GG and *N*-α-trinitrophenyl-[12-homoarginine]GG are weak antagonists that exert a partially agonistic effect. Antagonists include de-His^1-[Glu^9]GG-amide [1497], de-His^1-[Glu^9,Orn^{12}]GGamide, de-His^1-[Glu^9,Glu^{16},Lys^{29}]GG-amide, de-His^1-[Glu^9,Lys^{29}]GG-amide [1498], [Asp^3,D-Phe^4, Ser^5, $Lys^{17,18}$,Glu^{21}]GG, and [D-Phe^4,Tyr^5,I_2-Tyr^{10},Arg^{12}, $Lys^{17,18}$,Glu^{21}]GG [1499].

Uses. Glucagon (Glucagon Novo, Novo Ind.; Glukagon Lilly, Eli Lilly) is used to treat hypoglycemia. In McArdle's disease, subcutaneous injection of glucagon increases the blood sugar level and thus improves muscle function.

The spasmolytic activity of GG facilitates endoscopic and radiological investigations of the gastrointestinal tract and computer tomography of the liver, pancreas, and kidney. GG can also bring relief to persons suffering from esophageal cramps [1500] and hiccups [1501]. Its spasmolytic activity in combination with the enhanced flow of bile has been useful in the treatment of biliary colic.

As a result of its effect on the plasma Ca^{2+} level, glucagon has been used in combination with calcitonin and mithramycin in the treatment of Paget's disease.

GG can be applied parenterally (also with a portable pump [1502] or as a depot preparation [1503]), nasally [1504], or as eye drops [1505].

7.4.2. Glicentin and Enteroglucagon

Glicentin [*71567-776*], M_r 8101, [h-prepro-GG-(1 – 69)] and enteroglucagon [*62340-29-8*], M_r 4450, also called oxyntomodulin [h-prepro-GG-(33 – 69)], are synthesized in the small intestine from prepro-GG [1464]. The plasma level of glicentin rises during hypoglycemia [1506] and physiological doses of both peptides decrease the secretion of gastric acid [1507]. Glicentin has only 10 % of the cAMP-stimulating activity of GG. Thus, glicentin and enteroglucagon assume the function of a hypothetical enterogastrone in the intestine (an enterogastrone is a substance formed in the intestine which decreases gastric acid release). The common C-terminal octapeptide appears to be especially important for activity and can inhibit the pentagastrin-stimulated secretion of gastric acid (although 100 times weaker) [1508].

Enteroglucagon is also 10 times more active than GG in inhibiting cerulein-stimulated exocrine pancreatic secretion in rats [1509].

7.4.3. Glucagon-Like Peptide-1-(7 – 36)amide

The peptide GLP-1 is processed primarily in the small intestine to give the C-terminal amide GLP-1-(7 – 36)amide, M_r 3298 (h-GLP-1-(7 – 36)-amide) [1465].

Release. Intestinal release of GLP-1-(7 – 36)amide is stimulated by orally applied glucose, D-galactose, D-glucuronic acid, 3-*O*-methyl-D-glucose, maltose, sucrose, or maltitol [1510]. Fructose, D-fucose, D-mannose, D-xylose, and lactose have no effect [1510]. The postprandial increase in the plasma level of GLP-1-(7 – 36)amide is enhanced in the post-gastrectomy dumping syndrome [1511].

Receptors. GLP-1-(7 – 36)amide binds specifically to a cloned, insulin-producing cell line from the rat. Even micromolar concentrations of GLP-1-(1 – 36)amide or GIP do not displace GLP-1-(7 – 36)amide from the receptor [1512]. Dexamethasone reduces the number of GLP-1-(7 –36)amide receptors on the B-cells [1513].

Biological Effects. GLP-1-(7 – 36)amide is more effective in stimulating cAMP synthesis in rat preparations than GLP-1 or glucagon [1514], [1515]. GLP-1-(7 – 36)amide is more active than GLP-1 in stimulating the glucose-dependent release of insulin from pancreatic islet cells of the rat [1515]. Tests on humans have shown that GLP-1-(7 – 36)amide has the properties of inkretin (a hypothetical substance that stimulates the glucose-dependent endocrine secretion from the pancreas): its formation in the gastrointestinal tract is enhanced after an oral glucose meal, it releases insulin in a glucose-dependent process, and lowers the plasma levels of glucose

and glucagon [1511]. GLP-1-(7 – 36)amide and GIP have a synergistic effect on the release of insulin [1516] but not in the stomach. In the pancreas GLP-1-(7 – 36)amide inhibits the secretion of glucagon by post-transcriptional mechanisms and decreases the secretion of somatostatin [1517], [1518].

Structure – Activity Relationships. The amide structure of GLP-1-(7 – 36)amide is apparently not essential for its insulin-releasing activity. GLP-1-(7 – 37) is as active as GLP-1-(7 – 36)amide [1519].

Uses. In type II diabetes (non-insulin-dependent), GLP-1-(7 – 36)amide stimulates the release of insulin and inhibits the secretion of glucagon and somatostatin. Both in type I (insulin-dependent) and type II diabetes, the isoglycemic-meal- related insulin requirement is lowered by GLP-1-(7 – 36)amide [1520].

7.5. Gastrin Inhibiting Peptide

Occurrence. The gastrin inhibiting peptide (GIP) [*59392-49-3*] also known as the glucose-dependent insulinotropic peptide, M_r 4984, contains 42 amino acids and is found in the duodenum, jejunum, and together with glucagon in the A-cells of the pancreas.

Release. GIP is secreted in the duodenum and jejunum in response to a diet rich in fat and carbohydrate. The pancreatic enzymes are also involved in stimulating this release [1521]. During physical exercise, the secretion of GIP in response to orally administered glucose is greatly reduced [1522]. The effect of GIP is enhanced by bombesin or low doses of motilin. High doses of motilin, somatostatin, and dexamethasone inhibit the release of GIP [1523].

The raised level of GIP in insulin-dependent (type I) diabetes can be lowered with insulin [1524]. This is, however, not the case with non-insulin-dependent diabetes mellitus [1525]. The postprandial plasma level of GIP is low in obese, non-insulin-dependent diabetics [1526].

Biological Effects. GIP inhibits the gastrin-stimulated secretion of gastric acid, as well as the secretion of pepsin, and motor activity in the stomach. It stimulates the intestinal secretion and increases the glucose-dependent release of insulin, glucagon, pancreatic peptide, and somatostatin in the pancreas. cAMP is activated in this process. The GIP-mediated release of insulin is dependent on phospholipase A_2 [1527] and is linked to a cholinergic mechanism [1528].

In adipocytes, GIP increases the receptor affinity for insulin [1529] and potentiates the inhibitory effect of insulin on hepatic glycogenolysis [1530].

The physiological importance of GIP is probably based on its insulinotropic activity rather than on the reduction of the gastric acid concentration. Like GLP-1-(7 – 36)amide, GIP is an important candidate for the role of the postulated inkretin.

Structure – Activity Relationships. Shortening the molecule at the N-terminus by only two amino acids reduces the insulin-releasing effect to about 2 %. Endo-29a-Gln-GIP-(1 – 38) and b-GIP-(1 – 39) [1531] still possess full insulinotropic activity. GIP-(1 – 28) and GIP-(3 – 24) do not release insulin [1532].

7.6. Growth Hormone Releasing Hormone and the Growth Hormone Cascade

The growth hormone cascade is shown in Figure 6. The growth hormone releasing hormone or somatoliberin (GHRH or SRH) and galanine (GAL) stimulate the release of the growth hormone (also known as somatotropin, STH) from the pituitary which, in turn, releases the somatomedins (SMs). Somatostatin (SRIH), which inhibits the release of GHRH and GH, is of great importance in the regulation of growth hormone. GAL inhibits the release of SRIH, which promotes the release of GHRH and GH. GHRH and GH stimulate the synthesis of SRIH and thus permit a negative feedback. A synthetic hexapeptide (SK&F 110 679) also releases GH from the pituitary, independently of GHRH receptors.

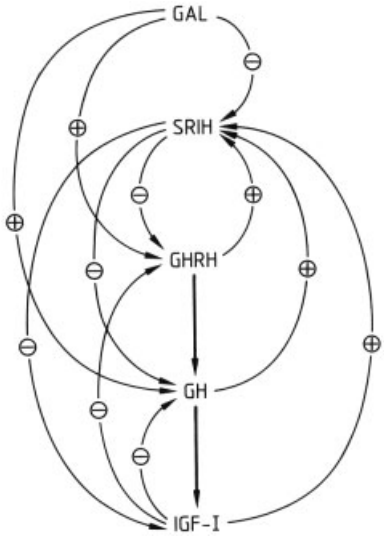

Figure 6. The growth hormone cascade
GAL = galanine; GH = growth hormone (somatotropin); GHRH = growth hormone relesing hormone (somatoliberin); IGF-I = insulin-like growth factor I; SRIH = somatotropin release inhibiting hormone (somatotropin)

There is a high degree of structural similarity between GH and placental lactogen, somatolactin, prolactin, proliferin, and prolactin-related or -like proteins. The insulin-like growth factors, IGF-I (somatomedin C) and IGF-II, as well as somatomedin B belong to the somatomedins. Insulin and relaxin are structurally related to IGF-I and IGF-II.

7.6.1. Galanine [1533]

Occurrence. Galanine (GAL) [*119418-04-1*], M_r 3211 (p-GAL), was first isolated from the pig intestine [1534]. Its name is derived from its N- and C-terminal amino acids (glycine and alanine). It is a 29 amino acid peptide amide with a C-terminal that is similar to the neurokinins, e.g., neurokinin A or substance P. Both p- and b-GAL are processed in the form of a prepropeptide containing 123 amino acids [1535].

In humans, GAL is not only found in the nerve cells of the small intestine, but also in the brain, spinal cord, pancreas, bladder, pituitary [1536], adrenal gland [1537], and the genital tract [1538]. In the central nervous system, the highest concentrations are found in the hypothalamus [1539] and the median eminence. Unusually high concentrations of GAL are found in pheochromocytomas [1540].

GAL is localized with gonadoliberin in neurons of the hypothalamus [1541].

Release. The release of GAL in the adenohypophysis and hypothalamus is stimulated by estrogen [1536] and the thyroid hormones, e.g., T_4 [1542]. The formation of GAL decreases after the withdrawal of food [1543] and treatment with capsaicin.

Biological Effects. GAL contracts preparations from the rat fundus, ileum, colon, bladder [1534], and jejunum [1544]. It potentiates the contraction of the rat vas deferens that is induced by electrical stimulation or noradrenaline [1545]. GAL inhibits, however, contraction induced by electrical stimulation in the guinea pig colon. It also inhibits the motility of the small intestine in the rabbit. GAL inhibits the release of acetylcholine and substance P from motor nerves [1545] and transmission at the cholinergic nicotinic synapses in the myenteric plexus [1546].

In humans, GAL inhibits gastrointestinal motiliy, delays gastric emptying, and increases the mouth to cecum transit time. It also suppresses the postprandial increase in glucose, PYY, insulin, neurotensin, enteroglucagon, glucagon, pancreatic peptide (PP), and somatostatin (SRIH) levels [1547]. GAL directly inhibits the exocytosis of insulin [1548]. Glucagon is stimulated in vitro. In mice, GAL inhibits the basal and stimulated insulin level, increases the plasma glucagon level temporarily and causes hyperglycemia [1549]. In the perfused porcine pancreas, however, GAL increases the release of insulin and inhibits the secretion of SRIH [1550].

GAL inhibits the release of gastrin and SRIH [1551] and the pentagastrin-stimulated secretion of gastric acid from the rat stomach [1552].

The intracerebroventricular application of GAL to rats or the intravenous infusion in humans increases the plasma levels of growth hormone and prolactin [1539] via the release of adrenaline [1553] and GHRH [1554], or via the inhibition of SRIH. It acts synergistically with GHRH [1555].

In rats, injections of GAL into the paraventricular nucleus cause an increase in the food

uptake. It protects hippocampal neurons from the functional effects of anoxia [1556].

Structure – Activity Relationships. The molecule can be shortened at the C-terminus without loss of activity; in contrast, deletions at the N-terminus result in complete loss of activity.

GAL-(1 – 10) strongly contracts the rat jejunum, but it does not inhibit the electrically stimulated contraction of strips of guinea pig colon [1544]. The action of GAL-(1 –20) is species and muscle dependent, and is similar to or weaker than that of GAL [1557]. While GAL-(1 – 16) has an agonistic effect on the rat hippocampal GAL receptor, GAL-(3 – 29) [1558] and GAL-(17 – 29) [1559] are inactive. The effect on the guinea pig ileum is as follows: GAL > GAL-(2 – 9) > GAL-(1 – 15) [1560]. In rats, the pentagastrin-stimulated secretion of gastric acid is inhibited by GAL-(9 – 29) with 5 % of the GAL activity, while GAL-(15 – 29) is devoid of activity [1552].

7.6.2. Growth Hormone Releasing Hormone [1561]

Occurrence. In humans, growth hormone releasing hormone (GHRH) [*9034-39-3*], also called somatotropin releasing hormone (SRH) or somatoliberin, M_r 5040 (h-GHRH), is a peptide amide containing 44 amino acids. Although GHRH is mainly synthesized in the hypothalamus, immunoreactive GHRH has also been found in the gastrointestinal tract, placenta, human and rat milk, and spinal fluid [1562].

In the neurons of the arcuate nucleus, GHRH is associated with NPY [1563], dopamine [1564], and tyrosine hydroxylase [1565].

During the processing of prepro-h-GHRH, prepro-h-GHRH-(78 – 107)amide with a pyroglutamyl residue at the N-terminus (CTPG or anorectin) is formed. This peptide reduces the intake of food when injected into the third ventricle [1566].

Release. The release of GHRH is stimulated by central α_2-adrenergic stimuli, high potassium concentrations, GABA, opioids, galanine [1567], dopa [1568], thyroid hormones, and by a glucose meal [1569]. More GHRH is found in the hypothalamus of male rats than in female rats [1570]. The plasma level of immunoreactive GHRH increases greatly during puberty (more so in girls than in boys) [1571].

Somatostatin (SRIH) and pentobarbital anesthesia inhibit the secretion of GHRH. Growth hormone [1572], insulin-like growth factors IGF-I and IGF-II [1573], GHRH itself [1574], and dopamine also inhibit the release of GHRH by increasing SRIH release. The GHRH-stimulated secretion of SRIH is presumed to be regulated by β-endorphin-containing neurons [1575].

The plasma level of GHRH is lower in children suffering from dwarfism and in acromegaly [1576]. Less GHRH is found in the hypothalamus of old rats than in young rats [1577].

Receptors. Chronic infusion (in vivo) of GHRH [1578] and adrenalectomy [1579] lead to the down regulation of GHRH receptors in the rat.

Biological Effects. GHRH stimulates transcription of the GH gene and the release of GH [1580] in the pituitary via an increase in cAMP, arachidonic acid, and PGE_2. Chronic administration does not totally inhibit the release of GH [1581]. It is assumed that melatonin [1582], cholinergic agonists [1583], β_1-adrenergic receptor blockers [1584], and hypoglycemia [1561] increase the GHRH-stimulated GH level via the inhibition of SRIH release. The GHRH-stimulated release of GH is inhibited by SRIH, cholinergic antagonists [1583], α-receptor blockers [1585], dopamine [1586], hyperglycemia [1587], and by IGF-I and II [1573]. The GHRH-induced release of GH is reduced in children with thalassemia major [1588], in patients suffering from non-insulin-dependent (type II) diabetes [1589], hypothyroidism [1590], Cushing's syndrome [1591], depression [1592], and obesity [1561]. The GHRH-stimulated secretion of GH is raised in diabetics with retinopathy [1561], in acromegaly, and in young Alzheimer patients [1593]. In patients with hypothyroidism, T_4 potentiates the GHRH-induced release of GH [1590].

Glucocorticoids exert a biphasic effect on the release of GH. They potentiate GHRH activity

by stabilizing the GHRH receptors and enhancing the transcription rate of the GH gene. In cases of chronic misuse of glucocorticoids, however, an inhibition of the release of GH is observed in humans and experimental animals [1594].

Other effects of GHRH include release of calcitonin and neurotensin from rat C-cells [1595], proliferation of endothelial cells in the mucosa of the fundus and antrum [1596], increase in the production of milk in cows and sheep, and increase in REM sleep [1597].

Structure – Activity Relationships. GHRH can be shortened at the C-terminus without loss of activity. In rats, h-GHRH-(1 – 29)amide has 50 % of the activity of h-GHRH-(1 – 44)amide. However, h-GHRH-(1 – 29)amide has the same activity as h-GHRH-(1 – 44)amide in children [1598]. Further shortening at the C-terminus significantly reduces activity [1599]. The N-terminal amino group is of no importance for biological activity because its acetylation gives more potent compounds.

Position 1 should be occupied by an aromatic amino acid [1600]. Nle^{27} and D-Ala^2 increase activity. Although [D-Ala^2]GHRH-(1 – 29)amide had 50 times the activity in some tests and Ac-[D-Tyr^1,D-Ala^2]-GHRH-(1 – 29)amide had 38 times the GHRH-(1 – 29)amide activity in the rat, these compounds do not exhibit a higher activity in humans [1601]. The replacement of Arg^{29} by agmatin (Agm) increases activity. Potent long-acting analogues are [D-Ala^2,Nle^{27}]h-GHRH-(1 – 28)Agm [1602] and [deamino-Tyr^1,Ala^{15},Nle^{27}]GHRH-(1 – 28)Agm [1603].

[D-Arg^2]h-GHRH-(1 – 29)amide exhibits GHRH antagonistic activity in concentrations of 10^{-7} mol/L [1604]. More potent antagonists are [D-$Arg^{2,29}$,Arg^{30}]h-GHRH-(1 – 30)amide and Ac-[His^1,D-Arg^2,Ala^{15}]r-GHRH-(1 – 29)amide [1605].

An antagonist that acts in the 10 µmol/L range is [Ser^9-ψ-CH_2NH-Tyr^{10}]GHRH-(1 – 29)amide [1606].

The activity of h-GHRH-(1 – 27)amide can also be increased: [D-Asp^3,D-Asn^8,Leu^{27}]h-GHRH-(1 – 27)amide has twice the activity of h-GHRH-(1 – 29)amide [1607].

Uses. GHRH can be used as a diagnostic aid to differentiate between hypothalamic- and hypophyseal-dependent growth hormone deficiency diseases. Only patients with a defective hypothalamus usually react to GHRH by producing GH. The older the patient (i.e., the longer the endogenous GHRH deficiency disease exists) the poorer the reaction of the hypophysis.

GHRH is used in the treatment of children with GH deficiency. The linear growth of children was improved by subcutaneous administration of 250 µg of GHRH-(1 – 29)amide twice a day [1608] or after pulsatile treatment (subcutaneous application of 1 – 3 µg/kg every 3 h) [1609]. Children suffering from dwarfism responded better to h-GH than to h-GHRH-(1 – 29)amide [1610]. In postmenopausal women GHRH increases bone formation and plasma osteocalcin [1611].

Of the large number of smaller synthetic peptides [1612], one of the most active is the hexapeptide amide SK&F 110 679 (sequence: H w A W f Ka), which has about 10 % of the GHRH activity [1613]. There is no structural similarity between this peptide and GHRH. It releases GH from hypophyseal cell cultures in vitro and acts synergistically with GHRH. Thus, it evidently does not bind to the same receptor as GHRH [1614]. In humans, doses as low as 0.25 – 1 µg kg^{-1} 30 min^{-1} stimulate the release of GH. The compound appears to be more effective in this experiment than GHRH [1615].

7.6.3. Growth Hormone, Placental Lactogen, and Prolactin

The pituitary hormones prolactin (PRL) and growth hormone (GH) are structurally related to placental lactogen (PL) and two GH variants (GH-V and GH-V2) from the placenta. PL has lactogenic properties and some somatotropic activity (tibia test). PRL has no somatotropic activity, but GH has weak PRL activity.

Other substances related to PRL are somatolactin (pl-SL) [1616], proliferin (m-PLF), prolactin related cDNA (b-PRC I, II, and III), proliferin related protein (m-PRP), and prolactin-like proteins (r-PLP A and B, b-PLP I, II, and III) [1617].

There is great similarity in the structures of secretin-(5 – 20), h-GH-(1 – 16), and h-PL-(1 – 16). There is also amino acid homology

between h-GH-(30 – 38), gastrin, and the C-terminus of the B-chain of insulin.

7.6.3.1. Growth Hormone

Occurrence. Growth hormone (GH) [*9002-72-6*], also known as somatotropin (STH), M_r 22 125 (h-STH), has 191 amino acids and is formed in the pituitary as pro-GH. Circulating GH occurs in several forms: monomeric (little GH), dimeric (big GH) and oligomeric (big-big GH, $M_r > 60\,000$) [1618]. In addition, 20Kh-GH, a de-(32 – 46)GH, and the two GH variants from the placenta, GH-V and GH-V2 [1619] have also been found. The amino acid sequences of GH in various species are described in [1620], [1621].

Release [1622]. GH is regulated primarily by the hypothalamic hormones galanine (GAL, see Section 7.6.1), growth hormone releasing hormone (GHRH, see Section 7.6.2), and somatostatin (SRIH, see Section 7.6.5). GAL and GHRH stimulate and SRIH inhibits the pituitary release of GH. Treatment with SRIH enhances the GHRH-stimulated release of GH [1623]. GH is also released in response to stimulation of α_1-adrenergic receptors [1624] and cholinergic muscarinic receptors [1625], inhibition of α_2-adrenergic receptors [1626], SKF-110 679, insulin, hypoglycemia, glucagon, arginine, physical stress [1627], α-melanotropin, opioids [1628], angiotensin II [1629], pentagastrin [1630], GIP, neurotensin, combined application of thyroliberin with ACTH [1631], slow wave sleep [1632], intracerebroventricular application of delta-sleep-inducing peptide, interleukin-1β [1633], serotonin [1634], dopamine and dopamine agonists [1635], and indomethacin [1636].

In women, the release of GH is dependent on estrogen. Thus, the plasma level of GH falls during menopause [1637] and in suppression therapy with high doses of GnRH agonists [1638]. In young hypogonadal males, testosterone stimulates the GH formation and growth [1639]. Glucocorticoids also regulate the synthesis of GH. Hypoadrenalism [1640] and high doses of glucocorticoids [1641] lower the release of GH. However, glucocorticoids and triiodothyronine (T_3) enhance the release of GH in vitro through GHRH.

GH release is inhibited by recombinant Met-GH [1642], centrally applied gastrin releasing peptide [1643], corticoliberin [1644], calcitonin [1645], calcitonin gene related peptide [1646], and by TRH [1647], mainly via the secretion of SRIH.

Plasma GH levels in persons with acromegaly and type I diabetes [1648] are higher than in healthy persons, but are reduced in depressive patients with an increased level of somatomedin C [1649]. TRH has no influence on the release of GH in normal subjects, but it causes a paradoxical increase in the plasma level of GH in patients suffering from acromegaly [1650], depression [1651], or metastatic testicular cancer [1652]. GnRH also raises the plasma GH level in type I and type II diabetics [1653].

Receptors. The amino acid sequences of the GH receptors from humans, rabbits [1654], and rats [1655] have been elucidated via their cDNA.

GH binding proteins are mainly the extracellular fragments of GH receptor [1656]. In patients with Laron-type dwarfism (high plasma GH level and low IGF-I level), the GH receptor is defective [1657] and the high-affinity GH binding protein is missing [1658]. African pygmies have a low plasma level of GH-binding protein [1659].

Biological Effects. GHRH is species specific and promotes the longitudinal growth and regeneration of bone. It has a strong mitogenic effect on osteoblasts [1660] and pancreatic β-cells [1661], increases the osteocalcin level [1662], and enhances DNA, RNA, and protein synthesis. It possesses lipolytic, diabetogenic, and insulinotropic properties; it also has an antidiuretic and antinatriuretic effect because it activates renin and increases the plasma level of aldosterone [1663].

In a negative feedback mechanism, GH inhibits the release of GHRH in the hypothalamus [1664] via the release of SRIH [1665].

Pulsatile intravenous GH stimulates the formation of somatomedin C [1666] in the liver and skeletal tissue. This substance is identical to the insulin-like growth factor I (IGF-I) and is largely responsible for the growth effect of GH [1667].

In vitro, GH stimulates the release of insulin from pancreatic islet cells [1668] and the insulin-dependent uptake and oxidation of glucose in adipose and muscle tissues [1669]. The insulin-like effect of GH is only found in vivo in GH deficiency models or children suffering from GH deficiency. Chronic therapy with GH or very high doses of GH have a diabetogenic effect in hypophysectomized animals. After prolonged GH treatment the fasting plasma glucose level in humans is raised [1670] and in patients with acromegaly, glucose metabolism is reduced in spite of the increased release of insulin [1671].

Physiological doses of GH also stimulate the production of peroxide anions in macrophages. It increases the activity of cytolytic T-cells and of natural killer cells, stimulates the formation of antibodies and thymulin synthesis [1672], and regenerates the thymus in dogs [1673].

Structure – Activity Relationships. Similar to h-GH, recombinant Met-h-GH, which acts like natural h-GH in children with GH deficiency [1674], has diabetogenic and insulin-like activity [1675]. Met-h-GH has a somewhat higher tendency to initiate antibody formation [1676]. Other acidic GH derivatives (deamidated h-GH) are as active as h-GH [1620].

The epiphyseal plate of the long bones of the rat declines following hypophysectomy and increases after treatment with GH. This increase in the width of the plate of the proximal end of the tibia has been used for the quantitative measurement of GH. Various GH fragments exhibit biological activity in the tibia test in hypophysectomized animals, e.g., GH-(88 – 124), GH-(125 – 156), b-GH-(96 – 133), and GH-(151 – 191). h-GH, but not 20K-h-GH, reduces the concentration of free fatty acids in children deficient in GH [1677]. The binding of 20K-h-GH to lactogenic receptors is about 500 times less than that of GH [1678]. The lactogenic activity of h-GH and o-PRL is competitively inhibited by [Met^{14}]h-GH-(14 – 191) [1621].

h-GH-(4 – 15) and h-GH-(6 – 13) stimulate glucose oxidation, glucose uptake in adipose tissue, and the synthesis of glycogen [1679]. The minimal GH sequence with this insulin-potentiating effect is h-GH-(8 – 13). An artifact, the aspartoylimide variant of these peptides, appears to be responsible for this effect [1680]. Other insulinotropic GH fragments are h-GH-(31 – 44), h-GH-(32 – 38) [1681], and h-GH-(32 – 46) [1682]. Both h-GH-(177 – 191) and h-GH-(178 – 191) have diabetogenic activity.

Uses. Met-h-GH and h-GH are produced by genetic engineering and are now used instead of natural h-GH in the therapy of hypophyseal dwarfism. Trade names include Genotropin (Pfrimmer Kabi), Grorm, Saizen (Serono), Humatrop (Lilly), and Norditropin (Nordisk).

h-GH has also been used to activate growth in persons with Turner's syndrome where growth is reduced in spite of normal GH formation [1683]. h-GH also finds application for the healing of wounds [1684] and for increasing the mass [1685] and density of bones [1686]. In older patients with low plasma GH levels, GH increases bone density, lean body mass, and skin thickness, and reduces the amount of adipose tissue [1687].

7.6.3.2. Prolactin

Occurrence. Prolactin (PRL) [*9002-62-4*], M_r 22 806 (h-PRL, protein part), has 198 amino acids and is formed in the pituitary where it is glycosylated at Asn^{31} [1688] or dimerized via disulfide bridges [1688]. Other sites of synthesis are the placenta, myometrium, ovaries [1689], and hypothalamus. The sequences of PRL from various species are listed in [1690].

Release [1691]. The pulsatile secretion of PRL from the pituitary is regulated by hypothalamic factors. It is stimulated by TRH and VIP and inhibited by dopamine. The C-terminal glycopeptide of vasopressin-neurophysin, which contains 39 amino acids, is a specific PRL-releasing factor [1692]. PRL is also released in response to K^+ [1693], hypocalcemia [1694], calcitriol [1695], cholecystokinin, insulin, arginine [1696], epidermal growth factor [1697], the sucking stimulus through endogenous opioids [1698] and VIP [1699], stress and physical exercise [1700] through serotoninrgic mechanisms [1701], β-endorphin [1702], AT II [1703], gonadoliberin [1704] through the formation of AT II [1705], neurotensin [1706], and to glutamate [1707]. It is also released by the stimulation of α_1-adrenergic receptors [1708] or histamine H_1-receptors [1709].

Dopamine from the hypothalamus and pituitary [1710] which is released by PRL is probably one of the most important inhibitors of PRL release.

The secretion of PRL is also inhibited by ascorbic acid [1711], δ opiates [1712], hypercalcemia [1713], the prolactin inhibiting factor (PIF), PRL itself [1714], stimulation of α_2-adrenergic [1708], GABA-A- [1715], or histamine H_2-receptors [1709], chronic nicotine intake [1716], somatostatin and its analogues [1717], calcitonin [1718], IL-1 [1719], T_3 [1720], melatonin [1721], glucocorticoids [1722], vasopressin (presumably via a dopaminergic pathway [1723]), and central neurotensin. A competitive antagonist of PRL is a recombinant h-GH, which lacks the 13 amino acids at the N-terminus [1724].

The steroid sex hormones modulate the release of PRL. Estrogen potentiates the secretion of PRL in experimental animals [1725], but can also potentiate the dopamine-stimulated inhibition of PRL release in humans. 17-β-estradiol potentiates the somatostatin-stimulated inhibition of PRL in pituitary cells [1726]. Testosterone appears to inhibit the secretion of PRL because the TRH-stimulated release of PRL is greatly enhanced in delayed puberty (males) [1727].

A period of breast feeding is followed by a reduced PRL level for at least 12 – 13 years [1728], which could be related to the low risk of breast cancer after pregnancies. Women who are not lactating have a pronounced PRL deficiency [1729]. In humans, the PRL response to TRH falls with age [1730]. Patients with chronic kidney damage and rats with experimentally induced kidney damage develop hyperprolactinemia. The plasma level of PRL is raised in spontaneously hypertensive rats and in patients with hypertension.

Receptors. Receptors for PRL are widely distributed: in the mammary gland, liver, kidneys, brain, prostate, testes, ovaries, and lymphocytes [1731]. The receptors have a similar structure to that of the GH receptors [1732], [1733].

In the rat, the number of PRL receptors in the mammary gland increases on the day of estrus. Together with estradiol, LH stimulates the formation of PRL receptors [1734]. The number of testicular PRL receptors in the golden hamster is increased by PRL, FSH, and LH; ovariectomy and hypophysectomy reduce the synthesis of PRL receptors [1735].

PRL is displaced from the receptors on lymphocytes by cyclosporin [1731].

Biological Effects. In female mammals, PRL increases the production of milk and initiates maternal behavior. The PRL-stimulated biosynthesis of casein and lipid is reduced by inhibitors of phospholipase A_2 [1736].

PRL stimulates the synthesis of dopamine and increases the density of dopamine receptors [1737]. Thus, the dopamine level in the central nervous system and the dopaminergic inhibition of aldosterone are increased in patients suffering from hyperprolactinism [1738].

Similar to placental lactogen, PRL also has luteotropic properties and stimulates the synthesis of progesterone [1739], but inhibits the formation of estradiol and testosterone [1739]. In men with chronically elevated PRL levels, the plasma level of testosterone is low and the excretion of LH in the urine is reduced [1740]. In women, hyperprolactinemia leads to irregular menstruation or even amenorrhea [1741].

PRL increases the blood pressure of experimental animals which is dependent on the adrenal glands and presumably occurs via the potentiation of noradrenaline activity. As a stress hormone, PRL potentiates the antidiuretic activity of Arg-vasopressin [1742] and the AT IIinduced uptake of water and retention of fluids [1743].

Hyperprolactinemia increases the production of androgens [1744] in the adrenal gland, DNA synthesis in the liver [1745], the susceptibility to chemically induced tumors [1746] and spontaneous mammacarcinoma, and the number of estrogen receptors in the mammary gland [1747]. It also prevents stress-induced stomach ulcers [1748], induces hypercalcemia in several species [1749], and reduces the mineral content of the bone [1750].

PRL is also immunostimulatory. Thus, the reactivity of the lymphocytes is reduced by blocking pituitary PRL release with bromocryptin. The immunosuppressive effect of cyclosporin can be abolished by PRL [1731].

7.6.3.3. Placental Lactogen

Placental lactogen (PL) [*9035-54-5*], M_r 22 212 (h-PL), is formed in the placenta in the first trimester of pregnancy and can be detected in the blood and urine of pregnant women. Secretion of PL decreases after birth.

h-PL first stimulates the development of the breast tissue. When the level of h-PL decreases after birth, pituitary h-PRL stimulates the secretion of milk.

h-PL has luteotropic properties: it stimulates the release of progesterone and estrogen from the corpus luteum. This effect is potentiated by h-choriogonadotropin (h-CG). The growth-promoting effect of h-PL is somewhat weaker than that of h-STH, but appears to play a more important role in the development of the fetus.

7.6.4. Somatomedins, Insulin, and Relaxin

Somatomedins of the IGF type (insulin-like growth factors I and II) are structurally related to insulin and relaxin. They are synthesized as single-chain prepropeptides. In the case of insulin and relaxin, three disulfide bridges are formed and the C-peptide (connecting peptide) between the N-terminal B-chain and the C-terminal A-chain is enzymatically removed to give two chains joined together by two disulfide bridges. In the case of IGF-I ($M_r = 7648.8$), and IGF-II ($M_r = 7469.5$), the C-peptide is not cleaved. Relaxin has no influence on the level of glucose but IGF-I, IGF-II, and insulin are functionally related.

7.6.4.1. Somatomedins [1751]

Occurrence. The somatomedins C [*67763-96-6*] and A [*62046-94-0*] (SM-C and SM-A) were previously called the sulfation factor, thymidine factor, nonsuppressible insulin-like activity in serum (NSILA-S), or the multiplication stimulating activity (MSA). The insulin-like growth factor I (IGF-I) is identical to SM-C, SM-A appears to be a deamidated IGF-I [1752]. Somatomedin B [*63774-77-6*] (SM-B) has a completely different structure which is similar to the N-terminal 44 amino acids of vitronectin.

The somatomedins are found in the liver, lungs, prostate, testes, seminal and follicular fluid [1753], thymus, thyroid gland [1754], skeleton, brain, heart, kidneys [1755], gastrointestinal tract, saliva [1756], milk [1757], fibroblasts, chondrocytes [1758], skeletal muscles, and in adipose deposits. A large number of tumors produce the messenger RNA for IGF-I and/or IGF-II [1759].

h-IGF-I, M_r 7649, and h-IGF-II, M_r 7470, each consists of a peptide chain linked with three disulfide bridges. Analogous to proinsulin, the peptide chain is divided into the N-terminal B-chain, the connecting C-peptide, the A-chain, and the C-terminal D-chain. The propeptides also contain the C-terminal E-peptide and the prepropeptides, the N-terminal prepeptide. The structures of IGF-I and IGF-II in various species are described in [1760–1764].

Release. The release of SM is stimulated by growth hormone, prolactin (in the liver [1765]), insulin, injuries [1766], exercise [1767], increased vascular strain [1768], gonadoliberin, follitropin, lutropin, h-choriogonadotropin [1769], estradiol [1770], [1771], ACTH, androgens, parathyroid hormone [1759] and by cAMP (in osteoblast cultures [1772]). Prior to birth, neither IGF-I nor IGF-II appears to be dependent on GH [1773]. Here, placental lactogen, but not GH, insulin, or other growth factors stimulates the synthesis of IGF-II.

SM secretion is reduced by insufficient food intake [1774], low-protein diet [1773], chronic alcohol consumption [1775], and suppression treatment with GnRH agonists [1776].

Serum SM levels fall in patients with hypophyseal insufficiency, diabetes mellitus, and after the age of sixty. Women taking contraceptives and patients with hyperthyroidism have elevated SM levels [1777]. In Cushing's syndrome, the SM level in the spinal fluid is raised [1777].

Receptors [1759]. IGF-I and IGF-II bind with high affinity to specific receptors [1778].

Like the insulin receptor, the IGF-I receptor has two α- and two β-subunits. The α-subunit binds the hormones in the cysteine-rich domain between His^{223} and Met^{274} [1779].

The β-subunit contains tyrosine kinase which phosphorylates intracellular proteins after ligand binding.

The IGF-II receptor is very similar to the mannose 6-phosphate receptor and also binds mannose 6-phosphate [1780]. It consists of a single polypeptide chain, has no kinase activity, does not bind insulin, and binds only poorly to IGF-I.

An IGF binding component found in the serum is the extracellular part of the IGF-II receptor [1781]. In addition, four structurally related IGF binding proteins (IGFBP 1, 2, 3, and 4) have been characterized [1782], [1783]. IGFBP-1 binds IGF-I and IGF-II with similar affinity. IGFBP-2 preferentially binds IGF-II. The majority of serum IGF is bound in a complex consisting of IGF-I or IGF-II (γ-subunit), IGFBP-3 (β-subunit), and an α-subunit that does not bind IGF.

The serum IGF bioactivity is inhibited by h-IGFBP-1 and h-IGFBP-2 [1784], [1785]. In contrast, IGFBP-3 has a stimulating effect on the activity of IGF-I [1759].

Biological Effects. IGF-I stimulates intracellular synthesis of 1,2-diacylglycerol and induces the degradation of phosphatidyl choline, a process which involves protein kinase [1786].

In hypophysectomized animals, administration of IGF-I, like GH, increases body weight and bone growth. GH stimulates the proliferation of osteoblasts via the local synthesis of IGF-I [1787]. The GH-dependent increase in calcitriol after a low-phosphate diet is mediated by IGF-I [1788].

IGF-I and IGF-II bind 100 – 200 times less tightly to insulin receptors than insulin but, like insulin, they increase glucose uptake and oxidation in adipose and muscle tissue. In humans, they have about 6 % of the hypoglycemic activity of insulin [1789] and inhibit endogenous insulin release [1790].

The basal and GHRH-stimulated release of GH is inhibited by a negative feedback mechanism by IGF-I. It acts directly on the pituitary, where it inhibits transcription of the GH gene [1791]. It also releases somatostatin in the hypothalamus [1773]. IGF-II does not have a negative feedback effect on the basal or GHRH-stimulated GH [1751]. IGF-I inhibits the secretion of PRL in the pituitary [1792], but stimulates the release of PRL in cells of the human decidua [1793].

In Leydig cells, IGF-I stimulates the growth of h-CG receptors [1794] as well as the basal and h-CG-stimulated secretion of testosterone [1795]. IGF-I stimulates progesterone synthesis and pregnenolone in ovarian cells [1796].

IGF-I and IGF-II also act as autocrine growth factors in muscle development [1797] and cause proliferation of tumor cells [1798]. They increase the motility of melanoma cells and, therefore, could also increase the metastatic potential [1799].

IGF-I has a modulating effect on the immune system: it stimulates the growth of the thymus [1800], but inhibits the IL-2-induced proliferation and antibody formation of splenocytes in vitro [1801].

Structure – Activity Relationships. The A-chain domains of IGF-I are of importance for receptor binding and the growth effect [1802], the B-chain sequences play a role in the interaction with IGFBPs [1803].

Tyr^{24}, Tyr^{31}, and Tyr^{60} are important for binding of IGF-I to the IGF-I receptor and Tyr^{60} for binding to the IGF-II receptor [1804]. IGF-I-(6 – 70) has only one hundredth of the activity of IGF-I (in stimulating protein synthesis in rat myoblasts), IGF-I-(4 – 70) and IGF-I-(5 – 70) are ten times and twice as active as IGF-I, respectively [1805].

Uses. IGF-I reduces the blood sugar and insulin levels in Mendenhall's syndrome with severe insulin resistance [1806] and in Laron-type dwarfism with a defective GH receptor [1807].

7.6.4.2. Insulin

Occurrence. Insulin [*9004-10-8*], M_r 5808 (h-insulin), was first used in 1921 as a therapeutic agent in diabetes mellitus. It is synthesized in the β-cells of the pancreatic islets of Langerhans. However, immunoreactive insulin has also been found in other tissues (e.g., brain, testes, and liver).

Insulin consists of two peptide chains which are joined together by two disulfide bridges. It is formed from single-chain proinsulin by enzymatic removal of the C-peptide (connecting

peptide) [1808] between the C-terminal A-chain and the N-terminal B-chain. In most species the A-chain (acid chain) contains 21 amino acids, while the B-chain (basic chain) contains 30 amino acids. The C-chain with its N- and C-terminal basic amino acids contains 35 amino acids. In vivo, the C-chain is found without the N- and C-terminal basic amino acids. In physiological concentrations, insulin occurs in human plasma as a monomer which does not bind to serum proteins [1809]. Insulin sequences of various species are given in [1810].

Insulin-like molecules are also observed in the silkworm *Bombys mori* [1811], in the mollusk *Lymnaea stagnalis* [1812], and in the brain of the insect *Locusta migratoria* [1813].

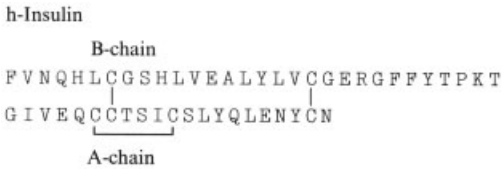

Release [1814]. Insulin is released by an increased plasma level of glucose or amino acids, cAMP, and Ca^{2+} (Mg^{2+}). The plasma insulin level increases after ingestion of food. Incretins are postulated to stimulate the secretion of insulin from the pancreas in a glucose-dependent process. Incretin release from the intestine is increased by oral glucose and inhibited by insulin. The criteria for the postulated incretin are best fulfilled by the glucagon-like peptide 1-(7 – 36)-amide [1815], the glucose-dependent insulinotropic peptide, and oxyntomodulin [1816]. Glucose-dependent secretion of insulin is also stimulated by the gastrin releasing peptide, neurotensin, substance P, gastrin, pentagastrin, CCK-8 (sulfated or unsulfated), the C-terminal tetrapeptide of CCK, cholinergic substances, α_2-adrenoreceptor antagonists [1817], TGF-β_1, TGF-β_2 [1818], oxytocin [1819], ACTH, glucagon, secretin [1820], calcitriol [1821], estradiol [1822], thyroid hormones [1823], prostaglandin D_2, lipoxygenase products [1824], and phospholipase A_2 products [1825].

Low doses of endogenous opioids and exorphins stimulate insulin secretion while high doses have a more inhibitory effect.

The glucose-stimulated release of insulin is inhibited by adrenaline [1826], prostaglandin E_2 [1827], somatostatin, pancreastatin, pancreatic peptide, galanine [1828], calcitonin gene related peptide [1829], amylin, calcitonin, exogenous insulin [1830], growth hormone, interferon [1831], interleukin 1β, and interleukin 6 [1832].

The cell-specific destruction of the pancreatic β-cells, which precedes insulin-dependent diabetes mellitus occurs by an autoimmune mechanism. Autoantibodies have been found in more than 80 % of newly diagnosed patients. The target of the autoantibodies is a β-cell autoantigen which has been identified as glutamic acid decarboxylase [1833] or as heat shock protein 65 [1834].

Receptors [1835]. The insulin receptor consists of two α- and two β- subunits joined together by disulfide bridges to give the configuration: (β-chain-S-S-α-chain)-S-S-(α-chain-S-S-β-chain). The cysteine-rich domain between Asn^{230} and Ile^{285} of the α-chain is important for insulin binding [1836]. The β-subunit is a transmembrane protein and contains a tyrosine-specific protein kinase. Insulin stimulates phosphorylation of the β-subunit. Autophosphorylation of the insulin receptor is restricted by triiodothyronine [1837].

The receptor is synthesized in the form of a prepropeptide.

Diabetes, resistance to insulin, can result from receptor defects (e.g., a deletion in the tyrosine kinase domain [1838] or the subtitution of valine for Gly^{996} [1839]). Trp^{1200} is also important for the normal function of insulin receptor kinase [1840]. In Hodgkin's disease, antibodies against the insulin receptors can cause hypoglycemia because of their insulin-like activity [1841].

The formation of insulin receptors is stimulated by glucose [1842] and down regulated by insulin and proinsulin [1843]. In cancerous breast tissue, the concentration of insulin receptors is more than six times that in normal tissue and correlates with the size of the tumor and with the estrogen receptor concentration [1844].

The following positions in insulin appear to be important for receptor binding: A-chain: Gly^1, Gln^5, Tyr^{19}, and Asn^{21}; B-chain: Val^{12}, Tyr^{16}, Arg^{22}, Gly^{23}, Phe^{24}, Phe^{25}, and Tyr^{26}.

Estrogen, progesterone [1845], h-GH, h-GH-(7 – 13) [1846], and relaxin [1847] increase the binding of insulin to the receptor; prolactin [1847], heparin [1848], or contraceptive therapy in women [1849] reduces binding.

Biological Effects. Insulin lowers blood glucose by promoting glucose uptake in skeletal muscle, heart, and adipose tissue, by activating glycogen synthesis [1850], and by inhibiting glycogenolysis in the liver [1851]. Glucose transport proteins are involved in the uptake of glucose [1852]. Taurine [1853], the basic peptides Arg-Lys and Arg-Arg [1854], and somatostatin potentiate the insulin effect on the blood glucose level.

Apart from stimulating glucose and amino acid transport, insulin also increases the secretion of somatomedins, the affinity of the IGF-II receptors [1855], the proliferation and differentiation of cells (healing of wounds), ion transport, renal Na^+/K^+ ATPase [1856], lipogenesis, potentiates the FSH-stimulated synthesis of estrogen and progestin [1857] and the h-CG-stimulated synthesis of testosterone [1858], increases food intake [1859], stimulates the secretion of gastric acid [1860] and bile flow, and is important for postprandial cholecystokinin- or secretin-stimulated exocrine pancreatic secretion [1861].

Hypoglycemia induced by insulin is a stress reaction which leads to the release of ACTH and cortisol [1862]. Insulin also increases the K^+-stimulated production of aldosterone [1863]. In type I diabetics, insulin lowers the level of cortisol raised by fasting [1864].

Insulin inhibits the synthesis and release of somatostatins and enteroglucagon [1865], reduces the basal and PTH-stimulated excretion of phosphate in the urine [1866] as well as renal Ca^{2+} reabsorption, stimulates the synthesis of collagen [1867], and potentiates the PTH-stimulated synthesis of calcitriol in the kidney [1868].

Insulin inhibits the noradrenaline- or AT II-stimulated contraction of arteries and veins (hypotensive effect [1869]), increases the heart rate [1870], and is positively inotropic [1871].

The inability of a given concentration of insulin to produce an expected biological effect is called insulin resistance. This resistance can be caused by circulating antagonists of insulin activity (antibodies, prostaglandins, and glucocorticoids [1872]), down regulation of the receptors, or receptor defects.

Structure – Activity Relationships. The biological activity of insulin is expressed in international units (IU). The activity is measured in the mouse cramp test or by determining the decrease in the blood sugar level in the rabbit. Crystalline insulin contains 26 – 28 IU/mg.

proinsulin is only about one-third as active as insulin in lowering the blood sugar level, but its action is significantly prolonged [1873]. The antilipolytic activity of h-proinsulin is higher than that of h-insulin [1874].

hydrophobic interactions are responsible for the formation of insulin dimers. The hexamer of insulin is formed from three dimers with two or four zinc atoms [1875]. Monomeric insulin is assumed to be the biologically active form, i.e., binds to the receptor. The regions around A1, A2, A19, A21, and b22 – B25 are important for biological activity.

fully active insulin is obtained when the B-chain is shortened at the N-terminus by two amino acids and at the C-terminus by three amino acids.

[aspb^{10}]h-insulin [1876] and de-(b^{26}–B^{30})-[AspB^{10}, tyr-nh$_2$B^{25}]insulin [1877] are more active than h-insulin. GlyB^{23} can be replaced by D-Ala without loss of activity, but [alab^{23}]insulin has a greatly reduced biological activity [1878]. Both [D-PheB^{24}]b-insulin and [D-AlaB^{24}]b-insulin still possess activity [1879]. Substitution of Glu for TyrB^{26} or thrb^{27} or of Asp for ProB^{28} leads to monomeric analogues having a slightly higher activity [1880]. ThrB^{29} can be replaced by many different amino acids and still retain activity [1881].

the n-terminal Gly of the A-chain can only be replaced by D-amino acids without loss of activity. [D-LeuA^1]insulin possesses full insulin activity and [D-TrpA^1]insulin is even slightly more potent than native insulin. The hydroxyl group of tyra^{19} forms a hydrogen bond with the carboxyl group of glya^1 and stabilizes the molecule. Modifications of asna^{21} lead to less active compounds. The cysteines are exremely important for the tertiary structure.

less soluble analogues with prolonged activity are obtained by introducing Arg into the molecule [1882–1884], amidating the Glu-γ-

carboxyl groups and the C-terminus of the B-chain [1885], and by the palmitoylation of the a-chain [1886]. Long-acting analogues include [glya^{21},ArgB^{27},ThrB^{30}-amide]h-insulin (NovoSol-Basal) [1882]. [aspb^{9},GluB^{27}]h-insulin [1887] and [aspb^{10}]h-insulin [1887] are monomeric derivatives that act especially quickly.

Uses [1888], [1889]. Many insulin preparations are available for the parenteral treatment of diabetes mellitus. Bovine, porcine, and human insulin are generally used. commercially available h-insulin preparations are either enzymatically modified p-insulin [H-Insulin, Depot-H-Insulin, Komb-H-Insulin (Hoechst); Insulin Monotard hm, Insulin Actrapid hm (novo); Insulin velasulin Human, Insulin mixtard Human (Nordisk)] or chain recombinant insulin produced by genetic engineering in bacteria (Huminsulin, Eli Lilly). In enzymatically modified products the C-terminal alanine of p-insulin is substituted by a threonine ester with the aid of trypsin. the ester function is then hydrolyzed to give h-insulin. In chain recombinant h-insulin, the A- and B-chains are synthesized separately and subsequently linked by oxidation, see also → Genetic Engineering.

de-pheb^{1}-insulin derivatives (Insulin Defalan, hoechst) have full insulin activity, but are more soluble and less antigenic than insulin. In patients with insulin allergy, de-pheb^{1}-p-insulin has been used successfully because it has a low antigenicity [1890]. However, de-Phe-insulins are no longer on the market.

insulin pumps have been developed for better insulin adjustment, pulsatile infusion appears to be more efficient than continuous infusion [1891].

treatment with insulin is also recommended in total parenteral feeding [1892], in cancer-induced anorexia, and loss of weight [1893]. Topically applied insulin promotes the healing of wounds [1894].

7.6.4.3. Relaxin [1895]

Occurrence. Relaxin (RLX) [*9002-69-1*], M_r 6333 (h-RLX 2), is found primarily in the serum and tissues of pregnant mammals. It occurs in especially high concentrations in the corpus luteum of pregnant animals, and is isolated from this source. In addition, RLX is found in the endometrium during pregnancy, breast [1896], testes, prostate, and seminal fluid.

RLX, like insulin, consists of an A- (24 amino acids in humans) and a B-chain (32 amino acids in humans) linked by sulfide bridges and is synthesized as a single-chain prohormone. h-RLX occurs in two forms, h-RLX-2 is formed in the ovaries during pregnancy [1897]. p-RLX occurs in three forms. RLX sequences from various species are described in [1897].

h-Relaxin-2

```
*QLYSALANKCCHVGCTKRSLARFC
  SWMEEVIKLCGRELVRAQIAICGMSTWSKRSL
```

Release. The plasma level of RLX increases greatly at the end of pregnancy and falls rapidly after birth. In hypophysectomized animals, endogenous RLX can be activated by estrogens and progesterone. the release of RLX is increased by h-CG [1898], oxytocin, and prostaglandin F_2. Indomethacin prevents the increase in the RLX level and delays birth.

In trophoblastic disease, which is accompanied by an increase in the plasma h-CG level, RLX is detectable in the plasma [1899].

Biological Effects. Purified RLX (NIH standard, RXN-P1) has about 3000 U/mg. One unit is the threshold dose which causes a softening of the interpubic disc in two-thirds of the guinea pigs tested.

In advanced pregnancy, RLX reduces the contraction of the uterus. This effect is supported by progesterone and is important for the maintenance of pregnancy. RLX has a growth-promoting and relaxing effect on the uterus and cervix and a growth-promoting and lactation-inhibiting effect on the mammary gland. RLX alone inhibits progesterone synthesis, but increases estrogen synthesis [1900]. RLX also appears to prepare the myometrium for parturition because without it oxytocin cannot induce birth. RLX inhibits prostaglandin synthesis during pregnancy, but stimulates it during birth [1901].

The relaxing effect of RLX on the interpubic disc (pelvic symphysis) occurs because the compact cartilaginous tissue between the two pubic bones is converted to a more flexible structure. the water content of the tissue is increased and the secretion of proteolytic enzymes is stimulated, loosening the network of collagen and glycosaminoglycan [1902]. Estrogen and hypophyseal hormones (e.g., growth hormone) are also involved in this deaggregation of collagen. Intravaginal RLX (2 mg/patient) causes cervical maturation in humans and induces birth.

RLX increases sperm motility and exerts a hypotensive effect [1903]. It inhibits the pressor activity of angiotensin II, vasopressin, and noradrenaline [1904].

Structure – Activity Relationships. The shortening of the A-chain of p-RLX by two amino acids at the N-terminus does not affect biological activity, whereas deletion of five or six amino acids halves the biological activity [1905].

Oxidation of $TrpB^{27}$ in p-RLX with *N*-bromosuccinimide has no influence on biological activity. However, p-RLX becomes inactive when $TrpB^{18}$ is oxidized.

7.6.5. Somatostatin [1906]

Occurrence. Somatostatin [*38916-34-6*], also called somatotropin release inhibiting hormone or factor (SRIH or SRIF), M_r 1638, is formed as prepro-SRIH. The main product of gene expression in the stomach and pancreas of mammals is pro-SRIH-(1 – 64), which is cleaved at the C-terminus to form SRIH-28 with 28 amino acids and SRIH-14 with 14 amino acids [1907].

SRIH-14 and SRIH-like substances have been found in the hypothalamus, nervous system, pancreatic islet D-cells, gastrointestinal mucosa, thymus, extrahepatic bile ducts, plasma [1908], in the C-cells of the thyroid gland, ovaries, and in plants (e.g., spinach [1909]). Somatostatin sequences of various species are described in [1910].

Release. SRIH is released in the antrum (lower part of the stomach) in response to fasting [1911], acid in the antrum, intraduodenal infusion of fat and HCl [1908], stress [1912], cholinergic and β-adrenergic mechanisms, dopamine D_2 agonists [1913], estrogen and androgens [1913], and numerous peptide hormones including growth hormone releasing hormone and somatomedins [1908], [1914–1917].

The secretion of SRIH is reduced shortly after meals [1911], by electrical stimulation of the vagus, sham feeding, α-adrenergic stimuli, serotonin [1918], and various peptides (e.g., endogenous opioids, galanine, pancreatic peptide, and insulin).

The plasma level of SRIH is lower at night [1919]. The plasma SRIH level in children below the age of two is raised. Above two years its level is normal and increases continuously with age [1920].

The plasma SRIH level is increased in patients with chronic renal failure, cirrhosis [1921], and primary hypothyroidism [1922]. The level of SRIH is raised in the cerebrospinal fluid of patients with brain tumors, metabolic encephalopathy, spinal cord compression, and intracranial hypertension [1923].

The plasma level of SRIH is reduced in patients with pernicious anemia, achlorhydria [1920], cluster headache [1924], psoriasis [1925], obesity [1926], and in women who are taking estrogen-containing contraceptives [1927]. The level of SRIH in the cerebrospinal fluid is reduced in patients with Alzheimer's disease [1928], unipolar depression, and active multiple sclerosis [1912]. The level of SRIH is reduced in the cortex and hippocampus of patients suffering from Parkinson's disease [1929].

Receptors. Several SRIH types of receptors are found in the brain.

SRIH receptors are detected in most GH- and TSH-producing pituitary adenomas, carcinoids, islet cell carcinomas, EGF-receptor-negative glial tumors, and small cell lung carcinomas [1930]. The number of SRIH receptors is reduced in the cerebral cortex of patients with Alzheimer's disease [1931].

Biological Effects. SRIH-14 [1932] stimulates phosphodiesterase, inhibits cAMP synthesis, blocks calcium influx, and the turnover of phosphoinositides in the cytosol [1933]. It also inhibits the release and action of hormones (e.g.,

growth hormone (GH), prolactin, thyrotropin, stress-induced corticotropin [1934]), calcitonin, insulin, glucagon, GIP, VIP, secretin, pancreatic polypeptide, gastrin releasing peptide [1908], gastrin, cholecystokinin, and motilin. GH is the most sensitive to SRIH inhibition, followed by thyrotropin [1935].

SRIH inhibits exocrine secretion of gastric acid, pepsin (anti-ulcer effect), pancreatic enzymes, pancreatic bicarbonate, colonic fluid (reduction of secretory diarrhoea [1936]), bile flow (patients with somatostatinoma usually have gall-stones [1937]). It also potentiates human platelet aggregation stimulated by collagen and arachidonic acid [1938] (control of upper gastrointestinal hemorrhage).

SRIH enhances gastric emptying [1939] and intestinal transit [1940]. It may be useful in the management of the dumping syndrome [1941]. SRIH inhibits cell proliferation in the exocrine pancreas [1942], rabbit jejunum [1943], liver after hepatectomy [1944], and tumor growth. The antitumor effect of SRIH is mediated by the stimulation of a tyrosine phosphatase [1945].

Immunomodulatory actions include increase of leukocyte migration inhibitory factor [1946], enhancement of human lymphocyte natural killer activity [1947], and mouse spleen lymphocyte proliferation [1948].

Structure – Activity Relationships. The substitution of D-Trp for Trp^8 produces an analogue that has 8 times the activity of SRIH in inhibiting GH release.

Potent, long-acting SRIH analogues of the urotensin II type are of tremendous interest [1949–1952]. For instance, octreotide [*49474-41-2*], [*79517-01-4*] is 70 times more effective than SRIH-14 in inhibiting GH. It has a long duration of action when given intramuscularly [1949].

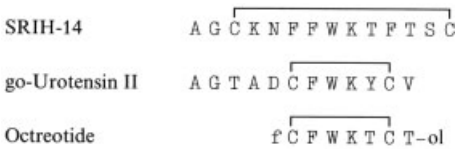

Cyclic peptides without a disulfide structure are also potent, long-acting compounds. Examples are cyclo-(*N*-Me-Ala-Tyr-D-Trp-Lys-Val-Phe) (seglitide) and cyclo-(Pro-Phe-D-Trp-Lys-Thr-Phe).

Cyclic retropeptides of the latter, e.g., cyclo-(Phe-Thr-Lys-Trp-Phe-D-Pro), inhibit the uptake of cholate and phallotoxin in liver cells [1953] and protect against taurocholate- and ceruletide-induced pancreatitis [1954], and against ethanol-induced stomach lesions [1955], without possessing true SRIH activity.

Analogues of octreotide which usually contain penicillamine instead of Cys^7 are poor ligands for SRIH receptors, but exhibit stronger binding to opioid receptors [1956]. They act as selective antagonists for μ-opioid receptors, i.e., they inhibit morphine induced analgesia [1957].

Uses. SRIH-14 infusions (Stilamin, Serono; Aminopon, UCB; Somatostatin Ferring, Ferring; Somatostatin Labaz, Sanofi) have proved useful for the treatment of bleeding peptic ulcers and gastrointestinal lesions, for preventive treatment against stress-induced ulcers, and in the healing of fistulae of the small intestine and gallbladder.

Octreotide (Sandostatin, Sandoz) [1958] is used in the treatment of GH- and thyrotropinsecreting pituitary tumors, carcinoid tumors, glucagonomas, insulinomas, and gastrinomas. The use of octreotide in acute esophageal variceal bleeding, pancreatic pseudocysts, external gastrointestinal and pancreatic fistulae, short bowel syndrome, dumping syndrome, and refractory hypersecretory diarrhoea related to the acquired immunodeficiency syndrome has provided encouraging results. Preliminary reports indicate the efficiency of octreotide in the treatment of psoriasis, autonomic neuropathy [1959] and of carcinoid flush [1960], and its ability to reduce growth in tall adolescents. Other potential applications include the control of pain [1961], rheumatoid arthritis, and diabetic microangiopathy as well as the prevention of postoperative pancreatic complications (fistulae, pancreatitis, and abscess).

7.6.6. Pancreastatin

Occurrence. Pancreastatin (PST) [*117148-67-1*] from the pancreas of the pig contains 49 amino acids [1962]. Sequences from

other species are given in [1963]. h-PST contains 52 amino acids, M_r 5509.

The precursor of pancreastatin is chromogranin A, an acidic glycoprotein which is widely distributed in the neuroendocrine system and released in large amounts from the chromaffin cells of the adrenal medulla [1963]. Pancreastatin contains the penta-Glu sequence of gastrin (GT) and has a C-terminus that is similar to that of gonadoliberin (GnRH).

In patients with non-insulin-dependent diabetes mellitus, the plasma pancreastatin level increases after the digestion of glucose, which is not the case in healthy persons [1964].

Biological Effects. Pancreastatin inhibits the first phase of glucose- or arginine-induced release of insulin and the arginine-induced release of somatostatin, but potentiates the arginine-induced secretion of glucagon [1965] from the rat pancreas. These changes are reminiscent of non-insulin-dependent diabetes.

In the rat, pancreastatin lowers the plasma insulin level and increases the plasma level of glucose in response to intragastric glucose [1966] and raises the plasma glucagon level in response to intravenous arginine [1967]. Like glucagon, it stimulates hepatic glycogenolysis [1968].

Pancreastatin also inhibits the CCK-stimulated secretion of protein and fluid from the pancreas [1967] and the CCK-stimulated growth of human pancreatic adenocarcinoma [1969].

Structure – Activity Relationships. The inhibitory effect on the glucose-stimulated release of insulin is restricted to the C-terminal region of the molecule [1962]. The shortest, active C-terminal fragment is p-PST (35 – 49) [1970]. The C-terminal amide function is required for biological activity [1970].

8. Neurotensin [1971]

Occurrence. Neurotensin (NT) [*39379-15-2*], M_r 1673 (h-NT), contains 13 amino acids and was first isolated from bovine hypothalamic extracts. Immunoreactive NT was later found in the gastrointestinal tract. The structurally and functionally related hexapeptide p-neuromedin N was isolated from the porcine spinal cord [1972]. NT and neuromedin N are coded on the same gene. Preprohormones from the rat, cow, and dog have also been characterized [1973].

h-Neurotensin	*Q L Y E N K P R R P Y I L
p-Neuromedin N	K I P Y I L

In humans, immunoreactive NT has been isolated primarily from the hypothalamus, median eminence, substantia nigra, and brain stem structures. It occurs with dopamine in dopaminergic neurons [1974]. NT and its main metabolite NT-(1 – 8) are also found in the adrenal medulla [1975].

The level of immunoreactive NT is raised by about 250 % in hot flushes experienced during menopause, but this material is not identical to NT or one of its known metabolites.

Release. NT is released from hypothalamic preparations by depolarizing K^+ and dopaminergic mechanisms. Methamphetamine, dopamine [1976], the dopamine antagonist haloperidol [1977], and neuroleptic agents [1978] release NT in the brain and increase the number of NT-containing neurons.

NT is released from the N-cells of the small intestine (particularly the ileum) into the blood in response to meals rich in fat and oleic acid, bombesin infusions, adrenaline [1979], hydrocortisone [1980], nerve growth factor, and insulin [1980]. Secretion of NT from the intestine is inhibited by somatostatin (SRIH) [1981] and dexamethasone [1982].

The level of NT may be elevated in tumor patients [1983][1984][1985]. The spinal fluid of children suffering from febrile aseptic meningitis has a low content of immunoreactive NT [1986].

Biological Effects. NT and C-terminal fragments such as NT-(8 – 13) raise the levels of cytosolic calcium [1987] and cGMP [1988] and stimulate the formation of inositol phosphates [1989].

NT has neuroleptic properties when applied centrally. It reduces dopamine synthesis within the nerve ends, increases the turnover of dopamine [1990], and inhibits the hyperactivity of dopamine receptor agonists [1991].

When centrally applied, NT stimulates the release of SRIH, luteotropin [1992], and corticosterone [1993]. It also enhances the effect of barbiturates [1985], exerts an antinociceptive effect [1994], reduces the secretion of thyrotropin, growth hormone, and prolactin, inhibits the carbachol, 2-deoxy-D-glucose and pentagastrin-stimulated gastric acid secretion [1995], exerts a cytoprotective effect on stress-induced gastric ulcers, and has a hypothermic effect [1996].

The most sensitive target organs for intravenously applied NT are the upper gastrointestinal tract and adipose tissue [1997]. NT reduces sphincter tone, inhibits gastric secretion, gastric emptying, and intestinal motility. It has a vasoconstrictive effect in adipose tissue.

In the stomach, NT inhibits the pentagastrin-stimulated release of pepsin and gastric acid secretion. It increases the exocrine $NaHCO_3$ secretion in the pancreas and the endocrine secretion of pancreatic peptide, insulin, and glucagon.

Relatively high doses of NT (> 100 ng kg^{-1} min^{-1}) raise the plasma glucose concentration and reduce blood pressure, probably through the release of histamine and other vasoactive substances from mast cells. In concentrations of $10^{-11} - 10^{-7}$ mol/L, NT enhances phagocytosis of macrophages and neutrophils and, consequently, could play a role in inflammatory processes.

Structure – Activity Relationships. [Gln^4]NT, g-NT, and [monoiodo-Tyr^3]NT [1998] appear to be identical to b-NT with respect to biological activity (hypotensive and hyperglycemic assays in the rat, radioimmunoassay). In vitro the C-terminal hexapeptide NT-(8 – 13) has almost the same activity as the entire molecule. The C-terminal pentapeptide and the carboxyl group are essential for activity. The C-terminal carboxyl group is not required for the hypotensive, vasodilating activity because cyclopeptides are potent hypotensive agents [1999].

The entire molecule appears to be required for the inhibition of the pentagastrin-stimulated secretion of gastric acid and pepsin and for the motility-inhibiting effect.

NT-(8 – 13) has the same central antinociceptive effect as NT. However, even shorter sequences such as NT-(9 – 11), NT-(8 – 10), or NT-(8 – 11)NH-C_2H_5 exhibit this activity [2000].

The cytoprotective activity of intracerebroventricularly applied NTs depends on the N-terminal sequence. NT-(8 – 13) and NT-(1 – 6) do not have cytoprotective activity, whereas NT-(1 – 8) and NT-(1 – 10) exhibit significant cytoprotection [2001].

The central hypothermic activity is especially pronounced in [D-Tyr^{11}]NT and [dopa^{11}]NT. [D-Tyr^{11}]NT also inhibits the hyperactivity induced by dopamine agonists when applied centrally [2002].

Depending on the dose, the NT-induced histamine-release, hypotensive effect [2003], synthesis of inositol phosphate [2004], and secretion of glucagon can be inhibited by [D-Trp^{11}]NT.

9. Motilin [2005]

Occurrence. p-Motilin (MT) [*9072-41-7*], M_r 2699.1 (h-MT), is a peptide containing 22 amino acids. The precursor of h-MT has been characterized via the cDNA [2006].

Motilin is found primarily in the jejunum but immunoreactive motilin has also been found in the mucosa of the esophagus, stomach, duodenum, ileum, and colon, in the pancreas, kidneys, adrenals, gallbladder, and brain.

Erythromycin A [2007] and the erythromycin derivative EM-523 [2008] are macrolide antibiotics which greatly increase the gastrointestinal motor activity and are motilin receptor agonists.

Release. The release of motilin is stimulated by oral fat, exogenous and endogenous bile acids, bombesin, alkalization of the duodenum (in the dog), acidification of the duodenum (in humans and pigs), stimulation of the vagus, exogenous motilin, and atropine [2009].

Release is inhibited by oral glucose, insulin, somatostatin, pancreatic peptide, VIP, and by sham feeding [2010].

In humans, elevated levels of MT have been observed in diarrhoea, kidney failure, after abdominal surgery and in alcoholics who do not have liver disease [2011].

Biological Effects. Motilin increases smooth muscle Ca^{2+} transport and stimulates intracellular cGMP synthesis. On a molar basis,

motilin has 50 times the contracting effect of acetylcholine in the duodenum.

The motilins appear to be species specific [2012]. Motilin induces contraction of the gastrointestinal tract and accelerates gastric emptying and intestinal transit. MT also contracts the lower esophageal sphincter and the gallbladder, and activates the Oddi sphincter [2013]. The cyclic pancreatic secretion of water, bicarbonate, and protein is increased by MT [2014].

In fasting rats, intraperitoneally applied MT (5 – 10 µg/kg) increases food uptake. This effect can be abolished by naloxone or pentagastrin. Thus, MT appears to be responsible for hunger during fasting [2015]. An infusion of low doses of motilin accelerates the absorption of glucose, increasing the plasma levels of glucose, GIP, and insulin. An infusion of high doses of motilin has the opposite effect.

Structure – Activity Relationships p-MT-(1 – 16) has 3 % of the p-MT activity (contraction of duodenum). Gly^8 can be replaced by D-Ala without loss of activity (i.e., contraction of the duodenum); Met^{13}, Gln^{14}, and Glu^{15} do not appear to be important for biological activity [2016], whereas substitution of Ser or Lys for Phe^1 produces compounds that are 250 and 1000 times less active, respectively [2017].

10. Pancreatic Spasmolytic Peptide

Occurrence. Pancreatic spasmolytic peptide (PSP) [*82987-77-7*], M_r 11 711, is a pyroglutamyl peptide containing 106 amino acids and seven disulfide bridges [2018]. Immunoreactive PSP is found in low concentrations in many organs, but substantial amounts only occur in the pancreas. There is about 50 % sequence homology between p-PSP-(14 – 49) and p-PSP-(63 – 98). There is great similarity between the sequences of p-PSP and the protein pS_2 which was isolated from human gastric juice [2019] and is induced by estrogen in the human breast cancer cell line MCF-7 [2020]. pS_2 is identical to the human epidermal growth factor-like immunoreactive factor [2021]. A protein from the skin gland of *Xenopus laevis*, known as the APEG protein [2022], has a sequence at its C-terminus that is very similar to p-PSP and h-pS_2. Another protein from *Xenopus laevis*, which was characterized via the cDNA structure, contains, after possible processing, two basic sequences (spasmolysin I and spasmolysin II) that are very similar to PSP [2023].

Biological Effects. p-PSP inhibits the activity of adenylate cyclase in the intestinal mucosal cell membranes of the rat [2024]. It inhibits gastrointestinal motility and the secretion of gastric acid in laboratory animals after parenteral and oral application. It also reduces the pressure and the wave frequency of the Oddi sphincter [2025]. PSP acts like a growth factor on MCF-7 cells [2019].

11. References

1. A. R. Green, *Lancet I* (1989) 705.
2. W. König in A. Kleemann et al. (eds.): *Arzneimittel, Fortschritte 1972 bis 1985,* VCH Verlagsgesellschaft, Weinheim 1987, pp. 572.
3. R. Geiger, *Coll. Soc. Fr. Etudes Fertil.* **26** (1988) 63.
4. E. Bürger, *Arzneim.-Forsch./Drug Res.* **38** (1988) 754.
5. IUPAC-IUB Commission on Biochemical Nomenclature, *J. Biol. Chem.* **247** (1972) 977.
6. IUPAC-IUB Commission on Biochemical Nomenclature, *Eur. J. Biochem.* **5** (1968) 151.
7. IUPAC-IUB Commission on Biochemical Nomenclature, *J. Biol. Chem.* **250** (1975) 3215.
8. IUPAC-IUB Commission on Biochemical Nomenclature, *Hoppe-Seyler's Z. Physiol. Chem.* **348** (1967) 262.
9. J. P. Adelman et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 179.
10. K. Nikolics et al., *Nature (London)* **316** (1985) 511.
11. S. S. Smith, S. R. Ojeda, *Endocrinology (Baltimore)* **115** (1984) 1973.
12. J. A. King, R. P. Millar, *Peptides (N.Y.)* **6** (1985) 689.
13. J. A. King, R. P. Millar, *Peptides (N.Y.)* **7** (1986) 827.
14. D. D. Rasmussen, *J. Endocrinol. Invest.* **9** (1986) 427.
15. B. A. Eipper et al., *Endocrinology (Baltimore)* **116** (1985) 2497.

16. P. A. Melrose, *Endocrinology (Baltimore)* **121** (1987) 200.
17. M. J. Kelly et al., *Neuroendocrinology* **49** (1989) 88.
18. C. R. Parker, Jr., J. C. Porter, *J. Clin. Endocrinol. Metab.* **58** (1984) 488.
19. S. Ohtsuka et al., *Acta Endocrinol. (Copenhagen)* **117** (1988) 399.
20. L. G. Allen et al., *Endocrinology (Baltimore)* **121** (1987) 1953.
21. K. Gerozissis et al., *Neuroendocrinology* **40** (1985) 272.
22. A. K. Dubey, T. M. Plant, *Biol. Reprod.* **33** (1985) 423.
23. T. J. Cicero et al., *Life Sci.* **37** (1985) 467.
24. F. Petraglia et al., *Endocrinology (Baltimore)* **120** (1987) 1083.
25. H. Rosen et al., *Endocrinology (Baltimore)* **122** (1988) 2873.
26. K. E. Nikolarakis et al., *Neuroendocrinology* **44** (1986) 314.
27. Z. Naor, *Endocr. Rev.* **11** (1990) 326.
28. A. Jagannadha Rao, N. R. Moudgal, *IRCS Med. Sci.* **12** (1984) 1105.
29. T. A. Bramley et al., *J. Endocrinol.* **113** (1987) 317.
30. N. Dekel et al., *Endocrinology (Baltimore)* **123** (1988) 1205.
31. M. Fekete et al., *J. Clin. Lab. Analysis* (1988) in press.
32. D. R. Pieper et al., *Endocrinology (Baltimore)* **115** (1984) 1190.
33. P. M. Conn et al., *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **43** (1984) 2351.
34. P. C. Wynn et al., *Endocrinology (Baltimore)* **119** (1986) 1852.
35. A. Starzec et al., *Endocrinology (Baltimore)* **119** (1986) 561.
36. W.-S. A. Wun, I. H. Thorneycroft, *Mol. Cell. Endocrinol.* **54** (1987) 165.
37. S. R. Swartz, G. P. Moberg, *Endocrinology (Baltimore)* **118** (1986) 2425.
38. W. R. Crowley et al., *Endocrinology (Baltimore)* **120** (1987) 941.
39. H. Dobson, *Acta Endocrinol. (Copenhagen)* **115** (1987) 63.
40. D. E. Suter, N. B. Schwartz, *Endocrinology (Baltimore)* **117** (1985) 849.
41. P. S. Li, *Life Sci.* **41** (1987) 2493.
42. A. P. N. Themmen et al., *J. Endocrinol.* **108** (1986) 431.
43. S. Y. Bin et al., *Contraception* **35** (1987) 79.
44. J. A. Foekens et al., *Biochem. Biophys. Res. Commun.* **140** (1986) 550.
45. B. D. Greenstein et al., *J. Endocrinol.* **112** (1987) 345.
46. M. J. Karten, J. E. Rivier, *Endocr. Rev.* **7** (1986) 44.
47. J. J. Nestor, Jr., B. H. Vickery, *Annu. Rep. Med. Chem.* **23** (1988) 211.
48. K. Folkers et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 1070.
49. J. J. Nestor, Jr., et al., *J. Med. Chem.* **27** (1984) 320.
50. T. L. Ho et al., *Int. J. Pept. Protein Res.* **24** (1984) 79.
51. W. König et al. in T. Shiba, S. Sakakibara (eds.): *Peptide Chemistry 1987,* Protein Research Found., Osaka 1988, p. 591.
52. J. Seprodi et al., *Biochem. Biophys. Res. Commun.* **144** (1987) 1214.
53. J. J. Nestor, Jr., et al. in F. Labrie et al. (eds.): *LH-RH and its Analogues,* Elsevier, Excerpta Medica, Amsterdam – New York – Oxford 1984, p. 24.
54. J. J. Nestor, Jr., et al., *J. Med. Chem.* **31** (1988) 65.
55. W. König et al. in G. Jung, E. Bayer (eds.): *Peptides 1988,* Walter de Gruyter, Berlin – New York 1989, p. 334.
56. J. E. Rivier et al., *J. Med. Chem.* **29** (1986) 1846.
57. The Salk Institute for Biological Studies, EP 0 102 492, 1985.
58. A. Ljungqvist et al., *Biochem. Biophys. Res. Commun.* **148** (1987) 849.
59. S. Bajusz et al., *Int. J. Pept. Protein Res.* **32** (1988) 425.
60. M. V. Nekola et al., *Int. Arch. Allergy Appl. Immunol.* **84** (1987) 3161.
61. K. Folkers et al., *Z. Naturforsch. B. Anorg. Chem. Org. Chem.* **42B** (1987) 101.
62. S. Bajusz et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 1637.
63. A. Phillips et al., *Life Sci.* **41** (1987) 2017.
64. A. Phillips et al., *Life Sci.* **43** (1988) 883.
65. A. V. Schally et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 851.
66. A. Bertelsen et al., *Eur. J. Pediatr.* **146** (1987) Suppl. 2, 40.
67. S. Zgliczynski et al., *Acta Endocrinol. (Copenhagen)* **105** (1984) 161.
68. K. E. Anderson et al., *N. Engl. J. Med.* **311** (1984) 643.
69. *Drugs Future* **9** (1984) 295.
70. *Drugs Future* **3** (1978) 645; **11** (1986) 812.
71. *Drugs Future* **8** (1983) 1038; **9** (1984) 930.
72. *Drugs Future* **10** (1985) 355; **12** (1987) 404.
73. *Drugs Future* **11** (1986) 623; **12** (1987) 720.

74. G. C. Doelle et al., *Horm. Metab. Res.* **18** (1986) 201.
75. *Drugs of the Future* **11** (1986) 745.
76. A. Lemay et al., *Fertil. Steril.* **47** (1987) 60.
77. M. W. van Maarschalkerweerd, M. Gons, *Eur. J. Pediatr.* **146** (1987) 272.
78. V. Perl et al., *Fertil. Steril.* **48** (1987) 383.
79. J. L. Andreyko et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 854.
80. K. E. Anderson et al., *Clin. Pharmacol. Ther. (St. Louis)* **41** (1987) 180.
81. J. L. Gabrilove et al., *N. Engl. J. Med.* **318** (1988) 580.
82. W. H. Matta et al., *Br. Med. J.* **294** (1987) 1523.
83. J. F. Wilber et al., *Clin. Res.* **34** (1986) 272A.
84. S. M. Cockle et al., *FEBS Lett.* **252** (1989) 113.
85. M. J. Kubek et al., *Life Sci.* **36** (1985) 315.
86. T. M. Engber et al., *Regul. Pept.* **12** (1985) 51.
87. B. M. Lewis et al., *J. Endocrinol.* **115** (1987) 419.
88. M. T. Lin et al., *Neuroendocrinology* **50** (1989) 177.
89. M. Mori, M. Yamada, *J. Endocrinol.* **114** (1987) 443.
90. T. Mitsuma et al., *Horm. Metab. Res.* **21** (1989) 301.
91. T. Mitsuma et al., *Peptides (N.Y.)* **8** (1987) 473.
92. S. Bhasin et al., *Endocrinology (Baltimore)* **114** (1984) 946.
93. P. Lamberton et al., *Endocrinology (Baltimore)* **117** (1985) 1834.
94. S. Aratan-Spire et al., *Acta Endocrinol. (Copenhagen)* **106** (1984) 209.
95. A. Dutour et al., *Peptides (N.Y.)* **10** (1989) 523.
96. T. Mitsuma et al., *Horm. Res.* **25** (1987) 223.
97. S. Wilk, *Life Sci.* **39** (1986) 1487.
98. L. D. Keith et al., *Neuroendocrinology* **43** (1986) 445.
99. M. Mori et al., *Neuroendocrinology* **48** (1988) 153.
100. H. N. Bhargava et al., *Arch. Int. Pharmacaodyn. Ther.* **297** (1989) 247.
101. H. N. Bhargava, *Ann. N. Y. Acad. Sci* **553** (1989) 526.
102. H. N. Bhargava, *Ann. N. Y. Acad. Sci* **553** (1989) 528.
103. A. H. Drummond, *Biochem. Biophys. Res. Commun.* **127** (1985) 63.
104. R. K. Rinehart et al., *J. Pharmacol. Exp. Ther.* **238** (1986) 178.
105. A. H. Drummond, P. J. Hughes, F. Ruiz Larrea, L. A. Joels, *Ann. N. Y. Acad. Sci.* **553** (1989) 197.
106. E. C. Grtiffiths, *Clin. Science* **73** (1987) 449.
107. A. Sanchez, E. Montoya, *Horm. Metab. Res.* **19** (1987) 604.
108. I. Zofkova, J. Bednar, *Exp. Clin. Endocrinol.* **83** (1984) 263.
109. H. Watanobe et al., *J. Endocrinol. Invest.* **8** (1985) 459.
110. W. Schäfgen et al., *Horm. Metab. Res.* **16** (1984) 615.
111. J.-R. Attali et al., *Endocrinology (Baltimore)* **116** (1985) 561.
112. L. O. D. Koskinen, *Ann. N. Y. Acad. Sci.* **553** (1989) 353.
113. L. O. D. Koskinen, *Ann. N. Y. Acad. Sci.* **553** (1989) 608.
114. M. Burnier et al., *Clin. Res.* **37** (1989) 249A.
115. S. Reichlin, *Acta Endocrinol. (Copenhagen) Suppl.* **276** (1986) 21.
116. T. Suzuki et al., *Gen. Pharmacol.* **20** (1989) 239.
117. T. Mitsuma et al., *Exp. Clin. Endocrinol.* **89** (1987) 55.
118. J. Mattila, R. D. Bunag, *J. Pharmocol. Exp. Ther.* **238** (1986) 232.
119. T. Garrik et al., *Life Sci.* **40** (1987) 649.
120. J.-P. Roussel et al., *Acta Endocrinol. (Copenhagen)* **114** (1987) 314.
121. *Drugs Future* **9** (1984) 807.
122. *Drugs Future* **8** (1983) 1007; **11** (1986) 1059; **12** (1987) 1153; **14** (1989) 1224.
123. H. Maeda et al., *Chem. Pharm. Bull.* **36** (1988) 190.
124. *Drugs Future* **14** (1989) 203; **15** (1990) 285.
125. *Drugs Future* **7** (1982) 21; **8** (1983) 56; **9** (1984) 62; **10** (1985) 75; **11** (1986) 6; **12** (1987) 70; **13** (1988) 73; **14** (1989) 78; **15** (1990) 92.
126. H. Maeda et al., *Int. J. Pept. Protein Res.* **33** (1989) 403.
127. *Drugs Future* **7** (1982) 167; **8** (1983) 282; **14** (1989) 295; **15** (1990) 319.
128. *Drugs Future* **13** (1988) 420; **14** (1989) 48.
129. M. Oka et al., *Arzneim. Forsch.* **39** (1989) 297.
130. R. Lanzara et al., *Ann. N. Y. Acad. Sci.* **553** (1989) 559.
131. J. G. Pierce, T. F. Parsons, *Ann. Rev. Biochem.* **50** (1981) 465.
132. D. N. Ward et al., *J. Protein Chem.* **1** (1982) 263.
133. F. S. Esch et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 6618.

134. C.-S. Liu et al., *J. Biochem.* **186** (1989) 105.
135. A. M. Naylor et al., *Neuroendocrinology* **49** (1989) 531.
136. S. S. Stojilkovic et al., *Science (Washington, D.C.)* **248** (1990) 1663.
137. G. Saade et al., *J. Endocrinol.* **114** (1987) 95.
138. G. Emons et al., *Acta Endocrinol. (Copenhagen)* **113** (1986) 219.
139. G. A. Schuiling et al., *Acta Endocrinol. (Copenhagen)* **110** (1985) 329.
140. L. N. Sidneva et al., *Probl. Endocrinol.* **33** (1987) 67.
141. I. Porsova et al., *Horm. Res.* **28** (1987) 288.
142. L. Kiesel et al., *Mol. Cell. Endocrinol.* **51** (1987) 1.
143. L. Wogensen, J. Warberg, *Acta Endocrinol. (Copenhagen)* **106** (1984) 30.
144. P. M. Conn et al., *Biochem. Biophys. Res. Commun.* **126** (1985) 532.
145. C. G. Brown, N. L. Poyser, *J. Endocrinol.* **103** (1984) 155.
146. M. Filicori et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 1136.
147. P. A. Steele, S. J. Judd, *Clin. Endocrinol.* **29** (1988) 1.
148. T. Nishi et al., *Acta Obstet. Gynecol. Scand.* **66** (1987) 309.
149. A. Tsatsoulis et al., *Clin. Endocrinol.* **32** (1990) 73.
150. F. R. A. Cagampang et al., *Horm. Metab. Res.* **22** (1990) 269.
151. E. M. McColl et al., *Clin. Endocrinol.* **31** (1989) 617.
152. F. Petraglia et al., *Life Sci.* **38** (1986) 2103.
153. S. E. Lenahan et al., *Neuroendocrinology* **44** (1986) 89.
154. M. S. Smith et al., *Neuroendocrinology* **50** (1989) 308.
155. J. H. Mendelson et al., *J. Pharmacol. Exp. Ther.* **237** (1986) 862.
156. I. R. Reid, *Ann. Intern. Med.* **106** (1987) 639.
157. K. Seki, I. Nagata, *Acta Endocrinol. (Copenhagen)* **122** (1990) 211.
158. C. R. Pohl et al., *Endocrinology (Baltimore)* **120** (1987) 849.
159. S. H. Sheika et al., *Life Sci.* **44** (1989) 1363.
160. S. Taleisnik, C. H. Sawyer, *Neuroendocrinology* **44** (1986) 265.
161. H. Loosfelt et al., *Science (Washington, D.C.)* **245** (1989) 525.
162. K. C. McFarland et al., *Science (Washington, D.C.)* **245** (1989) 494.
163. R. Farookhi, J. Desjardins, *Mol. Cell. Endocrinol.* **47** (1986) 13.
164. X.-C. Jia et al., *Endocrinology (Baltimore)* **117** (1985) 13.
165. P. E. Gottschall et al., *Biochem. Biophys. Res. Commun.* **149** (1987) 502.
166. J. S. Davis et al., *Biochem. J.* **238** (1986) 597.
167. V. V. Patwardhan, A. Lanthier, *J. Endocrinol.* **101** (1984) 305.
168. H. T. Keutmann et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 2038.
169. T. A. Santa Coloma et al., *Biochemistry* **29** (1990) 1194.
170. T. A. Santa Coloma, L. E. Reichert, Jr., *J. Biol. Chem.* **265** (1990) 5037.
171. M. C. Charlesworth et al., *J. Biol. Chem.* **262** (1987) 13409.
172. B. B. Saxena, P. Rathnam, *Biochemistry* **24** (1985) 813.
173. J.-M. Bidart et al., *Science (Washington, D.C.)* **248** (1990) 736.
174. L. Liu et al., *Endocrinology (Baltimore)* **124** (1989) 175.
175. A. Traub, Jr., *J. Med. Sci.* **155** (1986) Suppl. 30.
176. A. Lanzone et al., *Fertil. Steril.* **48** (1987) 1058.
177. W. G. Rossmanith et al., *Int. J. Fertil.* **32** (1987) 460.
178. J. A. Magner, *Endocrine Reviews* **11** (1990) 354.
179. J. A. Franklyn, M. C. Sheppard, *J. Endocrinol.* **117** (1988) 161.
180. K. Törnquist, C. Lamberg-Allardt, *Horm. Metab. Res.* **18** (1986) 69.
181. N. W. Kasting, J. B. Martin, *Neuroendocrinology* **39** (1984) 201.
182. M. D. Lumpkin et al., *Science (Washington, D.C.)* **235** (1987) 1070.
183. V. Rettori et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 4789.
184. S. W. J. Lamberts et al., *J. Endocrinol. Invest.* **7** (1984) 313.
185. F. Sanchez-Franco, L. Cacicedo, *Hormone Res.* **24** (1986) 55.
186. R. J. Comi et al., *N. Engl. J. Med.* **317** (1987) 12.
187. M. Parmentier et al., *Science (Washington, D.C.)* **246** (1989) 1620.
188. M. Misrahi et al., *Biochem. Biophys. Res. Commun.* **166** (1990) 394.
189. E. Bone et al., *Biochem. Biophys. Res. Commun.* **141** (1986) 1261.
190. P. P. Roger, J. E. Dumont, *Biochem. Biophys. Res. Commun.* **149** (1987) 707.
191. G. Zerek Melen et al., *Res. Exp. Med.* **187** (1987) 415.

192. A. Nilsson, R. Axelsson, *Curr. Ther. Res.* **46** (1989) 85.
193. J. C. Morris III et al., *Endocrinology (Baltimore)* **123** (1988) 456.
194. J. C. Morris et al., *J. Biol. Chem.* **265** (1990) 1881.
195. N. Ling et al., *Vitam. Horm. (N.Y.)* **44** (1988) 1.
196. S.-Y. Ying, *Endocr. Rev.* **9** (1988) 267.
197. R. I. McLachlan et al., *Biochem. Biophys. Res. Commun.* **140** (1986) 485.
198. P. E. Sawchenko et al., *Nature (London)* **334** (1988) 615, 2406.
199. T. Suzuki et al., *Mol. Cell. Endocrinol.* **54** (1987) 185.
200. C. G. Tsonis et al., *J. Endocrinol.* **112** (1987) R11.
201. F. Petraglia et al., *Science (Washington, D.C.)* **237** (1987) 187.
202. Z. Zhiwen et al., *Endocrinology (Baltimore)* **120** (1987) 1633.
203. C. Rivier, W. Vale, *Endocrinology (Baltimore)* **124** (1989) 195.
204. R. I. McLachlan et al., *J. Clin. Endocrinol. Metab.* **65** (1987) 954.
205. C. B. Sheckter et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 1221.
206. R. I. McLachlan et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 1305.
207. R. E. Lappöhn et al., *N. Engl. J. Med.* **321** (1989) 790.
208. C. P. Channing et al., *Proc. Soc. Exp. Biol. Med.* **178** (1985) 339.
209. S. A. Shanbhag et al., *J. Endocrinol.* **103** (1984) 389.
210. J. E. Mercer et al., *Mol. Cell. Endocrinol.* **53** (1987) 251.
211. Q. F. Wang et al., *Endocrinology (Baltimore)* **124** (1989) 363.
212. B. V. Bapat et al., *Int. J. Fertil.* **31** (1986) 71.
213. S.-Y. Ying et al., *Biochem. Biophys. Res. Commun.* **136** (1986) 969.
214. W. Vale et al., *Nature (London)* **321** (1986) 776.
215. N. Ling et al., *Biochem. Biophys. Res. Commun.* **138** (1986) 1129.
216. N. Ling et al., *Nature (London)* **321** (1986) 779.
217. T. Nakamura et al., *Science (Washington, D.C.)* **247** (1990) 836.
218. L. A. Hutchinson et al., *Biochem. Biophys. Res. Commun.* **146** (1987) 1405.
219. J. Massague, *Cell* **49** (1987) 437.
220. M. B. Sporn et al., *Science (Washington, D.C.)* **233** (1986) 532.
221. R. Derynck et al., *Nature (London)* **316** (1985) 701.
222. R. Pircher et al., *Biochem. Biophys. Res. Commun.* **136** (1986) 30.
223. L. M. Wakefield et al., *J. Biol. Chem.* **263** (1988) 7646.
224. T. A. McCaffrey et al., *J. Cell. Biol.* **109** (1989) 441.
225. J. Pfeilschifter, G. R. Mundy, *Proc. Natl. Acad. Sci. USA* **84** (1987) 2024.
226. R. O. C. Oreffo et al., *Biochem. Biophys. Res. Commun.* **158** (1989) 817.
227. M. Centrella et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 5889.
228. S. Y. Ying et al., *Biochem. Biophys. Res. Commun.* **135** (1986) 950.
229. E. Y. Adashi, C. E. Resnick, *Endocrinology (Baltimore)* **119** (1986) 1879.
230. M. Hotta, A. Baird, *Proc. Natl. Acad. Sci. USA* **83** (1986) 7795.
231. Y. Totsuka et al., *Biochem. Biophys. Res. Commun.* **158** (1989) 1060.
232. E. B. Leof et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 2453.
233. H. L. Moses et al., *Cancer Cells* **3** (1985) 65.
234. T. Nakamura et al., *Biochem. Biophys. Res. Commun.* **133** (1985) 1042.
235. A. B. Roberts et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 119.
236. J. Massague et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 8206.
237. E. Balza et al., *FEBS Lett.* **228** (1988) 42.
238. J. Liu, J. M. Davidson, *Biochem. Biophys. Res. Commun.* **154** (1988) 895.
239. A. Bassols, J. Massague, *J. Biol. Chem.* **263** (1988) 3039.
240. S. Chandrasekhar, A. K. Harvey, *Biochem. Biophys. Res. Commun.* **157** (1988) 1352.
241. J. R. Gamble, M. A. Vadas, *Science (Washington, D.C.)* **242** (1988) 97.
242. S. Tsunawaki et al., *Nature (London)* **334** (1988) 260.
243. H. J. Andrews et al., *Biochem. Biophys. Res. Commun.* **162** (1989) 144.
244. M. Centrella et al., *Endocrinology (Baltimore)* **119** (1986) 2306.
245. M. Noda et al., *J. Biol. Chem.* **263** (1988) 13916.
246. C. Chenu et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 5683.
247. R. Montesano, L. Orci, *Proc. Natl. Acad. Sci. USA* **85** (1988) 4894.
248. M. R. Urist, *Science (Washington, D.C.)* **150** (1965) 893.

249. R. W. Katz, A. H. Reddi, *Biochem. Biophys. Res. Commun.* **157** (1988) 1253.
250. H. Heumann, K. Schmidt, *Bild Wiss.* (1989) no. 5, 46.
251. J. M. Wozney et al., *Science (Washington, D.C.)* **242** (1988) 1528.
252. F. P. Luyten et al., *J. Biol. Chem.* **264** (1989) 13377.
253. E. özkaynak et al., *EMBO J.* **9** (1990) 2085.
254. T. K. Sampath et al., *Biochem. Biophys. Res. Commun.* **124** (1984) 829.
255. A. H. Reddi, *Isr. J. Med. Sci.* **21** (1985) 312.
256. N. Muthukumaran, A. H. Reddi, *Clin. Orthop. Relat. Res.* **200** (1985) 159.
257. E. A. Wang et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 2220.
258. J. J. Vandersteenhoven et al., *Calcif. Tissue Int.* **42** (1988) 39.
259. H. S. Talwar et al., *Kidney Int.* **29** (1986) 1038.
260. D. P. Desimone, A. H. Reddi, *Calcif. Tissue Int.* **35** (1983) 732.
261. R. Howes et al., *Calcif. Tissue Int.* **42** (1988) 34.
262. F. W. Thielemann et al., *Z. Orthop. Ihre Grenzgeb.* **122** (1984) 843.
263. R. L. Cate et al., *Cell* **45** (1986) 685.
264. J. P. Coughlin et al., *Mol. Cell. Endocrinol.* **49** (1987) 75.
265. R. B. Pepinsky et al., *J. Biol. Chem.* **263** (1988) 18961.
266. S.-Y. Ying, *Endocr. Rev.* **9** (1988) 267.
267. S. Shimasaki et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 4218.
268. T. Nakamura et al., *Science (Washington, D.C.)* **247** (1990) 836.
269. M. E. Joyce et al., *J. Cell. Biol.* **110** (1990) 2195.
270. H. M. Harms et al., *J. Clin. Endocrinol. Metab.* **69** (1989) 843.
271. J. F. Habener et al., *Physiol. Rev.* **64** (1984) 985.
272. T. J. Martin, L. J. Suva, *Clin. Endocrinol.* **31** (1989) 631.
273. J. J. Orloff et al., *Endocr. Rev.* **10** (1989) 476.
274. L. J. Suva et al., *Science (Washington, D.C.)* **237** (1987) 893.
275. E. af Ekenstam et al., *Clin. Endocrinol.* **32** (1990) 323.
276. J. Russell, L. M. Sherwood, *Mol. Endocrinol.* **3** (1989) 325.
277. L. J. Suava et al., *Science (Washington, D.C.)* **237** (1987) 893.
278. K. L. Insogna et al., *N. Engl. J. Med.* **313** (1985) 1126.
279. M. E. Markowitz et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 1068.
280. S. Ljunghall et al., *Exp. Clin. Endocrinol.* **84** (1984) 319.
281. M. S. Calvo et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 823.
282. L. A. Fitzpatrick et al., *Biochem. Biophys. Res. Commun.* **138** (1986) 960.
283. A. Goulding, E. Gold, *Horm. Metab. Res.* **20** (1988) 743.
284. M. S. Seshadri et al., *Endocrinology (Baltimore)* **117** (1985) 2417.
285. E. W. Seely et al., *Acta Endocrinol. (Copenhagen)* **121** (1989) 174.
286. S. Balabanova et al., *Horm. Res.* **24** (1986) 302.
287. S. Ljunghall et al., *Exp. Clin. Endocrinol.* **84** (1984) 313.
288. B. Duarte et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 584.
289. H.-M. Harms et al., *J. Clin. Endocrinol. Metab.* **69** (1989) 843.
290. J. Russell et al., *Endocrinology (Baltimore)* **119** (1986) 2864.
291. O. Fermet et al., *J. Endocrinol.* **113** (1987) 117.
292. G. A. Williams et al., *Metab. Clin. Exp.* **34** (1985) 612.
293. S. A. Rabbani et al., *J. Biol. Chem.* **259** (1984) 2949.
294. J. Lefebvre, *Horm. Res.* **32** (1989) 104.
295. L. E. Mallette, *Am. J. Med. Sci.* **293** (1987) 239.
296. S. Hirschel-Scholz et al., *Calcif. Tissue Int.* **40** (1987) 103.
297. E. F. Rittinghaus et al., *Acta Endocrinol. (Copenhagen)* **111** (1986) 62.
298. H. Yoshida et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 488.
299. T. Andersen et al., *Metab. Clin. Exp.* **35** (1986) 147.
300. S. Ljunghall et al., *Exp. Clin. Endocrinol.* **88** (1986) 95.
301. K. Ikeda et al., *J. Biol. Chem.* **264** (1989) 15743.
302. D. Wynick et al., *J. Endocrinol.* **123** (1989) Suppl., 36.
303. A. K. Hall, I. R. Dickson, *Acta Endocrinol. (Copenhagen)* **108** (1985) 217.
304. T. Okazaki et al., *Metab. Clin. Exp.* **33** (1984) 710.
305. A. van der Plas et al., *Biochem. Biophys. Res. Commun.* **129** (1985) 918.
306. J. A. Yee, *J. Bone Mineral. Res.* **3** (1988) 211.

307. C. W. G. M. Löwik et al., *Biochem. Biophys. Res. Commun.* **162** (1989) 1546.
308. P. M. J. McSheehy, T. J. Chambers, *Endocrinology (Baltimore)* **119** (1986) 1654.
309. S. F. Silverton et al., *Am. J. Physiol.* **253** (1987) E670.
310. G. E. Hall, A. D. Kenny, *J. Pharmacol. Exp. Ther.* **238** (1986) 778.
311. E. M. Spencer et al., *Acta Endocrinol. (Copenhagen)* **121** (1989) 435.
312. M. Centrella et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 5889.
313. K. Hove et al., *Endocrinology (Baltimore)* **114** (1984) 897.
314. J. Wortsman et al. . *J. Clin. Endocrin. Metab.* **62** (1986) 1305.
315. H. N. Hulter, J. C. Peterson, *Metab. Clin. Exp.* **33** (1984) 662.
316. L. L. S. Mok et al., *Endocr. Rev.* **10** (1989) 420.
317. P. K. T. Pang et al., *Can. J. Physiol. Pharmacol.* **64** (1986) 1543.
318. M. C. M. Yang et al., *J. Pharmacol. Exp. Ther.* **252** (1990) 840.
319. S. Anderson et al., *Hypertension (Dallas)* **5** (1983) Suppl. I, 59.
320. K. Iseki et al., *Am. J. Physiol.* **250** (1986) F924.
321. M. A. Gorozhankina et al., *Patol. Fiziol. Eksp. Ter.* **2** (1986) 71.
322. G. Clementi et al., *Eur. J. Pharmacol.* **166** (1989) 549.
323. N. Horiuchi et al., *Science (Washington D.C.)* **238** (1987) 1566.
324. D. D. Thompson et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 5673.
325. M. P. Caulfield et al., *Endocrinology (Baltimore)* **123** (1988) 2949.
326. A. L. Frelingerl, III., J. E. Zull, *J. Biol. Chem.* **259** (1984) 5507.
327. R. L. Shew, P. K. T. Pang, *Peptides (Fayetteville, N.Y.)* **5** (1984) 485.
328. S. H. Doppelt et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 7557.
329. J. W. Hamilton et al., *Mol. Cell. Endocrinol.* **44** (1986) 179.
330. M. E. Goldman et al., *Endocrinology (Baltimore)* **123** (1988) 2597.
331. J. Reeve et al., *Eur. J. Clin. Invest.* **17** (1987) 421.
332. E. F. Rittinghaus et al., *Acta Endocrinol. (Copenhagen)* **117**(1988) Suppl. 287, 167.
333. L. E. Mallette, *Ann. Intern. Med.* **109** (1988) 800.
334. D. Goltzman, G. S. Tannenbaum, *Brain. Res.* **416** (1987) 1.
335. E. Bucht et al., *Acta Physiol. Scand.* **126** (1986) 289.
336. E. Bucht et al., *Acta Endocrinol. (Copenhagen)* **113** (1986) 529.
337. J. M. Le Moullec et al., *FEBS Lett.* **167** (1984) 93.
338. J. Glowacki et al., *Endocrinology (Baltimore)* **116** (1985) 827.
339. F. Raue et al., *Endocrinology (Baltimore)* **115** (1984) 2362.
340. C. Greenberg et al., *Endocrinology (Baltimore)* **118** (1986) 2594.
341. H. Scheruebl et al., *Acta Endocrinol. (Copenhagen)* Suppl. 117 (1988) no. 287, 66.
342. O. Torring et al., *Horm. Metab. Res.* **17** (1985) 536.
343. B. Rasmusson, *Acta Endocrinol. (Copenhagen)* **106** (1984) 112.
344. F. Tagliaro et al., *J. Endocrinol. Invest.* **7** (1984) 331.
345. A. Lundin, K. öberg, *Acta Med. Scand.* **215** (1984) 281.
346. J. J. Body et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 700.
347. G. Luisetto et al., *J. Endocrinol. Invest.* **9** (1986) 239.
348. J. Zseli et al., *Calcif. Tissue Int.* **39** (1986) 297.
349. J.-Y. Reginster et al., *Gynecol. Endocrinol.* **2** (1988) 195.
350. C. Foresta et al., *Horm. Metab. Res.* **19** (1987) 275.
351. N. Takahashi et al., *Endocrinology (Baltimore)* **123** (1988) 1504.
352. H.-G. Schneider et al., *Mol. Cell. Endocrinol.* **58** (1988) 9.
353. C. M. Rotella et al., *Eur. J. Pharmacol.* **107** (1985) 347.
354. J. Barsony, S. J. Marx, *Endocrinology (Baltimore)* **122** (1988) 1218.
355. A. N. K. Yusufi et al., *Am. J. Physiol.* **252** (1987) F598.
356. F. Pansini et al., *J. Endocrinol.* **116** (1988) 155.
357. D. R. Roy, *Can. J. Physiol. Pharmacol.* **63** (1985) 89.
358. J. Puig et al., *Acta Pharmacol. Toxicol. Suppl.* **59** (1986) no. 5, part 2, 126.
359. M. Yamada et al., *Endocrinology (Baltimore)* **116** (1985) 693.
360. T. J. Chambers et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 1080.
361. A. Malgaroli et al., *J. Biol. Chem.* **264** (1989) 14342.

362. H. I. Khouja et al., *J. Endocrinol.* **119 Suppl.** (1988) 152.
363. J. R. Farley et al., *Calcif. Tissue Int.* **45** (1989) 214.
364. W. M. Burch, *Endocrinology (Baltimore)* **114** (1984) 1196.
365. G. P. Ceda et al., *Acta Endocrinol. (Copenhagen)* **120** (1989) 416.
366. C. Beglinger et al., *Gut* **29** (1988) 243.
367. G. H. Greeley, Jr., et al., *Regul. Pept.* **24** (1989) 259.
368. M. Yamaguchi, M. Toyoizumi, *Horm. Metab. Res.* **18** (1986) 378.
369. M. Yamaguchi, K. Momose, *Horm. Metab. Res.* **18** (1986) 22.
370. B. J. Looij, Jr., et al., *Ned. Tijdschr. Geneeskd.* **132** (1988) 1558.
371. I. Zofkova, J. Nedvidkova, *Exp. Clin. Endocrinol.* **92** (1989) 262.
372. T. S. Mahrous, A. M. Nakhla, *Acta Endocrinol. (Copenhagen)* **119** (1988) 525.
373. I. Zofkova, J. Bednar, *Exp. Clin. Endocrinol.* **92** (1988) 268.
374. Y. Goto et al., *Can. J. Physiol. Pharmacol.* (1986) July, Suppl., 110.
375. W. Woloszczuk et al., *Horm. Metab. Res.* **18** (1986) 197.
376. P. Demol et al., *Gastroenterology* **86** (1984) part 2, 1060.
377. D. V. Kleist et al., *Gastroenterology* **86** (1984) part 2, 1290.
378. H. J. Lenz, *Am. J. Physiol.* **254** (1988) G920.
379. Y. Tache et al., *Can. J. Physiol. Pharmacol.* (1986) July, Suppl., 53.
380. P. Bauerfeind et al., *Dig. Dis. Sci.* **30** (1985) 368.
381. M. J. Fargeas et al., *Regul. Pept.* **11** (1985) 95.
382. R. C. Orlowski et al., *Eur. J. Biochem.* **162** (1987) 399.
383. D. M. Findlay et al., *Endocrinology (Baltimore)* **117** (1985) 801.
384. D. M. Findlay et al., *Endocrinology (Baltimore)* **112** (1983) 1288.
385. R. M. Epand et al., *J. Med. Chem.* **31** (1988) 1595.
386. R. M. Epand et al., *Eur. J. Biochem.* **159** (1986) 125.
387. G. R. Moe et al., *J. Am. Chem. Soc.* **105** (1983) 4100.
388. R. M. Epand et al., *Biochem. Biophys. Res. Commun.* **152** (1988) 203.
389. M. J. Twery et al., *Eur. J. Pharmacol.* **155** (1988) 285.
390. T. Buclin et al., *Calcif. Tissue Int.* **41** (1987) 252.
391. G. Thamsborg et al., *Lancet I* (1988) 413.
392. M. T. McDermott, G. S. Kidd, *Endocr. Rev.* **8** (1987) 377.
393. L. Nogarin, *Clin. Trials J.* **23** (1986) Suppl. 1, 87.
394. K. Krueger, M. Schattenkirchner, *Z. Rheumatol.* **47** (1988) 287.
395. G. Schiraldi et al., *Eur. J. Respir. Dis.* **69** (1986) Suppl. 146, A150.
396. L. H. Breimer et al., *Biochem. J.* **255** (1988) 377.
397. T. Inui et al., *Life Sci.* **45** (1989) 1199.
398. A. Bjurhom et al., *Peptides (N.Y.)* **9** (1988) 165.
399. P. H. Steenbergh et al., *FEBS Lett.* **209** (1986) 97.
400. X.-Y. Hua, J. M. Lundberg, *Acta Physiol. Scand.* **128** (1986) 453.
401. A. F. Russo et al., *J. Biol. Chem.* **263** (1988) 5.
402. P. Popper, P. E. Micevych, *Neuroendocrinology* **50** (1989) 338.
403. M. Pohl et al., *Neuropeptides (Edinburgh)* **14** (1989) 151.
404. H. Kawasaki et al., *Nature (London)* **335** (1988) 164.
405. A. Haegerstrand et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 3299.
406. J. B. D. Palmer et al., *Br. J. Pharmacol.* **91** (1987) 95.
407. J. Wallengren, R. Hakanson, *Eur. J. Pharmacol.* **143** (1987) 267.
408. R. Oku et al., *Brain Res.* **403** (1987) 350.
409. C. Netti et al., *Neuroendocrinology* **49** (1989) 242.
410. C. Elie et al., *Neuropeptides (Edinburgh)* **16** (1990) 109.
411. J. P. Hinson, G. P. Vinson, *Neuropeptides (Edinburgh)* **16** (1990) 129.
412. H. J. Lenz et al., *Gut* **26** (1985) 550.
413. W. S. Helton et al., *Am. J. Physiol.* **256** (1989) G715.
414. L. Bartho et al., *Eur. J. Pharmacol.* **135** (1987) 449.
415. H. J. Lenz, *Am. J. Physiol.* **254** (1988) G920.
416. H. M. Cox et al., *Br. J. Pharmacol.* **97** (1989) 996.
417. B. Ahren et al., *Diabetologia* **30** (1987) 354.
418. B. Leighton et al., *FEBS Lett.* **249** (1989) 357.
419. Y. Umeda et al., *Biochem. Biophys. Res. Commun.* **154** (1988) 227.
420. J. R. Tippins et al., *Biochem. Biophys. Res. Commun.* **134** (1986) 1306.

421. C. A. Maggi et al., *Eur. J. Pharmacol.* **179** (1990) 217.
422. T. Dennis et al., *J. Pharmacol. Exp. Ther.* **251** (1989) 718.
423. M. Nishi et al., *J. Biol. Chem.* **265** (1990) 4173.
424. S. Mosselman et al., *FEBS Lett.* **239** (1988) 227.
425. A. N. Roberts et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 9662.
426. A. Clark et al., *Lancet II* (1990) 231.
427. A. Ogawa et al., *J. Clin. Invest.* **85** (1990) 973.
428. E. Hartter et al., *Lancet I* (1990) 854.
429. G. J. S. Cooper et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 7763.
430. H. K. Datta et al., *Biochem. Biophys. Res. Commun.* **163** (1989) 876. References for Chapter 4
431. J. R. Dave et al., *Endocrinology (Baltimore)* **120** (1987) 305.
432. M. Fenger, A. J. Johnsen, *Biochem. J.* **250** (1988) 781.
433. F. E. Estivariz et al., *J. Endocrinol.* **116** (1988) 207.
434. H. P. J. Bennett, *Peptides (N.Y.)* **7** (1986) 615.
435. D. C. Cumming et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 810.
436. C.-L. C. Chen, M. B. Madigan, *Endocrinology (Baltimore)* **121** (1987) 590.
437. J. N. Wilcox, J. L. Roberts, *Endocrinology (Baltimore)* **117** (1985) 2392.
438. A. Poteliakhoff, *Lancet II* (1985) 326.
439. A. C. Y. Chang et al., *Proc. Natl. Acad. Sci. USA* **77** (1980) 4890.
440. J. Drouin et al., *FEBS Lett.* **193** (1985) 54.
441. S. Nakanishi et al., *Nature (London)* **278** (1979) 423.
442. G. J. M. Martens et al., *J. Biol. Chem.* **260** (1985) 13685.
443. P. Petrusz et al., *Proc. Natl. Acad. Sci. USA* **80** (1983) 1721.
444. T. O. Bruhn et al., *Endocrinology (Baltimore)* **120** (1987) 25.
445. A. Sasaki et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 768.
446. D. J. Yoon et al., *Endocrinology (Baltimore)* **122** (1988) 759.
447. H. Jingami et al., *FEBS Lett.* **191** (1985) 63.
448. M. Patthy et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 8762.
449. T. Watabe et al., *Life Sci.* **40** (1987) 1651.
450. S. Suemaru et al., *Life Sci.* **39** (1986) 1161.
451. C. Rivier, W. Vale, *Endocrinology (Baltimore)* **114** (1984) 2409.
452. P. M. Plotsky, *Endocrinology (Baltimore)* **121** (1987) 924.
453. E. Redel et al., *Endocrinology (Baltimore)* **123** (1988) 2736.
454. S. Suemaru et al., *Endocrinol. Jpn.* **33** (1986) 441.
455. T. Suda et al., *Life Sci.* **40** (1987) 673.
456. B. Bateman et al., *Endocr. Rev.* **10** (1989) 92.
457. T. Suda et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 909.
458. F. Yajima et al., *Life Sci.* **39** (1986) 181.
459. J. C. Buckingham, T. A. Cooper, *Br. J. Pharmacol.* **90** (1987) Suppl., 167P.
460. T. Laatikainen et al., *Neuropeptides (Edinburgh)* **10** (1987) 343.
461. E. B. De Souza et al., *Nature (London)* **319** (1986) 593.
462. T. Shibasaki et al., *Life Sci.* **43** (1988) 1103.
463. A. E. Calogero et al., *Clin. Res.* **35** (1987) 393A.
464. G. Aguilera et al., *Neuroendocrinology* **43** (1986) 79.
465. M. C. Holmes et al., *Endocrinology (Baltimore)* **121** (1987) 2093.
466. J. R. Dave, R. L. Eskay, *Biochem. Biophys. Res. Commun.* **134** (1986) 255.
467. A. N. Margioris et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 922.
468. J. Kraicer et al., *Neuroendocrinology* **41** (1985) 363.
469. J. H. Meador-Woodruff et al., *Neuropeptides (Edinburgh)* **9** (1987) 269.
470. J. V. Conaglen et al., *J. Clin. Endocrinol. Metab.* **58** (1984) 463.
471. C. Maser-Gluth et al., *Life Sci.* **35** (1984) 879.
472. E. De Souza, G. R. Van Loon, *Experientia* **40** (1984) 1004.
473. D. O. Sobel, *Peptides (N.Y.)* **7** (1986) 443.
474. S. Guild, T. Reisine, *J. Pharmacol. Exp. Ther.* **241** (1987) 125.
475. D. M. Gibbs, *Endocrinology (Baltimore)* **116** (1985) 723.
476. M. Keller-Wood et al., *Am. J. Physiol.* **250** (1986) R396.
477. A. C. Hale et al., *J. Endocr.* **102** (1984) R1.
478. B. I. Baker et al., *J. Endocr.* **106** (1985) R5.
479. M. V. Graf et al., *Neuroendocrinology* **41** (1985) 353.
480. R. S. Rittmaster et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 891.
481. K. E. Nikolarakis et al.], *Brain Res.* **399** (1986) 152.

482. J. G. Kiang, E. T. Wei, *J. Pharmacol. Exp. Therap.* **243** (1987) 517. E. T. Wei et al., *J. Pharmacol. Exp. Therap.* **247** (1988) 1082.
483. K. M. Hargreaves et al., *Eur. J. Pharmacol.* **170** (1989) 275.
484. C. Rivier, W. Vale, *Endocrinology (Baltimore)* **114** (1984) 914.
485. B. Miskowiak et al., *Exp. Clin. Endocrinol.* **88** (1986) 25.
486. F. Petraglia et al., *Endocrinology (Baltimore)* **120** (1987) 1083.
487. A. Kooy et al., *Neuroendocrinology* **51** (1990) 261.
488. H. Katakami et al., *Neuroendocrinology* **41** (1985) 390.
489. A. R. M. M. Hermus et al., *Clin. Endocrinol. (Oxford)* **21** (1984) 589.
490. M. Vlaskovska et al., *Endocrinology (Baltimore)* **115** (1984) 895.
491. S. J. Konturek et al., *Peptides (N.Y.)* **8** (1987) 575.
492. Y. Tache et al., *Gastroenterology* **86** (1984) 281.
493. R. Murison et al., *Psychoneuroendocrinology (Oxford)* **14** (1989) 331.
494. M. Burlage et al., *Klin. Wochenschr.* **66** (1988) Suppl. 13, 207.
495. C. L. Williams, T. F. Burks, *Gastroenterology* **92** (1987) part 2, 1808.
496. D. Oppermann et al., *Klin. Wochenschr.* **63** (1985) 231.
497. G. Laux, K. P. Lesch, M. Schwab, *Neuropsychobiology* **19** (1988) 40.
498. G. L. Bernardini et al., *Peptides (N.Y.)* **5** (1984) 57.
499. G. F. Koob, F. E. Bloom, *Proc. Fed. Am. Soc. Exp. Biol.* **44** (1985) 259.
500. D. R. Britton et al., *Life Sci.* **38** (1985) 211.
501. A. Mele et al., *Peptides (N.Y.)* **8** (1987) 935.
502. C. R. DeBold et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 836.
503. P. C. Avgerinos et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 816.
504. D. Desir et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 1292.
505. H. Vierhapper, W. Waldhäusl, *Exp. Clin. Endocrinol.* **88** (1986) 355.
506. O. A. Müller et al., *Horm. Res.* **25** (1987) 185.
507. D. N. Orth et al., *Proc. Fed. Am. Soc. Exp. Biol.* **44** (1985) 197.
508. P. W. Gold et al., *N. Engl. J. Med.* **314** (1986) 1329.
509. Y. Hayashizaki et al., *Horm. Metabol. Res.* **18** (1986) 849.
510. L. Proulx et al., *J. Endocrinol. Invest.* **7** (1984) 257.
511. M. M. Valenca, A. Negro-Vilar, *Endocrinology (Baltimore)* **118** (1986) 32.
512. J. Knudtzon, *Horm. Metabol. Res.* **18** (1986) 579.
513. A. I. Smith et al., *J. Endocrinol.* **115** (1987) R5.
514. M. Carnes et al., *Peptides (N.Y.)* **9** (1988) 325.
515. P. M. Horrocks et al., *Clin. Endocrinol. (Oxford)* **32** (1990) 127.
516. J. C. Buckingham, *Br. Med. Bull.* **41** (1985) 203.
517. M. Familari, J. W. Funder, *Aust. N. Z. J. Med.* **17** (1987) Suppl. 1, 164.
518. J. Dohanics et al., *Peptides (N.Y.)* **11** (1990) 59.
519. J. F. Mortola et al., *J. Clin. Endocrinol. Metab.* **65** (1987) 962.
520. K. L. Bost, J. E. Blalock, *Mol. Cell. Endocrinol.* **44** (1986) 1.
521. E. M. Smith et al., *N. Engl. J. Med.* **317** (1987) 1266.
522. A. Penhoat et al., *Biochem. Biophys. Res. Commun.* **165** (1989) 355.
523. W. E. Rainey et al., *J. Biol. Chem.* **264** (1989) 21474.
524. R. Cheitlin et al., *J. Biol. Chem.* **260** (1985) 5323.
525. S. W. J. Lamberts et al., *Mol. Cell. Endocrinol.* **52** (1987) 243.
526. G. Ramirez et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 46.
527. L. J. Valenta et al., *Horm. Res.* **23** (1986) 16.
528. A. Singh et al., *Biochem. Biophys. Res. Commun.* **155** (1988) 524.
529. A. Bertolini et al., *Eur. J. Pharmacol.* **130** (1986) 19.
530. A. Bertolini et al., *Eur. J. Pharmacol.* **122** (1986) 387.
531. S. Guarini et al., *Pharmacol. Res. Commun.* **19** (1987) 511.
532. J. Chariot et al., *Seances Acad. Sci. Ser. 3,* **307** (1988) 235.
533. M. Kollind et al., *Diabetologia* **27** (1984) 298A.
534. A. J. Dunn, G. Vigle, *Neuropharmacology* **24** (1985) 329.
535. R. Poggioli et al., *Peptides (N.Y.)* **7** (1986) 843.
536. N. Chastrette et al., *Neuropeptides (Edinburgh)* **15** (1990) 61.
537. S. J. Dunmore et al., *Biochem. J.* **244** (1987) 797.

538. S. L. Wardlaw, C. E. Markowitz, *Clin. Res.* **34** (1986) 435A.
539. N. Colbert et al., *Presse Med.* **12** (1983) 1077.
540. R. C. Bell, J. M. Lipton, *Am. J. Physiol.* **252** (1987) R1152.
541. K. A. Gruber, M. F. Callahan, *Am. J. Physiol.* **257** (1989) R681.
542. J. Wortsman et al., *J. Clin. Endocrinol. Metab.* **61** (1985) 355.
543. A. M. de Lauro Castrucci et al., *Drugs Future* **15** (1990) 41.
544. G. G. Nussdorfer et al., *Biochem. Biophys. Res. Commun.* **141** (1986) 1279.
545. C. B. Newman et al., *Life Sci.* **36** (1985) 1661.
546. R. L. Reid et al., *J. Clin. Endocrinol. Metab.* **58** (1984) 773.
547. R. A. Daynes et al., *J. Invest. Dermatol.* **88** (1987) 483.
548. F. Di Giovine et al., *Br. J. Rheumatol.* **26** (1987) Suppl. 2, 62.
549. S. Y. Lin et al., *Hypertension (Dallas)* **10** (1987) 619.
550. M.-T. Pham-Huu-Trung et al., *J. Clin. Endocrinol. Metab.* **61** (1985) 467.
551. K. A. Gruber, M. F. Callahan, *Am. J. Physiol.* **257** (1989) R681.
552. V. J. Hruby et al., *J. Med. Chem.* **30** (1987) 2126.
553. M. Lebl et al., *Collect. Czech. Chem. Commun.* **49** (1984) 2680.
554. N. Levine et al., *J. Invest. Dermatol.* **87** (1986) 416.
555. D. G. Klemes et al., *Biochem. Biophys. Res. Commun.* **137** (1986) 722.
556. D. B. Richards, J. M. Lipton, *Peptides (N.Y.)* **5** (1984) 815.
557. F. Al-Obeidi et al., *J. Med. Chem.* **32** (1989) 2555.
558. W. L. Cody et al., *J. Med. Chem.* **28** (1985) 583.
559. *Drugs Future* **7** (1982) 319; **8** (1983) 468; **9** (1984) 381; **10** (1985) 431.
560. *Drugs Future* **11** (1986) 432; **12** (1987) 505; **13** (1988) 489; **14** (1989) 480.
561. *Drugs Future* **14** (1989) 514; **15** (1990) 621.
562. G. Wiemer et al., *Arch. Pharmacol.* **335** (1987) Suppl., R94.
563. G. Wolterink, J. M. van Ree, *Brain Res.* **421** (1987) 41.
564. T. K. Sawyer et al., *Peptides (N.Y.)* **11** (1990) 351.
565. M. E. Hadley et al., *Life Sci.* **40** (1987) 1889.
566. W. R. Kiessling, *Arch. Neurol. (Chicago)* **44** (1987) 995.
567. P. K. Sorensen et al., *Prog. Neuropsychopharmacol. Biol. Psychiatry* **10** (1986) 479.
568. I. Kitchen, *Gen. Pharmac.* **16** (1985) 79.
569. A. T. McKnight et al., *Neuropharmacology* **24** (1985) 1011.
570. R. S. Rapka, *Life Sci.* **39** (1986) 1825.
571. A. H. Mulder et al., *Neuropeptides (Edinburgh)* **14** (1989) 99.
572. J. Gairin et al., *J. Pharmacol. Exp. Ther.* **245** (1988) 995.
573. R. Quirion, A. S. Weiss, *Peptides (N.Y.)* **4** (1983) 445.
574. P. Sanchez-Blazquez et al., *Neuropeptides (Edinburgh)* **5** (1984) 181.
575. G. W. Pasternak, *Biochem. Pharmacol.* **35** (1986) 361.
576. T. F. Burks et al., *Life Sci.* **43** (1988) 2177.
577. A. H. Mulder et al., *Nature (London)* **308** (1984) 278.
578. S. B. Weinberger et al., *Regul. Pept.* **26** (1989) 323.
579. Z. Khalil, R. D. Helme, *Neuropeptides (Edinburgh)* **17** (1990) 45.
580. A. Jackson, S. J. Cooper, *Neuropharmacology* **25** (1986) 653.
581. A. Pfeiffer et al., *Endocrinology (Baltimore)* **116** (1985) 2688.
582. T. P. Blackburn et al., *Br. J. Pharmacol.* **89** (1986) 593.
583. I. Kitchen, *Gen. Pharmacol.* **16** (1985) 79.
584. F. A. Aleem et al., *Fertil. Steril.* **45** (1986) 507.
585. R. Singer et al., *Experientia* **41** (1985) 64.
586. T. Wahlström et al., *Life Sci.* **36** (1985) 987.
587. I. Shimizu et al., *Regul. Pept.* **16** (1986) 331.
588. M. Dennis et al., *J. Chromatogr.* **266** (1983) 163.
589. O. Vuolteenaho, *Acta Physiol. Scand. Suppl.* **531** (1984) 1.
590. C. W. Wilkinson, D. M. Dorsa, *Neuroendocrinology* **43** (1986) 124.
591. W. Knepel, G. Meyen, *Neuroendocrinology* **43** (1986) 44.
592. D. Copolov et al., *Clin. Endocrinol. (Oxford)* **19** (1983) 575.
593. C. Gianoulakis et al., *Endocrinology (Baltimore)* **122** (1988) 817.
594. F. R. Caspari, S. Alapin-Rubillovitz, *J. Clin. Endocrinol. Metab.* **60** (1985) 34.
595. C. Shaba et al., *Endocrinology (Baltimore)* **115** (1984) 378.
596. A. Fabbri et al., *Endocrinology (Baltimore)* **122** (1988) 749.

597. F. Petraglia et al., *Clin. Endocrinol. (Oxford)* **24** (1986) 609.
598. T. Barreca et al., *Life Sci.* **38** (1986) 2263.
599. C. Gianoulakis et al., *Life Sci.* **45** (1989) 1097.
600. F. Facchinetti et al., *J. Endocrinol. Invest.* **9** (1986) 11.
601. A. R. Genazzani et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 36.
602. W. O. Richter et al., *Endocrinology (Baltimore)* **120** (1987) 1472.
603. S. Itow et al., *Can. J. Physiol. Pharmacol.* **63** (1985) 81.
604. J. C. Buckingham, *Neuroendocrinology* **42** (1986) 148.
605. S. L. Wardlaw et al., *Endocrinology (Baltimore)* **119** (1986) 112.
606. D. Giugliano et al., *Metab. Clin. Exp.* **36** (1987) 974.
607. A. Y. S. Lee et al., *Int. J. Pept. Protein Res.* **24** (1984) 525.
608. L.-F. Tseng, *J. Pharmacol. Exp. Ter.* **239** (1986) 160.
609. J. P. Blanc, E. T. Kaiser, *J. Biol. Chem.* **259** (1984) 9549.
610. D. Yamashiro et al., *Int. J. Pept. Protein Res.* **24** (1984) 520.
611. P. Nicolas et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 3074.
612. D. Yamashiro et al., *Int. J. Pept. Protein Res.* **24** (1984) 516.
613. J. Blake et al., *Int. J. Pept. Protein Res.* **25** (1985) 575.
614. D. Collado-Escobar et al., *Eur. J. Pharmacol.* **129** (1986) 385.
615. H. H. Suh et al., *Peptides (N.Y.)* **8** (1987) 123.
616. A. E. Panerai et al., *Eur. J. Pharmacol.* **99** (1984) 341.
617. M. D. Hirsch, T. L. O'Donohue, *J. Pharmacol. Exp. Ther.* **237** (1986) 378.
618. W. O. Richter et al., *Endocrinology (Baltimore)* **120** (1987) 1472.
619. K. Yoshikawa et al., *J. Biol. Chem.* **259** (1984) 14301.
620. G. J. M. Martens, E. Herbert, *Nature (London)* **310** (1984) 251.
621. N. Zamir et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 6886.
622. I. Kitchen, *Gen. Pharmac.* **16** (1985) 79.
623. E. A. Singer et al., *Endocrinology (Baltimore)* **119** (1986) 1527.
624. T. Kanamatsu et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 9245.
625. J. D. Watson et al., *J. Endocrinol.* **111** (1986) 329.
626. J. S. Hong et al., *Neuropeptides (Edinburgh)* **5** (1985) 557.
627. L.-F. Tseng et al., *Regul. Pept.* **14** (1986) 181.
628. K. Yoshikawa, S. L. Sabol, *Biochem. Biophys. Res. Commun.* **139** (1986) 1.
629. J. A. Angulo et al., *Eur. J. Pharmacol.* **130** (1986) 341.
630. S. L. Lightman, W. S. Young III, *Proc. Natl. Acad. Sci. USA* **86** (1989) 4306.
631. D. C.-C. Wan et al., *Neuropeptides (Edinburgh)* **16** (1990) 141.
632. D. Kew, D. L. Kilpatrick, *Mol. Endocrinol.* **3** (1989) 179.
633. D. K. Sommers et al., *Eur. J. Clin. Pharmacol.* **38** (1990) 391.
634. A. Pfeiffer, A. Herz, *Horm. Metab. Res.* **16** (1984) 386.
635. A. Tsunoo et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 9832.
636. M. Bickl, *Dig. Dis. Sci.* **29** (1984) Suppl., 10S.
637. J. C. Reynolds et al., *Am. J. Physiol.* **246** (1984) part 1, G130.
638. A. Dray et al., *Pain (Amsterdam)* **20** (1984) Suppl. 2, S230.
639. R. F. Crochelt et al., *Gastroenterology* **84** (1983) part 2, 1403.
640. J. Behar, P. Biancani, *Gastroenterology* **86** (1984) 134.
641. P. A. Crooks et al., *Biochem. Pharmacol.* **33** (1984) 4095.
642. A. Dray, R. Metsch, *Eur. J. Pharmacol.* **98** (1984) 155.
643. M. D. Owen et al., *Peptides (N.Y.)* **5** (1984) 737.
644. G. E. Sander et al., *Peptides (N.Y.)* **6** (1985) 133.
645. D. M. Wright, G. Clarke, *Neuropeptides (Edinburgh)* **5** (1984) 273.
646. K. Hashimoto et al., *Endocrinol. Jpn.* **33** (1986) 813.
647. P. G. Andreis et al., *Neuropeptides (Edinburgh)* **12** (1988) 165.
648. B. Allilio et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 1427.
649. F. Facchinetti et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 1222.
650. I. Gerendai et al., *Regul. Pept.* **27** (1990) 107.
651. T. Mitsuma et al., *Horm. Metab. Res.* **21** (1989) 301.
652. N. F. Plotnikoff et al., *Fed. Proc. Fed. Am. Soc. Exp.* **44** (1985) 118.
653. Y. Shavit et al., *Science (Washington, D.C.)* **223** (1984) 188.
654. J. L. Torres et al., *Experientia* **45** (1989) 574.

655. M. Maeda et al., *Chem. Pharm. Bull.* **32** (1984) 4157.
656. G. W. Hardy et al., *J. Med. Chem.* **31** (1988) 960.
657. *Drugs Future* **11** (1986) 657.
658. N. Zamir et al., *Nature (London)* **307** (1984) 643.
659. S. Lemaire et al., *Can. J. Physiol. Pharmacol.* **62** (1984) 484.
660. E. Weihe et al., *Neuropeptides (Edinburgh)* **5** (1985) 453.
661. H. J. Wolter, *Peptides (N.Y.)* **7** (1986) 389.
662. S. Spampinato et al., *Life Sci.* **38** (1986) 403.
663. H.-K. Jiang et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 2948.
664. A. I. Faden et al., *Regul. Pept.* **11** (1985) 35.
665. A. Z. Zhang et al., *Neuropeptides (Edinburgh)* **5** (1985) 553.
666. R. Schick et al., *Dig. Dis. Sci.* **29** (1984) Suppl., 74S.
667. T. Mitsuma et al., *Horm. Metab. Res.* **21** (1989) 301.
668. N. Falke, *Neuropeptides (Edinburgh)* **11** (1988) 163.
669. J. A. T. Haaf et al., *Acta Endocrinol. (Copenhagen)* **114** (1987) 96.
670. A. Rezvani, E. L. Way, *Eur. J. Pharmacol.* **102** (1984) 475.
671. M. A. Della-Fera et al., *Am. J. Physiol.* **258** (1990) R946.
672. H. Shimizu et al., *Life Sci.* **45** (1989) 25.
673. H. Tilson et al., *Neuropeptides (Edinburgh)* **8** (1986) 193.
674. A. I. Faden et al., *Peptides (N.Y.)* **4** (1983) 631.
675. P. Sanchez-Blazquez et al., *Neuropeptides (Edinburgh)* **4** (1984) 369.
676. A. Turcotte et al., *Int. J. Pept. Protein Res.* **23** (1984) 361.
677. H. Yoshino et al., *J. Med. Chem.* **33** (1990) 206.
678. J. E. Gairin et al., *J. Med. Chem.* **29** (1986) 1913.
679. J. E. Gairin et al., *Br. J. Pharmacol.* **95** (1988) 1023.
680. Y. Kiso et al., *J. Pharmacobio Dyn.* **7** (1984) S-91.
681. H. L. Wen et al., *Peptides (N.Y.)* **8** (1987) 191.
682. *Drugs Future* **16** (1991) 203.
683. K. Richter et al., *Science (Washington, D.C.)* **238** (1987) 200.
684. A. Mor et al., *FEBS Lett.* **255** (1989) 2690.
685. A. Mor et al., *Neuropeptides (Edinburgh)* **13** (1989) 51.
686. D. Giudici et al., *Neuroendocrinology* **39** (1984) 236.
687. E. C. degli Uberti et al., *Horm. Res.* **24** (1986) 251.
688. E. Roti et al., *J. Endocrinol. Invest.* **7** (1984) 211.
689. H. G. Gullner et al., *Arch. Int. Pharmacodyn. Ther.* **266** (1983) 155.
690. E. C. Degli Uberti et al., *J. Clin. Endocrinol. Metab.* **61** (1985) 1018.
691. F. Petraglia et al., *Peptides (N.Y.)* **6** (1985) 869.
692. S. J. Konturek et al., *Gastroenterology* **84** (1983) part 2, 1213.
693. F. Pavone et al., *Peptides (N.Y.)* **11** (1990) 591.
694. A. Rossi et al., *Peptides (N.Y.)* **4** (1983) 577.
695. K. M. Sivanandaiah et al., *Int. J. Pept. Protein Res.* **33** (1989) 463.
696. K. Darlak et al., *Peptides (N.Y.)* **5** (1984) 687.
697. K. Kisara et al., *Br. J. Pharmacol.* **87** (1986) 183.
698. M. A. Cervini et al., *Peptides (N.Y.)* **6** (1985) 433.
699. S. Salvadori et al., *Eur. J. Med. Chem. Chim. Ther. (Paris)* **18** (1983) 489.
700. S. Salvadori et al., *Hoppe Seyler's Z. Physiol. Chem.* **365** (1984) 1199.
701. K. Suzuki et al., *Chem. Pharm. Bull.* **33** (1985) 4865.
702. K. Chaki et al., *Peptides (N.Y.)* **11** (1990) 139.
703. Y. Sasaki et al., *Biochem. Biophys. Res. Commun.* **120** (1984) 214.
704. K. Suzuki et al., *Chem. Pharm. Bull.* **36** (1988) 4834.
705. S. Salvadori et al., *J. Med. Chem.* **29** (1986) 889.
706. M. Marastoni et al., *J. Med. Chem.* **30** (1987) 1538.
707. G. Balboni et al., *Biochem. Biophys. Res. Commun.* **169** (1990) 617.
708. R. Greenberg et al., *J. Biol. Chem.* **259** (1984) 5132.
709. K. J. Chang et al., *J. Biol. Chem.* **260** (1985) 9706.
710. L. Graf et al., *Neuropeptides (Edinburgh)* **9** (1987) 113.
711. V. Schusdziarra et al., *Diabetologia* **24** (1983) 113.
712. D. Tome et al., *Am. J. Physiol.* **253** (1987) G737.
713. V. Schusdziarra et al., *Endocrinology (Baltimore)* **112** (1983) 885.
714. J. Nedvidkova et al., *Exp. Clin. Endocrinol.* **85** (1985) 249.

715. T. Mitsuma et al., *Exp. Clin. Endocrinol.* **84** (1984) 324.
716. M. Yoshikawa et al., *Agric. Biol. Chem.* **48** (1984) 3185.
717. K.-J. Chang et al., *J. Pharmacol. Exp. Ther.* **227** (1983) 403.
718. H. Matthies et al., *Peptides (N.Y.)* **5** (1984) 463.
719. M. Schaefer et al., *Cell (Cambridge Mass.)* **41** (1985) 457.
720. R. H. M. Ebberink et al., *Peptides (N.Y.)* **8** (1987) 515.
721. D. A. Price et al., *Peptides (N.Y.)* **8** (1987) 533.
722. R. B. Raffa, *Peptides (N.Y.)* **9** (1988) 915.
723. C. S. Barnard, G. J. Dockray, *Regul. Pept.* **8** (1984) 209.
724. T. Austin et al., *Can. J. Physiol. Pharamcol.* **61** (1983) 949.
725. H. Y. T. Yang et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 7757.
726. M. Kavaliers, M. Hirst, *Peptides (N.Y.)* **6** (1985) 847.
727. X. Z. Zhu, R. B. Raffa, *Neuropeptides (Edinburgh)* **8** (1986) 55.
728. G. Telegdy, I. Bollok, *Neuropeptides (Edinburgh)* **10** (1987) 157.
729. A. Ottlecz, G. Telegdy, *Neuropeptides (Edinburgh)* **9** (1987) 161.
730. K. Kouge et al., *Bull. Chem. Soc. Jpn.* **60** (1987) 4343.
731. A. Fukamizi et al., *J. Biol. Chem.* **265** (1990) 7576.
732. E. T. Ben-Ari et al., *J. Biol. Chem.* **264** (1989) 13074.
733. N. Itoh et al., *FEBS Lett.* **244** (1989) 6.
734. E. Coezy et al., *Am. J. Physiol.* **257** (1989) C888.
735. T. Inagami et al., *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **45** (1986) 1414.
736. M. Naruse et al., *J. Clin. Endocrinol. Metab.* **61** (1985) 480.
737. C. F. Deschepper et al., *Endocrinology (Baltimore)* **119** (1986) 36.
738. M. C. Chappell et al., *J. Biol. Chem.* **264** (1989) 16518.
739. M. Bumpus et al., *Am. J. Med. Sci.* **295** (1988) 406.
740. D. G. Changaris et al., *Biochem. Biophys. Res. Commun.* **138** (1986) 573.
741. O. Kurauchi et al., *Horm. Metab. Res.* **21** (1989) 558.
742. D. D. Smyth, H. Y. M. Fung, *Eur. J. Pharmacol.* **102** (1984) 55.
743. J. C. Lee et al., *Am. J. Physiol.* **251** (1986) R258.
744. J. Mann, *Acta Endocrinol. (Copenhagen)* **102** (1983) Suppl. 253, 10.
745. N. R. Levens, *Am. J. Physiol.* **246** (1984) Part 1, G634.
746. H. F. Janssen et al., *Circ. Shock* **23** (1987) 197.
747. A. T. Chiu et al., *Biochem. Biophys. Res. Commun.* **165** (1989) 196.
748. R. S. L. Chang et al., *Biochem. Biophys. Res. Commun.* **171** (1990) 813.
749. R. Tabrizchi et al., *Life Sci.* **43** (1988) 537.
750. A. Israel et al., *Neuroendocrinology* **42** (1986) 57.
751. E. Castren, J. M. Saavedra, *Endocrinology (Baltimore)* **122** (1988) 370.
752. I. Luoveau et al., *Biochem. Biophys. Res. Commun.* **163** (1989) 32.
753. B. Bouscarel et al., *J. Biol. Chem.* **263** (1988) 14920.
754. B. J. Ballermann et al., *Am. J. Physiol.* **247** (1984) F110.
755. G. Andoka et al., *Biochem. Biophys. Res. Commun.* **121** (1984) 441.
756. P. D. Carriere et al., *Neuroendocrinology* **43** (1986) 49.
757. J. Bingham Smith et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 7812.
758. J. A. Gilbert et al., *Biochem. Pharmacol.* **33** (1984) 2527.
759. H. Satoh et al., *Biochem. Biophys. Res. Commun.* **126** (1985) 464.
760. I. Antonipillai et al., *Clin. Res.* **35** (1987) 135A.
761. A. Nakamura et al., *Am. J. Physiol.* **258** (1990) E1.
762. M. E. Olson et al., *Clin. Sci.* **72** (1987) 429. K. Yamaguchi et al., *Acta Endocrinol. (Copenhagen)* **118** (1988) 82.
763. V. L. Brooks et al., *Circ. Res.* **58** (1986) 829.
764. M. C. Khosla et al., *Hypertension (Dallas)* **11** (1988) Suppl. I, 1.
765. J. E. Hall, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **45** (1986) 1431.
766. V. L. Schuster, *Federation Proc. Fed. Am. Soc. Exp. Biol.* **45** (1986) 1444.
767. N. R. Levens, *Am. J. Physiol.* **245** (1983) G511.
768. G. G. Nussdorfer et al., *Exp. Clin. Endocrinol.* **88** (1986) 158.
769. L. L. Espey et al., *Proc. Soc. Exp. Biol. Med.* **193** (1990) 249.
770. D. Stirling et al., *J. Biol. Chem.* **265** (1990) 5.
771. T. Kono et al., *Endocrinol. Jpn.* **34** (1987) 737.

772. M. T. Schiavone et al., *Neurobiolog.* **85** (1988) 4095.
773. K. Y. Hui et al., *Int. J. Pept. Protein Res.* **34** (1989) 177.
774. T. Kono et al., *Life Sci.* **37** (1985) 365.
775. G. Guillemette et al., *J. Med. Chem.* **27** (1984) 315.
776. A. Husain et al., *J. Pharmacol. Exp. Ther.* **239** (1986) 71.
777. J. Samanen et al., *J. Med. Chem.* **32** (1989) 466.
778. K. L. Spear et al., *J. Med. Chem.* **33** (1990) 1935.
779. K.-H. Hsieh et al., *J. Med. Chem.* **32** (1989) 898.
780. J. Samanen et al., *J. Med. Chem.* **31** (1988) 510.
781. O. Nyeki et al., *J. Med. Chem.* **30** (1987) 1719.
782. M. N. Scanlon et al., *Life Sci.* **34** (1984) 317.
783. M. H. Goghari et al., *J. Med. Chem.* **29** (1986) 1121.
784. J. Samanen et al., *J. Med. Chem.* **32** (1989) 1366.
785. P. R. Bovy et al., *J. Med. Chem.* **32** (1989) 520.
786. P. R. Bovy et al., *J. Med. Chem.* **33** (1990) 1477.
787. P. C. Wong et al., *Hypertension (Dallas)* **12** (1988) 340.
788. J. V. Duncia et al., *J. Med. Chem.* **33** (1990) 1312.
789. J. Samanen et al., *J. Med. Chem.* **31** (1988) 737.
790. J. E. Taylor et al., *Drug. Dev. Res.* **16** (1989) 1.
791. N. B. Oza et al., *J. Clin. Invest.* **85** (1990) 597.
792. D. I. Diz, *Peptides (Fayetteville, N.Y.)* **6** (1985) 57.
793. W. E. Wilson et al., *J. Biol. Biochem.* **264** (1989) 17777.
794. H. Kato et al., *FEBS Lett.* **232** (1988) 252.
795. K.-I. Enjyoji, H. Kato, *FEBS Lett.* **238** (1988) 1.
796. H. Maeda et al., *J. Biol. Chem.* **263** (1988) 1605.
797. K.-I. Enjyoji et al., *J. Biol. Chem.* **263** (1988) 973.
798. A. Barlas et al., *FEBS Lett.* **218** (1987) 266.
799. N. Gutman et al., *Eur. J. Biochem.* **171** (1988) 577.
800. J. A. Veenstra, *FEBS Lett.* **250** (1989) 231.
801. G. Gäde, R. Kellner, *Peptides (Fayetteville, N.Y.)* **10** (1989) 1287. Y. Kuroki et al., *Biochem. Biophys. Res. Commun.* **167** (1990) 273. L. Schoofs et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 427.
802. H. Wilkens et al. in A. Bertelli, N. Back (eds.): *Shock, Biochemical, Pharmacological and Clinical Aspects,* Plenum Press, New York 1970, pp. 201.
803. D. Proud, A. P. Kaplan, *Ann. Rev. Immunol. (Palo Alto)* **6** (1988) 49.
804. M. Wicklmayr et al., *Horm. Metab. Res.* **20** (1988) 535.
805. L. L. Espey et al., *Am. J. Physiol.* **251** (1986) E362.
806. G. M. Patrassi et al., *Thromb. Haemostasis* **50** (1983) 228.
807. C. P. Quilley et al., *Hypertension (Dallas)* **10** (1987) 294.
808. H. Okunishi et al., *Hypertension* **8** (1986) Suppl. 1, 114.
809. J. Spragg et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 203.
810. G. Caspritz et al., *Arzneim. Forsch.* **36** (1986) 1605.
811. S. Kumakura et al., *J. Pharmacol. Exp. Ther.* **243** (1987) 1067.
812. R. E. Ferner et al., *Br. J. Clin. Pharmacol.* **25** (1988) 105P.
813. E. G. Erdös, *J. Cardiovasc. Pharmacol.* **15** (1990) Suppl. 6, S20.
814. J. M. Stewart, R. J. Vavrek, *J. Cardiovasc. Pharmacol.* **15** (1990) S69.
815. I. A. Sheikh, A. P. Kaplan, *Biochem. Pharmacol.* **38** (1989) 993.
816. K. Shimamoto et al., *J. Cardiovasc. Pharmacol.* **15** (1990) S83.
817. P. Maddedu et al., *J. Hypertens.* **5** (1987) 645.
818. I. Miyamori et al., *Exp. Clin. Endocrinol.* **87** (1986) 169.
819. D. Proud et al., paper presented at the congress *Kinin 84,* Savannah, Ga., 1984.
820. R. Kayegama et al., *J. Biol. Chem.* **260** (1985) 12060.
821. A. Barlas et al., paper presented at the congress *Kinin 84,* Savannah, Ga., 1984.
822. F. Sierra et al., *Mol. Cell. Biol.* **9** (1989) 5610.
823. K. M. Braas et al., *Br. J. Pharmacol.* **94** (1988) 3.
824. A. A. Roscher et al., *J. Cardiovasc. Pharmacol.* **15** (1990) Suppl. 6, S39.
825. O. A. Carretero, A. G. Scicli in J. H. Laragh et al. (eds.): *Endocrine Mechanisms in Hypertension,* Raven Press, New York 1989, p. 219.

826. D. Regoli et al., *J. Cardiovasc. Pharmacol.* **15** (1990) Suppl. 6, S30.
827. J. M. Ritter et al., *Br. J. Pharmacol.* **96** (1989) 23.
828. C. R. Mantione, R. Rodriguez, *Br. J. Pharmacol.* **99** (1990) 516.
829. B. S. Dixon et al., *Am. J. Physiol.* **258** (1990) C299.
830. B. Waeber et al., *J. Cardiovasc. Pharmacol.* **15** (1990) Suppl. 6, S78.
831. K. Inokuchi, K. U. Malik, *Am. J. Physiol.* **246** (1984) F387.
832. E. T. Whalley et al., *Naunym-Schmiedeberg's Arch. Pharmacol.* **336** (1987) 652.
833. B. A. Schölkens et al., *J. Hypertens.* **5** (1987) Suppl. 5, S7.
834. Y. Yoshimura et al., *Endocrinology (Baltimore)* **122** (1988) 2540.
835. P. Geppetti et al., *Br. J. Pharmacol.* **94** (1988) 288.
836. G. Dietze et al., Paper presented at the congress *Kinin 84,* Savannah, Ga., 1984.
837. N.-E. Rhaleb et al., *Br. J. Pharmacol.* **99** (1990) 445.
838. J. Barabe et al., *Can. J. Physiol. Pharmacol.* **62** (1984) 627.
839. G. Drapeau et al., *Eur. J. Pharmacol.* **155** (1988) 193.
840. R. J. Vavrek et al. in E. Giralt, D. Andersen (eds.): *Peptides 1990,* ESCOM Sci. Publ., Leiden 1991, p. 642.
841. F. J. Hook et al., *Br. J. Pharmacol.* **102** (1991) 769.
842. R. J. Vavrek, J. M. Stewart in C. M. Deber et al. (eds.): *Peptides. Structure and Function,* Pierce Chem. Comp., Rockford, Ill., 1985, p. 655.
843. J. M. Stewart, R. J. Vavrek, *J. Cardiovasc. Pharmacol.* **15** (1990) Suppl. 6, S69.
844. D. Regoli et al., *Eur. J. Pharmacol.* **123** (1986) 61.
845. D. Regoli et al., *Eur. J. Pharmacol.* **127** (1986) 219.
846. P. Devillier et al., *Eur. J. Pharmacol.* **149** (1988) 137.
847. A. S. Dutta, *Drugs Future* **12** (1987) 781.
848. Y. Kawaguchi et al., *Biochem. Biophys. Res. Commun.* **139** (1986) 1040.
849. A. J. Harmar et al., *FEBS Lett.* **208** (1986) 67.
850. C.-R. Martling et al., *Life Sci.* **40** (1987) 1633.
851. J. M. Conlon et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 859.
852. E. Theodorsson-Norheim et al., *Eur. J. Biochem.* **166** (1987) 693.
853. A. Bucsics et al., *Peptides (Fayetteville, N.Y.)* **4** (1983) 451.
854. A. Ericsson et al., *Mol. Endocrinol.* **4** (1990) 1211.
855. B. Beding-Barnekow, E. Brodin, *Regul. Pept.* **25** (1989) 199.
856. J. M. Conlon et al., *FEBS Lett.* **200** (1986) 111.
857. L. Schoofs et al., *FEBS Lett.* **261** (1990) 397.
858. B. A. Yankner et al., *Science (Washington D.C.)* **250** (1990) 279.
859. J. L. Vaught, *Life Sci.* **43** (1988) 1419.
860. F. Angel et al., *Gastroenterology* **84** (1983) Part 2, 1092.
861. J. Wiley, C. Owyang, *Am. J. Physiol.* **252** (1987) G431.
862. A. Mandahl, *Eur. J. Pharmacol.* **114** (1985) 121.
863. P. Tonnesen, O. B. Schaffalitzky de Muckadell, *Allergy (Copenhagen)* **42** (1987) 146.
864. M. J. Bannon et al., *J. Biol. Chem.* **261** (1986) 6640.
865. I. Norheim et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 605.
866. P. Devillier et al., *N. Engl. J. Med.* **314** (1986) 1323.
867. J. M. Stewart, M. E. Hall in T. Shiba, S. Sakakibara (eds.): *Peptide Chemistry 1987,* Protein Research Foundation, Osaka 1988, p. 449.
868. J. M. Conlon, L. Sheehan, *Regul. Pept.* **7** (1983) 335.
869. D. J. Dusser et al., *J. Clin. Invest.* **84** (1989) 900.
870. W. E. Siems et al., *Pharmazie* **42** (1987) 153.
871. S. Lavielle et al., *Biochem. Pharmacol.* **37** (1988) 41.
872. Y. Yokota et al., *J. Biol. Chem.* **264** (1989) 17649.
873. Y. Masu et al., *Nature (London)* **329** (1987) 836.
874. D. Regoli et al., *Trends Pharmacol. Sci. TIPS* **9** (1988) 290.
875. S. H. Buck, S. A. Shatzer, *Life Sci.* **42** (1988) 2701.
876. S. H. Buck, J. L. Krstenansky, *Eur. J. Pharmacol.* **144** (1987) 109.
877. I. Haro et al., *J. Pharm. Sci.* **79** (1990) 74.
878. R. M. Hagan et al., *Br. J. Pharmacol.* **98** (1989) Suppl., 717P.
879. P. Rovero, V. Pestellini, *Neuropeptides (Edinburgh)* **10** (1987) 355.
880. P. Rovero et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 593.

881. S. H. Buck et al., *Life Sci.* **47** (1990) PL-37.
882. M. Chorev et al., *Eur. J. Pharmacol.* **127** (1986) 187.
883. C. J. Mussap, E. Burcher, *Peptides (Fayetteville, N.Y.)* **11** (1990) 827.
884. J.-M. Qian et al., *J. Biol. Chem.* **264** (1989) 16667.
885. T. Sakurada et al., *Eur. J. Pharmacol.* **174** (1989) 153.
886. P. Ward et al., *J. Med. Chem.* **33** (1990) 1848.
887. J. Mizrahi et al., *Eur. J. Pharmacol.* **118** (1985) 25.
888. D. Regoli et al., *Eur. J. Pharmacol.* **109** (1985) 121.
889. P. Rovero et al., *Eur. J. Pharmacol.* **175** (1990) 113.
890. G. Drapeau et al., *Regul. Pept.* **31** (1990) 125.
891. C. A. Maggi et al., *Eur. J. Pharmacol.* **177** (1990) 81.
892. E. Naline et al., *Am. Rev. Respir. Dis.* **140** (1989) 679.
893. S. Brunelleschi et al., *Br. J. Pharmacol.* **100** (1990) 417.
894. S. Evangelista et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 79.
895. C. A. Maggi et al., *Br. J. Pharmacol.* **100** (1990) 588.
896. I. Rollandy et al., *Neuropeptides (Edinburgh)* **13** (1989) 175. M. Arisawa et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 7290.
897. M. Lotz et al., *Science (Washington D.C.)* **235** (1987) 893.
898. J. E. T. Fox, E. E. Daniel, *Am. J. Physiol.* **250** (1986) G21.
899. L. D. Hirning, T. F. Burks, *Gastroenterology* **84** (1983) part 2, 1188.
900. J. P. McGillis et al., *Fed. Proc. Fed. Am. Soc. Exp.* **46** (1987) 196.
901. I. Paegelow et al., *Pharmazie* **44** (1989) 145.
902. M. E. Hall, J. M. Stewart, *Peptides (Fayetteville, N.Y.)* **5** (1984) 85.
903. T. Sakurada et al., *Jpn. J. Pharmacol.* **43** (1987) Suppl., 272P.
904. K. D. Yench et al., *Patol. Fiziol. Eksp. Ter.* 1984, no. 3, 75.
905. R. Michelot et al., *Eur. J. Med. Chem. Chim. Ther.* **23** (1988) 243.
906. C. Poulos et al., *J. Med. Chem.* **30** (1987) 1512.
907. Y. Higuchi et al., *Eur. J. Pharmacol.* **160** (1989) 413.
908. E. Munekata et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 169.
909. A. Ljungquist et al., *Regul. Pept.* **24** (1989) 283.
910. K. Folkers et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 4833.
911. R. Hakanson et al., *Regul. Pept.* **31** (1990) 75.
912. Z. Wiesenfeld-Hallin et al., *Regul. Pept.* **29** (1990) 1.
913. J.-M. Qian et al., *J. Biol. Chem.* **264** (1989) 16667.
914. P. J. Woll, E. Rozengurt, *Proc. Natl. Acad. Sci. USA* **85** (1988) 1859.
915. G. G. Glenner, *Cell* **52** (1988) 307.
916. C. L. Joachim et al., *Nature (London)* **341** (1989) 226.
917. J. Kang et al., *Nature (London)* **325** (1987) 733.
918. R. E. Tanzi et al., *Science (Washington D.C.)* **235** (1987) 880.
919. D. Goldgaber et al., *Science (Washington D.C.)* **235** (1987) 877.
920. M. R. Palmert et al., *Science (Washington D.C.)* **241** (1988) 1080.
921. D. Schubert et al., *Science (Washington D.C.)* **241** (1988) 223.
922. W. E. van Nostrand et al., *Science (Washington D.C.)* **248** (1990) 745.
923. F. Prelli et al., *Biochem. Biophys. Res. Commun.* **151** (1988) 1150.
924. D. Allsop et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 2790.
925. T. Dyrks et al., *EMBO J.* **7** (1988) 949.
926. A. I. Bush et al., *J. Biol. Chem.* **265** (1990) 15977.
927. T. Saitoh et al., *Cell* **58** (1989) 615.
928. B. A. Yankner et al., *Science (Washington D.C.)* **250** (1990) 279.
929. V. Geenen et al., *Science (Washington D.C.)* **232** (1986) 508.
930. R. Ivell, D. Richter, *EMBO J.* **3** (1984) 2351.
931. S. Lundin, G. Toresson, *Horm. Metab. Res.* **19** (1987) 629.
932. B. Liu et al., *J. Biol. Chem.* **263** (1988) 72.
933. B. Liu, J. P. H. Burbach, *J. Neuroendocrinology* **1** (1989) no. 1.
934. M. Mühlethaler et al., *Experientia* **40** (1984) 777.
935. V. A. Batra et al., *Regul. Peptides* **30** (1990) 7.
936. D. P. Brooks et al., *Experientia* **42** (1986) 1012.
937. P. Soelberg Sorensen, M. Hammer, *Am. J. Physiol.* **248** (1985) R78.
938. B. C. Wang, L. Share, K. L. Goetz, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **44** (1985) 72.
939. B. Waeber et al., *Am. J. Physiol.* **247** (1984) H895.

940. M. R. Murphy et al., *J. Clin. Endocrinol. Metab.* **65** (1987) 738.
941. S. W. T. Cheng, W. G. North, *Neuroendocrinology* **42** (1986) 174.
942. C. M. Gregg, *Neuroendocrinology* **40** (1985) 423.
943. M. T. Shiavone et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 4095.
944. R. Cacabelos et al., *Neuroendocrinology* **45** (1987) 368.
945. T. Ivanyi et al., *Exp. Clin. Endocrinol.* **88** (1986) 303.
946. O. Willoughby et al., *Neuroendocrinology* **45** (1987) 219.
947. T. Kimura et al., *Endocrinology (Baltimore)* **114** (1984) 1426.
948. I. Os et al., *Acta Med. Scand.* **220** (1986) 195.
949. G. Geelen et al., *Am. J. Physiol.* **247** (1984) R968.
950. H. Ishizawa et al., *Eur. J. Pharmacol.* **189** (1990) 119.
951. J. R. Sladek, Jr., C. D. Sladek, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **44** (1985) 66.
952. B. Bardrum et al., *Endocrinology (Baltimore)* **123** (1988).
953. T. O. Bruhn et al., *Endocrinology (Baltimore)* **119** (1986) 1558.
954. R. Landgraf et al., *Am. J. Physiol.* **259** (1990) R1056.
955. N. W. Kasting, *Can. J. Physiol. Pharmacol.* **64** (1986) 1575.
956. D. Schams et al., *J. Endocrinol.* **102** (1984) 337.
957. S. Thornton et al., *Br. Med. J.* **297** (1988) 167.
958. K. Uvnäs-Moberg et al., *Acta Physiol. Scand.* **124** (1985) 391.
959. M. S. Carmichael et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 27.
960. P. Chang, K. Kimura, *Jpn. J. Pharmacol.* **43** (1987) Suppl., 184P.
961. A. E. Johnson et al., *Neuroendocrinology* **50** (1989) 199.
962. M. Shumacher et al., *Science (Washington D.C.)* **250** (1990) 691.
963. S. Chin Mah, K. G. Hofbauer, *Drugs Future* **12** (1987) 1055.
964. E. A. Woodcock et al., *Endocrinology (Baltimore)* **118** (1986) 2432.
965. L. Share, J. T. Crofton, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **43** (1984) 103.
966. M. J. S. Miller et al., *Clin. Exp. Pharmacol. Physiol.* **13** (1986) 577.
967. J. Maslyszko et al., *Thromb. Haemostasis* **61** (1989) 537.
968. P. A. Flordal et al., *Thromb. Haemostasis* **61** (1989) 541.
969. F. A. Laszlo et al., *Acta Endocrinol. (Copenhagen)* **106** (1984) 56.
970. L. Hary et al., *Fundam. Clin. Pharmacol.* **3** (1989) 452.
971. J. C. Buckingham, *J. Endocrinol.* **113** (1987) 389.
972. P. Mormede et al., *Regul. Pept.* **12** (1985) 175.
973. J. J. Legros et al., *Prog. Neuropsychopharmcol. Biol. Psychiatry (Oxford)* **12** (1988) Suppl., S71.
974. J. McMurray, A. D. Struthers, *Br. Heart J.* **59** (1988) 627.
975. L. M. Graczak et al., *Am. J. Physiol.* **256** (1989) H925.
976. T. Wilson et al., *Prostaglandins* **35** (1988) 771.
977. D. P. Brooks et al., *Neuroendocrinology* **38** (1984) 382.
978. S. Stock et al., *Regul. Pept.* **30** (1990) 1.
979. N. Altszuler et al., *Proc. Soc. Exp. Biol. Med.* **182** (1986) 79.
980. A. N. Margioris et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 922.
981. A. Buku et al., *J. Med. Chem.* **30** (1987) 1509.
982. R. Brown, M. G. King, *Peptides (Fayetteville, N.Y.)* **5** (1984) 1135.
983. S. Chin Mah, K. G. Hofbauer, *Drugs Future* **12** (1987) 1055.
984. M. R. Melis et al., *Pharmacol. Res. Commun.* **20** (1988) 1117.
985. P. Rekowski et al., *Acta Chem. Scand. Ser. B* **B39** (1985) 453.
986. Z. Grzonka et al., *J. Med. Chem.* **29** (1986) 96.
987. A. Buku et al., *Int. J. Peptide Res.* **23** (1984) 551.
988. Z. Prochazka et al., *Collect. Czech. Chem. Commun.* **53** (1988) 2604.
989. B. Lammek et al., *Experientia* **43** (1987) 1211.
990. Z. Prochazka et al., *Collect. Czech. Chem. Commun.* **49** (1984) 642.
991. M. Lebl et al., *Collect. Czech. Chem. Commun.* **50** (1985) 418.
992. G. Sweeney et al., *Curr. Ther. Res.* **47** (1990) 528.
993. Z. Grzonka et al., *J. Med. Chem.* **26** (1983) 1786.
994. *Drugs Future* **9** (1984) 830.
995. M. Lebl et al., *Collect. Czech. Chem. Commun.* **50** (1985) 132.

996. R. A. Johns, *Anesthesiology* **71** (1989) A487.
997. B. Lammek et al., *J. Med. Chem.* **32** (1989) 244.
998. L. B. Kinter et al., *Am. J. Physiol.* **254** (1988) F165.
999. W. Y. Chan et al., *Proc. Soc. Exp. Biol. Med.* **185** (1987) 187.
1000. M. Manning et al., *J. Med. Chem.* **32** (1989) 382.
1001. D. W. Hahn et al., *Am. J. Obstet. Gynecol.* **157** (1987) 977.
1002. P. Rekowski et al., *Pol. J. Pharmacol. Pharm.* **39** (1987) 303.
1003. P. Melin et al., *J. Endocr.* **111** (1986) 125.
1004. *Drugs Future* **13** (1988) 25.
1005. B. Lammek et al., *Pol. J. Pharmacol. Pharm.* **40** (1988) 423.
1006. B. Lammek et al., *J. Med. Chem.* **31** (1988) 603.
1007. M. Manning et al., *J. Med. Chem.* **27** (1984) 423.
1008. M. Manning et al., *J. Med. Chem.* **26** (1983) 1607.
1009. B. Lammek et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 1109.
1010. M. Manning et al., *Nature (London)* **308** (1984) 652.
1011. F. El-Fehail Ali et al., *J. Med. Chem.* **30** (1987) 2291.
1012. W. F. Huffman et al., *J. Med. Chem.* **28** (1985) 1759.
1013. M. L. Moore et al., *J. Med. Chem.* **31** (1988) 1487.
1014. M. Manning et al., *J. Med. Chem.* **30** (1987) 2245.
1015. W. H. Sawyer et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 157.
1016. M. Manning et al., *Nature (London)* **329** (1987) 839.
1017. F. Kasprzykowski et al., *Collect. Czech. Chem. Commun.* **53** (1988).
1018. D. J. Pettibone et al., *Endocrinology (Baltimore)* **125** (1989) 217.
1019. R. M. Freidinger et al., *J. Med. Chem.* **33** (1990) 1845.
1020. M. G. Bock et al., *J. Med. Chem.* **33** (1990) 2321.
1021. P. Mannucci et al., *Ann. N.Y. Acad. Sci.* **509** (1987) 71.
1022. P. R. Martens, *Lancet I* (1989) 1322.
1023. B. Waeber et al., *J. Cardiovasc. Pharmacol.* **8** (1986) Suppl. 7, S111.
1024. M. Akerlund et al., *Br. J. Obstet. Gynaecol.* **94** (1987) 1040.
1025. J. E. Zadina et al., *Brain Res.* **409** (1987) 10.
1026. G. Campbell Teskey, M. Kavaliers, *Peptides (Fayetteville. N.Y.)* **6** (1985) 165.
1027. H. N. Bhargava, *Neuropharmacology* **25** (1986) 737.
1028. G. Szabo et al., *Acta Physiol. Hung.* **69** (1987) 115.
1029. G. L. Kovacs et al., *Pharmacol. Biochem. Behav.* **31** (1988) 833.
1030. R. L. Johnson et al., *J. Med. Chem.* **33** (1990) 1828.
1031. L. Pulvirenti et al., A. J. Kastin, *Eur. J. Pharmacol.* **151** (1988) 289.
1032. R. L. Johnson et al., *J. Med. Chem.* **29** (1986) 2100.
1033. K.-L. Yu et al., *J. Med. Chem.* **31** (1988) 1430.
1034. R. K. Mishra et al., *Drugs Future* **11** (1986) 303.
1035. M. Yanagisawa et al., T. Masaki, *Biochem. Pharmacol.* **38** (1989) 1877.
1036. T. Kosaka et al., *FEBS Lett.* **249** (1989) 42.
1037. T. Yoshizawa et al., *Science (Washington D.C.)* **247** (1990) 462.
1038. H. C. Lam et al., *FEBS Lett.* **261** (1990) 184.
1039. A. Inoue et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 2863.
1040. M. Yanagisawa et al., *Nature (London)* **332** (1988) 411.
1041. Y. Kloog et al., *Science (Washington D.C.)* **242** (1988) 268.
1042. S. Ohkubo et al., *FEBS Lett.* **274** (1990) 136.
1043. H. Onda et al., *FEBS Lett.* **261** (1990) 327.
1044. M. Yoshizumi et al., *Biochem. Biophys. Res. Commun.* **161** (1989) 859.
1045. T. Emori et al., *Biochem. Biophys. Res. Commun.* **160** (1989) 93.
1046. K. Ohta et al., *Biochem. Biophys. Res. Commun.* **169** (1990) 578.
1047. C. Boulanger, T. F. Lüscher, *J. Clin. Invest.* **85** (1990) 587.
1048. H. Rakugi et al., *Biochem. Biophys. Res. Commun.* **169** (1990) 973.
1049. N. Suzuki et al., *Biochem. Biophys. Res. Commun.* **167** (1990) 941.
1050. Y. Saito et al., *N. Engl. J. Med.* **322** (1990) 205.
1051. P. Cernacek, D. J. Stewart, *Biochem. Biophys. Res. Commun.* **161** (1989) 562.
1052. Y. Masuda et al., *FEBS Lett.* **257** (1989) 208.
1053. Y. Hirata et al., *FEBS Lett.* **239** (1988) 13.
1054. P. Roubert et al., *Biochem. Biophys. Res. Commun.* **164** (1989) 809.
1055. N. Fukuda et al., *Biochem. Biophys. Res. Commun.* **167** (1990) 739.

1056. R. K. Minkes et al., *Eur. J. Pharmacol.* **164** (1989) 571.
1057. R. M. Eglen et al., *Br. J. Pharmacol.* **97** (1989) 1297. Y. Matsumura et al., *Life Sci.* **44** (1989) 149.
1058. G. Mazzocchi et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 763.
1059. J. R. Hu et al., *Eur. J. Pharmacol.* **158** (1988) 177.
1060. Y. Fukuda et al., *Biochem. Biophys. Res. Commun.* **155** (1988) 167.
1061. S. S. Stojilkovic et al., *Science (Washington D.C.)* **248** (1990) 1633.
1062. Y. Shibouta et al., *Life Sci.* **46** (1990) 1611.
1063. R. K. Minkes et al., *Am. J. Physiol.* **259** (1990) H1152.
1064. T. Kashiwabara et al., *FEBS Lett.* **247** (1989) 73.
1065. K. Nakajima et al., *Biochem. Biophys. Res. Commun.* **163** (1989) 424.
1066. A. Locei et al., *Neuropeptides (Edinburgh)* **16** (1990) 21.
1067. M. D. Randall et al., *Br. J. Pharmacol.* **98** (1989) 685.
1068. J. Gutowska, M. Nemer, *Endocr. Rev.* **10** (1989) 519.
1069. G. Thibault et al., *Biochem. J.* **241** (1987) 265.
1070. P. Needleman, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **45** (1986) 2096.
1071. A. Miyata et al., *Biochem. Biophys. Res. Commun.* **142** (1987) 461.
1072. H. Ioth et al., *Biochem. Biophys. Res. Commun.* **143** (1987) 560.
1073. R. Eskay et al., *Science (Washington D.C.)* **232** (1986) 636.
1074. L. Meleagros et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 545.
1075. C. J. Winters et al., *Biochem. Biophys. Res. Commun.* **150** (1988) 231.
1076. G. P. Vlasuk et al., *Biochem. Biophys. Res. Commun.* **136** (1986) 396. R. Takayanagi et al., *Biochem. Biophys. Res. Commun.* **142** (1987) 483.
1077. N. Zamir et al., *Brain Res.* **365** (1986) 105.
1078. J. Gutkowska et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 461. S. H. Kim et al., *Life Sci.* **45** (1989) 1581. A. M. Vollmar et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 965.
1079. O. Vuolteenaho et al., *FEBS Lett.* **233** (1988) 79.
1080. A. Miyata et al., *Biochem. Biophys. Res. Commun.* **155** (1988) 1330.
1081. Y. Takei et al., *Biochem. Biophys. Res. Commun.* **164** (1989) 537.
1082. J.-I. Sakata et al., *Biochem. Biophys. Res. Commun.* **155** (1988) 1338.
1083. T. Sudoh et al., *Nature (London)* **332** (1988) 78.
1084. K. Maekawa et al., *Biochem. Biophys. Res. Commun.* **157** (1988) 410.
1085. T. Sudoh et al., *Biochem. Biophys. Res. Commun.* **168** (1990) 863.
1086. H. Ruskoaho et al., *Biochem. Biophys. Res. Commun.* **139** (1986) 266.
1087. K. Naruse et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 285.
1088. Y. Shenker et al., *J. Cardiovasc. Pharmacol.* **8** (1986) 1300.
1089. D. Lachance et al., *Hypertension (Dallas)* **9** (1987) 524.
1090. J. A. Sanfield et al., *Am. J. Physiol.* **252** (1987) Part 1, E740.
1091. M. G. Currie, W. H. Newman, *Biochem. Biophys. Res. Commun.* **137** (1986) 94.
1092. B. Baranowska et al., *Biochem. Biophys. Res. Commun.* **143** (1987) 159.
1093. B. Baranowska et al., *Hypertension (Dallas)* **9** (1987) 545.
1094. J. R. Hu et al., *Eur. J. Pharmacol.* **158** (1988) 177.
1095. O. Sergev et al., *J. Endocrinol. Invest.* **13** (1990) 649.
1096. M. Weissel et al., *Klin. Wochenschr.* **64** (1986) Suppl. VI, 93.
1097. K. Horky et al., *Biochem. Biophys. Res. Commun.* **129** (1985) 651.
1098. H. Matsubara et al., *Biochem. Biophys. Res. Commun.* **145** (1987) 336.
1099. G. Wambach et al., *Klin. Wochenschr.* **64** (1986) Suppl. VI, 53.
1100. K. M. Verburg et al., *Am. J. Physiol.* **251** (1986) R947.
1101. W. Rascher et al., *Horm. Res.* **28** (1987) 58.
1102. D. R. J. Singer et al., *Lancet II* (1987) 1394.
1103. M. R. Arendt et al., *Klin. Wochenschr.* **64** (1986) Suppl. VI, 97.
1104. T. Yamaji et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 348.
1105. T. Yamaji et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 815.
1106. A. L. Gerbes et al., *N. Engl. J. Med.* **313** (1985) 1609.
1107. M. Kohno et al., *Am. J. Med.* **83** (1987) 648.
1108. M. Ohashi et al., *Horm. Metab. Res.* **20** (1988) 705.
1109. A. E. G. Raine et al., *N. Engl. J. Med.* **315** (1986) 533.
1110. Y. Shenker et al., *Life Sci.* **41** (1987) 1635.

1111. N. Zamir et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 541.
1112. R. S. Zimmerman et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 353.
1113. R. Garcia et al., *Biochem. Biophys. Res. Commun.* **136** (1986) 510.
1114. M. Sakamoto et al., *Biochem. Biophys. Res. Commun.* **135** (1986) 515.
1115. J. L. Sonnenberg et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 173.
1116. A. Jardine et al., *J. Am. Coll. Cardiol.* **13** (1989) Suppl. A, 76A.
1117. D. B. Northridge et al., *J. Am. Coll. Cardiol.* **13** (1989) Suppl. A, 76A.
1118. M.-S. Chang et al., *Nature (London)* **341** (1989) 68.
1119. M. B. Anand-Srivastava et al., *J. Biol. Chem.* **265** (1990) 8566.
1120. D. L. Vesely et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 193.
1121. K. L. Lanier-Smith, M. G. Currie, *Eur. J. Pharmacol.* **178** (1990) 105.
1122. J. M. Saavedra et al., *Endocrinology (Baltimore)* **120** (1987) 426.
1123. E. L. Schiffrin et al., *Circulation* **72** (1985) part 2, III–293.
1124. Y. Hirata et al., *Biochem. Biophys. Res. Commun.* **138** (1986) 405.
1125. A. G. Jardine et al., *Klin. Wochenschr.* **67** (1989) 902.
1126. N. C. Trippodo, R. W. Barbee, *Am. J. Physiol.* **252** (1987) R915.
1127. P J. S. Chiu et al., *Eur. J. Pharmacol.* **124** (1986) 277.
1128. B. L. Pegram et al., *Fed. Proc. Am. Soc. Exp. Biol.* **45** (1986) 2382.
1129. R. J. Schiebinger, J. Linden, *Am. J. Physiol.* **251** (1986) H1095.
1130. A. Petrillo et al., *Gastroenterology* **92** (1987) part 2, 1764.
1131. J. McMurray et al., *Eur. J. Clin. Pharmacol.* **35** (1988) 409.
1132. K. Schafferhans et al., *Klin. Wochenschr.* **66** (1988) Suppl. 13, 223.
1133. A. H. Morice et al., *Am. Rev. Respir. Dis.* **135** (1987) Part 2, A300.
1134. S. Adnot et al., *Am. Rev. Respir. Dis.* **137** (1988) Part 2, 108.
1135. M. B. Anand-Srivastava, M. Cantin, *Biochem. Biophys. Res. Commun.* **138** (1986) 427.
1136. K. Iitake et al., *Endocrinology (Baltimore)* **119** (1986) 438.
1137. T. Takao et al., *Life Sci.* **42** (1988) 1199.
1138. M. A. Dillingham, R. J. Anderson, *Science (Washington D.C.)* **231** (1986) 1572.
1139. Z. Zukowska-Grojec et al., *J. Pharmacol. Exp. Ther.* **239** (1986) 480.
1140. P. Nambi et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 8492.
1141. J. P. Briggs, J. Schnermann, *J. Cardiovasc. Pharmacol.* **8** (1986) 1296.
1142. H. Vierhapper et al., *Hypertension (Dallas)* **8** (1986) 1040.
1143. F. Bex, A. Corbin, *Eur. J. Pharmacol.* **115** (1985) 125.
1144. P. R. Bovy, *Med. Res. Rev.* **10** (1990) 115.
1145. T. Kimura, S. Sakakibara, *Jpn. J. Pharmacol.* **40** (1986) Suppl., 77P.
1146. J. G. De Mey et al., *J. Pharmacol. Exp. Ther.* **240** (1987) 937.
1147. H. Saxenhofer et al., *Am. J. Physiol.* **259** (1990) F832.
1148. Y. Kiso et al., *J. Pharm. Sci.* **76** (1987) S172.
1149. J. Rittenhouse et al., *J. Biol. Chem.* **261** (1986) 7607.
1150. C. F. Hassman et al., *Biochem. Biophys. Res. Commun.* **152** (1988) 1070.
1151. P. W. Schiller et al., *Biochem. Biophys. Res. Commun.* **143** (1987) 499.
1152. D. L. Song et al., *Eur. J. Pharmacol.* **160** (1989) 141.
1153. J. D. Mogannam et al., *J. Cardiovasc. Pharmacol.* **8** (1986) 1317.
1154. R. Deghenghi et al., *Drug. Dev. Res. (New York)* **15** (1988) 87.
1155. J. Knolle et al., lecture held at Hoechst AG 1990.
1156. R. F. Nutt et al., *Abstr. Pap. Am. Chem. Soc.* **195** (1988) BIOL1.
1157. M. S. King, A. J. Baertschi, *Endocrinology (Baltimore)* **124** (1989) 286.
1158. D. L. Vesely et al., *Biochem. Biophys. Res. Commun.* **148** (1987) 1540.
1159. D. R. Martin et al., *Am. J. Physiol.* **258** (1990) F1401.
1160. Y. Kitajima et al., *Biochem. Biophys. Res. Commun.* **164** (1989) 1295.
1161. R. Corder et al., *Biochem. J.* **219** (1984) 699.
1162. K. Tatemoto et al., *Biochem. Biophys. Res. Commun.* **157** (1988) 713.
1163. W. B. Rhoten, *Experientia* **43** (1987) 428.
1164. A. B. Leiter et al., *J. Biol. Chem.* **259** (1984) 14702.
1165. D. M. McKay et al., *Regul. Pept.* **31** (1990) 187.
1166. J. B. Jaspan et al., *Am. J. Physiol.* **251** (1986) E215.
1167. S. J. Konturek et al., *Gastroenterology* **86** (1984) Part 2, 1140.

1168. R. D. Beauchman et al., *Gastroenterology* **90** (1986) Part 2, 1341.
1169. G. R. Greenberg et al., *Dig. Dis. Sci.* **30** (1985) 946.
1170. L. R. Gingerich et al., *Metab. Clin. Exp.* **34** (1985) 25.
1171. E. Hidvegi et al., *Acta Physiol. Hung.* **63** (1984) 237.
1172. A. J. L. de Jong et al., *Regul. Pept.* **17** (1987) 285.
1173. C. Beglinger et al., *Am. J. Physiol.* **246** (1984) G286.
1174. S. Ito et al., *J. Pharmacol.* **33** (1983) Suppl., 129P.
1175. R. L. Conter et al., *Clin. Res.* **32** (1984) 24A.
1176. W. H. Coltharp et al., *Gastroenterology* **90** (1986) Part 2, 1379.
1177. S. J. Konturek et al., *Am. J. Physiol.* **248** (1985) G687.
1178. T. S. Gray, J. E. Morley, *Life Sci.* **38** (1986) 389.
1179. P. Ciofi et al., *Neuroendocrinology* **45** (1987) 425.
1180. E. Ekblad et al., *Regul. Pept.* **10** (1984) 47.
1181. C. D. Minth et al., *J. Biol. Chem.* **261** (1986) 11974.
1182. J. Allen et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 2532.
1183. J. M. Allen et al., *Neuropeptides (Edinburgh)* **6** (1985) 95.
1184. K. Kitamura et al., *Biochem. Biophys. Res. Commun.* **169** (1990) 1164.
1185. M. Haass et al., *Am. J. Physiol.* **259** (1990) R925.
1186. J.-T. Cheng, C. L. Shen, *Eur. J. Pharmacol.* **123** (1986) 303.
1187. K. Takahashi et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 433.
1188. A. Rudehill et al., *Acta Physiol. Scand.* **131** (1987) 517.
1189. V. Castagne et al., *Regul. Pept.* **19** (1987) 55.
1190. L. S. Brady et al., *Neuroendocrinology* **52** (1990) 441.
1191. T. Ogawa et al., *Biochem. Biophys. Res. Commun.* **165** (1989) 1399.
1192. A. Balasubramaniam, S. Sheriff, *J. Biol. Chem.* **265** (1990) 14724.
1193. J. M. Lundberg et al., *Regul Pept.* **13** (1986) 169.
1194. J. D. Watson et al., *J. Endocrinol.* **116** (1988) 421.
1195. N. Pages et al., *Neuropeptides (Edinburgh)* **17** (1990) 141.
1196. A. Sahu et al., *Endocrinology (Baltimore)* **122** (1988) 2199.
1197. S. P. Sheikh et al., *FEBS Lett.* **245** (1989) 209.
1198. S. P. Sheikh, J. A. Williams, *J. Biol. Chem.* **265** (1990) 8304.
1199. N. A. Scott et al., *Am. J. Physiol.* **259** (1990) H174.
1200. L. Edvinsson et al., *Trends Pharmacol. Sci. TIPS* **8** (1987) 231.
1201. J. Fuhlendorff et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 182.
1202. S.-I. Mihara et al., *FEBS Lett.* **259** (1989) 79.
1203. B. B. Fredholm et al., *Acta Physiol. Scand.* **124** (1985) 467.
1204. T. Ishiguro et al., *Chem. Pharm. Bull.* **36** (1988) 2720.
1205. J. L. Morris, *Peptides (Fayetteville, N.Y.)* **11** (1990) 381.
1206. J. F. Aubert et al., *J. Pharmacol. Exp. Ther.* **244** (1988) 1109.
1207. G. H. Greeley, Jr. et al., *Am. J. Physiol.* **254** (1988) E513.
1208. Y.-S. Guo et al., *Am. J. Physiol.* **253** (1987) G298.
1209. J. F. Rodriguez-Sierra et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 539.
1210. W. B. Wehrenberg et al., *Neuroendocrinology* **49** (1989) 680.
1211. W. R. Crowley et al., *Endocrinology (Baltimore)* **120** (1987) 941.
1212. H. E. Albers et al., *Am. J. Physiol.* **258** (1990) R376.
1213. B. G. Stanley et al., *Peptides (Fayetteville, N.Y.)* **6** (1985) 1205.
1214. T. N. Pappas et al., *Gastroenterology* **90** (1986) Part 2, 1578.
1215. H. Härfstrand, *Acta Physiol. Scand.* **128** (1986) 121.
1216. J. Esteban et al., *Life Sci.* **45** (1989) 2395.
1217. J. L. Krstenansky et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 4377.
1218. A. Beck et al., *FEBS Lett.* **244** (1989) 119.
1219. A. G. Beck-Sickinger et al., *Eur. J. Biochem.* **194** (1990) 449.
1220. J. L. Krstenansky et al., *Neuropeptides (Edinburgh)* **17** (1990) 117.
1221. B. Baranowska et al., *Biochem. Biophys. Res. Commun.* **145** (1987) 680.
1222. L. Grundemar, R. Hakanson, *Br. J. Pharmacol.* **100** (1990) 190.
1223. G. W. Aponte et al., *Am. J. Physiol.* **249** (1985) G745.
1224. T. E. Adrian et al., *Gastroenterology* **92** (1987) Part 2, 1289.
1225. T. E. Adrian et al., *Regul. Pept.* **9** (1984) 320.

1226. F. Lluis et al., *Gastroenterology* **90** (1986) Part 2, 1525.
1227. T. N. Pappas et al., *Am. J. Physiol.* **248** (1985) G118.
1228. T. N. Pappas et al., *Gastroenterology* **89** (1985) 1387.
1229. W. W. Pawlik et al., *Gastroenterology* **90** (1986) Part 2, 1580.
1230. T. E. Adrian et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 803.
1231. R. J. Playford et al., *Lancet* **335** (1990) 1555.
1232. R. K. Harding, T. J. McDonald, *Peptides (Fayetteville, N.Y.)* **10** (1989) 21.
1233. N. Minamino et al., *Biochem. Biophys. Res. Commun.* **156** (1988) 355.
1234. J. Domin et al., *J. Biol. Chem.* **264** (1989) 20881.
1235. R. Murphy et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 613.
1236. J. Domin et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 779.
1237. C. A. Maggi et al., *Br. J. Pharmacol.* **99** (1990) 186.
1238. M. V. Graf, A. J. Kastin, *Peptides (Fayetteville, N.Y.)* **7** (1986) 1165.
1239. G. A. Schoenenberger, *Eur. Neurol.* **23** (1984) 321.
1240. A. Bjartell et al., *Neuroendocrinology* **50** (1989) 564.
1241. A. Nakamura et al., *Eur. J. Pharmacol.* **121** (1986) 157.
1242. A. M. J. Young, B. J. Key, *Neuropharmacology* **23** (1984) 1347.
1243. J. E. Taylor, *IRCS Med. Sci.* **14** (1986) 1122.
1244. K. V. Sudakov et al., *J. Biol. Sci.* **18** (1983) 1.
1245. M. V. Graf, A. J. Kastin, *Neurosci. Biobehav. Rev.* **8** (1984) 83.
1246. T. Okajima, G. Hertting, *Horm. Metab. Res.* **18** (1986) 497.
1247. M. V. Graf et al., *Neuroendocrinology* **41** (1985) 353.
1248. K. S. Iyer, S. M. McCann, *Neuroendocrinology* **46** (1987) 93.
1249. K. S. Iyer, S. M. McCann, *Neuroendocrinology* **46** (1987) 93.
1250. A. Sahu, S. P. Kalra, *Life Sci.* **40** (1987) 1201.
1251. W. Larbig et al., *Eur. Neurol.* **23** (1984) 372.
1252. K. Iwai et al., *J. Biol. Chem.* **262** (1987) 8956.
1253. S.-I. Fukuoka et al., *Biochem. Biophys. Res. Commun.* **145** (1987) 646.
1254. Y. Takahashi et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 1931.
1255. U. Gubler et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 4307.
1256. R. J. Deschenes et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 726.
1257. V. E. Eysselein et al., *Am. J. Physiol.* **258** (1990) G253.
1258. Z.-Z. Zhou et al., *Peptides* **6** (1985) 337.
1259. J. Shively et al., *Am. J. Physiol.* **252** (1987) G272.
1260. A. Takashima, S. Itoh, *Can. J. Physiol. Pharmacol.* **67** (1989) 228.
1261. T. Wakabayashi et al., *Gene* **31** (1984) 295.
1262. Y. F. Chen et al., *Am. J. Physiol.* **249** (1985) G29.
1263. J. B. J. M. Jansen, C. B. H. W. Lamers, *Life Sci.* **33** (1983) 2197.
1264. J. J. Benoliel et al., *Fundam. Clin. Pharmacol.* **3** (1989) 141.
1265. I. Koop, *Eur. J. Clin. Invest.* **20** (1990) Suppl. 1, S51.
1266. I. C. Forgacs et al., *Clin. Sci.* **66** (1984) 61P.
1267. R. A. Siegel et al., *Neuroendocrinology* **46** (1987) 75.
1268. J. B. M. J. Jansen, C. B. H. W. Lamers, *Gut* **27** (1986) A1260.
1269. I. Nakano et al., *Regul. Pept.* **14** (1986) 229.
1270. S. Kanayama et al., *Life Sci.* **41** (1987) 1915.
1271. A. Schafmayer et al., *Digestion* **30** (1984) 95.
1272. D.-H. Yu et al., *Am. J. Physiol.* **258** (1990) G86.
1273. M. Praissman, M. Walden, *Gastroenterology* **92** (1987) part 2, 1579.
1274. M. Otsuki, J. A. Williams, *Gastroenterology* **84** (1983) part 2, 1266.
1275. Y. K. Leung et al., *Dig. Dis. Sci.* **30** (1985) 981.
1276. M. Otsuki et al., *Gastroenterology* **87** (1984) 882.
1277. B. E. Evans, *Drugs Future* **14** (1989) 971.
1278. L. L. Iversen et al., *J. Endocrinol.* **124** (1990) Suppl., 18.
1279. J. Hughes et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 6728.
1280. B. Charpentier et al., *Peptides* **9** (1988) 835.
1281. B. Charpentier et al., *J. Med. Chem.* **32** (1989) 1148.
1282. I. Maraseigne et al., *J. Med. Chem.* **31** (1988) 966.
1283. R. A. Liddle et al., *J. Clin. Invest.* **84** (1989) 1270.
1284. M. F. O'Rourke et al., *Am. J. Physiol.* **258** (1990) G179.
1285. S. J. Konturek et al., *Am. J. Physiol.* **248** (1985) G687.
1286. J. E. Kellow et al., *Am. J. Physiol.* **253** (1987) G650.
1287. A. Brown et al., *Gut* **26** (1985) A580.

1288. L. Gullo et al., *Dig. Dis. Sci.* **31** (1986) 1345.
1289. L. Marzio et al., *Am. J. Gastroenterol.* **80** (1985) 1.
1290. G. Schmidt et al., *Klin. Wochenschr.* **64** (1986) Suppl. 5, 56.
1291. F. Makovec et al., *Regul. Pept.* **16** (1986) 281.
1292. K. A. Zucker et al., *Am. J. Physiol.* **257** (1989) part 1, G511.
1293. S. Karlsson, B. Ahren, *Acta Physiol. Scand.* **135** (1989) 271.
1294. K. C. K. Lloyd et al., *Gastroenterology* **96** (1989) part 2, A305.
1295. T. W. Vickroy et al., *Eur. J. Pharmacol.* **152** (1988) 371.
1296. C. T. Dourish et al., *Eur. J. Pharmacol.* **176** (1990) 35.
1297. E. K. Hong, A. E. Takemori, *J. Pharmacol. Exp. Ther.* **251** (1989) 594.
1298. M. Rattray et al., *Neuropeptides* **14** (1989) 263.
1299. D. S. K. Magnuson et al., *Neuropeptides* **16** (1990) 213.
1300. J. E. Kellow et al., *Am. J. Physiol.* **252** (1987) G345.
1301. G. Zetler, *Peptides* **5** (1984) 729.
1302. S. Rattan, R. K. Goyal, *Gastroenterology* **90** (1986) 94.
1303. D. Pelaprat, M. Reibaud, *J. Med. Chem.* **32** (1989) 445.
1304. C. Beglinger et al., *Regul Pept.* **8** (1984) 291.
1305. M. Ruiz-Gayo et al., *Peptides* **6** (1985) 415.
1306. M. Knight et al., *Peptides* **6** (1985) 631.
1307. C. Mendre et al., *J. Biol. Chem.* **263** (1988) 10641.
1308. B. Penke et al., *J. Med. Chem.* **27** (1984) 845.
1309. I. Marseigne et al., *Int. J. Peptide Protein Res.* **33** (1989) 230.
1310. M.-C. Galas et al., *Am. J. Physiol.* **254** (1988) G176.
1311. J. D. Gardner et al., *Am. J. Physiol.* **248** (1985) G98.
1312. *Drugs Future* **13** (1988) 34.
1313. M.-F. Lignon et al., *J. Biol. Chem.* **262** (1987) 7226.
1314. B. E. Evans, *Drugs Future* **14** (1989) 971.
1315. R. T. Jensen et al., *Biochim. Biophys. Acta* **761** (1983) 269.
1316. R. Tritapepe, C. di Padova, *Br. Med. J.* **293** (1986) 1102.
1317. N. Basso et al., *Gastroenterology* **89** (1985) 605.
1318. A. Pap, V. Varro, *Digesetion* **30** (1984) 118.
1319. B. S. Barbaz et al., *Neuropharmacology* **25** (1986) 823.
1320. E. R. Spindel et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 19.
1321. D. Barra et al., *FEBS Lett.* **182** (1985) 53.
1322. J. Price et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 1097.
1323. G. D. Jahnke, L. H. Lazarus, *Proc. Natl. Acad. Sci. USA* **81** (1984) 578.
1324. C. Scarpignato, B. Micali, *Gut* **27** (1986) 499.
1325. F. Porecca et al., *Eur. J. Pharmacol.* **114** (1985) 167.
1326. T. von Schrenck et al., *Am. J. Physiol.* **259** (1990) G468.
1327. C. B. H. W. Lamers et al., *Dig. Dis. Sci.* **31** (1986) Suppl., 428S.
1328. A. E. Pontiroli et al., *Hormone Res.* **23** (1986) 129.
1329. A. J. L. de Jong et al., *Regul. Pept.* **17** (1987) 285.
1330. C. E. Sievert, Jr. et al., *Am. J. Physiol.* **254** (1988) G361.
1331. T. Lehy et al., *Gastroenterology* **90** (1986) 1942.
1332. C. Stock Damge et al., *Gut* **28** (1987) Suppl., 1.
1333. F. Cuttitta et al., *Nature (London)* **316** (1985) 823.
1334. G. de Caro et al., *Peptides* **5** (1984) 607.
1335. E. Widerlöv et al., *Peptides* **5** (1984) 523.
1336. F. Girard et al., *Eur. J. Pharmacol.* **102** (1984) 489.
1337. P. Heinz-Erian et al., *Am. J. Physiol.* **252** (1987) G439.
1338. Z. A. Saeed et al., *Peptides* **10** (1989) 597.
1339. R. Camble et al., *Life Sci.* **45** (1989) 1521.
1340. L.-H. Wang et al., *Biochemistry* **29** (1990) 616.
1341. D. C. Heimbrook et al., *J. Biol. Chem.* **264** (1989) 11258.
1342. T. von Schrenck et al., *Am. J. Physiol.* **259** (1990) G468.
1343. D. H. Coy et al., *J. Biol. Chem.* **263** (1988) 5056.
1344. D. H. Coy et al., *J. Biol. Chem.* **264** (1989) 14691.
1345. G. Bonora et al., *Gastroenterology* **84** (1983) Part 2, 1111.
1346. C. A. Helman, B. I. Hirschowitz, *Gastroenterology* **92** (1987) 1926.
1347. J. C. M. Hafkenscheid et al., *Clin. Chim. Acta* **136** (1984) 235.
1348. B. N. Andersen et al., *Regul. Pept.* **10** (1985) 329.
1349. R. Dimaline et al., *FEBS Lett.* **205** (1986) 318.

1350. R. J. Nachman et al., *Science (Washington D.C.)* **234** (1986) 71.
1351. M. M. Wolfe et al., *Clin. Res.* **31** (1983) 477.
1352. T. Kurose et al., *Life Sci.* **42** (1988) 1995.
1353. T. Chiba et al., *Horm. Metab. Res.* **15** (1983) 516.
1354. S. F. Alino et al., *Acta Physiol. Scand.* **126** (1986) 1.
1355. K. C. Christensen, *Scand. J. Gastroenterol.* **19** (1984) 339.
1356. J. DelValle, T. Yamada, *J. Clin. Invest.* **85** (1990) 139.
1357. E. J. Dial et al., *Gastroenterology* **90** (1986) 1018.
1358. A. Rentis et al., *Horm. Metabol. Res.* **18** (1986) 423.
1359. B. I. Hirschowitz, E. Molina, *Peptides* **5** (1984) 35.
1360. C. H. Cho et al., *IRCS Med. Sci.* **13** (1985) 629.
1361. H. J. de Aizpurua et al., *N. Engl. J. Med.* **313** (1985) 479.
1362. M. K. Müller et al., *Horm. Metabol. Res.* **18** (1986) 675.
1363. J. Puurunen, U. Schwabe, *Br. J. Pharmacol.* **90** (1987) 479.
1364. L. Holm Rutili, T. Berglindh, *Am. J. Physiol.* **250** (1986) part 1, G575.
1365. A. J. Blair III et al., *J. Clin. Invest.* **78** (1986) 779.
1366. R. Hakanson et al., *Scand. J. Gastroenterol.* **21** (1986) Suppl. 118, 18.
1367. T. E. Solomon et al., *Gastroenterology* **92** (1987) 429.
1368. A. Imdahl et al., *J. Cancer Res. Clin. Oncol.* **115** (1989) 388.
1369. C. J. Kusyk et al., *Am. J. Physiol.* **251** (1986) G597.
1370. D. L. Morris et al., *Gut* **30** (1989) A739.
1371. C. de Montigny, *Arch. Gen. Psychiatry* **46** (1989) 511.
1372. G. Katsuura, S. Itoh, *Peptides* **7** (1986) 809.
1373. V. E. Eysselein et al., *J. Clin. Invest.* **73** (1984) 1284.
1374. W. Göhring et al., *Hoppe-Seyler's Z. Physiol. Chem.* **365** (1984) 83.
1375. J. Martinez et al., *J. Med. Chem.* **28** (1985) 1874.
1376. N. V. Sadovnikova et al., *Khim. Farm. Zh.* **21** (1987) 1424.
1377. D. G. Patel et al., *Diabetes* **38** (1989) Suppl. 2, 206A.
1378. J. Martinez et al., *J. Med. Chem.* **27** (1984) 1597.
1379. M. Rodriguez et al., *J. Med. Chem.* **32** (1989) 522.
1380. T. M. Chang et al., *Peptides (Fayetteville, N.Y.)* **6** (1985) 193.
1381. G. Gafvelin et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 6781.
1382. A. S. Kopin et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 2299. D. Gossen et al., *Biochem. Biophys. Res. Commun.* **160** (1989) 862. Y. Shinomura et al., *Life Sci.* **41** (1987) 1243. D. Gossen et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 123.
1383. T. Yamamoto et al., *Chem. Pharm. Bull.* **34** (1986) 3803.
1384. J. Leonora et al., *Am. J. Physiol.* **252** (1987) E477.
1385. G. E. Pratt et al., *Biochem. Biophys. Res. Commun.* **163** (1989) 1243.
1386. A. Argiolas, J. J. Pisano, *J. Biol. Chem.* **260** (1985) 1437.
1387. P. Li et al., *J. Clin. Invest.* **86** (1990) 1474.
1388. S. N. S. Murthy, A. Lavy, *Biochem. Pharmacol.* **37** (1988) 1027.
1389. S. Iwakawa et al., *J. Pharmacobio–Dyn.* **8** (1985) S136.
1390. M. V. Singer et al., *Am. J. Physiol.* **248** (1985) G532.
1391. D. R. Fletcher et al., *Regul. Pept.* **11** (1985) 217.
1392. L. Bolondi et al., *Dig. Dis. Sci.* **29** (1984) 802.
1393. G. E. Feurle et al., *Pancreas* **2** (1987) 422.
1394. H. Haarstad et al., *Scand. J. Gastroenterol.* **21** (1986) 589.
1395. I. S. Chung et al., *Clin. Res.* **36** (1988) 394A.
1396. M. Murakami et al., *Dig. Dis. Sci.* **30** (1985) 346.
1397. F. Yamagishi et al., *Eur. J. Pharmacol.* **118** (1985) 203.
1398. T. Hashimoto et al., *Can. J. Physiol. Pharmacol. Suppl.* 1986, 40.
1399. J. H. Kleibeuker et al., *Gastroenterology* **94** (1988) 122.
1400. D. L. Kaminski, Y. G. Deshpande, *Gastroenterology* **85** (1983) 1239.
1401. H. Olson et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 301.
1402. S. Kiyama et al., *Chem. Pharm. Bull.* **33** (1985) 3205.
1403. T. Tsuda et al., *J. Pharm. Sci.* **79** (1990) 53.
1404. J. Glaser et al., *Scand. J. Gastroenterol.* **24** (1989) 179.
1405. B. Krakamp et al., *MMW Münch. Med. Wochenschr.* **130** (1988) 47.
1406. Y. S. Malov, *Ter. Arkh.* **57** (1985) 6.

1407. B. Glaser et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 1138.
1408. S. I. Said, *J. Endocrinol. Invest.* **9** (1986) 191.
1409. G. Gafvelin et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 469.
1410. G. Gafvelin, *Peptides (Fayetteville, N.Y.)* **11** (1990) 703.
1411. K. Tatemoto et al., *FEBS Lett.* **174** (1984) 258.
1412. M. Nishizawa et al., *FEBS Lett.* **183** (1985) 55.
1413. D. Gossen et al., *Peptides (Fayetteville, N.Y.)* **11** (1990) 123.
1414. S. Björck et al., *Acta Physiol. Scand.* **128** (1986) 639.
1415. M. A. Arnaout et al., *Endocrinology (Baltimore)* **119** (1986) 2052.
1416. D. E. Burleigh, J. B. Furness, *Neuropeptides (Edinburgh)* **16** (1990) 77.
1417. M. Eriksson et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 411.
1418. P. K. Opstad, *Peptides (Fayetteville, N.Y.)* **8** (1987) 175.
1419. T. A. Crozier et al., *Horm. Metab. Res.* **20** (1988) 352.
1420. E. Giladi et al., *Isr. J. Med. Sci.* **23** (1987) 924.
1421. G. Burns et al., *Life Sci.* **40** (1987) 951.
1422. T. V. Nowak et al., *Gastroenterology* **96** (1989) part 2, A368.
1423. K. S. L. Lam et al., *Neuroendocrinology* **52** (1990) 417.
1424. P. Heinz-Erian et al., *Science (Washington D.C.)* **229** (1985) 1407.
1425. F. E. Bauer et al., *MMW Münch. Med. Wochenschr.* **126** (1984) 1097.
1426. L. Woie et al., *Gen. Pharmacol.* **18** (1987) 577.
1427. M. S. O'Dorisio, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **46** (1987) 192.
1428. K. McArthur et al., *Gastroenterology* **92** (1987) part 2, 1524.
1429. M. Murakami et al., *Gastroenterology* **92** (1987) part 2, 1544.
1430. P. Wiik et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 181.
1431. H. Houchi et al., *Biochem. Pharmacol.* **36** (1987) 1551.
1432. M. Raiteri et al., *Eur. J. Pharmacol.* **133** (1987) 127.
1433. M. S. Barnette, B. Weiss, *J. Neurochem.* **45** (1985) 640.
1434. Y. Karasawa et al., *Eur. J. Pharmacol.* **187** (1990) 9.
1435. S. I. Said, *N. Eng. J. Med.* **320** (1989) 1271.
1436. A. Bennett et al., *J. Pharm. Pharmacol.* **36** (1984) 787.
1437. J. Fontaine et al., *Br. J. Pharmacol.* **89** (1986) 599.
1438. N. Yanaihara, EP 0 225 639, 1985.
1439. D. W. Pincus et al., *Nature (London)* **343** (1990) 564.
1440. B. Carlsson et al., *Acta Physiol. Scand.* **129** (1987) 437.
1441. H. Abe et al., *Endocrinology (Baltimore)* **116** (1985) 1383.
1442. L. A. Cunningham, M. A. Holzwarth, *Endocrinology (Baltimore)* **122** (1988) 2090.
1443. B. Ottesen et al., *Regul. Pept.* **16** (1986) 299.
1444. F. W. George, S. R. Ojeda, *Proc. Natl. Acad. Sci. USA* **84** (1987) 5803.
1445. C. E. Ahmed et al., *Endocrinology (Baltimore)* **118** (1986) 1682.
1446. B. G. Kasson et al., *Mol. Cell. Endocrinol.* **48** (1986) 21.
1447. M. Hamada et al., *Am. Rev. Respir. Dis.* **137** (1988) part 2, 35.
1448. J. T. Turner et al., *Peptides (Fayetteville, N.Y.)* **7** (1986) 849.
1449. G. F. Musso et al., *Biochemistry* **27** (1988) 8174.
1450. G. Hallden et al., *Regul. Pept.* **16** (1986) 183.
1451. P. Robberecht et al., *Eur. J. Biochem.* **159** (1986) 45.
1452. S. J. Pandol et al., *Am. J. Physiol.* **250** (1986) G553.
1453. D. C. Thompson et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 443.
1454. P. Robberecht et al., *Eur. J. Biochem.* **165** (1987) 243.
1455. A. Miyata et al., *Biochem. Biophys. Res. Commun.* **164** (1989) 567.
1456. A. Miyata et al., *Biochem. Biophys. Res. Commun.* **170** (1990) 643.
1457. C. Kimura et al., *Biochem. Biophys. Res. Commun.* **166** (1990) 81.
1458. K. Ogi et al., *Biochem. Biophys. Res. Commun.* **173** (1990) 1271.
1459. H.-C. Lam et al., *Eur. J. Biochem.* **193** (1990) 725.
1460. G. Heinrich et al., *Endocrinology (Baltimore)* **115** (1984) 2176.
1461. G. I. Bell et al., *Nature (London)* **304** (1983) 368.
1462. L. C. Lopez et al., *Proc. Natl. Acad. Sci. USA* **80** (1983) 5485.
1463. T. Buhl et al., *J. Biol. Chem.* **263** (1988) 8621.
1464. L. Thim, A. J. Moody, *Regul. Pept.* **2** (1981) 139.

1465. C. Orskov et al., *J. Biol. Chem.* **264** (1989) 12826.
1466. D. S. Weigle, C. J. Goodner, *Endocrinology (Baltimore)* **118** (1986) 1606.
1467. G. Skoglund et al., *Eur. J. Pharmacol.* **143** (1987) 83.
1468. G. Paolisso et al., *Acta Endocrinol. (Copenhagen)* **115** (1987) 161.
1469. L. J. Klaff, G. J. Taborsky, Jr., *Clin. Res.* **34** (1986) 60A.
1470. D. L. Kaminski et al., *Am. J. Physiol.* **254** (1988) G864.
1471. D. G. Pipeleers et al., *Endocrinology (Baltimore)* **117** (1985) 817.
1472. A. Ohneda et al., *Diabetologia* **29** (1986) 397.
1473. C. G. Ostenson et al., *Diabetologia* **29** (1986) 861.
1474. A. A. R. Starke et al., *Diabetes* **33** (1984) 277.
1475. K. F. Hanssen et al., *Diabetologia* **27** (1984) 285A.
1476. A. E. Farah, *Pharmacol. Rev.* **35** (1983) 181.
1477. A. Akatsuka et al., *J. Biol. Chem.* **260** (1985) 3239.
1478. K. Yamauchi, K. Hashizume, *Endocrinology (Baltimore)* **119** (1986) 218.
1479. S. Marubashi et al., *Acta Endocrinol. (Copenhagen)* **108** (1985) 6.
1480. R. Holland et al., *Eur. J. Biochem.* **140** (1984) 325.
1481. R. J. Howland, A. D. Benning, *FEBS Lett.* **208** (1986) 128.
1482. O. G. Bjornsson et al., *Eur. J. Clin. Invest.* **14** (1984) Part 2, 39.
1483. C. Guettet et al., *Biochim. Biophys. Acta L* **1105** (1989) 233.
1484. B. G. Weick, S. Ritter, *Am. J. Physiol.* **250** (1986) R676.
1485. D. L. Carr Locke et al., *Dig. Dis. Sci.* **28** (1983) 312.
1486. D. L. Kaminski et al., *Am. J. Physiol.* **254** (1988) G864.
1487. M. S. Sherman et al., *J. Allergy Clin. Immunol.* **81** (1988) 908.
1488. K. Corey Flanders et al., *J. Biol. Chem.* **259** (1984) 7031.
1489. J. F. Rey et al., *Dig. Dis. Sci.* **31** (1986) 355.
1490. G. Jacobson et al., *Lancet II* (1984) 1149.
1491. A. Mallat et al., *Nature (London)* **325** (1987) 620.
1492. P. Blacke et al., *J. Biol. Chem.* **265** (1990) 21514.
1493. P. Robberecht et al., *Regul. Pept.* **21** (1988) 117.
1494. J. Sueiras-Diaz et al., *J. Med. Chem.* **27** (1984) 310.
1495. J. L. Krstenansky et al., *Biochemistry* **25** (1986) 3833.
1496. K. A. Cornely et al., *Arch. Biochem. Biophys.* **240** (1985) 698.
1497. C. G. Unson et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 4083.
1498. C. G. Unson et al., *J. Biol. Chem.* **264** (1989) 789.
1499. B. Gysin, V. J. Hruby, *Experientia* **42** (1986) 680.
1500. S. W. Trenkner et al., *Radiology (Easton, Pa.)* **149** (1983) 401.
1501. A. M. N. Gardner, *Br. Med. J.* **290** (1985) 822.
1502. L. van Gaal et al., *Acta Clin. Belg.* **40** (1985) 266.
1503. J. Buch, A. Buch, *Acta Pharmacol. Toxicol.* **53** (1983) 188.
1504. A. E. Pontiroli et al., *Eur. J. Clin. Pharmacol.* **37** (1989) 427.
1505. L. M. Chuang et al., *Pharmacologist* **31** (1989) 125.
1506. A. Ohneda et al., *Horm. Metab. Res.* **19** (1987) 85.
1507. B. T. G. Schjoldager et al., *Eur. J. Clin. Invest.* **18** (1988) 499.
1508. C. Jarrousse et al., *FEBS Lett.* **188** (1985) 81.
1509. T. M. Biedzinski et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 967.
1510. K. Shima et al., *Acta Endocrinol. (Copenhagen)* **123** (1990) 464.
1511. B. Kreymann et al., *Lancet II* (1987) 1300.
1512. C. Orskov, J. H. Nielsen, *FEBS Lett.* **229** (1988) 175.
1513. G. Richter et al., *Acta Endocrinol. (Copenhagen)* **120** (1989) Suppl. 1, 191.
1514. A. B. Hansen et al., *FEBS Lett.* **236** (1988) 119.
1515. D. J. Drucker et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 3434.
1516. H.-C. Fehmann et al., *FEBS Lett.* **252** (1989) 109.
1517. E. Yamato et al., *Biochem. Biophys. Res. Commun.* **167** (1990) 431.
1518. C. Orskov et al., *Endocrinology (Baltimore)* **123** (1988) 2009.
1519. G. K. Hendrick et al., *Diabetes* **37** (1988) Suppl. 1, 49A.
1520. M. Gutniak et al., *Diabetologia* **33** (1990) Suppl., A73.
1521. W. A. Rogers et al., *Dig. Dis. Sci.* **28** (1983) 345.

1522. P. C. S. Blom et al., *Acta Physiol. Scand.* **123** (1985) 367.
1523. L. Groop et al., *Diabetes* **33** (1984) Suppl. 1, 166A.
1524. T. Krarup et al., *Diabetologia* **25** (1983) 173.
1525. C. K. Lardinois et al., *Diabetes* **33** (1984) 110A.
1526. F. J. Service et al., *J. Clin. Endocrinol. Metab.* **58** (1984) 1133.
1527. R. B. Richeson et al., *Clin. Res.* **33** (1985) 66A.
1528. A. J. McCullough et al., *Am. J. Physiol.* **248** (1985) E299.
1529. G. Hansen Starich et al., *Am. J. Physiol.* **249** (1985) E603.
1530. H. Hartmann et al., *Diabetologia* **29** (1986) 112.
1531. E. Sandberg et al., *Acta Physiol. Scand.* **127** (1986) 323.
1532. W. E. Schmidt et al., *Eur. J. Clin. Invest.* **16** (1986) part 2, A9.
1533. A. Rökaeus, *Trends Pharmacol. Sci.* **10** (1987) 158.
1534. K. Tatemoto et al., *FEBS Lett.* **164** (1983) 124.
1535. A. Rökaeus, M. Carlquist, *FEBS Lett.* **234** (1988) 400.
1536. L. M. Kaplan et al., *Proc. Natl. Acad. Sci. USA* **85** (1988) 7408.
1537. F. E. Bauer et al., *FEBS Lett.* **201** (1986) 327.
1538. F. E. Bauer et al., *Peptides (Fayetteville, N.Y.)* **7** (1986) 5.
1539. F. E. Bauer et al., *Lancet II* (1986) 192.
1540. F. E. Bauer et al., *J. Clin. Endocrinol. Metabol.* **63** (1986) 1372.
1541. I. Merchenthaler et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 6326.
1542. S. C. Hooi et al., *Neuroendocrinology* **51** (1990) 351.
1543. L. S. Brady et al., *Neuroendocrinology* **52** (1990) 441.
1544. E. Ekblad et al., *Br. J. Pharmacol.* **86** (1985) 241.
1545. L. W. Haynes, *Trends Pharmacol. Sci.* **7** (1986) 214.
1546. K. Tamura et al., *Eur. J. Pharmacol.* **136** (1987) 445.
1547. F. E. Bauer et al., *Gastroenterology* **97** (1989) 260.
1548. G. W. G. Sharp et al., *J. Biol. Chem.* **264** (1989) 7302.
1549. S. Lindskog, B. Ahren, *Acta Physiol. Scand.* **129** (1987) 305.
1550. T. Messell et al., *Regul. Pept.* **28** (1990) 161.
1551. Y. N. Kwok et al., *Eur. J. Pharmacol.* **145** (1988) 49.
1552. W. J. Rossowski, D. H. Coy, *Life Sci.* **44** (1989) 1807.
1553. S. G. Cella et al., *Endocrinology (Baltimore)* **122** (1988) 855.
1554. Y. Murakami et al., *Eur. J. Pharmacol.* **136** (1987) 415.
1555. S. M. Gabriel et al., *Life Sci.* **42** (1988) 1981.
1556. Y. Ben Ari, M. Lazdunski, *Eur. J. Pharmacol.* **165** (1989) 331.
1557. J. E. T. Fox et al., *Peptides (Fayetteville, N.Y.)* **9** (1988) 1183.
1558. B. Gallwitz et al., *Biochem. Biophys. Res. Commun.* **172** (1990) 268.
1559. X.-J. Xu et al., *Eur. J. Pharmacol.* **182** (1990) 137.
1560. A. Kuwahara et al., *Regul. Pept.* **29** (1990) 23.
1561. L. A. Frohman, J.-O. Jansson, *Endocr. Rev.* **7** (1986) 223.
1562. K. Chihara et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 269.
1563. P. Ciofi et al., *Neuroendocrinology* **45** (1987) 425.
1564. B. Meister et al., *Acta Physiol. Scand.* **124** (1985) 133.
1565. H. Okamura et al., *Neuroendocrinology* **41** (1985) 177.
1566. K. Arase et al., *Endocrinology (Baltimore)* **121** (1987) 1960.
1567. Y. Murakami et al., *Eur. J. Pharmacol.* **136** (1987) 415.
1568. K. Chihara et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 466.
1569. Y. Kashio et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 92.
1570. S. M. Gabriel et al., *Neuroendocrinology* **50** (1989) 299.
1571. J. Argente et al., *J. Clin. Endocrinol. Metab.* **63** (1986) 680.
1572. I. Ganzetti et al., *Peptides (Fayetteville, N.Y.)* **7** (1986) 1011.
1573. G. P. Ceda et al., *Endocrinology (Baltimore)* **120** (1987) 1658.
1574. H. Katakami et al., *Endocrinology (Baltimore)* **118** (1986) 1872.
1575. M. C. Aguila, S. M. McCann, *Endocrinology (Baltimore)* **120** (1987) 341.
1576. I. Ganzetti et al., *J. Endocrinol. Invest.* **10** (1987) 241.
1577. V. de Gennaro Colonna et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 705.
1578. W. B. Wehrenberg et al., *Neuroendocrinology* **43** (1986) 266.

1579. H. Seifert et al., *Endocrinology (Baltimore)* **117** (1985) 424.
1580. W. B. Wehrenberg, *Endocrinology (Baltimore)* **118** (1986) 489.
1581. C. E. Brain et al., *J. Endocrinol.* **117** (1988) Suppl., 82.
1582. R. Valcavi et al., *J. Endocrinol.* **112** (1987) Suppl., 154.
1583. V. Locatelli et al., *J. Endocrinol.* **111** (1986) 271.
1584. N. Mauras et al., *Metab. Clin. Exp.* **36** (1987) 369.
1585. K. Chihara et al., *Endocrinology (Baltimore)* **114** (1984) 1402.
1586. M. C. White et al., *J. Endocrinol.* **105** (1985) 269.
1587. M. Press et al., *Diabetes* **33** (1984) Suppl. 1, 13A.
1588. C. Pintor et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 263.
1589. N. T. Richards et al., *Diabetologia* **27** (1984) 529.
1590. R. Valcavi et al., *Clin. Endocrinol.* **24** (1986) 693.
1591. M. Hotta et al., *Life Sci.* **42** (1988) 979.
1592. K.-P. Lesch et al., *J. Clin. Endocrinol. Metab.* **65** (1987) 1278.
1593. R. Cacabelos et al., *Acta Endocrinol. (Copenhagen)* **117** (1988) 295.
1594. F. L. Culler et al., *Clin. Res.* **35** (1987) 221A.
1595. F. Zeytin, P. Braceau, *Biochem. Biophys. Res. Commun.* **123** (1984) 497.
1596. T. Lehy et al., *Gastroenterology* **100** (1986) 646.
1597. F. Obal, Jr. et al., *Am. J. Physiol.* **255** (1988) R310.
1598. Z. Josefsberg et al., *Isr. J. Med. Sci.* **23** (1987) 837.
1599. N. Ling et al., *Biochem. Biophys. Res. Commun.* **123** (1984) 854.
1600. N. Ling et al., *Biochem. Biophys. Res. Commun.* **122** (1984) 304.
1601. T. J. Aitman et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 1.
1602. M. Kovacs et al., *Life Sci.* **42** (1988) 27.
1603. L. Bokser et al., *Life Sci.* **46** (1990) 999.
1604. K. Sato et al., *Biochem. Biophys. Res. Commun.* **149** (1987) 531.
1605. K. Sato et al., *Biochem. Biophys. Res. Commun.* **167** (1990) 360.
1606. S. J. Hocart et al., *J. Med. Chem.* **33** (1990) 1954.
1607. D. H. Coy et al., *J. Med. Chem.* **30** (1987) 219.
1608. R. J. M. Ross et al., *Lancet I* (1987) 5.
1609. M. O. Thorner et al., *N. Engl. J. Med.* **312** (1985) 4.
1610. O. Butenandt, B. Staudlt, *Eur. J. Pediatr.* **148** (1989) 393.
1611. P. Franchimont et al., *Acta Clin. Belg.* **42** (1987) 143.
1612. Beckmann Instruments, Inc., EP 0018072, 1979.
1613. C. Y. Bowers et al., *Endocrinology (Baltimore)* **114** (1984) 1573.
1614. O. Sartor et al., *Endocrinology (Baltimore)* **116** (1985) 952.
1615. B. E. Ilson et al., *J. Clin. Endocrinol. Metab.* **69** (1989) 212.
1616. M. Ono et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 4330.
1617. M. Yamakawa et al., *J. Biol. Chem.* **265** (1990) 8915.
1618. I. A. MacFarlane et al., *Acta Endocrinol. (Copenhagen)* **112** (1986) 547.
1619. N. E. Cooke, J. Ray, J. G. Emergy, S. A. Liebhaber, *J. Biol. Chem.* **263** (1988) 9001.
1620. A. Skottner et al., *Acta Endocrinol. (Copenhagen)* **118** (1988) 14.
1621. A. Gertler et al., *Acta Endocrinol. (Copenhagen)* **118** (1988) 720.
1622. J.-O. Jansson et al., *Endocr. Rev.* **6** (1985) 128.
1623. H. Soya, M. Suzuki, *Endocrinology (Baltimore)* **122** (1988) 2492.
1624. P. Maertens, C. Denef, *Mol. Cell. Endocrinol.* **54** (1987) 203.
1625. H. Yagi et al., *Horm. Metab. Res.* **18** (1986) 723.
1626. A. D. Struthers et al., *Neuroendocrinology* **44** (1986) 22.
1627. F. E. Chang et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 551.
1628. M. Marastoni et al., *Arzneim. Forsch.* **39** (1989) 639.
1629. E. C. degli Uberti et al., *Metabolism* **39** (1990) 1063.
1630. L. Altomonte et al., *Exp. Clin. Endocrinol.* **88** (1986) 334.
1631. P. A. Graham et al., *Clin. Res.* **35** (1987) 396A.
1632. J. R. Peters et al., *Clin. Endocrinol.* **25** (1986) 213.
1633. E. W. Bernton et al., *Science (Washington, D.C.)* **238** (1987) 519.
1634. F. Lopez et al., *Acta Endocrinol. (Copenhagen)* **113** (1986) 317.
1635. M. L. Vance et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 1136.
1636. J. K. Schmitt, *Clin. Res.* **37** (1989) 9A.

1637. K. Y. Ho et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 51.
1638. M. J. Mansfield et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 3.
1639. N. Mauras et al., *Metabolism* **38** (1989) 286.
1640. A. Giustina et al., *Horm. Metab. Res.* **21** (1989) 693.
1641. A. Giustina et al., *Acta Endocrinol. (Copenhagen)* **122** (1990) 206.
1642. R. J. M. Ross et al., *Clin. Endocrinol.* **26** (1987) 117.
1643. S. Kentroti et al., *Endocrinology (Baltimore)* **122** (1988) 2407.
1644. N. Ono et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 7787.
1645. G. P. Ceda et al., *Acta Endocrinol. (Copenhagen)* **120** (1989) 416.
1646. C. Netti et al., *Neuroendocrinology* **49** (1989) 242.
1647. H. Katakami et al., *Endocrinology (Baltimore)* **117** (1985) 1139.
1648. N. C. Schaper, *Acta Endocrinol. (Copenhagen)* **122** (1990) 7.
1649. K.-P. Lesch et al., *J. Clin. Endocrinol. Metab.* **65** (1987) 1278.
1650. A. Zarate et al., *Horm. Metab. Res.* **18** (1986) 400.
1651. C. Marchesi et al., *Pharmacopsychiatry* **20** (1987) 64.
1652. P. Pietschmann et al., *Horm. Metab. Res.* **22** (1990) 109.
1653. O. Giampietro et al., *Metabolism* **36** (1987) 1149.
1654. D. W. Leung et al., *Nature (London)* **330** (1987) 537.
1655. L. S. Mathews et al., *J. Biol. Chem.* **264** (1989) 9905.
1656. R. L. Hintz, *Horm. Res.* **33** (1990) 105.
1657. P. J. Godowski et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 8083.
1658. A. Aguirre et al., *Horm. Res.* **34** (1990) 4.
1659. G. Baumann et al., *N. Engl. J. Med.* **320** (1989) 1705.
1660. M. C. Slootweg et al., *J. Endocrinol.* **116** (1988) R11.
1661. J. H. Nielsen et al., *Mol. Endocrinol.* **13** (1989) 165.
1662. M. E. Markowitz et al., *J. Clin. Endocrinol. Metab.* **69** (1989) 420.
1663. J. K. Damm, *Acta Endocrinol. (Copenhagen)* **114** (1987) 124.
1664. I. Ganzetti et al., *J. Endocrinol. Invest.* **10** (1987) 241.
1665. D. M. Maiter et al., *Neuroendocrinology* **51** (1990) 174.
1666. J. Isgaard et al., *Endocrinology (Baltimore)* **123** (1988) 2605.
1667. C. T. Roberts, Jr. et al., *J. Biol. Chem.* **22** (1986) 10025.
1668. J. H. Nielsen et al., *Acta Endocrinol. (Copenhagen)* **110** (1985) Suppl. 273, 74.
1669. L. G. Frigert et al., *Horm. Metab. Res.* **19** (1987) 464.
1670. F. Salomon et al., *Diabetologia* **33** (1990) Suppl., A218.
1671. M. B. Davidson, *Endocr. Rev.* **8** (1987) 115.
1672. C. K. Edwards III et al., *Science (Washington, D.C.)* **239** (1988) 769.
1673. B. Lawier Goff et al., *Clin. Exp. Immunol.* **68** (1987) 580.
1674. S. L. Kaplan et al., *Lancet I* (1986) 697.
1675. J. L. Kostyo et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 4250.
1676. S. L. Kaplan et al., *Lancet I* (1986) 697.
1677. F. L. Culler et al., *Horm. Metab. Res.* **20** (1988) 107.
1678. N. Emoto et al., *Acta Endocrinol. (Copenhagen)* **114** (1987) 283.
1679. F. M. Ng, J. A. Harcourt, *Diabetologia* **29** (1986) 882.
1680. V. M. J. Robson et al., *Hoppe-Seyler's Z. Physiol. Chem.* **371** (1990) 423.
1681. R. W. Stevenson et al., *Acta Endocrinol. (Copenhagen)* **117** (1988) 457.
1682. C. E. Mondon et al., *Endocrinology (Baltimore)* **123** (1988) 827.
1683. C. Singer-Granick et al., *Horm. Res.* **24** (1986) 246.
1684. H. Waago, *Lancet I* (1987) 1485.
1685. E. A. van der Veen, J. C. Netelenbos, *Horm. Res.* **33** (1990) 65.
1686. F. Greig et al., *Horm. Res.* **31** (1989) Suppl. 1, 20.
1687. D. Rudman et al., *N. Engl. J. Med.* **323** (1990) 1.
1688. Y. A. Pankov, V. Y. Butnev, *J. Pept. Prot. Res.* **28** (1986) 113.
1689. R. Einspanier et al., *FEBS Lett.* **204** (1986) 37.
1690. M. Watahiki et al., *J. Biol. Chem.* **264** (1989) 5535.
1691. S. H. Shin et al., *Can. J. Physiol. Pharmacol.* **65** (1987) 2036.
1692. G. Nagy et al., *Biochem. Biophys. Res. Commun.* **151** (1988) 524.
1693. J. E. Merritt, B. L. Brown, *Life Sci.* **35** (1984) 707.
1694. F. Bigi et al., *Exp. Clin. Endocrinol.* **95** (1990) 224.

1695. K. Törnquist, *Acta Endocrinol. (Copenhagen)* **116** (1987) 459.
1696. I. S. Kampa et al., *Horm. Metab. Res.* **18** (1986) 419.
1697. J. S. Ramsdell, A. H. Tashjian, Jr., *Endocrinology (Baltimore)* **117** (1985) 2050.
1698. P. G. Knight et al., *Neuroendocrinology* **44** (1986) 29.
1699. H. Abe et al., *Endocrinology (Baltimore)* **116** (1985) 1383.
1700. E. Rolandi et al., *Horm. Res.* **21** (1985) 209.
1701. K. D. Meirleir et al., *Horm. Metab. Res.* **17** (1985) 380.
1702. P. Buydens et al., *Horm. Metab. Res.* **18** (1986) 575.
1703. P. L. Canonico, R. M. MacLeod, *Endocrinology (Baltimore)* **118** (1986) 233.
1704. H. Minakami et al., *Endocrinol. Jpn.* **33** (1986) 511.
1705. T. H. Jones et al., *J. Endocrinol.* **111** (1986) Suppl., 134.
1706. M. Memo et al., *J. Neurochem.* **47** (1986) 1689.
1707. I. S. Login, *Life Sci.* **47** (1990) 2269.
1708. P. Tatar, M. Vigas, *Neuroendocrinology* **39** (1984) 275.
1709. U. Knigge et al., *J. Clin. Endocrinol. Metab.* **62** (1986) 491.
1710. I. Murai, N. Ben-Jonathan, *Neuroendocrinology* **43** (1986) 453.
1711. S. H. Shin, R. Stirling, *J. Endocrinol.* **118** (1988) 287.
1712. N. Yonehara, D. H. Clouet, *J. Pharm. Exp. Ther.* **231** (1984) 38.
1713. S. Röjdmark et al., *J. Endocrinol. Invest.* **7** (1984) 635.
1714. L. S. Frawley, C. L. Clark, *Endocrinology (Baltimore)* **119** (1986) 1462.
1715. J.-P. Loeffler et al., *Neuroendocrinology* **43** (1986) 504.
1716. K. Fuxe et al., *Acta Physiol. Scand.* **125** (1985) 437.
1717. T. Karashima, A. V. Schally, *Proc. Soc. Exp. Biol. Med.* **185** (1987) 69.
1718. I. Zofkova, J. Nedvidkova, *Exp. Clin. Endocrinol.* **92** (1988) 262.
1719. E. W. Bernton et al., *Science (Washington, D.C.)* **238** (1987) 519.
1720. D. F. Wood et al., *J. Endocrinol.* **115** (1987) 497.
1721. A. I. Esquifino et al., *J. Endocrinol. Invest.* **12** (1989) 171.
1722. J. D. Wark, V. Gurtler, *Biochem. J.* **241** (1987) 397.
1723. L. V. dePaolo et al., *Peptides (Fayetteville, N.Y.)* **7** (1986) 541.
1724. A. Gertler et al., *Endocrinology (Baltimore)* **118** (1986) 720.
1725. C. L. Bethea, E. Yuzuriha, *Endocrinology (Baltimore)* **119** (1986) 771.
1726. N. Kimura et al., *Endocrinology (Baltimore)* **119** (1986) 1028.
1727. W. D. Hetzel et al., *Horm. Metab. Res.* **22** (1990) 648.
1728. V. C. Musey et al., *N. Engl. J. Med.* **316** (1987) 229.
1729. A. Kauppila et al., *J. Clin. Endocrinol. Metab.* **64** (1987) 309.
1730. B. B. Arnetz et al., *Life Sci.* **39** (1986) 135.
1731. P. C. Hiestand et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 2599.
1732. J.-M. Boutin et al., *Cell* **53** (1988) 69.
1733. M. Edery et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 2112.
1734. P. Guillaumot et al., *Biochem. Biophys. Res. Commun.* **135** (1986) 1076.
1735. H. G. Klemcke et al., *Endocrinology (Baltimore)* **118** (1986) 773.
1736. J. A. Rillema et al., *Horm. Metab. Res.* **18** (1986) 672.
1737. R. E. Hruska, *J. Neurochem.* **47** (1986) 1908.
1738. E. Jungmann et al., *Horm. Metab. Res.* **18** (1986) 704.
1739. T. H. Welsh, Jr. et al., *Biol. Reprod.* **34** (1986) 796.
1740. F. T. Murray et al., *J. Clin. Endocrinol. Metab.* **59** (1984) 79.
1741. A. Canfriez, *Horm. Res.* **22** (1985) 209.
1742. W. G. North et al., *Horm. Res.* **15** (1981) 55.
1743. S. Kaufmann, *J. Physiol. (London)* **310** (1981) 435.
1744. G. J. Pepe, E. D. Albrecht, *Endocrinology (Baltimore)* **117** (1985) 1968.
1745. A. R. Buckley et al., *Biochem. Biophys. Res. Commun.* **138** (1986) 1138.
1746. A. R. Buckley et al., *Life Sci.* **37** (1985) 2569.
1747. T. G. Muldoon, *Endocrinology (Baltimore)* **121** (1987) 141.
1748. F. Drago et al., *Life Sci.* **36** (1985) 191.
1749. J. P. Barlet, *J. Endocrinol.* **107** (1985) 171.
1750. M. P. Caraceni et al., *Calcif. Tissue Int.* **37** (1985) 687.
1751. R. E. Humbel, *Eur. J. Biochem.* **190** (1990) 445.
1752. W. H. Daughaday et al., *Endocrinology (Baltimore)* **121** (1987) 1911.
1753. K. Ramasharma et al., *Biochem. Biophys. Res. Commun.* **140** (1986) 536.
1754. L. K. Bachrach et al., *Biochem. Biophys. Res. Commun.* **154** (1988) 861.

1755. M. R. Hammerman, *Am. J. Physiol.* **257** (1989) F503.
1756. D. C. Costigan et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 1014.
1757. A. N. Corps et al., *J. Clin. Endocrinol. Metab.* **67** (1988) 25.
1758. J. Wroblewski et al., *Acta Endocrinol. (Copenhagen)* **115** (1987) 37.
1759. E.-M. Rutanen, F. Pekonen, *Acta Endocrinol. (Copenhagen)* **123** (1990) 7.
1760. A. Shimatsu, P. Rotwein, *J. Biol. Chem.* **262** (1987) 7894.
1761. A. Honegger, R. E. Humbel, *J. Biol. Chem.* **261** (1986) 569.
1762. P. P. Zumstein et al., *Biochemistry* **82** (1985) 3169.
1763. P. Rotwein, *Proc. Natl. Acad. Sci. USA* **83** (1986) 77.
1764. V. R. Sara et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 4904.
1765. L. J. Murphy et al., *Endocrinology (Baltimore)* **122** (1988) 2027.
1766. E. Jennische et al., *Acta Physiol. Scand.* **129** (1987) 9.
1767. P. Bang et al., *Eur. J. Clin. Invest.* **20** (1990) 285.
1768. H.-A. Hansson et al., *Acta Physiol. Scand.* **129** (1987) 165.
1769. K. Ramasharma, C. H. Li, *Proc. Natl. Acad. Sci. USA* **84** (1987) 2643.
1770. L. J. Murphy, A. Ghahary, *Endocr. Rev.* **11** (1990) 443.
1771. T. K. Gray et al., *Biochem. Biophys. Res. Commun.* **158** (1989) 407.
1772. T. L. McCarthy et al., *J. Biol. Chem.* **265** (1990) 15353.
1773. C. G. Scanes et al., *IRCS Med. Sci.* **14** (1986) 515.
1774. S. Goldstein, L. S. Phillips, *Metabolism* **38** (1989) 745.
1775. W. E. Sonntag, R. L. Boyd, *Life Sci.* **43** (1988) 1325.
1776. D. A. Harris et al., *J. Clin. Endocrinol. Metab.* **61** (1985) 152.
1777. A. Tham et al., *Horm. Metab. Res.* **18** (1986) 706.
1778. R. G. Rosenfeld et al., *Biochem. Biophys. Res. Commun.* **143** (1987) 199.
1779. T. A. Gustafson, W. J. Rutter, *J. Biol. Chem.* **265** (1990) 18663.
1780. R. G. MacDonald et al., *Science (Washington, D.C.)* **239** (1988) 1134.
1781. R. L. Hintz, *Horm. Res.* **33** (1990) 105.
1782. A. L. Albiston, A. C. Herington, *Biochem. Biophys. Res. Commun.* **166** (1990) 892.
1783. S. Shimasaki et al., *Mol. Endocrinol.* **4** (1990) 1451.
1784. A. M. Taylor et al., *Clin. Endocrinol.* **32** (1990) 229.
1785. M. Ross et al., *Biochem. J.* **258** (1989) 267.
1786. I. Kojima et al., *J. Biol. Chem.* **265** (1990) 16846.
1787. M. Ernst, E. R. Froesch, *Biochem. Biophys. Res. Commun.* **151** (1988) 142.
1788. B. P. Halloran, E. M. Spencer, *Endocrinology (Baltimore)* **123** (1988) 1225.
1789. H.-P. Guler et al., *N. Engl. J. Med.* **317** (1987) 137.
1790. H.-P. Guler et al., *Proc. Natl. Acad. Sci. USA* **86** (1989) 2868.
1791. S. Yamashita et al., *J. Clin. Endocrinol. Metab.* **43** (1986) 730.
1792. C. G. Goodyer et al., *Endocrinology (Baltimore)* **115** (1984) 1568.
1793. K. M. Thrailkill et al., *Endocrinology (Baltimore)* **123** (1988) 2930.
1794. M.-H. Perrard-Sapori et al., *Eur. J. Biochem.* **165** (1987) 209.
1795. T. Lin et al., *Biochem. Biophys. Res. Commun.* **137** (1986) 950.
1796. J. D. Veldhuis et al., *J. Biol. Chem.* **261** (1986) 2499.
1797. S. E. Tollefsen et al., *J. Biol. Chem.* **264** (1989) 13810.
1798. A. J. Stewart et al., *J. Biol. Chem.* **265** (1990) 21172.
1799. M. L. Stracke et al., *J. Biol. Chem.* **264** (1989) 21544.
1800. K. Binz et al., *Proc. Natl. Acad. Sci. USA* **87** (1990) 3690.
1801. P. Hunt, D. D. Eardley, *J. Immunol.* **136** (1986) 3994.
1802. L. Y.-H. Tseng et al., *Biochem. Biophys. Res. Commun.* **149** (1987) 672.
1803. M. L. Bayne et al., *J. Biol. Chem.* **263** (1988) 6233.
1804. M. L. Bayne et al., *J. Biol. Chem.* **265** (1990) 15648.
1805. C. J. Bagley et al., *Biochem. J.* **259** (1989) 665.
1806. J. D. Quin et al., *Diabetologia* **32** (1989) 531A.
1807. Z. Laron et al., *Lancet II* (1988) 1170.
1808. D. J. Gross et al., *J. Biol. Chem.* **264** (1989) 21486.
1809. E. Helmerhorst, G. B. Stokes, *Diabetes* **36** (1987) 261.
1810. W. König in A. Kleemann et al., (eds.): *Arzneimittel, Fortschritte 1972 bis 1985,* VCH Verlagsgesellschaft, Weinheim 1987, p. 728.

1811. T. Adachi et al., *J. Biol. Chem.* **264** (1989) 7681.
1812. A. B. Smit et al., *Nature (London)* **331** (1988) 535.
1813. M. Lagueux et al., *Eur. J. Biochem.* **187** (1990) 249.
1814. J. N. Fain, *Metabolism* **33** (1984) 672.
1815. B. Kreymann et al., *Lancet II* (1987) 1300.
1816. T. Matsuyama et al., *Diabetes* **36** (1987) Suppl. 1, 164A.
1817. W. H. Hsu et al., *Proc. Soc. Exp. Biol. Med.* **184** (1987) 345.
1818. Y. Totsuka et al., *Biochem. Biophys. Res. Commun.* **158** (1989) 1060.
1819. S. B. Richardson et al., *Diabetes* **37** (1988) 103A.
1820. H. Kofod et al., *Am. J. Physiol.* **154** (1988) E454.
1821. M. C. d'Emden et al., *Biochem. Biophys. Res. Commun.* **164** (1989) 413.
1822. A. Faure et al., *Horm. Metab. Res.* **17** (1985) 378.
1823. T. Ikeda et al., *Biochem. Pharmacol.* **40** (1990) 1769.
1824. S. B. Pek, M. F. Walsh, *Proc. Natl. Acad. Sci. USA* **81** (1984) 2199.
1825. N. G. Morgan et al., *Biochim. Biophys. Acta C* **845** (1985) 526.
1826. G. Holm, *Acta Med. Scand. Suppl.* **672** (1983) 21.
1827. I. C. Green et al., *Diabetologia* **27** (1984) 282A.
1828. B. Ahren et al., *Biochem. Biophys. Res. Commun.* **140** (1986) 1054.
1829. K. Hermansen, B. Ahren, *Regul. Pept.* **27** (1990) 149.
1830. S. Raptis et al., *Diabetes* **32** (1983) Suppl. 1, 94A.
1831. F. Shimizu et al., *Endocrinology (Baltimore)* **117** (1985) 2081.
1832. C. Southern et al., *Biochem. J.* **272** (1990) 243.
1833. S. Baekkeskov et al., *Nature (London)* **347** (1990) 151.
1834. D. B. Jones et al., *Lancet* **336** (1900) 583.
1835. S. Zeuzem, *Aktuel. Endokrinol. Stoffwechsel* **8** (1987) 132.
1836. T. A. Gustafson, W. J. Rutter, *J. Biol. Chem.* **265** (1990) 18663.
1837. C. Correze et al., *Biochem. Biophys. Res. Commun.* **126** (1985) 1061.
1838. M. Taira et al., *Science (Washington, D.C.)* **245** (1989) 63.
1839. M. Odawara et al., *Science (Washington,D.C.)* **245** (1989) 66.
1840. D. E. Moller et al., *Mol. Endocrinol.* **4** (1990) 1183.
1841. E. G. Walters et al., *Lancet I* (1987) 241.
1842. P. Briata et al., *Biochem. Biophys. Res. Commun.* **169** (1990) 397.
1843. M. L. Standaert, R. J. Pollet, *J. Biol. Chem.* **259** (1984) 2346.
1844. V. Papa et al., *J. Clin. Invest.* **86** (1990) 1503.
1845. G. B. Willars, T. W. Atkins, *J. Endocrinol.* **104** (1985) Suppl., 100.
1846. J. D. McArmstrong et al., *Acta Endocrinol. (Copenhagen)* **102** (1983) 492.
1847. J. C. Jarrett II et al., *Am. J. Obstet. Gynecol.* **149** (1984) 250.
1848. K. Kriauciunas, C. R. Kahn, *Clin. Res.* **32** (1984) 401A.
1849. G. Kasdorf, R. K. Kalkhoff, *Clin. Res.* **34** (1986) 965A.
1850. P. Dent et al., *Nature (London)* **348** (1990) 302.
1851. J. S. Marks, L. H. Parker Botelho, *J. Biol. Chem.* **261** (1986) 2781.
1852. D. E. James et al., *Nature (London)* **338** (1989) 83.
1853. W. G. Lampson et al., *Can. J. Physiol. Pharmacol.* **61** (1983) 457.
1854. H. S. Glauber et al., *N. Engl. J. Med.* **316** (1987) 443.
1855. C. L. Oppenheimer et al., *J. Biol. Chem.* **258** (1983) 4824.
1856. M. P. Cohen et al., *Biochim. Biophys. Acta M* **856** (1986) 182.
1857. J. Ben Davoren, A. J. W. Hsueh, *Mol. Cell. Endocrinol.* **35** (1984) 97.
1858. T. Lin et al., *Endocrinology (Baltimore)* **119** (1986) 1641.
1859. J. Rodin et al., *Metab. Clin. Exp.* **34** (1985) 826.
1860. C. J. H. Ingoldby et al., *Am. J. Gastroenterol.* **79** (1984) 16.
1861. K. Y. Lee et al., *Am. J. Physiol.* **258** (1990) G268.
1862. E. Mezey et al., *Science (Washington, D.C.)* **226** (1984) 1085.
1863. J. H. Pratt et al., *Clin. Res.* **31** (1983) 763A.
1864. A. J. Spijker et al., *Eur. J. Clin. Invest.* **15** (1985) part 2, A16.
1865. V. Schusdziarra et al., *Neuropeptides* **7** (1986) 51.
1866. J. Guntupalli et al., *Am. J. Physiol.* **249** (1985) part 2, F610.
1867. B. E. Kream et al., *Endocrinology (Baltimore)* **116** (1985) 296.
1868. N. Wongsurawat, H. J. Armbrecht, *Acta Endocrinol. (Copenhagen)* **109** (1985) 243.

1869. S. Yagi et al., *Diabetes* **37** (1988) 1064.
1870. A. Siani et al., *Eur. J. Clin. Pharmacol.* **38** (1990) 393.
1871. R. Sethi et al., *J. Mol. Cell. Cardiol.* **22** (1990) Suppl. 1, S33.
1872. C. Grunfeld, D. S. Jones, *Diabetes* **32** (1983) 128A.
1873. H. Schatz, S. Ammermann, *Horm. Metab. Res., Suppl. Ser.* **18** (1988) 1.
1874. C. Rosak et al., *Horm. Metab. Res., Suppl. Ser.* **18** (1988) 16.
1875. G. D. Smith et al., *Proc. Natl. Acad. Sci. USA* **81** (1984) 7093.
1876. G. P. Schwartz et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 6408.
1877. G. T. Burke et al., *Biochem. Biophys. Res. Commun.* **173** (1990) 982.
1878. S.-C. Chu et al., *Sci. Sin. (Engl. Ed.)* **16** (1973) 71.
1879. R. G. Mirmira, H. S. Tager, *J. Biol. Chem.* **264** (1989) 6349.
1880. J. Brange et al., *Nature (London)* **333** (1988) 679.
1881. H. P. Neubauer et al., *Diabetologia* **27** (1984) Suppl., 129.
1882. S. Jorgensen et al., *Diabetologia* **32** (1989) 500A.
1883. P. Balschmidt et al., *Diabetologia* **33** (1990) Suppl., A117.
1884. L. D. Monti et al., *Diabetologia* **33** (1990) Suppl., A60.
1885. P. Balschmidt et al., *Diabetes* **38** (1989) Suppl. 2, 157A.
1886. M. Hashimoto et al., *Pharm. Res.* **6** (1989) 171.
1887. S. Kang et al., *Diabetologia* **32** (1989) 502A.
1888. J. N. MacPherson, J. Feely, *Br. Med. J.* **300** (1990) 731.
1889. M. Rodier, *Presse Med.* **19** (1990) 959.
1890. W. Foertsch, *Aktuel. Endokrinol. Stoffwechsel* **7** (1986) 33.
1891. G. Paolisso et al., *J. Clin. Endocrinol. Metab.* **66** (1988) 1220.
1892. A. Roza et al., *Clin. Res.* **31** (1983) 622A.
1893. W. T. Chance et al., *JNCI J. Natl. Cancer Inst.* **77** (1986) 497.
1894. S. J. Whang, F. L. Greenway, *Clin. Res.* **33** (1985) 67A.
1895. B. E. Kemp, H. D. Niall, *Vitam. Horm. (N.Y.)* **41** (1984) 79.
1896. G. Mazoujian, G. D. Bryant-Greenwood, *Lancet* **335** (1990) 299.
1897. P. Hudson et al., *EMBO J.* **3** (1984) 2333.
1898. J. S. Ottobre et al., *Biol. Reprod.* **31** (1984) 1000.
1899. K. Seki et al., *Endocrinol. Jpn.* **33** (1986) 727.
1900. A. I. Musah et al., *Endocrinology (Baltimore)* **120** (1987) 317.
1901. A. L. Bernal et al., *Br. J. Obstet. Gynaecol.* **94** (1987) 1045.
1902. B. Viell, H. Struck, *Horm. Metab. Res.* **19** (1987) 415.
1903. J. St-Louis, G. Massicotte, *Life Sci.* **37** (1985) 1351.
1904. G. Massicotte et al., *Clin. Exp. Hypertens. Part B* **B6** (1987) 255.
1905. E. E. Büllesbach, C. Schwabe, *Biochemistry* **25** (1986) 5998.
1906. C. H. S. McIntosh, *Life Sci.* **17** (1985) 2043.
1907. M. Bersani et al., *J. Biol. Chem.* **264** (1989) 10633.
1908. P. G. Burhol et al., *Acta Physiol. Scand.* **121** (1984) 223.
1909. D. LeRoith et al., *Endocrinology (Baltimore)* **117** (1985) 2093.
1910. S. M. Cutfield et al., *FEBS Lett.* **214** (1987) 57.
1911. S. V. Wu et al., *Metabolism* **39** (1990) 125.
1912. R. M. Post, D. R. Rubinow, *Lancet II* (1986) 810.
1913. R. Zorrilla et al., *Neuroendocrinology* **52** (1990) 527.
1914. T. Chiba et al., *Am. J. Physiol.* **253** (1987) G62.
1915. M. C. Aguila et al., *Neuroendocrinology* **52** (1990) 238.
1916. M. Sato et al., *Neuroendocrinology* **50** (1989) 139.
1917. D. E. Scarborough, *Metabolism* **39** (1990) 108.
1918. H. Koop, R. Arnold, *Gastroenterology* **84** (1983) part 2, 1214.
1919. K. Uvnäs-Moberg, L. Wetterberg, *Acta Physiol. Scand.* **120** (1984) 517.
1920. S. M. Webb et al., *Acta Endocrinol. (Copenhagen)* **110** (1985) 145.
1921. S. E. Shoelson et al., *Am. J. Physiol.* **250** (1986) E428.
1922. G. Strazzulla et al., *Horm. Metab. Res.* **20** (1988) 126.
1923. H. Stepien et al., *Horm. Metab. Res.* **18** (1986) 555.
1924. F. Sicuteri et al., *IRCS Med. Sci.* **13** (1985) 308.
1925. D. Neugebauer, G. Weber, *Z. Hautkrankh.* **63** (1988) 585.
1926. V. Schusdziarra et al., *Diabetes* **34** (1985) 595.

1927. K. Uvnäs-Moberg et al., *Acta Obstet. Gynecol. Scand.* **68** (1989) 165.
1928. S. Gomez et al., *Life Sci.* **39** (1986) 623.
1929. J. M. Radke et al., *Eur. J. Pharmacol.* **134** (1987) 105.
1930. J. C. Reubi et al., *Metabolism* **39** (1990) 78.
1931. M. Flint Beal et al., *Science (Washington, D.C.)* **229** (1985) 289.
1932. S. Reichlin, *N. Engl. J. Med.* **309** (1983) 1556.
1933. M. J. Toro et al., *Horm. Res.* **29** (1988) 59.
1934. M. R. Brown et al., *Endocrinology (Baltimore)* **114** (1984) 1546.
1935. A. Skamene, Y. C. Patel, *Clin. Endocrinol.* **20** (1984) 555.
1936. S. Mulvihill et al., *N. Engl. J. Med.* **310** (1984) 467.
1937. S. Bjorck, J. Svanvik, *Scand. J. Gastroenterol.* **19** (1984) 173.
1938. P. P. Gazzaniga et al., *Experientia* **44** (1988) 892.
1939. M. Mogard et al., *Gastroenterology* **86** (1984) Part 2, 1186.
1940. G. J. Krejs, *Diabetes* **33** (1984) 548.
1941. R. G. Long et al., *Br. Med. J.* **290** (1985) 886.
1942. J. Morisset, *Regul. Pept.* **10** (1984) 11.
1943. E. F. Stange et al., *Horm. Metab. Res.* **16** (1984) 74.
1944. N. Altorki et al., *Gastroenterology* **86** (1984) part 2, 1014.
1945. C. Liebow et al., *Metabolism* **39** (1990) 163.
1946. M. Pawlikowski et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 951.
1947. M. Pawlikowski et al., *Neuropeptides* **13** (1989) 75.
1948. M. Pawlikowski et al., *Biochem. Biophys. Res. Commun.* **129** (1985) 52.
1949. *Drugs Future* **9** (1984) 342; **10** (1985) 435; **11** (1986) 430; **12** (1987) 500; **13** (1988) 482; **14** (1989) 479.
1950. R.-Z. Cai et al., *Proc. Natl. Acad. Sci. USA* **83** (1986) 1896. *Drugs Future* **14** (1989) 1052; **15** (1990) 1144.
1951. R.-Z. Cai et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 2502.
1952. W. A. Murphy et al., *Life Sci.* **40** (1987) 2515.
1953. K. Ziegler et al., *Biochim. Biophys. Acta C* **845** (1985) 86.
1954. G. Rohr et al., *Klin. Wochenschr.* **64** (1986) Suppl. 7, 90.
1955. S. Szabo et al., *Can. J. Physiol. Pharmacol.* (1986)Suppl., 8.
1956. J. T. Pelton et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 236.
1957. J. M. Walker et al., *Peptides (Fayetteville, N.Y.)* **8** (1987) 869.
1958. P. E. Battershill, S. P. Clissold, *Drugs* **38** (1989) 658.
1959. J. L. Abelson et al., *J. Clin. Psychopharmacol.* **10** (1990) 128.
1960. H. J. Balks et al., *Eur. J. Clin. Pharmacol.* **36** (1989) 133.
1961. R. D. Penn et al., *Lancet* **335** (1990) 738.
1962. K. Tatemoto et al., *Nature (London)* **324** (1986) 476.
1963. A. L. Iacangelo et al., *Endocrinology (Baltimore)* **122** (1988) 2339.
1964. A. Funakoshi et al., *Regul. Pept.* **30** (1990) 159.
1965. S. Efendic et al., *Proc. Natl. Acad. Sci. USA* **84** (1987) 7257.
1966. A. Funakoshi et al., *Regul. Pept.* **24** (1989) 225.
1967. A. Funakoshi et al., *Regul. Pept.* **25** (1989) 157.
1968. V. Sanchez et al., *Biosci. Rep.* **10** (1990) 87.
1969. J. P. Smith, S. T. Kramer, *Gastroenterology* **96** (1989) part 2, A479.
1970. T. Zhang et al., *Biochem. Biophys. Res. Commun.* **173** (1990) 1157.
1971. M. Mogard et al., *Gastroenterology* **86** (1984) part 2, 1186.
1972. R. Markenstein, P. Emson, *Europ. J. Pharmacol.* **152** (1988) 147.
1973. E. Kislauskis et al., *J. Biol. Chem.* **263** (1988) 4963.
1974. J.-M. Studler et al., *Neuropeptides (Edinburgh)* **11** (1988) 95.
1975. A. Rökaeus et al., *Acta Physiol. Scand.* **120** (1984) 373.
1976. K. M. Merchant et al., *J. Pharmacol. Exp. Ther.* **255** (1990) 775.
1977. K. W. Eggerman, D. S. Zahm, *Neuropeptides (Edinburgh)* **11** (1988) 125.
1978. G. R. Uhl, M. J. Kuhar, *Nature (London)* **309** (1984) 350.
1979. T. F. Lee et al., *Am. J. Physiol.* **252** (1987) E102.
1980. D. L. Barber et al., *Gastroenterology* **92** (1987) part 2, 1305.
1981. A. Rökaeus, *Acta Physiol. Scand.* **122** (1984) 261.
1982. E. Theodorsson-Norheim et al., *Acta Physiol. Scand.* **123** (1985) 331.
1983. N. A. Colliere et al., *Lancet* (1984) 538.
1984. T. W. Moody et al., *Life Sci.* **36** (1985) 1727.
1985. V. E. Eysselein, *Klin. Wochenschr.* **62** (1984) 523.
1986. K. Muraki et al., *Life Sci.* **40** (1987) 1365.

1987. J. Staley et al., *Peptides (Fayetteville, N.Y.)* **10** (1989) 1217.
1988. S. Amar et al., *Biochem. Biophys. Res. Commun.* **129** (1985) 117.
1989. P. L. Canonico et al., *Mol. Cell. Endocrinol.* **42** (1985) 215.
1990. A. D. Drumheller et al., *Neuropeptides (Edinburgh)* **15** (1990) 169.
1991. F. B. Jolicoer et al., *Neuropeptides (Edinburgh)* **6** (1985) 143.
1992. C. F. Ferris et al., *Neuroendocrinology* **38** (1984) 145.
1993. G. A. Gudelsky et al., *Neuroendocrinology* **49** (1989) 604.
1994. S. Furuta et al., *Br. J. Pharmacol.* **83** (1984) 43.
1995. L. Zhang et al., *Gastroenterology* **92** (1987) part 2, 1709.
1996. D. E. Hermandez et al., *Regul. Pept.* **8** (1984) 41.
1997. S. Rosell, *Acta Endocrinol. (Copenhagen)* **103** (1983) Suppl. 256, 44.
1998. J.-L. Sadoul et al., *Biochem. Biophys. Res. Commun.* **120** (1984) 812.
1999. I. V. Grinshteine et al., *Bioorg. Khim.* **11** (1985) 1167.
2000. S. Furuta et al., *Br. J. Pharmacol.* **83** (1984) 43.
2001. D. Hernandez et al., *Brain Res.* **301** (1984) 153.
2002. F. B. Jolicoer et al., *Neuropeptides (Edinburgh)* **6** (1985) 143.
2003. F. Rioux et al., *Neuropeptides (Edinburgh)* **3** (1983) 345.
2004. P. L. Canonico et al., *Mol. Cell. Endocrinol.* **42** (1985) 215.
2005. J. E. T. Fox, *Life Sci.* **35** (1984) 695.
2006. Y. Seino et al., *FEBS Lett.* **223** (1987) 74.
2007. T. Peeters et al., *Am. J. Physiol.* **257** (1989) G470.
2008. N. Inatomi et al., *J. Pharmacol. Exp. Ther.* **251** (1989) 707.
2009. K. E. Hall et al., *Gastroenterology* **87** (1984) 76.
2010. M. Lemoyne et al., *Can. J. Physiol. Pharmacol.* **62** (1984) 1092.
2011. N. S. Track, P. L. Carlen, *Peptides (Fayetteville, N.Y.)* **7** (1986) 1193.
2012. P. Poitras et al., *Gastroenterology* **92** (1987) 658.
2013. E. L. Muller et al., *Clin. Res.* **32** (1984) 692A.
2014. K. Y. Lee et al., *Dig. Dis. Sci.* **30** (1985) 980.
2015. T. L. Garthwaite, *Peptides (Fayetteville, N.Y.)* **6** (1985) 41.
2016. S. Kuno et al., *Chem. Pharm. Bull.* **34** (1986) 4811.
2017. K. Kitagawa et al., *Chem. Pharm. Bull.* **33** (1985) 3307.
2018. L. Thim et al., *Biochim. Biophys. Acta P* **827** (1985) 410.
2019. L. Thim, *FEBS Lett.* **250** (1989) 85.
2020. S. B. Jakowlew et al., *Nucleic Acids Res.* **12** (1984) 2861.
2021. K. Mori et al., *Biochem. Biophys. Res. Commun.* **155** (1988) 366.
2022. M. Gmachl et al., *FEBS Lett.* **260** (1990) 145.
2023. W. Hoffmann, *J. Biol. Chem.* **263** (1988) 7686.
2024. E. K. Frandsen et al., *Regul. Pept.* **16** (1986) 291.
2025. J. F. Rey et al., *Gut* **26** (1985) A581.

Peptide Synthesis

Wolfgang Heinzel, Bissendorf Biochemicals GmbH, Hannover, Federal Republic of Germany

Michael S. Verlander, Bachem Inc., Torrance, California 90505, United States

1. Introduction 1
2. Fundamentals of Peptide Synthesis 1
3. Protecting Groups 2
3.1. Amine Protecting Groups 2
3.2. Carboxyl Protecting Groups 4
3.3. Sulfhydryl Protecting Groups 5
3.4. Hydroxyl Protecting Groups 5
3.5. Imidazole Protecting Groups 5
3.6. Protecting Groups for Guanidine . 6
3.7. Other Protecting Groups 7
4. Formation of the Peptide Bond . . . 7
4.1. Active Ester Method 7
4.2. Anhydride Method 9
4.3. Azide Method 9
4.4. Carbodiimide Method 10
5. Strategies for Peptide Synthesis . . . 10
5.1. Synthesis in Solution 10
5.2. Solid-Phase Synthesis 11
5.3. Other Synthetic Methods 12
5.3.1. Liquid-Phase Synthesis 12
5.3.2. Semisynthesis 12
5.3.3. Enzymatic Synthesis 12
5.3.4. Recombinant DNA-Based Synthesis . 12
5.3.5. Concluding Remarks 13
6. Cleavage of Protecting Groups . . . 14
7. Purification of Peptides 14
8. Analytical Control of Peptides . . . 15
9. Important Commercial Peptides . . 15
10. References 15

1. Introduction

Peptides are chains of amino acids linked together by amide bonds. Peptides with up to ten amino acids are generally termed *oligopeptides* ($M_r \leq$ ca. 1000). *Polypeptides* contain 10 – 100 amino acids (M_r ca. 1000 – 10 000). *Proteins* contain > 100 amino acids ($M_r > 10\,000$).

Peptides have a broad range of applications, particularly in diagnostic and therapeutic areas, because they play important roles in the regulation and maintenance of biological processes. Demand for these products is already considerable and will probably increase dramatically in the future. Peptides may be synthesized by both chemical and biotechnological methods. Much of the emphasis in the field has been directed toward the development of methods for the synthesis and purification of peptides in high yield and in the most rapid and economical way.

2. Fundamentals of Peptide Synthesis [1–8]

Peptides are formed by the condensation of α-amino acids as in the following general reaction:

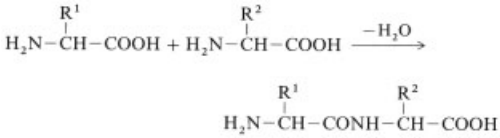

Since amino acids contain at least two functional groups (the α-amino and α-carboxyl groups), certain precautions must be taken to obtain a well-defined peptide. Thus, the amino group of one amino acid and the carboxyl group of the other must be blocked or "protected" to prevent undesirable side reactions. For the same reason, other functional groups in the amino acid side chains (R^1 and R^2) should also be protected.

The protecting groups used for side-chain functional groups must be stable during the procedures which are used to form the peptide bonds in the chain. Conversely, the protecting groups used for the α-amino and α-carboxyl groups must be more temporary in nature, so that they can be removed without cleaving the groups protecting the side chains. Only if these conditions are met can the α-amino and α-carboxyl groups react unequivocally in the condensation step. Once assembly of the peptide has been com-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_157

pleted, all of the protecting groups must be removed without significant damage to the final, desired peptide.

Formation of the peptide bond requires activation of one of the two functional groups, generally the carboxyl group. Activation is usually accomplished by conversion to a reactive species, in which the carboxyl group is susceptible to nucleophilic attack by the amine group of the second amino acid residue.

All α-amino acids except glycine have at least one asymmetric center – the α-carbon atom. Peptides and proteins isolated from natural sources almost always contain amino acids of the L- configuration. Unfortunately, racemization can occur during the activation step, the extent depending on the nature of the amino acid side chain and the protecting groups used. Thus the suppression of racemization is an extremely important concern in planning the synthesis of a peptide.

3. Protecting Groups

An enormous variety of protecting groups have been developed to solve particular problems arising during peptide synthesis. However, of these, only a relatively small number are currently in use.

3.1. Amine Protecting Groups

Some amine protecting groups are listed in Table 1 (see also [1–3]). The most important are the *tert*-butoxycarbonyl (Boc), benzyloxycarbonyl (Z), and fluorenylmethoxycarbonyl (Fmoc) groups. These groups are cleaved under a wide variety of conditions thus allowing a very broad range of applications.

The development of the *Z protecting group* by ZERVAS and BERGMANN in 1932 [9] was a milestone in peptide chemistry. Z-Protected amino acid derivatives are prepared by the Schotten–Baumann reaction:

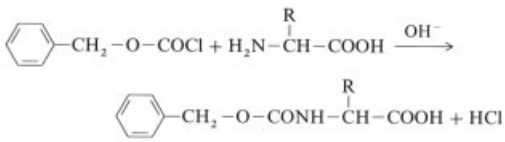

Nowadays the Z group is introduced by using *N*-(benzyloxycarbonyloxy)succinimide [*13139-17-8*], because the formation of oligopeptides can then be avoided. The protecting group may be removed either by hydrogenation or under anhydrous, strongly acidic conditions (e.g., with hydrobromic acid in acetic acid), giving innocuous byproducts which are removed easily:

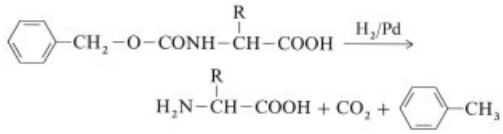

Boc-amino acids were originally prepared by the treatment of amino acids with *tert*-butyl azidoformate. However, this reagent is no longer available commercially and has been replaced by *tert*-butyl dicarbonate [*24424-99-5*] (**1**):

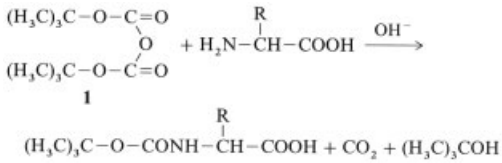

The Boc group is cleaved under acidic conditions, most commonly with trifluoroacetic acid (TFA).

Attempts to prepare *Fmoc-amino acid derivatives* using 9-fluorenylmethyl chlorocarbonate [*28920-43-6*] often lead to the formation of substantial amounts of Fmoc-oligopeptides [10]. However, this side reaction may be avoided by using Fmoc-succinimidyl carbonate [*82911-69-1*], the reagent of choice today.

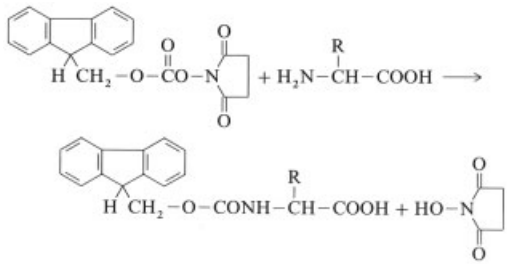

Removal of the Fmoc group is accomplished by base-catalyzed elimination:

Table 1. Amine protecting groups

Protecting group (abbreviation)	Structure	Cleavage agents	Stability
Trifluoroacetyl (TFA)	F_3C-CO-	base	acids, H_2/Pd
Phthaloyl (Pht)	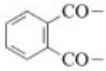	hydrazine	H_2/Pd, CF_3COOH
Dithiasuccinoyl (Dts)	$S-CO-$ $S-CO-$	thiols	H_2/Pd
Benzyloxycarbonyl (Z)	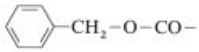	H_2/Pd, HBr/CH_3COOH	CF_3COOH, base
4-Methoxybenzyloxycarbonyl (Moz)	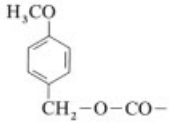	CF_3COOH, H_2/Pd	base
4-Dihydroxyboronatobenzyloxy-carbonyl (Dobz)	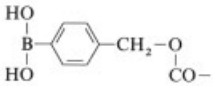	H_2/Pd, H_2O_2, HBr/CH_3COOH	CF_3COOH, base
tert-Butoxycarbonyl (Boc)	$(H_3C)_3C-O-CO-$	CF_3COOH	H_2/Pd, base
9-Fluorenylmethoxycarbonyl (Fmoc)	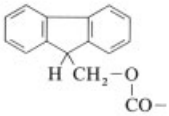	base, H_2/Pd	acids
2-(3,5-Dimethyloxyphenyl)propyl-2-oxycarbonyl (Ddz)	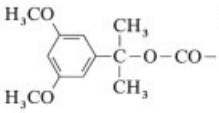	5% CF_3COOH, CH_2Cl_2, $h\nu$	H_2/Pd, base
2-Nitrophenylsulfenyl (Nps)	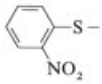	weak acids	
Trityl (Trt)	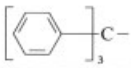	H_2/Pd, acids	base

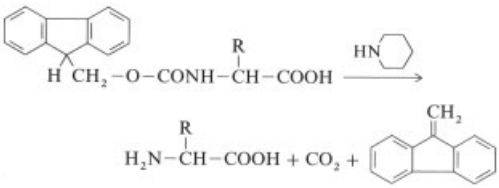

The Fmoc group has become increasingly popular in solid-phase peptide synthesis because of its facile removal under so-called "orthogonal" conditions (i.e., those under which sidechain protecting groups are completely stable).

The incorporation of lysine into the growing peptide chain requires the protection of both the α-amino group and the ε-amino group of the side chain. In most cases, two different protecting groups are required. The most widely used combinations are Boc-Lys (2-Cl-Z)-OH [*54613-99-9*] (**2**), Z-Lys (Boc)-OH [*2389-60-8*] (**3**), and Fmoc-Lys (Boc)-OH [*71989-26-9*] (**4**).

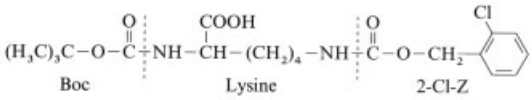

2

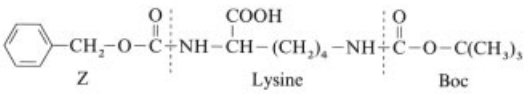

3

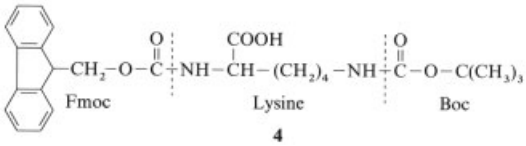

4

The 2-chlorobenzyloxycarbonyl (2-Cl-Z) group is an example of a more acid-stable ε-amine protecting group. Such groups are more stable to the acid conditions used for cleavage of the Boc group in solid-phase synthesis and are only cleaved by hydrogen fluoride.

3.2. Carboxyl Protecting Groups [1–3]

Some of the wide variety of available carboxyl protecting groups are listed in Table 2. Although, in principle, peptide synthesis can be carried out without protecting the carboxyl group(s) of the amine component, protection of these groups is advisable in most cases.

Simple alkyl esters (e.g., methyl and ethyl esters) may be prepared by the Fischer esterification method (refluxing the amino acid in the alcohol saturated with hydrogen chloride). An alternative method involves the addition of thionyl chloride to the cold alcohol, followed by addition of the amino acid. For other preparations see [2] and literature cited therein. The protecting groups are removed by saponification; complications can occur with longer peptides because saponification may be incomplete and racemization may occur. *Benzyl and 4-nitrobenzyl esters* are also cleaved by hydrogenation, which is an attractive advantage compared with alkyl esters. Benzyl esters may be prepared by reaction of amino acids with benzyl alcohol in the presence of an acidic catalyst, such as hydrogen chloride or 4-toluene sulfonic acid:

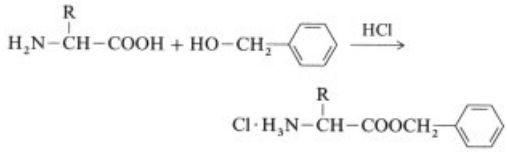

Another very important protection method is formation of the *tert-butyl ester,* which is especially useful when used in combination with the Fmoc and Z protecting groups. It is most widely used for protection of the acidic side chains of aspartic acid and glutamic acid. *tert*-Butyl esters are resistant to nucleophiles and are unaffected by treatment with base or hydrogenation. They may be prepared by acid-catalyzed addition of isobutene:

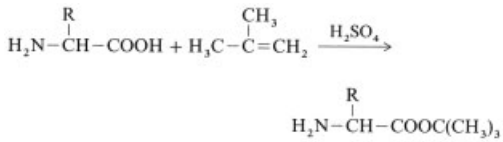

Cleavage of *tert*-butyl esters is accomplished under similar conditions to removal of the Boc group, i.e., by treatment with anhydrous acids, such as trifluoroacetic acid, HCl, or HBr in organic solvents.

3.3. Sulfhydryl Protecting Groups [1–3]

The sulfhydryl side chain of cysteine must be protected because of problems encountered with its nucleophilicity and also the possibility of its oxidation to disulfides. Sulfhydryl protecting groups are listed in Table 3. Preferred protecting groups are the benzyl (Bzl), trityl (Trt), and acetamidomethyl (Acm) groups. The trityl and acetamidomethyl groups can be removed under relatively mild conditions (i.e., with mercuric ions at pH 4, or with iodine). The trityl group is also removed by acid and may be selectively oxidized with iodine in the presence of the acetamidomethyl group. On the other hand, the acetamidomethyl group is stable to hydrogen fluoride, which cleaves the benzyl group. Substituted benzyl groups (e.g., 4-methyl- or 4-methoxybenzyl groups) are used widely in solid-phase synthesis. Peptides containing more than one disulfide bond require selectively cleavable sulfur protecting groups to ensure the correct

Table 2. Carboxyl protecting groups

Protecting group (abbreviation)	Structure*	Cleavage agents	Stability
Methyl ester (OMe)	$H_3C-O-CO-$	base	acids, H_2/Pd
tert-Butyl ester (OtBu)	$(H_3C)_3C-O-CO-$	CF_3COOH	base, H_2/Pd
Benzyl ester (OBzl)	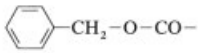	H_2/Pd, base	CF_3COOH
4-Nitrobenzyl ester (ONb)	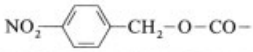	H_2/Pd, base	CF_3COOH, HBr/CH_3COOH

∗ After peptide synthesis the ester is cleaved to form the relevant alcohol and amino acid carboxyl group.

Table 3. Sulfhydryl protecting groups

Protecting group (abbreviation)	Structure	Cleavage agents	Stability
Benzyl (Bzl)	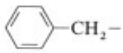	HF, Na/NH_3	acids, iodine
4-Methoxybenzyl (Mob)		HF, Na/NH_3	CF_3COOH, iodine
Acetamidomethyl (Acm)	$H_3C-CONH-CH_2-$	iodine, Hg^{2+}	HF, base
S-tert-Butylthio (StBu)	$(H_3C)_3C-S-$	RSH or other reducing reagents	CF_3COOH, HF, HBr/CH_3COOH
Trityl (Trt)	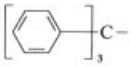	CF_3COOH, thiols, Hg^{2+}	base, weak acids
4-Methylbenzyl (MeBzl)	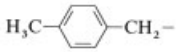	HF	CF_3COOH
9-Fluorenylmethyl (Fm)	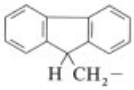	NH_3/CH_3OH, piperidine	acids

formation of the desired disulfide bonds. An example of this approach is the synthesis of insulin [11].

3.4. Hydroxyl Protecting Groups [1–3]

The side-chain hydroxyl groups of serine, threonine, and tyrosine need not be protected during peptide synthesis because of their poor nucleophilicity. By careful choice of reaction conditions, acylation of the hydroxyl groups (for example, by activated amino acids) can be avoided. However, additional problems may occur with tyrosine because of the aromatic ring system. To avoid possible problems, the hydroxyl groups are best protected, usually as benzyl or *tert*-butyl ethers: the choice depends on the type of amino protection used and the conditions used for final deprotection. The *tert*-butyl group can be removed by acids, such as trifluoroacetic acid, while the benzyl group is liberated by hydrogenolysis, hydrofluoric acid, and sodium in liquid ammonia.

3.5. Imidazole Protecting Groups [1–3]

Protection of the secondary amine in the imidazole ring of histidine is not essential, though often desirable. Even though nitrogen is a relatively weak nucleophile, acylation can occur and the unprotected imidazole ring can catalyze the hydrolysis of ester and amide bonds and also racemization, especially if the histidine is the C-terminal residue. Of the histidine (imidazole) protecting groups listed in Table 4, only the

Table 4. Imidazole protecting groups

Protecting group (abbreviation)	Structure	Cleavage agents	Stability
Benzyl (Bzl)	$-CH_2-$	H_2/Pd, Na/NH_3	acids, base
Trityl (Trt)	$[-C_6H_5]_3C-$	CF_3COOH, H_2/Pd, HBr/CH_3COOH	weak acids
2,4-Dinitrophenyl (Dnp)	$O_2N-C_6H_4-NO_2$	thiols, hydrazine	acids
tert-Butoxycarbonyl (Boc)	$(H_3C)_3C-O-CO-$	HF, HBr/CH_3COOH	base, HCl/dioxane
Adamantyloxycarbonyl (Adc)	$C_{10}H_{15}-O-CO-$	CF_3COOH	HCl/dioxane
4-Toluenesulfonyl (Tos)	$H_3C-C_6H_4-SO_2-$	HF, Na/NH_3	H_2/Pd
π-*tert*-Butoxymethyl (Bum)	$(H_3C)_3COCH_2-$	weak acids	H_2/Pd, base
π-Benzyloxymethyl (Bom)	$-CH_2OCH_2-$	H_2/Pd, HF	weak acids

π-benzyloxymethyl [12], π-*tert*-butoxymethyl [13], and trityl [14] groups seem to effectively suppress racemization.

The type of protecting group used for histidine depends on the planned strategy. Much of the peptide synthesis in solution is carried out without protection of histidine. In solid-phase synthesis, however, this amino acid is almost always protected, for example, by the tosyl (Tos) or 2,4-dinitrophenyl (Dnp) groups.

3.6. Protecting Groups for Guanidine [1–3]

The unusually basic guanidino group of arginine can be "protected" by protonation, for example, as a hydrochloride or hydrobromide salt. This is usually sufficient to prevent side reactions under the conditions used for peptide synthesis in solution. However, arginine salts have poor solubility in organic solvents, which often prompts the use of protected derivatives. For the same reason, protected arginine derivatives, such as the 4-toluenesulfonyl (Tos) or nitro derivatives, are usually used in solid-phase synthesis. Several new protecting groups have been developed, especially for use in combination with the Fmoc group [15–17] (see Table 5). The specific choice of protecting groups for the guanidino function depends on the strategy being used to synthesize the desired peptide.

3.7. Other Protecting Groups

The primary amide groups of *asparagine* and *glutamine* are relatively unreactive and, therefore, in theory, do not need to be protected. However, the use of protecting groups offers the advantage of increased solubility, as well as avoidance of dehydration of the amide function during activation. The most commonly used protecting groups for this purpose are the xanthyl and 4,4′-dimethoxybenzhydryl groups [1–3], both of which are removed under acidic conditions.

Methionine, although rarely thought of as being reactive, can undergo S-alkylation. This can be avoided through use of the sulfoxide, which is converted back to the thioether with 1,4-dimercapto-2,3-butanediol [*27565-41-9*] (Cleland's reagent) [18].

The secondary amine function of the indole group of *tryptophan* is oxidized relatively easily and may also be alkylated at acidic pH. The only useful method of protection is the formyl (For)

Table 5. Guanidino protecting groups

Protecting group (abbreviation)	Structure	Cleavage agents	Stability
Nitro (NO_2)	O_2N-	HF, H_2/Pd	acids, bases
Benzyloxycarbonyl (Z or Z_2)	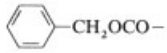	H_2/Pd	acids, bases
Adamantyloxycarbonyl (Adc)	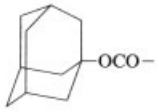	CF_3COOH	H_2/Pd, base
4-Toluenesulfonyl (Tos)	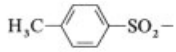	HF, Na/NH_3	H_2/Pd, CF_3COOH
4-Methoxybenzenesulfonyl (Mbs)	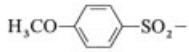	HF	H_2/Pd, CF_3COOH
2,4,6-Trimethylbenzenesulfonyl (Mts)	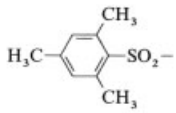	CF_3COOH	H_2/Pd, bases
2,3,4,5,6-Pentamethylbenzenesulfonyl (Pms)	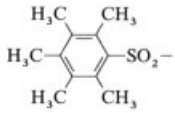	CF_3COOH	H_2/Pd, bases
4-Methoxy-2,3,6-trimethylbenzenesulfonyl (Mtr)	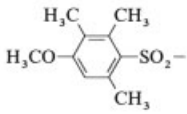	CF_3COOH	H_2/Pd, bases
2,2,5,7,8-Pentamethylchroman-6-sulfonyl (Pmc)	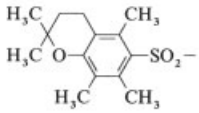	CF_3COOH	H_2/Pd, bases

group, which is stable under the acidic conditions used during peptide synthesis. Deprotection is accomplished through use of hydrazine in dimethylformamide [19].

4. Formation of the Peptide Bond

The principal reaction in peptide synthesis is the acylation of the amino group of one amino acid by the carboxyl group of a second amino acid, resulting in the formation of a peptide bond. However, because of the diversity of structure of amino acids, peptide bond formation is far more complicated than simple carboxamide formation. For this reason, a variety of methods of peptide bond formation have been developed, based on enhancement of the electrophilicity of the carboxyl carbon atom by using electronegative substituents. Aminolysis of the activated amino acid derivatives then results in peptide bond formation. This reaction may be divided into two steps:

$$\text{Activation:}\quad R-\overset{O}{\overset{\|}{C}}-OH \longrightarrow R-\underset{(\delta+)}{\overset{O}{\overset{\|}{C}}}-X\ (\delta-) \qquad (1)$$

$$\text{Aminolysis:}\quad R-\underset{(\delta+)}{CO}-X\ (\delta-) + H_2\ddot{N}-R' \xrightarrow{-HX} R-CONH-R' \qquad (2)$$

Practically, the reaction may be accomplished in three principal ways:

1) Activation of the carboxyl group followed by aminolysis
2) Isolation and purification of the activated carboxyl component, followed by aminolysis
3) Activation of the carboxyl component through use of a coupling reagent in the presence of the amine component

Some of the most commonly used methods follow.

4.1. Active Ester Method

Active esters of protected amino acids are prepared by the following general scheme:

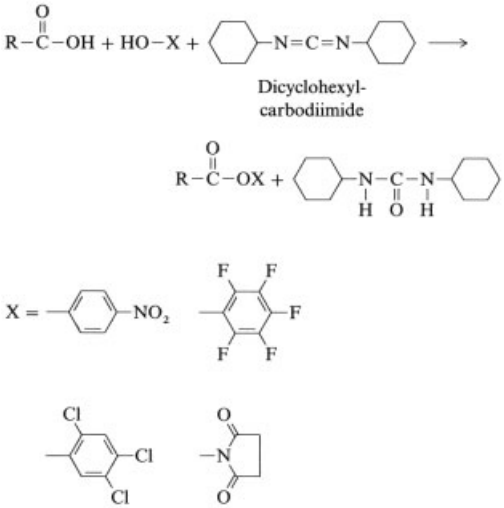

The greater the degree of activation of the carboxyl group, the faster the subsequent aminolysis. For this reason, virtually all of the useful active esters are of the aryl type; their activities vary widely as a function of the substituents on the aromatic ring. The most commonly used active esters are listed in Table 6. The *N*-succinimidyl ester has been used extensively because of its ability to suppress the racemization that can occur during activation and coupling. The sterically hindered pentachlorophenyl ester has largely been replaced by the pentafluorophenyl ester, especially in conjunction with Fmoc protection. However, both of these esters are expensive and they also liberate pentachloro- and pentafluorophenol as byproducts, which are a potential source of dioxins. 8-Oxo-1,8-dihydrobenzo-1,2,3-triazine esters are not often used on a large scale [20].

Table 6. Active esters

Active ester group (abbreviation)	Structure
4-Nitrophenyl ester (-ONp)	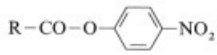
2,4,5-Trichlorophenyl ester (-OTcp)	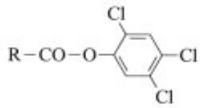
Pentachlorophenyl ester (-OPcp)	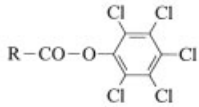
Pentafluorophenyl ester (-OPfp)	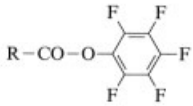
N-Succinimidyl ester (-OSu)	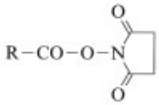
8-Oxo-1,8-dihydrobenzo-1,2,3-triazine ester (-ODhbt)	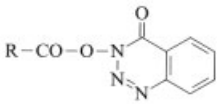
Benzo-1,2,3-triazole ester (-OBt)	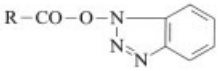

4.2. Anhydride Method

Two types of anhydride are used in peptide synthesis: symmetrical anhydrides and mixed anhydrides.

Symmetrical anhydrides (which are generally formed by reacting the protected amino acid with half an equivalent of carbodiimide) react in the following manner:

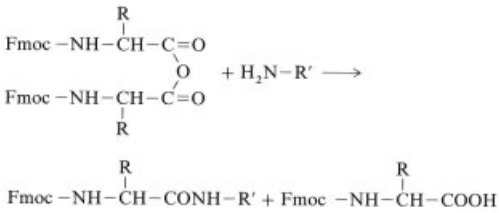

Although this procedure works well, it is relatively expensive because only one of the two amino acid molecules in the anhydride is incorporated into the growing peptide chain.

In the *mixed anhydride method*, two possible reactions can occur with the amine component:

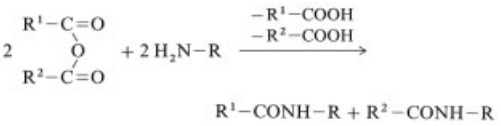

Since the second carboxyl component (R^2-COOH) of the mixed anhydride is not usually derived from an amino acid, reaction of the amine component at this carbonyl group does not give the desired peptide product. Furthermore, since reaction occurs preferentially at the more positively charged and less sterically hindered carboxyl group, mixed anhydrides can be designed to promote the formation of the desired peptide, for example:

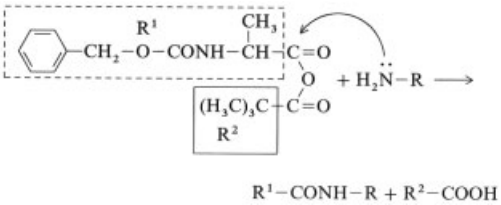

In this case, the R^1 group (derived from the protected amino acid) is electron-withdrawing, while the R^2 group is a sterically hindered, electron-donating group. If conditions are chosen carefully so that the reaction occurs rapidly, little or no racemization occurs.

4.3. Azide Method

The azide method has proved to be one of the most important developments in the peptide field and may be considered the starting point for modern peptide synthesis. Even today, it is still one of the most important coupling methods, because it has significant advantages:

1) Low potential for racemization
2) Ease of crystallization, allowing purification of intermediates
3) Useful in syntheses with minimum side chain protection
4) Useful for the assembly of large peptide fragments

The azide procedure is carried out in three steps:

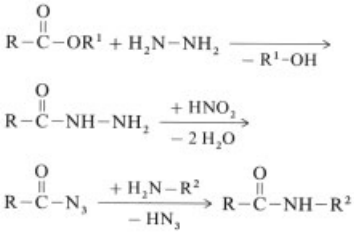

Despite possible side reactions, the azide method has been used widely in peptide fragment condensations, the optimal size range being peptides of up to about 50 amino acids. Special applications of this method include peptide cyclizations and the synthesis of polyamino acids.

4.4. Carbodiimide Method

The carbodiimide method, based on the use of dicyclohexylcarbodiimide (DCC) [*538-75-0*], effectively results in the elimination of water to form the peptide bond:

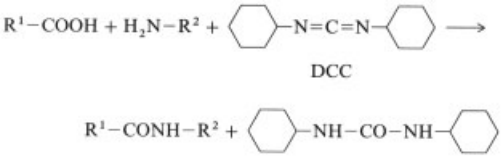

DCC is a highly reactive reagent, which reacts faster with the carboxyl groups than with the amine components. The *N,N′*-dicyclohexylurea formed during the reaction has low solubility in solvents used for coupling reactions and is therefore easily removed as a precipitate. Unfortunately, the method is accompanied by a number of side reactions, including:

1) Racemization
2) Partial dehydration of the carboxamide groups of asparagine and glutamine to nitrile groups
3) Formation of *N*-acylureas

Investigations of methods to suppress these side reactions have led to the use of α-dinucleophiles, such as *N*-hydroxysuccinimide (**5**) [*6066-82-6*], 1-hydroxybenzotriazole (**6**) [*2592-95-2*], and 1-hydroxy-8-oxo-1,8-dihydrobenzo-1,2,3-triazine (**7**) [*28230-32-2*]:

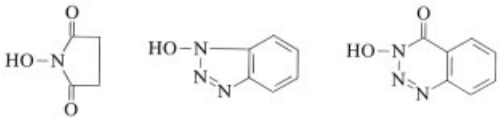

5 **6** **7**

The combination of dicyclohexylcarbodiimide and 1-hydroxybenzotriazole is still the most widely preferred coupling method, both in solid-phase and solution synthesis.

5. Strategies for Peptide Synthesis

Due to the wide variety of protecting groups and synthetic methods, the synthetic route for a particular peptide must be chosen carefully. The actual method used (solid-phase synthesis, synthesis in solution, stepwise or fragment condensation, or other methods) depends on a number of factors, including the length and sequence of the peptide and the scale of the synthesis.

5.1. Synthesis in Solution

Many of the peptides that are required in large quantities (e.g., glucagon, γ-endorphin) are still synthesized by solution methods. Once the synthetic route has been elaborated, production can be carefully controlled because intermediates are isolated and characterized at each step. Generally, scaleup of this type of synthesis is relatively straightforward. In principle, the method can be applied in two ways:

1) Repeated, stepwise addition of single amino acid residues to the growing peptide chain
2) Synthesis of segments of the peptide which are subsequently assembled to give the desired sequence (fragment condensation).

Stepwise Synthesis. Protecting groups of the urethane type effectively suppress racemization of amino acids during activation and coupling (Section 3.5), forming the basis of a highly efficient method for synthesizing peptides. The applicability of the method was first demonstrated successfully in the synthesis of oxytocin (9 amino acids) [21], in which *N*-α-Z-protected amino acid 4-nitrophenyl esters were used. Protected intermediates were purified by washing and crystallization. The process was relatively efficient – a new amino acid residue could be introduced every two days. The same process has been applied successfully to the synthesis of even larger peptides, such as secretin (27 amino acids) [22] and gastrin (17 amino acids) [23]. However, this approach does present some problems, especially as the peptide chain becomes longer. First, the effective molar concentration of the amine component decreases with the increase in molecular mass of the growing peptide chain thus necessitating longer coupling times. Secondly, as the size of the protected peptide increases, the solubility often decreases dramatically in the solvents commonly used for peptide synthesis (e.g., dichloromethane, dioxane, tetrahydrofuran, dimethylformamide).

Fragment Condensation. Because of some of the problems discussed above, an alternative strategy has been developed in which a series of fragments of the desired sequence (up to 10 or sometimes 15 amino acid residues) are synthesized in parallel. Even though this minimizes some of the problems, especially those relating to the insolubility of large, protected peptides, final assembly of the fragments still requires considerable preparative skill. The method is often limited by low yields in the fragment coupling steps (often due to insolubility), although use of amino acid derivatives with solubilizing protecting groups may be helpful. Unfortunately, it is virtually impossible to predict the efficiency of a synthesis because of the lack of knowledge concerning the relationship between sequence and physicochemical properties of protected peptides. Even so, the fragment condensation approach is useful for the synthesis of peptides of all sizes; optimization can lead to economical production methods in which reliable quality control of intermediates can be carried out.

5.2. Solid-Phase Synthesis

Synthesis of peptides in solution can be extremely time consuming due to the necessity of purifying and characterizing the intermediates, as well as the final product. In 1963, MERRIFIELD introduced a very powerful, alternative strategy, in which the growing peptide chain was attached to a swollen, insoluble polymeric support throughout the synthesis [24]. The matrix acts as a tool for purification; simple filtration and washing steps can be used to remove excess reagents and byproducts before continuing to the next step. Furthermore, reactions can be carried out under optimal, "forcing" conditions through use of relatively large molar excesses of reagents, resulting in reduced reaction times. Because the technique is simple and rapid and easily automated, it has found widespread application.

At the present time, the most commonly used polymeric supports are the polystyrene-based resins, originally introduced by MERRIFIELD, and the more recently developed polyamide supports. The C-terminal amino acid in the sequence is attached to the polymeric support via a benzyl ester linkage (or, in the case of a C-terminally amidated sequence, via a labile amide linkage) and amino acids are added repetitively to build up the desired sequence. In the original Merrifield approach, based on N-α-Boc and benzyl–based side-chain protection, the following cycle is used:

1) Cleavage of the Boc group with 50 % trifluoroacetic acid in dichloromethane
2) Neutralization with 10 % triethylamine in dichloromethane
3) Coupling of Boc-amino acids using dicyclohexylcarbodiimide in dichloromethane.

The assembled peptide is cleaved from the resin with anhydrous, liquid hydrogen fluoride, simultaneous cleavage of the side-chain protecting groups gives the crude product. However, hydrogen fluoride can lead to side reactions and many improvements have therefore been made to the original method, such as the development of new protecting groups and resins and methods for monitoring coupling steps and suppressing side reactions during the hydrogen fluoride treatment step [1], [4]. Alternative protecting group combinations which do not require the use of hydrogen fluoride (e.g., that based on the Fmoc group) are also receiving increased attention. Even though the solid-phase technique is rapid and convenient, it is not completely routine and has inherent limitations. Since there are no intermediate purification steps, each coupling step must be as complete as possible. Incomplete reaction or side reactions can lead to the formation of side products (e.g., "failure sequences" or "truncated peptides") which must be removed during the final purification step. Purification of the final product can therefore present a considerable challenge, although the introduction of preparative, reverse-phase high performance liquid chromatography (HPLC) has made the solid-phase technique extremely powerful. A number of products (e.g., calcitonin, somatostatin) are now manufactured on a large scale for use as pharmaceuticals using this technique.

5.3. Other Synthetic Methods

5.3.1. Liquid-Phase Synthesis

The liquid-phase method combines the advantages of solid-phase synthesis and synthesis in solution [25]. As with the solid-phase method, the C-terminal amino acid in the sequence is covalently attached to a polymer, in this case a soluble poly(ethylene glycol). Synthesis is carried out by stepwise addition of amino acid derivatives in homogeneous solution because of the soluble nature of the polymeric support. Because the physical properties of the peptide are dominated by the polymer, few of the solubility problems associated with solution syntheses are experienced in this method. Furthermore, intermediate purification can be effected by precipitation of polymer-bound peptides and washing to remove excess reagents. Aggregation problems are sometimes encountered when the growing peptide chain forms secondary structures. Although the method has attractive advantages, it has not gained widespread acceptance.

5.3.2. Semisynthesis

Selective modification of natural sequences can provide an approach for obtaining new

molecules. Although this method has largely remained a research tool, a notable example of its use on a commercial scale to manufacture a pharmaceutical product is the conversion of porcine to human insulin [26]. This transformation involves replacement of the C-terminal alanine residue of the B-chain of the porcine sequence with a threonine residue. This is accomplished in a two-step process by cleaving the alanine residue and subsequently attaching the threonine [27]; or directly, in a single-step transpeptidation reaction, catalyzed, for example, by trypsin [26]. The semisynthesis approach is likely to find its widest application in combination with the recombinant method (see Section 5.3.4) for the production of analogues of peptides or proteins which cannot be produced directly via the recombinant approach.

5.3.3. Enzymatic Synthesis

Although the use of proteases (e.g., papain and pepsin) as catalysts for the formation of peptide bonds was reported in 1937 [28], the method has only been studied extensively since the early 1980s. This method offers a number of advantages over chemical methods [29] including the use of unprotected or partially protected amino acids, without the need for side-chain protecting groups as well as racemization-free coupling because of the stereoselectivity of the enzymes used. However, in practice, a number of problems make this method less successful than it initially promised to be. Each coupling reaction in the enzymatic method is unique and must therefore be optimized. The method is not universal and certain couplings, such as that between cysteine and proline, cannot be carried out [30]. D-Amino acids are also difficult or impossible to couple. However, the fact that enzymatic synthesis is normally carried out under aqueous conditions or in a mixture of aqueous and organic solvents, may lead to important applications in which chemical and enzymatic methods are combined.

5.3.4. Recombinant DNA-Based Synthesis

One of the newest methods for peptide production, which has only been available since the early 1970s, is the "recombinant DNA approach". This method has been the subject of extremely intensive investigation because it offers tremendous potential for the production of both peptides and proteins. Genetically engineered microorganisms are used to produce large quantities of peptides and proteins through their normal metabolic processes (→ Genetic Engineering).

The sequence for a particular peptide is encoded in a gene which can either be obtained from a DNA library or synthesized rapidly in small quantities. This gene is then inserted into a host plasmid DNA by using highly specific enzymes. A suitable microorganism, such as yeast or *Escherichia coli*, must be found in which the gene is expressed, i.e., used as a template to synthesize the desired peptide or protein. The product must then be extracted and purified.

The majority of syntheses which have been carried out on a production scale appear to have concentrated on molecules which are not readily accessible by other routes, in particular proteins such as human growth hormone [31], interferons [32], and human insulin [33]. Although problems still exist, especially for the production of small peptides (poor expression, toxicity, bacterial inclusions, proteolytic degradation), these will probably be solved in the near future and this approach may therefore be the method of choice. Chemical synthesis is, however, still likely to remain the most practical approach for very small peptides and peptides containing unnatural sequence modifications.

5.3.5. Concluding Remarks

A tremendous number of peptides have been discovered (→ Peptide and Protein Hormones). These peptides, isolated from the pancreas, gut, lungs, and even the heart, were recognized as new regulators with great potential for pharmaceutical use. Thus, the need for synthetic peptides has become even greater than ever. Of the methods which are available for the production of peptides, there is no doubt that the recombinant approach will eventually play the most important role. Meanwhile, enzymatic synthesis has not achieved any significance as a production method. Semisynthesis is employed for the production of insulin (Nova and Hoechst). Chemi-

Table 7. Examples of peptides of pharmaceutical significance

Peptide	No. of amino acids	Use	Trade name (company)
Atrial natriuretic factor	28	hypertension	
Arginine-vasopressin	9	antidiuretic	
Calcitonin (salmon)	32	treatment of osteoporosis	Karil (Sandoz) Calsynar (Rorer)
Calcitonin (eel)	32	treatment of osteoporosis	
Corticotropin	39	treatment of inflammation	Acethropan (Hoechst)
Corticotropin releasing factor	41	diagnosis of Cushing's syndrome	
Insulin	51	treatment of diabetes	(Hoechst, Nordisk, Lilly, Novo)
Human parathyroid hormone	34	diagnosis of calcium metabolism disorders	
Glucagon	29	treatment of hypoglycemia	(Lilly, Novo)
Growth hormone releasing factor	29, 44	diagnosis and treatment of growth disorders	Somatobiss (Bissendorf Peptide)
Luteinizing hormone releasing factor	10	diagnosis of hypophyseal function	Retefac (Hoechst), GnRH-Serono (Serono)
Oxytocin	9	induction of labor in pregnancy	Orasthin (Hoechst), Syntocinin (Sandoz)
Somatostatin	14	control of gastric bleeding	Aminopan (Union Carbide), Stilamin (Serono)

cal synthesis plays the most important role. On a small scale (milligrams to grams) the solid-phase method is the most powerful and the most routine. However, as the scale of the synthesis increases (hundreds of grams to kilograms) solution synthesis will probably play an increasingly important role. At the present time solution synthesis is mostly used but solid-phase synthesis is becoming increasingly important.

6. Cleavage of Protecting Groups

Due to the wide variety of protecting groups, a correspondingly large variety of deblocking procedures are available.

Acidolysis. Deprotection by acidolysis is widely used because a large number of protecting groups are removed simultaneously by acids. The most commonly used acidic reagents are trifluoroacetic acid, liquid anhydrous hydrogen fluoride, and hydrochloric or hydrobromic acids in glacial acetic acid or organic solvents. The peptide can, for example, be dissolved in trifluoroacetic acid and left to stand for 1 h. The deblocked peptide is then precipitated with diethyl ether. Anhydrous hydrogen fluoride cleaves most of the protecting groups and is therefore often used during the final deprotection step, despite its inherent toxicity and corrosive properties. This reagent can, however, often be replaced by trifluoromethanesulfonic and fluorosulfonic acids.

Catalytic hydrogenolysis is also a popular method for deprotection, because of its mildness. It is used to remove benzyloxycarbonyl groups, nitro groups (used to protect the side chain of arginine), benzyl ester groups, and benzyl ether groups. The protected peptide is dissolved in methanol and treated with hydrogen in the presence of a catalyst (10 % palladium on charcoal). This method is ineffective when sulfur- containing residues are present.

Deprotection under alkaline conditions (e.g., saponification of peptide ethyl ester with 1 mol/L NaOH for 2 h) is relatively important, although it may be accompanied by several side reactions, including racemization. Deprotection may be carried out using hydroxyl ions, hydrazine, thiolysis, or rhodanolysis. Treatment of the peptide with sodium in liquid ammonia is also an effective method of deprotection, but is not widely used. *Other procedures* are described in [1–8].

7. Purification of Peptides

Purification of peptides by crystallization is difficult or often impossible because peptides rarely crystallize or, if they do, they usually cocrystallize with impurities. Other (chromatographic)

techniques must therefore be used, see also → Process-Scale Chromatography.

Gel chromatography is a simple, powerful technique because it can be applied both in aqueous and in organic solvents. Thus, both water-soluble and water-insoluble (protected) peptides can be purified. This technique separates components mainly on the basis of their molecular mass. The most common gels– cross-linked dextrans (Sephadex) and cross-linked polyacrylamides–are available in a number of different pore sizes, making them suitable for the purification of peptides and proteins with a wide range of molecular masses.

Liquid partition chromatography relies upon the distribution of materials between two (usually immiscible) liquid phases, one of which is a moving phase, while the other is dispersed on a carrier phase, such as Sephadex, cellulose, or aluminum oxide. The technique is often faster and more efficient than liquid – liquid extraction.

Liquid – liquid extraction is based on the fact that the main product and the side products have different distribution coefficients in two immiscible solvents (e.g., butanol and acetic acid) [34]. The technique can be automated and used to purify kilogram quantities of peptides. Examples of liquid phases are butanol – acetic acid – water (volume ratios 4 : 1 : 5) and butanol – pyridine – acetic acid–water (volume ratios 4 : 2 : 1 : 7).

Ion-exchange chromatography is widely used because of its efficiency and relatively easy scaleup. Chromatography is carried out on supports which bear negatively or positively charged functional groups on which the peptide is adsorbed and subsequently desorbed by varying the pH or salt concentration. Since the charge on molecules is sensitive to structure and microenvironment, this technique separates peptides and proteins on the basis of both charge and structure.

Reverse-phase high performance liquid chromatography (HPLC) is the most powerful method for the purification of peptides. It relies mainly on hydrophobic interactions between the products and a matrix (usually silica derivatized with a long carbon chain). Through careful choice of conditions, peptides can be purified extremely rapidly, and large-scale equipment is now available which allows purification on a kilogram scale.

8. Analytical Control of Peptides

Because of the complexity of peptides, no single analytical method can be used to assess the purity of a peptide after purification. Although traditional methods (e.g., thin layer chromatography and paper electrophoresis) have been used, analytical, reverse-phase HPLC is now largely employed. Larger peptides and proteins can also be analyzed by gel electrophoresis. It is also important to confirm the amino acid composition of the peptide, as well as the optical purity of the constituent amino acids by amino acid analysis [35]. The peptide is hydrolyzed with strong acid (usually 6 mol/L hydrochloric acid), and the resulting mixture of amino acids is quantitatively separated by ion-exchange chromatography. Analysis for racemization is carried out by using gas chromatography to separate enantiomers [36]. The correctness of the amino acid sequence may be confirmed by Edman degradation [37]. Mass and NMR spectroscopy provide additional confirmation of the structure and also confirm the absence of residual protecting groups from the synthesis.

9. Important Commercial Peptides

The list of peptides which are reaching the market place as both diagnostic and therapeutic products is increasingly constantly. Some peptides with important pharmaceutical applications are listed in Table 7.

10. References

1. E. Gross, J. Meienhofer (eds.): *The Peptides, Analysis, Synthesis, Biology,* vols. 1 ff, Academic Press, New York 1979.
2. *Houben-Weyl,* 4th ed., **15.**
3. K. Lübke, E. Schröder, G. Kloss: *Chemie und Biochemie der Aminosäuren, Peptide und Proteine,* vols. 1 and 2, Thieme Verlag, Stuttgart 1975.
4. J. M. Stewart, J. D. Young: *Solid Phase Peptide Synthesis,* 2nd. ed., Pierce Chemical Co., Rockford, Ill., 1984.
5. J. W. van Nispen, *Pure Appl. Chem.* **59** (1987) 331 –344.

6. I. J. Galpin, *Amino-Acid Pept. Protein Abstr. Spec. Publ.* **16** (1985) 272 – 341.
7. M. Bodansky, *Int. J. Pept. Protein Res.* **24** (1984) 449 –474.
8. G. Barany, N. Kneib-Cordonier, D. G. Mullen, *Int. J. Pept. Protein Res.* **30** (1987) 705 – 739.
9. M. Bergmann, L. Zervas, *Ber. Dtsch. Chem. Ges. B* **65** (1932) 1192 – 1201.
10. G. F. Sigler et al., *Biopolymers* **22** (1983) 2157 – 2162.
11. P. Sieber et al., *Helv. Chim. Acta* **57** (1974) 2617 –2621.
12. T. Brown, J. H. Jones, J. D. Richards, *J. Chem. Soc. Perkin Trans. 1* 1982, 1553 – 1561.
13. R. Colombo, F. Colombo, J. H. Jones, *J. Chem. Soc. Chem. Commun.* 1984, 292 – 293.
14. P. Sieber, B. Riniker in G. R. Marshall (ed.): *Peptides: Chemistry and Biology,* Escom, Leiden 1988, pp. 76 – 79.
15. E. Atherton, R. C. Sheppard, J. D. Wade, *J. Chem. Soc. Chem. Commun.* 1983, 1060 – 1062.
16. O. Nashimura, M. Fujino, *Chem. Pharm. Bull.* **24** (1976) 1568 – 1575.
17. R. Ramage, J. Green, *Tetrahedron Lett.* **28** (1987) 2287.
18. W. W. Cleland, *Biochemistry* **3** (1964) 480.
19. J. P. Tam, W. F. Heath, R. B. Merrifield, *J. Am. Chem. Soc.* **105** (1983) 6442 – 6455.
20. D. Hudson in J. E. River, G. R. Marshall (eds.): *Peptides: Chemistry, Structure and Biology,* Escom, Leiden 1990, pp. 914 – 915.
21. M. Bodansky, V. du Vigneaud, *Nature (London)* **183** (1959) 1324 – 1325.
22. M. Bodansky, N. J. Williams, *J. Am. Chem. Soc.* **89** (1967) 685 – 689.
23. J. S. Morley, *J. Chem. Soc. C* 1967, 2410 – 2421.
24. R. B. Merrifield, *J. Am. Chem. Soc.* **85** (1963) 2149 –2154.
25. M. Mutter, E. Bayer, *Angew. Chem.* **86** (1974) 97.
26. J. Markussen, A. Voland in V. J. Hruby, D. H. Rich (eds.): *Peptides: Structure and Function,* Pierce Chemical Co., Rockford, Ill., 1983, pp. 207 – 210.
27. K. Morihara, T. Oka, H. Tsuzuki, *Nature (London)* **280** (1979) 412 – 413.
28. M. Bergmann, H. Tränkel-Conrat, *J. Biol. Chem.* **124** (1937) 1 – 6.
29. H. D. Jakubke, P. Kuhl, A. Konnecke, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 85.
30. P. Thorbeck, J. Lauridsen, F. Widmer in G. R. Marshall (ed.): *Peptides: Chemistry and Biology,* Escom, Leiden 1988, pp. 279 – 281.
31. J. A. Martial in E. Gross, J. Meienhofer (eds.): *Peptides: Structure and Biological Function,* Pierce Chemical Co., Rockford, Ill., 1979, pp. 969 – 983.
32. R. Camble, M. D. Edge, V. E. Moore in C. M. Deber, V. J. Hruby, K. D. Kopple (eds.): *Peptides: Structure and Function,* Pierce Chemical Co., Rockford, Ill., 1985, pp. 375 – 384.
33. R. E. Chance et al. in D. H. Rich, E. Gross (eds.): *Peptides: Synthesis, Structure, Function,* Pierce Chemical Co., Rockford, Ill., 1981, pp. 721 – 728.
34. W. Hausman, L. C. Craig, *J. Amer. Chem. Soc.* **80** (1958) 2703.
35. S. Moore, W. H. Stein, *J. Biol. Chem.* **192** (1951) 663 – 681.
36. H. Frank, G. J. Nicholson, E. Bayer, *J. Chromatogr. Sci.* **15** (1977) 174 – 176.
37. P. Edman, *Acta Chem. Scand.* **4** (1950) 277.

Perchloric Acid → Chlorine Oxides and Chlorine Oxygen Acids

Perfumes

Wolfgang Sturm, Haarman & Reimer GmbH, Holzminden, Federal Republic of Germany

Klaus Peters, Haarman & Reimer GmbH, Holzminden, Federal Republic of Germany

1. Introduction 1
2. Raw Materials and Production 1
3. Perfume Compositions 4
4. Development of Extrait Perfumes . . . 4
4.1. Fragrance Concepts 4
4.2. Fragrance Notes 5
5. Perfumes for Cosmetics, Toiletries, and Household Products 5
6. Economic Aspects 6
7. References 6

1. Introduction

Although perfumes are alcoholic – aqueous solutions of perfume oils, the term "perfume" is often used to denote concentrated perfume oil. Perfume oils may be composed of complex natural products, as well as natural or synthetic fragrance substances. The art of perfumery consists of combining a wide range of raw materials to create a harmonious, pleasant, and expressive fragrance.

History. Mixtures of fragrance substances (odorants) in the form of incense (Latin *perfumum*) and perfumed unguents were used for ceremonial purposes in early times. Along with their ritual uses, perfumes became important among the Greeks and Romans as luxury items. After the isolation of essential oils from green plants, bark, wood, and herbs the art of perfumery advanced rapidly and became fully established during the Renaissance. French perfumery has held a dominant position in Europe since the reign of Louis XII (ca. 1500).

Around 1700, Giovanni Paolo Feminis, an Italian who had migrated to Cologne, began marketing what later became eau de cologne. The classic formula for this product consists of bergamot oil, lemon oil, orange oil, and orange flower oil in ethanol. The original name *aqua mirabilis* (marvellous water) was replaced by *aqua coloniensis* and *Kölnisch Wasser* (Colonia is the Latin name of the city of Cologne, and Köln is its German name), and finally, by eau de cologne. Several firms in the city, including Johann Maria Farina and 4711, made eau de cologne a world-famous product.

Modern perfumery began at the end of the 19th century when the first synthetic fragrance substances became available. The initially modest selection of natural products was subsequently extended in a few decades as a result of intensive research and advances in production technology. The importance of perfumery has increased greatly since then. It includes not only the manufacture of fine products (perfume and eau de cologne), but also the use of perfumes in cosmetics, toiletries, soaps, and household or industrial products.

In addition to French perfumery, North American perfumery has become very important in recent years. With respect to distinctive fragrance types as well as market volume, these two sources now enjoy a virtually equal status.

2. Raw Materials and Production

The raw materials used in perfumery are essential oils, floral oils, extracts from plants and from animal secretions, fragrance substances isolated from natural products, and purely synthetic aroma chemicals. The number of natural fragrance substances is limited, but the increasing refinement of analytical techniques continues to yield new and interesting natural products. Chemical synthesis is, however, mainly responsible for the steady stream of innovations in perfumery since the turn of the 20th century.

The production and properties of aroma chemicals and fragrance substances are described elsewhere (→ Flavors and Fragrances).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_171

Raw Materials of Plant Origin. The following products are isolated from plant raw materials:

1) *Essential oils*, obtained by pressing plant materials or by steam distillation
2) *Concretes*, obtained by extraction with volatile solvents such as petroleum ether, benzene derivatives, or alcohols
3) *Resinoids*, produced by extraction of aromatic resins, mosses, balsams, and roots with volatile solvents
4) *Absolutes*, derived by treating concretes and resinoids with ethanol

The starting materials for these products are

1) *flowers* such as lavender, rose, jasmine, and neroli (orange flower);
2) *stalks and leaves* such as geranium, patchouli, and petitgrain;
3) *fruit* such as anise, coriander, caraway, cumin, and juniper;
4) *peel of citrus fruits* such as bergamot, lemon, and orange;
5) *seeds* such as mace, angelica, celery, and cardamom;
6) *roots* such as angelica, costus, and iris;
7) *wood* such as sandalwood, guaiac, cedar, and bois de rose;
8) *herbs and grasses* such as tarragon (estragon), lemongrass, sage, and thyme;
9) *needles and twigs* of spruce, fir, and pine; and
10) *resins and balsams* such as those derived from galbanum, elemi, benzoin, myrrh, olibanum, and opopanax.

Both the natural products and the aroma chemicals isolated from them are employed in perfumery. Examples of the latter are geraniol and citronellal from citronella oil, citral from lemongrass or *Litsea cubeba* oil, eugenol from clove oil, and menthol from peppermint oil.

Raw Materials of Animal Origin. Raw materials of animal origin are becoming increasingly rare and expensive. The following are used:

1) *Ambergris*, a metabolic product excreted by sperm whales that may result from a pathological condition
2) *Musk*, a glandular secretion of a hornless deer of Central Asia
3) *Civet*, a glandular secretion of the civet cat
4) *Castoreum*, a glandular secretion of the beaver

Semisynthetic and Synthetic Aroma Chemicals. *Semisynthetic products* are obtained by chemical modification of a natural starting material. They include hydroxycitronellal from citronellal, citronellol from geraniol or citronellal, geranyl acetate from geraniol, and ionones and methylionones from citral.

Purely *synthetic fragrance substances* are produced from basic chemicals by complete synthesis. They can be divided into products that are identical to natural ones and products that do not occur in nature. Products identical to natural substances but obtained by chemical synthesis include benzyl acetate from toluene, phenethyl alcohol from benzene, menthol from thymol, and linalool, a product of acetylene synthesis.

Other synthetic aroma chemicals have molecular structures completely different from those of natural products. They can be produced only by chemical synthesis, and often imitate the olfactory impressions of natural raw materials. Examples are 4-*tert*-butylcyclohexyl acetate (woody note, violet note), α-amylcinnamaldehyde (jasmine), 4-*tert*-butyl-α-methylhydrocinnamaldehyde (cyclamen), musk ketone (musk), and ethylene brassylate (musk).

Many natural products will continue to be indispensable in perfumery. However, synthetic products are playing an increasingly important role in the perfumer's assortment of raw materials because of their virtually unrestricted availability, constant quality, and generally steady price. Modern analytical techniques allow identification of the components of natural mixtures, which are often very complex. Reconstitution of these products and of essential oils by synthetic materials has led to substitutes with olfactory distinction.

Fixatives. Fixatives are used to equalize the vapor pressures, and thus the volatilities, of the raw materials in a perfume oil, as well as to increase the tenacity.

Natural fixatives are resinoids (benzoin, labdanum, myrrh, olibanum, storax, and tolu

Table 1. Floral bouquet

Composition		Quantity, g	Compound
Top note		5	2-undecylene aldehyde
		5	ω-decenol
		20	2,6-dimethyl-2-heptanol
		60	dihydromyrcenol
		50	bergamot oil
		30	linalool
		10	neroli oil
Floral complex	rose	50	citronellol
	rose	30	geraniol
	rose	80	phenethyl alcohol
	lilac	25	terpineol
	jasmine	75	α-hexylcinnamaldehyde
	jasmine	25	α-amylcinnamaldehyde
	jasmine	30	benzyl acetate
	lily of the valley	120	4-*tert*-butyl-α-methylhydrocinnamaldehyde
	lily of the valley	40	4-(4′-hydroxy-4′-methylpentyl)-3-cyclohexene carboxaldehyde
	lily of the valley	10	4-isopropyl-α-methylhydrocinnamaldehyde
	gardenia	15	styrolyl acetate
	iris – violet	60	γ-methylionone
	iris – violet	25	4-*tert*-butylcyclohexyl acetate
	clove	10	eugenol
	clove	5	eugenol methyl ether
Fixation	sweet	25	cinnamic alcohol
	sweet	20	coumarin
	wood	60	acetylcedrene – acetylthujopsene
	ambergris	5	1,2-furano-2,5,5,8a-tetramethyldecalin (10 %) in diethyl phthalate
	musk	70	6-acetyl-1,1,2,4,4,7-hexamethyltetralin
	musk	10	cyclopentadecanolide
	musk	30	musk ketone (4-*tert*-butyl-2,6-dimethyl-3,5-dinitroacetophenone)
		1000	

balsam) and animal products (ambergris, castoreum, musk, and civet). Synthetic fixatives include substances of low volatility (cyclopentadecanolide, ambroxide, benzyl salicylate) and virtually odorless solvents with very low vapor pressures (benzyl benzoate, diethyl phthalate, triethyl citrate).

Solvents. The only solvent used in fine perfumery is extremely pure ethanol that is diluted with water to the required concentration. Weakly odorous synthetic fixatives are used as solvents for alcohol-free perfumes, especially bazaar oils. (Bazaar oils are perfume oils that are dissolved in an oily solvent; they are sold in bazaars in Arabian countries and used as perfumes.)

Production. Perfumes were originally prepared from alcoholic infusions, solutions, and tinctures of floral extracts, essential oils, natural and artificial aroma chemicals, drugs, and resins. The long storage times allowed complete clarification and maturing of the alcoholic solutions, so that the perfumes required a much shorter aging time. As a result of increasing industrialization, perfumery has departed from these classical methods of preparation. Today, perfume compositions prepared in the form of perfume oils are dissolved in ethanol, cooled to ca. 0 °C, filtered, packaged, stored, and marketed. Perfume (extrait) may be for example, a 10 – 20 vol% solution of a perfume oil in ca. 90 – 95 vol% ethanol. An eau de toilette or eau de parfum contains ca. 7 – 10 vol % perfume oil in 80 – 90 vol % ethanol; an eau de cologne (Kölnisch Wasser), up to 4 vol % perfume oil in ca. 70 – 85 vol % ethanol.

3. Perfume Compositions

In building up a composition, the perfumer must work systematically and observe certain rules of workmanship. A fragrance can be divided into three parts:

1) Top note or initial fragrance impression
2) Body note or middle note
3) End note or dry out

The odor must pass smoothly from one phase of this "cycle" to the next; the overall character of the perfume should be clearly perceptible from the top note to the end note, and must impart its own unmistakable note to the fragrance.

Linear Compounding. A sample formula is described to illustrate the formulation of a floral fragrance. This formula also demonstrates the technique of "linear" composition, which is based primarily on well-defined aroma chemicals, essential oils, and floral oils. "Complex" compounding, on the other hand, mainly uses predesigned and balanced perfume bases and fragrance complexes.

Sample Formula. The fragrance materials in the "floral bouquet" (Table 1) correspond to a variety of flowers; they give a basic, inherently rounded floral complex. Nuancing agents for the top note are fresh-flower components such as ω-decenol and 2,6-dimethyl-2-heptanol, whereas several sweet, woody, and musk odorants such as acetylcedrene – acetylthujopsene and 6-acetyl-1,1,2,4,4,7-hexamethyltetralin serve as fixatives.

Table 2. Rose base

Composition	Quantity, g	Compound
Top note	15	2-undecylene aldehyde (10 %) in diethyl phthalate
	5	ω-decenol
	5	phenylacetaldehyde dimethyl acetal
	20	decanal (1 %) in diethyl phthalate
Body	300	phenethyl alcohol
	200	citronellol
	100	geraniol
	20	geranyl acetate
	160	phenethyl diethylacetate
	25	nerol
	20	farnesol
	25	nerolidol
	5	eugenol
Fixation	20	trichloromethylphenyl carbinyl acetate
	30	cinnamic alcohol
	50	phenethyl phenylacetate
	1000	

Complex Compounding. In complex compounding, separately balanced perfume bases are used instead of individual fragrance materials or fragrance accords. These bases consist of numerous components; appropriate mixtures of synthetic products are also substituted for the fragrance complexes of natural essential oils.

The complex method yields a more well-rounded composition than a number of individual fragrance materials. For example, the simple rose accord in the "floral bouquet" (Table 1), consisting of citronellol, geraniol, and phenethyl alcohol, can be replaced by a rose base (Table 2) whose composition has been worked out in much more detail.

Replacement of fragrance notes composed of a few substances, as described in Table 1, by sophisticated perfume bases of appropriate odor type yields a composition consisting of several hundred components that has an exceedingly complex makeup and a very rounded odor.

4. Development of Extrait Perfumes

4.1. Fragrance Concepts

Extrait perfumes set the trends for all developments in perfumery. They are important for any creative perfumer, regardless of the application for which new fragrance compositions are being developed. The many perfumes now on the market can be classified according to three main fragrance concepts:

1) **Floral concept**
   a) Characteristic raw materials:

Benzyl acetate (jasmine)
α-Amylcinnamaldehyde (jasmine)
α-Hexylcinnamaldehyde (jasmine)
Indole (jasmine)
Jasmine absolute (jasmine)

Terpineol (lilac)
Anisaldehyde (lilac)

Phenethyl alcohol (rose)
Geraniol (rose)
Citronellol (rose)
Rose oil (rose)

Hydroxycitronellal (lily of the valley)
4-*tert*-Butyl-α-methylhydrocinnamaldehyde (lily of the valley)
4-(4-Hydroxy-4-methyl-pentyl)-3-cyclohexene carboxaldehyde (lily of the valley)

Ionones (iris – violet)
Methylionones (iris – violet)

Eugenol (clove)
Isoeugenol (clove)
Clove oil (clove)

b) Typical products:
Joy (Patou, 1935); L'Air du Temps (Ricci, 1948); Fidji (Laroche, 1966); Diorella (Dior, 1972); Charlie (Revlon, 1973); Chloé (Lagerfeld, 1975); Giorgio (Beverly Hills, 1981); Poison (Dior, 1985).

2) **Oriental Note Concept**
a) Characteristic raw materials:
Vanillin, coumarin, tonka bean extract, and musk aroma chemicals
b) Typical products:
Shalimar (Guerlain, 1925); Youth Dew (Lauder, 1952); Opium (St. Laurent, 1977); Must de Cartier (Cartier, 1981).

3) **Chypre Concept**
a) Characteristic raw materials:
Combination of bergamot oil, oakmoss extracts, and patchouli oil
b) Typical products:
Chypre (Coty, 1917); Cabochard (Grès, 1958); Miss Dior (Dior, 1947); Intimate (Revlon, 1955); Coriandre (Couturier, 1973); Ysatis (Givenchy, 1984).

4.2. Fragrance Notes

Within each fragrance concept, individual fragrance notes can dominate. Examples follow.

1) **Green Note**
a) Characteristic raw materials:
Hexanol, hexenol, and their esters; galbanum oil
b) Typical products:
Floral concept: Vent vert (Balmain, 1945)
Chypre concept: Alliage (Lauder, 1972)

2) **Aldehydic Note**
a) Characteristic raw materials:
Aliphatic aldehydes (C_6–C_{14})
b) Typical products:
Floral concept: Chanel No. 5 (Chanel, 1921)
Chypre concept: Ma Griffe (Carven, 1944)

5. Perfumes for Cosmetics, Toiletries, and Household Products

Perfumes and eaux de cologne are used specifically for their odor. In cosmetics, toiletries, and household products, the perfume is usually of secondary importance for the effectiveness of the product; it may, however, strongly influence the consumer's decision to buy a product and, in many cases, represents the only way of making the product's action "perceptible" to the consumer.

Although the model fragrance types are often created in the field of fine perfumery, chemical and physical aspects must also be taken into account in the development of perfume oils for toiletries and household products. For example, the perfume oil used in creams and white soaps must not cause discoloration; a perfume oil for aftershave lotion must be soluble in 50 – 60 % alcohol; the fragrance used in a powdered detergent must be alkali resistant; a fabric softener is expected to leave clothes with a pleasant odor; and even a household cleanser must have a pleasant and functional odor, although active chlorine places extraordinary demands on the stability of the perfume oil. Minimal perfume doses are expected to adequately mask the often strong and unpleasant odors of products such as insecticides, floor cleanser, paints, and varnishes.

Fragrances for soaps must satisfy the following requirements: chemical resistance, low volatility of raw materials (e.g., geraniol and ionone), strength of odor in soap (e.g., citronellal and γ-decalactone), and adhesion to the skin (e.g., 1,2-furano-2,5,5,8a-tetramethyldecalin and α-amylcinnamaldehyde). In

the past, well-defined perfume types such as lavender, citronella, rose, lilac, pine, and fern (fougère) were preferred. Today, however, perfumes are custom formulated, usually in fantasy notes with a fresh, invigorating impact or oriented to fine personal- care cosmetics.

Olfactory enhancement of all types of household products has become increasingly important in recent decades. The development of perfume oils for this wide-ranging field requires a special knowledge of perfumery, as well as information about the chemical and thermal stability of the perfumery materials used and their impact and substantivity in relation to the product being perfumed. [Substantivity denotes the affinity of a perfume material for a substrate, e.g., fibers (fabric conditioners) or hair (hair conditioners).]

6. Economic Aspects

Although no sales figures have been disclosed, worldwide consumption of perfume oils for all applications in 1984 was estimated at ca. 1.8×10^9 €(ca. 2.3×10^9 \$).

Detergents and household products account for ca. 30 % of this total, soaps for 20 %, cosmetics and toiletries for 30 %, and perfumes or eaux de cologne for 20 %.

7. References

1. S. van Toller, G. H. Dodd: *Perfumery, the Psychology and Biology of Fragrances,* Chapman and Hall, London 1989.
2. G. Ohloff: *Riechstoffe und Geruchssinn,* Springer Verlag, Heidelberg 1990.
3. Haarman & Reimer: *The H + R Book of Perfume,* Johnson Publications, London 1984.
4. C. Jellinek: *Praktikum des modernen Parfumeurs,* Urban & Schwarzenberg, Wien 1949.
5. C. Jellinek: *Die psychologischen Grundlagen der Parfümerie,* Hüthig Verlag, Heidelberg 1951.
6. J. S. Jellinek: *Parfümerien von Produkten,* Hüthig Verlag, Heidelberg 1976.
7. *Genealogie der Extrait-Parfums,* Haarmann & Reimer, Holzminden 1979.
8. W. E. Dorland, J. A. Rogers, Jr.: *The Fragrance and Flavor Industry,* Wayne E. Dorland Comp., Mendham, New Jersey 1977.

Perline → Silicates
Peroxides → Peroxo Compounds, Inorganic

Peroxo Compounds, Inorganic

Harald Jakob, Degussa GmbH, Hanau, Germany (Chap. 1, 2, 3, 4, 5, 6 and 7)
Stefan Leininger, Degussa GmbH, Hanau, Germany (Chap. 5)
Thomas Lehmann, Degussa GmbH, Hanau, Germany (Chap. 5)
Sylvia Jacobi, Degussa GmbH, Hanau, Germany (Chap. 8)
Sven Gutewort, Degussa GmbH, Pullach, Germany

Related Articles Hydrogen Peroxide is a separate keyword.

1. Introduction . . . 1
1.1. History . . . 2
1.2. Classification of Peroxo Compounds 2
1.3. Properties . . . 3
1.4. Analysis . . . 3
2. Peroxides, Hyperoxides, and Ozonides of the Alkali and Alkaline Earth Metals . . . 3
2.1. Peroxides . . . 3
2.1.1. Sodium Peroxide . . . 4
2.1.2. Other Peroxides . . . 5
2.2. Hyperoxides . . . 6
2.3. Ozonides . . . 7
3. Peroxoborates . . . 7
3.1. Sodium Peroxoborates . . . 7
3.1.1. Properties . . . 7
3.1.2. Production . . . 10
3.1.3. Uses . . . 11
3.1.4. Product Specifications . . . 12
3.1.5. Storage and Transportation . . . 12
3.2. Other Peroxoborates . . . 12
4. Peroxosulfuric Acids and Peroxosulfates . . . 12
4.1. Peroxomonosulfuric Acid . . . 13
4.2. Caro's Acid . . . 13
4.3. Peroxomonosulfates . . . 14
4.3.1. Properties . . . 14
4.3.2. Production . . . 14
4.3.3. Uses . . . 15
4.3.4. Analysis . . . 15
5. Peroxodisulfuric Acid and Peroxodisulfates . . . 16
5.1. Properties . . . 16
5.2. Production . . . 16
5.3. Uses . . . 18
5.4. Analysis . . . 18
6. Peroxophosphoric Acids and Their Salts . . . 18
7. Hydrogen Peroxide Addition Compounds . . . 19
7.1. Sodium Carbonate Peroxohydrate . 19
7.1.1. Properties . . . 19
7.1.2. Production . . . 20
7.1.3. Stabilization, Storage, Packaging, and Transportation . . . 22
7.1.4. Uses . . . 23
7.2. Urea Peroxohydrate . . . 23
7.2.1. Properties . . . 23
7.2.2. Production . . . 23
7.2.3. Stabilization and Packaging . . . 23
7.2.4. Uses . . . 23
8. Toxicology and Ecotoxicology . . . 24
8.1. Toxicology . . . 24
8.2. Ecotoxicology . . . 26
9. References . . . 26

1. Introduction

Peroxo compounds include hydrogen peroxide (→ Hydrogen Peroxide) and substances derived from hydrogen peroxide by substitution of one or both hydrogen atoms by a metal or a nonmetal such as sulfur, boron, nitrogen, or phosphorous. Peroxides in a wider sense also include the hyperoxides (MO_2), H_2O_2 addition compounds (e.g., sodium carbonate peroxohydrate), and the inorganic ozonides (M^IO_3). In the IUPAC nomenclature, the prefix peroxo is used for inorganic compounds, and peroxy for organic compounds.

A very large number of peroxo compounds have been described in the literature [1 – 5]. This article gives an overview on the properties and industrial applications of the most important peroxo compounds and products of commercial importance.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_177.pub2

Due to the low strength of the $-O-O-$ bonding (209 kJ/mol), all peroxo compounds exhibit powerful oxidizing properties. Inorganic peroxo compounds are therefore employed as oxidizing and bleaching agents in the detergents, textile, electronics, and chemical industries.

1.1. History

The first peroxo compounds discovered in the early 1800s by Gay-Lussac and Thenard were the peroxides of barium, sodium, and potassium. Hydrogen peroxide was isolated several years later in 1818 by Thenard. Due to the advantages in stability and handling, the solid peroxides were of greater practical use than hydrogen peroxide solutions. Sodium peroxide gained industrial importance mainly as a bleaching agent in the textile industry and was first produced on a commercial scale by the Castner process in 1899. Sodium peroxide lost its importance when hydrogen peroxide could be more cost-effectively manufactured by the anthraquinone process.

Today the most important industrial peroxo compound is sodium carbonate peroxohydrate. Although discovered in 1899 by Tanatar it only gained commercial importance in the 1990s due to the ecologically driven shift from sodium perborate to sodium carbonate peroxohydrate, especially in Europe. It has nowadays become the preferred bleaching agent in powder laundry and automatic dishwashing detergents since products with good shelf-life properties have become available through new economically viable coating technologies. Sodium peroxoborates, also discovered by Tanatar, in 1898, have been produced on a commercial scale since 1907 and still remain the preferred bleaching agents in powder laundry and automatic dishwashing detergents for climatically challenging regions such as the Middle East, northern Africa, and South and Middle America.

Other industrially important inorganic peroxo compounds are the sodium, potassium, and ammonium salts of peroxodisulfuric acid ($H_2S_2O_8$), the triple salt of potassium peroxomonosulfate ($2\ KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$), and calcium peroxide (CaO_2).

Of the other hydrogen peroxide addition compounds, only urea peroxide is produced on an industrial scale of several hundred tonnes; it has limited applications in pharmaceutical products.

1.2. Classification of Peroxo Compounds

Wannagat [6] proposed that the peroxo compounds should be classified according to their chemical bonding, i.e., whether the bond between the peroxo group and the substituent(s) is heteropolar, covalent, or dative. However, a precise distinction is often impossible, either because of continuous gradation between the bonding types or because the structure has not yet been fully elucidated.

Peroxo Compounds with Heteropolar Bonds. Compounds of this type crystallize in an ionic lattice and contain the colorless, diamagnetic O_2^{2-} anion. The distance between the oxygen atoms is 149 pm. Typical examples of this class are the alkali and alkaline earth metal peroxides $M^{I}_2O_2$ or $M^{II}O_2$.

Peroxo Compounds with Covalent Bonds. The typical example for this group is hydrogen peroxide itself. Whereas the paramagnetic oxygen molecule has a multiple bond (bond length 121 pm), the O–O distance for these compounds is 147 pm, indicating a covalent single bond. The two hydrogen atoms are linked to the O_2 moiety to form a nonplanar structure. Elements of Groups 13–16 of the periodic table, such as boron, carbon, silicon, phosphorus, and sulfur, also form covalent or mainly covalent bonds with the peroxo group. Either one or both hydrogen atoms of hydrogen peroxide can be substituted. Industrially important examples of these compounds are peroxoborates, most organic peroxy compounds (→ Peroxy Compounds, Organic), the silyl peroxides, and the peroxomono- and peroxodiphosphates and -sulfates. Peroxo compounds with covalent bonds to the O_2 group are colorless.

Peroxo Compounds with Dative Bonds. These are metal complexes with peroxo ligands and are formed by polyvalent transition metal cations such as Ti, Cr, Mo, W, and U. They are sometimes colored. An unambiguous assignment is, however, difficult in the case of these

compounds. They can sometimes be regarded as peroxohydrates (see below) or as peroxoacids with covalent bonds. They have no industrial importance.

Peroxohydrates. With many substances hydrogen peroxide forms crystalline adducts in which it behaves like water of crystallization, which may also be present in the crystal structure. The peroxohydrates are also regarded as peroxo compounds.

Hyperoxides are classified as peroxo compounds only in a broad sense. The paramagnetic O_2^- anion (O–O distance: 128 pm, measured in KO_2) is a derivative of the very unstable HO_2 radical. The only known pure hyperoxides are those of the alkali metals, with the exception of lithium.

Inorganic ozonides of the type M^IO_3 are also only classified as peroxo compounds in a broad sense. The name ozonides was proposed by Remy [7]. The O_3^- ion is paramagnetic, as is the O_2^- ion, and has a bent structure, with a bond angle of 100° and an O–O distance of 119 pm.

1.3. Properties

Oxygen is present in the peroxo group in the comparatively unstable oxidation state -1. Hydrogen peroxide can behave as an oxidizing or reducing agent, depending on the coreactant, with the oxygen gaining or losing electrons to attain the more stable oxidation state -2 or 0. Most inorganic peroxides, however, act predominantly as oxidizing agents.

Another typical property of many peroxo compounds is their ready decomposition to form radicals. The first step is homolytic cleavage of the O–O bond, which can be initiated either thermally or catalytically (e.g., by metal ions). The industrial importance of many organic and inorganic peroxides as polymerization accelerators is based on this behavior.

Another typical reaction is exothermic redox disproportionation. This type of reaction is the cause of the loss of active oxygen on prolonged storage or heating. The decomposition can be suppressed or considerably reduced by adding small quantities of stabilizers such as magnesium sulfate or sodium silicate. Many commercial peroxo compounds stabilized in this way have extremely good storage stability at room temperature. All aqueous solutions of peroxo compounds hydrolyze forming hydrogen peroxide.

The alkali metal hyperoxides are very powerful oxidizing agents and react vigorously with water to form hydroxides with liberation of oxygen. The alkali metal ozonides are unstable and readily decompose to form hyperoxides and oxygen; they also form alkali metal hydroxides and oxygen with water.

1.4. Analysis

The quantitative determination of peroxidic oxygen in peroxo compounds is usually carried out by hydrolysis in aqueous solution to form H_2O_2, followed by titration with permanganate, cerium(IV) sulfate or potassium iodide/thiosulfate, giving the so-called active oxygen content in weight percent. This terminology is historic and is based on the incorrect assumption that one of the oxygen atoms is initially split off from the peroxo group as atomic oxygen. Thus, in 1 mol H_2O_2 (34 g) the active oxygen content is 16 g or 47.06 wt %.

2. Peroxides, Hyperoxides, and Ozonides of the Alkali and Alkaline Earth Metals

2.1. Peroxides

The alkali and alkaline earth metals form peroxides with the general formulae $M_2^IO_2$ and $M^{II}O_2$, respectively, which crystallize with an ionic lattice and contain the diamagnetic O_2^{2-} anion. In the case of lithium and the elements of Group 2 of the periodic table, the preparation can be carried out in aqueous solution from the hydroxides and hydrogen peroxide, but the alkali metal peroxides and also barium peroxide are better produced by oxidation of the metal or metal oxide. All these compounds are colorless or light yellow, the color being caused by small

amounts of the hyperoxides. The alkali and alkaline earth metal peroxides are powerful oxidizing agents that usually have good stability in the pure state if stored with exclusion of moisture. If they come into contact with organic materials or other reductants, there is danger of spontaneous ignition or even explosion.

Most of all these peroxides form many hydrates and peroxohydrates. They dissolve in water exothermically, forming hydrogen peroxide and the hydroxide of the metal. The alkali metal peroxides form compounds with a number of penta- and hexavalent heavy metal, oxides, e.g., of vanadium, molybdenum, tungsten and uranium, with the general formulas $2\,M^{I}_2O_2{\cdot}M^{V}_2O_5$ and $M^{I}_2O_2{\cdot}M^{VI}O_3$ [8].

2.1.1. Sodium Peroxide

Sodium peroxide was produced on a large scale for many years, and was an important raw material in the synthesis of sodium peroxoborates and sodium carbonate peroxohydrate. In the early 1980s, most of the large-scale plants ceased production because the more cost-effective hydrogen peroxide produced by the autoxidation process replaced sodium peroxide in most industrial applications. Today sodium peroxide is used only in small quantities for specialty applications, e.g., as a fusion agent in chemical analysis.

Properties. Sodium peroxide, [*1313-60-6*], Na_2O_2, M_r 78.0, ϱ 2.6 g/cm^3, is a yellowish white powder with a theoretical active oxygen content of 20.5 wt %. It melts at ca. 660 °C and decomposes above 300 °C with liberation of oxygen. The oxygen dissociation pressure is 101.3 kPa (1 atm) [9]. Sodium peroxide is hygroscopic, but in the absence of moisture it exhibits good shelf life.

Sodium peroxide is a powerful oxidizing agent. In contact with organic materials or other reducing agents, spontaneous ignition and even explosion can occur. Sodium peroxide dissolves exothermically in water to form sodium hydroxide and hydrogen peroxide. Sodium peroxide reacts with hydrogen gas yielding sodium hydroxide [10].

Sodium peroxide forms several peroxohydrates and hydrates, namely: $Na_2O_2 \cdot 2\,H_2O_2$, $Na_2O_2 \cdot 2\,H_2O_2 \cdot 4\,H_2O$, $Na_2O_2 \cdot 2\,H_2O$, and $Na_2O_2 \cdot 8\,H_2O$ [8].

In Europe sodium peroxide is manufactured by Meteaux Speciaux in France.

Production. Gay-Lussac and Thenard first prepared sodium peroxide in 1811 by the combustion of sodium metal. This method formed the basis for the large-scale production of this compound. The oxidation to sodium peroxide takes place with intermediate formation of sodium oxide:

$$2\,Na + 1/2\,O_2 \rightarrow Na_2O \quad \Delta H = -431.6\,kJ/mol$$
$$Na_2O + 1/2\,O_2 \rightarrow Na_2O_2 \quad \Delta H = -79.6\,kJ/mol$$

The first industrial process for the manufacture of sodium peroxide was the block process, developed by Castner in 1891 and operated by Degussa in Rheinfelden since 1899, and later developed in several stages to give the classical continuous rotary furnace process [11]. Another continuous process which has long been in use is the U.S. Industrial Chemicals Process (USI Process) [12].

Rotary Furnace Process (Fig. 1). The sodium is oxidized in two stages. In the first stage, molten sodium is continuously fed countercurrently with an excess of dry air into a rotary furnace (the combustion furnace, b), which contains previously formed sodium oxide in powder form. The furnace rotates at ca. 2 rpm and is cooled externally with water because of the exothermic reaction. The reactor temperature is 600–700 °C near the entry point of the sodium, falling to 150–200 °C at the discharge end. The sodium oxide leaving the rotary furnace contains a few percent sodium peroxide and a maximum of 1 % of sodium metal. After intermediate storage in a bunker (e), it is fed by a screw conveyor or traveling containers via a lock to the next stage for further oxidation in a rotary reactor lined with sheet nickel (f). Because of the low exothermicity of the peroxide formation, this reactor is located in a brick-lined muffle furnace heated by gas or oil; it rotates at about the same speed as the combustion furnace. The oxidizing agent is oxygen from an air separation plant, which is introduced in cocurrent flow with the sodium oxide. It is important to maintain the reaction temperature at the optimum value of 350–400 °C, otherwise the product can stick on

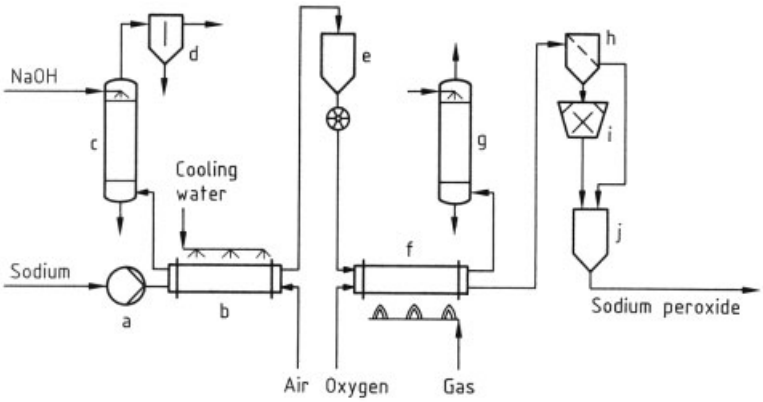

Figure 1. Production of sodium peroxide by the rotary furnace process
a) Pump; b) Combustion furnace; c) Scrubber; d) Separator; e) Bunker; f) Finishing reactor; g) Scrubber; h) Sieve; i) Mill; j) Bunker

the reactor wall and lose its free-flowing properties. After leaving the reactor, the sodium peroxide is cooled and then screened into the various particle size fractions. The product consists of small pale yellow spheres with a sodium peroxide content of ca. 99 %. The off-gases from the two furnaces are purified by scrubbers (c, g) and a separator (d).

USI Process [12]. In the USI process, like the rotary furnace process, sodium oxide is produced in the first stage by reaction of sodium metal with recycled off-spec sodium peroxide. The reactants are mixed at ca. 120 °C in a nitrogen atmosphere and then heated in a reactor at > 250 °C. Since excess sodium peroxide is used, a mixture of sodium peroxide and sodium oxide is obtained. This is oxidized by air in a heated rotary furnace to form the peroxide. The product is classified, and the fine and coarse material recirculated to the first stage. Since the process is operated exclusively with air as the oxidizing agent, 80–90 % of the peroxide formed must be recycled.

2.1.2. Other Peroxides

Lithium peroxide [*12031-80-0*], Li_2O_2, M_r 45.9, ϱ 2.36 g/cm^3, has good thermal stability and, unlike Na_2O_2, is not hygroscopic. The theoretical active oxygen content is 34.8 wt %, which is the highest of all metal peroxides. It has not attained industrial importance due to the comparatively high cost of lithium and its compounds and high manufacturing costs, among other reasons.

Lithium peroxide can be produced by the reaction of dilute aqueous solutions of lithium hydroxide with 30 wt % hydrogen peroxide to form lithium peroxide monoperoxohydrate trihydrate ($Li_2O_2 \cdot H_2O_2 \cdot 3\ H_2O$), which is then dehydrated over phosphorus pentoxide or under vacuum [13]. Other methods include the reaction of lithium alkoxides in alcoholic solution with concentrated hydrogen peroxide [14], and the treatment of solid lithium oxide hydrate with concentrated hydrogen peroxide, followed by washing the precipitate with methanol and drying in a conveying dryer [15]. Another process has been proposed in which hydrogen peroxide and saturated lithium hydroxide solution are sprayed together into a spray dryer [15].

Calcium and Strontium Peroxides. Calcium peroxide [*1305-79-9*], CaO_2, M_r 72.1, ϱ 2.92 g/cm^3, theoretical active oxygen content 22.2 wt %, and strontium peroxide [*1314-18-7*], SrO_2, M_r 119.6, ϱ 4.7 g/cm^3, theoretical active oxygen content 13.4 wt %, are produced by precipitation of the octahydrate from an aqueous solution of a salt or hydroxide, followed by dehydration at temperatures above 100 °C [16]. The thermal stability of strontium peroxide (decomp. ca. 410 °C) is marginally higher than that of calcium peroxide (decomp. ca. 375 °C). The active

oxygen content of the commercial products is only 60–90 % of the theoretical value. Calcium peroxide is sold under the trade name PermeOx by FMC. Calcium peroxide is also manufactured by Solvay.

Calcium peroxide is used as oxygen source in the leaching of gold and silver ores with cyanide to improve precious metal recovery and leaching kinetics [17]. A relatively new application for CaO_2 is as a peroxide source in toothpaste [18] and dental chewing gum products. Calcium peroxide is also used in limited quantities as a hardening accelerator for sulfide rubber, as dough conditioner in the baking industry, and as a nontoxic disinfectant for seed and grain [19 – 21].

Strontium peroxide is used in pyrotechnics and tracer bullet formulations.

Barium peroxide [*1304-29-6*], BaO_2, M_r 169.4, ϱ 5.4 g/cm^3, theoretical active oxygen content 9.4 wt %, was the earliest known peroxo compound. It was used until the mid-1900s as the starting material for the production of hydrogen peroxide by the Thenard reaction [22].

Barium peroxide is the only peroxide of the Group 2 metals that can be prepared at atmospheric pressure by air oxidation of the oxide. Air is passed over finely spread barium oxide at 500–550 °C. On further heating in air to ca. 700 °C, oxygen is liberated [23]. On the industrial scale, barium peroxide is produced in a rotary furnace in the presence of a small quantity of steam to accelerate the reaction. Today barium peroxide has only limited application in the manufacture of fireworks.

Magnesium and Zinc Peroxides. Magnesium peroxide [*14452-57-4*], MgO_2, M_r 56.3, and zinc peroxide [*1314-22-3*], ZnO_2, M_r 97.3, are manufactured in the same way as calcium and strontium peroxides [24]. However, well-defined compounds are not obtained, but mixtures of peroxohydrates and hydroxides with a peroxide content between 25 and 50 wt %. Zinc peroxide is used in dusting powders, as a deodorant, and in pharmacology for the treatment of infections and skin lesions. Recently, magnesium peroxide has been promoted as a oxygen-release compound for soil remediation [25]. Producers of magnesium peroxide are Solvay, Regenesis and FMC.

2.2. Hyperoxides

Hyperoxides, which are derived from the unstable HO_2 radical, have been described for some of the highly electropositive alkali and alkaline earth metals [26, 27]. Only potassium hyperoxide has a limited commercial importance.

Potassium hyperoxide [*12030-88-5*], KO_2, M_r 71.1, ϱ 2.14 g/cm^3, is a paramagnetic yellow solid for which melting points between 380 °C and 440 °C have been reported [30]. The physical properties of KO_2 are summarized in [28, 29]. It is used as an oxygen generator and carbon dioxide absorber in breathing apparatus and gas masks. This application is based on the ability of potassium hyperoxide to simultaneously give off oxygen and absorb carbon dioxide in the presence of water vapor. The following reactions take place above 20 °C:

$$4\,KO_2 + H_2O \rightarrow 2\,KOH + 3\,O_2$$

$$2\,KOH + CO_2 \rightarrow K_2CO_3 + H_2O$$

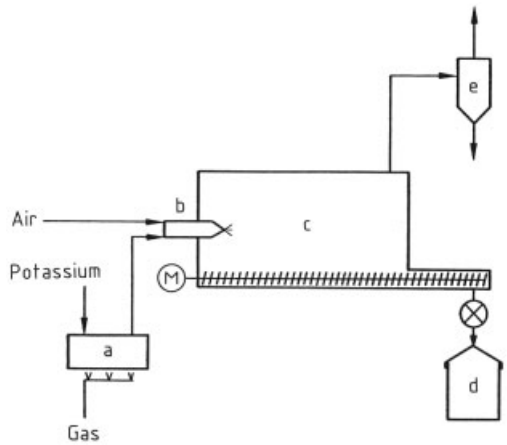

Figure 2. Production of potassium peroxide
a) Molten salt vessel; b) Spray nozzle; c) Reactor; d) Receiving vessel; e) Cyclone

Potassium hyperoxide can be prepared from potassium hydroxide and concentrated hydrogen peroxide, which are allowed to react under cooling at reduced pressure to give potassium

Table 1. The sodium peroxoborates

Common name	CAS registry no.	Historical formula	IUPAC name and constitutional formula
Sodium peroxoborate hexahydrate	[*10486-00-7*]	$NaBO_2 \cdot H_2O_2 \cdot 3\,H_2O$	disodium di-μ-peroxo-bis(dihydroxoborate) hexahydrate
Sodium perborate tetrahydrate	[*13517-20-9*]	$NaBO_3 \cdot 4\,H_2O$	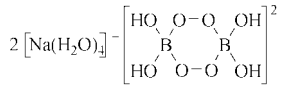
Sodium peroxoborate trihydrate	[*28962-65-4*]	$NaBO_2 \cdot H_2O_2 \cdot 2\,H_2O$	disodium di-μ-peroxo-bis(dihydroxoborate) tetrahydrate
Sodium perborate trihydrate	[*28108-09-0*]	$NaBO_3 \cdot 3\,H_2O$	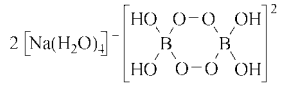
Sodium peroxoborate monohydrate	[*10332-33-9*]	$NaBO_2 \cdot H_2O_2$	disodium di-μ-peroxo-bis(dihydroxoborate)
Sodium perborate monohydrate	[*15120-21-5*]	$NaBO_3 \cdot H_2O$	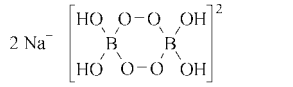
Anhydrous sodium peroxoborate	[*7632-04-4*]	$NaBO_3$	not a well-defined compound; sodium borate/boron–oxygen radical

peroxide diperoxohydrate $K_2O_2 \cdot 2\,H_2O_2$ [31, 32]. It decomposes at elevated temperature to form KO_2 (dehydration in a vacuum or spray drying). Another industrial process [33] uses potassium metal as the starting material (Fig. 2). Molten potassium is sprayed into a closed reactor (c) where it reacts with a large excess of preheated air at ca. 75 °C. The reaction is highly exothermic, and the temperature in the reactor reaches ca. 300 °C. The powdered product is removed by a screw conveyor and is then granulated.

KO_2 is manufactured by Mine Safety Appliance Company (USA).

2.3. Ozonides

Ozonides of the alkali metals with the general formula M^IO_3, can, with the exception of lithium ozonide, be prepared by the reaction of ozone–oxygen mixtures with the solid hydroxide, or preferably the hyperoxide, of the metal. These compounds are dark in color and paramagnetic. They are not very stable; the stability increases with increasing atomic mass of the metal. They decompose readily into the hyperoxide and oxygen [2]. Ozonides have no industrial importance.

3. Peroxoborates

3.1. Sodium Peroxoborates

3.1.1. Properties

Sodium peroxoborate exists as the hexahydrate, the tetrahydrate, the trihydrate, the water-of-crystallization-free substance, and the anhydrous substance. Anhydrous sodium perborate (trade name: Oxoborate) is not a peroxoborate compound. The naming of the sodium peroxoborates has long been confused. Table 1 lists the trivial names and systematic IUPAC nomenclature, CAS numbers, formulas, and molecular structures of the sodium peroxoborates and anhydrous sodium perborate. Most commonly used in the detergent industry are the names sodium perborate tetrahydrate and sodium perborate monohydrate. Sodium perborate trihydrate has no industrial importance.

Sodium perborate tetrahydrate was first prepared in 1898 by Tanatar and by Melikoff and Pissarjewski by the action of hydrogen peroxide on a solution of borax and sodium hydroxide [34]. The long-running controversy as to whether this is a peroxo salt or an addition compound of H_2O_2 was ultimately resolved by Hansson in 1961 by a single-crystal X-ray structure analysis [35].

The molecular structure (Fig. 3) consists of dimeric $[(HO)_2BOO]^-$ units, which form cen-

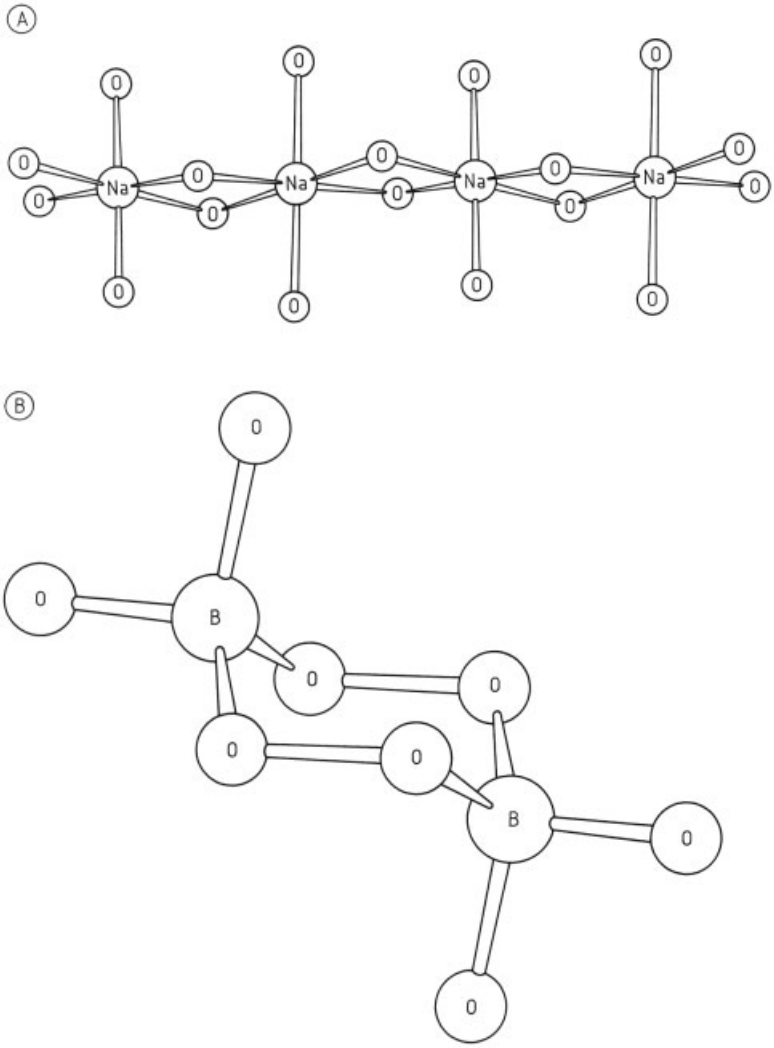

Figure 3. A) Part of the chain of hydrated Na^+ ions in sodium peroxoborate tetrahydrate; B) Structure of the $[BO_2(OH)_2]_2^{2-}$ anion

trosymmetric cyclic anions with two peroxo bridges. The sodium cations are coordinated by the water of crystallization and by certain hydroxide groups of the six-membered ring. The anions and cations form a polymeric structure.

An X-ray structural analysis of sodium perborate trihydrate [36], together with physicochemical investigations of the monohydrate, has shown that these compounds also contain the above-mentioned six-membered ring structure [37]. A molecular structure for the anhydride has been discussed in which the six-membered ring is a boron–oxygen radical [38].

Vapor-pressure studies on sodium perborate hydrates [26] have shown that the isotherms for compound formation and decomposition as a function of water vapor partial pressure exhibit marked hysteresis; that is, the sodium perborate hydrates do not behave as classical hydrates, for which the formation and decomposition reactions are reversible.

The thermal decomposition of sodium perborate hydrates has been investigated by thermogravimetric analysis. The decomposition curves for the tetrahydrate and trihydrate are very similar. At temperatures above 150 °C they corre-

spond to the curve obtained with sodium perborate monohydrate [37].

Physical Properties.

Sodium perborate tetrahydrate, M_r 153.9, ϱ 1.731 g/cm^3, theoretical active oxygen content 10.38 wt %, melts in its own water of crystallization at 65.5 °C.

The commercial product is a white, crystalline solid with a spherolitic or elongated particle shape, and a particle size distribution of 0.1–1.0 mm. Depending on the crystal shape, bulk densities of 0.65–0.9 kg/L can be obtained. When stored under cool, dry conditions the rate of loss of active oxygen is only 1 % per year. Decomposition is accelerated at higher temperatures and moisture content. There is a tendency for increased caking at higher temperatures caused by conversion of the tetrahydrate to the trihydrate [39].

The solubility in water is 0.152 mol/L at 20 °C and 0.388 mol/L at 40 °C.

Thermogravimetric analysis curves for sodium perborate tetrahydrate show that decomposition takes place in three stages: 20–150 °C, 150–165 °C, and 165–400 °C. In the first stage, the tetrahydrate is dehydrated to the monohydrate.

Sodium perborate trihydrate, M_r 135.9, ϱ 1.86 g/cm^3, theoretical active oxygen content 11.8 wt %, melts in its own water of crystallization at 81.7 °C.

Solubility and vapor pressure determinations indicate that above 15 °C sodium perborate trihydrate is the thermodynamically more stable phase. The tetrahydrate is stable below this temperature. However, the transition requires considerable activation energy and therefore only occurs in solution at elevated temperature or upon addition of substances which reduce the solubility [40].

Sodium perborate monohydrate, M_r 99.8, has a theoretical active oxygen content of 16.0 wt %. The anhydrous compound has no definite melting point and decomposes when strongly heated, liberating water and oxygen. The commercial product is manufactured by dehydration of sodium perborate tetrahydrate and is a white crystalline solid with a bulk density of 0.5–0.65 g/cm^3, depending on the starting material used. Modern methods of direct synthesis give a product with a bulk density between 0.4 and 1 g/cm^3 [41, 42]. A major advantage of sodium perborate monohydrate compared to the tetrahydrate, apart from its higher active oxygen content, is its much higher rate of dissolution, which results in better bleaching performance under short wash cycle conditions. The time required for 2 g tetrahydrate to dissolve in 1 L water at 15 °C is 6–8 min, while the same amount of monohydrate dissolves in < 1 min.

Anhydrous sodium perborate, also known under the trade name Oxoborate, has a theoretical molecular mass of 81.8. However, the commercial product is heterogeneous and contains ca. 2 wt % residual active oxygen. The releasable oxygen is liberated spontaneously as molecular oxygen when anhydrous sodium perborate comes into contact with water.

The releasable oxygen is determined by adding a weighed amount to dilute sulfuric acid, and measuring the oxygen liberated either gravimetrically or volumetrically. Oxoborate has a pale yellowish color. It is X-ray amorphous, and has paramagnetic properties [43].

Chemical Properties. Cryoscopic molecular mass determinations and conductivity measurements show that in aqueous solution the sodium perborate hydrates are partially hydrolyzed, reforming the starting materials metaborate and hydrogen peroxide according to the following equation [37]:

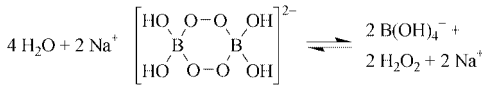

As can be seen from this equation, the amount of free H_2O_2 cannot be determined by titration, since this continually displaces the equilibrium.

The aqueous solution of sodium perborate reacts like an alkaline solution of hydrogen peroxide, the active oxygen content being catalytically decomposed by heavy metal ions. Stabilization can be achieved by adding magnesium salts. The pH is 10.1–10.4 and is virtually independent of the concentration. The active oxygen content can be determined by titration with potassium permanganate in dilute sulfuric acid. The sodium content is measured by titration with hydrochloric acid using methyl orange as the indicator. Finally, the boron content is determined by reacting the solution with a polyhydric alco-

hol, e.g., glycerine, leading to complex formation and producing a monobasic acid which can be titrated with sodium hydroxide solution using phenolphthalein as the indicator.

3.1.2. Production

The former electrochemical production of sodium perborate tetrahydrate is no longer of commercial importance [44]. The demand for higher product quality has led to the development of the process described below, almost the only one used today on a large scale.

Raw Materials. The boron-containing raw material is generally a water-soluble sodium tetraborate. Technical-grade borax pentahydrate is used, but impure minerals such as kernite and tincal are also employed.

The minerals ulexite (sodium calcium borate), colemanite and pandernite (calcium borates) are also used but to a lesser extent because they must be reacted with sodium carbonate, producing considerable quantities of sludge. The active oxygen component is usually hydrogen peroxide.

Sodium perborate tetrahydrate is produced in a two-stage process (Fig. 4). In the first stage, the boron minerals are reacted with aqueous sodium hydroxide to yield a highly concentrated solution of sodium tetrahydroxoborate:

$$Na_2B_4O_7 + 2\,NaOH + 7\,H_2O \rightarrow 4\,NaB(OH)_4$$

Recycled mother liquor from the crystallization stage is mixed with sodium hydroxide solution and the boron-containing raw material added, whereby the temperature rises from 60 to 95 °C [45]. Insoluble components are then removed by decantation or filtration. The resulting clear solution containing up to 800 g/L solid (expressed as $NaBO_2$) is diluted with more recirculated mother liquor and is subsequently cooled to below 40 °C. In the second stage, the diluted solution of sodium tetrahydroxoborate is treated with hydrogen peroxide with cooling at temperatures below 30 °C. This causes sodium perborate tetrahydrate to precipitate:

$$2\,NaB(OH)_4 + 2\,H_2O_2 + 2\,H_2O \rightarrow Na_2\left[BO_2(OH)_2\right]_2 \cdot 6\,H_2O$$

The perborate is stabilized by adding a solution of magnesium sulfate or alkali metal or alkaline earth metal silicate, either before or after the crystallization stage [46].

Both the chemical reaction and the crystallization can have several variations. The two components can be reacted continuously in the crystal-free external recirculation zone of a vacuum crystallizer, and then caused to flow through the salt layer, so that the solution is no longer supersaturated [47]. Depending on the raw materials used and the process, different crystal shapes and crystal size distributions can be obtained.

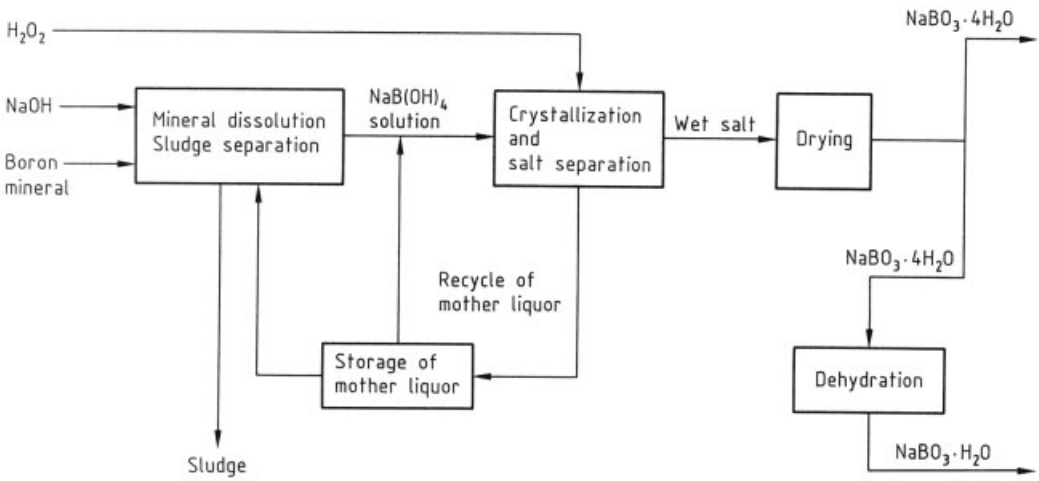

Figure 4. Flow diagram for the production of sodium peroxoborate

Degussa has described a process for the production of attrition-resistant sodium perborate tetrahydrate crystals. A highly supersaturated solution (200–700 g/L) of sodium perborate is passed into a previously prepared suspension of small crystals. Crystallization initiated by these nuclei yields coarse, compact crystal agglomerates [48].

Kali-Chemie has reported a continuous process in which controlling the stoichiometry produces attrition-resistant, free-flowing crystals of sodium perborate tetrahydrate with a bulk density of 0.75–0.85 kg/L [49]. Crystallization is carried out in an open, cooled, stirred vessel. The sodium perborate suspension formed is continuously removed and fed at a controlled rate directly into an automatic centrifuge. The mother liquor is recirculated. The separated crystals, containing 3–10 wt % water, are dried in a rotary or fluidized-bed dryer with air at ca. 100 °C [49].

During drying, the temperature of the salt must be kept below 60 °C to prevent melting. After drying, the product is cooled to ca. 25 °C by air in fluidized-bed coolers. Fine particles of perborate are removed from the exhaust air by wet scrubbers, integrated in the mother liquor recirculation system.

Major producers of sodium perborate tetrahydrate are Belinka (Slowenia), Degussa (Belgium, Austria), Etibank (Turkey), FMC Foret (Spain), Solvay (Australia, Germany).

Sodium perborate monohydrate is produced by dehydration of sodium perborate tetrahydrate. By removing the water of crystallization with warm air in a fluidized-bed dryer, an attrition-resistant product is obtained. The inlet temperature of the air is maintained between 180 and 210 °C [50].

A highly attrition resistant product is obtained by dehydration under vacuum [51], or in a fluidized-bed dryer operating at a relative humidity of 10–40 % and an air inlet temperature of 100–180 °C [52]. Alternatively a relative humidity of 40–80 % and an air exhaust temperature of at least 60 °C can be employed [53]. Besides the conventional processes with tetrahydrate as an intermediate, new production technologies have been described in which sodium perborate monohydrate is produced directly from aqueous solutions.

According to an Interox patent, attrition-resistant granules of sodium perborate monohydrate with a bulk density of 0.4–1 kg/L are obtained by spraying solutions of H_2O_2 and $NaBO_2$ onto a fluidized bed of crystal nuclei at up to 100 °C [42]. The rate of dissolution of these granules is rather low compared to that of monohydrate manufactured by conventional processes but can be improved by adding surface-active agents [54].

Degussa has described the direct synthesis of a sodium perborate monohydrate product with a bulk density of 0.4–1 kg/L. Sodium perborate solution is spray-dried to produce the monohydrate in the form of a fine powder. The fines are then granulated in a fluidized bed or granulated by compression to yield a product with good storage stability and dissolution properties [41].

Major producers of sodium perborate monohydrate are Degussa (Belgium, Austria), FMC Foret (Spain), Solvay (Germany).

Anhydrous sodium perborate is produced by further dehydration of sodium perborate monohydrate. Sodium perborate monohydrate is dehydrated in a fluidized bed with dry air at 150–160 °C [56]. The dehydration of the tetrahydrate requires a stepwise, slow temperature increase using dry, CO_2-free air to avoid agglomeration of the particles in the fluidized bed [57]. Both processes yield products with 15–17 wt % releasable oxygen and 1–2 wt % active oxygen.

3.1.3. Uses

Sodium perborate tetrahydrate and monohydrate are still used with decreasing tendency as active oxygen source in oxidizing and bleaching agents in cleaning, cosmetic, and pharmaceutical products, but mainly in detergents (→ Laundry Detergents). It is estimated that ca. 95 % of the perborates are consumed by this market. Detergents contain up to 30 wt % sodium perborate tetrahydrate and/or up to 15 wt % monohydrate. The monohydrate is preferred if rapid dissolution, improved storage stability, and compatibility with other constituents such as activators and water-softening agents are required. Other applications include detergents for automatic dishwashers, denture cleaners, color-safe bleach powders, and hair-treatment prepa-

rations. Sodium perborate has been used as a slow-release oxygen source in heap leaching of gold and silver ores [55].

Anhydrous sodium perborate is mainly used in denture cleanser tablets to enhance dissolution and to produce effervescence.

In 1998 the world production capacity for perborates (as sodium perborate tetrahydrate) was estimated at 550 000 t of which approximately 85 % was produced in Europe. With changes in European environmental policy and the discussion on the ecotoxicological properties of borates and perborates at the end of the 20th century, a continuous shift from sodium perborates to sodium carbonate peroxohydrate took place, and this entailed a significant reduction in production capacity estimated in 2004 to be less than 300 000 t. Especially Solvay converted existing sodium perborate production facilities to sodium carbonate peroxohydrate production.

3.1.4. Product Specifications

Commercial sodium perborate tetrahydrate and monohydrate have an active oxygen content of at least 10 and 15 wt %, respectively. They are white, odorless, free-flowing salts with a particle size distribution of 95 % between 0.1 and 1.0 mm. Modifications in the production process can give bulk densities for the tetrahydrate between 0.65 and 0.85 kg/L, and for the monohydrate between 0.50 and 0.65 kg/L. With storage under cool, dry conditions, the annual active oxygen loss rate is $< 2\,\%$ (relative). The dissolution time for 2 g substance in 1 L water is ca. 6–8 min for the tetrahydrate and < 1 min for the monohydrate. The residual moisture content, determined by drying for 30 min at 60 °C in a vacuum, should not exceed 1 wt % for the monohydrate. Attrition resistance is an important parameter and should not be higher than 10 % according to ISO 5937.

Commercial anhydrous sodium perborate has at least 13.0 wt % releasable oxygen and ca. 1.5–2.5 wt % active oxygen. It also contains small amounts of sodium perborate monohydrate and sodium metaborate, depending on the production process. The light yellow salt has a particle size distribution of ca. 90 wt % between 0.1 and 1.0 mm and a bulk density between 0.3 and 0.4 kg/L. It dissolves readily in water with evolution of oxygen.

3.1.5. Storage and Transportation

Sodium perborates must be stored under cool, dry conditions. Exothermic decomposition with liberation of oxygen can be caused by heat, moisture, and/or contact with decomposition catalysts, e.g., heavy metals, alkali, or dirt. Adiabatic storage conditions can lead to self-accelerating decomposition of sodium perborate monohydrate.

After manufacture, sodium perborate tetrahydrate and monohydrate may be stored in silos and transported in bulk by rail or truck or packed in paper sacks, optionally with polyethylene inner liners, or in polyethylene sacks. Oxoborate (anhydrous sodium perborate) is transported only in polyethylene-lined paper sacks.

Sodium perborate tetrahydrate is considered nonhazardous for international transportation purposes. Due to a change in testing methods for oxidizing substances, the UN classification for sodium perborate monohydrate changed recently to Oxidizer 5.1. [59]. Oxoborate has been assigned UN No. 3247 and is also classified as a Class 5.1 oxidizing substance.

3.2. Other Peroxoborates

Melikoff and Pissarjewski prepared peroxoborates of magnesium, calcium, barium, and strontium in 1898. Production processes for potassium peroxoborate [58, 60] and the peroxoborates of the alkaline earth metals have been described in patents [61]. Novel peroxoborates with monovalent cations and considerable molar excess of H_2O_2 have been published by Heller and Pawel [62]. However, these compounds now have no commercial importance.

4. Peroxosulfuric Acids and Peroxosulfates

Peroxomonosulfuric acid and peroxodisulfuric acid are intermediates in the former electrochemical production of hydrogen peroxide (→ Hydrogen Peroxide, Section 4.3). Peroxodisulfuric acid is not of commercial importance.

However, its salts, the peroxodisulfates, are widely used as active oxygen carriers and oxidizing agents. Industrially important salts of peroxodisulfuric acid are those of ammonium [*7727-54-0*], sodium [*7775-27-1*], and potassium [*7727-21-1*].

Peroxomonosulfuric H_2SO_5 in the form of Caro's acid has found industrial applications in the mining industry and in pulp bleaching. The salts of peroxomonosulfuric acid include the peroxomonosulfates of the alkali, alkaline earth, and other metals [63], e.g., $KHSO_5$ [*25482-78-4*] [64], and the triple salt $2\ KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$ [*37222-66-5*] [65], which is sold under the trade names Caroat (Degussa) and Oxone (DuPont).

4.1. Peroxomonosulfuric Acid

Peroxomonosulfuric acid [*7722-86-3*] is a colorless crystalline solid, *mp* 45 °C, that is difficult to prepare in high purity. Solutions of Caro's acid lose active oxygen significantly faster than hydrogen peroxide solutions. The stability of peroxomonosulfuric acid solutions increases with decreasing concentration and residual H_2O_2 content [66]. Peroxomonosulfuric acid is a strong acid. The proton bound to the peroxo group is, like that in H_2O_2, a weak acid with a pK_a of 9.4. In the presence of water it hydrolyzes rapidly yielding H_2O_2 and sulfuric acid.

Peroxomonosulfuric acid is a powerful oxidizing agent:

$$H_2SO_5 + 2\,H^+ + 2\,e^- \rightarrow H_2SO_4 + H_2O \quad E^\circ = -1.81\ V$$

The oxidation potential of peroxomonosulfuric acid was determined by a thermodynamic method to be 1.81 V [67, 68]. It oxidizes cyanide to cyanate, chloride to chlorine, sulfide to sulfate, and formaldehyde to H_2 and formic acid. Under neutral and alkaline conditions, peroxomonosulfate oxidizes hydrogen peroxide yielding oxygen gas. Peroxomonosulfuric acid exhibits significantly better bactericidal properties than H_2O_2 and has good overall effectiveness against bacteria and viruses, in particular under weakly acidic conditions [69].

4.2. Caro's Acid

Caro's acid is named after Heinrich Caro (1834–1910) who first described its synthesis and properties in 1898.

Production. Caro's acid obtained by treating sulfuric acid with hydrogen peroxide is an equilibrium mixture of peroxomonosulfuric acid, hydrogen peroxide, sulfuric acid, and water. The equilibrium constant for this reaction is 0.1 [70].

$$H_2O_2 + H_2SO_4 \rightleftharpoons H_2SO_5 + H_2O$$

The highly exothermic reaction requires removal of the heat of reaction to give high yields of peroxomonosulfuric acid. There are two manufacturing technologies: the isothermal and the adiabatic process.

In the *isothermal process*, 70 % H_2O_2 is mixed with 93–98 % sulfuric acid with efficient heat removal, so that the temperature does not exceed 15–30 °C. If equimolar quantities of reactants are employed, the product contains 42–50 % H_2SO_5 and ca. 10 % H_2O_2. H_2SO_5 yields can be significantly increased by increasing the amount of H_2SO_4 or by using oleum. Degussa describes an on-site manufacturing process in which H_2O_2 is added in separate reaction stages in series resulting in ca. 15 % higher yields over conventional single-stage reactors when a 2:1 molar ratio of H_2SO_4 (96–98 %) and H_2O_2 (70 %) is employed [71, 72].

In the *adiabatic process*, the reactants are mixed rapidly in a small-volume, high-throughput static mixer without cooling. Several patents describe various designs of equipment for adiabatic generation of Caro's acid [73 – 77]. Typically, the yields of Caro's acid with this method are in the range of 25–45 %. A process for the production of stable aqueous sodium peroxomonosulfate solutions has been described by Degussa [66].

Uses. Due to new and cost-effective on-site generation technologies, Caro's acid has found increasing use in mining and hydrometallurgy, pulp bleaching, and the electronics industry.

In the mining industry, Caro's acid is used as a detoxifying agent for residual cyanide in the tailings from gold and silver ore processing. Due to its faster oxidation kinetics and oxidizing power, Caro's acid has several advantages over

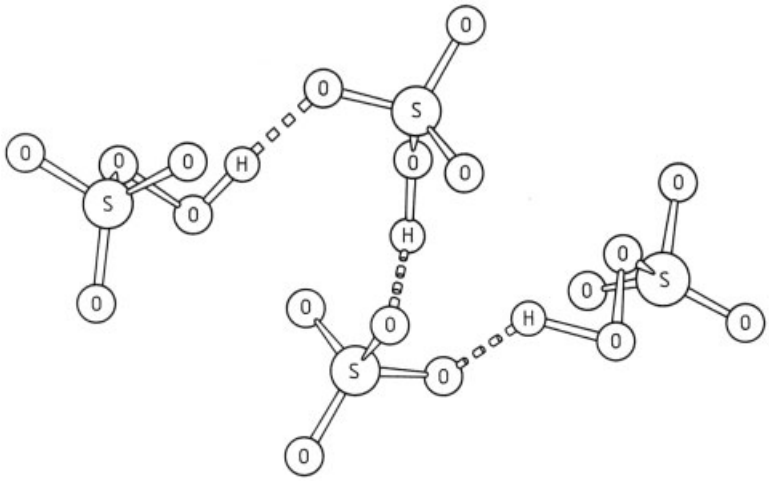

Figure 5. Arrangement of the complex anions in the compound 2 $KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$

catalyzed H_2O_2 (e.g., it also oxidizes thiocyanate). This application is practiced at several gold mines in North America [78 – 82]. Caro's acid also has been described as an oxidant in uranium ore processing.

In the pulp and paper industry, Caro's acid has proved to be an effective delignification chemical in totally chlorine free bleaching sequences; it replaces chlorine and chlorine dioxide. When employed under controlled pH and temperature conditions, the mechanical properties of the cellulose fibers are not impaired [83 – 86]. DuPont describes a delignification process using a combination of Caro's acid and peracetic acid [87, 88]. Caro's acid is also the starting material for the manufacture of peroxomonosulfates (see Section 4.3.2).

A proprietary form of Caro's acid is marketed in the electronics industry under the name of Nanostrip. It is used for cleaning and reclaiming of silicon wafers. Manufacturers are Cyantech (US), Micro-Image Technologies (UK), and RASA Industries (Japan).

4.3. Peroxomonosulfates

4.3.1. Properties

A number of sodium, potassium, ammonium, cesium, lithium, and rubidium salts of peroxomonosulfuric acid have been described in the literature. Only the potassium salt in the form of the triple salt 2 $KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$, M_r 614.8, ϱ 2.313 g/cm^3, theoretical active oxygen content 5.2 wt %, has gained commercial importance. Recent X-ray structure analysis investigations [89] have shown that it crystallizes in the monoclinic system, contrary to earlier reports describing an orthorhombic structure [65]. The crystal lattice consists of an array of K^+ cations and $H_3S_4O_{18}^{5-}$ anions in which HSO_5^-, HSO_4^- and SO_4^{2-} units are also included (Fig. 5).

The commercial product is a white, odorless, free-flowing material with a bulk density of 1–1.2 kg/L. About 90 wt % is in the particle size range 0.1–0.8 mm. Under proper storage conditions, the loss of active oxygen is < 1 % per month. The triple salt is soluble in water (25 g/L H_2O at 20 °C) and dissolves rapidly. The pH of an 1 wt % solution is ca. 2.3. The bactericidal properties of the triple salt are summarized in [90].

The triple salt is classified by the UN as corrosive solid under UN No. 3260.

4.3.2. Production

The starting materials are sulfuric acid or oleum, hydrogen peroxide, and potassium hydroxide. In the first step, hydrogen peroxide (65–90 wt %) and oleum (53–82 wt %) in a molar ratio of 1:1.3–2.7 are continuously mixed in a cooled reaction vessel at a temperature below 8 °C to obtain a 37–66 wt % peroxomonosulfuric acid containing H_2SO_4, H_2O, and H_2O_2 [91].

In the second stage the mixture is partially neutralized with potassium hydroxide and the triple salt is crystallized. This step requires careful control of the mother liquor composition to obtain high yields [65] (see Fig. 6).

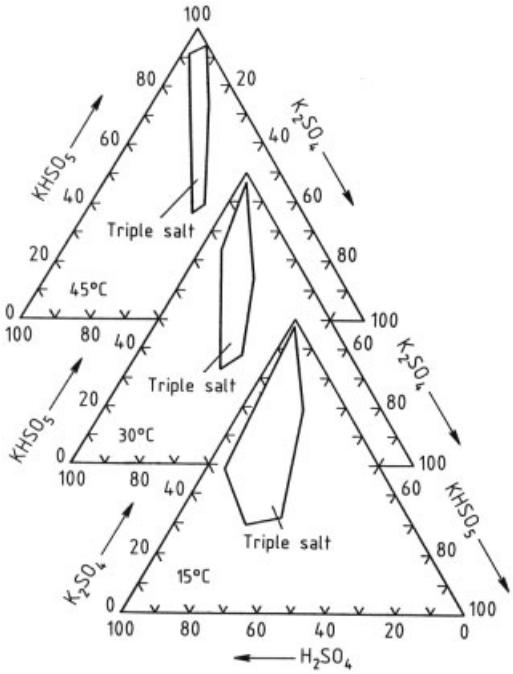

Figure 6. Region of stability of the triple salt 2 $KHSO_5 \cdot KHSO_4 \cdot K_2SO_4$ as a function of the composition of the mother liquor and temperature

According to a patent of the former Peroxid-Chemie (today Degussa), peroxomonosulfuric acid is produced from H_2SO_4 (90–100 wt %) and H_2O_2 (30–90 wt %), and then continuously fed, together with a concentrated solution of potassium hydroxide, into a circulating mother liquor containing 25–30 wt % $KHSO_5$ and ca. 15 wt % H_2SO_4 or K_2SO_4 in aqueous solution [92]. The water introduced with the raw materials and formed in the neutralization reaction is removed by passing the reaction mixture through vacuum treatment equipment at ca. 25 °C and 1 kPa. Some of the circulating liquor is continuously removed and cooled to ca. 5 °C in a crystallizing vessel, resulting in precipitation of the triple salt.

The salt is isolated by centrifugation, and the mother liquor is fed back into the circulating solution. The filter cake is washed while in the centrifuge, and the wet salt is dried in a fluidized-bed dryer [92]. The product is packed in polyethylene-lined plastic sacks and is stored under cool, dry conditions.

4.3.3. Uses

Peroxomonosulfate triple salt is used because of its powerful oxidizing properties and bactericidal effects in surface cleaners and dishwashing powders to improve bleaching and disinfecting action, especially at low temperatures [93 – 95]. Ketones have been reported to enhance the bleaching power of peroxomonosulfates by in situ formation of highly reactive dioxiranes [96 – 98].

A major use of the triple salt is in denture cleaners. In combination with sodium perborate monohydrate it gives an effervescence due to reaction with perborate yielding oxygen gas under alkaline conditions. In more recent formulations of denture cleaners Caroat is also combined with sodium percarbonate instead of sodium perborate monohydrate. The triple salt is used in the textile industry for nonshrinking and nonfelting treatment of wool, and in the electrical industry for etching printed circuits. In the electroplating and metal finishing industry, potassium peroxomonosulfate triple salt is employed for the treatment of cyanide-containing industrial wastewater and as a mild etchant and pickling agent [99, 100].

In water treatment it has found application as a shock treatment agent for swimming pool water [101]. It quickly oxidizes malodorous chloramines to nitrogen and nitrate.

Another application is its use in wastepaper recycling for destruction of wet-strength resin [102].

4.3.4. Analysis

Peroxomonosulfuric acid can be determined in the presence of H_2O_2 and peroxodisulfuric acid by redox titration, based on the reaction of the H_2SO_5 with arsenic acid and subsequent titration of the residual H_2O_2 with cerium(IV) sulfate solution. The peroxodisulfuric acid is determined by the addition of arsenic acid and ICl and

titration of excess arsenic acid with $Ce(SO_4)_2$ [103].

5. Peroxodisulfuric Acid and Peroxodisulfates

5.1. Properties

The dibasic peroxodisulfuric acid [*13445-49-3*] also known as Marshall's acid, HO_3SOOSO_3H, forms white crystals (*mp* 65°C decomp.).

The hydrolysis of peroxodisulfuric acid to sulfuric acid and peroxomonosulfuric acid is irreversible, while further hydrolysis of peroxomonosulfuric acid is reversible:

$$H_2S_2O_8 + H_2O \rightarrow H_2SO_4 + H_2SO_5$$

$$H_2SO_5 + H_2O \rightleftharpoons H_2SO_4 + H_2O_2$$

The hydrolysis of peroxodisulfuric acid is favored by increasing temperature and acid concentration [104 – 107]. Sulfate radicals are postulated as an unstable intermediate [108]. The peroxodisulfate ion in aqueous solution is one of the most powerful oxidizing agents (E^0 = 2.05 V) [109].

The peroxodisulfates of ammonium, sodium, and potassium are white, crystalline, and stable under dry conditions. The salts decompose on heating in two steps: first the formation of pyrosulfates and oxygen and then upon further heating the decomposition of the pyrosulfate into sulfate and sulfur tri- and dioxide [110, 111]. The decomposition temperatures increase in the series $NH_4^+ < K^+ < Na^+$, and decrease with increasing moisture content. Mixtures with combustible materials are easily ignited and burn vigorously even in the absence of oxygen.

In aqueous solution, the salts are hydrolyzed in a similar manner to peroxodisulfuric acid. Aqueous solutions of peroxodisulfates are slightly acidic (25 wt % $(NH_4)_2S_2O_8$, pH 2.30; 25 wt % $Na_2S_2O_8$, pH 4.35). Solubilities at 20 °C in 100 cm^3 water are: ammonium peroxodisulfate 55.9 g, sodium peroxodisulfate 55.6 g, and potassium peroxodisulfate 5.0 g.

5.2. Production

Pure peroxodisulfuric acid is formed under the same conditions as peroxomonosulfuric acid, i.e., from chlorosulfuric acid and hydrogen peroxide [112]. It has no industrial application.

Aqueous solutions of peroxodisulfuric acid are manufactured by anodic oxidation of sulfuric acid:

$$2\,HSO_4^- - 2\,e^- \rightarrow S_2O_8^{2-} + 2\,H^+ \quad E^0 = 2.057\,V$$

$$2\,SO_4^{2-} - 2\,e^- \rightarrow S_2O_8^{2-} \quad E^0 = 1.939\,V$$

This electrochemical synthesis followed by hydrolysis was formerly the most important industrial manufacturing method for hydrogen peroxide before the development of the autoxidation process.

The electrode process to form $S_2O_8^{2-}$ is complex [113, 114] and is accompanied by the undesired evolution of oxygen. To minimize the evolution of oxygen, electrolysis is performed at high current densities [115] with electrodes having a high oxygen overpotential, e.g., platinum. In addition, potential-increasing substances, e.g., halides or pseudohalides, are added to the electrolyte [116, 117]. The most important process that lowers the current yield is the thermodynamically favored acid-catalyzed hydrolysis of the peroxodisulfate anion to give peroxomonosulfate, whose anodic reaction leads to increased evolution of oxygen [118]. Therefore, the following requirements for the design have to be considered in order to achieve a high current efficiency:

1) Short residence time in the cell, no backmixing of the anolyte, but high final concentration of the product
2) High anodic current density with minimum heating of the electrolyte
3) Effective dissipation of heat through heat exchangers with large surface areas
4) Small anolyte gap to minimize voltage losses
5) Dimensionally stable anode construction with minimum use of platinum

The first successful industrial realization was the Degussa–Weißenstein cell [119]. A further development is a bipolar electrolyzer [120 – 122] with ion-exchange membranes [123], which is reviewed in [124, 125]. A bipolar electrode plate (see Fig. 7) is the basic element in the filter-press electrolysis cell, which consists of 26 electrodes [124, 126]. The characteristic

feature is the large number of vertical, parallel anolyte channels (9 × 3 mm). The anodes are ribbonlike platinum foils on a carrier, which is cooled from behind. Anolyte and catholyte are separated by a microporous polymer diaphragm or, in a more recent development, by a cation-exchange membrane [127].

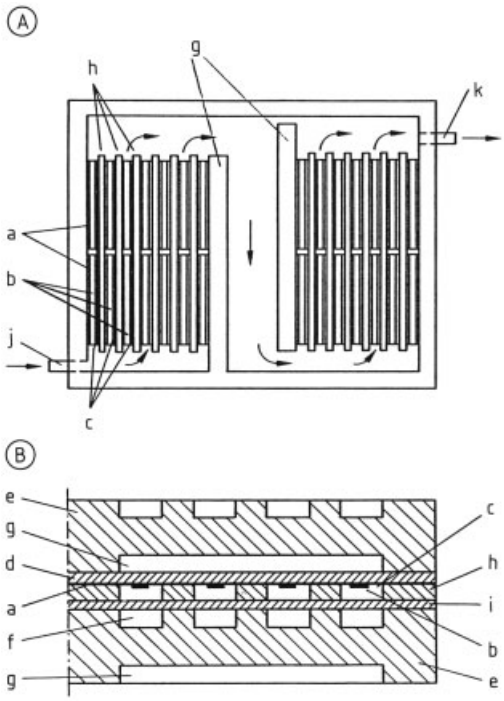

Figure 7. Schematic view of a frame and plate electrolyzer
A) Anode plate; B) Cross-sectional view
a) Tantalum foil; b) Platinum foil strip anodes; c) Anode channel; d) Supporting plate; e) Graphite cathode block; f) Cathode channel; g) Cooling channel; h) Spacer; i) Diaphragm; j) Anolyte inlet; k) Anolyte outlet

The electrolytic conditions for the production of crystalline peroxodisulfate salts are determined by the solubilities of the starting materials and products in the electrolytes, and these in turn are dependent on the proton activity. Fundamental investigations are reported in [128 – 138].

The current efficiency increases in the series $H^+ = Li^+ < Na^+ < NH_4^+$ $K^+ < Rb^+ < Cs^+$ due to the electrostatic effects of the cations in the electrolytic double layer at the electrode/electrolyte phase boundary [139]. The most favorable conditions with respect to solubility ratios and cationic effects are those associated with the production of ammonium peroxodisulfate. Therefore, the ammonium salt is still the most important peroxodisulfate produced electrolytically.

Figure 8 shows a flow diagram for the manufacture of ammonium and sodium peroxodisulfates. During electrolysis ammonium ions and water of hydration enrich the catholyte by migration through the diaphragm. The resulting catholyte excess is mixed with the anolyte to replace the consumed sulfate. After crystallization and separation of ammonium peroxodisulfate, the residual solution is neutralized with ammonia.

Typical operational data for the production of ammonium peroxodisulfate are as follows:

Current density, A/m^2	0.5–1
Cell voltage, V	4–6
Current efficiency, %	75–90
Specific energy consumption, kW · h/kg	1–2
Electrolyte temperature in anolyte, °C	30–45
Anolyte concentrations, g/L	
Inlet cell	
$(NH_4)_2SO_4$	380–440
$(NH_4)_2S_2O_8$	100–160
H_2SO_4	0–150
$(NH_4)SCN$	0.05–0.40
Exit cell	
$(NH_4)_2S_2O_8$	300–350
H_2SO_4	10–150
$(NH_4)_2SO_4$	200–240
Catholyte concentration, g/L	
H_2SO_4	300–700
$(NH_4)_2SO_4$	300–550

The electrolytic process is described in [140]. Details of the conversion of ammonium peroxodisulfate to the sodium [141] and potassium [142] salts are given in [143]. Additionally methods for direct electrochemical synthesis of the sodium and potassium salts have been published [144 – 148]. The advantage of these methods is that special equipment for handling ammonia is not required. However, in the direct electrochemical synthesis of sodium and potassium peroxodisulfates, special conditions for the electrolyte concentration and composition have to be considered to avoid solubility and electrolysis problems [146 – 150].

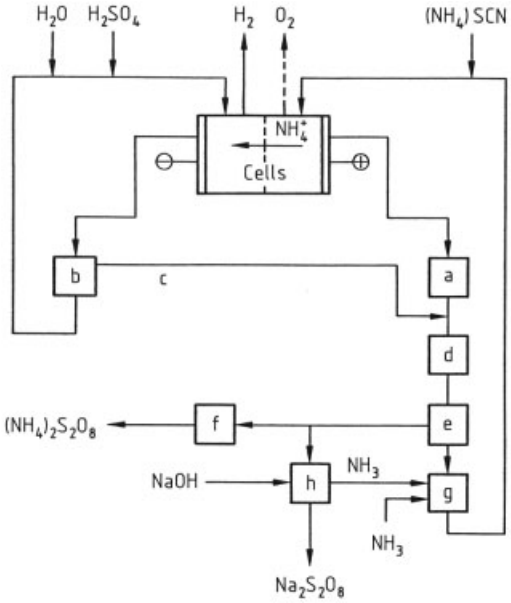

Figure 8. Flow sheet of an ammonium and sodium peroxodisulfate production plant
a) Anolyte; b) Catholyte; c) Overflow; d) Vacuum crystallization; e) Centrifuge; f) Drying; g) Neutralization of mother liquor; h) Reactor and ammonia stripping

5.3. Uses

The worldwide production capacity of peroxodisulfates in 2005 was estimated to be 165 000 t. The most important manufacturers are FMC (USA), Degussa (Germany), Rhein-PerChemie (Germany), and Mitsubishi Gas Chemical (Japan). Several smaller manufacturers are located in China, where additional capacities are about to be installed. The two largest market segments for peroxodisulfates are polymer production and metal etching, accounting for 85 % of the total market.

Peroxodisulfates are employed as radical initiators in combination with reducing agents and cocatalysts in emulsion polymerization processes for the manufacture of acrylonitrile–butadiene–styrene copolymers (ABS), high-impact polystyrene (HIPS), and styrene–acrylonitrile (SAN) [151, 152]. Redox systems that are commonly used in combination with peroxodisulfates are described in [153].

In the electronics industry peroxodisulfates are employed for etching printed circuit boards and removal of photoresists [154 – 156].

Peroxodisulfates have long been employed for desizing textile goods. A new application is as a bleaching agent in denim finishing [157 – 160].

A very important market are hair-coloring formulations which contain peroxodisulfates as the active ingredient [161, 162].

Sodium peroxodisulfate-based formulations are marketed as shock-treatment chemicals for swimming pool water [163, 164]. Peroxodisulfates are used in the preparation of adjustable-viscosity starch derivatives [165]. Peroxodisulfates have found also application in wastepaper recycling for the oxidation of wet strength resins [166]. An emerging new application is the purification of contaminated soil [167].

Peroxodisulfates have been described as oxidizing agents for photographic fixing baths [168 – 170].

5.4. Analysis

Peroxodisulfate is determined by oxidimetric titration with access of Fe^{2+} (e.g., iron(II) sulfate solution) and back-titration with potassium permanganate [171]. For mixtures with other sulfur compounds, separation by ion chromatography is used [172].

6. Peroxophosphoric Acids and Their Salts

The peroxophosphoric acids are structurally similar to the peroxosulfuric acids and behave similarly. They are hydrolyzed more rapidly than the peroxosulfuric acids, but the oxidation reactions of their anions are kinetically inhibited.

Peroxomonophosphoric acid [*13598-52-2*], H_3PO_5, is obtained by reacting phosphorus(V) oxide with hydrogen peroxide [173, 174]. The reaction proceeds less vigorously if the P_2O_5 has been heat-treated at 350–700 °C [175] or is suspended in an inert solvent [176]. A product containing 88 wt % H_3PO_5 and 9.6 wt % H_2O_2 is obtained. Peroxophosphoric acid is a colorless, viscous liquid. The three ionization constants are $pK_1 = 1.1$, $pK_2 = 5.5$ and $pK_3 = 12.8$ [177]. The salts can be obtained by

neutralization, e.g., with potassium hydroxide, which produces the relatively stable, crystallizable, but hygroscopic salt KH_2PO_5 [176 – 179].

Peroxodiphosphoric acid [*13825-81-5*], $H_4P_2O_8$, can be prepared by anodic oxidation of phosphoric acid. A mixture of peroxodiphosphoric and peroxoxomonophosphoric acid is obtained by reacting cold phosphoric acid with fluorine.

$$2\,H_3PO_4 + F_2 \rightarrow H_2P_2 + O_8 + 2\,HF$$

Peroxodiphosphoric acid is not produced commercially. Several salts have been reported, but only tetrapotassium peroxophosphate has been produced on a commercial scale.

Peroxodiphosphates Potassium peroxodiphosphate $K_4P_2O_8$ [*15593-49-4*], is prepared by electrolysis of potassium phosphate solutions containing added halide or pseudohalide anions [180]. In the older literature, fluorides are mainly recommended for this, but they cause corrosion problems [181]. Nitrates have been successfully tested as an alternative [182]. The electrolytic technique is analogous to that used in the manufacture of peroxodisulfates (see Section 4.3.2) [183, 184]. Sodium and ammonium peroxodiphosphates are obtained by reacting $K_4P_2O_8$ with the corresponding perchlorates, sometimes in the presence of free perchloric acid [185]. Only the secondary and quaternary peroxodiphosphates are obtainable in a crystalline state, with 6 or 18 molecules of water of crystallization. The hexahydrates can be dehydrated under vacuum at 65 °C. Mixed salts such as $K_2(NH_4)_2P_2O_8$ can be obtained by salting out after partial removal of the potassium ions as $KClO_4$ [186].

Potassium peroxodiphosphate is very soluble in water; a saturated solution of $K_4P_2O_8$ contains 42 wt % at 0 °C, 45 wt % at 25 °C, and 51 wt % at 45 °C. The peroxodiphosphates begin to lose oxygen on heating above 200 °C. The potassium salt is the most stable peroxodiphosphate compound [187 – 189]. In solution, the $P_2O_{84}^-$ ions undergo hydrolysis, which is accelerated in the presence of acids [190]. It is a strong oxidizing agent with an oxidation potential of 2.07 V.

The peroxophosphates have very few uses compared to the other peroxo compounds. They have been described as oxidizing agents in cosmetic hair preparations (bleaching, waving, and coloring agents) [191], as etchants for metal surfaces (e.g., aluminum) [192], and radical initiators in the graft polymerization of acrylic and vinylic monomers. However, today peroxophosphates have no commercial use due to their slower reaction kinetics and higher costs compared to the peroxosulfates.

7. Hydrogen Peroxide Addition Compounds

The hydrogen peroxide addition compounds of industrial importance are sodium carbonate peroxohydrate (sodium percarbonate) and urea peroxide (trade name: Percarbamid).

7.1. Sodium Carbonate Peroxohydrate

7.1.1. Properties

Sodium carbonate peroxohydrate [*15630-89-4*], sodium percarbonate, $Na_2CO_3 \cdot 1.5\,H_2O_2$, M_r 157.0, ϱ 2.14 g/cm^3, is a white salt with an orthorhombic unit cell [193]. Commercial crystallized sodium percarbonate consists of coarse, compact grains with a regular spherolitic crystal form [194]. The bulk density can range from 800 to 1200 g/L, depending on the synthesis conditions. The solubility in 100 mL water is 13.1 g at 10 °C, 15.4 g at 20 °C and 18.0 g at 30 °C [195]. The solubility of sodium percarbonate is reduced in the presence of other salts such as Na_2CO_3, Na_2SO_4, and NaCl.

The active oxygen content of technical sodium percarbonate varies between 12.0 wt % and 14.5 wt % (theoretical: 15.28 wt %), depending on the manufacturing process and stabilizing technology.

In contrast to sodium perborate, sodium carbonate peroxohydrate is an addition compound of hydrogen peroxide and sodium carbonate in which the hydrogen peroxide molecule is only weakly bound to the carbonate ion via hydrogen bonding, similar to water of crystallization. Thus, hydrogen peroxide is much more likely to be liberated from sodium carbonate peroxohydrate, for example, in the presence of moisture.

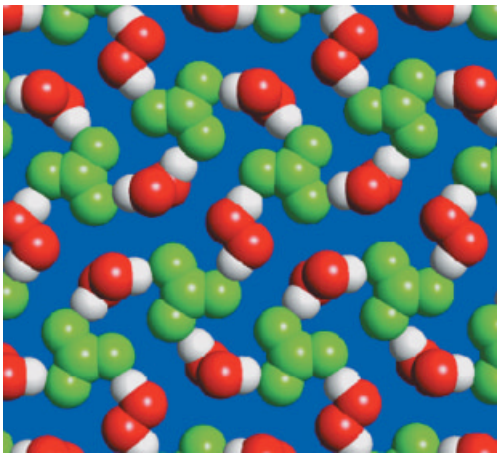

Figure 9. Computer model of crystalline sodium carbonate peroxohydrate

Figure 9 shows a computer model of crystalline sodium carbonate peroxohydrate.

When stored under cool, dry conditions, the rate of active oxygen loss is less than 0.3 % (relative) per month [195]. However, even in the presence of stabilizers, it is much more sensitive to higher temperatures and moisture than sodium perborate. Uncoated sodium percarbonate loses 82–86 % of the active oxygen within 10 d at a relative humidity of 96 % and a temperature of 30 °C [196], whereas sodium perborates loses < 2 %.

The stability of sodium carbonate peroxohydrate is mainly dependent on the presence of water. With increasing humidity the stability decreases as a result of accelerated exchange of the bonded hydrogen peroxide versus water. At the same time, exothermic decomposition of the liberated hydrogen peroxide in the sodium carbonate peroxohydrate, triggered by the caustic environment, yields molecular oxygen and water. Due to the weak bonding of hydrogen peroxide in the crystal lattice of sodium carbonate peroxohydrate, the process of lattice exchange is highly temperature dependent and thus can result in an autocatalytic decomposition reaction, especially if the material is not stabilized properly.

On dissolution of sodium carbonate peroxohydrate in water, the aqueous solution behaves like an alkaline solution of hydrogen peroxide.

Reliable methods for the determination of the decomposition properties of different sodium carbonate peroxohydrate qualities are highly sensitive microcalorimetric measurement methods such as thermal activity monitoring (TAM). Under isothermal conditions, the heat of reaction of the decomposition is measured at a defined temperature as a function of time [197]. The specific thermal output, normally measured in μW/g, is specified at less than 10 to ensure safe storage and handling.

The active oxygen content can be determined by titration with potassium permanganate solution in dilute sulfuric acid or by titration with cerium(IV) sulfate solution. The sodium content is obtained by titration with hydrochloric acid with methyl orange as indicator.

7.1.2. Production

Sodium percarbonate was first prepared by Tanatar in 1899 from hydrogen peroxide and soda ash. The product was isolated by precipitation with ethanol. It was not correctly identified as a hydrogen peroxide addition compound until 1909, when Riesenfeld and Reinold elucidated its structure [198]. Three major processes are used for the industrial manufacture of sodium percarbonate, yielding products with different properties with regard to bulk density, active oxygen content, and stability.

Crystallization Processes. In these processes the production of sodium percarbonate is mainly carried out by crystallization from aqueous solution in continuous or discontinuous manner. A concentrated and filtered solution of sodium carbonate is mixed with a concentrated aqueous solution of hydrogen peroxide, generally stabilized with magnesium salts or alkali metal silicates, in a crystallizer at 10–20 °C. The crystallizer has a clarification zone situated above a stirred zone across which an ascending current of supersaturated sodium carbonate peroxohydrate solution flows. Precipitating particles [199] move countercurrently as a result of gravity and are eluted into a classifier located directly below the crystallizer. The accumulated material at the bottom of the classifier is then removed with a rake. To improve the efficacy of the process salting-out agents like sodium chloride and crystallization aids such as polyphosphates or polyacrylates are used [194, 200 – 203].

The product is obtained as coarse, mechanically stable crystals (Fig. 10), which are cen-

trifuged followed by drying, preferably in a fluidized-bed dryer with an air inlet temperature of 75 – 130 °C. The mother liquor from the centrifuge can be further concentrated [204] and mixed with industrial-grade sodium carbonate (sometimes after adding a magnesium salt), filtered to remove iron compounds, and fed back into the crystallizer [205].

Processes which do not include a salting-out process with sodium chloride [204, 205] are usually less economic because of the lower space–time yield. Solvay describes a process without a salting-out agent, which yields a coarse sodium percarbonate product with enhanced thermal stability over products produced employing sodium chloride as salting-out agent [206].

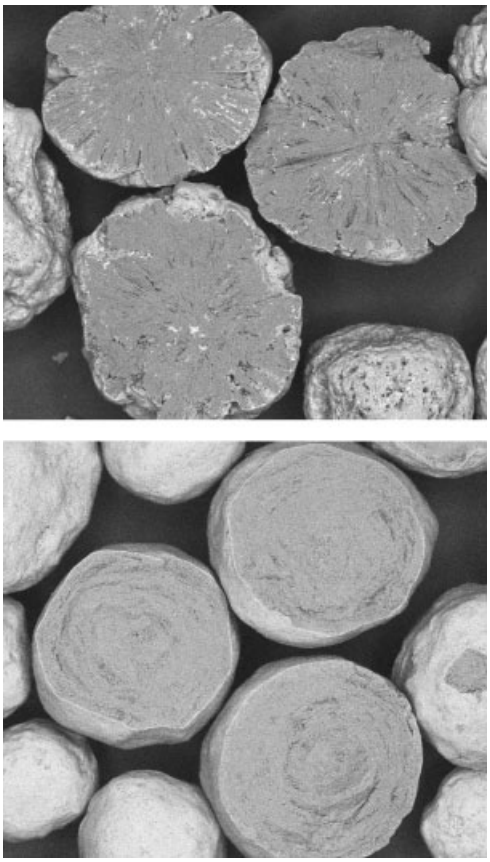

Figure 10. SEM pictures of sodium carbonate peroxohydrate
Top: Sodium carbonate peroxohydrate produced by a crystallization process with radial beamlike crystal growth lines
Bottom: sodium carbonate peroxohydrate produced by a spray-granulation process with onionlike growth structure

Spraying Processes. Several spraying processes for sodium percarbonate are described in the patent literature. These processes avoid the need for filtration and centrifugation of the crystalline product, and hence less waste is produced during production. In an early process described by Laporte, an aqueous solution or suspension of sodium carbonate and hydrogen peroxide is completely evaporated in a spray dryer [207, 208]. The product has a very low bulk density (ca. 0.35 kg/L) [208]. In further variations and optimizations of this process, described by Solvay, solutions of Na_2CO_3 and H_2O_2 are continuously sprayed separately onto a bed of sodium percarbonate fluidized with hot air [209]. Interox describes an improved continuous spray granulation process with higher hydrogen peroxide yields by using a single nozzle for both solutions. To ensure a long lifetime of the nozzle the addition of crystallization inhibitors like sodium hexametaphosphate to the feed solutions is critical [210].

In another process variation, Degussa describes a continuous spray-granulation process in a fluidized-bed reactor using a three-way nozzle. According to a patent [211], the previously necessary use of phosphorous-containing crystallization inhibitors to prevent blockages and incrustations in and on the nozzle can be avoided if the solutions are sprayed through a single three-substance pulverization nozzle with external mixing of the solutions. A highly attrition resistant product with an active oxygen content of $> 14\,\%$ is obtained [211]. Sodium carbonate peroxohydrate produced by a spray-granulation process is shown in Figure 10.

Dry Methods. In these processes, anhydrous or hydrated sodium carbonate (75–90 wt % Na_2CO_3) is treated with a concentrated solution (50–80 wt %) of hydrogen peroxide, so that only a small amount of water must be removed. This can be achieved in a mixer [212, 213] or in a fluidized-bed reactor [214], in which gaseous hydrogen peroxide can be used [215]. Apart from the long reaction time, these dry processes have the disadvantage that no purification of the sodium carbonate takes place, and therefore these sodium percarbonate products have only moderate storage properties. However, this problem can be solved by recrystalliz-

ing the sodium carbonate before use and by using sodium carbonate monohydrate [214]. FMC describes the use of phosphonic acids as additional stabilizers in the synthesis of sodium carbonate peroxohydrate, yielding a product with an active oxygen content of 11–14 % and improved stability [216].

7.1.3. Stabilization, Storage, Packaging, and Transportation

Active-oxygen-containing compounds have a pronounced tendency to decompose in an autocatalytic (self-accelerated) exothermic reaction. Sodium carbonate peroxohydrate is no exception and decomposes into molecular oxygen, water, and soda ash. There are two ways to stabilize sodium percarbonate. One is to improve the stability of the product (core or intrinsic stability) during the reaction process. The second is to stabilize the product by a separate protective coating to prevent moisture and other incompatible materials coming into direct contact with the peroxohydrate.

Intrinsic Stabilization. On the one hand, the intrinsic stability of sodium carbonate peroxohydrate is affected by the presence of heavy-metal impurities introduced by the raw materials sodium carbonate and hydrogen peroxide. Thus, precipitating agents such as silicates [217, 218] and/or magnesium salts [219], or complexing agents such as phosphates, and phosphonic acid [216], are commonly used. On the other hand, the intrinsic stability is also affected by the presence of excess water, mainly bound as crystal water in unconverted sodium carbonate hydrate or sodium hydrogencarbonate. Proper drying of the finished granulate can significantly improve the stability and reduce the decomposition tendency of sodium carbonate peroxohydrate [220].

Outer Stabilization. The outer stabilization of sodium carbonate peroxohydrate is necessary to improve the stability of the material in the presence of other detergent components. Especially protection against moisture brought into the detergent formulation by other ingredients is critical, since uncontrolled interaction of water with sodium carbonate peroxohydrate can result in loss of the active component. During production the absence of a coating or an insufficient coating can cause severe autocatalytic decomposition reactions accompanied by a strong temperature increase and volume expansion. A variety of coatings for percarbonate are described in the literature. Most of the coatings consist of inorganic substances such as borates [221, 222], perborax [223], sodium silicate [196, 224], sodium sulfate [225], and magnesium salts [226]. Mitsubishi describes a technology based on a two-layer coating with sodium silicate and a magnesium salt, which forms magnesium silicate on the surface of the granules [227]. Also organic coatings have been reported in the literature. Examples are chelating agents [228], fatty acids [229], and organic polymers [230, 231]. To control the liberation of hydrogen peroxide from sodium carbonate peroxohydrate Degussa developed a coating consisting of two layers: an inner layer of sodium sulfate for stability and an outer layer of waterglass for time-delayed release [232].

The degrees of inner and outer stability of sodium carbonate peroxohydrate can both be checked by measuring the microcalorimetric properties of the materials by TAM. Sodium carbonate peroxohydrate qualities with 40 °C TAM values of less than 10 μW/g after 48 h are considered to be stable during transportation, storage, and production, assuming proper handling of the material.

Sodium percarbonate must be stored under cool, dry conditions. Exothermic decomposition with liberation of oxygen can be accelerated by increased temperature, radiation caused by sunlight or steam piping, moisture, and/or contact with decomposition catalysts, e.g., heavy metals, alkali, or dirt. Adiabatic storage conditions can lead to self-accelerating decomposition. Exothermic decomposition may also occur during the manufacturing of detergents when sodium percarbonate is added to the still-warm base powder or when exothermic reactions, such as the rehydration of phosphates, proceed after the addition of sodium percarbonate.

Bulk storage of sodium percarbonate is more critical than storage of sodium perborates, in particular under hot and humid climatic conditions, due to its higher sensitivity to decomposition and caking.

Sodium percarbonate has been assigned recently to transport classification 5.1, oxidizing substances, by all major producers due to a change in the testing methods for oxidizing substances [59].

Sodium percarbonate is supplied in bulk by rail and truck, in big bags, and in 25 kg polyethylene sacks.

7.1.4. Uses

With the shift from sodium perborate to sodium carbonate peroxohydrate due to the ongoing discussion about the ecotoxicological properties of sodium perborate at the end of the 20th century, the most important areas of applications in terms of volume are nowadays all-purpose detergent powders and tablets. In 2004 it was estimated that the worldwide production capacity for sodium carbonate peroxohydrate was several hundred thousand tonnes. A smaller amount of ca. 10 % of sodium carbonate peroxohydrate is used in the growing market of automatic dishwashing detergents, especially tablets. Sodium carbonate peroxohydrate also experienced a revival as bleach additive, mainly as oxy products for stain removal and boosters for laundry and dishwashing detergents. In a niche application, sodium carbonate peroxohydrate plays an important role in denture cleaners, often in combination with peroxomonosulfates like Caroat.

Major producers are Solvay (UK, Germany, Italy, USA) Degussa (Germany, Italy), OCI (USA, Korea), Mitsubishi (Japan), Kemira (Sweden), FMC (Spain), JSC (Russia), Jinke (China) and Aktivsauerstoff Gesellschaft (Austria).

7.2. Urea Peroxohydrate

7.2.1. Properties

Urea peroxohydrate [*124-43-6*], $(NH_2)_2CO \cdot H_2O_2$, M_r 94.1, is a white, crystalline substance that melts with decomposition between 80 and 90 °C. At 20 °C, 100 cm^3 water dissolves 80 g urea peroxohydrate. The pH of aqueous percarbamid solution is 3.5–4.0. Urea peroxohydrate also exhibits good solubility in organic solvents, such as methanol (260 g/L), ethanol (120 g/L), glycerol (440 g/L), and various glycols. The bulk density is ca. 600 g/L.

The hydrogen peroxide content of the commercial products as powder is min. 35.0 wt %, and 33.4 wt % as tablets. The active oxygen content of these products is 16.5 and 15.7 wt %, respectively. When stored under cool and dry conditions, the active oxygen loss is less than 5 % per year. However, it decomposes on heating to form H_2O_2, NH_3, and CO_2. The H_2O_2 is released first with vigorous decomposition, especially in the presence of water [233].

7.2.2. Production

Urea peroxohydrate was first synthesized by Tanatar in 1908 by mixing 30 % H_2O_2 with an aqueous solution of urea. After concentrating the reaction mixture, the product was precipitated by cooling and separated by filtration.

The modern industrial process is based on the same principal. Powdered urea is dissolved in 35 % H_2O_2 in the presence of stabilizers, and the product is crystallized by cooling and concentration [234]. The precipitated solid material is separated from the mother liquor and dried at 30–40 °C. The mother liquor is further concentrated under vacuum and recycled. Alternatively, hydrogen peroxide can be sprayed onto urea in a fluidized bed. Excess moisture is removed by drying under vacuum at 30–40 °C [235].

7.2.3. Stabilization and Packaging

Urea peroxohydrate tends to decompose at elevated temperatures. The shelf life of the product can be improved by adding stabilizers such as NaH_2PO_4 or $Na_2H_2P_2O_7$ [236].

The product is supplied commercially in 25 kg packages as powder or tablets. Suitable packaging materials are polyethylene bags or polyethylene-lined containers.

7.2.4. Uses

Urea peroxide is mainly used in the cosmetics and pharmaceutical industry as a disinfecting and bleaching agent. Percarbamide has been employed as the active ingredient in tooth whiteners

and dentifrice formulations [237 – 240] which are sold in the United States as over-the-counter drugs. Worldwide consumption amounts to several hundred tonnes. The major producer of urea peroxide is Degussa (Germany).

8. Toxicology and Ecotoxicology

Metal peroxides, hydroperoxides, and ozonides cannot reasonably be tested for their toxicological and ecotoxicological properties due to their high reactivity to water with formation of oxygen and metal hydroxide or reaction with carbon dioxide to give metal carbonates. The toxicity is mediated by the high corrosivity to skin and mucous membranes of concentrated hydroxide solutions. Additionally, after direct contact with metal peroxides thermal burns can be developed due to the highly exothermic reaction with water.

8.1. Toxicology

Most inorganic peroxo compounds are expected to have a similar toxicity profile to hydrogen peroxide as they release hydrogen peroxide on contact with water, with the exception of the peroxodisulfates.

Available acute toxicity data of the peroxo compounds are summarized in Table 2. All substances tested are of low toxicity by the dermal route and of low to moderate toxicity after oral administration or inhalation exposure. The effects observed are characterized by irritation and inflammation of the skin and mucous membranes at the site of application.

The results of skin and eye irritation and skin sensitization studies are summarized in Table 3. Most of the compounds, except potassium peroxomonosulfate triple salt, which is corrosive, are nonirritating or only slightly irritating to skin, whereas eye irritation was evident for most of the substances. Some compounds cause irreversible corneal damage to the eye. Skin sensitization has only been reported for the peroxodisulfates. In humans, irritative and contact dermatitis but also asthmalike symptoms after exposure to higher airborne concentrations (> 1 mg/m^3) have been reported with peroxodisulfates [269, 270]. These effects are possibly

Table 2. Acute toxicity data for peroxo compounds

Substance	LD_{50} (rat, oral), mg/kg	LD_{50} (dermal), mg/kg	LC_{50} (rat, inhalation), mg/m^3
Ammonium peroxodisulfate	495 [a]–742[e] [241, 242]	$>$ 2000 (rat) [242][b]	$>$ 2950 (4 h) [242]
Sodium peroxodisulfate	895 [242, 245]	$>$ 10 000 (rabbit) [242][b]	$>$ 5100 [242][c, d]
Potassium peroxodisulfate	1130[f] [242]	10 000 (rabbit) [242]	42 900 (1 h) [242]
Potassium peroxomono sulfate triple salt	1204 [247]	$>$ 2000 (rat) [248] [b]	$>$ 5000 (4 h) [249] [cd]
Sodium percarbonate	1034 [250]	$>$ 2000 (rat) [251] [b]	$>$ 4580 (1 h) [252] [cd]
Sodium perborate monohydrate	1800	2000 (rabbit) [255][2]b	
Sodium perborate tetrahydrate	2243 [253]		1164[f] [254]

[a] Female.
[b] Limit test.
[c] No mortality occured.
[d] Maximum concentration technically obtainable.
[e] Male.
[f] For respirable particles [$<$ 10 µm mass medium aerodynamic diameter (MMAD)], lethality due to local irritant effects in the respiratory tract.

Table 3. Experimental findings on irritation and skin sensitization by peroxo compounds

Substance	Skin irritation (rabbit)	Eye irritation (rabbit)	Skin sensitization (guinea pig)
Ammonium peroxodisulfate	not irritating [256]	slightly irritating [257]	sensitizing (optimization test) [258]
Sodium peroxodisulfate	not irritating [259]	not irritating [242]	sensitizing (optimization test) [260], not sensitizing (Buehler Test) [242]
Potassium peroxodisulfate	not irritating [242]	not irritating [242]	
Potassium bis(peroxomonosulfate) bis sulfate (Caroat)	corrosive [261]	highly irritating/corrosive [262]	not sensitizing (maximization test) [263]
Sodium percarbonate	slightly irritating [264]	highly irritating/corrosive [265]	not sensitizing (Bühler Test) [266]
Sodium perborate monohydrate	slightly irritating [254]	highly irritating/corrosive [254]	not sensitizing (Bühler Test) [268]
Sodium perborate tetrahydrate	not irritating [267]	irritating [267]	

related to a direct histamine-liberating activity of peroxodisulfates. Sensory irritation studies in mice with sodium peroxodisulfate and sodium percarbonate resulted in RD_{50} concentrations (concentrations that resulted in a 50 % depression of the respiratory rate) of 2250 mg/m^3 [242] and 700 mg/m^3 [271], respectively.

After repeated exposure to inorganic peroxo compounds, toxic effects are normally limited to irritative and inflammatory changes at the site of contact. After repeated inhalation exposure of rats with ammonium peroxodisulfate aerosol (1–20 mg/m^3, 23.5 h/d, 7 d) effects indicative of irritation and inflammation of the respiratory tract including the lung were observed; the no observed adverse effect level (NOAEL) was 1 mg/m^3 [272]. A 90-d inhalation study in rats exposed 6 h per day for 5 d per week to concentrations of 5, 10, or 25 mg/m^3 of ammonium peroxodisulfate revealed an NOAEL of 10 mg/m^3, with the only effects being local irritation of trachea, bronchi/bronchioles, and lungs. In this study a NOEL of 5 mg/m^3 was identified based on sporadic rale and effects on the respiratory rate in the 10 mg/m^3 dose group. All effects were reversible during a 6-week recovery period. No systemic effects or histopathological changes in any organs distant from the site of first contact were observed [242]. Rats exposed to concentrations of 1.4–43 mg/m^3 of potassium bis(peroxomonosulfate) bis-sulfate dust by inhalation for 6 h/d, 5 d per week for 2 weeks showed severe ocular irritation as the main effect (NOAEL 1.4 mg/m^3); no pathological changes in other organs were observed [275].

No effects were observed in male rats receiving up to 1000 ppm sodium peroxodisulfate or potassium peroxodisulfate in the diet (corresponding to 137 mg/kg bw or 131.5 mg/kg bw, respectively) for 28 d [242]. In a 28-d feeding study in male rats with ammonium peroxodisulfate an NOAEL of 300 ppm in the diet, corresponding to 41 mg/kg bw was identified based on a decreased adrenal weight to body weight ratio at 82 mg/kg bw [242]. In a 13-week feeding study in rats with 3000 ppm sodium peroxodisulfate in the diet (corresponding to 300 mg/kg bw) local effects on the mucous membranes of the gastrointestinal tract were found, but no systemic toxicity was observed. The NOAEL in this study was 300 ppm or 30 mg/kg bw [269, 270]. In a 28-d study in rats receiving 1000 mg/kg bw of sodium perborate tetrahydrate by gavage, similarly the observed effects could be releated to irritation of the mucous membranes in the gastrointestinal tract [276].

The effects observed are consistent with the liberation of hydrogen peroxide by the peroxo compounds, which showed similar effects after repeated oral administration. In a 90-d drinking water study in mice receiving 100 to 3000 ppm of hydrogen peroxide, irritation of the mucous membranes of the gastrointestinal tract was observed (the NOAEL was 100 ppm, corresponding to 26 or 37 mg/kg/day for male and female mice, respectively). No systemic effects were observed [274]. Hydrogen peroxide is rapidly degraded in the exposed tissues and blood, catalyzed by metabolic enzymes such as catalases or peroxidases. Therefore systemic availability of peroxo compounds and hydrogen peroxide is thought to be low [273].

Ammonium and sodium peroxodisulfates were nonmutagenic in a variety of in vitro and in vivo test systems for point mutations and chromosomal aberrations with and without a metabolic activation system [242, 269, 277 – 280] and potassium bis(peroxomonosulfate) bis-sulfate was nonmutagenic in the Ames assay [281].

Results of mutagenicity studies with sodium perborate were comparable to those with hydrogen peroxide [273, 282 – 285].

While in vitro assays for gene mutation and chromosomal aberration were positive in the absence of a metabolic system (rat liver supernatant, catalase), the mutagenic activity was reduced or abolished when an exogenic metabolic system was added [273, 282, 283]. No in vivo data for sodium perborate are available, but hydrogen peroxide was nonmutagenic with regard to chromosomal aberrations or micronuclei in rats and mice as well as in a study for unscheduled DNA synthesis in rats [273, 284].

No carcinogenicity studies are available for peroxo compounds, but the carcinogenic effects of hydrogen peroxide were studied in rats and different strains of mice after oral administration. [272 – 288]. Available evidence suggests that hydrogen peroxide and peroxo compounds are not likely to cause cancer in humans except by uncommon or unlikely routes of exposure [289].

Table 4. Ecotoxicity data of peroxo compounds

Substance	Acute toxicity for fish LC_{50}, 96-h (mg/L)	Acute toxicity for daphnia EC_{50}, 48-h (mg/L)	Chronic toxicity for algae EC_{50}, 72-h (mg/L)
Ammonium peroxodisulfate	*Oncorhynchus mykiss* 76.3 [295]	*Daphnia magna* 120 [298]	*Scenedesmus quadricauda* 33 [a] (growth rate) [300] *Selenastrum capricornutum* (growth rate): 83.7 [242]
Sodium peroxodisulfate	*Oncorhynchus mykiss* 163 [242]	*Daphnia magna* 133 [299]	*Selenastrum capricornutum* (biomass): 116 [242]
Potassium peroxomonosulfate triple salt	*Brachydanio rerio* $> 32 < 56$ [301]	*Daphnia magna* 5.3 [b] [303]	
Sodium percarbonate	*Pimephales promelas* 71 [302]	*Daphnia pulex* 4.9 [304]	*Anabaena spec* (blue-green) 8 [c] [305] *Chlamydomonas spec* (green) 60 [d] [305]
Sodium perborate monohydrate	*Brachydanio rerio* 51 [306]	*Daphnia magna* 11 [307]	*Selenastrum capricornutum* 3.3 (biomass) [308] 6.3 (growth rate) [308]

[a] EC_{10} (96-h).
[b] 24-h.
[c] 140-h.
[d] 240-h.

No effects on male and female fertility were observed in a one-generation study in rats with ammonium peroxodisulfate up to the highest tested dose level of 250 mg (kg bw)$^{-1}$d^{-1} in the diet. No effects on embryo/fetal viability up to day 4 of lactation were observed in the same study. The NOAEL for both fertility and developmental toxicity was 250 mg (kg bw)$^{-1}$d^{-1} [242].

A potential for systemic toxicity of sodium perborates could be the formation of boric acid. However the uptake of boric acid in the body is limited by the simultaneous liberation of hydrogen peroxide that reduces local blood circulation. After repeated oral administration of high doses of boric acid to experimental animals, effects on male fertility (testicular atrophy) and developmental toxicity were observed [290]. For sodium perborate there is no indication of any effect on reproductive organs in a repeated dose study by the oral route (28 d, 7 d per week at 1000 mg (kg bw)$^{-1}$d^{-1}) in rats at dose levels corresponding to boron doses that already revealed considerable testicular toxicity in studies with boric acid. As these perborate doses corresponded to the maximum tolerated dose in the animals, it can be concluded that sodium perborate did not cause any testicular effect in rats at the maximum tolerated dose of 1000 mg (kg bw)$^{-1}$d^{-1} [254]. In a developmental toxicity study by the oral route in rats with sodium perborate tetrahydrate, fetotoxic effects were only observed at dose levels that were already maternally toxic [254].

No occupational exposure limits exist for perborates, sodium percarbonate, and peroxomonosulfate. Occupational exposure limits have been established for hydrogen peroxide in different countries, for example, by ACGIH [288], the Dutch Social Economic Council [290], the UK Health and Safety Executive (HSE) [291], Belgium [292], the French Ministry of Labour [293], and the German AGS (Auschuss für Gefahrstoffe) [294]: 8-h TWA: 1.4 mg/m^3 or 1.5 mg/m^3. A 15-min STEL of 2.8 mg/m^3 was established by HSE (max. 1 h in a 24 h period) [291], and a 15-min. STEL of 1.4 mg/m^3 by AGS [294].

For peroxodisulfates occupational exposure limits have been established by ACGIH (8-h TWA: 0.1 mg/m^3 [288]), the UK HSE, and Belgium (8-h TWA: 1 mg/m^3 [291, 292]).

8.2. Ecotoxicology

Ecotoxicological data are available for a number of peroxo compounds. The data are summarized in Table 4.

Peroxodisulfates are of low to moderate toxicity to aquatic organisms in standard tests [295 – 300].

Potassium peroxomonosulfate triple salt and sodium percarbonate exhibit moderate toxicity to fish, daphnia, and algae [301 – 305].

Sodium perborate monohydrate is of moderate toxicity to fish, daphnia, and algae [306 – 308]. Algae seem to be the most sensitive phylum.

Peroxo compounds tend to decompose in natural waters to give hydrogen peroxide and ultimately oxygen and water. In wastewater peroxide was shown to decompose before reaching the sewage treatment plant [309]. Therefore it is expected that peroxo compounds do not persist long enough in water to cause an impact on the environment in low concentrations. Given the polarity and high water solubility of peroxo compounds, a potential for bioaccumulation is not anticipated.

9. References

1. W. Machu: *Das Wasserstoffperoxyd und die Perverbindungen*, 2nd ed., Springer Verlag, Wien 1951.
2. I. I. Vol′nov: *Peroxides, Superoxides, and Ozonides of Alkali and Alkaline Earth Metals*, Plenum Press, New York 1966.
3. *Ullmann*, 4th ed., **17**, 691.
4. W. Weigert (ed.): *Wasserstoffperoxid und seine Derivate*, Hüthig Verlag, Heidelberg 1978.
5. *Winnacker-Küchler*, 4th ed., **2**, 563.
6. *Ullmann*, 3rd ed., **13**, 193.
7. H. Remy, *Angew. Chem.* **68** (1956) 613.
8. I. I. Vol′nov in [2] , p. 35.
9. L. Brewer, *Chem. Rev.* **52** (1953) 6.
10. S. Z. Makarov, I. I. Vol′nov, *Izu. Akad. Nank. SSR Otd., Khim Nank.* (1954) 765.
11. Roessler & Hasslacher Chemical, US 1 685 520, 1928; US 7 796 241, 1931 (H. R. Carveth).
12. H. R. Tennant, R. B. Schow, Handling and Uses of the Alkali Metals," *Adv. Chem. Ser.* 1957, no. 19, 118ff.
13. F. Feher et al., *Chem. Ber.* **86** (1953) 1429.
14. A. J. Coher, *J. Am. Chem. Soc.* **74** (1952) 3762.
15. R. O. Bach, W. W. Boardman, *Chem. Eng. News* **40** (1962) Nov. 19, 54.
16. I. I. Vol′nov in [2] , p. 58.
17. Interox Chemicals Ltd., GB 2 219 474, 1989 (D. B. Mobbs, S. C. Ball, A. J. Monhemius).
18. Colgate Palmolive, US 5 698 182, 1997 (M. Prencipe, J. Masters, V. O. Drago).
19. R. T. Tieckelmann, R. E. Steele, *Food Tech.* **45** (1991) no. 1, 106.
20. Campbell Baking Co., US 1 385 842, 1921 (C. J. Patterson).
21. Rumford Chemical Works, US 1 775 037, 1930 (A. H. Fiske).
22. L. J. Thenard, *Ann. Chim. Phys.* **8** (1818) 306.
23. I. I. Vol′nov in [2] , p. 66.
24. I. I. Vol′nov in [2] , p. 79.
25. ORC Technical Bulletin #5-2.2, http://www.regenesis.com/ORCtech/default.htm
26. I. I. Vol′nov in [2] , p. 96.
27. A. W. Petrocelli, *J. Chem. Educ.* **40** (1963) 146.
28. R. A. Sheldon, J. A. Van Doorn, *J. Organomet. Chem.* **94** (1975) 115.
29. *Kirk-Othmer*, 5th ed., **17, 22**
30. I. I. Vol′nov in [2] , p. 105.
31. L′Air Liquide, DE 2 313 116, 1973 (J. Malafosse).
32. Fire Research Institute, DE 2 208 347, 1972 (Takahashi et al.).
33. Handling and Uses of the Alkali Metals," *Adv. Chem. Ser.* 1957, no. 19.
34. S. Tanatar, *Z. Phys. Chem.* **26** (1898) 132–134. P. Melikoff, L. Pissadewsky, *Ber. Dtsch. Chem. Ges.* **31** (1898) 678–680.
35. A. Hansson, *Acta Chem. Scand.* **15** (1901) 934.
36. W. P. Griffith et al., *Chem. Ind. (London)* **1984**, no. 5, 185.
37. E. Koberstein et al., *Z. Anorg. Allg. Chem.* **374** (1970) 125–146.
38. J. O. Edwards et al., *J. Am. Chem. Soc.* **91** (1969) 1095.
39. D. W. van Gelder, *Recl. Trav. Chim. Pays-Bas* **75** (1956) 117–126.
40. Degussa, DE 1 078 101, 1955 (G. Bretschneider);
    Degussa, DE 1 079 603, 1955 (G. Bretschneider).
41. Degussa, EP-A 328 768, 1988 (B. Bertsch-Frank et al.).
42. Interox, DE 2 650 225, 1976 (J. Brichard, J.-C. Colery).
43. R. Bruce et al., *J. Am. Chem. Soc.* **87** (1965) 2057.
44. *Winnacker-Küchler*, 4th ed., **2**, 586.
45. Laporte, GB 761 371, 1953 (W. S. Wood, C. S. Goodfrey).
46. Solvay, DE-AS 1 138 381, 1960 (E. Ublon, F. L. Lambert).
47. W. Wöhlk, T. Adamski, *Chem.-Anlagen + Verfahren* **2** (1973) 45.
48. Degussa, DE 2 243 341, 1972 (G. Knippschild, H. Schuster).
49. Kali-Chemie, DE 204 373, 1972 (H. Honig et al.).
50. Solvay, DE 1 930 286, 1969 (J. L. Denaeyer).

51. Atochem, EP 155 894, 1985 (J. Dugua, J. P. Cuer).
52. Atochem, EP 194 952, 1986 (J. Dugua).
53. Peroxid-Chemie, DE 2 258 319, 1972 (H. Dillenburg et al.).
54. Interox, DE-OS 2 813 326, 1978 (L. Brichard, J.-C. Colery).
55. J. Lorösch, N. Steiner, H. J. Ullmann, *Expomin '92*, Santiago Chile, May 12–16, **1992.**
56. Degussa, DE 1 257 122, 1961 (H. Meier-Ewert et al.).
57. L′Air Liquide, FR 7 804 333, 1978 (M. Rignon).
58. Degussa, DE 940 348, 1955 (G. Bretschneider).
59. Recommendations on the transport of dangerous goods,UNNewYork and Geneva 1995, Manual of test and criteria, 2nd revised edition 1995, Test 0.1 34.4.1 "Test for oxidising solids".
60. Peroxid-Chemie, EP 150 432, 1984 (W. Dötsch, R. Siegel).
61. Degussa, DE 2 231 257, 1972 (G. Knippschild et al.).
62. G. Heller, A. Pawel, *Tenside Deterg.* **23** (1986) 73–75.
63. DuPont, US 2 802 722, 1954 (S. E. Stephanou).
64. DuPont, US 2 901 319, 1957 (A. A. D′Addieco).
65. DuPont, US 3 041 139, 1960 (A. A. D′Addieco).
66. Degussa, DE 40 20 856, 1992 (B. Bertsch-Frank et al.).
67. M. Spiro, *Electrochim. Acta* **24** (1979) 313.
68. W. V. Steele, E. H. Appelman, *J. Chem. Thermo.* **14** (1982) 337.
69. H.-P. Harke, *Hyg. + Med.* **12** (1987) 224–225.
70. *Kirk-Othmer*, 4th ed., **18**,
71. Degussa, US 5 141 731, 1992 (J. Meier et al.).
72. Solvay, EP 0 554 271, 1993 (J. Lane, C. Mc. Donogh, S. Woods).
73. FMC, EP 1 773 907, 1996 (J. Mangaro, M. Sacks, J. Carmichael).
74. FMC, WO 97/00225, 1997 (H. Castrantas, J. Mangaro, R. Midika, W. Johnson).
75. Interox Chemicals, WO 92/07791, 1992 (J. Lane, C. McDonoghs, S. Woods).
76. Interox Chemical Ltd., US 5 304 360, 1994 (J. R. G. Lane, C. F. McDonogh, S. E. Woods).
77. L′Air Liquide, GB 1 442 811, 1976 (E. Jourdan-Laforte).
78. FMC, US 5 397 482, 1995 (H. M. Castrantas et al.).
79. J. A. Cole, C. Stroiber: *SME Annual Meeting.* Phoenix Arizona, March 11–14, 1996.
80. E. N. Wilton, P. J. Wyborn, J. A. Reeve, *Proceedings of the 1st European Metal Conference*, Brussels 1991, Elsevier Applied Science, New York, 1951, p. 423.
81. H. M. Castrantas et al.,
82. R. Norcross: *Randol Gold Forum* at Squaw Creek Olympic Valley, CA, April 1996.
83. Degussa, US 5 091 054, 1992 (J. Meier, G. Arnold).
84. Degussa, US 5 246 543, 1993 (J. Meier, G. Arnold, O. Helmling).
85. N. Nimmerfroh, H. U. Süss: *International Non-Chlorine Bleaching Conference*, Orlando, FL, March 1996.
86. N. A. Troughton, F. Desperez, J. Devenyns, *Non-Chlorine Bleaching* Conference, Amelia Island, FL, March 1994.
87. DuPont, US 5 589 032, 1996 (H. Chang et al.).
88. DuPont, Technical Information: Pxa-high conversion peroxyacids, 7/94.
89. J. Flanagan, W. P. Griffith, A. C. Skapski, *J. Chem. Soc. Chem. Comms.* (1984) 1574.
90. M. G. C. Baldry, *J. Appl. Bact.* **58** (1985) 315–318.
91. DuPont, US 2 926 998, 1957 (A. A. D′Addieco, S. E. Stephanou).
92. Peroxid-Chemie, DE-OS 3 427 119, 1984 (W. Reh, S. Schelle).
93. Lever Brothers, US 5 501 812, 1996 (R. Vermeer et al.).
94. Procter & Gamble, EP 92 870 188, 1994 (G. O. Bianchetti et al.).
95. Schülke & Mayr, DE 3 046 769, 1993 (J. Sporkenbach et al.).
96. R. W. Murray, *Chem. Revs.* **89** (1989) 1187.
97. W. Adam, R. Curci, J. O. Edwards, *Accts Chem. Res.* **22** (1989) 205.
98. Colgate Palmolive, US 5 525 121, 1996 (R. Hefner, R. Steltenkamp).
99. *Chem.-Tech.* (Heidelberg) **17** (1988) 18.
100. Degussa, DE 2 352 856, 1973 (J. Fischer et al.).
101. Olin Corp., US 5 373 025, 1994 (W. A. Gay).
102. Henkel Corp., US 5 593 543, 1997 (B. Balos et al.).
103. L. J. Gramyi, F. Selymosi, *Journal Anal. Chem.* **142** (1954) 425–426.
104. D. L. Ball, J. O. Edwards, *J. Am. Chem. Soc.* **78** (1956) 1125–1129.
105. I. M. Kolthoff, I. K. Miller, *J. Am. Chem. Soc.* **73** (1951) 3055–3059.
106. J. Balej, M. Thumovä, *Collect Czech. Chem. Commun.* **46** (1981) 1229–1236.
107. J. Balej et al., *Chem. Prüm.* **3, 2** (1982) 225.

108. D. A. House, *Chem. Rev.* **62** (1962) 185–203.
109. W. M. Latimer, *The Oxidation States of the Elements and Their Potentials in Aqueous Solutions*, Prentice-Hall, Inc., New York 1938.
110. M. M. Barbooti, F. Jasmin, *Thermochim. Acta* **16** (1976) 402–406.
111. I. A. Vorsina, T. E. Grishakova, Y. Mikhailov, J. Mikhailov, *Inorg. Materials* **31** (1995) 1321–1324.
112. J. d'Ans, W. Friedrich, *Z. Anorg. Allg. Chem.* **73** (1912) 347.
113. W. Smit, J. G. Hoogland, *Electrochim. Acta* **16** (1971) 821–831.
114. W. Smit, J. G. Hoogland, *Electrochim. Acta* **16** (1971) 961–979.
115. J. Balej, M. Kaderavek, *Collect. Czech. Chem. Commun.* **44** (1979) 1510–1520.
116. J. Balej, M. Kaderavek, *Colleci. Czech. Chem. Commun.* **46** (1981) 2809–2817.
117. E. V. Kasatkin, G. F. Potapova, *Elektrokhimiya* **15** (1979) 1235–1239.
118. J. Balej, *Colleci. Czech. Chem. Commun.* **46** (1981) 462–466.
119. Degussa, US 2 795 541, 1957 (J. Müller); Degussa, DE 975 825, 1962 (J. Müller).
120. DD 99 548, 1972 (W. Thiele).
121. DD 152 266, 1981 (J. Balej et al.).
122. W. Thiele, H. Matschiner, *Chem. Tech. (Leipzig)* **29** (1977) 148–154.
123. L′Air Liquide, DE-OS 2 934 281, 1980 (M. Rignon, J. Malafosse).
124. J. Balej, H. Vogt, *Fortschr. Verfahrenstech. Abt.* **22** (1984) 361–389.
125. W. Thiele, M. Matschiner, *Dechema Monographie* **123** (1991) 133–165.
126. Eilenburger Elektrolyse- und Umwelttechnik, WO 93/20261, 1993 (W. Thiele et al.).
127. D. Hoormann, J. Joerissen, H. Puetter, *Chem. Ing. Tech.* **77** (2005) 1363–1376.
128. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **25** (1960) 1955–1957.
129. J. Balej, A. Regner, *Colleci. Czech. Chem. Commun.* **28** (1963) 1266–1272.
130. J. Balej, A. Regner, *Colleci. Czech. Chem. Commun.* **28** (1963) 3188–3193.
131. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **30** (1965) 1954–1967.
132. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **31** (1966) 938–942.
133. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **33** (1968) 2779–2791.
134. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **34** (1969) 1843–1853.
135. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **34** (1969) 2161–2168.
136. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **35** (1970) 1954–1963.
137. J. Balej, A. Regner, *Collect. Czech. Chem. Commun.* **47** (1982) 1539–1548.
138. J. Balej, M. Thumova, H. Spalkova, *Collect. Czech. Chem. Commun.* **59** (1994) 1606–1615.
139. J. Balej, M. Kaderavek, *Colleci. Czech. Clienz. Cotiimun.* **45** (1980) 2272–2282.
140. Mitsubishi Gas, JP 7 309 277, 1973 (T. Marushima et al.).
141. Food Machinery and Chemical, US 3 954 952, 1976 (M. J. McCarthy et al.).
142. Food Machinery and Chemical, DE 1 057 588, 1959 (G. G. Crewson, J. R. Ryan).
143. *Ullmann*, 4th ed., **17**, 723.
144. Peroxid-Chemie, DE 2 346 945, 1973 (E. Rossberger).
145. CS 145 656, 1972 (J. Balej).
146. DD 119 197, 1976 (W. Thiele, K. Wildner).
147. Eilenburger Elektrolyse- und Umwelttechnik, WO 97/07262, 1997 (W. Thiele).
148. Kemira, DE 4 426 246, 1994 (A. Kärki et al.).
149. Food Machinery and Chemical, DE-OS 2 757 861, 1978 (K. J. Radimer, M. J. McCarthy).
150. Degussa AG, unpublished results, 1989.
151. Nitto Electrical, JP-Kokai 62 100 511, 1987 (T. Watanabe et al.).
152. P. Bataille et al., *Coat. Technol.* **59** (1987) 71–75.
153. H. Warson, *ACS Symp. Ser.* **24** (1976) 228.
154. Buffalo Electro-Chem. Co., Uses of Persulfate, Bulletin no. 34, New York 1951.
155. C. F. Coombs, Jr. (ed.): *Printed Circuits Handbook*, Mc Graw-Hill Book Company, New York 1950, 6–17.
156. G. Hermann (ed.): *Leiterplatten*, Eugen G. Lenze Verlag, Saulgau 1978 115.
157. N. Steiner, B. Gec, *Annual Meeting of the American Association of Textile Chemists and Colorists* in Atlanta, GA, Oct. 1995.
158. Peroxid Chemie, DE 4 427 662, 1995 (C. E. Hartmann).
159. Solvay, WO 95 25195, 1995 (R. G. Jones, G. Frost, W. S. Stuart, M. A. Parry).
160. Novo Nordisk, WO 97 25469, 1997 (V. T. Den).
161. Henkel, DE 19 600 704, 1996 (H. Hoeffges, W. Neuhaus).
162. Wella, EU 778 020, 1996 (M. Schmitt et al.).
163. ZA 950 4754, 1995 (P. Cox).
164. Olin Corp, US 4 780 216, 1988 (J. A. Wojtowicz).
165. Degussa, US 5 116 967, 1992 (W. Fischer et al.).

166. FMC, US 5 718 837, 1998 (F. E. Caropreso, D. S. Thorp, R. H. Tieckelmann).
167. 2005.
168. ICI, DE 953 506, 1955/1956 (W. Maxwell).
169. Agfa, DE 1 051 117, 1957 (M. Heilmann, O. Wahl).
170. Fuji Photo Film, JP 0812 2991, 1996 (S. Yoshikava, T. Kajima, Y. Fujita).
171. P. P. Treadwell, W. T. Hall: *Analytical Chemistry*, 3rd ed., **2**, Wiley, New York 1945, 563.
172. M. Weidenauer et al., *Fresenius Z. Anal. Chem.* **331** (1988) 372–375.
173. Degussa, DE 1 096 339, 1959 (F. Beer, J. Müller).
174. Knapsack-Griesheim, DE 1 148 215, 1961 (G. Müller-Schiedmayer, J. Cremer).
175. DuPont, DE 1 037 434, 1958 (E. W. Heiderich).
176. DuPont, US 3 085 856, 1958 (D. B. Lake, G. Mamantov).
177. C. J. Battaglia, J. O. Edwards, *Inorg. Chem.* **4** (1965) 552.
178. G. Mamantov et al., *Inorg. Chem.* **3** (1964) 1043–1045.
179. Degussa, DE 1 143 491, 1960 (F. Beer, J. Müller).
180. E. Kasatkin et al., *Elektrokhimiya* **23** (1987) 679–681.
181. Ya. Kolotyrkin et al., *Elektrokhimiya* **17** (1981) 1339–1345.
182. Food Machinery and Chemical, US 4 626 326, 1986 (J. S. C. Chiang);
Food Machinery and Chemical, US 4 663 002, 1987 (J. S. C. Chiang, M. J. McCarthy).
183. SU 1 089 174, 1984;
E. Kasatkin et al., *Chem. Abstr.* **101** (1984) 80786.
184. L′Air Liquide, FR 2 261 255, 1974 (J. Pelloux-Prayer et al.).
185. A. Simon, H. Richter, *Z. Anorg. Allg. Chem.* **302** (1959) 165–174.
186. Food Machinery and Chemical, US 3 547 580, 1968 (B. Cohen).
187. J. J. Vol′nov et al., *Zh. Neorg. Khim.* **6** (1961) 268–270;
Chem. Abstr. **56** (1962) 4346b.
188. G. Rietz, E. Minke, *Z. Chem.* **9** (1969) 457–458.
189. B. Malinak et al., *Collect. Czech. Chem. Commun.* **39** (1974) 2549–2558.
190. P. Maruthamuthu, M. Santappa, *Indian J. Chem. Sect. A* **14 A** (1976) no. 1, 35–38;
Chem. Abstr. **84** (1976) 170247n.
191. Food Machinery and Chemical, DE 1 814 337, 1967 (B. Cohen, P. Mucenieks).
192. Mac Dermid, US 3 634 262, 1972 (J. J. Grunwald, E. E. Horner).
193. C. T. Carrondo et al., *J. Chem. Soc. Dalton Trans.* **1977**, 2323–2327.
194. Degussa, DE 2 644 147, 1976 (G. Knippschild et al.);
Degussa, DE 2 644 148, 1976 (G. Knippschild et al.).
195. E. Rommel, *Seifen Öle Fette Wachse* **103** (1977) 411–415.
196. Degussa, DE 2 651 442, 1976 (H. Klebe et al.).
197. M. Dreja, W. von Rybinski, *Tenside Surf. Det.* **39** (2002).
198. E. H. Riesenfeld, B. Reinhold, *Ber. Dtsch. Chem. Ges.* **42** (1909) 4377–4383.
199. Solvay Interox, EP0703190, 1995 (P. Romano, P. Soraya, B. Corrado).
200. Solvay & Cie, DE-OS 1 957 825, 1969 (J. Denaeyer, W. Kegelart).
201. Laporte, US 2 448 058, 1944 (V. W. Slater, W. S. Wood).
202. Degussa, DE-OS 2 800 760, 1978 (G. Knippschild et al.).
203. Peroxid-Chemie, DE 2 328 803, 1973 (H. Dillenburg et al.).
204. Allied Chemical, US 2 986 448, 1958 (W. H. Gates et al.).
205. DuPont, US 2 541 733, 1947 (J. H. Young).
KOA Soap, US 4 025 609, 1976 (K. Matsunaga et al.).
Mitsubishi Gas, US 4 409 197, 1983 (J. Sugano et al.).
206. Solvay Interox, WO 97 35 806, 1997 (A. P. James et al.).
207. Laporte, GB 722 351, 1952 (C. S. Godfrey et al.).
Degussa, US 3 555 696, 1969 (H. Pistor, H. Schuster).
208. Allied Chemical, US 3 801 706, 1971 (S. Milton).
209. Solvay, DE-OS 2 060 971, 1970 (J. L. Denaeyer, W. Kegelart).
210. Interox, DE-OS 2 733 935, 1977 (J. V. Brichard, J. C. Colery).
211. Degussa, DE 4 329 205, 1993 (M. Bewersdorf et al.).
212. DuPont, DE-OS 2 344 017, 1973 (M. Jayawant).
213. OCI, DE 19608000, 1997 (K. Sang-Ryul; K. Chong-Yun; H. Hwan-Kee; L. Jong-Pill).
214. Ugine, DE-OS 2 434 531, 1974 (P. Me′Iard, L. Mesaros).
215. Laporte, DE 2 133 566, 1971 (S. D. Williams et al.).

216. FMC, EP 429 321, 1991 (H. A. Pfeffer, C. Adams Jr.)
217. Degussa, DE 870 092, 1944 (H. Meier-Ewert); Degussa, DE-OS 2 748 783, 1977 (H. Klebe, G. Knippschild).
218. DuPont, DE-OS 2 448 453, 1974 (M. Jayawant, P. Yates).
219. Degussa, EP 1345845, 2001 (H. Jakob et al).
220. Degussa, EP1320511, 2001 (H. Jakob, B. Bertsch-Frank).
221. Interox, DE-OS 2 800 916, 1978 (J. Brichard, J. G. Colery).
222. Kao, DE 3 321 082, 1982 (K. Mutsumi et al.).
223. Degussa, DE 4 311 944, 1993 (B. Bertsch-Frank et al.), DE 4 306 399, 1993 (C. Klasen et al.).
224. Peroxid-Chemie, DE-OS 2 700 797, 1977 (H. Dillenburg et al.).
225. Degussa, DE 1 954 4293, 1995 (M. Bewersdorf et al.).
226. Interox, EP 279 282, 1987 (J.-M. Chauvier et al.).
227. Mitsubishi Gas Chemical, EP 623 553, 1993 (J. Kokubo et al.).
228. FMC, EP 363 852, 1990 (W. A. Hills et al.).
229. Henkel, DE 4 129 242, 1991 (H.-J. Beaujean et al.).
230. Nippon Peroxide, JP 63 215 502, 1988 (Y. Ito, E. Usu).
231. Henkel, DE 4 109 953, 1991 (R. Gonzales et al.).
232. Degussa, WO2004056954, 2002 (H. Jakob et al.).
233. B. Lorant, *Seifen Öle Fette Wachse* **92** (1966) 644–647.
234. S. Scholle, M. Kozak, *Chem. Prum.* **11** (1961) 417.
DD 212 493, 1982 (W. Thiele et al.).
235. Peroxid-Chemie, DE-OS 3 444 552, 1984 (E. Rossberger, H. Appl).
236. Degussa, DE 1 768 583, 1968 (F. Beer, G. Gilch).
237. Colgate Palmolive, US 5 766 574, 1997 (L. M. Christina-Beck et al.).
238. Chesebrough Pond′s, US 5 632 972, 1977 (D. R. Williams, C. W. Ryles, S. R. Barrow).
239. V. B. Haywood, Overview and Status of Mouthgard Bleaching," *J. Esthet. Dentistry* **1991**, no. 3, 157–161.
240. B. K. Yarborough, The safety and efficiency of tooth bleaching: A review of the literature 1988–1990. Compendium Contesi," *Educ. Dent.* **12** (1991) 191–196.
241. Unpublished report, FMC corporation, study No. I91-1201 (1991) as quoted in IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
242. SIDS Initial Assessment Report Persulfates, OECD, 2005, available at http://cs3-hq.oecd.org/scripts/hpv.
243. Unpublished report, FMC corporation, study No. I91-1200 (1991), IUCLID, International Uniform Information Chemical Data Base, European Commission, 1996.
244. Unpublished report, FMC corporation, study No. ICG/T-79-025 (1997), IUCLID, International Uniform Information Chemical Data Base, European Commission, 1996.
245. Unpublished report, Degussa AG, US-IT No. 79-0017-DKT, 1979.
246. Unpublished report, FMC corporation, study No. ICG/T-79-029-1201 (1979), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
247. Unpublished Report, Laporte Industries, Report No. 85/LAPO10/308, 1985.
248. Unpublished Report, Laporte Industries, Report No. 85/LAPO11/224, 1985.
249. Unpublished Report, DuPont, Haskell Laboratory Report, 18.12.1980, 1980.
250. Unpublished Report, Solvay Interox, Hazelton HLA 90903986, 1990.
251. Unpublished Report, Solvay Interox, Hazelton HLA 90903987, 1990.
252. Unpublished Information, ICI (1977), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
253. J. J. Dufour, C. Rogg, A. Cerioli
254. HERAreport on perboric acid, sodium salt, mono and tetrahydrate,CASNo. 11138-47-9 (10332-33-9, 10486-00-7), HERA (2002). Available at: http://www.heraproject.com
255. Unpublished Report, Solvay Interox, MB Research Laboratories Report No. MB87-8676B (1987) as quoted in A. Stringer (ed.), ECETOC, Technical Report No. 63, Reproductive and general toxicology of some inorganic borates and risk assessments for human beings, ECETOC, Brussels 1995.
256. Unpublished report, Degussa AG, US-IT No. 83–0023-DKT, 1983.
257. Unpublished report, Degussa AG, US-IT No. 83–0024-DKT, 1983.
258. Unpublished report, Degussa AG, US-IT No. 85–0014-DKT, 1985.
259. Unpublished report, Degussa AG, US-IT No. 79–0016-DKT, 1979.

260. Unpublished report, Degussa AG, US-IT No. 85–0048-DKT, 1985.
261. Unpublished report, Degussa AG, US-IT No. 83–0061-DKT, 1983.
262. Unpublished report, Laporte Industries Report No. 85/LAPO13/307, 1985.
263. Unpublished report, Degussa AG, US-IT No. 92–0084-DGT, 1992.
264. Unpublished report, Solvay Interox, Hazelton HLA 90903988, 1990.
265. J. Momma et al., *Shokuhin Eiseigaku Zasshi* **27** (1986) no. 5, 553–560.
266. Unpublished report, Solvay Interox, Hazelton HLA 90903990, 1990.
267. Unpublished report, ICI, Report No. CTL/T/2427 (1986) as quoted in A. Stringer (ed.),ECETOC, Technical Report No. 63, Reproductive and general toxicology of some inorganic borates and risk assessments for human beings, ECETOC, Brussels 1995.
268. Unpublished report, Solvay Interox,MBresearch laboratories, Report No. MB87–8676F(1987) as quoted in A. Stringer (ed.), ECETOC, Technical Report No. 63, Reproductive and general toxicology of some inorganic borates and risk assessments forhumanbeings, ECETOC, Brussels 1995.
269. Berufsgenossenschaft der chemischen Industrie, Toxikologische Bewertung Nr. 4 Ammoniumpersulfat (1994).
270. T. Cascieri, M. J. Fletcher, M. S. Weinberg, *The Toxicologist* **1** (1981) 149.
271. SIDS Initial Assessment Report Sodium Percarbonate, OECD, 2005, available at http://cs3-hq.oecd.org/scripts/hpv.
272. J. A. Last, P. K. Dasgupta, K. DeCesare, B. K. Tarkington, *Toxicol. Appl. Pharmacol.* **63** (1982) 257–263.
273. F. M. Carpanini (ed.): ECETOC Special Report No. 10 Hydrogen Peroxide OEL Criteria Document, ECETOC, Brussels 1996.
274. Unpublished report, CEFIC peroxygen sector group, hydrogen peroxide 13-week drinking water study with 6-week recovery period in C57BL/6NCrlBR mice, 1997.
275. Unpublished report, DuPont Haskell laboratory (1981) (refers to potassium bis (peroxomonosulfate) bis sulfate), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
276. Unpublished report, Degussa AG, US-IT No. 89–0013-DGT, 1989.
277. Unpublished report, FMC corporation, Report No. I90–1119 (1990), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
278. Unpublished report, FMC corporation, Report No. I90–1120 (1990), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
279. Unpublished report, FMC corporation, Report No. I90–1121 (1990), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
280. Unpublished report, FMC corporation, Report No. I90–1173 (1990), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
281. Unpublished report, DuPont (1976), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
282. J. P. Seiler, *Mutat. Res.* **224** (1989) 219–227.
283. H. S. Rosenkranz, *Mutat. Res.* **21** (1973) 171–174.
284. Unpublished report, CEFIC peroxygen sector group, Coming Hazelton Report No. 514/24 (1997).
285. T. Ishikawa, S. Takayama *Information Bulletin on the Survey of Chemicals being tested for carcinogenicity* **11** (1984) 86.
286. A. Ito, M. Naito, H. Wanatabe, *Ann. Rep. Hiroshima Univ. Res. Inst. Nuclear Medicine and Biology* **22** (1981) 147–158.
287. A. Ito, H. Wanatabe, M. Naito, Y. Naito, *Gann* **72** (1981) 174–175.
288. ACGIH, TLV0s and BEL0s, threshold limit values for chemical substances and Biological exposure indices, 2005.
289. European Union, Institute for Health and Consumer Protection, European Chemicals Bureau, Existing Substances 2nd priority list 38, 2003, Risk Assessment Report Hydrogen Peroxide, European Commission Joint Research Centre, EUR 20844 EN.
290. Sociaal Economische Raad (NL), MAC Waarden, Waterstofperoxide 90 %, 2005.
291. HSE, Health and Safety Executive, EH40/ 2005 Occupational Exposure Limits 2005.
292. Koninklijk besluit van 11 maart 2002 betreffende de bescherming van de gezondheid en de veiligheid van de werknemers tegen de risico's van chemische agentia op het werk (B.S. 14.3.2002, Ed. 2; erratum B.S. 26.6.2002, Ed. 2; erratum B.S. 4.12.2002) Bijlage I Grenswaarden voor beroepsmatige blootstelling.
293. Réferences franaises, INRS, ND2098, Valeurs limites d'exposition professionnelle aux agents chimiques en France, 2005.

294. TRGS 900, Technische Regeln für Gefahrstoffe, Grenzwerte in der Luft am Arbeitsplatz, Bekanntmachung des BMA § 52 Abs. 4 GefstoffV, 1997. Ausgabe Januar 2006: Ausschuss für Gefahrstoffe–AGS-Geschäftsführung–BAuA–www.baua.de.
295. Unpublished report, FMC corporation, Report No. I92–1246 (1992), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
296. Unpublished report, Degussa AG, US-IT No. 88–0016-DGO, 1988.
297. Unpublished report, FMC corporation, Report No. I92–1250 nd No. I92–1251 (1992), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
298. Unpublished report, FMC corporation, Report No. I92–1248 (1992), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
299. Unpublished report, FMC corporation, Report No. I92–1252 (1992), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
300. G. Bringman, R. Kuehn, Gesundheits-Ingenieur, vol. 4 (1959) 115–120, IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
301. Unpublished report, Degussa AG, US-IT No. 89–0014-DGO, 1989.
302. Unpublished report, Solvay Interox, Burlington Research Inc., November 4 (1989), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
303. Unpublished report, Degussa AG, US-IT No. 89–0015-DGO, 1989.
304. Unpublished report, Solvay Interox, Burlington Research Inc. August 25 (1989), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
305. C. A. Clarke, PHD Thesis, University of Bath (1991).
306. Unpublished report Solvay Duphar, Int. Doc. No. 56635/58/92 (1993), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
307. Unpublished report Solvay Duphar, Int. Doc. No. 56635/59/92 (1993), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
308. Unpublished report Solvay Duphar, Int. Doc. No. 56835/4/93 (1993), IUCLID, International Uniform Chemical Information Data Base, European Commission, 1996.
309. W. Guhl, A. Berends, *Surfactants* **2** (2001) 98–102.

Peroxy Compounds, Organic

Herbert Klenk, Degussa AG, Hanau, Federal Republic of Germany

Peter H. Götz, Degussa AG, Hanau, Federal Republic of Germany (Chaps. 1 – 3, 5, 7, 8)

Rainer Siegmeier, Degussa AG, Hanau, Federal Republic of Germany (Chaps. 4, 6, 9)

Wilfried Mayr, Degussa AG, Hanau, Federal Republic of Germany (Chap. 10)

1. Introduction 2
2. Alkyl Hydroperoxides 3
2.1. Properties 3
2.2. Production 5
2.3. Uses 6
3. Dialkyl Peroxides 6
3.1. Properties 6
3.2. Production 7
3.3. Uses 8
4. Peroxycarboxylic Acids 8
4.1. Properties 8
4.2. Production 9
4.2.1. Peroxycarboxylic Acids from Carboxylic Acids and Hydrogen Peroxide . . 10
4.2.2. Peroxycarboxylic Acids from Carboxylic Anhydrides and Hydrogen Peroxide 12
4.2.3. Peroxycarboxylic Acids from Carboxylic Acid Chlorides and Hydrogen Peroxide 12
4.2.4. Peroxycarboxylic Acids by Autoxidation of Aldehydes 13
4.3. Uses 13
5. Diacyl Peroxides 13
5.1. Properties 13
5.2. Production 14
5.3. Uses 15
6. Peroxycarboxylic Esters 15
6.1. Properties 16
6.2. Production 18
6.3. Uses 21
7. α-Oxyperoxides 21
7.1. Properties 21
7.2. Production 23
7.3. Uses 25
8. α-Aminoperoxides 25
8.1. Properties 25
8.2. Production 27
8.3. Uses 27
9. Safety Measures 28
9.1. Hazards 28
9.2. Transportation and Storage 28
9.3. Handling 29
10. Toxicology 30
11. References 32

1. Introduction

Peroxides are characterized by the presence of the peroxo group −O−O−. In organic peroxides this group is bound to at least one carbon atom, or is bound to a carbon atom via another atom [1–14].

The first synthesis of an organic peroxide was that of dibenzoyl peroxide in 1858 [15]. Since then, a large number of organic peroxides have been synthesized, mainly on a laboratory scale. Interest in organic peroxides arose mainly in connection with biological processes [16] and the aging of organic compounds such as rubber, fats [17], oils, paints, and solvents, in which peroxides are often formed as reactive intermediates. Since the mid-1970s a surprisingly large number, with regard to the relative instability of the peroxo group, of naturally occurring peroxo compounds have been isolated and characterized.

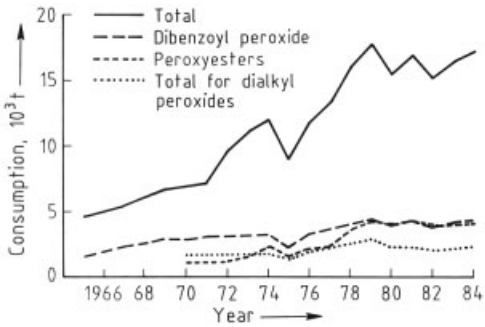

Figure 1. Consumption trend for organic peroxides in the United States

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_199

Industrial interest in organic peroxides began in the early 1900s with the introduction of benzoyl peroxide as a bleaching agent. The consumption of synthetic plastics has increased since the 1940s, resulting in the increasing importance of peroxides as initiators for the radical polymerization of vinyl monomers. Additional applications include their use as cross-linking agents for polyolefins, as hardeners for elastomers, as adhesion promoters for metal – plastic composites (e.g., polyethylene coatings), and in pharmaceuticals. Various organic peroxides have become industrially important as selective oxidation and epoxidation reagents (→ Epoxides). Peroxides are frequently generated in situ because of handling problems.

More than 65 organic peroxides in more than 100 formulations (liquid, solid, paste, powder, solutions, dispersions) are produced in industrial quantities today. An overview of the consumption of organic peroxides in the United States is given in Figures 1 [18] and Fig. 2.

Formally, organic peroxides can be regarded as derivatives of hydrogen peroxide in which the hydrogen atoms are substituted by various organic groups. The following classes can be distinguished:

Alkyl hydroperoxides	$R-O-O-H$
Dialkyl peroxides	$R^1-O-O-R^2$
Peroxycarboxylic acids	$R-\overset{O}{\overset{\|\|}{C}}-O-O-H$
Diacyl peroxides	$R^1-\overset{O}{\overset{\|\|}{C}}-O-O-\overset{O}{\overset{\|\|}{C}}-R^2$
	$R^1-\overset{O}{\overset{\|\|}{C}}-O-O-SO_2-R^2$
Peroxycarboxylic esters	$R^1-\overset{O}{\overset{\|\|}{C}}-O-O-R^2$
Peroxycarbonates	$R^1-O-\overset{O}{\overset{\|\|}{C}}-O-O-R^2$
α-Oxy-, α-amino peroxides	$R^1-\underset{X}{\overset{R^2}{C}}-O-O-R^3(H)$

$X = -OH,\ -OR^4,\ -O-O-,\ -NH_2,\ -NHR^5,\ -NR^5R^6$

Compounds in which the peroxy group is bonded to one or two atoms other than carbon are also regarded as organic peroxides, as long as additional organic substituents are present. Peroxides with the general structures

$$-\overset{|}{\underset{|}{C}}-O-O-M \quad \text{and} \quad -\overset{|}{\underset{|}{C}}-M-O-O-M-\overset{|}{\underset{|}{C}}-$$

where M is an element of groups 1, 2, 4, 5, 6, 8 or 12 – 16 of the periodic table have been synthesized in large numbers [19]. Furthermore, representatives of other substance classes, in which the peroxy group is connected to a heteroatom, are known. With few exceptions, compounds of this type are only of academic interest.

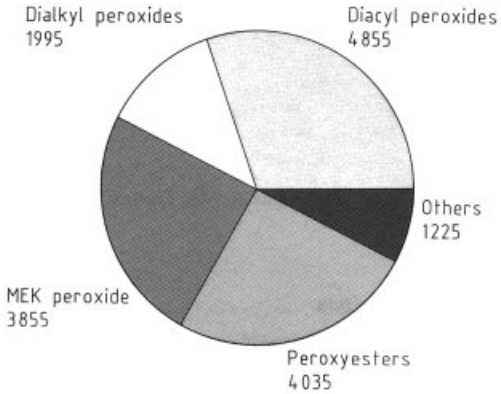

Figure 2. Consumption of organic peroxides (in tonnes) in the United States in 1983
MEK peroxide = methyl ethyl ketone peroxide

Homolytic cleavage of the O – O bond, readily induced by light or heat, is a typical reaction of organic peroxides. The reactivity depends strongly on the structure of the organic peroxy compound. In addition, heterolytic and hydrolytic cleavage, and rearrangements are also known. The oxidation potential decreases in the following order: peroxycarboxylic acids > hydroperoxides > diacyl peroxides > peroxycarboxylic esters > dialkyl peroxides.

Within a given class of compounds the stability increases with increasing carbon content. The thermal stability is often characterized by the half-life or the 10-h half-life temperature, whereby the radical initially formed upon decomposition is very short-lived, i.e., $t_{1/2} < 10^{-3}$ s [20]. Typical 10-h half-life temperatures of commercial peroxides lie between 25 and 170 °C. Decomposition is accelerated by various transition metals (promoters).

There are two main methods for the production of organic peroxides: (1) from hydrogen

peroxide and its derivatives and (2) by oxidation with oxygen.

Analytical methods for peroxides include the usual spectroscopic [21] and chromatographic procedures [22], as well as various methods based on oxidative titrations (iodometry) [23] and electrochemical techniques [24]. Particular attention must be paid to safety factors when dealing with peroxides. Many peroxy compounds are not only heat sensitive, but are also sensitive to shock, impact, and light. Peroxides tend to decompose explosively. These compounds are frequently handled in desensitized form.

2. Alkyl Hydroperoxides

The formal substitution of a hydrogen atom in hydrogen peroxide by an alkyl group gives a hydroperoxide with the general formula R–OOH. Primary, secondary, and tertiary hydroperoxides are known; vinyl hydroperoxides have not yet been synthesized.

Hydroperoxides occur widespread in nature. They occur mainly as intermediates during decomposition, and by aging of organic compounds under the influence of atmospheric oxygen (autoxidation):

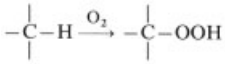

These compounds are oxidized further to frequently strong-smelling aldehydes, ketones, and carboxylic acids; an example is the smell of rancid fat [25]. Autoxidation is used industrially in the drying of paints (→ Drying Oils and Related Products).

2.1. Properties

Physical Properties. Hydroperoxides are generally sufficiently stable to be melted or distilled at a temperature below 80 °C. Decomposition increases markedly with increasing temperature; under certain circumstances explosions may occur. The thermal stability is mainly determined by the dissociation energy of the O–O bond; for homolytic cleavage a typical value is 180–184 kJ/mol [26], [27]. A summary of the physical properties of some important hydroperoxides is shown in Table 1.

The acidity of hydroperoxides is greater than that of the corresponding alcohols, so that in most cases salts can be prepared. This property is often exploited for isolation and purification [28], [29]. The acid strengths (pK_a at 20 °C) of some hydroperoxides are as follows [30]:

Methyl hydroperoxide	11.5
Ethyl hydroperoxide	11.8
Isopropyl hydroperoxide	12.1
Cumene hydroperoxide	12.6
tert-Butyl hydroperoxide	12.8
Triphenylmethyl hydroperoxide	13.07

When dissolved in noncoordinating solvents hydroperoxides form mainly dimers and trimers with six-membered ring structures.

Chemical Properties. Alkyl hydroperoxides can react with retention or cleavage of the O–O bond. In the first instance, substitution of the hydrogen atom on the peroxy oxygen atom leads to a whole series of products:

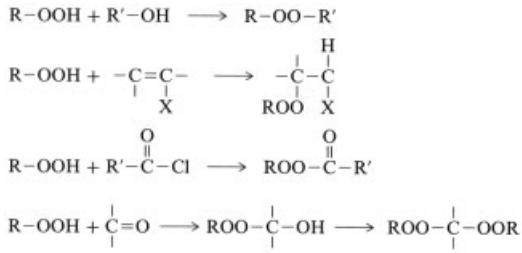

Reactions in which the O–O bond is cleaved are important for the quantitative analysis of hydroperoxides (e.g., by iodometry). Under basic conditions some hydroperoxides decompose to give carbonyl compounds, for example:

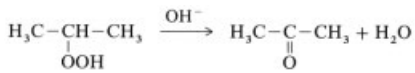

Tertiary hydroperoxides decompose under basic conditions to form the corresponding alcohol and oxygen:

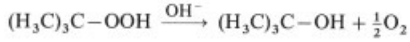

Under acidic conditions hydroperoxides decompose with rearrangement; a classic example

Table 1. Physical properties of some alkyl hydroperoxides-4

Name	CAS registry no.	Formula	M_r	*bp*, °C (kPa)	*mp*, °C	n_D, 20 °C	*d* (20 °C)	Active oxygen content, %
Methyl hydroperoxide	[*3031-73-0*]	H_3C-OOH	48.04	38 – 40 (8.6)		1.3641	0.9967 [a]	33.3
Ethyl hydroperoxide	[*3031-74-1*]	CH_3CH_2-OOH		43 – 44 (4)				
n-Propyl hydroperoxide	[*6068-96-8*]	$CH_3CH_2CH_2-OOH$	76.09	35 (2.7)		1.3890 [b]	0.9040 [b]	21.0
Isopropyl hydroperoxide	[*3031-75-2*]	$(H_3C)_2CH-OOH$	76.09	107 – 109 (101)		1.8861 [c]	0.8927 [c]	21.0
tert-Butyl hydroperoxide	[*75-91-2*]	$(H_3C)_3C-OOH$	90.12	35 (2.3)	4.0	1.4013	0.8960	17.7
Cyclohexyl hydroperoxide	[*766-07-4*]	$C_6H_{11}-OOH$	116.16	57 (0.16)	20	1.4622		
Cumene hydroperoxide	[*80-15-9*]	$Ph-C(CH_3)_2-OOH$	152.19	53 (0.013)		1.5242	1.0619	10.5
Octadecyl hydroperoxide	[*56537-19-0*]	$n\text{-}C_{18}H_{37}-OOH$	286.50		49 – 50			

[a] At 15 °C.
[b] At 25 °C.
[c] At 23 °C.

is the industrial synthesis of acetone and phenol from cumene hydroperoxide by the Hock process (→ Acetone , Chap. 4.1.). In the presence of a polymer with acidic groups, hydrogen peroxide and isobutene can be obtained from *tert*-butyl hydroperoxide [31]. The thermal decomposition of primary alkyl hydroperoxides takes place with homolytic cleavage, whereas for secondary and tertiary hydroperoxides homolytic cleavage can be observed for both thermal and photolytic decomposition [32]. The radicals formed in these processes are reactive species that can react, for example, with compounds containing C = C double bonds, either by addition or by substitution of a hydrogen atom. Peroxy radicals react with alkyl groups in the presence of metal ions (M) to give dialkyl peroxides and the corresponding alcohol in yields of up to 90 %:

$$2R-OOH + R'-H \xrightarrow{M^+} R-OOR' + R-OH + H_2O$$

In the absence of a suitable substrate, symmetrical dialkyl peroxides are formed with elimination of oxygen:

$$3R-OOH \xrightarrow{M^+} R-OH + R-OO-R + H_2O + O_2$$

2.2. Production

Several methods are known for the production of alkyl hydroperoxides. Essentially three major production routes exist, which can be differentiated according to the origin of the peroxy group:

1) From hydrogen peroxide
2) From oxygen (atmospheric oxygen, singlet or triplet O_2)
3) From another peroxy compound

Only a few processes can be carried out on a large scale and are of industrial importance.

Alkylation of Hydrogen Peroxide. Hydrogen peroxide can be alkylated under acidic or basic conditions by alkyl halides, alkyl sulfonates [33], carboxylic and sulfonic acid esters, ethers, tosylhydrazine, and cyclopropanes [34–36]. Dialkyl peroxides, which are often difficult to remove, are formed in addition to the monoalkylation product. A further limitation of this method is the safety measures required, as the lower molecular mass hydroperoxides in particular tend to undergo explosive decomposition. In another method, primary, secondary, and tertiary hydroperoxides are formed by reaction of an alkyl halide with hydrogen peroxide in the presence of silver trifluoroacetate [37]. The relatively stable tertiary hydroperoxides [38] (e.g., *tert*-butyl hydroperoxide) are of particular industrial importance.

$$(CH_3)_3C-OH \xrightarrow{H_2O_2/H_2SO_4} (CH_3)_3C-OOH$$

The remarkably stable acetylene mono- and dihydroperoxides can also be prepared in this way. These compounds have found some industrial interest as starting materials for the correspond-

ing dialkyl peroxides, which are used as high-temperature polymerization initiators [39]:

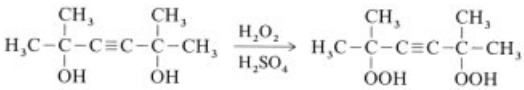

Primary and secondary hydroperoxides can only be obtained in moderate yields from hydrogen peroxide and an alkylating agent [40].

Halogen-substituted hydroperoxides can be obtained in good yields by the reaction of electron-rich olefins with hydrogen peroxide and a halogen [41]:

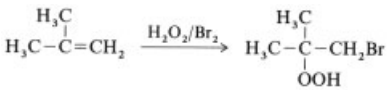

Halogen-substituted hydroperoxides can also be formed by peroxymercuration with subsequent substitution of the acetoxymercury(II) group with halogens [42].

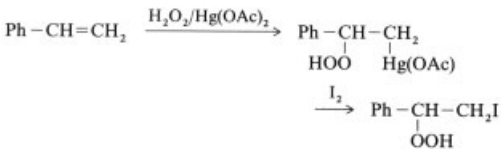

Oxidation with Oxygen. The insertion of oxygen into an alkane C – H bond leads initially to hydroperoxides. When this reaction is carried out with atmospheric oxygen it is known as autoxidation. It frequently leads to a multitude of products because of further spontaneous reactions. Therefore, this reaction is of limited synthetic use. Moreover, these reactions are generally very slow and require high temperature (usually > 100 °C).

The process involves a radical mechanism, which can be started by radical initiators (peroxides [43], azo compounds [44], halogens, or hydrogen halides [45]), photochemically, thermally [46], and by various transition metal ions.

Initiation:	$RH \longrightarrow R\cdot + H\cdot$
Chain propagation:	$R\cdot + O_2 \longrightarrow ROO\cdot$
	$ROO\cdot + RH \longrightarrow ROOH + R\cdot$
Chain termination:	$2\,R\cdot \longrightarrow R-R$
	$ROO\cdot + R\cdot \longrightarrow ROOR$
	$2\,ROO\cdot \longrightarrow O_2 + ROOR$

Tertiary carbon atoms are preferentially attacked because of the lower dissociation energy of the C – H bond compared with those of primary and secondary centers. Selective preparation of hydroperoxides can be achieved mainly by reaction of activated hydrogen atoms (benzylic hydrogen atoms, hydrogen atoms in the α-position to carbonyl groups). This method is of industrial importance for the production of cumene hydroperoxide, *tert*-butyl hydroperoxide, 1,4-diisopropylbenzene hydroperoxide, pinane hydroperoxide, and phenethyl hydroperoxide. In contrast to the oxidation of saturated hydrocarbons, autoxidation takes place even at room temperature.

tert-Butyl Hydroperoxide [*75-91-2*]. The oxidation of isobutane with oxygen in the gas phase at 160 °C and a residence time of ca. 3 min leads to a 70 % yield of *tert*-butyl hydroperoxide with a conversion of 80 %. Hydrogen bromide is used as the initiator [47]. Byproducts include di-*tert*-butyl peroxide, *tert*-butanol, and various alkyl bromides. Another process operates in the liquid phase at 130 °C and 3.5 MPa without a catalyst. The hydroperoxide is obtained in a yield of ca. 60 % with an isobutane conversion of < 25 % [48].

Cumene Hydroperoxide [*80-15-9*]. Cumene is oxidized with atmospheric oxygen at 120 °C. The selectivity is 95 % at a conversion of 15 % [49–53].

A method for the preparation of hydroperoxides that are not easily accessible by other routes is the reaction of organometallic compounds (Grignard reagents, zinc and cadmium compounds) with oxygen at 0 °C. This method leads to tertiary hydroperoxides in good yields, and primary and secondary hydroperoxides in moderate yields [54].

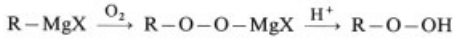

This method is particularly suitable for the production of low molecular mass hydroperoxides. Alternative synthetic routes via boranes [55] and ozonides are also known.

Peroxygenation of alkenes with triplet oxygen proceeds via the intermediate allyl radical; numerous allyl hydroperoxides can be formed due to resonance of the allyl radical:

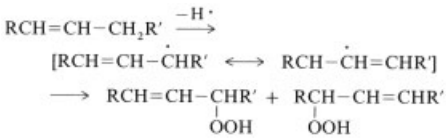

Various alkynes can be converted to propargyl hydroperoxides (2-alkynyl hydroperoxides) by treatment with triplet oxygen.

Alkenes that contain at least one allylic hydrogen atom undergo an allylic shift on reaction with singlet oxygen to give 2-alkenyl hydroperoxides [56–60]:

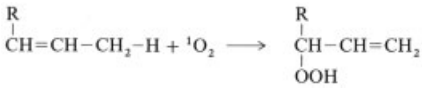

2.3. Uses

tert-Butyl hydroperoxide is used as an oxidant in the Halcon process for the production of propylene oxide. The standard commercial product (80 %, stabilized with water and phosphoric acid) is suitable for curing polyester resins and for the emulsion polymerization of styrene – butadiene rubbers.

In addition to the synthesis of phenol, cumene hydroperoxide is also used for the polymerization of dienes and for the cross-linking of sulfide rubbers. Furthermore, it is the starting material for dicumyl peroxide. The usual commercial product is an 80 % solution in cumene. Various heterocyclic hydroperoxides are employed as cross-linking agents for olefin polymers [61].

3. Dialkyl Peroxides

Dialkyl peroxides have the general formula R – O – O – R′, where R and R′ represent alkyl-, alkenyl-, or aryl groups. Both symmetrical (R = R′) and mixed (R ≠ R′) peroxides are known as well as numerous cyclic peroxides. Only the relatively stable tertiary dialkyl peroxides have so far gained any industrial importance.

3.1. Properties

Physical Properties. Several dialkyl peroxides can be distilled at atmospheric pressure or under vacuum without decomposition. Their dissociation energy is about 170 – 180 kJ/mol [2], [3]. Primary dialkyl peroxides such as dimethyl peroxide are shock sensitive and explosive. The physical properties of some dialkyl peroxides are listed in Table 2.

Chemical Properties. Dialkyl peroxides can be cleaved homolytically by the action of heat or light to give alkoxy radicals:

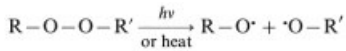

The subsequent reactions of these radicals depend on their structure, on the conditions, and on the temperature [62], [63]. In the absence of a substrate the main products result from disproportionation:

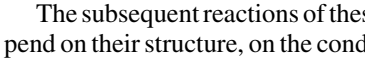

The reaction of the intermediate methoxy radical with formaldehyde then leads to methanol and carbon monoxide:

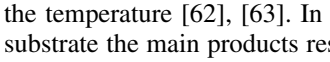

Secondary dialkyl peroxides lead analogously to the corresponding ketones and alcohols. Tertiary dialkyl peroxides undergo rearrangement to form ketones, alcohols, and epoxides, in addition to lower hydrocarbons. The main products of the decomposition of di-*tert*-butyl peroxide are acetone and ethane:

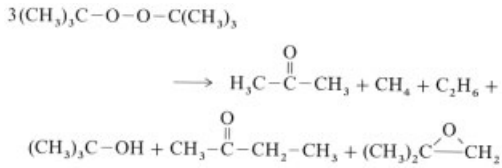

Numerous synthetic reactions are possible in the presence of reactive substrates [64], [65], of which the most important is the production of tertiary alkyl esters [66].

Table 2. Physical properties of some dialkyl peroxides-,4

Name	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (kPa)	n_D	d_4^{20}	Active oxygen content, %
Dimethyl peroxide	[*690-02-8*]	CH_3OOCH_3	62.07		13.5 (101)	1.3503 [a]	0.8680	25.8
Ethyl methyl peroxide	[*70299-48-8*]	$CH_3CH_2OOCH_3$	76.09		40 (98.5)	1.3590 [b]	0.8337 [b]	21.1
Diethyl peroxide	[*628-37-5*]	$CH_3CH_2OOCH_2CH_3$	90.12		64 (98.5)	1.3701 [b]	0.8266 [b]	17.8
tert-Butyl methyl peroxide	[*51392-67-7*]	$(CH_3)_3COOCH_3$	104.15	−102	23 (15.8)			15.4
Di-*tert*-butyl peroxide	[*110-05-4*]	$(CH_3)_3COOC(CH_3)_3$	146.23	−40	111 (101)	1.3860 [c]	0.791 [c]	10.9
Dicumyl peroxide	[*80-43-3*]	$PhC(CH_3)_2OOC(CH_3)_2Ph$	270.37	39		1.5630 [d]		5.9

[a] At 0 °C. [b] At 17 °C. [c] At 25 °C. [d] Supercooled melt at 21 °C.

3.2. Production

Dialkyl peroxides can be obtained from numerous starting materials by reaction with oxygen, hydrogen peroxide, alkyl hydroperoxides, and in some cases potassium peroxide.

Oxidation with Oxygen. For the production of acyclic dialkyl peroxides, the autoxidation of hydrocarbons with molecular oxygen is only suitable in certain cases. This method can only be used when the intermediate radical formed is stabilized, e.g., by alkyl or aryl substituents, or by allylic conjugation. In these reactions, either a C–H bond is cleaved initially [67], (e.g., in the case of 2-methylpropane) or formal insertion of oxygen can take place following cleavage of a C–C or a C–X bond (X = halogen):

$$Ph_3C-CPh_3 \longrightarrow 2\,Ph_3C^{\bullet} \xrightarrow{2O_2} 2\,Ph_3C-O-O^{\bullet}$$
$$Ph_3C-O-O^{\bullet} + Ph_3C^{\bullet} \longrightarrow Ph_3C-O-O-CPh_3$$

$$Ph_3C-O-O^{\bullet} + Ph_3C^{\bullet} \longrightarrow Ph_3C-O-O-CPh_3$$

With various alkenes, such as styrene, acrylates, acrylonitrile, and vinyl chloride, autoxidation leads to polymeric dialkyl peroxides with the general structure:

$$n\,HC(X)=CH_2 \xrightarrow{nO_2} \left[CH_2-CH(X)-O-O\right]_n$$

X = C_6H_5, COOR, CN, halogen

The oxidation of conjugated dienes (photooxygenation [68]) with oxygen leads either to polymers or to cyclo adducts (hetero-Diels–Alder adducts), i.e., the 1,4-epiperoxides, which are rarely accessible by other routes. The autoxidation of α-terpinene to give ascaridole is an example of the synthesis of an epiperoxide [69]:

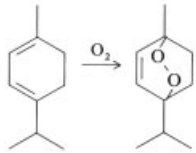

The photosensitized autoxidation of polynuclear aromatic hydrocarbons has been known for a long time [70].

The [2 + 2] cycloaddition of singlet oxygen to alkenes offers a simple route to 1,2-dioxetanes [71–76], which are of particular interest in connection with chemi- and bioluminescence.

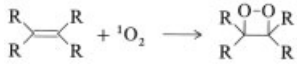

Alkylation of Hydrogen Peroxide. The dialkylation of hydrogen peroxide using alkyl halides, alcohols, alkenes, or alkyl sulfates gives symmetrical dialkyl peroxides and takes place under conditions similar to those described for monoalkylation in Section 2.2.

Dicumyl peroxide [*80-43-3*]. The reaction of an H_2O_2 –urea adduct with 2-phenyl-2-propanol at 35 °C in the presence of a mineral acid catalyst gives dicumyl peroxide in good yields [77].

Di-tert-amyl peroxide [*10508-09-5*] is formed in 68 % yield by slow addition of *tert*-amyl alcohol to a mixture of H_2O_2 and H_2SO_4 at 20 °C [78].

Alkylation of Hydroperoxides. The best method for the production of both symmetrical and unsymmetrical dialkyl peroxides is the alkylation of hydroperoxides with alcohols, olefins, esters, ethers, halides, or epoxides [79].

Dicumyl Peroxide. The reaction of cumene hydroperoxide with 2-phenyl-2-propanol in the presence of a mineral acid catalyst gives dicumyl peroxide in almost quantitative yield [49–53], [80–82].

Di-tert-butyl peroxide [*110-05-4*] is synthesized from *tert*-butyl hydroperoxide and *tert*-butanol [49–53], [80–83]. This reaction can also be carried out with an acidic ion-exchange resin as catalyst, the water produced in the reaction being removed by azeotropic distillation [84].

Bis(tert-amylperoxy)-2,5-dimethylhexane [*78-63-7*] is formed in 79 % yield from 2,5-dihydroperoxy-2,5-dimethylhexane and *tert*-amyl alcohol at 50 °C in the presence of sulfuric acid [85].

Phenols can be converted to dialkyl peroxides (quinone peroxides) by dehydrogenation [86–88]. These compounds have been investigated primarily in connection with the mechanism of action of phenolic antioxidants [89–91].

3.3. Uses

Due to their relatively high thermal stability, dialkyl peroxides are used mainly as catalysts for suspension and bulk polymerization [92], [93]. They also find use as hardeners for unsaturated polyester resins and for the cross-linking of polymers [94]. In addition to di-*tert*-butyl peroxide [95] and dicumyl peroxide, bis(*tert*-alkylperoxy) alkanes are employed. The dialkyl peroxides are commercially available as technical-grade material and as 40 – 50 % solutions.

4. Peroxycarboxylic Acids

Peroxycarboxylic acids have a wide range of applications in preparative and industrial chemistry on account of their high oxidation potential. Due to their instability some peroxycarboxylic acids are either generated in situ or used in desensitized form.

4.1. Properties

Physical Properties. Short-chain aliphatic peroxycarboxylic acids are infinitely miscible with water and possess an unpleasant, pungent odor. They are explosive when pure or highly concentrated. The tendency to explode decreases with increasing chain length. Long-chain peroxycarboxylic acids ($> C_6$) are increasingly water-insoluble solids. Aromatic peroxycarboxylic acids are only slightly soluble in water; both types are, however, soluble in various organic solvents.

In the crystalline state peroxycarboxylic acids associate to form dimers (**2**), whereas in the gas phase, in the molten state, and in solution in noncoordinating solvents monomers (**1**) with intramolecular hydrogen bonds are present [96].

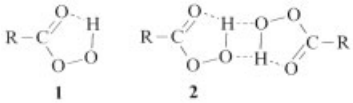

In the presence of donor compounds (Do) such as tetrahydrofuran, 1,4-dioxane, or *N*,*N*-dimethylacetamide, peroxycarboxylic acids form complexes with cleavage of the intramolecular hydrogen bonds [97], [98]:

Peroxycarboxylic acids exhibit pK_a values that are 3 – 4 units lower than those of the corresponding carboxylic acids [99] due to the absence of resonance stabilization in the peroxycarboxylate anions. As a result of the weak substituent effect, the pK_a values all lie in a small range. Table 3 lists physical properties for some peroxycarboxylic acids.

Under certain circumstances peroxycarboxylic acids can decompose explosively on heating or under impact. The dissociation energy of the O – O bond (80 – 90 kJ/mol) is considerably lower than that of the hydroperoxides or the dialkyl peroxides [100–102].

Chemical Properties. Peroxycarboxylic acids have the highest oxidation potential of all organic peroxides, whereby the corresponding carboxylic acid is formed with release of active oxygen. The stability of aqueous solutions of peroxycarboxylic acids depends on the pH. Bases cause decomposition with loss of active

Table 3. Physical properties of some peroxycarboxylic acids-,4

Name	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (kPa)	pK_a	Active oxygen content, %	Other
Peroxyformic acid (performic acid)	[*107-32-4*]	HCO_3H	62.0	−18 [a]	50 (13.3) [a]	7.1	25.8	
Peroxyacetic acid (peracetic acid)	[*79-21-0*]	H_3C-CO_3H	76.1	0	25 (1.6)	8.2	21.1	n_D^{20} 1.3974 d_4^{20} 1.0375
Peroxypropanoic acid	[*4212-43-5*]	$H_3C-CH_2-CO_3H$	90.1	−13	25 (2.7)	8.1	17.8	n_D^{20} 1.4041
Peroxybutanoic acid	[*13122-71-9*]	$H_3C-(CH_2)_2-CO_3H$	104.1	−10	26–29 (1.6)	8.2	15.4	n_D^{20} 1.4125 d_4^{21} 1.0745
Peroxynonanoic acid	[*3058-35-3*]	$H_3C-(CH_2)_7-CO_3H$	174.2	35		ca. 8.5	9.2	
Peroxydodecanoic acid	[*2388-12-7*]	$H_3C-(CH_2)_{10}-CO_3H$	216.3	50		10.8	7.4	
Diperoxyglutaric acid	[*28317-46-6*]	$HO_3C-(CH_2)_3-CO_3H$	164.1	80–100		ca. 7–8.5	19.5	
Diperoxyadipic acid	[*5824-51-1*]	$HO_3C-(CH_2)_4-CO_3H$	178.1	116–117		7.6	18.0	
Diperoxyoctanedioic acid	[*28317-47-7*]	$HO_3C-(CH_2)_6-CO_3H$	206.2	89		ca. 7.5–8	15.5	
Diperoxynonanedioic acid	[*1941-79-3*]	$HO_3C-(CH_2)_7-CO_3H$	220.2	90		ca. 8.0	14.5	
Diperoxydecanedioic acid	[*5796-85-0*]	$HO_3C-(CH_2)_8-CO_3H$	234.2	98		ca. 8–8.5	13.7	
Diperoxydodecanedioic acid	[*66280-55-5*]	$HO_3C-(CH_2)_{10}-CO_3H$	262.3	>100 (decomp.) [b]		9.7 [b, c]	12.2	
Peroxybenzoic acid (perbenzoic acid)	[*93-59-4*]	$C_6H_5-CO_3H$	138.1	41–42		7.8	11.6	
3-Chloroperoxybenzoic acid	[*937-14-4*]	*m*-$Cl-C_6H_4-CO_3H$	172.6	88		7.7	9.3	
4-Nitroperoxybenzoic acid	[*943-39-5*]	*p*-$NO_2-C_6H_4-CO_3H$	183.1	138 (decomp.)		7.4	8.7	
Monoperoxyphthalic acid	[*2311-91-3*]	*o*-$HO_2C-C_6H_4-CO_3H$	182.1	ca. 110 (decomp.) [a]		ca. 6–8	8.8	

[a] 90 % pure.
[b] 95 % pure.
[c] In H_2O/CH_3CN.

Table 4. Equilibrium composition and H_2O_2 conversion in the system acetic acid (AA) – hydrogen peroxide – peracetic acid (PAA) – water at 25 °C-,4

$[H_2O_2]$ (aqueous solution)	Molar ratio AA: H_2O_2	Composition, wt % (mol %)				H_2O_2 conversion, %
		AA	PAA	H_2O	H_2O_2	
50	1:1	22.0 (9.7)	29.1 (16.2)	31.5 (58.4)	16.0 (15.7)	37
	2:1	22.8 (11.6)	45.2 (29.2)	23.6 (50.9)	7.2 (8.2)	70
	4:1	18.6 (11.1)	62.4 (47.2)	15.3 (38.5)	2.4 (3.2)	76
	6.66:1	13.8 (9.0)	73.3 (60.8)	10.4 (28.8)	9.2 (1.3)	86
	10:1	10.1 (7.0)	81.0 (70.8)	7.4 (21.6)	0.5 (0.7)	90
90	1.5:1	45.0 (27.8)	35.0 (27.4)	14.0 (36.5)	6.0 (8.3)	77

oxygen, whereas in acidic media an equilibrium is established.

$$R-CO_3H + H_2O \overset{H^+}{\rightleftharpoons} R-CO_2H + H_2O_2$$

With the exception of performic acid, neutral solutions are largely hydrolytically stable. Peroxycarboxylic acids decompose thermally by two mechanisms [103]: radical and concerted:

$$R-CO_3H \rightarrow R-OH + CO_2 \quad (1)$$
$$R-CO_3H \rightarrow R-CO_2H + 1/2\,O_2 \quad (2)$$

In the absence of oxygen, the radical decomposition (Eq. 1), in which the first step is homolytic cleavage of the O – O bond, dominates. Heavy metal ions accelerate the radical decomposition even at low temperatures. In contrast, in the presence of radical inhibitors the concerted

reaction (Eq. 2) takes place. It is facilitated by the stereochemistry of the peroxycarboxyl group, which is able to form a five-membered cyclic transition state [103], [104].

In the presence of urea, starch, cholic acid, or zeolites, straight-chain peroxycarboxylic acids form inclusion compounds [105], [106]. Isolable 1 : 1 complexes are obtained with triphenylphosphine oxide, triphenylarsine oxide, and heteroaromatic *N*-oxides [107].

Table 5. Azeotropic mixtures of peroxycarboxylic acids and water-,4

Compound	Peroxycarboxylic acid, wt %	Water, wt %
Peracetic acid	61.0	39.0
Peroxypropanoic acid	51.5	48.5
Peroxybutanoic acid	42.0	58.0
Peroxyisobutanoic acid	55.5	44.5

Qualitative detection of peroxycarboxylic acids is facilitated with reagents that give colored products on oxidation, e.g., alkali-metal iodides, titanium sulfate, or phenolphthalein. For quantitative analysis potentiometry [3], polarography [5], and iodometric titration are suitable. IR and NMR spectroscopy are also useful for determination of peroxycarboxylic acids [108].

4.2. Production

Peroxycarboxylic acids can be produced by numerous methods, some of which are shown in Figure 3. Of these methods, however, only the first five are synthetically and industrially important. Production from diacyl peroxides is only of interest for perbenzoic acid, as the starting material, dibenzoyl peroxide, is commercially available.

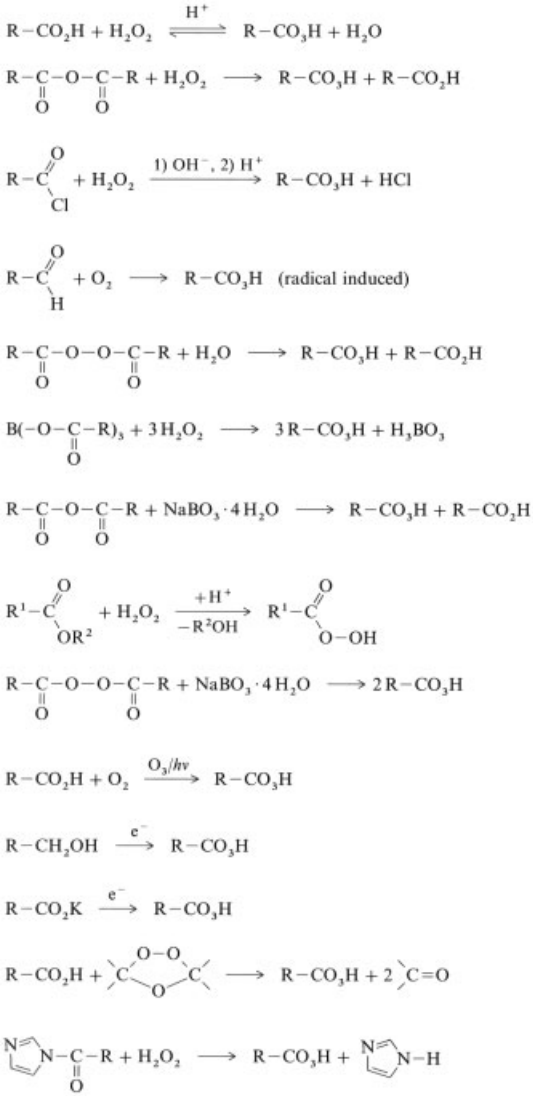

Figure 3. Reactions used for the synthesis of peroxyacids

4.2.1. Peroxycarboxylic Acids from Carboxylic Acids and Hydrogen Peroxide

The classical synthesis of peroxycarboxylic acids is the direct, usually acid-catalyzed, equilibrium reaction between carboxylic acids and hydrogen peroxide. With lower aliphatic carboxylic acids, aqueous solutions of peroxycarboxylic acids are obtained, which also contain carboxylic acid and hydrogen peroxide.

The equilibrium mixture is produced by simply mixing the components; the position of equilibrium depends on the molar ratio of the starting materials and on the concentration of the hydrogen peroxide. The equilibrium composition of the system acetic acid – hydrogen peroxide – peracetic acid – water is shown in Table 4 [109].

Equilibrium peracetic acid of this type, with a variable peracetic acid content of up to 40 %, is commercially available.

The reaction can also be carried out in the presence of organic solvents [110], [111]. Catalyst-free equilibrium mixtures can be obtained by the use of acidic, preferably perfluorinated, ion-exchange resins [112]. Strong aliphatic carboxylic acids such as formic acid or trifluoroacetic acid do not require a catalyst, particularly when used in a large excess relative to hydrogen peroxide.

The reaction between a carboxylic acid and hydrogen peroxide can also be carried out in the gas phase over a boric acid – phosphoric acid catalyst [113].

Pure aqueous solutions of peroxycarboxylic acids are obtained from the reaction mixtures by distilling off the peroxycarboxylic acid and water under vacuum [114]. The compositions of azeotropes of some lower peroxycarboxylic acids with water are listed in Table 5.

The production of anhydrous peroxycarboxylic acids can be achieved by three main routes:

1) By extraction of the equilibrium mixture or the pure aqueous solution of peroxycarboxylic acid with organic solvents such as phosphate esters [115], carboxylic esters [116], or chlorinated or aromatic hydrocarbons [117–122]. Residual water in the extracts can be removed by azeotropic distillation.
2) Removal of the water by azeotropic distillation with an organic solvent under vacuum [123–125]. For example, peroxypropionic acid [*4212-43-5*] is obtained in 89 % yield by heating a mixture of hydrogen peroxide, propionic acid, boric acid, phosphoric acid, and dichloroethane at 70 °C for 2 h and then distilling off the dichloroethane – water azeotrope under vacuum [126].
3) By selective absorption of the peroxycarboxylic acid and the corresponding carboxylic acid from the vapor in the distillative production of aqueous solutions of peroxycarboxylic acids in a phosphate ester, followed by desorption with a carboxylic ester [127].

It is not necessary for longer chain, water-insoluble mono- and dicarboxylic acids to be in solution to undergo reaction. In the presence of sulfuric acid the corresponding peroxycarboxylic acids can be obtained in good yields from aqueous suspensions [128–131]. The additional use of a solvent, from which the peroxycarboxylic acid formed will be isolated, is also possible [132]. The isolated compounds are frequently difficult to handle; in these cases they may be stabilized by addition of inorganic salts during the reaction (in situ desensitizing) [133], [134].

Diperoxydodecanedioic Acid [*66280-55-5*], **Desensitized.** Dodecanedioic acid is added to a mixture of 50 % hydrogen peroxide, 96 % sulfuric acid, and trioctylphosphine oxide. After 2 – 3 h at 60 °C the mixture is allowed to cool, diluted with sodium sulfate solution, the pH adjusted to 3.5 with sodium hydroxide, and, depending on the desired peracid concentration, conditioned with sodium sulfate. Diperoxydodecanedioic acid is obtained in 94 % yield.

Aromatic carboxylic acids can be converted to the corresponding peroxycarboxylic acids with methanesulfonic acid as catalyst [135–137]. The use of concentrated sulfuric acid as catalyst has also been described.

Diperoxyisophthalic acid [*1786-87-4*] [138] is obtained in 77 % yield as an 88 % solution from a mixture of 85 % hydrogen peroxide, 96 % sulfuric acid, and isophthalic acid.

4.2.2. Peroxycarboxylic Acids from Carboxylic Anhydrides and Hydrogen Peroxide

The reaction of lower aliphatic carboxylic anhydrides with hydrogen peroxide gives anhydrous solutions of peroxycarboxylic acids in the corresponding carboxylic acid, which is formed by reaction of the water introduced with the hydrogen peroxide with excess anhydride. The reaction is particularly suitable for the synthesis of monoperoxydicarboxylic acids, such as mono-peroxyphthalic acid or monoperoxysuccinic acid, from cyclic anhydrides [139], [140]. Inorganic bases, ammonia, or metal salts can be used as catalysts [141–143].

Monoperoxyphthalic Acid [*2311-91-3*]. The reaction of phthalic anhydride with hydrogen peroxide in 1,2-dichloroethane at 30 °C in the presence of a catalytic amount of ammonia gives, after acidification of the reaction mixture, 90 % monoperoxyphthalic acid in >96 % yield.

Furthermore, the diperoxycarboxylic acids can be obtained under acidic conditions by using a large excess of hydrogen peroxide [144–146].

Decylbutanediperoxycarboxylic Acid [*87458-73-9*]. The reaction of 1 mol of decylsuccinic anhydride in dichloromethane with 3.7 mol of 50 % hydrogen peroxide in the presence of 6.8 mol of 95 % sulfuric acid for 30 min gives decylbutanediperoxycarboxylic acid in 97 % yield with a purity of 98 %.

Explosive diacyl peroxides can be formed as byproducts of the acylation of hydrogen peroxide with lower aliphatic carboxylic acid anhydrides, the amount formed depending on the reaction conditions.

4.2.3. Peroxycarboxylic Acids from Carboxylic Acid Chlorides and Hydrogen Peroxide

The base-catalyzed reaction of carboxylic acid chlorides with hydrogen peroxide is the most important method for the production of aromatic peroxycarboxylic acids, which are difficult to synthesize by other routes [147–150]. Water [151], or water mixed with organic solvents [152], [153], can be used as solvent.

3-Chloroperbenzoic Acid [*937-14-4*] [154]. The reaction of 3-chlorobenzoyl chloride with hydrogen peroxide in alkaline, aqueous dioxane in the presence of $MgSO_4 \cdot 7\,H_2O$ gives, after acidification, 80 – 85 % 3-chloroperbenzoic acid.

The reaction of carboxylic acid chlorides with 85 % hydrogen peroxide in organic solvents such as dichloromethane, diethyl ether, or tetrahydrofuran in the presence of pyridine has also been described. However, the pyridine should not be added in equimolar amounts because of the formation of potentially explosive diacyl peroxides [155], [156].

Peroxydodecanoic Acid [*2388-12-7*]. The reaction of dodecanoyl chloride in tetrahydrofuran with 85 % hydrogen peroxide in the presence of pyridine gives 70 % peroxydodecanoic acid after 20 min reaction in an ice bath.

This method is also suitable for the production of sterically crowded peroxycarboxylic acids [157], [158].

4.2.4. Peroxycarboxylic Acids by Autoxidation of Aldehydes

Aldehydes react with oxygen in a radical chain process to give carboxylic acids. The intermediate peroxycarboxylic acids can be obtained as the main product by appropriate control of the reaction conditions. The reaction can be initiated photochemically or by the use of ozone and it can be catalyzed by metal salts, predominantly cobalt and iron compounds, or by 2,2′-bipyridyl [159–163]. The reaction can be carried out in the vapor phase at standard pressure [164], [165] or in the liquid phase [166], [167].

C_7-Peroxycarboxylic Acid Mixtures. The oxidation of isomeric C_7 aldehydes with oxygen in the presence of 2,2′-bipyridyl at 60 °C gives 62 % of the corresponding peroxycarboxylic acid mixture [168].

The method is suitable for the production of aromatic peroxycarboxylic acids when the autoxidation is carried out at room temperature in carboxylic esters as solvent. Substituted benzaldehydes, however, show considerable differences in reactivity [169].

4.3. Uses

The largest use of peroxycarboxylic acids is in the production of epoxides (→ Epoxides, Chap. 3.). Compounds with an isolated double bond that cannot be oxidized with hydroperoxides or hydrogen peroxide in the absence of a catalyst are epoxidized smoothly and stereospecifically by peroxycarboxylic acids. Cyclic ketones, such as cyclohexanone, are converted to lactones; the production of ε-caprolactone, an important intermediate for the production of polyesters and polyurethanes, is carried out on a large scale. Phenols can be hydroxylated with peroxycarboxylic acids to give dihydroxybenzenes [170–172].

Amine *N*-oxides and -sulfoxides, obtained by reaction of peroxycarboxylic acids with amines and sulfides are used as detergents and in the pharmaceutical industry. The use of peroxycarboxylic acids for disinfection [173–177], in environmental protection, and for bleaching is increasing due to their high oxidizing power and ecological safety. Low-temperature bleaches based on aromatic monoperoxycarboxylic acids

Table 6. Physical properties of some diacyl peroxides

Name	CAS registry no.	Formula	M_r	*mp*, °C	Decomposition temperature, °C	Active oxygen content, %
Diacetyl peroxide	[*110-22-5*]	$H_3C-\overset{O}{\overset{\|\|}{C}}-O-O-\overset{O}{\overset{\|\|}{C}}-CH_3$	118.09	30 (decomp.)	ca. 30	13.5
Acetyl benzoyl peroxide	[*644-31-5*]	$H_3C-\overset{O}{\overset{\|\|}{C}}-O-O-\overset{O}{\overset{\|\|}{C}}-Ph$	180.16	37–39	ca. 40	8.9
Dibenzoyl peroxide	[*94-36-0*]	$Ph-\overset{O}{\overset{\|\|}{C}}-O-O-\overset{O}{\overset{\|\|}{C}}-Ph$	242.23	106–7	ca. 110	6.6
Diisopropyl peroxydicarbonate	[*105-64-6*]	$(H_3C)_2CHO-\overset{O}{\overset{\|\|}{C}}-O-O-\overset{O}{\overset{\|\|}{C}}-OCH(CH_3)_2$	206.19	8–10	35–38	7.8
Acetyl cyclohexylsulfonyl peroxide	[*3179-56-4*]	$H_3C-\overset{O}{\overset{\|\|}{C}}-O-O-\underset{\underset{O}{\|\|}}{\overset{O}{\overset{\|\|}{S}}}-C_6H_{11}$	222.26	35–36	36	7.2

and long chain aliphatic diperoxycarboxylic acids are of interest for laundry detergents [178–182].

5. Diacyl Peroxides

Diacyl peroxides possess the following structural formula:

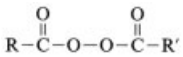

The best known compound of this type is dibenzoyl peroxide. Both symmetrical (R = R′) and unsymmetrical compounds are known. Furthermore, peroxydicarbonates of type (**3**) and sulfur-containing unsymmetrical diacyl peroxides of type (**4**), where R is an alkyl or cycloalkyl group, are also of interest.

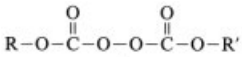

3

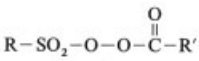

4

5.1. Properties

Physical Properties. As a result of the low dissociation energy of the O–O bond (ca. 120–130 kJ/mol [6], [7], [183]), diacyl peroxides are among the best radical-forming compounds. Most diacyl peroxides are low melting solids, which frequently decompose explosively even at low temperature. A summary of physical properties of some important diacyl peroxides is given in Table 6.

Extreme care must be taken when handling diacyl peroxides that have not ben desensitized as they are very sensitive toward shock, impact, and friction (see Chap. 9).

Chemical Properties. The products formed by thermal or photochemical decomposition [184] and the resulting homolytic cleavage of the O–O bond are shown below [185]:

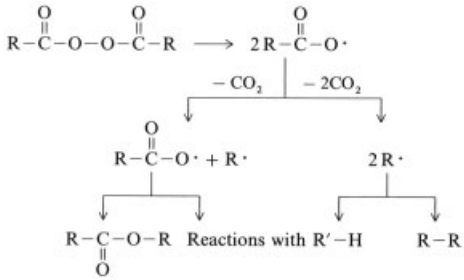

The radicals formed induce further homolytic cleavage, so that the decomposition is autocatalytically accelerated in a chain reaction.

The heterolytic decomposition of diacyl peroxides is catalyzed by strong Lewis acids and highly polar solvents, and leads to rearrangement products [186]:

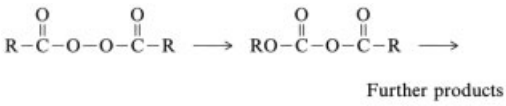

Acid hydrolysis of a diacyl peroxide gives a mixture of the acid and the peroxyacid [187]:

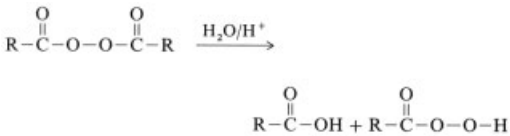

The reduction of diacyl peroxides leads to various products, depending on the reducing agent. Reaction with trialkylphosphines gives the corresponding carboxylic anhydride [188]; the carboxylic acid is formed with iodide solution, ammonium borohydrides, and with lithium aluminum hydride [189]. The reaction with chromium(II) perchlorate in aqueous ethanol gives a quantitative yield of carbon dioxide and the hydrocarbon [190].

Dialkyl peroxydicarbonates are heat, shock, and impact sensitive. They decompose at low temperature like the diacyl peroxides and at room temperature explosive decomposition may take place, particularly with the lower peroxycarbonates:

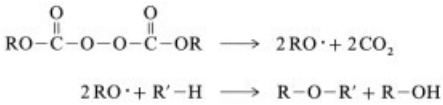

The risk of decomposition can be minimized by the use of additives such as iodine, oxygen, or substituted aromatics [191], [192].

5.2. Production

By Autoxidation. Only a few, mainly unsymmetrical, diacyl peroxides are produced by autoxidation. The sulfoxidation of hydrocarbons, leading to acetyl alkyl sulfonyl peroxides, is industrially important [193–195]. The reaction is induced by UV light [196–198] and by lithium and magnesium salts of carboxylic acids [199].

From Peroxycarboxylic Acids. The reaction of peroxycarboxylic acids with acid chlorides is used most frequently for the production of unsymmetrical diacyl peroxides [200–202].

3-Chlorobenzoyl 2-Chlorobutyryl Peroxide [*23128-41-8*]. Treatment of 85 % 3-chloroperbenzoic acid with 2-chlorobutyryl chloride in chloroform at −2 °C in the presence of aqueous sodium carbonate solution leads to the diacyl peroxide in almost quantitative yield with a purity of 90 %.

Chloroformate esters can be converted to acyl alkyl peroxycarbonates by reaction with peroxy acids in the presence of a base [203]:

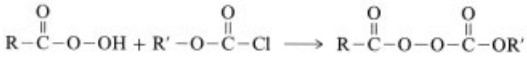

An analogous reaction with phosgene leads to diacyl diperoxycarbonates [204]:

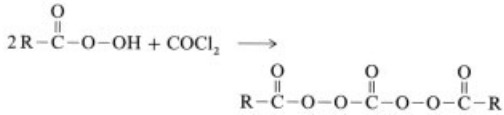

Diperoxydicarboxylic bis(carboxylic) dianhydrides can be obtained either by reaction of dicarboxylic acid dichlorides with peroxyacids, or by reaction of diperoxydicarboxylic acids with acid chlorides [205]:

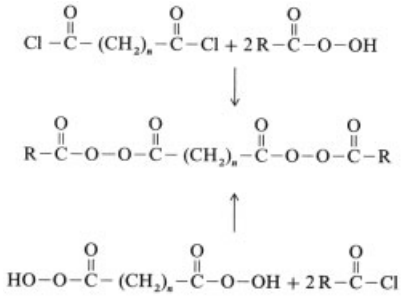

Acylation of Hydrogen Peroxide or Sodium Peroxide. Industrially, the most important method for the preparation of symmetrical diacyl peroxides is the reaction of carboxylic acid chlorides or anhydrides with hydrogen peroxide, sodium peroxide, or alkali-metal peroxoborates [206–214].

Diphthaloyl Peroxide [*37051-42-6*]. After 20 h reaction at 25 °C, 78 % diphthaloyl peroxide is

obtained from phthalic anhydride and hydrogen peroxide. About 14 % monoperoxyphthalic acid is formed as byproduct [215].

4,4′-Di-tert-butylbenzoyl Peroxide [*1712-79-4*]. Very pure crystalline product is obtained in quantitative yield by treatment of 4-*tert*-butylbenzoyl chloride with alkaline hydrogen peroxide (without solvent at 30 – 35 °C) in the presence of alkyl sulfonates to reduce the agglomeration of the peroxide particles [216].

For the preparation of hydrolytically sensitive diacyl peroxides, adducts of hydrogen peroxide (e.g., with urea) are used instead of the parent compound [217]. Dialkyl peroxydicarbonates are explosive when concentrated; they can be produced from chloroformate esters and alkaline hydrogen peroxide [218], [219].

Dicetyl peroxydicarbonate [*26322-14-5*] is obtained by treating 35 % alkaline hydrogen peroxide with cetyl chloroformate (hexadecyl chloroformate) at 30 °C, in 98 % yield after recrystallization from carbon tetrachloride [220].

5.3. Uses

Because of their facile thermal decomposition, diacyl peroxides are used, mainly as radical initiators for polymerization (especially diisopropyl peroxydicarbonate), and for cross-linking and hardening of polyester resins. Dialkyl peroxydicarbonates are gaining importance as low-temperature initiators for vinyl chloride polymerization.

The oxidizing power of diacyl peroxides is employed in organic synthesis and in bleaching additives for laundry detergents.

Diacyl peroxides are available commercially only in desensitized form (in solution or as a paste) because of their explosive nature. They should be stored and transported at low temperature.

6. Peroxycarboxylic Esters

The alkyl esters of peroxycarboxylic acids have been known since the beginning of the twentieth century. They can be separated into the following groups, based on the various possible structures:

Esters of peroxycarboxylic acids

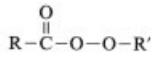

Esters of monoperoxydicarboxylic acids

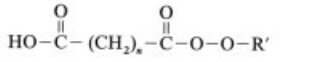

Esters of diperoxydicarboxylic acids

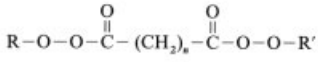

Esters of monoperoxycarbonic acid

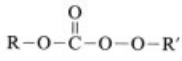

Esters of diperoxycarbonic acid

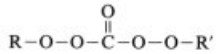

Alkylene bis(peroxy esters) of carboxylic acids

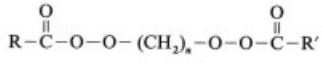

Esters of peroxycarbamic acid

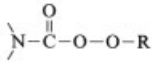

Esters of peroxysulfonic acids

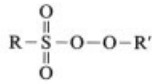

Monoperoxyesters of dibasic acids are best described by the *OO* or the *O* nomenclature, in order to indicate which alkyl group is attached to the peroxy group; hence the compound

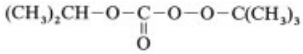

is named *OO-tert*-butyl-*O*-isopropyl monoperoxycarbonate.

6.1. Properties

Physical Properties. Peroxycarboxylic esters are thermally unstable because of the low dissociation energy of the O – O bond (ca. 147 – 164 kJ/mol). Therefore, they should be distilled only in exceptional cases and under very mild conditions. Purification by recrystallization or by chromatography is preferable. A summary of the important physical properties of some peroxyesters is given in Table 7.

Peroxyesters can be characterized by means of their refractive index and their density. Qualitative and quantitative determination is possible by polarography, gas chromatography, NMR

Table 7. Physical properties of some peroxyesters

Name	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (kPa)	n_D^{20}	d_4^{20}	Active oxygen content, %
Performic acid, *tert*-butyl ester	[*819-50-1*]	$HC(=O)-O-O-C(CH_3)_3$	118.1		28 (1.3)	1.3965[a]		13.5
Peracetic acid, *tert*-butyl ester	[*107-71-1*]	$H_3C-C(=O)-O-O-C(CH_3)_3$	132.2		50–51 (2.0)		0.888	12.1
Peroxyisobutyric acid, *tert*-butyl ester	[*109-13-7*]	$(CH_3)_2CH-C(=O)-O-O-C(CH_3)_3$	160.2	45				10.0
Perbenzoic acid, *tert*-butyl ester	[*614-45-9*]	$C_6H_5-C(=O)-O-O-C(CH_3)_3$	194.2	8	75–77 (0.27)	1.5007	1.043	8.2
Perbenzoic acid, cumyl ester	[*7074-00-2*]	$C_6H_5-C(=O)-O-O-C(CH_3)_2-C_6H_5$	256.3	45				6.2
Peroxynonanoic acid, *tert*-butyl ester	[*22913-02-6*]	$n\text{-}C_8H_{17}-C(=O)-O-O-C(CH_3)_3$	230.3		52–53 (0.001)	1.4271[b]	0.8868[c]	6.9
Monoperoxymaleic acid, *tert*-butyl ester	[*1931-62-0*]	$HO_2C-CH=CH-C(=O)-O-O-C(CH_3)_3$	188.2	114–116 (decomp.)				8.5
Monoperoxyphthalic acid, *tert*-butyl ester	[*15042-77-0*]	$HO_2C-C_6H_4-C(=O)-O-O-C(CH_3)_3$	238.2	104				6.7
Diperoxyadipic acid, di-*tert*-butyl ester	[*22158-52-7*]	$(H_3C)_3C-O-O-C(=O)-(CH_2)_4-C(=O)-O-O-C(CH_3)_3$	290.4	42–43				5.5
OO-tert-Butyl-*O*-isopropyl monoperoxycarbonate	[*2372-21-6*]	$(CH_3)_2CH-O-C(=O)-O-O-C(CH_3)_3$	176.2	58–59				9.1
Diperoxycarbonic acid, di-*tert*-butyl ester	[*3236-56-4*]	$(H_3C)_3C-O-O-C(=O)-O-O-C(CH_3)_3$	206.2		54–55 (0.07)	1.4106		15.5
2,5-Dimethyl-2,5-bis(benzoylperoxy)hexane	[*2618-77-1*]	$(H_3C)_2C(-O-O-C(=O)-C_6H_5)-(CH_2)_2-C(CH_3)_2(-O-O-C(=O)-C_6H_5)$	386.4	118				8.3
Peroxycarbamic acid, *tert*-butyl ester	[*18389-96-3*]	$H_2N-C(=O)-O-O-C(CH_3)_3$	133.1	51				12.0

[a] At 25 °C. [b] At 30 °C. [c] d_5^{30}.

and, in particular, IR spectroscopy, with characteristic bands in the carbonyl region at 1750–1820 cm^{-1} and in the fingerprint region between 800 and 900 cm^{-1} [221], [222]. They can also be used for quantitative determination of concentration for kinetic measurements.

Chemical Properties. Esters of peroxycarboxylic acids decompose thermally by radical, ionic, and concerted mechanisms. The most important process in their applications is the thermally or photochemically induced homolytic decomposition of *tert*-butyl peroxycarboxylates to give free radicals [223–225]. This process has been thoroughly investigated and can take place by way of two mechanisms. The actual mode of decomposition is determined by the polar and steric effects of the group R attached to the acyl group [226]:

1) Cleavage of the O – O bond and subsequent reactions

$$(CH_3)_3C-O-O-\underset{\underset{O}{\|}}{C}-R \longrightarrow (CH_3)_3C-O^{\bullet} + {}^{\bullet}O-\underset{\underset{O}{\|}}{C}-R$$

$$R-\underset{\underset{O}{\|}}{C}-O^{\bullet} \xrightarrow{} R^{\bullet} + CO_2 \quad \xrightarrow{+HX} R-COOH + X^{\bullet}$$

$$(CH_3)_3C-O^{\bullet} \xrightarrow{} (CH_3)_2CO + H_3C^{\bullet} \quad \xrightarrow{+HX} (CH_3)_3C-OH + X^{\bullet}$$

2) Simultaneous cleavage of the O – O and the C – R bonds

$$(CH_3)_3C-O-O-\underset{\underset{O}{\|}}{C}-R \longrightarrow (CH_3)_3C-O^{\bullet} + CO_2 + R^{\bullet}$$

The kinetics of thermal decomposition of tertiary peroxyesters is influenced by the structure of the peroxyester [227] and by the polarity of the solvent [228], [229]. The stability decreases in the following order: *tert*-butyl > *tert*-amyl > tert-octyl > cumyl. In addition, the stability depends on the degree of α-substitution of the carboxylic acid moiety [230], [231]. Primary and secondary peroxycarboxylic esters decompose rapidly by a concerted mechanism to give a carboxylic acid and a carbonyl compound [232], [233]:

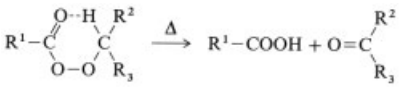

The radical-induced decomposition, in which a radical adds to the O – O bond resulting in cleavage, is usually undesired because fewer radicals are formed than expected for complete decomposition of a peroxycarboxylate. From one radical and one peroxyester molecule, which should normally give two radicals, only one radical is formed together with an ester molecule [234]:

$$R''^{\bullet} + R_3C-O-O-\underset{\underset{O}{\|}}{C}-R' \longrightarrow R_3C-O^{\bullet} + R''-O-\underset{\underset{O}{\|}}{C}-R'$$

Radical-induced decompositions are therefore not as effective as thermal homolysis, but can serve to extend the temperature range for the use of peroxycarboxylic esters.

Tertiary peroxyesters are more stable towards radical-induced decomposition than diacyl peroxides. This stability is more pronounced for peroxyesters that have an α-hydrogen atom or an unsaturated alkyl group [235].

Metal ions, especially copper(I) ions, catalyze the decomposition of peroxyesters [236], [237]:

$$R'-\overset{\overset{O}{\|}}{C}-O-O-R + Cu^{+} \longrightarrow R'-\overset{\overset{O}{\|}}{C}-O^{-} + R-O^{\bullet} + Cu^{2+}$$

This reaction is of preparative use in the acyloxylation of compounds with reactive C – H bonds.

The ionic decomposition of peroxycarboxylic acids occurs with rearrangement [238], [239]:

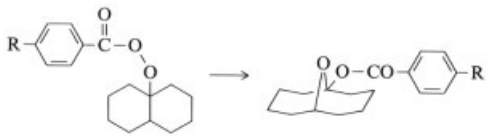

Peroxycarboxylic esters undergo hydrolysis much more readily than the corresponding carboxylic esters, with formation of carboxylic acids and hydroperoxides:

$$R_3C-O-O-\underset{\underset{O}{\|}}{C}-R' \xrightarrow{H_2O} R_3C-O-OH + R'-\underset{\underset{O}{\|}}{C}-OH$$

For the determination of peroxycarboxylic acids, conventional iodometric titration [240] is not usually applicable because it is always accompanied by an acid-catalyzed decomposition

that does not consume iodine. The reaction of the peroxycarboxylic ester with sodium ethoxide or methanolic potassium hydroxide, followed by either iodometric titration of the resulting hydroperoxide [241], or metal-catalyzed iodometric titration [242] is more suitable.

6.2. Production

Peroxycarboxylic esters cannot be produced by direct esterification of peroxycarboxylic acids. All of the literature methods start from the corresponding hydroperoxide. The most stable and, therefore, most readily accessible peroxyesters are the tertiary peroxycarboxylic esters. The most important production processes include the reaction of hydroperoxides with carboxylic acids, carboxylic acid halides, anhydrides, and *N*-acylimidazoles (Fig. 4).

From Hydroperoxides and Carboxylic Acids. Hydroperoxides can be esterified with carboxylic acids to give peroxyesters. The reaction is generally catalyzed with *p*-toluenesulfonic acid and pyridine [243], dicyclohexyldiimide [244], [245], or phosphate esters [246]. The reaction probably takes place via an intermediate activated carboxylic acid derivative. Mineral acids or boron trioxide are also suitable as catalysts for the production of peroxyesters of C_2-C_{10} aliphatic carboxylic acids and peroxyesters of aromatic carboxylic acids [247]. The water of reaction can be removed by azeotropic distillation with toluene, benzene, or cyclohexane.

From Hydroperoxides and Carboxylic Acid Halides. The reaction of carboxylic acid halides, in particular the chlorides, with hydroperoxides is the most important and industrially most frequently employed method for the synthesis of peroxycarboxylic esters. The process is usually carried out with high selectivity under Schotten–Baumann conditions in aqueous [248–251] or aqueous – organic media. Ether or a chlorinated hydrocarbon is used as the organic solvent [252–255]. In addition to the use of uniform starting materials, mixtures of carboxylic acid chlorides and/or hydroperoxides are also used, since the resulting mixtures of peroxyesters possess special properties for their application as polymerization initiators [256], [257].

tert-Butyl Peroxyoctadecanoate

[*2123-89-9*]. Octadecanoyl chloride (0.165 mol) is added at 10 °C to a mixture of 0.2 mol *tert*-butyl hydroperoxide in petroleum naphtha. Subsequently, 0.2 mol of pyridine is added to this mixture in such a way that the temperature does not exceed 15 °C. After standing for 40 min at room temperature, the mixture is shaken with 10 % hydrochloric acid. The product crystallizes from the organic phase, is washed with water and dried; yield 96 %.

The reaction takes place smoothly at low temperature in the presence of organic bases, preferably pyridine [258–265], but triethylamine [266–268], 2,6-dimethylpyridine [269] or 1,4-diazabicyclo[2.2.2]octane [270] have also been used. This method permits the use of an anhydrous system and is therefore suitable for acyl chlorides that are poorly soluble in aqueous media. In this way peroxyesters of carbamic acids, carbonic acid [271], and acetylene carboxylic acids can be prepared.

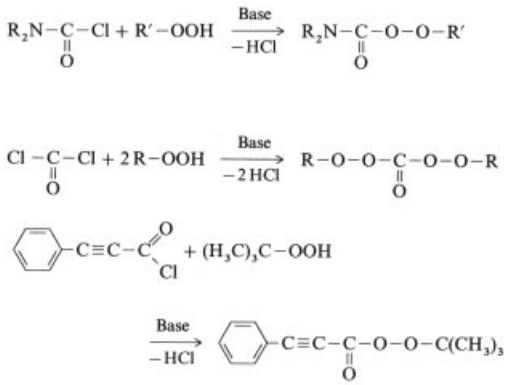

Di- or tetraperoxyesters can be produced both from the corresponding dicarboxylic acid dichlorides or tetracarboxylic acid tetrachlorides and hydroperoxides [272–274], as well as from bishydroperoxides and carboxylic acid chlorides [275]:

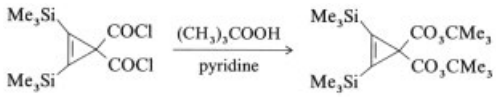

Primary and secondary alkyl peroxycarboxylic esters are best prepared from the alkali metal salts of the corresponding hydroperoxide [276–

280], since they are not readily accessible by other methods.

From Hydroperoxides and Carboxylic Anhydrides. Hydroperoxides, alkyl metal peroxides, or alkyl nonmetal peroxides can also be acylated by carboxylic anhydrides:

The reaction is carried out in the presence of alkali-metal hydroxide solution, organic bases such as pyridine, or alkali-metal acetates [281], [282]. This method is particularly suitable for the synthesis of dicarboxylic acid mono-*tert*-alkyl peroxyesters [283], [284]. The use of mineral acids as catalysts is only possible for acid-stable hydroperoxides.

tert-Butyl Monoperoxysuccinate [*28884-42-6*]. *tert*-Butyl hydroperoxide (0.13 mol) and 0.06 mol pyridine are added in succession to 0.12 mol succinic anhydride at 45–55 °C. After 15 min the mixture is cooled, extracted with chloroform, and the extract washed with hydrochloric acid and water. The product is isolated in 83 % yield by distillation from the dried organic phase.

Alkylperoxy organogermanes, -stannanes, -silanes, and -stibines are cleaved heterolytically at the O–M bond by carboxylic anhydrides [285–288]:

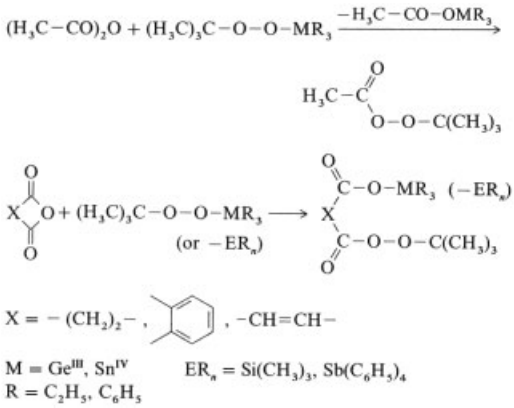

The corresponding dicarboxylic mono-*tert*-butyl peroxyesters can be obtained by treatment with 0.5 N sulfuric acid or by introducing dry hydrogen chloride [289].

From Hydroperoxides and *N*-Acylimidazoles. The acylation of alkylhydroperoxides with the very reactive *N*-acylimidazoles, available from carboxylic acids and *N,N′*-carbonylimidazoles, is particularly useful when the corresponding acyl chloride cannot readily be obtained pure [290–292]. Isolation and purification of the *N*-acylimidazole intermediate is not necessary. The reaction can be carried out from imidazole in a one-pot process [293]:

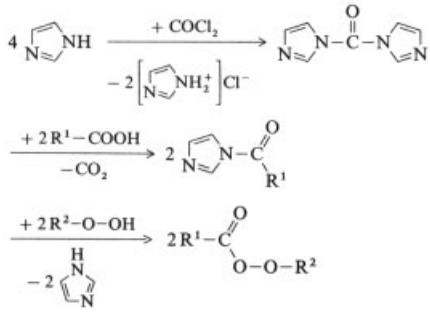

This method is of limited use for the production of peroxycarboxylic esters derived from carboxylic acids containing a C=C double bond [294].

Other Methods. The acid-catalyzed reaction of hydroxycarboxylic acids with hydrogen peroxide in ether leads to cyclic peroxycarboxylic esters [295]:

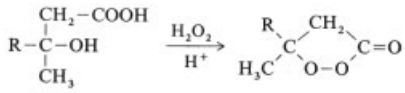

Esters of peroxychloroacetic acid are obtained by acidolysis of dialkyl peroxides with chloroacetic acid [296]:

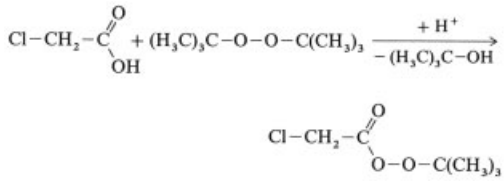

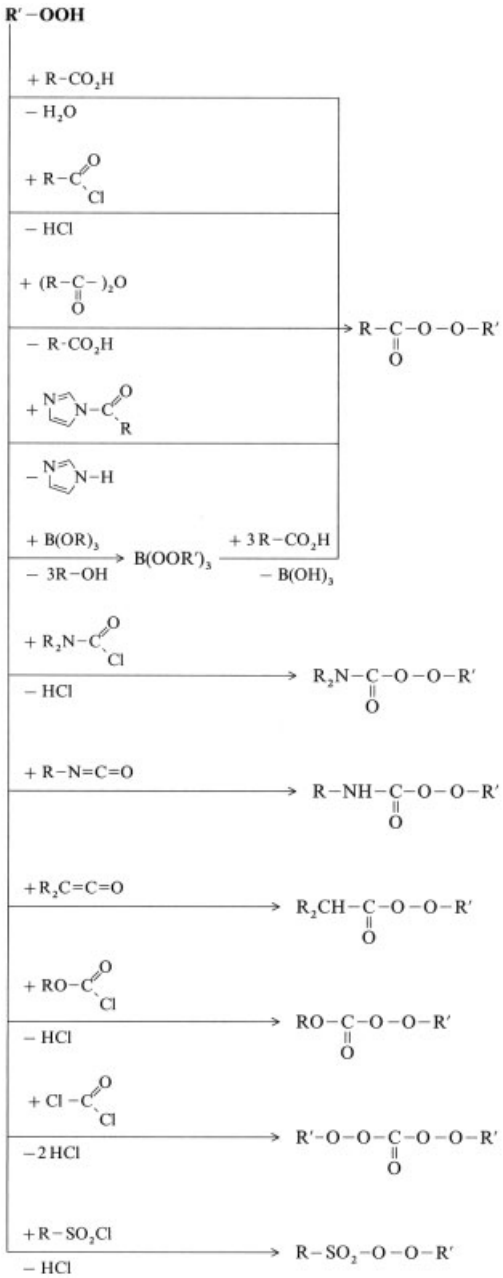

Figure 4. Reactions used for the synthesis of peroxyesters from hydroperoxides R′−OOH.

Ketenes and isocyanates add hydroperoxides under neutral or weakly acidic conditions to give the peroxyesters of carboxylic or carbamic acids [297–299]:

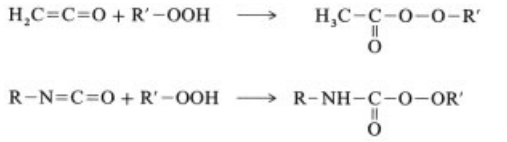

This method is particularly suitable for the production of base-sensitive peroxycarboxylic esters. The reaction of triperoxyborates, produced from borate esters and hydroperoxides, with carboxylic acids leads similarly to the corresponding peroxycarboxylic esters [300]:

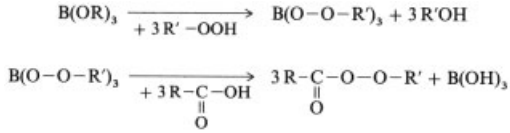

The acid-catalyzed reaction is carried out in aromatic or cycloaliphatic solvents such as benzene, toluene, or cyclohexane. However, the reaction is not suitable for primary hydroperoxides.

6.3. Uses

The most important industrial application of peroxyesters is in the initiation of polymerization reactions, and as hardeners for resins. The more reactive types (e.g., *tert*-butyl or cumyl peroxypivalate) are used mainly as initiators for the polymerization of ethylene or vinyl chloride. *tert*-Butyl perbenzoate is used predominantly as a hardener for polyester resins.

The usual commercial products are of technical grade or are product mixtures that have been desensitized by the addition of an organic solvent (aliphatic hydrocarbons, carboxylic esters). Frozen peroxide emulsions can be stored safely and exhibit very good storage stability [301].

Peroxycarboxylates play an important role in preparative organic chemistry in, for example, inter- and intramolecular epoxidation of olefins in the presence of copper salts [302]. The peroxyesters react with strong nucleophiles, such as Grignard reagents or phosphines, with transfer of the alkoxy group. The copper-catalyzed acyloxylation of C−H bonds occurs by a radical mechanism. In this way, α-acyloxy derivatives of olefins, ethers, and thioethers

can be obtained [303]. Thermal homolysis of peroxyesters in protic solvents can be used for the decarboxylation of carboxylic acids.

7. α-Oxyperoxides

α-Oxyperoxides possess the structural element

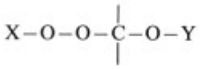

where X = −H, −R, and Y = −H, −R, −OH, −OR. The substituents X and Y may also be part of a cyclic group. A selection of possible structures is given in Table 8.

Table 8. Structure and nomenclature of α-oxyperoxides

Substituents		
X	Y	Name
H	H	α-hydroxyhydroperoxide
H	R	α-alkoxyhydroperoxide
H	OH	geminal bishydroperoxide
H	OR	α-hydroperoxyperoxide
R	H	α-hydroxyperoxide
R	R	α-alkoxyperoxide
R	OR	geminal diperoxide, peroxyketal

Cyclic alkoxyperoxides of the type

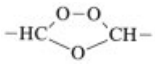

are known as ozonides. α-Oxyperoxides are unstable and highly reactive and therefore capable of numerous reactions.

7.1. Properties

Physical Properties. The stability of α-oxyperoxides varies widely depending on the structure, and range from explosive to being stable up to the melting or boiling point. Physical properties of some selected compounds are listed in Table 9.

Chemical Properties. As the chemical behavior of α-oxyperoxides depends markedly on the structure, only a few typical reactions are described here.

The thermolysis of α-oxyperoxides leads to a complex mixture of products [304]. Dicyclohexylidene diperoxide and tricyclohexylidene diperoxide give C_{12} or C_{15} cycloalkanes which are not readily accessible by other routes, as well as lactones with 12- or 15-membered rings [305], [306].

The decomposition of dimeric cyclopentanone and cyclohexanone peroxides (dicyclopentylidene diperoxide, dicyclohexylidene diperoxide) in the presence of a Cu(I) catalyst in aqueous hydrochloric acid leads to 5-chlorovaleric acid and 6-chlorocaproic acid, respectively, in high yield [307], [308].

Cycloaliphatic α-oxyperoxides decompose to form open-chain dicarboxylic acids in the presence of iron or copper salts as catalysts [305], [309], whereas dehydration of these compounds leads to cyclic geminal alkylidene triperoxides (1, 2, 4, 5, 7, 8-hexaoxycyclononanes).

Heating α-oxyperoxides in water causes elimination of hydrogen peroxide to give the dimeric α,α'-dihydroxyperoxides which can be dehydrated to give ozonides [310]:

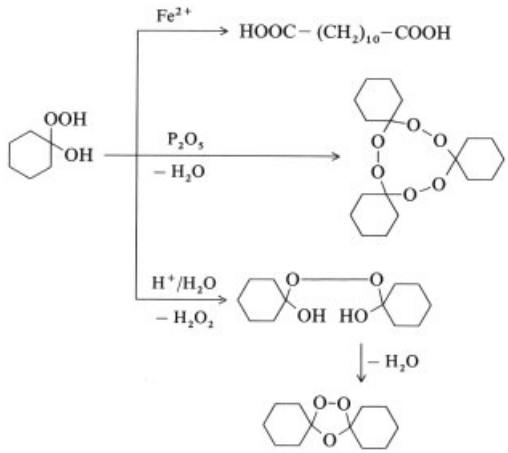

Ozonides (cyclic α-alkoxyperoxides) give a multitude of products upon thermal or radical induced decomposition [311]. The reactive species is the carbonyl oxide (**6**), formed from the primary ozonide (or 1,2,3-trioxolane) (**5**):

Table 9. Physical properties of some α-oxyperoxides

Name	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (kPa)	Active oxygen content, %
2-Chloro-1-hydroperoxy-cyclohexanol	[*15250-08-5*]	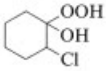	166.71	76		9.6
α,α′-Dihydroxydicyclo-hexyl peroxide	[*2407-94-5*]	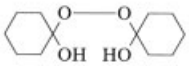	230.30	69–71		6.9
α-Hydroperoxy-α′-hydroxy-dicyclohexyl peroxide	[*78-18-2*]	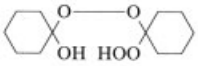	246.30	78		13.0
α,α′-Bishydroperoxydicyclo-hexyl peroxide	[*2699-12-9*]	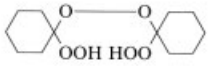	262.30	82–83		18.3
Dicyclohexylidene diperoxide	[*183-84-6*]	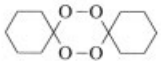	228.29	127–129		14.0
Tricyclohexylidene triperoxide	[*182-01-4*]	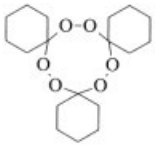	342.43	93		14.0
Dicyclopentylidene diperoxide	[*311-38-6*]	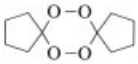	200.23	95–98		16.0
Diisopropylidene diperoxide	[*1073-91-2*]	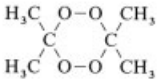	148.16	132		21.6
2,2-Bis(*tert*-butylperoxy)-propane	[*4262-61-7*]	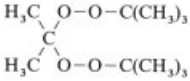	220.31		69–70 (2.0)	14.5
2,2-Bis(*tert*-butylperoxy)-butane	[*2167-23-9*]	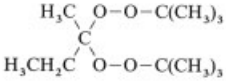	234.34		50 (0.27)	13.7
1,1-Bis(*tert*-butylperoxy)-cyclohexane	[*3006-86-8*]	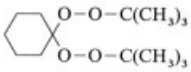	260.37		52–54 (0.02)	12.3
Ethylene ozonide (1,2,4,-trioxolane)	[*289-14-5*]	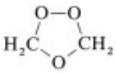	76.05		18 (2.1)	21.0
Isobutene ozonide (3,3-di-methyl-1,2,4-trioxolane)	[*22409-33-2*]	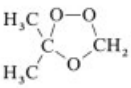	104.10		42–42.5 (18.0)	15.4
Tetramethylethylene ozonide (3,3,5,5-tetramethyl-1,2,4-trioxolane)	[*15592-50-4*]	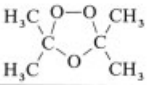	132.16	133–134		12.1

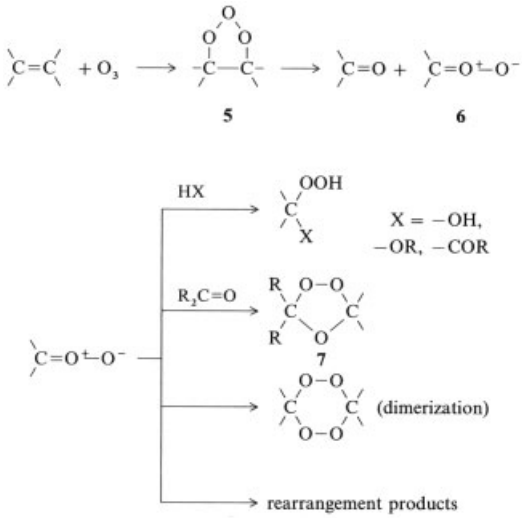

The primary ozonide (**5**) is unstable and generally not isolable [312], [313]. In contrast, ozonides of structure (**7**), with the exception of the simplest representatives such as ethylene ozonide which is explosive, are stable compounds that can be isolated. They only tend to decompose spontaneously under the catalytic influence of heavy metals [313].

Ozonides can be reduced to aldehydes and oxidized to carboxylic acids [304].

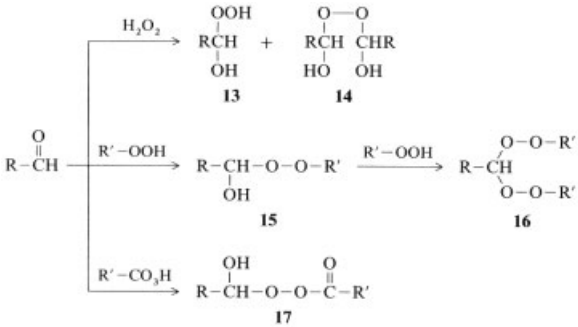

Figure 5. Reactions of aldehydes with hydrogen peroxide, alkyl hydroperoxides, and peroxycarboxylic acids

7.2. Production

Three main routes exist for the production of α-oxyperoxides:

1) Autoxidation of ethers, acetals, aldehydes, or alcohols
2) Addition of hydrogen peroxide, alkyl hydroperoxides, or peroxycarboxylic acids to carbonyl compounds
3) Ozonolysis of unsaturated compounds

α-Oxyperoxides by Autoxidation. Autoxidation is not suitable for the preparation of α-oxyperoxides because the reaction usually leads to mixtures of products. Ethers are readily oxidized by atmospheric oxygen to yield ether hydroperoxides, from which the explosive polymeric ether peroxides are formed [186], [314]:

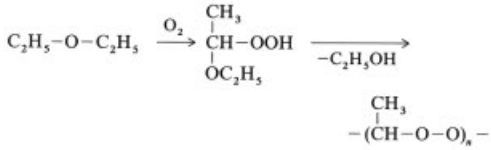

Acetals are easier to autoxidize than ethers. Dimers of structure **10** are formed from 1,3-dioxolanes (**8**) via the hydroperoxide (**9**), when R = H or CH_3 [315]. Bulky groups impede the dimeriza-tion. Autoxidation of vinyl-1,3-dioxolanes (R = −CH = CH_2) leads to heterocyclic α-oxyperoxides (**12**), via formation of an unstable allyl peroxide (**11**) intermediate [316].

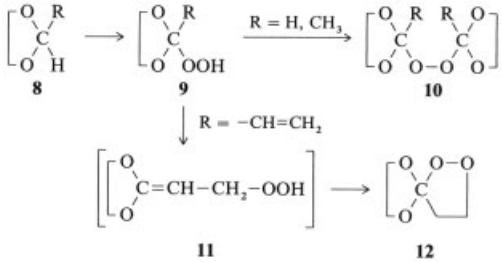

2-Hydroperoxy-2-isopropyl-1,3-dioxolane [*34594-19-9*] (R = $-CH(CH_3)_2$) is obtained in good yield by the reaction of 2-isopropyl-1,3-dioxolane with atmospheric oxygen at 75 °C [317].

Autoxidation of alcohols gives mixtures of α-hydroxyhydroperoxides and α,α'-bishydroperoxyperoxides [318–320]:

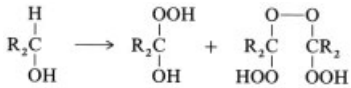

2-Hydroperoxy-2-propanol, the primary oxidation product of 2-propanol, decomposes in the presence of water to give hydrogen peroxide and acetone. This process is employed indus-

trially for the production of hydrogen peroxide (→ Hydrogen Peroxide) [321], [322].

The autoxidation of aldehydes proceeds via intermediate α-hydroxyalkyl peroxycarboxylates to give the corresponding peroxycarboxylic acids:

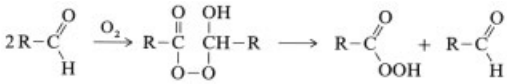

This reaction is used industrially for the production of peroxyacetic acid from acetaldehyde.

α-Oxyperoxides by Addition of Hydrogen Peroxide, Alkyl Hydroperoxides, or Peroxycarboxylic Acids to Carbonyl Compounds. Aldehydes react with hydrogen peroxide to give mixtures of α-hydroxyhydroperoxides (**13**) and α,α'-dihydroxyperoxides (**14**). Aldehydes react with alkylhydroperoxides under acid catalysis to form α-hydroxyperoxides (**15**), which react with excess hydroperoxide to form peroxyacetals (**16**) [323], [324]. With peroxycarboxylic acids, aldehydes and ketones form α-hydroxyalkyl peroxycarboxylic esters (**17**), which may also be obtained by acylation of α-hydroxyhydroperoxides [325]. These reactions are summarized in Figure 5.

In the presence of concentrated sulfuric acid, aldehydes react with hydrogen peroxide to form cyclic geminal alkylidenedi- and triperoxides (1,2,4,5-tetroxanes and 1,2,4,5,7,8-hexaoxacyclononanes) [326].

Treatment of ketones with hydrogen peroxide gives a multitude of products, most of which arise from rearrangements. For example, the following compounds are formed from cyclohexanone:

α-Hydroperoxycyclohexanol

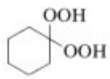

α, α-Dihydroperoxycyclohexane

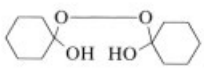

α, α′-Dihydroxydicyclohexyl peroxide

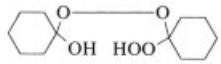

α-Hydroperoxy-α′-hydroxydicyclohexyl peroxide

α, α′-Dihydroperoxydicyclohexyl peroxide

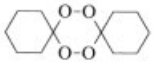

Dicyclohexylidene diperoxide (3,6-dicyclohexylidene-1,2,4,5-tetroxane)

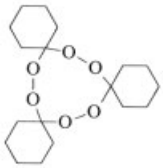

Tricyclohexylidene triperoxide (3, 6, 9-tricyclohexylidene-1, 2, 4, 5, 7, 8,-hexaoxacyclononane)

α,α'-Dihydroperoxydicyclohexyl Peroxide

[*2699-12-9*]. Treatment of cyclohexanone with hydrogen peroxide in acetonitrile in the presence of perchloric acid results in formation of α,α'-dihydroperoxydicyclohexyl peroxide in good yield. The compound can be recrystallized from hexane.

The α-oxyperoxides of acetone [327], cyclohexanone [305], [326], [328], [329], methyl ethyl ketone [327], and acetylacetone [330] have been studied in detail. Ketones react with alkyl hydroperoxides to give peroxyketals, whereby the peroxy hemiketals normally cannot be isolated [32], [327], [331], [332]:

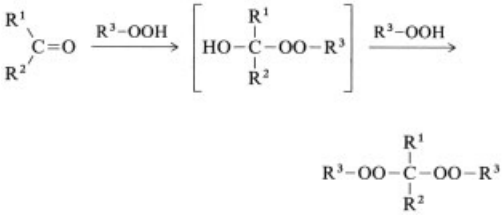

Reaction of α,α'-dihydroperoxyperoxides with ketones leads to formation of cyclic geminal alkylidenedi- and triperoxides, which are used as starting materials for the synthesis of macrocyclic dienes [305], [306], [333–335].

α-Oxyperoxides by Ozonolysis. In the ozonolysis of olefins, other α-oxyperoxides in addition to the ozonides are formed [304], [336]. The reaction is of little preparative use as it is generally not very selective for α-oxyperoxides,

and gives defined compounds only for tetrasubstituted olefins [311].

7.3. Uses

α-Oxyperoxides are mainly used as vulcanization accelerators, cross-linking agents [337], [338] and hardeners, in particular methyl ethyl ketone peroxide and methyl isobutyl ketone peroxide, as well as cyclohexanone peroxide and various peroxyketals. In combination with transition metal salts a wide temperature range can be covered (e.g., thermal hardening of polyesters), so that suitable catalyst systems are available for various materials [339–342].

Ozonides are used as disinfectants [343]. Ozonolysis is applied industrially using fatty acids, e.g., in the production of nonanedioic acid and nonanoic acid from oleic acid [344], [345], or of decanedioic acid from cyclodecene [346]. α-Oxyperoxides are commercially available only in desensitized form because of their explosiveness. Phosphates, dibutyl phthalate, and aliphatic hydrocarbons are typical additives.

8. α-Aminoperoxides

In addition to the hydroperoxy or peroxy group, α-aminoperoxides bear a nitrogen atom on the same carbon atom; both functional groups may be part of a ring.

8.1. Properties

Physical Properties. The physical properties of some α-aminoperoxides are listed in Table 10. Their stability varies widely, ranging from high shock and impact sensitivity with the lower homologues to a thermal stability up to the melting point.

Chemical Properties. Some typical reactions of α-aminoperoxides are as follows [347]:

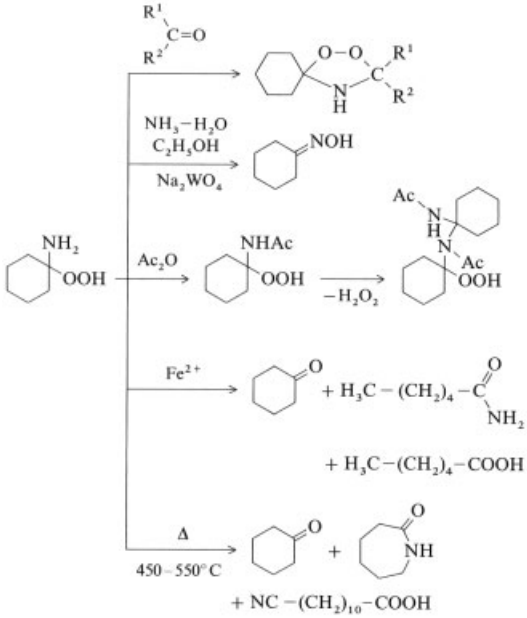

The unstable 1-amino-1-hydroperoxycyclohexane reacts with ketones to form cyclic α-aminoperoxides. With aqueous ammonia in alcohol and catalytic amounts of sodium tungstate, cyclohexanone oxime is formed. Acylation gives a dimeric product with elimination of hydrogen peroxide. Catalytic decomposition in the presence of iron leads to cyclohexanone together with small amounts of caproic acid and its amide. Cyclohexanone, together with ε-caprolactam, is also the pyrolysis product of 1-amino-1-hydroperoxycyclohexane.

1,1′-Peroxydicyclohexylamine rearranges, both thermally and photochemically, to decane-1,10-dicarboxylic acid imide; byproducts include ε-caprolactam and cyclohexanone. These reactions are shown in Figure 6.

Thermolysis at >400 °C leads mainly to 11-cyanoundecanoic acid.

Radical decomposition with iron or copper salts gives caproic acid amide and cyclohexanone.

ε-Caprolactam is formed from 1,1′-peroxydicyclohexylamine in good yield by treatment with sodium methoxide in methanol.

Table 10. Physical properties of some α-aminoperoxides

Name	CAS registry no.	Formula	M_r	*mp*, °C	*bp*, °C (kPa)	Active oxygen content, %
1-Amino-1-hydroperoxycyclohexane	[*24075-24-9*]	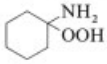	131.18	57–58 (decomp.)		12.2
1,1′-Peroxydicyclohexylamine	[*21842-28-4*]	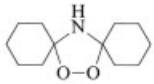	211.30	40–42	94–97 (0.05)	7.6
1,1′-Peroxydicyclopentylamine	[*21842-27-3*]	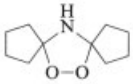	183.25	22–23		8.7
2,2′-Peroxydiisopropylamine	[*24075-03-4*]	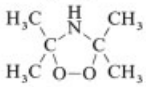	131.18		50 (1.6)	12.2
2,2′-Peroxydiisobutylamine	[*24075-04-5*]	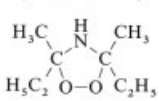	159.23		66–68 (1.6)	10.1
1-Amino-1-*tert*-butylperoxy-cyclohexane	[*24829-66-1*]	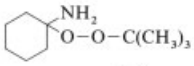	187.28		36–38 (0.03)	8.5
1-Cyclohexylamino-1-hydroperoxy-cyclohexane	[*2808-61-9*]	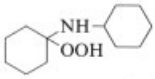	212.31	88–89		7.5
1-Hydroperoxy-1-azocyclohexane	[*32140-97-9*]	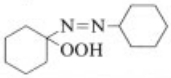	225.31	73–75		7.1

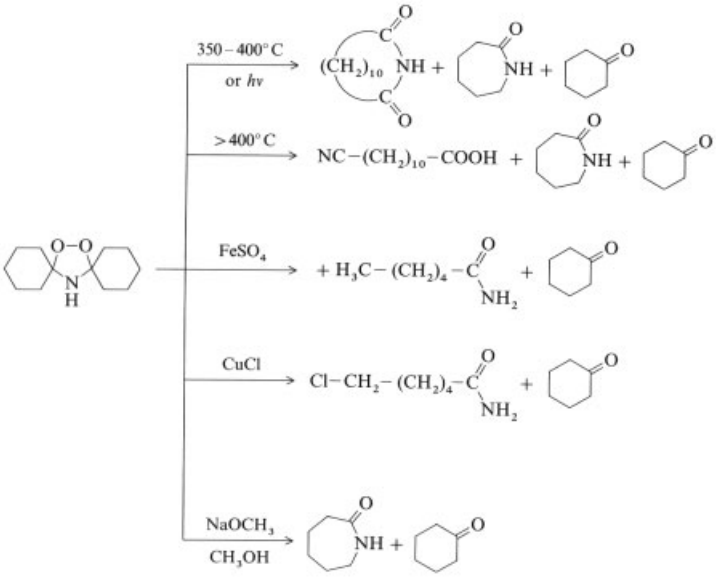

Figure 6. Reactions of 1,1′-peroxydicyclohexylamine

8.2. Production

α-Aminoperoxides can be produced by the same methods as α-oxyperoxides. Autoxidation is important only for Schiff's bases [348]. High yields of α-aminoperoxides are obtained by the Mannich reaction of amines with formaldehyde and hydroperoxides:

$$R_2NH + CH_2O + R'-OOH \xrightarrow{-H_2O} R_2N-CH_2-O-O-R'$$

The most useful synthetic method is the reaction of carbonyl compounds with hydrogen peroxide in the presence of ammonia or primary amines [349–358]:

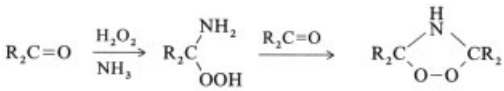

The initially formed α-aminoperoxides are frequently not isolable and react further with the carbonyl compound to give 1,2,4-dioxazolidines.

Cyclopentanone forms 1,1′-peroxydicyclopentylamine, whereas 1-aminohydroperoxides are formed from large-ring ketones.

1-Amino-1-hydroperoxycyclohexane [*24075-24-9*]. A solution of 30 % H_2O_2 is added to a mixture of methanol, ammonia, and cyclohexanone with stirring at 0 °C. The precipitated solid is removed by filtration and the crude product obtained in 88 % yield [359]. The use of alkyl hydroperoxides instead of hydrogen peroxide gives 1-aminoperoxides.

8.3. Uses

α-Aminoperoxides are important intermediates in the synthesis of ε-caprolactam [360–363], of cyclohexanone oxime [347], of derivatives of terminal dicarboxylic acids [347], [349], and in the production of nylon-12 from cyclohexanone [364]. More recently, they are also being considered for use as initiators in polymerization of styrene [365], as disinfectants [366], and as antimalarial agents [367].

9. Safety Measures

9.1. Hazards

Organic peroxides are frequently very unstable compounds which can decompose spontaneously and explosively under thermal or mechanical stress. Such decomposition can be caused by shock, impact, friction, or by the catalytic effect of impurities. Organic peroxides are therefore subject to the regulations for explosives [368].

Specific safety and protective measures must be observed when working with organic peroxides [369], [370]. To reduce the hazards involved, commercially available peroxides are usually desensitized by addition of inert solids or liquids. Typical desensitizing agents are water, inert solvents (aliphatic and halogenated hydrocarbons, phthalate and phosphate esters, acetals), and inert inorganic solids (sulfates, phosphates, borates, silicates, carbonates). Therefore peroxides are available in solid, liquid, or paste form. Investigation of the hazards associated with organic peroxides involves a wide range of test methods which are described in the relevant legal regulations. In Germany BAM is responsible for conducting and evaluating such tests. Similar institutions exist in other countries [371]. In the most important tests the following properties of peroxides are investigated:

1) Detonation or deflagration ability (BAM 2″ steel tube test)
2) Shock sensitivity (BAM falling hammer test)
3) Friction sensitivity
4) Heat sensitivity under defined confinement in a steel tube (Koenen test)
5) Explosive energy in a lead block (Trauzl method)

The explosive range of peroxide-containing mixtures can be displayed graphically, as shown in Figure 7 for the $H_2O_2-HCOOH-H_2O$ system [372], [373]. For industrial applications, knowledge of explosive ranges is essential, in particular for reactions in which performic or peracetic acid is generated in situ. In these cases the aqueous phase that is usually present is extremely explosive.

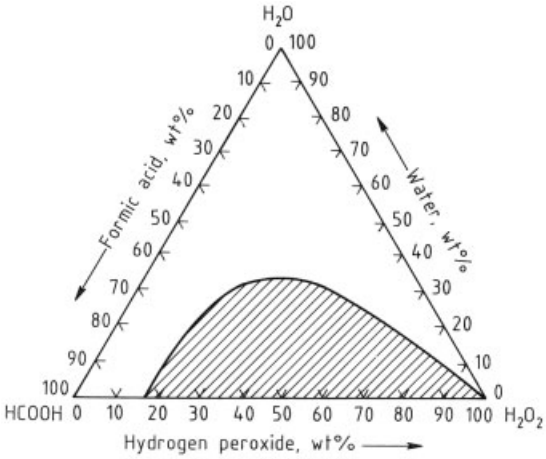

Figure 7. Explosive range (hatched area) for H_2O_2–HCOOH–H_2O.

The explosive range can be investigated, for example, by 2″ steel tube tests. Passage through this range must be avoided during a reaction by appropriate process control.

9.2. Transportation and Storage

Organic peroxides are subject to the regulations for dangerous materials [374]. The establishment of the maximum permitted ambient temperature for storage and transportation is based on the SADT (self-accelerating decomposition temperature), above which peroxide decomposition is self-accelerating. The SADTs of some commercial peroxides are listed in Table 11. Organic peroxides and their formulations must be transported and stored at a temperature significantly below the SADT. For thermally very unstable peroxides maximum transportation temperatures are specified.

In Germany, the "Safety Precautions for Organic Peroxides" are applied for storage [370]. National transportation regulations are based on the recommendations for the transportation of dangerous goods by the United Nations. The UN has reexamined the classification of organic peroxides and new regulations have been introduced [375]. Organic peroxides are assigned to the hazard groups A to G by means of a flow diagram, which also indicates the permitted amount for a given container and container type. Following the new regulations, some peroxides with a low active oxygen content may be exempted. Until then, the following dangerous goods regulations are valid for the Federal Republic of Germany, within which the organic peroxides are assigned mainly to Class 5.2:

Road transportation:	GGVS with Appendices A and B, ADR
Transportation by rail:	GGVE/RID
Maritime:	GGVSee with Appendix, IMDG
Air Freight:	ICAO Technical Instructions, IATA-DGR

9.3. Handling

Safety glasses and gloves must be worn when handling organic peroxides. Protective clothing is required when working with larger amounts of peroxides [369]. In the event that skin or eye contact with peroxides occurs, the afflicted area should be immediately rinsed with water for a sufficiently long period (at least 15 min). Contaminated clothing should be removed. A physician or eye specialist should be consulted without delay.

Cleanliness at the workplace is an essential prerequisite for working safely with peroxides because even small amounts of contaminants such as ash, dust, or traces of heavy metals could lead to decomposition. The containers and equipment employed must be stable against peroxides and must not promote decomposition. Suitable materials are glass, porcelain, earthenware, stainless steel, and certain plastics such as teflon or polyethylene. When setting up apparatus suitable precautions must be taken as decomposition can lead to a considerable pressure increase. Often it is best to work with open containers. Peroxide samples removed from their original containers should not be returned because of the risk of contamination. Preliminary experiments should be carried out with small quantities of material. Furthermore, peroxides should only be added in small amounts, in order to avoid high local concentrations. Transferring and weighing of peroxides must be carried out carefully, as even slight friction or impact with a spatula or a scoop could cause an explosion with the more sensitive compounds.

Spills of water-soluble peroxides can be cleared up without risk by the use of large amounts of water. Compounds that are insoluble in water must be diluted and desensitized

Table 11. Self-accelerating decomposition temperature (SADT) of commercial peroxides

Peroxide	CAS registry no.	Supplied form	SADT, °C
tert-Butyl hydroperoxide	[*75-91-2*]	70 % in water	88
		solution in phosphate	ca. 65
Cumene hydroperoxide	[*80-15-9*]	80 % water	80
Di-*tert*-butyl peroxide	[*110-05-4*]	technical grade (liquid)	79–80
2,5-Dimethyl-2,5-bis(*tert*-butylperoxy)-hexane	[*2618-77-1*]	liquid	82
Dicumyl peroxide	[*80-43-3*]	technical grade (powder)	88
		40 % (powder with chalk)	> 70
Peracetic acid	[*79-21-0*]	40 % in acetic acid/H_2O	55
		15 % in acetic acid/H_2O/H_2O_2	60
Dibenzoyl peroxide	[*94-36-0*]	aqueous, technical grade	68
		technical grade (flakes)	ca. 50
		50 % in plasticizer	52
		50 % in silicone oil	63
Diacetyl peroxide	[*110-22-5*]	25 % solution in phthalate	ca. 35
tert-Butyl perbenzoate	[*614-45-9*]	technical grade	ca. 60
Diisopropyl peroxydicarbonate	[*105-64-6*]	40 % solution in phthalate	ca. 5
Acetyl cyclohexanesulfonyl peroxide	[*3179-56-4*]	water-damp granules	ca. 15
		28 % solution in phthalate	ca. 5
Methyl ethyl ketone peroxide	[*1338-23-4*]	40 % solution in phthalate	63
		50 % solution in phosphate	ca. 65
Methyl isobutyl ketone peroxide	[*37206-20-5*]	60 % solution in phthalate	64
Cyclohexanone peroxide	[*12262-58-7*]	moistened with water	ca. 50
		50 % solution in phosphate	ca. 50
2,2-Bis(*tert*-butylperoxy)butane	[*2167-23-9*]	50 % solution in phthalate	63

with sufficiently large amounts of appropriate solvents. After being absorbed in an inert material they should be destroyed immediately. The hazard associated with moist peroxides increases considerably with increasing dryness. Before absorbing these materials they should therefore be dampened sufficiently with water. Small quantities of peroxides can be disposed of most simply by incineration. For larger quantities, information can often be obtained from the producer.

10. Toxicology

Organic peroxides generally possess a pronounced effect on body tissue, and marked damage must be expected at the site of direct contact. Nevertheless, there are many peroxide compounds whose effect on the organism is less pronounced or that lead to no noticeable effect on health. The chemical structure and physical properties such as water solubility, fat solubility, and vapor pressure, as well as the type and composition of the formulation (concentration, solvent, additives), can be important for penetration of human tissue and in reactions with cell compartments and macromolecules. In contrast, the influence of UV and ionizing radiation is not so predominant in the effect of organic peroxides on the mammalian organism, and is not comparable with the formation and effect of endogeneous radicals produced by radiation.

The main effects associated with contact with peroxides are local irritation and burns of the skin, the mucous membranes of the eyes and the respiratory tract [376] (see Table 12).

Local tissue reactions may appear immediately following exposure or may only become apparent after a few days. Healing of the damaged tissue is frequently very slow and irreversible effects in the tissue are posssible. In the skin, irritation appears in the form of reddening (erythema) and swelling (edema), as well as burns (necrosis, corrosion), which are characterized by severe degenerative changes and scarring. In direct contact with the eye, even very small amounts of peroxides can cause inflam-

Table 12. Acute toxicity data for organic peroxides

Compound	CAS registry no.	Irritation *		Acute Toxicity	
		Skin	Eye	LD_{50}, mg/kg (rat, oral)	LC_{50}, (4 h, inhalative)
tert-Butyl hydroperoxide	[*75-91-2*]	+ + +	+ + +	406, 790 **	500 ppm
Cumene hydroperoxide	[*80-15-9*]	+ + +	+ + +	382	220 ppm
Diisopropylbenzene peroxide	[*26762-93-6*]	+ + +			
p-Menthane hydroperoxide	[*37837-09-5*]	+ + +			
Di-*tert*-butyl peroxide	[*110-05-4*]	(+)	(+)	6750	> 4000 ppm
Dicumyl peroxide	[*80-43-3*]	(+)	(+)	4100	0.09 mg/L
Peracetic acid	[*79-21-0*]	+ + +	+ + +	17.5, 1410 **	0.45 mg/L
Diperoxydodecanedioic acid	[*66280-55-5*]	–	+ +	> 5000	
Diacetyl peroxide	[*110-22-5*]	+ + +			
Dicyclohexyl peroxide	[*1758-61-8*]	+ + +			
Dibenzoyl peroxide	[*94-36-0*]	(+)	+ +	> 5000	24 mg/L
Dilauroyl peroxide	[*105-74-8*]	–	+	> 500	100 mg L^{-1} d^{-1}
Diisopropyl peroxydicarbonate	[*105-64-6*]	(+)	(+)	2140	
Acetyl cyclohexyl-sulfonyl peroxide	[*3179-56-4*]	+ + +	+	(moderate)	
tert-Butyl 2,2-dimethylperoxypropionate	[*927-07-1*]	–	–	4300	
tert-Butyl peracetate	[*107-71-1*]	+ +	+ +		
tert-Butyl 2-ethylperoxyhexanoate	[*3006-82-4*]	–		675	> 8.2 mg/L
tert-Butyl perbenzoate	[*614-45-9*]	(+)	+ + +	10000	> 200 mg/L
Methyl ethyl ketone peroxide	[*1338-23-4*]	+ + +	+ + +	1012, 3800 **	> 200 mg/L
Cyclohexanone peroxide	[*12262-58-7*]	+ + +	+ + +	484	200 ppm

* Degree of irritation: – no effect, (+) weak, + slight, + + moderate, + + + severe (caustic). ** LD_{50} value after dermal application (rabbit).

mation of the conjuctiva, nacreous opacity of the cornea, or loss of the eye [377]. This can often be avoided by immediate and intensive rinsing with water. Nevertheless, this procedure must be carried out for a sufficiently long period, especially with contamination by solid or powdery material, as even strongly diluted solutions can cause irritation. For example, dry benzoyl peroxide possesses a smaller irritant effect than a 10 % aqueous solution.

Inhalation of the vapor of volatile peroxides also leads to irritation or burns of the mucous membranes of the eye or the upper respiratory tract (nose, throat, bronchia). Inflammation of the respiratory tract or even lung edema can result from prolonged exposure [378].

Sensitization of the skin by peroxides has been observed mainly with dibenzoyl peroxide [379], [380]. In animal experiments methyl ethyl ketone peroxide and cyclohexanone peroxide exhibited ambiguous positive results [381], [382], whereas diperoxydodecandioic acid was clearly negative [383]. Acute poisoning may result from the inhalation and ingestion of organic peroxides, but in practice the hazard of resorptive effects (systemic toxicity) is low. After swallowing, peroxides are moderately toxic (see Table 12). In animal experiments, the acute toxicity for substances administered orally, dermally and by inhalation was characterized by tremors, decrease in muscle tone, clonic convulsions, hyperactivity, and dyspnea. Furthermore, irritation, hemorrhages, and inflammation were detected in the mucous membranes of the small intestine or respiratory tract [384], [385]. After chronic exposure to peroxy compounds, effects on the general condition, on the red blood cells, and impairment of the immune system in the blood (SH groups, peroxidase activity) towards oxidative reagents (oxidative stress) may appear [386]. Data exist only for a few organic peroxide compounds with respect to their carcinogenic, mutagenic, or reproductive toxicological potential. These data are not always sufficiently unambiguous to allow estimation of the risk associated with chemicals of this group. Observations on humans are not available. In long-term studies with rats and mice no substance-related tumors appeared after oral or subcutaneous application of benzoyl peroxide for up to two years [387].

After repeated application of organic peroxides to the skin of mice (skin painting studies) none of the tested compounds were found to be an absolute carcinogen. In two-stage ex-

periments on the chemical carcinogenicity (initiating or promoting properties in the formation of tumors), none of the compounds examined showed a complete carcinogenic or tumor-initiating activity. Nevertheless, a number of compounds (benzoyl peroxide [388–390], *tert*-butyl hydroperoxide [391], lauroyl peroxide [391], peracetic acid [392], perbenzoic acid [393], chloroperbenzoic acid [394], and methyl ethyl ketone peroxide [381]) were found to promote skin tumors in mice and hamsters. For example, with dimethylbenzanthracene as tumor initator, subsequent treatment of the skin of mice with benzoyl peroxide led to skin carcinomas, whereas with benzoyl peroxide alone only hyperpigmentation, formation of skin scales, and a low level of hyperplasia of the skin was found [388–390]. When UV radiation was used as a tumor initiator, weak promoting activity was exhibited by methyl ethyl ketone peroxide [394], but not by benzoyl peroxide [395], [396]. Investigations of the mechanism of action of organic peroxides in vivo and in vitro showed an influence on the activity of the calcium-dependent protein kinase [397] and enzymes of the cytochrome P 450 peroxidase system [398], [399], as well as the participation of glutathione (SH groups) in the reduction [400] of the exogenously administered peroxides. In more recent investigations, the activation of oncogenes has been reported [401].

Testing of organic peroxides for mutagenicity by means of experiments with bacterial assays (*Salmonella typhimurium, Escherichia coli*) showed essentially no increased rate of mutation (dilauroyl peroxide [393], benzoyl peroxide [393], dibenzoyl peroxide [402], di-*tert*-butyl peroxide [393], peracetic acid [393]). However, both positive and negative results have been reported for *tert*-butyl hydroperoxide [393], [403], dicumyl peroxide [393], [404], monoperoxysuccinic anhydride [405], and disuccinoyl peroxide [404]. Direct interaction with DNA in the form of single strand breaks [406], [407] or as sister chromatid exchanges (SCE) [408] could be demonstrated by in vitro experiments. DNA transformation with *Bacillus subtilis* bacteria was observed with disuccinoyl peroxide [409], but not with di-*tert*-butyl peroxide [410]. In the mouse lymphoma forward test *tert*-butyl perbenzoate acted as a weak mutagen [408].

Reproductive toxicology experiments have been carried out with three-day-old chicken embryos. Cyclohexanone peroxide, cumene hydroperoxide, methyl ethyl ketone peroxide, benzoyl peroxide, acetylacetone peroxide, *tert*-butyl perbenzoate, dicumyl peroxide, and lauroyl peroxide led to embryotoxic effects. Malformation was only observed to a small extent [411].

11. References

General References

1. S. Patai: *The Chemistry of Peroxides,* John Wiley & Sons, Chichester 1983.
2. A. G. Davies: *Organic Peroxides,* Butterworth, London 1961.
3. E. G. E. Hawkins: *Organic Peroxides,* E. and F. F. Spon, London 1961.
4. J. O. Edwards: *Peroxide Reaction Mechanisms,* Interscience Publ., New York 1961.
5. D. Swern: *Organic Peroxides,* Wiley-Interscience, New York vol. I, 1970, vol. II, 1971, vol. III, 1972.
6. *Methodicum Chimicum,* **vol. 5,** Thieme Verlag, Stuttgart 1975, p. 723.
7. *Ullmann,* 4th ed., **17,** 661.
8. D. Swern in D. Barton, D. W. Ollis (eds.): *Comprehensive Organic Chemistry,* **vol. 1,** Pergamon Press, Oxford 1979, p. 909.
9. *Kirk-Othmer,* 3rd ed., **17,** 27.
10. W. M. Weigert, P. Kleinschmidt, "Organische Peroxide, Synthesen und Anwendungen," *Chem. Ztg.* **98** (1974) 583.
11. J. C. Mitchell, *Chem. Soc. Rev.* **14** (1985) 339.
12. W. M. Weigert et al., *Chem. Ztg.* **99** (1975) 106.
13. *Houben-Weyl,* 4th ed., **E 13.**
14. W. Weigert: *Wasserstoffperoxid und seine Derivate,* Hüthig Verlag, Heidelberg 1978.

Specific References

15. B. C. Brodie, *Justus Liebigs Ann. Chem.* **108** (1858) 79.
16. W. Adam, A. J. Bloodworth, *Annu. Rep. Prog. Chem. Sect. B* **75** (1978) 342.
17. E. N. Frankel, W. E. Neff, E. Selke, D. D. Brooks, *Lipids* **23** (1988) 295.
18. Modern Plastics International, Sept.–Oct. 1964 – 1984.
19. *Houben-Weyl,* 4th ed., **E 13/1,** 176 – 236.
20. D. Griller, K. U. Ingold, *Acc. Chem. Res.* **9** (1976) 13.

21. W. H. T. Davison, *J. Chem. Soc.* 1951 2456; S. S. Ivanchev, A. I. Yurzhenko, Y. N. Anisimov, *Zh. Fiz. Khim.* **39** (1965) 1900; G. D. Mendenhall, *Angew. Chem.* **89** (1977) 220; K. D. Gundermann, *Chem. Ztg.* **99** (1975) 279.
22. A. Rieche, M. Schulz, *Angew. Chem.* **70** (1958) 694; G. Dobson, A. Huges, *J. Chromatogr.* **16** (1964) 416; S. W. Bukata, L. L. Zabrocki, M. F. McLaughlin, *Anal.Chem.* **35** (1963) 885.
23. R. D. Mair, A. J. Graupner, *Anal. Chem.* **36** (1964) 194.
24. M. L. Whisman, B. H. Eccleston, *Anal. Chem.* **30** (1958) 1638.
25. E. N. Frankel, W. E. Neff, E. Selke, D. D. Brooks, *Lipids* **23** (1988) 295.
26. K. S. Pitzer, *J. Am. Chem. Soc.* **70** (1948) 2140.
27. S. W. Benson, *J. Chem. Phys.* **40** (1964) 1007.
28. J. D'Ans, H. Gold, *Chem. Ber.* **92** (1959) 2559.
29. N. A. Sokolov, Yu. A. Aleksandrov, *Russ. Chem. Rev. (Engl. Transl.)* **47** (1978) 172.
30. A. J. Everett, G. J. Minkoff, *Trans. Faraday Soc.* **49** (1953) 410; I. M. Kolthoff, A. I. Medalia, *J. Am. Chem. Soc.* **71** (1949) 3789; D. Barnard, K. R. Hargrave, G. M. C. Higgins, *J. Chem. Soc.* 1956, 2845.
31. Atochem, EP 237 402, 1987.
32. L. J. Durham, C. F. Wurster, Jr., H. S. Moscher, *J. Am. Chem. Soc.* **80** (1958) 327, 332; M. S. Kharasch, A. Fono, W. Nudenberg, *J. Org. Chem.* **16** (1951) 113.
33. H. Kropf, F. Woehrle, *J. Chem. Res. Synop.* 1987, 387.
34. G. Rücker, K. Frey, *Liebigs Ann. Chem.* 1987, 389.
35. W. Adam et al., *J. Org. Chem.* **43** (1978) 1154.
36. N. A. Porter, C. B. Ziegler, F. F. Khouri, D. H. Roberts, *J. Org. Chem.* **50** (1985) 2252.
37. P. G. Cookson, A. G. Davies, B. P. Roberts, *J. Chem. Soc. Chem. Commun.* 1976, 1022.
38. N. A. Milas, D. M. Surgenor, *J. Am. Chem. Soc.* **68** (1946) 205, 643; N. A. Milas, L. H. Perry, *J. Am. Chem. Soc.* **68** (1946) 1938.
39. N. A. Milas, O. L. Mageli, *J. Am. Chem. Soc.* **74** (1952) 1471.
40. H. R. Williams, H. S. Mosher, *J. Am. Chem. Soc.* **76** (1954) 2984, 2987.
41. M. Schulz, A. Rieche, K. Kirschke, *Chem. Ber.* **100** (1967) 370.
42. E. Schmitz, A. Rieche, O. Brede, *J. Prakt. Chem.* **312** (1970) 30.
43. E. J. Lorand, E. I. Edwards, *J. Am. Chem. Soc.* **77** (1955) 4035.
44. H. Hock, H. Kropf, *Angew. Chem.* **69** (1957) 313.
45. E. R. Bell et al., *Ind. Eng. Chem.* **41** (1949) 2597.
46. G. A. Russell, *J. Am. Chem. Soc.* **78** (1956) 1047.
47. Shell, US 2 403 772, 1946.
48. Shell, US 2 845 461, 1958.
49. Halcon, DE-OS 1 518 996, 1965.
50. Halcon, DE-OS 1 568 774, 1966.
51. Halcon, FR 1 484 567, 1966.
52. Halcon, DE-OS 1 668 200, 1967.
53. L. F. Martin: *Organic Peroxide Technology,* Noyes Data Corp., London 1973.
54. C. Walling, S. A. Buckler, *J. Am. Chem. Soc.* **77** (1955) 6032; C. F. Cullis, E. Fersht, *Combust. Flame* **7** (1963) 185.
55. G. Wilke, P. Heimbach, *Justus Liebigs Ann. Chem.* **652** (1962) 7.
56. R. W. Denny, A. Nickon, *Org. React. (N.Y.)* **20** (1973) 133.
57. A. P. Schaap: *Singlet Molecular Oxygen,* Dowden, Hutchinson & Ross, Stroudsburg, Pa., 1976.
58. K. Gollnick, H. J. Kuhn in H. H. Wasserman, R. W. Muray (eds.): *Singlet Oxygen,* Academic Press, New York 1979, p. 287.
59. I. Saito, T. Matsuura, *Tetrahedron* **41** (1985) 2237.
60. A. A. Frimer, L. M. Stephenson in A. A. Frimer (ed.): *Singlet O_2,* **vol. II,** CRC Press, Boca Raton, Fla., 1985, p. 67.
61. Akzo, EP 250 024, 1987.
62. J. H. Raley, F. F. Rust, W. E. Vaughan, *J. Am. Chem. Soc.* **70** (1948) 88, 95, 1338; P. L. Hanst, J. G. Calvert, *J. Phys. Chem.* **63** (1959) 104; L. J. Durham, H. S. Moscher, *J. Am. Chem. Soc.* **82** (1960) 4537; **84** (1962) 2811; E. S. Huyser et al., *J. Am. Chem. Soc.* **86** (1964) 2401, 4148; R. Hiatt, D. J. LaBlanc, C. Thankachan, *Can. J. Chem.* **52** (1974) 4090; E. S. Huyser, F. Tang, *J. Org. Chem.* **43** (1978) 1016.
63. E. G. E. Hawkins, *J. Chem. Soc.* 1959, 2169, 2801.
64. G. Baker et al., *J. Chem. Soc.* 1965, 6965, 6970.
65. Akzo, EP 273 990 A 1, 1988.
66. Texaco Dev. Corp., CA 743 559, 1966.
67. E. R. Bell et al., *Ind. Eng. Chem.* **41** (1949) 2597.
68. G. O. Schenk, *Angew. Chem.* **64** (1952) 12; *Nachr. Chem. Tech.* **19** (1971) 446; W. Adam, *Angew. Chem.* **86** (1974) 683; W. Adam, *Chem. Ztg.* **99** (1975) 142.

69. G. O. Schenck, K. Ziegler, *Naturwissenschaften* **32** (1944) 125.
70. C. Moureu, C. Dufraisse, P. M. Dean, *C. R. Hebd. Séances Acad. Sci.* **182** (1926) 1440; C. Dufraisse, E. Etienne, J. Rigaudy, *C. R. Hebd. Séances Acad. Sci.* **226** (1948) 1773.
71. A. A. Frimer, *Chem. Rev.* **79** (1979) 359.
72. A. P. Schaap, K. A. Zaklika in H. H. Wassermann, R. W. Murray (eds.): *Singlet Oxygen,* Academic Press, New York 1979, p. 174.
73. P. D. Bartlett, M. E. Landis in H. H. Wassermann, R. W. Murray: *Singlet Oxygen,* Academic Press, New York 1979, p. 244.
74. W. Adam, G. Cilento, *Angew. Chem.* **95** (1983) 525.
75. W. Adam, F. Yany in A. Hassner (ed.): *The Chemistry of Heterocyclic Compounds, Small Ring Heterocycles,* **vol. 42,** part 3, John Wiley & Sons, Chichester 1983, p. 351.
76. A. L. Baumstark in A. A. Frimer (ed.): *Singlet* O_2 , vol. II, part 1, CRC Press, Boca Raton, Fla., 1985, p. 1.
77. M. Schulz, O. Johansson, DL 63 491, 1967.
78. Union Rhein, Braunkohlen Kraftstoff, DE-OS 1 911 176, 1969.
79. JP – Kokai 01 61453 A 2 and 89 61453, 1989.
80. Bergwerksgesellschaft Hibernia, US 3 254 130, 1966.
81. Noury & van der Lande, US 3 337 639, 1967.
82. Pennwalt Corp., DE 3 034 235, 1989.
83. Massachusetts Institute of Technology, US 3 505 363, 1970.
84. Wallace & Tiernan, US 3 308 163, 1967.
85. FMC Corp., US 3 135 805, 1964.
86. E. R. Altwicker, *Chem. Rev.* **67** (1967) 511.
87. K. U. Ingold, *Chem. Rev.* **61** (1961) 575.
88. M. S. Kharasch, P. Pauson, W. Nudenberg, *J. Org. Chem.* **18** (1953) 322.
89. L. R. Mahoney, *Angew. Chem.* **81** (1969) 555.
90. L. Reich, S. S. Stivala: *Autoxidation of Hydrocarbons and Polyolefins,* Marcel Dekker, New York 1969, p. 139.
91. J. Pospísil, *Adv. Polym. Sci.* **36** (1980) 69.
92. T. Endo, K. Suga, Y. Orikasa, S. Kojima, *J. Appl. Polym. Sci.* **37** (1989) 1815.
93. Nippon Oils and Fats, JP 63 254 110 A 2 and 88 254 110, 1988.
94. Toray Silicone, EP 304 676, 1989.
95. Johnson, EP 303 057 A 2, 1989.
96. W. H. Richardson in S. Patai (ed.): *The Chemistry of Peroxides,* Wiley-Interscience, New York 1983, p. 129.
97. J. Skerjanc, A. Regent, B. Plesnicar, *J. Chem. Soc. Chem. Commun.* 1980, 1007.
98. R. Kavcic, B. Plesnicar, *J. Org. Chem.* **35** (1970) 2033.
99. W. E. Parker, L. P. Witnauer, D. Swern, *J. Am. Chem. Soc.* **79** (1957) 1929.
100. S. W. Benson, *J. Chem. Phys.* **40** (1964) 1007.
101. K. S. Pitzer, *J. Am. Chem. Soc.* **70** (1948) 2140.
102. A. H. Sehon et al., *Can. J. Chem.* **41** (1963) 1819, 1826.
103. D. Lefort et al., *Bull. Soc. Chim. Fr.* 1959, 1385;1961, 2219.
104. W. E. Parker, L. P. Witnauer, D. Swern, *J. Am. Chem. Soc.* **80** (1958) 323.
105. L. Heslinga, W. Schwaiger, *Rec. Trav. Chim. Pays-Bas* **85** (1966) 75.
106. W. Schlenk, Jr., *Fortschr. Chem. Forsch.* **2** (1951) 92; F. Cramer, *Angew. Chem.* **68** (1956) 115; F. Schwochow, L. Puppe, *Angew. Chem.* **87** (1975) 659.
107. B. Plesnicar, R. Kavcic, D. Hadzi, *J. Mol. Struct.* **20** (1974) 457.
108. *Houben-Weyl,* 4th ed., **E 13,** pp. 1418, 1438.
109. *Kirk-Othmer,* 1st. suppl. vol., 1957, p. 662.
110. Propylox, EP 4407, 1978.
111. Produits Chimiques Ugine Kuhlmann, EP 223 96, 1979.
112. Degussa, DE 2 018 713, 1970.
113. Degussa, DE-OS 2 125 160, 1971.
114. Degussa, DE 1 170 960, 1962.
115. Degussa, DE-OS 2 145 157, 1971.
116. Degussa, DE-OS 2 141 155, 1971.
117. K. Fancovic et al., CS 198 936, 1978.
118. Bayer, Degussa, DE-OS 2 519 287, 1975.
119. Bayer, Degussa, DE 2 519 293, 1975.
120. Bayer, Degussa, DE-AS 2 519 295, 1975.
121. Bayer, Degussa, DE 2 856 665, 1978.
122. Bayer, Degussa, DE-OS 2 519 295, 1975.
123. Bayer, DE-OS 2 038 318, 1970.
124. Mitsui Toatsu Chem., JP 63 159 365, 1988.
125. Produits Chimiques Ugine Kuhlmann, EP 83 894, 1982.
126. Mitsui Toatsu Chem., JP 63 115 857, 1988.
127. Degussa, DE-OS 3 720 562, 1987.
128. Procter & Gamble, US 4 233 235, 1979.
129. Procter & Gamble, US 4 244 884, 1979.
130. Degussa, Henkel, DE 3 418 450, 1984.
131. Degussa, Henkel, DE-OS 3 539 036, 1987.
132. Food Machinery, US 4 172 086, 1977.
133. Degussa, Henkel, DE 2 930 546, 1978.
134. Degussa, Henkel, DE 3 320 497, 1983.
135. L. S. Silbert, E. Siegel, D. Swern, *J. Org. Chem.* **27** (1962) 1336; *Org. Synth.* **43** (1963) 93.
136. Pittsburg Plate Glass, US 3 655 738, 1969.
137. Pittsburg Plate Glass, US 3 880 914, 1975.

138. Degussa, Henkel, DE 2 929 839, 1979.
139. Degussa, DE 3 426 792, 1984.
140. Interox Chem., GB 7 936 178, 1978.
141. G. B. Payne, *J. Org. Chem.* **24** (1959) 1354.
142. L'Air Liquide, FR 1 452 737, 1965.
143. Kharkov Polytechnic Inst., SU 615 066, 1976.
144. Degussa, Henkel, DE 3 438 529, 1984.
145. Degussa, Henkel, DE 3 539 036, 1985.
146. Monsanto Co., US 4 680 145, 1987.
147. Food Machinery, US 3 231 605, 1966.
148. Food Machinery, US 3 247 244, 1966.
149. Food Machinery, US 3 462 480, 1969.
150. B. A. Palei et al., *Zh. Prikl. Khim. (Leningrad)* **61** (1988) 123.
151. P. R. H. Speakman, *Chem. Ind. (London)* 1978, 579.
152. Y. Ogata, Y. Sawaki, *Tetrahedron* **23** (1967) 3327.
153. Food Machinery, US 3 485 869, 1967.
154. R. N. McDonald, R. N. Steppel, J. E. Dorsey, *Org. Synth.* **50** (1970) 15.
155. Agence Nationale de Valorisation, FR 2 324 626, 1975.
156. J. Y. Nedelec, J. Sorba, D. Lefort, *Synthesis* 1976, 821.
157. J. Rebek, L. Marshall, R. Wolak, J. McManis, *J. Am. Chem. Soc.* **106** (1984) 1170.
158. J. Rebek, L. Marshall, J. McManis, R. Wolak, *J. Org. Chem.* **51** (1986) 1649.
159. Mitsui Toatsu Chem., JP 73 13 527, 1973.
160. Daicel Co., DE 2 165 168, 1970.
161. M. R. Yazdanbakhch, Dissertation, Universität Hamburg 1977.
162. Mitsubishi, JP 55 27 904, 1976.
163. S. Levush, I. Garbuzyuk, Z. Prisyazhnyuk, V. Bryukhovetskii, SU 979 336, 1981.
164. V. A. Bryukhovetskii, S. S. Levush, F. B. Moin, V. U. Shevchuk, *Kinet. Katal.* **17** (1976) 1130.
165. D. Paronikyan, I. Vardanyan, A. Nalbandyan, *Arm. Khim. Zh.* **36** (1983) 20.
166. Mitsubishi, DE 2 515 033, 1975.
167. E. Sandler et al., *Neftekhimiya* **23** (1983) 674.
168. Mitsubishi, JP 55 43 058, 1978.
169. C. R. Dick, R. F. Hanna, *J. Org. Chem.* **29** (1964) 1218.
170. Bayer, DE 2 658 943, 1976.
171. Bayer, EP 46 534, 1980.
172. Oxiran Chem. Corp., DE 2 364 181, 1972.
173. Food Machinery, US 4 026 798, 1977.
174. J. Gutknecht, *Pharm. Technol.* **9** (1988) 46.
175. Peroxid-Chemie, EP 302298, 1989.
176. Interox, EP 233 731, 1987.
177. J. A. L. Fraser, *Spec. Chem.* **7** (1987) 178, 180, 182, 184, 186.
178. H. Krüger, P. Kuzel, H. Schwab, *Chem. Ztg.* **99** (1975) 132.
179. Degussa, DE-OS 3 709 347, 1988.
180. Degussa, DE-OS 3 709 348, 1988.
181. Kao Corp., JP 63 112 698, 1988.
182. T. Bolsman, R. Kok, A. D. Vreugdenhil, *J. Am. Oil Chem. Soc.* **65** (1988) 1211.
183. I. Jaffe, E. J. Prosen, M. Szwarc, *J. Chem. Phys.* **27** (1957) 416.
184. E. D. Skakovskii et al., *Vestsi Akad. Navuk BSSR, Ser. Khim. Navuk* 1987, 26.
185. C. E. Boozer et al., *Chem. Abstr.* **59** (1963) 2609;J. E. Guillet, US 3 232 921, 1966;J. E. Guillet et al., *Ind. Eng. Chem. Prod. Res. Dev.* **3** (1964) 257.
186. A. Rieche, *Angew. Chem.* **70** (1958) 251; R. C. Lamb, J. G. Pacifici, P. W. Ayers, *J. Am. Chem. Soc.* **87** (1965) 3928; R. C. Lamb et al., *J. Org. Chem.* **31** (1966) 147.
187. A. M. Clover, G. F. Richmond, *J. Am. Chem. Soc.* **29** (1903) 179.
188. L. Horner, B. Anders, *Chem. Ber.* **95** (1965) 2470.
189. E. A. Sullivan, A. A. Hinckley, *J. Org. Chem.* **27** (1962) 3731.
190. J. K. Kochi, P. E. Mocadlo, *J. Org. Chem.* **30** (1965) 1134.
191. S. G. Cohen, D. B. Sparrow, *J. Am. Chem. Soc.* **72** (1950) 611.
192. H. C. McBay, O. Tucker, *J. Org. Chem.* **19** (1954) 869.
193. Pennwalt Corp., DE-OS 1 954 965, 1970.
194. Pennwalt Corp., DE-OS 2 053 463, 1971.
195. Pennwalt Corp., DE 2 631 910, 1975.
196. Pennwalt Corp., US 3 580 955, 1971.
197. Pennwalt Corp., US 3 706 783, 1972.
198. Pfizer+Co., US 3 438 882, 1969.
199. Koppers Comp., US 3 397 245, 1968.
200. G. Schroeder, R. Lombard, *Bull. Soc. Chim. Fr.* 1964, 542.
201. Argus Chem. Corp., US 3 502 701, 1970.
202. Argus Chem. Corp., US 4 032 605, 1976.
203. Canadian Industries, US 3 108 093, 1963.
204. Eastman Kodak Co., CA 725 695, 1966.
205. Eastman Kodak Co., CA 725 694, 1966.
206. Montecatini, DE 1 192 181, 1965.
207. Elektrochem. Werke, NL 6 517 064, 1966.
208. Rhône-Progil, DE-OS 2 256 255, 1973.
209. Teknar Apex Co., US 3 728 402, 1973.
210. H. Spaeth, R. Kneissl, DE 1 668 355, 1974.
211. Akzo, DE-OS 2 628 272, 1976.
212. Peroxid-Chemie, DE 1 768 199, 1974.
213. Sanken Kako Co., JP 74 40 220, 1974.
214. Dow Chem. Co., US 3 674 858, 1972.
215. Interox, BE 848 246, 1975.

216. Société Chimiques de Carbonnages, DE-AS 2 334 269, 1974.
217. D. F. DeTar, L. A. Carpino, *J. Am. Chem. Soc.* **77** (1955) 6370.
218. Akzo, DE-OS 2 428 184, 1975.
219. Goodyear, DE 1 099 738, 1970.
220. Liljeholmens Stearinfabriks, DE-AS 1 947 434, 1969.
221. P. D. Bartlett, R. R. Hiatt, *J. Am. Chem. Soc.* **80** (1958) 1398.
222. Y. N. Anisimov, S. S. Ivanchev, *Zh. Anal. Khim.* **21** (1966) 113.
223. C. Rüchardt, V. Golzke, G. Range, *Chem. Ber.* **114** (1981) 2769.
224. J. J. Davies, J. G. Miller, *J. Am. Chem. Soc.* **99** (1977) 245.
225. L. Thijs, S. N. Gupta, D. C. Neckers, *J. Org. Chem.* **44** (1979) 4123.
226. S. O. Lawesson, G. Schroll in S. Patai (ed.): *The Chemistry of Carboxylic Acids and Esters,* Interscience, New York 1969, p. 670.
227. T. Komai, K. Matsuyama, M. Matsushima, *Bull. Chem. Soc. Jpn.* **61** (1988) 1641.
228. Y. Nakashio, T. Yamamoto, M. Hirota, *Nippon Kagaku Kaishi* 1988, 69.
229. E. M. Gavryliv, R. G. Makitra, Ya. N. Pirig, *Org. React. (Tartu)* **24** (1987) 3.
230. C. S. Sheppard, V. R. Kamath, *Polym. Eng. Sci.* **19** (1979) 597.
231. W. Duismann, C. Rüchardt, *Justus Liebigs Ann. Chem.* 1976, 1834.
232. B. G. Dixon, G. B. Schuster, *J. Am. Chem. Soc.* **101** (1979) 3116.
233. B. G. Dixon, G. B. Schuster, *J. Am. Chem. Soc.* **103** (1981) 3068.
234. C. Rüchardt, *Fortschr. Chem. Forsch.* **6** (1966) 251.
235. T. Ochiai, *Bull. Chem. Soc. Jpn.* **49** (1976) 2522, 2641.
236. S. Oliveri-Vigh, F. T. Hainsworth, *J. Pharm. Sci.* **67** (1978) 1035.
237. D. J. Rawlinson, M. Konieczny, G. Sosnovsky, *Z. Naturforsch. B Anorg. Chem. Org. Chem. 34 B* (1979) 76, 80.
238. I. S. Voloshanovskii, S. S. Ivanchev, *Zh. Obshch. Khim.* **44** (1974) 892.
239. R. Criegee, *Ber. Dtsch. Chem. Ges. B* **77** (1944) 22.
240. R. D. Mair, A. J. Graupner, *Anal. Chem.* **36** (1964) 194.
241. P. D. Bartlett, J. L. Kice, *J. Am. Chem. Soc.* **75** (1953) 5591; Technisches Merkblatt "Peroxyester", Luperox GmbH, Günsburg 1976.
242. L. S. Silbert, D. Swern, *Anal. Chem.* **30** (1958) 385.
243. V. V. Shibanov, V. A. Fedorova, T. I. Yurzhenko, *Visn. L'viv. Politekh. Inst.* 1974 24.
244. W. Adam, A. Alzerreca, J. C. Liu, F. Yani, *J. Am. Chem. Soc.* **99** (1977) 5768.
245. U. Aeberhard et al., *Helv. Chim. Acta* **66** (1983) 2740.
246. Y. Hamada, A. Mizuno, T. Ohno, T. Shioiri, *Chem. Pharm. Bull.* **32** (1984) 3683.
247. Halcon Int., DE-OS 2 013 469, 1970.
248. Akzo, NL 852 261, 1985.
249. P. M. Kohn, *Chem. Eng.* **85** (1978) 16, 88.
250. Pennwalt Corp., DE 3 004 330, 1979.
251. Argus Chem. Corp., EP 20 846, 1979.
252. Nippon Oils & Fats, JP 51 076 213, 1975.
253. Pennwalt Corp., US 4 129 586, 1977.
254. Nippon Oils & Fats, JP 52 156 828, 1976.
255. Kenobel, EP 271 462, 1988.
256. Pennwalt Corp., DE 2 757 440, 1977.
257. Nippon Oils & Fats, JP 82 92 005, 1980.
258. P. D. Bartlett, R. R. Hiatt, *J. Am. Chem. Soc.* **80** (1958) 1398.
259. P. Lorenz, C. Rüchardt, E. Schacht, *Chem. Ber.* **104** (1971) 3429.
260. B. Maillard, A. Kharrat, C. Gardrat, *Tetrahedron* **40** (1984) 3531.
261. M. Schulz, J. Römbach, *Z. Chem.* **21** (1981) 404.
262. C. Rüchardt, V. Golzke, G. Range, *Chem. Ber.* **114** (1981) 2769.
263. P. E. Eaton, Y. S. Or, S. J. Branca, B. K. R. Shankar, *Tetrahedron* **42** (1986) 1621.
264. Standard Oil Co., US 4 220 746, 1978.
265. M. Shiozaki, N. Ishida, H. Maruyama, T. Hiraoka, *Heterocycles* **20** (1983) 279.
266. L. Thijs, S. N. Gupta, D. C. Neckers, *J. Org. Chem.* **44** (1979) 4123.
267. P. Gottschalk, D. C. Neckers, *J. Org. Chem.* **50** (1985) 3498.
268. D. C. Neckers, US 4 416 826, 1980.
269. C. Rüchardt, G. Hamprecht, *Chem. Ber.* **101** (1968) 3957.
270. G. A. Olah, D. G. Parker, N. Yoneda, *J. Org. Chem.* **42** (1977) 32.
271. Nippon Oils & Fats, EP 266 966, 1988.
272. Pennwalt Corp., US 4 075 236, 1978.
273. G. Maier et al., *J. Am. Chem. Soc.* **109** (1987) 5183.
274. V. A. Federova, V. A. Donchak, V. A. Puchin, Y. Y. Musabekov, *Khim. Tekhnol.* 1980, 23.
275. Pennwalt Corp., US 4 097 408, 1978.
276. H. Hock, H. Kropf, *Chem. Ber.* **88** (1955) 1544.
277. W. Duismann, C. Rüchardt, *Justus Liebigs Ann. Chem.* 1976, 1834.

278. M. R. Vilenskaya, G. A. Petrovskaya, Y. V. Panchenko, V. A. Puchin, *Zh. Org. Khim.* **20** (1984) 2112.
279. G. A. Razuvaev, T. N. Brevnova, N. F. Cherepennikova, S. K. Ratushnaya, *Zh. Obshch. Khim.* **46** (1976) 857.
280. D. E. Falvey, G. B. Schuster, *J. Am. Chem. Soc.* **108** (1986) 7419.
281. W. E. Cass, A. K. Bahl, *J. Org. Chem.* **39** (1974) 3602.
282. Nippon Oils & Fats, JP 75 007 055, 1966.
283. I. I. Artym, S. S. Ivanchev, M. M. Soltys, M. A. Kovbuz, *Zh. Prikl. Khim. (Leningrad)* **49** (1976) 2056.
284. G. S. Bylina, L. R. Uvarova, *Zh. Org. Khim.* **11** (1975) 2503.
285. V. A. Dodonov, V. V. Chesnokov, E. B. Eresov, *Zh. Obshch. Khim.* **45** (1975) 1524.
286. V. A. Dodonov, V. V. Chesnokov, T. A. Urchenko, *Zh. Obshch. Khim.* **46** (1976) 1293.
287. V. A. Dodonov, T. I. Starostina, *Zh. Obshch. Khim.* **47** (1977) 843.
288. G. A. Razuvaev et al., *Izv. Akad. Nauk SSSR Ser. Khim.* 1980, 821.
289. V. A. Dodonov, T. I. Starostina, *Zh. Obshch. Khim.* **47** (1977) 843.
290. S. Sustmann, C. Rüchardt, *Chem. Ber.* **108** (1975) 3043.
291. A. Kharrat, C. Gardrat, B. Maillard, *Bull. Soc. Chim. Belg.* **95** (1986) 535.
292. A. L. J. Beckwith, V. W. Bowry, G. Moad, *J. Org. Chem.* **53** (1988) 1632.
293. R. Hecht, C. Rüchardt, *Chem. Ber.* **96** (1963) 1281.
294. E. Montaudon, M. Campagnole, M. J. Bourgeois, B. Maillard, *Bull. Soc. Chim. Belg.* **91** (1982) 725.
295. F. D. Greene, W. Adam, G. A. Knudsen, *J. Org. Chem.* **31** (1966) 2087.
296. A. M. Kukova, L. V. Zhitina, V. L. Antonovskii, G. S. Kirichenko, *Khim. Promst. (Moscow)* 1980, 464.
297. Shell, US 2 608 570, 1952.
298. E. J. Pedersen, *J. Org. Chem.* **23** (1958) 252.
299. B. G. Dixon, G. B. Hunter, *J. Am. Chem. Soc.* **101** (1979) 3116.
300. Halcon Int., DE-OS 1 814 344, 1967.
301. Pittsburg Plate Glass, US 3 988 261, 1975.
302. I. Saito, T. Mano, R. Nagata, T. Matsuura, *Tetrahedron Lett.* **28** (1987) 1909.
303. G. Bouillon, C. Lick, K. Schank in S. Patai: *The Chemistry of Peroxides,* John Wiley & Sons, New York 1983.
304. P. S. Bailey, F. J. Garcia-Sharp, *J. Org. Chem.* **22** (1957) 1008; A. D. Jenkins, D. W. G. Style, *J. Chem. Soc.* 1953, 2357; A. I. Kirillov, *J. Org. Chem. USSR (Engl. Transl.)* **1** (1965) 1226, 1411; C. W. Cooper, *J. Chem. Soc.* 1951, 1340; C. S. Marvel, V. E. Nichols, *J. Org. Chem.* **6** (1941) 269; P. S. Bailey, *Chem. Rev.* **58** (1958) 925.
305. P. R. Story et al., *J. Am. Chem. Soc.* **90** (1968) 817; P. R. Story et al., *J. Org. Chem.* **35** (1970) 3059; P. Busch, P. R. Story, *Synthesis* 1970, 181; J. R. Sanderson, P. R. Story, *J. Org. Chem.* **39** (1974) 3463.
306. Research Corp., DE 1 807 337, 1968.
307. Montecatini, DE 1 178 055, 1958.
308. Montecatini, DE 1 216 282, 1959.
309. Distillers Comp., GB 740 747, 1965.
310. G. O. Schenck, H. D. Becker, *Angew. Chem.* **70** (1958) 504; A. Rieche, K. Meister, *Chem. Ber.* **72** (1939) 1938.
311. R. Criegee, G. Werner, *Justus Liebigs Ann. Chem.* **564** (1949) 9.
312. R. Criegee, G. Schröder, *Chem. Ber.* **93** (1960) 689.
313. R. Criegee, *Angew. Chem.* **87** (1975) 765; R. Criegee, *Chem. Unserer Zeit* **7** (1973) 75.
314. H. Rein, R. Criegee, *Angew. Chem.* **62** (1950) 120.
315. A. Rieche, E. Schmitz, E. Beyer, *Chem. Ber.* **91** (1958) 1935; A. Rieche, H. E. Seyfarth, A. Hesse, *Angew. Chem.* **78** (1966) 269; H. E. Seyfarth, A. Rieche, A. Hesse, *Chem. Ber.* **100** (1967) 624.
316. C. K. Ikeda, R. A. Braun, B. E. Sorenson, *J. Org. Chem.* **29** (1964) 286.
317. J. Wolpers, W. Ziegenbein, *Tetrahedron Lett.* 1971, 3889.
318. Distillers Comp., GB 740 747, 1965.
319. N. Brown, A. W. Anderson, C. E. Schweitzer, *J. Am. Chem. Soc.* **77** (1955) 1760; C. Dufraisse, S. Ecary, *C. R. Hebd. Seances Acad. Sci.* **223** (1946) 735.
320. G. O. Schenk, H. D. Becker, K.-H. Schulte-Elte, C. H. Krauch, *Chem. Ber.* **96** (1963) 509.
321. Shell, US 2 871 104 (1955).
322. *Eur. Chem. News,* Jan. 1969, 24.
323. F. H. Dickey, F. F. Rust, W. E. Vaughan, *J. Am. Chem. Soc.* **71** (1949) 1432; F. H. Dickey et al., *Ind. Eng. Chem.* **41** (1949) 1673.
324. Montecatini, DE 1 259 894, 1965; DE-AS 2 025 974, 1970.
325. B. Phillips, F. C. Frostick, P. S. Starcher, *J. Am. Chem. Soc.* **79** (1957) 5982; J. d'Ans, K. Dossow, J. Mattner, *Angew. Chem.* **66** (1954) 633.

326. M. S. Kharasch, G. Sosnovsky, *J. Org. Chem.* **23** (1958) 1322.
327. N. A. Milas, A. Golubovic, *J. Am. Chem. Soc.* **81** (1959) 3361, 5824, 6461.
328. T. Ledaal, *Acta Chem. Scand.* **21** (1967) 1656.
329. R. Criegee, W. Schorrenberg, J. Becke, *Justus Liebigs Ann. Chem.* **565** (1949) 7.
330. A. Rieche, C. Bischoff, *Chem. Ber.* **95** (1962) 77.
331. Pennwalt Corp., DE-OS 1 923 085, 1969.
332. Montecatini, DE-AS 1 293 771, 1966.
333. Research Corp., DE-OS 2 034 736, 1970.
334. Story Chem. Corp., DE-OS 213 316, 1971.
335. K. Wulz, H. A. Brune, *Tetrahedron* **27** (1971) 3669.
336. K. Griesbaum, M. Meister, *Chem. Ber.* **120** (1987) 1573.
337. Showa Denko K. K., JP 62 121747 A 2 and 87 121747, 1987.
338. JP-Kokai 63 22816 A 2 and 88 22816, 1988.
339. Pennwalt Corp., EP 273 090, 1988.
340. Q. T. Wiles et al., *Ind. Eng. Chem.* **41** (1949) 1679; Pennwalt Lucidol, product bulletin "Keton Peroxides" Philadephia, Pa.
341. Montecatini, GB 1 034 266, 1966.
342. Wallace & Tiernan, GB 1 047 830, 1966.
343. K. Gäbelein, *Umsch. Wiss. Tech.* **75** (1975) 478.
344. Emery, US 2 813 113, 1953.
345. A. Rieche, DE 565 158, 1931.
346. S. D. Rasumowskij et al., DE-OS 2 217 661, 1972.
347. E. G. E. Hawkins, *J. Chem. Soc. C* 1969, 2663, 2671, 2678, 2682, 2686.
348. R. Criegee, G. Lohaus, *Chem. Ber.* **84** (1951) 219.
349. BP Chemicals, DE-OS 1 643 640, 1967.
350. BP Chemicals, DE-OS 1 670 246, 1967.
351. BP Chemicals, DE-OS 1 768 387, 1968.
352. Degussa, DE-OS 2 003 269, 1970.
353. Degussa, DE-OS 2 004 440, 1970.
354. BP Chemicals, DE-OS 2 049 560, 1970.
355. BP Chemicals, DE-OS 2 109 923, 1971.
356. BP Chemicals, DE-AS 1 695 503, 1967.
357. BP Chemicals, DE-AS 2 116 500, 1971.
358. BP Chemicals, US 3 576 817, 1971.
359. E. G. E. Hawkins, *J. Chem. Soc.* 1969, 2663.
360. BP Chemicals, DE-OS 1 803 872, 1968.
361. BP Chemicals, DE-OS 1 814 288, 1968.
362. BP Chemicals, DE-OS 1 814 289, 1968.
363. Rhône-Poulenc, DE-OS 2 048 575, 1970.
364. *Jpn. Chem. Week,* June 1978, 1; *Chem. Age,* June 1978, 14; Ube, DE-AS 2 504 323, 1975.
365. M. V. Pokhmurskaya, V. N. Nosan, N. I. Klim, *Vysokomol. Soedin. Ser. B* **30** (1988) 781.
366. Atochem, EP 241 390 A 2, 1987.
367. J. L. Vennerstrom, *J. Med. Chem.* **32** (1989) 64.
368. Gesetz über explosionsgefährliche Stoffe (Sprengstoffgesetz) from Sept. 13, 1976; Bundesgesetzblatt I (1976), pp. 2737 – 2787; 1st – 5th Verordnung zum Sprengstoffgesetz.
369. Organische Peroxide, Merkblatt M 001 4/80 der Berufsgenossenschaft der chemischen Industrie, Jedermann-Verlag, Heidelberg 1989.
370. Unfallverhütungsvorschrift "Organische Peroxide" der Berufsgenossenschaft der chemischen Industrie, Heidelberg (Entwurf 1989).
371. OECD-IGUS, Organisation for Economic Cooperation and Development, International Group of Experts on the Explosion Risks of Unstable Substances.
372. R. Siegmeier, unpublished results.
373. K. Eckwert, E. Peukert, L. Jeromin, *Fat Sci. Technol.* **90** (1988) 511.
374. Gefahrstoffverordnung of Aug. 1986, Bundesgesetzblatt I, p. 1470.
375. 8th Recommendations on the Transport of Dangerous Goods, 6th rev. ed., United Nations, New York 1988.
376. Berufsgenossenschaft Chemie, Merkblatt "Organische Peroxide", Verlag Chemie, Weinheim 1980.
377. H. J. Küchle, *Zentralbl. Arbeitsmed. Arbeitsschutz* **8** (1958) 25.
378. Register of Safety Information of Chemical Products, National Board of Labour Protection, Tampere, Finland, Nov. 1981.
379. F. A. Bahmer, A. Schulze-Dirks, H. Zaun, *Dermatosen Beruf Umwelt* **32** (1984) 21.
380. H. J. Bandmann, M. Agathos, *Hautarzt* **36** (1985) 670.
381. L. Aringer: *Arbetarskyddsstyrelsen,* Publikationssevice 17 184, Solna, Sweden 1985.
382. R. E. Gosselin, H. C. Hodge, R. P. Smith, M. N. Gleason: *Clinical Toxicology of Commercial Products,* The Williams & Wilkins Co., Baltimore 1976.
383. Degussa, unpublished results.
384. E. P. Floyd, M. S. Stokinger, H. E. Stockinger, *Ann. Ind. Hygiene Assoc. J.* **19** (1958) 205.
385. Degussa, unpublished results.
386. F. I. Orlova, *Sb. Nauchn. Tr. Kuibyshev. Nauchno Issled. Inst. Gig* **6** (1971) 101.
387. M. Sharratt, A. C. Frazer, O. C. Forbes, *Food Cosmet. Toxicol.* **2** (1964) 527.
388. C. J. Molloy, M. A. Gallo, J. D. Laskin, *J. Soc. Cosmet. Chem.* **35** (1984) 197.

389. J. Schweizer, H. Loehrke, L. Edler, K. Goerttler, *Carcinogenesis (London)* **8** (1987) 479.
390. L. W. Updyke, A. Chuthaputti, R. W. Pheifer, G. K W. Yim, *Carcinogenesis (London)* **9** (1988) 1943.
391. A. J. P. Klein-Szanto, T. J. Slaga, *J. Invest. Dermatol.* **79** (1982) 30.
392. F. G. Bock, H. K. Myers, H. W. Fox, *J. Nat. Cancer Inst. (U.S.)* **55** (1975) 1359.
393. T. Yamaguchi, Y. Yamashita, *Agric. Biol. Chem.* **44** (1980) 1675.
394. M. K. Logani, C. P. Sambuco, P. D. Forbes, R. E. Davies, *Food Chem. Toxic.* **22** (1984) 879.
395. O. H. Iversen, *J. Invest. Dermatol.* **86** (1986) 442.
396. J. H. Epstein, *J. Invest. Dermatol.* **91** (1988) 114.
397. T. E. Donnelly, Jr., J. C. Pelling, C. L. Anderson, D. Dalbey, *Carcinogenesis (London)* **8** (1987) 1871.
398. M. Ando, A. L. Tappel, *Toxicol. Appl. Pharmacol.* **81** (1985) 517.
399. M. Ando, *Nippon Eiseigaku Zasshi* **41** (1986) 587.
400. D. Galaris, E. Cadenas, P. Hochstein, *Free Radicals Biol. Med.* **6** (1989) 473.
401. J. C. Pelling et al., *Carcinogenesis (London)* **8** (1987) 1871.
402. M. Ishadate, Jr., T. Sofuni, K. Yoshikawa, *Hen'Igen to Dokusei* **3** (1980) 80.
403. *Bibra P. Watts Bull.* **20** (1981) 514.
404. M. Chevallier, D. Luzzatti, *C. R. Hebd. Séances Acad. Sci.* **250** (1960) 1572.
405. F. H. Sobels, *Dros Info Serv.* **32** (1958) 159.
406. H. L. Gensler, G. T. Bowden, *Carcinogenesis (London)* **4** (1983) 1507.
407. H. C. Birnboim, *Carcinogenesis (London)* **7** (1986) 1511.
408. Phillips Petroleum Co., TSCA-Mitteilung, Sect. 8 e (1985).
409. G. B. Freese et al., *Mutat. Res.* **4** (1967) 517.
410. K. Heindorff, A. Michaelis, O. Aurich, R. Rieger, *Mutat. Res.* **142** (1985) 23.
411. A. Korhonen, K. Hemminki, H. Vainio, *Environ. Res.* **33** (1984) 54.

Pest Control → Acaricides
Pest Control → Crop Protection
Pest Control → Insect Control
Pest Control → Molluscicides
Pest Control → Nematicides
Pest Control → Rodenticides
Pet Food → Foods, 1. Survey
Petrol → Automotive Fuels
Petroleum → Oil Refining

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Petroleum Coke

Standard Article

Heinrich Predel[1]

[1]MiRO GmbH & Co. KG, Karlsruhe, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a19_235.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Physical and Chemical Properties
3. Production
3.1. Production Processes
3.1.1. Delayed Coking
3.1.2. Fluid Coking
3.1.3. Flexicoking
3.2. Calcination
3.2.1. Rotary Kiln Calciner
3.2.2. Rotary Hearth Calciner
4. Uses
4.1. Green Petroleum Coke
4.2. Calcined Petroleum Coke
5. Quality Aspects
5.1. Green Coke
5.2. Regular Calcinate
5.3. Needle Coke
6. Environmental and Safety Aspects
6.1. Green Coke
6.2. Calcined Petroleum Coke

The petroleum coke production started about 80 years ago and will reach the 100 × 10^6-t/a level in the next years. The yearly increase rates are up to 10%.

Characteristics, localization, and production methods including the Delayed Coker process of petroleum coke are shown. The world market profile for petroleum coke is presented, and the amount and prices of the petroleum coke species, “green coke”, “regular calcinate”, and “needle coke” are given. Particular interesting is the production process of green coke, which is increasingly used in power generation, and the production of metallurgical coke for steel processes (blast furnace). The use of regular calcinate for the carbon anode in the aluminium production stagnates because of the impetuous development of the aluminium recycling. Likewise, the needle coke has only marginal increase for the production of graphite electrodes used in electric-arc furnaces by steel processes. The reason is the rapidly reduced consumption of electrodes/t steel due to process modifications.

The environmental and safety aspects are addressed and the available results are shown.

[Top of Page]

1. Introduction

Petroleum Coke is a grayish-black solid residue with a high carbon content. It is obtained during the thermal conversion process in crude oil processing. Although petroleum coke has been known for more than 100 years, both production and its specific application and improvement have seen an increasing boom only during the last 80 years. It is anticipated that

up to the year 2010 ~100 × 10^6 t/a of petroleum coke will be produced in oil refineries. The majority of this quantity is obtained in delayed coking plants. In other facilities for processing crude oil residue components, investments are made at present in flexicoking units, whereas fluid cokers are still run but hardly set up in new production.

[Top of Page]

2. Physical and Chemical Properties

Physical Properties. Petroleum coke is a grayish-black solid matter of very varied composition. Generally, a distinction is made between petroleum coke from fluid and flexicoking plants and that obtained from delayed coking processes. The petroleum coke from delayed cokers is also called *green coke*. The fluid and flexicokers supply a fine-grained petroleum coke that has a high level of isotropy and in general an ash content of between 1 and 4 wt %. The grain consistency of these cokes is high, the content of volatile combustible matter (VCM) is low and hardly exceeds 5 wt %. In contrast petroleum coke from delayed cokers is normally characterized by both high porosity and anisotropy. The VCM content is between 7 and 14 wt %; it is only set to < 5 wt % for needle coke production. The ash content is low, generally < 0.1 wt %. The bulk density of normal-quality petroleum coke is around 0.8 – 0.9 g/cm^3. If however, the feed stream to the delayed coker is rich in high molecular asphalt, a non-porous isotropic petroleum coke is the result which has a high bulk density of 1.2 – 1.4 g/cm^3. The formation of this *shot coke* quality can be counteracted by machine adjustment.

Another basic form of petroleum coke is *calcined coke*. Calcined coke is a product derived from green coke. Here, the hydrocarbons have been removed by heating green coke under reducing conditions in kilns or hearths to over 1300 °C. Classification of petroleum coke types derived from delayed cokers into:

- Shot coke
- Sponge coke (fuel coke)
- Sponge/honeycomb coke (regular or anode-grade coke)
- Needle coke (graphite or premium coke)

is not defined precisely [1]. A differentiation with regard to isotropy and crystal structure is logical. There is a graded transition from:

> Shot coke (mainly isotropic; amorphous, no crystal structures formed; hardly any pores)

to

> needle coke (high anisotropy; regular crystal structures, containing a great number of fine pores, crystallite size of the order of 4 – 7 nm).

Petroleum coke results from a thermal cracking and polymerization reaction. After going through the mesophase a solid crystalline form is formed, which, contrary to indications [2] does not plastify again. Coking only occurs in the VCM particles contained. If their fraction exceeds 15 %, the release of this petroleum coke at high temperatures gives the impression of a plastic state of aggregation. By treating petroleum coke with steam and hot water, those components that are slightly water soluble can be washed and steamed out of the petroleum coke.

Chemical Properties and Composition. Under normal conditions petroleum coke has a low reactivity. At higher temperature (up to 1400 °C, calcination conditions) an expansion of the crystallite size up to 2.5 – 3.5 nm takes place because of the cracking of side chains. At the same time the original hydrogen content of approx. 3.5 wt % is reduced to below 0.1 wt %. Under graphitization conditions (ca. 2900 °C) good crystallized forms of petroleum coke (needle coke character) are transformed into graphites. These graded changes in quality of petroleum coke with temperature demonstrate most suitable the structure, character, and classification of the product.

The dominant element in petroleum coke is carbon. The hydrogen content is around 3.5 wt %. An overproportional amount of sulfur and nitrogen reaches the petroleum coke from the feed components (S content of petroleum coke 0.1 – 5 wt %; N content 0.2 – 1.5 wt %). The V and Ni contents of the feed streams remain entirely in the petroleum coke. The Na, Ca, and Mg content in the petroleum coke is the sum of the content of these cations in the crude oil and of their concentration in the cutting water (Section Green Coke). According to the 21st ATP (Adaptation of Technical Progress of the European Substance Directive) and CONCAWE, product dossier no. 93/105 [3], petroleum coke is considered a nondangerous product. The level of polycyclic aromatics in petroleum coke is about 250 mg/kg. Benzo[*a*]pyrene makes up about 25 mg/kg. The amount of heavy metals apart from V and Ni is insignificant (< 1 mg/kg). In general the quality of the feed components and the production conditions determine the quality of the product.

[Top of Page]

3. Production

In 2004, about 60 × 10^6 t of petroleum coke were produced worldwide, which are available in the market. This is about two-thirds of the production rate. One third will be revalorized by the producer [4]. The market-available petroleum coke is about 10 % of the quantity of blast-furnace coke (hard coke) produced from hard coal [5]. More than 95 % of the petroleum coke is produced in delayed cokers, about a third of the feed streams is incurred as petroleum coke. Residues and crack components (see Section Delayed Coking) are used in delayed cokers in more than 65 refineries. Due to the reaction conditions only about 5 – 10 wt % of the feed material is converted into petroleum coke in case of fluid and flexicokers. The number of these production units is much lower. In 1996 13 fluid cokers and 5 flexicokers were in production worldwide. Most petroleum coke is produced in North America, followed by South America and Western Europe. Anode grade is the most important product with regard to margin and value. Table 1 and 2 list the worldwide locations for

petroleum coke and projects in construction [4].

Table 1. Refineries with Petroleum Coke Production

Refinery	Estimated Production Mta 2005	Typical Analysis %S	Current FOB Price HGI	US $
NORTH AMERICA				
East Coast				
Motiva - Delaware City*	0.70	5.6	34	15-18
Valero Energy - Paulsboro	0.50	5.0	100	No export
West Coast				
BP - Watson*	0.30	6.5-7.0	50	35-
BP - Cherry Point	1.00	3.5	45	anode grade
Chevron - El Segundo	1.10-1.20	4.0-4.5	65-70	49-51
Bakersfield*	0.25	0.9	30	
Shell - Martinez	0.42-0.45	1.1-1.5	40	50-55
Shell - Wilmington	0.630	2.6	53-55	55-
ExxonMobil Torrence	1.10	1.1-1.5	50	60-
ConocoPhillips-Santa Maria*	0.35	4.8	65	-
ConocoPhillips-Carson*		2.6-3.0	40-42	48-50
ConocoPhiBips-Rodeo*	0.40	0.9	85	anode grade
Tesoro-Golden Eagle	0.43	2.5	28	37-40
Valero- Wilmington	0.60	4.2-4.5	60-65	49-51
Valero Benicia	0.40	3.5	30-35	30-
USGulf				
BP - Texas City*	0.90	5.0-5.5	80	28-29
Chevron - Pascagoula*	2.50	6.3	35	25-26
CITGO - Corpus Christi	0.90	3.9-4.1	46-47	29-30
CITGO - Lake Charles	1.40	3.7	60-65	29-30
CITGO - Lyondell	2.00	4.0-4.3	45-48	29-30
ConocoPhillips - Lake Charles	1.30	6.0-6.2	40-42	26-27
Valero - Corpus Christi	0.40	5.8-6.2	77-80	25-26
ExxonMobil - Baton Rouge	1.30	4.5-5.0	50	28-29
				Dom 15-20
ExxonMobil - Chalmette	0.65	4.5	55-60	29-30
				Dom 15-20
ExxonMobil - Beaumont	1.10	5.8	65-67	27-28
ExxonMobil - Baytown	1.00	6.2-6.3	47-50	24-25
Koch Petroleum - Corpus Christi	0.40	3.0-4.0	90	40-
Marathon - Garyville	0.90	5.8	42-44	24-25
Motiva - Port Arthur*	0.90-1.0	6.7-6.9	57-62	26-27
Valero - St Charles*	1.40	4.0-4.5	42-	28-29
Valero -Texas City	1.00	6.5	40	24-25
ConocoPhillips -Sweenty	1.40	4.2	44-45	29-30
Premcor USA - Port Arthur	1.40	5.8-6.3	38-39	25-26
Shell - Deer Park*	2.00	5.8	42-50	25-26
Inland				
BP - Toledo (US)*	0.55	4.5	30-35	Dom 15-
BP - Whiting	0.55	5.0-5.5	80	Doml5-

CITGO - Lemont	0.40	4.0-5.0	80	
ConocoPhillips - Billings*	0.35	5.7	38-45	
El Dorado'	0.40	2.1	50	
ExxonMobil - Billings*	0.20	6.0	30	
ExxonMobil - Joilet	1.10	5.0-5.5	50	Dom 15-
Koch Petroleum - Pine Bend	1.80	5.1	50	28-30
ConocoPhillips - Wood River	0.35	5.0	35-40	Dom 15-
CENTRAL AMERICA				
Pemex-Refinación Madeira	0.80	7.0	40	
Pemex-Refinación Cadereyta J.	1.00	8.0	40	
SOUTH AMERICA				
El Paso - Aruba*	1.50	5.0-5.5	48-54	27-28
PDVSA - Lagoven	0.65	4.3	60-65	29-30
PDVSA - Maraven	1.20	4.2	38-42	28-29
Petrozuata	1.20	4.1	70-75	29-30
Sincor	2.30	4.2-4.6	52-60	29-30
PDVSA Cerro Negro	0.85	4.2	65-70	29-30
HOVENSA	1.20	4.0	37-39	27-28
Hamaca	1.30	4.0-4.3	60-65	29-30
ASIA & EUROPE				
Agip - Gela*	0.70	4.5	60	
Midor	0.30	5.5	55	50-
Malaysia Refining Co Sdn Bhd*	0.35	5.0	45	40-
MiRO - Karlsruhe	0.29	3.0	60	50-
MOL*	0.20	3.5	70	
Reliance Oil*	2.50	6.1-6.3	43-50	60-
Repsol - La Coruña	0.45	4.8	80	45-
Repsol - Puertollano*	0.45	4.5	55	*
BP - Gelsenkirchen	0.26	3.0	60	50-
BP - Lingen	0.15	1.5	65	60-
OMV – Burghausen	0.15	1.5	65	60-
Coroco – Immingham	2.1			(calcination)

*S = Sulfur; HGI = Hardgrove; * = No availability; Dom= Domestic

Table 2. Delayed Coker projects in construction

Company Name	Coker Name	Location	Expected Petcoke Production million t/a	Start-up Q/year	Expected Quality
Husky Oil Operations		Canada	0.2	I/2005	4%S/60HGI
Premcor	Port Arthur	Port Arthur	1.0	2005	6.3% S/ 35 HGI
Indian Oil Corporation		Domestic	0.7	I/2005	6.5% S/35-40 HGI
Repsol	Cartagena	Cartagena, Spain			uncertain
Petrozuata 1)	Petrozuata	Jose, Venezuela	1.2	2006	less 48 HGI Uncertain
Valero	Aruba	Aruba	1.5	2007	
Pemex-Refinación	Gral Lázaro Cárdenas	Minatitlán, Mexico	1.2	III-IV/2007	

Pemex-Refinación	Antonio Dovali Jaime	Salina Cruz, Mexico	1.0	2010

*S = Sulfur; HGI = Hardgrove; Mta = million tonnes per annum; Q = Quarter; *No availability
1) Joint venture between Conoco and Maraven, a subsidiary of PDVSA

3.1. Production Processes

3.1.1. *Delayed Coking*

Coking is an effective way for an integrated refinery to process the residues of crude oil. Coking processes are normally carried out to minimize coke and maximize liquid products. The first atmospheric coking processes were started in 1928. Coking, in general, has become more important over the years because tradtional heavy fuel oil demand has declined and the demand for lighter products has increased. This has made it more attractive to convert the heavy fuel oil components into better marketable light, clean components. The present major source of production of marketable coke is the delayed coking process developed by Standard Oil. Conventional delayed coking is shown in the flow diagram in Figure 1. A delayed coke complex is depicted in Figure 2.

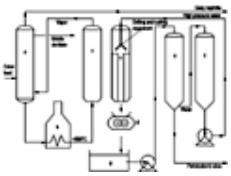

Figure 1. Flow sheet of delayed coking

a) Fractionator; b) Furnace; c) Coke drum; d) Crusher; e) Coke dewatering bin; f) Water tank; g) Coke – water pit

Figure 2. Photograph of a refining delayed coker complex: furnace, coke drums, fractionator, and dewatering bins

The refinery residues are heated in a furnace (b) up to about 485 – 500 °C. After passing a fractionator tower where middle distillates are topped off, they are fed into one of the coking chambers. The energy gained in the furnace passages is sufficient to release the cracking reaction in the filling time of the coking chamber (delayed coking). The reaction zone is also called mesophase, because the transition from liquid residue molecules to solid crystalline coke structure occurs here. The pressure in the coking chamber can be varied between 2 and 6 bar. In the furnace, steam is dosed to the residue to inhibe the formation of coke. The vapors in the coking chamber — gases, naphtha, middle distillates, and a heavy coker gas oil — are led to the fractionator for separation, fractionated and forwarded for stabilizing post-treatment. The heavy coker gas oil is recycled as coker feed or is used in other refinery processes: e.g., after go-fining (a hydrogenation process) as feed for fluid catalytic cracking [6].

While one coking chamber is being filled in the cycle time, the other chamber is emptied. A typical volume of a modern coke drum is about 1000 m^3 with a range in diameter of 5 to 8 m and in height of 25 to 35 m. After steaming and cooling of the coke chambers, the coke is removed by drilling and cutting with high pressure water (up to 300 bar pressure). The green coke obtained is conveyed to its application after the draining step: either as fuel of feed for calcining. Table 3 shows the different green coke qualities in relation to use.

Table 3. Green coke qualities in relation to use

	Heating medium	Regular calcinate feed	Needle coke feed
Sulfur content, wt %	> 2	1.0 – 3.0	0.1 – 0.8
V/Ni content, mg/kg	200 – 500	50 – 200	< 50
Ca/Na content, mg/kg	50 – 300	20 – 100	< 50
VCM, wt %	9 – 14	7 – 9	5 – 7

The following components can be used as feed for delayed cokers [7]:

1. Crude residues (atmospheric or vacuum),
2. Crack components (visbroken tar, cycle oil, decant oil, or thermal crack tar),
3. Deasphalting residues,
4. Coal oils, and
5. Used plastic materials (recycling)

The products of delayed coking processes are (basis: vacuum residue or crack components):

1. Gas/LPG (ca. 13 wt %)
2. Naphtha (ca. 11 wt %)
3. Middle distillate (ca. 45 wt %)
4. “Green coke” petroleum coke (ca. 31 wt %)

Delayed coking is carried out in order to free refinery from the product “heavy fuel oil” which is not easily marketable. In terms of investment delayed coking is a much cheaper technology than flexicoking or hydrocracking (factor 3 to 5). A further argument in favor of delayed coking can be the revalorization potential for the carbon products “regular calcinate” or “needle coke” produced. Moreover, delayed coking shows good product results, which, however, like all thermal conversions, have “middle distillates” as main product.

Figure 3 shows the product yields for the North Sea crude “Brent”. All thermal conversion processes reduce the crude residues higher than middle distillates, especially the delayed coker unit.

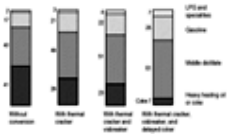

Figure 3. Product yields with thermal conversion processes

Goal of nearly all delayed coker units is to minimize coke yield and obtain as many distillate fractions as possible. However, when the objective is to produce green coke that shall serve as raw material for higher quality coke products (regular calcinate and needle coke) a specific feed optimum must be found that by no means results in a minimum coke yield.

The distillate yields and thus the capacities of delayed coking can be increased by [7], [8]:

1. Preconversion of feed components [6]
2. Addition of light middle distillate to the feed stream
3. High coke drum outlet temperature / high coker furnace outlet temperature (COT)
4. Low operating pressure
5. Low fresh feed: recycle ratio
6. Addition of coil steam

Some quantitative effects are shown in Figure 4.

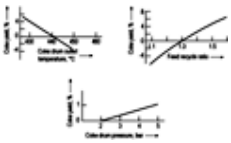

Figure 4. Relationship between plant operating conditions and plant yield; parameters: coke drum pressure = 2.5 bar; feed/recycle ratio = 1 : 1.3; coke drum outlet temperature = 440 °C [9]

The following technical improvements have been achieved for the production of petroleum coke in modern delayed coking plants. They result in increased plant reliability, improved work safety and better coke quality as well as increased production capacity [10-13]:

1. Onstream decoking (allows decoking of one of the four furnace coils during current production)
2. Homogeneous petroleum coke quality (requires temperature and recycle ratio ramps over the cycle time to adjust continuous coking conditions)
3. Optimum coke chamber filling (requires mass balance of feed and products and measurement of the coke chamber level)
4. Batch computerization (allows automatic checking and starting of operation steps without losing time)
5. Automatic coke cutting with vibration alarms at coke drum wall
6. Automatic hydraulic coke drum bottom head
7. Combination tool for drilling and cutting
8. Hydraulic feed line
9. Hydraulic coke drum bottom head closing

With these technical measures the cycle time for delayed coking can be reduced from about 24 to 12 h [14-17]. The time required for different steps in typical 12-h and 24-h coking cycles is listed in Table 4.

Table 4. Time allotted for different steps in typical 12-h and 24-h coking cycles

Operation step	Time, h	
	24-h cycle	12-h cycle
Switch drum	0.5	a

Steam out to fractionator	0.5	a
Steam out to blowdown	1.0	
Slow water cooling	1.0	b
Full water rate	5.0	b
Drain water and remove top head	3.5	1.5
Drop bottom head and feed line	0.5	0.5
Hydraulic decoking	4.0	2.0
Replace heads and feed line	1.0	0.5
Steam purge and pressure test	1.0	0.75
Drum heat up	4.0	2.25
Slippage allowance	2.0	1.0
Total	24 h	12 h

[a] Total time for both operation steps: 0.5.
[b] Total time for both operation steps: 3.0.

3.1.2. *Fluid Coking*

The continuous fluid coking process utilizes fluidized solids technique developed for fluid catalytic cracking (⟶ Oil Refining), except that no catalyst is used. A flow scheme is shown in Figure 5.

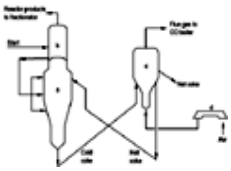

Figure 5. Flow sheet of fluid coking

a) Reactor; b) Scrubber; c) Burner; d) Air blower

Residue is feed to the reactor (a). The cracking reactions occur at 500 – 550 °C in a fluidized bed of coke particles. The coke formed in the reactor flows continuously to the heater, where it is heated up to 600 – 650 °C by partial combustion in a second fluidized bed. Part of the heated coke particles are returned to the reactor to supply the energy to the endothermic crack reactions and to maintain the reactor temperature. The remainder of the coke after cooling is removed as a stream of fine particle “petroleum coke”. Fluid coke is only occasionally used for production of anodes [18].

3.1.3. *Flexicoking*

The flexicoking process combines fluid coking with gasification. A flow scheme of flexicoking is shown in Figure 6.

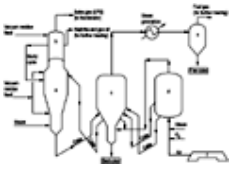

Figure 6. Flow sheet of flexicoking

a) Reactor; b) Scrubber; c) Heater; d) Gasifier; e) Cyclones; f) Airblower

Flexicoking is a residue conversion process that combines coke gasification with conventional fluid coking. Three fluidized beds are in operation at the reactor, heater, and gasifier. Shrinking markets for the coke obtained in substantial amounts as byproduct of delayed coking and fluid coking due to restrictions of sulfur emission encouraged the development of the flexicoking process. In flexicoking most of the coke produced is converted to low-Btu fuel gas in the gasification reactor. This fuel gas can be desulfurized to approximately the equivalent of a 0.25 wt % sulfur fuel oil. Thus, particularly for feed stocks with a high metal and sulfur content, flexicoking can be an advantage, because it produces marketable liquid products plus fuel gas.

Thermal conversion of the feed takes place in the reactor (a) where the feed is atomized through a series of nozzles into a bed of fluidized coke particles. Coke occurs in a thin film on the fluidized particles within the bed. The coke bed is fluidized by the product vapors and by stripping steam injected at the bottom of the reactor. Typical bed temperatures are 510 – 540 °C. Vapor products from the conversion reactions in the bed go overhead through cyclone separators, which remove most of the entrained coke and return it to the reactor bed. The cyclone outlets discharge the vapor product directly into a scrubber (b) where heavy liquid is used to scrub out the remaining coke dust and condense unconverted high boiling fractions. The dust laden liquid is recycled as “slurry cycle” to the reactor with the feed. The scrubbed vapor is sent to the coker fractionator, where the stream is split into gas, naphtha, distillate and heavy gas oil streams. Coke produced in the reactor is gasified by steam and air in the gasifier (d), which operates at a bed temperature of approximately 955 °C. The heater (c) serves as a surge vessel and transfers heat from the gasifier overhead gas to coke which in turn supplies the heat of reaction in the reactor. The heater bed temperature is approximately 610 °C. Coke is continuously circulated between the three vessels to transfer heat and maintain vessel inventories. Heat for the endothermic cracking reactions in the reactor is supplied by coke combustion in the gasifier. In addition to the coke fines recovered from the heater overhead gas, product coke can be withdrawn from the heater bed. All liquid or gas streams require further treating [19].

3.2. Calcination

About 25 % of the petroleum coke derived from delayed coking processes is stabilized by calcination. Calcination includes

the elimination, burning, and coking of water and higher boiling hydrocarbons contained in petroleum coke. Finally a carbon product remains that contains sulfur and nitrogen plus small amounts of vanadium, nickel, sodium, calcium, iron, and silicon.

At present calcining is carried out at about 40 calciner sites (see Table 5). While the calciners in North America are usually run by independent companies or attributed to aluminum works, the calciners in other regions are production units of the refineries.

Table 5. Petroleum coke calciners [19]

Owner	Location	State	Type	Production capacity, 10^3 t/a
North America				
Alcan	Arvida	Quebec	kiln	215
Alcan	Kitimat	British Columbia	kiln	80
Alcan/Intalco	Edmonton	Alberta	kiln	165
Arco	Cherry Point	Washington	hearth	675
Arco	Los Angeles	California	kiln	315
Calciner Industries	Chalmette	Louisiana	kiln	200
Calciner Industries	Gramercy	Louisiana	kiln	230
Calciner Industries	Norco	Louisiana	kiln	170
Calciner Industries	Purvis	Mississippi	kiln	80
Conoco	Lake Charles	Louisiana	kiln	260
Great Lakes Carbon	Enid	Oklahoma	kiln	470
Great Lakes Carbon	Port Arthur	Texas	kiln	635
Kaiser Aluminium	Mead	Washington	kiln	80
Landmark Petroleum	Fruita	Colorado	kiln	240
Reynolds Metals	Baton Rouge	Louisiana	kiln	620
Reynolds Metals	Lake Charles	Louisiana	kiln	200
Union Chemicals	Rodeo	California	kiln	295
Venco	Lake Charles	Louisiana	kiln	400
Venco	Moundsville	West Virginia	kiln	400
South America				
Copetro	La Plata	Argentina	kiln	200
Petrocoque	Cubatao	Brazil	kiln	400
Africa				
Egyptalum	Nag Hamadi	Egypt	hearth	135
Asia/Oceania				
Government	unknown	China	kiln	455
KOA Oil	Marifu	Japan	kiln	115
Petrocokes	Mizushima	Japan	kiln	80
Various	various	India	kiln	240
Petramina	Dumai	Indonesia	kiln	270
Europe				
Conoco	Immingham	United Kingdom	kiln	1500
OMV	Burghausen	Germany	hearth	200
Esso AG	Karlsruhe	Germany	hearth	130
Government	Sisak	Yugoslavia	kiln	50
Government	various	CIS	hearth/kiln	2015
Rafinor	Mongstad	Norway	hearth	170
Repsol	La Coruna	Spain	hearth	180
BP	Gelsenkirchen	Germany	kiln	200
BP	Lingen	Germany	kiln	130

Calcining is carried out in rotary kiln and rotary hearth plants. While the rotary kiln largely predominates in America, equal amounts of kiln and hearth plants are used in Europe.

In both types of calcining complexes the water and VCM contents of green coke are eliminated by heating up to 1400 °C. The heat required for calcining the green coke is primarily provided by burning the VCM released from the coke calcination process. Combustion air is admitted to the calciner. In normal operation only the quantity of air sufficient for complete combustion of the VCM release from the coke is added, no free oxygen occurs in the off-gas.

3.2.1. *Rotary Kiln Calciner*

The rotary kiln calciner (Fig. 7) consists of a rotating heated steel tube (e) with a refractory lining. Because of the gradient and the rotation of the tube, the green coke descends in the calciner and heat flow rises in countercurrent. The calcined coke leaving the kiln is discharged into a rotary cooler (f). The flue gases are cooled in a waste-heat boiler (i) and cleaned in an electrostatic precipitator (j).

Figure 7. Schematic of the rotary kiln calciner

a) Screen; b) Crusher; c) Weigh feeder; d) Green coke sampling; e) Rotary kiln; f) Cooler; g) Sampling calcined coke; h) Incinerator; k) Stack

3.2.2. *Rotary Hearth Calciner*

The rotary hearth calciner (Fig. 8) was originally developed by the Wise Coal and Coke company. Green coke charged at the periphery of the hearth travels around the calciner (b) in a series of concentric paths. At the completion of each revolution the coke is moved by a rabble arm set in the roof into the next concentric path. This process is repeated, until at the final rabble, the coke is discharged into the soaking pit. From the soaking pit the coke can be discharged into a rotary cooler (c). During normal operation the retention time for the coke in the hearth and soaking pit is approximately 2 h [20].

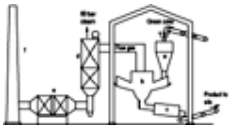

Figure 8. Schematic of the rotary hearth calciner

a) Feed bin; b) Rotary hearth; c) Calcinate cooler; d) Waste-heat boiler; e) Electrostatic precipitator; f) Stack

[Top of Page]

4. Uses

Worldwide annual consumption of industrial carbon is about 25×10^6 t/a, excluding metallurgical coke (see Table 6). Two-thirds of this margin is calcinate from petroleum coke [21], [22].

Table 6. Production of industrial carbon (World 1994)

Carbon Products	Production, t/a	Average product value, $/t	Market value, 10^9 $/a
From petroleum coke			
Carbon anodes	9×10^6	700	6.3
Söderberg anodes	4×10^6	500	2.0
Graphit electrodes	1.1×10^6	5500	6.05
Reduction medium/TiO2-Process	2×10^6	600	1.2
Partly from petroleum coke			
Carbon cathodes	6×0^5	1300	0.78
From other products			
Carbon black	8×10^6	1500	12
Activated carbons	5×10^5	3000	1.5
Carbon construction materials, etc.	1×10^5	800	0.8
Electrographite powder	8×10^4	1500	0.12
Carbon fibers	6×10^4	7×10^4	4.2
Synthetic diamonds	50	15×10^6	0.75
Total	~ 24×10^6		~ 36

Figure 9 shows the world market profile for petroleum coke in 1999 [23]. The relations are the same as now: More than 70 % of petroleum coke is burned as fuel.

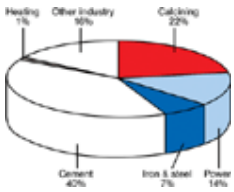

Figure 9. World market profil for petroleum coke

4.1. Green Petroleum Coke

More than 70 % of the petroleum coke produced is marketed as green coke. This green coke is a cheap fuel used primarily in the cement industry or for electricity generation. Qualities with a sulfur content below 1.5 % are used as a supplementary substance for hard coke production, if the coal has too high a VCM content. In addition, this green coke is used in the production of carbides, too.

There is increasing use of petroleum coke as addition for metallurgical coke [24] primarily used in blast ovens for producing iron.

Green coke is sold worldwide. The biggest trans-shipment center with some million tons of storage capacity is Rotterdam.

4.2. Calcined Petroleum Coke

Anode-Grade Coke (Regular Calcinate). Most calcined petroleum coke is used for production of carbon and Söderberg anodes (Table 6). Moreover, as aluminum production with Söderberg anodes is declining due to environmental aspects, the requirement will shift more to carbon anodes. The coke quality required is called anode grade coke or regular calcinate [21].

Worldwide aluminum production was 22.6×10^6 t in 2004; 60 % of the aluminum production capacities are located in North America and Europe. Aluminum production is expected to increase about 9 % per year. With a consumption of about 430 kg regular calcinate per tonne of aluminum the worldwide regular calcinate demand in the aluminum industry is more than 10×10^6 t/a. The recycling rate for aluminum is increased to about 30 % [25], [24]. Therefore, the world aluminum consumption is higher than 30×10^6 t/a.

The Alcoa Test (Alcoa = biggest Al producer in the world) with “inert” anodes is currently running. Inert anode means: Reduction without carbon. The success of this test will be sceptically criticized because of process economics and Al purity [24].

Regular calcinate in the fraction 0.2 – 1.0 mm is used for titanium dioxide production. More than 70 % of the worldwide TiO_2 production of 4×10^6 t/a is produced using the Chloride Process. This process needs about a third of that using rutil as reduction medium.

Needle Coke. Petroleum coke is also used for the production of graphite electrodes (Table 6). The quantity employed for this purpose has been stagnating for years. This high-grade quality is called needle coke. Although the proportion of electrosteel production in overall steel production is steadily increasing, a series of optimizing steps have continuously reduced the electrode consumption per tonne of steel. Whereas in 1985, 7 kg electrode material was needed per tonne of steel, modern furnaces can manage with less than 2 kg/t. For this reason the consumption of electrodes worldwide is still around 1.1×10^6 t/a (2005).

Needle coke qualities are used as part components for cathode blocks, sidewall blocks, and ramming pastes for the production of aluminum. A further application of this quality is its use in steel carburization.

A series of special industrial uses of carbon material, which are negligible in terms of quantity, make use the thermal and chemical resistance, the thermal and electrical conductivity, and graphitization of petroleum coke. The amount required for this purpose is normally covered by carbon dust separated during the production of carbon anodes and electrodes.

[Top of Page]

5. Quality Aspects

The quality of coke is chiefly influenced by the feed material used in the petroleum refining. Of second and third importance are the coking and the calcining conditions. Structure, density and VCM contents can vary significantly as can the sulfur, metals, and ash contents.

Over the last years an international standardization of testing methods for carbon materials has been developed. ISO, EN, DIN, or ASTM specifications or drafts exist for nearly all parameters. Standardization has almost been achieved even in the case of coefficient of thermal expansion measurement (CTE), for a long time tested by proprietary methods of electrode manufacturers. The situation that national standards (ASTM, DIN) are incorporated in the international ISO specifications makes itself positively noticeable.

5.1. Green Coke

Compared with coal, the petroleum coke has

- higher energy content
- lower ash residue
- lower VCM content (volatile combustion matter).

The quality of normal green coke mainly used as fuel is defined by the sulfur content. Additional parameters are the water content (for the quantity calculation), the ash, and VCM content. In the case of green coke used for higher grade products the level of metals and the grain size distribution are further quality criteria.

5.2. Regular Calcinate

The quality of calcined petroleum coke produced in Europe is superior to that produced in other continents. Due to authority requirements and improvements in efficiency, regular calcinate in Europe has undergone a quality improvement towards needle coke: the sulfur contents, reactivity results, and CTE figure are lower, and the vibrated bulk density is higher. Some non-European manufacturers are already producing similar qualities. Typical European quality characteristics are listed below. The figures in brackets refer to quality requirements outside Europe [26-29].

Sulfur content, wt %	max. 1.7 (3)
Ni/ V content, ppm	max. 120 – 150 (200)
Na/ Ca/ Si content, ppm	max. 100
Vibrated bulk density (1 – 2mm), g/cm^3	0.80 – 0.85 (0.75 – 0.80)
Air reactivity (1 – 1.4 mm), wt %/min	max. 0.15
CO_2 reactivity (1 – 1.4 mm), wt %	max. 12
CTE factor (carbon anodes), 10^{-6} /°C	3.5 – 4.5 (5.5)
Hydrogen content, wt %	0.05 – 0.09
Real density, g/cm^3	2.06 – 2.10
Ash content, wt %	0.05 – 2.20
Crystallite size L_c, nm	2.5 – 3.2
Specific electric resistivity, μΩ m	460 – 520
Grain stability, wt %	min. 70
Dust forming factor, wt %	max. 0.02
Water/oil content, wt % (from antidust additive)	max. 0.3

Except for the SO_2 emission the sulfur content only has slight side effects on aluminum electrolysis.

The goal in aluminum electrolysis is to have a small consumption of anodes with a good electricity yield. Good production results are 95 % electricity yield with an anode consumption of < 410 kg per tonne aluminum. These figures can be reached if the anodes meet the specifications for:

1. High carboxy (CO_2) reactivity residue
2. High air reactivity residue
3. Low specific electric resistivity
4. Low thermal conductivity

With the capacity increase of aluminum smelters, larger anodes have come into use. Cracking of the anodes due to thermal shock has been observed occasionally in these cases. Good thermal shock resistance of the anodes is therefore desirable. Improvements in thermal shock behavior will occur if CTE, the Young's modulus, and the flexural strength are decreased, and the thermal conductivity is increased.

5.3. Needle Coke

In the production of steel by the arc furnace process about 1×10^6 t/a graphite electrodes are needed. Charging stock for the electrodes are needle coke (about 85 %), and pitch binder for backing and impregnation. Needle coke is the highest grade product of calcinated petroleum coke. Because of the massive thermal and mechanical stresses in the arc furnaces, the highest demands are placed on the electrode properties. Moreover, subsequent modified graphitization due to the baking of the electrodes result in additional stricter quality requirements.

Modern lengthwise graphitization is considerably more economic and only needs a third of the time required by conventional Acheson graphitization (⟶ Carbon – Manufacture), but the more drastic reaction conditions demand superior qualities from the needle coke. For this reason needle coke specifications have greatly improved over the last 15 years. Room for further improvement seems only marginal.

For needle coke the chemical properties (low contents of S, Ni, V, Na, Ca, Si) and the following physical characteristics must be optimum:

1. High real density and high vibrated bulk density
2. High grain stability and high proportion of coarse grain
3. Low coefficient of thermal expansion

In addition the microcrystalline structure is important, because all physical parameters depend on it. A high level of anisotropy, regular micropores between the graphite lattice plane, a high crystallite size, especially after graphitization, and crystal areas, which contrary to graphite have to be offset against each other, are indispensable. In order to obtain these needle coke qualities, three manufacturing areas have to be optimized:

1. Feed selection for delayed coking (low sulfur and low metal residues with high contents of 3, 4, and 5-ring aromatics and low asphaltene content)
2. Delayed coking conditions (longer cycle, higher recycle ratio, right furnace coil out temperature ramp)
3. Calcination conditions (heating rate and final calcination temperature)

The quality analysis of needle coke is special because two parameters are specified that can only be analyzed after graphitization: The CTE and the puffing factor; i.e., the irreversible lengthwise expansion during graphitization (requirement: < 2 %). The quality comparison has been more complicated up to now by the fact that the preparation, graphitization, and measurement of CTE were proprietary company methods. Standardization however is nearly concluded (last step: proposal of CTE–DIN method, an English translation exists as ISO draft) [30].

The quality requirements of electrodes depend on the electrode diameter and also on the type of arc furnace (d.c. or a.c. operation). That is why different types of needle coke are available. Some properties of “super premium” and “intermediate premium” qualities are listed in Table 7. The CTE values refer to the measuring range 20 – 200 °C after preparation and graphitization according to the DIN method.

Table 7. Typical range of needle coke types

Property	Super Premium needle coke	Intermediate Premium needle coke
CTE (20 – 200 °C), 10^{-6}/°C	0.35	0.6
Sulfur content, wt %	0.2 – 0.4	0.4 – 0.6
Vibrated bulk density, g/cm^3	0.83 – 0.90	0.83 – 0.90
V/ Ni content, mg/kg	< 10	< 10
Real density, g/cm^3	2.13 – 2.15	2.13 – 2.14
H content, wt %	0.03 – 0.05	0.03 – 0.05

There are ongoing tests to produce “ultra premium” needle coke with CTE levels of 0.25 × 10^{-6}/ °C. Production constancy, possible production quantity and processing characteristics of this quality still need to be evaluated.

An additional challenge in the use of electrodes in the arc furnace process is the design and production of the nipples (screw connections between the individual electrodes). Thorough pilot tests, test runs, and a guaranteed production constancy for needle coke are the basis for delivery agreements.

[Top of Page]

6. Environmental and Safety Aspects

6.1. Green Coke

Although green coke contains polycyclic aromatics their total concentration of about 250 mg/kg is below the critical level of 1000 mg/kg. Also the content of the key component benzo[*a*]pyrene (ca. 25 mg/kg) is not critical. Green coke is therefore classified as a nontoxic substance [3].

No classification and labeling as regards toxicity, carcinogenicity, flammability, and environmental effects is required. Contents of Ni, V, Na, Ca, and Fe originating from the crude oil residues used as feed are accumulated in green coke. The levels amount to ca. 10 – 500 mg/kg. Small amounts of metals and anions are introduced with the cutting water, which makes up 5 – 15 % of green coke (cutting water content: Na, Ca, Cl, SO_4^{2-}). Other heavy metals (Hg, Cd, Pb, Cr, Cu, Co, As) are not traceable (level < 1 mg/kg). Hydrocarbons and metals in green coke cannot be elutriated with water (S4 eluate method).

The following classification is employed for the combustion and explosion behavior [measurement of Deutsche Montan-Technologie (DTM), mining test circuit, Dortmund]:

Combustion behavior	BZ 3 (local combustion or glowing without spreading)
Glow temperature	300 °C
Ignition temperature	700 °C
Dust explosiveness	contents > 60 g/m^3 and with ignition source of > 10 000 J

Green coke is normally stored in open piles with side walls as protection against the wind. In order to reduce the dust

discharge, green coke piles are sprinkled with water. The potential dust emission is visibly improved by means of dosing antidust additives (polyacrylate types) to the sprinkling water.

The general explosive boundary value at work of 6 mg/m^3 for fine dust concentrations is to be complied with.

6.2. Calcined Petroleum Coke

All hydrocarbons in green coke are eliminated by calcination. The remaining hydrogen content is < 0.1 %. Therefore, no polycyclic aromatics are traceable. The content of S, N, and metals is virtually unchanged. With water, only part of the salt content of the cutting water can be elutriated. No classification and labeling is required.

Due to the dryness of the calcinate, irritation of eyes and throat by the dust is possible. The general boundary value at work (max. 6 mg/m^3 for fine dust concentration) is to be complied with.

Calcined coke is usually stored in closed silos or bins to keep the material dry. To reduce dust formation, the calcinate is sucked out during transport and loading and the dust held back by filter stations. Sprinkling the calcinate with residue parts of 0.2 – 0.4 wt % (old process) or an aqueous emulsion of dust binder prevents environmental problems from dust. The dust forming factor ("R+D-Carbon"-method) must be < 0.02 wt %. The combustion and explosive behavior of calcined coke presents no problem:

Combustion behavior	BZ1 (no combustion)
Glow temperature	no glowing up to 450 °C
Ignition temperature	no ignition up to 850 °C
Dust explosive	no

Since no combustible dust – air mixture is formed, calcined dust is a material that has no security zone classification (zone 11) and does not come under the risk management regulation.

[Top of Page]

References

1. R. De Biase, J. D. Elliot, "Coking Process Reflects Trends, Innovations," *Oil Gas J.* (1982) April 19, 81. Links
2. Ch. L. Montell, *Carbon Graphite Handbook*, Interscience Publishers, New York 1968.
3. CONCAWE, reports nos. 93/105 and 95/95, "The Classification and Labelling of Petroleum Substances according to the EU dangerous Substances directive," Brussels, July 1995.
4. CEM. *Review* (2005) no. 220, 4. February; Chembureau Energy Market, 1–22, J. M. Chandelle.
5. Statistik der Kohlewirtschaft, Essen 1995.
6. H. Predel, "Delayed Coking," *Erdöl Kohle Erdgas Petrochem.* **43** (1990) 145 – 146.
7. Conoco Inc., US 44 55 219, "Method of reducing coke yield", H. R. Janssen, G. L. Poffenbarger, 1984
8. Conoco Inc., US 45 18 487, "Process for improving product yields from delayed coking", H. G. Graf, H. R. Janssen, 1985
9. H. Predel, "Theoretical and Practical Methods for Coke Yield Calculations and Optimisation," *Light Metals 1992*, TMS, San Diego, p. 601 – 609.
10. N. Liebermann, *Troubleshooting in Process Operations*, Pen Well Publishing Comp., 1985, 16 – 63.
11. H. Predel, "New Developments in Delayed Coker Operations," *Carbon Conference*, Essen, 22 – 26.6.1992.
12. A. Stefani, "Debottleneck Delayed Cokers for greater Profitability," *Hydrocarbon Process.* (1996) June, 99 – 103.
13. Mobil Oil Corporation, US 4 118 281, "Conversion of solid wastes to fuel coke and Gasoline", T.-Y. Yan, 1978
14. J. D. Elliot, R. Klick, "Trends in Delayed Coker Design," *Hydrocarbon Technology International*, Summer 1994, p. 17 – 23.
15. J. D. Elliot, "Latest Coker Designs Increase Liquid Yields, Reduce Emissions," *Oil Gas J.* (1993) Nov. 8.
16. J. D. Elliot, "Maximize Distillate Liquid Products," *Hydrocarbon Process.* (1992) Jan.
17. NPRA (National Petroleum Refiners Association), "Refiners exchange experiences on FCC problems, coking operations," *Oil Gas J.* (1994) May, 110 – 115.
18. M. R. Parrish, D. G. Hammond, V. A. Citarella, "Fluid Coking: A continuous, flexible and reliable conversion process," Hydrocarbon Technology International, Spring 1996, p.25 – 31.
19. Refining 1996, *Hydrocarbon Process.* **75** (1996) Special Edition, Nov., 100.
20. H. Predel, "Steps for Green Coke Calcination-Mathematic Model and Practical Tests and Experiences," *Light Metals 1996*, TMS, p. 499 – 507.
21. W. K. Fischer, G. Collin, J. W. Stadelhofer (1990), "Technischer Kohlenstoff— Herstellung und Anwendung," DGMK-Haupttagung, Sept. 1921, p. 423.
22. J. Mochida, K. Fujimoto, T. Oyama in P. A. Thore (ed.): "Chemistry in the Production and Utilization of Needle Coke," *Chemistry and Physics of Carbon*, vol. **25**, Marcel Dekker, New York – Basel – Hongkong 1997, p. 111 – 212.
23. H. Predel, *Erdöl Kohle Erdgas*, **121** (2005) 348 – 352.
24. Roskill Information Services, "Petroleum Coke Market Research", Global Information, Inc., Feb. 2005
25. International Aluminium Institute, Statistical Report, London, April 2005

26. H. Blumberg, “Fibers for Composites Status Quo and Trends,” *Chemical Fibers Internat.* **47** (1997) Feb., 36 – 41.
27. C. Jonville, J. C. Thomas, C. Dreyer, “The Influence of Coke Source on Anode Performance,” *JOM* **47** (1995) Aug. 23 – 24.
28. S. M. Hume, W. K. Fischer, R. C. Perruchoud, B. Welch, “A Model for Petroleum Coke Reactivity,” *Light Metals 1993*, TMS, p. 525 – 534.
29. F. Keller, U. Mannweiler, “Anode Performance: The Link between Coke Properties and Aluminum Production Cost,” *PACE, Petroleum Coke Quarterly* 1995.
30. M. W. Meier: *Cracking Behavior of Anodes*, R+D Carbon Ltd., Sierre 1996.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Pharmaceutical Dosage Forms

Surgical Materials is a separate keyword.

PAUL ZANOWIAK, Temple University, Health Sciences Center, Philadelphia, Pennsylvania 19140, United States

1. Dosage Forms as Drug Delivery Systems 2
2. Routes of Administration 4
3. Dermal application 4
3.1. Parenteral Administration . . . 4
3.2. Oral Administration 5
3.3. Dermal Administration 5
3.4. Rectal Administration 6
3.5. Other Routes of Administration 6
4. Types of Dosage Forms 7
4.1. Liquid Solution Dosage Forms . 7
4.1.1. Solutions 7
4.1.2. Extractive Solutions 7
4.1.3. Parenteral Solutions (Injections) . 8
4.1.4. Inhalations 8
4.1.5. Ophthalmic and Nasal Solutions (Drops) 8
4.1.6. Otic Solutions 8
4.1.7. Liniments 8
4.1.8. Enemas and Douches 8
4.1.9. Mouthwashes and Gargles 9
4.1.10. Collodions 9
4.2. Liquid Dispersions 9
4.3. Semisolid Dosage Forms 9
4.4. Plastic Dosage Forms (Suppositories) 10
4.5. Solid Dosage Forms 10
4.6. Prolonged-Action Dosage Forms 17
4.7. Pressurized Aerosol Dosage Forms 17
4.8. Nonpressurized Gaseous Dosage Forms (Inhalants) 17
4.9. Radiopharmaceutical Dosage Forms 17
5. Pharmaceutical Excipients . . . 17
6. Quality Assurance 20
7. Pharmaceutical Containers . . . 21
8. Development of the Ideal Drug Delivery System 23
8.1. Controlled-Release Drug Delivery Systems 24
8.2. Sustained- or Prolonged-Release Technology 26
8.2.1. Oral Sustained-Release Dosage Forms 26
8.2.2. Other Sustained- and Prolonged-Release Drug Delivery Systems . 28
8.3. Targeted Drug Delivery Systems 31
9. References 32

1. Dosage Forms as Drug Delivery Systems [1–3]

Contemporary dosage forms are designed to achieve a safe and therapeutically effective response each time they are administered within appropriate regimens that reflect accurate doses and dosing intervals. Thus, drug products are drug delivery systems that accomplish the prescriber's intent and which complete the Physician – Patient – Pharmacist triad when used "as directed" in a compliant manner.

Each type of dosage form involves several concepts and components:

1) The active ingredient (drug) in correct and accurately measured amount (dose) and of appropriate purity
2) Nontherapeutic ingredients (adjuvants, excipients) needed for the safe and effective preparation and delivery of the drug moiety
3) Unit manufacturing processes (technologies) designed so as to ensure that the product elicits safe and effective responses each time (e.g., tablet-to-tablet) and for each batch
4) Container or packaging designed for patient convenience and to maintain stability and therapeutic performance
5) Comprehensive quality assurance to safeguard the drug during shelf-life, during appropriate storage and use

The *drug moiety* must be available in pure form with an appropriate particle size, crystalline structure, and physicochemical form and in an accurate amount.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_241

Appropriate *excipients* must be selected to insure stability (e.g., antioxidants, antimicrobial preservatives, buffers), accurate dose delivery (diluents), and product performance. The latter excipients include disintegrating agents for compressed tablets and polymers for enteric coatings. Some excipients are processing aids, e.g., glidants to ensure effective flow of granules from hoppers and lubricants for tablet compression. Attention must also be given to excipients for patient acceptance and compliant use (flavors, colors). Excipients should be compatible with the drug moiety and should not form complexes with the drug that lessen therapeutic availability during production and storage. Similarly, excipients must not affect the stability of the drug product or any of its components.

Production technologies must guarantee reliable therapy, i.e., equally safe and effective responses each time a dosage form is administered. The technologies depend on the dosage form in question. They must not be deleterious to product stability, nor should they decrease the availability of the drug moiety upon administration.

The equipment used also varies in size from small laboratory units for research and development to large-scale manufacturing units. Each manufacturer utilizes a pilot plant or scale-up operation in which development pharmacists and engineers have to transform laboratory-scale batches to full production scale without altering product performance, accuracy, precision, or stability. The major operations utilized in the manufacture of drug products are listed below:

1) Dissolution
2) Filtration
3) Dispersion or wetting of solids
4) Homogenization
5) Lyophilization
6) Encapsulation of powders (hard gelatin shells and soft gelatin capsules)
7) Comminution of solids, including micronizing to ultrafine particle size
8) Blending of powders
9) Granulation of powders
10) Compression and tabletting
11) Coating of powders, granules or tablets (pan coating, compression coating, air suspension coating)
12) Microencapsulation (coacervation, etc.)
13) Heating and cooling of liquids and semisolids
14) Molding (suppositories)
15) Drying
16) Pressurized filling (of aerosols)
17) Sterilization (steam autoclave, ethylene oxide autoclave, membrane filtration, radiation, laminar flow hood)

The *design of the package*, including selection of appropriate container and sealing materials, also affects the performance of the drug product. The container must assure stability throughout the product storage and use. Its shape and size must insure ease of use and efficient delivery to foster patient compliance.

Throughout the preparation of the dosage form, *quality assurance* must be guaranteed by good manufacturing practice (e.g., cleanliness, record keeping, process control). Quality assurance also includes quality-control testing of raw materials and final products to comply with relevant specifications. In-process testing (e.g., of tablet hardness and weight) is performed at critical stages of production. Each process must be validated to show that it accomplishes its intended function(s) with the same results from batch-to-batch.

During the development of a new dosage form *stability testing* is performed to verify that degradation does not occur during processing and storage. A shelf life is determined for each product and expiration dates are obtained.

Product labeling insures product identity, provides appropriate storage and dating information, and cites necessary warnings.

Bioavailability [4–6] In order to achieve a therapeutic response after administration, the drug moiety must be made available to the body and efficiently released from the dosage form. For example, a compressed tablet designed for oral administration must first disintegrate in the gastrointestinal tract. The drug from the resulting particles must dissolve in the gastrointestinal fluids and be absorbed into the blood circulation for distribution throughout the body. The drug may, however, bind to plasma proteins in the blood. The drug is metabolized in the liver but a certain proportion is excreted unchanged through the kidney. Absorption, distribution, metabolism, and excretion (ADME) oc-

cur simultaneously and at different rates that depend on the drug involved; the sex, race, age, and state of health of the patient; and other factors (e.g., presence of other drugs).

For a given drug dose to be effective, a sufficient amount must be absorbed (i.e., be available to the blood) at an appropriate rate. The kinetics of the subsequent distribution, metabolism, and excretion processes must ensure that the concentration of the drug at the site of action reaches the threshold level required to achieve the desired therapeutic response. This concentration must be maintained for a desired period. Subsequent administration of the drug must be timed appropriately to maintain this concentration at the target organ site. If too high a concentration is reached, safety may be compromised. If a subminimal concentration results, a period of non-therapy occurs.

A peak plasma concentration ($c_{\max}$) occurs at a time $t_{\max}$ after each administration of a drug product (Fig. 1). This concentration should be somewhat greater than the minimum effective plasma concentration needed to achieve the desired pharmacologic effect, but not so high as to cause side effects and toxicity. The minimum toxic concentration (MTC) is the concentration of drug in the blood that initiates an untoward response (toxicity, side effect). The *therapeutic index* is the ratio of the minimum toxic concentration of drug (MTC) in the blood to the minimum effective concentration (MEC) in the blood (MTC/MEC); the greater the number the safer the drug is.

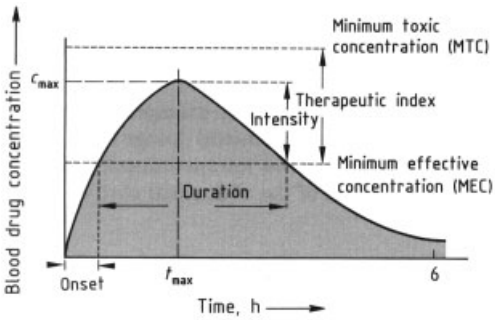

Figure 1. Blood drug concentration – time curve in a pharmacologic or therapeutic response
The shaded area is the "area under the curve" (AUC).

The parameter $t_{\max}$ is important in developing a rational regimen for dosage form administration. The point at which the threshold concentration (i.e., the MEC) is reached is termed the *onset of activity*. The time from the onset to the point when the concentration falls below the threshold level is termed the *duration of activity*.

Bioavailability is best defined as the amount of an administered drug that reaches the general circulation and the rate at which this occurs. Attention must be given to bioavailability during the design of a new drug product because it may be adversely affected by the selection of excipients and unit manufacturing operations. The area under the curve (AUC) is the total area enclosed by the blood drug concentration – time curve. It represents the extent of bioavailability of an administered dose of drug.

Excess compression pressure in tabletting may, for example, result in tablets that pass intact through the gastrointestinal tract. The relationship of formulation design and manufacture to bioavailability is generally termed biopharmaceutics. Biopharmaceutic concerns are important when a new (generic) dosage form is being designed to be pharmacologically and pharmacokinetically equivalent (i.e., bioequivalent) to an already marketed product that is no longer protected by patent.

2. Routes of Administration [7]

Drug products are administered to the human body by different routes and in various physical forms such as solutions, emulsions, suspensions, ointments, creams, suppositories, powders, granules, capsules, and compressed tablets. The most important administration route for achieving systemic therapeutic effects is the *enteral route*; the unit is swallowed, the drug dissolves in the gastrointestinal fluid and is absorbed through the mucous membrane of the gastrointestinal tract into the general circulation. Other major routes include *rectal administration* for local medication of the mucous membranes or for a systemic effect after absorption through the mucous membrane of the colon.

Parenteral injection through the skin, utilizing a syringe and needle, requires strict attention to sterility during administration and in the preparation and packaging of the dosage forms and equipment.

3. Dermal application

is used for treatment of topical conditions (e.g., skin infections, psoriasis, acne, contact dermatitis). Drugs can, however, also be made available to the systemic circulation by percutaneous absorption.

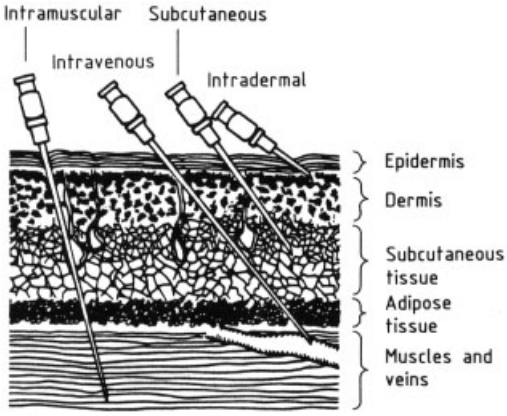

Figure 2. The most important types of parenteral administration [10]

3.1. Parenteral Administration [8], [9]

Dosage forms for parenteral administration are designed for injection through the skin with syringes and needles of appropriate size and diameter. When small doses are involved the drug is supplied in single-dose glass ampules or multi-dose vials. The vials have rubber stoppers through which the needle of a syringe is inserted to withdraw the desired dose.

For large-volume administration in hospitals or clinic settings, the drug is delivered intravenously from glass flasks or pliable plastic containers hung near the patient. Metered pumps are often used to insure accurate dosing (number of drops per minute) over long periods. Portable systems are also available in which the patient "wears" a self-contained unit consisting of a metering device and drug supply.

Parenteral injections can deliver the drug directly to the general circulation intravenously (i.v.), subcutaneously (s.c.), or intramuscularly (i.m.), see Figure 2. The *intravenous route* can be used to administer a single dose (bolus) or for continuous administration over a long period (12 – 24 h). In the latter case, several drugs can be administerd in a sequential system ("piggybacked"). Hyperalimentation solutions containing carbohydrates, amino acids, vitamins, and other nutrients are large-volume parenteral dosage forms that are administered intravenously to sustain debilitated patients for long periods of time. These nutrient solutions are administered continuously through indwelling catheters.

Parenteral dosage forms are manufactured as sterile units. They must also be free of pyrogens and foreign particles. Parenteral administration has the major disadvantage of being a potential source of infection, irritation, or pain at the site of injection. It is the route of choice when a quick therapeutic response is needed (e.g., in emergency situations or with comatose patients). It is also useful when oral drug administration cannot be tolerated (i.e., with nausea or after gastrointestinal surgery). Some drugs can only reach the general circulation by parenteral injection, for example, because of chemical instability in the gastric fluid or because they cannot be absorbed via the gastrointestinal tract (e.g., insulin).

Oily solutions or suspensions may be administered by *subcutaneous or intramuscular injection* to create "depot" doses. The drug slowly leaches from the injection site to extend the period of therapy. Small, sterile, rodlike pellets can also be implanted beneath the skin to achieve therapy periods of several months. Such implantation requires lancing and local anesthesia.

Specialized parenteral routes include intraarterial, intrathecal, intraspinal, intradermal, and intracisternal administration. Intra-arterial injection may require surgery to reach the artery and is used to administer radioopaque contrast media for viewing a specific organ. An intrathecal injection involves direct injection into the cerebrospinal fluid. Intraspinal injections are administered into the dural membrane surrounding the spinal cord. Intradermal administration occurs when the drug solution is injected into the dermis and epidermis of the skin (e.g., with solutions used to test for allergies). Intracisternal injection involves introduction of the drug solution into one of the subarachnoid cisternae of the brain.

Peritoneal dialysis is an example of a specialized large-volume parenteral dosage form. Drug solutions are injected intraperitoneally to treat the membranes of the abdominal cavity.

3.2. Oral Administration [11]

The oral route is popular because of its convenience. Oral dosage forms do not require specialized equipment, are noninvasive, and are generally more economical to prepare and administer. The main type of oral drug administration is the swallowing of solid (e.g., capsules, tablets) or liquid (e.g., syrups, solutions, emulsions) dosage forms. The drugs either medicate the gastrointestinal tract itself (e.g., antacids, anthelmintics) or are absorbed through the gastrointestinal mucous membranes into the systemic circulation to achieve a therapeutic effect at a distant target site or organ.

Orally administered dosage forms can also be held under the tongue (sublingual) or in the cheek pouch (buccal). The drug is absorbed through the oral mucous membrane into the blood stream. It enters the jugular vein and passes directly to the heart, avoiding the liver on its first pass through the body. This is in contrast to absorption through the gastrointestinal tract where the absorbed drug first passes through the portal vein to the liver before distribution to the rest of the body. Thus, drugs that are metabolized quickly by the liver (first-pass effect) can be administered more effectively in sublingual dosage forms. Also, drugs needed to treat the heart (e.g., nitroglycerin for angina) are delivered rapidly to the site of action from sublingually administered solid dosage forms that dissolve quickly.

Some dosage forms are designed for local treatment of the mucous membrane and gums of the mouth and throat (i.e., mouthwashes, gargles, lozenges).

3.3. Dermal Administration

Certain dosage forms are designed to treat the skin and its appendages (the sebaceous and sweat glands). Therapy is localized and nonsystemic; percutaneous absorption of the drug should be avoided. Dermal dosage forms can contain greasy base ingredients (pastes, ointments) or can be emulsified semisolids (cold creams, vanishing creams). Some topical applications are delivered from pressurized aerosol containers as solutions, suspensions, emulsions, foams, or powders. Solution dosage forms (soaks) are also used as astringent applications to weeping wounds.

Recent developments in drug delivery systems include the transdermal route of drug administration for prolonged periods of therapy. The drug is present either in a reservoir of drug and excipient ingredients or in a drug – excipient matrix that are attached to adhesive backings ("patches") for application to the skin. The drug is then released from the unit and transferred through the skin to the systemic circulation (see also Section 7.2.2).

3.4. Rectal Administration

Drugs can be administered rectally as enemas (solutions or suspensions) or suppositories to treat localized conditions of the rectum or colon (cathartics, hemmorhoid treatment). Such dosage forms can also be used for systemic therapy if the drug is absorbed through the mucous membrane of the colon. Treatment of comatose patients is possible by such rectal administration. This route is also useful in the treatment of infants, debilitated elderly patients, nauseated patients, and those having undergone recent gastrointestinal surgery. For example, debilitated cancer patients can gain relief from pain after administration of suppositories containing narcotic drugs.

3.5. Other Routes of Administration

Ophthalmic, Nasal, and Otic Routes. Various ophthalmic conditions (e.g., glaucoma) can be treated by local application of sterile, isotonic solutions or suspensions delivered from appropriate containers (glass dropper or plastic squeeze bottles). Sterile ophthalmic ointments are applied under the lower eye lid from specially designed tubes; any powdered drug present in an ophthalmic ointment or suspension must be of a very fine particle size ($\leq 10\,\mu m$) to avoid irritation.

Nasal solutions, drops, or sprays from appropriately designed plastic containers are used to treat local conditions of the nasal mucosa (e.g., vasoconstrictor drugs to alleviate "stuffy nose" and anti-inflammatory agents to treat perennial rhinitis). Pressurized aerosols can also be used to treat nasal conditions.

Transnasal drug delivery via absorption through the nasal mucosa has gained strong interest for systematic therapy [12]. Absorption of polypeptides by this route is the subject of considerable research. It may be an alternative route of administration of insulin for diabetic patients. The assurance of accurate dosing is, however, still a problem.

Otic (aural) administration of drug products is used only for local treatment of the external auditory canal. Since this canal is lined with dermal tissue and not mucous membrane, infections are treated much like skin infections.

Intravaginal Administration. Dosage forms applied vaginally are intended for treatment of local infections. Solutions (douches), aerosol foams, creams, or suppositories are used. Spermicidal agents can be administered similarly. Intrauterine devices that release small amounts of contraceptive drugs for long periods of time are recent developments [13]. One such commercially available product (Progestasert, ALZA) acts as a progesterone reservoir that releases effective amounts of the drug for up to one year. Effects are local, thus minimizing the systemic side effects that accompany oral administration.

Oral Inhalation [14]. Pressurized aerosol drug products are available for inhalation and are used with adapters that are placed between the upper and lower teeth during actuation. Such solution or suspension products are intended for local pulmonary use (e.g., asthma attacks) or for systemic therapy. Rather complex technologies are required to ensure delivery of accurate doses (metered valves) and to control the particle size of the delivered products. Inhalation aerosols are also available for sublingual medication. Nitroglycerin, for example, can be delivered in metered doses directly under the tongue from pressurized aerosol containers for quick absorption and therapy of angina pain.

4. Types of Dosage Forms [15]

Dosage forms can be classed according to their use or physical form. In the former case the classes correspond to the routes of administration (Chap. 2): topical or external use (creams, ointments, pastes); oral use dosage forms (syrups, tablets, capsules); rectal suppositories; parenteral injections; transdermal patches; and localized drops (ocular, nasal, otic). In the present chapter, currently used dosage forms are classified according to their physical form as follows:

1) Liquid solutions
2) Liquid dispersions
3) Semisolid dispersions
4) Plastic dispersions
5) Solid dosage forms including coated solids
6) Prolonged action products
7) Pressurized aerosols
8) Nonpressurized gaseous dosage forms (inhalants)
9) Radiopharmaceuticals

4.1. Liquid Solution Dosage Forms [16]

Liquid solutions can be used for various pharmaceutical purposes:

1) As oral medication: sweetened, flavored, with aqueous or hydroalcoholic (water – ethanol) solvent system
2) For external application (soaks, baths) to skin wounds
3) Application to body cavities or apertures: enemas, douches, bladder irrigations
4) As inhalations, by mouth, from atomizers, nebulizers, and pressurized aerosols
5) As gargles or mouthwashes to treat the mucous membranes of the mouth and throat
6) As eye, ear, or nasal drops for local application
7) Parenteral injection

4.1.1. Solutions

Early compendia identified solutions as liquid preparations of one or more substances dissolved in liquid solvents, which by reason of their ingredients or method of preparation could not be classified in some other group of official dosage forms. They were termed “liquors” and were generally made by simple dissolution or by chemical reaction.

Official requirements for pharmaceuticals and their dosage forms are specified in national

and international pharmacopeias. The Pharmacopea Internationalis and the European Pharmacopoeia are international works. National pharmacopeias include the British Pharmacopoeia, the Deutsches Arzneibuch (DAB 7, 1987), and the United States Pharmacopeia (U.S.P. XXII, 1990). In this article specifications are illustrated with examples from the U.S.P.

The United States Pharmacopeia and National Formulary (U.S.P./N.F.) currently identifies solutions according to their intended use (e.g., oral solutions, ophthalmic solutions, and inhalation solutions). Sterile parenteral solutions are termed injections.

4.1.2. Extractive Solutions

Several types of drug solution dosage forms that were originally developed years ago by extraction of plant (crude drug) materials with various solvents (menstrua) are still identified in compendia. Thus, various syrups, spirits, tinctures, and fluid extracts are extractive solutions that are listed in U.S.P. XXII and N.F. XVII. Some of these contain therapeutic ingredients (e.g., Ipecac Syrup, Senna Fluid Extract, Cascara Sagrada Fluid Extract, and Belladonna Tincture). Extracts that do not contain therapeutic ingredients were developed as flavored diluting vehicles for compounding admixture prescriptions (e.g., Peppermint Spirit, Vanilla Tincture, Tolu Balsam Tincture, and Tolu Balsam Syrup in N.F. XVII). Aromatic waters also belong to this type of solution. The last official waters in the U.S.P./N.F. are Peppermint Water, Rose Water, and Orange Flower Water.

Syrups. Aqueous solutions containing high concentrations of sucrose or other sugars are termed syrups. U.S.P. XXII lists the item "Syrup" as a near-saturated solution of sucrose in purified water. Certain polyols (e.g., glycerol, sorbitol) and flavoring agents can be added. Some syrups can be made by extraction; some are nonmedicated flavoring agents.

Aromatic waters are clear, saturated aqueous solutions of volatile oils or aromatic solids. They were used in the 1930s and 1940s as flavoring vehicles and had a minor carminative effect.

Elixirs. Sweetened, hydroalcoholic solutions for oral use are termed elixirs. Some are medicated; others were developed as official flavoring vehicles. Iso-alcoholic Elixir is an example. When this elixir was official it appeared in a high-alcoholic (75 vol%) and a low-alcoholic (10 vol%) form. By mixing proportions of each, a pharmacist could obtain varying percentages (10 – 75 vol%) of ethanol to insure solubility of prescribed drug entities in the extemporaneous compounding of prescriptions.

In commercial use the term "elixir" (unless followed by "U.S.P." or "N.F." in the United States) does not imply the official form. This is also true of terms such as "syrup" and "tincture".

Spirits and essences were alcoholic or hydroalcoholic solutions of volatile substances. Some were medicated, while others were flavoring vehicles to ensure a high ethanol content.

Tinctures were developed as alcoholic or hydroalcoholic solutions of plant drugs, synthetic drugs, or flavoring ingredients. Orginally, tinctures existed in three general types:

1) *Potent tinctures*: 100 mL contained the active ingredient content of 10 g of crude dried drugs (belladonna tincture)
2) *Nonpotent tinctures*: 100 mL contained the active ingredient content of 20 g of crude dried drug
3) *Flavoring tinctures*: 100 mL could contain the flavor principles of up to 50 g of crude, dried vegetable or fruit powder (sweet orange peel tincture, lemon tincture)

Fluid Extracts. Fluid extracts of plant materials were high in ethanol and could be viewed as standardized tinctures. They were designed so that each milliliter contained the extracted principle of 1 g of crude, dried drug.

4.1.3. Parenteral Solutions (Injections) [17]

Parenteral solutions are meant for injection through the skin and must be sterile, pyrogen-free, and isotonic. Their pH must be close to that of blood.

4.1.4. Inhalations

Inhalations are aqueous solutions for treatment of the lower respiratory tract. The smaller the droplet size delivered, the deeper into the respiratory tract the medication will reach. Administration necessitates use of inhalation devices, with nebulizers delivering smaller droplets than atomizers. Pressurized aerosols with mouthpiece adapters and actuators can deliver fine mists by means of specially designed valves. Such valves can deliver metered doses, this is an advantage over nebulizers and atomizers.

4.1.5. Ophthalmic and Nasal Solutions (Drops)

Ophthalmic and nasal solutions are used for local medication. Both should be isotonic and ophthalmic solutions must be sterile.

4.1.6. Otic Solutions

For application to the ear canal, otic solutions are best prepared with nonaqueous solvents. Anhydrous glycerol is a popular solvent.

4.1.7. Liniments

Liniments are liquid preparations meant for topical application with rubbing. They may be hydroalcoholic solutions, oily solutions, suspensions, or emulsions.

4.1.8. Enemas and Douches

Enemas and douches are solutions used to treat local rectal conditions or as vaginal medication, respectively.

4.1.9. Mouthwashes and Gargles

Mouthwashes and gargles generally are aqueous or hydroalcoholic solutions of aromatic principles and/or antiseptics. They are used to treat conditions of the mouth and throat or to refresh the buccal cavity. Antiseptic effects are, however, rather transient.

4.1.10. Collodions

Collodions are polymeric solutions (pyroxylin) that utilize diethyl ether and acetone as solvents. These viscous liquids are applied to the skin with applicator bottles. Evaporation of the solvent(s) leaves a protective film on the skin. Plasticizers can be added to provide a more flexible film. Drugs can also be added (e.g., salicylic acid as a keratolytic agent to treat corns and warts).

4.2. Liquid Dispersions [15], [16], [18] , [19]

Liquid dispersions include suspensions of solids in liquids. Such suspensions can be formulated as oral products to reduce the bad taste of a drug or to control chemical degradation of the drug. When applied topically they generally are termed lotions.

This group also includes pourable oil-inwater (o/w) or water-in-oil (w/o) emulsions. Emulsification, like suspension, can be used to overcome bad taste or chemical degradation. Oral emulsions are o/w, whereas external-use emulsions (also called lotions) can be o/w (the vanishing type) or w/o (the cleansing or "cold cream" type). Suspensions and emulsions can be packaged for use as aerosol products.

4.3. Semisolid Dosage Forms [15], [18]

Dosage forms that are semisolids at room temperature include mucilages, gels (jellies), magmas, ointments, creams, and pastes.

Gels, Magmas, and Mucilages. This group of semisolids consists of colloidal particles dispersed in a liquid medium. They can be thixotropic systems that are semisolid upon standing but become liquid with shaking or stirring. Gels (jellies) are systems that contain either organic polymers (e.g., methyl cellulose, alginates) with entrapped liquid or inorganic particles (e.g., aluminum hydroxide) suspended in a liquid. The liquid phase of gels is generally water, but it may also be an oil. Mineral oil, for example, can form a gel structure with polyethylene for use as a topically applied base.

If the dispersed macromolecules are distributed throughout the liquid without specific boundaries or without a network between them and the liquid, the system is considered single phase. If the gel consists of larger aggregates of dispersed particles, it is considered a two-phase system and termed a magma (bentonite magma).

Single-phase gels can consist of synthetic organic molecules (carbamer or carboxypolymethylene, a carboxyvinyl polymer) or may be dispersions of natural gums (acacia, tragacanth). Such gum dispersions can be termed mucilages.

Gels can be applied topically as in tretonoin gel (U.S.P.), an acne treatment. Sodium fluoride and phosphoric acid gel (U.S.P.) is used as a dental application. Gels and magmas can be used orally; aluminum hydroxide gel and magnesia magma (U.S.P.) are antacid preparations. Some mucilages can be used as pharmaceutical adjuvants to increase the viscosity of preparations or to adjust mouthfeel.

Ointments, Creams, and Pastes. These types of dosage forms are designed for application to the skin. However, ophthalmic ointments are applied under the lower eyelids from specially designed tubes; they must be sterile and should not contain gritty powders.

Ointment bases can be classified in the following manner:

1) *Oleoginous, hydrocarbon bases* have a greasy consistency, cannot contain significant amount of admixed water, and are occlusive to the skin, i.e., they restrict water evaporation from the skin surface
2) *Absorptive bases* are either semisolid mixtures that form w/o emulsions upon sorption of water (Hydrophilic Petrolatum, U.S.P. XXII) or w/o emulsions that can take up more water (Cold Cream, U.S.P. XXI)
3) *Water-removable bases* are o/w emulsions (Hydrophilic Ointment, U.S.P. XXII; Vanishing Creams) that are water-washable and less occlusive than absorptive bases or oleoginous bases (water-washable denotes bases that are insoluble in water but are o/w emulsions and can therefore be washed off with water)
4) *Water-soluble bases* that are termed greaseless ointments are least occlusive, they are blends of poly(ethylene glycols) (Polyethylene Glycol Ointment, U.S.P. XXII)

The choice of an ointment base depends upon the condition being treated, the degree of desired occlusiveness, and the stability of the drug to be incorporated. Ophthalmic ointments are generally nonemulsified.

Pastes are semisolids that contain powders. Their base ingredients can be gels or oleoginous mixtures. Pastes can thus absorb more water than greasy ointments and can be used to absorb serous secretion in lesion oozing.

The term *cream* denotes an emulsified semisolid such as cold cream (w/o) and vanishing cream formulations (o/w). As dermal applications, they have more cosmetic appeal to users than greasy ointment bases. Creams can also be used to medicate body cavities (rectum, vagina). They can be delivered from pressurized aerosol containers.

4.4. Plastic Dosage Forms (Suppositories) [15], [20]

Suppositories are semirigid, plastic dosage forms designed for insertion into body cavities (rectum, vagina, or urethra). They can be used for systemic therapy (rectally) or for localized treatment.

Rectal suppositories are the route of choice for comatose patients or after gastrointestinal surgery. They also find use in pediatric and geriatric medication.

Depending upon the base used, suppositories either melt at body temperature upon insertion (cocoa butter) or dissolve in the cavity (poly(ethylene glycol) blends, glycerinated gelatin). Glycerinated gelatin base suppositories have a rubbery consistency (70 % glycerol, 20 % gelatin, 10 % water plus drug).

Glycerin Suppositories (U.S.P. XXII) have a unique base: glycerin that is gelled or solidified with sodium stearate.

Suppositories based on cocoa butter can be prepared manually by pharmacists. However, heating mixtures of such base and drug, followed by pouring into calibrated molds and chilling, is the preferred method of preparation.

Poly(ethylene glycol) and glycerinated gelatin base suppositories must be made by fusion and molding. On a large industrial scale, all

suppositories are made by heating and molding procedures.

4.5. Solid Dosage Forms [15], [21], [22]

Solid dosage forms have progressed from powder blends dispensed as bulk powders for administration by teaspoonful dosages in water or in measured amounts wrapped in folded paper squares (folded powders, chartulae) to modern compressed tablets. Pharmacists have also dispensed cachets, capsules, tablet triturates, granules, troches, and pills as solid dosage forms.

Bulk Powders and Effervescent Powders. Bulk powders are blends of drug and water-soluble diluents for dissolution in water. The resulting solution can be administered orally or applied externally. Dosing of such powder mixtures by teaspoonful measurement affords poor accuracy and precision. Small-dose, potent drugs should therefore be weighed accurately as a dose and enclosed in a folded paper. Alternately they can be dispensed in "zip-lock" plastic envelopes/bags or encapsulated in gelatin capsules.

Bulk powders can be formulated with ingredients that effervesce on addition of water (e.g., sodium bicarbonate plus blends of citric and tartaric acid). Such powders provide carbonated, aqueous solutions to help mask bad-tasting drugs. Flavoring and sweetening agents can also be added to the powder blend.

Dusting Powders. Dusting powders are applied externally to treat skin lesions or rashes; they should have an impalpable particle size (100 – 150 μm). They are sprinkled onto the affected skin from sifter-top containers. They generally contain absorbent substances (starch, kaolin) to absorb serous fluid and talc or metallic stearates as skin lubricants.

Insufflations are obsolete. They were powder blends of fine particle size that were propelled from atomizer-like devices (insufflators) into body cavities.

Lyophilized Powders. Several drug products exist as lyophilized (freeze-dried) powders for reconstitution into solutions or syrups by the pharmacist or patient. This dosage form is generally used when the drug would otherwise decompose in the presence of the solvent system during warehousing and distribution. The reconstituted product should be used in a relatively short period of time and should be refrigerated by the patient. An example of such a product is an antibiotic syrup for pediatric use. Parenteral dosage forms can also exist as lyophilized powders to which sterile solvents must be added under sterile conditions directly before use.

Capsules and Cachets. Cachets are mentioned here only for historical reasons. They were used before gelatin capsule shells became available. They consisted of two thin disks usually made of rice flour. Each disk had a raised edge creating a hollow area. Powder was placed in one of the disks, its edges were then carefully wetted, and the second disk was placed on top of it. On drying, a sealed edge was formed. Cachets, also called Konseals, were designed to hold ca. 0.5 – 1.0 g of powder. They were intended to be swallowed with water.

Gelatin capsules may be either hard or soft (→ Gelatin, Chap. 4.3.) and are utilized to deliver small volumes of powder orally. The formulations of gelatin capsules do not require binders or disintegrating agents and are less complicated to formulate than compressed tablets. During the manufacture of the gelatin shells coloring agents, opaque pigments, and flavors can be added to the gelatin mixtures. Upon swallowing, the capsule's shell dissolves relatively quickly (< 10 min), making the enclosed powders available for gastrointestinal dissolution. Thus, bioavailability should be more complete and faster than that of compressed tablets.

Hard gelatin capsules are made from water and gelatin and can be produced in several sizes. They are of two-part construction (cap and base) and can be filled and dispensed by pharmacists extemporaneously or mass-produced. When the capsules are filled by high-speed machines, lubricants (metallic stearates, talc) can be added to the formulations to ensure uniform powder flow from the filling hoppers to each capsule shell. Such capsules range in size from number 000 to 5. The latter can hold ca. 65 mg of aspirin powder; number 00 can hold ca. 650 mg.

Soft gelatin capsules contain glycerol, as well as water and gelatin [23]. They cannot be filled

extemporaneously because they have to be made by a heat-sealing process (Fig. 3). Two rolled sheets of gelatin are kept pliable with mild heat (37 – 40 °C). The rolled sheets meet in an injection portion of the machine where the ingredient is injected into the resulting capsular form. This is sealed and ejected as a finished capsule. Excess gelatin is reprocessed. The encapsulated ingredients can be a powder, semisolid, or nonaqueous drug solution. Soft gelatin capsules can take any form from a globule (pearl) to the traditional capsular shape.

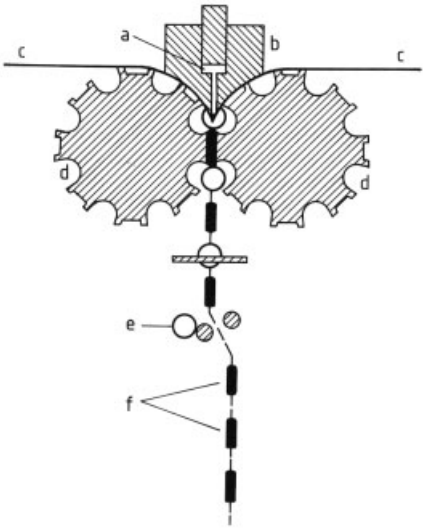

Figure 3. Rotary die process for production of soft gelatin (elastic) capsules (courtesy of R. P. Scherer Corp., Clearwater, Florida)
a) Fill material; b) Wedge; c) Gel ribbon; d) Die roll; e) Capsule; f) Waste gelatin

Pills. Although compressed tablet dosage forms are often called pills, this is a misnomer since pills were at one time a popular solid dosage form in themselves. Pills were an attempt to deliver a unit dose of a drug in a convenient, portable solid form. Their preparation entailed addition of adhesive liquids (e.g., syrup or gum mucilages) to medicated powder blends. Upon suitable manipulation the powders then became a pliable mass that was rolled into pencil-like cylinders called pipes. The pipes were each divided into an appropriate number of units on a gauged pill tile. Each unit was then rolled into a sphere, dried, and dusted with powder to minimize stickiness. Although attempts were made to add ingredients that would swell when exposed to gastrointestinal fluids to aid disintegration of such pills upon swallowing, bioavailability of the drug was poor. Solid dosage forms in pill shape are now made by granulation and compression (i.e., as a compressed tablet).

Tablet Triturates and Hypodermic Tablets. Tablet triturates were originally small, round tablets made in a small, hard, rubber mold that accomodated 50 – 100 tablets. They were intended to deliver a potent drug in an accurate amount that dissolved quickly and completely. They were composed of the drug, lactose, and/or sucrose. Binding of the ingredients was achieved by adding small amounts of 50 % ethanol to the drug – sugar blend before manual compaction into the mold.

Hypodermic tablets were similar products, intended for use by hypodermic injection after dissolution in sterile water. Today such products can be made with tablet compression equipment. One of the few such products used today is Nitroglycerin Tablets (U.S.P. XXII) for sublingual use to achieve quick relief of chest pain in angina patients. These tablets are very friable and crumble easily during handling and transport.

Troches and Lozenges are small solid disks or cubes that are designed to dissolve slowly in the mouth without disintegration to provide medication to the mucous membranes of the mouth and throat. Originally an adhesive kneadable mass was prepared, rolled into a sheet, cut into appropriate size and shape, and dried. Today such dosage forms are made by granulation (without disintegrating agents) and compression. Crystalline candy bases can also be formulated to which drug solutions can be added. Such syrup bases can then be molded into appropriate dosage units.

Granules are powder blends that have been either lyophilized or massed. Massing (wet granulation) involves addition of an adhesive binding solution (e.g., gelatin or polymer dispersion, gum, starch paste) to the powder blend. The mixture is then passed mechanically through appropriate sieves to obtain granules of the required size.

Granules can be used as such (in teaspoonful or tablespoonful doses) in water. Effervescent

ingredients (see page 10) can also be included in granule formulations to add flavoring and carbonation to the medicated solution.

Granules prepared by massing can be used to prepare compressed tablets by the wet granulation method (see page 12). The granules are dried, resieved, and blended with glidant and ejection lubricating powders (talc, metallic stearates) prior to compression.

Compressed Tablets. The compressed tablet is the most popular dosage form. It offers convenience, ease of administration, stability, accurate and precise dosing, and good bioavailability. Once the formulation has been developed and proved clinically efficacious and safe, production validation and specifications can be developed to ensure batch-to-batch reliability. Millions of tablets can be produced daily, insuring economical preparation.

Appropriate quality control procedures during tablet manufacture and of the final product assure tablet-to-tablet consistency and reliability. Such tests include weight uniformity, content identity and uniformity, tablet hardness, disintegration time, and friability. Each manufacturer establishes an appropriate set of tests with specification limits. Bioavailability is determined during clinical testing. During preformulation studies, the properties of the drug are investigated to insure stability, safety, and therapeutic effectiveness.

Tableting Procedures and Equipment. Tableting may be performed by direct compression of powder blends or by preparation of suitable granulations followed by compression.

Direct compression involves the admixture of the necessary powders followed by compression with tablet compression presses. Powder blends of small particles must have acceptable flowability and compressibility properties. This is not always the case. Modern equipment is, however, more adaptable to direct compression than in the past because mechanical delivery to the dies has been improved. Excipient ingredients (diluents, binders, disintegrating agents) are now available in granular form to insure better flow and compressibility.

A granulation blend of 90 – 95 % granules (ca. 10 – 40 mesh, 2 mm – 425 µm) plus 5 – 10 % fine particles generally provides better compressibility and flowability than fine powder blends.

When direct compression is not possible, wet or dry granulation (slugging) can be used. Granulation is required when the powdered drug with diluent etc. does not possess sufficient compressiblity calling for addition of a binding solution and granulation. Granulation is also advisable, if the powder blend does not flow well to ensure uniform content and weight of the tablets.

Wet granulation is widely used for tablets when the active ingredient is stable to water and heat. *Dry granulation (slugging)* is used when the drug is not stable to the conditions of wet granulation and the ingredients cannot be directly compressed. The ingredients are blended and compressed on a heavy-duty tablet press or slugger. Large hard tablets are produced, these "slugs" are 28 – 50 cm in diameter and weigh ca. 25 – 30 g. The slugs are then ground and sieved to appropriate mesh size and recompressed.

Various *excipient ingredients* are used in tablet production:

1) Diluents to insure appropriate size and convenient volume (e.g., lactose, calcium phosphate)
2) Disintegrating agents to aid tablet disintegration and dissolution of active ingredients (e.g., starch, microcrystalline cellulose)
3) Lubricating agents and glidants to ensure flowability of granulation, as well as antiadherents to ensure release from dies and punch faces (e.g., magnesium stearate, talc)
4) Binders to increase powder adhesion and ensure appropriate granule formation and compression (e.g., gelatin starch paste, methyl cellulose)

The major components of a *tablet press* are as follows (Fig. 4):

1) The *dies* are the units to which the granulation is delivered and whose shape determines the shape of the tablet
2) The *upper punch* fits snugly in the die and delivers the pressure
3) The *lower punch* also fits snugly in the die, it controls the volume of granulation delivered to the die and therefore tablet weight

The compression cycle (Fig. 5) begins with the delivery of the granulation from the hopper to the die by a feeder device that allows uniform

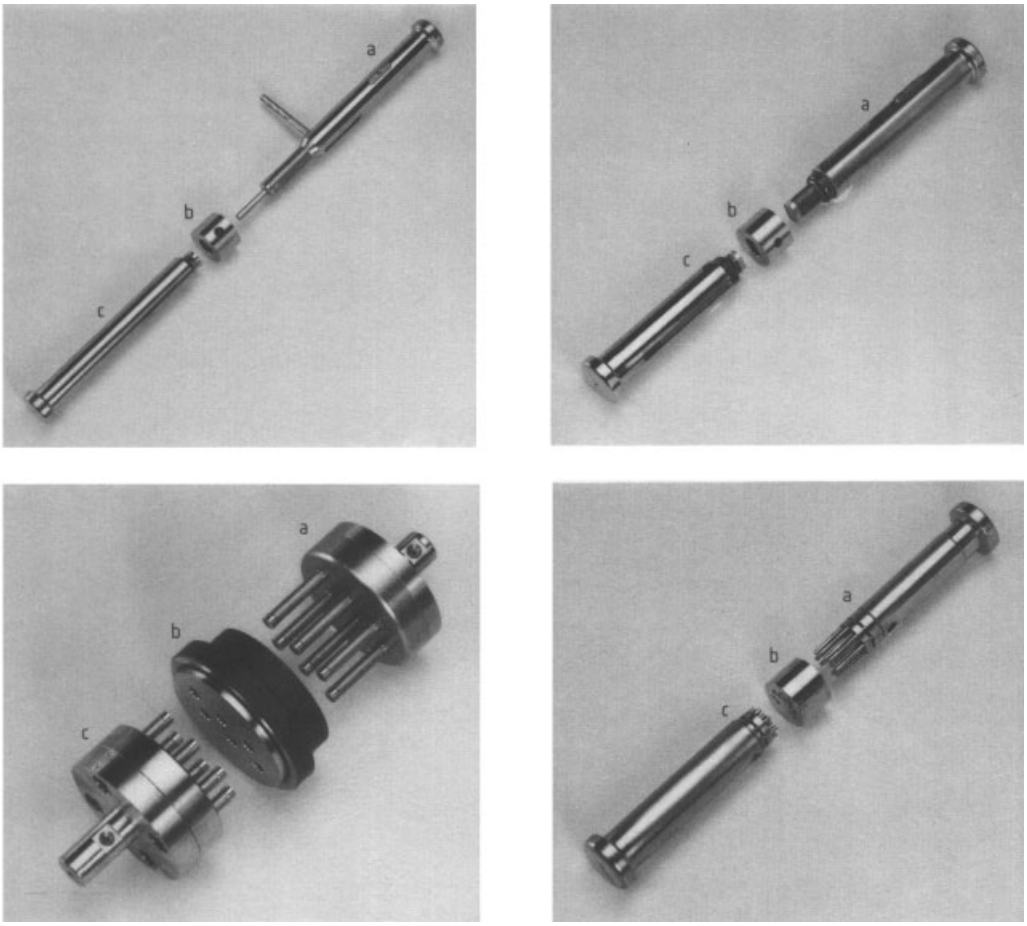

Figure 4. Tableting tools (courtesy of Fette GmbH, Hamburg, Germany)
a) Upper punch; b) Die; c) Lower punch

filling. The lower punch is at its lowest position at this point. The upper punch is then lowered to deliver a defined pressure. It then moves upward together with the lower punch to eject the tablet. Single punch presses have only one die – punch unit and compression is fast. On a multistation rotary press (one station consists of one die, one upper punch, and one lower punch), more than 60 dies can be assembled in a rotating circular table (Fig. 6). The punches ride over and under the dies in well-tooled cam tracks or channels in the head and foot areas of the press to achieve the synchronized upward and downward stroking action.

Newer high-speed presses can produce $>10\,000$ tablets per minute. Such units are equipped to measure compression and ejection forces and monitor tablet weight. Weight variations can be adjusted automatically by computer program. Feeding of granulation in sufficient quantity and rate to the die can be a problem. Standard gravity feed from hopper to feeder frame is not sufficient. Rotary agitator mechanisms within the feeder-frame units are used to induce faster feeding to the dies.

Types of Compressed Tablets. Compressed tablets can be divided into three main groups (Fig. 7):

1) Single-layer, uncoated tablets
2) Layered tablets
3) Coated tablets

Single-layer, uncoated tablets may be classified as follows:

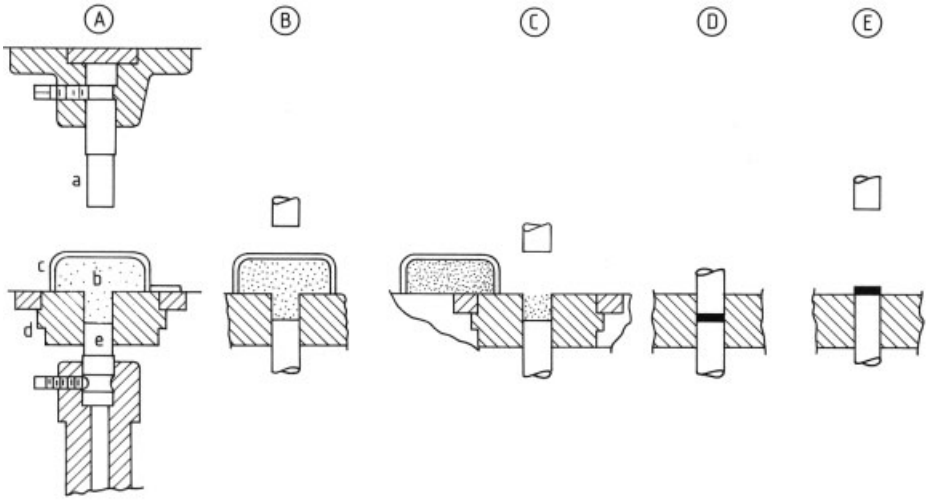

Figure 5. Compression on a single-punch tableting press (courtesy of Vector Corp., Marion, Iowa)
a) Upper punch; b) Granulation in die cavity; c) Feed shoe; d) Die; e) Lower punch
The feed shoe is filled with granulated materials from a hopper and is positioned over the die cavity (A) which then fills (B). The feed shoe retracts (C) and the upper punch lowers to compress the material (D). The upper punch retracts and the lower punch rises to eject the compact (E).

1) Tablets to be swallowed for systemic use or to medicate the gastrointestinal tract
2) Tablets that are to be chewed before swallowing; these are popular with young children (vitamins) and are generally flavored and sweetened
3) Buccal or sublingual tablets for quick dissolution and absorption and used for drugs that are not stable in gastrointestinal fluid, that are needed for quick cardiac therapy (angina), or have significant first-pass liver metabolism (see page 6)
4) Effervescent tablets
5) Lozenges or troches
6) Vaginal inserts
7) Pellets for parenteral implantation

Layered tablets may contain two or more layers and are prepared on special presses in which two or more different granulations are delivered from several hoppers. Layered tablets are used for two main reasons:

1) To separate ingredients that may be chemically or physically incompatible during processing and storage
2) To provide an initial (loading) dose from one layer and a prolonged-release dose from another; a nonmedicated layer (e.g., lactose granulation) can be used to separate the drug-containing layers

In *coated tablets* the coating protects the ingredients from atmospheric oxygen, water, carbon dioxide, and/or light [24]. Tablets are also coated to mask bad taste and to control the site of disintegration and drug dissolution. In the case of enteric coatings, for example, the tablet remains intact in the stomach but disintegrates with drug dissolution in the duodenum. Powders, granules, and tablets are also coated to control the rate of drug release.

Sugar coatings are applied in rotating pans. The core tablets are somewhat harder and more spherical than tablets that are not to be coated so that they can withstand tumbling and thus facilitate application of the coating. Batches of core tablets are initially dusted to remove powder from their surface and are sealed with a thin coat of food-grade shellac. Alternate coatings of syrup and powdered sugar are then added with intermediate drying until a rounded geometry is obtained. Smoothing coats of syrup are applied with coloring agents, if called for. Finally, the tablet is polished with a thin coating of wax and imprinted with a logo.

Film coating is faster than sugar coating and allows tablet surface engravings to be

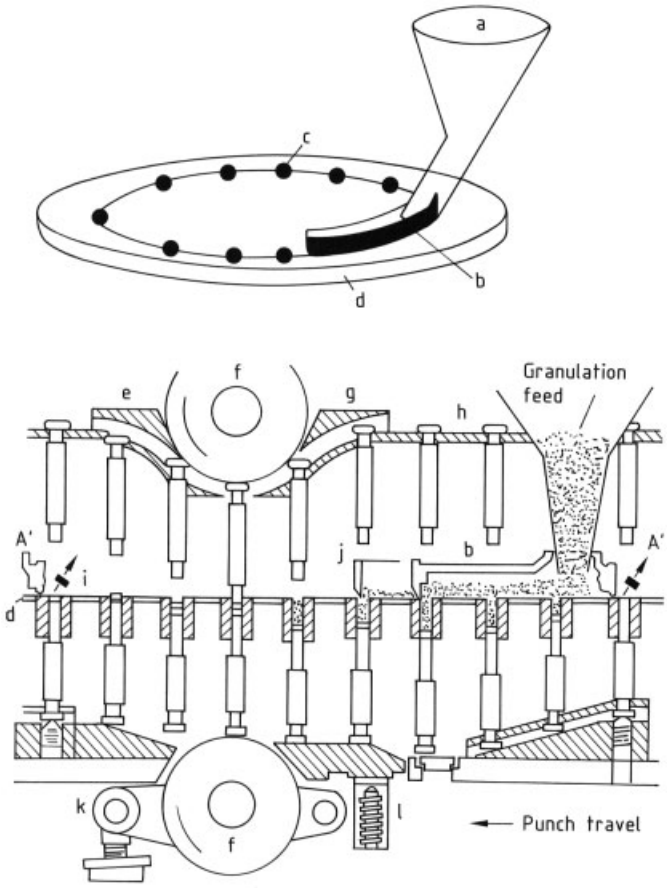

Figure 6. A tableting cycle (A′ – A′) on a rotary press (courtesy of Stokes – Merril Inc., Bristol, Pensylvania)
a) Hopper; b) Feed frame; c) Die position; d) Die table; e) Raising cam; f) Pressure roll; g) Lowering cam; h) Dwell cam; i) Knock-off; j) Scrape-off; k) Roll carrier; l) Lower punch access plug

seen. Unlike sugar coating it does not increase tablet weight significantly. Polymeric solutions (e.g., cellulose acetate phthalate for enteric coatings, sodium carboxymethyl cellulose for water-soluble coatings) are sprayed onto the tablets as they tumble in rotating pans (Fig. 8). Such solutions can contain plasticizers (e.g., poly(ethylene glycols), phthalates), coloring agents, and opaquing ingredients. Originally, organic solvents were used (e.g., acetone, chloroform). Modern technology now allows the use of aqueous polymeric dispersions. The polymers can be used to produce enteric film coatings, as well as coatings that dissolve in gastric fluids.

Film and sugar can be accomplished in equipment other than rotating pans. Air-suspension coating (Wurster apparatus) employs a cylindrical chamber in which the cores are suspended in a stream of air (→ Microencapsulation, Chap. 4.5.). Coating solutions are injected into this air stream with quick solvent evaporation. This technique can also be used to coat powders or beads. The coated beads or powders are then encapsulated or compressed into tablets to provide prolonged-release dosage forms (see page 30).

In *pressure coating*, coatings are compressed around core tablets. Two tablet presses are used in tandem: one to make the core tablet and the other to receive the core and compress the coating around it. The coating can be a protective one. However, it can include the initial-release portion of a prolonged-action tablet, with the core being a sustained-release unit. This situa-

tion also can be accomplished with sugar coating technology: the initially released drug portion added in the sugar coating and the sustained-release portion in the core.

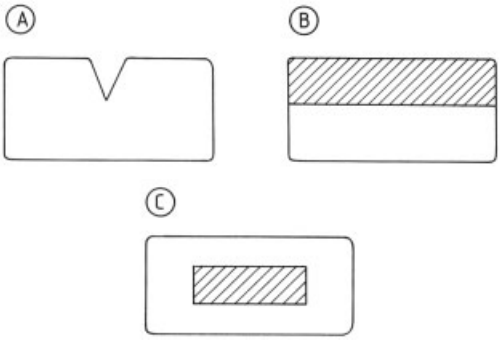

Figure 7. Compressed tablet types
A) Single-layer, uncoated, scored (bisected); B) Bilayered, uncoated; C) Compression coated with core: different drugs can be located in the shell and the core can also be a delayed-release or sustained-release component

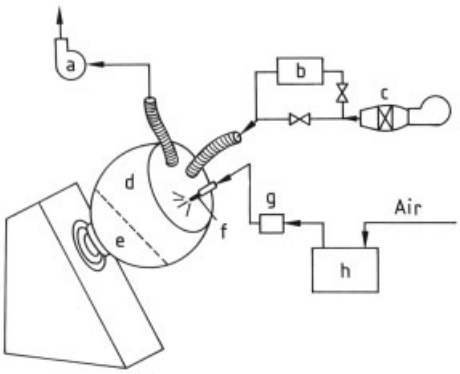

Figure 8. Hydraulic film coating system [25]
a) Exhaust; b) Heater; c) Cold/warm-air blower; d) Dragee vessel; e) Position of the tablet cores; f) Spraying gun; g) Filter; h) Pump

4.6. Prolonged-Action Dosage Forms [26–29]

The purpose of prolonged-action dosage forms is to extend periods of therapy from a single administration to two or three times those produced by conventional dosage forms. This can be accomplished by using orally administered dosage forms that release a portion of their drug content relatively quickly to induce onset of the therapeutic response and thereafter release the remaining drug content at a rate that approximates the rate of elimination of the drug from the body (Section 7.2.1). Prolonged therapy can also be achieved by transdermal drug systems [30], [31], ocular [32] and intrauterine [33] inserts, and parenteral dosage forms [34] (Section 7.2.2).

4.7. Pressurized Aerosol Dosage Forms [19]

Aerosols are dispersions of liquid droplets or small solid particles in gases. Gaseous dispersions can be produced from atomizers, nebulizers, or insufflators. Pressurized containers utilizing various propellant gases for delivering drug products through appropriately designed valve systems and actuator devices have been available since the 1950s (see also Chap. 6). These propellants include liquefied gases or gas mixtures. Examples are the Freon series of fluorinated hydrocarbons (e.g., Freon 12, CCl_2F_2) and hydrocarbons (e.g., butane, propane). Nonliquefied compressed gases are also used (e.g., carbon dioxide, nitrogen monoxide, dinitrogen monoxide). Medications can be produced in the form of suspensions, emulsions, pastes, gels, solutions, and foams depending on the formulation, valve system, dip-tube, propellant, actuator, and container. They can be used in inhalation therapy, for external application, in dental care, for vaginal therapy, and for cosmetic and toiletry products (e.g., antiperspirants and shaving foams).

4.8. Nonpressurized Gaseous Dosage Forms (Inhalants) [35]

Some drugs have low vapor pressures at room temperature and atmospheric pressure (e.g., propylhexedrine and amyl nitrite). Inhalants are prepared by impregnating suitable inert (fibrous) materials with such drugs. Cylindrical rolls of the impregnated material are then fabricated into a size that can fit into a rigid plastic tube (inhaler), for use by gentle insertion of the inhaler tip into the nostrils and sniffing (e.g., Propylhexedrine Inhalant and Amyl Nitrite Inhalant, U.S.P. XXII). The gaseous drug then medicates the nasal membranes. Aromatic ingredients can be added.

4.9. Radiopharmaceutical Dosage Forms [36]

A radiopharmaceutical is a chemical or drug that contains a radioactive isotope for human use in diagnosing or treating a disease state. These specialized drugs are dispensed as traditional dosage forms. For example, cyanocobalamin ^{57}Co capsules are used as diagnostic aids for pernicious anemia; sodium ^{131}I iodohippurate injection is used for diagnosis of renal function conditions and sodium ^{32}P phosphate solution in diagnosis and localization of ocular tumors. See also → Diagnostic Reagents, Chap. 6.

5. Pharmaceutical Excipients [37], [38]

Excipient ingredients, also called adjuvants or pharmaceutical necessities, are nontherapeutic ingredients that expedite the manufacture of dosage forms, make the delivery of the active ingredient possible, and/or ensure appropriate bioavailability (see page 3). For example, compressed tablets containing very small doses of drugs (<10 mg) also contain glidant/lubricant excipients (talc, starch) to aid the uniform delivery of the granulation from hopper to die for accurate dosing and tablet weight. Lactose, as a diluent, gives the needed bulk to such formulations. Gelatin can provide necessary adhesion as a binder agent, and a disintegrating agent (microcellulose) speeds up the rates of dissolution and gastrointestinal absorption of the drug.

Excipients should be pure, nontherapeutic, chemically stable, nonreactive with the drug, atmospheric oxygen, and/or water vapor, nonallergenic, odorless with an acceptable color and taste, and inexpensive. The physical nature of the excipient varies with the dosage forms and its physical state: liquid, semisolid, rigid but plastic (suppositories), and solid.

Solvents and Cosolvents. Purified Water (U.S.P./N.F.) is suitable for general use in oral and dermal drug products. Several U.S.P. types of water are used with parenteral dosage forms (Water for Injection, Bacteriostatic Water for Injection and Sterile Water for Injection, U.S.P./N.F).

Several water-miscible solvents are used in liquid dosage forms: ethanol (both pure and denatured), glycols, polyhydric alcohols (glycerol, sorbitol), and acetone.

A number of water-immiscible solvents are also employed: vegetable oils (peanut, soybean, cottonseed oils), mineral oil. Diethyl ether is used to solubilize pyroxylin for thickening in collodions.

Stabilizers. Several types of stabilizer are utilized to guarantee the required shelf-life: antioxidants (e.g., sodium sulfite), antimicrobial preservatives (e.g., parabens), and buffers (e.g., acetate – acetic acid, Sorensen phosphate solutions).

Excipients of Palatability [39]. Such ingredients assure product elegance and better patient compliance. They include colorants, flavors, sweeteners, and perfumes. Such ingredients must not harm product stability nor the patient (allergy, primary irritation).

Colorants include water-soluble and oil-soluble, coal-tar-derivative dyes. In the United States dyes are certified (or decertified) by the FDA. Colorants from natural sources include carmine from insects and caramel. Lakes are certified aluminum or calcium salts of coal-tar (aniline) dyes, that are water- and oil-insoluble. They are used in solid dosage forms. Pigment colorants [e.g., white titanium dioxide; red, yellow, and brown iron(III) oxide] f ind wide use in cosmetic formulas (e.g., face makeup).

Flavors can be complex mixtures of synthetic organic compounds formulated by specialty companies. They can also be of natural origin (e.g., peppermint, vanilla, glycyrrhiza, and anise). The U.S.P./N.F. still lists certain nonmedicated products (syrups, elixirs, fluid extracts, tinctures, aromatic waters, and spirits) as official entities, for use as flavored diluting solutions (e.g., Tola Balam Syrup, Aromatic Elixir, Peppermint Water, and Peppermint Spirit); see also Section 3.1.2.

Sweeteners may be sugars but several polyhydric alcohols also have a sweet taste (glycerol and sorbitol). Saccharin and aspartame are nonsugar sweetening agents that are used in preparations for diabetic patients.

Natural *perfume ingredients* and blends of synthetic aromatic compounds can be considered as excipients in cosmetic products. However, most dermal dosage forms omit such compounds to reduce the risk of irritation and allergic reactions.

Surfactants. Various surfactants (anionic, cationic, nonionic, and amphoteric) are used in several types of cosmetics and dosage forms [40]:

Anionic surfactants

1) Soaps: sodium oleate, potassium stearate, trolamine laurate
2) Fatty alcohol sulfates: sodium lauryl sulfate
3) Alkyl aryl sulfonates: sodium dioctyl sulfosuccinate

Cationic surfactants

Quaternary ammonium salts: benzalkonium chloride, c etyltrimethylammonium bromide, cetylpyridinium chloride

Amphoteric surfactants

1) Natural: lecithin
2) Synthetic: dodecyl-β-alanine, *N*-dodecyl-*N*,*N*-dimethylglycine, *N*-alkyl-β-aminopropionate

Nonionic surfactants

1) Sorbitan esters: sorbitan trioleate
2) Glycerin esters: glycerylmonostearate
3) Poly(ethylene glycol) esters: polyoxyethylene 50 stearate
4) Polyoxyethylene ethers: polyoxyethylated lauryl ether, polyoxyethylated octyl phenol
5) Polyoxyethylene – polyoxypropylene block copolymers: polaxamers
6) Polyoxyethylated sorbitan ester: polysorbate 80

They serve as primary emulsifying agents (both o/w and w/o), wetting agents, and solubilizing agents (micellar solubilization). Thus, they are used in liquid (lotions) and semisolid (creams) emulsions, liniments, suspensions, mouthwashes, and solid dosage forms. Less than critical micellar concentrations aid in the dissolution of poorly water-soluble drugs from solid dosage forms in the gastrointestinal tract, thereby increasing the bioavailability of the drug. Cationic agents are seldom used as emulsifiers or wetting agents but are widely employed as antimicrobial preservatives. The synthetic amphoteric surfactants are generally expensive and are used mainly in cosmetic formulations.

Polymers can increase the viscosity of emulsions and suspensions, thereby increasing stability while delivering a more palatable or luxurious-feeling product.

Other polymers are used in solution to coat powders, granules, or compressed tablets; to alter the site of dissolution (in the small intestine rather than in the stomach); to slow the release of the drug for prolonged periods of therapy; or to increase product stability against environmental chemicals or other drugs that are present in the dosage form. An example of the latter would be the presence of two drugs in a tablet or capsule formulation which, when intimately mixed, would form a eutectic at room temperature. Polymeric coating of the powdered particles of one of the drugs prevents intimate contact with the other drug particles, thereby preventing eutectic formation and increasing product stability. More commonly used pharmaceutical polymers include sodium carboxymethyl cellulose, methyl cellulose, ethyl cellulose, hydroxypropylmethyl cellulose, cellulose acetate phthalate, carbamer, hydroxypropyl cellulose, hydroxyethyl cellulose, poly(vinyl alcohol), and polyvinylpyrrolidone.

Tonicity adjusters are nontherapeutic solutes that are added to aqueous drug solutions to increase their osmotic pressure to levels that are isotonic with body fluids (e.g., in parenteral, ophthalmic, or nasal solutions). The most common tonicity adjusters are dextrose (parenteral solutions), sodium chloride, and boric acid (ophthalmic solutions).

Emulsion excipients are ingredients that are added to the oil phase of emulsions to increase physical stability and obtain a more viscous or acceptable feel/taste. Such dosage forms can be used for dermal application (lotion, cream) or for oral medication. This type of excipient can also be added to greasy ointments bases. They include natural and mineral waxes (beeswax, paraffin); fatty acids (stearic, oleic), fatty alcohols (cetyl alcohol), and fatty acid esters (isopropyl palmitate); petrolatum; vegetable and mineral oils; and cocoa butter.

Polymeric excipients (protective colloids) are also used to increase the viscosity of the water phase of emulsions (e.g., sodium carboxymethyl

cellulose, methyl cellulose, alginic acid derivatives from seaweed, and gelatin). Emulsified creams and lotions also include poly(ethylene glycols) and polyhydric alcohols that act as humectants for creams. Primary emulsifying agents for o/w or w/o emulsions are surfaceactive agents that can be anionic, cationic, amphoteric, or nonionic (see page 18). They can also be used as wetting agents in solid – liquid suspensions. Protective colloid dispersions are also used in suspension technology to affect stability and taste or "mouth feel".

Suppository Base Excipients. Cocoa butter (theobroma oil) is available as a natural product for oil-base suppositories. Some commercially available bases can be considered to be semisynthetic cocoa butter; they are more stable and yield a whiter product than cocoa butter itself. Examples include Cotomer, a partially hydrogenated cottonseed oil (Procter + Gamble); Suppositol, a hydrogenated coconut oil triglyceride (Fritz Wetz); Wecobees, a series of triglycerides derived from coconut and palm kernel oil (Drew Chemical Corporation); and Witepsol, a series of triglycerides (Riches-Nelson Inc.).

Water-dispersible suppository bases include glycerogelatin (ca. 20 % gelatin, 70 % glycerol, 10 % water and drug) and blends of poly(ethylene glycols) of appropriate molecular mass to achieve appropriate consistency and form upon cooling.

Propellant Gases. Aerosol dosage forms for inhalation therapy or topical use utilize various propellant gases including hydrocarbons and fluorinated hydrocarbons (see section 3.7). Fluorinated hydrocarbons have been limited in use to medicated aerosols for metered-dose inhalers and contraceptive vaginal foams. Compressed nitrogen and carbon dioxide have also been used as aerosol propellants.

Solid Dosage Form Excipients. *Diluents*, such as lactose (water-soluble) and calcium phosphate (water-insoluble), are used to obtain convenient bulk and weight and as aids for homogeneous blending of solids.

Binders or granulating agents (e.g., gelatin, starch, and gum dispersions) add adhesiveness to powders to create granules for better flow and compression.

Disintegrating agents absorb water after the tablet or granule has been swallowed, causing swelling and eventual bursting or disintegration of the solid unit to powdered form. This increases the dissolution rate of the drug. Examples of such agents are corn starch, sodium alginate, and microcrystalline cellulose.

Glidant and lubricants ensure smooth and uniform flow of granulation blends from hoppers to dies (tableting) or capsule filling units. This is necessary for uniform dose and product weight. Anti-adherent (ejection) lubricants ensure the release of tablets from the metal surfaces of dies and compression punches. Examples of such excipients are talc, calcium, zinc, and magnesium stearate, stearic acid, corn starch, and colloidal silica.

Solid dosage forms (powders, granules, compressed tablets) that effervesce in water to create a carbonated solution before swallowing contain solid *effervescent excipients* that serve as sources of carbon dioxide and safe and nontherapeutic acid. Sodium bicarbonate is generally the carbon dioxide source; citric acid, tartaric acid, and/or sodium acid phosphate provide the acid source.

Colors, flavors, and surfactants (see page 18), can also be used.

6. Quality Assurance [41–43]

In the case of prescription drug products, the quality of a dosage form is measured by the degree to which it successfully delivers the prescriber's intent. In the case of over-the-counter or nonprescription drug products, quality is the functional result expected by the person selecting the product for self-medication. Quality is thus a gauge of the product's acceptability to the patient or user, its safety, and therapeutic effectiveness.

Several departments within a pharmaceutical company are involved in the design, formulation, and monitoring of quality parameters in dosage forms to assure satisfactory quality to patients, pharmacists, prescribers, and regulatory governmental agencies.

The *research and development units* establish the physical and chemical specifications for all of the product ingredients and packaging components. In conjunction with pilot-plant or scale-up personnel, they develop acceptable param-

eters for production processes and equipment. This includes their validation to ensure that they function as expected.

The major responsibility of the *production unit* is the preparation of the final product in a quality manner: measure and handling of ingredients, use of appropriate processes and equipment as documented in batch formulae, and packaging of the correct product in the correct package with correct labeling.

Quality control units are responsible for systems that measure the desired standards for a product. These include physical and chemical assays and measurements at three key points: raw materials, in-process, and final product. They also include documentation and audit of all pertinent records and usually the control of label distribution.

The *quality assurance unit* has the responsibility for the overall quality of its products. While this includes quality control's inspection, monitoring, and analytical responsibilities, it is more pervasive. It encompasses all personnel activities, processes, equipment, and facilities involved in the manufacture of drug products.

Regulatory concerns established by federal, state, and/or local agencies must also be satisfied, see also → Pharmaceuticals, General Survey and Development. Thus, for example, national compendia (e.g., U.S.P./N.F.) establish specifications for various ingredients and processes. In the United States, the Food and Drug Administration (FDA) approves all new drugs and monitors the quality of prescription and nonprescription drug products marketed in the United States. Several other U.S. regulatory items that affect the assurance of quality of drug products are:

1) Establishment of good manufacturing practices (GMPs) to control all aspects of manufacture that must be attended to in order to deliver high-quality, safe, effective drug products (equipment; building/facilities; control of containers, closures, packaging and labeling; production and processes; personnel; warehousing and distribution; records and reports; and salvaging of returned products)
2) Good laboratory and good clinical procedures (GLPs and GCPs) to ensure these respective procedures
3) State health and licensing agencies with parallel requirements for the manufacture of drug products within their boundaries, which generally are very similar to federal laws and regulations
4) Local government regulations

Ecological environmental regulations affect pharmaceutical production at all levels of governance. A major concern is the quality of sewage a manufacturing unit produces and its effect on water pollution. Similarly, air quality must be monitored in terms of gaseous or aerosol effluent.

7. Pharmaceutical Containers [44], [45]

Dosage form containers, including their closures (seals), should not react with nor sorb any of the product ingredients. Similarly, package components should not leach into the drug product and the containers should be impervious to the environment. Plastic containers should not be permeable to atmospheric gases nor to product ingredients. The closure should also be impervious to gases entering or leaving the product.

In some cases the drug product must be protected against light by the use of opaque or amber glass. The product should not alter the physical state of the container: for example, plastics should not soften, become tacky, nor become disformed. Parenteral containers and closures must maintain sterility until used. All this must be assured for the entire designated shelf-life of the product, throughout storage and use.

Plastics used for the packaging of pharmaceutical dosage forms include: low-density polyethylene, high-density polyethylene, poly(vinyl chloride), poly(vinylidene chloride), polypropylene, polystyrene, and polyterephthalates. They differ in barrier and permeability properties, plasticity, and resistance to fracture.

Glass used for pharmaceutical containers, including parenteral dosage forms that may be exposed to autoclave temperatures, has been classified by the U.S.P./N.F. (U.S.P. XXII). The classes differ in chemical composition and resistance to hydrolysis:

1) Type I: borosilicate glass, highly resistant

2) Type II: soda-lime glass, internal surfaces are treated (dealkanized) to make them more resistant to acid pHs, than Type III
3) Type III: soda-lime glass
4) Type NP: soda-lime glass for general purpose use (nonparenteral glass)

Type I is intended for parenteral dosage forms at all pHs. Type II is used to package acidic or neutral pH parenteral products. Type III is used for nonparenteral drug products, but may be used for parenteral packaging if appropriate tests show it to be suitable. Type NP is used widely for oral or external-use dosage forms.

Every new container and closure material must be tested thoroughly with the drug product it is intended to contain before wide-scale manufacturing and marketing occur. Such tests should stress the product container over a suitable range of temperature, relative humidity, and light exposure to assure stability and, hence, safety and effective therapeutic activity.

The U.S.P./N.F. specifies various tests for establishing acceptable standards for container materials. In parenteral drug products, for example, materials are tested for light transmission (glass and plastic), chemical resistance of glass (the powdered glass test), container permeation, and foreign particulate matter.

Liquid and solid dosage forms for nonparenteral use are generally NP glass bottles or bottles made of a plastic that does not allow excessive gas permeation. The closure should similarly allow very little or no gas leakage into the container nor leakage of water vapor or other product components (e.g., perfume ingredients) out of the container. These containers are designed either for multiple use or as one-time, unit-dose containers. Amber glass is generally used to reduce the exposure of product to light. Clear glass is used when the drug product is light stable and its clarity improves product appearance. Plastic containers are lighter in weight and thus less expensive to ship. However, plastics are usually poorly biodegradable whereas glass can be recycled.

Parenteral and ophthalmic drug products utilize both glass and plastics as container components. Small-dose injections are delivered from glass ampules or vials whereas large-volume parenteral products are packaged in glass bottles and flasks or nonrigid plastic bags. Prefilled glass or rigid plastic syringes and syringe – vial combinations are sometimes used. Ophthalmic containers contain dropper units attached to the cap closure, or the dropper unit may be part of a squeezable plastic bottle. Such containers should not contribute particulate matter to the drug product; their form and chemical composition should also be stable to the sterilization procedure used.

External-use products are also packaged in glass or plastic containers. These can be bottles for the delivery of liquid solutions, suspensions, or emulsified lotions. They can also be squeezable plastic containers with specially designed plugs to aid product delivery. Many lotions are thixotropic and the design of the container should assist product flow.

Semisolid *ointments, creams, and pastes* can be delivered from glass or plastic jars of various diameters. The product is applied with the fingers, and widemouthed jars facilitate product access. Collapsible tubes (plastic, metal; coated, uncoated) are also used to deliver semisolid products. Less contact occurs between product and fingers, product contamination during use is thus lower. Viscous liquids and creams can be delivered from bottles with pump devices.

Suppositories can be packaged in cardboard boxes and may be wrapped individually in foil. Some suppositories are molded in rigid, plastic, torpedo-shaped molds, which then act as part of the final container.

Pressurized pharmaceutical aerosols are packaged in various containers. They must be able to "contain" their contents safely under the pressures that develop during production, storage, and use. Glass bottles coated with thermoplastic poly(vinyl chloride) are used for solution drug products. The external plastic coating is a protective measure to prevent shattering if the bottle is dropped. Suspensions and semisolids are generally enclosed in opaque containers. Aerosol containers may also be made of metal, with aluminum used for oral products and tin-plated steel for topical drug production. To ensure stability, metal containers may be internally coated. These coatings must be acceptable in terms of safety and product compatibility.

The type of valve used depends upon the product use. It can be designed to deliver metered doses or to control the size of the delivered liquid droplets or solid particles. Various actua-

tors or adaptor units can be attached for different purposes (e.g., to obtain different spray patterns). Oral inhalations to medicate an asthma episode need specially designed mouthpieces. Some aerosols include a flexible plastic dip tube or standpipe as an attachment to the valve unit (Fig. 9). Semisolid products (pastes, viscous creams) may utilize a rigid, plastic, dome-shaped diaphragm that is pushed upward by the expanding propellant gases. Such products may also be enclosed in a flexible plastic pouch that is squeezed by expanding propellant. These adaptations ensure a more economical delivery of the product.

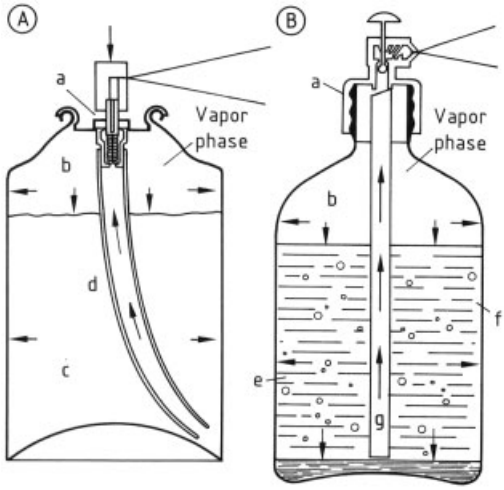

Figure 9. Two types of aerosol product containers (courtesy of Du Pont, Wilmington, Delaware)
A) Two-phase system; B) Three-phase system
a) Valve; b) Freon gas (fluorohydrocarbon) at 0.24 MPa, 294.26 K in container A and at 0.1 MPa, 294.26 K in container B; c) Solution of propellant and active ingredients; d) Standpipe; e) Aqueous solution of active ingredients; f) Liquid Freon; g) Dip tube

8. Development of the Ideal Drug Delivery System

Probably the most effective drug delivery system in use today consists of the intravenous administration of an appropriately formulated aqueous drug solution in a two-step procedure:

1) Intravenous administration of an initial or bolus dose of the solution to reach a therapeutically effective blood plasma concentration of the drug quickly
2) Subsequent intravenous administration of the drug solution at a delivery rate (drops per minute) that is adjusted to maintain the therapeutic blood plasma level for a long period of time (12 – 24 h)

Such a drug delivery system overcomes reduced bioavailability associated with the absorption of drug through membranes.

Delivery of a drug solution that is 100 % bioavailable by this method avoids the "peak and trough" effect in blood concentration produced by conventional, orally administered tablets taken every 4 – 6 h. The difference between the maximum and minimum blood levels achieved by the intravenous drug can be minimized by control of the drug concentration in the delivered solution and its delivery rate. This effect reduces the potential for toxicity and results in a lessening of the "peak and trough" effect and a steadier therapeutic response throughout the duration of pharmacologic activity.

Several drawbacks exist in the utilization of the intravenous bolus – drip drug delivery system as a widely employed, patient-accepted system, especially for the ambulatory ill:

1) Patient inconvenience and discomfort
2) Need for special equipment and personnel
3) Potential source of infection and vein irritation
4) Expense

However, the temporal pattern of such administration (short onset with an extended duration of effect) has served as a model for noninvasive drug delivery systems that overcome such disadvantages by controlled release (Section 7.1).

The use of the oral route of administration to achieve a systemic effect is much more acceptable than the parenteral route, except in cases such as emergency medication, during surgery, and in comatose patients. An oral dosage form that mimics the intravenous bolus – drip by delivering a steady and controlled release of drug to the blood would have a great advantage (Section 7.2.1). Other routes of administration mimic this intravenous system (e.g., transdermal patches, subcutaneous implants, and medicated intravaginal inserts, Section 7.2.2).

All conventional oral dosage forms have a more delayed onset of activity than an

intravenous bolus – drip administration since they must be transported across gastrointestinal membranes to reach the blood prior to distribution throughout the body. The sublingual route requires a somewhat shorter time to achieve the onset than the oral route. In order to mimic intravenous administration, release of drug from an oral dosage form and absorption (bioavailability) have to be designed so as to reduce the onset time.

In the drug delivery systems discussed above, the drug is distributed throughout the body before achieving its effective concentration at the desired site of pharmacologic activity. Tissues other than target tissues are therefore also exposed to various concentrations of the drug and pharmacologic responses may be elicited at other nondesired sites. This can result in unwanted side effects. For example, if the drug was to achieve a therapeutic effect at a particular organ innervated by the vagus nerve, undesirable responses could also occur at other organs controlled by vagal stimulation. This raises another concept in drug product development: targeted drug delivery in which the drug is released from the dosage form at a certain organ or tissue (site-specific release) or at a specific binding site within a tissue (receptor binding). Thus, systemically administered dosage forms would produce a highly localized effect (see Section 7.3). In the case of cancer chemotherapy, for example, side effects are many and serious, and such a drug delivery system would be extremely advantageous.

Ideal drug delivery systems should:

1) release a drug moiety to a target site, maintain therapeutic levels there for prolonged periods of time, and do so safely without side effects, by means of controlled-release technology;
2) have an onset of activity that approaches that of an intravenous bolus – drip;
3) be administered orally once a day, and be completely bioavailable to the target site;
4) be acceptable in terms of product elegance and utility to ensure compliance; and
5) be inexpensive.

While the development of drug delivery systems continues with the study of new technologies and materials, older concepts are being reinvestigated (e.g., medicated lollipops). The glycerogelatin base could perhaps be adapted for pediatric medication. Medicated, gummy "chews" could be used for administering vitamins, antipyretics, and cough and cold therapy to children too young to swallow or chew hard tablets or capsules and who refuse drops.

8.1. Controlled-Release Drug Delivery Systems

The ideal dosage form should deliver enough of its drug content to a specific pharmacodynamic site to achieve rapid initial onset of therapy. The therapeutic concentration should be maintained in a uniform, constant mode by slow release of the remaining portion of the drug over a period such that no more than two daily administrations (but preferably one) are needed. Furthermore, such a dosage form should be designed for oral administration and controlled release in the gastrointestinal tract to ensure good patient acceptance and compliance. In some instances where the biological half-life of the drug is $<$ ca. 6 h, the "loading" portion of the dose may not be necessary since the slowrelease technology should reach effective therapeutic concentrations effectively without it.

In some instances, a longer period of therapy ($>$24 h) may be desired from a single dose. This is not currently possible with orally administered dosage forms involving gastrointestinal release.

Transdermal controlled-release delivery from appropriate devices is a possible, noninvasive method for accomplishing such a therapy. Subcutaneous pellet implants, intravaginal insert devices, and portable or implanted infusion pump devices could be used as drug delivery systems for achieving therapy over several months. The infusion devices would need to have a built-in system of easily refillable drug reservoirs.

The design and manufacture of a controlled-release dosage form should produce a therapeutic plasma drug concentration at the site of activity followed by the maintenance of such a blood/target tissue level for an extended time period. Initial release of the drug to reach the therapeutic threshold concentration should be followed by constant, uniform release over a prolonged period. Once the initial threshold level is reached,

the amount of drug released, absorbed, and distributed to the site of action should equal the rate of drug removal resulting from metabolism and excretion.

The most appropriate means to accomplish controlled release is to develop a technology in which the drug release rate is constant and independent of the amount of drug at the release site throughout the therapy, i.e., zero-order release. Most currently available oral dosage forms designed for extended periods of therapy are not true zero-order systems but slow, first-order release systems that depend on the concentration of available drug at the release site (e.g., release from core matrices and coated pellets). (The "osmotic pump" tablet discussed in Section 7.2.1 provides zero-order release throughout most of the extended duration of activity it is designed to deliver.)

Research and development on controlled-release, oral drug delivery products began in the 1950s. Since then, the associated terminology has developed into a plethora of marketing phrases without strict standardization and sufficient attention to specificity of definition. Many terms have been used synonymously and, thus, inappropriately: sustained release/action (SR/SA); prolonged release/action (PR/PA); repeat action (RA); delayed action (DA); timed release (TR); and extended release/action (ER/EA). To avoid confusion, however, it is important to use these terms accurately within pharmaceutical and medical literature and especially in communication with health care personnel and patients [26], [27].

Initial differentiation must be made between conventional dosage forms and controlled-release products. The former are designed to release all of their active ingredients immediately for dissolution in body fluids; the latter release the drug in a controlled mode to extend the therapeutic response time and to deliver the drug to a desired target organ or tissue site. (Some controlled-release dosage forms may, however, release portions of the drug immediately to reach the onset of therapy more quickly.)

The term *controlled release* describes any drug delivery system that is used to adjust and control a therapeutic response, either temporally (extended duration) or spatially (site-specific, targeted delivery).

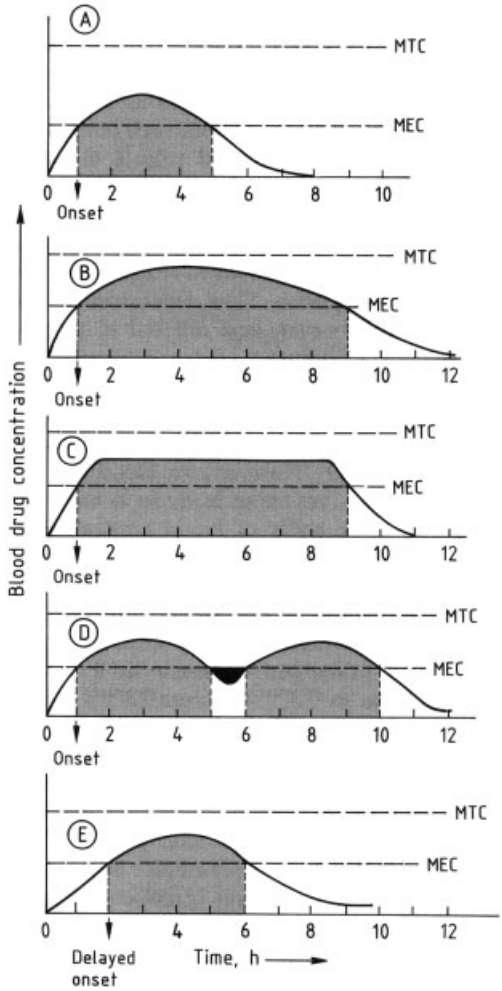

Figure 10. Blood drug concentration – time curves
A) Conventional oral product; B) Sustained release/prolonged release; C) Sustained release/controlled release; D) Repeat action with period of nontherapeutic response; E) Delayed action (coated enteric product)MTC = median toxic concentration;
MEC = median effective concentration

Sustained-release systems are designed to develop and maintain safe therapeutic responses at constant levels for extended periods that are some multiples of those achieved by conventional products (Fig. 10 A, B). With oral administration such periods can be 8 – 12 h, transdermal "patches" are effective for 24 h, and subcutaneous or intramuscular implant products can sustain therapy for 6 months or more. These systems are also termed controlled-release within the currently popular terminology if they in-

volve some precise technology that achieves or approximates a zero-order rate of drug release (Fig. 10 C).

The term *prolonged release* also refers to products designed to achieve extended periods of therapeutic response. Sometimes this term is used to refer to systems that provide a lower degree of precision than a zero-order drug release rate, and cannot therefore be considered true controlled-release systems. However, they can be considered and identified as sustainedrelease products. Hence, within such terminology sustained and prolonged release are synonyms when the rate of release is not zero order (Figs. 10 B, C).

Zero-order release technologies are most accurately described as sustained-release and controlled-release systems. Slow, first-order release systems are, however, best referred to as prolonged-release products.

Repeat-action oral tablets can be viewed as prolonged-release products, since they provide longer periods of therapy than conventional tablets (Fig. 10 D). However, the blood concentration – time curves for such products have two (or three) peak levels of blood concentration with intermediate periods where the blood concentration falls below the median effective concentration (MEC). A concentration above the MEC is not sustained throughout the entire expected therapeutic period; hence the term sustained-release to describe repeat-action tablets could be questioned.

Delayed-release or *delayed-action* oral products (Fig. 10 E) are conventional dosage forms that have been designed to release all of the drug content at one time after a delay period. Thus, enteric coated tablets are examples of delayedaction/delayed-release oral products because they begin releasing their drug content in the duodenum and not in the stomach.

Some researchers in the field feel that zero-order release has been overemphasized in the development of controlled-release products. Likewise, the overabundance of terms currently in use may confuse and distract attention from the objective of the controlled-release concept: the delivery of drugs by systems that provide excellent bioavailability to the systemic circulation for extended periods of time with targeted delivery to specific sites to eliminate both untoward side effects for the patient and undesirable effects of endogenous biochemicals upon the drugs [46].

Biotechnology has drawn attention to polyproteins, both endogenous and exogenous to the human body, as entities of therapy (e.g., insulin and growth hormone). They pose major problems of stability and effective delivery through biological membranes (i.e., bioavailability).

8.2. Sustained- or Prolonged-Release Technology

The advantages sustained- or prolongedrelease dosage forms include reduction of large fluctuations in blood concentration levels of the drug and longer duration of activity. This reduces side effects and increases patient compliance because fewer dosings are required each day.

8.2.1. Oral Sustained-Release Dosage Forms

The oral route of administration has received the greatest amount of developmental attention for sustained-release delivery systems. Oral sustained-release products mainly utilize diffusion-controlled or dissolution-controlled drug release technologies. Other systems include osmotic pump tablets, ion-exchange systems, and complexation.

The best candidates for inclusion in sustained- and prolonged-release systems are drugs that are uniformly well-absorbed throughout the gastrointestinal tract, that do not have long or very short biological half-lifes (i.e., between 2 and 8 h), and that are meant for maintenance therapy of chronic disease states (rather than for acute therapy).

Dissolution-Controlled Oral Dosage Forms. In these dosage forms the dissolution rate of the drug is decreased to maintain a sustaining concentration in the blood and at the site of activity for extended durations. This is achieved in three ways:

1) *Formation of Small Drug Beads with Differing Coating Thicknesses.* The beads dissolve in the gastrointestinal fluids at varying rates depending upon the coating thickness. The

coatings generally consist of lipoidal ingredients. The coated beads are encapsulated in gelatin shells or compressed into tablets.

2) Formation of Small, Layered Drug Beads. Drug layers are alternated with layers of rate-controlling cellulosic coatings (e.g., methyl cellulose). The outside layers can be the drug needed to reach the initial onset of activity; the remaining alternately coated layers produce the "pulsed intervals" of sustained release.
3) *Embedding of Drugs in Polymeric or Waxy Biodegradable Matrices.* Matrix dissolution products are obtained by compressing the drug with polymers (e.g., methyl cellulose, sodium carboxymethyl cellulose) or with hydrophobic waxy materials (e.g., glyceryl tristearate, cannuba wax) that are slowly biodegraded.

Diffusion-Controlled Oral Dosage Forms. Two major types of diffusion-controlled, sustained-release technologies have been described for oral administration: reservoir and matrix units.

Reservoir products utilize a central core (reservoir) of drug that is coated with a water-insoluble polymer (e.g., ethyl cellulose) through which the drug solution diffuses at a slow rate.

In the *matrix system* the drug is dispersed throughout a water-insoluble polymer matrix. The drug located on the outermost layer of the matrix is the portion that is initially exposed to the gastrointestinal fluids and dissolves first. As the gastrointestinal fluid penetrates the matrix more drug dissolves and diffuses out. The matrix may be a water-insoluble cellulosic polymer or a plastic [e.g., polymethacrylate, polyethylene, or poly(vinyl acetate)]. Gradumet (Abbott) is a tablet using a plastic matrix. Both the spent matrix of the matrix diffusion system ("ghost") and the empty membrane "sac" of the reservoir diffusion system pass through the gastrointestinal tract undissolved.

Hydrocolloid Matrix [29]. In this system the drug is incorporated within the hydrocolloid matrix of a tablet. The powdered drug is dispersed in a solid hydrocolloid (e.g., cellulolose polymers). Upon exposure to gastrointestinal fluids the drug – matrix combination floats, having a bulk density <1. The outer layer of the matrix swells, slowing inward penetration of water. The outer layer gradually erodes and a second layer then swells and erodes. During this process the drug continually dissolves and diffuses into the gastrointestinal fluid. Such a design allows the drug unit to stay in the stomach for a longer period than with a conventional tablet and release continues in the small intestine. This technology is termed "Valrelease" and was developed by Roche Laboratories for diazepam.

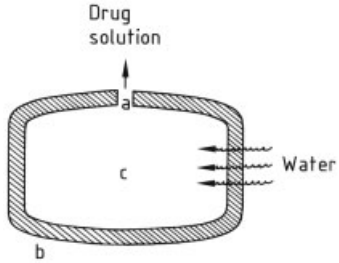

Figure 11. The OROS oral tablet system for prolonged and sustained therapy (courtesy of ALZA Corp., Palo Alto, California)
a) Delivery orifice; b) Semipermeable membrane; c) Osmotic core containing drug

Oral Sustained-Release Dosage Forms with Osmotic Control [29], [30], [47]. In this system osmotic pressure provides a driving force that produces zero-order controlled release. Tablets utilizing this technology contain a core of solid drug that is surrounded by a semipermeable membrane (e.g., ethylene – vinyl acetate copolymer or polypropylene) through which water can pass, but not the drug (Fig. 11). In the gastrointestinal tract, water flows through the membrane and dissolves the drug. The drug solution then exits at a constant rate through a laser-drilled hole in the tablet to the gastrointestinal lumen. This rate is controlled by the osmotic pressure developed by the drug solution.

The osmotic pressure concept for oral controlled-drug release was originally developed in the United States by the ALZA Corporation (OROS). In Elan's modification of this system (the multidirectional osmotic drug absorption system , MODAS), the enclosing membrane is permeable both to water and drug. Thus, water enters the unit to dissolve the drug and the drug solution then passes out through the membrane in multiple directions, not through a single port.

In another variation of this system (Alzet osmotic pump, Fig. 12) a drug solution (e) is encased in an impermeable but flexible membrane (b) within a small capsular device. An electrolyte (c) surrounds the internal portion of the device. Water enters the device through a rigid semipermeable outer membrane (d) and dissolves the electrolyte. This creates a constant osmotic pressure that presses on the enclosed drug solution, causing it to exit through an orifice. This osmotic pump device has not yet been used in oral dosage forms, but as subcutaneous implants in experimental animals. The device has about the same size as a 000 capsule (diameter 6.5 mm, length 25 mm) and can deliver 2.5 μL/h for four weeks.

Ion-Exchange Dosage Forms for Oral Sustained Release [48]. Cationic or anionic drugs can be complexed to appropriate ion-exchange resins and then encapsulated in gelatin shells or compressed into tablets. The drug is slowly released from the resin in exchange for ions in the gastrointestinal fluid, the rate depending on pH and electrolyte concentration. An example of this technology is the encapsulation of such resin complexes of amphetamine and dextroamphetamine. Recently products have become available (Pennkinetic, PennWalt Co., now owned by Fisons Corp.) that utilize very small drug – resin beads coated with a polymer (ethyl cellulose) that is not soluble in gastrointestinal fluids but through which water, electrolytes, and drug can diffuse. The beads can be suspended in viscous liquid formulations to provide up to 12 h of therapy (e.g., cough syrups). The coating also masks the somewhat unpleasant taste of the resins (Delsym Syrup, Fisons Corp.).

Chemical Complexation Dosage Forms for Oral Sustained Release [29]. A few drugs have been complexed with organic compounds (tannic or polygalacturonic acid, for example) and formulated into oral sustained-release products. These complexes have a low solubility that usually depends on pH. Slow dissolution in the gastrointestinal fluid provides the prolonged release pattern.

Bioadhesives [49]. Recent research concerned with achieving prolonged duration of activity in oral dosage forms has included investigation of bioadhesive compounds for attachment to the mucous membranes. Various polymers have been tested [e.g., sodium carboxymethyl cellulose, carbomer, hydroxypropyl cellulose, cross-linked poly(meth)acrylates]. Dosage forms utilizing such technology are envisaged as being attached to gastrointestinal or oral mucosal membranes and slowly releasing their drug content for systemic absorption. Thus, buccal tablets of various size and shape could be used, as well as dosage form units meant to be swallowed. If such units or portions of them could remain in place for long periods, the problem of gastric emptying time and overall gastrointestinal transit time as a limiting factor could be controlled.

8.2.2. Other Sustained- and Prolonged-Release Drug Delivery Systems [14–34]

There are several other sustained- and prolonged-release dosage forms that are not administered by the oral route:

1) Transdermal drug delivery
2) Vaginal inserts
3) Ophthalmic inserts
4) Parenteral implants and injections

Transdermal Drug Delivery Systems [30], [31], [50], [51]. Transdermal systems are dosage forms that are applied to intact, healthy skin to deliver their drug content percutaneously to the blood for a systemic, therapeutic effect with an extended duration of activity (e.g., 24 h). Until recently, transdermal systems were of two types: (1) drug release from a reservoir through a rate-controlling microporous semipermeable membrane made of ethylene – vinyl acetate copolymer or polypropylene (Fig. 13) and (2) direct drug release from a matrix directly to the epidermis (Fig. 14). These medicated reservoirs and matrices are attached to plastic "band-aid" type strips and can be used, for example, to administer therapeutic amounts of nitroglycerin to angina patients for periods >12 h. The drug diffuses from a matrix or through a membrane to the skin surface and then through the skin. As long as the concentration of the drug at the skin – drug interface remains constant,

the amount of drug in the unit does not affect the blood concentration. Such dosage forms are vastly different from the older, traditional plasters intended for local use (not systemic therapy). In such plasters the drug is embedded in an adhesive backing and no attempt is made to control the rate of drug release for prolonged action. For example, external analgesic drugs can be formulated so as to bring relief to some muscles.

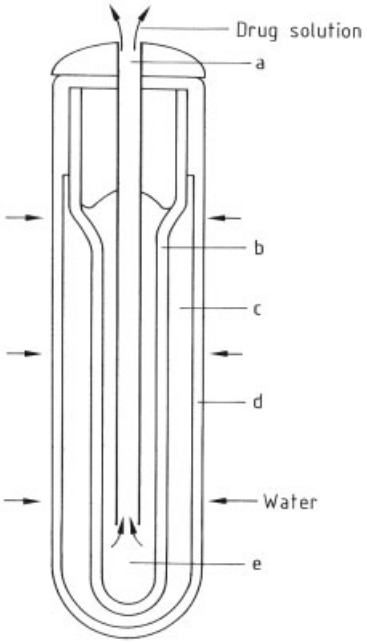

Figure 12. The Alzet osmotic pump (courtesy of ALZA Corp., Palo Alto, California)
a) Flow moderator; b) Flexible, impermeable reservoir wall; c) Saturated solution of osmotic agent; d) Rigid semipermeable membrane; e) Drug reservoir

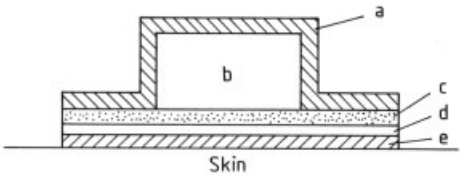

Figure 13. Transdermal patch using a semipermeable membrane to control drug release (Transderm Nitro, ALZA Corp. and Ciba)
a) Backing; b) Drug and excipient reservoir (fluid); c) Rate-release control membrane; d) Adhesive; e) Peel-away covering

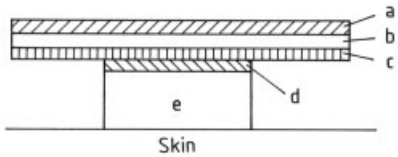

Figure 14. Transdermal patch using a matrix to control drug release (Nitrodisc, G. D. Searle; original form of Nitro-Dur, Schering – Key)
a) Occlusive outer backing; b) Absorbent pad; c) Adhesive; d) Foil base; e) Matrix (drug and excipient, semisolid formed)

The ALZA Corporation (1980) developed the first transdermal drug delivery system for scopolamine (Fig. 13). The product was designed to control nausea and motion sickness and was marketed by Ciba as Transderm Scop. With several excipients, the drug was contained in a fluid reservoir between an outer backing layer of aluminized polyester film and a microporous polypropylene membrane [52]. The membrane is covered externally with an adhesive for skin application. The adhesive contains sufficient impregnated drug to saturate the epidermis in contact with it and achieve a rapid onset of activity. Thereafter, the membrane controls the rate of drug diffusion to the skin to achieve prolonged therapy. This marketed unit has the appearance of a medium-sized "band aid" with an affixed pillow-shaped reservoir, resulting in the name "transdermal patch".

The transdermal system has since been used with several other drugs (e.g., nitroglycerin and estradiol). Other membrane materials have been used (e.g., an ethylene – vinyl acetate copolymer with the nitroglycerin product). Membrane polymers also include cellulose nitrate, cellulose acetate, polycarbonate, and polytetrafluoroethylene.

Another marketed version of a transdermal drug delivery product involves the use of a matrix and not a microporous membrane. One such product (Nitrodisc, G. D. Searle) utilizes a drug-impregnated, solid, silicone matrix that is bonded to an outer backing (Fig. 14). The drug is released by controlled diffusion to the skin surface. In a similar product developed by Key – Schering (Nitro-Dur) the matrix was a gellike structure made of polyvinylpyrrolidone poly(vinyl alcolhol), glycerol, water, lactose, and sodium citrate. This original Nitro-Dur form has been replaced by a new form in which

the nitroglycerin is dispersed in acrylic-based polymer adhesives that are applied to an outer backing, and the product has the appearance of a piece of tan-colored adhesive tape. A resin cross-linking agent is also added to the adhesive to obtain the desired rate of controlled, diffusive release of nitroglycerin.

Intravaginal/Intrauterine System (Inserts) [33]. ALZA Corporation has developed the "Progestasert", which is designed to release progesterone by slow diffusion through a rate-controlling membrane for several months. The progesterone reservoir is built in a T-shaped intrauterine device.

Ophthalmic Controlled-Release System [32]. ALZA Corporation also has developed an ocular sustained-release device for the treatment of glaucoma. Pilocarpine diffuses from a drug reservoir through rate-controlling membranes composed of an ethylene – vinyl acetate copolymer. Marketed by Ciba-Geigy, these "Ocusert" disks are placed under the lower eyelid and intended to release medication for several days (Fig. 15).

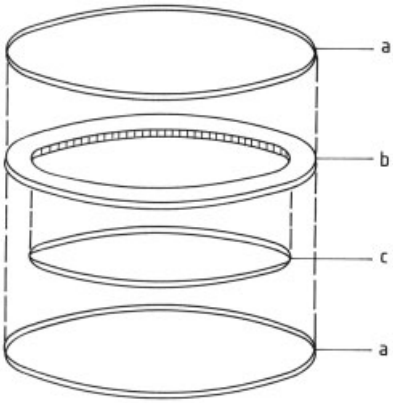

Figure 15. The Ocusert ocular therapy system for prolonged and sustained drug release after installation under the lower eyelid (courtesy of ALZA Corp., Palo Alto, California)
a) Transparent rate-controlling membrane; b) White ring of polymeric material; c) Drug reservoir film containing pilocarpine

Parenteral Sustained- and Prolonged-Release Systems [34]. Surgery or specialized injection equipment is required for *subcutaneous implants*. The implants are drug – polymer units designed for slow release of drug over prolonged periods (months). The polymers should be biodegradable; hydrogels and silicones have been used.

Parenteral administration of a drug as an intramuscular or subcutaneous *depot injection* can be used for extended periods of therapy. Such a formulation can be an oleoginous suspension or solution. The injected volume remains as a depot at the injection site and the drug slowly diffuses from it. The rate of drug release depends on the properties of the vehicles used (e.g., vegetable oils) and the drug.

Recent research concerning parenteral depot injections has involved use of colloidal dispersions for the depot unit. Depot injections of microencapsulated drug have been investigated; the polymeric coatings of the microcapsules act as membranes to control the dissolution and diffusion of drug.

Liposomes are of interest as components of site-specific or targeted drug delivery systems [27]. They are phospholipids that hydrate to form tiny sacs or vesicles (1 – 10 μm) with aqueous layers separating lipid layers. Liposomes can have a single internal aqueous compartment surrounded by a lipoid membrane or may be multilamellar structures composed of concentric lipid and aqueous layers. A drug added during liposome formation can be located within the aqueous compartment (hydrophilic drug) or become part of the lipoid layer (hydrophobic drug). Control of liposome vesicle size, as well as the chemical structure and properties of the phospholipid, can influence organ or tissue site disposition and drug release.

Microspheres (nanoparticles) are another area of current investigation as site-specific drug delivery systems [27], [53]. They have a diameter of 100 – 500 μm and are similar to liposomes in that they form colloidal dispersions which can be administered intravenously. They can be formed by a coacervation technique as employed in microencapsulation. Microspheres have been prepared from diverse matrix materials: lipoproteins, ion-exchange resins, ferromagnetic albumin, serum albumin, ethyl cellulose, gelatin, polystyrene, starch, polyacrylamide, and ferromagnetic alkylcyanoacrylate. The properties of microspheres (e.g., matrix material, size, density, cross-linking, molecular mass, drug concentration) affect the site, rate, and nature of drug release.

No liposome product is marketed currently. However, drug delivery is being investigated for parenteral (see Section 7.3), ocular, and inhalation administration. Similarly, no clinically proven drug product based on microsphere technology is commercially available yet. Proposed uses for microspheres and liposomes are similar, involving diseases of the reticuloendothelial systems (i.e., infections and cancers).

Circulating macrophages may also be able to assimilate such drug-rich particles for use within the vascular system. All such structures must be shown to exhibit matrix and lipid stability and purity, as well as drug stability within the structure. Particle size must also be strictly controlled.

Aqueous suspensions can be injected intramuscularly or subcutaneously to achieve prolonged therapy. For example, protamine, zinc, and insulin form a complex that slowly dissolves and acts for much longer periods than injections of simple insulin solutions. Control of the particle size of the suspended materials and their crystalline nature also contribute to the prolonged therapeutic pattern.

The parenteral use of a combination of an intravenous bolus injection with an intravenous drip (see section 7) achieves the goal of controlled, prolonged therapy as well.

8.3. Targeted Drug Delivery Systems [53–58]

A drug delivered by a conventional dosage form interacts with various complex molecules and cellular components and traverses different membranes before it accumulates a threshold concentration at the desired site of therapeutic activity. This diminishes its potential effectiveness and may cause undesirable results in other organ tissues. Thus, an ideal drug delivery system must include the concept of drug delivery to a specific (target) site, and only that site. Targeted delivery could possibly be achieved by attaching the drug to a "carrier". The drug – carrier unit should remain intact and not undergo reaction with any endogenous metabolites or cellular structures until it reaches its target site: a particular organ, type of cell, or subcellular unit.

Site-specific (targeted) drug delivery is different from conventional dosage forms that are delivered to an anatomical site to achieve a localized effect (e.g., suppositories to relieve local symptoms of hemmorhoids, eye drops for local ocular therapy, and creams or lotions to treat skin rashes). Site-specific drug delivery involves systemic administration of a drug – carrier/dosage form with the anticipated therapeutic activity targeted at a site away from the point of administration.

The drug – carrier unit should be able to "recognize" its target unit and release the drug entity in a controlled manner to achieve prolonged duration of activity. The carrier should then be biodegraded and pass out of the body without any further biochemical or pharmacologic intervention. An "ideal" carrier for targeted drug delivery should therefore have the following properties:

1) biodegradability,
2) high "carrying" capacity,
3) selective drug delivery,
4) reproducible properties,
5) nontoxic,
6) nonimmunogenic,
7) controlled release of drug,
8) long shelf-life, and
9) low production costs.

Recent research on site-specific drug delivery has involved a variety of carrier substances including plasma proteins, glycoproteins, polymers, monoclonal antibodies, and microparticulate carriers. This last group includes liposomes, microspheres, nanoparticles, and erythrocytes. Site specificity can vary depending on the delivery level achieved.

The term *passive targeting* is used when the drug – carrier unit passes through the same natural distribution pattern in the body that the carrier substance would normally take. An example of passive targeting is the intravenous injection of microparticles of a drug – carrier for delivery to subcellular lysosomes in macrophages within the organs of the reticuloendothelial system (liver, spleen, bone marrow).

Active targeting refers to drug – carrier distribution to the target site that is in some way altered from the normal distribution in the body of the carrier substance. The use of magnetic, biodegradable colloidal particles as carriers might be viewed as an example of the latter.

After intravenous injection, such a drug – carrier unit could be directed to the target site by an external magnetic device [59–61]. Once a suitable concentration has been achieved, the device then could release the drug at a controlled rate to achieve prolonged therapy.

9. References

1. G. S. Banker: "The Drug Product As A Delivery System" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 15 – 20.
2. P. Zanowiak: "Dosage Forms: Nonparenterals" in J. Swarbrick, J. C. Boylan (eds.): *Encyclopedia of Pharmaceutical Technology,* vol. 4, Marcel Dekker, New York 1990, in press.
3. *Kirk-Othmer,* **vol. 17,** pp. 272 – 290.
4. H. C. Ansel, N. G. Popovich: "Dosage Form Design: Biopharmaceutic Considerations" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 51 – 91.
5. D. W. Bourne, L. W. Dittert: "Bioavailability (Extent of Absorption)" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 91 – 142.
6. A. R. DiSanto: "Bioavailability and Bioequivalency Testing" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing Co., Easton 1990, pp. 1451 – 1458.
7. H. C. Ansel, N. G. Popovich: "Routes of Drug Administration" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 68 – 78.
8. J. C. Boylan, A. L. Fites: "Routes of Parenteral Administration" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 493 – 502.
9. S. J. Turco, R. E. King: *Sterile Dosage Forms,* 3rd ed., Lea & Febiger, Philadelphia 1987, pp. 6, 7.
10. P. H. List: *Arzneiformlehre,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1973.
11. J. L. Colaizzi, W. H. Pitlick: "The Oral Route of Administration" in G. S. Banker, R. K. Chalmers (eds.): *Pharmaceutics and Pharmacy Practice,* J. B. Lippincott, Philadelphia 1982, pp. 184 – 190.
12. V. H. L. Lee: "Changing Needs in Drug Delivery in the Era of Peptide and Protein Drugs" in V. H. L. Lee (ed.): *Peptide and Protein Drugs,* Marcel Dekker, New York 1990, pp. 1-56.
13. G. M. Grass IV, J. R. Robinson: "Intravaginal and Intrauterine Systems", in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 666 – 668.
14. J. J. Sciarra, A. J. Cutie: "Inhalation Aerosols" in [13]pp. 626 – 630.
15. "Pharmaceutical Dosage Forms" in: *USP XXII, U. S. Pharmacopeial Convention,* Rockville 1990, pp. 1688 – 1697.
16. J. G. Nairn: "Solution, Emulsions, Suspensions and Extractives" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1519 – 1544.
17. S. J. Turco: "Dosage Forms: Parenterals" in J. Swarbrick, J. C. Boylan (eds.): *Encyclopedia of Pharmaceutical Technology,* Marcel Dekker, New York 1990, in press.
18. H. C. Ansel, N. G. Popovich: "Oral Suspensions, Emulsions, Magmas and Gels" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 225 – 254.
19. J. J. Sciarra, A. J. Cutie: "Aerosols" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1694 – 1714.
20. H. C. Ansel, N. G. Popovich: "Suppositories, and Other Rectal, Vaginal and Urethral Preparations" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 373 – 390.
21. R. E. O'Conner, E. G. Rippie, J. B. Schwartz: "Powders as a Dosage Form" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1629 – 1632.
22. E. Rudnic, J. B. Schwartz: "Oral Solid Dosage Forms" in [20]pp. 1633 – 1665.
23. H. Senger: "Soft Gelatin Capsules: A Solution to Many Tableting Problems", *Pharm. Technol.* **9** (1985) no. 9, 36.
24. S. C. Porter: "Coating of Pharmaceutical Dosage Forms" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1666 – 1675.

25. H. List, L. Hörhammer (eds.): *Hagers Handbuch der Pharmazeutischen Praxis,* vol. VII, "Arzneiformen und Hilfsstoffe" part A and B, Springer Verlag, Berlin 1977.
26. M. A. Longer, J. R. Robinson: "Sustained-Release Drug Delivery System" in [25]pp. 1676 – 1691.
27. G. M. Grass IV, J. R. Robinson: "Sustained- and Controlled-Release Drug Delivery Systems" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 635 –671.
28. F. Theeuwes, W. Bayne: "Controlled-Release Dosage Form Design" in J. Urquhart (ed.): *Controlled-Release Pharmaceuticals,* American Pharmaceutical Association, Washington, D.C., 1981, pp. 61 – 93.
29. H. C. Ansel, N. G. Popovich: "Rate-Controlled Dosage Forms and Drug Delivery Systems" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 183 –191.
30. C. D. Black: "Update: Programmed Drug Delivery Systems", *U. S. Pharmacist* **8,** November (1983) 49 – 60.
31. Y. W. Chien: "Transdermal Controlled-Release Drug Administration" in: *Novel Drug Delivery Systems,* Marcel Dekker, New York 1982, pp. 149 – 218.
32. "Ocular Controlled-Release Drug Administration" in [31]pp. 13 – 50.
33. "Intravaginal/Intrauterine Controlled-Release Drug Administration" in [31]pp. 51 – 148.
34. "Parenteral/Implantable Controlled-Release Drug Delivery Systems" in [31]pp. 219 – 412.
35. H. C. Ansel, N. G. Popovich: "Aerosols, Inhalations, Inhalants and Sprays" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea & Febiger, Philadelphia 1990, pp. 390 – 405.
36. G. D. Chase: "Radiopharmaceuticals Used in Medicine" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 642 – 652.
37. J. C. Boylan et al.: *Handbook of Pharmaceutical Excipients,* American Pharmaceutical Association, Washington, D.C., and the Pharmaceutical Society of Great Britain, London 1986, pp. 1 – 375.
38. E. A. Swinyard, W. Lowenthal: "Pharmaceutical Necessities" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1286 – 1329.
39. J. L. Colaizzi, W. H. Pitlick: "Palatability – A Special Requirement for Oral Medications" in G. S. Banker, R. K. Chalmers (eds.): *Pharmaceutics and Pharmacy Practice,* J. B. Lippincott, Philadelphia 1982, pp. 191 –195.
40. H. Schott, personal communications, Temple University, Philadelphia, Pennsylvania.
41. A. M. Martin, R. V. Smith, V. A. Yanchick: "Product Quality Evaluation" in [39]pp. 33 – 40.
42. S. L. Hem, R. G. Stoll: "Appraisal of Drug Product Quality and Performance" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 741 – 802.
43. C. R. Erskine, Jr.: "Quality Assurance and Control" in A. R. Gennaro (ed.): *Remington's Pharmaceutical Sciences,* 18th ed., Mack Publishing, Easton 1990, pp. 1513 – 1518.
44. D. C. Liebe: "Packaging of Pharmaceutical Dosage Forms" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 695 – 740.
45. H. C. Ansel, N. G. Popovich: "Packaging, Labeling and Storage of Pharmaceuticals" in: *Pharmaceutical Dosage Forms and Drug Delivery Systems,* 5th ed., Lea and Febiger, Philadelphia 1990, pp. 116 – 133.
46. J. Robinson, personal communications, University of Wisconsin, Madison, Wisconsin.
47. F. Theeuwes: "Oral Dosage Form Design: Status and Goals of Oral Osmotic Systems Technology," *Pharm. Int.* **5** (1984) no. 12, 293 – 296.
48. C. H. Nightingale: "The Technology of Controlled-Release Drug Delivery Systems" at a symposium: *Controlled-Release Drug Delivery and PennKinetic,* Omega Communications Inc., Springfield, New Jersey and PennWalt Corp., Rochester, New York 1983, pp. 26 – 31.
49. K. Park, H. S. Ch'ng, J. R. Robinson: "Alternative Approaches to Oral Controlled Drug Delivery: Bioadhesives and In Situ Systems" in J. M. Anderson, S. W. Kim (eds.): *Recent Advances In Drug Delivery Systems,* Plenum Press, New York 1984, pp. 163 – 183.
50. A. Karim: "Transdermal Delivery Systems" in: *Drug Delivery Systems,* Aster Publishing Corp., Springfield, Oreg., 1983, pp. 28 – 32.
51. J. W. Fara: "Short- and Long-Term Transdermal Drug Delivery Systems" in [50]pp. 32 – 37.
52. H. C. Ansch, N. G. Popovich: "Transdermal Drug Delivery Systems", *Pharmaceutical*

Dosage Forms and Drug Delivery Systems, 5th ed., Lea and Febiger, Philadelphia 1990, pp. 310 – 320.
53. E. Timlinson: "Site-Specific Drug Delivery" in G. S. Banker, C. T. Rhodes (eds.): *Modern Pharmaceutics,* 2nd ed., Marcel Dekker, New York 1990, pp. 673 –694.
54. S. N. Mills, S. S. Davis: "The Targeting of Drugs/Controlled Drug Delivery" in L. Illum, S. S. Davis (eds.): *Polymers in Controlled Drug Delivery,* IOP Publishing, Bristol 1987, pp. 4 – 6.
55. P. Arthurson: "Site Specific Drug Delivery/The Fate of Microparticulate Drug Carriers after Intravenous Administration" in [54]pp. 15 – 24.
56. R. L. Juliano, D. Layton: "Liposomes as a Drug Delivery System" in R. L. Juliano (ed.): *Drug Delivery Systems,* Oxford University Press, New York 1980, pp. 189 – 236.
57. R. C. Oppenheim: "Nanoparticles" in [56]pp. 177 –188.
58. B. E. Ryman: "Liposomes: Possible Potential in Drug Delivery" in T. J. Roseman, S. Z. Mansdorf (eds.): *Controlled-Release Delivery Systems,* Marcel Dekker, New York 1983, pp. 27 – 42.
59. D. S. T. Hsieh, R. Langer: "Zero-Order Drug Delivery Systems with Magnetic Control" in [58]pp. 121 –132.
60. R. M. Morris et al.: "Magnetic Microspheres in Drug Delivery" in J. M. Anderson, S. W. Kim (eds.): *Recent Advances In Drug Delivery Systems,* Plenum Press, New York 1984, pp. 221 – 228.
61. R. Langer, L. Brown, E. Edelman: "Controlled Release and Magnetically Modulated Systems" in [60]pp. 249 – 258.

Pharmaceuticals, General Survey

John L. McGuire, Johnson & Johnson, New Brunswick, New Jersey, United States (Chap. 1 and 2)

Horst Hasskarl, Ludwigshafen, Germany (Chap. 3)

Gerd Bode, University of Goettingen, Germany (Chap. 4.1)

Ingrid Klingmann, Pharmaplex bvba, Wezembeek-Oppem, Belgium (Chap. 4.2)

Manuel Zahn, AstraZeneca, Plankstadt, Germany (Chap. 4.3)

1. Introduction 2
2. Companies, Markets, and Products 2
3. Legal Requirements and Guidance 4
3.1. General 4
3.2. Definitions 6
3.3. Requirements Relating to the Development of Pharmaceutical Products 6
3.4. Marketing Authorization 6
3.5. Requirements for the Manufacture and Quality Control of Pharmaceutical Products 7
3.6. Requirements for Placing a Pharmaceutical Product on the Market 7
3.7. Liability 8
3.8. International Non-Proprietary Names (INN) 8
4. Development of Pharmaceutical Products 9
4.1. Preclinical Safety/Toxicology . . . 9
4.1.1. Preclinical Safety 9
4.1.2. Toxicological Strategies 10
4.1.3. ICH Guidelines for Carcinogenicity 11
4.1.4. ICH Guidelines for Genotoxicity (ICH S2A and ICH S2B) 13
4.1.5. ICH Guidelines for Toxicokinetics/Pharmacokinetics (ICH S3A and ICH S3B) 13
4.1.6. ICH Guidelines for Reproductive Studies (ICH S5A and ICH S5B) . . 14
4.1.7. Preclinical Safety Evaluation of Biotechnology-Derived Products (ICH S6) 15
4.1.8. Safety Pharmacology Studies for Human Pharmaceuticals (ICH S7A) 15
4.1.9. Nonclinical Studies for Assessing Risk of Repolarization-Associated Ventricular Tachyarrhythmia for Human Pharmaceuticals (ICH S7B) 16
4.1.10. Immunotoxicology Studies (ICH S8) 17
4.1.11. Nonclinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals (ICH M3) 19
4.1.12. Inclusion of special populations (ICH M3) 20
4.1.13. The Common Technical Document (ICH M4) 20
4.2. Clinical Trials 22
4.2.1. Phases of Clinical Drug Development 24
4.2.2. Methods for Proving Effectiveness . 27
4.2.3. Adverse Drug Reactions 28
4.2.4. Good Clinical Practice 29
4.2.5. Clinical Application Dossier 30
4.3. Quality of Pharmaceutical Products 31
4.3.1. Manufacturing of Substances – Test of Starting Materials 32
4.3.2. Control of Active Substances: Impurities 32
4.3.3. Control of Pharmaceutical Products: Impurities 32
4.3.4. Specifications 32
4.3.5. Validation of Analytical Procedures 33
4.3.6. Process Validation and Qualification 33
4.3.7. Batch Release 34
4.3.8. Stability Testing 34
4.3.9. Documentation 35
4.3.10. Variations – Changes 38
4.3.11. Change Control 38
4.3.12. Compliance 38
4.3.13. Changes Requiring New Applications – Line Extension 38
4.3.14. Inspections 38
4.3.15. The Mutual Recognition Agreement (MRA) 39
4.3.16. Pharmaceutical Inspection Convention (PIC) 39
4.3.17. Pharmaceutical Inspection Cooperation Scheme (PIC/S) 39
4.3.18. Pharmacopoeias 40
4.3.19. The EP Certificate of Suitability . . . 41
5. References 42

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_273.pub2

1. Introduction

The pharmaceutical industry is largely a research-based industry. Not only has it been historically successful in providing valuable new medicines for treatment of diseases, but also its research and development laboratories have made significant contributions to our basic understanding of those diseases. In significant part because of the success of this industry, people today look forward to longer and healthier lives than ever before. Excellent reviews of this industry have been published [1, 2].

Historically, today's pharmaceutical industry emerged from a number of different sciences and approaches to finding therapeutics. Examples are: (1) the early discovery that extracts of plants such as salicylates from willow tree bark could provide products with medicinal value; (2) the discovery that proteins could be used as drugs, i.e., Banting and Best's demonstration that insulin could reverse the ravages of diabetes in 1938, and commercial expression of recombinant proteins such as tissue plasminogen activator (TPA) by Genentech; (3) Lister's small-pox vaccine and the development of the first monoclonal antibody therapeutic, OKT3, by Johnson & Johnson for prevention of tissue transplant rejection; (4) research in gene therapy; (5) use of genomics to find new targets for pharmaceutical intervention; and (6) the scientific approach which spawned most of today's pharmaceutical armamentarium – modification of organic molecules to provide therapeutically active agents through medicinal chemistry.

Active substances based on small organic molecules comprise the largest segment of available pharmaceutical products today. Other sources for active ingredients are biotechnological and biological materials, herbals, and plant extracts, homoeopathic as well as anthroposophical medicines. In some regions of the world, traditional medicinal products contribute to a large extent to the therapy of diseases, e.g., ayurveda in India or traditional chinese medicine (TCM).

In 2005, global sales for pharmaceutical products were estimated to be approximately US $ 602×10^9 [3]. This includes both medicinal products sold via prescription and directly by consumers, so-called "self-medication" or "over-the-counter" (OTC).

The pharmaceutical industry has features that distinguish it from other industries. First, this industry is research-intensive, having close interaction with universities and other research facilities. R & D is expensive, and the time needed for the discovery and development of new products typically long. Secondly, because of the enormous cost of pharmaceutical R & D, companies strive to market their products worldwide in order to recoup their R & D costs. Thirdly, this industry is heavily regulated, both in terms of requiring regulatory approval to market medicinal products and, in many countries, requiring price approvals. Price regulation is important not only because it affects profits and the ability to reinvest in new research within the pharmaceutical industry, but also because it relates to the cost of healthcare generally. Finally, physicians tend to be the customers rather than patients who actually use the pharmaceutical industry's products. As a result, a complex interaction exists with the medical profession, governments, and insurance groups that actually pay much of the cost of this industry's products in many countries.

Pharmaceutical products are one of the most efficient forms of therapy available today. Because of this success and the aging of the population, the percentage share of healthcare costs in GNP has grown.

2. Companies, Markets, and Products

The global pharmaceutical market is divided into two segments. One segment is medicinal products available only on prescription by a physician or other healthcare professional. The second segment is OTC products. These are generally well-established treatments for minor diseases, where safety and the ability to self-diagnose are of no concern. In this discussion of the pharmaceutical market, both segments are combined.

On a geographical basis, the sales figures for the major markets of medicinal products are shown in Table 1. These data show that the top six countries account for about 70 % of the worldwide market of US $ 601.3×10^9.

Table 1. World pharmaceutical market by country in 2005 [3]

Country	Sales (US \$ 10^9)
United States	252.2
Japan	67.7
Germany	31.9
France	30.3
Italy	19.8
United Kingdom	19.5
Spain	15.1
Canada	13.5
China	11.6
Brazil	9.1

If a breakdown of worldwide sales is made among companies, it can be seen that the top ten companies account for about 47% of the total audited market (Table 2). The remainder is supplied by a large number of smaller companies. Currently, the largest company in terms of worldwide sales covers just about 8 % of the total audited market. Obviously, many smaller companies still represent very significant businesses.

The major pharmaceutical companies maintain research facilities in a number of countries in North America, Europe and Asia and have multiple commercial operations throughout the world.

The global pharmaceutical market is divided into many therapeutic segments and classes. The largest class (Table 3) are cholesterol and triglyceride regulators, which accounted for approximately 5.7 % of the global market. Antiulcerants is the next largest segment, accounting for 4.7 % of the global market.

The top 10 branded medicinal products worldwide by sales in 2005 are shown in Table 4. Several points should be made. First, these sales figures are worldwide sales. The leading products in individual countries vary considerably because of differences in the incidence of diseases, medical practice, and the particular strengths of international and local companies in that particular market. Secondly, if prescription number rather than sales ranked the top 10 medicinal products, a somewhat different picture would emerge.

Table 3. Top five therapeutic classes of the pharmaceutical market in 2005 [3]

Therapeutic classes	Sales (US \$ 10^9)	% of Sales
Cholesterol & triglyceride regulators	32.6	5.7
Antiulcerants	26.9	4.7
Antidepress. & mood stabilizers	19.9	3.5
Antipsychotics	16.2	2.9
Erythropoietin products	12.4	2.2

Finally, most of the sales volumes discussed above relate to sales of pharmaceutical products based on prescriptions. There are also OTC products that are sold without a prescription. These products are commonly used for relatively minor diseases and have not only been shown over time to be safe and simple to use, but it has been concluded by regulatory authorities that no medical guidance beyond instructions on the package are needed by the patient to use the product correctly. The OTC company typically does not do innovative research to discover new active substances, but rather distributes well-established patent-free medicines directly to the consumer.

In the case of *prescription pharmaceuticals*, the research-based pharmaceutical company is the innovator of new pharmacologically active compounds known as new chemical entities (NCEs) or, to cover both small molecules and biotechnology products, new active substances (NAS). The research to discover a useful new active substance is expensive and lengthy, and the probability of success on any single project is

Table 2. Top ten pharmaceutical companies in 2005 [3]

Rank	Company	Country	Pharmaceutical sales [US \$ 10^9]
1	Pfizer	USA	47.7
2	GlaxoSmithKline	United Kingdom	34.9
3	Sanofi-Aventis	France	30.4
4	Novartis	Switzerland	28.7
5	Johnson & Johnson	USA	25.4
6	AstraZeneca	UK/Sweden	24.2
7	Merck & Co	USA	23.6
8	Roche	Switzerland	19.9
9	Abbott	USA	15.7
10	Bristol-Myers S.	USA	14.8

Table 4. Top 10 branded products by sales in 2005 [3]

Product	Active ingredient	Company	Sales (US \$ 10^9)	Disease
Lipitor	Atorvastatin Calcium	Pfizer	12.9	high cholesterol
Plavix	Clopidogrel bisulfate	Bristol-Myers Squibb & Sanofi-Aventis	5.9	heart disease
Nexium	Esomeprazole	AstraZeneca	5.7	heartburn
Seretide / Advair	Salmeterol xinafoate + fluticasone propionate	GlaxoSmithKline	5.6	asthma
Zocor	Simvastatin	Merck & Co.	5.4	high cholesterol
Norvasc	Amlodipine besylate	Pfizer	5.0	high blood pressure
Zyprexa	Olanzapine	Eli Lilly	4.7	schizophrenia
Risperdal	Risperidone	Johnson & Johnson	4.1	schizophrenia
Ogastro / Prevacid	Lansoprazole	Abbott & Takeda	4.0	heartburn
Effexor	Venlafaxine HCl	Wyeth	3.8	depression

low. Once a compound has been found, lengthy development studies are required to prove to regulatory agencies that the compound is safe and effective for use (Chap. 4) and of high quality. The time from initiation of a new research project to the marketing authorization can easily take 12–15 years and consume US \$ 850×10^6. This business requires high research investment over a prolonged period of time and will see a long delayed return on investment. The risks of failure are high, but the rewards for a successful product can also be high. These medicines are promoted primarily to medical professionals, citing claims approved by regulatory agencies, although direct to consumer advertising is sometimes also being used.

The *generic pharmaceutical industry* is another part of the pharmaceutical industry. Products that are the subject of current valid patents can be manufactured and marketed only by the patent holders or by a licensee. When the patents have expired, after 20 years of patent life in most countries, other manufacturers are free to manufacture and market the product if approved by regulatory agencies. Companies can sell these off-patent products as "generics". The same active ingredient can be used in generic products, but not with the same trade name because the innovator will typically own that. The generic business involves little research since it relies upon the research of others, and concentrates only on commercially successful products whose patents have expired. These products are still sold on prescription, usually at prices lower than the innovator since generic companies do not have to recoup extensive R& D expenses.

Finally, there are "active pharmaceutical ingredient" (API) suppliers, which manufacture and supply active substances and excipients to the pharmaceutical industry.

The pharmaceutical industry also can be said to include wholesale distributors and retail outlets.

3. Legal Requirements and Guidance

3.1. General

In almost all countries throughout the world, the development of pharmaceutical products, their manufacture, and marketing depend on governmental approval. Pharmaceutical products are regarded as being potentially dangerous. A comprehensive national legal framework, therefore, exists to avoid risks in connection with the development, manufacture, distribution, marketing, and use of medicines.

This chapter gives an overview of internationally and generally accepted legal requirements. It is based on the main laws of the United States (US) and the European Union (EU). In the following, law means legal provisions adopted by the legislative organ (i.e., parliament), regulation means legal provisions made by an institution of the executive (i.e., the government) on the basis and within the limits of an authorization given by the law. Administrative regulations are governmental measures directly binding not the individual or a company but the competent authorities.

United States Law. The basic law in the US is the *Federal Food, Drug, and Cosmetic Act* (FFDCA), as amended, the latest amendment being the *Food and Drug Administration Modernization Act of 1997* (FDAMA). Many regu-

lations have been issued relating to drugs based on this Act. The main regulations are:

1) Investigational New Drug Application (IND) [21 Code of Federal Regulations (CFR) Part 312]
2) Applications for FDA Approval to Market a New Drug (NDA) (21 CFR Part 314)
3) Protection of Human Subjects (21 CFR Part 50)
4) Institutional Review Boards (21 CFR Part 56)
5) Good Laboratory Practice for Non-Clinical Laboratory Studies (21 CFR Part 58)
6) Current Good Manufacturing Practice for Finished Pharmaceuticals (GMP) (21 CFR Part 211)

European Union. The legal system of the EU differentiates between regulations and directives, both issued by the Council of Ministers, the European Parliament, and occasionally by the Commission. Regulations have direct and binding effect in the member states, whereas directives have to be transferred separately into national law. The main regulations and directives are:

1) Consolidated Directive 2001/83/EC of the European Parliament and of the Council of 6 November 2001 on the Community code relating to medicinal products for human use as amended by Directive 2002/98/EC, Directive 2004/24/EC and Directive 2004/27/EC (http://ec.europa.eu/-enterprise/pharmaceuticals/eudralex/vol-1/-consol_2004/human_code.pdf)
2) Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use (Official Journal L 121, 1/5/2001 p. 34 – 44)
3) Commission Directive 2003/94/EC of 8 October 2003 laying down the principles and guidelines of good manufacturing practice in respect of medicinal products for human use and investigational medicinal products for human use (Official Journal L 262, 14/10/2003 p. 22 – 26)
4) Commission Directive 2005/28/EC of 8 April 2005 laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products for human use, as well as the requirements for authorization of the manufacturing or importation of such products (http://ec.europa.eu/-enterprise/pharmaceuticals/eudralex/vol-1/-dir_2005_28/dir_2005_28_en.pdf)
5) Regulation (EC) No 726/2004 of the European Parliament and of the Council of 31 March 2004 laying down Community procedures for the authorization and supervision of medicinal products for human and veterinary use and establishing a European Medicines Agency (Official Journal L 136, 30/4/2004 p. 1–33)
6) Commission Regulation (EC) No 1084/2003 of 3 June 2003 concerning the examination of variations to the terms of a marketing authorization for medicinal products for human use and veterinary medicinal products granted by a competent authority of a Member State (Official Journal L 159, 27/6/2003 p. 1–23)
   Commission Regulation (EC) No 1085/2003 of 3 June 2003 concerning the examination of variations to the terms of a marketing authorization for medicinal products for human use and veterinary medicinal products falling within the scope of Council Regulation (EEC) No 2309/93 (Official Journal L 159, 27/6/2003 p. 24–45)

The most relevant guidance documents for marketing authorizations for medicinal products for human and veterinary use in the member states of the EU have been published by the European Commission under the title "The Rules Governing Medicinal Products in the European Union" in 9 volumes (http://ec.europa.eu/enterprise/pharmaceuticals-/eudralex/index.htm).

International Harmonization A very important attempt has been made to harmonize the scientific/technical standards defining the type and amount of studies necessary to confirm the quality, safety, and efficacy of new medicines. This attempt is called "International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use" (ICH). Since 1991, ICH has successfully developed and published numerous guidelines in a joint effort by regulators and industry experts from the European

Union, Japan, and USA (http://www.ich.org/-cache/compo/276-254-1.html). ICH guidelines do not legally bind the national authorities but are adopted and transferred into national guidelines in many countries both within and outside the ICH region.

3.2. Definitions

A *pharmaceutical product* is defined as the dosage form in the final immediate packaging intended for marketing (ICH Q1A).

A *dosage form* is a pharmaceutical product type (e.g., tablet, capsule, solution, cream) that contains an active substance generally, but not necessarily, in association with excipients (ICH Q1A).

An *active substance* is defined (Directive 2001/83/EC) as either

a) any substance or combination of substances presented as having properties for treating or preventing disease in human beings; or
b) any substance or combination of substances which may be used in or administered to human beings either with a view to restoring, correcting, or modifying physiological functions by exerting a pharmacological, immunological, or metabolic action, or to making a medical diagnosis.

An active substance can be any matter irrespective of origin, which may be:

- human, e.g., human blood and human blood products;
- animal, e.g., microorganisms, whole animals, parts of organs, animal secretions, toxins, extracts, blood products;
- vegetable, e.g., microorganisms, plants, parts of plants, vegetable secretions, extracts;
- chemical, e.g., elements, naturally occurring chemical materials and chemical products obtained by chemical change or synthesis.

Quality is the suitability of either an active substance or a pharmaceutical product for its intended use (ICH Q8).

Safety means that, in the light of currently prevailing scientific knowledge, the pharmaceutical product has no harmful effects exceeding medical acceptability.

Efficacy means a reasonable expectation that, in a significant proportion of the target population, the pharmacological effect of the pharmaceutical product, when used under adequate directions for use and warnings against unsafe use, will provide clinically significant relief of the type claimed [21 CFR 330.10].

Side effects mean the occurrence of undesired reactions.

3.3. Requirements Relating to the Development of Pharmaceutical Products

Before a pharmaceutical product can be put on the market, various prerequisites have to be fulfilled notwithstanding the necessary final marketing authorization. These prerequisites relate to the testing of the pharmaceutical product. The performance of testing is not within the free discretion of the pharmaceutical company or the investigator. Internationally accepted standards and methods have to be applied. The testing program is divided into chemical-pharmaceutical development (see Section 4.3), pharmacological and toxicological studies (see Section 4.1), and clinical trials (see Section 4.2).

3.4. Marketing Authorization

The placing of pharmaceutical products on the market needs governmental approval as defined in relevant national laws. Such an approval generally needs proof of quality, safety, and efficacy of the pharmaceutical product. The competent authority has to give its approval on the basis of evidence provided during the development phases. Within the EU, a distinction has to be made between two procedures: a centralized procedure, which exclusively applies for medicinal products, derived from biotechnology and for new active substances. A marketing authorization granted under the centralized procedure is valid for the entire community market, which means the medicinal product may be put on the market in all EU member states. In addition, the procedure of mutual recognition exists meaning that normally a member state is bound by an approval of another member state in the sense that it recognizes that approval as basis for its own

identical decision. The approval, however, remains national (The Rules governing Medicinal Products in the European Union, Volume 2A, Notice to Applicants – Procedures for marketing authorization: http://ec.europa.eu/enterprise/-pharmaceuticals/eudralex/homev2).

If minor changes are made to the documentation already submitted to the health authority, the applicant has to notify the authority of these changes. Major changes (e.g., dosage, duration of application, limitation of contraindications, side effects, or interactions) have to be approved by the authority.

The competent authority is entitled to withdraw, revoke, or suspend the authorization given at any time if there are special reasons relating to the quality, especially to the safety or to the efficacy of the pharmaceutical product.

Unless previously prolonged, the marketing authorization expires in some countries after a defined period.

3.5. Requirements for the Manufacture and Quality Control of Pharmaceutical Products

The manufacture of pharmaceutical products may itself involve risks. The quality of the product has to be identical to that accepted by the authority responsible for the marketing authorization. Consequently, the manufacturer needs special governmental authorization to produce pharmaceutical products. The requirements to be met by the applicants are suitable and sufficient premises, technical equipment, and control facilities as well as, according to EU regulations, at least one appropriately qualified person responsible for production, i.e., the production director. The qualified person is usually a pharmacist, but chemists, biologists, physicians, and veterinarians are also admitted provided that they have acquired additional knowledge and experience specified in the relevant law. The qualified person has to be permanently present during formulation.

An additional, very important requirement for the manufacturer of pharmaceutical products is observance of the rules of Good Manufacturing Practice (GMP). These rules relate to quality assurance system, personnel, premises, equipment, documentation, production, quality control, contract manufacture and analysis, complaints, product recall, and self-inspection. These principles are laid down in Volume 4 of the Rules Governing Medicinal Products in the European Union "Medicinal Products for Human and Veterinary Use: Good Manufacturing Practice" (http://-ec.europa.eu/enterprise/pharmaceuticals/-eudralex/homev4.htm), and for the US in 21 CFR Part 210 (http://www.access.gpo.gov/-nara/cfr/waisidx_06/21cfr210_06.html) and 211 (http://www.access.gpo.gov/nara/cfr/-waisidx_06/21cfr211_06.html).

A GMP guideline for active substances has been developed by ICH: Good Manufacturing Practice Guide for Active Pharmaceutical Ingredients – Q7 (http://www.ich.org/-LOB/media/MEDIA433.pdf)

The manufacturing, control, storage, packaging, and marketing of pharmaceutical products are subject to supervision and inspection by the competent authority. This authority varies from country to country; it may be a central governmental authority or the authority of a federal state. The manufacturer has to give the inspectors of the authority access to the premises in which pharmaceutical products are manufactured. This also holds for the manufacture of active substances. The development of pharmaceutical products and their clinical testing are also supervised by the competent authority, which has to check whether the legal requirements for clinical trials have been met. The law gives the competent authority various authorizations to ensure adherence to the laws and regulations.

3.6. Requirements for Placing a Pharmaceutical Product on the Market

The most important legal prerequisite for placing a pharmaceutical product on the market is the marketing authorization. However, many other stringent provisions also have to be observed. Most pharmaceutical products can only be sold to the patient by a pharmacy. This applies to pharmaceutical products prescribed by a physician and to non-prescription (over-the-counter, OTC) products. The OTC products are generally recognized as safe and effective and not misbranded if certain conditions are met, especially

with respect to any applicable monograph (21 CFR Part 330). Some information has to be declared on the package (e.g., name and address of the manufacturer, name of the pharmaceutical product, registration number, batch number, type and amount of active ingredients, expiry date). The most important source of information for the patient, however, is the package insert, which contains instructions for use and a description of indications, contraindications, side effects, and interactions. Comprehensive scientific written information must also be available in written form for doctors and physicians. If the pharmaceutical company wants to promote the pharmaceutical product, compulsory information on the product has to be given. This information is almost identical to that contained in the package insert.

All these and additional information instruments are designed to avoid risks in connection with the administration of pharmaceutical products, and to guarantee maximum safety. The competent authority is entitled and authorized to control and inspect the company to make sure that it complies with the law.

An additional legal obligation of the pharmaceutical company is to report all possible side effects of the pharmaceutical product, whether previously known or unknown, to the competent authority without undue delay. Every pharmaceutical company has to have a special safety monitor whose function is to collect and to evaluate reports on side effects and to coordinate the necessary measures. The safety monitor is personally responsible for the fulfillment of the notification obligations toward the health authority and must be a physician, veterinary physician, or pharmacist with at least two years professional experience. The name of the person must be notified to the competent authority. Another specialist to be appointed by the pharmaceutical company has to guarantee for the authenticity of the product information used with the data contained in the registration file.

3.7. Liability

If, as a result of the administration of a pharmaceutical product, a person is killed or the body or the health of a person is considerably injured, then the pharmaceutical company that has placed the pharmaceutical product on the market is obliged to compensate for the harm caused by the injury. This liability is independent of fault and not imposed on the manufacturer, but on the company placing the product on the market. Liability only exists if the pharmaceutical product, despite correct use, produces harmful effects that exceed the bounds considered justifiable in the light of scientific knowledge available in medical science and which have their origin in the development of the pharmaceutical product or the manufacturing processes, or if damage has occurred as a result of formulation or instructions for use that do not comply with the knowledge available in medical science. In general liability insurance is compulsory to cover the potential financial risk.

Finally many penal provisions and administrative fines complete the instrumentarium of the law; in many cases violation of the law leads to personal responsibility and liability.

3.8. International Non-Proprietary Names (INN)

International Non-Proprietary Names (INN) were introduced in 1950 for the purpose of creating an unambiguous single non-proprietary name for active substances worldwide. (WHO Drug Information vol. 13, no. 1, 1999, pp. 13 – 14). The World Health Organization (WHO) is in charge of selecting a single name, which is acceptable all over the world for each active substance, which is to be marketed as a pharmaceutical product. Trademarks for medicines are not allowed to be similar to INNs so as not to confuse the patients or doctors.

In order to select an INN, a request is usually submitted to the WHO on an INN request form. In certain countries where national nomenclature commissions exist, this is done through the national nomenclature authority, e.g., in the United States it is the United States Adopted Names (USAN) Council. Precise information on the substance's chemistry, pharmacological action and therapeutic use must be provided. The proposed non-proprietary names and the name and address of the manufacturer are also entered in the form.

The names proposed by the originator are examined and one name is selected in accor-

dance with the "Procedure for the selection of recommended international non-proprietary names (INN) for pharmaceutical substances" and the "General principles for guidance in devising INNs". All the members of the WHO Expert Panel on International Pharmacopoeia and Pharmaceutical Preparations designated to select non-proprietary names have to agree to the name selected. Then it is first published as a *proposed INN*. During a four-month period, any person can forward comments, or lodge a formal objection to a name, e.g., on grounds of similarity with a trade name. If no objection is raised, the name will be published a second time as a *recommended INN*.

About 8000 INNs have been published to date (2006) in the "WHO Drug Information" journal. The names are given in English, French, and Spanish, as well as Latin. A cumulative list of all INNs is published periodically and includes INNs in Russian, as well as references to other common names (http://www.who.int/medicines/services/inn/en/).

4. Development of Pharmaceutical Products

4.1. Preclinical Safety/Toxicology

Preclinical disciplines as pharmacology, pharmacokinetics, and toxicology contribute to the evaluation of efficacy and safety of pharmaceutical compounds. There is no absolute safe administration of pharmaceutical drugs possible, but the experimental studies of toxicology facilitate to find the right path between safety issues for human use.

The toxicologist designs in vitro and in vivo studies to support the identification of unexpected and undesirable adverse effects of pharmaceutical or chemical compounds; the hazards should be described, the type and severity of toxicity should be analyzed, and the possible risks involved assessed. These objectives in toxicology help clinicians to design their clinical trials optimally, to use preclinical recommendation which biomarkers should be applied for human conditions and therefore support to manage the risks in humans when participating in clinical trials or later as patients after market authorization. An additional aim today in toxicology is to investigate the mechanisms behind changes observed, e.g., to explain why alterations appear in animals and possibly not in humans. This knowledge is relevant for doctors and facilitates the approval process by regulatory reviewers.

To identify risks and for using compounds safely, a great number of in vitro and/or in vivo experiments needs to be conducted. Test strategies for toxicological investigations have been refined and the extrapolation of preclinical results to humans improved. The prevention of toxic events in man is constantly the main goal in preclinical areas.

Today, toxicology is complex, expensive, and time-consuming. In former decades, considerable regulatory differences existed among different regions which led to unnecessary repetitions of studies without improving the safety profiles but increasing often unnecessarily the number of animals being used. Accordingly, public resistance and concerns among researchers increased the desire to harmonize test procedures and recommendations worldwide. Therefore, the International Conferences of Harmonization (ICH) were founded in 1989, which focus on the creation of guidelines to harmonize the testing procedures on an international basis.

Herein, the detailed requirements for preclinical studies to move the development of compounds into human clinical trials and finally to submit an approvable dossier to the regulatory agencies is described. The ICH guidelines outline the basic principles. For pragmatic recommendations when designing individual studies, additional guidelines like those published by the European Medicines Agency (EMEA) of the Food and Drug Administration (FDA) or WHO need to be considered too.

4.1.1. Preclinical Safety

The international discussions among experts have led to constructive solutions and recommendations, which are published in the ICH Guidelines (Table 5). These guidelines have been implemented practically worldwide, they provide excellent basic information, and stimulate new thinking to improve strategies or techniques, all with one objective: to be more predictive in regard to safety for human conditions.

Table 5. guidelines for the area of preclinical safety

ICH No.	Title	CPMP Doc. No.	Step
S1A	The Need for Carcinogenicity Studies of Pharmaceuticals	CPMP/ICH/140/95	Step 5
S1B	Carcinogenicity: Testing for Carcinogenicity of Pharmaceuticals	CPMP/ICH/299/95	Step 5
S1C	Carcinogenicity: Dose Selection for Carcinogenicity Studies of Pharmaceuticals	CPMP/ICH/383/95	Step 5
S1C (R)	Dose Selection for Carcinogenicity Studies of Pharmaceuticals: Addition of a Limited Dose and related Notes	CPMP/ICH/366/96	Step 5
S2A	Genotoxicity: Guidance on Specific Aspects of Regulatory Genotoxicity Tests for Pharmaceuticals	CPMP/ICH/141/95	Step 5
S2B	Genotoxicity: A Standard Battery for Genotoxicity Testing of Pharmaceuticals	CPMP/ICH/174/95	Step 5
S3A	Toxicokinetics: A Guidance for Assessing Systemic Exposure in Toxicology Studies	CPMP/ICH/384/95	Step 5
S3B	Pharmacokinetics: Guidance for Repeated Dose Tissue Distribution Studies	CPMP/ICH/385/95	Step 5
S4A	Duration of Chronic Toxicity Testing in Animals (Rodent and Non-Rodent Toxicity Testing)	CPMP/ICH/300/95	Step 5
S5A	Reproductive Toxicology: Detection of Toxicity to Reproduction for Medicinal Products	CPMP/ICH/386/95	Step 5
S5B (M)	Reproductive Toxicology: Toxicity on Male Fertility (Modification)	CPMP/ICH/136/95	Step 5
S6	Preclinical Safety Evaluation of Biotechnology-Derived Products	CPMP/ICH/302/95	Step 5
S7A	Safety Pharmacology Studies for Human Pharmaceuticals	CPMP/ICH/539/00	Step 5
S7B	Safety Pharmacology Studies Assessing the Potential for Delayed Ventricular Repolarization (QT Interval Prolongation) by Human Pharmaceuticals	CPMP/ICH/433/02	Step 5
S8	Immunotoxicity		Step 5

Table 6 summarizes the guidelines which have been agreed upon multidisciplinary subjects. This multidisciplinary guidance helps to find useful concentrated information about what to do during which timeframe to decide which studies should be conducted when and finally where in the dossier to report them.

4.1.2. Toxicological Strategies

In general, investigations are done always as early as possible, but compound availability limits these objectives. In reality, strategies follow the principle of a step-by-step procedure. When small amounts of compounds are available then the safety of those people being exposed first with the drug (in the chemical laboratories) is in the foreground. Accordingly, short-term tests to analyze the toxicity and exclude genetic damages are being conducted very early on and only later when grams and kilograms of compound are available, more complex and long-term studies such as carcinogenicity studies are being performed. This strategic principle is not followed in the sequence of issues being dealt with within the ICH process: here the most urgent issues were discussed first and guidelines finalized.

The requirements for the first step, the New Investigational Drug (IND), to move the drug from preclinical activity into the clinics are:

- Toxicokinetics/pharmacokinetics

Table 6. Guidelines for multidisciplinary subjects

ICH No.	Title	CPMP Doc. No.	Step
M1	Medical Terminology: Medical Dictionary for Regulatory Activities Terminology (MedDRA)		
M2	Recommendations on Electronic Transmission of Individual Case Safety Reports Message Specification	CPMP/ICH/285/95	Step 5
M3 (M)	Non-Clinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals (Modification)	CPMP/ICH/286/95	Step 5
M4	Common Technical Document for the Registration of Pharmaceuticals for Human Use	CPMP/ICH/2887/99	Step 4

- Short-term toxicity studies (single- and up to four-week studies)
- Local tolerance
- Genotoxicity in vitro (in vivo)
- Phototoxicity

The recommendations for the final process, the New Drug Application (NDA), in order to gain the market authorization include

- single- and repeated-dose toxicity studies
- reproduction toxicity studies
- genotoxicity studies
- assessment of carcinogenic potential
- local tolerance studies
- pharmacokinetic (absorption, distribution, metabolism, excretion; ADME) studies
- immunotoxicology studies
- studies on special populations, e.g., children

4.1.3. ICH Guidelines for Carcinogenicity

Guideline on the Need for Carcinogenicity Studies of Pharmaceuticals (ICH S1A). The objectives of carcinogenicity studies are:

- To identify a tumorigenic potential in animals
- To assess the relevant risk for humans.

Any cause for concern derived from laboratory investigations, animal toxicology studies, and data in humans may lead to a need for carcinogenicity studies.

Carcinogenicity studies should be performed for any pharmaceutical when the expected clinical use is continuous for at least six months or repeatedly in an intermittent manner. Examples of such conditions include allergic rhinitis, depression, and anxiety.

Pharmaceuticals administered infrequently or for short duration of exposure (e.g., anesthetics and radiolabel imaging agents) do not need carcinogenicity studies unless there is cause for concern, e.g., previous demonstration of carcinogenic potential in the product class or long-term tissue retention of parent compound or metabolite(s) resulting in local tissue reactions or other pathophysiological responses. Further, no cancerogenicity studies are needed for unequivocally genotoxic compounds, because it is assumed that such compounds will be positive for the cancerogenic potential also and these characteristics do not need to be proven. Further, pharmaceuticals showing poor systemic exposure from topical routes in humans may not need studies by the oral route to assess the carcinogenic potential to internal organs. Finally, carcinogenicity studies are not generally needed for endogenous substances, when given essentially as replacement therapy (i.e., physiological levels).

Testing for Carcinogenicity of Pharmaceuticals (ICH S1B). The regulatory requirements to assess the carcinogenic potential was the conduct of long-term carcinogenicity studies in two rodent species, usually the rat and the mouse. It was the mission of ICH to examine whether this practice could be reduced without compromising human safety for new compounds.

The results of a number of databases forced to rethinking of the historical strategies and led to the concept of a new paradigm as a new experimental approach to testing for carcinogenic potential. Thus, the basic principle for testing the carcinogenic potential is

- one long-term rodent carcinogenicity study plus

- one short- or medium-term study that supplements the long-term carcinogenicity study and provides additional information not readily available from the long-term assay.

Additional tests may include either short- or medium-term in vivo rodent test systems, in most cases using the mouse. These models of carcinogenesis may use transgenic or neonatal rodents. The ICH guideline includes also models of initiation/promotion in rodents; these models are considered today as useful models for hepatic carcinogenesis or adequate mechanistic studies but not as assays appropriate as general screen for drug-induced carcinogenesis.

A long-term carcinogenicity study in a second rodent species (e.g., mice) is still considered acceptable.

In the "Notes", the guideline describes some important information about the new models, e.g., Note 2 sets the conditions to limit testing with one species only. If the findings of a short- or long-term carcinogenicity study, genotoxicity tests, and other data would indicate that a pharmaceutical clearly poses a carcinogenic hazard to humans, usually a second carcinogenicity study would not be useful.

Note 3 provides details about the short- or midterm models. Representative approaches for testing the carcinogenic potential that may meet these criteria are:

- Several transgenic mouse assays including
  - p53+/− deficient model
  - Tg.AC model
  - TgHras2 model
  - XPA deficient model, etc.
- The neonatal rodent tumorigenicity model

According to their mechanism, these new models are characterized as follows:

- *Activated oncogenes*
  - TgHras2 model (Japan)
  - Tg.AC skin model (USA); also gavage
- *Inactivated tumor suppressor gene*
  - p53 "knock out" (=p53+/−) model (USA)
- *Inactivated DNA repair gene*
  - XPA-/- (NL)

Evidence of tumorigenic effects of the drug in rodent models should be evaluated in the light of the tumor incidence and latency, the pharmacokinetics of the drug in the rodent models as compared to that in humans, and data from any ancillary or mechanistic studies that are informative with respect to the relevance of the observed effects to humans.

The results from any test cited above should be considered as part of the overall "weight-of-evidence" evaluation, which takes into account the scientific status of the test systems.

These new models have been accepted by regulatory agencies. The assessment of the FDA of the new models in respect to the assay selection is as follows:

p53+/−	if clearly or equivocally genotoxic
TgAC	for dermally applied products
TgRasH2	for genotoxic or nongenotoxic products
Neonatal	if clearly or equivocally genotoxic

The assessment of the EMEA of the new models (EU: evaluation in the Safety Working Party of EMEA) is as follows:

- p53 and Tg.RasH2 are equally sensitive to genotoxic compounds (some false positives and false negatives)
- TgRasH2 is more sensitive to peroxisome proliferators
- p53 and TgRasH2 are acceptable in a regulatory context

Dose Selection for Carcinogenicity Studies of Pharmaceuticals (ICH S1C). The maximally tolerated dose or minimum toxic dose (MTD) method is still the standard method for high-dose selection.

The MTD is generally chosen based on data derived from toxicity studies of three months duration. The testing options for dose-range finding studies are:

- Usually:
  - 3 months for long-term studies
  - 1 month for neonatal- or transgenic-mice studies
- Range of different dose levels, often five
- Focus on toxicity endpoints, determination of MTD
- Profiling of the Area Under the Concentration-versus-Time Curve (AUC)
  - e.g., 1 and 3 months for rats
  - e.g., 1 and 4 weeks for alternatives
  - transgenic = wild type

The long-term carcinogenicity studies are designed as follows:

Species:	rats, mice, rarely hamster
Duration:	24 months rats, 24 months mice
Route:	as in clinical conditions, in feed, water, by gavage, inhalation
Dose levels:	mostly 3, mostly with a factor (e.g., 1:3:9)
Animal numbers	50/group/sex
Results available	prior to NDA
Costs	€ $1.0–1.5 \times 10^6$

The selected dose levels for rodent bioassays should provide a high enough exposure to the agent, and an adequate margin of safety. No significant physiological dysfunction should be induced and the toxicity should be compatible with two years survival. The criteria of choosing the high dose are:

- MTD
- 25-fold AUC ratio (rodent:human)
- Dose-limiting pharmacodynamic effects
- Saturation of absorption
- Maximum feasible dose

The MTD is the most frequent criteria used for the top dose. A minimum toxic effect over the course of the carcinogenicity study is the objective. Such MTD is defined, e.g., by no more than 10 % decrease in body weight gain relative to that of controls.

The second alternative for relatively well tolerated drugs is a systemic exposure representing a 25-fold multiple of the human AUC. Saturation of absorption measured by systemic availability of drug-related substances is also acceptable as a limit dose. Finally, the pharmacodynamic effects may limit the high-dose selection. If a compound, for example, reduces the blood glucose level, then exaggeration is detrimental.

The maximum feasible dose by dietary administration was considered 5 % of diet, but this amount of drug was considered to be too high. Therefore, a new and more reasonable solution was formulated in the Guideline ICH S1C(R).

Addendum to Dose Selection for Carcinogenicity Studies of Pharmaceuticals: Addition of a Limit Dose and Related Notes of Pharmaceuticals (ICH S1C(R)). In determining the high dose for carcinogenicity studies, it may not be necessary to exceed a dose of 1500 mg/kg/d. This limit dose applies only in cases where there is no evidence of genotoxicity and where the maximum recommended human dose does not exceed 500 mg/d.

4.1.4. ICH Guidelines for Genotoxicity (ICH S2A and ICH S2B)

The Genotoxicity Guidelines ICH S2A and ICH S2B (Table 7) deal with the avoidances of DNA damage. Compounds used for human treatment should not induce tumors or teratogenicity.

The recommendations help to conduct studies with acceptable meaningful approaches. It is requested to provide good information on the potential to induce gene mutations, structure chromosome aberrations (clastogenicity), and numerical chromosome aberrations (aneugenicity).

No single test is capable of detecting all relevant genotoxic agents and therefore, a battery of tests is considered appropriate.

The Standard Battery consists of three assays:

1) Test for gene mutations in bacteria (Ames)
2) An in vitro test with cytogenetic evaluation of chromosomal damage with mammalian cells *or* an in vitro mouse lymphoma *tk* assay
3) An in vivo test for chromosomal damage using rodent hematopoietic cells.

The in vitro tests are carried out with and without extrinsic metabolic activation, usually provided by a rat liver S9-mix.

The ICH Guidelines inform about the top concentrations for in vitro tests and the expected levels of cytotoxicity, about repeat or confirmatory testing, and the proof of target tissue exposure to the test article in the in vivo studies. Support can be found for the interpretation of positive test results and considerations on conditions leading to false positive data.

4.1.5. ICH Guidelines for Toxicokinetics/Pharmacokinetics (ICH S3A and ICH S3B)

With the data of toxicokinetics, one describes the systemic exposure achieved in animals and humans and its relationship to dose level and the time course of the toxicity study. The exposure is related to the toxicological findings and to the relevance of these findings related to the clinical safety.

The kinetic information of the parent compound and its metabolites contribute to the extrapolation of safety from animals to humans.

Table 7. Guidelines for genotoxicity aspects

ICH No.	Title	CPMP Doc. No.	Step
S2A	Genotoxicity: Guidance on Specific Aspects of Regulatory Genotoxicity Tests for Pharmaceuticals	CPMP/ICH/141/95	Step 5
S2B	Genotoxicity: A Standard Battery for Genotoxicity Testing of Pharmaceuticals	CPMP/ICH/174/95	Step 5

Species differ considerably in regard to their kinetic conditions, measured as Cmax, Tmax, AUC, $t\frac{1}{2}$, and ADME.

As far as possible, kinetic analyses should be integrated into preclinical safety studies. Therefore, the provision of kinetic data should be considered in the following study types: repeat-dose toxicity studies, genotoxicity in vivo studies when there are differences between in vitro and in vivo results, possibly reproductive toxicity studies, and cancerogenicity studies where monitoring is recommended to be appropriate but not beyond six months duration. It is important to demonstrate exposure and, e.g., inhibition or accumulation. The latter may explain toxicity at the site of increased compound accumulation.

Single-dose studies provide usually sufficient information about tissue distribution but there may be cases where assessments after repeated dosing may provide better information. Such studies are necessary when

- single-dose distribution studies suggest that the half-life of the test compound and/or metabolites in organs or tissues significantly exceeds the half-life of the elimination phase in plasma;
- steady-state levels of a compound/metabolite in the circulation, determined in repeated dose pharmacokinetic or toxicokinetic studies, are markedly higher than those predicted from single-dose kinetic studies;
- histopathological changes were observed that were not predicted from short-term toxicity studies.

This information is provided in the ICH guideline S3B: "Pharmacokinetics: Guidance for Repeated-Dose Distribution Studies (CPMP/ICH/395/95)".

4.1.6. ICH Guidelines for Reproductive Studies (ICH S5A and ICH S5B)

The special toxicology discipline "Reproductive and Developmental Toxicity" focuses on undesirable adverse effects on male and female fertility, birth defects (developmental toxicity, malformations, teratogenicity), and nonphysiological changes that appear shortly before, during, and after birth and during the weaning period.

In general, there is a distinction between the following time periods:

Segment I: Fertility and early embryonic development to detect drug-induced toxic effects before mating, through mating, and implantation (fertilized egg or zygote, prior to implantation, and prior to formation of the three primary germ layers). This includes changes in estrous cycle, tubular transport, determination of number of corpora lutea and implantation sites, live and dead conceptuses, etc.

Segment II: The embryo–fetal development studies (embryo during major organ development, i.e., organogenesis, malformations, fetal growth and functional deficits, and postembryonic period).

Segment III: The prenatal and postnatal development (neonate or postnatal offspring) including maternal function such as gestation or lactation through weaning. On the other hand, such effects may appear late in development and therefore observations are continued through sexual maturity. Females are allowed to deliver and rear offspring to weaning at which time one male and one female offspring per litter should be selected for rearing to adulthood and mating to assess reproductive competence. The most important investigations are: necropsy of all adults, histological evaluation of all organs with macroscopic findings, live offspring at birth, dead offspring at birth, body weight at birth, peri- and

postweaning survival, physical development, sensory functions and development, and behaviour.

These segments can be tested separately or in a combined manner. All stages of development from conception to maturity and the detection of acute and delayed effects of exposure through one complete life cycle should be examined. The standard species are rodents, rats as the preferred rodent species for all study types, and the rabbit as the second nonrodent species for the embryo toxicity studies. Rarely, mice or monkeys are being used too if special conditions, usually kinetic data, justify such species. The route of administration should be similar to the intended human usage.

4.1.7. Preclinical Safety Evaluation of Biotechnology-Derived Products (ICH S6)

This guidance is applicable for recombinant DNA proteins, vaccines, peptides, plasma derived products, endogenous proteins extracted from human tissues, or oligonucleotide drugs, etc., while, e.g., heparin, vitamins, and cellular blood components are not covered.

For biotechnology products, high flexibility for the development is recommended and steps should be reconsidered on a case-by-case basis.

Usually, two species should be used, but when the biological activity is well understood or when in short-term toxicity studies the effects were similar in both species, then the long-term testing could be restricted to one species only. When no relevant species can be identified, then one should consider the use of homologous proteins or transgenic animals.

Recommended Types of Safety Studies *Safety Pharmacology*, i.e., testing of the vital function such as cardiovascular, respiratory, and central nerve system functions is recommended, but S6 mentions also renal function, which from today's viewpoint would not be necessary before first administration of a compound in humans.

Some information on *absorption*, *disposition*, and *clearance* in the animal models is desirable before clinical trials and systemic exposure of the compound should be monitored, as well as the appearance of antibodies and their ability to neutralize the intended effect.

The *immune response* could also alter the pharmacokinetic or -dynamic effects. Anaphylactic responses tested in the guinea pig is not predictive for humans and therefore not necessary. The same holds true for the standard testing batteries for immunotoxicity, these are not recommended.

Studies with a duration of two weeks up to three months are often sufficient and only with chronic use in humans six-months studies should be considered. Flexibility may allow a further reduction of the traditional testing program for the reproductive endpoints. Such studies may not be necessary if a new compound shows similar effects to related well-known compounds. The same holds true for genotoxicity studies: the standard studies are not appropriate but testing on a case-by-case basis of impurities or promoter studies might be helpful.

Long-term carcinogenicity studies are usually not appropriate, but when there is cause for concern then studies with a single rodent species are sufficient. A cause for concern may have appeared in general toxicity studies, or a stimulation of the growth of normal as well as malignant cells can be assumed.

4.1.8. Safety Pharmacology Studies for Human Pharmaceuticals (ICH S7A)

The guideline ICH S7A requires testing of the vital functions usually in single-dose studies with focus on the *Core Battery*:

- Investigation of the effects of a test substance on vital functions
  - Central nervous system
  - Cardiovascular system
  - Respiratory system
  - Other systems as appropriate
- Exclusion of a system or function should be justified

and additional assays when needed:

- Follow-up studies, which provide a greater depth of understanding or additional knowledge compared to those studies provided by the Core Battery (e.g., mechanistic studies)
- Supplemental studies, which evaluate effects of the test substance on systems which are not addressed by the Core Battery, when there

is cause for concern not addressed elsewhere (e.g., in toxicology)

Timing of Studies The Core Battery is usually required before the first administration to humans and follow-up or supplemental studies are required when supporting safety, e.g., during clinical development, to clarify observed or suspected undesirable effects in animals or humans. Before the approval, the effects on all organ systems, if not covered elsewhere (e.g., in toxicology), should be addressed.

Good Laboratory Practice (GLP) As in all safety studies, GLP should be applied to the Core Battery, while follow-up and supplemental studies should be conducted according to GLP as far as possible and deviation should be justified and their impact discussed. Primary and secondary pharmacodynamic studies need not to be conducted according to GLP.

Special focus is given to the cardiovascular system. According to ICH S7A, blood pressure, heart rate, and electrocardiograms should be assessed for the Core Battery. Further in vivo, in vitro, and/or ex vivo evaluations should be considered, including methods for repolarization and conductance abnormalities.

4.1.9. Nonclinical Studies for Assessing Risk of Repolarization-Associated Ventricular Tachyarrhythmia for Human Pharmaceuticals (ICH S7B)

This text was finalized at a time when the details of ICH S7B were not yet outlined. With the increase of awareness that noncardioactive drugs, used for sometimes nonlife threatening diseases, can cause QT prolongation and serious dysrhythmias such as Torsade de Pointes, there appeared the need for further precision. Accordingly, the ICH Expert Working Group created ICH/S7B.

The QT interval (time from the beginning of the QRS complex to the end of the T wave) of the electrocardiogram (ECG) is a measure of the duration of ventricular depolarization and repolarization.

QT interval prolongation can be congenital or acquired (e.g., pharmaceutical-induced). When the QT interval is prolonged, there is an increased risk of ventricular tachyarrhythmia, including Torsade de pointes, particularly when combined with other risk factors (e.g., hypokalemia, structural heart disease, bradycardia. To cope with these issues, a relatively flexible approach is given with the ICH S7B nonclinical testing strategy as seen in Figure 1.

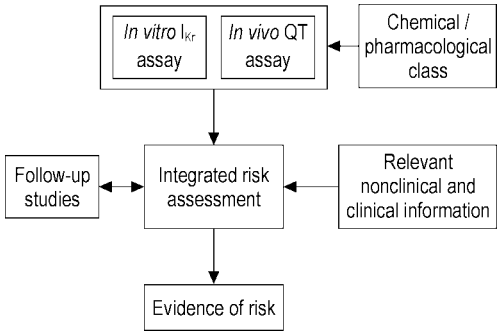

Figure 1.

The basis for the integrated risk assessment are in vitro and in vivo assays, supported by any knowledge about the chemical/pharmaceutical class.This first risk assessment may later be modified when results from follow-up studies or relevant nonclinical or clinical information become available. The evidence of risk summarizes the preclinical evaluation of the proarrhythmic potential for the clinicians.

In parallel with the development of ICH S7B, a clinical guideline (ICH E14) was drafted and reached Step 4 also in 2005. During the discussion between these two expert groups, the question was often raised whether toxicologists could exclude any risk for QT prolongation for humans with their testing strategies. Of course, they cannot because clinicians cannot exclude any of these risks for future patient generations based on their clinical trial results only. In addition, the FDA is currently comparing preclinical data with clinical results and a few cases could be identified already where QT prolongation was observed under clinical conditions while the preclinical tests were negative. This discrepancy is the basis for the diplomatic text in regard to the need for availability of S7B QT studies:

- Timing of S7B Non-Clinical Studies and Integrated Risk Assessment in Relation to Clin-

ical Development / Step 4, June 2005, Brussels: Conduct of S7B non-clinical studies assessing the risk for delayed ventricular repolarization and QT interval prolongation prior to first administration in humans should be considered.

- These results, as part of an integrated risk assessment, can support the planning and interpretation of subsequent clinical studies.

The term "should be considered" opens all flexibility to carry out the studies before it is done for the first time in humans or at a later stage of development. In practice, in most cases these studies are available before IND, because the industry wants to provide best safety to volunteers and patients. However, in conclusion is has to be stated that ICH/S7B proposes a series of nonclinical tests which are believed to predict the potential that a compound may prolong cardiac repolarization in vivo, in animals, and in humans.

Currently, these data have little impact on the clinical development proposals which are included in the draft guideline E14.

4.1.10. Immunotoxicology Studies (ICH S8)

The guideline ICH S8 is restricted to unintended immunosuppression and immunoenhancement, excluding allergenicity or drug-specific autoimmunity. The guideline applies to new pharmaceuticals intended for use in humans as well as to marketed drug products proposed for different indications or other variations on the current product label, but does not apply to biotechnology-derived pharmaceutical products covered by ICH S6 and other biological pharmaceuticals.

Immunosuppression or -enhancement can be associated with two distinct groups: 1) drugs intended to modulate immune function for therapeutic purposes (e.g., to prevent organ transplant rejection) where adverse immunosuppression can be considered exaggerated pharmacodynamics and 2) drugs not intended to affect immune function but cause immunotoxicity due to, for instance, necrosis or apoptosis of immune cells or interaction with cellular receptors shared by both target tissues and nontarget immune system cells. All new human pharmaceuticals should be evaluated for the potential to produce immunotoxicity.

Methods include standard toxicity studies (STS) and additional immunotoxicity studies conducted as appropriate. Whether additional immunotoxicity studies are appropriate should be determined by a weight-of-evidence review of cause(s) for concern. The immunotoxicity evaluation follows a straight-forward strategy which is detailed in Figure 2 and 3.

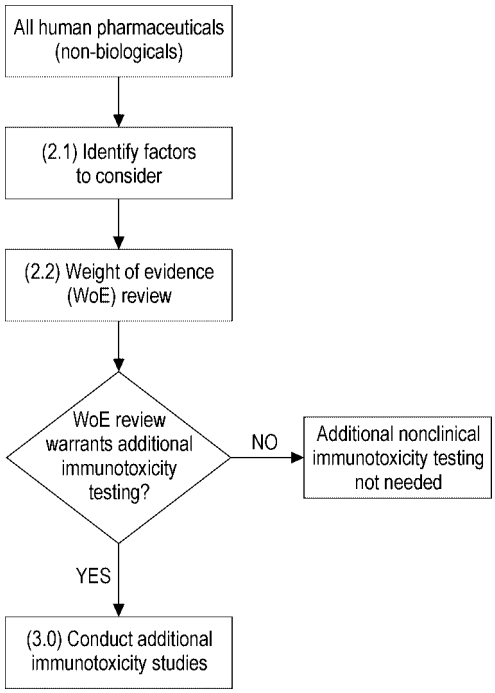

Figure 2.

The assessment of immunotoxicity should include the statistical and biological significance of the changes, the severity of the effects, dose–exposure relationship, a safety factor above the expected clinical dose, treatment duration, number of species and endpoints affected, and changes that may occur secondarily to other factors (e.g., stress), etc.

Additional immunotoxicity testing should be considered:

- If the pharmacological properties of a test compound indicate the potential to affect immune function (e.g. anti-inflammatory drugs).

- If the majority of the patient population for whom the drug is intended is immune-compromised by a disease state or concurrent therapy.
- If a compound is structurally similar to compounds with known immunosuppressive properties.
- If the compound and/or its metabolites are retained at high concentrations in cells of the immune system.
- If clinical findings are suggestive of immunotoxicity in patients exposed to the drug.

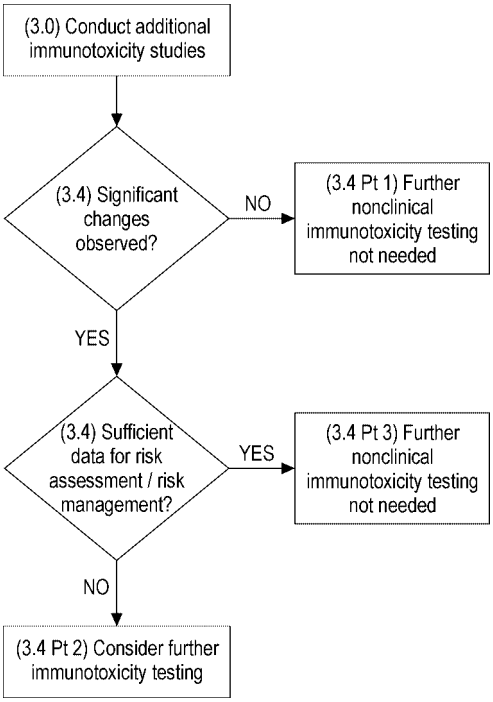

Figure 3.

If the weight-of-evidence review indicates that additional immunotoxicity studies are needed, then there are a number of assays which can be used, such as a T-cell dependent antibody response (TDAR). If specific cell types are affected in STS not involving cells participating in a TDAR, assays that measure function of that specific cell type might be conducted. Immunophenotyping of leukocyte populations, a nonfunctional assay, may be conducted to identify the specific cell populations affected and may provide useful clinical biomarkers.

Studies with 28 consecutive daily doses in rodents are generally accepted. Adaptations of immunotoxicity assays have been described using nonrodent species. The species, strain, dose, duration, and route of administration used in immune function assays should be consistent with the nonclinical toxicology study in which an adverse immune effect was observed.

Usually both sexes should be used in these studies, excluding nonhuman primates. The high dose should be above the No Observed Adverse Effect Level (NOAEL) but below a level inducing changes secondary to stress. Multiple-dose levels are recommended in order to determine dose–response relationships and the dose at which no immunotoxicity is observed.

Additional studies may show that no risk of immunotoxicity can be detected and no further testing is needed.

Additional studies may demonstrate a risk of immunotoxicity but there is not sufficient data to make a reasonable risk–benefit decision. In this case, further testing may be of benefit, which provides sufficient information for the risk–benefit decision. If the overall risk–benefit analysis suggests that the risk of immunotoxicity is considered acceptable and/or can be addressed in a risk management plan, then no further testing in animals might be called for.

If the weight-of-evidence review indicates the need for additional immunotoxicity studies, these should be completed before exposure of a large population of patients, usually Phase III. If the target patient population is immunocompromized, immunotoxicity testing can be initiated at an earlier time point in the development of the drug.

The different methods which are recommended for the evaluation of immunotoxic effects are summarized as follows.

Standard toxicity studies are:

Hematology	total leukocyte counts and absolute differential leukocyte counts
Clinical chemistry	globulin levels and A/G ratios
Gross pathology	lymphoid organs/tissues
Organ weights	thymus, spleen, (optional: lymph nodes)
Histology	thymus, spleen, draining lymph node, and at least one additional lymph node, bone marrow, Peyer's patch* , NALT**, BALT**

* Oral administration only.
** Inhalation administration only

Additional immunotoxicity studies are:

- T-cell dependent antibody response (TDAR)
- Immunophenotyping (lymphocyte subsets)
- Natural killer cell activity assays
- Host resistance studies
- Macrophage/neutrophil function
- Assays to measure cell-mediated immunity

This new ICH guideline replaces all guidelines from EU, USA, and Japan. It represents a very pragmatic approach and uses studies, e.g., standard toxicity studies, which are conducted anyhow. There is high confidence in the prediction of these assays for any potential of new compounds to induce immune suppression or stimulation. This guideline helps to reduce the number of animals and requires additional studies only in special cases for concern.

4.1.11. Nonclinical Safety Studies for the Conduct of Human Clinical Trials for Pharmaceuticals (ICH M3)

ICH M3 provides information which studies and of which duration are needed before the different clinical phases of development can be started. In addition to this timing of preclinical studies in dependence of clinical development plans, there are recommendations under which conditions different population can be included into clinical trials, populations such as men, women of childbearing potential or pregnant women, and finally pediatric populations.

The objectives of M3 are: To reduce differences between regions; to facilitate timely conduct of clinical trials; to reduce unnecessary use of animals and other resources; to promote early availability of new drugs. This guidance was implemented because regulatory recommendations differed among regions of Europe, USA, and Japan. Would it be possible to develop a mutually acceptable guidance?

The guidelines provide general guidance for drug development. A guideline is not a legal requirement. Approaches should be scientifically and ethically appropriate. The different endpoints are addressed in detail in this note.

Study requirements before the first starting dose in humans are:

- Safety pharmacology studies
- Kinetic/exposure data
- Single- and repeat-dose studies
- Local-tolerance studies
- Genotoxicity studies (in vitro)
- Phototoxicity studies

The requirement for the Core Battery of **safety pharmacology** is described in Section 4.1.9.

For **kinetics**, an information on exposure data (AUC) in animals prior human clinical trials is needed, while ADME data are needed at completion of phase I (human pharmacology) studies.

The duration for **repeat-dose toxicity studies** is related to the duration of clinical trials and their therapeutic indication. In principle, the duration of animal studies is equal to or exceeds the duration of the human clinical trials. In general, there is a relationship of 1:1 for studies in two mammalian species (one nonrodent). The details can be found in Table 8.

Table 8. Duration of repeated-dose toxicity studies to support Phase III trials in the EU*

Duration of clinical trials dose	Minimum duration of repeated toxicity studies	
	rodent	nonrodents
Up to 2 weeks	1 month	1 month
up to 1 month	3 months	3 months
up to 3 months	6 months	6 months
$>$ 3 months	6 months	6–12 months

* This table reflects the marketing recommendations in the three regions except a chronic nonrodent study is recommended for clinical use $>$ 3 months.

This table was updated by thorough additional Japanese studies in 2000 during the fifths ICH conference in San Diego, USA. The Japanese scientists compared the utility of routine four-week toxicity studies with two-week studies and concluded that in regard to the prediction of toxicities to the male reproductive organs two-week studies were as valid as four-week studies, therefore the former regional Japanese requirement to ask for a minimum duration of four-week studies before starting trials in men was dropped for a global consensus that the minimum duration of nonclinical studies is two weeks in rodents and two weeks in nonrodents.

On the other hand, regional differences continue in regard to **single-dose studies** in animals supporting single-dose studies in humans: In the United States, single-dose toxicity studies with extended examinations can support single-dose human trials. This concept encouraged the EU to offer comparable options with the microdosing

concept. This principle may especially be valuable for gaining on kinetic data very early for go/no go decisions, when several candidates are being developed in parallel.

Results from two in vitro **genotoxicity** studies are recommended to be available prior to first administration to humans, while the standard battery should be completed prior the initiation of phase II studies.

Carcinogenicity studies do not need to be completed in advance of the conduct of clinical trials unless there is cause for concern (ICH/S1A). For pharmaceuticals to treat certain serious diseases, carcinogenicity testing, if needed, may be concluded post-approval.

4.1.12. Inclusion of special populations (ICH M3)

For the inclusion of different patient populations, there exist regional differences, especially for **women with childbearing potential**:

- High level of concern for unintentional exposure of the embryo/fetus
- Currently regional differences in the timing of reproduction toxicity studies to support the inclusion
  Japan: assessment of female fertility and embryo fetal development should be completed prior to the inclusion of women of childbearing potential using birth control in any type of clinical trial
  EU: assessment of embryo fetal development should be completed prior to Phase I trials in women and female fertility studies prior to phase III trials
  US: women of childbearing potential may be included in early, carefully monitored studies without reproduction toxicity studies, provided appropriate precautions are taken to minimize risks
  US: assessment of female fertility and embryo fetal development should be completed before phase III trials

The inclusion of children into clinical trials has gained further tremendous interest:

- Safety data from previous adult human exposure:
  Most relevant information
  Necessary before pediatric clinical trials
- Prior trials:
  Appropriate repeated-dose toxicity studies
  All reproduction toxicity studies
  Standard battery of genotoxicity tests
  Juvenile animal studies should be considered
- Carcinogenicity testing:
  Prior to long-term exposure of children
  Cause for concern

4.1.13. The Common Technical Document (ICH M4)

The Common Technical Document (CTD, ICH M4) is another very important multidisciplinary guideline, which combines information for the three disciplines within the ICH process: for quality, safety, and efficacy. The following illustration focuses predominantly on the preclinical safety issues (Figure 4). For detailed information on the other aspects and the general organization, see Sections 4.2 and 4.3.

The objectives of M4 are to assist authors in the preparation of nonclinical pharmacology, pharmacokinetics, and toxicology written summaries in an acceptable format. The CTD is not intended to indicate what studies are required, but provides an appropriate format for the nonclinical data. The CTD is nothing else than a placeholder for the different parts of a documentation for the market authorization process.

No guideline can cover all eventualities; common sense and a clear focus on needs of regulatory authority assessor are the best guides to constructing an acceptable document. Therefore, modify the format if needed with the aim to provide the best possible presentation and facilitate the understanding for the evaluation of the results.

The **Nonclinical Summary** is organized as follows:

1) Pharmacology
   a) Written Summary
   b) Tabulated Summary
2) Pharmacokinetics
   Written Summary
   Tabulated Summary
3) Toxicology
   Written Summary
   Tabulated Summary

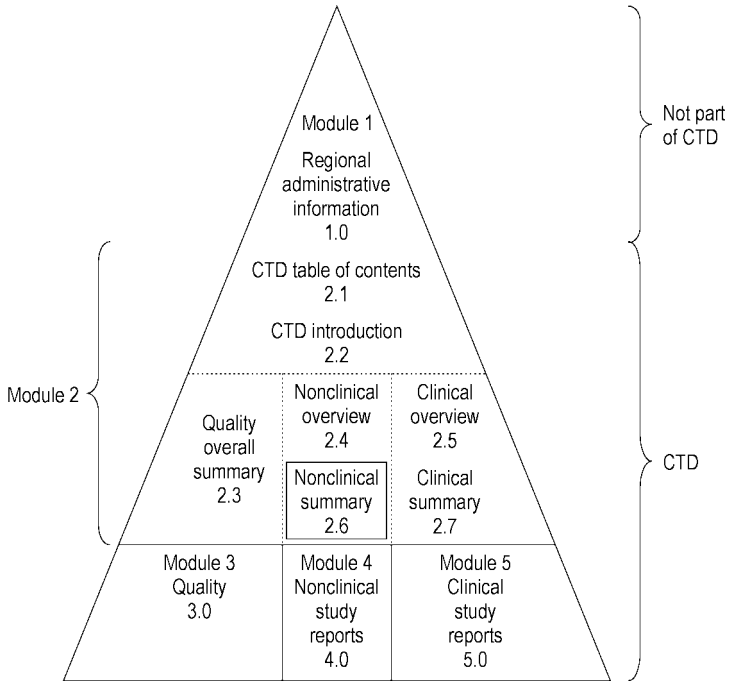

Figure 4.

This organization is kept up in all parts of the dossier, it is repeated for the overview, the summaries and the reports.

Details are:

1.a. Pharmacology, Written Summary
- Brief Summary
- Primary Pharmacodynamics
- Secondary Pharmacodynamics
- Safety Pharmacology
- Pharmacodynamic Drug Interactions
- Discussion and Conclusions

3.a. Toxicology: Written Summary
- Brief Summary
- Single-Dose Toxicity
- Repeat-Dose Toxicity
- Genotoxicity
- Carcinogenicity
- Reproduction Toxicity Local Tolerance
- Other Toxicity Studies
- Discussion and Conclusions

The kinetic information should be integrated across studies and across species, exposure in the test animals should be related to exposure in humans given the maximum intended doses.

In vitro studies should precede in vivo studies. Where multiple studies of the same type need to be summarized within the Pharmacokinetics and Toxicology sections, studies should be ordered by species, by route, and then by duration (shortest duration first).

Species should be ordered as follows:

1) Mouse
2) Rat
3) Hamster
4) Other rodent
5) Rabbit
6) Dog
7) Nonhuman primate
8) Other nonrodent mammal
9) Nonmammals

It is also recommended to limit the information in the summaries and overview: The overview should contain the essential and critical results on approximately 30 pages. The length of the nonclinical Written Summaries in general should not exceed 100-150 pages. The Brief Summaries for pharmacology should be written on two to three pages, for pharmacokinetics also two to three pages and for toxicology approximately six pages.

The **repeat-dose toxicity** studies (including supportive toxicokinetics evaluation) should be

summarized in order by species, by route, and by duration, a brief detail of the methodology is needed, the important findings (e.g., nature and severity of target organ toxicity, dose (exposure)/response relationships, no observed adverse effect levels, etc.) should be highlighted. The nonpivotal studies can be summarized in less detail (pivotal studies are the definitive GLP studies specified by ICH M3).

For the **carcinogenicity** studies (including supportive toxicokinetics), a brief rationale should explain why the studies were chosen and what is the basis for high-dose selection. Individual studies should be summarized in the following order: Long-term studies (in order by species; including range-finding studies that cannot appropriately be included under repeat-dose toxicity or pharmacokinetics), then short- or medium-term studies (including range-finding studies that cannot appropriately be included under repeat-dose toxicity or pharmacokinetics). Other studies such as antigenicity, immunotoxicity, mechanistic studies (if not reported elsewhere), dependence, studies on metabolites, studies on impurities or other studies should be summarized.

Module 2 contains in addition 34 templates and 31 examples. The examples have been taken from real dossiers and give good orientation how to fill out the templates. Module 4 finally contains all nonclinical study reports and is organized in an granular manner with the same principles for all types of compounds and studies.

This CTD has been tested in practice for a couple of years. Although not perfect for every case, it has proven its usefulness. Industry knows where to place specific information and data and regulators know where to find them. This has facilitated the review process tremendously, only one dossier is necessary for international registration, a lot of resources can be diverted to more important issues.

But the CTD is a living document, its weaknesses are apparent. In order to improve its quality and practicability, suggestions and proposals for improvements are invited from the public. These modifications are being dealt with within the process of "questions and answers", the ICH experts will continue to publish repeatedly their conclusions.

4.2. Clinical Trials

Before broad application of any treatment to patients suffering from a particular disease, the reliable efficacy and safety of the treatment needs to be demonstrated. As single-case studies and retrospective comparison of effects are not suitable to predict the success of a treatment in future patients, prospective, controlled, randomized, blinded and well-documented clinical trials are the worldwide recognized, scientific approach to achieve marketing authorization and clinical acceptance.

A clinical trial is any investigation in human subjects intended to discover or verify the clinical, pharmacological, and/or other pharmacodynamic effects of one or more medicinal products, medical devices or treatment procedures and/or to identify any adverse reactions to these treatments and/or to study absorption, distribution, metabolism, excretion or interaction of these treatments with the object of ascertaining its (their) safety and/or efficacy.

Methodology and quality standards of clinical trials have been rapidly developed in the last 50 years. The planning, performance and supervision of clinical trials are highly regulated on national and international levels. Scientific guidelines for the development program of drugs and the designs of required clinical trials in particular indications, released by the European and American health authorities, define the state-of-the-art and their recommendations should be followed. The protection, safety, and integrity of patients and healthy subjects participating in clinical trials are guaranteed through review and approval of each clinical trial by health authorities and independent ethics committees as well as by extensive subject data protection, information and consent requirements.

Especially the development of a new drug (Investigational Medicinal Product or IMP) requires the generation of in depth pharmaceutical, pharmacological, preclinical and clinical information within a development plan. The results of the different trials have to be compiled and presented in a dossier in an internationally accepted order and submitted to the respective health authorities for detailed review and evaluation of efficacy and safety of the IMP before a marketing authorization for use of the IMP in defined indications, dosage and conditions of use

is granted. Also after introduction to the market, safety and efficacy are further observed and evaluated in specially designed clinical trials or non-interventional trials.

1) Medical and ethical justification of the trial
2) Information of the human subject participating on the risk involved
3) His or her freely given consent (informed consent)

Thus, the goal of the special legal requirements is the protection of the human subject. Special parts of the U.S. Code of Federal Regulations (CFR) have been devoted to this aim, Part 50, Protection of Human Subjects (http://www.access.gpo.gov/nara/cfr/-waisidx_06/21cfr50_06.html) and Part 312, Investigational New Drug Application (http://-www.access.gpo.gov/nara/cfr/waisidx_06/-21cfr312_06.html). Basic elements of informed consent are designed to guarantee a free and self-responsible decision of the trial subject. This involves a statement that the study involves research, an explanation of the purpose of the research, and the expected duration of the subject's participation. A description of benefits and any reasonably foreseeable risks or discomforts must be given to the subject. Benefits obviously do not exist for trials in voluntary healthy subjects. The subject must be informed of appropriate alternative procedures or courses of treatment, especially if these might be to his or her advantage. Existing confidentiality and exceptions thereto also have to be described; an exception, for example, is that the records may be inspected by the competent authority. Furthermore information on possible financial compensation in case of injury or medical treatment must be provided. The fact that participation in the trial is voluntary and that the subject can refuse to participate without penalty must be included in the information.

Further elements pertaining to informed consent may include information on currently unforeseeable risks, on possible termination of the investigation, and the consequences of a subject's decision to withdraw from the trial. The patients should be notified of the number of subjects participating in the trial and also of significant new findings to let them decide whether or not to continue their participation. The subject must give written consent to participation in the trial.

Before the investigator is allowed to start a clinical trial, an application has to be approved by the competent authority. In the EU, the requirements are listed in Directive 2001/20/EC (http://ec.europa.eu/enterprise/-pharmaceuticals/eudralex/vol-1/dir_2001_20/-dir_2001_20_en.pdf). In the US, the sponsor of an investigation has to submit an *Investigational New Drug Application* (IND) to the FDA, if a clinical investigation is to be conducted with a new drug. Unless the FDA notifies the sponsor of any objections, IND goes into effect and the trial may be started thirty days after the FDA has received the IND.

A written positive opinion of an *Ethical Committee* has to be received by the sponsor before the trial can commence. In the United States, the ethical committee is called the *Institutional Review Board* (IRB), 21 CFR Part 56 (http://-www.access.gpo.gov/nara/cfr/waisidx_06/-21cfr56_06.html). The respective regulations describe the general standards for the composition, operation, and responsibility of an IRB that reviews clinical investigations. The term IRB denotes any board, committee, or other group formally designated by an institution to review, to approve the initiation of, and to conduct periodic review of biomedical research involving subjects. The primary purpose of such a review is to assure the protection of the rights and welfare of the human subjects. Clinical investigations relating to new drugs can only be initiated after the investigation has been reviewed by, approved by, and remains subject to continuing review by an IRB. An IRB has to have at least five members with varying professional backgrounds necessary to review specific research activities. The persons must be knowledgeable in these areas. No IRB may consist entirely of men or women or entirely of members of one profession. One member has to work in a lay area (e.g., lawyer, ethicist, member of the clergy). One member of the IRB is not to be affiliated with the institution whereas the other members may belong to the institution. No IRB member may have a conflicting interest in the research project (e.g., because the IRB member is also an investigator).

The IRB has to approve, require modification of, or disapprove all research activities covered

by the investigation. The Informed Consent of the participant in the trial is subject to inspection by IRB. The criteria for IRB approval of research are contained in 21 CFR Part 56.111 (http://a257.g.akamaitech.net/7/257/2422/-10apr20061500/edocket.access.gpo.gov/-cfr_2006/aprqtr/pdf/21cfr56.111.pdf). The institution responsible for the trial or the IRB must prepare and maintain adequate documentation of IRB activities. The parent institution is presumed to be responsible for the operation of an IRB.

4.2.1. Phases of Clinical Drug Development

The generation of information in the development of a new drug occurs in four phases:

Phase I	Human pharmacological trials
Phase II	Tolerability and dose finding trials
Phase III	Proof of efficacy and safety
Marketing authorization	
Phase IV	Evaluation of long-term safety and efficacy in the indication(s) and patient population(s) for which the drug received a marketing authorization

Theoretically, the phases are performed in consecutive order. However, the development of each drug requires a specific plan with different types and numbers of trials based on the results of the preclinical and earlier trials. Thus, the phases can overlap, be combined, or further divided, e.g., into Phase Ia, Ib, IIa, IIb, III, and IIIb, which indicates the types of trials considered belonging to the earlier or later stage of a particular Phase. Human pharmacological trials are mostly performed in Phase I but often in addition along Phase II, III, and IV.

Phase I The human pharmacological investigation of new drugs in small subject populations under strictly controlled conditions in specialized Phase I units is an efficient way to explore the suitability of a new drug for human use in general and to detect the potential for efficacy of the drug in the envisaged indication. Different types of trials are conducted:

- The first administration of a new drug to man and the evaluation of the pharmacological properties of a new drug are mostly performed in healthy male volunteers. Healthy female volunteers can only be included once the required reproductive toxicology tests have been performed. For safety reasons, drugs with a potential for cell damage such as anticancer drugs or for a specific immune response can only be tested in patients with the respective disease.
  Beginning with doses well below the no-effect level in animals and the expected effective dose range in humans, groups of 3 to 8 subjects get exposed to increasing single doses and placebo until the maximum tolerated dose level (MTD) is reached. This is followed by exploration of the MTD under multiple dosing in comparison to placebo.
- With a dose in the tolerability range, the pharmacological properties of the drug such as extent of absorption, distribution in the body, ways of metabolization, as well as route and extent of excretion (ADME trials) are investigated in 3 to 4 subjects after single-dose administration of radio-labeled drug in authorized institutions.
- The pharmacokinetic behavior for different potential dosage forms (e.g., infusion, oral solution, fast- and/or slow-release tablet, etc.) is compared in absolute (intravenous versus oral formulation) or relative (oral solution versus oral formulation) "bioavailability studies".
- To explore potential differences in the pharmacokinetic profile between healthy young and healthy elderly subjects (65 years and older) a parallel-groups trial with single- or multiple-dose administration of the drug to those two populations is performed with drugs indicated for these age groups.
- "Bioequivalence studies" are performed with the same type of formulation to compare, e.g., the pharmacokinetic behavior under fasting and fed conditions or to compare the formulations from two different manufacturing sources.
  Highly sensitive and specific analytical methods for the mother compound and its active metabolite(s) are used in these different types of pharmacokinetic trials to determine the drug levels over time in body fluids such as blood (plasma or serum), urine, or saliva. Comparing the pharmacokinetic conditions intra-individually by performing trials with randomized cross-over designs as well as the use of sophisticated biostatistical methods for

the calculation of results and their extrapolation allow to gain reliable information on the drug with small populations of 8 to 24 subjects per group.

- Depending on the enzymatic pathway of metabolization of the drug, the pharmacokinetic behavior of the drug in subjects with different pharmacogenetic and genotype conditions ("fast and slow metabolizers") are performed, especially to predict the potential for decreased tolerability of the drug in patients with slow metabolization capacity resulting in accumulation of the drug in the body after multiple administration.
- Drugs with an excretion predominantly via the kidney *or* the liver have the potential for accumulation in patients with renal or liver insufficiency. Evaluation of the pharmacokinetic behavior in patients with a different degree of renal or liver insufficiency is performed by comparing their drug plasma or serum levels with those in healthy subjects.
- In an attempt to predict efficacy of the drug or particular dose levels in patients, trials using special validated pharmacodynamic and psychometric methods or models, and comparing the effects of the drug with those of placebo administration, are also performed in healthy volunteers in Phase I.
- As many patients have to take different drugs for the same disease or for different diseases in parallel, a potential pharmacokinetic or pharmacodynamic interaction of the new drug with a selection of other drugs has to be investigated in "interaction studies".

Usually, 200 to 500 subjects are enrolled in Phase I trials for the development of a new drug.

Phase II. In Phase II, the new drug is administered for the first time to patients suffering from the disease intended to be treated. The objectives of clinical trials in this phase are to obtain information on

- efficacy in comparison to placebo
- therapeutic dose range
- optimal dose and dosing scheme
- dose–response relationship
- short-term tolerability
- short-term safety

The trials are performed in selected patients fulfilling narrow in- and exclusion criteria, with treatment duration of only few days to 6 weeks and under strong medical supervision in specialized hospitals. A larger number of diagnostic tests and patient visits than under normal treatment conditions are required.

The following types of trials are performed:

- A first administration of a new drug to patients with the respective disease is often performed in a pilot trial with 8 to 20 highly selected patients, in which the drug is administered unblinded, as a single dose or for a very short period of time, and under very controlled conditions, to get an impression of the tolerability of the drug in patients with the respective pathological condition.
- Proof-of-Concept trials (Phase IIa), ideally investigating the drug and placebo effects by intra- and inter-individual comparisons, using a double-blind, randomized cross-over design if the disease condition is such that the preconditions for all treatment periods are comparable. Mostly, however, these trials are performed in parallel groups with different dose levels (including expected no-effect dose, optimal dose and maximum-effect dose) investigating the inter-individual differences to placebo and sometimes to a standard treatment, double-blind, randomized and – if necessary – with a dose titration period at the beginning. Usually 40 to 100 patients per treatment arm are required to show statistically significant differences between the treatments. This sample size depends on the variability of the chosen efficacy parameters, the placebo responder rate in this particular indication and the efficacy of the new drug. Special emphasis is put on the evaluation of tolerability and safety of the different treatments. Due to the relatively short duration of the treatment and the fact that the patients may receive an insufficient dose level or placebo, the benefit for patients in these trials is limited.
- Comparative trials (Phase IIb) are also performed in patients with well-defined disease conditions and without concomitant other diseases. The trials are mostly performed double-blind, randomized, in parallel groups, comparing selected dose levels of the new drug with established drugs in this indication, sometimes also including placebo. Dif-

ferent dosing conditions, different diagnostic parameters and different efficacy parameters are included into the study designs of trials in that phase. The tolerability and safety of the different treatments is also compared.

All-together, the number of patients in Phase II does not exceed a few hundred.

Phase III. The purpose of Phase III is to investigate the new drug under conditions as close as possible to the daily life situation of patients. The exclusion criteria are less stringent, defined concomitant diseases and treatments are permitted, the treatment duration is according to the need of the disease. The number of test parameters and of the trial-requested patient visits is reduced. However, the studies are again performed under controlled, double-blind, randomized conditions. But 100 patients or more are invited to continue their trial with an open, uncontrolled extension for up to one year.

The objectives of this phase are

- to identify and demonstrate superiority of the new drug over existing comparators or to establish comparability with existing treatments
- to confirm the optimal dose
- to prove efficacy and tolerability for a treatment duration adequate for the disease
- to establish the safety profile of the new drug and to identify relatively rare adverse drug reactions
- to substantially increase the information available for estimating the risk–benefit as well as cost–benefit ratio of the new treatment in comparison to existing treatments
- to investigate special questions like interaction with other drugs or impact of different conditions such as age, impaired renal or liver function on safety, and efficacy of the drug
- to investigate the new drug in special populations

The clinical trials in Phase III include several hundred patients, when the primary efficacy parameter is prevention of the disease, then even thousands of patients are enrolled. These trials are performed in many different hospitals or practices, often in different countries or even globally. Standardization of trial conditions, assessments, and documentation are a logistic challenge but vital for the chance to show statistically significant differences between the treatments.

Phase III trials performed before and their results included in the marketing authorization application dossier are defined as belonging to Phase IIIa; trials performed after submission but before granting the marketing authorization are defined as Phase IIIb. The objectives of these trials are generation of additional information potentially required by the competent authorities before granting the marketing authorization, to collect more safety information and to generate interest for the new drug in the investigator community.

Phase IV. The purpose of Phase IV is to investigate the safety of the drug and efficacy under normal clinical conditions in patients with the respective indication, to collect broad safety information or pharmacoeconomic and epidemiological data. The trials have either short protocols with requirements for minimal additional diagnostic activities to normal procedures and with nearly no exclusion criteria or they are purely observational, so-called "Non-interventional trials", defined as follows:

"A study where the medicinal product(s) is (are) prescribed in the usual manner in accordance with the terms of the marketing authorization. The assignment of the patient to a particular therapeutic strategy is not decided in advance by a trial protocol but falls within current practice and the prescription of the medicine is clearly separated from the decision to include the patient into the study. No additional diagnostic or monitoring procedures are applied to the patients and epidemiological methods are used for the analysis of collected data."

For these trials, no authorization from the health authorities is required. The number of patients included in Phase IV is unlimited. Phase IV begins with granting of the marketing authorization and ends with the discontinuation of sales.

Another approach to classifying clinical trials is presented in a guideline of the EMEA entitled "General Considerations for Clinical Trials":

4.2.2. Methods for Proving Effectiveness

As with all scientific experiments, clinical trials should provide generally valid evidence from a limited number of observations. The reproducibility of a result can be determined after repetition of the test and the probability of general validity can be calculated with statistical methods. Clinical trials must therefore be designed and evaluated on the basis of biostatistical criteria.

The most reliable way of providing evidence of efficacy and safety is the performance of a confirmatory trial - mostly performed in Phase III. In such trial, the key hypothesis of interest follows directly from the primary objective of the trial, is always pre-defined, and is the hypothesis that is subsequently tested when the trial is complete. It is equally important to estimate with due precision the size of the effects attributable to the treatment of interest and to relate these effects to their clinical significance.

In such trial, unavoidable changes are explained and documented, and their effect is examined. A justification of the design of such trial, and of other important statistical aspects such as the principal features of the planned analysis, is set out in the protocol. Only a limited number of key clinical questions relevant to efficacy or safety are allowed to be addressed and can be answered. A single primary variable should be selected.

In contrast to confirmatory trials, exploratory trials – mostly performed in early drug development phases - have objectives which may not always lead to simple tests of predefined hypotheses, however, they may also be clear and precise. In addition, exploratory trials require sometimes a more flexible approach to design so that changes can be made in response to accumulating results. Their analysis may entail data exploration; tests of hypothesis may be carried out, but the choice of hypothesis may be data-dependent. Such trial cannot be the basis of the formal poof of efficacy, but may contribute to the total body of relevant evidence.

Any individual trial may have both confirmatory and exploratory aspects as in most confirmatory trials the data are also subject to exploratory analyses which serve as a basis for explaining or supporting their findings and for suggesting further hypotheses for later research.

An important element of a successful trial is the right selection of subjects. In the earlier phases, the choice of subjects is heavily influenced by the wish to maximize the chance of observing specific clinical effects of interest. Therefore, the protocol defines a narrow subgroup of the respective total patient population through the inclusion criteria and eliminates patients with potentially interfering factors and risk factors through the exclusion criteria. Structural uniformity can be further increased by stratification of each subject group according to certain characteristics like age range, sex or degree of illness of the subjects. However, by the time confirmatory trials are performed the trial population should more closely mirror the target population while maintaining sufficient homogeneity to permit precise estimation of the treatment effects.

To increase the comparability of the treatment groups and to avoid bias, the subjects must be assigned to the different treatment groups according to a randomization list. Preferably, this list should consist of randomization blocks to improve the chance that treatment groups will be of equal size.

Another technique to avoid bias in a clinical trial is *blinding*. In a single-blind trial the subjects are not informed about the treatment they receive, however, the investigator is aware of the type of treatment administered. In a double-blind trial the assignment of the patient to the treatment groups remains also unknown to the investigator. Precondition for a double-blind trial is the preparation of identically looking and tasting formulations for the test and the reference product. If this is technically not possible, the "double dummy technique" has to be applied: for test and reference medication an identically looking placebo version has to be prepared. The subjects get either the active version from the test medication plus the inactive version from the reference medication or vice versa, depending on their randomization assignment. Also the labeling of the blinded trial medication does not reveal which type of treatment is in which box. Individual sealed envelopes with the description of the trial medication for each subject are deposited with the investigator and can be opened in the case of occurrence of a serious adverse event in which the medically required action is dependent on the trial medication received.

The choice of the most suitable study design is depending on the objective of the trial and the conditions of the disease. Group (inter-individual) comparison (parallel design) can be used to in most medical conditions and is suitable for determination of optimal doses and comparison of efficacy. However, this design requires large patient numbers, especially if the efficacy parameters have a high variability or if small differences have to be identified.

Variability and thus the number of observations can be significantly reduced if the study is conducted as an intra-individual comparison (cross-over design) in the same subject. Each subject receives each of the test preparations in a random order with a sufficiently long wash-out phase between the periods. In pharmacokinetic trials in healthy subjects this design is mostly used, however, in patients it can only be applied in chronic illnesses with stable characteristics and when the termination of the first trial medication leads to a rapid reestablishment of the disease conditions before the start of the second treatment period. If this is not ensured, when the effect of one therapy continues into the subsequent treatment period (carry-over effect), the comparability of the two treatments is not given.

4.2.3. Adverse Drug Reactions

Despite state-of-the-art performed drug development, the efficacy of a medication in an individual patient cannot be guaranteed. On the other hand, already Paracelsus defined in the early 16th century that each drug has the potential of being a poison – depending on the dose administered. Therefore, physician and patient need to decide whether the relevance of the potential benefit of taking the medication outweighs the potential risk. With the increase of the seriousness of the disease and/or the impact of the symptoms on the patient's quality of life, patients are more willing to accept adverse drug reactions.

Participating in a clinical trial does not only contain the risk inherent in the medications administered but also the trial activities like blood withdrawals etc. include a certain risk for the subject to be harmed. For the evaluation of the safety profile of a drug administered in a clinical trial, it is very important to find out whether the occurrence of a negative experience is related to the study medication or not. Different terms have therefore to be distinguished:

"Adverse Event" (AE)	any untoward medical occurrence in a patient or clinical trial subject administered a medicinal product and which does not necessarily have a causal relationship with the treatment
"Adverse Drug Reaction" (ADR)	all untoward and unintended responses to an investigational medicinal product related to any dose administered
"Serious Adverse Event" or "Serious Adverse Reaction" (SAE) or (SAR)	any untoward medical occurrence or effect at any dose results in death is life-threatening, requires hospitalization or prolongation of existing hospitalization, results in persistent or significant disability or incapacity, or is a congenital anomaly or birth defect
"Suspected Unexpected Serious Adverse Reaction" (SUSAR)	an adverse reaction, the nature or severity of which is not consistent with the applicable product information (e.g., investigator's brochure for an unauthorized investigational product or summary of product characteristics for an authorized product) and which is suspected to be unexpected

An investigator is obliged to report every SAE within 24 hours to the sponsor of the clinical trial. In clinical trials performed outside the European Union, the sponsor is obliged to report all SAEs within 15 days, in case of life-threatening SAEs or death within 7 days, to the relevant health authorities and ethics committees. In European studies, only SUSARs need to be reported by the sponsor within these delays. In European trials the national health authority of the country where the sponsor of the trial is located is responsible for entering the information into the clinical trials section of the European safety database EudraVigilance. Also all investigators need to be informed by the sponsor about all SUSARs occurred with the drug.

According to internationally accepted standards, adverse events are classified according to the following categories:

"Intensity"	mild – moderate – severe
"Outcome"	resolved – resolving - not resolved - resolved with sequelae – fatal – unknown
"Causality"	certain – probable/likely – possible – unlikely - not related - unassessable/unclassifiable

In addition, onset and duration of AEs as well as required countermeasures are reported and analyzed in the safety evaluation of a clinical trial.

In an Annual Safety Report all AEs, all SAEs, and in Europe all SUSARs occurred - in all clinical trials performed and ongoing - must be pre-

sented and analyzed by the sponsor of a trial and an update of the risk-benefit analysis must be performed to justify the further development of the drug.

4.2.4. Good Clinical Practice

The need for quality standards and the assurance of their maintenance has been accepted for drug production as well as for preclinical and clinical drug development and has been defined in guidelines such as Good Manufacturing Practice (GMP), Good Laboratory Practice (GLP), and Good Clinical Practice (GCP).

The International Conference on Harmonization (ICH), a collaboration process between health authorities from USA, Europe, and Japan and pharmaceutical industry from these three regions, aiming to harmonize the standards, prerequisits, procedures, and documentation for drug development in the three regions, has developed guidelines on many aspects of drug development. The ICH Guideline on Good Clinical Practice (ICH E6 GCP) was adopted in 1996 by all ICH parties and in the following years implemented in the legislation in these regions and most other countries in the world. It represents the international ethical and scientific quality standard for designing, conducting, recording, and reporting clinical trials in human subjects. Compliance with GCP provides assurance that the rights, safety and well-being of trial subjects are protected and that the results of the clinical trials are reliable. The guideline is to provide a unified standard for the three regions to facilitate the mutual acceptance of clinical data by the regulatory authorities in these jurisdictions.

The following principles were agreed in ICH GCP:

1) Clinical trials should be conducted in accordance with the ethical principles that have their origin in the Declaration of Helsinki, and that are consistent with GCP and the applicable regulatory requirement(s).
2) Before a trial is initiated, foreseeable risks and inconveniences should be weighed against the anticipated benefit for the individual trial subject and society. A trial should be initiated and continued only if the anticipated benefits justify the risks.
3) The rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society.
4) The available nonclinical and clinical information on an investigational product should be adequate to support the proposed clinical trial.
5) Clinical trials should be scientifically sound, and described in a clear, detailed protocol.
6) A trial should be conducted in compliance with the protocol that has been received prior institutional review board (IRB)/independent ethics committee (IEC) approval/favorable opinion.
7) The medical care given to, and medical decisions made on behalf of, subjects should always be the responsibility of a qualified physician or, when appropriate, of a qualified dentist.
8) Each individual involved in conducting a trial should be qualified by education, training, and experience to perform his or her respective task(s).
9) Freely given informed consent should be obtained from every subject prior to clinical trial participation.
10) All clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation, and verification.
11) The confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirement(s).
12) Investigational products should be manufactured, handled, and stored in accordance with applicable good manufacturing practice (GMP). They should be used in accordance with the approved protocol.
13) Systems with procedures that assure the quality of every aspect of the trial should be implemented.

The preparation and implementation of a quality assurance systems with standard Operating Procedures (SOP) defining the roles and responsibilities of all parties involved, the processes, procedures, and documentation required and the approach to supervision of adherence to these SOPs through auditors, ensure that these

principles are reliably followed. In addition, health authorities perform inspections during the performance of a clinical trial, during review of the marketing authorization application or after granting of a marketing authorization to ensure that these principles are rigorously followed.

The protection of trial subjects is a central element in GCP. The principles for physicians involved in biomedical research have been established in the Declaration of Helsinki, adopted by the World Medical Association in 1964 and amended several times until 2002. The WHO guidelines, CIOMS guidelines, and many national guidelines and legislations emphasize as well on the fundamental need for subject protection in biomedical research.

GCP requires the establishment of independent, multidisciplinary ethics committees/institutional review boards to ensure the rights of the research subject, especially also in case of involvement of vulnerable populations, and to safeguard the integrity of the subjects. The IEC/IRB has to review all the ethical aspects of the trial protocol and accompanying documents, especially concerning the rationale and objective of the trial, the benefit-risk assessment, the subject information and informed consent process, the suitability of investigator and site, the compensation of trial subjects and investigators as well as the insurance coverage in case of injury.

4.2.5. Clinical Application Dossier

An especially important achievement of the ICH process was the adoption of Topic ICH M4, the "Common Technical Document" in 2000. This guideline presents the agreed upon common format for the preparation of a well-structured application dossier for applications supposed to be submitted to regulatory authorities of the three regions. It is aiming to significantly reduce the time and resources needed to compile applications for registration of new drugs and to ease the preparation of electronic submissions. Regulatory reviews and communication with each other and the applicant are facilitated.

The CTD is organized in five modules. While *Module 1* – containing administrative information and prescribing information (Summary of Product Characteristics = SmPC) is supposed to be region-specific, the other four Modules are common for the three regions.

Module 2 contains the CTD Summaries (see also Section 4.1.13):

- CTD Table of Content
- CTD Introduction
- Quality Overall Summary
- Nonclinical Overview
- Clinical Overview
  - Product Development Rationale
  - Overview of Biopharmaceutics
  - Overview of Clinical Pharmacology
  - Overview of Efficacy
  - Overview of Safety
  - Benefits and Risks Conclusions
  - Literature References
- Nonclinical Written and Tabulated Summaries
  - Pharmacology
  - Pharmacokinetics
  - Toxicology
- Clinical Summary
  - Biopharmaceutic Studies and Associated Analytical Methods
  - Clinical Pharmacology Studies
  - Clinical Efficacy
  - Clinical Safety
  - Literature References
  - Synopses of Individual Studies

Module 3 presents the information on quality, *Module 4* contains the nonclinical study reports.

Module 5 presents the clinical study reports and related information:

- Table of Contents of Module 5
- Tabular Listing of All Clinical Studies
- Clinical Study Reports
- Literature References

Topic ICH M4E is the guideline on the clinical documentation within the CTD. It explains in detail the type of information and assessments requested in the different sections of the CTD.

In the meantime, the three regions have implemented the CTD format into their national legislations including the standards for the electronic version (Table 9).

Table 9. Approach to classifying clinical studies according to objectives

Type of Trial	Objective of Trial	Trial Examples
Human pharmacology	assess tolerance define/describe pharmacokinetics and pharmacodynamics explore drug metabolism estimate activity	dose-tolerance trials single and multiple dose pharmacokinetic or pharmacodynamic trials drug interaction trials
Therapeutic exploratory	explore use for the targeted indication estimate dosage for subsequent trials provide basis for confirmatory study designs, endpoints, methodologies	earliest trials of relatively short duration in well-defined narrow patient populations, using surrogate or pharmacological endpoints or clinical measures dose-response exploration trials
Therapeutic confirmatory	demonstrate/confirm efficacy establish safety profile provide an adequate basis for assessing benefit/risk relationship to support licensing establish dose-response relationship	adequate and well controlled trials to establish efficacy randomized, parallel dose-response trials clinical safety trials trials of mortality/morbidity outcomes large simple trials comparative trials
Therapeutic use	refine understanding of benefit/risk relationship in general or special populations and/or environments identify less common adverse reactions redefine dosing recommendations	comparative efficacy trials trials of mortality/morbidity outcomes trials of additional endpoints large simple trials pharmacoeconomic trials

Source: ICH Topic E8 – General Considerations for Clinical Trials, 17 July 1997

4.3. Quality of Pharmaceutical Products

Quality cannot be tested into products; i.e., quality should be built in by design. Pharmaceutical companies develop a product and its manufacturing process to consistently deliver the intended performance of the product. The information and knowledge gained from pharmaceutical development studies and manufacturing experience provide scientific understanding to support the establishment of criteria for the release of the products in line with current quality standards. An agreement on what is regarded as a current quality standard is reached between the manufacturer and the regulatory authorities prior to marketing pharmaceutical products. This agreement, which forms a key element of the "Marketing Authorization" is similar to a legal contract in which the company producing and controlling the product is committed to meeting quality criteria in future, i.e., prospectively, which are based on experience gained in the past while developing the product, i.e., retrospectively.

Quality standards are published by national regulatory authorities. Since the adoption of the first European Directive on pharmaceuticals (65/65/EEC) in 1965, the Member States of the European Union (EU) have been working towards harmonization of their requirements. These requirements have been published in "The Rules Governing Medicinal Products in the European Union" Volume 3" Medicinal Products for Human Use: Guidelines, Vol 3A "Quality and biotechnology" (http://ec.europa.eu/enterprise/-pharmaceuticals/eudralex/homev3.htm)

In the US, the quality standards are available through FDA's (FDA = Food and Drug Administration) homepage (http://-www.fda.gov/cder/guidance/index.htm) or via the Code of Federal Regulations (CFR) Title 21 (http://www.access.gpo.gov/cgi-bin/-cfrassemble.cgi?title=200621)

Some manufacturers achieve their goal of ensuring consistent quality by implementing Process Analytical Technology (PAT), a system for designing, analyzing, and controlling manufacturing through measurements of critical attributes of raw and in-process materials and processes during the manufacturing process (ICH Guideline Pharmaceutical Development – Q8 – http://www.ich.org/-LOB/media/MEDIA1707.pdf).

Product quality should be maintained throughout the product lifecycle, i.e., all phases in the life of a product from the initial development through marketing until the product's discontinuation. Attributes that are important to the quality of the pharmaceutical product should remain consistent with those used in the clinical trials. An effective quality risk management approach can ensure the high quality of the pharmaceutical product by identifying and controlling potential quality issues during development and manufacturing. Information from pharmaceutical development studies can be a basis for quality risk management. Quality risk management is a systematic process for the assessment,

control, communication and review of risks to the quality of the pharmaceutical product across the product lifecycle. (ICH Guideline Quality Risk Management – Q9 – http://www.ich.org/-LOB/media/MEDIA1957.pdf).

4.3.1. Manufacturing of Substances – Test of Starting Materials

All starting materials have to be tested according to adequate quality standards applying suitable analytical procedures before manufacturing or synthesizing an active substance. A chain of documents is generated as chemical synthesis proceeds. This documentation includes, amongst others, a written description of the process and appropriate production records, records of the raw materials used, records of batch numbers, records of the critical processing steps accomplished, and intermediate test results.

4.3.2. Control of Active Substances: Impurities

Impurities in active substances can be classified as organic impurities, inorganic impurities, and residual solvents. Process- and substance-related *organic impurities* may arise during the manufacturing process and/or storage of the active substance. They may be identified or unidentified, volatile or non-volatile, and include starting materials, byproducts, intermediates, degradation products, reagents, ligands, and catalysts. *Inorganic impurities* are normally known and identified and include reagents, ligands and catalysts, heavy metals, inorganic salts, and other materials, such as, filter aids and charcoal. *Residual solvents* are inorganic or organic liquids used as vehicles for the preparation of solutions or suspensions in the synthesis of an active substance.

The amount of impurities, identified or unidentified, must be limited to ensure the safety of patients. The procedure used to establish acceptable thresholds is described in detail in the ICH *Guideline "Impurities in New Drug Substances (Q3A)"*, (http://www.ich.org/LOB/-media/MEDIA422.pdf).

The levels of residual solvents in medicinal products should be no higher than can be supported by the safety data. Solvents that are known to cause unacceptable toxicity, such as, benzene, should be avoided in the production of active substances, excipients, or medicinal products unless their use can be strongly justified in a risk-benefit assessment. Solvents associated with less severe toxicity, e.g., chloroform, methanol, or toluene, should be limited in order to protect patients from potential adverse effects. Ideally, less toxic solvents, such as acetone, ethanol, or ethyl ether, should be used where viable.

A method for classifying residual solvents by risk assessment is provided in the *ICH Guideline "Impurities: Guideline for Residual Solvents (Q3C)"* (http://www.ich.org/LOB/media/-MEDIA423.pdf).

4.3.3. Control of Pharmaceutical Products: Impurities

Impurities in pharmaceutical products may be produced by degradation of the active ingredient or by reaction of the active ingredient with an excipient and/or with the immediate container/closure system. Consequently, the amount of degradation products allowed in a pharmaceutical product should be such that the therapeutic efficacy of the product is not affected. The toxicity of degradation products must also be evaluated and assessed.

Levels of degradation products can be measured, for example, by comparison of an analytical response for a degradation product to that of an appropriate reference standard or to the response of the active ingredient itself. Methods for setting up limits for degradation products are described in detail in the *ICH Guideline " Impurities in New Drug Products (Q3B)"* (http://-www.ich.org/LOB/media/MEDIA421.pdf).

4.3.4. Specifications

A specification is defined as a list of tests, references to analytical procedures, and appropriate acceptance criteria, which are numerical limits, ranges, or other criteria for the tests described. It establishes the set of criteria to which a drug substance or drug product should conform to be considered acceptable for its intended use. "Conformance to specifications" means that the

active substance and/or pharmaceutical product, when tested according to the listed analytical procedures, will meet the listed acceptance criteria. Specifications are critical quality standards that are proposed and justified by the manufacturer and approved by regulatory authorities as conditions of approval.

ICH Guideline "Specifications: Test Procedures and Acceptance Criteria for New Drug Substances and New Drug Products: Chemical Substances – Q6A (http://www.ich.org/LOB/-media/MEDIA430.pdf). Decision trees: http://-www.ich.org/LOB/media/MEDIA431.pdf

ICH Guideline "Specifications: Test Procedures and Acceptance Criteria for Biotechnological/Biological Products – Q6B (http://-www.ich.org/LOB/media/MEDIA432.pdf).

4.3.5. Validation of Analytical Procedures

The objective of analytical procedure validation is to demonstrate that the method in question is suitable for its intended purpose. The following analytical procedures have to be validated in particular:

- Identification tests intended to ensure the identity of an analyte in a sample by comparing a property of the sample to that of a reference standard
- Quantitative tests for the content of impurities
- Limit tests for the control of impurities
- Quantitative tests of the active moiety in the active substance or pharmaceutical product
- Dissolution testing for pharmaceutical products
- Particle size determination for active substances

The following characteristics should be considered during the validation of analytical procedures:

Accuracy	closeness of agreement between value accepted and value found;
Precision	closeness of agreement between a series of measurements obtained from multiple sampling of the same homogeneous sample;
Repeatability	precision under same operating conditions over a short interval of time;
Intermediate Precision	within-laboratories variations: different days, different analysts, different equipment;
Reproducibility	precision between different laboratories;
Specificity	ability to assess unequivocally the analyte in the presence of components, which may be expected to be present;
Detection limit	lowest amount of analyte in a sample which can be detected;
Quantitation limit	lowest amount of analyte in a sample which can be quantitatively determined;
Linearity	results directly proportional to the concentration of analyte in the sample;
Range	interval between the upper and the lower concentration of analyte in the sample for which it has been demonstrated that the analytical procedure has a suitable level of precision, accuracy and linearity;
Robustness	capacity to remain unaffected by small variations in method parameters.

Relevant information on validation of analytical procedures is provided in the *ICH Guideline "Validation of Analytical Procedures: Text and Methodology – Q2(R1)"* (http://-www.ich.org/LOB/media/MEDIA417.pdf).

4.3.6. Process Validation and Qualification

Since it is important to ensure that the defined quality standards of the pharmaceutical product are met, the manufacturers conduct "Process Validation and Qualification" on the product prior to marketing thereby establishing and documenting that:

- The facilities, equipment and processes have been designed in compliance with the requirements of the current GMP guidelines.
- The facilities and equipment have been built and installed as specified. This constitutes the Installation Qualification (IQ).
- The facilities and equipment are operated as specified. This constitutes the Operational Qualification (OQ).
- The facilities and equipment operate as specified and repeatedly and reliably produce a finished product of the requisite quality. This constitutes the Process Validation (PV).

Any aspect of the premises, facilities, equipment or processes, which may affect the quality of the product, either directly or indirectly, is subject to qualification and validation (*Convention for the Mutual Recognition of Inspections: "Principles of Qualification and Validation in Pharmaceutical Manufacture," Document PH 1/96*, January 1996).

Continuous Process Verification is an alternative approach to process validation in which manufacturing process performance is continuously monitored and evaluated.

4.3.7. Batch Release

If an active substance or a pharmaceutical product meets all quality criteria as previously agreed with the regulatory authorities, the Qualified Person (QP) or the member of the Quality Assurance organization in charge within the organization of the manufacturer can release the batch for further processing or marketing.

Some types of pharmaceutical products, e.g., vaccines, sera, toxins, antitoxins, antigens, blood products, etc., have to be tested and released on a batch-by-batch basis by an independent state laboratory specializing in this particular analytical area. These products contain substances whose purity and potency cannot be measured by chemical methods alone.

4.3.8. Stability Testing

Stability studies are a major expense when developing new products. This applies in particular to pharmaceutical products that are to be marketed in several strengths and package types. Multiple strengths and package types combined with multiple batches, various storage conditions, test parameters, and test intervals require a great number of samples to be tested at considerable cost. Further to this, the different requirements of the regulatory agencies must be taken into account. As a result, an enormous amount of stability testing, much of it redundant, was performed by multinational pharmaceutical companies seeking approvals in several countries. Therefore, the compilation of a common set of stability requirements for marketing authorizations was considered to be a top priority for the pharmaceutical industry when the International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use (ICH) was formed in 1991. Regulators and pharmaceutical industry representatives from the EU, Japan, and the United States with observers from the Canadian and Swiss Health Authorities and the WHO chose stability testing as one of the first issues to be discussed and harmonized. An ICH Guideline on Stability Testing was subsequently developed and published in 1993, after which it was adopted throughout the ICH region [4]. This guideline describes the stability testing requirements for a registration application within the "ICH territory": EU, Japan, and the USA *(ICH Harmonized Tripartite Guideline "Stability Testing of New Drug Substances and Products – Q1A")* of 27 October 1993, as amended in 2003 (http://www.ich.org/-LOB/media/MEDIA419.pdf)

Meanwhile, additional ICH stability guidelines have been developed:

- *Photostability Testing of New Drug Substances and Products (Q1B)* http://-www.ich.org/LOB/media/MEDIA412.pdf
- *Stability Testing for New Dosage Forms (Q1C)* http://www.ich.org/LOB/media/-MEDIA413.pdf
- *Bracketing and Matrixing Designs for Stability Testing of New Drug Substances and Products (Q1D)* http://www.ich.org/LOB/media/-MEDIA414.pdf
- *Evaluation of Stability Data (Q1E)* http://-www.ich.org/LOB/media/MEDIA415.pdf
- *Stability Testing of Biotechnological/Biological Products (Q5C)* http://www.ich.org/-LOB/media/MEDIA427.pdf

In Europe, additional note for guidance has been published by regulators, which address specific aspects not covered by ICH:

- *Note for Guidance on Inclusion of Antioxidants and Antimicrobial Preservatives in Medicinal Products* (CPMP/CVMP/QWP/115/95) http://www.emea.eu.int/pdfs/human/qwp/-011595en.pdf
- *Note for Guidance on Start of Shelf –Life of the Finished Dosage Form* (CPMP/QWP/072/96) http://www.emea.eu.int/pdfs/human/qwp/-007296en.pdf
- *Note for guidance on Maximum Shelf-Life for Sterile Products after First Opening or following Reconstitution* (CPMP/QWP/159/96 corr) http://www.emea.eu.int/pdfs/human/-qwp/015996en.pdf
- *Guideline on Stability Testing for Applications for Variations to a Marketing Authorization* (CPMP/QWP/576/96/Rev 1) http://www.emea.eu.int/pdfs/human/qwp/-057696en.pdf
- *Note for guidance on Declaration of Storage Conditions A) in the Product Information of Medicinal Products, B) for Ac-*

Table 10. The CZs, Definition and criteria and long-term testing conditions

CZ	Definition	Criteria	
		Mean annual temperature measured in the open air/Mean annual partial water vapor pressure	**Long-term testing conditions**
I	Temperate climate	$\leq 15°C / \leq 11 \times 10^2$ Pa	21°C / 45% RH
II	Subtropical and Mediterranean climate	> 15 to $22°C / > 11$ to 18×10^2 Pa	25°C / 60% RH
III	Hot and dry climate	$> 22°C / \leq 15 \times 10^2$ Pa	30°C / 35% RH
IVA	Hot and humid climate	$> 22°C / > 15$ to 27×10^2 Pa	30°C / 65% RH
IVB	Hot and very humid climate	$> 22°C / > 27 \times 10^2$ Pa	30°C / 75% RH

tive Substances (CPMP/QWP/609/96/Rev 1) http://www.emea.eu.int/pdfs/human/qwp/-060996en.pdf

– *Note for guidance on In-Use Stability Testing of Human Medicinal Products* (CPMP/QWP/2934/99) http://www.emea.eu.int/pdfs/human/qwp/-293499en.pdf
– *Stability Testing of Existing Active Substances and Related Finished Products* (CPMP/QWP/122/02, rev 1) http://www.emea.eu.int/pdfs/human/qwp/-012202en.pdf

The US FDA has adopted meanwhile all ICH stability guidelines.

As the ICH guidelines are limited to new active substances and products for human use marketed in ICH countries in which climatic conditions are moderate, the World Health Organization (WHO) developed its own guideline to include established substances, and also hot or hot and humid climatic zones (CZs). The WHO guideline was published in 1996 (*"Guidelines for stability testing of pharmaceutical products containing well-established drug substances in conventional dosage forms" WHO Technical Report Series, No. 863*, 1996) [5]. According to this guideline, the recommended long term testing for hot and humid countries is 30°C/70% RH, a condition, which was changed in 2001 to 30°C/65% RH (*Report of the 37th meeting of the WHO Expert Committee on Specifications for Pharmaceutical Preparations, Geneva, 22-26 October 2001. WHO Technical Report Series 908, 2003, page 13).*

The 40th WHO Expert Committee on Specifications for Pharmaceutical Preparations met in Geneva in October 2005, and decided to split the current CZ IV (hot and humid countries) into CZ IVA – for which 30°C/65% RH will remain the standard long-term testing condition – and CZ IVB, for which, if justified, 30°C/75% RH will become the testing condition

Table 10 gives criteria [6] and long-term testing conditions assigned to the CZs of the world.

National regulatory authorities in countries outside the ICH region developed stability requirements, which are either adopting ICH guidelines, e.g., Switzerland, Norway, Iceland, Canada, Australia, or are based on the original ICH text but with additional or different testing conditions, e.g., Brazil, and ASEAN.

4.3.9. Documentation

Format of Registration Documentation. Regulatory authorities in various regions require applicants to submit marketing authorization applications according to their national guidelines and regulations. In *Europe*, the format for presenting quality related documents is outlined in detail in *The Rules governing Medicinal Products in the European Union, Volume 2B: Notice to Applicants*, June 2004, to be found at. http://pharmacos.eudra.org/F2/eudralex/vol-2/B/ctd_06-2004.pdf.

The format for applications in the *USA* is described in the *Code of Federal Regulations, Title 21, Chapter I, Subchapter D, Part 314.50* (http://a257.g.akamaitech.net/7/257/-2422/10apr20061500/edocket.access.gpo.gov/-cfr_2006/aprqtr/pdf/21cfr314.50.pdf).

In *Japan*, the Ministry of Health and Welfare regularly publishes updates of their Technical Requirements for New Drug Applications, updated versions are available from Yakuji Nippo, Ltd., Japan, editorial supervision by the Pharmaceuticals and Cosmetics Division, Pharmaceutical Affairs Bureau, Japanese Ministry of Health and Welfare.

The Common Technical Document – Quality The ICH Expert Working Group "Common Technical Document (CTD) Quality" developed a common structure (ICH M4 Q – http://-www.ich.org/cache/compo/276-254-1.html) for the chemical/pharmaceutical section of dossiers for marketing authorization applications. The final ICH guideline, which has been released in 2000, is regarded as a major achievement, and the result of a concerted effort on the part of the regulatory reviewers and pharmaceutical industry (see also Section The Common Technical Document).

Many regulatory authorities worldwide have adopted the ICH CTD as the standard format for submitting applications for clinical trials, new marketing authorizations, and variations.

In the following, the structure of the CTD Quality module is listed.

Drug Master Files. Drug Master Files (DMFs) are documents that are submitted to the U.S. FDA providing confidential detailed information on the facilities, processes, or articles used in the manufacturing, processing, packaging, and storage of medicinal products intended for human use. DMFs allow parties other than the holder of the DMF to reference information not disclosed to them (Center for Drug Evaluation and Research, FDA *Guideline for Drug Master Files*, September 1989, to be found at http://www.fda.gov/cder/guidance/dmf.htm).

In the EU, the use of Drug Master Files is restricted to active substances only (*CHMP Guideline on Active Substance Master File Procedure, London, 15 April 2005 – Rev 2 published on 24 May 2006* (http://www.emea.eu.int/pdfs/-human/qwp/022702en.pdf)

Drug Substance	
General information	
Nomenclature	Recommended INN[a], compendial name, if relevant, chemical name(s), company or laboratory code, national name, e.g., BAN[b], USAN[c], JAN[d], and CAS registry number.
Structure	NCE: structural formula, including relative and absolute stereochemistry, the molecular formula, and relative molecular mass. Biotech: schematic amino acid sequence indicating glycosylation sites and relative molecular mass.
General properties	Physicochemical and other relevant properties.
Manufacture	
Manufacturer(s)	Name, address, and responsibility of each manufacturer and each proposed production site or facility.
Description of manufacturing process and process controls	Description of the manufacturing process and process controls: flow diagram, quantities of raw materials, solvents catalysts and reagents, identification of critical steps, operating conditions such as temperature, pressure, pH, and time. Additionally for Biotechnology: Batch and scale definition; flow diagram from original inoculum up to the last harvesting operation; description of each process step including purification and modification reactions; filling procedure; container closure system; storage and shipping conditions.
Control of materials	List of raw materials, starting materials, solvents, reagents, catalysts; information on their quality and control. Additionally for biotech products: viral safety information, source of the cell substrate and analysis of the expression construct; cell banking system, quality control activities, cell-line stability.
Control of critical steps and intermediates	Critical steps: Tests and acceptance criteria performed at critical steps of the manufacturing process. Intermediates: Quality and control of intermediates.
Process validation and/or evaluation	Studies for aseptic processing and sterilization.
Manufacturing process development	Description of the significant changes made to the manufacturing of the drug substance used in producing non-clinical, clinical, scale-up, pilot, and production scale batches.
Characterization	
Elucidation of structure and other characteristics	NCE: Confirmation of structure based on synthetic route and spectral analyses, for example. Information on the potential for isomerism, the identification of stereochemistry, or the potential for forming polymorphs. Biotech: details on primary, secondary, and higher-order structure post-translational forms, biological activity, purity, and immunochemical properties.
Impurities	
Control of Drug Substance	
Specification	
Analytical Procedures	
Validation	
Batch Analyses	
Justification	

Reference standards or materials	Information on the reference standards or reference materials used for testing of the drug substance.
Container closure system	Description of the container closure system identity of materials of construction and their specifications; suitability with respect to choice of materials, protection from moisture and light, compatibility of the materials of construction with the active substance.
Stability	
Stability summary and conclusions	Types of studies conducted, protocols used, and the results obtained; conclusions with respect to storage conditions and retest date or shelf-life.
Stability data	Results of the stability studies in an appropriate format such as tabular, graphical, narrative; information on analytical procedures used and their validation.

[a] INN = Recommended International Non-Proprietary Names.
[b] BAN = British Approved Names.
[c] USAN = United States Adopted Names.
[d] JAN = Japanese Accepted Names.

Drug product

Description and composition	Description of the dosage form, list of all components, and their amount, their function, reference to their quality standards, type of container and closure used.
Pharmaceutical Development	Information on the development studies conducted to establish that the dosage form, the formulation, manufacturing process, container closure system, microbiological attributes and usage instructions are appropriate for the purpose specified.
Components of the drug product	Compatibility of the active substance with excipients; key physicochemical characteristics of the active substance; choice of excipients.
Drug product	Summary describing the development of the drug product; justification of overages; physicochemical parameters; and biological properties.
Manufacturing process development	Selection and optimization of the manufacturing process.
Container closure system	Suitability of the container closure system, choice of materials, protection from moisture and light, compatibility of the materials of construction with the dosage form.
Microbiological attributes	Microbiological attributes of the dosage form; selection and effectiveness of preservative systems.
Compatibility	Compatibility of the drug product with reconstitution diluent(s) or dosage devices.
Manufacture	
Manufacturer(s)	Name, address, and responsibility of each manufacturer, and each site or facility involved.
Batch formula	List of all components, their amounts, and a reference to their quality standards.
Description of manufacturing process and process controls	Flow diagram, description of the process, and type of equipment; process parameters.
Controls of critical steps and intermediates	Critical steps: tests and acceptance criteria, with justification, including experimental data, performed at critical steps. Intermediates: information on the quality and control of intermediates isolated during the process.
Process validation and/or evaluation	Description, documentation, and results for critical steps or critical assays used in the manufacturing process.
Control of excipients	
Specifications	
Analytical procedures	
Validation	
Justification	
Excipients of human or animal origin	
Novel excipients	For excipients used for the first time, full details of manufacture, characterization, and controls.
Control of drug product	
Specification(s)	
Analytical procedures	
Validation	
Batch analyses	
Characterization of impurities	
Justification	
Reference standards or materials	Information on the reference standards or reference materials used for testing of the product
Container closure system	Description of the container closure systems , including identity and specification of each component.
Stability	
Stability summary and conclusion	Types of studies conducted, protocols used, and the results. Conclusions with respect to storage conditions and shelf-life, and, if applicable, in-use storage conditions and shelf-life.
Stability data	Results of the stability studies conducted in an appropriate format such as tabular, graphical, narrative; information on analytical procedures used and their validation.

4.3.10. Variations – Changes

Manufacturers are required to amend and update manufacturing and control methods to take account of scientific and technical progress within the lifetime of a medicinal product. In Europe, this is detailed in Article 23 of Directive 2001/83/EC/EC and Article 16(2) of Regulation 726/2004. Since the level of scientific development is constantly subject to change, modifications in a product's composition, manufacturing and control methods are unavoidable, and even desirable as they may improve quality. However, every change that is made entails revision and amendments to the documentation. Sometimes approval has to be obtained from the health authorities before the proposed changes can be implemented (http://ec.europa.eu/-enterprise/pharmaceuticals/eudralex/vol-2/a/-v2a_chap5_r1_2004-02.pdf).

In the USA, regulation 21 CFR 314.70 (http://a257.g.akamaitech.net/7/257/2422/-10apr20061500/edocket.access.gpo.gov/-cfr_2006/aprqtr/pdf/21cfr314.70.pdf) provides regulatory requirements for changes subject to FDA approval prior to implementation, and for changes that can be reported to the FDA in an annual report.

4.3.11. Change Control

To ensure that the systems concerned are continuously validated, the manufacturer must monitor all changes with respect to facilities, materials, equipment, and processes used in the manufacture of active substances and pharmaceutical products. This commitment should be declared in the Validation Master Plan. As part of its Quality Management System, the manufacturer must establish a defined, formalized change control procedure.

4.3.12. Compliance

Manufacturers must make sure that the CMC documentation approved by the regulatory authorities complies with the methods applied in the manufacturing and quality control of the active substances and products.

4.3.13. Changes Requiring New Applications – Line Extension

Certain changes to a marketing authorization are considered to fundamentally change the terms of the authorization in question and therefore cannot be considered as a variation. For example:

- Changes to the active substance(s)
- Changes to strength (= the quantitative composition in terms of active substance)
- Changes to pharmaceutical form/dosage form
- Changes to route of administration

http://ec.europa.eu/enterprise/pharmaceuticals/-eudralex/vol-2/c/v2c_ea_v__10_2003.pdf

An abridged application for a new marketing authorization must be submitted for these changes and, in certain cases, the existing marketing authorization may have to be withdrawn. Otherwise, the new product is regarded as a "line extension" to the existing formulations and dosage forms.

4.3.14. Inspections

Regulatory authorities regard the marketing authorization as a contract between the manufacturers and themselves, based on the CMC information submitted and the satisfactory GMP status of the manufacturing site. In order to verify that data, processes and control methods submitted are accurate and complete, external inspectors are visiting the facilities and functions on a regular basis. Differences observed between submitted data and reality are cause for concern as the pharmaceutical product on the market may be "out of compliance". The inspectors must establish whether the control methods are applied and the facilities adequate, the personnel properly trained and manufacturing procedures followed. The inspectors also have to check whether cross-contamination can be excluded, i.e. whether the substance or product is subject to contamination by material stored or used in the same building. Other factors to be considered are:

- The degree of exposure of the material to adverse environmental conditions;
- Relative ease and thoroughness of clean-up;

– The risk of microbial contamination during aseptic processing of a sterile product.

The inspections may take place during the clinical development of a new substance, prior to granting marketing authorization (pre-approval inspection), or during the marketing phase.

4.3.15. The Mutual Recognition Agreement (MRA)

To cut down on the need for GMP inspections by regulatory authorities in foreign countries, agreements came into operations between the European Community and Canada, Australia, New Zealand, Switzerland and Japan (http://-ec.europa.eu/enterprise/pharmaceuticals/mra/-doc/statusreport20041018.pdf). It has been agreed to exchange certificates of GMP compliance for manufacturers, and batch certificates.

The main advantage of an MRA is the assurance that imported pharmaceutical products meet GMP requirements without resource expenditures on the part of the authorities of the importing country. Another important benefit of an MRA is a reduction of costs for the pharmaceutical manufacturer, mainly as a result of inspection reports being accepted by more than one regulatory authority. This eliminates the need for duplication of inspections and imply fewer inspections of the manufacturing sites. An MRA also reduces regulatory review times as it increases the efficiency of pre-marketing approval inspection activities.

4.3.16. Pharmaceutical Inspection Convention (PIC)

Another initiative to cut down on foreign GMP inspections is the "Convention for the Mutual Recognition of Inspections in respect of the Manufacture of Pharmaceutical Products" (shortened to Pharmaceutical Inspection Convention or PIC). This convention was signed in October 1970 by all EFTA countries at the time (i.e., Austria, Denmark, Finland, Iceland, Liechtenstein, Norway, Portugal, Sweden, Switzerland, and the United Kingdom). Since coming into force the convention has been extended to include Hungary, Ireland, Romania, Germany, Italy, Belgium, France, and Australia.

Similar to the MRAs mentioned above, the "PIC provides that an inspection of a pharmaceutical manufacturer, whose intention it is to export products in one of the Contracting States, is carried out by its national authority and shall be regarded and assessed by the health authority of the country of importation as if it had been carried out by its own inspectors."

A Committee of Officials supervises the operation of the PIC concerning the technical aspects of the manufacturing and control of pharmaceuticals. In the meantime, the convention's original scope has been extended to the training of laboratory staff, the mutual recognition of the quality standards of inspectorates, and also to cooperation concerning blood products, radiopharmaceuticals, medicinal gases, etc. Working groups meet regularly to develop guides, guidelines, and recommendations. Also seminars on technical subjects are organized, and the proceedings of these seminars and experts' meetings are published

4.3.17. Pharmaceutical Inspection Cooperation Scheme (PIC/S)

Some of the contracting states within the PIC are also Member States of the European Union with the result that certain rules and regulations were incompatible. Therefore, a new arrangement, which may eventually replace the PIC, called "Pharmaceutical Inspection Cooperation Scheme" (or PIC Scheme) has been set up and commenced operating in 1995. This new scheme is meant to be less formal and more flexible concerning cooperation between the inspectorates of the present PIC contracting states. Also the PIC Scheme allows for inspectorates of other countries to join. In addition to the recognition of foreign inspections, the PIC Scheme is in harmony with the main tasks of the PIC, namely, "the exchange of information, networking and confidence building between the national inspection authorities, the development of quality standards systems, the training of inspectors and other related experts and its work towards the global harmonization of Good Manufacturing Practice (GMP)".

The contracting members of the PIC Scheme are the following inspectorates: Australia, Austria, Belgium, Canada , the Czech Republic,

Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Malaysia, the Netherlands, Norway, Poland, Portugal, Romania, Singapore, the Slovak Republic, Spain, Sweden, Switzerland, and the United Kingdom. Observers are Estonia, the EMEA, UNICEF, and WHO. In addition, the following applications for membership are being considered: Estonia, South Africa, and Taiwan.

http://www.picscheme.org/indexnoflash.php

4.3.18. Pharmacopoeias

National Pharmacopoeias: In order to establish and maintain a scientific/technical standard for the quality of medicines, lists of active substances, excipients, and pharmaceutical products, their preparation, control, and use were developed and published in a "pharmacopoeia". The earliest of these was probably the "New Compound Dispensatory" issued in 1498 by the "Florentine Guild of Physicians and Pharmacists". Various pharmacopoeias were generated in the ensuing centuries: Barcelona (1535), Nürnberg (1546), Cologne (1565), Rome (1583), and London (1618). The first attempt to harmonize different standards and monographs was made by the Pharmacopoeia Committees of London, Edinburgh and Dublin. As the result of their efforts, the first "British Pharmacopoeia" documenting the agreement on harmonized international quality standards was published in 1864 [7].

International Pharmacopoeia: The project of developing an "International Pharmacopoeia" started in 1929 with an agreement signed by 19 countries. The next step was not made until 1948, when the first World Health Assembly established an expert committee to adapt and harmonize the existing monographs of major pharmacopoeias, and also to develop new standards. As the committee also kept developing countries in mind, some of the active substances included in the first edition of the International Pharmacopoeia were for the treatment of tropical diseases. Analytical methods are adapted for the practice of quality control testing in countries that need inexpensive methods, such as on-the spot identification of counterfeit pharmaceuticals (WHO Drug Information Vol. 13, No. 1, 1999, pp. 13 – 14).

European Pharmacopoeia: The need to unify the various national pharmacopoeias in Europe arose as a result of the proposal to allow the free movement of medicinal products within the European Community. Both public health and international trade require common manufacturing and quality control standards concerning active substances for human and veterinary use. It also requires these standards be updated to stay in line with scientific progress. In 1964, the "European Pharmacopoeia Convention" was signed by eight Member States of the Council of Europe. It is open to European countries (members and, under certain conditions, non-members of the Council of Europe), as well as to an international European organization. Observer status is also possible.

Thirty-five parties (in 2006) are members of the Convention including 24 EU member states): Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Macedonia, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Norway, Portugal, Romania, Serbia and Montenegro, the Slovak Republic, Slovenia, Spain, Sweden, Switzerland, Turkey, the United Kingdom; and the Commission of the European Union.

A further 18 parties participate as observers: Albania, Georgia, Poland, and the Ukraine, as well as the non-European countries Algeria, Australia, Brazil, Canada, China, Israel, Madagascar, Malaysia, Morocco, Senegal, Syria, Tunisia, United States of America, and the WHO. Some of these observers have implemented all or parts of the European Pharmacopoeia in their national pharmaceutical regulations (http://www.pheur.org/-site/page_608.php).

The European Pharmacopoeia (EP or Ph. Eur.) features more than 280 general analytical methods on which the specifications prescribed in the monographs are based. The methods range from physical, chemical and biological to microbiological, phytochemical, and so on. The EP also includes technical methods used to test the functional parameters of pharmaceutical forms such as the dissolution test for tablets.

The texts are updated at regular intervals to take into account scientific progress and any changes in the commercial products. The prin-

ciples of monograph compilation are adapted to accommodate regulatory requirements in the public health area (licensing, control and inspection authorities), industrial constraints and recent advances in technology and science (http://www.pheur.org/).

The monographs and methods published in the European Pharmacopoeia are normally transferred to the national pharmacopoeias, e.g., the *Deutsches Arzneibuch* (DAB), the British Pharmacopoeia (BP), the Pharmacopeé Francaise, the Farmacopeia Ufficiale della Republica Italiana.

In marketing authorization application dossiers submitted to regulatory authorities in the EU Member States, the monographs of the European Pharmacopoeia are obligatory.

The European Pharmacopoeia Commission has set the goals of rapidly adding to the number of existing European monographs (2000 so far), reducing the time taken to compile them, and, if necessary, generating monographs on recent substances still protected by patent.

Cooperation between the Council of Europe and the EU was further advanced on 26 May 1994, when the "European Pharmacopoeia Secretariat" took over additional objectives, e.g., setting up a European network of laboratories involved in the quality control of pharmaceutical products. Consequently, the "European Pharmacopoeia Secretariat" changed its name to the "European Department for the Quality of Medicines" (EDQM). This European network, called the Official Medicines Control Laboratories (OMCLs), is open not only to EU countries but also to members and observers of the European Pharmacopoeia Commission. Its main objectives are to achieve the mutual recognition of tests carried out at the national level for countries that belong to the EU, and the sharing of expertise, standardization and international collaboration for the other countries.

The European Pharmacopoeia is one of the co-founders of the Pharmacopoeia Discussion Group (PDG) that was set up in 1990 with the Japanese Pharmacopoeia (JP) and the United States Pharmacopeia (USP). Monographs and general methods of analysis proposed by national associations of manufacturers of pharmaceutical products are selected for harmonization in the three pharmacopoeias.

Unites States Pharmacopeia: In contrast to other pharmacopoeias, the Unites States Pharmacopeia (USP) is a corporation, not a governmental institution. Founded in 1820, it became the official public standards-setting authority for all prescription and over-the-counter medicines, dietary supplements, and other healthcare products manufactured and sold in the United States. The USP is funded through revenues from the sale of products, e.g. reference standards, and services (http://www.usp.org/aboutUSP/).

4.3.19. The EP Certificate of Suitability

Global expansion of international trade has given rise to a wide range of qualities of active substances within the territory covered by the Convention on the Elaboration of a European Pharmacopoeia. The great variety of supply sources has made it necessary to take additional measures to protect public health so that the European Pharmacopoeia can continue to play its role as a reliable reference in defining the quality of medicines. It is essential to have a procedure that allows the manufacturer of active substances to prove that the purity of the substance is suitably controlled by the EP monograph (http://www.pheur.org/site/page_625.php).

The Procedure is based on a Resolution of the Public Health Committee (Partial Agreement, Resolution AP-CSP (99) 4) (APCSP9904E). Under this official procedure, manufacturers or suppliers of substances for pharmaceutical use can apply for a certificate concerning

- the evaluation of the suitability of the monograph for the control of the chemical purity and microbiological quality of their substance according to the corresponding specific monograph, or
- the evaluation of reduction of Transmissible Spongiform Encephalopathy (TSE) risk, according to the new general monograph, or both.

This procedure is aimed at facilitating and simplifying exchanges between the partners to ensure that the quality of substances is guaranteed and that these substances comply with the European Pharmacopoeia.

5. References

1. J. S. Bindra, D. Lednicer: *Chronicles of Drug Discovery*, John Wiley, New York 1983.
2. C. Thromber: "The Pharmaceutical Industry," in Alan Heaton (ed.): *The Chemical Industry*, Blackie Academic and Professional, London 1986.
3. IMS World Review Analyst 2006, www.imshealth.com.
4. P. Jeffs: "The Importance of Stability Testing in the Registration of Pharmaceutical Products" in D. Mazzo (ed.): *International Stability Testing*, Interpharma Press Inc., Buffalo Grove, IL 1999.
5. S. Kopp-Kubel, M. Zahn: "The WHO Stability Guideline" in D. Mazzo (ed.): *International Stability Testing*, Interpharma Press Inc., Buffalo Grove, IL 1999.
6. Zahn M., et al.: "A risk-based approach to establish stability testing conditions for tropical countries", *Journal of Pharmaceutical Sciences* **95** (2006) 946 – 965.
7. A. Cartwright, B. Matthews (eds.): *Pharmaceutical Product Licensing – Requirements for Europe*, Ellis Horwood Ltd., Chichester 1991.

Phase-Transfer Catalysis

MARC HALPERN, Sybron Chemicals, Inc., Birmingham, NJ 08011, United States

1. **Introduction** 1
2. **Mechanism and Optimization Ramifications** 1
3. **Catalysts** 2
4. **Industrial Aspects** 4
5. **Uses** 4
5.1. **Nucleophilic Displacement** 4
5.2. **Dehydrohalogenation** 6
5.3. **Oxidation – Reduction** 6
6. **References** 6

1. Introduction

Phase-transfer catalysis (PTC) is a synthetic organic method and manufacturing process technology that is applied successfully in a wide range of organic reactions. Products manufactured in 1989 by processes that include at least one major PTC step are estimated to represent approximately $\$ 10\times10^9$ per year in sales. The major areas of application are the production and modification of polymers and monomers ($\$ 5\times10^9$), pharmaceuticals ($\$ 3\times10^9$), and agrochemicals ($\$ 2\times10^9$). Lesser commercial applications include the preparation of dyes, flavors, and fragrances.

Table 1. Applications of phase-transfer catalysis

PTC/X^- *	PTC/OH^-
Halide exchange	Alkylation (C-, O-, N-, S-)
Esterification	Asymmetric induction
Other nucleophilic substitutions (CN^-, RS^-, N_3^-, SCN^-, etc.)	Dehydrohalogenation
Oxidation	Condensation (Wittig, aldol, Darzens)
Reduction	Michael addition
Benzoin condensation	Hydrolysis
Polymerization	Deuterium exchange
	Isomerization
	Carbene addition
	Organometallic applications

* X^- = Anion other than OH^-.

Most PTC reactions involve the transfer of an anionic species from an aqueous or solid phase or interface to an organic phase. The transferred species then reacts with enhanced reactivity due to decreased hydration, increased concentration, or enhanced proximity of the reactants. In successful applications the result is high yield under mild reaction conditions. Because numerous organic reactions involve anions, either as reactants or catalysts (see Fig. 1), PTC is a highly attractive process option for a wide scope of reactions (see Table 1). Many of the attributes of PTC systems are advantageous from the standpoint of process engineers and chemists.

2. Mechanism and Optimization Ramifications

The mechanism of PTC reactions is illustrated in Figure 2. According to the extraction mechanism [1], a typical catalyst, such as a quaternary ammonium cation (symbolized by Q^+), forms an ion pair with an anionic reactant X^-, which equilibrates between an organic and an aqueous or solid phase. Once in the organic phase, the anion reacts with enhanced reactivity (due to reduced hydration) with a typical substrate RY to yield the product RX and an anionic leaving group Y^- that pairs with the catalyst. In the aqueous or solid phase the catalyst exchanges the leaving group for another anionic reactant to complete the catalytic cycle.

Two extremes of representative behavior may be observed in these systems: (1) chemical reaction control and (2) mass-transfer control (intermediate combinations of these two extremes are possible and do occur). In most classical PTC reactions the rate-determining step is the chemical reaction (e.g., Q^+X^-+RY in the example in Fig. 2). *Chemical reaction controlled* reactions are characterized by increasing reactivity as the lipophilicity of the catalyst increases, the hydration of the anionic reactant decreases, and the polarity of the organic solvent decreases

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_293

Bases:	OH^-, HCO_3^-
Oxidizing agents:	MnO_4^-, $Cr_2O_7^{2-}$, ClO^-, O_2^-, IO_4^-, ClO_4^-, $[Fe(CN)_6]^{3-}$
Reducing agents:	BH_4^-, BH_3CN^-, $HCOO^-$, $[Fe(CO)_{12}]^{2-}$, $[Co(CN)_5H]^{3-}$, $[Co(CO)_4]^-$, $S_2O_4^{2-}$
Miscellaneous:	HCl, HBr, HI, H_2O_2, $RhCl_4^-$, CCl_3^-
Nucleophiles:	F^-, Cl^-, Br^-, I^-, CN^-, RO^-, ArO^-, $RCOO^-$, RS^-, N_3^-, SCN^-, OCN^-, NO_2^-, NO_3^-

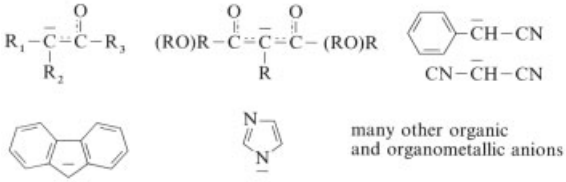

Figure 1. Anions successfully transferred and/or reacted under phase-transfer catalysis conditions

[2]. They also show a tendency toward catalyst poisoning by "soft" anions (definition of "hard" and "soft" acids and bases according to PEARSON) [3]. Some PTC reactions are *mass-transfer controlled*, for example, when the rate-determining step is the transfer of the reacting anion into the bulk organic phase. Mass-transfer controlled reactions occur often when the anion to be transferred has a high charge density and the subsequent reaction is rapid. For example, many alkylations of activated methylene or N−H groups, preceded by hydroxide ion extraction, are limited in rate by the transfer of the hydroxide ion into the organic phase. These reactions were previously thought to proceed via a different mechanism [4] and were termed "interfacial mechanism reactions." Mass-transfer limited reactions involving the transfer of anions with high charge densities are characterized by (1) increasing reactivity as the "accessibility" of the positive charge on the quaternary nitrogen of the catalyst increases (until the catalyst is too hydrophilic to function); (2) a strong dependence on agitation efficiency [5]; and (3) increasing reactivity if polar solvents are used (such as methyl isobutyl ketone).

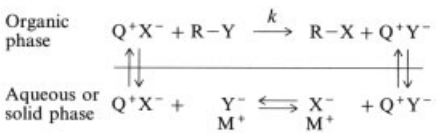

Figure 2. Extraction mechanism of phase-transfer catalysis

Empirically, PTC base-catalyzed reactions are often found to be mass-transfer limited for organic substrates that are not water soluble and have $pK_a < 23$. Not all PTC reactions involving transfer of hydroxide ions are mass-transfer limited [6], and mass-transfer control can also be observed with anions other than hydroxide (e.g., fluoride).

3. Catalysts

Most commercially available catalysts are quaternary ammonium and phosphonium salts, crown ethers, and open-chain polyethers (see Fig. 3).

Quaternary Ammonium Salts. Among the most useful quaternary salts ("quats") is tetrabutylammonium bromide [*1643-19-2*] (TBAB). Tetrabutylammonium bromide costs approximately $ 8.80/kg in tank car quantities and is recyclable by extraction [7]; its use has been reported in 450 patents describing both chemical reaction controlled and mass-transfer limited reactions. Other commercial quats include benzyltrimethylammonium chloride [*56-93-9*], methyltricaprylammonium chloride [*13275-89-3*] (Arosurf PT 64, trademark of Sherex; Aliquat 336 PTC, trademark of Henkel), and methyltributylammonium chloride [*56375-79-2*] (MTBC). The major ingredient of Aliquat 336 and Arosurf PT 64 is methyltrioctylammonium chloride. Both quats

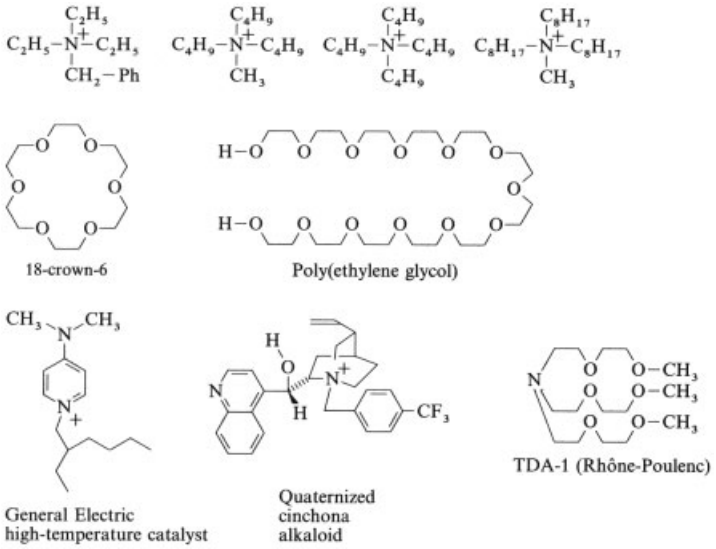

Figure 3. Representative phase-transfer catalysts

are effective in both chemical reaction controlled and mass-transfer limited reactions. One drawback of this catalyst is that the product must usually be distilled to separate the catalyst, or the catalyst may be separated by adsorption on a laboratory scale. Methyltributylammonium chloride and triethylbenzylammonium chloride are excellent mass-transfer limited catalysts for phase transfer of hydroxide ions. Methyltributylammonium chloride costs approximately \$ 4.80/kg, has a low molecular mass, is recyclable by extraction, and has an LD_{50} of 933 mg/kg (rat, oral [8]).

Quaternary Phosphonium Salts. Phosphonium salts based on tributylphosphine can be considered for cases in which the use of quaternary ammonium salts may be limited because of elevated reaction temperature (120 – 150 °C) (decomposition); it cannot be used with hydroxide ions.

Crown Ethers and Open-Chain Polyethers. Among the *crown ethers* (→ Crown Ethers) 18-crown-6 and dicyclohexano-18-crown-6 effectively solubilize salts in organic media via coordination of inorganic cations by the many lone electron pairs on the oxygen atoms [9]. The cost of crown ethers for industrial use is high, although some specialty applications are likely that employ crude crown ethers (ca. \$ 100/kg) in small but effective quantities. In addition, crown ethers are thought to be toxic. Openchain polyethers such as poly(ethylene glycols) (PEGs) and tris(dioxa-3,6-heptyl)amine [*70384-51-9*] (TDA-1) (Rhône-Poulenc) [10] are phase-transfer catalysts that act by coordination similar to crown ethers. The biological properties of these open-chain polyethers are more suitable for industrial use; the LD_{50} of PEG-400 (i.e., PEG with M_r=400) is 33 g/kg [11]; of TDA-1, 4 g/kg [12]. The PEGs are competitive in price with quaternary salts (ca. \$ 1.30/kg), whereas TDA-1 is more expensive (ca. \$ 42.10/kg in 1 200-kg lots).

The PEGs are sometimes as active as crown ethers, but larger quantities of PEGs are usually needed [13].

Newer Developments. *Chiral catalysts* and *high-temperature catalysts* are examples of novel catalysts with great potential commercial application. Merck, Sharp & Dohme has successfully employed chiral quaternized cinchona alkaloids to attain desired enantiomers with high enantiomeric excess and yield (both > 94 %) [14]. Significant understanding of this system has been achieved [15]. General Electric has shown the utility of *N*-(2-ethylhexyl)-4-(*N′*,*N′*-dimethylamino)pyridinium chloride and related compounds as effective high-temperature cata-

lysts [16]. *Polymer-bound phase-transfer catalysis* ("triphase catalysis") has been studied [17], [18] and used in fixed-bed reactors [19]. The commercial potential of these catalysts lies in their ease of separation from the product.

Choice of Catalyst. Quaternary salts are generally the catalysts of choice, except in certain cases, such as the transfer of certain metal species or at higher temperature. In these cases, TDA-1, 18-crown-6, poly(ethylene glycol), or *N*-alkyldimethylaminopyridinium salts may be effective.

4. Industrial Aspects

The *advantages* of PTC are high yields, milder reaction conditions relative to classical conditions, and increased reaction rates. This results in higher reactor turnover; more reproducible and reliable processes; better control of temperature for enhanced safety; and, often fewer impurities, less waste, and fewer recycle streams. The use of PTC usually alleviates the need for anhydrous polar aprotic solvents. A broad range of solvents may be chosen for a PTC process. Selection of the appropriate solvent can be based on properties such as good recoverability, low flammability, low toxicity, and environmental considerations. Approximately one-third of PTC applications use sodium hydroxide as a base to replace the classical alkoxides, amides, and hydrides. Aqueous sodium hydoxide is less corrosive, does not add organic materials to the waste stream, is inexpensive, and is readily available. Special precautions, such as maintaining anhydrous conditions, are not necessary with PTC. Because of the two-phase nature of most PTC reactions, separation can often be performed easily by decantation. Catalyst recovery and recycle have been achieved for many PTC processes.

The *limitations* of PTC for industrial processes relate to the catalyst. Quaternary ammonium and phosphonium salts usually decompose at ca. 120 – 150 °C, and even lower in the presence of bases or certain nucleophiles. The two major modes of quat decomposition are Hofmann degradation and nucleophilic attack. Separation, recovery, and recycle of the catalyst from the product may be difficult. Other attributes of various catalysts that may represent disadvantages include cost, toxicity, and lack of local waste treatment capability.

5. Uses

5.1. Nucleophilic Displacement

The primary strength of PTC is the ability to react anions of many types with organic substrates. Most of the applications of anions in organic chemistry involve nucleophilic attack in general and nucleophilic displacement in particular. The following are representative examples of reactions described in the patent literature, some of which are currently on-line in commercial manufacturing processes.

Polymers. The commercial production of many polymers involves the attack of a difunctional oxygen nucleophile on a difunctional electrophile. Bisphenol A has been reacted under PTC conditions with a variety of electrophiles to prepare polycarbonate, polyester, polyether, epoxy resin, polyethersulfone, and the polyetherimide monomer [20–23] (see Fig. 4). These reactions are good examples of nucleophilic displacement (nucleophilic aromatic substitution, nucleophilic aliphatic substitution, and acylation) and represent the largest volume of commercial PTC applications. Phase-transfer catalysis offers many advantages for polymerization. Among them are the ability to achieve high molecular masses; the ability to achieve nonstoichiometric results by using stoichiometric quantities (stoichiometry is regulated by the quantity of the quaternary cation – anion pair in the organic phase); the ability to obtain polymers with >99 % electrophilic end groups (particularly useful for production of block copolymers); and the ability to obtain polydispersities well below 2.

Polymers may be *modified* under PTC conditions. For example, glycidyl azide polymer (solid rocket fuel) may be made from polyepichlorohydrin and azide [24]. Tributylamine serves as an in situ phase-transfer catalyst for the preparation of a poly(vinyl chloride) plasticizer [25].

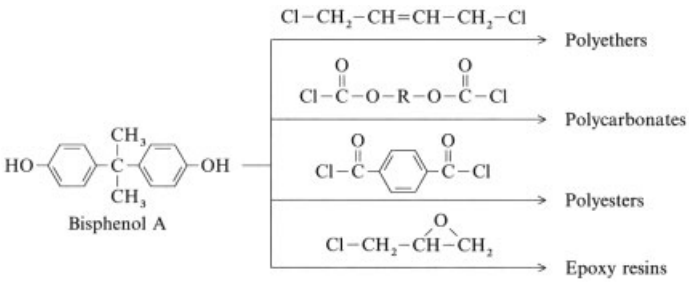

Figure 4. Examples of polymerizations catalyzed by PTC

Pharmaceuticals and Agrochemicals. The pharmaceutical and agrochemical industries utilize PTC for many processes, among which alkylation of carbon, nitrogen, oxygen, or sulfur is the most widespread.

Pharmaceuticals. One of the most outstanding projects with PTC involves the Merck *chiral alkylation* and related reactions (see Fig. 5) [26]. A chiral phase-transfer catalyst was also used to perform a chiral Michael addition [27]. Another group employed a cinchona alkaloid for the preparation of amino acids in $> 99\,\%$ e.e. (enantiomeric excess), via alkylation of an activated imine followed by hydrolysis and recrystallization [28]. *N-Alkylation* is another important reaction in the pharmaceutical industry. Dozens of nitrogen-containing heterocycles have been alkylated under PTC conditions, ranging from the very lipophilic dibenzazepine [29] to the very hydrophilic pyrimidine bases [30].

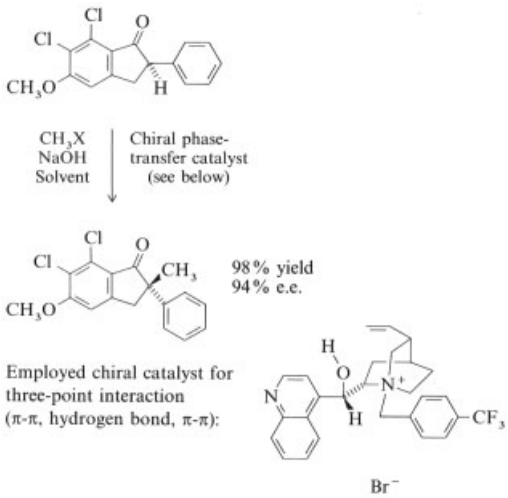

Figure 5. Chiral alkylation

Esterification of penicillin derivatives may have been one of the first commercial pharmaceutical applications of PTC [31]. Esterification of estradiol has been patented on a 9-kg scale [32]. A precursor to acemetacin is prepared in very high yield by esterification [33]. Nucleophilic attack is the primary reaction in the condensation of benzaldehyde with glycine to form higher amino acids [34]. Nucleophilic aromatic substitution has been used to prepare 4-nitrophenetole [35].

Three consecutive PTC reactions (esterification, *C*-alkylation, and hydrolysis) are used to convert phenylacetic acid derivatives to compounds such as ibuprofen and naproxen [36]. *Short-lived radioisotopes* may be synthesized by using PTC, because in these cases, high yields and particularly rapid workup procedures are required. For example, methylation of an activated imine with $^{11}CH_3I$ followed by oxime formation and hydrolysis yields a highly specific radiolabeled amino acid with a total synthesis and workup time of 35 min [37].

Agrochemicals. Pyrethroid insecticides have been prepared by PTC *esterification* [38]. Halohydroxypyridines and pyrimidines have been *alkylated* [39] and *phosphorothiolated* [40]. *Fluorination of aromatic rings* is an important nucleophilic substitution in both agrochemicals and pharmaceuticals. Aromatic fluoride – chloride exchange has been demonstrated under PTC conditions [41] (most likely mass-transfer limited). Another example of nucleophilic aromatic substitution is preparation of the insecticide intermediate cyanotetrachlorobenzene by $CN - NO_2$ exchange [42]. Mono-*O*-alkylation of catechol has been achieved under PTC conditions [43].

5.2. Dehydrohalogenation

Phase-transfer catalysis has been used for the production of chloroprene by dehydrohalogenation in cascade reactors; the feed rate was 4310 kg/h [44]. Poly(vinyl chloride) has been dehydrochlorinated under PTC conditions [45]. *N*-Vinylcarbazole monomer has been produced by two PTC reactions: *N*-alkylation of carbazole by dichloroethane, followed by dehydrochlorination [46]. The agrochemical intermediate 3,5-dichloro-α-methylstyrene has been prepared by dehydrobromination of the corresponding α-bromoalkyl compound [47].

5.3. Oxidation – Reduction

In a classic paper [48] and patent [49], Lee and Freedman described the oxidation of primary and secondary alcohols and amines to aldehydes, ketones, and nitriles, respectively, by using hypochlorite and PTC. Alper and co-workers have performed many oxidations and related reactions by combining PTC with organometallic catalysis [50]. Selective reduction with dithionite under PTC conditions has been reported [51]. A combination of PTC and transition-metal chemistry has been patented to oxidize toluene to benzoic acid in $>99\,\%$ yield and purity by using atmospheric oxygen as the oxidant [52].

Others. The wide applicability of PTC to sometimes surprising reactions is illustrated by the following examples: free-radical polymerization [53], scission of polysulfide cross-links in the reclamation of rubber [54], Wittig reaction in the presence of 50 % sodium hydroxide [55], decomposition of polychlorinated biphenyls in transformer oil [56], and by the use of membranes in PTC [57]. Phenol can be removed from dilute aqueous waste streams by extraction of the phenoxide and reaction with benzoyl chloride in a continuous reactor – settler system [58]. Cyanation has been patented for the preparation of dyes [59] and *C*-alkylation has been used to augment the aroma of perfume compositions [60].

6. References

1. C. Starks, *J. Am. Chem. Soc.* **93** (1971) 195.
2. D. Landini, A. Maia, F. Montanari, *J. Am. Chem. Soc.* **100** (1978) 2796.
3. J. Gordon, R. Kutina, *J. Am. Chem. Soc.* **99** (1977) 3903.
4. M. Makosza, *Pure Appl. Chem.* **43** (1975) 439.
5. M. Rabinovitz, Y. Cohen, M. Halpern, *Angew. Chem. Int. Ed. Eng.* **25** (1986) 960.
6. M. Halpern, Y. Sasson, M. Rabinovitz, *J. Org. Chem.* **48** (1983) 1022.
7. K. Sjoberg, *Aldrichimica Acta* **13** (1980) 55.
8. Ethyl Corp., *Material Safety Data Sheet,* 1985.
9. G. Gokel et al. in C. Starks (ed.): "Phase-Transfer Catalysis. New Chemistry, Catalysts and Applications," *ACS Symp. Ser.* **326** (1987) 24.
10. G. Soula, *J. Org. Chem.* **50** (1985) 3717.
11. N. Sax: *Dangerous Properties of Industrial Materials,* Van Nostrand Reinhold, New York 1984, p. 2254.
12. Rhône-Poulenc, *Material Safety Data Sheet,* 1985.
13. E. Santaniello, A. Manzochi, P. Sozanni, *Tetrahedron Lett.* **20** (1979) no. 47, 4581 – 4582.
14. Merck, US 4 578 509, 1986 (U. Dolling, E. Grabowski, S. Pines).
15. D. Hughes et al., *J. Org. Chem.* **52** (1987) 4745.
16. D. Brunelle, D. Singleton, *Tetrahedron Lett* **25** (1984) no. 32, 3383 – 3386.
17. S. Regen, *J. Am. Chem. Soc.* **98** (1976) 6270.
18. W. Ford in L. Mathias, C. Carreher (eds.): *Crown Ethers and Phase-Transfer Catalysis in Polymer Science,* Plenum, New York 1984, p. 201.
19. V. Ragaini et al., *J. Org. Chem.* **27** (1988) 1382.
20. V. Percec in C. Starks (ed.): "Phase-Transfer Catalysis. New Chemistry, Catalysts and Applications," *ACS Symp. Ser.* **326** (1987) 96.
21. F. Cook, R. Brooker, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **23** (1982) 149.
22. General Electric, US 4 595 760, 1986 (D. Brunelle).
23. D. Brunelle, D. Singleton, *Tetrahedron Lett.* **25** (1984) no. 32, 3383 – 3386.
24. Rockwell, US 4 379 894, 1983 (M. Frankel, E. Witucki, D. Woolery).
25. M. Yeh, H. Go, L. Chen, Y. Shih, *Ind. Eng. Chem. Prod. Res. Dev.* **27** (1988) 1582.
26. D. Hughes et al., *J. Org. Chem.* **52** (1987) 4745.
27. A. Bhattacharya et al., *Angew. Chem. Int. Ed. Eng.* **25** (1986) 476.
28. M. O'Donnell, W. Bennett, S. Wu, *J. Am. Chem. Soc.* **111** (1989) 2353.

29. I. Gozlan et al., *J. Heterocycl. Chem.* **19** (1982) 1569.
30. G. Phillopposian, US 4 450 163, 1984.
31. Astra AB Lakemedel, SE 397 981, 1977. Pfizer-Taito, JP 5 9 53 491, 1984.
32. Astra Aktiebolegat, US 4 081 461, 1978 (H. Fex, S. Kristensson, R. Stamvik).
33. Bayer, US 4 600 783, 1986 (S. Samaan, H. Horstmann).
34. Mitsui Toatsu, US 4 605 759, 1986 (C. Higuchi, T. Katoh, R. Mita, A. Yamaguchi).
35. Hoechst, US 4 454 355, 1984 (H. Schubert, K. Baessler).
36. J. Canicio, A. Ginebreda, R. Canelia, *An. Quim. Ser. C* **81** (1985) 181.
37. K. Fasth, G. Antoni, B. Langstrom, *J. Chem. Soc. Perkin Trans. 1* 1988, 3081.
38. Shell, US 4 409 150, 1983 (D. Wood).
39. Dow Chemical, US 3 969 360, 1976 (H. Freedman).
40. Ciba-Geigy, US 4 326 059, 1982 (R. Gargano, D. Perez, D. Williams).
41. Mallinckrodt, US 4 642 399, 1987 (C. White).
42. ICI, US 4 517 129, 1985 (D. Milner).
43. Brichima, US 4 599 461, 1986 (M. Correale, P. Maggioni, F. Minisci).
44. Du Pont, US 4 418 232, 1983 (L. Maurin).
45. Ihara Chem. Ind., JP 62 22 780, 1981.
46. Syntex, US 4 332 723, 1982 (M. Fleming).
47. Bayer, US 4 594 467, 1986 (H. Diehl, K. Henneke, K. Wedemeyer).
48. G. Lee, H. Freedman, *Tetrahedron Lett.* **17** (1976) no. 20, 1641 – 1644.
49. Dow Chemical, US 4 079 075, 1978 (H. Freedman, G. Lee).
50. H. Alper, *Adv. Organomet. Chem.* **19** (1981) 183.
51. O. Louis-Andre, G. Gelbard, *Tetrahedron Lett.* **26** (1985) no. 7, 831 – 832.
52. Gadot, EP-A 88 401 921.7, 1988 (J. Dakka, A. Zoran, Y. Sasson).
53. J. Rasmussen, S. Heilmann, L. Krepski, H. Smith in C. Starks (ed.): "Phase-Transfer Catalysis. New Chemistry, Catalysts, and Applications," *ACS Symp. Ser.* **326** (1987) 116.
54. P. Nicholas in C. Starks (ed.): "Phase-Transfer Catalysis. New Chemistry, Catalysts, and Applications," *ACS Symp. Ser.* **326** (1987) 155.
55. W. Tagaki, I. Isao, Y. Yano, T. Okonogi, *Tetrahedron Lett.* **15** (1974) no. 30, 2587 – 2590.
56. General Electric, US 4 410 422, 1983 (D. Brunelle).
57. Y. Okahata, K. Ariga, *J. Org. Chem.* **51** (1986) 5064.
58. V. Krishnakamur, M. Sharma, *Ind. Eng. Chem. Process Des. Dev.* **23** (1984) 410.
59. Bayer, US 4 348 319, 1982 (R. Hamprecht).
60. International Flavors and Fragrances, US 4 582 634, 1986 (M. Hanna, M. Sprecker).

Phenanthrene → Hydrocarbons
Phenethyl Alcohol → Flavors and Fragrances

Phenol

Manfred Weber, INEOS Phenol GmbH, Gladbeck, Germany
Markus Weber, INEOS Phenol GmbH, Gladbeck, Germany
Michael Kleine-Boymann, INEOS Phenol GmbH, Gladbeck, Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Cumene Oxidation (Hock Process) 3
4.2. Toluene Oxidation 7
4.3. Phenol from Coal 9
4.4. Other Processes 10
4.4.1. Sulfonation of Benzene 10
4.4.2. Chlorination of Benzene 11
4.4.3. Direct Oxidation of Benzene 11
5. Environmental Protection 12
6. Quality Specifications. 13
7. Storage and Transportation 13
8. Uses and Economic Aspects 14
9. Toxicology and Occupational Health 14
10. References 16

1. Introduction

Phenol is the specific name for hydroxybenzene, C_6H_5OH *[108-95-2]*. It is the generic name for any molecule containing one or more hydroxyl groups attached to an aromatic ring. Phenol was discovered in 1834 by Runge, who isolated it from coal tar and named it carbolic acid. In 1841 Laurent obtained it in crystalline form, determined its composition, and named it phenolic acid, which was later abbreviated to phenol. Finally Kekule formulated its molecular structure in 1858. Till the end of the 1800s all phenol was recovered from coal. It was mainly used as a disinfectant. In 1897 Hoffmann was the first to synthesize acetylsalicylic acid, the active ingredient in Aspirin, from phenol in a pure and stable form. In the 1900s demand for phenol grew with the commercialization of the phenolic resins after World War I, bisphenols after World War II, and polycarbonates during the 1960s. The first synthetic phenol was produced commercially by the benzenesulfonate process. Nowadays more than 95 % of phenol, around 7.3×10^6 t in total in 2003, is produced by the cumene oxidation process, and the rest is produced via toluene oxidation or recovered from coal tar. A direct process for oxidation of benzene to phenol has still not been commercialized.

2. Physical Properties

Phenol has a melting point of 40.9 °C. At lower temperatures it crystallizes as a white mass. In the molten state it is a clear, colorless liquid. The melting and solidification point of phenol is lowered considerably by traces of water (see Fig. 1).

At temperatures up to 68.4 °C it is partially miscibile with water [1] (see Fig. 2).

Phenol is soluble in most organic solvents, for example aromatic hydrocarbons, alcohols, ketones, ethers, acids, and halogenated hydrocarbons. It is less soluble in aliphatic hydrocarbons. Phenol forms azeotropic mixtures with water and other substances [2] (see Table 1).

Phenol has the following physical properties [3–6]:

M_r	94.11
bp (101.3 kPa)	181.75 °C
mp	40.9 °C
Relative density	
0 °C	1.092
20 °C	1.071
50 °C	1.050
Vapor density (air = 1)	3.24
Refractive index	
n_D^{41}	1.5418
n_D^{50}	1.5372

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_299.pub2

(Continued)

Dynamic viscosity	
20 °C	11.41 mPa · s
50 °C	3.42 mPa · s
100 °C	1.5 mPa · s
150 °C	0.67 mPa · s
Cubic expansion coefficient (50 °C)	8.8×10^{-7} L g^{-1} K^{-1}
Dissociation constant in water	
(20 °C)	1.28×10^{-10}
Heat of fusion (41 °C)	120.6 kJ/kg
Heat of vaporization (182 °C)	511 kJ/kg
Heat of formation (20 °C)	− 160 kJ/kg
Heat of combustion	− 32 590 kJ/kg
Specific heat	
0 °C	1.256 kJ kg^{-1} K^{-1}
20 °C	1.394 kJ kg^{-1} K^{1}
50 °C	2.244 kJ kg^{-1} K^{-1}
100 °C	2.382 kJ kg^{-1} K^{-1}
Specific heat (vapor state)	
27 °C	1.105 kJ kg^{-1} K^{-1}
527 °C	2.26 kJ kg^{-1} K^{-1}
Flash point (DIN 51 758)	82 °C
Lower explosion limit in air at 101.3 kPa (corresponds to 73 °C product temperature i.e., vapor – liquid equilibrium)	1.3 vol % (50 g/m^3)
Saturation concentration in air (20 °C)	0.77 g/m^3
Ignition temperature (DIN 51 794)	595 °C
Temperature class (VDE, EN)	T 1 (> 450 °C)
Autoignition temperature	715 °C
No allocation to a hazard class according to VbF because of a solidification point > 35 °C	
Flash point of mixtures with water	
1 – 4 % water	ca. > 82 °C
> 4.5 % water	> 100 °C
Specific resistance (45 °C)	$> 1.8 \times 10^7$ Ω · cm
Mixture with 5 % water	4.9×10^6 Ω · cm
Vapor pressure	
20 °C	0.02 kPa
36.1 °C	0.1 kPa
48.5 °C	0.25 kPa
58.3 °C	0.5 kPa
69 °C	1 kPa
84.8 °C	2.5 kPa
98.7 °C	5 kPa
114.5 °C	10 kPa
137.3 °C	25 kPa
158 °C	50 kPa
181.5 °C	100 kPa
207.5 °C	200 kPa
247.7 °C	500 kPa
283 °C	1000 kPa

3. Chemical Properties

The chemical properties of phenol are characterized by the mutual influence of the hydroxyl group and the aromatic ring upon each other.

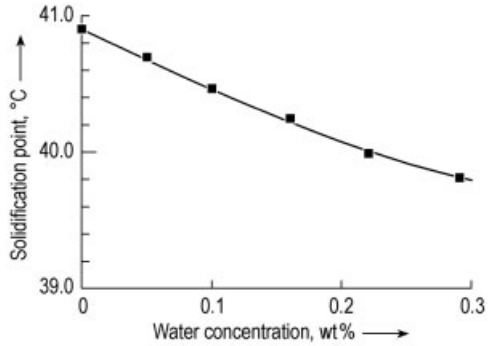

Figure 1. Effect of water on the solidification point of phenol

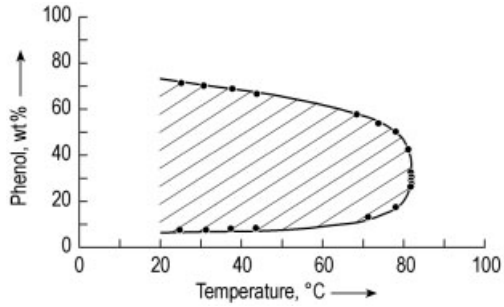

Figure 2. Miscibility of phenol and water

Reactions Involving the O–H Bond. In contrast to aliphatic alcohols phenol is a weak acid and therefore reacts with strong bases to form salts which are soluble in water. This effect is used to recover phenol from organic liquids such as coal tar. The pK_a value in aqueous solution is 10.0. The phenolate anion is stabilized by the aromatic ring system. At room temperature phenol can be liberated from the salts even with carbon dioxide. At temperatures near the boiling point of phenol, it can displace carboxylic acids such as acetic acid from their salts, whereby phenolates are formed. The hydoxyl group can easily be esterified (e.g., to phenyl acetate) and etherified (e.g., diphenyl ether).

Reactions of the Aromatic Ring. The acidity of phenol is due to the fact that the unshared electron pair is delocalized over the aromatic ring and thus stabilized [7].

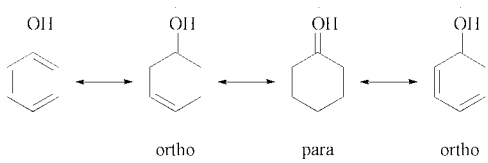

Table 1. Azeotropic mixtures with phenol

Second component	Pressure, kPa	bp, °C	Proportion of phenol, %
Water	101.3	94.5	9.21
	70.8	90.0	8.29
	39.2	75.0	7.2
	16.9	56.3	5.5
Isopropylbenzene	101.3	149	2
n-Propylbenzene	101.3	158.5	14
α-Methylstyrene	101.3	162	7

The excess of electrons in the aromatic ring at the *ortho* and *para* positions makes phenol very susceptible to electrophilic substitution. An example is the reaction with bromine in aqueous solution to give 2,4,6-tribromophenol in high yield. Another important reaction is the coupling of carbon dioxide with sodium phenoxide at 125 °C to give the sodium salicylate (Kolbe reaction). At 125 °C carbon dioxide is substituted in the *ortho* position, whereas at temperatures above 250 °C the *para* isomer is favored. Other reactions following electrophilic substitution are chlorination, sulfonation, and nitration.

Liquid-Phase Oxidation. The oxidation of organic substances by molecular oxygen leads to a large number of intermediate products, for example, cumene hydroperoxide. These processes proceed with a complex radical mechanism [8]. Phenol acts as an inhibitor in such reactions by reacting with free radicals to form phenoxy radicals. These are relatively stable radicals which do not propagate the chain mechanism. Therefore any unwanted ingress of phenol into such oxidation processes must be avoided.

4. Production

Today the oxidation of cumene (Hock process) is by far the dominant synthetic route to phenol with a production rate of more than 6.7×10^6 t/a of phenol and about 4.1×10^6 t/a of the coproduct acetone. It is the most competitive process, supported by the demand for acetone. As the demand for phenol may increase more strongly in the future, there is still a commercial interest in finding new, coproduct-free routes to phenol, for example, the direct oxidation of benzene to phenol. The oxidation of toluene with benzoic acid as intermediate product is used for phenol production with an annual production of about $300\,000 \times 10^6$ t. About $300\,000 \times 10^6$ t of phenol is still recovered from coal tar. Since ca. 1995 the direct oxidation of benzene to phenol with nitrous oxide has been developed by Solutia but has still not been commercialized. Classic synthetic routes such as the benzene sulfonation process and hydrolysis of chlorobenzene with caustic soda no longer have any economic importance.

4.1. Cumene Oxidation (Hock Process)

General Overview. The process consists of two fundamental chemical reactions: cumene is at first oxidized with oxygen to cumene hydroperoxide (CHP). CHP is then cleaved to phenol and acetone by using a strong mineral acid as catalyst. Both reactions are very exothermic with heats of reaction of -117 kJ/mol for the oxidation of cumene and -252 kJ/mol for the cleavage of the peroxide.

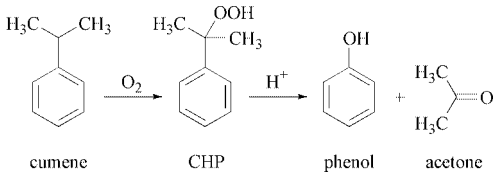

These reactions were first reported in 1944 by H. Hock and S. Lang [9]. The industrial cumene oxidation process was developed and commercialized after World War II by The Distillers Co. in Great Britain and the Hercules Powder Company in the USA [10, 11]. The first plant started up in 1952 at Shawinigan, Canada, with a capacity of 8000 t/a. Today the biggest producers of phenol using cumene oxidation are INEOS Phenol (Europe and USA), Sunoco (USA), Mitsui (Japan, Singapore), Shell (USA), Polymeri Europa (Italy), and Ertisa (Spain). Some developed their own technology. Others use a licensed

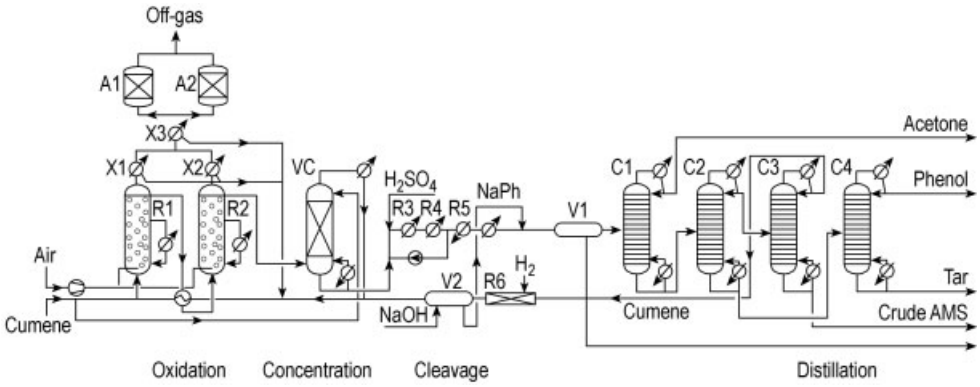

Figure 3. Production of phenol from cumene
R1, R2 cumene oxidation reactor; X1 – X3 condensers; A1, A2 activated-carbon adsorbers; VC vacuum distillation column; R3 – R5 cleavage reactors; V1 neutralization drum; C1 acetone column; C2 cumene column; C3 AMS column; C4 phenol column; R6 hydrogenation reactor; V2 cumene scrubber

process from, for example, Kellogg Brown and Root or UOP. Although all use the principles of the Hock synthesis, there are differences in the technical design of the reaction and fractionation sections. The key item for the overall process yield is the formation of the byproducts dimethylbenzyl alcohol (DMBA) and acetophenone (ACP) in the oxidation step. ACP leaves the process with the high boilers from distillation. DMBA is dehydrated in the cleavage stage to α-methylstyrene (AMS), which partially forms high-boiling dimers and cumylphenols in the acid-catalyzed process step. The unconverted AMS is hydrogenated back to cumene later in the process.

The flow sheet of the cumene oxidation process is shown in Figure 3. The oxidation of cumene as the first reaction step is usually carried out in a series of bubble column reactors R1 and R2. Fresh and recycle cumene are fed to the first reactor. Oxygen is used in each reactor as oxidant. The bubble columns are operated at pressures ranging from atmospheric to about 700 kPa and temperatures in the range 80 – 120 °C. The temperature is controled by internal or external heat exchangers [12]. The final CHP concentration varies typically from 20 to 40 % [13]. The off-gas from the reactor is purified by two-step condensation of organic components with cooling water in condensers X1, X2 and chilled water or other refrigerants in X3. The organic components are mainly cumene and methanol besides some smaller amount of water. The residual impurities in the off-gas are finally removed in activated carbon adsorbers A1, A2. Layers of different types of activated carbon may be used for better co-adsorption of cumene and methanol [14]. The condensed cumene is recycled to the oxidation reactors. The CHP concentration in the oxidation product is increased by vacuum distillation in column VC to about 65 – 90 %.

The first part of the cleavage section is a series of heat exchangers R3, R4 within a circulation loop. Heat exchangers are used as reactors because the reaction is very fast and exothermic. The CHP is spontaneously diluted in the circulation product, which consists of phenol, acetone, cumene, and several byproducts as well as residual amounts of CHP. Normally sulfuric acid is used as the catalyst at a concentration of several hundred parts per million by weight (wppm). As the acid is completely dissolved in the reaction mass this type of process is called homogeneous cleavage. The temperature is in the range 40 – 80 °C at atmospheric pressure. The product from the first step, so-called cleavage product, is heated in R5 to temperatures above 100 °C. In this second step residual CHP is cleaved, as is the intermediate dicumyl peroxide (DCP), which is formed in the first step by condensation of CHP and DMBA. Finally residual DMBA is dehydrated to AMS and water. The product is cooled and neutralized with aqueous sodium phenolate NaPh. An aqueous phase is formed containing sodium sulfate and salts of organic acids. The aqueous phase is separated in neutralization drum V1 and sent to wastewater treatment.

The organic phase, consisting of phenol, acetone, cumene, AMS, dissolved water, and low- and high-boiling byproducts, is separated in a series of distillation columns.

More details on several process steps follow.

Oxidation. The oxidation of cumene with oxygen in the liquid phase proceeds by a complex radical mechanism [15]. The reaction is autocatalyzed by cumene hydroperoxide. Thus, the reaction rate to CHP not only depends on the cumene but also on the CHP concentration. Phenol is a strong inhibitor, so any recycled cumene must be treated to remove even traces of phenol. Formation of the byproducts DMBA and ACP depends on the CHP concentration [16, 17]. The activation energy for these side reactions are higher than for the main reaction. Thus, the oxidation should be run at low temperatures to suppress the formation of DMBA and ACP.

As the reaction rate depends on the CHP concentration, backmixed reactors are preferred for cumene oxidation to give high CHP concentrations. Bubble columns, in which the liquid phase is well backmixed, are therefore suitable for this process. On the other hand, high CHP levels cause more DMBA formation. Thus a series of several bubble columns is the optimal reactor concept [17].

The oxidation is operated either under overpressure at temperatures between 90 and 120 °C, or under atmospheric pressure at temperatures below 100 °C. Depending on temperature, the residence time in the oxidizer train is between 4 and 20 h [18]. Hattori et al. [19] investigated the influence of the partial pressure of oxygen on the reaction rate to CHP. As a rule of thumb, the partial pressure of oxygen in the gas phase should be above 0.1 bar for temperatures around 100 °C to ensure sufficient mass transfer and thus avoid hindrance of the reaction rate. This is an important issue because atmospheric oxygen is consumed from ca. 21 vol % to some residual value at the top of the bubble column. Thus, depending on the absolute pressure, the oxygen partial pressure also decreases but should not fall below the above-mentioned value. Lower temperatures than 100 °C may require lower partial pressures of oxygen. The oxidation of cumene with pure oxygen in stirred vessels has been already described [11].

In the early years of the cumene oxidation process, the oxidation was carried out as a three-phase reaction [20–22], so-called wet oxidation. In addition to cumene and air an aqueous sodium carbonate solution was continuously added to the reactors to extract and neutralize organic acids, which are formed during oxidation. Phenolchemie, now INEOS Phenol, was the first to operate the reactors without adding any caustic soda or sodium carbonate [23], a process known as dry oxidation.

CHP is a relatively stable organic peroxide. Nevertheless the thermal decomposition of CHP is the main source for DMPC formation in the oxidation process. Especially in large reactors the exothermic decomposition of CHP has to be taken into account during a shutdown when the reactors are not mixed by aeration. The reactor must be cooled down to below the self-accelerating decompositon temperature (SADT). Below the SADT heat removal by heat losses is higher than heat production by thermal decomposition, so that the reactor will not heat up [17, 24].

Concentration is performed by one- or multiple-step vacuum distillation [11]. Cumene is separated from concentrated CHP at temperatures below 100 °C to minimize thermal decomposition of the organic peroxide. As traces of organic acids are present in the overheads the cumene must be washed with aqueous caustic soda or sodium carbonate before it is recycled to the oxidation. To minimize any acid-catalyzed decomposition of CHP in the concentration step, fresh cumene is taken as reflux [25].

Cleavage. The first stage of homogeneous cleavage is performed either in a circulation loop with one or more heat exchangers in series, as shown in Figure 3 [26–28] or in a backmixed vessel with internal or external cooling devices [18, 29] (see Fig. 4).

In the backmixed vessel with internal cooling the temperature is, for example, maintained by vaporizing and refluxing acetone (Fig. 4a). Recycle acetone from the distillation is added to the CHP before the reaction vessel [11]. Another possibility is to remove the heat from the reaction vessel with an external heat exchanger in a loop (Fig. 4b).

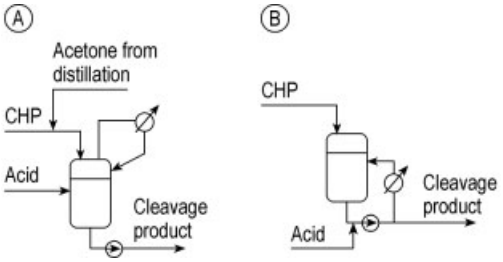

Figure 4. Cleavage of cumene hydroperoxide (CHP) in backmixed vessels
A: Removal of heat by vaporizing and refluxing acetone;
B: Removal of heat by external cooling

In heterogeneous cleavage the first stage is carried out in a circulation loop of 40 wt % aqueous sulfuric acid [11, 30, 31] (Fig. 5). The heat is removed by a series of heat exchangers in the loop. Cumene hydroperoxide (CHP) from the concentration step is added to the acid on the suction side of the circulation pump in a ratio of 1:4. After the last cooler the organic phase is separated from the aqueous acid, and the acid is recycled to the pump.

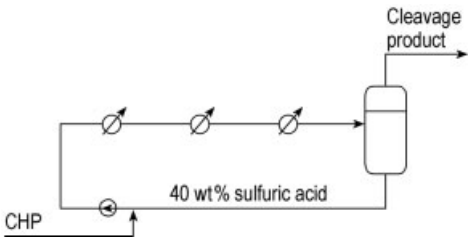

Figure 5. Heterogeneous cleavage of cumene hydroperoxide (CHP)

The homogeneous cleavage of CHP to phenol and acetone is first-order both in CHP and in acid [32]. Only small amounts of acid are continuously added to the reactor. The acid activity is considerably dependent on the solvent system [13]. Especially the amount of dissolved water, which is formed by dehydration of dimethylbenzyl alcohol (DMBA), has a major influence on the acid activity [33]. In the first stage of cleavage CHP partially condenses with DMBA to form the intermediate dicumyl peroxide (DCP) and water [29]. Other byproducts are hydroxyacetone from the reaction of CHP with acetone, 2-methylbenzofurane from the reaction of hydroxyacetone with phenol, diacetone alcohol from the aldol condensation of acetone, and *mesityl oxide* from the dehydration of diacetone alcohol. An overview of these reactions is given in [34].

The product from the first cleavage stage contains residual amounts of CHP and DMBA as well as the intermediate DCP. The reaction is completed in a short-residence-time plug-flow reactor in the second stage of the cleavage (see Fig. 3). The product is heated to temperatures above 100 °C [27, 29, 35]. DCP and water react back to CHP and DMBA, CHP is cleaved spontaneously to phenol and acetone, and DMBA is completely dehydrated to α-methylstyrene and water.

After cooling, all reactions are stopped by adding a neutralizing agent such as aqueous sodium phenolate to the cleavage product (see Fig. 3). Salts from the neutralization are extracted with the aqueous phase. It is important to neutralize the sulfuric acid and the organic acids to avoid corrosion in the distillation stage. On the other hand, salts from the neutralization should be extracted thoroughly with the aqueous phase to avoid fouling and plugging of the evaporators in distillation. Another possibilty is to neutralize the cleavage product on a caustic-exchanged sulfonic ion-exchange resin [13, 36].

Distillation. In the first distillation column C1 (see Fig. 3), acetone is separated as the top product. This acetone stream contains mostly water and low boiling aldehydes as impurities. For high-purity acetone, the top stream must be separated in an additional reactive distillation column [37]. Caustic soda is introduced over the feed position in this reactive distillation column. The caustic soda catalyzes condensation reactions of the aldehydes to form high boilers. A side stream above the feed contains pure acetone. The top stream of this column is recycled back to the cleavage. Cumene and AMS are taken as overheads in column C2. In column C3 the cumene – AMS fraction is separated into cumene as a top product and crude AMS as the bottom stream. This crude AMS can be further purified to obtain pure AMS. The top product contains mostly cumene and some AMS. This stream is sent to the hydrogenation, where AMS is hydrogenated to cumene. The hydrogenation takes place in fixed-bed reactor R6. After removal of residual phenol by extraction with caustic soda (cumene scrubber

V2), the reactor outlet is recycled back to oxidation. The aqueous sodium phenolate from the cumene scrubber is used as neutralization agent for the cleavage product. Pure phenol is taken as the overhead from column C4. The bottom stream of column C4 consists mainly of acetophenone, AMS dimers, cumylphenols, and tar and may be disposed of by incineration. In older processes the tar is treated thermally to recover cumene, AMS, and phenol [11]. A new application is the use of the desalted tar for production of carbon black [38]. Several minor impurities which are formed in cleavage, such as hydroxyacetone and mesityl oxide, can not be easily separated from phenol by simple distillation. For ultrahigh-purity phenol some subsequent purification steps are needed. The phenol from the top stream of column C4 can be treated with an acidic ion exchange resin in a fixed-bed reactor in the fluid phase. This catalyzes reactions of ketones and aldehydes. These reactions generate low and high boilers, so that further distillation columns are needed to finally obtain pure phenol [39]. As an alternative the phenol can be separated from impurities by extractive distillation [40, 41]. Polyalkylene glycol, alkanediols, and ethanolamines are used as extractive agents. The extractive agent is introduced in a first column above the feed. The interaction between the phenol and the extractive agent results in their being in the bottom stream. The impurities leave the column over the top stream. A second column is used to separate the extractive agent from the pure phenol. Water can also be added to enhance the separation of minor impurities [42]. Another possibility for the purification of the phenol stream is to add amines [43]. The amine reacts in a first step with the ketones and in a second step with aldehydes to high boilers that can be easily separated from phenol by distillation.

4.2. Toluene Oxidation

General Overview. In the 1950s it appeared that toluene would be available in large quantities and it would be a cheaper source for the aromatic ring than benzene. One suggestion was to use toluene for the production of phenol. This process was developed by the California Research Corp. [44] and Dow Chemical [45]. The two-steps of the process are first the oxidation of toluene to benzoic acid, which was already well known, and second the oxidation of benzoic acid to phenol. In principle the conversion of toluene to phenol is a complete oxidation of the methyl group with the byproducts carbon dioxide and water.

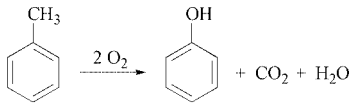

The oxidation of toluene as the first step is carried out in the liquid phase at temperatures about 100 – 150 °C and a gauge pressure of about 2 bar [46–48].

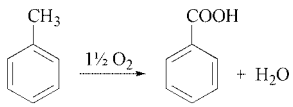

Atmospheric oxygen is used. A cobalt naphthenate is used as a soluble catalyst at concentrations between 0.1 and 0.3 %. The oxidation of toluene to benzoic acid has a complex radical chain mechanism. Benzyl alcohol and benzaldehyde are intermediates in the successive oxidation to benzoic acid. Byproducts such as benzyl benzoate and biphenyl are formed, as well as formic acid, acetic acid, and, in smaller amounts, carbon dioxide and carbon monoxide [46, 47].

For the oxidation of benzoic acid as the second step molten benzoic acid is used as the reactant and solvent medium. The temperature is about 230 – 240 °C at atmospheric pressure [46–48]. Benzoic acid reacts with atmospheric oxygen, steam, and soluble copper(II) benzoate as catalyst [46, 49]. Magnesium salts may be added to act as a promotor [50]. The formation of phenol is the net result of several reactions which occur simultaneously in the system. Copper(II) benzoate thermally decomposes to copper(I) benzoate and benzoylsalicylic acid:

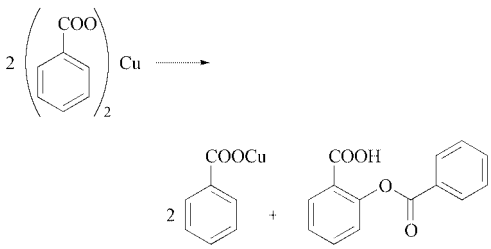

The copper(I) benzoate from the first reaction step is regenerated with atmospheric oxygen:

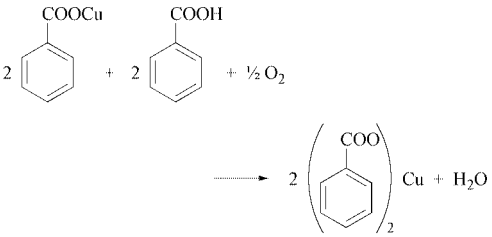

In the presence of water or steam, benzoylsalicylic acid is hydrolyzed to benzoic acid and salicylic acid:

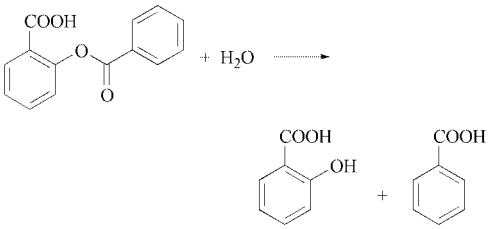

The benzoylsalicylic acid is decarboxylated rapidly to give phenol and carbon dioxide:

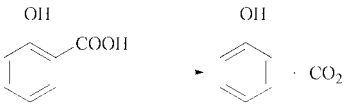

Thus, the net reaction for the oxidation of benzoic acid is:

Byproducts include benzene and diphenyl ether.

Dow Process. The first commercial plant based on the Dow process was started up in the early 1960s at Landar, British Columbia, operated by Dow-Canada (shut down), and the second at Kalama, Washington, operated by Dow Chemical Company (capacity in 2002 ca. 32 000 t/a phenol). The third plant was built at Rosenburg, Netherlands, operated by DSM (capacity in 2002 ca. 110 000 t/a phenol). In 1991 a new plant was brought on line in Japan by Nippon Phenol (capacity in 2002 ca. 120 000 t/a phenol).

Simplified process flow diagrams of the Dow process are given in [46–48, 51, 52]. Figure 6 gives an overview of the process. Fresh and recycle toluene as well as air and the cobalt catalyst are charged to gas – liquid reactor R1, which can be, for example, a bubble column. The reactor has cooling devices to remove the reaction heat, for example, in an external circulation loop as shown in Figure 6. The toluene conversion is limited to 40 %. The yield of benzoic acid is about 90 %. The off-gas from the reactor is cooled by passing it through a water-cooled condenser. The condensed toluene is recycled to the reactor, and the water phase is withdrawn for further wastewater treatment. The noncondensable gas, principally nitrogen, is treated in activated-carbon adsorption unit A1 to remove traces of toluene and other organic components. The liquid phase from the reactor is at first fractionated in column C1. Toluene and the intermediate benzaldehyde are taken overhead for recycle. The crude benzoic acid as the bottom product is purified in the second column C2. The catalyst containing high boilers (tar) is taken as the bottom product, which is partially recycled to the reactor.

Purified benzoic acid is oxidized in the second reactor R2, which is a stirred gas – liquid reactor. In the presence of the copper(I/II) catalyst the acid is contacted with air and recycled vaporized water. In addition steam may be added to the air. Small amounts of magnesium benzoate may be added as a promoter. The benzoic acid conversion is 70 – 80 % with a yield of phenol of about 90 %. Phenol, benzoic acid, water, and other organic components are vaporized continuously. The vapors are charged to column C3. Water and hydrocarbons are taken as the overhead fraction and separated into two liquid phases. The remaining gas phase is purified in adsorption unit A2. The separated water is used to extract benzoic acid from the tar which has to be removed from reactor R2. This is done in extraction unit V1, which consists of a series of several mixer – settlers. The aqueous phase from V1 is vaporized and recycled to reactor R1. The

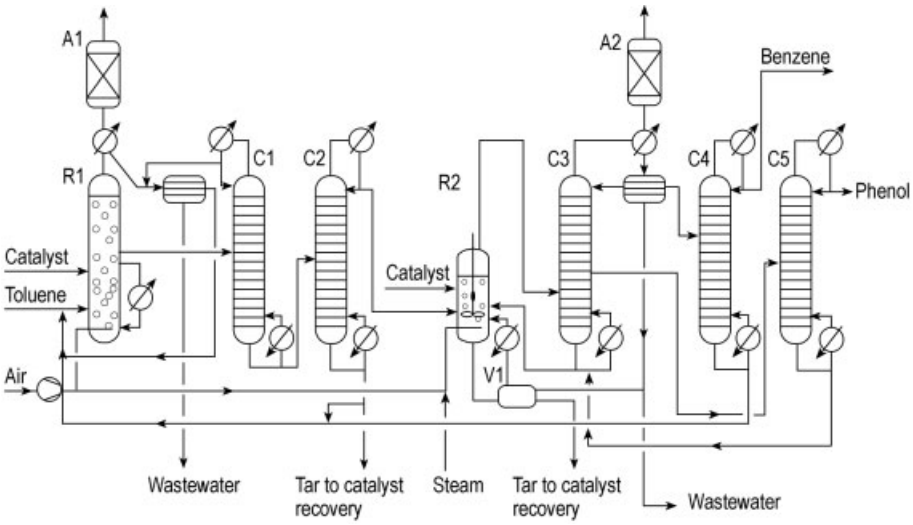

Figure 6. Production of phenol from toluene
R1 toluene oxidation reactor; A1 adsorption unit; C1 toluene column; C2 benzoic acid column; R2 benzoic acid oxidation reactor; A2 adsorption unit; C3 crude phenol column; C4 benzene column; C5 pure phenol column; V1 tar extraction unit

tar contains the copper salts. If it is burned special incineration equipment is required to insure that no heavy metal emissions occur. The organic phase from the overheads of column C3 contains mainly toluene and benzene. Benzene is separated from toluene in column C4. Toluene is recycled to the oxidation reactor R1. Crude phenol is taken as a side cut from column C3. In Column C5 pure phenol is taken as the overhead. Benzoic acid is withdrawn as the bottom product and recycled to reactor R2. Because of the corrosive character of benzoic acid apparatus must be constructed from stainless steel 1.4571 or even from Hastelloy.

Lummus Process. In 1978 The Lummus Company announced a new process for the oxidation of benzoic acid [53]. The oxidation is conducted in the vapor phase over a copper-containing catalyst [54]. Instead of using a gas – liquid reactor (R2 in Figure 6), the reaction is conducted in a vapor-phase reactor on a heterogeneous catalyst; this is the main difference from the Dow process. Lummus announced that there is no formation of tars or nonvolatile residues. This is probably due to the fact that phenol is rapidly removed from the reaction zone. It is believed that the residues in the gas – liquid process are derived from phenol. As there is no need for tar treatment and incineration Lummus claims that the fixed investment is significantly lower for their process [53]. However, the process has not yet been commercialized.

4.3. Phenol from Coal

The recovery of "natural" phenol from coal is the oldest commercial process for producing phenol and it is still practiced. Nowadays, this process is only of little importance as a source of phenol.

Phenol, as well as many other aromatic compounds, is recovered from the vapors of coal carbonization (Fig. 7). Coal is heated to above 900 °C with the exclusion of air in a coke oven. Condensable hydrocarbons and gases are removed as vapors leaving coke as the solid residue. The vapors are condensed to form coal tar. The gas may be washed with aqueous sulfuric acid or phosphoric acid to recover ammonia. It can also be washed with a wash oil to further remove benzene. The remaining gas, which consists of hydrogen and methane, is used as a fuel on site [55].

The fractionation of coal tar is described in detail in [56]. The simplied process diagram shown Figure 7 is based on information about the Rütgers Process given in [57]. The coal tar is further processed by continuous distillation. In

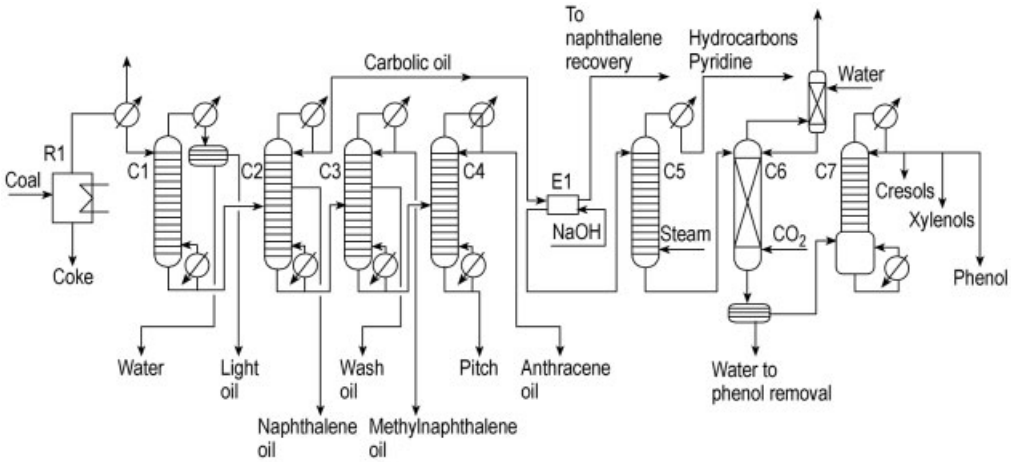

Figure 7. Production of phenol from coal
R1 coke oven; C1 light oil/water column; C2 carbolic oil column; C3 naphthalene oil column; C4 anthracene oil column; E1 extraction unit; C5 hydrocarbon removal column; C6 precipitation column; C7 fractionation column

a first column C1 light oil, which contains benzene, toluene, and xylenes, is separated together with water. In the second column C2 carbolic oil, containing phenol and phenolic compounds, is taken as the overhead and naphthalene oil as a side cut. The bottom product from C2 is further fractionated in columns C4 – C6 to give wash oil for benzene recovery from the off-gas, methylnaphthalene oil, anthracene oil, and pitch. Plate columns or columns with random packings are used with high reflux ratios and a high degree of thermal integration.

The carbolic oil, which constitutes ca. 2 – 3 % of the coal tar, contains about 30 – 35% of phenolic compounds such as phenol and cresols. Other components are non-phenolic benzene derivatives and pyridine bases. The carbolic oil is treated with 8 – 9 % caustic soda in extraction unit E1. Phenol and cresols, which are known as tar acids, are extracted as salts into the aqueous phase, which is called phenolate lye. The organic phase is further treated to recover naphthalene by crystallization. The crude phenolate lye is treated by steam distillation in column C5 to separate residual hydrocarbons and pyridine bases. The phenolate lye is then treated with carbon dioxide to liberate the acid tars from their salts in column C6. The aqueous phase containing sodium carbonate is further treated to remove dissolved phenol. The organic phase is fractionated into pure phenol (*bp* 181.8 °C), *o*-cresole (*bp* 191.0 °C), a mixture of *m*- and *p*-cresoles (*bp* 202.0 – 203.0 °C), and a xylenol fraction (*bp* up to 210 °C) by batch distillation.

4.4. Other Processes

4.4.1. Sulfonation of Benzene

The production of phenol via sulfonation of benzene is the oldest process to produce synthetic phenol. It is based on research by Wurtz and Kekule in 1867, who found that heating sodium benzenesulfonate in molten sodium hydroxide gives sodium phenolate:

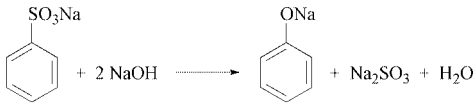

Sodium benzenesulfonate is produced by reaction of benzene with concentrated sulfuric acid to give benzenesulfonic acid and subsequent neutralization with sodium sulfite or sodium carbonate. Phenol is liberated from the sodium phenolate by treating the aqueous solution with carbon dioxide.

Phenol is no longer produced by this process, details of which can be found in [50, 51, 58].

4.4.2. Chlorination of Benzene

Alkaline Hydrolysis. DURANT and BARDY discovered in 1872 that the treatment of chlorobenzene with aqueous caustic soda affords sodium phenolate and sodium chloride.

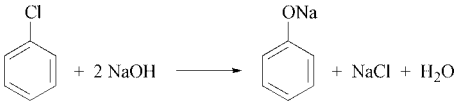

Chlorobenzene is produced from benzene and chlorine. Typically the process was combined with chlor-alkali electrolysis to recover chlorine from the sodium chloride.

Raschig-Hooker Process. The reaction of chlorobenzene with steam to give phenol and hydrochloric acid was discovered by MEYER and BERGIUS in 1914. It was commercialized as the Raschig – Hooker process at BASF.

Chlorobenzene is produced via oxychlorination of benzene with hydrochloric acid and oxygen. Thus the hydrochloric acid is recycled and the process is therefore also called the regenerative Raschig process.

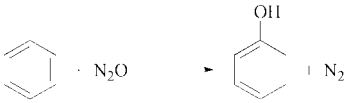

Details about these processes, in which phenol is made from the intermediate chlorobenzene, are given in [50, 51, 58]. They are no longer used for commercial production of phenol.

4.4.3. Direct Oxidation of Benzene

General. The direct oxidation of benzene would be the simplest process to produce phenol. For example, benzene can easily be oxidized by air, but the phenol which is formed is further oxidized much more easily than benzene. It reacts very fast with oxygen or any other oxidizing agent in successive reactions that finally form carbon dioxide and water. Therefore, the conversion of benzene must be limited to only a few per cent. Direct oxidation of benzene has not yet been used for commercial production of phenol. However several attempts were made since the mid-1990s to use nitrous oxide as the oxidzing agent in a heterogeneous gas-phase process.

Direct Oxidation with Oxygen or Hydrogen Peroxide. In [58] a survey is given on the first patents, granted in the 1920s and 1930s. For example, benzene can be oxidized to phenol with air in the gas phase at 600 – 800 °C without any catalyst. A commercial plant of Solvay Process Company using this process was operated between 1940 and 1947. In [59–61] an overview is given on the use of oxygen or hydrogen peroxide on different heterogeneous catalysts. With oxygen the selectivity of phenol formation with respect to benzene can reach more than 90 %, but only if the benzene conversion is limited to a few per cent. When hydrogen peroxide is used the conversion may reach more than 30 % with a selectivity of more than 95 %. Titanium silicalites (TS-1) are typically used as a catalysts for this. However, because of the high price of hydrogen peroxide, it is not economic to produce phenol in this way.

Direct Oxidation with Nitrous Oxide. The direct oxidation of benzene with nitrous oxide N_2O was mainly developed by Solutia, formerly Monsanto, in cooperation with the Boreskov Institute of Catalysis in Russia [62].

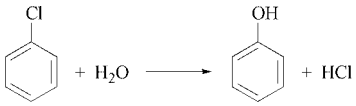

N_2O is a major component in the off-gas from adipic acid production plants. Solutia produces 400 000 t/a of adipic acid from cyclohexane [63], which is associated with the formation of about 200 000 t/a of nitrous oxide [59]. Nitrous oxide is a greenhouse gas and must be decomposed to avoid emission [64]. The use of nitrous oxide for the direct oxidation of benzene would therefore be ideally suited to any adipic acid producer.

The oxidation is carried out in the gas phase with zeolites of the ZSM-5 type as the heterogeneous catalyst. The heat evolved is 286 kJ/mol [61]. The reaction temperature is in the range between 300 and 450 °C at atmospheric pressure. The reaction time is 1 – 2 s. The conversion is up to 40 % with selectivities of benzene to phenol of more than 95 %. The space – time yield is 0.24 – 0.64 g phenol per gram catalyst per hour.

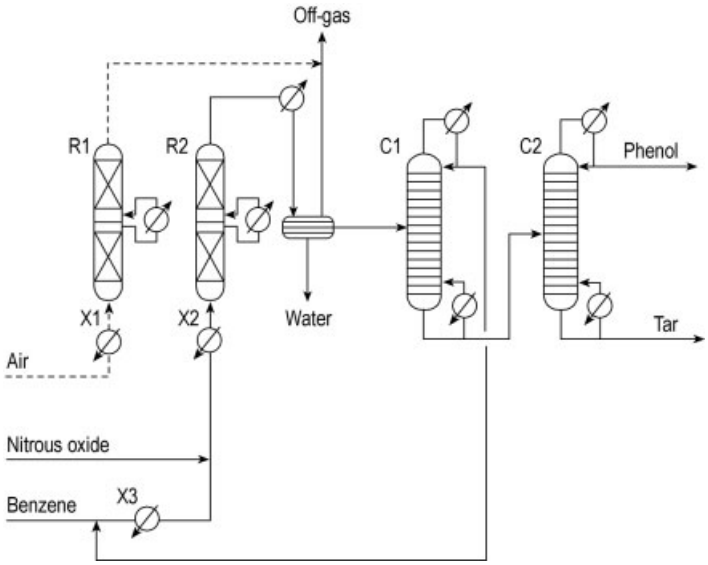

Figure 8. Production of phenol by direct oxidation of benzene with nitrous oxide
R1, R2 reactors; X1, X2 gas heaters; X3 benzene vaporizer; C1 benzene column; C2 phenol column

Byproducts are carbon dioxide and water from total oxidation, as well as higher boiling phenol derivates such as hydroquinone. Many patents deal with the preparation of the zeolite; a survey is given in [61]. Coking deactivates the catalyst after several hours. The catalyst can be regenerated by passing an oxygen-containing gas through the catalyst bed at elevated temperatures [62].

Based on the literature and patents a simplified process flow diagram was evaluated (Fig. 8). Reactor R2 is in operation while reactor R1 is regenerated. Fresh and recycle benzene is evaporated in heat exchanger X3. After adding nitrous oxide the gas is heated to 300 °C. An adiabatic process would limit the benzene conversion to about 10 %. Therefore, the reactors may have cooling devices, for example, external interbed exchangers. The off-gas is cooled to condense benzene, water, phenol, and higher boiling byproducts. The off-gas, consisting mainly of nitrogen and carbon dioxide, is purged. In the first column C1 benzene is separated and recycled. Phenol is taken as the overhead from the second column C2.

Solutia has operated a pilot plant at their site in Pensacola, Florida, first in 1996 [62]. It is not clear whether they used pure nitrous oxide or nitrous oxide from adipic acid off-gas. Nevertheless, there may be some problems with commercialization of the process. The construction of a new plant for 136 000 t/a of phenol was delayed to post 2002 [65].

5. Environmental Protection

Wastewater Purification. Because of its toxicity phenol must be removed from wastewater. Phenol is classified in the water-hazard class 2 (hazardous to water). For example, the wastewater from neutralization in the cumene oxidation process contains about 1 – 2 % phenol. The extraction of phenol from water with an aromatic solvent is well known. In [66] the performance of cumene, acetophenone, and mesityl oxide is compared. Other suitable solvents are a mixture of cumene and α-methylstyrene or a mixture of acetophenone and dimethylphenylcarbinol [67]. For example, the partition coefficient of phenol for cumene against water is about 1.5. For effective extraction a countercurrent process is used, such as a multistep mixer – settler system or an extraction column [68]. For nearly complete phenol removal the pH of the wastewater should be below 7 to release all phenol from sodium phenolate. In this way the phenol content in the wastewater can be reduced to some 10 ppm by weight. The remaining phenol is removed in the subsequent biological waste-

Table 2. Quality specifications for phenol

Pharmacopeia	Water content %	Color	Content, %	Nonvolatile impurities, %	Solidification point, °C
USP 26/2003	≤ 0.5		99.0 – 100	0.05	≥ 39.0
Ph. Eur., 4th ed., 2002		colorless, faintly pink, or faintly yellow	99.0 – 100	0.05	≥ 39.0
ASTM D 2439	≤ 0.1				≥ 40.6

Table 3. Quality specifications for acetone

Pharmacopeia	Water content, %	Density	Content, %	Nonvolatile impurities	Free acid (as acetic acid), %	Permanganate test, t/min
USP 26/2003		0.789	≥ 99.0	≤ 50 mg/kg		15
Ph. Eur. 4th ed., 2002	≤ 0.38	0.790 – 0.793		≤ 50 mg/kg		120
ASTM D 329	≤ 0.5	0.7910 – 0.793	≥ 99.5	5 mg/100 mL	0.002	30
DIN 53247	≤ 0.3	0.7900 – 0.792	≥ 99.5	1 mg/100 mL		
BS 509/1987	≤ 0.3	0.7890 – 0.791	≥ 99.5	5 mg/100 mL	0.002	180

water treatment. Other solvents used for the extraction of phenol from wastewater are butyl acetate, methyl isobutyl ketone, and diisopropyl ether, which are typically used in the recovery of phenol from gas condensates in gas production (Phenosolvan process). The removal of phenol from wastewater can also be carried out by adsorption on, for example, activated carbon.

Waste-Gas Purification. According to TA Luft phenol is classified in emissions class I [emission must be limited to a mass stream of 0.1 kg/h or a mass concentration of 20 mg/m^3 (STP)]. Phenol in gas streams can be removed by absorption. Aqueous caustic soda is used as the scrubbing agent. The absorption columns have several layers of random packings to improve mass transfer. Adsorption on activated carbon is another possibility for removing phenol from gas streams. Phenol, as well as other organic components, can be completely incinerated. Combustion devices for this treatment are thermal or catalytic incinerators [69] and flares.

6. Quality Specifications.

The quality specifications of the various pharmacopeias are generally met by synthetically produced phenol. These specifications are limited to identity tests, assays, determinations of acidically reacting impurities, nonvolatile impurities, color, water solubility, density, solidification point and water content. Quality specifications for phenol and the coproduct acetone are listed in Tables 2 and 3, respectively. The quality specifications of the various users of phenol go beyond these criteria to some extent. Impurities such as hydroxyacetone, mesityl oxide, 2-methylbenzofuran, and other carbonyl compounds can result in discoloration of downstream products in further processing. The content of lower carboxylic acids can increase the extent of undesired side reactions. Metal ions can damage catalysts or lead to discoloration of reaction products. The concentrations of these components in phenol must therefore be limited.

Synthetic phenol is normally colorless. Color is determined by comparison with calibrated color scales.

The solidification point of phenol is a sensitive criterion for impurities. It is, however, strongly affected by the water content (see Fig. 2). The effect of water on this method of testing can be excluded by the addition of gypsum or dry 4A molecular sieves. ASTM methods (D 2439-96, D 329, D 1363, D 1686-96, etc.) are used as the methods of investigation. Factory specifications of the producer and manufacturers are also used and are much stricter. Especially when phenol is used for Bisphenol A production it has less than 100 ppm impurities.

7. Storage and Transportation

Large quantities of phenol are transported as liquids in ships (with double walls) and road or rail tank cars, all of which are equipped with heating coils. Steel is generally used as the container

material. If the phenol is to remain colorless for as long as possible, the walls of the containers must be made of stainless steel. Pure phenol is therefore stored predominantly in stainless steel tanks, which are insulated to avoid heat losses. On storing liquid phenol it must be ensured that the temperature remains below 70 °C. The lower explosion limit for phenol – air mixtures is reached at a saturation temperature of 73 °C, corresponding to a phenol vapor fraction in the mixture of 1.3 vol %. At storage temperatures below 70 °C the tank gas volume is blanketed with nitrogen.

Smaller quantities of phenol are transported in steel drums. If the phenol contains water, steel drums are not suitable because of the coloration of the phenol which occurs. In this case drums of galvanized sheet metal are necessary.

For phenol the following transport regulations are in force:

Land transportation ADR/RID (cross-border/inland)	
ADR/RID	Class 6.1 (T1) toxic substances
UN No.	2312
Packaging group	II
Inland waterway transportation ADN/ADNR	
ADN/ADNR	Class 6.1
UN No.	2312
Packaging group	II
Sea transportation IMDG/GGVSee	
IMDG/GGVSee	Class 6.1
UN-Number	2312
Packaging group	II
EMS No.	6.1–02

8. Uses and Economic Aspects

In 2003 about 7.3×10^6 t of phenol was produced. The total production capacity worldwide is more than 9×10^6 t. Around 37 % of global output, is used for the production of bisphenol A [70]. Bisphenol A is especially used for the production of high-grade polycarbonates for compact discs, for glazing, and for the automotive industry. Bisphenol A is also used for the production of epoxy resins. The second largest consumption of phenol is for the production of phenolic resins with formaldehyde. They are mainly used for underseal applications in the automotive industry. Phenol is also used for the production of caprolactam via cyclohexanol – cyclohexanone. Many other derivatives of phenol are produced, such as aniline, alkylphenols, diphenols, and salicylic acid. Table 4 shows the distribution of phenol capacity and demand in 2000 and 2003 and the forecast for 2007 and 2010 [70]. As the automotive industry is an important consumer of phenol derivatives such as polycarbonates and phenolic resins, the entire phenol demand still tends to be cyclical and closely tied to this market.

The dominant synthetic route is cumene oxidation, which accounts for about 95 % of the total phenol production. This process will maintain its leading position as long as a market exists for the coproduct acetone. Production from toluene oxidation and from coal each amount to about 300 000 t/a. Toluene oxidation is only economical at a low toluene price relative to that of benzene. The latest plant based on this technology was built in 1991 in Japan. This indicates that the economics of the toluene oxidation process are not competitive with the cumene oxidation process. As long as coal tar is used as the natural source for aromatics phenol will be recovered in this way, too. The classic synthetic routes such as sulfonation or chlorination of benzene are no longer used because of the formation of considerable quantities of sodium salts as byproducts. Commercial production of phenol by direct oxidation of benzene with byproduct nitrous oxide does not seem to be viable in the near future.

9. Toxicology and Occupational Health

Phenol occurs as a metabolic product of animals and humans in the blood, urine, and feces in free form as well as in conjugated form as a glucuronide and sulfate. Excretion via urine is the main elimination pathway of phenol metabolites. An estimated 100 000 t/a of phenol (metabolites) are released worldwide in human excrement [71].

Phenol is harmless in small quantities according to present knowledge. Because of its germicidal effect phenol was long used in the 1800s as a very effective disinfectant (carbolic acid, Lister spray). However, because of its protein-degenerating effect, it often had a severely corrosive action on the skin and mucous membranes.

Table 4. Worldwide production capacity and demand for phenol (in 1000 t/a)

	2000	2003	2007[a]	2010[a]
Production capacity from				
Cumene	7641	8490	9879	10079
Toluene	307	307	307	307
Coal tar	291	291	291	291
Total	8239	9088	10477	10677
Demand for				
Bisphenol A	2306	2678	3446	3864
Phenolic resins	2236	2419	2564	2709
Caprolactam	739	705	724	728
Others	1335	1454	1576	1686
Total	6616	7256	8309	8987

[a]Forecast.

Today phenol has only limited use in pharmaceuticals because of its toxicity.

Toxicology. Phenol is toxic in high concentrations. It can be absorbed through the skin, by inhalation, and by swallowing. The typical main absorption route is the skin, through which phenol is absorbed relatively quickly, simultaneously causing caustic burns on the affected area of skin. Besides the corrosive effect, phenol can also cause sensitization of the skin in some cases. Resorptive poisoning by larger quantities of phenol, which is possible even over small affected areas of skin, rapidly leads to paralysis of the CNS with collapse and a severe drop in body temperature. Skin that is wetted with phenol or phenolic solutions must therefore be decontaminated immediately. After removal of contaminated clothing, polyglycols (e.g., lutrol) are particularly suitable for washing the skin.

On skin contamination local anesthesia sets in after an initial painful irritation of the affected area of skin. Hereby the danger exists that possible resorptive poisoning is underestimated. If phenol penetrates deep into the tissue, this can lead to phenol gangrene through damage to vessels. The effect of phenol on the CNS, namely, sudden collapse and loss of consciousness, is the same for humans and animals. In animals a state of cramp precedes these symptoms because of the effect phenol has on motor activity controlled by the CNS.

Caustic burns on the cornea heal with scarred defects. Possible results of inhalation of phenol vapor or mist are dyspnea, coughing, cyanosis, and lung edema. Swallowing phenol can lead to caustic burns inn the mouth and esophagus and stomach pains. Severe, though not fatal, phenol poisoning can damage inner organs, that is, kidneys, liver, spleen, lungs, and heart. In addition, neuropsychiatric disturbances have been described after survival of acute phenol poisoning. Most of the phenol absorbed by the body is excreted in urine as phenol and/or its metabolites. Only smaller quantities are excreted with the feces or exhaled [72].

Acute Toxicity. Signs and symtoms of acute toxicity for animals and humans are similar independent of the route of administration. Absorption is rapid. Phenol can produce symtoms of toxicity within minutes. In animals the LD_{50} was determined for different species and forms of administration. For oral administration the following values are reported: mouse 500 mg/kg; dog 500 mg/kg; rabbit 420 mg/kg [73]. Concentrations of 200 and 20 ppm phenol in aqueous solution were administered to rats. The LD_{50} values determined were 340 and 620 mg/kg respectively [74]. In the summary and evaluation of data on the lethal dose for humans, a value of 140 mg/kg was reported as the minimum lethal oral dose [75].

Subacute to Chronic Toxicity. Repeated administration of phenol leads to neurotoxic, immunotoxic, nephrotoxic, and hepatotoxic effects, as well as changes in hematological parameters [72].

Carcinogenicity. The published reports on the carcinogenic and cocarcinogenic investigations give ambiguous indications [72]. According to EU criteria on carcinogenic potential, there is neither need for classification and labeling nor for further investigations.

Mutagenic Effects and Reproductive Toxicity. According to TRGS 905 (the German technical regulations for handling hazardous substances, 10/2002) phenol is classified as a class 3 mutagen. The discussion whether phenol should be classified as a mutagen or not is still on-going.

Occupational Health. The following limiting values (8 h time-weighed average) were laid down for work protection (2003):

FI, ES, UK	20 mg/m^3
D, B, F, CH	19 mg/m^3
USA (NIOSH/OSHA AGGH):	19 mg/m^3
NL	8 mg/m^3
AU, IR, IT	7.8 mg/m^3
S, DK, NO	4 mg/m^3

Short-term exposure levels:

FI, UK	38 mg/m^3
D, CH	19 mg/m^3
S	8 mg/m^3

Phenol exposure can be determined from the phenol concentration in urine. The permitted BAT value is 300 mg/L.

Classification According to the EU Directives/Ordinance on Hazardous Materials. Phenol is classified with the hazard symbols T (toxic) and C (caustic) and must be labeled with the R phrases R24/25 (toxic in contact with skin and if swallowed) and R 34 (causes burns). Phenol-containing formulations require the following labeling, depending on the concentration c:

$5\,\% < c$	T, R24/25–34
$1 < c < 5\,\%$	Xn, R21/22–36/38

10. References

1. A. E. Hill, W. M. Malisoff, *J. Am. Chem. Soc.* **48** (1926) 918.
2. L. H. Horsley: Azeotropic Data I and II, American Chemical Society, Washington, D.C., 1952, 1962.
3. Verein Deutscher Ingenieure: *VDI-Wärmeatlas*, 8th ed., VDI-Verlag, Düsseldorf 1997.
4. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe*, 2nd ed., Dtsch. Eichverlag, Berlin 1963, 6th supplement, 1.7.1990.
5. E. Brandes, W. Möller: *Sicherheitstechnische Kenngrößen*, Band 1, Wirtschaftsverlag NW, Bremerhaven 2003.
6. C. I. Hilado, S. W. Clark, *Chem. Eng. (N.Y.)* **75** (1972) Sept. 4, 75.
7. J. D. Roberts, M. C. Caserio: *Basic Principles of Organic Chemistry*, W. A. Benjamin, New York, Amsterdam 1964.
8. N. M. Emanuel: *Liquid-Phase Oxidation of Hydrocarbons*, Plenum Press, New York 1967.
9. H. Hock, S. Lang, *Ber. Dtsch. Chem. Ges.* **B77** (1944) 257.
10. *The Industrial Chemist*, May 1960, 215.
11. H. Kropf, *Chem. Ing. Techn.* **B36** (1964) 749.
12. J. B. Fleming, J. R. Lambrix, J. R. Nixon, *Hydrocarbon Processing* (1976) 185.
13. J. J. McKetta, W. A. Cunningham: Phenol in *Encyclopedia of Chemical Processing and Design*, Marcel Dekker, New York.
14. M. Weber, *Umweltpraxis* **9** (2003) 35.
15. G. H. Twigg, *Chem. Ind.*, Jan. 6, 1962, **4**.
16. P. R. Pujado, J. R. Salazar, C. V. Berger, *Hydrocarbon Processing* (1976) 91.
17. M. Weber, *Chem. Eng. Technol.* **B25** (2002) 553.
18. *Ullmann's Encyclopedia of Industrial Chemistry*, 5th ed., **A1**, p. 79.
19. K. Hattori, Y. Tanaka, H. Suzuki, T. Ikawa, H. Kubota, *J. Chem. Eng. Jpn.* **B3** (1970) 72.
20. Distillers Co., GB 610293, 1948.
21. Hercules Powder, US 2547938, 1951.
22. Hercules Powder, US 2548435, 1951.
23. Phenolchemie, DE 1131674, 1960 (H. Sodomann, B. Hauschulz).
24. M. Weber, Proceedings of the DGMK-Conference "The Future Role of Aromatics in Refining and Petrochemistry", Oct. 13 – 15, 1999, Erlangen, Germany.
25. Distillers Co., DE 1114817, 1961.
26. Societa Italiana Resine S.p.A., DE 1443329, 1970 (A. Mategazza, C. Reni).
27. General Electric Company, US 5254751, 1993 (V. M. Zakoshansky).
28. *Hydrocarbon Processing* (1999) no. 3, 126.
29. Allied Corporation, US 4358618, 1982 (S. Sifniades, A. A. Tunick, F. W. Koff).
30. Phenolchemie, DE 944013, 1956 (E. Rindtorff, K. Schmidt, H. Sodomann).
31. Phenolchemie, DE 1112527, 1962 (H. Sodomann, B. Hauschulz, G. Berg).

32. G. Jung, G. Just, *J. prakt. Chem.* **313**, (1971) 377.
33. G. Burtzlaff, U. Felber, H. Hübner, W. Pritzkow, W. Rolle, *J. prakt. Chem.* **28**, (1965) 305
34. F. Kiessling, I. Krafft, K.-K. Moll, K. Peizing, *Chem. Techn.* **B23** (1971) 423.
35. Allied Chemical & Dye Corporation, US 2757209, 1956 (G. G. Joris).
36. UOP, GB 1108233, 1968.
37. Allied, US 4722769, 1988.
38. INEOS Phenol, US 6489518, 2002.
39. INEOS Phenol, DE 19951373, 1999.
40. Sumitomo, EP 0573070, 1993.
41. Mitsui, US 4351967, 1983.
42. General Electric, EP 299239, 1989.
43. Allied Signal, US 5064507, 1991.
44. California Research Corp., US 2762838, 1951.
45. Dow Chemical Corp., US 2727326, 1955.
46. W. W. Kaeding, *Hydrocarbon Processing* **B43** (1964) 173.
47. *Hydrocarbon Processing* **B40** (1961) 280.
48. *Hydrocarbon Processing* **B44** (1965) 255.
49. W. W. Kaeding, R. O. Lindblom, R. G. Temple, *Ind. Eng. Chem.* **B53** (1961) 805.
50. K. Weissermel, H.-J. Arpe: *Industrial Organic Chemistry*. 4th ed., Wiley-VCH, Weinheim 2003.
51. W. L. Faith, D. B. Keyes, R. L. Clark: *Industrial Chemicals*. 3rd ed., Wiley, New York, 1965.
52. M. Wells: *Handbook of Petrochemicals and Processes*. 2nd ed., Ashgate, Vermont 1999.
53. *Hydrocarbon Processing* **B57** (1978) 125
54. The Lummus Company, US 4277630, 1981 (A. P. Gelbein, A. M. Khonsari).
55. H. H. Schobert: *Coal – The Energy Source of the Past and the Future*, American Chemical Society, Washington, D.C., 1987.
56. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung*, Verlag Glückauf, Essen 1951.
57. H.-G. Franck, A. Knop: *Kohleveredelung – Chemie und Technologie*. Springer-Verlag, Berlin, Heidelberg, 1979.
58. "Phenole und Alkylphenole" in *Ullmanns Encyclopädia der technischen Chemie*, 3rd ed., Urban & Schwarzenberg, München, Berlin, 1962.
59. B. Lücke, A. Martin, Proceedings of the DGMK-Conference "The Future Role of Aromatics in Refining and Petrochemistry", Oct. 13 – 15, 1999, Erlangen, Germany.
60. M. Orschel: "Konventionelle und kombinatorische Suche nach neuen Katalysatoren für die selektive Oxidation von Benzol zu Phenol", PhD Thesis, Universität-Gesamthochschule Essen, 2000.
61. A. Reitzmann: "Direktoxidation von Benzol zu Phenol mit Distickstoffmonoxid an modifizierten Zeolithkatalysatoren", PhD Thesis, Universität Erlangen-Nürnberg, 2001.
62. A. K. Uritarte, M. A. Rodkin, M. J. Gross, A. S. Kharitonov, G. I. Panov, 3rd World Congress on Oxidation Catalysis, San Diego, 1997, Elsevier, 1997.
63. *Eur. Chem. News*, March 25 – 31, 2002.
64. R. A. Reimer, C. S. Slaten, M. Seapan, M. W. Lower, P. E. Tomlinson, *Environ. Progr.* **B13** (1994) 143
65. *Eur. Chem. News*, May 17 – 24, 1999.
66. P. A. Witt, M. C. Forbes, *Chem. Eng. Prog.* **B67** (1971) 90.
67. Phenolchemie, DE 2235493, 1974 (B. Hauschulz, J. Mertmann).
68. Enichem S.p.A., US 5811598, 1998 (V. Alessi, S. Astori).
69. J. M. Klobucar, *Chem. Eng.* **2002**, no. 2, 62.
70. Tecnon OrbiChem Newsletter, Phenol Cumene, May 21, 2003.
71. G. Rippen, *Handbuch Umweltchemikalien*, 2. Aufl. 1987, 28. Erg.-Lfg. 03/95.
72. BUA Report 209, Phenol, GDCh Advisory Committe on Existing Chemicals of Environmental Relevance, May 1997
73. Datenbank für wassergefährdende Stoffe, version June 22, 1988.
74. W. B. Deichmann, S. Witherup, *J. Pharmacol. Exp. Ther.* **80** (1944) 233 400.
75. R. M. Bruce, J. Santodonato, M. W. Neal, *Toxicol. Ind. Health* **3** (1987) 535 – 568.

My Profile

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Phenol Derivatives

Standard Article

Helmut Fiege[1], Heinz-Werner Voges[2], Toshikazu Hamamoto[3], Sumio Umemura[4], Tadao Iwata[5], Hisaya Miki[6], Yasuhiro Fujita[7], Hans-Josef Buysch[8], Dorothea Garbe[9], Wilfried Paulus[10]

[1]Bayer AG, Leverkusen, Federal Republic of Germany
[2]Hüls Aktiengesellschaft, Marl, Federal Republic of Germany
[3]Ube Industries, Ltd., Ube, Japan
[4]Ube Industries, Ltd., Ube, Japan
[5]Mitsui Petrochemical Industries, Ltd., Yamaguchi, Japan
[6]Mitsui Petrochemical Industries, Ltd., Yamaguchi, Japan
[7]Mitsui Petrochemical Industries, Ltd., Yamaguchi, Japan
[8]Bayer AG, Leverkusen, Federal Republic of Germany
[9]Haarmann & Reimer GmbH, Holzminden, Federal Republic of Germany
[10]Bayer AG, Uerdingen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a19_313
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Alkylphenols
1.1. Physical Properties
1.2. Chemical Properties
1.3. Occurrence, Formation, Isolation
1.4. Production, General
1.4.1. Alkylation of Phenols
1.4.2. Other General Processes
1.5. Industrially Important Alkylphenols
1.5.1. Polymethylphenols
1.5.2. Ethylphenols
1.5.3. Isopropylphenols
1.5.4. *sec*-Butylphenols
1.5.5. *tert*-Butylphenols
1.5.6. *tert*-Pentylphenols
1.5.7. Higher Alkylphenols
1.5.8. Cycloalkylphenols
1.5.9. Aralkylphenols
1.5.10. Alkenylphenols
1.5.11. Indanols
2. Catechol
2.1. Physical Properties
2.2. Chemical Properties
2.3. Production
2.4. Uses
2.5. Economic Aspects
2.6. Toxicology
3. Trihydroxybenzenes
3.1. Pyrogallol
3.2. Hydroxyhydroquinone
3.3. Phloroglucinol
4. Bisphenols (Bishydroxyarylalkanes)
4.1. Physical Properties
4.2. Chemical Properties
4.3. Production
4.4. Analysis, Testing, Storage
4.5. Uses, Economic Aspects
4.6. Toxicology
5. Hydroxybiphenyls
5.1. Physical Properties
5.2. Chemical Properties
5.3. Production
5.4. Analysis, Quality Specifications, Storage
5.5. Uses
5.6. Toxicology
6. Phenol Ethers
6.1. Properties
6.2. Production
6.3. Representative Phenol Ethers
7. Halogen Derivatives of Phenolic Compounds
7.1. Introduction
7.2. Representative Compounds

[Top of Page]

1. Alkylphenols

Helmut Fiege and Heinz-Werner Voges

In this article alkylphenols are defined as monohydric phenols in which one or more ring hydrogen atoms have been replaced by the same or a different alkyl, alkenyl, cycloalkyl,

cycloalkenyl, or aralkyl group, but which have no other substituents. Indanols are treated as disubstituted alkylphenols. For cresols and xylenols, see ⟶ Cresols and Xylenols.

In the nomenclature, alkylphenols are usually considered as phenol derivatives; the alkyl groups are ordered alphabetically and numbered starting from the hydroxyl group. In industry, products from cresol or xylenol are often described as derivatives of the corresponding cresols or xylenols, e.g., 6-*tert*-butyl-2,4-dimethylphenol is called 6-*tert*-butyl-2,4-xylenol. A very complicated system which is, e.g., found in handbooks considers alkylphenols as benzene derivatives. Here, the substituents are ordered alphabetically, and numbered such that the sum of all numbers gives the smallest possible value. 6-*tert*-Butyl-2,4-dimethylphenol is thus referred to as 1-*tert*-butyl-2-hydroxy-3,5-dimethylbenzene. According to this system cresol derivatives are also occasionally described as toluene derivatives, e.g., 2,6-di-*tert*-butyl-4-methylphenol is called 3,5-di-*tert*-butyl-4-hydroxytoluene (BHT). Some alkylphenols have common names.

1.1. Physical Properties

Pure alkylphenols are colorless liquids or form low-melting crystals, which frequently darken with time. They show a tendency to supercooling. Within groups of substitutional isomers the para isomers have the highest melting points. In some cases two melting points are observed (occurrence of a metastable phase).

The boiling points are > 200 °C at normal pressure and increase with molecular mass. Whereas the boiling point difference between meta and para isomers is very small, the para compounds usually boil 10 – 20 °C higher than the isomeric ortho compounds. The sharper rise in boiling point on para substitution remains if more alkyl groups are present. A particularly sharp rise in boiling point results if a 3- or 5-alkylphenol is *tert*-butylated in the 4-position [6]. This effect is used industrially for the separation of alkylphenols which cannot be separated by distillation.

The relatively small rise in boiling point on ortho substitution is caused by steric hindrance of hydrogen bond formation. This steric hindrance is reflected in many other physicochemical properties of alkylphenols (e.g., spectra, dissociation energies, acidity of the O – H bond, association, dipole moments [7], [8]).

Physical properties of individual alkylphenols are listed in the corresponding sections. Further data for ethylphenols can be found in [9], for isopropylphenols in [10-13], and for *tert*-butyl- and other phenols in [14-16]. Vapor pressure curves for many alkylphenols are given in [17] and thermochemical data under normal conditions in [18]. The relationship between vapor pressure and association of alkylphenols is given in [19].

Alkylphenols are readily soluble in polar organic solvents (lower alcohols, ketones, esters, carboxylic acids); their solubility in hydrocarbons increases with the number and size of the alkyl substituents on the ring. The opposite trend occurs with regard to solubility in water, higher alkylphenols are virtually insoluble and lower alkylphenols are soluble in water only at increased temperatures to a noticeable extent. Their water-solubility shows a temperature-dependence similar to that of phenol, but, with considerably higher critical solution temperatures (i.e., the temperature where unlimited miscibility with water occurs) [20], [21]. Within a group of substitutional isomers the water-solubility decreases with increasing proximity of the alkyl substituents to the hydroxyl group. The presence of other water-soluble organic compounds increases the solubility of alkylphenols and lowers the critical solution temperature [22].

Water can have considerable solubility in alkylphenols [20], [21] and hydrates may be formed, e.g., 6-*tert*-butyl-3-methylphenol, $C_{11}H_{16}O \cdot 1/4\ H_2O$, *mp* 37 °C [15].

1.2. Chemical Properties

The chemical behavior of alkylphenols deviates from the reactions typical of phenol with increasing number, size, and degree of branching of the alkyl groups. This change in properties is particularly pronounced in 2,6-di-*tert*-alkylphenols [23]. The alkyl groups themselves can also show typical reactions.

Acidity. Alkylphenols are less acidic than phenol. The acidity decreases in the direction of meta > para > ortho substitution. The decrease in acidity due to the presence of ortho alkyl groups becomes more pronounced the more these groups are branched in their α-positions. This is mainly because of steric hindrance of solvation in the phenolate anion formed [7].

These effects are reflected in the solubility of alkylphenols in alkali: 4-*tert*-butylphenol, 2-isopropylphenol, 2,6-diethylphenol, 4-*tert*-butyl-2,5-dimethylphenol, and 4-*tert*-butyl-2,6-dimethylphenol are soluble in dilute aqueous sodium hydroxide solution, 2-*tert*-butyl-4-methylphenol, 6-*tert*-butyl-2,4-dimethylphenol, 4,6-di-*tert*-butyl-2-methylphenol, 2,6-di-*sec*-butyl-4-methylphenol, and 2,4,6-triisopropylphenol are insoluble in alkali. The latter are, however, soluble in the so-called Claisen solution (350 g potassium hydroxide in 250 g water made up to 1 L with methanol). In contrast to these partially sterically hindered alkylphenols, phenols with two bulky alkyl groups (e.g., *tert*-butyl, *tert*-pentyl) in the ortho and one alkyl group in the para position are also insoluble in the Claisen solution [24]. Examples of such sterically hindered (shielded) phenols are 2,6-di-*tert*-butyl-4-methylphenol, 2,6-di-*tert*-pentyl-4-methylphenol, and 2,4,6-tri-*tert*-butylphenol. The phenolates of these compounds can only be produced under special conditions [25], e.g., by treatment with solid potassium hydroxide and azeotropic removal of the water by an inert organic entrainer [26]. o-*tert*-Butylphenols without para alkyl groups behave differently; for example 2-*tert*-butylphenol and 6-*tert*-butyl-3-methylphenol are soluble in aqueous sodium hydroxide, and 2,6-di-*tert*-butylphenol is soluble to a certain extent in the Claisen solution. This is presumably because alkali salts of the corresponding tautomeric cyclohexadienones can be formed relatively easily. The insolubility of pentamethyl-, 4-*tert*-octyl, 4-nonyl,- or 4-dodecylphenol in cold aqueous alkalis is apparently due to the increase in the hydrophobic nature of the molecule because of the increasing number or length of the alkyl groups. These phenols are readily soluble in the Claisen solution. Alkylphenols which are insoluble in aqueous alkali but readily soluble in the Claisen solution are called cryptophenols.

The different solubility of the alkylphenols in alkalis can be used for their separation. As a result of the low acidity of the alkylphenols, they can even be liberated from their alkali phenolate solutions by carbon dioxide (p*K* 6.4).

Ether Formation. The hydroxyl group can be etherified, for example, with alkyl halides, halogenated carboxylic acids, or dialkyl sulfates in aqueous alkaline media. Bulky ortho alkyl groups hinder the reaction. In these cases the etherification can be forced by, for example, treating the anhydrous alkali phenolate with dimethyl sulfate in a solvent. If phase-transfer catalysts are used, it is possible to work in the presence of water. An example is the etherification of 2,4,6-tri-*tert*-butylphenol with dimethyl sulfate in aqueous sodium hydroxide – dichloromethane in 93 % yield in the presence of phase-transfer catalysts [27]. Alkylene oxides add to the hydroxyl group in the presence of basic catalysts to form alkylphenolmono- and polyglycol ethers (alkoxylation). The hydroxyl group can be etherified with alkenes in the presence of Friedel – Crafts catalysts under mild conditions.

Esterification. The hydroxyl group can generally be esterified successfully with acid anhydrides or acid chlorides. In the case of sterically hindered phenols, conditions have to be modified as with etherification [24]. With the exception of 2,6-di-*tert*-alkylphenols the hydroxyl groups of the alkylphenols react with isocyanates to form urethanes, which can be used for identification of the alkylphenols [28].

Substitution of the Hydroxyl Group. At 250 °C and in the presence of palladium – charcoal catalysts the OH group can be exchanged to yield the corresponding dialkylaniline [29]. This reaction occurs even with dialkylphenols such as 6-*tert*-butyl-2-methylphenol.

Hydrogenation. The hydrogenation of alkylphenols with hydrogen in the presence of Raney nickel or noble-metal catalysts leads to alkylcyclohexanols and/or alkylcyclohexanones. The latter are formed particularly in the case of sterically hindered alkylphenols [30], [31]. At high temperatures (250 – 350 °C) and in the presence of transition-metal – aluminum oxide catalysts, hydrogenolytic elimination of the alkyl groups and the hydroxyl group can also occur [32].

Oxidation. The coloration of alkylphenols in the presence of air indicates that they are sensitive to oxidation. Oxidation generally begins with the formation of phenoxy radicals. These radicals are relatively stable in the case of sterically hindered alkylphenols and are the reason why these phenols are particularly strong autoxidation inhibitors (⟶ Antioxidants). In the further course of the oxidation, many different compounds, some intensely colored, can be formed: hydroquinones, quinols, quinones, ketones, aldehydes, diphenoquinones, di-, tri-, and polyphenols, di-, tri-, and polyethers, and other C – C and/or C – O-linked products. Depending on the reaction conditions, oxidizing agent, catalyst, and nature and position of the alkyl group(s), certain products are formed preferentially [33-35]. Thus the formation of diphenoquinones is observed particularly in the case of 2,6-dialkylphenols with bulky alkyl groups. The 4-methyl group of 2,6-dialkyl-4-methylphenols can be often oxidized selectively to an aldehyde group (e.g., with pentyl nitrite, *tert*-butyl hydroperoxide – bromine, or oxygen) [33], [36], [37]. Strong oxidizing agents degrade the phenyl ring, e.g., the oxidation of alkylated phenols with potassium permanganate gives fatty acids.

Electrophilic Aromatic Substitution. Alkylphenols readily undergo electrophilic substitution. Substitution generally occurs in the positions ortho or para to the hydroxyl group, provided that they are not already occupied by other substituents. In particular when several or bulky alkyl groups are present, substitution (displacement) of alkyl groups by the electrophile can occur in addition to the normal substitution of hydrogen, or the reaction may stop at the stage of the cyclohexadienone, formed initially by addition of the electrophile. These types of anomalies are observed, for example, in nitrations [38], chlorinations [39], and in the reaction with trichloromethane – alkali (Reimer – Tiemann aldehyde synthesis) [40]. The nitration occurs even with dilute nitric acid and requires inert dilution agents and the careful maintenance of low temperatures to avoid undesired secondary reactions or a dangerous reaction process [38]. Nitrosation [41], chlorination [39], alkylation [42], and sulfonation, for example, take place with similar ease. The nitroso group preferably enters the ring para to the OH group. The nitrosophenols can also be present in the tautomeric quinone oxime form.

On heating alkylphenols, particularly in the presence of Friedel – Crafts catalysts, isomerization, transalkylation, and dealkylation can occur [43]. The tendency towards these reactions increases sharply in the order primary < secondary < tertiary alkyl groups. The isomerizations proceed until an equilibrium state is reached in which the thermodynamically more stable 3- and 3,5- (i.e., meta) isomers prevail. Groups in the para and ortho positions are cleaved particularly easily, those in the meta position less easily. Dealkylation can also occur during distillation of alkylphenols, if acids or potential acid-forming substances in the starting material are not completely removed or rendered ineffective (by addition of alkali). 1,1,3,3-Tetramethyl-5-indanol—which can be considered as ring-closed 3,4-di-*tert*-butylphenol and generally behaves as a typical phenol—however, shows remarkable stability towards dealkylation even in the presence of acids. The thermal stability of other indanols is also much higher than that of alkylphenols.

Heating alkali phenolates of alkylphenols with free ortho or para positions in the presence of carbon dioxide in polar, aprotic solvents such as dimethylformamide leads to hydroxyalkylbenzoic acids (Kolbe – Schmidt synthesis) [44]. Aldehyde groups can be very selectively introduced into the position para to the hydroxy group with zinc cyanide – hydrochloric acid. If this position is occupied, introduction into the ortho position occurs (Gattermann – Adams aldehyde synthesis) [45]. In the presence of alkali, formaldehyde reacts at room temperature with free ortho and para positions of alkylphenols to form hydroxymethyl-alkylphenols. Under acidic conditions or at higher temperatures these hydroxymethyl compounds condense with alkylphenols to form bis(hydroxyalkylphenyl)-methanes, and undergo autocondensation to yield high molecular mass alkylphenol – formaldehyde resins [46], [47]. The condensation of alkylphenols with formaldehyde in the presence of primary or secondary amines leads to Mannich bases [47]. The acidic condensation of alkylphenols with other aldehydes or ketones gives bis(hydroxyalkylphenyl)alkanes (see Section Production). Thiobisphenols are readily obtained with sulfur dichloride [48] and with disulfur dichloride, dithiobisphenols [49].

Alkenylphenols show a tendency towards dimerization and polymerization.

1.3. Occurrence, Formation, Isolation

Some alkylphenols such as 4-ethylphenol, 4-isopropylphenol (*p*-cumenol), 6-isopropyl-3--methylphenol (thymol), 5-isopropyl-2-methylphenol (carvacrol), 2,6-diisopropyl-3-methylphenol, 2,4-diisopropyl-5-methylphenol, 2,5-diisopropyl-4-methylphenol, and 4-allylphenol (chavicol) are components of natural essential oils, some of which have been used for centuries for the production of perfumes (⟶ Flavors and Fragrances). 2,3,4-Trimethylphenol, 5-ethyl-2-methylphenol, and 2,3-xylenol were detected in the defensive secretions of opilionides [50], and 2,6-di-*tert*-butyl-4-methylphenol is contained in Australian brown coal [51]. 3-(Pentadecen-8-yl)phenol, 3-(pentadecadien-8,11-yl)phenol, and 3-(pentadecatrien-8,11,14-yl)phenol have been identified as the main components of cardanol extracted from cashew nutshell oil on an industrial scale [52]. 3-Pentadecylphenol (hydrocardanol) is obtained from it by hydrogenation [53].

Alkylphenols are obtained in great number in the *purely thermal cracking* (low-temperature carbonization, coking, see ⟶ Coal Pyrolysis), in the *oxidizing thermal cracking* (gasification, ⟶ Gas Production), and the *hydrogenating thermal cracking* (⟶ Coal Liquefaction, ⟶ Coal Pyrolysis) of natural products such as bituminous coal, brown coal, peat, wood, lignin, and in the *coking thermal cracking* of crude oil fractions (⟶ Cresols and Xylenols – Formation and Isolation). Thus many different alkylphenols were detected in the various coal tars, some of them with several methyl, ethyl, *n*-propyl, isopropyl, *n*-butyl, isobutyl, *sec*-butyl, allyl, 1-cyclopentenyl, cyclohexyl, and/or indanoyl groups and various indanols and indenols. The unsaturated compounds are found in small quantities, mainly in low-temperature tars. Without taking the cresols and xylenols into account, about a hundred different monohydric phenols in total have been found in the various coal tars [54]. For the history of tar phenols see [55], [56].

The product distribution of the various cracking processes differs in the nature, but more in the ratio and particularly in the quantity in which the alkylphenols are produced, which depends on the starting material and the operating conditions. The mass ratio range of alkylphenols isolated from bituminous coal tars is given in Table 1.

Table 1. Alkylphenol content in bituminous coal tars

Alkylphenols	Low-temperature carbonization, wt %	Coking, wt %
Ethylphenols	42	47
Trimethylphenols	24	13
Ethylmethylphenols	13	15
Propylphenols	10	15
Methylpropylphenols	0.5	5
Tetramethylphenols	1	0.5
Diethylphenols		<0.5
Butylphenols		<0.5
Indanols	9	4

More details concerning product composition in bituminous coal coking are given in [57], and in various bituminous coal low-temperature carbonization processes in [58], [59].

Brown coal low-temperature tar contains a relatively high proportion of ethylphenols [59], [60]. Alkylphenols isolated from crude oil have a higher content of phenols with longer side chains, e.g., propyl- and butylphenols [61], [62]. In the hydrogenation of coal many indanols are formed [63], [64].

The strong influence of the cracking process on the total amount of monohydric phenols produced and on the proportion of alkylphenols among them is shown in Table 2.

Table 2. Phenol yields in the hydrogenation, low-temperature carbonization, and coking of bituminous coal, and in the cracking of gas oils

Reforming process	kg phenols per t H_2O-free raw material	
	Total *	Alkylphenols **
Hydrogenation	75	25
Low-temperature carbonization	15	3
Coking	0.7	0.04
Cracking	1	0.04

* Phenol + cresols + xylenols + alkylphenols (without phenol ethers etc.).
** Excluding cresols and xylenols.

In the distillative workup of the crude phenols obtained in the different cracking processes (⟶ Cresols and Xylenols – Isolation from Coal Tars), 2-ethylphenol (*bp* 204.3 °C) is distilled over with the *m*-/*p*-cresol fraction. The residue, which has a very complex composition, is usually worked up by batch distillation. Since the boiling points of the alkylphenols lie very close together and overlap to some extent with those of the xylenols and other compounds which may be present (e.g., phenol ethers), the production of certain alkylphenols in pure form is not possible by distillation. The products in the distillation range of ca. 205 – 285 °C are therefore separated only into particular fractions which depend on the application purposes and are characterized by their boiling ranges.

The distillate boiling between ca. 205 – 225 °C is generally called the xylenol fraction. If it originates from high-temperature bituminous coal tar, it can contain up to 30 % alkylphenols, predominantly 3- and 4-ethylphenol and methylethylphenols [57]. In xylenol fractions from East German brown coal low-temperature tars the xylenol content can be below 30 % and the 3-/4-ethylphenol content 40 % [65]. A 216 –218 °C xylenol fraction obtained from brown coal low-temperature tar phenols for the varnish industry contains predominantly 3- and 4-ethylphenol.

For the separation of 3- and 4-*ethylphenol* the same processes as for the separation of *m*- and *p*-cresol must be used (⟶ Cresols and Xylenols). Similar to *p*-cresol, 4-ethylphenol can be separated by esterification with oxalic acid [66]. The aforementioned 216 – 218 °C brown coal xylenol fraction, for example, is a suitable starting material for isolation of 3- and 4-ethylphenol.

2-*Ethylphenol* can be separated from the fraction boiling between 203 and 207 °C (in which it is enriched) as the quinaldine addition compound [67]. Other alkylphenols which can be enriched in narrow-boiling fractions and isolated pure by special methods, include 2,3,5-trimethylphenol (isopseudocumenol) [68], 2,3,6-trimethylphenol [69], 2,4,6-trimethylphenol (mesitol), 5-ethyl-3-methylphenol, 4-indanol, and 5-indanol [70].

Most of the alkylphenols with a boiling point > 230 °C are used as a mixture. The composition of an alkylphenol fraction boiling between 230 and 285 °C obtained from vertical retort bituminous coal tar is given in [71].

1.4. Production, General

1.4.1. *Alkylation of Phenols*

Alkylphenols are produced industrially mainly by catalytic alkylation of phenol, cresols, or xylenols with olefins which are readily accessible petrochemicals. Alcohols or halogenated hydrocarbons are only used as the alkylating agent if they are more readily available (e.g., cyclohexanol) or if the corresponding olefin does not exist (e.g., methanol, benzyl chloride). Other production processes for alkylphenols are less important than alkylation.

Ortho/Para Alkylation and Meta Isomerization in the Presence of Friedel – Crafts Catalysts [72-77]. The following Friedel – Crafts catalysts are mainly used in industry: sulfuric acid, phosphoric acid, *p*-toluenesulfonic acid, strongly acidic cation-exchange resins (e.g., Amberlite JR-112, Lewatit S 100), acid-activated silicates of the Montmorillonite type (bleaching clays), synthetic aluminum silicates of the crack catalyst type, boron trifluoride, aluminum chloride, iron(III) chloride, and zinc chloride.

The catalysts react with the alkylating agent to form an alkyl cation which is as branched (stable) as possible. The alkyl cation enters the nucleus preferentially in the positions ortho and para to the OH group according to the rules of electrophilic substitution. In addition, particularly under mild conditions, etherification of the OH group occurs. For the mechanism see [72-74].

The ratio of isomers formed initially is kinetically controlled and determined (1) by the position and inductive effect of the methyl (alkyl) group(s) already on the ring; (2) the reactivity of the alkyl cation; (3) the steric situation on the ring; and (4) the incoming alkyl group. Since the monoalkylation products are generally more reactive than the starting phenol, they

are further alkylated, provided that there is no steric hindrance. The ratio of mono- to di- and trialkylphenols is proportional to the ratio of phenol to alkylating agent.

During the further course of the reaction, rearrangements (isomerizations, transalkylations, disproportionations) occur and the product composition shifts in the direction of a thermodynamically more stable equilibrium state, in which phenols with meta alkyl groups finally prevail. During this process phenol ethers first rearrange to alkylphenols, the amount of di- and trialkylphenols is lowered, and the proportion of o-alkylphenols is shifted in favor of the p-alkylphenols, provided that the para position is free and is not sterically hindered.

Possible side reactions are (1) dealkylations; (2) di-and oligomerizations of the olefin used as alkylating agent or formed by dealkylation; (3) isomerization and cleavage in the alkyl groups (particularly in the case of larger and branched alkyl groups or olefins); (4) readdition of the olefins formed by cleavage, di- or oligomerization, (5) dehydrogenation of the alkyl to alkenyl groups and their reaction with olefins to form indanols; and (6) reactions of the catalyst with reaction components, e.g., the formation of phenolsulfonic acids and alkylsulfonates if sulfuric acid is used as catalyst. Ortho and para alkyl groups are much more prone to rearrangement and cleavage reaction than meta alkyl groups. The tendency towards rearrangement and cleavage reactions of ortho and para alkyl groups increases in the order of primary < secondary < tertiary alkyl groups. Methyl groups can only be made to migrate with great difficulty.

How fast a reaction mixture reaches thermodynamic equilibrium and the amount of side reactions that occur, depends on the nature of the phenol and that of the alkylating agent and on the severity of the reaction conditions, in particular on the activity (type) and quantity of the catalyst, on the temperature, the reaction time, and occasionally also on the water content of the reaction mixture. The reaction conditions must be carefully adjusted to achieve optimal product compositions.

High yields of a particular alkylphenol are often possible by making use of the rearrangement equilibria, so that the undesired ortho-/para-mono-, di-, and/or trialkylphenols formed as byproducts are separated from the desired product and recycled to the alkylation reaction. This is, however, not the case for 2-alkylphenols and 2,6-dialkylphenols: these compounds can only be obtained with the coproduction of considerable quantities of the corresponding phenols alkylated in the 4-position even under carefully controlled alkylation conditions and skillfully exploited rearrangement conditions. This is because, as previously indicated, ortho and para alkylations always occur simultaneously in the presence of Friedel – Crafts catalysts and the ortho isomer readily rearranges to the para isomer which is often thermodynamically more stable. This behavior has been intensively investigated in the *tert*-butylation of phenol with isobutene [78-83]. Good yields of 2-alkyl- or 2,6-dialkylphenols are possible if the para position is blocked by an alkyl group which will not migrate under the reaction conditions, e.g., a methyl group.

Ortho-Alkylation. In the presence of *aluminum phenolates*, phenols can be selectively alkylated with olefins in the ortho position. The catalyst is produced in the alkylation reactor before the addition of the olefin by heating the dry phenol with 1 – 2 wt % aluminum sand, powder, or turnings to 120 – 180 °C (with hydrogen formation). Depending on the olefin – phenol ratio, 2-alkyl- or 2,6-dialkyl-phenols are preferentially formed in the alkylation; at very high ratios 2,4,6-trialkylphenols are also formed. Currently 2,6-dialkylphenols, in particular, are produced using this process, which was developed independently by both Bayer [84] and Ethyl Corp. [85] in 1954. An overview of the reaction is given in the publications [86-89].

It is assumed that ortho selectivity is caused by the formation of a strongly acidic complex (**2**) from the aluminum phenolate (**1**) and the olefin, which directs the electrophilic attack of the olefin to the sterically more favorable position ortho to the OH group. The reaction ends with the reversal of the aluminum phenolate formation by proton exchange with excess phenol [90]:

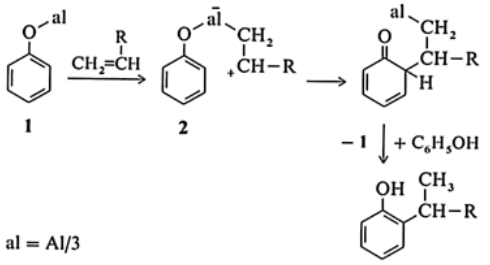

Apart from the special ortho-directing effect and a somewhat lower activity of the aluminum phenolate, which can be compensated for by using higher temperatures, the process is very similar to the alkylation in the presence of Friedel – Crafts catalysts. Thus the reactivity of the olefins decreases in the order $R_2C = CH_2$ > R–CH = CH_2 ≈ R–CH = CH–R > CH_2= CH_2, and the alkyl group with maximum possible branching is formed. The ortho alkylation is also kinetically controlled and at the beginning of the reaction is accompanied by reversible ether formation which increases with increasing temperature. At too high a temperature and with longer reaction times dealkylation and the formation of para isomers also occur, especially if the alkylation has taken place more readily. Therefore, if a high yield of the ortho isomer is desired, it is necessary to work at the lowest possible temperature (which must nevertheless be sufficiently high for rapid ether cleavage) and at a high olefin concentration, and to stop the reaction at the optimum point in time. If lower olefins are used as alkylating agents, it is necessary to work under pressure in order to achieve a high olefin concentration [86], [87]. To avoid dealkylation and isomerization in the subsequent distillative workup, the aluminum phenolate catalyst is deactivated by, for example, addition of water or sodium hydroxide to the reaction mixture after the end of the reaction [91]. For ethyl- and *sec*-alkylphenols it may be possible to do without the catalyst removal, if distillation is carried out very carefully.

The ortho-selective effect of the aluminum phenolate is also seen in transalkylations. Thus under kinetically controlled conditions 2,4-di-*tert*-butylphenol and phenol react to form predominantly 2-*tert*-butylphenol [92]. In the case of dealkylation in the presence of aluminum phenolate, ortho-*tert*-alkyl groups are eliminated more rapidly than those in the para position [87], [93].

The aluminum phenolate catalyst can be obtained by treating phenol with aluminum or with reactive aluminum compounds such as aluminum alkoxides, aluminum trialkyls [94], dialkylaluminum chloride [95], or aluminum chloride [96]. To achieve smooth dissolution of aluminum, it can be activated by addition of mercury(II) chloride [86]. If toluenesulfonic acid is added during the aluminum phenolate, a hydrolysis-resistant mixed catalyst of similar ortho-selectivity is obtained [97]. Aluminum salts of the ortho phenolsulfonic acids are also suitable catalysts. Since they are not as readily soluble in the reaction mixture as the aluminum phenolates, they can be filtered off after the reaction has finished and reused [98]. Also compounds such as aluminum thiophenolate [99], aluminum diphenyldithiophosphate [100], polyphenoxy aluminum oxanes [101], and aluminum phenolate bonded to phenol – formaldehyde resin [102] have been proposed as ortho-alkylation catalysts.

In addition to aluminum phenolate which is used exclusively in industry, the phenolates of, e.g., magnesium, zinc, iron [84], [85], zirconium, hafnium, niobium, tantalum [103], gallium [104], and boron (aryl borates) [105], also have a pronounced ortho-selective effect. Titanium only shows this effect as the catecholate [106], or as the titanium phenolate – arylsulfonate mixed salt [107]. Zirconium, hafnium, niobium, and tantalum phenolates split off the 6-*tert*-butyl group from 3-alkyl-4,6-di-*tert*-butylphenols very selectively, whereas titanium phenolates are unsuitable for this partial dealkylation [108], [109].

γ-Aluminum oxide [110] and other modifications of aluminum oxide (apart from α-Al_2O_3) [111] also direct the alkyl group very selectively into the ortho position, if olefins are reacted with phenols in the liquid phase at 200 – 350 °C and 2 – 20 MPa. In industry these types of aluminum oxide are used particularly for the production of ortho-monoalkylphenols, because they lead to a clearly higher ortho-mono/ortho-dialkylphenol ratio than aluminum phenolate if the same olefin/phenol ratio is used (e.g., 1 : 1). Moreover, they have the advantage over aluminum phenolate — unless it is fixed in a resin [102] — that they can be readily separated after the reaction by, for example, filtration, and reused or used in a fixed-bed form in continuous processes. For the effect of the double bond position in the olefin on the yield of ortho-alkylphenols on γ-Al_2O_3, see [112].

High selectivities were observed, for example, with oxides of the elements of groups 15 and 17 of the Periodic Table [113], and with high-purity silicon dioxide [114]. The activity of the metal compounds can be increased by applying them to a carrier such as Al_2O_3 or silicates [113], [115].

Relatively high ortho selectivities are also observed if olefins are heated with phenols *in the absence of catalysts* in the liquid phase to 300 – 400 °C under pressure [116]. This so-called *thermal alkylation* has been known for a long time [117], but has so far remained unimportant industrially. The conversions are moderate and the yields based on olefin are poor. Long-chain linear α-olefins become bonded to the phenol preferentially at their C-2 atoms to give ortho-monoalkylphenols. Unlike the situation with Friedel – Crafts catalysis, bonding at C-1 has been observed, but hardly any at the middle C-atoms of the olefin chain, and no para substitution [118]. It is assumed that the thermal ortho-alkylation takes place via a 6-membered, cyclic transition state, into which the phenolic OH group is incorporated [116], [118].

Alkylation Rules. The position that an alkyl group can occupy in the phenyl ring depends on the size of the entering group and on the alkyl group already present. For cresols and xylenols the following rules are generally valid:

1. *Secondary alkyl groups* such as isopropyl, *sec*-butyl, cyclopentyl, and cyclohexyl can enter the ring between the hydroxyl and a meta-methyl group, but not between two methyl groups meta to each other. For example, propylation of 3,5-xylenol can yield 2,6-diisopropyl-3,5-dimethylphenol, but not 4-isopropyl-3,5-dimethylphenol.
2. *Tertiary alkyl groups* such as *tert*-butyl, *tert*-pentyl, and *tert*-octyl are unable to enter the ring between the hydroxyl and a methyl group meta to it. This means that, for example, 3,5-xylenol cannot be *tert*-butylated.
3. Unlike *sec*-alkyl groups, *tert*-alkyl groups do not migrate to positions next to a para methyl group. 4-Methyl-3-*tert*-butylphenol therefore cannot be synthesized by isomerization of 4-methyl-2-*tert*-butylphenol in the presence of alkylation catalysts.

For the polyalkylation of phenols with various alkyl groups, the most sensitive (tertiary) group must be introduced last. For example, for the production of 4-*tert*-butyl-2-cyclopentylphenol the *sec*-cyclopentyl group is first introduced (with aluminum phenolate) and then the *tert*-butyl group (with a Friedel – Crafts catalyst).

General Information on Alkylation of Phenol with Olefins. Alkylation is carried out predominantly in the liquid phase without a solvent. In the presence of Friedel – Crafts

catalysts, normal or slightly increased pressure and a temperature range of ca. 30 – 150 °C are generally used. Higher pressures and/or temperatures can be necessary for the meta isomerization or for alkylations in the presence of ortho-selective catalysts (see below). The plant is generally inerted with nitrogen before alkylation.

To be able to adapt to the different starting materials and changing market demands with regard to the nature and quality of the alkylphenol, most alkylphenols are produced batchwise in multipurpose plants. For compounds produced in large quantities such as thymol, 2-*sec*-butylphenol, 4-*tert*-butylphenol, 2,6-di-*tert*-butylphenol, 2,6-di-*tert*-butyl-4-methylphenol, octylphenol, and nonylphenol there is a trend towards specialized continuous or semibatch plants.

The choice of catalyst depends, e.g., on the reactivity of the phenol and of the alkylating agent, the desired product composition (isomer distribution, byproducts), the cost, or the available possibilities with regard to the corrosion resistance of the plant, catalyst removal from the reaction mixture, disposal of the catalyst and/or its decomposition products, workup of the product mixture, the quantity of the product, and its value.

Protonic acids such as H_2SO_4 are readily available, cheap, and so active that small quantities (1 – 2 wt %) are sufficient to obtain high space – time yields at relatively low temperatures. However, they are corrosive and must be removed or rendered ineffective in order to obtain a stable, distillable product. This requires careful scrubbing of the alkylphenol with alkali hydroxides and water, thus producing wastewater which must subsequently be treated. A similar situation exists if metal halides and metal phenolates are used.

Solid inorganic catalysts (acid-activated bleaching clays, oxides of aluminum and silicon etc.) must be added in finely divided form at higher temperatures and in larger quantities (2 –5 wt %, for isomerizations even up to 10 wt %) to compensate for their lower activity. They are, however, less corrosive than protonic acids and have the advantage in process technology that they can be separated simply by filtration. However, filtration must be carried out carefully because catalyst particles which have passed through the filter can lead to dealkylation at high distillation temperatures. The disposal of the depleted catalyst loaded with phenol can be problematic.

Cation exchangers [76] are reusable (long service life), highly selective for monoalkylations, noncorrosive, can be readily separated from the alkylated phenol, and do not produce any wastewater. These advantages are best realized if they are used in continuous alkylations as bead polymers arranged in a fixed bed. When used in suspended form they can lead to formation of abrasion particles, which must be completely separated or neutralized by addition of alkali to the alkylated phenol, so that dealkylation does not occur in the subsequent distillations. The upper operating temperature of the cation exchanger is ca. 140 °C; at this temperature, however, clear loss of activity must be taken into account in continuous operation. Powder cation exchangers have a higher activity and are more effective for dialkylation than the bead exchangers [74].

After separation of the catalyst the reaction mixture is worked up by fractional distillation in vacuum. Unreacted starting material, undesired ortho and para isomers, and high-boilers are recycled to the reaction as far as possible. Isolation of a particular isomer is often laborious and frequently unnecessary: for many industrial applications alkylphenols are only separated from starting material and high-boilers by distillation, and are used as a mixture of isomers. Occasionally, the separation of high-boiling substances is not required, and stripping of the alkylation product is sufficient as the workup step. In special cases even the catalyst separation and distillation are not performed and the crude alkylation product is led directly into the subsequent reaction [119].

1.4.2. *Other General Processes*

Alkylphenols which are not readily accessible or totally inaccessible by alkylation, e.g., those with long primary alkyl groups (≥ 3), can be obtained as follows:

Reduction of acylphenols (phenolketones) which are obtained by Friedel – Crafts acylation of phenols or by the Fries rearrangement of phenol esters [120], [121]:

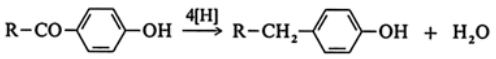

Alkali fusion of alkylbenzenesulfonates, which are obtained by sulfonation of alkylbenzenes [122] (⟶ Cresols and Xylenols – Alkali Fusion of Toluenesulfonates):

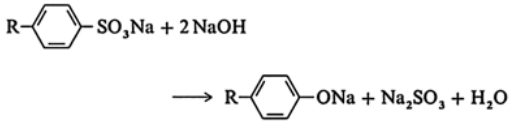

Cleavage of bis(4-hydroxyphenyl)alkanes, either alkali-catalyzed [123], [124] with hydrogen in the presence of a hydrogenation catalyst [125], [126], or with a hydrogen donor such as cyclohexanol:

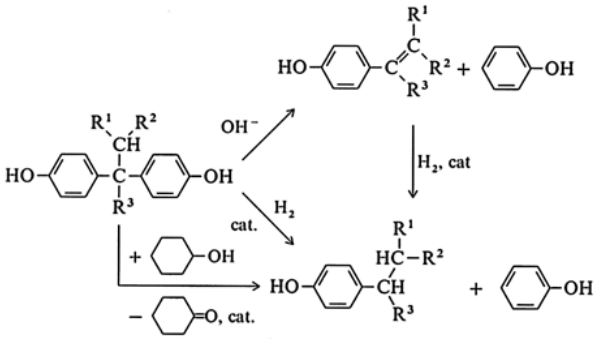

Claisen rearrangement of allyl aryl ethers into allylphenols [127], if necessary with subsequent rearrangement into vinylphenols [128] or hydrogenation to alkylphenols:

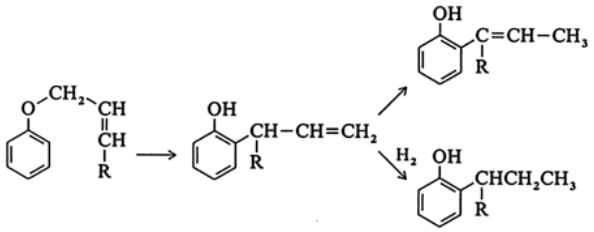

The Claisen rearrangement is particularly suitable for the introduction of alkyl groups into the position ortho to an OH group, the other methods are preferentially used for the production of para-alkylphenols.

Other possibilities can be found in *Houben – Weyl* (see general references). Alternatives to phenol alkylation are also discussed in the following section.

1.5. Industrially Important Alkylphenols

1.5.1. *Polymethylphenols*

2,3,6-Trimethylphenol, 2,4,6-trimethylphenol, and other polymethylphenols are formed in small quantities as byproducts in the production of cresol and particularly xylenol by methylation of phenol (⟶ Cresols and Xylenols – Methylation of Phenol, ⟶ Cresols and Xylenols).

2,3,6-Trimethylphenol [2416-94-6] (physical properties see Table 3) is produced selectively by gas phase methylation of *m*-cresol with methanol at 300 – 460 °C under normal pressure on ortho-selective metal oxide catalysts, as used for the selective methylation of phenol (⟶ Cresols and Xylenols – Methylation of Phenol). The reaction occurs in multitube reactors with a fixed catalyst. The reaction temperature to be maintained depends on the catalyst used in each case. Iron oxide catalysts modified with oxides of other metals (e.g., Zn [129], Cr and Sn[130], or Mg and Si [131]) are particularly suitable for the process. At temperatures of ca. 350 °C and a liquid hourly space velocity (LHSV) of ca. 1 h_{-1}, an initial mixture of *m*-cresol, methanol, and steam in a molar ratio of 1 : 6 : 1 gives 2,3,6-trimethylphenol in yields of 90 – 95 % (relative to *m*-cresol) at virtually complete *m*-cresol conversion. Small quantities of 2,5-xylenol and 2,4,6-trimethylphenol are the main byproducts. About half the methanol decomposes to H_2, CH_4, and CO_2. The reaction mixture is worked up by distillation. For the workup by recrystallization from aliphatic hydrocarbons, see [132]. The commercial product is ≥ 99 % pure.

Table 3. Physical properties of some polymethylphenols and ethylphenols

Name (common name)	Formula	CAS registry number	M_r	*mp*, °C	*bp* at 101.3 kPa, °C	*bp* at 2.66 kPa, °C	ΔH_V at *bp* 101.3 kPa, kJ/mol	d_4, g/cm^3 (°C)	n_d (°C)	p*K* at 20 °C
Polymethylphenols										
2,3,4-Trimethylphenol	$C_9H_{12}O$	[526-85-2]	136.19	81	235 – 237					
2,3,5-Trimethylphenol (isopseudocumenol)	$C_9H_{12}O$	[697-82-5]	136.19	93.7	235.3	125.6	50.0			10.72
2,3,6-Trimethylphenol	$C_9H_{12}O$	[2416-94-6]	136.19	62	226	158[a] (13.3)		0.963 (80)		
2,4,5-Trimethylphenol (pseudocumenol)	$C_9H_{12}O$	[496-78-6]	136.19	72	232	124			1.5070	10.72
2,4,6-Trimethylphenol (mesitol)	$C_9H_{12}O$	[527-60-6]	136.19	69	220.4	110			1.5105 (65)	
3,4,5-Trimethylphenol	$C_9H_{12}O$	[527-54-8]	136.19	109	248	141.5				10.51
2,3,4,5-Tetramethylphenol (prehnitenol)	$C_{10}H_{14}O$	[488-70-0]	150.21	82 – 87	256 – 260					
2,3,4,6-Tetramethylphenol (isodurenol)	$C_{10}H_{14}O$	[3238-38-8]	150.21	79 – 81	250					
2,3,5,6-Tetramethylphenol (durenol)	$C_{10}H_{14}O$	[527-35-5]	150.21	119	248.6	126.6				
Pentamethylphenol	$C_{11}H_{16}O$	[2819-86-5]	164.24	128	267					
Ethylphenols										
2-Ethylphenol	$C_8H_{10}O$	[90-00-6]	122.16	– 3.3; – 28 [b]	204.5	100.6	48.2	1.0146 (25)	1.5335 (20)	10.47
3-Ethylphenol	$C_8H_{10}O$	[620-17-7]	122.16	– 4.5	218.4	114.4	50.9	1.0076 (25)		10.17
4-Ethylphenol	$C_8H_{10}O$	[123-07-9]	122.16	45.1	218	114.3	50.9	1.054 (25) [c]	1.5239 (25) [d]	10.38
2,4-Diethylphenol	$C_{10}H_{14}O$	[936-89-0]	150.21		229			0.9794 (25)	1.5218 (25)	
2,6-Diethylphenol	$C_{10}H_{14}O$	[1006-59-3]	150.21	38 – 39	218	112			1.5105 (50)	
3,5-Diethylphenol	$C_{10}H_{14}O$	[1197-34-8]	150.21	77	248	131			1.5033 (75)	10.31
3-Ethyl-2-methylphenol	$C_9H_{12}O$	[1123-73-5]	136.19	71	227					
4-Ethyl-2-methylphenol	$C_9H_{12}O$	[2219-73-0]	136.19		224.5					
6-Ethyl-2-methylphenol	$C_9H_{12}O$	[1687-64-5]	136.19	147	213					
4-Ethyl-3-methylphenol	$C_9H_{12}O$	[1123-94-0]	136.19	26	228 – 230					
5-Ethyl-3-methylphenol	$C_9H_{12}O$	[698-71-5]	136.19	51.6	235.7	129	51.3		1.5293 (20) [d]	
6-Ethyl-3-methylphenol	$C_9H_{12}O$	[1687-61-2]	136.19	43.4	224.2					
2-Ethyl-4-methylphenol	$C_9H_{12}O$	[3855-26-3]	136.19	25.7	222.3					
3-Ethyl-4-methylphenol	$C_9H_{12}O$	[6161-67-7]	136.19	33	234 – 235					

[a] At 13.3 kPa.
[b] Metastable.
[c] Solid.
[d] Undercooled.

2,3,6-Trimethylphenol can also be obtained by methylation of 2,6-xylenol on γ-aluminum oxide, in particular if the reaction is carried out in a trickle bed [133]. Under optimum conditions (a molar ratio 2,6-xylenol – methanol of 2 : 1, 355 °C, 3.2 MPa, LHSV 4 h^{-1}, multitube reactor), however, a selectivity of only 50 % is achieved at 32 % conversion [134]. From the xylenol – trimethylphenol fractions thus obtained the 2,3,6-trimethylphenol must be isolated in pure form [126] by relatively sophisticated procedures, i.e., by alkaline countercurrent extraction [135] or, after spreading out the boiling points by *tert*-butylation of the mixture, by fractional distillation, debutylation, and a second fractionation [136].

Another industrial route to 2,3,6-trimethylphenol developed by BASF involves the condensation of diethyl ketone in the presence of base (NaOH) with methyl vinyl ketone [137] or crotonaldehyde [138] to give 2,3,6- and 2,5,6-trimethyl-2-cyclohexen-1-ones, respectively. These compounds are subsequently dehydrogenated in the gas phase at 250 – 300 °C on noble metal carrier catalysts [139]:

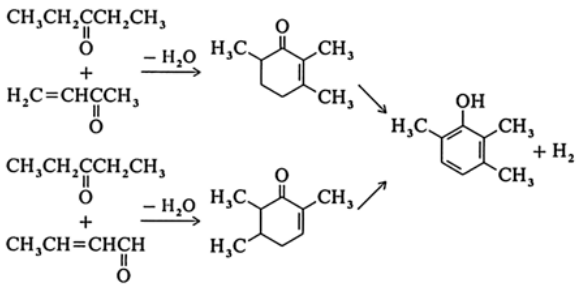

The condensation to give the trimethylcyclohexenones takes place in excess boiling diethyl ketone in yields of ca. 80 % [137], [138]. The selectivity of the dehydrogenation is 97 % on palladium – spinel catalysts, if the trimethylcyclohexanone that is formed as a byproduct in up to 10 – 15 % yield by hydrogenation is recycled into the dehydrogenation [139]. The dehydrogenation can also be performed at 220 – 225 °C on 1 % palladium – charcoal catalysts [140].

2,4,6-Trimethylphenol [527-60-6] (physical properties see Table 3) is the main product (> 80 %) of the gas phase alkylation of phenol with excess methanol (molar ratio 1 : 5) at 450 °C on magnesium oxide catalysts, which are doped with alkali or other metal oxides (e.g., V_2O_5, SrO). 2,6-Xylenol and 2,3,4,6-tetramethylphenol are the main byproducts [141], [142].

2,3,5-Trimethylphenol [697-82-5] and *2,3,5,6-tetramethylphenol* [527-35-5] (physical properties of both compounds are given in Table 3) are obtained by methylation of 3,5-xylenol in the presence of ortho-selective catalysts. A mixture of 3,5-xylenol – methanol vapors (molar ratio 1 : 5) is continuously reacted in a tubular reactor on a mixed MgO – UO_3 catalyst at 400 °C and an LHSV of 1 h^{-1}. The conversion of 3,5-xylenol is 75 %; a 2,3,5-trimethylphenol selectivity of ca. 75 % and a 2,3,5,6-tetramethylphenol selectivity of ca. 25 % are obtained [143].

In the same way *p*-cresol and 2,4-xylenol can be converted into *2,4,6-trimethylphenol*; 2,3- and 2,5-xylenol into *2,3,6-trimethylphenol*; and 3,4-xylenol into a mixture of *2,3,4-trimethylphenol*, *2,4,5-trimethylphenol* [496-78-6] (physical properties see Table 3) and *2,3,4,6-tetramethylphenol* [3238-38-8] (physical properties see Table 3) on ortho-selective

methylation catalysts.

2,4,6-Trimethylphenol can be isomerized in the presence of $AlCl_3$ – HCl at 85 °C to a mixture of 62 % 2,3,6-, 26 % 2,3,5-, and 12 % 2,4,6-trimethylphenols [144]. Under more stringent conditions all trimethylphenols rearrange in the presence of $AlCl_3$ – HCl to form the thermodynamically stable 2,3,5-trimethylphenol in good yields [145].

Producers. Producers of 2,3,6-trimethylphenol are, for example, Synthetic Chemicals in the United Kingdom, BASF and Lowi in Germany, and Honshu Chem. Ind. in Japan. 2,4,6-Trimethylphenol is produced by Lowi in Germany, General Electric in the United States, and Konan Chemicals, Mitsubishi Gas, and Nippon Kagaku in Japan. 2,3,5-Trimethylphenol is produced by Shell Chemicals in the United Kingdom.

Uses. *2,3,6-Trimethylphenol* is an important starting material for vitamin E synthesis (⟶ Vitamins); it also serves as a comonomer for the modification of polyphenyleneoxide resins (PPO). *2,3,5-Trimethylphenol* and *2,4,6-trimethylphenol* [142] are also starting materials for vitamin E. A mixture of the *N*-methylcarbamates of 2,3,5-trimethylphenol and 3,4,5-trimethylphenol is used as an insecticide.

The *tar alkylphenols* with a boiling point > 230 °C (to which the trimethylphenols and methylethylphenols belong) show bactericidal, fungicidal, and insecticidal action. Therefore these alkylphenols are contained in formulations of impregnating oils used for wood protection.

1.5.2. *Ethylphenols*

2-Ethylphenol [90-00-6] and *2,6-diethylphenol* [1006-59-3] (for physical properties of both compounds see Table 3) are produced industrially by ortho-alkylation of phenol with ethylene in high pressure autoclaves at 320 – 340 °C and 20 MPa in the presence of 1 – 2 % aluminum phenolate. If a molar ratio of phenol – ethylene of ca. 1 : 2 is used, the yield (related to converted phenol) is ca. 32 % 2-ethylphenol and ca. 39 % 2,6-diethylphenol after 6 h reaction time. About 17 % 6-*sec*-butyl-2-ethylphenol is formed as a byproduct [86].

Cresols and xylenols with free ortho positions can be ethylated similarly. The ethylation of 3,5-xylenol leads very smoothly to *2-ethyl-3,5-dimethylphenol* and *2,6-diethyl-3,5-dimethylphenol* [86].

Phenol can also be ethylated with ethanol , [146-149]. The conditions are similar to those for the methylation of phenol with methanol. A yield of 60 % 2-ethylphenol and 22 % 2,6-diethylphenol is achieved using an Fe_2O_3 catalyst doped with silicon and magnesium oxides in the gas phase at 360 °C (molar ratio phenol – ethanol – H_2O 1 : 6 : 1, LHSV of 0.6 h^{-1}) [131]. 2-Ethyl- and 2,6-diethylphenol are also formed preferentially on Al_2O_3 in the gas phase at ca. 300 °C [147]. With increasing temperature and residence time the ortho-selectivity decreases and the yield of 3- and 4-ethylphenol increases [148]. At 400 – 450 C° mixtures of isomers with a relatively high proportion of meta product can be obtained, in particular after increasing the isomerization and transalkylation activity of the catalyst by combination of the catalyst with other compounds [149]. If phenol is reacted with ethylene instead of ethanol, ca. 77 % conversion of phenol is achieved. Phenol and ethylene react, for example, in the gas phase at 420 °C on a phosphoric acid – SiO_2 catalyst to form an ethylphenol mixture, of which ca. 44 % is 3-ethylphenol [150]. The processes give product mixtures from which the pure meta or para isomers can only be isolated at relatively high cost. An especially high 4-ethylphenol selectivity (ca. 96 %) is possible according to [151] with ethylene – phenol – H_2O vapor mixtures (molar ratio 1 : 0.86 : 1.3) at 400 °C and low phenol conversions (ca. 7 %) on H-ZSM-5-zeolite modified with tetramethoxysilane. As far as is known these processes have not been exploited industrially.

4-Ethylphenol [123-07-9] (for physical properties see Table 3) is produced industrially by sulfonation of ethylbenzene under mild (kinetically controlled) conditions [152] and subsequent alkali fusion of the 4-ethylbenzenesulfonic acid obtained [153], [154]. The purity of the commercial product is 98 %.

3-Ethylphenol [620-17-7] (for physical properties see Table 3) is also produced by sulfonation of ethylbenzene and subsequent alkali melt of the 3-ethylbenzenesulfonic acid obtained. By sulfonation under severe (thermodynamically controlled) conditions (200 °C) the 4-ethylbenzenesulfonic acid formed initially isomerizes to form a mixture of 2-, 3-, and 4-ethylbenzenesulfonic acids in a ratio of ca. 2 : 58 : 40. The 2- and 4-ethylbenzenesulfonic acids are selectively hydrolyzed back to ethylbenzene and H_2SO_4 by blowing steam through the mixture at 150 – 170 °C. 3-Ethylphenol is then obtained by alkali melt of the remaining 3-ethylbenzenesulfonic acid at 330 – 340 °C in up to 98 % purity [155-157].

Producers. Producers of 2-ethylphenol are, for example, EMS-Dottikon in Switzerland and Coalite in the United Kingdom. 4-Ethylphenol is produced, e.g., by Synthetic Chemicals and Coalite in the United Kingdom and Maruzen Oil Co. and Konan Chemicals in Japan. Producers of 3-ethylphenol are, e.g., Coalite and Synthetic Chemicals in the United Kingdom and Taoka Chem. Co. in Japan.

Uses. 3-Ethylphenol and 4-ethylphenol are contained in the xylenol fraction boiling between ca. 205 and 225 °C obtained from tar phenols, which is used, for example, for the production of phenolic resins. A xylenol fraction boiling between 216 and 218 °C containing predominantly 3- and 4-ethylphenol is obtained from brown coal low-temperature tar and used in the varnish industry.

Pure 2-ethylphenol and 3-ethylphenol are starting materials for photochemicals. Pure 4-ethylphenol is the starting material for the production of 4-vinylphenol and of various antioxidants [e.g., 2,6-di-*tert*-butyl-4-ethylphenol and 2,2′-methylenebis(6-*tert*-butyl-4-ethylphenol)], which are used in rubber and polymers. 4-Ethylphenol is an intermediate for pharmaceuticals and dyes.

1.5.3. *Isopropylphenols*

Properties. Some physical properties of isopropylphenols mentioned in this article are listed in Table 4.

Table 4. Physical properties of isopropylphenols

Name (common name)	CAS registry no.	*mp*, °C	*bp*, °C
2-Isopropylphenol	[88-69-7]	16.8	215.5
3-Isopropylphenol	[618-45-1]	25.7	223.5
4-Isopropylphenol	[99-89-8]	63	223
2,4-Diisopropylphenol	[2934-05-6]	21.6	258
2,6-Diisopropylphenol	[2078-54-8]	19.6	236.5
3,5-Diisopropylphenol	[26886-05-5]	53.1	257.5
2,4,6-Triisopropylphenol	[2934-07-8]	28.7	255
5-Isopropyl-2-methyl-phenol (carvacrol)	[499-75-2]	2.5	237.4
6-Isopropyl-2-methyl-phenol	[3228-04-4]	–16.5	225
2-Isopropyl-3-methyl-phenol	[3228-01-1]	70 – 70.5	229
4-Isopropyl-3-methyl-phenol	[3228-02-2]	114	245
5-Isopropyl-3-methyl-phenol (isothymol)	[3228-03-3]	51	241
6-Isopropyl-3-methyl-phenol (thymol)	[89-83-8]	51.5	232.4
2-Isopropyl-4-methyl-phenol	[4427-56-9]	34	234

Production. Isopropylphenols are formed by reaction of phenol or alkylphenols with propene or isopropanol in the presence of acid catalysts of the Lewis or Brønsted type or of aluminum phenolate. An extensive study of the alkylation of phenol, both with propanol and isopropanol in the presence of alumina has been published [158].

Industrially, *2-isopropylphenol* is produced from phenol and propene using ϒ-Al_2O_3 as the catalyst. The alkylation is carried out continuously in a steel tube filled with ϒ-Al_2O_3 pressed into tablets, with graphite as a tabletting aid. The reaction temperature is 250 – 300 °C and the pressure is about 10 MPa, to keep both the phenol and the propene safely in the liquid phase. If a phenol : propene molar ratio of 1.3:1 is used, 2-isopropylphenol is obtained with a selectivity of about 85 % at complete propene conversion [159].

2,6-Diisopropylphenol is formed as a byproduct (10 – 12 %) in this process along with smaller amounts of *4-isopropylphenol* (1 – 2 %) and *isopropyl phenyl ether* [2741-16-4] (*bp* 176.8 °C). The proportion of ether formed increases considerably at temperatures below 250 °C. Isopropyl phenyl ether, once formed, remains unchanged over the alumina catalyst even at 250 – 300 °C. Isopropyl phenyl ether can be transformed mainly to 2-isopropylphenol by treatment at 85 °C with BF_3 [160] or with strongly acidic ion-exchange resins.

Monoisopropylphenol mixtures enriched in the 4-isomer can be synthesized from phenol and propene or isopropanol in the presence of zeolite catalysts [161], [162]. When a 1 : 1 molar mixture of phenol and isopropanol is passed over a ZSM-5-catalyst at 250 °C and LHSV = 1 h^{-1}, a phenol conversion of 20 % and a selectivity for 4-isopropylphenol of 63 % is achieved [161]. 1 : 1 Mixtures of 2- and 4-isopropylphenols are obtained by transalkylation of phenols with cumene in a HF – BF_3 medium at 30 °C [163].

Isomerization and transalkylation occur if (mixtures of) 2- and/or 4-isopropylphenol(s) are heated to 110 °C for 6 h in BF_3-containing hydrogen fluoride. The main reaction product is

3-isopropylphenol which is obtained in 61 % yield [164].

3- and 4-Isopropylphenol are minor byproducts of commercialized resorcinol and hydroquinone syntheses (⟶ Resorcinol ; ⟶ Hydroquinone) that use 1,3- and 1,4-diisopropylbenzene, respectively as starting materials. Production of 3- or 4-isopropylphenol via this autoxidation route is, however, not economical.

2,6-Diisopropylphenol is produced from phenol and propene (molar ratio 1 : 2.1) with homogeneously dissolved aluminum phenolate as catalyst, which is formed by dissolution of aluminum granules (1 wt %) in the educt phenol. The alkylation is carried out in an autoclave or pressure tube and takes place at 220 °C within 3 h. The workup procedure involves: (1) treatment of the reaction product with aqueous mineral acid to destroy the phenolate catalyst; (2) phase separation, (3) neutralization of the acid entrained in the organic phase with a base; and (4) a water wash. The product is subsequently distilled to yield 77 % 2,6-diisopropylphenol and 10 % 2-isopropylphenol [165], [166]. The wastewater is contaminated with phenolic compounds; therefore the procedure needs to be improved because of ecological considerations.

2,4,6-Triisopropylphenol is obtained in 70 % yield by reacting 1 mol phenol with 4 mol propene at 250 °C under pressure in the presence of fluorided alumina [167]. Diisopropylphenols are formed as byproducts. Mixtures of mono- di-, and triisopropylphenols can be produced from phenol and propene with aluminum chloride (at 50 – 150 °C) [168], [169] or with activated clays (at 100 – 150 °C) as catalysts [170]. These mixtures find various applications in the plastics industry. If desired, the proportion of 3- or 3,5-isomers can be increased by (1) raising the temperature (up to 200 °C), (2) increasing the amount of catalyst (up to 20 %) and/or (3) increasing the reaction time (> 20 h) [171]. The reaction mixtures are fractionated and fractions with desired composition are recycled to the alkylation reactor. If $AlCl_3$ is used as the catalyst, the mixture can be reacted with phosphorus oxychloride to form phosphoric acid tris(isopropyl phenyl ester) [168], [172]. Phenol alkylation and subsequent esterification of the alkylphenol can thus be carried out in one reaction vessel without intermediate separation of the catalyst and fractionation of the reaction mixture.

Several isopropylation products of the methylphenols (cresols) are used industrially. The most important is the *m*-cresol-derived *6-isopropyl-3-methylphenol* (*thymol*), a precursor of l-menthol (⟶ Flavors and Fragrances). *m*-Cresol with a purity ≥ 98.5 % must be used as the starting material for thymol production, because 2-isopropyl-4-methylphenol, derived from *p*-cresol (the main contaminant of commercial *m*-cresol), cannot be easily separated from thymol.

In the industrial process employed by Bayer [173], *m*-cresol and propene, both in the liquid state, are pumped through a pressure tube filled with activated alumina. The process is performed at a molar ratio of *m*-cresol – propene of 1 : 0.7, at 350 – 365 °C, ca. 5 MPa, and an LHSV of 0.25 h^{-1}. The reaction product consists of ca. 25 % *m*-cresol, 60 % thymol, and 15 % other products. Thymol of ≥ 99.5 % purity is obtained by rectification of the crude product. Undesired alkylates of *m*-cresol (2-, 4-, and 5-isopropyl-*m*-cresol; 2,6- and 4,6-diisopropyl-*m*-cresol) are recycled to the reaction zone to achieve a high overall yield of the desired thymol.

The influence of the reaction conditions on the formation of thymol from *m*-cresol and propene in the liquid phase in the presence of γ-Al_2O_3 has been described [175]. Impregnation with metal sulfates, e.g., $FeSO_4$, increases the activity of γ-Al_2O_3 [176], [177]. Aluminum phenolate [178] and aqueous zinc halide – hydrogen halide solution [179] have also been used as catalysts. With these systems selectivities for thymol formation are about 75 –80 %.

In 1988, a gas phase process for thymol manufacture was developed using medium-pore sized zeolites (e.g., erionite, mordenite, or ZSM-23) as heterogeneous catalyst; 80 – 100 % of the exchangeable cations in the zeolite are present as protons. Reaction temperatures are lower (230 –270 °C) than in the older liquid-phase process and the reaction pressure is normal or slightly elevated [174].

5-Isopropyl-3-methylphenol (isothymol) can be produced by employing catalysts that possess high isomerization and transalkylation activities. Examples include finely ground synthetic silica –alumina [180] and synthetic zeolites with faujasite structure, especially the wide-pored Y-type [181]. Acid-activated clays are less suitable as catalysts because they can lead to corrosion at the relatively high temperatures that are required for isomerization. Usually, the isothymol synthesis is conducted in two steps. First, 1 mol *m*-cresol is reacted with < 1 mol propene at 100 –200 °C in the presence of one of the aforementioned catalysts (10 wt % silica – alumina or 5 wt % zeolite Y). The alkylate is then kept in an autoclave at 300 °C until the thermodynamic equilibrium is established (ca. 5 – 10 h). The equilibrated mixture contains 5-, 6- and 4-isopropyl-*m*-cresols in a 75 : 20 : 5 ratio together with unreacted *m*-cresol and higher alkylates. The higher the excess of *m*-cresol used initially, the lower is the proportion of higher alkylates obtained. Using an efficient fractionating column, isothymol is distilled from the reaction mixture after removal of the catalyst by filtration. Undesired isomers of isothymol and higher alkylates can be recycled to the isomerization – transalkylation step for optimization of the yield. Isothymol can also be separated from its isomers, by alkylation (with isobutene) [182] or sulfonation (with sulfuric acid) of the isomeric mixture [183]. In contrast to its isomers, isothymol does not react with both compounds and thus can be separated from its alkylated/sulfonated isomers by a less efficient separation column. Isopropylation of *o*- and *p*-cresol yields rather complex product mixtures, depending on reaction conditions.

o-Cresol is isopropylated under kinetic control and with aluminum phenolate as catalyst to form *6-isopropyl-2-methylphenol* in good yield [178], which in turn can easily be isomerized to *4-isopropyl-2-methylphenol* by treatment with catalytic amounts of sulfuric acid at 60 °C [184]. Under thermodynamic control, at 300 °C, and with an activated clay as catalyst, the main product is *5-isopropyl-2-methylphenol (carvacrol)* [185].

p-Cresol reacts with propene in the presence of aluminum phenolate [178] or BF_3 – phosphoric acid [186] as catalysts to give 2-isopropyl-4-methylphenol and 2,6-diisopropyl-4-methylphenol under kinetic control. Conditions which establish the thermodynamic equilibrium lead to the additional formation of *3-isopropyl-4-methylphenol* and *2,5-* and *3,5-diisopropyl-4-methylphenols*. Product distributions resulting from the reaction of cresols with isopropanol in the presence of $AlCl_3$ and phosphoric acid have been studied [187-189].

If *xylenols* are to be isopropylated, the substitution rules given on page Alkylation Rules. can be applied to predict the composition of the resulting product mixtures.

Uses. Isopropylphenols have found relatively few commercial applications. As nontoxic substitutes the phosphate esters of isopropylphenol mixtures are used for cresylphosphates as secondary plasticizers and flame retardants in vinyl polymers, mainly poly(vinyl chloride) [167], [168]. 2-Isopropylphenol is an efficient antiskinning aid in air-drying coatings. 6-Isopropyl-3-methylphenol (thymol) is the starting material for an industrial l-menthol synthesis [190] (⟶ Flavors and Fragrances). Thymol and carvacrol (5-isopropyl-2-methylphenol) are used as antiseptics. The *N*-methylcarbamate derived from 5-isopropyl-3-methylphenol is a very efficient insecticide [181]. 2,4,6-Triisopropylphenol can be used as an antioxidant in motor fuels and lubricants [167]; its reaction products with alkylene oxides are reported to be useful nonionic surfactants with good surface activity and low foaming properties [191].

1.5.4. *sec-Butylphenols*

2-sec-Butylphenol [89-72-5], *mp* 12 °C, *bp* 226 – 228 °C, is obtained with high selectivity by reaction of phenol with *n*-butenes at 250 – 300 °C using γ-Al_2O_3 as catalyst and at a pressure of 3.5 – 8.0 MPa which keeps both reactants in the liquid phase. If the reaction is performed batchwise, the catalyst is finely ground and kept in suspension by stirring. In a continuous operation the catalyst is pressed into tablets with 10 – 20 % graphite as a tabletting aid, and positioned in a pressure tube. Typically, phenol and *n*-butene(s) in a 4 : 1 molar ratio are pumped over the catalyst at 250 °C, 3.5 MPa, with a residence time of 1 h. Conversions are 22 % (phenol) and 90 – 96 % (*n*-butenes); the selectivities for 2-*sec*-butylphenol and 4-*sec*-butylphenol [99-71-8] (*mp* 56 –61 °C, *bp* 135 – 136 °C at 3.33 kPa) are 95 % and 2.5 %, respectively. *sec*-Butyl phenyl ether [10574-17-1] (*bp* 101 – 103 °C at 5.8 kPa) is formed as a byproduct [192]. A gas-phase alkylation employing a multitubular reactor has been described [193].

Aluminum phenolate (prepared in situ by dissolving about 3 wt % of aluminum granules or chips in the educt phenol) can also be used as the alkylating catalyst. Alkylation takes place with *n*-butene(s) at temperatures of 150 – 180 °C and by increasing pressure up to ca. 5 MPa. Depending on the amount of butene(s) introduced, either 2-*sec*-butylphenol or 2,6-di-*sec*-butylphenol [5510-99-6] (*mp* 42 °C, *bp* 225 – 260 °C) is obtained as the principal product. The usual workup (treatment with aqueous mineral acid to destroy the catalyst) can be circumvented by flash distillation of the crude reaction product at temperatures < 225 °C (vacuum). Under these conditions dealkylation is prevented and the catalyst-containing distillation residue can be recycled back to the alkylation step [194]. Alternatively, the distillation residue can be subjected to a transalkylation reaction after addition of phenol and heating to 250 – 350 °C [195].

Producers. 2-*sec*-Butylphenol is commercially produced by Hüls, Ethyl, Pearson, Shell, and Schweizerische Teerindustrie AG (STIAG). On request 2,6-di-*sec*-butylphenol will also be provided. Product purities are about 98 %. The main use of 2-*sec*-butylphenol is in the synthesis of insecticides, acaricides, and herbicides.

1.5.5. *tert-Butylphenols*

Properties. Boiling and melting points of selected *tert*-butylated phenols as well as a qualitative note on their solubility in alkali are listed in Table 5. Solubility in 2.5 N NaOH can be useful for separating substances with similar boiling points.

Table 5. Physical properties of *tert*-butylphenols

Name	CAS registry no.	*mp*, °C	*bp*, °C 101.3 kPa	*bp*, °C (kPa)	Solubility in 2.5 N NaOH *
2-*tert*-Butylphenol	[88-18-6]	–7	224	99 (1.33)	s
3-*tert*-Butylphenol	[585-34-2]	43	240	129.5 (2.66)	s
4-*tert*-Butylphenol	[98-54-4]	99.5	239.8	114 (1.33)	s
2,4-Di-*tert*-butylphenol	[96-76-4]	56.5	264	146 (2.66)	i
2,6-Di-*tert*-butylphenol	[128-39-2]	39	253	133 (2.66)	i
3,5-Di-*tert*-butylphenol	[1138-52-9]	87 – 89			
2,4,6-Tri-*tert*-butylphenol	[732-26-3]	131	278	158 (2.66)	i
4-*tert*-Butyl-2-methylphenol	[98-27-1]	28	247	174 (13.3)	s

6-*tert*-Butyl-2-methylphenol	[2219-82-1]	30 – 32	230	159 (13.3)	i
4-*tert*-Butyl-3-methylphenol	[2219-72-9]	72		152 (2.66)	s
5-*tert*-Butyl-3-methylphenol	[4892-31-3]	50		142 (2.66)	s
2-*tert*-Butyl-5-methylphenol	[88-60-8]	23		129.5 (2.66)	i
2-*tert*-Butyl-4-methylphenol	[2409-55-4]	51 – 52	232.7	127 (2.66)	s
4,6-Di-*tert*-butyl-2-methylphenol	[616-55-7]	52	269	149.5 (2.66)	i
4,6-Di-*tert*-butyl-3-methylphenol	[497-39-2]	62	282	167 (2.66)	i
2,6-Di-*tert*-butyl-4-methylphenol	[128-37-0]	70 – 71	266	147 (2.66)	i
6-*tert*-Butyl-2,4-dimethylphenol	[1879-09-0]	22 – 23	249	131 (2.66)	i
4-*tert*-Butyl-2,5-dimethylphenol	[17696-37-6]	71	264	193 (13.3)	s
4-*tert*-Butyl-2,6-dimethylphenol	[879-97-0]	82 – 83	248	176 (13.3)	s
2-*tert*-Butyl-4,5-dimethylphenol	[1445-23-4]	46		187 (13.3)	s

* s = soluble; i = insoluble.

Production. *2-tert-Butylphenol* is obtained as the main alkylation product if phenol is reacted with isobutene in the presence of sulfonated polystyrene – polydivinylbenzene ion-exchange resins. The acidity of the ion-exchange resins is depressed to an exchange capacity of about 2.5 meq/g (dry weight) by having been used for related reactions, e.g., for the production of higher alkylphenols [196]. For a definition of exchange capacity, see ⟶ Ion Exchangers – Exchange Capacity. A preconditioned catalyst of this kind that is placed in a tower reactor on a supporting sieve, permits production of 2-*tert*-butylphenol on an industrial scale. Reaction of phenol and isobutene in a 1 : 0.42 molar ratio at 95 °C gives a product mixture which has the following composition after removal of unconverted phenol:

2-*tert*-butylphenol	39.7 %
4-*tert*-butylphenol	38.5 %
2,4-di-*tert*-butylphenol	21.4 %
highboilers	0.4 %

2-*tert*-Butylphenol is obtained in a purity of 99 % by distillation; 4-*tert*-butylphenol and 2,4-di-*tert*-butylphenol are marketable byproducts.

2-*tert*-Butylphenol is also formed with relatively high selectivity by reaction of phenol and isobutene over a γ-Al_2O_3-catalyst that has been tabletted with 10 % graphite and preconditioned by heating to 650 °C in an air stream. Phenol and isobutene conversions are 52 and 56 % at 150 – 160 °C, a molar ratio phenol : isobutene of 1 : 1, a pressure of 3.5 MPa, and a residence time of 1 h. The selectivity for 2-*tert*-butylphenol is 88 % based on phenol conversion; 4-*tert*-butylphenol (2.5 %) and 2,4-di-*tert*-butylphenol (5 %) are formed as byproducts. *tert*-Butyl phenyl ether [6669-13-2], *bp* 185 °C, is isolated as a byproduct (or intermediate). The amount of *tert*-butyl phenyl ether formed decreases with increasing residence times and reaction temperatures, and increases with age of catalyst. The ether can be recycled to the reaction zone.

4-tert-Butylphenol is obtained by reaction of phenol with isobutene at normal or slightly elevated pressure and ca. 80 – 140 °C in the presence of strongly acidic catalysts such as sulfuric [197] or phosphoric acids [198], boron trifluoride [199], activated clays [200-202], [209], zeolites [207], [208] or strongly acidic ion-exchange resins [203-206]. 4-*tert*-Butylphenol is obtained in ca. 70 % yield if *tert*-butyl phenyl ether is heated to 100 °C in the presence of a clay catalyst that has been activated with sulfuric acid [209]. Silica – alumina (75 % SiO_2) which has been sintered at 600 °C is said to catalyze isomerizations – trans-alkylations. Mixtures of phenol with 2-*tert*-butylphenol and 2,4-di-*tert*-butylphenol give 4-*tert*-butylphenol at 165 °C [210]. In the industrial process (Fig. 1), dehydrated strongly acidic sulfonated polystyrene – polydivinylbenzene ion-exchange resins of the macroporous form (⟶ Ion Exchangers) are employed which are placed on supporting sieves in column-type reactors. For maximum yields two reactors in series, held at different temperatures, must be used. The first reactor, where the exothermic alkylation takes place (– 90 kJ per mole isobutene reacted), is held at 90 – 100 °C; the second reactor (isomerization – transalkylation reactor) is held at 120 °C. (Isomerization and transalkylation reactions are thermoneutral to slightly endothermic.) The reactants are introduced to the top of the first reactor in a phenol – isobutene molar ratio of 1.4 : 1. Composition of the reaction products leaving the alkylation and the isomerization – transalkylation zones are given in Table 6. The reaction product leaving the isomerizer is worked up by continuous distillation under reduced pressure (4 – 10 kPa). A flow sheet of the industrial process [203] is depicted in Figure 1. Product yield is ≥ 95 % and product purity is > 98 %. 2,4-Di-, 2,4,6-tri-*tert*-butyl-, and 4-*tert*-octylphenol are formed as byproducts in the 4-*tert*-butylphenol production process. Their proportion increases with decreasing phenol – isobutene molar ratio and/or less stringent reaction conditions. 2,4-Di-*tert*-butylphenol can be made from phenol [211] or 4-*tert*-butylphenol [212] by alkylation with isobutene under mild reaction conditions and in the presence of small amounts of sulfuric acid at temperatures below 80 °C. 2,4,6-Tri-*tert*-butylphenol can be obtained from either of the three aforementioned phenols under these reaction conditions [213].

Table 6. Composition (wt %) of a phenol – isobutene alkylation product. Catalyst: cation-exchange resin; molar ratio 1.4 : 1 (phenol – isobutene)

Component	Alkylation	Isomerization–transalkylation
Isobutene	0.19	0.07
Phenol	30.17	22.16
4-*tert*-Butylphenol	33.92	75.08
2-*tert*-Butylphenol	10.19	1.62
2,4-Di-*tert*-butylphenol	23.71	0.82
2,4,6-Tri-*tert*-butylphenol + 4-*tert*-octylphenol	0.41	0.22

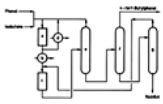

Figure 1. 4-*tert*-Butylphenol plant

a) Alkylator; b) Cooler; c) Isomerizer-transalkylator; d) Cooler; e) Distillation (phenol, 2-*tert*-butylphenol); f) Distillation; g) Vaporizer (2,4-di-*tert*-butylphenol recovery)

3-tert-Butylphenol is obtained from (1) 4- or 2-*tert*-butylphenol [214], [215], or (2) mixtures of 2,4- and 2,6-di-, or 2,4,6-tri-*tert*-butylphenol with phenol by heating to 170 – 200 °C in the presence of acid-activated clays [216], (synthetic) silica – alumina, modified zeolite of the Y-type [217], [218], or trifluoromethanesulfonic acid [219]. In the equilibrium state the monoalkylated fraction contains ≥ 70 % 3-*tert*-butylphenol and ca. 20 % of the 4-isomer; the dialkylated fraction contains ca. 80 % 3,5-di-*tert*-butylphenol [217]. A selective debutylation of the 4-isomer can be achieved. Thus, heating a 70 : 30 mixture of 3- and 4-*tert*-butylphenols to 160 – 170 °C for 5 h, after addition of 1 – 2 wt % H_2SO_4, followed by neutralization of the mineral acid and washing, yields 3-*tert*-butylphenol in > 90 % purity [220].

3-*tert*-Butylphenol can also be synthesized by heating 2-chloro-*tert*-butylbenzene with potassium-*tert*-butylate in *tert*-butanol to 240 °C for 4 h in an autoclave [221].

3,5-Di-tert-butylphenol is formed by heating isomeric mixtures of di-*tert*-butylphenols with acid-activated clays to 175 – 190 °C [222]. 3-*tert*-Butylphenol and/or 3,5-di-*tert*-butylphenol (depending on the chosen degree of alkylation) are obtained by reaction of phenol with isobutene at –78 to +25 °C in hydrofluoric acid [223]. Rearrangements of mono-*tert*-butylphenols in hydrofluoric acid as the solvent have been studied [224].

2,6-Di-tert-butylphenol (and 2-*tert*-butylphenol) is produced from phenol and isobutene in the presence of aluminum phenolate which is homogeneously dissolved in the phenol. The process, which has considerable industrial importance, was originally developed by Bayer [225].

Batchwise processes are preferred for industrial production using a stainless steel stirred-tank reactor provided with an efficient cooling device, e.g., an internal cooling coil. The reactor is charged with molten phenol, and flushed with pure nitrogen to remove all atmospheric oxygen. Metallic aluminum (1 wt %) in the form of granules or powder is added, and the reactor is heated to 120 – 180 °C for 0.5 – 1 h until the metal has dissolved to form aluminum phenolate under evolution of hydrogen which is vented. After cooling to 100 °C, liquid isobutene (1.8 – 2 mol per mole of phenol) is introduced into the reactor with stirring. During the reaction time of about 2 h, the pressure rises from 0.1 – 0.4 MPa to ca. 1.5 MPa. After stirring for a further 30 min, the pressure decreases to about 0.5 MPa. The reactor is depressurized and the reaction mixture is carefully washed with dilute sodium hydroxide and water. Rectification in vacuum yields about 75 % 2,6-di-*tert*-butylphenol, 10 % 2-*tert*-butylphenol, and a distillation residue consisting mainly of 2,4,6-tri-*tert*-butylphenol. 2-*tert*-Butylphenol can be recycled to the reactor. If desired, a higher amount of the monoalkylated phenol can be produced as a marketable product by using less than 1.8 mol isobutene per mole of phenol. The composition of the reaction products is dependent on the amount of isobutene used (see Table 7).

Table 7. Compositions (wt %) of phenol alkylation products using aluminum triphenolate as the catalyst (1 wt % Al dissolved in the educt phenol)

Component	Molar ratio isobutene : phenol	
	1.8	2.1
Phenol	4.4	0.6
2-*tert*-Butylphenol	10.3	1.7
4-*tert*-Butylphenol		0.2
2,6-Di-*tert*-butylphenol	73.5	75.9
2,4-Di-*tert*-butylphenol	2.2	0.9
2,4,6-Tri-*tert*-butylphenol	9.6	19.4

Exact control of temperature is important during the alkylation reaction, because at temperatures much lower than 100 °C *tert*-butyl phenyl ether formation prevails, whereas at higher temperatures ortho selectivity is progressively lost.

Industrially, the classical process provides an overall selectivity of about 85 %. Improvements have been proposed. According to an invention of Ethyl [226], selectivity can be as high as 95 % if 2-*tert*-butylphenol is used as starting material instead of phenol and if aluminum tris(2-*tert*-butylphenolate) is used as catalyst which is prepared in situ by reaction of 2-*tert*-butylphenol with triethylaluminum (ethane is evolved and vented). The reaction temperature should be preferably 0 – 10 °C. *tert*-Butyl phenyl ether is not formed because the catalyst is not acidic. The process is advantageous for producers that have an independent source of 2-*tert*-butylphenol. A simpler nonaqueous workup procedure has also been described [227]. Addition of small amounts of glycols and glycol ethers to the reaction mixture safely deactivates the aluminum phenolate catalyst and allows for direct recovery of products by distillation. Disposal of alkaline wash waters that are contaminated by phenolic bodies is therefore unnecessary.

Aluminum phenolate catalysis can be applied to produce 2-methyl-6-*tert*-butylphenol from *o*-cresol and isobutene.

Mono- and dibutylations of mixed *m*- and *p*-cresols, and of mixed 2,4- and 2,5-xylenols with gaseous or liquefied isobutene are carried out industrially in the presence of concentrated sulfuric acid (0.5 – 2 wt %). In the first step the corresponding phenolsulfonic acids are formed as the actual alkylation catalysts. A water content of more than 0.1 % in the educt phenols can be compensated for by using a higher proportion of sulfuric acid or by using oleum. Reactions occur at 50 – 80 °C and atmospheric to slightly elevated pressures.

For monobutylation of cresols an isobutene –cresol molar ratio of 0.7 – 1 is applied; the maximum yield of 80 % is obtained at a molar ratio of 1 [228], [229]. *p*-Cresol is converted to 2-*tert*-butyl-4-methylphenol (with some 2,6-di-*tert*-butyl-4-methylphenol), whereas *m*-cresol forms 2-*tert*-butyl-5-methylphenol (and some 4,6-di-*tert*-butyl-3-methylphenol). 4-*tert*-Butyl-3-methylphenol is formed from *m*-cresol under very mild, kinetically controlled conditions (very little H_2SO_4, < 40 °C) [230]. *o*-Cresol gives a mixture of 6-*tert*-butyl-2-methyl- and 4-*tert*-butyl-2-methylphenols, along with some 4,6-di-*tert*-butyl-2-methylphenol. The butylation products are usually worked up by neutralization of the catalyst with a slight excess of concentrated sodium hydroxide solution and subsequent distillation. Unconverted cresols and dibutylated compounds are recycled. The yield of monobutylated products is ≥ 90 %.

6-*tert*-Butyl-3-methylphenol can also be made by transbutylation of *m*-cresol with 4,6-di-*tert*-butyl-3-methylphenol, which is an intermediate of the *m/p*-cresol separation process (see below). The transbutylation is achieved by heating the mixture to 80 – 100 °C in the presence of acid-activated clays [231].

4-*tert*-Butyl-3-methylphenol is obtained only with difficulty. It can be formed by controlled debutylation of 4,6-di-*tert*-butyl-3-methylphenol by heating with the phenolates of aluminum [232] or (better) of zirconium [233], [234] under reflux conditions. Trans- and debutylation mixtures are worked up by fractionation, after deactivation of the catalyst by addition of alkali. 4-*tert*-Butyl-3-methylphenol can then be separated from 2-*tert*-butyl-5-methyl- and 2,6-di-*tert*-butyl-3-methylphenols because of its solubility in alkali [235].

5-*tert*-Butyl-3-methylphenol can be prepared by isomerization of 2-*tert*-butyl-5-methylphenol using substantial quantities of aluminum chloride at 30 °C [236].

2,6-Di-*tert*-butyl-4-methylphenol, frequently abbreviated as BHT (butylated hydroxytoluene), is a widely used antioxidant (⟶ Antioxidants – Hindered Phenols). Production of BHT involves reaction of pure and anhydrous (< 0.1 wt % H_2O) *p*-cresol with two or slightly more than two moles of isobutene in the presence of 2 wt % concentrated sulfuric acid at 70 °C and at an isobutene partial pressure of ca. 0.1 MPa. The reaction time is about 5 h. The reaction mixture contains about 95 % of BHT; some monobutylated phenols are formed as byproducts [237]. The usual workup procedure includes removal of acid by several washings with water at 70 °C, crystallization of the crude product, and a recrystallization from ethanol – water. The mother and washing liquids are first distilled at normal pressure for recovery of the ethanol. The remaining still bottoms separate into (1) an organic phase which is distilled under vacuum to recover 2-*tert*-butyl-4-methylphenol (to be recycled); and (2) an aqueous phase which is subjected to heat treatment (200 °C) to hydrolyze alkylphenolsulfonic acids into alkylphenols (which are recycled) and dilute sulfuric acid (which is neutralized and discarded) [238].

Alternatively, *m-/p*-cresol mixtures which cannot be separated by distillation, are first dibutylated and, after neutralization of the acid catalyst, are fractionated under vacuum to yield 2,6-di-*tert*-butyl-4-methylphenol (*bp* 147 °C at 2.66 kPa) and 4,6-di-*tert*-butyl-3-methylphenol (*bp* 167 °C at 2.66 kPa). The latter is subsequently reconverted to *m*-cresol by heating with 1 wt % oleum at 160 – 200 °C. The butyl group in the 4-position is eliminated first to yield 2-*tert*-butyl-5-methylphenol as intermediate. If required, this compound can be recovered by distillation of the partially debutylated product after neutralization of the acidic catalyst.

Several syntheses for 2,6-di-*tert*-butyl-4-methylphenol have been developed which are not based on *p*-cresol. At Shell, 2,6-di-*tert*-butylphenol was condensed with formaldehyde to yield 4,4′-methylenebis(2,6-di-*tert*-butylphenol) [118-82-1] which was subsequently heated to 200 °C with NaOH – methanol to give BHT in 80 % yield [239]. This synthesis can be conducted in one step by heating 2,6-di-*tert*-butylphenol with formaldehyde in the presence of NaOH – methanol [240]. In the Soviet Union, 2,6-di-*tert*-butyl-4-dimethylaminomethylphenol [88-27-7] (a Mannich base) is produced from 2,6-di-*tert*-butylphenol, formaldehyde, and dimethylamine, which in turn can be hydrogenated (Raney nickel, 120 °C) to give BHT in excellent yields [241], [242].

According to an invention of Ethyl, primary straight-chain alkyl groups can be introduced to the (ortho- or) para-position of phenols by heating them with *n*-alcohols that are presaturated with potassium hydroxide. However, this process does not take place with methanol. 2,6-Di-*tert*-butyl-4-ethylphenol [4130-42-1] and 4-butyl-2,6-di-*tert*-butylphenol [5530-30-3] became available by this “alkaline alkylation method” [243].

Mixtures of 2,4- and 2,5-xylenol can be alkylated with isobutene in the presence of concentrated sulfuric acid at 40 – 60 °C [244], [245]. 2,4-Xylenol reacts faster than 2,5-xylenol. Usually the reaction is terminated when all the 2,4-xylenol is converted. Fractionation of the neutralized and washed product mixture yields unconverted 2,5-xylenol, 6-*tert*-butyl-2,4-dimethylphenol, and a certain amount of the higher boiling 4-*tert*-butyl-2,5-dimethylphenol. Alternatively, the reaction mixture is first washed with dilute alkali to remove unchanged 2,5-xylenol (and, if present, 2,4-xylenol) and its mono-*tert*-butyl derivative. The insoluble residue is then distilled to give 6-*tert*-butyl-2,4-dimethylphenol [245].

Uses. *2-tert-Butylphenol* is a starting material for the synthesis of antioxidants and agrochemicals. Fragrance compounds are made from *cis*-2-*tert*-butylcyclohexanol [7214-18-8] which is obtained by hydrogenation of 2-*tert*-butylphenol in the presence of Pd – Al_2O_3 or Ru – Al_2O_3 catalysts [246]. Thermodynamic aspects of the stereoselective hydrogenation have been studied [247]. The acetate of the cis isomer [20298-69-5] has an orris-like fragrance.

3-tert-Butylphenol has found little if any commercial application. *4-tert-Butylphenol* is used on a large scale for the production of a variety of phenol – formaldehyde resins (⟶ Phenolic Resins – Aldehydes). If used as the single phenolic component, oil-soluble resins result; if used in admixture with phenol and *m*-cresol, resins with controllable cross-linking are obtained. These resins, which are applied as binders in surface coatings, lacquers, and varnishes, have good air-drying properties and durability against the deteriorating influences of light, weather, and chemicals. 4-*tert*-Butylphenol-derived novolac resins with low viscosity are used as tackifiers for oil-extended SBR – natural rubber blends [248] and as plasticizers for cellulose acetate. Under slightly modified conditions, formaldehyde and 4-*tert*-butylphenol (as well as other 4-alkylated phenols) condense to form cyclic condensation products called calixarenes because of their beaker- or cup-like shape [249], [250]. These compounds possess clathrating properties for cations.

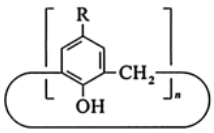

R = *tert*-butyl, higher alkyl; *n* = 4, 6, 7, 8

An oil-soluble calixarene (R = *n*-hexyl, *n*=6) [117397-61-2] is a very efficient extraction agent for uranyl ions from aqueous solutions [251].

4-*tert*-Butylphenol has antioxidant properties and is used as a stabilizer in rubber and chlorinated hydrocarbons, and, formerly, in soaps. The esters of 4-*tert*-butylphenol with phosphorous and phosphoric acids are UV stabilizers for plastics and lacquers. Esters of 5-*tert*-butylsalicylic acid [16094-31-8], which is obtained from 4-*tert*-butylphenol by Kolbe synthesis, are used for the same purpose. 4-*tert*-Butylphenol serves as a chain-length regulator in the production of polycarbonate resins.

The alkoxylates (i.e., reaction products with alkylene oxides) of 4-*tert*-butylphenol are emulsifiers and wetting agents which can also act as demulsifiers, e.g., in crude oil. Hydrogenation of 4-*tert*-butylphenol yields *cis*- and *trans*-4-*tert*-butylcyclohexanols [98-52-2], whose esters with acetic acid are widely used as a substitute for lemon grass oil, especially in soap perfumes. 4-*tert*-Butylcyclohexanone [98-53-3] has a strong camphor-like odor and is used in scenting detergents and soaps.

The main use of *2,4-di-tert-butylphenols* is in the manufacture of its triphosphite [31570-04-4] that is employed as a (co-)stabilizer for poly(vinyl chloride), and of its benzotriazole derivatives that are used as UV absorbers in polyolefins.

2,6-Di-tert-butylphenol is an indispensable building block in the synthesis of higher molecular mass antioxidants and light-protection agents for plastics, especially polyolefins. Frequently used intermediates are 3,5-di-*tert*-butyl-4-hydroxybenzyl alcohol [88-26-6] (**3**), 3,5-di-*tert*-butyl-4-hydroxybenzyl chloride [955-01-1] (**4**), methyl-3-(3,5-di-*tert*-butyl-4-hydroxyphenyl)- propionate [6386-38-5] (**5**), and 3,5-di-*tert*-butyl-4-hydroxybenzoic acid [1421-49-4] (**6**). These compounds are prepared from 2,6-di-*tert*-butylphenol by mild hydroxymethylation (**3**) [252],

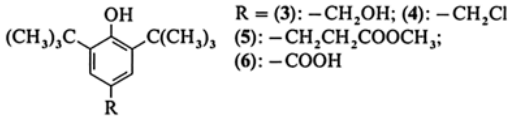

chloromethylation (**4**) [253], conjugate addition to methyl acrylate (**5**) [254], and carboxylation with carbon dioxide (**6**) [255]. 2,6-Di-*tert*-butylphenol reacts with formaldehyde to give 4,4′-methylenebis(2,6-di-*tert*-butylphenol) [118-82-1], an effective antioxidant for lubricating oils.

In the 1980s much research was focused on the oxidative coupling of 2 molecules of 2,6-di-*tert*-butylphenol to form 3,3′,5,5′-tetra-*tert*-butyldiphenoquinone which is then reduced to the corresponding 3,3′,5,5′-tetra-*tert*-butyl-4,4′-diphenol [2455-14-3]. After removal of the butyl groups this compound yields 4,4′-dihydroxydiphenyl [92-88-6] which is a building block for thermoresistant polymers.

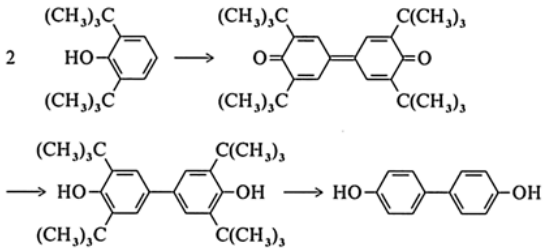

2,4,6-Tri-tert-butylphenol is the starting material for the synthesis of 2,6-di-*tert*-butyl-4-methoxyphenol [489-01-0] which is a powerful antioxidant. The first step involves oxidation of 2,4,6-tri-*tert*-butylphenol with Cl_2 in methanol in the presence of Na_2CO_3 followed by debutylation with sulfuric acid in methanol.

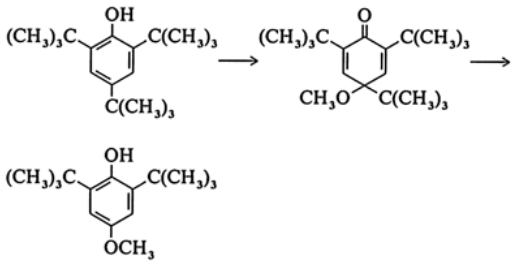

Monotert-butyl derivatives of m- and p-cresol are precursors for a number of commercially important antioxidants and light protection agents of the bisphenol and thiobisphenol type. Examples are 2,2′-methylenebis(6-*tert*-butyl-4-methylphenol) [119-47-1], 4,4′-methylenebis(6-*tert*-butyl-3-methylphenol) [2872-08-4], as well as the corresponding 2,2′-thiobis(6-*tert*-butyl-4-methylphenol) [90-66-4] and 4,4′-thiobis(2-*tert*-butyl-5-methylphenol) [96-69-5].

An important UV absorber, 2-(3-*tert*-butyl-2-hydroxyphenyl-5-methyl)-2 *H*-5-chlorobenzotriazole [3896-11-5], is produced from *2-tert-butyl-5-methylphenol*. 2-*tert*-Butyl-5-methylphenol is also used for the preparation of musk ambrette (6-*tert*-butyl-3-methyl-2,4-dinitroanisole) [83-66-9] which is a perfume fixative. The *N*-methylcarbamate of 5-*tert*-butyl-3-methylphenol is an insecticide. *2,6-Di-tert-butyl-4-methylphenol* (BHT) has found broad application as an antioxidant in polymers, in technical oils, as well as in edible oils and fat-containing foods where it retards the development of rancidity.

2,6-Di-tert-butyl-4-ethylphenol and *4-butyl-2,6-di-tert-butylphenol* are also used as antioxidants. *2-tert-Butyl-4-ethylphenol* is used as a precursor for the synthesis of other antioxidants.

Economic Aspects. The worldwide annual production capacity for *tert*-butylphenols is about 128 000 t, which can be broken down as follows:

Western Europe	53 000 t
United States	60 000 t
Japan	12 000 t

About 25 000 t/a of 2,6-di-*tert*-butylphenol and about 15 000 t/a of 2,4-di-*tert*-butylphenol are produced worldwide. The consumption of BHT as a stabilizer of polymers is about 15 000 t/a; the total consumption of BHT is higher by a factor of 2 – 3.

Toxicology and Occupational Health. Toxicological data for butylphenols are listed in Table 8. 4-*tert*-Butylphenol was found to cause leucoderma, a depigmentation of the skin (vitiligo), in a few sensitive individuals [256], [257]. The condition is likely to be long-lasting and cannot be treated. The appearance of vitiligo is greatly reduced if skin contact is avoided and there is good ventilation. The MAK value for 4-*tert*-butylphenol is 0.08 mL/m^3 (0.08 ppm), equivalent to 0.5 mg/m^3.

Table 8. Toxicological data of *tert*-butylphenols

	LD_{50} (oral, rat),	Effect on	
	mg/kg	Skin	Eyes
2-*tert*-Butylphenol	868	severe irritation	cauterization
4-*tert*-Butylphenol	4000	moderate irritation	severe irritation
2,4-Di-*tert*-butylphenol	3250	moderate irritation	moderate irritation
2,6-Di-*tert*-butylphenol	9200	irritation not reported	irritation not reported

Storage, Packaging, Transportation. In the presence of atmospheric oxygen, *tert*-butylphenols, like other alkylphenols, are prone to discolorations which are enhanced under the influence of light, elevated temperatures, traces of heavy metals (copper, iron), dissolved alkali, and moisture. During storage and transportation these influences must be kept to a minimum. *tert*-Butylphenols are therefore stored under an atmosphere of pure nitrogen.

In the liquid (molten) state, *tert*-butylphenols are stored and transported in stainless steel containers under a blanket of nitrogen. 2-*tert*-Butylphenol is transported in road or rail tank cars and in 200 L drums. 4-*tert*-Butylphenol is transported in the liquid state in thermostated road or rail containers at a temperature of about 110 °C. In the solid state it is stored under nitrogen in polyethylene bags of about 20 kg weight each that are palleted to a maximum weight of 1 t. As a solid 4-*tert*-butylphenol must be stored at < 40 °C. 2,4-Di-*tert*-butylphenol is transported in thermostated road tank cars at a temperature of 70° in the liquid state and iron drums (200 L) are used for the solid material. Transport classifications

for butylphenols are given in Table 9.

Table 9. Classification of butylphenols in land-, sea-, and air-transportation

Name	State [*]	*mp*, °C	Flash point, °C	Land		Sea				Air		
				GGVS/GGVE, ADR/RID		GGVSee, IMDG code (Amdt 25–89)				ICAO code, IATA-DGR		
				Class	No.	Class	UN no.	Page	PG	Class	UN no.	PG
2-*sec*-Butylphenol	l	ca. 14	ca. 101	8	66 b	8	1760	8147	II	8	1760	II
2-*tert*-Butylphenol	l	ca. –38	ca. 80	6.1	14 c	6.1	2228	6091	III	6.1	2228	III
4-*tert*-Butylphenol	s	ca. 100	ca. 115	6.1	14 c	6.1	2229	6091	III	6.1	2229	III
2,4-Di-*tert*-butylphenol	s	ca. 53	ca. 129									
2,6-Di-*tert*-butylphenol	l/s	ca. 39	ca. 118									

[*] l = liquid; s = solid.

1.5.6. *tert-Pentylphenols*

Properties. Only a few physical constants of the *tert*-pentylphenols are available (see Table 10). Synonyms for the *tert*-pentyl group are 2-methyl-2-butyl or *tert*-amyl.

Table 10. Physical properties of *tert*-pentylphenols

Name	CAS registry no.	*mp*, °C	*bp*, °C	
			101.3 kPa	2.66 kPa
2-*tert*-Pentyl-phenol	[3279-27-4]			124
4-*tert*-Pentyl-phenol	[80-46-6]	95	266	142
3-*tert*-Pentyl-phenol	[27336-20-5]			143[*]
2,4-Di-*tert*-pentylphenol	[120-95-6]	27		166
2,6-Di-*tert*-pentylphenol	[3279-20-7]		283	165

[*] At 3.33 kPa.

Production. *tert*-Pentylphenols are obtained from the reaction of phenol with 2-methyl-1-butene, 2-methyl-2-butene, and/or 3-methyl-1-butene in the presence of acidic catalysts, since methylbutenes isomerize to form the most stable cation which is the tertiary pentyl. When phenol and olefin are employed in equimolar amounts, and boron trifluoride [258], phosphoric acid [259], or strongly acidic ion-exchange resins are used as catalyst [260], yields as high as 90 – 95 % 4-*tert*-pentylphenol are obtained at 100 – 120 °C. With a molar excess of olefin, 2,4-di-*tert*-pentylphenol is also formed, especially if acid-activated clays are used as catalysts [261], [262]. The residual C_5-fraction of an isoprene extraction plant can be used as a source of methylbutenes, provided that the cyclopentene fraction has been selectively hydrogenated on a Pd/Al_2O_3 catalyst prior to the alkylation [263].

Uses. Condensates of 4-*tert*-pentylphenol with formaldehyde are compounded with a drying oil and used as oil-soluble resins for application in surface coatings. Ethoxylation of 4-*tert*-pentylphenol forms surface-active substances which can be used as demulsifiers in crude oils. 2,4-Di-*tert*-pentylphenol is used for the preparation of photographic color developers, textile processing aids, and light stabilizers.

1.5.7. *Higher Alkylphenols*

Properties. This group of compounds includes alkylphenols containing at least one alkyl group with six or more carbon atoms. Physical data and specifications of higher alkylphenols of commercial importance are given in Table 11. The data represent standards set by producers with modern plants, e.g., Hüls. Higher alkylphenols are generally mixtures of isomers, due to the varying composition and chemical structure of the higher olefins employed for their production. The alkyl groups (C_8, C_9, C_{12}) are bonded via a tertiary (or, less frequently, secondary) carbon atom to the aromatic ring. In addition, the alkyl chains are branched.

Table 11. Physical properties and specifications of commercial higher alkylphenols

Alkylating olefin	Product name	CAS registry no. of product	EINECS/ TSCA [h]	Molecular formula	*mp*/setting point, °C	*bp* (2.0 kPa), °C	Flash point, °C	Density (20 °C), g/mL	Viscosity, mPa · s 20 °C	Viscosity, mPa · s 80 °C	Color (state)
Diisobutene (DIB)	4-(1,1,3,3-tetramethyl-butyl) phenol (4-*tert*-octylphenol)	[140-66-9]	+/+	$C_{14}H_{22}O$	84 – 85	158		0.89 (90 °C)			white (s)
Diisobutene (DIB)	*tert*-octylphenol [a]	[27193-28-8]	+/+	$C_{14}H_{22}O$	80 (min)	162	ca. 147	0.901 (100 °C)		7 (100 °C)	colorless (l), white (s)
Di-*n*-butene (DNB)	isooctylphenol [b]	[11081-15-5]	+/–	$C_{14}H_{22}O$	ca. –10	163	ca. 152	0.966	1180	9	colorless (l)
Propene trimer (PTri)	nonylphenol [c]	[25154-52-3]	+/+	$C_{15}H_{24}O$	ca. –8	173	ca. 155	0.949	2500	12	colorless (l)
Propene trimer (PTri)	4-nonylphenol, branched [d]	[84852-15-3]	+/+	$C_{15}H_{24}O$	ca. –14	174	ca. 166	0.952	3160	14	colorless (l)
Propene trimer (PTri)	2,4-dinonylphenol [e]	[137-99-5]	+/+	$C_{24}H_{42}O$	ca. –13	224	ca. 180	0.912	3300		colorless – pale yellow (l)
Propene tetramer (PTet)	dodecylphenol, branched [f]	[121158-58-5]	+/–	$C_{18}H_{30}O$	ca. –7	187	ca. 165	0.944	12 000	25	colorless – pale yellow (l)
Tri-*n*-butene (TNB)	4-isododecylphenol [g]	[27459-10-5]	–/+	$C_{18}H_{30}O$	ca. –7	195	ca. 180	0.940	7000	22	colorless – pale yellow (l)

[a] With > 90 % 4-*tert*-octylphenol and ≤ 2 % 2-*tert*-octylphenol.
[b] With ≥ 90 % 4- and ≤ 10 % 2-isomer.
[c] With ca. 90 % 4- and ca. 10 % 2-isomer.
[d] With 98 % 4-nonylphenol.
[e] With 95 % (min) dinonylphenol.
[f] Physical data for TNB product, with < 10 % 2-isomer.
[g] Physical data for PTet product, with ca. 5 % 2-isomer.
[h] EINECS = European Inventory of Existing Chemical Substances; + = registered, – = not registered.

Production. Higher alkylphenols are obtained by alkylation of phenol with "higher olefins" (⟶ Hydrocarbons – Olefins). Linear or branched α-olefins as well as linear or branched olefins with internal double bonds can be used as alkylating agents. Industrially, the most important higher olefins are

1. diisobutene, a 4 : 1 mixture of 2,4,4-trimethyl-1-pentene [107-39-1], *bp* 101.4 °C, and 2,4,4-trimethyl-2-pentene [107-40-4], *bp* 104.5 °C;

2. di-*n*-butene [25377-83-7] (mixture of octenes) with boiling range 116 – 126 °C;
3. propene trimer [27215-95-8] (mixture of nonenes) with boiling range 134 – 143 °C;
4. propene tetramer [9003-07-0] (mixture of dodecenes) with usual boiling range 185 – 205 °C; and
5. tri-*n*-butene [72317-18-1] (mixture of dodecenes) with boiling range 197 – 203 °C

A review of phenol alkylations with higher olefins has been published [264]. The preferred catalysts are ion-exchange resins, acid-activated clays, and synthetic aluminosilicates [265]. Occasionally boron trifluoride, *p*-toluenesulfonic acid, or phosphoric acid [198] are employed.

5-tert-Octylphenol, produced from phenol and diisobutene, consists mainly of 4-(1,1,3,3-tetramethylbutyl)phenol; commercial products always contain small amounts of the ortho isomer. The ortho – para isomer ratio depends on the alkylation conditions.

Boron trifluoride or its complexes with phenol or ethers are recommended as catalysts for the alkylation of phenol (and the cresols) with diisobutene and with other higher olefins. A reaction temperature of 50 – 85 °C and 1 – 2 wt % of BF_3 with respect to the phenol is required. A phenol – diisobutene molar ratio of 1.2 – 1.4 : 1 yields 90 –95 % octylphenol in less than 30 min at 50 °C [266]. Catalysis with BF_3 is suitable for batchwise production. Processes in continuously operated stirred-tank and tube reactors have been also described [267], [268]. The BF_3 catalyst can be removed from the reaction product by extraction with hot water [267] or by precipitation with milk of lime or concentrated ammonia [269]. Various procedures for a (partial) recovery of BF_3 have been described [268], [270], [271]. If further reactions are to be carried out with the primary alkylate, the removal of BF_3 can sometimes be postponed to a later stage, for example, if the alkylate is reacted with sulfur dichloride to give thiobisphenols [272].

Sulfonated polystyrene – polydivinylbenzene ion-exchange resins of the macroporous type are currently the catalysts of choice. They are applied in the form of beads of 0.3 – 1.5 mm in diameter. Column-type reactors filled with the catalyst resin are employed. The reaction of higher olefins with phenol is slower than that of isobutene, permitting easier control of the heat of reaction and the temperature gradient along the catalyst bed. Sometimes the catalyst activity is deliberately reduced by special measures. Nonetheless, high space velocities can be maintained because isomerization reactions are less important with higher olefins than with isobutene. Alkylations with higher olefins can be carried out at or close to normal pressure because of their high boiling points. Plants are similar to those described for the production of 4-*tert*-butylphenol (two column-type reactors in series, connected by a heat exchanger allowing for an intermediate cooling of the primary reaction product leaving the first reactor). Depending on the olefin, the alkylations are run adiabatically or close to isothermal conditions.

In a process specially designed for the production of 4-octylphenol [273], phenol and diisobutene (1.5 : 1 molar ratio, LHSV = 1 h^{-1}) are passed over an ion-exchange resin that is placed in a single column reactor. The reaction temperature is maintained at 100 – 105 °C by internal cooling coils. The resin, with an exchange capacity of ca. 400 meq H^+/100 g, is kept in the hydrated form (water content 10 – 15 %) by using phenol with a water content of 1 – 2 wt %. The water decreases the catalyst activity to a desired level, thus spreading the evolution of reaction heat and effectively suppressing the depolymerization of diisobutene, and, consequently, the formation of 4-*tert*-butylphenol. The reaction is carried out to a diisobutene conversion of about 95 %. The reaction product is distilled to remove water, diisobutene, and phenol and has the following composition: 0.5 – 1 % 4-*tert*-butylphenol, 2 – 3 % 2-*tert*-octylphenol, 93 – 96 % 4-*tert*-octylphenol, and 2 – 3 % dialkylphenols. The product can be purified further by fractionation under vacuum. 4-*tert*-Octylphenol is marketed in the liquid state in thermostated containers or in the form of pastilles or flakes.

Hüls, a major producer of higher alkylphenols in Western Europe, operates plants with a total annual capacity of 50 000 t that produce *nonylphenol* (the major product; propene trimer as alkylating olefin), *octylphenols* (di-*n*-butene and diisobutene as alkylating olefins), and *dodecylphenol* (propene tetramer and tri-*n*-butene as alkylating olefins). The process uses two equally-sized reactors in series which are both operated adiabatically. The heat of reaction that is generated in the first reactor is partially removed by a heat exchanger connecting the reactors. Macroporous, anhydrous ion-exchange resins of different acidity are employed as catalysts. The exchange capacities are about 80 meq H^+/100 mL in the first, and about 140 meq H^+/100 mL in the second reactor. The nonylphenol process is carried out as follows [274]:

Process Description. A mixture of phenol and propene trimer (molar ratio 1.7 : 1) is preheated to 70 °C and pumped to the top of the first reactor at an hourly rate of about 30 mass units per 1 mass unit of catalyst. The product leaving the first reactor with a temperature of 120 °C is cooled to 100 °C and led to the second reactor. The temperature of the reaction product leaving the second reactor is about 125 °C. Product compositions after the first and second step are (in wt %): propene trimer 28.9 and 3.8, phenol 46.1 and 27.5, nonylphenol 22.4 and 65.8, and dinonylphenol 2.6 and 2.9. The product of the second reaction step is worked up by distillation. Recovered phenol and propene trimer, as well as dinonylphenol, are recycled. Nonylphenol thus can be obtained in > 95 % yield. In general, the product purity is > 98 % and the ratio of 4 – 2-nonylphenol is ca. 90 : 10. A nonylphenol with a purity of 99.7 % and a 4 – 2-nonylphenol ratio of at least 95 : 5, obtained by careful distillation, is also being commercialized by Hüls. Dinonylphenol is formed as a byproduct and is normally recycled to the reaction zone. It can be isolated from the nonylphenol bottoms by further distillation under vacuum in 50 – 98 % purity, depending on market requirements.

Other nonylphenol processes that are performed in stirred-tank reactors, with a suspension of a finely ground ion-exchange resin as the catalyst, are becoming obsolete.

The production of *dodecylphenol* from phenol and propene tetramer is carried out with a comparatively large excess of phenol; molar ratios of 3 : 1 (phenol – propene tetramer) are common. In this case the alkylation can be conducted adiabatically at moderate temperatures, rising from about 50 °C to about 100 °C along the catalyst bed. An ion-exchange resin with an exchange capacity of 200 meq H^+/100 g serves as catalyst. Distillation of the reactor effluent and recyclization of unconverted phenol gives product yields of 90 – 95 %, depending on the quality of the propene tetramer.

Comparable yields of dodecylphenol are obtained if activated clays are used as catalysts. The reaction times, however, are considerably longer (several hours) with clays than with ion-exchange resins or boron trifluoride. Batch [275] and continuous [276] processes have been described. The batch process, working at 80 °C, uses a finely ground catalyst, that is maintained in suspension by stirring; the continuous process, which is operated at about 150 °C, uses a granulated catalyst in a column-type reactor.

Uses. Higher alkylphenols, mainly 4-octyl-, 4-nonyl-, and 4-dodecylphenols are predominantly reacted with ethylene oxide to form the corresponding ethoxylates that find many applications as technical nonionic surfactants. Extensive use in household detergents is prohibited because of insufficient biodegradability. Ethoxylates of nonylphenol with 4 – 5 ethylene oxide units are oil-soluble, those with 8 – 9 ethylene oxide units are used as wetting agents, e.g., in textile scouring. The main application of ethoxylates of dodecylphenol, dioctyl-, and dinonylphenols is the field of technical emulsifiers. Sulfonated dodecylphenol ethoxylates have been suggested as foam-stabilizing surfactants in miscible-gas enhanced oil recovery. The alkali salts of the esters of sulfuric and phosphoric acid with nonylphenol ethoxylates are anionic surfactants, which are used in metalworking and as emulsifiers for insecticides and herbicides.

Condensates of higher alkylphenols with formaldehyde are the basis for oil-soluble resins and lacquers.

Higher alkylphenols are employed as starting materials for the synthesis of antioxidants (⟶ Antioxidants – Phosphorus Compounds); an important example is tris(4-nonylphenyl) phosphite [31631-13-7]. Derivatives of 4-octylphenol are used as light stabilizers. 4-Octyl-, 4-nonyl-, and 4-dodecylphenol, as well as their mixtures with the corresponding di- and trialkylphenols are used as age-protection additives. Dodecylphenol is the starting material for lubricating oil additives: dodecylphenol is first sulfurized to give a bisphenol disulfide, and subsequent carboxylation (Kolbe synthesis) yields the corresponding salicylic acids. The calcium and magnesium salts of these acids are widely used as dispersants and antioxidants in lubricating oils.

Economic Aspects. The annual worldwide capacity for higher alkylphenols is about 540 000 t (Western Europe 166 000 t, Eastern Europe 110 000 t, North America 233 000 t, South America 7000 t, Japan 24 000 t).

Toxicology and Occupational Health. Octyl-, nonyl-, and dodecylphenols are relatively nontoxic; LD_{50} values (oral, rat) are 2000 mg/kg or higher. Contact with skin and eyes must be avoided because irritation or even burns can occur.

Storage, Packaging, and Transportation. Octyl-, nonyl-, and dodecylphenols will discolor if oxygen is not excluded. Storage and transportation facilities are therefore blanketed with nitrogen. The phenols are stored at 50 – 70 °C to permit sufficient pumpability.

The products are shipped in stainless steel or rail tank cars, or in 200 L drums. Transport classifications for higher alkylphenols are given in Table 12.

Table 12. Classification of higher alkylphenols in land-, sea-, and air-transportation

Name	State	Land		Sea				Air		
		GGVS/GGVE, ADR/RID		GGVSee, IMDG code				ICAO code, IATA-DGR		
		Class	No.	Class	UN no.	Page	Packaging group	Class	UN no.	Packaging group
4-*tert*-Octylphenol										
[mainly 4-(1,1,3,3-tetramethylbutyl)phenol]	s	8	65 c	8	1759	8151	III	8	1759	III
Octyl-, nonyl-, dodecyl-phenols, and	l	6	66 b	8	1760	8147	II	8	1760	II
dinonylphenol (lower grade) [a]										
Dinonylphenol, conc. [b]	l									

[a] ca. 1 : 1 mixtures of nonyl- and dinonylphenols.
[b] Dinonylphenol concentration 95 % or higher.

1.5.8. *Cycloalkylphenols*

Properties. Physical properties of cyclopentyl- and cyclohexylphenols are listed in Table 13.

Table 13. Physical properties of cycloalkylphenols

Name	CAS registry no.	*mp*, °C	*bp*, (kPa) °C
2-Cyclopentylphenol	[1518-84-9]	40	152 (2.66)
4-Cyclopentylphenol	[1518-83-8]	64	166 (2.66)
2,6-Dicyclopentyl-4-methylphenol	[41505-40-2]	36 – 37	102 – 103 (0.003)
2-Cyclohexylphenol	[119-42-6]	55.7	169 (3.3)
			238 (101.3)
4-Cyclohexylphenol	[1131-60-8]	128	181 (3.3)
			295 (101.3)
2-Cyclohexyl-4-methylphenol	[1596-09-4]	56	172 (2.27)
			293 (101.3)
2-Cyclohexyl-6-methylphenol	[4855-68-9]	70.5	192 (6.67)
6-*tert*-Butyl-2-cyclohexyl-4-methylphenol	[51806-69-0]		183 (1.86)
			320 (101.3)
4-Methyl-2-(1-methylcyclohexyl)phenol	[16152-65-1]	60 – 61	155 (1.33)
2,4-Dimethyl-6-(1-methylcyclohexyl)phenol	[77-61-2]	28 – 29	179 (2.93)

Production.
Cyclopentylphenols are produced by reaction of phenols with cyclopentene. Monoalkylation in the position ortho to the phenolic hydroxyl group is achieved with γ-Al_2O_3 as the catalyst at about 300 °C [254], whereas aluminum phenolate at 200 °C leads to 2,6-disubstitution [197], [277]. Disubstitution at the 2,4-positions occurs with boron trifluoride at 80 °C, with *p*-toluenesulfonic acid at 100 – 200 °C, or with active clays at about 150 °C [277]. Extended reaction times lead to isomerizations. Thus, dialkylation of *p*-cresol will give a mixture of 4-methyl-2,5-, 4-methyl-2,6-, and 4-methyl-3,5-dicyclopentylphenols after 15 h on active clay at 150 °C [277] or after 20 h on ion-exchange resins at 80 °C [278].

Cyclohexylphenols can be obtained from cyclohexene and phenols, under the conditions described for cyclopentylphenols. In addition to the aforementioned catalysts, phosphoric acid can also be used [237]. However, for reasons of price and availability cyclohexanol rather than cyclohexane is predominantly employed as the alkylating reagent. With aluminum silicates as catalysts, *p*-cresol reacts with cyclohexanol to give 2-cyclohexyl-4-methylphenol, whereas with phenol a mixture of 2- and 4-cyclohexylphenols is formed, which can be separated by distillation. Likewise, 2-methylcyclohexanol forms 4-methyl-2-(1-methylcyclohexyl)phenol with *p*-cresol and a mixture of 2- and 4-(1-methylcyclohexyl) phenol, with phenol. Alkylation of 2,4-xylenol with 2-methylcyclohexanol yields 2,4-dimethyl-6-(1-methylcyclohexyl)phenol. Industrially, cycloalkylphenols are produced by Bayer and ICI.

Uses.
2-Cyclopentylphenol is a precursor for toner components in xerography. The *N*-methylcarbamate derived from 3-cyclopentylphenol is used as an insecticide. 2,6-Dicyclopentyl-4-methylphenol is reported to be an efficient antioxidant [279].

2- and 4-Cyclohexylphenols are intermediates in the synthesis of pharmaceuticals and agrochemicals. The most important application of cyclohexylphenols is in the field of rubber chemicals (⟶ Rubber, 4. Chemicals and Additives). Thus, 2,4-dimethyl-6-(1-methylcyclohexyl)phenol is an excellent antioxidant that is especially suitable for white or light-colored rubber products. The condensates of 2-cyclohexyl-4-methylphenol and of 4-methyl-2-(1-methylcyclohexyl)phenol with formaldehyde are well-known nonstaining antioxidants used in natural and synthetic rubbers and in thermoplastics such as polyolefins, ABS resins, and high-impact polystyrene.

1.5.9. *Aralkylphenols*

Properties. Physical properties of selected aralkylphenols, described in this article, are listed in Table 14.

Table 14. Physical data of aralkylphenols

Name	CAS registry no.	*mp*, °C	*bp*, °C
2-Benzylphenol	[28944-41-4]	53 – 54.5	312
4-Benzylphenol	[101-53-1]	83 – 85	198 – 200 (1.33 kPa)
p-Cumylphenol	[599-64-4]	70 – 73	335

Production. Mixtures of *2-* and *4-benzylphenols* (ratio about 3 : 2) are obtained by heating benzyl chloride in an excess of phenol to 150 °C. The product mixture is obtained in 50 % yield if a molar ratio phenol – benzyl chloride of 2 : 1 is used; with a ratio of 10 : 1 the yield is 90 % of the theoretical yield [280]. Dibenzylphenols are the main byproducts. In the presence of Friedel –Crafts catalysts, preferably zinc chloride, the reaction rate is enhanced and the temperature can be reduced to below 100 °C [281]. The reaction product can be distilled. Generally, a separation of single isomers is not carried out. Cresols and xylenols can be benzylated in the same way. It was found, that the 3- and 4-benzylated derivatives of 2,6-xylenol are formed in parallel, not consecutive reactions [282].

Benzylations of phenolic compounds can also be achieved with benzyl alcohol under acid catalysis, preferably with *p*-toluenesulfonic acid. This has been used for ortho benzylation of bisphenols, whose para-positions are substituted [283].

(α-Methylbenzyl)phenols (styrenated phenols) [61788-44-1] are produced on a relatively large scale by reaction of phenol with styrene in the presence of sulfuric or *p*-toluenesulfonic acid or activated clays [284]. Phosphoric acid [285] and oxalic acid are also suitable catalysts. Depending on the amount of styrene applied, mixtures of mono-, di-, and tristyrenated phenol are obtained. As an example, when 2.1 mol styrene are slowly added, with stirring, to 1 mol phenol in the presence of 0.002 mol *p*-toluenesulfonic acid and at 110 – 120 °C, a product is obtained with less than 1 % of unreacted phenol, and with mono-, di-, and tristyrenated phenol concentrations of ca. 12, 46, and 41 %, respectively. Styrenated phenol is produced by many manufacturers and traded as a colorless to pale yellow viscous oil. Styrenations, as described, can also be applied to the common alkylphenols, such as the cresols and xylenols.

The styrenated phenols are predominantly used as such. Sometimes the reaction mixtures are used as intermediates. For example, the reaction product from *p*-cresol and styrene is first distilled under vacuum to remove unreacted starting materials, and the residue, mainly 4-methyl-2-(α-methylbenzyl)phenol, is condensed with aldehydes to give the corresponding bisphenols. The alkylation of partially styrenated phenol with isobutene or diisobutene is another example. After an appropriate workup, which includes neutralization or filtration of the catalyst and distillation of low-boilers, the product can be used as an antioxidant.

If α-methylstyrene is used instead of styrene, α, α-dimethylbenzyl groups (cumyl groups) are introduced into the phenol moiety [286], [287]. The catalyst of choice is an acidic ion-exchange resin arranged as a fixed bed in a column reactor as described for the production of the higher alkylphenols. Very pure *p*-cumylphenol can be obtained by this method, which is required for certain applications in the polymer field.

The alkylation of phenols with styrene, α-methylstyrene, and indene can also be performed with aluminum phenolate as catalyst [178], [288]. Substitution of phenol or *o*-cresol with styrene takes place predominantly in the ortho-position, whereas with α-methylstyrene, mainly para substitution is found.

p-Cumylphenol is formed to a certain extent during acid-catalyzed cleavage of cumene hydroperoxide (⟶ Phenol – Cumene Oxidation (Hock Process)). The high boiling still bottoms of phenol – acetone plants are a potential source for this aralkylphenol, which is, however, not exploited.

Uses. *2-Benzylphenol* and the mixtures of *2-* and *4-benzylphenol* have a certain commercial significance as disinfectants.

(α-Methylbenzyl)phenols (styrenated phenols) find extensive use as protective agents against thermal and oxidative aging of natural rubber and synthetic styrene – butadiene rubbers and latices, especially for application in white and light-colored articles.

4-(α, α-Dimethylbenzyl)phenol (*p*-cumylphenol) can be condensed with formaldehyde to give acid- and base-resistant resins. *p*-Cumylphenol of high purity, as supplied, for

instance, by Hüls, serves as a molecular-mass controlling component in the production of polycarbonates.

1.5.10. *Alkenylphenols*

Production.
4-Vinylphenol [2628-17-3] (*mp* 73.5 °C, sublimation range 70 – 80 °C at 0.004 kPa) can be obtained by gas-phase dehydrogenation of 4-ethylphenol at 550 – 600 °C and 0.1 MPa on an iron oxide catalyst and in the presence of steam and an aromatic hydrocarbon as diluents [289]. The alkenylphenol, which polymerizes readily, can be separated from the less acidic unconverted 4-ethylphenol by extracting the condensed reactor effluent with alkali at 30 °C [290]. The Maruzen Oil Co. is a supplier of 4-vinylphenol.

4-Isopropenylphenol [4286-23-1] and other 4-(alken-2-yl)phenols are obtained in yields up to 95 % of the theoretical by alkali-catalyzed cleavage of the corresponding bis(4-hydroxyphenyl)-alkanes (bisphenols) [291], [292]. As an example, bisphenol A is cleaved by heating to 190 –240 °C at a pressure of ≤ 5 kPa in the presence of 0.1 wt % sodium hydroxide. This reaction is carried out on a small technical scale by Mitsui Toatsu. The resulting phenol – 4-isopropenylphenol mixture is removed from the resinous material by distillation, followed by separation of phenol by rectification. The residual material is a fairly stable mixture of oligomers of 4-isopropenylphenol [64054-77-9], consisting of ≥ 80 % of the linear dimers (**7**) and (**8**), with dimer (**7**) prevailing [293].

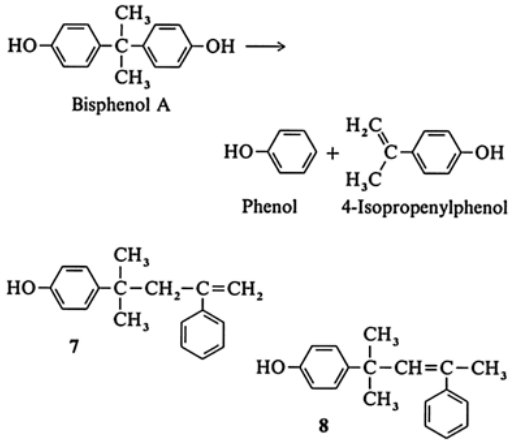

The addition of radical scavengers does not prevent the formation of dimers and oligomers, which show the reactive behavior of the underlying monomer in many consecutive reactions. When the oligomers are heated to 150 – 200 °C under slightly reduced pressure and the evolving vapors are absorbed in high-boiling alcohols (*n*-octanol or ethylene glycol), a solution of monomeric 4-isopropenylphenol with remarkable stability is obtained [294].

Alkylations of phenol with conjugated dienes take place primarily as 1,4-additions. However, these reactions are rarely carried out on an industrial scale. At low temperatures (20 – 25 °C), and with phosphoric or *p*-toluenesulfonic acid as catalyst, 1,3-butadiene and isoprene react to give the expected alk-3-en-2-ylphenols. These tend to undergo cyclization reactions: ortho-alkenylphenols cyclize to form chromans and para-alkenylphenols form indanols. A concise review on phenol alkylation with conjugated dienes has been published [295]. A modified aluminum phenolate catalyst has been described for the reaction of various phenols with 1,3-pentadiene to form the expected pent-3-en-2-ylphenols [296].

"Cardanol" [37330-39-5], also known as "anarcardol", is a mixture of meta-C_{15}-alkenylphenols, and is obtained by decarboxylation of anarcardic acid, which is the main constituent of cashew nut-shell oil. The alkenylphenols have been identified as 3-(penta-dec-8-en-1-yl)-, 3-(pentadeca-8,11-dien-1-yl)-, and 3-(pentadeca-8,11,14-trien-1-yl)phenols [297].

Uses. Mixtures of *4-isopropenylphenol oligomers* with phenol(s) and self-drying oils are used in the lamination of paper [293]. A hydroquinone synthesis has been suggested involving the addition of hydrogen peroxide to the double bond of 4-isopropenylphenol, followed by acid-catalyzed cleavage of the 4-hydroxycumene hydroperoxide.

4-Isopropenylphenol serves as the starting material for the synthesis of indanols (see below). *Cardanol* is a component of special phenol – formaldehyde resins, which have self-drying properties because of the unsaturated character of the C_{15}-substituent.

1.5.11. *Indanols*

5-Indanol [1470-94-6], *mp* 54 – 55 °C, *bp* 255 °C, is produced from indan [496-11-7] by a sulfonation – alkali-fusion process [298]. A synthesis for 1,1,3,3-tetraalkylindanols was developed by Bayer [299-301], which involves the reaction of isoolefins with secondary alkenylphenols (or compounds generating them under the reaction conditions) at 100 – 250 °C and in the presence of acidic catalysts. An example is the reaction of isobutene with bisphenol A in the presence of acid-activated clays which yields 1,1,3,3-tetramethyl-5-indanol [53718-26-6], *mp* 115 °C, *bp* (2.66 kPa) 156 °C, in a yield of 80 % of the theoretical, and 4-*tert*-butylphenol as a coproduct.

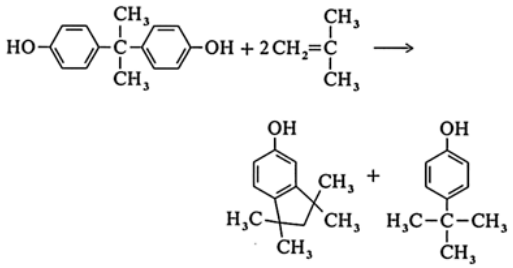

Uses. 5-Indanol is a precursor for the synthesis of drugs. 1,1,3,3-Tetramethyl-5-indanol and other alkylindanols are used as photochemicals [302].

[Top of Page]

2. Catechol

Toshikazu Hamamoto and Sumio Umemura

Catechol [120-80-9], (pyrocatechol, 1,2-dihydroxybenzene, 1,2-benzenediol), $C_6H_4(OH)_2$, M_r 110.11 is a crystalline compound with a phenolic odor. It was first obtained in 1839 by REINSCH by dry distillation of catechin. Currently, it is produced industrially from phenol. It is mainly used as raw material for the synthesis of polymerization inhibitors, perfumes, drugs, pesticides, dyes, in fur dyeing and leather tanning, as well as in photographic developers, deoxygenating agents, and analytical reagents.

Occurrence. Catechol has been found in, e.g., crude beet sugar, crude wood tar, some species of eucalyptus, onion, and coal. Catechol and many of its derivatives are obtained by dry distillation and other processes from, e.g., tannin, lignin, woods, and bituminous coal. It also occurs in tobacco smoke and is present in human and horse urine in the form of its sulfuric ester.

2.1. Physical Properties

Catechol forms colorless monoclinic crystals. It sublimes, is steam volatile, and discolors in air or light. It readily dissolves in cold and warm water and in cold and warm hydrophilic organic solvents (e.g., ethanol and acetone). Catechol is readily soluble in warm, but only sparingly soluble in cold hydrophobic solvents (e.g., benzene and chloroform). Physical properties of catechol may be summarized as follows [303]:

mp	104 – 105 °C
bp (at 0.1 MPa)	246 °C

d_4^{15}	1.37
d_4^{121} (fusion)	1.15
Solubility in water	
at 15 °C	25 g/100 g solution
at 20 °C	31.2 g/100 g solution
at 100 °C	98.2 g/100 g solution
Specific heat	
at 25 °C	0.14 kJ mol^{-1} K^{-1}
at 104.3 °C (fusion)	0.24 kJ mol^{-1} K^{-1}
Heat of fusion	22.78 kJ/mol
Heat of sublimation at 36 °C	80.81 kJ/mol
Heat of vaporization	103.6 kJ/mol
Heat of formation	1.2 kJ/mol
Flash point (closed cup)	127 °C
Dissociation constant	
K_1 (18 °C)	3.3×10^{-10}
K_1 (30 °C)	7.5×10^{-10}
K_2 (30 °C)	8.4×10^{-13}
Vapor pressure	
168 °C	10 kPa
188 °C	20 kPa
211 °C	40 kPa
235 °C	80 kPa

2.2. Chemical Properties

Catechol gives a green coloration with iron-(III) chloride and turns red upon addition of a small amount of sodium hydroxide or ammonia. These colorations are specific for catechol and can be used for its detection and identification. Catechol can form stable coordination compounds with almost all metals and is therefore used as an analytical reagent for metals. Catechol is a weak acid and forms mono- and disalts with alkali hydroxides or carbonates. Many heavy metal salts of catechol, in particular its lead salts, are virtually insoluble in water. Thus, the reaction of catechol with lead acetate is used for quantitative analysis and separation of catechol from its isomers hydroquinone and resorcinol. Catechol is the strongest reducing agent of the three benzenediol isomers and can react with a solution of heavy metal salts to form fine precipitates of the elemental metals. 1,2-Benzoquinone is formed by careful oxidation of catechol with silver oxide or silver carbonate on celite. *cis,cis*-Muconic acid monomethyl ester [61186-96-7] can be obtained by oxidative cleavage of catechol with oxygen in the presence of copper(I) chloride and methanol [304].

Catechol undergoes many typical reactions of phenol. It reacts with acyl halides to form the corresponding mono- and diesters, which are converted into phenolic ketones by Fries rearrangement with aluminum chloride as catalyst. An aldehyde group can be introduced into the aromatic nucleus by Reimer – Tiemann reaction with chloroform and alkali, or by the addition of glyoxylic acid and subsequent oxidative decarboxylation.

Mono- and diethers of catechol can be prepared by the usual methods. Catechol can undergo cyclization reactions because of its two adjacent hydroxyl groups. It reacts with dichloromethane to form methylenedioxybenzene and with bis(2-chloroethyl)ether to yield dibenzo-18-crown-6-polyether [14605-55-1] [305]. Ammonolysis of catechol gives 2-aminophenol. Catechol couples with aryldiazonium salts to form azo compounds, which can be reduced to 4-aminocatechol.

Catechol forms bis(dihydroxyphenyl)methane by condensation reaction with formaldehyde, and polycyclic compounds (e.g., hystazarin and alizarin) by condensation with phthalic anhydride.

Ring mono-substitutions occur at the 3- and 4-positions. Catechol can thus be alkylated, halogenated, nitrated, carboxylated, and sulfonated.

Analysis. Catechol can be determined directly in mixtures containing phenol and other benzenediols by gas chromatography without prior separation [306].

2.3. Production

Catechol was formerly produced by low-temperature carbonization of coal, but this process is currently carried out only in rare cases.

Hydrolysis of 2-Chlorophenol.

Industrially, catechol can be produced by hydrolysis of 2-chlorophenol with an aqueous solution of barium hydroxide and sodium hydroxide. Barium is recovered as carbonate and recycled to the process by converting the carbonate into hydroxide, which complicates the process. Therefore, improved processes in which only caustic alkali is used have been developed. For example, 1 mol 2-chlorophenol is reacted with 2.3 mol 4 – 8 % sodium hydroxide solution in the presence of a copper(II) sulfate or copper(I) oxide at 190 °C for 3 h in a copper autoclave. The conversion of 2-chlorophenol is 96 – 99 %, the selectivity for catechol is 81 – 86 % [307]. The reaction mixture is then neutralized with sulfuric acid and the crude catechol formed is extracted, the solvent is recovered, and catechol is separated and purified by distillation. The industrial production of catechol was carried out by this method up to 1973.

Hydroxylation of Phenol. Catechol is currently manufactured together with hydroquinone by direct hydroxylation of phenol with peroxides. At present, three plants using this process are in operation worldwide; each plant uses a somewhat different type of peroxide or catalyst. As the reaction is exothermic and the benzenediols formed are more easily oxidized than phenol, the reaction is carried out in a large excess of phenol.

In France, *Rhône-Poulenc* react phenol with 70 % hydrogen peroxide (molar ratio 20 : 1) in the presence of phosphoric acid and catalytic amounts of perchloric acid at 90 °C; catechol and hydroquinone are obtained in a ratio of ca. 3 : 2 [308]. This reaction proceeds electrophilically and phosphoric acid serves as a masking reagent preventing side reactions (formation of resorcinol by a radical reaction which gives a lower yield) caused by trace amounts of metallic ions. After the reaction, phosphoric and perchloric acids are removed by washing with water, then the reaction mixture is simultaneously extracted by diisopropyl ether, distilled, and continuously separated.

Brichima SpA (now Enichem) in Italy uses heavy metal compounds (e.g., small quantities of ferrocene and/or cobalt salts) as catalyst and reacts phenol with 60 % aqueous hydrogen peroxide at 40 °C. Catechol and hydroquinone are produced in the ratio of 1.5 – 4.1 [309]. This reaction occurs via a free-radical chain mechanism and is thus very fast.

In Japan, *Ube Industries* produces catechol together with hydroquinone by hydroxylation of phenol with ketone peroxides formed in situ from a ketone and hydrogen peroxide in the presence of an acid catalyst [310], [311]. The process is carried out by adding a trace amount of acid (e.g., sulfuric or sulfonic acid), a small volume of ketone, and 60 % aqueous hydrogen peroxide to phenol at 70 °C. The ketone peroxide that is formed in situ reacts rapidly and electrophilically with phenol, and catechol and hydroquinone are obtained in a molar ratio of about 3 : 2 in more than 90 % yield (based on phenol reacted). When a solid acid such as clay is used as a catalyst, the molar ratio of catechol and hydroquinone is about 1 : 1 [312]. As only a small amount of catalyst is used, no corrosion occurs and the reaction mixture can be distilled without removing the catalyst after the reaction. The added ketone can be recycled to the process by recovering it by distillation.

The *separation and purification methods* used in the three aforementioned processes are basically the same. That is, the reaction mixture is separated by distillation in different distillation columns. Water is removed, low-boiling fractions (solvents, ketone etc.) and unreacted phenol are recovered and recycled, the catechol fraction is made into a flaked product, and hydroquinone is purified by recrystallization of the corresponding fraction from water.

Processes, in which hydroxylation of phenol is carried out in the presence of catalysts such as strong acids or sulfur dioxide with nonaqueous hydrogen peroxide solution, have been patented [313], [314].

Peracids formed in situ from carboxylic acids and hydrogen peroxide or synthesized by air oxidation of aldehydes are utilized as hydroxylating agents [315]. In the case of hydroxylation using peracetic acid (synthesized through autoxidation of acetaldehyde), chelating reagents and phosphates are used as additives which improve the yield of benzenediols. Nevertheless, because of problems such as complicated separation of products and corrosion of apparatus, the peracid processes have not been realized industrially.

Dehydrogenation of 1,2-Cyclohexanediol. Recently, a patent has been published claiming that catechol can be obtained in 90 % yield by dehydrogenation of 1,2-cyclohexanediol with a Pd/Te catalyst system at 300 °C [316]. Catechol is reportedly formed as the single product.

In addition, catechol may be obtained via the following synthetic routes: (1) alkali fusion of 2-phenolsulfonic acid and phenol-2,4-disulfonic acid; (2) oxidation of salicylaldehyde with

hydrogen peroxide in aqueous alkaline solution; (3) demethylation of guaiacol with hydrobromic acid or aluminum chloride; and (4) hydrolysis of 2-aminophenol using a hydrogen halide.

2.4. Uses

Catechol itself is used as a photographic developer, analytical reagent, and oxygen scavenger (antioxidant); most catechol is used in the form of its derivatives: guaiacol (2-methoxyphenol) [90-05-1] and veratrole (1,2-dimethoxybenzene) [91-16-7] can be synthesized by *O*-methylation of catechol. Vanillin (4-hydroxy-3-methoxybenzaldehyde) [121-33-5], derived from guaiacol is used as a flavoring agent. Ethylvanillin(3-ethoxy-4-hydroxybenzaldehyde) [121-32-4] derived from guethol (2-ethoxyphenol, a homologue of guaiacol) is not found in nature, and its flavor is 3 – 4 times stronger than that of vanillin. It is thus a valuable aroma compound. Eugenol (2-methoxy-4-allylphenol) [97-53-0], safrole (5-allyl-1,3-benzodioxole) [94-59-7], and piperonal [3,4-(methylenedioxy)benzaldehyde] [120-57-0], which can be prepared from the latter, are useful fragrances in perfumery.

In medicine, potassium guaiacol sulfonate (4-hydroxy-3-methoxy-benzenesulfonic acid monopotassium salt) [16241-25-1] and guaiacol glyceryl ether [3-(2-methoxyphenoxy)-1,2-propanediol] [93-14-1] are used as expectorants. L- α-Methyldopa (3-hydroxy- α-methyl-L-tyrosine) [555-30-6] and L-dopa (3-hydroxy-L-tyrosine) [59-92-7] derived from vanillin are used as an antihypertensive and as an antiparkinsonism drug, respectively. In addition, trimethoprim [738-70-5] derived from vanillin is used as an antiinfective; carbazochrome [69-81-8] and carbazochrome sodium sulfonate [51460-26-5] are used as hemostatics; papaverine [58-74-2] is used as an antispasmodic, vasodilator drug and smooth muscle relaxant.

Two important carbamate insecticides that are produced from catechol are used as agricultural chemicals — carbofuran [1563-66-2] (2,3-dihydro-2,2-dimethyl-7-benzofuranyl methylcarbamate, trade name Furadan), and propoxur [114-26-1] (2-isopropoxyphenyl-*N*-methylcarbamate, trade name Baygon) developed by Bayer.

4-*tert*-Butylcatechol [98-29-3], which is prepared by ring alkylation of catechol, is applied as a polymerization inhibitor during manufacturing and storage of monomers such as styrene and butadiene.

2.5. Economic Aspects

The consumption of catechol in the world is estimated to be about 20 000 t/a in 1990. It is assumed that 50 % of that is used as starting material for insecticides, 35 – 40 % for perfumery and drugs, and 10 – 15 % for polymerization inhibitors and others. As the vanillin process is changing from production using lignin in pulp waste liquor to the synthesis from catechol, the demand for catechol used in the production of vanillin is expected to increase in the future [317]. The market price of catechol in May 1990 was 4.2 – 5.7 $/kg [318].

2.6. Toxicology

Catechol is readily absorbed from the gastrointestinal tract and through the skin and readily metabolized and excreted into urine. General symptoms of catechol intoxication are similar to those of phenol intoxication. Administration of a large dose of catechol induces strong suppression of the central nervous system and long-term vascular hypertension. The prolonged contact of catechol with skin leads to eczematous dermatitis or ulcers. In addition, it causes severe eye irritation.

Comprehensive references exist on catechol toxicity [319]. Some characteristic values of catechol toxicity are as follows [320]:

LD_{50} (oral, rat)	3890 mg/kg
LDLo (subcutaneous, rat)	200 mg/kg
LD_{50} (subcutaneous, mouse)	247 mg/kg
Threshold limit value	5 ppm

[Top of Page]

3. Trihydroxybenzenes

Tadao Iwata, Hisaya Miki, Yasuhiro Fujita

3.1. Pyrogallol

Pyrogallol [87-66-1], $C_6H_3(OH)_3$, M_r 126.11 is generally called 1,2,3-trihydroxybenzene, 1,2,3-benzenetriol (IUPAC name), or pyrogallic acid. Pyrogallol was first isolated by SCHEELE by dry distillation of gallic acid (3,4,5-trihydroxybenzoic acid) in 1786. Pyrogallol derivatives occur in many natural products, such as tannin, anthocyanin, and alkaloids [321].

Physical Properties. Pyrogallol forms colorless needles or leaflets, the color gradually changes to dark gray on contact with air or light. It is soluble in water and polar solvents, and slightly soluble in chloroform and carbon disulfide. It sublimes without decomposition upon gradual heating. Important physical data are as follows:

mp	133 – 134 °C
bp (101.3 kPa)	309 °C (partial decomp.)
(17 kPa)	171 °C
d_4^4	1.453
Dissociation constant (20 °C)	
K_1	1.5×10^{-10}
K_2	4.5×10^{-12}
Solubility (25 °C)	
water	38 wt %
ethanol	50 wt %
pyridine	soluble
ether	45 wt %
benzene	slightly soluble
chloroform	slightly soluble
carbon disulfide	slightly soluble

Chemical Properties. Pyrogallol is the strongest reducing agent among the polyhydroxybenzenes. The basic aqueous solution quickly absorbs gaseous oxygen to form a dark brown precipitate. Because of this property, pyrogallol is often used for quantitative analysis of oxygen.

Reactions of pyrogallol are quite similar to those of phenols. The hydroxy groups are etherified and esterified by the conventional methods to give mono-, di-, and trisubstituted products. Heating of pyrogallol in aqueous potassium bicarbonate solution mainly gives pyrogallol-4-carboxylic acid (2,3,4-trihydroxybenzoic acid); gallic acid is formed as a byproduct. Formylation, acylation, and the Mannich reaction produce 4-substituted cyclohexenones as a major product. Bromination with bromine eventually yields 1,2,6,6-tetrabromocyclohexene-3,4,5-trione. Reaction with phosgene gives pyrogallol carbonate, and with thionylbromide 4,5,6-tribromopyrogallol. The highly sensitive color reactions of aqueous pyrogallol solution with heavy metal ions are useful for the quantitative analysis of metals. Precious metal ions (e.g., Ag^+, Au^{2+}, Hg^{2+}) are reduced to elemental metals.

A complex prepared from diethylzinc and pyrogallol is used for a catalyst for alternating polymerization of carbon dioxide and propylene oxide [322].

Production. Pyrogallol is commercially produced by decarboxylation of gallic acid, which is prepared by hydrolysis of tannin.

The reaction is carried out batchwise. A 50 % aqueous solution of gallic acid is heated to 175 – 200 °C in an autoclave. During the reaction, the pressure increases to 1.2 MPa. After the evolution of carbon dioxide is completed, the solution is cooled. The reaction proceeds quantitatively. Decolorization of the reaction mixture with charcoal followed by distillation gives a crude product which is purified by sublimation or vacuum distillation [323].

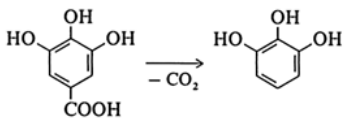

Pyrogallol is expensive to obtain because the starting material gallic acid is prepared from a natural product of limited availability. Therefore, new methods for producing gallic acid have been sought. Base-catalyzed condensation of trimethyl propane-1,2,3-tricarboxylate with the dimethyl ketal of mesoxalic acid, followed by hydrolysis and decarboxylation gives gallic acid in 74 % yield [324]. The reaction of the ketal and glutaric ester gives pyrogallol without generation of gallic acid [325].

Other methods for producing pyrogallol include oxidation of resorcinol with hydrogen peroxide [326], hydrolysis of 2,6-diamino-4-butylphenol [327], demethylation of 4-substituted 2,6-dimethoxyphenols [328], oxidation of 2,6-dimethylphenol [329], hydrolysis of 2,2,6,6-tetrachlorocyclohexanone [330], deoximation of 1,2,3-cyclohexanetrion-1,3-dioxime [331], and dehydrogenation of 1,2,3-trihydroxycyclohexane [332].

Uses. Pyrogallol is used in photography, lithography, and in hair dyes. It is also used as an antioxidant and stabilizer. The use of pyrogallol in the field of cosmetics and medicines is currently declining because of its pronounced toxicity.

Toxicology. Ingestion of pyrogallol may cause severe irritation. Toxicity data are as follows [333]:

LD_{50} (oral, mouse)	300 mg/kg
LD_{50} (subcutaneous, mouse)	566 mg/kg

3.2. Hydroxyhydroquinone

Hydroxyhydroquinone [533-73-3], $C_6H_3(OH)_3$, M_r 126.11 is usually called 1,2,4-trihydroxybenzene, 1,2,4-benzenetriol (IUPAC name), or hydroxyquinol. Its derivatives are found as ethers or quinoids in plants.

Physical Properties. Hydroxyhydroquinone forms colorless platelets or prisms that quickly colorize in air. It is soluble in water and polar solvents, and slightly soluble in chloroform and carbon disulfide. Important physical data are as follows:

mp	140.5 – 141 °C
Dissociation constant (20 °C)	
K_1	8.3×10^{-10}
K_2	1.5×10^{-12}
Solubility	
water	soluble
ethanol	soluble
pyridine	soluble
ether	soluble
benzene	slightly soluble
chloroform	slightly soluble
carbon disulfide	slightly soluble

Chemical Properties. Hydroxyhydroquinone is, like other phenols, a strong reducing agent. Its basic aqueous solution absorbs gaseous oxygen to produce black precipitates of the humic acid type. Hydroxyhydroquinone shows typical reactions of phenols. Derivatives of the tautomeric keto form are also known: halogenation of hydroxyhydroquinone, for example, gives 1,2,5,5-tetrahalocyclohexene-3,4,6-trione. Condensation with ethylacetoacetate produces dihydroxycoumarin; reaction with phthalic anhydride yields hydroxyhydroquinonephthalein. Monosubstitution of a hydroxy group by an amino group occurs easily at room temperature and gives 2,4-dihydroxyaniline [334].

Production. Hydroxyhydroquinone is synthesized by hydrolysis of 1,2,4-triacetoxybenzene, which is prepared by the acid-catalyzed reaction of *p*-benzoquinone with acetic anhydride [335].

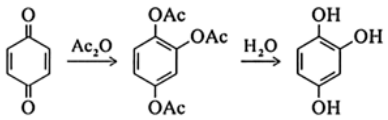

Other production methods for hydroxyhydroquinone are oxidation of resorcinol with hydrogen peroxide [336] and Dakin oxidation of 2,4- or 3,4-dihydroxybenzaldehydes or 2,4- or 3,4-dihydroxyacetophenones with alkaline hydrogen peroxide solution.

Uses. Hydroxyhydroquinone is used as a stabilizer, antioxidant, polymerization inhibitor and in hair dyes.

Toxicology. Toxicity data are as follows [333]:

LDLo (intraperitoneal, mouse)	125 mg/kg
LD_{50} (subcutaneous, mouse)	500 mg/kg

3.3. Phloroglucinol

Phloroglucinol [108-73-6], $C_6H_3(OH)_3$, M_r 126.11 is also called phloroglucine, 1,3,5-trihydroxybenzene, 1,3,5-benzenetriol (IUPAC name), or cyclohexane-1,3,5-trione. In 1855, HLASIWETZ found phloroglucinol in the hydrolysis products of phloretin, which was obtained from the bark of fruit trees. The name phloroglucinol originates from the Greek and means “sweet bark”. Its derivatives (e.g., flavones, anthocyanides, and xanthins) are widely distributed in the plant kingdom.

Physical Properties. Phloroglucinol crystallizes in the anhydrous form or as the dihydrate. Recrystallization from water gives the dihydrate, which is colorless, oderless, sweet in taste, and forms rhombic crystals. The dihydrate is converted into the anhydrous form by heating at 110 °C. Phloroglucinol sublimes at higher temperatures with partial decomposition. Important physical data are as follows:

mp (anhydrous form)	218 – 220 °C
mp (dihydrate)	116 – 117 °C
Dissociation constant (30 °C)	
K_1	3.56×10^{-9}
K_2	1.32×10^{-9}
Solubility	
water	1 wt %
ethanol	9 wt %
pyridine	75 wt %
diethyl ether	soluble
benzene	slightly soluble

In conductometric titration, the behavior of phloroglucinol resembles that of benzenediols. The third proton is not dissociated. According to ^{1}H NMR, the dianion of phloroglucinol forms stable keto-type tautomers in solution. Phloroglucinol itself, however, is present in the enolic form, which has been confirmed by UV, IR, NMR, and X-ray spectra [337].

Chemical Properties. Because of the keto – enol tautomerism, phloroglucinol reacts to form triphenol derivatives and triketone derivatives.

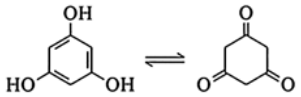

Phloroglucinol behaves as a triketone, e.g., towards hydroxylamine to form the trioxime, and adds sodium bisulfite to yield the corresponding mono-, di-, and tri-substituted products. Triphenol behavior is shown in etherification and esterification that form mono-, di-, and trisubstituted phenols. Friedel – Crafts reaction and nitration also take place.

Phloroglucinol is also a reducing agent. The alkaline aqueous solution absorbs gaseous oxygen. The absorption of oxygen by phloroglucinol proceeds more slowly than by pyrogallol. Phloroglucinol reduces Fehling's solution and precious metal ions (e.g., Au^{2+}, Ag^{+}, and Pt^{2+}). Reaction with methyl iodide causes ring methylation under basic conditions. Reaction with aqueous ammonia produces 5-aminoresorcinol or 3,5-diaminophenol. Phloroglucinol forms phloroglucide upon heating. Hydrogenation gives 1,3,5-trihydroxycyclohexane. Coupling with diazonium salts takes place easily.

Production. Phloroglucinol can be prepared by various processes. 2,4,6-trinitrotoluene is oxidized with sodium dichromate to give 2,4,6-trinitrobenzoic acid, which is subsequently decarboxylated. The nitro groups are then reduced with iron – hydrochloric acid to form 1,3,5-triaminobenzene. Subsequent nucleophilic substitution of the amino groups by hydroxyl groups produces phloroglucinol. However, waste-disposal of the acidic filtrates which contain chromium and iron poses a problem [338].

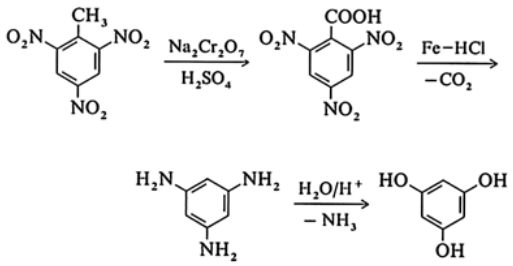

In the 1980s, a process was developed to produce phloroglucinol from 1,3,5-triisopropylbenzene via the corresponding trihydroperoxide [339]. The yield is increased by addition of hydrogen peroxide to the reaction mixture in the acidolysis step [340]. Another method for producing phloroglucinol is the Beckmann rearrangement of triacetylbenzene trioxime followed by hydrolysis [322].

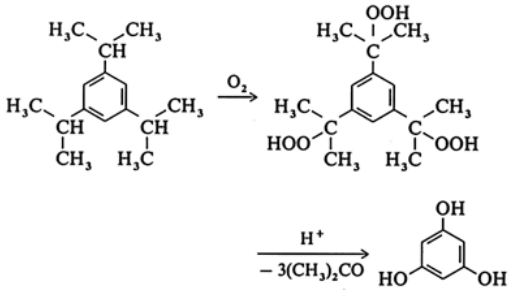

Other methods include Hofmann rearrangement of benzene-1,3,5-tricarboxylic acid triamide [342]; reaction of 1,3,5-tribromobenzene [343] or hexachlorobenzene [344] with alkoxide followed by acidolysis; and hydrolysis of 4-chlororesorcinol with potassium hydroxide [345].

Uses. Phloroglucinol is used as a coupler in diazotyping. Phloroglucinol and the diazo compound react with each other to form a large molecule which has a black color. In the drying process, the diazo and hydroxy compounds are both contained in the coating (⟶ Imaging Technology – Processing). Phloroglucinol is also used for photocopying [338].

Toxicology. Toxicity data are as follows [333]:

LD_{50} (intraperitoneal, mouse)	4050 mg/kg
LD_{50} (oral, mouse)	4550 mg/kg
LD_{50} (subcutaneous, mouse)	991 mg/kg

[Top of Page]

4. Bisphenols (Bishydroxyarylalkanes)

Hans-Josef Buysch

In this review bisphenols are defined as compounds consisting of two phenol nuclei linked by a hydrocarbon bridge.

History. The importance of bisphenols has increased continually with the growth of the plastics industry from about the end of the 1930s. They are used as intermediates for high-grade thermosets, thermoplastics, and raw materials for varnishes; and as antioxidants for rubber, plastics, oils, and fats. 2,2-Bis(4-hydroxyphenol)propane (bisphenol A), first described in 1891 by A. DIANIN, is by far the most important compound in this series. It has been produced industrially from phenol and acetone to be used as a starting material for epoxy resins since about 1945. The development of polycarbonates, the most important thermoplastics based on bisphenol A, began in the 1950s.

4.1. Physical Properties

Bisphenols are colorless, odorless substances, and most of them are solid at room temperature. The melting points of the most important bisphenols are between 100 and 200 °C [346]. Bisphenols are virtually insoluble in water. Their solubility in organic solvents is determined by their substituents. Whereas bisphenol A is only readily soluble in polar media such as ethers and alcohols, bisphenols with large aliphatic groups in the molecule are soluble in araliphatic and aliphatic hydrocarbons. The alkali salts of bisphenols are water-soluble. However, their solubility decreases drastically with increasing substitution. The boiling points of bisphenols are very high because of the size of the molecule and its polarity. For this reason and because of the decomposition frequently observed during boiling, bisphenols are rarely distilled. Some important physical properties of bisphenols are summarized in Table 15. Physical properties for the industrially important bisphenol A may be summarized as follows:

Density at 20 °C	1.04 g/cm^3
at 160 °C	1.065 g/m^3
Bulk density	0.492 g/cm^3
Heat of fusion	147 J/g
bp at 101.3 kPa	360 °C
at 1.4 kPa	240 °C
at 0.4 kPa	222 °C
Heat of varporization at 101.3 kPa	404 J/g
Flash point	227 °C
Ignition temperature	510 °C
Solubility in H_2O at 83 °C	0.344 wt %
Solubility in acetone, alcohol	good

Solubility in methylene chloride ca. 1 wt %

Table 15. Physical properties of some bisphenols

	CAS registry no.	Structural formula	Molecular formula	M_r	*mp*, °C	*bp*, °C
Bis-(4-hydroxyphenyl)methane	[*620-92-8*]	**9**	$C_{13}H_{12}O_2$	200.2	158	
2,2-Bis(4-hydroxyphenyl)propane (bisphenol A)	[*80-05-7*]	**10**	$C_{15}H_{16}O_2$	228.3	156–157	190 (130 Pa)
4,4-Bis(4-hydroxyphenyl)valeric acid	[*126-00-1*]	**11**	$C_{17}H_{18}O_4$	286.3	173	
2,2-Bis(4-hydroxy-3-methylphenyl)propane	[*79-97-0*]	**12**	$C_{17}H_{20}O_2$	256.3	136	
2,2-Bis(4-hydroxyphenyl)cyclohexane (bisphenol Z)	[*843-55-0*]	**13**	$C_{18}H_{20}O_2$	268.4	188	
2,2′-Methylene bis(6-*tert*-butyl-3-methylphenol)	[*119-47-1*]	**14**	$C_{23}H_{32}O_2$	340.5	131	
α,α′-Bis(4-hydroxyphenyl)-*p*-diisopropylbenzene	[*2167-51-3*]	**15**	$C_{24}H_{26}O_2$	346.3	196	230 (13 Pa)
Bis(2-hydroxy-3-cyclohexyl-5-methylphenyl) methane		**16**	$C_{27}H_{36}O_2$	392.6	128	
Bis(4-hydroxy-3,5-di-*tert*-butylphenyl)methane	[*118-82-1*]	**17**	$C_{29}H_{44}O_2$	424.7	154	

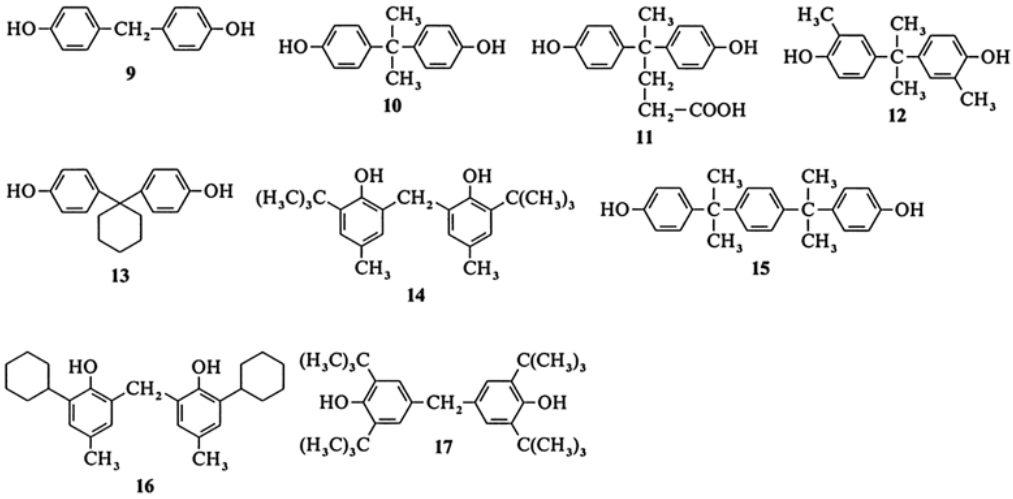

4.2. Chemical Properties

The chemical properties of the bisphenols are determined by the phenolic OH groups, the aromatic rings, and the alkyl bridge. They therefore undergo the same reactions as the corresponding substituted monophenols (esterification and etherification at the OH group, substitution and hydrogenation on the aromatic ring, see Section Chemical Properties). They are also suitable as building blocks for high molecular mass linear polyesters and polyethers because of their bifunctionality [346-348].

Bisphenols which are alkylated ortho to the OH group readily trap radicals and are therefore suitable as stabilizers (⟶ Antioxidants – Hindered Phenols).

Bisphenols fragment purely thermally, or better catalytically, according to:

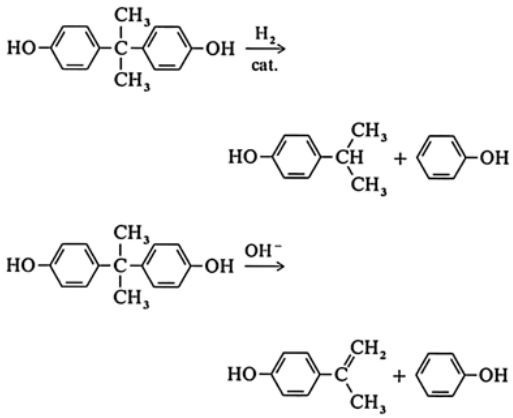

Under hydrogenation conditions bisphenol A is cleaved to give 4-isopropylphenol [349]; alkali-catalyzed cleavage gives 4-isopropenylphenol [350] in good yields. Both compounds are difficult to obtain by other methods. The alkali-catalyzed cleavage of various bisphenols has been investigated [351]. The cleavage can also be catalyzed by acids to form indans and spirobisindans [352]:

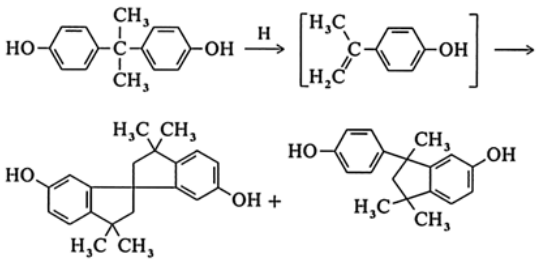

The purely thermal cleavage is generally less straightforward [350].

4.3. Production

Bisphenols can be produced (1) by condensation of ketones or aldehydes with phenols, (2) by reaction of, e.g., diols or dihalides with phenols, and (3) by condensation of hydroxymethylphenols with phenols.

Condensation of Ketones or Aldehydes with Phenols. Acid- or alkali-catalyzed condensation of 1 mol ketone or aldehyde with 2 mol phenol yields bisphenols, whose aromatic nuclei are separated by one carbon [353]. Acetone and cyclohexanone give particularly smooth para addition and high yields of 4,4′-dihydroxydiphenylalkanes [350]:

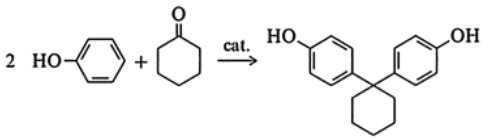

Aldehydes generally react with un- and monosubstituted phenols to give mixtures of isomers and polynuclear phenols of the novolac type (⟶ Phenolic Resins). 2,4-, 2,5-, 2,6-, 2,3,6-, and 2,3,5,6-substituted phenols, however, react readily and often in excellent yields with aldehydes (particularly formaldehyde) to form the corresponding bisphenols [354-358]. These reactions are frequently alkali-catalyzed. In other cases, however, alkali gives less pure products and poorer yields [350], [358]. Highly alkylated bisphenols are also obtained by subsequent alkylation of preformed bisphenols in the presence of acid catalysts [359]. Suitable acid catalysts are hydrogen chloride, highly concentrated hydrochloric acid, sulfuric acid (up to 75 %), boron trifluoride, acid-activated clays and sulfonated cross-linked polystyrenes [349], [350], [360-364].

Cocatalysts containing SH groups e.g., thiols and mercaptoacids, are used to increase the reaction rate [349], [350], [362], [365], [366].

Organic solvents such as aromatic and chlorinated hydrocarbons or glacial acetic acid serve as the reaction medium. Aqueous emulsions may also be used. To prevent side reactions, the phenol which is being reacted must frequently be added in excess and thus functions simultaneously as the reaction medium.

Reaction of Diols, Diolefins etc. with Phenols. Bisphenols can also be produced by the acid-catalyzed reaction of diols and dihalides [364], [367], [368] or alkynes or allenes [369] with phenols. Divinylbenzenes react with phenols at elevated temperature without catalysts [370].

Condensation of Hydroxymethylphenols with Phenols. Unsymmetrical dihydroxydiphenylmethanes are formed by condensation of hydroxymethylphenols (from phenol and formaldehyde) with phenols with a different substitution pattern [357]. The synthesis of unsymmetrical bisphenols substituted at the central carbon atom can be achieved by alkylation of phenols with vinylphenols [346], [371].

The workup of the bisphenol reaction mixture depends very much on the structure of the products and the purity requirements. Basic purification is not necessary for bisphenols used as stabilizers, whereas bisphenol A for polycondensations must be of high purity. Possible workup procedures are, for example, the simple distillation of volatile components from the reaction mixture, washing out starting materials, or recrystallization from suitable solvents. Distillations of bisphenols are generally not carried out because of their thermal lability.

The mechanism for the formation of bisphenols, in particular that for the formation of bisphenol A and its byproducts, is essentially understood [350].

Bisphenol A. There are many processes for the industrial synthesis of this most important bisphenol [362], [372-376]. Currently, hydrogen chloride or sulfonated cross-linked polystyrenes [372], [377-379] are used as the catalyst which are usually arranged as a fixed bed over which the reaction mixture is passed (Fig. 2). The reaction of phenol with acetone takes place at 50 – 90 °C, the molar ratio phenol – acetone is up to 15 : 1. Bisphenol A crystallizes as an adduct with 1 mol phenol, after separation of the hydrogen chloride by distillation or neutralization. The use of ion exchangers is preferred to that of hydrogen chloride because they are less corrosive. The yield is normally 80 –95 %.

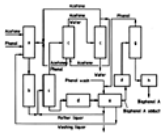

Figure 2. Flowchart of a bisphenol A plant according to [383]

a) Storage tank; b) Reactor; c) Column; d) Crystallizer; e) Separator; f) Melt; g) Desorber; h) Flaking off

To isolate the bisphenol, the whole reaction mixture can be fractionally distilled, whereby bisphenol A itself is distilled over particularly carefully under high vacuum, separated from resinous byproducts, and subsequently recrystallized under pressure at elevated temperature [380]. Crude bisphenol A can also be purified by extracting into hot heptane or mixtures of aromatics [381] or by recrystallization from aromatics [382]. A very high purity product with polyester quality is obtained if the bisphenol A – phenol 1 : 1 adduct is separated, recrystallized from phenol, and the phenol removed by distillation [375], [383]. These processes are mostly carried out continuously [372], [374], [376], [380], [383]. The acetone – phenol mixture produced in the Hock phenol synthesis from cumene hydroperoxide can be used as starting material for bisphenol A synthesis [384].

4.4. Analysis, Testing, Storage

Many bisphenols can be analyzed by gas chromatography. The quality of bisphenols can also be tested by measurement of the solidification point and the color index of solutions (bisphenol A in 30 % methanolic solution: APHA color index ≤ 50).

Bisphenols are usually sold as solids, rarely as liquids; bisphenol A is supplied as white flakes which sometimes have a yellow-brown coloration. It can be stored in this form, but not in the molten stage for an unlimited time in stainless steel containers. The pneumatic conveying of bisphenol A must be performed under nitrogen because of the danger of dust explosions.

4.5. Uses, Economic Aspects

Bisphenol A is the most important bisphenol which is used industrially. Since it is mainly used for the production of high-grade plastics, the demand has increased sharply with industrial development.

In 1976, production of bisphenol A in the United States (producers: Shell, General Electric, Dow, Aristech, Union Carbide) amounted to 168 000 t, and in Western Europe (producers: Bayer, Shell, Rhône-Poulenc, General Electric, Dow) 136 000 t were consumed. In line with the rapid industrial growth of the 1980s, world bisphenol A production capacity rose to 10^6 t, a figure which will presumably be exceeded in 1990. About 50 % of this capacity is located in the United States, over 30 % in Western Europe, and the remainder mostly in Japan. However, due to the current economic decline, the new overcapacity is only gradually being utilized.

Whereas in the mid-1970s ca. 50 % of the bisphenol A was converted to epoxy resins and only 40 % to polycarbonates [346], [385], in 1989 over 50 % was used for polycarbonates and ca. 40 % for epoxy resins. The remainder is used for a wide range of products such as aromatic polyesters with phthalic acid; polysulfones; polyetherketones; unsaturated, hydrolysis-resistant polyesters [386]; antioxidants; phenoxy resins from bisphenol A and epichlorohydrin [387]; and phenolic resins elastified by incorporation of bisphenol A [388]. However, the majority is used for flame retardants, especially tetrabromobisphenol A [389].

Polyester resins can be linked with phenolic resins via 4,4 bis(4-hydroxyphenyl)pentanoic acid (see Table 15, [387]. Bisphenol Z is used in the production of polycarbonate films [390].

Bisphenol Z and bisphenol A substituted with various ortho-alkyl groups counteract coccidiosis in poultry [391] and bisphenols containing nitrogen heterocycles are said to act as laxatives [392].

4.6. Toxicology

Bisphenols are far less toxic than phenol itself. Therefore some of them can even be used as antioxidants in cosmetics and foods.

Bisphenol A has an LD_{50} (mouse) of 4 g/kg, limits the weight gain in rats in feeding experiments, and has a clearly estrogenic effect. The permeation of bisphenol A through the skin is practically negligible compared with phenol. It is, however, reported that bisphenol A leads to minor skin irritation .

[Top of Page]

5. Hydroxybiphenyls

Hans-Josef Buysch

5.1. Physical Properties

Hydroxybiphenyls are colorless solids with a weak odor (if any). They are only sparingly soluble in water, but most of them dissolve readily in sodium hydroxide to form salts. Polar organic solvents such as alcohols, ethers, esters, and ketones have an outstanding solvent power for hydroxybiphenyls, whereas that of nonpolar solvents such as aromatic and aliphatic hydrocarbons is lower. The para-hydroxybiphenyls are generally less soluble than the *o*-hydroxybiphenyls. The solubilities can be affected considerably by substitution. Thus, e.g., (**22**) is soluble in various organic solvents but insoluble in water and 10 % sodium hydroxide.

Some physical properties are given in Table 16. 2-Hydroxybiphenyl (**18**) has the greatest industrial importance; further physical properties for 2-hydroxybiphenyl are summarized as follows:

Solubility (g/100 g solvent at 25 °C) in	
methanol	800
ethanol	590
acetone	660
white spirit	20
water	0.07
at 100 °C	0.13
10 % NaOH	50
Heat of fusion	91 J/g
Flash point	138 °C
Ignition temperature	520 °C
Vapor pressure (100 °C)	70 Pa

Table 16. Physical properties of some hydroxybiphenyls

Name	CAS registry no.	Structural formula	Molecular formula	M_r	*mp*, °C	*bp*, °C	d_4^{20}
2-Hydroxybiphenyl	[*90-43-7*]	**18**	$C_{12}H_{10}O$	170.1	57	286	1.219
4-Hydroxybiphenyl	[*92-69-3*]	**19**	$C_{12}H_{10}O$	170.1	165	312	
2,2′-Dihydroxybiphenyl	[*1806-29-7*]	**20**	$C_{12}H_{10}O_2$	186.2	109		
4,4′-Dihydroxybiphenyl	[*92-88-6*]	**21**	$C_{12}H_{10}O_2$	186.2	274–275		
3,5,3′,5′-Tetra-*tert*-butyl-4,4′-dihydroxybiphenyl	[*2455-14-3*]	**22**	$C_{28}H_{42}O_2$	410.6	185		

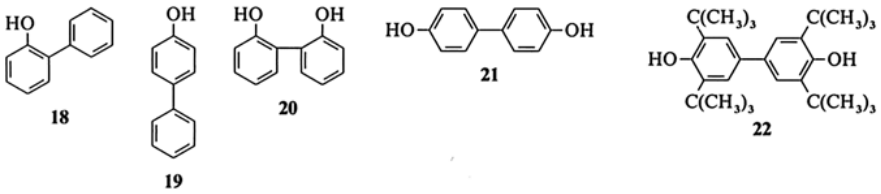

5.2. Chemical Properties

The chemical properties of hydroxybiphenyls are very similar to those of phenol itself. Phenyl substitution has a small but insignificant deactivating effect on the other phenyl nucleus; in the ortho-position it also causes steric hindrance, which is, however, still lower than that of the *tert*-butyl group.

Accordingly, hydroxybiphenyls can undergo reactions at the OH group involving etherification, esterification, and salt formation in the usual way. Because of their bifunctionality, dihydroxybiphenyls, especially 4,4-dihydroxybiphenyl, can be converted into polycondensation products which have achieved a limited industrial importance as liquid crystal polymers.

Substitution at the ring (halogenation, sulfonation, nitration, and alkylation) also takes place as expected. In addition, 4,4′-dihydroxybiphenyls can be oxidized to yield diphenoquinones [394] and 2,2′-dihydroxybiphenyls can undergo ring-closure reactions to form dibenzofurans, carbazoles, and dibenzothiophenes [395].

5.3. Production

2-Hydroxybiphenyl (2-Phenylphenol). The starting material for the synthesis of 2-hydroxybiphenyl (**18**) is cyclohexanone, which is first autocondensed to form 2-cyclohexenylcyclohexanone. This reaction is catalyzed by acidic ion exchangers or by metal salts of higher aliphatic carboxylic acids [396]. 2-Cyclohexenylcyclohexanone is subsequently dehydrogenated to (**16**) in the gas phase with noble metal catalysts on an inert carrier material [396-398].

2-Hydroxybiphenyl is formed as a byproduct in the pressure phenol synthesis from chlorobenzene (⟶ Phenol – Chlorination of Benzene).

4-Hydroxybiphenyl (**19**) is formed alongside (**18**) as a byproduct in the alkaline pressure hydrolysis of chlorobenzene to phenol in 5 – 6 % yield. The mixture consists of 2/3 ortho and 1/3 para isomer which can be separated by fractional distillation.

4-Hydroxybiphenyl can also be produced by sulfonation of biphenyl and alkaline fusion of the *para*-biphenylsulfonic acid [399].

2,2′-Dihydroxybiphenyl (**20**) is obtained by alkaline fusion of biphenylene oxide at ca. 300 °C (sometimes catalyzed by fluorene or carbazole [400], [401]) or by debutylation of 3,3′,5,5′-tetra-*tert*-butyl-2,2′-dihydroxybiphenyl, which itself is produced by the oxidative coupling of 2,4-di-*tert*-butylphenol [402]. 2,2′-Dihydroxy-5,5′-dimethylbiphenyl can be synthesized accordingly.

4,4′-Dihydroxybiphenyl (**21**) can also be synthesized in excellent yields by debutylation of 3,3′,5,5′-tetra-*tert*-butyl-4,4′-dihydroxybiphenyl [403]. Other production processes are the alkaline fusion of 4,4′-biphenyldisulfonic acid, the boil down of benzidinebisdiazonium salts [404], and the decarboxylation of 4,4′-dihydroxybiphenyl-2-carboxylic acid [405]. Symmetrically alkylated 4,4′-dihydroxybiphenyls such as (**22**) or 4,4′-dihydroxy-2,2′,3,3′,5,5′,6,6′-hexamethylbiphenyl are preferably obtained by the oxidative coupling of the corresponding substituted monophenols [403], [406]. Here, diphenoquinones are generally formed in the first step which must be reduced. Phenol itself always gives isomeric mixtures of dihydroxybiphenyls on coupling with various oxidants [407]. These mixtures are also formed by alkaline fusion of 4,4′-dihydroxybiphenylsulfone [408].

5.4. Analysis, Quality Specifications, Storage

The purity of 2-hydroxybiphenyl can be determined by gas chromatography or by titration [409]. The melting point should be 56 – 58 °C.

Commercially 2-hydroxybiphenyl is supplied as virtually colorless and odorless flakes or in the form of the sodium salt (NaOH content < 1.0 %). In the absence of light and moisture it can be stored indefinitely.

5.5. Uses

2-Hydroxybiphenyl shows a broad effectiveness against bacteria, yeasts, and mold, but low toxicity towards warm-blooded organisms [410] and very good biodegradability [411]. Therefore it is increasingly used as a preservative and fungicide and for the production of disinfectants in the leather, paper, glue, textile, plastics, and cosmetics industries. 2-Hydroxybiphenyl and its sodium salt are used for the preservation of citrus fruits in their skins (approval according to EC guideline no. E 321).

2-Hydroxybiphenyl is also used as a dyeing accelerant for dyeing synthetic fibers, e.g., polyethylene terephthalate and cellulose triacetate (Levegal OPS, Bayer).

Sulfonation of 2-hydroxybiphenyl and condensation with formaldehyde yields a dispersion agent that is used in the formulation of plant protection agents in aqueous solution [412].

Trade names for 2-hydroxybiphenyl are:

Preventol O extra (Bayer, content > 99.5 %)
Preventol ON extra (Bayer, sodium salt)
Dowicide 1 (DOW), Dowicide A (DOW, sodium salt)
Cotane (Coalite)

4-Hydroxybiphenyl is an intermediate for the production of varnish resins and nonionogenic emulsifiers (trade names: Emulgator W and Emulvin W, Bayer) which are used in the plant protection, polyurethane, and dyeing sectors.

Other Hydroxybiphenyls. A range of possible applications have been proposed for **20**, **21**, and **22**: e.g., as a fungicide (**20**), an antiseptic for feeds (**20**), (**21**), a toner for xerography

(**20**), starting materials for polyesters, polycarbonates, and powder coatings (**21**). Mixtures of isomers of chlorinated and brominated dihydroxybiphenyls (e.g., chlorinated **20**) are used for flameproofing plastics and have been proposed for combating algae and bacteria. However, dihydroxybiphenyls are actually only used as stabilizers and antioxidants, principally **22**, in oils, rubber, and polyolefins.

5.6. Toxicology

The toxicity of 2-hydroxybiphenyl towards warm-blooded organisms is usually low for a phenolic compound. The LD_{50} for rats is 2700 – 3000 mg per kilogram body weight. Accumulation in rat tissue was not established even after feeding over a period of 2 years with feeds containing 2 % 2-hydroxybiphenyl. Daily doses of 200 mg/kg did not produce any harmful effects even after administration for several weeks. Human skin is neither irritated nor sensitized by application of a 5 % solution in oil [409], [410]. 2-Hydroxybiphenyl is toxic to fish in concentrations > 2 mg/L. However, since it is rapidly and completely degraded in biological sewage treatment plants, fishing grounds are not endangered [411].

[Top of Page]

6. Phenol Ethers

Dorothea Garbe

Phenol ethers are organic compounds of the general formula Ar – O – R, in which Ar represents a substituted or unsubstituted phenyl ring and R is a substituted or unsubstituted aliphatic or aromatic hydrocarbon group.

6.1. Properties

Physical Properties. Most phenol ethers have a pleasant, characteristic smell. Some phenol ethers are important in the flavoring and fragrance industry (⟶ Flavors and Fragrances – Phenols and Phenol Derivatives). The ethers described in this article are those, whose importance lies mainly in other areas. The lower phenol ethers are colorless liquids that are insoluble in water and soluble in organic solvents.

Chemical Properties. The ether linkage in phenol ethers is very stable. The carbon – oxygen bond in completely aromatic phenol ethers such as diphenyl ether is cleaved only under drastic conditions, e.g., with alkalis at high temperature and pressure; in contrast, the aliphatic carbon – oxygen bond in anisole, for example, is split by treatment with strong acids such as hydrogen iodide to form phenol and an alkyl iodide. Sterically hindered methoxy groups can be cleaved selectively with BCl_3. The phenyl ring is susceptible to substitution reactions, including hydroxylation, nitration, halogenation, and sulfonation. Phenol ethers can be alkylated or acylated by Friedel – Crafts reactions. Aromatic – aliphatic ethers, however, can also be cleaved in the presence of Friedel – Crafts catalysts at high temperatures.

6.2. Production

Nearly all phenol ethers can be prepared by the reaction of phenols with alkyl or aryl halides (preferably chlorides) in weakly basic aqueous media:

$$C_6H_5ONa + RCl \rightarrow C_6H_5OR + NaCl$$

For the preparation of phenol ethers with short-chain aliphatic groups, dialkyl sulfates are often employed as alkylating agents:

$$C_6H_5OH + (RO)_2SO_2 \xrightarrow{NaOH} C_6H_5OR + (RO)SO_3Na + H_2O$$

The second alkyl group present in the monoalkyl sulfate can be reacted using higher temperatures. Another method for the preparation of aromatic – aliphatic ethers is the etherification of phenols with aliphatic alcohols in the presence of an acidic ion exchanger [413]. This reaction is performed using excess phenol to suppress formation of the dialkyl ether [414].

6.3. Representative Phenol Ethers

Phenyl methyl ether [100-66-3] (**23**), anisole, C_7H_8O, M_r 108.15, *mp* –37.5 °C, *bp* (101.3 kPa) 155 °C, d_4^{20} 0.9954, is a colorless liquid with a spicy-sweet smell. Anisole can be produced by all

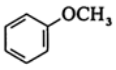

23

the aforementioned methods. It is an important intermediate in the synthesis of organic compounds, for example fragrances and pharmaceuticals. The compound is also used as a solvent and as a heat transfer medium.

Phenyl ethyl ether [103-73-1] (**24**), phenetole, $C_8H_{10}O$, M_r 122.17, *mp* –29.5 °C, *bp* (101.3 kPa) 170 °C, d_4^{20} 0.966, is a colorless liquid which is prepared by etherification of phenol with ethyl chloride or diethyl sulfate. Phenyl ethyl ether is an intermediate in numerous organic syntheses.

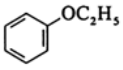

24

Diphenyl ether [101-84-8] (**25**), $C_{12}H_{10}O$, M_r 170.21, *mp* 26.8 °C, *bp* (101.3 kPa) 257.9 °C, *bp* (1.34 kPa) 121 °C, d_4^{20} 1.0748, occurs, e.g., in lemon balm oil. The compound forms colorless

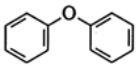

25

crystals that possess a geranium-like aroma. Diphenyl ether is formed as a byproduct in the high-pressure hydrolysis of chlorobenzene. Because of its high thermal stability, a eutectic mixture of diphenyl ether and biphenyl is employed as a heat transfer medium. Diphenyl ether is the starting material for organic intermediates and is used as a fragrance component. The compound is also employed as a processing aid in the production of polyesters [415]; several polybromine derivatives of diphenyl ether are used as fire retardants.

2-Phenoxyethanol [122-99-6] (**26**), $C_8H_{10}O_2$, M_r 138.17, *mp* –2 °C, *bp* (101.3 kPa) 237 °C, *bp* (2.4 kPa) 134 – 135 °C, d^{20} 1.0020, is a colorless liquid with a faint rose-like aroma.

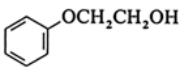

26

2-Phenoxy-ethanol is produced by the hydroxyethylation of phenol, for example, in the presence of alkali-metal hydroxides or alkali-metal borohydrides [416]. The compound is known under the trade name Cellosolve as a solvent for cellulose acetate. 2-Phenoxyethanol is a synthetic intermediate in the production of plasticizers, pharmaceuticals, and fragrances. In the perfume industry the compound finds use as a fixing agent.

Phenoxyacetic acid [122-59-8] (**27**), $C_8H_8O_3$, M_r 152.16, *mp* 98 – 99 °C, *bp* (101.3 kPa) 285 °C (decomp.), forms colorless crystals that possess a faint honey-like odor and are soluble in the common organic solvents.

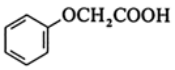

27

Phenoxyacetic acid can be prepared by reaction of phenol with chloroacetic acid in dilute aqueous sodium hydroxide, or by hydrolysis of its methyl ester. The methyl ester is obtained by reaction of phenol and methyl chloroacetate in the gas phase with zinc oxide – aluminum oxide as catalyst [417]. Phenoxyacetic acid is an intermediate for fragrances and dyes and is a starting material for the synthesis of penicillin V.

4-Methylanisole [104-93-8] (**28**), $C_8H_{10}O$, M_r 122.17, *bp* (101.3 kPa) 176.5 °C, d_{25}^{25} 0.9698, is a colorless liquid with a pungent odor, which in dilute solution has a flower-like

fragrance. 4-Methylanisole is a starting material for the preparation of *p*-anisaldehyde.

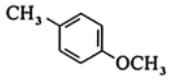

28

2-Methoxyphenol [90-05-1] (**29**), guaiacol, $C_7H_8O_2$, M_r 124.15, *bp* (101.3 kPa) 205 °C, *bp* (3.2 kPa) 106.5 °C, d_4^{21} 1.128, forms colorless prisms that have a sweet, faint phenol odor. Guaiacol is present in many essential oils and is also found as a flavor compound in many foods. Guaiacol is prepared by monomethylation of catechol using methyl halides or dimethyl sulfate. Monomethylation of catechol can also be accomplished with methanol in the presence of phosphoric acid, phosphates, or an ion exchanger [418]. Guaiacol is an important synthetic intermediate in the flavoring and fragrance industries (e.g., for the synthesis of vanillin), as well as in the pharmaceutical industry.

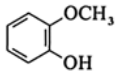

29

2-Ethoxyphenol [94-71-3] (**30**), $C_8H_{10}O_2$, M_r 138.17, *mp* 29 °C, *bp* (101.3 kPa) 217 °C, *bp* (5.3 kPa) 68 °C, is prepared analogously to guaiacol by monoetherification of catechol. The compound is an intermediate in the synthesis of ethyl vanillin.

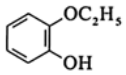

30

2-Isopropoxyphenol [4812-20-8] (**31**), $C_9H_{12}O_2$, M_r 152.20, *bp* (101.3 kPa) 217 °C, *bp* (14.7 kPa) 100 – 102 °C, is prepared by monoetherification of catechol with isopropyl halides in the presence of alkali-metal salts and either a phase transfer catalyst [419] or a polyhydroxyalkyl monoether in an autoclave [420]. This compound is the starting material for the synthesis of the pesticide Propoxur (2-isopropoxyphenyl methylcarbamate).

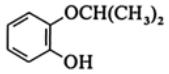

31

Veratrole [91-16-7] (**32**), $C_8H_{10}O_2$, M_r 138.17, *mp* 22.5 °C, *bp* (101.3 kPa) 206 °C, *bp* (1.3 kPa) 90 °C, d_{25}^{25} 1.084, is prepared by methylation of catechol or guaiacol. This compound is an intermediate in the synthesis of alkaloids and pharmaceuticals.

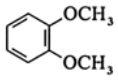

32

1,3-Benzodioxole [274-09-9] (**33**), $C_7H_6O_2$, M_r 122.13, *bp* (101.3 kPa) 172 °C, is a colorless liquid. This compound is prepared by the reaction of catechol and dihalomethanes, for example, by use of concentrated aqueous alkaline solution in the presence of tetraalkylammonium or phosphonium salts [421], sometimes in the presence of alkyl iodides [422]. 1,3-Benzodioxole is an important intermediate in organic synthesis, especially for the preparation of alkaloids.

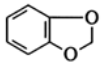

33

[Top of Page]

7. Halogen Derivatives of Phenolic Compounds

Wilfried Paulus

The phenolic derivatives described in this chapter include halogenated alkylphenols, cycloalkylphenols, arylphenols, alkylarylphenols, phenoxyphenols, dihydroxydiphenyl sulfides (thiobisphenols) and halogenated salicylanilides. They all possess antimicrobial properties and are therefore widely used as microbicides. Depending on their chemical and physical properties, the compounds can be used (1) as preservatives for functional fluids, (2) to control slime in process water systems, (3) to protect various materials e.g., textiles, paper, and wood against microbial decay, and (4) as active ingredients in disinfectants (for halogenated phenol derivatives used in disinfectants, see ⟶ Disinfectants). Pentachlorophenol (PCP) is not discussed in this chapter even though it can be widely used as a biocide. The use of PCP as a biocide is rapidly declining because of its ecotoxicity and general toxicity. The production, use, and marketing of PCP and PCP-containing products has been banned in various countries, e.g., in the Federal Republic of Germany from January 1, 1990.

Similarly, 2,4,5- and 2,4,6-trichlorophenol are not included in this chapter although they possess strong microbicidal properties. There is pressure to find suitable substitutes for these compounds because they are the starting materials in the formation of the highly toxic tetrachlorodibenzo-*p*-dioxins.

7.1. Introduction

Phenol (formerly known as carbolic acid) was discovered in 1834. In 1860, the antiseptic properties of phenol were recognized and used for antibacterial treatment of surgical instruments and bandaging material and for treatment of wounds. This was the beginning of the use of chemical methods to kill microbes or to inhibit their growth. At the same time, the toxic and skin-cauterizing properties of phenol itself led to the isolation and syntheses of hundreds of phenol derivatives with the aim of finding less toxic phenol derivatives and more effective compounds which required lower concentrations for use. With the discovery that chemicals could be used to protect materials against microbial attack and decay, phenol derivatives with suitable chemical and physical properties were investigated for such uses. This development, which can now be considered complete, also led to discovery of the correlation between structure and effectiveness and the mechanism of action of phenol derivatives [431].

Phenol derivatives are so-called membrane-active microbicides. Initially, they are reversibly adsorbed on the surface of microbe cells and inhibit metabolic processes. If applied in higher concentrations, the phenol derivatives dissolve fairly quickly and readily in lipids according to their chemical and physical properties and penetrate into the microbe cell. They react with the protoplasma and cell proteins and inhibit enzymes, which leads to destruction of the microbial cell. The applied concentration determines whether the phenol derivatives act microbistatically or microbicidally.

Phenol derivatives show optimal antimicrobial effect in neutral and acid media, i.e., in the undissociated state, since only in this state are they membrane-active and exhibit affinity for the negatively charged surface of the microbial cell wall; the negatively charged phenolate anions are repelled. Therefore, if optimal advantage is to be taken of the antimicrobial power of phenol deravatives during application, their pK_a values must be taken into account. The pK_a values for some important phenolic microbicides are listed in the following:

2-Hydroxybiphenyl	11.6
6-Isopropyl-3-methylphenol	10.6
4-Chloro-3,5-dimethylphenol	9.7
4-Chloro-3-methylphenol	9.6
2,4,6-Trichlorophenol	8.5
4-Chloro-2-(2,4-dichloro)phenoxyphenol	7.9
2,2′-Methylenebis(4-chlorophenol)	
pK_a (1)	8.7
pK_a (2)	12.6

Halogenation of phenol derivatives produces active substances with substantially increased antimicrobial effect. The acidity of the compounds is also more pronounced; the degree of dissociation of halogenated phenol derivatives increases with the number of halogen atoms (see above). The combination of alkylation and halogenation — the latter preferably in

the 4-position — produces phenolic microbicides which have great practical importance as active ingredients in disinfectants and for the preservation of materials. These include 2-benzyl-4-chlorophenol, 4-chloro-3-methylphenol, and 4-chloro-3,5-dimethylphenol. The minimum inhibitory concentrations listed in Table 17 give an indication of the range of effect of halogenated phenol derivatives which have particular practical importance as microbicides.

Table 17. Minimum inhibitory concentration (MIC) of halogenated phenol derivatives in nutrient agar

Test organisms	MIC, mg/L				
	PCMC [a]	PCMX [b]	Chlorophene [c]	Dichlorophene [d]	Fenticlor [e]
Bacillus subtilis	150	75	10	100	
Bacillus punctatum	200	100	10	50	
Escherichia coli	250	200	3500	100	75
Leuconostoc mesenteroides	200	100	10	5	3500
Proteus vulgaris	200	200	100	50	
Pseudomonas aeruginosa	800	1000	5000	>5000	
Staphylococcus aureus	200	100	20	5	35
Desulfovibrio desulfuricans	35	50	50	20	
Candida albicans	200	75	50	50	50
Torula rubra	50	100	50	50	50
Aspergillus niger	100	100	100	100	50
Aureobasidium pullulans	30	50	20	35	<20
Chaetomium globosum	80	50	20	20	75
Cladosporium herbarum	200	100	100	200	35
Coniophora puteana	100	35	5	2	<20
Penicillium citrinum	100	50	75	50	20
Penicillium glaucum	100	35	50	50	10
Polyporus versicolor	5000	75	50	50	35
Trichoderma viride	140	100	100	50	50
Trichophyton pedis	100	50	10	10	10

[a] PCMC = 4-chloro-3-methylphenol.
[b] PCMX = 4-chloro-3,5-dimethylphenol.
[c] Chlorophene = 2-benzyl-4-chlorophenol.
[d] Dichlorophene = 2,2′-methylenebis(4-chlorophenol).
[e] Fenticlor = 2,2′-thiobis(4-chlorophenol).

It is often claimed that phenols, especially chlorinated phenols, show high oral toxicity, are toxic percutaneously, and are in general difficult or impossible to degrade biologically. This disqualification of phenols is a very unreliable generalization and is not verified by the facts. The whole class of phenolic microbicides must not be looked at uniformly and unjustly disqualified, just because some representatives of this class of compounds indeed possess properties such as high ecotoxicity, are persistent, and can contain dioxins. Even some of the halogenated phenol derivatives have toxicological properties and ecological effects which satisfy the standards for microbicides used for preservation of materials and for disinfectants that are set by a civilization with an increased sense of responsibility for the environment. As shown by the curves in Figure 3, there are phenol derivatives including halogenated phenols which can be biodegraded quickly and effectively, and others, e.g., pentachlorophenol which are only biodegraded slowly and only if present in very low concentrations [423]. The biodegradation of phenols does not only occur in biological sewage plants; surface waters also possess a considerable potential to biodegrade phenol derivatives including pentachlorophenol [424].

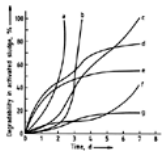

Figure 3. Biodegradability of various phenol derivatives

a) 2-Hydroxybiphenyl, 100 ppm; b) 4-Chloro-3-methylphenol, 50 ppm; c) Benzylphenol, 20 ppm; d) 2-Benzyl-4-chlorophenol, 50 ppm; e) Dichlorophene, 20 ppm; f) 4-Chloro-3,5-dimethylphenol, 10 ppm; g) Pentachlorophenol, 10 ppm

7.2. Representative Compounds

4-Chloro-3-methylphenol [59-50-7] (**34**), *p*-chloro-*m*-cresol, PCMC, C_7H_7ClO, M_r 142.59, has a density ϱ 1.37 g/cm^3 at 20 °C, a vapor pressure of 8 Pa at 20 °C, of 0.7 kPa at 100 °C, a *mp* of 63 – 65 °C, a *bp* of ca. 239 °C, the flash point is ca. 118 °C, and the ignition temperature ca. 590 °C. Its solubility in water is 4 g/L. 4-Chloro-3-methylphenol is readily soluble in alkali solutions and organic solvents, and has a pK_a of 9.6.

Toxicological Data [431], [432]. LD_{50} (oral, rat) 5129 mg/kg, LD_{50} (dermal, rat) >500 mg/kg (exposure 7 d). PCMC irritates the skin and mucous membranes but is not significantly sensitizing and not mutagenic. LCLo for fish 0.5 mg/L (96 h exposure).

Preparation and Uses. 4-Chloro-3-methylphenol is prepared by chlorination of 3-methylphenol with, e.g., SO_2Cl_2. PCMC is of major importance as a preservative for aqueous functional fluids and as a raw material for disinfectants (⟶ Disinfectants). This is due to its strong and wide-ranging antimicrobial effect combined with favorable chemical and physical properties and good biodegradability [431].

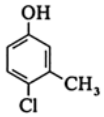

34

2-Benzyl-4-chlorophenol [120-32-1] (**35**), Chlorophene, $C_{13}H_{11}ClO$, M_r 218.69, has a density ϱ 1.22 g/cm^3 at 20 °C, a vapor pressure of 10 Pa at 100 °C, a *mp* of 48 – 49 °C, and a *bp* (1.5 kPa) of 190 – 196 °C. Its flash point is ca. 188 °C, and the ignition temperature ca. 490 °C. 2-Benzyl-4-chlorophenol has a solubility in water of 0.4 g/L at 20 °C. It is readily soluble in alkali solutions and organic solvents and has a pK_a of 11.0.

Toxicological Data. LD_{50} (oral, rat) > 5000 mg/kg, LD_{50} (dermal, rat) > 2500 mg/kg, LC_{50} (inhaled for 4 h, rat) 2.5 mg/L. Chlorophene strongly irritates the skin and mucous membranes but is not toxic percutaneously and is neither mutagenic nor teratogenic. Its LCLo for *Leuciscus idus* is 0.5 mg/L (48 h exposure). Organisms in activated sludge tolerate 10 – 50 mg/L Chlorophene and are able to degrade it [431], [432].

Preparation and Uses. Chlorophene can be prepared by reaction of 2-benzylphenol with SO_2Cl_2 at 100 °C. It is used as an active agent in disinfectants (⟶ Disinfectants).

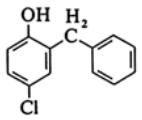

35

4-Chloro-2-cyclopentylphenol [13347-42-7] (**36**), $C_{11}H_{13}ClO$, M_r 196.68, *mp* 23.7 °C, *bp* 303 °C, has a solubility in water of 0.24 g/L at 20 °C, and is readily soluble in organic solvents.

Toxicological Data. LD_{50} (oral, rat) 2460 mg/kg, LD_{50} (oral, rabbit) 420 mg/kg, LD_{50} (dermal, rabbit) 850 mg/kg. 4-Chloro-2-cyclopentylphenol strongly irritates the skin and mucous membranes but is not sensitizing.

Preparation. 4-Chloro-2-cyclopentylphenol is prepared by shaking 4-chlorophenol and cyclopentene in concentrated sulfuric acid; 2,6-dicyclopentyl-4-chlorophenol (*bp* 235 – 240 °C at 1.8 kPa) is produced as a byproduct. 4-Chloro-2-cyclopentylphenol is no longer used as a microbicide in disinfectants because of its toxicity.

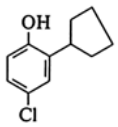

36

5-Chloro-2-hydroxybiphenyl [1331-46-0] (**37**), monochloro-*o*-phenylphenol, MCOPP, $C_{12}H_9ClO$, M_r 204.66, *mp* 36 – 37 °C, *bp* (1.5 kPa) 178 °C, is very sparingly soluble in water and readily soluble in organic solvents.

Toxicological Data. LD_{50} (oral, rat) 3500 mg/kg. 5-Chloro-2-hydroxybiphenyl is toxic percutaneously, irritates the skin and mucous membranes, and can cause photoallergic contact dermatitis. 5-Chloro-2-hydroxybiphenyl is no longer used as a microbicide in disinfectants because of these properties.

Preparation. 5-Chloro-2-hydroxybiphenyl is obtained along with 2-chloro-6-hydroxybiphenyl when chlorine is introduced into 2-hydroxybiphenyl.

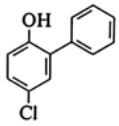

37

4-Chloro-3,5-dimethylphenol [88-04-0] (**38**), *p*-chloro-*m*-xylenol, PCMX, C_8H_9ClO, M_r 156.61, *mp* 114 – 116 °C, *bp* 246 °C, has a vapor pressure of 0.3 kPa at 100 °C, its solubility in water is 0.33 g/L at 20 °C. 4-Chloro-3,5-dimethylphenol is soluble in alkali solutions and organic solvents, and has a pK_a of 9.7.

Toxicological Data. LD_{50} (oral, rat) 3830 mg/kg. PCMX mildly irritates the skin and mucous membranes, has a low sensitization potential, is not mutagenic in the Ames test, and is not teratogenic in rats; the LCLo for *Leuciscus idus* is 1 mg/L. Organisms in activated sludge tolerate 15 mg/L PCMX and slowly degrade the substance [431], [432].

Preparation and Uses. 4-Chloro-3,5-dimethylphenol is prepared by reaction of 3,5-dimethylphenol with chlorine at 25 – 30 °C or with SO_2Cl_2 starting at 30 – 40 °C [425]. It is used as a raw material for disinfectants (⟶ Disinfectants), as an active component in deodorants, soaps, skin preparations for dermatological disorders, and antiseptics, and as a preservative for aqueous functional fluids.

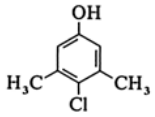

38

4-Chloro-6-isopropyl-3-methylphenol [89-68-9] (**39**), 4-chlorothymol, $C_{10}H_{13}ClO$, M_r 184.67, *mp* 64 °C, *bp* 259 – 263 °C, has a solubility in water of 0.3 g/L at 20 °C, is soluble in alkali solutions and readily soluble in organic solvents.

Toxicological Data. LD_{50} (oral, mouse) 2460 mg/kg.

Preparation and Uses. 4-Chloro-6-isopropyl-3-methylphenol is prepared by reaction of equimolar amounts of propene and 4-chloro-3-methylphenol in the presence of bleaching clay under pressure at 140 – 160 °C or by chlorination of thymol with SO_2Cl_2 in chloroform. On account of its microbicidal properties [426], chlorothymol is of importance as a raw material for disinfectants [427] (⟶ Disinfectants).

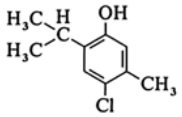

39

4-Benzyl-2-chloro-6-methylphenol [65053-91-0] (**40**), $C_{14}H_{13}ClO$, M_r 232.72, has a *bp* (27 Pa) of 128 – 129 °C, and an n_D^{20} of 1.5944.

Toxicological Data. LD_{50} (oral, mouse) 1700 mg/kg, LD_{50} (subcutaneous, mouse) 1000 mg/kg, LD_{50} (intravenous, mouse) 96 mg/kg [428].

Preparation. 4-Benzyl-2-chloro-6-methylphenol is prepared from 2-chloro-6-methylphenol (or the forerunnings formed in the production of 4-chloro-2-methylphenol by chlorination of 2-methylphenol) by reaction with benzyl chloride in the presence of anhydrous zinc chloride starting at 100 °C and reaching completion after 5 h at 150 – 155 °C [428].

Uses. The substance shows a strong antimicrobial effect especially towards bacteria and fungi, and in particular dermatophytes. The microbicide is used mostly for epidermal applications, e.g., in pharmaceuticals, cosmetics, and soaps because it is tolerated well by skin and has a low skin-sensitization potential (⟶ Disinfectants).

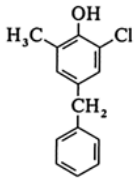

40

4-Bromo-2,6-dimethylphenol [2374-05-2] (**41**), 4-bromo-2,6-xylenol, C_8H_9BrO, M_r 201.08, has a *mp* of 80 – 81 °C.

Toxicological Data. LD_{50} (i.p., mouse) 650 mg/kg.

Preparation and Uses. 4-Bromo-2,6-dimethylphenol is prepared by bromination of 2,6-dimethylphenol in glacial acetic acid at 15 °C and used as a microbicide in disinfectants (⟶ Disinfectants).

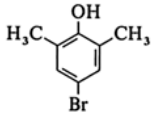

41

2,4-Dichloro-3,5-dimethylphenol [133-53-9] (**42**), dichloro-*m*-xylenol, DCMX, $C_8H_8Cl_2O$, M_r 191.06, *mp* 95 – 96 °C (sublimation), *bp* 250 °C, flash point 134 °C, has a solubility in water of 0.2 g/L at 20 °C, is soluble in alkali solutions, and readily soluble in organic solvents.

Toxicological Data. LD_{50} (oral, rat) is 2810 – 4120 mg/kg.

Preparation and Uses. DCMX is obtained by reaction of 4-chloro-3,5-dimethylphenol with *N*-chloroacetamide in glacial acetic acid with concentrated HCl. It is used as a preservative in aqueous functional fluids and as a microbicide in disinfectants, but in comparison with PCMX (**38**) is of lesser importance because it has a more intense smell and is less soluble in water.

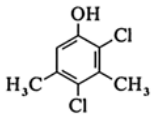

42

Tetrabromo-2-methylphenol [576-55-6] (**43**), 3,4,5,6-tetrabromo-*o*-cresol, $C_7H_4Br_4O$, M_r 423.75, *mp* 205 – 208 °C, is practically insoluble in water. The solubility (g/L) in various solvents is as follows: ethanol 85, isopropanol 80, ethylene glycol 9, propylene glycol 8, vaseline oil 4. Tetrabromo-2-methylphenol discolors in light and in contact with traces of iron.

Toxicological Data. Tetrabromo-2-methylphenol has an LD_{50} (oral, rat) > 6400 mg/kg and is well tolerated by skin and mucous membranes.

Preparation and Uses. Tetrabromo-2-methylphenol is obtained by reaction of *o*-cresol with bromine in tetrachloromethane in the presence of aluminum and iron powder. It is used as a preservative for cosmetics, as an active component in deodorants and disinfectants (⟶ Disinfectants), and as an active agent for antimycotic finishing of textile materials; it is, however, of minor importance [432].

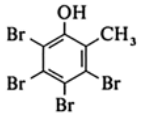

43

5-Chloro-2-(2,4-dichlorophenoxy)phenol [3380-34-5] (**44**), 2,4,4′-trichloro-2′-hydroxydiphenylether, Triclosan, $C_{12}H_7Cl_3O_2$, M_r 289.55, *mp* 60 – 61 °C, decomposes at 280 – 290 °C. The vapor pressure is 2.6 Pa at 100 °C. The solubility (g/L) in various solvents is as follows: water 0.01 at 20 °C, 0.039 at 50 °C, in 0.1 mol/L NaOH 23; it is readily soluble in organic solvents. Triclosan has a pK_a of 7.9, is unstable in light, and discolors in contact with traces of heavy metals. Under certain conditions such as higher alkalinity and heat, conversion to chlorinated dibenzo-*p*-dioxines can occur.

Toxicological Data [432]. LD_{50} (oral, rat or dog) > 5000 mg/kg, LD_{50} (intravenous, rat) 29 mg/kg, LD_{50} (intraperitoneal, rat) 198 mg/kg. Triclosan is neither mutagenic nor teratogenic, is tolerated well by the skin, and has shown no evidence of sensitization potential, LCLo for fish is ca. 0.6 mg/L (exposure 48 h). Activated sludge organisms tolerate up to 10 mg/L Triclosan and degrade the substance slowly.

Preparation and Uses. Triclosan is produced by treatment of 2,4,4′-trichloro-2′-methoxydiphenyl ether with $AlCl_3$ in benzene under reflux. It is highly active against staphylococci and is used as an active agent in deodorants and antiseptic soaps (⟶ Disinfectants).

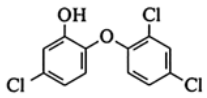

44

2,2′-Methylenebis(4-chlorophenol) [97-23-4] (**45**), Dichlorophene, $C_{13}H_{10}Cl_2O_2$, M_r 269.14, *mp* 176 °C, has a density ϱ 1.5 g/cm^3 at 20 °C, a vapor pressure < 1 Pa at 100 °C, and is not steam-volatile. The solubility (g/L) in different solvents is as follows: water 0.07 at 20 °C, 0.2 at 50 °C; 10 % NaOH 500, butanol 400; toluene 12. Dichlorophene has a pK_a (1) of 8.7 and pK_a (2) of 12.6.

Toxicological Data [431], [432]. LD_{50} (oral, rat) 3300 mg/ kg. Dichlorophene does not irritate the skin, mildly irritates the mucous membranes, and skin sensitization is not significant. Dichlorophene is neither mutagenic nor teratogenic. LCLo for *Leuciscus idus* 0.5 mg/L (72 h exposure); data on biodegradability are given in Figure 3.

Preparation and Uses. Dichlorophene is produced by reaction of 4-chlorophenol with 0.5 M aqueous formaldehyde solution with the addition of sulfuric acid at 50 – 65 °C. On account of its bactericidal and fungicidal properties, the pure nearly odorless compound can be used in deodorants and antiseptic soaps, for antimicrobial finishing of textiles and paper, and as a preservative for metalworking lubricants.

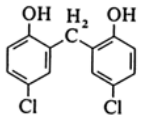

45

4,4′-Methylenebis(2,6-dichlorophenol) [3933-88-8] (**46**), $C_{13}H_8Cl_4O_2$, M_r 338.04, *mp* 184 –185 °C, is produced by introducing chlorine into a solution of 4,4′-methylenebis(2-chlorophenol) in glacial acetic acid at room temperature or by treating 2,6-dichlorophenol with formalin in concentrated sulfuric acid at –10 °C. It has antimicrobial properties, but because of its high toxicity it no longer has practical importance as a microbicide.

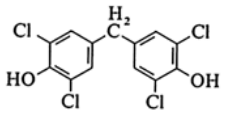

46

2,2′-Methylenebis(3,4,6-trichlorophenol) [70-30-4] (**47**), Hexachlorophene, $C_{13}H_6Cl_6O_2$, M_r 406.92, has an *mp* of 164 – 165 °C; its solubility (g/L) in various solvents is as follows: acetone 1010, ethanol 500, toluene 56. Hexachlorophene is soluble in alkalis but practically insoluble in water. Its pK_a values are 5.4 and 10.9.

Toxicological Data [431], [432]. LD_{50} (oral, rat) 59 mg/kg, LD_{50} (oral, mouse) 80 mg/kg. Hexachlorophene is neither mutagenic nor teratogenic; it is however, neurotoxic, irritates the skin and mucous membranes, and is absorbed through the skin.

Preparation and Uses. Hexachlorophene is formed by reaction of 2,4,5-trichlorophenol with paraformaldehyde in fuming sulfuric acid (20 % SO_3) at 65 – 135 °C. Hexachlorophene is a strong bacteriostat which also has a microbicidal effect at high concentrations. The use of Hexachlorophene as an antimicrobial agent in cosmetics, medicinal soaps, detergent solutions, and textiles has, however, been discontinued because of its toxicity. For details on the correlation between chemical structure and biological activity of chlorinated methylenebisphenols and their physical, chemical, and pharmacological properties, see [429].

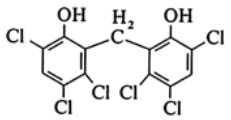

47

2,2′-Methylenebis(6-bromo-4-chlorophenol) [15435-29-7] (**48**), bromochlorophene, $C_{13}H_8Br_2Cl_2O_2$, M_r 426.95, *mp* 188 – 191 °C, its solubility (g/L) in different solvents is as follows: water < 0.01, 95 % ethanol 0.5, propanol 0.7, isopropanol 0.4, 1.2-propylene glycol 2.5, paraffin oil 0.05.

Toxicological Data [432]. LD_{50} (oral, rat) 3700 mg/kg, LD_{50} (oral, mouse) 1550 mg/kg, LD_{50} (dermal, rat) > 10 000 mg/kg. Bromochlorophene does not irritate the skin and there is no evidence of sensitization.

Preparation and Uses. Bromochlorophene is produced by bromination of 2,2′-methylenebis(4-chlorophenol) in glacial acetic acid in the cold. It is used as a microbicide in cosmetics, e.g., deodorants, mouthwashes, and toothpastes.

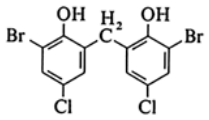

48

2,2′-Thiobis(4-chlorophenol) [97-24-5] (**49**), Fenticlor, $C_{12}H_8Cl_2O_2S$, M_r 287.18, *mp* 174 °C, has a solubility in water of 30 mg/L and is readily soluble in alkalis and organic solvents.

Toxicological Data. LD_{50} (oral, rat) 3250 mg/kg; photosensitization of the skin is possible.

Preparation and Uses. Fenticlor is produced (1) by the reaction of 4-chlorophenol with sulfur dichloride in the presence of $AlCl_3$ or (2) by chlorination of 2,2′-thiobisphenol with chlorine in glacial acetic acid or with SO_2Cl_2 in dichlorobenzene. It shows a tendency to discoloration. Fenticlor has a broad range of antimicrobial action; it is, e.g., effective against bacteria, fungi, yeasts, and algae. Fenticlor is therefore used as a preservative for aqueous functional fluids and to control algae and slime in process water circulation systems [431].

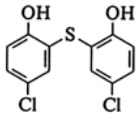

49

2,2′-Thiobis(4,6-dichlorophenol) [97-18-7] (**50**), Bithionol, $C_{12}H_6Cl_4O_2S$, M_r 356.07, *mp* 188 °C, is virtually insoluble in water, readily soluble in ethanol, diethyl ether, acetone, glacial acetic acid, and dilute alkalis.

Toxicological Data. LD_{50} (oral, rat) 1430 mg/kg, LD_{50} (oral, mouse) 2100 mg/kg.

Preparation and Uses. Bithionol is prepared by reaction of 2,4-dichlorophenol with sulfur dichloride and $AlCl_3$ in carbon disulfide or tetrachloromethane. It is used as a microbicide to protect aqueous functional fluids from biodeterioration and to control algae and slime formation in process water systems [431].

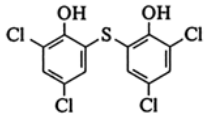

50

2,2′-Thiobis(3,4,6-trichlorophenol) [3161-14-6] (**51**), $C_{12}H_4Cl_6O_2S$, M_r 424.97, *mp* 157 – 159 °C, is practically insoluble in water, readily soluble in ethanol, diethyl ether, acetone, and glacial acetic acid. It is prepared by reaction of 2,4,5-trichlorophenol with SCl_2 and $AlCl_3$ in 1,2-dichloroethane at 45 °C. Its use as a microbicide is declining because of its toxicity.

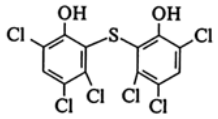

51

3,3′-Thiobis(2,4,6-trichlorophenol) [104294-11-3] (**52**), $C_{12}H_4Cl_6O_2S$, M_r 424.97, *mp* 285 °C, is prepared by reaction of 2,4,6-trichlorophenol with SCl_2 in the presence of $AlCl_3$ in carbon disulfide or tetrachloromethane. Its use as a microbicide is declining.

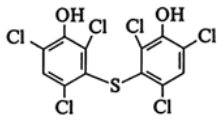

52

5,2′-Dichlorosalicylanilide [6626-92-2] (**53**), $C_{13}H_9Cl_2NO_2$, M_r 282.14, *mp* 188 °C, is produced by heating equimolar quantities of 3-chlorosalicylic acid and 2-chloroaniline in the presence of PCl_3 in chlorobenzene or toluene.

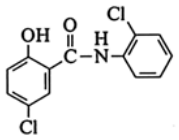

53

5,3′,4′-Trichlorosalicylanilide [642-84-2] (**54**), Anobial, $C_{13}H_8Cl_3NO_2$, M_r 316.59, *mp* 244 –246 °C, is virtually insoluble in water, soluble in acetone and methanol. It is produced by reaction of 3-chlorosalicylic acid and 3,4-dichloroaniline with PCl_3 in boiling toluene. It is more effective as a microbicide than DBS (**55**), but its use as a microbicide [430] is declining.

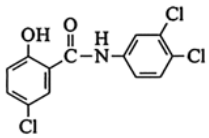

54

4,3′,4′-Trichlorosalicylanilide [58622-66-5] (**54a**) $C_{13}H_8Cl_3NO_2$, M_r 316.59, *mp* 221 –

222 °C, is soluble in acetone and methanol and is light-sensitive. It is produced by reaction of 4-chlorosalicylic acid and 3,4-dichloroaniline with $AlCl_3$ and PCl_3 in nitrobenzene. It was formerly used a microbistat in deodorants and soaps [430].

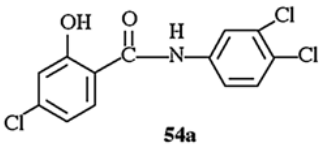

54a

5,4′-Dibromosalicylanilide [87-12-7] (**55**), DBS, Disanyl, $C_{13}H_9Br_2NO_2$, M_r 371.21, *mp* 238 – 245 °C, is very sparingly soluble in water, sparingly soluble in ethanol, and soluble in acetone and dimethylformamide. Its LD_{50} (oral, mouse) is 1700 mg/kg; photosensitization of the skin is possible. 5,4′-Dibromosalicyclanilide is produced from salicylanilide and bromine in acetic acid. DBS has broad antimicrobial activity [430] but its use as a microbicide in, e.g., paper and plastics is declining.

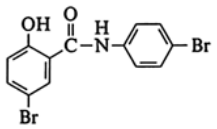

55

3,5,4′-Tribromosalicylanilide [87-10-5] (**56**), TBS, Trisanyl, $C_{13}H_8Br_3NO_2$, M_r 449.96, *mp* 227 – 228 °C, is practically insoluble in water, soluble in hot acetone, and readily soluble in dimethylformamide. Its LD_{50} (oral, rat) is 570 mg/kg; photosensitization of the skin is possible. 3,5,4′-Tribromosalicylanilide is produced by direct bromination of salicylanilide in aqueous acetic acid at 50 – 55 °C or in water containing an emulsifier at 50 – 65 °C. Its use as a microbicide for the protection of materials is declining [430], [432].

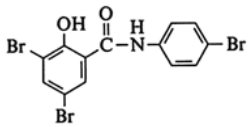

56

[Top of Page]

References

General References

1. *Beilstein*, **6 H**, 470 ff; **6 I**, 234 ff; **6 II**, 442 ff, **6 III**, 1655 ff, **6 IV**, 3011 ff.
2. *Houben-Weyl*, 6/1 c, parts 1 and 2.
3. *Rodd's Chemistry of Carbon Compounds*, 2nd. ed., Elsevier, Amsterdam vol. III, A, pp. 289 ff., suppl. III A, pp. 161 ff.
4. A. Dierichs, R. Kubička: *Phenole und Basen*, Akademie-Verlag, Berlin 1958.
5. *Ullmann*, 4th ed., **18**, 191 – 215.

Specific References

6. R. S. Bownman, D. R. Stevens, *J. Org. Chem.* **15** (1950) 1172 – 1176.
7. G. A. Nikiforov, V. V. Ershov, *Russ. Chem. Rev. (Engl. Transl.)* **39** (1970) 644 – 654.
8. P. Demerseman *et al.*, *Bull. Soc. Chim. Fr.* 1971, 201 – 210.
9. D. P. Biddiscombe *et al.*, *J. Chem. Soc.* 1963, 5764 – 5768.
10. G. Bertholon, *Bull. Soc. Chim. Fr.* 1967, 2977 – 2982.
11. R. Larmatine, R. Perrin, *Bull. Soc. Chim. Fr.* 1969, 443 – 445.
12. J. Bassus *et al.*, *Bull. Soc. Chim. Fr.* 1974, 3031 – 3038.
13. G. Bertholon, R. Perrin, *Bull. Soc. Chim. Fr.* 1975, 1537 – 1544. Links
14. W. A. Pardee, W. Weinrich, *Ind. Eng. Chem.* **36** (1944) 595 – 603.
15. R. Hanley *et al.*, *J. Chem. Soc.* 1964, 4404 – 4406.
16. H. Stage, E. Müller, P. Faldix, *Erdöl Kohle* **6** (1953) 375 – 380.
17. S. Ohe: *Computer Aided Data Book of Vapour Pressure*, Data Book Publishing Co., Tokyo 1976.
18. G. Bertholon *et al.*, *Bull. Soc. Chim. Fr.* 1971, 3180 – 3187.
19. A. Ksiazczak, J. J. Kosinski, *Fluid Phase Equilib.* **55** (1990) 17 – 37.
20. E. Terres *et al.*, *Brennst. Chem.* **36** (1955) 289 – 301.
21. L. U. Erichsen, E. Dobbert, *Brennst. Chem.* **36** (1955) 338 – 345.
22. E. Terres *et al.*, *Brennst. Chem.* **36** (1955) 359 – 372.
23. T. H. Coffield *et al.*, *J. Am. Chem. Soc.* **79** (1957) 5019 – 5023.
24. G. H. Stillson *et al.*, *J. Am. Chem. Soc.* **67** (1945) 303 – 307; **68** (1946) 722.
25. C. A. Brown, *Synthesis* 1974, 427 – 428.
26. *Houben-Weyl*, **6/2**, 35 – 41.
27. A. McKillop, J.-C. Fiaud, R. P. Hug, *Tetrahedron* **30** (1974) 1379 – 1382.
28. J. B. McKinley *et al.*, *Ind. Eng. Chem. Anal. Ed.* **16** (1944) 304.
29. Ethyl Corp., DE 2 208 827, 1972 (J. C. Wollensak).
30. *Houben-Weyl*, **7/2 a**, 849 – 853.
31. O. M. Kut, U. R. Dätwyler, G. Gut, *Ind. Eng. Chem. Prod. Res. Dev.* **27** (1988) 215 – 225.
32. V. Pencev *et al.*, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **23** (1970) 571 – 574.
33. *Houben-Weyl*, **6/1 c**, 1121 – 1129.
34. D. C. Nonhebel *et al.*, *J. Chem. Res. Synop.* 1977, 12 – 16, *J. Chem. Res. Miniprint*, 0201 – 0283.
35. F. R. Hegwill, G. B. Howie, *Aust. J. Chem.* **31** (1978) 907 – 917.
36. *Houben-Weyl*, **7/1**, 152, 158.
37. Bayer AG, EP 330 036, 1989 (A. Schnatterer, H. Fiege).
38. *Houben-Weyl*, **10/1**, 576 – 584; 825 – 828.
39. *Houben-Weyl*, **5/3**, 679 – 688.
40. *Houben-Weyl*, **7/2 b**, 1413 – 1422.
41. *Houben-Weyl*, **10/1**, 1028 – 1037.
42. *Houben-Weyl*, **6/1 c**, 925 – 1019.
43. *Houben-Weyl*, **6/1 c**, 1073 – 1090.
44. *Houben-Weyl*, **6/1 c**, 1086 – 1090.
45. *Houben-Weyl*, **7/1**, 20 – 29.
46. *Houben-Weyl*, **14/2**, 193 – 230.
47. *Houben-Weyl*, **6/1 c**, 1021 – 1060.
48. M. Weinberg, *Rev. Chim. (Bucharest)* **21** (1967) 669 – 671.
49. V. A. Parfenova, V. I. Isagulyants, *Khim. Tekhnol. Topl. Masel* **1** (1974) 7 – 8.
50. B. Roach, T. Eisner, J. Meinwald, *J. Chem. Ecol.* **6** (1980) 511 – 516; *Chem. Abstr.* **93**, 12 90 24 k.

51. P. D. Swann *et al.*, *Fuel* **52** (1973) 154 – 155.
52. J. H. P. Tyman, *J. Chromatogr.* **111** (1975) 277 – 284, 285 – 292.
53. V. S. Pansare, A. B. Kulkarni, *J. Indian Chem. Soc.* **41** (1964) 251.
54. J. Mačak, P. Buryan, *Chem. Listy* **69** (1975) 457 –518.
55. M. Sy, G. Lejeune, *Chim. Ind. (Paris)* **78** (1957) 619 – 628.
56. K. F. Lang, I. Eigen, *Fortschr. Chem. Forsch.* **8** (1967) 93 – 130.
57. P. Buryan, J. Mačak, V. M. Nabivach, *J. Chromatogr.* **148** (1978) 203 – 210.
58. H. Pichler, G. Schwarz, *Brennst. Chem.* **50** (1969) 72 – 78. H. Pichler, P. Hennenberger, *Brennst. Chem.* **50** (1969) 341 – 346.
59. A. Siller, *J. Prakt. Chem. Ser. 4* **1** (1955) 209 –224.
60. V. Kusy, *Erdöl Kohle Erdgas Petrochem. Brennst. Chem.* **23** (1970) 575 – 580.
61. R. S. Aries, S. A. Savitt, *Chem. Eng. News* **28** (1950) 316 – 321.
62. R. E. Maple: *Symposium on Refining Petroleum for Chemicals*, Am. Chem. Soc. New York City Meeting, 7 – 12, Sept. 1969, pp. D 105 – 113.
63. E. O. Woolfolk, M. Orchin, *Ind. Eng. Chem.* **42** (1950) 552 – 556.
64. Y. Sugimoto, Y. Miki, S. Yamada, M. Oba, *Nippon Kagaku Kaishi* 1984, 755 – 763.
65. S. Preis, *J. Prakt. Chem. Ser. 4* **1** (1955) 157 – 171, 172 – 176, 177 – 186.
66. VEB-Leuna-Werke, DE 1 003 222, 1955 (S. Preis).
67. Yorkshire Tar Distillers, GB 736 604, 1952 (D. W. Milner, R. Flathers).
68. Gesellschaft für Teerverwertung, DE 1 127 908, 1959 (H. Wille, L. Rappen). Shell, US 2 370 554, 1941 (D. B. Luten, S. B. Thomas).
69. Shell, US 2 336 720, 1941 (A. De Benedictis, D. B. Luten).
70. J. G. M. Thorne, *Chem. Process. (Chicago)* (1970) March, 29 – 31.
71. R. E. Dean, E. N. White, D. Mc Neil, *J. Appl. Chem. (London)* **9** (1959) 629 – 641.
72. G. A. Olah (ed.): *Friedel – Crafts and Related Reactions*, vol. II/1, Interscience Publ., New York 1964.
73. *Houben-Weyl*, **6/1 c**, 925 – 1019.
74. O. N. Tsvetkov *et al.*, *Int. Chem. Eng.* **7** (1967) 104 – 113, 113 – 121.
75. P. S. Belov, K. D. Korenev, A. Y. Estigneev, *Khim. Tekhnol. Topl. Masel* **4** (1981) 58 – 61.
76. J. Klein, H. Widdecke, *Chem. Ing. Tech.* **51** (1979) 568; **53** (1981) 954 – 957.
77. M. M. Sharma *et al.*, *Ind. Eng. Chem. Prod. Res. Dev.* **29** (1990) 29 – 34, 1025 – 1031.
78. Shell, US 2 923 745, 1957 (V. W. Buls, R. S. Miller).
79. Shell, US 3 177 259, 1963; US 3 116 336, 1960 (J. L. Van Winkle).
80. J. I. De Jong, *Recl. Trav. Chim. Pays-Bas* **83** (1964) 469 – 476.
81. Consolidation Coal, US 3 461 175, 1965 (M. D. Kulik, R. J. Laufer).
82. Consolidation Coal, US 3 408 410, 1965 (R. J. Laufer, M. D. Kulik).
83. Consolidation Coal, US 3 418 380, 1965 (R. J. Laufer, M. D. Kulik).
84. Bayer, DT 944 014, 1954 (R. Stroh, R. Seydel).
85. Ethyl Corp., US 2 831 898, 1954 (G. G. Ecke, A. J. Kolka).
86. R. Stroh, R. Seydel, W. Hahn in W. Foerst (ed.): *Neuere Methoden der präparativen organischen Chemie*, vol. **II**, Verlag Chemie, Weinheim, Germany 1960, pp. 231 – 246.
87. A. J. Kolka *et al.*, *J. Org. Chem.* **22** (1957) 642 – 646.
88. G. G. Knapp *et al.*: *Proc. 7th World Petroleum Congress*, vol. **5**, Elsevier, London 1967, pp. 403 –413.
89. K. N. Kulieva, *Khim. Tekhnol. Topl. Masel* **10** (1989) 38 – 40.
90. R. Gompper, *Angew. Chem.* **76** (1964) 412 – 423.
91. Hoechst, DE-OS 2 039 062, 1970 (E. Reindl).
92. F. R. J. Willemse *et al.*, *Recl. Trav. Chim. Pays-Bas* **90** (1971) 5 – 13.
93. Koppers Comp. Inc. US 3 091 646, 1960 (G. Leston).
94. Ethyl Corp., DE 1 493 622, 1964 (T. H. Coffield, G. G. Knapp, J. P. Napolitano).
95. Shell, US 3 268 595, 1962 (K. L. Mai).
96. Shell, DE 1 154 484, 1960 (V. W. Buls).
97. Koppers Comp. Inc., US 3 267 154, 1963 (T. Hokama).
98. Koppers Comp. Inc., US 3 267 153, 1963 (G. Leston).
99. Consolidation Coal, US 3 032 595, 1958 (N. B. Neuworth, R. J. Laufer, E. P. Previc).
100. Shell, FR 1 398 153, 1964 (J. C. E. Bolle, J. M. Tomaszewski).
101. M. V. Kurashev, N. N. Korneev, *Neftekhimiya* **28** (1988) 176 – 182.
102. Ethyl Corp., EP 206 085 (D. E. Goins, S. W. Holmes, E. A. Burt) S. W. Holmes, E. A. Burt).
103. Koppers Comp. Inc., US 3 331 879, 1965 (G. Leston).
104. M.-F. Berny, *Bull. Soc. Chim. Fr.* 1969, 973 – 976.
105. Monsanto, GB 1 008 592, 1963 (B. B. Millward).
106. Koppers Comp. Inc., US 3 267 155, 1963 (G. Leston).
107. Koppers Comp. Inc., US 3 267 152, 1963 (T. Hokama).
108. Koppers Comp. Inc., US 3 346 649, 1966 (G. Leston).
109. Koppers Comp. Inc., DE-AS 1 230 432, 1964 (G. Leston).
110. Bayer, DE 1 142 873, 1961 (W. Hahn).
111. Ethyl Corp., US 3 367 981, 1963 (J. P. Napolitano).
112. I. T. Golubtschenko *et al.*, *Dokl. Akad. Nauk SSSR* **261** (1981) 891 – 894.
113. Bayer, DE 1 159 960, 1960 (W. Hahn).
114. UK Wesseling, DE-OS 2 040 228, 1970 (C. Kalav).
115. T. Yamanaka, *Bull. Chem. Soc. Jpn.* **49** (1976) 2669 – 2673.
116. E. A. Goldsmith *et al.*, *J. Org. Chem.* **23** (1958) 1871 – 1876.
117. S. Skraup, W. Beifuss, *Ber. Dtsch. Chem. Ges.* **60** (1972) 1070.
118. R. G. Anderson, S. H. Sharman, *J. Am. Oil Chem. Soc.* **48** (1971) 107 – 112.
119. Geigy (UK) Ltd., GB 1 146 173, 1966 (W. Pickels, D. R. Randell).
120. *Houben-Weyl*, **6/1 c**, 1060 – 1072.
121. J. Becht, W. Gerhardt, Akademie Wiss. DDR, Int. Tagung Grenzflächenaktive Stoffe, Berlin 1977, part 1, pp. 143 – 149.
122. *Houben-Weyl*, **6/1 c**, 202 – 208.
123. *Houben-Weyl*, **6/1 c**, 1081 – 1083.
124. Bayer AG, EP 297 391, 1988 (R. Dujardin, W. Ebert).
125. Schering-Kahlbaum, DT 467 640, 1928; DT 479 352, 1929 (H. Jordan).
126. Bayer, DT 1 105 428, 1958 (G. Schuckmann, H. Schnell).
127. *Houben-Weyl*, **6/1 c**, 502 – 550.
128. *Houben-Weyl*, **6/1 c**, 1081.
129. Kuraray, JP 50 76 033, 1973.
130. Mitsui Toatsu Chemicals, JP 55 07 731, 1978; *Chem. Abstr.* **93 (15):** 14 99 78c.
131. Bayer AG, EP 102 493, 1983 (J. Käsbauer, K. F. Wedemeyer).
132. Kuraray, JP 62 255 445, 1986 (H. Tamai, T. Kunitomi); *Chem. Abstr.* **109 (19):** 17 00 21a.
133. B. E. Leach, *J. Org. Chem.* **43** (1978) 1794 – 1797. Continental Oil Co., US 3 979 464, 1975.
134. Continental Oil Co., DE 2 811 471, 1978 (R. Poe, J. F. Scamehorn, C. R. Schupbach).
135. Continental Oil Co., US 3 985 812, 1974 (E. Del Bel, D. C. Jones, M. B. Neuworth).
136. Continental Oil Co., US 3 862 248, 1968 (D. C. Jones, M. B. Neuworth).

137. BASF, DE 1 668 874, 1968 (L. Arnold, H. Pasedach, H. Pommer).
138. BASF, DE 1 793 037, 1968 (L. Arnold, H. Pasedach, H. Pommer).
139. BASF, EP 0 123 233, 1984 (N. Goetz, H. Laas, P. Tavs, L. Hupfer, K. Baer).
140. N. G. Baranova *et al.*, *Khim.-Farm. Zh.* **22** (1968) 736 – 739; *Chem. Abstr.* **110** (13) : 114 380u.
141. Teijin Ltd., US 3 968 172, 1973/76 (Y. Ichikawa, Y. Yamanaka, T. Naruchi, O. Kobayashi, K. Sakota).
142. Y. Ichikawa *et al.*, *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1979) 373 – 375.
143. Rütgerswerke AG, DE-AS 1 254 155, 1965 (M. Froitzheim, K. F. Lang, L. Rappen, E. Schweym, J. Turowski).
144. General Electric Co., US 4 503 270, 1983 (J. J. Talley).
145. D. E. Pearson, C. A. Buehler, *Synthesis* 1971, 471 –472.
146. T. Kotinagawa, K. Shimokawa, *Bull. Chem. Soc. Jpn.* **47** (1974) 1535 – 1536.
147. Rütgerswerke AG, DE 2 346 498, 1973 (H.-D. Hausigk, G. Löhnert).
148. K. Aomura *et al.*, *Kogakubu Kenkyu Hokoku (Hokkaido Diagaku)* **76** (1975) 147 – 153; *Chem. Abstr.* **84**, 16 688r.
149. M. Inoue, S. Emoto, *Chem. Pharm. Bull.* **20** (1972) 232 – 237.
150. M. Inoue *et al.*, *Yaki Gosei Kagaku Kyokai Shi* **28** (1970) 1127 – 1132.
151. Maruzen Petrochemical Co., EP 320 936, 1988 (T. Yamagishi, T. Idai, E. Takahasi).
152. Maruzen Oil Co., JP 53 063 346, 1976 (F. Suganuma, Y. Hirose, M. Hayano, T. Nakamura).
153. Maruzen Oil Co., JP 52 106 825, 1976 (M. Sato *et al.*).
154. Maruzen Oil Co., JP 53 119 839, 1977 (F. Suganuma *et al.*).
155. Taoka Chemical Co., EP 80 880, 1982 (M. Wada, S. Maki).
156. Taoka Chemical Co., JP 59 108 730, 1982.
157. Taoka Chemical Co., JP 59 076 033, 1982.
158. L. H. Klemm, D. R. Taylor, *J. Org. Chem.* **45** (1980) 4320 – 4329.
159. Bayer DE 1 142 873, 1961 (W. Hahn).
160. F. J. Sowa, H. D. Hinton, J. A. Nieuwland, *J. Am. Chem. Soc.* **54** (1932) 2019.
161. Mobil Oil, US 4 391 998, 1981 (M. M. Wu).
162. Sh. G. Sadykhov, Ch. K. Rasulov, Yu. K. Dzhaforov, Ch. K. Salamanova, *Azerb. Khim. Zh.* 1981, no. 5, 53 – 55; *Chem. Abstr.* **96** (1982) 162 239u.
163. N. Yoneda, Y. Takahashi, C. Tajiri, A. Suzuki, *Nippon Kagaku Kaishi* 1977 no. 6, 831 – 836; *Chem. Abstr.* **87** (1977) 134 237n.
164. PCUK Ugine Kuhlmann, US 4 423 254, 1982 (G. A. Olah).
165. R. Stroh, R. Seydel, W. Hahn, *Angew. Chem.* **69** (1957) 699.
166. Ethyl Corp., US 3 766 276, 1979 (L. E. Goddard).
167. UOP, US 4 275 248, 1980 (B. E. Firth).
168. Geigy (UK), GB 1 146 173, 1966 (W. Pickles, D. R. Randell).
169. FMC Corp., US 3 936 410, 1973 (F. L. Terhune, G. A. Rampy).
170. FMC Corp., US 3 859 395, 1973 (F. L. Terhune, G. A. Rampy).
171. Bayer, DE 1 280 255, 1965 (W. Schulte-Huermann, A. Kersting).
172. FMC Corp., BE 830 497, 1975 (R. A. Aal, N. Chen, J. K. Chapman).
173. Bayer, DE 2 528 303, 1975 (W. Biedermann, H. Köller, K. Wedemeyer).
174. Bayer, DE 3 824 284, 1988 (P. Wimmer, H. J. Buysch, L. Puppe).
175. M. Nitta, *Bull. Chem. Soc. Jpn.* **47** (1974) 2360 –2364.
176. M. Nitta, *Bull. Chem. Soc. Jpn.* **47** (1974) 2897 –2898.
177. T. Yamanaka, *Bull. Chem. Soc. Jpn.* **49** (1976) 2669 – 2673.
178. R. Stroh, R. Seydel, W. Hahn in W. Foerst (ed.): *Newer Methods of Preparative Organic Chemistry*, vol. **2**, Academic Press, New York 1963, p. 337 ff.
179. Union Rheinische Kraftstoff, DE-OS 2 139 622, 1971 (E. Biller, D. Kühne).
180. Bayer, DE-OS 2 242 628, 1972 (A. Klein, K. Wedemeyer).
181. Koppers Co., US 3 992 455, 1976 (G. Leston).
182. Sumitomo Chem. Co., DE-OS 2 340 218, 1973 (H. Suda, N. Kotera, S. Hasegawa).
183. Koppers Co., US 4 046 818, 1975 (G. Leston).
184. H. Hart, E. A. Haglund, *J. Org. Chem.* **15** (1950) 396.
185. Rhein. Kampfer Fabrik, GB 325 855, 1928.
186. L. S. Charcenko, V. Zavgorodnij, *Ukr. Khim. Zh. (Russ. Ed.)* **30** (1964) 187 – 190.
187. J. Bassus, R. Perrin, *C. R. Hebd. Seances Acad. Sci. Ser. C*, **264** (1967) 1444 – 1446.
188. J. Bassus, *Bull. Soc. Chim. Fr.* 1974, 3031 – 3038.
189. G. Bertholon, C. Decoret, *Bull. Soc. Chim. Fr.* 1975, 1530 – 1536.
190. Chem. Eng. (N.Y.) **85** (1978), no. 12, 62 – 63.
191. Ajinomoto KK, JP 57 046 931, 1980; *Chem. Abstr.* **97** (1982) 164 951d.
192. Bayer DE 1 142 873, 1961 (W. Hahn).
193. Combinatul Petrochimic, Ploiesti, RO 66 499, 1978 (J. Hersccvici et al.); *Chem. Abstr.* **98** (1983) 109 290u.
194. Ethyl Corp. US 3 766 276, 1970 (L. E. Goddard).
195. Ethyl Corp. US 3 933 927, 1973 (L. E. Goddard).
196. Hüls AG, DE 3 443 736, 1984 (H. Alfs).
197. Bayer, DE 944 014, 1954 (R. Stroh, R. Seydel).
198. Nalco Chem. Co., US 4 092 367, 1976 (B. W. Bridwell, C. E. Johnson).
199. Hoechst, DE-AS 1 813 840, 1968 (J. Bohunek).
200. Progil, FR 1 336 080, FR 1 336 081, 1981 (M. E. DeGeorges, J. Berthoux).
201. Progil, DE-OS 2 034 369, 1971 (M. Berthoux, M. Jean, G. Schwachhofer).
202. Union Rheinische Kraftstoff, DE-OS 2 021 525, 1970 (C. Kalav).
203. Union Carbide, DE-AS 1 443 346, 1961 (S. Kaufmann, R. E. Nicolson).
204. B. Loev, J. T. Massengale, *J. Org. Chem.* **22** (1957) , 988 – 989.
205. BASF, DE-OS 2 526 644, 1975 (F. Merger *et al.*).
206. Chem. Werke Hüls AG, DE-OS 2 745 589, 1977 (H. Alfs, G. Boehm, H. Steiner).
207. Sh. G. Sadykhov *et al.*, *Azerb. Khim. Zh.* 1982 no. 1, 26 – 28; *Chem. Abstr.* **97** (1982) 162 490q.
208. A. Corma, H. Garcia, J. Primo, *J. Chem. Res, Synop.* 1988 no. 1, 40 – 41.
209. I. I. de Jong, *Rec. Trav. Chim. Pays-Bas* **83** (1964) 472.
210. Hitachi Chemical Co., JP 7 400 823, 1970; *Chem. Abstr.* **81** (1974) 25 364g.
211. ICI, GB 701 264, 1950 (W. Kunz).
212. ICI, GB 1 068 693, 1965 (A. S. Briggs).
213. Union Rheinische Kraftstoff, DE-OS 2 608 241, 1976 (B. Haas, E. Meisenburg).
214. Coalite & Chem. Prod., DE 1 186 873, 1961 (H. F. Bondy, F. R. Moore).
215. Pennsalt Chem. Corp., US 3 014 079, 1958 (J. F. Olin).
216. Chevron Res. Co., DE-OS 1 965 165, 1969 (G. K. Kohn).
217. A. P. Bolton *et al.*, *J. Org. Chem.* **33** (1968) 3415 –3418.
218. Goi Kasei KK, JP-Kokai 6 019 739, 1983; *Chem. Abstr.* **102** (1985) 220 560x.
219. Stauffer Chem. Co., US 4 103 096, 1977 (S. L. Giolito, S. B. Mirviss).
220. Hodogaya Chem. Co., DE-OS 1 960 747, 1969 (N. Onodera, H. Mitsuta, M. Wataya).
221. Ethyl Corp., US 4 774 368, 1987 (D. R. Brackenridge).
222. Coalite & Chem. Prod., BE 661 950, 1965; *Chem. Abstr.* **64** (1966) 6 561g.

223. Phillips Petr. Co., US 3 878 255, 1972 (J. R. Norell).
224. J. R. Norell, *J. Org. Chem.* **38** (1973) 1929.
225. R. Stroh, R. Seydel, W. Hahn, *Angew. Chem.* **69** (1957) 699.
226. Ethyl Corp., US 3 355 504, 1967 (T. H. Coffield, G. G. Knapp, J. P. Napolitano).
227. Maruzen Petrochemical Co., JP-Kokai 6 150 935, 1984; *Chem. Abstr.* **105** (1986) 114 725g.
228. D. R. Stevens, *Ind. Eng. Chem.* **35** (1943) 655 – 660.
229. J. K. Gehlawat, M. M. Sharma, *J. Appl. Chem.* **20** (1970) 93 – 98.
230. Gulf Res. & Dev., US 2 560 666, 1948 (D. R. Stevens, R. S. Borrman).
231. ICI, DE-OS 2 423 356, 1973 (J. Atkinson, D. Ball).
232. Koppers Co., US 3 091 646, 1960 (G. Leston).
233. Koppers Co., US 3 346 649, 1963 (G. Leston).
234. Koppers Co., DE-AS 1 230 432, 1963 (G. Leston).
235. *Houben-Weyl*, **E 4**, vol. 6/1 c, 959.
236. Sumitomo Chem. Co., GB 1 296 179, 1972 (K. Kamoshita, S. Nakai).
237. C. L. Zundel, L. Choron, DE-OS 1 518 460, 1963.
238. Bayer, DE-OS 2 602 149, 1976 (H. Fiege *et al.*).
239. Shell, US 2 841 622, 1957 (D. G. Norton, F. C. Davis).
240. Shell, US 2 841 623, 1957 (D. G. Norton, R. C. Morris).
241. Sterlitamak Petrochem., GB 1 512 941, 1977 (N. V. Zakharova *et al.*).
242. Novokuib. Giprokauchuka, DE-OS 2 749 278, 1977 (N. V. Zakharova *et al.*).
243. Ethyl Corp., US 3 919 333, 1974 (J. C. Wollensak).
244. Shell, GB 557 519, 1942 (D. B. Luten, A. DeBenedictis).
245. ICI, GB 582 057, 1945 (H. A. Basterfield).
246. Chem. Werke Hüls AG, DE 3 401 343, 1983 (W. Otte, R. Nehring, M. ZurHausen).
247. O. M. Kut, U. R. Daetwyler, G. Gut, *Ind. Eng. Chem. Res.* **27** (1988) no. 2, 215 – 225.
248. F. F. Wolny, J. J. Lamb, *Elastomerics* **116** (1984) no. 4, 40 – 43.
249. C. D. Gutsche, B. Dhawan, K. H. No, R. Muthukrishan, *J. Am. Chem. Soc.* **103** (1981) no. 13, 3782 – 3792.
250. B. Dhawan, S. Chen, C. D. Gutsche, *Makromol. Chem.* **188** (1987) no. 5, 921 – 950.
251. Konebo Ltd., JP-Kokai 6 399 031, 1986 (Y. Kondo *et al.*); *Chem. Abstr.* **109** (1988) 210 747u.
252. Shell, US 3 052 728, 1962 (R. C. Morris, A. L. Rochlin).
253. N. P. Neureiter, *J. Org. Chem.* **28** (1963) no. 12, 3486.
254. J. R. Geigy AG, FR 1 343 301, 1963 (E. A. Meier, M. Dexter).
255. Ciba Geigy AG, EP 186 629, 1986 (S. D. Pastor, J. D. Spivack, P. Odorisio).
256. H. J. Florian, H. M. Schian, R. E. Tiller in: *Tagungsbericht Verband deutscher Betriebs- und Werksärzte e.V.*, Gentner-Verlag, Stuttgart 1989, pp. 267 – 274.
257. C. D. Calnan, *Proc. R. Soc. Med.* **66** (1973) no. 3, 258 – 260. Links
258. A. V. Topchiev, J. M. Paushkin, M. V. Kurosev, *Dokl. Akad. Nauk SSSR* **130** (1960) 559; *Chem. Abstr.* **54** (1960) 10 921d.
259. Ethyl Corp., US 3 933 927, 1973 (L. E. Goddard).
260. V. I. Isagulyants, N. A. Slavskaja, *Z. Prikl. Khim. (Leningrad)* **33** (1960) 953; *Chem. Abstr.* **54** (1960) 16 414a.
261. I. F. Radzevenchuk, *Zh. Org. Khim.* **1** (1965) , 1017 – 1020; *Chem. Abstr.* **63** (1965) 11 407d.
262. I. F. Radzevenchuk, E. Koltsova, *Zh. Obshch. Khim.* **38** (1968) 204; *Chem. Abstr.* **69** (1968) 16 019y.
263. Mitsubishi Petrochemical, JP-Kokai 62 153 235, 1985 (M. Imanari, H. Iwane, T. Sugawara); *Chem. Abstr.* **108** (1988) 21 490y.
264. O. N. Tsevtkov *et al.*, *Int. Chem. Eng.* **7** (1967) 104 – 121.
265. V. A. Soldatova *et al.*, *Azerb. Khim. Zh.* 1985 no. 4, 40; *Chem. Abstr.* **105** (1986) 190 578b.
266. G. G. Trigo, *Chim. Ind. (Paris)* **86** (1961) 549 – 556.
267. Centre de Technologie Chimique, FR 1 209 863, 1956.
268. BASF, FR 1 419 289, 1964 (E. Rotter, H. Buelow).
269. Nopco Chem. Co., US 3 168 577, 1961 (R. Weinstein, I. M. Rose, W. R. Christian).
270. Jefferson Chem. Co., US 3 000 964, 1953 (J. G. Milligan).
271. Thompson-Hayward Chem. Co., GB 1 142 233, 1966.
272. Standard Oil, US 2 655 544, 1953 (G. M. McNulty, T. Cross).
273. Soc. des Produits Chim. du Sidorbe-Sinova, FR 2 228 749, 1973.
274. Chemische Werke Hüls AG, DE-AS 2 346 273, 1973 (H. Alfs *et al.*).
275. California Research Corp., US 2 732 408, 1953 (J. K. Foote).
276. Calumet Petrochem., US 4 055 605, 1977 (C. L. Jarreau).
277. Agfa-Gevaert, DE-OS 2 440 678, 1974 (A. Klein *et al.*).
278. Nippon Zeon Co., DE-OS 2 527 402, 1975 (K. Goto, T. Natsumme, H. Asai).
279. *Chem*, Mark. Rep. (1978, 17.04.) 7, 36.
280. L. McMaster, W. M. Bruner, *Ind. Eng. Chem.* **28** (1936) 505 – 506.
281. W. F. Short, M. L. Stewart, *J. Chem. Soc.* 1929, 553.
282. P. McLaughlin, V. M. Greedom, B. Miller, *Tetrahedron Lett.* 1978, 3537 – 3540.
283. Bayer, DE 2 804 215, 1978 (H. J. Buysch, M. Matner, H. Freese).
284. N. N. Alekseev *et al.*, *Zh. Prikl. Khim. (Leningrad)* **46** (1973) 218 – 220; *Chem. Abstr.* **78** (1973) 110 716.
285. Ya. M. Paushkin *et al.*, *Neftekhimiya* **9** (1969) 842 –847; *Chem. Abstr.* **72** (1970) 78 572.
286. Naucno-Issledovatelskij Institut, DE-OS 1 543 791, 1966 (A. A. Grinberg *et al.*).
287. Goodyear, DE-OS 2 600 204, 1975 (W. S. Hollingshead),
288. J. A. M. Laan, J. P. Ward, *Chem. Ind. (London)* 1987 no. 1, 34 – 35.
289. Maruzen Oil Co., DE-OS 2 608 407, 1976 (K. Kanezaki).
290. Maruzen Oil Co., DE-OS 2 637 923, 1976 (M. Sato *et al.*).
291. Bayer, DE 1 235 894, 1958 (H. Krimm, H. Schnell).
292. J. Kahovec *et al.*, *Collect. Czech. Chem. Commun.* **36** (1971) 1896 – 1994.
293. H. Schnell, H. Krimm, *Angew. Chem.* **75** (1963) 662.
294. Mitsui Toatsu Chem. Inc., DE-OS 2 438 432, 1973 (K. Mimaki *et al.*).
295. *Houben-Weyl*, **E 4, vol. 6/1 c**, 986 – 990.
296. F. B. Gershanov et al., SU 1 191 445, 1983; *Chem. Abstr.* **104** (1986) 224 590d.
297. J. H. P. Tyman, *J. Chromatogr.* **111** (1975) 277 –292.
298. Rütgerswerke AG, DE-OS 2 208 253, 1972 (E. Pastorek, H. Miele, W. Orth).
299. Bayer, DE-OS 2 304 588, 1973 (A. Klein, K. Wedemeyer, J. Thies).
300. Bayer, DE-OS 2 319 079, 1973 (A. Klein, K. Wedemeyer, J. Thies).
301. Bayer, DE-OS 2 356 813, 1973 (A. Klein, K. Wedemeyer).
302. *Chem. Eng. (N.Y.)* **85** (1978) 99.
303. *Ullmann*, 4th ed., **18**, 219 – 222.
304. J. Tsuji, H. Takayanagi, *J. Am. Chem. Soc.* **96** (1974) 7349 – 7350. Links
305. C. J. Pedersen, *Science (Washington, D.C.)* **241** (1988) 536 – 540.
306. P. Buryan, J. Macak, *J. Chromatogr.* **150** (1978) 246 – 249.
307. Ube Industries, JP-Kokai 47 39 039, 1971 (T. Nagaoka *et al.*).
308. J. Varagnat, *Ind. Eng. Chem. Prod. Res. Dev.* **15** (1976) 212 – 215.

309. P. Maggioni, F. Minisci, *Chim. Ind. (Milan)* **59** (1977) 239 – 242.
310. Ube Industries, US 4 078 006, 1976, JP-Kokai 50 130 727, 1974 (S. Umemura, N. Takamitu, T. Hamamoto, N. Kuroda).
311. T. Hamamoto, N. Kuroda, N. Takamitu, S. Umemura, *Nippon Kagaku Kaishi* 1980, 1850 – 1854.
312. Ube Industries, JP-Kokai 52 118 436, 1976 (S. Umemura, N. Takamitu, T. Hamamoto, N. Kuroda).
313. Bayer AG., US 4 053 523, 1975 (H. Seifert, H. Waldmann, W. Schwerdtel, W. Swodenk).
314. Degussa AG, US 4 618 730, 1984 (K. Drauz, A. Kleemann).
315. Agency of Industrial Science and Technology,Oxirane Chemical Co. Ltd., DE 2 364 181, 1973 (J. Imamura, M. Ando, K. Sasaki, T. Iio).
316. Mitsubishi Kasei Co., Ltd., JP-Kokai 58 55 439, 1981 (T. Maki, K. Murayama).
317. *Chem. Mark. Rep.* **235** (1989) Jan. 23, 4.
318. *Chem. Mark. Rep.* (1990) May 7, 33.
319. M. A. Liebert, *J. Am. Coll. Toxicol.* **5** (1986) 123 – 164.
320. N. I. Sax: *Dangerous Properties of Industrial Materials*, 6th ed., Van Nostrand Reinhold Co., New York 1984, p. 2339.
321. A. Critechlow, R. D. Haworth, P. L. Pauson, *J. Chem. Soc.* 1951, 1318.
322. P. Gorecki, W. Kuran, *J. Polym. Sci. Polym. Lett. Ed.* **23** (1985) 299.
323. Przedsiebiorstwo przemyslowo-Handlowe “Polskie Odczynniki Chemiczne”, PL 83 989, 1979; PL 87 534, 1979 (C. Osnowski).
324. M. T. Shipchandler, C. A. Peters, C. D. Hurd, *J. Chem. Soc. Perkin Trans. 1* 1975, 1400.
325. IMC Chemical Group, Inc., US 4 046 817, 1977 (M. T. Shipchandler).
326. Degussa AG, DE 3 607 924, 1987 (G. Prescher, G. Ritter, H. Sauenstein).
327. Mitsubishi Chem. Ind. Ltd., DE 2 445 336, 1975 (H. Obara, J. Onodera, A. Matsukuma, K. Yoshida).
328. Fisons Ltd., DE 2 627 874, 1979 (D. Baldwin, P. S. Gates).
329. Fisons Ltd., EP 25 659, 1981 (J. F. Harris).
330. Fisons Ltd., US 4 268 694, 1981 (J. F. Harris, B. J. Magill).
331. Fisons Ltd., EP 13 085, 1980 (J. F. Harris).
332. Mitsubishi Chem. Ind. Ltd., EP 31 530, 1981 (T. Maki, K. Murayama).
333. NIOSH: “Registry of Toxic Effects of Chemical Substances”, Cincinnat, Ott, 1985 – 1986.
334. R. L. Lautz, E. Michel, *Bull. Soc. Chim. Fr.* 1961, 2402.
335. E. B. Vliet, *Org. Synth. Coll. Vol.* **1** (1964) 317.
336. Degussa AG, DE 3 607 924, 1986 (G. Prescher, G. Ritter, H. Sauerstein).
337. R. J. Hight, T. J. Batterham, *J. Org. Chem.* **29** (1964) 475.
338. M. L. Kastens, J. F. Kaplan, *Ind. Eng. Chem.* **42** (1950) 402.
339. Mitsui Petrochem. Ind. Ltd., JP-Kokai 83 150 529, 1983.
340. Mitsui Petrochem. Ind. Ltd., EP 88–302 689, 1988.
341. AKZO GmbH, DE 2 621 431, 1976 (H. Zengel, M. Bergfeld).
342. AKZO GmbH, DE 2 502 429, 1975 (H. G. Zengel, M. Bergfeld).
343. Kalle AG, DE 1 195 327, 1965 (S. Pietsch).
344. AKZO GmbH, DE 2 840 597, 1980 (R. Zielke, H. Maegerlein).
345. Andeno N. V., DE 2 231 005, 1973 (A. J. J. Hendrickx, N. A. De Heij).
346. H. Schnell: *Chemistry and Physics of Polycarbonates Polymer Reviews.*, vol. **9**, Wiley Interscience, New York 1964, pp. 77 – 98.
347. G. Bier, *Polymer* **15** (1974) 527 – 535. K. Hazama, *Jpn. Plast. Age* **14** (1976) 39 – 44.
348. H. Kittel: *Kunststoffjahrbuch*, 10th ed., Pansegrau Verlag, Berlin 1968, p. 10.
349. Bayer, DE-AS 1 105 428, 1958.
350. H. Schnell, H. Krimm, *Angew. Chem.* **75** (1963) 662 – 668.
351. A. Burawoy, J. T. Chamberlain, *J. Chem. Soc. (London)* **1949**, 626. Bayer, DE-AS 1 768 749, 1968.*Houben-Weyl*, vol. VI/1 c, p. 1081 ff.
352. Bayer, DE-OS 2 645 020, 1976.
353. *Houben-Weyl*, vol. **VI/1 c**, p. 995, 1021 – 1060; vol. **XIV/2**, p. 195 ff.
354. A. Burawoy, J. T. Chamberlain, *J. Chem. Soc. (London)* 1949, 624 – 626.
355. Consolidation Coal, US 3 027 412, 1959.
356. Ethyl Corp., US 2 807 653, 1955.
357. Shell, Int. Res., BE 593 606, 1961.
358. A. H. Filbey, T. H. Coffield, *J. Org. Chem.* **22** (1957) 1435.
359. Goodyear Tire & Rubber Co., DE-AS 1 495 985, 1964.
360. Union Carbide Corp., US 2 858 342, 1955.
361. Esso Res. Eng. Co., DE-AS 1 068 270, 1957.
362. Rhône-Poulenc, DE-AS 1 071 713, 1958.
363. Bayer, DE-AS 1 051 864, 1953.
364. Bayer, US 3 689 572, 1966.
365. Goodrich Co., US 2 468 982, 1946.
366. Dow, DE 905 977, 1950.
367. Bayer, DE-OS 2 204 380, 1972.
368. Union Carbide Corp., US 3 419 624, 1964. BASF, DE-OS 2 050 800, 1970.
369. Union Carbide Corp., DE 1 056 620, 1956.
370. Bayer, US 3 808 279, 1971.
371. Bayer, BE 611 184, 1961.
372. Dow, DE-AS 1 244 796, 1962.
373. Dow, US 2 191 831, 1938.
374. Hooker Chem. Corp., DE-AS 1 238 038, 1959.
375. Dow, US 2 623 908, 1951.
376. Bataafsche, DE-AS 1 030 836, 1953.
377. Hercules Powder, US 3 172 916, 1960.
378. Union Carbide Corp., FR 1 237 656, 1959; DE-AS 1 242 237, 1958; R. A. Reinicker, B. C. Gates, *AIChE J.* **20** (1974) no. 5, 933 – 940.
379. Dow, DE-OS 2 164 339, 1971. *Eur. Chem. News* 1965, July 16, 38, 40.
380. Hooker Chem. Corp., DE-AS 1 254 637, 1963.
381. Gulf Oil Canada Ltd., US 3 493 622, 1967. Shell Dev. Co., US 2 845 464, 1954.
382. Koningklijke Zwavelzuurfabrieken, DE-AS 1 272 302, 1961.
383. Union Carbide Corp., DE-AS 1 244 796, 1962.
384. Hibernia, DE-AS 1 025 418, 1953.
385. J. Bussink, *Kunststoffe* **66** (1976) 600.
386. *Kunststoffhandbuch*, vol. **8**, Hanser Verlag, München 1973, pp. 258, 261, 283.
387. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, vol. **I/1**, Verlag Heenemann, Berlin 1971, pp. 355 –357.
388. *Kunststoffhandbuch*, vol. **10**, Hanser Verlag, München 1973, pp. 53, 109, 124.
389. General Electric, DE-OS 2 243 226, 1972. Bayer, DE-OS 2 148 598, 1971.
390. Bayer, DE-OS 2 354 533, 1973.
391. Dow, US 2 538 725, 1949; US 2 535 015, 1948.
392. Dr. H. Thomae, US 2 753 351, 1953.
393. G. Bornmann, A. Loeser, *Arzneim. Forsch.* **9** (1959) 9.
394. R. Willstätter, L. Kalb, *Chem. Ber.* **38** (1905) 1235. P. P. T. Sah, *Re. Trav. Chim. Pays-Bas* **59** (1940) 454. N. M. Cullinane, C. G. Davies, G. J. Davies, *J. Chem. Soc.* 1936,

1435.
395. N. M. Cullinane, C. G. Davies, *Re. Trav. Chim. Pays-Bas* **55** (1936) 881.
396. ICI, DE-AS 1 913 182, 1969.
397. Bayer, DE-AS 2 049 809, 1970.
398. Agency of Industrial Science and Technology, Tokio, DE-AS 2 211 721, 1972.
399. Monsanto Chem. Co., US 2 368 361, 1942.
400. ICI, GB 922 679, 1960.
401. Rütgerswerke, GB 529 936, 1939.
402. M. Tashiro, H. Watanabe, O. Tsuge, *Org. Prep. Proced. Int.* **6** (1974) no. 3, 117 – 122.
403. General Electric, US 3 631 208, 1971. A. S. Hay, *J. Org. Chem.* **34** (1969) 1160 – 1161.
404. Hoechst, DE-AS 1 148 326, 1968. H. Musso, W. Steckelberg, *Justus Liebigs Ann. Chem.* **693** (1966) 187 – 196. J. Bourdon, M. Calvin, *J. Org. Chem.* **22** (1957) 101 – 116.
405. *Chem. Abstr.* **52**, 10 991g (1958).
406. Sun Oil, US 3 281 435, 1962. C. R. H. J. De Jonge, H. M. van Dort, L. Vollbracht, *Tetrahedron Lett.*, 1970. 1881 – 1884.
407. UCC, US 3 322 838, 1963. General Electric, NL 6 410 238, 1963.
408. UBE Industries Ltd., JP 73 36 152, 73 34 153, 1973.
409. *WHO Tech. Rep. Ser.* 1965, 309.
410. H. C. Hodge *et al.*, *J. Pharmacol. Exp. Ther.* **104** (1952) 202. Links F. C. Mac Intosh, *Analyst (London)* **70** (1945) 334.
411. O. Pauli, G. Franke: *Biodeterior*, Mater. Proc. Int. Biodeterior. Symp. 2nd (1972) 52 – 60.
412. Bayer, DE-OS 1 719 417, 1965.
413. G. D. Kharlampovich, V. N. Vinogradova, SU 197 613, 1967.
414. Bayer AG, DE-OS 2 655 826, 1976 (R. Neumann, H.-H. Schwarz, K.-H. Arnold).
415. Union Carbide, EP 73 492, 1983 (L. M. Maresca, M. Matzner).
416. National Distillers and Chemical Corp., DE-OS 3 312 684, 1982 (E. G. Harris).
417. Dow Chemical Co., US 4 613 682, 1986 (K. A. Eickholt).
418. Brichima SpA., BE 890 589, 1980.
419. Brichima SpA., EP-A 151 392, 1985 (P. Maggione, F. Minisci, M. Correale).
420. Bayer AG, EP-A 52 314, 1982 (G. Buettner, A. Jufat, U. Allenbach, M. Lenthe).
421. Brichima SpA., DE-OS 2 703 640, 1976 (P. Maggione).
422. Takasago Perfumery, JP 84 46 949, 1976.
423. W. Paulus, H. Genth in T. A. Oxley, S. Barry (eds.): *Biodeterioration 5*, John Wiley & Sons, New York 1983, pp. 701 – 712.
424. D. Liu, *Toxicity Assessment* **4** (1989) 115 – 127.
425. W. W. Cocker, US 2 350 677.
426. E. Klarmann *et al.*, *J. Am. Soc.* **55** (1939) 2576 – 2583.
427. R. S. Law, *J. Soc. Chem. Ind. London Trans. Commun.* **60** (1941) 66.
428. VEB Jenapharm, DE-OS 2 706 747, 1977.
429. G. F. Reddish: *Antiseptics*, Disinfectants, Fungicides and Sterilization, Lea & Febiger, Philadelphia 1954, p. 250.
430. K. H. Wallhäusser: *Praxis der Sterilisation – Desinfektion – Konservierung*, Georg Thieme, Stuttgart 1984, pp. 416 – 420.
431. W. Paulus in *Microbicides for the Protection of Materials – a Handbook*, Chapman and Hall, London 1993, pp. 141 – 198.
432. H. P. Fiedler in *Lexikon der Hilfsstoffe für Pharmazie*, Kosmetik und angrenzende Gebiete, Editio Cantor, Aulendorf 1989.

[Top of Page]

Related Articles

, , Cresols and Xylenols, Hydroquinone, and Resorcinol are separate keywords. Chlorohydroxybenzenesulfonic Acids and Aminohydroxybenzoic Acids , Nitrophenols , Vanillin Flavors and Fragrances

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Phenolic Resins

WOLFGANG HESSE, Hoechst AG, Werk Kalle – Albert, Wiesbaden, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 4
3. Raw Materials 5
3.1. Phenols 5
3.2. Aldehydes 5
3.3. Catalysts 6
4. Production 6
4.1. Novolacs 7
4.2. Resols 7
4.3. Phenolic Resins Modified by Natural Resins 9
4.4. Phenolic Resins with Special Properties 9
4.5. Wastewater 10
5. Storage and Transportation 10
6. Testing and Analysis 10
7. Uses 11
7.1. Novolacs 11
7.1.1. Cross-Linked Novolacs 11
7.1.2. Novolacs without Cross-Linking . . . 12
7.2. Resols 13
7.2.1. Water-Soluble Resols 13
7.2.2. Resols in Organic Solvents 14
7.2.3. Alkylphenol Resols 14
7.3. Phenolic Resins Modified by Natural Resins 14
7.4. Waterborne Paints, Phenol Ether Resins 15
7.5. High-Temperature Coking 15
8. Economic Aspects 15
9. Toxicology and Occupational Health 15
10. References 16

1. Introduction

History. The first synthetic resins and plastics were produced by polycondensation of phenol with aldehydes. In 1872 VON BAYER first reported the reaction between phenol and aldehydes. The resins formed were, however, not of industrial and certainly not of scientific interest. The phenol resin condensation was used industrially in 1902 by BLUMER for the production of novolacs, which served as a substitute for shellac.

In 1909 BAEKELAND made the first plastics. He carried out the polycondensation of phenol and formaldehyde to form cross-linked thermosets over several steps.

Besides the production of plastics, phenolic resins were sought as a replacement for natural resins, which were then used on a large scale for oil varnishes. In 1910 oil-soluble modified phenolic resins were produced by BEHRENDS by polycondensation of phenols, formaldehyde, and rosin.

Between 1928 and 1931 phenolic resins gained increased importance through the treatment of resols with fatty oils to give air-drying varnishes. The main problem, an inadequate compatibility of phenolic resins with other varnish raw materials, was solved by using alkylphenols or by etherification of the hydroxymethyl groups of resols with monohydric alcohols.

These varnish applications and the use of phenolic resins as thermosets and electrical insulating materials were the main application areas. However, other polycondensates and, above all, polymers increasingly limited the market for phenolic resins from the mid 1930s onwards. Theoretical work on the constitution and mechanism of formation of phenolic resins was being carried out at that time by VON EULER, HULTZSCH, MEGSON, ZIEGLER, and others, which led to the development of new application areas for phenolic resins, i.e., as adhesives, printing-ink binders, waterborne paints, temperature-resistant binders, and laminated plastics.

The industrial development of phenolic resins is still continuing despite the long history. Their importance is likely to remain considerable because the raw materials can be obtained at reasonable cost from both petroleum and coal. Phenolic resins can be used as raw materials for synthetic fibers and in photoresists for the production of microchips which characterizes the continuing relevance of this group of resins [7].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_371

Classification. Phenolic resins are polycondensation products of phenols and aldehydes, in particular phenol and formaldehyde. DIN 16 916 (ISO 10 082) attempts to define the relevant terms and properties.

The ring hydrogens in the *para*- and both *ortho*-positions relative to the hydroxyl group can react with formaldehyde and thus cross-link to form a three-dimensional network. If at least one of these three positions bears a substituent other than hydrogen, cross-linking is no longer possible and comparatively low molecular mass compounds are formed as the end products of the polycondensation.

Phenolic resins are classified as novolacs and resols. In *resols* the polycondensation is base-catalyzed and has been stopped deliberately before completion. Characteristic functional groups of this class of resins are the hydroxymethyl group and the dimethylene ether bridge. Both are reactive groups. During processing the polycondensation can be restarted by heating and/or addition of catalysts i.e., resols are self-cross-linking. In the case of *novolacs* the polycondensation is brought to completion. The molecular growth of these thermoplastic synthetic resins is limited by addition of a substoichiometric amount of the aldehyde component. Novolacs are phenols that are linked by alkylidene (usually methine) bridges, without functional groups (apart from the phenolic hydroxyl groups), and cannot cure on their own. However, novolacs can be cross-linked by addition of curing agents, such as formaldehyde or hexamethylenetetramine, and give end products similar to resols.

The classification of phenolic resins into novolacs and resols is only strictly valid if phenols which are trifunctional towards formaldehyde are used as starting material, because resols from bifunctional phenols cannot cross-link by themselves. Nevertheless, the polycondensates from substituted phenols are differentiated according to their characteristic groups as *alkylphenol novolacs* (alkylidene bridge) or *alkylphenol resols* (hydroxymethyl group, dimethylene ether bridge).

The third large group are *phenolic resins modified by natural resins*. Besides phenolic hydroxyl groups, they contain double bonds, ester links, and carboxyl groups.

Novolacs. The first step in phenolic resin polycondensation is always the electrophilic attack of a carbonyl compound (generally formaldehyde) on the *para*- and/or *ortho*-positions of a phenol molecule (acid catalysis, Eq. 1) or a phenolate anion (base catalysis, Eq. 2).

Since hydroxymethyl-substituted phenols are more reactive than phenol itself, the hydroxymethylation continues. The hydroxymethyl compounds formed are unstable in acidic medium and are rapidly converted into compounds linked by methylene bridges (Eq. 3). This reaction also occurs in both the *ortho*- and *para*-positions. In basic media hydroxymethyl groups can be stable. At higher temperatures, however, they react with the formation of methylene bridges according to Equation (3). To hinder polyalkylation by cross-linking which would make further processing more difficult or impossible, less than one mole of formaldehyde must be added per mole of phenol.

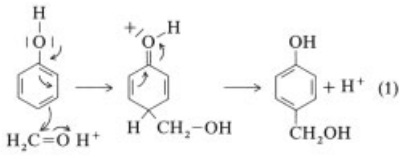

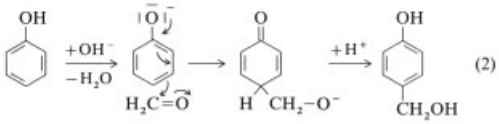

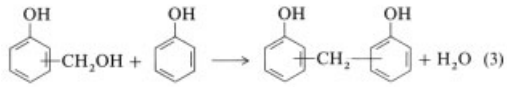

Novolacs are sometimes used as chemically unmodified synthetic resins. Their main application is based, however, on their capability to undergo cross-linking with hexamethylenetetramine. The reaction occurs at ca. 150 °C according to Equation (4) [8].

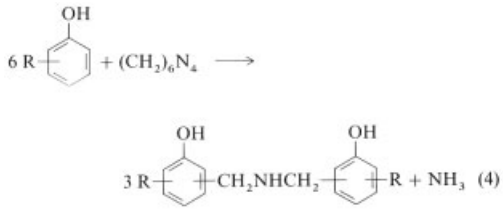

Resols. In a strongly acidic medium, hydroxymethyl groups are rapidly converted into methylene bridges. Therefore, the synthesis of resols can only be catalyzed by bases or salts of weak acids or bases. In analogy to the novolacs, the hydroxymethyl groups are formed in the *ortho*- or *para*-positions (Eq. 2). At temperatures above ca. 40 °C the hydroxymethyl groups can react to form dimethylene ether bridges with elimination of water, according to Equation (5). No catalyst is needed for this reaction. The dimethylene ether bridges formed are more stable in the *ortho*- than in the *para*-position. They can be converted into methylene bridges with elimination of formaldehyde (Eq. 6). The formaldehyde liberated is then available for the formation of new hydroxymethyl groups, if a suitable catalyst is present.

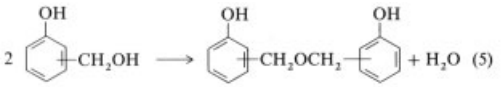

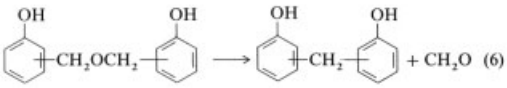

The hydroxymethyl groups of resols can also condense directly with other phenol molecules according to Equation (3). In resols therefore, three different types of formaldehyde-derived moieties occur: (1) relatively stable methylene bridges (the only type of linkage in novolacs), (2) hydroxymethyl groups, which are both capable of condensation reactions, and (3) dimethylene ether bridges.

The structure of resols depends not only on the choice of raw materials and their molar ratios, but also on the temperature of formation, concentration of raw materials, presence or absence of solvents, type of catalyst, and catalyst concentration. These parameters determine the structure to a much greater extent than in the case of novolacs.

The catalyst-containing synthetic resins formed are ready to be used industrially. However, the catalyst can also be removed, either before application or during production, because no catalyst is required for the conversion of hydroxymethyl groups into dimethylene ether bridges, a rise in temperature being sufficient.

To affect compatibility (i.e., miscibility) properties the hydrophilic character of resols can be lowered by etherification of the hydroxymethyl group with alcohols according to Equation (7).

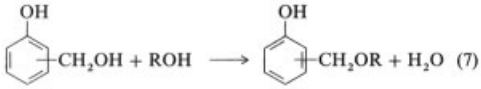

Modified Phenolic Resins. Rosin, a natural resin (→Resins, Natural), contains abietic acid and its double bond isomers as main components. Resols react with the unsaturated centers of these resin acids to form polycarboxylic acids with methylene bridges [9].

These condensation products from rosin and phenol – formaldehyde resin are known as albertol acids. They can be converted by esterification with polyols, or by salt formation, into higher molecular mass products which are readily soluble in nonpolar solvents but can release the solvent rapidly.

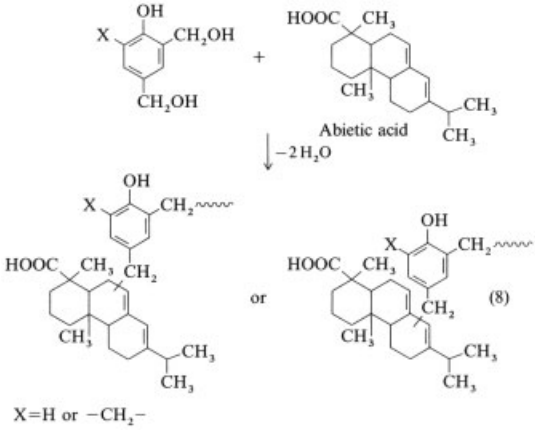

Albertol acids are also obtained by direct condensation of rosin, phenols, and formaldehyde.

Resols also undergo analogous reactions with other natural or synthetic unsaturated compounds such as fatty oils, rubbers, and polymer oils. A limited increase in the molecular mass or cross-linking can thus be achieved. Whether a particular reaction can be carried out successfully depends on the ratio of the rates of auto-condensation of the starting materials to co-condensation with the other reaction partners, but particularly on their mutual compati-

bility. Phenolic resins are therefore often classified as "water-soluble", "alcohol-soluble", "oil-soluble" etc.

Compatibility of resols with other components can be produced in many ways, e.g., (1) by using ring-alkylated phenols as the raw materials; (2) by etherification of the hydroxymethyl groups with alcohols; (3) by co-condensation of the resol with natural resins; or by a combination of these measures.

2. Physical Properties

Phenolic resins are yellow to brown in color, and the coloration can be very intense. Pale phenolic resins become colored immediately after production during storage or processing. The coloration is less intense only in the case of phenolic resins from *para*-alkyl-substituted phenols. Characteristic UV absorption maxima lie at 254 nm and 280 nm. The IR spectra of phenolic resins are described in [10].

Phenolic resins which are not cross-linked are commercially available as solids or solutions. For particular applications, e.g., in thermosets, the polycondensation can be driven so far that the resins are no longer soluble but can only be swelled by organic solvents. The *softening point* of solid resins can be determined by the capillary melting point according to DIN 53 244, by the ring and ball method, or similar procedures. These temperatures are not melting points in the thermodynamic sense. They characterize a lowering of viscosity caused by a rise in temperature, as a result of which previously crushed resin particles can be observed to coalesce or another change in form occurs.

The *miscibility* with solvents, usually described as "solubility," depends on the structure of the resin ranging from solubility in water to that in naphtha. Resins often have limited miscibility with certain solvents.

The *viscosity* of phenolic resins or their solutions is measured at high concentrations, e.g., in 30 – 80 % solution. Estimation of the degree of condensation from the viscosity is only possible for the same type of resin, because the molecular structure, in particular the presence of hydroxymethyl groups and dimethylene ether bridges, has a great effect on the viscosity. Usually, the results obtained do not depend on the type of viscometer used. Soluble phenolic resins have a broad *molecular mass distribution*: determination (usually by gel permeation chromatography) gives values of more than 50 000 depending on the type of phenol monomer.

Cross-linked phenolic resins are hard substances which only have a small fracture strain and cannot be melted. Decomposition reactions begin at 120 – 250 °C, depending on the molecular structure. There are, however, also types of phenolic resin (phenolic ether resins) which are stable for some time up to 300 °C.

Phenolic resins can be plasticized. Their compatibility with plasticizers can be adjusted by introduction of hydrophilic or hydrophobic groups.

3. Raw Materials

3.1. Phenols

Phenol (→ Phenol). Unsubstituted phenol is the most important of this group. Phenol produced from tar has a greater tendency towards coloration than synthetic phenol. This is, however, not important because phenolic resins tend to darken during production, storage, and use.

Cresols (→ Cresols and Xylenols). Resins produced from cresols are more compatibile to hydrophobic substances than those produced from phenol. Cresols are therefore important principally for the production of varnish resins. The cresol fractions obtained in tar distillation are mixtures of isomers which contain all three cresol isomers and usually also 2,6-dimethylphenol. For the application of these fractions, their content of *meta*-cresol, which is the most reactive component, is particularly important. As only the *m*-position is substituted, *meta*-cresol is trifunctional towards formaldehyde. Of the pure isomers, *ortho*- and *para*-cresols are the usual raw materials for special applications.

para-Alkylphenols (→ Phenol Derivatives). *para*-Alkylphenols are bifunctional towards formaldehyde, i.e., resols produced from *para*-alkylphenols cannot cross-link directly on heating. If a mixed condensation of different phenols with formaldehyde is carried out, their different reaction rates towards formaldehyde must be taken into account. 4-*tert*-Butylphenol (*mp*

99 °C), 4-isooctylphenol (*mp* 72 – 74 °C), 4-hydroxybiphenyl (*mp* 167 °C), and 4-nonylphenol (liquid at room temperature) are mainly used. Other alkylphenols such as isopentylphenol, cyclohexylphenol, dodecylphenol, and cashew oil are also sometimes used. The latter mainly contains phenols with C_{14}-alkenyl substitutents in the *meta*-position.

[*2,2-bis(4-Hydroxyphenyl)propane*] (diphenylolpropane, bisphenol A, Dian, *mp* 155 – 156 °C) is tetrafunctional towards formaldehyde. Polycondensation of bisphenol A and formaldehyde gives relatively pale resins.

Resorcinol (1,3-dihydroxybenzene, *mp* 117 °C) reacts particularly rapidly with formaldehyde and gives resins which can be cross-linked even in an alkaline medium at room temperature.

Other phenol derivatives which can be used as starting materials are, for example, phenol ethers, phenols containing carboxyl groups, and phenoxyacetic acid.

3.2. Aldehydes

Formaldehyde. The most important aldehyde for the production of phenolic resins is formaldehyde (Formaldehyde), which is used as an aqueous solution at concentrations between 30 and 60 %. Besides formaldehyde, the aqueous solutions contain formic acid (oxidation product of formaldehyde) and methanol. Methanol is added to inhibit formation of higher molecular mass polyoxymethylenes, which would precipitate from the aqueous solution. Methanol can intervene in the polycondensation reaction between phenol and formaldehyde by the formation of ethers, acetals, and hemiacetals and promotion of side reactions. Thus, knowledge of the methanol concentration (which should be as low as possible) is important for carrying out polycondensation reactions. Formaldehyde solutions can also be stabilized by certain amines. The methanol concentration can then be kept low.

Paraformaldehyde is a polyoxymethylene with a relatively low molecular mass. Paraformaldehyde depolymerizes in the reaction mixture. The quantity of formaldehyde liberated per unit time increases with decreasing molecular mass of the paraformaldehyde used as starting material. Since phenolic resin condensations can often take place very vigorously, it is important to assess the reactivity of the paraformaldehyde correctly.

Higher aldehydes are less important. In the presence of basic catalysts they preferably undergo side reactions and can therefore generally only be reacted with phenol in an acidic medium. Besides butyraldehyde, benzaldehyde, salicylaldehyde, acrolein, and crotonaldehyde, acetaldehyde and glyoxal are also important. Glyoxal is tetrafunctional towards phenol and gives novolacs with a low content of low molecular mass condensation products.

Hexamethylenetetramine forms adducts with phenols and novolacs very easily, which cross-link on heating. Hexamethylenetetramine is particularly important in the production of thermosetting materials.

3.3. Catalysts

The catalytic action of *acids* on the condensation to produce novolacs is essentially a function of the hydrogen ion concentration. The nature of the anion is less important but must be taken into account because of possible side reactions (sulfonation, ester formation). Besides mineral acids, oxalic acid is often used. It decomposes on heating above 180 °C and thus allows the production of catalyst-free novolacs.

Bases or *salts of weak acids and bases* are suitable as catalysts for resols and novolacs. Catalyst cations are particularly important: whereas alkali hydroxides catalyze the reaction of phenol with formaldehyde in both the *ortho*- and *para*-positions, alkaline-earth metal hydroxides particularly favor formation of *ortho*-products.

Ammonia, amines and hexamethylenetetramine have a catalytic effect. They can be incorporated into the resin by the Mannich reaction.

4. Production

The production of phenolic resins can involve several separate stages. The substitution and condensation reactions between phenols and aldehydes can be carried out separately in terms of both time and location if required, and also

with various catalysts. In addition, hydroxymethyl groups or phenolic hydroxyl groups can be etherified and reactions with unsaturated compounds can be carried out. Resins are frequently worked up by distillation to give concentrated solutions or solid resins. Examples relating to production can be found in [11] and [12].

Corresponding to the great variety of phenolic resins, they are produced over a very broad temperature range which extends from 20 °C in the case of hydroxymethylphenols to more than 280 °C for phenolic resins modified with natural resins. The reactions can be carried out in aqueous solution, organic solvents, or in melts. The reaction is sometimes carried out under pressure to avoid sublimation of, for example, paraformaldehyde or to increase the temperature and thus the reaction rate. Distillations are often performed under vacuum. The phenolic resin plant shown in Figure 1 can be used for all steps in batchwise production. Subsitution and condensation reactions between phenol and formaldehyde are strongly exothermic (the heat of reaction for the formation of methylene bridges in novolacs is −98.7 kJ/mol [11]) and can proceed very fast and vigorously. Therefore, the access to the condensers and overhead and outlet pipes must be kept very wide and the cooling capacity must be designed for rapid condensation and removal of the condensate. To achieve safe industrial production, phenol is always charged to the reactor first and formaldehyde is then added, not vice versa. Because of the low MAK and TLV values (phenol MAK and TWA =5 ppm, formaldehyde MAK=0.5 ppm, TWA=1 ppm, STEL=2 ppm) the working areas must be well ventilated.

Corrosion-resistant or enameled steel is used as the *construction material* for phenolic resin plants.

Emptying the reactor vessel is not problematic for dissolved, low viscosity resins. For solid phenolic resins – particularly resols with a high melt viscosity – rapid emptying and cooling of the resin is very important to avoid further condensation. Besides emptying into pans, removal by cooling conveyors or Pastillier machines has come into use. Phenolic resins which cannot cross-link, such as novolacs or alkylphenol resols, are also worked up in thin-layer evaporators or other continous evaporation plants.

In general, phenolic resins are produced in very small quantities of a very great variety of types, so that batchwise processes are more economic. Some types, e.g., novolacs for thermosets, or resols for the production of wood glues, are, however, produced in large quantities so that a continuous process is more favorable from the point of view of both obtainable throughput and properties of the substance. Processes described in the literature suggest the use of tubular reactors and series of tank reactors [13], [14].

4.1. Novolacs

In the formation of novolacs, substitution and condensation reactions occur simultaneously. In large reaction vessels formaldehyde is metered to a phenol – catalyst mixture and the rate of addition is controlled depending on the heat evolved for safety reasons. When using smaller vessels, even under laboratory conditions, care must be taken because of the exothermic reaction.

Under industrial conditions, analytical control of the course of the reaction is not necessary, because the degree of polycondensation and molecular structure of the resin are determined by the reaction conditions and the concentration of the starting materials. A temperature – time program is usually used.

Example [11]: 1880 parts of phenol are melted in a reaction vessel to which 50 parts of oxalic acid are added and heated to 100 °C. 1560 parts of an aqueous 30 % formaldehyde solution are then gradually metered in over ca. 2 h. The reaction mixture is held at 98 °C for 5 h and the volatile parts are then distilled off under reduced pressure, until a residue temperature of 160 °C is attained at a pressure of 2.0 kPa. 1968 parts of a novolac with a softening point of 124 °C remain in the vessel.

4.2. Resols

Resols from Trifunctional Phenols. The production of resols differs from that of novolacs in that the reactions between phenol and formaldehyde are not allowed to go to completion but are stopped at the stage where auto-cross-linking

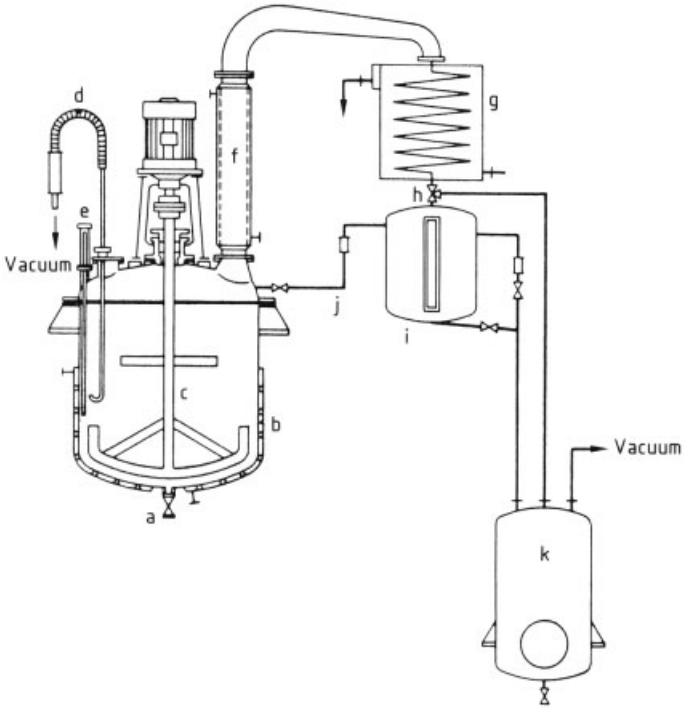

Figure 1. Phenolic resin plant
a) Outlet tap; b) Heating; c) Stirrer; d) Evacuation system; e) Thermometer; f) Reflux condenser; g) Main condenser; h) Three-way tap; i) Reflux and separation vessel; j) Reflux return with siphon; k) Receiver

resols are still liquid or soluble. The continuation of the condensation reactions beyond the resol stage (*A-stage*) leads to resins which are no longer soluble but can only be swelled and which are known as *resitols* (*B-stage*). The final cross-linking to form *resites* (*C-stage*) gives completely cross-linked plastics.

In resol production, the concentration of formaldehyde and the degree of condensation must be controlled during the reaction. The degree of condensation is controlled by measurement of the viscosity, if necessary, after dilution with solvents. Other possibilities include the determination of solubility, melting point, and B-time (see section 6). The reaction temperatures for the production of resols are between room temperature and ca. 100 °C. The molar ratio of phenol to formaldehyde can vary widely. The upper limit is determined by the functionality of the phenol towards formaldehyde (i.e., a molar formaldehyde – phenol ratio of 3:0). Resols are also formed, however, if an excess of phenol is used. In this case a major portion of the phenol does not react.

The reaction is controlled by the type and quantity of catalyst used and the temperature. High alkali concentrations (up to 1 mol per phenolic hydroxyl group) in combination with low temperatures promote the formation of hydroxymethyl groups and stabilize them. Low catalyst concentrations and high temperatures particularly promote the condensation reactions.

The catalyst can remain in the resin and – in the case of alkali hydroxides – render it water-soluble. These catalyst-containing resols can be used industrially. The catalysts can also be neutralized and, if necessary, removed by washing out or filtration. The resols thus lose their water-solubility partially or completely and must then be employed in polar organic solvents.

The heat of formation of hydroxymethyl groups is only −20.3 kJ/mol, as compared with −98.7 kJ/mol for the formation of methylene bridges [11]. Thus, the heat of formation of resols is theoretically lower than that of novolacs. However, hydroxymethyl groups are only stable at low temperatures and can be converted into methylene bridges at higher temper-

atures. Therefore, in the production of resols a great quantity of heat is evolved if the reaction gets out of control. As a result of the high formaldehyde–phenol ratio, resol production also has high safety requirements.

To *etherify* the hydroxymethyl groups, a solution of the resol is heated in, for example, butanol. The water formed is distilled off azeotropically with the organic solvent as entrainer which is recycled to the reactor. Condensation reactions between the hydroxymethyl groups to form ether bridges occur simultaneously. The reaction is controlled by measurement of the viscosity and the solubility. Etherification of alkyd- and oil-modified phenolic resins should be carried out in the presence of a plasticizer, because the compatibility between the reaction components is thus established more quickly. The distillative removal of solvents or water may be necessary during the various reaction stages or as a final step. During the distillation the condensation reactions continue. Distillation must therefore be carried out under mild conditions.

Example [11]. *Condensation*: 846 g (9 mol) phenol and 1600 g formaldehyde solution (30 %, 16 mol) are mixed at 20 °C and the pH adjusted to 9.5 – 10 by addition of concentrated aqueous sodium hydroxide solution with stirring (ca. 110 mL 38 % sodium hydroxide). The mixture is heated to 60 °C and kept for 4 h at this temperature. The decrease in formaldehyde concentration is monitored by titration with hydroxylamine hydrochloride. At the end of the reaction only ca. 1 % formaldehyde is present. The mixture is cooled, and carefully neutralized almost completely with ca. 5 % hydrochloric acid with stirring. The reaction solution must not become acidic and should have a pH of ca. 7.2 – 7.5. Water and excess phenol are subsequently distilled off in vacuum, such that the inner temperature may reach 60 °C at maximum (at 1.5 kPa). 1360 g of viscous resin is obtained as residue.

Etherification: 1360 g butanol and 125 g toluene are added to the residue and the mixture is stirred at 40 °C until the components have completely dissolved. After cooling, precipitated sodium chloride is filtered off. 900 g butanol are then added to the filtrate and the pH ist adjusted to 6 with a concentrated solution of phosphoric acid in ethanol with vigorous stirring. The reaction mixture is then heated to boiling with stirring in a three-necked flask fitted with a Dean–Stark head which is filled with a 1:1 butanol–toluene mixture. In the Dean–Stark head, the water which has been distilled off is separated and the solvent is recycled into the reaction vessel. After ca. 72 mL water have been distilled off, a sample of the reaction solution can be diluted at 20 °C with 4 parts of xylene or 1.4 parts of petroleum ether without the resulting solution becoming cloudy (before the etherification a sample could only be diluted by 0.7 or 0.4 parts, respectively). The solution is concentrated to 2620 g in vacuum and traces of salt are filtered at 20 °C after 48 h.

Alkylphenol Resols. *para*-Alkylphenols can only react with formaldehyde in the *ortho*-positions. Hydroxymethyl groups or dimethylene ether bridges in the *ortho*-position differ from those in the *para*-position by their greater stability. In the production of alkylphenol resols the hydroxymethyl groups therefore remain relatively stable in alkaline medium even at elevated temperatures of between 60 and 70 °C. To prevent the conversion of dimethylene ether bridges to methylene bridges according to Equation (6), see page 3, the production of alkylphenol resols is often performed in two steps. In the first step, a resin rich in hydroxymethyl groups is produced in the presence of alkali. The catalyst is subsequently removed by neutralization and washing out. The substance is then heated to >100 °C to carry out the actual polycondensation, the etherification of hydroxymethyl groups to form dimethylene ether bridges. Catalysts are not required in this step.

4.3. Phenolic Resins Modified by Natural Resins

The reaction of preformed resols with rosin to give albertol acids (see page 3) is carried out in a melt at 100 – 200 °C. The albertol acids are subsequently reacted with polyols to give polyesters or with metal oxides to give resinates (both procedures can also be combined). These reactions must be carried out at temperatures of up to 280 °C because the tertiary carboxyl groups of the abietic acids are only esterified slowly and the melt viscosities and softening

points of the rosin-modified phenolic resins can be very high. Suitable resol components are the reaction products of formaldehyde with phenol, cresols, alkylphenols, or bisphenol A, or mixtures of these. The properties of the resins are determined by the composition and structure of the resol, the resol – rosin ratio, and the nature and quantity of the polyols or metal oxides used for the subsequent reaction. The modified phenolic resins vary greatly with respect to viscosity, softening point, miscibility with solvents, and compatibility with varnish or printing ink raw materials.

The reaction sequence between rosin, resol, and esterification components can also be reversed. In this case the natural resin acids are first esterified and the esters are then reacted with resols.

4.4. Phenolic Resins with Special Properties

Resols for Water-Soluble Varnishes. Water-soluble resols are water-soluble in the form of their phenolates. These phenolates are not suitable for varnishes because they do not form adhesive varnish films. For the production of self-cross-linking binders for water-soluble varnishes, carboxyl groups are introduced into the preformed resols. These carboxyl groups are rendered water-soluble by salt formation with amines or ammonia. The phenolic hydroxyl groups are reacted with chloroacetic acid according to the Williamson ether synthesis usually at the temperature of formation of the resol. Alkali is used as the condensation agent. To achieve water-solubility it is not necessary to introduce carboxyl groups into all resol molecules. The reaction is generally carried out in the presence of auxiliary solvents, mostly glycol ethers, which then also remain in the varnish formulation [15].

Water-soluble phenolic resins for varnishes are mostly applied in combination with plasticizers, for example epoxy resins or fatty oils. It is thus often sufficient to incorporate the groups that impart water-solubility into one of the components. The other component is carried into the aqueous phase by the former component.

Ether Resins from Resols and Novolacs. The phenolic hydroxyl groups can be etherified with, e.g., benzyl chloride or allyl chloride. The etherification always lowers the rate of curing. On the other hand, it reduces disadvantages which are due to the presence of phenolic hydroxyl groups, e.g., oxidizability, tendency towards coloration, and sensitivity towards alkalis. A strong decrease in viscosity is often associated with etherification. This is particularly the case with allyl ethers. The phenolic hydroxyl groups of resols and novolacs can also be etherified with epoxy compounds such as ethylene oxide, propylene oxide, or styrene oxide in the presence of catalytic quantities of alkali. The nucleophilic opening of the oxacyclopropane ring by the phenolic hydroxyl groups is very fast. It is thus possible to work in aqueous solution without significant nucleophilic attack of water at the epoxy group. Reaction of the epoxy compounds with hydroxymethyl groups only occurs if the epoxy compound is added in a nearly stoichiometric ratio to the phenolic hydroxyl group.

Condensation Reactions between Phenol Ethers or Hydrocarbons with Formaldehyde and Phenol. Preformed phenol ethers can form polycondensates with formaldehyde. Known examples are diphenyl ether, phenoxyacetic acid, diphenylglycerol ethers, and phenoxyisopropanol. The condensation reactions can only be carried out with acid catalysts. Depending on the polarity of the ether, strong acids must be used, sometimes in very large quantities. Aromatic hydrocarbons, e.g., xylene, are even less reactive than phenol ethers. The hydroxymethyl groups and dimethylene ether bridges are fairly stable in the phenol ether resins in spite of the high acid concentration and remain as reactive groups in the resin.

Because of the low tendency towards self-curing, these resins are not used for self-cross-linking, but are usually reacted with (excess) phenol to form methylene bridges between phenol ether (or hydrocarbons) and phenol.

4.5. Wastewater

In the production of phenolic resins phenol-containing wastewater is produced, which can also be contaminated by formaldehyde, salts, acids

and alkali, and organic solvents. They must be rendered harmless in a way that will not pollute the environment, for example by incineration. In addition, phenol can be degraded very readily by bacteria. The purification in biological sewage plants – after dilution with other wastewater, if necessary – is therefore very effective. It can be economic to extract the phenol before the incineration of the wastewater.

5. Storage and Transportation

Solid phenolic resins which are sold in broken pieces, flakes, pastilles, or granules, can be packed in sacks, drums, or large containers. If the resins (e.g., rosin-modified types), have a tendency towards autoxidation, they should be stored in the absence of oxygen. *Phenolic resin solutions* are not corrosive and can therefore be stored in iron containers, if a possible coloration of the resin is acceptable. Otherwise the containers must be coated on the inside. *Liquid phenolic resins* are transported in tank cars, containers, or tank pallets. When stored at room temperature, aqueous alkaline resol solutions often lose their viability after a few days or weeks. These solutions must be kept cool during transportation and stored in refrigerated areas. The storage life of solid resols at room temperature is generally longer: it is nevertheless advisable to keep all resols cool during storage. Certain phenolic resins have relatively low softening temperatures and compact during storage which impairs their storage stability.

6. Testing and Analysis

A summary of the testing methods is contained in DIN 16 914 (draft version).

General Investigations. Color index, viscosity, flash point, density, melting point (softening point), and free formaldehyde content are measured according to DIN 53 244. When *free formaldehyde* is determined by the hydroxylamine method, solvents can be used for dissolution of the resins. The blank value of the solvent must be taken into account.

Free phenol is measured photometrically according to DIN 53 704 after steam distillation. Phenol can also be determined bromometrically by bromination of the aromatic nucleus, however side-chain bromination can occur as well and can give values for the phenol concentration which are too high.

Determination of *water* is carried out by the Karl Fischer method.

Hydroxyl Groups. The total quantity of hydroxyl groups is determined by acetylation according to DIN 53 240. The presence of unsaturated groups, for example in modified natural resins, can give rise to values which are too high.

Reactivity Determination for Resols: B-Time. The cross-linking rate of resols, the transition from the A- into the B-stage is determined on a thermostatted heating plate at a predetermined temperature, usually 120 or 150 °C.

The content of hydroxymethyl groups and dimethylene ether bridges in alkylphenol resols is determined by reaction with an excess of phenol and subsequent determination of the water liberated (Karl Fischer method) or the quantity of phenol that has reacted [16].

Differentiation between the Various Types of Phenolic Resin. Neutral resol solutions in polar solvents give a blue or purple coloration with iron(III) chloride. With novolacs this coloration only appears after the addition of a small amount of hexamethylenetetramine. Alkylphenol novolacs often give a black-violet coloration. Modified natural resins do not show the color reaction, but can be identified by the Storch – Morawski color reaction (→ Resins, Natural) [17], [18].

Chromatographic Methods. Thin layer chromatography is very suitable for the determination of phenolic resins (e.g., monomer content) and can also be used for the quantitative determination of components by comparison of the spot size. The sodium salt of diazotized sulfanilic acid in aqueous solution, or iodine vapor are used to develop the spots.

Gel permeation chromatography on polystyrene columns is applied to determine the molecular mass distribution of phenolic resins. In the low molecular mass range it also gives information on the concentration of phenols, hydroxymethylphenols etc. The columns must be calibrated with phenolic compounds [19].

Gas chromatography is suitable for the determination of free phenols. Phenols up to and including butylphenol can be detected by direct injection. Higher phenols are determined after esterification (acetylation, silylation).

Spectroscopic Methods. IR spectroscopy of phenolic resins is described in detail in [10]. The groups typical of phenolic resins can be determined and analyzed semiquantitatively.

Phenolic resins show UV absorption bands at 254 and 280 nm. UV spectroscopy is sometimes used for product monitoring, whereby however, the shift of the bands due to other resin components or because of substitution must be taken into account.

NMR spectroscopy enables detailed examination of the molecular structure of phenolic resins and at the same time kinetic studies of the formation reactions under different conditions [20].

7. Uses

7.1. Novolacs

7.1.1. Cross-Linked Novolacs

For cross-linking, novolacs are processed together with *curing agents*, mainly hexamethylenetetramine. Cross-linking occurs at a sufficient rate at 140 – 160 °C and can be carried out within a few minutes, particularly if a part of the polycondensation reaction has already taken place.

Novolacs are sometimes cross-linked with *resols*. This reaction gives resins with high hardness, high stability, but a very low fracture strain.

The deflection under load is tested according to Martens. According to DIN 53 458 and DIN 53 462, the fatigue strength at elevated temperatures is determined from properties such as the weight loss after storage in air or under an inert gas. Cross-linked phenolic resins are much less inflammable than thermoplastics. Phenolic resins can be rendered virtually nonflammable by the addition of usual plastics additives such as phosphates, borates, halo compounds, red phosphorus, phosphoric acid esters, or antimony trichloride. This is particularly important if other inflammable substances are used as fillers or for reinforcement.

Thermosets. Phenolic resin molding materials are produced with heated rollers or extruders. Treatment of a mixture of novolac, hexamethylenetetramine, fillers, and additives leads to an intermediate, the actual molding material. The polycondensation is driven so far that the molding material in the final molding step can still flow, but so that the final cross-linking occurs very quickly. Phenolic resin thermosets are standardized according to DIN 7708, part 2. The application properties of the final products depend very much on the choice of filler material. The processing properties, in contrast, are affected primarily by the novolacs used. Especially important processing properties are good flow characteristics and rapid curing. Molding technology which predominated until the end of the 1960s has in the meantime been substituted by transfer molding and injection molding, particularly in the mass production of less complicated preforms. Thus the expression "thermosetting molding materials" has come into common use.

Grinding Wheels. In the production of phenolic resin-bound grinding wheels, the grinding material (usually corundum of various granularities) is impregnated with a liquid phenol resol and mixed with a ground, pulverized mixture of phenol novolac and hexamethylenetetramine. The binder cures to a three-dimensional cross-linked resite under a carefully controlled temperature and pressure program. No defects may occur in this process because grinding wheels are exposed to high thermal and mechanical stresses (DIN 69 100, DIN 69 111).

Friction Linings. Brake linings and clutch facings are made of reinforced phenolic resin-bound thermosets. Novolac – hexamethylenetetramine mixtures can also be used here as binders. For better heat conduction, copper wire or copper gauze is often incorporated. Novolacs modified with alkylphenols or cashew oil are used to adjust hardness and lubricating properties, particularly in brake linings.

Reinforcing Resin for Rubber (→ Rubber, 4. Chemicals). The hardness of rubber is increased by the incorporation of novolac – hexamethylenetetramine mixtures. This hardness can be so great as to enable fabrication of solid, tough molded articles, for example for bodywork parts. Often, however, as in car tire mixtures, only slight improvement of the rubber hardness is aimed at. The reinforcing effect is thought to be caused by strong intermolecular interactions between the cured, duroplastic

phenolic resin and the rubber-elastic vulcanized product. Both types of macromolecule are not bonded by covalent bonds.

7.1.2. Novolacs without Cross-Linking

The use of novolacs which are not cross-linked is less important and is based particularly on their solubility and compatibility properties. As polyalkylidenephenols they have strong inter- and intramolecular interactions and can be adapted to the most varying requirements by alkyl substitution.

Varnishes (→ Paints and Coatings, Chap. 2.13.2.). The use of non-cross-linked novolacs in varnishes is the oldest application of phenolic resins. Because of their high susceptibility towards oxidation and coloration, particularly in the case of unsubstituted phenol novolacs, this application area has hardly developed further. Non-cross-linked novolacs are now only important in specialist areas such as model lacquers and bituminous paints.

Printing Technology. Novolacs are often added to aniline printing inks because of their high affinity for these dyes. Because aniline printing inks are often used for printing on food packaging materials, the novolacs must be phenol-free. Novolacs are also used in the formulation of ballpoint pastes.

Novolacs are applied together with diazo compounds to zinc or aluminum plates in a thin layer for production of positive offset printing plates (→ Imaging Technology, Chap. 3.2.). Similar systems are used as photoresists for the photolithographic production of integrated circuits (→ Imaging Technology, Chap. 4.).

Raw Materials for Epoxy Resins. The reaction of epichlorohydrin with novolacs (instead of bisphenol A) leads to polyfunctional epoxy resins (→ Epoxy Resins, Chap. 2.2.2.). Their curing products are characterized by their particularly high usage temperatures.

Alkylphenol novolacs, particularly those based on 4-hydroxybiphenyl, 4-isooctylphenol, and formaldehyde are used for the production of copying paper.

Tackifiers for Rubber (→ Rubber, 4. Chemicals). Nonvulcanized rubber mixtures are often not tacky enough to adhere to each other. This is a considerable disadvantage in the production of laminated rubber articles, e.g., car tires. Tackifiers are used in rubber technology to improve this. Besides mineral oils and natural and synthetic resins, alkylphenol novolacs are particularly effective and when used with other tackifiers have a synergistic effect. In addition to alkylphenol – formaldehyde novolacs, *para*-alkylphenols – acetaldehyde resins, and *para*-alkylphenols – acetylene resins are often used as tackifiers.

Crude Oil Separators. Ethoxylation (i.e., reaction with ethylene oxide) of alkylphenol novolacs gives water-soluble, surface-active substances. If the degree of ethoxylation is suitably adjusted, they have de-emulsifiying properties. Addition of small quantities to the crude oil that is extracted as an oil – saltwater emulsion effects a faster separation into an oil and an aqueous phase.

7.2. Resols

7.2.1. Water-Soluble Resols

Condensation products of phenol and formaldehyde, which are present as water-soluble phenolates, are used mainly as adhesives for wood or wood chips, as binders for organic and inorganic fibers and for granular material such as corundum or sand. Curing is performed thermally at 100 – 180 °C and, if possible, under pressure. If curing is carried out under atmospheric pressure, the cross-linking process must be carefully controlled and performed slowly to avoid the formation of bubbles or cavities.

The waste-gas from the thermal curing process contains phenol, formaldehyde, and sometimes solvent vapors. When legally stipulated limiting values are exceeded, the waste-gases must be incinerated.

Cross-linked phenolic resins have a low flammability. They can be rendered virtually nonflammable by additives. In the event of fire the flue gases are relatively nontoxic and, compared to the flue gases from other burning plastics, are relatively transparent. This makes them particularly suitable for the interiors of vehicles, places of assembly etc.

Chipboard Adhesives. The greatest quantity of aqueous, alkaline resol solutions are used in the production of wooden chipboard. Chipboard

bound by phenolic resin is moisture-resistant in contrast to that bound by urea resins, and is therefore suitable for the production of exterior walls in buildings and related uses.

Construction Adhesives. Resorcinol – formaldehyde resins are used as wood glues. Resorcinol is treated with a substoichiometric amount of formaldehyde in an alkaline medium and the stable solution is treated with more formaldehyde (usually in the form of paraformaldehyde) immediately before use. The polycondensation with resorcinol, which is very fast at, and even below, room temperature is thus restarted. The resorcinol – formaldehyde resin cross-links three-dimensionally. Glueings containing resorcinol resins have high mechanical stability and are highly water-resistant.

Binding Agents for Molding Sand (→ Foundry Technology, Chap. 2.3.4.). Phenol resols or novolac – hexamethylenetetramine mixtures which can be sufficiently diluted with water, are used as binders for casting molds, which consist mainly of sand. Mixtures of both resins are sometimes used for full molds. For hollow and thin-walled molds (Croning process), alcoholic solutions of novolac – hexamethylenetetramine adducts are used. To achieve special properties or to make the process cheaper, phenolic resins are sometimes modified with furan or urea resins.

Laminates. High quality laminates for electrical equipment are obtained by impregnation of paper with phenolic resins and subsequent thermal cross-linking. They are used as insulation and construction materials. Unmodified resols and oil-plasticized resols are used together for impregnation. Both resins have different functions. The unmodified resins – with as low a molecular mass as possible – penetrate into the fiber and the oil-plasticized resins remain on the fiber surfaces and bind them together.

In *decorative laminates*, water-soluble, alkaline resol solutions can only be used for the production of core resins because of their dark coloration. Melamine resins are used as decorative facings.

Fiber Bonding. Organic or inorganic fibers or rock wool can be bonded to plates or nonwoven fabrics by impregnation with an aqueous resol solution or this can be sprayed on them. Curing is carried out continuously with heated rollers or in presses. Matting or insulating plates bound with phenolic resin have heat-insulation properties similar to thermoplastic foams, but have the advantage of being virtually nonflammable and at the same time soundproof.

Foamed Plastics (→ Foamed Plastics, Chap. 4.4.). Water-soluble resols can be used for the production of foamed plastics after neutralization and, if required, removal of the basic catalyst. For foaming, the addition of surfactants and the use of blowing agents (i.e., low-boiling or halogenated hydrocarbons) is necessary. Phenolic resin foams have the advantage of low flammability. However, the corrosive effect of the strong acids used as hardeners limits their general use.

Abrasives. To produce abrasives such as sandpaper, abrasive cloth, or flexible sanding disks on vulcanized fiber, abrasive grains are applied to a moving backing cloth on one side. These grains are fixed together and to the surface of the backing cloth with a binder. For heat-resistant abrasives, phenol resols are preferably used as binders.

Since the carrier material (usually cellulose) is sensitive to high temperatures, a prolonged application of temperatures $>130\,^{\circ}C$ must be avoided. Therefore, cross-linking of the phenol resol to resite takes a long time. Phenolic resins that are suitable for the production of abrasives must therefore be very reactive, to limit the curing time to a few hours. This naturally limits their storage life. The cross-linking reactions are carried out under atmospheric pressure in sloping dryers and the carrier materials, coated with phenolic resin, and abrasive grains are introduced in sloping loops through channels, whose temperature and humidity must be carefully controlled.

7.2.2. Resols in Organic Solvents

The resols produced as aqueous alkaline solutions can be dissolved in polar solvents after neutralization of the catalyst. Basically such solutions do not differ from aqueous solutions. Catalysts can be added to accelerate cross-linking. Since, however, phenolic resins are often used in combination with other components, these resins must be compatible with the other substances, and for this purpose their solution properties must be altered. This can be done using alkylated phenols as raw materials or in the

etherification of the methylol groups with alcohols.

Epoxy Resin – Phenolic Resin Coatings. Resols based on cresols are used in cross-linking together with expoxy resin. The enamels baked at temperatures around 200 °C stand out by virtue of their very good ductility and very high chemical stability. They are principally used for the inner coating of food and drink cans, and other containers.

Oil-Plasticized Resols. Resols – mainly based on cresol or bisphenol A – are modified by fatty oils or alkyds and are used as baking enamels. These varnishes have good chemical stability and very high ductility and are used, in particular, as inner coatings for containers. Other uses of oil-plasticized resols lie in the area of cylinder-heat gaskets and electric insulation materials.

7.2.3. Alkylphenol Resols

Resols from *para*-alkylphenols are also soluble in nonpolar solvents. The solubility increases with increasing size of the alkyl group. *para*-Alkylphenol resols differ from other resols in their bifunctionality towards formaldehyde. Thus, *para*-alkylphenol resols cannot cross-link by themselves, but are used as cross-linking agents for other, polyfunctional substances. Examples are fatty oils, rubbers, long-chain epoxy resins, alkyd resins, and phenol novolacs.

Vulcanization. Rubbers can be vulcanized by alkylphenol resols. As phenolic resins are considerably more expensive than sulfur, this possibility is only exploited for special cases, e.g., butyl rubber.

Adhesive Resins. The properties of polychloroprene adhesives can be improved by natural and synthetic resins. Alkylphenol resols not only increase the stickiness of the adhesive but also its binding and heat stability. Since the amount of resin needed to be added is about half the quantity of polychloroprene, adhesive resins have considerable economic importance.

The alkylphenol resols do not act as such in polychloroprene adhesives but in the form of their magnesium and zinc complexes. These are formed with the magnesium and zinc oxides present in the polychloroprene adhesive. The complexes described as chelate polymers have a very high molecular mass, but are nevertheless readily soluble in aromatic hydrocarbons and even in naphtha [21].

7.3. Phenolic Resins Modified by Natural Resins

Printing Inks. The rosin-modified phenolic resins are used as binders for book or offset printing inks for the following reasons: (1) They are soluble in mineral oil and aromatics, (2) compatible with other printing-ink raw materials, and (3) have a high melting point, which allows a rapid transition from the liquid to the solid state on printing. Phenol resins modified by natural resins are very varied so that they can be adopted to the continually increasing rate of working in the production of printed articles.

7.4. Waterborne Paints, Phenol Ether Resins

Waterborne Paints. The amine and ammonium salts of resols containing carboxyl groups are water-soluble. They are used as baking enamels and give coatings which are resistant to chemicals but are brittle. To plasticize them they are mixed with fatty oils, alkyds, or unsaturated hydrocarbons (which have been rendered water-soluble by the addition of maleic anhydride and subsequent salt formation). Their main use lies in the electrodeposition coating of car bodies and vehicle parts. See also, → Paints and Coatings, Chap. 2.13..

Allyl Ether Resols. Resols etherified with allyl groups at the phenolic hydroxyl group have a very low viscosity and can be used as solvent-free or low-solvent binders in varnishes. Higher baking temperatures must be used for cross-linking than in the case of simple phenolic resins.

Phenol Ether and Hydrocarbon – Formaldehyde Polycondensates. Mixed condensates from (1) phenol ethers – formaldehyde – phenol, or (2) hydrocarbons – formaldehyde – phenol are used in thermosets which have an increased stability towards chemicals and oxygen.

Table 1. Phenolic resin production in the United States

Year	Production, t/a	Proportion of duro-plastics, %	Proportion of all plastics, %
1978	728	36	6.1
1980	675	37	5.6
1982	902	47	7.4
1984	1126	45	7.4
1986	1231	47	7.3
1988	1380	47	6.9

7.5. High-Temperature Coking

Phenolic resins, whether resols or novolacs, coke upon heating in the absence of oxygen. They can therefore be used as binders for refractories based, for example, on magnesite or dolomite, which are used for the lining of blast furnaces and similar constructions. Phenolic resins can also be used as high-temperature materials after coking and in some cases graphitization. These materials can be reinforced with temperature-resistant fibers.

8. Economic Aspects

In 1978, 728 000 t phenolic resin were produced in the United States [22] and in 1994, 730 000 t [23]. The development of production since 1978 and the proportion of phenolic resins in the production of all plastics in the United States is shown in Table 1.

In the Federal Republic of Germany the phenolic resin market did not develop to the same extent. Excluding phenolic resins modified by natural resins and thermosets, production was ca. 230 000 t in 1989 compared with 180 000 – 200 000 t in 1978. Of this ca. 35 % is used by the wood industry in the form of glues. 20 % are needed for the production of casting molds and molding materials (free market) and 10 % each for the production of fiber matting, brake linings, and grinding agents. 3 % each go into the paints and coatings industry, the adhesives industry, and the production of battery separators.

No conclusions about the turnover value can be drawn from the quantity data, for although phenolic resins for mass applications are cheap, very high prices are often demanded for specialist applications.

9. Toxicology and Occupational Health

Cured phenol – aldehyde resins are classified as nontoxic. The toxicity of phenolic resins is therefore assessed according to their content of unreacted phenol [24]. In the Federal Republic of Germany phenolic resin formulations with a content of more than 5 % free phenol must be labeled with a skull and crossbones and those with more than 1 % with a St. Andrew's cross. Substances liberated in the thermal cross-linking of phenolic resins, particularly formaldehyde, are toxic and lead at least to irritation. Contact with phenolic resins can lead to allergies in sensitized people.

Occupational Health. No TLV or MAK values have been established for phenolic resins. The corresponding values of the most important monomers are as follows: formaldehyde TLV-TWA 1 ppm (1.2 mg/m^3), TLV-STEL 2 ppm (2.5 mg/m^3), MAK 0.5 ppm; phenol TLV-TWA 5 ppm (19 mg/m^3), MAK 5 ppm.

10. References

General References

1. A. Bachmann: *Phenoplaste,* Dtsch. Vlg. Kunststoffind., Leipzig 1973.
2. K. Hultzsch: *Chemie der Phenolharze,* Springer Verlag, Berlin 1950.
3. A. Knop, L. Pilato: *Phenolic Resins,* Springer Verlag, Berlin 1985.
4. R. Martin: *The Chemistry of Phenolic Resins,* Wiley, New York 1956.
5. N. Megson: *Phenolic Resin Chemistry,* Butterworths, London 1958.

6. H. Wagner, H. F. Sarx: *Lackkunstharze,* Hanser, München 1971.

Specific References

7. N. Megson, *Chem. Ztg.* **96** (1972) no. 1/2, 14 – 19.
8. K. Hultzsch, *Kunststoffe* **42** (1952) no. 11, 385.
9. K. Hultzsch, *Kunststoffhandbuch, vol. X: Duroplaste,* Hanser, München 1968.
10. D. Hummel: *Atlas der Kunststoff-Analyse,* 2nd ed., Hanser, München 1984.
11. *Houben-Weyl,* 4th ed., **XIV/2,** 272 – 291.
12. *Ullmann,* 3rd ed., **13,** 453 – 479.
13. K. H. Jung, DE-OS 2 620 488, 1977.
14. F. Wilhelm, DE 1 570 349, 1972.
15. W. Brushwell, *Farbe + Lack* **82** (1976) no. 10, 917.
16. Hoechst AG, Kunstharz-Lieferprogramm, Analytische Bestimmungsmethoden, Data sheet, Feb. 15, 1982.
17. *Dtsch. Farben-Z.* **20** (1966) 363.
18. D. Braun, *Kunststoffe* **62** (1972) no. 1, 41.
19. E. Wagner, R. Greff, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 2193 – 2207.
20. J. Woodbrey et al., *J. Polym. Sci. Polym. Chem. Ed.* **3** (1965) 1079 – 1106.
21. K. Hultzsch, *Farbe + Lack* **77** (1971) 1165 – 1172.
22. *Chem. Eng. News* **67** (1989) June 19, 45.
23. K. Weissermel, H.-J. Arpe: *Industrielle organische Chemie,* Wiley-VCH, Weinheim 1998, p. 394.
24. H. Babisch et al., *Regul. Toxicol. Pharmacol.* **1** (1981) 90 – 109.

Phenolsulfonic Acids → Benzenesulfonic Acids and Their Derivatives

Phenothiazine and Derivatives

GÉRARD TAURAND, Rhône – Poulenc Rorer CRVA, Vitry Sur Seine, France

1. **Introduction** 1
2. **Uses and Biological Activity** 1
2.1. **Therapeutic Uses** 1
2.2. **Biological Activity** 3
2.3. **Other Uses** 5
3. **Syntheses and Reactions** 5
3.1. **Ring Syntheses** 5
3.2. ***N*-Substituted Phenothiazines** 8
3.3. ***C*-Substituted Phenothiazines** 8
3.4. **Analogues and Miscellaneous Compounds** 8
3.5. **Other Chemical Properties of the Phenothiazine Nucleus** 9
4. **Physical Properties** 9
5. **Toxicology** 11
6. **References** 11

1. Introduction

Phenothiazine (**1**) [*92-84-2*] (10*H*-dibenzo-[*b, e*]-1,4-thiazine) was first synthesized by BERNTHSEN in 1883 [1]. The chemistry of phenothiazine derivatives has since assumed considerable importance due to the diversity and number of compounds synthesized and the many products exhibiting extremely important therapeutic properties.

The atoms in the phenothiazine ring (**1**) are numbered as follows [2]:

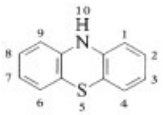

1

Phenothiazines have been reviewed [3]. More detailed information is to be found in [4–29].

2. Uses and Biological Activity

Physical properties of the compounds discussed in this chapter are listed in Chapter 4.

2.1. Therapeutic Uses

Work carried out by LAUTH [30] and BERNTHSEN [31] revealed evidence of the anthelmintic properties [32], [33] of phenothiazine (**1**) and the antimalarial properties [34–36] of derivatives such as methylene blue (**2**) [*61-73-4*], which were also used as dyes (→ Azine Dyes, Chap. 4.1.) and antiseptics [37–39].

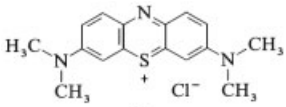

2

However, it was only with the work of CHARPENTIER [40], [41] that any real interest was shown in 10-aminoalkyl phenothiazine derivatives, thereby revealing the versatility of phenothiazine as a pharmaceutical.

Antihistamines. The discovery of the role of histamine in biological processes [42] allowed the development of bioassays (e.g., isolated ileum spasm) that led to the discovery of the first synthetic antihistamines such as phenbenzamine [43].

Aminoalkyl groups were shown to have various biological activities (e.g., antimalarial, antihistamine, antispasmolytic). These groups were therefore condensed onto phenothiazine to obtain new antimalarial agents. However, the resulting derivatives were found to be of no use as antimalarials [44], but had useful antihistamine activity [45].

These investigations terminated in the synthesis of promethazine (**3**) [*60-87-7*], followed by many other products; some examples are given in Table 1 (see also → Antiallergic Agent).

Oxomemazine (**7**) [*3689-50-7*] (RP 6847, Dysedon, Doxergan) [50] and isothipendyl (**8**)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_387

Table 1. Phenothiazines used as antihistamines

No.	Names	CAS registry no.	Trade names (examples)	Substituents ∗		Other uses	Ref.
				$-R^2$	$-R^{10}$		
3	Promethazine	[*60-87-7*]	RP 3277, RP 3389, Phenergan [*58-33-3*]	$-H$	$-CH_2CH(CH_3)N(CH_3)_2$		[46]
4	Thiazinamium	[*2338-21-8*]	RP 3554, Multergan [*58-34-4*]	$-H$	$-CH_2CH(CH_3)N^+(CH_3)_3$	spasmolytic	[47]
5	Trimeprazine	[*84-96-8*]	RP 6549, Alimemazine, Theralene [*4330-99-8*]	$-H$	$-CH_2CH(CH_3)CH_2N(CH_3)_2$	antiemetic	[48]
6	Dimetothiazine	[*7456-24-8*]	RP 8599, Fonazine, Migristene [*13115-40-7*]	$-SO_2N(CH_3)_2$	$-CH_2CH(CH_3)N(CH_3)_2$	antimigraine agent	[49]

∗ The superscripts indicate the position of the substituents in the phenothiazine ring.

[*482-15-5*] (D 201, Andantol) [51] may conveniently be added to this list of products.

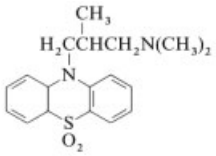

7

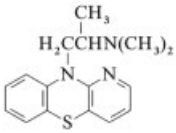

8

Antiparkinsonism Agents. Utilization of the inhibitory properties of certain phenothiazines on the extrapyramidal motor system [52] has led to the development of products with anticholinergic properties that are used to treat Parkinson's disease. Examples are diethazine (**9**) [*60-91-3*] (RP 2987, Diparcol) [53], and ethopromazine (**10**) [*522-00-9*] (RP 3356, Parsidol) [54] (see → Parkinsonism Treatment, Chap. 2.1.3.).

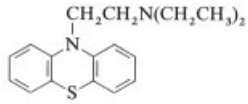

9

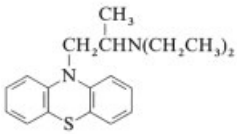

10

Neuroleptics. The discovery of neuroleptics resulted from numerous investigations on the physiopathology of the autonomic nervous system [55], [56]. Laborit perfected anesthesia and artificial hibernation by using " cocktail-lytics" [57], [58]. He recognized the advantage conferred by certain antihistamines (in particular promethazine) in these mixtures on account of their "hypometabolizing, hypnotic, analgesic, and antithermic properties." This advantage and the powerful potentiating activity of these products on the action of barbiturates [59], [60] led to the investigation of phenothiazines with a pronounced action on the central nervous system (CNS), but with a greatly reduced antihistamine activity.

The usefulness of chlorpromazine (**11**) [*50-53-3*] [61] was first demonstrated in anesthesiology and in the treatment of operative shock [57] as an " autonomic stabilizer" [61]. However, it was soon recognized that the major importance of this product was in psychiatric therapeutics [62–65]. Chlorpromazine was thus the first representative of a new class of drugs used to treat psychoses, namely the major tranquillizers (neuroleptics).

Table 2 gives some examples of neuroleptics (see also → Psychopharmaceutical Agent).

Some neuroleptic phenothiazines have a very prolonged action. Examples include the heptanoate ester of fluphenazine (**20**) (Moditen-Retard [*2746-81-8*]) and the palmitate ester of pipotiazine (**22**) (Piportil M 2). The latter compound can be administered as a single monthly intramuscular injection [82].

The field of neuroleptics has stimulated research aimed at elucidating and characterizing pharmacological profiles for the discovery of

Table 2. Phenothiazines used as neuroleptics

No.	Names	CAS registry no.	Trade names (examples)	Substituents*		Ref.
				$-R^2$	$-R^{10}$	
11	Chlorpromazine	[*50-53-3*]	RP 4650, Largactil, Largactil monohydrochloride [*69-09-0*]	$-Cl$	$-CH_2CH_2CH_2N(CH_3)_2$	[66]
12**	Promazine	[*58-40-2*]	RP 3276, Liranol, promazine hydrochloride [*53-60-1*]	$-H$	$-CH_2CH_2CH_2N(CH_3)_2$	[67]
13	Acepromazine	[*61-00-7*]	1522 CB, Vetranquil, acepromazine maleate [*3598-37-6*]	$-C(=O)CH_3$	$-CH_2CH_2CH_2N(CH_3)_2$	[68]
14**	Methotrimeprazine	[*60-99-1*]	RP 7044, Levomepromazine hydrogen maleate, methotrimeprazine maleate [*7104-38-3*]	$-OCH_3$	$-CH_2CH(CH_3)CH_2N(CH_3)_2$	[69]
15	Cyamemazine	[*3546-03-0*]	RP 7204, Tercian	$-CN$	$-CH_2CH(CH_3)CH_2N(CH_3)_2$	[70]
16	Trifluomeprazine	[*2622-37-9*]	RP 7746, Nortran [*71609-19-3*]	$-CF_3$	$-CH_2CH(CH_3)CH_2N(CH_3)_2$	[71]
17	Perazine	[*84-97-9*]	P 725, Taxilan	$-H$	$-CH_2CH_2CH_2-N$ 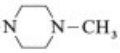 $N-CH_3$	[72]
18	Prochlorperazine	[*58-38-8*]	RP 6140, SKF 4657, Stementil [*84-02-6*]	$-Cl$	$-CH_2CH_2CH_2-N$ 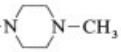 $N-CH_3$	[73]
19	Thioproperazine	[*316-81-4*]	RP 7843, SKF 5883, Majeptil [*2347-80-0*]	$-SO_2N(CH_3)_2$	$-CH_2CH_2CH_2-N$ 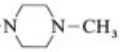 $N-CH_3$	[42]
20	Fluphenazine	[*69-23-8*]	S 94, Sq 4918, Permitil [*146-56-5*]	$-CF_3$	$-CH_2CH_2CH_2-N$ 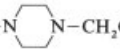 $N-CH_2CH_2OH$	[74], [75]
21	Pericyazine	[*2622-26-6*]	RP 8908, Neuleptil, propericiazine	$-CN$	$-CH_2CH_2CH_2-N$ 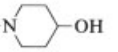 $-OH$	[76], [77]
22	Pipotiazine	[*39860-99-6*]	RP 19366, Piportil	$-SO_2N(CH_3)_2$	$-CH_2CH_2CH_2-N$ 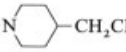 $-CH_2CH_2OH$	[78]
23	Perimethazine	[*13093-88-4*]	RP 9159, Leptryl	$-OCH_3$	$-CH_2CH(CH_3)CH_2-N$ $-OH$	[79]
24	Mesoridazine	[*5588-33-0*]	TP 523, Serentil, mesoridazine benzene sulfonate [*32672-69-8*]	$-SOCH_3$		[80]
25	Thioridazine	[*50-52-2*]	TP 21, Melleril	$-SCH_3$		[81]

* The superscripts indicate the positon of the substituents in the phenothiazine ring. ** Also used as a sedative.

other types of psychotropic drugs. Neuroleptic phenothiazines can interact with a large number of aminergic (e.g., dopaminergic, histaminergic, adrenergic) or enzyme systems, and have been used as tools in attempting to understand the physiopathology of psychotic disease [83–97]. The interactions involved have still not been elucidated in detail.

Antiemetics. The exploitation of the activity of certain phenothiazines in inhibiting apomorphine-induced vomiting in dogs [98] has resulted in excellent antiemetic products for use

in humans (hepatodigestive system, pediatrics, drug intolerance, etc.). Examples are metopimazine (**26**) [*14008-44-7*] (RP 9965, Vogalene) [99], [100] and prochlorperazine (**18**) [101] (see also → Antiemetics, Chap. 3.).

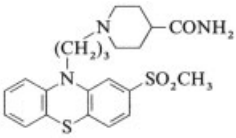

26

Anti-Inflammatories. Certain phenothiazine products such as metiazinic acid (**27**) [*13993-65-2*] (RP 16091, Soripal) [102] are used as nonsteroidal anti-inflammatory agents.

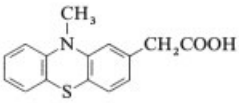

27

Other Therapeutic Effects. Some phenothiazine derivatives possess activities that have resulted in their use in special therapeutic applications. Table 3 gives examples.

Other products have also been the subject of clinical cardiovascular studies. They include chloracizine (**31**) [*800-22-6*] (G-020, RP 11504, Chlorocyzin) as a coronary vasodilator [106], and moricizine (**32**) [*31883-05-3*] (EN-313, Ethmosine) as an antiarrhythmic agent [107–109].

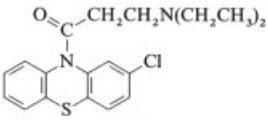

31

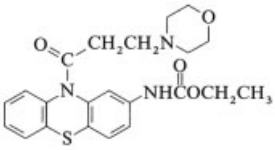

32

Numerous studies have been published on the potential use of phenothiazines as anticancer drugs [110–112]. Veterinary applications have also been described [113], [114].

Certain phenothiazines with antidepressant properties (uptake of noradrenaline, serotonin) have stimulated research on other tricyclic compounds, resulting in the discovery of powerful antidepressants such as imipramine.

2.2. Biological Activity

A brief summary of the numerous biological fields in which phenothiazines are of interest will now be given to explain both their versatility as pharmaceuticals and the motivation of future research.

Binding to Receptors. The pharmacological profiles of phenothiazines that have been the subject of a therapeutic application are often correlated with their agonistic or antagonistic interaction with neuronal transmitter receptors: dopamine D_1 [115] and D_2 [116], histamine H_1 [117], adrenergic α_1 and α_2 [118], muscarinic (cholinergic) [117], and serotoninrgic [118] receptors.

Binding to other receptors of the neuronal system has also been noted, including the opiate receptor [119]. The product RP 60180 (**33**) [*128182-96-7*], for example, has a good affinity for the κ opiate receptor [120], [121].

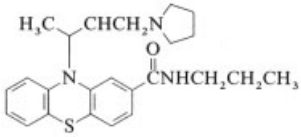

33

Some phenothiazines have a high affinity for nonneuronal receptors, such as plasma proteins (albumin [122], α_1-AG [123]), lipoproteins, tissues (e.g., muscle, liver), membranes, or proteins (e.g., calmodulin [124]).

Enzyme Inhibition. The interaction of phenothiazines with numerous enzyme systems has been studied [16]. Phenothiazines bind extremely easily to proteins in cell membranes [125] and affect the activity of the enzymes or the accessibility of substrates and cofactors. Numerous physicochemical studies have been

Table 3. Phenothiazines used in special therapeutic applications

No.	Names	CAS registry no.	Trade names (examples)	Substituents [a]		Indications for use	Ref.
				$-R^2$	$-R^{10}$		
28	Amino-promazine	[*58-37-7*]	RP 3828, Lispamol [*2278-27-5*]	$-H$	$-CH_2CH(N(CH_3)_2)CH_2N(CH_3)_2$	spasmolytic	[b]
29	Chlorpro-ethazine	[*84-01-5*]	RP 4909, Neuriplege [*4611-02-3*]	$-Cl$	$-CH_2CH_2CH_2N(C_2H_5)_2$	decontrac-tant, muscle relaxant	[c]
30	Dimethoxa-nate	[*477-93-0*]	Tussidin	$-H$	$-C(=O)-OCH_2CH_2OCH_2CH_2N(CH_3)_2$	antitussive agent	[d]

[a] The superscripts indicate the position of the substituents in the phenothiazine ring.

[b] [103]. [c] [104]. [d] [105]

carried out on a wide variety of systems (e.g., ATPase, calmodulin-dependent enzymes, "redox" enzymes). These interactions are often associated with the complex-forming or redox properties of phenothiazines.

Antitumor Activities. Numerous phenothiazines have been studied in oncology. They include dyes [126], psychotropic agents [127], products opposing pleiotropic resistance (resistance to penetration of drugs into the cells) [128] and, in particular, "mustard"-type alkylating agents [129], [130].

Structure – Activity Relationships and Molecular Modeling. Several reviews have been published on the structure – activity relationships of phenothiazines [131–133] using physicochemical approaches (e.g., X-ray analysis, NMR spectroscopic studies in solution). A study of the different molecular conformations (Chap. 4) and geometrical parameters is useful for interpreting selectivities observed in various biological fields. For example, it has been shown that the conformation of the chains in the 10-position plays an important role in neuroleptic activity [134], [135].

Conformational analysis using molecular modeling enables hypotheses to be proposed concerning the interaction of phenothiazines with their biological targets [136].

2.3. Other Uses

Phenothiazine dyes are covered elsewhere (→ Azine Dyes, Chap. 4.1. and → Sulfur Dyes). Phenothiazines are also used for analytical purposes as redox indicators [137–139] and as reagents in spectrophotometry [139], [140] and gravimetry [17].

3. Syntheses and Reactions

3.1. Ring Syntheses

Phenothiazines can be synthesized by cyclization of diphenylamines or diphenyl sulfides (Fig. 1).

Cyclization of Diphenylamines. The first synthesis of phenothiazine (**1**) was performed by reacting diphenylamine with sulfur [1]. This synthesis was substantially improved by using iodine as catalyst [141], which enabled the first synthesis of 2-chlorophenothiazine, the key intermediate in the preparation of chlorpromazine (**11**) [66], [142–144].

Cyclization of Diphenyl Sulfides. The cyclization of diphenyl sulfides is by far the most important method of phenothiazine synthesis: In the "classical" route, Y in Figure 1 is a halogen atom leading to Ullmann-type syntheses (Fig. 2). In the route where Y is a nitro group, synthesis involves a Smiles rearrangement (Fig. 2).

The starting materials of choice for the two methods are 2-halonitrobenzenes (**34**) and 2-aminothiophenols (**35**), respectively.

Reaction of 2-halonitrobenzene (**36**) with **35** in an alkaline medium yields the corresponding 2-amino-2′-nitrodiphenyl sulfides, which

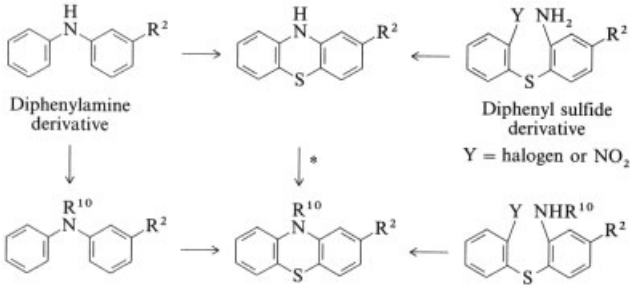

Figure 1. Synthesis of phenothiazines by cyclization of diphenylamines and diphenyl sulfides
∗ see Section 3.2

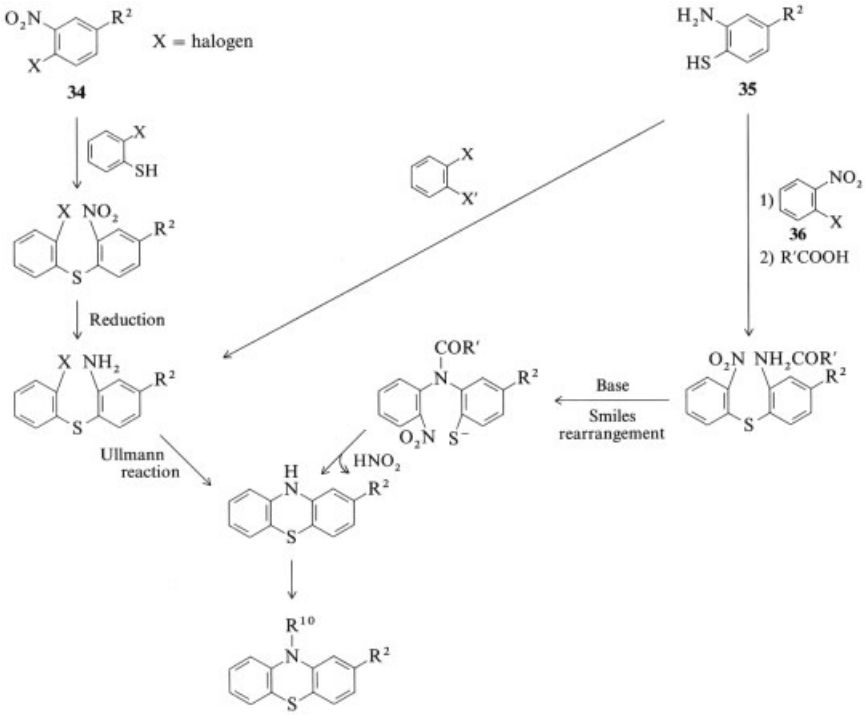

Figure 2. Cyclization of diphenyl sulfides

are acylated (generally with formic or acetic acid). The phenothiazine is then obtained by the Smiles rearrangement [145–148]. The main factors governing the reaction are steric hindrance [149], dependency on the electrophilic nature of carbon [150], the acidity of the acylamino group that attacks the aromatic nucleus [147], as well as the nature of the base.

As an example, 2-chlorophenothiazine can be prepared by this method [151]:

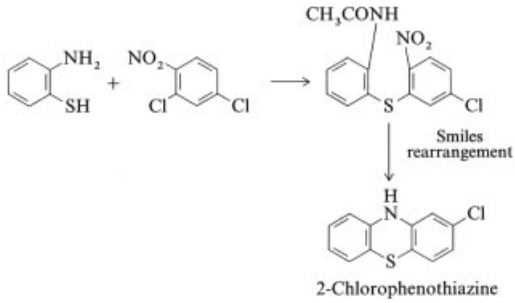

The second main cyclization route for diphenyl sulfides starts by reacting 2-halothiophenols with 2-halonitrobenzenes (**34**) (Fig. 2) [152]. Reduction of the product yields 2-halo-2′-aminodiphenyl sulfides which give the phenothiazine by an Ullmann reaction via formation of a benzyne intermediate, or by a Smiles rearrangement [153].

2-Chlorophenothiazine is synthesized by this route [154]:

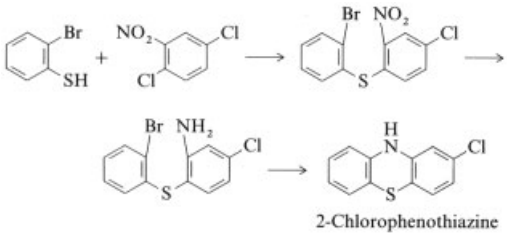

Finally, phenothiazines may be obtained by cyclizing 2-aminothiophenols (**35**) with 1,2-dihalogenated cyclic compounds. The main application of this method is the synthesis of azaphenothiazines [155–160]. For example [159]:

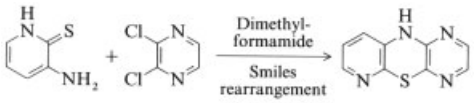

Cyclization with Benzoquinones. Another important synthesis route to the phenothiazine ring is cyclization of 2-aminothiophenols with chloroquinones [161–164]:

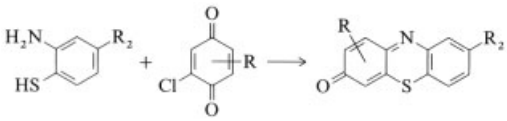

An example is [165]:

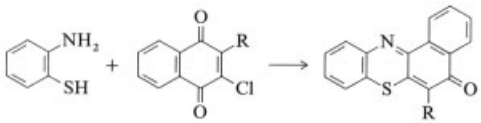

Other Cyclizations. Phenothiazinones may also be obtained by cyclization of 2-aminothiophenols with 1,3-cyclohexanediones [166]:

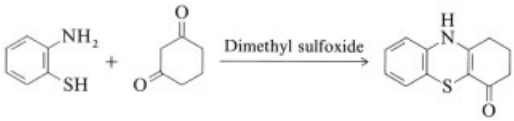

The reductive cyclization of 2-nitrodiphenyl sulfides with triethyl phosphite also yields phenothiazines [167], [168]:

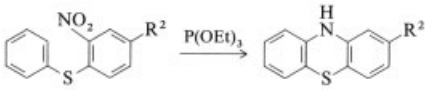

This cyclization involves a rearrangement similar to the Smiles rearrangement. The same applies to cyclizations starting from azide derivatives [169], [170], which probably proceed via a nitrene intermediate and involve electrophilic attack:

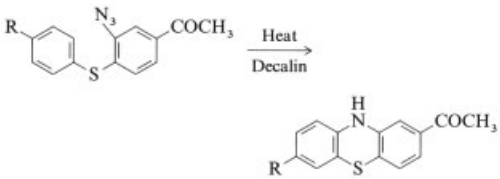

Other reports of phenothiazine formation cited in the literature include the reaction of isopropyl dithiocarbonate with 4-chloro-3,5-dinitrotrifluoromethylbenzene [171] or syntheses of azaphenothiazines by thermal cyclization of 1,4-benzo[*b*]thiazine derivatives [172].

3.2. *N*-Substituted Phenothiazines

The main routes for the substitution of nitrogen in the 10-position of phenothiazines are summarized in Figure 3.

The first substitutions involving acyl (route 1) or alkyl groups (route 2) were described by Bernthsen [173] and Fränkel [174], see also [12]. These substitutions are performed with alkyl halides, acid anhydrides, or acid chlorides in an alkaline medium. The first phenothiazines having a tertiary alkylamine substituent in the 10-position were described by Gilman and Shirley [44]. Charpentier [40], [175], [176]

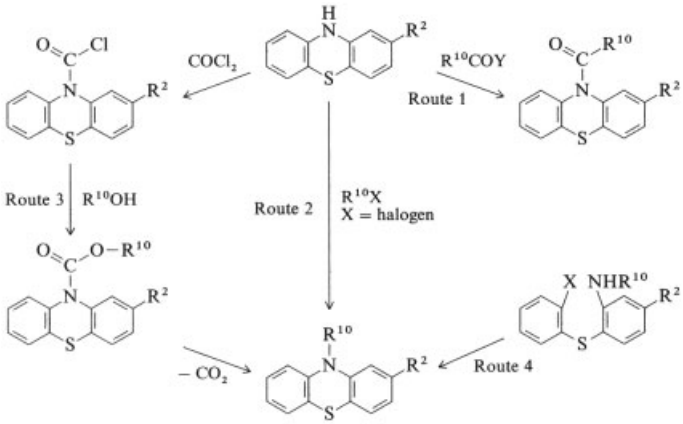

Figure 3. Routes for the substitution of nitrogen in the 10-position of phenothiazines

then synthesized this type of product and demonstrated their valuable pharmacological properties [177].

Route 3 involves decarboxylation of carbamates [178], [179].

In route 4, diaryl sulfides are alkylated, and the *N*-substituted products are then cyclized by the methods discussed in Section 3 [180].

As an example, chlorpromazine (**11**) has been synthesized by these four routes as well as via the 10-cyanoethyl derivative [181].

In route 2 (direct alkylation), the C_2 or C_3 chains of the alkylating agents may undergo rearrangement via a mechanism involving the formation of an aziridinium or an azetidinium compound. For example [40]:

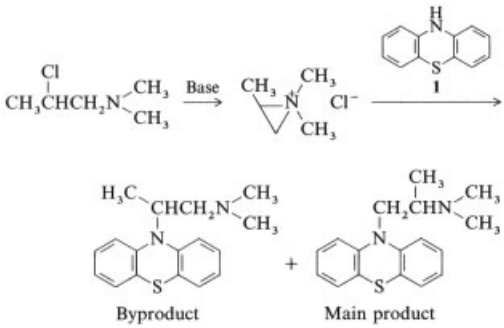

A similar mechanism is observed in route 3 with the formation of an intermediate cyclic ammonium compound [67], [182]. Route 2, which proceeds by direct alkylation of the phenothiazine nucleus, is the most important reaction in synthesis.

3.3. *C*-Substituted Phenothiazines

Electrophilic substitution occurs preferentially at the 3,7-C>1,9-C positions for 10-alkyl phenothiazines, and at the 2,8-C positions for 10-acyl phenothiazines. This generally holds for alkylation [183], [184], Friedel – Crafts acylation [185], [186], electrophilic halogenation [187], [188], nitration via "NO_2^+" [189], and metallation. Metallation yields compounds that are difficult to obtain by other methods [190–193].

Nucleophilic substitution is governed by the fact that phenothiazine has a high electron density and can thus yield oxidized cationic species that are susceptible to nucleophilic attack in the 3,7-C positions [194].

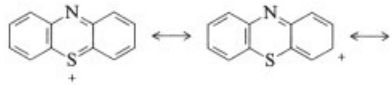

Examples of such reactions are nitration via "NO_2^-" [195], [196], halogenation via "X^-" [197–199], and thiocyanation by the Kaufmann method [200].

3.4. Analogues and Miscellaneous Compounds

Recyclized analogues may be obtained by Friedel – Crafts reactions [201–203], for example [204]:

Table 4. Physical properties of phenothiazines

No.	Trade names	Form	Empirical formula	M_r	Appearance	*mp*, °C	Solubility
1	Phenoxur	base	$C_{12}H_9NS$	199.3	yellow needles	185	ether
2	Methylene blue	chloride	$C_{16}H_{18}ClN_3S$	319.9	dark green crystals		1 g/2.5 cm^3 water
3	Phenergan	hydrochloride	$C_{17}H_{20}N_2S \cdot HCl$	320.9	colorless crystals	230 – 232	water
4	Multergan	methyl sulfate	$C_{19}H_{26}N_2O_4S_2$	410.6	colorless crystals	206 – 210	ethanol
5	Theralene	hemitartrate	$C_{18}H_{22}N_2S \cdot C_2H_3O_3$	373.5	colorless crystals	68 (base)	water
6	Bonpac	mesylate	$C_{19}H_{25}N_3O_2S_2 \cdot CH_4O_3S$	487.7	colorless crystals	214 (decomp.)	water
7	Doxergan	hydrochloride	$C_{18}H_{22}N_2O_2S \cdot HCl$	366.9	colorless crystals	250	water
8	Andantol	hydrochloride	$C_{16}H_{19}N_3S \cdot HCl$	321.9	colorless crystals	222 – 223	water
9	Diparcol	hydrochloride	$C_{18}H_{22}N_2S \cdot HCl$	334.9	colorless crystals	184 – 186	water
10	Parsidol	hydrochloride	$C_{19}H_{24}N_2S \cdot HCl$	348.9	colorless crystals	223 – 225 (decomp.)	water
11	Largactil	hydrochloride	$C_{17}H_{19}ClN_2S \cdot HCl$	355.3	colorless crystals	279 – 280 (decomp.)	water
12	Talofen	hydrochloride	$C_{17}H_{20}N_2S \cdot HCl$	320.9	pale yellow crystals	181	water
13	Notensil	maleate	$C_{19}H_{22}N_2OS \cdot C_4H_4O_4$	442.5	pale yellow crystals	135 – 136	water
14	Nozinan	maleate	$C_{19}H_{24}N_2OS \cdot C_4H_4O_4$	444.6	pale yellow crystals	190 (decomp.)	0.3 % at 20 °C in water
15	Tercian	base	$C_{19}H_{21}N_3S$	323.5	yellow powder	89 – 96	HCl/water
16	Nortran	maleate	$C_{19}H_{21}F_3N_2S \cdot C_4H_4O_4$	482.5	fine crystals		water
17	Taxilan	base	$C_{20}H_{25}N_3S$	339.5	colorless crystals	51 – 53	HCl/water
18	Stemetil	dimaleate	$C_{20}H_{24}ClN_3S \cdot C_8H_8O_8$	606.1	very fine crystals	228	water
19	Majeptil	dimethanesulfonate	$C_{22}H_{30}N_4O_2S_2 \cdot C_8H_8O_6S_2$	638.8	yellow crystals	140 (base)	water
20	Permitil	dihydrochloride	$C_{22}H_{26}F_3N_3OS \cdot 2\,HCl$	510.5	colorless to pale yellow crystals	235 – 237	water
21	Pericyazine	base	$C_{21}H_{23}N_3OS$	365.5	colorless to pale yellow crystals	116 – 117	HCl/water
22	Pipothiazine	base	$C_{24}H_{33}N_3O_3S_2$	475.7	yellow crystals	120 – 122	water
23	Leptryl	base	$C_{22}H_{28}N_2O_2S$	384.5	colorless to pale yellow crystals	137 – 138	water
24	Serentil	phenyl sulfonate	$C_{21}H_{26}N_2OS_2 \cdot C_6H_6O_3S$	544.8	yellow crystals		water
25	Melleril	base	$C_{21}H_{26}N_2S_2$	370.6	colorless to pale yellow crystals	72 – 74	HCl/water
26	Vogalene	base	$C_{22}H_{27}N_3O_3S_2$	445.6	yellow crystals	170 – 171	water
27	Soripal	base	$C_{15}H_{13}NO_2S$	271.3	colorless to pale yellow crystals	146	NaOH/water
28	Lispamol	hemifumarate	$C_{19}H_{25}N_3S \cdot C_2H_2O_2$	385.5	colorless to pale yellow crystals	166 – 170 (decomp.)	water
29	Neuriplege	hydrochloride	$C_{19}H_{23}ClN_2S \cdot HCl$	383.4	colorless to pale yellow crystals	178	water
30	Tussidin	base	$C_{19}H_{22}N_2O_3S$	358.5	colorless to pale yellow crystals	161 – 163 (HCl)	HCl/water
31	Chlorocyzin	base	$C_{19}H_{21}ClN_2OS$	360.9	colorless to pale yellow crystals	168 – 169 (HCl)	water
32	Ethmosine	hydrochloride	$C_{22}H_{25}N_3O_4S \cdot HCl$	464.0	colorless to pale yellow crystals	189 (decomp.)	water
33	RP 60180	hydrochloride	$C_{23}H_{29}N_3OS \cdot HCl$	432.0	colorless to pale yellow crystals	187 – 189	water

Benzo-condensed analogues, oxidized quinone forms, and polyhydro forms may be obtained by cyclization.

Many "aza" derivatives have been synthesized; examples are described in [159], [160].

Isothipendyl (**8**) is synthesized as follows [51], [205]:

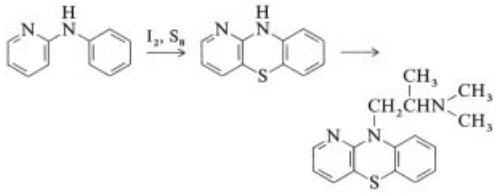

8

Finally, phenothiazines can be oxidized (see Section 3.5) to sulfones with potassium permanganate [206] or hydrogen peroxide in acetic acid [207], [208], or to sulfoxides with hydrogen peroxide in ethanol [209] or with nitric acid [210].

Phosphorylated derivatives have also been synthesized [211].

3.5. Other Chemical Properties of the Phenothiazine Nucleus

Oxidation of phenothiazines produces free radicals and cations [212–215]. A large amount of work has been carried out on the elucidation of the electrochemical [216] and photochemical [217] reaction mechanisms and the relationships between the various oxidized species and products [218], [219]. The most important sequence may be formulated as [220]:

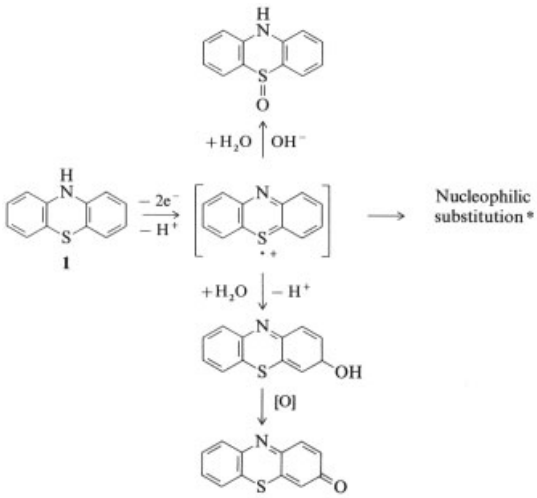

∗ See Section 3.3.

Since phenothiazines have a high electron density, they may form charge-transfer complexes [212] in the presence of electron acceptors such as iodine [221], quinones [222], or tetracyanoethylene [223]. Numerous radical cation salts may also be formed, for example with picric acid [224].

4. Physical Properties

Molecular Conformation. Various studies (principally X-ray diffraction [225], [226] and nuclear magnetic resonance [227]) have shown that the phenothiazine ring may have a planar, axial-boat (extra), or equatorial-boat (intra) conformation. These forms are all in equilibrium with one another in solution (Fig. 4) [228], [229].

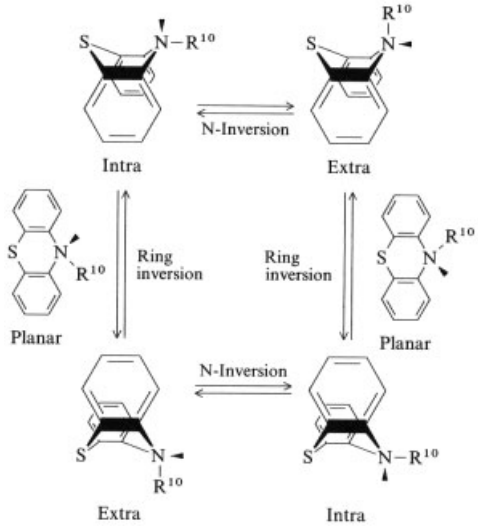

Figure 4. Conformations of phenothiazines

Spectral Properties. The spectral properties of phenothiazines are described in the literature: NMR [228–230], X-ray diffraction [225], IR spectroscopy [231–233], UV spectroscopy [185], [231], electron spin resonance [234] (and orbital calculations), mass spectrometry [235–237], and coupling of mass spectrometry with chromatography [238].

Analysis. Phenothiazines are mainly separated by chromatography (gas liquid chromatography, high performance liquid chromatography, or thin layer chromatography) [239–242]. Methods of determination are based on chemical properties (redox reactions, complex formation [243]) or spectroscopy (UV absorption, mass spectrometry [244]).

Physical Properties of Products. The physical properties of products mentioned in Chapter 2 are listed in Table 4.

5. Toxicology

Phenothiazines are generally regarded as being well tolerated. The toxicological effects are

largely due to their modes of action, while some specific toxicity phenomena are associated with the chemical properties of the phenothiazine nucleus.

A large number of studies have been carried out on the metabolism of phenothiazines [245–250] and their pharmacokinetics [245]. The phenothiazine nucleus is mainly metabolized by oxidation at the 5- and 7-positions:

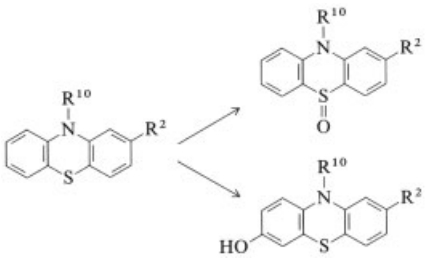

Metabolism depends on the nature of the R^2 and R^{10} substituents and, in turn, may also affect the substituents (certain R^2 substituents, but particularly the R^{10} chain [245]).

The toxicity [251], [252] and side effects [253], [254] associated with the mode of action of phenothiazines (Section 2.2) have been investigated in detail [255–257]. Side effects in the CNS are mainly manifested in the extrapyramidal system (Parkinsonism [258], acathesia, delayed dyskinesia [259]) as well as in the dopaminergic [260] and noradrenergic [261] systems.

Effects on fertility have been noted [262], [263]. Mutagenic effects are very slight [264] and no teratotoxicity has been observed [265]. Tolerance problems and interaction with other drugs have been observed but no dependence symptoms [266].

Phenothiazines may cause poor local tolerance [267] or lead to dermal photosensitization [268] and dermatitis on exposure to light [269]. Several other rare effects have been detected [270], [271].

6. References

1. A. Bernthsen, *Ber. Dtsch. Chem. Ges.* **16** (1883) 2896.
2. *Ring System Handbook,* Chemical Abstracts Publ. 1984.
3. R. R. Gupta (ed.): "Phenothiazines and 1,4-Benzothiazines: Chemical and Biomedical Aspects," *Bioactive Molecules,* **vol. 4,** Elsevier, Amsterdam 1988, p. 992 ff.
4. S. Saraf, F. Al-Omran, B. Al-Saleh, *Heterocycles* **26** (1987) no. 1, 239 – 273.
5. D. K. Yarmukhametova, K. Dzhundubaev, B. V. Kudryavtsev: "Chemistry of Phenothiazine, Bibliographic Index, Ilim: Founze," *Kirg. SSR* (1979) p. 410 ff.; Chem. Abstr. 93, 71 747 e.
6. D. Lednicer, L. A. Mitscher: *The Organic Chemistry of Drug Synthesis: Phenothiazines,* Wiley, New York 1977, pp. 379 – 392.
7. P. A. J. Janssen in C. J. Cavallito (ed.): *Industrial Encyclopedia of Pharmacology and Therapeutics,* **vol. 1,** Section 5, Pergamon Press, New York 1973, p. 37.
8. C. Bodea, J. Silberg: "Recent Advances in the Chemistry of Phenothiazines," in A. R. Katritsky, A. J. Boulton (eds.): *Advances in Heterocyclic Chemistry,* vol. 9, Academic Press, New York 1968, pp. 321 – 460.
9. J. Sykulski, *Wiad. Chem.* **20** (1966) 1 – 23; *Chem. Abstr.* **64** 11 206 d.
10. C. R. Ramage in E. H. Rodd (ed.): *Heterocyclic Compounds,* Elsevier, Amsterdam 1960.
11. D. E. Pearson in R. C. E. Elderfield (ed.): *Heterocyclic Compounds,* **vol. 6,** Wiley, New York 1967, p. 624 ff.
12. S. P. Massie, *Chem. Rev.* **54** (1954) 797 – 823.
13. R. J. Baldessarini in L. S. Goodman, A. Gilman, A. G. Gilman (eds.): *The Pharmacological Basis of Therapeutics,* McMillan, New York 1980, Chap. 19, pp. 395 – 414.
14. C. Kaiser, P. E. Setler in M. E. Wolff (ed.): *Burger's Med. Chem. 4th ed. 1980,* **Part III,** 874 – 889.
15. C. L. Zirkle, C. Kaiser in A. Burger (ed.): *Medical Chemistry,* Wiley, New York 1970.
16. C. L. Zirkle, C. Kaiser in M. Gordon (ed.): *Psychopharmacological Agents,* **vol. 3,** Academic Press, New York 1974, pp. 39 – 128.
17. E. Usdin, H. Eckert, I. S. Forrest (eds.): *Developments in Neuroscience,* **vol. 7,** Elsevier, New York 1980, p. 376 ff.
18. I. S. Forrest, C. J. Carr, E. Usdin (eds.): *Adv. Biomed. Psychopharmacol. Ser.* **9** (1974) p. 840 ff.
19. I. S. Forrest, E. Usdin, *Psychopharmacology* **2** (1977) Part 2, 699 – 753.
20. U. Breyer, B. Müller-Oerlinghausen, W. Mauruschat, *Psychopharmacology* **2** (1977) Part 2, 755 –793.

21. C. D. Burrell, *Psychopharmacology* **2** (1977) Part 2, 795 – 826.
22. P. Deligne, M. Bunodière, *Ann. Anesthesiol. Fr.* **19** (1978) 7 – 9 (1 H–11 H, I-1/I-13, 1-J/13-J).
23. E. F. Domino, R. D. Hudson, G. Zografi in A. Bürger, M. Dekker (eds.): *Drugs Affecting the Central Nervous System,* **vol. 2,** M. Dekker, New York 1968, pp. 327 – 397.
24. G. M. Simpson, J. W. S. Angus, *Curr. Ther. Res. Clin. Clin. Exp.* **9** (1967) 265.
25. P. B. Bradley in W. J. Root, F. G. Hofmann (eds.): *Physiological Pharmacology,* **vol. 1,** Academic Press, New York 1963, p. 417.
26. E. Schenker, H. Herbst, *Prog. Drug Res.* **5** (1963) 269 – 627.
27. D. S. Folse, *Agric. Exp. Stn. Tech. Bull. Kansas* **115** (1961) p. 216 ff.
28. M. W. Parkes in G. P. Ellis, G. B. West (eds.): *Prog. Med. Chem.* **1** (1961) 72.
29. S. Courvoisier et al., *Arch Int. Pharmacodyn. Ther.* **92** (1952) 306 – 336.
30. C. Lauth, *Ber. Dtsch. Chem. Ges.* **9** (1876) 1035.
31. A. Bernthsen, *Ber. Dtsch. Chem. Ges.* **45** (1912) 1987 – 2042.
32. E. G. Erdös, N. Baart, S. P. Shanor, F. F. Földes, *Arch. Int. Pharmacodyn. Ther.* **117** (1958) 163.
33. F. De Eds, A. B. Stockton, J. O. Thomas, *J. Pharmacol. Exp. Ther.* **65** (1939) 353.
34. P. Guttman, P. Ehrlich, *Berl. Klin. Wochenschr.* **28** (1891) 593.
35. L. Goldman: *Synthetic Antimalarials,* College Parks, Maryland, USA, 1944, p. 209 ff.
36. W. Kikuth, J. Schilling, *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Hyg. Abt. 1: Orig.* **151** (1944) 293.
37. E. Poulson, *Lehrb. Pharmakol.* **16** (1949) 350.
38. B. L. Freedlander, *Proc. Soc. Exp. Biol. Med.* **57** (1944) 106.
39. J. J. Jouin, N. P. Buu-Hoi, *Ann. Inst. Pasteur Paris* **72** (1946) 580.
40. P. Charpentier, *C. R. Acad. Sci. (France)* **225** (1947) 306 – 308.
41. P. Charpentier, US 2 645 640, 1953.
42. H. H. Dale, P. P. Laidlaw, *J. Physiol. London* **41** (1910) 318.
43. Rhône-Poulenc, FR 913 161, 1946 (J. Mosnier).
44. H. Gilman, D. A. Shirley, *J. Am. Chem. Soc.* **66** (1944) 888 – 893.
45. B. N. Halpern, R. Ducrot, *C. R. Seances Soc. Biol. Ses Fil.* **140** (1946) 361.
46. Rhône-Poulenc, US 2 607 773, 1952 (S. S. Berg, J. N. Ashley).
47. Rhône-Poulenc, GB 641 452, 1950.
48. Rhône-Poulenc, US 2 837 518, 1958 (R. Jacob, J. Robert).
49. Rhône-Poulenc, GB 814 512, 1959.
50. Rhône-Poulenc, US 2 972 612, 1961 (R. Jacob, J. Robert).
51. H. L. Yale, F. Sowinski, *J. Am. Chem. Soc.* **80** (1958) 1651.
52. M. Fink, *E. E. G.,* Clin. Neurophysiology 12 (1960) 359.
53. Rhône-Poulenc, US 2 530 451, 1950 (P. Charpentier).
54. Rhône-Poulenc, US 2 607 773, 1952 (S. S. Berg, J. N. Ashley).
55. J. Reilly, E. Rivalier, A. Compagnon, R. Laplane, *C. R. Seances Soc. Biol. Ses Fil.* **116** (1934) 24.
56. H. Selye, *Endocrinology (Baltimore)* **21** (1937) 169.
57. H. Laborit, P. Huguenard, *Presse Med.* **59** (1951) 1329.
58. H. Laborit, *Presse Med.* **62** (1954) 359, 936.
59. D. Bovet, J. Fournel, P. Charpentier, *Therapie* **2** (1947) 115.
60. R. Ducrot, P. Decourt, *C. R. Seances Soc. Biol. Ses Fil.* **144** (1950) 908.
61. H. Laborit, P. Huguenard, R. Alluaume, *Presse Med.* **60** (1952) 206.
62. J. Sigwald, *Presse Med.* **59** (1951) 819.
63. J. Sigwald, D. Bouttier, J. Solignac, *Sem. Hop.* **36** (1960) 40.
64. J. Delay, P. Deniker, *J. Ther. Paris* **8** (1953) 347 –364.
65. J. Delay, P. Deniker: *Methods Chimiotherapeutiques en Psychiatrie,* Masson, Paris 1961.
66. P. Charpentier et al., *C. R. Hebd. Seances Acad. Sci.* **235** (1952) 59.
67. Rhône-Poulenc, US 2 519 886, 1950 (P. Charpentier).
68. J. Schmitt et al., *Bull. Soc. Chim. Fr.* 1957, 938.
69. S. Courvoisier et al., *C. R. Seances Soc. Biol. Ses Fil.* **151** (1957) 1378.
70. Rhône-Poulenc, US 2 877 224, 1959 (R. Jacob, J. Robert).
71. SKF, GB 813 861, 1959 (G. E. Ullyot).
72. H. Hromatka et al., *Monatsh. Chem.* **88** (1957) 56.
73. Rhône-Poulenc, GB 780 193, 1957 (R. J. Horclois); FR 1 167 627, 1958 (R. J. Horclois).
74. H. L. Yale, F. Sowinski, *J. Am. Chem. Soc.* **82** (1960) 2039.
75. SKF, GB 829 246, 1960 (P. J. C. Buisson, P. Gaillot, J. Gaudechon).

76. Scherico, GB 833 473, 1960 (P. Gaillot, J. Gaudechon).
77. K. Florey, *Anal. Profiles Drug Subst.* **2** (1973) 245 – 294.
78. Rhône-Poulenc, FR M 7835, 1970 (J. C. Blonde, J. Fouché).
79. Rhône-Poulenc, GB 904 210, 1962 (R. Jacob, J. Robert).
80. Sandoz, US 3 084 161, 1963 (J. Renz et al.).
81. J. P. Bourquin et al., *Helv. Chim. Acta* **41** (1958) 1072.
82. J. Brown-Thomson, *Acta Psychiatr. Scand.* **241** (1973) 241. Olin Mathieson Chem. Co., DE 1 165 602, 1964 (R. C. Merril, H. L. Yale).
83. P. Radiguet de La Bastaie: *Bases Pratiques d'Anesthésie à l'Usage des Aides et Techniciens en Anesthésie-Reanimation,* Arnette, Paris 1984, p. 264 ff.
84. J. C. Blanchard, A. Boireau, C. Garret, L. Julou, *Biochem. Pharmacol.* **29** (1980) 2933 – 2942.
85. C. E. Speicher, M. J. Walters: "Phenothiazines," *Clin. Lab. Med.* 1981, 547 – 555.
86. M. Krause-Sternberg: "Psychotropic Drugs, III. Phenothiazine," *Monatskurse ärztl. Fortbild.* **30** (1980) no. 8, 332.
87. L. Rivera Calimlim, *Ration. Drug Res. Ther.* 1977, 1 – 8.
88. J. O. Cole, *Drug Treat. Ment. Disord.* 1976, 13 –30.
89. A. E. Hartshorn, *Drug Intell. Clin. Pharm.* **9** (1975) no. 10, 536 – 550.
90. E. Lear, *Clin. Anesth.* **10** (1973) Part 1, 185 – 206; *Chem. Abstr.* **83,** 141 591 u.
91. D. N. Teller, *Mod. Pharmacol. Toxicol.* **2** (1975) 1 – 62; *Chem. Abstr.* **83,** 90 533 s.
92. S. C. Kaim, *Adv. Biochem. Psychopharmacol.* **9** (1974) 685 – 690.
93. H. Goldenberg, V. Fishman, *Princ. Psychopharmacol.* 1970, 179 – 191; *Chem. Abstr.* **76,** 107 712 d.
94. J. M. Davis, N. S. Kline in P. Black (ed.): *Drugs Brain,* J. Hopkins Press, Baltimore 1969, pp. 173 – 184.
95. J. C. Le Douarec, G. Régnier, *Ann. Anesthesiol. Fr.* **9** (1968) no. 2, 287 – 297.
96. R. Dequeker, *Verh. K. Vlaam. Acad. Geneeskd. Belg.* **30** (1968) nos. 1 – 2, 5 – 49.
97. J. Crisman, *Psychopharmacol. Bull.* **4** (1967) 151 pp.
98. J. H. Moyer, *Med. Clin. North Am.* **41** (1957) 405.
99. R. Ducrot, P. Decourt, *C. R. Seances Soc. Biol. Ses Fil.* **145** (1951) 356.
100. Rhône-Poulenc, DE 1 092 476, 1962 (R. Jabob, J. Robert).
101. W. L. Wilson, L. L. Shane, J. H. Moyer, *Antibiot. Med. Clin. Ther. (N.Y.)* **6** (1959) 421.
102. Rhône-Poulenc, FR M 4163, 1966 (M. Messer et al.).
103. R. Jacob et al., *C. R. Hebd. Seances Acad. Sci.* **243** (1956) 1637.
104. Rhône-Poulenc, US 2 769 002, 1956 (P. J. C. Buisson et al.).
105. American Home Products, US 2 778 824, 1957 (P. Seeman).
106. Rhône-Poulenc, FR 1 060 715, 1954 (R. J. Horclois, R. Métivier).
107. Acad. Med. Sci. USSR, DE 2 014 201, 1971 (A. Grisenko).
108. H. Cousse, G. Mouzin, B. Bonnaud et al., *Boll. Chim. Farm.* **119** (1980) no. 1, 31 – 40.
109. A. M. Habibzadeh, *Ariz. Med.* **36** (1979) no. 7, 518 – 520; *Chem. Abstr.* **91,** 133 621 d.
110. R. R. Gupta (ed.): *Bioactive Molecules,* **vol. 4,** Elsevier, Amsterdam 1988, pp. 705 – 774.
111. G. Jones, N. Robert, *Tumori* **71** (1985) no. 6, 563 – 569; *Chem. Abstr.* **108,** 68 140 z.
112. D. Farge, M. Messer, *Bull. Soc. Chim. Fr.* 1969, 4395 – 4402.
113. E. J. Tweedale, J. T. Edwards: "Phenothiazines in Veterinary Medicine," *Imp. Bur. Animal Health* 1947, p. 36 ff.
114. D. Zivanov, *Veterinaria (Sarajevo)* **37** (1988) 369 – 385; *Chem. Abstr.* **111,** 166 655 c.
115. M. Cincotta, D. J. DeVries, P. M. Beart, *Neurosci. Lett.* **61** (1985) 85 – 89.
116. J. Creese, D. R. Burt, S. H. Snyder, *Science (Washington, D.C., 1883)* **194** (1976) 546 – 548.
117. E. Richelson, *J. Clin. Psychiatry* **45** (1984) 331 – 336.
118. S. J. Peroutka, S. H. Snyder, *Am. J. Psychiatry* **137** (1980) 1518 – 1522.
119. J. Creese, A. P. Feinberg, S. H. Snyder, *Eur. J. Pharmacol.* **36** (1976) 231 – 235.
120. Rhône-Poulenc-Santé, EP 346 238, 1989 (C. Garret, C. Guyon, B. Plau, G. Taurand).
121. V. Fardin et al., *Pain Suppl.* **5** (1990) 192.
122. J. Krieglstein, F. Lier, J. Michaelis, *Naunyn-Schmiedeberg's Arch. Pharmacol.* **272** (1972) 121 – 130.
123. M. L. Korguth, L. G. Hutchins, B. S. Eichelmann, *Biochem. Pharmacol.* **30** (1981) 2435 – 2441.
124. B. Weiss, W. C. Prozialeck, T. L. Wallace, *Biochem. Pharmacol.* **32** (1982) 2217 – 2226.
125. A. Rosen, S. Y. Tham, *J. Pharm. Pharmacol.* **32** (1980) 357 – 359.

126. H. Gilman, D. A. Shirley, *J. Am. Chem. Soc.* **66** (1944) 888 – 893.
127. T. Hirata, G. Peng, J. S. Driscoll, *J. Pharm. Sci.* **67** (1978) 157 – 162.
128. J. M. Ford, W. C. Prozialeck, W. N. Hate, *Proc. AACR* **29** (1988) 1245.
129. R. B. Boss, P. E. Swartzentruber: *Literature Survey of Nitrogen Mustards,* Cancer Chemotherapy National Service Center, Bethesda, Maryland, 1959.
130. A. Loveless, *Nature (London)* **167** (1951) 338 – 342.
131. A. Felmeister, *J. Pharm. Sci.* **61** (1972) no. 2, 151 – 164.
132. J. Schmutz, C. W. Picard in F. Hoffmeister, G. Stille (eds.): *Psychotropic Agents,* Part 1, Springer Verlag, Berlin 1980, pp. 3 – 26.
133. P. R. Andrews, E. J. Lloyd, *Med. Res. Rev.* **21** (1982) no. 4, 355 – 393.
134. G. Fronza, E. Ragg, R. Mondelli, *Actual. Chim. Ther. 8th* 1981 245.
135. W. C. Prozialeck, M. Cimino, B. Weiss, *Mol. Pharmacol.* **19** (1981) 264 – 269.
136. J. Barbe et al., *Eur. J. Med. Chem. Chim. Ther.* **18** (1983) 531 – 534.
137. H. Puzanowska-Tarasiewicz, M. Tarasiewicz, *Wiad. Chem.* **37** (1983) no. 3, 231 – 253; *Chem. Abstr.* **100,** 68198 j.
138. I. Nemeova, N. Zimova-Sulcova, I. Nemec, *Chem. Listy* **76** (1982) no. 2, 142 – 157.
139. J. Bartos, M. Pesez: Practique de l'Analyse Organique Colorimetrique et Fluorimetrique, **vol. 2**, Masson, Paris 1984, p. 398 ff.
140. H. Basinska, M. Tarasiewicz, *Chem. Anal. (Warsaw)* **17** (1972) 469 – 478.
141. E. Knoevenagel, *J. Prakt. Chem.* **89** (1913) 1 – 50.
142. P. Charpentier, FR 1 029 987, 1950.
143. E. A. Nodiff et al., *J. Heterocycl. Chem.* **7** (1970) 203 – 208.
144. E. A. Nodiff, P. N. Craig, *J. Org. Chem.* **26** (1961) 824 – 828.
145. A. A. Levy, H. C. Rains, S. Smiles, *J. Chem. Soc.* 1931, 3264 – 3269.
146. J. F. Bunnet, *Q. Rev. Chem. Soc.* **12** (1958) 1 – 16.
147. W. E. Truce, E. M. Krier, W. W. Brand: "The Smiles and Related Rearrangements of Aromatic Systems," *Org. React. (N.Y.)* **18** (1970) 99 – 125.
148. J. A. Van Allan, G. A. Reynolds, R. F. Adel, *J. Org. Chem.* **27** (1962) 1659 – 1664.
149. A. A. Levy, S. Smiles, *J. Chem. Soc.* 1932, 1488 – 1492.
150. F. Galbraith, S. Smiles, J. Chem. Soc. **1935,** 1234 – 1238.
151. J. Evans, S. Smiles, *J. Chem. Soc.* 1935, 1263.
152. E. A. Nodiff, M. Hausman, *J. Org. Chem.* **31** (1966) 625 – 626.
153. G. E. Bonvicino, L. H. Yogodzinski, R. A. Hardy, Jr., *J. Org. Chem.* **27** (1962) 4272 – 4280.
154. Rhône-Poulenc, BE 556 475, 1960.
155. C. O. Okafor, *Int. J. Sulfur Chem. Part B* **6** (1971) no. 3, 237 – 265.
156. C. O. Okafor, *Int. J. Sulfur Chem. Part B* **7** (1972) no. 2, 109 – 119.
157. C. O. Okafor, M. L. Steenberg, J. P. Buckley, *Eur. J. Med. Chem. Chim. Ther.* **12** (1977) 249 – 256.
158. C. O. Okafor, *Phosphorus Sulfur* **4** (1978) no. 1, 79 – 99.
159. C. O. Okafor, *J. Org. Chem.* **47** (1982) 592 – 594.
160. C. O. Okafor, *J. Heterocycl. Chem.* **17** (1980) 1587 –1592.
161. S. K. Jain, Dissertation, Rajasthan University, Jaipur, India, 1970.
162. V. I. Oprean, W. Schäfer, *Justus Liebigs Ann. Chem.* **765** (1972) 1 – 7.
163. Y. Ueno, *Pharmazie* **39** (1984) 355 – 356.
164. S. K. Jain, Dissertation, Rajasthan University, Jaipur, India, 1975.
165. Y. Ueno, Y. Takeuchi, J. Kashitani, T. Yoshida, *J. Heterocycl. Chem.* **19** (1982) 167 – 169.
166. S. Miyano, N. Abe, K. Sumoto, *J. Chem. Soc. Chem. Commun.* 1975, no. 18, 760.
167. J. I. G. Cadogan, R. K. MacKie, M. J. Todd, *Chem. Commun.* **15** (1966) 491.
168. J. I. G. Cadogan, *Synthesis* **1** (1969) no. 1, 11 – 17.
169. M. Messer, D. Farge, *Bull. Soc. Chim. Fr.* **7** (1968) 2832 – 2842.
170. M. Messer, D. Farge, *Bull. Soc. Chim. Fr.* **12** (1969) 4395 – 4402.
171. J. J. D'Amico, C. C. Tung, N. E. Dahl, *Phosphorus Sulfur* **7** (1979) 143 – 148.
172. R. J. Chorvat et al., *J. Med. Chem.* **26** (1983) 845 – 850.
173. A. Bernthsen, *Justus Liebigs Ann. Chem.* **230** (1885) 88 – 95.
174. N. Fränkel, *Ber. Dtsch. Chem. Ges.* **18** (1885) 1845.
175. Rhône-Poulenc, DE 824 944, 1945 (P. Charpentier).
176. Rhône-Poulenc, DE 841 750, 1949 (P. Charpentier).
177. P. Charpentier, *C. R. Hebd. Seances Acad. Sci.* **232** (1951) 415.

178. BASF, DE 939 630, 1953 (H. H. Friederich, O. A. Grossinsky, A. Amann).
179. J. Schmitt et al., *Bull. Soc. Chim. Fr.* 1957, 938 – 947.
180. Rhône-Poulenc, DE 1 037 460, 1953 (P. J. C. Buisson, P. Gaillot, J. Gaudechon).
181. Rhône-Poulenc, BE 547 133, 1955.
182. O. C. Dermer, G. E. Ham: *Ethyleneimine and Other Aziridines,* Academic Press, New York 1969, p. 207 ff.
183. D. R. Randell, GB 1 036 696, 1966.
184. J. S. Elliot, E. D. Edwards, A. D. Brazier, GB 889 341, 1962.
185. G. Cauquil, A. Casadevall, *Bull. Soc. Chim. Fr.* 1955, 1061 – 1075.
186. S. P. Massie, I. Cooke, W. A. Hills, *J. Org. Chem.* **21** (1956) 1006 – 1008.
187. C. Bodea, M. Raileanu, *Acad. Repub. Pop. Rom. Fil. Cluj Stud. Cercet. Chim.* **9** (1958) 159 – 166.
188. C. Bodea, M. Terdic, *Acad. Repub. Pop. Rom. Fil. Cluj Stud. Cercet. Chim.* **13** (1962) 81 – 87.
189. V. A. Skorodumov, E. N. Il'chenko, S. V. Zhuravlev, *Zh. Obsh. Khim.* **30** (1960) 1680 – 1683.
190. G. Cauquil, A. Casadevall, E. Casadevall, *Bull. Soc. Chim. Fr.* 1960, 1049 – 1066.
191. T. Dahlgren, A. Hallberg, A. R. Martin, *Heterocycles* **20** (1983) 33 – 38.
192. A. Hallberg, A. Svenson, A. R. Martin, *Tetrahedron Lett.* **27** (1986) no. 18, 1959 – 1962.
193. A. Hallberg, A. R. Martin, *J. Heterocycl. Chem.* **19** (1982) 433 – 435.
194. S. S. C. Chu, H. T. Yang, *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 32** (1976) 2567 –2570.
195. J. Cymerman-Craig, M. F. Tate, *Prog. Drug Res.* **3** (1961) 75 – 150.
196. J. Daneke, H. W. Wanzlick, *Justus Liebigs Ann. Chem.* **740** (1970) 52 – 56.
197. M. V. Diudea, *Rev. Roum. Chim.* **26** (1981) 725 – 732.
198. E. R. Biehl, M. V. Jovanovic, *J. Org. Chem.* **49** (1984) 1905 – 1908.
199. A. C. Schmalz, A. Burger, *J. Am. Chem. Soc.* **76** (1954) 2577 – 2579.
200. H. P. Kaufmann, W. Oehring, *Ber. Dtsch. Chem. Ges.* **598** (1926) 187 – 194.
201. G. Cauquil, A. Casadevall, E. Casadevall, *Bull. Soc. Chim. Fr.* 1960, 1566 – 1572.
202. S. H. Khim, A. R. Martin, *J. Heterocycl. Chem.* **15** (1978) 1507 – 1508.
203. E. J. Corey, M. Ohno, R. B. Mitra, A. P. Vatakencherry, *J. Am. Chem. Soc.* **86** (1964) 478 – 485.
204. E. J. Elliot, H. Guzik, *J. Heterocycl. Chem.* **18** (1981) 861 – 862.
205. Degussa, DE 964 050, 1954 (W. A. Schuler, H. Klebel).
206. H. Bernstein, L. Rothstein, *J. Am. Chem. Soc.* **66** (1944) 1886 – 1888.
207. K. Zehnder, F. Kelberer, W. Kreiss, J. Rutschmann, *Biochem. Pharmacol.* **11** (1962) 535 – 550.
208. E. A. Nodiff, P. C. Taunk, *J. Heterocycl. Chem.* **19** (1982) 1313 – 1314.
209. F. Kehrmann, L. Diesesens, *Ber. Dtsch. Chem. Ges.* **48** (1915) 318 – 328.
210. C. O. Okafor, *J. Heterocycl. Chem.* **17** (1980) 149 – 153.
211. H. Cierpka, DE 1 812 849, 1968.
212. R. Pummerer, F. Eckert, S. Gassner, *Ber. Dtsch. Chem. Ges.* **47** (1914) 1494 – 1507.
213. F. Kehrmann, M. Sandoz, *Ber. Dtsch. Chem. Ges.* **50** (1917) 1667 – 1673.
214. L. Michaelis, S. Granick, M. P. Schubert, *J. Am. Chem. Soc.* **63** (1941) 351 – 355.
215. D. Clarke, B. C. Gilbert, P. Hanson, C. M. Kirk, *J. Chem. Soc. Perkin Trans. 2* 1978, 1103 – 1110.
216. J. P. Billon, *Ann. Chim. (Paris)* **7** (1962) 183 – 206.
217. D. C. Borg, G. C. Cotzias, *Proc. Natl. Acad. Sci. USA* **48** (1962) 617 – 623.
218. H. J. Shine, E. E. Mach, *J. Org. Chem.* **30** (1965) 2130 – 2139.
219. P. Hanson, R. O. C. Norman, *J. Chem. Soc. Perkin Trans. 2* 1973, 264 – 271.
220. G. Dryhurst, K. M. Kadish, F. Scheller, R. Renneberg: *Biological Electrochemistry,* **vol. 1,** Academic Press, New York 1982, p. 548 ff, letter 82 445 840 618.
221. F. Gutman, H. Keyser, *Agressologie* **9** (1968) 27 –35.
222. M. Saucin, A. V. Vorest, J. Duchesne, *Bull. Cl. Sci. Acad. R. Belg.* **59** (1968) 1006 – 1014.
223. P. C. Dwivedi, K. G. Rao, S. N. Bhat, C. N. R. Rao, *Spectrochim. Acta Part A* **31 A** (1975) 129 – 135.
224. Y. Iida, *Bull. Chem. Soc. Jpn.* **44** (1971) 663 – 667.
225. H. Chiou, P. C. Reeves, E. R. Biehl, *J. Heterocycl. Chem.* **13** (1976) 77 – 82.
226. J. J. C. Chu, D. Von Der Helm, *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 32** (1976) 1012.
227. G. Fronza, R. Mondelli, *J. Magn. Reson.* **23** (1976) 437.

228. G. Fronza, E. Ragg, R. Mondelli, *Actual. Chim. Ther. 8th* 1981 245 – 264.
229. J. Fouché, S. Comorisson, *Bull. Soc. Chim. Fr.* 1973, 1693 – 1698.
230. S. Saraf, M. A. Khan, S. Al-Mousawi, *Heterocycles* **19** (1982) 1149 – 1187.
231. S. Saraf, M. A. Khan, S. Al-Mousawi, *Heterocycles* **19** (1982) 935 – 964.
232. N. L. Agrawal, R. L. Mital, *Ann. Soc. Sci. Bruxelles Ser. 1* **89** (1975) 415 – 426.
233. S. Saraf, M. A. Khan, S. Al-Mousawi, *Heterocycles* **20** (1983) 283 – 324.
234. H. J. Haink, J. R. Huber, *Chem. Ber.* **108** (1975) 1118 – 1124.
235. A. M. Duffield, J. C. Craig, L. R. Kay, *Tetrahedron* **24** (1968) 4267.
236. G. McKay, S. F. Cooper, K. K. Midha, *Methodol. Surv. Biochem. Anal.* **16** (1986) 159 – 171.
237. I. Taulov, *Izv. Khim.* **15** (1982) no. 3, 343 – 355.
238. G. C. Hammar, B. Holmstedt, *Anal. Biochem.* **25** (1968) 532 – 548.
239. C. Cimbura, *J. Chromatogr. Sci.* **10** (1972) no. 5, 287 – 293.
240. F. Leterrier, J. Meunier, *Rev. Corps Sante Armees: Terre Mer Air* **12** (1971) no. 4, 427 – 441; *Chem. Abstr.* **76,** 135 415 d.
241. A. M. Chodakowski, *Farm. Pol.* **25** (1969) no. 8, 637 – 641.
242. J. P. Thomas, A. Brun, J. P. Bounine, *Analysis* **17** (1989) 221.
243. J. P. Billon, G. Cauquis, J. Combrisson, *J. Chim. Phys. Phys. Chim. Biol.* **61** (1964) 374 – 381.
244. R. E. Ardrey et al. (eds.): *An Eight Peak Index of Mass Spectra of Compounds of Forensic Interest,* The Scottish Academic Press, Edinburgh 1983, 65 ff.
245. S. H. Curry, *Drugs Psychiatry* **3** (1985) 79 – 87.
246. G. S. Dahl, *Clin. Pharmacol. Psychiatry* 1980, 125 – 137.
247. S. C. Mitchell, *Drug Metabol. Rev.* **13** (1982) no. 2, 319 – 343.
248. J. J. H. McDowell, *Jerusalem Symp. Quantum Chem. Biochem.* **7** (1974) 269 – 300; *Chem. Abstr.* **83,** 77 812 n.
249. J. Paul, C. Buisson, P. Gaillot, J. Gaudechon, FR 1 211 304, 1960.
250. P. A. J. Janssen, C. J. E. Niemegeers, K. H. L. Sdellekens, *Arzneim. Forsch.* **15** (1965) 104.
251. S. Courvoisier et al., *Arch. Int. Pharmacodyn. Ther.* **92** (1953) 305 – 361.
252. L. Julou et al., *Acta Psychiatr. Scand. Suppl.* **241** (1973) 9 – 30.
253. S. L. Tompsett, *Prog. Chem. Toxicol.* **5** (1974) 151 – 185; *Chem. Abstr.* **82,** 132 651 z.
254. A. Cerletti, W. Meier-Ruge, *Proc. Eur. Soc. Study Drug Toxic.* **9** (1969) 170 – 188.
255. R. E. Regal, J. E. Billi, H. M. Glazer, *Clin. Pharm.* **6** (1987) no. 10, 787 – 794.
256. S. Dobric, *Pharmaca* **24** (1986) no. 3, 151 – 155.
257. A. A. Barbely, M. Leepfe-Hinkkanen, *Mod. Pharmacol. Toxicol.* **16** (1979) 403 – 426.
258. I. Jurna: "Neurophysiological Properties of Neuroleptic Agents in Animals," in F. Hoffmeister, G. Stille (eds.): *Handbook of Experimental Pharmacology 55-I, Psychotropic Agents Part I,* Springer Verlag, Berlin 1980, pp. 111 – 175.
259. R. J. Baldessarini: "Drugs in the Treatment of Psychiatric Disorders," in A. G. Gilman, L. S. Goodman, T. W. Roll, F. Murad (eds.): *The Pharmacological Basis of Therapeutics,* 7th ed., MacMillan, New York 1985, pp. 387 – 413.
260. A. Carlsson, *Am. J. Psychiatry* **135** (1978) 164 – 173.
261. D. Naber et al., *Neuropsychobiology* **6** (1980) 181 – 189.
262. L. Julou et al., *Proc. Eur. Soc. Study Drug Toxicol.* **9** (1968) 11 – 51.
263. J. Julou et al., *Arzneim. Forsch.* **19** (1969) 1198 – 1206.
264. F. Gebhart, *Fortschr. Med.* **97** (1979) 45 – 48.
265. F. Leuschner, W. Neumann, R. Hempel: "Toxicology of Antipsychotic Agents," in F. Hoffmeister, G. Stille (eds.): *Handbook of Experimental Pharmacology 55-I, Psychotropic Agents Part I,* Springer Verlag, Berlin 1980, pp. 225 – 265.
266. G. M. Simpson, E. H. Pi, J. J. Srametr, Jr., *Excerpta Med. Sect. 10* **45** (1984) 95 – 108.
267. J. E. Starmark, A. Forsman, J. Wahlström, *Acta Psychiatr. Scand.* **62** (1980) 154 – 157.
268. J. H. Epstein, *J. Am. Acad. Dermatol.* **8** (1983) 141 – 147.
269. E. Cronin: *Contact Dermatitis,* Churchill Livingstone, London 1980, pp. 415 – 460.
270. B. M. Mullock, D. E. Hall, L. J. Shaw, R. H. Hinton, *Biochem. Pharmacol.* **32** (1983) 2733 – 2738.
271. J. L. McGee, M. R. Alexander, *Am. J. Hosp. Pharm.* **36** (1979) no. 5, 633 – 640.

Phenyl Hydrazine → Hydrazine
Phenylalanine → Amino Acids

Phenylene- and Toluenediamines

Robert A. Smiley, Wilmington, Delaware 19880-0336, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 3
5. **Environmental Protection** 4
6. **Quality Specifications** 4
7. **Storage and Transportation** 4
8. **Uses** 4
9. **Toxicology and Occupational Health** . 5
10. **References** 6

1. Introduction

Phenylenediamine is the common name for diaminobenzene, which exists as three isomers: 1,2-, 1,3-, and 1,4-diaminobenzene. The 1,2-isomer is commonly called *o*-phenylenediamine; the 1,3-isomer is *m*-phenylenediamine; and the 1,4-isomer is *p*-phenylenediamine.

The *toluenediamines*, with six possible isomers, are commonly named by using numbers to designate the position of the amino groups with respect to the toluene methyl group, i.e., 2,3-diaminotoluene or 2,3-toluenediamine (3-methyl-1,2-phenylenediamine) [*2687-25-4*]; 2,4-diaminotoluene or 2,4-toluenediamine (4-methyl-1,3-phenylenediamine) [*95-80-7*]; 2,5-diaminotoluene or 2,5-toluenediamine(2-methyl-1,4-phenylenediamine) [*95-70-5*]; 2,6-diaminotoluene or 2,6-toluenediamine (2-methyl-1,3-phenylenediamine) [*823-40-5*]; 3,4-toluenediamine (4-methyl-1,2-phenylenediamine) [*496-72-0*]; and 3,5-toluenediamine (5-methyl-1,3-phenylenediamine) [*108-71-4*]. An 80 : 20 mixture of the 2,4- and 2,6-isomers is used and sold as *m*-toluenediamine. Commercially available *o*-toluenediamine is a 60 : 40 mixture of the 3,4- and 2,3-isomers.

Like most simple aromatic amines, the phenylene- and toluenediamines can be prepared by electrophilic aromatic nitration, followed by reduction of the introduced nitro group (usually catalytic hydrogenation) to an amino group. This chemistry was developed in the early 1800s. Until the mid-1900s, phenylene- and toluenediamines were used principally for the preparation of dyes, but they are currently more important in the production of high-performance textile fibers, agricultural chemicals, and diisocyanates for the preparation of a variety of polyurethane products.

2. Physical Properties

The melting and boiling points of the phenylene- and toluenediamines are listed in Table 1. All of these compounds are white crystalline solids when pure, except 3,5-toluenediamine which is a liquid at room temperature. They are all readily soluble in hot water and common organic solvents such as alcohols, ketones, and ethers. Because of their limited solubility in cold water, *o*- and *p*-phenylenediamines can be recrystallized from water.

Table 1. Melting and boiling points of phenylene- and toluenediamines

Compound	CAS registry no.	*mp*, °C	*bp*, °C
o-Phenylenediamine	[*95-54-5*]	102 – 103	256 – 258
m-Phenylenediamine	[*108-45-2*]	62 – 63	284 – 287
p-Phenylenediamine	[*106-50-3*]	145 – 147	267
2,3-Toluenediamine	[*2687-25-4*]	63 – 64	255
2,4-Toluenediamine	[*95-80-7*]	97 – 99	283 – 285
2,5-Toluenediamine	[*95-70-5*]	64	273 – 274
2,6-Toluenediamine	[*823-40-5*]	104 – 106	
3,4-Toluenediamine	[*496-72-0*]	91 – 93	265 (sub.) ∗
3,5-Toluenediamine	[*108-71-4*]	>0	283 – 284

∗ sub.=sublimation.

3. Chemical Properties

The chemical properties of the phenylene- and toluenediamines are typical of aromatic amines; they react as bases in aqueous solution and form stable salts with acids. Their most notable chemical characteristic is their ease of oxidation to

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_405

complex colored compounds, which is the basis for their use in dyes and hair colorants. Oxidation also occurs if they are stored in the presence of air. The various isomers differ greatly in chemical reactivity due to the relative positions of the two amino groups. Phenylene- and tolueneamines with *ortho* amino groups, for example, readily form heterocyclic compounds, whereas the *m*- and *p*-isomers do not.

The aromatic rings of phenylenediamines can be catalytically hydrogenated to the corresponding diaminocyclohexanes; for example, hydrogenation of *p*-phenylenediamine over a ruthenium on alumina catalyst produces *cis*- and *trans*-1,4-cyclohexanediamine [*15827-56-2*] (cis), [*2615-25-0*] (trans)] in a ratio of 1:3 [1].

***o*-Phenylenediamine.** Benzimidazole [*51-17-2*] is formed by a ring-closure reaction of *o*-phenylenediamine with formic acid. Alkyl-substituted benzimidazoles are obtained by reaction of *o*-phenylenediamine with nitriles or aldehydes in the presence of copper(II) acetate. When *o*-phenylenediamine is reacted with vicinal dicarbonyl compounds, quinoxalines result; reaction with glyoxal [*107-22-2*], for example, gives quinoxaline [*91-19-0*]. Benzotriazole [*95-14-7*] is obtained by the reaction of *o*-phenylenediamine with nitrous acid in dilute sulfuric acid [2].

The oxidation of *o*-phenylenediamine with aqueous iron(III) chloride produces 2,3-diaminophenazine [*655-86-7*]; oxidation with silver oxide in diethyl ether gives *o*-benzoquinonediimine [*4710-40-1*]; and oxidation with air in a solution of copper(I) chloride in pyridine is reported to give 1,4-dicyano-1,3-butadiene [*821-60-3*] [3].

***m*-Phenylenediamine** can be readily acylated with acid chlorides to form substituted amides. With diacid chlorides it forms polyamides; for example, with isophthaloyl chloride a poly(*m*-phenyleneisophthalamide) [*24938-60-1*] is produced, which is the starting material for textile fibers. When *m*-phenylenediamine is reacted with excess nitrous acid, both amino groups are diazotized and couple with unreacted *m*-phenylenediamine to give 5,5′-[(4-methyl-*m*-phenylene)bisazo]bis(toluene-2,4-diamine), Bismarck Brown, one of the first azo dyes.

***p*-Phenylenediamine** is more readily oxidized than the *o*- or *m*-isomer. Acidic manganese dioxide or potassium dichromate produces 1,4-benzoquinone [*106-51-4*]. When *p*-phenylenediamine is oxidized with hydrogen peroxide, for example in a hair dye, the first product is *p*-benzoquinonediimine [*4377-73-5*], which can react with unreacted *p*-phenylenediamine to form a brown dye known as Bandrowsky Base [*20048-27-5*]. The diimine will also react rapidly with other aromatic amines and phenols to form a variety of colorants.

The amino groups of *p*-phenylenediamine undergo reactions typical of aromatic amines, (e.g., diazotization, acylation, alkylation, and phosgenation).

The chemical properties of the toluenediamines are identical to those of the phenylenediamines, i.e., the same type of compounds can be made from them. The methyl group may slow down the reaction rate of an adjacent amine group due to steric hindrance, but this effect is minor in most cases.

4. Production

Preparation of phenylene- and toluenediamines almost always starts with aromatic nitration followed by reduction, although some exceptions exist. The initial starting material in most cases is either benzene or toluene.

o-Phenylenediamine. The principal route to *o*-phenylenediamine is by the reduction of 2-nitroaniline [*88-74-4*], which is produced by the amination of 2-chloronitrobenzene [*88-73-3*] with ammonia [4]. Hydrogenation can be achieved with a variety of reducing agents, such as iron powder [5], hydrazine [6], or hydrogen sulfide [7], but commercially it is performed catalytically in the liquid phase over a palladium catalyst [8], [9]. Other routes to *o*-phenylenediamine include the amination of 1,2-dichlorobenzene [*95-50-1*] with ammonia [10], [11] and the direct amination of aniline [12]. The yield in the latter process is, however, low. A method starting from cyclohexane, sulfur, and ammonia has also been described [13].

m-Phenylenediamine. The *m*-isomer is synthesized by catalytic hydrogenation of 1,3-dinitrobenzene [*99-65-0*], which is obtained as the principal product in the dinitration of benzene with a mixture of sulfuric and nitric acids [14–16]. A certain amount of *o*- and *p*-phenylenediamines (about 10 %) are also produced as byproducts, but these can be removed by chemical treatment of the crude nitration product either prior to hydrogenation [17] or during isolation of the *m*-diamine by distillation. Hydrogenation is usually performed in water or methanol over a supported palladium or Raney nickel catalyst [18], [19].

p-Phenylenediamine. Routes to *p*-phenylenediamine include reaction of 1,4-dichlorobenzene [*106-46-7*] [10], [11], hydroquinone [*123-31-9*], [20], [21], or aniline [*62-53-3*] [22] with ammonia; hydrogenation of 4-nitroaniline [*100-01-6*] [23]; Hoffmann degradation of terephthalamide [*3010-82-0*] [24]; and diazotization of aniline to a triazine followed by rearrangement [25], [26].

The most important commercial routes are the catalytic hydrogenation of 4-nitroaniline and aniline diazotization. The *4-nitroaniline route* is identical to the preparative method for the *o*-isomer; 4-chloronitrobenzene is aminated with ammonia to the nitroaniline, followed by hydrogenation over a palladium catalyst. A disadvantage of this process is the 65 : 35 *ortho – para* ratio obtained during nitration of chlorobenzene because of the difficulty of matching the demand for the fixed ratio of isomers. The diazotization process used by Du Pont in the United States, overcomes this problem because it is based only on aniline.

In the *Du Pont process*, aniline is diazotized using nitrogen oxides [25]. The diazo compound reacts with excess aniline to produce 1,3-diphenyltriazine [*136-35-6*], which rearranges under acidic conditions to 4-aminoazobenzene [*60-09-3*]. 4-Aminoazobenzene is catalytically cleaved and hydrogenated to give *p*-phenylenediamine and aniline. The aniline is recovered and recycled.

One route to *p*-phenylenediamine that does not involve any nitration and does not start with benzene is the *Akzo process*, which can start with dimethyl terephthalate [*120-61-6*] or waste poly(ethylene terephthalate). Either ester can be reacted with ammonia to produce terephthalamide. The diamide is chlorinated to give a chloroamide that is then reacted with a base such as sodium hydroxide to produce *p*-phenylenediamine (Hoffmann degradation) [24].

Toluenediamines. From a commercial standpoint, the most important toluenediamine isomers are the *meta*, or 2,4- and 2,6-isomers, which are obtained as a mixture by hydrogenation of the corresponding dinitrotoluenes. The processes involved are similar to those used for the manufacture of *m*-phenylenediamine, but using toluene as the starting material instead of benzene. About 4 % of the 2,3- and 3,4-isomers are formed as byproducts during toluene dinitration, and the same isomeric diamines subsequently appear as impurities in the mixture of *m*-diamines. They can be removed by distillation and are sold as commercial *o*-toluenediamine.

Table 2. Phenylenediamine product specifications

	Isomer		
	Ortho	*Meta*	*Para*
Water, % (min.)	0.2	0.1	0.1
fp, °C	2.7		
Purity, %	99.5	99.8	99.5
Content *ortho* isomer, ppm		200	1000
Meta isomer, ppm	1 ∗		1000
Para isomer, ppm	1 ∗	100	
Insolubles, % (in water)	0.1		
Nitrobodies as dinitrobenzene, %		0.1	
Aniline, ppm			500

∗ Combined *meta* and *para* content = 0.25 %.

5. Environmental Protection

Aquatic Toxicity. The LC_{50} of *o*-phenylenediamine determined with fathead minnows is 44 mg/L (96 h exposure). For the *p*-isomer it is only 0.028 mg/L, which indicates that *p*-phenylenediamine is very toxic to aquatic life. The LC_{50} of the *m*-isomer is greater than 1 mg/L but less than 50 mg/L [27–29]. The LC_{50} for toluenediamine is 430 mg/L for golden shiners and 1000 mg/L for minnows [30].

6. Quality Specifications

All three *phenylenediamine isomers* are available commercially as flaked solids; the *m*-isomer

is also available as cast solid and molten liquid. The commmercial products are light tan to brown in color. They tend to discolor during storage, and the presence of air, moisture, or elevated temperature hastens color deterioration. However, color alone is not a good criterion of purity because highly colored material may still be above the minimum purity specifications. Typical sales specifications are shown in Table 2.

Toluenediamines are available as cast solids in drums and molten liquids in tank cars. The solid *m*-isomer mixture is yellow to tan, whereas the *o*-mixture is light gray to purple. Like the phenylenediamines, these compounds tend to darken on storage. Product specifications are shown in Table 3.

Table 3. Toluenediamine product specifications

	Isomer	
	Ortho	*Meta*
Purity, % (min.)	97.0	99.0
Normalized isomer ratio, wt %		
2,4-		79 – 81
2,6-		19 – 21
o-	99.5 (min.)	
Other	0.5 (max.)	
Ortho isomers (2,3- and 3,4-)		0.3 (max.)
Para and *meta* isomers (2,5- and 3,5-)		1.0 (max.)
Nitrobodies as dinitrotoluene, %		0.03 (max.)
Moisture, %	1.0 (max.)	0.1 (max.)
Toluidines, %		0.05 (max.)

7. Storage and Transportation

All aromatic diamines should be stored in tightly closed, upright containers in a cool, well-ventilated area away from heat and sources of ignition.

In the United States, the shipping of *o*-phenylenediamine other than by air (DOT) is not regulated. *m*-Phenylenediamine and *p*-phenylenediamine are shipped under hazard class ORM-A and UN no. 1673. For international shipping (IMO) by water or air, all three isomers are shipped under the name phenylenediamine and hazard class Poison β, 6.1.

8. Uses

For about a hundred years, phenylene- and toluenediamines were used mainly in the manufacture of dyes and dye intermediates. However, in the mid-1900s these compounds became very important starting materials for a variety of polymers, which now accounts for most of their consumption.

o-Phenylenediamine. A significant proportion of *o*-phenylenediamine is used for the preparation of benzimidazole-derived agricultural fungicides such as benomyl [*17804-35-2*]. This is produced in the United States by Du Pont and sold under the trade name Benlate, and in the former Soviet Union under the name Uzgen. Other fungicides made from *o*-phenylenediamine include thiophanate methyl (Topsin), a turf fungicide made by Nippon Soda for sale by others, and fuberidazole produced by Bayer.

A small but high-value use of *o*-phenylenediamine is in the preparation of a variety of substituted benzimidazoles used as veterinarian anthelmintics. These compounds are available under trade names such as Equizole (thiabendazole) (Merck, Sharpe & Dohme) and Panacur (fenbendazole) (Hoechst), [31].

Benzotriazole and tolyltriazole are prepared, respectively, from *o*-phenylenediamine and an isomeric mixture of *o*-toluenediamines, by reaction with nitrous acid. These compounds are widely used as corrosion inhibitors in automobile antifreezes and other aqueous solutions because of their specific inhibition of copper and copper alloys.

Mercaptobenzimidazole [*583-39-1*] obtained from *o*-phenylenediamine and potassium ethyl xanthate is used as an antioxidant for rubber (Antioxidant MB, Mobay). A substituted benzotriazole derived from *o*-phenylenediamine is sold by Ciba – Geigy under the trade name Tinuvin P as an ultraviolet light absorber for stabilizing plastics and other organic materials against discoloration and deterioration [32].

m-Phenylenediamine. The fastest growing use for *m*-phenylenediamine is in the production of the aramid fiber poly(*m*-phenyleneisophthalamide), which is obtained by reaction of *m*-phenylenediamine with isophthaloyl chloride [*99-63-8*] [33]. This material is used in areas where high-temperature resistance or fire retardancy is required. Examples are protective clothing and aircraft interiors. The polymer is made in the United States by Du Pont (Nomex), in Japan

by Teijin, and in the former Soviet Union. *m*-Phenylenediamine is also used for curing epoxy resins to impart high-temperature strength and good resistance to chemical solvents. Such cured resins are used in filament-wound casings and adhesives.

m-Phenylenediamine is still used as the starting material for many dyes. Examples include Basic Brown 1 (Bismarck Brown), Basic Orange 2, Direct Black 38, and Developed Black BH.

p-Phenylenediamine. Aramid textile fibers are obtained from poly(*p*-phenylene terephthalamide) [*24938-64-5*], which is produced by the reaction of *p*-phenylenediamine with terephthaloyl chloride [*100-20-9*] [33–35]; the two suppliers are Du Pont (Kevlar) and Akzo (Twaron). These fibers are used for reinforcing tires and V-belts, as a substitute for asbestos in brake linings, in reinforcing plastics for aerospace and sporting applications, and for a variety of other uses where high strength and tensile modulus are required [36].

Reaction of *p*-phenylenediamine with phosgene produces *p*-phenylene diisocyanate (PPDI), a symmetrical diisocyanate that gives highly crystalline polyurethanes reported to have good high temperature properties as both thermoplastics and cast elastomers [37].

p-Phenylenediamine is still employed in hair dyes and is the main component of oxidation hair dyes. *N,N′*-disubstituted *p*-phenylenediamines are used extensively as photographic chemicals, as fuel additives, and in the compounding of rubber. However, most of these compounds are made by the reductive alkylation of *p*-nitroaniline rather than from the parent diamine.

Toluenediamines. The largest use for toluenediamines is in the production of diisocyanates for the polyurethane industry. These diisocyanates are synthesized by reaction of *m*-toluenediamine (80 : 20 mixture of the 2,4- and 2,6-isomers) with phosgene. The resulting diisocyanates (trade name TDI) are used principally in polyurethanes as flexible foams for upholstery and other types of padding.

9. Toxicology and Occupational Health

All aromatic diamines are toxic when ingested and may cause skin and eye irritation. In some individuals, inhalation can lead to respiratory problems and asthma. Thus, as a class these diamines should be treated as very hazardous materials, and all contact with the body should be avoided by use of appropriate protective equipment.

***o*-Phenylenediamine.** In addition to causing skin rashes, evidence exists that *o*-phenylenediamine can permeate the skin in toxic amounts. A single application to the skin of rabbits produced weight loss, loss of coordination, weakness, darkening of the urine, and death. Currently, this compound is not considered a carcinogen, but it produces noninheritable genetic damage in bacterial and mammalian cell cultures [27].

***m*-Phenylenediamine.** The toxic effects observed in animals from short exposure to *m*-phenylenediamine by inhalation, ingestion, or skin contact include liver and kidney damage. Since no carcinogenic activity has been noted in animal tests, *m*-phenylenediamine is not considered a carcinogen, although it is mutagenic in bacterial cultures and may have embryotoxic activity [28].

***p*-Phenylenediamine.** Toxic effects observed in animals from single exposures to *p*-phenylenediamine by inhalation include tremors, cyanosis, and severe weight loss. Toxic effects by ingestion include weight loss and lethargy. This compound does not have carcinogenic effects on animals. Tests for reproductive activity have not yet been performed [29].

p-Phenylenediamine is such a strong dye-former in the presence of air and moisture that everything in contact with it, whether skin, clothes, or walls, is almost immediately dyed a dark blueish black. Thus, exposure to *p*-phenylenediamine in even small amounts or lack of good industrial hygiene in operations that involve its use can be readily detected.

***o*-Toluenediamines.** The toxic inhalation and ingestion effects of the commercial mixture of *o*-toluenediamines can be assumed to be similar to those of phenylenediamines, but they do not appear to be primary skin irritants. The Ames mutagenic assay of *o*-toluenediamines is positive. Because the mixture can contain up to 0.5 %

of the *meta* isomers, the toxic effects of these compounds must be considered when working with commercial *o*-toluenediamine.

***m*-Toluenediamine.** Application of *m*-toluenediamine to the skin in hair dyes does not appear to produce toxic or carcinogenic effects [38], but it does appear to be a carcinogen when ingested by rats [39]. It is an active mutagen in the Ames test.

10. References

1. Du Pont, US 3 636 108, 1977 (L. Brake).
2. J. B. Wright, *J. Am. Chem. Soc.* **71** (1949) no. 5, 203.
3. D. Barton, W. D. Ollis in I. O. Sutherland (ed.): *Comprehensive Organic Chemistry,* **vol. 2,** Pergamon Press, Oxford 1979, p. 1 77.
4. Du Pont, US 3 929 889, 1975 (E. N. Squire).
5. GAF, US 2 956 082, 1960 (L. M. Schenck).
6. C. Budeanu, RU 60 417, 1976.
7. S. Bogdal, M. Milewska, *Przem. Chem.* **58** (1979) no. 9, 481.
8. Hoechst, BR 1 344 796, 1974 (R. Mees, J. Ribka).
9. UOP, US 3 230 259, 1966 (B. Levy).
10. Toyo Soda, JP 5 3 77 023, 1978 (M. Ohshio et al.).
11. Japan Synthetic Rubber, US 4 521 622, 1985 (N. Andoh, S. Fujiwara).
12. Du Pont, BR 1 327 494, 1973 (E. N. Squire).
13. F. J. Weigert, *J. Org. Chem.* **46** (1981) no. 9, 1936.
14. Atlantic Refining, US 2 951 571, 1960 (G. A. Bonetti).
15. A. D. Mesurc & J. G. Tillett, *J. Chem. Soc. Phys. Org.* **7** (1966) 669.
16. A. Ujhidy, M. Magyar, R. Berkes, *Acta Chem. (Budapest)* **69** (1971) no. 1, 107.
17. Du Pont, US 3 329 719, 1967 (D. J. Crowley).
18. Du Pont, US 3 328 465, 1967 (L. Spiegler).
19. Teijin, JP 45 490, 1970 (T. Yamanaka).
20. Olin Mathieson, BR 1 019 750, 1966.
21. Du Pont, US 4 031 106, 1977 (T. W. Del Pesco).
22. Asahi, JP 7 7 42 829, 1977 (S. Enomoto, T. Kamiyama, Y. Takahashi, E. Fujimoto).
23. Sumitomo, DE 2 331 900, 1974.
24. Akzo, BR 1 364 229, 1974 (H. Zengel, M. Bergfeld); DE 2 216 117, 1973 (H. Zengel, M. Bergfeld).
25. Du Pont, US 4 020 052, 1977 (J. K. Detrick).
26. Du Pont, US 4 279 815, 1981 (F. F. Herkes).
27. Du Pont, *o*- phenylenediamine, Material Safety Sheet, E. I. du Pont de Nemours & Co. Inc. (Environmental Affairs), Wilmington, Del., 1987.
28. Du Pont, *m* -Phenylenediamine, Material Safety Data Sheet, E. I. du Pont de Nemours & Co. Inc. (Chemicals & Pigments Dept.), Wilmington, Del., 1987.
29. Du Pont, *p*- phenylenediamine, Material Safety Sheet, E. I. du Pont de Nemours & Co. Inc. (Environmental Affairs), Wilmington, Del., 1987.
30. Air Products, Nitro/Amino Aromatic Intermediates, Product Bulletin, Air Products and Chemicals, Inc., Allentown, Pa., 1983.
31. *Veterinary Pharmaceuticals and Biologicals,* Harwal Publishing Co., Media, Pa., 1980 – 1981.
32. *Merck Index,* 11th ed., Merck & Co, Rahway, N. J., 1989.
33. Du Pont, US 375 361, 1973 (J. H. Jensen).
34. Teijin, JP 4 8 23 198, 1973 (S. Ozawa).
35. Du Pont, US 3 869 430, 1975 (H. Blades).
36. Akzo, BR 1 547 802, 1979.
37. Du Pont, Paraphenylenediisocyanate, Product Bulletin, E. I. du Pont de Nemours & Co., Inc. (Petrochemicals Dept.), Wilmington, Del., 1990.
38. C. Burnett et al., *Food Cosmet. Toxicol.* **13** (1975) 353.
39. R. H. Cardy, *JNCI, J. Natl. Cancer Inst.* **26** (1979) no. 4, 1107.

Pheromones → Biological Control
Pheromones → Crop Protection
Pheromones → Insect Control
Phloroglucinol → Phenol Derivatives

Phosgene

Wolfgang Schneider, Bayer AG, Dormagen, Federal Republic of Germany (Chaps. 1 – 9)

Werner Diller, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 10)

1. Introduction 1
2. Properties 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.2.1. Reactions with Organic Compounds . 2
2.2.2. Reactions with Inorganic Compounds 3
3. Production 3
4. Storage and Transportation 4
5. Quality and Analysis 5
6. Waste-Gas Treatment and Monitoring 5
7. Safety Precautions 6
8. Uses 6
9. Economic Aspects 8
10. Toxicology and Occupational Health 8
11. References 9

1. Introduction

Phosgene [*75-44-5*], $COCl_2$, (carbonic acid dichloride, carbonyl chloride), M_r 98.82, was discovered in 1812 by J. Davy from the action of sunlight on a mixture of carbon monoxide and chlorine. The name phosgene is derived from the Greek words for "light" and "generate". Phosgene is a colorless gas, which condenses under normal pressure at 7.56 °C to form a colorless liquid.

The importance of phosgene, apart from its use as a warfare agent in World War I [1], was initially its various uses in the production of intermediates in many different fields of chemistry. Currently, now that the polyurethane plastics have become accepted, most phosgene produced is used in the manufacture of diisocyanates.

Because of its high toxicity and wide range of commercial applications, exceptional effort has been devoted worldwide to the safe handling of phosgene. This high level of safety technology is being developed further. To avoid safety problems, costly work has been performed to develop phosgene-free processes for isocyanate production. Up to now, however, these have gained only partial acceptance on the commercial scale (see Chap. 8).

2. Properties

2.1. Physical Properties

Important physical properties of phosgene may be summarized as follows:

mp	−127.8 °C
bp (101.3 kPa)	7.56 °C
Critical data	
Critical pressure	5.674 MPa
Critical temperature	182 °C
Critical density	0.52 g/cm^3
Density at 20 °C	1.38 g/cm^3
Surface tension	
at 16.7 °C	2.11×10^{-2} N/m
at 34.5 °C	1.76×10^{-2} N/m
at 46.1 °C	1.59×10^{-2} N/m
Specific heat	
Gas at 25 °C	6.17 J kg^{-1} K^{-1}
Liquid at 7.5 °C	1026 J kg^{-1} K^{-1}
Enthalpy of formation at 25 °C	
from the elements	− 220 kJ/mol
from CO and Cl_2	− 107.6 kJ/mol
Heat of evaporation	
at *bp*	24.2 kJ/mol
Heat of fusion	5.740 kJ/mol
Ebullioscopic constant	29

The phosgene molecule is planar, with atomic distances of 1.166×10^{-10} m for C – O and 1.746×10^{-10} m for C – Cl. The Cl – C – Cl bond angle of 111.3° is close to the tetrahedral angle (method of determination: microwave spectroscopy [2]).

Phosgene is readily soluble in many organic solvents and itself dissolves many organic substances. The only inorganic substances which

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_411

it dissolves are the halogens and a few halides ($AlCl_3$, $SbCl_3$, $SbCl_5$) [3]. Activated carbon adsorbs large amounts of phosgene.

Vapor pressure and density of phosgene are given in Table 1 [4]. The following properties are compiled in graphs over fairly wide temperature ranges in [5]: vapor pressure, heat of vaporization, specific heat, density, viscosity, surface tension, and thermal conductivity. Spectra of phosgene are given in the following references: IR spectra [6], [7]; UV spectra [8]; Raman spectra [6], [9]; photoelectron spectra [10].

Table 1. Vapor pressure and density of liquid and gaseous phosgene according to [4]

ϑ, °C	p, MPa	ϱ_L, g/cm^3	ϱ_G, g/cm^3
10	0.109	1.405	0.005
50	0.404	1.306	0.016
100	1.38	1.165	0.046
150	3.49	0.966	0.142

2.2. Chemical Properties

Dissociation of phosgene to carbon monoxide and chlorine is noticeable at 200 °C and practically complete at 800 °C. At 300 °C it is about 5 %, and at 400 °C about 20 % (for further values, see [11]). The temperature dependence of the equilibrium constant of the dissociation reaction and other thermodynamic functions are given in [12].

A survey of the reactions of phosgene is given in [13]. Literature sources cited there are repeated here only if they are survey articles.

In order to avoid the difficulties associated with the toxicity of phosgene, substitutes for phosgene can sometimes be used in the laboratory e.g., trichloromethyl chloroformate ("diphosgene") [14–17] and bis(trichloromethyl) carbonate ("triphosgene") [18].

2.2.1. Reactions with Organic Compounds

Being the dichloride of carbonic acid, phosgene is very reactive and reacts with many other substances to form a wealth of products, which are usually also very reactive. The importance of phosgene for the production of synthetic intermediates is based on the large number of possible coreactants.

The reaction with *primary amines* is by far the most important reaction of phosgene, which made it become a large-scale product of the chemical industry. The first step is the formation of carbamoyl chlorides, from which isocyanates are formed by elimination of HCl at elevated temperature:

$$R-NH_2 \xrightarrow[-HCl]{COCl_2} R-NH-COCl \xrightarrow[-HCl]{} R-NCO$$

This scheme is used in the production of the isocyanates used for intermediates and pesticides (→ Isocyanates, Organic, Chap. 4.1.) and of all diisocyanates important for polyurethane chemistry, such as toluene diisocyanate [*1321-38-6*] and methylenediphenyl diisocyanate monomer [*101-68-8*] and polymer [*9016-87-9*] [19]. Aromatic *ortho*-diisocyanates can also be obtained via intermediate formation of the corresponding benzimidazolone [20]:

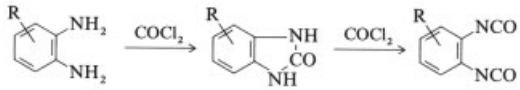

With *secondary amines*, phosgene forms the fairly stable *N, N* - dialkylcarbamoyl chlorides:

$$R_2NH \xrightarrow[-HCl]{COCl_2} R_2N-COCl$$

Secondary amides react with phosgene to form imidoyl chlorides, which are of preparative importance because of their reactivity:

$$RCONHR^1 \xrightarrow[-HCl,\ -CO_2]{COCl_2} R-\overset{\overset{Cl}{|}}{C}=NR^1$$

Tertiary amines add to phosgene in the cold to form cationic complexes, which eliminate alkyl chlorides upon heating:

$$R_3N \xrightarrow{COCl_2} [R_3NCOCl]^+Cl^- \longrightarrow R_2NCOCl + RCl$$

With excess tertiary amine, tetrasubstituted ureas are formed in excellent yields:

$$2\,R_3N \xrightarrow{COCl_2} R_2N-CO-NR_2 + 2\,RCl$$

Nitriles react with phosgene to form heterocyclic compounds with the parent ring systems

of isoquinoline, pyrimidine, and triazine (cf. also [21]).

Phosgene reacts fairly rapidly with *alcohols,* and more slowly with *phenols*, to form chloroformic esters (→ Chloroformic Esters, Chap. 4.). This reaction is of considerable industrial importance (a survey is given in [22]). The reaction is usually carried out in the presence of acid acceptors such as amines or inorganic hydroxides. It can also be accelerated catalytically (with *N, N*-dialkylamides or quaternary ammonium salts as catalysts).

$$\mathrm{ROH} \xrightarrow[-\mathrm{HCl}]{\mathrm{COCl_2}} \mathrm{RO-COCl}$$

Carboxylic acids react with phosgene to form acid chlorides in the presence of catalysts such as *N*, *N*-dialkylcarbamides or tertiary amines.

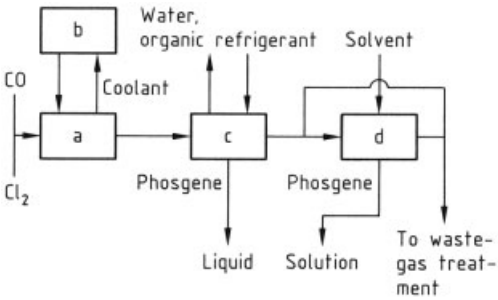

Figure 1. Simplified flow sheet for the production of phosgene
a) Reaction; b) Recooling of the coolants; c) Phosgene liquefaction; d) Phosgene absorption

2.2.2. Reactions with Inorganic Compounds

Phosgene reacts slowly with *water* in the cold, but more rapidly when hot to yield carbon dioxide and hydrochloric acid. The reaction is accelerated by acids [23–30].

Ammonia reacts with phosgene to form urea and ammonium chloride.

A large number of *metallic oxides* are converted by phosgene to their chlorides at elevated temperatures. This reaction is commercially important for the production of aluminum chloride (→ Aluminum Compounds, Inorganic, Chap. 4.1.2.).

Compounds analogous to phosgene are obtained by reaction of phosgene with certain *halides*; carbonyl fluoride is formed with nitrogen trifluoride (NF_3), carbonyl bromide with aluminum bromide, and " carbonyl isocyanate" in a salt melt with potassium isocyanate [31]:

$$\mathrm{COCl_2 + 2KNCO} \xrightarrow{400\,^\circ\mathrm{C}} \mathrm{O{=}C(NCO)_2 + 2KCl}$$

Phosgene reacts with calcium fluoride at 500 °C to form fluoromethane and the various fluorochloromethanes.

3. Production

Phosgene is obtained commercially by passing carbon monoxide and chlorine over activated carbon. The enthalpy of formation is −107.6 kJ/mol, i.e., the reaction is strongly exothermic (for the kinetics of catalytic phosgene formation, see [32], [33]). Carbon monoxide is added in stoichiometric excess to keep the free chlorine content of the phosgene as low as possible, since chlorine can lead to the formation of undesirable products during further processing.

Process Description. The reaction is carried out in a multitubular reactor of carbon or stainless steel, whose tubes are filled with the granulated catalyst (particle size 3 – 5 mm). The tube diameter is 50 – 70 mm. The temperature of the activated carbon bed is up to ca. 400 °C in the very lowest region of the tubes, but it then falls rapidly along the tubes to product temperatures of 40 – 150 °C.

The heat of reaction is dissipated by water or an organic heat - carrier in the shell of the reactor. A flow sheet of the production process is given in Figure 1.

In a variant, high-boiling cooling liquids (*bp* ca. 200 – 300 °C) are used as coolants by the evaporative cooling principle. The vapors of these heat carriers are condensed with water, which is then vaporized in turn and used in this form as a heating medium, e.g., in other plants. Since the carbon monoxide – chlorine conversion is incomplete at the high reaction temperatures the reaction product is subsequently led through conventional reactors [34].

The reaction is carried out under normal pressure or slight excess pressure. The gaseous phos-

gene is obtained at a temperature of 40 – 150 °C. It is either processed directly, or condensed with water or a coolant. Noncondensable parts of phosgene are absorbed directly with solvent beforehand to form phosgene solution. Nonabsorbable gaseous components (inert gas, CO, phosgene) are fed to waste-gas treatment (see Chap. 6). The activated carbon catalyst can have a service life of up to several years. The service life depends on the purity of carbon monoxide.

Chlorine is usually produced from chlor – alkali electrolysis and HCl electrolysis or from evaporators in which liquid chlorine is evaporated. Usually chlorine is fed through a preheater before mixing with carbon monoxide in order to avoid introducing liquid chlorine into the phosgene generator.

Carbon monoxide is usually produced from natural gas or naphtha in steam reforming plants, but also partly from coke, from which it is obtained by reaction with CO_2 and oxygen. Interfering byproducts are methane and hydrogen in CO from the reformer, and sulfur compounds and hydrogen (from the water content of the coke) in CO from coke. The content of methane and hydrogen is of great importance for safety, because it can lead to spontaneous exothermic HCl formation on mixing with CO and chlorine. The temperature rise associated with this involves the risk of reaction of chlorine with the pipe material (chlorine – iron fire). The presence of sulfur compounds (e.g., COS, $CSCl_2$) can lead to deterioration in the quality of the resulting products during the further use of the phosgene.

Phosgene can be formed at high temperatures from chlorinated hydrocarbons with atmospheric oxygen. This reaction can lead to possible health hazards, above all in welding technology [35–38].

4. Storage and Transportation

Storage and transportation of phosgene are avoided as far as possible for safety reasons. Generally, only the minimum amounts necessary in the course of the process are stored at intermediate stages. Steel vessels are suitable containers if corrosive impurities, especially water, can be excluded. Phosgene is shipped almost exclusively on the kilogram scale in seamless drawn steel cylinders or in pressure drums which are regularly checked at specific intervals.

Transport classifications for phosgene are as follows:

IMDG Code (GGVSee)	Class 2 D 2227
	E-F 2106
Emergency Schedule (EmS)	2-08
Medical First Aid Guide for Use in Accidents Involving Dangerous Goods (MFAG)	600
RID (GGVE)	Class 2 Rn 201 no. 3at
ADR (GGVS)	Class 2 Rn 2201 no. 3at
ADNR	Class 2 Rn 6201 no. 3at
EC Directive	No. 006-002-00-8
USA CFR 49	§ 172.101 Poison A
ICAO/IATA-DGR	Class 2
UN no.	1076 forbidden

5. Quality and Analysis

The phosgene obtained from production in the presence of activated carbon is so pure that an additional purification step is usually unnecessary. Possible impurities are e.g., free chlorine, iron chloride, carbon tetrachloride, polychlorinated hydrocarbons, hydrogen chloride, and sulfur compounds. Phosgene can be purified by distillation when required.

The quality of phosgene (liquid or in concentrated solution) can be determined by the following methods:

1) By gas chromatography [39]
2) By IR spectroscopy [40]
3) By titration. Iodine is liberated from a NaI – acetone solution by phosgene and from an aqueous NaI solution by chlorine which may be present as an impurity. The iodine is then titrated with thiosulfate. Other possible impurities are determined as follows: $FeCl_3$ is determined by reacting the sample with potassium iodide solution after pretreatment. Hydrogen chloride is determined by titration with NaOH [41].
4) Total chlorine is determined as chloride after saponification with NaOH. Separate determination is possible of *free chlorine*; e.g., by evaporation of the phosgene, absorption

in a solution of arsenious acid, and back-titration with iodine solution; and *free HCl* by Liebig titration of the hydrogen cyanide released from mercury(II) cyanide.

5) Phosgene can be reacted with an aniline–toluene solution to form diphenylurea which can then be determined qualitatively and quantitatively [42].

For the determination of phosgene in gaseous mixtures see Chapter 6.

6. Waste-Gas Treatment and Monitoring

The residual gases from phosgene production and the off-gases from plants for phosgene processing must be freed from phosgene in a decomposition unit. This is achieved by various methods:

Decomposition with Alkaline Scrubbing Agents, mainly Sodium Hydroxide Solution. The residual gas is led through packed towers, at the top of which sodium hydroxide solution is introduced. Make-up sodium hydroxide solution is added under pH control. Venturi scrubbers can also be used.

Decomposition in Activated Carbon Towers. The residual gas is passed through towers that are filled with activated carbon and fed with water. Water and residual gas are usually fed cocurrently to avoid flooding of the packing. The water is usually recycled. If the concentration of hydrochloric acid which is formed by decomposition of phosgene exceeds 10 %, the decomposition effect is greatly reduced. Therefore a sufficient supply of fresh water must be assured [43]. For safety reasons frequently two towers in series are used.

Combustion. The residual gases are burnt so as to convert phosgene to CO_2 and HCl only. An advantage of this method is that all remaining pollutants in the residual gas, such as solvents and carbon monoxide, are burnt together with phosgene [44].

Normally, the outlet gas from the phosgene destruction equipment is continuously monitored for residual phosgene content. The following methods of detection are used:

1) Gas detectors with test tubes [45]
2) Photometric methods: (1) IR spectrometry, preferably for waste-gas monitoring in the concentration ranges 0 – 100 and 0 – 1000 mL/m^3 [46], [47]. (2) Devices whose measuring principle is based on the colorimetric evaluation of the Dixon and Hands [48] color reaction [49–52]. The measuring ranges are 0 –250 µL/m^3 and 0 – 100 µL/m^3. These devices are also used for room air monitoring.
3) Ionization methods [53], [54]
4) Gas chromatography [39]
5) Electrochemical methods [55], for use in the range of 0 – 300 µL/m^3

Methods in development will permit line and area monitoring ("fence line monitoring") in the area around phosgene plants using Fourier Transform IR spectrometry (FTIR) [50], [56].

7. Safety Precautions

When handling phosgene, extensive safety precautions are required to prevent exposure to phosgene. These measures can be divided into

1) Primary technical measures
2) Secondary technical measures
3) Organizational measures
4) Personal precautions

Primary Technical Measures. The primary technical measures are aimed at preventing phosgene emission from the closed equipment. This is achieved by e.g., selection of resistant, high-grade materials for equipment and lines; careful testing of this equipment and lines, especially the welds; leak tests before resuming operation of plant units; the use of pumps which are hermetically sealed to the outside (canned motor pumps, magnetic pumps); the use of high-grade materials for flange gaskets; and regular inspections of equipment and lines. In the field of process control engineering, particularly high requirements are set for the quality and design of process control equipment. Plant equipment which are particularly important for safety have duplicated process control instrumentation. Automatic systems ensure immediate safe shutdown of the plant when important process parameters move outside a narrow bandwidth. In addition, the probability of phosgene emission

and the emission potential are reduced further by continual optimization of the process system. In this way, e.g., the number of vessels containing phosgene and their phosgene content are reduced.

Secondary Technical Measures. Secondary technical measures are used to detect leaks as soon as possible and to combat escaped phosgene. These measures include, e.g., continuously operating alarm systems for monitoring room air and ambient air, systems for combating escaped phosgene by chemical reaction (e.g., steam – ammonia curtains in the case of gaseous emissions), jacketed pipes, and complete containments for phosgene processing plants or plant units. Leaks of liquid phosgene or phosgene solution can effectively be combated by covering the phosgene-containing liquid with sorption and destruction agents.

Organizational Measures. The organizational measures ensure that the staff operate the plants safely and initiate rapid countermeasures following malfunctions or the detection of leaks. These measures include special training, routine plant inspection, training for exceptional situations, clear operating instructions, and regular safety discussions.

Personal Safety Precautions. The personal safety precautions ensure that the staff are protected from exposure during a possible release of phosgene into the atmosphere and that a possible exposure is detected early enough to prevent inhalation of a dose which would be dangerous to health. The staff are therefore equipped with escape masks or respirators. For other personal safety precautions see section 10.

8. Uses

Its many different reactions makes phosgene an important starting compound in the production of intermediates and end products in many branches of large-scale industrial chemistry.

Diisocyanates. Most phosgene is used for the production of diisocyanates as starting materials of polyurethane chemistry. The most important compounds are toluene diisocyanate (TDI) [*1321-38-6*] as well as monomeric [*101-68-8*] (MDI) and polymeric methylenediphenyl diisocyanate [*9016-87-9*] (PMDI) (→ Isocyanates, Organic, Chap. 3.).

The production of *toluene diisocyanate* (TDI) starts from toluene, which is converted in a continuous process to a mixture of 2,4- and 2,6-dinitrotoluene. This mixture is hydrogenated with hydrogen to give 2,4- and 2,6-toluenediamine (TDA). The reaction with phosgene is usually carried out in a continuous process, using e.g., toluene, chlorobenzene, or 1,2-dichlorobenzene as solvents, at 65 – 180 °C and at excess pressure. In the course of this process hydrogen chloride (reaction product), phosgene (excess reactant), and finally the solvent are successively separated. The end product TDI is purified by distillation.

For the production of *methylenediphenyl diisocyanate* (MDI), aniline is used as the starting material. It is condensed with formaldehyde to form a mixture of isomeric methylenediphenyldiamines (MDA). The subsequent reaction of this amine mixture with phosgene is similar to that described above for TDI. The end product is a mixture of isomeric MDI and various PMDI types. Distillative purification is not carried out.

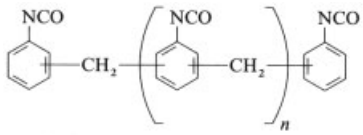

$n = 0$ MDI
$n > 0$ PMDI

The high toxicity of phosgene has resulted in intensive work on the development of phosgene-free processes for diisocyanate manufacture. These processes start from the corresponding dinitro compounds, which react with CO to form diisocyanates [57–62]. The reaction of the amines with compounds other than phosgene such as dimethyl carbonate [63] or methanol and carbon monoxide with oxygen [64], is another line of development. The development of phosgene-free processes has been successful for only a few aliphatic diisocyanates; the production of other diisocyanates by this route suffers from a lack of profitability or process engineering difficulties (→ Isocyanates, Organic, Chap. 4.3.) [65], [66].

Aliphatic diisocyanates are important starting materials for polyurethanes used in coatings. Monoisocyanates, such as isopropyl isocyanate, phenyl isocyanate, and chlorophenyl

isocyanates, are important intermediates in the manufacture of herbicides and insecticides.

Polycarbonates. The next largest phosgene consumers are polycarbonate plastics (→ Polycarbonates). They are produced by reaction of 2,2-bis(4-hydroxyphenyl)propane (bisphenol A) with phosgene, either directly or indirectly via diphenyl carbonate.

Chloroformic Esters. The reaction of phosgene with alcohols to form chloroformic esters is very important for industrial applications. These esters are exceptionally versatile intermediates, e.g., for the production of carbonic esters, polycarbonates, and polyurethanes, and for many other applications (→ Chloroformic Esters).

Urea Azo Dyes. The treatment of aminoazo dyes with phosgene in aqueous solution to produce symmetrical urea azo dyes, has been practiced since the end of the 19th century (→ Azo Dyes, Chap. 4.1.6.).

Carbonate Esters. Among the carbonate esters, dimethyl and diethyl carbonates in particular are used in the pharmaceutical industry and in the production of pesticides. Diphenyl carbonate is an intermediate in the production of polycarbonates.

Friedel – Crafts Acylations. Friedel – Crafts acylations with phosgene are important reactions, e.g., for the dye industry, yielding important ketones such as benzophenone (→ Ketones, Chap. 7.3.) or 4,4′-bis(dimethylamino)benzophenone (Michler's ketone) (→ Acylation and Alkylation, Chap. 2.2.). Benzoyl chloride can also be obtained in this way from benzene. The production of 2,5-dichlorobenzoic acid from 1,4-dichlorobenzene and phosgene in an $AlCl_3$ – sodium formate melt is described in [67].

Other Uses. Phosgene undergoes commercially important cyclization reactions with aminocarboxylic acids, aminophenols, and diamines. Acid chlorides are often prepared by reaction of carboxylic acids with phosgene (for suitable catalysts see [68]). Examples are the production of isophthaloyl dichloride and terephthaloyl dichloride on the large industrial scale.

Phosgene disproportionates to form carbon tetrachloride and carbon dioxide at high temperatures (→ Chlorinated Hydrocarbons).

In *inorganic chemistry*, phosgene is used as an intermediate for the large-scale production of aluminum chloride (→ Aluminum Compounds, Inorganic, Chap. 4.1.2.). The production of zirconium chloride is described in [69].

9. Economic Aspects

In 1989 the estimated world production of phosgene was 2.74×10^6 t. This was divided between fields of application approximately as follows (figures in 10^6 t):

TDI	1.18
MDI	0.96
Polycarbonate	0.30
Aliphatic diisocyanates	0.10
Monoisocyanates	0.05
Chloroformic esters	0.05
Other fields of application	0.10

The main producers of phosgene are identical with the large TDI and MDI manufacturers. In the United States these are Dow, Mobay (Bayer), Rubicon (ICI), BASF, and Olin, and in Western Europe Bayer, BASF, Dow, ICI, Rhône-Poulenc and Montedipe. About 40 % of the world production of phosgene is manufactured in Western Europe, ca. 37 % in the United States, ca. 11 % in Japan, and ca. 7 % in Eastern Europe. Latin America and the Far East (excluding Japan) make up the remaining 5 %. The growth rate of production largely corresponds to the growth rate of TDI and MDI production and is estimated at 4 % per year for the 1990s.

10. Toxicology and Occupational Health

Phosgene is a typical lung irritant gas without generalized systemic absorption. Because of its low water solubility, phosgene reacts only to an insignificant extent with the aqueous film on the mucous membranes of the upper respiratory tract. Most of the inhaled phosgene reaches the pulmonary alveoli, where gas exchange occurs. Here the phosgene reacts with NH_2, OH, and SH groups of the blood – air barrier [70]:

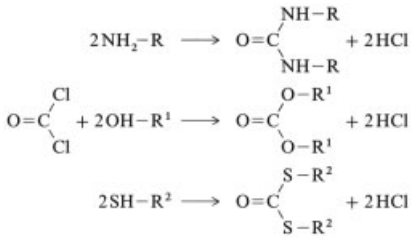

In the course of this, nucleic acid bases and enzymes are modified and the blood – air barrier loses its function as a membrane between the blood vessels and the pulmonary alveoli. As a result, blood plasma is able to pass from the blood vessels into the pulmonary alveoli (lung edema) thereby increasingly disturbing the gas exchange. Death can result from "internal suffocation".

In the very rare cases of inhalation of extremely high doses of phosgene, unreacted phosgene may penetrate the blood – air barrier, reach the blood vessels, and react with blood constituents. Red blood cells break down and the fragments block up the lung capillaries. The cessation of capillary circulation can lead to death within a few minutes [71].

The clinical picture of phosgene poisoning varies according to the inhaled concentration or dose [71], [72]. Death by heart failure due to blockade of the pulmonary circulation can occur within a few minutes after inhaling excessively high doses. Typically, however, medium-size doses are inhaled, and the poisoning then follows a prolonged course. Characteristic is a clinically silent "latency phase" in which the liquid gradually accumulates in the lungs until finally signs and symptoms of pulmonary edema result. These include cough with production of large amounts of liquid phlegm. Finally, a liquid escapes from the mouth and nose which becomes increasingly frothy and pink- colored.

The length of the clinically silent latency period is inversely proportional to the inhaled dose [73]: after high doses symptoms and discomfort can occur after 2 – 3 h, but after smaller doses these occur after 12 – 20 h. Early indications of a gradually developing pulmonary edema can be provided by a radiography of the lungs, even in the clinically silent latent phase [73]. Initial transient irritation phenomena, such as burning eyes, burning throat, and dry cough, occur only for a short while after inhalation of quite high concentrations (> 3 ppm).

Direct contact with liquid phosgene is said to lead to severe irritation or burning of the skin and mucous membranes [74], [75].

For humans the following concentration- or (dose-) effect relations apply [76]:

TLV, MAK value	0.1 ppm
Odor detection	ca. 0.4 ppm
Odor identification	ca. 1.5 ppm
Irritation in eyes, nose, throat, and bronchial tubes	> 3.0 ppm
Beginning of lung impairment (for concentrations > 0.15 ppm)	ca. 30 ppm · min
Clinical pulmonary edema	> 150 ppm · min
Lethal dose (concentration × time, Lct)	
Lct minimum	ca. 300 ppm · min
Lct_{50}	ca. 500 ppm · min
Lct_{100}	ca. 1300 ppm · min

The characteristic phosgene odor is said to be reminiscent of freshly mown grass or fresh hay. With increasing phosgene concentration, however, a pungent odor dominates. Since odor detection or odor identification is only possible at phosgene concentrations above the TLV or MAK value, the phosgene odor does not provide sufficient warning. Life-threatening poisoning is possible without detection of the odor.

The insidious feature of phosgene poisoning is the substantial freedom from major initial discomfort; there is at first no clinical clue to the dose inhaled. Therefore, a so-called "phosgene indicator badge" was developed at the end of the 1970s [77]. This is a simple, visually readable, passive sampler based on a phosgene indicator paper that is impregnated with a solution of 4-*p*-nitrobenzylpyridine and *N*-benzylaniline. These phosgene indicator badges are now worn routinely in many plants by all employees. When the indicator paper becomes visibly colored, the persons affected should immediately consult a doctor. No specific antidote to phosgene is yet known. Hexamethylenetetramine has only a prophylactic effect, i.e., an effect if administered before phosgene inhalation [71]. Phosgene poisoning can be treated using corticosteroids and positive pressure respiration [71].

In the course of a case of phosgene poisoning, the pulmonary edema usually reaches its climax about 24 h after inhalation. The lethality depends on the dose inhaled. Fairly extensive statistics for fairly long periods of time usually give a lethality of 1 – 4 %. In general the pulmonary edema largely recedes within 3 – 5 d. Care must be taken to avoid pneumonia superinfection.

After surviving phosgene poisoning with pulmonary edema, the affected persons often complain for weeks (sometimes for months) of physical weakness, sweating, and an accelerated

heart beat (tachycardia) accompanying physical exertion. This is apparently a transient circulatory regulation disorder. Permanent reduction of the lung function or ensuing chronic bronchitis is extremely rare [78].

Repeated slight (subclinical) inhalation of phosgene can lead to an increased tolerance for a period of several days. This is explained by the edematous-inflammatory reaction of the blood – air barrier after the first exposure [79]. Thus the thickening of the blood – air barrier works like a protective dam in front of the capillaries; the diffusion path for phosgene becomes greater for later inhalations; more phosgene reacts with organic cell compounds or hydrolyses in the interstice and only a part of the inhaled phosgene is able to reach the capillary wall and exert its harmful effect there.

A chronic phosgene exposure with concentrations below the TLV (MAK) value of 0.1 ppm (8 h/d) is tolerated by healthy persons without harm [80]. The chronic inhalation of higher concentrations, on the other hand, can lead to pulmonary emphysema, without acute poisoning phenomena necessarily appearing in the meantime [81].

Mutagenicity studies with the *S.typhimurium* test system gave negative results [82].

11. References

1. L. F. Haber: *The Poisonous Cloud. Chemical Warfare in the First World War,* Clarendon Press, Oxford 1986.
2. *Beilstein,* **suppl. vol. I, part 4,** 31.
3. H. Spandau, V. Gutmann, *Angew. Chem.* **64** (1952) 93.
4. *Int. Crit. Tables,* **vol. III,** McGraw Hill, New York 1928, p. 237.
5. R. W. Gallant, *Hydrocarbon Process.* **47** (1968) no. 7, 141.
6. A. H. Nielsen, T. G. Burke, P. J. H. Woltz, G. A. Jones, *J. Chem. Phys.* **20** (1952) 596 – 604.
7. R. H. Pierson, A. N. Fletscher, E. St. Clair Gantz, *Anal. Chem.* **28** (1956) 1218.
8. V. Henri, O. R. Howell, *Proc. R. Soc. London A* **128** (1930) 190.
9. R. Anathakrishnan, *Proc. Indian Acad. Sci. Sect. A* **5** (1937) 285.
10. D. Chadwick, *Can. J. Chem.* **50** (1972) 737.
11. M. Bodenstein, H. Plaut, *Z. Phys. Chem. Stoechiom. Verwandtschaftsl.* **110** (1924) 402. H. I. Ingleson, *J. Chem. Soc.* **79** (1927) 2244.
12. J. S. Gordon, D. Goland, *J. Chem. Phys.* **27** (1957) 1223.
13. H. Babad, A. G. Zeiler, *Chem. Rev.* **73** (1973) 75 – 91.
14. K. Kurita, T. Matsumura, Y. Iwakura, *J. Org. Chem.* **41** (1976) no. 11, 2070.
15. H. C. Ramsperger, H. Waddington, *J. Am. Chem. Soc.* **55** (1933) 214.
16. G. Skorna, I. Ugi, *Angew. Chem.* **89** (1977) 267; *Angew. Chem. Int. Ed. Engl.* **16** (1977) 259.
17. K. Kurita, Y. Iwakura, *Org. Synth.* **59** (1979) 195.
18. H. Eckert, B. Forster, *Angew. Chem.* **99** (1987) 922.
19. R. G. Arnold, J. A. Nelson, J. J. Verbanc, *Chem. Rev.* **57** (1957) 47.
20. W. J. Schnabel, E. Kober, *J. Org. Chem.* **34** (1969) 1162.
21. M. Ohoka et al., *Bull. Chem. Soc. Jpn.* **46** (1973) 1275.
22. M. Matzner, R. P. Kurkjy, R. J. Cotter, *Chem. Rev.* **64** (1964) 645.
23. H. Keiser, E. Reichold, *Staub-Reinhalt. Luft* **48** (1988) 239 – 241.
24. A. M. Prentiss: *Chemicals in War,* Mc Graw Hill, New York 1947, pp. 262 – 294.
25. Drägerwerk Lübeck, "Phosgen-Prüfröhrchen 0.25 b, 38," *Mitteilung zum Gasspürgerät,* May 1960.
26. M. S. Gaisinovich, A. N. Ketov, *Russ. J. Inorg. Chem. (Engl. Transl.)* **14** (1969) 1218.
27. H. B. Singh, D. Lillian, A. Appelby, *Anal. Chem.* **47** (1975) 860 – 864.
28. R. Butler, A. Snelson, *J. Air Pollut. Control Assoc.* **29** (1979) no. 8, 833 – 837.
29. K. Kirchner, *Chimia* **37** (1983) no. 1, 1 – 9.
30. W. H. Manogue, R. L. Pigford, *AIChE J.* **6** (1960) 494 – 500.
31. C. Jäckh, W. Sundermeyer, *Chem. Ber.* **106** (1973) 1752.
32. C. Potter, S. Baron, *Chem. Eng. Prog.* **47** (1951) 473.
33. E. N. Shapatina, V. L. Kuchaev, B. E. Penkovoi, M. I. Temkin, *Kinet. Katal.* **17** (1976) 652 – 664.
34. Bayer, EP 0 134 506, 1984 (H. Sauer, H. Porkert, D. Liebsch).
35. A. Farwer, *Technica* **32** (1983) no. 7, 536 – 537.
36. R. Frant, *Weld. World* **12** (1974) nos. 11 – 12, 276 – 279.
37. R. Frant, *Schweißtechnik (Zürich)* **65** (1975) no. 4, 112 – 115.

38. T. Yabuta, *Yosetsu Gakkaishi* **39** (1970) no. 8, 719–730.
39. L. J. Priestley, Jr., F. G. Critchfield, N. H. Ketcham, J. D. Cavender, *Anal. Chem.* **37** (1965) 70–71.
40. H. Volkmann: *Handbuch der IR-Spektroskopie,* Verlag Chemie, Weinheim 1972.
41. C. F. Rush, C. E. Danner, *Anal. Chem.* **20** (1948) 644–646.
42. *Houben-Weyl,* 4th. ed., **2,** 542–543.
43. Bayer, DE 961 681, 1953 (H. Wollthan, W. Groves).
44. K. Hess, R. Kotkamp, *Chem. Ing. Tech.* **45** (1973) 873.
45. Auergesellschaft, Berlin; Drägerwerk, Lübeck, FRG.
46. Hartmann & Braun, "URAS", Frankfurt/Main. Perkin-Elmer, "SPECTRAN", Bodenseewerk, Konstanz, FRG.
47. Rosemound, "BINOS", Hanau, FRG.
48. B. E. Dixon, G. C. Hands, *Analyst (London)* **84** (1959) 463.
49. Bayer Diagnostic, Optotox, München, FRG.
50. MDA Inc., Lincolnshire, Ill., USA.
51. GMD Systems, Inc., Hendersonville, Pa., USA.
52. NIOSH: *Manual of Analytical Methods,* 2nd ed., **part 1** (Meth. No.: P + CAM 219).
53. Auergesellschaft, Billion-Air, Berlin, FRG.
54. Perkin-Elmer, Bodenseewerk, "IKG", Konstanz, FRG.
55. Bayer Diagnostic, Statox, München, FRG.
56. Miles Inc., Pittsburgh, Pa., USA.
57. F. J. Weigert, *J. Org. Chem.* **38** (1973) 1316–1319.
58. VEB Schwarzheide, DD 123 807, 1975 (K. Schwetlick et al.).
59. Arco, DE-OS 2 343 826, 1974 (J. G. Zajacek, J. J. McCoy, K. E. Fuger).
60. Mitsui-Toatsu, DE-OS 2 635 490, 1977 (R. Tsumura, U. Takaki, A. Takeshi).
61. K. Schwetlick, K. Unverferth, H. Tietz, *SYSpur Rep.* **19** (1981) 3–12; *Chem. Abstr.* **97** (1982) 39416.
62. Arco, US 4 146 727, 1979 (E. T. Shaw, J. G. Zajanek).
63. Dow Chem., US 4 268 683, 1981 (A. E. Gurgiolo).
64. S. Fukuoka, M. Chono, M. Kohno, *J. Org. Chem.* **49** (1984) 1458–1460.
65. K. A. Henderson, Jr., V. A. Alexanian, *Org. Prep. Proced. Int.* **18** (1986) 149–155.
66. Akzo, AU 523 314, 1982 (P. Hentschel, H. Zengel, M. Bergfeld).
67. BASF, DE 878 647, 1951 (W. Braun).
68. C. F. Hauser, L. P. Theiling, *J. Org. Chem.* **39** (1974) 1134.
69. FIAT, no. 774. Bayer, DE 943 403, 1942 (E. Podschus).
70. R. W. Gerard in E. C. Andrus (ed.): *Science in World War II,* vol. II, Little Brown Comp., Boston 1948.
71. W. F. Diller, *Toxicol. Ind. Health* **1** (1985) 73–133.
72. W. F. Diller, *Pneumologie* **150** (1974) 139–148.
73. W. F. Diller: "Radiologische Untersuchungen zur verbesserten Frühdiagnose von industriellen Inhalationsvergiftungen mit verzögertem Wirkungseintritt," *Zentralbl. Arbeitsmed. Arbeitsschutz Prophyl.,* **vol. 2,** Verlag für Medizin Dr. E. Fischer, Heidelberg 1976.
74. Amer. Ind. Hyg. Assoc., *Hygienic Guide Series: Phosgen,* Acron, Oh. 1968.
75. PPG Industries Inc., Bulletin no. 600, Chicago 1981.
76. W. F. Diller, R. Zante, *Zentralbl. Arbeitsmed. Arbeitsschutz Prophyl. Ergon.* **32** (1982) 60–368.
77. W. F. Diller, E. Drope, E. Reichold, *Ber. Int. Kolloq. Verhütung Arbeitsunfällen Berufskrankh. Chem. Ind.* **6th** (1979) 137–150; *Chem. Abstr.* **92** (1980) 168366x.
78. W. F. Diller, F. Schnellbächer, F. Wüstefeld, *Zentralbl. Arbeitsmed. Arbeitsschutz Prophyl.* **29** (1979) 5–16.
79. D. Henschler, W. Laux, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmokol.* **239** (1960) 433–441.
80. D. Henschler (ed.): *Gesundheitsschädigende Arbeitsstoffe, Toxikol. und arbeitsmedizinische Begründung der MAK-Werte,* Verlag Chemie, Weinheim 1971.
81. R. G. Rossing, *Am. J. Physiol.* **207** (1964) 265–271.
82. D. Reichert, T. Neudecker, U. Spengler, D. Henschler, *Mutat. Res.* **117** (1983) 21–29.

Phosphanes → Phosphorus Compounds, Inorganic

Phosphate Fertilizers

Gunnar Kongshaug, Norsk Hydro Research Centre, Porsgrunn, Norway (Chaps. 1 – 4 and 13 – 15)

Bernard A. Brentnall, British Sulphur, London, United Kingdom (Chap. 5)

Keith Chaney, Levington Agriculture, Levington, United Kingdom (Chap. 6)

Jan-Helge Gregersen, Norsk Hydro Research Centre, Porsgrunn, Norway (Chap. 7)

Per Stokka, Norsk Hydro Research Centre, Porsgrunn, Norway (Chap. 7)

Bjørn Persson, Hydro Supra, Landskrona, Sweden (Chap. 8)

Nick W. Kolmeijer, Hydro Agri Rotterdam, Vlaardingen, The Netherlands (Chaps. 9 and 10)

Arne Conradsen, Hydro Landbruk, Porsgrunn, Norway (Chap. 11)

Torbjørn Legard, Norsk Hydro Research Centre, Porsgrunn, Norway (Sections 12.1 and 12.2)

Harald Munk, Landwirtschaftliche Versuchsanstalt, Kamperhof Mülheim Ruhr, Federal Republic of Germany (Sections 12.3 and 12.4)

Øyvind Skauli, Norsk Hydro Research Centre, Porsgrunn, Norway (Chap. 16)

1. History 2
2. Terminology 3
3. Composition of Phosphate Fertilizers 4
4. Phosphate Rock 6
5. Economic Aspects 6
5.1. Phosphate Rock 6
5.2. Phosphate Fertilizer Consumption . 8
5.3. Phosphate Fertilizer Production . . 8
6. Phosphorus Uptake by Plants 9
7. Chemical Equilibria in Fertilizer Production 12
8. Superphosphates 16
8.1. Single Superphosphate 16
8.1.1. Chemistry 16
8.1.2. Production 17
8.1.3. Fluorine Recovery 19
8.1.4. Granulation 20
8.2. Triple Superphosphate 20
8.3. Double Superphosphate 20
8.4. PK Fertilizers 20
9. Ammonium Phosphates 20
9.1. Fertilizer Grades and Applications 21
9.2. Production 21
9.2.1. Ammonium Phosphate Powder 21
9.2.2. Granular Ammonium Phosphates . . . 23
9.2.3. Off-Gas Treatment 25
10. Compound Fertilizers by the Sulfur Route 25
10.1. Granulation of Mixtures of Dry Materials 26
10.2. Granulation of Dry Materials with Additives Producing Chemical Reactions 27
10.3. Slurry Granulation 27
10.4. Melt Granulation 29
11. Compound Fertilizers by the Nitro Route 30
11.1. Chemistry 31
11.2. Product Specification 31
11.3. Nitrophosphate Process with Calcium Nitrate Crystallization (Hydro) 31
11.4. Nitro Process with Calcium Nitrate Crystallization (BASF) 34
11.5. Nitro Process with Ion Exchange (Kemira Superfos) 36
11.6. Nitro Process with Sulfate Recycle (DSM) 36
11.7. Effluent Control of Nitrophosphate Process 37
12. Other Straight Phosphate Fertilizers 37
12.1. Phosphate Rock for Direct Application 37
12.2. Partially Acidulated Phosphate Rock 38
12.3. Basic and BOF Slag Fertilizers . . . 39
12.4. PK Mixed Fertilizers with Basic Slag 41
13. Energy Consumption 41
14. Effluents from Phosphate Fertilizer Production 41
15. Heavy Metals in Phosphate Fertilizers 41
16. Safety in Storage, Handling, and Transport 42
17. References 43

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_421

The article was coordinated by GUNNAR KONGSHAUG.

Abbreviations used in this article:

AN ammonium nitrate, NH_4NO_3
BOF basic oxygen furnace (slag)
BPL bone phosphate of lime
DAP diammonium phosphate, $(NH_4)_2HPO_4$
DCP dicalcium phosphate, $CaHPO_4$
MAP monoammonium phosphate, $NH_4H_2PO_4$
MCP monocalcium phosphate, $Ca(H_2PO_4)_2$
PAPR partially acidulated phosphate rock
SSD self-sustaining decomposition
SSP single superphosphate
TPL total phosphate of lime
TSP triple superphosphate (known as concentrated superphosphate in North America)

1. History [1]

Farmers have always been anxious to improve crop yields. Some thousand years ago, Chinese farmers used calcined bones and the Incas in Peru used phosphoguano to increase crop output. In Europe, bones have been applied for centuries to French vineyards.

Several seventeenth century publications in Europe mention the beneficial effect of bones as a fertilizer for plant growth.

The German alchemist HENNING BRANDT discovered phosphorus in 1669 by isolating it from urine. In 1769 the Swedish scientist J. G. GAHN discovered that calcium phosphate is the main component of bones. About 30 years later, the conclusion was reached that the fertilizing effect of bones is due mainly to calcium phosphate and not to organic material. In 1797 the British physician GEORGE PEARSON gave the name superphosphate to the phosphate compound (calcium dihydrogenphosphate) found in bone; this name was later applied to fertilizers. Field trials demonstrated that bones should be crushed and applied in very small pieces.

Merchants then moved into the fertilizer business and established local powder mills for bone grinding. Attempts were made to improve fertilizer efficiency by composting bones in earth, animal waste, or plant waste; by boiling bones in water; or by treating them with steam under pressure.

Increased understanding of the fertilizer effect of phosphorus and a rapid increase in the use of bones in the early nineteenth century led to the idea of using chemical treatment of bones to improve fertilizer efficiency. Developments occurred in many countries. HEINRICH W. KÖHLER of Bohemia was probably the first to suggest and file a patent for using acids (especially sulfuric acid) in the processing and commercial production of phosphate fertilizers (1831).

In 1840 JUSTUS VON LIEBIG's theory on phosphorus uptake in plants contributed greatly to acceptance of the product and to rapid worldwide growth of the phosphate fertilizer industry. In the early 1840s the lack of bones as a raw material led to the export of phosphoguano from Peru. The discovery of low-grade mineral phosphates in France and England eased the raw material situation, but the development of the phosphate industry was secured by the discovery of large sedimentary phosphate deposits in South Carolina. The deposits were rediscovered as phosphate rock in 1859; mining began in 1867, and in 1889 the mine supplied 90 % of the worldwide phosphate fertilizer production.

The production of ammonium phosphate fertilizers by ammoniation of phosphoric acid began around 1917 in the United States. The Haber – Bosch process boosted this product line, and in 1926 the IG Farbenindustrie in Germany announced the development of a series of multinutrient (compound) fertilizers based on crystalline ammonium phosphate.

Separation of calcium sulfate from the superphosphate slurry by increasing the sulfuric acid/rock ratio and use of phosphoric acid for acidulation led to the development of concentrated (triple) superphosphates and commercialization in ca. 1890.

The treatment of phosphate rock with nitric acid (nitrophosphate process) was developed in the late 1920s by the Norwegian ERLING B. JOHNSON. The IG Farbenindustrie, DSM in The Netherlands, and Norsk Hydro in Norway commercialized this route for complex fertilizer production in the 1930s.

In Europe, calcium silicophosphate fertilizer is produced as a byproduct of the steel industry. Iron ore may contain phosphorus, which can be removed by slagging out with lime. The product is sold under the name Thomas phosphate or basic slag. Many attempts have been made

to produce similar fertilizer products by thermal treatment of phosphate rock with additives but most were unsuccessful due to high energy costs. Small amounts of fused magnesium phosphate and calcined defluorinated phosphate are produced in Brazil, China, Korea, and Japan [2].

2. Terminology

Phosphorus Content. The phosphorus-containing component of phosphate rock is apatite. *Fluorapatite* [*1306-05-4*] is the most common phosphate rock mineral. The correct formula of fluorapatite is $Ca_{10}F_2(PO_4)_6$, but it can be simplified to $Ca_5F(PO_4)_3$.

In commercial trading of phosphate rock, the phosphorus content is calculated as the weight percentage of tricalcium phosphate, $Ca_3(PO_4)_2$, and expressed as the bone phosphate of lime (BPL) or the total phosphate of lime (TPL). Tricalcium phosphate is not present as such in phosphate rock, but to simplify the relationship between BPL and fluorapatite, the formula of fluorapatite is sometimes expressed as $3\,Ca_3(PO_4)_2 \cdot CaF_2$.

The phosphorus in fertilizers is supplied as orthophosphate (referred to here as phosphate), PO_4^{3-}, but the content is generally expressed as the weight percentage of phosphorus pentoxide (P_2O_5) or, incorrectly, as phosphoric acid. Similarly, the contents of potassium, magnesium, and calcium are given as the weight percentage of their oxides: potassium oxide (K_2O), magnesium oxide (MgO), and calcium oxide (CaO). In the Scandinavian countries (except Iceland) and Ireland, however, the nutrient contents are expressed as elements (P, K, Mg, Ca).

$$\begin{aligned}
BPL &= 2.185 \times P_2O_5 = 5.008 \times P \\
P_2O_5 &= 0.458 \times BPL = 2.291 \times P = 0.724 \times H_3PO_4 \\
P &= 0.200 \times BPL = 0.436 \times P_2O_5 \\
&= 0.316 \times H_3PO_4
\end{aligned}$$

PO_4^{3-} = (ortho)phosphate ion
HPO_4^{2-} = hydrogen(ortho)phosphate ion
$H_2PO_4^{-}$ = dihydrogen(ortho)phosphate ion
H_3PO_4 = phosphoric acid

Types of Fertilizer. The production routes for phosphate fertilizers are summarized in Figure 1.

Superphosphate fertilizers (Chap. 8) are produced by treating phosphate rock with acid to give calcium dihydrogenphosphate.

The product obtained with sulfuric acid is called *single* (*or normal*) *superphosphate* (SSP). It is produced as superphosphate powder (run-of-pile), which later can be granulated to form 2 – 5 mm particles. Single superphosphate contains mainly calcium dihydrogenphosphate and calcium sulfate.

The most common product is, obtained by treatment of phosphate rock with phosphoric acid, is called *triple superphosphate* (TSP) (or concentrated superphosphate in North America). Triple superphosphate is produced either by use of run-of-pile powder as an intermediate or by a direct slurry granulation process.

Of minor importance is *double superphosphate*, which is a mixture of single and triple superphosphates. The term double superphosphate may cause some confusion because it was used earlier as a name for triple superphosphate.

Partly acidulated phosphate rock (PAPR, see Section 12.2) is produced in the same way as superphosphates, but with less sulfuric acid, to obtain ca. 50 % water-soluble P_2O_5. Annual production is negligible compared to other superphosphates.

Ammonium phosphate fertilizers (Chap. 9) contain ammonium dihydrogenphosphate [*7722-76-1*], $NH_4H_2PO_4$ (referred to in the fertilizer trade as monoammonium phosphate, MAP), and diammonium hydrogenphosphate [*7783-28-0*], $(NH_4)_2HPO_4$ (referred to as diammonium phosphate, DAP).

Compound (multinutrient) NP or NPK fertilizers are made by acidulating rock with sulfuric acid (Chap. 10) or nitric acid (Chap. 11).

Ground phosphate rock (Section 12.1) may also be used directly as a fertilizer. Basic slag (Section 12.2) is a fertilizer byproduct of the steel industry.

3. Composition of Phosphate Fertilizers

The major chemical components of the most common phosphate fertilizers are given in Table 1.

In phosphate rock, the F^- in fluorapatite may be replaced by OH^- and Cl^-; PO_4^{3-} by

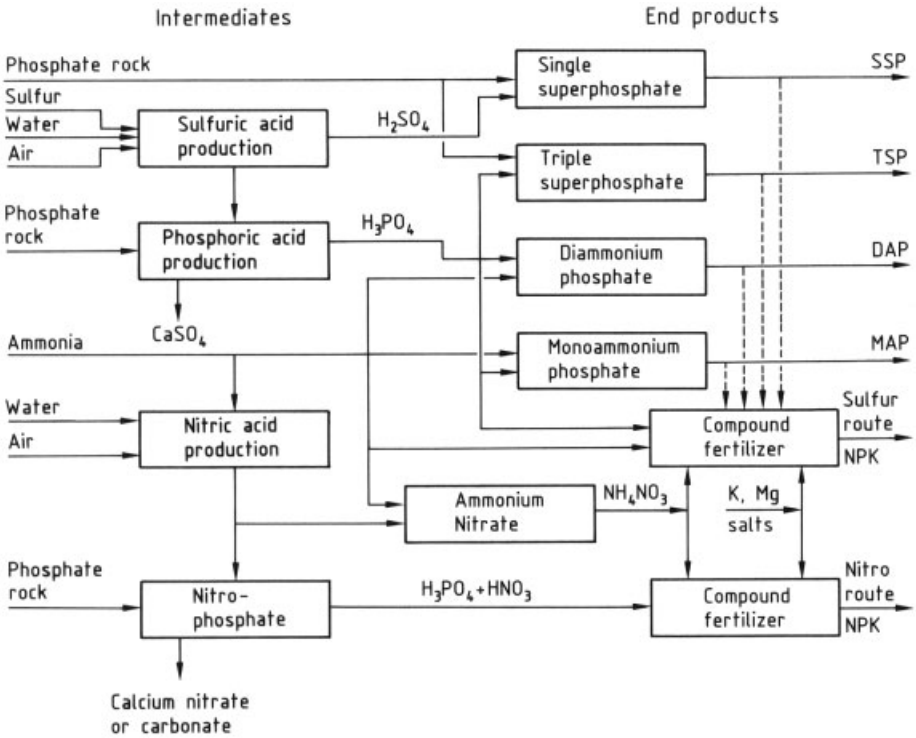

Figure 1. Primary production routes for phosphate fertilizers

CO_3^{2-}, SO_4^{2-}, CrO_4^{2-}, and SiO_4^{4-}; and Ca^{2+} by Na^+, K^+, Mg^{2+} and heavy metals. Possible metal compounds in phosphate fertilizers have been reported [3][4][5][6][7]. Their formation depends on process conditions and concentrations.

Dissolved iron and aluminum in superphosphate precipitate slowly as complex salts. High concentrations of aluminum and iron form amorphous aluminum and iron phosphate in superphosphates, which revert to crystalline calcium metal phosphate during production and storage. The most common components are $CaFe_2(HPO_4)_4 \cdot nH_2O$ and $CaAl_2(HPO_4) \cdot nH_2O$; $(Fe,Al)CaH(PO_4)_2 \cdot nH_2O$ may also be present. The same calcium metal phosphates are formed when phosphoric acid containing free calcium ions is ammoniated.

When phosphoric acid is ammoniated, the metals always seem to be present as metal ammonium phosphates in crystalline form, amorphous form, or as pyrophosphate gel.

Crystalline form:
$FeNH_4(HPO_4)_2$
$AlNH_4(HPO_4)_2$
Amorphous form:
$FeNH_4(HPO_4)_2 \cdot nH_2O$
$AlNH_4(HPO_4)_2 \cdot nH_2O$
$Mg(NH_4)_2(HPO_4)_2 \cdot nH_2O$
$AlNH_4F_2(HPO_4) \cdot nH_2O$
$FeNH_4F_2(HPO_4) \cdot nH_2O$
Pyrophosphate gel:
$(Mg, Al, Fe)NH_4FHP_2O_7 \cdot nH_2O$

Phosphate solubilities are measured in different ways to indicate plant availability (→ Fertilizers, Chap. 6.3.):

1) Extraction with water
2) Extraction with citrate solution (neutral ammonium citrate solution)
3) Extraction with 2 % citric acid
4) Extraction with 2 % formic acid
5) Determination of total phosphate content

Calcium dihydrogenphosphate $Ca(H_2PO_4)_2$ (monocalcium phosphate, MCP), MAP, and DAP are water soluble; calcium hydro-

Table 1. Main chemical components of commercial fertilizers

Chemical composition	Type of fertilizer								
	Single superphosphate (SSP)[a]	Triple superphosphate (TSP)[a]	MAP	DAP	S-route NPK[b]	N-route NPK[c]	PAPR[d]	Basic slag	Phosphate rock
P_2O_5[e], wt%	21	48	53	46	15	12	21	14	29
P_2O_5, % (min.-max.)	18–22	46–50	48–61	46–53	5–30	5–23	18–24	13–15	27–33
$NH_4H_2PO_4$, MAP, P:P[f] (total)	0	0	0.80–0.89	0.01	0–0.89	0.74	0	0	0
$(NH_4)_2HPO_4$, DAP, P:P (total)	0	0	0.05	0.84–0.93	0–0.93	0.01	0	0	0
$Ca(H_2PO_4)$, MCP, P:P (total)	0.81–0.91	0.81–0.91	0	0	0–0.71	0	0.45	0	0
$CaHPO_4$, DCP, P:P (total)	0	0	0.01	0.01	0–0.2	0.2	0	0	0
Fluorapatite, P:P (total)	0.04	0.04	0	0	0–0.04	0	0.5	0	1.0
Calciumsilicophosphate, P:P (total)	0	0	0	0	0	0	0	1.0	0
Metal phosphates (mainly Al, Fe, Mg), P:P (total)	0.05–0.15	0.05–0.15	0.05–0.2	0.05–0.2	0.05–0.2	0.05	0.05	0	0
CaO, wt%	32	22	0.7	0.7	0.7–15	1.5–7	34	>35	47
$CaO:P_2O_5$, weight ratio	1.52	0.46	0.01	0.01	0.01–0.5	0.3	1.6	>2.5	1.6
N, wt%	0	0	10–12	16–21	0–30	12–30	0	0	0
$NO_3:NH_4$, molar ratio			0	0	0–0.9	0.7–0.9			
K_2O, wt%	0–30	0–30	0	0	0–21	0–21	0	0–30	0

[a] Range for P_2O_5 given without K addition. [b] Compound NP and NPK fertilizers based on the sulfuric acid route; the P component may be based on MAP, DAP, SSP, TSP, or a mixture; [c] Compound NP and NPK fertilizers based on the nitrophosphate route; specific values given for 75% water solubility P_2O_5. [d] Partially acidulated phosphate rock. [e] Estimated average P_2O_5 content in given type of fertilizer. [f] P:P (total) denotes the ratio of the phosphorus content in the specified chemical component to the total phosphorus content of the fertilizer.

genphosphate, $CaHPO_4$ (dicalcium phosphate, DCP), is citrate soluble (not water soluble).

Calcium metal phosphates as crystalline precipitates are citrate soluble.

Metal ammonium phosphates as amorphous precipitates or pyrophosphate gel are citrate soluble. Metal ammonium phosphates as crystalline precipitates are not usually citrate soluble.

Superphosphates normally contain >90 % water-soluble and >98 % citrate-soluble P_2O_5. Ammonium phosphates contain >85 % watersoluble and >99 % citrate-soluble P_2O_5. Compound NPK fertilizers contain >70 – 75 % water-soluble and >99 % citrate-soluble P_2O_5.

In partly acidulated rock fertilizers, the P_2O_5 solubility in water and citrate is ca. 45 – 50 % and the solubility in citrate solution somewhat higher.

Ground phosphate rock has no citrate-soluble P_2O_5 but solubility in 2 % citric acid may be 5 – 53 % and solubility in 2 % formic acid may be as high as 86 %. Basic slag (calcium silicophosphate) has no water-soluble P_2O_5, but citrate solubility is reported to be up to 90 % and the citric acid solubility up to 97 % [8].

4. Phosphate Rock

Phosphate rock is virtually the sole raw material for phosphate fertilizers. The primary source is sedimentary phosphate rock (phosphate precipitated from seawater and bones) but magmatic (igneous) phosphate rocks are also important.

Phosphate rock contains calcium phosphates as apatites, mainly fluorapatites. It is mined in almost 30 countries, and phosphate fertilizers are produced in 26 countries in Europe, 13 in America, 9 in Africa, 13 in Asia, and 2 on the Australian continent [9].

About 90 % of mined phosphorus is used in fertilizers, being divided almost equally among superphosphates, ammonium phos-

phates, and compound fertilizers (including nitrophosphates).

5. Economic Aspects

5.1. Phosphate Rock

Phosphate rock reserves of varying composition and quality are widely distributed. Exploitation of deposits occurs in many countries, but large-volume production for captive use and export is limited to a few countries. Total world rock production in 1989 was 161.4×10^6 t and was divided as follows (10^6 t):

United States	49.9
Former Soviet Union	35.0
People's Republic of China	18.5 *
Morocco	18.0
Jordan	6.7
Tunisia	6.6
Israel	3.9
Brazil	3.7
Togo	3.4
Republic of South Africa	3.0
Senegal	2.3
Syria	2.3

* Down from 24.8×10^6 t in 1988 due to market disruption

The four major producers of rock account for 75 % of world output. The output of former Soviet and Chinese phosphate rock mines was essentially used for domestic consumption.

Since the early 1970s, phosphate rock has become a less important export product in comparison with processed phosphate products (mainly phosphoric acid, ammonium phosphates, and triple superphosphate). World export of phosphate rock was 43.5×10^6 t in 1989 (down from a peak of 53.4×10^6 t in 1979) and was divided as follows (10^6 t):

Morocco	12.4
United States	8.3
Jordan	5.9
Togo	3.3
Former Soviet Union	3.2
Israel	2.4
Syria	1.7
Senegal	1.4
Nauru	1.2
Republic of South Africa	1.1
Tunisia	1.1

The four leading export countries accounted for 69 % of world trade.

The shift away from phosphate rock as the major vehicle for trade in phosphates results from the development of vertically integrated industries at or near mine sites. The key element in these operations is phosphoric acid capacity, which allows a concentrated, gypsum-free product to be transported to market.

The shift toward captive processing of phosphate rock is reflected in a significant change in the pattern of phosphate rock consumption and world trade. The total trade in rock continues to decrease and appears to have fallen below 40×10^6 t/a in 1990 for the first time since 1970. On the other hand, the tonnages delivered to downstream processing industries are now substantial. Total worldwide consumption of phosphate rock in 1989 was 158.9×10^6 t and was divided as follows (10^6 t):

Western Europe	17.0
Eastern Europe	42.2
Former Soviet Union	31.8
North America	42.4
Africa	14.4
Morocco	5.2 *
Republic of South Africa	2.1
Tunisia	5.1
Central America	2.4
Mexico	2.4
Australasia	2.5
Australia	1.7
Middle East	3.9
Iraq	1.1
Israel	1.4
Jordan	1.0
Asia	28.7
People's Republic of China	18.8
India	3.0
Indonesia	1.1
Japan	1.6
South Korea	1.5
South America	4.6
Brazil	4.1

* Down from 10.6×10^6 t in 1988 as a result of market disruption.

A comparison of phosphate rock consumption statistics with production and export statistics reflects the development of downstream processing facilities in many rock-exporting countries. Demand for phosphate rock has thus receded in a number of regions, notably Western Europe and Japan.

Phosphate rock statistics are generally expressed as tonnes of product. The phosphorus content of the rock varies. Rock of sufficiently high phosphorus content (grade) to be acceptable commercially contains 30–40 % P_2O_5 (66–87 BPL).

The majority (ca. 90 %) of phosphate rock produced is converted into fertilizer. Smaller, nonfertilizer end uses account for ca. 10 % of phosphate rock consumption; the most important are the manufacture of detergents and animal feed additives.

5.2. Phosphate Fertilizer Consumption

Worldwide consumption of phosphate fertilizers continues to grow, but during the 1980s growth was erratic and less dynamic than in the 1960s and 1970s. Consumption (in 10^6 t P_2O_5) was as follows (in 10^6 t):

1960	9.7
1965	14.0
1970	19.7
1980	31.6
1985	33.0
1986	34.8
1987	36.4
1988	37.7
1989	36.6

A number of factors will contribute to the future consumption of phosphate fertilizers:

1) The agricultural sector in the export-oriented countries of the developed world has suffered consistently from overproduction and mounting subsidy bills. In addition to poor farm economics, government programs have been aimed at reducing crop production.
2) In many developed countries, years of regular phosphate application have led to a buildup of phosphate soil reserves. As farm economics have deteriorated, farmers have been content to exploit the soil reserves by reducing phosphate application.
3) Environmental concern about the consequences of overfertilization and runoff of nutrients in groundwater has led to a more accurate application of nutrients in many countries.

In Western Europe, phosphate fertilizer consumption peaked in 1979 at 6.8×10^6 t of P_2O_5; pressures on agriculture from the Common Agricultural Policy (CAP), the General Agreement on Tariffs and Trade (GATT) negotiations, and environmental groups indicate that consumption levels are unlikely to recover and may stabilize close to 5×10^6 t of P_2O_5.

In Central Europe and the former Soviet Union, phosphate fertilizer consumption increased steadily during most of the 1980s in line with economic policies. A peak of 11.6×10^6 t of P_2O_5 was recorded in 1988. Changed economic circumstances may lead to reduction.

Consumption of phosphate fertilizer in North America peaked in 1980 at 5.6×10^6 t of P_2O_5. In the latter half of the 1980s, annual use fluctuated between 4.3 and 4.6×10^6 t. Decline in consumption is not anticipated. At 2.4×10^6 t P_2O_5 in 1989, phosphate fertilizer usage in South America was down some 0.4×10^6 t P_2O_5 over 1980. Demand in Brazil has proved the single most important variable, fluctuating by as much as 1×10^6 t/a in the 1980s. In Africa, annual demand fluctuated between 1.3×10^6 t (1981) and 0.9×10^6 t (1989) due to civil conflict, economic crises, and drought.

Asia had consistent growth in phosphate fertilizer consumption during the 1980s. This reflects the constantly increasing demand for food by the rapidly growing population. Chinese consumption almost doubled during the 1980s, reaching 5.3×10^6 t of P_2O_5 in 1989. In the Middle East, P_2O_5 consumption (including Australasia) in 1989 stood at 12.2×10^6 t, 33 % of the world total.

5.3. Phosphate Fertilizer Production

Changes in the phosphate industry toward vertical integration and the trading of processed P_2O_5 products rather than phosphate rock have resulted in significant shifts in phosphate fertilizer product patterns. The importance of low-grade phosphate fertilizers (i.e., products with a low P_2O_5 content) has declined significantly. These include single superphosphate (typically 18–22 % P_2O_5) and basic slag (typically 13–15 % P_2O_5). Not only does the low nutrient content make such products more expensive to transport, handle, and apply, but production economics have made some of them unattractive.

High-nutrient fertilizers now dominate the world phosphate fertilizer market and fall into three categories:

1) Ammonium phosphates
2) Triple (or concentrated) superphosphate
3) Binary and ternary compound (or complex) fertilizers

The ammonium phosphate category comprises two main fertilizer products: diammonium phosphate, DAP ($N-P_2O_5$ content: 18 – 46; i.e., 18 wt % N and 46 wt % P_2O_5) and monoammonium phosphate, MAP ($N-P_2O_5$ content: 11 – 53).

Triple (or concentrated) superphosphate is a straight, single-nutrient phosphate product (typical P_2O_5 content 48 wt %).

Compound (multinutrient) fertilizers are produced in various combinations (→ Fertilizers, Chap. 3.1.). The NPK fertilizers contain all three main fertilizer nutrients, i.e., nitrogen, phosphorus, and potassium (typical N-P_2O_5-K_2O content in weight percent: 16-16-16, NP fertilizers contain no potash (nutrient content typically 22-20-0), PK fertilizers are produced mainly in Western Europe (nutrient content typically 0-25-25).

In 1989, world phosphate fertilizer production amounted to 38.6×10^6 t of P_2O_5. Production breakdown follows (in 10^6 t):

Single superphosphate	7.0
Triple superphosphate	4.6
Ammonium phosphates	13.7
Compound fertilizers	10.3
Other low-analysis phosphates	3.0

Single superphosphate retains its importance in some developing countries and in Eastern Europe. In 1989 the People's Republic of China accounted for 2.5×10^6 t of P_2O_5 production, Eastern Europe for almost 2×10^6 t, and India for 0.5×10^6 t. As economic conditions in Central Europe change, single superphosphate will probably be displaced by more concentrated products.

Production of *triple superphosphate* (known as concentrated superphosphate in North America) is widely dispersed. More than 1×10^6 t of P_2O_5 was produced in Eastern Europe in 1989, but this is likely to decrease. Producers in the United States accounted for ca. 0.8×10^6 t of P_2O_5, and North African and Middle Eastern industries for an additional 1.1×10^6 t. Not all triple superphosphate is applied directly to soil. It is often blended at the distributor or even the farm level with other nutrients and applied as compound fertilizer. The same is true of single superphosphate, although to a lesser degree.

Triple superphosphate is the least important export product. In 1989, world trade amounted to 1.5×10^6 t of P_2O_5 (3.3×10^6 t of product). Approximately 0.9×10^6 t of these exports originated in North Africa and the Middle East, and 0.25×10^6 t in the United States.

Ammonium phosphates were developed in the United States primarily as components of blended fertilizers (bulk blends): DAP is the most important product, world production amounting to 9.3×10^6 t of P_2O_5 in 1989, 5.8×10^6 t of this being produced in North America. In 1989, U.S. exporters shipped ca. 3.6×10^6 t. Morocco is now the second most important exporter (0.5×10^6 t P_2O_5), and output is growing. The production of MAP is concentrated primarily in the former Soviet Union and North America. In 1989, world output was 4.4×10^6 t of P_2O_5, of which 2.5×10^6 t was manufactured in the former Soviet Union and 1.3×10^6 t in North America. These products have become so attractive in terms of cost effectiveness on the world market that an increasing proportion of ammonium phosphate is being used as a direct application fertilizer, particularly in less-developed countries.

Compound fertilizers are a development of the European fertilizer industry. They may be produced by treating phosphate rock with sulfuric acid (the sulfur, wet process phosphoric acid, or ammonium phosphate route) or with nitric acid (the nitrophosphate route). In addition, fertilizer components such as ammonium phosphate may be mixed physically to produce blended compound fertilizers. Blending often occurs at the wholesale or retail level; the total volume of blended compounds is therefore not reflected in consumption (delivery) statistics.

In 1989, the chemical production of compound fertilizers exceeded 2×10^6 t of P_2O_5 in both Eastern and Western Europe and 1×10^6 t in North America. However, considerable pressure has been exerted on European industries from imported DAP and triple superphosphate (which will lead to continuing rationalization of this sector).

Although a significant amount of regional trade in P_2O_5 fertilizers such as NPKs occurs within Western Europe, between the former Soviet Union and Central Europe, interregional trade in phosphate fertilizers is now dominated by vertically integrated producers. The United States dominates the international market for DAP (see above).

6. Phosphorus Uptake by Plants

[10–12]

Role in Plant Nutrition. Phosphorus is essential for vital growth processes in plants because it is a constituent of nucleic acids. It is also a constituent of phospholipids. Phosphorus compounds (coenzymes) are involved in respiration, energy transfer, and the efficient utilization of nitrogen. See also → Fertilizers, Chap. 2.1.2.

Phosphorus is of special importance in root development and in the ripening of seeds and fruit. The application of phosphate fertilizer to soils that are low or deficient in available phosphate improves root development and seedling growth, giving the crops a better start.

Phosphorus is taken up by plants as the dihydrogenphosphate ($H_2PO_4^-$) or the hydrogenphosphate (HPO_4^{2-}) ion.

Phosphorus Deficiency Symptoms. In phosphorus deficiency the growth of plant tops and roots is greatly restricted: shoots are short and thin, growth is upright and spindly; leaves are small, and defoliation is premature; lateral shoots are few in number, and lateral buds may die or remain dormant; blossoming is greatly reduced, resulting in poor yields of grain and fruit. These symptoms also apply to nitrogen deficiency, but with phosphate deficiency the leaf color is generally a dull, bluish green, usually with a purple tint (rather than yellow or red). Leaf margins may also show brown scorching.

Forms of Phosphate in the Soil. Most phosphate in the soil is present in the solid phase; only a small amount is dissolved in the soil solution. Solid-phase phosphate occurs partly in organic form (30 – 50 %), but in mineral soil it exists mainly in inorganic form (50 – 70 %). The three major components of *organic soil phosphate* are nucleic acids, phospholipids, and phytin with its derivatives. Organic phosphate is continuously released as HPO_4^{2-} ions by microbial degradation of organic soil matter. However, this phosphate release does not always coincide with maximum plant uptake.

Most *inorganic soil phosphates* fall into two groups: those containing calcium, and those containing iron and aluminum. The calcium compounds of greatest importance are apatites with the general formula $Ca_5(F, Cl, OH, 1/2\,CO_3)(PO_4)_3$; calcium hydrogenphosphate, $CaHPO_4$ (dicalciumphosphate, DCP); and calcium dihydrogenphosphate, $Ca(H_2PO_4)_2$ (monocalciumphosphate, MCP).

Fluorapatite is the most insoluble of these compounds and is therefore unavailable to plants. In contrast, MCP and DCP are readily available for plant growth. Except in recently fertilized soil, these two compounds are present in very small quantities because they readily react to form more insoluble compounds.

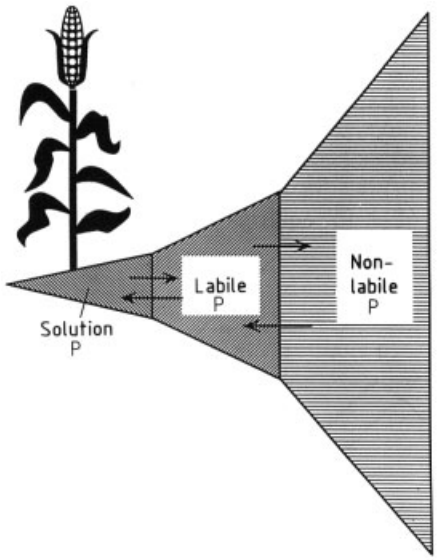

Figure 2. Main soil phosphate fractions

The iron and aluminum phosphates in soil are probably hydroxyphosphates. They are very stable in acid soil and are extremely insoluble.

Three inorganic soil phosphate fractions are important in terms of plant nutrition (Fig. 2):

1) Phosphate in soil solution: dissolved phosphates available for plant uptake

2) Phosphate in the labile pool: often absorbed on mineral soil particles and can rapidly go into soil solution
3) Phosphate in the nonlabile fraction: only very slowly released into the labile pool

Phosphate dissolved in soil solution can be taken up by plants. Phosphate removed from this pool is replenished by phosphate from the labile pool. Nonlabile (insoluble) phosphate is released very slowly into the labile pool–too slowly to make a significant contribution to the crop in one growing season.

Effect of Soil on Phosphate Availability. The pH determines the phosphate form in soil solution. In very acid solution, only $H_2PO_4^-$ ions are present. As pH increases the HPO_4^{2-} ions and then the PO_4^{3-} ions dominate. At pH 7.0 the proportions of $H_2PO_4^-$ and HPO_4^{2-} ions are roughly equal:

$$\underset{\text{Acid solution}}{H_2PO_4^-} \underset{}{\overset{H^+}{\rightleftharpoons}} HPO_4^{2-} \underset{}{\overset{H^+}{\rightleftharpoons}} \underset{\text{Very alkaline solution}}{PO_4^{3-}}$$

Although the $H_2PO_4^-$ ion is generally considered to be more available than the HPO_4^{2-} ion, both can be taken up by plants. However, in soil this relationship is complicated by other compounds and ions.

In acid soil, soluble iron, aluminum, and manganese react with $H_2PO_4^-$ ions to form insoluble phosphates that are unavailable for plant growth:

$$\underset{\text{Soluble}}{Al^{3+} + H_2PO_4^-} + 2\,H_2O \longrightarrow 2\,H^+ + \underset{\text{Insoluble}}{Al(OH)_2H_2PO_4}$$

In most acid soils the concentration of iron and aluminum ions exceeds that of $H_2PO_4^-$ ions, and only minute quantities of phosphate remain immediately available to plants. Phosphate can also be fixed by hydrous oxides, kaolinite, montmorillonite, and illite. Alkaline soil containing an excess of exchangeable calcium (e.g., calcium carbonate) also precipitates phosphates:

$$\underset{\text{Soluble}}{Ca(H_2PO_4)_2} + 2\,CaCO_3 \longrightarrow \underset{\text{Insoluble}}{Ca_3(PO_4)_2} + 2\,CO_2 + 2\,H_2O$$

Because insoluble phosphates are formed in both acid and alkaline soils, maximum phosphate availability is obtained at a soil pH of 6.0–7.0. Even in this range, phosphate availability may be low, and added soluble phosphates can be readily converted into unavailable forms.

Use of Phosphate as a Fertilizer. Phosphate ions are relatively immobile in soil compared to nitrate, and are taken up by plants from the soil solution at the root surface. The critical time for phosphate supply to the plant is in the seedling stage, when seed reserves of phosphate have been exhausted and the root system has not developed sufficiently to supply phosphate needs.

Two strategies are employed with phosphate fertilizers:

1) Sufficiently high phosphate levels are built up in soil to meet the crop requirement for phosphate ions. Crops with shallow root systems (e.g., potatoes) require a higher concentration of phosphate in soil solution to sustain growth than crops with extensive root systems (e.g., cereal).
2) A small reservoir of phosphate ions is provided near the site at which seedling roots develop (a hot spot of phosphate-enriched soil). Enriched hot spots can be created only if the phosphate fertilizer dissolves rapidly in the soil. For most soil the fertilizer must be water- or citrate-soluble.

In some situations the association of roots with microorganisms (endotropic mycorrhizae) increases phosphate uptake by a number of crops. However, at present this seems to be of only agronomic importance in soil with a very low phosphate content. Similarly, root exudates can influence chemical and microbial activity in the root zone and can increase phosphate uptake where soil phosphate supply is marginal.

High-yield crops used in modern agriculture demand high levels of soluble phosphate in the soil. This can be achieved only by adding materials (fertilizers) containing soluble phosphates. The natural soil phosphate level is usually not sufficient.

The quantity of soluble phosphate fertilizer added for a particular crop depends on the phosphate removed in the harvested yield, the response of the crop to the fertilizer, and the available phosphate level of the soil in which the crop is grown. The amounts of phosphate removed by

crops are as follows (kilogram of P_2O_5 per tonne of fresh material) [13]:

Sugar beets, swede roots, cabbage, carrots, onions	0.7 – 0.9
Potato, kale, maize, French beans, beetroot, cauliflower, silage grass, cereal straw, vining peas, broad beans	1.0 – 1.6
Brussel sprouts	2.1 – 2.6
Grass, hay, cereal grain, dried peas	5.9 – 8.8
Field beans	11
Oilseed rapeseed	16

7. Chemical Equilibria in Fertilizer Production

In acidic soil, fluorapatite in porous phosphate rocks may slowly be broken down to form water-soluble phosphate ions. In fertilization, a strong acid is used to break down the apatite to hydrogenphosphates.

The system of phosphate fertilizers can be understood by means of acid – base and solid –liquid chemical equilibria. The primary acid –base equilibria involve phosphoric acid, sulfuric acid, nitric acid, and ammonia. The main solid – liquid equilibria involve ammonium phosphates, calcium phosphates, and water.

Fertilizer Acids. Most important is the phosphoric acid equilibrium [14]:

$$H_3PO_4 \rightleftharpoons H^+ + H_2PO_4^- \quad K = 10^{-2.13}$$

$$H_2PO_4^- \rightleftharpoons H^+ + HPO_4^{2-} \quad K = 10^{-7.2}$$

$$HPO_4^{2-} \rightleftharpoons H^+ + PO_4^{3-} \quad K = 10^{-12.4}$$

These equilibria can be represented in different ways: as a function of mole percent and pH (→Phosphoric Acid and Phosphates, [15] or as a function of logarithmic concentration and pH (also called a Bjerrum diagram, Fig. 3) [16–21].

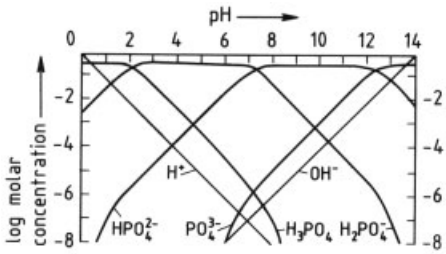

Figure 3. Phosphoric acid equilibria in solution [17]

Sulfuric acid and nitric acid are also important acids in phosphate fertilizer production. The equilibrium reactions are [14]

$$H_2SO_4 \rightleftharpoons H^+ + HSO_4^- \quad K = >1$$

$$HSO_4^- \rightleftharpoons H^+ + SO_4^{2-} \quad K = 10^{-1.99}$$

$$HNO_3 \rightleftharpoons H^+ + NO_3^- \quad K = 10^{1.37}$$

These equilibria can also be illustrated in a logarithmic diagram.

Ammonium Phosphates. Ammoniation of phosphoric acid involves the following reactions [14]:

$$NH_4^+ \rightleftharpoons H^+ + NH_3 \quad K = 10^{-9.24}$$

$$NH_4^+ + H_2PO_4^- \rightleftharpoons NH_4H_2PO_4 \text{ (MAP)}$$

$$2\,NH_4^+ + HPO_4^{2-} \rightleftharpoons (NH_4)_2HPO_4 \text{ (DAP)}$$

These equilibria can be represented in a logarithmic diagram by combination of the ammonia and phosphoric acid constants (Fig. 4).

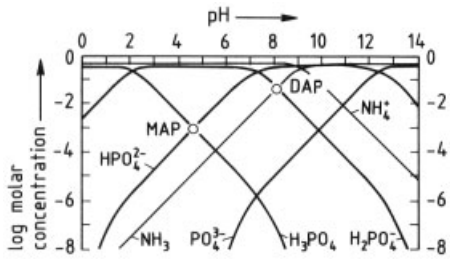

Figure 4. Ammonium phosphate equilibria in solution

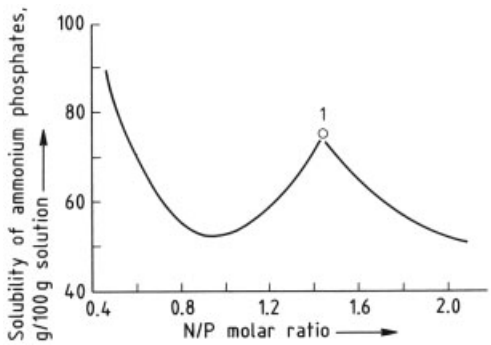

Figure 5. Effect of NH_3/H_3PO_4 molar ratio on the solubility of ammonium phosphates at 75 °C [23]

The maximum concentrations of NH_4^+ and $H_2PO_4^-$ ions necessary for MAP production are present in the pH range 2.5 – 7.0 (parallel lines at the top of Fig. 4). The optimum point for MAP production lies where the concentrations of H_3PO_4, (phosphoric acid) and HPO_4^{2-} (the DAP component) are minimized. This is the point where the concentrations of H_3PO_4 and HPO_4^{2-} equal each other, their concentrations lines cross at pH 4.65. A 1 % aqueous solution of MAP has a pH of 4.5 [22].

The maximum concentrations of NH_4^+ and HPO_4^{2-} ions necessary for DAP production are present at about pH 8.0 (parallel lines at the top of Fig. 4). The optimum point for DAP production lies where the concentrations of $H_2PO_4^-$ (the MAP component) and NH_3 (ammonia) are minimized. This is the point where the concentrations of $H_2PO_4^-$ and NH_3 equal each other, their concentrations lines cross at pH 8.1. A 1 % aqueous solution of DAP has a pH of 8.0 [22].

The concentration curves of HPO_4^{2-} and $H_2PO_4^-$ in Figure 4 give the fractions of DAP and MAP present as a function of pH in the neutralization process.

In fertilizer production, high pH must be avoided to minimize formation of free ammonia in the liquid, which is a potential source of loss.

Crystallization. Neutralization and evaporation are used in most fertilizer-producing processes. Knowledge of solubility and crystallization in the system is important. In neutralizing phosphoric acid, the water solubility of a mixture of H_3PO_4, MAP, and DAP is usually presented as shown in Figure 5.

In Figure 6 the N/P molar ratio is plotted as a function of MAP – DAP content.

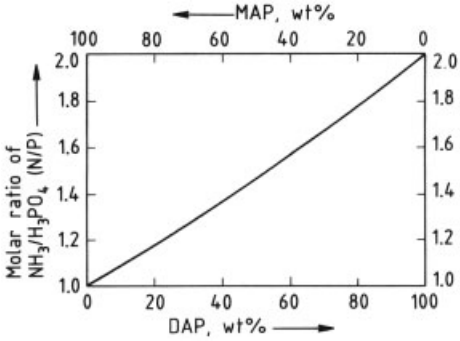

Figure 6. N/P molar ratio as a function of MAP – DAP content

If only water, MAP, and DAP are considered, point 1 in Figure 5 shows that at 75 °C, a saturated solution consists of 75 wt % MAP – DAP, and 25 wt % water. The N/P molar ratio is 1.45, corresponding to 52 wt % MAP and 48 wt % DAP (Fig. 6). The composition at point 1 is therefore 25 wt % water, 39 wt % MAP, and 36 wt % DAP. For this composition a reduction in temperature below 75 °C results in crystallization.

A more informative way of presenting the water – MAP – DAP system is to use a phase diagram (Fig. 7). Points 1 in Figures 5 and 7 represent the same composition. The temperature curves in Figure 7 give the crystallization temperature for a given concentration of MAP, DAP, and water. Point 1 is on the borderline between the crystallization areas for MAP and DAP; MAP and DAP will both crystallize as the temperature is reduced below 75 °C.

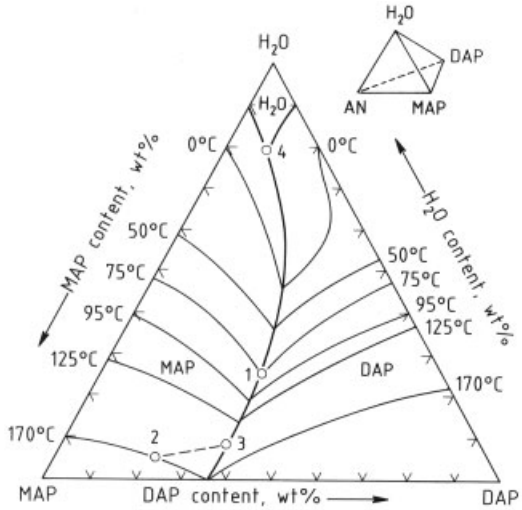

Figure 7. Phase diagram of MAP – DAP – water [24]

A solution given by point 2 in Figure 7 (73 % MAP, 22 % DAP, 5 % water) starts to crystallize at 170 °C. If the temperature is reduced, MAP begins to crystallize. As MAP crystallizes, the liquid composition changes along a straight line through point 2, with its origin at the MAP corner toward point 3. When the liquid phase composition reaches point 3, both MAP and DAP crystallize. Further cooling results in the crystallization of MAP and DAP. The liquid phase composition moves along the borderline between the MAP and DAP crystallization areas. At point 4 the

last liquid phase crystallizes at the eutetic temperature well below −20 °C (to give ice, MAP, and DAP).

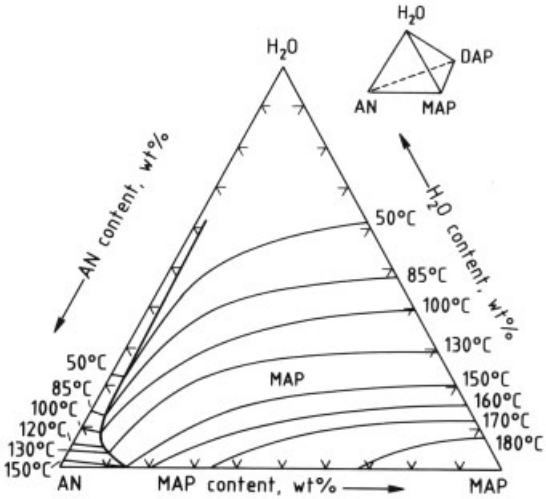

Figure 8. Phase diagram of AN – MAP – water [25]

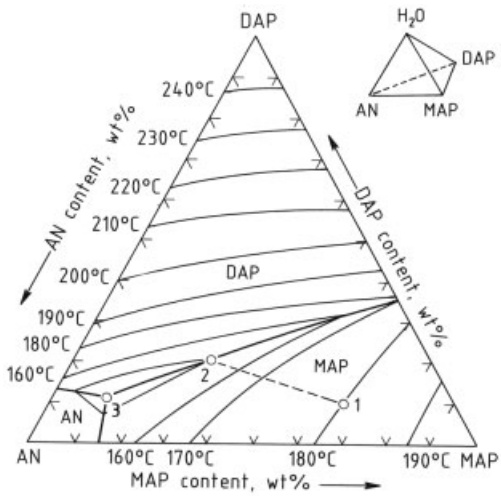

Figure 9. Phase diagram of AN – MAP – DAP [26]

Compound Fertilizers. In compound fertilizers the nitrogen content is usually increased by the addition of ammonium nitrate (AN). A system without DAP can be represented by the AN – MAP – water phase diagram shown in Figure 8. The MAP area is dominant. Compositions with a high MAP content and a low water content tend to supercool (undercool).

If the liquid phase of the compound fertilizer melt also contains DAP, a quaternary tetrahedral phase diagram must be used. Figures 7 and 8 represent two sides of the tetrahedron. Figure 9 gives the AN – MAP – DAP diagram representing the "base" of the tetrahedron. An anhydrous melt is difficult to obtain in compositions with a high DAP content.

For a melt with the composition given by point 1, MAP starts to crystallize if the temperature is reduced below 180 °C (Fig. 9). Only MAP crystallizes until the temperature reaches 140 °C (point 2); then both MAP and DAP begin to crystallize. On further cooling the composition of the liquid phase follows the borderline between the MAP and DAP crystallization areas. The last liquid crystallizes at the eutectic temperature of ca. 130 °C (point 3).

With increasing water content, the system becomes a quaternary system. To simplify the tetrahedral phase diagram, the water level is kept constant while the three other components are varied. This corresponds to cutting the tetrahedron at a given water content. Figure 10 shows a section through such a tetrahedron, where the water content for all compositions is 10 wt %.

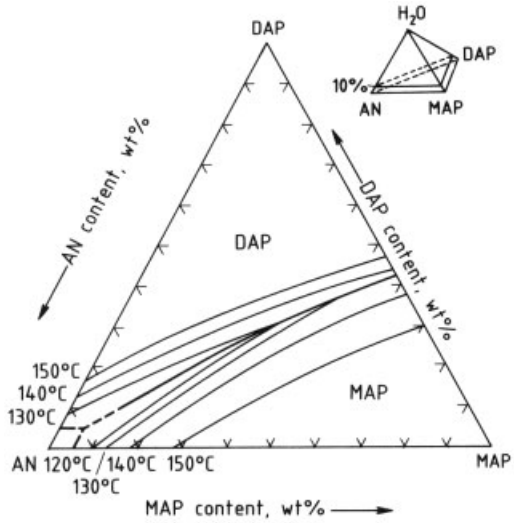

Figure 10. Phase diagram of AN – MAP – DAP with 10 % water [27]

Three primary crystallization areas exist. The lines represent compositions at which the liquid phase is in equilibrium with two solid phases. For one composition the liquid phase is in

equilibrium with the three solid phases, AN(s), MAP(s), and DAP(s).

Calcium Phosphates. The phosphate component in superphosphates, MCP, is produced as follows:

$$Ca_5F(PO_4)_3 + 7\,H_3PO_4 \rightleftharpoons 5\,Ca(H_2PO_4)_2 + HF$$

Phosphoric acid is produced by the sulfuric acid treatment of phosphate rock, gypsum ($CaSO_4$) is formed as a byproduct. In triple superphosphate production, gypsum is separated from phosphoric acid prior to MCP formation. If calcium sulfate ends up in the product, it can be considered to be an inert solid.

In nitrophosphate processing, ca. 50 % of the remaining calcium in the phosphoric acid mother liquor reacts to give DCP, $CaHPO_4$, during neutralization.

Both superphosphate and nitrophosphate processing involve calcium phosphate solid – liquid equilibria:

$$Ca^{2+} + 2\,H_2PO_4^- \rightleftharpoons Ca(H_2PO_4)_2\,(s)\ \text{MCP}$$

$$Ca^{2+} + HPO_4^{2-} \rightleftharpoons CaHPO_4(s)\ \text{DCP}$$

Because these equilibria depend on ion concentration and temperature, the phase diagram described by GMELIN may be used to determine which calcium phosphate precipitates (Fig. 11) [28].

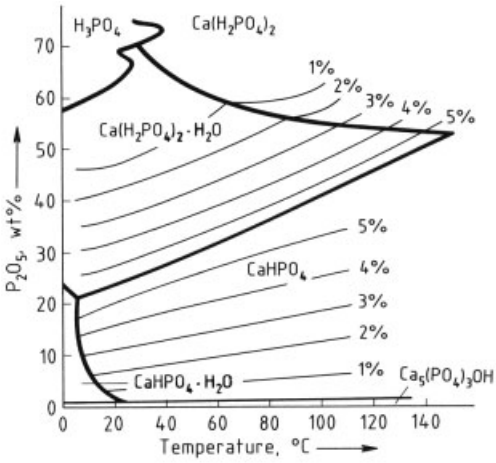

Figure 11. Phase diagram for calcium phosphates [28] The lines indicate the minium CaO content in the liquid necessary for calcium phosphate precipitation

MCP is produced under the conditions used in superphosphate production (high liquid P_2O_5 concentration); DCP is produced under the conditions used in nitrophosphate processing (high temperature, lower liquid P_2O_5 concentration).

When the compound-fertilizer route is based on superphosphates, MCP is partially converted to DCP and MAP during neutralization with ammonia:

$$Ca(H_2PO_4)_2 + NH_3 \longrightarrow CaHPO_4 + NH_4H_2PO_4$$

In compound-fertilizer granulation, both MCP and DCP are in the solid state and can be considered as inert.

Additives in Compound-Fertilizer Production. Potassium salts may be added in compound-fertilizer production:

$$KCl + NH_4NO_3 \rightleftharpoons KNO_3 + NH_4Cl$$

$$K_2SO_4 + 2\,NH_4NO_3 \rightleftharpoons 2\,KNO_3 + (NH_4)_2SO_4$$

To achieve steady state, the potassium salt has to be dissolved in a multicomponent melt. This may affect salt crystallization and melt viscosity (Fig. 12). If this reaction is not controlled, the liquid phase may increase during granulation or storage.

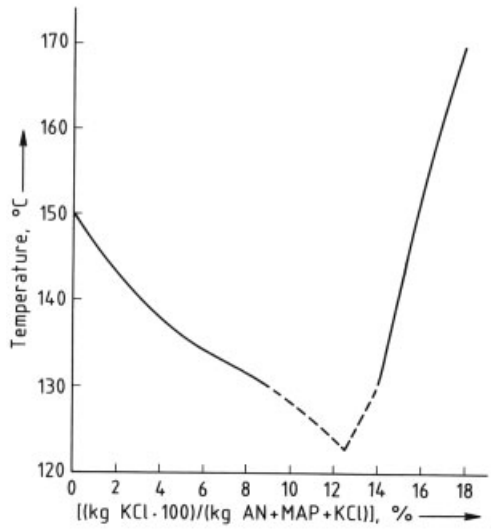

Figure 12. Crystallization points for potassium chloride dissolved in an 86 % AN – 14 % MAP melt with 0.5 % water [29]

Other salts added to MAP- and DAP-containing melts (e.g., ammonium sulfate, magnesium sulfate, dolomite, borax, micronutrients) can be considered as inert salts in phosphate fertilizer production.

8. Superphosphates

For definitions, see Chapter 2.

8.1. Single Superphosphate

8.1.1. Chemistry

Single superphosphate is produced by acidulation of finely ground phosphate rock with sulfuric acid. Many chemical reactions occur when the phosphate mineral is mixed with sulfuric acid (70 %). The mixture is liquid for 2 – 10 min but solidifies in the next 5 – 10 min.

After 40 – 60 min the superphosphate product is transported to storage where reaction is completed after 3 – 10 days. The reaction rates depend on the type and particle size of the ground phosphate rock, the type and concentration of trace elements in the rock, and the concentration and amount of sulfuric acid.

The primary overall reaction for fluorapatite is

$$2\,Ca_5F(PO_4)_3 + 7\,H_2SO_4 \longrightarrow 3\,Ca(H_2PO_4)_2 + 7\,CaSO_4 + 2\,HF$$

It has two consecutive stages:

$$Ca_5F(PO_4)_3 + 5\,H_2SO_4 \longrightarrow 3\,H_3PO_4 + 5\,CaSO_4 + HF$$

$$Ca_5F(PO_4)_3 + 7\,H_3PO_4 \longrightarrow 5\,Ca(H_2PO_4)_2 + HF$$

The first reaction is complete in 5 – 20 min. Components of the rock, such as calcium fluoride, calcium carbonate, silica, iron, and aluminum, all affect reaction rate.

Silica reacts with hydrogen fluoride from the above reaction to give fluosilicic acid, which dissociates to silicon tetrafluoride:

$$6\,HF + SiO_2 \longrightarrow H_2SiF_6 + 2\,H_2O$$

$$H_2SiF_6 \longrightarrow SiF_4(g) + 2\,HF$$

Phosphate rock contains 2 – 5 % fluorine, of which 10 – 30 % is rapidly evolved as silicon tetrafluoride during acidulation. The SiF_4-enriched off-gas is used for the production of fluosilicic acid (→ Fluorine Compounds, Inorganic, Chap. 5.2.; see also Section 8.1.3). The rest of the fluorine remains in the superphosphate product as unreacted calcium fluoride, water-soluble calcium silicofluoride, or free fluosilicic acid.

Practical experience shows that phosphates rich in carbonate are acidulated more quickly and evolve more silicon tetrafluoride. The silicon tetrafluoride yield increases with higher sulfuric acid concentration, higher temperature, more finely ground rock, and longer mixing time. However, some of these conditions are harmful for the production of a good superphosphate.

The effects of particle size and acid concentration on the reaction rate for Morocco rock have been studied in the laboratory [30]. In the example shown in Figure 13 the free sulfuric acid is consumed after 60 min. The concentration of free phosphoric acid reaches a maximum of ca. 58 %. After 1 h, approximately 80 % of the phosphate is water-soluble, about 70 % of this being free phosphoric acid. The amount of water-soluble phosphate increases to 92 % after one week and to 93 % after three weeks. At this time, 14 % is free phosphoric acid (corresponding to 3 % P_2O_5 in superphosphate).

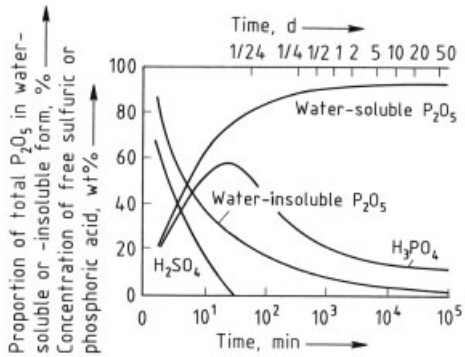

Figure 13. Reaction of Morocco phosphate rock with 76 % sulfuric acid

On a technical scale the mixture goes to storage for curing (i.e., complete reaction) after 0.5 – 2 h. The mixture solidifies to form a gel consisting of colloidal MCP and containing anhy-

drous calcium sulfate crystals. The gel structure gives the fresh superphosphate a high plasticity. Superphosphate becomes fluid under pressure. These properties make handling difficult but facilitate granulation.

With increasing storage time and decreasing content of free acid, the gel structure becomes crystalline and plasticity decreases.

Byproducts. Superphosphate contains anhydrous calcium sulfate (gypsum). Small amounts of the dihydrate occur after prolonged storage when the content of free acid is very low. Hydration of calcium sulfate is inhibited strongly in the highly viscous aqueous phase. The stability limit for $CaSO_4 \cdot 2\,H_2O$ is reached rather early because hydration results in concentration of the liquid phase (Fig. 14).

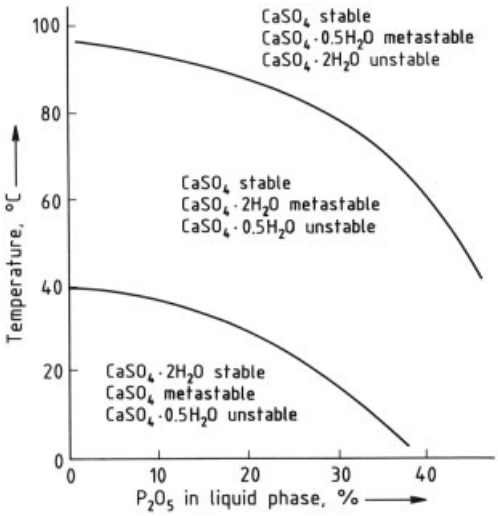

Figure 14. Different forms of gypsum in superphosphate [31]

When the concentration of free phosphoric acid decreases below a certain limit, citrate-soluble DCP is formed:

$$Ca(H_2PO_4)_2 \rightleftharpoons CaHPO_4 + H_3PO_4$$

Formation of DCP can be reduced or avoided by the use of more concentrated acid or by drying the superphosphate product to increase the acid concentration in the liquid phase.

The iron and aluminum compounds in phosphate rock can form water-soluble compounds during acidulation, but they react to produce water-insoluble compounds when the amount of free acid is reduced.

8.1.2. Production [31]

The production of superphosphate consists of grinding the phosphate rock, mixing the reaction components, acidulation, curing in a den, transportation to storage for final curing, and granulation or further processing to compound fertilizers.

Grinding Phosphate Rock. Phosphate rock is ground before being mixed with sulfuric acid. The particle size depends on different factors: reactive phosphate rock can be ground more coarsely; the use of concentrated sulfuric acid demands a more finely ground rock. Modern continuous superphosphate plants with a short curing time also need a more finely ground rock, in general >90 % through 100 mesh (<150 μm).

Batch Processes. Up to about 1970 the batch process dominated because it allowed a more accurate feed of phosphate rock and sulfuric acid. Mixing time could also be adjusted to give a good quality product. The primary disadvantage was the need for manual cleaning.

Examples of batch processes are the Beskow den process (developed in Europe, Fig. 15) and the Sturtevant den process (developed in the United States).

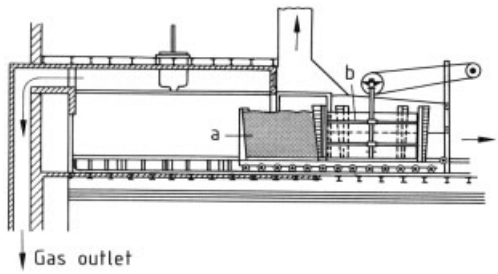

Figure 15. Beskow den batchwise superphosphate process a) Reaction vat; b) Rotating cutter for removing superphosphate product

Beskow Den Process. The chamber floor is mounted on rollers. The curved back wall of the chamber is fixed to the floor and the side walls can be folded away. The front wall consists of two folding doors that are opened prior to emptying the chamber. The chamber wagon is then

pulled toward a fixed revolving cutter equipped with vertical knives that scrape off the superphosphate. Beskow wagons can hold 50 – 100 t of product. Mixing of the components must be rapid and thorough so that the mixture remains thin enough to run over the surface of the chamber contents.

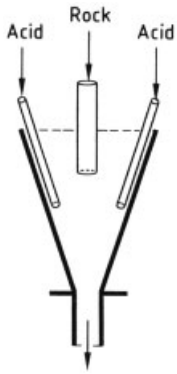

Figure 16. The TVA mixer for superphosphate production

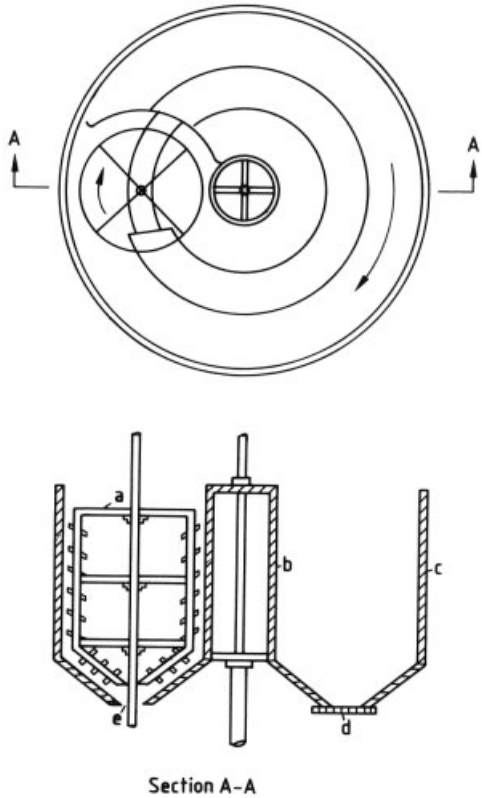

Figure 17. Moritz – Standaert superphosphate den process [32]
a) Cutter; b) Inner cylinder; c) Outer cylinder; d) Stationary bottom; e) Superphosphate outlet

Continuous Processes. Continuous processes are now dominant. Sulfuric acid is measured and controlled by magnetic flow meters. Different types of paddle mixers are used.

The *Tennessee Valley Authority (TVA)* has developed a simple cone mixer (Fig. 16) without moving parts. The acid is fed in by two to eight tangential nozzles, and phosphate rock is fed into the center of the swirling acid. This mixer is usually combined with a den belt and used mainly to produce triple superphosphate.

The *Moritz – Standaert den process* was often used in Western Europe and involves a movable, circular den (Fig. 17).

Another common European process is the *Broadfield den process* (Fig. 18). The Broadfield den consists of a slat conveyor mounted on rollers, with a long stationary box over it and a revolving cutter at the end. A variable-speed drive is used for the conveyor, giving a retention time of ca. 30 min [32]. A similar process was developed by *Nordengren*. Here, a reaction chamber is built into the top of the den to allow sufficient time for solidification of the superphosphate mixture [32].

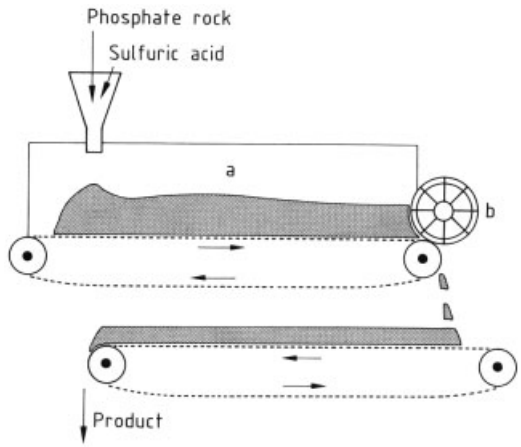

Figure 18. Broadfield den process
a) Den; b) Cutter

In the *Kuhlmann den process* (Fig. 19), denning is accomplished by transporting a thin layer of the mixture on a long, flexible belt conveyor. The den is troughed for more than half its length to form its own side walls. A hood is provided to collect the off-gas [32].

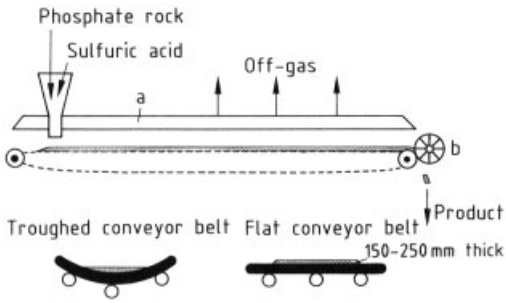

Figure 19. Kuhlmann den superphosphate process
a) Hood; b) Cutter

A high-speed mixer used in combination with a den has been developed by *Agrimont* in Italy (Fig. 20). The very brief and intense mixing gives a short curing time and a high-quality superphosphate.

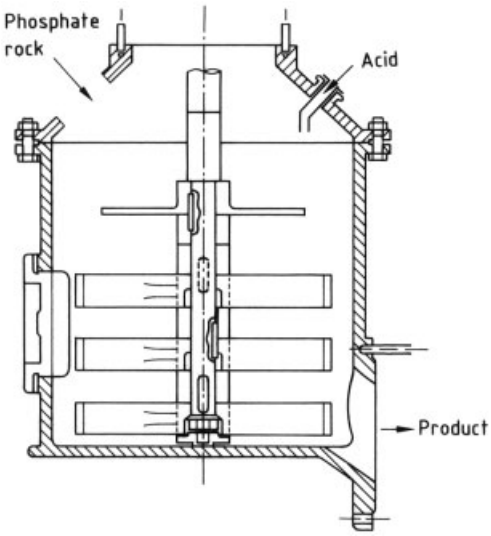

Figure 20. Agrimont mixer for superphosphate production (courtesy of Agrimont Italy)

8.1.3. Fluorine Recovery [33]

Off-gases from the mixer and the den contain air, steam, carbon dioxide, and silicon tetrafluoride. The fluorine content is normally 0.5 –2 vol % but depends on process conditions and air flow.

When the off-gases are washed in scrubbers, silicon tetrafluoride reacts with water to form fluosilicic acid [*16961-83-4*], which is used as a commercial product (→ Fluorine Compounds, Inorganic, Chap. 5.2.):

$$3\,SiF_4 + 2\,H_2O \longrightarrow 2\,H_2SiF_6 + SiO_2$$

The silica is at first colloidal and therefore impossible to filter. It solidifies in an uncontrollable fashion.

The first cleaning units consisted of Venturi scrubbers and spinning disk scrubbers (Fig. 21). Hydro Supra has now developed a process giving clean fluosilicic acid (20 – 25 % H_2SiF_6 containing <100 mg of P_2O_5 per liter) suitable as a raw material for the production of aluminum fluoride from alumina (Fig. 22). The first stage is a Venturi scrubber (a), where solid P_2O_5 is absorbed. The fluosilicic acid produced in absorption towers (b) is fed to an aging tank (c) where silica particles grow. The silica particles are then filtered from the product acid (d) and mixed with the Venturi liquid (e) before being recycled to the process.

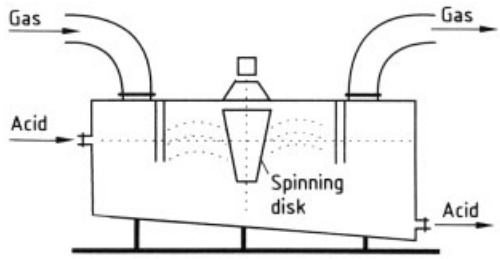

Figure 21. Spinning disk scrubber

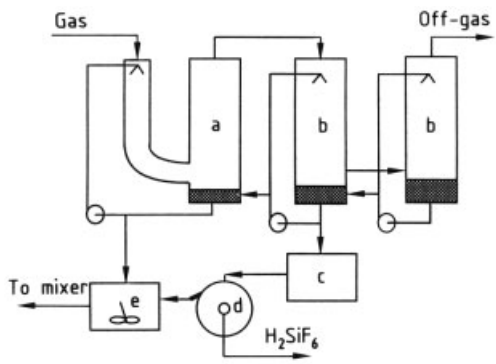

Figure 22. Hydro Supra fluosilicic acid production
a) Venturi scrubber; b) Absorption tower; c) Aging tank; d) Filter; e) Recycle tank

8.1.4. Granulation [34]

After 3 – 6 weeks storage, superphosphate may be crushed, bagged, and sold as powder superphosphate. The product is dusty and has a high caking tendency. This disadvantage can be avoided by granulation to 2 – 5-mm particles. Granulated superphosphates have been on the market since 1950 and now account for all commercial superphosphates in the Western world.

Run-of-pile powder superphosphate is granulated with steam or water in a drum (length 6 – 12 m, diameter 3 – 5 m). Another possibility is use of a pan granulator with a diameter of 3 – 5 m.

Freshly produced superphosphate is more easily granulated but may cause handling problems due to its plasticity (see Section 8.1.1).

After granulation, the superphosphate product is dried in rotating drums, cooled in a similar drum, and screened to a product size of 2 – 5 mm. The fine undersize particles are directly recirculated to the granulator, whereas the coarse oversize particles are crushed before recycling. The final product can be cooled further before storage and dispatch.

Dust formed in the dryer, the cooler, the screens, and the crushers is separated from the off-gases in dust-collecting equipment.

8.2. Triple Superphosphate [35]

Single superphosphate contains up to 50 % calcium sulfate. If sulfuric acid is replaced by phosphoric acid for acidulation of phosphate rock, triple superphosphate is produced with a higher content of MCP and without calcium sulfate. The principal chemical reaction is

$$Ca_5F(PO_4)_3 + 7\,H_3PO_4 \longrightarrow 5\,Ca(H_2PO_4)_2 + HF$$

Triple superphosphate has similar properties to single superphosphate but lower plasticity. Optimal process conditions and product quality are reached with a phosphoric acid concentration of 50 – 54 % P_2O_5. The same equipment is used as for single superphosphate production (see Section 8.1.2). However, the mixing time is shorter, due to faster chemical reaction (10 – 20 s). The reaction heat is one-third that for single superphosphate. The same temperature (80 – 100 °C) is reached, but less water vapor and silicon tetrafluoride are evolved.

Granulation based on run-of-pile powder triple superphosphate is similar to granulation of single superphosphate (Section 8.1.4).

A slurry granulation process can also be used. The Dorr – Oliver process route is usually employed. Phosphoric acid (40 % P_2O_5) is mixed with ground phosphate rock in a large tank reactor with a retention time of 3 – 5 h. The slurry is then pumped directly to a granulator. As in the run-of-pile route, the granulation process consists of drying, crushing, recycling, and cooling.

8.3. Double Superphosphate

In the production of double superphosphate, phosphate rock is treated with a mixture of sulfuric and phosphoric acids. The phosphate content of double (enriched) superphosphate thus corresponds to mixtures of single superphosphate and triple superphosphate containing 18 – 45 % P_2O_5. The product is used mainly as an intermediate for PK products [36].

By mixing sulfuric and phosphoric acids, more dilute phosphoric acid can be used. This reduces the energy requirement for concentration of phosphoric acid. The balance between sulfuric acid and phosphoric acid depends on the acid concentration [32].

8.4. PK Fertilizers [36]

Triple superphosphate or double superphosphate can be granulated together with potassium chloride, potassium sulfate, and magnesium sulfate. The granulation process is the same as for single superphosphate (see Section 8.1.4), but chloride salts release hydrogen chloride gas during drying. The hydrogen chloride gas must be recovered in a scrubber.

The PK fertilizers are normally applied in the fall.

9. Ammonium Phosphates

For further details of ammonium phosphates, see → Phosphoric Acid and Phosphates.

9.1. Fertilizer Grades and Applications

Fertilizer-grade ammonium phosphates are normally produced from nonpurified wet process phosphoric acid (WPPA, or the so-called sulfur route) and consequently have a purity of only 70 – 85 %. In spite of relatively large deviations from the stoichiometric N/P ratio of the pure products, in the fertilizer grade they are called monoammonium phosphate (MAP), diammonium phosphate (DAP), and ammonium polyphosphate (APP). Triammonium phosphate, $(NH_4)_3PO_4$, has no significance as a fertilizer because it is unstable at 20 °C.

The nutrient content of solid ammonium phosphate fertilizers depends greatly on the quality of the phosphoric acid (Table 2).

Table 2. Composition of ammonium phosphate fertilizers

Type	N, wt %	P_2O_5, wt %
MAP from fertilizer acid	10 – 12	57 – 48
MAP from purified acid	12	61
DAP from fertilizer acid	16 – 18	48 – 46
DAP from purified acid	21	53

Granular MAP and DAP are used as straight fertilizers or applied after mechanical blending with other granular fertilizers (bulk blending). Ammonium phosphate salts are also formed in situ or mixed into processes for the production of compound NP or NPK fertilizers from phosphoric acid or nitrophosphates.

Solid MAP is used extensively in the United States for the manufacture of suspension fertilizers (see → Fertilizers, Chap. 3.2.3.).

Ammonium polyphosphates are normally a constituent of liquid fertilizers (see → Fertilizers, Chap. 3.2.2.1.). The solid products were also expected to have a very important potential in connection with their high plant nutrient value and other special features, but in practice this did not turn out to be true. Only small quantities of solid products have been produced with a polyphosphate content up to 25 % of the ammonium phosphate [37], [38].

Pure MAP and DAP are used as fertilizers in irrigated greenhouses (completely water-soluble compounds).

9.2. Production

For a long time solid ammonium phosphate fertilizers were produced only in granular form in an integrated neutralization – granulation process [39]. This is still the case for DAP, but since the mid-1960s, considerable capacity has existed for producing MAP in powder form. This product is used as a relatively concentrated, low-cost intermediate for compound-fertilizer manufacture.

Crystallization processes for technically pure ammonium phosphates are not described here.

The high ammonia vapor pressure of DAP has important consequences for its industrial production. The same applies to the pronounced maximum solubility of a mixture of MAP and DAP at an N/P molar ratio of ca. 1.4 [40] (eutectic mixture containing 17 % water, *fp* 110 °C). The point at which the borderline in Figure 7 crosses the 110 °C isotherm corresponds to the composition 46 % MAP, 37 % DAP, and 17 % H_2O. This is equal to an N/P molar ratio of 1.4 (Fig. 6).

9.2.1. Ammonium Phosphate Powder

MAP Powder. The processes used for the production of MAP powder (particle size 0.1 – 1.5 mm) are relatively simple. All processes are claimed to give good storage, handling, and granulation properties when used for the manufacture of MAP, DAP, or other NP/NPK compound fertilizers.

In the *PhoSAI process* of Scottish Agricultural Industries [41] (Fig. 23), ammonia reacts with phosphoric acid (minimum concentration 42 % P_2O_5) in a stirred-tank reactor (d) to produce a slurry with an N/P ratio of 1.4 (point of maximum solubility) under atmospheric pressure. The slurry flows to a pin mixer (e) in which the N/P ratio is brought back to ca. 1 by the addition of more concentrated phosphoric acid. During this step the solubility decreases and more water vaporizes due to the heats of reaction and crystallization. A solid product is formed that typically contains 6 – 8 % moisture. This is screened (f) and the oversize particles are ground, after which the product can be stored without further treatment.

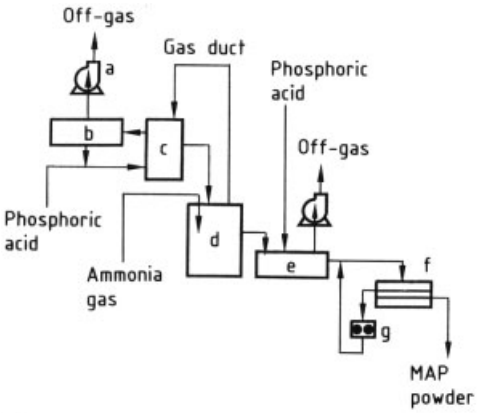

Figure 23. PhoSAI process for nongranular MAP
a) Fan; b) Separator; c) Gas scrubber; d) Reactor; e) Pin mixer; f) Screen; g) Oversize crusher

In the *Minifos process* [42] of Fisons (now Hydro Fertilizers), ammonia and phosphoric acid (45 – 54 % P_2O_5) react (Fig. 24, b) at 0.21 MPa to give an N/P ratio of 1. The liquid product is flash-sprayed in a natural-draught tower (c), where the droplets solidify to form a powder with 6 – 8 % residual moisture.

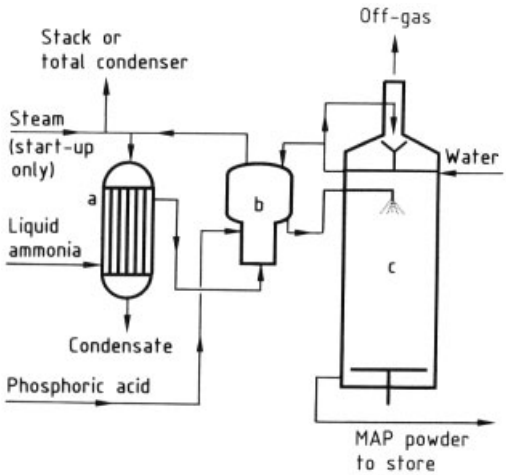

Figure 24. Hydro Minifos process for MAP powder
a) Ammonia evaporator; b) Pressure reactor; c) Spray tower

Alternative processes based on the use of a pipe reactor instead of a tank reactor were developed by Gardinier S.A. [39], Swift Agricultural & Chemical Corporation (Fig. 25) [43], and ERT – Espindesa [44], [45]. Product humidity is reported to be as low as 2 %, which results in less caking.

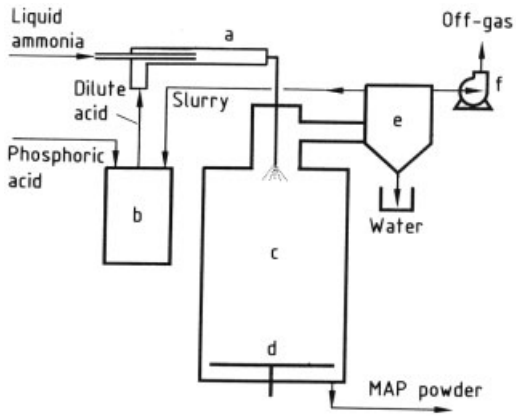

Figure 25. Swift process for MAP powder
a) Pipe reactor; b) Surge tank; c) MAP tower; d) Discharge scraper; e) Scrubber system; f) Fan

More recent MAP powder plants are based on the use of pipe reactors (including those designed by Hydro Fertilizers).

Considerable quantities of MAP powder are produced from "sludge acids" from phosphoric acid clarification with solids contents up to 20 %.

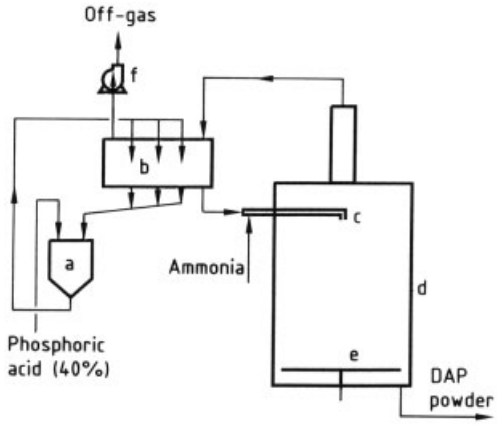

Figure 26. ERT – Espindesa process for DAP powder
a) Buffer vessel; b) Scrubber; c) Pipe reactor; d) DAP tower; e) Discharge scraper; f) Fan

DAP Powder. Based on the favorable results obtained with a pipe reactor for MAP powder manufacture, ERT – Espindesa experimented with the production of DAP powder [44], [45] (Fig. 26). The DAP product has a remarkably high water- and citrate-soluble P_2O_5 content, probably due to the short residence time in the reactor, which does not yield the insoluble

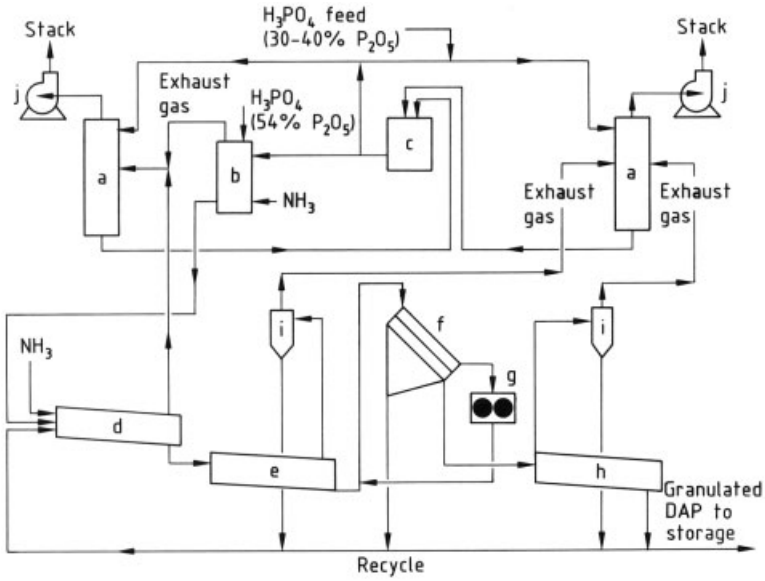

Figure 27. The TVA process for granulated DAP
a) Scrubber; b) Preneutralizer; c) Surge; d) Granulator; e) Dryer; f) Screens; g) Crusher; h) Cooler; i) Cyclone; j) Fan

salts that are normally formed in long-retention-time DAP plants. The same effect has been observed with MAP formed in a pipe reactor.

9.2.2. Granular Ammonium Phosphates

Since the 1960s, granular ammonium phosphates have belonged to some of the most important fertilizer products [46]. They are applied directly or used as intermediates in bulk blends and for the manufacture of compound fertilizers. More than five times as much DAP is produced as MAP. Since the beginning of large-scale production in the 1930s, considerable developments have been made to optimize raw material and energy efficiencies, to reduce the capital cost of the plants, to improve their reliability, and make them more environmentally friendly.

Granular ammonium phosphate plants consist of a wet and a dry section, the dry section being a granulation loop (→ Fertilizers, Chap. 5.).

Granular DAP. In the 1950s the *Dorr – Oliver process* was most commonly used. Phosphoric acid and ammonia react in two or three vessels in series at atmospheric pressure under conditions that take advantage of solubility and vapor pressure properties.

Initially, three reactor vessels were used. Neutralization to MAP was carried out in the first. In the second, more ammonia was injected until ca. 80 % DAP was formed. Final ammoniation occurred in the third reactor but was later carried out in the blunger (pugmill) of the granulation loop. Because of the low solubility of DAP the process must be run with a high slurry water content to maintain fluidity. High recycle ratios are also required to avoid overgranulation in the blunger where the slurry and recycle are mixed to give a moist granulate containing 4 – 6 % water. Typically, these recycle ratios are between 11 : 1 and 8 : 1, depending on where final ammoniation occurs. The granules are dried in a concurrent rotary dryer and screened. The fines, smaller granules, and broken oversize particles are recycled to the pugmill. Cooling of the on-size product (1.5 – 3 mm) is not essential but is recommended to minimize loss of ammonia and ensure good storage stability.

Off-gases from reactors, granulator, and dryer are scrubbed with filter-strength (28 % P_2O_5) acid. Stronger acid (40 – 52 % P_2O_5) is used in the reactor section.

In the early 1960s the TVA [47], [48] (Fig. 27) introduced a modified process that became very

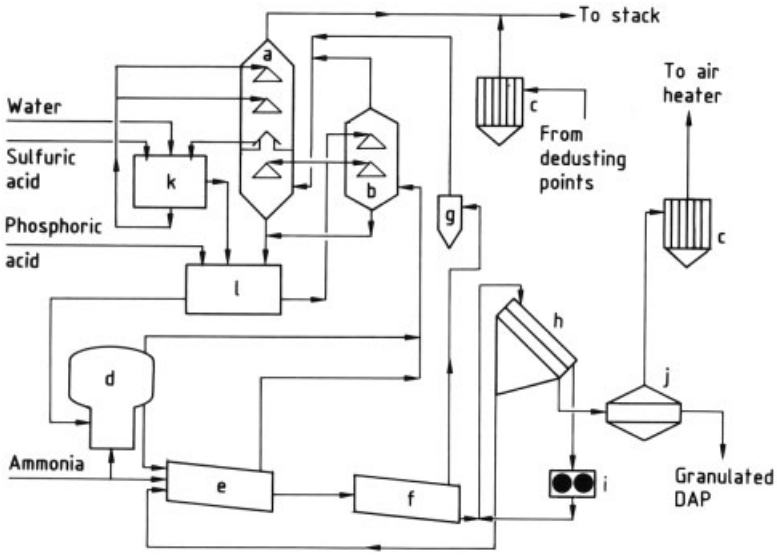

Figure 28. Hydro Fertilizers DAP granulation process with pressure reactor
a) Tail-gas scrubber; b) Ammonia scrubber; c) Bag filter; d) Pressure reactor; e) Granulator; f) Dryer; g) Cyclone; h) Screens; i) Crusher; j) Fluidized-bed cooler; k), l) Surge tanks

popular because the recycle ratio dropped to ca. 5:1.

Acid reacts with ammonia in a preneutralizer (b) where the N/P ratio is controlled at about 1.4. The heat of reaction raises the slurry temperature to the boiling point (ca. 115 °C) and evaporates water. The hot slurry contains 15 – 20 wt % water and is pumped to a rotary drum granulator (d) where it is sprayed on a relatively thick bed of recycled material. Additional ammonia is sparged underneath this bed to increase the molar ratio to 2.0. Additional heat is generated, resulting in the evaporation of more moisture. The decrease in solubility obtained in going from a molar ratio of 1.4 to 2.0 assists granulation.

Further reduction of the recycle ratio (e.g., to 4:1) is attained by carrying out preneutralization to an N/P ratio of 1.4 in a pressure reactor (Hydro Fertilizers) [47], [49] operating at 0.1 MPa overpressure; so that the boiling point of the reaction mixture is elevated by ca. 20 °C compared to operation at atmospheric pressure (Fig. 28). Thus, the steep temperature dependence of the solubility of ammonium phosphate in water allows operation at a lower water content while maintaining the ammonium phosphate in solution. The amount of water fed to the granulator is therefore minimized.

In an attempt to further simplify DAP plants and reduce the recycle ratio, TVA, CROS S.A., and ERT – Espindesa have tried to use pipe reactors [50] (also called T reactors or pipe-cross reactors) directly releasing slurry in the rotary granulator, instead of applying a preneutralization tank.

CROS and ERT – Espindesa [51] report successful operation, but experiences in several U.S. plants were disappointing because of high heat input in the granulator.

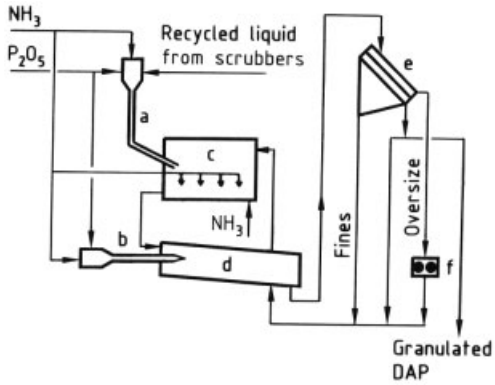

Figure 29. The AZF dual pipe process for granulated DAP
a, b) Pipe reactors; c) Granulator; d) Dryer – cooler; e) Screen; f) Mill

CdF – Chimie AZF [51–53] originally had the same experience, but succeeded by introducing a second pipe reactor spraying into the dryer (Fig. 29).

Approximately half of the phosphoric acid is fed to a pipe reactor (a) that operates at N/P = 1.4 and releases its product into the granulator (c). The remaining acid is fed into a second pipe reactor (b) at N/P = 1.1 that sprays into the dryer (d). Extra ammonia is introduced underneath the rolling bed in the rotary drum granulator.

The pipe reactor (b) in the dryer produces MAP powder. Part of it is carried away into the cyclones of the dedusting loop; the remainder crystallizes on the DAP product.

The recycle ratio of this process is lower than with a conventional preneutralizer; fuel consumption of the dryer is also reduced.

Granular MAP. Granular MAP is produced in the same types of plants as DAP. The most common process is the TVA, with a preneutralizer and rotary drum ammoniation (Fig. 27). The preneutralizer is operated at an N/P ratio of 0.6, again a point of high solubility, and more ammonia is added in the granulator to increase the molar ratio to 1.0.

In contrast to DAP production, ammonia recovery by acid scrubbing is not necessary, but all off-gases are scrubbed to recover dust and fumes. Granular plants can thus be simpler, but most units are designed so that both products can be made in the same plant.

The pipe reactor processes [43] have been very successful in MAP production. They operate with a very slight ammonia loss, and little or no additional heat is required for drying. In some plants, preneutralizers have been replaced by pipe reactors.

9.2.3. Off-Gas Treatment

An essential process step in all ammonium phosphate plants is scrubbing of the off-gas. The preneutralizer and pipe reactors give ammonia slip due to vapor pressure and "mechanical losses" caused by imperfect mixing. Absorption of ammonia in the rotary drum or blunger is not 100 %; ammonia is released from DAP in the dryer and cooler due to its high vapor pressure. Off-gases may also contain dust and ammonium fluoride aerosols.

In an environmentally friendly plant, all off-gases are treated [54]. In DAP plants, relatively large quantities of ammonia must be recovered. This is normally done by using filter-strength (28 % P_2O_5) phosphoric acid, which is subsequently sent to the preneutralizer or pipe reactor. Dilute sulfuric acid scrubbing may also be employed. Sulfuric acid can also increase the N/P ratio of the product.

Passage of large quantities of air through scrubbers that operate on wet process phosphoric acid results in the stripping of considerable quantities of fluorine-containing gases, which must be absorbed in tail-gas scrubbers. In the production of granular MAP, off-gas treatment is simpler due to the lower ammonia slip.

In modern powder MAP plants, entrained dust is recovered in a scrubber with circulating process water that is sent to the reactor section.

10. Compound Fertilizers by the Sulfur Route

Compound fertilizers (also called complex or mixed fertilizers) contain more than one plant nutrient element. They may be ternary (N + P + K) or binary (N + P, P + K, or N + K). They may also contain considerable quantities of magnesium or trace elements.

The reason for applying compound fertilizer is convenience: farmers no longer have to apply several straight (single-element) fertilizers separately. However, even if compound fertilizers are used, nitrogen is often still applied separately because supplementary dressings of nitrogen may be needed during the growing season. Also, some straight nitrogen materials (e.g., urea – ammonium nitrate solution) are sometimes considerably cheaper than the same amount of nitrogen in compound fertilizers.

The first compound fertilizers were powder mixtures (e.g., of ammonium sulfate, superphosphate, and potash). In the 1930s, producers began to ammoniate superphosphate and thus introduced a cheap form of nitrogen, while improving the physical properties of the superphosphate.

To lower transportation costs, higher-grade products were used, which, however, caused se-

rious caking problems. Increased mechanical application called for free-flowing material, and in the 1950s granulation therefore started to become popular.

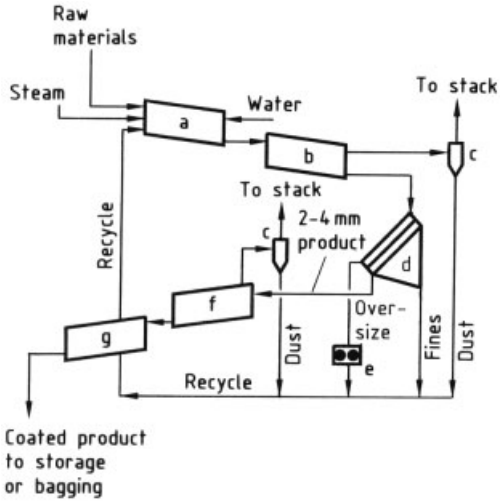

Figure 30. Typical plant for granulation of dry materials a) Rotary drum granulator; b) Dryer; c) Cyclone; d) Screen; e) Crusher; f) Cooler; g) Coating drum

The concentration of compound fertilizers increased as single superphosphate was replaced by triple superphosphate, and ammonium nitrate or urea replaced ammonium sulfate. Later, even higher grades were attained by using phosphoric acid and ammonia. Production was no longer a simple, mainly mechanical operation, but rather a complex chemical processing operation.

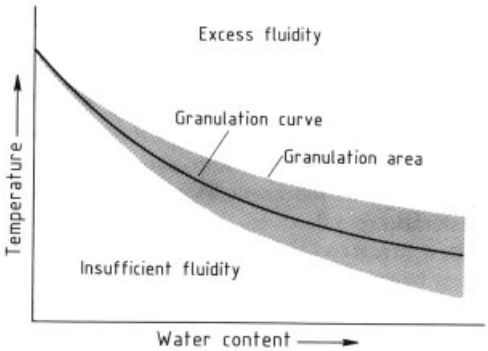

Figure 31. Relationship of temperature and water content for the granulation of fertilizers

Bulk blends (see → Fertilizers, Chap. 5.5.) became popular in some countries as a method by which a local dealer with a simple mixing plant could supply whatever mix the farmer needed. Granular materials with an approximately equal size distribution (e.g., DAP, compacted potassium chloride, ammonium nitrate, or urea) are used. These compounds are shipped in bulk to the blenders who offer prescription mixing, custom application, and other services to the farmers.

Dry mixing of nongranular materials is no longer used and therefore not described here.

10.1. Granulation of Mixtures of Dry Materials

Agglomeration with Steam or Water. Mixtures of dry powders or ground components can be granulated by agglomeration with the aid of steam or water.

Several devices are used, but rotary drum granulation is the most popular method (Fig. 30).

Incoming materials are screened to remove coarse lumps, which are crushed. The materials are fed in by batchwise weighing or with continuous weighing belts. In the rotary drum granulator (a) the mixture of recycle and raw materials is adjusted for agglomeration by injecting steam underneath the bed and, if necessary, by spraying water on top. Granulation is performed so that optimum quantities of on-size granules (in Europe mostly 2 – 4 mm, in the United States normally 1 – 3.3 mm) are formed, with a minimum of oversize particles [55]. Obtaining more than 60 – 70 % of the product in the desired size range is difficult, but the recycle ratio for such operations is normally $<1:1$. Optimum granulation is obtained at a defined "liquid-phase" content (i.e., the water content plus the salts that dissolve in this amount of water). Because the solubility of salts increases with temperature, a higher temperature means that less water is required (Fig. 31). Although such a curve can be used to predict the optimum point at which granulation occurs, it does not describe granulation efficiency, which depends on the degree of plasticity of the mixture. Pure inorganic salts (e.g., KCl, NH_4NO_3) have little plasticity and are relatively difficult to granulate. In contrast, fresh single superphosphate shows excellent agglomeration properties (see section 8.1.2). The binding tendency of ammonium phosphates is highly dependent on their impurity content [56].

Steam granulation is normally better than the use of water alone because the damp granules are hotter, have a lower moisture content, and, therefore dry more easily. The product is usually more dense and stronger. The curve shown in Figure 31 depends not only on the composition of the salt mix, but also on the physical properties of the raw materials. The particle size of the feed and the potash coating may affect binding ability [57].

After the granulator, the remainder of the plant is similar for all granulation processes (→ Fertilizers).

The grades produced in such a plant depend on the raw materials. The product can be very low in concentration (N-P_2O_5-K_2O = 9-9-9) if it is based on single superphosphate and ammonium sulfate, and very concentrated (19-19-19) if it is based on urea and nongranular MAP or DAP.

Compaction and Extrusion [58–62].
Roll compaction is described elsewhere (→ Fertilizers, Chap. 5.3.5.) and is used relatively little for the manufacture of compound fertilizers. Compaction is used for materials that are difficult to granulate (e.g., straight KCl). It operates at a low moisture level (0.5 – 1.5 %) and at ambient temperature, so that heat-sensitive materials can be admixed.

Particles made by compaction tend to be angular compared to the spherical granules produced by other shaping techniques. A further disadvantage is that the compacted particles may react with each other, resulting in disintegration.

10.2. Granulation of Dry Materials with Additives Producing Chemical Reactions

Chemically reacting materials can be introduced to increase the temperature. This allows granulation to proceed with a minimum water content and with reduced amounts of steam or water.

Ammoniation of superphosphates by injection of liquid or gaseous ammonia underneath the bed or by use of aqueous solutions of ammonium nitrate or urea containing free ammonia is such a technique. The ammonia initially neutralizes free acid in the superphosphate and then reacts with the water-soluble MCP to form non-water-soluble, but citrate-soluble P_2O_5 (mainly DCP) [63–65].

Whether the decrease in water-soluble P_2O_5 content is acceptable depends on farmers' requirements. The United Kingdom and The Netherlands traditionally prefer water-soluble P_2O_5, whereas in the United States, France, and Germany "available" P_2O_5 is fully accepted as an equally valuable plant nutrient.

When nongranular MAP is present in the formulation it is often beneficial in terms of improving the granulability of the material that is to be ammoniated in the granulator. This increases the solubility of the ammonium phosphate, and heat is generated due to the chemical reaction of ammonia with MAP.

If heat released during ammoniation of the superphosphate or MAP is insufficient, sulfuric or phosphoric acid and more ammonia can be added directly to the granulator [66]. The development of the continuous ammoniator granulator by TVA has been very important in this context [65], [67]; see → Fertilizers, Chap. 5.3.2. This technique has been used very successfully in the granulation of solid urea-based NPK grades [66].

The TVA process was subsequently extended with a preneutralizer [68], [69], a reactor in which a slurry is produced by reacting ammonia or ammoniating solution with sulfuric or phosphoric acid [70]. It is used for formulations in which the heat of reaction becomes too great for release in the granulator.

The preneutralizer operates at the optimum N/P ratio of 1.4 to minimize introduction of water. Preneutralizers have recently been replaced by pipe reactors, which allow direct release into the granulator of slurry or melt with an even lower water content. An alternative is to introduce all or part of the P_2O_5 in the form of MAP powder; the crystallizing ammonium nitrate solution then provides the necessary plasticity. For solid urea-based NPK grades the pipe reactor process is a viable method.

10.3. Slurry Granulation

In slurry granulation, all or most of the ingredients enter the granulation system in slurry form (Fig. 32). See also → Fertilizers, Chap. 5.2.

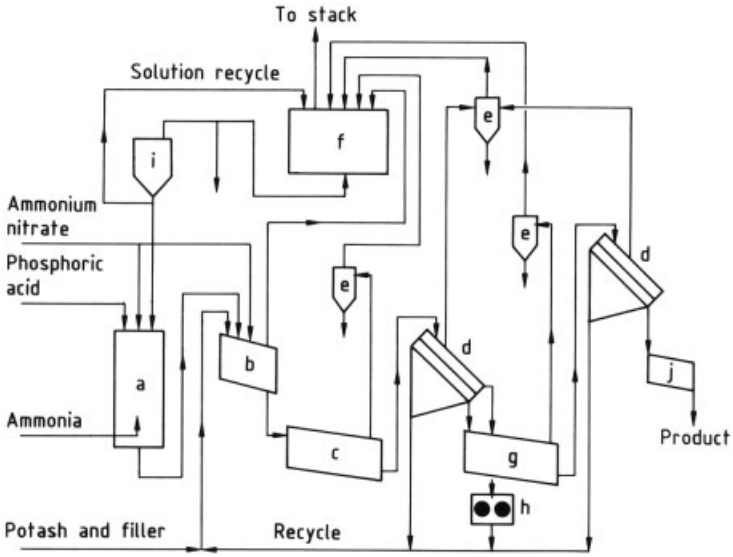

Figure 32. Slurry granulation in the production of NPK fertilizer
a) Preneutralizer; b) Blunger; c) Dryer; d) Screen; e) Cyclone; f) Scrubber; g) Cooler; h) Crusher; i) Cyclone; j) Coating drum

The slurry is usually manufactured by neutralization of nitric, phosphoric, or sulfuric acid–or mixtures of these acids–with ammonia. Potash may be added to the neutralized slurry before granulation or in the granulator. If potassium chloride is required, premixing with slurries containing ammonium nitrate is beneficial for product quality because the reaction

$$KCl + NH_4NO_3 \longrightarrow KNO_3 + NH_4Cl$$

is virtually complete before the product goes to storage.

The granulator may be of the rotary drum or pugmill (blunger) type (→ Fertilizers, Chap. 5.3.), and recycled material (fines, broken oversize, and part of the on-size) is added in a proportion that ensures good granulation efficiency.

If the preneutralizer is used only to produce ammonium phosphate, the slurry must contain 15 – 20 % water so that it remains sufficiently fluid even when operating at the optimum N/P molar ratio. This creates conditions of excessive fluidity in the granulator, particularly when high-concentration grades such as 17-17-17 or 23-23-0 are produced. Large amounts of recycle are therefore required to control granulation, particularly when ammonium nitrate is added in the form of a concentrated solution rather than a solid.

As in the manufacture of granular DAP or MAP (see Section 9.2.2), processes have been modified to reduce the recycle ratio and thus save energy and boost the capacity of existing units or save capital when constructing new units. Employment of a cooled recycle by positioning the screening units downstream from the cooler allows granulation to occur at lower temperature and thus higher water content (Fig. 31) [71].

In a plant that produces grades containing both ammonium nitrate and ammonium phosphate, the phosphoric and nitric acids can be neutralized separately or together (coneutralization). Although separately neutralized ammonium nitrate can be concentrated easily (e.g., up to 95 –98 %), the ammonium phosphate slurry still contains 15 –20 wt % water. However, addition of ammonium nitrate to the ammonium phosphate increases its fluidity which allows the water content of the slurry to be decreased [72], [73]. This is the reason why nitric and phosphoric acids are often coneutralized. The coneutralized liquor is pumped to the granulator, if necessary, after further evaporation.

Other techniques to reduce the recycle ratio include the use of a pressure reactor to increase the boiling point and thus increase the solubil-

ity of ammonium phosphate. Fluid slurries are therefore obtained at a low water content [74], [75].

A pipe reactor can also be used that operates at a higher temperature and a lower water content than the preneutralizer. The pipe discharges directly into the granulator. If the ventilation capacity is sufficient to remove all the process steam, the recyle requirement decreases [76]. Pipe reactors have been fitted in blungers [77], rotary drum granulators [76], [78], [79], dryers, or both in rotary drum granulators and dryers [80], [81].

The use of a relatively high recycle in a slurry granulation plant is not always a disadvantage [82], [83]. Granule formation takes place largely by layering. The water is on the outside of the granule which facilitates drying. Because only a small portion of the granules can be exported, the screening operation is not very critical and the size-distribution quality is good. Finally, plants operating with a high recycle ratio and where granulation is "water-balance controlled" are normally more stable than low-recycle granulation plants whose capacity is determined by granulation efficiency, which depends on fine tuning by the operators.

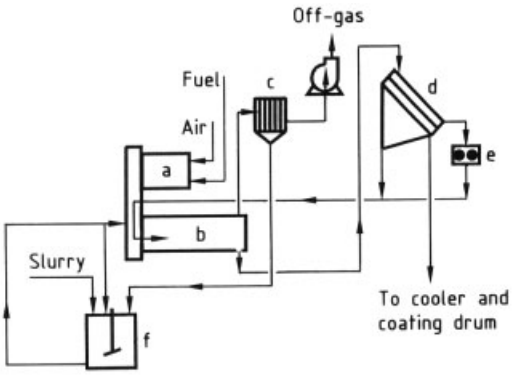

Figure 33. C & I – Girdler spherodizer granulation process for production of compound fertilizers
a) Air heater; b) Spherodizer granulator; c) Dust removal; d) Screen; e) Crusher; f) Surge tank

Slurry granulation is also carried out in spherodizers (Fig. 33) – spray-drum granulators in which granulation and drying are combined (→ Fertilizers, Chap. 5.2., → Fertilizers, Chap. 5.3.) [84].

Neutralized slurry containing all the potash is sprayed against a falling curtain of recycle solids produced by flights in a rotating drum. Hot air is blown in cocurrently with the slurry spray. Granulation takes place by layering. Granulation efficiency is high and the need for recycle is relatively low. The granules are well rounded, hard, and generally considered to be of excellent quality.

Drying during granulation is also a feature of the Scottish Agricultural Industries double-drum granulation process with a very large internal recycle (→ Fertilizers) [85]. Granulation and drying of compound fertilizers in fluidized-bed and spouted-bed equipment are also reported [86].

10.4. Melt Granulation

See also → Fertilizers.

Mixtures of ammonium nitrate and ammonium phosphate maintain their fluidity at very low moisture content provided the temperature is sufficiently high (Figs. 9 and 10). The concentrated melt can therefore be allowed to solidify and dry by the heat of crystallization.

If the acid is concentrated and/or preheated, the heat of neutralization may be sufficient to drive out all the water. If not, the N – P solution can be concentrated by using external heat in an evaporator. Removal of the water before granulation is thermally more efficient than using a conventional dryer. Hydro Fertilizers operated such a melt granulation plant for several years (Fig. 34) [87].

Another way to produce melts is to use a pipe reactor operating at high temperature [88]. The pipe produces virtually anhydrous melts, which are blown directly onto a cooled recycle bed in the granulator (Fig. 35).

Normally phosphoric-acid-containing solutions mixed with nitrates are not ammoniated in pipe reactors because chlorine and organic impurities may reduce the safety limits for ammonium nitrate explosion. The practical use of pipe reactors is thus limited to processing of either pure ammonium nitrate or ammonium phosphates and sulfates.

The TVA ran a demonstration plant in which 15 – 25 % of the P_2O_5 was condensed to nonorthophosphates, mainly pyrophosphates [89].

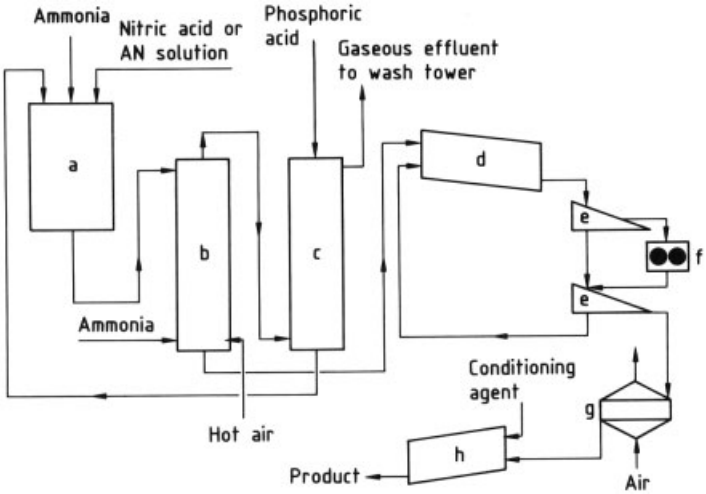

Figure 34. Hydro Fertilizers melt granulation process for the production of NPK fertilizer
a) Reactor; b) Dehydrator; c) Scrubber; d) Granulator; e) Screen; f) Mill; g) Cooler; h) Coating

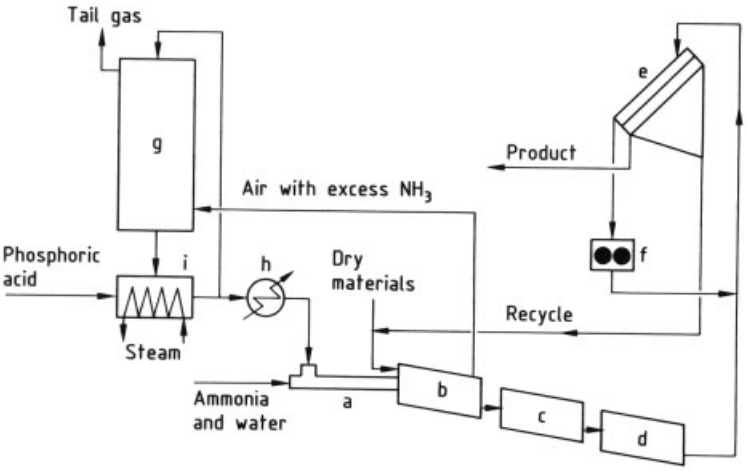

Figure 35. Pipe reactor – drum granulator process
a) Pipe reactor; b) Drum granulator; c) Dryer; d) Cooler; e) Screens; f) Crusher; g) Scrubber; h) Heater; i) Heat exchanger

Potassium salts and other solid raw materials can be added to the melts to produce NPK products.

Methods to form particles from melts include pan, pugmill, and rotary drum granulation, as well as prilling [90] and flaking.

In a conventional melt granulation loop, the recycle requirement is determined not by the water balance, but by the need to remove sufficient heat. The recycle ratio is so low that the plant will be "granulation-efficiency" controlled.

11. Compound Fertilizers by the Nitro Route

Nitrophosphates are nitrogen- and phosphate-containing fertilizers that are produced by digestion of phosphate rock with nitric acid. Examples of known production processes are the mixed acid and Odda processes.

The nitrophosphate process was invented by Erling B. Johnson of Odda, Norway [91] to avoid the diluting effect of the sulfate ion in superphosphate. The process had been developed to the commercial stage by the 1930s, and

Norsk Hydro introduced its own technology in 1938. However, not until the sulfur shortage in the early 1950s did nitrophosphates become of interest, and a number of variants of the process were developed [92][93][94].

Approximately 10 % of the world's phosphate fertilizers are produced by the nitrophosphate route.

11.1. Chemistry

The overall acidulation reaction is given by

$$Ca_5F(PO_4)_3 + 10\,HNO_3 \longrightarrow 5\,Ca(NO_3)_2 + 3\,H_3PO_4 + HF$$

The reaction is normally carried out with an excess of 10 – 20 % nitric acid [95], [96]; acid-insoluble materials in the phosphate rock may be partly removed.

Most nitrophosphate processes include removal of calcium nitrate from the solution to increase the nutrient content of the product, to avoid precipitation of water-insoluble DCP during neutralization and to obtain a product that meets thermal stability requirements for transportation.

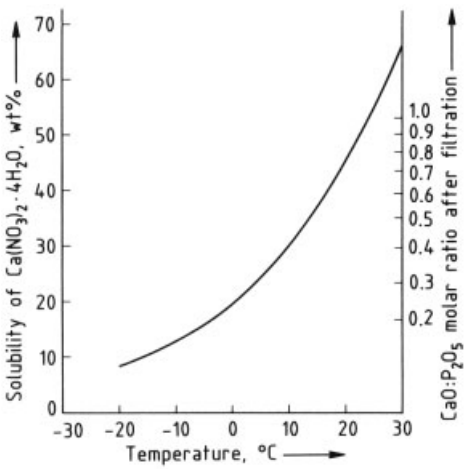

Figure 36. Crystallization points for calcium nitrate in the digestion liquor (digestion of phosphate rock with nitric acid)

About 60 % of the calcium nitrate crystallizes at 20 °C and 80 – 85 % at −5 °C (Fig. 36) [97], [98]. Alternatively, calcium can be precipitated as calcium sulfate by addition of ammonium sulfate to the solution. If the reason for using the nitrophosphate process is to remain independent of a sulfur source, sulfur can be recycled in the Merseburg process:

$$CaSO_4 + 2\,NH_3 + CO_2 + H_2O \longrightarrow CaCO_3 + (NH_4)_2SO_4$$

Calcium can be precipitated as calcium carbonate with carbon dioxide:

$$Ca(NO_3)_2 + 2\,NH_3 + CO_2 + H_2O \longrightarrow CaCO_3 + 2\,NH_4NO_3$$

Care must be taken to suppress formation of tricalcium phosphate, which is citrate insoluble:

$$2\,CaHPO_4 + CaCO_3 \longrightarrow Ca_3(PO_4)_2 + H_2O + CO_2$$

Another nitrophosphate process is the ion-exchange method in which dissolved calcium ions are replaced by potassium ions from a potassium-loaded ion-exchange resin [99]. An advantage of this process is that the resulting phosphoric acid – potassium nitrate solution can be used for making chloride-free NPK fertilizers.

11.2. Product Specification

High-grade fertilizers containing 50 – 90 % water-soluble P_2O_5 are produced, depending on the amount of calcium removed. Citrate-soluble P_2O_5 accounts for 99 – 100 % of the total phosphorus. The N/P_2O_5 ratio obtained by the normal nitrophosphate route ranges from about 0.7 to 5.0 in actual plant operation. Above a ratio of 2, ammonium nitrate from external sources must be added. If no ammonium sulfate is added, ca. 55 – 60 % of the nitrogen is in the form of ammonium and 40 – 45 % in the form of nitrate.

Today's nitrophosphate products are of a quality comparable to those based on the sulfur route [100].

11.3. Nitrophosphate Process with Calcium Nitrate Crystallization (Hydro)

Norsk Hydro has carried out extensive research and development on nitrophosphate process technology. The principles of the Odda process have been retained, i.e., phosphate rock digestion and calcium nitrate crystallization.

The process consists of three main parts:

1) Mother liquor production (phosphoric acid with nitric acid, Fig. 37)
2) Production of prilled or granulated products
3) Conversion of calcium nitrate to ammonium nitrate and calcium carbonate

The calcium nitrate can also be processed to granulated calcium nitrate fertilizers.

Digestion. Phosphate rock is dissolved in an excess of 60 % nitric acid in two stirred-tank reactors in series (Fig. 37). Because of the heat of reaction, the digestion temperature increases to 50 – 70 °C.

The main reactions for phosphate rock containing silica, carbonates, and metal oxides are:

$$Ca_5F(PO_4)_3 + 10\,HNO_3 \longrightarrow 3\,H_3PO_4 + 5\,Ca(NO_3)_2 + HF$$

$$6\,HF + SiO_2 \xrightarrow{-2\,H_2O} H_2SiF_6 \rightleftharpoons SiF_4 + 2\,HF$$

$$YCO_3 + 2\,HNO_3 \longrightarrow Y(NO_3)_2 + CO_2 + H_2O$$
$$Y = Ca \text{ or } Mg$$

$$X_2O_3 + 6\,HNO_3 \longrightarrow 2\,X(NO_3)_3 + 3\,H_2O$$
$$X = Al \text{ or } Fe$$

Acidulation of phosphate rock liberates small amounts of nitrogen oxides (NO_x), water vapor, hydrogen fluoride, and silicon tetrafluoride. These gases are vented to a scrubbing system.

The amount of NO_x formed is highly dependent on the amount of iron, sufides, organic material, and other reducing agents in the rock. Urea is added to the first digester to reduce the amount of NO_x formed. No urea is found in the final product [101]. An antifoaming agent is also added to the first digester.

Some phosphate rocks contain large amounts of acid-insoluble material, which can be removed with sand traps or hydrocyclone (Fig. 37, d) systems. The insolubles are normally washed in a drum (e) and filtered on a belt before disposal.

Crystallization is accomplished by cooling. The digestion liquor is cooled and forms $Ca(NO_3)_2 \cdot 4\,H_2O$ crystals (Fig. 37, f). Crystallization begins at ca. 23 °C, and solubility decreases rapidly with decreasing temperature. The end temperature of the crystal slurry is adjusted (normally to 0 to −5 °C) to obtain the desired composition (CaO/P_2O_5 ratio) in the final product.

Crystallization is carried out batchwise to avoid buildup of deposits on the cooling coils. Cooling is performed by circulating cold ammonia – water solution through coils. Some of the heat removed is used to vaporize the ammonia required in the process. The rest of the heat load is removed by mechanical refrigeration or by cold water.

Calcium Nitrate Filtration. The crystal slurry enters a feed tank and is fed continuously to filters (Fig. 37, g), where $Ca(NO_3)_2 \cdot 4\,H_2O$ crystals are separated from the mother liquor. In a second filter stage, the crystals are washed with nitric acid and water. The wash acid is returned to the first digester.

The mother liquor is pumped to the neutralizers. The calcium nitrate crystals are flushed to a melting tank (i) before being processed in the conversion plant or to calcium nitrate fertilizers.

Neutralization. The mother liquor contains phosphoric acid, nitric acid, some hydrofluoric acid, dissolved calcium and magnesium nitrates, small amounts of dissolved impurities (e.g., iron, aluminum, silicon), and suspended insoluble particles (e.g., quartz). Neutralization takes place in one or two reactors by addition of gaseous ammonia under strict pH control (Fig. 38, a). Nitric acid or ammonium nitrate solution is added to adjust the N/P_2O_5 ratio in the final product.

When the acids are neutralized with ammonia, calcium and most of the dissolved impurities precipitate as fluorides and phosphates:

$$Ca(NO_3)_2 + 2\,HF + 2\,NH_3 \longrightarrow 2\,NH_4NO_3 + CaF_2$$

$$Ca(NO_3)_2 + H_3PO_4 + 2\,NH_3 \longrightarrow 2\,NH_4NO_3 + CaHPO_4$$

Ammonium nitrate and ammonium phosphates remain in solution. Ammonium nitrate solution is added to give the desired N/P_2O_5 ratio.

Due to the heat of reaction, water and ammonia are evolved during operation. Ammonia is normally recovered as ammonium nitrate in the off-gas treatment system.

Nitrophosphate Evaporation. The nitrophosphate liquor from the neutralizers is concentrated in two stages in evaporators under reduced pressure (Fig. 38, b). The water content in the concentrated slurry is ca. 2.5 % for granulation and 0.5 % for prilling.

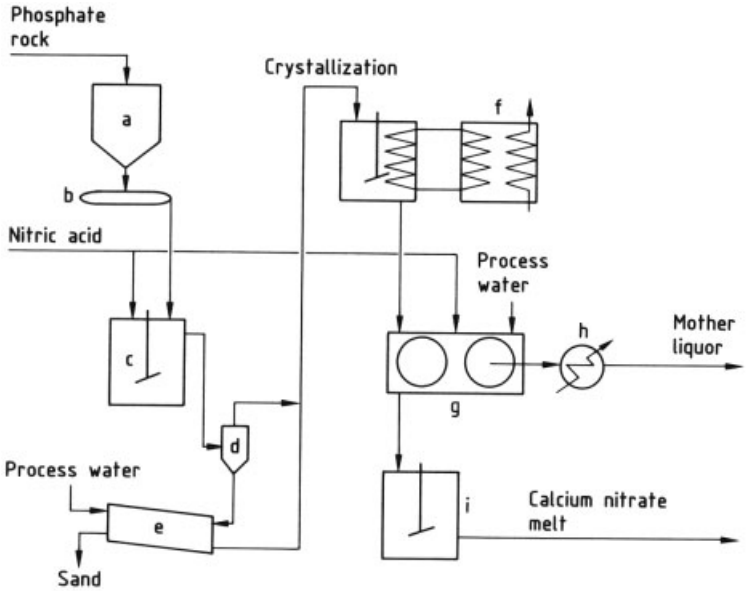

Figure 37. Nitrophosphate mother liquor (NP acid) production
a) Rock bin; b) Belt scale; c) Digestion; d) Hydrocyclone; e) Washing; f) Cooling system; g) Calcium nitrate filtration; h) Heater; i) Melting tank

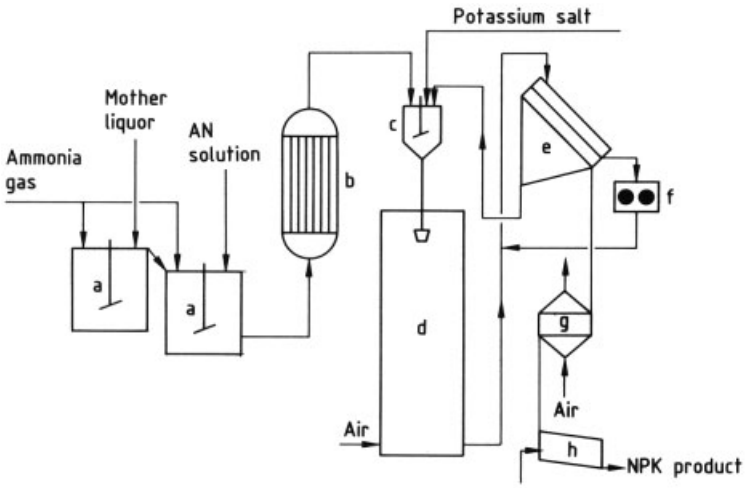

Figure 38. Nitrophosphate neutralization, evaporation, and prilling
a) Neutralization; b) Nitrophosphate evaporator; c) Mixing vessel; d) Prilling tower; e) Screen; f) Mill; g) Cooler; h) Coating drum

During evaporation, some ammonia escapes with the water vapor. Ammonia is recovered by scrubbing with acidic ammonium nitrate solution or by indirect condensation and subsequent stripping from the condensate. Prior to ammonia recovery, fluorine is removed from the vapor by scrubbing to avoid plugging and corrosion.

Mixing and Prilling/Granulation. The concentrated nitrophosphate liquor flows by gravity to the mixing – prilling section (Fig. 38, c, d). For the production of NPK, concentrated nitrophosphate liquor is mixed with potassium chloride (or sulfate) in a high-speed mixer. Recycled material from the dry handling

system containing dust, crushed oversize particles, and fines is also added to the mixer.

In the prilling process [102–104] the mixer overflows to a rotating prill bucket from which the slurry is sprayed into the prill tower. Fans at the top of the tower cause cold air to flow countercurrent to the droplets and result in solidification. The solid prills fall onto a rotating tower bottom and are scraped toward the edge by a scraper. The percentage of off-size normally varies between 10 and 20 %.

The granulation process is similar to that presented for the sulfur route (see Figs. 31, 32, 33, 34).

After screening (e), coarse particles are crushed (f) and recycled, together with fine particles, to the mixer. The product flows to a rotating drum or a fluidized-bed cooler (g) to adjust product temperature before being sent to storage. Solid and liquid anticaking agents are added to the product in the coating drum (h).

Salt Preparation. The primary steps of salt preparation are weighing, drying, and screening. The main salts added to the process are kieserite, dolomite, potassium chloride, and potassium sulfate. These fairly pure salts contain only small quantities of other metals or insolubles.

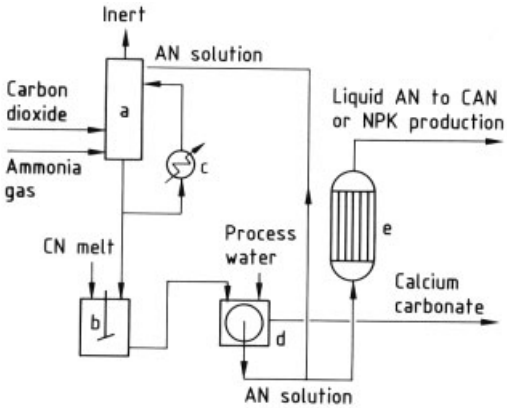

Figure 39. Conversion of calcium nitrate
a) Absorption; b) Conversion; c) Cooler; d) Filtration; e) Evaporation
AN = ammonium nitrate; CAN = calcium ammonium nitrate; CN = calcium nitrate

Calcium Nitrate Conversion. The calcium nitrate byproduct can be either used for the production of calcium nitrate fertilizers or converted into ammonium nitrate solution and calcium carbonate [105–108].

Calcium nitrate crystals from the filters are melted before being reacted with ammonium carbonate solution. The slurry is filtered on a vacuum filter, and the filter cake is washed with water.

In the Norsk Hydro process (Fig. 39), absorption of ammonia and carbon dioxide, and subsequent precipitation of calcium carbonate, take place in separate units. Ammonium carbonate is formed by reaction of gaseous ammonia and carbon dioxide in a recirculated stream of 60 % ammonium nitrate solution.

Because of the heat of reaction, external coolers (c) are installed to control the absorber temperature and to reach the desired temperature for the ammonium carbonate solution entering the conversion reactor.

Precipitation of calcium carbonate takes place in the conversion reactor (b):

$$Ca(NO_3)_2 + (NH_4)_2CO_3 \longrightarrow 2\,NH_4NO_3 + CaCO_3$$

To ensure complete precipitation the reactor is operated with a slight excess of carbon dioxide. Because the heat of reaction is relatively low, no cooling is required.

The reactor slurry can be filtered on a rotary vacuum filter or a belt filter (d); the filter cake should be washed with hot water. The calcium carbonate is used both for technical and agricultural purposes. Another alternative is to make calcium ammonium nitrate fertilizer (28 % N).

Because the ammonium nitrate solution contains a slight excess of ammonia, it is neutralized with nitric acid prior to evaporation (e).

11.4. Nitro Process with Calcium Nitrate Crystallization (BASF)

The Odda process has been used by BASF since 1953 at Ludwigshafen and until recently at its subsidiary in Köln.

The company designed and built a new 700 000-t/a plant at Antwerp, which went into operation in 1985; BASF now offers the process for license. Contracts have been awarded to Agrolinz (Austria) and Bharuch (India).

The principles of the BASF process (Fig. 40) are similar to those of the Norsk Hydro process; the following description focuses on the

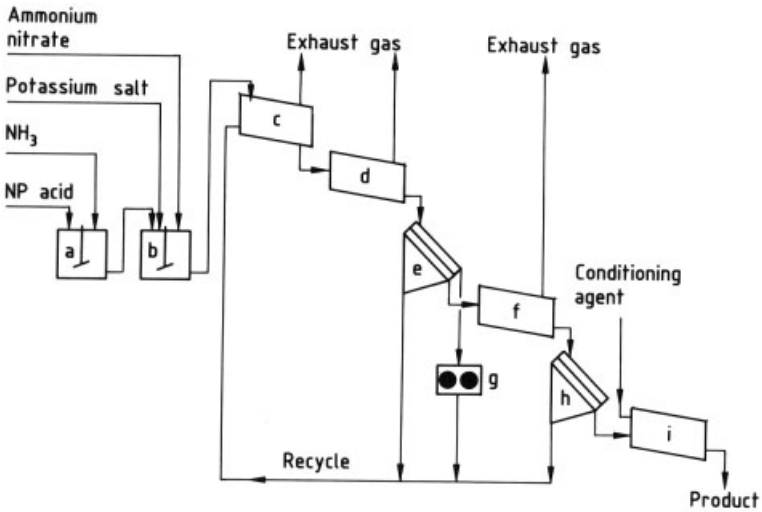

Figure 40. The BASF granulation process for production of NPK fertilizers
a) Neutralization; b) NPK slurry; c) Granulation; d) Dryer; e) Hot screen; f) Cooling; g) Crusher; h) Cold screen; i) Finishing

main features specific to the BASF process [98], [109].

Digestion is carried out in a cascade of reactors. The temperature must be kept above 68 °C to ensure complete acidulation and below 72 °C to prevent corrosion. Residence time is 2 h. Foaming is usually controlled by specially designed stirrers, but antifoaming agents may be required. Only rock particles larger than 4 mm–usually <10 % of the total–require crushing.

The process employs sedimentary rocks containing 68 – 75 % tricalcium phosphate in the form of fluorapatite.

Crystallization and Separation of Calcium Nitrate. Crystallization is carried out in tanks with integral cooling coils. The cooling medium is chilled brine. In more recent systems the crystallizers are operated batchwise, but the cooling coils are connected in series.

The following conditions should be satisfied to achieve optimum crystal separation:

1) High separation efficiency with regard to $Ca(NO_3)_2 \cdot 4\,H_2O$ and nitrophosphate solution
2) Low sensitivity to changing process conditions
3) Minimal costs for operation, maintenance, and repair

The shear centrifuges used in older plants operate in conjunction with downstream hydrocyclones and only partially meet these requirements. The centrifuges are subjected to a great deal of wear and are extremely sensitive to the penetration of sand from the digestion stages. The screen slots must be wide; they allow large amounts of salt to slip through, which cannot be recovered adequately with downstream hydrocyclones. A sharp boundary between the separation zone and the washing zone cannot be achieved. Belt filters are therefore used in new plants and meet the above-mentioned requirements more satisfactorily.

Neutralization and Evaporation. The N/P_2O_5 ratio of the nitrophosphate solution from the mother liquor is adjusted with concentrated ammonium nitrate obtained by the conversion of separated calcium nitrate. The slurry is then ammoniated in two stages at 120 – 130 °C. Overammoniation, which would result in ammonia loss and cause reversion of phosphate to a citrate-insoluble form, is prevented by an automatic pH monitor.

If necessary, the slurry is evaporated at atmospheric pressure in heated, single-stage evaporators with forced circulation.

Granulation. Final correction of the N/P ratio and incorporation of the potassium salt are

carried out in mixers. The slurry is then sprayed into a drum granulator (spherodizer) or a pugmill granulator. The spherodizer type of drum operates with slurries containing 15 – 20 % water, and only a small amount of material is recycled. In standard drum or pugmill units a water content <8 % is required, and the granulation and drying units are separate. Production also involves screening, crushing, cooling, and coating units.

Calcium Nitrate Conversion. As in the Norsk Hydro process (see page 34), calcium nitrate conversion is carried out by treatment with ammonium carbonate. An ammonium carbonate solution is made by absorbing carbon dioxide and ammonia into ammonium nitrate solution. A belt filter is again used for separation. The calcium carbonate either can be used for making lime ammonium nitrate fertilizer or can be used in the cement industry. The ammonium nitrate solution passes through pressure filters to remove any calcium carbonate present due to malfunction upstream; it is then concentrated in steam-heated evaporators.

11.5. Nitro Process with Ion Exchange (Kemira Superfos)

The ion-exchange process (Fig. 41) permits the production of high-grade, chloride-free NPK fertilizers based on the cheapest possible raw materials [99].

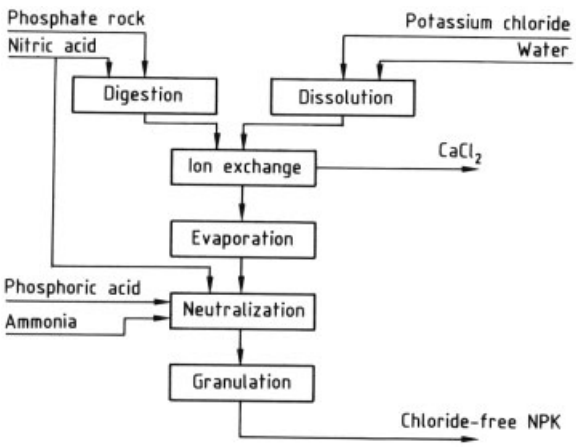

Figure 41. Kemira Superfos ion-exchange NPK process

Acidulation of calcium phosphate with nitric acid gives a solution containing calcium ions, phosphoric acid, and nitrate ions:

$$Ca_5F(PO_4)_3 + 10\,HNO_3 \longrightarrow 5\,Ca^{2+} + 3\,H_3PO_4 + 10\,NO_3^- + HF$$

If the acidulated solution is exposed to a potassium-loaded cation exchanger, calcium ions are exchanged for potassium ions. The cation exchanger is regenerated with potassium chloride solution. The product solution contains only the valuable constituents potassium, phosphate, and nitrate; the byproduct solution from regeneration contains the unwanted calcium and chloride, which are removed.

Production. The product solution is treated in the usual way to give granular NPK fertilizers.

The process was developed in the late 1960s, and an industrial plant began operating at the end of 1971 in Fredericia, Denmark.

Conventional chloride-free NPK fertilizers based on potassium sulfate contain ammonium nitrate, ammonium phosphate, and potassium sulfate. In NPK fertilizers based on ion exchange, the potassium sulfate is replaced by potassium nitrate.

Products based on ion exchange can be made with a higher total content of nitrogen, phosphorus, and potassium nutrients because they do not contain "ballast" sulfate. A typical N-P_2O_5-K_2O grade is 17-17-17; the corresponding grade in potassium-sulfate-based processes is 15-15-15.

The presence of potassium nitrate instead of chloride or sulfate has a favorable effect on physical properties. Ammonium chloride formation, which often causes caking problems, does not occur. The presence of nitrate instead of chloride or sulfate facilitates production of hard, uniformly sized granules with good storage properties in a spherodizer.

11.6. Nitro Process with Sulfate Recycle (DSM)

Phosphate rock is acidulated with nitric acid at 65 °C (Fig. 42, c). Phosphoric acid may also be added. Calcium nitrate in the slurry is then reacted with ammonium sulfate at 55 °C (d) to form calcium sulfate dihydrate [110].

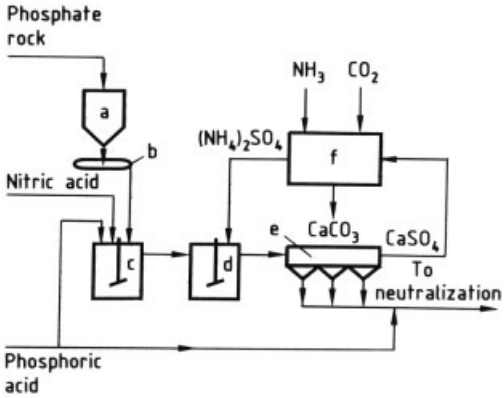

Figure 42. The DSM nitro process for production of NPK fertilizer with sulfate recycle
a) Rock bin; b) Belt weigher; c) Digestion; d) Reactor; e) Filter; f) Conversion plant (see Fig. 39)

The slurry is filtered (e) and the gypsum filter cake is fed to the ammonium sulfate recovery section where 52 – 53 % ammonium carbonate solution is used to produce a slurry of ammonium sulfate solution and solid calcium carbonate. A 40 % ammonium sulfate solution is obtained for recycle, after filtration of the calcium carbonate (Fig. 39). The mother liquor from the filter is mixed with phosphoric acid to increase the P_2O_5 water solubility of the final product. After neutralization the remaining moisture is removed by steam evaporators before the product is mixed with other nutrients and prilled or granulated.

A feature of this process is the absence of straight nitrogen fertilizer coproduct, although low N/P_2O_5 ratios are obtained only by the addition of other P_2O_5 materials (phosphoric acid, triple superphosphate, or MAP).

The use of a mixture of acids to produce nitrophosphates was developed to eliminate calcium nitrate as a final coproduct. Some of the calcium nitrate formed during acidulation of the phosphate rock with nitric acid remains in the slurry after filtration and ends up as DCP in the final product. To increase the P_2O_5 water solubility above 30 %, large amounts of phosphoric acid must be added.

11.7. Effluent Control of Nitrophosphate Process

Gaseous emissions from the nitrophosphate process are especially low in fluorine; <1 % of the fluorine in the rock evolves during digestion. The concentration of nitrous gases can be reduced by addition of urea [101]. The NO_x concentration of the stack gas is <150 ppm. Furthermore, ammonia-containing effluents can be collected and treated before recycling to the process as ammonium nitrate or gaseous ammonia.

Emissions from the dry section of the plant handling solid materials are treated in cyclone or filter systems. For granulation units, treatment similar to that described for the sulfur route and MAP production is required (Chaps. 9 and 10). The prilled product consists of even particles with a smooth surface; dust generation from a prilling process is therefore considerably lower than in a granulation plant. The dust collected in these systems is recycled to the process.

Dust emission from the prilling tower is very low in nitrophosphate production, compared to the production of straight fertilizers (e.g., urea or ammonium nitrate). Thus treatment systems are not needed for the cooling air.

12. Other Straight Phosphate Fertilizers

12.1. Phosphate Rock for Direct Application

The agronomic effect of directly applied phosphate rock is assumed to be related to the degree of carbonate substitution for phosphate within the apatite structure. Carbonate fluorapatite is less resistant against weathering and may thus release P_2O_5 into the soil [111]. These rocks are often called soft phosphate rocks.

Only sedimentary rocks are suitable for direct application. Rock from North Carolina (30.1 wt % P_2O_5) and Tunisia (28.7 wt % P_2O_5) is most reactive (86.3 and 74.6 % P, respectively, extracted by 2 % formic acid) [112], [113].

World use of rock for direct application (excluding China) in 1989 was ca. 1.2×10^6 t P_2O_5 (i.e., about 3.5 % of total phosphate use). The main users worldwide are the former Soviet Union, France, Malaysia, and Sri Lanka.

Application. Direct application of reactive rock is most useful for gradually building up

the phosphate status of an acid soil or for maintaining phosphate levels in soil that already has a high status.

Reactive rocks can be used to provide an "insurance" for crops that are not particularly responsive to phosphate (e.g., perennial grazed grassland, established tree crops, and paddy rice). Crops that have a high calcium uptake (e.g., legumes) can often extract more phosphate from rock sources. Rock is least satisfactory for small-seeded annual crops or crops grown at low soil temperature.

Phosphate rock is not readily soluble in soil and can therefore only be used to build up soil reserves. However, the slow dissolution means that phosphate ions are still being released several years after application [114]. The best chemical test of the potential dissolution rate of phosphate rock is provided by extraction with 2 % formic acid.

Dissolution is aided if the phosphate rock is "soft" or if it is finely ground and applied as a powder.

Dissolution also improves if the soil is moist and acidic (pH<5.5), but the proportion of phosphate taken up by a crop decreases with increasing soil acidity [115]. Models of dissolution rate in soil are described in [116], [117].

Production. Run-of-mine rock is ground, screened, and applied directly to soil without further treatment. The rock must be ground to a particle size of 0.15 – 0.075 mm [118]. Handling problems associated with this finely ground, dusty material can be alleviated by granulation.

The minigranulation process developed by IFDC compacts finely ground phosphate rock into spherical granules by use of a binder [119]. Binders include soluble salts, mineral acids, and organic materials. Granules made with a soluble salt binder disintegrate in soil and are almost as effective as finely ground rock.

12.2. Partially Acidulated Phosphate Rock [118]

Partially acidulated phosphate rock (PAPR) consists of a mixture of immediately available MCP and gypsum (calcium sulfate). The amount of MCP present is determined by the degree of acidulation. Typically, PAPR is produced by using 50 % of the acid required for full conversion to superphosphate.

Very little PAPR is in commercial production. The main use is in New Zealand where ca. $(75-100)\times10^3$ t of P_2O_5 are produced in PAPR form.

Application. A PAPR mixture has potential agronomic advantages because it combines the immediate effect of water-soluble MCP with the prolonged release of the rock. In a responsive situation, the water-soluble fraction is most beneficial if the fertilizer is granulated and placed near the crop roots, but such placement reduces the value of the rock component.

The agronomic value of PAPR is affected by the "softness" of the original rock as well as by soil and crop factors that influence the agronomic value of directly applied rock. This value is often surprisingly high under laboratory conditions [120].

The PAPR is most valuable in moist, acid soil with moderate phosphate status where the crop is not particularly responsive. The crop should then obtain its immediate phosphate needs from MCP, whereas the rock fraction maintains or builds up the general phosphate status over time.

Partial acidulation improves the agronomic performance of almost all rocks but probably has economic advantages over full acidulation (i.e., production of superphosphates) only in areas where crops show little or no immediate response and in soils where the rock component dissolves rapidly enough to maintain the soil at a satisfactory phosphate status.

Production. Sulfuric and phosphoric acids are most commonly used for partial acidulation. The former is generally preferred for economic reasons. Two processes have been evaluated for the production of sulfuric-acid-based PAPR [121].

The *run-of-pile (ROP) process* is simpler and may be operated in one or two stages. Ground phosphate rock, sulfuric acid, and water (if required) are fed continuously to a mixer (usually of the pugmill type). Retention time in the mixer is 30 – 60 s. Continuous operation is preferable to batch operation to avoid poor product handling as well as excessive agglomeration and caking. Other materials can be added to the

mixer to improve acidulation or provide additional nutrients.

Depending on the characteristics of the rock and the degree of acidulation, the product can be transferred directly from the mixer to storage without curing or denning. It is typically stored for two weeks before being reclaimed, crushed to pass through a 4-mm screen, and prepared for dispatch.

In the *single-step acidulation – granulation process* the feed includes recycled material (product screen undersize and some product). Depending on the rock and the operating conditions, the recycle to product ratio in the feed is ca. 2. Retention time in the granulator is usually 5 – 8 min.

The moist, plastic product leaving the granulator is fed to a rotary drum dryer. After being dried, the granules are screened in a closed circuit with a crusher to obtain a product size of 1 – 4 mm.

Table 3. Composition of basic and BOF slag fertilizers

Composition or property	Basic slag (Thomas phosphate), wt %	BOF slag (Thomas kalk), wt %
Typical values		
P_2O_5 (citric acid soluble)	13–15	4–8
$CaO + MgO$	45–50	45–55
SiO_2	8–12	ca. 12
$Fe(FeO + Fe_2O_3)$	8–12	12–18
Mn, Al, and trace elements	2–4	3–6
German regulation requirements		
Minimum citric acid soluble P_2O_5	10	3
Minimum CaO		35
<0.16 mm	>75	
<0.63 mm	>96	
<0.315 mm		>80
<1.0 mm		>97
Color	brownish gray	
Density	ca. 3.2 g/cm^3	
Bulk density	ca. 1.6 g/cm^3	
Tapped density	ca. 2.1 g/cm^3	

12.3. Basic and BOF Slag Fertilizers

The origin of basic fertilizer slag and basic oxygen furnace (BOF) slag is apatite from magmatic iron ores. The slag is separated during steel production by addition of lime. Excess lime and silica (sand) can be added to increase the fertilizer value and give porous crystals of calcium silicophosphate, $Ca_5P_2SiO_{12}$ (silicocarnotite). Slag composition varies with origin and process; BOF slag has a lower P_2O_5 content. Typical values and German fertilizer regulation requirements are given in Table 3.

Since the mid-1960s the production of phosphorus-containing fertilizer slag has declined because the steel industry requires high-quality ores with a low phosphorus content. Today, only iron ores from Kiruna (Sweden) and Lorraine (France) yield phosphorus-containing fertilizer slag. Annual EC production of basic slag is ca. 220 000 t of P_2O_5 (1.6×10^6 t of slag) and of BOF slag ca. 12 000 t of P_2O_5 (220 000 t of slag).

Application. The fertilizing effect of slag increases with its silica content. Many years ago, extra sand was sometimes added to liquid slag either during the slagging process or while the slag stream was being poured off. At present, nearly all basic slag fertilizers have a high citric acid solubility (>80 %).

Production. Basic slag is produced as a byproduct in the basic Bessemer process for melting phosphorus-containing iron ores.

In all steel-making processes based on blowing with air or oxygen, oxidation of phosphorus-containing iron in the converter produces both steel and phosphorus-containing slag. All elements that are initially reduced in the blast furnace along with the iron oxide are subsequently oxidized in the converter at high temperature. They accumulate in the slag to form calcium silicophosphate. At the end of the blowing process, liquid slag floats on the molten steel surface. This slag consists mainly of a lime – phosphate – silicate melt, as well as certain amounts of iron and manganese oxides.

Addition of phosphate rock to the converter as a replacement for lime has frequently been recommended to increase the P_2O_5 content. Such addition cannot be used if the phosphate contains fluorine because this severely reduces the solubility of the basic slag in 2 % citric acid due to apatite formation. Use of fluorine-free phosphates gives the desired increase in the citric-acid-soluble phosphate fraction of the slag [122].

In the basic Bessemer process for oxidizing carbon, phosphorus, silicon, etc., oxygen is introduced in the form of air blown through the bottom of the converter. The use of pure oxygen instead of air has led to a series of other steel production processes (→ Steel). A basic slag containing bound P_2O_5 is formed by blowing the oxygen through a lance from above (Linz – Donawitz Arbed Centre process, LDAC) or through nozzles on the floor of the converter (Oxygen Bottom Maximilianshütte process, OBM). The oxygen blast can be interrupted before complete decarburization of the melt (ca. 0.7 %) occurs; most of the "first slag" is then poured off. The "second slag", formed after final blowing of the melt, remains in the converter and is reprocessed with the next melt (two-slag process). The first slag resembles basic Bessemer slag, but has a lower iron and a higher P_2O_5 content (5 – 7 % Fe, 18 – 22 % P_2O_5). It is marketed as basic slag [123]. In the one-slag process the charge is blown until dephosphorization is complete, and mixed slag with a lower phosphate content is produced.

Separation of the Slag. The converter containing liquid slag floating on the molten steel is tilted. Slag flows into a trolley where it solidifies to a solid block, which is then taken to the slag heap where it cools. When the slag solidifies, considerable segregation occurs in the interior of the block, resulting in virtually complete separation of lime-rich phosphates, and iron and manganese oxides [124].

The LDAC and BOF slags are poured into flat beds for technical reasons. This allows rapid cooling and thus speeds up further processing.

The slag begins to solidify even at high temperature and thickens rapidly. Complete dissolution of the added sand is therefore difficult. Moreover, the melting point of silicophosphate in the slag increases with increasing silica content. Sand is therefore sometimes dried and mixed with ground coke before being added. Thorough mixing results in reaction of the coke with iron oxides in the slag.

Constitution. The shaded area in the phase diagram in Figure 43 shows the composition of commercial basic Bessemer slags, when the iron oxides and manganese oxide present are ignored. The line connecting $3\,CaO \cdot P_2O_5$ and $2\,CaO \cdot SiO_2$ clearly separates the lime-rich part of the ternary system from the rest because the melting points that occur on this line are all much higher than those of the neighboring phosphorus-rich compositions. The orthophosphate $3\,CaO \cdot P_2O_5$ and the orthosilicate $2\,CaO \cdot SiO_2$ are extremely stable and therefore dominate the crystal phases in the basic slag. The two phases have similar crystal structures, leading to formation of a solid solution series during crystallization. Calcium silicophosphates are almost 100 % soluble in 2 % citric acid and are present in concentrations of 60 – 80 wt % in normal basic slag or in slag produced by the LDAC or the OBM process [125].

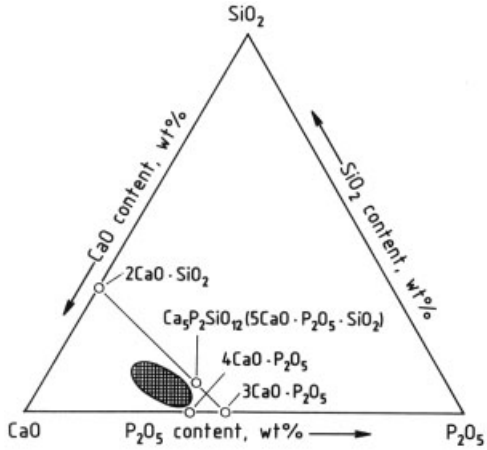

Figure 43. Phase diagram for calcium silicophosphates

The crystal phase changes that occur during cooling contribute largely to the physical properties of the slag.

Size Reduction. The blocks of slag disintegrate during the course of weeks or months. The resulting particle sizes vary greatly between extremely fine powder and larger pieces. Rapid cooling in a slag bed prevents spontaneous breakdown of the slag so that further grinding is necessary.

The slag is ground in ball mills or tube mills to a fineness of at least 75 % <0.16 mm. In the mill, efficient dust removal is necessary because prolonged inhalation of basic slag phosphate dust can damage the respiratory tract owing to the presence of silicates and silicophosphates (basic slag is covered by the "2. Verordnung zur änderung der Gefahrstoffverordnung," Bundesgesetzblatt dated April 23, 1990, Part 1).

Attempts to manufacture granulated basic slag phosphate directly on a large scale have so far not been successful for economic reasons [125].

12.4. PK Mixed Fertilizers with Basic Slag

Increasing quantities of granulated mixtures of basic slag and potash salts are used as PK mixed fertilizers with $P_2O_5 : K_2O$ ratios between 1 : 1 and 1 : 3. They are usually produced by agglomerating the powdered basic slag and potassium salts in double-shaft granulators, followed by granulation by rolling in cylindrical drums and drying with hot gases. Inorganic substances are used as granulation aids [126].

The addition of fused phosphates or DCP is legally permitted to allow the P_2O_5 content of the slag to be adjusted to maintain a constant P/K ratio [127].

13. Energy Consumption [128]

Phosphate rock is the raw material for phosphate fertilizers. Approximately 5 t of ore must be mined and beneficiated to produce 1 t of commercial phosphate rock with an average phosphorus content of 14 % (32 % P_2O_5). This process requires 0.9 GJ/t or 2.9 GJ/t of P_2O_5. If the rock is calcined, the total energy requirement is doubled. Sulfuric acid is required for the production of phosphoric acid, which is an intermediate for 80 % of the phosphate fertilizers worldwide. The energy contribution for sulfuric acid varies from an export of 8 GJ/t of P_2O_5 to a requirement of 7 GJ/t of P_2O_5 depending on the raw materials.

Production of phosphoric acid (54 % P_2O_5) from sulfuric acid and uncalcined rock requires between 10.8 GJ/t of P_2O_5 (hemihydrate) and 12.9 GJ/t of P_2O_5 (dihydrate).

Processing the acid to ammonium phosphate fertilizers requires a further 1.5 – 1.75 GJ/t of P_2O_5. Processing to triple superphosphate requires a further 2.6 – 3.5 GJ/t of P_2O_5.

Complex fertilizers based on the nitrophosphate route require a minimum of 5.5 GJ/t of P_2O_5 to process the rock to complex phosphate fertilizers.

14. Effluents from Phosphate Fertilizer Production

Disposal of gypsum-containing effluents in the sea is the main environmental concern in phosphate fertilizer production. Some companies have established land disposal and eliminated all gypsum effluents.

Emissions from phosphate fertilizer production vary considerably, normally depending on the age of the plant (Table 4).

Table 4. Emission levels from phosphate fertilizer plants

Emitted substance	Modern plants	Old plants
To air		
NH_3, kg N/t N	0.01	1–10
NO_x, kg N/t N	0.3	0.5–4
F, kg F/t P	0.01	<0.12
To water		
$NH_4^+ + NO_3^-$, kg N/t N	0.01	0.6–5
PO_4^{3-}, kg P/t P	0.15	1.0–4
F, kg F/t P	0.06	0.9

Nutrient losses during fertilizer production are <1 % of the nutrients handled in old plants and <0.04 % in modern plants. Thus, the main environmental concerns associated with fertilizers are not at the production site but on the farm (→ Fertilizers, Chap. 9.2.1.).

15. Heavy Metals in Phosphate Fertilizers

Many elements that do not appear to be necessary for plant nutrition occur in the raw materials used for fertilizer production and thus in small amounts in fertilizers.

Table 5 lists the heavy metals found in phosphate rocks and compares typical contents with those found in soil. Contents vary widely with origin and type of rock [129].

In the sulfur route, 80 – 90 % of the mercury and lead and 30 – 50 % of the cadmium end up in the gypsum waste. In the other processes, almost all the trace elements end up in the phosphate fertilizer.

Table 5 also gives the input of heavy metals to a topsoil after a century's use of phosphate

Table 5. Heavy-metal content of phosphate rock and soil [129]

Element	Average content*, mg/kg P	Range, mg/kg rock	Average content, mg/kg		Input to soil after 100 years of fertilizer use, mg/kg soil**
			Rock	Soil	
Arsenic	45	1–300	7	6	0.04
Cadmium	170	0.01–120	25	0.35	0.14
Chromium	1000	0.3–460	150	70	0.83
Cobalt	13	0.5–6	2	8	0.01
Copper	200	6–80	30	30	0.17
Lead	40	3–40	6	35	0.03
Manganese	200	6–300	30	1000	0.17
Mercury	0.2	0.01–0.10	0.03	0.06	0.0017
Molybdenum	33	1–10	5	1.2	0.03
Nickel	230	1–85	35	50	0.19
Zinc	660	3–800	100	90	0.55

* 15% P in rock. ** Topsoil (20 cm soil), soil volume weight 1.2 kg/L, annual application of 20 kg P/ha for 100 years.

fertilizer. This input is small compared with the levels naturally present in the average soil, with the notable exception of cadmium.

Cadmium is highly toxic to humans but less so to plants. Concern exists that the use of phosphate fertilizers will slowly increase the cadmium content of arable land and that within about a century this might eventually result in an unacceptably high cadmium level in agricultural produce.

The average daily cadmium intake in the European diet is 20 µg; WHO recommends a maximum daily limit of 70 µg.

Table 6. Phosphorus and cadmium content of phosphate rocks [129]

Type of rock	Phosphorus content, wt%	Cadmium content	
		mg/kg rock	mg/kg P
Magmatic origin			
Kola, Soviet Union	17.2	0.15	0.9
Palfos, Republic of South Africa	17.2	0.15	0.9
Sedimentary origin			
Bou Craa, Morocco	15.9	35	220
Togo	15.7	55	350
Youssofia, Morocco	14.6	40	274
Jordan	14.6	5	34
Texas Gulf, USA	14.4	40	278
Florida, USA	14.4	8	56
Negev, Israel	14.2	20	140
Khouribga, Morocco	14.2	16	113
Khneifiss, Syria	13.9	6	43
Gafsa, Tunisia	13.2	50	380

Fertilizers can contain cadmium if sedimentary rock phosphate is used as a raw material. Table 6 shows typical values for cadmium content in some important rock phosphates.

No commercial process is currently available for removing cadmium from phosphate fertilizers. Several companies are, however, carrying out research into the development of cadmium removal processes. The most promising processes involve separating cadmium from phosphoric acid for the sulfur route and from the mother liquor for the nitro route.

16. Safety in Storage, Handling, and Transport

Phosphate fertilizers without free phosphoric acid are regarded as nontoxic, with no dangerous properties.

Phosphate substances that are handled as fertilizers do not present a specific health risk nor do they have undesirable chemical effects. Their transport is therefore not covered by dangerous goods regulations.

Phosphatic raw materials and many fertilizer products are transported in bulk. The physical stability of the contents of the ship hold in marine transport is a risk factor during loading and voyages in rough seas. The International Maritime Organization (IMO) takes these risks into consideration and specifies values for cohesiveness and angle of repose [130].

Phosphate compounds are present in compound NP or NPK fertilizers, often with nitrates as the nitrogen source. Nitrates have some sensitivity to heat. Different grades of these fertilizers

are listed in the UN's "Recommendations on the Transport of Dangerous Goods" and in the specific codes for all transport modes.

Self-sustaining decomposition (SSD) is linked predominantly to NPK grades but may also be observed in NK grades. The thermal stability of fertilizer products is tested in a standardized procedure (the trough test) [131]. If the product exhibits SSD after the heat source is removed, the product is assigned to Class 9. If the velocity of SSD exceeds 25 cm/h, bulk transport is prohibited. For NPK grades SSD may be favored by a low water solubility of phosphorus in KCl-based, nitrate-containing commodities.

17. References

1. K. D. Jacob: *Superphosphate: Its History, Chemistry, and Manufacture,* U.S. Department of Agriculture and Tennessee Valley Authority (TVA), Washington, D.C. 1964, pp. 8 – 94.
2. J. Ands: "Thermal Phosphate", in F. T. Nielsson (ed.): *Manual of Fertilizer Processing,* Marcel Dekker, New York 1987, pp. 93 – 124.
3. K. D. Jacob: *Superphosphate: Its History, Chemistry and Manufacture,* U.S. Department of Agriculture and Tennessee Valley Authority (TVA), Washington, D.C., 1964, pp. 8 – 94.
4. E. F. Dillard, A. W. Frazies, T. C. Woodis, F. P. Achorn: "Precipitated Impurities in 18-46-0 Fertilizers Prepared from Wet-Process Phosphoric Acid," Bulletin Y-162 TVA/OACD-83/12, National Fertilizer Development Center and TVA, Muscle Shoales, Ala, 1981.
5. E. F. Dillard, A. W. Frazier: "Precipitated Impurities in MAP and Their Effect on Chemical and Physical Properties of Suspension Fertilizers," Bulletin Y-183 TVA/OACD-83/17, National Fertilizer Development Center and TVA, Muscle Shoales, Ala, 1983.
6. G. M. Lloyd, F. P. Achorn, R. M. Scheib: *Diammonium Phosphate Quality-1988,* American Institute of Chemical Engineers, Clearwater, Fla., 1988, pp. 1 – 43.
7. E. F. Dillard, J. R. Burnell, J. Gautney: "Stable Suspension Fertilizers from MAP," paper presented at the *American Chemical Society Meeting,* Washington, D.C., Aug. 26 – 31, 1990.
8. J. Ands: "Thermal Phosphate" in F. T. Nielsson (ed.): *Manual of Fertilizer Processing,* Marcel Dekker, New York 1987, pp. 93 – 124.
9. *World Fertilizer Atlas,*7th ed., British Sulphur. Corp., London 1983.
10. J. Archer: "Phosphorus," in *Crop Nutrition and Fertilizer Use,* Farming Press, Ipswich 1985, pp. 57 – 64.
11. N. C. Brady: "Supply and Availability of Phosphorus and Potassium," in *The Nature and Properties of Soils,* 8th ed., Macmillan Publ. Co., New York 1974, pp. 456 – 475.
12. K. Mengel, E. A. Kirby: "Phosphorus," in *Principles of Plant Nutrition,* 3rd ed., International Potash Institute, Bern 1982, pp. 387 – 409.
13. Ministry of Agriculture, Fisheries and Food (MAFF): *Phosphate and Potash for Rotations,* Booklet 2496, Her Majesty's Stationery Office, London 1985.
14. G. H. Aylward, T. J. V. Findlay: *SI Chemical Data,* John Wiley & Sons, Sydney 1971, pp. 122 – 123.
15. A. F. Hollemann, E. Wiberg: *Lehrbuch der anorganischen Chemie,* De Gruyter, Berlin 1976, p. 404.
16. K. S. Førland: *Kjemisk Likevekt,* Tapir, Trondheim 1974, pp. 1 – 60.
17. G. Hägg: *Kemisk Reaktionslära,* Almqvist & Wiksell, Stockholm 1969, pp. 67 – 105.
18. J. N. Butler: *Ionic Equilibrium,* Addison – Wesley, Reading 1964, pp. 206 – 260.
19. L. G. Sillen, "Graphic Presentation of Equilibrium Data," in I. M. Kolthoff, P. I. Elving, E. B. Sandell (eds.): *Treatise on Analytical Chemistry,* vol. 1, Interscience, New York 1959, pp. 277 – 291.
20. L. G. Sillen, P. W. Lange, C. O. Gabrielson: *Fysikalsk-Kemiska Räkneuppgifter,* Almqvist & Wiksell, Stockholm 1951, pp. 193 – 213.
21. L. G. Sillen, P. W. Lange, C. O. Gabrielson: *Problems in Physical Chemistry,* Prentice-Hall, New York 1952, pp. 193 – 213.
22. *Ullmann,* 4th ed., **18,** p. 332.
23. C. H. Davis, R. G. Lee: TVA-Publication, presented at the *American Chemical Society Meeting,* Atlanta, Ga, Nov. 1967.
24. P. Stokka: *Three Component Phase Diagram for MAP, DAP and Water,* Norsk Hydro Internal Report, Porsgrunn 1985.
25. P. Stokka: "Recent Developments in the Pan Granulation Process," *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Edmonton, Canada, 1988, pp. 5.1 – 5.12.

26. P. Stokka: *Three Component Phase Diagram for AN, MAP, DAP at Water Content* $\leqq 1\,\%$, Norsk Hydro Internal Report, Porsgrunn 1985.
27. P. Stokka: *Three Component Phase Diagram for AN, MAP, DAP at Water Content 10 %,* Norsk Hydro Internal Report, Porsgrunn 1985.
28. *Gmelin,* 8th ed., **Calcium (System no. 28)** 1127.
29. P. Stokka: *Viscosity of AN/MAP Melt with KCl,* Norsk Hydro Internal Report, Porsgrunn 1986.
30. R. J. Nunn, T. P. Dee, *Proc. ISMA Tech. Conf.,* paper LE 388, Cambridge 1953.
31. "Super Flo Process for Superphosphate," *Phosphorus Potassium* **1** (1962) 34.
32. K. D. Jacob: *Superphosphate: Its History, Chemistry, and Manufacture,* U.S. Department of Agriculture and Tennessee Valley Authority (TVA), Washington, D.C., 1964, pp. 116 – 250.
33. P. Monaldi, G. Venturino: "Process for Recycling H_2SiF_6 Solutions Recovered by Gas Washing to the Superphosphate Den," *Proc. ISMA Tech. Conf.,* 1976, Paper 1, 1 – 16.
34. "The Market Role of GTSP – Key Features of Plant Design," *Proc. Br. Sulphur Corp. Int. Conf. Fert.,* 1977, Paper 21.
35. "Triple Superphosphate Manufacture," *Phosphorus Potassium* **80** (1975) 33.
36. A. Sinte Maartensdijk: "Direct Production of Granulated Superphosphates and PK-Compounds from Sulphuric Acid, Phosphoric Acid, Rock Phosphate and Potash," *Proc. ISMA Tech. Conf.,* 1976, Paper 12, 200 – 214.
37. *Proc. Annu. Meet. Fert. Ind. Round Table,* 1968, 123 – 127;1976, 93 – 98, 120 – 130. I. W. McCamy, M. M. Norton, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.,* 1977, 68 – 94.
38. B. R. Parker et al., *Proc. ISMA Tech. Conf.,* Orlando 1978, Paper 19, 255 – 282.
39. *Phosphorus Potassium* **81** (1976) 37 – 38.
40. *Proc. Annu. Meet. Fert. Ind. Round Table* **16** (1966) 69 – 73. I. C. Brosheer, J. F. Anderson, *J. Am. Chem. Soc.* **68** (1946) 902 – 904.
41. I. A. Brownlie, R. Graham, *Proc. ISMA Tech. Conf.,* Helsinki, 1963, Paper 1, 2 – 26.
42. J. D. C. Hemsley, *Proc. Annu. Meet. Fert. Ind. Round Table,* 1968, 115 – 123.
43. *Phosphorus Potassium* **49** (1970) 18 – 19.
44. *Phosphorus Potassium* **157** (1988) 36.
45. L. M. Marzo, J. L. Lopez-Nino, *Proc. Fert. Soc.* **245** (1986) 1 – 13.
46. E. L. Newman, L. H. Hull, *Farm Chem.* **I** (1965) no. 6, 48, 49;**II** (1965) no. 8, 25 – 30; **III** (1965) no. 9, 48, 50, 70.
47. G. H. Wesenberg in F. T. Nielsson (ed.): *Manual of Fertilizer Processing,* Marcel Dekker, New York 1987, pp. 258 – 265.
48. R. R. Heck, *Proc. ISMA Tech. Conf.,* Brussels, 1968, V.
49. D. M. Ivell, N. D. Ward, *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Paris 1984, TA/84/3, Paper 3, 3.1 – 3.10.
50. G. H. Wesenberg in F. T. Nielsson (ed.): *Manual of Fertilizer Processing,* Marcel Dekker, New York 1987, pp. 251 – 258.
51. P. Chinal, Y. Cotonea, *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Paris 1984, TA/84/1, Paper 1, 1.1 – 1.19.
52. A. Constantinides, J. Lainé, D. Bellis, *Proc. Tech. Conf.,* Port el Kantaoui 1986, TA/86/12.
53. P. Chinal, Y. Cotonea, C. Debayeux, J. F. Priat, *Proc. Annu. Meet. Fert. Ind. Round Table,* London, April 17, 1986.
54. G. M. Hebbard, *Proc. Annu. Meet. Fert. Ind. Round Table* vol. 129, Washington, D.C. 1979.
55. A. T. Brook, *Proc. Fert. Soc.,* 1957, no. 47.
56. O. Bognati, L. Buriani, I. Innamorati, *Proc. ISMA Tech./Agr. Conf.* Stresa, 1967, Paper IX.
57. H. Rug, K. Kahle, *Proc. Fert. Soc.* 1990, no. 297.
58. A. Seixas, J. D. Ribeiro Marcal, J. Correia, *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Paris 1984, TA/84/11.
59. S. Maier, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.,* 1977, 283 – 295.
60. W. B. Pietsch, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.,* 1983, 467 – 479.
61. R. Zisselmar, *Proc. Fert. Soc.* 1985, no. 238.
62. A. Stephenson, *Proc. Fert. Soc.* 1985, no. 238.
63. F. G. Keenan, *Ind. Eng. Chem.* **22** (1930) no. 12, 1378, 1382.
64. L. M. White, J. O. Hardesty, W. H. Ross, *Ind. Eng. Chem.* **27** (1935) no. 5, 562 – 567.
65. L. D. Yates, F. T. Nielsson, G. C. Hicks, *Farm. Chem.* **117** (1954) no. 7, 38 – 48;no. 8, 34 – 41.
66. D. M. Ivell, *IFDC-FADINAP-FAI Workshop,* Madras, India 1988.
67. Tennesse Valley Authority, US 2 741 545, 1953 (F. T. Nielsson).
68. G. H. Wesenberg in F. T. Nielsson (ed.): *Manual of Fert. Processing,* Marcel Dekker, New York 1987, 251 – 258.
69. R. R. Heck, *Proc. ISMA Tech. Conf.,* Brussels, 1965, Paper V.
70. *Phosphorus Potassium* **49** (1970) 18 – 19.
71. M. D. Pask, I. Podilchuk, *ISMA/ANDA Tech. Conf.,* Sao Paolo, Brazil 1975.

72. G. Hunter, J. L. Hawksley, *Proc. FAI – ISMA Semin. Technol. Comp. Fert. Based Urea, Use Benefic, Low Grade Phosphate Rock,* New Delhi 1975, III/3, 1 – 34.
73. J. L. Hawksley, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.* 1977, 110 – 124.
74. K. J. Barnett, D. M. Ivell, *Proc. Fert. Soc.* 1983, no. 216.
75. D. M. Ivell, N. D. Ward, *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Paris 1984, TA/84/3.
76. B. R. Parker, M. M. Norton, I. W. McCamy, D. G. Salladay, *Proc. Int. Fert. Ind. Tech./Econ. Conf.,* Orlando 1978, 255 – 282.
77. R. J. Milborne, D. W. Philip, *Proc. Fert. Soc.* 1986, no. 244.
78. F. G. Membrillera, J. L. Toral, F. Codina: *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.* 1977, Paper 12.
79. I. S. Mangat, J. M. Toral, *Proc. Int. Fert. Ind. Assoc. Tech./Econ. Conf.,* Orlando 1978, 239.
80. P. Chinal, Y. Cotonea, C. Debayeux, J. F. Priat, *Proc. Fert. Soc.* 1986, no. 245.
81. P. Chinal, Y. Cotonea, *Proc. Int. Fert. Ind. Assoc. Tech. Conf.,* Paris 1984, TA/82/2.
82. D. W. Leyshon, I. S. Mangat, *Proc. ISMA Tech. Conf.,* The Hague 1976, TA/76/21, 368.
83. I. S. Mangat, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.* 1977, Paper 11.
84. E. Peletti, J. C. Reynolds, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.,* 1977, 95 – 109.
85. J. W. Baynham, *Proc. ISMA Tech. Conf.,* Edinburgh 1965, Paper 12.
86. Y. F. Berquin, *Proc. Br. Sulphur Corp. Int. Conf. Fert. Technol.,* 1977, 296 – 303.
87. S. J. Porter, W. F. Sheldrick, *Proc. Annu. Meet. Fert. Ind. Round Table* **16** 1966, 104 – 109.
88. *TVA-Bulletin Y-107,* 1976, 11th Demonstration, 44 – 48.
89. R. G. Lee, R. S. Meline, R. D. Young, *Ind. Eng. Chem. Process Des. Dev.* **11** (1972) no. 1, 90 – 94.
90. T. A. Mitchel, B. H. Fotherfill, W. J. Kelly, *Proc. ISMA Tech. Conf.,* Sandelfjord 1970, Paper 6.
91. Odda Smelteverk, US 1 816 285, 1929 (E. B. Johnson).
92. G. Langhans, B. Bieniok, *Proc. Tech. Conf. ISMA 1976* 1977, paper 13, 215 – 233.
93. K. Tesche, "New Experiences with the Kampka Nitro Process," *Proc. FAI – ISMA Semin. Technol. Comp. Fert. Based Urea Use Benefic Low Grade Phosphate Rock* ITE/71/16 (1971).
94. J. Kofiseh, L. Hellmer, H. P. Bethke: "Nitrophosphate Process Using Direct Cooled Continuous Crystallization," *UNIDO Second Interregional Fert. Symposium,* Kiev 1971.
95. S. I. Vol'Gkovich, A. A. Sokolovskii, *Russ. Chem. Rev.* **43** (1974) no. 3, 224 – 234.
96. A. V. Slack: "Phosphoric Acid," *Fert. Sci. Technol. Ser.* **1** (1968) 654 ff.
97. J. F. Steen, S. G. Terjesen: "Norsk Hydro Nitrophosphate Process," *UNIDO Second Interregional Fert. Symposium,* Kiev 1971.
98. L. Diehl, K. F. Kummer, H. Oertel: *Nitrophosphates with Variable Water Solubility: Preparation and Properties,* The Fert. Society of London, April 1986.
99. K. C. Knudsen: *Production of Chloride-Free NPK Fertilizer and Feed Grade Dicalcium Phosphate,* The Fertilizer Society of London, Oct. 1985.
100. O. Kjøhl: "Product Quality Requirements in Bulk Shipments of Fertilizers," *New Dev. Phosphate Fert. Technol. Proc. Tech. Conf. ISMA, 1976.*
101. R. Ringbakken, O. Lie, G. T. Mejdell: "Urea as an Agent in the Destruction/Recovery of NO_x in the Nitric Acid and Nitrophosphate Fertilizers Production," *Proc. ISMA Tech. Conf.,* Prague 1974, Paper 22, 22.1 – 22.20.
102. Norsk Hydro, US 3 617 235, 1967 (I. A. Friestad, O. Skanli).
103. Norsk Hydro, US 3 561 678, 1967 (I. A. Friestad).
104. Norsk Hydro, US 3 900 164, 1973 (I. A. Friestad).
105. Hoechst, DE 1 016 723, 1955 (F. Moosbrugger et al.).
106. BASF, DE 1 064 536, 1957 (K. Huberich).
107. Chemische Fabrik Kalk, DE 1 065 436, 1957 (H. Nees, K. Geiersberger).
108. J. F. Steen, T. Heggebø, E. Aasum: "The Norsk Hydro Nitrophosphate Process," *Manual of Fertilizers Processing,* chap. 15, Marcel Dekker, New York 1987, pp. 393 – 419.
109. *Phosphorus Potassium* **155** (1988) 25 – 33.
110. Stamicarbon BV, Technical information, Jan. 1991.
111. G. H. McClellan : "Mineralogy and Reactivity of Phosphate Rock," *Seminar on Phosphate Rock for Direct Application,* IFDC publication, Muscle Shoales, Ala., 1978, pp. 57 – 81.
112. "Potential for Use of Unacidulated and Partially Acidulated Phosphate Rock," *Phosphorus Potassium* **168** (1990) 15 – 19.
113. "Ground Phosphate Rock as a Direct Application Fertilizer in New Zealand," *Phosphorus Potassium* **128** (1983) 17 – 21.

114. F. E. Khawawneh, E. C. Doll: "The Use of Phosphate Rock for Direct Application," *Adv. Agron.* **30** (1978) 195 – 206.
115. N. S. Bolan, M. J. Hedley: "Dissolution of Phosphate Rocks in Soils," *Fert. Res.* **24** (1990) 125 – 134.
116. G. J. D. Kirk, P. H. Nye: "A Simple Model for Predicting the Rate of Dissolution of Sparingly Soluble Calcium Phosphates in Soil," *J. Soil. Sci.* **37** (1986) 541 – 554.
117. A. D. Mackay et al.: "A Simple Model to Describe the Dissolution of Phosphate Rock in Soils," *Soil Sci. Soc. Am. J.* **50** (1986) 291 – 296.
118. "Partial Acidulation of Phosphate Rock," *Phosphorus Potassium* **150** (1987) 48 – 53.
119. O. W. Livingston: "Minigranulation – A Method for Improving the Properties of Phosphate Rock for Direct Application," *Seminar on Phosphate Rock for Direct Application,* IFDC publication, Muscle Shoales, Ala., 1978, pp. 367 – 377.
120. L. L. Hammond et al.: "Agronomic Value of Unacidulated and Partially Acidulated Phosphate Rocks Indigenous to the Tropics," *Adv. Agron.* **40** (1986) 89 – 140.
121. J. J. Schultz: "Sulphuric Acid-Based Partially Phosphate Rock – its Production, Cost and Use," *IFDC Tech. Bull.* IFDC-T-31, Ala., April 1986.
122. Fisons, DE-AS 1 592 644, 1966 (G. G. Brown, D. C. Harper).
123. G. Trömel, E. Görl, *Arch. Eisenhüttenwes.* **35** (1964) 287 – 298.
124. G. Trömel, *Stahl Eisen* **63** (1943) 21 – 30.
125. IRSID, FR 1 309 365, 1961. ATH, DE 1 237 592, 1961 (M. T. Blauel, T. Kootz, A. Michel).
126. Kali & Salz, DE 973 396, 1953 (H. Keitel, W. Jahn-Held, W. Appel).
127. Kali & Salz, DE 2 263 334, 1972 (A. Singewald, W. Jahn-Held).
128. M. S. Mudahar, T. P. Hignett: "Energy and Fertilizer. Policy Implications and Options for Developing Countries," *IFDC Tech. Bull.,* IFDC T-20, Ala., 1982.
129. O. C. Bøckman, O. Kaarstad, O. H. Lie, I. Richards: *Agriculture and Fertilizers. Fertilizers in Perspective,* Agricultural Group, Norsk Hydro, Oslo 1990.
130. International Maritime Organization: Code of Safe Practice for Solid Bulk Cargoes, Appendix C, 1987, p. 94.
131. International Maritime Organization: Code of Safe Practice for Solid Bulk Cargoes, Appendix D 4, 1987, p. 121.

Phosphates, Organic → Phosphorus Compounds, Organic
Phosphating → Metals, Surface Treatment
Phosphazenes → Phosphorus Compounds, Inorganic
Phosphazenes → Inorganic Polymers
Phosphines → Phosphorus Compounds, Organic
Phosphonium Compounds → Phosphorus Compounds, Organic

My Profile Lo
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Phosphoric Acid and Phosphates

Standard Article

Klaus Schrödter[1], Gerhard Bettermann[1], Thomas Staffel[2], Thomas Klein[2], Thomas Hofmann[3]
[1]Hoechst Aktiengesellschaft, Werk Knapsack, Hürth, Germany
[2]BK Giulini, Ladenburg, Germany
[3]Sanofi-Aventis Deutschland GmbH, Drug Evaluation, Hattersheim, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a19_465.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Monophosphoric Acid
1.1. Properties
1.2. Production
1.2.1. Thermal Processes
1.2.2. Wet Processes
1.2.3. Concentration of Wet Phosphoric Acid
1.2.4. Purification of Wet Phosphoric Acid
1.3. Uses and Economic Aspects
1.4. Environmental Aspects
2. Condensed Phosphoric Acids
3. Phosphates
3.1. Sodium Monophosphates
3.2. Condensed Sodium Phosphates
3.2.1. Properties
3.2.2. Sodium Diphosphates
3.2.3. Pentasodium Triphosphate
3.2.4. Higher Condensed Sodium Polyphosphates
3.2.5. Sodium Metaphosphates
3.2.6. Uses of Condensed Sodium Phosphates
3.3. Other Phosphates
3.3.1. Potassium Phosphates
3.3.2. Ammonium Phosphates
3.3.3. Magnesium Phosphates
3.3.4. Calcium Phosphates
3.3.5. Boron Phosphate
3.3.6. Aluminum Phosphates
3.4. Economic Aspects
4. Toxicology

[Top of Page]

1. Monophosphoric Acid

Klaus Schrödter and Gerhard Bettermann

Phosphoric acid was first prepared in 1694 by BOYLE from P_4O_{10} soon after the discovery of elemental phosphorus.

1.1. Properties

Physical Properties. Pure, anhydrous phosphoric acid , H_3PO_4, M_r 98, *mp* 42.35 °C [1], ϱ 1.88 g/cm^3, is a colorless, crystalline compound, that is readily soluble in water.

A hemihydrate of phosphoric acid, $H_3PO_4 \cdot ½ H_2O$, *mp* 29.25 °C, is also known.

Figure 1 shows the H_3PO_4 – H_2O phase diagram [2].

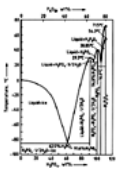

Figure 1. H_3PO_4–H_2O phase diagram

Phosphoric acid is infinitely miscible with water. It is commercially available in three standard concentrations:

75 % H_3PO_4 with 54.3 % P_2O_5, *mp* –20 °C
80 % H_3PO_4 with 58.0 % P_2O_5, *mp* 0 °C
85 % H_3PO_4 with 61.6 % P_2O_5, *mp* +21 °C

The 85 % acid (ϱ 1.687 g/cm^3) has syrupy consistency. It tends to supercool, especially when transferred to glass bottles. It is usually stored in heated containers in order to avoid crystallization. The anhydrous acid can be obtained from the 85 % acid by evaporation in vacuum at 80 °C.

Figures 2 and 3 show plots of the density and viscosity of phosphoric acid against concentration and temperature. The density is practically constant in the temperature range 25 – 100 °C. As shown in Figure 4, the same is true of the specific heat.

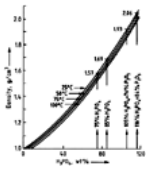

Figure 2. Density of phosphoric acid as a function of concentration and temperature

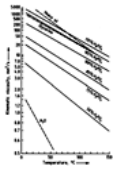

Figure 3. Kinematic viscosity of phosphoric acid as a function of concentration and temperature

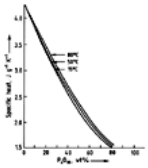

Figure 4. Specific heat of phosphoric acid as a function of concentration and temperature

The boiling point of phosphoric acid as a function of concentration and the composition of the corresponding vapor phases are shown in Figure 5 [3]. Only traces of P_4O_{10} are present in the gas phase at temperatures up to 300 °C. P_4O_{10} appears in the vapor phase only above this temperature after condensation of phosphoric acid to form polyphosphoric acids.

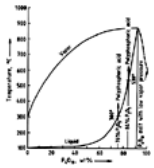

Figure 5. Liquid – vapor equilibrium diagram of the P_4O_{10} – H_2O system

Chemical Properties. Monophosphoric acid is tribasic. The dissociation constants are $K_1 = 7.1\times10^{-3}$, $K_2 = 6.3\times10^{-8}$, and $K_3 = 4.7\times 10^{-13}$.

Accordingly, primary, secondary, and tertiary salts are formed with bases. Aqueous solutions of dihydrogenphosphates are weakly acidic, while those of hydrogenphosphates and tertiary phosphates are weakly or strongly alkaline, respectively. As shown in Figure 6, mixtures of dihydrogenphosphates and hydrogenphosphates act as buffers, e.g., in the range pH 6 to 8.

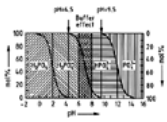

Figure 6. Ion concentration of a phosphoric acid (phosphate) solution as a function of pH

1.2. Production

Two methods are used for the industrial production of phosphoric acid. The wet digestion of phosphate rock with mineral acids is the most important process in terms of volume. Digestion is usually carried out with sulfuric acid, but nitric acid [4] or hydrochloric acid [5] are also used to a lesser extent. The acid obtained in this process is called wet phosphoric acid.

The second process for the production of phosphoric acid starts with elemental phosphorus, which is obtained from crude phosphate, coke, and silica in an electric resistance furnace (⟶ Phosphorus), and subsequently oxidized to P_4O_{10}. The acid obtained by hydration of the oxide is generally termed thermal phosphoric acid.

1.2.1. *Thermal Processes*

The production of phosphoric acid from elemental phosphorus is carried out in two stages: combustion with excess air to give P_4O_{10} and hydration of P_4O_{10} to form H_3PO_4:

$$P_4 + 5\,O_2 \rightarrow P_4O_{10}\ \Delta H = -3053\,\mathrm{kJ/mol} \quad (1)$$

$$P_4O_{10} + 6\,H_2O \rightarrow 4\,H_3PO_4\ \Delta H = -377\,\mathrm{kJ/mol} \quad (2)$$

The IG and Tennessee Valley Authority (TVA) processes for the production of phosphoric acid are distinguished by the method of absorption and hydration of P_4O_{10}. The Hoechst process differs from these in that it utilizes the heat of phosphorus combustion for steam generation.

IG Process [6]. In the IG process (Fig. 7), phosphorus combustion and P_4O_{10} absorption are carried out in a single stage in a common tower. The tower walls are slightly conically inclined so that the recycled acid, overflowing inside from a cup located at the tower head, can completely cover the wall with an acid layer. This protects the wall from the phosphorus flame that descends from the tower head and reaches a temperature > 2000 °C. The recycling phosphoric acid dissolves the phosphorus pentoxide formed and dissipates the heat from the reaction zone. It is collected below the acid tower and recycled to the tower cup after passing through the heat exchanger.

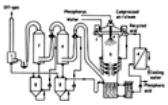

Figure 7. Production of thermal phosphoric acid (IG process)

a) Nozzle; b) Combustion tower; c) Overflow cup; d) Heat exchanger; e) and f) Venturi scrubbers; g) Receiver for wash acid; h) Off-gas fan; i) Separator

Phosphoric acid overflows from the tower cup at ca. 60 °C and is heated on the way along the tower walls to up to 85 °C. About 80 % of the heat generated is dissipated by the cooling water, 17 % by off-gas and water evaporation, and 3 % by the sensible heat of the acid.

The phosphorus is sprayed into the tower from a binary nozzle by pressurized air or steam at up to 1.5 MPa. It is ejected from the nozzle at such a high speed that the phosphorus flame is kept away from the head and cannot damage it. A slight underpressure is maintained in the tower, generated by a fan which draws the combustion air from an annular aperture around the nozzle at the tower head. The off-gas is discharged in the wash stage. Here, P_4O_{10} that is entrained in the off-gas in the form of aerosol or acid droplets is separated. It accounts for ca. 30 % of the total conversion. The wash acid is added to the tower acid. Venturi systems, wash towers operated with or without packing, or demister elements packed with glass fibers are used as separators [7].

All parts of the equipment which come into contact with P_4O_{10}-containing off-gases are made of rubber- coated steel, stainless steel, or plastic. A stainless steel with very low carbon content, stable against concentrated phosphoric acid at up to 100 °C, has prevailed as construction material for the acid tower.

In addition to the phosphoric acid flowing down on the tower walls, phosphoric acid is sprayed into the tower through nozzles in the lower tower half below the phosphorus flame and perpendicular to the off-gas flow. This decreases the off-gas temperature to ca. 100 °C.

The orthophosphoric acid obtained is virtually free of lower phosphorus oxides and its H_3PO_3 content is 0.1 %.

Thermal phosphoric acid contains only traces of impurities due to the use of pure phosphorus for combustion. Only arsenic, a natural companion of elemental phosphorus which is contained in the acid at a concentration of 5 –30 ppm, need be removed in a further purification step. This is accomplished by precipitation of arsenic sulfide with hydrogen sulfide followed by filtration. The hydrogen sulfide is generated by injecting Na_2S solution into the phosphoric acid. This is done either in a precipitation tower with subsequent filtration after addition of activated charcoal and kieselgel [8], or more directly and simply in a centrifugal pump which immediately conveys the acid to the filter [9].

TVA Process [10]. In the Tennessee Valley Authority (TVA) process phosphorus combustion and P_4O_{10} absorption are carried out separately (Fig. 8).

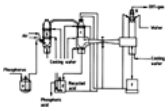

Figure 8. Production of thermal phosphoric acid (TVA process)

a) Storage tank; b) Combustion chamber; c) Absorption tower; d) Cooled receiver; e) Venturi scrubbers; f) Dilute acid tank; g) Separator

The vertical, cylindrical, stainless steel phosphorus combustion chamber is externally cooled. The phosphorus burner is installed laterally and is directed downwards toward the bottom of the combustion chamber at an angle of ca. 40°.

The gas discharge and the connection to the absorption tower are located at the top of the combustion chamber. A layer of polyphosphoric acid is formed from generated P_4O_{10} and atmospheric moisture, which protects the inner walls against corrosion. The polyphosphoric acid contains ca. 92 % P_2O_5, and its vaporization point is 867 °C. Phosphorus and air are introduced in a manner similar to the IG process.

The stainless steel cylindrical absorption tower with a conical head is also externally cooled. The inner tower walls are sprayed with dilute phosphoric acid from the off-gas purification stage and with recycled, concentrated phosphoric acid such that an unbroken acid film is formed. The phosphoric acid so produced contains 54 – 83 % P_2O_5.

A significant part of the P_4O_{10} product is entrained in the gas flow from the absorption tower. Thus, depending on whether the product is orthophosphoric acid (54 % P_4O_{10}) or superphosphoric acid (83 % P_4O_{10}), 40 – 70 % of the acid produced accumulates in the gas purification stage. Venturi scrubbers, wash towers, and demisters with webs made of special steel are used for flue-gas purification.

Hoechst Process [11]. The heat generated by combustion of phosphorus is unused and is dissipated by cooling in the IG and TVA processes. In the Hoechst process (Fig. 9), operated industrially since 1989, this energy is used for steam generation. Phosphorus combustion and P_2O_5 hydration are performed separately.

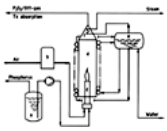

Figure 9. Schematic of the Hoechst process

a) Storage tank; b) Air dryer; c) Burner; d) Combustion chamber; e) Separator; f) Superheater

Liquid phosphorus and dried air are injected into the bottom of a vertical stainless steel combustion chamber and are combusted to give P_4O_{10}. The walls of the combustion chamber are made of pipe coils through which deionized water is pumped. The steam – condensate mixture formed in the coils by dissipation of the heat of combustion is kept at a pressure of 15 – 18 MPa. The mixture is separated into gas and liquid in a separator. The remaining condensate is returned to the water storage tank, and the evaporated portion is replenished by addition of deionized water. The high-pressure steam is transferred from the separator to the combustion chamber, where the steam temperature is increased to 400 – 500 °C at 15 – 18 MPa in a superheater. The pressure is then released, and the steam is saturated with water and used as a heat-transfer medium. The heat dissipation in the phosphorus combustion is controlled so that the temperature of the P_4O_{10}-containing off-gas is high enough to avoid separation of P_4O_{10} on the way to the absorption tower, where hydration of P_4O_{10} is carried out. An acid film is formed on the tower walls by concentrated recycled acid, similar to the IG process.

1.2.2. *Wet Processes*

The digestion of phosphate rock with sulfuric acid has been known since the mid-1880s. However, intensive research and development in this area started only after World War II, due to increased use of mineral fertilizers.

The digestion of fluorapatite with sulfuric acid proceeds according to Equation 3:

$$Ca_5(PO_4)_3F + 5\,H_2SO_4 + 5xH_2O \rightarrow 5\,CaSO_4 \cdot xH_2O + 3\,H_3PO_4 + HF \qquad (3)$$

$$x = 0, 1/2, 2$$

However, concurrent side reactions occur with other, secondary components of the crude phosphate. Their reaction products are then largely dissolved in the wet phosphoric acid. The typical composition of crude wet phosphoric acid (in wt %) is as follows:

P_2O_5	30.0
SO_3	2.1
F	2.17
Si (mainly as SiF_6^{2-})	0.41
Ca	0.45
Mg	0.13
Fe	0.08
Al	0.04
Cr	0.022
Zn	0.019
V	0.013
Ni	0.0034
Cu	0.0027
Mn	0.0009
As	0.0007
Pb	0.0002

Calcium sulfate can precipitate in three different modifications, depending on the temperature and the phosphoric acid concentration (Fig. 10). The designation of the various wet processes is derived from the modification formed in the digestion stage. Thus, dihydrate (DH) and hemihydrate (HH) processes are distinguished. In other processes, two crystalline forms are formed sequentially and separated by recrystallization (HH/DH or DH/HH processes).

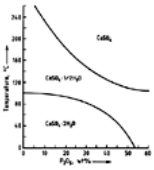

Figure 10. Stability range of calcium sulfate and its hydrates in phosphoric acid

So far, the development of an anhydrite process has failed because the high reaction temperature of 120 – 130 °C required for anhydrite formation causes insurmountable corrosion problems.

Two basic operations are common to all wet phosphoric acid processes: digestion in a reactor and removal of $CaSO_4$ by filtration.

Reactors. Originally digestion was carried out in several large stirred tanks arranged in series, with recirculation of the digestion slurry by pumps. The introduction of single tanks brought an essential simplification as well as an improvement of the circulation for compensation of temperature and concentration fluctuations, and thus for optimization of the crystallization conditions [12], [13]. Round models (see Fig. 11) as well as rectangular ones were constructed. In the round single tank, the reaction components are slowly moved in a circle in the outer, annular space with vigorous, local agitation. Part of the digestion slurry overflows into the central space and flows to filtration. In medium-sized facilities (200 – 400 t/d P_2O_5), air is blown onto the surface for dissipation of the heat of reaction (2 GJ/t P_2O_5 in a DH process). In larger plants, the heat is dissipated by vacuum evaporation of water from a recycled partial stream of the slurry in a flash cooler [14], [15].

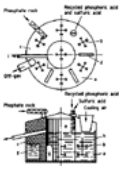

Figure 11. Single tank round model

a) Outer annular section; b) Agitators; c) Partition; d) Cooling air nozzles; e) Center section; f) Outer passage; g) Passage to the center cylinder; h) Carbon bricks; i) Overflow channel

Filters. About 4 – 6 t of $CaSO_4$ per tonne of P_2O_5 produced must be filtered from the wet phosphoric acid. After initially using box band filters [16], producers switched to the use of plane and tilt-cell filters (Prayon, Bird, Eimco, UCEGO [13], [15], [17], [18]). The effective filter area can approach up to 200 m^2. Later, band filters with larger outputs and filter areas exceeding 100 m^2 became available. In the filtration, the wet phosphoric acid is drawn off from the solids in a partial vacuum. Adhering acid is washed off the solids in two or three stages in countercurrent. The wash acid is discharged separately and recycled to the reactor. The discharge of the filter cake is carried out by tilting the filter pans, by means of a screw conveyor, or by stripping from the band.

Due to technical progress in reactor and filter construction, production units for wet phosphoric acid now have capacities up to 1000 t/d of P_2O_5.

Dihydrate Process. The dihydrate process is the classical method for the production of wet phosphoric acid (Fig. 12). Prior to digestion, the phosphate rock must usually be ground to a fineness of 75 % < 150 µm. Phosphate rock and concentrated sulfuric acid are charged spatially separated. In this way, the concentration of sulfuric acid is lowered so that the digestion occurring at the phosphate rock charging point is not hindered by formation of a gypsum layer on the ore particles. In addition, the formation of too many gypsum seed crystals due to supersaturation of Ca^{2+} and SO_4^{2-} ions is avoided. The calcium sulfate is present as the dihydrate at a reaction temperature of 70 – 80 °C and a phosphoric acid concentration of 28 – 31 %. Usually, the filtered acid is further concentrated (see Section Concentration of Wet Phosphoric Acid).

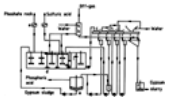

Figure 12. Flow diagram for the DH process

a) Metering device; b) Scrubber; c) Filters; d) Single tank; e) Thickener; f) Discharge tank

Advantages of the dihydrate process are the low reaction temperature and, therefore, fewer corrosion problems; its applicability to most grades of phosphate rock; and its suitability for large-volume production. Some disadvantages are the production of a comparatively dilute phosphoric acid which must subsequently be concentrated; the need to grind the phosphate rock; and the relatively high P_2O_5 losses of 4 –6 % in the gypsum.

Hemihydrate Process. Rising energy costs initiated the development of processes that allow the production of phosphoric acid with 40 – 50 % P_2O_5, without using additional concentration steps, for direct use in fertilizer production. These processes must be carried out at higher temperature in the stability range of calcium sulfate hemihydrate (see Fig. 10). Thus, single-stage (HH) and two-stage (HH/DH, DH/HH) processes were developed in which the hemihydrate is formed as end or intermediate product.

The first commercial HH/DH process was introduced in Japan. The major interest in this process was the production of pure gypsum that could be used for cement production since there are no natural gypsum deposits in this region.

There are currently 60 hemihydrate plants operating worldwide; more than half of them use a HH/DH process. The single-stage HH process has seldom been used because of the unfavorable process yield.

The various processes and some typical features are as follows:

1. *HH Process [19]*

   Features:
   - Single filtration stage
   - Impure $CaSO_4 \cdot 1/2\ H_2O$
   - Yield, 90 – 94 %
   - Phosphoric acid, 40 – 48 % P_2O_5
2. *HH/DH Process [20]*

   a. Without intermediate separation of the hemihydrate
      Features:
      - Single filtration stage
      - Pure $CaSO_4 \cdot 2\ H_2O$
      - Phosphoric acid, 30 – 32 %P_2O_5

   b. With intermediate separation
      Features:
      - Two filtration stages
      - Very pure gypsum
      - Yield, 98.5 %
      - Phosphoric acid, 40 – 52 % P_2O_5

Figure 13 shows the flow sheet for the Fison process as an example of a HH/DH process with intermediate separation of the hemihydrate [14]. In this two-stage processes, the higher digestion temperature of 85 – 110 °C increases the reaction rate and allows the use of unground phosphate rock.

Figure 13. Flow diagram for the HH/DH process

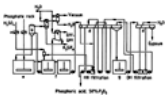

a) Metering device; b) Flash cooler; c) Gas scrubbing; d) Filters; e) Premixer; f) Digester; g) Hydration tank

Depending on the process, the conversion of the hemihydrate is carried out under different conditions. In the Nissan process, which dispenses with the intermediate separation of hemihydrate, the dihydrate is crystallized after cooling of the digestion slurry from 85 – 90 °C to 55 –65 °C [20]. In the Fison process the hemihydrate is recrystallized after filtration from a washing acid (a mixture of sulfuric and phosphoric acid: 4 – 8 % SO_3, 10 – 20 % P_2O_5) slurry.

3) DH/HH Process [14], [21]

Features:

Repeated filtration with intermediate conversion
Very pure $CaSO_4 \cdot 1/2\ H_2O$
Yield, 96 – 98 %
Phosphoric acid, 32 – 36 % P_2O_5

This process is only of interest where very high purity hemihydrate is required, for example, in plaster.

1.2.3. *Concentration of Wet Phosphoric Acid*

Most of the wet phosphoric acid produced worldwide is made by the dihydrate process. For further processing to fertilizers, the major application (⟶ Phosphate Fertilizers), the concentration of the phosphoric acid must be increased from 28 – 30 % P_2O_5 to 40 – 55 % or even 70 % P_2O_5 (superphosphoric acid). Submerged combustion burners and vacuum circulation evaporators are customary concentration devices. In the older submerged burner process, hot combustion gases from an oil burner are blown through a dip tube or directly into phosphoric acid [22]. At an acid temperature of 350 °C, the off-gases entrain high proportions of phosphoric acid as aerosols as well as volatile fluorine compounds. Purification of these off-gas streams is expensive.

Today, submerged combustion burners are still in use only in older plants. Currently, vacuum evaporation with forced circulation is the concentration method generally used (Fig. 14). Depending on the amount of water to be evaporated, one or more evaporators are used in parallel or in series. Multistage evaporators arranged in tiers of usually three stages have the advantage of lower energy costs. The energy savings result from effective heat utilization. The vapors of an evaporator stage operated at a higher temperature are used for heating another operated at a lower temperature [23]. In the concentration of phosphoric acid from 30 to 50 % P_2O_5, the steam requirements can be reduced from 1.25 t to 0.7 t per tonne of evaporated water. Rubber-proofed steel is the predominant material of construction for containers, piping, and pumps at operating temperatures below 120 °C. At higher temperatures, e.g., in the production of superphosphoric acid at up to 200 °C, alloys such as Hastelloy G 3 or Sanicro 28 are used. The heat exchanger tubes are usually made of graphite or nickel alloys.

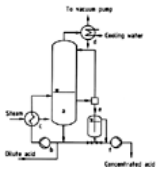

Figure 14. Recycling evaporator for concentration of wet phosphoric acid

a) Recycle evaporator; b) Circulation pump; c) Heat exchanger; d) Condenser; e) Product takeoff; f) Pump

Apart from water vapor, a mixture of SiF_4 and HF is generated during the concentration of wet phosphoric acid. About 50 – 60 % of the fluorine content of wet phosphoric acid is volatilized on concentration from 30 to 55 % P_2O_5. In order to comply with the fluorine emission limits of the plant or to obtain a salable coproduct, concentration units are equipped with facilities for the production of H_2SiF_6. Thus, 90 – 95 % of the generated fluorine can be converted to a 15 – 25 % H_2SiF_6 solution [24].

1.2.4. *Purification of Wet Phosphoric Acid*

Wet phosphoric acid produced by wet digestion contains variable amounts of inorganic impurities, depending on the origin of the phosphate rock.

Depending on the further application, these impurities must be partially or completely removed from crude phosphoric acid. Precipitation and extraction processes are used.

Precipitation Processes. The separation of arsenic and, to an increasing extent, of cadmium is required for the use of wet phosphoric acid in fertilizer production. For separation of arsenic, a Na_2S solution is added to the crude acid, and the precipitated arsenic sulfide is removed by filtration.

The separation of cadmium by means of a complexing agent such as alkyldithiophosphoric acid alkyl ester [25] can also be classified as a precipitation reaction because the precipitated cadmium complex is separated in solid form either directly or after addition of a filter aid or absorbent. Such a process is operated on a commercial scale by Tessenderlo Chemie in Belgium.

Further purification of wet phosphoric acid by precipitation of cationic impurities, especially Fe, Al, Mg, and Ca, is possible by neutralizing the acid with caustic soda. The precipitated metal phosphates are filtered under pressure from the resulting sodium phosphate solution, which contains 18 – 20 % P_2O_5. In order to decrease P_2O_5 losses in the filter cake from > 10 % to < 5 %, [26] the filter cake is worked up with more caustic soda to give a trisodium phosphate solution and precipitatable metal hydroxides. Since the phosphoric acid in this neutralization process is converted to a phosphate salt solution, its uses are limited. For instance, wet phosphoric acid is used for the production of detergent phosphates (pentasodium triphosphate, see Section Pentasodium Triphosphate) via this route in Western Europe.

Extraction processes are the predominant methods for the purification of wet phosphoric acid. Processes are operated for the separation of single components as well as of practically all impurities in wet phosphoric acid. However, this complete purification includes measures (e.g., acid discoloration and removal of traces of arsenic and fluorine) which supplement the extraction process. The quality of such purified phosphoric acid equals that of thermally produced acid.

In 1990, production capacities for extractively purified P_2O_5 amounted to 740 000 t/a worldwide, including 300 000 t/a in Western Europe.

The basic concept for extractive purification of wet phosphoric acid is the same in all processes. Figure 15 shows a schematic of the purification steps. Crude wet phosphoric acid containing 40 – 50 % P_2O_5 is prepurified in order to largely separate, if necessary, arsenic, sulfate, and organic components by precipitation or adsorption. In the subsequent extraction process, prepurified acid is contacted with solvent countercurrently in several stages. Together with small amounts of impurities, the phosphoric acid is transferred to the organic phase, while most of the impurities remain in the aqueous phase. If a high yield of P_2O_5 (95 – 98 %) is desired, sulfuric acid is added during the extraction so that the cationic impurities remain in the aqueous phase as sulfates. In addition, so-called splitting processes are customary. In these, only 50 – 70 % of the phosphoric acid are extracted in simple plant. Phosphoric acid contained in the aqueous phase, together with the impurities, is treated further in fertilizer production.

Figure 15. Schematic of extractive purification of wet phosphoric acid

The crude extract containing the phosphoric acid is washed countercurrently with small amounts of water, phosphoric acid, or phosphate salt solutions. Depending on the wash liquid, anionic components such as SO_4^{2-} or F^- are also removed with the metal cations. The wash extract—phosphoric acid enriched in impurities—is recycled to the extraction stage. The phosphoric acid product is obtained from the purified extract by addition of water or by distillation of solvent. If further processing of the acid to commercial phosphates is intended, a phosphate salt solution can be obtained directly by addition of alkalies to the extract. The recovered solvent is recycled to the extraction stage. Usually, a partial stream is purified, for example, by distillation.

Generally, stripping for complete solvent separation and concentration of the phosphoric acid to 50 – 70 % P_2O_5 are among the posttreatment steps. In these operations, the fluoride content of the acid is also further decreased. Pressure stripping of the acid with steam to decrease the fluoride content to <10 ppm and oxidative degradation of organic

trace impurities are used for the production of food-grade phosphoric acid.

The choice of solvent is an important feature of the extraction processes. With solvents that are infinitely miscible with water, such as isopropanol [27], extraction may be complete in one or two stages. Alkali-metal or ammonium ions are added for better development of the organic and aqueous phases. Phosphoric acid is recovered from the extract by solvent distillation.

Frequently, alcohols that are partially miscible with water, such as butanol [7] or amyl alcohol [28], [29], are used. Six- to ten-stage mixer –settler units are used to attain essentially quantitative phosphoric acid extraction.

The IMI process, developed for the extraction of phosphoric acid from phosphate rock digested with hydrochloric acid, is a special case [28]: $CaCl_2$ is added as extraction aid to increase the equilibrium concentration of phosphoric acid in the alcoholic phase.

Methyl isobutyl ketone (MIBK) [30], tributyl phosphate [31], diisopropyl ether [7], or mixtures of the latter two [32] are immiscible with water. Usually, the processes based on these solvents are splitting processes without quantitative P_2O_5 extraction. The use of diisopropyl ether for extraction and reextraction is based on the temperature dependent solubility of phosphoric acid in this solvent. Extraction is carried out at 0 – 5 °C, and the acid is reseparated after addition of a small amount of water at 30 °C.

Extraction of Single Components. The first commercial processes for the separation of single components involved the production of uranium, which can be present in wet phosphoric acid at concentrations of up to 200 ppm. Large industrial plants were first operated in the 1950s with alkyl diphosphate esters [33], and in the 1970s improved processes with octyl phenyl phosphate esters, octyl diphosphate esters [34], or mixtures of di-2-ethylhexyl phosphate and tri-*n*-octyl phosphine oxide [35], [36] were introduced. However, these extraction plants were closed after several years of operation, mainly because of declining uranium prices. Only in Belgium is uranium still produced extractively from wet phosphoric acid by Prayon [37].

A process developed for the removal of cadmium involves prepurification of the acid to separate solids and subsequent extraction of cadmium into an organic phase [38]. Alkylamines are used as extractive agents in the presence of halide ions (Cl^-, I^-, Br^-). In the reextraction stage, cadmium is transferred back to the aqueous phase. Cadmium is precipitated and separated from this solution as the sulfide.

1.3. Uses and Economic Aspects

After sulfuric acid, phosphoric acid is the most important mineral acid in terms of volume and value. This is mainly due to the enormous demand for wet phosphoric acid for further processing to fertilizers. As shown in Figure 16 for the United States, the world's largest producer and consumer of wet phosphoric acid, its share of the total phosphoric acid production amounts to more than 90 %. Besides the fertilizers area, industrial phosphates (see Chap. Phosphates) and feedstuff are other areas of application of wet phosphoric acid.

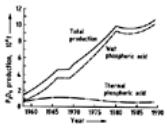

Figure 16. Phosphoric acid production in the United States

About 80 – 90 % of the thermal phosphoric acid produced is used for the production of industrial phosphates, especially the sodium, potassium, calcium, and ammonium salts. Pentasodium triphosphate is always the largest single product, although purified wet phosphoric acid is being used increasingly for its production. Thermal acid is used in metal surface treatment (⟶ Metals, Surface Treatment – Phosphating) and, in food-grade quality, for the acidulation of beverages.

After decades of strong growth, a decrease of worldwide phosphoric acid production occurred in the 1980s (see Fig. 16). Ecological problems related to rising phosphate consumption are an essential cause of this development. Among these are the impairment of subsoil water by phosphate fertilizers and the eutrophication of surface waters by phosphates from fertilizers and detergents. These ecological pressures have already resulted in extensive replacement of phosphate-containing detergents and a decrease in the use of mineral fertilizers in agriculture.

1.4. Environmental Aspects

Phosphoric acid is now mainly produced (ca. 95 %) by the wet process. Because of the high energy consumption, only ca. 5 % is produced by the thermal process. An overview of the worldwide capacity in 1989 (40.6×10^6 t P_2O_5) is given in [39]. Gypsum and fluoride are the main byproducts of the wet process; phosphorus slag is a byproduct of phosphorus production by the electric furnace process.

Phosphogypsum. A quantity of 4.5 – 5.5 t impure calcium sulfate is obtained as byproduct during the production of 1 t P_2O_5 by the wet process. In 1988, 27.7×10^6 t P_2O_5 were produced and hence ca. 140×10^6 t gypsum. Disposal of these large amounts of gypsum is causing considerable problems, especially from an environmental and ecological viewpoint [40]. The gypsum contains production- and source-specific constituents (Table 1) [41]. Three possibilities for the disposal of phosphogypsum exist:

1. Dumping at sea (currently 10 %)
2. Dumping on land (ca. 88 %)
3. Use as a raw material

Table 1. Typical phosphogypsum analyses (%)

	Production process		
	Dihydrate	Hemihydrate	Hemi-/ dihydrate
CaO	32.50	36.9	32.20
SO_3	44.00	50.3	46.50
P_2O_5	0.65	1.5	0.25
F	1.20	0.8	0.50
SiO_2	0.50	0.7	0.40
Fe_2O_3	0.10	0.1	0.05
Al_2O_3	0.10	0.3	0.30
MgO	0.10		
H_2O	19.00	9.0	20.00
Process yield, t/t P_2O_5	4.9	4.3	4.9

The disposal of gypsum from production facilities near the sea is usually by outflow into the sea. Three aspects for the evaluation of the ecological consequences must be considered: (1) the large amounts of gypsum involved, (2) traces of heavy metals in the gypsum, and (3) changes in the pH of the water due to residual phosphoric and hydrofluoric acids.

The solubility of calcium sulfate in water is 2.4 g/L; in seawater it is ca. 3.5 g/L [42]. Seawater naturally contains 1.6 g/L calcium sulfate, so that, given sufficient tidal and natural movement of the water, a rapid dissolution of the gypsum is observed. SiO_2, Al_2O_3, and organic constituents remain undissolved.

The amount of heavy metals added to seawater due to phosphogypsum is relatively small in comparison to the existing levels; see Tables 2 and 3 [43], [44]. Hence, there is no enrichment of heavy metals in seawater due to this disposal method. The enrichment of the highly toxic heavy metal cadmium in the environment as a result of industrial emission is, however, viewed critically. As the result of a study [45], cadmium will be subject to increasing strict emission limits, for example, in the Netherlands [46]. The

amount of cadmium in phosphogypsum discharged into seawater in the EC in 1979 was 62 t [47].

Table 2. Trace elements in phosphate rock from Morocco and Florida and percentages thereof in phosphogypsum

Element	Morocco, ppm	Florida, ppm	Percentage in phosphogypsum
Cadmium	17	10	15 – 25
Chromium	150	≤ 90	5 – 10
Copper	28	5 – 20	4 – 8
Manganese	10	min. 100	10 – 15
Nickel	16	25 – 75	25 – 50
Uranium	120	200	15 – 20
Zinc	200	10	4 – 8

Table 3. Trace elements in seawater

Element	Concentration, kg/km^3
Chromium	48
Cadmium	120
Copper	3 360
Manganese	2 160
Nickel	2 160
Uranium	3 360
Zinc	11 280

The pH of a 10 % phosphogypsum suspension in water is 3.3. Seawater is buffered by hydrogencarbonate ions and has a pH of 7.9. Therefore, the introduction of even large amounts of slurry cannot cause the pH to drop below the critical level of 5.5 [48]. However, particular attention must be paid to the effects of dumping gypsum in the sea. Permanent surveillance of the effects of the dumping (especially with regard to accumulation) on marine flora and fauna must be carried out.

The dumping of phosphogypsum proceeds in two ways:

1. The gypsum is slurried with water and then pumped to settling vessels. The used water is generally recycled [49]. This disposal method is used especially in the United States. Problems can arise due to the difficulty of evaluating the emission of fluoride and of seepage water containing phosphate and other constituents [50]. Here, too, supervision of the ecological effects should be considered [51]. In Sweden an artificial island has been developed for the dumping of gypsum [52].
2. The dried gypsum (residual water content 20 % min.) is stored in a dump. Another possibility is the filling of empty brown coal mines with protection of the groundwater by confining the dried gypsum with polyethylene sheets [53].

For the disposal of dried phosphate sludge, special isolated disposal sections and treatment of seepage water should be undertaken to avoid environmental risks.

Due to the large amounts of phosphogypsum, the sensitivity with regard to environmental problems, and the scarcity of dumping areas, the following attempts have been made to use it as a raw material [54]:

1. Use as building material (e.g., plasterboard)
2. Cement additive
3. Use in agriculture
4. Use as a filler in the paper industry
5. Recycling to sulfuric acid and cement

The use of large amounts of phosphogypsum as raw material is generally only of economic interest in countries that do not have any natural source of gypsum. Examples are Japan and South Korea, where ca. 3×10^6 t of phosphogypsum is used, mainly in the cement industry. Otherwise only small amounts are utilized. Reasons include the impurities, the water content, the transport costs, the low price of high-purity natural gypsum, and the increased accumulation of gypsum from flue gas desulfurization [55].

Fluoride. Phosphate ore contains ca. 3.6 – 4.1 % fluoride that—depending on the process—is distributed in various amounts among the acid, in the phosphogypsum, the reactor off-gas, and the off-gases from cooling or concentration of the acid (Table 4) [56].

Table 4. Distribution of fluoride in the production of wet phosphoric acid

	Dihydrate process, %	Hemihydrate process, %
Acid	15	12
Phosphogypsum	45	50
Reactor off-gas	5	8
Flash cooler vapor		30
Concentration vapor	35	

Because of its damaging effects on the environment [57] fluoride from off-gases is either precipitated in a basic wash as CaF_2 or in an acid wash as fluorosilicate [58]. Depending on the demand the CaF_2 or fluorosilicate is either utilized or dumped. In Germany the maximum emission values for the off-gas [59] are 50 mg/m^3 for solid particles and 5 mg/m^3 for gaseous fluorine compounds.

Phosphorus Furnace Slag. Eight tonnes of slag are obtained during the production of 1 t of phosphorus by the furnace process. The slag, which consists mainly of calcium silicate, is obtained as lumps or granules.

Typical percentage slag composition is as follows:

P_2O_5	1.0
CaO	48.4
SiO_2	42.5
F	3.3
Al_2O_3	2.5
Fe_2O_3	0.2
MgO	0.6

Na_2O	0.9
K_2O	0.5

Uses of the phosphorus furnace slag are in waste-disposal technology, in road building, and as a filler and insulating material [60].

[Top of Page]

2. Condensed Phosphoric Acids

Gerhard Bettermann

During concentration of aqueous orthophosphoric acid, formation of pyrophosphoric acid can already be detected chromatographically at an orthophosphoric acid concentration of 95 %, corresponding to 68.8 % P_4O_{10}. In liquid, anhydrous orthophosphoric acid, 12.7 % is present as pyrophosphoric acid, which however, reverts to orthophosphoric acid upon crystallization [61]. Upon further concentration, a mixture of straight- chain polyphosphoric acids, $H_{n+2}P_nO_{3n+1}$, is formed initially; eventually, it is converted to a mixture of polymetaphosphoric acids, $H_nP_nO_{3n}$ [62]. The acidity increases during this endothermic condensation process.

The contents of the individual polyphosphoric acids present in a concentrated phosphoric acid mixture for various P_4O_{10} concentrations are compiled in Table 5.

Table 5. Percentage contents of individual polyphosphoric acids in a concentrated polyphosphoric acid mixture as a function of P_4O_{10} content [63]

Total	Number of P atoms in the low-molecular polyacids														Highly	Trimeta	Tetra-
															condensed		meta
P_4O_{10}, %	1	2	3	4	5	6	7	8	9	10	11	12	13	14			
67.4	100.0																
68.7	99.7	0.33															
70.4	96.2	3.85															
71.7	91.0	8.86	trace														
73.5	77.1	22.1	0.79														
73.9	73.6	25.1	1.34														
75.7	53.9	40.7	4.86	0.46													
77.5	33.5	50.6	11.5	2.86	0.74	trace											
79.1	22.1	46.3	20.3	7.82	2.26	1.02	0.34										
80.5	13.8	38.2	23.0	13.0	6.86	3.38	1.67	1.03	0.22								
81.0	12.2	34.0	22.7	14.6	8.42	4.36	2.27	1.41	0.56	trace							
81.2	10.9	32.9	22.3	15.0	9.36	5.41	2.85	1.75	0.97	0.36	0.05						
82.4	7.32	23.0	19.3	15.9	12.3	8.21	5.73	3.89	2.52	1.36	0.91	0.14			trace		
84.0	3.92	11.8	12.7	12.0	10.5	8.97	7.99	6.62	5.63	4.54	3.72	3.03	2.46	1.68	6.63		
85.0	2.28	6.36	7.32	8.01	8.17	7.67	7.22	6.93	6.42	5.89	5.27	4.69	3.99	3.83	16.9		
85.3	1.87	4.73	6.33	6.58	6.66	6.71	6.36	6.11	5.88	5.46	5.07	4.90	4.64	4.38	25.6		
86.1	1.46	2.81	3.74	4.43	4.52	4.77	4.79	4.93	4.67	4.54	4.67	4.63	4.38	4.17	43.5	trace	
87.1	0.83	1.81	2.17	2.53	3.09	3.39	3.46	3.33	3.55	3.47	3.45	3.52	3.26	3.24	61.1	trace	
87.9	0.50	0.82	1.56	1.76	1.72	2.03	2.13	2.26	2.07	2.26	2.06	2.20	1.99	2.30	76.4	0.42	0.11
89.4	1.88	1.52	0.7	0.61	0.62	0.68	0.54	0.71	0.36	1.03	0.98	1.16	1.23	1.37	86.8	1.17	0.41

Diphosphoric acid [2466-09-3], pyrophosphoric acid, $H_4P_2O_7$, M_r 177.98, exists in two crystalline modifications with *mp* 54.3 and 71.5 °C. It is readily soluble in water and is converted to orthophosphoric acid, especially upon boiling or in the presence of strong acids. Pure pyrophosphoric acid cannot be obtained by concentration of orthophosphoric acid or by addition of dry P_4O_{10} to the latter (cf. Table 5). It can be prepared from sodium pyrophosphate by ion exchange, or by reaction of insoluble lead pyrophosphate with hydrogen sulfide.

Unlike monophosphoric acid, aqueous solutions of pyrophosphates do not give a yellow precipitate with silver nitrate but a white one. Compared with metaphosphoric acids, which also form a white precipitate, pyrophosphate differs in that it does not coagulate albumin.

Polyphosphoric and Metaphosphoric Acids. In low-molecular, linear polyphosphoric acids (n = 2 – 15) such as tetraphosphoric acid, $H_6P_4O_{13}$,

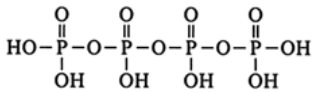

each phosphorus atom bears a strongly acidic OH group. In addition, the two terminal P atoms are each bonded to a weakly acidic OH group. Cyclic metaphosphoric acids, $H_nP_nO_{3n}$, which are formed from low-molecular polyphosphoric acids by ring closure, have a comparatively small number of ring atoms (n = 3 – 8). Each P atom in the ring is bound to one strongly acidic OH group, i.e., all H atoms are equally acidic.

In high-molecular polyphosphoric acids, $H_{n+2}P_nO_{3n+1}$ (n = 16 – 90), the two terminal H atoms exhibit only low acidity and cannot be replaced by metal atoms. The composition of these acids are close to that of cyclic metaphosphoric acids $H_nP_nO_{3n}$. Upon continual concentration they are initially converted to cross-linked chain structures. These are further transformed from a two-dimensional network to three-dimensionally cross-linked, glassy metaphosphoric acid [62], [64].

Production. Polyphosphoric acid can either be produced from wet phosphoric acid by evaporative concentration, or thermally by combustion of elemental phosphorus.

Wet phosphoric acid is concentrated by direct heating with hot gases in reactors lined with carbon bricks [65] or by indirect heating under vacuum [66]. In the former process, the heating gas is usually injected below the acid surface.

Concentration of wet phosphoric acid also improves its transportability. Commercial wet phosphoric acid with ca. 50 % P_2O_5 contains relatively high concentrations of impurities, which deposit on pipes and containers unless during shipment and storage the acid is constantly agitated. Because polyphosphoric acid has a higher lime binding strength, the impurities are held in solution and cannot deposit.

Phosphoric acid plants in which combustion of phosphorus and absorption of P_2O_5 vapors are performed separately are best suited to the thermal production of polyphosphoric acid (see TVA process, TVA Process). An externally water-cooled tower made of stainless steel is used for combustion. A hydration tower, also made of stainless steel, cooled externally with water and internally with recycled acid, is used for the absorption of P_2O_5 mist. The phosphoric acid recycled in the hydration tower can be kept at the concentration of the desired polyphosphoric acid if an appropriately intensive cooling facility is available [67]. A polyphosphoric acid plant differs from an orthophosphoric acid plant by more expensive cooling and circulation equipment because, in contrast to orthophosphoric acid production, water evaporation cannot be used as a contribution to cooling capacity during production of polyphosphoric acid, and the viscosity of polyphosphoric acid is much higher than that of orthophosphoric acid.

In another process, polyphosphoric acid is produced in a tower in which phosphorus is combusted and P_2O_5 mist is absorbed (cf. IG process, IG Process). The tower consists of two sections of different diameters so that a step is formed at the joint, which serves as an acid collection cup. A second collection cup is located below the tower.

The tower walls are protected by two phosphoric acids of different concentrations. In the upper section, polyphosphoric acid containing 76 % P_2O_5 flows from an overflow cup down the tower walls. In the lower section, polyphosphoric acid with 84 % P_2O_5 is sprayed onto the tower walls. Both acid recycle streams are cooled by cooling facilities. It is advantageous to operate this plant in combination with other production units for phosphoric acid. Such a plant can for instance use the flue gases from production of dry P_2O_5 and pass its own flue gases on to the absorption section of the orthophosphoric acid plant [68].

Commercial Products and Uses. The fertilizer industry was the largest area of application of polyphosphoric acids until about 1973. Especially in the United States, large amounts of ammonium phosphate, triple superphosphate, and liquid fertilizers were made from polyphosphoric acids. This application has declined strongly. Polyphosphoric acid is used in organic chemistry as an acidic catalyst and dehydrating agent as well as for cyclization reactions [69].

[Top of Page]

3. Phosphates

Thomas Staffel and Thomas Klein

The salts of phosphoric acid are technically and commercially the most important group of phosphorus compounds besides the acid itself. According to their applications, fertilizer phosphates (⟶ Metals, Surface Treatment) are distinguished from other commercial phosphates discussed here. While the use of phosphates as fertilizers started in the mid-1800s with the introduction of mineral fertilization, other phosphates were used industrially at the turn of the century only sporadically. They attained greater commercial importance since ca. 1930 with the production of condensed phosphates, especially for the detergent industry.

Structure and Nomenclature. All phosphate salts contain PO_4 units as characteristic structural element. The central phosphorus atom is surrounded by a tetrahedral arrangement of oxygen atoms. The tetrahedral symmetry is based on formation of sp^3 hybrid orbitals on the phosphorus atom, which is linked to the four oxygen atoms by sigma bonds. The P – O bond strength is enhanced by π bonding with participation of phosphorus *d* orbitals. Ideal tetrahedral structure with a theoretical O – P – O bond angle of 109°28′ can only be expected for a free PO_4^{3-} ion. In most crystalline phosphates, especially in condensed phosphates, deformation of the PO_4 tetrahedra occurs due to unsymmetrical lattice forces, so that bond angles of 95 – 125° are found. The bond lengths are shorter than expected for a single bond due to the double bond character. Average P – O distances of 154 pm are found.

Polymers with chain, ring, or cross-linked structure are obtained by linkage of several PO_4 tetrahedrons via common oxygen atoms. The tetrahedra are joined exclusively by vertex sharing.

Since polyphosphates are prepared from acidic monophosphates by condensation, they are usually referred to as condensed phosphates. Three types of condensed phosphates are distinguished according to the structure of their anions [62]: polyphosphates with chain structure, cyclic metaphosphates, and cross-linked or ultraphosphates with branched chains and/or rings.

All phosphate salts can be described as combinations of oxides, e.g., $Na_2H_2P_2O_7$ as $Na_2O \cdot H_2O \cdot P_2O_5$. The ratio of cationic oxides (Na_2O, CaO, H_2O) to P_2O_5 is characteristic for the individual structural types. Table 6 shows the structural relations and classifications.

Table 6. Structure and nomenclature of phosphates

Oxide ratio	Group name	General formula	Anion structure	Examples
3	monophosphate (isolated PO_4 tetrahedra)	$M^I_3PO_4$	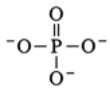	Na_3PO_4 trisodium phosphate $Ca(H_2PO_4)_2$ calcium dihydrogen phosphate
> 1–2	polyphosphates (chain form)	$M^I_{n+2}P_nO_{3n+1}$	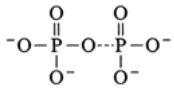	$Na_2H_2P_2O_7$ disodium dihydrogen phosphate $Na_5P_3O_{10}$ pentasodium triphosphate
1	metaphosphates (cyclic)	$M^I_3P_nO_{3n}$	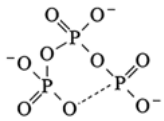	$Na_3(PO_3)_3$ sodium trimetaphosphate $Na_6(PO_3)_6$ sodium hexametaphosphate
<1[a]	cross-linked or ultraphosphates (branched chains or rings)		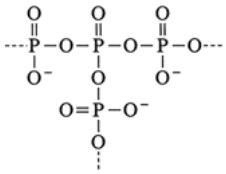	CaP_4O_{10} $Ca_2P_6O_{17}$

[a] Special cases: isopolyphosphate, oxide ratio >1 to 2; isometaphosphate, oxide ratio 1.

The polyphosphates ($M^I_{n+2}P_nO_{3n+1}$) are named according to the IUPAC rules as di-, tri-, tetraphosphates, etc. Frequently, the older designations pyro-, tripoly-, tetrapolyphosphate, etc., are also still used. Moreover, it is customary to distinguish between chromatographically separable, oligomeric phosphates with $n < 12$ and high-molecular polyphosphates with n = 13 to > 10 000. Both glassy and crystalline forms are known in both groups. While mixtures of various chain lengths are always present in glassy phosphates, crystalline oligomeric phosphates are uniform with regard to anion size. It is debatable whether crystalline, high-molecular polyphosphates are also uniform [62].

The oxide ratio in metaphosphates, $M^I_nP_nO_{3n}$, is exactly 1. Polyphosphates with very high molecular masses have practically the same oxide ratio, and are therefore often incorrectly called “metaphosphates” (e.g., sodium hexametaphosphate for Graham's salt). According to the newer nomenclature, the term “metaphosphate” is reserved for condensed phosphates with cyclic anions. So far, metaphosphates are known for n = 3 – 8. There is no evidence for the existence of a dimeric anion with n = 2. Although there are indications for the occurrence of ring structures with n > 8 in some phosphate glasses, they have so far not been isolated. Cross-linked or ultraphosphates contain P atoms which are linked with each other via three O atoms. Additional designations are commonly used in special cases: isometaphosphate when a ring is linked with one or more chains, and isopolyphosphate if only branched chains are present. Compounds with rings linked directly or via chains are more specifically called ultraphosphates. Commercially, cross-linked phosphates are not very important. So far, only a few well-defined examples of this structural type are known.

3.1. Sodium Monophosphates

As a tribasic acid, monophosphoric acid (orthophosphoric acid) forms three series of sodium salts, i.e., mono-, di-, and trisodium monophosphate. Besides the anhydrous compounds, defined hydrates and some double salts are known. Table 7 lists the monophosphates characterized in the $Na_2O – P_2O_5 – H_2O$ system.

Table 7. Sodium monophosphates

Compound	Density, g/cm^3	*mp*	Crystal system
NaH_2PO_4	2.36	200 *	
$NaH_2PO_4 \cdot H_2O$	2.04	100	orthorhombic
$NaH_2PO_4 \cdot 2\ H_2O$	1.91	60	orthorhombic

$2\ NaH_2PO_4 \cdot Na_2HPO_4 \cdot 2\ H_2O$			
$2\ NaH_2PO_4 \cdot Na_2HPO_4$			
$NaH_2PO_4 \cdot H_3PO_4$		127 – 131	
Na_2HPO_4	1.10		monoclinic
$Na_2HPO_4 \cdot 2\ H_2O$	2.07	95	orthorhombic
$Na_2HPO_4 \cdot 7\ H_2O$	1.68	48	monoclinic
$Na_2HPO_4 \cdot 8\ H_2O$			
$Na_2HPO_4 \cdot 12\ H_2O$	1.52	48	monoclinic
Na_3PO_4, low temp.	2.63	700 **	hexagonal
Na_3PO_4, high temp.	2.54	1583	cubic
$Na_3PO_4 \cdot {}^{12}\ H_2O$		180	monoclinic
$Na_3PO_4 \cdot 6\ H_2O$		110	
$Na_3PO_4 \cdot 8\ H_2O$		86	triclinic
$Na_3PO_4 \cdot 12\ H_2O$	1.64	73 – 76 (decomp.)	hexagonal

* Decomposition.
** Transformation.

Figure 17 shows the solubility of sodium monophosphates as a function of temperature. As salts of a tribasic acid whose dissociation decreases greatly from the first to the third stage ($pK_1 = 2.15$, $pK_2 = 7.1$, $pK_3 = 12.4$), alkali-metal monophosphates exhibit buffer properties.

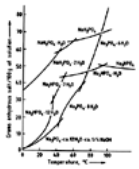

Figure 17. Solubility of sodium monophosphates [70]

Sodium monophosphates are usually produced by neutralization of phosphoric acid with soda ash or caustic soda. The latter is predominantly used in Germany, while in the United States, the less expensive soda ash prevails. The basicity of sodium carbonate is insufficient for the formation of trisodium phosphate, so that caustic soda must be used in this step. The phosphates crystallize from the solutions as hydrates and are separated by centrifugation. Anhydrous salts are obtained by dehydration in rotary dryers or directly from the solutions by spray drying or in rotary kilns.

Both thermal phosphoric acid and the cheaper wet phosphoric acid are used as starting materials. The very pure thermal phosphoric acid is generally used in the production of food-grade phosphates.

Monosodium dihydrogenmonophosphate [7558-80-7], monosodium phosphate, NaH_2PO_4, M_r 119.98, crystallizes in anhydrous form and as the mono- and dihydrates. The pH of a 1 % aqueous solution is 4.5. The anhydrous salt decomposes upon heating from 169 °C to 240 °C, with elimination of water and formation of $Na_2H_2P_2O_7$.

Thermal phosphoric acid is used preferentially for production of monosodium phosphate. Wet phosphoric acid is less suitable because the impurities are only partially precipitated in the neutralization with soda ash or caustic soda up to the first stage, so that additional recrystallization would be required to purify the phosphate. The dihydrate and the anhydrous salt are the usual commercial forms. Due to its acidity, monosodium phosphate is used as phosphating agent for metal surface treatment and as component of acidic detergents. Mixtures of mono- and disodium phosphate are used as buffers in the pH range 5 – 8, especially in textile finishing processes. In water conditioning, monosodium phosphate is frequently added for correction of the alkalinity of water purified by ion exchange. Monosodium phosphate serves as a mineral supplement for feeds and as a pH stabilizer for soups and juices in the food area. Monosodium phosphate is used as raw material for the production of disodium diphosphate and higher condensed phosphates.

Disodium hydrogenmonophosphate [7558-79-4], disodium phosphate, Na_2HPO_4, M_r 141.97, crystallizes in the anhydrous form as well as with 2, 7, 8, and 12 H_2O. The crystal structure is composed of HPO_4^{2-} anions in a hexagonal close-packed arrangement, in which Na^+ cations occupy octahedral and tetrahedral interstices.

Both thermal and purified wet phosphoric acid are used in production. In a process developed in the 1980s, acidic phosphate solution ($CaHPO_4$ in the presence of H_3PO_4) is reacted with $NaHSO_4$, whereby purification occurs through precipitation of gypsum [71]. Neutralization with NaOH is carried out in a second stage. Salts are obtained from the neutralized solutions by evaporation, crystallization, and centrifugation. The desired hydrate is obtained by appropriate temperature control. The anhydrous product is also produced by dehydration of the solution in a spray drying tower. Careful temperature control is required to avoid formation of tetrasodium diphosphate.

Anhydrous Na_2HPO_4, $Na_2HPO_4 \cdot 2\ H_2O$, and $Na_2HPO_4 \cdot 12\ H_2O$ are the commercial forms. Disodium phosphate is blended with monosodium phosphate to prepare buffer mixtures. Other applications are treatment of boiler feedwater (see Trisodium Monophosphate Uses.), addition to detergents and cleansers, and preparation of mineral feedstuff supplements. In the food industry, disodium phosphate is used in the production of evaporated milk and other concentrated dairy products such as milk powder and cream to prevent coagulation on heating. It is also used as a component of instant desserts and puddings. Another application of Na_2HPO_4 is for silk weighting in the textile industry. Disodium phosphate is used as starting material for the production of tetrasodium diphosphate and, together with monosodium phosphate, for the production of higher condensed phosphates.

Trisodium monophosphate [7601-54-9], trisodium phosphate, Na_3PO_4, M_r 163.95, crystallizes in the anhydrous form and with 1/2, 6, 8, or 12 H_2O. The pH of a 1 % aqueous solution is ca. 11.9. The dodecahydrate always crystallizes from aqueous solutions with an excess of NaOH. Products with compositions ranging from $Na_3PO_4 \cdot 12\ H_2O \cdot ¼$ NaOH to $Na_3PO_4 \cdot 12\ H_2O \cdot {}^1/_7$ NaOH are obtained, depending on the composition of the starting solution and the temperature. The stoichiometric compound $Na_3PO_4 \cdot 12\ H_2O$ is unstable in aqueous solution and can therefore only be obtained by isothermal hydration of the other hydrates, which are free of NaOH. Depending on composition, the dodecahydrate dissolves at 70 – 74 °C in its water of hydration.

In addition to NaOH, several other sodium compounds such as NaOCl, NaCl, $NaNO_3$ form double salts with trisodium phosphate [72]. Of these, the addition compound with sodium hypochlorite (chlorinated trisodium phosphate) is commercially important.

Production. Trisodium phosphate is mainly produced as the dodecahydrate by complete neutralization of phosphoric acid with caustic soda or soda ash and subsequent crystallization. Since neutralization with soda ash does not proceed beyond the stage of the disodium salt, caustic soda must be added to complete the reaction. Frequently, mother liquors of mono- or disodium phosphate production or crystalline disodium phosphates are used as starting materials. In another process, trisodium phosphate is obtained by digestion of the phosphate-rich neutralization sludges, which accumulate in the purification of wet phosphoric acid by precipitation, with caustic soda and with addition of water glass [73]. In this process, aluminum and iron phosphates are converted to $Fe(OH)_3$ and $Na_2Al_2SiO_6$ and separated from the Na_3PO_4 solution by filtration. On a commercial scale, it is also possible to react $CaHPO_4$ with H_2SO_4 to form $CaSO_4$ and H_3PO_4. Caustic soda can be added to the latter to give $Na_3PO_4 \cdot 12\ H_2O$ [74].

For production of the dodecahydrate, the most widely used form, the solutions are concentrated. The solid salt is obtained in the form of flakes on cooled rollers or as powder in spray towers. Cooling the solution gives the salt as needles, which are centrifuged and dried with hot air. The anhydrous salt is mostly produced from the molten dodecahydrate by spray drying.

The dodecahydrate tends to cake during storage. In order to obtain a free-flowing and storage-stable product it is mixed with hydrophilic materials such as magnesium salts or silicates [75].

Uses. The most widely used commercial product is the dodecahydrate, in the form of flakes, needles, or spray dried powder. The anhydrous salt is marketed as a powder. Due to hydrolysis, trisodium phosphate reacts strongly alkaline in aqueous media, and thus causes saponification of fats.

$$PO_4^{3-} + H_2O \rightleftharpoons HPO_4^{2-} + OH^-$$

Due to its fat- and dirt-solving properties it is used as an industrial cleansing agent for machine parts, occupational clothing, milk cans, dishes, and bottles. Trisodium phosphate is also used for surface treatment of aluminum to obtain a gloss effect.

Trisodium phosphate is the standard agent for water softening, especially of feedwater for steam boilers. Ions such as Ca^{2+} or Mg^{2+} react with trisodium phosphate to form water insoluble phosphates as voluminous, sludgy precipitates, which are removed by sedimentation or filtration before the purified water enters the boiler:

$$3\,Ca(HCO_3)_2 + 2\,Na_3PO_4 \rightarrow Ca_3(PO_4)_2 + 6\,NaHCO_3$$

$$3\,MgSO_4 + 2\,Na_3PO_4 \rightarrow Mg_3(PO_4)_2 + 3\,Na_2SO_4$$

Trisodium phosphate is also used to remove residual hardness from water treated by ion exchange or precipitation processes. An advantage is that the phosphate precipitates do not form deposits and are easily removed. Moreover, a small excess of phosphate in the boiler water decreases corrosion by formation of a strongly adhering protective layer of iron phosphate. For complete softening, 14 g $Na_3PO_4 \cdot 12\,H_2O$ is needed for 1° d carbonate hardness (German hardness scale), and 42 g for 1° d non-carbonate hardness per cubic meter of water.

Chlorinated trisodium phosphate, $Na_3PO_4 \cdot ¼\,NaOCl \cdot 11\,H_2O$; the pH of a 1 % aqueous solution is ca. 11.8. The addition compound of trisodium phosphate with sodium hypochlorite is usually produced by the same process as trisodium phosphate itself but with replacement of a part of the required caustic soda by NaOCl. Special processes for production of a storage-stable product are described in the patent literature [76]. Due to its active chlorine content, chlorinated trisodium phosphate has bleaching and bactericidal properties in addition to the fat- and dirt-solving effect resulting from its high alkalinity. Therefore, it is used as a disinfecting cleanser. The product has gained considerable importance especially in the United States where it is used in detergents for automatic dishwashers and in the sanitary area. However, in Europe it is only of minor importance.

Sodium monofluorophosphate [10163-15-2], Na_2PO_3F, M_r 144, is a white odorless powder. Its solubility in water is 25 g per 100 g H_2O at 25 °C. The $NaPO_3$ – NaF phase diagram (Fig. 18) shows polymorphic transformations of Na_2PO_3F at 370 °C and 410 °C. The salt melts congruently at 635 °C; prolonged heating of Na_2PO_3F gives pyrophosphate. Hydrolysis occurs in the presence of moist air or in solution:

$$Na_2PO_3F + H_2O \rightleftharpoons HF + Na_2HPO_4$$

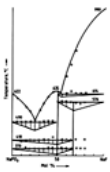

Figure 18. Phase diagram of the $NaPO_3$ – NaF system [77]

Structure. The NMR spectrum of sodium monofluorophosphate shows that the monofluorophosphate anion has the structure of the tetrahedral orthophosphate group in which an oxygen atom is substituted by a fluorine atom [78].

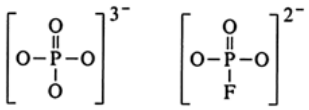

The PO_3F^{2-} ion is isoelectronic with SO_4^{2-} and hence possesses similar chemical properties to the sulfate group.

Production. There are two main processes used in the commercial production of Na_2PO_3F. The most common process involves the reaction of sodium fluoride with sodium (meta)phosphates by melting at the eutectic point [79]; the reaction is exothermic:

$$NaF + NaPO_3 \rightleftharpoons Na_2PO_3F$$

Since the molten mixture of Na_2PO_3F, NaF, and $NaPO_3$ can react with many metals and metal oxides, the choice of reactor material is important. Platinum, platinum alloys, and glass – carbon composites are usually used for the reactor walls [80-82]. Other reactor materials used are graphite [81], silver, and iron.

With oligophosphates the rate of reaction increases with increasing chain length. Cyclic tri- and tetrametaphosphates react even faster. Maddrell's salt can also be used to prepare Na_2PO_3F, as can a mixture of tetrasodium pyrophosphate and phosphorus pentoxide [80], [83]:

$$Na_4P_2O_7 + 4\,NaF + P_2O_5 \rightleftharpoons 4\,Na_2PO_3F$$

The second process for the production of Na_2PO_3F involves the reaction of tetrasodium pyrophosphate or disodium phosphate with preheated hydrogen fluoride at 200 – 460 °C in a nickel or aluminum reactor [84], [85]:

$$Na_4P_2O_7 + 2\,HF \rightleftharpoons 2\,Na_2PO_3F + H_2O$$

$$Na_2HPO_4 + HF \rightleftharpoons Na_2PO_3F + H_2O$$

Uses. Sodium monofluorophosphate is used in applications where fluoride ions are needed for human physiological processes. It has lower toxicity than sodium fluoride [187]. Sodium monofluorophosphate is mostly used in toothpastes, where it serves as a carrier for the fluoride ion, which is a caries prophylactic [86]. It performs a similar function in pharmaceuticals for the treatment of osteoporosis. In special concrete materials, sodium monofluorophosphate is used as a setting retarder [188], [189].

Analysis. Detection of PO_3F^{2-} is based on the fact that Ag_2PO_3F, unlike AgF, is insoluble in water. Hence determination of PO_3F^{2-} can be carried out gravimetrically or complexometrically. Detection of PO_3F^{2-} by ion chromatography is also possible [86].

3.2. Condensed Sodium Phosphates

Condensed sodium phosphates are formed by condensation of two or more hydrogenphosphate units to form chain or cyclic phosphates, with elimination of water, usually thermally. For example, tetrasodium diphosphate (commercial name, sodium pyrophosphate) is formed from two moles of Na_2HPO_4 upon heating:

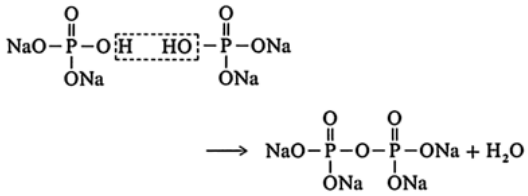

The composition and structure of the products formed depend on temperature and the $Na_2O : P_2O_5$ ratio. Figure 19 shows the possible products by means of the phase diagram.

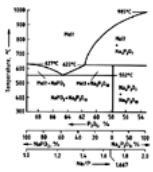

Figure 19. Phase diagram of the Na_2O – P_2O_5 system [87]

3.2.1. ***Properties***

Condensed sodium phosphates are all water soluble except for high-molecular, crystalline salts such as Maddrell's and Kurrol's salts. The acids from which polyphosphates are derived contain two types of OH groups that differ greatly in their degree of dissociation. Each P atom bears a strongly acidic OH group, while the terminal P atoms additionally have a weakly acidic OH group. Since with increasing chain length the number of strongly acidic groups becomes very large compared with the two weak terminal groups, higher condensed polyphosphates react like salts of a strong, monobasic acid. The buffer capacity, strongly pronounced in monophosphates, decreases with increasing chain length and is absent in higher condensed products such as Graham's salt. Accordingly, the pH of 1 % aqueous solutions of “neutral salts” decreases from trisodium phosphate (11.5) through tetrasodium phosphate (10.2) and pentasodium phosphate (9.7) to Graham's salt (6.0). Due to their cyclic structure, metaphosphates contain no terminal acid groups. Consequently, they react in aqueous solution like salts of strong, monobasic acids.

Hydrolytic Decomposition. In aqueous solution, polyphosphates decompose hydrolytically in reversal of the condensation reaction, eventually forming monophosphates [88]. The decomposition rate is strongly dependent on the temperature and pH of the solution. It usually increases with increasing temperature and decreasing pH. At pH 7 and 25 ° C the half-life for the degradation into shorter chains has been calculated to be in the range of 3.5 years [190].

The decomposition is influenced by metal ions. Alkali metal, alkaline earth metal, Al^{3+}, Mn^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+} and Zn^{2+} cations retard the hydrolysis of phosphates in acid media. The efficiency increases from univalent to trivalent ions and within the first group of the periodic table from bottom to top. In alkaline solution, however, metal cations accelerate the hydrolysis of phosphates and the catalytic effectiveness decreases with increasing ionic radius [191]. The degree of acceleration increases in the following order :

$$Al^{3+} > Mg^{2+} > Ca^{2+} > Sr^{2+} > Ba^{2+}$$

Sodium triphosphate in aqueous solution has a half-life of more than a year at pH 8 and room temperature. However, its half-life is in the range of minutes at pH 3 and 100 °C. Under neutral to alkaline conditions, decomposition mainly involves successive cleavage of monosphosphate units from the chain ends, whereas in strongly acidic medium, internal chain cleavage, with formation of low molecular mass fragments, also occurs. In addition to cleavage of phosphate units from the chain ends, a second decomposition reaction occurs for polyphosphates of higher molecular mass, which is accelerated in acidic solution and in which metaphosphates are formed initially [62].

In neutral aqueous solutions at room temperature metaphosphates are also very stable. At elevated temperature and low pH they decompose rapidly, giving initially the corresponding polyphosphates, which, however, themselves undergo rapid hydrolysis to give monophosphates. The metaphosphate rings are also cleaved practically quantitatively in alkaline solution, albeit at a relatively slow rate.

In acidic solution, linear oligomers are hydrolyzed faster than cyclic oligomers, except for cyclotriphosphate. For the linear phosphates, the rate gradually increases with increasing degree of polymerization. In alkaline solution, large differences are observed between the rates for cyclic (fast hydrolysis) and linear oligomers (slow). However, with increasing degree of polymerization the differences become smaller. The first-order hydrolysis rate constants for cyclic and linear penta- and heptaphosphate at 50 °C in 0.1 mol/L HCl are 0.128 h^{-1} (*cyclo*-P_5), 0.764 h^{-1} (P_5), 0.136 h^{-1} (*cyclo*-P_7), 1.54 h^{-1} (P_7), and at the same temperature in 0.5 mol/L NaOH: 0.0166 h^{-1} (*cyclo*-P_5), 6.7×10^{-3} h^{-1} (P_5), 7.4×10^{-3} h^{-1} (*cyclo*-P_7), 0.0234 h^{-1} (P_7) [192].

Tri- and tetrametaphosphate can be converted completely to the corresponding polyphosphates with caustic soda because the latter are comparatively stable under these conditions. However, a similar preparation of higher polyphosphates is not possible because the stability of metaphosphates increases greatly with ring size, so that the polyphosphates are largely decomposed before all the metaphosphate is cleaved [62], [89].

The hydrolytic degradation is easily monitored by ^{31}P NMR spectroscopy or ion chromatography. The ^{31}P NMR spectra of polyphosphates show two major peaks, with chemical shifts in the range of 5 – 10 and 15 – 20 ppm. The high-field signal can be attributed to the end-group phosphate units, and the other signal to the middle groups [193]. The middle to end group integral ratio is related to the average degree of polymerization, which decreases during hydrolysis.

Ion chromatography of polyphosphate solutions can be performed with a anion-exchange column (Dionex AS7) and gradient elution. The eluent is a NaOH/NaCl solution whose content of NaCl is increased during elution. Detection by hydrolysis to orthophosphoric acid and post-column derivatization (as molybdatovanadatophosphate complex) has proved best [194]. Each polyphosphate species (e.g., pentaphosphate, hexaphosphate, etc., up to P_{50}) gives a peak whose integral corresponds to the concentration. Hence this method allows a detailed look at the alteration of the concentration of distinct molecules. Figure 20 gives an example for a highly condensed polyphosphate.

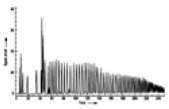

Figure 20. Ion chromatograph of a highly condensed polyphosphate $Na(PO_3)_n$ with a P_2O_5 content of 68 % [195]

Cation-Binding Capacity. The cation-binding capacity of polyphosphates is of considerable commercial importance. This term denotes the capability to bind cations of multivalent metals, especially of Ca, Mg, and Fe, and to hold them in solution such that they can no longer be detected with the usual precipitation reagents. For example, calcium ions are not precipitated by oxalate in an aqueous solution containing sodium triphosphate. Numerous investigations have shown that the effect of polyphosphates is not due to the formation of well-defined, soluble complexes but that they behave like soluble ion exchangers [90], [91]. In aqueous solution, polyphosphates are at most 30 % dissociated [92], although all OH groups of the corresponding polyacids, apart from the two terminal groups, can be titrated as strong acids.

Bonding of the cation is primarily attributed to electrostatic attraction by the negatively charged polyanions. In addition, formation of relatively stable chelates, in which the cations are bonded to several PO_3 groups in the flexible polyphosphate chain, also must be assumed [62]. Since multivalent cations are bound much more strongly than the monovalent ions of alkali metals, the latter can readily be exchanged for alkaline earth or heavy metal ions. This exchange and binding capacity is used for numerous purposes to mask and inactivate interfering cations.

Various methods are used to determine the exchange capacity. The lime-binding capacity or lime-binding value denotes the amount of CaO kept in solution by a specific amount of alkali-metal polyphosphate in the presence of reagents which precipitate Ca^{2+}. For the determination, a predetermined amount of polyphosphate is titrated with $CaCl_2$ in the presence of sulfate ions to the onset of turbidity. Alternatively, the end point is determined by formation of sparingly soluble calcium polyphosphate without addition of calcium-precipitating anions [93].

The lime-binding capacity of polyphosphates increases with increasing chain length up to n = 8. For higher degrees of condensation it remains essentially unchanged. Figure 21 shows the amount of phosphate which is necessary to prevent precipitation of an insoluble calcium compound (e.g., calcium oxalate) as a function of chain length. The lime-binding capacity is also dependent on concentration, pH, temperature, and the presence of electrolytes [94]. In contrast to polyphosphates, cyclic metaphosphates exhibit only minor exchange capability toward multivalent cations.

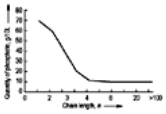

Figure 21. Lime-binding capacity of polyphosphates as a function of chain length

The figure shows the amount of polyphosphate which is necessary to complex all calcium ions in 10 L of water with a hardness of 10° dH

Colloidal Behavior. Many applications of polyphosphates are based on colloidal effects. These are secondary effects because polyphosphates themselves exhibit no colloidal properties in pure, aqueous solution. Essentially, the effect is due to adsorption of multiply charged polyphosphate anions on suspended or emulsified particles, electrical charging of the latter, and consequent change of the physicochemical characteristics [95].

The dispersive action of polyphosphates is one of the most prominent effects. Viscosity reduction, which always occurs concurrently, is directly related to this effect. For

instance, stiff, tenacious pastes such as blanc-fixe ($BaSO_4$) can be converted to stable, free-flowing suspensions by addition of small amounts of Graham's salt. Kaolin suspensions of high solids content and low viscosity can be prepared with small amounts of polyphosphate.

The improvement of the soil antiredeposition capacity in the laundry process by polyphosphates can also be regarded as suspension stabilization. Although polyphosphates themselves are not surface-active and thus are not really emulsifiers, they increase the surface activity of wetting agents. Presumably, the effect is due to a dispersion of the surface-active substances.

The peptization of proteins is another colloid chemical effect. It plays an important role in the production of processed cheese and in meat processing. Protein is partially digested by polyphosphate and converted to a sol.

In many cases the dispersing effect of polyphosphates is due to adsorption and is supplemented by stabilization due to elimination of interfering factors (calcium binding).

3.2.2. *Sodium Diphosphates*

The following sodium diphosphates (pyrophosphates) are known as crystalline compounds [96]:

$NaH_3P_2O_7$	$Na_3HP_2O_7 \cdot H_2O$
$Na_2H_2P_2O_7$	$Na_3HP_2O_7 \cdot 9\ H_2O$
$Na_2H_2P_2O_7 \cdot 6\ H_2O$	$Na_4P_2O_7$
$Na_3HP_2O_7$	$Na_4P_2O_7 \cdot 10\ H_2O$

Of these, only disodium diphosphate and tetrasodium diphosphate are produced on a large scale.

Disodium dihydrogendiphosphate [7758-16-9], disodium diphosphate, acidic sodium pyrophosphate, $Na_2H_2P_2O_7$, M_r 221.97, *d* 2.31. Its solubility in water is 13 g $Na_2H_2P_2O_7$/100 g H_2O at 20 °C, and 20 g at 80 °C. The pH of a 1 % aqueous solution is 4.1. The usual commercial product is the anhydrous, nonhygroscopic salt in powder form. The hexahydrate, $Na_2H_2P_2O_7 \cdot 6\ H_2O$, *d* 1.85, crystallizes from aqueous solution below 27 °C. Above this temperature, it is converted to the anhydrous form.

Disodium dihydrogendiphosphate is produced from sodium dihydrogenmonophosphate by heating at 200 – 250 °C:

$$2\,NaH_2PO_4 \rightarrow Na_2H_2P_2O_7 + H_2O$$

Careful control of reaction temperature and time is required to achieve complete conversion of the monophosphate and to avoid formation of higher condensed phosphates. Reaction times of 1 – 6 h at 215 – 245 °C are used [97]. The degree of condensation can be influenced by regulating the partial pressure of water vapor [98].

Usually, monophosphate is converted to diphosphate continuously in rotary kilns. Smaller plants also use batch processes with plate dryers or electrically heated pans.

A monophosphate solution made from phosphoric acid and caustic soda in situ is used in the so-called “return” process. The crystal slurry obtained by evaporation is combined with recycled material in a twin screw and charged to a rotary dryer [99].

In another process, NaH_2PO_4 solution is spray dried at 140 – 195 °C, giving 10 – 40 % $Na_2H_2P_2O_7$. Subsequent calcination at 200 – 220 °C gives pure $Na_2H_2P_2O_7$ [100].

Disodium diphosphate is used as a (tropically stable) acid carrier in baking powder, for improvement of flow properties in flour, for pH regulation, and in dental care products for prevention of tartar formation.

Tetrasodium diphosphate [7722-88-5], tetrasodium pyrophosphate, neutral sodium pyrophosphate, $Na_4P_2O_7$, M_r 265.91, *d* 2.50, *mp* 985 °C. Its solubility in water is 5 g $Na_4P_2O_7$/100 g H_2O at 20 °C and 35 g at 80 °C. The pH of a 1 % aqueous solution is 10.4. The usual commercial product is the anhydrous salt in the form of a white, nonhygroscopic powder. However, in contact with moist substances it acts as a drying agent with formation of $Na_4P_2O_7 \cdot 10\ H_2O$.

The decahydrate, $Na_4P_2O_7 \cdot 10\ H_2O$, *d* 1.83, *mp* 79.5 °C (with loss of H_2O), is also marketed but on a much smaller scale. It crystallizes from aqueous solution below the melting point. Anhydrous $Na_4P_2O_7$ is known in five different modifications with transition temperatures between 400 and 545 °C. Only the low-temperature form is stable below 400 °C.

Tetrasodium diphosphate is produced by calcination of disodium hydrogenphosphate at 300 – 900 °C:

$$2\,Na_2HPO_4 \rightarrow Na_4P_2O_7 + H_2O$$

In the two-stage process, a concentrated monophosphate solution is dehydrated to give anhydrous disodium hydrogenphosphate in a spray drying tower and subsequently condensed to diphosphate in a rotary kiln. For improvement of the heat economics, the hot off-gases from the second stage can be used for spray drying. In another process, phosphoric acid and caustic soda are metered separately into a spray drying tower in order to utilize the heat of neutralization [101].

Both rotary kilns and spray drying towers are used in the single-stage production of $Na_4P_2O_7$. In the rotary kiln process, the phosphate solution is sprayed onto a bed of already formed, solid monophosphate (return process). A single-stage spray tower process is described below (see Pentasodium Triphosphate Properties.). Generally, the processes for production of sodium triphosphate can also be used for sodium diphosphate.

Tetrasodium diphosphate is used in detergents and cleansers, in metal surface treatment, and in oil well drilling fluids for viscosity adjustment. In the food area, $Na_4P_2O_7$ is used as a component of cold puddings and ice cream for regulating the consistency. The stabilization of toothpastes based on calcium hydrogenphosphate is another important application.

3.2.3. *Pentasodium Triphosphate*

Pentasodium triphosphate [7758-29-4], sodium triphosphate, sodium tripolyphosphate, is one of the commercially most significant phosphorus compounds.

Properties. Pentasodium triphosphate, $Na_5P_3O_{10}$, M_r 367.93; solubility of the hexahydrate in water, 14.6 g $Na_5P_3O_{10}$/100 g H_2O at 20 °C, 23.7 g at 80 °C; pH of a 1 % aqueous solution, 9.7.

Pentasodium triphosphate exists in two anhydrous modifications, i.e., the low-temperature (LT) form (*d* 2.57, monoclinic) and the high-temperature (HT) form [*d* 2.62, monoclinic, *mp* 622 °C (decomp.)], and as the hexahydrate, $Na_5P_3O_{10} \cdot 6\ H_2O$ (*d* 2.12, triclinic). In the technical literature, the HT form is often referred to as form II, and the LT form as form I. The LT form is converted completely to the HT form on heating to 500 °C. However, the HT form is transformed to the LT form only very slowly on tempering below 400 °C (transition temperature, 417 ± 8 °C [102]). At 622 °C the anhydrous salt decomposes to a melt and crystalline $Na_4P_2O_7$ (see also Fig. 19). Above 865 °C a homogeneous melt is obtained, which on rapid cooling yields a glassy product. This can be converted to the crystalline modifications by heating under specific conditions [103], [104].

The two modifications differ in their reaction with water. Both forms are metastable in water and undergo complete conversion to the hexahydrate in an exothermic reaction. However, the hexahydrate is formed much faster from the HT modification than from the LT form. The anhydrous modifications are more soluble than the hexahydrate. Therefore, on dissolution a concentration of ca. 32 % $Na_5P_3O_{10}$ is attained initially which is considerably higher than that of a solution in equilibrium with the hexahydrate (ca. 13 % $Na_5P_3O_{10}$). This occurs especially on dissolution of the LT form, from which the hexahydrate is formed at a slower rate.

On heating to 100 – 120 °C, the hexahydrate hydrolyzes to give mono- and diphosphate:

$$Na_5P_3O_{10} \cdot 6\,H_2O \rightarrow Na_3HP_2O_7 + Na_2HPO_4 + 5\,H_2O$$

$$Na_3HP_2O_7 + Na_2HPO_4 \rightarrow Na_4P_2O_7 + NaH_2PO_4$$

On further heating to 140 – 220 °C the hydrolyzed products recondense to form anhydrous $Na_5P_3O_{10}$ in the LT modification. Commercial products contain a mixture of the HT and LT forms, depending on application and processing requirements. Formerly products with a high content of the LT form were preferred for the manufacture of detergent powders, the major application of sodium triphosphate. The reason was that the rapidly hydrated HT form tends to cake during the production of detergent slurries and is thus

more difficult to process. Later it was recognized how much the hydration properties of the triphosphate are also affected by the content of hexahydrate crystal nuclei. Now, the detergent industry tends not to specify certain contents of the two modifications but rather overall hydration characteristics measurable by the heat of hydration as a function of time [105].

Production. Pentasodium triphosphate is usually produced from a mixture of two moles of disodium phosphate and one mole of monosodium phosphate:

$$2\,Na_2HPO_4 + NaH_2PO_4 \rightarrow Na_5P_3O_{10} + 2\,H_2O$$

The reaction is carried out in the temperature range 300 – 550 °C, whereby the reaction temperature determines the LT/HT form ratio. For the production of triphosphate with a minimum content of $Na_4P_2O_7$ and higher condensed polyphosphates, the monophosphates must be present in precise stoichiometric ratio in a homogeneous, finely crystalline mixture. In most cases, a stoichiometrically adjusted monophosphate solution, obtained by neutralization of phosphoric acid with caustic soda or soda ash, is used as starting material. Both single-stage and two-stage processes are used for dehydration of the solution and subsequent condensation to triphosphate.

In the two-stage processes (Fig. 22), the solution is dehydrated in the first stage to give an anhydrous monophosphate mixture. Frequently, partial condensation with formation of diphosphate already occurs in this stage. The actual condensation to triphosphate then follows in the second stage.

Figure 22. Two-stage process for production of pentasodium triphosphate [106]

a) Receiver; b) Heavy-duty pump; c) Combustion chamber; d) Spray tower; e) Wet dust removal from flue gas; f) Sieve; g) Mill; h) Pneumatic conveyor; i) Silo; j) Rotary kiln; k) Flue gas purification; l) Condenser; m) Bucket chain; n) Sieve; o) Bagging scale

Generally, spray dryers are used for dehydration of the monophosphates, and rotary kilns for the conversion to triphosphate. The required energy is supplied by hot gases from a combustion chamber or by an oil or gas flame in the rotary kiln.

A fluidized-bed reactor has also been recommended for the conversion of dried monophosphates to triphosphate [107]. In another process rotary kilns are used in both stages, and a granular end product is obtained by grinding the intermediate product and adding water before calcination [108].

In the single-stage process, rotary kilns or spray drying towers are used. In the Hoechst-Knapsack process (Fig. 23), the monophosphate solution is sprayed at 1 – 2 MPa [109] into a stainless-steel spray tower [110]. Burners, arranged concentrically around the nozzle, generate a flame zone directed conically towards the center of the tower. The sprayed solution moves downward cocurrent with the burner gases and is rapidly dehydrated and converted to triphosphate. Most of the $Na_5P_3O_{10}$ is collected in the tower cone and discharged. Finer particles are separated from the off-gas by cyclones. The off-gas is then further purified by a liquid washing.

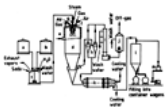

Figure 23. Production of pentasodium triphosphate by the Hoechst-Knapsack process [106]

a) Caustic soda; b) Phosphoric acid; c) Neutralization; d) Spray tower; e) Cooling jacket; f) Burner; g) Spray nozzle; h) Cyclone; i) Cooling pipe; j) Wash tower; k) Rotary cooling drum; l) Product silo; m) Bag filling and weighing machine

Free-flowing, powdery products are obtained, with triphosphate contents of up to 98 %. The ratio of LT/HT form can be varied over a wide range by controlling the temperature and the water content in the gas phase. Moreover, a particular crystalline form can be obtained by simultaneously spraying seed crystals of the desired modification [111]. Coarse $Na_5P_3O_{10}$ particles (> 0.15 mm) can also be prepared by this seed crystal method. The required crystallites are taken from the residues of the dust filter bag because these flue dusts have favorable spray and suspension characteristics.

In the FMC process, the monophosphate solution is sprayed upwards from nozzles at the bottom of the tower [112]. The particles move upward to the flame zone at the top of the tower and are then carried downward with the hot gas stream.

Single-stage rotary kiln processes are carried out both according to the cocurrent and countercurrent principle. The monophosphate solution is sprayed through a flame directed into the rotary kiln and onto a hot bed of previously calcined material [113]. The product then moves cocurrently with the hot combustion gases to the discharge port and thus is posttempered at a reduced temperature. Advantages of this mode of operation are the avoidance of caking and the possibility to produce a product with a high content of the LT form. The high energy consumption is, however, a disadvantage.

A countercurrent rotary kiln process is described by FMC [114]. The material is conveyed countercurrently to the heating gases, which are generated in a combustion chamber outside of the rotary kiln. This process permits a substantially better heat utilization than the cocurrent process. However, the frequent occurrence of caking at the charging point is a disadvantage.

Numerous suggestions for modification of the commercial processes are found in the patent literature. They are aimed at the preparation of pentasodium phosphate with special application properties and concern especially production of the low-temperature modification [115], [116], attainment of low bulk densities [117], and increase of abrasion resistance [118]. Partially hydrated products are prepared for improvement of the solubility characteristics [119], [120].

Sodium triphosphate hexahydrate is produced by hydration of anhydrous $Na_5P_3O_{10}$. It is important that too high a temperature rise is avoided in the exothermic hydration reaction because the triphosphate anion is cleaved hydrolytically by the water of hydration at elevated temperatures. In the Hoechst-Knapsack process an excess of water is sprayed onto $Na_5P_3O_{10}$ in a rotary kiln. The excess water is evaporated by simultaneously blowing air over the product, dissipating part of the heat of hydration and maintaining the product temperature at 60 – 70 °C [121]. In the Monsanto process crushed ice is used for the hydration of $Na_5P_3O_{10}$ to the hexahydrate [122].

3.2.4. *Higher Condensed Sodium Polyphosphates*

If mixtures of NaH_2PO_4 and Na_2HPO_4 in a ratio > 1 : 2 are heated until the water of constitution is released, a mixture of polyphosphates of different chain lengths is formed. The degree of condensation *n* (average number of P atoms per molecule) is used for characterization. Glassy, amorphous products are obtained by fast cooling of the melts of such mixtures. They are commercially important as glassy phosphates or fused phosphates. Products with *n* = 4 (tetraphosphate) to *n* = 25 (Graham's salt) are produced, depending on the application.

Frequently, the incorrect designation “hexametaphosphate” is used for Graham's salt in the literature, especially in the United States. However, it consists essentially of high-molecular, straight-chain polyphosphates with ca. 6 % trimetaphosphate.

Upon tempering, the glassy Graham's salt is converted to crystalline, high-molecular forms (Kurrol's and Maddrell's salt) or to cyclic sodium trimetaphosphate, $Na_3P_3O_9$. Figure 24 shows the various condensed phosphates which can be obtained from NaH_2PO_4 and their conditions of preparation [62].

Figure 24. Condensed phosphates which can be prepared from NaH_2PO_4 by heating and their interrelationship [62]

Fused phosphates (glassy polyphosphates), $Na_{n+2}P_nO_{3n+1}$, are all hygroscopic and water soluble. The pH depends on the degree of condensation. In 1 % aqueous solutions it ranges from 8.4 for hexasodium tetraphosphate to 4.0 for long-chain polyphosphates (*n* > 20).

High-molecular fused phosphates are produced from acidic monophosphates in tank furnaces lined with zircon mineral bricks (Fig. 25). The average chain length depends on temperature, duration of fusion, and water vapor pressure of the melt. The melt temperatures are 600 – 800 °C, depending on the $Na_2O : P_2O_5$ ratio. The outflowing melt is chilled on water-cooled barrels or drums with formation of glassy clods, which are subsequently broken and ground.

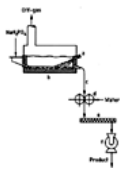

Figure 25. Production of fused phosphate [95]

a) Burner; b) Melting furnace; c) Melt; d) Cooling roller; e) Worm gear; f) Mill

In another process, fused phosphates are obtained directly from elemental phosphorus [123]. Phosphorus and caustic soda are sprayed by air into a spray tower lined with graphite bricks. Molten polyphosphate is formed in the phosphorus flame at 1500 – 1800 °C, accumulates at the tower bottom, and runs onto a cooled drum. About 30 % of the combusted phosphorus is entrained by the off-gas in the form of fused phosphate and small amounts of P_4O_{10}. These are converted to an aqueous solution of sodium phosphate in a downstream wash tower and returned to the combustion tower. The energy required for dehydration of caustic soda and wash solution as well as for production of the polyphosphate melt is supplied entirely by the heat of phosphorus combustion.

A posttreatment of the ground products by addition of Na_2CO_3 or $Na_4P_2O_7$ or by spray coating with alkaline solutions is recommended for decreasing the hygroscopicity [124].

Maddrell's Salt. Crystalline sodium polyphosphate with a chain length of ca. 2000 P units exists in two modifications: the high-temperature triclinic form (*d* 2.669) and the low-temperature form (*d* 2.52). Both are essentially insoluble in water. For application as a polishing compound in fluoridated toothpastes, it is important that formation of water-soluble trimetaphosphate during production is avoided by maintaining a low partial pressure of water vapor. Sodium dihydrogenphosphate is conveyed at a constant rate through a tunnel furnace while the generated water vapor is continuously drawn off [125]. In the Monsanto process, NaH_2PO_4 is compressed into pellets, which are then dehydrated in a rotary kiln at 300 – 460 °C in a hot air stream [126].

Kurrol's Sodium Salt. This crystalline, long-chain sodium polyphosphate is formed from Graham's salt by tempering at 580 – 590 °C: form A, *d* 2.58, monoclinic; form B, *d* 2.56, monoclinic. It has a fibrous structure similar to asbestos and is only sparingly soluble in water. Upon longer storage in water the fibers swell and slowly form a highly viscous solution.

3.2.5. *Sodium Metaphosphates*

Sodium tri- and tetrametaphosphate, $Na_3P_3O_9$ and $Na_4P_4O_{12}$, are the best known examples of cyclic sodium metaphosphates. In addition, metaphosphates with 5, 6, 7, and 8 phosphorus atoms have also been isolated from Graham's salt [127], [128].

Sodium trimetaphosphate can be prepared by cooling of a corresponding polyphosphate melt and subsequent tempering above 350 °C [129] or by condensation of sodium dihydrogen-phosphate with addition of a small amount of ammonium nitrate [130].

Sodium tetrametaphosphate can be prepared by reaction of the corresponding copper salt with Na_2S or by hydrolysis of hexagonal P_4O_{10} with caustic soda [131].

Sodium trimetaphosphate is used as a phosphorylating agent for ascorbic acid [198] to stabilize vitamin C mixtures against thermal decomposition.

3.2.6. *Uses of Condensed Sodium Phosphates*

Water Treatment. Monophosphates form sparingly soluble precipitates with water hardeners and are therefore used as precipitation agents in water softening. However, the water softening effect of condensed phosphates is due to their ion-exchange capacity for cations. The water softening effect and the capability to mask heavy metal ions are utilized in various ways.

For instance, polyphosphates, especially fused phosphates, are used in the *textile industry* for bleaching, dyeing, and textile finishing. In these applications, additional advantages are attained by the other properties of phosphates, e.g., their dispersing capacity [132].

Fused phosphates are also used in *photography* to prevent spots and shadows caused by precipitation of water hardeners and heavy metal ions in the baths.

Fused phosphates are also used in place of monophosphates in the conditioning of *boiler feedwater*. Compared with monophosphates, polyphosphates have the advantage that, after addition to the feedwater, they initially keep the hardeners in solution. The hydrolytic degradation to monophosphates occurs only at elevated temperature in the boiler. Thus, the hardeners are precipitated there and can be discharged as sludge.

Another property of polyphosphates important in practice is their capability, even at very low concentrations, to prevent precipitation of calcium carbonate. This effect is utilized in the phosphate threshold treatment of tap water, e.g., in communal water conduits, in warm water facilities, or in industrial cooling water systems. The required amounts of phosphate, usually 2 – 5 g/m^3 of water, are much lower than the amount necessary for complexation of the hardeners. In this way, the precipitation of calcium carbonate can be completely prevented at moderate temperatures and water hardnesses and normal pH values. At elevated temperatures and higher water hardnesses, crystal growth is hindered to such an extent that no firmly adhering deposits can be formed. It is assumed that growth of the crystal nuclei is stopped or strongly retarded by adsorption of polyphosphate anions [133], [134].

It is important in the threshold process that the phosphate concentration does not exceed a certain limit because otherwise precipitation of sparingly soluble phosphates begins. These can only be redissolved at substantially higher concentration by complexation. Water-soluble, mostly long-chain fused phosphates such as Graham's salt are used in large waterworks. They are added to the water carrying system as a 2 – 5 % solution. Slow-dissolving, glassy calcium sodium polyphosphates are used mostly in smaller works and in private residences. They are charged in form of lumps to a lock installed in the main line or in a bypass, where they dissolve slowly in the passing water [95].

An additional advantage of the phosphate threshold treatment is the anticorrosion effect due to the formation of a firmly adhering, dense, protective layer of iron calcium phosphate. Phosphates also act as corrosion inhibitors for copper, zinc, and aluminum.

Detergents and cleansers (⟶ Cleansing Agents; ⟶ Laundry Detergents) are major applications of polyphosphates. Pentasodium triphosphate, mostly marketed under the name sodium tripolyphosphate (STPP), is by far the most important product. Many industrial cleansers contain 20 – 35 % pentasodium triphosphate as builder. On a much smaller scale, tetrasodium diphosphate is also used for these applications but its significance has declined steadily in the last 10 years. The higher condensed fused phosphates are also used, i.e., in specialty cleansers and personal hygiene preparations.

Triphosphate performs several functions in the laundry process. Primarily, it softens the water by masking alkaline earth metal ions and preventing their deposition on the fabric. In addition, triphosphate prevents precipitation in the washing machine, especially on the heating elements [135]. Furthermore, the pronounced dispersing capacity of triphosphate and its emulsion stabilizing effect are important in the laundry process. Condensed phosphates increase the surface activity of wash active substances and thus their detergency (synergistic effect).

To an increasing extent, $Na_5P_3O_{10}$ is also used in automatic dishwasher detergents, which contain 20 – 50 % pentasodium triphosphate. In this application, the triphosphate is used especially for its water softening and dispersing properties and for its advantageous pH range [136].

Food Industry. The versatile applicability of condensed phosphates in the production and processing of food is essentially due to their ion-exchange and colloid properties.

In the production of processed cheese, condensed phosphates are used as rectifying or melt salts. In the production process, calcium caseinate is converted by sodium polyphosphates to sodium caseinate and is peptized. A stable suspension is formed from this product together with milk fat, finely dispersed during melting, and the substances dissolved in the aqueous phase. Polyphosphates of medium chain length are preferred for this application. An important application of long-chain polyphosphates is their bacteriostatic effect in cheese products. To a small extent, mono- and diphosphates are added as corrective and buffer salts [95].

Polyphosphates are also used in the production of concentrated milk products such as evaporated milk and milk powder, especially to prevent aggregation of the casein (heat coagulation) by deactivation of calcium ions.

Furthermore, the dispersing and calcium-binding properties of polyphosphates are utilized in the preparation of instant puddings, ice cream, skimmed milk products, and instant sauces. In end-products, they act as stabilizers for batters and as gelation regulators.

Polyphosphates are used for prevention of blood coagulation in the preparation of blood plasma from the blood of slaughtered animals in the Fibrisol process.

Polyphosphates effect disaggregation and peptization of muscular albumin. Thus, they act similarly to natural adenosine triphosphate, which is present in living organisms but not in meat. This effect of polyphosphates, especially of diphosphates, is utilized in the preparation of cooked sausage and boiled, cured meats to avoid loss of quality due to a change in the protein structure and the resulting loss of tissue water.

Baking Aids. Disodium dihydrogendiphosphate is widely used as an acid carrier in baking powders, especially in Europe. At moderate temperatures it does not react with $NaHCO_3$ contained in baking powder and thus permits the preparation or storage of large amounts of dough. The acidic diphosphate reacts with sodium hydrogencarbonate only during the baking process with evolution of carbon dioxide and formation of tasteless disodium monophosphate. In the United States, $Na_2H_2P_2O_7$ is used only for fine baked products; otherwise monocalcium phosphate, $Ca(H_2PO_4)_2 \cdot H_2O$, is used predominantly. Because of its high reactivity towards sodium hydrogencarbonate, monocalcium phosphate is often mixed with sodium aluminum sulfate, $NaAl(SO_4)_2$, which reacts only at baking temperature. In addition, it is added to flour as "bread improver" in bread baking [137].

Sodium aluminum phosphates of composition $NaH_{14}Al_3(PO_4)_8 \cdot 4\ H_2O$ and $Na_3H_{15}Al_2(PO_4)_8$ are now used in the United States as acid carriers in baking powders [137]. They are less reactive than monocalcium phosphate and develop their full effect as expanding agents only at baking heat. This is advantageous especially for the durability of prefinished mixtures.

Leather Industry. In the production of leather, polyphosphates act in various ways. During the soaking of hides and skins, they prevent coagulation of blood and reduce the hardness of the water. In the liming step, the properties of a dispersing agent are necessary. The polyanions stabilize the lime slurry in the state of single solid particles, which are able to penetrate quicker. The result is a short and effective liming step. Another important function is the reaction with protein chains in the skin. It is comparable with a mild tanning process, so polyphosphates are used as pretanning agents [95]. These functions are important in the commonly used tannage with chromium salts and in classical vegetable tannage. In the latter process the iron-binding capacity is also essential, because the complexation of iron ions by polyphosphates prevents the formation of iron stains. During the complete beamhouse work, many washing steps are necessary. Addition of polyphosphates to the water prevents the formation of lime stains or lime shadows by complexing Ca ions.

Paint Industry. In the production of paints, especially emulsion paints, the main components water, pigments, fillers, and binder are stirred together to form a paint with good application properties. Manufacturers aim for high solids contents to give good covering power and for good storage stability of the paint. Both goals can be achieved by adding dispersing agents to the water in the first production step. Polyphosphates are well-known dispersing agents. Comparative tests show that Polyphosphates with chain lengths between $n = 6$ and $n = 10$ give the best results. The dispersing effect is due to adsorption of anionic phosphate molecules on the surface of the pigments and fillers, which eliminates the attractive forces between the particles [196]. In current emulsion paint formulations, a combination of polyphosphates and polyacrylates is used, because in addition to optimal dispersing and storage results advantages in scrubbing resistance can be obtained.

Construction Industry. Short-chain polyphosphates, especially the di- and triphosphates, are used as setting retarders for concrete and cement-based mortar. Addition of combinations of tartraric or citric acid with long-chain polyphosphates to gypsum plaster results in optimum workability and plasticity. In cement-based self-leveling floor compounds, condensed phosphates increase the open time and the fluidity of the formulation.

Ceramics Industry. The production of large quantities of ceramic materials (e.g., tiles) requires very homogeneous raw materials. These so-called ceramic slips are formulations of different clays and additives and are prepared in a wet stage by mixing and grinding the components in ball mills. After homogenization the slips must be dried in spray towers to give granular materials. Normal ceramic slips have a solids content of 55 – 60 %. The energy consumption of the spray drier could be lowered by increasing the solids content. By adding small amounts (0.5 – 1.0 %) of polyphosphates to the slips during the ball mill process, solids contents of 70 % can be attained. The polyphosphate anions are adsorbed on the surface of the silicate particles and act as a deflocculant. Calculations show that the decreased energy demand in the spray drier is economically favorable for the overall process [197].

Other Uses. Polyphosphates are also used as dispersing agents and stabilizers in textile dyeing and printing, as well as in the paper industry for the prevention of resin excretions and for improving coatings and fillers [95]. They are also added to drilling muds in oil well drilling because of their emulsion stabilizing effect [95].

3.3. Other Phosphates

A range of inorganic phosphates, including zinc and chromium phosphates, are used as corrosion-inhibiting pigments; these are discussed elsewhere, ⟶ Metals, Surface Treatment. For phosphating of metal surfaces, see ⟶ Metals, Surface Treatment – Phosphating.

3.3.1. *Potassium Phosphates*

Potassium Monophosphates. Potassium forms the following monophosphates:

KH_2PO_4
$KH_2PO_4 \cdot H_3PO_4$
$KH_2PO_4 \cdot 2\ K_2HPO_4 \cdot H_2O$
$KH_2PO_4 \cdot 3\ K_2HPO_4 \cdot 2\ H_2O$
K_2HPO_4
$K_2HPO_4 \cdot 3\ H_2O$
$K_2HPO_4 \cdot 6\ H_2O$
K_3PO_4
$K_3PO_4 \cdot 3\ H_2O$
$K_3PO_4 \cdot 7\ H_2O$
$K_3PO_4 \cdot 9\ H_2O$

Properties. Generally, the potassium salts are more water soluble than the corresponding sodium monophosphates. Monopotassium dihydrogenphosphate, KH_2PO_4, is an exception; it is substantially less soluble than the other potassium monophosphates and monosodium phosphates. Figure 26 shows the solubilities of the potassium monophosphates.

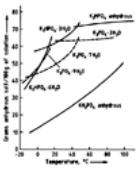

Figure 26. Solubility of potassium monophosphates [70]

The pH of 1 % aqueous solutions are 4.5 for KH_2PO_4, 9.1 for K_2HPO_4, and 11.8 for K_3PO_4. Like the corresponding rubidium and cesium salts, potassium dihydrogenphosphate forms ferroelectric crystals (T_C of KH_2PO_4, –151 °C). The alkali dihydrogenphosphates (K, Rb, Cs, NH_4) exhibit piezoelectric behavior [70].

Production. Potassium monophosphates are usually produced by the same process as the corresponding sodium salts, i.e., from phosphoric acid and potassium hydroxide. For KH_2PO_4, the cheaper potassium chloride is also used instead of KOH [138]. Various methods have been suggested for quantitative removal of the hydrogen chloride formed: use of ion exchangers [139], long-chain tertiary amines [140], or distillation with octane [141]. The economics depend on the possibility of using the hydrogen chloride.

Uses. The potassium phosphates KH_2PO_4 and K_2HPO_4 are used in the food area for mineral enrichment of fruit juices and as stabilizers in milk products. Potassium dihydrogenphosphate is used as a nutrient salt for mushrooms and cormophytes. Crystals of KH_2PO_4 are used as piezoelectric vibrators [142]. Dipotassium phosphate is suitable as a buffer in the range pH 9, e.g., for pH adjustment of water – glycol mixtures in cooling systems of automotive engines in order to avoid corrosion [143]. Mono- and dipotassium phosphate are further processed to potassium polyphosphates.

Tetrapotassium diphosphate [7320-34-5], tetrapotassium pyrophosphate, $K_4P_2O_7$, M_r 330.35, crystallizes in the anhydrous form and the mono- and trihydrates. The anhydrous salt and the monohydrate are strongly hygroscopic. The solubility of $K_4P_2O_7$ in water is 187 g/100 g H_2O at 25 °C, and the pH of a 1 % aqueous solution is 10.3.

Tetrapotassium diphosphate is produced by the same methods as the corresponding sodium salt, i.e., by dehydration of K_2HPO_4 at 350 – 460 °C. Like the other alkali-metal polyphosphates, $K_4P_2O_7$ also has pronounced dispersive properties as well as the capacity to bind multivalent cations and to keep them in solution. Despite the higher price, tetrapotassium diphosphate is used preferentially as builder in liquid detergent formulations containing 20 – 25 % diphosphate and ca. 50 % water in addition to wash active substances. The reasons are the substantially higher water solubility compared with the sodium salts and the hydrolytic stability of the diphosphate anion.

Potassium diphosphate is also used as a stabilizer in foods and for the prevention of tartar formation. Use as a catalyst in the production of Buna is mentioned in some product information brochures.

Pentapotassium triphosphate [13845-36-8], potassium tripolyphosphate, $K_5P_3O_{10}$, M_r 448.43, *mp* 614.5 °C (decomp.), crystallizes in the anhydrous form and as the dihydrate. The anhydrous salt is strongly hygroscopic. Its solubility in water is ca. 200 g/100 g H_2O at 20 °C. The pH of a 1 % aqueous solution is 11.5. Like the corresponding sodium compound, pentapotassium triphosphate is produced from 2 mol of K_2HPO_4 and 1 mol of KH_2PO_4. It is used as a builder in liquid detergents.

Kurrol's salt [29057-10-1], (high-molecular potassium polyphosphate) $(KPO_3)_x$, M_r $(118.09)_x$, where *x* can vary from 400 to 20 000 depending on the conditions of preparation, *d* 2.26, *mp* 813 °C. Although it is a true polyphosphate with chain structure, the incorrect designation potassium metaphosphate is frequently still found in the literature. Kurrol's salt is insoluble in pure water. However, in the presence of other alkali-metal ions, especially Na^+, it dissolves to form a viscous gel.

Kurrol's salt is generally produced by heating KH_2PO_4 at 350 – 500 °C in a rotary kiln. In another process, $(KPO_3)_x$ is produced as a spray dried product from aqueous KH_2PO_4 solution in a hot gas stream at 330 – 390 °C [144]. Due to its good dispersing action, Kurrol's salt is used as an additive in soaps and cleansing agents.

3.3.2. *Ammonium Phosphates*

The most important application of ammonium phosphates is as fertilizers (⟶ Phosphate Fertilizers). Other applications are known, especially as flame retardants.

The following crystalline monophosphates exist in the $NH_3 – P_2O_5 – H_2O$ system [145]:

$NH_4H_2PO_4$	$(NH_4)_2HPO_4$
$NH_4H_2PO_4 \cdot H_3PO_4$	$(NH_4)_3PO_4 \cdot 3\ H_2O$
$NH_4H_2PO_4 \cdot H_3PO_4 \cdot H_2O$	$(NH_4)_3PO_4 \cdot 2\ (NH_4)_2HPO_4$

Only the mono- and diammonium phosphates are commercially important. The compound $(NH_4)_3PO_4$ and the double salt $(NH_4)_3PO_4 \cdot 2\ (NH_4)_2HPO_4$ are unstable at room temperature. They decompose with release of ammonia and formation of $(NH_4)_2HPO_4$.

Ammonium dihydrogenphosphate [7722-76-1], monoammonium phosphate, $NH_4H_2PO_4$, M_r 115.03, *d* 1.80, tetragonal, *mp* 190 °C; solubility in water, 36 g $NH_4H_2PO_4$/100 g H_2O at 20 °C, and 120 g at 80 °C; pH of a 1 % aqueous solution, 4.5. The crystals have piezoelectric properties. In contrast to the tri- and diammonium salts, monoammonium phosphate is very stable (pNH_3, 7 Pa at 125 °C) and melts at 190 °C without decomposition.

Single crystals of monoammonium phosphates are used, e.g., in underwater sounding equipment instead of quartz oscillators. Moreover, monoammonium phosphate has applications in ceramic binders and for the production of special fertilizers (⟶ Phosphate Fertilizers).

For production, see Diammonium Hydrogenphosphate.

Diammonium hydrogenphosphate [7783-28-0], diammonium phosphate, $(NH_4)_2HPO_4$, M_r 132.07, *d* 1.62, monoclinic, *mp* 155 °C (decomp.); solubility in water, 72 g $(NH_4)_2HPO_4$/100 g H_2O at 25 °C; pH of a 1 % aqueous solution, 8.0; pNH_3, 27 Pa at 50 °C and 1.21 kPa at 100 °C.

Ammonium monophosphates are produced by neutralization of phosphoric acid with ammonia. Crude or partially purified wet phosphoric acid is mostly used for fertilizer applications. However, pure thermal phosphoric acid is used preferentially for the production of commercial ammonium phosphates with an assay > 99 %. Commercial ammonium monophosphates are used for flame-retardant textile finishes (not washfast), as flame-retardant impregnating agents for wood, and as components of fire-extinguishing powders. Matches are impregnated with ammonium phosphate solutions to prevent smoldering. The flame-retardant effect of ammonium phosphates is due to their decomposition to phosphoric acid and ammonia. The generated phosphoric acid catalyzes carbonization of combustible material such as cellulose, so that the formation of readily combustible, gases (e.g., CO) is decreased. A film of condensed phosphoric acids or ultraphosphates is formed, which shields the substrate against heat and protects it from access of atmospheric oxygen [146].

Ammonium polyphosphates, $(NH_4)_{n+2}P_nO_{3n+1}$, cannot be produced from ammonium monophosphates by thermal condensation because the ammonia partial pressure is too high at the required temperatures. Short-chain ammonium polyphosphates are produced by neutralization of polyphosphoric acids with ammonia [147]. They are water soluble and used predominantly as components of liquid fertilizers (⟶ Phosphate Fertilizers).

High-molecular ammonium polyphosphate with a minimum of water-soluble fractions is being used to an increasing extent in flame retardants. Five modifications with a chain length of *n* > 50 are described in the literature [148]. Form I or form II products are produced preferentially for commercial applications. The NH_3 partial pressure of form I is 0.053 kPa at 100 °C, 3.19 kPa at 200 °C, and 49.5 kPa at 300 °C [148]. In order to carry out the reaction at low temperatures, urea or similar compounds such as melamine are used as ammonia sources [148], [149].

$$x\,H_3PO_4 + x\,CO(NH_2)_2 \rightarrow (NH_4PO_3)_x + x\,CO_2 + x\,NH_3$$

In another process, phosphorus pentoxide is reacted with diammonium phosphate and ammonia in a mixing and kneading machine [150].

$$x/2\,P_4O_{10} + x(NH_4)_2HPO_4 + x\,NH_3 \rightarrow 3(NH_4PO_3)_x$$

The phosphorus pentoxide/diammonium phosphate/urea combination is also suitable for the production of high-molecular products [151].

$$x/2\,P_4O_{10} + x(NH_4)_2HPO_4 + x/2\,CO(NH_2)_2 + x/2\,H_2O \rightarrow 3(NH_4PO_3)_x + x/2\,CO_2$$

In this reaction, some product properties can already be influenced by the choice of starting materials. An ammonium polyphosphate with a high proportion of modification I and chain lengths of *n* = 40 to *n* = 200 is formed predominantly if urea is used as ammonia source. However, if gaseous ammonia is used, a modification II product with chain lengths of *n* >1000 is obtained. The reasons for these reaction characteristics are related to the reaction rate. If urea is used, then ammonia supply is optimal provided that the components are thoroughly mixed. Accordingly, the reaction proceeds very quickly, and the kinetically favored product is formed. Thermodynamically stable ammonium polyphosphate of modification II is formed only if the reaction proceeds slowly such as in a kneading machine and at temperatures of 240 – 290 °C.

All manufacturing processes have to be carried out in an ammonia atmosphere so that the NH_3 partial pressure exceeds the decomposition pressure of ammonium polyphosphate.

Long-chain ammonium polyphosphates are used as fire-extinguishing agents [152], as flame retardants for paints and varnishes, for flame-retardant finishing of polyurethane foams [153], and in fireproof coatings on steel beams and wooden partitions [154].

3.3.3. *Magnesium Phosphates*

With the exception of magnesium-containing mixed phosphates used in feeds, magnesium phosphates have only minor commercial significance.

Magnesium dihydrogenphosphate [13092-66-5], monomagnesium phosphate, $Mg(H_2PO_4)_2$, M_r 218.28, crystallizes in the anhydrous form and as the di- and tetrahydrate. On dissolution in water, partial hydrolysis occurs with formation of monophosphoric acid and crystalline magnesium hydrogenphosphate, $MgHPO_4 \cdot 3\ H_2O$. Acidic magnesium diphosphate is formed on heating to above 170 °C. It condenses at higher temperatures to give higher-molecular polyphosphates. The hydrated salts are obtained by dissolution of magnesium oxide, magnesium carbonate, or di- or trimagnesium phosphate in phosphoric acid and subsequent crystallization. The anhydrous compound is formed by dehydration at 100 – 170 °C.

Magnesium hydrogenphosphate, dimagnesium phosphate, $MgHPO_4$, M_r 120.29. Only the trihydrate, $MgHPO_4 \cdot 3\ H_2O$, *d* 2.12, orthorombic, *mp* 205 °C (with loss of H_2O), is stable at 25 °C in the $MgO – H_2O – P_2O_5$ system. It is also the usual commercial product. In addition, the anhydrous salt and the mono- and heptahydrate are described in the literature. The trihydrate occurs as the natural mineral newberyite. It dissolves in dilute acids but is practically insoluble in water. It is produced by double decomposition of disodium phosphate with magnesium salts in aqueous solution or by neutralization of solutions of magnesium salts in phosphoric acid with caustic soda.

Magnesium hydrogenphosphate is used for mineral enhancement in foods for babies and athletes as well as in dietetic foods. It is also utilized as a feedstuff additive, in the ceramic industry, and as a stabilizer for $CaHPO_4 \cdot 2\ H_2O$ in toothpastes [155].

Trimagnesium phosphate [7757-87-1], $Mg_3(PO_4)_2$, M_r 262.88, *mp* 1184 °C, crystallizes with 5, 8, and 22 H_2O. The octahydrate, *d* 2.19, monoclinic, occurs as the mineral bobbierite. $Mg_3(PO_4)_2 \cdot 22\ H_2O$, *d* 1.66, also crystallizes in the monoclinic form. The salts are practically insoluble in water but readily soluble in acids. The anhydrous salt is

formed by heating the hydrates at 400 °C. Hydrated trimagnesium phosphate can be prepared from solutions of magnesium salts by addition of trisodium phosphate solution or phosphoric acid and caustic soda. It is used as anticaking agent for silicate-containing cleaning agents [156] and table salt and as a stabilizer for $CaHPO_4 \cdot 2\ H_2O$.

3.3.4. *Calcium Phosphates*

For production and use of calcium phosphates in fertilizers see ⟶ Phosphate Fertilizers. Pure calcium phosphates are used predominantly as feedstuff additives and as polishing compounds in toothpastes.

Calcium dihydrogenphosphate, monocalcium phosphate, $Ca(H_2PO_4)_2 \cdot H_2O$, M_r 252.09, *d* 2.22, triclinic, loses its water of crystallization starting at 109 °C, with formation of anhydrous $Ca(H_2PO_4)_2$ [7758-23-8], *d* 2.55, triclinic.

Monocalcium phosphate dissolves in water with precipitation of calcium hydrogenphosphate and formation of an aqueous phase with increased P_2O_5 content:

$$Ca(H_2PO_4)_2 \cdot H_2O \rightleftharpoons CaHPO_4 + H_3PO_4 + H_2O$$

Therefore, preparation of pure $Ca(H_2PO_4)_2 \cdot H_2O$ [10031-30-8] by crystallization from aqueous solution succeeds only in the presence of excess phosphoric acid [157]. Adhering phosphoric acid gives hygroscopic products which tend to decompose further by absorption of atmospheric moisture. Therefore, commercial products are made with an excess of CaO and contain up to 20 % calcium hydrogenphosphate.

Monocalcium phosphate is produced by neutralization of phosphoric acid with calcium oxide or hydroxide [158]. Only thermal phosphoric acid is used for production of food-grade quality (e.g., for baking aids). Defluorinated wet phosphoric acid is used mostly for feed phosphate. The starting materials are reacted in vessels made of stainless steel by means of high-intensity agitators or in continuously operating mixers. Some of the water is evaporated by the heat of reaction. A pulpy to crumbly product is obtained, which is subsequently dried and ground.

Products with low bulk density are obtained from a slurry of the starting components by spray drying with hot air. Predominantly monohydrate is produced by these processes because the anhydrous salt is less stable towards atmospheric moisture.

A relatively water-stable, less reactive product is obtained by addition of small amounts of various metal salts, e.g., of potassium or aluminum [132].

The anhydrous salt is formed by carrying out the reaction at 140 °C. In a posttreatment step at 200 – 220 °C, the crystals become coated with a glassy, essentially water-stable, protective layer (anhydrous coated calcium phosphate, ACCP).

In the United States, monocalcium phosphate is used mainly as a leavening aid in the production of self-raising flour. Further applications are in mineral enrichment of foods, as a stabilizer for milk products, and as a feedstuff additive in place of calcium hydrogenphosphate.

Calcium hydrogenphosphate [7757-93-9], dicalcium phosphate, $CaHPO_4$, M_r 136.06, crystallizes in the anhydrous form, *d* 2.89, triclinic, and as the dihydrate, *d* 2.32, monoclinic. The dihydrate is slowly converted to the anhydrous salt at 36 °C (incongruent melting point). Substantially higher temperatures are required for production of the anhydrous salt. $CaHPO_4 \cdot 2\ H_2O$ dissolves in water with precipitation of hydroxyapatite, $Ca_5(PO_4)_3OH$, and formation of an aqueous phase with increased P_2O_5 content. However, its solubility is low (220 mg/L CaO and 530 mg/L P_2O_5 at 25 °C).

The dihydrate is produced by neutralization of phosphoric acid with calcium hydroxide, with cooling to avoid formation of anhydrous $CaHPO_4$. Thermal phosphoric acid is used as starting material in the production of $CaHPO_4 \cdot 2\ H_2O$ for use in toothpastes. However, feedstuff phosphates are produced predominantly from defluorinated wet phosphoric acid.

For defluorination, the acid is partially neutralized with calcium hydroxide or carbonate. The fluoride precipitates as calcium hexafluorosilicate together with phosphates of bi- or multivalent metal ions such as Fe and Al. The P_2O_5-containing precipitate is filtered and used for fertilizer production. Na_2SiF_6 or K_2SiF_6 can also be precipitated by addition of sodium or potassium salts. After subsequent addition of calcium hydroxide, the precipitated $CaHPO_4 \cdot 2\ H_2O$ is filtered, dried, and ground.

For a large decrease of the fluoride content, an additional defluorination can be carried out by adding SiO_2 in the form of bentonite [159].

In a more recent process, stoichiometric quantities of $CaCl_2$, NaOH, and H_3PO_4 are added simultaneously to an intensively agitated reactor [160]. In a variation of this process, $CaCl_2$ and NaOH are partially replaced by an aqueous suspension of calcium carbonate [161].

The water-adsorptive capacity of $CaHPO_4 \cdot 2\ H_2O$ also has a function in its application as a polishing compound in toothpastes. Highly adsorptive products can be prepared by adding H_3PO_4 to a suspension of $CaCO_3$ and homogenizing the reaction mixture at the drip spot with a disperser [162].

The anhydrous salt is produced by the same processes as are used for the dihydrate, but precipitation is carried out above 80 °C. Calcium hydrogenphosphate, in the form of its dihydrate and sometimes also as anhydrous salt, is used as a mineral supplement in feedstuff. The dihydrate is used as a component of toothpastes because of its advantageous polishing properties and low abrasiveness [163]. Stabilizers such as $MgCl_2$, $Mg_3(PO_4)_2$, $MgHPO_4 \cdot 2\ H_2O$, or $Na_4P_2O_7$ are added to prevent hardening of toothpastes prepared with $CaHPO_4$ due to dehydration during storage.

Calcium hydrogenphosphate is used in the food industry for mineral enrichment of various products, and in the pharmaceutical industry as a pelletizing aid and thickening agent. Use as fertilizer component for tropical soils is a special application of $CaHPO_4$.

Tricalcium phosphate [7758-87-4], $Ca_3(PO_4)_2$, M_r 310.20, *d* 3.14 (whitlockite, rhombohedral); solubility in water 2.5 mg $Ca_3(PO_4)_2$/100 g H_2O at 25 °C. Two crystalline modifications are known, the high-temperature α- and the low-temperature β-form, transition temperature 1180 °C, *mp* 1730 °C. In production, hydroxyapatite, $Ca_5(PO_4)_3OH$, is formed by complete neutralization of phosphoric acid with slaked lime and subsequent filtration, drying, and grinding. The Ca/P ratio depends on the process conditions. Pure $Ca_3(PO_4)_2$ with a Ca/P ratio of 1.5 can be obtained from hydroxyapatite by calcination above 900 °C. However, commercial tricalcium phosphate consists mostly of hydroxyapatite.

Tricalcium phosphate is added to various substances such as table salt, sugar, and fertilizers as anticaking agent in amounts of 1 – 2 %. It is used in toothpastes containing calcium hydrogenphosphate for adjustment of the polishing strength.

Tricalcium phosphate and calcium halophosphates such as $Ca_5(PO_4)_3(F,Cl)$ are used in the manufacture of luminescent materials.

Dicalcium diphosphate [7790-76-3], calcium pyrophosphate, $Ca_2P_2O_7$, M_r 254.11, *d* 3.09, orthorhombic, *mp* 1353 °C, is practically insoluble in water but soluble in acids. It is produced by thermal dehydration of $CaHPO_4$ or $CaHPO_4 \cdot 2\ H_2O$ at 700 – 900 °C. Dicalcium diphosphate is used as a polishing compound in fluoridated toothpastes because of its low solubility and its inertness towards fluoride ions [164].

Calcium triphosphate [26158-70-3], calcium tripolyphosphate, $Ca_5(P_3O_{10})_2$, M_r 706.47, is practically insoluble in water. It is produced as the octahydrate from calcium chloride solutions by precipitation with pentasodium triphosphate. It is added to hygroscopic salts such as table salt for preservation of the pourability and as a stabilizer for $CaHPO_4 \cdot 2\ H_2O$ [165].

3.3.5. *Boron Phosphate*

Boron phosphate [13308-51-5], BPO_4, M_r 105.79, *d* 2.80, tetragonal, is a white, crystalline solid that sublimes at 1450 °C without decomposition. In addition to the anhydrous product, hydrates with 3, 4, and 6 H_2O have also been reported. The hydrates dissolve readily in water with hydrolysis to give the corresponding acids. The anhydrous salt also hydrolyzes but at a much slower rate. Under ambient conditions, BPO_4 has a cristobalite structure consisting of alternating BO_4 and PO_4 tetrahedra linked via common oxygen atoms to form a three-dimensional network.

Boron phosphate is produced by reaction of boric acid with 85 % phosphoric acid and subsequent dehydration at up to 1000 °C [166]. A product with very large specific surface area is obtained by heating tri-*n*-propyl borate with phosphoric acid at 120 – 130 °C [167]. Addition of aluminum hydroxide in combination with ethylenediamine and ethyl acetate or of the combination aluminum oxide chloride/urea has the same effect [168].

Boron phosphate is used in organic chemistry as a catalyst in various syntheses, especially in hydration, dehydration, isomerization, and amination reactions [169], and in the

ceramic industry as a carrier for boric acid and phosphate.

3.3.6. *Aluminum Phosphates*

The following aluminum monophosphates are known in the Al_2O_3 – P_2O_5 – H_2O system: $Al(H_2PO_4)_3 \cdot n\ H_2O$, $Al_2(HPO_4)_3 \cdot n\ H_2O$, and $AlPO_4 \cdot n\ H_2O$. Acidic salts of composition $Al(H_2PO_4) \cdot (HPO_4) \cdot n\ H_2O$ are also known.

Aluminum dihydrogenphosphate [13530-50-2], monoaluminum phosphate, $Al(H_2PO_4)_3$, M_r 317.96, water soluble, exists in the anhydrous form (hexagonal) and as the trihydrate. On heating to 250 °C, acidic aluminum triphosphate ($AlH_2P_3O_{10}$) is formed, which condenses above 500 °C to give long-chain polyphosphate and cyclic tetrametaphosphate. The trihydrate is obtained by reaction of aluminum hydroxide with phosphoric acid and subsequent crystallization. It is used as binder for highly fireproof materials [170].

Aluminum phosphate [7784-30-7], $AlPO_4$, M_r 121.95, *d* 2.57, *mp* 1850 ± 50 °C, crystallizes in the anhydrous form as hexagonal berlinite. Cubic and orthorhombic modifications are also known. It also crystallizes as a dihydrate (variscite, orthorombic); a monoclinic modification also exists. It is practically insoluble in water and sparingly soluble in acids. It dissolves in alkalies with formation of aluminates. $AlPO_4$ is isotypic with SiO_2 and forms polymorphous modifications with quartz, tridymite, and cristobalite structures.

Aluminum phosphate is produced by addition of alkali-metal phosphates to solutions of aluminum salts or by treating sodium aluminate solution with phosphoric acid.

It is used in the manufacture of molecular sieves and optical glasses, in ceramics, in the enamel industry, and as an anticorrosion powder in coatings.

Aluminum polyphosphate, $[Al(PO_3)_3]_n$, is insoluble in water and acids. Upon heating at 900 – 1000 °C it is converted to a cyclic tetrametaphosphate with *mp* 1230 –1250 °C [171]. Aluminum polyphosphate is produced from monoaluminum phosphate by heating at 600 °C. It is used in the enamel industry in heat-resistant glasses and as component of optical glasses. Aluminum polyphosphates with a molar ratio P_2O_5:Al_2O_3 of ca. 2.5 can be used as hardeners for water glass putties [172].

3.4. Economic Aspects

Tables 8, 9, 10, 11, and 12 list regional and world demand for phosphates in various applications. The following abbreviations are used:

MSP. monosodium phosphate
DSP. disodium phosphate
TSP. trisodium phosphate
TSPP. tetrasodium pyrophosphate
TKPP. tetrapotassium pyrophosphate
STPP. sodium tri(poly)phosphate
SHMP. sodium hexametaphosphate
MCP. monocalcium phosphate
DCP. dicalcium phosphate
DAP. diammonium phosphate
TCP. tricalcium phosphate
SAPP. acidic sodium pyrophosphate
SAlP. sodium aluminum phosphate
IMP. Maddrell's salt (insoluble)
SMFP. sodium monofluorophosphate
APP. ammonium polyphosphate
ZnP. zinc phosphates
AlP. aluminum phosphates

Table 8. Phosphate demand for soaps and detergents (in 10^3 t P_2O_5) *

Region	1988	1990	1992	1994	1996 **
Europe	720	700	690	500	490
America	460	470	450	350	340
Asia	210	230	250	350	380
World	1500	1520	1580	1330	1340

* Main phosphates: STPP (95 %), TSP, TKPP.
** Estimated.

Table 9. Phosphate demand for the food, beverage, and baking industry (in 10^3 t P_2O_5) *

Region	1988	1990	1992	1994	1996 **
Europe	27	28	29	35	40
America	78	80	82	95	96
Asia	6	7	8	10	11
World	120	125	130	145	150

* Main phosphates: STPP, SHMP, TSP, SAPP, SAlP, MCP, DSP (without H_3PO_4).
** Estimated.

Table 10. Phosphate demand for water treatment (in 10^3 t P_2O_5) *

Region	1988	1990	1992	1994	1996 **
Europe	37	38	40	40	42
America	40	41	43	43	46
Asia	11	12	13	25	30
World	90	94	99	115	126

* Main phosphates: Glassy phosphates, SHMP, STPP, TSPP, MSP, DSP.
** Estimated.

Table 11. Phosphate demand for toothpastes (in 10^3 t P_2O_5) *

Region	1994	1996 **

Europe	16	17
America	11	13
Asia	17	18
World	45	49

[*] Main phosphates: DCP, IMP, SMFP.
[**] Estimated.

Table 12. Phosphates for other uses [*] (in 10^3 t P_2O_5) [**]

Region	1994	1996 [***]
Europe	100	100
America	110	112
Asia	42	45
World	255	262

[*] Ceramics, leather, anticaking, setting retarders, flame retardents, paper, anticorrosion pigments, textiles, rubber manufacture.
[**] Main phosphates: Glassy phosphates, STPP, TCP, APP, DAP, ZnP, AlP.
[***] Estimated.

[Top of Page]

4. Toxicology

Thomas Hofmann

Phosphoric Acid. Analogous to phosphorus pentoxide, phosphoric acid does not possess any specific toxic effects. Phosphoric acid irritates the eyes, respiratory tract, and mucous membranes. Aqueous solutions containing phosphoric acid in concentrations > 10% are irritating, solutions with concentrations > 25% are corrosive. After swallowing larger quantities, nausea, vomiting, diarrhea, hematemesis, and hypovolemic shock occur. Ingestion of concentrated solutions causes burns of the mucosa of the mouth, esophagus, and stomach. Additionally, hyperphosphatemia, hypocalcemia and acidosis may be present [173], [174]. Washing of the skin with warm water, irrigation of the eyes with warm water or physiological saline solution is recommended in case of contamination. First aid after ingestion consists of support of respiration and treatment of shock by intravenous replacement of lost fluid volume [174]. The systemic toxicity of phosphoric acid is low. No effect on growth and reproduction was observed in three generations of rats that received 0.75% phosphoric acid in the diet for 90 weeks. No pathological findings attributable to the administration of phosphoric acid were observed in this study. Dental attrition was somewhat higher than in controls [175]. Phosphoric acid was not mutagenic in the Ames test [176]. Phosphoric acid can be used in concentrations of 500 – 1000 mg/L in beverages. No changes which could indicate any detrimental effect on metabolism could be established in voluntary subjects who took 2000 – 4000 mg/kg per day orally for 10 d or 3900 mg/kg per day for 14 d [177]. The TLV-TWA value for phosphoric acid is 1 mg/m^3 [178].

Phosphates and Polyphosphates. A summary of the toxicological properties of phosphates and polyphosphates is given in [179], [180]. There was no evidence for genotoxicity of inorganic phosphates in various test systems. Embryo-fetal toxicity studies in rats, mice, rabbits, and hamsters did not reveal any evidence of teratogenicity. The toxicological profiles phosphates and polyphosphates are largely similar, provided no toxic cations are present. Hyperphosphatemia, secondary hyperparathyreoidism, skeletal changes (deformations, bone loss), and calcification were observed in animal studies after administration of high phosphate dosages. In subchronic and chronic feeding studies with rats, primarily nephrocalcinosis occurred, whereby this species is particularly sensitive to acids that form calcium chelates or complexes. The effective dosages overlap in the different studies; the lowest concentration which caused nephrocalcinosis was 10 g per kg of feed. In view of these results, it was concluded that the oral intake of 70 mg of phosphate (as phosphorus) per kilogram of body weight per day is tolerable in humans (70 mg is the MTDI, maximum tolerable daily intake) [180].

Hypocalcemia, hyperphosphatemia, and dehydration have been reported in studies where sodium phosphate/biphosphate was administered as an enema [181]. Sodium hexametaphosphate, which is used for softening water, can cause serious, or even fatal poisoning when taken orally. The most prominent symptoms are massive shock, irregular pulse, eventually bradycardia, and severe hypocalcemia with tetanic symptoms. Lavaging the stomach with chalk and repeated intravenous injection of calcium gluconate are recommended as therapy [182].

Chlorinated trisodium phosphate, which is used in various cleaning compounds, has a low acute and subchronic toxicity in rats and mice. Toxic effects after chronic administration were observed in rats and were attributable to local irritation/corrosion, which led to termination of the two-year carcinogenicity bioassay. The compound was not carcinogenic after administration for two years to mice. Like sodium hypochlorite, it is a weak base-pair substitution mutagen in bacteria [183].

Sodium monofluorophosphate is used in dentifrices [184] and for treatment of osteoporosis in postmenopausal women [185]. Toxic effects after administration of higher doses (fluorosis) are related to fluoride [186].

[Top of Page]

References

1. *CRC Handbook of Chemistry and Physics*, 62nd ed., 1981/82, p. B 127.
2. *Kirk-Othmer*, 3rd ed., **17**, p. 433.
3. E. H. Brown, C. H. Whitt, *Ind. Eng. Chem.* **44** (1952) 615.
4. S. J. Vollfkovich, A. A. Sokolovski, *Russ. Chem. Rev. (Engl. Transl.)* **43** (1974) no. 33, 224 – 234.
5. Allied Chemical Corp., US 3 723 606, 1970.
6. F. Ritter, *Chem. Ing. Tech.* **22** (1950) 253.
7. A. V. Slack: *Phosphoric Acid*, vol. **1**, "part II", Marcel Dekker, New York 1968, p. 927.
8. H. Pieper, G. Hartlapp, *Fette Seifen Anstrichm.* **59** (1957) 960.
9. Hoechst, US 4 777 028, 1988.
10. M. M. Striplin, *Chem. Eng. Rept.* **2** (1948) 09.
11. Hoechst, US 4 603 039, 1986.
12. H. G. Sheldrick, *Chem. Eng. (N.Y.)* **73** (1966) no. 26, 82.
13. *Phosphorus Potassium* **88** (1977) 30 – 35.
14. W. E. Blumrich, H. J. Koenig, E. W. Schwehr, *Chem. Eng. Prog.* **74** (1978) no. 11, 58.
15. *Chem. Eng. News* **45** (1967) no. 13, 54.
16. B. F. Greek, F. W. Bless, R. S. Sibley, *Ind. Eng. Chem.* **52** (1960) 638.
17. *Phosphorus Potassium* **106** (1980) 41 – 47.
18. L. Rivoirard, M. Lonannic, P. Tiberghien, *Phosphorus Potassium* **147** (1987) 33 – 36.
19. W. Blumrich, *Chem. Anlagen+Verfahren* (1971) no. 11, 59.
20. *Inf. Chim.* **20** (1972) no. 3/4, 19.
21. S. Nakajima, *CEER, Chem. Econ. Eng. Rev.* **10** (1978) no. 4, 41.
22. D. R. Stern, J. D. Ellis, *Chem. Eng. (N.Y.)* **77** (1979) no. 6, 78.
23. G. Kleinmann, *Chem. Eng. Prog.* **74** (1978) no. 11, 37.
24. J. Bidder, J. Hallsworth, *Phosphorus Potassium* **73** (1974) 35.
25. H. von Plessen, G. Schimmel, *Chem. Ing. Tech.* **59** (1978) no. 10, 772 – 778.
26. Hoechst, US 3 764 655, 1973.
27. Chemische Fabrik Budenheim, DE 2 050 008, 1970.
28. Israel Mining Ind., FR 1 396 077, 1964.
29. Hoechst, US 3 953 581, 1976.
30. Albright & Wilson, US 3 914 382, 1975.

31. Rhône-Poulenc, US 3 607 029, 1971.
32. H. Davister, M. Peeterbrock, *Chem. Eng. Prog.* **78** (1982) no. 3, 35 – 39.
33. R. C. Ross, *Eng. Min. J.* **176** (1975) no. 12, 80. Links
34. *Phosphorus Potassium* **99** (1979) 31.
35. F. J. Hurst, W. D. Arnold, A. D. Ryon, *Chem. Eng. (N. Y.)* **84** (1977) no. 1, 56.
36. W. W. Berry, *Chem. Eng. Prog.* **78** (1981) no. 2, 76 – 82.
37. Chemische Fabrik Budenheim, DE-OS 3 327 394, 1983.
38. Current Industrial Reports, *Fertilizer Materials and Related Products*, U.S. Dept. of Commerce, Bureau of the Census.
39. D. Savostianoff, *Inf. Chim.* **314** (1990) 127.
40. *Green Market* 2., 23. March, 7th April, 1987.
41. A. P. Kouloheris, *Int. Symp. Phosphogypsum* 1980, 8 – 35.
42. H. Beijer, W. Mak: "European Practice," in A. V. Slack (ed.): *Phosphoric Acid*, vol. 1, part 2, Marcel Dekker, New York 1968, chap. 6.
43. "Use and Disposal of Wastes from Phosphoric Acid and Titanium Dioxide Production," UN, New York 1988, Economic Commission for Europe, Geneva.
44. Mero, Goldberg, *Chem. Ind. (London)* 9th Oct., 1971.
45. "Cadmium in the European Community," *MARC Rep.* **28** (1982)
46. *Phosphorus Potassium* **155** (1988) 25.
47. *Phosphorus Potassium* **120** (1982) 36.
48. In [43] p. 10 ff.
49. *Phosphorus Potassium* **153** (1988) 25.
50. P. Becker: "Phosphates and Phosphoric Acid," in: Fertilizer Science and Technology Series, 2nd ed., vol. 6, Marcel Dekker, New York 1989, p. 537. *U.S. Environ. Prot. Agency Res. Dev. [Rep.] EPA* 1989, no. 600/9–89/072. R. Szpadt, Z. Augustyn, W. Grysiewicz, *Int. Conf. New Front. Hazard. Waste Manage. 3rd.* 1989, 518 –529. A. Smith, I. Hutchinson, J. Caldwell, *Annu. Madison Conf. Appl. Res. Pract. Munic. Ind. Waste 7th* 1984.
51. A. E. Z. Wissa, N. F. Fuleihan: "Control of Groundwater Contamination from Phosphogypsum Disposal Sites," *Int. Symp. Phosphogypsum* 1980, 482 –539. A. May, J. W. Sweeney: "Assessment of Environmental Impacts Associated with Phosphogypsum in Florida," *Rep. Invest. U.S. Bur. Mines* 1982, no. 8639.
52. *Phosphorus Potassium* **105** (1980) 30.
53. G. Jaekel, *Müll Abfall Abwasser* 1970, no. 15, 2.
54. *Phosphorus Potassium* **87** (1977) 37; **89** (1977) 36; **94** (1978) 24; **96** (1978) 30; **155** (1988) 25.
55. F. Wirsching, *Chem. Unserer Zeit* **19** (1985) 137.
56. *Phosphorus Potassium* **152** (1987) 36.
57. G. Bayer, H.-G. Wiedemann, *Chem. Unserer Zeit* **19** (1985) 33.
58. *Phosphorus Potassium* **152** (1987) 36.
59. VDI-Richtlinien, VDI 3450, May 1989.
60. A. van Bentum, O. Böhme, *Silikattechnik* **29** (1978) 245.
61. B. Keisch, J. W. Kennedy, A. C. Wahl, *J. Am. Chem. Soc.* **80** (1958) 4778 – 4782. Links
62. E. Thilo, *Angew. Chem.* **77** (1965) 1056 – 1066. Links
63. *Mellor's Comprehensive Treatise on Inorganic and Theroretical Chemistry*, vol. **VIII**, suppl. III, Longmann, London 1971, p. 729.
64. R. N. Bell, *Ind. Eng. Chem.* **40** (1948) 1464.
65. W. C. Scott, G. G. Patterson, A. B. Phillips, *Commer. Fert.* **113** (1966) no. 2, 32 ff.
66. W. E. Rushton, *Phosphorus Potassium* 1966 no. 23, 12.
67. TVA, US 2 999 010, 1961 (M. M. Striplin Jr., D. McKnight, E. C. Marks); US 3 015 540, 1962 (M. M. Striplin Jr., D. McKnight, E. C. Marks).
68. Knapsack, DE 1 300 527, 1967 (H. Harnish, F. Krahl, F. Thomas, H. Panter).
69. F. D. Popp, E. E. McEwan, *Chem. Rev.* **58** (1958) 321. Links
70. J. R. Van Wazer: *Phosphorus and Its Compounds*, vol. **1 and 2**, Interscience, New York 1958 and 1961.
71. Neyer Phosphates, EP 24 0635, 1986.
72. R. N. Bell, *Ind. Eng. Chem.* **41** (1949) 2901.
73. Knapsack, DE 2 035 505, 1970.
74. S. K. Metha, J. K. Dhar, *Res. Ind.* **31** (1986) 324.
75. Swift & Comp., US 2 468 448, 1946.
76. S. L. Madorsky, US 2 524 394, 1946. Kelite Prod., US 2 536 456, 1946. Henkel, DE 1 004 148, 1955.
77. D. E. C. Corbridge: *The Structural Chemistry of Phosphorus*, Elsevier, Amsterdam 1974.
78. J. R. van Wazer: *Phosphorous and Its Compounds*, vol. **1**, Wiley Interscience, New York 1958, pp. 813 –814.
79. *Gmelin*, Sodium, Suppl. **21**, 1675 ff.
80. Chem. W. Albert, AU 289 033, 1968.
81. Chem. W. Albert, DE 1 667 413, 1967.
82. Chem. W. Albert, DE-OS 1 792 648, 1968.
83. Albright & Wilson, DE 2 601 085, 1976.
84. Central Glass, JP 80 159 442, 1980.
85. Central Glass, JP 79 150 757, 1979.
86. W. E. White, *Caries Res.* **17** (1983) 2 – 8. Links
87. *Winnacker-Küchler*, 4th ed., **2**, p. 238.
88. E. Thilo, W. Wieker, *Z. Anorg. Allg. Chem.* **291** (1957) 164; Links **306** (1960) 48; **313** (1961) 296.
89. U. Schülke, *Angew. Chem.* **80** (1968) 80. Links
90. J. R. van Wazer: C. F. Callis, *Chem. Rev.* **58** (1958) 1011. Links
91. E. Thilo, *Naturwissenschaften* **46** (1958) 367.
92. H. Schwindewolf, K. F. Bonhoeffer, *Z. Elektrochem.* **57** (1953) 216.
93. K. R. Andress, K. Wüst, *Z. Anorg. Allg. Chem.* **237** (1938) 113. Links
94. H. Rudy, H. Schloesser, R. Watzel, *Angew. Chem.* **53** (1940) 525. Links
95. H. Rudy: *Altes und Neues über kondensierte Phosphate*, J. A. Benckiser GmbH, Ludwigshafen 1960.
96. L. Steinbrecher, J. F. Hatzel, *Inorg. Nucl. Chem. Lett.* **4** (1968) 559.
97. Monsanto, US 2 408 258, 1946; US 2 408 258, 1952.
98. Swann Res. Inc., US 2 021 012, 1952.
99. I. G. Farbenindustrie, DE 745 307, 1942.
100. Zaklady Chemiczne "Alwernia", PL 142 169, 1988.
101. J. A. Benckiser, BE 708 767, 1966.
102. G. W. Morey, *J. Am. Chem. Soc.* **80** (1958) 775. Links
103. E. Thilo, *Adv. Inorg. Chem. Radiochem.* **4** (1962) 1.
104. R. N. Bell, *Inorg. Synth.* **3** (1950) 101.
105. H. D. Nielen, H. Landgräber, *Tenside Deterg.* **14** (1977) 205.
106. *Winnacker-Küchler*, 2nd ed., **1**, p. 394.
107. Stauffer, US 3 210 154, 1962.
108. Monsanto, US 3 233 967, 1962.
109. Hoechst, US 4 534 946, 1985.
110. Knapsack-Griesheim, DE 1 018 394, 1954.
111. FMC, US 4 656 019, 1987.

112. FMC, US 3 661 514, 1970.
113. Blockson Chem.Corp., US 2 419 147, 1942; US 2 419 148, 1944.
114. R. J. Fuchs, *Proc. Int. Congr. Phosphorus Comp. 1977*, p. 201.
115. Chem. Fabr. Budenheim, DE 965 126, 1950.
116. Comp. de Saint-Gobain, US 3 030 180, 1958.
117. Knapsack-Griesheim, DE 1 002 742, 1955.
118. El. Reduction Comp., US 3 356 447, 1963.
119. FMC, US 3 054 656, 1958.
120. Knapsack, FR 1 484 560, 1965.
121. Knapsack, DE 1 244 743, 1964.
122. Monsanto, US 3 046 092, 1960.
123. Knapsack-Griesheim, DE 1 194 383, 1963; DE 1 194 384, 1963.
124. Monsanto, US 2 414 969, 1943.
125. Knapsack, DE 1 667 561, 1967.
126. Monsanto, US 2 356 799, 1943.
127. E. Thilo, U. Schülke, *Angew. Chem.* **75** (1963) 1175. Links
128. S. Ohasi *et al.*, *J. Inorg. Nucl. Chem.* **39** (1977) 1513.
129. Knapsack, DE 1 225 621, 1964.
130. Chem. Werke Albert, US 3 393 974, 1968.
131. R. Bell, L. Audrieth, O. Hill, *Ind. Eng. Chem.* **44** (1952) 570.
132. J. A. Benckiser GmbH: *Calgon in der Textil-Industrie*, Konradin-Verlag, Stuttgart 1957.
133. G. B. Hatch, O. Rice, *Ind. Eng. Chem.* **37** (1945) 710.
134. H. Seifert, *Chem. Ing. Techn.* **27** (1955) 140.
135. K. Merkenich, W. Gohla, *Seifen Öle Fette Wachse* **105** (1979) 39.
136. T. Altenschöpfer, *Fette Seifen Anstrichm.* **74** (1972) 36.
137. A. D. F. Toy: "Phosphorus Chemistry in Everyday Living," *J. Am. Chem. Soc.*, Washington 1976.
138. Lonza-Werke, CH 272 501, 1948.
139. Nissen Chem. Ind., JP-Kokai 40 14 293, 1965.
140. Kaiser Aluminium & Chem. Corp., FR 1 443 004, 1966.
141. M. Israel, R. Blumberg, A. M. Baniel, P. Melzer, IL 21 072, 1966.
142. Brush Dev. Comp., US 2 546 305, 1947.
143. Dow Chem. & Co, US 3 291 741, 1966.
144. L. I. Klochina et al., SU 850 579, 1981.
145. B. Wendrow, K. A. Kobe, *Chem. Rev.* **54** (1954) 891. Links
146. R. W. Little: *Flameproofing Textile Fabrics*, Reinhold, New York 1957.
147. J. G. Getsinger, M. R. Siegel, H. C. Mann, J. Agric, *Food Chem.* **10** (1962) no. 4, 341.
148. C. Y. Shen, N. E. Stahlheber, D. R. Dyroff, *J. Am. Chem. Soc.* **91** (1969) 62. Links
149. Chem. Werke Albert, US 3 419 349, 1965; DE-AS 1 216 856, 1963.
150. Hoechst, DE 2 330 174, 1973.
151. Knapsack, CH 502 264, 1969.
152. Hoechst, DE-AS 1 771 540, 1968.
153. Monsanto, US 3 423 343, 1965.
154. H. L. Vandersall, *J. Fire Flammability* **2** (1971) 97.
155. Monsanto, US 3 012 852, 1956.
156. Blockson Chem. Corp., US 2 468 448, 1946.
157. *Gmelin*, 8 ed., **28-B 3**, 1126.
158. H. W. Easerwood: *Phosphoric Acid*, Phosphates and Phosphatic Fertilizers, Reinhold, New York 1952, p. 443.
159. A. V. Slack: *Phosphoric Acid. Fertilizer Science & Technology Series*, vol. 1, Marcel Dekker, New York 1968.
160. Knapsack, DE-AS 1 216 264, 1963; DE 1 567 609, 1963.
161. Knapsack, DE 1 567 616, 1964.
162. Hoechst, DE 3 705 845, 1988.
163. Stauffer, US 3 169 096, 1962.
164. Procter & Gamble, US 3 112 247, 1962.
165. Victor Chemical Works, US 2 852 341, 1954.
166. H. L. Goltz, J. B. Moffat, *J. Catal.* **22** (1971) 85.
167. J. B. Moffat, J. F. Neelemann, *J. Catal.* **31** (1973) 274. Links
168. "BOR" Ind. Enterprises, Primorsk, SU 1 375 562, 1988.
169. *Gmelin*, 8 ed., **19**, 99.
170. I. L. Rashkovan, L. N. Kuzminskaya, V. A. Kopeikin, *Izv. Akad. Nauk SSSR Neorg. Mater.* **2** (1966) no. 3, 541.
171. F. D' Yvoire, *C. R. Hebd. Seances Acad. Sci. 251* (1960) 2182.
172. Hoechst, DE 1 252 835, 1964.
173. R. von Burg, *Toxicol. Appl. Pharmacol.* **12** (1992) 301 – 303
174. W. N. Harrison, S. M. Bradberry, J. A. Vale: Phosphoric Acid, UK PID Monograph, National Poisons Information Service (Birmingham Centre), West Midland Poisons Unit, City Hospital NHS Trust, Dudley Road, Birmingham, United Kingdom 1998.
175. World Health Organisation, *Tech. Rep. Ser.* **683** (1982) 25.
176. F. Y. Al-Ani, S. K. Al-Lami, *Mutat. Res.* **206** (1988) 467–470. Links
177. F. Lauerson, *Z. Lebensm. Unters. Forsch.* **96** (1953) 418 – 440. Links
178. American Conference of Governmental Industrial Hygienists: TLVs and BEIs. Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices, Cincinatti, Ohio 2005.
179. JECFA (Joint FAO/WHO Expert Committee on Food Additives): Toxicological evaluations of certain food additives. Twenty-sixth report of the joint FAO/WHO Expert Committee on Food Additives. WHO Food Additives Series 17 (1982).
180. EFSA (European Food Safety Authority): Opinion of the Scientific Panel on Dietetic Products, Nutrition and Allergies on a request from the Commission related to the Tolerable Upper Intake Level of Phosphorus. *The EFSA Journal* **233** (2005) 1–19.
181. R. F. Davis, J. M. Eichner, W. A. Bleyer, G. Okamoto, *J. Pediatr.* **90** (1977) no. 3, 484 – 485. Links
182. S. Moeschlin: *Klinik und Therapie der Vergiftungen*, Thieme Verlag, Stuttgart 1986.
183. National Toxicology Program: NTP Technical Report Series No. 294 (1986).
184. G. L. Vogel, Y. Mao, L. C. Chow, H. M. Proskin, *Caries Res.* **34** (2000) 404–411. Links
185. J. Y. Reginster, L. Meurmans, B. Zegels, L. C. Rovati, H. W. Minne, G. Giacovelli, A. N. Taquet, I. Setnikar, J. Collette, *Ann. Intern. Med.* **129** (1998) 1–8. Links
186. World Health Organisation, *Environmental Health Criteria* **227** (2002).
187. J. K. Lim *et al.*, *Caries Res.* **12** (1978) no. 3, 177 – 179. Links
188. Benckiser-Knapsack, DE 2 914 664, 1979.
189. T. Staffel, G. Brix : "Untersuchung des Abbindeverhaltens von Zementmischungen", *Con. Chem. J.* **3** (1996) no. 3, 93 – 95.
190. J. B. Gill, S. A. Riaz, *J. Chem. Soc. A* **1969**, 183 – 187.
191. M. Watanabe, M. Matsura, T. Yamada, *Bull. Chem. Soc. Jpn.* **54** (1981) 738 – 741. Links

192. G. Kura, *Polyhedron* **10** (1968) 738 – 741.
193. J. R. Van Wazer: *Phosphorus and Its Compounds*, vol. **1** Interscience, New York 1958, p. 640. T. Glonek, J. R. Costello, T. C. Myers, J. R. Van Wazer, *J. Phys. Chem.* **79** (1975) 1214 – 1218.
194. N. Yoza *et al.*, *J. Chromatogr.* **186** (1980) 471 – 480. Y. Baba *et al.*, *J. Chromatogr.* **350** (1985) 119 – 125. Links E. Vaeth *et al.*, *Fresenius' Z. Anal. Chem.* **329** (1987) 584 – 589.
195. T. Klein, unpublished results.
196. T. Staffel, D. Sondermann, G. Brix, "Einsatz von kondensierten Phosphaten als Dispergiermittel", *Farbe + Lack* **7** (1994) 523 – 527.
197. A. Nebelung, Sprechsaal 7/85.
198. M.-L. Liao, P. Seib : "A Stable Form of Vitamin C", *J. Agric. Food Chem.* **38** (1990) 355 – 366.

[Top of Page]

Related Articles
Superphosphates ⟶ Phosphate Fertilizers; Zinc and Chromium Phosphates ⟶ Pigments, Inorganic

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Phosphorus

HERBERT DISKOWSKI, Hoechst Aktiengesellschaft, Werk Knapsack, Hürth, Federal Republic of Germany (Chaps. 1 – 8)

THOMAS HOFMANN, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 9)

1. Introduction 1
2. Properties 3
2.1. Physical Properties 3
2.2. Chemical Properties 4
3. Occurrence of Phosphates in Nature . 4
4. Production of White Phosphorus . . . 6
4.1. Raw Materials 7
4.2. The Phosphorus Furnace 10
4.3. Purification of Crude Phosphorus . . 14
4.4. Treatment of Phosphorus-Containing Wastewater 14
4.5. Slag 15
4.6. Ferrophosphorus 15
4.7. Yields 16
5. Production of Red Phosphorus 17
6. Quality Specifications, Storage, and Transportation 18
6.1. White Phosphorus 18
6.2. Red Phosphorus 19
7. Uses 20
8. Economic Aspects 20
9. Toxicology 21
10. References 22

1. Introduction

Phosphorus, P [*7723-14-0*], atomic number 15, A_r 30.98, is a member of group 15 of the Periodic Table. In its compounds, phosphorus exhibits oxidation numbers between −3 and +5.

Phosphorus occurs in three main allotropic forms: white (yellow when slightly impure) [*7723-14-0*], red, and black. Other forms of solid phosphorus are derived from these three main allotropes. Only white and red phosphorus are of industrial importance.

The only stable isotope of phosphorus is ^{31}P. Of the seven known unstable isotopes of phosphorus, only ^{32}P and ^{33}P, with half-lifes of 12.26 d and 25.3 d, respectively, are suitable for isotopic labelling [1].

History. White phosphorus was first prepared by HENNIG BRAND in 1667 by evaporating urine and strongly heating the residue in the absence of air. Soon afterwards, phosphorus received its name (light bearer). In 1715, GMELIN published a method of preparation based on the reaction of the residue from the evaporation of urine with charcoal and sand. The sodium metaphosphate formed ($NaPO_3$) reacted with the carbon to form white phosphorus.

In the 1800s, PELLETIER, further developed a process invented by SCHEELE for the production of white phosphorus for matches. Bone ash was first reacted with sulfuric acid to produce $Ca(H_2PO_4)_2$ [*7758-23-8*], which was then reacted at high temperatures with wood charcoal to form white phosphorus, $Ca_3(PO_4)_2$ [*7758-87-4*], and CO. Only two-thirds of the phosphorus in the $Ca(H_2PO_4)_2$ was recovered as white phosphorus.

In the mid-1800s, Albright & Wilson started the production of white phosphorus in England. Syrupy phosphoric acid was reacted with wood charcoal in clay retorts, and the phosphorus was distilled off. The development of the electric furnace at the end of the century enabled a method of manufacture originally invented by WO in 1829 to be developed into an industrial process. Tricalcium phosphate, coke, and silica were reacted to produce white phosphorus, calcium metasilicate [*10101-39-0*], and carbon monoxide:

$$2\,Ca_3(PO_4)_2 + 6\,SiO_2 + 10\,C \longrightarrow P_4 + 6\,CaSiO_3 + 10\,CO$$

This process, including the use of an electric furnace, was patented by the Englishman READ-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_505

MAN. A single-phase furnace of 80 kW was constructed in England in 1893, and soon phosphorus was being produced in tonnage quantities in electric furnaces in the United States, France, and Germany.

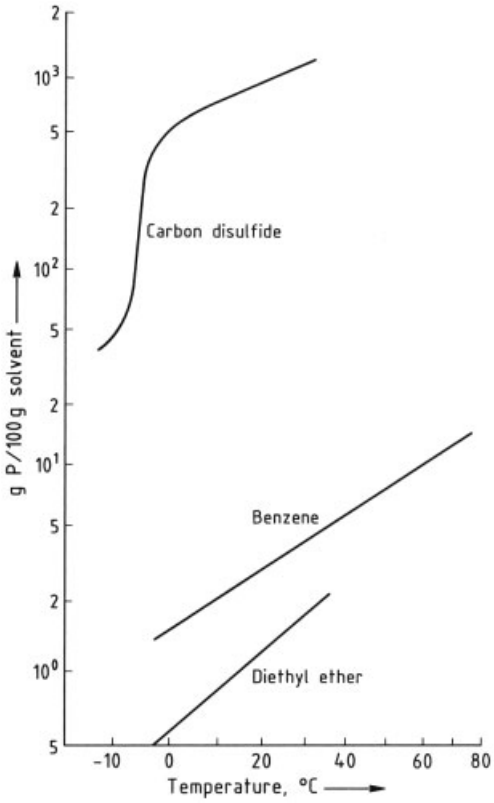

Figure 1. Solubility of white phosphorus

Although white phosphorus was prohibited in matches in the early 1900s because of its toxicity, consumption continued to increase (see Chap. 7).

In 1925 the first three-phase electric furnace (3 MW) was built in Bitterfeld, Germany. In 1927, four 10 MW furnaces were built in Piesteritz, and similar large furnaces were built in the United States in the 1930s.

After World War II, phosphorus production worldwide grew very rapidly. Many phosphorus furnaces with power consumptions of up to 70 MW per furnace were built in the United States, Canada, the Netherlands, the Federal Republic of Germany, the former Soviet Union, and China.

2. Properties

2.1. Physical Properties

Of the three main allotropic forms, only black phosphorus is thermodynamically stable at room temperature; white and red phosphorus are metastable under normal conditions.

White phosphorus is the most reactive of the solid modifications. It is formed on condensation of phosphorus vapor under water as the α-or high-temperature form (cubic) – a transparent, waxy mass with a density of 1.82 g/cm^3. Below −76.9 °C, phosphorus changes into the hexagonal β-form (ϱ 1.88 g/cm^3). Both these forms and the liquid and gaseous phases consist of tetrahedral P_4 molecules.

White phosphorus, *mp* 44.1 °C, *bp* 280.5 °C, is a nonconductor. It is soluble in many liquids, especially in phosphorus trichloride, phosphorus tribromide, and carbon disulfide, and also in organic solvents such as benzene and diethyl ether (see Fig. 1). Further physical data and constants for white phosphorus are given in [2].

Red Phosphorus. The general term red phosphorus includes numerous crystalline or amorphous forms with varying depths of color. The densities are in the range 2.0 – 2.4 g/cm^3 and the melting points between 585 and 610 °C. All the forms of red phosphorus appear to be made up of P_4 molecules polymerized in varying degrees. Above 200 °C, yellow phosphorus changes into red phophorus, the transformation being catalyzed, among others, by water vapor, iodine, and sulfur. Ultraviolet light also accelerates the process. Unlike white phosphorus, red phosphorus has low solubility, is not spontaneously flammable, and is nontoxic.

Commerical red phosphorus is X-ray amorphous and has a density of ca. 2.16 g/cm^3.

Violet or Hittorf's phosphorus is the best defined form of red phosphorus. It was prepared by HITTORF in 1865 by heating phosphorus in contact with molten lead, and by SCHENCK in 1905 from a phosphorus–bismuth melt. The crystal structure of violet phosphorus is rhombohedral or triclinic; the density is 2.34 g/cm^3.

Light red phosphorus, described by SCHENCK, is not a definite modification, but a very finely divided form of red phosphorus. A 10 % solution of white phosphorus in phosphorus tribromide is boiled for 10 h, yielding a product which still contains some solvent. It can be obtained in a pure state by reduction of phosphorus tribromide with mercury. Schenck's phosphorus is appreciably more reactive than the product obtained by the transformation of white phosphorus; for example, it reacts vigorously with alkali , forming phosphine, hydrogen, and hypophosphite.

Other Phosphorus Modifications. The relationships between the known phosphorus modifications are given in [3].

Black phosphorus exists in three crystalline and one amorphous form. All these are highly polymerized, insoluble, and practically nonflammable. BRIDGMAN prepared orthorhombic black phosphorus (ϱ 2.69 g/cm^3) by heating white phosphorus to 200 °C at 12 kbar [4]. Further increase of the pressure leads to reversible formation of the even denser rhombohedral and cubic forms [5]. Orthorhombic black phosphorus can also be prepared at slightly elevated pressures if white phosphorus is heated to 380 °C with mercury and crystal nucleii of black phosphorus [6]. The orthorhombic form can be prepared from a solution of phosphorus in bismuth [7].

Amorphous black phosphorus (ϱ 2.25 g/cm^3) has a partly crystalline structure, but becomes crystalline at higher pressure and temperature.

Brown phosphorus [8] is formed by the condensation of phosphorus vapor containing P_2 molecules (> 1000 °C) onto a surface cooled by liquid nitrogen. Above 0 °C, brown phosphorus changes into a mixture of the white and red forms.

Vitreous, gray phosphorus can be obtained as an intermediate product in the preparation of orthorhombic black phosphorus. It is a dark gray, brittle substance that exhibits conchoidal fracture. It can be ignited by impact and is a nonconductor.

Liquid and Gaseous Phosphorus. The liquid phase consists of P_4 molecules, and is the same irrespective of how it was formed, e.g., by melting of some form of solid phosphorus or by condensation of phosphorus vapor. The density of liquid phosphorus at 50 °C is 1.74 g/cm^3, and its boiling point is 280.5 °C. The vapor pressure curve of liquid phosphorus is shown in Figure 2. In the region of the broken line (between 350 and 510 °C), liquid phosphorus does not exist due to its transformation into red phosphorus. Liquid phosphorus can be supercooled to below − 71.3 °C. Its critical point (triple point) is at 695 °C and 83.3 bar.

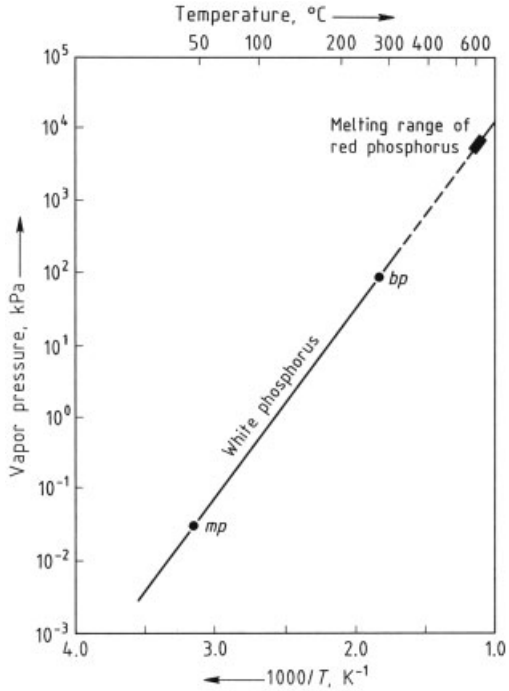

Figure 2. Vapor pressure curve of phosphorus

Above 800 °C, the P_4 molecules in the gas phase begin to dissociate, there being equal numbers of P_4 and P_2 molecules at ca. 1200 °C. The extent of dissociation into atomic phosphorus in the vapor is so low that at 1500 °C the concentration of P atoms is < 0.1 %.

2.2. Chemical Properties

White phosphorus burns to form P_2O_5 or the lower oxides such as P_2O_3 and PO_2, depending on the amount of oxygen present. Finely divided white phosphorus is extremely reactive and ignites spontaneously at room temperature in air. In bulk form it burns at ca. 60 °C.

The glow of phosphorus in air can be observed in the dark. The chemical energy of the slow oxidation is converted not only into heat, but also partly into light (chemiluminescence). The Mitscherlich reaction for the determination of small amounts of white phosporus is based on this light emission.

White phosphorus does not react with water under normal conditions, but phosphorus vapor is oxidized by water vapor to form hydrogen and phosphorus pentoxide.

White phosphorus acts as a strong reducing agent towards concentrated oxidizing acids. Sulfuric acid is reduced to sulfur dioxide, and nitric acid to oxides of nitrogen.

The compounds formed by the reaction of phosphorus with halogens or sulfur are described in [9] (see also → Phosphorus Compounds, Inorganic). Carbon monoxide does not react with phosphorus or its oxides, but carbon dioxide has an oxidizing action above 650 °C.

With alkali , white phosphorus undergoes disproportionation to form phosphine and hypophosphite. Phosphorus combines with many metals, e.g., aluminum, zinc, cadmium, iron, and copper, to form well-defined phosphides (→ Phosphorus Compounds, Inorganic). Aqueous solutions of easily reducible metal salts react with white phosphorus on gentle heating, and the metals are precipitated either as such (e.g., gold, silver, lead) or as the phosphides (e.g. Cu_3P_2).

Red phosphorus ignites in air at ca. 300 °C. At room temperature, friction or impact are sufficient for ignition.

3. Occurrence of Phosphates in Nature

Phosphorus does not occur naturally in the free state. It is almost always found in its highest oxidation state (+5) in the form of orthophosphates. These are very widespread: in the earth, in rocks, on the ocean floor, and in the bones of animals. Of the many phosphate minerals, virtually the only one used in industry as a source of phosphorus is apatite. The deposits of aluminum phosphates in Africa and Brazil are of minor industrial importance. Guano is nearly all used in fertilizers.

The following apatites represent the most important raw materials for thermal phosphorus manufacture (in an electric furnace) and for wet dissolution (phosphoric acid production by reaction of phosphate rock with sulfuric acid; → Phosphoric Acid and Phosphates): fluorapatite $Ca_5(PO_4)_3F$ [*1306-05-4*], hydroxyapatite $Ca_5(PO_4)_3OH$ [*1306-06-5*], and carbonatoapatite $Ca_{10}(PO_4)_6(CO_3) \cdot H_2O$ [*12415-02-0*]. They are of either magmatic or sedimentary origin.

The *sedimentary apatites* represent by far the largest quantity of phosphate rock both as regards size of the reserves and amount extracted. Because they mainly originate from animal remains from earlier geological epochs, they also contain sand, silicates, carbonates, and even organic compounds.

In the United States, sedimentary phosphate deposits are spread across the entire country. The phosphate rock is extracted in opencast mines, clay and sand are removed as sludge, and flotation then gives products which are sorted and stored according to their percentage of tricalcium orthophosphate (% BPL: bone phosphate of lime, or % TPL: triphosphate of lime), which is calculated from the P_2O_5 content. The usual industrial grades of phosphate rock in the United States have BPL values of 62 – 75 %.

The phosphate rock is sometimes calcined to decompose the organic compounds, and this to some extent also removes combined CO_2 from the carbonatoapatite and calcite. Calcination also renders the phosphate rock less reactive towards mineral acids.

Morocco is the world's largest exporter of phosphate rock. The sedimentary phosphates in the three deposits Khouribga, Youssuffia, and Bu-Craa (extraction by subsurface mining) vary in the amount of metallic impurities and chloride. The BPL content of the phosphate rock after treatment is 68 – 78 %.

In addition to these two most important producer nations, there are other countries with similar types of deposit, including several Mediterranean countries, the former Soviet Union and China (these two countries being nonexporters), other African countries, and Brazil. However, with the exception of the former Soviet Union and China, these are not of great economic significance.

Magmatic apatites are extracted in only two countries: in the former Soviet Union as Kola phosphate and in South Africa as Phalaborwa

phosphate. Both these ores must be separated from other minerals by flotation. The BPL value for Phalaborwa phosphate is 80 – 88 %, and, for Kola phosphate, 86 %. These phosphates do not contain organic components.

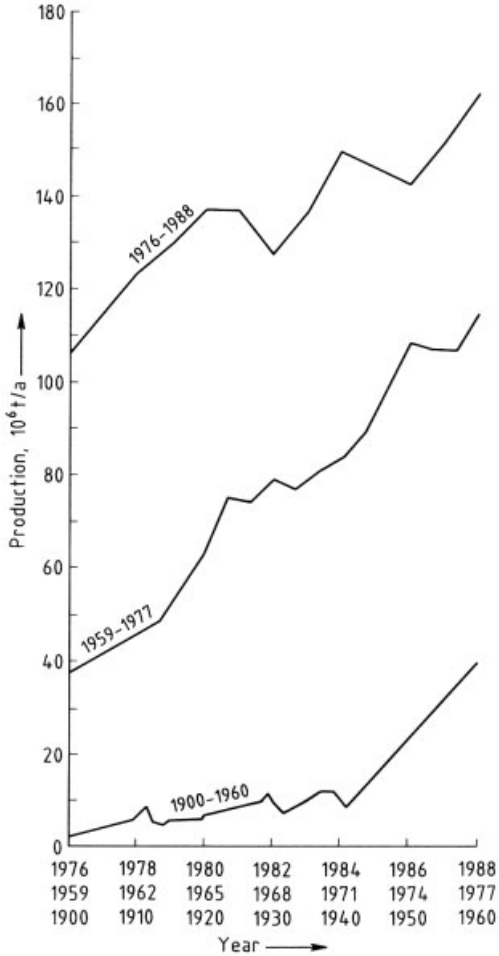

Figure 3. Phosphate output 1900 – 1988 (10^6 t/a)

Table 1 lists typical analytical figures for the important phosphate rocks. All these can in principle be used in electrothermal phosphorus production. The P_2O_5 content should not be too low, as a high content of mineral impurities leads to increased electric power consumption. Metallic impurities can have a detrimental effect on the reaction process. Small amounts of such impurities (e.g., zinc) present in the phosphate rock can build up during the recycling of dusts, etc.

For the wet dissolution process, magmatic phosphate rocks are preferred because of the absence of organic impurities. Depending on the dissolution process, Kola phosphate and Phalaborwa phosphate may differ with respect to ease of filtration of the gypsum.

Economic Aspects. The annual output of phosphate rock increased slowly but steadily from 1900 until the 1940s, and then rapidly from ca. 10×10^6 t at that time to 40×10^6 t in 1960. From 1960 to 1977 the annual output increased to 116×10^6 t. From 1977 to 1988 there was again a rapid increase to ca. 165×10^6 t (see Fig. 3). It is estimated that the natural reserves of phosphate exceed the annual consumption by a factor of 1000.

The United States is still the world's largest producer of phosphate rock, but its share of total world production has decreased. In 1959 it was 44 %, in 1977 ca. 41 %, and in 1988 ca. 28 %. The share held by the former Soviet Union has continually increased: in 1959 it was ca. 16 %, in 1977 ca. 21 %, and in 1988 ca. 24 %. Morocco maintains a constant share: in 1977 it was ca. 15 %, and the same in 1988. China has a 9 % share, and so has become one of the important producing nations. Twelve other nations produced over 10^6 t phosphate rock each in 1988 (see Table 2).

The amount of phosphate rock used in electrothermal production of elemental phosphorus in 1988 has decreased. In 1959, ca. 11 % was used, in 1977 ca. 6 %, and in 1988 ca. 4 %. The reason is the increasing use for fertilizers. In 1988, ca 95 % of the phosphate rock produced was reacted with mineral acids, and of this ca. 95 % was used in fertilizers. The difference of ca. 5 % was used to produce orthophosphoric acid, which was purified and then used in applications where phosphoric acid manufactured by combustion of elemental phosphorus was previously used.

4. Production of White Phosphorus

Large-scale production of white phosphorus uses exclusively the Wöhler process, in which apatites (nearly always fluorapatite) are reacted with coke and silica (flint pebbles or gravel) in low shaft electric furnaces at 1400 – 1500 °C. The reaction proceeds according to the equation

Table 1. Composition of phosphate rocks from various sources (wt % based on dry material)

Type of phosphate rock	P_2O_5	CaO	F	SiO_2	Al_2O_3	Fe_2O_3	CO_2	SO_3
Kola concentrate (USSR)	39.1	51.5	3.4	2.0	1.2	0.7	0.0	0.0
Phalaborwa (South Africa)	38.4	52.5	2.3	1.9	0.2	0.4	1.3	0.1
Florida phosphate (USA)	31.6	47.7	3.9	9.0	1.0	1.6	3.7	1.4
Idaho phosphate (USA)	31.2	45.7	3.6	8.5	0.6	0.5	3.1	1.2
Khouribga phosphate (Morocco)	34.8	52.5	4.2	0.9	0.5	0.1	4.1	1.5
Togo phosphate (Togo)	29.7	40.3	3.9	10.2	4.4	5.3	1.6	0.1

$$4\,Ca_5(PO_4)_3F+18\,SiO_2+30\,C \longrightarrow 3\,P_4+30\,CO+18\,CaSiO_3+2\,CaF_2$$

Experimental production of phosphorus in blast furnaces was carried out in the United States before World War II. Phosphate–carbon briquettes were reacted with silica and coke in the presence of preheated air such that the carbon was oxidized only as far as CO. The phosphorus formed was either condensed or immediately oxidized to P_4O_{10} [11].

Table 2. Phosphate rock production in 1988 [10]

Country [a]	Production, 10^3 t	Average P_2O_5 content, wt %
America	**52 117.4**	
United States	46 596.1	30.8
Brazil	4 671.6	35.6
Africa	**42 113.1**	
Morocco	24 783.3	31.8
Algeria	1 332.1	30.3
Tunisia	6 103.3	29.7
Egypt	1 145.8	25.6
Togo	3 464.0	36.3
Senegal	2 295.6	36.2
Republic of South Africa	2 850.2	37.9
Europe	**39 530.0**	
Former Soviet Union	38 820.0	30.9
Asia	**29 650.1**	
China	15 000.0	27.0
Iraq	1 273.0	30.0
Israel	3 479.0	31.4
Jordan	5 666.2	33.0
Syria	2 342.0	30.5
Oceania	**1 546.4**	
Nauru	1 540.4	38.5

[a] Countries producing $> 10^6$ t.

The simultaneous production of phosphorus and calcium carbide was also investigated. The reaction is

$$2\,Ca_3(PO_4)_2+28\,C \longrightarrow P_4+16\,CO+6\,CaC_2$$

However, this process did not become industrially established [12].

Reaction Mechanism. sPhosphate and quartzite form a melt in the furnace at the reaction temperature of 1400 – 1500 °C, and the carbon reacts with the ionic phosphate–silicate–fluoride melt according to

$$2\,PO_4^{3-}+Si_6O_{15}^{6-}+5\,C \longrightarrow P_2+5\,CO+2\,Si_3O_9^{6-}$$

The phosphate ion is reduced to gaseous P_2, transferring its charge to the polymeric silicate ions, whose degree of polymerization is thereby reduced.

It is not certain whether this is a solid–liquid reaction [13], or whether CO acts as a reducing agent and the CO_2 formed reacts with carbon to form more CO. It is known that the solid components apatite, silica, and carbon react at 1100 °C to form phosphorus [14]:

$$Ca_3(PO_4)_2+5\,CO \longrightarrow 3\,CaO+5\,CO_2+P_2$$

$$5\,CO_2+5\,C \longrightarrow 10\,CO$$

4.1. Raw Materials

Trouble-free operation of a phosphorus furnace requires that the three raw materials–sintered phosphate, coke, and silica–should have the same grain size distribution, so that the solid burden in the furnace has virtually constant gas permeability throughout its volume. The reaction gases, phosphorus and CO, formed in the lower part of the furnace, should give up as much of their heat as possible to the burden as they pass through the burden bed (see Section 4.2), and this movement of gases should be accompanied by as few pressure fluctuations as possible. A

high dust content in the burden leads to blockages in the burden bed and consequent pressure surges and temperature peaks.

Any significant differences in the grain sizes of the three raw materials lead to segregation in the furnace and hence variability of the specific electrical resistance of the burden. This has unwanted effects on the electrode heights in the furnace.

It is preferable that the raw materials used should have an unvarying chemical composition over a long period of time. This does not usually present any problems as far as the coke and silica are concerned. The variations in the P_2O_5, CaO, and SiO_2 contents of the phosphate sinter should be minimized.

Phosphate Sinters. As a rule, the phosphate rock is received from the mine in finely ground form, and then converted into lump form (sintered phosphate) by the phosphorus producer. Phosphate rock in broken pieces (produced by sieving) from Florida is now practically unobtainable in Europe. Lump phosphates from North Africa have a relatively high CO_2 content, which leads to unwanted bursting of the lumps in the furnace with consequent dust formation.

There are several processes for the manufacture of lump phosphate from finely ground phosphate rock. The earthy phosphates from Idaho can be fairly easily compressed into briquettes, and then heated to the sintering point in a Dwight–Lloyd bed (FMC). This causes the organic components of the phosphate to be burnt off, producing a sinter with good resistance to pressure and abrasion.

In the Monsanto process, the Idaho phosphate is ground and fed into a heated rotary furnace with 10 % dust from sintered phosphate. Pellets (nodules) are formed at the entrance to the furnace. These become sintered when the organic substances are burnt.

A process applicable to almost all types of phosphate rock is pelletization with a disk granulator followed by roasting on a moving grate (Hoechst, Allis Chalmers). Finely ground phosphate rock, to which a slurry for pelletizing has been added, is agglomerated into pellets of 1 – 2 cm diameter in a disk granulator. The slurry includes various amounts of, e.g., clay, dust from electrostatic precipitators used in phosphorus production, sodium silicate solution, sulfite liquor, and alkali-metal phosphates, depending on the type of phosphate rock used.

Roasting of the green pellets produced is carried out with Lepol roasters, Dwight – Lloyd bands, shaft furnaces, or rotary furnaces. Figure 4 shows a Lepol roaster [15]. The moist phosphate pellets pass via the feeder (a) into the drying zone (b) where they are dried by the exhaust gases from the combustion zone (c) and by the hot air from the cooling zone (e). The heat for roasting the dried pellets is provided by combustion of phosphorus furnace gas (85 – 90 % CO) in the burners (f), in which other fuels such as brown coal dust can also be used if there is insufficient gas. After the roasted pellets have passed through a dwell zone (d), they pass to the cooling zone (e) where they are cooled by drawing cold air through the bed of pellets. The process makes efficient use of the heat of combustion of the fuel gas.

The gases from the sintering process are removed by the extraction fan (j) via a spinningflow wet scrubber (i) or other gas washing equipment. The water used is slightly alkaline, containing $Ca(OH)_2$, and so absorbs HF and dust.

These gas cleaning methods, when used in new installations [16], reduce the impurities in the waste gases to levels not exceeding the following (mg/m^3):

Dust	50
Phosphoric acid	20
Fluorine compounds	5
Chlorine compounds	10
Sulfur dioxide	100
Phosphine	1

The filtration equipment that dedusts the air extracted from conveying equipment and mills gives dust levels in the exhaust air not exceeding 50 mg/m^3 in modern plants [16].

No wastewater is produced in this process, as the water used in the spinning-flow wet scrubber goes into the mix for granulation. The fines from the washer, together with the precipitated CaF_2, are recycled to the sintering process.

Coke. In general, dried metallurgical coke is used. The maximum moisture content should be 1 %, and the ash content as low as possible. A good quality coke has the composition: 90 % fixed carbon, 8.5 % ash, 1 % H_2O, and 0.5 %

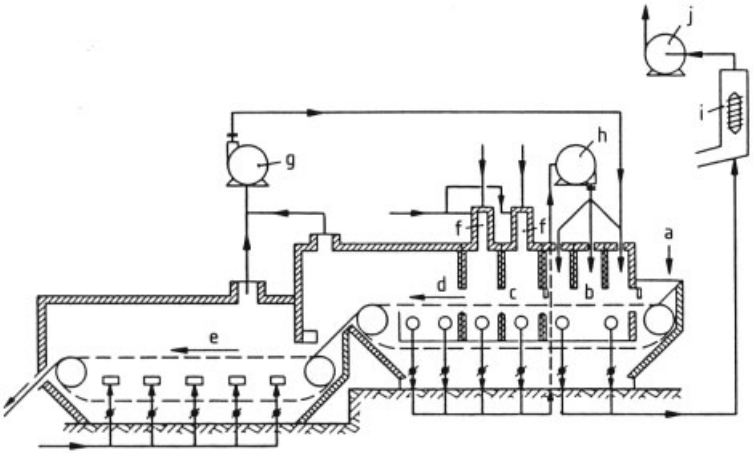

Figure 4. Sintering plant with Lepol roasters
a) Raw materials feed; b) Drying zone; c) Combustion zone; d) Dwell zone; e) Cooling zone; f) Burners; g) Hot gas blower; h) Combustion gas blower; i) Spinning-flow wet scrubber; j) Extraction fan

other volatiles. It is possible to replace a large proportion of this coke by cheaper petroleum coke. In this case, a higher proportion of organic impurities in the phosphorus must be accepted.

Silica. Deposits of silica as pebbles, gravel, or coarse sand containing > 97 % SiO_2 are fairly common, but not every phosphorus installation has such a deposit in the vicinity. Either higher costs of transporting the silica would have to be incurred, or the disadvantages of lower quality silica (high impurity levels, especially Fe_2O_3 content) would have to be accepted.

Composition of the Burden. The C/P_2O_5 ratio in the burden determines the extent of "complete reduction" of the phosphate, and the SiO_2/CaO ratio determines the flow properties of the slag. The carbon is consumed not only by the reduction of the pentavalent phosphorus in the phosphate, but also in reducing impurities such as oxides of iron. Therefore, it is not possible to calculate the coke requirement solely from the P_2O_5 content of the sintered phosphate. The P_2O_5 content of the slag must also be taken into account.

In phosphorus plants using phosphates (also coke and silica) from several suppliers, the effect of the composition of these phosphates on the production cost of the phosphorus is investigated before purchase. It is advantageous to use a computer to perform rapid calculations of quantities of raw materials required, furnace throughput, and amounts of byproducts, basing this on thermal data for the furnace and raw materials, and chemical compositions of the raw materials. For example, using

8.00 t phosphate pellets (31 % P_2O_5)
2.80 t silica (97 % SiO_2)
1.25 t coke (90 % fixed carbon)
0.05 t electrode material (uncalcined)

produces

1.00 t phosphorus
7.70 t slag
0.15 t ferrophosphorus (22 % P)
0.10 t dust from electrostatic precipitators
2500 m^3 off-gas (85 % CO)

4.2. The Phosphorus Furnace

The plant for production of white phosphorus (Fig. 5) consists of three parts:

A) the furnace with its electrodes, the raw material feeding equipment, the slag and ferrophosphorus tapping points, and the two vents for the furnace gas;
B) the electrostatic precipitators for removal of dust; and
C) the condensation equipment with the receivers for the phosphorus produced.

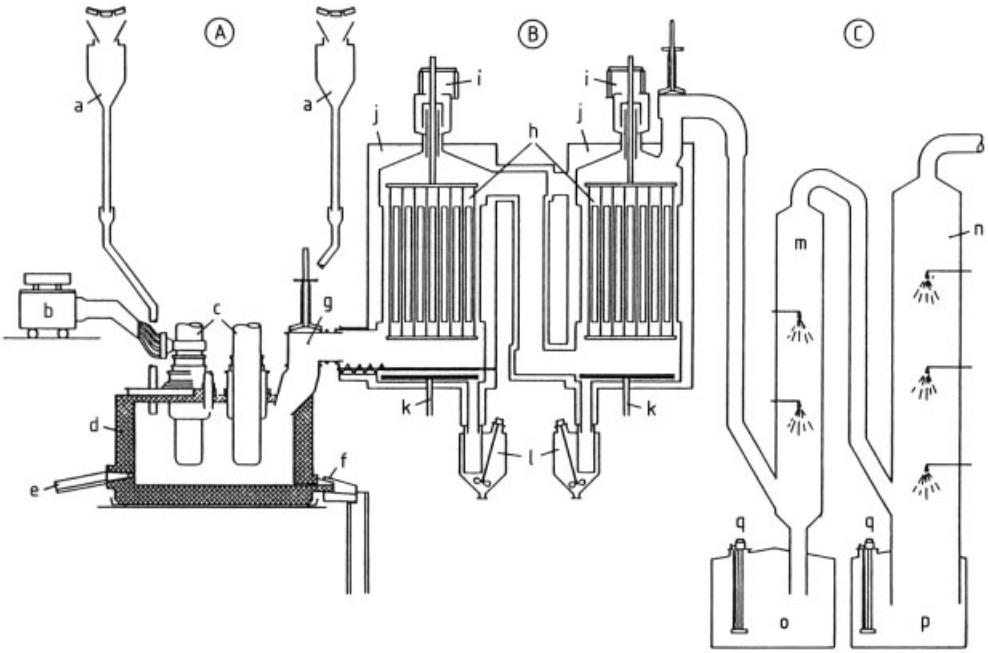

Figure 5. White phosphorus production plant
A) Furnace installation
a) Raw material feed from bunkers via chutes; b) Transformers (3); c) Electrodes (3); d) Furnace casing; e) Slag tapping point; f) Ferrophosphorus tapping point; g) Furnace gas exit
B) Gas purification
h) Electrostatic precipitators with sparking and precipitating electrodes; i) Insulated, gas-tight electrical connections; j) Heating jacket, k) Dust collection equipment; l) Containers for dust – water mixture
C) Condensation plant
m) Hot condensation; n) Cold condensation; o) Liquid phosphorus collection zone; p) Solid phosphorus collection zone; q) Submerged pumps

Small phosphorus furnaces have simpler equipment, i.e., dust settling chambers instead of electrostatic precipitators, and simplified condensation equipment without the high- and lowtemperature condensers shown in Figure 5. The liquid byproducts (slag and ferrophosphorus) are tapped off at different heights above the furnace floor.

The dust is removed from the furnace gas, which contains ca. 400 g/m^3 phosphorus and ca. 40 g/m^3 dust, by electrostatic precipitators. The temperature in this equipment must exceed the boiling point of phosphorus (280 °C), or the precipitators must be continuously heated. The furnace-gas temperature is usually higher in larger furnaces. This temperature should not exceed 550 °C, as the efficiency of electrostatic dedusting decreases at higher temperatures. The next stage in the process is the condensation of the phosphorus from the furnace gas. After the dust and phosphorus have been removed from the furnace gas, it can be combusted to provide heat for calcining the pellets and for the production of condensed phosphates.

The phosphorus furnace (Fig. 6) is a symmetrical alternating current furnace of the type used for the production of calcium carbide [17].

The furnace casing is made of sheet steel, the lower part lined with hard blocks of strongly calcined carbon, and the upper part with firebrick. The floor and lower part of the furnace are water-cooled. The furnace is circular, or shaped like an equilateral triangle with rounded corners. The three electrodes are situated at the corners of a regular triangle. The furnace top is made of corrosion-resistant steel, protected from the

heat of the furnace by a system of water-cooled coils and a covering made of clay and refractory cement. Furnace tops constructed of steel-reinforced refractory concrete without cooling coils have also proved to be satisfactory.

The furnace is equipped with Söderberg electrodes and has a flat furnace top which ensures that the submerged holder through which the current flows to the electrodes [18] and the electrodes themselves need not be of too great a length. The material of the Söderberg electrodes (20 – 25 % tar or pitch, 30 – 35 % coke dust, and 40 – 50 % granular calcined anthracite, softening point 110 °C) is heated by the electrical current and the furnace heat and is baked solid in the region of the contact plates. The electrode must become baked solid over its entire cross section when it is fed downwards to compensate for its consumption in the furnace (a few centimeters per hour). If the electrodes are not completely baked, there is a risk of breakage, especially with long electrodes. New electrode casings are welded on and electrode paste is added while the furnace is operating.

The electrode cross section depends on the electrode current, which for large phosphorus furnaces is 50 – 70 kA. For a current density of 4 A/cm^2, the corresponding electrode diameter is 1.30 – 1.50 m.

To avoid electrical losses in the secondary circuit, the transformers are situated close to the furnace. When single phase transformers are used, each one supplies two electrodes. The three electrodes of the furnace can be operated with a star or a delta connection, the latter being preferred. The secondary voltage is 200 – 650 V. The electrode current is kept fairly constant by automatically raising and lowering the electrodes; i.e., if the current increases, the electrodes are raised, increasing the electrical resistance between the electrodes and the furnace floor and hence reducing the current, and vice versa.

The furnace current is taken directly from the high voltage circuit [19]. Modern furnace transformers are provided with stepwise controllers which can increase the secondary voltage from a lowest value of ca. 130 V in 20 V steps up to a maximum of ca. 650 V.

The required quantities of sintered phosphate, coke, and silica are measured out by automatic weighing equipment, mixed by conveyors, and delivered into bunkers (Fig. 5), from which the burden slides down chutes into the furnace.

There are four feed chutes surrounding each electrode. The middle three combine to form a wider pipe (see Fig. 6). Since the furnace is operated under slight overpressure, nitrogen is continuously fed into the charging bunker, so that no carbon monoxide can reach the charging platform.

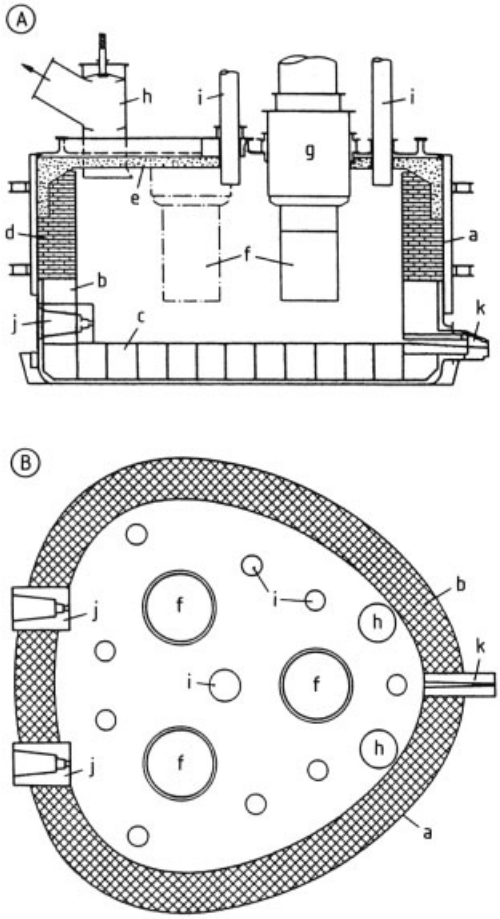

Figure 6. Phosphorus furnace
A) Side view; B) Plan view
a) Furnace casing; b) Carbon wall lining; c) Carbon floor lining; d) Fireclay brick wall; e) Furnace top; f) Electrodes; g) Submerged holder; h) Gas exit; i) Feed chutes; j) Slag tapping point; k) Ferrophosphorus tapping point

The lifetime of a phosphorus furnace is mainly determined by the durability of the carbon lining. Detailed monitoring of the state of the lining does not necessitate shutting down the furnace and completely emptying it. It is

achieved by incorporating radioactive sources at points where erosion is known to occur, a technique already used in steelworks [20]. Two or three ^{60}Co sources ($t_{1/2}$ 5.3 years) are used. These are inserted at various depths in the carbon bricks [21]. The intensities of the sources are chosen such that the radiation intensity outside the furnace decreases by one-third or one-half when the innermost source disappears. In this way, the thickness of the carbon lining can be continuously monitored. Furnaces with water-cooled floors have a longer working life than those without [22].

American Phosphorus Furnaces. American furnaces usually have a domed top, originally lined with refractory bricks, but usually made of cast concrete. Such furnaces can only be operated with prebaked electrodes. The original straight-line arrangement of electrodes gives rise to live and dead phases with three-phase electric current, which is a major disadvantage compared with furnaces with a symmetrical (i.e., triangular) electrode arrangement. The Tennessee Valley Authority (TVA) therefore also built furnaces with carbon or graphite electrodes in a triangular arrangement.

The baked carbon or graphite electrodes are bolted together with carbon nipples. However, the electrode is weakened at the junction point, and the furnace must be shut down to fit an extension piece by this method.

Whereas European furnaces usually have a carbon lining consisting of large blocks that must be fitted to within a millimeter, American furnaces floors are lined with carbon bricks. This type of lining is easier to install but less durable. It is very important to have a layer of rammed material between the carbon lining and the steel furnace casing.

In 1950 the TVA constructed a slowly rotating phosphorus furnace, which, however, did not become established [23].

After World War II, domed phosphorus furnaces with prebaked electrodes were built with power consumptions up to 70 MW.

Energy Consumption. Phosphorus manufacture in an electric furnace is energy intensive. The electrical current consumption for a 50 MW furnace breaks down as follows:

Energy consumption by the chemical reactions	ca. 45 %
Heat content of slag, ferrophosphorus, and furnace gases	ca. 40 %
Electrical losses, radiation losses, and cooling losses (cooling water)	ca. 15 %

Larger furnaces have smaller radiation losses. The composition of the phosphate rock and the purity of the other raw materials determine the amounts of slag and ferrophosphorus formed; these materials account for a considerable proportion of the power consumption. The specific energy consumption is minimized if the P_2O_5 content of the phosphate rock is high and the molar ratio of CaO to P_2O_5 is minimized.

Furnaces with a power consumption of 50 MW that use calcined phosphate rock have a specific power consumption of 12.5 – 13.5 kWh/kg phosphorus (excluding the phosphorus in the ferrophosphorus). Lump phosphate rock also contains calcium carbonate, which requires additional energy for its decomposition. Also, the CaO content must be controlled by the addition of silica, which increases the quantity of slag.

Gas Purification. The large-scale production of white phosphorus was first made possible by the use of electrostatic precipitators [24] to remove the dust from the furnace gases. After leaving the furnace, these gases first pass into an electrostatic gas purification system (usually two chambers in series) which removes 98 – 99 % of the dust particles. The gas velocity in the precipitators is ca. 0.5 – 0.8 m/s, and the dwell time in the electric field is ca. 10 s per precipitator. The vessels are heated to prevent condensation of phosphorus.

The separated dust settles on the floor of the electrostatic filter and is removed by a raking arm. It then falls into a slurrying vessel through tubes that dip into the slurry. The dust is slurried with water, and the slurry is used as a binder for the agglomeration of the phosphate rock in the sintering plant.

The composition of the dust (Table 3) shows that it is primarily formed in side reactions in the furnace.

Table 3. Examples of the compositions of dusts from electrostatic precipitators

Component	Dust composition, wt %		
	Hoechst	TVA	FMC
P_2O_5	16.9	28.5	38.4
CaO	12.2	10.7	16.3
F	3.6	5.9	1.1
SiO_2	36.6	18.1	19.6
Al_2O_3	2.4	3.4	1.2
Fe_2O_3	1.4	1.5	0.9
Na_2O	4.5	3.3	10.0
K_2O	11.3	16.2	10.0
C_{fix}	0.8	4.6	1.6

A small number of plants still operate without electrostatic precipitators, and the phosphorus is isolated by means of filter presses from the mixture of phosphorus and mud produced by the condensation process.

Condensation. The virtually dust-free furnace gases enter the condensation system at ca. 400 °C. There are hot and cold condensation stages. The hot stage operates above the melting point of phosphorus, and the cold stage below it. In both cases, the gas is washed with water, causing the phosphorus to condense from the gas stream.

The hot condensation equipment consists of one washing tower, or two in parallel, into which water at 50 – 60 °C is sprayed. The water collects with the condensed phosphorus in the lower part of the tower below the washing zone. The phosphorus is periodically pumped or siphoned off into heated storage vessels where it awaits further treatment while covered with a water layer. Part of the circulating water is continuously bled off and replaced, so as to remove impurities which would affect the quality of the product. These impurities consist mainly of fluorine compounds derived from SiF_4. This first forms hexafluorosilicic acid, which then hydrolyzes in the condensation equipment and stabilizes the mud formed by the dust and phosphorus. This interferes with the subsequent purification of the phosphorus (see Section 4.3).

The wash tower of the cold condensation stage uses water at 10 – 25 °C; the phosphorus condenses as a solid. The quantity of water required depends on the vapor pressure of the phosphorus in the hot condensation stage. The efficiency of recovery increases as the temperature of the water decreases (Fig. 7). The solid phosphorus that collects in the lower part of the tower is liquefied periodically by a heating system and pumped into storage tanks. The condensation water can be cooled by vacuum evaporation using steam jets [25].

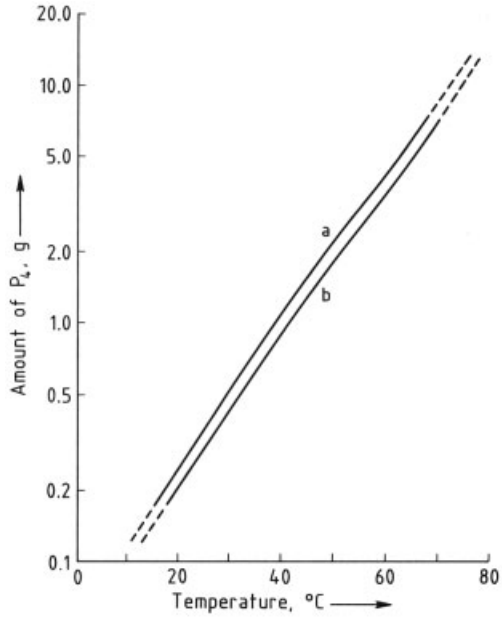

Figure 7. Phosphorus content of furnace gas as a function of temperature
a) Amount of P_4 per cubic meter (STP) of furnace gas; b) Amount of P_4 per kilogram furnace gas

By efficient operation of the condensation equipment, the loss of phosphorus in the furnace gas can be less than 0.1 % of the total product. About 90 % of the furnace gas is used as a fuel in the sintering and phosphate salt plants, and 10 % is flared off. The dry furnace gas contains 85 – 90 % CO, 3 – 8 % N_2, 4 – 6 % H_2, 1 % CH_4, and 1 % CO_2. The calorific value is ca. 11.3 kJ/m^3.

Leaks in the equipment can allow CO to escape, endangering operating personnel. In the area of the furnace and up to the start of the hot condensation process, the danger is slight because phosphorus is also released, which ignites and causes combustion to form CO_2. The P_2O_5 that is also formed can lead to increased corrosion. Between the hot condensation process and the flaring off, and also at the open parts of the feed bunker, escapes of CO can occur. There are numerous sampling points where the gas is sampled continuously and automatically analyzed for CO. If the MAK value is exceeded,

a visible alarm is activated on the spot, and an audible one in the control room.

4.3. Purification of Crude Phosphorus

The phosphorus produced in the hot condensation equipment contains mud as an impurity. When the phosphorus is stored in liquid form, an intermediate layer appears between the phosphorus and the water layer. This contains up to 80 % elemental phosphorus, the remainder being water, and is stabilized by colloidal silicic acid.

All phosphorus plants face the problem of disposing of the phosphorus-containing mud. The practice of sending it to ponds where it oxidizes slowly is now obsolete. A certain amount can be used in the production of thermal phosphoric acid if the mud is homogeneously stirred into the phosphorus [26]. However, the quality of the phosphoric acid is reduced by this procedure, and the process is therefore seldom used.

The mud impurities can be removed from the crude phosphorus by centrifugation. A water-free mud with a high phosphorus content is obtained, which is recycled to the furnace (FMC).

Another proven method is to filter off all the phosphorus-containing product in a filter press. A wet residue with up to 30 % P is produced. This can be treated in three ways:

1) The residue is steam distilled in lined pressure vessels at a steam pressure of up to 10 bar [27]. The phosphorus is condensed as a liquid in tubular coolers, and the phosphorus-free residue is used in the phosphate pelletizing process along with the dust from the electrostatic precipitators.
2) The residue is fed into the furnace exit, whereby the hot gas vaporizes the water and phosphorus, and the solids are added to the dust from the electrostatic precipitators [28].
3) The mud is fed into the furnace. In this case, very strict safety precautions must be taken, and the rate of addition must be automatically controlled because pressure surges may occur if too much water enters the furnace.

In the United States, phosphorus distillation is also carried out in gas-heated rotary furnaces made of corrosion-resistant steel. Distillation processes that operate in the absence of water vapor have not proved successful owing to the formation of red phosphorus, which remains in the residue.

Phosphorus produced using Söderberg electrodes can contain up to 0.3 % organic impurities. This amount can be reduced to < 0.01 % by using adsorption media [29]. The residues so produced are treated in the way described above.

4.4. Treatment of Phosphorus-Containing Wastewater

The wastewater from the hot and cold condensation processes, the water from the storage tanks, and the water used for pressure transfer of the phosphorus must be treated to remove most of their phosphorus content before they are conveyed to the central wastewater treatment installation. Since wastewater always carries with it phosphorus mud, the purification problems resemble those which arise with crude phosphorus.

Wastewater is generated at a rate of 5 – 7 m^3 per tonne of phosphorus produced, and contains up to 1 g/L phosphorus, usually as a stable emulsion. Purification can be carried out by two methods:

1) The wastewater is passed through a settling tank to remove phosphorus, and another to remove coarse mud particles. The sediments are fed to the phosphorus filtration process. The water from the second settling tank is adjusted almost to neutrality, and a flocculating agent is added. All the solid materials are then separated out in a Dorr thickener. These solids are used in the pelletization of phosphate rock, or treated in the same way as the phosphorus mud. The clarified water from the Dorr thickener can be used for various purposes instead of fresh water (see below). Amounts surplus to requirements can be treated with chlorine solution to oxidize the dissolved phosphorus (3.6 mg/L) before discharge.
2) Another effective technique is filtration of the wastewater, a process similar to the filtration of crude phosphorus. Substances that stabilize phosphorus mud are kept at a low level in the filter. The phosphorus and water pass through the filter and separate quantitatively in a settling tank. The clarified

water contains essentially dissolved phosphorus only. It can be used without further treatment in place of fresh water: (1) as granulation water for slag granulation (making up for evaporation losses); (2) as the protective layer for storage of phosphorus; (3) to produce a slurry from the dust from the electrostatic precipitators; (4) as rinse water in the filtration of phosphorus; and (5) for cleaning storage vessels and tank cars. Water that is surplus to these requirements can be passed directly to a central wastewater treatment installation. If the waste-water is sent directly to the outfall, the dissolved phosphorus (3.6 mg/L) is most simply oxidized by chlorine bleaching. The residue is treated in the same way as that from phosphorus filtration.

4.5. Slag

For each tonne of phosphorus, ca. 8 t slag is also produced. The smooth tapping off of this slag at ca. 1350 °C determines to a considerable extent the production rate of a furnace. The furnace described above has two slag tapping holes to enable the slag to be tapped "symmetrically". The holes are changed over every few hours, and must therefore be wear-resistant. In the furnace described, they are water-cooled metal devices built into the carbon blocks of the furnace lining, and can be closed with a water-cooled metal plug.

As the slag leaves the furnace, it is granulated by quenching in a water jet, forming vitreous pseudo-wollastonite, $CaSiO_3$, and cuspidine, $Ca_4Si_2O_7F_2$. This granulated silicate slag is a valuable byproduct. It is used as a construction material for dumping grounds, and a small amount is milled for use as a foundry chemical.

If there is an outlet for the slag in lump form for the construction of embankments or roads, it is poured while molten into beds where it cools slowly.

Conversion of the slag into lightweight building blocks is not feasible because of the slight radioactivity of the uranium originating from the phosphate rock.

Various fume emissions are associated with slag tapping and granulation. Fluorine compounds, phosphoric acid, chlorine compounds, and water vapor are emitted. The use of wet scrubbing in modern plants enables emissions to be kept below the following levels [16]:

Phosphoric acid (as P_2O_5)	20 mg/m^3
Fluorine compounds (as F^-)	5 mg/m^3
Chlorine compounds (as Cl^-)	10 mg/m^3

The yield of phosphorus is critically dependent on the P_2O_5 content of the slag. Therefore, not only must the analysis of the raw material be monitored, but continual analysis of the slag produced is also essential. With good furnace operation, the P_2O_5 content of the slag can be held to 1.0 %. The slag modulus (weight ratio of SiO_2 to CaO) is 0.8 – 0.9. Examples of slag compositions are given in Table 4.

The heat of the slag accounts for ca. 30 % of the electrical energy consumed by the furnace. Attempts to recover this energy have so far been unsuccessful.

4.6. Ferrophosphorus

As well as phosphorus, the furnace produces ferrophosphorus, which collects under the slag. The ferrophosphorus is tapped off, usually once per day, through a hole in one of the carbon blocks, which is opened with an oxygen lance and closed again with a clay plug. The ferrophosphorus, together with some slag, flows into a crucible, from which the lighter slag overflows into a bed of sand. When the tapping off is completed the ferrophosphorus is poured from the crucible into a bed of sand, and, after cooling, is broken into pieces [30].

Every time the ferrophosphorus is tapped off, its radioactivity is measured. If a ^{60}Co source becomes dislodged as the furnace lining wears away, the ^{60}Co is transferred quantitatively into the ferrophosphorus. The radiation supervisor decides in each case whether the ferrophosphorus should be sold, placed in short term storage, or stored indefinitely.

Fume extraction operates in the areas surrounding the furnace tapping and sand beds. The waste gases are purified by wet methods. For new installations, the following emission levels are not exceeded [16]:

Dust	50 mg/m^3
Phosphoric acid (as P_2O_5)	20 mg/m^3

Table 4. Composition of furnace slags (wt %)

Process	P_2O_5	CaO	F	SiO_2	Al_2O_3	Fe_2O_3	Na_2O	K_2O
Hoechst	1.0	48.2	2.8	43.3	2.4	0.2	0.5	0.2
TVA	1.0	50.0	2.6	39.5	6.2	0.3	0.6	0.7
FMC	1.0	48.2	2.6	38.3	5.9	0.1	0.0	0.0

Depending on the method of furnace operation, the ferrophosphorus contains between 15 and 28 % P, corresponding approximately to the formula Fe_2P. The more extensively the phosphate rock is reduced, the more silica is converted into silicon, which reduces the phosphorus content of the ferrophosphorus. The combined content of P and Si is 25 – 30 %. Ferrophosphorus with a low silicon content (< 3 %) has a fairly good market potential, and is used in the manufacture of phosphorus-containing alloys. The grades with a low phosphorus content are not in great demand, but can be used in smelting low-phosphorus iron ores or to increase the P_2O_5 content of basic Thomas slag.

Ferrophosporus is stored in the open, and supplied either loose or in drums. The other metals present in the ferrophosphorus in addition to the Fe, P, and Si depend on the composition of the phosphate rock. Examples of compositions of ferrophosphorus are given in Table 5.

Valuable metals such as vanadium, can be recovered economically from the ferrophosphorus if they are present in unusually high concentrations. Vanadium is recovered together with chromium by blowing with oxygen. Both metals accumulate in the slag.

Older processes for treating ferrophosporus, which are however no longer of importance, are described in [31].

4.7. Yields

Modern pelletizing and sintering installations operate with a 99 % yield based on the phosphate rock used. Only dust losses and losses during transport and handling occur, the former being very small with modern dust filters. The recycling of the dust from the electrostatic precipitators can involve slight yield losses if it contains elemental phosphorus which burns to form P_2O_5. This is not completely absorbed by the water in the scrubbers attached to the sintering furnace.

If ZnP becomes concentrated in the dust from the electrostatic precipitators (due to a high Zn content in the phosphate rock), to such an extent that it cannot be recycled, it must be used in other ways; e.g., it is oxidized and then used in fertilizers. This also represents a loss of yield.

Furnace operation involves higher losses than sinter production. Good reduction of the P_2O_5 in the sinter gives a slag containing on average 1.3 % P_2O_5, corresponding to a loss of yield of ca. 5 %. The ferrophosphorus byproduct represents an average loss of yield of ca. 3 %.

When the excess furnace gas is flared off, phosphorus is burnt as well as CO, the amount depending on the temperature of the cold condensation water. As only ca. 10 % of the total furnace gas is flared off, the associated yield loss only amounts to a maximum of 0.01 %. Most of the gas (ca. 90 %) is used in the phosphate sintering plant or for the production of condensed phosphates. Here, the phosphorus in the gas is burnt and used again.

Modern equipment for dewatering and treating phosphorus mud produces phosphorus mainly in elemental form. In earlier processes, a large proportion of the recovered phosphorus was oxidized. The loss of yield is <0.1 %.

Overall, the total yield in a modern phosphorus production plant is ca. 91 %.

The largest loss of yield is associated with the slag, so that it might appear advantageous to increase the coke rate in order to reduce the P_2O_5 content of the slag. However, this causes increased reduction of silica, thus increasing the Si content of the ferrophosphorus and reducing the P content. Ferrophosphorus of this quality is practically unmarketable, and, moreover, does not separate well from the slag. This leads to difficulties when tapping and granulating the slag. The increase in electricity consumption for the reduction of the SiO_2 must also be taken into

Table 5. Composition of ferrophosphorus (wt %)

Process	P	Si	Ti	Cr	V	Mn
Hoechst high percentage	26.6	0.3	4.5	0.2	0.6	0.5
Hoechst low percentage	20.0	9.5	5.8	0.2	0.5	0.7
TVA	24.0	1.7	1.7	0.2	0.3	1.1
FMC	27.1	0.2	0.3	4.5	5.4	0.5

account. In practice, it is not possible to reduce the P_2O_5 content of the slag below 1 %.

5. Production of Red Phosphorus

The conversion of white phosphorus into the more stable red form is an exothermic reaction. It must therefore be carried out under controlled conditions.

The process which has been used in Germany since 1936 in Bitterfeld and since 1953 in Knapsack is based on a Russian invention in which the phosphorus is converted in an open reactor. However, in Bitterfeld and Knapsack, closed reactors with charges of ca. 6 t are used (Fig. 8).

The reactor consists of a thick-walled iron rotary furnace resting on two hollow shafts with a driving mechanism, and surrounded by an air bath. It contains a small charge of steel balls. The white phosphorus in the reactor is first heated by the gas-heated air bath to 100 – 180 °C to drive off the water completely. Evaporating phosphorus is recovered in a condenser. After removal of the water, the reactor is sealed and heated to 270 – 275 °C over a period of ca. 24 h, whereby ca. 70 % of the white phosphorus is converted to the red form. A further increase in temperature causes the remainder to be converted very rapidly, with accompanying increases in pressure and temperature. The reactor is then cooled and filled with water, and the phosphorus is ground to the desired fineness.

In addition to the fact that the charge is relatively large, an important advantage of this process is that the phosphorus is not produced in the form of hard cakes, as in earlier autoclave processes, but as a powder.

In a modified version of this process, the full charge is not added to the cold reactor. Instead, white phosphorus is pumped in small batches into the dry reactor, which has been brought up to the reaction temperature. A large fraction of the phosphorus is converted directly into the red form. The advantages of this method are higher throughput and greater safety because the quantity of white phosphorus present in the reactor at any time is small [32].

The aqueous suspension of red phosphorus is steam heated in stirred vessels, and the remaining white phosphorus (ca. 0.1 %) is decomposed by adding sodium hydroxide solution. The phosphorus is filtered off and washed in rotary cellless filters or filter presses. The phosphorus is then continuously dried in steam-heated rotary or shelf dryers under a slight positive pressure of nitrogen. The nitrogen is circulated to remove the water vapor from the drying equipment. Air-entrained dust is removed in scrubbing towers and returned to the process.

The conversion and drying of the red phosphorus must be precisely controlled to avoid severe accidents. The pressure and temperature in the reactor can rise sharply if the conversion is uncontrolled. The oxygen content of the nitrogen in the shelf dryer must be controlled to avoid spontaneous explosive ignition of the red phosphorus. The explosion limit is at 3.5 % oxygen; in practice, the concentration of oxygen in the nitrogen is usually limited to ca. 1.5 %.

In the course of the treatment process (after heating to remove the last of the white phosphorus and up to the drying process), the red phosphorus can be treated to coat the surface of the grains with substances to perform various functions, e.g., to stabilize the red phosphorus (against P_2O_5 formation), to consolidate or "phlegmatize" (to prevent dust formation), to give hydrophobic properties, or to produce special properties to enable the material to be incorporated into plastics. The major producers of red phosphorus have developed special grades for the various applications in collaboration with their customers (see Section 6.2. and Chap. 7).

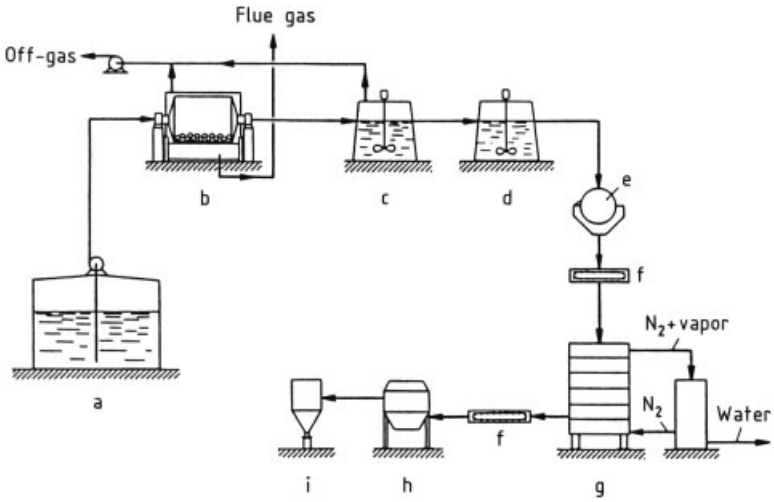

Figure 8. Manufacture of red phosphorus
a) Phosphorus storage; b) Reactor; c), d) Treatment of phosphorus – water mixture; e) Filtration; f) Conveyor; g) Dryer; h) Air classification and sieving; i) Predelivery storage

The finished and dried red phosphorus is packed in polyethylene bags and steel drums containing up to 50 kg, or in stainless steel containers (DIN 1.4301) containing up to 1.5 t. Grinding, sieving, air classification using nitrogen, and filling are all carried out in a nitrogen atmosphere.

To reduce the conversion time and produce an extremely fine product, TVA has suggested that the reaction should be carried out only to ca. 50 % completion, the unchanged white phosphorus then being removed by vacuum distillation [33].

Albright & Wilson [34] and Rinkagaku Kogyo [35] use a similar method to obtain a red phosphorus that is easier to handle. Erco Industries have applied for patent protection in Europe for a process in which the heat of the conversion of white to red phosphorus is removed by the evaporation of the white phosphorus during the conversion. About 40 % of the white phosphorus is recovered as red phosphorus, which is produced in a special amorphous form [36].

During the production of red phosphorus, wastewater and fumes are produced which have to be treated. If the production plant is part of a factory where white phosphorus is manufactured and treated, this is fairly simple.

The wastewater that has been in contact with white phosphorus is taken to the wastewater treatment plant attached to the phosphorus production plant, and the wastewater from the red phosphorus plant is filtered to remove the red phosphorus and taken to the central water treatment plant where reducing compounds are removed, e.g., by oxidation with air.

Air is extracted from the area of the reactors and from the vats where white phosphorus is treated with hot NaOH. This air must be oxidized, and the P_2O_5 produced must be absorbed. This is achieved by using the air for the combustion of phosphorus in a phosphoric acid plant.

6. Quality Specifications, Storage, and Transportation

6.1. White Phosphorus

Commercial white phosphorus, as produced by the electrothermal process using Söderberg electrodes, contains ca. 99.4 % P. The main impurities are water (0.3 %), which is taken up during condensation and storage, and benzene-soluble organic components, mainly produced during the baking of the Söderberg electrodes. The other impurities originate from the raw materials (Table 6). Arsenic is not dissolved in phosphorus, but replaces it in the phosphorus lattice, so that it is practically impossible to remove this impurity on an industrial scale.

Table 6. Analytical data for white phosphorus

Impurity	Content, wt %	
	Technical grade	Purified
Benzene-soluble organics	0.3	0.01
H_2O	0.25	0.10
As	0.01	0.01
S	0.05	0.05
Fe	0.005	0.005
Pb	0.0005	0.0005

Purified phosphorus with a reduced content of benzene-soluble organics is also produced (Table 6).

White phosphorus is usually stored as a liquid in heated vessels. It is constantly kept covered with warm water, and the free space above this is ventilated with nitrogen, so that no air is present in any part of the vessel.

Tonnage quantities are transported in heated tank cars with special connectors and valves. These tank cars normally contain only nitrogen at a slight positive pressure, but, if required by the customer, a protective water layer is also included.

For export, phosphorus is poured into internally coated steel drums with bungs. Small quantities are packed in wedge-shaped pieces (ca. 1 kg each) in drums with swaged-on lids. These drums always contain a protective water layer. The space above the liquid must in all cases amount to 2 % of the total volume at 60 °C.

The following transport regulations apply:

Road:	GGVS/ADR Class 4.2, No. 1
Rail:	GGVE/RID Class 4.2, No. 1
Sea:	ADNR Class 4.2
Air:	IATA-DGR not permitted
UN no.:	1381/I (phosphorus as cast blocks)
UN no.:	2447/I (phosphorus in wedge form)

6.2. Red Phosphorus

Red phosphorus, after the conversion process but before any special treatment, has the composition shown in Table 7.

The content of benzene-soluble organics and water is decreased in the red phosphorus compared with the original white phosphorus (furnace quality), while the iron content increases due to abrasion of the steel balls and of the reactor itself.

Heating with sodium hydroxide solution reduces the white phosphorus content to < 0.01 % (absence of Mitscherlich reaction [37]).

Depending on the intended application of the red phosphorus, the surface of the particles is coated by treating with various substances:

1) $Mg(OH)_2$ or $Al(OH)_3$ for stabilization (by combining with any P_2O_5 that may be formed).
2) Silicones for prevention of dust formation
3) Waxes to confer hydrophobic properties
4) Synthetic resins to condition for use in fireproofing plastics.

Table 7. Analytical data for red phosphorus

Component	Content, wt %
Benzene-soluble organic substances	0.1
H_2O	0.04
As	0.01
S	0.05
Fe	0.08
Pb	0.0005
H_3PO_4	0.05
White phosphorus [a]	0.1

[a] Decreases to 0.01 % white phosphorus on heating with sodium hydroxide solution.

The treatment of red phosphorus can involve various combinations of these possible techniques to suit customer demand, and affects the composition of the end product in a manner for which no general analysis can be given.

For many kinds of red phosphorus "products" (including various size distributions), storage is not possible. Each product is immediately packed under nitrogen in the customer's container.

Red phosphorus is packed in polyethylene bags under nitrogen, and these are packed in steel open-topped drums with lids. For tonnage quantities of red phosphorus, stainless steel containers are used. The following regulations apply:

Road:	GGVS/ADR	Class 4.1, No. 8
Rail:	EVO/RID (Appendix C)	Class 4.1.
Sea:	IMDG Code	Class 4.1.
Air:	IATA-(DGR)	No. 1413
UN no.:	1338	

7. Uses

White Phosphorus. In 1988, world production of white phosphorus was 750 000 t, of which ca. 83 % was used in the production of phosphoric acid, and of this ca. 50 % was used to manufacture phosphates for detergents. The remaining 50 % of the phosphoric acid was used to make phosphates for foodstuffs, pharmaceuticals, and animal feeds, a further fraction was used directly in foodstuffs, and some was used for surface treatment of metals.

The remaining 17 % of the phosphorus produced was used in the manufacture of phosphorus halides (PCl_3, $POCl_3$, PCl_5), P_2S_5, red phosphorus, P_2O_5, and metallic phosphides (e.g., magnesium phosphide and phosphor bronzes).

Red Phosphorus. The classical uses for red phosphorus are in the striking surface of match boxes and in pyrotechnics.

A considerable proportion is converted to metal phosphides such as aluminum phosphide. This has an important application as a pest control agent in grain silos (liberation of PH_3), especially in India and China.

A continually increasing application is in the manufacture of various flameproofing preparations, replacing chlorine-containing substances. It is incorporated in synthetic polymers such as condensations resins, polyamides, polyesters, polyepoxides, and polyurethane foams.

High-purity red phosphorus produced by distillation of PCl_3, reduction to white phosphorus, and conversion to the red form, is important in the manufacture of semiconductors, especially gallium phosphide (for laser diodes), although the quantities are very small.

8. Economic Aspects

White Phosphorus. World production of white phosporus increased from only a few thousand tonnes at the end of the 1900s to 10^5 t in 1939. By 1974, the year of maximum world production, this had increased to ca. 8.2×10^5 t. The reason for this large increase was the demand for tetrasodium diphosphate and especially pentasodium triphosphate, which were manufactured exclusively from thermal phosphoric acid and used principally in detergents.

During the 1970s, but initially in the United States in 1969, the use of phosphates in detergents decreased in the western World as a consequence of new regulations limiting the use of phosphates to reduce the eutrophication of inland waters. In the same period, energy costs (electricity and coke) increased, so that the cost of white phosphorus increased sharply, and purified wet phosphoric acid then became strongly competitive with thermal phosphoric acid produced from elemental phosphorus. Both these events caused a slowing down and eventual stagnation of the production of white phosphorus, so that phosphorus furnaces were closed down (see Fig. 9 and Table 8 [38]).

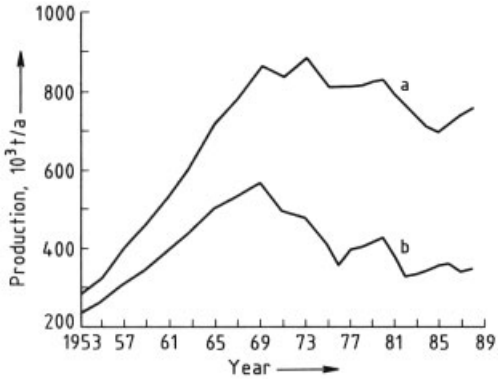

Figure 9. Phosphorus production, 1953 – 1988
a) World production; b) U.S. production

According to available information, the former Soviet Union and China, which are not linked to the world market, have considerably increased their capacity during the last 12 years (Table 8) [38].

The world market price of white phosphorus in 1988 was ca. 1640 €/t. European manufacturers are now, as in the past, increasingly at a disadvantage compared with the United States because of the large distances over which the phosphate rock is transported and the high costs of electricity and coke.

Red Phosphorus. World production of red phosphorus in 1988 was ca. 6500 t, and the world market price was ca. 4600 €/t. No reliable information concerning world capacity is available. Red phosphorus is manufactured in seven coun-

Table 8. Producers of white phosphorus and their capacities (1989)

Country	Company	No. of furnaces	Power, MW	Capacity, 10^3 t/a
United States	FMC	4	220	124
	Oxychem	3	115	54
	Monsanto	5	300	173
	Rhône-Poulenc	5	150	68
Canada	Albright & Wilson	4	152	90
Federal Republic of Germany	Hoechst	2	105	65
German Democratic Republic	VEB Stickstoffwerk Piesteritz	2	28	14
France	Atochem	1	28	12
Netherlands	Hoechst Holland	3	210	100
Italy	Ausidet	1	40	16
Former Soviet Union		ca. 30	ca. 1300	650
China		ca. 6	ca. 130	75
India	Excel Industries	2	10	5
	Star Chemicals	1	5	3
	United Phosphorus	1	5	2
Republic of South Africa	Chemfos	1	18	12
Total		71	2816	1463

tries, the three largest producers being Germany, India, and China.

9. Toxicology

Whereas red phosphorus is hardly absorbed at all on oral ingestion and therefore has very low toxicity, fat-soluble white phosphorus is highly toxic. However, experiments on animals have shown that the effects of red phosphorus on the liver and kidneys are similar to those of white phosphorus [39].

Kinetics and Metabolism. White phosphorus can be absorbed through the skin and via the respiratory and gastrointestinal tracts. In rats, the highest retention five days after oral administration was found in the liver, skeletal musculature, gastrointestinal tract, blood, and kidneys. In the body, phosphorus is oxidized to phosphate [39]. Following a single oral ingestion of isotopically labelled phosphorus, rats excreted about half of the radioactivity in the urine and about one-third in the feces within five days [40].

Acute Toxicity. The toxicity of white phosphorus is probably a result of its powerful reducing properties [39], which interfere with intracellular oxidation processes. In humans, as little as 15 mg can cause toxic effects. The lethal dose is reported to be 50 – 100 mg, corresponding to ca. 1 mg/kg of body weight [41–43]. The course of poisoning following oral ingestion is characterized by initial severe irritant effects on the digestive tract. The symptoms consist of stomach pains, nausea, vomiting, and shock. The stomach contents have a garliclike odor and may phosphoresce in the dark [42]. At this stage, death can occur through cardiovascular failure. Otherwise, the symptoms generally become less severe within one to three days. Following this recovery period, the symptoms of resorptive toxicity appear. Severe damage to liver cells occurs, and within a few days acute yellow hepatic atrophy can lead to death. The first symptoms manifest themselves as diarrhoea, vomiting (frequently blood-stained), sensitivity to pressure in the region of the liver, and increase in transaminases and bilirubin [44]. Kidney damage and cardiovascular changes also occur [39], [42]. Vascular damage leads to petechial and ecchymotic hemorrhages in the skin and mucous membranes, as well as hemorrhages in other organs [45], [46]. Collapse ultimately occurs, due partly to the toxic effects on the muscles of the heart [47], [48], and partly to liver and vascular damage.

Investigations of the acute oral toxicity in rats and mice gave LD_{50} values of 3 – 5 mg/kg [40]. Here also the picture of acute poisoning is principally characterized by liver damage (fatty degeneration) [49].

If white phosphorus comes into contact with the skin or eyes, serious burns result [40]. Experiments on animals have shown that dermal application can lead to liver and kidney damage [39]. Phosphorus vapor has an irritant action on the respiratory tract [40].

The treatment for acute oral phosphorus poisoning consists of immediate gastric lavage with 0.1 % potassium permanganate solution (oxidation of phosphorus to phosphoric acid) and high doses of animal charcoal, or gastric lavage with 0.2 % copper sulfate solution, which forms insoluble copper phosphide. The resorption of phosphorus can be prevented by administration of liquid paraffin. Additionally, treatment for symptoms of liver damage, shock, and dehydration is advisable [44]. The results of animal experiments [50], [51] suggest that 5 % sodium hydrogencarbonate with 3 % copper sulfate in 1 % hydroxyethylcellulose with the addition of 1 % lauryl sulfate could be suitable for the local treatment of burns caused by white phosphorus.

Subchronic and Chronic Toxicity. Chronic poisoning by phosphorus has become seldom. The first indications of chronic phosphorus poisoning are anemia, loss of appetite, gastrointestinal distress, chronic cough, garlic-like odor of the breath, and pallor [40]. More prolonged exposure leads to chronic atrophy of the bones. If the bones (usually of the jaw) then become infected, necrosis can occur [52]. The first signs of damage to the jaw consist of toothache and salivation, followed by tooth loss and swelling of the jaw [40].

Subchronic intake of phosphorus in the feed causes pathological changes to the bones of rats and rabbits (retardation of longitudinal bone growth and disturbance of dentine calcification in rats). Changes to the bones varying with dose rate also take place after chronic subcutaneous application of phosphorus to rats and guinea pigs, and also to a lesser extent after chronic administration in the feed in rats. Rats exhibited dose-related reduction of survival rates in both studies. With dogs, subchronic subcutaneous application of phosphorus led to fatty degeneration of the liver, kidney degeneration, changes in clinical-chemical parameters, and loss of weight [40]. Liver damage, including cirrhosis, was also observed following chronic oral administration [39].

Reproductive Toxicity. The administration of phosphorus to male and female rats before and after mating, and during gestation, led to the death of some female animals, apparently caused by birth difficulties. The number of live births was reduced [40].

Legal Requirements. The MAK value for white phosphorus is 0.1 mg/m^3 [53]. In the United States, a TLV-TWA value of 0.1 mg/m^3 has been prescribed [54].

10. References

1. K. Schmeiser: *Radionuclide,* Springer Verlag, Berlin 1963.
2. A. D. F. Toy: *The Chemistry of Phosphorus,* Pergamon Press, London 1975.
3. D. E. C. Corbridge: *Phosphorus,* Elsevier, Amsterdam 1978.
4. P. W. Bridgman, *J. Am. Chem. Soc.* **36** (1914) 1344; **38** (1916) 609.
5. D. E. C. Cordridge, *Top. Phosphorus Chem.* **3** (1966) 57 – 394.
6. H. Krebs in I. Kleinberg (ed.): *Inorganic Synthesis,* **vol. VII**, Mc Graw-Hill, New York 1963.
7. A. Brown, S. Rundgvidt, *Acta Cristallogr.* **19** (1965) no. 4, 684.
8. F. O. Rice, R. Potocki, K. Gosselin, *J. Am. Chem. Soc.* **75** (1953) 2003.
9. *Ullmann,* 4th ed., **18,** 365 ff.
10. *Phosphate Rock Statistics 1988,* IFA, Paris 1988.
11. *Ullmann,* 3rd ed., **13,** 516.
12. I. Kamlet, US 2 860 037, 1952.
13. Y. I. Sukkarnikow, A. M. Kuaner, A. A. Riev, *ISSLED. Obl. Neorg. Tekhnol.* 1972, 122.
14. F. W. Dorn, H. Harnisch, *Chem. Ing. Tech.* **42** (1970) 1209.
15. W. Goldmann, *Chem. Ing. Tech.* **32** (1960) 210.
16. VDI-Richtlinie: *Phosphor und anorganische Phosphorverbindungen, ausgenommen Düngemittel,* VDI 3450, May 1989.
17. *Ullmann,* 4th ed., **3,** 528 ff.
18. *Ullmann,* 4th ed., **3,** 530.
19. *Ullmann,* 4th ed., **3,** 531.
20. R. Mintrop, E. Roemer, *Tech. Mitt. Hüttenwerk Rheinhausen,* **4** (1955) 218.
21. K. Schmeiser, *Atompraxis* **6** (1960) 133.

22. M. M. Striplin, Jr., I. M. Potts, E. C. Marks, *J. Electrochem. Soc.* **106** (1959) 146.
23. *Ullmann,* 3rd ed., **13,** 512.
24. *Ullmann,* 4th ed., **2,** 240 ff.
25. Knapsack Griesheim, DE 1 048 885, 1957 (H. Harnisch, W. Bielenberg, F. Rodis).
26. Knapsack Griesheim, DE 1 002 743, 1955.
27. Knapsack Griesheim, DE 1 045 994, 1954 (H. Harnisch, F. Ritter, F. Rodis).
28. Hoechst, DE 2 228 636, 1972 (H. Ebert, U. Thümmler, H. Werner).
29. Knapsack, DE 2 135 546, 1971 (G. Hartlapp, H. Haas, K. H. Steudenbach).
30. Knapsack, DE 2 127 251, 1971 (U. Thümmler, J. Rothkamp).
31. *Ullmann,* 3rd ed., **13,** 515.
32. Hoechst, EP 15 384, 1979 (K. Lehr, G. Heymer, H. W. Stephan, U. Thümmler).
33. TVA, US 2 397 951, 1943.
34. Albright & Wilson, DE-OS 3 741 517 A1, 1987.
35. Rinkagaku Kogyo, EP 249 723, 1987.
36. Erco Industries, EP 144 194, 1984.
37. W. Fresenius, G. Jander: *Handbuch der analytischen Chemie,* **Part II, vols. IV b, V a, b, Part III, vols. V a, b,** Springer Verlag, Heidelberg 1953/56.
38. *Ullmann,* 4th ed., **18,** 298, Table 7.
39. R. P. Beliles in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A.** Wiley-Interscience, New York 1981, p. 2121.
40. Monsanto Co., *Material Safety Data Sheet,* St. Louis 1989.
41. S. Rao, R. H. Brown, *Ill. Med. J.,* Feb. 1974, 128–130.
42. H. J. Rubitsky, R. M. Myerson, *Arch. Intern. Med.* **83** (1949) 164–178.
43. C. L. Winek, W. D. Collom, E. P. Fusia, *Clin. Toxicol.* **6** (1973) no. 4, 541–545.
44. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* Thieme Verlag, Stuttgart–New York 1986.
45. K. Salfelder et al., *Beitr. Pathol.* **147** (1972) 321–340.
46. W. Wirth, C. Gloxhuber: *Toxikologie,* Thieme Verlag, Stuttgart–New York 1985.
47. R. J. Pietras, C. Stavrakos, R. M. Gunnar, J. R. Tobin, *Arch. Intern. Med.* **122** (1968) no. 5, 430–434.
48. R. C. Talley et al., *Am. Heart J.* **84** (1972) no. 1, 139–140.
49. J. M. Warnet, J. R. Claude, R. Truhaut, *J. Eur. Toxicol.* **2** (1973) 57–64.
50. N. Ben-Hur, A. Giladi, J. Appelbaum, Z. Neuman, *Br. J. Plast. Surg.* **25** (1972) 245–249.
51. N. Ben-Hur, J. Appelbaum, *Isr. J. Med. Sci.* **9** (1973) no. 1, 40–48.
52. P. Michaelis, *Arch. Gewerbepathol. Gewerbehyg.* **7** (1936) 477–485.
53. DFG: *Maximale Arbeitsplatzkonzentrationen und biologische Arbeitsstofftoleranzwerte 1989,* VCH Verlagsgesellschaft, Weinheim 1989.
54. ACGIH: *TLV's – Threshold Limit Values and Biological Indices for 1989–1990,* Cincinatti, Ohio, 1989.

Phosphorus Compounds, Inorganic

GERHARD BETTERMANN, Hoechst Aktiengesellschaft, Werk Knapsack, Hürth, Federal Republic of Germany (Chap. 1)

WERNER KRAUSE, Hoechst Aktiengesellschaft, Werk Knapsack, Hürth, Federal Republic of Germany (Chaps. 3 – 6)

GERHARD RIESS, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 2)

THOMAS HOFMANN, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Chap. 7)

1. Phosphorus – Oxygen Compounds . 1
1.1. Phosphorus Pentoxide 2
1.2. Other Phosphorus Oxides 3
1.3. Phosphorus Acids and their Salts . 4
1.3.1. Phosphinic Acid 5
1.3.2. Phosphonic Acid 6
2. Phosphorus – Halogen Compounds 6
2.1. Phosphorus Tri- and Pentahalides . 7
2.1.1. Phosphorus Trichloride 7
2.1.2. Phosphorus Pentachloride 8
2.1.3. Phosphorus Tribromide 8
2.1.4. Phosphorus Fluorides 9
2.2. Phosphoryl and Thiophosphoryl Halides 9
2.2.1. Phosphoryl Chloride 9
2.2.2. Thiophosphoryl Chloride 10
3. Phosphorus Sulfides 10
4. Phosphides 13
5. Phosphorus – Hydrogen Compounds 14
6. Phosphorus – Nitrogen Compounds 15
7. Toxicology 16
8. References 18

1. Phosphorus – Oxygen Compounds

Approximately one decade after the discovery of phosphorus, R. BOYLE in 1680 obtained its most important oxide, the pentoxide, by combustion of phosphorus. In 1777, LE SAGE prepared phosphorus trioxide mixed with other phosphorus oxides, and in 1886 phosphorus tetroxide was described by T. E. THORPE and A. E. TUTTON as a compound distinct from P_2O_5 and P_2O_3.

The existence of P_4O, P_2O, PO, and PO_2 in an argon matrix has been demonstrated spectrographically. The PO molecule has also been detected in the vapor phase. However, these compounds are unstable under normal conditions.

A suboxide of phosphorus with the approximate formula $(P_4OH)_n$ is produced by the action of oxygen on phosphorus in a water vapor atmosphere at 80 – 160 °C.

The phosphorus oxides are formed in strongly exothermic reactions between the elements, and occur in dimeric and polymeric form. The only industrially important oxide of phosphorus is the pentoxide.

1.1. Phosphorus Pentoxide

Physical Properties. Phosphorus pentoxide [*1314-56-3*], tetraphosphorus decaoxide, P_4O_{10}, M_r 141.95, the anhydride of orthophosphoric acid, is at room temperature a fluffy, white, snowlike, odorless powder that sublimes under standard pressure at 360 °C, and separates as hexagonal crystals (H form) on slow cooling of the vapor. More rapid cooling produces an amorphous form. Commercial phosphorus pentoxide consists mainly of the H form, though it usually also contains some of the amorphous

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_527

powder form. Both the vapor and the solid consist of P_4O_{10} molecules.

Each phosphorus atom is tetrahedrally surrounded by four oxygen atoms.

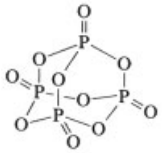

P_4O_{10}

At a pressure of 0.5 MPa, the crystalline H form can be melted to give a colorless, clear liquid of low viscosity and high vapor pressure. Liquid phosphorus pentoxide also consists of P_4O_{10} molecules.

If hexagonal P_4O_{10} is heated to 400 °C in a closed system, after 2 h a metastable orthorhombic modification (O form) is produced. This has a melting point of 562 °C, and consists of chalky aggregates of small crystals. If the H form is heated to 450 °C for 24 h under the same conditions, a stable orthorhombic O′ modification is obtained consisting of readily cleavable aggregates and having a melting point of 580 °C. Both crystal forms are composed of $(P_2O_5)_x$ layers in which each phosphorus atom is tetrahedrally surrounded by four oxygen atoms. They melt to form a viscous involatile liquid, which on quenching gives a vitreous modification with a melting point of 565.5 °C.

After exposure to light, phosphorus pentoxide phosphoresces with a green color. As the temperature is reduced, the intensity of the light increases, reaching a maximum at −180 °C.

Chemical Properties. The best known property of P_4O_{10} is its remarkable tendency to take up water. Below 100 °C, it is the strongest known dehydrating agent. The water vapor partial pressure of the hydrolysis product is $<10^{-4}$ Pa at room temperature. However, the surface of the phosphorus pentoxide soon becomes covered with a film of polyphosphoric acid, completely isolating the reactive oxide underneath. Therefore, when P_4O_{10} is used as a drying agent, it should be distributed over a large surface area (e.g., on glass wool). Reaction with water first forms cyclic tetrametaphosphoric acid $(HPO_3)_4$, which, on taking up a further H_2O molecule, undergoes cleavage to give the straight-chain tetraphosphoric acid $H_6P_4O_{13}$. Further hydrolytic decomposition produces diphosphoric acid $H_4P_2O_7$ and then orthophosphoric acid. Reaction with ethers causes cleavage of different P−O bonds to those cleaved by water, giving esters of isotetrametaphosphoric acid.

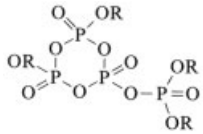

Alcohols react with P_4O_{10} to form mono- and diesters of orthophosphoric acid.

Phosphorus pentoxide is also used for the dehydration of chemical compounds. For example, anhydrous nitric and sulfuric acid yield the acid anhydrides N_2O_5 and SO_3, while acid amides are dehydrated to form nitriles.

Production. The combustion of phosphorus in an excess of dry air or oxygen produces phosphorus pentoxide:

$$4\,P + 5\,O_2 \rightarrow P_4O_{10} \quad \Delta H = -2994.6\,\text{kJ}$$

On the industrial scale, pure liquid phosphorus is introduced in finely divided form into a special burner containing dry, oxygen-enriched air at elevated pressure. Sometimes further dry air is added during combustion.

The combustion chambers can be the same as those used in the production of phosphoric acid (→ Phosphoric Acid and Phosphates), or can be of a special type [1]. They are made of stainless steel, and the walls of the combustion chamber are protected from the intense heat of reaction by intensive water cooling. The combustion temperature is ca. 2000 °C. Some of the phosphorus pentoxide vapor deposits on the cold wall of the combustion chamber, forming a vitreous coating which protects the wall from corrosion. This coating sublimes off at temperatures above 605 °C.

The phosphorus pentoxide vapor leaves the combustion chamber at ca. 900 °C, and passes through a pipe into a dwell chamber. The output of the production unit depends to a large extent on the intensity of cooling of this part of the apparatus. Up to 95 % of the P_4O_{10} produced collects in the dwell chamber as a white,

dry powder. Approximately 5 % of the product is lost in the off-gases, which contain 7 % oxygen and have a temperature of 120 °C at the dwell chamber exit. The gases are scrubbed with phosphoric acid, e.g., in a Venturi apparatus, to remove the P_4O_{10}. If the phosphorus pentoxide plant is operated near to a phosphoric acid plant, the off-gases from phosphorus pentoxide production can be fed directly to the acid plant [2].

Coarse particles of agglomerated phosphorus pentoxide of low reactivity can be obtained if the vapor is condensed in a fluidized bed in which water-cooled stainless steel cold fingers or spiral condensing tubes are immersed [3]. The conventional condensation system described above can also be used to give various pentoxide qualities by varying the condensation conditions [4].

The reactivity of phosphorus pentoxide can be reduced by heat treatment [5].

In all the cases described above, the phosphorus pentoxide produced can be shown by X-ray analysis to have a hexagonal crystalline form.

Quality. Phosphorus pentoxide produced in modern plants contains 99.8 % P_4O_{10} and < 0.01 % P_4O_6. It does not even decolorize a pale pink solution of $KMnO_4$ after several hours, showing that oxidizable substances (elemental phosphorus, lower phosphorus oxides) are not present. Other data:

Arsenic	< 50 ppm
Iron	< 5 ppm
Fluorine	< 5 ppm
Heavy metals (e.g., Pb)	< 1 ppm
Residue insoluble in water	<0.02 %
Color	pure white
Bulk density	ca. 0.7 kg/L

The technical product was formerly purified by sublimation, but is now produced directly with the required purity.

Uses. Apart from its use in the production of phosphoric acid, phosphorus pentoxide is mostly used in organic chemistry for dehydration reactions. It is also used as a drying agent for solids, liquids, and gases with which it does not react. It is a convenient medium for absorption of moisture in vacuum systems, e.g., in the production of electric light bulbs.

The production of phosphoric acid esters from phosphorus pentoxide is described in → Phosphorus Compounds, Organic. The addition of ca. 0.5 % phosphorus pentoxide to asphalt increases its softening temperature and prevents brittleness at low temperature [6]. The production of ammonium polyphosphate for fireproofing is described in [7], [8].

1.2. Other Phosphorus Oxides

Phosphorus trioxide [*12440-00-5*], tetraphosphorus hexaoxide, P_4O_6, M_r 219.88, the anhydride of phosphonic acid, forms soft, waxy, colorless, monoclinic crystals, ϱ 2.135 g/cm^3, *mp* 23.8 °C, *bp* 175.3 °C. It has an unpleasant smell and is very poisonous. It can be crystallized from carbon disulfide, and is slightly soluble in benzene, carbon tetrachloride, chloroform, and dioxane. White phosphorus is soluble in phosphorus trioxide (ca. 1.7 g/100 g P_4O_6 at 25 °C). Phosphorus trioxide consists of P_4O_6 molecules in both the vapor and the solid.

Phosphorus trioxide is slowly oxidized to phosphorus pentoxide by air at room temperature, and spontaneously ignites on heating to 70 °C. When dissolved in cold water with vigorous agitation to give a dilute solution, exclusively H_3PO_3 is formed; without agitation, H_3PO_4 and an undefined solid yellow substance are formed in addition to H_3PO_3. In hot water, the reaction becomes violent, forming phosphine, phosphoric acid, and red phosphorus.

Phosphorus trioxide reacts with chlorine and bromine to form halides. Reaction with iodine in a sealed tube gives P_2I_4. With HCl, phosphonic acid and phosphorus trichloride are formed [9].

Production. When white phosphorus is burnt at 46 – 50 °C in a silica tube in the presence of a restricted amount of air, phosphorus trioxide is formed in a highly exothermic reaction.

$$4\,P + 3\,O_2 \rightarrow P_4O_6 \quad \Delta H = -1646.4\,\text{kJ}$$

This method, described by Thorpe and Tutton, was later improved to give yields of 50 – 60 % [10]. The phosphorus trioxide produced always contains residual elemental phosphorus, which can be converted into the red form by irradiating with ultraviolet light [11], and can

then easily be removed. A continuous process for the production of P_4O_6 is described in [12].

Attempts to produce phosphorus trioxide by reaction of phosphorus trichloride with phosphonic acid have so far been unsuccessful.

Phosphorus trioxide has not yet gained industrial importance.

Phosphorus Tetroxide [*12137-38-1*], diphosphorus tetraoxide, P_2O_4, M_r 125.95, can be sublimed at high temperatures, forming colorless, transparent crystals, ϱ 2.54 g/cm^3. It is the mixed anhydride of phosphonic and orthophosphoric acid, and dissolves readily in water with liberation of heat to form H_3PO_3 and H_3PO_4. Like P_2O_5, it is dimeric in the vapor phase.

Phosphorus trioxide disproportionates on heating at 210 °C for several days into red phosphorus and phosphorus tetroxide. Other phosphorus(III,V) oxides are also formed.

Tetraphosphorus heptaoxide P_4O_7 and tetraphosphorus nonaoxide P_4O_9 are III/V oxides formed as side products in the disproportionation of phosphorus trioxide to phosphorus tetroxide [13], or can be obtained by the controlled reduction of phosphorus pentoxide with red phosphorus [14]. Of the two crystalline forms, the rhombic α-form contains P_4O_8 and P_4O_9 molecules, while the monoclinic β-form consists of P_4O_7 and P_4O_8 molecules. The pure compounds P_4O_7, P_4O_8, and P_4O_9 have not been isolated.

A blue-violet solid with the composition P_2O_6, probably a phosphorus peroxide, is produced when a mixture of P_2O_5 vapor and hydrogen is subjected to an electric discharge under reduced pressure at 150 – 220 °C. The peroxide reacts with water to form peroxodiphosphoric acid ($H_4P_2O_8$). The violet solid is stable in the absence of water [15].

1.3. Phosphorus Acids and their Salts

(For phosphoric acid, see → Phosphoric Acid and Phosphates)

The oxo acids with the general formulae H_3PO_n (ortho acids) and $HPO_n{}^{-1}$ (meta form) and the diacids with the general formula $H_4P_2O_n$ ($n = 4 - 8$) are known, and are listed in Table 1 in order of increasing oxidation state of the phosphorus.

In aqueous solution, the anions of phosphinic, phosphonic, and phosphoric acids are present in the following forms, as shown spectroscopically.

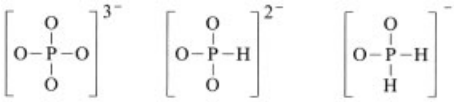

Phosphate Phosphite Hypophosphite

The hypophosphite (phosphinite) anion contains two hydrogen atoms that cannot be neutralized by alkalies. Phosphinic acid is thus monobasic. Likewise, phosphonic acid is dibasic and phosphoric acid tribasic.

In addition to the simple phosphoric acids, there are many polyphosphoric acids with the general formula $H_{n+2}P_nO_{3n+1}$ with three or more phosphorus atoms in the molecule and containing P–O–P bonds, as well as acids with a lower oxygen content that contain P–P bonds:

Diphosphonic acid, $H_4P_2O_5$, with the anion

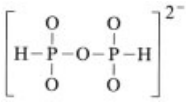

Diphosphoric acid, pyrophosphoric acid, $H_4P_2O_7$, with the anion

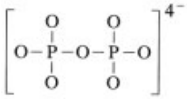

Triphosphoric acid, $H_5P_3O_{10}$, with the anion

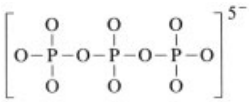

A representative of the low oxygen polyphosphoric acids is hypodiphosphoric acid, $H_4P_2O_6$, with the anion

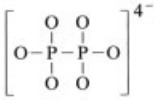

Table 1. Mono- and diphosphoric acids

Oxidation state	Acids of type H_3PO_n and HPO_n^{-1}				Acids of type $H_4P_2O_n$		
	Formula		Name	Salts	Formula	Name	Salts
	Ortho-	Meta-					
+1	H_3PO_2		phosphinic acid, hypophosphorous acid, phosphoric(I) acid	hypophosphites, phosphates(I)			
+2					$H_4P_2O_4$	hypodiphosphonic acid, diphosphoric(II) acid	hypodiphosphites, diphosphates(II)
+3	H_3PO_3	HPO_2	phosphonic acid, phosphorous acid, phosphoric(III) acid	phosphites, phosphates(III)	$H_4P_2O_5$	diphosphonic acid, diphosphoric(III) acid	diphosphites, diphosphates(III)
+4					$H_4P_2O_6$	hypodiphosphoric acid, diphosphoric(IV) acid	hypodiphosphates, diphosphates(IV)
+5	H_3PO_4	HPO_3	phosphoric acid, phosphoric(V) acid	phosphates, phosphates(V)	$H_4P_2O_7$	diphosphoric acid, pyrophosphoric acid, diphosphoric(V) acid	diphosphates, diphosphates(V)

1.3.1. Phosphinic Acid

Phosphinic acid [*6303-21-5*], hypophosphorous acid [*14332-09-3*], H_3PO_2, M_r 66.0, *mp* 26.5 °C, forms colorless leaflets. It is obtained by reaction of its salts with sulfuric acid. The salts are usually prepared by heating phosphorus with alkali or alkaline earth metal hydroxides, whereby a disproportionation reaction occurs. On an industrial scale, calcium hydroxide is used:

$$2\,P_4 + 3\,Ca(OH)_2 + 6\,H_2O \rightarrow 2\,PH_3 + 3\,Ca(H_2PO_2)_2$$

$$Ca(H_2PO_2)_2 + H_2SO_4 \rightarrow CaSO_4 + 2\,H_3PO_2$$

On evaporating the aqueous solution, the phosphinic acid crystallizes with a lamellar structure. Oxidation of PH_3 by iodine in aqueous solution also produces phosphinic acid.

$$PH_3 + 2\,I_2 + 2\,H_2O \rightarrow H_3PO_2 + 4\,HI$$

Phosphinic acid forms esters with two tautomeric structures.

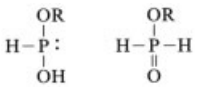

Phosphinic acid is a considerably stronger reducing agent than phosphonic acid, being converted firstly into phosphonic and then phosphoric acid. On heating to 130 – 200 °C, it disproportionates into phosphine and phosphoric acid.

$$2\,H_3PO_2 \rightarrow PH_3 + H_3PO_4$$

Both the hydrogen atoms bound to the phosphorus in phosphinic acid can be replaced by alkyl groups to give mono- and dialkyl phosphinic acids.

Phosphinic acid can be distinguished from phosphonic acid by its reaction with copper sulfate solution. It reduces $CuSO_4$ not just to metallic copper, but to copper hydride CuH.

All salts of phosphinic acid, MH_2PO_2, are very soluble in water.

Hypodiphosphoric acid, $H_4P_2O_6$, can be prepared by first oxidizing red phosphorus with H_2O_2 or other oxidizing agents in strongly alkaline solution to form $Na_2H_2P_2O_6$ or $BaH_2P_2O_6$. Good yields are obtained when the red phosphorus is oxidized at 12 – 15 °C by $NaClO_3$ [16]. The free acid is obtained by reacting the salts with sulfuric acid. On evaporation, the acid is obtained as the hydrate $H_4P_2O_6 \cdot H_2O$ (*mp* 62 °C). The anhydrous compound $H_4P_2O_6$ (*mp* 55 °C) can be obtained by vacuum drying over phosphorus pentoxide. However, diphosphoric(IV) acid, $(HO)_2OP{-}PO(OH)_2$, is unstable, and gradually changes into the isomeric diphosphoric (III,V) acid, $(HO)_2P{-}O{-}PO(OH)_2$.

1.3.2. Phosphonic Acid

Phosphonic acid [*13598-36-2*], phosphorous acid [*10294-56-1*], H_3PO_3, M_r 82.00, was obtained as a mixture with other oxoacids by the oxidation of white phosphorus by moist air in 1777 by Le Sage. In the anhydrous state, it forms colorless crystals, ϱ 1.65 g/cm^3, *mp* 70.1 °C, which are very soluble in water. It is a dibasic acid, and so forms both hydrogenphosphites and phosphites.

Phosphonic acid can be obtained by dissolving P_4O_6 in cold water, or by reaction of its calcium salt with sulfuric acid. It is now produced industrially by spraying phosphorus trichloride with water vapor into a hot reaction zone where the excess water and the hydrochloric acid formed are evaporated off [17].

$$PCl_3 + 3\,H_2O \rightarrow H_3PO_3 + 3\,HCl$$

This reaction has been operated as a continuous process by Hoechst for many years. In a Monsanto process [18], phosphonic acid is produced by indirect electrolytic oxidation of elemental phosphorus.

The trialkyl derivatives (esters) exist in two forms: $P(OR)_3$ can be converted to the more stable form $OPR(OR)_2$.

Phosphonic acid is a strong reducing agent. On heating in the dry state, it disproportionates to form phosphoric acid and phosphine.

The salts of phosphonic acid are not oxidized in air, whereas the free acid is slowly converted into orthophosphoric acid.

Phosphonic acid is used as the starting material for preparing basic lead phosphite, which is added to chlorine-containing plastics (e.g., PVC) to stabilize them against light and heat. Its principal use is in the synthesis of various phosphorus(III) compounds [19].

Phosphonic acid is supplied in flake form in polyethylene-lined paper sacks. The H_3PO_3 content of the commercial product is ca. 98.5 % (0.5 % H_3PO_4, the rest water).

2. Phosphorus – Halogen Compounds

Of the many compounds of phosphorus with the halogens, only those of the type PX_3, PX_5, POX_3, and PSX_3 (X = Cl or Br) are of industrial and economic importance. The corresponding fluorine, iodine, and polymeric halogen compounds are of only academic interest. The most important phosphorus – halogen compound is phosphorus trichloride, for which world production capacity in 1989 was estimated at >300 000 t. All the industrially important derivatives such as phosphorus pentachloride, phosphoryl chloride, and thiophosphoryl chloride are produced from phosphorus trichloride. Some figures for world consumption in 1989 are: PCl_3, 168×10^3 t; PCl_5, 4×10^3 t; $POCl_3$, 82×10^3 t. The stability of the tri- and pentahalides decreases with increasing atomic mass of the halogen atom. All the phosphorus halides are extremely sensitive to atmospheric moisture. They react very vigorously with water, the halogens being replaced by OH groups. Hydrogen halides, phosphonic and phosphoric acids, and sometimes phosphine (spontaneously flammable) are formed. A large excess of water must be used when dealing with spillages [20]. The phosphorus chlorides are neither flammable nor explosive.

Properties. Some physical properties of the most important phosphorus halides are given in Table 2 [21–23]. Phosphorus trichloride, phosphoryl chloride, and thiophosphoryl chloride are liquids at room temperature; phosphorus pentachloride is a solid. The molecules of the phosphorus trihalides are trigonal pyramidal, with the phosphorus atom at the apex. The structure of the phosphorus pentahalides is a trigonal bipyramid with the phosphorus atom at the center. This is true for phosphorus pentafluoride in all states of aggregation, but for the other pentahalides only when in the gas phase or in solution. Crystalline phosphorus pentachloride and pentabromide are composed of ions, i.e., $(PCl_4)^+$ $(PCl_6)^-$ and $(PBr_4)^+$ Br^-. In addition to the compounds listed in Table 1, the following mixed phosphorus trihalides are known:

	mp	*bp*
PF_2Cl	− 164.8 °C,	− 47.3 °C
$PFCl_2$	− 144.0 °C,	+ 13.85 °C
PF_2Br	− 133.8 °C,	− 16.1 °C
$PFBr_2$	− 115.0 °C,	+ 78.4 °C

Table 2. Some physical properties of phosphorus – halogen compounds

Substance	PF_3	PF_5	PCl_3	PCl_5	PBr_3	PBr_5	$POCl_3$	$PSCl_3$
Appearance	colorless	colorless	colorless	white	colorless	red/yellow	colorless	colorless
Physical state	gas	gas	liquid	solid	liquid	solid	liquid	liquid
Density (20 °C), g/cm^3			1.575	2.114	2.880		1.675	1.668
Vapor density, g/L	3.907	5.805						
bp, °C	− 101.2	− 84.5	76.1	159 (sublimes)	173.2	> 106 (decomp.)	105.1	125
fp, °C	− 151.5	− 93.8	− 93.6	160.5	− 41.5	83.8 (sublimes)	1.25	− 36
Heat of formation, kJ/mol	− 918.8	− 1595.8	− 319.7	− 443.5	− 184.5	− 269.9	− 597.1	− 344

2.1. Phosphorus Tri- and Pentahalides

2.1.1. Phosphorus Trichloride

Phosphorus trichloride [*7719-12-2*], PCl_3, M_r 137.33, was discovered in 1808.

Some physical properties are as follows:

Heat of evaporation		31 kJ/mol
Molar heat capacity		120.9 J mol^{-1} K^{-1}
Vapor pressure:	20 °C	12.97 kPa
	70 °C	85.75 kPa
Viscosity:	0 °C	0.65 cP
	50 °C	0.438 cP

Phosphorus trichloride is a clear, volatile liquid, which is rapidly hydrolyzed by water with evolution of heat to form phosphonic acid $HP(O)(OH)_2$, and hydrochloric acid. Phosphorus trichloride has an irritant effect on the mucous membranes. It is soluble in benzene, chloroform, carbon tetrachloride, and diethyl ether, and is miscible with phosphoryl chloride. Addition of chlorine, oxygen, and sulfur to the phosphorus electron lone pair occurs readily, giving phosphorus(V) compounds.

Production. Phosphorus trichloride is formed by the combustion of phosphorus in a dry stream of chlorine.

$$P_4 + 6\,Cl_2 \rightarrow 4\,PCl_3 \quad \Delta H = -1276\,\text{kJ/mol}$$

In a Hoechst continuous process, molten white phosphorus and gaseous chlorine react in previously produced phosphorus trichloride. The formation of phosphorus pentachloride is prevented by the presence of a small excess of phosphorus. The heat of reaction, ca. 10 times the heat of evaporation, keeps the system at its boiling point, and the phosphorus trichloride distills off. The vapors are fractionated as reflux takes place, and the PCl_3 condenses in air-cooled condensers. The rate of removal of PCl_3 (i.e. the production rate) is equivalent to the feed rate of phosphorus and chlorine. The process is controlled by the boiling temperature, which is determined by the phosphorus content [24]. The rate of chlorine flow is fixed, and the phosphorus feed rate is adjusted manually. The phosphorus trichloride produced contains no free phosphorus and needs no further treatment. The impurities from the phosphorus remain in the reactor, and are removed periodically, i.e., the phosphorus trichloride is completely driven off by distillation, and the tarry mass remaining is removed and can be burnt.

In another process, liquid phosphorus and gaseous chlorine are reacted in a combustion chamber to form phosphorus trichloride. The product is then distilled [25].

The commercial product contains ca. 99 % phosphorus trichloride, the main impurity being phosphoryl chloride. The boiling range is 74.5 – 76.5 °C.

Phosphorus trichloride is transported in steel drums or tank cars, U.N. no. 1809, RID/ADR/IMO Class 8, Section 11 a. Road transport is in tank trucks, for which pure nickel has good corrosion resistance.

Uses. Phosphorus trichloride is mainly used for the manufacture of phosphoryl chloride, phosphorus pentachloride, thiophosphoryl chloride, and phosphonic acid. It is the starting material in the production of organophosphorus

compounds. Phosphite esters are formed by reaction with alcohols and phenols. Phosphorus trichloride is an important intermediate in the production of phosphate ester insecticides. Hydrocarbons, particularly alkenes, react with phosphorus trichloride to form phosphinic and phosphonic acids (see → Phosphorus Compounds, Organic). In organic chemistry, phosphorus trichloride is used as a chlorinating agent and catalyst, e.g., fatty acid chlorides are obtained by the reaction of fatty acids with phosphorus trichloride, phosphonic acid being the byproduct [26].

2.1.2. Phosphorus Pentachloride

Phosphorus pentachloride [*10026-13-8*], PCl_5, M_r 208.23, is a white, crystalline mass when pure. It fumes in air and has a strong, pungent odor. It was discovered in 1810 by Davy. Phosphorus pentachloride has a vapor pressure of 1.6 Pa at 20 °C, and is 40 % dissociated into phosphorus trichloride and chlorine at 180 °C, 80 % at 250 °C, and in the vapor phase at 300 °C is fully dissociated.

Production. Phosphorus pentachloride can be prepared by the reaction of excess chlorine with phosphorus trichloride.

$$PCl_3 + Cl_2 \rightarrow PCl_5 \quad \Delta H = -125\ \text{kJ/mol}$$

Small quantities of phosphorus pentachloride can be prepared by the reaction of chlorine with phosphorus trichloride dissolved in carbon tetrachloride. The phosphorus pentachloride crystallizes out and is purified by vacuum distillation. Industrial batch manufacture is performed by passing chlorine into phosphorus trichloride. The heat of reaction is removed by cooling the vessel and by the latent heat of evaporation. With careful reaction control in a stirred stainless steel reactor, a granular, free-flowing product is obtained. This can be packed and transported in plastic containers, U.N. no. 1806, RID/ADR Class 8, Section 11 a, IMO Class 8. When working with phosphorus pentachloride, it is important to keep it cool, to keep the storage time short, and to prevent contact with moisture.

Uses. Phosphorus pentachloride is used in organic chemistry as a mild chlorinating agent. Two of the three chlorine atoms are easily split off. It is used in the pharmaceutical industry in the manufacture of penicillin and cephalosporin antibiotics. Phosphonic acids are produced on reaction with alkenes. In aluminum metallurgy, it is used as a grain refiner for Al – Si alloys. With compounds containing carboxyl groups, phosphorus pentachloride reacts to form phosphoryl chloride and the corresponding acid chlorides, with liberation of hydrogen chloride.

2.1.3. Phosphorus Tribromide

Phosphorus tribromide [*7789-60-8*], PBr_3, M_r 270.7, U.N. no. 1808, is a strongly corrosive liquid which fumes in air. All methods of preparation start from the elements. The reaction with white phosphorus is very violent, and the phosphorus is therefore dissolved in carbon disulfide or submerged in benzene. Phosphorus tribromide is mainly used as a brominating agent in organic chemistry.

2.1.4. Phosphorus Fluorides

Phosphorus trifluoride [*7783-55-3*], PF_3, M_r 87.96, is a colorless gas which does not fume in air when pure. It is not prepared from the elements but by exchange reactions between other phosphorus and fluorine compounds, e.g., by dropwise addition of arsenic trifluoride to phosphorus trichloride. This forms gaseous phosphorus trifluoride, which is washed with water, dried, and isolated by freezing. Other methods of preparation are the reaction of copper phosphide with lead fluoride or of phosphorus tribromide with zinc fluoride. Phosphorus trifluoride has little commercial importance. Reaction with hydrocarbons produces fluorocarbons [27].

Phosphorus pentafluoride [*7647-19-0*], PF_5, M_r 125.96, U.N. no. 2198, is a colorless gas, which has an unpleasant odor and fumes in air. It is the most stable pentahalide of phosphorus, and is prepared in a similar way to phosphorus trifluoride, i.e., by double decomposition, for example as follows:

$$3\,PCl_5 + 5\,AsF_3 \rightarrow 3\,PF_5 + 5\,AsCl_3$$

Other methods of preparation include the reaction of phosphorus trichloride with fluorine [28], or with chlorine and calcium fluoride [29]. The gaseous product is purified by passing it over an alkali metal to remove the residual chlorides. Phosphorus pentafluoride is a relatively mild fluorinating agent of very minor industrial importance.

2.2. Phosphoryl and Thiophosphoryl Halides

In addition to the compounds thiophosphoryl fluoride, phosphoryl chloride, and thiophosphoryl chloride, many mixed substituted phosphoryl and thiophosphoryl halides have been described, e.g., POF_2Cl (*mp* −96.4 °C, *bp* +3.1 °C), $POFCl_2$ (*mp* −80.1 °C, *bp* +52.9 °C), POFClBr (*bp* +79 °C), and PSFClBr.

Condensed compounds such as pyrophosphoryl chloride $P_2O_3Cl_4$, pyrophosphoryl fluoride $P_2O_3F_4$, and tetraphosphoryl decachloride $P_4O_5Cl_{10}$ are also known, all of which contain P−O−P groups and should be regarded as halides of condensed phosphoric acids.

Some properties of the phosphoryl and thiophosphoryl halides are given in Table 3. They are all colorless at room temperature.

All these compounds are very reactive, the reactivity increasing from the fluorides to the iodides. Like the binary halides, they are sensitive to moisture and hydrolyze to give the corresponding phosphorus acids. In the phosphoryl and thiophosphoryl halides, the phosphorus atom is tetrahedral.

2.2.1. Phosphoryl Chloride

Phosphoryl chloride [*10025-87-3*], $POCl_3$, M_r 153.33, was discovered in 1847 by WURTZ. It is a colorless liquid which fumes in air and has a pungent odor. It is obtained by oxidizing phosphorus trichloride with air or oxygen.

$$PCl_3 + {}^{1}\!/_{2}\,O_2 \rightarrow POCl_3 \quad \Delta H = -279.5\,\text{kJ/mol}$$

It can be produced industrially either batchwise or in a continuous process. A finely divided stream of oxygen is reacted with phosphorus trichloride until the freezing point of the product is 1 °C. The reaction is accelerated by carrying it out under pressure. The presence of orthophosphoric acid has a catalytic effect, but traces of metallic impurities such as copper and cobalt, or of sulfur compounds, decrease the reaction rate. In a Hoechst process, phosphorus trichloride and oxygen are passed in cocurrent through a stainless steel water-cooled packed column. The oxidation process, i.e., the uptake of oxygen, can be followed by the characteristic temperature changes. The first product, which leaves the column at a maximum temperature of 35 °C, has a freezing point of ca. 0.9 °C, and is taken from the top of the column into a second column where it is again treated with oxygen. This procedure avoids large temperature increases. The excess oxygen leaving the plant passes through a cold trap where entrained phosphoryl chloride is condensed out and recovered. The off-gas is then scrubbed with water. Commercial phosphoryl chloride contains ca. 99.5 % $POCl_3$ and traces of phosphorus trichloride.

It is transported in drums with polyethylene liners, and in lead-lined or stainless steel heated tank cars or tanker trucks (U.N. no. 1810, Class 8, Section 12).

Uses. Phosphoryl chloride is used on a large scale to produce phosphate esters by reaction with alcohols and phenols. It is an ionizing solvent, dissociating to a small extent according to:

$$POCl_3 \rightleftharpoons POCl_2^+ + Cl^-$$

2.2.2. Thiophosphoryl Chloride

Thiophosphoryl chloride [*3982-91-0*], $PSCl_3$, M_r 169.51, U.N. no. 1837, is a colorless liquid with an unpleasant odor. It is hydrolyzed by water to form hydrogen sulfide, orthophosphoric acid, and hydrochloric acid. It is soluble in benzene and carbon disulfide.

Production. Thiophosphoryl chloride is prepared by the direct reaction of excess sulfur with phosphorus trichloride at ca. 180 °C, either as a batch process in an autoclave or under reflux. The product is purified by distillation. Yields are up to 98 %. The reaction is accelerated and the

Table 3. Properties of phosphoryl and thiophosphoryl chlorides

Property	POF_3	$POCl_3$	$POBr_3$	PSF_3	$PSCl_3$
Physical form at room temperature	gas	liquid	crystal	gas	liquid
Density, g/cm^3		1.675	2.82		1.668
mp, °C	−39	1.2	55	− 148.8	− 36.2
bp, °C	−39.7	105.1	191.7	−52.2	125

reaction temperature reduced by the addition of catalysts such as halides of aluminum [30] or iron [31], or disulfur dichloride [32]. Alternatively, thiophosphoryl chloride can be prepared by the reaction of phosphorus pentachloride with phosphorus pentasulfide, or of phosphorus pentachloride with antimony trisulfide.

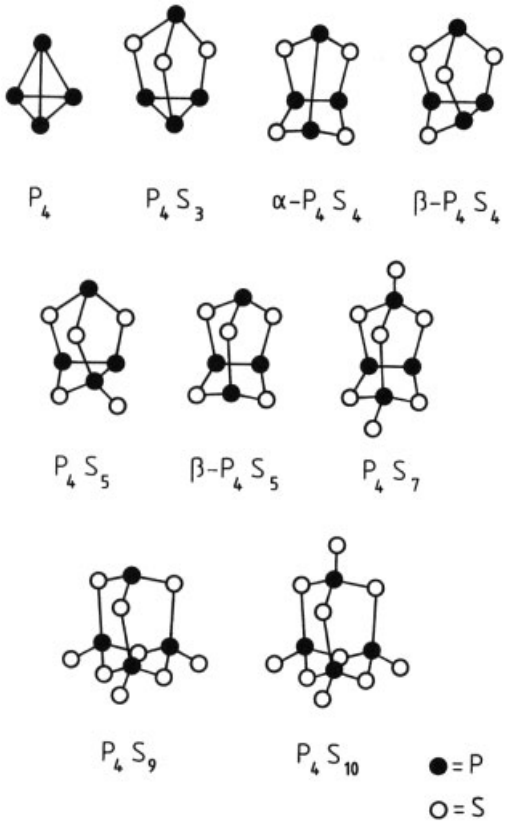

Figure 1. Structure of the phosphorus sulfides

Uses. Thiophosphoryl chloride is used to introduce the P−S group into organic compounds. Thiophosphate esters are prepared by reacting thiophosphoryl chloride with alcohols, sodium alkoxides, phenols, or trialkyl phosphites. Inorganic thiophosphates are produced by the alkaline hydrolysis of thiophosphoryl chloride.

3. Phosphorus Sulfides

The sulfides of phosphorus form a series of compounds with the general formula P_4S_n ($n = 2-10$) [33]. They may be formally derived by the successive introduction of sulfur atoms into the P−P bonds of the P_4 tetrahedron, and by addition of sulfur atoms to the P atoms. Further addition of sulfur to P_4S_{10} leads to incompletely characterized substances containing $P-(S)_n-P$ groups [34], [35]. The structures of some phosphorus sulfides are shown in Figure 1. Thermochemical data and the phase diagram of the phosphorus – sulfur system are given in [36–38]. The reaction of P_4S_{10} with P_4O_{10} gives mixed phosphorus oxide sulfides [39].

Properties. Table 4 lists physical properties of some phosphorus sulfides. They are yellow, crystalline substances with an odor of hydrogen sulfide due to hydrolysis when handled in air. With the exception of P_4S_3, which has only a slight tendency to hydrolyze, all the phosphorus sulfides decompose relatively rapidly in the presence of water, acids, or alkalies. When they are melted or vaporized, decomposition takes place. P_4S_3 is the only phosphorus sulfide that can be melted and evaporated unchanged. All phosphorus sulfides are flammable. The ignition temperature of P_4S_3 in air is only ca. 100 °C; that of P_4S_{10} is ca. 270 °C. There is a danger of dust explosions, particularly when grinding or on filling or discharging the material. Dust – air mixtures containing $> 3\,\%$ oxygen and > 50 g/m^3 P_4S_{10} (particle size < 0.07 mm) are explosive. Containers and items of plant should be filled with an inert gas (nitrogen), and all metallic parts should be earthed.

Table 4. Physical properties of phosphorus sulfides

	P_4S_3	α-P_4S_4	P_4S_5	β-P_4S_5	P_4S_7	P_4S_9	P_4S_{10}
CAS no.	[*1314-85-8*]	[*39350-99-7*]	[*12137-70-1*]	[*20419-10-7*]	[*12037-82-0*]	[*25070-46-6*]	[*1314-80-3*]
mp, °C	172	134 [a]	163 [a]	103 [a]	303	257 [a]	288
bp, °C	406				523		515
Crystal system	orthorhombic	monoclinic	monoclinic	monoclinic	monoclinic	cubic	triclinic
ϱ, g/cm^3	2.03	2.22	2.15	2.26	2.20	2.08	2.09
Solubility [b]	100		10		0.03	0.3	0.22

[a] Decomposition.
[b] Grams per 100 g carbon disulfide.

Preparation. The simplest method of preparing phosphorus sulfides is by heating stoichiometric mixtures of the elements to temperatures above their melting points, followed by cooling. The compounds P_4S_3, P_4S_7, P_4S_9, and P_4S_{10} can be prepared in this way. Either white or red phosphorus can be used. As a rule, pure products are not obtained, and further treatment or purification is necessary. When the reactions are carried out in organic solvents (usually carbon disulfide), phosphorus sulfides can be produced that are not obtainable by the fusion method on account of decomposition reactions. These include, P_4S_5, synthesized by reaction of P_4S_3 with sulfur, and β-P_4S_5, obtained from P_4S_7 by removal of sulfur with triphenylphosphine.

Production. Industrial production of *tetraphosphorus decasulfide*, P_4S_{10} (known in industry as phosphorus pentasulfide or "pentasulfide"), are continuous processes that use white phosphorus and sulfur as starting materials. Molten phosphorus and sulfur are fed into a vessel containing liquid phosphorus pentasulfide at 280–515 °C. Some producers purify the crude liquid product by distillation [40], [41], the heat of evaporation being provided by the heat of reaction. However, the required purity of the final product can also be attained by using specially purified starting materials [42]. Organic impurities are removed from white phosphorus by treatment with concentrated sulfuric acid, and the purified phosphorus is reacted with highpurity sulfur. In this way, distillation of the pentasulfide can be dispensed with, and the liquid reaction product can be fed directly to cooling rollers and then ground.

If the molten product is cooled slowly, fully crystalline phosphorus pentasulfide of low reactivity is obtained, whereas the product obtained by quenching on a cooling roller contains amorphous material and is very reactive. Material of lower reactivity can be prepared from this reactive form by a tempering process [43]. The reactivity of the pentasulfide is an important product characteristic, which is of particular importance for further processing. The reactivity is measured by a simple alcoholysis test, in which the temperature increase due to the heat of reaction is measured when a sample of pentasulfide is reacted with an alcohol (e.g., 2-ethylhexanol). Commercial-quality phosphorus pentasulfide does not consist of a single phase, although the analysis corresponds almost exactly to the stoichiometric compound P_4S_{10}. In addition to the main component P_4S_{10}, a proportion of P_4S_9 is present, and some uncharacterized polymers containing $P-(S)_n-P$ groups [35], [44].

Tetraphosphorus trisulfide [*1314-85-8*], phosphorus sesquisulfide, P_4S_3, is produced by adding phosphorus to liquid sulfur, and distilling the product.

Transportation. Phosphorus pentasulfide is transported in iron drums or aluminum containers. It is a flammable solid with the following transport classifications:

Road∗:	ADR (GGVS) Class 4.1, Rn 2401, Section 8
Rail∗:	RID (GGVE) Class 4.1, Rn 401, Section 8
Sea:	IMDG-Code (GGVSee) Class 4.1, U.N. no. 1340, Packing Group II
Air:	IATA Class 4.3, U.N. no. 1340, Packing Group II

∗ Will presumably be changed by the end of 1992 to Class 4.3, Section 20 b.

Uses. Phosphorus pentasulfide is mostly reacted with alcohols and phenols to form dialkyl

and diaryl dithiophosphate esters. These are important intermediates in the manufacture of insecticides, lubricating oil additives, and flotation agents (→ Phosphorus Compounds, Organic). It is also used to a very small extent in organic chemistry as a sulfurizing reagent.

Tetraphosphorus trisulfide, P_4S_3, is used in the match industry in the manufacture of strike-anywhere matches (→ Matches).

Table 5. Physical properties of phosphides

Phosphide	CAS no.	Crystal system [a]	*mp*, °C	*ρ*, g/cm^3	Color
Na_3P	[*12058-85-4*]	hex		1.74	gray
NaP	[*72490-22-3*]	ort		2.07	black
Na_3P_7	[*82584-48-3*]	ort	942	1.97	yellow
Na_3P_{11}	[*39343-85-6*]	ort		2.14	orange
NaP_{15}	[*87345-78-6*]	mon		2.35	red
Mg_3P_2	[*12057-74-8*]	cub		2.06	yellow-green
Ca_3P_2	[*1305-99-3*]	tet	1600	2.51	red-brown
BP	[*20205-91-8*]	cub		2.91	brown
AlP	[*20859-73-8*]	cub		2.40	green
GaP	[*12063-98-8*]	cub	1477	4.16	orange
InP	[*22398-80-7*]	cub	1070	4.79	gray
SnP_3	[*37367-13-8*]	tri		4.33	light gray
SnP	[*25324-56-5*]	tri		4.98	light gray
Sn_4P_3	[*12286-33-8*]	tri		5.76	light gray
Zn_3P_2	[*1314-84-7*]	tet	>420	4.54	gray
ZnP_2	[*12037-79-5*]	tet		3.51	red
CuP_2	[*12019-11-3*]	mon	900	4.20	black
Cu_2P_7	[*84861-49-4*]	mon		3.56	black
Cu_3P	[*12019-57-7*]	hex	1022	7.41	black
FeP_2	[*12022-85-4*]	ort		5.11	gray
FeP	[*26508-33-8*]	ort		6.07	gray
Fe_2P	[*1310-43-6*]	hex	1365	6.90	gray
Fe_3P	[*12023-53-9*]	tet	1166	6.92	gray
Fe_4P	[*122014-50-0*]	ort		7.89	gray
Ni_3P	[*12059-19-7*]	tet	970	7.9	gray
Ni_5P_2	[*11103-55-2*]	tri	1175	3.80	gray
$Ni_{12}P_5$	[*12035-78-8*]	tet	1150	7.44	gray
Ni_2P	[*12035-64-2*]	hex	1110	7.2	gray
Ni_5P_4	[*12035-76-6*]	hex	850	6.26	gray
NiP	[*12035-46-0*]	ort		5.85	black
NiP_2	[*12035-47-1*]	mon		4.58	gray
NiP_3	[*12192-07-3*]	cub		4.21	gray

[a] cub = cubic, tet = tetragonal, hex = hexagonal, tri = trigonal, ort = orthorhombic, mon = monoclinic.

Economic Aspects. Tetraphosphorus decasulfide is economically important, as is tetraphosphorus trisulfide, though to a much lesser extent. None of the other phosphorus sulfides are of economic importance. In 1989, world capacity for phosphorus pentasulfide was estimated to be a little over 200 000 t, and the actual production ca. 150 000 t. Consumption was divided up as follows: ca. 50 % for lubricating oil additives, ca. 45 % for insecticides, and ca. 5 % for other uses, mainly as flotation agents.

Tetraphosphorus trisulfide is of comparatively minor importance; world production in 1989 was 150 – 200 t.

4. Phosphides

Phosphides are compounds of phosphorus with elements of lower electronegativity. Most metals form compounds with phosphorus. These are mainly binary metal phosphides, but a few are ternary. Some nonmetals and semimetals such as boron or silicon are also known to form phosphides. In total, well over 500 different phosphides have been described, and in many instances have been characterized by determining their crystal structure [45], [46]. Physical data for some phosphides are given in Table 5.

The phosphides may be classified in various ways. The usual classification is based on the stoichiometry, i.e., into metal-rich phosphides (M/P > 1), monophosphides (M/P = 1), and phosphorus-rich phosphides (M/P < 1). Metal-rich phosphides are usually hard, brittle substances with a metallic luster. They possess relatively high thermal and electrical conductivity and are thermally stable and very unreactive. They are stable towards water and dilute acids. Some are decomposed only by molten alkali. The industrial products are often regarded as metal–phosphorus alloys. The monophosphides have similar properties to the metal-rich phosphides. For example, those of the elements of group 3 have marked semiconductor properties. Phosphorus-rich phosphides typically have lower melting points and lower thermal stabilities than the monophosphides or the metal-rich phosphides. With increasing phosphorus content, the phosphorus atoms increasingly cross-link, forming polyphosphide anions. In some of these polyphosphides, structural elements typical of violet (Hittorf's) phosphorus appear. Alkali metals, alkaline earth metals, and lanthanides form phosphides with largely ionic character. A typical reaction of these salt-like compounds is their smooth hydrolysis to give PH_3.

Production. The simplest method of production is direct combination of the elements at elevated temperature. White, red, or gaseous phosphorus can be used. Some phosphides can be prepared by reaction of phosphine with metals or their salts. Metallic phosphides can also be prepared by the electrothermal reduction of metal oxides in the presence of carbon and phosphates. Ferrophosphorus is produced in this way from the iron-containing impurities present in the phosphate rock used in the production of phosphorus (→ Phosphorus).

Uses. Phosphine, produced in large quantities by hydrolyzing metal phosphides, is used in pest control (see Chap. 5). The most important compounds used are aluminum phosphide AlP (Phostoxin, Degesch), and magnesium phosphide Mg_3P_2 (Magtoxin, Degesch). Compared with pure PH_3, these have the advantage of being much safer to use. The products are protected against uncontrollably rapid hydrolysis, such as could occur by accidental contact with water, by coating their surfaces with hydrophobic substances such as paraffin or metal stearates. By mixing the material with ammonium carbamate, which produces ammonia and carbon dioxide on decomposition, the possible spontaneous ignition of the PH_3 produced is prevented [47]. The annual world production of AlP and Mg_3P_2 together amounts to ca. 3000 t. Aluminum phosphide and calcium phosphide (Polytanol) Ca_3P_2 are used as rodenticides, especially for voles. Zinc phosphide Zn_3P_2 is used to prepare poison bait for rodent control.

Gallium phosphide GaP, indium phosphide InP, and the mixed crystals with their respective arsenides, are important semiconductors, e.g., in the manufacture of light-emitting diodes (→ Electroluminescent Materials and Devices).

The phosphides of iron, copper, and tin are used in metallurgical processes. Iron phosphide (ferrophosphorus [*8049-19-2*]) is used in by far the largest quantities. It has an average phosphorus content of 20 – 27 %, so that the principal constituent is diiron phosphide Fe_2P (phosphorus content 21.7 %). Ferrophosphorus is produced as a byproduct in the manufacture of yellow phosphorus. The production rate therefore directly depends on the phosphorus production rate and on the iron content of the phosphate rock. World production is ca. 50 000 t. The main application is in the foundry and steel industries. For certain applications, the phosphorus content of cast iron and steel must be increased by the addition of ferrophosphorus. The phosphorus content (up to 1.5 %) improves the flow properties of cast iron and increases its wear resistance. When manufacturing free-cutting steel, the presence of a small percentage of phosphorus (up to 0.3 %) improves machinability.

The phosphides of copper and tin are mainly used as deoxidants in the production of the corresponding metals (→ Copper Alloys, Chap. 2.1.) and as constituents of phosphor bronzes (→ Copper Alloys, Chap. 6.1.) and hard solders (brazing materials). The usual commercial products are copper – phosphorus alloys with 10 or 15 % P (principle constituent Cu_3P) and tin –phosphorus alloys containing 5 % P. Microcrystalline nickel phosphide Ni_3P is present in the nickel layer produced by the electrodeless deposition of nickel (→ Electrochemical and Chemical Deposition).

5. Phosphorus – Hydrogen Compounds

Phosphorus and hydrogen form a large number of compounds, which have been characterized mainly by spectroscopic methods [48]. The known hydrides of phosphorus are exceeded in number only by those of carbon. However, the only phosphorus hydride that is of economic importance is phosphine [49].

Phosphine [*7803-51-2*], phosphane, PH_3 is a colorless, very poisonous gas (*mp* −134 °C; *bp* − 88 °C; vapor pressure 3.6 MPa at 20 °C). Above 375 °C, it decomposes into the elements. Phosphine – air mixtures are explosive over a wide range. The lower explosion limit is 1.6 %, and the upper limit is close to 100 % [50]. Pure phosphine should presumably therefore be regarded as spontaneously flammable like phosphine that contains P_2H_4 as an impurity. Although phosphine has a garliclike odor, the threshold of detection can exceed several hundred ppm if the gas is very pure. The characteristic odor of acetylene is due to phosphine, produced when the acetylene is generated from calcium carbide that contains calcium phosphide

as an impurity. With strong acids, phosphine forms phosphonium salts; e.g., with hydrogen iodide, phosphonium iodide [*12125-09-6*], PH_4I, is formed, from which pure phosphine may be liberated by adding water or solutions of alkalimetal hydroxides.

Production. Phosphine can be synthesized by a number of different reactions [49]. On the laboratory scale, it may be prepared simply by the hydrolysis of metal phosphides or phosphonium salts. The two most common industrial methods are based on disproportionation reactions of elemental phosphorus catalyzed by alkali or acids. A well-known example is the reaction of white phosphorus with sodium hydroxide solution, which is used to produce both phosphine and sodium hypophosphite [*7681-53-0*], NaH_2PO_2 (sodium phosphinite).

$$P_4 + 3\,NaOH + 3\,H_2O \rightarrow PH_3 + 3\,NaH_2PO_2$$

In sodium hypophosphite production plants, the phosphine byproduct is usually burnt off. In the presence of alcohols, the reaction may be controlled such that the product formed in addition to phosphine is almost entirely sodium phosphite Na_2HPO_3 [51]. At the same time, the maximum yield based on the amount of phosphorus used increases from 25 % to 50 %. In acid-catalyzed disproportionation, white phosphorus is converted into phosphine and phosphoric acid at ca. 280 °C, red phosphorus being formed as an intermediate [52].

$$2\,P_4 + 12\,H_2O \rightarrow 5\,PH_3 + 3\,H_3PO_4$$

This reaction gives the highest yield of PH_3 based on the amount of phosphorus used (theoretical yield 62.5 %). Due to the high reaction temperature and the corrosiveness of the medium, the reactor must be made of graphite.

Uses. World production of phosphine is ca. 1500 t/a. The phosphine is mostly used captively as a starting material in the production of organophosphorus compounds, mainly by addition of aldehydes or alkenes to the P–H bonds (→ Phosphorus Compounds, Organic). Phosphine is also used for pest control (see Chap. 4) where foodstuffs and animal feeds are stored. Almost all dry products of this type are suitable for gassing with phosphine without any harmful effects on such properties as smell, taste, and appearance. The main area of use is in gassing grain silos. In the semiconductor industry, specially purified phosphine (arsine-free in particular) is used as a dopant.

Polyphosphanes. The simplest polyphosphane is diphosphane [*13445-50-6*], P_2H_4, a spontaneously flammable liquid with an unpleasant smell (*bp* 63 °C). Diphosphane is formed together with phosphine in the hydrolysis of calcium phosphide Ca_3P_2. Diphosphane belongs to the straight-chain phosphanes. It is part of a homologous series with the general formula P_nH_{n+2}, in which the higher members have very low thermal stability. Thermal decomposition of diphosphane yields a large number of different polyphosphanes, some of which have been characterized by means of mass and NMR spectroscopy [48]. In addition to the straight-chain structures, monocyclic and, to a greater extent, polycyclic structures exist. The individual compounds can be allocated to various homologous series with the general fomulae P_nH_n (cyclic compounds), P_nH_{n-2}, and P_nH_{n-4} to P_nH_{n-18} (polycyclic compounds). Polyphosphanes that have been prepared in the pure state are diphosphane; triphosphane(5) [*13597-70-1*], P_3H_5; pentaphosphane(5) [*6798-45-4*], P_5H_5; and heptaphosphane(3) [*51273-53-1*], P_7H_3.

6. Phosphorus – Nitrogen Compounds

Although a large number of inorganic phosphorus – nitrogen compounds exist, hardly any of these are of industrial importance, with the exception of the phosphazenes. Many compounds that contain a P–N bond can be regarded as formal derivatives of the oxoacids of phosphorus, in which P–OH or P=O groups are replaced by the isoelectronic $P-NH_2$ or P=NH groups. The phosphorus nitrides with various P–N ratios described in the literature are mostly amorphous substances. Only triphosphorus pentanitride [*12136-91-3*], P_3N_5, is known in the crystalline state. It is formed by the thermal decomposition of the product of the reaction between PCl_5 and NH_3. Phospham [*33849-97-7*], phosphorus nitride imide $[PN(NH)]_x$ is the intermediate product, which is converted at 500 °C to P_3N_5, a white, insoluble powder that can be

used for flameproofing cellulose fibers [53] or wood surfaces [54]. The halogenophosphazenes (phosphonitrilic halides), especially the chlorine compounds, are of greater importance [55]. Phosphazene is a general term for unsaturated phosphorus – nitrogen compounds with alternating P and N atoms. The halogenophosphazenes form a large group of oligomeric and polymeric compounds with both cyclic and linear structures, with the general formula $(PNX_2)_n$ ($n \geq 3$, X = halogen).

The chlorocyclophosphazenes are white, crystalline solids, while the short-chain linèar compounds are oily liquids. The simplest compound, which has been known for many years, is hexachlorocyclotriphosphazene [*940-71-6*], $(PNCl_2)_3$ (*mp* 114 °C, *bp* 256 °C, ϱ 1.99 g/cm^3), which is produced on an industrial scale as the starting material for the manufacture of organosubstituted polyphosphazenes. It forms an almost planar six-membered ring with equal P–N bond lengths. The monomeric $PNCl_2$ (phosphorus dichloride nitride) is unknown.

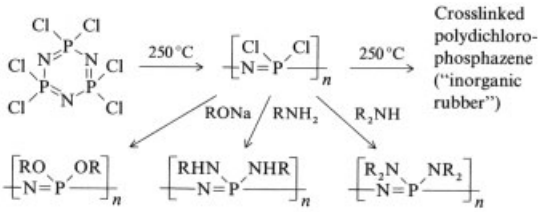

Figure 2. Synthesis of polyorganophosphazenes

The simple chlorocyclophosphazenes polymerize on heating to 250 – 350 °C to form rubber-like products ("inorganic rubber"). These are cross-linked polymers which are insoluble in organic solvents and have no direct application on account of their marked tendency to hydrolyze. However, hydrolysis can be prevented if the halogen atoms are completely replaced by organic groups. Complete substitution can only be achieved if the only polymer present is linear poly(dichlorophosphazene), [*25034-79-1*] $(PNCl_2)_n$ ($n \leq 15\,000$) [56]. This substance, not being cross-linked, can be dissolved in organic solvents such as toluene, and then converted into chlorine-free products by reaction with alkoxides, phenoxides, or amines (Fig. 2). The polyorganophosphazenes produced in this way are hydrolytically stable elastomers with a wide range of possible applications [55], [57–59], (→ Inorganic Polymers, Chap. 4.).

Production. The starting material for the production of chlorophosphazenes is phosphorus pentachloride, which is reacted with finely divided ammonium chloride in an inert organic solvent such as chlorobenzene or 1,1,2,2-tetrachloroethane at elevated temperature. Catalysts are usually added, e.g., metal salts or quinoline derivatives [60].

$$n\,PCl_5 + n\,NH_4Cl \rightarrow (PNCl_2)_n + 4\,n\,HCl$$

In addition to the trimer, the crude reaction product contains higher cyclic oligomers and a number of linear phosphazenes. The isolation of the desired trimer is usually carried out by washing with water and vacuum distillation [61]. The trimer can be polymerized either in the molten state or in solution. Excess PCl_5 or traces of water and other impurities can impede polymerization and lead to cross-linking reactions. By adding catalysts (mainly Lewis acids such as BCl_3), the temperature required for polymerization can be reduced from 250 – 300 °C to 200 – 220 °C, and the yield increased [62], [63]. These catalysts are assumed to remove a chlorine atom from the phosphazene ring, forming an ionic intermediate compound and thereby initiating polymerization.

Although the cost of the starting materials PCl_5 and NH_4Cl is low, the production cost of pure trimeric chlorophosphazene is relatively high due to the high cost of purification. This means that industrial applications are limited to areas in which the outstanding material properties of the end products are of overriding importance (→ Inorganic Polymers, Chap. 4.).

7. Toxicology

Phosphorus Oxides. In contrast to the oxides of arsenic, the phosphorus oxides P_4O_6 and P_4O_{10} do not possess any specific toxicity [64]. Phosphorus pentoxide is irritating to corrosive to the skin and the mucous membranes, respectively [65]. The MAK value for P_4O_{10} is 1 mg/m^3 [66].

Phosphorus Halides. The phosphorus halides possess a local irritative or corrosive effect, since they form hydrogen halides with water. A detailed description of the toxicological data on phosphorus trichloride and phosphoryl chloride is given in [67].

On contact with moisture, *phosphorus trichloride* is hydrolyzed in an exothermic reaction forming phosphonic acid and hydrogen chloride. Therefore, it is strongly irritating to the skin, eyes, and respiratory tract. Severe corrosion of the skin and mucous membranes may occur upon contact with liquid PCl_3. Irritation is five to six times stronger than would be expected from the hydrochloric acid formed. Due to the high solubility in water, the irritation is highest in the upper respiratory tract. Therefore, the main symptoms consist of irritation of the eyes, throat, and trachea, lacrimation, and coughing. Upon inhalation of higher concentrations, injury of the deeper regions of the lung, and formation of a toxic pulmonary edema can occur. Upon chronic exposure to lower concentrations, cough, dyspnea, and bronchitis have been described. In humans, systemic toxic effects have not been observed till now. In rats, the LC_{50} for a 4-h exposure is 590 mg/m^3 [67]. The MAK value is 3 mg/m^3 [66]; the TLV-TWA value is 1.1 mg/m^3 [68].

Phosphoryl chloride hydrolyzes slowly but exothermically to give phosphoric acid and hydrogen chloride. Therefore, phosphoryl chloride has primarily an irritant action, whereby deeper penetration into the respiratory tract is possible due to the delayed hydrolysis. Often, no complaints are felt in the beginning; only after a latent period of several hours does damage to the deeper respiratory tract occur. This is often of an irreversible nature. For this reason, $POCl_3$ is regarded toxicologically as the most dangerous of the phosphorus chlorides. Characteristic of the inhalative poisoning of humans is the serious state of the lungs, which is in contrast to the relatively good general condition of the affected person. The after-effects of poisoning in those who survive consist frequently of obstructive respiratory tract diseases. Skin contact with liquid phosphoryl chloride can lead to acid burns. The 4-h LC_{50} value in rats is 310 mg/m^3 [67]. The MAK value is 1 mg/m^3 [66], and the TLV-TWA value, 0.63 mg/m^3 [68].

For *phosphorus pentachloride*, whose vapors are similarly strongly irritating to the mucous membranes and respiratory tract [65], the MAK value is 1 mg/m^3 [66] and the TLV-TWA value is 0.85 mg/m^3 [68].

Phosphorus Sulfides. *Tetraphosphorus trisulfide* is not sufficiently volatile to cause an inhalation risk at normal temperatures. In the form of dust or vapor, it irritates the skin, eyes, and respiratory tract. The toxicity is regarded as slight in comparison to white phosphorus [65]. Upon contact with the skin, tetraphosphorus trisulfide can cause eczema [69–72], whereby serious, edematous eczemas of the eyelids and cheeks have been described in some cases. The cutaneous tests were positive to P_4S_3. The oral lethal dose in rabbits is reported to be 100 mg/kg of body weight [73].

Phosphorus pentasulfide, P_4S_{10}, decomposes in the presence of water with the formation of phosphoric acid. Therefore, it acts as an irritant [65]. At the same time, poisonous hydrogen sulfide is formed. In rats, the oral LD_{50} for P_4S_{10} is reported to be 389 mg/kg of body weight [74]. The MAK and TLV-TWA values are both 1 mg/m^3 [66], [68].

Phosphides are decomposed by water or acids with the formation of phosphine, which is the actual cause of phosphide poisoning.

Aluminum, magnesium and zinc phosphides, which are used as rodenticides, react in the gastrointestinal tract with water and hydrochloric acid to give phosphine. In rats, the oral LD_{50} value is 8.7 (AlP), 20.7 (Mg_3P_2, male animals), 10.4 (Mg_3P_2, female animals) and 45 mg/kg (Zn_3P_2) of body weight [75]. In humans, even quantities of 5 g zinc phosphide have caused death, although survival has been reported for doses of 50 g [76]. The relative insensitivity of dogs and cats towards zinc phosphide is attributable to the emetic effect of zinc, which prevents the absorption of larger quantities [77]. In humans, the symptoms of poisoning include exhaustion, nausea, tinnitus aurium, chest complaints, and anxiety. Elementary help, gastric lavage with 0.1 % $KMnO_4$ solution, and symptomatic treatment are recommended as therapy following oral intake [75]. In aluminum phosphide poisoning, treatment consisting of extracorporal hemodialysis as well as sup-

portive cardiovascular-active therapy after gastric lavage with $KMnO_4$ and $MgSO_4$ have been found to be useful [78]. In a long-term study on rats, feeds treated with aluminum phosphide (nominal maximal concentration 30 mg PH_3/kg) caused no toxic or carcinogenic effects [79].

Phosphine. The most important source of poisoning is technical, impure acetylene, which is formed by the action of water on calcium carbide that contains calcium phosphide as an impurity. Phosphine is also formed when ferrosilicon comes into contact with moisture. Phosphine is used as a rodenticide. It is formed from aluminum, magnesium, or zinc phosphide in the presence of moisture and can cause fatal accidents. Quantitatively, poisoning with phosphine does not play any major role today [64], [80].

Phosphine is a highly poisonous gas which acts as a poison in cell metabolism and which can lead to death through internal suffocation. Intake into the organism can occur following inhalation or, after oral intake of phosphides, through the gastrointestinal tract. Phosphine is not absorbed through the skin [67]. Numerous cases of poisoning, sometimes fatal, have been described in humans. The pattern of poisoning resembles poisoning with arsine; however, there are no hemolytic effects. In lower concentrations, symptoms of poisoning may occur after a latent period of several hours. Major symptoms are headache, giddiness, nausea, vomiting, stomach pain, and perhaps diarrhea. After the victim has become unconscious, death usually occurs within 24 – 48 h not only due to pulmonary edema and respiratory paralysis, but also due to cerebral edema and cardiovascular failure. In higher concentrations, sudden death due to pulmonary edema and respiratory paralysis are possible. If the victim survives the poisoning, then damage to the liver and kidneys can appear after a latent period of 24 – 48 h [67], [80].

For humans, concentrations of 400–840 mg/m^3 over a period of 30 – 60 min are extremely dangerous or even lethal, while concentrations of 140 – 260 mg/m^3 are tolerated under these conditions. However, one death has been observed following exposure to 10 mg/m^3 for 6 h [81]. In cats and rats, deaths occur already after acute exposure to 35 mg/m^3. In subacute exposures, death occurs already from concentrations of 7 mg/m^3 onwards [82].

Apparently, a continuous detoxification process occurs in the organism following longer exposure to low phosphine concentrations ($\leq$ 3.5 mg/m^3), so that no chronic cumulative poisoning takes place [82].

Therapy: Symptomatically, treatment of toxic pulmonary edema. A specific therapy for phosphine poisoning is not known [75], [80].

The MAK value is 0.15 mg/m^3 [66]; the TLV-TWA value is 0.42 mg/m^3 [68].

Higher Phosphanes. Diphosphane, a highly poisonous, spontaneously flammable liquid, acts in a similar way to phosphine [80].

8. References

1. Hoechst, DE – OS 2 736 765, 1979 (G. Hartlapp, W. Kowalski, R. Queck, T. Dahmen).
2. Hoechst, DE 1 300 527, 1967 (H. Harnish, F. Krahl, F. Thomas, H. Panter).
3. Stauffer Chem. Comp., DE 1 147 924, 1961 (G. I. Klein, L. B. Post, R. E. Newby).
4. Hoechst, US 4 777 025, 1988 (W. Kowalski, T. Dahmen, H. Haas, G. Schimmel).
5. Hoechst, US 4 705 529, 1987 (G. Schimmel, W. Kowalski, G. Heymer, R. Gradl).
6. W. H. Shearon, A. J. Hoiberg, *Ind. Eng. Chem.* **45** (1953) 2122 ff.
7. Hoechst, DE 1 767 205, 1968 (G. Heymer, W. Gerhardt, H. Harnish); DE 2 330 174, 1973 (H. Schrödter, H. W. Stephan, G. Heymer).
8. Hoechst, DE 2 330 174, 1973 (H. Schrödter, H. W. Stephan, G. Heymer).
9. D. Heinz, *Pure Appl. Chem.* **44** (1975) 141.
10. L. Wolf, H. Schmanger, *Ber. Dtsch. Chem. Ges.* **62** (1929) 771 – 786.
11. L. Wolf, W. Jung, M. Tschudnowsky, *Ber. Dtsch. Chem. Ges.* **65** (1932) 488 – 491.
12. VEB Kombinat Agrochemie Piesteritz, DE 3 319 606, 1983 (D. Heinz et al.).
13. D. Heinz, R. G. Asijew, I. F. Müller, G. Kauschka, *Pure Appl. Chem.* **52** (1980) 825.
14. D. Heinz, *Z. Anorg. Allg. Chem.* **336** (1965) 137 – 155.
15. P. W. Schenk, H. Rehag, *Z. Anorg. Allg. Chem.* **233** (1937) 403 ff.
16. H. E. Stokinger et al., *J. Occup. Med.* **5** (1963) 491 ff.
17. Hoechst, DE 1 206 406, 1964 (J. Cremer, U. Thümmler, F. Schulte, H. Harnisch); DE 2 200 653, 1977 (J. Cremer).

18. Monsanto, US 2 670 274, 1951 (O. C. Jones).
19. G. Heymer, H. D. Wasel-Nielen, *Chem. Ztg.* **96** (1972) 475 – 487.
20. H. Weichardt, *Chem Ztg.* **81** (1957) 421 – 423.
21. T. E. Daubert, R. P. Danner: "Data Compilation –Tables of Properties of Pure Compounds," AIChE/DIPPR, 1987.
22. The NBS Tables of Chemical Thermodynamic Properties, *J. Phys. Chem. Ref. Data* **11** (1982) Suppl. no. 2.
23. *Gmelin,* System no. 16, part C, 458 – 480.
24. Hoechst, DE 1 104 931, 1959.
25. Bayer, DE 1 767 949, 1968.
26. A. R. Galbraith, *J. Am. Oil Chem. Soc.* **41** (1964) no. 2, 104 – 107.
27. Du Pont, US 2 709 186, 1954.
28. Du Pont, DE 1 044 777, 1957.
29. Du Pont, US 2 810 629, 1955.
30. Bayer, DE 834 243, 1948.
31. VEB Farbenfabrik Wolfen, DD 10 041, 1953.
32. F. F. Knotz, US 2 915 361, 1954.
33. H. Hoffmann, M. Becke-Goehring, *Top. Phosphorus Chem.* **8** (1976) 193 – 271.
34. G.-U. Wolf, M. Meisel, J. Neels, DD 233 365, 1984.
35. M. C. Demarcq, *Phosphorus Sulfur* **33** (1987) 127 –134.
36. R. Blachnik, A. Hoppe, *Z. Anorg. Allg. Chem.* **457** (1979) 91 – 104.
37. H. Vincent, *Bull. Soc. Chim. Fr.* 1972, 4517 – 4521.
38. R. Förthmann, A. Schneider, *Z. Phys. Chem. NF* **49** (1966) 22 – 37.
39. G.-U. Wolf, M. Meisel, *Z. Anorg. Allg. Chem.* **509** (1984) 101 – 110.
40. Saffa, US 4 374 706, 1981 (P. Molla).
41. Monsanto, US 2 794 705, 1954 (R. B. Hudson).
42. Knapsack-Griesheim, DE 1 141 625, 1961 (J. Cremer, F. Rodis).
43. Hooker, US 3 146 069, 1959 (S. Robota).
44. M. C. Demarcq, *Phosphorus Sulfur* **11** (1981) 65 – 69.
45. H. G. von Schnering, W. Hoenle, *Chem. Rev.* **88** (1988) 243 – 273.
46. *Landolt-Börnstein,* **New Series, III/7 c,** 1 – 71.
47. W. Freyberg, W. Friemel, DE 1 143 053, 1961.
48. M. Baudler, *Angew. Chem.* **99** (1987) 429 – 451; *Angew. Chem. Int. Ed. Engl.* **26** (1987) 419.
49. E. Fluck, *Top. Curr. Chem.* **35** (1973) 1 – 64.
50. H. Ohtani et al., *Combust. Flame* **76** (1989) 307 – 310.
51. Hoechst, DE 2 840 147, 1978 (G. Elsner, W. Klose, C. May, G. Heymer).
52. Albright & Wilson, GB 990 918, 1962 (E. J. Lowe, F. A. Ridgway).
53. Hoechst, EP 0 015 546, 1980 (J. Cremer, H. Harnisch).
54. Dow Chemical, US 4 741 971, 1988 (H. N. Beck, D. C. MacWilliams).
55. H. R. Allcock: *Phosphorus Nitrogen Compounds,* Academic Press, New York 1972.
56. H. R. Allcock, R. L. Kugel, *J. Am. Chem. Soc.* **87** (1965) 4216 – 4217.
57. H. R. Allcock, *Angew. Chem.* **89** (1977) 153 – 162; *Angew. Chem. Int. Ed. Engl.* **16** (1977) 147.
58. H. R. Allcock, *ACS Symp. Ser.* **360** (1988) 250 – 267.
59. J.-F. Labarre, *Top. Curr. Chem.* **102** (1982) 1 – 87.
60. Shin Nisso Kako, DE 3 626 820, 1986 (T. Okamoto, S. Ueyama, K. Fujikawa).
61. Firestone, EP 0 004 878, 1979 (J. W. Fieldhouse).
62. Firestone, US 4 226 840, 1979 (J. W. Fieldhouse, D. F. Graves).
63. Firestone, US 4 327 064, 1980 (J. W. Fieldhouse, S. L. Fenske).
64. W. Wirth, C. Gloxhuber: *Toxikologie,* Thieme Verlag, Stuttgart 1985.
65. *Patty,* 3rd ed., **2 A,** 2121.
66. Deutsche Forschungsgemeinschaft: *Maximale Arbeitsplatzkonzentrationen und biologische Arbeitsstofftoleranzwerte 1989,* VCH Verlagsgesellschaft, Weinheim 1989.
67. D. Henschler: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* VCH Verlagsgesellschaft, Weinheim 1989.
68. American Conference of Governmental Industrial Hygienists: *TLV's – Threshold Limit Values and Biological Indices for 1989 – 1990,* American Conference of Governmental Industrial Hygienists, Cincinatti, Ohio, 1989.
69. W. Nikolowski, *Dtsch. Med. Wochenschr.* **74** (1949) 637 – 641.
70. J. F. Burgess, *Can. Med. Assoc. J.* **65** (1951) 567 – 568.
71. I. R. White, R. J. G. Rycroft, *Contact Dermatitis* **9** (1983) no. 2, 162.
72. S. M. Burge, S. M. Powell, *Contact Dermatitis* **9** (1983) no. 5, 424.
73. C. G. Santesson, R. Malmgren, *Skand. Arch. Physiol.* **15** (1904) 259 – 327.

74. *Registry of Toxic Effects of Chemical Substances,* U.S. Department of Health, Education and Welfare, Public Health Service, Center of Disease Control NIOSH, Cincinnati 1989.
75. Industrieverband Pflanzenschutz e.V.: *Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmitteln, Physikalisch-chemische und toxikologische Daten,* Pressehaus Bintz-Verlag, Offenbach 1982.
76. J. B. P. Stephenson, *Arch. Environ. Health* **15** (1967) 83 – 88.
77. L. J. Casarret: *Casarret and Doull's Toxicology: The Basic Science of Poisons,* Macmillan Publishing, New York 1980.
78. K. E. Zipf, T. Arndt, R. Heintz, *Arch. Toxicol.* **22** (1967) 209 – 222.
79. U. Hackenberg, *Toxicol. Appl. Pharmacol.* **23** (1972) 147 – 158.
80. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* Thieme Verlag, Stuttgart 1986.
81. F. Flury, F. Zernik: *Schädliche Gase,* Springer Verlag, Berlin 1931.
82. O. R. Klimmer, *Arch. Toxicol.* **24** (1969) 164 – 187.

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Phosphorus Compounds, Organic

Standard Article

Jürgen Svara[1], Norbert Weferling[2], Thomas Hofmann[3]
[1]Aventis Pharma Deutschland GmbH, Drug Evaluation, Hürth-Knapsack, Germany
[2]Hoechst AG, Werk Knapsack, Hürth-Knapsack, Germany
[3]Sanofi-Aventis Deutschland GmbH, Hattersheim, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a19_545.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Phosphines
2.1. Properties
2.2. Production
2.2.1. Primary and Secondary Phosphines
2.2.2. Tertiary Phosphines
2.3. Uses
3. Halophosphines
3.1. Properties
3.2. Production
3.3. Uses
4. Phosphonium Salts
4.1. Properties
4.2. Production
4.3. Uses
5. Phosphine Oxides and Sulfides
5.1. Properties
5.2. Production
5.3. Uses
6. Phosphonous Acid Derivatives
6.1. Production
6.2. Properties and Uses
7. Phosphinic Acids and their Derivatives
7.1. Properties
7.2. Production
7.3. Phosphinic Acids and Phosphinate Esters
7.4. Thiophosphinates
7.5. Other Phosphinic Acid Derivatives
8. Phosphites and Hydrogenphosphonates
8.1. Properties
8.2. Production
8.3. Trialkyl Phosphites
8.4. Triaryl Phosphites and Alkyl Aryl Phosphites
8.5. Dialkyl and Diaryl Phosphonates
8.6. Alkyl Phosphonates
9. Phosphonic Acids and their Derivatives
9.1. Properties
9.2. Production
9.3. Phosphonic Acids and Phosphonocarboxylic Acids
9.4. Esters of Phosphonic Acid
9.5. Other Derivatives of Phosphonic Acid
10. Esters of Phosphoric Acid
10.1. Properties
10.2. Production
10.3. Trialkyl Phosphates
10.4. Triaryl and Alkyl Aryl Phosphates
10.5. Mono- and Dialkyl Phosphates
11. Esters of Thiophosphoric Acid
11.1. Properties
11.2. Production
11.3. Monothiophosphates
11.4. Dithiophosphates
12. Economic Aspects
13. Toxicology

[Top of Page]

1. Introduction

Jürgen Svara and Norbert Weferling

For the purpose of this article, organophosphorus compounds include not only compounds containing a phosphorus – carbon bond, but all compounds of phosphorus with an organic group in the molecule.

Organophosphorus compounds are characterized by an unusually large variety of structures and a wide range of uses. The reason is that phosphorus is able to form stable compounds with a coordination number of 1 to 6, and an oxidation state of – III to + V. Of the wide range of possible compounds, only those of industrial importance are treated here.

The organic chemistry of phosphorus has been described in many reviews. In particular, the chemistry of phosphorus with a low coordination number has been intensively investigated in

recent years [1-9].

[Top of Page]

2. Phosphines [10]

Jürgen Svara and Norbert Weferling

Phosphines are compounds in which phosphorus has a coordination number of 3 and an oxidation state of – 3. They are derived from phosphine [7803-51-2], PH_3. Successive formal substitution of 1, 2, or 3 hydrogen atoms of PH_3 by alkyl or aryl groups gives primary, secondary, and tertiary phosphines.

2.1. Properties

Under normal conditions, PH_3, methylphosphine CH_3PH_2, trifluoromethylphosphine [420-52-0], CF_3PH_2, and bis(trifluoromethyl)phosphine [460-96-8], $(CF_3)_2PH$, are gases. Most of the known phosphines are colorless liquids at 0.1 MPa and 20 °C, usually immiscible with water, and have a density of 0.8 – 0.85 g/cm^3. *tert*-Alkylphosphines with a chain length > C_{12} tend to crystallize at room temperature. Phosphines that contain a rigid organic group, e.g., 9*H*-9-phosphabicyclononane [13887-02-0], [13396-80-0] are crystalline, as are compounds in which the van der Waals forces resulting from the functional groups leads to formation of crystalline phases [e.g., tris(2-cyanoethyl)phosphine].

Phosphines have an intense and penetrating odor, often described as resembling garlic, even in the ppb range. For phosphine itself, the odor threshold is extremely dependent on the purity of the gas. High-purity PH_3 (5 N), as used in the semiconductor industry, is detectable by smell only when the concentration is at a dangerously toxic level (> 1000 ppm), but when PH_3 is produced by the hydrolysis of AlP or Mg_3P_2 (see also ⟶ Phosphorus Compounds, Inorganic), it can be detected even at ca. 20 ppb. The MAK value for PH_3 is 0.1 ppm and the TLV value, 0.3 ppm. There are no MAK or TLV limits for any other phosphines, with the exception of phenylphosphine (see Chap. Toxicology).

Phosphine, PH_3, like some other low molecular mass phosphines (e.g., phenylphosphine, methylphosphine, dimethylphosphine, and trimethylphosphine), is very toxic. The LC_{50} and LD_{50} values show that the acute toxicity decreases with increasing molecular mass.

The presence of an electron lone pair gives rise to the three principal properties of the phosphines:

1. Oxidizability
2. Ability to function as a ligand in complexes
3. Nucleophilicity

The physical properties of some primary, secondary, and tertiary phosphines are given in Tables 1-3.

Table 1. Physical properties of primary phosphines

Formula	CAS no.	M_r	Density, g/cm^3	^{31}P NMR, ppm	*bp* at 103 kPa, °C	n_D^{20}	Flash point, °C
CH_3PH_2	[593-54-4]	48.02		– 160.2	– 14		[a]
$C_2H_5PH_2$	[593-68-0]	62.05		– 128	25		[a]
n-$C_4H_9PH_2$	[1732-74-7]	90.10	0.769	– 140	86.7	1.4477	[a]
c-$C_6H_{11}PH_2$	[822-68-4]	116.13	0.875	– 110	145	1.4860	[a]
$C_6H_5PH_2$	[638-21-1]	110.09	1.001	– 118.7	160	1.5736	[a]

[a] Spontaneously flammable.

Table 2. Physical properties of secondary phosphines

Formula	CAS no.	M_r	Density, g/cm^3	^{31}P NMR, ppm	*bp* (*p*), °C (Pa)	*mp*, °C	n_D^{20}	Flash point, °C
$(CH_3)_2PH$	[676-59-5]	62.05		−99.5	21.1 (1.03×10^5)	< −64	1.4960	[a]
$(C_2H_5)_2PH$	[627-49-6]	90.10	0.7862	−55.5	85 (1.03×10^5)	−51	1.4470	[a]
$(n\text{-}C_4H_9)_2PH$	[1732-72-5]	146.20	0.808	−69.5	183 (1.03×10^5)		1.456	[a]
$(c\text{-}C_6H_{11})_2PH$	[829-84-5]	198.27	0.98	−28.3	281 (1.03×10^5)		1.514	[a]
$(C_6H_5)_2PH$	[829-85-6]	186.17	1.07	−41	116 (346)		1.6270	[a]
$C_8H_{14}PH$ [b]	asym. [13396-80-0]	142.17	1.06[c]	−48.9	220 (1.03×10^5)[c]	50[c]		50[c]
	sym. [13887-02-0]			−55.1				

[a] Spontaneously flammable. [b] asym. ; sym. . [c] Mixture of isomers.

Table 3. Physical properties of tertiary phosphines

Formula	CAS no.	M_r	Density, g/cm^3	^{31}P NMR, ppm	*bp* at 103 kPa, °C	*mp*, °C	n_D^{20}	Flash point, °C
$(CH_3)_3P$	[594-09-2]	76.07	0.748	– 62	40	– 85		[a]
$(C_2H_5)_3P$	[554-70-1]	118.15	0.801	– 20.4	127		1.456	[a]
$(n\text{-}C_4H_9)_3P$	[998-40-3]	202.31	0.817	– 32.2	240		1.4635	37.2
$(c\text{-}C_6H_{11})_3P$	[2622-14-2]	280.41		+ 7		76		
$(C_6H_5)_3P$	[603-35-0]	262.27	1.2	– 5.4	360	80		182
$(NCCH_2CH_2)_3P$	[4023-53-4]	193.19		– 23		98 – 99		278

[a] Spontaneously flammable.

2.2. Production

2.2.1. *Primary and Secondary Phosphines*

The radical-induced addition of alkenes to PH_3 produces primary and secondary phosphines with good selectivity, provided that linear α-alkenes are not used. Commercially, 9*H*-9-phosphabicyclononane (American Cyanamid, Hoechst) and diisobutylphosphine [1732-72-5] (American Cyanamid) are produced by this method [11]. Azobisisobutyronitrile (AIBN) is used as the initiator.

Primary phosphines are produced selectively, though only in moderate yield, by the acid-catalyzed reaction of PH_3 with alkenes [12]. In contrast to the radical-catalyzed reaction, this reaction yields Markovnikov products. *tert*-Butylphosphine can be obtained from PH_3 and isobutene in this way. This substance has been proposed as a substitute for high-purity PH_3 for use in the semiconductor industry (American Cyanamid) [13].

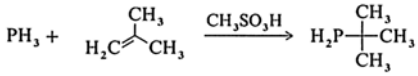

2.2.2. *Tertiary Phosphines*

Reaction of PCl_3 with Organometallic Compounds. The Grignard reaction with phosphorus trichloride offers an industrial route to tertiary phosphines (M & T or Atochem, Hokko) [14], [15].

$$PCl_3 + 3\,RMgCl \xrightarrow{-MgCl_2} PR_3$$

R = phenyl, cyclohexyl, butyl, octyl, tolyl

Triphenylphosphine, which is important as a catalyst ligand and as an intermediate in the synthesis of Wittig reagents, is also produced by BASF in a reaction analogous to the Wurtz reaction [16].

$$PCl_3 + 3\,C_6H_5Cl + 6\,Na \longrightarrow P(C_6H_5)_3 + 6\,NaCl$$

Addition of Unsaturated Compounds to PH_3. Trialkylphosphines are obtained by the radical-induced addition of alkenes to PH_3 (American Cyanamid, Hoechst, Nippon Chemical Industries) [17-19].

$$PH_3 + 3\,H_2C{=}CHR \xrightarrow{I} P(CH_2CH_2R)_3$$

R = H, C_nH_{2n+1} (n = 1–18)

I = Initiator (e.g., azobisisobutyronitrile)

Both the Grignard process and the PH_3 route have advantages and disadvantages. For example, trimethylphosphine and triarylphosphines cannot be obtained by the PH_3 process. Only linear α-alkenes react smoothly to give tertiary phosphines, and phosphines with sterically hindered groups are generally not obtainable in this way. On the other hand, the Grignard process has a potential pollution problem due to the byproduct magnesium salts, meaning that extra cost is incurred.

Tris(2-cyanoethyl)phosphine, which is of interest as an intermediate, can be obtained either directly by radical- or base-catalyzed reaction of PH_3 with acrylonitrile [20], or by reacting tris (hydroxymethyl)phosphine [2767-80-8] with acrylonitrile [21]. Tris(hydroxymethyl)phosphine can be obtained by reacting PH_3 with formaldehyde in the presence of metal catalysts (e.g., Cd^{2+}), but is more conveniently produced from the commercially more readily available tetrakis(hydroxymethyl)phosphonium chloride or sulfate (see Chap. Phosphonium Salts) by reaction with aqueous NaOH with loss of one equivalent of formaldehyde.

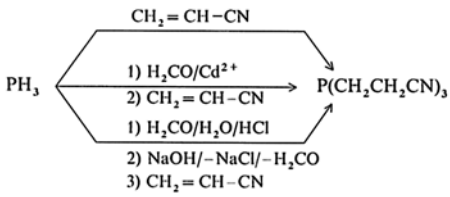

The industrial production of tertiary phosphines with differing groups on the phosphorus atom is carried out by a number of methods.

Primary or secondary phosphines can be further alkylated with alkenes (sometimes bearing functional groups) under free-radical catalysis.

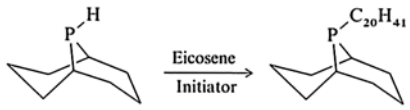

(1)

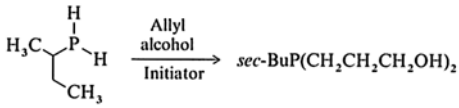

One cyanoethyl group of tris(2-cyanoethyl)phosphine can be cleaved under basic conditions if the phosphine compound is first converted to a quaternary compound by treatment with an alkyl halide:

$$P(CH_2CH_2CN)_3 + BrCH_2CH_2CH_2SO_3^-Na^+ \longrightarrow$$

$$(NCCH_2CH_2)_3P^+CH_2CH_2SO_3^- \xrightarrow{CH_3ONa} \qquad (2)$$

$$Na^{+-}O_3SCH_2CH_2P(CH_2CH_2CN)_2$$

This method can be used to produce chiral phosphines with the phosphorus atom as chirality center.

Synthesis from Arylphosphines. Tertiary phosphines with one or two phenyl groups can be obtained either from the corresponding chloroarylphosphines or from triphenylphosphine [22].

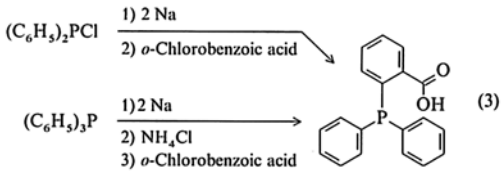

2.3. Uses

Primary and secondary phosphines are intermediates, not usually isolated, in the production of tertiary phosphines and their derivatives. Also, phosphinic and dithiophosphinic acids (see Chap. Phosphinic Acids and their Derivatives), obtained by oxidation of secondary phosphines, have been used to a limited extent for solvent extraction and flotation processes (Cyanex 272, Aerofine 3418 A, American Cyanamid). High-purity *tert*-butylphosphine [2501-94-2] (Cypure, American Cyanamid) can partially replace the more toxic PH_3 in the production and doping of semiconductors in the electronics industry [13]

Triphenylphosphine (TPP) has the greatest industrial importance of all the tertiary phosphines as a ligand in homogeneous catalysis (hydroformylation, hydrogenation, and oligomerization), and as a starting material in the preparation of Wittig reagents, especially for the synthesis of vitamin A and in the production of β-carotene [23]. TPP can be sulfonated by a process developed by Rhône-Poulenc and Hoechst. The rhodium complexes of sulfonated TPP used in hydroformylation are water soluble and can therefore easily be separated from the organic phase [24], [25].

The phosphinoethanesulfonate salt, obtainable by Equation (2), is also water-soluble. It is used as an antioxidant and complexing agent for silver in the production of positive transparencies [26].

The tertiary bicyclic phosphines, produced according to Equation (1), enable long chain α-alkenes to be hydroformylated and hydrogenated in high yield to give alcohols for detergents [27].

The ligand formed in Equation (3) is of major industrial importance. It is used as a nickel complex to catalyze the oligomerization of ethylene to form linear α-alkenes in the Shell Higher Olefin Process (SHOP) [28].

Trialkylphosphines are used as starting materials in the production of phosphonium salts (see Chap. Phosphonium Salts) as well as tertiary phosphine oxides and sulfides (see Chap. Phosphine Oxides and Sulfides). They are also used directly as catalysts for the trimerization of diisocyanates [29].

Triethylphosphine functions as a ligand in a gold(I) complex for oral treatment of arthritis (Auranofin, Smith-Kline & French) [30].

[Top of Page]

3. Halophosphines

Jürgen Svara and Norbert Weferling

The halophosphines RPX_2 (phosphonous acid dihalides) and R_2PX (phosphinous acid halides) are important starting materials in the production of numerous organophosphorus compounds.

3.1. Properties

Halophosphines are generally colorless liquids of high density. Due to their high reactivity and water and air sensitivity, they are used only as intermediates in the production of plant protection agents, stabilizers for plastics, and catalysts. Some physical properties of the industrially important compounds are listed in Table 4.

Table 4. Physical properties of halophosphines

Formula	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	*mp*, °C	n_D^{20}	Flash point, °C
CH_3PCl_2	[676-83-5]	116.92	1.32	81 (1.03×10^5)	<–64	1.4960	23 [a]
$C_6H_5PCl_2$	[644-97-3]	178.99	1.318	224 (1.03×10^5)	–51	1.5958	102
$Cl_2PC_6H_4C_6H_4PCl_2$	[35346-33-9]	355.96		175 (665)	63		
$(C_6H_5)_2PCl$	[1079-66-9]	220.64	1.19	320 (1.03×10^5)	15	1.634	138 [b]

[a] Pensky – Martens closed cup.
[b] Cleveland open cup.

3.2. Production

From Hydrocarbons and PCl_3. Alkyl- and aryldichlorophosphines are obtained by the free radical reaction of PCl_3 with hydrocarbons in the gas phase at high temperature. Halogen compounds are used as catalysts [31-35]. In the case of aromatic compounds, yellow phosphorus may be added as an additional component [36].

$$CH_4 + PCl_3 \xrightarrow[CCl_4]{550-650\,°C} CH_3PCl_2$$

$$C_6H_6 + PCl_3 \xrightarrow[C_6H_5Cl]{450-600\,°C} C_6H_5PCl_2$$

By the Friedel – Crafts Reaction. The reaction of phosphorus trichloride with aromatic hydrocarbons in the Friedel – Crafts reaction takes place under considerably milder conditions than the gas-phase reaction. However, complexes of halophosphine and $AlCl_3$ are produced, which must be decomposed by adding $POCl_3$, ketones, or esters, forming more stable aluminum chloride complexes [37], [38].

$$C_6H_6 + PCl_3 + AlCl_3 \longrightarrow C_6H_5PCl_2 \cdot AlCl_3 + HCl$$

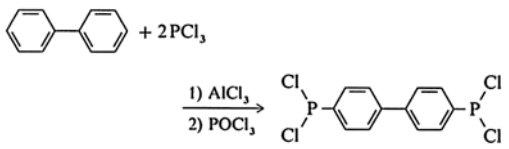

By the Kinnear – Perren Reaction. Alkyltrichlorophosphonium tetrachloroaluminates can be obtained from phosphorus trichloride, alkyl chlorides, and aluminum trichloride. Alkyldichlorophosphines can be obtained from the phosphonium salts by reduction and complexing the $AlCl_3$.

$$RCl + PCl_3 + AlCl_3 \longrightarrow RPCl_3^+AlCl_4^-$$

$$3\,RPCl_3^+AlCl_4^- \xrightarrow{+2P + 3KCl} 3\,RPCl_2 + 2\,PCl_3 + 3\,KAlCl_4$$

$$3\,RPCl_3^+AlCl_4^- \xrightarrow{+2Al + 5KCl} 3\,RPCl_2 + 5\,KAlCl_4$$

By Disproportionation or Comproportionation. Chlorodiphenylphosphine and PCl_3 are obtained by the disproportionation of dichlorophenylphosphine in the gas phase at elevated

temperature [39]. Similarly, triphenylphosphine or triphenylphosphine oxide can be comproportionated with PCl_3 or dichlorophenylphosphine [40-43].

$$2\,C_6H_5PCl_2 \longrightarrow (C_6H_5)_2PCl + PCl_3$$

$$(C_6H_5)_3P + 2\,PCl_3 \longrightarrow 3\,C_6H_5PCl_2$$

$$(C_6H_5)_3PO + 2\,PCl_3 \longrightarrow 2\,C_6H_5PCl_2 + C_6H_5P(O)Cl_2$$

From Elemental Phosphorus. The reaction of red phosphorus with perfluoroiodoalkanes to give a mixture of perfluoroalkyldiiodophosphine and bis(perfluoroalkyl)iodophosphine is of industrial importance [44].

$$2\,P_{red} + 3\,RI \longrightarrow RPI_2 + R_2PI$$

$$R = C_nF_{2n+1},\ n = 6-12$$

3.3. Uses

Dichloromethylphosphine (Hoechst) [45] is used in the production of methylphosphinic acid [4206-94-4], $CH_3PH(O)OH$, which is used in the production of the total herbicide Gluphosinate. Another application is in the production of a cyclic phosphinocarboxylic anhydride, which is used in the flameproofing sector (see Sections Phosphinic Acids and Phosphinate Esters and Other Phosphinic Acid Derivatives). Dichlorophenylphosphine (Akzo, Ferro Corp.) is converted to phenylphosphinic acid [1779-48-2] by hydrolysis (see Section Phosphinic Acids and Phosphinate Esters), or disproportionated to give chlorodiphenylphosphine. It has largely lost its importance as a precursor for the insecticide EPN (see Section Dialkyl and Diaryl Phosphonates). The bifunctional biphenyl derivative $Cl_2P–C_6H_4–C_6H_4–PCl_2$ [35346-33-9] (Sandoz) is used in the production of the plasticizer Sandostab P-EPQ (see Section Properties and Uses).

Chlorodiphenylphosphine (Akzo) is an important starting material for the production of UV-hardening paint systems (see Section Uses). Mixtures of perfluoroalkyliodophosphines are converted by hydrolysis to perfluoroalkylphosphonic and perfluoroalkylphosphinic acids, which are used for cleaning and degreasing (Fluowet PP, Hoechst).

[Top of Page]

4. Phosphonium Salts [46]

Jürgen Svara and Norbert Weferling

Phosphonium salts are four-coordinate phosphorus compounds of the type $[R_nPH_{4-n}]^+ X^-$, where the phosphorus atom carries a positive charge. In the industrially important phosphonium compounds the group R is alkyl, benzyl, or aryl. The counterion X^- can be organic or inorganic. If both charges are present on the same molecule, the compound is referred to as a phosphabetaine. Phosphonium salts cannot in all cases be easily distinguished from phosphoranes with a five-coordinate phosphorus atom, because they can exist in either form, depending on the solvent and/or the state of aggregation.

4.1. Properties

Phosphonium salts, with few exceptions, are colorless, crystalline solids with good to very good solubility in water. Compounds with short-chain (C_1 – C_4) and long-chain (> C_8) alkyl groups in the same molecule exhibit surfactant properties. Like the corresponding ammonium compounds, they are biocides (see Section Uses), though in general their bactericidal properties are weaker. Their toxicity to fish is lower than that of the corresponding ammonium salts. Many phosphonium salts have an extremely irritant effect on skin and mucous membranes and moderate acute toxicity. They are biologically degraded only very slowly.

Tetraalkylphosphonium salts are oxidized by bases (OH^-, RO^-) with elimination of alkane to give tertiary phosphine oxides only upon prolonged reaction at elevated temperature [47]. Phosphonium salts are thus more stable than the corresponding ammonium compounds under the conditions of phase-transfer catalysis.

Table 5 lists the physical properties of some phosphonium salts.

Table 5. Physical properties of phosphonium salts

Formula	CAS no.	M_r	Density, g/cm^3	^{31}P NMR, ppm	*bp* at 103 kPa, °C	*mp*, °C	Flash point, °C
$[(CH_3)_4P]Cl$	[1941-19-1]	126.57		+24.7	400		> 300
$[(CH_2OH)_4P]Cl$	[124-64-1]	190.56	1.33	+25.2		150	
$[(C_4H_9)_4P]Br$	[3115-68-2]	339.34		+34		95	290
$[(C_4H_9)_3PC_{16}H_{33}]Br$	[14937-45-2]	507.66		+32.6		55	
$[(C_4H_9)_3PCH_3]I$	[1702-42-7]	344.26		+32.0		145	285
$[(C_6H_5)_4P]Br$	[2751-90-8]	419.31		+20.8		295	295 [a]
$[(C_6H_5)_3PCH_2OCH_3]Cl$	[4009-98-7]	342.80				180	

[a] Thermal decomposition.

4.2. Production

From Phosphines by Formation of Quaternary Compounds. Phosphines, R_nPH_{3-n}, like the corresponding nitrogen compounds, form quaternary salts with acids.

$$R_nPH_{3-n} + HX \longrightarrow [R_nPH_{4-n}]\ X^-$$

$$R = \text{alkyl, aryl};\ X = \text{Cl, Br, I}$$

Owing to their nucleophilicity, phosphines react with alkyl and aryl halides to give the corresponding phosphonium salts; the rate of reaction decreases in the order R′I > R′Br > R′Cl.

$$R_3P + R'X \longrightarrow [R_3PR']^+ X^-$$

$$R = \text{alkyl, aryl, H}$$
$$R'X = \text{alkyl or benzyl halide}$$

(Methoxymethyl)triphenylphosphonium chloride is produced by a one-pot variation of this reaction in which triphenylphosphine (TPP) is reacted with acetyl chloride and dimethoxymethane [48]. Chloromethyl methyl ether, formed as an intermediate, alkylates the TPP in a subsequent reaction.

$$Ph_3P + Cl-\overset{O}{\overset{\|}{C}}-CH_3 + CH_3OCH_2OCH_3 \longrightarrow$$
$$CH_3C(O)OCH_3 + [Ph_3PCH_2OCH_3]^+\ Cl^-$$

In the formation of quaternary compounds from aryl bromides or chlorides, the so-called complex salt method is particularly suitable [49], whereby the molar ratio of phosphine to metal salt is 2:1. Nickel halides, and chlorides of cobalt, zinc, and copper(II) can be used.

$$(C_6H_5)_3P + C_6H_5Br \xrightarrow[200°\,C]{NiBr_2} [(C_6H_5)_4P]^+\ Br^-$$

Addition of Phosphines to Carbonyl Compounds. A very effective method of preparing phosphonium salts is the reaction of phosphines with carbonyl compounds in the presence of at least one equivalent of acid. The carbonyl group is inserted into one or more P–H bonds, forming α-hydroxyalkylphosphonium salts.

$$PH_3 + H_2C{=}O + HCl \longrightarrow [P(CH_2OH)_4]^+\ Cl^-$$

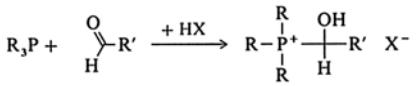

HX = mineral acid, carboxylic acid

Addition of 1,3-Dienes to Halophosphines (McCormack Reaction). 1,3-Dienes react with, e.g., dichlorophosphines, to give chlorophospholenium chlorides in a reaction analogous to the Diels – Alder reaction. The formation of the adduct takes place very slowly under normal conditions, but can be accelerated by the use of, e.g., $POCl_3$ as a solvent. Solvolysis of the product yields phospholene oxides (see Chap. Phosphine Oxides and Sulfides).

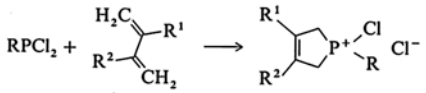

R = alkyl, aryl; R^1, R^2 = H, alkyl

By Anion Exchange. In cases where the presence of halide ions is undesirable (e.g., where they may cause corrosion), strongly basic ion exchangers are used to replace them by hydroxyl ions, which can themselves be easily exchanged for other anions. In this way, the readily accessible tetraalkylphosphonium halides can be converted to phosphonium acetates, which are used as latent hardeners and accelerators in the production of epoxy resins [50].

4.3. Uses

Few other classes of compounds can match the phosphonium salts for the remarkable variety of their technical applications, though the quantities involved are not always large. These include the use of phosphonium salts derived from triphenylphosphine as Wittig reagent precursors [51], and the use of tetraalkylphosphonium salts as reaction accelerators (phase-transfer catalysts [52] or promoters). The latter are used in several processes, including the production of acetic acid and acetic anhydride by the carbonylation of methanol [53], in the production of stabilizers for polypropylene, and the manufacture of isocyanurate esters [54] and polysiloxanes [55]. Owing to their high thermal stability, they are also suitable for chlorine – fluorine exchange in chloronitro aromatic compounds [56]. The quality of dithiophosphoric acids produced from P_4S_{10} and alcohols is significantly improved by the presence of catalytic amounts of quaternary phosphonium compounds [57].

The biocidal properties of tetraalkylphosphonium salts are exploited in cooling water additives (Belacide, formerly Belclene 350, Ciba-Geigy [58]), in antifouling paints [59], and as additives in drilling oils.

2,4-Dichlorobenzyltributylphosphoniumchloride (Phosfleur (Perifleur Products), Chlorphonium [115-78-6]) is used as a growth regulator for plants of the genus chrysanthemum [60].

Technochemie has specialized in the production of quaternary compounds from triarylphosphines. Albright & Wilson produce tetrakis(hydroxymethyl)phosphonium salts as intermediates for the production of flameproofing agents for cotton textiles (Proban process [61]). Also, all phosphine producers produce phosphonium salts (American Cyanamid, Hoechst, Hokko, M & T, and Nippon).

[Top of Page]

5. Phosphine Oxides and Sulfides

Jürgen Svara and Norbert Weferling

The terms phosphine oxide and phosphine sulfide denote four-coordinate phosphorus compounds with only organic groups or hydrogen atoms present in addition to the P=O or P=S double bond.

5.1. Properties

Primary phosphine oxides, $RH_2P{=}O$, are only stable under normal conditions if the R group stabilizes the molecule by its steric bulk, e.g., 2,4,6-tri(*tert*-butyl)phenylphosphine. Even under mild conditions, disproportionation products (primary phosphine and phosphonous acid) are formed during oxidation of primary phosphines.

Primary phosphine sulfides are also unstable and tend to form cyclic polyphosphines at elevated temperature, with liberation of H_2S.

Secondary phosphine oxides (phosphinous acids) $R_2HP{=}O$ are generally crystalline, colorless, odorless substances that are stable at room temperature. However, they disproportionate at high temperature to form secondary phosphines and phosphinic acids.

$$2\ (CH_3)_2\overset{O}{\overset{\|}{P}}H \xrightarrow{T > 100\,°C} (CH_3)_2PH + (CH_3)_2\overset{O}{\overset{\|}{P}}OH$$

Tertiary phosphine oxides are, with few exceptions, crystalline substances. They are odorless, resistant to hydrolysis, and of low acute toxicity, but are not easily biologically degradable. The physical properties of some phosphine oxides are given in Table 6.

Table 6. Physical properties of phosphine oxides

Formula	CAS no.	M_r	Density g/cm³	^{31}P NMR, ppm	*bp* (*p*), °C (Pa)	*mp*, °C	Flash point, °C
$(CH_3)_3PO$	[*676-96-0*]	92.08		+ 36.2	214 (1.03×10^5)	139.5	
$(C_8H_{17})_3PO$	[*78-50-2*]	386.64	0.8 (50 °C)	+ 42.0	360 (1.03×10^5)	50	240
sec-$C_4H_9PO(C_8H_{17})_2$	[*114415-78-0*]	330.53	0.88	+ 51.1	165 (70)	< − 10	> 200
	[*930-38-1*]	116.10	1.094 (50 °C)	+ 71.5[a]	80 (100)	36	113
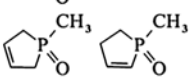 isomer mixture	[*31563-86-7*]	116.10	1.12	+ 71.5, + 73.8[a]	150 (3300)	< 0	125

[a] In alkaline medium: 60.2/62.6 ppm.

5.2. Production

Oxidation of Secondary and Tertiary Phosphines. Secondary and tertiary phosphines react smoothly with equimolar amounts of H_2O_2 in aqueous solution or with elemental sulfur in an aprotic organic solvent under mild conditions to form the corresponding secondary and tertiary phosphine oxides or sulfides. These methods are by far the most important for the production of these classes of substances.

$$R_nH_{3-n}P \xrightarrow{+\ H_2O_2/S_8} R_nH_{3-n}P{=}X$$

X = O, S; n = 2, 3

In the reaction of secondary phosphines with H_2O_2, the reaction conditions— especially the reaction temperature—must be chosen so that formation of phosphinic acids is minimized.

Water-soluble tertiary phosphines can be oxidized with water, with liberation of hydrogen.

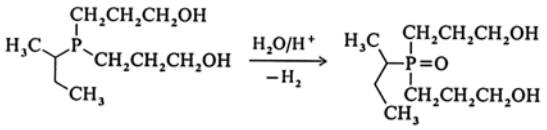

$$P(CH_2OH)_3 \xrightarrow[-H_2]{H_2O/OH^-} (HOCH_2)_3P{=}O$$

The use of oxygen or air instead of H_2O_2 can be recommended only for the oxidation of secondary phosphines. When tertiary phosphines react with oxygen, the P–C bond is broken, leading to the formation of secondary phosphine oxides, phosphonates, and phosphinates, the tertiary phosphine oxides being obtained only as side products.

By Grignard Reaction. Phosphoryl chloride, phosphinic acid dichlorides, and phosphonic acid chlorides react with Grignard reagents with substitution of the halogen to form tertiary phosphine oxides.

$$POCl_3 \xrightarrow[-3\,MgXCl]{+3\,RMgX} R_3P{=}O$$

$$R^1POCl_2 \xrightarrow[-2\,MgXCl]{+2\,RMgX} R^1R_2P{=}O$$

$$R^1R^2POCl \xrightarrow[-MgXCl]{+RMgX} R^1R^2RP{=}O$$

The industrial production of trioctylphosphine oxide is carried out by this process, among others.

The reaction of the corresponding thio compounds with Grignard reagents often gives diphosphine disulfides as side products.

Secondary phosphites (dialkyl phosphonates) and tertiary phosphites can also be reacted with Grignard reagents to give tertiary phosphine oxides.

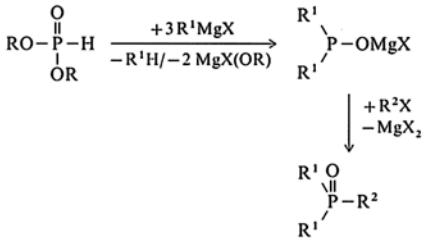

Solvolysis of P – Cl Compounds. Monohalophosphines react with water or alcohols to give secondary phosphine oxides, whereas monohalophosphonium salts (dihalophosphoranes) form tertiary phosphine oxides under these conditions.

$$R_2PCl \xrightarrow[-RCl]{+ROH} R_2\overset{O}{\overset{\|}{P}}{-}H$$

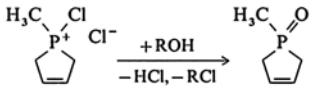

(4)

By the Michaelis – Arbusov Reaction. Esters of phosphinous acids react with alkyl halides, generally smoothly and with high yields, to form the metathesis products, with rearrangement of the organophosphorus product to give the tertiary phosphine oxide. Acylphosphine oxides may be obtained by using acyl halides.

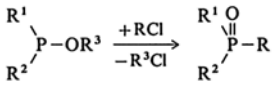

R = alkyl, benzyl, benzoyl

By Reaction with Carbonyl Compounds. Carbonyl compounds react readily with secondary phosphines and phosphine oxides with substitution at the P–H bond to give the corresponding tertiary phosphines or phosphine oxides with an α-hydroxyl group.

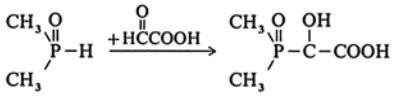

By Reaction of Phosphinate Esters with Activated Alkenes. Esters of phosphinous acids react smoothly with reactive alkenes in the presence of proton donors (not carboxylic acids) to give tertiary phosphine oxides. Phosphonous acid esters are converted to phosphinic acid esters.

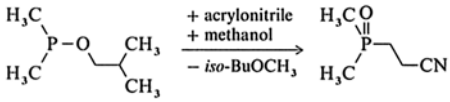

5.3. Uses

Tertiary phosphine oxides and sulfides are used almost entirely as solvent extraction media [62]. Trioctylphosphine oxide (TOPO) is of particular importance (Cyanex 921, American Cyanamid; Hostarex PX 324, Hoechst; Hokko), being used for the extraction of metal ions [63-65] or carboxylic acids, alcohols, and phenol from aqueous solutions [66], [67]. Products such as Cyanex 923 (American Cyanamid, a mixture of tertiary phosphine oxides containing hexyl and octyl groups), and Hostarex PX 320 (Hoechst, *sec*-butyldioctylphosphine oxide) were also developed for this application [68], [69]. Cyanex 471 X (American Cyanamid, triisobutylphosphine sulfide) is used for the extraction of noble metals (silver, platinum, and palladium) and mercury from strongly acidic solutions [70]. The use of tertiary phosphine oxides as catalysts is still restricted to a few reactions, though these are of economic importance. Phospholene oxides [71], which can be obtained, for example, by Equation (4), catalyze the formation of carbodiimides from isocyanates [71], and of polyamides and polyimides [72]. A further application is the production of lactones from acid chlorides [73]. Tertiary phosphine oxides can also be used for stabilizing peroxycarboxylic acids in detergents [74].

Bifunctional phosphine oxides, such as *sec*-butyl-bis(3-hydroxypropyl)phosphine oxide or 1-hydroxy-1-dimethylphosphinoacetic acid are reactive flameproofing agents [75] and herbicides [76], respectively.

Mono- and bis-acylphosphine oxides are used as initiators for the hardening of photopolymerizable materials by UV radiation [77]. They are especially suitable for hardening white paints [78] and dental materials [79].

[Top of Page]

6. Phosphonous Acid Derivatives

Jürgen Svara and Norbert Weferling

Organic derivatives of phosphonous acid are formally derived from the hypothetical phosphonous acid $H–P(OH)_2$. Structures containing three-coordinate phosphorus are only stable if both hydroxyl groups are blocked, for example, by esterification (**1**) or amide formation (**2**). Monoesters (**3**), however, contain four-coordinate phosphorus and are regarded as derivatives of phosphinic acid $H_2P(=O)OH$. Phosphonous acid dihalides (dichlorophosphines) $R–PX_2$ are treated in Chapter Halophosphines.

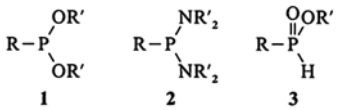

1 **2** **3**

6.1. Production

Diesters of phosphonous acid are formed by reacting dihalophosphines with alcohols in the presence of bases, or by reacting phosphorous acid diester chlorides with Grignard reagents.

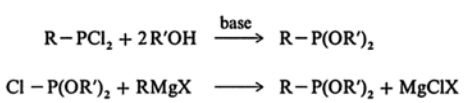

Phosphonous acid diamides can be obtained from dihalophosphines and amines, or from phosphorous acid diamide halides and Grignard reagents.

6.2. Properties and Uses

Phosphonous acid derivatives are generally readily oxidized to phosphonic acid derivatives. Although the phosphorus – carbon bond is very stable, hydrolytic cleavage of the bond between the phosphorus atom and the hetero atoms takes place even under mild conditions. The tetraester of a diphosphonous acid **4** is used under the names Sandostab P-EPQ (Sandoz) and Irgafos P-EPQ (Ciba-Geigy) for the thermal stabilization of plastics [80], [81].

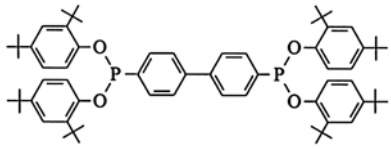

4 [*38613-77-3*]

[Top of Page]

7. Phosphinic Acids and their Derivatives

Jürgen Svara and Norbert Weferling

Organophosphinic acids are the organic derivatives of phosphinic acid (hypophosphorous acid), $H_2P(=O)OH$, in which one or both of the hydrogen atoms on the phosphorus atom are replaced by organic groups. The free acids are of industrial importance, as are the esters, halides, and thio derivatives.

7.1. Properties

The phosphorus – carbon bonds in diorganophosphinic acids (with the exception of α-hydroxyalkyl, alkoxy, or trihalomethyl compounds) are very stable to hydrolysis under both acid and alkaline conditions. Long-chain substituted phosphinic acids and their dithio analogues form stable complexes with divalent cations, and are used in flotation and metal extraction. A number of phosphinic acid derivatives are intermediates in the production of flameproofing and crop protection agents.

7.2. Production

By Hydrolysis of Halophosphines. Organophosphinic acids are obtained by the hydrolysis of dihalophosphines under nonoxidizing conditions.

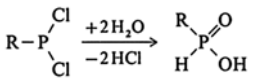

By Oxidation. Secondary phosphines, phosphine oxides, or halophosphines are oxidized to phosphinic acid derivatives by oxygen, sulfur, halogens, or other oxidizing agents. The most important industrial processes are the oxidation and sulfurization of secondary phosphines [82].

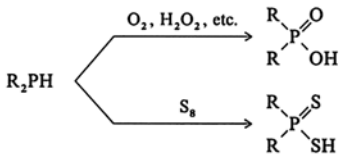

By the Michaelis – Arbusov Reaction (see Section By the Michaelis – Arbusov Reaction.). Alkyl diorganophosphinates are obtained by reacting dialkyl esters of phosphonous acid with alkyl halides.

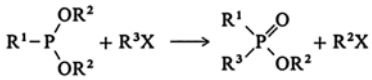

By Addition to Double Bonds. Organophosphinates RHP(=O)OH and their esters can add to compounds containing C=C, C=O, or C=N double bonds in the presence of catalysts [83-85]. Phosphinic acid (hypophosphorous acid) H_3PO_2 [6303-21-5] undergoes double addition to yield symmetrically substituted diorganophosphinates.

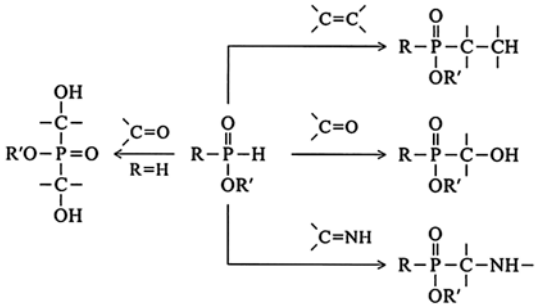

Dihalophosphines react with α, β-unsaturated carboxylic acids to form the dichlorides of carboxyalkylalkylphosphinic acids, which can be converted to cyclic anhydrides [86-88].

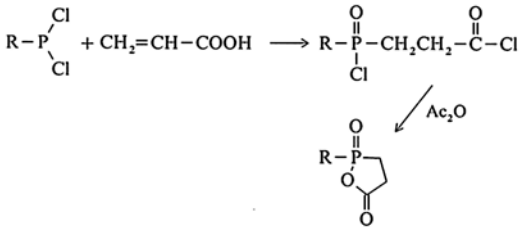

From Tertiary Phosphine Oxides. Treatment of tertiary phosphine oxides with alkali-metal hydroxides at elevated temperature usually causes the most electronegative substituent to be cleaved as an alkane or arene, forming alkali-metal salts of diorganophosphinic acids [89].

$$R_2R'P{=}O + NaOH \xrightarrow{200-300°C} R_2P(=O)ONa + R'H$$

7.3. Phosphinic Acids and Phosphinate Esters

Methylphosphinic acid [4206-94-4], $CH_3PH(=O)OH$, is used in the production of the total herbicide Gluphosinate (Phosphinothricin) [53369-07-6] (Basta, Hoechst), which is used in the form of its ammonium salt [77182-82-2] [90], [91]. The synthesis takes place via the intermediate compound isobutyl methylphosphinate [25296-66-6] or the isoamyl ester [87025-51-2].

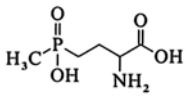

[53369-07-6]

Phenylphosphinic acid [1779-48-2], $C_6H_5PH(=O)OH$, and its sodium salt [4297-95-4] (Akzo, Ferro Corp.) are used to improve the stability of polyamides towards light and heat, and as antioxidants and promoters for emulsion polymerization.

Bis(hydroxymethyl)phosphinic acid [2074-67-1] is an intermediate in the synthesis of crop protection agents [85]. The calcium or magnesium salts (Chemische Fabrik Budenheim) are used as binders in basic fire-resistant materials [92]. The longer chain compounds bis(2,4,4-trimethylpentyl)phosphinic acid [83411-71-6] (Cyanex 272, American Cyanamid) and bis(2-ethylhexyl)phosphinic acid [298-07-7] (P-229, People's Republic of China) are used as extractants in the hydrometallurgical separation of cobalt and nickel [93].

7.4. Thiophosphinates

Thio derivatives of phosphinic acids are mainly used in ore and metal processing. The sodium salt of bis(2-methylpropyl)dithiophosphinic acid [13360-80-0] (Aerophine 3418 A) is supplied by American Cyanamid as a flotation agent for sulfidic ores. Bis(2,4,4-trimethylpentyl)dithiophosphinic acid [107667-02-7] (Cyanex 301, American Cyanamid) has better solubility in organic solvents and is suitable for extracting zinc and other heavy metals [93].

7.5. Other Phosphinic Acid Derivatives

The cyclic phosphinic anhydride 2-methyl- 2,5-dioxo-1,2-oxaphospholane (or 2-methyl-1,2-oxaphospholane-5-one 2-oxide) [15171-48-9] is of major importance in the production of washable fireproofed textiles, and are incorporated into polyester fibers by copolymerization (Trevira CS, Hoechst).

[Top of Page]

8. Phosphites and Hydrogenphosphonates

Jürgen Svara and Norbert Weferling

Phosphites are organic esters of the hypothetical phosphorous acid $P(OH)_3$ in which all three OH groups are esterified. In the mono- and diesters, the phosphorus is four-coordinate. They are therefore formal derivatives of phosphonic acid $(HO)_2P(O)H$ (phosphonates). However, both the properties and the uses of the phosphites and the hydrogenphosphonates are so closely related that treatment together is appropriate.

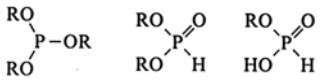

Hydrogenphosphonates and phosphites are mainly used as raw materials for insecticides (⟶ Insect Control) as flameproofing agents, and as stabilizers and antioxidants for plastics (⟶ Antioxidants – Phosphorus Compounds).

8.1. Properties

Hydrogenphosphonates and phosphites are colorless liquids or solids. Owing to their structure, phosphites and phosphonic acid diesters are neutral compounds, while the monoesters have strongly acidic properties. All these compounds are readily hydrolyzable and are susceptible to oxidation. These properties make them suitable as stabilizers for plastics, in which they act by reducing peroxides and by capturing reactive radicals.

8.2. Production

Phosphites and hydrogenphosphonates are usually produced by reacting alcohols or phenols with phosphorus trichloride.

$$PCl_3 + 3\,ArOH \longrightarrow (ArO)_3P + 3\,HCl$$

Whereas reaction of PCl_3 with 3 mol phenol yields the desired triaryl phosphite with liberation of hydrogen chloride, alcohols react to form dialkyl phosphonates and alkyl chloride.

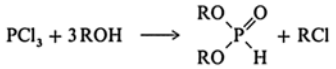

To produce the trialkyl esters, it is necessary to add an equimolar amount of a substance that will neutralize acid. As a rule, ammonia or a volatile amine is used in an inert solvent in which the ammonium chloride formed is sparingly soluble and can be removed by filtration.

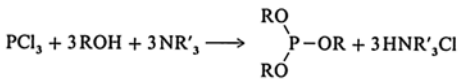

The esterification of H_3PO_3 with alcohols to give dialkyl phosphonates is of less importance [96]. The addition of ethylene oxide to PCl_3 in the presence of catalytic quantities of chloroethanol, iron chloride, or aluminum chloride produces tris(2-chloroethyl) phosphite.

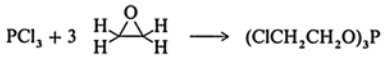

Triphenyl phosphite can be reacted with alcohols to give a mixture of phenyl dialkyl and alkyl diphenyl phosphites.

$$(ArO)_3P + ROH \longrightarrow (ArO)_2POR + ArOH$$

$$(ArO)_3P + 2\,ROH \longrightarrow ArOP(OR)_2 + 2\,ArOH$$

8.3. Trialkyl Phosphites

Some physical properties of the most important trialkyl phosphites are listed in Table 7.

Table 7. Physical properties of trialkyl phosphites $(RO)_3P$

R	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point,[a] °C
CH_3	[*121-45-9*]	124.08	1.053	112 (1.03×10^5)	1.4080	27
C_2H_5	[*122-52-1*]	166.16	0.963	156 (1.03×10^5)	1.4127	54
CH_2CH_2Cl	[*140-08-9*]	269.51	1.353	115 (400)	1.4876	190
$CH(CH_3)_2$	[*116-17-6*]	208.24	0.914	63 (1460)	1.4110	74
$CH_2(CH_2)_2CH_3$	[*102-85-2*]	250.32	0.925	122 (930)	1.4322	121
$CH_2CH(C_2H_5)(CH_2)_3CH_3$	[*301-13-3*]	418.64	0.902	163 (40)	1.445	185
iso-C_8H_{17}	[*25103-12-2*]	418.64	0.891	161 (40)	1.450	146
n-$C_{12}H_{25}$	[*3076-63-9*]	586.97	0.879		1.459	222

[a] Cleveland open cup.

The short-chain compounds are mainly used as starting materials in the production of dialkyl esters of phosphonic acid (Michaelis – Arbusov reaction, see Section Production).

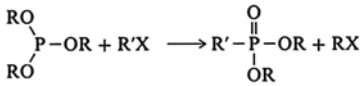

In a further important application, they are converted into insecticidal vinyl esters of phosphoric acid [97]. Manufacturers of trialkyl phosphites include Albright & Wilson, Bayer, and Ciba-Geigy.

The long-chain compounds are mainly used as antioxidants for plastics [98], [99]. Spirocyclic compounds are also used in this application. These are obtained from pentaerythritol [115-77-5] and long- chain alcohols. Manufacturers: Albright & Wilson, Akzo, and Borg-Warner.

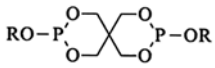

R = iso-$C_{10}H_{21}$ [*26544-27-4*]
$C_{18}H_{37}$ [*3806-34-6*]

[*26741-53-7*]

8.4. Triaryl Phosphites and Alkyl Aryl Phosphites

The physical properties of the more important of these compounds are given in Table 8. These compounds are used as plastics additives.

Table 8. Physical properties of triaryl phosphites $(R^1O)_3P$ and alkyl aryl phosphites $(R^1O)_2(R^2O)P$

R^1	R^2	CAS no.	M_r	Density, g/cm^3	n_D^{20}	Flash point[a], °C
C_6H_5		[*101-02-0*]	310.29	1.180	1.590	146
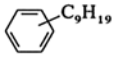		[*26523-78-4*]	689.02	0.990	1.528	207
		[*31570-04-4*]	646.94	1.03		245
C_6H_5	iso-C_8H_{17}	[*26401-27-4*]	346.41	1.045	1.522	182
C_6H_5	iso-$C_{10}H_{21}$	[*26544-23-0*]	374.46	1.025	1.517	154
iso-$C_{10}H_{21}$	C_6H_5	[*25550-98-5*]	438.63	0.94	1.480	160

[a] Pensky–Martens closed cup.

Apart from tris(nonylphenyl) phosphite (TNPP), a product derived from a mixture of nonylphenol and dinonylphenol is used.

Manufacturers (trade names): Akzo (Phosclere), Borg-Warner (Weston, Ultranox), Ciba-Geigy (Irgafos, Irgastab), Dover Chemical Corp. (Doverphos), Hoechst (Hostanox), M&T (Stavinor), Rhône-Poulenc (Garbefix), Uniroyal (Naugard).

8.5. Dialkyl and Diaryl Phosphonates

Table 9 gives a survey of some industrially important dialkyl and diaryl phosphonates.

Table 9. Physical properties of dialkyl and diaryl phosphonates $(RO)_2P(=O)H$

R	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
CH_3	[*868-85-9*]	110.05	1.201	170 (1.03×10^5)	1.4036	64[a]
C_2H_5	[*762-04-9*]	138.10	1.072	50 (270)	1.4080	74[b]
$CH_2(CH_2)_2CH_3$	[*1809-19-4*]	194.21	0.995	118 (930)	1.4238	115[a]
$CH_2CH(C_2H_5)(CH_2)_3CH_3$	[*3658-48-8*]	306.43	0.973	163 (400)	1.4423	168[c]
$C_{12}H_{25}$	[*21302-09-0*]	418.64	0.90		1.452	145[a]
C_6H_5	[*4712-55-4*]	234.19	1.223	135 (150)	1.558	143[a]

[a] Pensky–Martens closed cup. [b] Tag closed cup. [c] Cleveland open cup.

The main areas of application are in the production of phosphonic acid derivatives, insecticides, and plastics additives. Technical-grade diphenyl phosphite is produced by hydrolysis of triphenyl phosphite with water, and usually contains considerable quantities of free phenol. A low-phenol product is obtained by reacting triphenyl phosphite with phosphonic acid.

8.6. Alkyl Phosphonates

Alkyl phosphonates are not of great industrial importance, an exception being monoethyl phosphonate, which is marketed as a fungicide in the form of the aluminum salt $Al[C_2H_5OPH(=O)O]_3$ [39148-24-8] (Fosetyl-Aluminum, Aliette, Rhône-Poulenc) [100], [101].

[Top of Page]

9. Phosphonic Acids and their Derivatives

Jürgen Svara and Norbert Weferling

Of all the various types of organophosphorus compounds, phosphonic acids and their derivatives are outstanding due to their structural variety and great economic importance.

Phosphonic acid derivatives are used as crop protection agents (⟶ Weed Control) in water treatment, in metal processing, and as flameproofing agents.

9.1. Properties

The dibasic phosphonic acids are mostly weaker acids than phosphoric acid. The P–C bond in phosphonic acid derivatives is generally very stable towards oxidation or hydrolysis, so that many reactions can be carried out on the organic part of the molecule. Phosphonic acids are often only slowly biodegradable, but are usually rapidly destroyed by a series of photolytic and biological degradation steps. However, compounds with electronegative substituents on the α-carbon atom are considerably less stable.

Several di- and polyphosphonic acids exhibit complexing (sequestering) properties towards polyvalent cations, and, when added in substoichiometric amounts, prevent the precipitation of low-solubility salts such as alkaline earth sulfates or carbonates (threshold effect).

Esters of phosphonic acid are usually more stable to hydrolysis than the corresponding phosphates, but are converted into free phosphonic acids on heating under acid conditions.

9.2. Production

By Oxidation. The oxidation of low-valency organophosphorus compounds is important mainly in the case of the thiophosphonic acid chlorides. This method is used in the production of methylphosphonic acid dichloride [676-97-1] by reaction of methyldichlorophosphine with sulfuryl chloride or chlorosulfuric acid [102]. Elemental sulfur reacts with $MePCl_2$ in the presence of catalytic quantities of tetraalkylphosphonium salts [103] to form methylthiophosphonic acid dichloride [676-98-2].

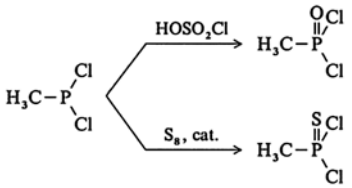

From Organometallic Compounds. Ethylthiophosphonic acid dichloride [993-43-1], $C_2H_5P(=S)Cl_2$, is prepared by the reaction of thiophosphoryl chloride with $(C_2H_5)_3Al_2Cl_3$ or $(C_2H_5)_3Al$ (Ethyl process) [104], [105].

By the Michaelis – Arbusov Reaction. One of the most important reactions in organophosphorus chemistry is the Michaelis – Arbusov reaction, in which compounds containing three-coordinate phosphorus and at least one alkoxy or alkylthio group react with an alkylating agent R′X, forming a phosphorus – carbon bond in compounds with coordination number 4 [106], [107].

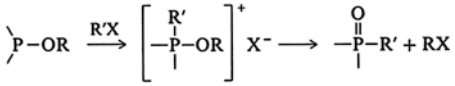

The intermediate phosphonium compounds can only be isolated in exceptional cases. They react further by loss of alkyl halide RX. If R=R′, isomerization occurs with only catalytic quantities of alkylating agent. The method offers a great variety of possibilities, owing to the large range of possible groups R and R′.

Industrially important starting materials for the Michaelis – Arbusov reaction are trialkyl phosphites $P(OR)_3$, which are converted to dialkyl esters of alkanephosphonic acid and to esters of phosphonocarboxylic acids, from which the free phosphonic acids are obtainable by hydrolysis.

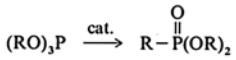

$$(C_2H_5O)_3P + ClCH_2COOC_2H_5 \longrightarrow (C_2H_5O)_2\overset{O}{\overset{\|}{P}}CH_2COOC_2H_5 + C_2H_5Cl$$

If α-halocarbonyl compounds are used as the alkylating agents, vinyl esters of phosphoric acid are obtained (Perkow reaction, see Chap. Esters of Phosphoric Acid).

By Addition to C=C Double Bonds. Dialkyl phosphonates undergo base-catalyzed addition to activated C=C double bonds. The reaction is mainly used for the production of phosphonocarboxylate esters, and can be carried out as a continuous process [108]. In the presence of a radical initiator, nonactivated alkenes can also be used.

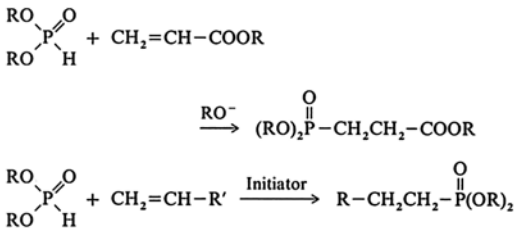

By the Quasi-Mannich Reaction. Phosphonic (phosphorous) acid H_3PO_3 or its dialkyl esters react with mixtures of formaldehyde with ammonia or amines under acidic conditions to give oligo(methylenephosphonates). When oligo-(ethylene)amines are used, the industrially important poly(methylenephosphonic acids) are formed.

$$NH_3 + CH_2O + H_3PO_3 \xrightarrow{H^+} N(CH_2PO_3H_2)_3$$

By Acylation of Phosphonic Acid. The reaction of phosphonic acid with acetyl chloride or acetic anhydride yields 1-hydroxyethane-1,1-diphosphonic acid [2809-21-4] after hydrolysis.

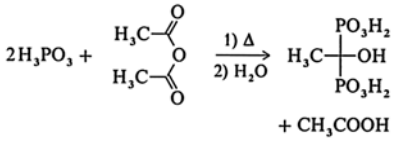

Alternative starting materials are PCl_3 and acetic acid in the presence of water [109].

The reaction can also be carried out with other carboxylic acid derivatives [110-113]. For example, with acetonitrile, 1-aminoethane-1,1-diphosphonic acid [15049-85-1] is obtained.

9.3. Phosphonic Acids and Phosphonocarboxylic Acids

Methylphosphonic acid [993-13-5] (Ciba-Geigy) is used in the production of lubricant additives and for treating textiles [114], [115]. Octylphosphonic acid [4724-48-5] (Hoechst) is used as a selective collector for the flotation of cassiterite (tin ore). Vinylphosphonic acid [1746-03-8] (Hoechst) [116], [117] or its polymers are used for the surface treatment of aluminum in the manufacture of printing plates [118]. Phenylphosphonic acid [1571-33-1] (Akzo) is used as a catalyst in the production of resins and for manufacturing stabilizers for plastics.

Salts of phosphonoformic acid and phosphonoacetic acid (**5**) have virostatic properties. Trisodium phosphonoformate (**6**) (Foscarnet) is used in the treatment of herpes (Triapten, formerly VEB Germed) or cytomegalic infections in humans (Foscavir, Astra).

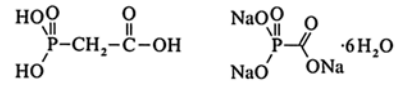

5 [*4408-78-0*] **6** [*34156-56-4*]

2-Phosphonobutane-1,2,4-tricarboxylic acid (**7**) [37971-36-1] (NH_4 salt [70233-62-4], Bayhibit AM, Bayer) is used in industrial water treatment on account of its complexing properties.

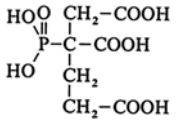

7 [*37971-36-1*]

Geminal diphosphonic acids are used in the same application, and in washing and cleaning agents, peroxide stabilizers, and many other areas. Compounds of this type include 1-hydroxyethane-1,1-diphosphonic acid [2809-21-4] (Briquest ADPA, Albright & Wilson; Sequion 10 H 60, Bozetto; Turpinal SL, Henkel; Jaypol 210, M & J Polymers; Dequest 2010/2016, Monsanto; Mykon P60, Warwick) and 1-aminoethane-1,1-diphosphonic acid [15049-85-1] (BK Ladenburg), as well as the large group of poly(methylenephosphonic acids) produced from ammonia or ethylene amines. Some important products are nitrilotris(methylenephosphonic acid) [6419-19-8], ethylenediaminetetrakis(methylenephosphonic acid) [1429-50-1] and diethylenetriaminepentakis(methylenephosphonic acid) [15827-60-8]. Producers (trade names) are: Albright & Wilson (Briquest), Bozetto (Sequion), M & J Polymers (Jaypol), Monsanto (Dequest).

The total herbicide *N*-carboxymethylaminomethanephosphonic acid (*N*-phosphonomethylglycine) [1071-83-6], Glyphosate (Roundup, Monsanto) is outstanding for its effectiveness and ease of biodegradability. It is used in the form of its isopropylammonium salt [38641-94-0]. The sodium salt of 2-chloroethylphosphonic acid [16672-87-0] Ethepon (Ethrel, Union Carbide) causes accelerated ripening of fruit by releasing the ripening hormone ethylene in the plant.

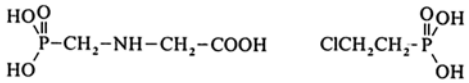

[*1071-83-6*] [*16672-87-0*]

9.4. Esters of Phosphonic Acid

Monoesters of phosphonic acid are used in industry only to a limited extent. Salts of short-chain alkyl alkylphosphonates have a flameproofing effect [119], [120]. The 2-ethylhexyl ester of 2-ethylhexylphosphonic acid (**8**) (PC-88A, Daihachi; Ionquest 801, Albright & Wilson) is used as an extractant for lanthanides and for the separation of cobalt and nickel [121].

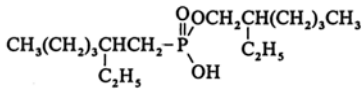

8 [*14802-03-0*]

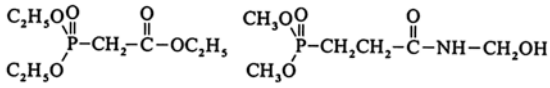

9 [*867-13-0*] **10** [*20120-33-6*]

Dialkyl alkylphosphonates are used as flameproofing agents, metal extractants, plasticizers, lubricant additives, and chemical intermediates. The more important compounds are listed in Table 10 with their physical properties. Producers of the methyl compound DMMP include Akzo, Albright & Wilson, Bayer, Ciba-Geigy, and Courtaulds. Other compounds are produced by Albright & Wilson and Daihachi.

Table 10. Physical properties of dialkyl alkylphosphonates RP(=O)(OR)$_2$

R	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
CH_3	[*756-79-6*]	124.08	1.174	62 (1330)	1.4137	92[a]
C_2H_5	[*78-38-6*]	166.16	1.025	82 (1433)	1.4148	90[b]
$CH_2CH_2CH_2CH_3$	[*78-46-6*]	250.32	0.948	127 (330)	1.4310	125[c]
$CH_2CH(CH_2)_3CH_3$ | C_2H_5	[*126-63-6*]	418.64	0.908	160 (33)	1.4481	158[a]

[a] Pensky–Martens closed cup. [b] Tag closed cup. [c] Cleveland open cup.

Diethyl ethoxycarbonylmethylphosphonate (**9**) (triethyl phosphonoacetate; PEE, Hoechst) is used as a stabilizer for polyesters [122]. Esters of phenylmethylphosphonic acid with bulky substituents (Irganox 1222 [976-56-7] and 1093 [3135-18-0], Ciba-Geigy) are sold as antioxidants and high-temperature stabilizers for polyamides, polyesters, polyolefins and polyurethane foams.

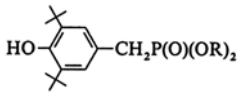

R = C_2H_5; [*976-56-7*]
R = $C_{18}H_{37}$; [*3135-18-0*]

A wide range of other phosphonate esters, often containing chlorine, are used as flameproofing agents. Producers (trade names) are: Akzo (Fyrol, Victastab), Albright & Wilson (Amgard, Antiblaze), Sandoz (Sandoflam 5087). Dimethyl 3-hydroxymethylamino-3-oxopropanephosphonate (**10**) is of particular importance in the production of washable flameproofed cellulose textiles (Pyrovatex CP, Ciba-Geigy) [123-125].

9.5. Other Derivatives of Phosphonic Acid

Halides of phosphonic and thiophosphonic acids are mainly used in the synthesis of crop protection agents. Ehylthiophosphonic acid dichloride is used in the production of the insecticide Fonofos [944-22-9] (Dyfonate, ICI) and Trichloronat [327-98-0] (Agrisil, Phytosol, Bayer), and benzenethiophosphonic acid dichloride [126] in the production of *O*-2,4-dichlorophenyl *O*-ethyl phenylphosphonothioate (S-Seven) [3792-59-4] and EPN [2104-64-5] (both Nissan Chemical Industries).

The acid halides are produced by Akzo, Ethyl, Ferro, and Nissan Chemical Industries. Physical properties are listed in Table 11.

Table 11. Physical properties of phosphonic and thiophosphonic acid halides

Formula	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
$CH_3P(O)Cl_2$	[676-97-1]	132.91	1.39 [a]	163 (1.03×10^5)	1.462 [a]	>110
$C_6H_5P(O)Cl_2$	[824-72-6]	194.99	1.394	258 (1.03×10^5)	1.5600	204 [b]
$CH_3P(S)Cl_2$	[676-98-2]	148.98	1.434	154 (1.03×10^5)	1.548	76 [b]

$C_2H_5P(S)Cl_2$	[993-43-1]	163.01	1.35	172 (1.03×10^5)	1.541	74 [b]
$C_6H_5P(S)Cl_2$	[3497-00-5]	211.04	1.376	205 (1.73×10^4)	1.6240	>110

[a] At 38 °C.
[b] Cleveland closed cup.

[Top of Page]

10. Esters of Phosphoric Acid

Jürgen Svara and Norbert Weferling

The esters of phosphoric acid constitute an important group of organophosphorus compounds with a broad application as cleaning agents and emulsifiers, textile improvers, plasticizers and flameproofing agents for plastics, anticorrosion agents, and extractants in hydrometallurgy. Phosphoric acid ester chlorides are starting materials in the production of crop protection agents. Many vinyl esters of phosphoric acid have insecticidal properties (⟶ Insect Control – Vinyl Phosphates).

10.1. Properties

Phosphate esters are colorless liquids or crystalline or waxy solids, depending on the substituents. The triesters are neutral compounds, while the mono- and diesters are strongly acidic. The esters of phosphoric acid hydrolyze in the presence of water. The rate of hydrolysis varies widely and depends on the degree of esterification and the nature of the substituents.

10.2. Production

The important raw materials for the production of esters of phosphoric acid are phosphoryl chloride, phosphorus pentoxide, polyphosphoric acid, and trialkyl phosphites.

Phosphorus oxychloride reacts with phenols to form triaryl phosphates. With aliphatic alcohols, the hydrogen chloride formed must be removed by vacuum or by purging with an inert gas. Alternatively, bases may be added.

$$POCl_3 + 3\,ROH \longrightarrow (RO)_3PO + 3\,HCl$$

If a substoichiometric amount of alcohol is used, ester chlorides of phosphoric acid are formed. However, diester chlorides of phosphoric acid are better obtained by reacting dialkyl phosphonates with chlorine or sulfuryl chloride.

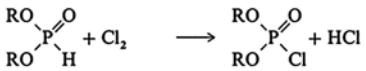

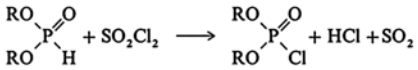

The addition of ethylene oxide or propylene oxide to $POCl_3$ at 40 – 100 °C affords tris(2-chloroethyl) or tris(2-chloropropyl) phosphate. To ensure complete reaction, catalysts must be added ($AlCl_3$, $TiCl_4$, PCl_3).

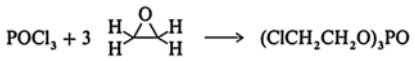

A mixture of mono- and diesters is formed when alcohols or phenols are reacted with phosphorus pentoxide, while the reaction with polyphosphoric acid yields monoesters of phosphoric acid and free phosphoric acid.

$$P_4O_{10} + 6\,ROH \longrightarrow 2(RO)_2PO(OH) + 2\,ROPO(OH)_2$$

$$H_6P_4O_{13} + 3\,ROH \longrightarrow 3\,ROPO(OH)_2 + H_3PO_4$$

An important method for the production of insecticidal vinyl phosphates is the reaction between trialkylphosphites and α-halocarbonyl compounds (Perkow reaction) [127].

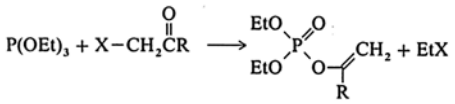

10.3. Trialkyl Phosphates

Table 12 gives a survey of the important trialkyl phosphates.

Table 12. Physical properties of trialkyl phosphates $(RO)_3PO$

R	CAS no.	M_r	Density, g/cm³	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
CH_3	[*512-56-1*]	140.08	1.197	197 (1.03×10^5)	1.3960	none
C_2H_5	[*78-40-0*]	182.16	1.064	215 (1.03×10^5)	1.4039	116
CH_2CH_2Cl	[*115-96-8*]	285.51	1.414	210–220 (2700)	1.4720	252[a]
$CH(CH_3)CH_2Cl$	[*13674-84-5*]	327.57	1.294	[b]	1.4625	218[a]
$CH(CH_2Cl)_2$	[*13674-87-8*]	430.91	1.513	[b]	1.5019	251[a]
$CH_2(CH_2)_2CH_3$	[*126-73-8*]	266.32	0.976	148–153 (1330)	1.4249	165[a]
$CH_2CH_2OC_4H_9$	[*78-51-3*]	398.48	1.006	215–228 (532)	1.4380	224[a]
$CH_2CH(CH_2)_3CH_3$ (with C_2H_5)	[*78-42-2*]	434.64	0.924	196–200 (133)	1.4440	193[a]

[a] Cleveland open cup. [b] Decomposes.

Triethyl phosphate is used as a catalyst in the production of acetic anhydride by the ketene process, as a desensitizing agent for peroxides, and as a solvent and plasticizer for cellulose acetate. Tributyl phosphate is used for solvent extraction in hydrometallurgy. Tris(2-ethylhexyl) phosphate is used as a solvent in the production of hydrogen peroxide and is also used as a plasticizer with flameproofing properties. The important producers of trialkyl phosphates are Akzo, Albright & Wilson, Bayer, Daihachi Chemical Industry, and FMC.

Tris(chloroalkyl) phosphates are used as fire retardants for polyurethane foams, polyesters, and other plastics [128]. In addition to the compounds listed in Table 12, a large number of other halogen-containing, sometimes oligomeric trialkyl phosphates with mixed ester groups are marketed as flameproofing agents. Producers (trade name) are: Albright & Wilson (Amgard), Akzo (Fyrol), Bayer (Disflamoll), Courtaulds, Great Lakes Chemical Corp. (Firemaster), Hoechst (Genomoll), Olin (Thermolin), and Sandoz (Sandoflam).

10.4. Triaryl and Alkyl Aryl Phosphates

Some important aromatic triesters of phosphoric acid are listed in Table 13.

Table 13. Physical properties of triaryl phosphates $(R^1O)_3PO$ and alkyl aryl phosphates $(R^1O)(R^2O)_2PO$

R^1	R^2	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
C_6H_5		[*115-86-6*]	326.29		244 (1330)		223
$C_6H_4CH_3$		[*1330-78-5*]	368.37	1.17	249 (533)	1.555	252
$C_6H_3(CH_3)_2$		[*25155-23-1*]	410.45	1.147	248–265 (533)	1.553	263
$C_6H_4CH(CH_3)_2$		[*68937-41-7*]	452.53	1.16	220–270 (533)	1.552	254
$CH_2CH(CH_2)_3CH_3$ (with C_2H_5)	C_6H_5	[*1241-94-7*]	362.41	1.09	239 (1330) decomp.	1.508	224
iso-$C_{10}H_{21}$	C_6H_5	[*29761-21-5*]	390.46	1.07	245 (1330) decomp.	1.507	241
$C_6H_4CH_3$	C_6H_5	[*26444-49-5*]	340.32	1.20	235–255 (1330)	1.561	230
$C_6H_4C(CH_3)_3$	C_6H_5	[*56803-37-3*]	385.42	1.18	258 (1330)	1.555	263

These compounds are mainly used as flameproofing and plasticizing agents for plastics, as lubricants, and as low-flammability hydraulic fluids. Mixtures of isomers are mainly used, based on technical-grade alkyl phenols.

Some producers (trade names) are: Akzo (Phosflex), Albright & Wilson (Pliabrac), Bayer (Disflamoll), Ciba-Geigy (Reofos), FMC (Kronitex), Monsanto (Santicizer).

10.5. Mono- and Dialkyl Phosphates

Only a small number of pure compounds in this class are of industrial importance. Bis(2-ethylhexyl) phosphate [298-07-7] DEHPA, is used in hydrometallurgy as an extraction solvent for a large number of metals. It is produced by chlorinating bis(2-ethylhexyl) phosphonate [3658-48-8] to give the phosphate diester chloride, followed by hydrolysis [129], or by saponification of tris(2-ethylhexyl) phosphate [78-42-2]. It is produced by Albright & Wilson, Bayer, Daihachi, and Hoechst. Products with a high monoester content, obtained from ethoxylated alcohols, are used as defoamers in washing powders (Sokalan S, BASF).

Esters of thiophosphoric acid are used in plant protection (⟶ Insect Control; ⟶ Fungicides, Agricultural) and as flotation agents in ore preparation, lubricant additives, for solvent extraction of metals, and in rubber production.

Mixtures of mono- and diesters of phosphoric acid are of major industrial importance. Their properties can be varied over a wide range by choice of the alcohol or phenol. Ethoxylated alcohols are most commonly used. The wide range of applications includes emulsifiers in plant protection, cleaning agents, cosmetics, and in the paper and textile industries. Mixtures of mono- and diesters are also used as acid hardeners for resins. Manufacturers (trade names) include Akzo (Dapral, Victawet), Albright & Wilson (Albrite, Briphos, Duraphos), GAF (Gafac), Henkel (Disponil), Hoechst (Hostaphat, Leomin, Knapsack phosphate esters), Hüls (Marlophor), and Rhône-Poulenc (Celanol, Soprophor).

Mixtures of monoesters of phosphoric acid with free phosphoric acid are also marketed. These are mainly used as components of detergents (Hoechst).

[Top of Page]

11. Esters of Thiophosphoric Acid

Jürgen Svara and Norbert Weferling

The esters of thiophosphoric acid are formally derived by the replacement by sulfur of one or more of the oxygen atoms bonded to phosphorus in esters of phosphoric acid. Thiophosphoric acids form mono-, di-, and triesters, the location of the organic substituent being indicated as *O*- or *S*-. Of the large range of possible compounds, the derivatives of mono- and dithiophosphoric acid have attained the greatest industrial importance.

11.1. Properties

Thiophosphate esters generally have an unpleasant smell and in the pure state are colorless liquids or solids. The triesters are neutral, while the mono- and diesters are strongly acidic. Technical-grade *O*,*O*-dialkyl dithiophosphates usually have a greenish or yellowish color due to traces of heavy metals. The short-chain compounds in particular cannot be stored indefinitely, and are usually immediately converted into other products. The thermal and hydrolytic stability increases with increasing alkyl chain length and degree of esterification and upon salt formation, so that diaryl dithiophosphates are more easily hydrolyzed than alkyl esters of comparable chain length [130].

Many esters of thiophosphoric acid have high biological activity as acetylcholine esterase inhibitors and are therefore useful insecticides. However, the toxicity to warm-blooded animals is very high. The compounds are decomposed to less toxic products on contact with the soil, as they are readily hydrolyzed.

11.2. Production

Important raw materials for the production of esters of thiophosphoric acid include phosphorus pentasulfide, thiophosphoryl chloride, and dialkyl hydrogenphosphonates.

Phosphorus pentasulfide reacts with alcohols or phenols with loss of H_2S to form the industrially important *O*,*O*-diesters of dithiophosphoric acid. Addition of catalytic nitrogen, phosphorus, or sulfur compounds (e.g., phosphonium halides) improves the yield of the reaction and the product quality [131-133].

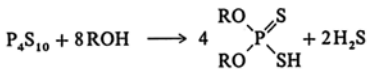

In practice, a suspension of phosphorus pentasulfide in the end product is reacted with excess alcohol at 60 – 80 °C. Due to their lower reactivity, phenols must be reacted at 80 – 120 °C. The hydrogen sulfide produced is burnt, and the sulfur dioxide formed is removed by scrubbing with alkali.

O,*O*-Dialkyl esters of dithiophosphoric acid are converted to *O*,*O*-diester chlorides of thiophosphoric acid by reaction with chlorine. Depending on the conditions, the byproducts are hydrogen chloride and disulfur dichloride [10025-67-9] or sulfur.

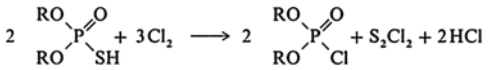

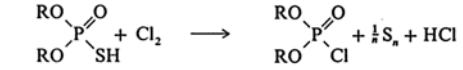

Another route to this class of compounds is the reaction of thiophosphoryl chloride with alcohols in the presence of acid acceptors.

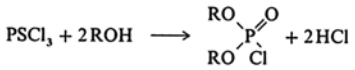

Salts of thiophosphoric acid are produced by reacting dialkyl hydrogen phosphonates with sulfur in the presence of a base (e.g., ammonia). The *O*,*O*-diesters of thiophosphoric acid are liberated from these compounds by reaction with phosphoric acid.

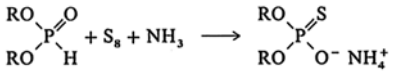

11.3. Monothiophosphates

O,*O*-Diesters of thiophosphoric acid are used as extraction solvents for heavy metals, especially zinc, or as lubricant additives. The *O*,*O*-diester of thiophosphoric anhydride (**11**) (Sandoflam 5060, Sandoz) is used as a halogen-free flameproofing agent for low-density polyethylene films.

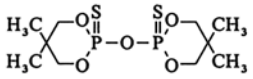

11 [*4090-51-1*]

The alkylation of *O*,*O*-dialkyl thiophosphates usually gives mixtures of *O*,*O*,*O*- and *O*,*O*,*S*-triesters of thiophosphoric acid, some of which are used in crop protection.

The *O,O*-diester chlorides of thiophosphoric acid are very important in the synthesis of crop protection agents, and are usually reacted with phenols or their salts to give insecticidal or fungicidal *O,O,O*-triesters of thiophosphoric acid (Parathion [56-38-2], Methylparathion [298-00-0] ⟶ Insect Control; Pyrazophos [13457-18-6] ⟶ Fungicides, Agricultural). The most important intermediates are *O,O*-dimethoxylthiophosphoryl chloride [2524-03-0], $(CH_3O)_2PSCl$, (M_r 160.56, *bp* 70 – 72 °C (270 Pa), ϱ 1.305 g/cm^3, n_D^{20} 1.4807, flash point 101 °C, colorless liquid (Cheminova), and *O,O*-diethoxylthiophosphoryl chloride [2524-04-1], $(C_2H_5O)_2PSCl$, (M_r188.61, *bp* 71 – 72 °C (93 Pa), ϱ 1.200 g/cm^3, n_D^{20} 1.4688, flash point 106 °C, colorless liquid (Albright & Wilson, Cheminova).

11.4. Dithiophosphates

The *O,O*-diesters of dithiophosphoric acid and aqueous solutions of sodium or ammonium dithiophosphates are used as flotation agents for sulfidic copper and cadmium ores. A new application of increasing importance is the extraction of cadmium from wet phosphoric acid by formation of insoluble complex salts [134-136]. The *O,O*-dialkyl dithiophosphates of short to medium chain length (C_2 – C_8) are particularly effective in this application. The amount added is ca. 0.3 %, depending on the origin of the crude phosphoric acid and the temperature. Preliminary reduction of the acid with iron improves the efficiency of separation of the cadmium [137].

The zinc salts of long-chain esters are used as lubricant additives to improve high-pressure lubrication properties, and useful life of the lubricant [138] (⟶ Lubricants and Lubrication). They are produced by reacting zinc oxide with the alcohol-containing *O,O*-diesters of dithiophosphoric acid. The crude product is recovered by pressure filtration, and the water liberated in the reaction and the excess alcohol are then removed by vacuum distillation.

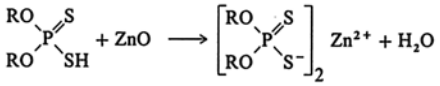

These products protect the metal surface by forming a sulfide and thiophosphate layer, by removing oxidation products from the lubricant, and by corrosion inhibition.

Water-insoluble *O,O*-dialkyl esters of dithiophosphoric acid can be used as extractants for the hydrometallurgical removal of heavy metals such as zinc from dilute acidic solutions, and are therefore of interest for the treatment of wastewater and polluted groundwater.

An important reaction of *O,O*-dialkyl dithiophosphates is the addition of an activated C= C double bond to form *O,O,S*-triesters of dithiophosphoric acid. Thus, *O,O*-dimethyl dithiophosphate [756-80-9] and diethyl maleate react to produce the insecticide Malathion [121-75-5].

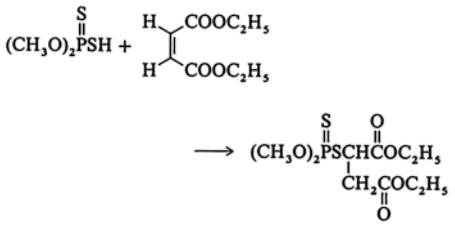

The addition to C = O double bonds is the key reaction in the synthesis of the insecticide Terbufos [13071-79-9] from *O,O*-diethyldithiophosphoric acid [298-06-6], formaldehyde, and 2-methyl-2-propanethiol [75-66-1].

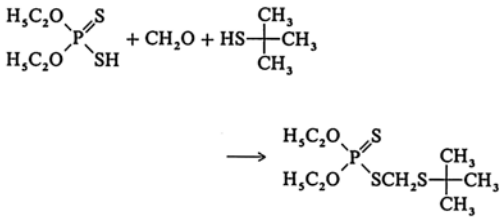

O,O-Dialkyl dithiophosphates are alkylated at the sulfur atom by alkyl halides. This reaction is used to produce numerous insecticidal *O,O,S*-triesters of dithiophosphoric acid.

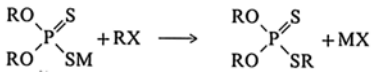

Alkyl Dithiophosphates. The physical properties of the most important *O,O*-dialkyl dithiophosphates are given in Table 14.

Table 14. Physical properties of *O*, *O*-dialkyl dithiophosphates $(RO)_2P(S)SH$

R	CAS no.	M_r	Density, g/cm^3	*bp* (*p*), °C (Pa)	n_D^{20}	Flash point, °C
CH_3	[*756-80-9*]	158.19	1.28	48 (270)	1.5350	92
C_2H_5	[*298-06-6*]	186.23	1.17	85 (650)	1.5070	82[a]
$CH(CH_3)_2$	[*107-56-2*]	214.29	1.10	90 (400)	1.4933	112[a]
$CH_2CH(CH_3)_2$	[*2253-52-3*]	242.34	1.05	107 (300)	1.4908	136[a]
$CH(CH_3)C_2H_5$	[*107-55-1*]	242.34	1.07	100 (200)	1.4928	79[b]
$CH_2CH(CH_2)_3CH_3$ (with C_2H_5)	[*5810-88-8*]	354.56	0.98		1.4855	180[a]

[a] Cleveland open cup. [b] Pensky–Martens closed cup.

Most long-chain compounds are produced from mixtures of alcohols. Manufacturers include Albright & Wilson, American Cyanamid, Cheminova, and Hoechst. Aqueous solutions of the salts are marketed as flotation agents by American Cyanamid (Aerofloat) and Hoechst (Hostaflot). Lubricant additives based on the zinc salt are supplied by Chevron and Lubrizol.

Aryl Dithiophosphates. An important compound of this type is dicresyl dithiophosphoric acid [27157-94-4] $(CH_3C_6H_4O)_2P(S)SH$ (M_r 279.41, brown liquid), used as a flotation agent for sulfidic ores. The commercial product is not usually a pure substance, being the reaction product of a mixture of phenol, cresols, and xylenols with phosphorus pentasulfide. Both the free acid and an aqueous solution of the ammonium salt are commercial products (Aerofloat 25, 31, 242, American Cyanamid; Phosokresol B, C, E, Hoechst).

[Top of Page]

12. Economic Aspects

Jürgen Svara and Norbert Weferling

Measured by the tonnage of inorganic phosphorus-containing base materials, e.g., PCl_3 and $POCl_3$, used in their production, the organophosphorus products discussed here are of minor importance. Based on value, the comparison is more favorable, as they are all value-added products, the value of some specialized products being very high. However, for reasons of industrial secrecy, there is very little information available from the manufacturers concerning production quantities and values, so that some of the figures given here are only estimates.

Of the phosphines, triphenylphosphine is the largest product from the point of view of both tonnage and value, being the starting material for Wittig reagents and catalyst ligands. The output in 1990 in Europe was ca. 2000 t. Products of comparable economic importance include the fire-retardant precursors tetrakis(hydroxymethyl)phosphonium chloride and sulfate.

The organophosphorus product with by far the largest sales value is still phosphonomethyl glycine (Glyphosate, Monsanto), at ca. 10^9 DM/a. Halogen-free phosphonic and phosphinic acid derivatives are of increasing importance in the production of nonflammable textiles. Approximately 1000 t/a of these products are produced.

The market for plastics additives (flame retardants, plasticizers and antioxidants) expressed as PCl_3 in the USA for 1990 is estimated at 30 000 t phosphorus trichloride, corresponding to

6800 t/a phosphorus. Of the total PCl_3 produced, 10 % (15 000 t/a) is converted to phosphorous acid for the production of industrial water-treatment additives. In Europe, the quantity of halogen-containing flameproofing agents based on esters of phosphoric acid was ca. 9000 t/a in 1990. Halogen-free products are estimated at 1500 t/a, and other organophosphorus plastics additives amounted to ca. 5000 t/a. In Japan in 1987, ca. 3600 t of phosphite esters, > 10 000 t halogen-free phosphate esters, and ca. 1400 t halogen-containing flameproofing agents were used. The total Japanese market for organophosphorus products was ca. 30 000 t, with a value of 180×10^9 Yen.

Apart from the esters of thiophosphoric acid used in plant protection, the most important compounds of this class are the zinc dithiophosphates used in lubricants, for which over 50 % of phosphorus pentasulfide production (ca. 65 000 t) was used in the United States in 1990. The quantity of phosphorus pentasulfide used in the production of flotation agents was < 2000 t/a. Production in developing countries with phosphate ore deposits is increasing. For example, in 1989 Chile consumed ca. 2400 t dithiophosphates, of which ca. 50 % was from its own production.

[Top of Page]

13. Toxicology

Thomas Hofmann

Phosphines. Tributylphosphine, a liquid with a garliclike odor, has an LD_{50} of 750 mg/kg in rats following oral administration, whereas the LD_{50} is slightly more than 2000 mg/kg in rabbits upon dermal application. The substance is irritating to the skin and mucous membranes. Inhalation causes nausea and, at higher concentrations, damage to the lungs and kidneys [139]. In rats, 4-h LC_{50} values of 0.17 and 12.5 mg/L have been reported for phenylphosphine and triphenylphosphine, respectively. Symptomatically, mild irritation of the respiratory tract was prominent [140]. Hyperemia, hypersensitivity to touch and sound, dermatitis, and anemia were observed in a 3–month inhalation toxicity study with phenylphosphine in rats and dogs [141]. The TLV-TWA value for phenylphosphine is 0.23 mg/m^3 [142]. For triphenylphosphine, the LD_{50} is 700 mg/kg in rats following oral administration, and in rabbits > 5000 mg/kg following dermal application [143]. It was not mutagenic in the Ames test [144].

Halophosphines. Halophosphines are corrosive to the skin due to their high reactivity with water. For dichloroethylphosphine, the lethal dose in mice following a 10-min inhalation is 1990 mg/m^3. For dichloro(2-fluoroethoxy)phosphine, it is 100 mg/m^3 in guinea pigs [145].

Phosphonium Compounds. Dodecyltriphenylphosphonium bromide possesses antimycotic properties and has been used in invert soaps. In higher concentrations, it is irritating to the skin [146].

Tetrakis(hydroxymethyl) phosphonium salts (chloride, sulfate, acetate/phosphate) are used to produce crease-resistant and flame-retardant finishes on cotton textiles and cellulosic fabrics. Tetrakis(hydroxymethyl)phosphonium chloride and tetrakis(hydroxymethyl)phosphonium sulfate were tested for carcinogenicity by oral administration in rats and mice and did not cause a dose-related increase in the incidence of tumors. However, male rats administered with the low dose level showed an increased incidence of leukemia, and an increased incidence of malignant lymphomas occurred in male mice of the low dose level. Tetrakis(hydroxymethyl) phosphonium chloride was negative in a two-stage skin initiation-promotion bioassay in mice. The mixed acetate/phosphate salt exhibited a weak promoting activity on mouse skin. Results of genotoxicity studies with these compounds were not consistent [147].

Phosphorous – Nitrogen Compounds. *N*,*N*-Bis (2-chloroethyl)tetrahydro-2*H*-1,3,2-oxazaphosphorin-2-amine 2-oxide monohydrate (cyclophosphamide) is used as a cytostatic agent in tumor therapy. In doses comparable to those given in clinical treatment, the compound causes an increased incidence of tumors in mice and rats and it is also genotoxic. Only in the organism is it metabolized to the active compound.

1,1′,1″-Phosphinylidenetris(2-methylaziridine) (METEPA) is used as a cross-linking agent in dye, textile, and polymer chemistry. The acute oral toxicity in rats (LD_{50}) is 136 mg/kg (male animals) and 213 mg/kg (female animals). TEPA (1,1′,1″-phosphinylidenetrisaziridine), which was formerly used as a flameproofing agent in the textile industry, found only limited application in the treatment of cancer, since the thio analogue was equally effective and had a lower toxicity. The LD_{50} in rats is 37 mg/kg following oral administration and 87 mg/kg upon dermal application [148]. Thio-TEPA (1,1′,1″-phosphinothioylidenetrisaziridine) is used as an antineoplastic agent. The LD_{50} in mice is 46 mg/kg following oral administration. The substance is genotoxic and leads to a higher incidence of malignant tumors in mice following intraperitoneal application and in rats following intraperitoneal or intravenous administration. Several cases of leukaemia have been reported in humans following treatment with Thio-TEPA [148], [178]. In contrast, hexamethylphosphoric triamide, which is used as solvent, catalyst, and stabilizer for polystyrene, is only slightly toxic after single dose administration (LD_{50}, oral: male rats, 2650 mg/kg; female rats, 3360 mg/kg). However, upon chronic inhalation, the substance caused a dose-dependent increase in the incidence of nasal tumors in rats. Moreover, pathological changes occurred in the entire respiratory tract [149-151]. In the United States the substance is classified as a suspected human carcinogen [142].

Phosphate Esters.
Trimethyl phosphate is metabolized to dimethyl phosphate in the body; the latter is excreted as a conjugate. The substance is only moderately toxic in acute doses (LD_{50} acute, oral, rat 2000 mg/kg). Hyperexcitability and tremor are prominent symptoms, and at higher doses death occurs due to cessation of respiration. Following repeated administration, flaccid and partly spastic paralysis of the extremities have been observed in rabbits [152]. In dogs, too, trimethyl phosphate is neurotoxic. The peripheral nerve fibers were found to be damaged [153]. Increased incidences of fibroma in male rats and increased incidences of endometrial tumors in female mice were observed after two years of oral administration of trimethyl phosphate to rats and mice for two years. Female rats and male mice showed no increased incidence of tumors [154]. Trimethyl phosphate was mutagenic in most test systems; however, high doses or concentrations were required in most cases [155-159], [179]. A sterilizing effect was found in mice, rats, and rabbits for trimethyl phosphate, whereby the target cells were the spermatozoa in the epididymis [160]. *Dimethyl phosphate*, a metabolite of trimethyl phosphate, shows no antifertility effects in male animals, in contrast to trimethyl phosphate. It has been found to be nongenotoxic in various test systems [170].

Triethyl phosphate, which is metabolized to diethyl phosphate and is excreted in the urine as such together with *S*-ethyl cysteine, has an LD_{50} (oral, rat) of 1310 – 1600 mg/kg. Poisoning is characterized by narcosis, excitability followed by CNS depression, loss of coordination, paresis of the hind extremities, disturbances in breathing, hypotonia, reduced cardiac activity, and reduced muscle tone. Triethyl phosphate is slightly irritating to the mucous membranes. Skin irritation and sensitizing effects have not been observed. The inhibition of choline esterases is low in comparison to other phosphate esters. Following repeated administration, triethyl phosphate causes an increase in the weight of the liver and adrenal glands with associated hepatocellular hypertrophy as the histological correlation. Triethyl phosphate is mutagenic in drosophila (recessive lethal test, eye mosaic assay). However, Ames test, cell transformation studies, in vivo cytogenic studies, and dominant lethal tests have given no indication of genotoxicity. The alkylating action is much weaker than that of trimethyl phosphate [161], [179].

Tributyl phosphate blocks cholinesterase only weakly [162]. The major metabolites are dibutyl hydrogenphosphate, butyl dihydrogenphosphate, and butyl bis(3-hydroxybutyl) phosphate [163]. The oral LD_{50} in rats is between 1390 and 3350 mg/kg. Upon subchronic oral administration of tributyl phosphate, damage to the liver, kidneys and urinary bladder were observed [180]. Tributyl phosphate was not neurotoxic in rats after acute or subchronic oral exposure [181]. Tributyl phosphate was administered via the diet to rats for two years. Dose levels were equivalent to 9, 33, and 143 mg/kg per day for males and to 12, 42, and 182 mg/kg per day for females. An increased incidence of urinary bladder hyperplasia and papillomas was observed at the high dose group in males and females. The incidence of transitional cell carcinomas of the urinary bladder was increased in males [182]. Mice were treated in the same way with dietary doses equivalent to 29, 169, and 585 mg/kg per day (males) and 24, 206, and 711 mg/kg per day (females). An increased incidence of hepatocellular adenomas was observed in males of the high dose group [183]. The mechanism of the bladder tumorigenesis in rats is probably nongenotoxic. High dietary doses of tributyl phosphate, given for 10 weeks, produced reversible urothelial cytotoxicity with regenerative hyperplasia. This cytotoxicity is likely due to a direct effect of tributyl phosphate or its metabolites rather than an indirect consequence of crystalluria [184]. Tributyl phosphate was not genotoxic in several test systems. It is not teratogenic in rats or rabbits and did not impair fertility in rats [180], [185]. Irritation of eyes and respiratory tract were observed after inhalation exposure in humans [180]. The TLV-TWA value is 2.5 mg/m^3 [142].

2-Ethylhexyl diphenyl phosphate has only a very low acute toxicity (LD_{50}, oral, rat and rabbit, > 24 000 mg/kg). No indications of carcinogenic effects were found in rats after two years of administration of the substance in the feed (up to 1 %) [164]. 2-Ethylhexyl diphenyl phosphate was not genotoxic in the Ames test, the HGPRT test, and the in vivo chromosome aberration test in rats. No teratogenic effects were observed in rats, and development of the offspring was only impaired at maternally toxic doses [186].

Tris(2-chloroethyl) Phosphate. The toxicological data of tris(2-chloroethyl) phosphate, which is used as a flameproofing agent, are summarized in [147], [187-189]. LD_{50} values of 390 – 1410 mg/kg (and > 2000 mg/kg in one study) have been reported for rats following oral administration. The substance is nonirritating to very mildly irritating to the skin and mucous membranes. It was not sensitizing to the skin in guinea pigs. Subchronic administration of tris(2-chloroethyl) phosphate caused increased liver weight in rats and mice. Additionally, rats showed increased kidney weights and degenerative changes in the hippocampus region of the brain. Results of mutagenicity studies on tris(2-chloroethyl) phosphate are not consistent, but the majority of tests were negative. Results of carcinogenicity studies suggest that tris(2-chloroethyl) phosphate possesses a weak tumorigenic activity, possibly via an epigenetic mechanism. When pregnant rats were treated with tris(2-chloroethyl) phosphate, toxic effects were found in the mother, but no fetotoxic or teratogenic effects were observed. Impairment of reproduction was observed in a two-generation study in mice.

Tris(2,3-dibromopropyl) phosphate, which was mainly used as a flameproofing agent in synthetic fibers for children's nightdresses, possesses only a low acute toxicity (LD_{50}, oral, rat: 5240 mg/kg) [147]. Renal inflammation and atrophy of the testicles were observed following subchronic dermal application in rabbits [166]. The compound was found to be genotoxic in several test systems [147], [167-169]. It was carcinogenic in mice and rats following chronic administration in the feed [147].

Aromatic Phosphate Esters.
Cresyl Phosphates. Among the aromatic phosphate esters, tricresyl phosphates with one or more *o*-cresyl groups are of toxicological importance. They give a completely different clinical picture following poisoning in human beings and some animal species than that resulting from alkyl phosphate poisoning. Motor paralysis, initially of the feet, tibia and thighs, and later of the hands and arms, occurs one to three weeks after intake. The initial flaccid paralysis progresses to a spastic condition. In the past, tricresyl phosphates have caused numerous mass poisoning incidents, for example, as softeners for PVC and in contaminated edible oils and machine-gun oil. The largest incidence of poisoning occurred in 1929/1930 in the United States following consumption of ginger spirit which was adulterated with tricresyl phosphate (“ginger paralysis”) [171]. In the organism, tricresyl phosphate is metabolized to the active agent by hydroxylation of the α-C atoms in the *ortho* position, followed by ring closure. The active agent irreversibly inhibits a nonspecific esterase in the CNS (neuropathy target esterase, formerly called neurotoxic esterase, NTE). Other compounds producing this type of neurotoxicity, which is known as organophosphate-induced delayed polyneuropathy (OPIDP) or neuropathy (OPIDN), are related phosphates, phosphonates, and phosphoramidates. Primarily, the axons are damaged. The demyelinisation of the affected nerves, which is later visible on histological examination, should be regarded as a secondary effect [172], [190]. Target organs after subchronic administration of tricresyl phosphate in rats and mice were adrenals, kidneys, ovaries, and testes. Reduced grip strength occurred in mice. Tricresyl phosphate was not carcinogenic in rats and mice. It was not mutagenic in the Ames test and did

not induce chromosomal aberrations or sister chromatid exchanges [191]. Tri-*o*-cresyl phosphate was not teratogenic in rats [192]. It caused a dose-dependent decrease in sperm motility and sperm population in the epididymis in male rats [173], [174]. The TLV-TWA value has been set at 0.1 mg/m^3 [142].

Triphenyl phosphate itself is not neurotoxic. The effects of this type that are occasionally observed are attributed to impurities [190], [193]. Triphenyl phsophate has a low acute oral and dermal toxicity and is not irritating to the skin. Likewise, repeat-dose studies showed low toxicity. It did not affect reproduction when administered to rats during gametogenesis, through mating and gestation. No teratogenic effects were seen. Triphenyl phosphate was not genotoxic in several test systems [194]. The TLV-TWA value for triphenyl phosphate is 3 mg/m^3 [142].

Phosphites. Due to its tendency to hydrolyze, *trimethyl phosphite* is irritating to the skin, eyes, and respiratory tract. Studies on the acute oral toxicity gave LD_{50} values of 1600 – 2890 mg/kg in rats. Following single administration of high doses (in the LD_{50} range), flaccid paralysis of the extremities was observed. Local irritating properties of the substance are prominent when the substance is inhaled repeatedly. After repeated oral administration to rats, changes of the lungs, fatty degeneration of the liver, and reduced sperm production were observed. The results of mutagenicity tests are not consistent. Upon oral administration to pregnant rats, trimethyl phosphite causes an increase in fetal resorption rate as well as anomalies in the offspring [170]. The TLA-TWA value has been set at 10 mg/m^3 [142]. *Diethyl phosphite* (oral LD_{50} in rats 3500 – 5560 mg/kg) and *triethyl phosphite* (oral LD_{50} in rats 1840 – 4000 mg/kg) are mildly irritating to the skin in rabbits and induced skin sensitization in guinea pigs. These compounds were not mutagenic in the Ames test and micronucleus test. Lung damage was observed after repeated oral administration of diethyl phosphite [195], [196]. The following LD_{50} values have been reported for other aliphatic phosphites following oral administration to rats: dimethyl phosphite, 3800 mg/kg; tri-*n*-butyl phosphite, 3200 mg/kg; triisooctyl phosphite, 19 200 mg/kg [175].

The minimum lethal dose in rats following subcutaneous administration is ca. 2400 mg/kg for triphenyl phosphite, 3400 mg/kg for tri-*p*-cresyl phosphite, 3400 – 5700 mg/kg for tri-*m*-cresyl phosphite, and 11 300 mg/kg for tri-*o*-cresyl phosphite. Poisoning was characterized initially by mild, and then severe tremors, normally of the larger muscle groups. Hyperexcitability, spasticity, ataxia, and flaccid paresis of the extremities appear several days after administration. Tri-*o*-cresyl phosphite and triphenyl phosphite exhibited the highest activity. Upon oral administration to chickens, tri-*o*-cresyl phosphite causes neurotoxic effects similar to tri-*o*-cresyl phosphate. Probably, it is oxidized to the latter in the body [176]. In the case of triphenyl phosphite, which causes ataxia in rats, chickens, and cats, the pattern of injury caused differs from that described for the tricresyl phosphates [197].

[Top of Page]

References

1. G. M. Kosolapoff, L. Maier (eds.): *Organic Phosphorus Compounds*, Wiley-Interscience, New York 1972 – 1976.
2. *Houben-Weyl*, 4th ed., **XII/1, XII/2**.
3. *Houben-Weyl*, 4th ed., **E 1, E 2**.
4. J. Emsley, D. Hall: *The Chemistry of Phosphorus*, Harper & Row, London 1976.
5. D. E. C. Corbridge: *Phosphorus*, 4th ed., Elsevier, Amsterdam 1990.
6. J. I. G. Cadogan: *Organophosphorus Reagents in Organic Synthesis*, Academic Press, London 1979.
7. M. Grayson, E. J. Griffith: *Topics in Phosphorus Chemistry*, vol. **1 – 11**, Interscience Publishers, New York 1964 – 1983.
8. *Organophosphorus Chemistry*, vol. **1 – 20**, The Royal Society of Chemistry, London 1970 – 1989.
9. R. S. Edmundson: *Dictionary of Organophosphorus Compounds*, Chapman and Hall, London 1987.
10. G. M. Kosolapoff, L. Maier (eds.): *Organic Phosphorus Compounds*, vol. 1, Wiley-Interscience, New York 1972.
11. Hoechst, DE 2 703 802, 1977 (G. Elsner, G. Heymer, H.-W. Stephan).
12. M. C. Hoff, J. Hill, *J. Org. Chem.* **24** (1959) 356. Links
13. S. H. Li, C. A. Larsen, N. I. Buchan, G. B. Stringfellow, *Conf. Ser. Inst. Phys.* **96** (1989) no. 3, 153.
14. Deutsche Advance Produktion, DE 1 265 746, 1965 (H. J. Lorenz, A. R. Zintl, V. Franzen).
15. M&T Chemicals, DE 2 638 720, 1975 (V. M. Chopdekar).
16. BASF, DE 2 050 095, 1970 (H. Müller, A. Stü-binger).
17. M. M. Rauhut, H. A. Currier, A. M. Semsel, V. P. Wystrach, *J. Org. Chem.* **26** (1961) 5138. Links
18. Hoechst, US 4 324 919, 1980 (G. Elsner, H. Vollmer, E. Reutel).
19. Hoechst, EP 0 258 693, 1987 (N. Weferling, G. Elsner, H.-W. Stephan, F.-K. Frorath).
20. American Cyanamid Co, US 2 822 376, 1957; CA 1 151 212, 1981 (A. J. Robertson, J. C. Oppelt).
21. W. J. Vullo, *Ind. Eng. Chem. Prod. Res. Dev.* **5** (1966) 346.
22. W. Wolfsberger, *Chem. Ztg.* **115** (1991) 7.
23. H. Pommer, *Angew. Chem. Int. Ed. Engl.* **16** (1977) 423.
24. Rhône-Poulenc, FR 2 314 910, 1975 (E. G. Kuntz).
25. Ruhrchemie, EP 0 175 919, 1985 (L. Bexten, B. Cornils, D. Kupies).
26. Ciba-Geigy US 4 138 256, 1979 (C. Chylewski *et al.*).
27. Shell Int. Res., NL 6 604 094, 1965 (R. C. Morris, J. L. van Winkle, R. F. Mason).
28. Shell Oil, US 3 644 563, 1972 (R. S. Bauer *et al.*).
29. Mobay Co., GB 1 244 416, 1971 (R. L. Sandridge).
30. Smith-Kline Beckman, CH 634 581, 1978 (D. T. Hill, I. Lantos, B. M. Sutton).
31. Stauffer Chemical, DE 2 046 314, 1969 (E. H. Uhing).
32. Stauffer Chemical, EP 50 841, 1981 (A. E. Skrzec).
33. Mobil Oil, DE 2 456 095, 1975 (A. T. Jurewicz, W. W. Kaeding).
34. Toyama Chemical Industry, JP 44 003 354, 1969 (Y. Toyama, M. Nakabayashi, T. Uehara); *Chem. Abstr.*
35. Hoechst, DE 2 629 299, 1976 (K. Gehrmann, A. Ohorodnik, K. H. Steil, S. Schaefer).
36. Stauffer Chemical, US 4 409 152, 1981 (F. H. Lawrence).
37. Nissan Chemical Industries, JP 50 0 120 054, 1975 (M. Aramaki *et al.*).
38. Sandoz DE 2 152 481, 1972 (K. Hofer, G. Tscheulin).
39. Koppers, US 3 078 304, 1963 (H. Niebergall).
40. Hoechst, EP 93 418, 1983 (H.-J. Kleiner).
41. Hoechst, EP 93 419, 1983 (H.-J. Kleiner).
42. Hoechst, EP 93 420, 1983 (H.-J. Kleiner).
43. Hoechst, EP 110 305, 1983 (H.-J. Kleiner).
44. Cassella, DE 2 233 941, 1972 (C. Heid, D. Hoffmann, J. Polster).
45. K. Weissermel, H.-J. Kleiner, M. Finke, U.-H. Felcht, *Angew Chem.* **93** (1981) 256 – 266. Links
46. P. Beck in G. M. Kosolapoff, L. Maier (eds.): *Organic Phosphorus Compounds*, vol. **2**, Wiley-Interscience, New York 1972, pp. 189 – 508. E. V. Dehmlow, *Angew. Chem. Int. Ed. Engl.* **13** (1974) 170.
47. M. Grayson, P. T. Keough, *J. Am. Chem. Soc.* **82** (1960) 3919 – 3924. Links D. Landini, A. Maia, A. Rampoldini, *J. Org. Chem.* **51** (1986) 3187 – 3191. Links
48. Schering, EP 0 377 850, 1989 (H. Häuser).
49. S. Affandi, *et al.*, *Synth. React. Inorg. Met. Org. Chem.* **17** (1987) 307 – 317.
50. Dow Chem., EP 0 019 852, 1980 (G. A. Doorikian, J. L. Bertram).
51. I. Gosney, A. G. Rowley in [6]
52. H. H. Freedman, *Pure Appl. Chem.* **58** (1986) 857 –868. Links
53. Hoechst, EP 0 240 703, 1987 (H. Erpenbach, K. Gehrmann, W. Lork, P. Prinz).
54. Akzo, EP 0 078 567, 1982 (J. Brand).
55. Dow Corning, US 4 433 096, 1984 (G. N. Bokerman, N. R Langley).
56. Boots, US 4 287 374, 1980 (R. A. North).
57. Hoechst, EP 0 036 485, 1981 (W. Krause, J. Große, W. Klose).
58. Ciba-Geigy, US 4 874 526, 1989 (R. Grade, B. M. Thomas).
59. Ciba-Geigy, EP 0 105 843, 1983 (W. Wehner, R. Grade).
60. Mobil, US 3 268 323, 1966 (L. E. Goyette).
61. D. J. Daigle, A. W. Frank, *Text. Res. J.* **52** (1982) 751 – 755. Links
62. W. A. Rickelton, R. J. Boyle, *Sep. Sci. Technol.* **23** (1988) 1227; P. R. Danesi, *Solvent Extr. Ion Exch.* **3** (1985) 435.
63. F. J. Hurst, D. J. Crouse, US 3 711 591.
64. R. Marr et al., EP 9 106 118 A1, 1983.

65. *Phosphorus Potassium* **111** (1981) 31 (part I); **112** (1981) 31 (part II).
66. R. R. Grinstead, DE 2 423 272, 1974/1983.
67. CTP, EP 0 038 317, 1981 (W. Kanzler, J. Schedler).
68. American Cyanamid, EP 0 132 700, 1984 (A. J. Robertson, W. A. Rickelton).
69. Hoechst, EP 0 259 583, 1987 (N. Weferling *et al.*).
70. Mitsui Cyanamid KK, JP 19 3 846, 1985.
71. ICI, DE OS 2 614 323, 1976 (A. Ibbotson).
72. The Upjohn, DE-OS 2 801 701, 1978.
73. BASF, DE 3 927 146, 1989 (W. Franzischka *et al.*).
74. Degussa, DE 3 740 899, 1987 (M. Dankowski, P. Nagler, T. Lieser).
75. FMC Corp., EP 0 095 453, 1983 (D. P. Braksmeyer).
76. Hoechst, EP 0 196 026, 1986 (H. J. Löher, K. Bauer, H. Bieringer).
77. Ciba-Geigy, EP 0 413 657, 1990 (W. Rutsch, K. Dietliker, R. G. Hall).
78. BASF, US 4 710 523, 1987 (P. Lechtgen, I. Buethe, F. Jacobi, W. Trimborn).
79. Espe, EP 0 184 095, 1985 (K. Ellrich, C. Herzig).
80. Sandoz, DE 2 152 481, 1972 (K. Hofer, G. Tscheulin).
81. Hoechst, EP 374 761, 1989 (M. Böhshar, H.-J. Kleiner, K. Waldmann, G. Pfahler).
82. Cyanamid Canada, GB 2 068 381, 1981 (A. J. Robertson, T. Ozog).
83. R. Engel in A. S. Kende (ed.): *Organic Reactions*, vol. **36**, J. Wiley & Sons, New York 1988, pp. 175 –248.
84. K. Weissermel, H.-J. Kleiner, M. Finke, U.-H. Felcht, *Angew. Chem.* **93** (1981) 256 – 266. Links
85. Ciba-Geigy, DE 2 805 074, 1978 (L. Maier).
86. V. K. Khairullin, T. I. Sob čuk, A. N. Pudovik, *Zh. Obshch. Khim.* **37** (1967) 710.
87. V. K. Khairullin, R. M. Kondrat'eva, A. N. Pudovik, *Zh. Oshch. Khim.* **38** (1968) 288.
88. Hoechst, DE 2 529 731, 1975 (E. Lohmar, K. Gehrmann, A. Ohorodnik, P. Stutzke).
89. Hoechst, EP 159 656, 1984 (H.-J. Kleiner).
90. Hoechst, CH 620 812, 1976 (W. Rupp *et al.*).
91. P. Langelueddeke *et al.*, *Meded. Fac. Landbouwwet. Rijksuniv. Gent* **47** (1982) 95.
92. Chemische Fabrik Budenheim Rudolf A. Oetker, DE 3 616 168, 1986 (K. Sommer, K. Dorn, G. Scheuer, K. Götzmann).
93. W. A. Rickelton, R. J. Boyle, *Sep. Sci. Technol.* **23** (1988) 1227 – 1250.
94. Hoechst, DE 2 346 787, 1975 (H.-J. Kleiner, M. Finke, U. Bollert, W. Herwig).
95. Hoechst, DE 2 454 189, 1976 (U. Bollert, A. Ohorodnik, E. Lohmar).
96. Hoechst, DE-OS 1 668 031, 1967 (K. Schimmelschmidt, H.-J. Kleiner).
97. C. Fest, K.-J. Schmidt: *The Chemistry of Organophosphorus Pesticides*, 2nd ed., Springer-Verlag, Berlin 1982, pp. 68 – 80.
98. R. Gächter, H. Müller (eds.): *Plastics Additives Handbook*, Hanser Verlag, München 1989.
99. T. J. Henman: *World Index of Polyolefin Stabilizers*, Kogan Page, London 1982.
100. Philagro, DE-OS 2 751 035, 1977 (J. Abblard, M. Decines, R. Viricel).
101. Philagro, DE-OS 2 911 516, 1979 (A. Bernard, A. Disdier, M. Royer).
102. Hoechst, EP 49 343, 1981 (J. Große).
103. Hoechst, DE 3 131 249, 1981 (W. Krause).
104. Ethyl Corporation, DE 1 793 376, 1972 (J. B. Hinkamp, V. F. Hnizda).
105. Ethyl Corporation, US 3 879 454, 1975 (J. B. Hinkamp, V. F. Hnizda).
106. B. A. Arbusov, *Pure Appl. Chem.* **9** (1964) 307.
107. A. K. Bhattacharya, G. Thyagarajan, *Chem. Rev.* **81** (1981) 415. Links
108. Bayer, EP 358 022, 1989 (R. Kleinstück *et al.*).
109. B. Blaser, K.-H. Worms, H.-G. Germscheid, K. Wollmann, *Z. Anorg. Allg. Chem.* **381** (1971) 247. Links
110. W. Plöger, N. Schindler, K. Wollmann, K.-H. Worms, *Z. Anorg. Allg. Chem.* **389** (1972) 119. Links
111. Hoechst, DE 3 047 107, 1980 (W. Klose, T. Auel).
112. Henkel, EP 332 068, 1989 (H. Blum, S. Hemmann).
113. H. Blum, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **43 B** (1988) 75.
114. Ciba-Geigy, DE 2 903 928, 1979 (H. Nachbur, P. Rohringer).
115. Ciba-Geigy, EP 157 731, 1985 (H. O. Wirth, K. Müller).
116. Hoechst, EP 77 460, 1980 (H.-J. Kleiner, W. Dursch).
117. Hoechst, EP 281 122, 1987 (G. Roscher, H.-J. Kleiner, G. Ihl, H. Leipe).
118. Kalle, US 3 468 725, 1969 (F. Uhlig).
119. General Electric, EP 321 002, 1989 (C. A. A. Claessen).
120. Ciba-Geigy, EP 310 559, 1989 (C. D. Weis, P. Sutter).
121. Daihachi Chemical Industry, US 4 196 076, 1980 (A. Fujimoto, I. Miura, K. Noguchi).
122. Akzo, DE 2 708 790, 1977 (J. Kowallik, A. Brandner).
123. Ciba-Geigy, DE 2 013 665, 1970 (H. Nachbur, J. Kern, A. Maeder).
124. Ciba-Geigy, DE 2 147 481, 1973 (P. Hofmann, H. Nachbur, A. Maeder).
125. R. Aenishänslin *et al.*, *Text. Res. J.* **39** (1969) 375.
126. Nissan Chemical Industries, DE 2 527 650, 1974 (Y. Ura, H. Takamatsu, C. Funabashi).
127. R. G. Harvey, E. R. De Sombre in M. Grayson, E. J. Griffith (eds.): *Topics in Phosphorus Chemistry*, vol. **1**. Wiley-Interscience, New York 1964, p. 57.
128. G. Bohlmann: "Flame Retardant Chemicals," *Process Economics Programm*, Supplement A, SRI International, Menlo Park, Ca.
129. Albright & Wilson, Mobil Oil, EP 33 999, 1980 (R. P. Napier, T. N. Williams Jr., B. E. Johnston, J. M. Horn).
130. G. Côte, D. Bauer, *Anal. Chem.* **56** (1984) 2153.
131. Produits Chimiques Ugine Kuhlmann, FR 7 220 487, 1972 (M. Démarcq).
132. Hoechst, EP-A 36 485, 1981 (W. Krause, J. Große, W. Klose).
133. ICI America, EP-A 285 073, 1988 (S. B. Mirviss).
134. Hoechst, EP 85 344, 1983 (R. Gradl, G. Schimmel, W. Krause, G. Heymer).
135. E. Jdid, J. Bessière, P. Blazy, *Ind. Miner. Tech.* 1984, 389.
136. H. v. Plessen, G. Schimmel, *Chem. Ing. Tech.* **59** (1987) 772.
137. Hoechst, US 4 713 229, 1985 (G. Schimmel, R. Gradl).
138. G. E. Russell: *Process Economics Program*, Report No. 113, Stanford Research Institute, Menlo Park, Ca. 1977, p. 1, 75.
139. W. B. Deichmann, H. W. Gerarde: *Toxicology of Drugs and Chemicals*, Academic Press, New York 1969, pp. 600.
140. R. S. Waritz, R. M. Brown, *Am. Ind. Hyg. Assoc. J.* **36** (1975) 452 – 458. Links
141. R. M. Brown, R. S. Waritz, *Am. Ind. Hyg. Assoc. J.* **31** (1970) no. 2, 47.
142. American Conference of Governmental Industrial Hygienists: TLVs and BEIs. *Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices*, Cincinatti, Ohio, 2005
143. The Registry of Toxic Effects of Chemical Substances (1997).
144. E. Zeiger, B. Anderson, S. Haworth, T. Lawlor, K. Mortelmans, W. Speck, *Environ. Mutagen*. **9** (1987) 1–110. Links
145. *Ullmann*, 5th ed., **A19**, 567.
146. J. Kimmig, D. Jerchel, *Klin. Wochenschr.* **28** (1950) 429 – 431. Links
147. IARC: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man*, vol. 71, Lyon 1999.
148. IARC: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man*, vol. 9, Lyon 1975.
149. IARC: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man*, vol. 15, Lyon 1977.
150. K. P. Lee, H. J. Trochimowicz, *Toxicol. Appl. Pharmacol.* **62** (1982) 90 – 103. Links

151. K. P. Lee, H. J. Trochimowicz, *JNCI J. Natl. Cancer Inst.* **68** (1982) no. 1, 157 – 172.
152. W. M. B. Deichmann, S. Witherup, *J. Pharmacol. Exp. Ther.* **88** (1946) 338 – 342. Links
153. U. Schaeppi, G. Krinke, W. Kobel, *Neurobehav. Toxicol. Teratol.* **6** (1984) 39 – 50. Links
154. NCI: "Bioassay of Trimethylphosphate for Possible Carcinogenicity," *DHEW Publ. (NIH) (U.S.)*, NCI-CG-TR-81, 1978.
155. T. H. Connor, *Mutat. Res.* **65** (1979) 121 – 131. Links
156. J. Van't Hof, L. A. Schairer, *Mutat. Res.* **99** (1982) 303 – 315. Links
157. J. F. Sina *et al.*, *Mutat. Res.* **113** (1983) 357 – 391. Links
158. N. Degraeve, M. C. Chollet, J. Moutschen, *J. Toxicol. Clin. Exp.* **6** (1986) 5 – 11. Links
159. U. Graf *et al.*, *Mutat. Res.* **222** (1989) 359 – 373. Links
160. R. D. Harbison, C. Dwivedi, M. A. Evans, *Toxicol. Appl. Pharmacol.* **35** (1976) 481 – 490. Links
161. Beratergremium für umweltrelevante Altstoffe(BUA) der GDCh: *Triethylphosphat (Phosphorsäure-tri-ethylester)*, VCH Verlagsgesellschaft, Weinheim, Germany 1989.
162. J. C. Sabine, F. N. Hayes, *AMA Arch. Ind. Hyg. Occup. Med.* **6** (1952) 174 – 177.
163. T. Suzuki, K. Sasaki, M. Takeda, M. Uchiyama, *J. Agric. Food Chem.* **32** (1984) 603 – 610. Links
164. J. F. Treon, F. R. Dutra, F. P. Cleveland, *AMA Arch. Ind. Hyg. Occup. Med.* **8** (1953) 170 – 184.
165. IARC: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man*, vol. 20, Lyon 1979.
166. R. E. Osterberg, G. W. Bierbower, R. M. Hehir, *J. Toxicol. Environ. Health* **3** (1977) 979 – 987. Links
167. A. Blum, B. N. Ames, *Science* **195** (1977) 17 – 23. Links
168. M. J. Prival, E. C. McCoy, B. Gutter, H. S. Rosenkranz, *Science* **195** (1977) 76 – 78. Links
169. E. Zeiger, D. A. Pagano, A. A. Nomeir, *Environ. Mutagen.* **4** (1982) 271 – 277. Links
170. D. Henschler: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten*, VCH Verlagsgesellschaft, Weinheim, Germany 1995.
171. W. Forth, D. Henschler, W. Rummel: *Allgemeine und spezielle Pharmakologie und Toxikologie*, Bibliographisches Institut & F. A. Brockhaus AG, Mannheim 1987.
172. M. K. Johnson, *CRC Crit. Rev. Toxicol.* **4** (1975) 289 – 316.
173. S. G. Somkuti *et al.*, *Toxicol. Appl. Pharmacol.* **89** (1987) 49 – 63. Links
174. S. G. Somkuti *et al.*, *Toxicol. Appl. Pharmacol.* **89** (1987) 64 – 72. Links
175. L. Levin, K. L. Gabriel, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 286 – 291. Links
176. M. I. Smith, R. D. Lillie, E. Elvove, E. F. Stohlmann, *J. Pharmacol. Exp. Ther.* **49** (1933) 78 – 99. Links
177. S. S. Padilla, T. B. Grizzle, D. Lyerly, *Toxicol. Appl. Pharmacol.* **87** (1987) 249 – 256. Links
178. IARC: *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Man*, vol. 50, Lyon 1990.
179. E. W. Vogel, M. J. M. Nivard, *Mutagenesis* **8** (1993) 57 – 81. Links
180. Berufsgenossenschaft der chemischen Industrie: *Toxicological evaluations. No. 170: Tributylphosphat*, Springer Verlag, Berlin 1993.
181. C. E. Healy, P. C. Beyrouty, B. R. Broxup, *Am. Ind. Hyg. Assoc. J.* **56** (1995) 349 – 355. Links
182. C. S. Auletta, M. L. Weiner, W. R. Richter, *Toxicology* **128** (1998) 125–134. Links
183. C. S. Auletta, L. A. Kotkoskie, T. Saulog, W. R. Richter, *Toxicology* **128** (1998) 135–141. Links
184. L. L. Arnold, W. R. Christenson, M. Cano, M. K. St. John, B. S. Wahle, S. M. Cohen, *Fundam Appl Toxicol*. **40** (1997) 247–255. Links
185. T. Noda, T. Yamano, M. Shimizu, S. Morita, *Fd. Chem. Toxic.* **32** (1994) 1031 – 1036.
186. Berufsgenossenschaft der chemischen Industrie: *Toxicological Evaluations, no. 194: Diphenyl-2-ethylhexylphosphat*, Springer Verlag, Berlin 1995.
187. Beratergremium für umweltrelevante Altstoffe (BUA) der GDCh: *Tris(2-chlorethyl)phosphat*. VCH Verlagsgesellschaft, Weinheim 1987.
188. Berufsgenossenschaft der chemischen Industrie: Toxicological evaluations. no. 33: *Tris(2-chlor)-ethyl phosphat*. Springer Verlag, Berlin 1995.
189. World Health Organisation, *Environmental Health Criteria* 209 (1998).
190. M. Lotti, *CRC Crit. Rev.* **4** (1992) 465 – 487.
191. *National Toxicology Program: NTP Technical Report Series No.* 433 (1994).
192. D. R. Tocco, J. L. Randall, R. G. York, M. K. Smith, *Fundam. Appl. Toxicol*. **8** (1987) 291–297. Links
193. A. Nakamura, *Environ. Health Criter.* **111** (1991) 1 – 80.
194. *World Health Organisation*, Environmental Health Criteria 111 (1991).
195. Berufsgenossenschaft der chemischen Industrie: *Toxicological Evaluations, no. 192: Triethylphosphit*, Springer Verlag, Berlin 1995.
196. Berufsgenossenschaft der chemischen Industrie: *Toxicological Evaluations, no. 193: Diethylphosphit*, Springer Verlag, Berlin 1995.
197. E. J. Lehning, D. Tanaka Jr., S. J. Bursian, *Fundam. Appl. Toxicol*. **29** (1996) 110–118. Links

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Photochemistry

Karl-Heinz Pfoertner, Basel, Switzerland

1. Introduction . . . 1
2. Primary Photophysical Processes . . . 2
3. Performance of Photochemical Reactions . . . 5
3.1. Photoreactor Design . . . 5
3.2. Light Sources . . . 8
3.3. Materials . . . 10
3.4. Scaleup . . . 13
3.5. Economic Considerations . . . 13
4. Light-Induced Chain Reactions . . . 15
4.1. Photohalogenation . . . 15
4.2. Photosulfochlorination . . . 17
4.3. Photosulfoxidation . . . 17
4.4. Type I Photooxidation . . . 18
4.5. Photopolymerization . . . 18
5. Quasi-Stoichiometric Photoreactions 19
5.1. Photonitrosation . . . 19
5.2. Photosynthesis of Previtamin D . . . 20
5.3. Photoreactions in Supercritical Fluids 21
6. Photosensitized Reactions . . . 22
6.1. (*E/Z*)-Isomerization of Tachysterol . . . 22
6.2. Type II Photooxidation . . . 23
7. Selected Useful Photoreactions . . . 26
7.1. Enamide Cyclization . . . 26
7.2. Cyclobutane Formation . . . 27
7.3. Oxetane Formation . . . 27
7.4. Oxadi-π-Methane Rearrangement . . . 28
7.5. 1,3-Dipolar Cycloaddition . . . 28
8. Photochemical Wastewater Purification . . . 28
9. References . . . 30

1. Introduction

Photochemistry comprises the chemical reactions of atoms or molecules that have been electronically excited by absorption of light with wavelength in the range of 200 nm to ca. 700 nm. In thermal reactions heat is transferred to the respective ground state species, generating a Maxwell – Boltzmann energy distribution. In contrast, photochemical excitation proceeds via the absorption of discrete quanta of energy. The advantages of photochemistry, in particular concerning one-quantum or quasi-stoichiometric photochemistry, can be summarized as follows:

1) Light may be considered as a highly specific and ecologically clean "reagent"
2) Photochemical methods offer less "aggressive" routes in chemical synthesis than thermal ones
3) Photochemical reactions may contribute to the safety of industrial processes, because they are generally performed at, or below, room temperature
4) Photochemically reacting organic compounds do not require any protecting groups
5) Many conventional syntheses can be shortened by inserting photochemical reaction steps

About 8000 preparative organic photoreactions, classified in ca. 550 reaction types, are known [1] which may be economically feasible if they lead in one step to products of unusual structure, hereby shortening conventional syntheses considerably [2]. In contrast to this fact, industrial photochemistry has been ill-famed for many decades, and successful processes and production units are kept out of publicity because the developed know-how is of considerable value. Thus, it is quite difficult to collect and review data on photochemical technology and to discuss the evolution in this field. Nevertheless, in this article, the problems which always accompany the scaleup of photoreactions are treated as well as the relevant basic photochemical knowledge.

2. Primary Photophysical Processes

The energy E (in joules) of a single photon is given by the Planck equation

$$E = h\nu = \frac{hc}{\lambda} \quad (1)$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a19_573

where h is Planck's constant (6.6265×10^{-34} J s), ν the frequency of radiation (s^{-1}), c the velocity of light (2.9979×10^{8} m/s), and λ the wavelength (m). One mole of photons is also defined as one einstein. Usually the energies transferred in photochemical reactions exceed the typical activation energies of thermal reactions and also the dissociation energies of chemical bonds (Table 1).

When a system is irradiated, the light may be transmitted, scattered, refracted, or absorbed. The first law of photochemistry, formulated by Grotthus (1817) and Draper (1843), states that only light which is absorbed by a molecule is effective in producing a reaction which changes the molecule. This implies that there must be some overlap between the range of wavelengths of the light entering the system and those absorbed by the starting material. Therefore, the absorption spectrum of the latter and the spectral energy distribution of the light source should be known. It is also advantageous to know the wavelength-dependent transparency of the wall in front of the light source and the absorption spectra of the solvents and photoproducts.

The probability of absorbing a photon depends on the capability of the incident radiation to transform the electronic distribution of the molecule's ground state into that necessary for the excited state. According to the Beer–Lambert law, the absorbance A of a beam of collimated monochromatic radiation in a homogeneous isotropic medium is proportional to the absorption path length l (cm) and to the concentration c (mol/L), or–in the gas phase–to the partial pressure of the absorbing species. The law can be expressed as

$$A = \log(P_o/P) = \varepsilon c l \text{ or } P = P_o 10^{-\varepsilon c l} \tag{2}$$

where P_o is the incident radiant power, and P the power of the penetrating light; the proportionality constant ε ($L\,mol^{-1}\,cm^{-1}$) is called the molar decadic absorption coefficient.

The spectral radiant power P_{abs} (W/nm) [3] of a collimated beam absorbed along the path length l is given by the difference of the incident and penetrating radiant power:

$$P_{abs} = P_o - P \text{ or } P_{abs} = P_o\left(1 - 10^{-\varepsilon c l}\right) \tag{3}$$

Frequently, the incident radiant power P_o cannot be measured exactly. In many cases it is, however, sufficient to express P_o as 100 %, replacing Equation (3) by

$$P'_{abs} = 100\left(1 - 10^{-\varepsilon c l}\right)\% \tag{3a}$$

Furthermore, in industry it is common practice to characterize a light-absorbing substance with unknown molecular mass M_r by the absorbance $A^{1\,\%}_{1\,cm}$ of a 1 wt % solution with a 1 cm path-length. This term is related to ε by the equation

$$A^{1}_{1} = \frac{\varepsilon 10}{M_r} \tag{4}$$

When more than one light-absorbing component is present at the beginning of a photoreaction or developed in its course, the concentration c_i of any light-absorbing component i present must be considered as well as its molar decadic absorption coefficient ε_i. In this case Equation (2) must be expanded to [4]:

$$P_{abs,i} = P_o \frac{\varepsilon_i c_i \left(1 - 10^{-l \sum_{n=1}^{n} \varepsilon_n c_n}\right)}{\sum_{n=1}^{m} \varepsilon_n c_n} \tag{5}$$

The exponential term becomes negligible when $10^{-l \sum_{n=1}^{m} \varepsilon_n c_n} << 1$, to give

$$P_{abs,i} = P_o \frac{\varepsilon_i c_i}{\sum_{n=1}^{m} \varepsilon_n c_n} \tag{5a}$$

which can be simplified further by again expressing P_o as 100 %.

If the radiant power P is of the order of magnitude which is common in conventional photochemical systems, the concentration of electronically excited molecules is very low and the probability that an excited molecule absorbs a second photon is also low. In laser photochemistry, however, the concentration of excited molecules may increase to an extent which renders biphotonic processes possible.

A universal integral quantum yield φ defines the number of events which occur per photon absorbed by the system [3]

$$\varphi = \frac{\text{number of events}}{\text{number of photons absorbed}} \tag{6}$$

In this generalized form, φ can be used to describe photoexcitation and physical deactivation processes of a typical organic molecule, as

Table 1. Photon energies at different wavelengths compared with dissociation energies of selected single bonds

Wavelength, nm	Energy, kJ/einstein	Single bond	Dissociation energy, kJ/mol
200	598.2	HO – H	498
250	478.6	H – Cl	432
300	398.8	H – Br	366
350	341.8	Ph – Br	332
400	299.1	H – I	299
450	265.9	Cl – Cl	240
500	239.3	Me – I	235
550	217.5	HO – OH	213
600	199.4	Br – Br	193
650	184.1	$Me_2N - NMe_2$	180
700	170.9	I – I	151

shown in Figure 1. These physical deactivation processes are summarized in the second law of photochemistry

$$\varphi_{IC} + \varphi_{ISC(S)} + \varphi_{ISC(T)} + \varphi_F + \varphi_P = 1 \quad (7)$$

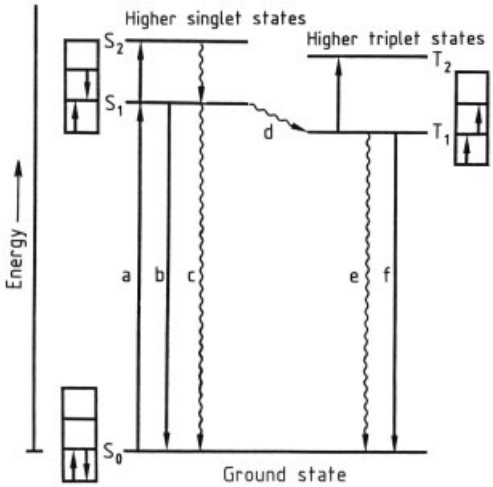

Figure 1. Modified Jablonski diagram for a typical organic molecule. Vibrational levels are omitted for clearness.
The transitions a to f are defined as follows (the lifetime in s is given in parentheses):
a) Excitation $S_o + h\nu \rightarrow S_1$ (10^{-15}); b) Internal conversion (IC) $S_1 \rightarrow S_o + \Delta$ ($10^{-9} - 10^{-7}$); c) Fluorescent emission (F) $S_1 \rightarrow S_o + h\nu_F$ ($10^{-11} - 10^{-6}$); d) Intersystem crossing [ISC(S)] $S_1 \rightarrow T_1 + \Delta$ ($10^{-11} - 10^{-8}$); e) Phosphorescent emission (P) $T_1 \rightarrow S_o + h\nu_P$ ($10^{-3} - 10^{+2}$); f) Intersystem crossing [ISC(T)] $T_1 \rightarrow S_o + \Delta$ ($10^{-3} - 10^{+2}$)

At room temperature excitation starts from the lowest vibrational level of the S_o state leading to vibrational sublevels of the S_1 state. Electron spins, characterized by arrows in boxes in Figure 1, are antiparallel to its "partner" in singlet states S and parallel in triplet states T. According to Hund's rule, triplet excited states are lower in energy than the corresponding singlet excited states. Therefore, after very rapid vibrational equilibration to the zeroth vibrational level of the excited states, the electronically excited molecule can partly redistribute its energy by radiationless intersystem crossing [ISC(S)] to a lower lying triplet state T_1. In most organic molecules the only strong absorptions and emissions are singlet – singlet in character, the transitions

$$S_o \underset{-h\nu}{\overset{h\nu}{\rightleftharpoons}} T_1 \quad (8)$$

are prohibited [5]. T_1 therefore has a much longer lifetime than S_1 (Fig. 1), hence the probability that photochemical reactions (including sensitized ones) occur from T_1 is much higher. Photosensitized reactions are those in which light is not absorbed by the substrate which undergoes the chemical reaction, but by an added sensitizer. The sensitizer undergoes no chemical change, but merely transfers electronic energy to the substrate or undergoes a reversible redox process with the substrate.

According to the Wigner spin conservation rule the energy of an electronically excited molecule can only be transferred to a ground-state molecule by the following processes:
Singlet energy transfer

$$^aS_1\,(\uparrow\downarrow)* + {}^bS_o\,(\uparrow\downarrow) \longrightarrow {}^aS_o\,(\uparrow\downarrow) + {}^bS_1\,(\uparrow\downarrow)* \quad (9)$$

Triplet energy transfer

$$^aT_1\,(\uparrow\uparrow)* + {}^bS_o\,(\uparrow\downarrow) \longrightarrow {}^aS_o\,(\uparrow\downarrow) + {}^bT_1\,(\uparrow\uparrow)* \quad (10)$$

where a and b denote different molecules. A third form of energy transfer is electron transfer,

a true photochemical reaction which can occur from the singlet as well as from the triplet state

$$D_1 + A_o \longrightarrow (D^{\cdot +} A^{\cdot -}) \quad (11)$$
$$D_3 + A_o \longrightarrow (D^{\cdot +} A^{\cdot -}) \quad (12)$$

D_1 and D_3 are donors and A_o is the acceptor.

To complete the description of primary photophysical processes, two types of electron transitions which are related to basic chemical structures can be distinguished (Fig. 2). In unsaturated compounds with carbon – carbon multiple bonds, light absorption produces $\pi \rightarrow \pi*$ transitions, whereas excitation of compounds containing heteroatoms such as N,O,S in general produce $n \rightarrow \pi*$ transitions. The electronic transition $(\pi, \pi*)$ may be described approximately as the promotion of an electron from a "bonding" π orbital to an "antibonding" π orbital, designated $\pi*$. In contrast, the $n \rightarrow \pi*$ transition is an electronic transition described approximately as the promotion of an electron from a "nonbonding" (lone pair) n orbital to an "antibonding" π orbital, also designated $\pi*$.

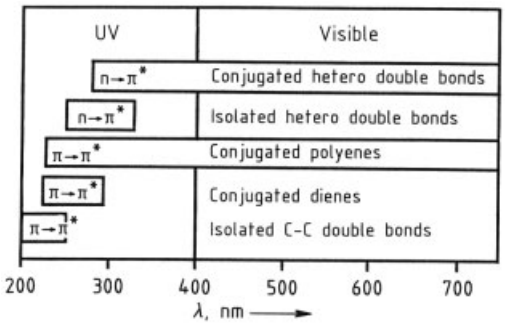

Figure 2. Absorption regions associated with the different types of electron transitions

These light-induced electron transitions exhibit different properties. The first $(\pi, \pi*)$ absorption band of aromatic hydrocarbons is more intense than the $(n, \pi*)$ transitions of aromatic carbonyl compounds. The $(n, \pi*)$ states generally have longer lifetimes than $(\pi, \pi*)$ states. The energy differences $\Delta E\,(S_1 - T_1)$ between the lowest excited $(\pi, \pi*)$ singlet and triplet states are much larger than those of the $(n, \pi*)$ states, e.g., in carbonyl systems. The variation in molecular geometry of the various excited states derived from $(\pi, \pi*)$ and $(n, \pi*)$ transitions determine their photochemical behavior.

Ketones are often used as sensitizers because their lowest excited singlet and triplet states are favorably situated between the singlet and the triplet state of many acceptors. However, the triplet state of ketones may bring about different processes, e.g., radical reactions started by hydrogen abstraction. Sensitization has advantages over direct irradiation because it enables irradiation at convenient wavelengths; several simple chromophores absorb at very short wavelengths at which mercury lamps, which are commonly used, are ineffective. Sensitizers in contrast absorb at the wavelengths emitted by mercury lamps.

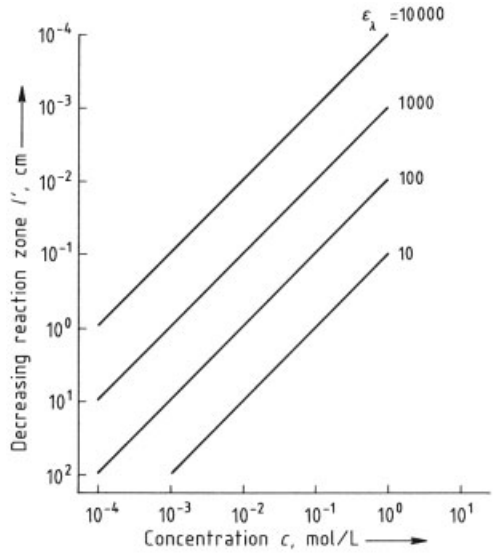

Figure 3. Thickness of the reaction zone l' (90 % absorption) as a function of the concentration c for different values of ε_λ

3. Performance of Photochemical Reactions

3.1. Photoreactor Design

According to the Beer – Lambert law photoreactions are characterized by their spatial inhomogeneity. For a reaction zone l' defined by an absorbance $A = 1$ (90 % light absorption) the term $1/\varepsilon c$ (cm) determines its thickness (Fig. 3). High molar absorption coefficients ε (L mol^{-1} cm^{-1})

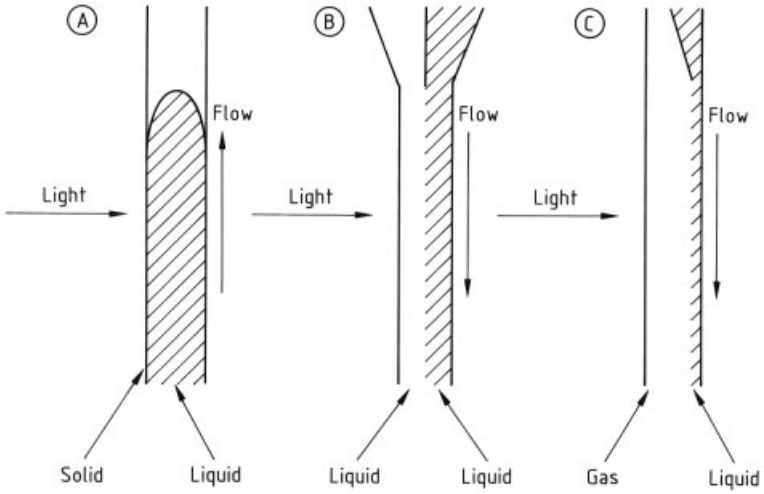

Figure 4. Fundamental arrangements of the most important photoreactor types [6] (reproduced with permission)
For description of A, B, and C see text

and/or high substrate concentrations c (mol/L) may reduce the thickness of l' to several micrometers, i.e., to a region where even the turbulent flow of a liquid becomes laminar due to adhesive forces from the wall. This results in a decreasing contribution of macroscopic homogenization processes to the exchange of reactants out of and into the reaction zone. It is therefore not always favorable to use light which corresponds to the substrate's absorption maximum. Irradiation of less absorbing parts of the chosen absorption band may extend the reaction zone l' significantly [6].

In industrial photochemistry, high substrate concentrations are preferred because of the cost of solvent losses and recycling. The disadvantages that arise from small reaction zones l' must be compensated for by an adequate photoreactor design. There are three technically suitable types of reactor which are outlined in Figure 4. In Figure 4A the photoreaction zone l' is located in the solid – liquid contact region where $1/\varepsilon\, c << 1$ cm, the zone l' can only be removed from the wall by a flow rate of twice the velocity of sound. To overcome this difficulty, in this case different methods may be applied, e.g., using the turbulent flow of the irradiated liquid by tangential injection into the reactor (Fig. 5) [7]. Another very effective approach is the gas-bubble photoreactor Fig. 6) [8]. Here, the solution flows fast in upward direction and entrains many small gas bubbles. These bubbles avoid filming and/or destructive overirradiation by producing local turbulences while passing the solution along the light permeable reactor wall (Fig. 6). This system is also useful for the photosensitized production of singlet oxygen (see Section 6.2) and for photochlorinations (see Section 4.1). It also has to be considered that solvent vapor taken along with the gas has to be condensed when leaving the photoreactor. In Figure 4B the reaction zone l' is located in the liquid – liquid contact region, provided that both liquids flowing downwards are nonmiscible and that the photoreactive light penetrates the liquid in front of the light source, whereas the light-absorbing substrate should only be soluble in the liquid that is more remote from the light source [9]. In this reactor type the two liquids must flow downwards. The proportions of the following parameters must be defined:

1) Density and viscosity of both liquids
2) Slit width which limits the space available for the two nonmiscible liquids flowing between the two concentric tubes which surround the tube-like lamp
3) The flow rates of both liquids
4) The pressure gradient

The literature only deals with one liquid flowing between two concentric tubes [10]. The flow of two nonmiscible liquids can also be treated mathematically, assuming that the concentric tubes are infinite in length and that the flow is

laminar. In this case combined differential equations can be formed, taking into account the gravitational, frictional, and compressive forces of the two liquids. The relevant limiting conditions are that the flow rate on the walls is zero, and that the flow rate and shear stress in the separation layer between the two liquids are of the same magnitude. The resulting system of differential equations can then be resolved by iteration methods using a computer.

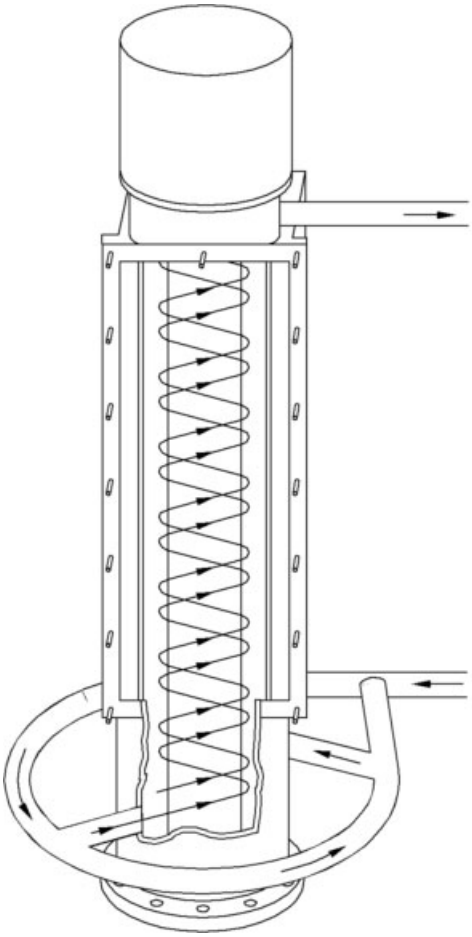

Figure 5. Example of a type 4A photoreactor: Free hanging UV source inside an open quartz glass tube; stainless steel jacket; injection of the liquid to produce a tangential flow; removable front plate. (VitaTec UV-Systeme GmbH) [7]

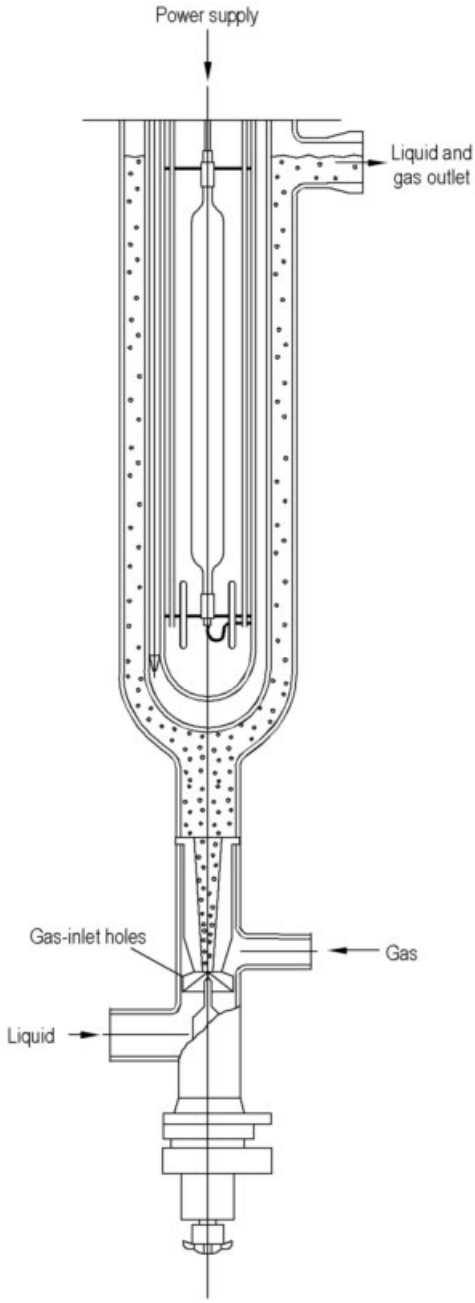

Figure 6. Example of a modified type A photoreactor: The QVF – Pfoertner photoreactor, creating local turbulences by small gas bubbles in the upwardly fast flowing solution [8] (QVF Glastechnik GmbH)

The structural features of the reactor shown in Figure 7 allow circulation of the liquids by pumping just the heavier phase. The less dense liquid automatically rises in the tubes outside the reactor and is forced to move down inside the reactor by friction with the falling heavier phase. It is possible to vary the motion of the

contact layer, i.e., of the photoreaction zone l', by altering the flow rate within limits [9].

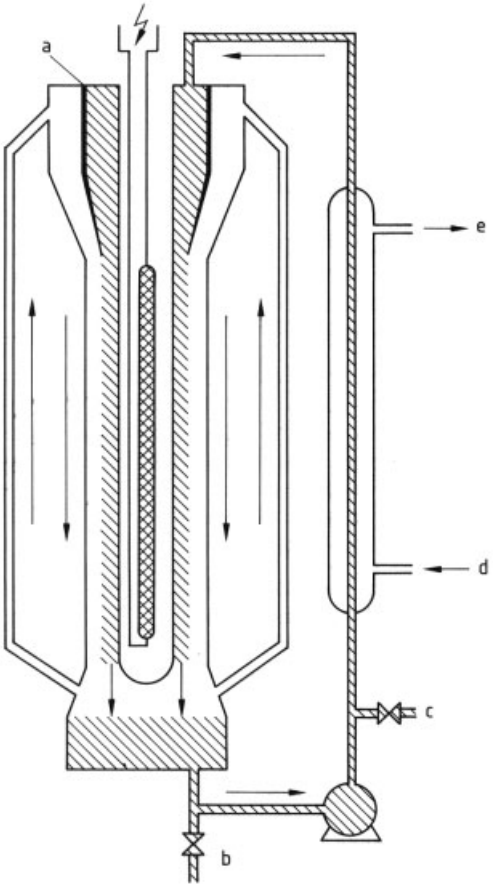

Figure 7. Example for a complete irradiation system containing a photoreactor of type 4 B
a) Partition wall; b) Outlet; c) Inlet; d) Heat exchanger inlet; e) Heat exchanger outlet

In Figure 4C the reaction zone l' is located in the gas – liquid contact zone, i.e., the surface of a falling film containing the substrate is irradiated. This can be realized with a variety of reactor types. One example is given in Figure 8, where the falling film is formed at the top of the reactor. Any splashes occurring on the lamp's housing are washed with the same solvent used in the irradiated solution which is thus continuously diluted [11]. Figure 9 shows a very special type of photoreactor. In this UV autoclave, photoreactions can be performed under pressures up to 15 MPa; mixing of the irradiated system is achieved by a magnetic stirrer [12]. Problems connected with photoreactor design have been intensively studied by J. C. André and coworkers [13].

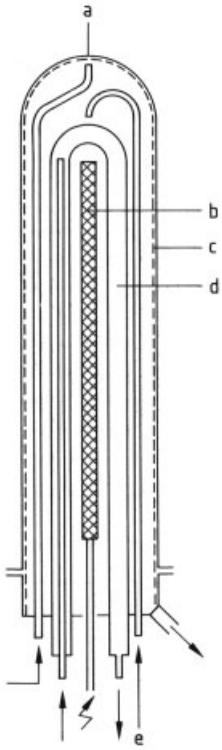

Figure 8. Example for a falling film photoreactor of type 4 C [11]
a) Formation of the falling film; b) Light source; c) Falling film; d) Cooling; e) Solvent for cleaning the lamp housing

3.2. Light Sources

In photochemistry two fundamental arrangements of light sources exist. In one the lamp, a long cylinder, is surrounded by a cooling jacket that is immersed in the reaction solution (Fig. 10); in the other, the reaction mixture is irradiated from outside using a focussing reflector. The energy output of a lamp can best be utilized by performing the photochemical reaction within the space between two concentric tubes surrounding the lamp, or by applying the aforementioned immersion mode. The choice of light source is dictated by a suitable combination of spectral energy distribution, radiant power, and

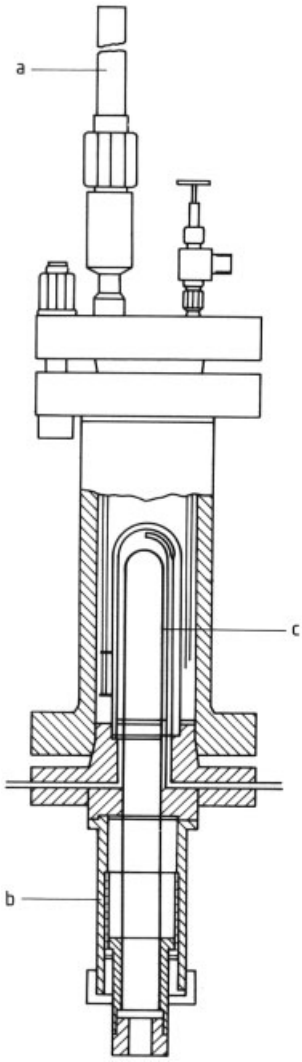

Figure 9. Example of an UV autoclave [12]
a) Housing for magnetic stirrer; b) 15 MPa resistant steel tube with attachment; c) Light source

filter for the special case considered. If the reactant absorbs in more than one region of the spectrum, as nearly all organic molecules do, irradiation in different bands may induce different photochemical reactions. Thus, irradiation is mainly restricted to absorption in the longest wavelength region by applying the usual polychromatic lamps in combination with a suitable cutoff filter. For a typical medium pressure mercury lamp with an electrical power input of 10 – 60 kW, the total output of photochemically utilizable light (200 – 600 nm) averages 44.2 %. The insertion of boron silicate glass decreases the light output (300 – 600 nm) to 19.6 %.

Furthermore, is it noteworthy that the light output of mercury medium-pressure lamps (Figure 10) at $\lambda = 200 - 600$ nm, expressed as percentage of the electrical power input, increases more than proportional relative to the latter, as can be recognized from the following comparison (Table 2) [2].

Table 2. Percentage of light output relative to the electrical power input

Electrical power input, kW	Light output (220 – 600 nm), %
2	19.5
4	27.4
10	46.3
20	43.5

Reason for this effect is the formation of ″van der Waals″ collision complexes consisting of two mercury atoms. Starting with 2 kW their number increases with increasing electrical power input and remains nearly constant from 10 – 60 kW [2]. The light emitted by these collision complexes adds to the light output of excited single mercury atoms which are the original emitting species of mercury lamps.

For technical purposes it is useful to know the conversion of the radiant power P in watts emitted by the lamp into the photon flow φ_{P} in einstein per hour as a function of the wavelength λ in nanometers:

$$\varphi_{\mathrm{P}} = 3.01 \times 10^{-5} \lambda P \tag{13}$$

The majority of light sources are of the immersion lamp type (see Fig. 10). The spectral energy distributions of the widely used medium-pressure mercury lamps with a power input of 2 – 60 kW are given in Table 3. Medium-pressure mercury lamps can be doped with other metals resulting in a weakening of their original polychromatic light in favor of new emission lines which are characteristic of the added metal. Antimony offers light in the range 270 – 300 nm where undoped mercury lamps emit poorly. TlI and GaI create very strong lines at 535 nm and 417 nm, respectively.

Newly developed incoherent excimer sources are now available for performing photochemical reactions (Fig. 11) [14]. Because excimer sources do not possess a stable ground state, self-absorption cannot occur. This should lead to an intrinsic efficiency of at least 30 %. But

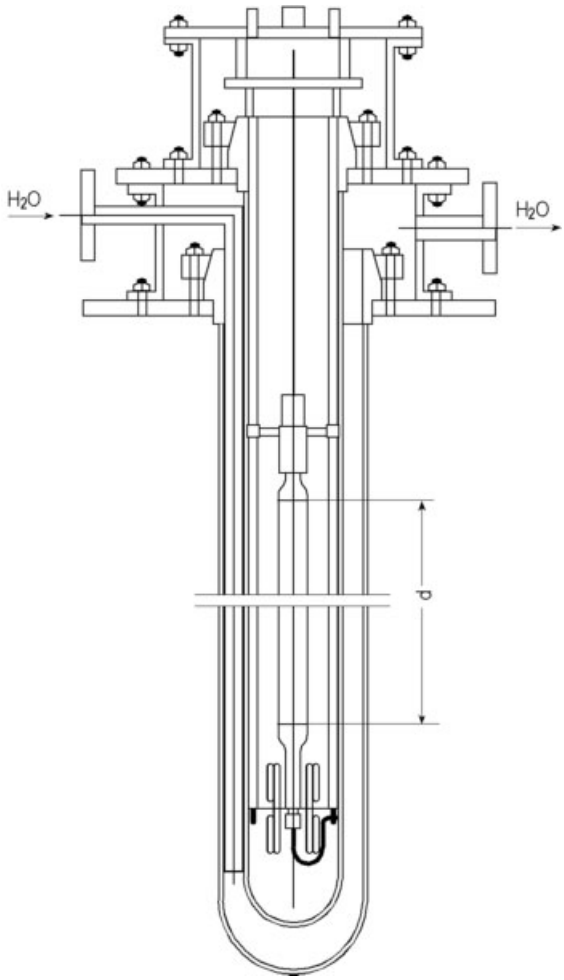

Figure 10. Fundamental arrangement of a mercury medium-pressure lamp with an electrical input 2 – 60 kW, equipped with cooling jacket and gasket.

the surrounding metal net which is necessary for the dielectric-barrier discharge lowers the efficiency to ca. 10 % by absorption and reflection of the generated light. The excimer light sources are operated with frequencies of about 50 kHz, and voltage amplitudes of 5 – 10 kV are required to ignite the discharge. Under those conditions, bursts of light pulses with a repetition frequency of about 100 kHz are produced. Compared with mercury or xenon lamps excimer sources generate considerable less heat. As immersion lamps they are delivered with an already installed cooling system which makes further cooling of the irradiated solution unnecessary. With proper choice of the gas mixture, excimers can be found that radiate only in a narrow specified wavelengths range, e.g., at 172, 222, and 308 nm. Disadvantage for an upscaling of photoreactions performed with excimer sources is the limitation of their electrical power input to 1.5 kW. The most striking difference to other UV sources, however, is the wide choice of geometries. The possibility of constructing large flat radiating panels as well as hollow cyclinders filled with UV radiation could lead to a variety of new photochemical applications [14].

In addition to artificial light sources, concentrated sunlight has been taken into consideration for the photochemical production of fine chemicals [15]. Here, the photoreaction occurs in a solution flowing through a glass tube in the focal line of a parabolic trough with suitable light

Table 3. Spectral energy distribution of medium-pressure mercury lamps (Heraeus)

Radiating part d, mm	1044	1044	1535	1520	1520	2000
Electrical power input, kW	2	4	10	20	40	60
Wavelength λ, nm		Photon flow φ_p, einstein/h				
248	0.045	0.140	0.76	1.40	2.24	3.36
254	0.280	0.730	3.06	7.90	17.58	26.22
265	0.096	0.420	1.44	2.10	3.59	4.23
270	0.023	0.059	0.37	0.59	1.46	1.46
275	0.020	0.053	0.25	0.56	0.75	1.16
280	0.040	0.140	0.96	1.50	2.53	3.71
289	0.015	0.045	0.37	0.52	1.48	1.65
292	0.009	0.023	0.16	0.15	0.44	0.53
296	0.070	0.200	1.63	2.30	4.83	7.15
302	0.150	0.480	2.07	3.20	6.64	8.71
313	0.430	1.050	3.72	6.30	15.83	18.65
334	0.020	0.073	0.58	1.00	2.11	2.71
366	0.730	1.820	7.66	12.70	27.98	34.92
391	0.006	0.016	0.12	0.26	0.36	0.59
405/8	0.320	0.800	2.52	4.70	14.70	15,42
436	0.530	1.490	4.36	8.40	23.23	27.82
492	0.009	0.025	0.21	0.07	0.44	0.44
546	0.850	2.100	5.95	13.10	37.14	45.85
577/9	0.330	1.180	9.64	19.80	39.49	56.72

reflecting properties. The sun tracking is performed automatically three-dimensional on two axes, e.g., with a Helioman Module from MAN (Fig. 12) [16]. Irradiation with sunlight initiates not only photoreactions but also heats up the solution. This limits the concentration of sunlight by parabolic mirrors, even when the solution is rapidly flowing through external heat exchangers. In the Helioman Module (Fig. 12) some reflecting parts are deleted, in order to show how they are arranged. The Helioman Module is part of the so-called PROPHIS system [17] and the relevant data are given in Table 4.

Table 4. Data for the Helioman Module

Length of one parabolic mirror	4.5 m
Width of one parabolic mirror	1.81 m
Aperture angle	70.5°
Focal distance	0.64 m
Total reflecting area	ca. 30 m^2
Reactor volume	40 – 80 L
Maximum solar power	up to 30 kW

3.3. Materials

Apart from light-transmitting quartz or glass (Table 5), nonmetallic construction materials

Table 5. Thickness of optical materials determining a transmission limitof 10 %

Material	Thickness, mm	Cutoff λ, nm
Poly(methyl methacrylate)	2.5	297
	5	311
	10	326
Window glass, standard	1	307
	3	314
	10	330
Pyrex (Corning 774)	1	280
	2	297
	4	310
Vycor 791	1	212
	2	213
	4	217
Suprasil I (Heraeus)	1	160
	10	167
	30	170
Suprasil-W (Heraeus)	1	156
	10	172
	30	178

are preferred because traces of polyvalent metal ions, e.g., from stainless steel, may change the direction of a photoreaction considerably. Titanium vessels are required only if there is contact with extremely corrosive solutions which occur in photonitrosation processes. Pumps are another source of possible contamination. Therefore, centrifugal pumps with magnetic coupling are recommended, provided that all parts in con-

tact with the pumped medium consist of materials with suitable properties.

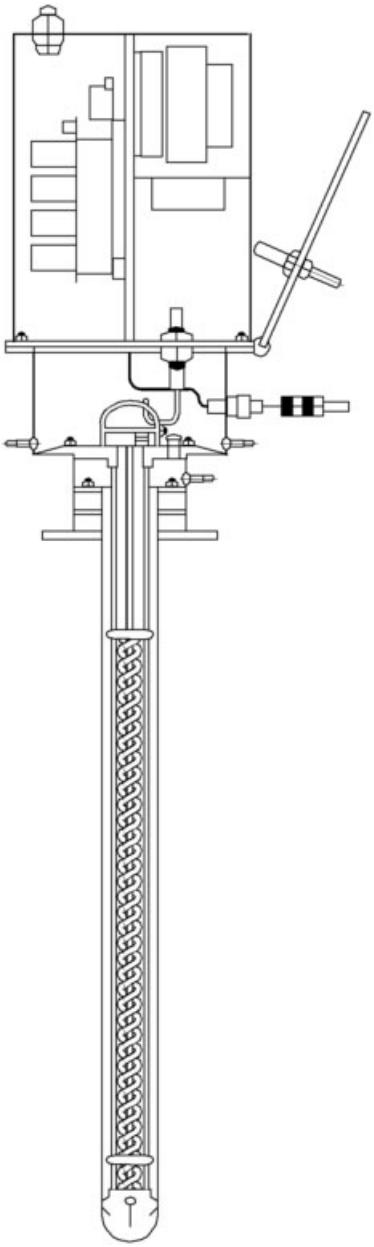

Figure 11. Excimer immersion lamp [14] (Heraeus Noblelight)

Also important is the elimination of UV attenuating films which may result from photoreactions and/or from algae in the cooling water and which are frequently invisible. These films are best removed with an aqueous solution of 10 % HF. In addition, the appearance of hazardous electrostatic charges is observed when 1,4-dioxane or other ethers, and tetrachloromethane circulate as solvents through tubes made of glass or polyfluorinated plastics. Electrostatic charging can be avoided not by earthing the apparatus but by using an admixture of protic solvents, e.g., 2 vol % *tert*-butanol, which do not participate in the photoreaction.

3.4. Scaleup

Photoreactions can either be batch or continuously operated. The typical *batch reactor* for photoreactions in a homogeneous liquid phase consists of a large vessel containing an appropriate number of lamps and a stirring device, and can be maintained under inert gas if necessary (Fig. 13). A great part of the electrical power supplied to the lamp is transferred into heat; the proportion decreases from 63 % for a 20 kW lamp to 37 % for a 60 kW lamp. This heat is usually removed by water flowing between the inner and outer wall (quartz or glass) surrounding the lamp. The cooling water may be replaced by a filter solution flowing through an external heat exchanger. If the filter solution does not come into contact with metals it is advantageous to replace the cooling water by an aqueous solution of 35 g/L $CuSO_4 \cdot 5\,H_2O$ (optical path length 1 cm, cutoff at 300 nm). In this way, convection heat as well as radiant heat is removed, because in addition to the UV light of 200 – 300 nm, the Cu^{2+} ion absorbs strongly in the infrared region.

An apparatus using an immersion lamp to perform gas – liquid reactions such as photonitrosation is shown in Figure 14.

Continuously working photoreactors are mainly used for light-induced chain reactions in vapor-, liquid-, or mixed-phase reactions. Because of the very high reaction rates only short contact times are necessary. These can be best achieved in a flow system, e.g., in concentric tubes surrounding the lamp [18].

Scaleup Problems. Though theoretical studies concerning practical photochemistry given in [13] contribute to the understanding of the problems involved, the best approach to scaleup

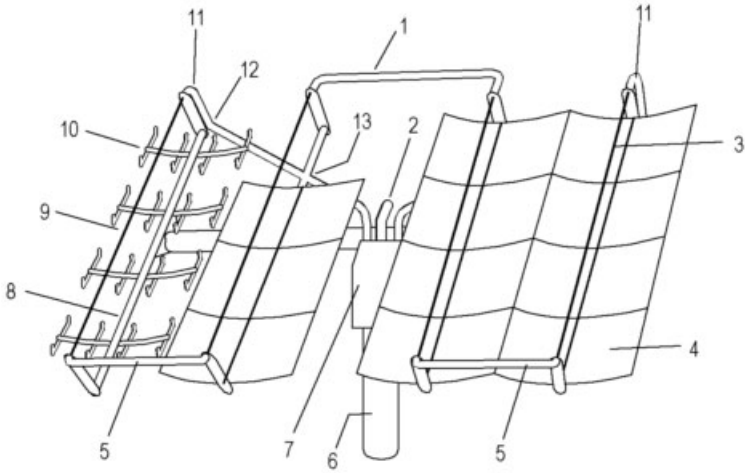

Figure 12. Arrangement for the photochemical application of sunlight (MAN Helioman Module [16])
1, 5, 11, 12, 13 = System of tubes for transportation of the irradiated liquid; 2 = Solar sensor; 3 = Light absorbing tube; 4 = Reflector, 8, 9, 10 = Reflector support

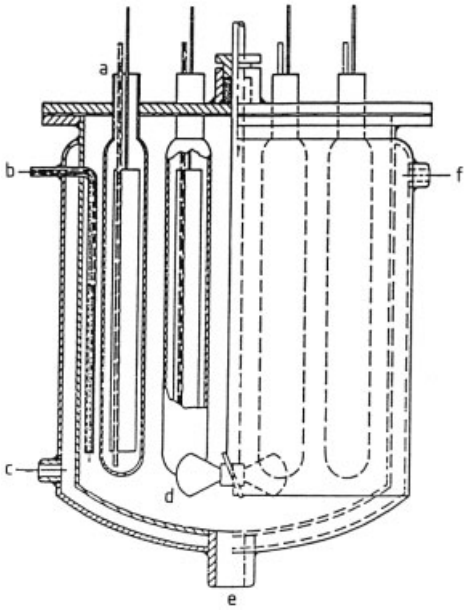

Figure 13. Batch reactor for photochemical manufacturing processes in a homogeneous liquid phase
a) Lamp unit with N_2 cooling; b) Gas inlet; c) Heat-exchange input; d) Stirrer; e) Solution exit; f) Heat-exchange output

a photoreaction seems to be the enlarging of a laboratory-type prototype reactor with the important parameters being changed as little as possible. For the same reason initial pilot-plant studies should be carried out using a reactor with only one production-size lamp. Once the characteristics of a flow reactor with one lamp are understood, further scaleup is simply a matter of adding more lamps and/or more reactors.

3.5. Economic Considerations

With respect to light-induced chemical transformations, the universal quantum yield φ (see Eq. 6) must be replaced by the quantum yield φ':

$$\varphi' = \frac{\text{moles of product obtained}}{\text{einsteins of light absorbed}} \tag{14}$$

which is also measured in terms of monochromatic light by complicated actinometric systems. In industrial photochemistry this method is not useful for determining the product-specific electrical cost of running a light source. In most cases the starting material absorbs more than one of the lamp's emission lines. Therefore, to consider the absorption of an entire spectral region rather than one wavelength a quantum yield φ'' must be defined which can be determined by simple measurement of the moles of product obtained per kW · h of electrical energy consumed by the lamp. Such an estimation should include all utilizable radiation filling up the whole solid angle. This is the case when an immersion lamp is used.

$$\varphi'' = \frac{\text{moles product / kW·h}}{\text{einsteins / kW·h}} \tag{15}$$

Assuming that about 10 % of the energy input is utilized as photochemically active light, most of the total energy input has to be removed as heat. This increases the electrical cost and must be taken into account by multiplying the energy input of the lamp by a factor of about 3. In addition to the cost of equipment and the cost arising from the limited lifetimes of the lamps (i.e., 2000 – 6000 h for Hg low pressure lamps and 1000 – 3000 h for Hg medium pressure lamps), it must be considered that a radiant power of 1 einstein/h is equivalent to 130.8 W at 254 nm but only to 60.8 W at 546 nm (both are Hg emission lines).

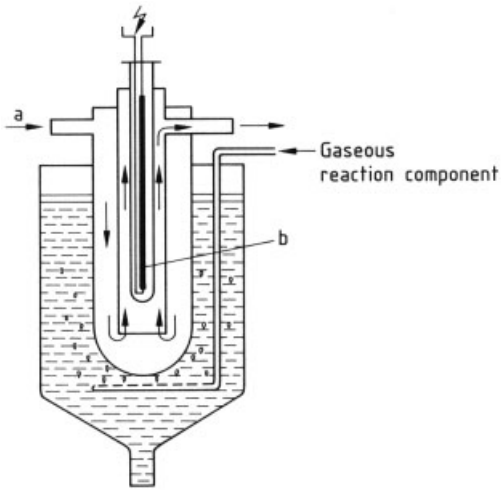

Figure 14. Photoreactor for gas-liquid reactions
a) Cooling; b) Light source

The costs resulting from solvent losses and recycling may exceed the electrical cost of running the lamp. For the complete photochemical conversion of a compound dissolved in ether with a concentration of 1 wt % into the desired product, ca. 100 t of solvent need to be recycled to obtain 1 t of product. This requires 1.054 $\times 10^4$ kW · h (calculated from the specific heat of ether C_p (25 °C) = 172.5 $\times 10^{-3}$ J/mol and from its heat of vaporization = 26.508 $\times 10^3$ J/mol).

Many chemists who are not specialized in photochemistry are not aware of the fact that especially in quasi-stoichiometric photoreactions the molecular mass of the photoproduct must be taken into account, as it determines the amount of material in kg or tons produced by photoreactions with the same molar conversion in the same irradiation time (Fig. 15) [2].

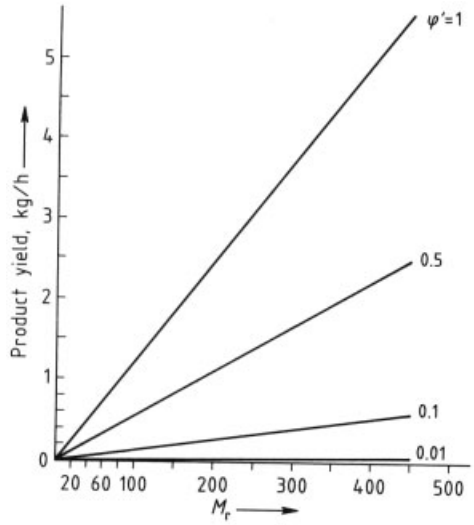

Figure 15. Molecular mass and yield of the photoproduct using a 20 kW Hg medium pressure lamp radiating 12.7 einstein/h at 366 nm at various molar conversions (φ'). Reproduced with permission [6]

A theoretical treatment of the determination of the most suitable irradiation conditions can be found in [13]. One way to characterize a photoreaction is to consider its quantum yield φ'. Light-induced chain reactions are indicated by $\varphi' > 1$, whereas in photochemical conversions with $\varphi' \leq 1$ light plays the role of a reagent. Therefore, such photoreactions may be considered as "quasi-stoichiometric". Examples of both types of photoreactions are given in Chapters 4 and 5.

4. Light-Induced Chain Reactions

4.1. Photohalogenation

The values of λ_{max} and ε_{max} of the halogens X_2 increase with increasing atomic number (Table 6). The small value of ε_{max} reflects the forbidden character of the transition which becomes progressively more allowed as one moves to heavier atoms. The transition responsible for these absorption bands is presumed to be a singlet – triplet promotion which largely results in the formation of halogen atoms [5]:

Table 6. Characteristic light absorption data of the halogen molecules in the gas phase

Halogen X_2	λ_{max}, nm	ε_{max}, L mol^{-1}cm^{-1}
F_2	284.5	6
Cl_2	330	66
Br_2	420	200
I_2	520	950

$$X_2(\uparrow\downarrow) \xrightarrow{h\nu} X_2(\uparrow\uparrow)^* \longrightarrow 2X\cdot \quad (16)$$

The greater the dissociation energy of X_2 (Table 1), the greater is the reactivity of the radicals formed by dissociation; in contrast, cleavage of weak bonds leads to less reactive radicals. Thus, light-induced chain reactions may occur when Cl_2 or Br_2 are involved, but not with I_2.

Photochlorinations. Industrial photochlorinations are determined by the quantum yields, which depend on the structure of the hydrocarbon to be photochlorinated and are in the range $10^3 - 10^6$. Even traces of oxygen act as an inhibitor. The reaction chains may be interrupted by removal of either a chlorine atom or a hydrocarbon radical:

$$Cl\cdot + O_2 \rightarrow ClO_2 \quad (17)$$
$$R\cdot + O_2 \rightarrow RO\cdot_2 \quad (18)$$

The ClO_2 molecules and $RO_2\cdot$ radicals enter into secondary reactions which mainly result in the formation of stable molecules without regeneration of chain-propagating free radicals.

Since the probability of chain termination increases with the concentration of newly formed chain-starting radicals, the economy of photochlorination decreases with increasing photon irradiance E_P (einstein m^{-2} s^{-1}). Here, the use of weak lamps is preferred. Additionally, the speed of photochlorination, e.g., of toluene, depends on the chlorine excitation wavelength. The best conditions are given when the maximum light emission of the lamp lies in the blue region where the reaction zone l' (cm) becomes larger. This is due to the variation of the molar absorption coefficient ε of chlorine. Thus, at 350 nm 99 % of the light is absorbed within a very short distance leading to high local concentrations of chlorine atoms which favor recombination and thereby reduce the chain reactions (see Section 3.1).

Photochlorination can occur according to two different mechanisms:

1) Addition of chlorine
2) Substitution by chlorine

Addition of Chlorine. The synthesis of the pesticide *lindane* (γ-hexachlorocyclohexane) is a well-known example of photochemical additions of chlorine to benzene. Because of its toxicity, lindane is no more commonly used and the production of lindane will not be discussed here. More important is the industrial preparation of oxalyl chloride which is a reagent with many applications in the production of pharmaceuticals, agrochemicals, etc. Since syntheses of oxalyl chloride starting from oxalic acid have many disadvantages, such as low yields, high amounts of waste, purification and safety problems, Tessenderlo Chemie in Belgium has developed a suitable photochemical procedure by improving the method of producing oxalyl chloride from ethylene carbonate, already known since 1936. This procedure combines the photochemical addition of chlorine with photochlorination by substitution, yielding the corresponding tetrachloro compound which can be decomposed into oxalyl chloride and phosgene (see Eq. 19).

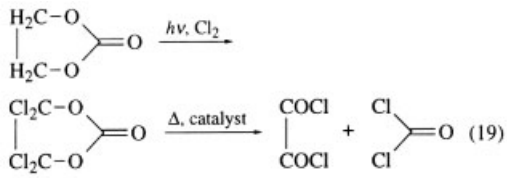

Photochlorination by Substitution. Photochlorination by substitution proceeds via the following mechanism:
Initiation: see Equation (16)
Propagation:

$$RH + Cl\cdot \longrightarrow R\cdot + HCl \quad (20)$$
$$R\cdot + Cl_2 \longrightarrow RCl + Cl\cdot \quad (21)$$

Termination:

$$R\cdot + R\cdot \longrightarrow R_2 \quad (22)$$
$$R\cdot + Cl\cdot \longrightarrow RCl \quad (23)$$
$$Cl\cdot + Cl\cdot \longrightarrow Cl_2 \quad (24)$$

An important example of the substitution mechanism is the *production of benzyl chloride*

by photochlorination of toluene. Here, the heat from the uncooled lamps together with the heat of reaction maintain the toluene under reflux. Thus, photochlorination of the aromatic ring can be avoided. Photochlorination of the aromatic ring is also catalyzed by metallic impurities in the toluene; this effect can be inhibited by adding complexing agents because purification of the toluene by distillation prior to side-chain photochlorination is economically not feasible. The ratio $Ar-CH_2Cl$: $Ar-CHCl_2$: $Ar-CCl_3$ is 10 : 1 : 0.1.

The chain reaction during substitution of hydrogen atoms by chlorine proceeds because the energy of 432 kJ/mol obtained through the formation of each mol HCl exceeds the dissociation energy of Cl_2 (see Table 1). This concerns also the side-chain photochlorination of *p*-xylene to $\alpha,\alpha,\alpha,\alpha',\alpha',\alpha'$-hexachloro-*p*-xylene (Eq. 25).

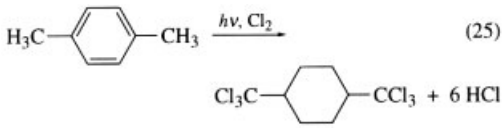

(25)

This is the first step in the aramid fiber production (→ High-Performance Fibers, Chap. 3.2.1.) at Aramid Products V.o.F./Netherlands, a subsidiary of Akzo Nobel N.V. The semi-batch reaction is performed in two photoreactors, each of them equipped with three UV lamps of 10 kW. Similar to the production of benzylchloride the reaction temperature is 80 – 140 °C, in order to avoid addition of chlorine to the benzene ring. At the end of the reaction the concentration of hexachloro-*p*-xylene is 98.5 % without purification. In the presence of a catalyst hexachloro-*p*-xylene reacts with terephthalic acid to terephthalyl chloride which is polymerized with *p*-phenylene diamine to give poly-*p*-phenylene-terephthalylamide (PPTA) with a capacity of ca. 10 000 t/a.

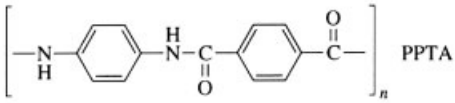

Part of the polymer PPTA is then spun to Twaron, a fiber with high thermal and chemical resistance whose tensile strength is five times higher than that of steel of the same weight.

Phosgene is a highly toxic gas which has been used in the chemical industry in enormous quantities for chloroformylations, carbonylations, formation of isonitriles, etc. Because of its toxic character substitutes for phosgene are used whenever possible. The liquid trichloromethyl chloroformate behaves like 2 mol phosgene in chemical reactions and is therefore called "diphosgene". Its use is safer and easier than that of phosgene. Even safer, especially during transport and storage, is the crystalline bis(trichloromethyl)carbonate or "triphosgene" (→ Phosgene, Chap. 2.2.; → Isocyanates, Organic, Chap. 4.).

In a suitable medium (e.g., dissolved in toluene) it reacts like 3 mol phosgene. During the photochlorination of methyl chloroformate, or dimethylcarbonate, the hydrogen atoms of the methyl groups are stepwise substituted by chlorine atoms. (Eqs. 26 and 27). Both reactions are exothermic and their $-\Delta H$ exceeds the dissociation energy of Cl_2, hereby initiating chain reactions which guarantee relatively short irradiation times, even with sunlight

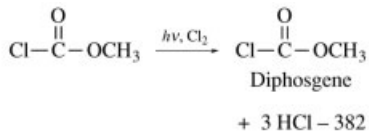

+ 3 HCl – 382 kJ (26)

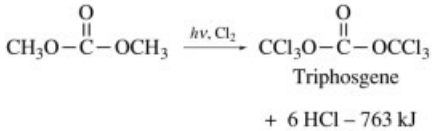

+ 6 HCl – 763 kJ (27)

Photobromination. The photobromination of toluene proceeds according the following reactions:

$$Br_2 + h\nu \longrightarrow Br_2* \quad (28)$$

$$Br_2* \longrightarrow Br\cdot + Br\cdot \quad (29)$$

$$Br\cdot + Br\cdot \longrightarrow Br_2 \quad (30)$$

$$PhCH_3 + Br_2 \longrightarrow PhCH\cdot_2 + HBr \quad (31)$$

$$\Delta H = -10\,\text{kJ/mol}$$

$$\text{PhCH}\cdot_2 + \text{Br}_2 \longrightarrow \text{PhCH}_2\text{Br} + \text{Br}\cdot \tag{32}$$

$$\Delta H = -21\,\text{kJ/mol}$$

The chain reactions (31) and (32) are exothermic processes. The heats of formation and dissociation involved in the photobromination of toluene are as follows:
Heats of dissociation

$$\Delta H\,(\text{Br}-\text{Br}) = 193\,\text{kJ/mol}$$
$$\Delta H\,(\text{PhCH}_2-\text{H}) = 356\,\text{kJ/mol}$$

Heats of formation

$$\Delta H\,(\text{H}-\text{Br}) = -366\,\text{kJ/mol}$$
$$\Delta H\,(\text{PhCH}_2-\text{Br}) = -214\,\text{kJ/mol}$$

The dissociation energy of the Br_2 molecule (193 kJ/mol) corresponds to a wavelength of 620 nm. Although Br_2 absorbs efficiently at 511 nm, irradiation with light of shorter wavelengths is necessary to start the chain reaction:

λ, nm	Absorption	Photolysis
620	scarce	no
550	efficient	no
511	efficient	yes

4.2. Photosulfochlorination

Sulfochlorination of *n*-alkanes produces almost random substitution and starts, like photochlorination, with the formation of chlorine atoms (Eq. 16) and the formation of alkyl radicals R· (Eq. 20). The propagation is then supplemented by

$$\text{R}\cdot + \text{SO}_2 \longrightarrow \text{RSO}_2\cdot \tag{33}$$
$$\text{RSO}_2\cdot + \text{Cl}_2 \longrightarrow \text{RSO}_2\text{Cl} + \text{Cl}\cdot \tag{34}$$

and the termination by

$$\text{RSO}_2\cdot + \text{Cl}\cdot \longrightarrow \text{RSO}_2\text{Cl} \tag{35}$$

Analogous to photochlorinations, oxygen is also a powerful inhibitor for sulfochlorinations. The alkylsulfonyl chlorides are starting materials for the preparation of detergents, emulsifiers, corrosion inhibitors, and auxiliaries for the textile and paper industries.

4.3. Photosulfoxidation

The $(n, \pi*)$ excitation of SO_2 dissolved in hydrocarbons exhibits an absorption maximum at 290 nm and its singlet energy lies below 380 kJ/mol. Since the dissociation energies for C – H bonds in ethyl – CH_2 and in methyl groups are given by 410 kJ/mol and 435 kJ/mol, respectively, the observed hydrogen abstraction might be considered to proceed via an $(n, \pi*)$ triplet as well as via an exciplex $(SO_2 \cdot\cdot RH)*$. Therefore, the reaction mechanism can be generalized as follows:
Initiation

$$\text{SO}_2 + h\nu \longrightarrow \text{SO}_2* \tag{36}$$

$$\text{RH} + \text{SO}_2* \longrightarrow \text{R}\cdot + \text{SO}_2\text{H} \tag{37}$$

Propagation

$$\text{R}\cdot + \text{SO}_2 \longrightarrow \text{RSO}_2\cdot \tag{38}$$

$$\text{RSO}_2\cdot + \text{O}_2 \longrightarrow \text{RSO}_2\text{OO}\cdot \tag{39}$$

$$\text{RSO}_2\text{OO}\cdot + \text{RH} \longrightarrow \text{RSO}_2\text{OOH} + \text{R}\cdot \tag{40}$$

Termination

$$\text{RSO}_2\text{OOH} + \text{H}_2\text{O} + \text{SO}_2 \longrightarrow \text{RSO}_3\text{H} + \text{H}_2\text{SO}_4 \tag{41}$$

In the absence of water the peroxysulfonic acid RSO_2OOH is cleaved into $RSO_2O\cdot$ and HO· ; both radicals are able to enter the propagation sequence:
Propagation

$$\text{RSO}_2\text{OOH} \longrightarrow \text{RSO}_2\text{O}\cdot + \text{HO}\cdot \tag{42}$$

$$\text{RSO}_2\text{O}\cdot + \text{RH} \longrightarrow \text{RSO}_2\text{OH} + \text{R}\cdot \tag{43}$$

$$\text{HO}\cdot + \text{RH} \longrightarrow \text{R}\cdot + \text{H}_2\text{O} \tag{44}$$

The radicals $RSO_2O\cdot$ and HO· may abstract hydrogen from the hydrocarbon, thereby forming alkyl radicals R·. The sulfoxidation becomes autocatalytic and irradiation is only necessary to initiate the reaction sequence.

In practice the peroxysulfonic acids are continuously *extracted with water.* This process prevents formation of polysubstituted alkanes. Further reaction with SO_2 produces the sulfonic acids shown in the termination step (Eq. 41). In this case the quantum yields obtained are very much lower than for sulfochlorination, being in the range of only $\varphi' = 7$ or 8. Nevertheless, Hoechst runs three plants operating on the light-water procedure with a total capacity of 72 000 t/a. The lamps consume 0.2 kW · h per kilogram of sulfonate produced, i.e., 5 kg/kW · h.

Such "photosulfonates" are produced and marketed, e.g., by Hoechst as detergents with good biodegradability.

4.4. Type I Photooxidation

Type I photooxidations start with hydrogen abstraction from suitable substrates, e.g., from alcohols, by photoexcited diarylketones, acetophenone, aromatic aldehydes, and their heterocyclic analogs. Further sensitizers are α-diketones such as diacetyl and benzil, camphor-1,2-diketone, and orthoquinones e.g., 9,10-phenanthrenequinone. Most substrates can only undergo Type I photooxidation with ketones or quinones with $(n, \pi*)$ excited states but not with quinoid dyes, because their $(\pi, \pi*)$ states cannot abstract H atoms.

Type I photooxidations are not very useful for industrial syntheses; on the contrary, it is important to avoid them in industrial processes by means of triplet quenchers or antioxidants, e.g., to suppress the yellowing of colorless plastics etc.

4.5. Photopolymerization

Photopolymerization by UV irradiation almost always requires the presence of a chemical photoinitiator. Depending on the type of photopolymerization, the photoinitiator system can be free-radical or ionic. Two types of free-radical photoinitiator systems exist: those that produce radicals by intermolecular hydrogen abstraction and those that produce them by intramolecular cleavage. Ionic photoinitiator systems utilize the photolysis of diazonium, sulfonium, or iodonium salts. Strong acids are thus formed which are able to initiate polymerization of electron-rich olefins, e.g., enolic ethers, or of cyclic ethers, e.g., oxiranes. Good photoinitiators are colorless and do not give rise to light-induced yellowing of the prepared polymer.

Photopolymerizations are characterized by a significant increase in the molecular mass of the starting compound. This can be achieved by linear polymerization of a monomer and/or by cross-linking. Since air-pollution problems resulting from heat-cured polymer solutions can be eliminated by photopolymerization, light-cured polymers have become more popular. In addition to further irreversible (color formation) and to reversible (photochromism) reactions, photopolymerizations also play an important role in stereolitography, resists for microlithography, and solder masks. Specially designed photocatalysts for ring-opening metathesis polymerization were developed by Ciba and applied together with unsaturated, strained cyclic monomers, oligmers, or polymers for use in imaging technique. This method also opens a wide range of new applications in microelectronics, tooling, and in the printing wiring board industry. The new photocatalysts are Ru(II) complexes and high-valent polyalkyl complexes of lower transition metals. Involved are the photochemical solvatation of Ru(II) in an aqueous 0.05 mol/L H_2SO_4 – ethanol mixture (3 : 2), e.g. (see top of next page)

and the photoinduced H_α-abstraction shown in

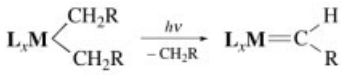

(47)

where L are the ligands and M is one of the above mentioned metals.

Because the world consumption of UV-curable products presents a multibillion dollar market with an annual average growth of 15 – 25 % for the coming years, radiation curing is in the midst of a profitable period which has been awaited for a long time. Development of new reactive oligomers and monomers, water-borne formulations, improved photoinitiators, and more efficient photosensitizers will presumably increase market growth.

$$Ru(NCCH_3)_6\,(Tos)_2 \xrightarrow{h\nu,\ EtOH,\ H_2O} Ru(H_2O)_6\,(Tos)_2 \qquad (45)$$

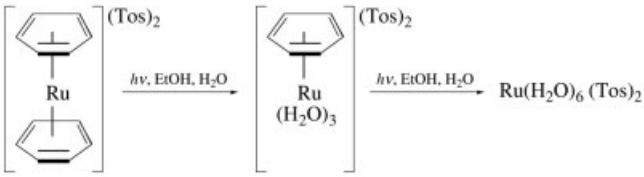

(46)

Polymerization of acrylate monomers is strongly inhibited by oxygen. Therefore, oxygen is removed from the curing environment by blanketing with inert gas.

5. Quasi-Stoichiometric Photoreactions

5.1. Photonitrosation

Photonitrosation of alkanes by light-induced cleavage of NOCl has been known since 1919. In the 1950s, Toray Industries/Japan developed the process for production of ε-caprolactam by photonitrosation of cyclohexane and installed the first production unit in 1963. This was the first process for production of nylon 6 starting from cyclohexane. Today, only 10 % of the world's ε-caprolactam capacity of ca. 2 900 000 t/a comes from photonitrosation of cyclohexane, 90 % are prepared by catalytic oxidation of cyclohexane with air. In 1975 the French company ATO Chimie used the photonitrosation of cyclododecane also for the production of nylon 12.

Nitrosyl chloride absorbs in the whole spectral range from 200 to 600 nm. Up to 500 nm the absorption spectrum consists of a continuum, whereas in the range 500 – 600 nm it is evidently structured indicating a zone of predissociation. Therefore, two mechanisms for the primary photochemical step are proposed, depending on the NOCl excitation below and above 500 nm:

NOCl excitation below 500 nm:

$$NOCl + h\nu \longrightarrow NO\cdot + Cl\cdot \qquad (48)$$

$$R^1R^2CH_2 + Cl\cdot \longrightarrow R^1R^2CH\cdot + HCl \qquad (49)$$

$$R^1R^2CH\cdot + NO\cdot \longrightarrow R^1R^2CHNO \qquad (50)$$

NOCl excitation above 500 nm:

$$NOCl + h\nu \longrightarrow NOCl* \qquad (51)$$

$$R^1R^2CH_2 + NOCl* \longrightarrow R^1R^2CH\cdot + NO\cdot + HCl \qquad (52)$$

$$R^1R^2CH\cdot + NO\cdot \longrightarrow R^1R^2CHNO \qquad (53)$$

Excess HCl is fed into the system with NOCl, transforming the photochemically obtained nitroso compound into an oxime which undergoes a Beckmann rearrangement in the presence of oleum to give ε-caprolactam and dodecyl lactam. To prevent deposition of tar on the lamps, radiation below 366 nm is removed, e.g., by adding cutoff filters to the cooling water. Continuous cleaning of the lamp housing with a falling film of concentrated H_2SO_4 is another method. Light sources are 60 kW Hg lamps doped with TlI which create a strong emission line at 535 nm. The reaction then proceeds mainly according to the second mechanism. The advantage of the second mechanism is its higher selectivity since the formation of chlorocyclohexane or chlorocyclododecane is suppressed, resulting in a higher quantum yield of the desired products. For nitrosocyclohexane the quantum yield increases from 0.7 at 350 nm to ca. 1 at 578 nm. The choice of optimal reaction conditions guarantees a cyclohexanone oxime yield > 90 %.

5.2. Photosynthesis of Previtamin D

The term previtamin D covers different types of precalciferols. Provitamin D stands for the corresponding precursors and vitamin D for the resulting calciferols. The most important components are the compounds **1a**, **1b–5a**, **5b** (see Fig. 16).

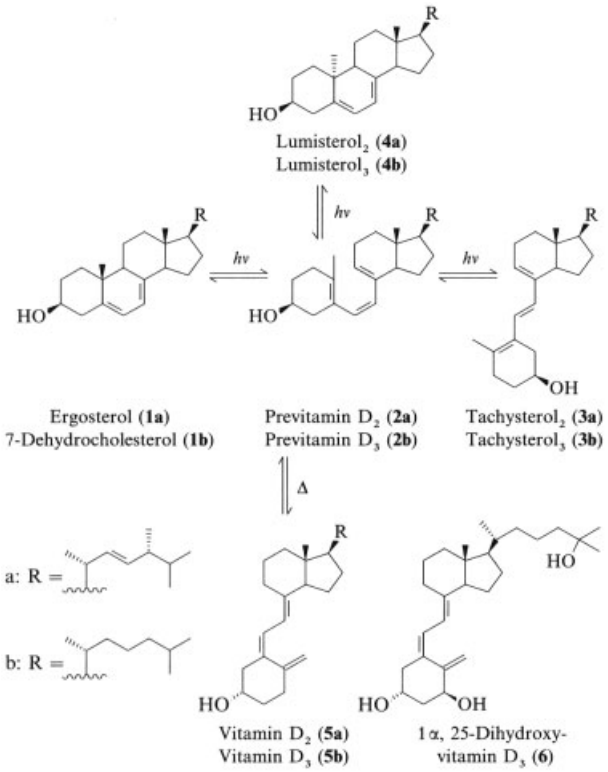

Figure 16. Important intermediates in the photosynthesis of previtamin D

The manufacturing process of vitamin D (**5a**, **5b**) involves a photochemical step in which the 9,10-bond of provitamin D (**1a**, **1b**) is cleaved by irradiation with appropriate UV light to give the (*Z*)-triene previtamin D (**2a**, **2b**). With increasing consumption of **1a**, **1b**, the photochemically formed compounds **2a**, **2b** compete with **1a**, **1b** for light absorption. Hence, **2a**, **2b** is partly photoisomerized into the (*E*)-triene tachysterol (**3a**, **3b**) and to a minor extent into the ring-closed products lumisterol (**4a**, **4b**) and provitamin D (**1a**, **1b**). The quantum yields of all these photoreactions are wavelength- dependent. Thus, the highest quantum yields are obtained for the (*Z*/*E*) isomerization of **2a**, **2b** in the region 248 – 265 nm and for the formation of lumisterol (**4a**, **4b**) above 290 nm. The conversion of **2a**, **2b** into vitamin D (**5a**, **5b**) proceeds thermally. Starting from the corresponding steroid ring B dienes, the preparation of hydroxylated vitamin D_3 metabolites such as **6** follows the same mechanism.

The D vitamins are involved in the regulation of calcium homeostatis. Vitamin deficiency leads to several diseases such as rickets which is caused by failure to deposit calcium in bones. Vitamin D_2 (ergocalciferol) is only active in mammals, whereas vitamin D_3 (cholecalciferol) is active in mammals and birds, where it also has an effect on the eggshell stability. Hence, vitamin D_3 is produced on a large scale as a feed additive with an estimated worldwide output of 200 – 300 t/a. Calcitriol (**6**), 1,25-dihydroxycholecalciferol, (5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-1α,3β,25-triol, is used as a drug against renal osteodystrophy. Since the daily dose for calcitriol is only 0.25 µg, the worldwide production of **6** including 1α-hydroxy vitamin D_3 comes up to ca. 25 kg/a.

The efficiency of the photochemical step in the vitamin D_3 manufacturing process depends on irradiation of the bromine-free starting material **1b** after having introduced the 7,8 double bond into chlolesterol by bromination and dehydrobromination. When more than 400 ppm bromine- containing byproducts are present, the irradiation time must be prolonged by about

20 % to reach the same conversion as with pure **1b**. This is because the photochemical ring B opening proceeds via the (π, $\pi*$) excited singlet state. The heavy bromine atom accelerates intersystem crossing to the corresponding triplet state. Formation of previtamin D_3 (**2b**) is thus delayed and undesired byproducts appear in the reaction mixture.

As outlined in Section 3.5, it is not economical to perform quasi-stoichiometric photoreactions in highly diluted solutions. Thus, the vitamin D_3 manufacturing process involving photoreaction with 1 – 2 wt % **1b** in ether, tetrahydrofuran, 1,4-dioxane, or ethanol which is commonly stated in the literature does not reflect the present state of the art. In fact, **1b** is sparingly soluble in most solvents, especially in methanol. In order to achieve high concentrations (Section 3.5), solutions containing 5 wt % **1b**, e.g., in *tert*-butyl methyl ether, are irradiated with 20 kW mercury medium pressure lamps in falling film photoreactors similar to that in Figure 8. The high concentration of **1b** reduces the thickness of the photoreaction zone l' (Section 3.1), resulting in exchange of the reactants out of and into the reaction zone by diffusion. For this reason the flow along the light-emitting part d of a 20 kW lamp (see Fig. 10), which is 152 cm long, has to be adjusted to a residence time < 5 s. To avoid overirradiation products, the total conversion of **1b** should be kept below 40 %. In a falling film photoreactor the residence time is limited to a minimum of 3 s, corresponding to the maximum flow rate of ca. 160 L/h. Thus, increased production capacity cannot be achieved by using stronger lamps. The length of the light-emitting part of a 40 kW mercury lamp is also 152 cm and that of a 60 kW lamp is 200 cm (see Table 5). The only way to increase previtamin D_3 output is to multiply the aforementioned photoreaction unit working with 20 kW lamps connected in series or in parallel. Protic solvents should not be used since they are unable to dissolve **1b** sufficiently. They also add photochemically to previtamin D (D_2 and D_3) and even more so to the corresponding tachysterols [19] leading to byproducts which reduce the yield of the final product vitamin D.

5.3. Photoreactions in Supercritical Fluids

Humulones (α-acids) and lupulones (β-acids) are constituents of hop and are of great importance for giving beer its bitter taste. Lupulones differ from humulones by having a side chain instead of an OH group on C-3 of the six-membered ring. In an industrial process both compounds are extracted from hop using supercritical CO_2. Humulone is photochemically transformed into *trans*-isohumulone via oxadi-π-methane rearrangement (ODMP, see Section 7.4) in supercritical CO_2 as solvent (see bottom of page).

Figure 17 shows the flow sheet of the whole process [20].

6. Photosensitized Reactions

6.1. (*E/Z*)-Isomerization of Tachysterol

In the vitamin D_3 production process (Section 5.2) the most important unavoidable byproduct is (*E*)-triene tachysterol$_3$ (**3b**), usually formed in 12 – 15 % yield. Formerly it was separated from the reaction mixture by Diels – Alder reaction with maleic anhydride and subsequent extraction with NaOH. A double irradiation method was then introduced to reconvert **3b** into **2b** by a photosensitized (*E/Z*)-isomerization. 9-Fluorenone was added as photosensitizer to the reaction mixture resulting from the first irradiation. Afterwards a second irradiation was performed using light of longer wavelengths which

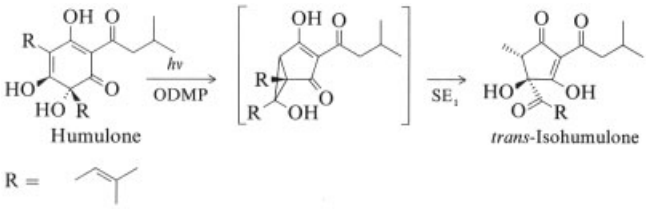

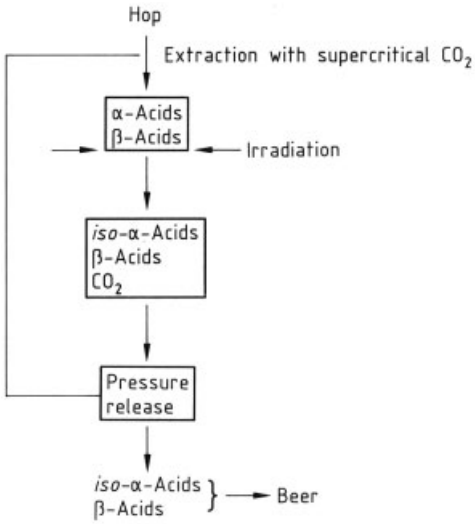

Figure 17. Enrichment of the bitter taste in beer by irradiation of the extract obtained from hop using supercritical CO_2 [20]

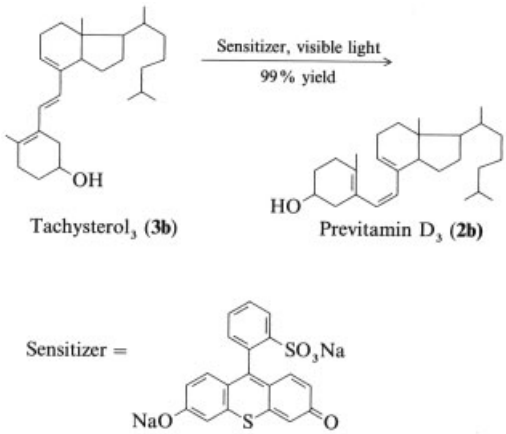

is not absorbed by the components of the original reaction mixture. Since only partial (*E*/*Z*)-isomerization occurred, more effective sensitizers such as anthracene were applied [21]. All these photosensitizers can only be eliminated from the reaction mixture by column chromatography, but this method cannot be used for large-scale production of vitamin D_3. To overcome this difficulty liquid polymer-bound anthracene sensitizers were developed which, after photoisomerization of **3b** into **2b**, are precipitated as an oil by adding methanol to the reaction mixture [21]. The starting material **1b** is also sparingly soluble in methanol. Because less than 40 % **1b** is consumed in the first irradiation step, the remaining **1b** would be precipitated together with the liquid sensitizer. Due to this contamination it could not be recycled directly. To prevent additional purification steps, 2-(6-hydroxy-3-oxo-3*H*-thioxanthen-9-yl)benzene sulfonic acid disodium salt was introduced as photosensitizer dissolved in small amounts of methanol (Hoffmann-La Roche process for the production of vitamin D_3). It is very effective in the (*E*/*Z*)-isomerization of **3b** and has the advantage that it can be separated from the final photoreaction mixture by extraction with water [22]or by ion exchangers. The photosensitized transformation of tachysterol **3b** into previtamin D_3 reduces the tachysterol content in the final product vitamin D_3 (**2b**) only to 1 %, whereas the aforementioned Diels — Alder reaction with maleic anhydride and subsequent extraction with aqueous NaOH decreases the tachysterol content to 0.1 %.

6.2. Type II Photooxidation

The 16 electrons of O_2 are distributed among the available molecular orbitals. Since triplet states with their parallel electron spins have lower energies than corresponding singlet states with antiparallel electron spins, molecular oxygen has a triplet state as its ground state symbolized by $^3O_2$. Direct absorption of light by $^3O_2$ to produce singlet oxygen, symbolized by $^1O_2$, is forbidden by the Wigner spin conservation rule [23].

Type II photooxidation generally predominates with sensitizers absorbing visible light. Energy can be transferred to $^3O_2$ from both the singlet and triplet excited states of the sensitizer. The triplet state of the sensitizer, however, usually has a much longer lifetime than its singlet state (see Fig. 1). Thus, with very few exceptions photoproduction of $^1O_2$ proceeds via this state:

$$\text{sens}(S_o) \xrightarrow{h\nu} \text{sens}^*(S_1) \xrightarrow{[ISC(S)]} \text{sens}^*(T_1) \xrightarrow{^3O_2} {}^1O_2 + \text{sens}(S_o) \quad (54)$$

Two species of $^1O_2$ exist with different energies. They are symbolized spectroscopically by $^1\Sigma_g$ and $^1\Delta_g$ and their lifetimes are solvent-

Table 7. Three electronic states and selected properties of molecular oxygen

Molecular state	Configuration of $\pi*_{2p}$		Energy, kJ/mol	Lifetime, s		Fluorescence, nm
				Gas phase	Solution	
$^1\Sigma_g^+$	↑	↓	156.9	7 – 12	10^{-9} – 10^{-12}	762
$^1\Delta_g$	↓↑		94.2	45	10^{-3} – 10^{-6}	1268.7
$^3\Sigma_g^-$	↑	↑	0		infinite	

dependent (Table 7). Because of its longer lifetime, only the $^1\Delta_g$ state is dealt with when using the term singlet oxygen ($^1O_2$).

The emission of the $^1\Delta_g$ state in the near IR at 1268.7 nm can be used as physical evidence for the presence of $^1O_2$. Useful chemical evidence is the photooxidative transformation of 2,3-dimethyl-2-butene into 3-hydroxyperoxy-2,3-dimethyl-1-butene which is stable enough to be measured by means of gas chromatography. Furthermore, unlike Type I photooxidations, a sharp reduction in the reaction yield is observed in the presence of 1,4-diazabicyclo[2.2.2.]octane (DABCO) or azide ion due to a sharp reduction of $^1O_2$ production.

Singlet oxygen was discovered in a photochemical experiment performed in 1931 by KAUTSKY. Nevertheless, convincing chemical proof for the existence of an excited oxygen species identical with $^1O_2$ was first found in 1964 by FOOTE and WEXLER. SCHENCK and coworkers previously performed many photooxygenations without recognizing that their oxidizing agent was $^1O_2$.

Sensitizers with a triplet energy $E_T > 95$ kJ/mol are able to transform $^3O_2$ into $^1O_2$ (Table 8). Their solubilities differ according to their chemical structure. Most sensitizers are soluble in protic and/or polar aprotic solvents. The problem of solubilizing anionic dyes such as rose bengal in aprotic solvents can be circumvented by the use of their tetramethylammonium salts or by addition of crown ethers. When strong light sources are applied, rose bengale the most used sensitizer, is rapidly decomposed. Therefore, to assure a constant light absorption a concentrated sensitizer solution must be continuously added to the reaction mixture. Because of its stability, lower molecular mass, stronger absorption, and resistance against acids, the photosensitizer TPP (Table 8) is much better adapted to technical photooxidations than rose bengal, as can be seen from Table 9.

Table 8. Suitable triplet sensitizers for the production of $^1O_2$

Sensitizer	Triplet energy, kJ/mol
Methylene blue	140.2 – 142.3[a]
meso-Tetraphenylporphyrin[b] (TPP)	142.3
ZnTPP	
Hematoporphyrin	155.7
Rose bengal	164.1 – 176.6[a]
Eosin	180.8 – 192.5[a]

[a] Depending upon the solvent used.
[b] 5,10,15,20-Tetraphenyl-21*H*,23*H*-porphyrin.

Table 9. Relevant data of rose bengal and TPP

Characteristic	Rose bengal	TPP
M_r	1017.65	614.76
λ_{max}, nm	557 (EtOH)	416 (CH_2Cl_2)
ε_{max}	80 950	478 480
Absorbed Hg-lines, nm	546	405/408
in acidic solution	decomp.	436
Soluble in solvents	protic, polar	aprotic, apolar

The most convenient method for a rapid and reproducible synthesis of crystalline TPP of high purity from low-priced benzaldehyde and pyrrole is described in [24]. The usefulness of the TPP photosensitized $^1O_2$ production in the presence of acids is exemplified by the preparation of intermediates for the technical syntheses of antimalarial agents which are similar to artemisine, a constituent of the traditional Chinese medicinal herb *Artimisia annua L.Compositae*, known as ″Qinghaosu″ [25]. The following reaction scheme shows how the primarily formed hydroperoxides of methylcarvone are immediately transformed into cyclic hydroperoxides by an acid-catalyzed Michael addition. This method of formation of ″saturated″ cyclic peroxides is remarkable because the 1,4-cycloaddition of $^1O_2$ to dienes leads usually to "unsaturated" cyclic peroxides (endoperoxides, see next page).

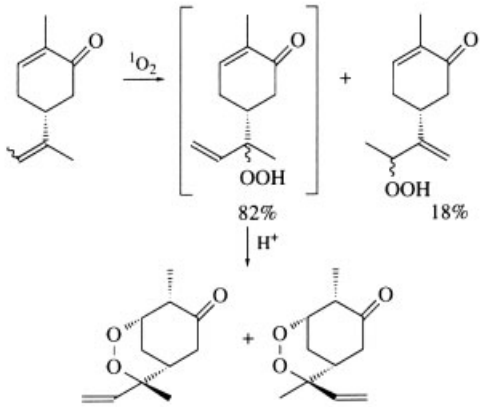

Saturated cyclic peroxides

The irradiated solution contains 11.8 % methylcarvone, 18.0 % CH_2Cl_2, 10 % methanol, 0.14 % *p*-toluenesulfonic acid and 0.02 % TPP. The oxidant is air which is bubbled through the solution. To protect the sensitizer TPP against decomposition due to UV radiation, all UV light is eliminated by a liquid cut-off filter with 90 % absorption at 400nm [26]. A 10 kW mercury medium pressure lamp (Fig. 10) transforms 0.3 – 0.4 kg/h methylcarvone into the peroxide mixture.

Various structural evidence confirms the different reactivities of $^3O_2$ (Type I) and $^1O_2$ (Type II); see the following reactions [6], reproduced with permission:

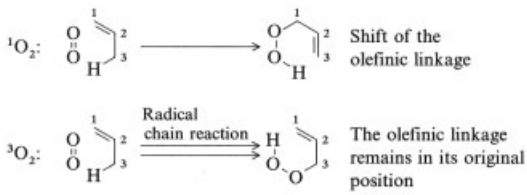

Singlet oxygen is an electrophilic species. Its addition to dienes to generate 1,4-endoperoxides may be viewed as a Diels – Alder reaction with $^1O_2$ as dienophile. In the $^1O_2$ "ene" reaction, olefins containing allylic hydrogens are oxidized to the corresponding allylic hydroperoxides accompanied by 1,2-shift of the double bond. In the final mode, the $^1O_2$ oxygenation of activated double bonds to yield dioxetanes is analogous to the formation of cyclobutanes by a $[2_s + 2_a]$ cycloaddition of two ethylenes.

Singlet oxygenation of conjugated dienes yields different types of endoperoxides depending on the structure of the starting material [6] (the following reactions are reproduced with permission):

From nonhomoannular aliphatic dienes

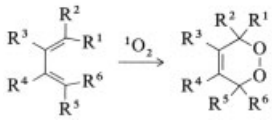

From homoannular aliphatic dienes

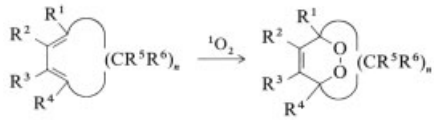

From alkenylarenes and related systems

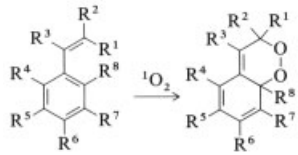

From heterocyclic dienes

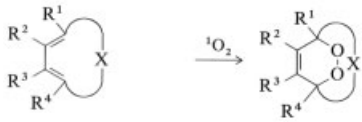

X = O (furan), X = CH_2O (pyran)
X = C(O)O (2-pyrone)

From nonheterocyclic aromatic dienes

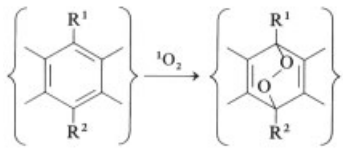

The synthetically useful transformations of 1,4-endoperoxides are as follows [6] (reproduced with permission):

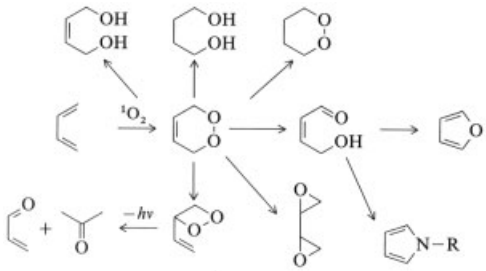

In sharp contrast to Type I photooxidations the $^1O_2$ "ene" reaction proceeds stereospecifically in a suprafacial manner with respect to the "ene" unit, i.e., $^1O_2$ attack and hydrogen removal occur on the same side of the olefin [6] (the following reaction is reproduced with permission):

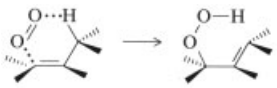

When applying the $^1O_2$ "ene" reaction to pure enantiomers of chiral olefins no racemization is observed. Also (*E*/*Z*)-isomerization does not occur during the course of an "ene" reaction. This stereospecifity makes allylic functionalization by $^1O_2$ attractive to synthetic chemists; its regiospecifity is also useful. Singlet oxygen approaches the substrate predominantly from the sterically less hindered side. Also, in contrast to Type I photooxidations allylic hydrogen abstraction by $^1O_2$ in an "ene" reaction proceeds independently of whether the H atom is a primary, secondary, or tertiary one. As an electrophilic species, $^1O_2$ is sensitive to the nucleophilicity of the olefinic bond. As a general rule, alkyl substitution increases the reactivity of olefins 10 to 100 fold per group. Some of the possible transformations of allylic hydroperoxides, including a silatropic reaction to give an allylic silyl peroxide, include the following [6] (reproduced with permission):

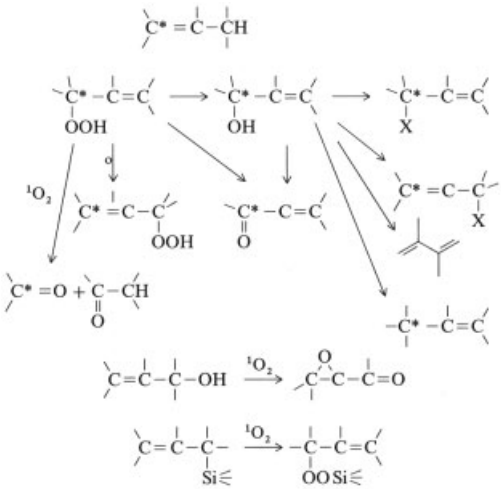

Because of the instability of hydroperoxides in general and of their allylic derivatives in particular, it is common to reduce them to the corresponding alcohols prior to workup and isolation. Unstable hydroperoxides can also be reduced in situ with excess phosphite, especially with $(PhO)_3P$, which reacts quite sluggishly with $^3O_2$ or $^1O_2$, and is in this case the reagent of choice.

The acid-catalyzed heterolysis of the peroxide bond is also depicted in the above scheme; this reaction is known as Hock cleavage and results in the formation of two carbonyl fragments. The base-catalyzed Kornblum – DeLaMare dehydration yielding *α,β-unsaturated ketones*– a synthetically useful transformation–is also shown in this scheme.

A general route to *1,2-dioxetanes* is the cycloaddition of $^1O_2$ to enol ethers, enamines, and other electron-rich alkenes. The following reactions (reproduced with permission) demonstrate the cycloaddition of $^1O_2$ to enamines and the possible transformations of the resulting 1,2-dioxetanes [6]:

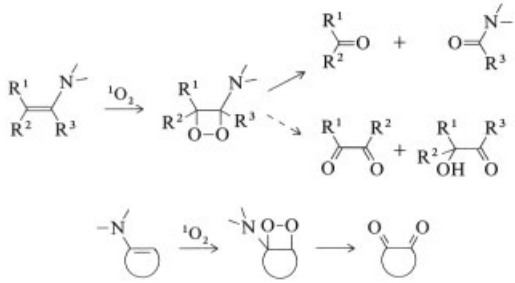

Nonluminescent reactions of isolable dioxetanes proceed via catalytic cleavage to give 1,2-dicarbonyl or 2-hydroxycarbonyl compounds. The O – O bond can also be reduced by insertion of, e.g., trivalent phosphorus compounds leading to phosphoranes.

Photosensitized formation of $^1O_2$ from air is one of the cleanest and most convenient methods for the introduction of oxygen functionalities into unsaturated hydrocarbons. Thus, it is expected that $^1O_2$ will be utilized by industry more intensively in the future. Oxygenations using $^1O_2$ are already applied in the manufacturing processes of fragrant chemicals, e.g., of rose oxide and nerol oxide [6] (the following reactions are reproduced with permission):

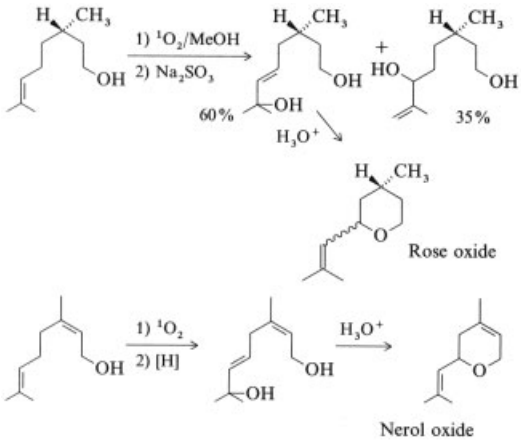

7. Selected Useful Photoreactions

7.1. Enamide Cyclization

A variety of alkaloids and many pharmacologically active compounds related to them have been synthesized mostly via routes such as Diels – Alder reactions. Enamide photocyclization provides a facile and stereoselective approach for forming compounds such as phenanthridine, benzo[c]phenanthridine, aporphine, protoberberine, and yohimbine alkaloids, and others which were previously accessible only from natural sources. Photochemical formation of a typical basic structure is as follows:

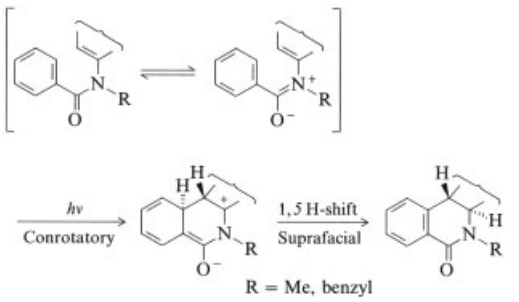

Acrylanilides, enaminoketones, and benzanilides exhibit analogous photochemical behavior. Photocyclization of benzanilides proceeds only under oxidative conditions, e.g., in the presence of I_2 or $^3O_2$.

7.2. Cyclobutane Formation

The photoinduced [2 + 2] cycloaddition of alkenes is important for organic synthesis. The significance of this reaction is derived in part from the stereochemical control for synthesis which enables conjunction of two alkenes to produce a cyclobutane with up to four stereogenic centers:

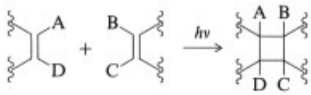

Regioselectivity is a crucial factor in intermolecular [2 + 2] photocycloaddition, because the photoreaction between unsymmetrical olefins can lead to sixteen products. In some systems, photocycloaddition produces a complex mixture of products. It was, however, soon discovered that substituents can control the regioselectivity to some extent. The intramolecular [2 + 2] photocycloaddition offers many advantages over its intermolecular counterpart. In general, the excited state subunit is more efficiently trapped, and constraints introduced by the intramolecular nature of the reaction frequently allow for greater control over the regio- and stereochemical course of the reaction. Therefore, the intramolecularity of the reactions serves to regulate stereoinduction, regioselectivity, and endo – exo selectivity. The versatility of this reaction is seen in representative studies involving syntheses of small, medium, and large rings as well as of bridged and spirocyclic systems such as longifolene, epi-bulnesene, pentalenic acid, epi-precapnelladiene, hibiscone C, isoamijiol, $\Delta^{9(12)}$-capnellene, β-panasinsene, α-acoradiene, and phoracantholide [6]. The [2 + 2] photoadduct can undergo ring expansion, ring contraction, and one- or two-bond cleavage reactions:

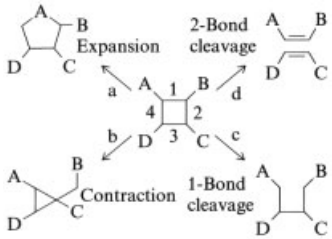

7.3. Oxetane Formation

The light absorbing species responsible for oxetane formation

(Paterno – Buchi reaction) is usually the carbonyl group. A bimolecular reaction is one of the several known paths for nonradiative decay of an excited carbonyl species. Carbonyl compounds produce triplet excited species in high yield on excitation by ISC(S). Due to short singlet lifetimes for carbonyl compounds, triplet states are more likely to be involved in bimolecular reactions. For singlet bimolecular reactions, simple aliphatic aldehydes and ketones are used more than α,β-unsaturated carbonyl compounds. The electronic structure of excited carbonyl groups dictate that the $(n, \pi*)$ reagent acts as an electrophilic radical. Thus, the excited carbonyl can be defined as an "umpolung" reagent, i.e., a reactant having a reverse electronic structure with respect to the functional group that it represents (in this case the carbonyl ground state).

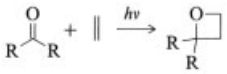

Oxetanes have a variety of potential uses. Photoreactions leading to carefully selected oxetanes are impressive tools in the construction of complex molecules. Among the classes of interesting target compounds are chemicals with pronounced and specific biological activity such as drugs, crop protection agents, and fragrants. Perhaps more interesting to the synthetic chemist is the role of oxetanes as intermediates. The combination of a basic site and ring strain produces the driving force for a variety of selective transformations. Electrophiles (E^+) and nucleophiles (N^-) react with the oxetane yielding ring opening.

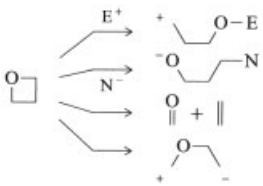

Regioselective cracking can take place; heterolytic cleavage providing an intermediate for reaction with dipolarophiles is also possible.

7.4. Oxadi-π-Methane Rearrangement

Triplet sensitization of β,γ-unsaturated ketones causes them to rearrange to cyclopropyl ketones via a formal 1,2-acyl shift. This photosensitized reaction is designated oxadi-π-methane (ODPM) rearrangement and occurs from the lowest excited $(\pi, \pi*)$ triplet state of the β,γ-unsaturated ketone. The reaction path can be described by a sequence of biradical intermediates:

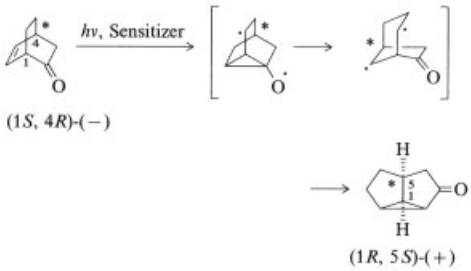

The reactive triplet of the β,γ-unsaturated ketone corresponds to an energy level of ca. 289 – 310 kJ/mol. This means that sensitizers with triplet energies equal to or above 310 kJ/mol must be applied, e.g., acetophenone or 2-propanone, using light with $\lambda > 300$ nm.

Chemical yields are high (86 – 91 %), provided that the conversion of the starting enone is in the range 94 – 96 %. The reaction is most effective in 2-propanone, since the latter serves as solvent and also as sensitizer, making workup easy. However, the substrate concentration in 2-propanone should not exceed ca. 4 wt %, because competitive direct light absorption by the starting material itself leads to an undesired 1,3-shift via Norrish Type I cleavage. Additionally, enhanced light absorption of the starting material lowers the yield of the photoproduct to about 34 % when the 2-propanone solution is irradiated with 254 nm light.

The ODPM rearrangement of bicyclic bridged substrates is enantiospecific. Enantiomeric target materials can thus be obtained by simply choosing the appropriate optically pure precursors to be irradiated. One application is the approach of *cis*,*anti*,*cis*-tricycloundecane type targets e.g., antitumor- and antibacterial-active (–)-coriolin.

7.5. 1,3-Dipolar Cycloaddition

Photolysis of 2*H*-azirines, particularly that of 3-aryl-2*H*-azirines, produces intermediates of the nitrile ylide type which react with numerous electron-deficient alkenes or alkynes as well as with carbonyl compounds, heterocumulenes, and activated esters to yield a variety of five-membered heterocycles [27]. Addition of activated carbon – carbon double bonds to the photochemically generated nitrile ylides forms 1-pyrrolines, whereas addition of carbonyl compounds yields 3-oxazolines:

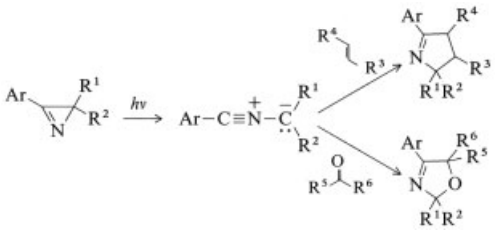

Although the light absorption of 3-aryl-2*H*-azirines above 300 nm is weak, they react rapidly with the aforementioned components when they are irradiated with a mercury medium pressure lamp through pyrex glass. This method is especially advantageous because the photoproducts absorb at shorter wavelengths, and are thus protected from subsequent photoreactions by the pyrex cutoff filter.

8. Photochemical Wastewater Purification

Vacuum Ultraviolet (VUV) Process. The VUV process is limited to the strong UV absorption bands of oxygen, e.g., in air at 190 nm. Excitation in the VUV spectral region leads in most cases to a homolysis of chemical single bonds and became feasible with the availability of excimer light sources (Fig. 11) which emit in this UV range. Besides photohomolysis of the target substance, VUV photolysis of H_2O generates very efficiently hydroxy radicals (Eq. 55) which then attacks the dissolved or dispersed substrate.

$$H_2O \xrightarrow{h\nu} 1/2\,H_2 + HO^{\bullet} \qquad (55)$$

Xe excimer lamps can be used for water photolysis on a preparative scale [28]. Suitable reactors were developed for the VUV production of ultrapure water for pharmaceutical and microelectronic industries. Disadvantages are the short penetration depth of the VUV (several μm) and the maximum electrical power input of 1.5 kW what makes xenon excimer lamps less suitable for ground and wastewater decontamination.

H_2O_2 – UV Process. Efforts using the Hg emission at 254 nm, produced in particular by low pressure mercury lamps, demonstrate that irradiation at 254 nm is not effective in removing organics from water. However, this kind of radiation is quite efficient for disinfection purposes. In contrast, oxidation by a combination of UV light and H_2O_2 implies the generation and subsequent reaction of hydroxy radicals which perform very successfully the desired mineralization of organic pollutants. A suitable technical arrangement is shown in Figure 5. The oxidation potential of hydroxy radicals (2.80 V) is only outmatched by that of fluorine (3.03 V). Although short-lived, the hydroxy radical is capable of oxidizing organic compounds by hydrogen abstraction (Eq. 56). This generates organic radicals which form peroxy radicals by addition of molecular oxygen (Eq. 57). These intermediates initiate thermal (chain) reactions of oxidative degradation, finally leading to CO_2, H_2O and inorganic salts. Therefore, the photooxidative degradation (mineralization) of organic pollutants is accelerated when the temperature of the irradiated wastewater is increased [29]. However, at temperatures above 70 – 80 °C the light output of medium pressure mercury lamps not only decreases but is also shifted to longer wavelengths. Thus, the temperature of the lamp encircling polluted water should not exceed 70 °C.

$$HO^{\bullet} + RH \longrightarrow R^{\bullet} + H_2O \qquad (56)$$

$$R^{\bullet} + O_2 \longrightarrow RO_2^{\bullet} \qquad (57)$$

$$RO_2^{\bullet} + RH \longrightarrow ROOH + R^{\bullet} \qquad (58)$$

$$ROOH \xrightarrow{\Delta} RO^{\bullet} + HO^{\bullet} \qquad (59)$$

In the H_2O_2 – UV process the rates of degradation increase with increasing H_2O_2 concentration and UV light intensity and are highly dependent on the chemical struc-

ture of the substrates. The reactivities of volatile organic halogen compounds are in the order of $Cl_2C{=}CCl_2 > Cl_2C{=}CHCl > CHCl_3 > CH_2Cl_2 >$ (benzene) > chlorobenzene > chlorophenol > $Cl_2HC\text{-}CHCl_2 > CCl_4$. Chlorine is abstracted and converted into chloride, indicating that the chlorinated structures were destroyed. The use of H_2O_2 is advantageous compared to other methods of photooxidative water treatment: The oxidant is available commercially, it can be stored safely, and is infinitely soluble in water. Hydrogen peroxide is a very cost-effective source of hydroxy radicals and the procedure of water treatment with H_2O_2 is simple. A disadvantage is the rather low absorption coefficient of H_2O_2 at 254 nm. A higher percentage of hydroxy radical formation may be obtained using Xe-doped mercury light sources exhibiting a strong emission at 210 – 240 nm, where the absorption coefficients of H_2O_2 are considerably higher. A further disadvantage exists in the efficient trapping of hydroxy radicals by HCO_3^- and CO_3^{2-} ions (Eqs. 60 and 61).

$$HO\cdot + HCO_3^- \longrightarrow H_2O + CO_3\cdot \quad (60)$$
$$HO\cdot + CO_3^{2-} \longrightarrow HO^- + CO_3\cdot \quad (61)$$

Although the generated carbonate radical anions are oxidants themselves, their oxidation potential is less positive than that of hydroxy radicals. By this way a certain selectivity may be introduced with respect to the compounds to be degraded (i.e., those compounds that are less prone to oxidation will not be attacked) [30].

Ozone – UV Process. Ozone has a much higher light absorption than H_2O_2, and the inner filter effects, e.g., by aromatics present in the wastewater, play a minor role. On the other hand, the low solubility of ozone in water together with the necessity to install further apparatus for ozone production represent additional problems in the technical application of the O_3 – UV process. In general, the photolysis of ozone dissolved in water leads to the production of H_2O_2 (Eq. 62)

$$O_3 + H_2O \xrightarrow{h\nu} H_2O_2 + O_2 \quad (62)$$

Finally, by a sequence of reactions including the interaction of organic substrates present in the water, HO· radicals are also formed in the O_3 – UV process.

The analytical control determines the total organic carbon in milligrams per liter treated water. For this purpose, a wastewater sample is led in a stream of oxygen over a catalyst, e.g., Pt/SiO_2. Here, the organic water pollutants are quantitatively oxidized to CO_2. The subsequent determination of CO_2 can be performed in many ways, e.g., with an infrared detector.

9. References

1. E. Müller (ed.): *Houben-Weyl, Methoden der Organischen Chemie,* **vol. IV,** Parts 5a, 5b (Photochemistry, Parts 1, 2), Thieme, Stuttgart 1975.
2. K. H. Pfoertner, *J. Photochem.* **25** (1984) 91.
3. J. C. Scaiano (ed.): *CRC Handbook of Organic Photochemistry,* **vol. II,** CRC Press, Boca Raton 1989.
4. O. Kling, E. Nikolaiski, H. L. Schläfer, *Ber. Bunsen-Ges. Phys. Chem.* **67** (1963) 883.
5. J. C. Calvert, J. N. Pitts, Jr.: *Photochemistry,* J. Wiley & Sons, New York 1966.
6. K. H. Pfoertner in J. D. Coyle (ed.): *Photochemistry in Organic Synthesis,* special publication no. 57, The Royal Society of Chemistry, London 1986.
7. K. Seikel, M, Tragesser in *Abbau industrieller Schadstoffe,* TÜV Rheinland 1993.
8. K. H. Pfoertner, A. Richter unpublished results QVF Glastechnik GmbH, Wiesbaden 1997.
9. K. H. Pfoertner, J. Leisibach, unpublished results 1968.
10. W. Tiedt: "Berechnung des laminaren Reibungswiderstandes konzentrischer und exzentrischer Ringspalte," *Chem. Ztg. Chem. Appar.* **90** (1966) 813; **91** (1967) 17, 299, 569, 968; **92** (1968) 76.
11. Ciba, CH 439 268, 1970 (N. Tarköy, C. Campana).
12. T. Gascard, A. Saus, *J. Phys. E.* **15** (1982) 627.
13. J. C. André: *Industrial Photochemistry I – XVII,* Groupe de Recherches et Applications en Photophysique et Photochimie, Nancy, France. Articles published from 1982 – 1991 in various journals.
14. B. Eliasson, U. Kogelschatz, H. J. Stein, *EPA Newsletter* **32** (1988) 29.
15. P. Esser, B. Pohlmann, H.-D Scharf, *Angew. Chem.* **106** (1994) 2093.
16. P. Kesselring, C. S. Selvage (eds.): *Solar Power Plants,* Springer, Berlin 1986.

17. Deutsche Forschungsanstalt für Luft- und Raumfahrt e.V., Energy Technology Division, Köln 1994.
18. Monsanto, US 4 087 342, 1978 (J. J. Bloomfield).
19. J. N. Pitts, Jr., G. S. Hammond, K. Gollnick (eds.): *Advances in Photochemistry,* **vol. II,** J. Wiley & Sons, New York 1979, p. 305.
20. FR 8 517 436, 1985 (J. C. André, A. Said, M. L. Viriot).
21. Solarchem Research, Division of Brolor Investments, US 4 686 023, 1985 (R. D. S. Stevens); EP-A 0 252 740, 1987 (C. E. Slemon).
22. Hoffmann-La Roche, US 4 551 214, 1985 (K. H. Pfoertner, H.-J. Hansen).
23. A. A. Frimer (ed.): *Singlet O_2,* **vol. II**, CRC Press, Boca Raton 1985.
24. A. D. Adler, F. R. Longo, J. D. Finarelli, *J. Org. Chem.* **32** (1967) 476.
25. K. H. Pfoertner, unpublished results 1994.
26. K. H. Pfoertner, G. Englert, P. Schönholzer, *Tetrahedron* **44** (1988) 1039.
27. A. Padwa (ed.): *1,3-Dipolar Cycloaddition Chemistry,* **vol. 1,** J. Wiley & Sons, New York 1984.
28. G. Baum, Th. Oppenländer, *Chemosphäre,* **51** (1995) 1781.
29. A. Gulde, Diploma Thesis, FH Furtwangen 1995.
30. A. M. Braun in M. L. Viriot, J. André, A. Braun (eds).: *CPIC-ENSIC,* Nancy 1990.

Photoconductive Polymers → Imaging Technology
Photoconductive Polymers → Plastics, General Survey
Photodetectors → Photoelectricity

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Photoelectricity

Standard Article

Manfred Lutz[1]

[1]AEG Elektrofotografie GmbH, Warstein, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a19_599.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. External Photoeffect
3. Internal Photoeffect
4. Applications
4.1. Light Sensors and Detectors
4.2. Solar Cells
4.3. Imaging Systems

Photoelectricity is one of the key phenomena of today's imaging and communication technologies. In general, photoelectricity is understood as the interaction of photons with solids, resulting either in a liberation of electrons (external photoeffect) or in an enhanced electrical conductivity (internal photoeffect). Both effects are discussed particularly with regard to the basic principles and materials involved.

Photoelectricity with its link between light flux and electrical current was the basis for a huge technical and scientific effort directed to enhance, extend, and diversify devices in their functions as light sensors and detectors and later as image pick-up systems which enable transmission, processing, display, and printout of the received optical information; e.g., solar cells and sensor arrays for digital still and video cameras.

[Top of Page]

1. Introduction

Generally photoelectricity is understood as the interaction of photons with solids, resulting either in a liberation of electrons (*external photoeffect*) or in an enhanced electrical conductivity (*internal photoeffect*).

Depending on the spectral range of the interacting electromagnetic radiation various different photoeffects are observed (Table 1).

Table 1. Radiation and photoeffects

Radiation	Photoeffect
High energy γ-radiation, > 1 MeV	photoproduction of electron – positron pairs
γ-Radiation	nuclear photoeffect, Auger effect
Low energy γ-radiation or X-rays	ionization of atoms or molecules
UV-light	external photoeffect
Visible and infrared light	internal photoeffect, photoconductivity

The *external photoeffect* was detected in 1887 by HALLWACHS. The following studies of LENARD (1899 – 1902) on the material and wavelength dependence of this phenomenon and its explanation by EINSTEIN in 1904 gave a significant impulse to the development of modern quantum theory.

Further work on photoemission from metal surfaces resulted not only in a good understanding of the mechanism itself but lead also to applications such as vacuum photocells, photomultipliers, vidicon tubes and in addition to the exciting new analytical method of photoelectron spectroscopy for materials research and surface analysis (⟶ Surface and Thin-Film Analysis, ⟶ Surface and Thin-Film Analysis – Ultraviolet Photoelectron Spectroscopy (UPS)).

The *internal photoeffect* — light impact on a solid state material gives rise to a higher current of electrons — was observed by HITTORF on selenium early in 1852. However, until 1930 these experiments were ignored as being not reproducible and therefore of little scientific value.

Finally, the publications of GUDDEN, POHL, HIELSCH, and STÖCKMANN between 1928 and 1947 draw the attention to photoconductivity as a common property of insulators and semiconductors.

From that time up to now the investigation of photoconductivity has proved to be a useful tool to understand the physics of the solid state.

In parallel to these basic studies a broad variety of technologies and devices utilizing the effect of photoconductivity have been brought into application.

[Top of Page]

2. External Photoeffect

Electrons released by incident light can escape from a solid if the photon energy $h\nu$ is equal or higher than a given minimum energy E_0 which is a material characteristic. Furthermore, the kinetic energy E_{kin} is independent of the light intensity, but a linear function of the photon energy. The energy balance is given by the Lenard – Einstein equation:

$$h\nu = E_0 + E_{kin} \quad (1)$$

E_{kin} represents the maximum kinetic energy observed in the emerging electron flux. It may be lowered by collision processes in the surrounding atmosphere.

The external photoeffect was first observed with metals. Later work showed that certain semiconductors offered much higher photoemissive yield and a response to smaller photon energies than metals.

The minimum energy E_0 which light must possess to liberate an electron depends on the type of bonding encountered. In the case of metals E_0 is practically equal to the electron work function E_w (Fig. 1) which is in the order of several electron volts [3].

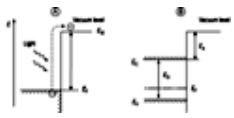

Figure 1. Electron energy level diagram for photoemission of an electron from a metal (A) and a semiconductor (B)

Thus, in the case of metals, the external photoeffect is primarily a phenomenon of the ultraviolet and visible region. The work functions (Table 2) are correlated with the position of the corresponding metals in the periodic table [4].

Table 2. Work functions and threshold wavelength of some metals

Metal	Work function, V	Threshold wavelength, nm
Cesium	1.94	639
Sodium	2.29	541
Barium	2.56	484
Magnesium	3.67	338
Chromium	4.37	284
Tungsten	4.54	274
Copper	4.36	284
Platinum	5.55	223

In semiconducting materials, the existence of a forbidden gap E_g between an almost filled valence band and an almost empty conduction band gives conditions for the photoeffect which differ considerably from the behavior of metals. The most significant feature of semiconductors is the absence of inelastic scattering between electrons. Thus, in contrast to the metallic transport processes, the photoelectrons in semiconductors encounter energy losses only by lattice scattering, i.e., phonon generation. Since much lower energy losses are involved in this case, the photoemissive yield can be high. However, to obtain a high photoemissive yield, a sufficiently small electron affinity E_a is required, which is the energy difference between the vacuum level and the bottom of the conduction band E_c. If the gap $E_g = E_c - E_v$ (E_v top of the

valence band) is reasonably larger than the electron affinity, the probability to escape from the semiconductor is high enough for most of the photoelectrons even after inelastic collisions with phonons. Thus the range in which the electrons are able to escape corresponds to the penetration depth of the incident light (Fig. 1). Therefore, quantum efficiency can be close to the theoretical maximum, if reflection losses of light are taken into account. Compounds with such a high quantum efficiency include several alkali antimonides such as Cs_3Sb and Na_2KSb (see Table 3), and certain I – VI compounds such as Cs_2Te and Rb_2Te. They have practical importance as photocathode materials [4].

Table 3. Band gap E_g, electron affinity E_a, and photoemissive yield of alkali antimonide photocathodes

Material	E_g, eV	E_a, eV	Quantum efficiency, electrons/photons
Cs_3Sb	1.6	0.45	0.30
Na_2KSb	1.0	1.0	0.30
$[Cs]Na_2KSb$	1.0	0.7	0.40

[Top of Page]

3. Internal Photoeffect

The variation of electrical conductivity with the intensity of incident light, called photoconductivity, is the result of an internal photoeffect in which carriers are excited by radiation energy from lower to higher energy levels, supposed the quantum energy is high enough to cause the transition. To observe photoconductivity, however, the number of carriers generated by light must be measurable in comparison to the free carriers normally available in the material. This is the reason why photoconductivity is not observed in metals. As mentioned above, selenium was the first material in which photoconductivity was detected.

However, since no suitable technology was available to purify the material and to grow crystals with defined structure, it took nearly 100 years before an intensive study on the internal photoeffect started. The studies were carried out on crystalline covalent solids such as silicon, germanium, III – V and II – VI compounds, e.g., InSb, GaAs, PbS, CdS, and resulted in a fundamental understanding of the phenomenon, including a description of the role of defects and impurities, which control carrier lifetime and hence photoconductor characteristics. Other important areas of investigation dealt with the complex issues of surface states including chemical and physical adsorption. The studies benefited from the advances in experimental techniques of surface analysis such as X-ray and UV photoelectron spectroscopy and Auger electron spectroscopy.

Subsequently, attention turned to disordered systems such as amorphous solids, molecular solids, molecular doped polymers, and liquids.

All these different materials have in common that the conductivity depends on the number of photogenerated carriers. When the light is switched on, the conductivity at first increases linearly with time until saturation is reached because of the fact that carrier recombination occurs after a lifetime τ.

For stationary illumination of a sample where the light is uniformly absorbed the number of photoelectrons Δn amounts to

$$\Delta n_e = \eta_e \cdot n_{phot} \cdot \tau_e \quad (2)$$

the corresponding change in the electrical conductivity σ is:

$$\Delta\sigma = e_0 \cdot \eta_e \cdot n_{phot} \cdot \tau_e \cdot \mu_e \quad (3)$$

where η_e is the photoelectric yield, n_{phot} the number of absorbed electrons, μ_e the mobility of the electrons and τ_e denotes the time span in which the excess carriers contribute to the photoconductivity before they recombine or are trapped in impurity centers.

Excitation may take place either across the energy gap between the valence band and the conduction band, or alternatively via intermediate energy levels due to the presence of impurities which may be either donors or acceptors.

Photoconducting materials of the first type are known as intrinsic photoconductors (Fig. 2), those of the second type extrinsic photoconductors (Fig. 3) [5]

Figure 2. Band to band transitions in direct (A) and indirect (B) semiconductors

In indirect semiconductors, e.g., c-Si the transition is phonon assisted.

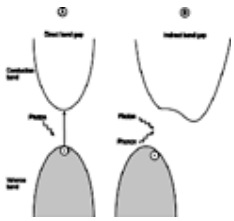

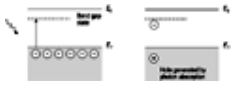

Figure 3. In an extrinsic semiconductor acceptor or donor levels exist in the band gap and photons can cause transitions to these levels. The threshold energy for the transition is normally smaller than the gap.

The long wavelength threshold of most *intrinsic photoconductors* is in the range between 0.5 and 10 µm (Fig. 4). An exception exist with a mixture of HgTe and CdTe compounds to form an intrinsic photoconductor which, depending on composition has an adjustable threshold between 0.8 and 40 µm [5]. The alloy systems GaAsP, InGaAs, and InGaAsP can be tailored to respond at a wavelength range from the visible up to 1.5 µm. But the choice of semiconductor materials for detectors is limited to those that match the available substrates on which the active devices are grown. Substrates with adequate quality, technical sizes, and reasonable costs are Si, GaAs, Ge, and InP. The most widely used detector material for long haul communication is InGaAs for λ = 1.55 µm and InGaAsP for λ = 1.55 µm and λ = 1.3 µm. Both detector materials match very well with InP substrates.

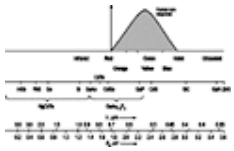

Figure 4. Band gap and cutoff wavelengths of several semiconductors

In *extrinsic photoconductors*, presence of donors or acceptors requires activation energies which are by an order of magnitude, or more, lower than the energy gap, resulting in a sensitivity to medium or far infrared radiation. However, the photoelectric yield is significantly smaller since the impurities must necessarily be diluted by a factor of 10^4 – 10^7 in the host material (Table 4).

Table 4. Ionization energies and cutoff wavelengths for doped germanium and silicon

Semiconductor	Impurity	Ionization energy, eV	Wavelength cutoff for detection, µm
Ge	Au	0.15	8. 3
	Cu	0.041	30
	Zn	0.0333	38
	B	0.0104	120
Si	In	0.155	8
	Bi	0.0706	18
	B	0.044	28

In contrast to crystalline semiconductors the photoconductive phenomenon in amorphous solids, molecular solids, and liquids show quite different features. In the amorphous or disordered state the concept of allowed energy bands is no longer valid. Here, excitations localized on individual atoms or molecules, intramolecular relaxation and electron – phonon interaction play the dominant role. Activities in this field are increasing, not only because of scientific interest but also under the aspects of technical applications for electrophotographic imaging and solar cells, since these materials offer the possibility to manufacture large area photoconductors either by vacuum deposition, chemical vapor deposition, or cast of films from solutions [6].

In order to use photoconductivity as a tool to investigate excitation mechanism and electrical transport parameters experiments have to performed under special conditions or together with other independent experiments. The dependence of the photoconductivity on the wavelength under stationary illumination gives information whether intrinsic or extrinsic transitions are involved. In some cases the number of photoelectrons Δn_e and the drift mobility μ_e whose product appears in Equation 3 can be determined separately by Hall effect and electron spin resonance.

The response of a photoconductor to pulsed or modulated light allows to evaluate the lifetime and the mobility of the majority carrier. A very powerful kind of pulsed photoconductivity measurement is the “time of flight” method which enables the observation of the transit of a group of charge carriers through the photoconductor. Thus from a single experiment of this type the polarity, density, and drift mobility of the charge carriers can be determined. Fig. 5 shows the experimental

configurations mostly used to investigate transient photoconductivity in disordered materials such as amorphous and molecular solids. The sample is sandwiched between electrodes. Through the upper one a short pulse of strongly absorbed light excites a thin layer of carrier pairs. Depending on the direction of the electric field either electron or holes will be drawn into the bulk and if carrier lifetime permits, they will reach the opposite electrode in the transit time t_T. From t_T the carrier mobility μ can be calculated. The time constant of the external circuit, RC' determines whether a rectangular pulse (upper part of Fig. 5 A) or its integral (upper part of Fig. 5 B) can be observed. A rectangular pulse is only expected for semiconductors in which a unique mobility or mean velocity per unit field defines the drifting charge sheet. For disordered systems the analysis of the carrier transport signal as obtained from the experiment is more complex but very still useful to investigate the transport phenomenon. Another experimental configuration of the time of flight method uses a corona charging of the photoconductor instead of a transparent counterelectrode (lower part of Fig 5). This arrangement closely resembles the experimental conditions in typical applications of high resistivity photoconductors such as in electrophotography or vidicon tubes. The surface potential V is monitored by an electrostatic probe. The light pulse which is absorbed in a narrow region below the surface results in the transfer of the charge Q to the bottom contact producing a linear drop in the surface potential V. The transit time can be obtained from either V (lower part of Fig 5 A) or its derivative (lower part of Fig. 5 B).

While pulse excitation is usually easier to implement in practice, step-function excitation (Fig. 5 C) also allows determination of the transit time t_T. Typical transit times are in the order of nanoseconds for amorphous silicon and selenium and in the range of micro- to milliseconds for molecularly doped polymers which are caused by the low mobilities (Fig. 6) [7], [8].

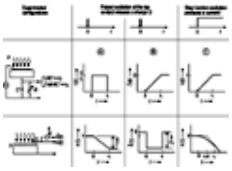

Figure 5. Time of flight method

Unipolar transits: Specimen geometries and idealized waveforms for pulse excitation (A, B) and step-function excitation (C)

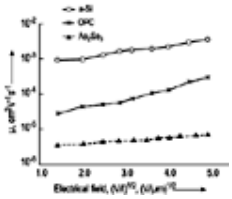

Figure 6. Field dependence of charge carrier mobility (*p*-type) in disordered systems [amorphous As_2Se_3, amorphous silicon and molecularly doped polymer (OPC = organic photoconductor)]

[Top of Page]

4. Applications

Photoelectricity with its link between light flux and electrical current was the basis for a huge technical and scientific effort directed to enhance, extend, and diversify devices in their functions as light sensors and detectors and later as image pick-up systems which enable transmission, processing, display, and printout of the received optical information.

4.1. Light Sensors and Detectors

Table 5 gives a summary of commercial photoelectric devices, their technology, composition, structure and application. It is beyond the scope of this article to describe each of the devices in detail. They are treated extensively in the standard textbooks on electronic or optoelectronic devices [9], [10].

In the following emphasis is being put on solar cells and imaging systems.

Table 5. Light sensors and detectors

Device	Photoeffect	Material	Application
Photocell	external	alkali, multialkali or semiconductor photocathode in vacuum tube	measuring of light intensity short response time ≈ ns
Photomultiplier	external	as above, but higher sensitivity by additional dynode chains for secondary electron emission. Gain up to 10^8	measuring of weak light intensities photon counting, response time ≈ ps – ns
Photoconductor	internal intrinsic and extrinsic	photoconductive material (Fig. 3) between ohmic contacts	measuring of light intensity, visible to far infrared
Photodiode	internal/intrinsic	silicon, germanium *p-n*, *p-i-n*, M-S[a] structures	measuring of light intensity in the photovoltaic mode as solar

			cell (⟶ Photovoltaic Cells) light detector in fiber optics
Phototransistor	internal/intrinsic	silicon *p-n* photodiode as base-collector in a bipolar transistor or MOSFET[b]	measuring of weak light intensities light controlled switch of electrical power optoelectronic coupler
Photothyristor	internal/intrinsic	silicon	switching of high voltage and high currents

[a] M-S = Metal-semiconductor.
[b] MOSFET = Metal-oxide semiconductor field-effect transistor.

4.2. Solar Cells (⟶ Photovoltaic Cells, ⟶ Solar Technology – Photovoltaic Systems)

A solar cells is basically a large-area photodiode operated in the photovoltaic mode, where upon illumination the diode feeds electrical energy into an external load. In principle the solar cell consists of a depletion layer (*p-n* or *p-i-n* junction). *p* (base) and *n* (emitter) regions are connected to the external load by ohmic contacts. In order to achieve maximum illumination of the active zone the front contact to the emitter has to be very small and the free surface of the emitter must be covered by an antireflex coating (Fig. 7).

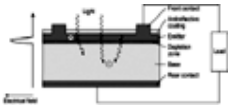

Figure 7. Schematic of a solar cell

Photons, whose energies are equal or higher than the band gap of the semiconductor are absorbed in the cell and generate electron – hole pairs which are separated by the electrical field across the *p-n* junction. Electrons and holes are accumulated in the *p*- and *n*-region giving rise to a photovoltage across the external contacts which is known as open circuit voltage V_{OC} [5] (Fig. 8).

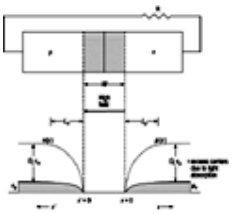

Figure 8. Schematic of a *p-n* diode and the minority carrier concentration in absence and presence of light. The minority charge goes to zero at the depletion region edge due to the high field which sweeps the charge away.

The equilibrium minority charge is p_n and n_p on the *n*- and *p*-sides respectively.

At high light intensities V_{OC} can approach E_g/e_0. In the case of Si solar cells V_{OC} is roughly 0.7 V. A plot of the photocurrent as a function of photovoltage provides the curve shown in Figure 9. The point of maximum power output P_m can be taken from that curve to calculate the conversion efficiency η of a solar cell, defined as the ratio of the output electrical power P_m to the input optical power P_{in}. Efficiency depends quite critically on the match of the band gap E_g to the solar energy spectrum. Photons with energy lower than E_g will not produce electron – hole pairs, whereas photons with energies greater than E_g generate only electrons and holes of the same energy E_g regardless of the amount of excess energy. The excess energy is dissipated as heat and may contribute to the decrease of efficiency by warming up the device.

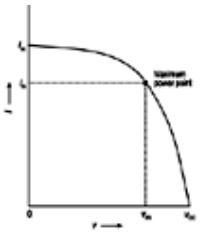

Figure 9. The relationship between current and voltage delivered by a solar cell

V_{OC} = open circuit voltage, I_{SC} = short circuit current

From this point of view GaAs-based solar cells are better matched to the solar spectrum and provide greater efficiencies than Si-based cells. However they are more expensive than Si technology but very useful when high efficiency and stability against cosmic radiation as for example in space applications are required.

Shockley and Queisser [11] calculated in 1961 an upper limit of 30% for the yield outside the atmosphere und 33% inside for unfocussed sunlight. Peak values for c-Si cells from lab production are currently at 24,7% , which is about 75% of the maximum achievable.

Up to now, only solar cells on the basis of c-Si and a-Si have reached the status of commercial industrial production. But yields are to low and costs of production are still to high to compete with fossil energy. It is expected that by using thin-film

technologies production costs can be reduced significantly, assumed that yields remain adequate. First thin-film a-Si and CIS cells from pilot plant promise a yield of 11% and ever higher.

Table 6 gives a summary of efficiencies of commercial cells and lab samples of various materials and technologies.

Table 6. Efficiency of current solar cell technologies

Material	η, %	
	Industrial production	Laboratory production
Crystalline silicon	13 – 17	24.7
Amorphous silicon	7 – 8	13.5
CDTe – CdSe		16.1
$CuInGaSe_2$ (CIS)	11	17.7
GaAs		25.1
GaInP – GaAs, tandem		30.3
GaAs – GaSb, tandem		31.1
TiO_2 – organic dye		10.4

4.3. Imaging Systems

In principle imaging systems consist of a device that captures an image and converts it into electrical signals, processes and transmits the image, and finally displays it or records it on archivable media.

On both sides — image pick up and recording — photoelectric devices contribute as key components (Table 7).

Table 7. Photoelectric devices in imaging systems

Device	Used in
Photodiodes	drum type image scanners
Photocells	
Photomultipliers	
Photoconductive semiconductors materials	target in TV pick up tubes
Highly integrated charge coupled devices (CCD)	one or two dimensional sensor arrays; flatbed image scanners, fax machines, digital copiers, digital photographic cameras, TV-cameras, and camcorders
Photoconductive layers of amorphous chalcogenides and silicon and molecularly doped polymers	photoceptors in office copiers, nonimpact printers, and plain paper fax machines

In highly integrated image pick-up systems for video cameras and digital photo cameras, charge coupled device (CCD) sensors and complementary metal oxide semiconductors (CMOS) are widely used.

The basic steps of a CCD sensor array are the generation of electrical charge carriers when photons are absorbed, the charge collection at each pixel (picture element), the charge transfer from each pixel to the single output amplifier, and the charge measurement for each pixel.

Charge transfer takes place sequentially by shift registers.

CCDs allow the fabrication of image sensors with up to 8 megapixel in digital photo cameras, exceeding in the meantime the resolution of 35-mm silver halide film. Cameras with CMOS sensors go even beyond this, reaching 16 megapixel on an array size of 24 × 36 mm. CMOS technology integrates charge generating, signal amplification, and read-out functions on each pixel. Common for both types of sensors is that a color filter is used on each pixel to obtain the RGB information for image processing. Usually, the ratio of the number of the filters is R=1, G=2, B=1. The only exception is a special CMOS sensor having three separate pn-junctions at different depths in the device. Color separation occurs by the wavelength-dependent absorption coefficient of silicon [12].

With electrophotography — besides ink-jet printing one of the dominating technologies for production of hard copies output (⟶ Imaging Technology – Electrophotography) — the charge pattern formed by imagewise exposure is developed by charged pigment particles, called toner. The toner image is then transferred and fixed by fusing the toner on the paper. The advantage of this successful imaging process relies on the ability of amorphous chalcogenides, silicon, and molecular doped polymers

1. to hold electrical charge in the dark;
2. to discharge exposed areas efficiently; and
3. to be capable of being fabricated into large-area and defect-free films.

[Top of Page]

References

General References

1. W. Wettling, *Phys. Bl.* **53** (1997) 12, 1197 – 1202.
2. Conference Reports: IEEE Photovoltaic Specialist Conference USA, (PVSC) European Photovoltaic Solar Energy Conference and Exhibition, (E.C.PVSEC) International Photovoltaic Science and Engineering Conference, Japan, (PVSEC)

Specific References

3. M. Cardona, L. Ley (eds.): *Photemission in Solids*, General Principles, Springer, Berlin 1978.
4. M. Garbuny, *Optical Physics*, Academic Press, New York 1965.
5. J. Singh, *Semiconductor Optoelectronics, Physics and Technology*, McGraw-Hill, New York 1995.
6. J. Mort, D. M. Pai (eds.): *Photoconductivity and related Phenomena*, Elsevier Scientific Publishing Company, Amsterdam 1976.
7. F. K. Dolezalek, “Experimental Techniques”, in [6]
8. D. Adam, H.-J. Humpert, S. Dreihöfer, H. Pinsler, M. Lutz, Proc. 13. Int. Conf. on Digital Printing Technologies, Seattle 1997, 245 – 247.
9. R. Paul, *Optoelektronische Halbleiterbauelemente*, B. G. Teubner, Stuttgart 1997.
10. J. Seymour, *Electronic Devices and Components*, 2nd ed., Longman Scientific & Technical, Harlow, UK 1988.
11. W. Shockley, H. J. Queisser, *J. Appl. Phys.* **32** (1961) 510. Links
12. R. F. Lyon, P. M. Hubel, *The IS&T Reporter* **17** (2003) no. 6, 1–6

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Photography

Imaging Technology and Holography are separate keywords.

Karlheinz Keller, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Chaps. 1 and 10, Section 5.1)

Helmut Kampfer, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 2.2)

Reinhard Matejec, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 2.3)

Otto Lapp, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 2.4 and 3.1)

Werner Krafft, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 3.2 and 3.2.1)

Hans Frenken, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 3.3)

Hermann Lührig, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 3.4)

Rainer Scheerer, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 4.1 and 4.2)

Max Heilmann, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 4.3, 5, 5.1, 5.2.1, 5.2.2 and 5.2.2.1)

Heinz Meckl, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 4.3, 5.2.2.1, 5.3 – 5.3.2.4)

Peter Bergthaller, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 4.4 and 5.5)

Dirk Hübner, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 5, 5.2.1, 5.2.2, and 5.2.2.6)

Erich Wolff, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 5.2.2.2 – 5.2.2.5)

Bernhard Morcher, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 5.2.2.7)

Wolfgang Zahn, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 5.3.3 and 5.6)

Hans Buschmann, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 6.1, 6.2, and 8.4, Chap. 7)

Rudolf Blank, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 6.1)

Rudolf Tromnau, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Sections 6.2, 8.3, and 8.3.1)

Jürgen Plamper, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 6.3)

Adolf Seiler, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 6.4)

Klaus Nieswandt, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Section 8.2.1)

Immo Boie, Agfa – Gevaert AG, Leverkusen, Federal Republic of Germany (Chap. 9)

Erik Moisar, Universität Frankfurt, Frankfurt, Federal Republic of Germany (Section 2.1)

Robert Winiker, Felix Schoeller jr GmbH & Co. KG, Osnabrück, Federal Republic of Germany (Section 3.2.2)

Matthias Schellenberg, Ilford AG, Fribourg, Switzerland (Section 5.4)

Leendert Ketellapper, Agfa – Gevaert NV, Mortsel, Belgium (Sections 8.1, 8.2.2 – 8.2.7)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_001

1. History 4
2. The Photographic Sensor 6
2.1. Elementary Process 6
2.1.1. Basic Principles 6
2.1.2. The Silver Halide Crystal 7
2.1.3. Interaction with Light 7
2.1.4. Interaction with Ionizing Radiation 8
2.1.5. Theories of Latent-Image Formation 9
2.1.6. Photographic Effects of Exposure . 11
2.2. Spectral Sensitization 12
2.2.1. Theory 12
2.2.2. Blue Sensitization 14
2.2.3. Green Sensitization 15
2.2.4. Red Sensitization 16
2.2.5. Infrared Sensitization 19
2.2.6. Application Technology 19
2.3. Theory of Development 20
2.3.1. Principle Types of Development . . 20
2.3.2. Development as an Electrochemical Redox Process 22
2.3.3. Development as an Electrode Process 23
2.3.3.1. Models of Single-Grain Development Kinetics 23
2.3.3.2. Development Kinetics of Emulsions 25
2.3.4. Development Effects 26
2.3.4.1. Superadditivity 26
2.3.4.2. Adjacency Effects 26
2.3.4.3. Direct Positive Development 27
2.4. Photographic Properties and Their Control 28
2.4.1. Sensitivity, Granularity, Sharpness . 28
2.4.2. Gradation, Development Kinetics, Reciprocity 29
2.4.3. Stabilization 30
3. Manufacture of Photographic Materials 31
3.1. Emulsions 31
3.1.1. Precipitation and Physical Ripening 31
3.1.2. Removal of Soluble Salts 38
3.1.3. Chemical Sensitization 39
3.1.4. Emulsion-Making Technology . . . 40
3.1.4.1. Raw Materials 41
3.1.4.2. Emulsion-Making Equipment 41
3.1.4.3. Test and Measurement Methods for Photographic Emulsions 43
3.2. Emulsion Supports 44
3.2.1. Films 45
3.2.1.1. Properties of Film Materials 45
3.2.1.2. Manufacture of Photographic Film Bases 46
3.2.1.3. Subbings 47
3.2.1.4. Curl 48
3.2.1.5. Antistatic Coatings 48
3.2.1.6. Antihalation Coatings 49
3.2.1.7. Friction-Control and Protective Coatings 49
3.2.2. Papers 49
3.2.2.1. Properties 50
3.2.2.2. Manufacture and Coating 50
3.2.2.3. Packaging, Storage, Transportation, Manufacturers 53
3.3. Coating 53
3.3.1. Coating Machinery 53
3.3.2. Coating Methods 54
3.3.3. Chilling and Drying 56
3.4. Packaging 57
4. Black-and-White Photography . . 58
4.1. Negative – Positive Processes . . . 58
4.2. Reversal Processes 62
4.3. Processing of Exposed Black-and-White Materials 62
4.3.1. Developing Agents 62
4.3.1.1. Inorganic Developing Agents 62
4.3.1.2. Organic Developing Agents 63
4.3.2. Developer Solutions 65
4.3.2.1. Developer Formulas 65
4.3.2.2. Developer Additives 65
4.3.2.3. Special-Purpose Developers 67
4.3.3. Postdevelopment Treatment 68
4.3.3.1. Stopping 68
4.3.3.2. Fixation 68
4.3.3.3. Washing 69
4.3.3.4. Stabilization 69
4.3.3.5. Reduction, Intensification 70
4.3.3.6. Toning 70
4.3.3.7. Black-and-White Reversal Processing 70
4.3.3.8. Replenishment 71
4.4. Silver Halide Diffusion Processes 72
5. Color Photography 72
5.1. Additive Color Processes 72
5.1.1. Screen Color Plates and Films . . . 73
5.1.2. Lenticular-Screen Film (Agfa, 1932) 73
5.2. Subtractive Color Processes 73
5.2.1. Principle 73
5.2.2. Processes with Chromogenic Development 74
5.2.2.1. Color Developers 74

5.2.2.2. Dye-Forming Couplers 75
5.2.2.3. Special Dye Couplers 79
5.2.2.4. Anchoring of Dye Couplers in Layers 80
5.2.2.5. Couplers Added to Developer 82
5.2.2.6. Chromogenic Negative – Positive Processes 82
5.2.2.7. Reversal Color Process 85
5.3. Processing of Color Photographic Materials 86
5.3.1. Replenishment and Rejuvenation . 87
5.3.2. Processes and Formulas for Color Photographic Materials 88
5.3.2.1. Processing of Color Negative Film 88
5.3.2.2. Color Reversal Film Processing . . 89
5.3.2.3. Color Negative Paper Processing . . 90
5.3.2.4. Color Reversal Paper Processing . . 90
5.3.3. Developing Equipment 91
5.4. Silver-Dye Bleach Process 92
5.5. Dye Diffusion Processes 95
5.5.1. The Polaroid Instant Photography Process 95
5.5.2. Kodak and Fuji Instant Photography Processes 97
5.5.3. Agfachrome Speed and Ektaflex . . 99
5.5.3.1. Agfachrome Speed 99
5.5.3.2. Kodak Ektaflex PCT 100
5.5.4. Additive Instant Color Photography Systems 100
5.5.4.1. Instant Color Transparency 100
5.5.4.2. Narrow-Gauge Instant Color Film . 100
5.6. Making Color Prints 100
5.6.1. Printing Processes 100
5.6.2. Exposure Control 101
5.6.3. Types of Color Printers 102
5.6.4. Laboratories for Amateur Color Photographs 103
5.6.5. Copying of Motion-Picture and Television Films 104
6. Photographic Properties and Their Measurement 104
6.1. Sensitometry and Densitometry . 104
6.1.1. Optical Density 104
6.1.2. Characteristic Curve and Evaluation 106
6.1.3. Measurement of Density 108
6.1.4. Sensitivity and Evaluation 109
6.1.5. Sensitometry of Printing 112
6.1.6. Sensitometry of Interimage Effects 114
6.2. Tonal Reproduction and Image Structure 114
6.2.1. Tonal Reproduction 114
6.2.2. Granularity 116
6.2.3. Sharpness 119
6.2.4. Resolving Power 123
6.3. Color Reproduction 123
6.4. Stability of Images 125
6.4.1. Introduction 125
6.4.2. Storage 126
6.4.3. Processing 129
6.4.4. Materials 129
7. Nonphotographic Properties and Their Measurement 132
7.1. Hardening 132
7.2. Dry Properties 132
7.2.1. Brittleness 133
7.2.2. Curl 133
7.2.3. Dry Adhesion 133
7.2.4. Dimensional Stability 134
7.2.5. Friction 134
7.2.6. Dry Abrasion 135
7.3. Wet Properties 135
7.3.1. Swelling 135
7.3.2. Wet Abrasion 135
7.3.3. Reticulation 136
7.3.4. Wet Adhesion 136
8. Photographic Materials and Applications 136
8.1. Photographic Plates 136
8.2. Films 138
8.2.1. Films for Professional and Amateur Photography 138
8.2.2. Motion Picture Films 140
8.2.3. Reprographic Films 142
8.2.4. Microphotography 142
8.2.5. Radiographic Materials 142
8.2.6. Infrared Films and Plates 143
8.2.7. Films for Aerial Photography 143
8.3. Photographic Papers 144
8.3.1. Papers for Continuous-Tone Reproduction 144
8.3.1.1. Black-and-White Papers 144
8.3.1.2. Color Papers 147
8.3.2. Special Applications 152
8.4. Holography and Holographic Recording Materials 154
8.4.1. Introduction 154
8.4.2. Principles of Holography 155
8.4.3. Recording Materials 156
9. Environmental Aspects of Photographic Processing 157
10. Economic Aspects 158
11. References 159

1. History [51–58]

It has been known since ancient times that many substances undergo visible changes on exposure to light. The first attempts to create images with the help of light, however, were reported in 1727 by JOHANN HEINRICH SCHULZE, a German professor of medicine. He glued stencils with letters or words onto a bottle containing a mixture of chalk and silver nitrate, exposed the assembly to sunlight, removed the stencils, and obtained metallic silver images of the stencils in the contents of the bottle. These images could not be permanently fixed in daylight. The first camera images were described in 1802 by an Englishman, THOMAS WEDGWOOD, who used paper or leather moistened with silver nitrate and exposed it in a camera obscura, but these images could also not be fixed.

The earliest surviving photograph is one made approximately 1826–1827 by a Frenchman, JOSEPH NICÉPHORE NIÉPCE. A tin plate coated with photocurable asphalt was exposed for ca. 8 h in a camera obscura, and the unilluminated, uncured areas were removed by washing with a mixture of petroleum and lavender oil (heliography) [51], [53], [56]. In 1829 NIÉPCE commissioned LOUIS JACQUES MANDÉ DAGUERRE to develop the process.

In 1839 DAGUERRE announced the first photographic process having a practicable sensitivity (daguerreotype process). A silver- coated copper plate was treated with iodine vapor to generate a thin layer of light-sensitive silver iodide. After exposure in a camera, the image was rendered visible by treatment with mercury vapor, which formed light-scattering particles of silver amalgam in the areas exposed to light. The image was fixed by bathing in a sodium thiosulfate (hypo) solution, which removed any photosensitive silver iodide still present. At first, the exposure time was ca. 8 – 20 min, but later improvements in the process shortened this to a few seconds.

At about the same time, WILLIAM HENRY FOX TALBOT in England devised a process that he called "photogenic drawing." A paper impregnated with silver chloride, was exposed imagewise for about an hour and black, finely dispersed silver was produced in the illuminated areas (printing-out paper). In 1840 TALBOT created a negative – positive process that could generate several positive prints from a single negative original (calotype). Paper impregnated with silver iodide was exposed for ca. 1 min, developed with a gallic acid – silver nitrate solution, and fixed with a solution of sodium chloride (later replaced by sodium thiosulfate). This negative was made transparent with wax and then used for production of positive copies by exposing silver chloride-impregnated printing-out paper through the negative.

In 1851 – 1880, the "wet collodion process" of the Englishman FREDERICK SCOTT ARCHER found wide acceptance in photography. A glass plate was coated with silver iodide or bromide in cellulose nitrate solution, exposed for ca. 10 s to 1.5 min while still wet, and developed with pyrogallol or iron(II) sulfate. The developing action of iron(II) sulfate was discovered in 1844 by ROBERT HUNT, that of pyrogallol in 1851 by JUSTUS VON LIEBIG. The expensive wet collodion process required a photographic laboratory in a tent on-site, because the plate could not be coated until immediately before use. From 1854 on, a simpler "dry collodion process" devised in France by MARC ANTOINE GAUDIN came into use, but at first the sensitivity of this method was low. The albumen paper created by LOUIS DÉSIRÉ BLANQUART-EVRARD was introduced in 1850; it used egg white to bind the silver chloride grains. Albumen paper gave particularly good tonal reproduction and was favored by professional photographers until around 1930.

A crucial advance took place in 1871 when the English physician RICHARD LEACH MADDOX reported the preparation of a *dry gelatin plate*. The emulsion of silver halide crystals in gelatin is still used in silver halide photography today; at that time, it made the first amateur photography possible by separating the steps of exposure and development. Dry plates were mass-produced since 1877, first in the United Kingdom, later also in other countries (Germany, Austria, the Netherlands) [56].

Flexible backings are easier to handle than plates; paper roll film was introduced in 1884 (GEORGE W. EASTMAN), cellulose nitrate film in 1887 (HANNIBAL GOODWIN), nonflammable cellulose acetate film in 1923, and polyester film, which has a better dimensional stability, in 1955 (Du Pont). A paper backing is still used for photos to be viewed in reflected light; from 1866 to 1969 the paper stock was coated with baryta

(barium sulfate) in order to prevent the image sinking into it. Since 1969, paper stock has been coated with polyethylene on both sides.

Spectral Sensitizers.

The advance of photography depended critically on the discovery of spectral sensitizers by the German HERMANN WILHELM VOGEL in 1873. Unsensitized silver halides are not sensitive to green and red light. Only the discovery of spectral sensitizers enabled the reproduction of all colors in black-and-white photography in a tonally faithful manner; development of spectral sensitizers was also the prerequisite of color photography. VOGEL discovered, e.g., the green-sensitizing action of the triphenylmethane dye corallin. Later research on sensitizers focused on the methine dyes, especially the cyanines. Since 1906, WALTER KÖNIG (Dresden), W. H. MILLS (University of Cambridge), and FRANCES M. HAMER and LESLIE G. S. BROOKER (Kodak) have synthesized and identified many cyanines that have been used as sensitizing dyes. Supersensitization (see Section 2.2.1) was discovered in 1920 by F. FORSTER RENWICK (Ilford).

In 1936 ROBERT KOSLOWSKY discovered that the sensitivity could be enhanced by adding traces of gold salts and thiocyanates to the silver halide emulsion. This "gold effect" made it possible to create the first highly sensitive fine-grained films.

Developers.

At first, exposed silver halide – gelatin materials (plates, films, papers) were developed with an alkaline solution of pyrogallol. Later developers include hydroquinone (WILLIAM DE WIVELESLIE ABNEY, 1880), *p*-phenylenediamine and 4-aminophenol (MOMME ANDRESEN, 1888 and 1891), Metol (ALFRED BOGISCH and JULIUS HAUFF, 1891), and Phenidone (J. D. KENDALL, 1940). The *acidic fixing bath* (sodium thiosulfate, sodium sulfite, acid) now used was devised simultaneously by ALEXANDER LAINER and M. ANDRESEN in 1889.

Around 1940, ANDRÉ ROTT (Gevaert) and EDITH WEYDE (Agfa) independently devised the silver halide diffusion transfer process. During development, unexposed silver halide is dissolved by an Ag^+ complexing agent. These complexes diffuse into an image-receiving layer containing development nuclei, where they form a positive silver image. In 1948, Polaroid (EDWIN LAND) marketed a black-and-white "instant photography" system based on this process, and Agfa and Gevaert created office duplicating materials and offset printing films (→ Imaging Technology, Chap. 2.2.3.).

Color Photography.

The search for true color photography began soon after the invention of photography. Hand-tinting of daguerreotypes and silver halide images were initially used.

The first attempts at color photography did not lead to any practical results. In 1852, HERMANN VON HELMHOLTZ proposed a process based on the color sensation theory of THOMAS YOUNG. JAMES CLERK MAXWELL discovered *additive color photography* in 1855: black-and-white separation images were produced with red, blue, and green filters; the images were converted to transparencies and projected onto the same surface through the same color filters (see Section 5.1). LOUIS DUCOS DU HAURON proposed in 1862 that a screen could be employed for additive color photography. In 1891, GABRIEL LIPPMANN discovered interference color photography in Paris. In this process, a silver halide – albumen coating on a glass plate is exposed while in contact with a mirror-like mercury surface which results in interference phenomena.

The first successful color process was the additive grain screen process, represented by the Autochrome plates (1907) of LOUIS and AUGUSTE LUMIÈRE, the Agfa mosaic plate of 1916, and the Agfacolor mosaic film of 1932. This process was based on the ideas of MAXWELL (additive process) and DUCOS DU HAURON (mosaic). Another additive process, using a lenticular screen, was proposed in 1896 by JULIUS R. E. LIESEGANG and since 1908 further developed by R. BERTHON. Kodak in 1928 and Agfa in 1932 brought lenticular films for amateur cinematography on the market. The Berthon – Siemens color motion-picture projection process – also based on a lenticular screen – was in use in 1936 for a short time (developed by Siemens & Halske and Perutz).

Virtually all modern color processes, however, are *subtractive*, the image being formed with yellow, magenta, and cyan dyes. The subtractive process was discovered in 1868 by L. DUCOS DU HAURON and independently

in 1869 by CHARLES CROS (see Section 5.2.1). Color motion-picture films using colored washed-out reliefs were manufactured in 1922 (Technicolor; see Section 5.5). Other processes based on the principles of washed-out and swollen reliefs were the Pinatype, Duxochrome, and Kodak dye-transfer processes.

The bleach-out process, in which nonlightfast dyes selectively absorb incident light and are thereby decolored, was proposed by LIESEGANG in 1889. It attracted the attention of several researchers (e.g., EMILE VALLOT, KARL WOREL, and RICHARD NEUHAUSS), but never led to a practicable method.

Among the processes utilizing the principle of image formation by dye destruction, only the silver-dye bleach process has become commercially important (see Section 5.4). The principle, proposed by KARL SCHINZEL (1905) and J. H. CHRISTENSEN (1918), was the basis for Pantachrom-Tripofilm (Agfa, 1939), Gasparcolor (according to BELA GASPAR, 1942), the Cibachrome and Cibacopy materials (Ciba-Geigy, 1963) today Ilfochrome (Ilford).

Today's subtractive color films and papers, however, are chiefly based on the principle of chromogenic development, in which the dyes in the image are produced when oxidized color developers react with "dye couplers" (RUDOLF FISCHER, 1911). The first commercial product was Kadachrome narrow-gauge film (1935), which employed dye couplers in the processing baths and controlled diffusion from dye-destroying baths (LEOPOLD MANNES and LEOPOLD GODOWSKY Jr.). The simplified second Kodachrome process, based on selective re-exposure, followed in 1938. Agfacolor transparency film appeared on the market in 1936. Here hydrophilic dye couplers are embedded in separate layers of the material so that they cannot diffuse (GUSTAV WILMANNS and WILHELM SCHNEIDER). The corresponding Agfacolor negative film was marketed in 1939 and Agfacolor paper in 1942. Kodacolor paper and film were introduced in the same year; oil droplets containing the couplers are emulsified into the layers of these materials (EDWIN E. JELLEY, PAUL W. WITTUM). The Ektachrome and Ektacolor films (Kodak), also employing this principle, followed in 1946 and 1948.

Virtually all color films and papers now on the market incorporate hydrophobic couplers in the emulsion.

Today's black-and-white and color materials may utilize inventions dating back several decades, but their sensitivity, sharpness, granularity, color reproduction, and durability far surpass those of the earlier materials. Important contributions include core – shell emulsions (EBERHARD KLEIN and ERIK MOISAR, Agfa, 1962) for high-speed films; the use of development-inhibitor-releasing (DIR) couplers to improve granularity, sharpness, and color reproduction (KEITH E. WHITMORE and coworkers, Kodacolor II film, 1972); and the creation of tabular-grain emulsions to improve sensitivity and granularity (J. T. KOFREN and coworkers, Kodak, 1981).

In 1963, Polaroid introduced a novel subtractive color material (Polacolor) based on a *dye diffusion transfer process*. In the development step, embedded color developers in the exposed regions are oxidized to nondiffusing quinones, while the dye developer from the unexposed regions migrates into an image-receiving material. Kodak in 1976 and Fuji in 1981 followed with a dye diffusion material based on dye-releasing redox compounds (see Section 5.5).

After 1945, color printing materials containing no silver halides were also devised (→ Imaging Technology, Chap. 3.3.4.2.).

2. The Photographic Sensor

2.1. Elementary Process

2.1.1. Basic Principles

In all conventional photographic materials, silver halide microcrystals (grains) serve as both the light-sensitive sensors and the elements for the storage of optical information. Silver atoms are formed upon absorption of electromagnetic radiation or interaction with ionizing particles; these atoms eventually combine to form small nuclei of silver, the latent-image specks. When grains bearing such specks come into contact with a reducing agent (the developer), they are rapidly and completely reduced to metallic silver. Unexposed crystals do not bear such latent-image specks; nevertheless the reduction

of these crystals starts after a considerable delay and leads to the formation of fog. Under usual development conditions, the minimum size of a silver nucleus required for immediate developability is approximately four silver atoms [59], [60]. For larger silver aggregates, the "critical size" r of the aggregate and the redox potential ΔE of the developer can be related by a relation similar to the Gibbs – Thomson equation:

$$r = \frac{2 \cdot \sigma \cdot V_{\mathrm{Ag}}}{F \cdot \Delta E} \tag{2.1}$$

This relation has been confirmed experimentally [61], [62]. Here σ is the specific surface energy of the aggregate, V_{Ag} is the molar volume of silver, $\Delta E = E_{\mathrm{redox}} - E_{\mathrm{Ag^+/Ag}}$ is the difference between the redox potential of the developer and the electrochemical potential of silver, and F is the Faraday constant. It seems certain that a relation similar to Equation (2.1) can appropriately describe the redox properties of extremely small silver clusters down to the size of latent-image specks. However, bulk properties such as radius and surface energy lose their physical meaning for particles consisting of just a few atoms and must be replaced by atomistic properties such as binding energy and dissociation energy [63], [64].

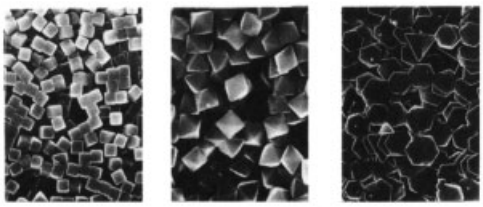

Figure 1. Electron micrographs (SEM) of typical silver bromide microcrystals
Left to right: cubes, octahedra, tabular crystals

2.1.2. The Silver Halide Crystal

Silver bromide and *silver chloride* form lattices with the face-centered cubic (fcc, rock salt) structure. Their equilibrium shape, therefore, is cubic with (100) faces. Under certain conditions, silver bromide in particular may appear in octahedra bounded by (111) planes, for example when crystal growth occurs in the presence of excess bromide ions which are mainly adsorbed to (111) faces [65][66][67]. Crystal planes of even higher order have been observed [68]. If twinning due to the appearance of stacking faults occurs during nucleation, anisotropic shapes such as tabular crystals will be obtained (see Fig. 1) [69–71].

Silver iodide is never used alone in photographic materials, but always in mixed crystals with silver bromide. Silver iodide exhibits trimorphism: below 147 °C, it crystallizes either as fcc γ-AgI or as hexagonal β-AgI, depending on the growth conditions. A phase transition to body-centered cubed (bcc) α-AgI occurs above 147 °C.

Temperature-dependent lattice vibrations give rise to point defects of the Frenkel type in AgBr and AgCl. Some of the silver ions are displaced from their lattice positions to form interstitial silver ions (Ag_i^+) and silver ion vacancies (Ag_v^-), both of which are mobile. The presence of Ag_i^+ is one of the essentials for the photographic process. The concentration of these two charge carriers in silver halide microcrystals is much higher than the equilibrium value in the bulk silver halide. They are also not uniformly distributed in the small crystals. There is an excess of interstitials near the surface, offset by excess vacancies in the surface. Adsorption effects or doping of the crystals with, for example, multivalent ions can further change the distribution of the Frenkel defects [72], [73].

2.1.3. Interaction with Light

The absorption of a photon of sufficiently high energy (> 2.6 eV) liberates an electron from a halide ion and the electron is promoted from the valence to the conduction band (→ Semiconductors). During its lifetime, the photoelectron thus created (Fig. 2 A) has a mobility of ca. $50\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$. After losing an electron, the halide ion is converted to a neutral halogen atom. The excess unbalanced charges of the surrounding lattice silver ions induce a positive charge at the site of the halogen atom; the resulting positive hole h^+ is mobile by virtue of electron exchange with adjacent halide ions.

$$\mathrm{Br^-} \xrightarrow{h\nu} \mathrm{e^-} + \mathrm{h^+} \tag{2.2}$$

The hole mobility of $1.7\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ is, however, much lower than that of the photoelectron.

After this primary electronic step, an ionic process can take place, in which the photoelectron combines with an interstitial silver ion to form a silver atom.

$$Ag_i^+ + e^- \longrightarrow Ag \quad (2.3)$$

After several subsequent absorption events, the successively produced silver atoms can aggregate until a silver cluster of sufficient size – the latent-image speck – has formed. Ideally, every photon absorbed should lead to the formation of one silver atom. Since about four silver atoms would be enough to build up a latent-image speck, a grain should consequently be expected to become developable after the absorption of four light quanta. This requires, however, that loss reactions must not occur, that all silver atoms formed must gather at one site only, and finally that the silver aggregate must be located at the surface and not in the interior of the grain, because the developer reacts only with latent-image specks located at the grain surface.

In practice, these requirements are fulfilled only in part . Loss reactions do occur, e.g., by recombination of photoelectrons and positive holes by the reverse reaction of Equation (2.2). This can happen if a charge carrier is temporarily immobilized at a trap (Fig. 2 C) and a charge carrier of opposite sign approaches it because of, for example, Coulomb forces. A hole may also oxidize a previously formed silver atom or silver aggregate; in either case, the direct or indirect result is the loss of a photoelectron and a decrease in the quantum efficiency of the photographic process. Irreversible removal of holes is therefore of paramount importance.

A more-or-less random fine dispersion of photolytic silver in the silver halide grains, instead of a concentration at one (surface) site, is another possible cause of inefficiency in the photographic process. A considerable increase in speed occurs after doping of the crystal surface with "sensitivity centers" (see Section 3.1.3) such as Ag_2S or $(Au, Ag)_2S$. The increase in speed is accompanied (or, rather, caused) by the capability of such centers to direct latent-image deposition to their site [74].

Silver halides absorb light only in the short-wavelength range. Their intrinsic sensitivity is thus restricted to UV or blue light. It can be extended to the entire visible spectrum and beyond, into the IR, if certain dyes (spectral sensitizers) are adsorbed to the grains (see Section 2.2 and Fig. 2 B). The dye absorbs light and is converted to an excited state. This ultimately leads to the formation of a photoelectron in the conduction band of the silver halide and causes photolytic silver formation just as if an absorption event had taken place in the intrinsic absorption band of the silver halide. From thermodynamic considerations it follows that an energy of ca. 1 eV corresponding to a wavelength of ca. 1200 nm is required to break the Ag–Br bond under standard conditions; this represents the long-wavelength limit of spectral sensitization.

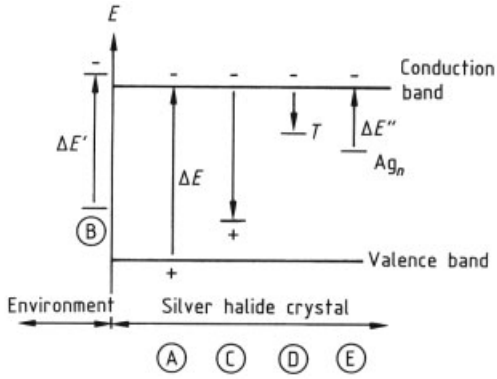

Figure 2. Energy-band diagram of photographic effects
A) Absorption of light ($\Delta E \geq 2.6$ eV) in AgBr with the formation of a photoelectron and a positive hole; B) Absorption of light ($\Delta E' < \Delta E$) in a spectral sensitizer with electron transfer from the excited state to the conduction band of AgBr; C) Recombination of a conduction-band electron with a trapped positive hole; D) Capture of a conduction-band electron by an electron trap; E) Herschel effect: Absorption of light ($1.2\,\text{eV} \leq \Delta E'' < \Delta E$) in a silver cluster leads to its ionization and decay

2.1.4. Interaction with Ionizing Radiation

Ionizing Particles. During passage through matter, ionizing particles continually lose energy along their path (the rare case of direct collisions with atoms is not considered here). The energy loss increases with the charge of the particle and decreases with its velocity. The energy transferred to the material may cause ionization; in the case of silver halides, electrons will be liberated, and these will react with silver interstitial ions in the same way as photoelectrons after light absorption. Special photographic materials (nuclear track materials) with a very high vol-

ume concentration of sensitive grains are used to record the tracks of ionizing particles. Information on the identity and energy of the particles can be derived from the number of developed grains per unit path length and the particle scattering [75].

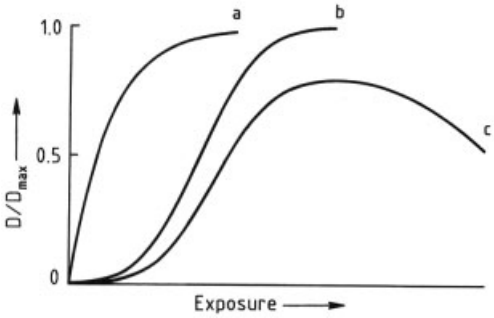

Figure 3. Characteristic curves (density after development versus exposure)
a) Exposure to high-energy radiation (e.g., γ or X-rays); b) Exposure to light; c) Exposure to light with solarization at high exposure values

High-Energy Photons. High-energy photons such as X- and γ-rays, are rather poorly absorbed by silver halides. If, however, an absorption event occurs, a large amount of energy is transferred. Successive ionization events follow, and many secondary electrons are produced. These may escape from the grain which has absorbed the high-energy photons and render adjacent crystals developable as well. The response of a photographic layer to high-energy radiation therefore differs from the response to light (Fig. 3): to become developable by exposure to *light*, i.e., to form a latent-image speck consisting of several silver atoms, a grain must absorb several photons. The response curve (*characteristic curve*), which relates the optical density after development (or the fraction of developable grains) to the exposure, thus has a threshold below which the exposure is not sufficient to make any grains developable. In contrast, the very first *high-energy photon* absorbed by a grain releases so many electrons that latent-image specks are formed and the grain becomes developable. While the characteristic curve for light exposure is sigmoidal, the corresponding curve for exposure to high-energy photons increases from the origin and displays no threshold.

2.1.5. Theories of Latent-Image Formation

The processes starting with exposure and terminating in latent-image formation are not directly observable in detail. Theories have therefore been devised that attempt to describe the steps involved in the formation of a latent-image speck with the aid of known solid-state data on the silver halides along with certain assumptions. Broadly and chronologically speaking, there are *atomistic theories*, which chiefly focus on reaction pathways (how the process takes place), and theories that try to explain why certain events occur and others do not. The second group of theories treat latent-image formation in the framework of *phase-formation* terminology.

Atomistic Theories. The first approach based on solid-state physics was published in 1938 by Gurney and Mott [76]. They assumed that silver halide crystals contain electron traps that can capture photoelectrons (see Fig. 2 D). The initially uncharged trap thus acquires a negative charge, which attracts an interstitial silver ion. The latter reacts with the trapped electron to form a silver atom, and the trap (T) is reset for the next trapping event:

$$(\mathrm{T}) + \mathrm{e}^- \longrightarrow (\mathrm{T}\cdot\mathrm{e})^- \quad (2.4\,\mathrm{a})$$

$$(\mathrm{T}\cdot\mathrm{e})^- + \mathrm{Ag}_\mathrm{i}^+ \longrightarrow \mathrm{Ag}\,(\mathrm{T}) \quad (2.4\,\mathrm{b})$$

$$\mathrm{Ag}\,(\mathrm{T}) + \mathrm{e}^- \longrightarrow \mathrm{Ag}\,(\mathrm{T}\cdot\mathrm{e})^- \quad (2.4\,\mathrm{c})$$

$$\mathrm{Ag}\,(\mathrm{T}\cdot\mathrm{e})^- + \mathrm{Ag}_\mathrm{i}^+ \longrightarrow \mathrm{Ag}_2\,(\mathrm{T}) \quad (2.4\,\mathrm{d})$$

At a certain size, the growing silver aggregate becomes a latent-image speck. Traps can be point or structural crystal defects, but the most efficient ones are assumed to be the sensitivity centers (Ag_2S, etc.) introduced during chemical sensitization. The original theory has been repeatedly modified [77], [78], and several experimental observations seem to agree well with the concept of traps. Examples are (1) the topographic relationship between the location of the sensitivity centers and the site of latent-image deposition [74] and (2) the decrease in photoelectron lifetime after the incorporation of sensitivity centers by chemical sensitization [79], which can be taken as evidence for faster

electron removal. A closer look at this theory, however, reveals certain undesirable details: (1) The negatively - charged transient species occurring in Equations (2.4 b and 2.4 d) could attract highly mobile positive holes instead of slower silver interstitial ions, with resulting recombination and electron loss (Fig. 2 C). (2) If the sensitivity centers are regarded as traps, then increasing their number could give rise to competition in electron trapping, which could lead to dispersed deposition of silver atoms and growing nuclei, so that more photons would have to be absorbed before one of the nuclei could reach the minimum size of a latent-image cluster. This would result in a waste of photons and hence a loss of sensitivity.

In the *Mitchell model* the initial presence of traps is not assumed [80–82]. Electrons are neither accepted by small silver aggregates (up to Ag_3) nor by sensitivity centers. The latter, however, can absorb positive holes. This has two consequences: eliminating the hole lowers the hazard of recombination, and the transfer of the positive charge of the hole to the sensitivity center favors the reaction of an electron with an interstitial silver ion. Once an Ag_3 has been formed, it adsorbs an interstitial and becomes an Ag_4^+. Because of its positive charge, this concentration speck (and the successively larger ones) is now the first deep electron trap. A silver cluster grows at this site through alternating electron trapping and neutralization by interstitial silver ions.

In these two theoretical models, a charged species appears sooner or later; this species enables the growth of a silver phase through its Coulomb attraction for charge carriers with the opposite charge. In contrast, MALINOWSKI [83] assumes that no charged species occur at all. Photoelectrons immediately combine with interstitial silver ions to form neutral silver atoms, while holes form neutral complexes with silver ion vacancies. Silver atoms and hole complexes are mobile; the hole complexes diffuse to the surface, where they release a halogen atom to the environment, and the silver atoms migrate randomly over the surface until they gather into larger aggregates. This event takes place preferentially at condensation centers; the sensitivity centers can be regarded as playing this role [84]. The Malinowski model is a combination of atomistic and phase-formation models.

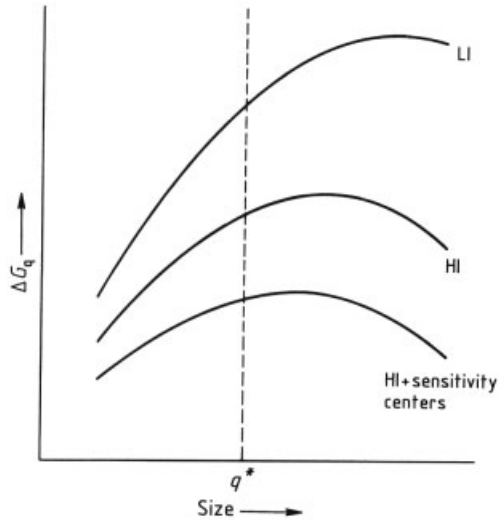

Figure 4. Free energy change vs. aggregate size
Curves LI and HI are for low and high exposure intensity, leading to low and high supersaturation. An aggregate of size $q*$ may be sufficient to cause developability; the energies of formation are then represented by the vertical distances from the abscissa to the intersection of the respective curves with the dotted vertical. The energy of formation decreases with increasing supersaturation (increasing exposure intensity) and is further reduced at some foreign condensation centers, such as sensitivity centers

Phase-Formation Concept. Here it is assumed that the formation of silver in the silver halide matrix is not essentially different from other processes in which a new phase forms and grows (e.g., condensation from vapor or nucleation from solution or melt) and can be described in principle by the same formalism. In all these cases, phase formation requires the existence of a supersaturated host phase. In silver halides, the supersaturation is created by exposure, which immediately increases the electron concentration from its extremely low equilibrium level [85] to a very high value.

Silver nuclei will be formed in the supersaturated system as described by Equation (2.3). The activation energy ΔG_r for the formation of a nucleus of size r depends in general on the supersaturation c_s:

$$\Delta G_r = 4\pi r^2\sigma - \frac{4}{3}\pi r^3 \frac{RT}{V_{Ag}} \ln c_s \qquad (2.5)$$

(For an explanation of the remaining symbols and the applicability of such equations to microscopic systems, see the discussion on Eq. 2.1).

The activation energy of silver formation is reduced at sensitivity centers (Fig. 4). The rates of the steps leading to the formation and destruction of silver are proportional to exp $(-\Delta G/kT)$. It then follows that the formation probability for a silver aggregate of given size increases with increasing exposure intensity (or supersaturation) and that, at low exposure intensities, not only does the formation probability decrease but – even worse – the decay probability increases. The decay of silver aggregates results from their increasing tendency to become oxidized by positive holes. This is the reason for the occurrence of low-intensity reciprocity failure, which can be reduced if the holes are eliminated by other means (Fig. 5) [86–88]. The lower activation energies at sensitivity centers lead to preferential and concentrated silver deposition at these sites. This fact explains why sensitivity centers increase the speed and also why latent-image specks are preferentially deposited at these centers.

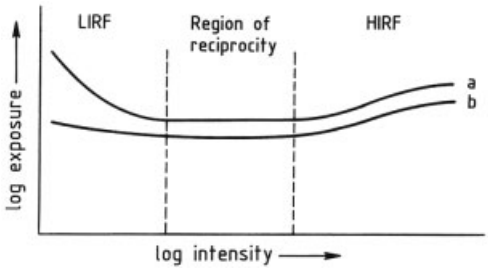

Figure 5. Reciprocity diagram showing log exposure required to reach a given density after development versus log intensity
Curve (a) exhibits three regions: low-intensity reciprocity failure (LIRF) at the left, where the isodense deviates from the horizontal, the horizontal section in the reciprocity range; and high-intensity reciprocity failure (HIRF) at the right. Addition of a halogen acceptor reduces the extent of LIRF (curve b).

2.1.6. Photographic Effects of Exposure

Exposure (to light) is defined as $I \cdot t$, where I is the intensity (irradiance) and t is time. Exposure is measured in energy per unit area, e.g., J/m^2, $W \cdot s/cm^2$, or lux · t, or less frequently (and in scientific work) in incident or absorbed quanta per unit area or per grain. The latter unit is most significant for describing the performance of photographic systems, though measurement is experimentally difficult. The response after development is the *optical density D*; to a first approximation, this is proportional to the fraction of developed grains. The response curve *(characteristic curve)* is obtained by plotting the density versus log exposure. Generally, three regions can be distinguished. At low exposure, no response is observed. As the exposure increases beyond a threshold value, the number of developed grains increases, finally reaching a saturation value (compare Fig. 3 where, however, the exposure axis is not logarithmic).

Certain deviations from this response behavior are known:

Reciprocity Failure (Schwarzschild Behavior). Reciprocity means that a given total exposure (equal to $I \cdot t$) always leads to the same response, i.e., to the same density, independent of the individual values of I and t. Reciprocity failure occurs when, at extreme values of I or t, the total exposure required to reach a given density is higher than in the reciprocal range. Reciprocity can be illustrated by plotting the exposure that results in a given density versus the intensity (Fig. 5) [89]. Reciprocity is obeyed in the region where the curve (the isodense) runs parallel to the abscissa. Simultaneously changing the intensity by a factor f and the exposure time by a factor $1/f$ does not affect the density value in this region.

At low intensities, an increasingly higher exposure is necessary to obtain the given density (*low-intensity reciprocity failure,* LIRF). LIRF is most commonly attributed to the decay of a nondevelopable predecessor of a latent-image speck before the next arriving photoelectron permits growth to the next larger aggregate. This decay is not, or not to a significant extent, a spontaneous process caused by the intrinsically limited lifetime of very small silver aggregates:

$$Ag_n \longrightarrow Ag_{n-1} + Ag_i^+ + e^- \quad (2.6)$$

Instead, it is initiated by a hole (i.e., a migrating bromine atom):

$$Ag_n + h^+ \longrightarrow Ag_{n-1} + Ag_i^+ + Br_L \quad (2.7)$$

which directly oxidizes a silver atom, reduces the size of the silver cluster by one silver atom, and is converted to a lattice bromide ion (Br_L); or it is caused by a bromine atom which is formed

when a positive hole migrates to the crystal surface:

$$4\,Ag_L + h^+ \longrightarrow 1/2\,Br_2 + Ag_i^+ \quad (2.8)$$

For charge compensation, an adjacent lattice silver ion (Ag_i) is released to the lattice as an interstitial silver ion. Lattice ions, although bearing a charge, appear uncharged with respect to the lattice because of the compensating action of adjacent lattice ions with the opposite charge (e.g., Br_L). The bromine atom liberated according to Equation (2.8) then reacts with a silver aggregate at the surface:

$$Ag_n + 1/2\,Br_2 \longrightarrow Ag_{n-1} + Ag^+ + Br^- \quad (2.9)$$

LIRF can therefore be suppressed by elimination of holes or bromine atoms by adding a halogen acceptor (a compound such as a hydrazine that reacts with bromine) to the layer (Fig. 5 B).

Reciprocity failure at high intensities (HIRF) is attributed to the simultaneous formation of silver aggregates at numerous sites, particularly at sensitivity centers [90]. It therefore occurs chiefly when the grains are doped with sensitivity centers having a high degree of dispersion.

Solarization. At high exposure levels, the characteristic curve sometimes bends downward (Fig. 3 C). This behavior results from the destruction of latent-image specks by photolytically liberated halogen at extended exposure times according to Equation (2.9). Solarization occurs when some of the photoelectrons are consumed by silver formation in the interior of the grains. Photolytic halogen can then appear in excess at the grain surface, leading to the oxidation of previously formed latent-image specks. The phenomenon of solarization is the basis for one kind of *direct positive materials*. These contain sensitivity centers in the grain interior (core – shell grains) [91], [92] and chemically produced silver specks at the grain surface. These prefogged (i.e., prereduced) grains are therefore immediately developed without exposure. After exposure, an increasing number of grains become undevelopable because the initially present silver fog specks are more and more oxidized by photolytic halogen. The developed density thus decreases with increasing exposure; a direct positive image results [93].

Herschel Effect. The density due to exposure or chemical fogging decreases after a postexposure to long-wavelength light. Phenomenologically similar to solarization, this effect is based on a different mechanism: Long-wavelength radiation is absorbed not by the silver halide but by previously formed silver aggregates (Fig. 2 E)

$$Ag_n \xrightarrow{h\nu} Ag_{n-1} + Ag_i^+ + e^- \quad (2.10)$$

This leads to their ionization, the transfer of an electron to the conduction band, the formation of an interstitial silver ion, and a reduction in the size of the silver aggregate, which ultimately disappears after sufficient exposure.

For a more detailed description and a listing of other effects, see [94], [95].

2.2. Spectral Sensitization

2.2.1. Theory [96]

Silver halides are sensitive to ultraviolet and, depending on the halide composition, blue light. The long-wavelength limit of spectral sensitivity lies at ca. 420 nm for AgCl, 470 nm for AgBr, and 520 nm for AgI.

The process of extending the sensitivity to long wavelengths is called "spectral sensitization." It involves the addition of dyes (spectral sensitizers) that adsorb to the silver halide. Various classes of sensitizers absorb in the blue, green, red, and infrared parts of the spectrum.

Essentially all photoconductors, whether organic or inorganic, can be spectrally sensitized. The classes of dyes employed depend on the type of semiconductor and the application.

All the spectral sensitizers used with silver halide materials are polymethine dyes [97] (→ Methine Dyes and Pigments), and of these only the ones with cationic (cyanines) or neutral chromophore systems (merocyanines). Dyes with an anionic chromophore (oxonols) are not suitable because they have the same negative charge on the crystal surface as the silver halide (negatively charged due to adsorption of excess halide ions at the surface) and are thus poorly adsorbed.

The preconditions for good adsorption are a planar structure of the dye molecule and the presence of heteroatoms with lone electron pairs, such as nitrogen, oxygen, sulfur, selenium, and tellurium. Suitable heterocyclic terminal groups are pyrroline, pyridine, oxazole, thiazoline, thiazole, selenazole, tellurazole, 1,3,4-oxadiazole, 1,3,4-thiadiazole, and heterocycles derived from these by annelation. In addition to these basic nuclei, the *merocyanines* have, as a second terminal group, at least one "acidic" heterocycle with an active methylene group. Examples of these heterocycles are thiohydantoin, thiooxazolidone, and rhodanine (2-thioxo-4-thiazolidinone). Polymethine dyes with other heterocyclic terminal groups have been described in the literature, but these have not gained much importance as spectral sensitizers.

The sensitizing dye, when adsorbed to the silver halide, has an absorption maximum ("sensitization maximum") wavelength ca. 30 – 50 nm longer than that of single molecules in the dissolved state. Cyanine dyes with cationic chromophores of planar conformation and hindered rotation of the heterocyclic terminal groups, however, tend to form associations or aggregates [98], whose formation is especially promoted if the dye molecules are adsorbed to surfaces. These aggregates have an absorption spectrum much different from that of the single molecule. The absorption maximum of the aggregate is shifted to either longer or shorter wavelengths relative to that of the single molecule. The *shift toward longer wavelengths* is up to 150 nm. It is accompanied by extremely narrow absorption bands that fall off steeply on the long-wavelength side. The cause of the long-wavelength shift in the absorption maximum is the association of molecules in the aggregate in such a way that the angle between the molecular axis and the aggregate axis is less than 54° (preferentially 30° and 19°) [99]. The absorption maximum is *shifted toward shorter wavelengths* relative to the single-molecule value if the angle between the molecular and aggregate axes is greater than 54°. The maxima shifted to longer wavelengths are called J-maxima; those shifted to shorter wavelengths, H-maxima. Both kinds of aggregate bands are important in spectral sensitization. Dyes with J-maxima are indispensable for the spectral sensitization of color materials because of the steep drop on the long-wavelength side; this is the only way to achieve selective sensitivity to individual spectral regions.

The mechanism of spectral sensitization is explained as follows. The dye is promoted from the ground state S_0 to the excited singlet state S_1 by absorption of a photon. In *negative materials*, electrons are transferred from the S_1 state of the dye to the conduction band of the silver halide, with latent-image formation by reduction. Direct electron transfer with the formation of a cationic semioxidized dye free radical ("dye hole") [100] and coupled transfer with the participation of the anionic semireduced dye free radical [101] are under discussion. Energy transfer by the Förster mechanism (radiationless transfer) is also possible but makes only a minor contribution to spectral sensitization [102]. In *fogged direct positive materials*, electron vacancies are transferred from the S_1 state of the dye to the valence band with oxidation of the silver nuclei and the formation of a positive image. The energy levels of the dye molecular orbitals involved relative to the valence and conduction bands of the silver halide determine whether a spectral sensitizer is suitable for negative materials or direct positive materials [100].

Along with its sensitizing action, every spectral sensitizer has a desensitizing action that is independent of light absorption. The desensitizing effect is governed by the constitution of the dye and its concentration on the crystal surface, the type of silver halide emulsion, and chemical digestion. Differences in desensitization are interpreted as resulting from the energy level of the dye molecular orbitals (electron acceptor action) and the statistical distribution of photoelectrons between the crystal and the dye [101], [103]. Desensitization is measured as the decrease of blue sensitivity (intrinsic sensitivity) of the silver halide emulsion in the presence of the spectral sensitizer.

The term *supersensitization* refers to the increase in sensitivity in the spectral range of a sensitizing dye due to substances having a superadditive action (supersensitizers). These can be sensitizing dyes or colorless compounds. Their effect can result from a variety of causes; one such effect is enhanced quantum efficiency of sensitization [104], [105].

A special case of supersensitization is *potentiation*, in which acid (carboxyl, sulfo, sulfon-

amido) derivatives of thiols, enhance the spectrally sensitized sensitivity [106]. This effect is based on increased adsorption of the sensitizing dye to the silver halide crystal.

2.2.2. Blue Sensitization

Sensitization in the blue range of the spectrum is important for silver halide emulsions that are sensitive only in the ultraviolet or ultraviolet plus short-wavelength blue, such as the silver chloride or silver chloride – bromide emulsions used chiefly for photographic paper materials. Blue sensitizers are also indispensable in the silver bromide – iodide film emulsions.

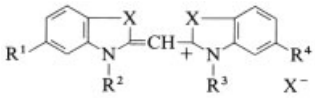

1

X, Y = O, S, Se

R^1, R^4 = H, CH_3, CH_3O, Cl, C_6H_5

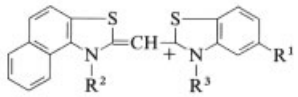

2

R^1 = H, CH_3, Cl,

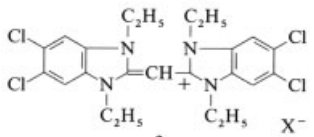

3

Cyanines. The most important cyanines in photography are those with one methine group and a cationic chromophore (monomethinecyanines), such as **1** and **2**.

This group includes, above all, derivatives of benzothiazole (**1**, X=Y=S) with R^1 and R^4=CH_3, Cl, or CH_3O, which are excellent blue sensitizers with a strong tendency to form a J-maximum at 465 – 475 nm. The naphthothiazole cyanines **2** give J-maxima at 480 – 485 nm [107]. These dyes are especially suitable for all color materials (film and paper). In order to prevent residual coloration after photographic processing, dyes with sulfoalkyl residues on the nitrogen [R^2, R^3=$(CH_2)_nSO_3H$] are preferred for their solubility in water.

Special-purpose silver chloride materials can be sensitized with the benzimidazole monomethine dye **3**, whose sensitization maximum is at 440 nm.

Merocyanines. Simple merocynanines (i.e., merocyanines with a zeromethine chain) from rhodanine (**4**–**7**) and open-chain dimethine merocyanines (**8** or **9**) [108] are suitable for blue sensitization of silver bromide and silver chloride – bromide papers in the 450 – 500 nm region.

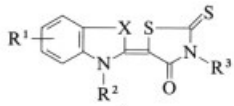

4

X = S, Se, NR^2

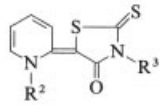

5

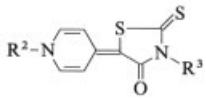

6

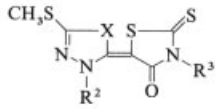

7

X = O,S

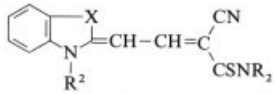

8

X = O, S, NR^2

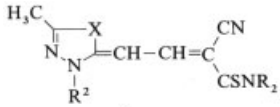

9

X = O, S

Hydrophilic groups such as carboxyl or sulfo groups in the R^2 and R^3 substituents prevent residual coloration after processing that is needed for paper materials.

2.2.3. Green Sensitization

The following kinds of materials require sensitizing dyes for the green part of the spectrum (500 – 600 nm):

1) Reprographic materials (e.g., line, continuous tone, masking, and contact screen films)
2) Direct positive films
3) Black-and-white papers with variable gradation (see Section 8.3.1.1)
4) Black-and-white films (e.g., amateur, professional, aerial-survey, and motion-picture films)
5) Medical X-ray films
6) Color films (negative, reversal, motion-picture, duplicating, and internegative films)
7) Color papers (negative, reversal)

In keeping with the many applications, the number of industrially important dyes is also large.

Merocyanines. Suitable merocyanines for the range up to 550 nm include dimethine merocyanines with five-membered heterocyclic rings or benzoxazole as basic residues, and with rhodanine, thiooxazolidone, or thiohydantoin as acidic residues, e.g., **10–12** [109], [110].

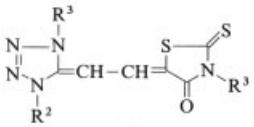

10

R^2, R^3 = CH_3, C_2H_5, C_6H_5
R^4 = C_2H_5, $CH_2CH = CH_2$

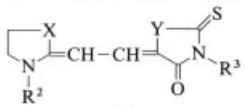

11

These merocyanines are employed for reprographic materials that are processed in red light. Polycyclic merocyanines (**13**) are also suitable for the same spectral range.

X = CH_2, S Y = O, S

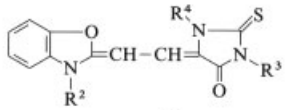

12

R^2 = $(CH_2)_3OSO_3Na$, $(CH_2)_4SO_2NHCOCH_3$
R^3 = CH_3, C_2H_3, C_6H_5
R^4 = CH_2COOH, C_2H_5

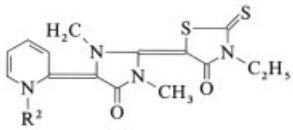

13

Cyanines. The most important green sensitizers in the industry are trimethinecyanines with benzoxazole or benzimidazole groups. The first, with 9-ethyl and 5- chloro or 5-phenyl substitution (**14**), are used in all color materials because they tend to produce narrow J-bands at 540 – 550 nm; they are the sole green sensitizers utilized in color papers.

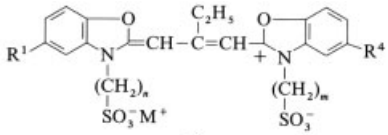

14

R^1, R^4 = Cl, C_6H_5; n, m = 2 – 4
M = Na, K, $(C_2H_5)_3NH$

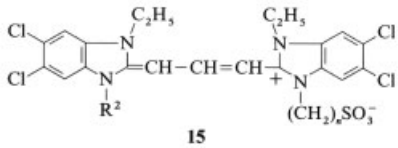

15

R^2 = CH_3, C_2H_5, $(CH_2)_nSO_3Na$
n = 3, 4

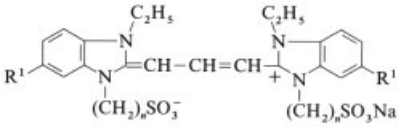

16

$R^1 = CN, COOC_4H_9$
$n = 3, 4$

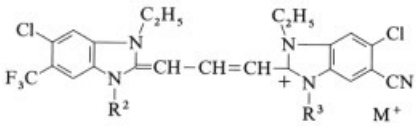

17

$R^2 = (CH_2)_nSO_3^-$; $n = 3,4$
$R^3 = R^2, (CH_2)_2$–C_6H_4–SO_3^-

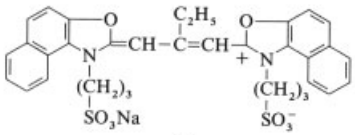

18

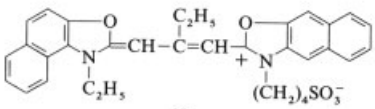

19

In color films, **14** is combined with benzimidazole trimethinecyanines (**15**, **16** and **17**) sensitizing at longer wavelengths (570 – 580 nm), or with naphthoxazole trimethinecyanines **18** or **19** [111], [112]. In **15**–**17**, substituents that lower the basicity of the benzimidazole are essential in order to control fog. Asymmetric trimethinecyanines **20** with two different heterocyclic bases can also be used as long-wavelength green sensitizers in color films.

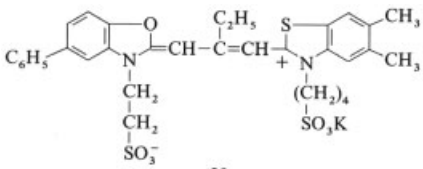

20

The 2,2′- cyanines **21** are suitable for black-and-white (B & W) films provided they are combined with supersensitizers [113], [114]. Only then can high sensitization quantum efficiencies be achieved at 575 nm.

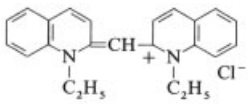

21

To improve the color reproduction, special sensitizers for the blue-green portion of the spectrum are used [112]; **22**, for example, gives a sensitization maximum at 520 nm.

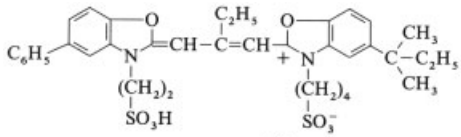

22

Green sensitizers **23** are employed for variable-gamma B & W papers (i.e., B & W papers with variable gradation) that can be rendered "extra hard" to "extra soft" by controlling the proportions of blue and green light [115].

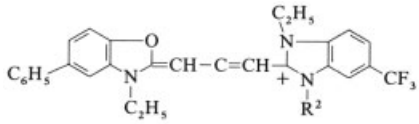

23

$R^2 = (CH_2)_nSO_3^-$; $n = 3,4$

Polymethine dyes that can break down existing silver nuclei by oxidation when exposed to light are used for green sensitization of fogged direct-positive emulsions. Electron-accepting cyanines such as indole dyes **24** [116] or quinoxaline derivatives **25** [117] are suitable for the green part of the spectrum.

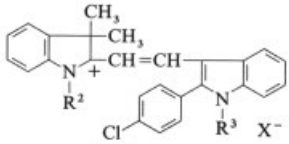

$R^2, R^3 = CH_3, C_2H_5$

24

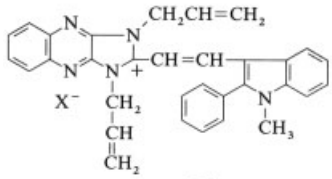

25

2.2.4. Red Sensitization

Sensitizers for the red, from 600 to 700 nm, can be classified by application:

1) Color negative and color reversal films, color reversal papers must be selectively sensitized between 600 and ca. 660 nm.
2) Color negative papers have a maximum red sensitivity at or above 700 nm, which is the region containing the absorption maxima of the blue-green imaging dyes in color negative films. Panchromatic B & W papers are sensitized for the same region.
3) In panchromatic B & W films, amateur films sensitized up to 650 nm are distinguished from technical films, which are sensitive up to the limit of the visible spectrum.
4) Films for aerial photography have panchromatic sensitization, especially in the long-wavelength red or the near infrared (to ca. 750 nm), to achieve a better penetration of atmospheric haze.
5) Graphic-arts films (lith films and masking films) and materials for laser exposure (holography, scanning).
6) Direct positive films.

Suitable red sensitizers are trimethine and pentamethinecyanines having benzo- or naphtho-1,3-thiazole, -selenazole, or -tellurazole [118] as basic terminal groups; tetramethine merocyanine dyes, rhodacyanines [119]; and polynuclear merocyanines [120]. The spectral sensitivity distribution needed for a given application can be produced by choosing the substituents in the methine chain and the heterocyclic nuclei. In the annelated benzene rings, substitution is usually restricted to small groups such as methyl, methoxy, cyano, fluoro, and chloro; substitution in the methine chain of trimethinecyanines to methyl or ethyl groups on the center methine group (the "meso," μ, or 9 position). Bulkier substituents impair the sensitizing action.

Cyanines. The most important red sensitizers for the range up to 660 nm are found among the 9-methyl- and 9-ethyltrimethinecyanines with benzothiazole end groups. The 9-methylmethines, such as **26** or its 5,5′-dimethyl derivative **27**, have the following effects on silver bromide – iodide emulsions: they give J-aggregate bands at 620 – 640 nm and simultaneously H-aggregate bands in the green at 520 – 540 nm, and they also produce a single-molecule sensitization band (monomer band) at 580 – 600 nm.

These dyes are excellent for use in panchromatic B & W films.

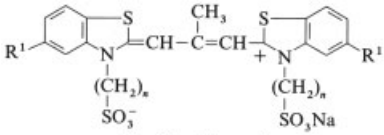

26 $R^1 = H$; $n = 4$
27 $R^1 = CH_3$; $n = 3$

The 9-ethyltrimethines **28** or **29** are characterized by intense, narrow J-aggregate sensitization bands at ca. 650 nm.

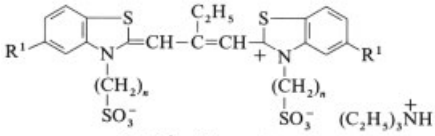

28 $R^1 = Cl$; $n = 3$
29 $R^1 = CH_3$; $n = 4$

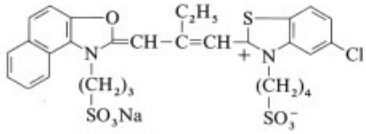

30

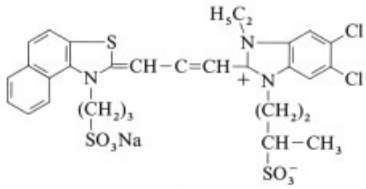

31

These dyes are employed for color films, especially in combination with cyanines **30** or **31** which have two different basic terminal groups, and sensitize in the short-wavelength red between 605 and 620 nm [112]. Such combinations often result in supersensitization and the formation of "mixed" maxima lying between the aggregate bands of the individual dyes [98], [104].

9-Ethylnaphthothiazole and 9-ethylbenzoselenazole dyes, such as **32** and **33**, can be used for sensitization in the longer-wavelength red (660 – 690 nm) in both B & W and color films. These dyes, in combination with sensitizers for the near infrared such as **34**, are of special interest for aerial-survey films [121].

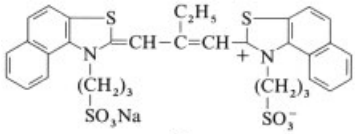

32

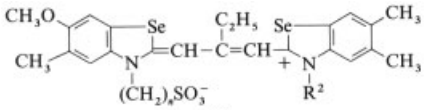

33

$R^2 = C_2H_5$, $(CH_2)_nSO_3Na$; $n = 3,4$

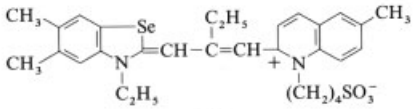

34

Benzothiazole and benzoselenazole pentamethinecyanines impart sensitization maxima between 690 and 720 nm to silver chloride and bromide emulsions. Better photographic sensitivity is attained with pentamethines where the methine chain forms part of a cycloaliphatic ring, such as the cyanines **35** that can be derived from isophorone (3,5,5-trimethyl-2-cyclohexenone) [122].

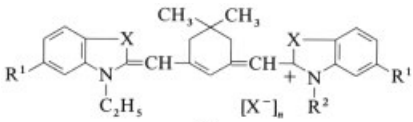

35

X = S, Se
R^1 = H, CH_3, CH_3O, Cl
$R^2 = (CH_2)_3$, SO_3^-; $n = 0$
$R^2 = C_2H_5$ $n = 1$

The sensitivity of materials utilizing these pentamethines (color negative paper, panchromatic B & W paper) can be further enhanced by supersensitization with optical brighteners [123] or heterocycles substituted with thiol groups [124]. This is an important point because the desensitizing action of pentamethines is greater than that of cyanines with shorter methine chains.

Imidazoquinoxaline dyes (**36**) are used as red sensitizers for fogged direct positive emulsions [125].

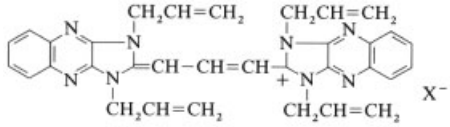

36

Merocyanines. Tetramethine merocyanines were important red sensitizers in color papers for a long time. Dyes such as **37** are currently being used only for color print films [126].

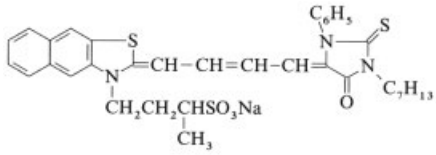

37

Polycyclic Polymethines. Trinuclear merocyanines have been described as red sensitizers, especially **38** for lith emulsions [127] and **39** for thermally developable materials [128].

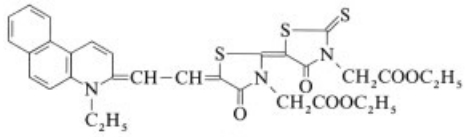

38

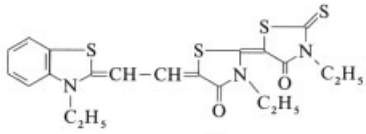

39

Tricyclic allopolar-isomeric cyanines (i.e., cyanines with mero- and cyanine structure) such as **40** are of some value as supersensitizers, chiefly in the red spectral range [129], [130].

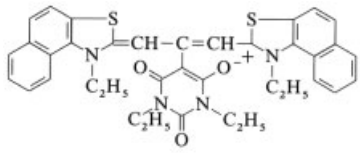

40

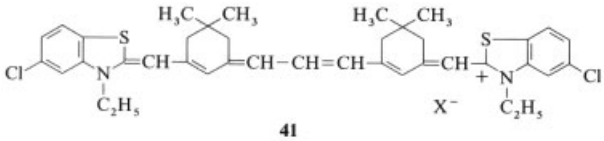

41

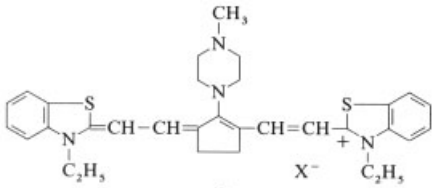

42

2.2.5. Infrared Sensitization

Hepta-, nona-, and undecamethinecyanines are employed for spectral sensitization above 700 nm (to over 1 µm). The infrared sensitivity can be enhanced by incorporating cycloalkenyl residues in the methine chain, e.g., in **41** [131] and **42** [132], or by the use of supersensitizing combinations [133]. Polymethine dyes with longer methine chains are not suitable as sensitizers, since they are unstable and tend to fog (see above).

2.2.6. Application Technology

Sensitizing dyes can be added to the photographic emulsion in various forms:

1) Solution in organic solvents or in water, with wetting agents if appropriate [134]
2) Dispersion obtained by wet grinding in the presence of wetting agents and grinding aids [135]
3) Emulsion prepared by dissolving a high concentration of the dye in a suitable solvent and emulsifying in water with emulsifiers [136]

Organic solvents used include

1) Low-boiling solvents such as methanol, ethanol, or acetone. These are completely removed in the drying operation after emulsion coating.
2) High-boiling solvents with a high solvent power for polymethine dyes and a low toxicity, such as glycols or glycol ethers (e.g., phenoxyethanol).

Dyes can be added at various stages during emulsion making and the photographic properties are influenced by the stage of addition. For example, introducing the dyes at a high temperature promotes epitaxial adsorption to the silver halide crystals with formation of certain absorption bands.

The optimal quantity of dye added depends on the intended application, the size of the emulsion grain, and the molecular structure of the dye. Fine-grained emulsions require a higher concentration per mole of silver than coarse-grained ones. The dye concentration at the crystal surface required for maximum sensitivity is far less than a monomolecular layer. At higher coverages, desensitization takes place (concentration desensitization). This effect is more pronounced the longer the methine chain in the dye. Infrared sensitizers must therefore be employed at especially low concentrations.

In addition to the spectral sensitivity needed for the material in question, a sensitizing dye must meet many other requirements. As much as possible of the dye added to the silver halide emulsion must be adsorbed to the crystal surfaces, and other additives, such as dye couplers, surfactants, stabilizers, antifoggants must not interfere with this adsorption. Sensitometric properties such as sensitivity, gradation, and fog must not be impaired in storage. The sensitizing dyes in multilayer color photographic materials must not diffuse into adjacent layers (This tendency is especially marked in warm, humid climates). After processing, as little residual sensitizer color as possible should remain in the material; i.e., the dyes must be largely washed out of the layers or bleached in the processing baths. For environmental reasons, application of organic solvents, should be largely dispensed with when adding the dyes to the molten emulsion.

Some of these requirements can be satisfied only partly, or at the cost of other problems. Dyes that are well adsorbed to the silver halide crystal and have a low diffusibility, for example, are difficult to dissolve and tend to leave residual coloration. These include cationic cyanines and merocyanines with benzoxazoles, benzothiazoles, benzoselenazoles, and their higher annelated analogues. The introduction of hydrophilic groups such as carboxyalkyl or sulfoalkyl on the nitrogen lowers the residual coloration, and also improves the solubility if the molecule is present as the anion; however, adsorption of the dye to the silver halide is impaired, and the danger of desorption due to other additives is increased. The ranking of parameters in relation to the material gives rise to the molecular design of a great variety of spectral sensitizers.

2.3. Theory of Development [109], [114]

The photographic (primary) process is not favored over other photochemical processes in terms of quantum efficiency; the maximum quantum efficiency φ of the primary process is always ≤ 1 (see Section 2.1). It is the intensification of the photochemical transformation in the silver halide grains by photographic development that lends the photographic process its high sensitivity to light and makes silver halide photography so superior to all other known photochemical processes. Depending on size a grain (silver halide crystals) of a photographic emulsion contains ca. $10^7 - 10^8$ silver ions and the same quantity of halide ions. For the most sensitive grains, the absorption of about four light quanta [137] is enough to make the entire grain developable. The amplification factor of photographic development is thus ca. 10^7.

When a photographic material is developed, those silver halide crystals bearing a latent-image center Ag_n are reduced to metallic silver by a reducing agent (developer). The latent-image centers have a catalytic action:

$$AgBr + Red^- \xrightarrow{Ag_n} Ag + Br^- + Ox$$

where Red^- is the developing agent (generally present as anion in the alkaline developing solution) and Ox is its oxidized form.

Photographic development is thus based on a kinetic effect. Because of the catalytic action of the latent-image centers, photographic developer solutions (see Section 4.3.2) reduce exposed silver halide grains to silver several orders of magnitude faster than unexposed grains, but after a (sufficiently) long time all the grains are developed, even those that not exposed.

2.3.1. Principle Types of Development

There are two distinct development mechanisms; although they occur in practice only as limiting cases, usually a mixture of both mechanisms is found:

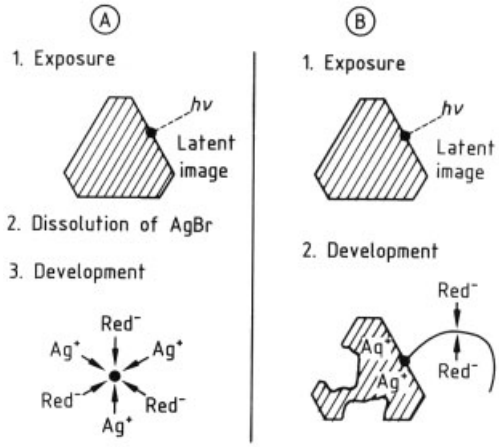

Figure 6. Reactions in physical (A) and chemical (B) development [142]

Physical Development. In physical development [138–141], both silver ions and developing agent from the solution react at the surface of the nuclei. In addition to the nuclei provided by the latent image there are also fog nuclei, whose development leads to the unwanted precipitation of silver at unexposed sites, and metal or metal chalcogenide nuclei deliberately added to the emulsion, for example in the silver halide diffusion process (see Section 4.4).

Two kinds of physical development exist:

Postfixation Physical Development (Fig. 6A). Here, the silver halide is first completely dissolved by complexing agents (see Section 4.3.3.2) and removed from the emulsion layer by washing. The "fixed-out" layer is then treated in a bath consisting of the developer solution and

added silver salt or silver complex. In this process silver formed by reduction is preferentially deposited on the centers formed by exposure (latent image), which are still present in the emulsion layer.

Prefixation Physical Development. Complexing agents are added to the developer solution so that dissolution of the silver halide grain and precipitation of silver at the nuclei in the unfixed layer take place at the same time.

In physical development the silver is precipitated in compact form with high luster and litte covering power [142].

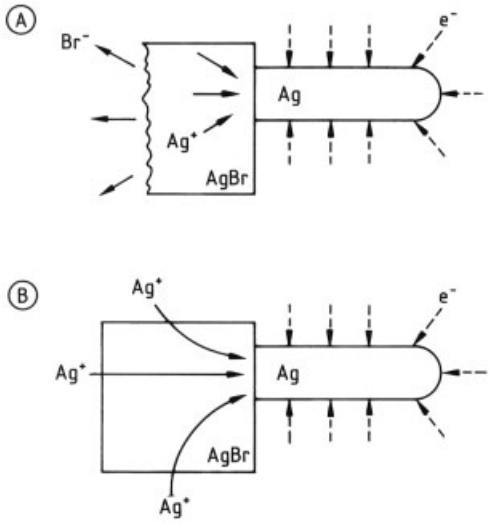

Figure 7. Chemical development
A) Normal case (low silver ion concentration in solution) [144]; B) Ideal case (high silver ion concentration in solution) [148]

Chemical Development. In chemical development (Fig. 6 B), the nucleus is in contact with the surface of the silver halide grain during development. As in physical development, the developing agent is oxidizing at the nucleus – solution interface, but because there is no dissolved silver halide, no silver ions from the solution can be reduced at this interface. Instead, reduction of silver ions occurs at the interface between the nucleus and the silver halide grain [143], [144]. The nucleus grows by silver deposition at this interface; the silver advances into the solution in filamentary form.

In the *normal case* (Fig. 7 A), a quantity of halide ion equivalent to that of deposited silver passes from the surface of the silver halide grain into the solution. Initially, etch pits result on the grain surface; the silver halide grain is completely dissolved in later stages of development. Electron micrographs reveal a jumble of silver filaments and bands; the original grain shape can no longer be perceived. The tangled filaments absorb nearly all the incident light; so little is reflected that the developed silver appears deep black. This deep black form of the developed silver is generally referred to as "Mohr form" after its discoverer [142].

In the *ideal case* (Fig. 7 B) [145], when the concentration of free silver ion in the solution is kept very high by addition of silver salts, the silver ions reduced at the silver nucleus – silver halide interface and thus removed from the grain are quickly replaced by ions from the solution crossing the silver halide – solution interface. The grain is then not dissolved during chemical development, and the mass of silver filaments can be several times as large as that of the silver halide initially present in the grain (see Fig. 8). By a combination of this development technique with a preceding "decoration" (size enlargement by physical predevelopment) of the latent-image specks, it can be shown that the filaments grow out of the AgBr surface like hairs, because the enlarged latent-image specks are found at the tips of the silver filaments.

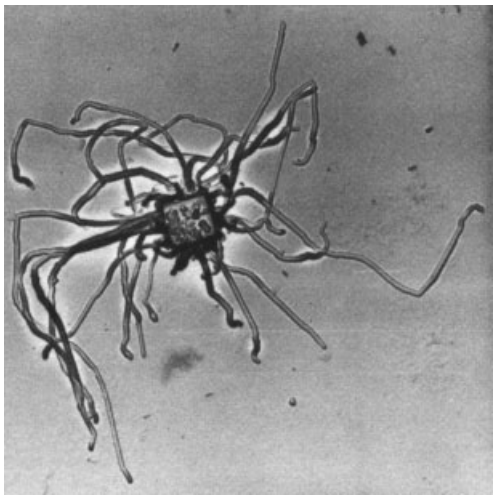

Figure 8. Silver bromide crystals after ideal chemical development [148]

Most developers now in use predominantly employ chemical development. But the presence of sulfite, which is commonly added to the de-

veloper as an antioxidant (see Section 4.3.2), causes a certain amount of dissolution of the silver halide (especially with slow-acting developing agents) and thus effects some physical development before fixing [146], [147]. Table 1 lists the principal types of development and the conditions under which they are employed.

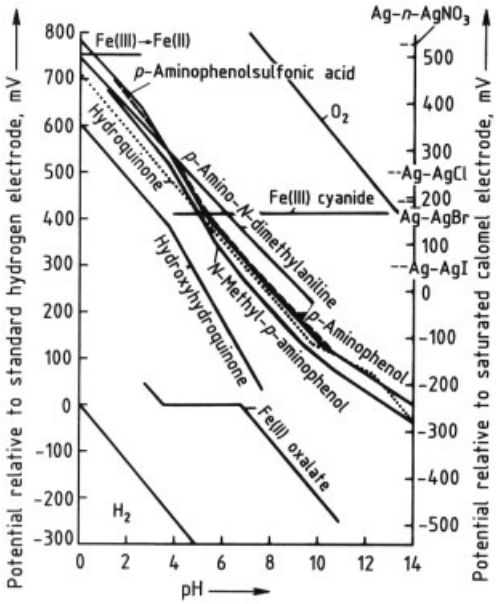

Figure 9. Dependence of standard redox potential of some developer compounds on pH

2.3.2. Development as an Electrochemical Redox Process

Photographic development can best be described as an electrochemical redox process, in which the latent-image center simultaneously acts as anode and cathode [139–141], [148–152].

At the boundary between the silver nucleus (or silver filament) and the developer solution, the developing agent (Red^-) transfers electrons (e^-) to the silver nucleus and is thereby oxidized:

$$Red^- \longrightarrow Ox + e^- \text{ (anode process)}$$

The excess electrons thus arriving at the silver nucleus are accepted by arriving Ag^+ ions:

$$Ag^+ + e^- \longrightarrow Ag \text{ (cathode process)}$$

In physical development, these silver ions stem from the solution phase; in chemical development, they are interstitial silver ions of the silver halide crystals (see Sections 2.1.2 and 2.3.1).

For thermodynamic reasons the redox potential of the developer (U_{redox}, corresponding to the anode process) must be more negative than the Ag^+/Ag potential ($U_{Ag^+/Ag}$) of the cathode process, which depends on the halide concentration via the solubility product of the silver halide.

The electrochemical potential difference

$$\Delta U = \left(U_{Ag^+/Ag} - U_{Redox}\right) > 0$$

is the driving force of photographic development.

The potentials can be determined with the Nernst equation:

$$U_{Redox} = U^0_{Redox} + \frac{RT}{nF} \ln \frac{[Ox]}{[Red^-]}$$

(anode process)

$$U_{Ag^+/Ag} = U^0_{Ag^+/Ag} + \frac{RT}{nF} \ln [Ag^+]$$

(cathode process)

and

$$U_{Ag^+/Ag} = U^0_{Ag^+/Ag} + \frac{RT}{nF} \ln [Ag^+]$$

(cathode process)

Here U_{redox} is the electrode potential of the developer solution, U^0_{redox} is the standard potential of the developing agent, $U_{Ag^+/Ag}$ is the potential of silver in the solution, and $U^0_{Ag^+/Ag}$ is the standard potential of silver.

For the very small silver nuclei of the latent image, the Ag^+/Ag equilibrium potential $U_{Ag^+/Ag}$ does not correspond to that of a massive silver electrode. Because of the large interfacial energy of these small silver particles, the equilibrium potential of the cathode process shifts toward negative values with decreasing particle size [153–155]. Tests with reversible redox systems have led to quantitative relations between this shift in $U_{Ag^+/Ag}$ and the Ag particle size [156–158]. Therefore with more negative values of U_{redox} (at given $U_{Ag^+/Ag}$) not only the rate of silver deposition at the development nuclei increases, but also smaller silver nuclei are developed, so that the light sensitivity of the photographic system increases.

Table 1. Types of photographic development

Process	Addition of silver halide solvent	Addition of silver ion or silver complexes	Anodic and cathodic potential (relative to the Ag^+/Ag-potential of the emulsion layer)
Postfixation, physical development	before development	Ag^+-complex added to developer	positive
Prefixation, physical development	during development	Ag^+-complex formed by dissolution of grain during development	negative
Normal, chemical development	none	none	negative
Ideal, chemical development	none	Ag^+-ions (high concentration) added to developer	positive

As soon, however, as the electrochemical potential difference of the developer system passes a critical value of $\Delta U' \approx 200$ mV, the internal energy of this redox-system is high enough to supply the energy necessary for nucleation so that new silver nuclei can appear. Spontaneous nucleation then begins, and the photographic system becomes spontaneously fogged.

The relation between the development activity of organic developing agents and the pH can be simply derived from the pH dependence of the developer's redox potential. This pH dependence is due to the removal of protons from the developer molecules in alkaline solution. Because of the greater number of single electron pairs, the anions of organic developing agents present at high pH are stronger reductants than the protonized developer molecules present at low pH values [159], [160]. Figure 9 shows the pH dependence of the standard redox potentials of several developing agents.

The pH dependence of the activity of a developer is not always determined solely by its transformation from one ionic form to another. Studies on *p*-phenylenediamines, for example, show that their development activities vary strongly in the pH range from 7 to 13, although no p*K* values exist in this range [161]. This is explained by the pH dependent inhibiting action of the oxidation products [162].

2.3.3. Development as an Electrode Process

2.3.3.1. Models of Single-Grain Development Kinetics

Treating the development process in terms of an electrode process and dividing it into partial processes has led to predictions about the growth kinetics of silver on the single grain [148–152]. The results obtained have been compared with determinations of the volumes of the developed silver particles from electron micrographs [163], [164].

Physical Development. In physical development, anodic and cathodic transfer takes place at the silver nucleus – developer solution interface. The application of electrode theory has given a relation between cathodic and anodic current – voltage curves, the working potential, and the kinetics of physical development (see Fig. 10).

In a model the nucleus can be regarded as a silver electrode connected by an insulated wire to a counterelectrode immersed into the solution. If the solution contains silver ions only, the electrochemical potential at the silver nucleus takes the value $U_{Ag^+/Ag}$ predicted by the Nernst equation and no current flows. If an external emf imposes a more negative potential on the nucleus, an Ag^+ ion current flows to the silver electrode, where the ions are neutralized in accordance with the cathodic current – voltage curve (c in Fig. 10) and silver is deposited.

If the silver nucleus in the same configuration is brought into contact with a solution that contains only the redox system of the developer, the anodic current – voltage curve (a) can be derived in a similar way. The magnitude of the current corresponds to anodic oxidation of developer molecules (Red^-).

In physical development the silver nucleus is in contact with a solution which contains both reaction partners. With respect to silver deposition, it acts as a cathode and must establish the potential $U_{Ag^+/Ag}$ if no external emf is imposed. With respect to the oxidation of the developing agent (Red^-), the silver nucleus acts as an anode and

tends to have the potential U_{redox}. In reality, a steady-state working potential $U*$ comes about; this is the abscissa of the intersection $P*$ of the two current – voltage curves. The corresponding current $i*$ is directly proportional to the silver deposition rate and the formation rate of the developer oxidation product ($dn/dt = i*/e$, where e is the electron charge). The potential difference $U_{\mathrm{Ag^+/Ag}} - U*$ is the driving force of the cathode process, while $U* - U_{\mathrm{redox}}$ is that of the anode process.

The slope of the current – voltage curves depends on the size of the nucleus, among other factors. For diffusion-limited physical development, $dn/dt \sim r$, where r is the radius of the nucleus; since $r \sim n^{1/3}$, the time dependence is $n \sim t^{2/3}$. If the rate is limited by a reaction at the surface of the nucleus (e.g., electron transfer from the developer molecule to the silver nucleus, dehydration of silver ion, dissociation of the Ag^+ complex), $dn/dt \sim r^2$; hence the time dependence for transfer-limited physical development of a silver nucleus is $n \sim t^3$ ([139–141]).

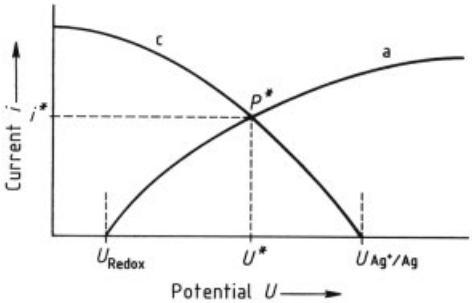

Figure 10. Cathodic (c) and anodic (a) current – voltage curves $i = f(U)$ for physical development [142]

Chemical Development. In chemical development, the outside of the silver center toward the developer acts as the anode (silver – solution interface). In the cathodic process, silver ions of the silver halide crystal lattice are transported through the crystal, via the interstitial lattice, to the nucleus – silver halide interface, where they are reduced to silver (see Fig. 11).

In the anode process the limiting step can be, e.g., the electron transfer or the diffusion of developing agents to the electrode (or the diffusion of developer oxidation product from the electrode); in the cathode process, it can be the electron transfer or the diffusion of silver ions in the crystal, the escape of halide from the crystal, or adsorption. In addition, the anode area increases in proportion to the silver deposition during filament growth, while the cathode area remains constant. It can be derived from this model that filament growth should follow an exponential function [165]. Kinematic investigations of development on specially grown large, tabular AgBr grains have shown, however, that the rate of silver filament growth is very often rather constant over time (see, e.g., [166]) after an induction period (see below). This is explained by the assumption that the constant inward migration of Ag^+ ions through the AgBr crystal to the Ag filament/AgBr crystal contact (i.e., the cathode process) becomes the rate-limiting step [139–141].

Regarding the proportion of physical and chemical components in development as a function of development conditions, see [167].

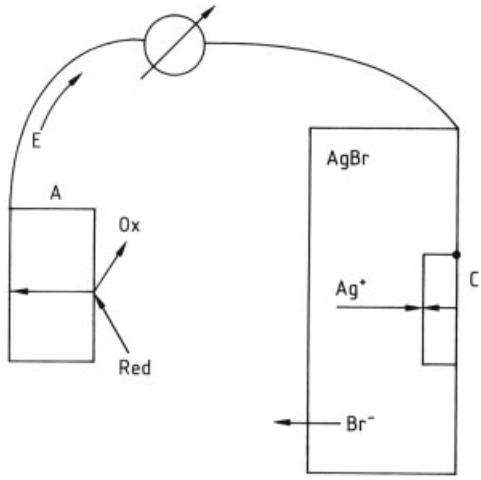

Figure 11. Electrochemical model of chemical development [152]
A=anode; C=cathode; E=electron; Red=reduced form of developer; Ox=oxidized form of developer

Induction Periods. Whether the development of a single grain begins immediately or after an induction or "latency" period depends on the developer and the process conditions. This behavior is accounted for by prior reactions that must first overcome a potential barrier created by the ionic charge of the developing agent and the electrostatic field around the silver halide grain [168]. The silver halide crystal in the emulsion layer is generally surrounded by a shell of

Br^- ions generated during manufacture. Therefore, the potential barrier becomes higher as the negative charge of the active developing agent Red^- increases. Another explanation for the occurrence of an induction period is the transition from spherical to filamentary growth of the nucleus [169]. Induction periods after short exposure to light with high intensity, whereby many very small nuclei are generated, are generally longer than after long exposure to light with low intensity, which produces fewer but larger latent-image centers (see Section 2.1.5). This implies that differences in the induction period can also be due to differences in the size or size distribution of the latent-image centers.

2.3.3.2. Development Kinetics of Emulsions

From the photographic standpoint, development rate of emulsions is defined as

1) The growth rate of the optical density D at a given exposure or
2) The rate of increase of gradation (i.e., of γ; see Section 6.1).

The first definition is more useful in basic studies of development kinetics in the emulsion. The rate of increase of optical density is described according to Hurter and Driffield (see e.g., [170]):

$$\frac{dD}{dt} = K \cdot (D_\infty - D)$$

where D is the optical density at time t, D_∞ that after infinitely long development, t is the development time, and K is a rate constant given by

$$K = k \cdot [\mathrm{RedH}]^a \cdot [\mathrm{OH}^-]^\beta$$

where k is another rate constant and a and β are exponents with values between 0 and 1.

In attempts to work out a detailed description of the development mechanism, the quantities k, a, and β have been determined under widely varying development conditions (cf., e.g., [168], [171], [172].

Mathematical models have also been devised for the development kinetics of emulsions [173–176].

The principal factors governing the development kinetics of emulsion layers are

1) The nature and concentration of the developing agent
2) The development temperature
3) The pH
4) The nature and concentration of special compounds in the developing solution (sulfite, halide, accelerators, restrainers, developer oxidation products)
5) The nature and properties of the silver halide grains in the emulsion (halide composition, grain size, packing density)
6) Exposure conditions (high- or low-intensity, etc.)
7) The properties of the layer, such as thickness, swelling, hardening, and contents of silver halide, color coupler, plastics etc.

Any emulsion layer is composed of very many silver halide grains. Two limiting cases can therefore be described (see Fig. 12):

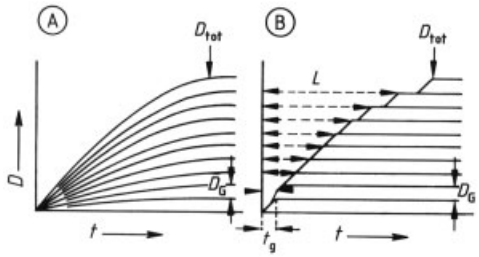

Figure 12. Increase in optical density D of an emulsion versus time, for (A) simultaneous development of all grains; (B) fast development of individual grains differing in induction period
D_{tot}=total density of the layer, D_G=single-grain density contribution, t_g=development time of the individual grain, L denotes several induction periods

1) Each grain is developed immediately at the start of development; there is *no induction period*. The time dependence of the process for this collective of grains is essentially determined by the (average) time dependence of the development of the single grains.
2) Any one grain is completely developed very rapidly, but the individual grains have *widely differing induction periods*. the development kinetics of the layer is then essentially governed by the statistical distribution of these induction periods.

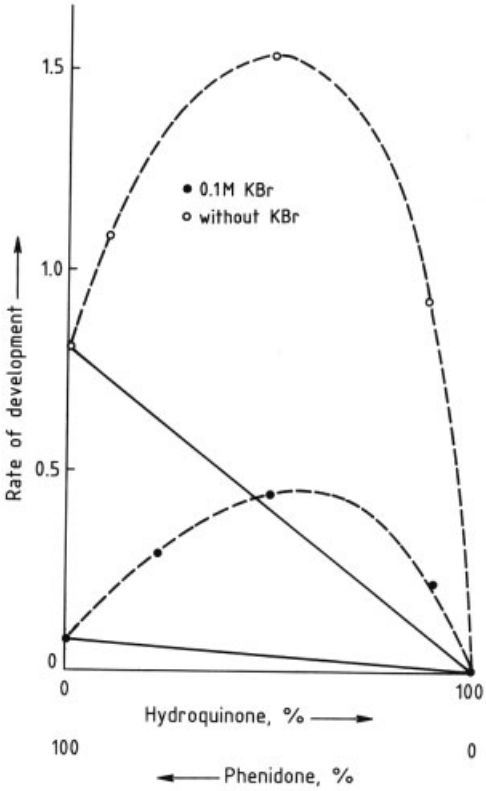

Figure 13. Rate of overall development versus proportion of the developing agents hydroquinone and Phenidone, for a constant total concentration of 0.01 M [179]
Solid lines=expected curves (additive behavior); Dashed curves=experimental curves

2.3.4. Development Effects

2.3.4.1. Superadditivity

A mixture of two developing agents exhibits superadditivity if a measured quantity (e.g., development rate, or optical density after a given exposure and development time) of the mixture has a higher value than that would be obtained as the sum of these quantities for the unmixed substances times their molar fractions.

An example is given in Figure 13. The tie lines between the development rates measured with the unmixed (100 %) developing agents represent the expected "normal" values for mixtures. Any values lying above these tie lines are superadditive [177–179]. If developer mixtures show superadditivity, they generally do so in both chemical and physical development [180].

The following explanations have been proposed for the mechanism of superadditivity:

The induction period of developer ions B, bearing a highly negative charge (e.g., hydroquinone, pyrogallol, catechol, or ascorbic acid in alkaline solution) is shortened by the less negatively charged or neutral developer molecules A because these are less sensitive to the negative charge barriers around the silver halide grain (see Section 2.3.3.1). In this way the development rate is enhanced [181], [182].

Developing agent B, which acts more slowly and/or with an induction period, regenerates faster developing agent A. For example, a slow electron transfer from the more active reducing agent B to the silver can be accelerated by the weaker reducing agent A (i.e., A acts as a redox catalyst) [183–185]. For example in the case of a superadditive mixture of hydroquinone and Phenidone, the reaction between the hydroquinone anion and the Phenidone radical cation is assumed to be the rate-limiting step [185].

In the superadditive hydroquinone/Phenidone mixture at pH$\geq$10.5, the Phenidone eventually also displaces adsorbed SO_3^{2-} ions that may block electron transfer from hydroquinone to Ag [185].

2.3.4.2. Adjacency Effects

The structure of a photographic image can be heavily influenced by adjacency effects [186], [187]. These effects, also called Eberhard effects after their discoverer, are particularly marked when

1) The developer concentration is low
2) Only a limited quantity of developer is available (e.g., in the development of instant pictures with viscous developers; see Section 4.4)
3) The developer solution has little pH buffering capacity

In color photography, development inhibitor releasing (DIR) couplers (see Section 5.2.2.5) are used to produce very strong adjacency effects in the color layers. Horizontal (edge) effects improve image sharpness, vertical (interimage) effects improve color quality.

Localized diffusion fluxes and restricted transport of developing agent and (development-inhibiting) byproducts are necessary for these adjacency effects (see Fig. 14).

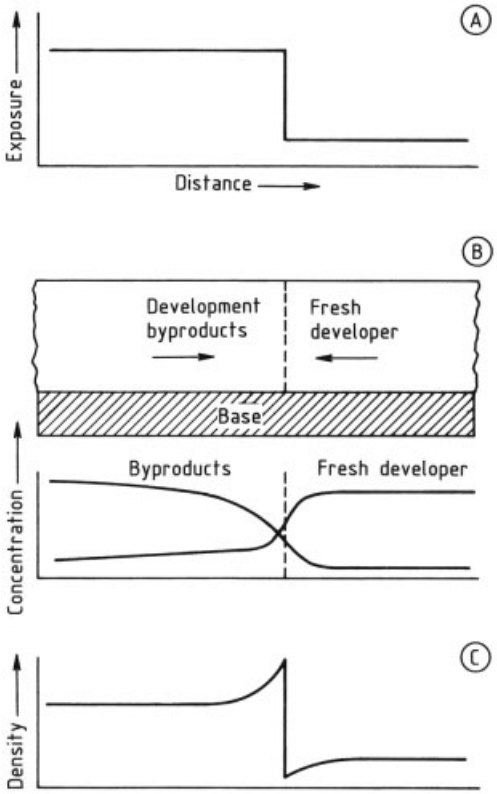

Figure 14. Horizontal adjacency effect (edge effect) due to developer depletion and inhibiting byproducts [19, p. 609] For explanation see text.

If the emulsion is exposed to an image having a detail with an edge (Fig. 14 A), then during development the developer becomes depleted in the exposed region of the layer, while halide or other development-inhibiting substances are produced there. As a result, development is retarded in this region.

Near the edge of the exposed region, however, fresh developer diffuses in from the adjacent unexposed part of the layer while inhibiting byproducts diffuse out. Mass transport between the layer and the developer solution is too slow to maintain the same concentrations of developer and inhibiting compounds in this edge region as in a homogeneous (exposed or unexposed) region (Fig. 14 B and C). The result is a higher optical density in the layer inside the exposed region next to the edge and a lower (fog) density in the unexposed region next to the edge. The contrast in optical density is thus enhanced on a microscopic scale; as a rule, this improves the sharpness of the image.

At the sprocket holes, however, the horizontal adjacency effect can have a very detrimental result. Fresh developer diffuses laterally into the photographic layer and can produce density streaks; these may generate interference frequencies in motion-picture films, especially in those with optical sound tracks.

Vigorous agitation of the developer solution promotes mass transport between the emulsion and the solution and thus weakens the adjacency effects. Adjacency effects can be quantitatively described by the line spread function [188], [189].

2.3.4.3. Direct Positive Development

In certain emulsions the latent-image centers are formed only inside the grains on exposure. These emulsions can be developed into direct positive images with special developer solutions [190], [191]. Such grains can be produced, for example, if after chemical digestion the digestion nuclei generated at the grain surface are "buried" in the grain by precipitation of more silver halide (see, e.g., Section 2.1.6).

The developer solutions used for this process contain, along with the developing agent, very small amounts of an especially strong reducing agent (e.g., a hydrazide) as a fogging agent. This reducing agent injects electrons from the developer solution into the silver halide grain [192]. In those grains where internal latent-image centers have been formed during exposure, these centers "draw" the injected electrons into the bulk of the grain, so that no fog nuclei (development centers) can arise on the surfaces of these exposed silver halide grains. As a result, only the grains that were not exposed are developed in such a fogging developer. This principle is utilized in instant color photography.

2.4. Photographic Properties and Their Control

Photographic materials based on silver halides are used in all areas of life. In addition to the broad assortment of B & W and color films and papers for amateur use, a wealth of special-purpose photographic materials exist, such as motion-picture films, graphic-arts films for use in the printing trades, X-ray films for use in medical diagnostic and materials testing, microfilm for archival use, and special films for scientific applications, e.g., in astronomy, holography, and nuclear research. All these materials have photographic emulsions as the light-sensitive component. This term "emulsion," which is physically incorrect, refers to the system of silver halide microcrystals (emulsion grains) in a matrix of a protective colloid, usually a dried and hardened gelatin layer (→ Gelatin).

All color photographic materials, and many B & W ones, have multiple emulsion layers that often contain mixtures of several emulsions, and in general any one emulsion can perform only the functions specific to it in a particular layer. The reasons for this limitation are the sensitometric requirements of the product, which differ sharply from one application to another; the complex interplay of the individual components in a photographic material; and the different developing and processing techniques employed with the several product groups. Accordingly, any photographic emulsion must be optimized for its special function, though compromises must always be made because certain properties vary contrarily. The modes of controlling the photographic properties include methods used in emulsion making and practices employed in the layer-by-layer construction of photographic materials with painstaking selection of all components.

The following sections mainly deal with the effect of the emulsion and to a first approximation the properties of the single grains are examined as representative of the emulsion. Virtually every innovation in emulsions is conceived of at the single-grain level and then extended to the greatest possible number of grains at once. There has been a trend toward more and more uniform emulsion grains in photographic materials for some years. In turn, modern physical analytical methods (see Section 3.1.4.3) on individual emulsion crystals have gained importance. In reality, any photographic image involves very many emulsion grains that react; for example, an ordinary ASA 400 color film has on the order of $10^9 - 10^{10}$ grains per square centimeter.

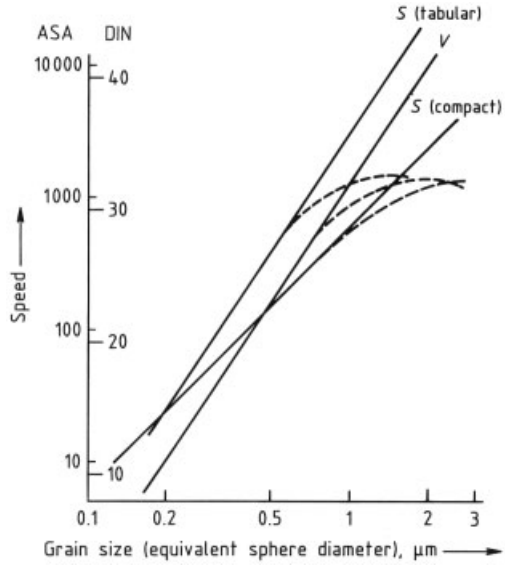

Figure 15. Theoretical sensitivity (quantum threshold 4 photons) as a function of grain size and shape [3]
Dashed curves show sensitivities attained in current practice

2.4.1. Sensitivity, Granularity, Sharpness

The sensitivity or speed of emulsion depends, above all, on the grain size. The larger the grains, the greater the number of incident photons per grain at a given exposure, and thus the higher the probability that a latent-image center will be formed. To obtain a numerical correlation between sensitivity and grain size, the intrinsic sensitivity of the emulsion grains must be distinguished from the spectral sensitivity imparted by dyes on the grain surface. The intrinsic sensitivity is proportional to the grain volume V; the spectral sensitivity, depending on absorption in the dye layer adhering to the grain, is proportional to the grain area S.

Figure 15 shows that these linear functions are valid only up to some limiting values; beyond these levels (dashed curves), the sensitivity increases more slowly and eventually decreases because of loss reactions in latent-image formation. If the sensitivity is enhanced by enlarge-

ment of the emulsion crystals (when enlargement is possible at all), the granularity of the developed silver in B & W films or the structure of the dye clouds in color films becomes coarser so that the image quality is impaired, especially in an enlarged print or transparency. In general, therefore, the desired sensitivity should be achieved with the smallest possible crystals, and the crystals must have optimal geometry. The emulsion grains for the blue part of the spectrum may be compact, while those to be sensitized for the green and red should ideally have a tabular shape. If the tabular grains are sufficiently thin (for instance, 0.2 μm in Fig. 15), they have a large area *S* but a relatively small volume.

Even with the minimum grain volume for a given sensitivity, the granularity can be substantially improved if development is retarded or even stopped before the entire grain is developed. This can be done with DIR couplers (see page 79); i.e., substances that react with part of the developer oxidation product and release an inhibitor. save

The halide composition is also important in the sensitivity – granularity relationship. The best results are achieved with AgBr – AgI mixed crystals with an iodide content of 2 – 15 %. The iodide should advantageously be incorporated in an interior zone of the grain to increase the sensitivity and to achieve proper interaction with the DIR coupler. High iodide concentrations at the crystal surface should be avoided as much as possible, because otherwise development is retarded from the very beginning and the sensitivity attained will not be high.

The sensitivity of an emulsion is greatly enhanced by chemical sensitization or ripening, in which aggregates of Ag_2S, or AgAuS are formed at the grain surface (see Section 2.1.3). The latent image is then formed on relatively few sensitivity specks on the grain surface, whereas in the unripened crystal, it forms preferentially on internal lattice defects and is inaccessible to the developer. The sensitivity can be increased by a factor of ca. 10 – 50 by chemical sensitization.

All measures aimed at increasing the sensitivity – in addition to those mentioned already there are many that are part of the carefully guarded know-how of every manufacturer – can be categorized under one of three basic concepts:

1) Enhanced light absorption (through halide composition, grain shape, dyes)
2) Improved efficiency of latent-image- center formation (through halide configuration, interfaces, doping, digestion)
3) Accelerated development [through emulsion additives such as polyalkylene oxides, development-accelerator releasing (DAR) couplers, halide scavengers]

The emulsion influences not only the sensitivity and granularity, but also the *sharpness* of the photographic layer. Sharpness is decreased first of all by the scattering of light on the individual emulsion grains. On the other hand, scattering increases sensitivity by lengthening the light path inside the emulsion layer and thus increasing the absorption. Scattering is at a maximum at grain sizes of ca. 0.30 μm and decreases in both larger and smaller grains. As a compromise between desired increase in sensitivity and undesired decrease in sharpness, the total amount of emulsion applied is kept as small as possible, and the strongly scattering grain-size fraction of ca. 0.3 μm is omitted, at least from the top layers, in order to hold down the sharpness loss in the layers beneath. Sharpness can also be improved by filter dyes added to the emulsion layer. These absorb in the same range of wavelengths where the emulsion is spectrally sensitized, thus reducing the diffusion halo but simultaneously also the sensitivity. The sharpness of a photographic layer also depends on its thickness. The thicker the layer, the larger the scattering cone (which widens along the light path) or the halo. The production of thin layers requires highly concentrated emulsions (high silver halide concentration, low gelatin content).

2.4.2. Gradation, Development Kinetics, Reciprocity

Gradation. The relationship between grain size and sensitivity implies that the characteristic curve of an emulsion with a broader grain-size distribution will cover a wider range of illuminance, that is, show a lower gradation (see Section 6.1). The grain-size distribution can be controlled over a wide range during emulsion making. Sometimes it is advantageous to synthesize a broad distribution by blending homodisperse emulsions (see Section 3.1.1) of different

grain size. Chemical sensitization always broadens the sensitivity distribution and thereby renders the gradation lower. Thus, even homodisperse emulsions do not have an arbitrarily steep gradation.

A fundamentally different approach to high gradation values is the doping of the emulsion grains with heavy-metal ions such as those of rhodium, cadmium, lead, and bismuth. Doping pushes back the toe of the characteristic curve and produces a steep gradation. For ecological reasons, however, cadmium salts are no longer employed in photographic materials.

Finally, the halide composition of the emulsion crystals affects not only the sensitivity but also the gradation. The gradation becomes lower with increasing iodide content. This is due to a broader sensitivity distribution and a slower rate of development.

Kinetics of Development. It is the rate of development, however, that becomes more and more important as processing speeds rise. This holds especially for photographic papers, in which fast-developing silver chloride emulsions are increasingly used.

Color film emulsions, in contrast, must not develop too rapidly; instead, their development must be retardable, for example by inhibitors released during development (see Section 5.2.2.3). In this way, not only can the granularity be improved, but edge and interimage effects that are important for image quality can also be attained. In general, the development kinetics of emulsions in color materials (especially color films) must be precisely matched – both to one another and with respect to the interplay with dye couplers and DIR couplers.

Reciprocity. The two kinds of reciprocity failure discussed in Section 2.1.6, LIRF and HIRF, can seriously impair the image quality. Since the individual grain-size fractions often differ in reciprocity, these failures manifest themselves not only in a loss of sensitivity but even more in gradation changes. Since reciprocity failure cannot be eliminated entirely, the changes in the color layers must be at least of the same kind, so that color imbalances will not occur. This sort of matching is essentially done through the emulsions.

Besides the addition of halogen acceptors, LIRF can most effectively be controlled by doping the emulsion grains with Ir, Ag_2, or Ag_2S nuclei. To prevent HIRF, chemical sensitization must be done in such a way as to ensure that the sensitivity specks are formed only at a few sites on the grain surface (see Section 2.1.6).

2.4.3. Stabilization

All photographic materials are chemically stabilized so that their sensitometric properties will change as little as possible during long storage, even in the presence of heat and high atmospheric humidity. The properties affected by such changes, usually include sensitivity, gradation, and fog; most commonly, sensitivity and gradation decrease and fog increases. These instabilities are caused by chemical reactions that continue inside the silver halide layers or by outside influences (e.g., atmospheric oxygen, humidity, cosmic and radioactive rays) that come into play during storage.

Important factors controlling the stability of an emulsion are the pH and pAg (negative logarithm of the concentration of silver ion). Suitable pH and pAg values are different for various photographic materials. Even so, an optimal choice of these two values will not alone result in adequate photographic stability. Additives called "stabilizers" are also necessary. These act chiefly by adsorbing to the silver halide crystals and forming sparingly soluble silver salts. The exact mechanism of stabilization is not known.

Typical classes of substances and examples of the associated stabilizers include the following:

1) Azaindenes or indolizines: 4-hydroxy-6-methyl-1,3,3a,7-tetraazaindene

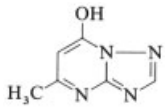

2) Imidazoles: 5-nitrobenzimidazole

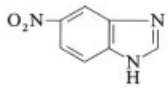

3) Triazoles: 1H-benzotriazole

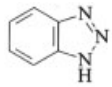

4) Tetrazoles: 2-mercapto-1-phenyltetrazole

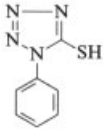

5) Thiazoles: 2-mercaptobenzothiazole

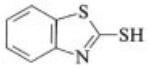

For a comprehensive bibliography on stabilizers, see [3]. Other compounds of practical value are those that release effective antifogging agents only during development (see Section 5.2.2.3).

Because of their gelatin content, photographic emulsions require bactericidal additives as preservatives. These include phenol, pentachlorophenol, thymol, and *p*-chloro-*m*-cresol. Biocides are usually added to the emulsion after chemical sensitization.

3. Manufacture of Photographic Materials

3.1. Emulsions

[14][15][19][27][31][193][194][195][196][197]

As discussed in Section 2.4, many photographic properties are directly influenced by the silver halide emulsion. The size, shape, and composition of the light-sensitive microcrystals allow a great variety of modifications. Despite this variety, the production of virtually every photographic emulsion can be broken down into three basic steps:

1) **Precipitation and Physical Ripening.** Silver halide crystals are formed in the desired size or size distribution with the desired chemical makeup. This step largely establishes the photographic qualities of the emulsion.
2) **Removal of Soluble Salts.** The soluble salts from precipitation are removed by washing. If coagulation or ultrafiltration is employed, the emulsion can be simultaneously concentrated to a high degree.
3) **Chemical Sensitization.** The emulsion is digested in the presence of special agents at ca. 45 – 65 °C. Sensitivity specks are formed on the surface of the emulsion grains; on subsequent exposure, these will facilitate the formation of latent-image centers. At the same time, this step markedly enhances the sensitivity.

3.1.1. Precipitation and Physical Ripening

To a prepared aqueous gelatin solution (ca. 2 – 5 wt %), aqueous solutions of silver nitrate and an alkali halide are added with vigorous stirring between 30 and 80 °C. Ammonium halides used to be employed frequently but are now avoided for environmental reasons. Because of the relatively low solubility of silver halides, the liquid phase becomes supersaturated as soon as the reactants come together, and silver halide nuclei form. The gelatin acts as a protective colloid and prevents the coagulation of the microcrystals. As precipitation continues, new nuclei can be formed but silver halide can also grow onto the ones already formed. The latter process is favored energetically. The dissolution of small unstable particles and reprecipitation onto larger crystals begins in the first instants of precipitation. This process is known as physical or Ostwald ripening. The driving force for this redissolving process is the increased solubility L_r/L_∞ of small particles which can be calculated with the Gibbs – Thomson equation. A plot of L_r/L_∞ versus particle size is shown in Figure 16 for AgBr. As a result of this increased solubility, the mean particle size of the disperse system increases while the number of particles declines, so that the total surface area and energy also decrease. Coagulation or coalescence processes, in which two or more microcrystals come into contact and grow together, are less important but give the same result. Finally, the process known as recrystallization occurs when crystals differing in halide composition are present. Again, the system tends toward the lower-energy state, in this instance, the formation of mixed crystals. Which of these growth mechanisms dominates (or excludes) all the others during emulsion making depends on the precipitation conditions, which usually vary over time. They include the order and rate of addition and the concentrations of the reactants, as well as the solubility of the silver halides, which depends on

temperature, pH, and complex formation. The nature and concentration of the gelatin can also play a crucial role in precipitation. All these factors imply a great diversity in the emulsion formulas. In view of the complex relationships between emulsion-specific and photographic properties, it is not surprising that the development of new formulations involves not only theoretical but also empirical knowledge. Some significant features of emulsions and emulsion making are described in what follows.

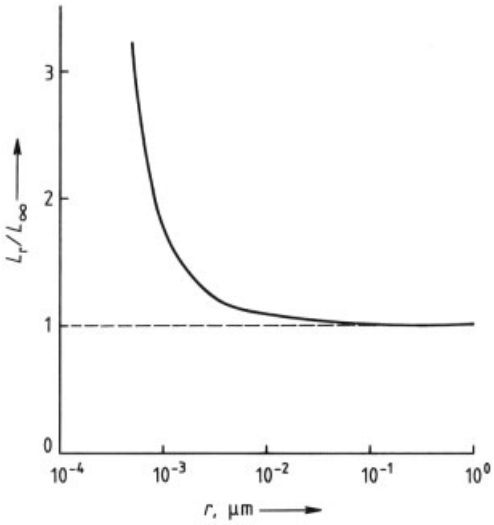

Figure 16. Relative increase in solubility L_r/L_∞ for AgBr in water at 50 °C versus particle size [14]

Single-Jet and Double-Jet Methods. If one reactant is present in the gelatin solution and the other is added, the precipitation scheme is called *single-jet process*. Precipitation is *direct* if the halide is initially in solution or *inverse* if the silver nitrate is. In a *double-jet method*, both reactants are added to the precipitation vessel at the same time. The rates of addition can differ greatly. Feed times commonly range from a few seconds (for some single-jet emulsions) to several hours for double-jet precipitations. The solutions can also be divided into several portions added at intervals. Emulsion making by the single-jet method, but also by some double-jet processes occurs with a high excess of halide ions during precipitation. The halide concentration $[X^-]$ which is related to the silver ion concentration $[Ag^+]$ through the solubility product of the silver halide $[Ag^+] \cdot [X^-] = K_{sp}(AgX)$ strongly influences the solubility of the emulsion grains. This relationship is plotted in Figure 17 for AgBr. The solubility is at a maximum at the equivalence point; as either the Br^- or the Ag^+ concentration increases from this point, it goes through a minimum, and at still greater ion concentrations the solubility rises again because in larger excess either ion can form soluble complexes with the corresponding counterion. In Figure 18 the solubilities of AgCl, AgBr, and AgI are plotted for the preferred working range of halide excess.

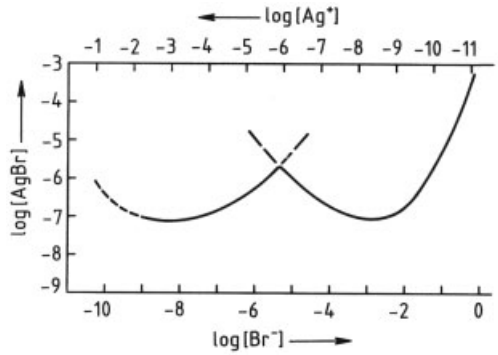

Figure 17. Solubility of AgBr in water at 50 °C versus concentrations of bromide and silver ions [14]

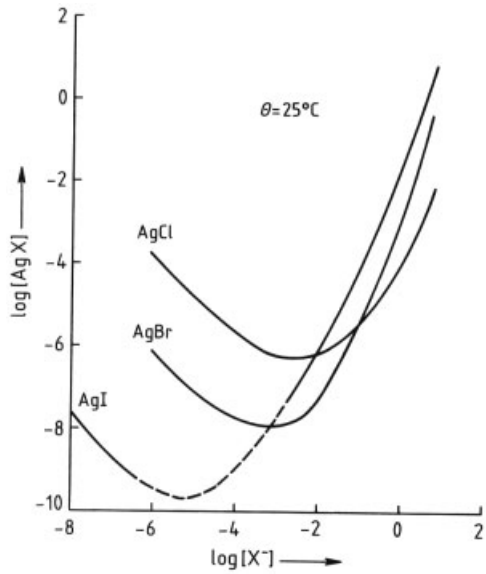

Figure 18. Solubilities of silver halides AgX (X=Cl, Br, I) in water at 25 °C versus halide concentration [14]

Photographic emulsions generally do not consist of a single silver halide but of a blend

of halides. The individual salts are precipitated in order of increasing solubility. When a solution of $AgNO_3$ is added to an initial mixture of alkali bromide and iodide, the sparingly soluble pure AgI precipitates first. This process continues until the iodide concentration in the solution has fallen to a value of

$$[I^-] = \frac{K_{sp}\,AgI}{K_{sp}\,AgBr}\,[Br^-]$$

A mixture of AgI and AgBr then precipitates; the ratio of the iodide and bromide concentrations $[I^-]$ and $[Br^-]$ is equal to the ratio of the solubility products. AgBr – AgI emulsions made by the single-jet method therefore have a higher iodide concentration in the interior than in the outer regions of the grain and large grains have a higher iodide content than smaller ones. While silver chloride and silver bromide, both of which crystallize in the fcc lattice, form mixed crystals in all proportions, silver iodide with its hexagonal lattice can be incorporated into AgCl and AgBr only up to some limiting concentration. A miscibility gap exists if the iodide content exceeds ca. 40 mol % in the AgBr – AgI system. The upper limit of miscibility i (in mol %) is a linear function of the temperature ϑ (in °C) [19]:

$$i = 34.5 + 0.165\,(\vartheta - 25)$$

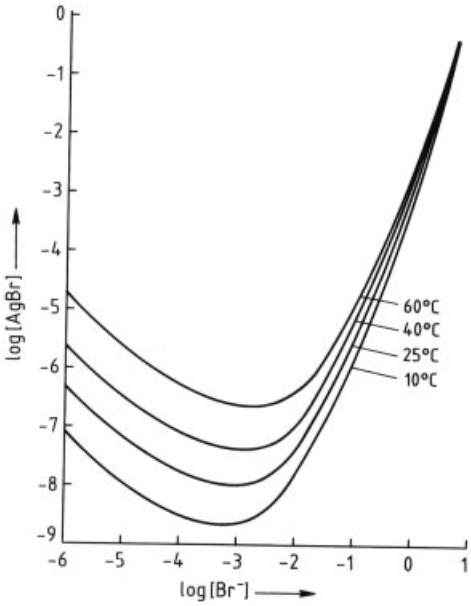

Figure 19. Solubility of silver bromide in aqueous solution versus bromide ion concentration at various temperatures [14]

Acid, Neutral, and Ammoniacal Emulsions. The first two types of emulsions are prepared with silver nitrate present in an acid or neutral solution. Both single-jet and double-jet methods are suitable, preferably a large halide excess is used at a temperature ca. 50 – 80 °C. The solubility of the silver halides increases with rising temperature (Fig. 19).

The solubility products $K_{sp}(AgX)$ as functions of temperature T (in K) can be approximately described as

$$K_{sp}\,(AgCl) = 2 \times 10^{-\frac{3000}{T}}$$
$$K_{sp}\,(AgBr) = 468 \times 10^{-\frac{4470}{T}}$$
$$K_{sp}\,(AgI) = 33 \times 10^{-\frac{5090}{T}}$$

Within certain limits, the grain size and grain-size distribution – in photographic terms, the sensitivity and the gradation – can be controlled through the precipitation temperature alone. Complexing in general, and in particular *complexing with ammonia*, increases the solubility of the silver halides even more than raising the temperature (Fig. 20).

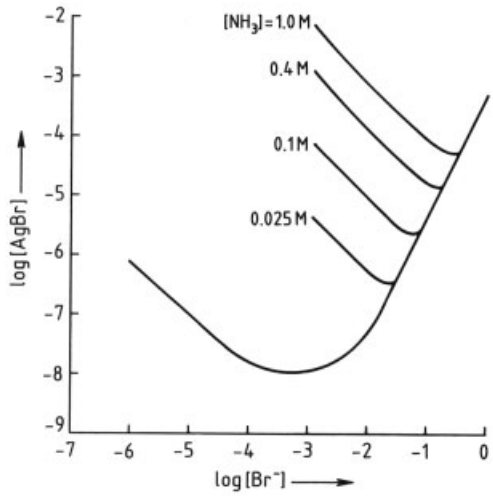

Figure 20. Solubility of silver bromide in aqueous solution at 25 °C versus $[Br^-]$ at various ammonia concentrations

Commonly, concentrated aqueous ammonia solution is added to the silver nitrate solution until the silver hydroxide, which precipitates first, is dissolved as a silver diammine complex $[Ag(NH_3)_2]^+$. The ammonia can also be added together with the halide. The precipitation method in the presence of ammonia is restricted to the sparingly soluble silver halides

AgBr and AgI and blends of these. Silver chloride emulsions cannot be precipitated from ammoniacal solution because the solubility is too high in this setting. Ammoniacal emulsions are prepared at a relatively low temperature during precipitation and physical ripening. Due to the high vapor pressure of ammonia the temperature must generally be kept below 50 °C. After physical ripening, the excess ammonia is neutralized with acid. The ammoniacal precipitation method still finds use in the preparation of coarse-grained X-ray emulsions. This process has the drawback of a relatively high ammonium ion load in the wastewater. Acid and neutral emulsions on the one side and ammoniacal emulsions on the other show characteristic morphological differences. The first, as a rule, show highly anisotropic crystal growth in the form of thin, often sharp-edged, three- or six- cornered tablets; the second exhibit no distinct growth orientation but take on a more and more spherical form in the course of precipitation and physical ripening.

These earlier precipitation methods are characterized by the fact that the crucial parameter for the quality of an emulsion made in these ways is the solubility during the emulsification process. As a rule of thumb, the higher the solubility (which depends on pAg, pH, temperature, and complexing) and the longer the precipitation and Ostwald ripening time, the coarser and more heterodisperse the emulsion becomes. And conversely, the emulsion becomes finer-grained and more homodisperse the lower the solubility and the shorter the precipitation time.

Converted Halide Emulsions. The conversion process was used especially for the production of paper emulsions. It is based on the transformation of more easily soluble silver halide grains to more sparingly soluble ones. Thus, if bromide is added to an AgCl emulsion, most of the halide is converted to an AgBr emulsion and chloride is liberated:

$$\mathrm{AgCl} + \mathrm{Br}^- \rightleftharpoons \mathrm{AgBr} + \mathrm{Cl}^-$$

The bromide concentration required for this reaction can be calculated from the solubility products and the free chloride ion concentration:

$$[\mathrm{Br}^-] > \frac{K_{\mathrm{sp}}\,(\mathrm{AgBr})}{K_{\mathrm{sp}}\,(\mathrm{AgCl})} \cdot [\mathrm{Cl}^-] \gtrsim 10^{-3}\,[\mathrm{Cl}^-]$$

In contrast, the conversion of AgBr to AgCl can take place only if $[\mathrm{Cl}^-] \gtrsim 10^3\,[\mathrm{Br}^-]$. As the extent of conversion increases, the crystal lattice becomes more and more deformed and is ultimately disrupted with the formation of a new phase having a different morphology.

Converted halide emulsions offer particularly good developability because of their highly disordered crystal lattice. They have the disadvantage of pressure sensitivity, because the internal lattice imperfections are reinforced by deformation so that the latent image appears inside the grains rather than at the surface. Pressure points may show up, for example, as bright areas on a black-and-white paper.

Monodisperse Emulsions. The double-jet method allows controlled emulsion making. Emulsions with narrow grain-size distributions result in particular when precipitation takes place near the solubility minimum, i.e., at relatively low halide excesses. Nucleation occurs only within the first few seconds, and a certain number of stable nuclei are generated. Afterwards, all the crystals grow in a uniform way. Since the physical ripening step is omitted, the initial grain-size distribution remains virtually unchanged in the course of precipitation, so that "monodisperse" emulsions are obtained [14].

pAg-Controlled Double-Jet Method. In process-engineering terms, the making of a monodisperse emulsion requires holding the halide concentration constant at a very low level ($10^{-2} - 10^{-3}$ mol/L) over a long period in which balanced double jet addition of silver nitrate and halide solutions occurs. The only way to achieve this is with electronic metering devices using the halide concentration or the corresponding silver ion concentration $[\mathrm{Ag}^+]$ as control variable; this latter value is more easily measured as the potential difference U_{Ag} between a silver electrode and a reference electrode. The potential difference is given by the Nernst equation

$$U_{\mathrm{Ag}} = U_{\mathrm{Ag}^\circ} + \frac{RT}{F} \ln\,[\mathrm{Ag}^+]$$

where U_{Ag° is the standard potential of silver, R is the universal gas constant ($R = 8.31\ \mathrm{J\,K^{-1}\,mol^{-1}}$), T is the temperature in K, and F is the Faraday constant ($F = 9.65 \times 10^4$ C/mol).

The Ag/AgCl-electrode in 3 M KCl solution is used as reference electrode, relative to which $U_{Ag^o} = 589$ mV.

Hence $U_{Ag} = 589\,\text{mV} + 59(= 2.303 \cdot 298\,\text{K} \cdot 8.31\,\text{J K}^{-1}\text{mol}^{-1} \times 10^3/9.65 \times 10^4\,\text{C/mol})$ log $[Ag^+]$ mV at 298 K. As $-\log[Ag^+] = \text{pAg}$, $U_{Ag} = 589\,\text{mV} - 59 \cdot \text{pAg}$ at 298 K. The saturated calomel electrode which is also used has a 37 mV lower potential than the Ag/AgCl electrode at 298 K.

The pAg-controlled precipitation method makes it possible to obtain emulsion crystals not only of the desired uniform size but also of well-defined shape. As already noted, AgCl and AgBr both form cubic crystals, but AgBr crystallizes in a cubic lattice only when precipitated at a very slight bromide excess, that is, at pAg values < 7.5. If the bromide ion concentration during precipitation corresponds to a pAg > 8.5, the grains become octahedral. The reason for this behavior is the difference in the strength of bromide adsorption to the cubic (100) and the octahedral (111) faces. At pAg < 7.5, more bromide is adsorbed to the (100) than to the (111) faces, so that growth is inhibited at (100) faces relative to (111) faces; the octahedral faces thus grow faster and eventually disappear, and only the cubic faces figure as growth forms. In contrast, at pAg > 8.5, the (111) faces are more strongly inhibited by bromide ion, so that the cubic faces grow faster, leaving only the octahedral ones. In the intermediate range, the tetradecahedra bounded by the cubic and octahedral faces are stable crystalline forms [198]. All seven kinds of possible crystal forms of silver halides allowed by the symmetry of the cubic lattice have now been observed. They can be obtained by growth on cubic or octahedral crystals in the presence of certain organic compounds called growth modifiers [199].

Precipitation at low halide excesses, that is, near the solubility minimum, must be carried out so slowly that no new stable nuclei can readily form when the reactant solutions are added. From the economic standpoint, the shortest possible precipitation time is desirable. The duration of this step can be shortened if the inlet rates of the silver and halide solutions are increased either stepwise or continuously in the course of precipitation. The metering rate can be increased in proportion to the total area of the growing emulsion grains, but it cannot be raised indefinitely. Supersaturation and new nucleation will occur, at the latest, when the spacing between grains has become so great (because of the greater dilution) that added silver and halide cannot diffuse rapidly enough to the growth sites on the grain surfaces. As a countermeasure, silver halide solvents can be added to enhance the solubility of the silver halide, so that the diffusion path around the individual grains becomes longer and the formation of new nuclei is shifted toward higher dilutions. Examples of AgX solvents include:

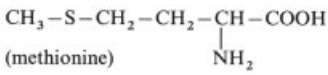

(methionine)

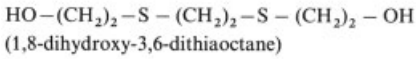

(1,8-dihydroxy-3,6-dithiaoctane)

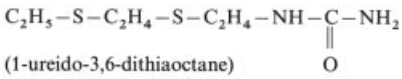

(1-ureido-3,6-dithiaoctane)

(imidazole)

The pAg-controlled process, the most modern used in emulsion making, offers high precision and even makes it possible to control within certain limits the number of stable nuclei formed in the nucleation phase (which is proportional to the reactant inlet rate and inversely proportional to the total solubility and the temperature). In this way, the precipitation time needed to attain a desired grain size can be calculated.

Core – Shell Emulsions. With pAg-controlled precipitation, grains of uniform size and shape can be obtained, and the internal structure of the crystals can be modified in a way that is largely uniform from grain to grain. This holds true especially for the halide composition, but also for photographically effective dopants of all kinds [200]. The simplest modification of the emulsion grain consists in a grain whose inner and outer zones differ in halide composition. The advantages of such a core – shell emulsion, with the core enriched in iodide and the shell depleted in iodide or free of that ion, were discussed in Section 2.3.4.3. The phase boundaries are especially important in this grain structure, since space-charge effects

at these boundaries are important in the formation of latent-image centers and can facilitate the separation of photoelectrons and positive holes. Emulsions have been developed that go beyond the simple core – shell principle, having multiple interfaces and multiple zones with different halide compositions and dopants. This approach makes it possible to vary the photographic behavior of emulsions widely and adapt them to other components of the photographic coatings. In addition to its effect on the sensitivity, gradation, and reprocity properties, the use of core – shell and multistructured emulsion grains also makes it possible to match the kinetics of development and coupling (e.g., of DIR couplers) to each other. This is a necessity with color films if high color brilliance is to be obtained through interimage effects, or extreme detail reproduction through edge effects.

Direct Positive Emulsions. The direct positive emulsions described in Sections 2.1.6 and 2.3.4 are a typical example of the pAg-controlled precipitation process with control of the grain structure. The incorporation of sensitivity specks in the grain interior is particularly successful when the crystals are regular in shape and homodisperse. Sensitivity centers need not always be present in the grain interior. An incorporated phase boundary such as AgCl – AgBr can have the same effect, that is, can topologically control the formation of latent-image centers [201], [202].

Redeposition of Lippmann Emulsions. In the older precipitation processes, there is always an overlap in time between nucleation, growth by direct deposition, and physical ripening. Ostwald ripening is suppressed with an approximately equimolar double-jet inlet of the reactants under pAg control, the silver halide grains grow to their final size solely by direct precipitation. In contrast, grain growth can also occur predominantly by Ostwald ripening if the silver halides are introduced not as salt solutions but as superfine so- called "Lippmann" or micrate emulsions with grain sizes smaller than 0.1 µm. If such an emulsion, which is relatively soluble, is blended with an emulsion having larger or more sparingly soluble grains, and if the mixing ratio, temperature, and complexing by the added silver halide solvents are selected appropriately, the micrate grains are dissolved and redeposited on the more sparingly soluble grains [203]. The pAg remains constant during deposition without any external control.

Tabular Grain Emulsions. Tabular emulsion grains, by virtue of their relatively large surface-to-volume ratio, are especially suitable for the adsorption of sensitizing dyes that impart the desired spectral sensitivity. In emulsion work, such tabular grains are characterized by the aspect ratio, the ratio of grain diameter to thickness, which can vary from grain to grain. The diameter is calculated by taking a circle having the same area as the projection of the crystal; the thickness is the distance between the two parallel (111) external faces [204]. Theoretically, it would be possible to attain an extremely high aspect ratio by making the tabular grains sufficiently thin. From the photographic standpoint, however, tabular grains are of interest only up to an aspect ratio of about 30 and with a thickness of about 0.1 µm. The reasons for this are: photoelectrons do not have an indefinitely long diffusion range; and large-area tabular grains become mechanically unstable, which leads to photographically problematic pressure phenomena such as fogged streaks where pressure has been applied. Even in the classical acid and neutral emulsions, the number of tabular crystals of various sizes is relatively high, and they normally occur along with other, irregular grain shapes. The newer tabular grain emulsions are distinguished from these by the fact that virtually all the grains are tabular. In order to produce such an emulsion, the maker must know the crystallographic structure of tabular grains. Despite their variety, nearly all observed crystal forms in commercial AgBr and AgBr – AgI emulsions can theoretically be attributed to just one type of possible structural defect in the AgBr lattice: stacking faults on (111) faces [205], [14].

The regular growth of a crystal of the rock salt type proceeds on (111) faces in such a way that the second and third lattice planes B and C are each displaced slightly from the first plane A, while the fourth lattice plane again lies exactly above the first (see Fig. 21).

A lattice plane of an octahedral face can "erroneously" fail to occupy its proper position, that of the third-from-last plane, taking up instead the position of the next-to-last. This can occur espe-

cially in the nucleation or initial growth phase. In other words, instead of the ABCABCABC arrangement, the lattice plane following B can grow in position A instead of position C. This "error" results in a stacking fault, with the consequence that the nucleus grows in a symmetrical manner, forming a mirror image relative to plane B. From its appearance, this crystal is called a twin. At the surface location indicated by arrow in Figure 21, where the twinning plane B intersects the surface, a trough forms.

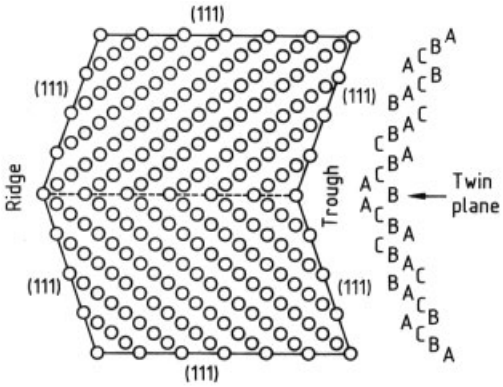

Figure 21. Stacking faults on a (111) face

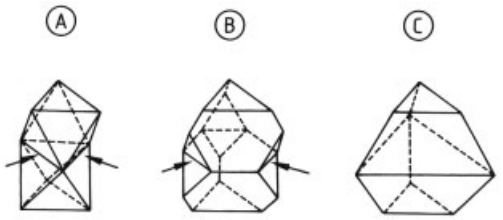

Figure 22. Simple octahedral twinning (schematic)
For explanation see text

Figure 22 A shows a crystal nucleus produced by twinning. The lower octahedron represents the regularly grown region; the upper one is the part grown after twinning. In reality, the upper portion is not an ideal octahedron but a region just a few lattice planes thick. The figure clearly shows, however, that troughs form at the surface locations pointed out by the arrows, where the twinning plane intersects the surface. There is a third trough behind the plane of the image (not visible). Further silver and halide ions can be easily deposited in these three troughs (Fig. 22 B); because this process is energetically favored. The trough faces therefore grow faster than the other faces, until finally the simple twin form (Fig. 22 C) is reached.

The situation is different if two stacking faults occur in quick succession on two parallel (111) faces. The resulting double twins have two troughs, which remain permanent features during the further deposition of silver and bromide ions and lead to strongly anisotropic lateral growth, that is, a tabular grain shape (Fig. 23).

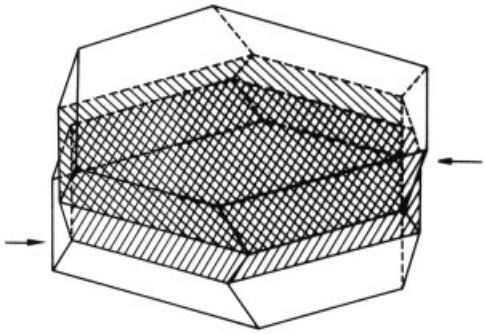

Figure 23. Double twin (twin lamella)

The crystallographic precondition for the formation of tabular grains is accordingly the presence of two stacking faults on parallel (111) faces, that is, the occurrence of a twin lamella in the nucleation phase. Octahedral faces are produced at high pAg values (see page 34), while stacking faults occur chiefly at high inlet rates. If both conditions are established at the beginning of a precipitation, the grains produced will be almost exclusively tabular grains.

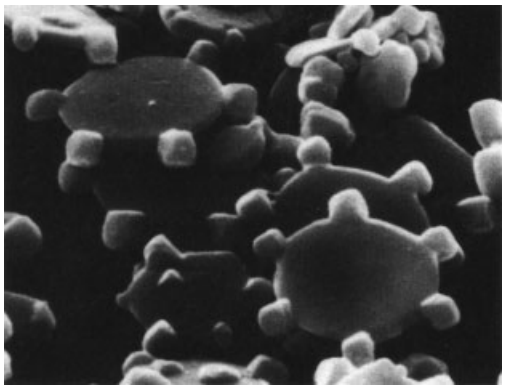

Figure 24. AgBrI crystals with epitaxially deposited AgCl

Epitaxy. The oriented growth of small quantities of a new silver halide phase on emulsion grains has long been employed to achieve photo-

graphically useful effects [206]. Such epitaxial growth can also be topologically controlled in several ways. Besides the crystallographic influences, the local halide composition of the host crystals can be decisive, and in addition "site directors," such as stabilizers and sensitizing dyes, can modify the epitaxial structures in a variety of ways [207]. Figure 24 shows, as an example, small AgCl cubes at the vertices of AgBr – AgI tabular grains. The cubes are produced by precipitation in the presence of a sensitizing dye. The formation of sensitivity specks and latent-image centers can be directed by those epitaxial overgrowths.

3.1.2. Removal of Soluble Salts

The soluble salts generated during precipitation, chiefly the excess halide, must be removed so that emulsions can be chemically ripened and after coating, crystallization of salts in the dried emulsion layers can be prevented.

Washing and Coagulation Processes. Formerly, the emulsion was solidified, cut into small pieces, and washed in running water down to a well-defined residual salt content. Currently, a coagulation process is generally used; this is more economic and also offers the photographically important advantage of concentrating the emulsion, since the coagulated precipitate can be dissolved in less water than was present in the initial volume.

For example, the gelatin can be precipitated with alcohol or high concentrations of inorganic salts, such as Na_2SO_4, $MgSO_4$, or $(NH_4)_2SO_4$. The amounts of salts required are smallest at pH values of 3 – 4. Low gelatin concentrations and high silver halide contents promote coagulation.

A more elegant method of coagulating the gelatin is to modify it chemically in such a way that it is only sparingly soluble in acids. This can be done by reacting the basic amino groups of the gelatin with acid chlorides, anhydrides, isocyanates, diketones, and similar reagents. The addition of 10 % of an acylated gelatin is enough to coagulate all the gelatin at pH 3. After the coagulated precipitate has been washed several times and decanted, the acylated gelatin is redissolved in water by raising the pH, and the original emulsion is reconstituted. The gelatin can also be coagulated at an acidic pH through the formation of coacervates with polymeric organic compounds that contain sulfo or carboxyl groups. The washed precipitate is then redispersed at $pH > 6$. Suitable polymers include polyesters of sulfosuccinic acid, poly(styrenepyridinesulfonic acid), or poly(vinylpyridinesulfonic acid); maleic acid – styrene copolymers; and copolymers of acrylic or methacrylic acid with their esters. In order that the coagulated precipitate can be uniformly washed, it must be in a solid but fine-grained form. The consistency of the precipitate can be controlled through the type of gelatin, the ratio of gelatin to silver halide, and the coagulation temperature. The degree of washing or the residual salt content of the precipitate can be determined by electric conductivity and pAg measurements. The pAg of each emulsion is adjusted to a specifically well-defined value with alkali halide after redispersion; it is crucial for the next operation, namely chemical sensitization.

Ultrafiltration. Ultrafiltration is a desalting method in which the gelatin need not be solidified or coagulated. It is a membrane process that can separate the alkali nitrate solution from the remaining emulsion in the liquid phase and is thus superior to the more complicated and hard-to-automate coagulation process from both engineering and economic standpoints. Ultrafiltration equipment with hollow-fiber modules has proven suitable for emulsion work. A hollow fiber module contains several hundred parallel capillary tubes made of polysulfone inside a cartridge. From a supply tank, which can also be supplied with fresh water, the emulsion is circulated through these hollow fibers. The salt solution permeates through the membranes and is withdrawn on the shell side of the cartridge. The separation limit of the membrane is around 30 – 50 kdaltons. The operation of ultrafiltration equipment is characterized by the pressure and the flow rates.

Figure 25 shows a flowsheet of an ultrafiltration unit employing the feed – bleed principle, in which two closed loops allow independent control of the flow velocity in the UF cartridge and in the feed vessel. Provision is also made for a reserve flow so as to permit cleaning. Of course, ultrafiltration can also be used for concentrating the emulsion. Low gelatin contents favor this

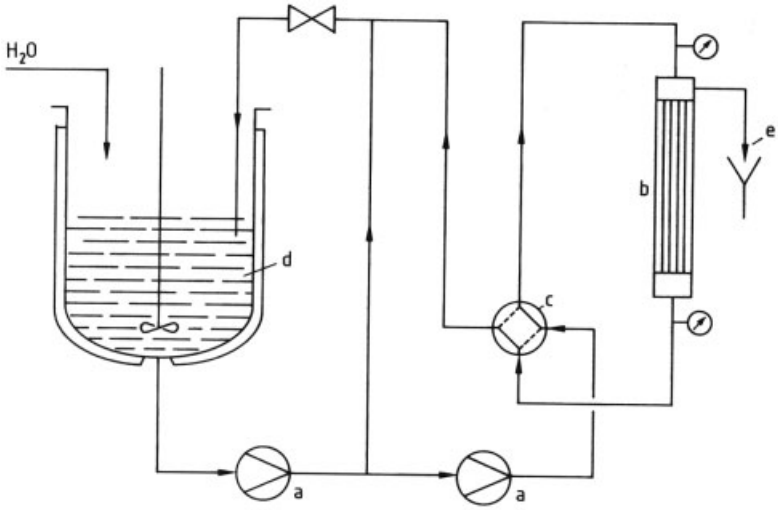

Figure 25. Ultrafiltration unit
a) Pump; b) Cartridge; c) Switch valve; d) Feed vessel; e) Salt solution

process. Ultrafiltration can also be carried out during precipitation or in between precipitation steps [208]. By concentrating the emulsion, this technique makes it possible to process relatively large quantities of silver halide per unit vessel volume.

3.1.3. Chemical Sensitization

The emulsion grains after precipitation and desalting are fairly insensitive. When they are exposed to light, latent-image centers are formed throughout the grain volume, preferentially at any lattice defects that are present. Early in the history of photographic emulsions, it was discovered that the sensitivity in the intrinsic absorption part of the spectrum can be enhanced by additional digestion of the washed emulsion at elevated temperature, especially if more gelatin is added.

Sulfur Sensitization. In 1925, S. E. Sheppard realized that the gain in sensitivity described above is due to labile sulfur compounds in the gelatin. After this discovery, it was found that thiosulfate, added in quantities of ca. 5 – 50 µmol per mole silver, is the compound that acts as sensitizer during chemical ripening or postdigestion. Initially thiosulfate is adsorbed to the surface of the silver halide as the complex $[Ag(S_2O_3)]^-$. The adsorbed complex then disproportionates to Ag_2S and sulfate. On exposure, latent-image centers consisting of a few silver atoms are formed preferentially at these Ag_2S sensitivity specks. Other sulfur compounds employed in digestion include thiourea, *N*-allylthiourea, and thiosemicarbazide. Sulfur sensitization depends on pAg, pH, and temperature. The usual operating ranges are as follows: pAg, ca. 7 – 9; pH, ca. 5 – 8; temperature, 45 – 65 °C. Digestion times are usually between 1 and 4 h. Much shorter ripening times of a few minutes in combination with higher temperatures of up to ca. 80 °C have also been cited in the literature.

Gold Sensitization. In 1936, R. Koslowsky achieved a markedly greater increase in sensitivity and a simultaneous improvement in latent-image stability by using gold(I) thiocyanate complexes in addition to sulfur sensitization. It is thought that a mixed gold – silver sulfide is produced in gold sensitization. This compound presumably lowers the number of silver atoms required for developability from about four to three or even two. The amount of gold needed is 2 – 20 µmol per mole silver. The combination of gold and sulfur sensitization with the stabilizer triazaindolizine, in particular, led to a marked improvement in emulsions.

The degree of chemical ripening is controlled by sampling the emulsion at intervals, submitting it in liquid state to sensitometric tests or coating it on a support, and performing photographic tests. Figure 26 illustrates the sensitometric changes occurring in digestion. The sensitivity increases, reaches a maximum, and may decrease again if digestion time continues. The gradation behaves in a similar way. The limiting factor is the photographically undesirable fog that occurs on prolonged digestion. The presence of fog suggests that larger silver aggregates, developable even without exposure, have been formed on some grains, Chemical ripening is stopped, e.g., by cooling or addition of a stabilizer when sensitivity is at a maximum and fog is still relatively slight.

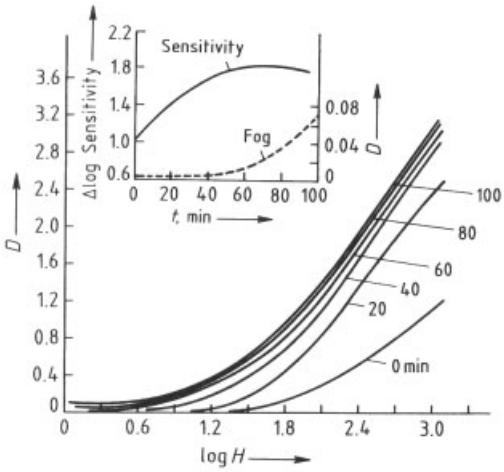

Figure 26. Developed density D versus exposure ($H=I \cdot t$) after various digestion times
Inset: sensitivity and fog versus ripening time

Reduction Sensitization. The sensitivity of an emulsion can be greatly enhanced even without thiosulfate and gold. Digestion with suitable reducing agents, such as tin(II) compounds, ascorbic acid, hydrazine, or formamidinesulfinic acid has a sensitizing action. Reduction sensitization by *Wood's silver digestion method* is also very effective. In this process, apparently, the reducing capacity of certain reducing agents present in the gelatin is utilized at high pH and high silver ion concentration, i.e., a low pAg. Such reduction sensitizers as carbohydrates and aldehydes can be found in nearly all types of gelatin, so that reduction sensitization is thought to be involved to some degree in all forms of chemical sensitization. The sensitivity specks formed during reduction sensitization consist of silver. Photographically, the behavior of small and large reduction centers is totally different: small reduction centers act as hole traps, thereby diminishing recombination and not controlling the topography of the latent image. For this reason, such centers should preferably be incorporated in the interior of the emulsion grains. Larger reduction centers, in contrast, act as electron traps, similarly to the sensitivity specks formed in sulfur and gold sensitization. More intensive reduction sensitization in conjunction with gold sensitization, results in fog. Reduction centers can be produced with hydrogen too; which leads to an increase in sensitivity for some hours, especially in dry, degassed photographic films, and low-intensity reciprocity failure can be minimized.

Digestion in the Presence of Dyes. It used to be common practice not to add sensitizing dyes until just before emulsion coating. Currently, emulsions are often precipitated or at least chemically ripened in the presence of one or more dyes. In the first case the adsorbed dyes serve as growth modifiers, in the second they cover large areas of the grain surface during digestion and thus cause the sensitivity specks to be concentrated on the uncovered areas. Lowering the number of sensitivity centers should combat the "waste" of silver deposition and thus counteract any loss of sensitivity. The use of dyes as site directors for the topology of sensitivity specks, and subsequently that of latent-image centers, does not work with all the commonly employed sensitizing dyes.

3.1.4. Emulsion-Making Technology

In keeping with more complicated emulsion recipes, the standards of emulsion-making technology have become more stringent. Good emulsion quality and reproducibility are preconditions for meeting the high quality standards imposed on photographic products. Emulsion technology includes the following essential process steps:

1) Tempering

2) Metering
3) Stirring – mixing
4) Ultrafiltration (optional)

One source of difficulty is the necessity of working under muted red light or, once the sensitizing dyes have been added to the emulsion, in complete darkness. For this reason, manual operations are eliminated whenever possible, and emulsion plants make use of automation and process control systems where these are workable and economically acceptable. Process control computers handle the batching (weighing) of all ingredients, including water; temperature and flow-rate control; and monitoring of the desalting process.

3.1.4.1. Raw Materials

Even the lowest level of impurities in an emulsion can markedly impair the photographic result. Process equipment, peripheral equipment, and all raw materials used must therefore meet strict cleanliness and purity requirements.

Appropriate filtration units must deliver *air* free of both solid and gaseous contaminants, especially hydrogen sulfide. *Water* is usually treated on ion-exchange resins and must not contain any reducing agents. The specification for *silver nitrate* and *halides* is stringent, especially for heavy-metal impurities: the concentration of iron, copper, and lead must be < 1 ppm. The suitability of all feedstocks is ultimately judged through photographic testing.

Gelatin is still irreplaceable for emulsion making because of its unique combination of good physical and chemical properties, such as protective colloid action, halogen-accepting capacity, its ability to solidify below 30 °C, its hardenability, and swelling. Despite the animal origin and thus a varying composition of the raw materials, great improvements in the quality and re-producibility of photographic gelatin have been achieved. Inert bone gelatin, which offers the maximum freedom from photographically active constituents, predominates in emulsion making. Special modified grades of such gelatins are sometimes employed for particular applications (see page 38). In addition to gelatin, synthetic polymers are also used. Examples are cellulose derivatives, poly(arylic acid) derivatives, poly (vinyl alcohol), and polyvinylpyrrolidone. However, they are generally applied only as additives to improve the physical properties of the emulsion coatings; they are not so much employed in emulsion making.

3.1.4.2. Emulsion-Making Equipment

Photographic emulsions are usually prepared batchwise. The vessels range from hundreds to thousands of liters in capacity. To protect against corrosion, all vessels used for precipitation, coagulation, and chemical ripening, as well as all other equipment that comes in contact with the starting solutions or the emulsion, are made of stainless steel or, more recently, titanium. The corrosion resistance of austenitic stainless steel increases with increasing molybdenum content. Suitable materials of construction include those identified by the following Material Numbers: DIN 1.4439 (4 % Mo); DIN 1.4529 (6 % Mo); and DIN 3.7035 (Ti).

Normally, separate vessels are used for each process step (precipitation, coagulation, or ultrafiltration, and chemical ripening). The vessels have double walls and are equipped with electronic water-temperature controls as well as pAg and pH electrodes. The preferred bottom shape is conical so that even small starting volumes can be stirred without undue foaming. The silver nitrate and halide solutions can be fed via piping from several (at least two) feed tanks (also with temperature control) to the *precipitation vessel*. In the simplest arrangement, the reactants are delivered by gravity and fed through fixed or adjustable gate valves; for more exact control, metering pumps or flow-rate control systems (which may be pAg- controlled if appropriate) are employed. Because good mixing is very important – i.e., of the reactants with one another and with the gelatin solution, or with emulsion already present – a variety of methods are used to achieve adequate stirring and mixing. In addition to classical stirrers in the vessel there are techniques where the reactants are metered into a separate mixing chamber installed either in the tank itself or in an external pumped- circulation loop [209–213] (Fig. 27). The spontaneous ionic reaction then takes place in a well-defined space under relatively reproducible conditions, while the slower process of physical ripening takes place chiefly in the tank.

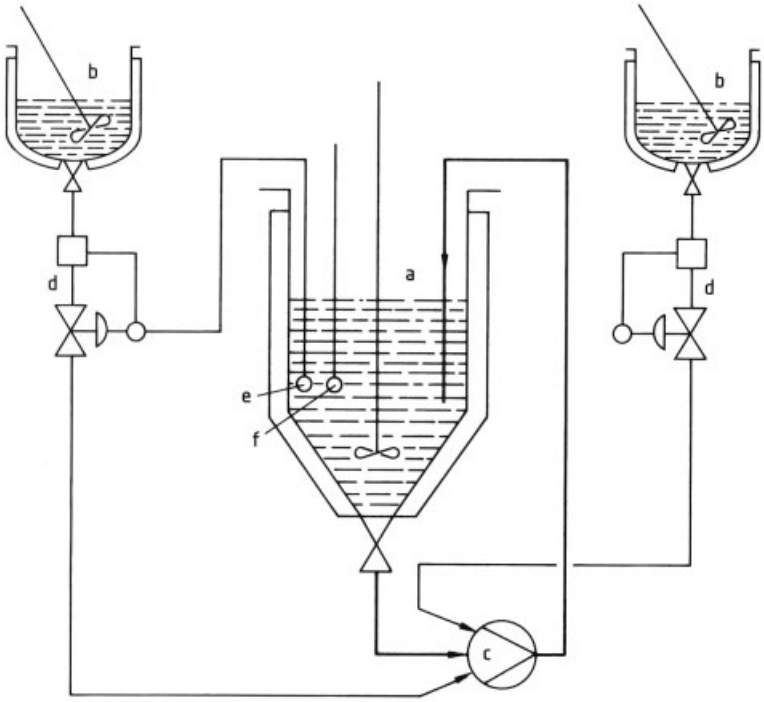

Figure 27. Precipitation system for photographic emulsions
a) Precipitation tank; b) Reactant feed tanks; c) Pump and mixing chamber; d) Metering controls; e) pAg electrode; f) pH electrode

The *coagulation tanks* also have high-performance stirrers so that the coagulated precipitates can be uniformly washed. Equipment in these tanks also includes pH, pAg, and electric conductivity measurement devices. After coagulation and washing water can be decanted or separated by filters, centrifuges, or hydrocyclones. Like all other water streams used for cleaning, it is intercepted and forwarded to silver recovery and clarification. For ultrafiltration, see Section 3.1.2.

Redispersion of the coagulated precipitates and chemical sensitiztion usually take place in a single operation. The *digestion tanks* feature precise temperature control; for controlled dissolution of the coagulated precipitates, they have pH control and, like the precipitation and coagulation tanks, are provided with pAg electrodes. They also permit the addition of sensitizing dyes. After chemical sensitization is complete, dyes, stabilizers, and preservatives are added to the emulsions. The emulsions are either filled in plastic bags and stored in cool rooms or rapidly solidified in cooling pans or on cooled belt conveyors then held in sealed metal or plastic tanks located in cool rooms (between 2 and 5 °C). Under these conditions, emulsions can be held for several months.

Continuous Emulsion Making. Continuous operation, at least in the *precipitation step*, is useful in view of the desired uniformity and exact reproducibility of photographic emulsions. In the simplest case, the reactants ($AgNO_3$, halide, and gelatin solutions) are continuously metered into a mixing vessel, while the same amount of emulsion is discharged from the vessel, for example by way of an overflow, so that the volume remains constant. In this continuous mixed-suspension, mixed-product-removal (CMSMPR) process a steady state is reached after a certain transient time period; the grain-size distribution and grain shape of the emulsion then remain constant [196].

In other continuous precipitation processes, the precipitation tank is replaced by a tubular reactor, which has at least one but usually several mixing chambers in a cascade arrangement;

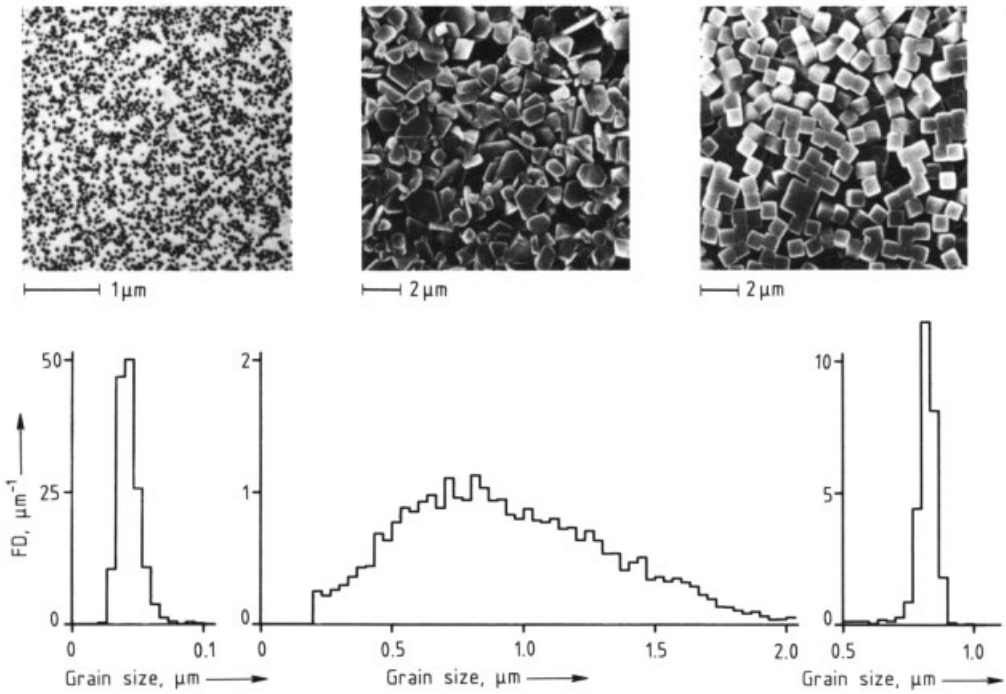

Figure 28. Grain size plots and electron micrographs of three emulsions

a wide variety of mixing chambers have been described. The precipitation reaction itself takes place in the reactor, while ripening occurs in pipes, which may be provided with static mixers. The length of piping corresponds to the needed ripening time [6].

Continuous processes have been described not only for precipitation and ripening but also for *coagulation* and *washing* as well as chemical sensitization. Continuous salt removal by ultrafiltration is particularly easy to manage.

The continuous processes offer some advantages over the batch processes: better reproducibility, higher production rates, capability of automation, quicker detection of malfunctions, and as a result lower emulsion losses. They have, however, the disadvantage of inflexibility. Different types of emulsions demand substantial equipment modifications; for example, the number of mixing chambers and the length of the ripening pipe runs. With regard to freedom of formulation, batch emulsion making is superior to the continuous processes. Another problem lies in the thorough cleaning required for emulsion-making equipment; a tubular reactor with its large wall area is inferior to a tank in this respect.

3.1.4.3. Test and Measurement Methods for Photographic Emulsions

In addition to the many photographic tests, which also include simulated long-term storage under extreme conditions, a number of nonphotographic measurement techniques are also used [214]. The most important characteristic of an emulsion is the *grain-size distribution*. Initially grain-size analysis was carried out by means of electron micrographs, but this method was relatively difficult. Today this analysis can be done easily and quickly by an *electrochemical technique* developed specially for silver halide emulsions [215], [216]. A highly diluted emulsion is used, the grains of which are allowed to settle onto a substrate, the preparation is placed in an electrolyte, and a platinum tip acting as

negatively charged electrode is guided over it. When the tip makes contact with a grain, the latter undergoes complete spontaneous reduction. The electric flux is a direct measure of the grain volume.

Figure 28 shows three emulsions with very different grain-size distribution, illustrating the capability of the electrochemical technique as a complement to electron microscopy.

In addition to the grain size and shape, the *halide distribution in the grains* is photographically relevant. It can be measured in integral form by X-ray diffraction techniques. Crystallographic phases differing in halide composition are often reflected in separate peaks. The resolution needed for this purpose can be improved substantially by using synchrotron orbital radiation (SOR) instead of X-rays for structural analysis [217]. Synchrotron radiation is emitted by electrons orbiting in an accelerator; it features high intensity and collimation. These two measurement methods are compared in Figure 29. While the principal peak (97 mol % AgBr, 3 mol % AgI) is measured by both techniques, only the SOR measurement reveals the existence of a second mixed phase (84 mol % AgBr, 16 mol % AgI).

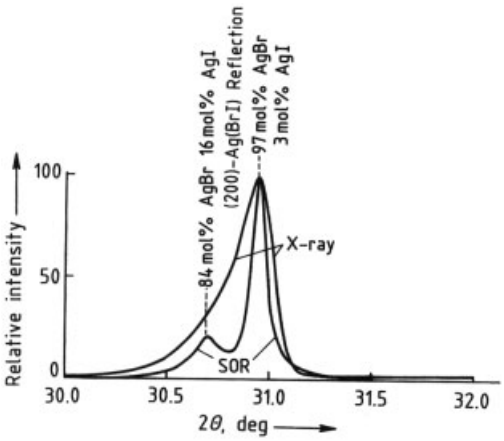

Figure 29. The (200) reflection of AgBrI microcrystals X-ray=Guinier measurement with Cu $K_{\alpha 1}$ radiation. SOR=synchrotron measurement with $\lambda = 0.1274$ nm [218] recalculated to Cu $K_{\alpha 1}$

In principle, the halide distribution in single emulsion grains could be determined as well; for example, by electron-beam microanalysis. In this technique, a grain or a thin-section specimen of a grain is scanned with an electron beam, and the concentration of halides is determined from the intensity of the corresponding X-rays emitted. Portions of the grain near the surface can also be investigated by SIMS and ESCA. Single-grain analyses are relevant only for homodisperse emulsions, because the results cannot be representative in other cases.

Measurements that can provide information on the formation of the latent image are of special interest. Important quantities are the concentrations and lifetimes of photoelectrons, positive holes, and interstitial silver ions. Microwave photoconductivity, dielectric loss, and the Dember effect are suitable analytical techniques. The measurement of luminescence, which can identify the energy level of lattice defects and recombination centers, give additional information. The technique is also suitable for the study of spectral sensitization and stabilization. Radicals of sensitizing dyes generated by exposure and dopants, often present at levels of a few atoms per microcrystal, can be detected by ESR spectroscopy [214].

3.2. Emulsion Supports

The substrates under emulsion layers must meet high standards, especially with respect to their physical properties. These include high transparency, high rupture and tearing strength, freedom from curl (wet and dry), stiffness, high modulus of elasticity, antistatic qualities (to guard against electrical discharges), good dimensional stability, and low coefficient of friction.

The constituents of the support must not affect the photographic properties of the various materials in the emulsion layer or the sensitometry of the photographic film. In addition the gelatin-based emulsion layer must also adhere well to the support during every stage of use and processing.

These diverse requirements are met through the proper selection and fabrication of the film base and the use of special, nonphotosensitive auxiliary layers on it.

The only bases now commonly used are *papers* [with baryta (barium sulfate) or polyethylene coating] and *plastic films* (cellulose triacetate or polyester). *Glass plates*, despite their undisputed advantages (dimensional stability,

freedom from curl), have lost much of their former importance because of their undeniable disadvantages (fragility, weight, space required). They are still used to a limited extent in some special fields. Emulsion- coated textiles find limited application as decorative materials (e.g., coated linen).

Films, chiefly polyester, opacified with white pigment such as TiO_2 are now coming into use for display materials (dimensionally stable substitute for paper).

3.2.1. Films

3.2.1.1. Properties of Film Materials

Cellulose Nitrate. When Kodak introduced cellulose nitrate as an emulsion support at the end of the 19th century, the basis was laid for the widespread use of photography today; ie., a light, flexible recording medium was used instead of the clumsy glass plates used up to that time. Motion-picture photography, above all, is possible only with flexible film as emulsion support (motion-picture and narrow-gauge films). Later, however, the high flammability of cellulose nitrate led to a number of catastrophic fires, which ultimately led to the development of less flammable films with no tendency to ignite spontaneously. Today, as far as is known, cellulose nitrate film base is no longer manufactured. For reasons of safety, most archival films, especially motion pictures, have been transferred from this medium to cellulose triacetate-based films.

Cellulose Acetate. The first "safety films," made from acetone-soluble cellulose acetate with 54 – 56 % combined acetic acid content (2^{12} acetate), were not competitive with cellulose nitrate, being inferior to it in mechanical properties and dimensional stability. These qualities were later improved by increasing the degree of acetylation. Today, virtually all miniature, roll, motion-picture, and narrow-gauge films use a triacetate base with ca. 61 % bonded acetic acid, which is soluble in a 9 : 1 solution of dichloromethane and methanol.

Cellulose is acetylated with acetic anhydride, either in a solvent that dissolves the product cellulose acetate or in a medium such as xylene, which acts as a nonsolvent, leaving the fibrous structure of the feed cellulose (e.g., cotton linters) unchanged. The two types of cellulose acetate are equally suitable for use as emulsion supports.

Cellulose Mixed Esters. Mixed esters of cellulose with acetic and homologous carboxylic acids are less water-sensitive than pure cellulose acetate. Cellulose acetates butyrates, in particular, found practical use. Today they are employed only for special-purpose films (e.g., duplicating sheet film, where the acetone solubility of the support allows special techniques such as butt splicing).

Polystyrene. Films made of polystyrene have very low water uptake, but increased temperature sensitivity limits their dimensional stability (theoretically better than cellulose triacetate). These films have a great sensitivity to kinking; irreversible clouding results from even slight kinking. Even though this support has now been improved, it is of little practical importance.

Polycarbonate. Because of their low water sensitivity, polycarbonate supports made from bisphenol A [2,2-bis-(4-hydroxyphenyl)propane] and phosgene appeared highly suitable for reprographic films when they first became available. But the relatively low modulus of elasticity meant that these bases had inadequate resistance to the tensile forces of emulsion layers, and hence poor dimensional stability. Only small amounts of polycarbonate films are now manufactured for use in photographic materials. Polycarbonate pigmented with TiO_2 has now been largely supplanted by similarly pigmented polyester.

Polyester. In this context, "polyester" always denotes poly(ethylene terephthalate) (PETP). Decades of effort aimed at improving dimensional stability have resulted in the worldwide use of this support, which combines all the desired features: high modulus of elasticity, low water absorption, and good performance at high temperatures. Special PETP supports for photographic materials are manufactured, e.g., by Du Pont (Cronar), ICI (Melinex), Kodak (Estar), Agfa-Gevaert (Gevar), 3 M, Fuji, Toray, and Teijin. These bases are absolutely insoluble

in common solvents. Polyester supports are employed for graphic-arts films and reprographic materials because of their optimal dimensional stability; they are used for X-ray films because they are stiffer than cellulose acetate. Finally, the high mechanical strength of PETP makes it possible to produce thinner supports for narrow-gauge films (Fuji). Their acetaldehyde content (resulting from thermal conversion of ethylene glycol during synthesis) can be detrimental to the potential use of PETP in color negative films; acetaldehyde can modify the color balance if it reacts with dye couplers, especially the magenta coupler. In order to minimize the acetaldehyde content of the substrate, the manufacturing conditions must be closely controlled.

3.2.1.2. Manufacture of Photographic Film Bases

Modern emulsion supports are produced by a wide variety of methods. *Cellulose triacetate* is fabricated by solution casting (→ Films, Chap. 2.1.). The operation is performed in drum or belt casting machines, the latter usually having belts made of highly polished stainless steel. The solvent is a mixture of dichloromethane and lower aliphatic alcohols such as methanol, 2-propanol, and butanol. Solvent recovery takes place in brine chillers or on activated carbon. Plasticizers (usually phosphates such as triphenyl phosphate) are added not to improve the mechanical properties but to reduce the flammability and water absorption (so as to minimize longitudinal expansion when the humidity is high). Dyes are added to prevent halation (see Section 3.2.1.6) and to combat *light piping*. (Light piping means that light may penetrate into the film inside its cartridge, passing through the edges of the leader, which protrudes from the cartridge, and into the transparent support. It experiences total reflection at the boundaries of the base, thus propagating further into the film and causing pre-exposure.) If a blue or, better, a black absorbing dye is incorporated into the support, the light is absorbed and light piping is suppressed.

Common film thicknesses are ca. 0.09 mm for roll films, 0.13 mm for miniature and narrow-gauge films, 0.14 mm for motion-picture films, and 0.19 mm for studio and sheet films.

In contrast, *polyester film base* is extruded from the melt at around 300 °C and biaxially stretched sequentially (→ Films, Chap. 2.3.2.), usually first in the longitudinal and then in the transverse direction. The film is then tempered at about 200 °C (thermofixation). Polyester film-base thicknesses are ca. 0.10 mm for graphic-arts and reprographic films and ca 0.18 for X-ray films. The methods used to produce polyester granules and films have continuously been improved in recent years. The price for PETP supports is now only about half that of equally thick cellulose triacetate supports. As a result, the question whether to use PETP bases for miniature, roll, and motion-picture films is being reexamined, even though some important technical problems remain to be solved. Examples are: precurvature, developing better cutting and perforating dies for this very tough base, devising antistatic coatings with adequate conductivity and mechanical strength. One problem not yet dealt with is how to control the cold creep of PETP film ("core setting" effect); i.e., the property of the film base to take on an irreversible curvature when rolled up. Poly(ethylene terephthalate) does not lose this curvature in wet development whereas cellulose triacetate does. Recently, however, PETP films have been subjected to tempering in the rolled-up stage in order to impart a negative precurvature; thus, the aforementioned difficulty has been avoided, at least over a long span of time; however, the remaining core setting is too strong, thus it is not appropriate to use PETP as a support for miniature films. PETP bases are already in use for some motion-picture films (because they have a larger winding diameter), for some roll films (e.g., ORWO negative films), and some sheet films.

Because photographic emulsions are sensitive to a great variety of contaminants, the machinery and fabrication rooms involved in the manufacture of film bases (including subbings and backings) have to meet extraordinarily stringent requirements on cleanliness. Special or even extreme purification and filtration of air used in the machines and present in the rooms – particularly the air used for drying the film and coatings – is an obvious measure, as are special lint-free clothing, clean rooms, and air-looks.

3.2.1.3. Subbings

The hydrophilic binders (nearly always gelatin) used with light-sensitive silver salts cannot adhere to hydrophobic emulsion supports. Good adhesion of the wet or dry emulsion is, however, an indispensable condition for the use and processing of photographic materials, especially the graphic-arts types. The needed adhesion is ensured by "subbing", in which the film surface is rendered hydrophilic by chemical or physical means, through the incorporation of hydrophilic colloids, or by multilayer systems in which the hydrophilicity increases from layer to layer. In difficult cases, especially with PETP, several approaches are combined. Achieving optimal adhesion during every stage of use and processing is one of the most serious problems in the manufacture of emulsion supports, and accordingly the patent literature in this area is quite extensive [219].

Cellulose Acetate Film Bases. All cellulose acetate bases now in use have hydrophilic colloids, usually gelatin, incorporated into the surface. These layers are applied as gelatin solutions that contain organic solvents as swelling agents for cellulose triacetate.

To facilitate drying, the swelling agents used have the lowest possible boiling points. Acetone is most common; sometimes it is mixed with dichloromethane. The solutions also contain a gelatin solvent, usually water (ca. 3 – 20 %), along with lower aliphatic alcohols such as methanol or ethanol. The crucial factor for surface swelling is the acetone – alcohol – water ratio. It varies usually between 60 – 70 : 10 – 20 : 2 – 3 vol %, depending on gelatin content which is in the range of 7 – 14 wt %.

To be stable at all in such predominantly organic solvent systems, the gelatin must be largely salt-free. Salt-free gelatin is produced by complete deionization, for example on ion-exchange resins. Acidic or alkaline stabilizing agents are also needed; examples are acetic, phthalic, salicylic, or maleic acids. Subbing solutions often contain hydrophobic polymers, such as cellulose nitrate or cellulose acetate, which improve the attachment of the subbing to the surface of the film base. Finally, gelatin hardeners are added.

If applied in quantities above a certain threshold level, cellulose nitrate catalyzes the hydrolytic decomposition of the triacetate substrate with the liberation of acetic acid ("vinegar syndrome" [220], [221]). This effect must be taken into account in the archiving of (motion-picture) films on a triacetate base.

The surface of cellulose acetate films can also be rendered hydrophilic by corona discharge or saponification [222], but these methods are no longer used.

Polyester Films. Because PETP films do not swell properly in the common organic solvents, the development of subbings for this material has been far more difficult. In practice, quite a variety of approaches are employed. Most frequently double-layer systems with increasing hydrophilicity are used, in which the first layer contains synthetic polymers and the second is essentially gelatin. As a first layer, for example, vinylidene chloride – methyl acrylate – itaconic acid copolymers (91 : 10 : 2) combined with poly(ethyl acrylate) in aqueous dispersion are applied before stretching [223]. The layer is then anchored to the film surface and incorporated into it during stretching. An adequate second layer is a 2 % gelatin solution applied after stretching. Other methods of subbing use soluble polyesters, such as poly(ethylene isophthalate) [224], applied from chlorohydrocarbon solutions after stretching, as the first coating. Such layers attach to PETP by adhesion. The second layers contain cellulose nitrate and gelatin.

The adhesion of the relative hydrophobic first coatings can be further improved by various pretreatments. Examples include oxidizing processes (e.g., with chromosulfuric acid, permanganate, peroxides, or a flame treatment of the film surface with an oxygen excess), corona discharges, UV and high-frequency irradiation, and swelling with special agents (e.g., phenol, resorcinol, chlorophenols [225], chloral hydrate, and chloroacetic acid). Swelling agents are sometimes added directly to the first subbing layer [226]. Most of the swelling agents suited to PETP are, however, not physiologically innocuous – in particular polyphenols and halogenated polyphenols so that special practices are necessary in manufacturing.

Efforts are being made to achieve good subbing with a single-layer system. In addition to

gelatin and a special polymer, such formulations contain one of the swelling agents listed above if the coating is to be applied after drawing [227]. If the subbing is applied before stretching or between the longitudinal and transverse stretching steps, no swelling agent is required [228].

3.2.1.4. Curl

After coating with a photographic emulsion, an initially flat film base will tend to curl toward the emulsion side. The amount of curl depends on the thickness and stiffness of the support as well as the thickness, composition, and drying of the emulsion. In practice, the tensile forces produced are exactly offset if the back of the support has a similar coating. In the simplest instance, such as X-ray films, the film is coated with *emulsion* on both sides. An "empty" gelatin layer (i.e., a plain gelatin layer without photographically sensitive components) on the back of the support is called a *noncurl (NC) coating*. Noncurl coatings are used on roll, studio, and reprographic films.

Because the gelatin NC coating is easily damaged, this system initially could not be employed on films processed in continuous-transport machines. Originally only motion-picture films were processed in these machines, but nowadays this also holds true for narrow-gauge and miniature films (i.e., 8- and 35-mm films). If the emulsion exerts too great a tensile force on these films, which causes a strong curvature of the film base to the emulsion side, the film base is precurved in the opposite direction by treating it on the back with swelling solvents such as acetone. The precurvature can be decreased and controlled through mixing with nonswelling solvents or through the use of high temperatures for post drying.

3.2.1.5. Antistatic Coatings

Film bases are extremely hydrophobic and tend to collect electrostatic charge. Charges are generated by friction, but especially when two surfaces separate after being in close contact; for example, when the emulsion and back sides separate during rewinding or unwinding. The discharges that follow lead to undesired and harmful fog, which takes the form of dots in the case of positive charging and branching lines in the case of negative charging [229]. There is a danger of creating such marks whenever the material is light-sensitive: during coating, cutting, packaging, use, and processing. It increases with increasing photographic sensitivity of the emulsion, increasing fabrication rate, and decreasing atmospheric humidity. The danger is especially great for single- coated films and for polyethylene- coated (resin- coated) photographic papers. Because gelatin is more hydrophilic, materials coated on both sides have much less tendency to static discharges.

Miniature, motion-picture, and narrow-gauge films are generally provided with antistatic backings for this reason. Charging can be prevented or minimized as follows:

1) Reduce the contact between the emulsion and backing through the incorporation of discrete particles ("spacers") into the emulsion; of i.e., photographically inactive inorganic pigments or small plastic spheres (usually methyl methacrylate), which protrude partway out of the uppermost layer.
2) Equalize the contact potentials of all possible contact partners.
3) Dissipate the electrical charges by lowering the surface resistance.

The most effective, and most widely used, approach is to lower the surface resistance [230] by application of conductive backings. Surface resistance values of $10^8 - 10^9\ \Omega$ per square at 50 % R.H. are generally sufficient. Most conductive backings contain polymeric film-forming electrolytes, usually alkali salts of polymeric carboxylic or sulfonic acids; examples are copolymers of acrylic or maleic acid [231], polystyrenesulfonic acid [232], or polymers with a quaternary nitrogen such as poly-2-*N*-methylpyridinium sulfate [233]. These antistatic coatings can be anchored to the back of the film base if appropriate swelling agents or swelling solvents are included in them. In the best case, a single solution can serve as antistatic and noncurl coating. Coatings with conductive carbon-black offer particularly good antistatic properties. They also have the advantages that their surface resistance is independent of the ambient humidity, and that they protect against halation (see Section 3.2.1.6).

But removal of such coatings requires an alkaline pretreatment in processing. Alkali-soluble high-molecular-mass compounds such as cellulose acetate hexahydrophthalate are used as binders [234].

Organic semiconductors, such as polyanilinium salts [235], vanadium pentoxide [236] and doped metal oxides, such as SnO_2 doped with antimony [237], have found use in humidity-independent permanent antistatic coatings on various microfilms. Important features of these materials include conductivity of the backing even after processing, which helps to prevent dirt adhering to the base.

3.2.1.6. Antihalation Coatings

Halation results from both reflection and diffusion (for diffusion halation). *Reflection halation* produces a halo effect in the developed film; it results from light rays reflected at the interface between media that differ in refractive index, mainly at the back of the film base. Suitable dyes, located in various layers of the overall system, can absorb this light. Characteristic of halation dyes are their content of acid groups in the molecule. By virtue of this they can be dissolved and eliminated from the layer during development.

Black layers of colloidal silver are often placed between the emulsion and the support to absorb reflected light. This approach, however, can be used only if the processing includes a bleaching bath in which this silver dissolves (black-and-white reversal films, color materials). Many black-and-white negative films have *antihalation undercoats* (AHU coatings), which contain dyes that are dissolved and bleached in the developer; hence, no bleaching bath is required.

The film base is dyed gray (*gray base*) when a neutral density of about 0.2 in the developed film will not pose a problem (miniature black-and-white negative films). The dyes used for this purpose have the most uniform absorption possible over the entire spectrum. Nearly all other miniature films, as well as roll and narrow-gauge films, have a lower gray-base density of ca. 0.04 in order to improve light-tightness in cartridges or cassettes (to prevent light piping; see above). Roll and sheet films with gelatin-based NC coatings (see Section 3.2.1.4) can have water-soluble dyes, such as acid green, pan yellow, or acid fuchsin, added to them in order to provide additional antihalation protection. When NC coatings or dyed undercoats cannot be employed, alkali-soluble backings containing carbon-black or appropriate dyes can be used as an alternative to black layers (e.g., in motion-picture films). Binders for these coatings are polymers with free carboxyl groups, which are again dissolved and removed during processing in alkaline baths.

3.2.1.7. Friction-Control and Protective Coatings

In order to avoid malfunctions in film transport in the camera, certain types of films, such as miniature and narrow-gauge films, must have a low coefficient of friction. Lubricants are coated on both the emulsion side and the back. Friction-control coatings on the back usually contain waxes (carnauba wax, montan waxes) and simultaneously improve the abrasion resistance of antistatic coatings and alkali-soluble antihalation coatings. Because the solvents involved (carbon tetrachloride, trichloroethylene) are toxic, attempts have been made to incorporate disperse lubricants directly into the antistatic or antihalation coatings.

3.2.2. Papers

Photographic paper bases can be classified as follows:

1) Resin-coated (polyethylene-coated) papers
2) Barium sulfate (baryta) coated papers
3) Lacquered papers
4) Uncoated papers for direct emulsion coating

Combinations of these treatments are also possible; for example, paper resin-coated on both sides can be given an additional baryta coating. The only treatments of importance are resin and baryta coating. The types of photographic paper bases common in the industry are as follows (basis weight of uncoated paper is given in grams per square meter).

Document paper, baryta- coated	50 – 110
Semi-weight paperboard, baryta- coated	120 – 159
Paperboard, baryta- coated	160 – 250
Papers for phototypesetting, resin- coated	65 – 90
Papers for diffusion transfer process, resin- coated	90 – 125
Semi-weight paperboard, resin- coated	120 – 159
Paperboard, resin- coated	160 – 200

3.2.2.1. Properties

Physical Properties. The most important physical properties are surface finish and formation, tint and opacity, stiffness and structural strength, thickness, curl, and freedom from physical contaminants.

The *surface finish*, which is produced by one of the treatments listed, is crucial for obtaining the desired surface quality in the final image (high-gloss, matt, lustre, etc.). The *formation* (as a property of the paper base) is among the factors controlling the uniformity of the final image.

The *tint* of the support must be precisely matched to the emulsion that will be applied, so that undesired color deviations can be avoided. High *opacity* in the support prevents the perceived color from changing when the photo is viewed on a dark background.

The *stiffness* of the paper is important chiefly to the consumer, especially in the case of large formats. The *structural strength* must be high when the material is to be processed in a modern developing machine.

Rigorous standards apply to *uniformity of thickness*, to guarantee the quality of emulsion coating and the trouble-free processing of roll and sheet papers.

Since the emulsion is coated on one side only, the gelatin usually exerts strong tensile forces. These must not produce curl in the print. Curl is opposed by the *stiffness of the support*, the *negative curl* in resin- coated papers (controlled by the density of the polyethylene resin), and gelatin- containing *noncurl coatings* on the back of the paper.

Physical contaminants can be directly visible in the finished print or may impair the surface of the paper in such a way that local disturbances of the emulsion that can be seen in the print. The same applies to defects such as impressions and other unevennesses on the paper surface.

Chemical Properties. The most important chemical properties are photochemical inertness, ability to adhere to the photographic emulsion, and stability with respect to developer solutions (hydrophobility).

All components of the support must meet strict requirements with regard to *photochemical inertness*. Constituents that initiate redox reactions in the emulsion ("chemical fog") or can influence other reactions must be excluded. Defects include not only local impurities containing metals or sulfur, but also volatile constituents or those able to migrate, which can lead to fog when photosensitive papers are stored for long periods (as they usually are). Ordinary methods of chemical analysis are hardly suitable for detecting such impurities, because photographic emulsions are far more sensitive to them than are the analytical methods [238]. Sensitometric tests with special test emulsions and suitable processing baths are carried out for a variety of impurities.

The *adhesion* of the photographic emulsion is improved by agents in the baryta coating or by treatment of the resin coating ("corona discharging" or subbing).

Stability to developing and fixing baths is achieved by rendering the paper base hydrophobic with resin-based or fatty substances; thus shortening the time required for processing baryta- coated paper. In resin- coated supports, the penetration of developer solution is prevented at the cut edge, where the paper base is exposed; otherwise, an unattractive brown streak would occur at the edge of the print.

3.2.2.2. Manufacture and Coating

All photographic paper bases, regardless of the treatment carried out before sensitization, begin as fiber-base paper made on endless-wire (foundinier) machines (→ Paper and Pulp, Chap. 7.2.3.).

They are made from highly bleached wood pulp (sulfite or kraft process, softwood or hardwood). Refining is optimized with regard to surface finish and mechanical strength; the freeness can be characterized as cut/short-fibered, between 30 and 45° SR (→ Paper and Pulp, Chap. 6.1.4.). The degree of internal sizing is

far greater than in other fine papers (writing and printing papers), so that the paper will be sufficiently hydrophobic for the subsequent processing operations.

Surface sizing is performed such as to match the subsequent operation (baryta or resin coating) and combines functions such as hardening, adhesion, and conductivity. The paper is compacted (calendered) in order to produce the required surface properties.

Important properties of the paper base include

1) Thickness and basis weight
2) Uniformity of thickness and basis weight profiles
3) Moisture content and uniformity of moisture content profile
4) Finish and smoothness (also after coating)
5) Stiffness and structural strength
6) Hydrophobicity
7) Tint
8) Adhesion of subsequent coating
9) Freedom from photochemical and mechanical contaminants and defects

The paper-base formulation is always matched to the special requirements of the end product, i.e., the photographic paper. The paper support is thus regarded as only one portion of a whole system and cannot be arbitrarily replaced [239].

Some typical paper-base formulations are illustrated in what follows.

Paperboard for baryta coating (240 g/m^2)

Pulp:
70 % softwood sulfite pulp
30 % hardwood kraft pulp

Internal sizing (relative to pulp):
1.5 % resin size
0.8 % sodium stearate
1.0 % gelatin
0.5 % melamine – formaldehyde resin
1.5 % aluminum sulfate
0.2 % optical brightener

Surface sizing:
4 % aqueous gelatin solution

Paperboard for resin coating (175 g/m^2)

Pulp:
30 % hardwood sulfite pulp
50 % hardwood kraft pulp
20 % softwood sulfite pulp

Acidic internal sizing (relative to pulp):
1.3 % sodium stearate
1.0 % polyacrylamide
1.0 % aluminum sulfate
3.0 % titanium dioxide
0.1 % optical brightener
0.4 % polyamineamide – epichlorohydrin resin

Alternatively :

Neutral sizing (relative to pulp):
1.5 % starch
0.5 % alkylketene dimer
1.0 % polyamineamide – epichlorohydrin resin
0.1 % optical brightener

Surface sizing:
Starch solution or poly(vinyl alcohol) solution, optical brightener, salt

Baryta Coating [240]. The baryta coating consists of barium sulfate, as pigment, and a binding agent, chiefly gelatin. Dyes and optical brighteners are added to produce the desired color location. Special agents ensure the adhesion of the subsequent emulsion coating.

The baryta coating is applied in one or two layers. These are leveled with brushes, air knives, or bars and dried in a festoon or channel dryer (→ Paper and Pulp, **A 18**, and Fig. 30). The baryta-coated paper is calendered or embossed, depending on the finish desired in the photographic paper.

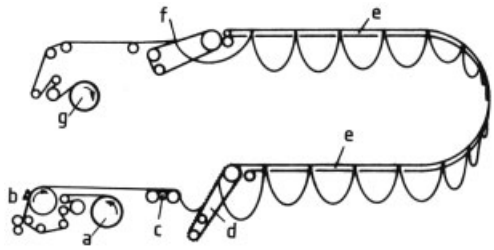

Figure 30. Schematic diagram of a brush-type coating machine with festoon dryer for the application of baryta coating to photographic paper base
a) Unwinding; b) Coating station; c) Tension rolls; d) Ascending pin conveyor; e) Festoon dryer; f) Descending pin conveyor; g) Windup

A typical baryta coating (50 % solids content) has the following composition:

45 kg barium sulfate
15 kg gelatin
0.04 kg citric acid for pH adjustment
0.2 kg chrome alum (hardener)
0.01 kg optical brightener
toning dyes

Figure 31 shows the structure of the product.

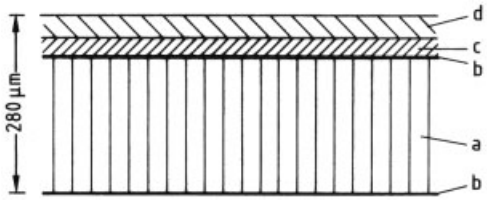

Figure 31. Schematic cross section of a baryta-coated paperboard base
a) Paper base; b) Top sizing; c) First baryta coating; d) Second baryta coating

Resin Coating. Fiber-base paper is laminated on both sides with polyethylene in an extrusion machine with a wide-slot die (Fig. 32) [241–244]. If necessary, the back side can be printed before coating (b). The front and back resin coatings are applied in separate laminators (d, e). The surface of the paper is treated with corona discharges (c) just before coating in order to improve the adhesion of the resin.

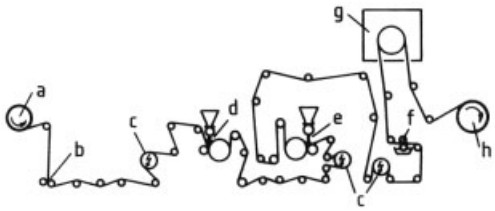

Figure 32. Schematic diagram of a tandem extrusion coating machine with post aqueous coating facilities for two-sided resin coating of photographic paper base
a) Unwinding; b) Back-side printing press; c) Corona treatment for improved adhesion; d) Back-side laminator with chill roll; e) Front-side laminator with chill roll; f) Postfinishing, i.e., application of functional coating; g) Dryer; h) Windup

Immediately after the polyethylene melt is applied, the film is rapidly solidified on a chill roll. The resin coat is imprinted with the surface of the chill roll; accordingly, this step determines the finish characteristics of the final print (glossy, matt, lustre) as well as the finish and smoothness of the back.

Low-density and high-density polyethylenes (LDPE and HDPE) of high purity or mixtures of these are used. The back of the support is generally coated with transparent polyethylene. The resin for the front coating usually contains titanium dioxide, which enhances the line sharpness of the final print. Toning pigments and optical brighteners are added to adjust the tint. Special grades (diffusion papers) have carbon-black pigment added to the resin for the front coating.

The curl of the finished photographic paper base is controlled by adjusting the density of the polyethylene resins used for the front and back coatings. If the support is to be coated with emulsions that generate particularly strong tensile forces but must nontheless remain flat over a wide range of relative humidity, noncurl coatings are commonly applied. They are applied especially to papers with raw basis weights between 60 and 100 g/m^2 and top of the back resin coating.

If necessary, other functional layers can be applied on the resin, for example subbings, writability aids, and antistatic coatings. Figure 33 shows the structure of the product.

A typical composition of the polyethylene coating of a paperboard is as follows:

Front , 32 g/m^2 :
80 % LDPE
10 % HDPE
10 % titanium dioxide and pigments

Back , 30 g/m^2 :
65 % LDPE
35 % HDPE

Subbing :
2 % aqueous gelatin solution

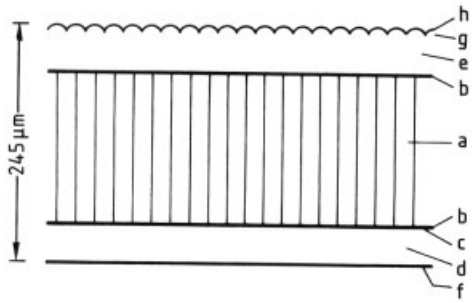

Figure 33. Schematic cross section of a resin-coated photographic paper base with embossed surface
a) Paperboard; b) Top sizing; c) Printing on back; d) Back-side resin coating; e) Front-side resin coating; f) Functional back-side coating; g) “Silk” pattern embossing; h) Subbing

3.2.2.3. Packaging, Storage, Transportation, Manufacturers

Photographic paper bases are packaged as machine-width rolls in polyethylene shrink-wrap; they are usually stored on end in climate-controlled rooms. The rolls are generally transported, on end, on special pallets in sealed containers, rail cars, and trucks. Packaging, storage, and transportation conditions must be such as to prevent variations in moisture, mechanical damage, and contamination [245], [246].

Suppliers. Major suppliers of photographic paper bases (without emulsion) include the following:

Fuji Photo Film Co., Tokyo, Japan (resin-coated bases)
James River Corp., High Wycombe, Bucks, United Kingdom (baryta- and resin- coated bases)
Mitsubishi Paper Mills, Tokyo, Japan (baryta- and resin- coated bases)
OJI Paper, Tokyo, Japan (resin- coated bases)
Felix Schoeller jr., Osnabrück, Germany (fiber-base, baryta- coated, and resin- coated bases)
Schoeller Technical Papers, Pulaski, N.Y., United States (fiber-base, baryta- coated, and resin- coated bases)
Freiberger Zellstoff- und Papierfabrik, Weissenborn, Germany (fiber-base and baryta-coated bases)
Serpukhov, former Soviet Union (fiber-base, baryta- coated, and resin- coated bases)

3.3. Coating

Exact Coating Thickness. Extremely small thicknesses, down to 10^{-4} mm (dry thickness), must be precisely reproduced. Unevenness in the thickness of a single layer can lead to variations in the density, that is, spots. In a color film, an error of a few percent in coating thickness can cause a color imbalance.

Freedom from Coating Defects. Even small flaws, such as bubbles, streaks, gaps, dust particles, agglomerations of coating material etc., can render a photographic material unserviceable, because enlargement (by a factor of 200 in the case of a motion-picture film) will make them distinctly visible. A high degree of cleanliness, especially the elimination of dust, is thus a vital requirement. In addition to finished product quality control, a number of flaw-detecting instruments are built into the coating machine in order to ensure defect-free operation.

Multiple Coatings and High Speeds. Nowadays color films, for example, can consist of as many as 15 emulsion layers. There is also a trend toward higher coating speeds (30 – 250 m/min) and greater support widths (1 – 2 m). For reasons of economy, the number of passes must be kept as small as possible; thus it is necessary to apply many coatings in a single pass through the coating machine.

Operation in Darkness. As photographic sensitivity increases and emulsions are sensitized far into the red portion of the spectrum, not even red light can be permitted during processing. Materials are therefore manufactured in the dark by using infrared light for automatic quality control and image converters for visual surveillance. When there are gaps in the spectral sensitivity of a material, light of a definite wavelength can be used.

3.3.1. Coating Machinery

The coating on the emulsion support consists of light-sensitive layers, which have an enormous range of photographic properties, and a number of auxiliary layers, which together or separately perform the following functions: light filtration, separation of differently sensitized emulsions containing different dye couplers (see Section 5.2.2.3), antihalation, mechanical protection, noncurl, subbing, antistatic, hardening, matt finish, and many others (see Section 3.2.1). The number of layers is steadily increasing, especially in color materials, because of more and more stringent requirements; at present, 10 to 15 is common. Generally these coatings are on one side of the material. Certain materials, however, such as X-ray films, have coatings on both front and back.

The coatings are usually applied in the form of an aqueous gelatin solution, dispersion, emulsion, or suspension; in some cases, solutions in organic solvents are employed. The coating material must be prepared before application; this stage may include melting, tempering, and the addition of "finals" to adjust surface tensions and viscosities, degassing, and filtration. The emulsion can be prepared batchwise or continuously. The design and sizing of the tanks and metering equipment depend initially on the lot size and secondly on the shelf life of the photographic emulsion, including finals.

Equipment for the *coating of long film or paper webs* (1 – 2 m wide and up to 5000 m long) includes devices similar to those employed in the paper converting or printing industry. Special machinery is used for the coating of plates made of glass or other materials, for example of optical filters, of photographic plates for scientific purposes, or of plates serving as the initial stage in the manufacture of printed circuits (→ Imaging Technology, Chap. 4.2.) and microelectronic devices (→ Imaging Technology, Chap. 4.3.).

A coating machine for film or paper webs is shown in Figure 34. The most important features include: the unwinding section (a) (usually there are two or even three unwinding stations in order to allow continuous operation); the device for splicing the webs (b); the web accumulator (c) which allow continuous running of the machine while, for example, a new web of material is being pasted to the old one or when a new web is begun at the windup station (h).

Modern machines with "flying attachment" to the new web do not have an accumulator ahead of the windup station; cleaning and pretreatment devices (d), such as a station for corona treatment to improve adhesion to resin-coated papers; the coating station (e); the emulsion feed tanks (f); the chilling and drying section (g); and the windup station (h).

Other components of a coating machine include equipment for the conveying, cooling, moistening, and cutting of the papers or films, dry air preparation, solvent recovery, and process measurement and control. Because of the large number of variables that must be determined and set, process-control computers are now used in large coating machines.

3.3.2. Coating Methods

Warm gelatin solutions (ca. 40 °C) can be applied by various methods (see Fig. 35 and [247]). All the coating systems have the following common features: they must produce a defect-free coating with a well-defined thickness that is highly uniform over the length and width of the web; they must be easy to operate and keep clean; and must offer high reliability.

Dip Coating. In the earliest process, dip coating (Fig. 35 A), the thickness of the coating applied depends on the speed of the web of material being coated, the viscosity and density of the coating solution, and the surface tension (web – coating solution). These quantities determine the quantity of liquid that the support picks up and carriers along against gravity and the quantity that runs back into the tank. In general, dip coating permits speeds of only 5 – 20 m/min. For a mathematical physics treatment of coating formation, see [30], [248][249][250][251][252][253][254].

Air Knife. In this process (Fig. 35 B) [255–258], excess coating material is removed with the help of a slot-shaped air nozzle; the nominal coating thickness is controlled by the intensity of the air blast. Emulsions coated by this method can have higher concentrations than in dip coating; as a result, drying capacity can be reduced or coating speed can be increased.

Roller Coater. Coating with pickup rolls (Fig. 35 C) is subject to conditions similar to those in dip coating (Fig. 35 A), but the coating thickness further depends on the spacing between pickup roll and emulsion support as well as the rotation speed and direction of the pickup roll. Higher coating speeds are possible with this process.

In the first three processes, the coating solution and the runoff must be recirculated to the coating tank in order to prevent concentration variations due to sedimentation and other factors. The remaining processes operate without solution return. The coating solution is precisely metered onto the support by way of measuring devices (capillaries, pumps). For this reason, the coating thickness does not depend on the viscos-

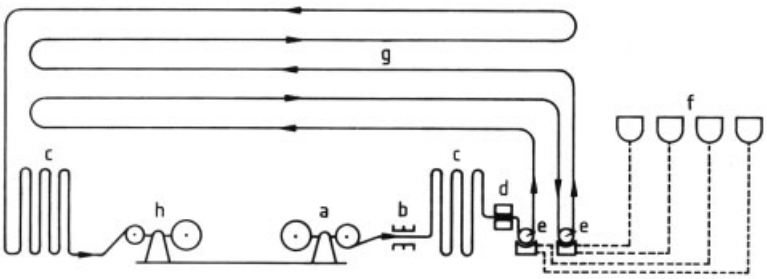

Figure 34. Schematic diagram of a coating machine for film or paper webs
a) Unwinding station; b) Web splicing; c) Storage rolls; d) Cleaning and pretreatment; e) Coating; f) Emulsion feed tanks; g) Setting and drying section; h) Windup section

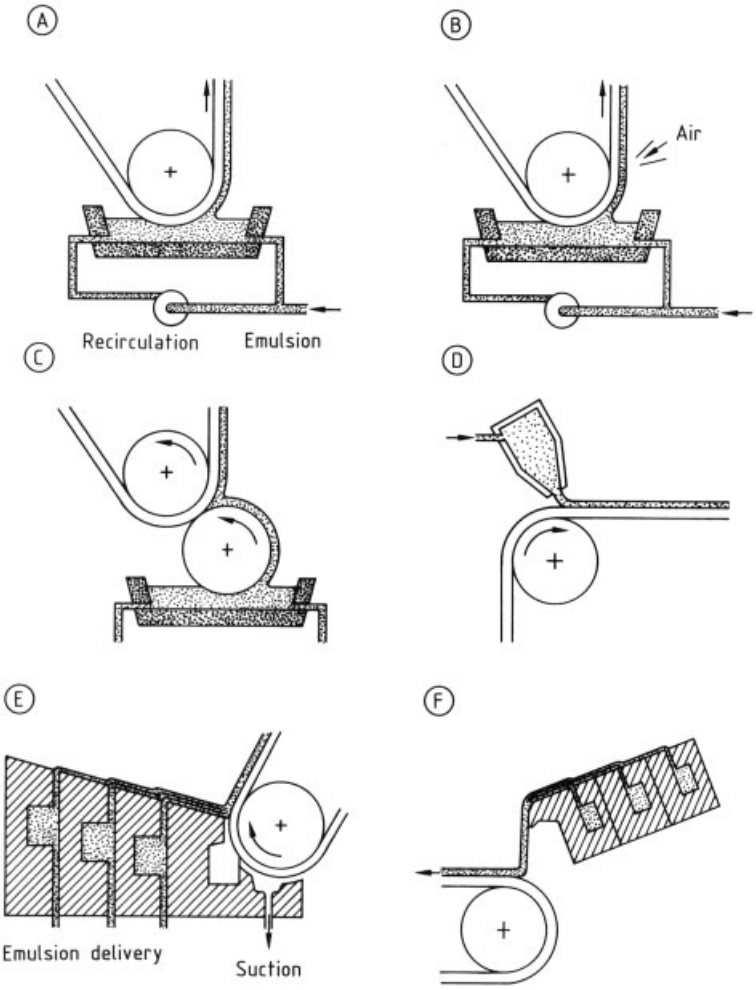

Figure 35. Schematic diagrams of processes for coating photographic emulsions on films or papers
A) Dip coating; B) Air knife; C) Roller coater; D) Extrusion coater; E) Slide hopper coater; F) Curtain coater

ity, but on three primary factors: metering rate, coating width, and speed of support.

Extrusion Coater [259], [260]. (Fig. 35 D). The uniformity of application across the web width depends strongly on the accuracy of the extruder die lips and the pressure distribution.

Slide Hopper Coater. The slide [261] (Fig. 35 E) and curtain coaters (Fig. 35 F) [253] can apply several coatings at once. The materials flow in superimposed layers down the cascade to the coating point. They do not mix because the flow is laminar. Metering devices convey the emulsions to the coating head; they flow in layers over the sloping top surfaces, form a "bead" (meniscus) at the point of contact with the moving support, and are stretched and picked up by the web of material. The wet thicknesses, viscosities, and surface tensions of the many emulsion layers must be suitably related to one another. The slide coating process is economic because it can produce multilayer coatings at high speeds (50 – 250 m/min) and it uses concentrated coating materials.

Curtain Coater. Curtain coating (Fig. 35 F) [253], [254], [262], [263], permits still higher speeds. Because of the wide spacing (20 – 200 mm) between the coating head and the emulsion support, it is possible to work past gaps or irregularly shaped supports without breaking up the curtain of falling material. The thin curtain must be specially protected, because it can very easily oscillate, which may cause process upsets and coating flaws. Lateral edge guides prevent the contraction of the falling curtain.

General. The coating systems described all require exact control of the *solution viscosity*. This can be achieved by varying the concentration, selecting the proper gelatin grades, adding other high-molecular-mass thickeners, or changing the temperature.

The *surface tension* is adjusted with wetting agents [252]. Alcohol is often used, but surface-active compounds such as saponins, potassium salts of acid alkenyl succinates, sodium salt of diisooctyl sulfosuccinate, salts of perfluorinated carboxylic acids, and other surfactants are also suitable for the purpose.

The operation is sometimes carried out at several coating stations, and thus possibly with a variety of coating equipment. In these cases, the layers first applied must be dried (setting may be sufficient) before the support comes to the next coating station, and so forth. If the coatings are not dry, the coating side must not be mechanically contacted.

Coating thickness is measured by radiography in a nondestructive manner during the coating step. The measured value can then be used to control the operation. The radiation dose involved is so low that it produces no developable blackening of the photographic material.

3.3.3. Chilling and Drying

The emulsion support with the gelatin solutions, applied warm, is conveyed from the coating station into the chilling zone, where the emulsions are soldified, often with the aid of cold, dry air (5 – 15 °C).

The geometry of the chilling section and the quantity and velocity of the cooling air must be selected to meet two criteria: Maximum heat transfer must take place, but the coating must not be damaged by the air blast. In order to maintain the coatings in a uniform state, it is crucial that they be transformed from a liquid state to a nonflowing state as quickly as possible. The capability of modern coating machinery to work with more highly concentrated solutions, dispersions, and the like is again an advantage, because these solutions set faster. Water is removed from the solidified emulsion in the downstream drying step. The time dependences of the emulsion temperature and moisture content, which are controlled through the parameters of the drying air, have a substantial effect on the photographic and mechanical properties of the photographic material.

On cooling, gelatin solutions go from the sol form – in which gelatin molecules are randomly distributed in the solution – to the gel form, in which the gelatin molecules take up a partly crystalline configuration [264]. This change usually takes more than two minutes and requires temperatures under 30 °C. The gel state has, for example, markedly better mechanical (e.g., lower brittleness), hardening, and swelling properties.

Drying involves two main steps [265–267]. In the first, water evaporates from the gelatin solution as from a free water surface. This first, physical drying step comes to an end when the gelatin concentration in the aqueous solution reaches ca. 30 wt %. In the next step, the evapo-ration rate is limited by diffusion of water (or other solvent) through the coating.

The temperature of the coating during the first drying stage is equal to the wet-bulb temperature of the drying air; in the second stage, when the surface of the coating becomes dry, it can reach the same temperature as the air.

When drying is complete, the photographic materials are carefully brought to predetermined moisture and temperature values in order to prevent static discharges during later operations or during use.

The cleanliness of the drying air is very important. Purification to remove gaseous impurities, such as hydrogen sulfide, nitrogen oxides, and ozone, is necessary, as is dust removal.

As in industrial climate-control systems, *moisture* is removed from the drying air by one of two methods: freezing followed by temperature adjustment, or with lithium chloride absorbers.

The moisture content of the traveling web of material in the several drying zones is often measured by IR absorption [268]. The *temperature* of the emulsion layer during the drying operation is also measured in a contactless manner, for example with pyrometers.

Increasing travel speeds and dryer capacities and the trend toward contactless conveying have led to drastic changes in the engineering design of the drying section. Film and paper webs were formerly conveyed slowly through the drying zones in festoons supported by recirculating pins, or drying wheels were used where the dryer length was restricted. Today, vertical or horizontal flat-layer dryers, several 100 m in length, some designed as floater dryers, are used. The drying air is blown onto the web through plates provided with round holes or slot-type nozzles.

Once the final coating and drying operation is done, the material is subjected to final inspection, put in temporary storage (as required for the specific product), and converted to salable form. The finished product is then obtained by cutting, stamping, winding, cementing, printing, packaging, and other operations.

3.4. Packaging

The term packaging here refers to all the operations carried out after the stock is coated in order to get the product ready for sale.

Film. Film is either cut into sheets or made into a wide variety of roll films. *Sheet film* [269], used chiefly by professional photographers, holds only a small share of the market. Cutting is performed with oscillating cutters or with single strip cross cutters working on film that has already been slit and wound. Notches are cut on the right hand side in the top edge of each sheet to identify the emulsion side and film type. Paper – aluminum – polyethylene composite film (Paalpo) is used for light-tight, climate-proof packaging.

Other than disk film [270], all films for amateur use are packaged in *roll form*. The most common sizes are 120, 220 [271], 135 [272], [273], 126 [274], and 110 [275] roll film. The sequence indicates the order in which these systems were devised. The Super-8 system of amateur motion-picture film now plays only a marginal role, because of the development of video technology. After some initial successes for the 126 and 110 systems the 135 system has achieved market dominance (more than 2×10^9 units per year worldwide). The manufacture of 135 films involves the following steps:

1) Production of film magazines
2) Slitting of film to 35 mm width
3) Perforating
4) Edge-mark printing
5) Loading of film into the film magazine
6) Packaging in plastic can or film, cardboard box, and shipping box

Modern high-performance equipment links these operations in a single series. Information in the edgeprint typically includes the frame number, the film type, the lot number, and even the image frame marks (126 and 110 sizes). Machine-readable barcodes, such as the DX code [273] and the frame number in coded form, have also been included. To edge-print the frame numbers with DX code, the film is brought into contact with specially prepared stencils and exposed. Edgeprints (e.g., film type and lot number) recurring at regular intervals and that are

independent of the film length are applied with egdge-printing drums. The film passes over the rotating drum, which has a light source inside it and a stencil between the light source and the film. Slides with a wide variety of characters make it possible to change film type or lot number information quickly. New edge-printing devices employ a matrix arrangement of many individual small light sources [276], which are illuminated by a control system in such a way that the desired message, consisting of single pixels, is imprinted on the film. The on-time of the individual light sources is only a few microseconds, so that the film can be imprinted on the fly. This type of edge-printing is freely programmable and thus very flexible, and it permits extremely short changeover times.

There are two fundamentally different ways of *loading film into magazines*. The most widely used is to wind the film on the spool outside the magazine, insert the spool and film into a magazine shell that has only one cap in place, and finally close the magazine with the other cap. In another technique [277], the magazine has two caps and the spool for the film, and the film is fed in through the magazine slot. This method can be used only with hook spools [278].

Photographic Paper. Photographic paper is also cut into sheets or slit and wound into rolls. Virtually all black-and-white papers are packaged in sheet form; color papers usually in the roll form.

As in the case of film, oscillating cutters are most often used for cutting the photographic paper into sheets. These machines have a high output (ca. 200 cuts per minute) and offer good cut quality. The single sheets are stacked (usually 10, 25, 50, or 100 sheets), enclosed in light-tight, climate-proof polyethylene or Paalpo pouches, and finally packaged in cardboard sleeves or boxes with identifying labels. Because of the huge number of possible combinations of format [279], gradation, surface finish, and number of sheets, there is a wide variety of stock items in sheet paper.

The preparation and packaging of photographic papers centers on color paper in roll form [280]. Most equipment used for this purpose is highly automated and very productive. High-performance roll slitting machines have an individual windup station for each paper strip. This design feature makes it possible to automate the entire operation, including the following steps: cut off when roll length is reached; remove rolls and convey to packaging; deliver, insert, and clamp new cores; and start winding strip on core. Using belt wrappers no glue or tape is needed [281]. The operations listed above are usually performed with the machinery stopped, but in modern systems they require only a few seconds. Travel speeds of 200 m/min and higher are common. All slitting operations are done with circular knives. Regular regrinding of the blades is vital if dust-free cutting and high edge quality are to be maintained. The top and bottom knife shafts are therefore integrated into a unit that can be changed out all at once; in this way, downtime for knife changing is kept extremely short, and maintenance and precise adjustment of the knives is simplified. Special equipment has been developed for splicing the photographic paper and then for punching long holes to mark the splices [282], [283]. Roll papers are usually packed in light-tight, climate-proof pouches made, for example, of Paalpo and kraft paper. Automatic horizontal bag forming – filling – sealing machines (e.g., those made by Omori, Japan) are used; bag materials are fed from rolls. The machines have adjusting motors so that they can be reset with a change in roll size. All the subsequent packaging and handling operations, up to the stage where the product is on pallets ready to ship, are also fully automated.

4. Black-and-White Photography

[284]

4.1. Negative – Positive Processes

Black-and-white (B & W) films are generally made with a cellulose triacetate support; PETP is used only for sheet films. Packaging and sizes are the same as for color films.

The structure of B & W films consists of one or two emulsion layers and protective overcoat, which may contain matting agents and friction-control substances.

Along with stabilizers, the *emulsion layers* may contain development accelerators, such as ionic surfactants based on poly(alkylene oxides) or quaternary salts of $C_{11} - C_{16}$ amines, for example laurylpyridinium bromide.

Some films, such as Agfa APX-25 and APX-100, have an *antihalation coating* between the emulsion support and the emulsion. These coatings must be completely decolorized in processing (by the action of the sulfite or of the alkali in the developer).

Roll and sheet films have an *NC coating* on the back. In general, this coating is dyed for antihalation protection and must satisfy the same bleachability requirements as the antihalation coating. APX-25 and APX-100 both have a separate antihalation coating and a dyed NC coating and thus exhibit excellent protection against halation.

The base for miniature films is dope-dyed gray and has a neutral density of 0.24. In many products, this dye is the sole provision against halation; it also prevents light piping (see Section 3.2.1.2).

Some B & W films are also *hardened* with formaldehyde; hardening in these materials continues for one to three years and results in sensitometric changes. The one-time hardening sometimes practiced does not have this drawback.

The gradation of B & W films can be varied over a wide range through the development time. Sensitivity and gradation also depend strongly on the developer composition. The manufacturers deal with this fact by publishing curves of gamma versus development time (for a definition of gamma see Section 6.1.2). For subjects having average contrast, the desired value is $\gamma = 0.65$. Densitometric measurements of B & W films are performed with diffuse light so as to eliminate the Callier effect, which leads to an increase in density as a function of the specularity of the light used for measurement.

Prolongation of the development time always makes the characteristic curve steeper and increases the sensitivity, since small latent-image centers become developable. Special methods of attaining maximal sensitivity are based on limiting the increase in steepness (retarded development; see Section 4.3.2).

The *"zone system"* was devised by A. Adams. Its basic principle is matching of the gradation of the negative to the subject contrast as measured with a spot exposure meter. Since development time differs from picture to picture, depending on subject contrast, the process can be used only with sheet film. When the resulting negatives are enlarged and printed on grade 2 paper, the prints show both deep shadow and detail in the highlights.

Excellent edge sharpness, resulting from adjacency effects (see Section 2.3.4.2), is achieved when technical films, which have $\gamma = 1.5 - 2.0$ in normal processing, are developed to a pictorial gamma value of ca. 0.65. Examples of such materials are Kodak Technical Pan 2415 and Agfa-Gevaert A O – 25. Development is carried out in Phenidone – sulfite or in highly diluted standard developer. All such processes with "restrained" development are susceptible to development fog, because convection at the film surface strongly influences the quantity of silver developed out.

In general, B & W films are *panchromatically sensitized*, with a range of ca. 400 – 640 nm. Special materials offer sensitivity at longer wavelengths, up to 690 nm (Technical Pan) and 740 nm (AP-200 S). The advantage of this sensitization is the high sensitivity when a red filter is used; the disadvantage is that some tones, such as red lips, are reproduced too pale. Under the heading of special materials, Kodak High Speed Infrared 4143 deserves mention. This film is sensitive to 900 nm and even behind IR filters, which appear black to the eye.

The transformation from colors to gray-scale values always has an element of arbitrariness. Well-defined standard colors are reproduced with different values by different products. The tool for comparison is the color step wedge, in which colored surfaces (yellow, blue, green, red) are each combined with a gray wedge.

In general, blue becomes lighter as the film sensitivity increases, since the volume sensitivity (intrinsically blue) increases more rapidly with grain diameter than does the sensitized sensitivity, which is proportional to the surface area. As a consequence, the sensitivity of high-speed films declines markedly in incandescent light, which contains less blue.

The gray values can be controlled in any desired way with filters used at exposure time. Yellow, orange, or red filters are most often used to cause blue sky to appear dark. Red filters give the impression of clearing haze, since long-wavelength light is scattered least on aerosol particles. These filters are frequently used in aerial photography. Aerial survey films are therefore sensitized in the long-wavelength red. Filters for

B & W are much more strongly colored than the correction filters used in color photography.

Figure 36 presents typical density curves for optimally ripened emulsions that differ in grain size but have the same coating weight. As the grain size increases, the sensitivity increases while the maximum density decreases. This behavior reflects the fact that the covering power is strongly dependent on the morphology of the developed silver. One compact grain absorbs less light than a number of small grains having the same total volume. Thus the Lambert – Beer law does not apply.

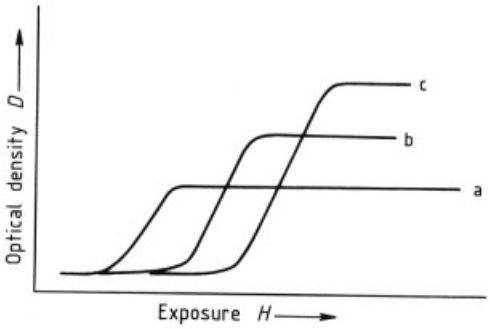

Figure 36. Sensitometric curves of different emulsions after B & W development; coating weight is 3 g silver per square meter for each emulsion
a) Mean diameter 1.2 µm; b) Mean diameter 0.7 µm; c) Mean diameter 0.3 µm

According to SIEDENTOPF and NUTTING, the granularity σ_D also increases with grain size (for a definition of granularity, see Section 6.2). The ratio of sensitivity to granularity is a measure of the efficiency of the photographic primary process and characterizes the quality of emulsion.

Since a maximum density of 3.0 is completely sufficient for pictorial photography, low-speed films can have a lower coating weight than high-speed ones. Typical values are 3 – 4 g $AgNO_3$ per square meter for ISO 25/15° films, 6 g for ISO 100/21°, and 9 g for the ISO 400/27° class.

The loss of sharpness with increasing sensitivity results from the heavier coating weight and, in practice, limits the coating weight. Higher sensitivity is thus achieved at the cost of signal to noise ratio, as in other detection systems. This situation has nothing to do with the quantum nature of visible light, for the "quantum noise" is far less than the granularity of ordinary B & W films. In contrast, the quantum nature of the radiation plays a role in producing the granular structure in X-ray films.

In all conventional B & W films, the granularity σ_D increases with the negative density D. The rise in σ_D is usually less than predicted by the Siedentopf relation $\sigma_D \sim \sqrt{D}$, because at higher densities even the smaller emulsion grains contribute to the image. The practical consequence is a fundamental rule of B & W photography: i.e., use the least exposure possible.

Overexposure also works to the detriment of sharpness. This effect depends on the type of developer because the image silver is unevenly distributed in the emulsion layer. At low densities, it is more highly concentrated at the film surface than near the emulsion support. Near the support scattering on silver halide grains has the greatest effect of reducing the light modulation. Special techniques for achieving maximum sharpness, namely the use of "surface developers," tend to produce the image chiefly at the grain surface. This principle always has the drawback that the characteristic curve departs from that for standard development, since higher densities cannot be formed without the grains in the bulk of the emulsion.

Three sensitivity classes have been established in B & W photography:

1) The ISO 25/15° class offers the least granularity and the greatest sharpness; Agfa APX-25 is an example. Such films are used when a stand is available; for example, to reproduce continuous tone images or in architectural photography.
2) The ISO 100/21° class permits the use of a handheld camera in normal sunlight and is therefore the most frequently used.
3) The ISO 400/27° class finds use not only with poor lighting, but also for subjects in motion, as in sports photography. These films are the standard materials of press photographers.

T-Max Films. Tabular emulsion grains, in comparison with those of a more spherical form, offer higher silver covering power as well as lower light scattering, provided the long axes of

the T-grains can be oriented parallel to the film plane. Kodak T-Max P3200 has a high-speed layer containing crystals with a lateral diameter of 5 – 6 µm and a thickness of ca. 0.2 µm (aspect ratio around 20 – 30). The material has a speed of ISO 1000/31° and a sharpness comparable to that of the ISO 400/27 °C class (see Fig. 37).

Chromogenic B & W Films. In 1980, Ilford and Agfa simultaneously presented B & W films (Ilford XP-1400 and Agfa Vario XL) that produced a monochrome, approximately gray negative in chromogenic development. The gradation of these films, when developeed by the standard process for color negative films, is ca. $\gamma = 0.90$.

The departure from the normal gamma value of 0.65 for conventional B & W films is due to the absence of the Callier effect in chromogenic films, which produces an increase in effective gradation on enlargement with certain types of light source. The advantage of color processing is that it is widely available and the conditions are generally precisely controlled. The user sacrifices some creativity in extending latitude of development, but gains better reproducibility of speed and gamma.

The decisive factor is that the granularity σ_D of chromogenic films decreases on overexposure, in contrast to conventional B & W materials (see Fig. 38). The shape of the granularity versus density curve is a general property of modern color negative materials containing DIR couplers. The reasons are DIR action (see Section 5.2.2.5) and double-coating.

Chromogenic B & W materials have dye couplers for cyan, magenta, and yellow, as well as DIR couplers, in two panchromatic emulsion layers. The higher-speed emulsion layer has a markedly lower content of dye couplers (relative to the silver content) than the lower-speed layer (double- coating principle). In all other details, the films are identical with conventional color negative materials.

Because of the lower content of dye couplers, all dye coupler available in the layer is coupled from a certain exposure value on and the statistical grain distribution of the high-speed layer is not imaged as a density variation. This configuration corresponds to a detection system whose sensitivity is matched to the signal intensity.

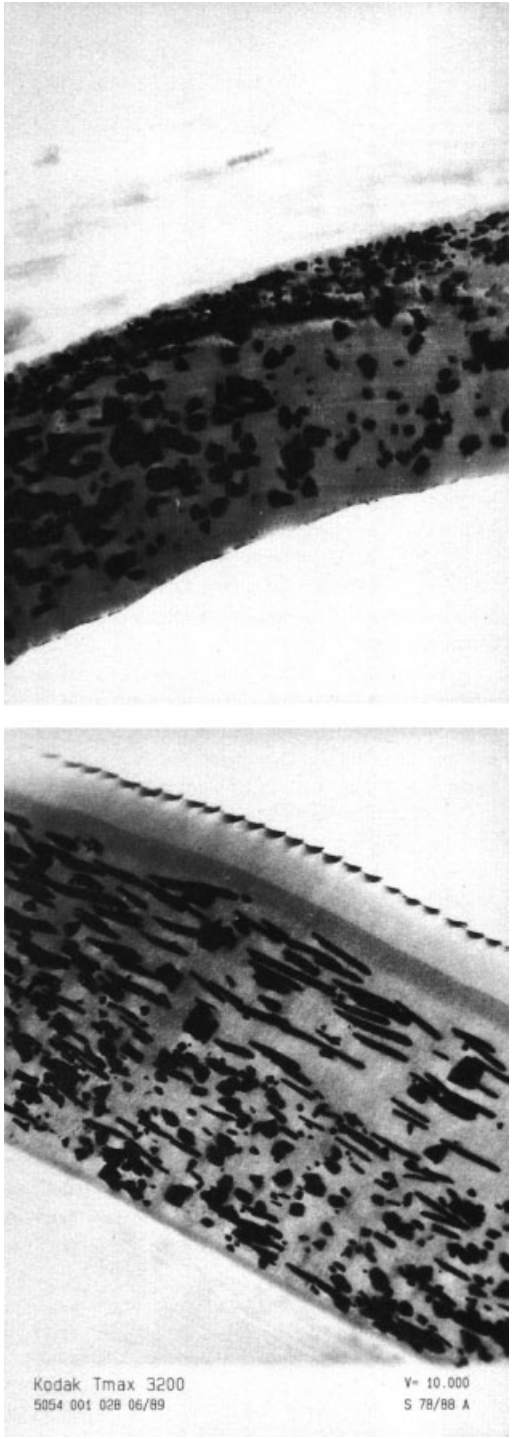

Figure 37. Electron micrographs of conventional ISO 400-B & W film (top) and Kodak T-Max P 3200 containing tubular grain emulsion (bottom)

A chromogenic B & W film with a speed of ISO 400/27°, exposed as if it were ISO 100/21°, gives images with significantly finer grain than the same film normally exposed, thus attaining the image quality of an ISO 100/21° B & W film.

4.2. Reversal Processes

Kodak and Agfa currently offer films specially designed for *B & W reversal processing* (see Section 4.3.3.7). The speeds are ISO 16/13° (Agfa – Gevaert Dia Direct) and ISO 100/21° (Kodak Direct Positive Panchromatic Film 5246). In principle, any B & W film can be developed by a reversal process; the quantity of silver halide solvents in the first developer must simply be matched to the film type.

Direct positive-working films, which give a positive image in normal developer, have speeds roughly equal to those of B & W paper and are suitable for reprographic applications (Kodak B/W Duplicating Film 4168).

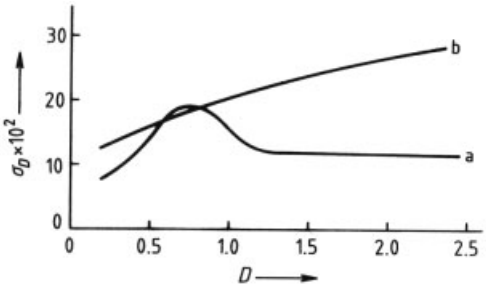

Figure 38. Typical plot of granularity vs. optical density for chromogenic (a) and conventional B & W films (b) of equal speed

4.3. Processing of Exposed Black-and-White Materials

Two types of development, chemical and physical, are distinguished (see Section 2.3.1). These historical names are not wholly correct and may be misleading. The chemical processes involved in them cannot always be sharply separated.

In its purest form, *physical development* takes place in the development of prefixed silver halide layers. The exposed emulsion is treated in a solution of sodium thiosulfate (hypo) containing silver bromide, until all the silver halide has dissolved and only the photolytically formed silver centers remain in the layer. After an intermediate treatment with a sodium sulfite solution, the layer is developed in a solution of Metol (**44**, see Section 4.3.1.2) and sodium sulfite, which contains a dissolved silver sulfite complex. At the sites where nuclei of photolytically formed silver are located, the silver sulfite complex is reduced to metallic silver [285].

Physical development is mainly of theoretical interest, although underdeveloped negatives can sometimes be intensified by subsequent physical development. Physical development is important as a subsidiary process to normal "chemical" development and occurs in developers that contain sufficient amounts of silver halide solvents, such as sodium sulfite.

4.3.1. Developing Agents

There are many reducing agents for silver halides, but only a few can be used as developing agents [11, vol. 2, pp. 813 – 926]. Developers must be capable of reducing only those silver halide grains that contain a latent-image center produced by exposure. Silver halide crystals having no latent-image center but containing fog nuclei or no image centers at all must be reduced very slowly or not at all. In order to attain this goal, two basic criteria must be met: (1) The *kinetic condition* is that crystals with latent-image centers must be developed (under the catalytic action of the centers) orders of magnitude faster than crystals having no latent-image centers. (2) The *thermodynamic condition* for development is that the potential of the developing redox system must be more negative than the Ag^+/Ag potential. Only in this way can a photographic material be developed to form an image.

Developing agents can be inorganic or organic, but only organic substances are of any practical value.

4.3.1.1. Inorganic Developing Agents

In 1877 C. Lea discovered the developing action of potassium dioxalatoferrate(II),

$K_2[Fe(C_2O_4)_2] \cdot 2\,H_2O$. This developer was frequently used in the past, but today it is merely of historical interest, along with the salts and complex salts of chromium, molybdenum, tungsten, uranium, copper, and titanium. Their loss of importance is due to their low activity, rapid exhaustion, and poor stability of the developer solution. The vanadium(II) developer first described in 1894 by A. and L. LUMIÈRE can be electrochemically reduced and continuously regenerated; it was proposed for the development of motion-picture film because of its very rapid action [286]. It has not come into widespread use because development is only possible in hydrogen bromide solutions which are highly corrosive.

A second group of inorganic developers, which are mainly of theoretical interest and useful at best for the development of special-purpose emulsions, contain as developing agents sulfite, reductants based on sulfinic acid derivatives, (marketed, for example, as Rongalit by BASF), hydrogen peroxide in alkaline solution, hydroxylamine, or hydrazine.

4.3.1.2. Organic Developing Agents

Organic developing agents possess hydroxyl or amino groups. Their general constitutional formula was published by LUMIÈRE and SEYEWETZ in 1891 and ANDRESEN in 1892 and expanded later by KENDALL and PELZ. The most general form is [287]:

$$Y\text{-}(A{=}B\text{-})_n\,Y'$$

where Y and Y′ an represent OH, NH_2, NHR, or NR_2 groups.

Most developing agents can be placed in one of the following four groups.

1) $n = 0$: hydrogen peroxide, hydroxylamine, hydrazine (see Section 4.3.1.1)
2) $n = 1$, A = B = carbon: ascorbic acid, catechol, pyrogallol, 2-aminophenol, *o*-phenylenediamine and its substitution products, and 4-amino-1-phenyl-5-pyrazolone derivatives.
3) $n = 2$, A = B = carbon: hydroquinone (**43**), 4-aminophenol, *N*-methyl-4-aminophenol (**44**), and *p*-phenylenediamine and its derivatives.
4) $n = 1$, A = carbon, B = nitrogen: 1-phenyl-3-pyrazolidone (**45**) and 1-(*p*-aminophenyl)-3-amino-2-pyrazoline.

The most important of these compounds for B & W photography are hydroquinone (**43**); N-methyl-4-aminophenol, whose hydrogensulfate (**44**) is known under the trade names Metol (Agfa) and Elon (Kodak); and 1-phenyl-3-pyrazolidone (Phenidone, **45**) and its derivatives.

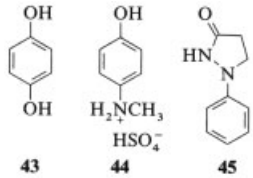

The development of exposed silver halide by hydroquinone can be represented as

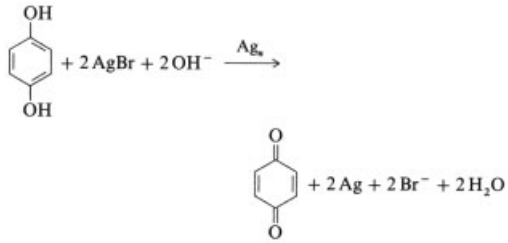

The reaction with Phenidone is

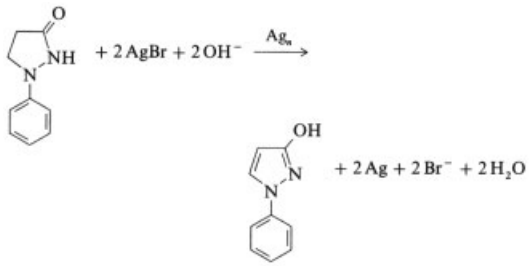

and the general reaction is

$$\text{Red} + 2\,\text{AgX} + 2\,\text{OH}^- \xrightarrow{\text{Ag}_n} \text{Ox} + 2\,\text{Ag} + 2\,\text{X}^- + 2\,\text{H}_2\text{O}$$

Black silver metal, or image silver, is produced from the nearly colorless silver halide. From the reduced form of the developer (Red), its oxidized form (Ox) is generated. In color de-

velopment, the oxidized form of the developer is responsible for the color, but in B & W development these oxidation products play no role. It has been suggested that the oxidation products of B & W developers can be used to intensify the silver image through intermolecular or intramolecular dye formation; one purpose of doing this would be to save silver [288].

Hydroquinone, one of the earliest B & W developers, is still important (→ Hydroquinone, Chap. 7. In high-contrast special-purpose developers, hydroquinone alone is used at a high pH. It is a principal ingredient, with Metol (**44**) or with Phenidone (**45**), in most B & W developers, where it has a superadditive effect (see below).

Catechol (1,2-dihydroxybenzene; → Phenol Derivatives, Chap. 2.) is employed in sulfite-free tanning developers, which not only reduce the silver halide but also (by means of catechol oxidation products) tan the unhardened gelatin in regions where development has taken place, thus rendering it insoluble.

N-Methyl-4-aminophenol sulfate (Metol) (**44**) was introduced by A. Bogisch in 1893 in the form of Metol – hydroquinone developer. It has lost much of its earlier importance to Phenidone. Metol is used by itself in soft-working developers. In combination with hydroquinone, it was commonly used for decades as a powerful, fast-working developer. Because it allowed development even in acidic media, it was used to initiate acidic physical development.

4-Aminophenol in strongly alkaline solution is employed as a one-time developer for amateur B & W films (tradename: Rodinal).

Developers containing 4-hydroxyphenylglycine (called glycin developers) have not retained their former importance, nor have the slow-working but fog-free Amidol developers (containing 2,4-diaminophenol) although the latter are still employed for nuclear track emulsions.

Phenidone (**45**) (1-phenyl-3-pyrazolidone) was discovered in 1940 [289]. It is the only heterocyclic compound that has achieved great value as a developer. It exhibits a very strong superadditive effect with hydroquinone, so that it can be used in small quantities. In comparison with Metol, Phenidone has the advantage that its primary oxidation product, a radical, does not react with sulfite and thus is not withdrawn from the development cycle. Phenidone has the disadvantage of being sensitive to hydrolysis in alkaline media. Derivatives of Phenidone that are mono- or disubstituted at the 4-position are, however, resistant to hydrolysis.

The phenimines, or substituted 1-phenylpyrazoline derivatives, have an amino group at the 3 position (**46**). They include developing agents that show a very strong superadditive effect with hydroquinone. In comparison with Phenidone, they have the advantage of short latency periods [290]. Developing agents taken from the class of 3-alkyl-4-amino-1-phenyl-5-pyrazolones (**47**) have proven useful in automated developing equipment that must handle a variety of emulsion, because these developers are relatively insensitive to variations in the bromide content. They have also been proposed for the development of nuclear track emulsions [291], [292], [293]. Neither of these types of developing agents has acquired any market importance.

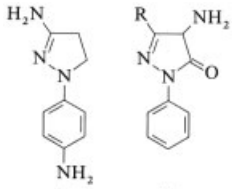

46 **47**

Heterocyclic developers containing oxygen include ascorbic acid (vitamin C) as well as the structurally similar hydroxytetronic acid and its imide [294]. These substances, whose oxidation products are colorless, are mixed with Metol or Phenidone to obtain special-purpose developers. For example, sulfite-free Metol – ascorbic acid developers are employed when only the latent image on the surface of the silver halide grain is to be developed.

Superadditivity. If an exposed silver halide emulsion is developed more strongly by a combination of two developing agents than the sum of the developments by the agents singly, this effect is called superadditivity. For decades the most important developer pair with superadditivity was Metol and hydroquinone, but this combination has been replaced by the even more strongly superadditive pair Phenidone and hydroquinone.

This effect is accounted for by a regeneration process or the removal of the charge barrier. The nitrogen- containing developing agent, Metol or Phenidone, is preferentially adsorbed by silver metal and silver halide. In exposed regions of the halide, the developing agent is oxidized to a radical while silver halide is reduced to silver metal. In this way the latent-image center becomes larger.

One explanation for the superadditive effect is that the radical adsorbed to the silver halide grain is reduced to the original developer molecule by the hydroquinone dianion, so that development can take place again (*regeneration process*).

But it is also conceivable that the increase in the size of the latent-image center locally breaks down a charge barrier at the surface of the silver halide grain, making it easier for the hydroquinone dianion to reach the crystal surface and thus promoting development by hydroquinone [295].

In the most recent theory, both reaction mechanisms occur in the Phenidone – hydroquinone pair. At high pH and low Phenidone concentration, the elimination of the charge barrier would play the major role; at low pH (ca. 7 – 8), regeneration would dominate [295].

4.3.2. Developer Solutions

The term "developer" generally denotes an aqueous solution containing all the substances needed for the development of a photographic material. The essential constituents of developer solutions are the developing agent proper and a variety of buffers and protective substances.

Photographic emulsions exist that already contain developing agents. The reason why this presents no difficulty is that most developing agents reduce exposed silver halide only when in an alkaline medium. These emulsions are made weakly acidic and thus cannot be reduced. Such emulsion layers are stable and must simply be placed in an aqueous alkali solution for development. Development begins as soon as the developing agent dissolves in the inward-permeating alkaline solution. The process has the advantages that development takes place very rapidly and the material cannot be overdeveloped.

4.3.2.1. Developer Formulas

A great number of developer formulas are known; often they differ in only small ways. For example, all manufacturers of photographic materials formerly published the formulas they considered best. Different formulas are naturally suggested for different intended results and processing conditions. While development used to take place only in trays and in tanks at temperatures of ca. 20 °C, processing temperatures have now risen (to 50 °C) and automated devices and machines have become more important.

In order to save labor and achieve constant results, premixed processing chemicals are preferred. Their compositions are often trade secrets, but environmental and occupational-safety regulations now force the disclosure of the composition of several of these premixed chemicals. Such premixed chemicals often contain additives comprising an improvement or technical advance not embodied in the corresponding published formulas.

4.3.2.2. Developer Additives

Besides developing agents proper, developers contain other ingredients to perform a variety of functions.

Buffers. The great majority of developing agents work only at pH > 9. Only in a few special cases is a pH of 7 – 8 high enough. Developers therefore contain bases, especially sodium and potassium carbonates but also borates, phosphates, sulfites, and sodium and potassium hydroxides. Nitrogenous bases can be used only if they have a low dissolving power for silver halides, since otherwise the emulsion will be fogged during development; triethanolamine is an example for a base with low dissolving power. Hydrogen halide is liberated in the development process, and so the buffering capacity of the developer solution must be high enough to keep the pH from dropping. Developers with high hydroquinone concentrations can even be used with an alkali hydroxide alone, since the hydroquinone dianion provides adequate buffering in the pH range from 10 to 12.

Preservatives. Being strong reducing agents, developing agents are very sensitive to oxygen, especially in alkaline solution. The addition of antioxidants is therefore essential in practice. A particularly good choice is sulfite. In the presence of hydroquinone, sulfite reacts with oxygen to form hydroquinonesulfonic acid, sulfate, and hydroxide. In this way, the formation of colored oxidation products of the developing agents is prevented even under incorrect processing conditions. In addition, the capacity of sulfite to dissolve silver halide often results in a desirable effect; i.e., promoting physical development and yielding a fine, compact silver grain. In fine-grain developers containing Amidol or Metol as developing agent, sulfite actually performs three functions: buffer (pH 7–8), preservative, and silver halide solvent helping to refine the grain structure. In direct positive emulsions where the sensitivity centers are inside the silver halide crystals grains (see Section 2.4.3), sulfite facilitates internal development, because it dissolves the outer, undeveloped layers of the silver halide crystal and releases the latent-image centers lying inside so that development can then take place on them. A developer without a solvent for silver halide cannot develop such special-purpose emulsions.

Restrainers. Photographic fog arises when unexposed silver halide is developed. The nonimage silver formed lowers the transparency in films or degrades the highlights in prints. Fog results from a variety of causes and therefore must be combatted in a variety of ways.

Excessive ripening or sensitization of the emulsion can generate silver halide crystals that are developable without exposure (ripening fog, sensitization fog). When the activity of a developer is increased by raising its concentration, the pH, or the temperature, even unexposed silver halide grains are developed. The higher the capacity of the developer to dissolve silver halide (e.g., due to sulfite, thiosulfate, ammonia, and similar agents), the more fog it will produce by physical development on the fog nuclei. The action of atmospheric oxygen on the developer can lead to the formation of hydrogen peroxide or its derivatives when trace heavy metals are present. The same process can initiate the development of unexposed silver halide.

The most important restrainers are sodium and potassium bromides. A number of organic substances (antifoggants) are also used; these include benzotriazole, benzimidazole, and phenylmercaptotetrazole. All these agents form sparingly soluble silver compounds and are strongly adsorbed to the surface of the silver halide grains. Presumably, they stop fogging by retarding physical development, whether through their adsorption to the grain surface (which inactivates the fog nuclei) or because they reduce the concentration of complexed silver in dissolved form. At the same time, however, these additives restrain development, and so the concentration of antifoggants in the developer should be held low and as constant as possible.

The restraining action of some dyes (phenosafranine, pinakryptol green) and some anthraquinone derivatives might be attributed to a direct or indirect redox reaction with the agent that causes fogging.

Water Softeners. Since deionized water is seldom available for preparation of the developer, a sequestrant must be added. In the alkaline developer solution, the magnesium and calcium salts from tap water form insoluble precipitates, which vary in composition and character (hydroxides, carbonates, sulfites, sulfates, borates, phosphates). Floccular precipitates are less of a problem than those that form incrustations. Incrustations occur especially in developing machines, plugging water lines or impairing heat exchange and thus interfering with the proper functioning of the equipment. Such precipitates can also scratch the photographic material.

Suitable sequestrants are both inorganic and organic substances that form highly water-soluble complexes with magnesium and calcium ions. The inorganic compounds include polymetaphosphates (e.g., Calgon) and polyorthophosphates. They have the disadvantage that they slowly hydrolyze to orthophosphates at elevated temperatures, thus forming sparingly soluble salts. Organic sequestrants include hydroxycarboxylic acids, such as citric acid, and aminoacetic acid derivatives, such as nitrilotriacetic acid and ethylenediaminetetraacetic acid. The latter group is gaining importance. Organic phosphonic acids are not sequestrants in the proper sense; they do not prevent precipitation

but merely impede the formation of adherent crusts. They are used for masking iron ions.

Image-Tone Modifiers. The image tone of a B & W image depends on the form of the precipitated silver. The image tone can also be controlled with inorganic and organic developer additives. For example, the addition of a small amount of thiosulfate shifts the tone toward blue, producing a colder image tone. The same effect can be achieved with substituted thioureas and with heterocyclic thioketones; for example, 4-bromo-2,3-dimethyl-1-phenyl-5-thiopyrazo- lone, like thiosulfate, has the ability to dissolve silver halide. Benzotriazole can also produce a blue tone. With some emulsions, the addition of ternary sulfur compounds and quaternary compounds of nitrogen and phosphorus to the developer can shift the tone toward brown. These ternary and quaternary substances are much more effective when added to the emulsion and are therefore added before coating the support.

Development Accelerators. The induction period (see Section 2.3.3.1) can often be shortened, by the addition of quaternary nitrogen compounds, such as laurylpyridinium chloride, or of poly(ethylene oxides), so that development is accelerated.

Organic Solvents. For ease of use, developers are often sold in the form of concentrated solutions, which are simply diluted with water before use. Because organic developing agents are sparingly soluble in water, they are frequently dissolved in organic solvents such as ethylene glycol, diethylene glycol, or triethanolamine. These additives simply act as solvents and have no photographic effect. Triethanolamine, however, can also provide buffering in the developer solution (see page 65).

4.3.2.3. Special-Purpose Developers

Hardening Developers. When photographic materials are processed at higher temperatures (e.g., ca. 50 °C) it is often necessary to add gelatin-hardening (tanning) agents to the developer. Proven additives for this purpose include dialdehydes, such as succinaldehyde or glutaraldehyde, which offer a hardening action even in the presence of sulfite but do not react with developing agents and hence do not interfere with development.

Lith Developers. Developers with extremely high- contrast working are used for lithographic purposes. They must contain only very small amounts of free sulfite, since otherwise the desired lith effect is not produced. However, it is impossible to do without sulfite altogether, for the developer would then be too unstable and oxidation-sensitive. The problem is solved through the addition of sodium hydroxymethanesulfonate ("formaldehyde bisulfite," $HOCH_2-SO_3Na$), or a mixture of paraformaldehyde and sulfite, (which in solution give sodium hydroxymethanesulfonate), or sodium cyclohexanone bisulfite. In the alkaline developer solution, these compounds dissociate partly into sulfite and formaldehyde or cyclohexanone. The sulfite concentration is then high enough to provide adequate protection against atmospheric oxygen without interfering with the lith development (sulfite buffer).

Fixing Developers or "Monobaths." For special purposes, such as the rapid processing of X-ray, logging, aerial-survey films (for military aerial reconnaissance), or microfilm, fixing developers have been devised. Fixing developers contain enough silver halide solvent to remove the unexposed silver halide (fixation) during the ordinary development time. Development must be largely complete before fixation commences. Some loss of sensitivity and contrast is unavoidable. In order to restrain physical development, which would result in unacceptable fogging, such monobaths contain relatively high levels of bromide (up to 10 g/L).

In addition to sodium thiosulfate, aliphatic and heterocyclic thiols, can be used as fixers which form highly water-soluble silver complexes in the high-pH developer. Modern fixing developers also have additives that prevent or at least retard the formation of silver sludge. Because fixing development is very susceptible to problems and can be applied to only a few materials, it has not come into general use [296].

Paste Developers. Black-and-white developers can be thickened with methylcellulose, carboxymethylcellulose, or synthetic polymers. Such developers are said to give sharper images and much more consistent development than unthickened ones. Despite these advantages, the processing of photographic materials with thickened baths has not gained acceptance except in the processing of optical sound films (thickened sound-track developer, thickened bleach bath). The removal of the paste materials involves serious difficulties and high costs. Significant amounts of thickening agents must be added to the baths, so that the wastewater has a high suspended solids content and a high oxygen demand. Both factors play a role in wastewater treatment cost calculations. Suitable replenishment processes free of nonrecoverable pastes are unknown.

4.3.3. Postdevelopment Treatment

4.3.3.1. Stopping

Stop baths are used to break off the development process quickly and as uniformly as possible. With adequate replenishment, this bath also provides an intermediate wash, preventing large amounts of developer getting into the fixing bath, and thus preventing additional fog.

Acidic Stop Bath. A proven stop bath is 0.5 – 2 wt % aqueous acetic acid, which may also contain a gelatin-hardening agent such as formaldehyde or an aluminum salt (hardening stop bath). Development is cut off by the nearly instantaneous drop in pH in the emulsion layer upon addition of the stop bath.

Neutral Stop Bath. Development can also be stopped by quick fixation of the silver halide with a sodium or ammonium thiosulfate solution (stop – fixing bath). Thiosulfate solutions may also contain an aluminum salt for hardening (hardening stop – fixing bath).

4.3.3.2. Fixation

Undeveloped silver halide must be dissolved out of the photographic material, or it will continue to develop (darken) in light. Of the known solvents for silver halide, such as ammonia, thiourea, heterocyclic mercapto compounds, concentrated alkali halide solutions, cyanides, thiocyanates, and thiosulfate, the only ones of any importance in practice are sodium and ammonium thiosulfates. They must be used in fairly high concentrations, since only the thiosulfate-rich silver complexes $[Ag(S_2O_3)_2]^{3-}$ and $[Ag(S_2O_3)_3]^{5-}$ are readily soluble in water. *High-speed fixing baths* always contain ammonium salts, such as ammonium chloride or ammonium thiosulfate, because these baths dissolve the silver halide more rapidly than pure sodium thiosulfate solutions. Potassium thiosulfate cannot be employed because the complex salts it forms with silver are not easily soluble. Fixing baths contain sulfite to improve their stability. Acidic fixing baths contain sodium hydrogensulfite or an alkali disulfite such as potassium metabisulfite (potassium pyrosulfite, $K_2S_2O_5$). They can be used as stop baths, so that no intermediate stop bath is necessary, but the fixing bath must be adequately replenished. *Acidic fixing baths* may also contain aluminum salts such as potassium aluminum sulfate to harden the gelatin, as well as acetates and borates as complex-forming buffer compounds to prevent the precipitation of basic aluminum salts.

B & W papers that give a warm, brown-black image tone when dried at room temperature often have the disadvantage of turning an unattractive gray when dried on a hot glazing drum. To avert this color change ("burning"), heterocyclic mercapto compounds are added to the fixing bath. These are adsorbed to the silver metal and block the color change by preventing the individual silver grains from sintering together. The substances used for this purpose must not interfere with fixation.

The time required for fixing can easily be determined on transparent material. Generally speaking, fixation is complete as soon as the last fog due to silver halide has vanished from the emulsion layer. In the case of opaque materials, the unfixed silver halide can be detected only through intense exposure (photolysis) or by a chemical test in which silver sulfide is formed. Excessively long fixing should be avoided, since fine-grained precipitated silver is dissolved under the action of atmospheric oxygen, causing loss of detail (erosion of highlights).

4.3.3.3. Washing

Thorough washing after fixation is essential if stable B & W silver images are to be obtained. Since any silver thiosulfate complex not removed will gradually decompose to silver sulfide, the highlights will become fogged. (Insufficiently washed-out thiosulfate slowly disproportionates into sulfite and sulfur, and the sulfur oxidizes the image silver to form silver sulfide.) If this happens, a yellow-brown color will appear especially in regions of low image-silver density.

Most photographic papers have resin coatings on both sides of the base (see page 52). Because the processing solutions cannot penetrate into the fiber base, only a relatively short wash is needed.

In developing machines, papers are washed with running tap water; a countercurrent arrangement is used in order to save water. The rate of washing depends on the water temperature (increase in diffusion rate) and the thickness of the swollen emulsion layer (diffusion length). It is also influenced by the quality of the fixing bath, especially in the case of non-resin- coated papers. Acidic fixing baths cannot be washed out of the fiber base as quickly as neutral and, especially, as alkaline baths. When acid fixing is practiced, the papers should pass a soda bath before washing. The soda raises the pH in the emulsion layer and the support. A similar effect is achieved with a sulfite bath after fixation. Thiosulfate residues are destroyed by nonoxidizing substances that react with thiosulfate but do not attack the silver metal; an example is sodium chloroacetate [297]. Thiosulfate and sodium chloroacetate react to form the sodium salt of an *S*-ester of thiosulfuric acid, i.e., a Bunte salt:

$$^{-}O_3S-S^{-} + Cl-CH_2-COO^{-} \longrightarrow$$
$$^{-}O_3S-S-CH_2-COO^{-} + Cl^{-}$$

4.3.3.4. Stabilization

If the image need not meet archival standards but rapid access is needed, the time- consuming process of fixing and washing can be replaced by stabilization of the image. In this operation, only part of the undeveloped silver halide is dissolved; most of it remains as a silver compound having little light sensitivity. These compounds are formed with the components of the stabilizing bath that have penetrated into the emulsion. Acidic stabilizing baths contain thiosulfate, thiocyanate, thiourea, or organic sulfur compounds, such as thioglycolic acid (mercaptoacetic acid), thiosalicylic acid, or mercaptopyrimidine derivatives at relatively high levels.

To prevent yellowing of the silver image, stabilizing baths further contain small amounts of heterocyclic mercapto compounds, such as phenylmercaptotetrazole, which yield sparingly soluble silver compounds. To prevent the formation of colored oxidation products from residual developing agents in the emulsion, these baths also contain reducing agents such as sulfite or formaldehyde. After stabilization, the material should be washed no more than necessary; film materials should get only a very brief wash in order to dissolve salts present on the film surface that would otherwise crystallize and cloud the film. Provided the stabilized image has not become discolored, it can be converted to a form with ordinary stability by subsequent fixing and washing [298].

This *two-bath-process* (tradename Rapidoprint) has lost ground in favor of the standard process of development, fixing, and washing. High-speed developing machines using the conventional sequence of baths, working on B & W papers with incorporated developing agents, yield high-quality images in such a short time that the stabilization process now offers virtually no time advantage.

The term "stabilization" also includes processes by which *already processed B & W materials can be given archival stability*. The stability of the silver image is threatened not just by residual thiosulfate in the image but also by oxidizing gases, which convert the image silver to finely dispersed silver oxide. When this happens, the gray tone, chiefly in regions of low image density, changes to a yellow-brown. Hydrogen peroxide has particularly aggressive action; this gas is mainly evolved from new wood furniture and from the boards on which pictures are mounted. This harmful effect can be suppressed by treatment with a silver stabilization bath after washing. Agents that can be used for this pur-

pose include substances containing sulfur and nitrogen; for example, thiocyanates (tradename Sistan) and thiourea derivatives [299].

4.3.3.5. Reduction, Intensification

Reduction. Negatives that are too dense (as a result of overexposure) to be printed can be rendered copiable through subtractive reduction. In this process, the image silver is diminished by the same amount everywhere. The negative retains its original gradation. The silver is oxidized with a strong oxidizing agent, such as permanganate, dichromate, or persulfate, and then dissolved and removed. The best-known agent is Farmer's reducer, which consists of potassium hexacyanoferrate(III) and sodium thiosulfate.

Proportional reducers can be used to diminish negative contrast. These substances mainly attack the silver in regions of high density; regions of low coverage are only slightly attacked. In this way, the gradation is flattened without any loss of detail in the highlights.

Proportional reducers contain ammonium peroxodisulfate or the sodium salt of the iron (III) – ethylenediaminetetraacetate complex as oxidizing agent. The latter oxidizing agent has the advantage that it does not attack thiosulfates. It reacts with them to form stable reducers; in color photography, such reducers are widely used as bleach – fixing baths.

Intensification. Negatives that have been underexposed or underdeveloped and do not have sufficient coverage to give usable positives can be intensified by treatment with a physical developer or with mercury(II), copper(II), or uranyl salts. Depending on which process is employed, a metal or metal-salt image is superimposed on the silver image.

4.3.3.6. Toning

Many special papers, when normally developed, give silver images with a blue (cold), green, yellow, or brown (warm) tone instead of a neutral black image. It is also possible to convert an already processed B & W silver image subsequently to a colored or toned image.

Toning processes have the drawback that the resulting color tones depend strongly on the quality of the silver image and on the treatment it has received. Some toning processes require bleaching with potassium hexacyanoferrate(III) and potassium bromide. In sulfur toning, in which the bleached silver image is treated with sodium sulfide to produce a silver sulfide image, the tone ranges from yellow to brown, depending on the bromide content of the bleaching bath. Similarly, antimony toning with sodium antimony sulfide (Schlippe's salt, $Na_3SbS_4 \cdot 9\,H_2O$) gives brown to ocher- colored tones, and copper(II) hexacyanoferrate(III) toning gives violet-brown to cherry-red tones. Gold toning baths with gold(III) chloride and ammonium thiocyanate yield blue-black tones without intermediate bleaching; the metallic silver is partly or wholly replaced by metallic gold. If the silver image is initially converted to silver sulfide, reddish tones are obtained. Pure blue images are obtained by transforming the silver image to a Berlin blue image; the image is first bleached with potassium hexacyanoferrate(III) and then treated with iron(III) chloride for this purpose.

Sodium and potassium selenosulfates convert a silver image directly to a silver selenide image with a brown to purplish-brown tone. In one toning process, the silver image is converted to an organic dye image or such an image is superimposed over it. The steps in this process involve converting the silver image to a silver halide image by reacting it with an oxidizing agent in the presence of chloride or bromide, then converting the halide image to a silver image and to a dye image by reaction with a color developer in the presence of a dye component. The resulting image has a blackish tone. Finally, removing the silver in a bleach – fixing bath leaves behind a pure dye image.

4.3.3.7. Black-and-White Reversal Processing

When a B & W film material is to be processed into transparencies, the film is first developed to a normal negative. The silver metal must then be totally removed without the residual silver halide being disturbed. For this purpose, the negative is treated with a dilute, halide-free solution of dichromate or permanganate and sulfuric acid.

The silver is oxidized and the product silver sulfate goes into solution. The film is then treated with sodium sulfite and exposed in bright light. The silver halide image that results is now developed to give a silver image.

4.3.3.8. Replenishment

Photographic baths in processing machines must be replenished if consistently good results are to be achieved. The developer loses activity for a variety of reasons. Part of the developing agent is oxidized in the development process, a second part by atmospheric oxygen, which is not entirely scavenged by the high concentration of sulfite. Developers containing Phenidone decline in activity as this compound is hydrolyzed. In development, bromide ion from the silver bromide is liberated, builds up in the developer, and restrains development. Protons generated in the development process partly neutralize the alkali, so that, for example, carbonate is transformed into hydrogencarbonate.

A "replenisher" is a solution that restores a conventional bath to its original composition.

Developer Replenisher. A developer replenisher for an exhausted developer should contain high concentration of developing agent and sulfite, pH- correcting agents, and a reduced amount or no bromide. A portion of the exhausted developer is discarded and replaced by the replenisher. In this way, the diminished contents of developing agent and sulfite are restored, while the bromide level in the developer, which has been increased through use, is lowered.

Developing machines operate continuously. Accordingly, replenisher is also added continuously, at a rate in keeping with the amount of photographic material developed. Some overflow usually occurs when the replenisher is added. With some developers, the rate of addition (replenisher rate in mL/m^2) can be matched to the rate of solution removal with the developed material. There is no overflow of developer in these systems. The downstream baths, in contrast, always overflow on replenishment, since the emulsion is already swollen on arrival there and leaves each bath with about the same amount of liquid that it brought in.

Such overflows generally cannot be discharged into sewers. They have to be regarded as liquid waste. It therefore makes sense to regenerate exhausted baths chemically (rejuvenation), so that they can be converted to replenishers and returned to the cycle. Manufacturers of photographic chemicals issue instructions and identify sources for automated preparation equipment (see Section 5.3.1).

Replenishment of Acidic Stop Baths. In order to hold the pH constant in acidic stop baths and prevent too much buildup of developer constituents, these baths are replenished with more strongly acidic solutions.

Fixing Bath Replenishers. Spent fixing baths contain large amounts of dissolved silver, which can be recovered by electrolysis (as metal) or by precipitation (as the sulfide or other sparingly soluble compound). The latter methods are not very important, since the electrolytic method yields high-quality silver in a technically advanced and easily run process. Even so, trimercaptotriazine (trade name TMT) holds some interest as a precipitating agent because its silver salt, in contrast to silver sulfide, is readily filterable [300].

The fixing baths in all developing machines are electrolytically desilvered; in most cases, they are freshened by the addition of a rejuvenator (regenerator), which raises the concentration and corrects the composition. There are two approaches. In the first (off-line desilvering), the overflow from the fixing bath is collected, desilvered, treated with rejuvenator, and returned to the process as replenisher. In the second (on-line desilvering), the fixing bath constantly circulates through the fixing tank and the silver-recovery cell. On the way back to the tank, a rejuvenator concentrate is metered into the desilvered solution to replenish it. This process has the advantage that the silver concentration is held fairly low, so that little silver is carried into the downstream wash-water tanks.

Silver carried in to the wash water in the form of its thiosulfato complexes has a high affinity for the anchor groups of anion-exchange resins. It can thus be significantly concentrated on ion-exchange resins. The wash water, entering with a mean silver content of 30 mg/L, is desilvered to around 1 mg/L. The silver is recovered by re-

generating the ion-exchange resin, then putting the regenerant through electrolysis, or by incinerating the resin [301], [302].

Both replenished and rejuvenated baths quickly come to a relatively steady state with respect to the concentrations of bath constituents, reaction products, and other contaminants. As a result, solutions can be recycled for long periods, often more than a year. They are usually not replaced until the developing machine is serviced.

4.4. Silver Halide Diffusion Processes

The term "silver halide diffusion processes" refers to all black-and-white image-forming processes in which a positive is formed by diffusion reversal [303]. If the silver halide in a donor layer is developed in the presence of a silver complexing agent or fixer, and if this dissolved silver complex is converted to a silver image by physical development on pre-existing metal or metal sulfide centers in an acceptor layer, the negative remains in the donor while a positive with high covering power is produced in the acceptor layer.

A suitable material [304] might take the form of a single-sheet or two-sheet material. The donor layer, like the image receiving (acceptor) layer, can be pulled off (stripping film principle) or transferred to another carrier. Gelatin is not suitable as the colloidal medium for the image receiving layer because of the unsatisfactory covering power of the silver; cellulose derivatives, poly(vinyl alcohol), or colloidal silica are better. It is even possible to make the viscous processing solution serve as the image receiver. Fundamental difficulties (e.g., unacceptable or bronzing image tone and instability of the extremely finely divided positive silver to chemical degradation) have now been remedied. Modern two-sheet systems, such as those for graphics use and for scientific or identification photography, no longer require protective varnishing for the positive.

In the silver halide diffusion process, development generally takes place with a limited amount of developer. This feature has a decisive effect on the photographic result; for example, reproducible development is achieved by the application of a paste film of a well-defined thickness or through the observance of a predetermined immersion time before contact with the receiving sheet. Gradation and sensitivity are also strongly influenced by the availability of silver complex formers (fixing agents). The quantity of fixer can be held surprisingly low, because the fixer is continuously regenerated by the formation of positive silver.

B & W instant-photo materials are manufactured by Polaroid and Fuji; those for technical use, also by Agfa – Gevaert. Polaroid further markets B & W instant transparency films for graphic-arts use and for continuous tone reproduction. The Polachrome instant transparency film is a black-and-white material with an additive linear screen. See, → Imaging Technology, Chap. 2.2.3.1. .

5. Color Photography

The purpose of color photography is to reproduce in pictures all the color impressions received by a normally sighted person. Color photography takes advantage of the fact that all perceived colors can be generated by the juxtaposition or combination of just three primary colors in various proportions.

There are two sets of primary colors: additive primaries (blue, green, and red; see Section 5.1) and subtractive primaries (cyan, magenta, and yellow; see Section 5.2).

5.1. Additive Color Processes

Each of the colors blue, green, and red transmits a third of visible light (a third of the solar spectrum) and blocks the other two-thirds. The same holds for light reflected from these colors. Accordingly, color images can be created with blue, green, and red light rays. Three projectors (blue, red, and green) are needed; their light is modulated by three distinct black-and-white transparencies and projected onto a white screen in a dark room (Fig. 39).

To create a color image, suitable B & W emulsions and color filters (blue, green, and red) must be used to make three separate images, which are processed into B & W transparencies. The transparencies are placed in three projectors together with the respective color filters, and the three images are projected one over another. Prints (for

viewing by reflected light) cannot be made by the additive process.

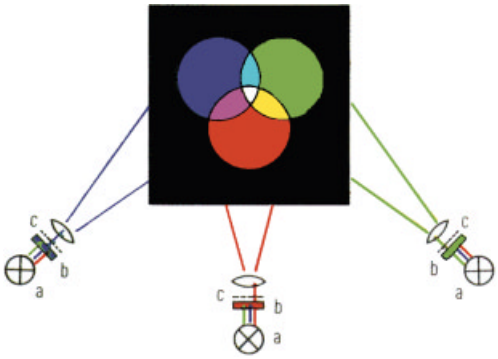

Figure 39. Principle of additive color mixing
Beams of the three primary colors (blue, green, and red) are projected onto a white wall (shown here as black) in a dark room. The addition of two primary colors gives the mixed colors cyan (blue and green), magenta (blue and red), and yellow (green and red). All three primaries together give white.
a) Light source; b) Colored filter; c) Black-and-white transparency

5.1.1. Screen Color Plates and Films

In order to decrease the amount of equipment required, the LUMIÈRE brothers marketed the first screen color plates in 1907. These "Autochrome" plates contained color filters in the form of individual starch grains, dyed blue, green, and red, in a layer just one grain thick on top of the panchromatically sensitized emulsion. The emulsion was exposed through the screen of dyed grains. After reversal processing, the color transparency could be viewed only by transmitted light. A single image, with gray-tinted colors, resulted.

The color screen had to be so fine that the color elements could not be separated by the eye. Only in this way was it possible to perceive yellow as a consequence of red and green dye particles lying next to each other.

In 1977, Polaroid put a narrow-gauge instant film with a linear color screen (Polavision) on the market, following it in 1982 with an instant transparency material (Polachrome); see Sections 5.5.4.1 and 5.5.4.2.

5.1.2. Lenticular-Screen Film (Agfa, 1932)

The lenticular-screen film bore a very fine screen of cylindrical lenticules on its back. The film was exposed through this screen with the aid of a special filter, which had the colors blue, green, and red next to one another. After reversal processing, the film had to be projected with the same filter.

5.2. Subtractive Color Processes

5.2.1. Principle

The subtractive primary colors cyan, magenta, and yellow each transmit two-thirds of the spectral colors and absorb only one-third. Cyan dye, for example, absorbs red light and passes blue and green without hindrance (B in Fig. 40). The same holds for reflected light. Accordingly, cyan is sometimes called "minus-red"; magenta, which absorbs only green, as minus-green (Fig. 40 C); and yellow, accordingly, as minus-blue (Fig. 40 D).

With the three subtractive primaries, all perceptible colors can be reproduced in either transmitted or reflected light. If magenta is superimposed on cyan, the magenta absorbs green light not absorbed by the cyan, and a pure blue is obtained in transmitted or reflected light (E in Fig. 40). If magenta and yellow are combined, pure red remains (F in Fig. 40). Yellow and cyan give pure green (G in Fig. 40). All modern color processes are based on these three subtractive primaries.

In subtractive color processes, white is therefore represented by a transparent film (Fig. 40 A), so that all the light from the projection lamp is passed. In other words, no color is subtracted from the white light. This is one of the main advantages of subtractive processes. In contrast, white in the additive process is obtained by the addition of all three primaries, but each of these contain only a third of the white light intensity from the projection lamp.

Subtractive color processes require three colored images: yellow, magenta, and cyan. They are not, however, prepared with the help of filters and an emulsion sensitive to all colors, as in the additive process. Instead, three emulsions are

used, each sensitive to only one of the three primary colors. In most subtractive processes, the sensitive emulsions are coated one over another on a support. In this way, the three layers can be exposed simultaneously in a single camera.

There are three major groups of subtractive color processes, all of which presuppose the development of a silver image:

1) *Chromogenic (dye-coupling) processes*, in which the dyes are developed. These are further subdivided into *negative–positive* and *reversal processes*.
2) *Silver dye bleach (chromolytic) processes*, in which dyes present at the outset are bleached in such a way as to form an image.
3) *Dye-transfer processes* (Technicolor, Dye Transfer, instant photography systems).

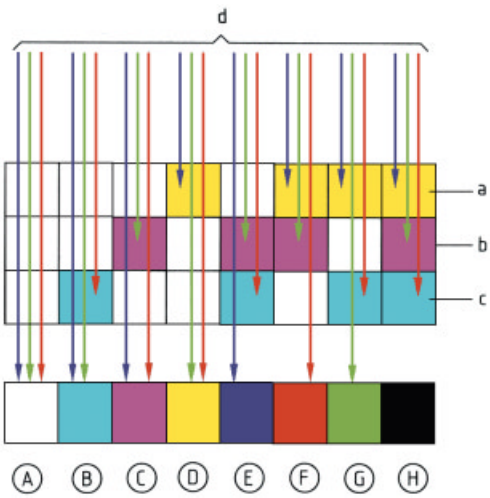

Figure 40. Principle of the subtractive color process
a) Yellow filter; b) Magenta filter; c) Cyan filter; d) White light
For explanation of A, B, C, D, E, F, G, H see text.

5.2.2. Processes with Chromogenic Development

The greatest advance in color photography was the discovery of chromogenic development. Chromogenic methods have become the most important ones in color photography. The color developer that is oxidized in developing the silver image usually reacts with colorless reactive substances (dye components, dye couplers) to form a variety of dyes. The principle of this method was described in 1911 by R. Fischer and H. Sigrist [305]. It can be schematically represented in the following way:

Exposed silver halide + color developer → silver metal + oxidized developer

Oxidized developer + dye coupler → dye

Two types of materials employ this process. *Reversal films* of the Kodachrome type have no couplers in the emulsion layers (see Section 5.2.2.5). Instead, the dye couplers are dissolved in various color developer baths. When these materials are processed, each dye coupler must diffuse, along with the color developer, into the emulsion layer for which it is intended. The development process is very complicated and hence can be carried out only with special machinery in special laboratories. It has the advantage of high sharpness because the emulsion layers in the material are very thin.

In materials of the *Agfacolor type* (Kodacolor, Fujicolor, etc.), the dye couplers are embedded in the three emulsion layers, from which they cannot escape by diffusion (see Section 5.2.2.2). Because these materials are developed in a single developer solution, the process is much simpler and can even be performed by amateurs.

5.2.2.1. Color Developers

Until recently, the only substances that could satisfy the many requirements that a color developing agent must meet were *N,N*-dialkylated *p*-phenylenediamine derivatives (**48**) [306]. Table 2 lists the most important developing agents now in use.

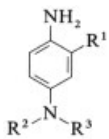

48

Of the agents in Table 2, only two are in common use: CD-3(**48d**) for color negative papers, reversal color films, and reversal color papers; CD-4 (**48e**) for color negative films. These substances, in alkaline solution, can reduce exposed silver halide. Initially, a semiquinonediimine is formed. After further oxidation, a quinonedi-

Table 2. The most important color developers

Compound	R^1	R^2	R^3	Trade name
48a	H	C_2H_5	C_2H_5	TSS
48b	H	C_2H_5	C_2H_4OH	T 32
48c	CH_3	C_2H_5	C_2H_5	CD-2
48d	CH_3	C_2H_5	$C_2H_4-NH-SO_2CH_3$	CD-3
48e	CH_3	C_2H_5	C_2H_4OH	CD-4
48f	CH_3	C_2H_5	$C_2H_4OCH_3$	CD-6
48g	H	C_4H_9	$C_4H_8-SO_3H$	T 60

imine cation is produced from this intermediate (**49** is the benzoid resonance structure with the electron sextet on the nitrogen). This quinonediimine reacts with the carbon atom of the active methylene group of the dye coupler to form a leuco dye (**50**).

The leuco dye is dehydrogenated to the dye (**51**) by a second developer molecule (also oxidized by silver halide) by atmospheric oxygen, or iron(III) compounds. To form one molecule of dye requires a total of four oxidation equivalents (electrons) (see nxt page, top).

If the coupler molecule has an easily removable leaving group (e.g., halogen, mercapto, sulfo, azo) at the point of attachment instead of a hydrogen atom, this group can be split off together with the hydrogen on the nitrogen. In this case, only two oxidation equivalents are required ("2-equivalent couplers").

Color Intensification. Processes have been devised in which the maximum quantity of dye formed depends not on the stoichiometric quantity of image silver formed, but only on the quantity of dye coupler that is present. Only a catalytic amount of image silver is needed for dye formation.

If a silver-poor color material, containing the normal quantity of dye coupler but only part of the normal amount of silver halide, is developed normally, the color in the resulting image is of low density. If the material after development, still saturated with developer, is placed in an oxidizing bath, for example with hydrogen peroxide [307] or hexaminecobalt(III) chloride [308], then the color developer will be oxidized by the oxidant at the locations where silver metal has been formed. Silver metal has a catalytic action. As a result, additional image dye is formed. The cobalt(III) complex can even be added to the developer itself, since this complex does not attack the developing agent in the absence of metallic catalysts, and color intensification begins after normal development of the exposed silver halide. The color intensification process has not gained much importance, but in the future it may become important depending on the silver economy.

5.2.2.2. Dye-Forming Couplers

p-Phenylenediamines can form various oxidation products in the photographic development of a latent image; for example in the following reaction:

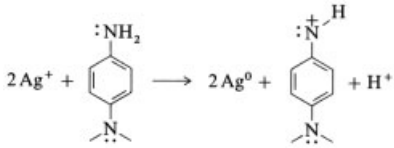

It gives up two electrons as well as one proton. The products are two silver atoms, one developer oxidation product (DOP), and one proton. If this DOP can now be utilized to produce a variety of dyes, it can form the basis of a color process. R. Fischer [305], [309] stated that oxidized *p*-phenylenediamines can be coupled with special organic compounds to form dyes. These compounds have long been referred to as dye-forming couplers.

$$\text{DOP} + \text{Dye-forming coupler} \xrightarrow[-H_2O]{OH^-} \text{Dye}$$

The requirement for obtaining a multicolored image is that the same DOP, combined with three distinct dye-forming couplers, gives yellow, magenta, and cyan coupling products:

Dye-forming coupler 1 + DOP ⟶ Yellow dye
Dye-forming coupler 2 + DOP ⟶ Magenta dye

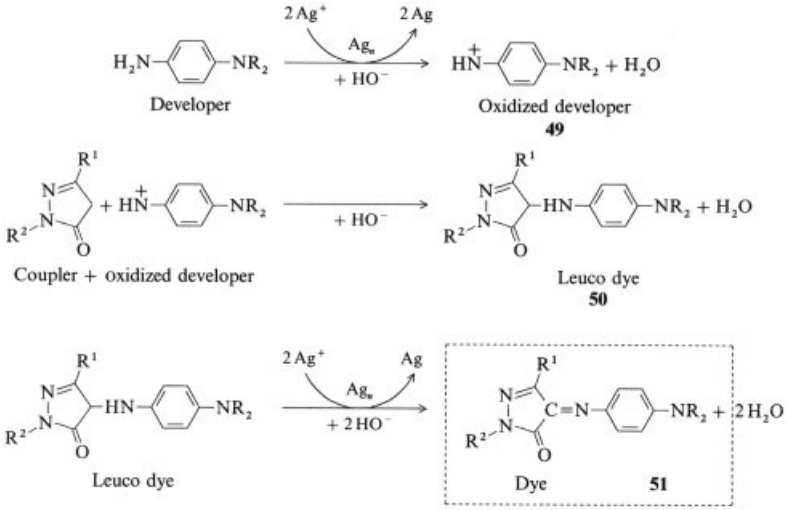

Dye-forming coupler 3 + DOP ⟶ Cyan dye

Dye-forming couplers themselves are usually colorless, whereas the dyes derived from them have the highest possible extinction in the desired portion of the spectrum.

A desirable feature in dyes is a narrow absorption curve (small half-width) with only small side absorptions in adjacent parts of the spectrum.

Dyes and dye-forming couplers for photographic materials must have a number of other qualities in addition to an absorption in the desired spectral range, e.g.:

- Color purity
- High coupling activity
- High extinction
- High sensitivity
- Low fog
- Good solubility
- Good emulsion stability
- High dark stability
- High stability under tropical climatic conditions
- Good light stability
- Low granularity
- Availability (low price)

All these special requirements can be controlled within certain limits through appropriate substitution in the dye-forming coupler molecule. In many cases, however, a compromise is necessary when the requirements are in conflict (e.g., high coupling activity versus low fog).

Yellow Couplers. Dye-forming couplers that form yellow azomethine dyes are called "yellow couplers." They have an absorption maximum (λ_{max}) in the blue region, preferably between 430 and 460 nm.

The methylene group at the coupling position is activated by two adjacent carbonyl groups; the simplest yellow couplers are β-diketones or acylacetamides.

By far the most widely used yellow couplers are benzoylacetanilides and pivaloylacetanilides (2,2-dimethylpropanoyl acetanilides). The following list gives a rough comparison of these two classes of couplers.

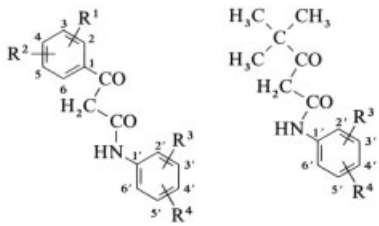

Coupling activity	higher	lower
Sensitivity	higher	lower
Fog	higher	lower
Color reproduction	longer wavelengths (redder)	shorter wavelengths (yellower)
Light stability	lower	higher
Dark fading stability	lower	higher
Stability under tropical conditions	lower	higher

Because of their better dye stability [310], pivaloylacetanilide yellow couplers are used in all color papers. The other properties can be modified within certain limits by appropriate substituents (R^1, R^2, R^3, R^4) in the benzoyl and aniline parts. In order to anchor the coupler in one particular layer of a photographic material, at least one of the substituents (R^1, R^2, R^3, R^4) must be a "ballast" group: i.e., an organic substituent that is large enough to prevent diffusion into an adjacent layer. The following substituents can be used:

R^1: 2–OCH_3, 2–$OC_{16}H_{33}$, 3–NH–CO–$C_{15}H_{31}$
R^2: 4–OCH_3, 4–$OC_{16}H_{33}$,
R^3: 2′–Cl, 2′–OCH_3, 2′–$OC_{16}H_{33}$

R^4: 5′–$COOC_{14}H_{29}$, 5′–NH–CO–$(CH_2)_3$–O–

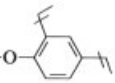

5′–OCH_3, 5′–SO_2–NH–CH_3, 5′–SO_2–N(CH_3)(CH_3)

5′–NH–SO_2–$C_{16}H_{33}$, 4′–SO_2–N(CH_3)(CH_3)

Extensive research has shown that the Hammett equation adequately describes the effects of various substituents and their position on pK_a [311], λ_{max} [312], coupling kinetics [313], and hydrolytic stability of dyes [314], [315].

Magenta Couplers. By analogy with yellow couplers, compounds that form magenta dyes on oxidative coupling are called magenta couplers. In contrast to open- chain methylene compounds, the oxidative coupling of cyclic compounds with an active methylene group leads to dyes having absorption maxima in the range of 530 – 580 nm. The major types of compounds can be summarized as follows (see page 78, top).

The absorption maximum shows a bathochromic shift of ca. 100 nm in comparison to that of yellow couplers and is preferably between 540 and 560 nm in the green part of the spectrum. Of the many known coupler structures, only a few have actually found use in photographic materials. Examples include the pyrazolones (**57**), the pyrazolobenzimidazoles (**58**) and (in 1983 – 1984) the two isomeric pyrazolotriazole structures (**59**, **60**).

Pyrazolone Magenta Couplers. For many years, all the magenta couplers used in photographic materials had a pyrazolone structure with a phenyl nucleus in the 1 position. The substituent in the 3 position of the pyrazolone ring, a phenylamino group or an acylamino residue, largely governs the properties as well as the potential applications. A comparison of the effect of these two substituents is given in the following list:

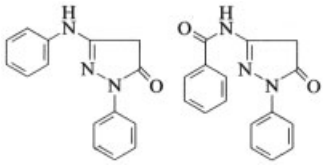

Coupling activity	higher	lower
Sensitivity	higher	lower
Light stability	higher	lower
Dark stability	higher	lower
Stability under tropical conditions	higher	lower
Granularity	higher	lower
Absorption	shorter	longer

Accordingly, anilinopyrazolones were preferred for color papers because of better dye stability, while acylaminopyrazolones were used in film materials (color negative and color reversal) because of their lower granularity. In both classes, the 1-phenyl nucleus has three chlorine substituents in the 2,4,6 positions. The reasons for this is that the residual coupler [310] in the emulsion showed less yellowing and the dye stability is markedly improved.

The light stability of anilinopyrazolones can be further improved by the use of special stabilizing residues in the ballast part of the molecule, for example:

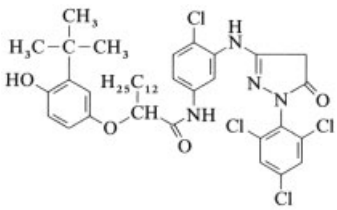

Pyrazolobenzimidazoles. The cyclization of 1-(2-aminophenyl) pyrazolones to pyrazolobenzimidazoles by elimination of water is easily carried out:

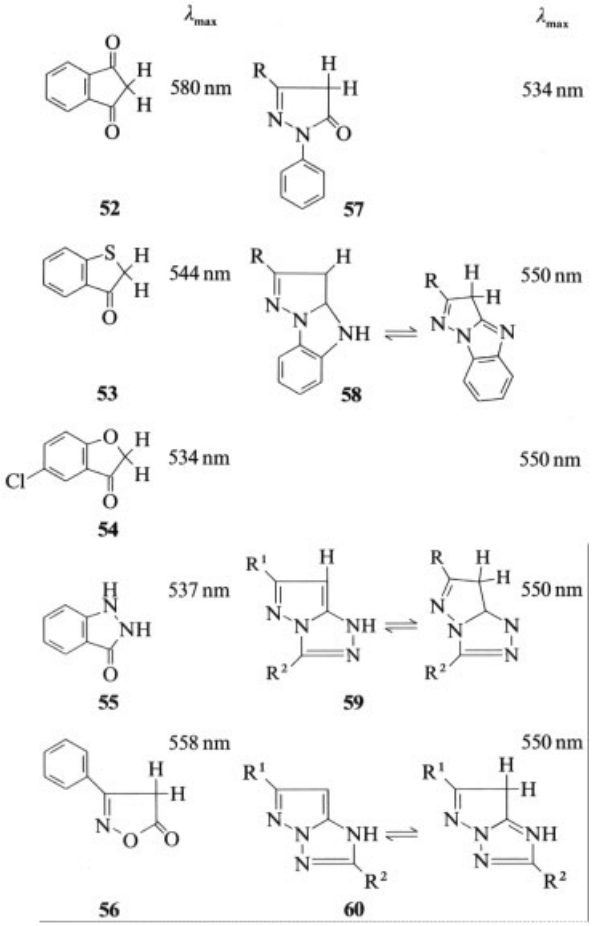

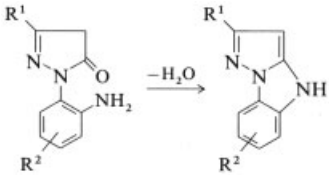

These compounds couple with developer oxidation products (DOP) of the phenylenediamine type to give magenta dyes of great brilliance. In contrast to the conventional pyrazolone magenta dyes, these do not have a secondary absorption maximum in the blue region. The usual substituents R^1 are alkyl, acylamino, arylamino, and alkoxy groups.

Pyrazolotriazole Magenta Couplers. Closely related to the pyrazolobenzimidazoles are magenta couplers of the pyrazolotriazole type. These again show high color brilliance, having no secondary absorption in the blue part of the spectrum. Of the two structural isomers available, the pyrazolo[3,.2,- c]-5-triazoles

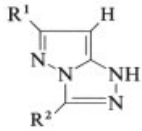

have been known since the mid 1970s [316], [317], but they were not used in photographic materials until 1982. Dyes from 1 *H*-pyrazolo[1,5-b] [1,2,4]triazole magenta couplers

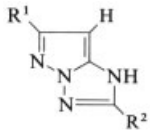

were described as magenta couplers in 1983 – 1984; they are used in modern color papers because of their good light stability [318].

Cyan Couplers. Compounds whose azomethine dyes absorb in the red region of the spectrum are called cyan couplers. In all, five distinct types of structure in the class of phenols and naphthols are common:

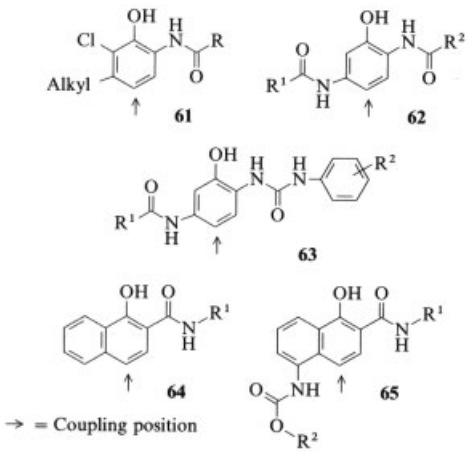

Dyes formed from CD-3 (see Table 2) and couplers **61** and **62** absorb at ca. 660 nm and are used in color papers (**61**) and color reversal film (**62**). The dyes from **61** offer especially good dark and light stability.

When color negative films are processed with developers containing CD-4, couplers **63–65** give dyes that have an absorption maximum at 690 – 700 nm. Important features are the good sensitivity of naphtholic coupler **65** and the good dark fading stability and good re-oxidizability of the leuco dyes from couplers **63** and **65**.

5.2.2.3. Special Dye Couplers

If a coupler has a suitable substituent X at the coupling position, the reaction with DOP does not give the leuco dye but goes directly to the azomethine compound (see Section 5.2.2.1).

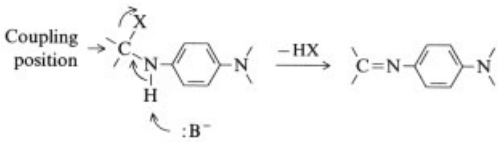

As only two equivalents of Ag^+ are needed per mole of dye formed, couplers with these leaving groups X≠H are called 2-equivalent couplers.

Normal 2-Equivalent Couplers. Substituents X with no special sensitomeric action include

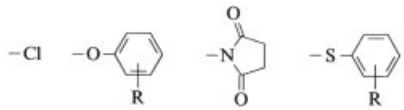

Compounds bearing these substituents have the advantage of a low silver consumption.

Development Inhibitor Releasing Couplers. Another group of special 2-equivalent couplers are development inhibitor releasing (DIR) couplers. Certain leaving groups X called inhibitors, such as

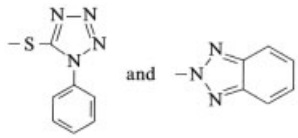

diffuse to the silver halide crystal after being split off. Because of its strong affinity to silver halide and silver metal, the inhibitor is adsorbed on the surface of the grain. In this way, the further reduction of silver halide to metallic silver is inhibited.

Inhibitors can develop their action in the layer where they are liberated or can diffuse into neighboring layers. Intralayer and interlayer effects are therefore distinguished.

Intralayer effects
Control of granularity
Control of sharpness through edge effects

Interlayer effects
Improvement of color reproduction through interimage effects

DIR couplers are assigned to one of two groups depending on whether the inhibitor is released immediately on coupling or by an additional chemical reaction (e.g., cyclization or elimination). An example of a DIR coupler of the second type is

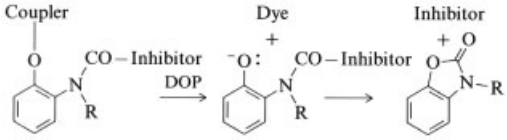

Development-Accelerator-Releasing Couplers [319]. In contrast to DIR couplers, development-accelerator-releasing (DAR) couplers split off a development accelerator or a fogging agent. An example of a DAR coupler is the following:

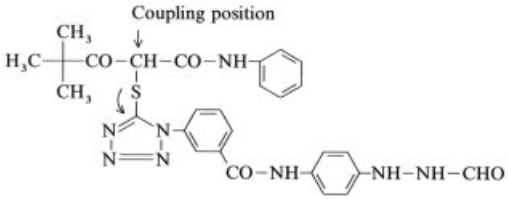

The formylphenylhydrazine group, once liberated, is oxidized to the acylphenyldiazene; the latter then forms a phenyldiazene that can serve as a nucleating agent for unexposed silver halide [320].

Non-Color-Forming Couplers. Couplers that bear an alkyl group (e.g., methyl or ethyl) at the coupling position can still form a coupling product in the usual way with DOP. However, subsequent base-catalyzed elimination of the alkyl group as an alkyl anion is not possible, and no dye formation takes place. For this reason, compounds of this type are called non-color-forming couplers (white couplers). They are included in interlayers as DOP traps in order to prevent its diffusion into the adjacent layer.

Masking Couplers [321], [322], [323]. Couplers with a dye attached via the leaving group (**66**) or that have an intrinsic color by virtue of the leaving group (**67**) are called masking couplers (see next page, top).

They are employed to correct undesired side absorptions of the image dyes. The action of the yellow masking coupler **67** is given as an example. In development, it recouples with DOP; the yellow density of **67** decreases and a magenta dye is formed upon coupling with DOP (Fig. 41). The sum of curves b and c gives the relatively constant density d, which practically corresponds to a magenta dye with no blue absorption. The uniform yellow fog can be filtered out with a blue filter in printing.

A similar discussion applies to the red masking coupler **66**, which couples to cyan and is used to compensate the side absorption of cyan dyes in the green region of the spectrum. The masking dye, which is present as a leaving group, has a high water solubility by virtue of its two sulfo groups and is washed out in processing.

Bleach Accelerator Releasing Couplers. Couplers having a bleach accelerator, for example –S–CH_2–CH_2–COOH, as leaving group are called bleach-accelerator-releasing (BAR) couplers [324].

The group liberated upon coupling effects the desilvering of the photographic material in rapid processing when the bleaching time is short or when exhausted bleaching baths are employed.

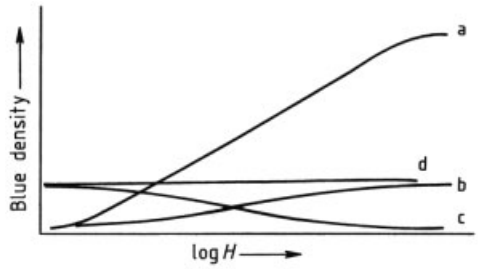

Figure 41. Action of a yellow masking coupler generating a magenta dye
a) Gradation of a magenta dye; b) Yellow gradation due to secondary density of magenta dye; c) Decrease in yellow density due to recoupling of **67**; d) Compensated yellow density (resultant of b and c)

5.2.2.4. Anchoring of Dye Couplers in Layers

The dyes must be generated in three superimposed emulsion layers without interfering with color reproduction. The layers must therefore be extremely thin; however, the coupler must not be allowed to diffuse from its own layer into an adjacent one (interlayer wandering).

This diffusion resistance, which is necessary for good color separation, can be attained in a variety of ways.

Aliphatic "Tail". If a long-chain hydrocarbon residue containing hydrophilic sulfonic acid

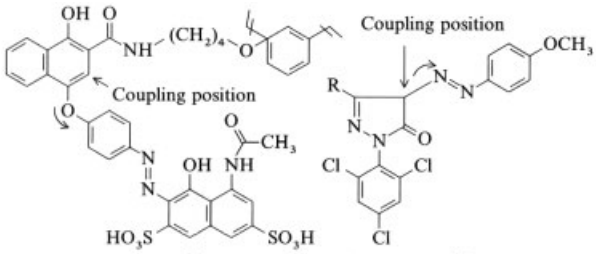

or carboxyl groups is attached to the dye-coupler molecule, it will prevent the coupler from escaping its layer. The hydrophilic groups facilitate the introduction of the coupler into the emulsion layer as a salt. The low diffusivity results not only from the size of the molecule and the "tail" (large hydrophobic group); these couplers also interact with the gelatin binder, forming micelles, which hinder diffusion. Examples of micelle-forming couplers are:

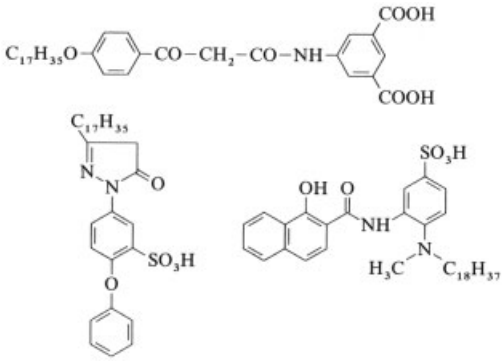

Solvent Dispersion Couplers. In 1942, Kodak introduced a film employing a different method for anchoring the couplers against diffusion. The hydrophobic couplers are dissolved in a highboiling organic solvent (e.g., tricresyl phosphate or dibutyl phthalate), dispersed in gelatin with suitable aids (e.g., Alkanol XC), and added to the silver halide emulsions. The resulting emulsion layers contain the couplers in the form of very finely divided oil droplets, into which the oxidized color developer penetrates and where the dye-forming reaction takes place. The chemical structure of the groups hindering diffusion is different from that of micelle-forming couplers, in keeping with the technical requirements: The straight- chain hydrocarbon group is replaced with a highly branched and bulky group such as

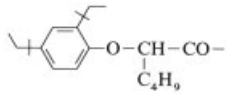

which is joined to the coupler molecule by a group. The high-boiling solvents used as the oil phase further protect the coupler and the dye from the action of moisture, heat, and light.

Coupler-Loaded Latex. The direct mechanical dispersion of hydrophobic couplers in the emulsion layer requires a relatively large amount of energy and often yields dispersions with a broad disperse-phase size distribution. These disadvantages can be avoided if a polymer dispersion is used and the hydrophobic coupler is deposited on the surface of the latex particles from a solvent miscible with water. In this way it is possible to obtain coupler dispersions with a very narrow particle-size distribution and good photographic qualities. On account of their affinity for the hydrophobic latex particles, the coupler molecules deposited in this way are fixed to the surface of the latex particles and are virtually prevented from diffusion.

Polymeric Couplers. Low-molecular-mass coupler molecules without a ballast group but having a polymerizable group can be anchored in a different way. They can be polymerized with suitable comonomers and small quantities of cross-linking monomers (e.g., divinylbenzene, ethylene glycol diacrylate, trimethylol propene triacrylate) into "latex couplers." The latex particles, which embody the coupling action, are much smaller than those of dispersed couplers and have no tendency to increase in size.

Some appropriate coupler monomers are as follows:

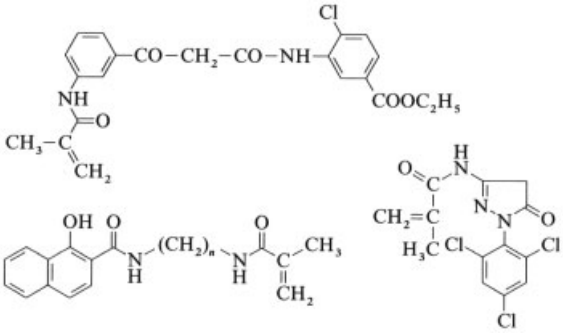

Examples of compounds used as comonomers are acrylates, acrylamides, and styrenes. The coupling activity, the granularity, the stability of the emulsion layer, and the stability of the dyes can be controlled within certain limits by the choice of comonomer.

Because the coupler monomer and the comonomer (ballast portion) are in a favorable ratio and no oil-phase former is needed, latex couplers make it possible to obtain much thinner emulsion layers than dispersed couplers.

5.2.2.5. Couplers Added to Developer

As an alternative to incorporating the couplers in the materials, Kodak devised another process for making positive color transparencies. The dyes are produced from conventional developers and the known couplers. In this process, however, the couplers are not fixed in the several emulsion layers but are added to the developing solutions. The dye-forming couplers are readily soluble in the alkaline baths, but upon chromogenic development they produce dyes that are sparingly soluble; hence, they cannot be washed out in subsequent processing. Suitable choices are low-molecular-mass compounds with enolizable carbonyl groups or phenolic OH groups, such as:

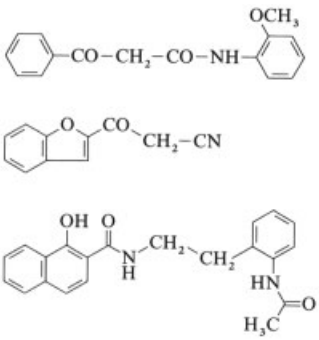

After exposure of the color transparency film, a black-and-white negative is developed. The dyes are then produced one after another, not simultaneously, and the positive color image is obtained by a combination of reexposures behind narrow-band filters:

1) The cyan image is formed in the red-sensitive emulsion by exposure through a red filter and development in the presence of the cyan coupler.
2) The yellow image is formed in the blue-sensitive layer by exposure through a blue filter and development in the presence of the yellow coupler.
3) The magenta image is formed by chemical fogging; that is, the silver nuclei are generated not by exposure but by treatment with sodium borohydride. Development in the presence of a magenta coupler follows.

This complicated process requires exact control of the conditions and can be performed only with special equipment. In spite of this difficulty, the process, which gives particularly sharp images, has remained important.

5.2.2.6. Chromogenic Negative – Positive Processes

In color negative film, a silver image and a dye image are obtained after exposure and development. Because each dye has the color complementary to the sensitivity of the emulsion where it is produced, bleaching and fixing of the silver image yields a color image that is complementary with respect to lightness and color (a negative). What was blue in the original is yellow in the negative; red appears as cyan; and green appears as magenta. What was black in the original

shows up as white or transparent in the negative, while what was white appears black.

Color Recording Material. The requirements imposed on color recording materials today are very stringent; they cover light sensitivity, reproduction of details and colors, and other qualities. As a result, these materials have a much more complex structure than color print materials. As an example of a modern, high-speed recording material, the structure of Agfacolor XRG-100 negative film will be discussed (see Fig. 42) [325]. The color-forming layers in this film are applied as double coatings. Double-coating involves combining a relatively coarse-grained, high-speed emulsion having a low ratio of couplers to silver with a relatively low-speed, fine-grained emulsion having a higher ratio [326]. The resulting film offers high light sensitivity with very fine color grain. Since each emulsion acts as a gray filter, diminishing the exposure of the emulsion beneath it, the high-speed coating is always applied on top of the lower-speed one. The uppermost emulsions are the two blue-sensitive yellow layers. Immediately below them is a yellow filter layer consisting of colloidal silver. This layer prevents the green-sensitive and red-sensitive emulsions from being exposed in the blue intrinsic sensitivity region of the silver halide. The green-sensitive magenta emulsions are separated from the red-sensitive cyan emulsions by a filter layer containing a red dye. This enhances the sharpness by reducing the back-scattering of green light and also contributes to improved color separation. Beneath the red-sensitive cyan emulsions is the black antihalation layer. This protective layer, consisting of colloidal silver, prevents the transmitted light from being reflected at the emulsion support. The colloidal silver in the yellow filter layer and the antihalation layer are oxidized in the bleaching bath and dissolved out in the fixing bath. The green-sensitive emulsions contain yellow masking couplers, whereas the red-sensitive emulsions contain red masking couplers. These masking couplers (see Section 5.2.2.3) are colored substances, which change their color in development in relation to the density of the negative and so compensate for undesired secondary absorptions of the image dyes. Masking makes the colors purer and more brilliant. DIR couplers (see Section 5.2.2.3) are also present in all the light-sensitive layers. This decreases the granularity (see Section 6.2.2) and improves the sharpness (see Section 6.2.3) and the color reproduction by virtue of edge and interimage effects (see Section 6.1.6), respectively.

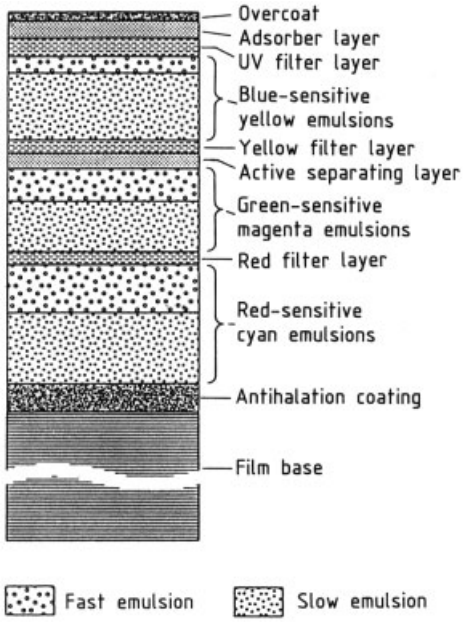

Figure 42. Schematic coating structure of a high-speed color negative film

The emulsion support of 135 size film consists of cellulose acetate 125 µm thick. The emulsion thickness is 23 µm. The film is processed by the Agfacolor AP-70, AP-71, or AP-72 process or the Kodak C-41, C-41 B, C-41 BNP, C-41 RA, or C-41 RANP process.

Positive Print. To obtain a color positive, the color negative must be printed onto a color positive material (emulsion side to emulsion side). After processing, the result is a positive corresponding to the original with respect to color and lightness gradation. A flow sheet of the entire chromogenic negative – positive process is shown in Figure 43. Since the positive material has a very high gradation and its blue-sensitive layer is markedly faster than its green-sensitive and red-sensitive emulsions, the yellow filter layer can be dispensed with. The arrangement of color emulsions can also be different. In color paper, in contrast to the recording material, the topmost light-sensitive layer forms the cyan par-

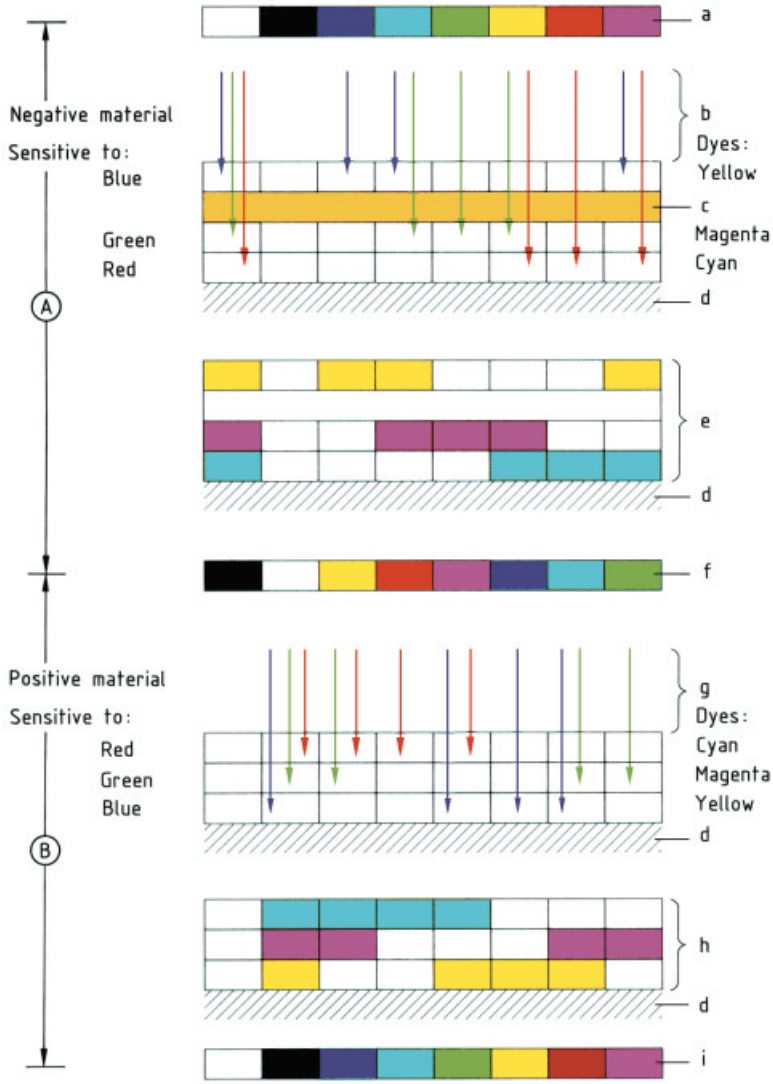

Figure 43. Chromogenic negative – positive process
A) Production of the negative; B) Production of the positive
a) Original; b) Transmitted or reflected light; c) Yellow filter layer; d) Emulsion support; e) Recording material after color development, silver bleaching, and fixation; f) Color negative viewed in transmitted light; g) Printing light transmitted by the negative; h) Positive material after color development and bleach fixing; i) Color positive viewed in reflected or transmitted light

tial image and the layer forming the yellow partial image is the undermost one.

The printing process allows the use of color correcting filters. If one color is reproduced too strongly in the positive relative to the other two in the positive, a correcting filter can be inserted in order to restore the color balance.

Color positive papers have "selective" sensitization, unlike the "overlapping" sensitization of negative films; that is, the spectral sensitivity curves of the three emulsions in papers have little or no overlap. The sensitization maxima must be in close agreement with the extinction maxima of the color negative dyes corresponding to the respective positive emulsions.

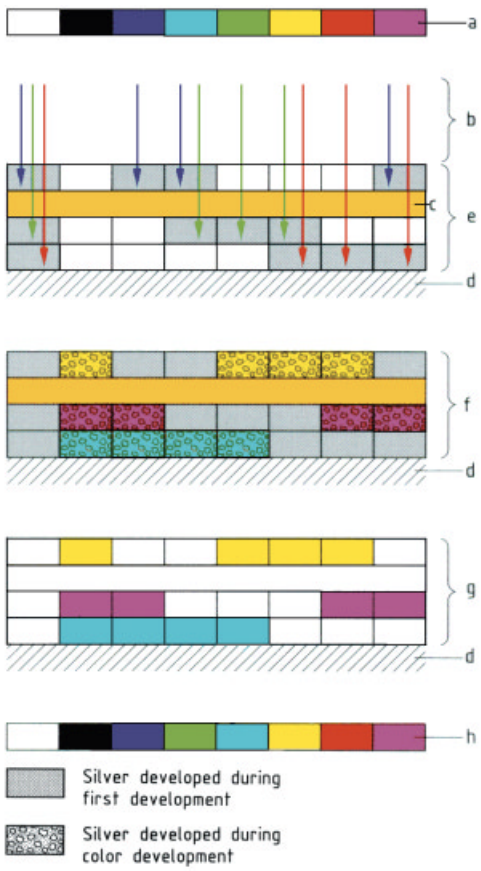

Figure 44. Reversal color process
a) Original; b) Transmitted or reflected light; c) Yellow filter layer; d) Emulsion support; e) Reversal color film after first development; f) Reversal color film after chromogenic development; g) Reversal color film after bleaching and fixation; h) Processed reversal color film viewed in transmitted light

5.2.2.7. Reversal Color Process

The reversal color process leads, in a single step, to a positive in which the lightness and color of the developed photographic material correspond to those of the original object. If a transparent support is used, a transparency is obtained: a positive intended for viewing in transmitted light (projector or viewer). To obtain a positive in one operation, the photographic material passes through two development steps in the reversal color process. In the first development, a black-and-white negative is obtained. The undeveloped silver halide is subsequently either exposed or chemically fogged, then put through a second, chromogenic development, which "reverses" it to give a positive silver and dye image. The silver formed in both the first and second developments is oxidized in the same bleaching bath and converted to silver bromide, which is complexed and dissolved in a downstream fixing bath and then washed out of the emulsion. Figure 44 is a schematic diagram of the reversal color process.

Reversal Color Recording Materials. Two types of reversal color films are on the market. They differ in layer structure and in processing. Films of the *Kodachrome type* contain no dye couplers; the couplers are introduced only during development, and three color developments are required in addition to the first development (see Section 5.2.2.5). In film materials of the *Agfachrome type* (Ektachrome, Fujichrome), the couplers are incorporated into the emulsions and cannot diffuse away from their location. The structure of such a film is explained for the example of Agfachrome CT-100 (see Fig. 45).

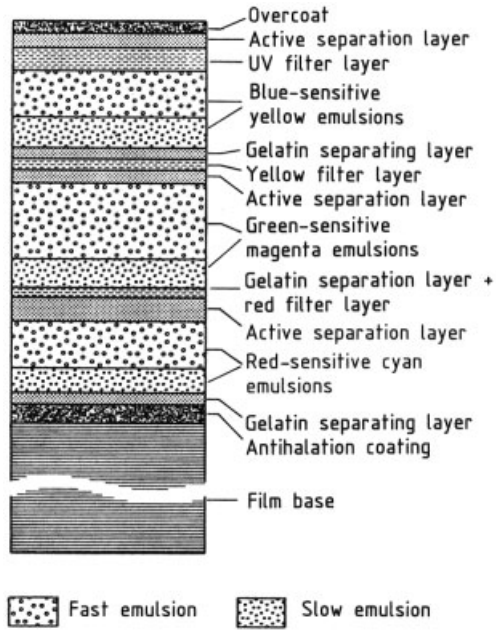

Figure 45. Structure of a color reversal film with incorporated couplers

In *Agfachrome CT-100*, the color-forming layers are double-coated; that is, the light-sensitive emulsions and the couplers for the three

subtractive primary colors yellow, magenta, and cyan are each applied in two coatings. For the same reasons as in the case of color negative film, the upper coating of each color element has a higher speed than the lower one. In contrast to color negative materials, the higher-speed emulsions in reversal color materials have a higher relative content of couplers than the lower-speed emulsions. This arrangement offers low granularity despite the high sensitivity. In addition to the color-forming emulsion layers, Agfachrome CT-100 further includes a number of separating and filter layers. The UV filter layer absorbs the ultraviolet portion of the light incident on the film during exposure, thus preventing color distortion due to UV radiation. Yellow and red filter layers and the antihalation layer have the same functions as in color negative films (see Section 5.2.2.6). The separating layers contain substances which prevent developer oxidation product from diffusing into adjacent emulsion layers and causing color distortion. Sometimes they also contain substances that control interimage effects. The colored masking couplers included in color negative films to improve color reproduction (see Section 5.2.2.3) cannot be employed in reversal color films, because the intrinsic colors of such couplers distort the color reproduction.

Development inhibitor releasing (DIR) couplers, which are very useful in color negative films, cannot exert their full effect in the reversal process. The reason lies in the very intensive chromogenic second development, in which complete development of the silver halide is sought and even inhibited silver halide cannot survive. Even so, under certain conditions DIR couplers can be advantageous in improving image quality in reversal color films. For example, Gevachrome II films, which are intended chiefly for television purposes, contain such couplers.

Films of the Agfachrome type today are virtually all compatible with the E-6 process; that is, they can be processed by the Kodak E-6 process or in similar processes marketed by various manufacturers (see Section 5.2.2.4).

Reversal Color Printing Materials. Film prints can be made from positive transparencies on reversal color copy films. These films have the same basic structure as recording films and are processed in the same way, but they have a lower contrast than recording films. Paper prints from positive transparencies can be made by way of an internegative. This process requires two operations, the preparation of a negative from the transparency on negative film and the printing of this negative on color negative paper. A better and faster method is to print the positive transparency on reversal color paper. In this way, a positive print is obtained in one operation (see Section 8.3.1.2).

5.3. Processing of Color Photographic Materials

The most important processing steps for color photographic materials are as follows.

Development. In this step, a silver image and a dye image are formed in each of the emulsion layers (see Section 5.2.2.1). Along with the silver and dye images produced in exposed regions, fog is also formed in unexposed regions under certain conditions. There can be several causes of fogging [327]. *Base fog* is the coloration of the support and emulsion layers due to sensitizers etc. *Development fog* is generated in the developer. *Postdevelopment fog* forms by reduction of unexposed silver halide grains in subsequent washing, if there is such an operation. *Bleaching fog* can arise in the emulsion layers as a result of residual color developing agent, which is oxidized in the bleaching bath or bleach – fixing bath and then couples with unconsumed coupler to form a dye. Bleaching fog can be largely suppressed by adequate washing or stopping before the oxidizing bath or by the maintenance of the lowest possible pH in the bath itself. "White couplers" (non- color-forming coupler) are effective in suppressing bleaching fog (see bleach – fixing below).

Silver Bleaching. Because the blackening inherent in the silver image distorts the color image, this silver must be bleached (oxidized) without disturbing the image dyes. This step involves oxidizing agents, such as the iron (III) – EDTA complex, hexacyanoferrate(III), or potassium peroxodisulfate ("persulfate"). Depending on the composition of the bleaching bath, a silver salt (halide, hexacyanoferrate, or sulfate) or a complex salt of silver is formed. In addition to

the developed silver, the metallic silver in the yellow filter layer and antihalation layer are also oxidized in the bleaching bath.

Fixation. Both the undeveloped silver halide and the silver salt formed in silver bleaching must be removed from the color material so that these light-sensitive silver salts do not affect the process at a later stage. Sodium or ammonium thiosulfate is used almost without exception for fixing. Ammonium thiocyanate is sometimes added to speed up fixation.

Bleach – Fixing. To save a step, it is possible to use a bath that carries out the functions of bleaching and fixing. Bleach – fixing baths contain the iron(III) complex of an aminopolycarboxylic acid (most commonly EDTA) as oxidizing agent for silver metal, along with a thiosulfate as complexing agent for oxidized silver. In addition, these solutions contain free complexing agents to stabilize the iron complex and sulfite to stabilize the thiosulfate. Bleach accelerators, such as mercapto compounds of triazole and thiadiazole, may also be present. White couplers, i.e., suitably substituted pyrazolones (see Section 5.2.2.3) are also incorporated into these solutions. White couplers react with developer oxidation products (carried into or generated in the bleach – fixing bath) to form colorless products, before these oxidation products can combine with the couplers in the emulsions to form nonimage dyes (bleaching fog).

Stabilization. The concept of a developing machine without a water supply or a water drain led to the use of stabilizing baths (not to be confused with those described in Section 4.3.3.4). These baths make the use of wash water superfluous. The iron complex originating from the bleach – fixing bath that remains in the emulsions must be prevented from altering the highlights and colors of the image after long storage. The principal ingredient of such a bath is a phosphonic acid, such as hydroxyethanediphosphonic acid (HEDP), which "masks" the iron ion; i.e., accepts it from the EDTA complex. The ferric complex of HEDP is not able to take part in redox reactions [328]. This step occurs in film and paper processing. A film processor and a paper processor with an exposure unit at the head end constitutes a "minilab" (see Section 5.6.3).

Because the color materials marketed by various manufacturers have similar physical and chemical properties, there is little difference between processing techniques. This principle of compatibility dominates the market. In the extreme case, it means that color films and color papers from all the major manufacturers can be processed alongside one another even if each bath comes from a different chemical manufacturer. Exceptions to this rule are few and apply only to special-purpose products.

The equipment used for processing influences the operating conditions: development can take place in a tray, a drum, or a roller transport or leader-belt machine. Machines with low throughput and long idle times at high temperature may require more than the usual amount of antioxidants (e.g., for the developer).

5.3.1. Replenishment and Rejuvenation

The baths used for the processing of color materials are chemically modified during processing. For example, halides, stabilizers, protective dyes, wetting agents, gelatin particles, and other soluble constituents of the emulsion layers build up in the *developer*. The *bleaching* and *bleach – fixing baths* become enriched in developer ingredients, and silver collects in the bleach – fixing and fixing *baths*. In principle, the same discussion applies to the replenishment and rejuvenation (regeneration) of color photographic baths as to those of B & W baths (see Section 4.3.3.8). The processing of color materials is almost completely automated, regardless of the size of the laboratory. The baths are continuously replenished and replaced. All the manufacturers of photochemicals strive to offer *replenishers* that can be added in the smallest possible amounts, so as to minimize the quantity of waste and thus the cost of shipping it.

A further reduction in processing waste is achieved by *rejuvenation* (regeneration). In larger operations especially, overflows from all the major baths are reprocessed and recycled (rejuvenated or regenerated). Color developers are rejuvenated on anion-exchange resins, which selectively adsorb the halide. The solution, now with little or no halide, is then topped up with

the rejuvenator, which replaces the substances consumed and makes up the volume of solution lost through entrainment. In this way a ready-to-use developer replenisher is obtained. When the ion-exchange resin is exhausted, it must be regenerated, usually with sodium chloride and/or sodium hydrogencarbonate solution. The entire rejuvenation process is carried out in mainly automated equipment [329].

Most *bleaching-bath rejuvenation processes* are simple. The principal bleaching agents used – iron complexes of aminopolycarboxylic acids – oxidize even on standing in the presence of air, so that the concentrated rejuvenator solution can be added without further treatment. The volume of developer solution or water brought into the bath with the materials is turned into bleaching solution by the rejuvenator. Bleaching baths with hexacyanoferrates are now seldom used, they require a peroxodisulfate for the reoxidation of the bleaching agent. The result is a buildup of sulfate, which can become a bleaching inhibitor after a long period of closed- cycle operation.

Rejuvenation of bleach – fixing baths involves one special difficulty: the iron(III) complex that has not been reduced during silver bleaching must be reduced before reduction of silver can occur. The steel-wool process originally used for silver recovery involved the deposition of silver on finely divided iron (cementation), but the reaction of the iron(III) compound with the metallic iron, which takes place first, led to an undesirable silver sludge product. The first satisfactory outcome was achieved with a modified process in which the iron(III) compound was reduced with sodium dithionite before the inlet to the steel-wool cartridge [330]. Since equipment providing a high electrolysis current has become available, however, the most widely used process is electrolytic silver deposition. Once the silver has been recovered, the iron complex, which has been completely reduced in the process, is oxidized with air. The rejuvenator then brings the solution to the desired replenisher concentration.

For the rejuvenation of fixing baths, see Section 4.3.3.8.

5.3.2. Processes and Formulas for Color Photographic Materials

The formulas set forth by the various manufacturers of photographic chemicals are essentially similar, because the world market honors the principle of compatibility; that is, materials sold by all manufacturers can be processed in a satisfactory way regardless of the supplier of the chemicals used. There are, however, marked differences between brands of chemicals as to handling, stability, and environmental impact.

To simplify the preparation of baths from prepackaged processing chemicals, what is supplied is usually just the replensiher chemicals. A ready-to-use tank charge is obtained by adding "starter" solutions to the replenisher solutions. The starter solutions act mainly to set the proper halide concentration and pH.

5.3.2.1. Processing of Color Negative Film

An example of a process for color negative film is the Kodak C-41 process. Variants are C-41 B and C-41 RA, intended chiefly for minilabs. Agfa markets a corresponding process as AP-70, with variants AP-71 and AP-72 for minilabs. The processing conditions of these processes are shown in Table 3 [331].

Table 3. Sequence of baths in the processing of color negative films

Bath	Processing time, s	Temperature, °C
Process C-41/AP 70		
Developer	195	37.8 ± 0.25
Bleaching	260	38 ± 3
Intermediate washing	65	38 ± 3
Fixing	260	38 ± 3
Final washing	195	38 ± 3
Final bath	65	24 – 41
Process C-41 B/AP 71		
Developer	195	37.8 ± 0.25
Bleaching	180	38 ± 3
Fixing	240	38 ± 3
Stabilizing	47	38 ± 3
Stabilizing	47	38 ± 3
Stabilizing	47	38 ± 3
Process C-41 RA/AP 72		
Developer	195	37.8 ± 0.25
Bleaching	45	38 ± 3
Fixing	90	38 ± 3
Stabilizing	20	38 ± 3
Stabilizing	20	38 ± 3
Stabilizing	20	38 ± 3

The processes for minilabs, C-41 B/AP-71 and C-41 RA/AP-72, are significant in that they have shorter bleaching times. In the first process, bleaching is cut to 3 min by mixing the iron complexes of EDTA and PDTA (1,3-propylenediaminetetraacetic acid) instead of using iron(III) EDTA alone. In the second case, the use of iron(III) PDTA alone brings the bleaching time down to 45 s. The use of a thiosulfate – thiocyanate mixture shortens the fixing time to 90 s.

The topic of color negative film processing also includes processes for motion-picture films. These differ from processes for color negative still films in the sequence of baths and the compositions.

Color Negative Motion-Picture Film Processing. The commonly used process for color negative motion-picture film is the Kodak ECN-2 process [332]. It exists in two variants using different bleaching agents. The recommended process employs peroxodisulfate ("persulfate"), but this agent requires a bleach accelerator bath immediately upstream of the bleaching bath (see Table 4).

Table 4. Sequence of baths in color negative motion-picture processing

Bath	Processing time, s	Temperature, °C
Pre-bath	10	27
Backing removal	5	27 – 38
Color developer	180	41
Stop bath	30	27 – 38
Bleach accelerator ∗	30	27
Washing ∗∗	30	27 – 38
Bleaching bath (persulfate ∗, hexacyanoferrate(III) ∗∗)	180	38 ∗, 27 ∗∗
Washing	60	27 – 38
Fixing	120	38
Washing	120	27 – 38
Stabilization	10	27 – 38

∗ First variant.
∗∗ Second variant.

The accelerating agent, an isothiuronium compound, has the Kodak designation PBA-1. The second, alternative variant uses hexacyanoferrate(III) as bleaching agent (see Table 4). Because persulfate is reduced to sulfate, its presence in wastewater poses no problem, in contrast to hexacyanoferrate [333].

Color Positive Motion-Picture Film Processing. The ECP-2 A process introduced by Kodak is also commonly used here [334] (see Table 5). Again, there are two variants, which differ in the bleaching bath [335].

Table 5. Sequence of baths in color positive motion-picture film processing

Bath	Processing time, s	Temperature, °C
Pre-bath	10	27
Backing removal	2	27
Color developer	180	36.7
Stop bath	40	27
Washing	40	27
First fixing	40	27
Washing	40	27
Bleach accelerator ∗	20	27
Bleaching bath (persulfate) ∗	40	27
Bleaching bath (hexacyanoferrate) ∗∗	60	27
Washing	40	27
Sound track development	20	room temp.
Spray washing	2	27
Second fixing	40	27
Washing	60	27
Stabilization	10	27

∗ First variant.
∗∗ Second variant.

This process differs from the one used for motion-picture negative film because the optical sound track on positive film must be brought up in the form of a pattern of silver metal.

Sound Track Development. After the negative and the optical sound track have been exposed, development takes place. Unexposed silver halide is removed by fixing, and a bleaching step follows. Since bleaching takes place in the light, the newly produced silver halide is also exposed. A narrow roller then applies the thickened sound track developer to the edge of the film, converting the sound track back into metallic silver. A second fixing bath finally removes the residual halide. The silver is needed to carry the optical sound signal, because the audio cells in projection equipment (cesium photocells) are sensitive only in the infrared, where the absorption of the film dyes is not adequate.

5.3.2.2. Color Reversal Film Processing

As an example, the E-7 process disclosed by Kodak in 1977 for the processing of Ektachrome films is described [336] (see Table 6).

Table 6. Sequence of baths in color reversal film processing

Bath	Processing time, s	Temperature, °C
First developer	360	38 ± 0.3
Washing	120	33 – 39
Reversal bath	120	33 – 39
Color developer ∗	360	38 ± 0.6
Conditioner ∗	120	33 – 39
Bleaching bath ∗	360	33 – 39
Fixing bath ∗	240	33 – 39
Countercurrent washing ∗	240	33 – 39
Stabilizing bath ∗	60	room temp.

∗ With lights on.

The first development is performed with hydroquinone monosulfonate and 4-hydroxymethyl-4-methyl-1-phenyl-3-pyrazolidone (Dimezone S). The subsequent reversal bath includes reversal agent RA-4 [contains tin(II) chloride]. CD-3 is used as color developer.

Table 7. Sequence of baths in the RA-5 process

Bath	Processing time, s	Temperature, °C
Color developer	45	35
Bleach – fixing bath	45	35
Washing	22.5	25 – 30
Washing	22.5	25 – 30
Washing	22.5	25 – 30
Washing	22.5	25 – 30

5.3.2.3. Color Negative Paper Processing

Since 1986, the composition of color paper has been changing. While all papers made before that time had silver chloride – silver bromide emulsions, Kodak began marketing a paper type with a pure silver chloride emulsion in that year. The new type and its processing have been adopted by all manufacturers as a new standard. The compatibility principle remains in effect.

Because silver chloride emulsions can be developed much faster than emulsions containing silver bromide, the processing time can be made much shorter than before: 45 s at 35 °C (versus 3 : 30 min at 33 °C). The bleach – fixing time is also greatly shortened through a reduction in pH, i.e., to pH 6.0 by addition of acetic acid. The process for color negative papers with silver chloride emulsions bears the Kodak designation RA-4 and the Agfa designation AP 94. The published RA-5 process is described as an example [337] (see Table 7).

Washing is carried out as a countercurrent operation in a four-stage cascade. Stabilizing baths replace the washing step in minilabs; there are many formulas for these, and no standard exists.

The processing of the still-existent silver chloride – silver bromide papers differs greatly from the method just described in the bath times and, above all, in the developer formula. This process is designated EP-2 (Kodak) and AP 92 (Agfa) (see Table 8).

Table 8. Sequence of baths in the EP-2 and AP-92 processes

Bath	Processing time, s	Temperature, °C
Color developer	210	33
Bleach – fixing bath	90	33
Washing	180	25 – 30

The developer differs in several principal ways: It contains benzyl alcohol and much more potassium bromide, and as antioxidants it has more sulfite, with hydroxylammonium sulfate used in place of diethylhydroxylamine.

Large laboratories commonly use a variant of the process, with bleaching, washing, and fixing, instead of a single bleach – fixing bath. The ease of rejuvenating the bleaching and fixing baths and of desilvering the fixing bath are advantages where enough space for larger processing machines is available. This variant is available for both types of color paper.

5.3.2.4. Color Reversal Paper Processing

The processing of color reversal paper uses a developer with significant elements of color reversal film processing. The first development is done with hydroquinone sulfonate and Dimezone S (see Section 5.3.2.2); color development is with developing substance CD-3 and the development accelerator is Hs 104 (ethylenedithiodiethanol). The fogging treatment is replaced by a diffuse reexposure, and the color developer also exhibits a much lower pH (10.4 instead of 12.4) and contains benzyl alcohol. The bleach – fixing bath is largely the same as in color negative paper processing. A wide range of materials and processes are again compatible. The processes are designated R-3 (Kodak) and AP-63 (Agfa); see Table 9.

Table 9. Sequence of baths in color reversal paper processing

Bath	Processing time, s	Temperature, °C
First developer	75	38
First washing (2 stages)	90	35 – 41
Reexposure	not specified ∗	not specified ∗
Color developer	135	38
Second washing	45	25 – 41
Bleach-fixing bath	120	38
Third washing (3 stages)	135	25 – 41

∗ Diffuse reexposure is performed with a lamp located above inlet to developer tank.

5.3.3. Developing Equipment [28], [338]

Amateurs and small-scale users develop materials in plastic trays. To achieve consistent development results, especially of color materials, small-scale users also employ developing tanks, some featuring motor drive and automatic temperature control. Industrial operations, X-ray laboratories, and large-scale processors usually have automated developing equipment. Wet processing consists of three steps – development, fixing, and washing (black-and-white materials) – up to a maximum of nine steps (color reversal materials) and generally requires between 90 s (X-ray films) and ca. 40 min (color reversal films). In the usual processes, the temperature is between 20 and 40 °C. It must be held precisely constant, particularly in the case of color materials (± 0.2 °C). In special cases, rapid development within a few seconds is possible if elevated temperatures and diminished quality are acceptable. The photographic material and the developer must be kept in motion relative to each other by mechanical agitation, squeezing, or introduction of nitrogen, so that development will proceed in a uniform way. The exact processing conditions depend on the model of equipment and are specified by the maker of the photographic materials.

In order to obtain consistent results over large quantities and long periods of time, developing equipment must meet the following conditions:

1) Constant temperature
2) Constant residence times in the individual baths
3) Constant chemical composition of the baths
4) Uniform agitation
5) Little carryover of solutions from bath to bath

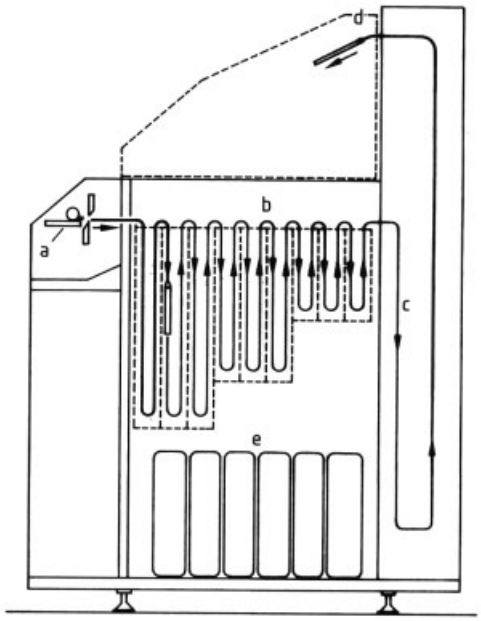

Figure 46. Continuous-flow film developing machine for a minilab
a) Film loading station with knives to separate film from cartridge; b) Tanks with developing baths and film transport units (racks); c) Hot-air dryer; d) Delivery of developed film, leader card first; e) Tanks for replenisher solutions (4) and spent baths (2)

The following types of equipment are used:

Plastic trays and tanks find use mainly for the processing of single sheets of B & W paper and film. *Developing tanks* can accomodate single sheets or roll films on spools. Motor drives for these devices can also provide uniform agitation, and the temperature can be held constant with water baths. In contrast to tanks, *developing drums* offer automation of movement and temperature control. They are used for processing of larger formats.

Developing tanks and drums have no replenishment devices to hold the chemistry of the solutions constant. The baths are used only once, and these types of equipment are therefore called "one-time" developing machines.

When large amounts of material are to be processed and the standards of processing quality and convenience are more rigorous, automatic developing machinery with mechanized transport of the photographic material is used.

In one class of machines, sheets or films are clamped to *hangers or frames* and transported through the baths. With appropriate hangers, single sheets and roll films of various formats can be developed consistently and safely in the same machine. These devices are therefore widely used. However, in large laboratories where there is a heavy flow of standard formats they are being displaced by continuous developing machines.

In continuous film developing machines, friction drives transport motion-picture films, spliced or individual still films through the tanks. In contrast to roller-transport machines (see below), these machines are not self-threading but require a leader strip or card (see Fig. 46).

The capacities of such machines are from 20 films per hour in minilab designs to 2000 films per hour in large-scale developing and printing plants. Modern film processes have a "dry to dry" time of ca. 8 min.

Roller-transport machines for films or papers have driven roll pairs that transport the photographic material through the baths without clamping. These devices are particularly well-suited to sheet films (such as X-ray films), single sheets, and short strips of film or roll paper (Figs. 47 and 48).

A special design is used for *roll-fed papers of limited length*. The roll pairs are widely spaced in their rack. For reliable conveyance of the leading edge from one roll pair to another, the paper is inserted in V-shaped (swallowtail) guides. This transport method is employed chiefly in the widely-used compact paper developing machines found in minilabs (see Sections 5.6.3 and 5.6.4). These machines, working in conjunction with a printer, have capacities of 500 – 2500 prints per hour; the processing time from exposure of the paper to the dry print is 4 – 5 min with the modern rapid process.

Conveyor-type paper developing machines, also called *roll-fed paper developing machines*, are used only for the development of paper in roll form. The leading edge of the paper is fastened to a clamp on the conveyor belt. These machines can be built with very large capacities (see Section 5.6.4). Up to six paper strips can be fed one beside another at a transport speed of up to 30 m/min. With six strips, the development capacity is ca. 80 000 prints (9 × 13 cm) per hour (see Fig. 49).

All automatic developing machines have thermostatically controlled bath temperatures. The solutions are continuously circulated and filtered. They are replenished in accordance with the area of photographic material processed. Doctor blades, air squeegees, or suction heads are used to prevent the photographic material from carrying over much of the solution from one bath into the next one or into the dryer. All these machines have integral warm-air or radiant-heat drying. The size and number of tanks depends on the processing width, dwell times, and processing speed. Special devices are available for silver reclamation and environmental cleanup of baths and wash waters.

Photographic baths, especially color developers. may contain toxic substances. Direct contact may result in inflammations and rashes, and inhalation of vapors can cause swelling of the eyes as well as allergic reactions. Developing equipment must therefore be designed and operated in such a way that direct contact with the chemicals is prevented and the vapors produced cannot be inhaled (exhaust fans or similar provisions). The instructions of the chemical manufacturers must be strictly complied with.

5.4. Silver-Dye Bleach Process

In contrast to chromogenic materials, silver-dye bleach materials in the unprocessed condition contain all the dyes needed to form the image. In processing, the portions of the dyes not required for the image are chromolytically decomposed and washed out.

Chemically, the process is based on the reduction of azo image dyes by metallic silver. Silver metal is produced, in an amount proportional to the exposure, in the first processing step; a classical B & W development. Next, in the strongly acidic bleach bath, the azo dyes are reduced by the silver and thus decomposed. Reduction can only occur if the bleaching bath contains strong ligands for Ag^+, which transform the silver into a strong reducing agent. In practice, iodide (or, less often, thiourea) is used; the product is then insoluble silver iodide (or soluble silver – thiourea complexes).

The dye bleach reaction can be represented as

$$R^1\text{–N=N–}R^2 + 4\,Ag + 4\,HI \rightarrow R^1\text{–}NH_2 + H_2N\text{–}R^2 + 4\,AgI$$

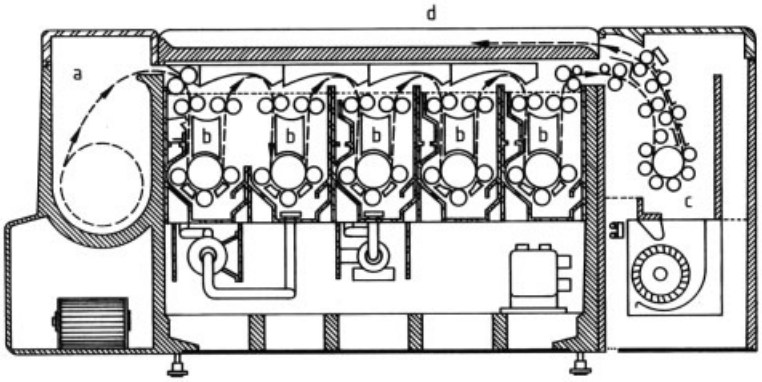

Figure 47. Sheet developing machine for color papers with roller transport system
a) Darkroom loading station; b) Tanks with developing baths and transport rollers; c) Hot-air dryer; d) Paper delivery

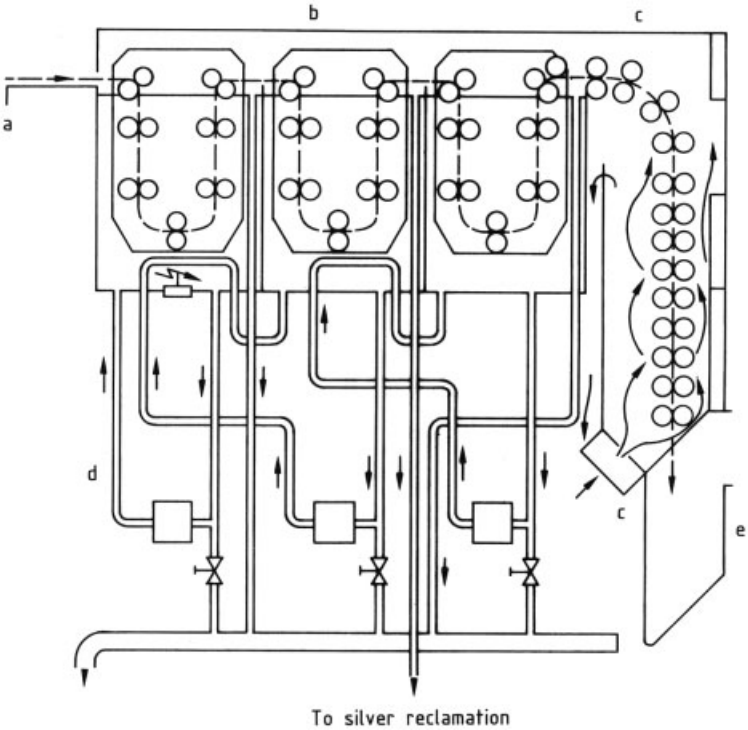

Figure 48. Developing machine for X-ray film
a) Loading tray; b) Processing tanks with transporting roller sets; c) Hot-air dryer; d) Recirculation and temperature- control system; e) Film delivery

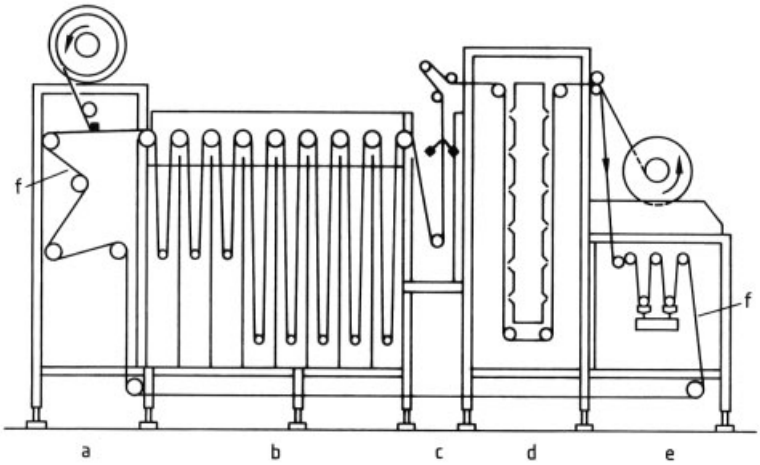

Figure 49. Conveyor-type paper developing machine
a) Loading section with paper unrolling and drive; b) Processing tanks; c) Drip tank with doctor blades; d) Hot-air dryer; e) Delivery section with paper winding; f) Endless conveyor belt to guide leading edge of paper

Since the azo dyes and the developed silver are both immovably embedded in the gelatin layers and thus cannot react directly with one another, suitable redox catalysts are needed to effect the reduction. The only agents used in practice are quinoxaline derivatives, which can be reduced to a reactive 1,4-dihydro form with consumption of two reduction equivalents. This 1,4-dihydro form is able to diffuse and reacts with the azo dyes, regenerating the quinoxaline derivative.

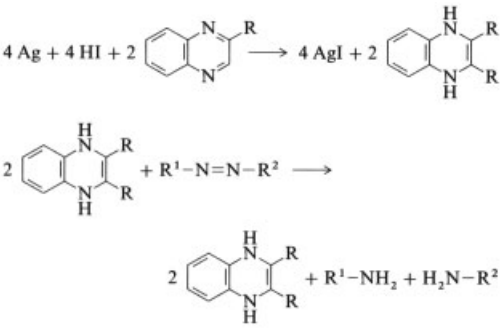

Silver-dye bleach materials have the following basic structure: A gelatin layer containing a cyan azo dye and silver halide that is sensitized in the red range is coated on an opaque or transparent support. Next, a green-sensitive silver halide emulsion with a magenta dye is applied, and on top of this a blue-sensitive emulsion containing a yellow dye is added. These layers are separated by pure gelatin interlayers in order to prevent unwanted long-range bleaching effects in adjacent layers. The yellow image dye protects to a certain extent the green-sensitive and red-sensitive emulsions against exposure in their intrinsic sensitivity regions (by blue light); the interlayer below the yellow layer may also contain an additional yellow filter dye for more complete suppression of such exposure.

The dyes are bisazo or trisazo compounds substituted with sulfo groups that are largely unable to diffuse through the gelatin. The amines produced from the dyes on bleaching are virtually colorless; they can diffuse in the gelatin and have relatively good solubility in water.

After exposure, these materials are *processed* in the following way: The negative silver images in the three emulsions are generated by a classical superadditive B & W developer. The next step is treatment in the bleaching bath, which contains the redox catalyst and a mild oxidizing agent (an aromatic nitro compound) along with ca. 0.5 M sulfuric acid and ca. 0.1 M potassium iodide. The nitro compound competes with the dyes as a reaction partner for the reduced catalyst. It limits the lifetime and range of the reduced catalyst and thus blocks interlayer effects, even when the interlayers are thin, and also ensures the complete oxidation of excess silver. Next the silver compounds are removed in a rapid fixing bath. Finally, the material is washed to dissolve away any dye fragments that have been left behind in the emulsions.

The result is a positive copy of the original. Since the absorption maxima of the image dyes roughly coincide with the sensitivity maxima of the respective silver halide emulsions, silver-dye bleach materials have only modest sensitivity and are therefore used chiefly as print materials on opaque or transparent supports for displays. The outstanding antihalation effect of the image dyes yields a very high sharpness, therefore silver-dye bleach materials are used as microfilms. The azo dyes used to form the image afford high light stability.

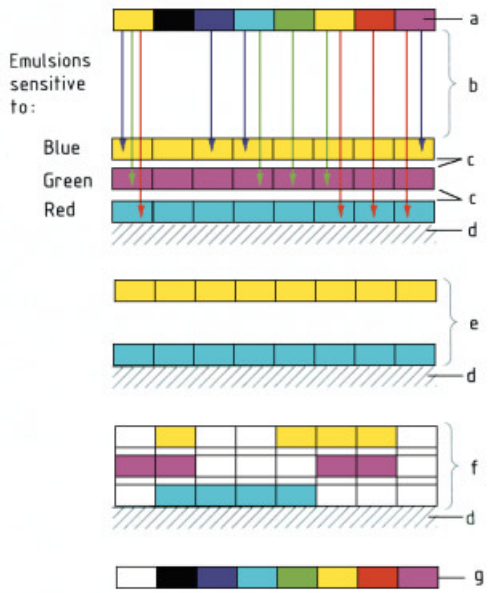

Figure 50. Silver-dye bleach process
a) Original; b) Light transmitted or reflected by original; c) Interlayers; d) Base; e) Emulsions with developed silver; f) Emulsion layers after bleaching and fixing; g) Copy of original for viewing by reflected or transmitted light

The problem of unwanted secondary absorptions of the dyes, chiefly the secondary blue absorption of the magenta dye, is corrected by *self-masking techniques* in modern silver-dye bleach materials. The yellow dye at a given point in the image is bleached in proportion to the secondary blue absorption of the magenta dye present at the point. This self-masking technique is based on a supplementary layer beneath the yellow layer, in which the silver image negative to the green exposure is developed. The corresponding fraction of the yellow dye is then rendered colorless by long-range bleaching in the bleaching bath. The silver dye bleaching bath is shown in Figure 50. Other general descriptions of the silver dye bleach process are given in [339–341]. *Trade names* include Ilfochrome Classic, Ilfochrome Rapid.

5.5. Dye Diffusion Processes

The *Technicolor process* for motion-picture film copying, now obsolete and still in use only in China, was based on the successive transfer of three positive separation images to a mordant layer. It was thus a predecessor of the image-forming system used in instant photography. The same holds for the Kodak dye-transfer process.

5.5.1. The Polaroid Instant Photography Process

In 1963, Polaroid brought out Polacolor instant film; thus the transition was made from instant B & W photography, which used the silver halide diffusion process, to instant color photography by dye diffusion [342] (see Fig. 51 A).

In dye diffusion [343], the photographic process causes a different diffusivity of the various dyes or dye precursors that form the image. Diffusion starts at the beginning of development and ends at roughly the same time as silver development is complete; i.e., when a drop in pH in the stack of layers is induced.

Of the many chemical systems that have been described for the production of color prints by the dye-diffusion method, only a few have been turned into commercial materials. All use the principle of subtractive color mixing and are thus fundamentally different from the additive processes and for instant color transparencies or instant color films.

Instant materials employ either a *two-sheet* (*peel-apart*) *system* (see Fig. 51 A), in which the negative sheet and the image-receiving sheet are united for the dye transfer, or a *single-sheet* (*integral*) *system* (see Fig. 51 B), in which the image-recording part with the image-generating chemicals, the processing chemicals, and the image-receiving part are integrated.

Commercial materials derived from Polacolor II, the improved version of Polacolor (large formats, television screen recording, identity

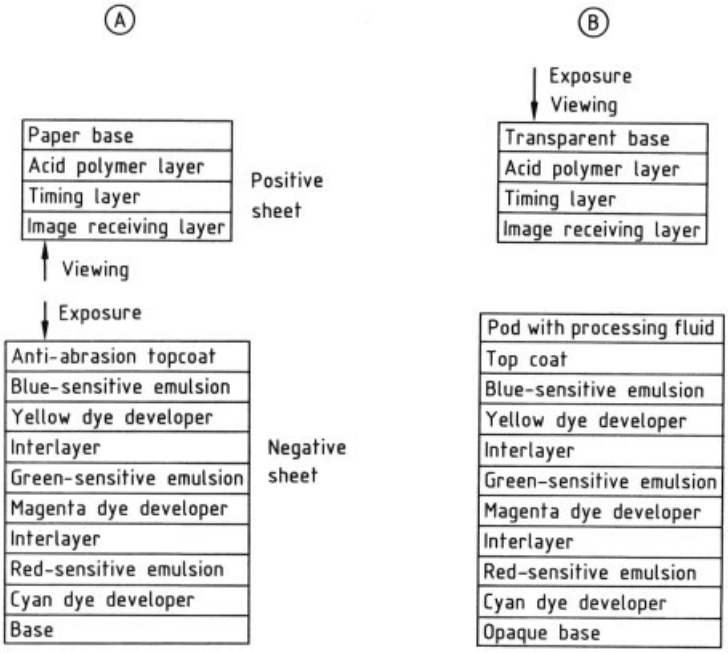

Figure 51. Schematic layered structure of Polacolor instant film (A) and SX 70 integral film unit (B)

photographs), are peel-apart materials. Single-sheet systems are predominantly used in amateur photography.

All integral materials are covered by basic patents held by Polaroid and are extraordinarily complicated. Obtaining an image whose quality is comparable to that of conventional color photographs requires not only the photographic emulsion layers proper but also a number of auxiliary layers, which control the process and compensate for the inherent temperature dependence of the system.

The image dyes (yellow from the blue-sensitive emulsion layer, magenta from the green-sensitized layer, and cyan from the red-sensitized layer) are fixed in the image-receiving layer by a mordant and viewed against a white background. Polymers with cationic groups serve as mordants. The dyes are held in place by ionic bonds.

Currently the time required to obtain a picture is in the range of 1 – 3 min, depending on the temperature. An acidic polymer layer stops silver development and dye diffusion. It is separated from the other layers by at least one timing layer consisting of an alkali-sensitive polymer. The acidic layer abruptly becomes alkali-permeable only after the proper action time. These layers can also contain additives to prevent both fogging and subsequent increases in density.

The processing chemicals are embodied in a highly viscous, strongly alkaline paste, which is squeezed out of a pod by rollers and spread through the integral film unit as a film when the unit is removed from the camera. Excess paste is collected in traps. The correct, uniform thickness of the paste film is ensured by the stiff backing, spacers, and image masks.

The Polaroid Image System (U.S. Brand Name: Spectra) The Polaroid Image System [344] (see Fig. 52) was developed from the SX-70 system (brought out in 1971) and consists of the Impulse series of cameras and P-600 Plus film.

The exposure and viewing directions are the same. The transferred image is examined through the backing, against the white-pigmented background of the developer paste. The image-generating part must therefore be rendered opaque from the moment is leaves the camera until the image-production time has

passed. This is done by a decolorizing indicator dye in the processing paste.

The Polaroid system operates with negative emulsions. For magenta and cyan, the reversal of the image is achieved through the use of dye developers, which were used in the SX-70 system and its predecessors. The *magenta* (Fig. 53 A) and *cyan dye* (Fig. 53 B) *molecules* [345] contain hydroquinone residues and therefore act as developers: they are dissolved by alkali and have the ability to diffuse. By transferring electrons to the auxiliary developer, which actually develops in silver halide, the dye developers in exposed regions of the image are converted to insoluble and immobile quinoid products. Only the unoxidized portion of the dye is transferred to the image-receiving layer.

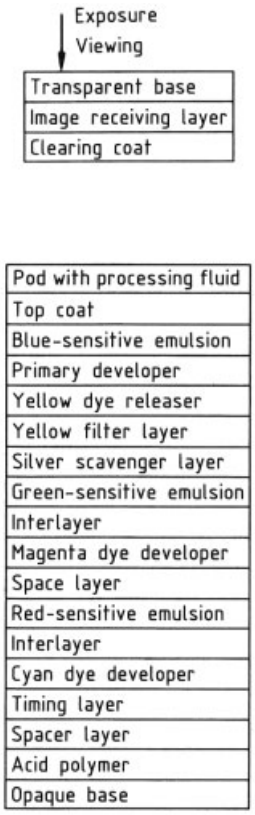

Figure 52. Schematic layered structure of a Polaroid image film unit

The *yellow image* in Polaroid P-600 film is based on a dye-releasing compound working on the argentolytic principle [346]. In other words, the compound reacts with silver ion to form a silver complex and release the yellow dye. This system, in which the silver ions of the undeveloped grains are utilized, also works to reverse the image. Diffusion control ensures that only dissolved silver ion from the blue-sensitive layer can react with the dye-releasing compound, a nondiffusing thiazolidine compound.

5.5.2. Kodak and Fuji Instant Photography Processes

Kodak PR-10 System. The Kodak PR-10 system [347], which also used integral film units, was introduced in 1976. Fundamentally new features that characterized the system were the use of nondiffusing dye-releasing compounds with developer properties (from which the highly diffusible image dyes – simple anionic monoazo dyes – were released by oxidation and subsequent hydrolysis) and the use of unfogged direct positive emulsions for image reversal. The dye-releasing compounds were 4-sulfonamido-1-naphthols [348]. The system underwent a number of improvements, which led to better shelf life and higher speed (ISO 320). The Kodamatic Trimprint system (1983) provided for the subsequent separation of the transfer red dye image from the negative. After losing a patent suit to Polaroid, Kodak withdrew the instant photography system from the market in 1987.

The Fuji System F 10 Fotorama. In 1981, Fuji brought out an *integrated* instant-picture material compatible with the Kodak PR-10 system (format 4×5 inches = 10.15×12.7 cm) but utilizing different dye-release chemistry [349]. The system was intended for the Japanese domestic market.

A further development of this system is the Fotorama system with F 1-800 G *peel-apart* film, which is for sale in Japan. Since the abandonment of instant photography by Kodak, there is no need for compatibility with the Kodak system. The Fuji instant film, like that in the Kodak system, uses negative-working dye-release chemistry.

The *ortho*-sulfonamidophenol dye releasers in the Fuji material (see Fig. 54), like the *para*-sulfonamidoaphthols in the PR-10 system, are strong reducing agents, which are converted

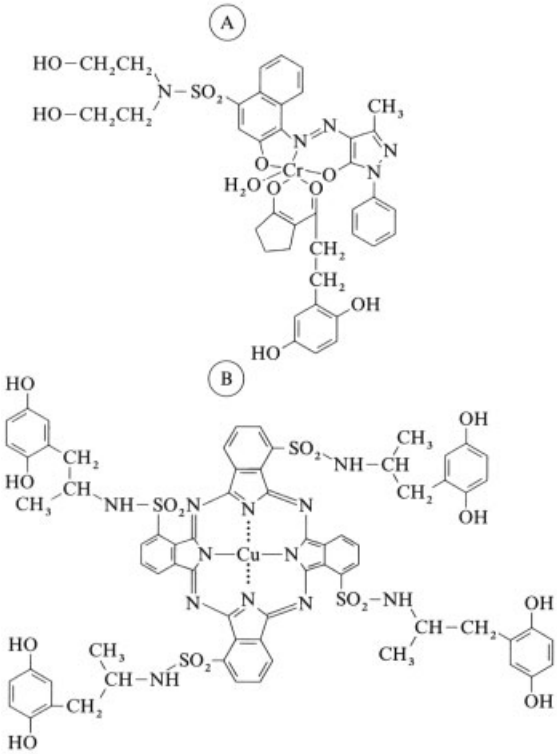

Figure 53. Dye developers in the SX-70-Time Zero and Polaroid image unit
A) Magenta dye developer; B) Cyan dye developer

to alkali-labile *ortho*-quinonesulfonylimines after oxidation. Because of the high alkalinity of the processing paste, the dye group – which is attached to the quinoid dye-releaser oxidation product by the sulfonylimine group – is split off from the nondiffusing carrier by hydrolysis, thus becoming diffusible (see Fig. 55).

The negative transfer characteristic of the *ortho*-sulfonamidophenols means that the required image reversal must be accomplished through the emulsions. Remarkable advances in emulsion technology have now made it possible to use high-speed, unfogged direct positive emulsions of the internal- center type (rated speed: ISO 800). These emulsions are not developed after exposure (see Section 3.1.1). Instead, the action of an incorporated hydrazine derivative in the developer causes subsequently developable reduction centers to form on unexposed grains. In this way, high dye-transfer densities are achieved in the unexposed regions, while exposed portions of the image remain virtually transfer-free. Electron transport between the emulsion and the nondiffusing dye-release compound is implemented by Phenidone-type auxiliary developers.

The integral film unit is taken out of the camera immediately after exposure, passing through a pair of rolls that rupture the pod and spread the paste into the space between the cover sheet and the image sheet (i.e., later separation of integral sheet). Unwanted re-exposure through the cover sheet is prevented by carbon-black in the paste. In contrast to the Polaroid system, the image-receiving layer is located on the back of the film unit. The image is viewed against a white-pigmented layer (containing titanium dioxide), which in turn is backed by an opaque carbon-black layer. The Fuji instant picture also has auxiliary process- control layers in the cover sheet.

The image-receiving part is structured as in "stripping film"; the transfer image is separated from the image-generating part after the process is complete.

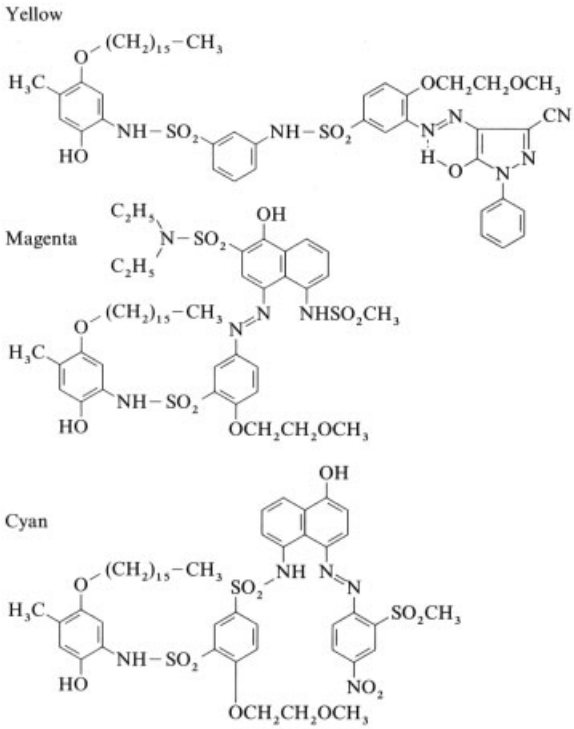

Figure 54. Dye-releasing compounds of Fuji instant color film Fl-800 G

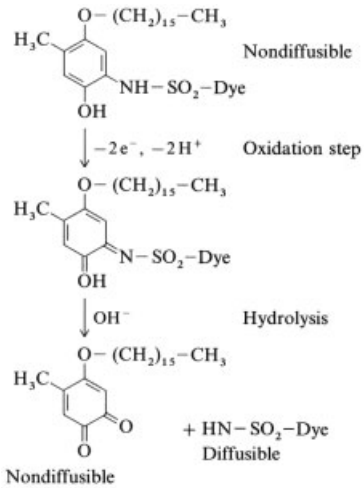

Figure 55. Dye-releasing mechanism of Fuji instant color film Fl-800 G

5.5.3. Agfachrome Speed and Ektaflex

Eastman Kodak, Agfa – Gevaert, Konishiroku, and Fuji have proposed newer dye-release systems [350–353] in which the bond between the dye and the nondiffusing carrier is broken by reduction instead of oxidation. The electron-donor compound used as reducing agent is accordingly consumed by the silver halide in regions where development takes place and by the dye-release compound in the unexposed regions.

These dye-release systems are positive-working and yield fog-free transfers when the process of silver development is ensured. The best way to do this is by alkali-labile blocking of the electron-donor compound.

5.5.3.1. Agfachrome Speed

The sheet material Agfachrome Speed [354] was introduced in 1983, simultaneously with the Copycolor graphic-arts material (→ Imaging Technology, Chap. 2.2.3.2.), but is no longer on the market. Agfachrome Speed was a print material for the home laboratory. It used dye-releasers of the quinone methide type, with which brilliant prints could easily be made from transparencies. The material was exposed through the backing but viewed from the emulsion side against

a white-pigmented background. Processing was done by immersing the material in an alkaline activator.

5.5.3.2. Kodak Ektaflex PCT [355]

Kodak has not brought out any material with the new dye-release systems. Both the negative version (1981) and the reversal version (1982) of the Ektaflex PCT peel-apart print material utilize the negative-working dye-release chemistry of the PR-10 system, i.e., the Ektaflex PCT Reversal Film in combination with direct positive emulsions. Both materials have now been withdrawn.

5.5.4. Additive Instant Color Photography Systems

Additive instant color systems do not produce and transfer a dye image. Instead, a threecolor additive linear screen, in the primary colors blue, green, and red, is overlain by a silver positive to form the image. The positive is obtained by silver halide diffusion in a receiving layer. Accordingly, the projected brightness of an additively produced instant picture is roughly two-thirds less than that of a conventional transparency. The technical challenge of the system lay in the costly process of obtaining an extremely fine linear screen from dyed, dichromated gelatin in three layers along the optical path; i.e., virtually invisible when projected [356].

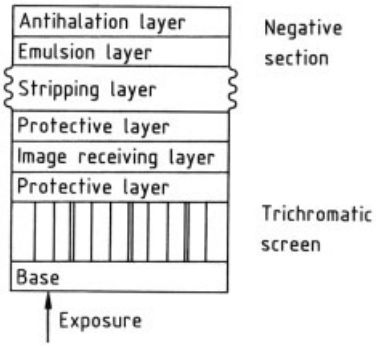

Figure 56. Schematic layered structure of a Polachrome film

5.5.4.1. Instant Color Transparency [357]

Two 35 mm transparency films, Polachrome CS (introduced in 1982, speed ISO 40) and Polachrome HCP (a high-contrast film), use an additive linear screen having 394 line triplets per centimeter (see Fig. 56). They are exposed through the base and the screen. The negative part is stripped off in the processor. The transparencies are put in slide frames and ready for projection immediately after they are processed.

5.5.4.2. Narrow-Gauge Instant Color Film [356]

The Polavision narrow-gauge film system (Super 8 format) consisted of player, camera, and film cassette. The film had a linear screen with 590 triplets per centimeter. The structure was not altered in processing; the individual image, with low light-transmitting power, was projected through a superimposed negative with extremely low covering power. Polavision therefore remained a specialty material for limited applications. It was taken off the market in 1988.

5.6. Making Color Prints [28], [338]

In black-and-white photography, the illumination in the printer must be matched to the sensitivity of the print material and the density of the original, such as a negative. In color printing, the illumination must be separately controlled for each of the three emulsion layers in the paper (sensitive to blue, green, and red) in order to obtain a print of the correct density without color distortion. The printing light is controlled as the color and exposure by the additive or subtractive process. Correct illumination can be determined empirically or with the help of measuring devices. Color prints can be produced from color negatives or, as direct positive prints, from transparencies or color prints (on color reversal paper).

5.6.1. Printing Processes

Additive Printing Process. The basic principle of the additive printing process is the separate exposure of each emulsion in the print material through filters in the additive primary colors

blue, green, and red. The three exposures can be performed one after another by inserting the three printing filters in the ray path one by one, or simultaneously by combining the properly filtered light from three printing lamps into a single beam.

The light in each of the three color channels is precisely metered to the process in one of two ways: at constant intensity and variable exposure time, or by spectrally neutral brightness modifications. For example, if the first positive print made from a color negative comes out too yellow, the exposure time or the intensity of the blue light must be reduced so that the blue-sensitive yellow emulsion receives less exposure.

The additive printing process has the advantages that it is easy to follow, the light control is cleanest from the spectral point of view, and the printing light can be matched to the absorption of the negative dyes and the sensitization of the positive material. Its principal drawback is the low intensity of the light passing through the additive filters, which leads to excessive exposure times and the necessity of using high-output printing lamps.

Subtractive Printing Process. In the subtractive process, the desired color balance is achieved by placing subtractive filters (yellow, magenta, cyan) in the projection beam of the printing equipment. For example, if a positive print made from a color negative without filters turns out too yellow, a yellow filter of suitable density is inserted to stop part of the blue light. The blue-sensitive yellow emulsion now receives less exposure, and the yellow in the print is de-emphasized. For fine control, either a color filter of the proper density can be selected or a color filter of maximum density can be inserted partway into the beam. In the second case, however, it is necessary to ensure especially good mixing of the light downstream of the filter.

The most commonly used method of subtractive exposure is a "white" main exposure followed by a subtractive second exposure. The first step is an exposure with white light, which contains the blue, green, and red components. When the exposure in a certain color is sufficient, the subtractive filter blocking that color is pivoted into the beam. For example, if the blue exposure is finished first, a yellow filter is inserted which blocks the blue light but lets the green and red light through. A similar procedure is followed when the exposure of another color is sufficient; that is, a second substractive filter is pivoted into the beam along with the first, so that the only light used for continued exposure is the color passed by these two filters in series. For example, a combination of yellow and cyan filters is transparent only for the color green, and green can be finally blocked by adding the magenta filter. For the principles by which the pivoted or sliding filters are controlled, see Section 5.6.2.

The subtractive printing process has the advantage of relatively simple design and high light intensity, but it is complicated by the secondary absorptions of most color filters. Most color printing machines use the subtractive method.

5.6.2. Exposure Control

Because the eye is more sensitive to slight differences in color than it is to density differences, the color matching in the print must be performed with special care, and for this purpose the correct color exposure must be determined precisely.

The simplest but most laborious method is to make test prints (test strips) with various filter arrangements, develop them, and select the combinations that give the optimal color match. When quality standards are high, these operations must be repeated several times. This primitive method is now used only in special cases. All modern color printers employ exposure control based on optical measurement of the negative and calculation of the quantities of blue, green, and red light needed for exposure of the paper. In simpler machines, the integral transmission densities of the negative in the colors blue, green, and red are measured, and on the basis of the values found the light intensities in the three colors are controlled in such a way that all three emulsion layers in the print material produce about the same color density. This method is based on the assumption that all three primary colors are represented in equal strength in an average amateur picture (neutral gray principle). Together with supplementary conditions for overexposed and underexposed pictures as well as large variations in color and density from region to region inside the negative, this principle results in almost 90 % color-correct prints. The remain-

der are reprinted with individual correction. A simplified version of this kind of machine is the *color analyzer*.

In color printers that must satisfy higher standards, the image analysis is carried out with much greater precision. The negatives are broken down into 100 or more pixels, and the densities at the individual scan points are measured in the blue, green, and red regions (*color scanner*). A complex analysis of these local color and lightness values permits a much more exact determination of the quantities of light needed for printing.

As in the integral method, however, data on typical printing properties of films from various manufacturers must be determined empirically and entered in a program (printer memory), so that the proper exposures can be calculated from the data for the given film together with the measurements on the negative. Films now on the market have the film type coded on the edge in machine-readable form (DX code) so that the type can be determined automatically.

The state of the art in measurement and analysis is *total film scanning*, in which the measurements on all negatives of the film being printed are lumped together and analyzed. In this way, statistical methods can be applied to determining the photographic properties of the film (film-specific data) separately from incidental features of the individual negatives (image-specific data).

This procedure not only permits the first-class reproduction of all exposures from a film, but also the processing of all film types from all manufacturers without preprogramming (one-channel processing). For average amateur negatives, the yield of prints which are correct with respect to color and brightness is ca. 99 % with this technique.

5.6.3. Types of Color Printers

The classical darkroom enlarger is currently used only for special tasks. Virtually all color prints are made on color printing machines, which can be classified by application as high-capacity, universal, special-purpose, and minilab printers.

The principal use of *high- capacity printers* is usually the fully-automated mass production of prints from freshly developed amateur films (first print). These machines have film scanning systems and offer capacities of 10 000 – 20 000 prints per hour. Photographic paper and films must be supplied in strip form already wound on spools; films are spliced together before develop-ment. These machines can make prints in formats up to 13 × 18 cm.

Universal printers are also used in printing amateur pictures. They can handle special jobs that do not lend themselves to automated processing (roll films, detail enlargements, mini-posters, etc.). Universal printers feature simple integral measuring systems and operate rather slowly, because the technical setup depends on an operator. These printers can print on the appropriate papers from both negative films and reversal films (transparencies).

Special-purpose printers are related to the universal type but can handle wide papers (up to 75 cm); most of them have a wide range of continuously variable enlargement ratios, which makes them suitable for all kinds of special jobs. Such a printer requires an operator and is used mainly for professional photographic work.

The *video printer* is a variety of special-purpose printers. In this device, the exposure process is first electronically simulated, and the result is displaced on a color monitor. The operator can modify the monitor image until its brightness and color are satisfactory. Because the printer can be controlled in such a way that the print exactly matches the display, a high-quality print can be made without subsequent costly testing.

The *minilab printer* (Fig. 57, see also Section 5.6.4) is directly connected to a paper developing machine. This type of printer is related to the universal type, but the higher-quality models have modern negative scanners such as those in high- capacity printers. These machines are designed for capacities of ca. 500 – 2000 prints per hour. By virtue of the expensive measuring instruments, even photographically untrained personnel can achieve very good print quality on a first attempt.

The applications of *digital printers* fall between those of universal and special-purpose printers. Also called *cathode ray tube (CRT) printers* or *laser printers* (depending on their operating principle), these devices represent an entirely new generation of color printers. While conventional color printers have the capability of

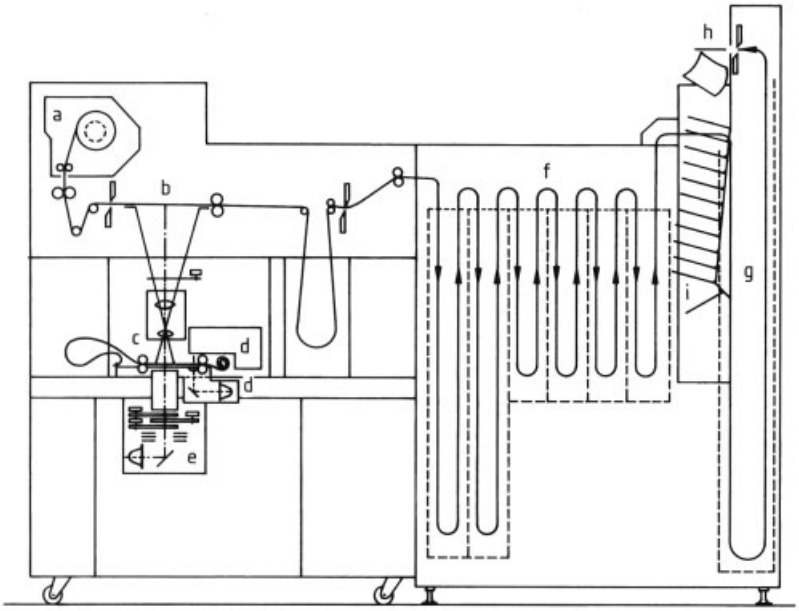

Figure 57. Minilab printer with paper developing machine
a) Paper supply cassette; b) Paperdeck; c) Variable-ratio objective and negative carrier; d) Color scanner with illumination system; e) Printing lamp house with exposure filters; f) Tanks with developing baths and roller transport system with V-guides; g) Hot-air dryer; h) Print cutter; i) Print stacker

modifying only the brightness and color of the print, the digital printer can also vary the contrast and color saturation as well as enhance the preceived sharpness. These advantages are achieved by placing the original (transparency or negative) in a high-resolution color scanner where it is broken down into several million image points (pixels). The pixels are stored electronically in "image memories"; a computer used to modify them in accordance with certain criteria (image processing) leads to the features just cited. The pictorial outcome of these operations can be continuously monitored on a color display. Finally, the signals representing the manipulated pixels control the intensity of a laser beam or of the electron beam in a CRT, which exposes the light-sensitive material (e.g., normal color negative paper) point by point in order to create the print.

Once the original has been scanned, the image exists in the form of electrical signals anyway; hence the electronics could also be supplied with external signals representing any electronically stored image. Such an image can be optimized and printed as described above. This procedure is significant chiefly in conjunction with electronic still photography (still video camera).

Another new picture-storage medium that has become commercially established is the *Photo Compact Disc* (*Photo CD or PCD*). The PCD is a plastic disc, similar to an audio CD, on which up to 100 color negatives or color transparencies (size 24×36) can be stored by means of a high-resolution optical scanner. The pictures are retrieved by a PCD player and can be used for various purposes. For example, they can be displaced on a television screen or used in preparing layouts in desktop publishing and – in combination with image processing – for producing professional graphics.

5.6.4. Laboratories for Amateur Color Photographs

The mass production of amateur photographs in large laboratories involves the logging, tagging, and splicing of the films and their development in continuous-flow developing machines. After development, the photos are printed on color paper in fully-automated printers. The exposed roll

paper is developed and cut. The finished pictures, along with the films, are shipped back to the outlets where the films were accepted. The centralization and automation of photo production has led to a steady decline in the price of amateur pictures since the 1960s despite increasing wages and material costs.

Since the end of the 1970s, an alternative to mass production in the large laboratory has been gaining importance: the *minilab*. The minilab, usually an in-store operation in a good shopping location, is a small laboratory set up for film development, printing, and paper development and having a capacity of 150 – 500 films per day. The equipment for such an operation is specially designed for this application, combining compactness with advanced technology (see Sections 5.3.3 and 5.6.3). These minilabs are attractive to customers because most offer one-hour service and the amateur can personally discuss the photographs with the laboratory personnel.

5.6.5. Copying of Motion-Picture and Television Films

Internegatives and interpositives from motion-picture and televison films are also made by the additive and subtractive methods. Because film transport is uniform, however, each frame must receive the same exposure time, so that the density and color can be controlled only through the filtering and the intensity of the copying light. Two approaches are used in determining the proper filters or intensity: test copies are made from the beginning or end of each scene, or video analyzers are employed.

The duplicates themselves are made by the additive or the subtractive process. The most commonly used *additive printers* have an exposure system in which the light from a copying lamp is split into the three primary colors blue, green, and red by dichroic mirrors. Each primary color beam has an electrically adjustable light-damping element ("light valve") inserted in it. After passing through the light valves, the three beams are reunited and guided to the printing station. Whenever the scene changes, a message carried on the original initiates a readjustment of the light valves in accordance with a program stored on punched tape or electronically. In *subtractive* printing, the color and gray filters needed for each scene are mounted on a black paper strip with punched holes and automatically brought into position scene by scene.

In the interest of consistent quality in the printing of television and motion-picture films, sensitometric monitoring of the process is particularly important. An immersion liquid can be used at the printing station in order to prevent loss of sharpness and contamination.

6. Photographic Properties and Their Measurement

This chapter describes the properties that govern the quality of photographs. "Quality" is thought of in very broad terms. In addition to properties such as tonal reproduction, sharpness, granularity, and color reproduction, the discussion also includes the stability of films and images. Improvements in stability, particularly of color materials, has been the focus of intensive research by manufacturers of photographic materials in recent years. The methods used to measure the properties cited are also described.

6.1. Sensitometry and Densitometry

*Sensitometry*currently means the study of the relationships between the developed photographic image and the exposure. Formerly, this term was restricted to determining the sensitivity of photographic emulsions. Scientific sensitometry was founded by HURTER and DRIFFIELD [358]. A detailed summary of sensitometric methods can be found in [34], [36]. *Densitometry* is the measurement of optical density in transmission or reflection.

6.1.1. Optical Density

The image as perceived is created by variations in the absorption of light in the emulsion, corresponding to the quantities of silver or dye formed after exposure and development. If Φ_0 is the radiant flux incident on a developed, transparent emulsion and Φ_τ is the radiant flux transmitted

after absorption, then $\tau = \Phi_\tau / \Phi_0$ is the transmittance and $D_\tau = -\log \tau$ is the transmittance density.

Similarly, for photographic papers viewed in reflected light, $\rho = \Phi_\tau / \Phi_0$ is the reflectance and $D_\rho = -\log \rho$ is the reflectance density.

In photographic papers, incident radiation is reflected in the baryta coating or, in resin- coated papers, in the titanium oxide- containing resin films; a portion of the incident radiation reaches the fiber base of the paper before being reflected. As it passes through layers containing silver, the light is scattered. The scattering depends on the size and structure of the silver grains. The measured density or diffuse density is therefore a function of the geometry of the measuring setup (see Section 6.1.3). The spectral makeup of the radiant flux remains virtually unchanged on passage through silver layers, which are thus said to be gray. In contrast, the absorption of the measuring beam in color emulsions is strongly dependent on wavelength. With monochromatic light of wavelength λ, the measured spectral density is symbolized by D_λ and the spectral transmittance by τ_λ (the reflectance by ρ_λ). With light having a known spectral distribution Φ_λ and a detector having a known sensitivity s_λ, the *integral density* D_i measured over a range of wavelengths is given by

$$D_\mathrm{i} = \log \frac{\int_0^\infty \Phi_\lambda e_\lambda s_\lambda \mathrm{d}\lambda}{\int_0^\infty \Phi_\lambda e_\lambda s_\lambda \tau_\lambda \mathrm{d}\lambda}$$

where Φ_λ is the incident flux at wavelength λ, e_λ is the relative spectral efficiency of the entire optical system, s_λ is the relative spectral sensitivity of the detector, and τ_λ is the transmittance of the specimen at wavelength λ.

The scattering power of color emulsions can generally be neglected in transmission measurements, and so the geometry is not critical. The same does not, however, hold for color materials viewed in reflected light (see Section 6.1.3).

Integral color densities are found by measurements on the complete assembly of emulsions in color materials. For routine measurements, optical filters in the colors blue, green, and red (corresponding to the three subtractive colors yellow, magenta, and cyan) are placed ahead of the material in the beam and the filter densities are determined. Films and papers whose images are examined with the human eye must be evaluated by measuring the *visual optical density*; this is performed with a detector whose relative spectral sensitivity matches that of the eye and with a beam of light whose spectrum corresponds to a temperature of 3000 K, i.e., tungsten light. In the case of negative films, the detector characteristic is matched to the spectral sensitivity of the print material. The density measured in this way is called the *printing density.*

Analytical densities are the color densities of individual emulsion layers. In practice, direct measurement of analytical density is possible only for the manufacturer of the photographic material. The spectral analytical density is defined by analogy with the spectral density (see above). In the assessment of the color-reproducing ability of a photographic material, the colors of the individual layers are combined into a hue with a neutral gray appearance, whose density is measured as the visual optical density. Figure 58 shows the integral (total) spectral density (curve a) for a specimen with a visual optical density of 1.0. The spectral analytical densities of the three layers are plotted as curves b, c, and d. The three dye layers are superimposed to form a gray tone "metameric with true gray." The spectral absorptions of individual dyes are not restricted to the regions of their "defined" colors, because there may be large secondary absorptions at nearby wavelengths; i.e., major color densities are distinguished from minor color densities. In the example of Figure 58, the measured integral density in the major absorption region of the yellow layer (436 nm) is due to three contributions: the analytical major color density (64 %), the analytical minor color density of the magenta layer (19 %), and that of the cyan layer (17 %). The minor color densities degrade color reproduction through chroma losses and hue shifts.

The analytical major color densities of the individual dyes that add to give the visual gray density are called "equivalent neutral densities." If the integral densities are measured in the region of the major color densities of the three layers with appropriate narrow- cut filters, both the analytical densities and the equivalent neutral densities can be calculated from the results [359], [34].

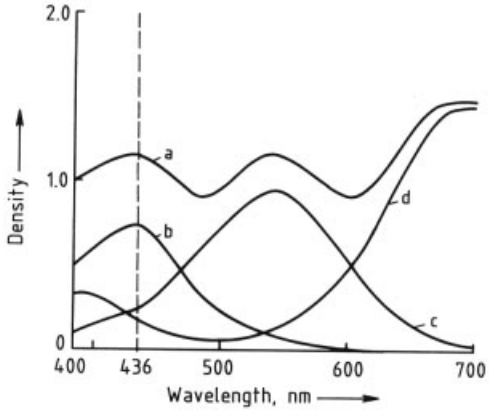

Figure 58. Spectral density of a color film at a neutral density of 1.0
a) Total spectral density of all dyes; b) Spectral analytical density of the yellow layer; c) Spectral analytical density of the magenta layer; d) Spectral analytical density of the cyan layer

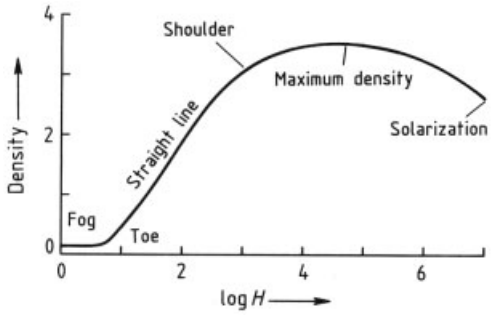

Figure 59. Sensitometric density curve for negative film (H = exposure in lux · s)

6.1.2. Characteristic Curve and Evaluation

If the optical density is plotted versus the logarithm of exposure $H = E \cdot t$ (in lux · s), the characteristic curve is obtained. This curve goes back to Hurter and Driffield [358] and remains the most important way of presenting sensitometric results. A characteristic curve always has a typical shape. For a photographic material that yields a *negative image* of the object represented (Fig. 59), the curve begins at low (not yet effective) exposures, showing a density called the fog density. Then comes the toe (underexposed region). The next part of the curve is nearly a straight line (normally exposed region). The shoulder represents the region of overexposure and is followed by the maximum density. As exposure increases further, the density may fall off again. This portion of the curve represents solarization. For a photographic material that yields a *direct image* of the object (e.g., a reversal film or an instant-picture material; see Fig. 60), a high density corresponds to the region of low "sub-toe" exposure, and the characteristic curve is inverted relative to that of the negative material.

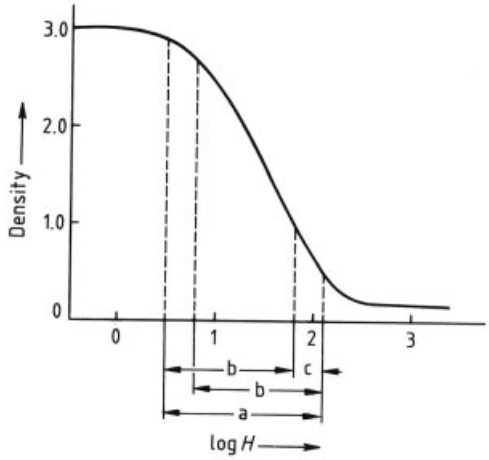

Figure 60. Exposure range of a reversal film
a) Maximum exposure range; b) Exposure range of normal objects; c) $\Delta \log H = 0.3$, that is $\pm 1/2$ stop

In practical sensitometry, the density curves are described in terms of parameters that convey information about the photographic properties. The most important is the exposure necessary for a stated density; this is equivalent to the position of the whole curve relative to the exposure axis. It determines the *sensitivity,* whose measurement is described in detail in Section 6.1.3. The slope of the characteristic curve [$dD/d(\log H)$] is a measure of how relative exposure differences are transformed into density differences or density contrasts. The slope is at a maximum on the straight-line portion of the curve. This maximum value is called the gamma (γ) and is a measure of gradation. In emulsion types that have a more S-shaped characteristic with no straight-line portion, the gamma is replaced by the slope of the secant passing through two points significant for image formation (this is sometimes called beta). The gradation of a photographic material increases with the development time, among other factors. Curves illustrating this relationship are called contrast index curves.

Depending on the numerical gradation value, materials in practice are said to be soft-, flat-, or weak-working ($\gamma \approx 0.4$) on the one hand or steep- or hard-working on the other. "Normal" gradation values are 0.6 – 0.8 for negative films and 1.5 – 2.5 for papers.

Fog, gradation, and maximum density define the range of exposure values that can be reproduced as differentiable densities. The brightness range of a scene is the ratio of the lowest to the highest luminance measured in the image plane of the camera. It is the ratio of luminous flux between the darkest and the brightest part of the scene (called object luminance ratio by GOLDBERG [37]), corrected for the contrast-reducing effect of camera flare. According to studies by JONES and CONDIT [360], landscapes in full sunlight have a luminance ratio between 1 : 10 and 1 : 40; portraits with black and white clothing, from 1 : 30 to 1 : 50; interiors with bright windows, from 1 : 100 to 1 : 1000; and printed illustrations, from 1 : 4 to 1 : 10.

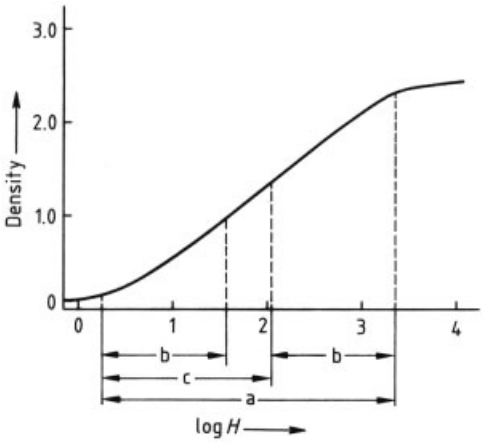

Figure 61. Exposure range of a negative film
a) Maximum exposure range; b) Exposure range of normal objects; c) $\Delta \log H = 1.8$, i. e., ± 3 stops

The luminance ratio is specified in logarithmic units on the exposure axis of the characteristic curve. An "average" object corresponds to 1.3 log units (1 : 20). A photographic material offers some freedom, depending on the shape of the characteristic curve, in the selection of the mean exposure (exposure latitude) so as to bring a given luminance ratio inside the useful range of the density curve. In the example of Figure 61, the maximum useful exposure range of a B & W negative film is ca. 3.1 log units. The exposure latitude is calculated as follows: maximum exposure range − exposure range of normal objects; i.e., 3.1 − 1.3 = 1.8, i.e., ± 0.9 log units or ± 3 stops. As Figure 60 shows, the exposure latitude of a B & W reversal film is only ± 0.5 stop. The maximum exposure range and thus the exposure latitude are still smaller for papers on account of the lower maximum density.

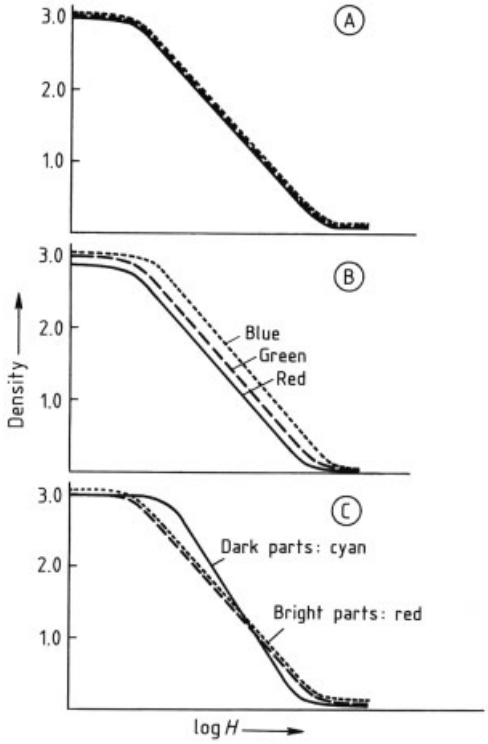

Figure 62. Characteristic curves of a color reversal film A) Neutral; B) Speed mismatch, yellow-red- cast; C) Higher red contrast (density)

The sensitometric behavior of color photographic materials is represented by three color-density curves corresponding to the three primary colors. The parameters listed in the preceding section also apply to color materials, but now it is necessary to consider the interaction of the three dye layers; hence, there are three times as many parameters. As already pointed out, a color material under given exposure conditions should reproduce gray object colors as a total density that appears gray to the eye or in the positive material. If the nominal "equivalent neutral" filter densities are known, then the position of

the color-density curves indicates whether this condition is satisfied. For clarity, it has been assumed in Figure 62 A that parallel color-density curves represent a neutral-gray density curve. A departure from this nominal shape corresponds to a color imbalance (cast) in the material. For example, if the color densities in the magenta and the cyan are relatively low, the image will display a yellow-red cast (Fig. 62 B). If the gradations of the three color curves do not match, the color imbalance may differ in the underexposed and overexposed parts of the curves, and indeed as a rule the casts are colorimetrically opposite. If the cyan curve has too high a slope (Fig. 62 C), the image will show a red cast at low densities (highlights) but a cyan cast at high densities (shadows). The film "tilts" from red highlights to cyan shadows.

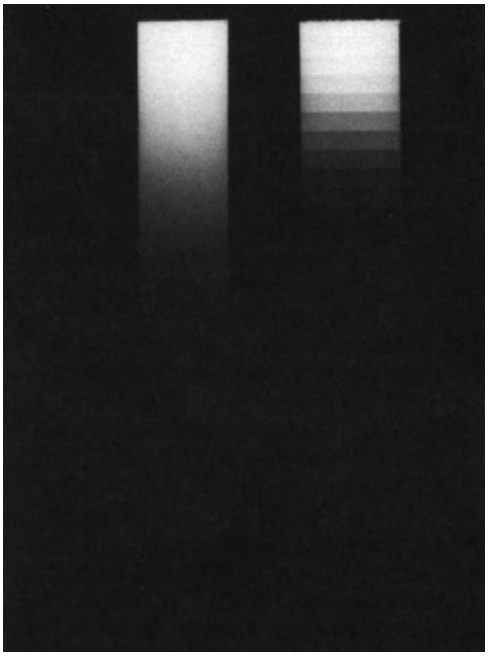

Figure 63. Filters
A) Continuous wedge; B) Step wedge, density difference = 0.15

6.1.3. Measurement of Density

The material to be tested is first exposed in a *sensitometer* with regularly increasing quantities of light. Either the time can be held constant and the illuminance varied (*intensity-scale densitometer*) or the illuminance can be held constant and the exposure time varied (*time-scale sensitometer*). Because the reciprocity law does not hold in photography with visible light (Schwarzschild effect), the two methods do not give equivalent results. Most sensitometers now in use are of the intensity-scale type. The light is attenuated by neutral-gray filters; such a filter consists of a glass flat with a silver, graphite, or dye layer. Filters of two types are used: discrete patches with constant increments of density (step tablet or step wedge with density increment 0.1, 0.15, or 0.3; see Fig. 63 A) or a wedge with continuously increasing density (0.1, 0.15, or 0.3 density difference per centimeter of length; see Fig. 63 B). Density increments of 0.1, 0.15, and 0.3 correspond to increases in exposure by factors of 1.26, 1.41, and 2. The exposure time is usually set by means of a falling-plate shutter or a slot moving at constant speed in front of the gray wedge. The sensitometer light source must be arranged in such a way so as to illuminate the wedge uniformly, and must be calibrated and constant in color temperature and intensity.

The exposed "sensitometer strip" is then carefully processed in the prescribed baths. The reproducibility of development is a crucial factor in the precision of the whole process.

Instruments for measuring optical density are called *densitometers.* The basic elements of such a device are a light source as near to a point source as possible, an aperture for the object being measured, a photoelectric detector (photocell or photomultiplier), a readout for the (measured) density, and a filter (several in the case of color materials) serving to match the spectral distribution of the measuring light to the detector sensitivity. The conditions for the standard determination of the optical density of transparent layers are set forth in ISO 5/3 [361]. For the measurement of diffuse density in B & W films, the specimen must be illuminated with parallel light and the entire half-space behind the specimen must be measured with an Ulbricht sphere or a diffuser between the specimen and the detector.

According to the reversibility principle of ray optics, this arrangement can be changed to diffuse illumination and measurement in parallel light.

To measure the color density of *color films,* filters matched to the three principal absorptions of the dye layers are used. Such filters

include interference filters for wavelengths of 435.8, 546.1, and 643.8 nm, or the Kodak Status A, M, or D filters, depending on the type of photographic materials being tested (see below).

Kodak Status A Reversal film, print films
Kodak Status M Color negative film
Kodak Status D Reflection densities, paper

Visual optical densities or printing densities are determined in accordance with ISO 5/3 and 5/5 [361] respectively, which establish numerical conditions relating to the color temperature of the lamp and filter versus the wavelength.

In *reflected-light densitometers,* a definite geometry is required for the measurement of reflected density, because of surface scattering effects [362].

Densitometer designs fall into two groups: direct-reading and comparison instruments. *Direct-reading instruments* use analog or digital instruments to measure directly the output current of the photoelectric detector. The logarithmic density is obtained from the reading with specimen in relation to the reading with no specimen in the beam. Light sources for direct-reading densitometers must be highly constant and detectors must be highly reproducible, but instruments on the market today easily meet these standards.

In *comparison instruments,* the detector serves as a null indicator. Behind the filter set, the measuring beam is split into a sample beam and a reference beam. The path for the reference beam contains a variable attenuating element (a continuous density wedge that can be repositioned), so that the intensity of the reference beam can be matched to that of the sample beam. The wedge is calibrated in density units. Such instruments are relatively independent of fluctuations in the lamp current, but do depend on the constancy of the gray wedge in the reference beam path and the precision with which it is adjusted. In a simple design, which has a restricted measuring range and limited precision (but is still in use), the eye serves as a nulling instrument; the luminances in the two beam paths are matched.

Recording densitometers directly record the density curves. By storing a large number of density values, modern instruments also allow computer calculation and printing of the parameters of the characteristic curve.

6.1.4. Sensitivity and Evaluation

A properly exposed photograph is the central concern of sensitometry. In practice, the mean luminance of an object is determined with an exposure meter; i.e., an indicating instrument with a photoelectric detector. The exposure meter must be calibrated so as to ensure a correct exposure – within the exposure latitude – for any type of film, given an average object luminance ratio, provided the camera (aperture and exposure time) is set to the indicated "exposure" value. The parameter "light sensitivity" must correspond to the meter reading [363], [364]. Modern cameras generally have built-in exposure meters.

According to DIN 4512 (ISO 5800, ISO 2240, ISO, ANSI PH 2) the light sensitivity (DIN 12 speed) is the quantity of exposure to which a photographic emulsion must be subjected in order to form an image. From the sensitometric standpoint, it is a measure of the location of the characteristic curve relative to the exposure axis. The numerical value is the exposure that results in a stated optical density. The smaller this exposure value, the higher the sensitivity of the material; accordingly, the sensitivity number (speed) is inversely proportional to the exposure H. The German standard employs a logarithmic scale defined by the formula

$$\mathrm{DIN} = 10 \cdot \log \frac{H_0}{H_\mathrm{R}}$$

The constant H_0 is given, and H_R is determined from the characteristic curve (see below). A difference of 1 in the speed, for example between 17 DIN and 18 DIN, represents a sensitivity change by a factor of $10^{0.1} = 1.26$; a DIN number 3 higher corresponds to double the sensitivity.

The international ISO/ASA standard is arithmetic:

$$\mathrm{ISO/ASA} = \frac{1 \cdot H_0}{1.26\, H_\mathrm{R}}$$

ISO/ASA speeds are converted to DIN speeds by the relation

$$\mathrm{DIN} = 1 + 10 \cdot \log\,(\mathrm{ISO})$$

Manufacturers have recently been citing both speeds together, e.g., ISO 100/21°. A comparison of the two sensitivity scales follows:

DIN	6	7	8	9	10	11	12	13	14
ISO	3	4	5	6	8	10	12	16	20
DIN	15	16	17	18	19	20	21	22	23
ISO	25	32	40	50	63	80	100	125	160
DIN	24	25	26	27	28	29	30	31	32
ISO	200	250	320	400	500	630	800	1000	1250
DIN	33	34	35	36					
ISO	1600	2000	2500	3200					

Establishing the speed in accordance with the standard involves using precisely specified conditions of exposure, development, and evaluation. The material is exposed in the sensitometer in daylight (color temperature 5500 K) or with tungsten light (3400 or 3200 K), depending on the intended use. The radiation incident on the emulsion is adjusted for the spectral transmittance of a representative camera objective. Matching of the incident radiation to the prescribed spectral distribution is performed with liquid or glass filters [365].

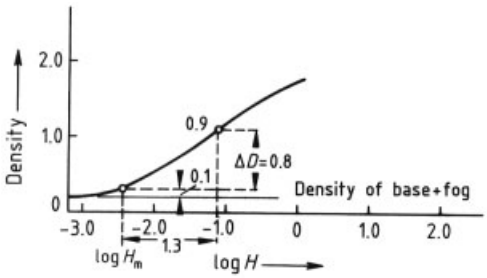

Figure 64. Method for determining speed of B & W films

B & W negative films used to be developed in a solution whose composition was laid down in ISO 6 DIS 1989 (DIN 4512/1) [366]. The development time was chosen to give quite a definite form of characteristic curve such that the difference between two specified density values (on the vertical axis) was $\Delta D = 0.8$. (The densities were 0.1 and 0.9 above base plus fog.) This difference corresponded to an exposure difference of 1.3 log units (20 : 1) on the horizontal axis (see Fig. 64). For such a characteristic curve, the sensitivity point log H_m is given by the abscissa at the density of 0.1 above base plus fog. The sensitivity is then $S = 0.8/H_m$, where H_m is in lux · s. This standard, based on a special developer, is now being reworked into ISO 6 DIS 1989. Presumably the above definitions will remain in use; the manufacturer can select any process (developer), but must then include statement such as, "Material PAN X has a speed of ISO 100/21 ° for developer X at 20 °C and 4 min development time."

Color negative films must be exposed as described above and processed in accordance with the manufacturer's instructions. The printing density is measured as set forth in ANSI PH 2.27-1981 [366], [367]. No definite gradation value is specified. Since three density curves exist, there are three sensitivity points, each of which refers to a color density of 0.15 above base plus fog (Fig. 65). The sensitivity is calculated from the exposure value of the green-sensitive layer H_G and that of the less sensitive of the other two layers; i.e., the slowest layer:

$$H_m = \sqrt{H_G \cdot H_{sl}}$$

The ISO (ASA) speed is then given as

$$s = \sqrt{2}/H_m$$

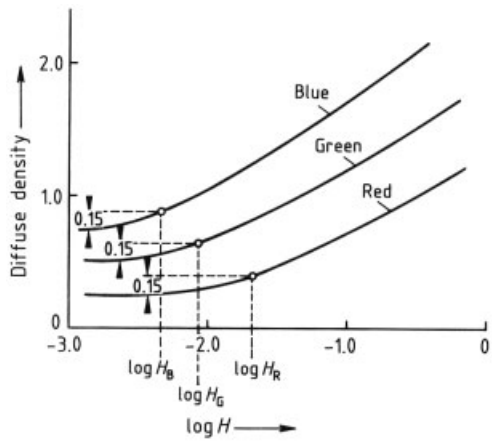

Figure 65. Method for determining speed of color negative films (filters: Status M)

Color Reversal Films. A completely different criterion is used in determining the sensitivity of color reversal films. After an exposure specified in the standard and processing in accordance with the manufacturer's instructions, the visual density is measured as provided in ANSI PH 2.21 [368] or ISO 2240 (DIN 4512/4) [366]. The corresponding sensitivity point H_m on the exposure axis is found as described in ISO 2240 (DIN 4512/4) [366], [368] (see Fig. 66). The

procedure involves taking the geometric mean of H_T, the exposure for a density of 0.2 above base plus fog, and H_S, the exposure corresponding to the point where the tangent from the density point 0.2 above base plus fog intersects the characteristic curve. If the second of these points has an optical density > 2.0, the sensitivity point exposure H_m is calculated as follows:

$$H_m = (H_S \times H_T)^{1/2}$$

The speed (ASA) is then calculated as $S = 10/H_m$.

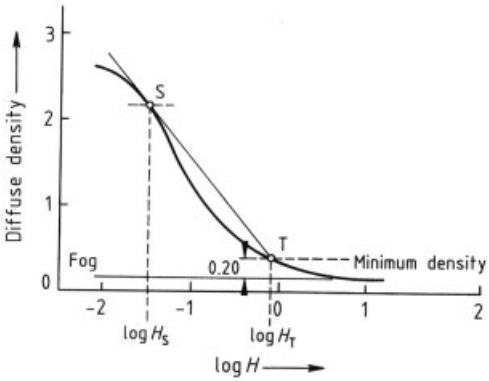

Figure 66. Method for determining speed of color reversal films (filter: visual)

The method based on an average object luminance ratio does not function for scenes having deviant luminance ratios. It is therefore recommended to give about one more stop of exposure when using reversal films during cloudy weather, that is, when the scene luminance ratio is reduced (bad-weather correction).

Measurement of Spectral Sensitivity. Photographic emulsion sensitivity depends on wavelength. The spectral sensitivity is the reciprocal of the radiant energy in a narrow range of wavelengths that produces a stated optical density.

The instrument used for measuring the spectral sensitivity is called a *spectrosensitometer*. It includes a monochromator and a device for the well-defined stepwise variation and measurement of the radiant energy per wavelength range. Another method, simpler but also less accurate, is to use a prism or grating spectrograph with a gray wedge inserted between the slit and the specimen. The sensitivity of the emulsion governs the density versus wavelength profile obtained after development (Fig. 67), which can be evaluated with a densitometer. To allow a quantitative comparison of the results, the energy distribution of the lamp spectrum must be taken into account and the measured spectral sensitivity must be converted to a "constant-energy spectrum."

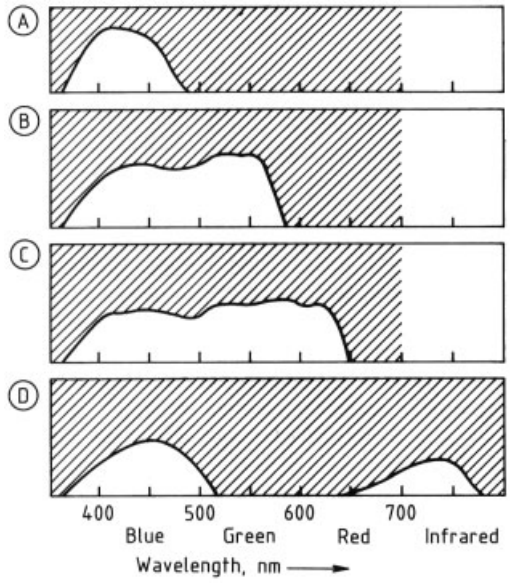

Figure 67. Spectral sensitivity of different sensitized materials
A) No optical sensitization: blue sensitive; B) Orthochromatic: blue and green sensitive; C) Panchromatic: blue, green, and red sensitive; D) Blue and infrared sensitive

Pure silver halide emulsions have no sensitivity except in the ultraviolet and blue regions (Fig. 67 A). The region of sensitivity expands into the long-wavelength infrared only after spectral sensitization with dyes. The following terminology has come into use for describing the spectral sensitivity of emulsions. *Orthochromatic*: additional sensitivity in the green to about 560 nm (Fig. 67 B). *Panchromatic*: intrinsic sensitivity plus additional sensitivity in the green and red to ca. 680 nm (Fig. 67 C).

The purpose of *sensitizing B & W films* for pictorial use is that the brightness values of surface colors occurring in nature should be reproduced as they are seen by the eye. Compromises are necessary for the sake of greater light sensitivity. The use of a yellow filter reduces the blue component of the spectral sensitivity and often results in a better visual impression.

In *color photography*, each emulsion layer is sensitized for a certain color. For example, the yellow layer is sensitized for blue light; the blue component of the light is subsequently removed by a yellow filter for the panchromatically sensitized emulsions lying beneath. The sensitivities of the emulsions in a color film are matched to the spectral energy distribution of the light illuminating the scene. Daylight contains more blue and less red than the light from an incandescent tungsten lamp. Accordingly, in daylight reversal film the blue layer is less sensitive than the green and the red layer is more sensitive than in films with tungsten light. Departures from the specified light color, such as daylight exposures with the sun low in the sky, may produce a marked color imbalance in the image, as can the use of photographic objectives with "atypical" spectral transmittances. Reducing the short- and long-wavelength limits of the spectral sensitivity makes it possible to diminish the effects on the photograph due to differences in the type of light (universal sensitization). Similarly, the use of a filter on the camera can render a film sensitized for one type of light (e.g., tungsten) usable with another (e.g., daylight).

6.1.5. Sensitometry of Printing

Negative films represent an intermediate stage in the creation of an image. They are printed onto a suitable positive or print material (a positive film or paper). The brightness distribution of an original object is reproduced (tonal reproduction) in the best possible way if the image contrasts match the luminance levels in the scene, that is, if the print has a gradation value of 1 relative to the original. Hence the corresponding portions of the characteristic curves for the negative and the positive should be related as follows (Goldberg condition):

$$\left[\frac{\mathrm{d}D}{\mathrm{d}\log H}\right]_{\mathrm{neg}} \cdot \left[\frac{\mathrm{d}D}{\mathrm{d}\log H}\right]_{\mathrm{pos}} = 1$$

The Goldberg condition [34], [37] can be exactly satisfied only in the straight-line portion of the characteristic curve. Exact calculations should, moreover, take into account the effect of flare light in the camera and printing apparatus. The result of a printing operation (printing curve) can be plotted by using the characteristic curves of the negative and positive materials [369]. To meet the Goldberg condition, the flat gradation of negative films must be offset by a markedly higher gradation of the print material. Hence papers must have gradation values > 1, which means that the maximum exposure latitude (called printing latitude) is limited. For positive film, the latitude is ca. 3.0 log units. Papers must be exposed so precisely that detail is still present in the highlights and shadows even when the nonlinear parts of the characteristic curve are included. B & W papers are therefore available with a range of gradations or "hardnesses" so that the range of negative density differences can be accommodated. The light sensitivity of the positive material is not critical unless economic factors such as printing time must be considered.

The measurement of speed in *B & W materials* is governed by an ISO/ANSI standard [370]; (see Fig. 68); H_{m} is the exposure that produces a net density of 0.6. The speed is then calculated as $S = 1000/H_{\mathrm{m}}$. The printing range is given by $R = \log H_{\mathrm{S}} - \log H_{\mathrm{T}}$, where H_{S} is the exposure at 90 % of maximum net density and H_{T} is the exposure at 0.04 above base and fog density. At present no standard exists for *color papers*.

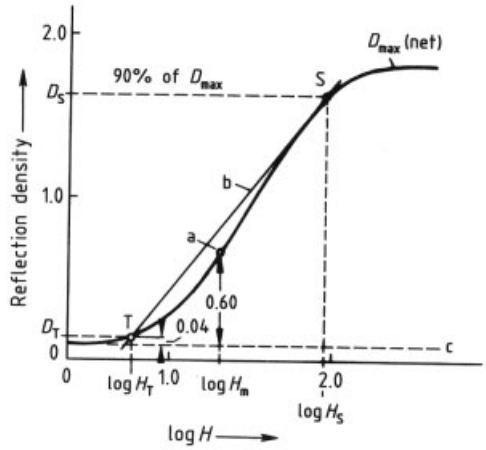

Figure 68. Determination of ISO speed and range of printing for black and white continuous-tone papers, Filter: visual
a) Speed point; b) Average gradient; c) Base plus fog density

In printing from color negatives, the dyes simultaneously act as filters. For example, the blue component in the printing light is modulated by the yellow color density of the negative before

exposing the blue-sensitive emulsion in the positive material. Aside from the secondary color densities of the individual layers, a given dye layer in the negative prints to the same dye layer in the positive. Under these idealized conditions, the Goldberg condition holds separately for the yellow, magenta, and cyan layers. If a pair of layers fails to satisfy the condition, a color error occurs in the print. For example, if the magenta in the negative is too steep relative to the other layers, the positive will show green highlights and magenta shadows.

Theoretically, a gradation departure in one layer of the negative material could be compensated by a departure in the opposite sense in the corresponding positive layer. In the example just described, a flat magenta in the print material would offset the error. In practice, the approach is not used. Negative and positive materials are separately made "equivalent neutral", so that products from different manufacturers can be used interchangeably.

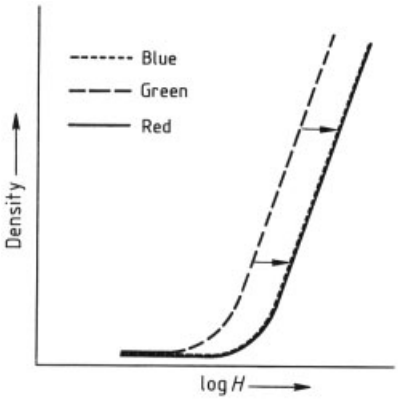

Figure 69. Effect of subtractive filter on print material

An important advantage in color copying is that color imbalances in the positive and the negative can be corrected by adjusting the color of the printing light. In the subtractive process, this is performed by inserting appropriately colored printing filters. Figure 69 shows how a magenta cast in the positive is remedied by a magenta filter, which reduces the exposure of the excessively fast green-sensitive emulsion. The use of the filter shifts the magenta curve to the right. An imbalance in the negative due to sensitivity mismatches between the emulsions cannot, however, be exactly filtered out in the fog region, so that a color tilt can occur in the dark portions of the image. Figure 70 illustrates the case of a cyan emulsion (a) that is less sensitive than the magenta and yellow emulsions (b). A cyan filter shifts curve (a) upward by the filter density (d) to c. Only the straight-line portion can be brought to coincide with curve (b). The fog plus base part of the curve, on the other hand, takes on a cast of the filter color.

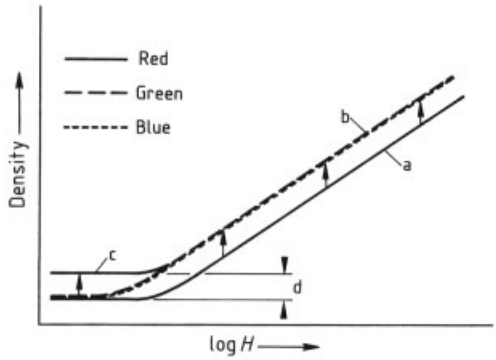

Figure 70. Unbalanced color negative film
a) Cyan emulsion; b) Magenta and yellow emulsions; c) Fog plus base part; d) Filter density

Because it is fairly easy to compensate for color imbalances in printing, color negative films need not be strictly balanced for the color temperature of the light on the scene. For this reason, most color negative films now on the market are daylight types, which can be used under any lighting conditions that occur in practice, given proper exposure. In motion-picture work, however, standards are higher and filters must be used on the camera under unusual lighting conditions. The use of camera and printing filters can remedy only overall casts. Mixed light sources, for example, sunlight and tungsten light, must therefore be avoided. Up to this point, the effect of minor color densities of the dye layers on the sensitometric result has not been taken into account. In the two-step negative – positive process, and especially in multistep processes, secondary absorptions of the negative dyes lead to inadequate dye formation or dye formation in the wrong layer in the positive material. Because the positive will show further reproduction errors, the multistep process will result in more hue shifts and chroma losses than the reversal process. Negative films do, however, offer the possibility of improving color reproduc-

tion through the use of "color masks." These are emulsion layers in the color of the minor color density. Their sensitometric action consists in compensating for the image produced in the "minor" color by an opposite (inverted) gradation, thus producing a uniform fog density. This fog in the minor color does not appear in the print, so that the chroma of the major color is improved. Modern color negative films therefore contain masking dyes incorporated in the emulsion layers (see Section 5.2.2.3).

The high yellow and magenta fog level in color negative films, which largely determines the position of the printing density curves (Fig. 65), is accounted for by the incorporated masking dyes.

6.1.6. Sensitometry of Interimage Effects
(see also Section 2.3.4)

In addition to the use of masks to minimize the minor densities (see Section 6.1.5) of the dyes, color reproduction can also be improved through interimage effects. Figure 71 illustrates the principle. This effect can be strongly promoted by DIR compounds. Such compounds flatten the gradation given a neutral exposure: roughly equal amounts of DIR couplers are released in all the emulsion layers on development, thus inhibiting the yellow, magenta, and cyan layers equally. The film remains neutral. If, on the other hand, the film is exposed in red light, only small amounts of DIR couplers appear in the yellow and magenta layers, while the quantity liberated in the cyan layers grows larger and larger as the exposure increases, so that these compounds can diffuse into adjacent layers. This means that the cyan layers are not as strongly inhibited as the other layers. The gradation becomes steeper, and in this way the color reproduction is improved. IN 1987 there is no ISO standard on the sensitometry of interimage effects [34–36]. The gamma is referred to a value of 1 for neutral exposure; if the selective red exposure gives 40 % interimage effects in the cyan, then $\gamma = 1.4$. Additive neutral exposure (red+green+blue) generates less cyan dye than the same amount of red exposure only. Thus the layers influence one another (interlayer effects). If, for example, the film is exposed with green light (wedge) and uniformly with red light, the action of the DIR couplers suppresses the cyan densities with increasing green exposure. This corresponds to masking and improves the purity of the colors; (see Fig. 72). The cyan density has a negative gradient: $\gamma = -0.2$. The same procedure can be followed for any other combination of layers, yielding an interimage-effect matrix as shown below. In this example it was assumed that all the layers (yellow, magenta, and cyan) have 40 % interimage effect, 20 % of which is due to each of the other layers.

$$M = \begin{array}{c} \text{(yellow)} \quad \text{(magenta)} \quad \text{(cyan)} \\ \begin{pmatrix} 1.4 & -0.2 & -0.2 \\ -0.2 & 1.4 & -0.2 \\ -0.2 & -0.2 & 1.4 \end{pmatrix} \end{array} \begin{array}{l} \\ \text{(blue)} \quad \text{(blue)} \\ \text{(green)} \quad \text{(green)} \\ \text{(red)} \quad \text{(red)} \end{array}$$

6.2. Tonal Reproduction and Image Structure (see also Section 2.4.1) [2], [4],[19], [38]

6.2.1. Tonal Reproduction

The task of the photographic emulsion is to record and store optical information. The information may comprise a portrait, a landscape, or a microfilm of a technical drawing. In any case, the intensity pattern generated in the plane of the film must be transferred without distortion. The emulsion can achieve this only to a limited extent because it is not a structureless image storage medium, but consists of a gelatin matrix in which light-sensitive silver halide crystals reside in a statistical distribution. The quality of image transfer is essentially described by the characteristic curves, the granularity, and the sharpness of the material employed.

Characteristic Curve. The characteristic curve (see also Section 6.1.2) is the most important graph used to assess the emulsion as an information medium. It shows how the optical density D (-log τ) resulting after development depends on the exposure H. In accordance with the scattering power of the developed emulsion, two kinds of transparency are distinguished: the diffuse transparency $\tau *$ and the transparency τ'' measured in parallel light (Callier coefficient $C = D''/D *$) [374].

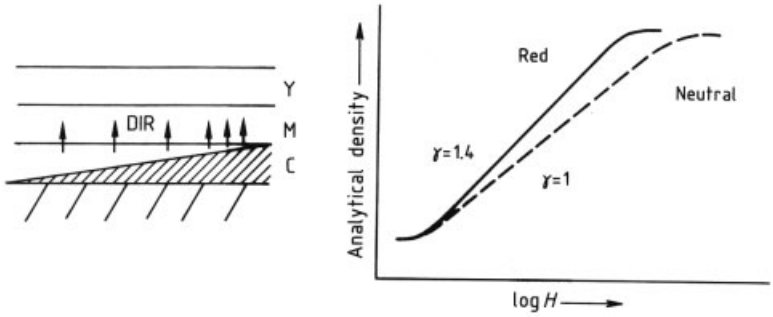

Figure 71. Interimage effect, red and neutral exposure
Y=Yellow layer; M=Magenta layer; C=Cyan layer

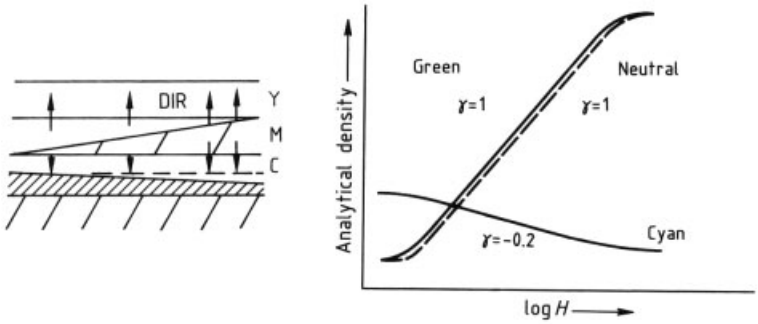

Figure 72. Interimage effect, green exposure and constant red exposure
Y=Yellow layer; M=Magenta layer; C=Cyan layer

The exposure H is usually plotted logarithmically on the horizontal axis (see Fig. 59). The shape of the characteristic curve strongly depends on the emulsion, its laydown on the support, and the development.

The increase in density with exposure generally results from an increasing number of developed grains per unit area. According to a simple but frequently applicable model (due to NUTTING)

$$D = 0.434 \cdot n_f \cdot \frac{a}{f}$$

where n_f is a number of grains per area f and a is the mean effective grain cross section [19], [375].

At the toe of the characteristic curve (Fig. 59), that is, near H_0, a limited number of large, sensitive grains are exposed. On the linear portion of the curve, the number of developable medium-sized and small grains becomes greater and greater. The onset of saturation, finally, indicates that greater exposures will not make any more grains developable. The fact that a certain exposure renders a certain fraction of the silver halide grains in an emulsion developable implies that each individual grain must have its own critical threshold number of photons, beyond which it becomes developable. This minimum number of photons that must be absorbed may depend on grain size or may vary within a single grain-size class (sensitivity distribution).

How many photons are actually absorbed by the individual grain at a given exposure partly depends on the grain size, which is generally not uniform in an emulsion (grain-size distribution), and partly on the location of the grain in the emulsion layer (depth distribution of exposure). Finally, the statistical distribution of incident photons has the consequence that grains of equal size and depth have different numbers of photons available for absorption (Poisson distribution of absorbed photons).

If these distribution functions are known, the number of developed grains can be calculated as a function of the exposure, if some simplifying assumptions are made. The characteristic

curve can be determined by Nutting's method if it is further taken into account that the density contributions of the developed grains depend on their size.

Detailed treatments of the theory of optical density are found in [2], [4], [19], [32], [34]. For the practical meaning of the characteristic curve and, in particular, the extension of this concept to color photography, see Section 6.1. The gradation in the finished photograph should be chosen so as to produce the most natural impression possible of the object imaged.

6.2.2. Granularity

If the developed emulsion is examined after uniform areal exposure, the structure appears homogeneous at first but becomes irregular as the enlargement is increased. This impression, which gives the examiner an impression of reduced quality, is called *graininess*. *Granularity*, in contrast, refers to the same phenomenon in the form of objectively measured local density fluctuations, such as can be recorded for the same specimen by a microphotometer with a sufficiently small measuring aperture. Both these terms relate to an important property of the photographic system as an information-transfer medium. This property results from the sensitivity distribution of the silver halide grains and the random arrangement and sizes of the developed silver grains [4], [19], [34], [38]. The grains themselves are too small to be detected at common magnifications, but the number of grains per unit volume fluctuates for the reasons stated above.

Graininess. Graininess comparisons between photographic materials are generally carried out by exposing a series of distinct, sufficiently large ($> 1\ \mathrm{cm}^2$) areas of each specimen, and then developing. On the test strips, negatives of equal density are then selected, enlarged by a factor of 12, examined, and compared.

Another common procedure is based on increasing the enlargement until the graininess becomes disturbing. The enlargement is then reduced until it is just impossible to perceive the random structure. The reciprocal of this enlargement limit is a measure of the graininess [376].

An essential condition for the success of such test methods is consistent processing of all materials; special care should be taken that all test strips are developed to the same gradation and that comparisons are performed only in regions of equal optical density, since graininess and its detectability depend on development and density.

The drawbacks of such methods are the rather laborious procedure of specimen preparation and the lack of objectivity in the numerical values or rankings obtained.

Granularity. Granularity comparisons between photographic materials can be done with a variety of objective methods. These techniques and the quantities obtained are quite similar to noise analysis techniques in electronics, for granularity is the counterpart of noise in electronic transmission channels. The starting point for all methods is the transformation of grain noise into electrical noise by photoelectric scanning in a microphotometer (Fig. 73). The scanned track x may be circular or straight, since granularity in photographic emulsions is isotropic. The position coordinate x is linked to the time coordinate t of the electrical signal through the scanning rate v [371].

The most common way of plotting granularity is to record the rms value σ (see below) as a function of optical density D at a constant aperture area (Fig. 74). The curve is derived from the density variations $D(x)$ recorded when scanning the longest possible track X:

$$\sigma_{D,f} = \left[\frac{1}{X} \int_0^X \left(D(x) - \bar{D} \right)^2 \mathrm{d}x \right]^{\frac{1}{2}}$$

where $\bar{D}$ is the mean density, defined as

$$\bar{D} = \frac{1}{X} \int_0^X D(x)\, \mathrm{d}x$$

The rms value is defined similarly to the effective value of an AC voltage, so that in the extended sense $(\mathrm{rms})^2$ can be regarded as a measure of the power of the grain noise. The dependence of the rms value on the selected aperture area f is represented by the subscript f. The two equations above are mathematically exact only as $X \to \infty$. In practice, the rms value is also determined as

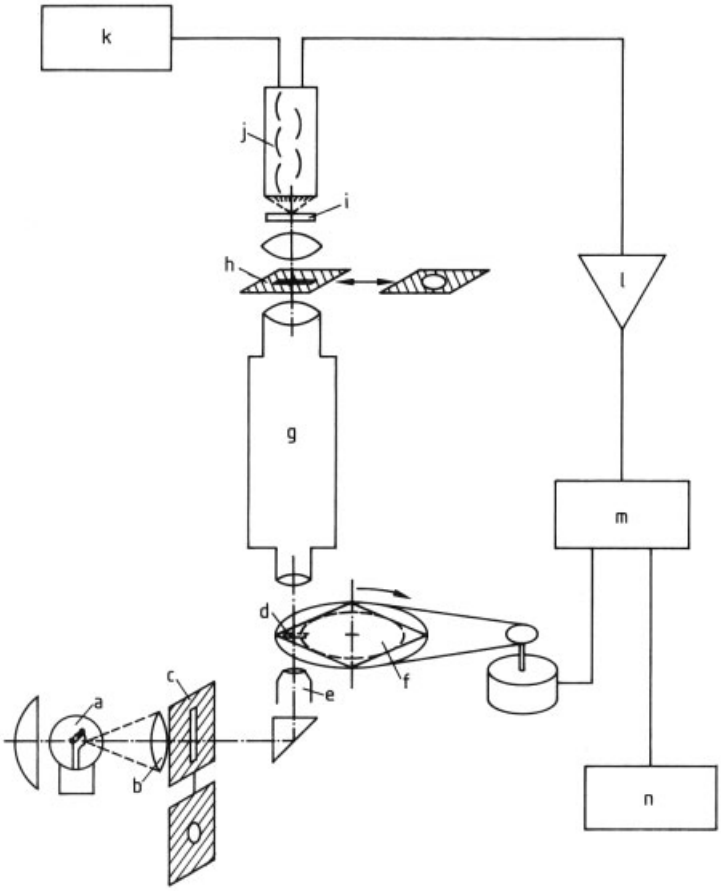

Figure 73. Optical system of a microdensitometer including electronic equipment
a) Lamp; b) Illuminating condenser; c) Field slit; d) Image of field slit or aperture; e) Illuminating microscope; f) Sample holder (rotating or translating) movable stage; g) Collector microscope; h) Scanning slit or aperture; i) Diffuser; j) Photomultiplier; k) High voltage power supply; l) Amplifier (plus log- converter and ADC); m) Computer (stage motion control and data analysis); n) Plotter, printer

$$\sigma_{D,f} = \left[\frac{1}{n}\sum_{i=1}^{n}(D_i - D)^2\right]^{\frac{1}{2}},$$

$$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i, n \geq 1000$$

Many formulas exist for calculating the granularity from the fundamental elements of the emulsion [377]. The simplest model leads to the Siedentopf formula [372]:

$$\sigma^2_{D,f} = 0.434 \cdot \bar{D} \cdot \frac{a}{f}$$

This equation holds only if the area f of the measuring aperture is much larger than the mean effective grain cross section a and if the grains have a statistically independent distribution. These conditions are fairly well satisfied for B & W films and apertures larger than 10 µm in diameter [373], [378], so that the Selwyn granularity G_D, defined as

$$G_D = \sigma_{D,f} \cdot \sqrt{f}$$

is constant here too provided the density D is constant [379]. The fact that G_D is independent of aperture was the reason that the Selwyn granularity was preferred as an emulsion parameter for a long time. But modern color materials are far from satisfying this condition, and so the rms values are now preferred.

Unfortunately the measuring apertures used in reported studies have different diameters (12 – 50 µm); hence, the rms values obtained cannot be compared easily. The diameter should be selected such that the microphotometer – aperture system simulates the viewing conditions for visual assessment as well as the optical properties of the human eye (48 µm diameter and 12 x enlargement). The rms value and the graininess rating could then be correlated.

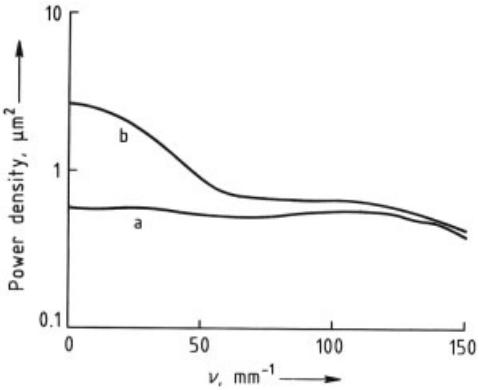

Figure 74. RMS-values as a function of density for
a) Color reversal film, 100 ASA; b) Color negative film, 100 ASA
Filter = visual

Frequency Analysis. The most complete and objective information on granularity comes from a frequency analysis of the grain noise. One way of performing such an analysis is to proceed as in the rms measurement but to feed the electronic noise from the microphotometer output to a filter having the narrowest possible passband $\Delta\nu*$ about the center frequency ν. By varying $\nu*$ while keeping the bandwidth $\Delta\nu*$ constant, a noise value, rms($\nu*$, $\Delta\nu*$), is obtained for each frequency band $\nu*$, $\Delta\nu*$. This noise value depends on the mean density $\bar{D}$ of the specimen and the area and shape of the measuring aperture. Since the electrical frequency $\nu*$ is generally equal to the product of scanning rate v and spatial frequency ν, the noise power per unit spatial frequency $\sigma^2{}_{D,f}\,(\nu, \Delta\nu)$ can be plotted versus the spatial frequency ν. The resulting curve is called the power-density spectrum $N_{D,a}\,(\nu)$ of the grain noise. The mathematically exact definition of this spectrum is

$$N_{D,f}\,(\nu) = \lim_{\Delta\nu\to 0} \frac{\sigma^2_{D,f}\,(\nu,\Delta\nu)}{\Delta\nu}$$

Because the granularity contributions of the individual spatial-frequency intervals simply add to give the total granularity, the following applies:

$$\sigma^2_{D,f} = \int_0^\infty N_{D,f}\,(\nu)\,\mathrm{d}\nu$$

The power-density spectrum $N_{D,f}\,(\nu)$ depends on the mean density $\bar{D}$ and on the area and shape of the measuring aperture. If constraints due to the microphotometer optics are neglected, then

$$N_{D,f}\,(\nu) = N_D\,(\nu) \cdot \left|M_f\,(\nu_x, \nu_y)\right|^2$$

Here $M_f\,(\nu_x, \nu_y)$ is the modulation transfer function associated with the measuring aperture f. In general, it depends on the aperture dimensions and its orientation is relative to the scanning direction (see Section 6.2.3). $N_D\,(\nu)$ is the true power-density spectrum (Wiener spectrum) of the photographic emulsion, which can be derived from the above equation with little mathematical difficulty, if a very long, narrow slit is used as the aperture and the scanning direction is perpendicular to the long axis of the slit.

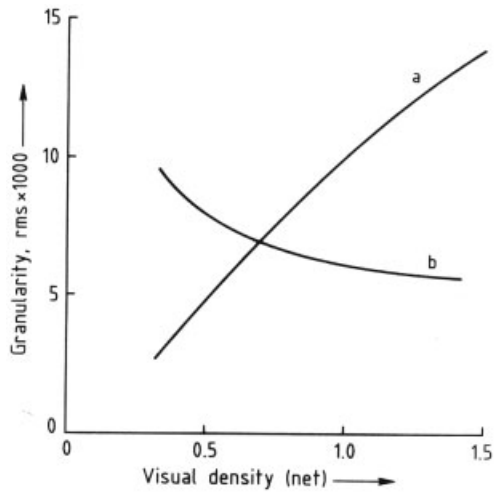

Figure 75. Power spectral density of a B & W film
a) Exposure by light; b) Exposure by X-rays

In *B & W films* exposed with visible light, all spatial frequencies make roughly equal contributions to the Wiener spectrum (Fig. 75, curve a). For this "white-noise" case, $N_D\,(\nu) = N_D\,(0)$, so that the Selwyn granularity is $G^2{}_D = N_D\,(0)$.

For exposure with X-rays, grain agglomerates are formed because an X-ray quantum renders a large number of grains developable. Corresponding to this frequently-encountered coarse structure in position space are low spatial frequencies in frequency space, so that the Wiener spectrum shows a marked exaggeration at low spatial frequencies (Fig. 75, curve b).

Detailed mathematical analyses (Fourier transforms, autocorrelation, etc.) and methods of measuring the Wiener spectrum are discussed in [19], [34], [38], [378].

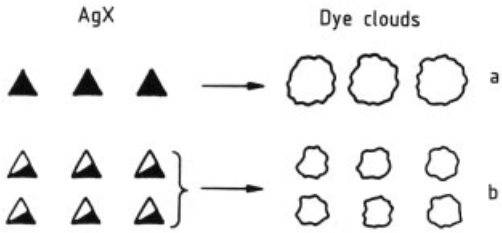

Figure 76. Improvement of granularity by DIR- couplers
a) Without DIR- couplers; b) With DIR- couplers

The techniques just described for measuring graininess and granularity in B & W materials can also be applied to *color materials.* Selecting appropriate color filters makes it possible to study the granularity and the Wiener spectrum of individual emulsion layers. The graininess rating and the measured granularity correlate well if the contributions to the total noise power are weighted with the sensitivity characteristic of the eye: magenta 60 %, cyan 30 %, and yellow 10 %. For color reversal materials, a filter is employed that combines with the spectral sensitivity of the photomultiplier to reproduce the sensitivity of the eye. Both interpretation and theory relating to the measured color granularity are far more complicated than in the case of B & W granularity [380], because the dye clouds are transparent, unlike the developed silver grains, and because the relationships between the silver halide grain and the dye cloud are indirect (with the diffusible developer oxidation product as intermediate stage). In addition, the three color-generating layers may interact.

The Nutting and Siedentopf formulas imply that the granularity decreases as the number of silver grains per unit area increases. Since 1972, DIR couplers have been used to improve (i.e., decrease) the granularity [381] (Fig. 76; see also Section 5.2.2.3). During color development, inhibitors are released and prevent the complete reduction of the silver halide grain. The result is that smaller dye clouds are formed. In order to achieve the same density, more grains are needed; this implies a lower granularity than in a film without DIR couplers.

6.2.3. Sharpness

If the developed emulsion is examined after an exposure containing small details, a sufficient enlargement will reveal that the boundaries between image regions differing in luminescence have become inexact and that details in the original of less than some definite size are no longer discernible in the image. If defects in image recording (e.g., blurring or incorrect focusing) are not present, the subjective impression gained from such a visual examination characterizes the sharpness of the developed emulsion. Avoiding errors in the assessment depends strongly on separating the effects of sharpness and granularity on the quality of image transfer, since grain noise also makes it impossible to distinguish details when the signal – noise ratio is not sufficiently high. The sharpness is the third important transfer property of photographic materials. It results from the behavior of the emulsion elements (light scattering) and from processes occurring in development (color formation, edge effects) [2], [4], [19], [34], [38]; see also Section 2.4.1.

"Visual sharpness" comparisons between photographic materials are carried out by essentially the same methods known from grain investigations, except that large-area exposure is replaced by suitable test objects.

An objective, numerical measure of sharpness, which is relatively easy to measure and correlates fairly well with the visual assessment, is obtained from the density curve of an exposed edge. After exposure and development, the edge pattern in the material is placed in the beam of a mirophotometer and scanned. In order to suppress the granularity, an aperture in the form of a slit, narrow in the x direction and as long as possible in the direction parallel to the edge, is used. The density curve $D(x)$ is illustrated in Figure 77. The density begins to increase even before x_0, that is, in the geometrical shadow re-

gion of the edge, but does not reach its maximum value until considerably beyond x_0. The edge is reproduced in smeared-out form. The reason for this loss of sharpness is light scattering, for the undeveloped emulsion can be regarded, from the optical standpoint, as a plane-parallel layer of a medium containing haze. The scattering inhomogeneities are the grains, whose refractive index (ca. 2.2) differs markedly from that of the surrounding gelatin matrix (1.5). During exposure, an illuminance distribution occurs at the emulsion surface (external exposure; see Fig. 77). Inside the layer, this distribution is modified by scattering, absorption, and reflection (internal exposure). As a result of scattering, light diffuses into the geometrical shadow of the edge (light spread), degrading the sharpness. Light quanta can also reach the back of the film base, where they are reflected, again reducing the sharpness (halation). In commercial materials, however, this effect is largely blocked by an antihalation layer.

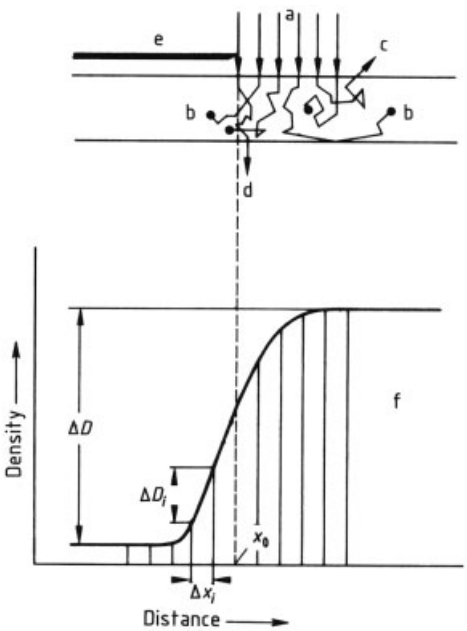

Figure 77. Light scattering in a photographic emulsion
a) Photons (exposure); b) Absorbed photons; c) Re-emitted photons; d) Transmitted photons; e) Knife edge exposure; f) Density distribution after development

The numerical sharpness value is derived from the measured density profile. The position coordinate is subdivided into small, equal intervals Δx_i, and the mean square slope $(\bar{D})^2$ in the edge is calculated with the formula

$$(\bar{D})^2 = \frac{1}{n}\sum_{i=1}^{n}(\Delta D_i/\Delta X_i)^2$$

If the density change ΔD in a comparison is not constant, the acutance, AC defined as

$$AC = \frac{(\bar{D})^2}{\Delta D}$$

correlates rather well with the visual assessment [382].

Modulation Transfer Function. The introduction of information theory and its associated frequency-domain representation into optics [19], [38] has made the modulation transfer function (MTF) the most important descriptor of the sharpness of photographic materials. This function $M(\nu)$ states quite generally how the modulation transfer of a locally sinusoidal illuminance distribution $E\left(\vec{r}\right)$ having spatial frequency $\vec{\nu}$ (sine-wave pattern) is changed in the imaging system. In principal, any optical system – a microphotometer, a circular aperture, or a measuring slit – has an MTF [2], [19], [34], [38]. In order to measure the MTF of a photographic material, sine-wave patterns having various spatial frequencies are exposed onto the material. The spatial distribution of the external illuminance is then

$$E(x) = E_0\left[1 + m(\nu)\cdot\sin(2\pi\nu x)\right]$$

where E_0 is the mean intensity and $m(\nu)$ is the input modulation. Light scattering in the emulsion causes the modulation of the illuminance distribution to change as a function of the spatial frequency. The higher the spatial frequency, the smaller the amplitude of the internal illuminance $E(x)$:

$$E(x) = E_0\left[1 + m'(\nu)\cdot\sin(2\pi\nu x)\right]$$

The ratio of the modulations of the internal and external exposures at a given spatial frequency ν gives the modulation transfer factor $M(\nu)$:

$$M(\nu) = \frac{m'(\nu)}{m(\nu)}$$

The functional relation between $M(\nu)$ and the spatial frequency is the MTF. The procedure

is illustrated in Figure 78 [383]. Any spatial exposure distribution $H\ \vec{\nu} = E\left(\vec{\nu}\right) \cdot t$ can be uniquely expanded in two-dimensional sine and cosine functions varying in amplitude and phase (Fourier transform). Thus, if the MTF is known, the photographic image of an object can be calculated in the following way: the MTF is multiplied by the Fourier transform (spatial-frequency spectrum) of the external exposure, and then the inverse Fourier transform is applied to the product (the spatial-frequency spectrum of the internal exposure). The result is the spatial distribution of the internal exposure. Conversely, any object can be used for measuring the MTF if the spatial-frequency spectra of the external and internal exposure are known. Thus, in addition to the classical method based on a sine-wave pattern [384], [385] (where the transfer of discrete frequencies is measured) other methods exist. These alternative methods employ square-wave patterns [386], knife-edges [387], [388], and noise patterns [389] of selected bandwidth. The last method is quite similar to the way the frequency response of an electronic transmission channel is measured.

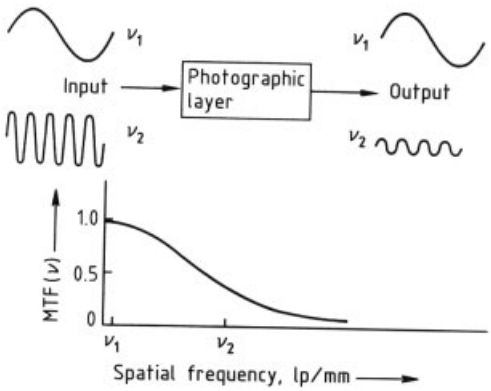

Figure 78. Modulation transfer function, principle
Unit of spatial frequency is line pairs per millimeter (lp/mm)

All the methods are similar in that the internal exposure is calculated from the characteristic curve derived from the microphotometric measurement of the transparency or density distribution in the developed image. In fact, the density distribution is the crucial output signal. This operation is necessary because, as shown by the characteristic curve, the relationship between the internal exposure and the developed optical density is nonlinear. It follows that, if a sinusoidal exposure distribution is transferred to the corresponding optical density, the sinusoidal form generally is not retained; the output contains distortions analogous to overtones in electronics. This effect can be suppressed by making the modulation of the external exposure sufficiently small. In the region under consideration, the characteristic curve is linear to a good approximation, so that the concept of the MTF still applies to the transfer from the illuminance distribution to the density distribution.

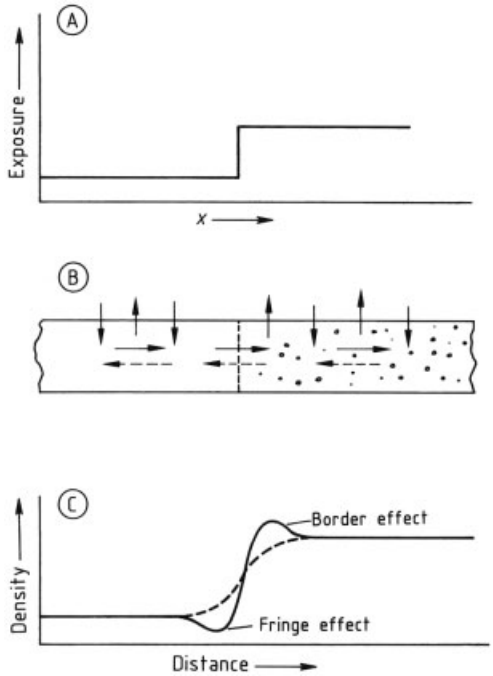

Figure 79. Schematic diagrams of adjacency effect
A) Exposure, knife edge; B) Diffusion of reaction products; C) Contrast without (dashed line) and with adjacency effect (solid line)

A completely different cause gives rise to a second nonlinearity between exposure and density. This is due to *diffusion effects in development* and therefore depends strongly on the developer composition and the agitation of the specimen while being developed. At the boundary of a uniformly exposed region, which consumes corresponding amounts of developer, fresh developer can diffuse into the edge of the region from the surrounding zone of lower exposure. In this way, the optical density inside the edge becomes higher than the density aver-

aged over the region. On the other hand, oxidation products from the region of high density can diffuse into the adjacent region, restraining the formation of density, so that regions of low density show a fringe of reduced density in front of the transition to darker regions (Fig. 79). The strength of this *adjacency* or *Eberhard effect* depends on the density difference as well as the absolute density [381], [390], [391]. It may be so pronounced that, e.g., in the region of a slit, a density is developed such that a back- calculation with the characteristic curve would show an internal exposure greater than what was actually present (see Fig. 80). This means that small details are reproduced with higher contrast. This effect is greatly intensified by DIR couplers, in modern color negative films. In general, adjacency effects work to the benefit of the perceived sharpness. These effects are also described by the MTF, which can take on values greater than unity (see Fig. 81). Where an edge is exaggerated, however, the MTF behaves similarly to the Eberhard effect; hence, to characterize the sharpness of the photographic material – which is governed by light scattering and the edge effect – it would be necessary to determine a family of MTF curves with the mean density and input modulation as parameters. When two materials are being compared, however, it is usually sufficient to measure an MTF for each at equal mean densities and not too great an input modulation, provided the two specimens are nearly equal in sensitometry [381].

Taking the constraints cited into consideration, the modulation transfer function of an entire imaging system containing the photographic emulsion can be determined (MTF system). With such an MTF system, parameters of the image sharpness can be obtained or – what is often more important – the object can be reconstructed from the existing image. The basis for the calculations is the multiplication theorem [19], [34], [38], which states that the MTF of an imaging chain equals the product of the MTFs of the links in the chain, provided all transfer the signal in a linear way. The assessment of the chain camera – film – projector – screen is given as an example. For each spatial frequency, the individual modulation transfer factors are multiplied with allowance for the enlargements between steps. To get a good correlation to the visual impression, the resulting MTF is multiplied by the MTF of the eye, again with allowance for the way the viewing conditions affect the spatial-frequency axis. The area under the overall MTF gives a measure of sharpness (CMT acutance) [392], [393].

Besides the sharpness measures obtained in a similar way [392], [393], a figure of merit based on the MTF (sharpness) and the Wiener spectrum (granularity) has been proposed. The calculations lead to good agreement with the visual impression [394].

The concepts and methods for sharpness measurement can be extended from B & W to color materials; the MTFs of the individual layers can be measured through the choice of suitable color filters. The MTFs not only represent the variation in external exposure due to light scattering and egde effect, but also include the effect of color formation on sharpness.

In color development, the light quantum absorbed in the silver halide is made detectable by means of developed density not at the point where it was absorbed, but where the developer oxidation product diffusing away from the grain reacts with the dye coupler. This process degrades the sharpness, so that the MTF of a given color layer, by the multiplication theorem, is the product of three factors: MTF(S) (= scattered light), MTF(C) (= color development), and MTF(E) (= edge effect). In the multilayer color film, the scattered-light component becomes more distinct from layer to layer (in the irradiation direction). In this way, edges and stripe structures may be edged in the color of the next lower layer.

For the same reasons as cited for granularity, the magenta MTF has the strongest influence on the visual impression of sharpness.

Monte Carlo calculations of the MTF on modern high-performance computers have become a powerful tool for optimizing complicated multilayer structures with respect to sharpness. This is important because average emulsions have grain diameters of the order of the wavelength of light, and this means that scattering takes place in the range governed by Mie scattering [394]. Mie scattering theory is particularly complicated, whereas the Monte Carlo technique allows the calculation of light scattering in structures, no matter how complex they are. Monte Carlo calculations are used to obtain

point or line spread functions and information on the sensitivty. The MTF is then derived from the line spread function by application of the Fourier transform [395–398].

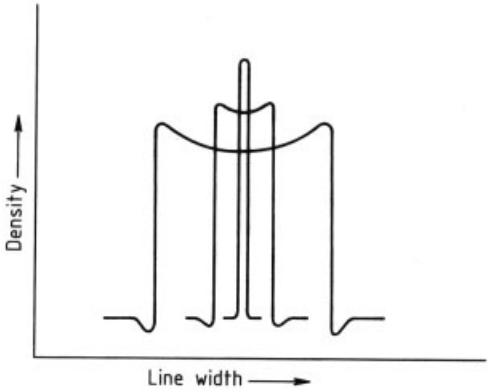

Figure 80. Microdensitometer traces of photographic line images of various width, all exposed with equal intensities

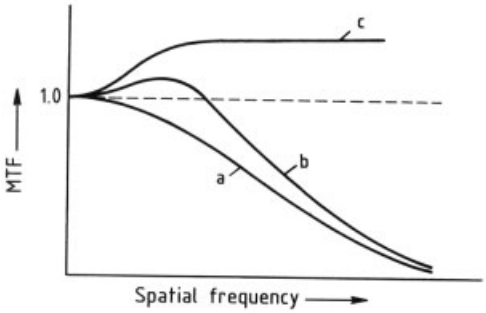

Figure 81. Influence of the adjacency effect on the modulation transfer function
a) MTF without adjacency effect; b) MTF with adjacency effect; c) MTF of the chemical spread function

6.2.4. Resolving Power [19], [38], [399]

The resolving power (RP) is the reciprocal of the distance between two details (e.g., bars in line grating) when the images of these details on the photographic emulsion are just distinguishable from each other. Despite is shortcomings, the resolving power continues to be used as a measure of the transfer properties of photographic materials. this is understandable because the measurement can be done without costly apparatus.

The resolving power is determined by exposing line gratings with increasing spatial frequency onto the emulsion, often in varying intensity (Fig. 82). Accordingly, to MTF the contrast between light and dark bars must become smaller and smaller as the spatial frequency increases. In addition to this light-scattering effect, the eye experiences more and more interference due to graininess in the microscopic examination, so that ultimately the density variation due to the exposure "drowns" in the graininess.

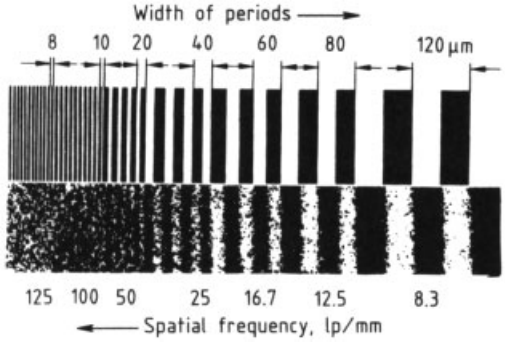

Figure 82. Line gratings for evaluating the resolving power

It follows that the resolving power depends on both the MTF and the granularity. In a given photographic material, the same RP can result from a good MTF and high graininess or from a poor MTF and low graininess. In the same way, two emulsions having equal modulation transfer functions can have completely different resolving powers. What is more, the sharpness of an image in practice is governed not just by the film, but also by the camera, whose MTF may be such that the film cannot display its full resolving power. In such cases, a high numerical RP has no predictive value and it is only important that the MTF has its maximum values in the range of spatial frequencies that the camera transfers well. But the major drawback of the resolving power is that the procedure is not objective [399].

6.3. Color Reproduction [35], [400], [401]

An important goal of color photography is to produce colors in the image as near as possible to those that the human eye saw in the object. Subjective factors, however, play a major role in the optimization of color reproduction (e.g., in the case of skin tone). Electromagnetic radiation

preceived as "light" or "color" is radiation incident on the retina of the human eye and having wavelengths between about 400 and 700 nm.

The retina contains three types of *stimulus centers* (cones) for color vision; these differ in spectral sensitivity. The impression of color is produced by the intensity of the three stimuli transmitted to the brain.

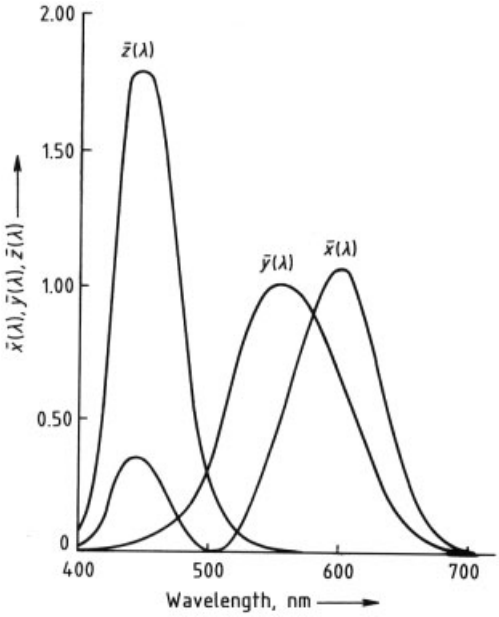

Figure 83. CIE tristimulus functions according to CIE Publication no. 15.1 and DIN 5033

In *colorimetry*, the vision process is simulated by metrological apparatus. The conventional spectral sensitivity distributions for the eye model are embodied in the spectral tristimulus values set forth internationally by the Commission Internationale de l'Eclairage (CIE); see also → Dyes, General Survey, Chap. 3. The radiation reflected or transmitted by an object – the product of the incident radiation $E(\lambda)$ (normally standard illuminant A for tungsten or D_{65} for daylight) and the spectral reflectance $R(\lambda)$ or transmittance $T(\lambda)$ – is integrated with the spectral tristimulus functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, $\bar{z}(\lambda)$ in order to obtain the CIE tristimulus values X, Y, Z where 380 and 775 nm are the lower and upper integration limits and k is a normalization factor:

$$X = k \int_{380}^{775} \bar{x}(\lambda) \cdot E(\lambda) \cdot R(\lambda)\, \mathrm{d}\lambda$$

$$Y = k \int_{380}^{775} \bar{y}(\lambda) \cdot E(\lambda) \cdot R(\lambda)\, \mathrm{d}\lambda$$

$$Z = k \int_{380}^{775} \bar{z}(\lambda) \cdot E(\lambda) \cdot R(\lambda)\, \mathrm{d}\lambda$$

The three tristimulus values provide a measure of the three stimuli sent to the brain (Fig. 83). The set of all color locations (X, Y, Z) forms a three-dimensional color space (CIE – XYZ system). This color space is not visually equispaced; an equal geometrical distance between two color locations almost never corresponds to an equal difference as visually perceived.

The CIE – XYZ system can be remapped into a visually equispaced system according to several international standard methods. After the remapping, an equal (Euclidean) distance between any two color locations corresponds to an approximately equal visually perceived difference. The CIE – $La*b*$ 1976 formula (CIE Publication no. 15.2) is given as an example in what follows:

$$\text{Luminance}\, L = 116 \cdot \left(\frac{Y}{Y_0}\right)^{1/3} - 16$$

Red – green portion $a*$

$$a^* = 500 \cdot \left[\left(\frac{X}{X_o}\right)^{1/3} - \left(\frac{Y}{Y_0}\right)^{1/3}\right]$$

Yellow – blue portion $b*$

$$b^* = 200 \cdot \left[\left(\frac{Y}{Y_o}\right)^{1/3} - \left(\frac{Z}{Z_0}\right)^{1/3}\right]$$

where X, Y, Z are the tristimulus values for the sample and X_o, Y_o, Z_o are the tristimulus values for the reference illuminant.

Color differences according to the CIE – $La*b*$ formula (CIE Publication no. 15.2, DIN 6174) are as follows: Given two specimens with L_1, a_1*, b_1* and L_2, a_2*, b_2*, then the CIE – $La*b*$ color difference is

$$\Delta E = \left[(L_2 - L_1)^2 + (a_2^* - a_1^*)^2 + (b_2^* - b_1^*)^2\right]^{1/2}$$

This can be broken up into lightness, chroma, and hue components, whose squares sum to the square of ΔE

$$\Delta E^2 = \Delta L^2 + \Delta C^2 + \Delta H^2$$

The luminance difference is

$$L = L_2 - L_1$$

The chroma difference is

$$\Delta C = \left(a_2^{*2} + b_2^{*2}\right)^{1/2} - \left(a_1^{*2} + b_1^{*2}\right)^{1/2}$$

The hue difference is

$$\Delta H = \left(\Delta E^2 - \Delta L^2 - \Delta C^2\right)^{1/2}$$

The three-dimensional character of the color space accordingly dictates three criteria for the characterization of color differences: luminance, chroma, and hue.

In the color plane (L = const), color differences are represented as shown in Figure 84.

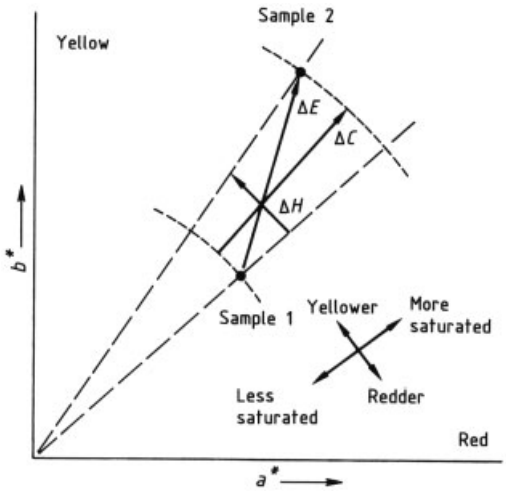

Figure 84. CIE-$La*b*$ color plane for L = const

Colorimetry can also be used to represent the color reproduction of photographic materials in the CIE – $La*b*$ color space. By way of example, Figure 85 shows the color rendition in a color negative – positive system. The measurements were performed on prints made on Agfa Type 8 color paper from two color negative films: the older XRi-100 (film 1) and the new XRG-100 (film 2). The object photographed was the Agfa color chart with its six colors (blue, green, red, yellow, magenta, cyan). Correct conditions were maintained as to exposure and filtration during both exposure and printing (see Section 6.1). For ease of visualization, the color locations are joined by lines to form a color hexagon in the color plane (color gamut). The lightnesses in the original and the prints are not considered here. Modern color films thus yield better results as to chroma and hue, partly because of strong and optimized interimage effects (see also Section 6.1.6).

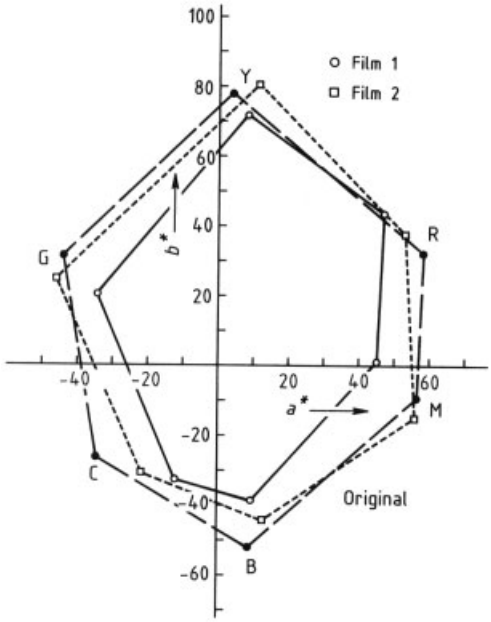

Figure 85. Color planes of two films
B = blue; C = cyan; G = green; M = magenta; R = red; Y = yellow

6.4. Stability of Images [34]

6.4.1. Introduction

The discovery of photography dates back to approxi mately 1826. In historical terms, this is a short span of time. Nonetheless, photographic images from this period are of great documentary value. Therefore, great efforts are made to store and preserve such *existing images* in the best way. Another problem is how to ensure the longest possible life for *newly made photographic documents* that are judged worthy of preservation.

Three fundamentally different factors influence the life of a photograph [402]:

1) Properties of the material itself
2) Processing
3) Storage conditions

In the case of historic photo documents, only the storage conditions can be controlled; the processing and choice of material were established long ago.

Optimal storage conditions can generally be described as follows: the rates of possible destructive chemical reactions are decreased by lowering the temperature and the access of harmful reactants is hindered. The most important reactant is water in the form of atmospheric humidity.

Processing must be carried out so that no residues of processing chemicals that could act as harmful reactants over a long time interval are left in the photographs. Special care should be taken with regard to thiosulfate from the fixing bath.

Materials differ in lifetime, chiefly in that silver images are much more stable and nonsilver materials, such as color photographs, are significantly less stable. Although even a gelatin – silver emulsion can be attacked, for example by bacteria or fungi more concern should be focused on emulsion support materials such as glass, cellulose acetate, cellulose nitrate, polyester, paper, and polyethylene. In color, diazo, and vesicular materials, (diazo and vesicular materials, see → Imaging Technology, Chap. 2.3.) on the other hand, the image substances are more important, especially the dyes [40].

Experience – mostly unpleasant – with stability is reflected in the provisions set forth in *standards* for storage conditions [403–405], processing [406], stability and materials for archival purposes [407–413]. These standards are checklists for the prevention of known harmful effects [40].

6.4.2. Storage

Stability specifications for photographic material are classed according to use. For example, documents of *archival quality* (to be preserved forever or for as long as it is at all possible), materials with *long-term stability* (up to about 100 years), those with *medium-term stability* (to 10 years), and *working copies for everyday use* are distinguished [402].

Working copies do not come within the scope of the standards. They are continuously used up, chiefly by mechanical damage due to handling. Familiar examples are exhibition prints of motion pictures and frequently-used microfilms, which become unusable because of scratches and dirt.

The intermediate levels of *medium- and long-term stability* are defined by laws that require stated periods of time for which materials must be stored (e.g., microfilmed business documents or medical X-ray photographs of significance in insurance cases). The periods differ from country to country.

The reason for these different storage periods is that protection of documents against harmful effects, which is necessary for long-term stability, is incompatible with ready access to the stored information. A feasible compromise between quick access and permanence must be found, in keeping with the intended use of the material. Corresponding to the three materials specifications are two *quality levels* of storage conditions: long-term storage for archival materials and long-term (to 100 years) stable documents, where access is difficult, and medium-term storage for continuing use, with the possibility of rapid access. This level of storage, less tightly restricted, is to ensure the survival of documents for the medium term, up to about 10 years.

The provisions of the standards aim to keep harmful volatile substances away from the archived photographic materials. Such substances include SO_2, H_2S, nitrogen oxides, H_2O_2 and organic peroxides, vapors of volatile acids, atmospheric humidity, and atmospheric oxygen [39], [40]. For this reason, requirements on filing enclosures, storage housings, and storage areas are specified. Separate provisions govern air conditioning, safeguards against fire and water, and the handling and inspection of documents in the archive under an established statistical inspection plan [403–405], [407–409], [411].

Storage Areas. Storage areas for long-term storage differ from those for medium-term stor-

age in that they must be separate from all workrooms, such as those for inspection or processing of films. Provisions relating to fireproof building materials and safeguards against water damage due to leaky pipes or floods (but also due to fire-extinguishing sprinkler systems as well as condensation on ceilings and walls when the outdoor temperature is low) are the same for both types of rooms. Finally, vapors and reaction products (e.g., peroxides) of unsuitable ceiling, wall, and floor paints can endanger the stability of stored photographic documents [414].

Storage Housings. The same warnings as to paints and coatings apply to storage housings in the storage rooms, such as shelving, cabinets, file drawers, and racks [414]. Open shelves and racks are permissible only for archived objects in closed containers. Enclosed cabinets must have vent openings if the rooms are air- conditioned to maintain the recommended atomospheric conditions. Storage housings must allow air circulation on all sides without impairment of the protection they afford against fire and the action of firewater.

The *materials* used for storage facilities must be at least fire-resistant; for long-term storage they must be entirely nonflammable (DIN 4102 part 1, class A 1). Anodized aluminum or stainless steel is recommended. Nonstainless steel can be used only with good surface protection. Materials such as wood, pressed board, plywood, hard board and particle board should not be used, not only because they are flammable but also because they release active agents that can fade the photographic images over a period of time [415]. In particular, formaldehyde from glues used in wood products is considered suspect; it is thought to be oxidized by atmospheric oxygen through peroxide and hydroperoxide intermediates to formic acid forming hydrogen peroxides as byproduct. Hydrogen peroxide oxidizes the silver image, converting it to silver formate in the presence of formic acid, which is soluble and diffusible in water that is absorbed even by dry gelatin in equilibrium with the humidity of the air. The result is bleaching of the silver image and possibly, if traces of H_2S are present, a general increase in fog.

Filing Enclosures. The selection of materials for filing enclosures (i.e., folders, pouches, sleeves, boxes, envelopes, jackets, tubes, film cores and reels, cassettes, magazines, film cans) is based on the same principles as that of storage housings. Selection of suitable materials, however, is especially critical because they come into intimate contact with the images [416].

In particular, folders, sleeves, window mounts, slide mounts and aperture cards made of wood-fiber paper with acidic alum sizing should not be used. The autoxidation of lignin in such papers yields agents that can damage photographic images. The specification for suitable wood-free papers includes a content of > 87 wt % α- cellulose, a pH of 7.5 – 9.5 for contact with black-and-white films and a pH close to 7.0 for contact with color films, and an alkali reserve of at least 2 wt % (e.g., through the use of calcium carbonate as filler). Such papers must be free of wood or metal particles, waxes, plasticizers, and other fillers.

Adhesives and tapes used to close sleeves and envelopes must also be chemically stable and free from active decomposition products (acidic, oxidizing, or reducing). Glues based on photographic gelatin, poly(vinyl acetate), and cellulose esters are acceptable; rubber-based cements are not suitable because of autoxidation products. The same holds for printing inks, because experience shows that they often contain unknown ingredients which can cause the formation of microscopically small colored dots on microfilms [417]. This is the reason why B & W paper prints should not be stored in the same envelope with newspaper clippings, even though this might be desirable in terms of subject [418].

Suitable materials for *cores, spools*, and *cans* used to store films in roll form are noncorroding metals such as anodized aluminum and stainless steel. Nonstainless steel can be used only with good surface protection, such as enamel, tinning, plating, or coating, but only with lacquers that do not release decomposition products. Larger rolls, for film lengths of ≥ 150 m, should be stored flat, with the core axis vertical, so that the weight of the whole roll is not imposed on a small bearing area; thus adhesion due to the pressure can be avoided.

Plastics that are best for protective enclosures (e.g., spools, cans, envelopes, bags, jackets, and slide mounts), consist of the same material as the base of the photographic material (e.g., cellulose acetate, poly(ethylene terephthalate), or polyethylene). Other plastics may also be suitable, but practical experience is lacking in most cases.

Definitely unacceptable, on account of chemical instability, are chlorinated plastics, cellulose nitrate, and celluloid as well as plastics containing plasticizers, residual solvents, or peroxides, and foamed polyurethane plastics.

Environmental Conditions. Even in the absence of decomposable materials whose products damage the images, environmental conditions in storage have a strong effect on lifetime. In general, low temperatures and low relative humidities extend the storage life. For the sake of easier access, temperatures and relative humidities may be higher when a medium-term storage life in the 10-year range is needed rather than when true long-term storage with a target life of 100 years (or longer if possible) is desired.

At relative humidities $> 60\,\%$ or temperatures $> 20\,°C$, the danger of fungal or bacterial attack on the gelatin is possible. Even for *short periods of time*, the peak temperatures should not be allowed to go above $32\,°C$. For accessible, *medium-term storage* of color films with chromogenic development, temperatures $< 10\,°C$ are desirable; $< 20\,°C$ is acceptable for B & W films because of the greater stability of the image substance. For *long-term storage*, the relative humidity should be lower than $60\,\%$; the proper value differs somewhat from one material to another, but ca. $30\,\%$ R.H. can be taken as an average value for the storage of various materials in the same room. Again, the temperature for B & W materials should be $< 20\,°C$, but for color materials the limit should be set at $\leq 2\,°C$. These conditions can scarcely be maintained without air-conditioning in the storage room. A special warning is given with respect to the use of cool cellar rooms as storage areas. The relative humidity can reach almost $90\,\%$ if there is air exchange with the surroundings in hot summer weather. The difference in temperature between storage areas and workrooms necessitates a warm up period in the packaged condition in order to prevent condensation of water on archival photographs when they are used, processed, or inspected. Short-term cycling of temperature and relative humidity should be avoided during these operations as well as when the materials are in storage. Films for photogrammetry (i.e., extraction of geometrical data from stereophotographs, mostly used in cartography) and for photofabrication respresent a special case in this respect. In order to retain their dimensional stability, such films must have the same standard atmosphere ($23\,°C$ and $50\,\%$ R.H., as provided in ISO 554-1976) during storage and use so that adaptation processes do not take place. Climate-protected packaging is recommended for the shipping of such films.

Conditioning. Apart from films for photogrammetry and photofabrication, further protection of films that are seldom used can be afforded by conditioning at equilibrium with air at the lower limit of the recommended relative humidity, and sealing the film. Moisture-tight enclosures include cans for film rolls; such a can is sealed with adhesive tape and heat-sealed into a foil bag. For sheet materials, a double-bagging technique is used: two foil bags, each heat sealed, one inside the other. Packages sealed and protected against moisture in this way are stored with their contents at low temperatures. Before they are opened, they must be warmed up for several hours at room temperature in order to prevent moisture condensation on the archival material.

If suitable air- conditioned storage rooms are available, the sealed containers can be dispensed with. In this case, however, special care must be taken to provide for air purification. A slight excess pressure in the storage room will exclude particulate and gaseous contaminants. If the air is treated in scrubbers or absorbers continuous and careful maintenance of this equipment is required, as well as special care against fungal and bacterial attack. The latter is statistically the most common cause of damage to photographic archives.

A further advantage of suitable protective enclosures is that they protect against fire as well as the action of steam produced from firewater. Silver-based archival materials can survive for several hours in such insulated record containers at a temperature of $150\,°C$ without becoming unusable, provided the seal prevents the entry of firewater and steam. For this reason, stability against degradation and deformation in a four-hour test at $150\,°C$ is a requirement for such insulated record containers.

An important point in long-term storage is the inspection of the materials under a statistical inspection plan that provides for a fairly large amount of samples. A two-year interval is considered acceptable. Inspections should be

performed at shorter intervals when deviation in the temperature and relative humidity from the recommended ranges have occurred. The handling of materials requires special care.

Testing of materials for filing enclosures is carried out by keeping them in close contact with a special detector with a coating of colloidal silver under conditions of 86 % R.H. and 70 ± 2 °C for 15 days [416]. There must be no transfer of structure from the enclosure material to the detector and the mean fade value of the detector must not be greater than the control value. As a control test, the same procedure is performed with the detector in contact with paper having a pH between 7.0 and 7.7, e.g., Whatman no. 1 filter paper.

6.4.3. Processing

One effect from which photographic materials cannot be isolated, is that due to the chemicals used in processing. These are removed as far as possible by intensive final washing [419]. As the last processing bath used with B & W materials, the fixing bath is critical [406]. Residual thiosulfate (hypo) was identified as a cause of the instability of photographs soon after the discovery of fixable silver halide materials [420], [421]. The quality of processing, especially the final washing is checked on representative samples by determination of the residual thiosulfate in the freshly processed material [406].

A suitable method is the photometric determination of the concentration of methylene blue in an extraction solution. This dye is formed by reaction of sulfide sulfur from the thiosulfate ion and *N,N*-dimethyl-*p*-phenylenediamine under standardized conditions.

A less specific method is the photometric determination of the iodine consumption in the iodine-starch inclusion compound with which an extract of the sample material is reacted. Both tests are not very useful in the case of aged materials, because part of the residual thiosulfate originally present has already been converted to other products by air oxidation. Instead, the increase in density in regions of minimum density is evaluated in representative samples after treatment with excess $AgNO_3$ solution has produced a silver sulfide fog. In this method, conversion products such as polythionates and sulfide are determined along with thiosulfate; the threshold of detection is not, however, as low as in the photometric dye determination [406].

A complementary test involving treatment with excess sulfide solution provides information about residual, unfixed silver salts in the archival material [406].

The test for effectiveness of final washing will become more important if environmental considerations make it undesirable to extend the final washing (make it longer and more intensive, with greater water consumption) beyond what is actually necessary.

For silver images that are to be safeguarded in an archive for as long as possible under carefully controlled storage conditions, the thiosulfate content must be as low as possible. The level is limited by the thickness of the gelatin layer, which always retains small amounts of thiosulfate.

In 1986 studies showed that a very small residue of thiosulfate can actually enhance stability under storage conditions not corresponding to recommended archival quality [422–424]. Such conditions apply to areas where peroxides and hydrogen peroxide may be present. It is thought that sulfide is released by the thiosulfate, forming a protective coating of silver sulfide on the image silver and thus guarding it against peroxide attack. A sulfide toning of the finished image after processing has a similar stabilizing effect on the silver image [425], [426]. The introduction (from baths) of substances such as thiocyanate, aminoguanidine, etc. [427] that intercept peroxides, also improves stability. This is recommended, e.g., for photographs that will be placed in display windows in inner cities where automobile exhaust concentrations are high.

6.4.4. Materials

When amateur photography got under way (Eastman Kodak), and especially after the invention of cinematography, original exposures were made on films with cellulose nitrate as support. A substantial body of photodocuments, especially historic motion pictures and newsreels, still exists on this base.

Because of the inherent instability of the cellulose nitrate base, the outlook for the survival of these historical documents is not good. Even

without fires, in which case this material cannot be saved, the autocatalytic degradation of celluloid cannot be prevented indefinitely. The degradation product nitric acid also attacks the silver image in the emulsion. In most cases, however, rolls of this film agglomerate into inextricable lumps when the base softens by the decompostition of cellulose nitrate [39].

For these reasons, a number of film archives have continuing programs with the aim of at least recopying the information onto safety film. The authentic originals are discarded; their disposal is governed by the safety precautions applicable to explosives [428]. Because of the nitric acid and nitrogen oxides given off which attack silver images, it is not even permissible to store films on safety bases [such as cellulose acetate or poly(ethylene terephthalate)] in the same rooms where films on cellulose nitrate base are present [429–431]. Cellulose acetate can also undergo hydrolysis with acid and alkaline catalysis. The rate of hydrolysis is a minimum at ca. pH 5. Because of the acetic acid released in this process, the pH drops when hydrolysis occurs. In a slightly alkaline environment, the result is an initial gain in stability, but if the environment is acidic at the start of degradation the process is accelerated by autocatalysis [432].

Another parameter affecting the rate of hydrolysis is the degree of acetylation. The lower the initial degree of acetylation, the faster the hydrolysis. The "vinegar syndrome" is known from experience in archives holding cellulose acetate materials [433]; the cellulose acetate base decomposes to a powder with a strong vinegar odor when film cans of foil bags are opened.

No such experience has yet been obtained with polyester bases, since storage times for this more recently introduced film base are still far shorter. A problem relating to the storage life of these films is the failure of adhesion of the gelatin layer under conditions of cycling humidity.

Testing of Film Bases. The suitability of emulsion supports for archival storage is tested in quality tests. Relevant properties include the *folding endurance* in cellulose acetate and the *tensile strength* and *percent elongation at break* in polyester, especially the impairment of these properties in high-temperature accelerated aging tests in comparison to samples that are not heated [407].

In addition, the *acid value* and the *viscosity* of solutions of isolated base materials are used as a figure of merit, as are the changes in these figures after accelerated aging of representative samples. For polyester bases, the adhesion of the emulsion is tested after cycling between high and low humidity. Pollakowski investigated many vinegar syndrome incidents with cellulose acetate-based films and found that subbings based on cellulose nitrate had been used [433]. These subbings were common until quite recently [434]. Because this cause could not be proved in some vinegar syndrome cases, there is still controversy as to whether noninduced, "spontaneous" hydrolysis of the cellulose acetate had taken place or whether the decomposition had perhaps been due to storage for some time in the same room with nitrate-based films [432].

Print Materials. No cases of damage due to instability of the emulsion support have been noted for prints on the classical material, baryta-coated paper. Wood-free paper used for prints does not show the decomposition that is known in wood- containing, alum-sized book papers of similar age. In the case of polyethylene- coated papers for rapid processing, there are instances where the resin coating, especially on the emulsion side where the coating is filled with the white pigment titanium dioxide, has shown brittleness, blistering, or fibrous splitting.

Damage has occurred, in particular, to exhibition prints that were exposed to light. Similar damage patterns have been produced in accelerated aging tests with high illumination, elevated temperatures, and cycling between high and low relative humidity [435].

In images of the *gelatin – silver type,* the image layer proper – the gelatin layer with the silver image distributed in it – has generally proven to be the more stable component provided the materials were correctly processed and harmful external effects (bacterial and fungal attack, moisture, hydrogen peroxide, organic peroxides, nitrogen oxides, etc.) were averted; however, the useful lifetime is limited by the emulsion support [436]. The stability of negatives on fragile *glass plates* is thought to be similar to the image layer proper under normal storage conditions,

the expected life of glass is longer than that of the other bases [405].

The situation is different for *color photographic materials*. Here the dyes in the image layers are inherently less stable than the support. Dye stability has improved markedly as the technology has developed, but this process can by no means be considered completed.

In the case of color photographic documents worthy of archival preservation – whether still photography or motion pictures – it is therefore recommended to prepare black-and-white separations on silver – gelatin material for archival purposes. The method used is the same as in the making of plates for four- color printing or in the production of color films by the Technicolor process [437].

The preservation of original color photographs and color films can also be useful if their limited life is kept in mind. The same principles of storage apply here as for silver images: low temperature and low relative humidity, or conditioning at low humidity and storage in sealed containers. But effects on the condition of color documents due to the storage conditions are expected to become perceptible within a human lifespan.

Improvements in the storage life of processed color images have resulted from the adoption of better couplers, which form more durable dyes in the photosensitive material [438][439][440]. Approaches to extending the useful life of exhibited color photographs, where photochemical decomposition governs the lifetime of the image dyes, include the incorporation of UV absorbers in the image layers and in protective overcoats. Both quality attributes – the dark or archival lifetime and the light or display stability – of modern color materials result in lifetimes that are longer than the time span required to launch improved product generations on the market [441]. Practical experience relating to storage life therefore cannot be employed in improving the storage qualities of new product generations, for example in selecting better couplers [410], [442]. Time-accelerated tests are needed for this purpose in which the processed material is subjected to more drastic conditions than will occur in its practical use. Life expectancies deduced from these tests should be regarded as highly theoretical. With regard to light stability, the bleaching of color densities and the yellowing of whites under intense illumination imply that similar color shifts will take place under conditions of low illumination that occur in practice [443]. As a first approxiamtion, equal effects are assumed to result from equal exposures.

For purposes of practical testing, a draft standard [410] sets forth increases in illuminance that differ from one material to another.

Archival tests involve time-accelerated tests at elevated temperature. The assumption is that the mechanism of the reactions leading to changes in color density will not change in the range between the storage temperature and the test temperature. The effect is quantified with the Arrhenius equation [444]:

$$\log k = \frac{\Delta G^*}{RT} + C$$

where k is the rate constant, $\Delta G*$ is the free energy of activation, R is the universal gas constant, T is the absolute temperature, and C is a constant of integration. This relationship also decribes reasonably well the behavior of older photographic dyes in a period of up to 10 years [442].

For the selection of image dyes in practice, it can be assumed that if two dyes have the same extrapolated life expectancy, the one with the lower free energy of activation $\Delta G*$ and the lower temperature coefficient of color-density decrease will have a high probability of being the better. In other words, this test selects the dye that is also more stable at low temperature. There is an increased probability that a competing reaction (with a lower free energy of activation) will become rate-determining at lower temperature, if the measured free energy of activation in the high-temperature test is higher. The presumed reliability of the extrapolated life expectancy increases as the lower value of the experimental temperature range declines, that is, as it becomes closer to room temperature. But this criterion implies a longer experiment. The draft standard [410] proposes at least four temperatures in the range $55 \pm 0.5 - 95 \pm 0.5$ °C; the span should be at least 20 °C and preferably larger. The recommended relative humidity is 50 ± 3 %. Higher and lower humidities are permitted if special humidity values are expected in use or in storage [440], [445]. R.H. values above 60 % should be avoided in the high-temperature test, because the properties of some constituents of the image lay-

ers, especially the gelatin, change abruptly when the humidity rises above ca. 60 %. This can lead to changes in the reaction mechanism, which do not occur under service conditions.

7. Nonphotographic Properties and Their Measurement

Both photographic film materials and photographic papers must resist damage from the stresses they experience during processing and use. A variety of loads are applied, beginning with cutting, perforation etc, at the packaging stage. Films are then subjected to partly discontinuous stress during transport and may fracture, especially in very dry climates when they give up water and become more brittle. Conversely, when films absorb water, for example, in tropical countries, they may adhere to materials in contact with them.

In processing, emulsion layers swell in the aqueous baths and suffer dramatic losses of strength. Nevertheless, films and papers being transported through processing machines must not receive scratches from the conveying and deflecting rollers or from the doctor-blade arrangements used to remove liquids from the surface of the emulsions. Hardening of the gelatin during the manufacture of photographic materials is the decisive factor for the increase in mechanical strength (see also Section 3.1.4.1).

7.1. Hardening

The gelatin binder, if not hardened, swells strongly in the processing solutions and has little mechanical strength, in addition, gelatin melts at ca. 25 °C. For these reasons, hardeners are added to emulsion layers: i.e., compounds that cross-link the polypeptide chains. These hardening agents should be as photographically inert as possible.

Processing, especially with the high-speed, high-temperature automatic methods currently used, imposes severe requirements. It is relatively easy to cross-link gelatin molecules to varying degrees, so that photographic emulsions can be selectively hardened. The degree of hardening, as expected, influences the sensitometry: hardening reduces the sensitivity and gradation, restrains the development kinetics, and diminishes the granularity. The swelling factor Q = thickness of wet layer/thickness of dry layer serves as a measure of hardening. It has a value of ca. 3 for films and 4 for printing materials.

Photographically inert compounds, both inorganic and organic, can be used as hardeners. For example, chrome alum and aluminum alum cross-link gelatin at the free carboxyl groups. On the other hand, organic compounds react chiefly with the free amino groups, examples include [19]:

1) Aldehydes, e.g., formaldehyde, glyoxal, glutaraldehyde
2) Aldehyde acids, e.g., 2,3-dichloro-4-oxo-2-butenoic acid
3) Diketones, e.g., diacetyl
4) Dihalides, e.g., 1,3-dichloropropanol
5) Compounds with activated double bonds, e.g.,divinylsulfones, *N,N′,N″*-trisacryloylperhydro-s-triazine
6) Diisocyanate – bisulfite adducts
7) Bisepoxides
8) Bisaziridines
9) Dichlorotriazine derivatives, such as 4,6-dichloro-2-hydroxy-5-triazine

Examples of substances that react with the peptide groups are: (1) carbodiimines, (2) *N*-carbamoylpyridinium and isoxazolium salts, (3) esters of 2-alkoxy-*N*-carbonyldihydroquinolines, and (4) chloroformamidinium salts.

Hardeners can be added to the emulsion just before coating if the hardening reaction goes slowly enough. Fast-reacting substances, must be applied as a separate gelatin-poor overcoat. Since the mid-1980s it has become possible to integrate instant hardeners(i.e.,hardeners whose action goes to completion in < 4 h) into all other layers.

7.2. Dry Properties

The mechanical properties in dry climatic conditions are determined by the properties of both the emulsion support and the emulsion. The water contents in the various elements of the emulsion play an important part. Table 10 gives a summary [34].

Table 10. Moisture capacity (wt %) of some film components at different relative humidities

	Relative humidity of air, %		
Material	30	50	70
PETP	0.12	0.2	0.4
Cellulose triacetate	0.9	1.4	2.4
Emulsion	5.9	8.1	12.0

The water content shows relatively little dependence on temperature but water absorption and desorption is complicated by hysteresis. For example, pure gelatin absorbs ca. 15 % water when the relative humidity is raised to 50 % from the dry side but ca. 17 % when it is decreased to the same level from the humid side.

7.2.1. Brittleness

Photographic materials are generally fairly flexible. Problems can occur at low moisture contents and low temperatures. For example, the film may break in the camera, or in the finishing laboratory after films have been spliced. The film may break in the emulsion support (triacetate or PETP), or cracks may develop in the emulsion. This can happen if, the film, for instance is rapidly drawn over small rollers with the emulsion side out, so that the emulsion is stretched. This kind of takeup mechanism is mainly used in newer cameras. In order to prevent fracturing, plasticizers are added to the gelatin; examples are polyethylacrylate, polyacrylates, and polyurethane latices.

Testing Methods. Methods that have been introduced for measuring brittleness can only approximately simulate the realistic situation.

The *wedge brittleness* test is used for film [446]. Strips of film are pulled through a wedge, emulsion side out, under well-defined dry climatic conditions (20 or 30 % R.H.). The length of the fragments gives the fracture radius. Other methods exist, such as the *folding endurance test* [447], in which well-defined flexural oscillations are counted up to failure, or the *parallel-jaw brittleness tester* [448]. In the last apparatus, film loops of specified length are stretched, emulsion side out, between two parallel jaws that move at a specified speed to break the sample. The device makes it possible to determine the fracture diameter and breaking strength.

The brittleness of films depends on the filler content of the gelatin binder; silver halide grains in particular degrade this property. In addition, the drying of the film during manufacture is important; the cooler and slower the drying process, the more flexible the film will be.

7.2.2. Curl

As was pointed out in Section 7.2.1, a photographic film or paper absorbs or releases moisture depending on the ambient conditions. Gelatin swells on absorbing water, and the resulting forces curl the film or paper. Conversely, in a dry climate gelatin gives up water, and tensile forces are produced that again can lead to curl. Expansion or shrinkage is greater in the emulsion than in the base. Along with the water balance, two other factors are important in curling: (1) the initial curl applied to the base (see also Section 3.2) by special manufacturing processes, and (2) cold creeping, especially of PETP, when the material is wound on small-diameter cores (core setting). Severe curling leads to problems in the camera (sharpness) or in finishing.

Curl Measurement. Curl is measured as set forth in ISO 4330 [449]. Film specimens as described in the standard are acclimatized at $23 \pm 2\,^{\circ}\mathrm{C}$ and 15, 30, 50, 70, and 85 % R.H.; then a quantity such as the bow height in mm is measured. The curvature $1/R$ (R in m) is then calculated or determined with prefrabricated templates. A flat specimen accordingly has an infinitely large radius R. The curvature value $1/R$ is prefixed with a + or − sign, + if the emulsion is inside (curl toward the emulsion) and − if the emulsion is outside (curl toward base).

7.2.3. Dry Adhesion

Special subbings have been devised in order to control the adhesion of the emulsion layers and any backings that may be present (on roll or sheet film) on the base before and after processing (see also Section 3.2.1.3). The layers must remain adherent even on climatic changes.

The measurement or test is covered by standards [450], [451]. Specimens are conditioned

for ≥ 15 h at 23 °C and 50 % R.H. A strip of adhesive tape ca. 15 cm long is applied to the test specimen with the most uniform possible pressure; one end is left free as a grip. The specimen is held on a flat surface and the tape is rapidly pulled off with its free end at an angle of 180 °C. If the adhesion force is to be sufficient, it must be at least 9 N for an adhesive tape 1 cm wide. An acceptable film will show no material pulled loose from the emulsion side or the back. As a rule, the specimen is scored in a checkerboard pattern before the test.

There is also a humidity-cycling adhesion test. Specimens are held, hanging freely, through 12 cycles each consisting of 8 h at 50 °C and 96 % R.H. followed by 16 h at 50 °C and 11 % R.H. They are then examined to see whether parts of the layers have become loose, peeled off, or separated.

7.2.4. Dimensional Stability

The dimensional stability of photographic materials is not so important for amateur products but very much so for professional ones used in the graphic arts or in aerial-survey photography. Such materials should exhibit the smallest possible changes due to moisture, temperature, processing, and aging. Possible consequences of such changes include the registration problem in color printing.

From this standpoint, PETP is better than cellulose triacetate, for the PETP base has a smaller coefficient of thermal expansion as well as a smaller coefficient of expansion due to moisture [361].

Dimensional stability is measured as given in ANSI [452], DIN [453], or ISO standards [454]. Changes in length are determined to the best precision possible, generally with a measuring microscope. The specimens are measured before and after a high-temperature moisture treatment or processing. The coefficient of linear thermal expansion is stated in $\mu m\, m^{-1}\, K^{-1}$, and similarly the coefficient of expansion due to humidity in $\mu m\, m^{-1}\, \%^{-1}$ (where % implies relative humidity in %). For the processing test, the change is simply stated in $\mu m m^{-1}$.

7.2.5. Friction

The frictional behavior of photographic films and papers must be appropriate to the application. For example, miniature films and motion-picture films must not cause transport malfunctions in the camera. These films have extremely thin wax coatings or dispersed silicone oils on the emulsion surface and, in some cases, on the back. Such treatments also improve the resistance to scratching. A photographic paper should have enough friction to prevent its escape from the rollers and keep stacks of material from collapsing. In order to reduce the contact areas and prevent adhesion between emulsion side and back, for example in miniature films, plastic spheres (1 – 10 µm in diameter) that are soluble or insoluble in the developer are added to the protective overcoat. Roughly half the diameter of these spheres protrudes from the topmost layer.

For photographic materials in general, the static and kinetic coefficients of friction are complicated functions of temperature, moisture content, and pressure. For example, the gelatin binder becomes tacky above 70 % R.H., and friction increases markedly.

Naturally, the simulation of practical conditions in testing is difficult. There are numerous methods that (sometimes in combination) give a good approximation. The determination of static and kinetic coefficients of friction is governed by DIN 15 586 [455] and ASTM-D 1894 [456]. A sled of specified size and a mass of 200 g is pulled over the contact partner and two force measurements are made: (1) the force required to "start" the motion and (2) the force needed to keep the sled moving at constant speed over the specimen. Measurement (1) yields the static coefficient and (2) the dynamic coefficient of friction. In each case the coefficient is the ratio of the measured force to the normal force (weight of the sled). The sled can be covered with any materials that come into contact under practical conditions. The test can also be done with the sled stationary and the contact partner moving; the results are the same. The DIN standard provides for testing on an inclined plane; the quantity measured is the angle at which the specimen begins to slide. There are other methods as well. For example, the forward and reverse transport forces in cameras can be measured.

Other standards apply to the determination of film pullout force for photographic films in miniature cartridges [457], [458]. After acclimatization of the film in the cartridge under stated conditions, the film is inserted into a gauge (weight 2.5 N) and the free end is pulled vertically upward. The gauge must not be picked up from the floor.

7.2.6. Dry Abrasion

Photographic materials have little resistance to scratching. Abrasion is harmful especially when films are to be enlarged. This means that films and work areas must be kept very clean. Some improvement can be achieved by adding oily substances, such as emulsified silicone oil, to the protective layers. Such additives also reduce friction (see also Section 7.2.5).

The measurement of scratching resistance is described in standards [459], [460]. The specimens are mounted flat on a smooth surface. The holder is moved either manually or by motor. Scratches are made by a sapphire rod with a radius of 75 µm held in a counterbalanced arm loaded with weights from 0.1 to 10 N. The test is done on acclimatized specimens. The test result states the force in N at which the first scratches visible to the naked eye in obliquely incident light are produced. The greater this weight, the better is the protection against abrasion. Typical values for emulsions are 0.5 – 2 N.

In addition to this visual examination other methods of assessment can be used depending on the application of the film, for example by projection or in a film scanner.

7.3. Wet Properties

As already mentioned, photographic materials can also suffer damage from roller-transport systems or stripping devices in processing, when they are wet and swollen. In particular, since 1972 advances have continued to shorten processing times and raise processing temperatures, imposing added mechanical stresses on the materials. For the manufacturers of photographic materials, this has meant that improvements in wet properties are also needed.

7.3.1. Swelling

Emulsion layers must swell in processing in order to facilitate the necessary diffusion processes. For example, fresh developers, bleaching chemicals, and fixers must be able to enter the emulsion, while reaction products such as bromide or silver complexes from fixation must be able to diffuse out of the film. The degree of swelling depends not only on the hardening of the layer (see Section 7.1) but also on the salt load and the pH of the processing solutions. Finally, the drying of the photographic material plays a role [1]. The emulsion experiences not just vertical but also lateral swelling. This lateral swelling, however, is hindered by adhesion to the base. The stresses produced can curl the material. When the degree of swelling is sufficiently high (i.e., ca. 3), however, the modulus of elasticity becomes small, and so the stresses are low too. The performance of the base (precurl) is then dominant.

There are many techniques for determining the degree of swelling or the swelling factor (for a definition of swelling factor see Section 7.1). The simplest method is weighing. Stripping off the "surface" water is a problem in this approach. Other methods are described in [9]. For example, the static pressure between two parallel plates with the swelling material between them can be used to determine the swelling height [461], [462]. Or a probe can be lightly pressed against the swelling surface and precision distance sensors can be used to measure the swelling height. These techniques can also be performed at intervals so as to determine the kinetics of swelling (as in the static-pressure method). In order to arrive at the swelling factor, the dry layer thickness must be determined, for example on a microsection examined under the microscope. Typical swelling factors are ca. 3.

7.3.2. Wet Abrasion

When photographic materials have swollen in the processing baths, they are very sensitive to surface damage.

Wet abrasion behavior is determined similarly to dry abrasion; the test is governed by an

ANSI standard [463]. The abrader is again a sapphire, but now the radius is larger, $r = 0.38$ mm. The weight ranges continuously from 0 to 5 N and can be further increased with supplementary weights. Measurements are performed in temperature- controlled water or processing solutions after specified times have been allowed for the specimen to swell. Producing the first visible damage requires a weight of ca. 0.5 – 1 N.

7.3.3. Reticulation

A photographic material may have up to 15 thin layers, which differ greatly in the contained ballast (silver halide, coupler, oil former, etc.). Such structures can be treated as laminates. The lateral swelling tendency of the emulsion layers can generate stresses that differ between adjacent layers. Cracks may result, and these may affect image quality. The danger of reticulation diminishes with increasing hardening.

7.3.4. Wet Adhesion

The wet abrasion resistance described in Section 7.3.2 strongly depends on the wet adhesion. Determination of wet abrasion resistance is meaningful only if the wet adhesion is adequate. The test is a primitive one: After the specimen has had a sufficient time to swell, the emulsion side is scratched with a cross-shaped stylus, then rubbed with a finger or a rubber stopper, and afterwards examined to see whether the emulsion layers have separated at the scratched mark. As a rule of thumb, the adhesion between emulsion layers is poor if one or more of the layers contain too little binder (≤ 45 vol % gelatin).

The subbing also contributes to adhesion between emulsion layers and the base. Everything mentioned in Chapter 7 about the dry and wet properties of emulsion layers also applies to backing layers when these are used, as in roll and sheet films.

8. Photographic Materials and Applications

8.1. Photographic Plates

Photographic emulsions are currently coated on paper or film base. Such flexible materials in roll form are easier to handle during production, packaging, and use. Modern supports also have excellent physical properties (see Section 3.2.1.1). During the first hundred years of photography, *glass plates* were more popular because of their availability, flatness, and dimensional stability. Except for paper, no flexible supports existed. Today, glass plates are employed only when the requirements on dimensional stability are extremely rigorous; this is almost exclusively the situation with scientific and technical applications. Most plates have a simple coating structure: subbing, emulsion, and protective overcoat.

Photography has always been a useful tool for scientific observation. For example, UV radiation was discovered as early as 1801 by exposing a paper coated with silver chloride to the light emerging from a prism.

Holographic Plates [464], [465]. The discovery of holography and the laser opened up a new field of application for photographic plates. Plates for this use must have extremely fine grain. Their micrate emulsions have grain diameters of 30 – 90 nm, high contrast ($\gamma = 5 - 8$), and extremely high resolving power (up to 3000 lines /mm). Their spectral sensitivtiy must be matched to the radiation from a variety of lasers (514.5 nm for argon, 694.3 nm for ruby, or 632.8 nm for He – Ne lasers).

Suppliers. Agfa – Gevaert (Holo-test) and Eastman Kodak (H.R. and Holographic Plate) supply an assortment of photographic plates for holography, but general-purpose high-resolution plates from the technical lines are also used.

Mask Making (→ Imaging Technology, Chap. 4.2.2.). Exposure masks for the fabrication of printed circuit boards and microelectronic devices are sometimes produced by the silver halide process. Such plates are of high

resolution. However, most masks are made with photoresists.

Spectroscopic Plates [465], [466]. Spectroscopic plates, with spectral sensitization ranging from the UV to the IR, are employed for the detection and recording of visible and invisible radiation. Originally these were used for spectrography. After the introduction of recording spectrophotometers, their application shifted to more general work. Speeds are in the low to middle range for camera use and can be varied somewhat through the processing technique.

The gradation ($\gamma = 0.8 - 4.0$) ranges from normal to ultra-hard; an effective antihalation protection is required. Special low-gelatin- content plates are used for the UV. Ultraviolet plates are sensitive to $\lambda \geq 200$ nm; Schumann plates to $\lambda \geq 180$ nm. Infrared plates are generally sensitive to thermal radiation and must therefore be stored at low temperatures.

Suppliers. The main suppliers of these special materials are Agfa – Gevaert (Sciencia) and Eastman Kodak (Spectroscopic Plate). Similar plates are avialable for photomicrography and electron micrography.

Nuclear Plates [467–469]. Nuclear plates were important in nuclear research and technology, where they found use in the detection of elementary particles and in the recording of the tracks left as they passed through the emulsion, interacted with other nuclei, or decayed (Fig. 86). These studies stimulated the understanding of elementary matter.

Many scientists in this area prepared their own emulsions. There was an intense exchange of knowledge about interesting and unconventional methods of preparation, and the effectiveness of the resulting materials.

Nuclear track emulsions are very silver-rich (ca. 80 % AgBr), have a moderate grain size (0.1 – 0.3 µm), a narrow grain-size distribution, and are not dye-sensitized. The sensitivity is determined by the maximum energy of a charged particle that just leaves a detectable track in the emulsion. The most sensitive grades of track emulsions detect particles of all energies.

The development of other kinds of detectors (cloud chambers, bubble chambers, coincidence counters, etc.) has made nuclear track emulsions much less important.

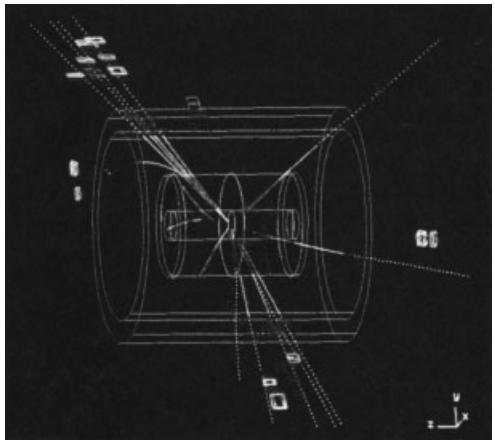

Figure 86. Decay of a Z^0 particle in a quark/anti-quark as seen in the DELPHI detector (CERN, Geneva)

Autoradiographic Plates [470], [471]. Autoradiographic plates are glass plates that are coated with a thin layer of nuclear track emulsion or another fine-grain emulsion. Often a releasing layer is incorporated in the emulsion. Then, it is possible to remove the light-sensitive layer from the glass support before or after processing (stripping film). A development of the 1980s in this area is the optical or thermal excitation of a phosphor layer (see Section 8.2.5), which is exposed to the radioactive specimen, substituting the photographic plate. The released light can be detected by electronic devices, and digital image processing is possible.

Astronomical Plates [467]. High-speed plates (DIN 17 – 25, ISO 40 – 250) are used in astronomical research. These materials have normal to hard gradation ($1 < \gamma < 1.5$) and a variety of spectral sensitizations (blue-sensitive, orthochromatic, or panchromatic).

The light, often very feeble, is "collected" during extremely long exposure times. Under these conditions, low-intensity reciprocity failure (see Section 2.1.6) becomes a problem and must be held to a minimum through the design of the material or through special processing. Astronomers sometimes employ pretreatment and special processing to push the speed. These plates usually form part of the manufacturer's technical line.

8.2. Films

8.2.1. Films for Professional and Amateur Photography

Color negative, color reversal, and black-and-white films used for still photography fall into a number of sensitivity classes. (For the determination of film sensitivity, see Section 6.1.4 and the following ISO standards: ISO 5, ISO 6, ISO 2240, and ISO 5800). (ISO standards can be obtained through the national standards organizations or directly from ISO).

Low-speed films (ISO speeds from ca. ISO 25/15°–ISO 50/18°) are fine-grained; they have high resolving power and relatively thin emulsion layers and therefore offer excellent sharpness. They are used, especially in professional photography, as originals for enlarging and for images to appear in printed materials.

High-speed films are those with speeds of ISO 800/30°–ISO 5000/38°. They are used for short exposure times under unfavorable lighting conditions (sports, theater, etc.). At the time of writing (1990), films with speeds of ISO 100/21°, ISO 200/24°, and ISO 400/27° are used most frequently.

Table 11 lists the most common sizes and packagings of photographic films. There are standardized film and cassette (cartridge, magazine) configurations which fit the body and transport system of different camera types.

Sheet, roll, and miniature films differ in the coating structure and the thickness of the base. The common base materials are cellulose esters and poly(ethylene terephthalate). Both materials satisfy the requirements of ISO 5438 and ISO 7830 for safety films; i.e., they are not easily ignited and do not burn readily [42]. Cellulose esters are employed chiefly in amateur films. Because of its greater dimensional stability and tensile strength, PETP finds use in sheet and other films for professional users.

Sheet films and roll films 120 and 220 have a nonphotosensitive gelatin backing, which counteracts the tensile forces developed in many photosensitive front coatings. Colored backings are occasionally still used for antihalation protection; such carbon black or dyed coatings must be removed in processing (see Section 3.2.1.6).

Most cellulose-ester-based films have the base slightly gray-colored (gray base) for antihalation purposes. In addition, silver–gelatin or dyed gelatin coatings (in B & W films) are applied as antihalation undercoats below the light-sensitive emulsion layers. These must be bleached or decolorized in processing so as to render the film transparent.

Figure 87 shows several coating structures for photographic films. Today, the sensitive emulsions are usually applied by multiple coating (up to 5 coating operations). A modern color film can have as many as 19 individual coatings (see Fig. 45).

When films in roll form are exposed in the camera, they are wound from a supply chamber to a takeup chamber. Cassette films have both chambers integrated into a plastic cassette. For *roll films* 120 and 220 size, the spool on the takeup side is identical to that on the supply side; after the last frame is exposed, the empty supply spool is moved over to the takeup position.

135 Size miniature films are generally transported, frame by frame, into a takeup chamber with a special permanent takeup spool. When the film has all been exposed, it must be rewound back into the cartridge before it is removed from the camera.

In some *motor-driven cameras*, the film is loaded and then, before exposure, its entire length is transported into the space supplied. As each frame is exposed, the film is rewound into the original cartridge.

In order to facilitate the adjustment of the amateur's camera to the speed of the film loaded in it, cassettes for the 126 and 110 and cartridges for 135 size films bear codes, accessible from outside, that regulate the exposure control in properly equipped cameras and also contain information about the length of film loaded.

Color reversal films are balanced either for daylight (5500 K) or tungsten light (3200 K). Color reversal films for tungsten light are specially balanced for longer exposure times. In contrast, films for daylight are balanced for short exposure times or flash. The manufacturer provides instructions on the use of filters for exact matching to the color of the light source. Most color negative films can be used with all types of illumination. Because of wide exposure latitude (see Section 6.1.2), exposure is less critical with color negative films than with color reversal films. When the negative film is printed on color paper, however, part of the exposure range

Table 11. Common film sizes and packagings for still photography

Film type	Cutting width *W*, mm; minimum length *L*, mm	Image area dimensions	Typical manufacturing and packaging	Dimensions according to ISO standards ∗
Sheet films		65×90 – 300×400	in boxes	1012
Roll films	*W* 61.75	45×60	unperforated film spooled	732, 897
120 and 220 Size	*L* 808 and	56×72	together with backing paper or	
	L 1545, respectively	60×60	leader paper on a special spool	
		60×90		
70 mm film	*W* 69.95	56×72	unperforated or perforated	3023, 2467
	L not determined		film on special spools	
35 mm film	*W* 34.97	24×36	film perforated according to	491
	L not determined		ISO 491 on spools or cores	
135-Size film	*W* 34.97	24×36 or	perforated film on spools	1007, 897
	L 639.5 (12 expos.)	17×24 (rare)	in standardized magazines	
	L 1149.5 (24 expos.)			
	L 1605.5 (36 expos.)			
126 Size film	*W* 34.97	28×29	spooled with backing paper	3029, 897
	L 470 (12 expos.)		in standardized cartridges,	
	L 851 (24 expos.)		size 126	
110-Size film	*W* 15.95	13×17	spooled with backing paper	7261, 7320, 7374
	L 455 (12 expos.)		in standardized cartridges,	897
	L 750 (24 expos.)		size 110	

∗ ISO standards can be obtained from the ISO Central Secretariat, Case postale 56, CH 1211 Genève, or from national standard organizations.

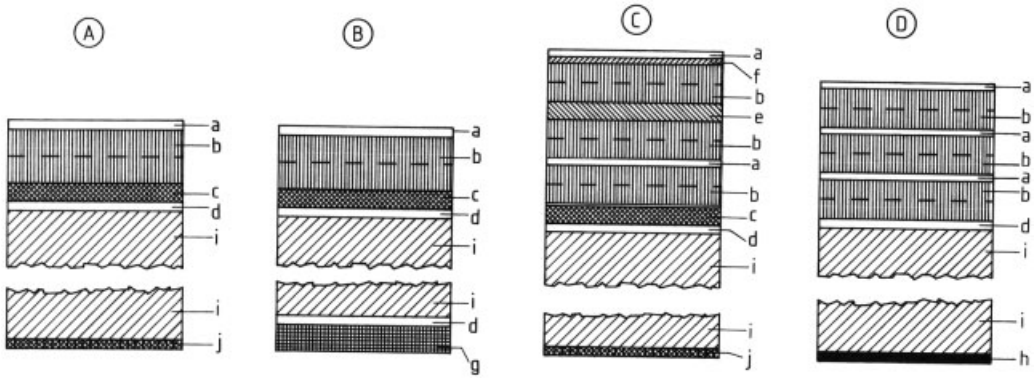

Figure 87. Examples of coating structures of photographic films (not to scale)
A) Black-and-white 135-size film (complying with ISO 1007); B) B & W roll or sheet film (complying with ISO 732 and ISO 6407); C) Color negative or color reversal 135-size film (complying with ISO 1007); D) Color positive motion-picture film (complying with ISO 491)
a) Overcoat (0.2 – 1.0 µm); b) One of three emulsions (3 – 12 µm) with panchromatic sensitization (B & W) or spectral sensitization (color films); total thickness of all emulsion layers in color films 12 – 25 µm; c) Silver – gelatin antihalation undercoat (2 – 5 µm); d) Subbing (< 0.3 µm); e) Yellow filter layer (3 – 6 µm); f) UV filter layer (0.5 – 2 µm); g) Gelatin NC backing (colorless or colored); h) Backing with carbon black or dye; i) Base: cellulose esters or polyethylene terephthalate (PETP), usually with a weak gray color (gray base), but occasionally colorless (clear base); thickness 170 – 200 µm (sheet film), 90 – 100 µm (roll film), 115 – 130 µm (films size 135, 126, 110 motion-picture film); j) Antistatic coating (< 0.3 µm)

is lost again. Reversal transparencies are superior to paper prints both in color reproduction and in brightness range; they can be printed on reversal paper (with some loss of quality).

Professional photographers use chiefly large-format sheet and roll films. Often called "professional films," these materials are adapted to the special needs of this user group; i.e., with respect to characteristic curve, Schwarzschild behavior (see page 11), color balance, response to pushing development, or mechanical properties (such as surface texture for retouching).

Color and B & W duplicates are made with reversal copy film or with internegative films

combined with positive films (print films). Print film (packaged as sheet film or 35 mm roll film) can also be used to obtain transparencies from negatives.

8.2.2. Motion Picture Films [472][473]

Camera Films. The professional production of motion picture film is based on a complete negative – positive system including several intermediate steps in order to facilitate editing and printing. 35-mm film has been the main standard material during the last hundred years (Table 12). 70-mm film [474], [475] became a second standard associated with the high-quality projection on wide screens combined with stereo sound. This application started with the introduction of CinemaScope and Cinerama in the 1950s. The use of 70-mm film as camera-sensitive original is, however, quite exceptional today. Most 70-mm copies have been blown up on print film from 35-mm originals. The smaller formats (16-mm and 8-mm) are employed for television commercials and films, educational purposes and amateur motion pictures. After World War II the color motion picture business gained momentum and acceptance. Subsequently black-and-white materials gradually disappeared.

Table 12. Figures of merit of motion picture films

Film width, mm	Frame format (for projection), mm	Speed of projection, frames/s	Film consumption, m/min
35	20.9 × 15.2	24 ∗	27.4
16	9.6 × 7.0	24 ∗	10.9

∗ The projection speed of TV films is 25 or $16\frac{2}{3}$ frames/s.

Motion-Picture Color Negative Film. This material is the basis of a process of several steps with the positive print as a final product. The layer structure is very similar to that of the color negative employed by photographers. However, the requirements concerning color reproduction and image quality are very high because several printing operations must be carried out between the original and the final positive. Moreover, a high image enlargement is usual for theater projection. Therefore every detail, starting from the shape and structure of the emulsion grains (see Section 3.1.1), is carefully controlled. Negative materials for camera use are available in the medium to high-speed range (≥ ISO 100). The lighting conditions in the studios are carefully balanced so that a large exposure latitude is not a necessary condition. The sensitivity is optimal for tungsten light (3200 K). Shooting under real daylight conditions requires a conversion filter. The support material is often cellulose triacetate for convenient editing and splicing.

Reversal Motion-Picture Films. Immediately after processing, reversal motion-picture films give a positive suitable for exhibition. If a reversal image is compared with a copy of a negative film having virtually equal sensitivity, the reversal image is finer-grained and sharper, and also less costly. If, however, a larger number of copies must be made from a film, the negative – positive process becomes more economic.

Black-and-white reversal films in narrow formats are hardly used any more. Color reversal films in these formats often have a back coating of carbon black to combat halation as well as improve the mechanical properties of the film.

Packaging. 35 mm negative motion-picture films are available in 60, 120, and 300 m rolls on plastic cores; 16 mm films, in 30 and 60 m lengths on metal daylight spools and in 120 m rolls on plastic cores. 16 mm films have perforations on one edge or both edges. 8 mm films (Super-8 perforations on one edge) are sold mainly in cassettes containing 15 m of film (DIN 15 823) and, in this packaging, are used chiefly by amateurs. The latter application is being superseded by video recording. The motion picture technique is well established and standards exist for, e.g., materials, formats, treatment, conservation, sound tracks, and film transport.

Sound Recording Films [476], [477]. In professional film-making, the original sound is usually recorded separately on magnetic tape, then synchronized with the scene by the use of sync markings on the film (pilot sound). The sound is then transferred to the film in one of two ways: to a magnetic stripe on the film material (*magnetic sound*) or, in the form of an optical copy, to an optical sound track on the edge of the film (*optical sound*). Because the projected frames are transported discontinuously (one at a time) by employing a loop, the sound must be

scanned continuously outside the loop. Hence, the image and the sound must be offset from each other in the direction of film travel.

The *optical sound track* is created with electro-optic transducers (e. g., Karolus cell), which vary either the track width (variable-area recording) or the image density (variable-density recording). When the copy is exhibited, the track modulates the light flux incident on a photocell.

In the optical sound system, high-resolution sound negative films are used as the intermediate step. In the case of variable-area recording, the film is developed to a steep gradation ($\gamma = 4 - 5$).

In the *magnetic system,* exhibition copies are made on films with magnetic stripes applied at the factory. In addition to the professional copy, direct sound recording on the camera film is possible; this technique is employed chiefly by amateurs. The width and location of the stripe depend on the film format and perforation.

Singly - perforated films have the recording stripe on the unperforated edge; a narrower "balancing stripe" on the perforated edge brings the two edges of the film to the same thickness so that the film will wind evenly on the roll. This second track is occasionally used for sound recording, such as for stereo playback (see Fig. 88).

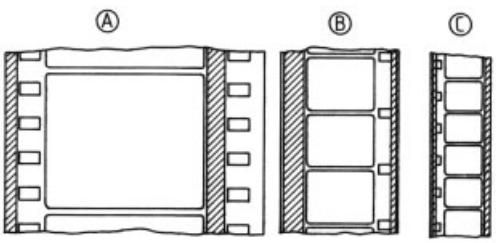

Figure 88. Arrangement of magnetic tracks in motion picture films
A) Double-perforated film; B), C) Singly perforated film

Both camera films and positive motion picture films (see below) with added magnetic stripe are available in all the usual formats and packagings. The sound track can also be applied in the developing plant after exposure.

Positive Motion Picture Films (Motion Picture Print Films). In the negative – positive process, not only the image quality of the negative but also the fineness of grain, resolving power, and sharpness of the positive film are crucial for the quality of the projected image. Color positive films often have the three dye layers in a different order from camera films; for example, the magenta layer may be on top (see Section 5.2.2.6). This is done to match the material to the sharpness perception of the human eye, which is greatest in the green layer. Fine grain emulsions and incorporated filter dyes are additional factors contributing to the resolution. Positive films are roughly a factor of 10 – 100 slower than camera films. They are developed to a gradation of 2.5 – 3.5. B & W positive films have no antihalation coating. Color positive films have a dye or carbon-black backing. The positive film is exposed in contact with the negative in special motion picture film copying machines, or by optical imaging of the negative (principally when normal-sized negatives are reduced to small-format film).

Packaging. 35 mm width with standard "positive perforation"; 16 mm with single or double perforation. Plastic core, rolls of 300 and 600 m.

Films for the Production of Duplicates. The motion picture industry uses negative duplicates when making exhibition prints, in order to avoid mechanical damage to the original negative in the copying process. Fades and dissolves as well as special effects are produced in the preparation of negative duplicates. The duplicate negative is obtained either by copying the original negative onto an intermediate positive film (interpositive, master positive master) and copying the interpositive onto duplicate negative film, or by copying the original negative onto a duplicate reversal film. From a reversal original, a duplicate negative can similarly be made on internegative film, so that exhibition copies can than be made on print film.

Reversal copying films are also used for making prints from reversal originals. These copying films are developed similarly to reversal films.

Special Films for Television. In addition to negative and positive motion picture films in 35 and 16 mm formats, television uses mainly 16 mm reversal films with gradation and color characteristics especially matched to the scanning properties of television. These special color reversal films can also be used for documentary, instructional, and other nontelevision purposes.

Along with films balanced for daylight, there are those balanced for tungsten light. They have speeds of ISO 50 – 400, but special processing makes it possible to "push" individual film types to three times the nominal speed if some loss of quality is acceptable. In addition to the camera films, there are reversal copy films suitable for TV use.

8.2.3. Reprographic Films

Reprographic Films are described in detail elsewhere (→ Imaging Technology, Chap. 3.1.1.).

8.2.4. Microphotography [478–481]

The idea of photography at a reduced scale was conceived as early as 1850. The systematic microfilming of printed matter, containing alphanumeric and/or graphical information, for documentation and archiving purposes started only in 1920 with the copying of bank checks on 16 mm film. Today, microfilm is a legally accepted storage medium, used in document management systems including computer-aided retrieval (CAR). Replacement of these systems by digital storage media is not anticipated, because of the permanency of the records and the very favorable material costs. Integration into hybrid systems can, however, be expected. A more recent application is computer output on microfilm (COM).

Depending on the specific application and the available equipment, several formats and microimage configurations are in use. Basic forms are film roll, fiche, and aperture card. About 1000 images can be stored on 10 m of 16 mm film. The camera original is almost always on low-speed silver halide material. High resolution and contrast are typical. The resolution can exceed 200 lines/mm, in order to ensure the legibility of the microimage when enlarged by a reader-printer (e. g., 384 lines/mm).

The type of microfilm used depends on the application:

1) For the recording of engineering and other documents, medium-speed films designed for maximum sharpness are generally used.
2) Higher-speed microfilms are employed in high-speed continuous-feed cameras for filming printed documents.
3) For use in COM systems (computer output on microfilm), there are special films whose spectral sensitivity matches the light emitted by the phosphors in the cathode-ray tube screen.

Certain types of microfilm can be developed by simple heating, without any chemical treatment. Duplicates are often made on inexpensive diazo or vesicular films (→ Imaging Technology, Chap. 2.3.).

Under proper storage and processing conditions, developed microfilms have an indefinite lifetime.

Packaging. 16, 35, 70, and 105 mm, commonly unperforated; 16 and 35 mm perforated; in lengths of 30, 60, 120, and 300 mm, and in sheet form.

8.2.5. Radiographic Materials [471], [482–486]

Radiographic images are principally shadow images created by the direct transmission of X-rays through the object. The image on the screen or on the developed film represents variations in X-ray absorption over the object. The local absorption is determined by intrinsic properties of the material (atomic number), the composition, and the thickness. Usually the radiograph is a negative image that can be viewed and interpreted in transmitted light. X-ray films are often coated on both sides of a dimensionally stable polyester base (colored or uncolored). Radiography requires high contrast in order to enhance details. Radiographic materials can be broadly divided into two classes: screen films and nonscreen films.

Screen Films. Screen films are preferentially used in medical diagnosis. During exposure, the film is sandwiched between two intensifying screens in a light-tight cassette, resulting in a signficant reduction of the radiation dose. An *intensifying screen* consists of a thin layer of phosphor crystals dispersed in a binder and coated on, for example, a polyester base. During exposure, the screen converts X-rays to visible light (blue, green), or UV. The output intensity of the screen is directly proportional to the intensity of the incident radiation. About 95 % of the

exposure of the film stems from the light emitted by the phosphor. Matching the spectral response of the emulsion to the color of this light results in optimal screen – film combinations.

Until 1975, blue-emitting calcium tungsten oxide [*12789-67-2*], with a conversion efficiency of about 5 %, was the main phosphor. Modern screens, having a quantum efficiency more than three times as high, are based on rare-earth compositions. The position of the peak spectral lines of these screens can be varied with dopants (activators). Current screens emit blue, green, and UV; they are manufactured for a variety of purposes, such as high resolution, fast or low dose. This diversity is also reflected in the selection of film materials. Orthochrome (green) and monochrome (blue) systems are common.

Double coating and relatively low silver content in the emulsion ($\leq 8\ g/m^2$) favor rapid automatic processing. A dry radiograph can be obtained within 90 s after insertion, and even shorter processing times can be expected.

The disadvantage of screen radiographs is a certain loss of image quality. The images are less sharp than direct X-ray images because of optical effects, and the image becomes mottled because of statistical fluctuations in the number of quanta incident on a given area of the screen.

The image on an intensifying screen or that displayed by a cathode-ray tube can also be photographed with a camera. The film used is a single- coated small-format sheet or roll film with a spectral response corresponding to the emission. This method is useful for routine examinations.

Nonscreen X-Ray Films. Direct nonscreen X-ray film is employed when fine details have to be observed (e.g., bone structure). The absorption of the emulsion layers is increased by coating with more silver and keeping the silver – gelatin ratio as high as possible. Such films are usually double- coated.

The *nondestructive testing* of materials may involve industrial radiography (→ Nondestructive Testing , Chap. 2.2.). The inspection of steel castings and pipeline joints is a well-known application. The detection of small defects requires high-resolution direct X-ray materials without phosphor screens. The films may, however, be provided with intensifying lead foil and placed in a heavy-duty packaging for use under difficult conditions, for example, in machine shops or outdoors. The lead screen acts as a source of secondary electrons and also reduces X-ray scattering. The range of X-ray energies employed is much broader than in medical applications, because of the great diversity of materials to be tested. The film must show a fairly constant response over a wide energy range, including high-energy radiation from radioisotopes. The films are double- coated and contain up to 25 g silver per square meter. Low-speed materials with very high resolution are also available.

8.2.6. Infrared Films and Plates [466], [487–489]

Infrared films have special dyes providing sensitivity up to 1300 nm, but more commonly to 900 – 1000 nm. Both B & W and color materials are available. The color film, a "false color" material, consists of three emulsion layers sensitized in the green (nearest the base), red, and infrared (top layer); after processing, however, these form yellow, magenta, and cyan images (in the same order). The image shows no natural colors but reveals details that differ in reemission quality.

Both color and B & W materials are used to penetrate haze in air and land surveys; for special applications such as agriculture, archeology, geology, and oceanography; for special effects such as "nighttime photography in daylight"; and for making multispectral images with which properties of an object can be better interpreted. Filters are almost always used during exposure in order to block the UV or visible part of the spectrum. For this reason, the practical sensitivity of these films depends strongly on the filters employed.

8.2.7. Films for Aerial Photography [490–492]

Both B & W and color materials, most of them with dimensionally stable polyester bases, are available for aerial photography.

Black-and-white aerial survey films are panchromatic films with uniform sensitization

to ca. 700 nm. To reduce scattering by haze, they are exposed behind yellow or orange filters.

The speed is comparable with camera-sensitive materials for pictorial photography. The contrast of the scene is generally lower than in the case of terrestrial photogaphy and tends to be decreased further when photographed from a higher altitude. The gamma may be varied from 0.8 to 2.0 by the developing time depending on the altitude. IR-sensitive black-and-white film is also available for specialized applications (see Section 8.2.6). *Color films* are preferentially used at low altitude because the low contrast at high altitude disturbs the effect of color. Materials are available as reversal films, negative films and infrared sensitive color films (false-color film see Section 8.2.6).

Reversal and negative films are structured essentially like ordinary amateur materials but with steeper gradation.

Aerial survey films are used, for example, in cartography, land use planning, oceanography, environmental monitoring, and military undertakings.

8.3. Photographic Papers

8.3.1. Papers for Continuous-Tone Reproduction

Demand for silver halide photographic papers used in professional and amateur photography has risen steadily over the years. Most of the growth is in color negative paper, with a growth rate of ca. 8 % per year. Black-and-white papers have been declining for years (stagnant at present), and color reversal paper is insignificant, with ca. 2 % of total color paper demand. In 1987 the three largest markets, Japan, Europe, and the United States, consumed ca. 500×10^6 m^2 of color negative paper, 60×10^6 m^2 of B & W paper, and 13×10^6 m^2 of color reversal paper. Nearly 50 % of color negative and B & W paper consumption is in the United States, ca. 30 % in Europe, and roughly 20 % in Japan.

8.3.1.1. Black-and-White Papers

Since a wide selection of B & W papers is available, specialists and amateur photographers are able to find the right paper for any practical application. The following different product qualities exist

1) Papers with cold, neutral-gray, or warm image tone
2) Papers with fixed or variable gradation (contrast)
3) Papers with panchromatic sensitization
4) Papers with resin-coated or baryta-coated bases
5) Papers with glossy, matt, or textured surfaces
6) Papers with or without incorporated developing agents

All B & W papers have the coating structure shown in Figure 89 in common. Emulsions vary widely and are optimized by application. Low-speed (ISO P 50) silver chloride emulsions are common for contact printing; for enlargement, emulsions with speeds between ISO P 100 and ISO P 500 (under ISO 6846-1983) are used (see Section 6.1.4).

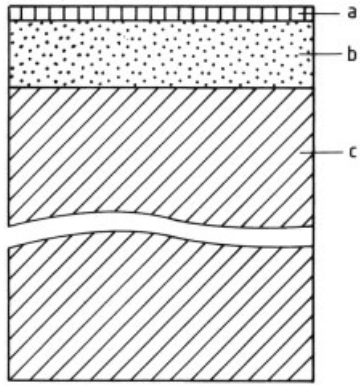

Figure 89. Layer structure of a B & W paper (schematic) Thickness of emulsion layer plus protective layer, 8 – 10 µm. Base thickness: ca. 220 µm (RC paper plus antistatic coating), ca. 145 µm (baryta-coated paper), or ca. 245 µm (baryta-coated paperboard)
a) Protective layer; b) Emulsion layer; c) Base

Pure silver chloride and silver bromide emulsions developed in Phenidone – hydroquinone developers give neutral-gray silver deposits. Image silver shifted more toward brown (warm-black to warm-brown image tones) is obtained with silver chloride bromide emulsions. The silver coating weight of B & W papers ranges between 1.0 and 1.8 g silver per square meter depending on the emulsion type and the desired gradation.

It is often useful to examine the image during the development process. For this purpose, the perceptible differences in continuous tone reproduction and shadow definition seen under darkroom illumination (safelight) must correspond as exactly as possible to the impression given in daylight or bright tungsten light. This requirement is satisfied by yellow-green light, which has its mean wavelength at the maximum sensitivity of the eye. Such illumination is produced, for example, with tungsten light combined with an Agfa – Gevaert no. G 7 filter (Fig. 90). B & W papers should be as insensitive as possible to this safelight; i.e., they should have a sensitivity gap in the maximum-transmission range of such filters.

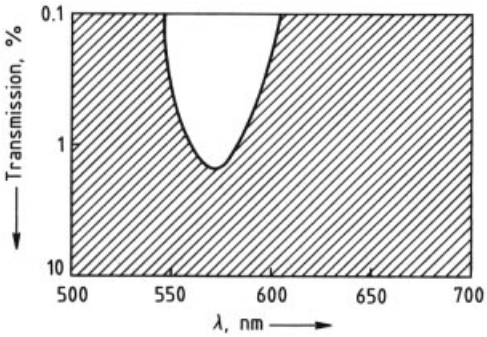

Figure 90. Transmission spectrum of Agfa-Gevaert no. G 7 safelight filter

Because pictorial negatives have density ranges between $D_{\text{neg}} = 0.4$ and 1.6, it is desirable (as explained in Section 6.1.5) for each type of B & W paper to vary in gradation while all the other properties remain unchanged.

The market share of *fixed-gradation B & W papers* is declining. These papers consist of up to six distinct materials, each representing just one gradation from "extra hard" ($\log ER \leq 0.5$) through "normal" ($\log ER = 0.8 - 0.9$) to "extra soft" ($\log ER = 1.35 - 1.50$, where ER denotes exposure range). Such a contrast "fan" (see Fig. 91) is obtained by modifying the emulsion composition for each gradation; in order to produce the proper $\log ER$ value (see Section 6.1.2 for definition), two or three emulsions differing in gradation and sensitivity are often mixed in the requisite proportions. If necessary, the sensitivity is controlled by the use of sensitizing dyes for the blue part of the spectrum.

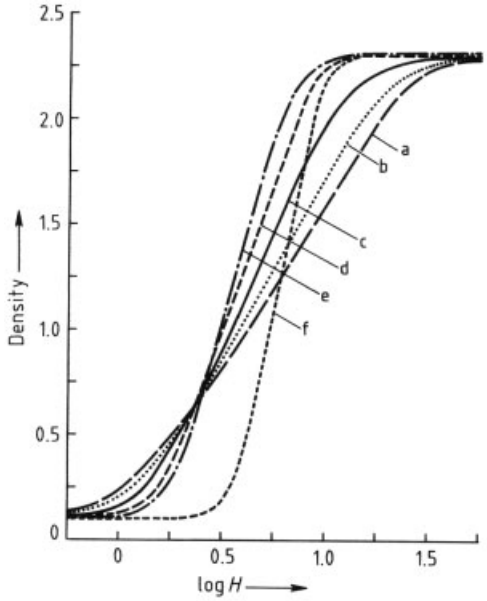

Figure 91. Density curves of Brovira papers (Agfa – Gevaert) with various gradations (H in lux · s)
a) Extra soft; b) Soft; c) Special; d) Normal; e) Hard; f) Extra hard

In order to simplify stock-keeping, there have been attempts to combine in a single B & W paper the gradations required most often in practice. Papers with *optical gradation control*, in which the contrast can be varied within certain limits through precisely metered sub-toe pre-exposure (flashing), are occasionally found on the market. These make use of the fact that the density curve of an emulsion becomes flatter and flatter with increasing pre-exposure, until the pre-exposure reaches the threshold value for blackening and the emulsion becomes fogged.

The filter method, patented as early as 1911 [493], has become the standard for *variable-contrast papers* and now represents a serious threat to the market for B & W papers with fixed gradation. In this technique, first realized by Frank F. Renwick [494], the $\log ER$ is varied over the range of 0.5 – 1.3 by altering the color of the printing light, so that a single paper can be used for the entire range of gradations.

In the simplest case, the emulsion consists of a mixture of two silver chloride bromide emulsions, one of which is not sensitized and responds only to blue light, while the other is sensitized to the short-wavelength green, thus leaving the sensitivity gap as required for B & W papers

and nonetheless being sensitive to blue and green light.

If the paper is exposed through a magenta filter, which passes blue light but blocks green, both emulsions respond. A high-contrast image is produced, because the density curves add (both emulsions being blue-sensitive; see Fig. 92 A). If the paper is exposed through a yellow filter, which passes mainly green light but blocks blue, initially only the green-sensitized emulsion responds and contributes to image formation mainly in the highlights. The natural blue sensitivity of both emulsions comes into play only at higher intensities, i. e., in the shadows, thus making a low-contrast contribution to the overall image produced by the superposition of the emulsions (Fig. 92 B). The result is a flatter overall gradation.

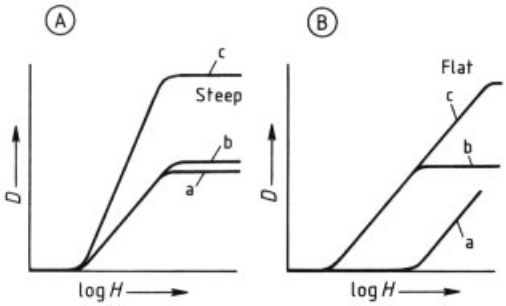

Figure 92. Principle of a variable-contrast B & W paper
A) Blue exposure; B) Green exposurea) Unsensitized emulsion; b) Green-sensitized emulsion; c) Total gradation

Between the maximum achievable contrasts (extra hard = exposure through magenta filter; extra soft = exposure through yellow filter), any intermediate gradation can be produced by mixing color components in the exposure light. Any enlarger having a color-mixing head or filter holder with an 11-filter set can be used. Since the sensitivity of the paper depends on gradation these filters differ in gray density, so that the exposure time can be kept as constant as possible at any gradation (Fig. 93).

In addition to the standard papers with fixed gradation and the variable-gradation paper, *panchromatic B & W paper* still has a role to play. The silver halide grains in the emulsion of this paper are sensitized to red in addition to blue and green, matching the three colors of a color negative film. In this way, all the subject colors stored in the color negative film are rendered as gray values that correspond to human perception (see Section 4.1). This material is used, for example, by photojournalists who work with color negative film but wish to transmit a first report in black and white. Because of its panchromatic sensitivity, this paper can be processed only under a safelight permitted for color negative papers (e.g., Agfa – Gevaert no. 08 safelight filter).

Every user of B & W papers desires deep shadows, good highlights, and optimal whites in the finished print. Many materials have *optical brighteners* incorporated in the emulsion to enhance the whites; large amounts of optical brighteners above the emulsion may be perceived as unpleasant because of a bluish shimmer in the blacks.

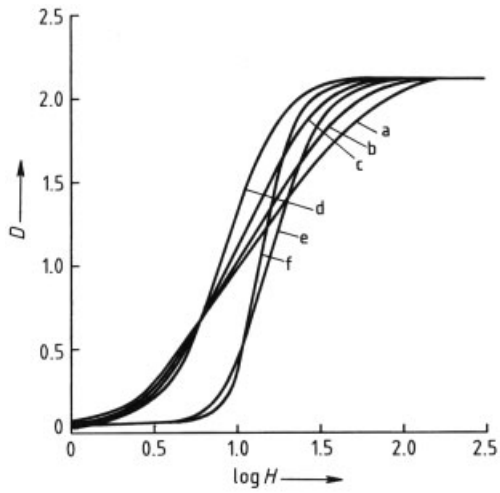

Figure 93. Density curves of Multigrade paper (Ilford)
a) Filter 0 (extra soft, yellow filter alone); b) Filter 1; c) Filter 2; d) Filter 3; e) Filter 4; f) Filter 5 (extra hard, magenta filter alone)

Matt surfaces are obtained by adding starch or plastics (e.g., ethylcellulose [495] or polyvinyl compounds such as polyacrylonitrile, polystyrene, poly(methyl methacrylate) and poly(vinyl acetates) [496]) to the emulsion or protective layer before coating. In resin-coated papers (see Section 3.2.2.2), *textured surfaces* are obtained by imparting a fine pattern to the polyethylene coating on the emulsion side with a calender.

Most photographic papers, after wet processing, are dried by the application of heat. *Baryta- coated papers* are pressed between a

stretched cloth apron and a metal surface heated to 80 – 100 °C until completely dry. High-gloss surfaces are obtained by pressing the emulsion side against a highly polished stainless steel surface that is either cold or else heated to 80 – 100 °C (glazing). Papers with matting agents in the emulsion or protective layer must have the emulsion side turned toward the apron. Heat drying requires heat-resistant hardening of the emulsion layer, but above all of the protective layer. Protection against a loss of the deep blacks in glazing (burning of the blacks) is afforded by heterocycles with sulfur- containing groups, such as mercaptobenzothiazole [497]. *Resin- coated photographic papers* are moved continuously through a hot-air or infrared drying zone, where they make no contact with the heating elements, or else they are dried in air.

The following methods, singly or in combination, are used in order to obtain the photographic result more quickly.

Increase in Developing Rate. Extra-short developing times are achieved if developing agents are incorporated in the layer structure. Such papers, after exposure, are mechanically dipped into a caustic alkali – sulfite solution (activator) for a few seconds, squeezed between rubber rollers, treated with a stabilizer, and again passed through rollers. They are then subjected to a (usually automatic) mechanical heat drying. The stabilizer bath contains substances that form nonphotosensitive silver complexes. Developing machines for this activator-stabilizer system are also called "two-bath" machines. The processing time for an 18 × 24 cm paper is ca. 60 s. These papers can be held for only a short time after processing. However, they have a lifetime of a few years if they are additionally fixed and washed and if care is taken that they are not exposed for an extended time to bright light (especially sunlight), excessive heat, or high atmospheric humidity.

Shortening of Washing and Drying Times. The washing time can be cut without impairing the lifetime of paper prints only if it is impossible for chemicals of any kind, but especially thiosulfate, to penetrate into the fiber base. This can be achieved by sealing the paper base on both sides with a plastic resin (resin- coated papers). Reagents can then be removed from the thin emulsion layer by washing for 1 – 5 min; washing of conventional papers requires 10 – 30 min, depending on the thickness of the paper and the efficiency of the washing equipment.

If a combination of developing agents in the photographic layers and resin- coated paper is used, papers can be processed in very compact roller-type developing machines with activating, stabilizing, fixing, and washing baths (four-bath equipment). Such papers have a lifetime of 5 – 10 years; the processing time for an 18 × 24 cm print is ca. 60 s. Because there are a number of machines (Kodak Royal Print) using this system, especially on the U.S. market, both standard B & W papers and variable-contrast B & W papers with incorporated developers on RC bases are available. If a roller-type developing machine with the conventional sequence of baths (developing, fixing, washing) is used for these papers, archival-quality prints are obtained in ca. 4 min.

Manufacturers of black-and-white papers include: Agfa – Gevaert, Leverkusen, Germany; Argenta, Munich, Germany; Fuji, Tokyo, Japan; Ilford, London, United Kingdom; Kodak, New York, USA; Konica, Tokyo, Japan; Labaphot, Flörsheim, Germany; Mitsubishi, Tokyo, Japan; and Oriental, Tokyo, Japan.

8.3.1.2. Color Papers

All color papers use the subtractive color process so that the finished print will have whites as bright and colors as brilliant as possible (see Sections 5.1 and 5.2). *Color negative papers* require a negative film from which to print. The class of *color reversal* (*color positive*) papers includes all materials that are printed from a positive (e.g., a color reversal film or a color print). Because the coating structures and processing are largely identical, display films (used to make large color images for viewing in transmitted light) are discussed together with color negative papers. For the same reasons, overhead film is grouped with the color reversal papers.

Image formation in color papers virtually always involves the creation of yellow, magenta, and cyan dyes (*chromogenic process*, see Section 5.2.2) through exposure and development. The *silver-dye bleach process* is still important in the small color reversal paper market but only in the professional sector. This process offers outstanding color brilliance, and it formerly yielded

images with better light stability than the chromogenic processes (for a more detailed account of this material, see Section 5.4).

Color Negative Papers. Figure 94 illustrates the layer structure of color negative papers. To achieve rapid processing, all supports are resin-coated paper (see Section 3.2.2.2), which is available with a variety of surface finishes, thicknesses, and backings (normal, self-adhesive, or suitable for writing with a pen). Color negative paper on a white-pigmented polyester base has been on the market for some years; it offers not only good dimensional stability but also an extraordinarily glossy finish.

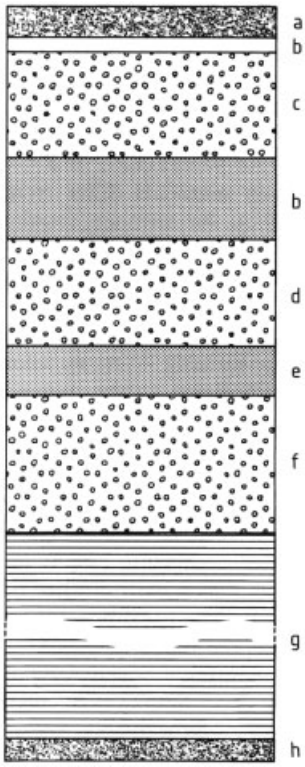

Figure 94. Layer structure of color negative paper (schematic)
Total thickness exclusive of base, ca. 10 µm. Base thickness, ca. 230 µm with ca. 170 g/m^2 paper base and ca. 30 g/m^2 polyethylene coated on either side.
a) Protective layer; b) Interlayer with UV absorber; c) Red-sensitive cyan layer; d) Green-sensitive magenta layer e) Interlayer; f) Blue-sensitive yellow layer; g) Base; h) Antistatic layer

The layers have essentially the same functions as in color negative film (see Section 5.2.2.6). The UV absorber protects the finished print against premature yellowing and color fading, but simultaneously impairs the action of the optical brighteners that are introduced during development or incorporated in the material during manufacturing. Optical brighteners beneath the UV absorber layers are less efficient while those above the color-producing layers give improved whites but cause the blacks to exhibit a bluish fluorescence.

For economic reasons, the emulsion layers are single, with a total silver coating weight of 0.6 – 0.9 g silver per square meter. The image dyes must exhibit as little secondary absorption as possible in order to prevent darkening of the colors. They should be stable for as long as possible regardless of whether the prints are stored in the dark (album) or in the light, so that the information content of the images is not lost within a few years. More than 100 years dark stability is claimed for modern color prints (see Section 6.4); a color print can hang for ca. 30 years on an interior wall not exposed to direct sunlight before it shows unacceptable color shifts [498].

The spectral sensitization of the emulsions is matched as exactly as possible to the absorption maxima of the image dyes in the color negative films; color reproduction is optimal when the sensitization curves have the least possible overlap. For the reasons discussed in Section 5.2.2.6, there is no need for a yellow filter. The granularity of color papers is not important; sharpness is enhanced by screening dyes, which are decolorized in processing. These dyes absorb scattered light during exposure, which decreases the sensitivity. There is as yet no standard for the speed of color papers; if, for example, a density of 0.6 above base plus fog is used as the sensitivity point, color negative papers have a speed of ca. ISO 5/8°.

The positions of the layers relative to the base virtually do not influence the sensitivity and sharpness of a color negative paper.

The red-sensitive cyan layer (see Fig. 94) is on top, because the dyes formed in this layer show the least light fading and the direct contact of this layer with fresh bleach-fixing bath (lowest Fe^{2+} concentration) counteracts any loss of cyan density due to leuco dye formation. The yellow layer, at the bottom of the stack, com-

pensates for some roughness of the base; the eye, which is less sensitive in this part of the spectrum, does not immediately register the density variations that result after development.

All color papers must be stored under low-temperature conditions (below 10 °C) until used.

Along with the most faithful reproduction of all colors, a crucial factor in the quality of a color paper is optimal reproduction of neutral grays without any hue. If a continuous gray wedge is exposed onto a color negative paper, the density curves for the three image dyes must virtually coincide given proper exposure and filtering (see Section 5.6). Because yellowish whites and brownish-red blacks do not suit the common tastes, the yellow density in minimum-density regions is less than the densities of the other colors, while the cyan density in high-density regions is greater than those of the other colors. This balance is established in order to avoid to a certain extent undesirable toning of white and black due to possible storage conditions (e.g., interruption in low-temperature storage) or processing effects (e.g., leuco dye formation).

For economic reasons, a given grade of color paper should give optimal results for all exposure times occurring in practice. For color negative papers, this means that optimal sensitometry must be attainable between times of 20 ms (high-capacity printer) and several minutes (large-format prints). This requirement presents a major technical challenge: The three emulsions in a color paper, which differ in speed and spectral sensitivity, must exhibit no reciprocity failure (or at least identical reciprocity failures; see Section 2.1.6) over this wide range of exposure times. If this is not the case, a type of paper must be subdivided into a short-exposure material and a long-exposure material. The latent image (see Section 2.1.4) in color papers must remain stable over a range from several seconds up to 24 h.

Color negative papers now on the market include those with *silver bromide emulsions* (> 70 % silver bromide), which are processed in 81/2 min (to the end of washing) if the chemistry corresponds to the Kodak EP-2 standard (see Section 5.3). These papers are offered in a normal version for amateur and industrial photography and a professional version with a somewhat flatter gradation (and therefore used especially for portrait work). As in the case of speed, there is no standard governing the measurement of contrast in color negative papers. If the standard for B & W paper is followed, the standard color negative paper has a log *ER* of ca. 1.0. Professional papers have log *ER* values of ca.1.2.

Silver bromide paper is increasingly being replaced by a material with*pure silver chloride emulsions* (> 98 % chloride), which is processed in just 3 min (to the end of washing) if the chemistry corresponds to the Kodak RA-4 standard (see Section 5.3). Again, a variety of gradations are available, but the material has had to be divided into short-exposure and long-exposure types.

Display Films. Display films are used for making large color images for transmission viewing. The color-producing layers have the same components as the corresponding color negative papers. These materials are therefore processed in the same baths used for papers (EP-2 or RA-4 standard). Transmission materials differ from paper prints in that they, for instance, require about twice as much dye for a brilliant image. In order to obtain optimal color reproduction, high color saturation, and deep blacks, display films must therefore be made with more than twice the coating weight of color-forming layers as color negative papers. Accordingly, they require longer processing times and large amounts of replenishers (see Section 5.3).

Figure 95 shows the coating structures of the two types of display films on the market. Figure 95 A illustrates the clear variant, which is completely transparent after processing and is intended for display cases with a matt diffuser or film. The type shown in Figure 95 B, with an opaque titanium dioxide coating, is designed for display in the case of no diffuser. The density of the TiO_2 layer is as low as possible so that a very good brightness range can be achieved, but as high as necessary so that the light is uniformly scattered and the light bulb behind the display cannot be located. The opaque layer does not, however, reflect nearly as much as a color negative paper base, so that the image in reflected light appears relatively soft.

The antihalation layer has the same purpose as in color negative films (see Section 5.2.2.6). The additional NC function of the antistatic coating is necessary because the polyester film used as base cannot adequately cope with the tensile

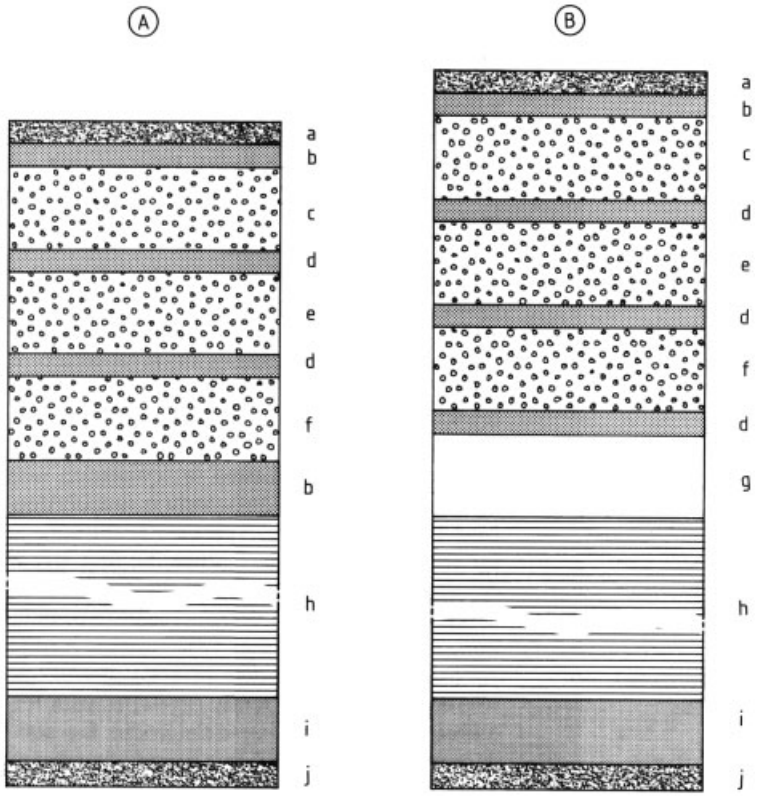

Figure 95. Layer structure of display films (schematic)
A) Clear type, total coating thickness ca. 208 µm; B) Trans type, total coating thickness ca. 218 µm (base is PETP film ca. 175 µm thick)
a) Protective layer; b) Interlayer with UV absorber; c) Red-sensitive cyan layer; d) Interlayer; e) Green-sensitive magenta layer; f) Blue-sensitive yellow layer; g) Titanium dioxide layer; h) Base; i) Antihalation layer; j) NC and antistatic layer

forces exerted by the emulsion layers and interlayers applied on top of it.
Manufacturers of color negative papers and display films include Agfa, Fuji, Kodak, Konica, Mitsubishi, and Oriental.

Color Reversal Papers. All color reversal papers described here employ the chromogenic process. Everything that has been said about bases, UV absorbers, optical brighteners, image dyes, sharpness, storage conditions, the reciprocity and latent-image behavior of the emulsions applies equally to these materials. The essential difference lies in the silver halide emulsions and in the layer structure (Fig. 96), which differs from that of color negative papers.

The emulsions used are of the silver bromide type containing a few percent of iodide; these give the strongest possible interimage effects. Interimage effects are always desirable in color prints, but whereas printing from modern color negative films can take advantage of masking (to compensate for secondary densities) and strong interimage effects (see Section 5.2.2) incorporated in the films, the originals used with color reversal papers feature no masking and only weak (color reversal film) or no (color print) interimage effects. For this reason, interimage effects in color reversal papers must make a substantial contribution to the color saturation.

In principle, the blue-sensitive and green-sensitive emulsions can employ the same sensitizing dyes used in color negative papers, but the maximum sensitivity of the red-sensitized emulsion must be at a shorter wavelength (ca. 650 nm) in order to ensure optimal spectral matching to the original. As in the case of color negative pa-

per, no standard exists for determining the speed. If the standard for color reversal film is applied, then color reversal papers have a speed of ca. ISO 12/12°.

Color reversal papers have a markedly flatter gradation than color negative papers, in accordance with the Goldberg condition (see Section 6.1.5). The exposure range is about 20 DIN (versus a maximum of 12 DIN for color negative papers), so that the blue sensitivity cannot be neglected when silver bromide emulsions are used. With regard to the layer arrangement (Fig. 96), this has the consequence that the blue-sensitive yellow layer is placed on top. As in color films the yellow filter layer that follows prevents the unwanted exposure of the greensensitive and red-sensitive layers in their intrinsic sensitivity bands.

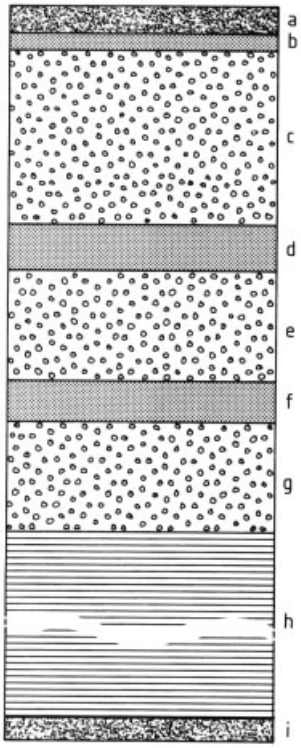

Figure 96. Layer structure of a color reversal paper (schematic)
Total coating thickness exclusive of base, ca. 13 µm. Base thickness, ca. 240 µm with ca. 180 g/m^2 paper base and ca. 30 g/m^2 polyethylene coated on either side
a) Protective layer; b) Interlayer with UV absorber; c) Blue-sensitive yellow layer; d) Yellow filter layer; e) Green-sensitive magenta layer; f) Interlayer; g) Red-sensitive cyan layer; h)Base; i) Antistatic layer

The silver coating weight of color reversal papers is ca. 1.3 g silver per square meter, i.e., higher than that of color negative papers. The reason is that up to 30 % of the silver halide grains are developed in the first B & W developer, even without exposure, so that this fraction makes no contribution to the maximum density in the subsequent chromogenic development (see also Section 5.3).

At present, two types of color reversal papers are on the market, namely papers with normal and those with a higher contrast. Color reversal papers with normal contrast are suitable for printing from color reversal films. If, for lack of a standard, the method used for B & W papers is employed for these materials, the log *ER* value is ca. 1.9. The type with higher contrast (log *ER* around 1.7) is available for copying color prints, because the density range of prints is less than that of color reversal films.

Overhead Films. Similar to the relationship between display film and color negative paper, overhead film is a special material related to color reversal paper. Regarding the layer structure, see Figure 96 for layers above the base and Figure 95 for the base plus backings. The basic components are the same as for color reversal papers. For the same reasons cited in connection with display film the emulsion content is roughly doubled. Processing, however, is exactly the same as for color reversal papers, since the first B & W developer is active enough and the chromogenic development is already designed to go to completion.

Processing. Both color reversal papers and overhead films are currently developed with formulations corresponding to the Kodak R-3 standard (processing time 10 min to end of washing). The situation is quite different for color reversal papers working with direct positive emulsions (see Section 3.1.1). These papers can be processed in the same series of baths (see Section 5.3) used for color negative papers. This comparatively simple process is well-suited to use in an office color copier (Konica) or in a multipurpose machine (Fuji), similar to a copier, that can print from color negative and reversal films as well as positive color originals. Either a color negative paper or a color reversal paper with direct positive emulsions is employed; either type passes through the same processing system.

Manufacturers of color reversal papers and overhead films include Agfa, Fuji, Kodak, and Konica.

8.3.2. Special Applications

Papers for Coin-Operated Photo Booths. In photo booths, highly sensitive photographic papers are automatically exposed, for example by a flash tube, and automatically processed to a finished positive image. These coin-operated machines usually employ reversal development without reexposure. Papers suitable for this application have a high-speed, soft-working silver bromide emulsion, which can also be sensitized for long-wavelength light. Some brands of B & W paper for reversal development in coin-operated machines exhibit fogging when the first development is carried out to obtain the negative. In this way, the shadows show up better in the final positive image. In strongly illuminated areas, which will be white tones in the final image, undevelopable residual silver halide is left behind after the first development. This residual halide would react with sodium sulfide later, in reversal processing, forming a brown fog and thus degrading the brightest parts of the image. Undevelopable silver halide is therefore removed by the addition of sodium thiocyanate to the first developer.

After development, the material is washed briefly and the reduced silver is removed with dichromate solution acidified with sulfuric acid. After again washing and clearing in sulfite solution, the residual silver bromide still present in the emulsion is directly converted to a positive image without reexposure. This is carried out with a sulfur or selenium toning bath (usually sodium sulfide solution with sublimed sulfur or selenium dissolved in it). Another brief wash follows, and the pictures are then dried in a stream of heated air. The entire processing operation takes some 4 – 5 min.

An aim is to simplify the processing and reduce the number of steps to two, by incorporating developing agents into the emulsion and modifying the bath composition. This kind of rapid processing is effective if the paper base has a plastic resin coating on both sides, so that the chemicals in the various baths cannot penetrate into the paper.

Reversal papers for coin-operated booths are *manufactured* by Agfa – Gevaert (Reversal Waterproof), Chemco (Photomaton), and Eastman Kodak (Direktaprint).

In addition to the coin-operated B & W systems just described, there are also systems for making color photos. A positive image is obtained by the reversal, direct positive, or silver-bleach processes (see Section 5.4).

Papers for Nondestructive Testing [471]. Not only X-ray films but also X-ray papers are employed in materials testing. The latter do not reproduce details as well as films, but they are much less expensive and quite suitable for test exposures and documentation. X-ray papers are coated with a high-speed-blue-sensitive silver bromide emulsion. A blue-fluorescing intensifying screen is used with the papers, as with films (see Section 8.2.5). The emulsion is therefore only sensitized in the wavelengths where this fluorescence takes place.

In Structurix IC paper (Agfa – Gevaert), the emulsion layer, supported on a polyethylene-coated base, contains developing agents. With the help of a developing machine employing the activator-stabilizer principle (see Section 5.3.3), usable images can be obtained in 10 s.

Photographic Chart Recording Papers [466], [499]. In medicine and in technology, fast oscillatory processes must often be recorded, or a number of processes must be recorded at once in parallel. "Direct" chart recorders cannot perform these tasks either because of their inertia or because the recording devices mechanically interfere with one another. Optical recorders are used instead. Relative writing speeds of up to ca. 2200 m/s are attained (depending on the oscillograph type and light source). In general, recording papers must show relatively great blackening even after minimal exposures, but the blackening should not increase much with greater exposure, so that both rapid zero crossings and relatively slow spot movement at the extrema (amplitudes) are equally well registered. Most such papers have high-speed emulsions with steep gradation and correspondingly reduced silver content. Bases vary widely; for example, paper with a weight of ca. 90 g, with or without baryta coating, and 40 – 50 g papers with the lowest possible opacity. There is also

a wide variety of packagings with regard to roll length, width, and core diameter. Some widths are perforated on both edges. The longitudinal lines (usually spaced 2 mm apart) required for some applications are printed during packaging. So that the oscillograms can be examined as soon as possible after exposure in the recording camera, a variety of approaches to shortening the wet processing have been proposed.

A combination of a recorder and processor may be used for rapid access so that the processing automatically follows after the exposure. Some of these developing papers contain developing agents making then suitable for activation and stabilization processing.

Recording without processing is possible with special recording papers, e.g., Linagraph (Kodak) and Oscilloscript (Agfa – Gevaert). A second exposure of the paper to roomlight of relatively low intensity is sufficient to produce a visible image with a maximum density of 0.4 – 0.75 on a sligthly tinted background. Inspection can occur immediately after the paper has left the recording unit and the delay due to processing is avoided. The latent image of the trace exposure is latensified (see Section 2.1.6) i.e., photodeveloped by this second exposure. The emulsions are orthochromatically sensitized to absorb as many photons as possible from the writing beam. Halogen accepters are added in order to obtain a high print-out density and also to protect to some extent the image.

Latensified oscillograms have an unlimited lifetime if stored in a dark, dry place and protected from heat. Exposure to direct sunlight or light with a high content of UV should be minimized.

If color printing paper is employed, traces in different colors can be obtained. For every channel a specific color filter must be inserted. A disadvantage of this process is the four-step processing.

Electron-Beam Recording. Instead of recording the traces of a moving light-beam the virtually inertialess displacement of the electron-beam from a cathode ray tube (CRT) may be recorded as follows:

1) Direct exposure of the film to the electrons
2) Photographing the fluorescent trace on the screen of a CRT

Direct Exposure. This method is less common because the technique is more difficult; however, the resolution is much better. The CRT for direct exposure is not a standard type, but is provided with a film chamber and a device to transfer the sensitive material from atmospheric pressure into vacuum and back, with minimum leakage.

Direct exposure avoids the energy losses and other limitations of CRT screens. Moreover, exposure of the film in a camera in indirect exposure introduces additional losses, requiring high speed material. In contrast, even rather slow materials are suitable for direct exposure, further improving the resolution. Films and papers can be chosen from the general and technical assortment. Electron-beam recorders and electron-sensitive dry silver materials for thermal processing are supplied by 3 M.

Indirect Exposure. The more common indirect method of photographing the CRT image has many applications in addition to oscillography, e.g., in phototypesetting (→ Imaging Technology, Chap. 3.1.1.), computer output microfilm (see Section 8.2.4), radar image recording, television screen photography etc.

The intensity and spectral composition of the output radiation of the CRT determine the choice of the recording material, which must have a well-matched speed and spectral sensitivity. The recording material has high contrast and high density. For some applications the recording of traces on paper or film can be replaced by other techniques such as transient-recording and realtime digital signal processing. Nevertheless, photographic recording remains in use for prolonged monitoring and fieldwork such as surveying the railroad-system, i.e., checking the flatness of the rails. An additional advantage is the convenient measurement and comparison of a collection of different records.

A great variety of photosensitive materials (paper and film) are available, among them Polaroid instant sheet or roll film and dry silver (photothermographic) materials supplied by 3 M. The latter require only a single heat-treatment to yield an image. Materials for wet processing are mostly adapted to the conditions of rapid access by prehardening.

Papers for Line and Screened Continuous Tone Originals [500], [501], [43]. In order to

make more efficient use of the light source, these papers generally have orthochromatic sensitization and steep gradation. They can be exposed either in contact or optically, in an enlarger or a repro camera. Optimal sharpness is achieved with a vacuum back and a point light source. Paper base weights range from 100 – 140 g/m^2. Most papers are processed by the activator – stabilizer method in compact roller-transport machines.

Papers with relatively steep gradation and high speed are used for reenlargement, such as from microfilm line negatives.

Photographic materials with widths up to and beyond 1 m are used in industrial drafting offices and also for photogrammetry and topographic work. These materials are intended for machine processing. Depending on the specific application, papers can be written on or retouched with pencil or ink. They have a special backing to improve flatness under various humidity conditions. Steeply graded papers of this type are used in negative and positive contact copying of text and line originals and in the preparation (by the reflex method) of negatives from opaque originals or those printed on both sides. Since these applications sometimes involve exposure through the paper base, uniform density of the cellulose mass is subject to especially stringent requirements so that shadows will not appear in the emulsion layer.

Reflex copies are made by placing the photographic paper with its emulsion side against the original to be copied, pressing the two together, and passing light through the photo paper. The light transmitted through the photo paper is reflected from white areas of the original and largely absorbed in dark areas (print, etc.) of the original; the action of the light on the emulsion is thus stronger next to white areas. The slight differences in exposure intensity (1 : 2 arithmetic) in the emulsion can produce a marked difference in blackening in the negative only if the emulsion is particularly hard-working. Contact photographic copying papers accordingly span an exposure range of 1 : 2 – 1 : 3 (arithmetic).

Producers of an extended assortment are Agfa – Gevaert (Copyline, Rapidoprint) and Kodak (Kodaline, Kodagraph, Kodak Rapid).

Direct Positive Papers. These papers give after exposure and normal processing, without any intermediate steps, a replicate of the original as opposed to the positive images resulting from reversal processing (see Section 4.3.3.7) and diffusion transfer (see Section 4.4 and → Imaging Technology, Chap. 2.2.3.1.). However, the material must be exposed in cameras provided with a prism to obtain a correct orientation. The direct positive emulsions are often based on the solarization (see Section 2.1.6), i.e., bleaching, of prefogged grains. Otherwise, the Herschel effect (Section 2.1.6) may be applied. Suitable low-speed silver chloride emulsions containing a sensitizer [502] can be fogged with precisely metered UV light after coating on the paper and before cutting and packaging. The emulsion can also be chemically fogged before coating, e.g., with thiourea dioxide at elevated temperature and pH > 8 [503]. Yellow filters ($\lambda > 500$ nm) are used for copying by the transmission or reflex method.

Direct positive-working papers with high speed and relatively steep gradation are employed in process camera work with line originals and changes of scale. Most such emulsions are coated on polyethylene- coated bases so that processing can be done in machines. They are developed in lith developers or fast Phenidone – hydroquinone developers.

Some materials are suitable for halftone images. Manufacturers are, e.g., Agfa – Gevaert (Autoreversal Rapid, Copyline Autreversal) and Kodak (Kodagraph Autopositive, Kodaprove).

8.4. Holography and Holographic Recording Materials (→ Holograph

8.4.1. Introduction

Holography was discovered by D. Gabor [504] in 1948 – 1949. At the time, no sufficiently monochromatic light source existed, and only small objects could be represented as holograms. Holography began to advance only after the invention of the laser in 1960. Gabor himself had made "in-line" holograms, in which the virtual and real images are not separated on reconstruction (reference and object beams coaxial). Leith and Upatnieks [505] introduced the obliquely incident reference beam, thus separating the virtual and real images.

8.4.2. Principles of Holography [48–50]

According to the Huyghens – Fresnel principle, every point of an illuminated object can be regarded as the point of origin of a spherical wave, which can interfere with all other such spherical waves originating from other points of this object. This principle leads to a simple treatment of holography. Figure 97 shows an object point (c) from which a spherical wave is propagating. At the recording point, this spherical wave is superposed on a plane wave (reference wave) which is produced with a mirror (b). Since each of these waves is coherent (monochromatic), an interference pattern (hologram) is generated; this is recorded by the photographic plate (d). If this hologram (b), after processing, is illuminated with the reference wave alone (Fig. 98), the object point (c) is reconstructed by diffraction. Besides this virtual image, which is seen directly by the eye, a real image (which can be examined on a ground glass) and an undiffracted component (g) are also produced. In the language of diffraction theory, the grating generates orders − 1, 0, and +1. Since every point of a complicated object gives rise to its own diffraction grating in this way, on reconstruction a great number of points are generated, creating an image of the object in three dimensions.

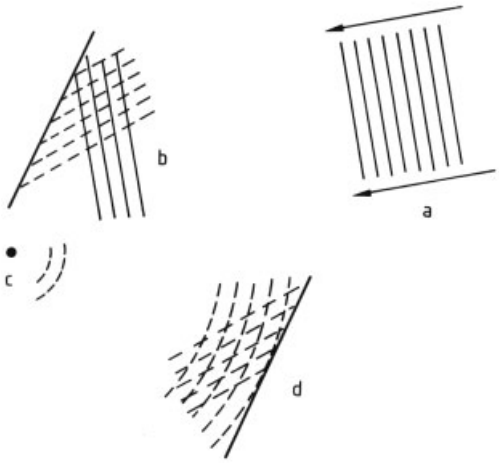

Figure 97. Recording a hologram of an object point (spherical wave)
a) Laser beam; b) Mirror (reference beam); c) Object point; d) Holographic recording medium

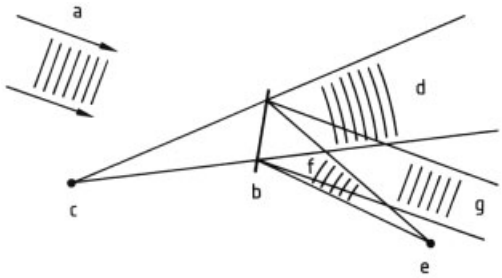

Figure 98. Reconstruction of the hologram with an object point
a) Laser beam; b) Hologram; c) Reconstructed object point (virtual image); d) Divergent wave (primary image); e) Real object point (real image); f) Convergent wave (conjugate image); g) Undiffracted beam

A spherical wave gives rise to the Fresnel lens (Fig. 99), which is a plane object having the properties of a lens (in-line technique).

Figure 99. Fresnel zone lens
The radius of each circle is proportional to $n1/2$, where n is an integer.

The information about an object is stored in the hologram not as a real image but as a wave field. Cutting the hologram up into separate pieces therefore does not erase part of the information but gradually diminishes it because the object can then be examined only through a correspondingly smaller window.

Because the hologram embodies the wave field emitted by the object at the time of exposure, this field can be made to interfere with the wave field of the same object at a later stage. If the object has undergone some slight change in the meantime, for example through thermal expansion or warping, then the superposition field will contain interference fringes, whose number and relative positions allow quanitative de-

termination of the magnitude of the change in the object. The fact that wave fields existing at different times can be coherently superposed is of enormous importance for the use of interferometry in measurement. Special interferometric measurements are possible only with holography. The superposition of wave fields can also be accomplished by multiple exposures on a single hologram.

Two types of holograms exist: absorption and phase holograms. In *absorption holograms*, the developed density distributions carry the information. By a bleaching process, the density can be converted to a pattern of "phase structures", that is, regions differing in refractive index, which are insensitive to light as far as possible (*phase holograms*) [506]. Alternatively, the residual silver halide is used in "reversal" processes. Another possibility for production of phase holograms is to create relief structures, for example by tanning development. Finally, thick and thin holograms are distinguished. In order to obtain optically thick holograms, as much as possible of the light in the reference wave I_r for reconstruction must be diffracted into the first order I_1. The ratio $\eta = I_1/I_r$ is called the diffraction efficiency. Table 13 lists values for the most important types. With thick phase holograms, reconstruction with white light (from the sun or a sufficiently point-like source) is possible. The carrier frequency of the hologram (Bragg lattice) itself generates the needed monochromatic light. In transmission holograms, the reconstructed image appears in transmission, and for reflection holograms in reflection.

Holograms can also be copied, so that mass production is possible. An example of a mass-produced hologram is the reflection hologram in a check card or credit card, which is difficult to falsify.

8.4.3. Recording Materials [464], [507]

Because the interference pattern of a hologram in general contains structures with dimensions of the order of the wavelength of the radiation used, materials for recording must have extremely high resolving power.

In addition to photographic materials, other recording media have been proposed. Examples are thermoplastics, photopolymers, bichromated gelatin, and photoresists. Because of their high sensitivity, however, silver halide photographic materials are still preferred. Lippmann emulsions are used; i.e., photographic plates with silver halide grains having diameters substantially smaller than 0.1 µm (0.03 – 0.09 µm).

Such emulsions only have a low light-scattering action. Because of their small silver halide grains, however, the emulsions have relatively low sensitivities, so that practical holography often requires exposures lasting from some seconds to minutes. A change in position (e.g., of the object or the plate) on the order of the wavelength would completely smear out the detailed interference pattern, and so the requirements on the mechanical stability of the apparatus used for recording holograms are rather stringent.

Several firms manufacture special *materials* for holography; examples are Agfa – Gevaert, Kodak, and Ilford. The materials have glass or PETP bases. Table 14 describes materials offered by Agfa – Gevaert and Kodak. Holographic recording materials have a steep gradation, $\gamma = 5 - 8$, and extremely high resolving power.

The diffraction efficiency at high spatial frequencies is determined with diffraction optics. Figure 100 A shows schematically the process of recording two-beam interferences and the reconstruction of the hologram after processing of the photographic material.

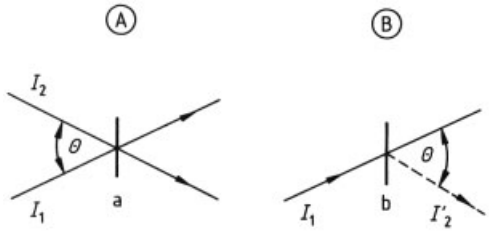

Figure 100. Recording and reconstruction of a hologram; two beam interferences
A) Recording; B) Reconstruction
a) Photograph plate; b) Interference pattern

In recording, two plane waves with intensities I_1 and I_2 are incident on the photographic plate; the directions of incidence are symmetrical with respect to the normal. The angle between the two beams is Θ. The spatial frequency ν is then

$$\nu = \frac{2}{\lambda} \sin \frac{Q}{2}$$

Table 13. Diffraction efficiencies (%) of different types of holograms

Diffraction efficiency values	Amplitude hologram			Phase hologram		
	Thick		Thin	Thick		Thin
	T [a]	R [b]	T a.R [c]	T [a]	R [b]	T a.R [c]
Theoretical	3.7	7.2	6.3	100	100	33.9
Observed	3.0	3.8	4.0	95	60	33.0

[a] Transmission hologram.
[b] Reflection hologram.
[c] Transmission and reflection holograms.

Table 14. Special materials for holography [467]

Supplier	Material	Spectral sensitivity	Resolving power by diffraction, lp/mm ∗
Agfa – Gevaert	Holotest 8 E 56 HD	blue and green	5000
	Holotest 10 E 56	blue and green	3000
	Holotest 8 E 75 HD	blue and red	5000
	Holotest 10 E 75	blue and red	3000
Kodak	649 F	panchromatic	4000
	120 – 01/02	blue and red	4000
	125 – 01/02	blue and green	1600
	131 – 01/02	blue and green	1600

∗ lp/mm=line pairs per millimeter.

where λ is the wavelength of the incident light in air ($\lambda = 633$ nm for the He – Ne laser).

If $\Theta = 90°$, the spatial frequency is thus $\nu = 2235$ lp/mm (line pairs per millimeter). The spacing between two lines, found as the reciprocal of the spatial frequency, is ca. 0.45 µm.

Table 15. Relative toxicity of silver compounds in comparison to the free silver ion [509] (30-day fathead minnow embryo-larval test)

Substance	Relative toxicity
Ag^+	1 ∗
AgCl	0.003
Ag_2S	$< 0.7 \times 10^{-4}$
Ag-thiosulfate complex	$< 0.6 \times 10^{-4}$

∗ The MATC (maximum acceptable toxicant concentration) reported for the free silver ion from tests with fathead minnows was between 0.00037 and 0.00065 mg/L.

The modulation m depends on the polarization of the laser light and the ratio of the intensities I_1 and I_2. In the case shown in Figure 100, the laser light was linearly polarized with the electric vector perpendicular to the plane of incidence. The intensity ratio I_1/I_2 was taken as 0.25, so that the corresponding modulation factor was 0.8. In general, for this polarization,

$$m = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}$$

The process of reconstruction is illustrated in Figure 100 B. Beam I_2' was reconstructed with beam I_1. The intensity I_2' diffracted by the grating into the first order, and thus the ratio I_2'/I_1, depend on the angle Θ, the modulation m, and the exposure.

9. Environmental Aspects of Photographic Processing

Silver. Because *free silver ion* is highly toxic to bacteria and fish, silver is classified as an environmentally hazardous heavy metal in all countries [508]. Free silver ion occurs only in emulsion making but not in the downstream production steps, in the photographic materials, or in the processing baths. In countries where wastewater treatment is advanced, the silver present in wastewater appears as silver sulfide in the clarifier sludge.

The aquatic toxicities of the *silver thiosulfate complex* and of the extremely sparingly soluble

silver sulfide are significantly lower than that of free silver ion (see Table 15) [509], [510].

Silver releases from photographic processing are chiefly from fixing baths, but also from bleaching baths and rinse water. Fixing and bleach – fixing baths are forwarded to electrolytic silver recovery in the processing plant itself or at a recovery firm. The desilvered fixing baths can be further treated with sodium sulfide or other precipitating agents, such as trimercapto-s-triazine trisodium salt [*17766-26-6*] [511]; the residual silver and other heavy metals are recovered as sulfides. After ultrafiltration, baths treated in this way are disposed of as wastewater (after further treatments as necessary). Other treatments include:

1) Precipitation of sulfate with milk of lime
2) Reduction of COD (with H_2O_2 or ozone; see below)
3) Reduction of ammonium load by stripping of ammonia in alkaline medium

Desilvered baths and developers are often burned in hazardous-waste incineration facilities. Before shipping to the incineration plant, water can be evaporated in (vacuum) evaporators to reduce the volume. Black-and-white images that are no longer needed (e.g., X-rays) are incinerated in special furnaces; the ashes are processed by special silver recovery firms.

Regulatory limits for the release of silver into sewer systems differ widely from country to country; often they are established at the municipal level. In general, limiting levels are 0 – 3 mg silver per liter. (In some countries, regulations do not cover the release concentration but rather the silver load that can be released per square meter of material processed.)

These low silver concentrations in wash water can be achieved through

1) Low silver content in the material
2) Low silver content in the fixing bath
3) Low carryover from fixing bath into wash water (low swelling of material, good stripping equipment)
4) Treatment of wash water on ion exchangers [512]

When these practices are employed, limits of 1 – 2 mg silver per liter of wash water can currently be attained. Mixing of bath residues and overflows with rinsing water is prohibited in most countries.

Other Environmentally Relevant Bath Constituents. *Cadmium*, formerly used in warm-tone papers and special-purpose films, is no longer in use for environmental reasons. Some *developing agents* have aquatic toxicities and biodegradability is difficult (Phenidone, CD-3, CD-4) [513], [514]. It has been found, however, that no harmful effects appear at the concentrations that occur in the practice. CD-3 and CD-4 are retained in sewage plants (partly in chemically modified form) in such a way that they do not develop fish toxicity except at very high discharge concentrations [515]. *Ammonium thiosulfate* is undesirable in wastewater because of the fertilizing action of the ammonium and the high COD. Modern treatment facilities are able to eliminate this substance. Thiosulfate, like sulfate, can attack concrete sewer pipes. Release limits depend on the local circumstances (piping material, dilution in the sewer); levels < 400 mg sulfate per liter are commonly permitted.

The COD values differ sharply between photographic processing baths:

Fixing baths	40 – 100 g/L
B & W developers	10 – 150 g/L
Color developers	10 – 80 g/L
Bleaching baths	90 – 120 g/L

The COD of wash water is 500 – 3000 mg oxygen per liter. If local conditions make it necessary, aeration can be used to lower the COD of wash water by 10 – 20 %. Treatment with H_2O_2 [516] or ozone [517] is also technically feasible, but normally not used in practice.

10. Economic Aspects [518]

Table 16 shows annual sales (ex factory) of photochemical products for the Federal Republic of Germany and worldwide. The category of photochemical products includes not only films, papers, and processing chemicals for amateur and professional photographers, but also specialty products such as X-ray films, motion picture films, microfilms, sensitized graphic arts products, and printing plate materials.

Table 16. Sales (ex factory) of photochemical products

Year	Manufacturer sales, Federal Republic of Germany, 10^9 \$	Manufacturer sales worldwide, 10^9 \$
1986	1.260	19.844 *
1987	1.532	23.812 *
1988	1.703	26.273 *
1989	1.678	26.046 *
1990	2.095	29.700 *
1991	2.143 *	30.100 *

* Estimated.

The estimated breakdown in percent of the world photo market by sales regions is as follows (1990; total volume ca. \$ 75×10^9 ; includes phototechnical products such as cameras and printers):

North America	42
Western Europe	32
Other European countries	2
Japan	16
Other Far East countries	5
Other countries	3

11. References

General References

1. E. v. Angerer: *Wissenschaftliche Photographie,* 7th ed., Akad. Verlagsges. Geest & Portig., Leipzig 1959.
2. F. Bestenreiner: *Vom Punkt zum Bild – Entwicklung, Stand und Zukunftsaspekte der Bildtechnik,* Wichmann, Karlsruhe 1988.
3. E. J. Birr: *Stabilization of Photographic Silver Halide Emulsions,* Focal Press, London 1974.
4. H. Böttcher, J. Epperlein: *Moderne photographische Systeme, Wirkprinzipien – Prozesse – Materialien,* 2nd ed., VEB Verlag für Grundstoffindustrie, Leipzig 1988.
5. A. Boni: *Photographic Literature,* Morgan & Morgan, Dobbs Ferry, N.Y., 1962 and 1972.
6. E. Brinckman, G. Delzenne, A. Poot, J. Willems: *Unconventional Imaging Processes,* Focal Press, London 1978.
7. L. P. Clerc: *Photography, Theory and Practice,* 3rd ed., Pitman & Sons, London 1954.
8. R. J. Cox (ed.): *Photographic Processing,* publ. for the Scientific and Technical Group of the Photographic Society by Academic Press, London – New York 1973.
9. R. J. Croome, F. G. Clegg: *Photographic Gelatin,* Focal Press, London 1965.
10. J. C. Dainty, R. Shaw: *Image Science,* Academic Press, London 1974.
11. G. F. Duffin: *Photographic Emulsion Chemistry,* Focal Press, London 1966.
12. J. Eder: Handbuch der Photographie, 3rd ed., 1911/12, vol. 1; 1927/28, vol. 2; 1930/32, vol. 3; 1922/29, vol. 4; (altogether 17 vols.), W. Knapp, Halle.
13. C. E. Engel: *Photography for the Scientist,* Academic Press, London 1968.
14. H. Frieser, G. Haase, E. Klein: *Die Grundlagen der photographischen Prozesse mit Silberhalogeniden,* **3 vols.,** Akad. Verlagsges., Frankfurt 1968.
15. P. Glafkidès: *Chimie et Physique Photographiques,* 5th ed., **2 vols.,** Editions de l'Usine Nouvelle, Paris 1987.
16. Y. N. Gorokhovskii, T. M. Levenberg: *General Sensitometry,* Focal Press, London 1965.
17. A. Hay: *Handbuch der wissenschaftlichen und angewandten Photographie,* 8 vols., J. Springer, Wien 1929 – 1933.
18. K. I. Jocobson, R. E. Jacobson: *Imaging Systems,* Focal Press, London 1976.
19. T. H. James: *The Theory of the Photographic Process,* 4th ed., Macmillan, New York 1977.
20. C. E. K. Mees: *The Theory of the Photographic Process,* 2nd ed., Macmillan, New York 1954.
21. G. Joos, E. Schopper: *Grundriß der Photographie und ihrer Anwendungen, besonders in der Atomphysik,* Akad. Verlagsges., Frankfurt 1958.
22. M. J. Langford: *Advanced Photography,* 4th ed., Focal Press, London 1980.
23. L. F. A. Mason: *Photographic Processing Chemistry,* 2nd ed., Focal Press, London 1975.
24. K. Michel (ed.): *Die wissenschaftliche und angewandte Photographie* (Fortführung des Hay – von Rohr'schen Handbuches [A 17]), Springer, Wien. Vol. 1, J. Flügge: Das photographische Objektiv (1955); Vol. 2, J. Stüper: Die photographische Kamera (1962); Vol. 3, H. Weise: Die kinematographische Kamera (1955); Vol. 4, E. Mutter: Farbphotographie (1967); Vol. 5, E. Mutter: Die Technik des Negativ- und Positivverfahrens (1955); Vol. 6, H. Tümmel: Laufbildprojektion (1973); Vol. 10, K. Michel: Die Mikrophotographie (1957).
25. J. W. Mitchell: *Die photographische Empfindlichkeit,* O. Helwich-Verlag, Darmstadt 1957.
26. H. Schober: *Photographie und Kinematographie,* Verlag Kurt Wesemeyer, Hamburg 1957.

27. J. Sturge et al. (eds.): *Imaging Processes and Materials,* Neblette's, Van Nostrand Reinhold, New York 1989.
28. G. Teicher: *Handbuch der Fototechnik,* 8th ed., VEB Fotokinoverlag, Leipzig 1983.
29. H. J. Wallis, G. G. Attridge: *Basic Photo Science,* Focal Press, London 1977.
30. W. Walther: *Fotografische Verfahren mit Silberhalogeniden,* VEB Fotokinoverlag, Leipzig 1983.
31. V. L. Zelikman, S. M. Levi: *Making and Coating Photographic Emulsions,* Focal Press, London 1964.
32. Mitteilungen aus den Forschungslaboratorien der Agfa Leverkusen – München, 4 vols., Springer Verlag, Berlin 1955 – 1964.
33. Veröffentlichungen des wissenschaftlichen Zentral-Laboratoriums der photographischen Abteilung Agfa Wolfen, 8 vols., S. Hirzel, Leipzig 1930 – 1954.
34. *SPSE Handbook for Photographic Science and Engineering,* John Wiley & Sons, New York 1973.
35. R. W. G. Hunt: *The Reproduction of Colour in Photography, Printing and Television,* 4th ed., Fountain Press, London 1987.
36. G. Vieth: *Meßverfahren der Photographie,* R. Oldenbourg Verlag, München – Wien; Focal Press, London – New York 1974.
37. E. Goldberg: *Der Aufbau des photographischen Bildes,* W. Knapp, Halle 1925.
38. H. Frieser: *Photographische Informationsaufzeichnung,* R. Oldenbourg Verlag, München – Wien; Focal Press, London – New York 1975.
39. G. Pollakowski: "Die Archivierung von Silberhalogenid-Filmmaterial", in: *DEFA Beiträge zur Filmtechnik,* vol. 10, DEFA Zentralstelle für Filmtechnik, Berlin 1985.
40. Kodak: Pamphlet F-40: "Conservation of Photographs", Rochester 1985.
41. in [32]vol. 3, p. 215.
42. G. Schröder: *Technische Fotographie,* Vogel-Verlag, Würzburg 1981.
43. P. Glafkidès: *Chimie et Physique Photographiques,* 4th ed., Paul Montel, Paris 1976.
44. J. M. Sturge (ed.): *Neblette's Handbook of Photography and Reprography,* 7th ed., R. Van Nostrand, New York 1977.
45. L. Stroebel, J. Compton, I. Current, R. Zakia: *Photographic Materials and Processes,* Focal Press, Boston – London 1986.
46. R. H. Herz: *The Photographic Action of Ionizing Radiation,* Wiley Interscience, New York 1969.
47. G. Spitzing: *Grenzbereiche der Photographie,* Heering-Verlag, Seebruck 1968.
48. *SPSE Handbook of Photographic Science and Engineering,* J. Wiley & Sons, New York 1973; see also [42] pp. 79 – 88.
49. H. Kiemle, D. Röss: *Einführung in die Technik der Holographie,* Akademische Verlagsgesellschaft, Frankfurt 1969.
50. H. M. Smith: *Principles of Holography,* 2nd ed., J. Wiley & Sons, New York 1975.
51. W. Baier: *Quellendarstellungen zur Geschichte der Fotografie,* Schirmer-Mosel, München 1980.
52. J. S. Friedman: *History of Color Photography,* 2nd ed., Focal Press, London – New York 1968.
53. H. and A. Gernsheim: *The History of Photography,* Mc Graw-Hill, New York 1969.
54. G. Koshofer: *Farbfotografie,* vols. 1 – 3, Verlag Laterna Magica, München 1981.
55. B. Newhall: *Geschichte der Photographie,* special ed., Schirmer-Mosel, München 1989.
56. U. Tillmanns: *Geschichte der Photographie,* Huber Frauenfeld, Stuttgart 1981.
57. E. J. Wall: *The History of Three-Color Photography,* Photogr. Publ. Comp., Boston, MA, 1925.
58. "Pioneers of Photography, Their Achievements in Science and Technology", Proc. Int. Congr. **1st,** June 22 – 25, 1986, Rochester, NY, publ. by SPSE, Springfield, VA, 1987.
59. R. K. Hailstone, J. F. Hamilton, *J. Imaging Sci.* **29** (1985) 125.
60. P. Fayet et al., *Phys. Rev. Lett.* **55** (1985) 3002.
61. I. Konstantinov, J. Malinowski, *J. Photogr. Sci.* **23** (1975) 1.
62. I. Konstantinov, J. Malinowski, *J. Photogr. Sci.* **23** (1975) 145.
63. A. Milchev, S. Stoyanov, R. Kaishew, *Thin Solid Films* **22** (1975) 255.
64. A. Bonnisent, B. Mutaftschiev, *J. Chem. Phys.* **58** (1973) 3727.
65. E. Moisar, E. Klein, *Ber. Bunsen-Ges. Phys. Chem.* **67** (1963) 949.
66. in [14] vol. II, p. 612.
67. in [19] p. 88.
68. J. E. Maskasky, *J. Imaging Sci.* **30** (1986) 247.
69. R. W. Berriman, R. H. Herz, *Nature (London)* **180** (1957) 293.
70. E. Klein, H. J. Metz, E. Moisar, *Photogr. Korresp.* **99** (1963) 99.
71. E. Klein, H. J. Metz, E. Moisar, *Photogr. Korresp.* **100** (1964) 55.

72. in [14] vol. II, p. 85.
73. in [19] p. 12.
74. E. Moisar, S. Wagner, *Ber. Bunsen-Ges. Phys. Chem.* **67** (1963) 356.
75. in [19]p. 645.
76. R. W. Gurney, N. F. Mott, *Proc. R. Soc. London A* **164** (1938) 151.
77. B. E. Bayer, J. F. Hamilton, *J. Opt. Soc. Am.* **55** (1965) 439.
78. R. C. Baetzold, *J. Solid State Chem.* **6** (1973) 352.
79. L. M. Kellogg, *Photogr. Sci. Eng.* **18** (1974) 378.
80. J. W. Mitchell, *Photogr. J. Sect. B* **92** (1952) 38.
81. J. W. Mitchell, *Z. Electrochem.* **60** (1956) 557.
82. J. W. Mitchell, *J. Phys. Chem.* **66** (1962) 2359.
83. J. Malinowski, *Photogr. Sci. Eng.* **23** (1979) 99.
84. V. Platikanova, K. Starbova, *Photogr. Sci. Eng.* **22** (1978) 6.
85. C. Wagner, *Z. Electrochem.* **63** (1959) 1027.
86. E. Moisar, F. Granzer, D. Dautrich, E. Palm, *J. Photogr. Sci.* **25** (1977) 12.
87. E. Moisar, F. Granzer, D. Dautrich, E. Palm, *J. Photogr. Sci.* **28** (1980) 71.
88. E. Moisar, *Photogr. Sci. Eng.* **26** (1982) 124.
89. E. Kron, *Ann. Phys. (Leipzig)* **41** (1913) no. 4, 751.
90. H. E. Spencer, R. E. Atwell, *J. Opt. Soc. Am.* **54** (1964) 498.
91. R. Matejec, E. Moisar, *Photogr. Sci. Eng.* **12** (1968) 133.
92. Agfa, DE 1 169 290, 1962 (E. Klein, E. Moisar).
93. Agfa-Gevaert, DE-OS 1 597 488, 1967 (E. Klein, E. Moisar).
94. in [14] vol. III, p. 1122.
95. in [19] pp. 133, 182.
96. F. Dietz: "Zum Stand der Theorie der spektralen Sensibilisierung," *J. Signalaufzeichnungsmat.* **6** (1978) no. 4, 245 – 266;no. 5, 341 – 361.
97. F. M. Hamer: *The Cyanine Dyes and Related Compounds,* Wiley – Interscience, New York 1964.
98. F. Dietz: "Die Aggregation der Cyaninfarbstoffe und ihre Bedeutung für die spektrale Sensibilisierung", *J. Signalaufzeichnungsmat.* **1** (1973) no. 3, 157 – 180; no. 4, 237 – 252.
99. K. Norland, A. Ames, T. Taylor, *Photogr. Sci. Eng.* **14** (1970) 295 – 307.
100. W. West, P. B. Gilman in [19] Chap. 10.
101. J. W. Mitchell, *J. Imaging Sci.* **30** (1986) 91 – 101.
102. R. Steiger et al., *Photogr. Sci. Eng.* **24** (1980) 185 – 195.
103. P. B. Gilman, *J. Signalaufzeichnungsmat.* **4** (1976) no. 1, 5 – 23.
104. P. B. Gilman, *Photogr. Sci. Eng.* **18** (1974) 418 –428.
105. W. West, B. H. Carrol, *J. Chem. Phys.* **19** (1951) 417 – 427; W. West, P. B. Gilman, *Photogr. Sci. Eng.* **13** (1969) 221 – 230.
106. O. Riester, *Photogr. Sci. Eng.* **18** (1974) 295 – 301.
107. Polaroid, DE 2 804 148, 1977 (A. L. Borror, R. L. Hill).
108. Agfa-Gevaert, US 3 617 296, 1967 (H. Öhlschläger et al.).
109. Hoechst, DE 1 019 418, 1954 (A. Sieglitz, L. Berlin).
110. Agfa-Gevaert, DE-OS 2 106 517, 1970 (H. Depoorter, T. H. Ghys).
111. Fuji, EP-A 0 302 528, 1987 (H. Mifune, S. Takada).
112. Fuji, EP-A 0 278 509, 1987 (A. Ikegawa, M. Okazaki, Y. Ohashi).
113. R. Brünner, A. E. Oberth, G. Pick, G. Scheibe, *Z. Elektrochem.* **62** (1958) 132 – 145.
114. A. A. Muenter, W. Cooper, *Photogr. Sci. Eng.* **20** (1976) 121 – 124.
115. Ilford, GB 1 390 247, 1972.
116. Agfa, DE 1 153 246, 1962 (J. Götze, A. Randolph, O. Riester).
117. Kodak, US 3 505 070, 1967 (R. A. Litzerman, D. W. Heseltine, J. D. Mee).
118. Kodak, EP 0 138 362, 1983 (W. H. Günther, R. Przyklek, R. E. Leone).
119. General Aniline, US 2 442 710, 1987 (O. Riester).
120. Kodak, GB 1 240 522, 1967 (P. W. Jenkins).
121. Agfa-Gevaert, DE-OS 1 547 641, 1967 (J. Götze et al.).
122. Ilford, GB 604 217, 1945 (J. D. Kendall, F. P. Doyle).
123. Kodak, US 2 875 058, 1955 (B. H. Carroll, J. E. Jones).
124. Agfa-Gevaert, DE 1 447 577, 1964 (O. Riester).
125. Kodak, GB 1 199 794, 1966 (L. G. S. Brooker, E. J. Van Lare); US 3 501 310, 1967 (B. D. Illingsworth, H. E. Spencer).
126. Kodak, US 3 642 786, 1967 (D. W. Heseltine, L. L. Lincoln).
127. Mitsubishi Paper, DE 2 401 368, 1973 (A. Tanaka et al.)
128. Minnesota Mining, DE-OS 2 048 561, 1969 (B. A. Lea).

129. Kodak, GB 740 770, 1952 (L. G. S. Brooker, F. L. White).
130. Kodak, US 2 704 714, 1954 (B. H. Caroll, J. E. Jones).
131. Kodak, US 2 734 900, 1953 (D. W. Heseltine).
132. Kodak, US 3 623 881, 1969 (A. F. Fumia, L. G. S. Brooker).
133. Kodak, DE 2 046 687, 1969 (G. L. Hiller, C. G. Ulbing, J. E. Starr).
134. Fuji, DE-OS 2 161 184, 1970 (T. Sakai, A. Sato, N. Yamamoto, S. Kondo).
135. Agfa-Gevaert, DE 2 609 741 and DE 2 609 742, 1976 (H. Langen et al.).
136. Kodak, DE 1 547 705, 1965 (J. M. Owen, J. L. Graham, E. A. Pilsworth).
137. P. C. Burton, W. F. Berg, *Photogr. J. Sect. B* **86** (1946) 19.
138. J. Eggert: *Wissenschaftliche Photographie,* Verlag O. Helwich, Darmstadt 1958, p. 328.
139. R. Matejec, R. Meyer, *Z. Wiss. Photogr. Photophys. Photochem.* **57** (1963) 45.
140. R. Matejec, *Photogr. Korresp.* **107** (1971) no. 3, 37.
141. R. Matejec, *J. Signalaufzeichnungsmat.* **3** (1975) 219.
142. L. R. Solman, *J. Photogr. Sci.* **18** (1970) 179–184.
143. N. F. Mott, *Photogr. J. Sect. B* **88** (1948) 119.
144. J. Eggers, E. Klein, R. Matejec, *Angew. Chem.* **69** (1957) 291–301.
145. H. J. Metz, *J. Photogr. Sci.* **20** (1972) 111.
146. E. Klein, *Z. Elektrochem.* **62** (1958) 505.
147. G. C. Farnell, *J. Photogr. Sci.* **21** (1973) 145–154.
148. W. Jaenicke, C. Schott, *Z. Elektrochem.* **59** (1955) 956.
149. W. Jaenicke, F. Sutter, *Z. Elektrochem.* **63** (1959) 722.
150. W. Jaenicke, *Prog. Photogr.* **3** (1958) 55.
151. W. Jaenicke, *Photogr. Sci. Eng.* **6** (1962) 85.
152. W. Jaenicke, *Photogr. Sci.* **20** (1972) 2–12.
153. W. Reinders, *J. Phys. Chem.* **38** (1934) 783.
154. M. C. F. Beukers, Dissertation, Delft 1934.
155. P. J. Hillson, *J. Photogr. Sci.* **6** (1958) 97.
156. E. P. Sencenkov, E. A. Galasin, *Int. Congr. Photogr. Sci. Moscow C* **23** (1970) 101–104.
157. I. Konstantinov, J. Malinowski, *J. Photogr. Sci.* **23** (1975) 1 and 145.
158. I. Konstantinov, *Chimia* **33** (1979) 23.
159. J. Eggers, R. Posse, G. Scheibe, *Z. Elektrochem.* **58** (1954) 731.
160. J. Eggers, H. Heckelmann, K. Lohmer, R. Posse, *Agfa-Mittg.* **1** (1955) 95.
161. T. H. James, *PSA J. Sect. B* **16** (1950) 83.
162. T. H. James, *J. Phys. Chem.* **43** (1939) 701.
163. R. B. Pontius, R. G. Willies, R. J. Newmiller, *Photogr. Sci. Eng.* **16** (1972) 406.
164. R. B. Pontius, R. G. Willis, *Photogr. Sci. Eng.* **17** (1973) 21, 157, 326.
165. K. S. Bagdasarjan, *Zh. Fiz. Khim.* **17** (1943) 336.
166. H. Eger, Dissertation, TH München 1969.
167. H. J. Metz, *Int. Congr. Photogr. Sci.* Cologne (1986) 122.
168. T. H. James, *Sci. Am.* **187** (1952) no. 5, 30–32.
169. A. Hoffman, J. Charkoudian, A. Ames, *Photogr. Sci. Eng.* **18** (1974) 12.
170. in [19] p. 406.
171. T. H. James, *Photogr. Sci. Eng.* **1** (1957) 141.
172. T. H. James, *J. Franklin Inst.* **240** (1945) 83.
173. D. D. F. Shiao, E. L. Dedio, *Photogr. Sci. Eng.* **25** (1981) no. 4, 145.
174. D. D. F. Shiao, *Photogr. Sci. Eng.* **28** (1984) no. 6, 233.
175. D. D. F. Shiao, *J. Imaging Sci.* **31** (1987) no. 2, 64.
176. A. Hoffman, *J. Photogr. Sci.* **30** (1982) no. 2, 39.
177. in [19] p. 432.
178. in [14] p. 898.
179. W. E. Lee, T. H. James, *Photogr. Sci. Eng.* **6** (1962) 33.
180. G. I. P. Levenson, P. J. Twist, *J. Photogr. Sci.* **26** (1978) no. 1, 44.
181. A. Weissberger in [20] p. 686 ff.
182. T. H. James, *J. Photogr. Soc. Am.* **9** (1943) 62.
183. G. I. P. Levenson, *Photogr. J. Sect. B* **88** (1948) 102.
184. G. I. P. Levenson, *Photogr. J. Sect. B* **89** (1949) 2, 13.
185. W. Jaenicke, H. Kobayashi, *J. Photogr. Sci.* **31** (1983) no. 2, 69.
186. in [19] p. 609.
187. in [14] p. 974.
188. C. N. Nelson, *Photogr. Sci. Eng.* **15** (1971) 82.
189. G. C. Higgins, *Photogr. Sci. Eng.* **15** (1971) 106.
190. Research Disclos. 1976, no. 151 (15 162), 75.
191. W. T. Hanson, *Photogr. Sci. Eng.* **20** (1976) 155.
192. T. H. James, *Photogr. Sci. Eng.* **10** (1966) 344.
193. B. H. Carroll, D. Hubbard, C. M. Kretschman: *The Photographic Emulsion,* Focal Press, London 1968.
194. N. Kirillov: *Problems in Photographic Research,* Focal Press, London 1967.
195. R. W. James: *Chemical Technology Review no. 21: Photographic Emulsions,* Noyes Data Corp., Park Ridge, New Jersey, 1973.

196. J. S. Wey, J. P. Terwilliger, *AIChE J.* **20** (1974) no. 6.
197. J. S. Wey in W. R. Wilcox (ed.): *Preparation and Properties of Solid State Materials,* vol. 6, Marcel Dekker, New York 1981.
198. E. Moisar, E. Klein, *Ber. Bunsen-Ges. Phys. Chem.* **67** (1963) 949.
199. J. E. Maskasky, *J. Imaging Sci.* **30** (1986) no. 6, 247.
200. E. Moisar, S. Wagner, *Ber. Bunsen-Ges. Phys. Chem.* **67** (1963) 356.
201. Agfa-Gevaert, US 3 935 014, 1973 (S. Klötzer, E. Moisar).
202. Agfa-Gevaert, US 3 957 488, 1973 (S. Klötzer, E. Moisar).
203. Agfa-Gevaert, EP 0 006 543, 1978 (O. Lapp, L. Endres, H. Hückstädt, F. Moll).
204. Kodak, DE-OS 3 241 634, 1981 (J. A. Haefner, W. H. Sedgwick).
205. E. Klein, H. J. Metz, E. Moisar, *Photogr. Korresp.* **99** (1963) 99; **100** (1964) 55.
206. Agfa-Gevaert, DE 2 144 127, 1971 (H. Hammerstein, H. Reuß, E. Moisar).
207. J. E. Maskasky, *J. Imaging Sci.* **32** (1988) no. 4, 160.
208. Kodak, EP 0 041 959, 1979 (A. G. E. Mignot).
209. Kodak, US 3 415 650, 1964 (G. F. Frame, B. A. Johnson).
210. Agfa-Gevaert, DE 1 472 745, 1965 (R. F. Posse et al.).
211. Kodak, US 3 782 954, 1969 (H. D. Porter, B. A. Johnson).
212. Kodak, US 3 897 935, 1972 (D. M. Forster, G. J. Culhane).
213. Fuji, GB 1 515 139, 1974 (T. Sato, J. Miyahara, S. Tada).
214. G. Roewer, A. Winzer, *Wiss. Z. Tech. Hochsch. "Carl Schorlemmer" Leuna – Merseburg* **26** (1984) 645.
215. G. Möller, *Int. Congr. Photogr. Sci.* Moscow 1970, no. A 14.
216. B. Holland, J. R. Sawers, *Photogr. Sci. Eng.* **17** (1973) no. 3, 295.
217. Konica, EP-A 0 039 119, 1987 (Y. Mochizuki, S. Ishikawa, T. Yagi, M. Inoue).
218. U. Löchner, TH Darmstadt, Synchrotron-Messung am Hasylab (Desy, Hamburg).
219. E. Fuchs, *Adhäsion* **8** (1964) 152 – 160.
220. A. T. Ram et al., *SMPTE J.* **97** (1988) June, 474.
221. G. Pollakowski, *Bild Ton* **40** (1987) no. 3, 90 – 91.
222. Perutz, DE 1 068 114, 1958 (K. Herrmann).
223. Du Pont, DE 1 572 119, 1964 (A. B. Cohen).
224. Gevaert, BE 604 869, 1961 (L. F. De Keyser, A. J. Conix, J. A. Herbots).
225. ICI, DE-OS 2 400 372, 1973 (D. R. Mann).
226. Kodak, US 3 245 937, 1961 (R. H. Wagner).
227. Agfa, DE 1 207 623, 1964 (P. Kruck, H. Klockgether).
228. Agfa-Gevaert, DE-OS 2 325 087, 1972 (A. J. Van Paesschen, L. J. Van Gossum, J. J. Priem).
229. W. I. Kisner, *Kino-Tech. (Berlin)* **16** (1962) 295 – 298; L. Berben, *Kino-Tech. (Berlin)* **22** (1968) 275 – 278.
230. E. Fuchs, *Coating* **3** (1969) 78 – 82.
231. Agfa-Gevaert, US 3 514 291, 1966 (J. C. Vanpoeke, L. F. De Keyser).
232. Agfa, DT 912 778, 1944 (L. Jacob).
233. Kodak, GB 1 475 685, 1973 (H. M. Wagner, N. D. Uffindell).
234. Kodak, FR 2 180 465, 1972.
235. Kodak, US 3 963 498, 1971 (D. J. Trevoy); US 4 237 194, US 4 308 332, US 4 526 706, 1979 (D. A. Upson, D. J. Steklenski); M. Angelopoulos, A. Ray, A. G. Macdiarmid, A. J. Epstein, *Synth. Met.* **21** (1987) 21 – 30.
236. Kodak, US 5 006 451, 1991 (C. C. Anderson et al.).
237. Fuji, DE-OS 3 150 514, 1980 (H. Kawaguchi, T. Inayama).
238. W. Nauck, E. Lehmann: *Fabrikation und Prüfung der photographischen Materialien,* Union Deutsche Verlagsgesellschaft, Berlin 1928, pp. 17 – 23.
239. R. Wanka, H. G. Jerenz: "Wechselwirkung zwischen photographischen Schichten und verschiedenen Trägern", Paper Presented at the 7th Conference on Scientific and Applied Photography, Budapest, 11 Nov. 1975.
240. H. Hentschel: *Chemische Technologie der Zellstoff- und Papierherstellung,* 3rd ed., VEB Fachbuchverlag, Leipzig 1967, p. 606.
241. Wiggins Teape, GB 1 043 703, 1962; FR 1 353 771, 1964 (G. Kemp).
242. Kodak, GB 971 058, 1962 (M. J. Alsup, W. J. Venor).
243. Kodak, GB 1 005 631, 1964 (J. E. Ratcliff, W. L. Johnson).
244. Kodak, US 3 833 380, 1965 (I. H. Crawford, F. M. Hartman).
245. F. Wentzel: *Die photographisch-chemische Industrie,* **vol. X,** Th. Steinkopf, Dresden 1926, pp. 102 – 109.
246. R. Korn, F. Burgstaller: *Papier und Zellstoff-Prüfung,* 2nd ed., Springer Verlag, Berlin 1953.
247. L. Booth, *Mod. Plast.* **36** (1958) no. 1, 90 – 95; **36** (1958) no. 2, 90 – 99.

248. B. V. Deryagin, S. M. Levi: *Film Coating Theory,* Focal Press, London 1964.
249. P. Groenveld, *Chem. Eng. Sci.* **25** (1970) 33, 1259, 1267.
250. R. P. Spiers, C. V. Subbaram, W. L. Wilkinson, *Chem. Eng. Sci.* **29** (1974) 289.
251. K. J. Ruschak, *Chem. Eng. Sci.* **31** (1976) 1057.
252. P. R. Schunk: "Surfactant and Polymer Additives in Coating and Related Flows," Dissertation, University of Minnesota 1989.
253. S. F. Kistler: "The Fluid of Mechanics of Curtain Coating and Related Viscous Free Surface Flows with Contact Lines," Dissertation, University of Minnesota 1984.
254. J. R. A. Pearson, S. M. Richardson (eds.): *Computational Analysis of Polymer Processing,* Appl. Sci. Publishers, London 1983, Chap. 8.
255. R. J. Jacobs, *Pap. Film Foil Converter* **37** (1963) no. 2, 2, 47; no. 3, 50; no. 4, 62; no. 5, 60; no. 6, 60;no. 7, 5.
256. Agfa-Gevaert, DE 1 652 273, 1967 (P. Herzhoff et al.).
257. Agfa-Gevaert, DE 1 652 317, 1963 (P. Herzhoff et al.).
258. Kodak, GB 726 783, 1951 (T. A. Russell, D. C. Bellis, D. B. Fogg).
259. Ch. F. Carlson, US 2 221 776, 1938.
260. Du Pont, DE 1 030 178, 1955 (R. H. Beck).
261. Kodak, DE 1 151 173, 1955(R. A. Russell).
262. Kodak, DE 1 928 025, 1968 (J. F. Greiler).
263. Kodak, DE 1 928 031, 1968 (J. J. Hughes).
264. J. E. Jolley, *Photogr. Sci. Eng.* **14** (1970) 169 – 177.
265. L. Taeschky, *Chem. Ing. Tech.* **36** (1964) 449.
266. O. Krischer: *Die wissenschaftlichen Grundlagen der Trocknungstechnik,* 3rd ed., Springer Verlag, Berlin 1978.
267. E. U. Schlünder, H. Martin: *Stoffübergang bei Prallströmung aus Schlitzdüsenfeldern,* Inst. für Thermische Verfahrenstechnik der Universität Karlsruhe 1970.
268. M. Roth, *Chem. Eng.* **73** (1966) no. 8, 83.
269. ISO 1012, Photography – Film dimensions – Pictorial Sheets, 1991.
270. ANSI PH 1.65, Disc Film Cartridge Dimensions and Functional Requirements, 1988.
271. ISO 732, Photography – Film Dimensions – 120 and 220 sizes, 1991.
272. ISO 1007, Photography – 135-size Film and Magazine – Specification, 1979.
273. ANSI PH 1.14, 135-Size Film Magazines and Film for 135-Size Still-Picture Cameras – Dimensions, 1990.
274. ISO 3029, Photography – 126-size Cartridges – Dimensions of Cartridge, Film and Backing paper, 1983.
275. ISO 7261, Photography – 110-size Cartridges – Dimensions, 1987. ISO 7330, Photography – 110-size Cartridges – Location and Dimensions of Film Exposure Identification Notches, 1988. ISO 7374, Photography – 110-size Cartridges – Dimensions and Format of Film and Backing Paper, 1987.
276. Fuji, DE 3 344 634, 1982 (T. Kanaoka, S. Shimizu).
277. Agfa-Gevaert, DE 2 305 383, 1973 (F. Hoffacker, H. Johanns, H. Schruff).
278. Agfa-Gevaert, EP 0 186 824, 1984 (F. Hoffacker, G. Kovacic, H. Lührig).
279. ISO 1008, Photography – Black and White Paper for General Use – Sizes of Sheet Material, 1973.
280. ISO 1009, Photography – Paper Dimensions – Rolls for Printers, 1973.
281. Nishimura, US 3 884 427, 1972 (K. Hashimoto, K. Wada).
282. Agfa-Gevaert, DE 2 049 427, 1970 (K. Taeffner, H. Lührig).
283. Agfa-Gevaert, DE 2 112 808, 1971 (H. Lührig, K. Taeffner, G. Pohl, G. Schmitz, R. Kreit).
284. R. J. Henry: *Controls in Black-and-White Photography,* 2nd ed., Focal Press, Boston – London 1986.
285. R. Matejec, *Photogr. Korresp.* **104** (1968) 153 – 157.
286. A. A. Rasch, J. I. Crabtree, *J. SMPTE* **62** (1954) 1.
287. W. Pelz, *Angew. Chem.* **66** (1954) 231.
288. L. Lugosy, M. Grossa, W. Pistor in [A 8] pp. 195,196.
289. Ilford, GB 542 502, 1940 (D. J. Kendall).
290. H. Ulrich, H. Schüler, *Photogr. Sci. Eng.* **4** (1960) 225 – 228; Agfa, DE 945 606, 1954 (H. Ulrich, R. Mersch, O. Wahl, D. Delfs).
291. Agfa, DE 955 025, 1954 (L. Burghardt, W. Pelz, O. Wahl, H. Sassmann); DE 1 029 229, 1956 (W. Pelz, L. Burghardt, O. Wahl).
292. L. Burghardt, W. Pelz in [32]vol. 2 (1958) 165 – 175.
293. K. S. Bogomolov, *Usp. Nauchn. Fotogr.* **12** (1966) 100 – 105.
294. R. W. Henn, T. A. Olivares: "Dihydrofuran Developing Agents", in [8] pp. 207 – 221.
295. W. E. Lee: "Superadditivity," in [19] pp. 432 – 436.
296. G. Haist: *Monobath Manual,* Morgan & Morgan, New York 1966.

297. Agfa-Gevaert, DE 1 522 428, 1967 (H. Meckl, H. Häseler).
298. A. Newman, *Br. J. Photogr.* **114** (1967) 1009 – 1012.
299. Agfa-Gevaert, DE-OS 2 218 387, 1972 (E. Weyde, M. Scheibitz, A. v. König).
300. "Silberrückgewinnung aus fotochemischen Abwässern," *Mod. Fototech.* **27** (1979) no. 3, 127.
301. H. Meckl: "Experience with Ion Exchange for Silver Recovery from Photographic Effluents," *J. Imaging Technol.* **11** (1985) 51 – 58.
302. P. R. Quinones: "In-Situ-Precipitation as the Regeneration Step in Ion Exchange for Silver Recovery," *SMPTE J.* **93** (1984) 800.
303. A. Rott, E. Weyde: *Photographic Silver Diffusion Processes,* Focal Press, London 1972.
304. V. Walworth, S. Mervis in [27] pp. 181 – 195.
305. R. Fischer, DE 253 335, 1912; Neue Photographische Ges., DE 257 160, 1911.
306. R. L. Bent et al., *J. Am. Chem. Soc.* **73** (1951) 3100 –3125.
307. Agfa-Gevaert, DE 1 813 920, 1968 (R. Matejec).
308. Kodak, DE 2 226 770, 1971 (V. L. Bissonette); DE 2 226 771, 1971 (V. L. Bissonette, W. B. Travis).
309. R. Fischer, H. Siegrist, *Photogr. Korresp.* **51** (1914) 18 – 22.
310. R. J. Tuite, *J. Appl. Photogr. Eng.* **5** (1979) 200.
311. E. Pelizetti, C. Verdi, *J. Chem. Soc. Perkin Trans. 2,* 1973, 808.
312. G. H. Brown et al., *J. Am. Chem. Soc.* **79** (1957) 2919.
313. *J. Chem. Soc. Perkin Trans. 2,* 1973, 1766.
314. K. Sano, *J. Org. Chem.* **34** (1968) 2076.
315. K. Sano, THL 1968, no. 28, 3203.
316. Kodak, DE 1 810 464, 1967 (P. A. Marr, E. B. Knott, J. B. Bailey).
317. *J. Chem. Soc. Perkin Trans. 1,* 1977, 2047.
318. A. Mitsui, N. Furutachi, T. Hirose, *SPSE 2nd Photofinishing Minilab. Technology Symp.* 1986; N. Mii, *Shashin Kogyo* **45** (1987) no. 2, 43.
319. Fuji, EP 117 511, 1983 (H. Kobayashi, K. Mihayashi, I. Itoh).
320. K. I. Shinohara, E. Bayer, *J. Photogr. Sci.* **35** (1987) 181.
321. W. T. Hanson, P. W. Wittum, *PSA J.* **13** (1947) 94.
322. W. T. Hanson, *J. Opt. Soc. Am.* **40** (1950) 166.
323. W. D. Sende in: Veröffentl. d. Wiss. Photo-Lab. Wolfen, vol. X, 1965, p. 237.
324. Fuji, EP 310 125, 1987 (S. Sakanoue, H. Kobayashi).
325. Agfa-Gevaert: "Agfacolor XRG Negativfilme für Farb-Papierbilder," Technical Data no. A 20-F, 1989.
326. Agfa-Gevaert DE 1 121 470, 1960 (E. Böckly, H. Ulrich, G. Kuhn); Fuji, DE 2 018 341, 1969 (A. Kumai, T. Usami).
327. K. Meyer, *Bild Ton* **8** (1954) 228.
328. Konishiroku Photo, DE-OS 3 412 684, 1983 (S. Koboshi, M. Kurematsu).
329. M. Meckl: "Recycling of Color Paper Developer," *J. Imaging Technol.* **5** (1979) 216 – 219.
330. Agfa-Gevaert, DE 2 137 549, 1971 (H. Häseler, H. Meckl, A. Seiler).
331. Eastman Kodak, Pamphlet C 42-301, Rochester 1984.
332. Agfa-Gevaert, Pamphlet 21.8820 (287), Mortsel 1987.
333. J. E. Crisante, W. A. Szafranski: "Kodak Persulfate Bleach for Process ECN-2," *SMPTE J.* **91** (1982) 1058 – 1065.
334. Agfa-Gevaert, Pamphlet 21.8821 (488), Mortsel 1988.
335. R. C. Stehlin: "Persulfate Bleach and Motion Picture Film Processes," *SMPTE J.* **91** (1982) 158 –163.
336. Eastman Kodak, Pamphlet: "Manual for Processing Kodak Ektachrome Films Using Process E-7", Rochester 1977, revised Jan. 1980.
337. Kodak, Pamphlet RA-5-301, Sept. 1987.
338. M. Huber: *Grundlagen der Fotografie, Farblabor-Praxis,* Verlag Laterna Magica, München 1984.
339. A. Meyer, *J. Photogr. Sci.* **20,** (1972)81 – 86.
340. A. Meyer, *Photogr. Sci. Eng.* **18,** (1974)530 .
341. M. Schellenberg, H. P. Schlunke, *Chem. Unserer Zeit* **10,** (1976)131 – 138 .
342. V. Walworth, S. Mervis in [27]pp. 202 – 203.
343. Ch. Van de Sande: "Farbstoffdiffusionssysteme in der Farbphotographie", *Angew. Chem.* **95** (1983) 165 – 256.
344. R. Lambert: "A Technical Description of Polaroid's Spectra Film," *J. Imaging Technol.* **15** (1989) 108 –113.
345. S. Bloom, M. Green, M. Idelson, M. Simon: "The Dye Developer in the Polaroid Photographic Process", *Chem. Synth. Dyes* **1978,** no. 8, 331 –387.
346. Polaroid, US 3 719 489, 1971 (R. Cieciuch, R. Luhowy, F. Meneghini, H. Rogers); Polaroid, US 4 740 448, 1986 (P. Kliem).

347. W. T. Hanson, Jr.: "A Fundamentally New Imaging Technology for Instant Photography," *Photogr. Sci. Eng.* **20** (1976) 155 – 160.
348. Kodak, US 3 928 312, 1971 (L. J. Fleckenstein).
349. Fuji, US 4 055 428, 1977 (K. Koyama, Y. Maekawa, M. Miyakawa).
350. Kodak, US 4 139 379, 1977 (R. A. Chasman, R. P. Dunlap, J. C. Hinshaw).
351. Agfa-Gevaert, DE 2 823 159, 1978 (E. Meier, W. Lässig); Agfa-Gevaert, EP 0 004 399, 1978 (W. Janssens).
352. Konishiroku, JP-Kokai 58 153 933, 1982 (N. Fujita).
353. Fuji, EP 0 220 746, 1985/86 (K. Nakamura, S. Nakamura).
354. M. Peters: "A Novel Dye-Diffusion Material for Prints from Color Slides," *J. Imaging Technol.* **11** (1985) 101 – 104.
355. V. Walworth, S. Mervis in [27]p. 221.
356. E. Land, *Photogr. Sci. Eng.* **21** (1977) 225 – 236.
357. S. Liggero, K. Mc Carthy, J. Stella: "The Polaroid 35 mm Instant Slide System", *J. Imaging Technol.* **10** (1984) 1 – 9.
358. F. Hurter, *J. Soc. Chem. Ind. (London)* **9** (1890) 455.
359. W. Behrendt in [33] vol. 3.
360. L. A. Jones, H. R. Condit, *J. Opt. Soc. Am.* **31** (1941) 651.
361. ISO 5/3, 1984: Photography, Density Measurements, Part 3: Spectral Conditions.
362. ANSI, PH 2.17, 1958.
363. DIN 19 010, Part 1, 1979: Lichtelektrische Belichtungsmesser.
364. ANSI-PH 3.49, 1979: Photographic Exposure Meters (Photoelectric Type).
365. G. Vieth, W. Heiland, *Appl. Opt.* **7** (1968) 1043.
366. a) DIN 4512, Parts 1 – 6, 1984 – 1985: Photographische Sensitometrie; b) ISO 6 DIS, 1989; c) ANSI-PH 2.5, 1979: Method for Determination Speed of Photographic Negative Materials (monochrome, continuous-tone).
367. ANSI-PH 2.27, 1981: Determination of ISO (ASA) Speed of Color Negative Films for Still Photography.
368. a) ANSI-PH 2.21, 1983: American National Standard for Photography (Sensitometry) – Color Reversal Camera Films Determination of Speed; b) DIN 4512, Part 4, 1985; see also [378]
369. L. A. Jones, *J. Franklin Inst.* **190** (1920) 39.
370. ANSI-PH 2.2, 1984: American National Standard for Photography (sensitometry) – Black-and-White Continuous Tone Papers – Determination of ISO Speed and Range of Printing.
371. ANSI PH 2.40, 1985: American National Standard for Photography (Film): Root Mean Square (rms) Granularity of Film (Images on one side only), Method for Measuring.
372. H. Siedentopf, *Z. Phys.* **38** (1937) 454.
373. G. C. Higgins, K. F. Stultz, *J. Opt. Soc. Am.* **49** (1959) 925.
374. A. Callier, *Z. Wiss. Photogr. Photophys. Photochem.* **7** (1909) 257.
375. P. G. Nutting, *Philos. Mag.* **26** (1913) 423.
376. L. A. Jones, G. C. Higgins, *J. Opt. Soc. Am.* **41** (1951) 61.
377. S. A. Benton, *Photogr. Sci. Eng.* **21** (1977) 177.
378. H. T. Buschmann, *Optik Stuttgart* **38** (1973) 196.
379. E. W. Selwyn, *Photogr. J.* **75** (1935) 571.
380. E. A. Trabka, *Photogr. Sci. Eng.* **21** (1977) 4.
381. E. Ranz, *Chem. Labor Betr.* **30** (1979) 229.
382. G. C. Higgins, R. N. Wolfe, *J. Opt. Soc. Am.* **45** (1955) 121.
383. ANSI PH 2.39, 1977: American National Standard: Method of Measuring the Photographic Modulation Transfer Function of Continuous-Tone, Black-and-White Photographic Films.
384. R. R. Lamberts, C. M. Straub, *Photogr. Sci. Eng.* **9** (1965) 335.
385. M. De Belder, J. Jespers, R. Verbrugghe, *Photogr. Sci. Eng.* **9** (1965) 314.
386. G. Langner, R. Müller in [32] vol. 4, pp. 339 – 353.
387. F. Scott, R. M. Scott, R. V. Shack, *Photogr. Sci. Eng.* **7** (1963) 345.
388. D. Hertel et al., *J. Signalaufzeichnungsmat.* **16** (1988) 185.
389. R. Tromnau, H. T. Buschmann, Kongreß für Reprographie und Information, Hannover 1975, p. 159.
390. M. A. Kriss, C. N. Nelson, F. C. Eisen, *Photogr. Sci. Eng.* **18** (1974) 131.
391. C. N. Nelson, *Photogr. Sci. Eng.* **15** (1971) 82.
392. G. C. Higgins, Kongreß für Reprographie and Information, Hannover 1975, p. 207.
393. R. G. Gendron, *J. Soc. Motion Pict. Eng.* **82** (1973) 1009.
394. C. N. Nelson, *J. Opt. Soc. Am.* **63** (1973) 1289.
395. M. De Belder, J. De Kerf, J. Jespers, R. Vergrugghe, *J. Opt. Soc. Am.* **55** (1965) 1261.
396. H. J. Metz, *Photogr. Korr.* **106** (1970)37, 56, 70.

397. J. J. De Palma, J. Gasper, *Photogr. Sci. Eng.* **16** (1972) 181.
398. J. Gasper: "Photographic and Electronic Image Quality," Symposium Cambridge, Sept. 1984, p. 1.
399. ANSI/ISO 6328, 1982: American National Standard for Photography (materials): Determination of ISO Resolving Power.
400. CIE Publication no. 15.1, DIN 5033: Farbmessung.
401. CIE Publication no. 15.2, DIN 6174, 1976: CIE-Lab.
402. P. Z. Adelstein: *J. Imaging Technol.* **12** (1986) 52 – 56.
403. ISO 5466, 1986: Processed Safety Photographic Film – Storage Practices.
404. ISO 6051, 1986: Processed Photographic Paper Prints – Storage Practices.
405. ISO 3897, 1986: Processed Photographic Plates – Storage Practices.
406. ISO 417, 1989: Determination and Measurement of Residual Thiosulphate and other Chemicals in Film, Plates and Papers.
407. ISO 10 602, 1991: Photography – Processed Silver – Gelatin Black and White Film–Specifications for Stability.
408. P. Z. Adelstein, *J. Imaging Sci. Technol.* **36** (1992) 37 – 41.
409. ISO 8225, 1987: Ammonia Processed Diazo Photographic Film – Specification for Stability.
410. ISO – DIS 10 977, 1991: Photography– Processed Photographic Colour Films and Paper Prints–Methods for Measuring Image Stability.
411. ISO 2803, 1974: Silver-Gelatin Type Microfilms –Processing and Storage for Archival Purposes.
412. DIN 19 070, Parts 1, 3, 5: Haltbarkeit verarbeiteter strahlungsempfindlicher Materialien.
413. ANSI PH 1.28, PH 1.41, PH 1.66, PH 1.60, PH 1.67, PH 1.43, PH 1.45, PH 1.48, PH 1.53.
414. L. H. Feldman, *J. Appl. Photogr. Eng.* **7** (1981) 1.
415. R. W. Henn, D. G. Wiest, *Photogr. Sci. Eng.* **7** (1963), 253 – 261.
416. ISO 10 214, 1991: Photography – Processed Photographic Materials–Filling Enclosures for Storage.
417. R. W. Henn, D. G. Wiest, B. D. Mack, *Photogr. Sci. Eng.* **9** (1965) 167.
418. K. B. Hendriks in J. M. Sturge, V. Walworth, A. Shepp (eds.): *Imaging Processes and Materials, Neblette's 8th ed.,* van Nostrand, New York 1989.
419. F. J. Drago, W. E. Lee, *J. Imaging Technol.* **12** (1986) 57 – 63.
420. E. Dunmore, *Br. J. Photogr.* **37** (1890) 327.
421. J. Spiller, *Photogr. News* **162,** 471.
422. D. F. Kopperl, Th. J. Huttmann, Jr., *J. Imaging Technol.* **12** (1986) 173 – 180.
423. D. F. Kopperl, R. Goetting, Th. J. Huttman, Jr., *J. Photogr. Sci.* **36** (1988) 40 – 42.
424. J. M. Reilly, D. W. Nishimura, K. M. Cupriks, P. Z. Adelstein, *Abbey Newsl.* **12** (1988) no. 5, 83 –88.
425. E. Weyde, *Chimia* **23** (1969) 42.
426. E. Weyde, *Photogr. Sci. Eng.* **16** (1972) 283.
427. Agfa-Gevaert, EP 0 309 873, 1987 (W. Berthold, P. Marx, H. Oehlschläger).
428. DIN 15 551, Part 3, 1985: Strahlungsempfindliche Filme – Zellhornfilm – Begriffe, Eigenschaften, Handhabung, Lagerung.
429. ISO 543, 1974: Motion Picture Safety Film – Requirements and Method of Testing.
430. ISO 7830, 1983: Safety Films other than Motion-Picture Films – Material Specifications.
431. DIN 15 551, Part 1: Strahlungsempfindliche Filme – Sicherheitsfilm – Begriffe, Anforderungen, Prüfung.
432. K. Brems, *SMPTE J.* **97** (1988) 991 – 993.
433. G. Pollakowski, *Bild Ton* **40** (1987) 90 – 91.
434. *Ullmann,* 4th ed., **18,** 436; *Ullmann,* 3th ed., **13,** 633.
435. T. F. Parsons, G. G. Gray, I. H. Crawford, *J. Appl. Photogr. Eng.* **5** (1979) 110 – 117.
436. S. I. Anderson, R. L. Ellison, *J. Imaging Technol.* **16** (1990) 27 – 32.
437. P. Z. Adelstein, C. L. Graham, L. E. West, *SMPTE J.* **79** (1970) 1011 – 1018.
438. Folienserie des Fonds der Chemischen Industrie, Frankfurt, **23** (1989) 41.
439. R. J. Tuite, *J. Appl. Photogr. Eng.* **5** (1979) 200 –207.
440. T. Miyagawa, Y. Shirai, *J. Imaging Sci.* **29** (1985) 216 – 218.
441. C. C. Bard, SPSE Symposium on Photofinishing Technology, Las Vegas 1986.
442. C. C. Bard, G. W. Larson, H. Hammond, C. Packard, *J. Appl. Photogr. Eng.* **6** (1980) 42 – 45.
443. E. Fuji, H. Fuji, Y. Hisanaga, *J. Imaging Technol.* **14** (1988) 29 – 37.
444. S. Arrhenius, *Z. Phys. Chem.* **4** (1889) 226 – 248.
445. Y. Seoka, S. Kubodera, T. Aono, M. Hirano, *J. Appl. Photogr. Eng.* **8** (1982) 79 – 82.

446. ISO 6077, 1980: Photography – Determination of Brittleness of Photographic Film – Wedge Brittleness Test.
447. ISO 8776, 1988: Photography – Photographic film – Determination of Folding Endurance.
448. P. Lalvani, D. Heppekausen: "Apparatus to Determine the Brittleness of Uncoated and Coated Film Base," *Res. Disclos.* **253** (1985) (25 302), 213 –215.
449. ISO 4330, 1987: Photography – Determination of the Curl of Photographic Film.
450. ANSI-IT 9.1, 1988: Silver Gelatine Type – Specifications for Stability.
451. DIN 19 070, Part 1, 1985: Haltbarkeit verarbeiteter strahlungsempfindlicher Materialien.
452. ANSI-PH 1.32, 1959: American National Standard Methods for Determining the Dimensional Change Characteristics of Photographic Film and Paper.
453. DIN 15 586, Part 3, 1985: Prüfung mechanischer Eigenschaften von strahlungsempfindlichen Materialien, Bestimmen der Maßänderungen von fotografischen Filmen und Papieren.
454. ISO 6221, 1980: Photography – Photographic Films and Paper, Determination of the Dimensional Change Characteristics.
455. DIN 15 586, 1977: Bestimmung der Gleitfähigkeit entwickelter Filme.
456. ASTM D 1894-87, 1987: Standard Test Method for Static and Kinetic Coefficients of Friction of Plastic Film and Sheeting.
457. ISO 1007, 1979: Photography – 135-size Filmmagazine – Specification.
458. DIN 4535, Part 2, 1977: Patrone für Film Nr. 135 (35 mm), Lehre für Kapselumriß und Filmauszugskraft.
459. ANSI-PH 1.37, 1977: Methods for Determining the Scratch Resistance of Processed Photographic Film.
460. DIN 15 586, Part 2, 1983: Prüfung mechanischer Eigenschaften von strahlungsempfindlichen Materialien, Bestimmen der Schrammfestigkeit von fotografischen Filmen.
461. E. J. Lewis, A. K. Soper: "Note on an Apparatus for Measuring the Vertical Swelling of Colloid Layers," *J. Sci. Instr.* **27** (1950) 242.
462. R. S. Pomeroy: "A Rapid Response Swellometer," *Photogr. Sci. Eng.* **16** (1972) 449.
463. ANSI-PH 4.35, 1972: Method for Determining the Resistance of Photographic Films to Abrasion during Processing.
464. H. M. Smith (ed.): *Holographic Recording Materials,* Springer Verlag, Berlin 1977.
465. in [42] p. 124.
466. H. W. Hoadley: "Scientific Applications of Photography," in [44] pp. 562 – 573.
467. in [43] pp. 506 – 520.
468. in [15] pp. 593 – 606.
469. P. Demers: *Ionographie, Les Emulsions Nucléaires, Principe et Applications,* Les Presses Universitaires, Montréal 1955.
470. in [15] pp. 590 – 592.
471. H. Seemann: "Radiography", in [44] pp. 550 – 561.
472. G. A. Chambers: "Cinematography," in [4]
473. in [30]Chap. 7.4, pp. 252 – 280.
474. J. Belton, *SMPTE J.* **99** (1990) 457 – 465.
475. J. Belton, *SMPTE J.* **97** (1988) 711 – 720.
476. in [15] pp. 374 – 392.
477. BKSTS Technical Manuals: Photographic Sound Technology, BKSTS, London 1980.
478. J. M. Sturge: "Micrographs", in [4] pp. 607 – 619.
479. G. W. W. Stevens: *Micrography,* Wiley Interscience, New York 1968.
480. D. Gadaly: *A Microfilm Handbook,* National Microfilm Assoc., Silver Spring, MD, 1974.
481. G. A. Plunke (ed.): *Micrographic Film Technology,* National Microfilm Assoc., Silver Spring, MD 1983.
482. H. E. Seemann: *Physical and Photographic Principles of Medical Radiography,* Wiley Interscience, New York 1968.
483. Th. Laubenberger: *Technik der Medizinischen Radiologie,* 4th ed., Deutscher ärzte-Verlag, Köln 1986.
484. D. P. Roberts, N. L. Smith: *Radiographic Imaging, a Practical Approach,* Churchill, Livingstone, Edinburgh 1988.
485. in [15] pp. 561 – 586.
486. in [43] pp. 484 – 502.
487. in [45] pp. 545 – 546.
488. in [43] pp. 968 – 775.
489. C. H. Blaquez, *J. Imaging Technol.* **15** (1989) 163 –168.
490. L. A. Mayer et al., "Aerial Imaging Systems," in [27] Chap. 15, pp. 423 – 464.
491. "Aerial Photography and Photogrammetry", in [287] pp. 528 – 549.
492. I. G. Clevers, *Int. J. Remote Sensing* **9** (1938) 319 –332.
493. Neue Photogr. Ges., DE 250 183, 1911; DE 262 036, 1912.
494. F. F. Renwick, GB 494 088, 1937.
495. Bayer, DE 825 346, 1950 (A. Leubner).

496. Agfa, DE 1 058 835, 1957 (B. Morcher, G. Kolb, I. Dennstedt).
497. Kodak, FR 1 123 458, 1954 (W. B. Kendall); FR 1 151 738, 1955 (H. C. Yutzy, A. W. Wise).
498. F. Granzer, E. Moisar: *Progress in Basic Principles of Imaging Systems,* Vieweg, Braunschweig – Wiesbaden 1987, p. 8.
499. in [15] pp. 357 – 362.
500. in [42] p. 125.
501. in [15] pp. 362 – 363.
502. Gevaert, BE 538 700, 1954.
503. Kodak, US 3 062 651, 1959 (P. J. Hillson).
504. D. Gabor, *Nature (London)* **161**(1948) 777; *Proc. R. Soc. (London) A* **197** (1949) 454.
505. E. N. Leith, J. Upatnieks, *J. Opt. Soc. Am.* **52** (1962) 1123; **53** (1963) 1377.
506. H. T. Buschmann: "Bleichprozesse zur Erzeugung rauscharmer, lichtstarker Phasenhologramme," *Optik (Stuttgart)* **34** (1971) 240.
507. Agfa-Gevaert, Nondestructive Testing Technical Information: Holography, Antwerpen.
508. EG-Richtlinie 76/464/EWG (1976), Appendix List II.
509. A. C. Cooley, T. J. Dagon, P. W. Jenkins, K. A. Robillard: "Silver and the Environment," *J. Imaging Technol.* **14** (1988) no. 6, 183 – 189.
510. European Council of Chemical Manufacturer's Federation (CEFIC): "Silver and the Environment."
511. Degussa AG: "TMT 15 für die Abtrennung von Schwermetallen aus Abwässern," Hanau, 1987.
512. Abwassertechnische Vereinigung: "Abwasser fotografischer Verarbeitungsbetriebe (Großlabors)," Merkblatt M 769, 1988.
513. K. Münger, W. Schefer: "ökologische Beurteilung von fotografischen S/W-Entwicklern," UGRA-Bericht 81, St. Gallen 1987.
514. National Association of Photographic Manufacturers, Hydroscience Inc.: "Environmental Effect of Photoprocessing Chemicals," vol. 1, Harrison, N. Y., 1974.
515. M. Watson: "Environmental Fate and Effects of a Color Developer," 6th International Symposium on Photofinishing Technology, Las Vegas 1990.
516. H. Knorre, D. Maennig, K. Stuetzel: "Chemical Treatment of Effluent from Photofinishing Plants," *J. Imaging Technol.* **14** (1988) no. 6, 154 – 156.
517. T. W. Bober, T. J. Dagon: "Treating Photographic Processing Solutions and Chemicals with Ozone," *J. Water Poll. Control Fed.* **47** (1975) 2114 – 2129.
518. Communication from the Photoindustrieverband, Frankfurt/M. 1992.

Photoresists → Imaging Technology
Phototubes → Photoelectricity

Photovoltaic Cells

FRITZ PFISTERER, Institute of Physical Electronics, University of Stuttgart, Stuttgart, Federal Republic of Germany

1. Introduction 1
2. The Need for Renewable Energies – History of Photovoltaics . . 2
3. Fundamentals 3
3.1. Features of Photovoltaic Cells 3
3.2. Principle of the Photovoltaic Effect and Characteristics of Photovoltaic Cells 3
3.3. Energy Conversion and Energy Losses in Photovoltaic Cells 4
3.4. Equivalent Circuit and Current – Voltage Characteristic of Solar Cells 6
3.5. Tandem Structures 7
3.6. Types of Semiconductor Junctions . . 8
4. Types of Photovoltaic Cells 9
4.1. Single- and Multicrystalline Solar Cells 9
4.2. Amorphous Silicon Solar Cells 16
4.3. Crystalline Silicon Thin-Film Solar Cells 18
4.4. Summary of the Status of Silicon Solar Cells 22
4.5. Compound Semiconductor Solar Cells 23
4.6. Other Types of Photovoltaic Cells . . . 30
4.7. Photovoltaic Modules 33
5. Uses 33
5.1. Stand-Alone Systems 34
5.2. Grid-Connected Power-Generating Systems 34
5.3. Consumer Products 37
6. Economic Aspects and Market 38
7. Future of Photovoltaic Cells 39
8. References 41

(This article is based on the article of YUKINORI KUWANO, Ullmann's Encyclopedia of Industrial Chemistry, 5ed., Vol. A 20.)

1. Introduction

Photovoltaic cells can convert sunlight directly into electric power. The way via any thermodynamic cycle of conventional power production is avoided, the conversion process operates noiseless without any moving parts. The main components of photovoltaic cells are semiconductors (→ Semiconductors) and surface and back electrodes for collecting the current. Photovoltaic power generation is one of the most fascinating methods of providing renewable energy, and the expectations for photovoltaic cells are high.

2. The Need for Renewable Energies – History of Photovoltaics

Concerns about the finiteness of fossil fuels and the environmental problems associated with the consumption of coal, oil, and natural gas have increasingly promoted solar energy as a limitless and pollution-free energy source. However, in spite of enthusiasm and impressive research and development (R & D) results in the field of photovoltaics (PV), this technique of regenerative energy use is, more than 40 years after its discovery, still at its infancy with respect to the total power of worldwide installations and contributions to the energy demand of mankind. The reason is very simple: from the economic point of view, PV up to now cannot compete with conventional large-scale electric power production. This situation might be different if external costs of conventional power production, especially environmental damage, would be "internalized". However, discussions concerning the quantification of external costs are very controversial [1], [2], and a multinational or global agreement about enhanced installation of renewable energy systems could not yet be achieved. An essential substitution of conventional power generation by solar technologies causes enhanced energy cost, a burden that is believed to reduce the international competitiveness of a country's economy.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_161

The history of photovoltaics is summarized in the following list:

1839	Photovoltaic effect observed at electrodes immersed in an alkaline electrolyte (A. E. Becquerel [3])
1876	Photovoltaic effect observed at selenium crystals (G. W. Adams and R. E. Day [4])
1930	Use of Cu_2O-Cu- and selenium photocells for light measurements
1954	First ″solar cell" based on crystalline silicon (c-Si) for power production; efficiency 6 % (D. M. Chapin, C. S. Fuller, G. L. Pearson [5])
1954	Solar cell made of copper-coated CdS (D. C. Reynolds et al., [6])
1954	Report about photovoltaic effect in CdS thin films (G. Nadjakov et al. [7])
1954	Terrestrial use of photovoltaics for telecommunications, military, and portable transistor radios; however, new technology not accepted by market
1956	GaAs solar cell
1956	Planning of extraterrestrial use of photovoltaics in USA (after start of "Sputnik" by the former USSR)
1958	Launching of first PV-powered satellite (Vanguard I)
1963	CdTe solar cell (D. A. Cusano [8])
1970	"Rediscovery" of photovoltaics for terrestrial use (8th International Photovoltaic Specialists Conference of the IEEE [9])
1973	Solar energy research program of the European Commission including PV at JRC, Ispra, Italy
1973	1st Oil embargo; start of larger PV R & D programs in several countries
1974	Solar cell based on monocrystalline $CuInSe_2$ (S. Wagner, J. L. Shay, P. Migliorato, M. Kasper [10])
1976	Polycrystalline – later called "semicrystalline" or "multicrystalline" – Si solar cell
1976	Amorphous silicon (a-Si) solar cell (D. E. Carlson, C. R. Wronski [11])
1978	Electronic calculator powered by c-Si solar cells
1980	Electronic calculator powered by a-Si solar cells
1980	Worldwide annual PV module production about 5 MW
1983	6.4 MW-PV power plant in USA (Carrisa Plains, CA)
1988	Efficiency of more than 30 % obtained with a GaAs – c-Si PV tandem structure
1990	Worldwide annual PV module production close to 50 MW
1996	Worldwide annual PV module production 90 MW

The year 1954 represents the beginning of the era of modern photovoltaics. Various kinds of photovoltaic cells have been developed and conversion efficiencies have been improved continuously to the present maximum levels given below, as confirmed by the standard measurement laboratories – National Renewable Energy Laboratory (NREL), Golden, CO, USA; Fraunhofer-Institut für Solare Energiesysteme (FhG-ISE), Freiburg, Germany; Sandia, USA; and the Japanese Quality Assurance Organization (JQA) [12]:

Single-crystalline Si cells	24 %
a-Si cells (not stabilized)	12.7 %
$CuInSe_2$-based cells	17.7 %
CdTe-based cells	15.8 %
Si concentrator cells (96 suns∗)	26.8 %
GaAs one sun∗ cells	25.3 %
GaAs/GaSb tandem concentrator cells (100 suns∗)	32.6 %

∗1 sun corresponds to the intensity of ordinary sunlight, 1 sun = 100 mW/cm^2.

3. Fundamentals

3.1. Features of Photovoltaic Cells

Photovoltaic cells differ from other power-generation methods by the following advantages:

1) Solar energy is limitless (approximately 75×10 GW of sunlight reaches the Earth's surface)
2) Photovoltaic energy is clean, noiseless, and does not require any moving parts. In principle it can be considered to be maintenance-free
3) Photovoltaic cells generate power with constant efficiency on a large or small scale (modularity: i.e, the same solar cell materials and structures are used, independent of system size) which contrasts with the efficiency of water and steam power generation systems which depends on the scale of the system
4) The irradiance threshold for start of output operation is very low, practically no energy inventory is needed within a PV system for operation
5) Photovoltaic cells can generate power at the place where it is to be used
6) Even diffuse light can be used to generate power (e.g., fluorescent and other indoor lighting)

However, the following disadvantages must also be considered:

1) The incident energy on a photovoltaic cell is sparse
2) There is no direct power storage function (power can only be generated under sunlight and cannot be stored in the dark.

3) Power output depends on weather conditions and incident light. Consequently, the reliability of power generation by PV power plants generally is not very high. This holds especially for Central European and other moderate climatic regions. In addition, in those regions the daily and seasonal variations of energy demand and power delivery by PV are out of phase. That means, the capacity factor and the capacity value of PV power plants are low. Thus, according to standard economical calculations, the value of PV power supplied to the public grid is low.

3.2. Principle of the Photovoltaic Effect and Characteristics of Photovoltaic Cells (→ Semiconductors, Chap. 3.)

In principle the photovoltaic effect is based on the transfer of the energy of light quanta (photons) to the electronic subsystem of a semiconductor (*internal photoeffect* → Photoelectricity, Chap. 3.), and the "collection" of this energy within a short time – before it is transferred to the atomic lattice, i.e,. converted into heat. For this energy collection an internal field in the semiconductor is required. Suitable devices are *p-n* homojunctions, *p-n* heterojunctions, *p-i-n* junctions, Schottky barriers, and metal – insulator – semiconductor junctions (MIS junctions). For introduction into basic semiconductor physics and *p-n* junction theory see, e.g., [13–15]. Figure 1 A shows a schematic diagram of a photovoltaic cell with a *p-n* junction. Since conversion of solar energy requires larger areas of the converters, a solar cell in principle must be a large-area diode with the *p-n* junction parallel to the radiation-exposed surface.

The energy band diagram of such a structure is shown in Figure 1 B. At the interface between *p*- and *n*-type conduction, carrier diffusion forms a new equilibrium state, and the negative charge of the ionized acceptors at the *p*-side as well as the positive charge of the ionized donors at the n-side create an electric field in this region (*space charge region* or *depletion layer*).

The basic processes of the photovoltaic effect are indicated in Figure 1 B: Absorption of light means that the energy of a photon is transferred to one electron in the semiconductor, regularly to one located in the valence band. If the quantum energy of the photon is greater than the energy gap E_g of the semiconductor, the electron is lifted up into the conduction band and leaves one hole in the valence band. Holes can be treated like positive carriers, though in reality they are nothing more than missing electrons. In total, one electron – hole pair is generated (see process I). The equivalent process also takes place at the *n*-side of the junction and within the space charge region. Thus, the energy of the light quanta is transferred to the electrons. The energy collection occurs by charge separation: photogenerated minority carriers (electrons in *p*-type material and holes in *n*-type material) diffusing to the junction are swept across the space charge region by the internal field and contribute to the building-up of a photovoltage, i.e., a positive charge at the *p*-side and a negative charge at the *n*-side of the junction. If an external load is connected to the cell, a current flows through the load and, consequently, the device operates as a generator.

3.3. Energy Conversion and Energy Losses in Photovoltaic Cells

During conversion of energy, especially of solar radiation, two inherent loss mechanisms occur (see Fig. 1 B):

1) The absorption of photons with quantum energies $h\nu$ considerably greater than E_g – those of the blue part of the spectrum – lifts electrons up to higher energetic regions of the conduction band. The excess energy ($h\nu$ - E_g) of the photons is converted into kinetic energy of the electrons. These electrons are named "hot" electrons. Hot electrons, however, are subjected to a very fast "thermalization" process within the conduction band, i. e., the excess energy is transferred to the atomic lattice within a time interval of less than 10^{-12} s (process II in Fig. 1 B) and will, therefore, not contribute to the photovoltaic effect.
2) light quanta with $h\nu < E_g$ cannot be absorbed by the semiconductor.

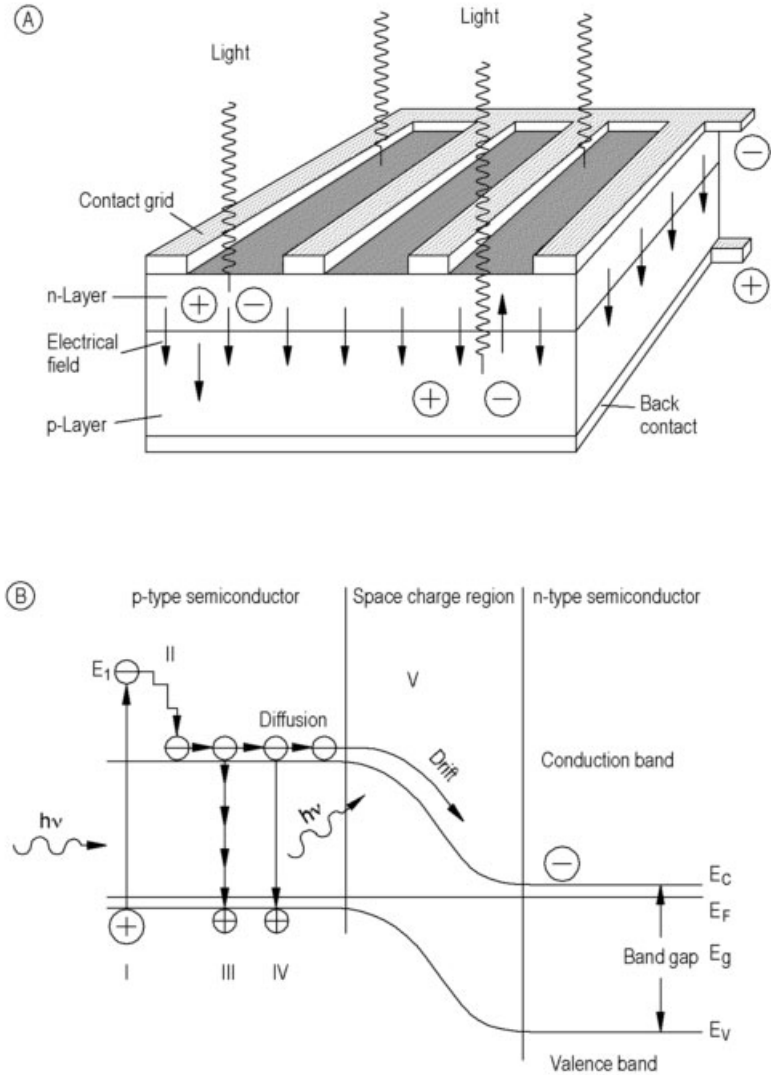

Figure 1. A) Schematic structure of a photovoltaic cell based on a *p-n* junction;
B) Energy band diagram of a semiconductor *p-n* junction, processes causing the photovoltaic effect, and loss mechanisms

Furthermore, it follows from Figure 1 B that the amount of energy of an absorbed photon that can be converted into electrical energy increases with increasing band gap E_g. The theory of the photovoltaic effect also shows that the open circuit voltage of a solar cell increases with increasing E_g (see, e.g., [16]).

Consequently, with respect to the choice of semiconductor materials for solar cells the following situation results:

To obtain high photovoltages, materials with large band gaps should be chosen
To absorb as many photons as possible and to obtain high photocurrents, materials with small band gaps should be chosen

These contradictory requirements indicate that for solar energy conversion an optimum band gap exists. According to calculations carried out by various authors (see, e.g., [17], [18]), the optimum band gap is in the range of 1.3 to 1.6 eV.

The electrons and holes generated by photon absorption will not all be collected at the contacts of the photovoltaic cell. While diffusing, they may recombine with carriers of opposite charge. These loss processes (III and IV in Fig. 1 B) depend on the diffusion length L_n of the electrons in the *p*-side and on the diffusion length L_p of the holes in the *n*-side of the device. The recombination rate decreases and the diffusion length increases with decreased defect density in semiconductors, i.e., with increased material quality (→ Semiconductors, Chap. 3.3.).

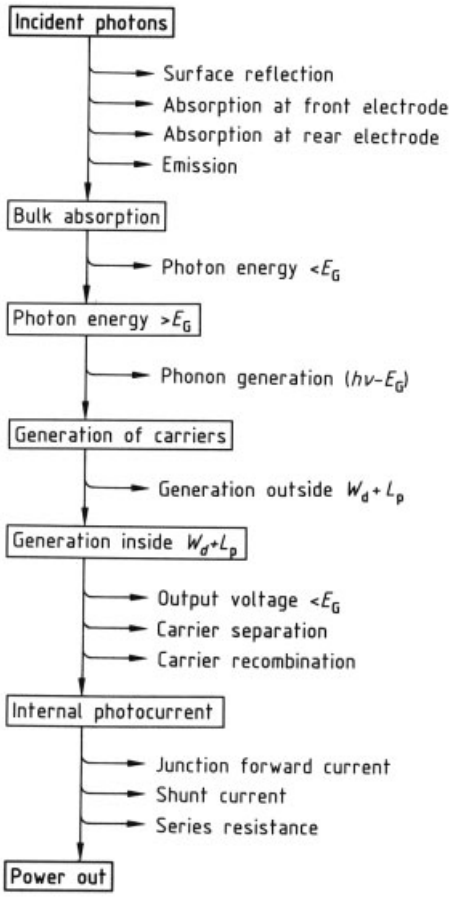

Figure 2. Energy conversion and energy losses in photovoltaic cells
E_G=band gap; $h\nu$=photon energy; W_d=width of depletion layer; L_p=diffusion length of charge carrier (electrons and holes)

Figure 2 summarizes the losses occurring in the energy conversion process of photovoltaic cells. In order to further categorize the losses in solar cells, there are two main energy losses in the process, namely the optical loss and the electrical loss. The *optical loss* includes, e.g., losses due to reflection at the surface of the solar cell and absorption of light at the rear electrode. As a first approximation, the photon flux Φ which can be absorbed in the semiconductor can be expressed as follows:

$$\Phi \approx \int_0 S(E)\cdot[1-R(E)] \cdot \{1-\exp[-\alpha(E)\cdot d]\}\,\mathrm{d}E \approx \int_{E_G} S(E)\cdot[1-R(E)] \cdot \{1-\exp[-\alpha(E)\cdot d]\}\,\mathrm{d}E \quad (1)$$

where $S(E)$, $R(E)$, and $\alpha(E)$ are the energy (E) dependencies of the photon density of sunlight S, of the surface reflectance R, and of the absorption coefficient α, respectively, and d denotes the thickness of the semiconductor. Photon density is the number of incident photons per unit area per second.

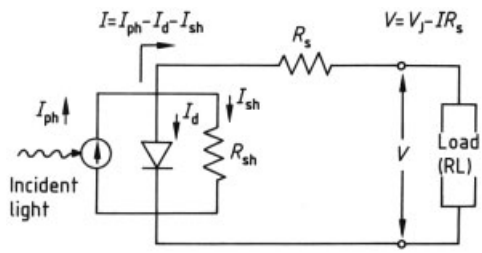

Figure 3. Equivalent circuit of a photovoltaic cell

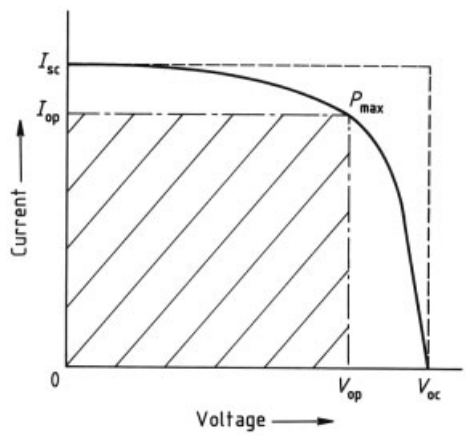

Figure 4. Current – voltage ($I-V$) relationship of photovoltaic cells

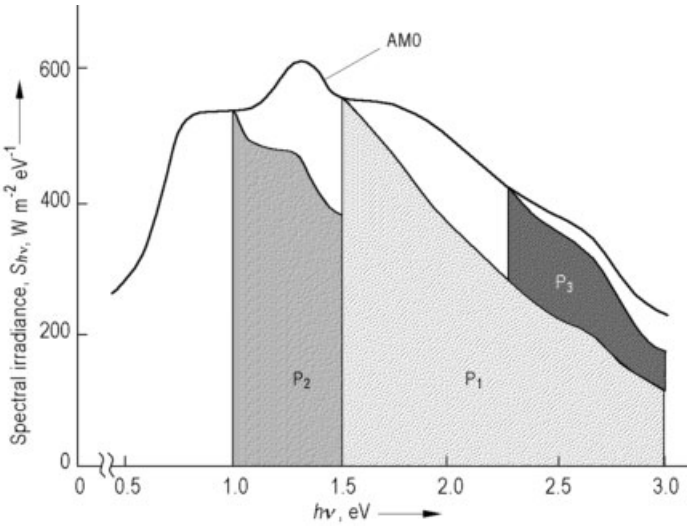

Figure 5. Solar spectral irradiance and usable energy (shaded areas) by a single solar cell (P_1) and by a 3-cell-tandem structure ($P_1 + P_2 + P_3$)

In addition to these optical losses, *electrical losses* must be considered. They are caused by recombination processes (processes III and IV in Fig. 1 B) and by series and shunt resistances in the solar cell. These are considered in the equivalent circuit of the cell (see below, Fig. 3).

3.4. Equivalent Circuit and Current – Voltage Characteristic of Solar Cells

The equivalent circuit of photovoltaic cells is shown in Figure 3 [19]. Here, I_{ph} is the photocurrent, I_{d} the diode current, I_{sh} the shunt current, and I the output current; R_{s} is the series resistance, R_{sh} the shunt resistance, V the output voltage, and V_{J} the junction voltage. Two kinds of photocurrent exist in a photovoltaic cell: diffusion current (current flow according to concentration gradients) and drift current (current flow under an electric field). The output current $I(V)$ under illumination is

$$I(V) = I_{\mathrm{ph}}(V) - I_{\mathrm{d}}(V) - I_{\mathrm{sh}}(V) \tag{2}$$

where the diode current I_{d} is defined as follows:

$$I_{\mathrm{d}}(V) = I_0 \left[\exp\left\{q\left(V + IR_{\mathrm{s}}\right)/AkT\right\} - 1\right] \tag{3}$$

where I_0 is the reverse saturation current, q the electronic charge, A the diode quality factor, k the Boltzmann's constant, and T the absolute temperature. The diode quality factor A indicates the quality of the semiconductor junction. When an ideal semiconductor junction is used, diffusion current dominates and $A = 1$; when a poor quality semiconductor junction is used, recombination current dominates and $A = 2$. In practise, both currents exist and $n = 1 - 2$.

The shunt current I_{sh} can be written as follows:

$$I_{\mathrm{sh}}(V) = \left(V + IR_{\mathrm{s}}\right)/R_{\mathrm{sh}} \tag{4}$$

Figure 4 shows an example of the current–voltage (I–V) relationship for a photovoltaic cell; I_{sc} is the short circuit current (i.e., when $V = 0$), V_{oc} the open circuit voltage (i.e., when $I = 0$), I_{op} the optimum operating current, V_{op} the optimum operating voltage, and $P_{\max}$ the maximum output power ($P_{\max} = I_{\mathrm{op}} \times V_{\mathrm{op}}$). The fill factor ($F.F.$) is determined as follows:

$$F.F. = P_{\max}/\left(I_{\mathrm{sc}} \times V_{\mathrm{oc}}\right) \tag{5}$$

The conversion efficiency η of the photovoltaic cell is

$$\eta = P_{\max}/\left(S \cdot P_{\mathrm{in}}\right) \tag{6}$$

where S is the area of the photovoltaic cell and P_{in} is the radiant power incident on the photovoltaic cell per unit area. The conversion efficiency η is the most important parameter of a photovoltaic cell.

3.5. Tandem Structures

A way to reduce the inherent losses mentioned above is to apply tandem structures. This has already been recognized and discussed by JACKSON [20] in 1955. The effect of a tandem configuration is twofold:

1) When a second solar cell with a lower band gap E_g is placed behind the first cell (which now has to have a transparent back contact), then this second cell converts light quanta which have not been absorbed by the first cell.
2) When a third cell with a higher band gap E_g is placed in front of the first cell, then the thermalization losses (process II of Fig. 1 B) are lower than in the first cell.

These effects are visualized in Figure 5: The solid line AM 0 ("air mass 0," see → Solar Technology, Chap. 1.2.) shows – in this case the extraterrestrial – solar spectral irradiance. The hatched areas indicate the maximum power that can be converted:

P_1: by a single solar cell (cell 1),
P_2: by an additional cell (cell 2) with $E_{g2} < E_{g1}$, and
P_3: by placing a third cell (cell 3) with $E_{g3} > E_{g1}$ in front of cell 1

Quantitatively, these tandem effects have been calculated by various authors (see, e.g., [21], [22]). With a three-cell tandem structure the theoretical efficiency goes up to about 50 %.

3.6. Types of Semiconductor Junctions

There are various kinds of semiconductor junctions for photovoltaic cells, the most important of which is the *p-n* junction. Two different types of *p-n* junction exist, the homojunction and the heterojunction. The *homojunction* joins *p*- and *n*-type layers of the same semiconducting material; examples are *p*-Si/*n*-Si and *p*-GaAs/*n*-GaAs junctions. *Heterojunctions* join different semiconducting materials; examples are *p*-GaAlAs/*p*-GaAs/*n*-GaAs and *n*-CdS/*p*-CdTe junctions (see Section 4.4).

p-i-n Junctions are modifications of the *p-n* junction. Here an intrinsic undoped layer is inserted between the *p*- and *n*- layers. Other kinds of photovoltaic cells are Schottky-type and MIS-type photovoltaic cells. Schottky-type cells consist of a semitransparent metal layer and an undoped or slightly doped semiconductor layer. In the MIS-type cells, a thin insulator layer is inserted between the metal and the semiconductor.

Heterojunctions join different semiconducting materials. Examples are *p*-GaAlAs/*p*-GaAs/*n*-GaAs and junctions between compounds such as chalcogenides (CdS, CdSe, CdTe,...) and chalcopyrites ($CuInSe_2$, $CuGaSe_2$, $CuInS_2$, and their quaternary and pentenary alloys). These materials generally are not amphoteric, i.e., they only show one type of conductivity, either *p* or *n*. Therefore, *p-n* and *p-i-n* structures as in case of crystalline or amorphous silicon are not feasible. Instead, heterojunctions or MIS-structures must be prepared in order to realize solar cells. This situation not necessarily is a disadvantage. Theoretically, efficiencies of heterojunction solar cells are equal to those of *p-n* homojunction cells (see, e.g., [23]). The semiconductor with the lower band gap determines the voltage of the solar cell. This semiconductor is the decisive partner in a heterojunction and is called *absorber*. Its optimum band gap is just the same as in the case of a homojunction solar cell. The wide-band gap partner is called *collector*, however, in some structures it also absorbs part of the incident radiation. Heterojunctions even have some features that are advantageous with respect to fabrication of solar cells. Three different structures of heterojunction solar cells have been applied:

Front-Wall Solar Cells. The absorber is placed to the front of the cell, it faces the sun. So, the density of photogenerated carriers is highest at the surface of the absorber. Since the carriers have to cross the whole absorber in order to be collected at the junction, the thickness of the absorber must be optimized with respect to absorption depth and diffusion length. If the absorber is too thin, a too large part of the incident radiation is not absorbed; if the absorber is too thick, a too large part of carriers generated at the surface will – due to recombination – not reach the junction. Semiconductors with high absorptivity such as direct semiconductors (see Chapter 4) are favorable in those

structures. If the band gap of the collector is only slightly wider than that of the absorber, then the thickness of the absorber can be reduced and the collector can contribute to the absorption of the incident light. The spectral sensitivity of such a structure covers the range $h\nu > E_g$ (absorber).

Back-Wall Solar Cells. In this case the collector faces the sun. Depending on the thickness and on the optical and electronic properties of the collector, it may or may not contribute to the photocurrent. If not, then the spectral sensitivity of this structure is limited to the range $E_g(\text{absorber}) < h\nu < E_g(\text{collector})$.

Window-Absorber Structure. The window-absorber structure is a special case of the back-wall solar cell. If the band gap of the collector is sufficiently wide, it operates as a "window." This structure is the optimum case of a heterojunction solar cell. The incident light is mainly absorbed in the immediate neighborhood of the *p-n* junction, so that carrier collection and charge separation operate under optimum conditions. Figure 6, showing a band diagram of a heterojunction, indicates the window-absorber-principle.

The two partners of a heterojunction must fit together with respect to lattice constant, thermal expansion coefficient, and electron affinity. Mismatch of these parameters results in the formation of so-called interface states, i.e., energy levels at the interface that are – in the energy scale – distributed across the energy band gap. They increase the probability of recombination of electrons in the conduction band with holes in the valence band. The higher the density of these interface states, the higher will be the recombination losses of photogenerated carriers and, consequently, the losses in photocurrent. In addition, mismatch of electron affinities results in a reduced photovoltage. As indicated in Figure 6, buffer layers might be necessary for compensating mismatch between window and absorber layer. The choice of an appropriate buffer layer technology is essential for high efficiencies.

4. Types of Photovoltaic Cells

Materials with band gaps in the range $1.3\,\text{eV} < E_g < 1.6\,\text{eV}$ are suited optimally for photovoltaic cells (see Section 3.3). However, crystalline silicon (c-Si) with a band gap of only 1.1 eV is the number 1 champion since the beginning of the era of photovoltaics. This is due to its availability, the advanced knowledge of its physical properties, and the know-how of material handling and device technology. The simultaneously growing electronic semiconductor industry stimulated and enforced the development of c-Si solar cells. A decisive property of semiconducting materials with respect to PV technology and cost reduction potential is the absorption coefficient of light α. In this concern *direct* and *indirect* semiconductors have to be distinguished. Direct semiconductors show α values that are about two orders of magnitude higher than those of indirect materials. Consequently, for light-absorbing devices such as solar cells, indirect semiconductors demand a thickness that is about 100 times as thick as in case of direct semiconductors. Crystalline Si is an indirect semiconductor, whereas amorphous Si as well as the most prominent chalcogenide and chalcopyrite solar cell materials are of direct character. Consequently, the realization of cost effective thin film solar cells is easily possible with direct materials. With crystalline Si very special measures are required; some promising approaches exist (see Section 4.3).

4.1. Single- and Multicrystalline Solar Cells (→ Silicon)

The production of crystalline silicon solar cells can be subdivided into three areas:

1) Preparation of high-purity silicon
2) Production of wafers, ribbons, or sheets
3) Production of solar cells (*p-n* junction, contacts, etc.)

Figure 7 shows the manufacturing process and the cell structure of "classical" single-crystalline Si solar cells.

Preparation of High-Purity Silicon. Up to now practically all crystalline Si solar cells are made from "electronic-grade" silicon (EG-Si), an ultrapure material (impurity content $< 10^{-9}$) produced by the so-called Siemens

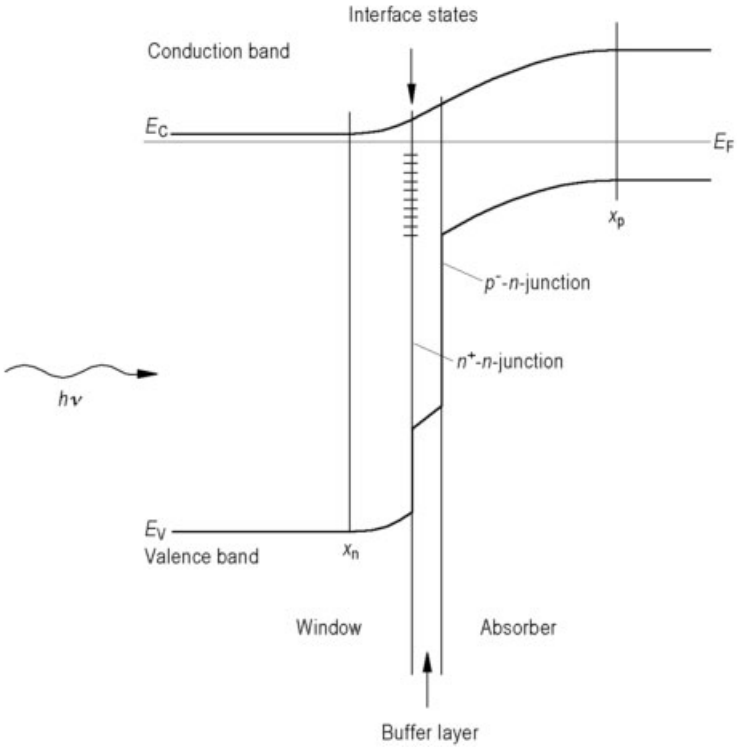

Figure 6. Band diagram of a heterojunction with a multilayer window

Process, which is based on fractional distillation of $HSiCl_3$ (trichlorosilane), a gaseous compound fabricated from metallurgical-grade silicon (MG-Si). A pyrolytic process retransforms the ultrapure trichlorosilane into polycrystalline silicon (poly-Si), the starting material for Si crystal growth (see Fig. 7), multicrystalline Si block casting, Si ribbon pulling, etc. Worldwide a number of efforts have been undertaken for substituting the Siemens Process by simpler and lower-cost processes [25–27]. The intention was to produce low-cost "solar-grade" silicon (SoG-Si) that is still good enough for solar cells. However, the relevance of these processes has decreased in recent years, since (a) the main efforts in Si solar cell research have been directed to efficiency enhancement, and (b) as long as the material consumption of the PV industry is $< 10\,\%$ of the total electronic-grade Si-production, the photovoltaic industry can make use of low-cost "off-specification" EG-Si and of "remelts." However, these resources are becoming scarce.

Crystal Growth. Silicon single-crystalline rods of standard quality are made using the Czochralski (CZ) method. Figure 7 shows the manufacturing process. Single crystals of top quality require higher efforts, they demand for the floating zone (FZ) method.

Czochralski Method (→ Crystal Growth). Starting material is Poly-Si obtained by pyrolytic reduction of trichlorosilane. In the Czochralski process (Fig. 7), a seed crystal of silicon on a holder is dipped into a silicon melt and then pulled up very slowly by mechanical control at a rate of about 2.5 cm every three hours. The molten material starts solidifying on the seed crystal, forming a crystalline pattern identical to that of the seed crystal.

Floating Zone Method. The floating zone method (→ Crystal Growth, Chap. 6.4.) consists of six process steps:

1) A zone of a solid rod of the nutrient material poly-Si (formation from trichlorosilane) is

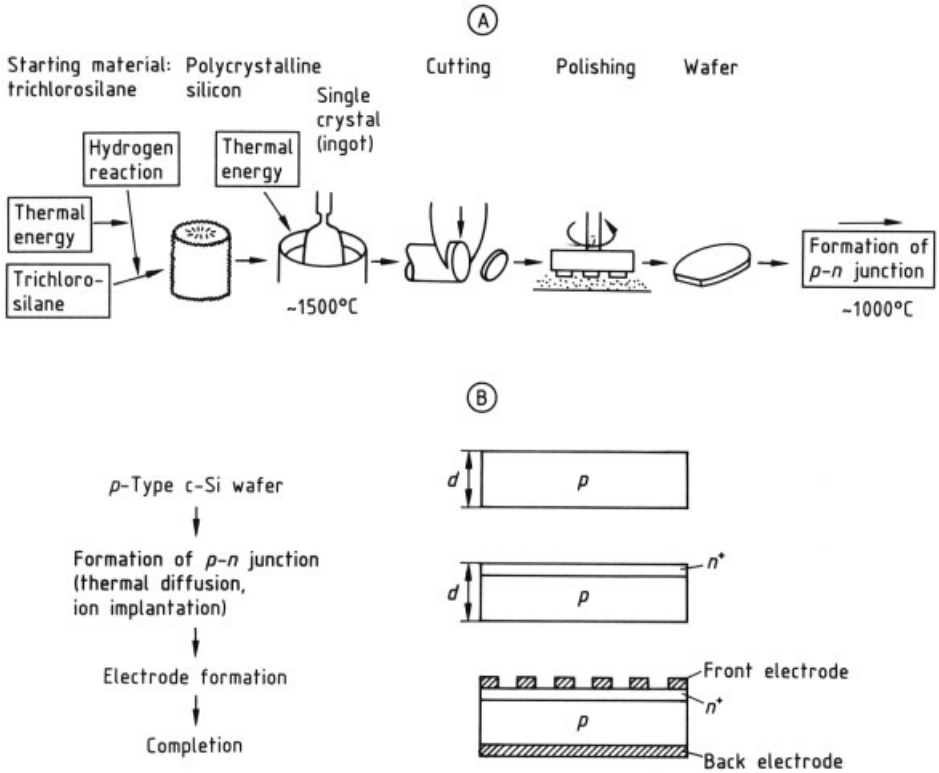

Figure 7. Manufacturing process and cell structure of c-Si solar cells
A) Manufacturing process of c-Si wafers; B) Manufacturing process of c-Si solar cell and the cell structure n^+ represents a highly doped layer

heated by a high-frequency power generator and a molten zone is formed
2) The molten zone is put into contact with a silicon seed crystal
3) The zone is moved away from the seed crystal
4) A single crystal with increasing diameter grows after the molten zone has moved
5) The growing crystal is supported
6) The growing crystal is separated from the nutrient material poly-Si

The floating zone method produces c-Si single crystals of the highest quality available, however, the production costs are so high, that this technology will never be a basis for mass production of Si solar cells.

Multicrystalline Ingot Production. In order to achieve cost reduction, methods to avoid the costly techniques of single crystal production were investigated. Many approaches to prepare both low-cost and high-quality materials have been attempted. The most successful technology is the casting method resulting in "multicrystalline" ingots. Conventional designations for various crystalline states are given below:

Single- or monocrystalline	grain size > 10 cm
Multicrystalline (also called "semicrystalline")	grain size 1 mm – 10 cm
Polycrystalline	grain size 1 µm – 1 mm
Microcrystalline	grain size < 1 µm
Nanocrystalline	grain size < 100 nm

In preparing multicrystalline Si ("multi-c-Si") ingots suitable for solar cell production, the following objectives must be achieved:

1) Enlargement of grain size.
2) Absence of columnar structure in the depth. Grains of multi-c-Si have a columnar structure and it is important that there is no grain boundary, since carriers would otherwise flow across the grain boundary resulting

in recombination loss due to defects in the boundary. Grain boundaries are barriers to majority carriers (i.e., electrons in the *n*-layer and holes in the *p*-layer) and are recombination centers for minority carriers (i.e., holes in the *n*-layer and electrons in the *p*-layer).

3) Passivation of grain boundaries.
4) Exclusion of impurities.

In order to solve these problems, several methods have been developed. Many methods exist for preparing multi-Si substrates [28].

Casting Method. The casting method is shown in Figure 8. A silicon ingot can be obtained by gradually cooling molten silicon in a special pot or crucible. A silicon source is melted in a furnace, then poured into a special carbon pot with a carefully controlled vertical temperature gradient. The ingot quality depends on the cooling rate and the pot shapes and material. The ingot is then carefully sliced into thin sheets or wafers with a grain size of ca. 0.5 – 1.0 cm.

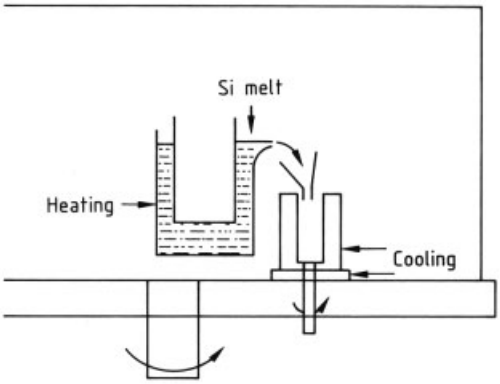

Figure 8. Casting of poly-Si substrates

The concentrations of impurities such as oxygen and carbon in the ingots are $10^{16} - 10^{17}\,\mathrm{cm}^{-3}$ and $\approx 10^{17}\,\mathrm{cm}^{-3}$, respectively. The concentration of impurities increases from the top to the bottom of the ingot ($\approx 1\times 10^{18}\,\mathrm{cm}^{-3}$ at the bottom). The conversion efficiency of solar cells using multi-c-Si obtained by the casting method decreases with increasing concentration of impurities ($\leq 1\times 10^{18}\,\mathrm{cm}^{-3}$ impurities should be present).

Casting of multicrystalline blocks is somewhat cheaper than single crystal growth, however, cells based on this material exhibit lower efficiencies. Cast material is produced by Wacker (SILSO) and Bayer (Germany), Photowatt (France), Crystal Systems and Solarex (United States), and Osaka Titanium (Japan).

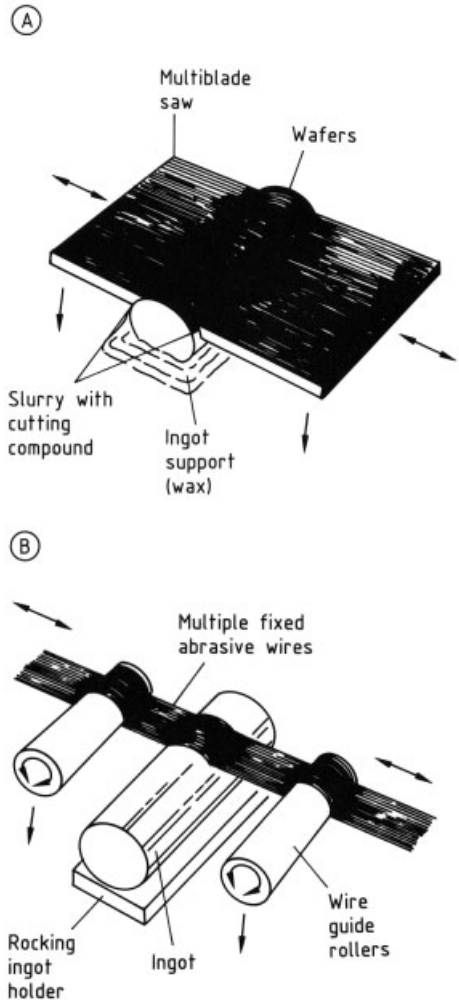

Figure 9. Sawing methods for ingots
A) Multiblade saw; B) Multiwire saw

Cutting. Single-crystalline rods or multicrystalline ingots are carefully sliced into wafers. The objective is to obtain as many wafers as possible from one rod and to minimize losses in the form of "kerf" or sawdust. Typical wafer thickness is 0.35 mm, kerf losses can amount to ca. 50 %. Much has been done to improve sawing techniques and speeds. Thinner sawing, multibladed circular saws, and different types of wire saws have been developed, some of which

can cut 1000 wafers simultaneously (see Fig. 9). These processes are complicated and require a large amount of energy. Therefore, the fabrication costs of c-Si solar cells are high.

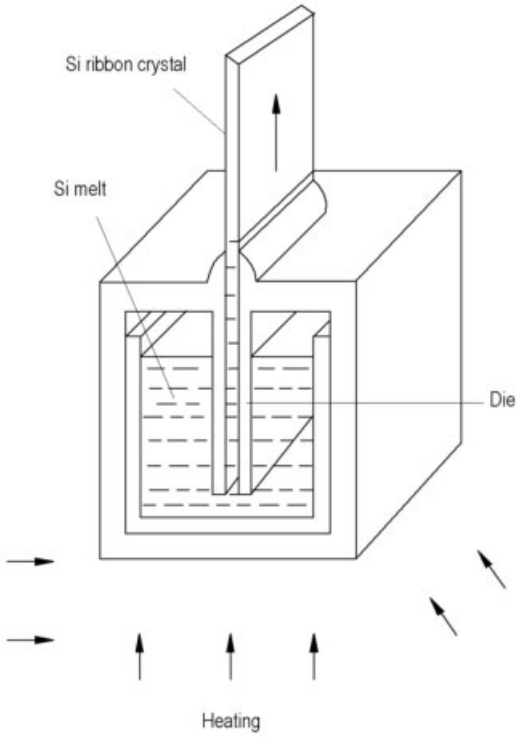

Figure 10. Manufacturing process of multi-c-Si substrates: ribbon growth method

With advanced multiwire sawing techniques remarkable reductions of kerf losses have been achieved. Furthermore, thinner wafers can be produced; the PV industry is on the way to introduce a 200-µm-wafer technology. The brittleness of the wafers and the resulting handling problems are the main obstacles

Ribbon and Sheet Growth Technologies. In order to dispense with sawing techniques that always cause material losses and require a number of production steps, various types of ribbon pulling and casting and foil preparation techniques have been investigated (altogether about 20 different methods). An overview of Si ribbon and sheet developments is given in [29]. Today, the edge-defined film-fed growth (EFG), is the only process used for the industrial production of ribbon-type Si solar cells (ASE Americas, formerly Mobil Solar). The ribbon growth method is shown in Figure 10. A graphite or ceramic die, which has a thin slit through the center (slit width ca. 400 – 600 µm) is immersed in a pool of molten silicon. A silicon sheet (ribbon) can be drawn up through the slit by capillary action.

At ASE Americas, multicrystalline silicon sheets are grown in this way in the form of large octagonal tubes which are cut into wafers by laser. The time and material consuming wafer slicing and etching steps can thus be omitted. Because of the special growth conditions (oxide and carbon content, surface roughness etc.) a special optimized solar cell technology had to be developed.

Fabrication of Solar Cells. The standard process for Si solar cell production based on Si-wafers (mostly *p*-type) is the formation of a *p-n* junction by phosphorous doping, using a thermal diffusion technique from the gaseous phase. A number of alternative techniques also have been approved.

Thermal Diffusion. After the wafers have been cut, they are coated on one side, then passed through a furnace containing phosphorus vapor (or another element with five outer shell electrons, e.g., As or Sb) for just enough time to allow the phosphorus to diffuse a short distance into the silicon, converting it into a very thin layer of *n*-silicon. At the front of deepest penetration of phosphorus, a static electricity barrier is formed spontaneously and immediately.

In the thermal diffusion technique, it is difficult to implement an abrupt decrease (to zero level) of the phosphorus concentrations at the *n-p* interface, usually the phosphorus distribution follows the shape of an error function.

Alternative Technologies. More abrupt donor distributions can be obtained, e.g., by ion implantation. Other techniques result in reduced cost when they are applied in conjunction with especially cheap technologies such as screen printing of doping and contacting materials. The most important alternative techniques are:

1) Diffusion out of a solid phase previously deposited by chemical vapor deposition (CVD), spin-on, or screen printing (screen printing is successfully applied in large-scale production)

2) Diffusion out of thin solid disks, assembled as a continuous stack (batch process with high throughput)
3) Ion implantation and thermal annealing (thermal annealing also by laser/electron beam annealing)
4) Laser-induced diffusion of dopants deposited to the Si surface by evaporation, sputtering, spin-on, or spray-on
5) Direct-coupled microwave thermal processing (results in quick heating of stacked wafers)

After formation of the *p-n* junction, metal contacts are applied to each side of the cells, usually a metal plate to the bottom or *p*-side, and a grid of metallic fingers to the top or *n*-side.

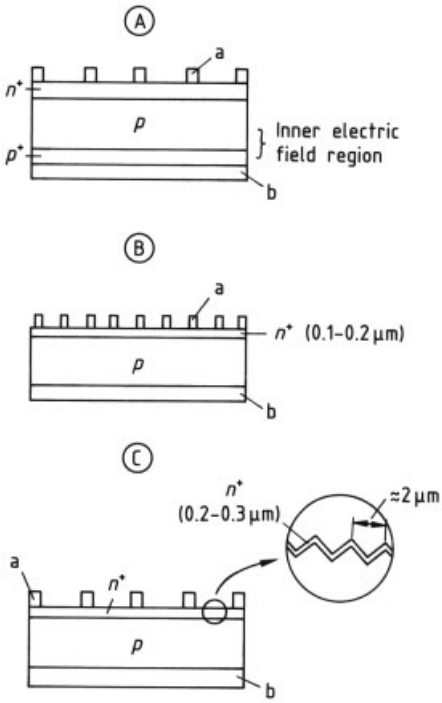

Figure 11. Various cell structures of c-Si solar cells
A) BSF solar cell; B) Violet cell; C) CNR cell
a) Front electrode; b) Back electrode

Milestones of Development. Based on the cell structure shown in Figure 7 B, various structures have been developed to increase conversion efficiency.

Back Surface Field Solar Cell. Figure 11 A shows the structure of a back surface field (BSF) solar cell which was first realized in 1972 [30]. Its principle of operation was theoretically calculated even in 1963 by Martin Wolf [31]. In the BSF cell, an inner field is formed by the incorporation of a highly doped p (a p^+)-layer. The p^+-layer prevents holes and electrons from recombining near the back surface. In BSF solar cells, I_{sc} and V_{oc} increase. The increase of output power by the BSF effect is ca. 15 %, and the response of the solar spectrum for long wavelengths is improved. BSF-type solar cells are usually fabricated by preparation of highly doped layers using thermal diffusion.

Violet Cell. Figure 11 B shows a violet cell which was developed in 1972 by J. Lindmayer [32]. The n^+-layer is very thin (0.1 – 0.2 µm) compared with that of a conventional c-Si solar cell. Because of this thin n^+-layer, a *p-n* junction can be formed very close to the front surface, which prevents photogenerated carriers from recombining near the front surface of the n^+-layer. Photoresponse in the short wavelength region of the solar spectrum is improved by using this structure.

Inversion-Layer Solar Cell. Another solar cell designed to improve the response in the short wavelength region is the inversion-layer solar cell. This solar cell is fabricated by deposition of silicon dioxide on the surface of *p*-type silicon. The positive charge in SiO_2 (e.g., Na^+ adsorbed from the water used in the surface wet-cleaning process before SiO_2 deposition) is fixed near the SiO_2 – Si interface by deposition of SiO_2 on *p*-type silicon. These positively charged ions attract electrons and repel holes. As a result, a thin *n*-type inversion layer is fabricated at the c-Si surface. In the late 1980s this technology was transferred to industrial production [33]. The technology offers the economic advantage of low temperature processing, however, quenching of the fixed charges – an effect that should not occur – is the cloven hoof of the technique.

Comsat Non-Reflecting Solar Cell. Another way to improve conversion efficiency is to reduce the optical losses due to surface reflection and to increase the number of penetrating photons. Comsat non-reflective (CNR) solar cells have been developed for this purpose [34], which have many microscopic pyramids 1 – 2 µm in size at their cell surface to reduce surface reflection (Fig. 11 C). Many pyramidal structures

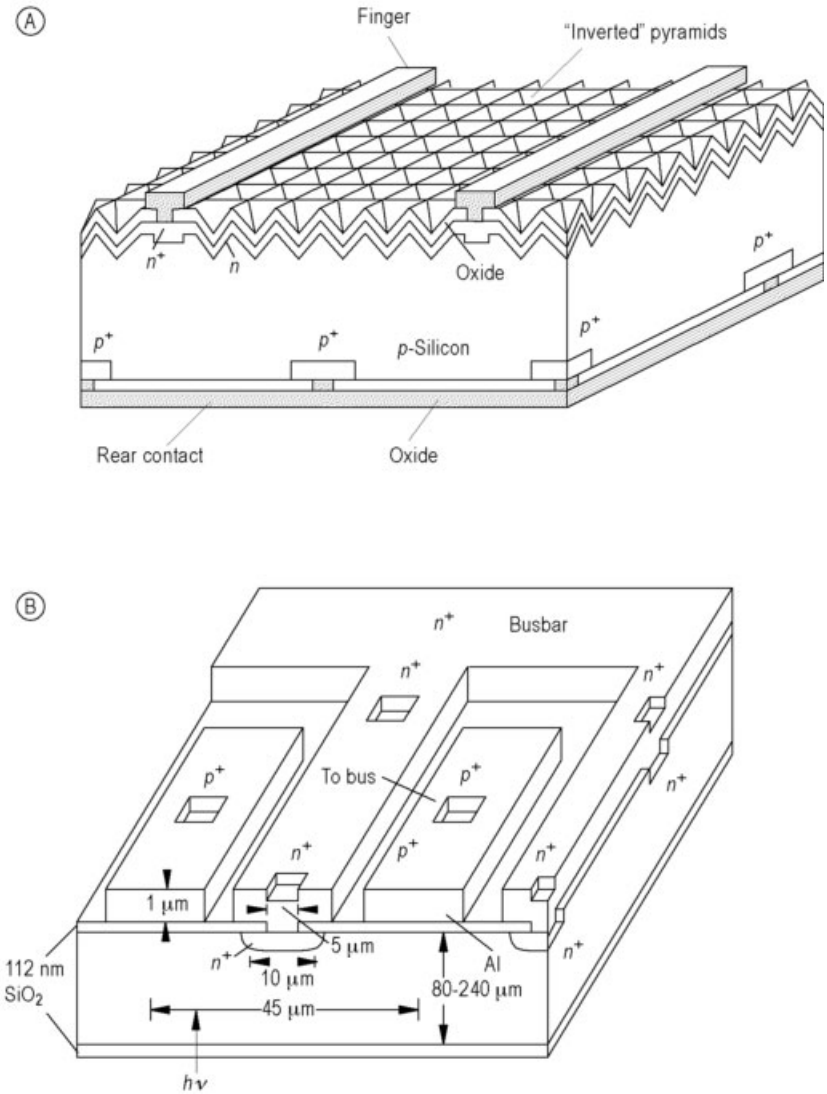

Figure 12. Highly efficient c-Si solar cells
A) Passivated emitter – rear locally diffused (PERL) solar cell [36]; B) Point-contact solar cell with interdigitated backside collectors [38]

with (111) planes are fabricated by the structural etching technique. The silicon wafer used for etching has a flat (100) plane which is etched using concentrated alkali to form microscopic pyramids with four (111) side planes.

Passivated Emitter Solar Cell (PESC) and Passivated Emitter – Rear Locally Diffused Cell (PERL). The idea to passivate Si surfaces and Si interfaces with thin films of SiO_2 has been introduced into PV science and technology by Martin Green, University of New South Wales (UNSW), Australia [35], [36]. It was the last real milestone on the way to highest efficiencies. Passivation reduces the losses due to surface and interface recombination of light-generated carriers in a decisive way. Even the interface between Si and the metallic contacts should be passivated; nonpassivated contact areas should at least be minimized. The PERL cell is shown in Figure 12 A. Inverted pyramid structures of the

surface of the cell allow effective light trapping. As a result, a conversion efficiency of 24.2 % at 1 sun has been reported [36]; a value of 24.0 % has been internationally accepted [12].

Point-Contact Solar Cells with Interdigitated Backside Collectors (Fig. 12 B) [37–39]. The conversion efficiency of photovoltaic cells could be significantly improved also with this novel device configuration. The structure features low metallization series resistance (low resistance in doped layers), lack of shadowing losses (no shadows caused by the electrode on the incident side surface), and lack of heavy doping effects (i.e., the heavily doped semiconductor cannot effectively generate carriers) in regions of maximum photogeneration. Furthermore, the principle of surface and interface passivation is also realized in this structure. A conversion efficiency of 28.4 % in concentrated sunlight of 79 suns has been reported [40]; a value of 26.8 % has been internationally accepted [12]. Those high efficiencies of PESC, PERL, and point-contact solar cells have been achieved with highest-quality FZ Si single crystals; they demonstrate the feasibility of efficiencies close to the maximum theoretical value (28 % at AM 1.5). For industrial production, however, FZ Si is far too expensive. Therefore, CZ-Si based efficiencies are more relevant in practice: the highest efficiency with CZ wafers (21.7 %) has been obtained at the Fraunhofer-Institute for Solar Energy Systems (FhG-ISE), Freiburg, Germany [41].

Summary of Achievements with Crystalline Silicon Solar Cells. Table 1 shows the performance of c-Si solar cells and c-Si PV modules obtained with the technologies described above.

4.2. Amorphous Silicon Solar Cells

Silicon in its amorphous hydrogenated state differs essentially in its physical properties from crystalline Si. It provides a direct optical band gap of about 1.7 eV that can be adjusted continuously from about 1.3 to 2.2 eV by forming alloys with germanium or carbon. The carrier mobility in amorphous silicon is strongly reduced as compared with crystalline Si. Effective charge collection requires implementation of a drift field in the *p-n* junction, i.e., *p-i-n* structures are required (see below).

Spear and LeComber first succeeded in the valency control (substitutional doping) of amorphous silicon (a-Si) (→ Silicon; → Semiconductors) in 1975 [42]. Production of a-Si solar cells was first reported in 1976 [43]. Since then intensive research efforts have concentrated on a-Si solar cell technology.

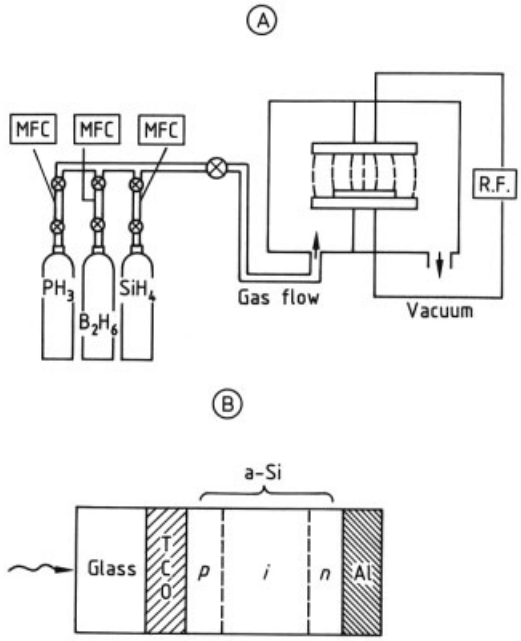

Figure 13. Basic fabrication method and cell structure of a-Si solar cells
A) Single chamber method; B) Basic cell structure
MFC=Mass flow controller

a-Si solar cells have many advantages that cannot be realized by other materials such as c-Si or multi-c-Si. Since the absorption coefficient of the direct semiconductor a-Si is higher than that of c-Si, even a very thin ($< 1\,\mu m$) a-Si solar cell can convert sunlight into electricity. The material costs of a-Si solar cells are therefore low. In addition, both fabrication costs and energy requirement of a-Si cells are low because of the simple and low-temperature fabrication process (see below). Large-area a-Si solar cells can be manufactured by using the plasma CVD method because the method is a gas-phase deposition. It is easy to obtain integrated-type a-Si solar cells using this fabrication process. A high

Table 1. Performance of c-Si solar cells and PV modules according to [12] and other sources

Type of cell/ module	V_{oc}, V	J_{sc}, mA/cm^2	Fill factor	Eff., %	Area, cm^2	Condi- tion	Group
PERL cell, FZ-Si	0.709	40.9	0.827	24.0	4	1 sun	UNSW
PERL-cell module, FZ-Si	5.6	3920	0.801	22.3	787	1 sun	UNSW
PERL/LBSF cell, CZ-Si	0.689	40.5	0.778	21.7	4	1 sun	FhG-ISE
CZ-c-Si modules	top values in laboratories			18	> 500	1 sun	several
CZ-c-Si modules	industrial production			16	> 500	1 sun	several
Multi-c-Si (HEM, USA)	0.636	36.5	0.804	18.6	1	1 sun	Georgia Tec
Multi-c-Si (Bayer)	0.622	33.4	0.790	16.4	98	1 sun	ASE
Multi-c-Si module (HEM)	14.6	1360	0.786	15.3	1017	1 sun	Sandia
Multi-c-Si modules	industrial production			13 – 14	> 500	1 sun	several
Rear contact mono-c-Si				26.8	1.60	96 sun	SunPower

voltage can be obtained from one substrate, as will be mentioned later.

Fabrication. Figure 13 A shows a principle fabrication method for a-Si solar cells. a-Si thin films are generally prepared by a radio-frequency plasma discharge in silane (SiH_4) gas [44]. The SiH_4 molecule is decomposed in the plasma (i.e., SiH_3, SiH_2 etc. radicals are formed) and a-Si films are deposited on a suitable substrate. *p*-Type or *n*-type a-Si films can easily be obtained by adding a small amount of diborane (B_2H_6) or phosphine (PH_3), respectively, to SiH_4 gas. By changing the reaction gases (doping gases e.g., B_2H_6 and PH_3), an a-Si solar cell can be continually deposited directly onto a substrate at relatively low temperatures of 200 – 250 °C. A *p*-layer is obtained with B_2H_6 and SiH_4, an *i*-layer with SiH_4 only, and an *n*-layer with PH_3 and SiH_4.

Figure 13 B shows the basic structure of an a-Si solar cell. An a-Si solar cell is usually deposited onto a glass-transparent conductive oxide (TCO, mainly SnO_2) substrate. a-Si solar cells have a *p-i-n* junction, i.e., an undoped layer (*i* layer) is inserted between the *p* and *n* layers, while c-Si solar cells consist of a *p-n* junction. The *p*-type and *n*-type layers induce an electrical field across the entire *i* layer and photogenerated charge carriers are separated by this electric field. The typical thickness of *p*, *i*, and *n* layers are ca. 10, 500 nm, and 40 nm, respectively.

Many efforts have been made to improve the conversion efficiency and the fabrication methods of a-Si solar cells. The consecutive, separated reaction chamber method is shown in Figure 14 [45]. Here, each layer (*p*, *i*, *n*) of a-Si solar cells is fabricated in a different chamber. This method is not only effective in preventing the mutual intermixing of dopants such as boron or phosphorus, but is also suitable for mass production.

Cell structures have also been improved. a-SiC, which is usually fabricated by glow discharge in $SiH_4 - CH_4$ mixtures, has a lower absorption coefficient than a-Si. Therefore, photoresponse in the short wavelength region has been improved by using *p*-type a-SiC as a window layer for an a-Si solar cell [46] (see Figs. 14 and 15). *p*-Type a-SiC has a higher transparency than *p*-type a-Si and sunlight can pass into the intrinsic region (*i*-layer) where it is absorbed and contributes to the cells output current. A textured TCO-substrate is used in order to increase the photoresponse in the long-wavelength region.

Stability Problems. The conversion efficiency of a-Si solar cells is still lower than that of other types of solar cells. Another serious problem is the degradation of efficiency (Staebler – Wronski Effect). This can be attributed to a light-induced increase of the number of "dangling bonds" in the bulk material acting as recombination centers, which may be caused by (1) breaking of weak bond (Si-H bonds, Si-Si bonds) [47] or (2) breaking of paired bonds [48]. These mechanisms are considered to be system-inherent. Up to now, outdoor stability tests of a-Si cells and generators always resulted in a decrease of efficiency, which is not tolerable for real power applications of a-Si modules. Progress with respect to stability ap-

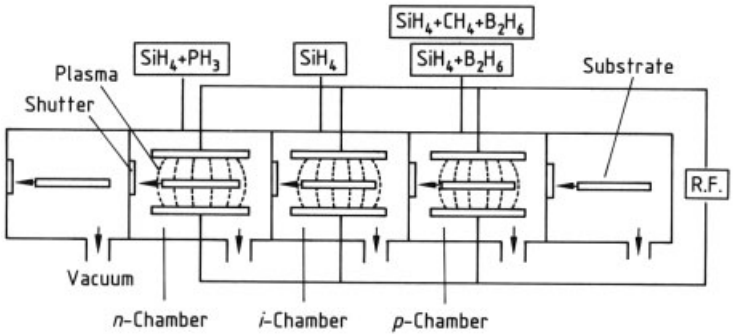

Figure 14. Consecutive, separated reaction chamber method

pears not to be possible by improvement of material quality; structural modifications are believed to be promising. For this purpose high drift fields are implemented, obtainable by reducing the *i*-layer thickness. The accompanying reduction of light absorption is compensated by multijunction cell structures as shown in Figure 15 D. With such structures further improvements of cell efficiencies have also been achieved due to the fact that they act like tandem structures with different band gaps. As discussed in Section 3.5, tandem structures have a potential of achieving higher efficiencies than single-junction solar cells. Initial efficiencies of 13.7 % have been obtained with a-Si-based triple junction cells. Furthermore, "stabilized" or "end of life" efficiencies of 10 to 11.83 % have been announced [49], [50]. This increase of stability is due to the higher electric fields within the *i*-layers of the triple junctions that are smaller than those of single-junction cells. The combination of a-Si and other materials such as multi-c-Si [51] and $CuInSe_2$ [52] has also been proposed (see Fig. 15 C – F).

Integrated-Type a-Si Solar Cells. An integrated-type a-Si solar cell is shown in Figure 16. Here a-Si solar cell units are arranged in cascade fashion on an insulated glass substrate [53]. This makes it possible to obtain a high voltage from one substrate. As a solar cell cannot generate voltages above 1 V, several cells are connected in series for practical use. Conventionally, series connections are made using lead wire, but unlike conventional solar cell modules, the individual cells need not be connected with a lead wire to increase voltage. This means that the structure is highly suited to various solar cell applications. During manufacturing, the TCO, the a-Si layer, and the metal electrode can be selectively patterned by a laser scribing technique [54]. A large-area a-Si solar cell module with this integrated structure can easily be obtained using a laser scribing method.

Conversion Efficiencies. Conversion efficiencies of a-Si solar cells have been greatly improved by combining many of the key technologies described above. The performance of a-Si single cells and a-Si related stacked cells are summarized in Table 2.

In recent years, the selling-prices (cost per peak Watt) of a-Si modules have been in the same range as those of crystalline Si modules. One of possible reasons is the slow deposition rate obtained with present-day industrial technologies (about 1 μm/h). The only way to overcome this problem and to increase the deposition rate by a factor of about 10, without reducing the quality of the deposited layers, at present appears to be the application of very high frequency (VHF) glow discharge techniques (very high frequency, 30 – 300 MHz). Concerning the cost reduction potential of module production, there is still optimism to bring the cost down to the range \$ 0.35 / W_p (W_p = output power, peak Watt).

4.3. Crystalline Silicon Thin-Film Solar Cells

The obstacles that standard-type crystalline Si solar cells are facing on their way to low-cost

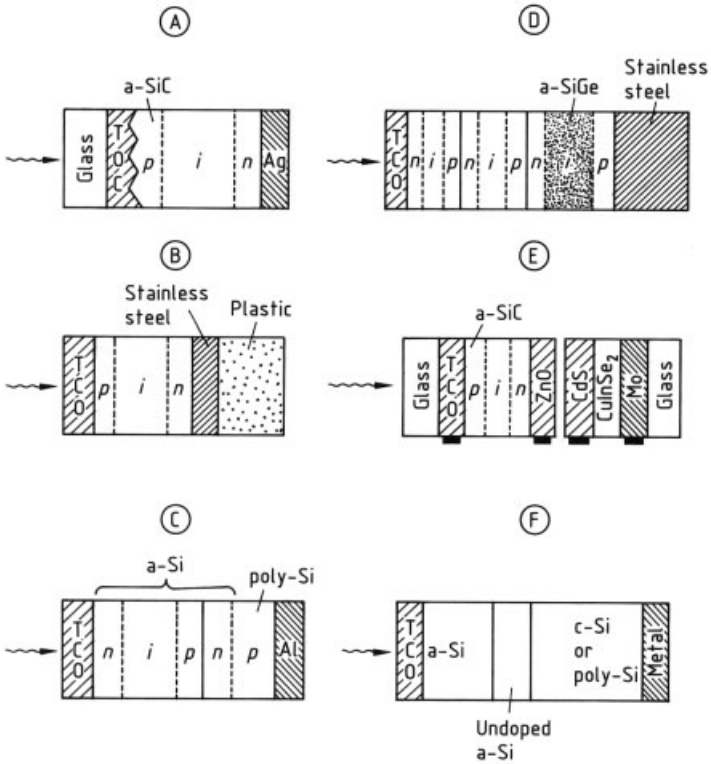

Figure 15. Cell structures of various types of solar cells based on a-Si
A) Textured TCO substrate a-Si solar cell; B) Plastic substrate a-Si solar cell; C) a-Si/poly-Si stacked cell; D) a-Si/a-Si/a-SiGe stacked cell; E) a-Si CdS/$CuInSe_2$ stacked cell; F) a-Si/c-Si or poly-Si stacked cell
TCO = transparent conducting oxide

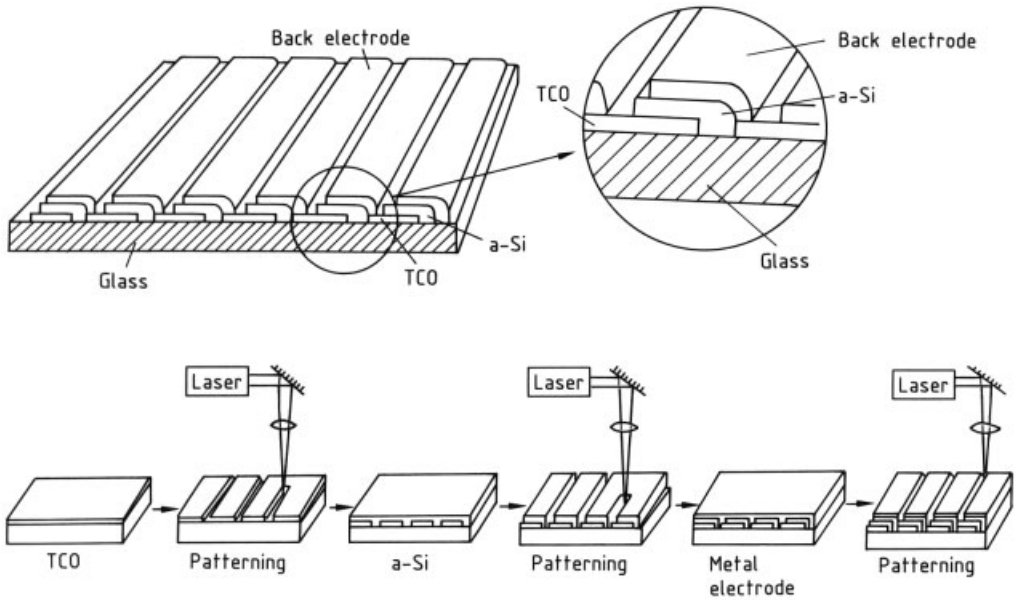

Figure 16. Manufacturing process and cell structure of integrated-type a-Si solar cells

Table 2. Performance of a-Si solar cells and modules; efficiencies marked (i) = initial and (st) = stabilized; after [12] and [55]

Type of cell	V_{oc}, V	J_{sc}, mA/cm$_2$	Fill factor	Effic., %	Area, cm$_2$	Group/ organization
a-Si single junction	0.887	19.4	0.741	12.7(i) ∗	1	SANYO
a-Si single junction				13.2(i)	1	Mitsui-Toatsu
a-Si single junction				8.9(st)	1	SANYO
a-Si single junction				10.2(i)	4050	KANEKA
a-Si/a-Si dual junction	39.3	441 mA	0.696	10.5(i)	1200	Fuji Electric
a-Si/a-Si dual junction				8.9(st)	1200	Fuji Electric
a-Si/a-SiGe dual junction	1.621	11.72	0.658	12.5(i) ∗	0.28	United Solar
a-Si/a-SiGe dual junction				11.2(st)	0.25	United Solar
a-Si/a-Si/a-SiGe triple junction	2.289	7.9	0.685	12.4(i) ∗	1	Sharp
a-Si/a-SiGe/a-SiGe triple junction	(see P. D. Maycock, [55])			14.5(i)		United Solar
a-Si/a-SiGe/a-SiGe triple junction	(see P. D. Maycock, [55])			12.8(st)		United Solar

∗ Confirmed by standard measurement laboratory.

mass production are a strong motivation for developing viable thin-film technologies for silicon solar cells. Nowadays, the most advanced thin-film silicon technology is that of hydrogenated amorphous Si as described above. Due to its optical properties a-Si represents a "true" thin-film solar cell material.

Further morphological modifications of silicon are also considered for thin-film solar cells, e.g., polycrystalline Si and microcrystalline Si (μc-Si) with grain sizes in the range of 1 µm and 1 to 50 nm, respectively. The decisive advantage of these modifications compared to amorphous Si is the expectation that there will be no stability problems in solar cells made from these materials. The main idea behind is to combine the advantages of crystalline Si – availability and technological know-how of the environmentally harmless material – with the economical advantages of thin-film technologies.

The design of thin-film solar cells with these materials strongly depends on their absorptivity. The absorption coefficients as a function of the quantum energy of incident light are shown in Figure 17. Polycrystalline and μc-Si have absorption curves being in between those of monocrystalline Si and of a-Si [56].

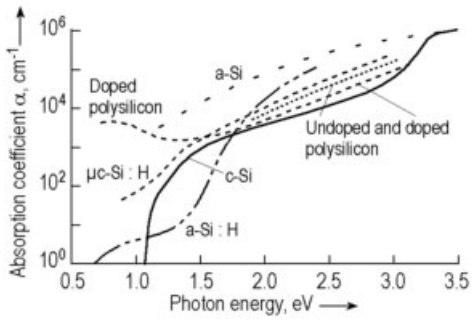

Figure 17. Absorption coefficients for different silicon-based photovoltaic materials (after [57])

Microcrystalline Silicon Thin-Film Solar Cells. Microcrystalline silicon (μc-Si, a material consisting of very small crystalline grains embedded in an amorphous Si matrix) has a relatively high absorption; layer thicknesses of 2 to 4 µm would be fully sufficient for solar cells. μc-Si has been successfully implemented in amorphous Si cells; major progress has been made using doped μc-Si for the *p*- and *n*-layers. The built-in voltage is enhanced (i.e., V_{oc} increases), the conductivity is increased, and the *n*-*p* contact

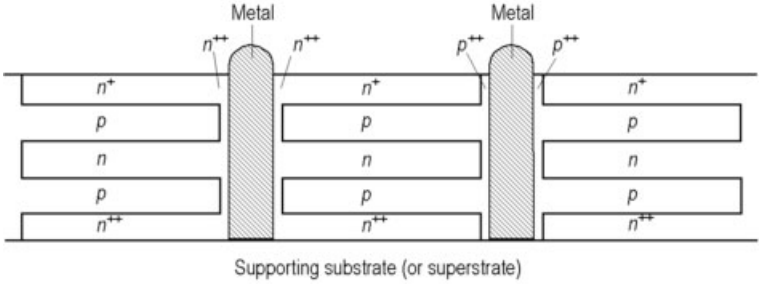

Figure 18. Cross-section of a multilayer, parallel multijunction buried contact solar cell (after [68])

in the center of a stacked multijunction structure act as a more effective tunnel junction if it is rendered microcrystalline.

A 1.7 µm thick entirely microcrystalline *p-i-n* cell deposited by VHF plasma deposition at moderate substrate temperatures (170 to 220 °C) has reached an efficiency of 4.6 % [19]. The structure resembles that of an amorphous Si cell, however, the short-circuit current is increased because of enhanced infrared absorption. Apparently, the physical properties of μc-Si depend on the volume fraction of the crystalline portion and are anywhere in between that of crystalline and that of amorphous Si. The energy gap can be as low as that of c-Si (1.1 eV), and the absorptivity approaches that of a direct semiconductor such as a-Si. The crystalline volume fraction can be controlled within a certain range by the choice of the deposition parameters. A very encouraging feature of entirely microcrystalline Si cells is that they so far do not show any degradation. With these findings, μc-Si solar cells show promising perspectives, however, the deposition rates of good-quality films are only in the range of 0.1 nm/s. They need to be increased at least by a factor of 10 in order to obtain solar cell processing times of less than one hour.

Polycrystalline Silicon Thin-Film Solar Cells. At present (in the 1990s), widespread efforts are being undertaken to realize polycrystalline Si thin-film solar cells. For summaries see [12], [56], [58]. In the case of crystalline Si thin films, the term "thin film" may be extended to film thicknesses of 20 µm or even more. Essentially it is the deposition technique that defines "thin films," for example vacuum evaporation, sputtering, chemical vapor deposition (CVD), or chemical bath deposition (CBD). Even "thick-film" techniques such as screen printing, spray-on, and spin-on may be considered. The film thickness as such is not the decisive parameter; finally, in PV it is always the price per unit of output power that counts.

The following aspects need to be considered when talking about crystalline thin-film Si solar cells:

1) Thin-film c-Si is polycrystalline with grain sizes depending on deposition technique, substrate type and temperature, and postdeposition treatments.
2) Grain boundaries cause substantial recombination losses when the flow of light-generated minority carriers has to cross them. Consequently, the grain size should be at least in the order of the film thickness, so that minority carriers can reach the *p-n* junction without crossing a grain boundary.
3) The light path in the film must be sufficiently long in order to absorb enough light to generate a short-circuit current of reasonable size (ca. 30 mA/cm^2). In case of a planar configuration the film thickness must be in the range of the absorption depth $1/\alpha$ with α being the absorption coefficient. For the major part of the solar spectrum $1/\alpha$ is in the range of 20 µm.
4) Films with a thickness of 20 µm should be deposited with high deposition rate in order to keep production time short. Furthermore, the grain size in films with planar structure should be of the order of 20 µm or more.
5) With low-cost substrates such as ordinary glass and low-cost deposition techniques such as vacuum deposition these requirements could not be fulfilled up to now [60].
6) Films of reasonable quality on amorphous substrates require substrate temperatures in

the range of 1000 °C during deposition or postdeposition annealing temperatures in the same range.

Many laboratories have published results that demonstrate the feasibility of high efficiencies even with thin-film c-Si structures. By using c-Si wafers as substrates [61–63] or by thinning c-Si wafers [64], efficiencies of more than 20 % have been achieved. Not clearly specified thin-film technologies based on "inexpensive Si feedstocks" (most probably waste of the semiconductor industry) with steel as substrates resulted in efficiencies of 9.5 % (AstroPower, [65]). With a different technology AstroPower even achieved 16.6 % [66]. However, the results of the first group can only be judged as a demonstration of the potential of thin-film c-Si solar cells. They do not have any relevance for low-cost industrial production of PV cells or modules. Furthermore, the technologies developed by AstroPower may allow a certain volume of PV business at present, but must be critically reviewed, if they depend on waste of the semiconductor industry. This previously discounted source is becoming more and more scarce. Firstly, waste is used or recycled for electronic applications, secondly, the annual increase of PV production volume results in a scarcity as such.

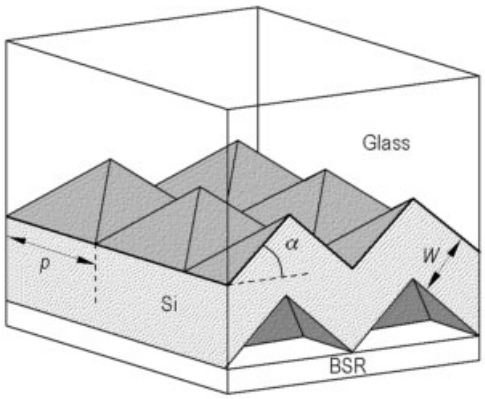

Figure 19. Light trapping structure with pyramidally structured glass (pyramidal film texture; after [69])
W = c-Si film (thickness = 4 µm); p = texture period = 15 µm, α = facet angle of the pyramids = 75°

Nevertheless, there are some promising approaches of producing low-cost thin-film c-Si solar cell, provided they can succeed in realizing medium or high efficiencies. Especially the texturing of the films for efficient light trapping is believed to be the key to successful developments:

- Researchers at SANYO [67] deposited an a-Si film on a textured metal substrate and crystallized it by annealing for up to 10 h at 550 – 650 °C. With a 10 µm thick polycrystalline cell a conversion efficiency of 8.5 % was obtained.
- At the University of New South Wales (UNSW) a different concept of poly-c-Si thin-film solar cells is investigated: since the problem of crystalline thin-film Si cells is the incomplete absorption of incident solar radiation, the idea was to use a stack of interlocking thin films of alternating conductivity as shown in Figure 18. The contacts to the films are metallic lines deposited into laser-eroded grooves. The left line in Figure 18 contacts all *n*-type layers, the right one all *p*-type layers. In total the arrangement between the metallic lines represents an interlocked *p-n* junction. This layered structure still provides the special advantage of thin film solar cells: the photogenerated carriers have to cover only a short distance to the next *p-n* junction, i.e., the probability of recombination on the way to the junction is low. This allows the use of thin films consisting of low-quality material like vapor-deposited polycristalline Si layers. A total thickness of 10 µm provides sufficient light absorption.
- Researchers at the Max-Planck-Institute at Stuttgart, Germany, are investigating the realization of thin-film poly-c-Si solar cells on pyramidally structured glass (see, e.g., [69]). Their light trapping studies demonstrate that Si films of only 4 µm thickness suffice to generate a maximum short-circuit current density of 35 mA/cm^2 with the "pyramidal film" texture as shown in Figure 19 having a texture period of 15 µm, a c-Si film of 4 µm thickness can generate short circuit current densities as high as 37 mA/cm^2.

4.4. Summary of the Status of Silicon Solar Cells

Figure 20 shows typical representatives of the various kinds of industrially produced silicon so-

lar cells. The standard size is still $10 \times 10\,\text{cm}^2$ for single- and multicrystalline Si cells, but larger cells are coming up for certain applications. The a-Si "cell" shown is of course a small module with a couple of cells interconnected in series.

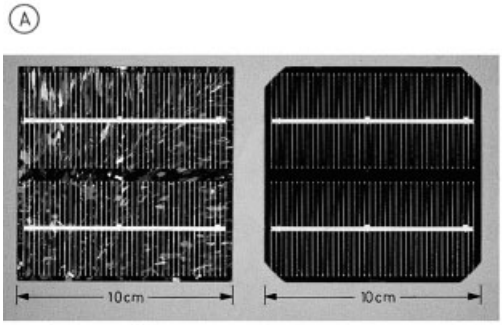

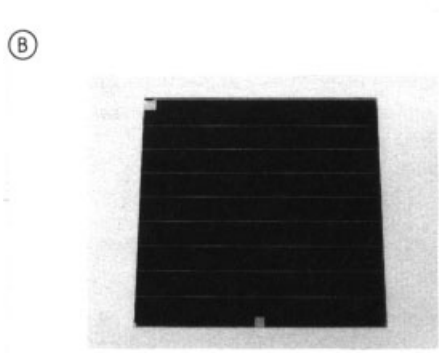

Figure 20. Photographs of various kinds of solar cells A) Multi-c-Si (left) and single-c-Si (right) solar cell (DASA); B) a-Si solar cell (SANYO)

The c-Si solar cell is still the "working horse" in the worldwide PV scene. Its market share is about 80 % of the worldwide PV production. The rest is covered by amorphous Si and by a small quantity (less than 1 %) of CdTe-based cells. Si cells have also been referred to as "the ultimate photovoltaic solution" [70]. The availability, the technological know how, and the environmental harmlessness of the material might justify such a statement.

At present, the annual consumption of Si in the PV industry is about 1000 t. This is roughly 10 % of the total Si production for the electronic industry. So the major source of PV-feedstock Si is the material "waste" (off-specification Si, remelts, etc.). An increase in annual production of photovoltaic cells, therefore, will not necessarily result in cost decrease. The development of solar-grade Si might be unavoidable, which may go along even with increasing cost of the production technology. Consequently, thin-film technologies may be of higher topicality than presently accepted.

After an impressive start of their carrier, amorphous Si solar cells are still suffering from problems associated with limited efficiencies and with instability problems. However, in recent years higher and higher values of increased "stabilized" or "end-of-life" efficiencies have been reported, and a-Si cells and modules may now slowly enter the "power market" after having been restricted to the consumer electronics sector in the past.

4.5. Compound Semiconductor Solar Cells

Before discussing the most relevant compound semiconductor solar cells, namely the thin-film cells based on chalcogenides and chalcopyrites, the III – V (i.e., elements from group 13 and group 15 of the Periodic Table) semiconducting compounds will be treated shortly.

III – V Compound Solar Cells. Gallium arsenide (GaAs) is the most commonly used material in III – V compound solar cells. Gallium arsenide has excellent characteristics: its band gap of 1.43 eV is suitable for effectively converting solar light into electric energy. Since the band gap of GaAs is wider than that of silicon, GaAs solar cells can operate at relative high temperatures. Therefore, GaAs solar cells are suitable for use as concentrator solar cells, i.e., configurations that use lenses or mirrors to focus sunlight onto very high efficiency cells. They are also suitable for use in space because they are highly resistant to cosmic rays.

Fabrication. The surface recombination velocity (i.e., the rate of carrier recombination near the surface) of gallium arsenide solar cells is higher than that of silicon solar cells (that of GaAs is $10^6 - 10^7$ cm/s, that of Si is $\approx 10^4$ cm/s). That is why the photoresponse of *p*-GaAs/*n*-GaAs homojunction solar cells in the short wavelength range is low. In order to solve this problem, *p*-$Ga_xAl_{1-x}As$ $(0 \leq x \leq 1)$/*p*-GaAs/*n*-GaAs heterojunction solar cells have been devel-

oped. The manufacturing process and cell structure of this type of solar cells is shown in Figure 21.

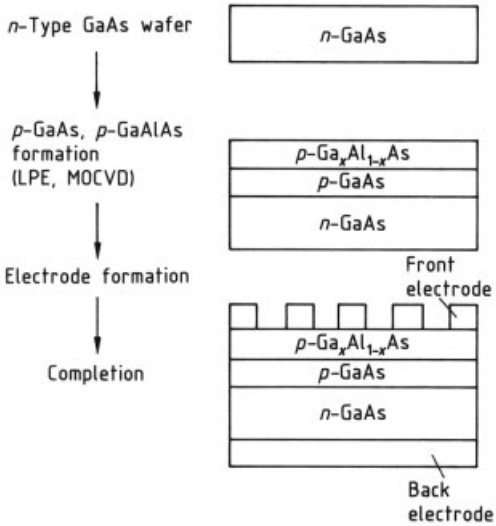

Figure 21. Manufacturing process and cell structure of GaAs solar cells
LPE=liquid-phase epitaxy; MOCVD=metal organic chemical vapor deposition

In p-$Ga_xAl_{1-x}As$ $(0 \leq x \leq 1)$/p-GaAs/n-GaAs heterojunction solar cells, the band gap of the p-GaAlAs commonly used is about 2.1 eV. The GaAs band gap varies with aluminum content; if excess Al is present the band gap becomes wider, deficiency in Al narrows the band gap. Photons with energies < 2.1 eV are not absorbed in the p-$Ga_xAl_{1-x}As$ $(0 \leq x \leq 1)$ layer but in the n-GaAs layer. Therefore, photovoltaic performance is not adversely affected by the p-$Ga_xAl_{1-x}As$ $(0 \leq x \leq 1)$ surface. Conversion efficiencies of more than 25 % have been achieved by using p-$Ga_xAl_{1-x}As$ $(0 \leq x \leq 1)$/p-GaAs/n-GaAs heterojunctions (see Table 3) .

The remarkable know how available in the production technology of III – V compounds allows the epitaxial growth of multilayer structures with varying composition. With the experience in producing semiconductor laser structures and integrated optoelectronic devices it was obvious also to fabricate multilayer PV devices like tandem structures and even multiple quantum well solar cells. Efficient tandem configurations are composed, e.g., of GaAs and GaInP or GaAs and GaSb (see Table 3).

In principle, these materials – GaAs, InP, and also ternary and even multinary compounds – show excellent photovoltaic properties, however, material and processing cost are high; III – V compound solar cells can be taken into account only for space applications and, possibly, for terrestrial concentrating PV systems.

CdTe-based Thin Film Solar Cells. The development of CdTe-based solar cell has created a large variety of cell structures (altogether more than ten different types; see, e.g., [24], [71]). The most successful ones have been CdTe-cells based on heterojunction structures. With n-CdS/p-CdTe heterojunctions efficiencies of about 10 to 12 % could be achieved. However, since CdS with its band gap of 2.4 eV is not an ideal window in combination with CdTe ($E_g = 1.5$ eV), the spectral sensitivity is limited to the range $1.5\,\text{eV} < h\nu < 2.4\,\text{eV}$. A further push of efficiencies towards values around 16 % has been initiated by applying real window – absorber structures such as n-SnO_2/n-CdS/p-CdTe or n-ITO/n-CdS/p-CdTe (ITO = Indium – Tin – Oxide) with very thin CdS-layers acting solely as buffer layers. This concept is shown in Figure 22, and the resulting extension of the spectral response to quantum energies beyond 2.4 eV is indicated in Figure 23. It is remarkable that the same window configuration – transparent conducting oxide (TCO) + CdS – is also appropriate for $CuInSe_2$ cells (see below).

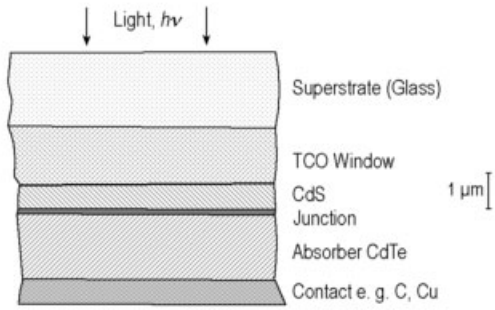

Figure 22. Structure of a thin-film TCO/CdS/CdTe solar cell

Table 3. Performance of III – V compound solar cells and PV tandem structures (after [12])

Type of cell	V_{oc}, V	J_{sc}, mA/cm$_2$	Fill factor	Effic., %	Area, cm$_2$	Group/ organization
One sun cells						
GaAlAs/GaAs	1.022	28.17	0.871	25.1	3.91	Kopin
GaAs (large area)	1.035	26.9	0.854	24.2	16.14	Spire
InP	0.878	29.29	0.854	21.9	4.02	Spire
GaInP/GaAs 2-terminal tandem	2.385	13.99	0.885	29.5	0.25	NREL
GaInP/GaAs 2-terminal tandem	2.488	14.22	0.856	30.3	4.00	Japan Energy
Concentrator cells						
GaAs/Si 4-terminal stack, 350 suns				29.6	0.317	Varian, Stanford, Sandia
GaAs/GaSb 4-terminal stack, 100 suns				32.6	0.053	Boeing

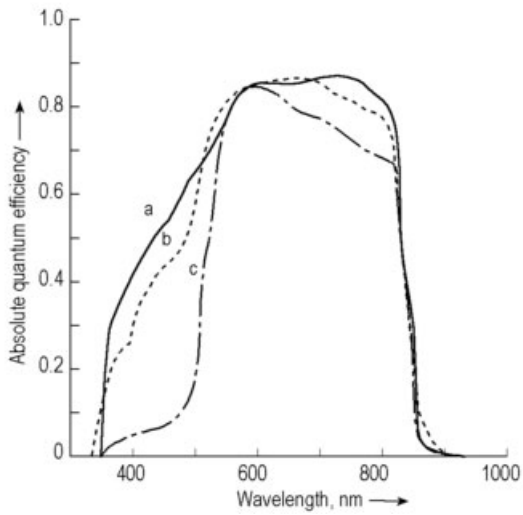

Figure 23. Spectral response of various CdTe solar cells (after [72])
a) CdS/CdTe cell made by Solar Cells, Inc. (SCI): $\eta = 8.5\,\%$; b) ANTEC-cell (TCO/CdS/CdTe): $\eta = 10.5\,\%$; c) TCO/CdS/CdTe-cell made by University of South Florida (United States): $\eta = 14.2\,\%$

For the deposition of the CdTe layer a large number of methods, resulting in high efficiency cells on the one hand and for economic production on the other hand, can be applied. In addition to close-spaced sublimation (CSS) and atomic layer epitaxy (ALE), also physical vapor deposition (PVD), chemical deposition methods such as spraying, electrodeposition and metallorganic chemical vapor deposition (MOCVD), and sintering of screen-printed films have been applied successfully. In most cases CdS is deposited in a chemical bath. Recent efforts for the deposition of CdS include RF sputtering and close-spaced sublimation. In the early 1980s MATSUSHITA commercialized the screen-printing technology [73], [74]; the process sequence is shown in Figure 24.

The following points are critical for the performance of CdTe solar cells:

1) It turned out to be difficult to find suitable high-quality ohmic contacts for *p*-type CdTe. Successful approaches include (1) the deposition of a Cu-doped ZnTe layer followed by the deposition of an appropriate metal such as Ni; (2) the use of a mixture of graphite, $CuCl_2$, and HgTe which is mechanically applied to the CdTe surface, and (3) vacuum evaporation of an Au contact on a chemically treated CdTe surface which modifies the stoichiometry leaving back a Te-rich surface (see, e.g., [75], [76]).
2) Postdeposition treatments must be carried out and optimized. They influence grain growth, grain orientation, and they cause interdiffusion of CdTe and CdS leading to the formation of Cd(S,Te) at the interface. $CdCl_2$-treatments and subsequent air annealing at 400 – 500 °C promote the grain growth of as-deposited thin films. The formation of mixed alloy Cd(S,Te) at the interface is believed to shift the actual *p-n* junction into the CdTe layer and to passivate the metallurgical heterointerface (see, e.g., [77], [78]).

Since a couple of years, the highest conversion efficiencies confirmed by NREL have been

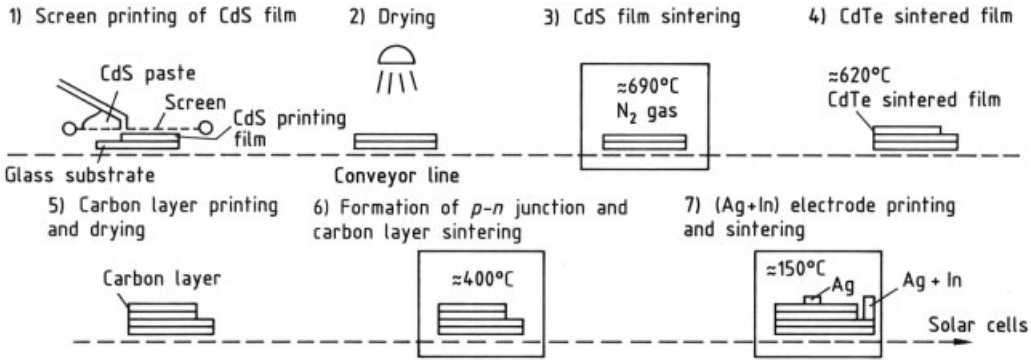

Figure 24. Manufacturing process for CdS – CdTe solar cells by the screen-printing method

those reported by the research group at the University of South Florida ($\eta = 15.8\,\%$, cell area = 1.05 cm^2; see [79]). In 1996, Golden Photon, Inc., Golden, Colorado, fabricated a 0.3205 cm^2 CdTe thin-film cell with an efficiency of 16.1 % [80]. This value has also been confirmed by NREL. Values very close to 16 % have been reported also by Matsushita Battery (15.12 %; see [81]) and by Japan Energy Corp. (16.0 %; see [82]).

At present, commercial production of CdTe solar cells with a capacity of about 1 MW/a is carried out by Matsushita Battery. CdTe-module production has also been started at Golden Photon (GPI), Solar Cells Inc. (SCI) and BP Solar. (Only recently, it has been announced that GPI closed down the module production again! [83]). Module efficiencies obtained by different commercial producers are listed in Table 4. The values presented in Table 4 indicate the maturity of CdTe-module technology and its high potential for large-scale production.

Table 4. Performance of CdTe PV modules

Group/company	Area, cm^2	Power, W	η, %
Solar Cells Inc.	6728	61	9.1∗
Golden Photon	3528	27	7.7∗
Matsushita Battery	1200	10	8.7
BP Solar	4500	37	8.4
	706	7	10.1

Values marked by ∗ have been confirmed by NREL.

Estimations within the European Union revealed that the prospective costs for CdTe module production depend on the assumed annual production volume and will be in the range of 1.0 – 1.5 ECU/W$_p$ (1 ECU = 1 European Currency Unit = \$ 1.1) for the year 2000 at an estimated annual production rate of 10 MW$_p$. A U.S. study even ends up at a production cost of only 0.28 ECU/W$_p$, provided one particular factory could be operated with a production volume of 500 W$_p$ (see, e.g., [84]).

The energy pay back time for frameless CdTe modules at an energy input of about 30 kWh/m^2, which is mainly consumed in the production of the (single) glass sheet, amounts to 1.6 months. CdTe – CdS modules operated under both external and internal accelerated tests showed stable efficiency.

Thin-Film Solar Cells Based on CuInSe$_2$ and Related Compounds. The development of CuInSe$_2$ ("CIS") solar cells dates back to the mid-1970s, to the work of S. Wagner et al. [10], who reported on production of CuInSe$_2$-based single crystalline solar cells with an efficiency of 12 %. It was mainly the high absorption coefficient of this direct-band gap material that attracted the researchers' attention: thin film devices with thicknesses $< 1\,\mu$m appeared to be feasible. The handicap of the low band gap ($E_g = 1.04$ eV) could be bypassed in the following years by applying alloys with related compounds such as CuGaSe$_2$ ($E_g = 1.7$ eV) and CuInS$_2$ ($E_g = 1.55$ eV). The possibility of designing the optical properties plus the special feature of these chalcopyrites to show a type of self-passivation of grain boundaries and surfaces led to efficiencies of more than 17 % in the mid-1990s [85], [86].

The basic structure of a Cu(In,Ga)(Se,S)$_2$ ("CIGS") PV cell is shown in Figure 25. It re-

presents a modified window – absorber heterojunction with the following features:

A molybdenum back contact is deposited by sputtering or electron-beam evaporation. Mo withstands the subsequent processing conditions of the chalcopyrite film.
The CIGS absorber film is deposited with a typical thickness of 2 µm.
Since chalcopyrites commonly are not amphoteric – they exhibit only *p*-type conduction, see below – heterojunctions must be applied.
The absorber film is coated with a thin buffer layer (CdS or Cd free materials) in a chemical bath. This improves the properties of the heterojunction.
The window layer consists of ZnO which is fabricated by RF sputtering or by a CVD process.

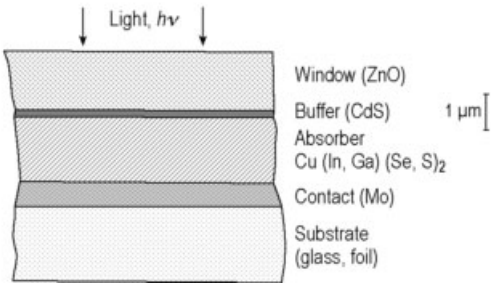

Figure 25. Structure of a thin-film CIGS solar cell

An idea of the complexity of this multinary material system can be obtained, e.g., from the ternary-phase field of the system Cu – In – Se, which is shown in Figure 26.

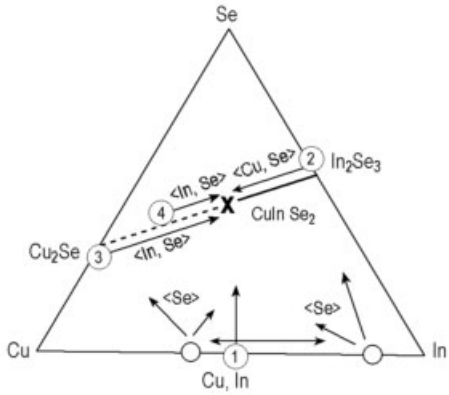

Figure 26. Ternary-phase field for Cu – In – Se (after CATALANO, [58])

A movement along the tie line connecting Cu_2Se and In_2Se_3 does not involve deviations from the normal valence of the intermediate compounds, i.e., the deviation ΔZ from the ideal "valence stoichiometry" is zero along the tie line. ΔZ is given by:

$$\Delta Z = \{2\,[\mathrm{Se}] / [(\mathrm{Cu}) + 3\,(\mathrm{In})]\} - 1$$

On the line $CuInSe_2$ and the so-called ordered vacancy compounds (OVCs) $CuIn_3Se_5$ and $CuIn_5Se_8$ are located [87], [88]. Deviations normal to the tie line, however, result in an excess or a deficiency in the number of bonding electrons. In principle, the tie line represents the pseudobinary $Cu_2Se - In_2Se_3$ phase diagram of $CuInSe_2$. Phase diagrams as elaborated by various research groups have been collected in [89]. They indicate the areas of existence of $CuInSe_2$ and of the ordered vacancy compounds. The improved knowledge of the phase diagram and the large amount of experimental work resulted in a number of findings that are important with respect to the production of efficient thin film solar cells [84], [89]:

1) The conductivity of $CuInSe_2$ ranges from low-resistivity *p*-type to high-resistivity intrinsic or *n*-type, depending on deviations from ideal stoichiometry ("intrinsic" inversion).
2) Cu-rich films – films with more than 25 at % Cu – are of *p*-type ("intrinsic doping").
3) OVCs occur on the indium-rich side of the pseudobinary phase diagram. Their conductivity appears to be intrinsic or slightly *n*-type. The surface of In-rich $CuInSe_2$ films stabilizes at a composition of $CuIn_3Se_5$, in the case of Ga compounds this composition tends to be $CuGa_5Se_8$.
4) In gallium-containing chalcopyrites the possibility of inversion of conductivity – from *p*- to *n*-type – decreases with increasing Ga content and vanishes for nearly pure $CuGaSe_2$.
5) The solubility of excess Cu in $CuInSe_2$ is very small, a larger excess Cu results in segregation of Cu_2Se.

Several processes for CIGS deposition led to devices with efficiencies exceeding 16 %:

Coevaporation/vapor deposition of the elements onto a heated substrate [85], [90]

Selenization of sputtered or evaporated Cu – In precursors in H_2Se or elemental Se atmosphere [91], [92]
Diffusion of Cu and Se into $(In,Ga)_2Se_3$ precursors [93], [94]

Table 5 summarizes the highest efficiencies of CIGS cells reported in recent years.

The success of the chalcopyrite thin-film solar cell is mainly due to:

1) The vapor – liquid – solid (VLS) growth of the chalcopyrite layer caused by the segregation of a liquid Cu_xSe phase at Cu excess. The melting point of CuSe is 513 °C. In thin films this segregation occurs mainly on the surface and in grain boundaries. A model for the growth of the chalcopyrite layer under these conditions is shown in Figure 27. Due to the very high mobility of metals in the binary chalcogenides, their transport and exchange within the films is assisted by this segregated phase and large polyhedral chalcopyrite grains are formed. This mechanism offers the possibility to achieve stoichiometric material by growing the films under an excess of copper and subsequent removal of the secondary phase, e.g., by a chemical treatment (with KCN). This implies that very accurate control of the evaporation rates is not required because the stoichiometry of the segregated solid chalcopyrite is determined by the field of existence according to the phase diagram and can be additionally adjusted by a subsequent chemical treatment. By application of this method, it was possible to grow absorber films for highest-efficiency devices [99].
2) The formation of an "intrinsic *p-n* junction" that is obtainable by adjusting the vapor beam to an indium-rich beam at the end of the coevaporation process. The unwanted Cu_xSe is thus converted into useful $CuInSe_2$ that grows epitaxially on the existing nuclei. Once Cu_xSe is consumed, the Cu-deficient vapor arriving at the surface may convert the surface to the OVC $CuIn_3Se_5$. Since the OVC shows *n*-type conductivity, a quasi *p-n* homojunction in the interior of the grains is formed [100]. This reduces interface states and interface recombination losses. To the inverted surface the heterojunction partners forming the buffer and the window layer only have to provide an electronic adaption. The remaining task of the buffer layer is to avoid a potential barrier in the conduction band that would hamper the flow of photocarriers across the junction.
3) The *self-passivation* of defects, grain boundaries and surfaces, that is typical for chalcopyrites. The diffusion length, the light absorption depth, the grain size, and the film thickness – all these parameters are roughly of the same order of magnitude. If all the imperfections mentioned above would actively behave as recombination centers, then the obtainable efficiencies of thin film solar cells based on these materials would be considerably lower than the values that have been already obtained.

Research and development activities are focused on deposition and crystallite growth processes, surface properties, the impact of impurity diffusion, novel buffer layers to replace CdS, and in particular on the establishment of the production of larger-area modules. Some production lines are under construction, however, commercially produced modules are not yet available. A sequence of fabrication steps for the manufacturing of CIS modules is illustrated in Figure 28. The deposition of the absorber is based on vacuum evaporation using an in-line evaporation process with line sources for the elements. The subsequent steps in principle resemble those of a-Si module production (see Fig. 16).

The present status of module manufacturing is summarized in Table 6.

Concerning production cost and cost development, the perspectives for $CuInSe_2$ have been considered to be very similar to those of CdTe (see above): at an annual production rate of 10 MW_p, module production cost might be in the range 1.0 – 1.5 ECU/W_p in the year 2000. In case of an annual production of 500 MW_p in a single factory, according to a U.S. study, the cost could also be as low as 0.28 ECU/W_p. The APAS Music FM study of the European Commission [109] also included cost studies: on the basis of a manufacturing facility producing 60 MW_p/a the cost of all thin-film solar cell technologies (a-Si, CdTe, $CuInSe_2$) are practically the same and are in the range 0.58 – 0.68 ECU/W_p.

Table 5. Efficiencies of CIGS laboratory cells

Material	Area, cm^2	V_{oc}, mV	η; %	Laboratory
$Cu(In,Ga)Se_2$	0.414	674	17.7	NREL [95]
$Cu(In,Ga)Se_2$	0.38	647	17.6	Uppsala Univ./ IPE [86]
$Cu(In,Ga)Se_2$	0.48	655	17.7	Matsushita [96]
$Cu(In,Ga)Se_2$ InOHS/ZnO	0.38	594	15.4	Uppsala Univ./ IPE
$CuGaSe_2$	0.38	850	8.9	IPE [97]
$CuInS_2$	0.38	735	12	IPE [98]

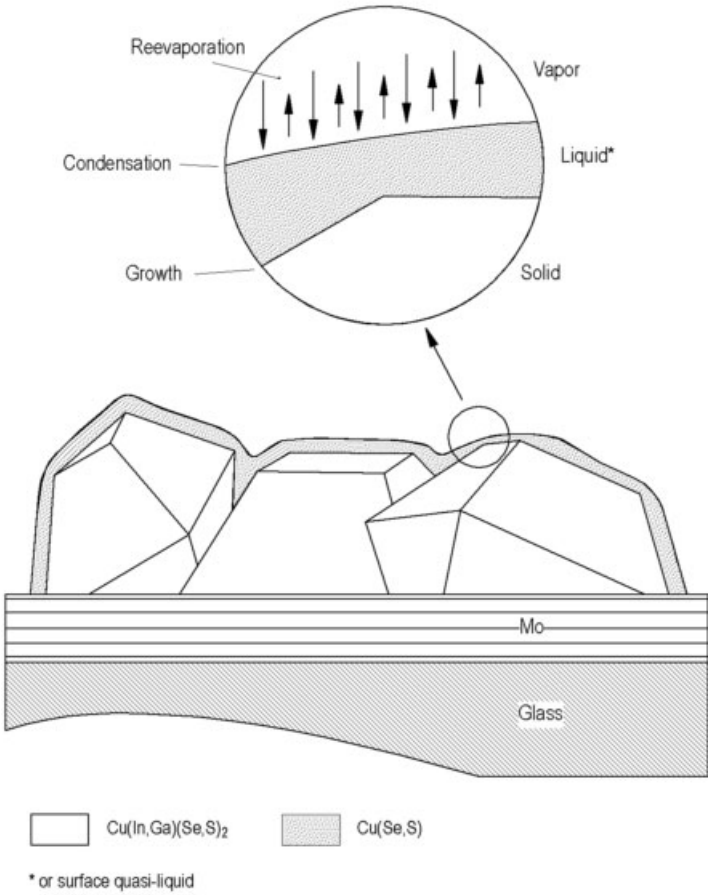

Figure 27. Model for the growth of polycrystalline $Cu(In,Ga)(S,Se)_2$ films by vacuum evaporation (after SCHOCK [89])

Table 6. Status of CIGS module development

Material	Area, cm^2	V_{oc}, V	η %	Laboratory/process
$CuIn(Ga)Se_2$	3830		11.2	Siemens Solar Ind./selenization [102]
$Cu(In,Ga)Se_2$	717	22.7	10	ZSW/evaporation [103]
$CuIn(Ga)(S,Se)_2$	50.2	6.7	14.1	Showa Shell (Japan)/selenization [104]
$Cu(In,Ga)Se_2$	90.6	9.3	13.9	IPE, Univ. Stuttgart; ZSW/evaporation [105]

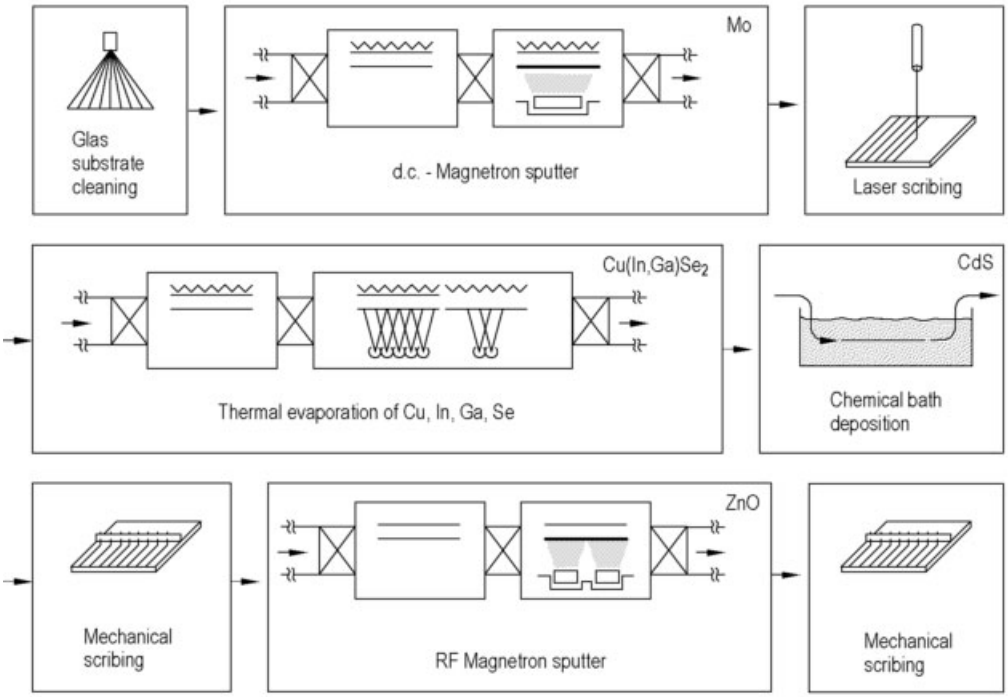

Figure 28. Fabrication steps for module integration of thin-film $Cu(In,Ga)(S,Se)_2$-based solar cells (ZSW, Stuttgart, Germany; see [101])

Environmental Aspects of Compound Semiconductor Solar Cells. Environmental, safety and health risks associated with the production and commercial use of CdTe and CIS modules have been studied in detail by Moskowitz [106] and Alsema [107]. Environmental aspects of photovoltaic module manufacturing, especially in view of CdTe and CdS, demand a minimum of material use, high yields and safe handling, as well as recycling of wastes. Apart from accidental fires there is no hazard in operation, however, concepts and measures have to be taken for disposal and recycling of modules at the end of their lifetime. Some health hazards exist with Cd and Te compounds which lead to concentration limits in air and water. CdTe – CdS modules have failed the toxicity characteristic leaching procedure CTCLP-test slightly. It is believed, however, that reduced material quantities in the deposition processes and careful recycling could overcome this problem.

Environmental, safety, and health issues also exist for manufacturing and commercial use of CIS. Concerns which are related to the use of CdS have been overcome by the development of Cd free buffer layers leading to a comparable device performance as described above. The toxicity of H_2Se during manufacturing demands special precautions. Strategies for recycling the modules are still in a designing phase.

4.6. Other Types of Photovoltaic Cells

Besides the "classical" purely solid-state semiconductor based solar cells using the internal photoeffect directly, some exotic devices are worth mentioning. They include photovoltaic devices based on organic semiconductors and electrochemical photovoltaic cells (→ Solar Technology, Chap. 5.). Furthermore, porous silicon and fullerenes have been investigated with respect to their photovoltaic properties, and also the thermophotovoltaic effect is considered for small-scale energy conversion (see, e.g., [12]). Spectacular promises have been announced for the electrochemical photovoltaic cell – termed

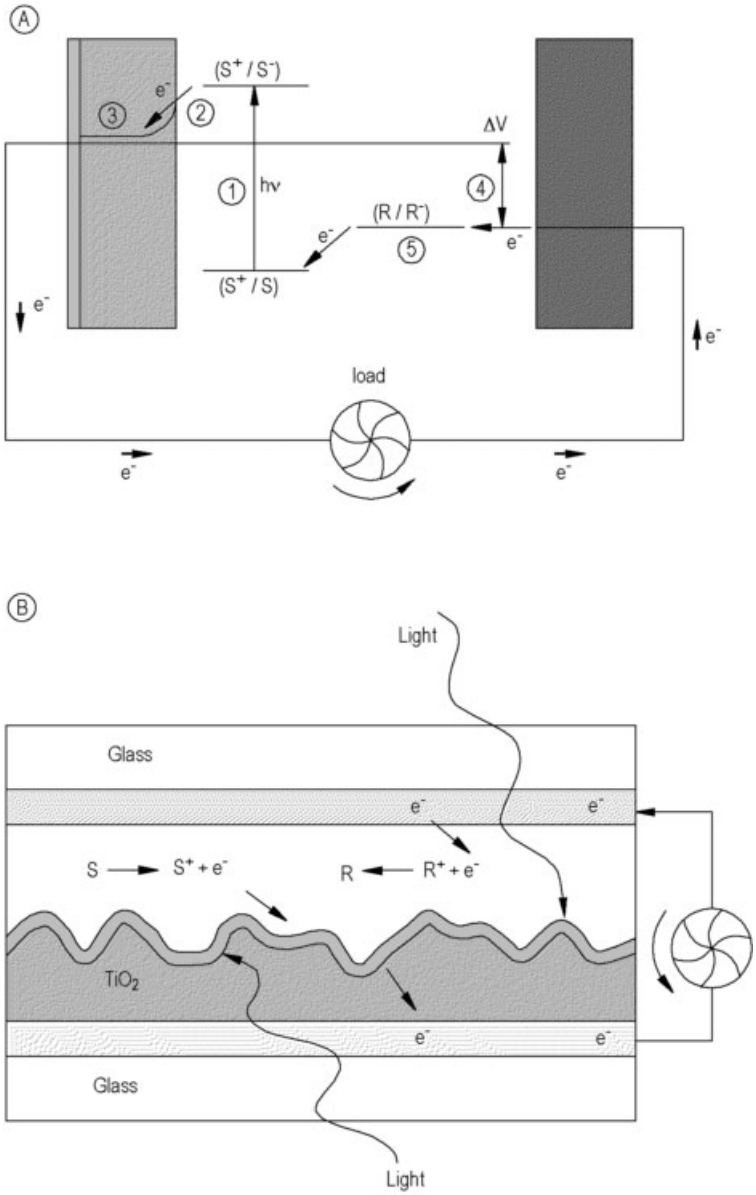

Figure 29. Schematic structure (A) and mechanism (B) of the nanocrystalline electrochemical photovoltaic cell (after M. Grätzel and A. J. McEvoi, [108])

Grätzel cell after its discoverer [108] – and much money has been invested in recent years in the development of this type of photovoltaic converter in Europe. Therefore, the operation principle and some major aspects of this system will be discussed here

Nanocrystalline Electrochemical Photovoltaic Cell (Grätzel Cell). Like in a conventional solar cell, the junction between an illuminated semiconductor and an electrolyte generates a photovoltage. In the 1970s and 1980s, this concept received some attention, but it has been given up because of material instability prob-

lems caused by photocorrosion. As a general experience it could be taken from literature that only devices with large band gap semiconductors exhibited a certain stability, yet they delivered only small efficiency values.

The introduction of oxide semiconductors reduced the effect of photocorrosion, particularly for *n*-type semiconductors: here the photogenerated minority carriers (holes) act as oxidants at the interface. However, oxide semiconductors such as TiO_2 have band gaps in the range of 3 eV, i.e., they are practically insensitive to solar radiation. Therefore, as shown in Figure 29, a monolayer of an electroactive dye (the sensitizer S) is introduced to sensitize the oxide semiconductor. The dye has an optical absorption band that extends across the visible spectrum. After absorption of light quanta (process 1 in Figure 29 A) charge separation occurs by electron transfer from the photoexcited dye to the semiconductor (processes 2 and 3). The dye molecule is left as an oxidizing species. The ground state of the dye is subsequently restored by a charge transfer reaction with the redox electrolyte R (process 5). The process is completed by a reduction reaction of the counter electrode with the redox system (process 4).

Early experiments with such structures resulted in low performance compared to that of their solid-state counterparts. When practical devices were fabricated, it turned out that perfect or planar surfaces are not required. Porous and polycrystalline surfaces provide even better performance, their large effective surface area is advantageous for charge transfer. No minority-carrier recombination deteriorates this cell because these carriers are not present in the semiconductor. After injection into the semiconductor, the electrons are separated from the oxidized semiconductor dye by a potential barrier. The reverse process – electron loss from the substrate to the oxidized surface species – is inhibited by chemical mechanisms.

The suitability of porous and polycrystalline semiconductor (TiO_2) layers is the big technological advantage of the system. Very cheap techniques such as screen printing and sintering can be applied for layer deposition. On the other hand, the instability of the dye could be the greatest disadvantage. Dye development, therefore, is one of the main tasks in conjunction with this system. The original photochemical dye is the natural porphyrin chlorophyll, which can sensitize TiO_2. Ruthenium porphyrins are used as sensitizers and chemisorption of ruthenium trisbipyridyl dyes resulted in structures of good performance. Further matters of concern of the Grätzel cell are electrolyte optimization and encapsulation of the complete structure.

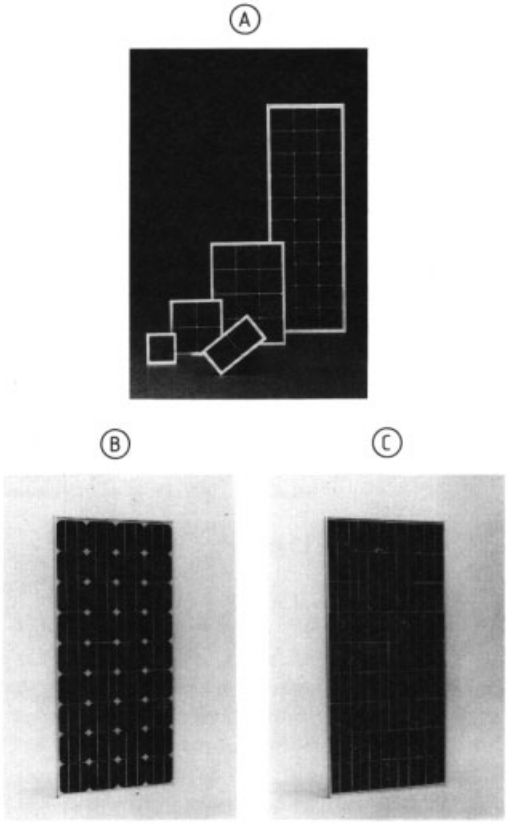

Figure 30. Photographs of large-area solar cell modules A) a-Si modules; B) c-Si modules; C) poly-Si modules

GRÄTZEL announced efficiencies of about 10 % [12]; at FhG-ISE Freiburg even 10.9 % has been confirmed for a small-area device [110]. Yet, efficiency measurements are more complicated than for conventional PV cells because of the time constant involved with the absorption and recombination process (see, e.g., [111]). Several research institutes and companies are currently working on the technological realization of these devices. However, new solutions have to be found for interconnecting the devices in monolithic modules. Furthermore, several is-

sues concerning encapsulation and stability have still to be addressed.

Figure 31. Crop drying tunnel with PV module for operating the ventilator; (Institut für Agrartechnik in den Tropen und Subtropen, Universität Hohenheim, Germany)

4.7. Photovoltaic Modules → Solar Technology, Chap. 4.

As described above, photovoltaic technologies have progressed remarkably. Large-area solar cell modules have been developed for power-generating systems. Figure 30 shows photographs of a-Si (Fig. 30 A), c-Si (Fig. 30 B), and poly-Si (Fig. 30 C) solar cell modules.

5. Uses

Commonly it is assumed that photovoltaic cells found their first application as power supply of space satellites. In reality, in the years following 1954 terrestrial applications were envisaged first, mainly in the consumer-electronics sector (e.g., PV-powered radios, see Chap. 2). However, PV cells were very expensive in those days. In 1958, with the launching of the first PV-powered satellite (Vanguard I) the career of photovoltaics as the standard energy source for satellites and space flight started. This monopoly still persists and provides good profits to the respective PV industry. In those early years, PV activities were practically limited to the United States, France, Germany, and to the former Soviet Union.

In the early 1970s – in fact even before the first oil crisis – the terrestrial application of photovoltaics was "reinvented" and several national programs were promoted, e.g., by the DOE in the United States, by the Commission of the European Communities (CEC) and by individual governments, especially in Europe. In 1974 also Japan joined the PV community with the creation of the Sunshine Project. For several years in the late 1980s, governmental PV funding in Germany was even higher than in the United States.

Figure 32. PV-powered buoy in the Baltic Sea near Stralsund

Space application of photovoltaics demands solar cells with high power: weight ratio, high reliability, and high resistivity to radiation. For terrestrial applications, however, the demand that had first priority was reduction of production costs. In this context, the development of thin-film solar cells became most important.

Terrestrial application of photovoltaics is usually subdivided into three main fields that are discussed in the following.

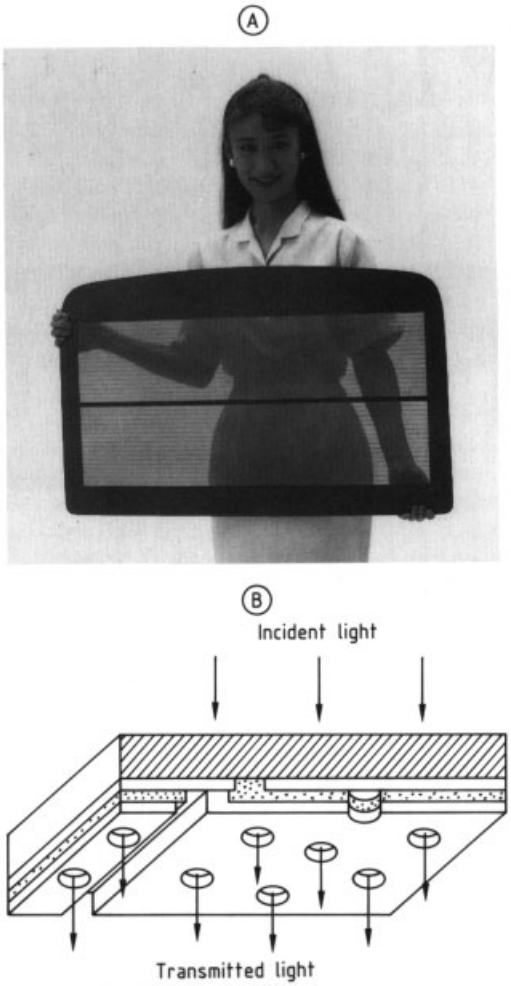

Figure 33. See-through a-Si photovoltaic cell A) for a car sunroof (Sanyo); B) Cell structure

5.1. Stand-Alone Systems

The "classical" application of PV are stand-alone systems, i.e., systems that provide electric power to facilities or applications in remote places, where no public grid is available. Prominent examples are listed in Table 7. In developing countries there is a tremendous need of such applications, but a rather huge market for stand-alone PV systems also exists in industrialized countries.

Figures 31 and 32 show examples of stand-alone PV systems.

Figure 33 A shows a photovoltaic cell for car sunroofs using see-through a-Si cell modules [12]. This resulted in a clear field of vision, self-ventilation and prevention of self-discharge of the car battery. The structure of this cell module is shown in Figure 33 B. Many microscopic holes are spaced uniformly on the active area of an integrated-type a-Si photovoltaic cell. It is also suitable for home windows. The power requirement of these stand-alone systems is of the order of tens of watts.

Stand-alone applications of photovoltaic generators are cost competitive or even much more economic than the conventional power supply methods and systems that had to be used before. Examples are repeated transport and recharging of batteries, use of diesel engines or – in case of larger systems – extension of the public electricity grid. In a number of cases PV first has enabled the application itself; e.g., the PV car sunroof (Fig. 33 A). PV systems can also help to create an infrastructure in places where there is none at all. The availability of electric power increases the value of land so much that it may compensate the cost of the PV system. In such a way building land in remote places can be developed.

Stand-alone systems commonly need a battery as a storage element. In particular cases, however, e.g., with PV-powered water pumps, the water itself can act as storage medium. Furthermore, the need for pumping and the solar irradiance are "in phase." Consequently, systems of that type can be designed without storage battery. The same holds for simple air conditioning systems, drying systems, etc.

5.2. Grid-Connected Power-Generating Systems (→ Solar Technology), Chap. 4.7.

Grid-connected PV systems are facing an economic situation completely different from that

Table 7. Main application of photovoltaics in stand-alone systems

Application scope	Application	Power range (peak Watt)
Telecommunications		10 – 7000
	– Telecommunication systems – Relay stations – TV repeaters – Transmitters – Educational TV – Mobile radio	
Water systems		400 – 6000
	– Surface water pumps / irrigation – Well pumps / irrigation and potabel water – Water treatment	
Village power supply		40 – 15000
	– Hut supply /solar home systems – Small working establishments – Streetlighting – Medical stations – Overall power supply	
Environmental systems		10 – 10000
	– Measurement stations – Cathodic corrosion protection of pipelines – Sea restoration – Sewage plants	
Traffic engineering		20 – 100000
	– Emergency calling points – Parking ticket automat – Mobile traffic lights – Buoys, warning beacons – Tunnel illumination	
Domestic supply		40 – 1000
	– Weekend houses – Detached houses – Mountain huts	
Leisure		1 – 200
	– Camping, caravans – Sailing – Car ventilation – Entertainment electronics	
Military		40 – 5000
	– Mobile power supply	

of stand-alone systems: the cost of PV energy is automatically compared to the cost of electrical energy as supplied by the grid. Up to now (in 1997), PV is nowhere in the world competitive with the public grid (see Chap. 6). That is why grid-connected PV systems always have been either experimental and demonstration projects sponsored by governments or utility companies, or they have been built in the frame of supporting programs for technology improvement and market introduction like the German “1000-PV-roofs” program and similar programs in other countries. The success of these programs, however, needed the enthusiasm of individuals, the

willingness to spend an extra money for renewable energies and for a clean environment, because in most cases the governmental funding does not cover the extra cost completely.

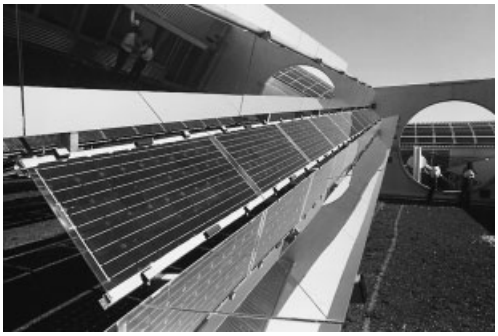

Figure 34. Part of the 435-kW grid-connected PV system on the new engine plant of Daimler – Benz in Stuttgart – Bad Cannstatt;
Center and left: mirror-concentrated ("V-trough") PV system with single-axis tracking;
Right, background: firmly installed PV on shed roof (photograph: Mercedes – Benz)

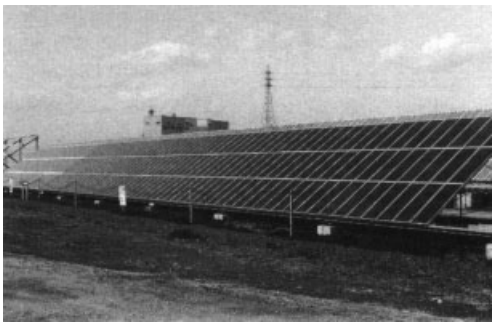

Figure 35. 1-MW power-generation system (Shikoku Electric Power Company)

Experimental large-scale PV systems utilizing nearly exclusively crystalline Si solar cells are being operated all over the world. Still the world's largest PV plant was the 7-MW plant at Carrisa Plains, California, dismantled in 1989. In Europe, a number of large-scale experimental PV systems have been built with the sponsorship of the Commission of the EC/EU, e.g., a 300-kW plant – meanwhile enlarged to 600 kW – for a recreation center on the island Pellworm in the North Sea. Substantial funding has been provided also by European utility companies, especially in Germany and Italy. The RWE (Rheinisch – Westfälisches Elektrizitätswerk) erected a 300-kW plant at Kobern – Gondorf and a 360-kW plant at the Lake of Neurath. In recent years the Italian ENEL built a PV plant of even 3.3 MW at Serre in Southern Italy. The largest project in Europe, a PV plant on the island of Crete initiated by the Government of Greece and by Greenpeace, is intended to finally have a power output of 50 MW. Only recently, Daimler – Benz put a 435-kW PV plant into operation. It has been installed on the roofs of the company's new engine plant in Stuttgart – Bad Cannstatt. Figure 34 shows a view of this plant.

Figure 36. Roof-mounted PV system as installed in the frame of the German 1000-PV-roofs program (photograph: RWE)

In Japan, a 1-MW power-generation system was constructed by the Shikoku Electric Power Company (see Fig. 35). The results are generating a great deal of interest because they indicate that photovoltaic cells are fully capable of system interconnection, i.e., the connection of solar-power-generation systems to the supply grids of electric power companies with electric wires. Hence, electricity can be transferred between solar-power-generation systems and electric power companies' supply grids.

If photovoltaics in industrialized countries should provide a considerable amount of electric power, then real large-scale installations and large areas are required. Therefore, decentralized installations on roofs and facades of buildings are envisaged. In principle, these areas are available without cost. The "1000-PV-roofs" program, that started in 1990 in Germany, and similar programs in other countries take advantage of this possibility. In Germany the total roof area of all buildings allow PV installations of about 50 GW [112]. Figure 36 shows a roof-

mounted PV system of the first generation. It is the typical standoff mount favored in those days. Future developments, however, should consider roof-integrated PV modules that substitute the conventional roofing material and thus follow the principle of multifunctional use. An example for this approach, realized in Japan with a-Si PV modules [113] is shown in Figure 37.

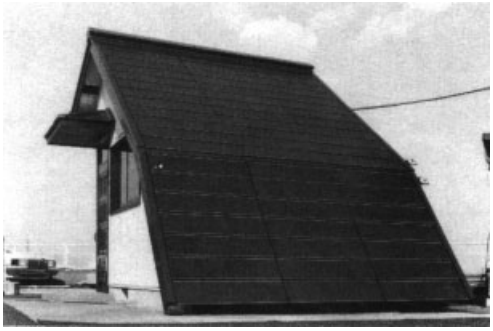

Figure 37. Model house with a roof comprised of shingle-type photovoltaic cells (SANYO)

A unique case is the incorporation of PV cells into buildings as photovoltaic facades. In contrast to PV roofs, where the substitution of PV modules for conventional roofing materials does not significantly reduce the cost, the substitution of conventional facade elements by PV modules can save a considerable amount of money, especially when luxurious facade materials such as polished granite or polished marble are considered. Such facade elements are often even more expensive than PV facades. As a consequence, in particular cases the electric power delivered by PV facades is free of cost! Of course, this does not hold for low-cost facades of industrial functional buildings. Figure 38 demonstrates the ornamental effect of PV modules integrated into a facade.

The possibilities of artistic designs of facades by integrating PV modules are large. Cell type, module shape and size, module background color around cells, semitransparent modules, modules with different types of frames (slight or sturdy; from aluminum, from stainless steel, or from plastic material) or even modules without any frame – there is an open field for creativity. It is even enlarged by the fact that cells and modules of different colors are available; only some changes in the design of the antireflection coating have to be made.

5.3. Consumer Products

Especially *a-Si thin-film modules* opened this market sector due to the flexibility in cell- and module design. The integrated manufacturing technology shown in Figure 16 allows the pro-

Figure 38. PV modules integrated into a facade in Gelsenkirchen (photograph: Pilkington Solar International)

duction of modules of the appropriate size, shape, and output voltage. In 1980, pocket calculators powered by a-Si photovoltaic cells appeared on the market for the first time, and since then a-Si photovoltaic cells have been used as power sources for various consumer products.

Some applications for consumer products are shown in Figure 39. After the commercialization of pocket calculators, watches, radios, and battery chargers have also been marketed.

Figure 39. Applications of solar cells in consumer products (SANYO)

6. Economic Aspects and Market

Figure 40 shows the world production volume of photovoltaics [114], [115]. In the past, there was a rapid increase from 1982 onward, a decline in 1985, and a continuous increase since the late 1980s with an average annual rate of 15 %. The decrease of the European share in recent years is a result of the closing of PV production in Germany. Only a short time ago plans have been announced to erect solar cell production facilities in Germany again.

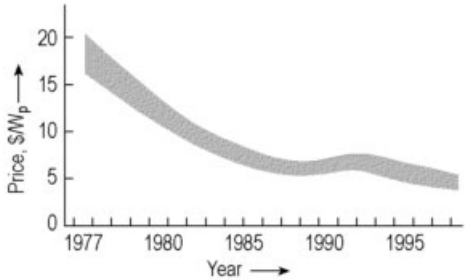

Figure 41. Development of PV module prices (average values worldwide)

The development of the cost of photovoltaic modules within the last 20 years is indicated in Figure 41. The minimum in 1989 resulted from the special offers of modules during the dismantling of the Carrisa Plains plant. These special offers forced the PV industry to set selling prices below cost. But in 1990 the prices were up again, especially in Europe. This was due to the German 1000-PV-roofs-program that stimulated the market considerably. Enhanced production capacities, however, brought the prices down again

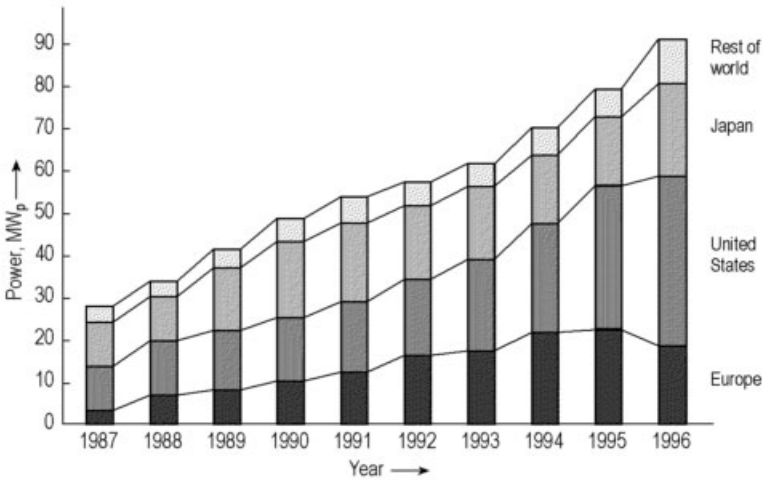

Figure 40. Worldwide shipments of the PV industry; after [114] and [115]

to the present values of \$ 4 to 5/W_p in the United States and \$ 5 to 6/W_p in Europe.

In the past, the cost reduction potential has often been estimated and cost reduction goals have been formulated, e.g., by the DOE and the Japanese Sunshine Project. These goals, coupled to proposed shipments, have always been too optimistic. Even for the late 1980s the cost goal for PV modules was set to values below \$ 1/W_p! The same figure was set again for the late 1990s; and most probably it will not be achieved either. In [114] the European Photovoltaic Industry Association (EPIA) in cooperation with the Commission of the European Communities (CEC) proposed much more realistic values: it is estimated that until 2005 the industrial cost of PV modules could be reduced to 1.8 ECU/W_p, i.e., \$ 2/W_p.

Considering the fields of application of PV, it is interesting to compare the share of these fields in the past and in recent years. Figure 42 shows that in 1983 demonstration projects financed by governments absorbed half of the PV production. In other words, during the first 30 years of the era of modern photovoltaics a substantial market could not be created despite permanent governmental funding. The application shares of 1993 (see Fig. 42) clearly indicate that the "market take-off" in the late 1980s is based on cost-effective stand-alone PV systems. Government demonstration projects as well as grid-connected PV systems only play a marginal role. However, part of the stand-alone systems indicated in Figure 42 are governmentally supported as well!

A further breakdown of PV applications is depicted in Figure 43. As indicated, the communications sector has been the major area of applications in the first half of the 1990s. A significant portion is utilized for solar home systems, defined as simple PV systems mainly for lighting purposes in developing countries; and for camping, boating and leisure applications, a typical application in industrialized countries. Grid-connected PV roof and facade systems represent the fastest growing market in Europe.

7. Future of Photovoltaic Cells

The GENESIS system (Global Energy Network Equipped with Solar cells and International Superconductor grids), discussed in [117] is based on the idea that, by the early 2000s, PV systems all over the world will be interconnected by superconductors to form a global photovoltaic power system. The GENESIS system is considered to integrate also superconductor power storage systems. This may be a scenario for the very far future; it is a fascinating vision with a touch of science fiction.

To develop a more realistic picture for the near- and medium-term future of photovoltaics, conclusions from the present situation should be drawn. A down-to-earth scenario is the "Diffusion Model" of the U.S. utility company PG&E (Pacific Gas and Electric; [118]) (see Fig. 44). The various elliptical areas or sectors represent markets or fields of applications. Today, PV is successful in the market of stand-alone services and systems for remote users. In addition, PV is currently competitive in the sector of power supply for remote communities and islands. These are still stand-alone applications. By gaining experience and increasing turnover in these fields, cost will decrease and PV will become more and more competitive in the next niche, which is the sector of grid-connected building-integrated systems, i.e., PV roofs and PV facades. Continuously increasing turnover will again reduce production cost, and the next niche will be commercially attainable. This mechanism is believed to promote PV in a natural way, based on the principles of a free market. Finally, after having gained enough experience, after having achieved the status of a large-scale industry, PV is expected to be competitive also in the field of bulk power generation.

From the history of photovoltaics and from the present situation the future of PV can be summarized as follows:

1) R & D will provide the PV industry with technologies for the production of higher-efficient and/or cheaper solar cells.
2) Progress in R & D and increased turnover will make PV systems more and more competitive in increasing fields of applications; a leap of cost reduction may occur when polycrystalline thin-film solar cells – those of crystalline silicon or those of compound semiconductors – will be transferred to industrial mass production. A drastic reduction in production cost for amorphous silicon is

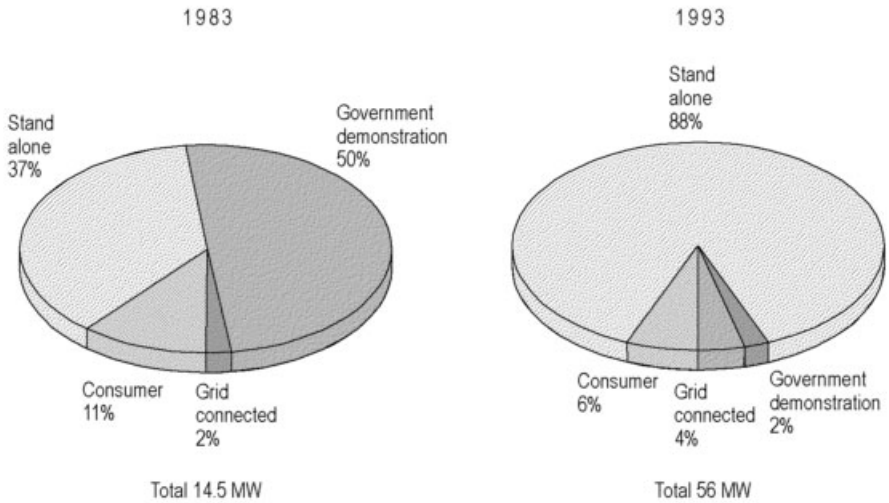

Figure 42. Application shares of PV modules at different times; after VIGOTTI [116]

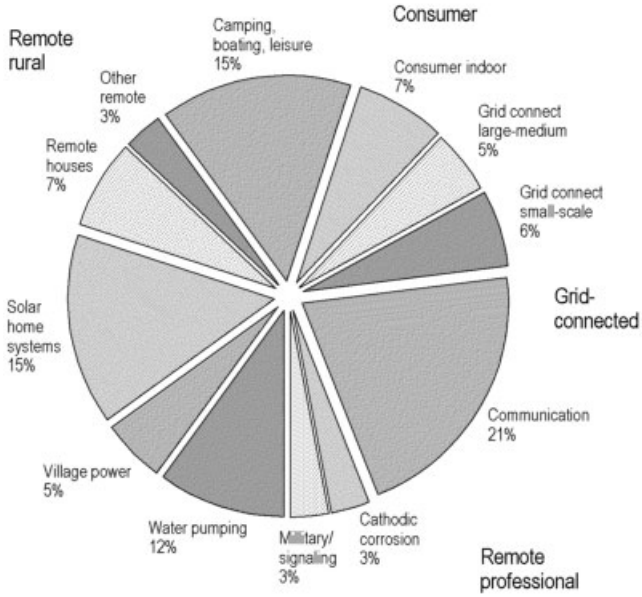

Figure 43. Main application segments of PV (average 1990 – 1994); after [114]

not anticipated. Amorphous Si is in industrial production since 15 years and progress will only be gradual.

3) The Third World market will continue to provide excellent prospects for PV. A rural population of more than 1×10^9 people without access to electricity represents a market potential of more than 16 GW.
4) Stand-alone PV for remote applications within Europe represents a potential market of some 150 MW. Here, it should be taken into account that within Europe money for financing PV installations is available,

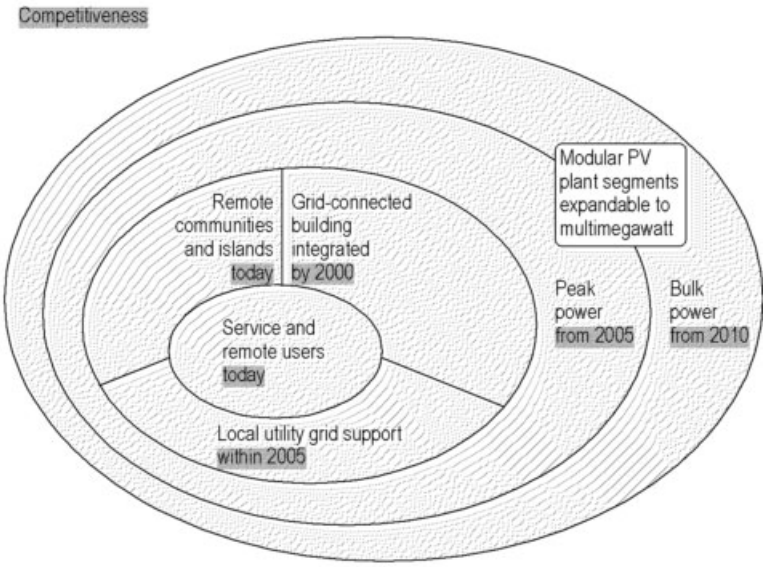

Figure 44. Diffusion model of PV utility applications by PG&E [118]

whereas in developing countries the need is huge, but money is scarce!

5) PV integration into buildings is a rapidly growing market. It is not primarily based on the purpose of an economic application of PV, but on engagement for environment and sustainable development, on the desire to support renewable energies and PV industry, and also on the wish to equip a building with exceptional features. This especially holds for PV facades, which can easily be in the economic range due to the saving of conventional facade elements (see Chap. 5).
6) If PV succeeds to lower system cost to the range \$ 2.00 – 2.50/W_p, then it will be competitive with diesel engines. According to experts, the diesel power market – for water pumping and for other purposes – is in the range of "thousands of megawatts."

8. References

1. O. Hohmeyer, *Energiewirtsch. Tagesfragen* **39** (1989) 579.
2. R. Friedrich, A. Voss, *Energiewirtsch. Tagesfragen* **39** (1989) 640.
3. E. Becquerel, *Comptes Rendues* **9** (1839) 561.
4. W. G. Adams, R. E. Day, *Proc. Royal Soc.* **A25** (1877) 113.
5. D. M. Chapin, C. S. Fuller, G. L. Pearson, *J. Appl. Phys.* **25** (1954) 676.
6. D. C. Reynolds, G. Leies, L. L. Antes, R. E. Marburger, *Phys. Rev.* **96** (1954) 533.
7. G. Nadjakov, R. Andreitchine, M. Borissov, *Izv. Bulg. Nauk.* **4** (1954) 10.
8. D. A. Cusano, *Solid State Electron.* **6** (1963) 217.
9. J. J. Loferski, *Progr. in Photovoltaics: Res. a. Appl.* **1** (1993) 67.
10. S. Wagner, J. L. Shay, P. Migliorato, M. Kasper, *Appl. Phys. Lett.* **25** (1974) 434.
11. D. E. Carlson, C. R. Wronsky, *Appl. Phys. Lett.* **28** (1976) 434.
12. L. L. Kazmerski, *Renewable Sustainable Energy Rev.* **1** (1997) 71.
13. W. Shockley: *Electrons and Holes in Semiconductors,* van Nostrand, New York 1950.
14. E. Spenke: *Elektronische Halbleiter,* 2nd ed.Springer-Verlag, Berlin – Heidelberg – New York 1965.
15. S. M. Sze: *Physics of Semiconductor Devices,* J. Wiley & Sons, New York – London – Sydney – Toronto 1969.

16. W. H. Bloss et al.: "Photovoltaic Power Stations," in C. J. Winter, R. L. Sizmann, L. L. Vant-Hull (eds.): *Solar Power Plants,* Springer-Verlag, Berlin 1991, pp. 283 – 335.
17. J. J. Loferski, *J. Appl. Phys.* **27** (1956) 777.
18. C. J. Henry, *J. Appl. Phys.* **51** (1980) 4494.
19. Y. Kuwano et al., *Jpn. J. Appl. Phys.* **21** (1982) 235.
20. E. D. Jackson, *Trans. Conf. "Use of Solar Energy," Tucson, Arizona, 1955*, p. 122.
21. N. A. Gokcen, J. J. Loferski, *Sol. Energy Mater.* **1** (1979) 271.
22. J. C. C. Fan et al., *Conf. Rec. 16th IEEE Photov. Spec. Conf.,* IEEE, New York, 1982, p. 692.
23. S. I. Fonash: *Solar Cell Device Physics,* Academic Press, New York 1983.
24. A. L. Fahrenbruch, R. H. Bube: *Fundamentals of Solar Cells,* Academic Press, New York 1983.
25. M. H. Leipold, *Proc. 4th EC Photov. Solar En. Conf., Stresa, 1982*, D. Reidel, Dordrecht, 1982 , p. 985.
26. H. A. Aulich, *Conf. Rec. 17th IEEE Photov. Spec. Conf., Orlando, 1984,* IEEE, New York, 1984, p. 390.
27. E. Sirtl, in W. H. Bloss, F. Pfisterer (eds.): *Advances in Solar Energy Technology'', Proc. ISES Solar World Congress, Hamburg, 1987* **vol. I,** Pergamon Press, Oxford 1988, p. 34.
28. Solar Energy Research Institute: *Basic Photovoltaic Principles and Methods,* Van Nostrand Reinhold, New York
29. A. Goetzberger, A. Räuber, *Proc. ISES Solar World Congr., Kobe, 1989* Pergamon Press, Oxford 1990, p. 43.
30. J. Mandelkorn, J. H. Lamneck, Jr., *Conf. Rec. IXth IEEE Photov. Spec. Conf. 1972,* IEEE, New York 1972, p. 66.
31. M. Wolf, *Proc. IEEE* 1963, p. 674.
32. J. Lindmayer, J. Allison, *Conf. Rec. IXth IEEE Photov. Spec. Conf. Silver Spring, 1972,* IEEE, New York 1972, p. 83.
33. R. Hezel: *Status Report Photovoltaik (Germany) 1990* (BMFT, Bonn, 1990) p. 15-1.
34. J. Haynos, J. Allison, R. Arndt, A. Meulenberg, *Intern. Conf. Photov. Power Generation, Hamburg, 1974* DGLR Köln, 1974, p. 487.
35. M. A. Green et al., *Conf. Rec. 17th IEEE Photov. Spec. Conf., Orlando, 1984* IEEE, New York 1984, p. 386.
36. M. A. Green et al., *Proc. Int. Photovoltaic Sci. Eng. Conf. 5th* 1990, 603.
37. R. M. Swanson, *Proc. Int. Photovoltaic Sci. Eng. Conf. 4th* 1989, 573.
38. R. R. King, R. M. Swanson, *Proc. Int. Photovoltaic Sci. Eng. Conf. 4th* 1989, 795.
39. J. L. Stone: "Present Status and Future Prospects of Solor Cells," *Proc. Int. Solar Energy Soc. Solar World Congr.* 1989, 53.
40. R. A. Sinton et al., *Proc. 8th EC Photovoltaic Solar Energy Conf., Florence, 1988,* Kluver, Dordrecht 1988, 1472.
41. J. Knobloch et al., *Conf. Rec. 25th IEEE Photov. Spec. Conf. 1996, Washington D.C.,* IEEE, Piscataway 1996, p. 405.
42. W. E. Spear, P. G. LeComber, *Solid State Commun.* **17** (1975) 1193.
43. D. E. Carlson, C. R. Wronsky, *Appl. Phys. Lett.* **28** (1976) 671.
44. J. I. Pankove (ed.): "Hydrogenated Amorphous Silicon," part A, in *Semiconductors and Semimetals,* vol. 21, Academic Press, Orlando, Fla, 1984.
45. Y. Kuwano et al., *Jpn. J. Appl. Phys. Part 1* **21** (1982) 413.
46. T. Tawada et al., *Appl. Phys. Lett.* **39** (1981) 237.
47. M. Stutzmann, W. B Jackson, C. C. Tsai, *Phys. Rev.* **B32** (1985) 23.
48. D. Adler et al., *J. Noncryst. Sol.* **66** (1984) 273.
49. T. Yoshida et al., *Proc. 10th EC Photov. Solar En. Conf., Lisbon, 1991*, Kluwer, Dordrecht 1991, p. 1193.
50. S. Guha, *Techn. Dig. Int. Photov. Sci. Engg. Conf. 6th, New Delhi* 1992, 55.
51. H. Takakura et al., *Tech. Dig. Int. Photov. Sci. Engg. Conf. 1st, Kobe* 1984, 641.
52. Mitchell et al., *Tech. Dig. Int. Photovoltaic Sci. Eng. Conf. 3rd., Tokyo* 1987, 443.
53. Y. Kuwano et al., *Jpn. J. Appl. Phys.* **20** (1981) no. 2, 213.
54. S. Nakano et al., *Jpn. J. Appl. Phys.* **25** (1986) 1986.
55. M. Konagai, *Tech. Dig. Int. Photov. Sci. Engg. Conf. 9th, Myazaki, Japan* 1996, 145. S. Guha et al., *Tech. Dig. Int. Photov. Sci. Engg. Conf. 9th, Myazaki, Japan* 1996, 283, P. D. Maycock, *Photovoltaic News* **15** (1996) no. 12.
56. A. V. Shah, R. Platz, H. Keppner, *Solar En. Mater. Solar Cells* **38** (1995) 501.
57. K. Prasad, Thesis, University of Neuchátel, Suisse 1991.
58. A. Catalano, *Conf. Rec. 24th IEEE Photov. Spec. Conf. / 1st World Conf. Photov. En. Conv., Hawaii 1994*, IEEE, Piscataway 1994, p. 52.

59. M. Spitzer et al., *Conf. Rec. 14th IEEE Photov. Spec. Conf., 1980*, IEEE, New York 1980, p. 375.
60. A. K. Gosh, C. Fisherman, T. Fang, *J. Appl. Phys.* **50** (1979) 3454.
61. R. Brendel, M. Hirsch, R. Plieninger, J. H. Werner, *Proc. 13th EC Photov. Solar En. Conf. 13th, Nice, 1995,* H. S. Stephens, Bedford 1995, p. 432.
62. J. H. Werner, R. Bermann, R. Brendel in R. Helbig (ed.), *Adv. Solid State Phys.* **34** (1994) 115.
63. G. F. Zheng, , A. B. Sproul, S. R. Wenham, M. A. Green, *Conf. Rec. 25th IEEE Photov. Spec. Conf., 1996*, IEEE, Piscataway 1996, p. 465.
64. M. A. Green et al., *Proc. 13th EC Photov. Solar En. Conf., Nice, 1995* S. Stephens, Bedford 1995, p. 13.
65. A. Barnett et al., *Conf. Rec. 18th IEEE Photov. Spec. Conf., Las Vegas 1985*, IEEE, New York 1985, p. 1094.
66. M. G. Mauk et al., *Conf. Rec. 25th IEEE Photov. Spec. Conf., 1996*, IEEE, Piscataway 1996, p. 147.
67. T. Matsuyama et al., *Proc. 12th EC Photov. Sci. Engg. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 1827.
68. S. R. Wenham et al., *Conf. Rec. 24th IEEE Photov. Spec. Conf., 1994*, IEEE, Piscataway 1994, p. 1234.
69. R. Bergmann et al., *Conf. Rec. 25th IEEE Photov. Spec. Conf., 1996*, IEEE, Piscataway 1996, p. 365.
70. M. A. Green, *Prog. Photov.* **2** (1994) 87.
71. K. Zweibel, A. Hermann, *J. Soc. Photo-Opt. Instrum Eng. (SPIE J.)* **543** (1985) 119.
72. D. Bonnet, H. Richter, K.-H. Jäger, *Proc. EC 13th Photov. Solar En. Conf., Nice, 1995*, H. S. Stephens, Bedford 1995, p. 1456.
73. Y. Tyan et al., *Conf. Rec.16th IEEE Photov. Spec. Conf., San Diego, 1982*, IEEE, New York 1982, p. 928.
74. S. Ikegami, *Tech. Dig. Int. PV Sci. Engg. Conf., Tokyo 1987*, Jap. Convention Serv., Tokyo 1987, p. 677.
75. T. A. Gessert, T. J. Coutts, *AIP Conf. Proc.* **306** (1993) 345.
76. M. E. Özsan et al., *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 1604.
77. J. M. Woodcock et al., *Proc. 12th EC Photov. Solar En. Conf.,Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 948.
78. G. Attar et al., *AIP Conf. Proc.* **306** (1993) 309.
79. T. L. Chu et al., *Proc. 11th EC Photov. Solar En. Conf., Montreux, 1992*, Harwood Acad. Publ., Chur 1993, p. 988.
80. R. Currey (ed.): *Photovoltaic Insider's Report*, vol. XV, no. 12, Dallas TX, 1996.
81. S. Kumazawa et al., *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p.113.
82. T. Takamoto et al., *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p.121.
83. R. Curey (ed.): *Photovoltaic Insider's Report*, vol. XVI, no.5, Dallas TX, 1997.
84. W. H. Bloss, F. Pfisterer, M. Schubert, T. Walter, *Prog. Photov.* **3** (1995) 3.
85. J. Hedström et al., *Conf. Rec. 23rd IEEE Photov. Spec. Conf., Louisville, 1993*, IEEE, New York 1993, p. 364.
86. L. Stolt et al., *Proc. 13th EC Photov. Solar En. Conf., Nice, 1995*, H. S. Stephens, Bedford 1995, p. 1451.
87. M. Contreras et al., *Mater. Res. Soc. Spring Meet.* 1996.
88. J. R. Tuttle, M. Contreras, *J. Appl. Phys.* **77** (1995) 1.
89. H. W. Schock, "$CuInSe_2$ and Related Materials for Thin Film Solar Cells," *Adv. Solid State Phys.* **34** (1995) 147.
90. R. Mickelsen, W. S. Chen, *Conf. Rec.15th IEEE Photov. Spec. Conf., Orlando, 1981*, IEEE, New York 1981, p. 800.
91. B. Basol, *Jpn. J. Appl. Phys.* **32** (1993) Suppl. 3, 35.
92. A. Delahoy et al., *AIP Conf. Proc.* **306** (1993) 370.
93. S. Zweigart, D. Schmid, J. Kessler, H. W. Schock, *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 648.
94. A. Gabor et al., *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 939.
95. J. R. Tuttle et al., *Mater. Res. Soc. Symp. Proc.* **426** (1996) 143.
96. T. Wada, *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p. 131.
97. H. W. Schock et al., *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p.123.
98. T. Walter et al., *Proc. 13th EC Photov. Solar En. Conf., Nice, 1995*, H. S. Stephens, Bedford 1995, p. 597.

99. T. Walter, R. Menner, C. Koeble, H. W. Schock, *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 1755.
100. D. Schmid, M. Ruckh, F. Grunwald, H. W. Schock, *J. Appl. Phys.* **73** (1996) 2902.
101. H. W.. Schock, *Appl. Surf. Sci.* **92** (1996) 606.
102. R. R. Gay, *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p. 149.
103. B. Dimmler, H. W. Schock, *Prog. Photov.* **4** (1996) 425.
104. K. Kushiya et al., *Tech. Dig. Int. Sci. Engg. Conf. 9th (PVSEC-9), Miazawa, Japan,* 1996, p. 143.
105. H. W. Schock, A. Shah, *Proc. 14th EC Photov. Solar En. Conf., Barcelona, 1997*, H. S. Stephens, Bedford 1997, p. 2000.
106. P. Moskowitz, K. Zweibel, V. Fthenakis, NREL, Golden CO, 1990.
107. E. Alsema, B. C. W. van Engelenburg, *Proc. 11th EC Photov. Solar En. Conf., Montreux, 1992*, Harwood Acad. Publ., Chur 1993, p. 995.
108. M. Grätzel, A. J. McEvoi, *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 579.
109. Final Report, MUSIC-FM, Contract No APAS RENA CT94 0008, Comm. Eur. Union, Brussels.
110. R. Curey (ed.): *Photovoltaic Insider's Report*, vol. XVI, no. 5, Dallas TX, 1997.
111. A. Hagfeldt et al., *Solar En. Mater.* **32** (1993) 481.
112. W. H. Bloss, F. Pfisterer, W. Kleinkauf, M. Landau, H. Weber, H. Hullmann, *Proc. 10th EC Photov. En. Conf., Lisbon 1991*, Kluwer, Dordrecht 1991, p. 1295.
113. Y. Kuwano et al., *J. Non-Cryst. Solids* **115** (1989) 189.
114. EPIA and CEC: *Photovoltaics in 2010*, vol. 1, Summary Report, CEC, DEG XVII, Sept. 1995.
115. P. D. Maycock (ed.): *Photovoltaic News,* PV Energy Systems, Casanova VA, yearly Feb. issues.
116. R. Vigotti, *Proc. 12th EC Photov. Solar En. Conf., Amsterdam, 1994*, H. S. Stephens, Bedford 1994, p. 22.
117. Y. Kuwano, *Ullmann's* 5th ed, **Vol. A20,** 1992, p. 161.
118. A. Taschini, J. J. Iannucci, *Proc. Exec. Conf. Photov. Sys. Electr. Util. Appl., Taormina 1990*, OECD/IEA, Paris 1992, p. 51.

Phthalic Acid and Derivatives

PETER M. LORZ, BASF AG, Ludwigshafen, Germany (Chap. 1, 2, 3 and 4)

FRIEDRICH K. TOWAE, BASF AG, Ludwigshafen, Germany (Chap. 1, 2, 3 and 4)

WALTER ENKE, BASF AG, Ludwigshafen, Germany (Chap. 5)

RUDOLF JÄCKH, BASF AG, Ludwigshafen, Germany (Chap. 6)

NARESH BHARGAVA, BASF Canada Inc., Cornwall, Ontario, Canada (Chap. 1, 2, 3, 4 and 5)

WOLFGANG HILLESHEIM, GUBDH Consultancy, Schwetzingen, Germany (Chap. 6)

Related Articles Isophthalic acid is treated under → Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid

1. Phthalic Acid 2
2. Phthalic Anhydride 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.3. Resources and Raw Materials . . 3
2.4. Production 3
2.4.1. Gas-Phase Oxidation 3
2.4.1.1. Catalyst and Reaction Mechanism . 4
2.4.1.2. Apparatus and Important Process Steps in the Gas-Phase Oxidation of *o*-Xylene 6
2.4.2. Fluidized-Bed Oxidation 8
2.4.3. Liquid-Phase Oxidation of *o*-Xylene 9
2.5. Environmental Protection 9
2.6. Quality Specifications and Analysis 9
2.7. Economic Aspects 10
2.8. Storage and Transportation 10
2.9. Uses 10
3. Phthalimide 10
3.1. Properties 10
3.2. Production 11
3.2.1. Production from Phthalic Anhydride and Ammonia 11
3.2.2. Production from Phthalic Anhydride and Urea 11
3.2.3. Production from *o*-Xylene 12
3.3. Uses 12
4. Phthalonitrile 12
4.1. Properties 12
4.2. Production 12
4.2.1. Production from *o*-Xylene 12
4.2.2. Production from Phthalic Acid Derivatives 13
4.3. Uses 14
5. Phthalates 14
5.1. Physical and Chemical Properties 14
5.2. Raw Materials 14
5.3. Production 14
5.4. Environmental Protection 17
5.5. Quality Specifications 17
5.6. Storage and Transportation 18
5.7. Uses 18
5.8. Economic Aspects 18
6. Toxicology 19
6.1. Use of and Exposure to Phthalic Acid and Derivatives 19
6.2. Toxicological Profiles 19
6.2.1. Phthalic Acid 19
6.2.2. Phthalic Anhydride 19
6.2.3. Phthalimide 20
6.2.4. Phthalonitrile 20
6.2.5. Phthalate Esters 21
6.2.5.1. Metabolism and Toxicokinetics . . . 21
6.2.5.2. Acute Toxicity 24
6.2.5.3. Irritation and Sensitizing Potential . 25
6.2.5.4. Repeated DEHP Dosing 25
6.2.5.5. Genotoxicity and Mutagenicity . . . 26
6.2.5.6. Carcinogenicity 27
6.2.5.7. Reproductive Toxicity 28
6.2.5.8. Effects of Phthalate Esters by Groups 33
6.3. Risk Assessment 36
6.3.1. Biomonitoring and Human Exposure to Phthalate Esters 37
6.3.2. Carcinogenicity 38
6.3.3. Toxicity to Reproduction 39
6.4. Risk Management 39
7. References 40

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_181.pub2

1. Phthalic Acid

Phthalic acid [*88-99-3*], 1,2-benzenedicarboxylic acid, has remained unimportant industrially. It is formed as a byproduct in the manufacture of phthalic anhydride.

Physical Properties. Phthalic acid, $C_8H_6O_4$, M_r 166.14, forms colorless, monoclinic crystals which melt at 191 °C (sealed tube) and are converted into phthalic anhydride with the elimination of water at 210–211 °C. Some physical properties of phthalic acid are as follows [1 – 3]:

Density (15 °C)	1.593 g/cm^3
Heat of fusion	315.3 J/g
Specific heat of solid (0–99 °C)	1.214 J g^{-1} K^{-1}
Heat of combustion	19 657.03 J/g
Heat of formation	43 714.34 J/g
Heat of solution at 25 °C	123.55 J/g
Flash point	168 °C

2. Phthalic Anhydride

Phthalic anhydride [*85-44-9*], isobenzofuran-1,3-dione, has been commercially produced continuously since 1872 when BASF developed the naphthalene oxidation process. It was the first anhydride of a dicarboxylic acid to be used commercially and is comparable in its importance to acetic acid.

The most important derivatives of phthalic anhydride are plasticizers and, to a lesser degree, polyester resins and dyes.

About 60 years after its discovery in 1836 by A. Laurent, a more effective commercial process for its production was introduced, which was based on mercury-catalyzed liquid-phase oxidation of naphthalene.

The breakthrough that led to commercial production of a quality product was the development of the gas-phase oxidation of naphthalene or *o*-xylene in an air stream with vanadium oxide as catalyst [4].

2.1. Physical Properties

Phthalic anhydride, $C_8H_4O_3$, M_r 148.12, forms colorless needles or platelets, with a monoclinic or rhombic crystalline form. Some important physical properties of phthalic anhydride are as follows [1, 5, 6]:

Density of solid (4 °C)	1.527 g/cm^3
Specific vapor density (1013 mbar)	6.61 kg/m^3
Melting point	131.6 °C
Boiling point (1013 mbar)	295.1 °C
Heat of fusion	159.1 J/g
Heat of combustion	22 160.7 J/g
Heat of formation from naphthalene	12 058 J/g
Heat of formation from *o*-xylene	8625 J/g
Heat of sublimation	601 J/g
Heat of evaporation	441.7 J/g
Flash point	152 °C
Ignition temperature	580 °C
Upper limit of flammability (1013 mbar)	10.5 vol %
Lower limit of flammability (1013 mbar)	1.7 vol %
Lower dust explosion limit	25 g/m^3

The density of liquid phthalic anhydride in the range 140–240 °C can be calculated by using the following equation [7]:

$$\varrho/\mathrm{kg\,m^{-3}} = 1321.55 - 0.6697\,(\mathrm{t}/^\circ\mathrm{C}) - 0.000905(\mathrm{t}/^\circ\mathrm{C})^2 \quad (1)$$

Table 1 lists the solubility of phthalic anhydride in various solvents.

Table 1. Solubility of phthalic anhydride

Solvent	Temperature, °C	Solubility, g/100 g
Water	20	1.64
Water	50	1.74
Water	100	19.0
Carbon disulfide	20	0.7
Formic acid	20	4.7
Pyridine	20	80
Benzene		soluble
Ethanol	20	soluble
Diethyl ether	20	slightly soluble

The reported explosion hazard data of phthalic anhydride in air vary significantly [4, 5]. Explosions can occur at concentrations below 100 g/m^3, depending on the impurities present. Recent incidents in production plants indicate that phthalic anhydride concentrations exceeding 35 g/m^3 in the reaction product gas are capable of ignition if heat-transfer salt enters the reactor due to broken reactor tubes.

2.2. Chemical Properties

As a cyclic anhydride, phthalic anhydride is a reactive compound, but in addition, the otherwise very stable aromatic ring is capable of reaction. Phthalic anhydride reactions which have

achieved commercial importance are summarized below. The most important is the reaction with alcohols or diols to give esters or polyesters.

Unsaturated polyester resins are obtained by polycondensation in the presence of maleic anhydride or fumaric acid.

One or both of the carboxy groups can react with ammonia to give phthalic monoamide and phthalimide [*85-41-6*] or phthalonitrile [*91-15-6*].

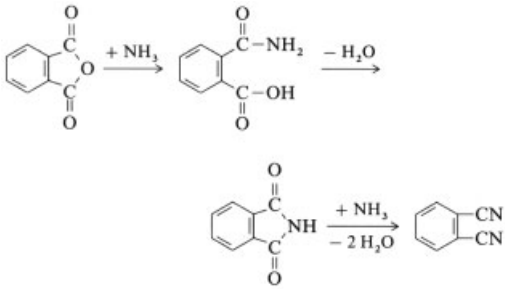

Phthalein and rhodamine dyes, some of which have been in production for over 100 years and have not yet lost their importance, are obtained by reaction of phthalic anhydride with phenols, aminophenols or quinaldine derivatives. The Friedel–Crafts reaction of phthalic anhydride with benzene derivatives followed by ring closure to form anthraquinone derivatives is of importance as a route to Indanthrene dyes.

Much attention has been devoted to the rearrangement of dipotassium phthalate to produce terephthalic acid [*100-21-0*] [8], but due to technical problems the process is no longer used.

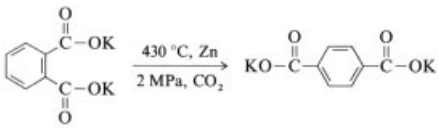

3,5-Dihydrophthalic acid can be produced by electrochemical hydrogenation of phthalic anhydride [4]. Hydrogenation with a nickel catalyst produces phthalide [*529-20-4*].

2.3. Resources and Raw Materials

In the 1960s a fundamental shift took place in the raw material base for phthalic anhydride. From 1960 to 1975 the production was switched from 100 % coal-tar naphthalene to about 75 % *o*-xylene. In 1991, 85 % of phthalic anhydride was produced from *o*-xylene [9]. Naphthalene derived from petroleum, which is produced mainly in the United States, has not gained importance as a raw material for phthalic anhydride.

The changeover to *o*-xylene as the feedstock was inevitable because the quantities of naphthalene derived from coal tar depend on the production of coke and these were unable to keep pace with the increasing demand for naphthalene. On the other hand, naphthalene is an inevitable byproduct of coke production which should be used commercially; therefore, the phthalic anhydride process will continue to use some naphthalene as feedstock in the future. *o*-Xylene, which can readily be separated from the mixture of xylenes containing roughly one third *o*-xylene and two thirds *p*-xylene, is nowadays available in adequate quantities from cracking plants and refineries. However, in the past, variations in the demand for *p*-xylene have often affected the availability and price of *o*-xylene, as has the need to increase the amount of alkyl aromatics in unleaded gasoline.

These facts and the need to be able to react to varying raw material prices have resulted in plants being planned which are capable of processing *o*-xylene or naphthalene or mixtures of the two with tailor-made catalysts.

2.4. Production

2.4.1. Gas-Phase Oxidation

Phthalic anhydride is predominantly produced on an industrial scale by gas-phase oxidation of *o*-xylene or naphthalene.

General Features. Preheated *o*-xylene is introduced into a stream of hot air. The *o*-xylene–air mixture is passed through a tubular reactor where the exothermic oxidation takes place on a highly selective catalyst. The heat of reaction is used to produce steam, only part of which is utilized in the plant itself. Any excess steam can be used elsewhere. The gases emerging from the reactor are precooled. At high loadings of phthalic anhydride in the product gases, some liquid phthalic anhydride can be won in a liquid condenser. The product gases are then fed

to a switch condenser system where the phthalic anhydride is condensed on the finned tubes as a solid. The switch condensers are cooled by a heat-transfer oil in an automated switching cycle. During the heating cycle the deposited phthalic anhydride is melted and collected in a storage tank.

After the phthalic anhydride has been separated, the exhaust gases still contain byproducts and small quantities of phtalic anhydride and must be cleaned by scrubbing with water, or catalytically or thermally incinerated. If scrubbing with water is employed, it is possible to concentrate the main byproduct, maleic acid, and the scrubbing solution can then be processed further to yield fumaric acid or maleic anhydride [10, 11]. If the scrubbing of the exhaust gases is combined with production of maleic anhydride, the discharge of any polluted wastewater from the plant can be avoided.

The crude phthalic anhydride is transferred to a continuous thermal/chemical treatment system, which converts the phthalic acid formed as a byproduct into the anhydride. The crude product is then purified in a continuous two-stage distillation system [12].

The BASF process can be operated with a wide range of o-xylene loadings up to 105 g/m^3 (STP). The reactor outlet gases are further treated catalytically in a finishing reactor to decrease the amount of byproducts and to improve product quality. Flexible operation of the reactor system enables the phthalic anhydride yield to be optimized and decreases the amount of residues and volatile organic compounds. Simultaneously a smoother operation of the catalyst in the tubular reactor prolongs its life. Addition of SO_2 to activate the catalyst is not required, and there is no need for a liquid condenser before collecting the crude phthalic anhydride in the switch condensers. Because of the low byproduct content, chemical treatment of the crude phthalic anhydride is unnecessary. The plant design is optimized for low energy consumption and a high net export of steam and electrictity [13].

The Wacker process makes it possible to use *o*-xylene or naphthalene or mixtures of the two. With *o*-xylene, loadings of 90–100 g/m^3 (STP) are possible [14, 15]. Modifications of the process are aimed at saving energy [16].

The Nippon Shokubai VGR Process. In patent literature, the Nippon Shokubai VGR (vent gas recycling) process is described. Its characteristic feature is, that exhaust gas is recycled and added to the mixture of *o*-xylene and process air to reduce the oxygen concentration to less than 10 vol %. This makes it possible to work outside the limits of flammability despite a high *o*-xylene loading (up to 85 g/m^3 STP). Despite the low O_2 concentration, yields of up to 116 g PA/100 g xylene (given as 116 % by PA producers) have been reported. This is attributed to a specially developed catalyst system. It is also possible to use naphthalene instead of *o*-xylene in this process [17 – 21].

The Alusuisse–Ftalital LAR Process. In the LAR (low air ratio) process, *o*-xylene loadings in the process air of up to 134 g/m^3 (STP) are suggested. This could make it possible to achieve an appreciable reduction in energy and the size of the equipment [22, 23]. In commercial application *o*-xylene loadings up to 80 g/m^3 (STP) are reached [107]. Catalysts in the form of rings or half rings are used. The catalyst is capable of processing both *o*-xylene and naphthalene or mixtures of the two [24, 25].

The Rhône-Poulenc process employs *o*-xylene as raw material. The crude product is subjected to a chemical posttreatment before being purified by two-stage distillation. The waste gas is incinerated [26, 27].

The ELF Atochem/Nippon Shokubai process uses a Nippon Shokubai catalyst for reacting *o*-xylene or naphthalene. With *o*-xylene, feed loadings of up to 75 g/m^3 (STP) are used industrially [107].

2.4.1.1. Catalyst and Reaction Mechanism

The oxidation of *o*-xylene and naphthalene is nowadays carried out almost exclusively in tubular reactors cooled by a molten salt. The spherical catalysts previously employed are only used in isolated cases in old plants. In modern plants ring-type catalysts have become established for

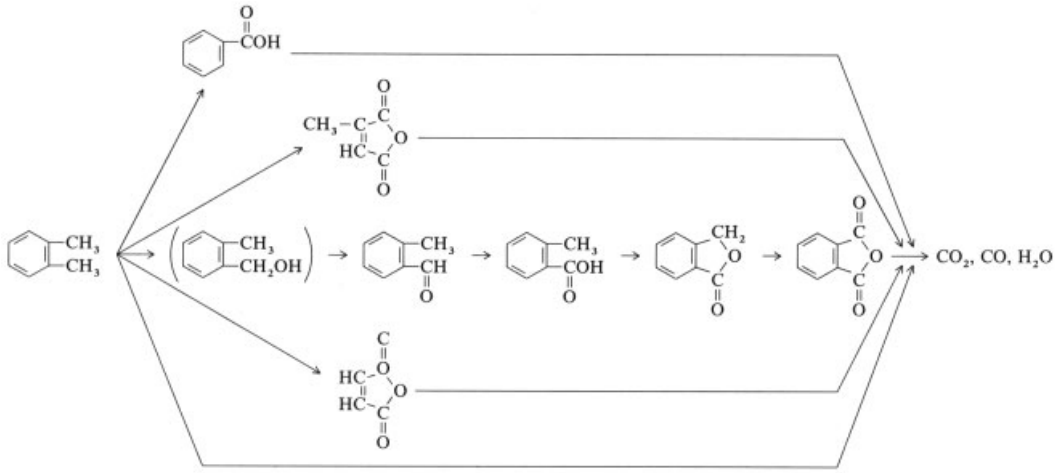

Scheme 1.

energy saving reasons [28], but catalysts on half-shell supports are also available [24, 25]. Commercial catalysts are normally composed of an inert ring-type support, made of silicate, silicon carbide, porcelain, alumina, or quartz measuring 5–10 mm in diameter, on which a thin active layer of a mixture of finely divided titanium dioxide and vanadium oxide is deposited. Compounds of antimony, rubidium, cesium, niobium, and phosphorus are added to improve the selectivity [29 – 35]. The use of a two-zone catalyst, i.e., a combination of catalysts with low activity with catalysts of higher activity, has become generally established [36]. For high loadings even three-zone catalysts are known. Some catalysts require addition of SO_2 for activation and longer service life [37, 38]. The reaction is driven to an almost quantitative turnover of *o*-xylene, to ensure low concentrations of disturbing byproducts (e.g., phthalide). It is summarized by the following equation:

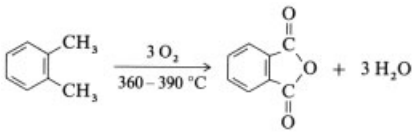

The heat of formation is 1108.7 kJ/mol. Total combustion of *o*-xylene yields 4380 kJ/mol, i. e., almost four times the heat of formation of phthalic anhydride. Oxidation in tubular reactors produces a heat of reaction between 1300 and 1800 kJ/mol of *o*-xylene.

Scheme 1 shows the important byproducts found in the reaction product. Although the intermediate alcohol has so far not been detected, it is probable that the formation of phthalic anhydride passes through it. A standard interpretation of the course of the reaction is not available in the literature. All the conclusions are very dependent on the method of measurement used and on the experimental setup. So far no commercially used catalyst has been investigated in depth. There are also no literature data relating to aging phenomena. It is generally accepted that a redox mechanism, in which the selective oxidation proceeds at oxygen atoms in the lattice, is involved. Attempts are made in [39 – 59] to describe the course of the reaction.

According to literature data, it is now possible to achieve yields of 120 kg PA /100 kg *o*-xylene [35], but it is generally accepted that reactor yields exceeding 112–114 kg PA/100 kg *o*-xylene (calc. 100 %) are improbable on an industrial scale. This is equivalent to 80–82 % of stoichiometric yield. In *o*-xylene based processes, after allowing for the losses during condensation, dehydration, and distillation, pure phthalic anhydride yields of 110–112 kg PA/100 kg *o*-xylene may be expected.

The oxidation of naphthalene proceeds in accordance with the following equation:

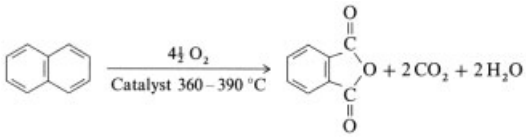

The heat of formation is 1788 kJ/mol; total combustion of naphthalene yields 5050 kJ/mol. Depending on the yield achieved and the byproducts formed, a heat of reaction of 2100–2500 kJ/mol is expected. Possible side reactions are shown in Scheme 2.

Besides maleic anhydride, naphthoquinone is also formed as a byproduct, which requires a high performance of the purification stage. Data on the course of the reaction can be found in [60 – 63].

According to the literature, yields of up to 102 kg/100kg naphthalene are achievable in commercial plants, but the final product yields generally do not exceed 98 kg/100kg naphthalene, i.e., 85 % of theory.

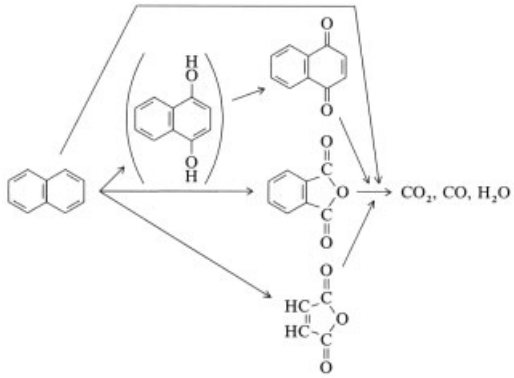

Scheme 2.

2.4.1.2. Apparatus and Important Process Steps in the Gas-Phase Oxidation of *o*-Xylene

Procedure Employing an Ignitable Xylene–Air Mixture. The current objective of phthalic anhydride technology is essentially to limit energy consumption as far as possible. This is supplemented by efforts to construct as complete a system as possible for utilizing the steam produced in the plant, or alternatively by considering driving the air blower with a steam turbine [64, 65]. It is possible to save energy costs primarily by increasing the concentration of *o*-xylene in the process air, thus reducing the quantity of process air for the same capacity and increasing the steam yield [66].

It is possible to operate existing older plants, which were originally designed for loadings of 40 g/m^3 (STP), with *o*-xylene loadings of up to 80 g/m^3 (STP) by employing modern catalysts and modifying the reactor cooling system. The attainable increase in loading depends primarily on the feasibility of removing safely the increased heat of reaction from the reactor.

The lower limit of flammability is exceeded with a loading of ca. 44 g/m^3. In order to exceed the limit of flammability, it is essential to estimate the risks to the safety of the personnel and of the plants. It is impossible to completely eliminate the risk of ignition of an explosive mixture, but the equipment may be designed to be inherently safe, i.e., to withstand the highest possible pressure. Alternatively the equipment may be safeguarded by pressure relief devices so that pressure surges are relieved safely. Probability of ignition inside the catalyst is minimal as the autoignition temperature for phthalic anhydride is 580 °C, which is far above the temperature of the reaction.

Reactors. Figure 1 shows the development of the reactor types used to produce most of the phthalic anhydride worldwide. The reactor (A) was originally designed for the von Heyden catalyst and has a salt bath circulation pump and cooler fitted internally. This reactor gave way to reactor (B), which has an external cooling system [67]; the latter makes it possible to achieve a more advantageous temperature distribution for larger diameters. Superheated steam at pressures up to ca. 70 bar is produced in both reactors. The reactor (C), with an external cooling system in which saturated steam is produced, was developed for processes with higher loadings [68]. Reactor (D) is the common type for high-load processes. It uses radial salt flow for improved heat transfer. Reactors can be machined in one piece up to a diameter of 8 m, but can also be fabricated in segments and assembled on site [69]. Such reactors may incorporate more than 20 000 catalyst tubes and the construction of single-line plants having phthalic anhydride capacities exceeding 60 000 t/a is possible.

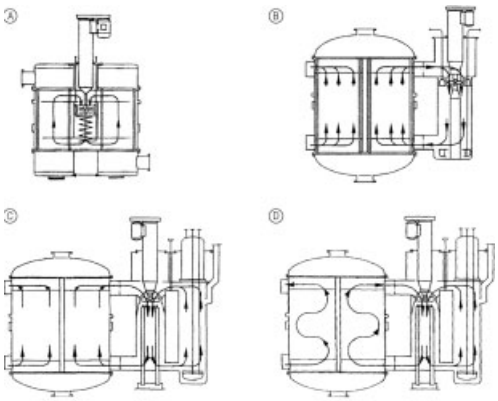

Figure 1. Reactor types used for phthalic anhydride production
A) Reactor designed for the von Heyden catalyst; B) Reactor with an external cooling system; C) Reactor with an external cooling system in which saturated steam is produced; D) Reactor with radial salt flow (DWE-type)

Switch Condenser. A decisive factor for the efficiency of a phthalic anhydride plant is the performance of the condenser system in which the phthalic anhydride (which is contained in the process gas at a concentration of less than 1 vol %) must be condensed as completely as possible. Several separating methods are discussed in the literature. Processes in which the phthalic anhydride reaction gas is extracted by absorption into solvents or liquid phthalic anhydride [70 – 72] have not been exploited commercially.

Generally, switch condensers equipped with finned tubes which are alternately heated or cooled with a hot or cold heat-transfer medium are used. Solid phthalic anhydride is desublimated on the surface of finned tube bundles and subsequently melted out and collected in a storage tank.

A milestone in the improvement of switch condenser technology was the changing of the design from the upflow-type to the downflow-type. In the old switch condensers the reaction gas enters from the bottom and passes through the internal tube bundles from the bottom to the top. Most of the phthalic anhydride is deposited on the lowest bundle.

In the downflow type, the reaction gas enters at the top and leaves the condenser at the bottom. Most of the solid product deposits on the uppermost bundle, and the lower tube bundles are washed by molten phthalic anhydride during each melting cycle. In this way maleic acid and phthalic acid, which are byproducts of the process, are removed from the tubes. Phthalic acid and especially maleic acid are highly corrosive, and their iron salts, being pyrophoric, can cause fires by self-ignition.

Cooling and heating are usually carried out by a heat-transfer oil, fed through the finned tubes. Cooling with water and heating with steam has been proposed but has no obvious advantage.

Modern condensers can separate more than 99.5 % of the phthalic anhydride from the reaction gas. Improvements in the arrangement of the finned tubes resulted in reduced oil throughput and hence lower energy consumption compared with older condensers. The mechanical flexibility has been improved by changing the design, leading to a reduction of thermal stresses when switching from hot oil to cold oil and vice versa.

The increase of the capacities in the recent years have also affected the size of equipment, and switch condensers with a surface area of up to 4000 m^2 are being used today. At the beginning of the melting cycle, when the cold oil is replaced by hot oil, the system requires the highest amount of steam to maintain the oil temperature due to the heating-up of the swith condenser and to the heat of fusion of the phthalic anhydride. By installation of an additional heat reservoir connected to the hot oil system, it is possible to eliminate peaks and make the steam demand more uniform [73 – 76]. This effect is especially pronounced when the air blower is driven by a turbine using the steam from the same plant.

Treatment of Crude Phthalic Anhydride. To obtain high-purity phthalic anhydride, the crude product is subjected to thermal treatment before purification by distillation. Treatment with chemicals is generally necessary in the case of the crude product obtained from naphthalene in order to remove the byproduct, naphthoquinone, but is also used for products manufactured from *o*-xylene. Apart from sulfuric acid, which has been used for a long time, the following substances are used to destroy naphthoquinone: sodium hydroxide [77]; boric acid/sulfuric acid [78, 79]; potassium disulfide [80]; sodium carbonate [81 – 83]; potassium maleate [84]; and aliphatic polydienes [85, 86]. The most widely used method is probably simple heat treatment, but this is only suitable for

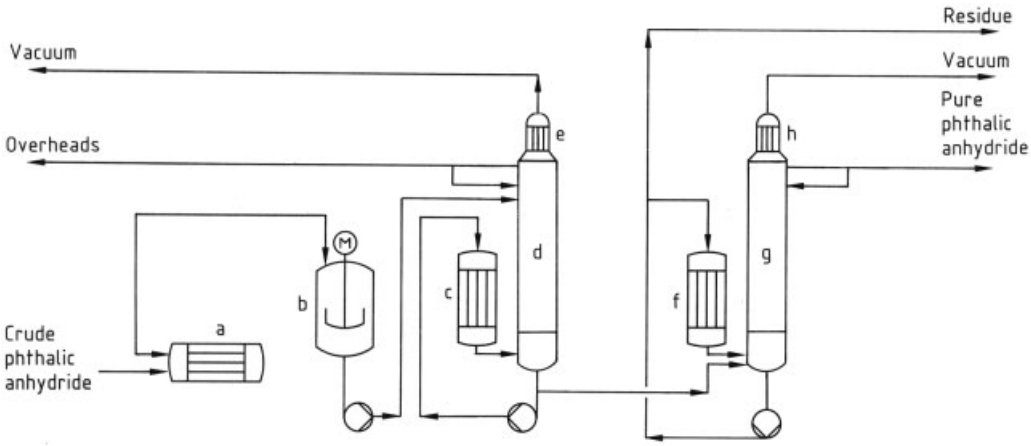

Figure 2. Purification of crude phthalic anhydride by distillation
a) Crude phthalic anhydride heater; b) Predecomposer; c) Reboiler; d) Stripper column; e) Condenser; f) Reboiler; g) Distillation column; h) Condenser

crude products with low contents of impurities. It involves heating the crude product to 100–400 °C in a cascade of vessels. Normally the temperature is limited to 230–300 °C, and the retention times are 10–24 h. In the case of phthalic anhydride produced from *o*-xylene, the treatment is intended to remove water and low-boiling contaminants such as maleic anhydride and *o*-tolualdehyde. The destruction of phthalide is also achieved in pretreatments [87 – 90].

Distillation. The final purification of the crude phthalic anhydride is normally carried out in a continuous distillation system. Figure 2 shows schematically a typical distillation plant. The crude phthalic anhydride is first pretreated in a predecomposer where byproduct phthalic acid is converted to phthalic anhydride. The product is then introduced into a column in which the low-boiling constituents such as maleic anhydride and benzoic acid are concentrated at the head and removed. The bottoms are introduced into a distillation column, from which pure phthalic anhydride is removed from the head while the residue is discharged from the bottom.

2.4.2. Fluidized-Bed Oxidation

Of the many fluidized-bed processes for the oxidation of naphthalene, the process developed by Badger may still be in use today [91 – 93]. The capacity of such plants has been increased by injecting oxygen and air into the fluidized bed, thus enabling the naphthalene concentration to be increased.

Liquid naphthalene is injected directly at the bottom of the catalyst bed and evaporates immediately, distributing itself over the entire bed. Here it comes in contact with the catalyst and reacts with the atmospheric oxygen fed in via a distributor plate.

The vigorous agitation and mixing in the fluidized bed results in a uniform temperature being maintained throughout the bed, with the temperature being in the range 345–385 °C. The heat is removed by cooling tubes installed in the bed and is used to produce high-pressure steam. Any entrained catalyst is separated by cyclones and, after being retained in specially constructed ceramic filters, is blown back into the reactor. Several filter units are used, and one of them is always blown back with a stream of air in order to remove catalyst from the filter surfaces. Up to 60 % of the phthalic anhydride can be recovered in a liquid condenser, the remainder being desublimated in switch condensers. The crude product is then purified in a distillation column under vacuum. A proposal for regenerating a deactivated catalyst with SO_2 has been published [94].

Kawasaki Steel Corporation is working on the fluidized-bed oxidation of *o*-xylene. According to patent literature a stoichiometric yield of

up to 84 % at an *o*-xylene conversion of 94 % is reached [95 – 99].

2.4.3. Liquid-Phase Oxidation of *o*-Xylene

In the liquid-phase oxidation of *o*-xylene, a mixture of acetic acid, *o*-xylene, and catalyst, which consists mainly of Co, Mn, and Br, is fed to the first vessel in a cascade. The reaction, which is completed in the subsequent vessels, is initiated under pressure while the air is injected. The water produced by the reaction is removed in the first vessel by azeotropic distillation with *o*-xylene. The isomers of phthalic acid are removed from the reaction mixture, and the phthalic anhydride is obtained by crystallization. The crude product is subjected to a pretreatment which imposes special requirements because of the bromine content. Purification can then be carried out by distillation. The *o*-xylene must have an ortho isomer content $> 99\,\%$ if the process is to achieve high yields of 130 kg of phthalic anhydride from 100 kg of *o*-xylene. Companies such as Hüls have investigated the process [100] and, in addition, further work has been done on it by Standard Oil of Indiana [101].

Sisas has described a two-step process: in the first liquid-phase oxidation step crude*o*-toluic acid is formed with a yield of 1.19 g per gram of *o*-xylene at partial *o*-xylene turnover; unreacted *o*-xylene is recycled. The second step is the gas-phase oxidation of the *o*-toluic acid to phthalic anhydride. The overall molar yield based on *o*-xylene is up to 88 % [102].

2.5. Environmental Protection

Exhaust Gas. The exhaust gas from the switch condensers still contains an appreciable quantity of organic substances and therefore must be purified before it is discharged from the plant. Scrubbing with water in a scrubbing tower is suitable for extracting maleic anhydride, the main byproduct in the exhaust gas. In a recovery process, the organic acids from the scrubbed gas are concentrated in several scrubbing towers in which the scrubbing water is circulated until the concentration is ca. 30 %. The scrubbing solution is then worked up to recover maleic anhydride or fumaric acid. It is also possible to produce sodium maleate by adding sodium hydroxide or sodium carbonate [103].

The efficiency of the scrubbing depends on the sophistication of the equipment and the number of scrubbing stages. The concentration of all organic compounds can be reduced to < 150 mg per m^3 exhaust air discharged from the plant.

The main substance emitted from phthalic anhydride plants is carbon monoxide, which cannot be removed by exhaust-gas scrubbers. A plant emits about 1100 t CO per 10 000 t of phthalic anhydride produced, depending on the operating conditions.

Apart from scrubbing towers, two processes are at present available for purifying exhaust gas:

1) Thermal combustion at 650–850 °C which involves burning the organics and carbon monoxide in the presence of an oil-fed or gas-fed reverberatory flame.
2) Catalytic combustion at 250–450 °C, which normally does not require additional fuel. This process employs noble-metal or mixed-oxide catalysts.

Such plants are supplied, for example, by Haldør Topsoe in Denmark who employ their Catox process or by Engelhard who market the Torvex process.

A comparison of thermal and catalytic exhaust-gas combustion for phthalic anhydride plants is discussed in [104].

Catalyst. Depending on the operating procedure, the catalyst normally has a service life of 2–8 years, after which it must be replaced because its activity is too low. Used catalysts may be deposited, for example, at a secure landfill site; this depends on local regulations owing to the toxicity of the additives. As a result of various problems, washing and reusing the active material has so far been uneconomical, even though proposals are repeatedly being made in this connection [105, 106].

2.6. Quality Specifications and Analysis

The specifications of technically pure phthalic anhydride are listed in Table 2. A suitable method of determining the acids by gas chromatography is to form derivatives, for example, with diazomethane, but direct determination on

Table 2. Specifications of technical-grade phthalic anhydride

Property	Average value	Method	Specification limit
Purity, %	99.9	GC	99.8 min
Solidification point, °C	131.0	ASTM 1493–67	130.8 min
Maleic anhydride, %	0.03	GC	0.05 max
Benzoic acid, %	0.03	GC	0.1 max
Color index of melt (APHA)	5–10	ASTM 3366–74	20 max
Heat stability (APHA)	10–20	ASTM 3366–74	40 max

packed columns or capillary columns is also possible.

2.7. Economic Aspects

In 1996 world production capacity for phthalic anhydride was $> 3.7\times10^6$ t/a and this figure is expected to exceed 4.6×10^6 t/a in the year 2000. A remarkable increase in capacity is occurring in south-east Asia.

The world capacities for phthalic anhydride (10^3 t) in 1996 can be broken down as follows:

Western Europe	869
Eastern Europe	641
North America	568
South America	256
Japan	320
South Korea	260
India	153
Taiwan	95
Remainder of Asia	155
Africa and Near East	124

2.8. Storage and Transportation

Pure phthalic anhydride is a very stable compound and may be stored for a prolonged period in the molten state without any change in its properties. To prevent fires, it is recommended that storage containers and transportation are blanketed with nitrogen.

Molten phthalic anhydride is subject to the following safety and transport regulations:

EC directive no.	607–009–00–4
Symbols	X_i
R phrases	36/37/38
UN no.	3256
Packaging group	III
Transportation regulations	
GGVE/RID	3.61 c
GGVS/ADR	3.61 c
ADN/ADNR	3.61 c
GGV See/IMDG code	3.3, UN-No. 3256-PG III
IATA Dangerous Goods Regulation	3, UN-No. 3256

2.9. Uses

The most important application of phthalic anhydride is the production of plasticizers. The main uses of phthalic anhydride are [13]:

Plasticizers	55 %
Unsaturated polyester resins	14 %
Alkyd resins	15 %
Other	16 %

3. Phthalimide

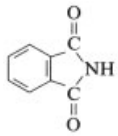

3.1. Properties

Phthalimide [*85-41-6*], 1,3-dioxoisoindoline, $C_8H_5O_2N$, M_r 147.14, crystallizes from solution as white needles or prisms, while sublimation produces platelets. The compound is sparingly soluble in water (0.3 g at 20 °C, 0.9 g at 50 °C, 2.2 g at 100 °C in 100 g of water) and readily soluble in acetic acid, sodium hydroxide solution, or potassium hydroxide solution. Some physical properties of phthalimide are as follows:

Melting point	238 °C
Heat of combustion	3560 kJ/mol
Heat of fusion	187.6 J/g
Specific heat at 100 °C	1.21 J g^{-1} K^{-1}
Vapor pressure at	
120 °C	0.10 mbar
150 °C	0.95 mbar
180 °C	5.93 mbar
220 °C	30.7 mbar
254 °C	187.6 mbar
Flash point	214 °C
Ignition temperature	530 °C

With bases, phthalimide forms water-soluble salts, which react with halogens to form the corresponding *N*-chloro, *N*-bromo, or *N*-iodo compounds. These *N*-halo compounds are also obtained if alkali-metal phthalimides are reacted with hypochlorous or hypobromous acid. When the *N*-halo compounds are heated, isatoic anhydride [*118-48-9*] or anthranilic acid (*o*-aminobenzoic acid) [*118-92-3*] is formed in a Hofmann degradation [108, 109]. The reaction of alkali-metal phthalimides with alkyl halides to give *N*-alkyl phthalimides, and subsequent hydrolysis or hydrazinolysis affords primary amines (Gabriel synthesis) [110]:

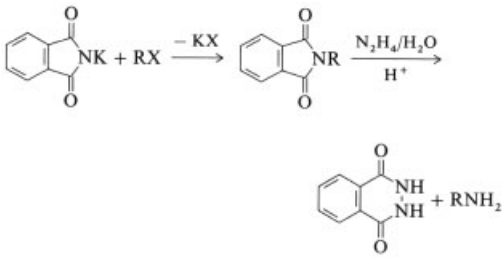

3.2. Production

Phthalimide is produced almost exclusively from phthalic anhydride and ammonia, but processes based on phthalic anhydride and urea or oxidative ammonolysis of *o*-xylene are also known.

3.2.1. Production from Phthalic Anhydride and Ammonia

The continuous processes for producing phthalimide are particularly important industrially. One of the continuous processes is carried out in an externally heated, vertical reaction tube filled with packing material [111]. A gas-tight connection links the bottom of the reaction tube to a sublimation chamber from which a discharge arm discharges the phthalimide through a slot by means of a screw conveyor. The exhaust gases are discharged from the sublimation chamber via built-in baffles.

Molten phthalic anhydride and excess ammonia are continuously fed into the reaction tube at the top and are reacted at 250–280 °C. The reaction gases are cooled to 170–180 °C in the sublimation chamber, and solid phthalimide is deposited and discharged. Water and excess ammonia are removed via an exhaust-gas pipe. A 98 % yield of phthalimide is obtained with a purity of 99 %.

Another continuous process for producing phthalimide from phthalic anhydride is a countercurrent process [112]. Molten phthalic anhydride is continuously fed to the head of a reactor while ammonia is continuously fed to the bottom. The temperature increases from $\geq$ 150 °C at the top of the reactor to a maximum of 270 °C at the bottom. The molten phthalimide emerging from the bottom of the reactor with a purity of 99 % is cooled and flaked or dissolved in aqueous alkali, and supplied as a starting material for further syntheses, e.g., to produce anthranilic acid.

The exhaust gas from the head of the reactor, which is composed of water vapor, sublimed phthalimide and phthalic anhydride, and ammonia, is fed to the bottom of a scrubbing column where it is scrubbed in countercurrent with molten phthalimide from the bottom of the reactor. The melt taken from the bottom of the scrubbing column, which contains unreacted phthalic anhydride, is returned to the reactor, while the exhaust gas from the scrubbing column, which contains predominantly water vapor and phthalimide, is dissolved in aqueous alkali, and the solution is used to produce anthranilic acid.

3.2.2. Production from Phthalic Anhydride and Urea

Solvent-Free Process. A mixture of phthalic anhydride and urea is introduced into a heatable, tiltable reaction vessel. Mounted on the cover of the vessel is an extraction tube for the reaction gases, CO_2 and water vapor. The reactor is heated to 130–140 °C until the contents are molten, and as soon as the elimination of CO_2 and water starts, heating is ceased. The heat of reaction causes the temperature to rise to ca. 160 °C. The reaction is complete when the contents of the reactor have swelled to several times their original volume as a result of foaming and have solidified.

The cover of the reactor is then removed. The pure phthalimide is discharged into a collecting vessel by tilting the reactor, is cooled and ground, and used without further treatment. The yield is > 90 % [113].

Production in a Solvent. Phthalic acid or phthalic anhydride can be reacted with urea in the presence of a solvent to form phthalimide. Suitable solvents are substituted and unsubstituted hydrocarbons and aromatics and heteroaromatics such as *n*-propylbenzene, cumene, 1,2-dichlorbenzene, and picoline. Urea must be insoluble in the reaction medium, while phthalic acid or phthalic anhydride must be only slightly soluble. The reaction is carried out below the boiling point of the solvent, normally at 160–170 °C. When the reaction is complete, the pure phthalimide is filtered off and washed with water. The yield is 95–100 % [114].

3.2.3. Production from *o*-Xylene

o-Xylene is reacted in the gas phase with ammonia in the presence of a metal-oxide catalyst as oxygen donor. Phthalimide, phthalamide, or phthalonitrile can be selectively produced, depending on how the reaction is controlled [115].

3.3. Uses

Phthalimide is of industrial importance as the starting material for producing anthranilic acid by Hofmann degradation, and a large number of primary amines can be produced by the Gabriel synthesis. Phthalimide is an intermediate in the production of agricultural pesticides and wood preservatives, pigments, and pharmaceuticals.

4. Phthalonitrile

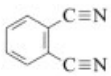

4.1. Properties

Phthalonitrile [*91-15-6*], 1,2-dicyanobenzene, 1,2-benzenedicarbonitrile, $C_8H_4N_2$, M_r 128.14, is a crystalline powder having a faint grayish yellow color and a slighty aromatic odor, similar to benzonitrile. Phthalonitrile was first described in 1896 when it was isolated during the diazotization of 2-aminobenzonitrile.

The compound is sparingly soluble in water (ca. 1 g/L) and soluble in acetone, nitrobenzene, and benzonitrile. It cannot be distilled and polymerizes if heated above the melting point. It is not explosive and is difficult to ignite, but the dust can explode. Some physical properties of phthalonitrile are as follows:

Melting point	141 °C
Heat of combustion	4013 kJ/mol
Heat of evaporation	67 kJ/mol
Specific heat at 30 °C	1.30 J g^{-1} K^{-1}
Vapor pressure at 20 °C	0.05 mbar
Density	1.238 g/cm^3
Apparent density	ca. 0.5 g/cm^3
Flash point	162 °C

4.2. Production

Phthalonitrile is produced commercially from *o*-xylene, phthalic acid, phthalic anhydride, phthalamide (**1**), or phthalimide (**2**).

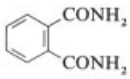

1

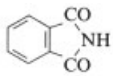

2

4.2.1. Production from *o*-Xylene

Ammoxidation. In a single-stage continuous process, *o*-xylene is converted to phthalonitrile by reaction with ammonia and oxygen in the gas phase in a fluidized-bed reactor. Generally, metal oxide mixtures containing vanadium, antimony, chromium, and molybdenum, with further active components such as iron, tungsten, and alkali-metal oxides, on an alumina or silica support are used as catalysts [116 – 123].

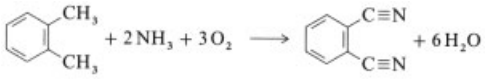

The course of the reaction and the industrial importance of ammoxidation have been described in depth in [124 – 126]. The commercial process of producing phthalonitrile from *o*-xylene by oxidative ammonolysis is as follows [127, 128]: A gaseous mixture of *o*-xylene, NH_3, and O_2 is passed through a distributor plate into a fluidized-bed reactor. Either a vanadium oxide–antimony oxide catalyst or a vanadium oxide–chromium oxide catalyst on an alumina support is used. The optimum temperature for pressureless ammoxidation is 480 °C. Temperatures below 480 °C lead to the formation of up to 10 % of the undesirable byproducts phthalamide and phthalimide. Above 500 °C, ammonia begins to burn. The cooling elements incorporated in the fluidized bed make it possible to keep the temperature constant despite the large heat of reaction. The hot reaction gases from the reactor are quenched in a product settler with an aqueous suspension of phthalonitrile. After settling and cooling, the phthalonitrile is separated in decanters and dried.

The phthalonitrile produced in this way, which contains < 0.1 % of acid and phthalimide and < 0.1 % water, can be processed directly without further purification to yield phthalocyanine pigments. The yield is 80–85 %.

Ammonia is removed from the gas mixture emerging from the product settler (the latter contains NH_3, CO_2, CO, N_2, and traces of HCN) in an ammonia recovery plant, and the exhaust gas is burnt. The ammonia is returned to the process. Unreacted *o*-xylene and the intermediate *o*-toluonitrile can be worked up to isolate *o*-toluonitrile or can be fed back into the fluidized-bed reactor for complete conversion to phthalonitrile.

Oxidative Ammonolysis. The continuous production of phthalonitrile from *o*-xylene by oxidative ammonolysis has been studied in a pilot plant [115, 129]. In this process *o*-xylene is reacted with ammonia in the presence of a metal-oxide catalyst in the gas phase at 350–450 °C using a vanadium oxide–molybdenum oxide catalyst on an alumina or silica support. The oxidation medium is the metal-oxide catalyst, part of which is continuously removed, reoxidized, and then fed back into the reactor. The phthalonitrile is quenched from the hot reaction gases in the same way as in the ammoxidation process, while byproducts such as *o*-toluonitrile, phthalamide, and phthalimide are fed back into the phthalonitrile reactor or can be converted into phthalonitrile in a second reactor at 400 °C in the presence of NH_3 on a boron phosphate–alumina catalyst.

Ammonia is recovered from the residual gas mixture, which contains NH_3, CO_2, CO, N_2, and traces of HCN, and fed back into the process. Noxious gases such as CO and HCN are burnt in the regenerator and N_2 and CO_2 are eliminated from the process.

4.2.2. Production from Phthalic Acid Derivatives

Phthalonitrile can be produced from phthalic acid, phthalic anhydride, phthalamide, or phthalimide by reaction with ammonia and elimination of water at 300–500 °C in the gas phase in the presence of a catalyst.

Catalysts mentioned in the patent literature are the oxides of thorium, copper, beryllium, zirconium, or tungsten on a silica, alumina, or phosphate support, or on silicates, borates, and basic alumina.

Molten phthalic anhydride heated to ca. 160 °C is evaporated in a heating apparatus into which preheated circulating gas containing NH_3 in a concentration of ca. 90 % is also fed. The gas mixture is reacted under virtually unpressurized conditions at 350–400 °C in a fixed-bed reactor downstream of the heating apparatus using an Al_2O_3 catalyst. The reaction gas emerging from the reactor is quenched with water, and the phthalonitrile is separated from the aqueous suspension in decanters and dried. The concentration of NH_3 in the circulating gas is kept at ca. 90 % by adding NH_3.

Numerous other production methods have been described [130, 131], but have no commercial significance. For example, phthalonitrile can be prepared by eliminating water from phthalamide in the presence of acid halides such as phosgene or thionyl chloride. To prevent hydrolysis of the phthalamide by the acid formed, diluents such as benzene or chlorobenzene are added

or the acids are bound by tertiary amines such as *N,N*-diethyl-*o*-toluidine or pyridine. It is also possible to use acylated secondary amines such as *N*-methylformamide.

4.3. Uses

Phthalonitrile is used as a starting material for producing phthalocyanine pigments (→ Phthalocyanines), fluorescent brightners, and photographic sensitizers.

5. Phthalates

A wide variety of dialkyl phthalates are produced and marketed (Table 3).

In the early 1900s, the availability of monohydric alcohols was limited to those with a chain length of up to four carbon atoms. This resulted in the development and manufacture of four esters; e.g., diethyl phthalate (DEP) was used as a heat-transfer oil, and dibutyl phthalate (DBP) was used to reduce the hygroscopicity of explosives. Both DEP and DBP were also used as carriers in the perfume industry [132].

Also at this time camphor was being used as a plasticizer for cellulose nitrate and cellulose acetate. In the early 1920s it was found that phthalates could replace the expensive plasticizer camphor. This increased the demand for phthalates [132]. In 1937, the "Chemiker Taschenbuch" [133] listed seven phthalates for use as plasticizers. The late 1920s saw the start of the development of poly(vinyl chloride), PVC. PVC is rigid and brittle at ambient temperature. However, when plasticized it was found to be suitable for replacing natural rubber [132]. In 1929 a U.S. patent for plasticized PVC was granted to Ostromislenski [134], and Kyrides applied for U.S. patent for the use of di-2-ethylhexyl phthalate (DEHP) as plasticizer for PVC, which was granted in 1933 [135].

The most important phthalates are DEHP, the standard general purpose plasticizer for PVC, followed by DINP and DIDP [136]. In trade, DEHP is also referred to as DOP. Besides the phthalates listed in Table 3, phthalate blends and coesters are used for special applications [137].

5.1. Physical and Chemical Properties

All phthalates listed in Table 3 are clear, oily liquids at room temperature. They are soluble in common organic solvents and are miscible with other PVC plasticizers. When added to plastics and resins, they improve the workability during fabrication, modify the properties, or give rise to new improved properties not exhibited by the original material [138].

Some of the important properties of plasticizers are listed below. These properties are influenced by the manufacturing process and can vary significantly. The values given are typical ranges in which commercial products are available. The analytical methods are listed in Table 4.

APHA color	5–60
Acid number, mg KOH/g	0.02–0.1
Water content, wt %	0.05–0.1
Purity, area % or wt %	99.0–99.5
Odor	odorless or low

5.2. Raw Materials

Phthalic anhydride (PA) and monohydric alcohols are generally the raw materials for phthalates of commercial significance. Phthalic anhydride is mainly used in the molten form. The alcohols are based on the C_2–C_4 olefins, which are produced in either the steam cracking process or in petrochemical plants. These alcohols can be produced by (1) oligomerization or Ziegler synthesis followed by hydroformylation and hydrogenation, or (2) by aldol condensation with hydrogenation (e.g., the production of 2-ethylhexanol from propylene). The linear or slightly branched C_6 to C_{11} alcohols are based on ethylene. Isononanols are produced from the C_4 fraction of the steam cracker output or from C_8–C_9 olefins from a poly gas unit (olefin oligomerization) [133].

5.3. Production

The formation of phthalates is as follows:

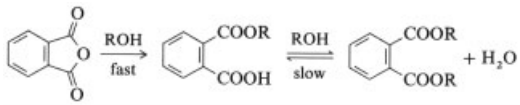

Table 3. Commercially important phthalates

Chemical denomination	Code 1* DIN 7723	CAS registry no.	M_r	Viscosity at 20 °C, mPa · s	Density at 20 °C, g/m^3	Refractive index at 20 °C
Dimethyl phthalate	DMP	[*131-11-3*]	194.19	18	1.191	1.516
Diethyl phthalate	DEP	[*84-66-2*]	222.24	13	1.117	1.502
Dibutyl phthalate	DBP	[*84-74-2*]	278.35	21	1.047	1.493
Diisobutyl phthalate	DIBP	[*84-69-5*]	278.35	41	1.039	1.490
Butyl benzyl phthalate	BBP	[*85-68-7*]	312.36	59	1.121	1.538
Diisopentyl phthalate	DIPP	[*84777-06-0*]	306.40	30	1.024	1.490
Diheptyl phthalate	DHP	[*68515-50-4*]	362.51	49	0.993	1.487
Di-2-ethylhexyl phthalate	DOP(DEHP)	[*117-81-7*]	390.56	80	0.983	1.487
Diisooctyl phthalate	DIOP	[*27554-26-3*]	390.56	75–84	0.984	1.488
Di-*n*-octyl phthalate	DNOP	[*117-84-0*]	390.56	39	0.980	1.485
Di(hexyl-octyl-decyl) phthalate	HXODP(610P)	[*68515-51-5*]		44	0.974	1.484
Di(octyl-nonyl-decyl) phthalate	ONDP(810P)	[*71662-46-9*]		49	0.969	1.483
Di(heptyl-nonyl) phthalate	HNP(79P)	[*68515-41-3*]		44	0.986	1.486
Di(heptyl-nonyl-undecyl) phthalate	HNUP(711P)	[*68515-42-4*]		50	0.972	1.484
Di(nonyl-decyl-undecyl) phthalate	NDUP(911P)	[*68515-43-5*]		73	0.962	1.484
Diisononyl phthalate	DINP **	[*28553-12-0*]	418.62	80	0.975	1.486
Diisononyl phthalate	DINP **	[*68515-48-0*]	418.62	105	0.974	1.486
Di(3,5,5-trimethylhexyl) phthalate	DINP **	[*14103-61-8*]	418.62	110	0.968	1.486
Diisononyl phthalate	DINP **	[*28553-12-0*]	418.62	165	0.978	1.488
Diisodecyl phthalate	DIDP	[*26761-40-0*]	446.67	130	0.966	1.486
Diundecyl phthalate	DUP	[*3648-20-2*]	474.72	70	0.955	1.482
Diisoundecyl phthalate	DIUP	[*85507-79-5*]	474.72	200	0.964	1.487
Diisotridecyl phthalate	DTDP	[*68515-47-9*]	530.83	310	0.948	1.482
Di(methoxyethyl) phthalate	(DMEP)	[*117-82-8*]	282.29	55	1.170	1.502
Di(butoxyethyl) phthalate	(DBEP)	[*117-83-9*]	366.45	42	1.061	1.486

* Codes in parentheses are used in English.
** DINP types are based on different branched isononyl alcohols.

The first step, alcoholysis of PA to give the monoester, is rapid and goes to completion. The reaction generally starts at elevated temperatures and proceeds exothermically.

The second step is the conversion of the monoester to a diester with the formation of water. This is a reversible reaction and proceeds more slowly than the first, thus determining the overall rate of reaction. To shift the equilibrium towards the diester, the water of reaction is removed by distillation. The rate of reaction can be influenced by the choice of catalyst and the reaction temperature. For fast conversion rates, high reaction temperatures are generally used. However, these are influenced by the boiling point of the alcohol and/or the type of catalyst.

Sulfuric acid [132], methanesulfonic acid [139, 140] and *p*-toluenesulfonic acid [132] are effective esterification catalysts [140]. Sodium hydrogensulfate [132] or the monoester itself [132, 141] can also be used. Acid catalysts can be used for esterification reactions at temperatures up to 140–165 °C [132, 141]. Their use above these temperatures results in undesirable side reactions such as alcohol degradation with the formation of olefins, ethers, and colored products [141]. Acid catalysts are the preferred catalysts for the manufacture of phthalates from low-boiling alcohols up to C_4. Acid catalysts have also been used for the production of phthalates from higher alcohols in the presence of a water entrainer, e.g., aromatics or chlorinated hydrocarbons. Alternatively, the esterification can be carried out at reduced pressure, which allows the distillation of a water–alcohol azeotrope at low temperature; however, longer reaction times are required.

The autocatalytic esterification [132] can be carried out at temperatures above 200 °C, but the conversion to diester is not complete, and the recovery of unreacted monoester is necessary.

Currently, nearly all major phthalate producers use amphoteric catalysts for the esterification of high boiling alcohols. The first claim for the use of titanates as esterification catalysts was made in 1955 in the United States [142]. Since then titanates and tin(II) compounds have

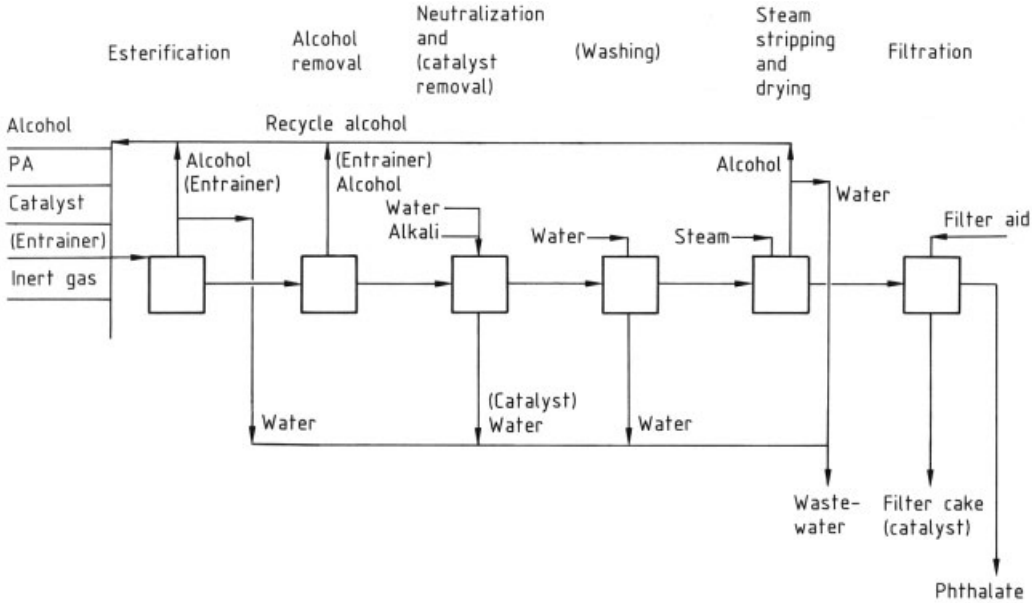

Figure 3. Simplified flow sheet for the production of phthalates. Terms in parentheses are optional.

been widely used as esterification catalysts [140, 143 – 145].

The reaction temperatures for the amphoteric catalysts are about 200 °C [143]. At this temperature side reactions are minimized, and the alcohol can be recycled without purification. By using this type of catalyst, over 99.5 % conversion to diester can be achieved [143 – 145].

Phthalates can be produced continuously [132, 141] or batchwise. The process steps outlined in Figure 3 are applicable to both methods. Variations are possible depending on the catalyst used, the reactant alcohol, and the process type.

For batch processes the reactor is usually a stirred vessel made of stainless steel, which can be rapidly heated to maintain a high distillation rate during diester formation for removing the water of reaction.

For the esterification of the water-soluble C_1–C_4 alcohols, the overhead vapors of the reactor are fed into a rectification column where the evaporated alcohol is dried before it is recycled to the reactor. If higher alcohols of very low water solubility are used, the vapors from the reactor can be fed directly into a condenser which drains into a alcohol–water separator. The water is separated here by settling and the alcohol is recycled to the reactor.

Usually a 20–25 % excess of alcohol is used [140]. The catalyst can be added with the alcohol. Amphoteric catalysts are preferably added after monoester formation has taken place at 160 °C [140]. PA can be charged as flakes or molten. The use of molten PA reduces the cycle time of a batch.

After charging the raw materials heating is continued, e.g., for DOP to a maximum of 224 °C [145]. In order to remove the water of reaction, a good boil up rate is maintained. This can be achieved by adjusting the pressure with a vacuum source. The water of reaction can also be stripped off by an inert gas sparge. When the required acid number is achieved, the excess alcohol is removed under reduced pressure [145].

The purification procedure required after the esterification is determined by the catalyst type. For acid catalysts, neutralization with aqueous caustic soda is necessary. However, traces of alkali remain in the organic phase, and therefore a water wash after the neutralization step is advisable.

Titanates are generally removed by alkaline hydrolysis. The product is cooled to ca. 100 °C [140], and lime or soda ash or sodium hydroxide are added with water. The mixture is stirred and

the water is then removed by vacuum distillation.

After neutralization, residual alcohol is removed by steam distillation [140]. The temperature, pressure, and steam for injection are adjusted so that the product after this treatment is nearly alcohol free and dry. The final filtration is performed with a rotary vacuum filter or plate and frame filters. If the hydrolyzed catalyst was not removed with the water phase in the neutralization step, the catalyst can be removed in the filtration step by adding a filter aid.

The continuous BASF process described in [132, 141] can be carried out autocatalytically or preferably by using an amphoteric catalyst. The esterification is carried out in a multistage cascade of agitated vessels. The reaction mixture leaving the last reactor is pumped to a vacuum flashing chamber to recover the excess alcohol [132]. To remove the catalyst, the crude ester is treated with a mild aqueous alkali solution in a mixer/settler system. To remove residual volatiles such as water, alcohol, and other low-boiling compounds a steam distillation is carried out under reduced pressure. For the final treatment a filter aid is added. The product is subjected first to coarse and then to final filtration.

5.4. Environmental Protection

The wastewater from phthalate processes is contaminated with phthalates, alcohols, and other organic compounds. Therefore, the process wastewater must be treated before discharging to a municipal system or a water body. Local regulations dictate the treatment and disposal of this water. Generally, biological treatment is required. Studies on phthalates have shown that they can be biodegraded by many species of bacteria [146 – 148]. In Germany only three phtha-lates (diethyl phthalate, diallyl phthalate, benzyl butyl phthalate) are listed in the catalog for water-endangering materials [149]. They are classified as materials with moderate hazard potential (WGK 2).

In the United States and Canada there are no current regulations limiting the concentration of phthalates in wastewater.

In the TA Luft [150], only DOP is mentioned as material of class II (intermediate danger potential). The other phthalates have been classified according to the hazard potential for the environment by the rules given in [150]. Exxon [146] lists the C_4–C_{11} phthalates, with the exception of DOP, in class III. Because of their low vapor pressure, the phthalates easily meet the concentration limits for phthalates specified by TA Luft.

Owing to the high working temperatures, air emissions from compounding and extrusion processes for flexible PVC generally have high phthalate concentrations. Therefore emission controls may be required for these discharges. Exxon [146] gives the following data for the environmetal levels of DBP and DEHP:

- Fresh and estuarine/sea waters: 10 and 0.7 ppb, respectively
- Sediments: a few ppb up to 100 ppm
- Air: 3 ng/m^3 over the sea and up to 130 ng/m^3 in cities.

Detailed information on DBP and DEHP distribution in the environment are given in [147, 148].

In North America, waste containing DEHP and DBP is regarded as hazardous waste and must be disposed off in a secured landfill site.

5.5. Quality Specifications

For manufacturing flexible plastics the following properties are important for the selection of a plasticizer [134]: (1) compatibility with plastics, (2) solvating temperature (a measure of the solvating capability on plastics), (3) efficiency for plasticizing, (4) volatility in plastics, (5) extractability and diffusion losses in plastics, and (6) electrical properties. The properties required in the final products determine the type and the quality specification of the plasticizer.

For manufacturing flexible PVC for blood bags or other goods in medical application the plasticizer used (e.g., DEHP) should have a purity of min. 99.5 %, a low alcohol content, a low acid number (max. 0.02 mg KOH/g), and no odor [151].

Flexible vinyl sheets used for automotive interior trim should be plasticized with a phthalate that ensures low fogging of the interior windshield. Also low-temperature flexibility, light and thermal stability, and processability

are required. BASF [152] recommends a low-branched DINP or a HNUP for this application.

Phthalates used for plasticizing PVC for wire, cable, and conductor sheathing need high electrical volume resistivity. Products with volume resistivities $> 1.0\times10^{11}\ \Omega\cdot$cm at 20 °C are required. Typical resistivity of DOP is 1.0×10^{11}, and of DIDP $1.0\times10^{12}\ \Omega\cdot$cm.

Quality tests are carried out both on the phthalate itself and on the plasticized polymers. Table 4 lists standards for testing phthalates as PVC plasticizers.

Nowadays, almost all demands of plasticizers for use in flexible plastics can be met by phthalates.

Table 4. Standards for testing of phthalates

Property	Methods	
	DIN	ASTM
Phthalates		
Viscosity at 20 °C	51 561	D 445
Density at 20 °C	51 757	D 1045
Refractive index at 20 °C	51 423	D 1218–61
APHA color	ISO 6271	D 1209
Acid value	53 402	D 1045
Pour point	ISO 3016	D 97–57
Water content	51 777	D 1364
Purity	gas chromatography	
Plasticized PVC		
Solvating temperature	53 408	
Tensile strength	53 455	D 412
Elongation at break	53 455	D 412
Brittleness temperature	53 372	D 746
Clash & Berg torsion stiffness at 310 N/mm^2	53 447	D 1043
Shore B hardness	53 505	D 2240
Volume resistivity at 25 °C	53 482	D 1169

5.6. Storage and Transportation

Because phthalates are neutral liquids they can be stored in carbon steel tanks. The more expensive stainless steel or aluminum alloy tanks are also used. All phthalates have a flash point above 100 °C. Thus bulk storage of phthalates does not require any special fire protection equipment [153]. German regulations [154] classify phthalates as having an endangering potential to water. Therefore storage tanks have to be installed in a diked area which has a containment capacity of the largest tank. Detailed proposals for storing phthalates are given in [155]. Phthalates can be stored indefinitely and can withstand tropical conditions.

Phthalates are transported in tank cars, tank trucks, ships, or mild steel drums. Canada and California have special labeling requirements for DBP and DEHP.

5.7. Uses

The C_1–C_4 phthalates including DMEP and DBEP are mainly used as plasticizers for cellulosic resins and some vinyl ester resins. DBP and BBP are fast-fusing plasticizers for PVC and are mostly used in combination with DEHP. C_4 phthalates are also appropriate plasticizers for nitrocellulose lacquers.

DEHP, DINP, DIDP are the general purpose plasticizers for PVC in most applications (see also → Plasticizers). For wire and cable, DIDP is preferred. The linear phthalates, e.g., 711P or 911P, are used for PVC in various applications. They are less volatile and confer better low-temperature properties than general-purpose phthalates. DUP and DTDP are recommended for use in wire insulation for use at ca. 100 °C.

87 % of the phthalates produced are used for formulating flexible PVC, which is consumed for manufacturing the following goods [156]:

Wire and cable	25 %
Film and sheeting	23 %
Flooring	15 %
Profiles and tubing	10 %
Plastisol spread coatings	11 %
Other plastisols	8 %
Miscellaneous (shoe soles, blood bags, gloves)	8 %

5.8. Economic Aspects

The major phthalate producers in Europe, the United States, and Japan are listed in Table 5. The capacities of these companies range from 70 000 to 350 000 t/a.

In 1988 world consumption of plasticizers, including nonphthalates, was 3.85×10^6 t [157]. Based on these figures and uses in other applications, world consumption for phthalates is estimated at 3.25×10^6 t, of which DEHP accounts for ca. 2.1×10^6 t. The estimated total consumption of all phthalates and consumption of DEHP (in 10^3 t) by geographic region is:

	Total	DEHP
Western Europe	900	465
North America	730	155
Eastern Asia	530	490
Japan	320	245
Others	720	765
World total	3250	2120

Prognoses for the world market show an annual increasing demand of ca. 2.5 %.

Table 5. Major phthalate producers and their trade names

Company	Location	Trade name
Aristech Chemical Corp.	United States	PX
ATOCHEM	Europe	Garbeflex
BASF	Europe, United States	Palatinol
BP Chemicals	Europe	Bisoflex
Chemische Werke Hüls	Europe, United States	Vestinol, Nuoplaz
Eastman Chemical	United States	Kodaflex
Enichimica	Europe	Sicol
Exxon Chemical	Europe, United States	Jayflex
Hoechst	Europe	Genomoll
ICI	Europe	Hexaplas
Kyowa Hakko	Japan	Kyowacizer
Mitsubishi Kasei Vinyl	Japan	
Neste Chemicals	Europe	
New Japan Chem.	Japan	

6. Toxicology

6.1. Use of and Exposure to Phthalic Acid and Derivatives

Phthalate esters, phthalic acid, phthalic anhydride, phthalimide, and phthalonitrile are intermediates that are produced and further used in industry under well-controlled conditions. Occupational exposure may occur but is expected to be low. Exposure of the general public is low and negligible. However, the case of phthalate esters is different because these are generally used as plasticizers, not only for a wide range of products for industrial use such as resins, paints, electrical cable, and equipment lining, but also for numerous consumer products like PVC floor tiles, shoes, rubber boots, gloves, and household plastic materials including food containers. Plasticizers are also contained in toys and in medical equipment such as blood bags, infusion lines, and devices for hemodialysis and extracorporeal membrane oxygenation (ECMO). Thus, both occupational exposure of workers during production and the exposure of the general public may occur. The highest levels of exposure are expected for children and patients undergoing intensive medical treatment. This promoted intense investigation of the exposure to and the mode of action of phthalate esters. The most prominent — diethylhexyl phthalate (DEHP) — may be regarded as a model phthalate ester and is therefore described below in more detail.

6.2. Toxicological Profiles

6.2.1. Phthalic Acid

Only few investigations exist; however, the toxicity appears to be low. Acute toxicity (LD_{50}, mouse, i.p.): 550 mg/kg [158].

Metabolism. Most of the material is excreted in the urine, either directly (as in dogs) or partially conjugated (as in rats and rabbits). A minor part may be decarboxylated and excreted as benzoic acid [159].

6.2.2. Phthalic Anhydride

Phthalic anhydride is an inhalation allergen and may cause occupational asthma in exposed persons [160 – 164]. Hexahydrophthalic acid showed also respiratory sensitizing properties [163].

Acute toxicity data indicate minor toxicity. Acute toxicity: LD_{50}, rat, oral: 1500–4000 mg/kg [165]; LC_{50}, rat inhalation: > 210 mg m^{-3} h^{-1} (dust) [166]; LD_{50}, rabbit, dermal: $>$ 10 000 mg/kg [166]. Dermal irritation: nonirritating or slightly irritating in rabbit skin [165]. Eye irritation (rabbit): irritating [165, 166]. Dermal sensitization: sensitizing in the mouse ear swelling test [162].

Subacute Toxicity. Feeding experiments with 250, 1000, and 3800 ppm in the diet (17, 67, and 253 mg kg^{-1} d^{-1}) were tolerated for 28 d without significant effects [166]. Gavage experiments in female rats, beginning with 20 mg kg^{-1} d^{-1} and subsequently increasing doses up to 4800 mg kg^{-1} d^{-1} over a period of 8 weeks, caused toxic effects on the kidneys and the gastric mucosa at doses exceeding 1200 mg kg^{-1}

d^{-1} [167]. Inhalation experiments with sublimed material in rats, mice, rabbits, and cats showed some irritation and impaired respiration at 10 000 mg/m^3 for 4 h per day for two weeks [165]. In female guinea pigs, vapor concentrations of 615 mg/m^3 for 30 min per day over 4 and 8 d caused irritative effects on the eyes and respiratory tract [167]. Repeated inhalation of pure phthalic anhydride dust in concentrations of 8.5 mg/m^3 for 3 h per day for several periods of 4 consecutive days, alternating with 10-d recovery periods for a total experimental period of 8 months, caused irritation of the respiratory tract and frequent pneumonias [167].

Genotoxicity. No mutagenic effects were detected in the Ames test. No sister chromatid exchanges or chromosome aberrations were found in CHO cells. All experiments were conducted with and without metabolic activation [168, 169].

Chronic Toxicity and Carcinogenicity. Long-term bioassays with 7500 and 15 000 ppm in the diet of rats (500 and 1000 mg kg^{-1} d^{-1}) did not lead to an increased tumor incidence compared to control animals. Similarly, mice dosed with 3500 and 7100 mg kg^{-1} d^{-1} for 32 weeks and with 900 and 1800 (females) or 1800 and 3600 (males) mg kg^{-1} d^{-1} for a further 72 weeks did not respond with an increase of tumors [170, 171].

Reproductive Toxicity. Only screening experiments with intraperitoneal injection in CD-1 mice have been reported. 80 mg/kg injections were carried out from gestation days 8 to 10 or 11 to 13. Malformations (cleft palates, malformations of ribs and vertebrae) were detected at doses $\geq$ 55.5 mg kg^{-1} d^{-1} [172 – 174]. The experiments are inconclusive for the quantitative assessment of a teratogenic risk, since the exposure route was artificial and mice are prone to respond to experimental stress with these types of malformations.

6.2.3. Phthalimide

Few data exist; however, a low toxicity may be assumed. The material is hydrolyzed to phthalic acid and ammonia [175].

Acute toxicity. LD_{50}, (rat, oral): $>$ 5000 mg/kg (Bayer AG (1978), cited in [176]); LD_{50}, (mouse, oral): 5000 mg/kg [177]. The material was not irritating to the rabbit skin and eye [176]. No data on sensitizing properties are known to exist.

Repeated-Dose Toxicity. There were no substance-related effects in male and female rats after inhalation of the material at 0.051, 0.154, and 0.523 mg/L (4 weeks, 6 h/d, 5 d/week) regarding mortality, body weight gain, clinical chemistry including hematology and urinalysis, or gross organ changes (IBT (1979); cited in [176]). No long-term studies are known to exist.

Genotoxicity. Phthalimide was not genotoxic in an Ames test performed according to the OECD test guideline using *S. typhimurium* strains TA98, TA100, TA1535, and TA1537 at concentrations of 8–5000 μg per plate, with and without metabolic activation (Bayer AG (1993); cited in [176]).

Reproductive Toxicity. No reliable studies are reported in [176].

6.2.4. Phthalonitrile

Acute Toxicity. LD_{50}; (rat, oral): 30–125 mg/kg [178, 179]. No irritation of rabbit skin (BASF AG (1969); cited in [178, 179] or eye (BASF AG (1993); cited in [178]) was noted. Medical records of workers involved in the production and use of phthalonitrile do not indicate sensitizing properties (BASF AG (1995); cited in [262]).

Repeated-Dose Toxicity. Neurotoxicity was noted in rats and mice following repeated dosing in the form of convulsions and increased excitation [178]. Increased activity and body weight decreases were seen in a 13-week rat neurotoxicity study at 10 and 25 mg kg^{-1} d^{-1}. There was no histopathological correlation to the behavioral changes. The NOAEL was 3 mg kg^{-1} d^{-1} in this study (Bio Research (1994), cited in [178]). Epileptiformic convulsions lasting minutes were observed in workers after occupational exposure with a latency period of several hours to days [180]. Clinical examinations,

clinical chemistry, hematological parameters, neurological studies, and EEG studies detected no abnormal findings in 81 workers who had been exposed for an average of 8.5 years, 11 of whom had suffered from acute intoxications (Kleinsorge et al. (1979); cited in [178]). Also dermal exposure may be relevant [179]. Since the material does not appear to be metabolized to cyanide, the typical measures taken against cyanide intoxications are not promising.

Genotoxicity. The material does not cause mutagenic effects in the Ames test [181] and the CHO/ HGPRT gene mutation assay [182]. The substance was negative in the mouse micronucleus test in vivo (LMP (1987); cited in [178]). Chromosomal aberrations were not increased in 20 workers, seven of whom had experienced acute intoxications (Fleig and Thiess (1979); cited in [178]).

Carcinogenicity. After oral and subcutaneous administration to rats and mice, and after percutaneous administration to mice, induction of leukemia was reported. The study description is, however, vague and does not allow a conclusive assessment [183].

6.2.5. Phthalate Esters

Diethylhexyl phthalate has been most extensively investigated, due to the large production volume and widespread use. The following sections on the general toxicological profile of phthalate esters therefore focus on DEHP as the lead substance. However, significant differences between groups of phthalate esters of varying chain lengths need consideration. Generally, according to the length of the linear alkyl backbone of the alcohol moiety, groups of low (C_1–C_3), intermediate (C_4–C_6), and high molecular weight (C_7–C_{13}) phthalate esters are distinguished.

6.2.5.1. Metabolism and Toxicokinetics

Phthalate esters in general are rapidly hydrolyzed to the corresponding monoesters and the respective alcohol in the intestine [184]. These are readily absorbed from the gut [185], with some species difference in absorption and excretion. Short-chain phthalates such as dimethyl phthalate are either excreted unchanged in the urine or completely hydrolyzed to phthalic acid and excreted [186].

The pattern of DEHP metabolites in the urine has been elucidated [187, 188]. MEHP appears in native or — depending on the species — in glucuronidated form (at the carboxylate moiety) in the urine along with degradation products from ω-and ω-1-oxidation in the alcohol side chain. ω-Oxidation leads to a carboxylic moiety at the C-terminal site of the alcohol chain. ω-1-Oxidation leads to the 5-hydroxy and 5-keto derivatives of MEHP [188]. The latter was shown to be the metabolite ultimately responsible for peroxisome proliferation [189]. The ω-1-oxidative pathway is more pronounced in the rat than in the cynomolgus monkey [190].

The metabolism of DEHP is shown in Figure 4.

The liberated primary alcohols may be oxidized to the respective aldehyde and carboxylic acid via enzymatic reactions catalyzed by alcohol and aldehyde dehydrogenases. The fatty acids are further degraded in the intermediary metabolism, i.e., the fatty acid and citrate cycles. This pathway holds true for linear primary alcohols and for primary alcohols which are methyl-branched at even-numbered positions, whereas branched primary alcohols with a methyl substituent at an odd-numbered position and 2-ethyl-branched alcohols do not undergo β-oxidation. Similarly, secondary and tertiary alcohols are not metabolized to the fatty acid and aldehyde, respectively. When β-oxidation does not allow rapid metabolism, alternative reactions (chain hydroxylation and conjugation) occur prior to urinary excretion [191].

Di-*n*-octyl phthalate was similarly metabolized, i.e., the monoester was formed and side-chain oxidations occurred. Phthalic acid was also detected in minor quantities [192].

Toxicokinetic studies using radiolabeled DEHP by gavage in pregnant and nonpregnant female Wistar rats and CD-1 mice appeared in unpublished reports sponsored by the European Council for Plasticizers and Intermediates [193 – 197]. Determinations were made after single oral doses of 200 or 1000 mg/kg b.w. and after five daily doses at these levels. Results indicated rapid absorption, linearity of peak ra-

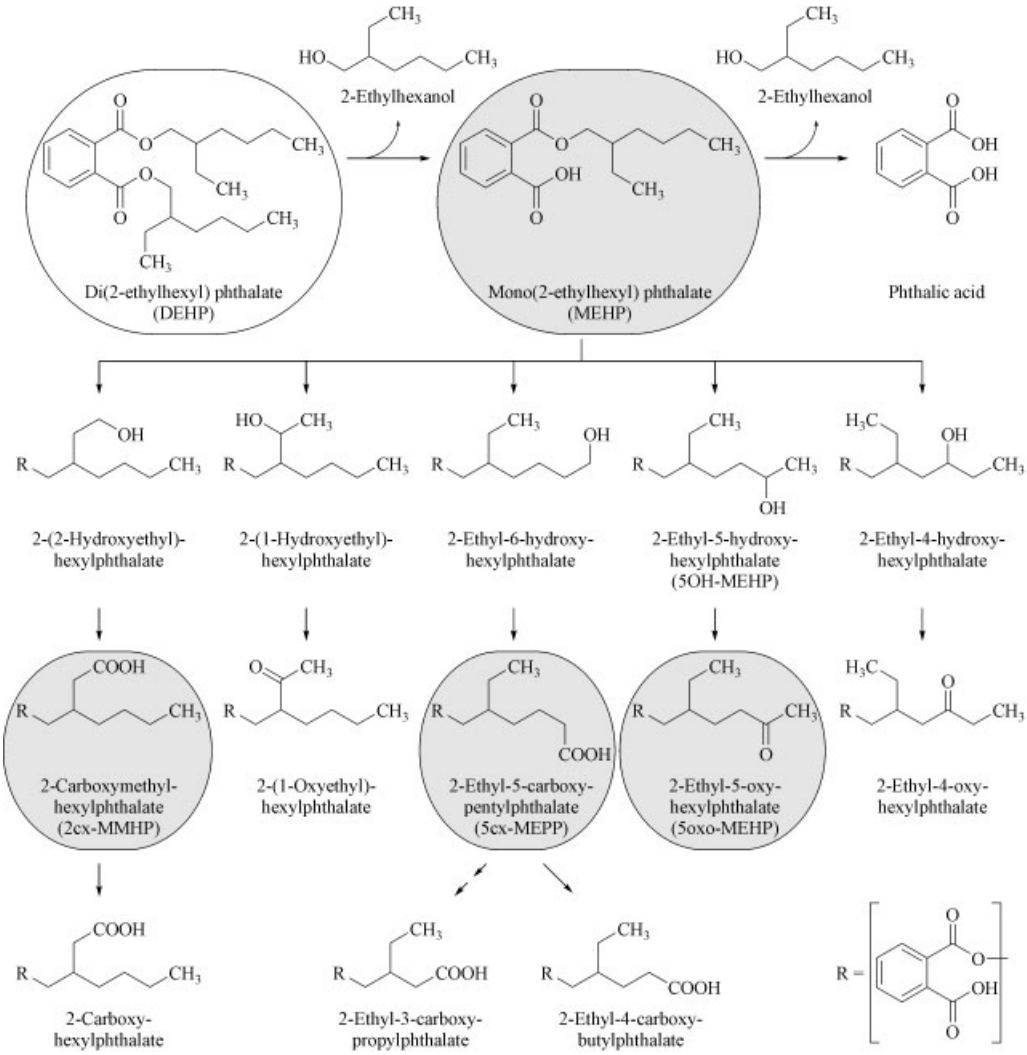

Figure 4. DEHP metabolism. The key metabolites are circled. (From Koch et al. [188], used with kind permission of Springer Science and Business Media and Prof. Dr. J. Angerer.)

diolabel plasma levels, and the area under the curve (AUC) with dose in both species irrespective of the pregnancy status. A rapid increase and biphasic radiolabel blood curve was noted in all mice but not in rats. This probably mirrors enterohepatic cycling of conjugates, e.g., glucuronides, which are formed in mice but not in rats. The plasma half-life times ranged between 7.1 and 10.3 h at the low dose and between 5.5 and 13.5 h at the high dose. The findings were confirmed when radiolabeled DEHP was given after five preceding doses. MEHP dominated DEHP in blood both at the low and the high dose level, as was seen from the AUC in rats (by a factor of 1.5–2) and mice (factor of ca. 2.5–3.5). In both species peak blood concentrations of DEHP and MEHP were comparable at the low dose (30–40 nmol/g); at the high dose peak MEHP concentration was approximately twice as high as that of DEHP. In rats, MEHP metabolites were ca. 10 % compared to the MEHP peak concentration and AUC. The MEHP metabolites were not detectable in mouse blood.

Species differences in metabolism between rats and mice were also noted by the urinary and fecal excretion pattern in a study using repeated

doses (10 doses, each 200 and 1000 mg/kg). In rats, the ω-1-oxidation metabolites, 5-OH- and 5-oxo-MEHP, accounted for 65 % compared to 35 % ω-oxidation metabolites, both in nonpregnant and pregnant rats, and at both dose levels. Excretion of metabolites occurred almost exclusively in urine. Unchanged DEHP and MEHP were excreted in comparable amounts: ca. 10 % of the administered dose each, and almost exclusively in the feces.

In mice, DEHP was also excreted exclusively in feces (6–10 % of the dose), whereas MEHP (15–25 % of the dose) was excreted 2/3 in feces and 1/3 in urine. Thus, much larger amounts of DEHP and MEHP were excreted in mice than in rats. The majority of the MEHP metabolites was excreted in urine (86 %; feces 14 %). ω-1-oxidation metabolites tended to decrease during the treatment (from 63 to 55 %) parallel to an increase in ω-oxidation (from 37 to 45 %) [193 – 197].

From these metabolism studies [193 – 197] with DEHP in pregnant and nonpregnant rats and mice it can be concluded that:

- DEHP and its metabolites are rapidly absorbed in both species as DEHP, MEHP, and MEHP-derived metabolites.
- The systemic exposure to MEHP is higher in pregnant mice than in pregnant rats, but DEHP exposure is equivalent.
- Metabolites from ω-1- (metabolites VI and IX, i.e. 5-OH- and 5-oxo-MEHP) and ω-oxidation pathways are excreted at higher levels in rats than in mice, but the subsequent β-oxidation of metabolite V (6-carboxy-MEHP) to form metabolite I (2-ethyl-4-carboxybutyl phthalate) is more active in mice than in rats.
- Repeated DEHP administration induced a modification of the metabolic pathway:
  - a decrease of DEHP excretion in nonpregnant and pregnant rats and in pregnant mice and an increase in nonpregnant mice
  - an increase of MEHP excretion in pregnant mice
  - an increase of ω-1-oxidation in rats and a decrease in mice
  - an increase of β-oxidation in pregnant mice at the high dose level.

An unpublished 65-week oral-dose toxicity study of DEHP in marmosets [198, 199] described pharmacokinetics and organ distribution following single oral doses of 100 and 2500 mg/kg DEHP to animals aged 3 and 18 months; a third group was treated at 18 months after a 65-week pretreatment with unlabeled DEHP. Ring-labeled ^{14}C-DEHP was used to determine radiolabel levels in blood (at 1, 2, 4, 8, 12, 24, 48, 72, 120, and 168 h after dosing) and in excreta. Organ distribution including male and female reproductive organs was examined 2 h after a second dosing, at least two weeks after the first treatment. Peak blood levels were reached within 1 or 2 h after dosing. Plasma half-life times of the radiolabel ranged between 2.7 and 8 h. The organ/plasma ratio was generally 0.2 to 0.4. The authors called particular attention to the small amounts of label distributed to the testes compared to rodents, in which large amounts of MEHP are distributed to the testes. Pretreatment for 65 weeks did not alter the distribution [198, 199].

No difference was seen when comparing blood concentration time courses of pregnant and nonpregnant rats and marmosets, respectively, receiving single or repeated doses of 30, 500, and 1000 mg DEHP per kg body weight. In both groups of rats MEHP AUCs were about two orders of magnitude higher than those of DEHP. Maximum MEHP blood levels were reached at 0.5, 2, and 4 h in animals receiving 30, 500, and 1000 mg/kg, respectively. The authors concluded that kinetics were linear for DEHP but saturated for MEHP, based on increased time to reach maximum blood levels [200].

The blood concentration time courses of pregnant and nonpregnant marmosets were similar except that the values of MEHP were lower in the high-dose pregnant group than in the nonpregnant group. Generally, the AUC of MEHP was ca. 10- to 20-fold higher than that of DEHP.

Comparison of species revealed that the AUCs of DEHP were similar in both species, while those of MEHP were 10-fold higher in rats than in marmosets [200].

The above confirms earlier studies in which rats dosed with 2000 mg kg^{-1} d^{-1} of DEHP for 14 d by gavage eliminated (on day 13) 56.2 % of the administered daily dose in the urine, and 58.7 % in the feces. The peak blood level on day 14 was 368 μg equivalents of DEHP per gram blood. In a parallel marmoset study the same dose regimen led to only 13 μg equivalents of DEHP per gram blood. Tissue levels were also

Table 6. Urinary metabolite excretion in a human 24 h after ingestion of DEHP, % of dose

Metabolite	Estimated elimination $t_{1/2}$, h	DEHP dose, μg/kg		
		4.7	28.7	650
MEHP	5	6.26	4.3	7.3
5-OH-MEHP	10	23.1	22.7	24.1
5-oxo-MEHP	10	17.3	13.0	14.6
Mono(2-ethyl-5-carboyxypentyl) phthalate	12–15	15.5	19.4	20.7
Mono[2-(carboxy-methyl)hexyl] phthalate	24	3.7	5.2	3.8
Total percentage of DEHP dose		65.8	64.6	70.5

much lower in marmosets, and no peroxisome proliferation or testicular effects could be detected [201].

In humans, urinary excretion of metabolites is also considered to reflect exposure and absorption. Two volunteers, each ingesting 30 mg of DEHP, excreted 11 and 15 % of the dose in the urine within 24 h. Four daily doses of 10 mg lead to a urinary recovery rate of 15–25 % [202].

One volunteer ingesting deuterium-labeled DEHP (48 mg, 0.64 mg/kg) excreted 47 % of the dose within 2 d as MEHP (7.34 %), 5-OH-MEHP (24.7 %), and 5-oxo-MEHP (14.9 %). Peak level of MEHP in urine occurred 2 h after dosing; the maximum urinary concentrations of the MEHP oxidation products occurred after 4 h. Serum levels of MEHP were higher than those of its oxidation products at all times, consistent with the more rapid elimination of polar metabolites. The estimated serum elimination half-life of all three metabolites was < 2 h [203].

The proportions of ω-1- and ω-oxidation were investigated in the 24-h urine of one volunteer ingesting three different doses (4.7, 28.7, and 650 μg/kg) of deuterium-labeled DEHP (Table 6). The ω-oxidation metabolites were mono(2-ethyl-5-carboyxypentyl) phthalate and mono[2-(carboxymethyl)hexyl] phthalate. The results indicate that excreted MEHP alone does not properly reflect the exposure because ω-1- and ω-oxidation contribute significant and varying proportions of the entire metabolism. The authors argued that serum MEHP level is not a useful marker for DEHP exposure due to its short half-life. However, they stated that serum levels were of the same orders of magnitude despite the fact that the human dose was 50–1000 times lower than in animal studies [378].

There are species differences in Phase-II metabolism between rats and mice (see above) and primates. After intravenous administration of DEHP to the African green monkey, 80 % of the urinary metabolites were excreted as glucuronide conjugates, in contrast to rats, which did not excrete conjugates in the urine [204]. In mice receiving 200 mg/kg, high proportions of the urinary MEHP (58 %) and MEHP-derived metabolites (49 %) were excreted as glucuronide conjugates [196]. This indicates that species differences in the rate of glucuronidation occur and may cause considerable variations in biological half-lives and tissue levels. As a consequence, the higher glucuronidation (detoxification) rate of the metabolites in primates may contribute to the observed lower MEHP peak concentrations and AUCs in marmosets compared to rats [198 – 200] and the higher resistance to DEHP-induced toxicity.

6.2.5.2. Acute Toxicity

Having entered the organism, one of the two ester bonds of phthalic diesters is cleaved and the alcohol is released [184, 185, 187, 202, 205]. Toxicological effects may result from the phthalic acid monoester and its sequel products and/or from the alcohol. In some cases (e.g., dimethoxyethyl phthalate) the impact of the alcohol is of greater importance [206 – 208]. The acute toxicity of diallyl phthalate is relatively high (LD_{50}, rat, oral: 800–1700 mg/kg) and it is irritating [215, 216], whereas most phthalates show very low acute toxicity. Acute oral toxicity in the rat and in the rabbit is summarized in Table 7.

Table 7. Acute toxicity of phthalate esters [209]

Phthalate ester	LD_{50} (rat, oral), mg/kg	LD_{50} (rabbit, dermal), mg/kg
Diallyl phthalate	≈ 1 500	
Dimethyl phthalate	> 5 000	> 12 000
Diethyl phthalate	> 9 000	> 20 000
Dibutyl phthalate	> 8 000	> 20 000
Diisobutyl phthalate	> 15 000	> 10 000
Diethylhexyl phthalate	> 30 000	> 24 500
Diisononyl phthalate	> 10 000	> 3 100
Diisodecyl phthalate	> 20 000	> 3 600
Diundecyl phthalate	> 15 800	> 7 900

6.2.5.3. Irritation and Sensitizing Potential

As a rule, most phthalates do not cause significant irritation [210 – 212] or sensitization of the skin [213, 214]. Diallyl phthalate [215, 216], however, may exert a weak irritating potential. Earlier reports of respiratory sensitizations appear to be related to contamination of the esters with phthalic anhydride and maleic anhydride.

6.2.5.4. Repeated DEHP Dosing

Effects in rodents on liver, kidneys, and testes were seen, as well as increased incidences of tumors in liver and testes (Section 6.2.5.6) and toxicity to reproduction (Section 6.2.5.7). Subacute and chronic effects as well as reproductive effects depend to some extent on chain length and structure of the alcohol moiety. Thus, a differentiation between short-, medium-, and long-chain esters appears to be justified (Section 6.2.5.8). Generally, the effects of phthalates are due to the monoester and metabolites derived therefrom, rather than the liberated alcohol and its metabolites. Exceptions to this rule are the more toxic allyl alcohol and the teratogenic methoxyethanol. 2-Ethylhexanol leads to 2-ethanoic acid, which is known to have a potential for teratogenicity. However, this is insufficient to explain the teratogenic effects seen with DEHP.

DEHP and several other phthalates act on the liver lipid metabolism in a qualitatively similar manner to a group of structurally unrelated compounds which are used or investigated as hypolipidemic drugs to reduce serum cholesterol and triglycerides. Increased enzyme activities involved in mitochondrial and peroxisomal fatty acid β- and ω-oxidation appear to be early effects of DEHP [217 – 219]. Initially, short transient inhibition of β-oxidation and lipid accumulation may occur, but subsequently can be overcome [220 – 222]. In rodents, but not in primates, the numbers and volumes of peroxisomes increase, and the liver undergoes a short period of DNA replication and cellular proliferation with a maximal mitogenic response after 24 h [223 – 226]. This results in a dose-related increase in liver weight which may double [224 – 231]. Plasma cholesterol values decrease in the same way as is observed with hypolipidemic drugs.

A study in Sprague–Dawley rats reports a continuous increase of peroxisomal enzyme activities upon prolonged treatment over two years [232]. In Fischer rats, increased peroxisomal enzyme activity was still present after two years, but not hepatomegaly or cellular proliferation [233]. 25 mg/kg of DEHP appears to be the no-effect level for the increase of peroxisome-related enzyme activity [228]. The no-effect levels for DINP and DIDP are somewhat higher; DBP is much less active [234].

Peroxisome proliferation and liver weight increase occurred in mice receiving DEHP in the diet (>1000 mg/kg, 24 weeks), but not in mice which were deficient in PPARα (PPARα-knock-out mice) [235]. Peroxisome proliferator-activated receptors (PPAR) are physiologically thought to play a role in lipid homeostasis in that they induce the expression of enzymes involved in lipid metabolism (β-oxidation, cytochromes P450). The three different types that are known to exist [PPARα, -δ (or -β), and -γ] respond to endogenous (fatty acids, prostaglandins, leucotrienes) and exogenous ligands (peroxisome proliferators: fibrates, phthalate esters, phenoxy herbicides). The sequence of action is the initial formation of a proliferator–receptor complex which becomes a transcription factor after association with the 9-*cis*-retinoic acid receptor [236, 237]. The transcription factor stimulates the expression of genes having specific promoter regions (peroxisome proliferator responsive elements, PPRE). The promoter regions are identical for all types of PPARs; differences in the enzyme induction pattern and intensity result from differences in the vicinity of the promoter region, cofactors, and the ligand itself, which, like

DEHP, may have overlapping affinity to all three types of PPAR [238, 239].

In rodent liver, PPARα mediates an increase of cell proliferation, inhibition of apoptosis, and increased expression of peroxisomal, mitochondrial, and microsomal enzymes [231, 240 – 242]. Under physiological conditions PPARγ is equally involved in the growth regulation of fat cells and epithelial cells of diverse organs [239], but its role in processes stimulating and inhibiting tumor growth are not yet fully understood [243]. PPARδ is also involved in cell cycle regulation and conversion of colon mucosa cells into carcinoma cells [243, 244].

The concentration of PPARs is high in rodents but low in primates including humans. In marmosets receiving 100, 500, and 2500 mg kg^{-1} d^{-1} DEHP for 13 weeks (oral, 91 consecutive days) peroxisome proliferation was not detected, neither by morphology nor by biochemistry, except for minor changes indicative of induction of some enzyme activities (increase in P450 and microsomal protein content) in all males and in females at the intermediate and high dose levels. The number of peroxisomes per cell was not affected. Only the mean volume of peroxisomes showed a slight increase (1.3–1.4-fold) at 500 and 2500 mg/kg b.w. Gross pathology, organ weights, histopathology, and electron microscopy did not show any substance-related findings in any of the treated groups. Especially the liver, pancreas, and testes or testicular parameters (Sertoli cells, Leydig cells, spermatogonia, spermatids, spermatozoa) were not affected. The NOAEL for systemic toxicity was 500 mg kg^{-1} d^{-1} in marmosets [245].

On the other hand, it was pointed out that DEHP can affect multiple signalling pathways by transcriptional activation of PPAR-regulated genes (suppression of apoptosis, protooncogene expression) and can induce biological effects (morphological cell transformation, decreased levels of gap junction intercellular communication) in several tissues independent of peroxisome proliferation and receptor binding [246].

Effects in kidneys, including changes in weight (both increase and decrease), function, and histopathology, were seen in several rodent studies (rat, oral, 2.1 mg kg^{-1}, 3 times per week, for 3–12 months; cystic changes and decreased creatinine clearance [247]; Fischer 344 rat, oral, $\geq$ 147 mg kg^{-1} d^{-1}, 104 weeks; increased weight, chronic nephropathy [248]; mouse, $\geq$ 99 mg kg^{-1} d^{-1}, 104 weeks; decreased weight, chronic nephropathy [249, 250] but not in primates [245].

6.2.5.5. Genotoxicity and Mutagenicity

DEHP and its metabolites have shown little if any genotoxic activity in a wide variety of short-term tests [251 – 257]. A few in vitro experiments on DEHP and MEHP were weakly positive and probably related to the cytotoxicity of MEHP [258 – 261]. The metabolites 5-OH- and 5-oxo-MEHP and mono(5-carboxyl-2-ethylpentyl) phthalate were also negative in the Ames test [262]. Positive results were seen with MEHP (39 μg/mL) in the Comet assay for DNA single-strand breaks using human leucocytes in the absence of metabolic activation [263]. Chromosomal aberration was seen with MEHP ($\geq$ 3.9 μg/mL) in Syrian hamster embryo cells [264].

Two dominant lethal studies [265, 266] using very high doses and exposure routes of little relevance (subcutaneous, intraperitoneal) showed positive results in vivo. However, the effects were of borderline significance and appear to reflect testicular toxicity rather than true genotoxicity. After oral administration of DEHP and MEHP, no dominant lethal effects were observed [267, 268].

DEHP has been shown not to react covalently with rat liver DNA after gavage administration following a 3-week prefeeding period [269]. When high doses of DEHP, known to cause peroxisomal proliferation, were repeatedly administered, an increase of 8-hydroxyguanosine in the DNA was found in the rodent liver but not in the kidney. This could reflect an oxidative DNA alteration [270], which, e.g., is also observed in rats fed on a choline-free diet [271], and other conditions of oxidative stress [272], but a second study on the same end point using more animals of the same strain, route, and dose (Fischer 344 rats, 12 000 mg DEHP per kg feed) but shorter duration (11 and 22 weeks, 6 rats per group) was negative [273]. Determination of 8-hydroxydeoxyguanosine in the DNA is very difficult to perform. In conclusion, there is no evidence that DEHP is genotoxic; the few positive results are contradictory or could result from cytotoxicity [274].

Other phthalates so far investigated showed similarly negative genotoxicity test results in most of the experiments [213, 227, 260].

6.2.5.6. Carcinogenicity

DEHP is again the phthalate ester most thoroughly investigated. Earlier chronic toxicity studies of DEHP showed no tumors in Wistar or Sherman rats over two-year periods at dietary levels of 4000 and 5000 ppm [275, 276]. Sprague-Dawley rats did not respond with liver tumors or preneoplastic lesions when fed with doses of 200, 2000, and 20000 ppm for 2 years; a continuous and sustained dose-and time-dependent peroxisomal proliferation was observed [232]. Due to the low animal numbers, all these studies are of limited value in assessing the carcinogenicity risk.

In more recent long-term rat and mouse feeding studies, increased incidences of liver tumors were seen. For example, increased incidences of hepatocellular carcinoma, adenoma, and preneoplastic foci were seen in Fischer 344 rats with 50 animals per dose group and sex, exposed to 6000 and 12000 ppm DEHP in the diet, and in B6C3F1 mice with 3000 and 6000 ppm [232]. The actual uptake from the feed was estimated to be 320–770 mg kg^{-1} d^{-1} in rats, depending on dose, sex, and food consumption, and 600–1800 mg kg^{-1} d^{-1} in mice [277]. A clear dose-dependent increase in the incidence of hepatocellular adenomas was noted in male/female rats fed 6000 and 12 000 ppm (322/394 and 674/774 mg kg^{-1} d^{-1}, respectively), and in male/female mice fed 3000 and 6000 ppm (672/799 and 1325/1821 mg kg^{-1} d^{-1}, respectively) for two years [278 – 280]. With a dose of 3000 mg kg^{-1} d^{-1}, 78.5 % of the treated rats developed liver tumors and 29 % pancreatic islet cell adenomas [281].

The relationship between hepatic peroxisome proliferation, cell proliferation, and carcinogenicity has been evaluated in chronic studies of DEHP in rats and mice [248 – 250]. Animals were fed a diet containing DEHP for 104 weeks; additional groups treated for only 78 weeks were subsequently placed on a DEHP-free diet for 26 weeks to examine reversibility of the effects. In rats, relative liver weight was significantly increased at 147 mg kg^{-1} d^{-1}, and this correlated with increased activity of palmitoyl CoA oxidase, a peroxisomal enzyme marker. Treatment with DEHP resulted in hepatocellular adenomas and carcinomas at weeks 78 and 104; in general, the incidence of tumors at week 78 was low except for the high-dose group. Cessation of treatment resulted in a one-third to one-half decrease in the incidence of total neoplasia relative to the groups treated for 104 weeks and in a reduction in liver weight and palmitoyl CoA oxidase activity. The increased incidence of hepatic tumors was significant at the 147 mg kg^{-1} d^{-1} dose level and higher in rats [248, 249].

Similar findings were reported in B6C3F1 mice. Cessation of treatment in mice resulted in a 50 % decrease in the incidence of total neoplasia in males relative to groups treated for 104 weeks, but the incidence in females decreased only slightly. The increased incidence in hepatocellular tumors achieved statistical significance at dose levels of 292 mg kg^{-1} d^{-1} DEHP [249, 250].

The role of peroxisome proliferation was further elucidated in an experiment with PPARα-null and wild-type [PPARα(+/+)] mice that were fed 0.1 % Wy-14,643, a model peroxisome proliferator. After 11 months of feeding 100 % of the wild-type animals had multiple hepatocellular neoplasms, including carcinomas and adenomas, while the PPARα(−/−) mice were unaffected [282].

Leydig cell tumors occurred in high incidences of 90–95 % in untreated F-344 rats and in animals at intermediate DEHP dose levels. The incidence was less than one-third in the high-dose groups [248, 249, 278, 280]. High incidences of this tumor are specific for this strain [283].

Similar results at comparable dose ranges (1.2–2.0 % in the diet) were obtained in consecutive studies with DEHP in smaller collectives of Fischer rats [230, 233, 284, 285] and in bioassays with two different forms of diisononyl phthalate and with a phthalate of a mixture of heptanols, nonanols, and undecanols (C_7–C_{11} fraction) [210, 286]. Other investigators [287] employing lower dose levels (up to 6000 ppm) did not find a statistically significant increase of liver tumors with a diisononyl phthalate. Under these conditions (6000 ppm over two years) the material did not exhibit peroxisome proliferation at the end of the 2-year observation pe-

riod. Different forms of diisononyl phthalates and related materials may exhibit slightly different profiles in relation to their structure and the type of branching and molecular shape.

Butyl benzyl phthalate, which is not a significant peroxisome proliferator, caused some mononuclear cell leukemia in Fischer rats at 1.2 % in the diet, but no liver tumors [210, 288].

Diallyl phthalate, similarly, did not cause liver tumors, but necrotic and fibrotic lesions of the liver and bile duct hyperplasia in rats, and gastric inflammation and hyperplasia in mice were observed. These effects were probably mediated via release of allyl alcohol. The doses in this study were 100 and 300 mg/kg in rats, and 50 and 150 mg/kg in mice [171].

In hamsters, inhalation of atmospheres saturated with DEHP vapor (ca. 15 mg/m^3; ca. 7–10 mg/kg) for two years or intraperitoneal injection up to a total dose of 54 g/kg, administered by weekly injections, did not increase the numbers of tumors [251]. Similarly, feeding of comparably low levels (300 and 1000 ppm) of DEHP did not lead to liver tumors [284].

Initiation/Promotion Experiments. No initiating activity was found with DEHP after single oral administration of 10 g/kg or after 12 weeks feeding of 1.2 %, followed by 0.05 % feeding of phenobarbital as promoter [289]. Generally, peroxisome proliferators have been recognized as a group of epigenetic rodent liver carcinogens which widely differ in their carcinogenic activity [290, 291]. Tumor-promoting activity of DEHP was investigated with several experimental procedures. The results were either negative or weak [292 – 296]; in some regimens antipromoting activities were also recorded [295, 297, 298]. There is evidence that peroxisome proliferators exert significant promoting effects only late in the rodent's life [299]. Di-*n*-octyl phthalate, which is not a peroxisome proliferator, was shown to be a liver tumor promoter [300], probably due to lipid accumulation and hepatomegaly [201].

There are several hypotheses on the mechanism of action. The most important and most complex seem to be the receptor-mediated mechanisms (Section 6.2.5.4), which are also associated with liver weight increase and replicative DNA synthesis [225]. Some authors focused on reduced expression of growth factor receptors [292, 301, 303]. Another hypothesis is that oxidative stress plays a role. Increased intracellular H_2O_2 and reactive O_2 species released from the peroxisomes [290, 304 – 306] may cause continuous accumulation of 8-hydroxydeoxyguanosine in the DNA [270]. Leakage of H_2O_2 from peroxisomes is reflected by increases in oxidized glutathione in the bile [291]. Conjugated dienes and accumulation of lipofuscin pigments in lysosomes were also observed and indicate oxidative stress to the cell [224, 277, 291, 307]. It is, however, not clear whether oxidative stress is cause or consequence in the course of phthalate ester tumorigenesis.

No direct genetic activity is observed with phthalate esters (see also Section 6.2.5.5). This also applies to hypolipidemic drugs, even if they stimulate peroxisomes at very low doses and are strongly carcinogenic in rats and mice [291]. These rodent liver tumors are therefore regarded as epigenetic and a sequel of the liver effects. The present knowledge allows the assumption that threshold levels for these early-appearing liver effects also reflect a threshold for the appearance of liver tumors in rodents. In addition, phthalate esters do not cause peroxisome proliferation in nonrodents including primates.

6.2.5.7. Reproductive Toxicity

Testicular Toxicity. Phthalate esters of medium and longer chain length (C_3–C_6, DEHP) cause testicular damage [290, 308 – 322]. The lesion is characterized by an initial effect on Sertoli cells [323, 324] leading to the exfoliation of spermatocytes and spermatids with atrophic degeneration of seminiferous tubules containing cytoplasmic remnants of Sertoli cells and spermatogonia [323, 325].

Minimal vacuolization of Sertoli cells was reported to occur in rats receiving 38 mg kg^{-1} d^{-1} DEHP for 13 weeks; at a 10-fold higher dose level, 375 mg kg^{-1} d^{-1}, mild vacuolization and testicular atrophy were noted. The authors derived a no-adverse-effect level (NOAEL) of 3.7 mg kg^{-1} d^{-1}. However, because the same effects were seen in control animals in a second experiment of the same study[333], the true NOAEL should be considered to be 37.6 mg kg^{-1} d^{-1}. Germ cell detachment may also be observed in vitro with MEHP [323, 334] but

Table 8. Effects of DEHP on male fertility

Species, strain	Dose, mg kg^{-1} d^{-1} (duration)	Effect (dose, mg kg^{-1} d^{-1})	Ref.
Rat, SD	0, 143, 737 (17 weeks)	testes weight ↓ (737)	[326]
Rat, F-344	0, 322, 674 (103 weeks)	atrophy of seminiferous tubules (674)	[280]
Rat, F-344	0, 330, 1000, 3000 (13 d)	slight atrophy of testes (1000)	[327]
Rat, F-344	0, 320, 1250, 5000, 20000 (60 d)	testes weight ↓, testosterone↓, LH ↑, FSH ↑ (284)	[328]
Rat, SD	0, 10, 100, 1000, 2000 (5 d)	palmitoyl CoA oxidase ↑ (10), liver weight ↑ (100); testes weight ↓, no. Sertoli cells ↓, spermatocytes ↓, spermatids ↓, testicular zinc ↓ (1000)	[324, 329]
Rat, Wistar	0, 50, 100, 250, 500 (30 d)	slight testes weight ↓, change of testicular enzymes (50); significant testes weight ↓ (100); histopathological changes, testes (250)	[330]
Rat, F-344	0, 6, 29, 147, 789 (104 wk)	bilateral aspermatogenesis in aged rats (29); testes weight ↓, bilateral aspermatogenesis (789)	[248]
Rat, Wistar	0, 113, 340, 1088	advanced atrophy of seminiferous tubules, aspermia in epididymes (1088)	[331]
Rat, SD	0, 0.12, 0.78, 2.4, 7.9, 23, 77, 592, 775 (unspecified weeks)	sperm density ↓; male accessory organ weights ↓; atrophy of seminiferous tubules; sloughed epithelial cells in epididymis (592)	[332]
Marmosets	0, 100, 500, 2500 (91 d)	testes and testicular parameters (Sertoli cells, Leydig cells, spermatogonia, spermatids, spermatozoa) not affected (2500)	[199]

not with 2-ethylhexanol [335]. Further hormonal sequels of the testicular atrophy are decreases of serum testosterone and increases of serum FSH (follicle stimulating hormones) and LH (luteinizing hormones) [328]. Some of the early and late effects and the respective dose levels following administration of DEHP are summarized in Table 8. Extensive studies have been conducted on the effects on Sertoli and Leydig cells, as well as testosterone and sex hormone regulation in male rodents. The absence of any effect in marmosets [199] indicates that marked physiological and endocrine differences exist between rodents and primates regarding Sertoli cell and Leydig cell physiology and susceptibility, and sex hormone regulation. Rodent Leydig cells are regarded as more sensitive than human Leydig cells [283]. Atrophy of the seminiferous tubules is generally seen in rodents at doses >250 mg kg^{-1} d^{-1}. The exact mechanism is still unknown [209].

Early functional effects, e.g., reduced sperm counts, may be seen at lower dose levels, whereas in acute studies relatively high (> 250 mg/kg) oral doses are required to induce visible testicular damage. In experiments of longer duration the effective levels are considerably lower. Serum testosterone [328] and male fertility in continuous breeding protocols appear to be sensitive parameters. The no-effect levels (NOAEL) are in the range 25–50 mg kg^{-1} d^{-1}. A dose of 1250 ppm (70 mg kg^{-1} d^{-1}) in the diet of rats decreased serum testosterone within 60 d [328]. This is a dose range which caused malformations in a mouse teratology study [304, 336].

Reversal of testicular effects is slow and incomplete. After administration of 2000 mg kg^{-1} d^{-1} DEHP for two weeks to rats, only slight reversal of the testicular lesions occurred after 45 d [318]. Incomplete reversal was also observed after 60-d feeding of 20 000 ppm of DEHP in the diet of rats and a 70-d postobservation period [328]. The effects are age- and, to some extent, species-related [321]. Young rats (4-weeks-old) appear to be more susceptible (lower doses, shorter duration required) than adult rats. This age-dependent susceptibility has been successfully modeled with MEHP in primary testicular cultures from rats of varying age [334].

As to species differences, DEHP can also induce testicular atrophy in mice [315, 337], guinea pigs [325], ferrets [338], and hamsters [229]. Marmosets appear to be more resistant; oral doses of DEHP (2000 mg/kg for 2 weeks [201]; 2500 mg kg^{-1} d^{-1} for 91 consecutive days [199]) caused no testicular toxicity. A lower intestinal resorption and higher rate of excretion may contribute to these findings [218].

The mechanism of testicular toxicity is still unclear, but zinc is thought to play a role. Testicular zinc content is decreased while urinary zinc increases [308, 314, 319]. Detailed studies with di-*n*-pentyl phthalate [339] have not been able to indicate whether the changes in testicular zinc content are solely a secondary phenomenon. A zinc-supplemented diet partially protected rats from DEHP-induced testicular damage, whereas rats on a zinc-deficient diet were more sensitive [327]. Hypolipidemia and hepatomegaly appeared to occur independently of zinc supplementation. On the other hand, there is evidence that reduced testicular zinc content was probably a result of testicular toxicity after DEHP treatment rather than a mediator of DEHP toxicity [340].

Diethyl and di-*n*-octyl phthalate showed no adverse effects on fertility and the testes [337, 345]. Comparative studies using several mono- and di-esters were conducted in male Sprague-Dawley rats, orally treated for four weeks with phthalate diesters at 500 mg kg^{-1} d^{-1} or phthalate monoesters at 250 mg kg^{-1} d^{-1} [313]. Animals were evaluated for body weight, testicular and epididymal weights, epididymal sperm count, and sperm motion parameters. There were no significant effects of treatment on body or reproductive organ weight. The order of effects on sperm motility for the diesters was DEHP > DnOP > DEP > DUP > DIDP > BBP, and for the monoesters MBuP > MEP > MEHP (MBuP = monobutyl phthalate; MEP = monoethyl phthalate) [313].

The relevance of rodent data to humans is still under debate. The absence of testicular effects in monkeys even at 2500 mg kg^{-1} d^{-1} [199], and the failure to confirm vacuolization of Sertoli cells in rats [331, 332] are major factors in this discussion because the NOAEL of 3.7 mg kg^{-1} d^{-1} for vacuolization was used by several institutions during their evaluations and risk assessments [274, 342, 343].

Effects on Fertility and Development. There are two high-quality rat studies available which deserve a more detailed description.

In a two-generation reproductive study in Wistar rats DEHP was administered in feed at 0, 1000, 3000, and 9000 ppm, resulting in estimated intakes of 0, 113, 340, and 1088 mg kg^{-1} d^{-1} [331]. Cauda epididymal sperm were assessed for motility, morphology, and head count in F0 and F1 males at necropsy. Testicular sperm head count was also performed. Male pups were evaluated for the presence of nipples and areolas. Day and weight at vaginal opening and preputial separation were assessed in pups. F2 pups were also evaluated by functional observation battery and motor activity testing, and in a water-maze test of learning and memory.

The high-dose level (9000 ppm) was associated with a decrease in feed consumption and weight gain at several intervals during the study. F2 pups in the high-dose group were smaller and gained less weight from birth through the assessment of functional observation battery and water-maze testing. Grip strength was reduced in males and hind-limb splay was reduced in both sexes in the high-dose group. There were no other treatment-related findings in the functional observation battery or in the water maze. Differential ovarian follicle counts of F0 and F1 adults showed a deficit in growing follicles and corpora lutea in the high-dose group.

It was concluded that reproductive performance and fertility were affected at the 9000 ppm dose level (1088 mg kg^{-1} d^{-1}) with a NOAEL of 3000 ppm (340 mg kg^{-1} d^{-1}), and that developmental toxicity was noted at 3000 and 9000 ppm, with an increase in stillbirth, an increase in postnatal day (PND) 0–4 pup mortality, retardation of F2 pup body weight, altered male anogenital distance, and retained nipples/areolas. A delay in sexual maturation was also noted in F1 offspring at 9000 ppm. The NOAEL for developmental toxicity was considered by the study authors to be 1000 ppm, as was the NOAEL for systemic toxicity (113 mg kg^{-1} d^{-1}) [331].

In a multigeneration continuous-breeding study Sprague-Dawley rats were fed diets containing 1.5 (control group exposed to background DEHP levels), 10, 30, 100, 300, 1000, or 7500 ppm DEHP from the first day of the

study until the day of necropsy [302, 332]. Due to a lack of reproductive effects in the first litter produced, the study was repeated with two additional doses, 1.5 (control) and 10 000 ppm. Ranges of DEHP intake in the F0, F1, and F2 animals were estimated at 0.09–0.12, 0.47–0.78, 1.4–2.4, 4.8–7.9, 14–23, 46–77, 392–592, and 543–775 mg kg^{-1} d^{-1}. At about five weeks of age, F0 rats were fed the DEHP-containing diets for six weeks prior to mating and were then cohabitated for nine weeks. The first two litters delivered during the cohabitation period (F1a and F1b) were counted, weighed, assessed for anogenital distance, and then discarded. The third litter (F1c) was raised by the dam. Following weaning of pups, vaginal cytology was monitored in F0 females for 14 d. After completion of crossover studies described below, some of the F0 rats were necropsied. Sperm analyses were conducted, and organs were collected for histopathological evaluation. F1 pups were counted, weighed, and examined for anogenital distance and nipple retention during the lactation period. On PND 16, one female per litter was evaluated for vaginal opening, and a second was selected for F1 mating. One male per litter was selected for mating, and 4 or 5 males per litter were evaluated for testicular descent and preputial separation. At weaning on PND 21, pups were given diets containing the same DEHP concentrations as their parents. On PND 81, the F1 rats chosen for mating (17 per sex per group) were randomly assigned to breeding pairs (preferably nonsibling) and cohabited for nine weeks. The study conducted in F0 parents and F1 offspring was repeated in F1 parents and F2 offspring, except that no F3 offspring were mated. Selected F3c males were necropsied on PND 63–64 and selected females on PND 60–74.

During numerous time periods of the study and especially at necropsy, body weight gains were decreased in rats from the 7500 and 10 000 ppm groups. Dam body weights during delivery and lactation were decreased by 8–20 % in the F0 10 000 ppm group. Increases and decreases in feed intake were observed at most dose levels. In the F0 7500 and 10 000 ppm groups, feed intake was decreased during lactation. The liver was identified as a target of toxicity with increases in liver weight and hepatocellular hypertrophy observed at dose levels $\geq$ 1000 ppm. Changes in organ weights and lesions were also observed in the kidney at $\geq$ 7500 ppm and the adrenal gland at 10 000 ppm. The only reproductive effects observed in the F0 parents occurred at 10 000 ppm and included decreases in sperm counts and velocity, reductions in testis and epididymis weights, and increased numbers of rats with small testes. Histopathological findings associated with small testes included minimal to marked atrophy of seminiferous tubules characterized by loss of germ cells.

The lowest dose level producing dose-related effects in F1 offspring was 7500 ppm, and those effects included decreases in number of live pups per litter, reduced male anogenital distance, and delays in vaginal opening, preputial separation, and age at testicular descent. Additional effects noted in the F1 offspring from the 10 000 ppm group included decreased live pup weight at birth and during the lactation period and increased ratio of female anogenital distance to body weight. Fertility was compromised in the F1 rats from the 10 000 ppm group, which did not produce any viable litter. Other effects observed in F1 parents were similar to those observed in F0 parents but occurred at lower dose levels. Histopathological findings observed in all animals of the 7500 and 10 000 ppm groups were consistent with those observed in the F0 generation and included minimal-to-marked seminiferous tubule atrophy and occasional sperm-release failure. Additional reproductive effects observed in F1 rats were reduced sperm counts at 7500 ppm and higher and increased uterus and ovary weights at 10 000 ppm.

In the F2 pups, delays in preputial separation and testicular descent occurred at each dose level. Other effects occurring in F2 pups of the 7500 ppm group included delayed vaginal opening and reductions in live pup weight at birth and during the lactation period, male anogenital distance, and survival during the lactation period.

In the F3 offspring, seminiferous tubule atrophy was observed in 10/10 males from the 7500 ppm group. All other reproductive effects in F2 adults occurred at 7500 ppm and included decreases in pregnancy index, the number of litters per pair, male reproductive organ weights, sperm counts, and sperm motility. The F3 generation was the only generation in which there was an increase in males with nipples. Other F3 effects included reduced sperm counts and weights of

dorsolateral prostate, testis, and epididymis at 7500 ppm.

The study authors discussed the relevancy of small male reproductive organ sizes observed in both F1 and F2 rats of the 300 ppm groups. They noted that although incidences were low, the effects were consistent with phthalate-induced toxicity. The incidence of small testes and epididymides exceeded historical control data from the laboratory. Therefore, the effects were considered as potentially treatment-related. However, the study authors concluded that the overall significance of the effects cannot be determined due to lack of histopathological data and lack of adverse reproductive effects at 300 and 1000 ppm. Using crossover studies, the male was identified as the sex in which adverse fertility effects occurred.

In conclusion, DEHP was clearly a reproductive and developmental toxicant at 7500 and 10 000 ppm based on changes in fertility and pregnancy indices, litter data, sperm parameters, sexual development, and/or histopathological changes in testes. Intake at 7500 ppm was estimated at 392–592 mg kg^{-1} d^{-1}, and that at 10 000 ppm at 543–775 mg kg^{-1} d^{-1}. Based on the incidence of testicular abnormalities, the CERHR Expert Panel considered 300 ppm to be an effect level, giving a NOAEL of 100 ppm, about 3–5 mg kg^{-1} d^{-1} [344]. Sertoli cell vacuolation, which was the end point driving the LOAEL in the original CERHR Expert Panel evaluation of DEHP [342, 344], was not increased by DEHP treatment in this study [302, 332].

In CD-1 mice exposed to 0.3 % DEHP in the feed in a continuous breeding protocol, complete suppression of fertility at the highest dose ($\approx$ 430 mg kg^{-1} d^{-1}) was observed. At 0.1 % a significant reduction in fertility was induced. No effects were seen at 0.01 % ($\approx$ 20 mg kg^{-1} d^{-1}). A cross-over mating employing treated males and females from the 0.3 % groups indicated significant effects on both the male and female reproductive function [304, 345].

DBP exposure in mice resulted in a reduction in the number of litters at 1.0 % in the diet, but not at 0.3 %. Females were more affected than males [345]. Di-*n*-hexyl phthalate affected both sexes of mice in a dose-related fashion [345] at 0.3, 0.6, and 1.2 %. Di-*n*-pentyl phthalate completely inhibited fertility at 1.25 % and 2.5 %. Di-*n*-propyl phthalate caused complete inhibition at 5.0 % and reduced fertility at 0.5 % and 2.5 %. Di-*n*-pentyl phthalate was more toxic to males, and di-*n*-propyl phthalate more toxic to females [337]. Diethyl and di-*n*-octyl phthalate showed no adverse effects on fertility and the testes [337, 345].

An abnormal sexual development was seen in male rats with in utero and lactational exposure to DEHP (oral gavage, dose levels 0, 375, 759, or 1500 mg kg^{-1} d^{-1}, dosing during gestational day 3 through postnatal day 21) [346]. Dose-related effects on male offspring included reduced anogenital distance, areola and nipple retention, undescended testes, and incomplete preputial separation. Testis, epididymis, glans penis, prostrate and seminal vesicle weights were reduced at postnatal days 21, 63, and/or 105–112. Further increased incidences of prostrate and seminal vesicle agenesis, reduced sperm counts, and testicular, epididymal and penile malformations were noted. The observation that many of the exposed males were sexually inactive was suggested to reflect an inhibition of the sexually dimorphic central nervous system development. DEHP was attributed an "anti-androgenic action" the anogenital distance reduction was only seen in male rats, together with the retention of areolas and nipples, whereas DEHP had no significant effect on vaginal opening or the first estrus. The exact mode of action is still unclear, but anti-androgenicity must not be confused with estrogenicity [346]. In an uterotrophic assay in ovariectomized female rats, DEHP (1000 mg kg^{-1} d^{-1} for 5 d) had no estrogen activity as it failed to change sex hormone levels or uterus estrogen receptor levels, or to increase uterus weight, and finally there was no difference compared to controls during histopathology in uterus [347, 348].

Prenatal Toxicity; Embryotoxicity; Teratogenicity. Embryotoxic and teratogenic effects of DEHP and DBP depend on species, dose, and route of administration. Mice are more sensitive than rats, and feeding appears to be more effective than gavage or inhalation.

In mice, DEHP and DBP caused teratogenicity after oral administration [337, 345, 349 – 353]: In CD-1 mice 0.1 and 0.15 % of DEHP in the diet produced both maternal and fetal toxicity [304, 352]. Malformations (eye and

tail defects, exencephaly, aortic and pulmonary arch defects, and skeletal defects) were observed at 0.05 % in the absence of maternal or fetal toxicity. No effects were reported at the lowest dietary level in this study (0.025 % DEHP, equivalent to ca. 35 mg kg^{-1} d^{-1}) [354].

DBP caused reduction of the numbers of litters and of live pups per litter at the 1.0 % level, but not at 0.3 % in a continuous breeding study [345]. In another feeding study in mice, with 1 % DBP in the diet, a slight increase in malformations (exencephaly and spina bifida) was observed [349, 350].

Butyl benzyl phthalate at 2 % in the diet of rats caused complete resorption of all litters (embryonic or fetal death), whereas 1 % and 0.5 % were without effect [190]. In another rat feeding study [355] 2 % of butyl benzyl phthalate in the diet did not lead to complete resorption of the litters; the number of malformations was greatly increased. In mice, malformations occurred, with a dose-dependent increase at 0.5 and 1.25 %. The maternal and developmental no-adverse-effect level was 0.1 % [355].

In rats, only embryotoxic effects were observed, with little indication of teratogenicity, even at dose levels that led to some maternal toxicity [331, 336]. Postimplantation loss was significantly increased in F0 and F1 dams at 1088 mg kg^{-1} d^{-1}, and viability of the F1 and F2 pups was reduced in the groups at 340 mg kg^{-1} d^{-1}. Genital organ malformations (hypospadia, cleft prepuce, small penis) were noted in only two high-dose F2 pups [331].

The number of fetuses with external, soft tissue, and skeletal malformations was increased in a Wistar rat study [0, 40, 200, 1000 mg kg^{-1} d^{-1}, exposure during gestational day (GD) 6–15]. Malformations were seen at the maternally toxic high-dose level, and were mainly confined to the tail, brain, urinary tract, gonads, vertebral column, and sternum [356].

Following exposure to DEHP in utero (GD 14 to PND 3, 750 mg kg^{-1} d^{-1}) malformations of reproductive organs were observed in 82 % of DEHP-treated males at necropsy. The types of malformations included permanent nipples, clefting of phallus and hypospadias, vaginal pouches, agenesis of prostate, seminal vesicles, or coagulating glands. Sperm production and sperm numbers were reported to be not affected by DEHP treatment, as was sexual behavior in adult rats, except that males with malformed penises were unable to achieve intromission. Testicular defects included hemorrhage, granuloma, fibrosis, reduced size or atrophy, and nondescent associated with abnormal gubernacula or ligaments. Dihydrotestosterone-dependent end points, i.e., reduced AGD and areolas, were affected [357]. Inhalation of DEHP aerosols up to 300 mg/m^3 did not cause embryotoxic or teratogenic effects [300].

As to the mechanism, zinc depletion and interference with sex hormone regulation have been discussed. Since DEHP in the diet might interfere with zinc resorption and metabolism the prenatal toxicity may partly be due to zinc depletion, which is known to cause teratogenic effects [358]. Dihydrotestosterone-dependent end points, i.e., reduced AGD and areolas, were affected by DEHP, BBP, and DINP, but not by DMP, DEP, and DOTP [357].

No teratogenicity or embryotoxicity was found with dimethyl [359] and diethyl phthalate [360] in rat feeding studies (0.25–5.0 % in the diet) or with dimethyl phthalate in rats after occlusive epicutaneous administration (0.5, 1.0 and 2.0 mL kg^{-1} d^{-1}; 2 h/d; gestation days 6–15) [361].

Di(methoxyethyl) phthalate is teratogenic due to the release of highly teratogenic methoxyethanol [206 – 208].

6.2.5.8. Effects of Phthalate Esters by Groups

Relations between structure and effects of phthalate esters seem to exist, and it appears to be justified to group them according to the structure of the alcohol moiety, i.e., linear chain length and structure (linear or branched).

Short-Chain Phthalate Esters (C_1–C_3). Dermal exposure to dimethyl phthalate in humans occurs widely due to its use in insect repellants; so far, no significant toxicological side effects were observed [213]. Dimethyl and diethyl phthalate are used in cosmetics at concentrations of < 10 % and are dermally resorbed. They are considered to be nonirritants, nonsensitizers, and nonphototoxic agents [213]. Feeding studies in rats and mice with diethyl phthalate caused liver weight increases at high doses

($\geq$ 0.5 % in the diet), but no testicular toxicity [308 – 310, 345]. Peroxisome proliferation is only marginal and requires high doses [223]. Developmental toxicity is regarded as insignificant [359 – 361].

Dimethoxyethyl phthalate has a teratogenic potential [206 – 208] due to the released methoxyethanol.

Diallyl phthalate is acutely relatively toxic (LD_{50}, rat, oral: 800–1700 mg/kg) and irritating [215, 216]. In a long-term bioassay with gavage administration there was no clear carcinogenic activity, but allyl alcohol-related toxicity was observed (see Section 6.2.5.6).

Medium-Chain Phthalate Esters (C_4–C_7). Di(2-ethylhexyl)phthalate (DEHP) is the substance examined most extensively, and countless publications on toxicological profile, biochemistry, and metabolism exist. Far fewer data have been generated on any other phthalate ester.

In rodents, but not in other species, hepatic alteration (peroxisome proliferation and hepatomegaly [210, 227, 228, 317]) is observed. The peroxisome proliferation is somewhat more pronounced in the group of longer chain phthalates, especially of the branched-chain type.

Branched-chain phthalates commonly share a potential to induce peroxisomes and hepatomegaly in livers of rats and mice at high doses. This was demonstrated with diethylhexyl phthalate (DEHP), diisononyl phthalate and diisodecyl phthalate. These materials may cause peroxisome proliferation and hepatomegaly in rodents (DEHP at doses exceeding 50 mg kg^{-1} d^{-1}) [201, 210, 223 – 230, 232, 234, 278]. In long-term feeding experiments, DEHP [362] and related phthalates [210] also led to an increase of liver tumors in rodents, mediated by receptor PPARα. This mechanism is specific to rodents and not relevant to humans. The role of peroxisome proliferator activated receptors (PPARs) is outlined in Section 6.2.5.4. Nonrodents and primates have no or very low levels of PPARs and do not show these liver effects [189, 201, 290, 363, 364]. Hence no carcinogenicity is expected for humans, though some uncertainty remains because there might be receptor-independent effects (see Section 6.3.2).

Apart from the liver, effects were also seen in kidneys and the thyroid gland. Chronic administration of $\geq$ 0.5 % of DEHP in the diet also led to kidney toxicity in rats and mice with organ weight increase, cystic changes, dysfunctions [365 – 368], and alterations in thyroids [366, 369], with a fall in plasma thyroxine levels associated with hyperactivity and changes of the thyroidal colloid spaces. Effects on the thyroid gland were also observed with other peroxisome proliferators [369].

The straight-chain analogues di-*n*-hexyl and di-*n*-octyl phthalate also show thyroid effects although they differ in terms of hepatic toxicity [226, 366]. Di-*n*-octyl phthalate causes a different type of renal toxicity than DEHP [370].

Medium-chain phthalate esters exhibit degenerative testicular effects upon repeated oral administration in mice, rats, and hamsters [308 – 320]. The testicular effects are also obtained with DEHP (C_8, branched) to a similar degree but are not observed with the linear longer chain di-*n*-octyl phthalate or with short-chain phthalates. Testicular effects were noted with DEHP after oral intake in rats and mice, [234, 309, 313, 317, 318, 321] but were lacking in monkeys [199]. Oral uptake of $>$ 50 mg kg^{-1} d^{-1} of DEHP also caused embryotoxic and teratogenic effects in mice [337, 345, 349 – 353]. In rats, some embryotoxicity was observed in feeding studies at maternally toxic levels [352], but not upon inhalation [300].

Dibutyl phthalate (DBP) also caused embryo toxicity and teratogenicity at a concentration of 1 % in the diet of ICR/ICL mice [315, 349, 350]. After a 6-month inhalation exposure of rats to 50 mg DBP/m^3 for 6 h per day, an accumulation of DBP in the brain and increase of brain weight were found; at 5 mg/m^3 these effects were marginal [371].

Butyl benzyl phthalate at 2.5 and 5.0 % in the feed caused testicular effects, enlargement of liver and kidneys and decrease of bone marrow cellularity and thyroid function in rats [372].

Longer Chain Phthalate Esters (C_8–C_{12}). Di-*n*-octyl phthalate did not exert testicular toxicity in vivo [337, 341, 345] nor significant peroxisome proliferation in the liver [362]. It was shown to be a liver tumor promoter in a rat feeding experiment [373], probably as a sequel of fat accumulation and hepatomegaly [226].

The straight-chain DEHP analogues di-*n*-hexyl and di-*n*-octyl phthalate also show thyroid effects although they differ in terms of hepatic

toxicity [366]. Di-*n*-octyl phthalate causes a different type of renal toxicity than DEHP [370].

The effects of a group of high molecular weight phthalate esters (HMWPEs) comprising C_7–C_{12} alkyl backbone esters (di-2-propylheptyl ester; bis(C_7–C_9) branched and linear alkyl esters; di-C_{11} branched and linear alkyl esters; di-C_{11} alkyl ester; bis(C_{11}–C_{14}) branched and linear alkyl esters, C_{13}-rich; di-C_{13} alkyl ester) have recently been assessed as a category of related chemicals in the course of the global OECD-ICCA-HPV program [374]

HMWPEs have a low order of acute toxicity by the inhalation, dermal, intraperitoneal, and oral routes of exposure. Chemicals of the HMWPE category are not irritating to the skin or eyes (only slight conjunctival irritation for di-C_{13} alkyl ester), nor are they skin sensitizers (maximization test or comparable, or Buehler method). Although some data for these end points are older, the weight of evidence is consistent.

The primary findings in the repeated-dose rat studies were in the liver and kidney and to a lesser degree in the thyroid. Effects on the liver are indicative of peroxisomal proliferation, including increased PCoA, liver weights, and liver hypertrophy, and are not relevant to humans. It has been shown that these effects are mediated through peroxisome proliferation activated receptor alpha (PPARα) and that levels of PPARα are much higher in rodents than humans. Thus, one would expect humans to be substantially less responsive than rodents to peroxisome-proliferating agents. Empirical evidence for this hypothesis has been provided by studies in primates, in which repeated administration of DINP had no effects on liver, kidney, or testicular parameters, including peroxisome proliferation. The kidney effects were a result of a dose-dependent α-2μ-globulin nephropathy. Such effects are sex- and species-specific to male rats and also are not relevant to humans. Thyroid effects are likely to be a compensatory effect associated with the peroxisomal proliferation in the liver. The results were consistent for all members of the category, with NOAELs ranging between 10 and 282 mg^{-1} kg d^{-1}. The 10 mg^{-1} kg d^{-1} value was delivered in an OECD 422 study in which rats were dosed for 45 d, and the effect observed at 50 mg^{-1} kg d^{-1} was increased liver weight. This is likely to be related to peroxisome proliferation. All the NOAELs are driven by liver and/or kidney effects.

HMWPE category members are nongenotoxic. All of them have been tested in the Ames reverse mutation assay using *Salmonella typhimurium* and all were nonmutagenic with or without metabolic activation. Similarly, a range of substances covering the majority of the carbon numbers in this category was found to be inactive in the mouse lymphoma tests. Additional testing of di-C_{13} phthalate ester showed that the test substance did not induce either structural chromosomal aberrations or polyploidy in Chinese hamster liver cells up to the limit concentration of 4.75 mg/mL, in the absence or presence of an exogenous metabolic activation system.

Although HMWPEs have not been tested for carcinogenic properties (i.e., chronic toxicity or bioassay studies), previous experience with a wide range of phthalates including di-*n*-octylphthalate and DINP suggests that high doses might produce liver changes in rodents via interaction with PPARα. However, these are considered to be not relevant to humans and not indicative of a potential human risk. Three chronic toxicity/carcinogenicity studies on DINP have been conducted: two in rats and one in the mouse. In the rat studies, the major findings were liver and kidney changes principally related to the induction of peroxisome proliferation. There was an increase in liver tumors in both male and female rats and also a small increase in kidney tumors in the male rats. Both of these tumors are considered to be rat-specific and without relevance to humans. In the mouse study, there were liver tumors as well, also the consequence of peroxisomal proliferation, but no tumors of other types. It should be noted that di-*n*-octyl PE was a liver tumor promoter in a rat feeding study [373], probably as a sequel of fat accumulation and hepatomegaly [226].

Although not all members of the category have been tested for reproductive toxicity (di-2-propylheptyl, PE, di-C_{11} PE, or di-C_{13} PE), there are data for the lower [bis(C_7–C_9) PE], intermediate [bis(C_9–C_{11}) PE], and higher (di-C_{13} PE) molecular weight representatives indicate no significant reproductive toxicity at doses up to 500 mg^{-1} kg d^{-1} (or 250 mg^{-1} kg d^{-1} di-C_{13} PE). Effects included transiently decreased body weights or slightly decreased ovary and epididymidal weights. These effects are minor

and are not directly related to reproductive toxicity. Furthermore, the category members bis(C_7–C_9) PE and bis(C_9–C_{11}) PE have been recently shown not to be associated with detectable reproductive effects and do not affect fertility, similarly to DINP and DIDP.

Data from the developmental toxicity tests for the HMWPE conducted in rats on di-2-propylheptyl, bis(C_7–C_9), bis(C_9–C_{11}), and di-C_{13} PEs, have shown minimal maternal toxicity at doses up to 1000 mg^{-1} kg d^{-1} (limit dose) or 250 mg^{-1} kg d^{-1} (di-C_{13} PE). Either no effects were produced or the effects were associated with decreased food consumption and body weight loss in dams. Only the di-2-propylheptyl PE showed maternal toxicity at the limit dose and associated effects of resorption, decreased litter size, or fetal survival associated with the above two symptoms of maternal toxicity. In the two-generation study on DIDP, a decrease in offspring survival, more marked in F2, was observed. In the di-C_{13} PE one-generation study (F1 generation) a decrease in survival indices was observed, leading to a NOAEL of 50 mg^{-1} kg d^{-1} for offspring rats, whereas NOAELs for parental rats were 250 mg^{-1} kg d^{-1}. These may be considered related to developmental effects. No such changes were seen in either generation of the separate studies on bis(C_7–C_9) and bis(C_9–C_{11}) PE; these results are consistent with those from DIDP studies. However, none of the HMWPEs tested produced developmental effects. Increased frequencies of developmental variants including dilated renal pelvis and supernumerary lumbar ribs were produced in the studies on bis(C_7–C_9) PE and bis(C_9–C_{11}) PE, but are common findings in rats. Although not all members of the HMWPE Category have been tested for developmental toxicity, there are data for the lower [di-2-propylheptyl and bis(C_9–C_{11}) PE], intermediate [bis(C_9–C_{11}) PE], and higher (di-C_{13}) molecular weight representatives. Like DINP and DIDP, they have shown no significant developmental toxicity. It is reasonable to conclude that other members of the category would behave similarly, as shown by the weight of evidence. Thus, it can be concluded that this category of HMWPE induces no biologically significant developmental effects in rodents.

In conclusion, the weight of evidence shows that the HMWPEs have a low acute and subchronic toxicity. They are not irritating to the skin or eyes. They are not skin sensitizers. They are not mutagenic. No or only minimal developmental toxicity and no adverse effects on reproductive capability have been observed in rodent studies. Thus, there is minimal concern about these PEs resulting in reproductive toxicity in humans. Although not tested for carcinogenicity, the members of this category do not show the potential for producing genetic effects. Also, the same mechanism of action through peroxisome proliferation can be anticipated for induction of liver tumors in rodents, and this is presumed not to be relevant to humans [374].

6.3. Risk Assessment

Extensive toxicological data on the effects of phthalic acid intermediates and on most of the phthalate esters have been collected and are available to the public [374, 375]. Final EU Risk Assessment Reports are also available for dibutyl-, diisononyl-, and diisodecyl phthalate [375]. In contrast, only draft versions exist for butyl benzyl phthalate and DEHP [375] which are still under discussion because of uncertainties for some key elements, which included the human exposure situation, carcinogenicity, toxicity to reproduction, and implications of species differences.

Carcinogenicity. IARC evaluated the short- and long-term studies regarding genotoxicity and carcinogenicity [376]. It was concluded that there is sufficient evidence in experimental animals for the carcinogenicity of DEHP, but that there is inadequate evidence in humans for the carcinogenicity of DEHP. DEHP was therefore not classifiable as to its carcinogenicity to humans (Group 3).

In making its overall evaluation of the carcinogenicity to humans of DEHP, it was taken into consideration that:

- DEHP produces liver tumors in rats and mice by a non-DNA-reactive mechanism involving peroxisome proliferation.
- Peroxisome proliferation and hepatocellular proliferation have been demonstrated under the conditions of the carcinogenicity studies of DEHP in rats and mice.
- Peroxisome proliferation has not been documented in human hepatocyte cultures exposed

to DEHP or in the liver of exposed nonhuman primates.

Therefore, the mechanism by which DEHP increases the incidence of hepatocellular tumors in rats and mice is not relevant to humans [376]. Similar results were obtained for the other phthalate esters. Consequently, none of these is classified as being carcinogenic to humans.

The overall evaluation is widely accepted [274, 343]. It was, however, pointed out that there are some effects of DEHP that are not receptor-mediated, that peroxisome proliferation is not established as an obligatory step in the carcinogenesis of DEHP, and that therefore the species differences are not viewed as a valid hypothesis for the noncarcinogenicity to humans. It was argued that, independent from peroxisome proliferation and PPARα, DEHP can affect multiple signaling pathways by transcriptional activation of PPARα-regulated genes (suppression of apoptosis, protooncogene experession) and can induce biological effects (morphological cell transformation, decreased levels of gap junction intercellular communication). All these effects could contribute to carcinogenicity, and the precise mechanism of carcinogenic action remains to be determined [246, 274].

Toxicity to Reproduction. Several of the phthalate esters are classified as reproduction toxicants according to European legislation (di-*n*-butyl, di-*n*-pentyl, *n*-pentyl isopentyl, diisopentyl, butyl benzyl, di-2-ethylhexyl, and di-2-methoxyethyl phthalate); however, none is listed as being mutagenic or carcinogenic [377, 401].

6.3.1. Biomonitoring and Human Exposure to Phthalate Esters

Urinary excretion of phthalate ester metabolites (such as monoesters in conjugated or unconjugated form) may be taken as an indirect measure for intestinal absorption. After in vitro hydrolysis of all metabolites phthalic acid may be taken as sum parameter. Urinary MEHP was extensively used in animal and human studies for DEHP exposure estimates. However, there were uncertainties, as during the complex DEHP metabolism MEHP is further metabolized. This suggests that MEHP in blood or urine underestimates the exposure to DEHP, and also that the sum of MEHP and secondary urinary metabolites, namely, 5-OH- and 5-oxo-MEHP, provides a better exposure estimate [188, 203, 378 – 380]. This is widely acknowledged, but debate is ongoing as uncertainties remain regarding the value of earlier studies where the oxidized MEHP metabolites were not measured. This issue was treated among others by the NTP-CERHR Expert Panel during the evaluation of DEHP-related risks to human reproduction in October 2005 [344]. Since the first NTP-CERHR report on DEHP identified data gaps [342] substantial progress was made regarding biomonitoring of phthalate esters in the general population [382].

In 85 urban Germans aged 7–34 years urinary MEHP predicted a median DEHP intake of 13.8 μg kg^{-1} d^{-1} (95th percentile: 38.3); inclusion of the secondary MEHP metabolites gave a median DEHP intake of 10.3 μg kg^{-1} d^{-1} (95th percentile: 52.1) [379, 380].

Urinary MEHP and, for the first time, also 5-OH- and 5-oxo-MEHP, were measured in an U.S. National Health and Nutritional Examination Survey (NHANES) for the period 2001/2002 that included 2782 individuals. It was noted that MEHP levels were comparable to previous reports for U.S. residents [383]. The levels of the secondary metabolites were roughly 5–10-fold higher than MEHP [384]. Thus, they were similar to or up to twofold higher than in German adults and children [379].

Exposure estimates are always higher in children than in adults and highest in medically treated infants, due to the high dose and their low body weights. More recent estimates indicate that in adults and children the exposure to different phthalates is comparable, with the exception of the use of medical devices in children reaching enormous doses. Though doses are higher in children than in adults, the more recent estimates are far lower than those used previously by the NTP-CERHR Expert Panel in exception of the use of DEHP in medical devices 2000. Based on the recent estimates and refined NOAELs the margins of safety (MOS) for all scenarios in Table 9 were reported to be $> 10\,000$, with the exception of use DEHP in medical devices (MOS ≈ 5) [382].

High exposure may result from DEHP leaking out of medical devices during medical treat-

Table 9. Phthalate ester exposure estimates [342, 382]

Phthalate	Exposed group	Exposure estimate [342] in μg kg^{-1} d^{-1}	More recent exposure estimates [382] in μg kg^{-1} d^{-1}
Dibutyl phthalate	adult	2–10	0.86 (3.86) [b]
	young child	not provided	2.65
Butyl benzyl phthalate	adult	2	0.43 (2.08)[b]
	young child	≤ 6	1.64
DEHP	adult	3–30	0.61 (3.51)[b]
	young child (healthy)	10–20	2.57
DEHP, medical-device use	critically ill infant	1800–3300	12 000
Diisononyl phthalate	adult	$<$ 3–30	$<$ LOD (0.73)[ab]
	children using toys	mean $<$ 320	2.91 (10.71)
Diisodecyl phthalate	adult	$<$ 3–30	best analog: diisononyl phthalate
	children	$>$ 3–30	

[a] LOD = Level of detection.
[b] Values are in the format: mean value (95th percentile).

ment of patients and parenteral nutrition of neonates [382, 385, 386]. Estimates are generally far below 1 mg^{-1} kg d^{-1} in adults, but may reach values of 1–8 mg^{-1} kg d^{-1}during extracorporeal membrane oxygenation (ECMO) and large-volume blood transfusions. On the other hand, the plasticizer most commonly employed for medical use, DEHP, and its corresponding monoester MEHP appear to stabilize the membranes of stored erythrocytes [387]. In neonates, dose estimates increase in the order: parenteral nutrition without lipids (0.03 mg kg^{-1} d^{-1}) and with lipids (2.5 mg kg^{-1} d^{-1}) $<$ ECMO (14 mg kg^{-1} d^{-1}) $<$ exchange transfusions (22 mg^{-1} kg d^{-1}) [388]. These exposures are far beyond the TDI value of 50 μg kg^{-1} d^{-1} [389], but only recently, a promising substitute has become available, that lacks toxicity to reproduction and testes [386]: diisononyl cyclohexane-1,2-dicarboxylate (DINCH). DINCH has been successfully introduced into the market and a name plate production capacity of 100 000 tons is currently under construction (BASF AG, personal communication).

In conclusion, the occupational exposure levels resulting from handling and workplace exposure are generally low and well controlled. Exposure estimates for subgroups of the general public covers a wide range for all of the phthalate esters, but more recent data indicate that the exposure levels are much lower than was assumed during the last evaluation of toxicity to reproduction in the year 2000. Only under special circumstances such as parenteral exposure (e.g., eluates from medical devices) may the systemic intake be considerable and of potential toxicological relevance [205, 342, 382, 390, 391]. Dermal resorption has been measured in vitro [341] and in vivo [392] for several phthalate esters and appears to decrease with increasing chain length. The contribution of this route to the total exposure to phthalate esters is considered to be minor compared to the oral and parenteral routes. The significantly enhanced knowledge of exposure situations will improve the assessment of the risks of reproduction toxicity, and it is reasonably expected that the MOS are satisfactory in most scenarios, with the exception of phthalate ester uptake from medical devices. An alternative to DEHP lacking toxicity, DINCH, has only recently been developed and is suited to solve the exposure problems associated with intense medical treatment of patients and neonates.

6.3.2. Carcinogenicity

Hepatomegaly, peroxisome proliferation and related liver effects in rodents are described in Section 6.2.5.4. DEHP exerted a much lower extent of peroxisome proliferation [229] in hamsters. Hepatocytes of guinea pigs, ferrets, marmosets, and humans did not respond with peroxisome proliferation [189, 363]. Marmosets did not show peroxisome proliferation after oral administration of DEHP (2000 mg kg^{-1} d^{-1} for 14 d), whereas in a parallel rat experiment marked peroxisome proliferation was observed [201]. Peroxisome proliferation was also not observed in cynomolgus monkeys which were dosed with 500 mg kg^{-1} d^{-1} of DEHP for 3 weeks [393].

Several reasons for this species difference are assumed: The primary metabolite, MEHP

(see Section 6.2.5.1), undergoes much less glucuronidation in rats. Therefore, renal excretion is much slower in rats, and tissue levels are higher. This has been demonstrated in experiments with rats and monkeys [201, 393]. Furthermore, it was shown that in rats, due to saturation of a final β-oxidation step in the ω-oxidation process, a secondary metabolic pathway (ω-1 oxidation) is induced to a far greater extent, leading to the 5-keto derivate of MEHP, which was shown to be the ultimate peroxisome proliferator [393]. In addition to these kinetic differences, there may also be intrinsic toxicodynamic differences among the species. The overall evidence indicates that human hepatocytes either do not respond to peroxisome proliferators or do so only to a much lesser extent [291]. The conclusion is that, in the case of phthalates, rodent experiments overestimate the risks of chronic toxicity to humans.

6.3.3. Toxicity to Reproduction

DEHP is the model substance to investigate adverse effects and end points in rodents, in primates, and in vitro. Evidently DEHP is a reproduction developmental and developmental toxicant in rodents, but not in primates, that exerts adverse effects on the male reproductive organs and reduces sperm parameters, reduces fertility (increased postimplantation loss) at elevated dose, exerts developmental toxicity to the progeny as substantiated by decreased viability and adverse effects on reproductive parameters of male progeny (adverse effects on reproductive organs, sperm parameters). The underlying mechanisms need further elucidation; toxicity of metabolites to Sertoli cells and interaction with sex hormone regulation are clearly involved, potentially via zinc depletion or receptor interaction. Species differences exist: primates are much more resistant than rodents.

However, mechanisms, potential threshold levels, and species differences are under intense investigation. An update of the evaluation by the CERHR Expert Panel in 2000 [342] was discussed in October 2005 and the final version of the update Expert Panel Report is available since November 2005 [344]. The database is much weaker for other phthalates, but for managing risks the precautionary principle was applied, i.e. many of the short-, intermediate-, and long-chain phthalate esters are considered to act through similar mechanisms as does DEHP, and lead to the same adverse effects.

6.4. Risk Management

Phthalates have been under regulatory scrutiny for more than 20 years because of questions about carcinogenicity and reproductive toxicity, and this has led to some regulatory measures for worker and consumer protection.

In Germany permissible exposure levels (PEL) existed for phthalic anhydride (1 mg/m^3) and 11 phthalate esters. The values stemmed from several European countries and ranged from 3 mg/m^3 (e.g., diethyl, butyl benzyl, dibenzyl, diisodecyl phthalate) and 5 mg/m^3 (e.g., diallyl, dicyclohexyl, diheptyl, dioctyl, dinonyl phthalate) to 10 mg/m^3 for DEHP [394]. However, the scientific basis of the PEL values was regarded to be insufficient, and consequently the PEL values were abolished with the exception of DEHP, which is currently the only phthalate for which a scientifically derived AGW[a] value of 10 mg/m^3 (short-term value: 8-fold) exists. Work is ongoing to establish AGW values for a number of phthalates for which the scientific basis is regarded to be sufficient. These are dicyclohexyl, diheptyl, dioctyl (excluding di-*n*-octyl and DEHP), and dinonyl phthalate, as well as isophthalic and terephthalic acids [395]. The MAK list gives additionally MAK values for isophthalic and terephthalic acids besides DEHP, but these values are not binding. The data basis for phthalic acid, phthalic anhydride, and diallyl phthalate was regarded as being insufficient for deriving an MAK value [396]. Inhalation of vapors, mists, and aerosols is regulated in the USA by TLVs (TWA 5 mg/m^3, STEL 10 mg/m^3 for DEHP, dibutyl phthalate, and dimethyl phthalate).

Regulatory scrutiny led to the voluntary removal of DEHP from children's teethers and toys in 1986, and subsequent pressure to remove others. Since December 1999, the EU has prohibited toys containing more than 0.1 % of six phthalates (di-*n*-butyl, butyl benzyl, di-2-ethylhexyl, dioctyl, diisononyl, and diisodecyl phthalate) that are intendend to be placed in the mouth of children under three years of age

[397]. This temporary ban was strengthened in July 2005 in that di-*n*-butyl, butyl benzyl, and di-2-ethylhexyl phthalate are prohibited in all toys and childcare articles; for the sake of the precautionary principle the regulation remained unchanged for dioctyl, diisononyl, and diisodecyl phthalate [398]. The more recent Directive 2005/84/EC prohibits the use of plasticized materials containing more than 0.1 % of DEHP, DBP, or BBP in toys and childcare articles; or more than 0.1 % of DINP, DIDP, or DNOP in toys or childcare articles which can be placed in the mouth by children. This new regulation was adopted by the EC Member States during 2006, and is effective since January 2007 [399].

There is a general call for substitution of hazardous materials by less hazardous or nonhazardous alternatives. Several applications have been identified or are under investigation in which DEHP could be substituted by less toxic plasticizers, e.g., diisononyl, diisodecyl, or di-2-propylheptyl phthalate, adipates, citrates, alkylated sulfonic esters, trimellitates (in which a third ester group is attached to the benzene ring of phthalates), terephthalates, and cyclohexane-1,2-dicarboxylic esters (DINCH) [400], but there may be limitations that are of technical (e.g., stability), economical (e.g., increased costs), or toxicological and ecological nature (e.g., unknown effect profiles of new substances). Nonetheless, research efforts already led to very promising new solutions and products, notably the substitution of DEHP by DINCH in medical-care products [386].

The former MAK and TRK-values are all abolished and will be stepwise replaced by science-focused AENs (= Arbeitsplatzgrenzwerte)

7. References

1. R. C. Weast: *Handbook of Chemistry and Physics*, 60th ed., CRC Press, Boca Raton, Fla., 1979/1980 , C 436.
2. *Handbuch des Chemikers*, VEB-Verlag Technik, Berlin 1956.
3. *Kirk-Othmer*, 2nd ed., **15**, 445.
4. H. Suter: *Phthalsäureanhydrid und seine Verwendung*, Dr. Dietrich Steinkopf-Verlag, Darmstadt 1972.
5. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe*, 2nd ed., Deutscher Eichverlag, 1970.
6. R. H. Perry: *Chemical Engineer's Handbook*, 6th ed., McGraw-Hill, New York 1973.
7. BASF, unpublished results, Ludwigshafen 1989.
8. B. Raeke, *Angew. Chem.* **70** (1958) 1.
9. K. Weissermel, H.-J. Arpe: *Industrielle organische Chemie*, 4th ed., VCH Verlagsgesellschaft, Weinheim, Germany 1994, p. 415.
10. F. Wirth, *Hydrocarbon Process.* **54** (1975) 107.
11. BASF, DE 2 356 049, 1973 (F. Wirth et al.).
12. BASF, DE 1 266 748, 1966 (G. Schaefer et al.).
13. BASF company brochure: "Phthalic Anhydride Advanced Catalytic PA-Technology," 1995.
14. *Hydrocarbon Process.* **76** (1997) no. 3, 146.
15. *Europa Chemie* (1997) no. 4, 12.
16. O. Wiedemann et al., *Chem. Eng. (N.Y.)* **86** (1979) 62.
17. Nippon Shokubai Kagaku Kogyo, DE 2 948 163, 1978 (T. Sato et al.).
18. Nippon Shokubai Kagaku Kogyo, US 4 284 571, 1978 (T. Sato et al.).
19. T. Sato et al., *Hydrocarbon Process.* **62** (1983) no. 10, 107.
20. Nippon Shokubai Kagaku Kogyo, US 4 046 780, 1975 (Y. Nakanishi et al.).
21. Nippon Shokubai Kagaku Kogyo, US 4 356 112, 1976 (Y. Nakanishi et al.).
22. L. Verde, A. Neri, *Hydrocarbon Process.* **63** (1984) no. 11, 83.
23. *Hydrocarbon Process.* **64** (1985) no. 11, 156.
24. Alusuisse Italia S.P.A., EP 37 492, 1980 (A. Neri et al.).
25. Alusuisse Italia S.P.A., US 4 489 204, 1980 (A. Neri et al.).
26. *Hydrocarbon Process.* **59** (1979) no. 11, 208.
27. Rhône Progil, DE 2 416 457, 1973 (R. Dumont et al.).
28. BASF, DE 2 510 994, 1975 (K. Blechschmitt et al.).
29. BASF, DE 2 547 624, 1975 (K. Blechschmitt et al.).
30. BASF, DE 2 914 683, 1979 (P. Reuter et al.).
31. Nippon Shokubai Kagaku Kogyo, DE 3 045 624, 1979 (Y. Nakanishi et al.).
32. Nippon Shokubai Kagaku Kogyo, JP 56 73 543, 1981.
33. Nippon Shokubai Kagaku Kogyo, JP 57 105 241, 1982.

34. Nippon Shokubai Kagaku Kogyo, JP 57 180 430, 1982.
35. Mitsui Toatsu Chemicals Inc., JP 54 5 891, 1979 (K. Oshima).
36. BASF, DE 2 546 268, 1975 (K. Blechschmitt et al.).
37. N. G. Glukhovskii et al., SU 978 910, 1981; *Chem. Abstr.* **98** (21): 178 962 b.
38. D. W. B. Westerman et al., *Appl. Catal.* **3** (1982) 151.
39. M. S. Wainwright et al., *Can. J. Chem. Eng.* **55** (1977) 557.
40. W. Fiebig et al., *Chem. Tech. (Leipzig)* **28** (1976) 673.
41. D. Vankove et al., *J. Catal.* **36** (1975) 6.
42. J. Herten, G. F. Froment, *Ind. Eng. Chem. Process Des. Dev.* **7** (1968) 516.
43. P. H. Calderbank, *Chem. Eng. Sci.* **32** (1977) 1435.
44. M. S. Wainwright et al., *Catal. Rev. Sci. Eng.* **19** (1979) 211.
45. D. D. McLean et al., *Can. J. Chem. Eng.* **58** (1980) 608.
46. A. G. Lyubarksii et al., *Kinet. Katal.* **14** (1973) 956.
47. K. Hertwig et al., *Chem. Tech. (Leipzig)* **23** (1971) 584.
48. J. Skrzypek et al., *Chem. Eng. Sci.* **40** (1985) 611.
49. M. Blanchard, D. Vanhove, *Bull. Soc. Chim. Fr.* 1971, 3291.
50. R. Römer et al., *Chem. Ing. Tech.* **46** (1974) 653.
51. G. Emig et al., *Chem. Ing. Tech.* **47** (1975) 717, 888, 1012.
52. I. E. Wachs et al., *Stnd. Surf. Sci. Catal.* **19** (1984) 275.
53. I. E. Wachs et al., *Appl. Catal.* **15** (1985) 339.
54. R. Saleh et al., *Appl. Catal.* **31** (1987) 87.
55. V. Nikolov et al., *Stnd. Surf. Sci. Catal.* **34** (1987) 173.
56. I. E. Wachs et al., *J. Catal.* **91** (1985) 366.
57. G. F. Froment et al., *Appl. Catal. A* **120** (1994) 17.
58. G. F. Froment et al., *Chem. Eng. Scie.* **51** (1996) 2091.
59. G. C. Bond et al., *J. Catal.* **164** (1996) 276.
60. N. L. Franklin et al., *Trans. Inst. Chem. Eng.* **34** (1956) 280.
61. F. de Maria et al., *Ind. Eng. Chem.* **53** (1961) 259.
62. F. Bernardini et al., *Chim. Ind. (Milan)* **48** (1966) 9.
63. A. Farkas, *Hydrocarbon Process.* **49** (1970) 121.
64. Davy Powergas, DE 2 602 895, 1976 (L. Hellmer et al.).
65. R. M. Dow et al., *Hydrocarbon Process.* **56** (1977) 167.
66. *Chem. Eng.* **84** (1977) 65.
67. BASF, DE 1 601 162, 1967 (F. Lorenz et al.).
68. Deggendorfer Werft und Eisenbau, DE 2 207 166, 1972 (O. Wanka et al.).
69. Deggendorfer Werft und Eisenbau, DE 2 543 758, 1975 (R. Vogl).
70. Davy McKee AG, DE 2 855 629, 1978 (G. Keunecke et al.).
71. Davy McKee AG, DE 2 855 630, 1978 (G. Keunecke et al.).
72. BASF, DE 3 207 208, 1982 (E. Danz et al.).
73. BASF, DE 3 007 627, 1980 (W. Muff).
74. Davy McKee AG, DE 3 339 392, 1983 (H. Saffran).
75. Metallgesellschaft AG, DE 3 411 732, 1984 (V. Franz et al.).
76. GEA Luftkühler GmbH, DE 3 512 903, 1985 (W. Rudowski).
77. Metallgesellschaft AG, DE 1 227 443, 1964 (W. Scheiber).
78. Pittsburgh Coke and Chemical Company, US 2 850 440, 1955 (M. O. Shrader et al.).
79. Mitsui Toatsu Chem., DE 1 948 374, 1968 (K. Hayakawa et al.).
80. Japan Gas Chem., JP 45/10 333, 1970.
81. Rhône-Progil, DE 2 452 171, 1976.
82. Nippon Shokubai Kagaku Kogyo, JP-Kokai 58 8074, 1981.
83. Nippon Steel Chem., JP-Kokai 58 45 434, 1977.
84. Nippon Shokubai Kagaku Kogyo, DE 3 225 079, 1981 (Y. Kita et al.).
85. Chemische Werke Hüls AG, EP 120 199, 1983 (F. Gude et al.).
86. Chemische Werke Hüls AG, DE 3 309 310, 1983 (F. Gude et al.).
87. BASF, US 3 507 886, 1966 (H. Suter et al.).
88. Chemiebau Dr. A. Zieren GmbH, US 3 655 521, 1968 (H. Gehrken et al.).
89. Nippon Steel Chem. Co, JP-Kokai 57 49 549, 1977.
90. Nippon Steel Chem. Co, JP-Kokai 58 45 435, 1977.
91. *Hydrocarbon Process.* **46** (1967) 215.
92. The Badger Company Inc., US 4 435 580, 1982 (C. D. Miserlis).
93. The Badger Company Inc., US 4 435 581, 1982 (C. D. Miserlis).
94. Kawasaki Steel Corp., JP-Kokai 63 141 645, 1986.

95. Kawasaki Steel Corp., US 5 252 752, 1991 (T. Aona et al.).
96. Kawasaki Steel Corp., JP 61 90 281, 1994 (Y. Asami et al.).
97. Kawasaki Steel Corp., EP 451 614, 1991 (T. Aona et al.).
98. Shokubai Kasei Kogyo, JP 80 99 040, 1994.
99. Kawasaki Steel Corp., JP 57 32 096, 1994.
100. Chemische Werke Hüls AG, DE 1 643 827, 1967 (F. List et al.).
101. Standard Oil Company, US 4 215 051, 1979 (H. Schroeder et al.); US 4 215 052, 1979 (H. Schroeder et al.); US 4 215 053, 1979 (D. A. Palmer et al.); US 4 215 054, 1979 (H. Schroeder et al.).; US 4 215 055, 1979 (D. A. Palmer et al.); US 4 215 056, 1979 (H. Schroeder et al.).
102. Sisas, EP 256 352, 1987 (F. Celeste et al.).
103. Mitsubishi Gas Chemical Co, JP 53/40 711, 1978.
104. V. N. Orlik, *Khim. Tekhnol.* **6** (1981) 51.
105. S. Slavov, *Khim. Ind. (Sofia)* **56** (1984) 252.
106. S. Slavov, *Geterog. Katal.* **1** (1983) 81.
107. Chem. Sys. Inc., Phthalic Anydride, 93–3, 05. 1995.
108. BASF, DE 2 902 978, 1979 (G. Kilpper et al.).
109. BASF, EP 4635, 1978 (G. Kilpper et al.).
110. *Beilstein* **E II 21** (1953) 348.
111. LENTIA GmbH, DE 2 056 891, 1970 (J. Schweighofer).
112. BASF, DE 2 334 379, 1973 (E. Hetzel et al.); DE 2 334 916, 1973 (E. Hetzel et al.); DE 2 911 245, 1979 (G. Kilpper et al.).
113. O. Joklik, AT 286 271, 1969.
114. Dawe's Lab Inc., US 3 819 648, 1971 (W. R. Boehme et al.).
115. M. C. Sze et al., *Hydrocarbon Process.* **55** (1976) no. 2, 103.
116. Sun Res. + Develop. Co, DE 2 256 596, 1972 (R. D. Bushick et al.).
117. Degussa, DE 1 948 714, 1969 (T. Lüssling et al.).
118. Allied Chem. Corp., US 2 833 807, 1956 (A. Farkas et al.).
119. Mitsubishi Gas. Chem. Co, DE 2 427 191, 1979 (M. Saito et al.).
120. Institut Nefttecnimitscheskich USSR, DE 2 627 068, 1979 (S. D. Mechtjev et al.).
121. BASF, EP 699 476, 1996 (L. Karrer et al.).
122. Takeda Yakuhin Kogyo, JP 55 17 360, 1980 (I. Onishi et al.).
123. BASF, EP 766 985, 1997 (L. Karrer et al.).
124. G. Stefani, *Chim. Ind.* **54** (1972) 984.
125. K. K. Moll et al., *Chem. Tech. (Leipzig)* **20** (1968) 600.
126. T. Kudo, *CEER Chem. Econ. Eng. Rev.* **2** (1970) 43.
127. BASF, DE 1 172 253, 1962. (H. Kröper et al.).
128. BASF, DE 1 643 630, 1967 (M. Decker et al.).
129. Lummus Corp., US 3 776 937, 1972 (A. P. Gelbein);
DE 2 008 648, 1970 (A. P. Gelbein).
130. *Ullmann*, 3rd ed., **13**, 733–735.
131. *Beilstein*, E III 9, 4199–4200.
132. H. Suter: *Phthalsäureanhydrid und seine Verwendung*, D. Steinkopff Verlag, Darmstadt 1972.
133. H. Rein, *Chemiker Taschenbuch*, 58th ed., Part 2, Springer-Verlag, Berlin 1937.
134. L. Meier: "Weichmacher für PVC," in G. W. Becker, D. Braun (eds.): *Kunststoff-Handbuch,* vol. 2/1: Polyvinylchlorid, Hanser-Verlag, München 1986.
135. Monsanto US 1 923 938, 1933 (L. P. Kyrides).
136. G. Lützel, J. Holzmann: "Ein neuer Weichmacher für PVC auf Basis DINP," *Kunststoffe* **74** (1984) no. 4, 1088–1091.
137. *Modern Plastics Encyclopedia*, McGraw-Hill, New York (1988).
138. US. International Trade Commission, Synthetic Organic Chemicals, US. Production and Sales,1987. US. Government Printing Office, Washington, 1987.
139. Bulletin S-107 B, Methane Sulfonic Acid, Pennwalt Corporation, Pa 19 102, USA.
140. TIL, Tioxide UK Ltd., Booklet TIL 12, Tilcom Catalysts For Ester Manufacture, Billingham 1982.
141. *Ullmann*, 4th ed. **18**, 534–538.
142. F. X. Werber, B. F. Goodrich, US 3 056 818, 1962.
143. Du Pont Company, Chemical & Pigment Dept., Booklet E-69 996–1 4/87 (2 M), Du Pont Performance Products, Wilmington, DE, 1987.
144. Dynamit Nobel, Technical Information, ZN 666/2.80/500/2439, Titanates as Catalysts for the manufacture of plasticizers, Troisdorf, 1980.
145. M & T, Chemicals Inc., How to make DOP and other Phthalate Plasticizers easier with FASCAT 2001, and M & T Technical Data Sheet No. 341, Rev. 9/78, and Material Safety Data Sheet 3/82 for Stannous Oxide, Rahway, NJ 1978.
146. Exxon Chemical International. A practical guide to protecting man and the environment, Kraainem, Belgium 1988.

147. BUA Report 22 on Di-butyl phthalate, VCH-Verlagsgesellschaft, Weinheim, Dec. 1987.
148. BUA Report 4 on Di-2-ethylhexyl phthalate, VCH-Verlagsgesellschaft, Weinheim, Jan. 1986.
149. Gemeinsames Ministerialblatt (GMBl), Bonn, Germany 1990.
150. Gemeinsames Ministerialblatt (GMBl), G 3191 A, Bonn, W. Germany, 28. Feb. 1986.
151. BASF, Technical Leaflet, Palatinol AH-L (med), Ludwigshafen, 1989.
152. B. L. Wadey, J. Holzmann, SAE Technical Paper Series, 870 318, Plasticizer for Automotive Interior Trim, SAE, 400 Commonwealth Drive Warrendale, PA 15 096, USA 1987.
153. Verordnung über Anlagen zur Lagerung, Abfüllung und Beförderung brennbarer Flüssigkeiten–VbF, Bonn, v. 27. Feb. 1980.
154. Regulations for water (Wasserhaushaltsgesetz-WHG) Bonn, last edition Sept. 23., 1986 (BGBl. I, S. 1529).
155. BASF, Hinweise zur Lagerung von Weichmachern für PVC, TI-CIW/ES 003 d, Ludwigshafen, Dec. 1987.
156. Verband Kunststofferzeugende Industrie, Argumente 3, Zum Thema Weichmacher, Frankfurt, 1988.
157. BASF, Marketing CIW, Ludwigshafen, 1988.
158. Registry of Toxic Effects of Chemical Substances (NIOSH; 1990), *Proc. Soc. Exp. Biol. Med.* **49** (1942) 471.
159. E. Sandmeyer, C. Kirwin, jr., in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology*, vol. **II A**, Wiley-Interscience, New York (1982) p. 2350.
160. M. Wernfors, J. Nielsen, A. Schütz, S. Skerfving: "Phthalic Anhydride-Induced Occupational Asthma," *Int. Arch. Allergy Appl. Immunol.* **79** (1986) 77–82.
161. J. Nielsen, H. Welinder, A. Schütz, S. Skerfving: "Specific Serum Antibodies Against Phthalic Anhydride in Occupationally Exposed Subjects," *J. Allergy Clin. Immunol.* **82** (1988) 126–133.
Alternative Dermal Sensitization Test: The Mouse Ear Swelling Test (MEST)," *Toxicol. Appl. Pharmacol.* **84** (1986) 93–114.
162. S. C. Gad et al.: "Development and Validation of an Alternative Dermal Sensitization Test: The Mouse Ear Swelling Test (MEST)," *Toxicol. Appl. Pharmacol.* **84** (1986) 93–114.
163. V. van Kampen, R. Merget, X. Baur: "Übersicht der atemwegssensibilisierenden und -irritativen Arbeitsstoffe", Arbeitsmed. Sozialmed. Umweltmed. 34 (1999) 232–247. http://www.bgfa.ruhr-uni-bochum.de/specials/irritativ.php.
164. Deutsche Forschungsgemeinschaft (DFG), Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe: "MAK- und BAT-Werte-Liste 2005" Mitteilung 4. Wiley-VCH, Weinheim, 2005.
165. BAYER AG, Institut für Toxikologie, unpublished observations, 1978.
166. Biofax Industrial Bio-Test Laboratories, Inc., Northbrook, Ill. 60 062 Data Sheets 13–4/70.
167. E. Gross, H. Friebel, Prüfung der Toxizität von reinem Phthalsäureanhydrid und Rohprodukten aus der industriellen Phthalsäuresynthese, unpublished results, Bayer, 1955.
168. E. Zeiger, S. Haworth, K. Mortelmans, W. Speck, "Mutagenicity Testing of Di(2-ethylhexyl) Phthalate and Related Chemicals in Salmonella," *Environm. Mutagen.* **7** (1985) 213–232.
169. I. Florin, L. Rutberg, M. Curvall, C. R. Enzell, "Screening of Tobacco Smoke Constituents for Mutagenicity Using the Ames Test," *Toxicol.* **15** (1980) 219–232.
170. NTP Bioassay of Phthalic Anhydride for Possible Carcinogenicity, Techn. Report 159, NCI, Bethesda, Md. 1979.
171. "NTP Technical Report on the Carcinogenesis Bioassays of Diallylphthalate in F 344/N Rats and B 6 C 3 F 1/N Mice," NTP-TR Report no. 284, NTP-81–083, US Department of Health and Human Services, 1985. NIH Publ. 85–2540 and 82–1798.
172. R. L. Dixon, G. E. Shull, S. Fabro, "Relative Teratogenicity of Selected Anhydrides," *Arch. Toxicol.* **39 Suppl. 1** (1978) 327 A.
173. N. A. Brown, G. E. Shull, R. L. Dixon, S. E. Fabro, "The Relationship between Acylating Ability and Teratogenicity of Selected Anhydrides and Imides," *Toxicol. Appl. Pharmacol.* **45** (1978) no. 361, 333 A.
174. S. Fabro, G. Shull, N. A. Brown, "The Relative Teratogenic Index and Teratogenic Potency: Proposed Components of Developmental Toxicity Risk Assessment," *Teratogen. Carcinogen. Mutagen.* **2** (1982) 61–76.
175. *Patty*, 3rd ed., **IIA**, 2710.
176. European Chemical Bureau (ECB). Phthalimide. IUCLID Dataset (2000). Available at http://ecb.jrc.it/existing-chemicals/.

177. Registry of Toxic Effects of Chemical Substances (NIOSH; 1990), *Farmaco Ed. Sci.* **20** (1965) 3.
178. Berufsgenossenschaft der chemischen Industrie (BG Chemie). *o*-Phthalodinitril. Toxicological Evaluation No. 28 (1995).
179. H. Zeller, H. T. Hofmann, A. M. Thiess, W. Hey, "Zur Toxizität der Nitrile," *Zentrallbl. Arbeitsmed. Arbeitsschutz* **19** (1969) 225.
180. A. M. Thiess, "Beobachtungen von Gesundheitsschädigungen und Vergiftungen durch Einwirkung von o-Phthalodinitril" *Zentrallbl. Arbeitsmed. Arbeitsschutz* **18** (1968) 303.
181. F. Oesch, Ames-Test for o-Phthalodinitrile, unpublished investigations BASF, Ludwigshafen, 1978.
182. H. G. Miltenburger: "Detection of Gene Mutations in Somatic Mammalian Cells in Culture: HGPRT-test with V 79 Cells; ortho-Phthalodinitrile," in *BUA-Ber.,* Benzoldicarbonitrile, VCH Verlagsgesellschaft, Weinheim, 1988.
183. G. B. Pliss, N. J. Wolfson: ber die leukämogene Wirkung des Phthalodinitrils (Russian). Vorposy Onkologii XVIII (1), 81–86, 1972 (cited in *BUA-Ber.,* Benzoldicarbonitrile, Dez. 1988, VCH, Weinheim, 1989).
184. P. W. Albro, R. O. Thomas: "Enzymatic Hydrolysis of Di(2-ethylhexyl) Phthalate by Lipases," *Biochim. Biophys. Acta* **360** (1973) 380–390.
185. J. W. Daniel, H. Bratt: "The Absorption, Metabolism and Tissue Distribution of Di(2-ethylhexyl) Phthalate in Rats," *Toxicology* **2** (1974) 51–65.
186. W. M. Kluwe: "Overview of Phthalate Ester Pharmacokinetics in Mammalian Species," *EPH Environ. Health Perspect.* **45** (1982) 3–10.
187. P. W. Albro, R. Thomas, L. Fishbein: "Metabolism of Diethylhexyl Phthalate by Rats. Isolation and Characterization of the Urinary Metabolites," *J. Chromatogr.* **76** (1973) 321–330.
188. H. M. Koch, R. Preuss, J. Angerer, "Di(2-ethylhexyl)phthalate (DEHP): Human Metabolism and Internal Exposure — An Update and Latest Results", *Int. J. Androl.* 2005.
189. A. M. Mitchell, J. C. Lhuguenot, J. W. Bridges, C. R. Elcombe: "Identification of the Proximate Peroxisome Proliferator(s) Derived from Di(2-ethylhexyl) Phthalate," *Toxicol. Appl. Pharmacol.* **80** (1985) 23–32.
190. M. Ema, T. Murai, T. Itami, H. Kawasaki: "Evaluation of the Teratogenic Potential of the Plasticizer Butyl Benzyl Phthalate in Rats," *JAT, J. Apll. Toxicol.* **10** (1990) 339–343.
191. WHO-JECFA: Safety evaluation of certain food additives and contaminants: Saturated aliphatic acyclic branched-chain primary alcohols, aldehydes, and acids. WHO Food Additives Series 40. Prepared by: The forty-ninth meeting of the Joint FAO/WHO Expert Committee on Food Additives (JECFA). 1st Draft, prepared by G. Semino. Geneva (1998).
192. M. J. Silva, K. Kato, E. L. Gray, C. Wolf, L. L. Needham, A. M. Calafat: "*Urinary Metabolites of Di-n-octylphathalate in Rats*", *Toxicology* **210** (2005) 123–133.
193. L. Laignelet, J. Lhuguenot: *Di-(2-ethylhexyl)phthalate (DEHP). Absorption, excretion, metabolism and pharmacokinetic profile in Wistar female rats*. Report No. 1/99. Dijon: Ecole Nationale Supérieure de Biologie Appliquée à la Nutrition et à l'Alimentation (ENSBANA), 2000.
194. L. Laignelet, J. Lhuguenot: *Di-(2-ethylhexyl)phthalate (DEHP). Absorption, excretion, metabolism and pharmacokinetic profile in pregnant Wistar rats*. Report No. 2/99. Dijon: Ecole Nationale Supérieure de Biologie Appliquée à la Nutrition et à l'Alimentation (ENSBANA), 2000.
195. L. Laignelet, J. Lhuguenot: *Di-(2-ethylhexyl)phthalate (DEHP). Absorption, excretion, metabolism and pharmacokinetic profile in pregnant CD1 mice*. Report No. 3/99. Dijon: Ecole Nationale Supérieure de Biologie Appliquée à la Nutrition et à l'Alimentation (ENSBANA), 2000.
196. L. Laignelet, J. Lhuguenot: *Di-(2-ethylhexyl)phthalate (DEHP). Absorption, excretion, metabolism and pharmacokinetic profile in female CD-1 mice*. Report No. 4/99. Dijon: Ecole Nationale Supérieure de Biologie Appliquée à la Nutrition et à l'Alimentation (ENSBANA), 2000.
197. L. Laignelet, J. Lhuguenot: *Di-(2-ethylhexyl)phthalate (DEHP). Absorption, excretion, metabolism and pharmacokinetic profile in pregnant and*

non-pregnant rats and mice. Synthesis of reports No. 1/99 to 4/99. Dijon: Ecole Nationale Supérieure de Biologie Appliquée à la Nutrition et à l'Alimentation (ENSBANA), 2001.

198. Mitsubishi Chemical Safety Institute: *Sixty-five week repeated oral dose toxicity study of di-(2-ethylhexyl)phthalate (DEHP) in juvenile common marmosets*, Ibaraki, Japan: Mitsubishi Chemical Safety Institute, 2003.
199. Y. Kurata, F. Makinodan, N. Shimamura, M. Okada, M. Katoh: "Metabolism of Di-(2-ethylhexyl) phthalate (DEHP) in Juvenile and Fetal Marmoset and Rat", *Toxicologist* **84** (2005) 1251.
200. W. Kessler, W. Numtip, K. Grote, G. A. Csanady, I. Chahoud, J. G. Filser: "Blood Burden of di(2-ethylhexyl) Phthalate and its Primary Metabolite Mono(2-ethylhexyl) Phthalate in Pregnant and Nonpregnant Rats and Marmosets", *Toxicol. Appl. Pharmacol.* **195** (2004) 142–153.
201. C. T. C. Rhodes et al.: "Comparative Pharmacokinetics and Subacute Toxicity of di(2-Ethylhexyl) Phthalate (DEHP) in Rats and Marmosets: Extrapolation of Effects in Rodents to Man," *EHP Environ. Health Perspect.* **65** (1986) 299–308.
202. P. Schmidt, C. Schlatter: "Excretion and Metabolism of Di(2-ethylhexyl) Phthalate in Man," *Xenobiotica* **15** (1985) 251–256.
203. H. M. Koch, H. M. Bolt, J. Angerer: "Di(2-ethylhexyl)phthalate (DEHP) Metabolites in Human Urine and Serum after a Single Oral Dose of Deuterium-Labelled DEHP", *Arch. Toxicol*. **78** (2004) 123–130.
204. P. W. Albro et al.: "Identification of the Metabolites of Di-(2-ethylhexyl) Phthalate in Urine from African Green Monkey," *Drug, Metab. Dispos.* **9** (1981) 23.
205. G. M. Pollack et al.: "Circulating Concentrations of Di(2-ethylhexyl) Phthalates and its De-esterified Phthalic Acid Products Following Plasticizer Exposure in Patients Receiving Hemodialysis," *Toxicol. Appl. Pharmacol.* **79** (1985) 257–267.
206. A. R. Singh, W. H. Lawrence, J. Autian: "Teratogenicity of Phthalate Esters in Rats," *J. Pharm. Sci.* **61** (1972) 51–55.
207. J. Yonemoto, N. A. Brown, M. Webb: "Effects of Dimethoxyethyl Phthalate, Monomethoxyethyl Phthalate, 2-Methoxyethanol and Methoxyacetic Acid on Post Implantation Rat Embryos in Culture," *Toxicol. Lett.* **21** (1984) 97–102.
208. E. J. Ritter, W. J. Scott, J. L. Randall: "Mechanistic Studies of the Teratogenic Action of Di(2-methoxyethyl) phthalate (DMEP) and 2-Methoxyethanol in Wistar Rats," *Teratology* **29** (1984) 54 A Suppl.
209. R. M. David, G. Gans: "Summary of Mammalian Toxicology and Health Effects of Phthalate Esters" in *The Handbook of Environmental Chemistry*, Vol. 3, Part Q, 2003, pp. 299–316.
210. J. W. Hirzy: "Carcinogenicity of General-Purpose Phthalates: Structure-Activity Relationships," *Drug Metab. Rev.* **21** (1989) 55–63.
211. D. Calley, J. Autina, W. L. Guess: "Toxicology of a Series of Phthalate Esters," *J. Pharm. Sci.* **55** (1966) 158–162.
212. W. H. Lawrence et al.: " A Toxicological Investigation of Some Acute, Short-Term, and Chronic Effects of Administering Di-2-Ethylhexyl Phthalate (DEHP) and Other Phthalate Esters," *Environ. Res.* **9** (1975) 1–11.
213. "Final Report on the Safety Assessment of Dibutyl Phthalate, Dimethyl Phthalate, and Diethyl Phthalate," *Am. Coll. Toxicol.* **4** (1985) 267–303.
214. G. Klecak, H. Geleik, J. R. Frey: "Screening of Fragrance Materials for Allergenicity in the Guinea Pig," *J. Soc. Cosmet. Chem.* **28** (1977) 53–64.
215. J. Autian: "Toxicity and Health Threats of Phthalate Esters: Review of Literature," *EHP Environ. Health Perspect.* **4** (1973) 3.
216. D. E. Carter, B. Feldmann, J. G. Sipes: "Liver and Lung Toxicity of Diallyl Phthalate," *Toxicol. Appl. Pharmacol.* **45 Abstr. 81** (1978) 254.
217. R. Z. Christiansen, H. Osmundsen, B. Borrebaek, J. Bremer: "The effects of Clofibrate Feeding on the Metabolism of Palmitate and Erucate in Isolated Hepatocytes," *Lipids* **13** (1978) 487–419.
218. C. R. Elcombe, A. M. Mitchell: "Peroxisome Proliferation due to Di-(2-ethylhexyl) Phthalate: Species Differences and Possible Mechanisms," *EPH Environ. Health Perspect.* **70** (1986) 211–219.
219. G. P. Mannaerts, L. J. Debeer, J. Thomas, P. J. Schepper: "Mitochondrial and Peroxisomal Fatty Acid Oxidation in Liver Homogenates and Isolated Hepatocytes from Control and Clofibrate-Treated Rats," *J. Biol. Chem.* **254** (1979) 4585–4595.

220. E. A. Lock, A. M. Mitchell, C. R. Elcombe: "Biochemical Mechanisms of Induction of Hepatic Peroxisome Proliferation," *Annu. Rev. Pharmacol. Toxicol.* **29** (1989) 145–163.
221. P. I. Eacho et al.: "Hepatic Peroxisomal Changes Induced by a Tetrazole-Substituted Alkoxyacetophenone in Rats and Comparison with other Species," *Toxicol. Appl. Pharmacol.* **83** (1986) 430–437.
222. S. P. Foxworthy, D. N. Perry, D. M. Hoover, P. I. Eacho: "Changes in Hepatic Lipid Metabolism Associated with Lipid Accumulation and its Reversal in Rats given the Peroxisome Proliferator LY 171 883," *Toxicol. Appl. Pharmacol.* **106** (1990) 375–383.
223. D. E. Moody, J. K. Reddy: "Hepatic Peroxisome (Microbody) Proliferation in Rats Fed Plasticizers and Related Compounds," *Toxicol. Appl. Pharmacol.* **45** (1978) 497–504.
224. F. E. Mitchell et al.: "Time and Dose-Response Study of the Effects on Rats of the Plasticizers Di(2-ethylhexyl) Phthalate," *Toxicol. Appl. Pharmacol.* **81** (1985) 371–392.
225. B. E. Butterworth, D. J. Loury, T. Smith-Oliver, R. C. Cattley: "The Potential Role of Chemically Induced Hyperplasia in the Carcinogenic Activity of the Hypolipidemic Carcinogens," *Toxicol. Ind. Health* **3** (1987) 129–148.
226. A. H. Mann et al.: "Comparison of the Short-Term Effects of Di(2-ethylhexyl) Phthalate, Di(n-hexyl) Phthalate, and Di(n-octyl) Phthalate in Rats," *Toxicol. Appl. Pharmacol.* **77** (1985) 116–132.
227. K. N. Woodward, A. M. Smith, S. P. Mariscotti, N. J. Tomlinson: "Review of the Toxicity of the Esters of o-Phthalic Acid (Phthalate Esters)," *Toxic. Rev.* **14** (1986) .
228. B. G. Lake, T. J. B. Gray, S. D. Gangolli: "Hepatic Effects of Phthalate Esters and Related Compounds–In Vivo and In Vitro Correlations," *EHP Environ. Health Perspect.* **67** (1986) 283–290.
229. B. G. Lake et al.: "Comparative Studies on Di(2-ethylhexyl) Phthalate-Induced Hepatic Peroxisome Proliferation in the Rat and Hamster," *Toxicol. Appl. Pharmacol.* **72** (1984) 46–60.
230. J. A. Popp, L. K. Garvey, R. C. Cattley: "In Vivo Studies on the Mechanism of di(2-Ethylhexyl) Phthalate Carcinogenesis," *Toxicol. Ind. Health* **3** (1987) 151–163.
231. D. S. Marsman, R. Cattley, J. Conway, J. Popp: "Relationship of Hepatic Peroxisome Proliferation and Replicative DNA Synthesis to the Hepatocarcinogenicity of the Peroxisomal Proliferators Di(2-ethylhexyl)phthalate and [4-Chloro-6-(2,3-xlidino)-2-pyrimidinylthio]acetic acid (WY-14,643) in Rats", *Cancer Res*. **48** (1988) 6739–6744.
232. A. E. Ganning, J. J. Olsson, U. Brunk, G. Dallner: "Effects of Prolonged Treatment with Phthalate Ester on Rat Liver," *Pharmacol. Toxicol.* (Amsterdam) **68** (1991) 392–401.
233. C. Cattley, J. G. Conway, J. A. Popp: "Association of Persistent Peroxisome Proliferation and Oxidative Injury with Hepatocarcinogenicity in Female F-344 Rats Fed Di-(2-ethylhexyl) Phthalate for 2 years," *Cancer Lett. (Shannon Irel.)* **38** (1987) 15–22.
234. *An Investigation of the Effect of Di-2-Ethylhexyl Phthalate (DEHP)*, Dibutylphthalate (DBP), Di-iso-decylphthalate (DIDP) on Rat Hepatic Peroxisomes, British Industrial Biological Research Association (BIBRA), Surrey 1990.
235. J. M. Ward, J. M. Peters, C. M. Perella, F. J. Gonzalez: "Receptor- and Non-Receptor-Mediated Organ-Specific Toxicity of Di(2-ethylhexyl)phthalate (DEHP) in Peroxisome Proliferator-Activated Receptora-null Mice", *Toxicol. Pathol*. **26** (1998) 240–246.
236. S. A. Kliewer, K. Umesono, D. J. Noonan, R. A. Heymann, R. M Evans: "Covergence of 9-cis-Retinoic Acid and Peroxisome Proliferators Signalling Pathways Through Heterodimer Formation of their Receptors", *Nature* **358** (1992) 771–774.
237. B. P. Leblanc, H. G. Stunnenberg: "9-Cis-retinoic Acid Signalling: Changing Partners Causes Some Excitement", *Genes Dev*. **9** (1995) 1811–1816.
238. J. W. Lawrence, Y. Li, S. Chen, J. G. DeLuca, J. P. Berger, D. R. Umbenhauer, D. E. Moller, G. Zhou: "Differential Gene Regulation in Human versus Rodent Hepatocytes by Peroxisome Proliferator-Activated Receptor (PPAR) α", *J. Biol. Chem.* **276** (2001) 31521–31527.
239. E. D. Rosen, B. M. Spiegelman: "PPAR γ: A Nuclear Regulator of Metabolism, Differentiation, and Cell Growth", *J Biol Chem* **27** (2001) 37731–37734.???
240. R. Schulte-Herrmann, W. Bursch, B. Marian, B. Grasl-Kraupp: "Active cell Death (Apoptosis) and Cellular Proliferation as

Indicators of Exposure to Carcinogens", IARC Scientific Publications No. 146, International Agency for Research on Cancer, Lyon 1999.
241. N. H. James, R. A. Roberts: "Species Differences in Response to Peroxisome Proliferators Correlate in vitro with Induction of DNA Synthesis rather than Suppression of Apoptosis", *Carcinogenesis* **17** (1996) 1623–1632.
242. P. P. Van Veldhofen, G. P. Mannaerts: "Role and Organization of Peroxisomal β-Oxidation", *Adv. Exp. Med. Biol*. **466** (1999) 261–272.
243. C. H. Park, B. Breyer, T.-C. He: "Peroxisome Proliferators-Activated Receptors: Roles in Tumorigenesis and Chemoprevention in Human Liver", *Mol. Pharmacol*. **53** (2001) 14–22.
244. K. W. Kinzler, B. Vogelstein: "Lessons from Hereditary Colorectal Cancer", *Cell* **87** (1996) 159–170.
245. Y. Kurata, F. Kidachi, M. Yokoyama, N. Toyota, M. Tsuchitani, M. Katoh: "Subchronic Toxicity of Di-(2-ethylhexyl)phthalate in Common Marmosets: Lack of Peroxisomal Proliferation, Testicular Atrophy or Pancreatic Acinar Cell Hyperplasia", *Toxicol. Sci*. **42** 49–56 (1998).
246. R. L. Melnick: "Is Peroxisome Proliferation an Obligatory Precursor Step in the Carcinogenicity of Di(2-ethylhexyl)phthalate?", *Environ. Health Perspect* **109** (2001) 437–442.
247. J. F. S. Crocker, S. H. Safe, P. Acott: "Effects of Chronic Phthalate Exposure on the Kidney", *J. Toxicol. Environ. Health* **23** (1988) 433–444.
248. R. M. David, M. R. Moore, D. C. Finney, D. Guest: "Chronic Toxicity of Di(2-ethylhyl)phthalate in Rats", *Toxicol. Sci*. **55** (2000) 433–443.
249. R. M. David, M. R. Moore, M. A. Cifone, D. C. Finney, D. Guest: "Chronic Peroxisome Proliferation and Hepatomegaly Associated with the Hepatocellular Tumorigenesis of Di(2-ethylhyl)phthalate and the Effects of Recovery", *Toxicol. Sci*. **50** (1999) 195–205.
250. R. M. David, M. R. Moore, D. C. Finney, D. Guest: "Chronic Toxicity of Di(2-ethylhyl)phthalate in Mice", *Toxicol. Sci*. **58** (2000) 377–385.
251. P. Schmezer et al.: "Various Short-Term Assays and Two Long-Term Studies with the Plasticizer Di-(2-ethylhexyl) Phthalate in the Syrian Golden Hamster," *Carcinogenesis (London)* **9** (1988) 37–43.
252. K. Yoshikawa, A. Tanaka, T. Yamaha, H. Kurata: "Mutagenicity Study of Nine Monoalkyl Phthalates and a Dialkyl Phthalate using Salmonella Typhimurium and Escherichia Coli," *Food Chem. Toxicol.* **21** (1983) no. 2, 221–223.
253. R. J. Rubin, K. Kozumbo, R. Kroll: "Ames Mutagenic Assay of a Series of Phthalic Acid Esters: Positive response of the Dimethyl and Diethylesters in TA 100," *Toxicol. Appl. Pharmacol.* **48** (1979) A 133.
254. D. K. Agarwal, W. H. Lawrence, L. J. Nunez, J. Autian: "Mutagenicity Evaluation of Phthalic Acid Esters and Metabolites in Salmonella Typhimurium Cultures," *J. Toxicol. Environ. Health* **16** (1985) 61–69.
255. B. E. Butterworth et al.: "Lack of Genotoxic Activity of Di(2-ethylhexyl) Phthalate (DEHP) in Rat and Human Hepatocytes," *Carcinogenesis (London)* **5** (1984) 1329–1335.
256. D. J. Kornbrust, T. R. Barfknecht, P. Ingram, J. D. Shelburne: "Effect of Di(2-ethylhexyl) Phthalate on DNA Repair and Lipid Peroxidation in Rat Hepatocytes and on Metabolic Cooperation in Chinese Hamster V-79 Cells," *J. Toxicol. Environ. Health* **13** (1984) 99–116.
257. G. Zeiger, S. Haworth, K. Mortelmans, W. Speck: "Mutagenicity Testing of Di(2-ethylhexyl) Phthalate and Related Chemicals in Salmonella," *Environ. Mutagen.* **7** (1985) 213–232.
258. J. M. Parry, N. Danford, E. M. Parry: "In Vitro Techniques for the Detection of Chemicals Capable of Inducing Mitotic Chromosome Aneuploidy," *ATLA Altern. Lab. Anim.* **11** (1984) 117–128.
259. B. J. Phillips, D. Anderson, S. D. Gangolli: "Studies on the Genetic Effects of Phthalic Acid Esters on Cells in Culture," *EHP Environ. Health Perspect.* **263** (1986) 263–266.
260. N. N. Woodward: *Phthalate Esters: Toxicity and Metabolism*, vol. **II**, CRC Press, Boca RatonFlorida, 1988, pp. 47–75.
261. B. J. Phillips, T. T. B. James, S. D. Gangolli: "Genotoxicity Studies of DEHP and its Metabolites in CHO Cells," *Mutat. Res.* **102** (1982) 297–304.
262. H. A. A. M. Dirven, J. L. G. Theuws, F. J. Jongeneelen, R. P. Bos: "Non-mutagenicity of 4 Metabolites of Di(2-ethylhyl)phthalate (DEHP) and 3 Structurally Related Derivatives of Di(2-ethylhyl)adipate (DEHA) in the Salmonella Mutagenicity Assay", *Mutat. Res.* **260** (1991) 121–130.

263. D. Anderson, T.-W. Yu, F. Hincal: "Effect of Some Phthalate Esters in Human Cells in the Comet Assay", *Teratogen Carcinogen Mutagen* **19** (1999) 275–280.
264. T. Tsutsui, E. Watanabe, J. C. Barrett: "Ability of Peroxisome Proliferators to Induce Cell Transformation, Chromosome Aberrations and Peroxisome Proliferation in Cultured Syrian Hamster Embryo Cells", *Carcinogenesis* **14** (1993) 611–618.
265. A. R. Singh, W. H. Lawrence, J. Autian: "Mutagenic and Antifertility Sensitivities of Mice to Di-2-ethylhexyl Phthalate (DEHP) and Dimethoxyethyl Phthalate (DMEP)," *Toxicol. Appl. Pharmacol.* **29** (1974) 35–46.
266. D. K. Agarwal, W. H. Lawrence, L. J. Nunez, J. Autian: "Antifertility and Mutagenic Effects in Mice from Parenteral Administration of Di-2-ethylhexyl Phthalate (DEHP)," *J. Toxicol. Environ. Health* **16** (1985) 61–69.
267. Y. Hamano et al.: "Dominant lethal test of DEHP in mice," *Food Hygiene Series* **10** (1980) 1–10 (Japan); "Chem. Abstr. Sect. Toxicol. 92 2 099 960, 145.
268. C. J. Rushbrook, T. A. Jorgenson, J. R. Hodgson: "Dominant Lethal Study of Di-2-(ethylhexyl) Phthalate and its Major Metabolites in ICR/SIM Mice," *Environ. Mutagen.* **4** (1982) 387 A.
269. A. von Däniken, W. K. Lutz, R. Jäckh, C. Schlatter: "Investigation of the Potential for Binding of Di(2-ethylhexyl) Phthalate (DEHP) and Di(2-ethylhexyl) Adipate (DEHA) to Liver DNA in Vivo," *Toxicol. Appl. Pharmacol.* **73** (1984) 373–387.
270. A. Takagi et al.: "Relationship between Hepatic Peroxisome Proliferation and 8-Hydroxyoxyguanosine Formation in Liver DNA of Rats following Long-Term Exposure to Three Peroxisome Proliferators; Di(2-ethylhexyl) Phthalate, Aluminium Clofibrate and Simfibrate," *Cancer Lett.* (Shannon Irel.) 53 (1990) 33–38.
271. L. I. Hinrichsen, R. A. Floyd, O. Sudilovsky: "Is 8-Hydroxydeoxyguanosine a Mediator of Carcinogenesis by a Choline-Devoid Diet in the Rat Liver?" *Carcinogenesis (London)* **11** (1990) 1879–1881.
272. R. A. Floyd: "The role of 8-Hydroxyguanine in Carcinogenesis," *Carcinogenesis (London)* **11** (1990) 1447–1450.
273. R.C. Cattley, S.E. Glover: "Elevated 8-hydroxydeoxyguanosine in Hepatic DNA of Rats Following Exposure to Peroxisome Proliferators: Relationship to Carcinogens and Nuclear Localisation", *Carcinogenesis* **14** (1993) 2495–2499.
274. Deutsche Forschungsgemeinschaft (DFG). Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe. "Di(2-ethylhexyl)phthalate (DEHP)". *Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten*. Wiley-VCH, Weinheim, 2002.
275. C. P. Carpenter, C. S. Weil, H. F. Smyth: "Chronic Oral Toxicity of Di(2-ethylhexyl) Phthalate for Rats, Guinea Pigs, and Dogs," *Arch. Ind. Hyg. Occup. Med.* **8** (1953) 219–226.
276. R. S. Harris, H. C. Hodge, E. A. Maynard, H. J. Blanchet Jr.: "Chronic Oral Toxicity of 2-Ethylhexyl Phthalate in Rats and Dogs," *AMA Arch. Ind. Health* **13** (1956) 259–264.
277. J. M. Makowska, F. W. Bonner, G. G. Gibson: "Hepatic Induction Potency of Hypolipidaemic Drugs in the Rat Following Long-Term Administration: Influence of Different Dosing Regiments," *Xenobiotica* **20** (1990) 1112–1128.
278. W. M. Kluwe, J. K. Haseman, J. F. Douglas. J. E. Huff: "The Carcinogenicity of Dietary Di(2-ethylhexyl)phthalate in Fischer 344 Rats and B6C3F1 Mice", *J. Toxicol. Environ. Health* **10** (1982) 797–815.
279. W. M. Kluwe: "Carcinogenic Potential of Phthalic Acid Esters and Related Compounds: Structure-Activity Relationships", *Environ. Health. Perspect.* **65** (1986) 271–278.
280. National Toxicology Program (NTP). Carcinogenesis bioassay of di(2-ethylhexyl)phthalate (CAS No. *117-81-7*) in F344 rats and B6C3F1 mice (feed study). Research Triangle Park, NC, U.S. Department of Health and Human Services, Public Health Services, National Institute of Health, NTP publication no. 217 (1982).
281. M. S. Rao, A. V. Yeldandi, V. Subbarao: "Quantitative Analysis of Hepatocellular Lesions Induced by Di(2-ethylhexyl)phthalate in F-344 rats", *J. Toxicol. Environ. Health* **30** (1990) 85–89.
282. J. M. Peters, R. C. Cattley, F. J. Gonzalez: "Role of PPARα in the Mechanism of Action of the Nongenotoxic Carcinogen and Peroxisome Proliferator Wy-14, 643". *Carcinogenesis* **18** (1997) no. 11, 2029–2033.
283. J. C. Cook, G. R. Klinefelter, J. F. Hardist, R. M. Sharpe, P. M. D. Foster: "Rodent Leydig

Cell Tumorigenesis: A Review of the Physiology, Pathology, Mechanisms, and Relevance to Humans", *Crit. Rev. Toxicol.* **29** (1999) 169–261.
284. G. Lake et al.: "Effect of Prolonged Administration of Clofibric Acid and Di-(2-ethylhexyl) Phthalate on Hepatic Enzyme Activities and Lipid Peroxidation in the Rat," *Toxicology* **44** (1987) 213–228.
285. M. S. Rao, N. Usuda, V. Subbarao, J. K. Reddy: "Absence of γ-Glutamyl Transpeptidase Activity in Neoplastic Lesions induced in the Liver of Male F-344 Rats by Di-(2-ethylhexyl) Phthalate, a Peroxisome Proliferator," *Carcinogenesis* **8** (1987) 1347–1350.
286. Biodynamics Inc., A Chronic Toxicity Carcinogenicity Feeding Study in Rats with Santicizer 900, Proj. no. 81–2572, Monsarto Company, 1986.
287. A. Lington, M. Bird, R. Blutnick, J. Quance: *The Toxicologist* **7 Abstr.** (1987) 405.
288. Carcinogenesis Bioassay of Butyl Benzyl Phthalate (CAS No. 85–68–7) in F 344/N Rats and B 6 C 3 F 1 Mice (Feed Study). NTP-Rep. 80–25, NIH Publication 82–1769, US Dpt. Health and Human Services, Washington, DC, 1982.
289. L. K. Garvey, J. A. Swenberg, T. E. Hamm, J. A. Popp: "Di(2-ethylhexyl) Phthalate: Lack of Initiating Activity in the Liver of Female F 334 Rats," *Carcinogenesis (London)* **8** (1987) 285–290.
290. J. J. Reddy, N. D. Lalwani: "Carcinogenesis by Hepatic Peroxisome Proliferators: Evaluation of the Risk of Hypolipidemic Drugs and Industrial Plasticizers to Humans," *CRC Crit. Rev. Toxicol.* **12** (1983) 1–58.
291. J. G. Conway, R. C. Cattley, J. A. Popp, B. E. Butterworth: "Possible Mechanisms in Hepatocarcinogenesis by the Peroxisome Proliferator Di-(2-ethylhexyl) Phthalate," *Drug Metab. Rev.* **21** (1989) no. 1, 65–102.
292. T. Yario et al.: "Effect of Di-(2-ethylhexyl) Phthalate (DEHP) on DNA Synthesis and on Promotion of Glucose-6-phosphatase Negative Preneoplastic Lesions in Rat Liver," *Carcinogenesis (London)* **27 148, Abstr. 585** (1986).
293. A. B. DeAngelo, C. T. Garrett, A. E. Queral: "Inhibition of Phenobarbital and Dietary Choline Deficiency Promoted Preneoplastic Lesions in Rat Liver by Environmental Contaminant Di(2-ethylhexyl) Phthalate," *Cancer Lett*, (Shannon Irel.) **23** (1984) 323–330.
294. J. M. Ward et al.: "Tumor-Initiating and Promoting Activities of Di(2-ethylhexyl) Phthalate in Vivo and in Vitro," *EHP Environ. Health Perspect.* **65** (1986) 279–291.
295. M. I. R. Perera, H. Shinozuka: "Accelerated Regression of Carcinogen-Induced Preneoplastic Hepatocyte Foci by Peroxisome Proliferators, BR 931, 4-Chloro-6-(2,3-xylidino)-2-pyrimidinylthio (N-B-hydroxyethyl)acetamide, and Di(2-ethylhexyl) Phthalate," *Carcinogenesis (London)* **5** (1984) 1193–1198.
296. D. Oesterle, E. Deml: "Promoting Activity of Di (2-ethylhexyl) Phthalate in Rat Liver Foci Bioassay," *J. Cancer Res. Clin. Oncol.* **114** (1988) 133–136.
297. A. B. DeAngelo, A. E. Queral, C. T. Garrett: "Concentration-Dependent Inhibition of Development of GGT Positive Foci in Rat Liver by the Environmental Contaminant Di(2-ethylhexyl) Phthalate," *EPH Environ. Health Perspect.* **60** (1985) 381–385.
298. A. B. DeAngelo, C. T. Garrett: "Inhibition of Development of Preneoplastic Lesions in the Livers of Rats Fed a Weakly Carcinogenic Environmental Contaminant," *Cancer Lett. (Shannon Irel.)* **20** (1983) 199–205.
299. B. Grasl-Kraupp, W. Huber, H. Taper, R. Schulte-Hermann: "Increased Susceptibility of Aged Rats to Hepatocarcinogenesis by the Peroxisome Proliferator Nafenopin and the Possible Involvement of Altered Liver Foci Occurring Spontaneously," *Cancer Res.* **51** (1991) in press.
300. J. Merkle, H. J. Klimisch, R. Jäckh: "Developmental Toxictiy in Rats after Inhalation Exposure of Di-2-ethylhexyl Phthalate (DEHP)," *Toxicol. Lett.* **42** (1988) 215–223.
301. R. R. Shukla, P. W. Albro: "In Vitro Modulation of Protein Kinase C Activity by Environmental Chemical Pollutants," *Biochem. Biophys. Res. Commun.* **142** (1987) 567–572.
302. G. Wolfe, K. Layton, L. Nehrebeckyj, Y. Wang, R. Chapin, S. Rousselle, J. Bishop: "Reproductive effects of DEHP in Sprague-Dawley rats when assessed by the continuous breeding protocol," presented at the US Society of Toxicology meeting, 20 March 2002.
303. C. Gupta, A. Hattori, H. Shinozuka: "Suppression of EGF Binding in Rat Liver by

the Hypolipidemic Peroxisome Proliferators, 4-chloro-6(2,3-xylidino)-2-pyrimidinylthio(N-β-hydroxyl)acetamide and Di-(2-ethylhexyl) Phthalate," *Carcinogenesis (London)* **9** (1988) 167–169.

304. R. L. Melnick.R. E. Morrissey, K. E. Tomaszewski: "Studies by the National Toxicology Program on Di(2-ethylhexyl) Phthalate," *Toxicol. Ind. Health* **3** (1987) 99–118.
305. K. E. Tomaszewski, D. A. Agarwal, R. L. Melnick: "In Vitro Steady-State Levels of Hydrogen Peroxide after Exposure of Male F 344 Rats and Female B 6 C 3 F 1 Mice to Hepatic Peroxisome Proliferators," *Carcinogenesis* (London) 7 (1986) 1871–1876.
306. H. Tamura, T. Iida, T. Watanabe, T. Suga: "Long-Term Effects of Hypolipidemic Peroxisome Proliferator Administration on Hepatic Hydrogen Peroxide Metabolism in Rats," *Carcinogenesis (London)* **11** (1990) 445–450.
307. J. G. Conway et al.: "Relationship of Oxidative Damage to the Hepatocarcinogenicity of the Peroxisome Proliferators Di(2-ethylhexyl) Phthalate and Wy-14,643," *Carcinogenesis (London)* **10** (1989) 513–519.
308. S. Oishi, K. Hiraga: "Testicular Atrophy Induced by Phthalic Esters: Effect on Testosterone and Zinc Concentrations," *Toxicol. Appl. Pharmacol.* **53** (1980) 35–41.
309. S. Oishi, K. Hiraga: "Effect of Phthalic Acid Esters on Gonadal Function in Male Rats," *Bull. Environ. Contam. Toxicol.* **21** (1979) 65–67.
310. P. M. D. Foster, L. V. Thomas, M. W. Cook, S. D. Gangolli: "Study of the Testicular Effects and Changes in Zinc Excretion Produced by some n-Alkyl Phthalates in the Rat," *Toxicol. Appl. Pharmacol.* **54** (1980) 392–398.
311. P. M. D. Foster et al.: "Differences in Urinary Metabolic Profile from Di-n-butyl Phthalate-Treated Rats and Hamster. A Possible Explanation for Species Differences in Susceptibility to Testicular Atrophy," *Drug. Metab. Dispos.* **11** (1983) 59–61.
312. J. A. Ward, H. Zenick, W. W. Wilfinger, R. J. Niewenhuis: "Effects of Dibutyl Phthalate (DBP) on the Reproductive System of Two Strains of Young Male Rats," *The Toxicologist* **4** (1984) 28 A.
313. S. C. Kang, B. M. Lee: "Comparative evaluation of phthalates for sperm motility and male fertility in Sprague-Dawley rats," *Birth Defects Res. Part A Clin. Mol. Teratol.* **70** (2004) 310.
314. B. R. Cater, M. W. Cook, S. D. Gangolli, P. Grasso: "Studies on Di-butyl Phthalate-Induced Testicular Atrophy in the Rat: Effect on Zinc Metabolism," *Toxicol. Appl. Pharmacol.* **41** (1977) 609–618.
315. J. R. Reel, A. D. Lawton, D. B. Feldmann, J. C. Lamb: "Di(n-butyl) Phthalate: Reproduction and Fertility Assessment in CD-1 Mice, when Administered in the Feed," Report 1984, NTP-84–411, *Gov. Rep. Announce. Index* (*U.S.*) **85** (1985) no. 7, 35.
316. T. J. B. Gray, K. R. Butterworth: "Testicular Atrophy Produced by Phthalate Esters," *Arch. Toxicol. Suppl.* **4** (1980) 452–455.
317. T. J. B. Gray et al.: "Short-Term Toxicity Study of Di(2-ethylhexyl) Phthalate in Rats," *Food. Cosmet. Toxicol.* **15** (1977) 389–399.
318. S. Oishi: "Reversibility of Testicular Atrophy Induced by Di(2-ethylhexyl) Phthalate in Rats," *Environ. Res.* **36** (1984) 160–169.
319. S. Oishi, K. Hiraga: "Testicular Atrophy Induced by Di-2-ethylhexyl Phthalate: Effect of Zinc Supplement," *Toxicol. Appl. Pharmacol.* **70** (1983) 43–48.
320. S. P. Srivastava et al.: "Testicular Effects of Di-n-butyl Phthalate (DBP): Biochemical and Histopathological Alterations," *Arch. Toxicol.* **64** (1990) 148–152.
321. P. Sjöberg, N. G. Lindquist, L. Plöen: "Age-Dependent Response of the Rat Testes to Di(2-ethylhexyl) Phthalate," *EHP Environ. Health Perspect.* **65** (1986) 234–242.
322. M. Fukuoka, Y. Zou, A. Tanaka: "Mechanism of Testicular Atrophy induced by Di-n-butyl Phthalate in Rats, part 2. The Effects on Some Testicular Enzymes," *JAT, J. Appl. Toxicol.* **10** (1990) 285–293.
323. T. J. B. Gray, S. D. Gangolli: "Aspects of the Testicular Toxicity of Phthalate Esters," *EPH Environ. Health Perspect.* **65** (1986) 229–235.
324. L. A. Dostal et al.: "Testicular Toxicity and Reduced Sertoli Cell Numbers in Neonatal Rats by Di(2-ethylhexyl) Phthalate and the Recovery of Fertility as Adults," *Toxicol. Appl. Pharmacol.* **95** (1988) 104–121.
325. S. D. Gangolli: "Testicular Effects of Phthalate Esters," *EPH Environ. Health Perspect.* **45** (1982) 77–84.
326. T. J. B. Gray, K. R. Butterworth, I. F. Gaunt, P. Grasso, S. D. Gangolli: "Short-Term Toxicity Study of Di(2-ethylhexyl)phthalate in Rats", *Food Cosmet. Toxicol.* **15** (1977) 389–399.

327. D. K. Agarwal et al.: "Influence of Dietary Zinc on Di(2-ethylhexyl) Phthalate-Induced Testicular Atrophy and Zinc Depletion in Adult Rats," *Toxicol. Appl. Pharmacol.* **84** (1986) 12–24.
328. D. K. Agarwal et al.: "Effects of Di(2-ethylhexyl) Phthalate on the Gonadal Pathophysiology, Sperm Morphology, and Reproductive Performance of Male Rats," *EPH Environ. Health Perspect.* **65** (1986) 343–350.
329. L. A. Dostal, W. L. Jenkins, B. A. Schwetz: "Hepatic Peroxisome Proliferation and Hypolipidemic Effects of Di(2-ethylhexyl)phthalate in Neonatal and Adult Rats", *Toxicol. Appl. Pharmacol.* **87** (1987) **81** 90.
330. D. Parmar, S. P. Srivastava, G. B. Singh, P. K. Seth: "Testicular Toxicity of Di(2-ethylhexyl)phthalate in Developing rats", *Vet. Hum. Toxicol.* **37** (1995) no. 4, 310–313 (1995).
331. K. Schilling, K. Deckardt, C. Gembardt: Di-2-ethylhexyl phthalate—Two-generation reproduction toxicity range-finding study in Wistar rats, continuous dietary administration. Laboratory Project ID: 15R091/997096: BASF Aktiengesellschaft (2001).
332. National Toxicology Program. Diethylhexylphthalate: Multigenerational reproductive assessment by continuous breeding when administered to Sprague-Dawley rats in the diet. Research Triangle Park NC: National Toxicology Program (2004). Abstract. Available at http://ntp-server.niehs.nih.gov.
333. R. Poon, P. Lecavalier, R. Mueller, V. E. Valli, B. G. Procter, I. Chu: "Subchronic Oral Toxicity of Di-n-octyl Phthalate and Di(2-ethylhexyl) Phthalate in the Rat", *Food Chem. Toxicol.* **35** (1997) 225–239.
334. T. J. B. Gray, J. A. Beamand: "Effect of some Phthalate Esters and other Testicular Toxins on Primary Cultures of Testicular Cells," *Food Chem. Toxicol.* **22** (1983) 123.
335. C. Rhodes et al.: "The Absence of Testicular Atrophy in Vivo and in Vitro Effects on Hepatocyte Morphology and Peroxisomal Enzyme Activities in Male Rats Following the Administration of Several Alkanols," *Toxicol. Lett.* **21** (1984) 103–109.
336. R. Wolowsky-Tyl, C. Jones-Price, M. C. Marr, C. A. Kimmel: "Teratologic Evaluation of Diethylhexyl Phthalate (DEHP) in Fischer-344 Rats," *Teratology* **27** (1983) 85 A.
337. J. J. Heindel et al.: "Reproductive Toxicity of Three Phthalic Acid Esters in a Continuous Breeding Protocol," *Fundam. Appl. Toxicol.* **12** (1989) 508–518.
338. B. G. Lake et al.: "Studies on the Effect of Orally Administered Di-(2-ethylhexyl)Phthalate in the Ferret", *Toxicology* **6** (1976) 341.
339. P. M. D. Foster et al.: "Changes in Ultrastructure and Cytochemical Localization of Zinc in Rat Testis Following the Administration of Di-n-pentyl Phthalate," *Toxicol. Appl. Pharmacol.* **63** (1982) 120–132.
340. J. D. Park, S. S. Habeebu, C. D. Klaassen: "Testicular Toxicity of Di-(2-ethylhexyl)phthalate in Young Sprague-Dawley Rats", *Toxicology* **171** (2002) 105–115.
341. R. C. Scott, P. H. Dugard, J. D. Ramsey, C. Rhodes: "In Vitro Absorption of some o-Phthalate Diesters through Human and Rat Skin," *EHP Environ. Health Perspect.* **74** (1987) 223–227.
342. National Toxicology Program (NTP). NTP-CERHR Expert Panel Report on DEHP. (2000). Available at http://cerhr.niehs.nih.gov/news/phthalates/DEHP-final.pdf.
343. Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological Profile for Di(2-ethylhexyl)phthalate. U.S. Department of Health and Human Services (2002). Available at http://www.atsdr.cdc.gov/toxprofiles.
344. National Toxicology Program (NTP). NTP-CERHR Expert Panel Update on the reproductive and developmental toxicity of Di(2-ethylhexyl)phthalate (2005). Available at http://cerhr.niehs.nih.gov.
345. J. C. Lamp IV. et al.: "Reproductive Effects of Four Phthalic Acid Esters in the Mouse," *Toxicol. Appl. Pharmacol.* **88** (1987) 255–269.
346. R. W. Moore, T. A. Rudy, T. M. Lin, K. Ko, R. E. Peterson: "Abnormalities of Sexual Development in Male Rats with in Utero and Lactational Exposure to the Antiandrogenic Plasticizer Di(2-ethylhexyl) Phthalate", *Environ. Health Perspect.* **109** (2001) 229–237.
347. Mitsubishi chemical safety institute, unpublished results, Study No. 7L432, 30 June 1998; sponsored by Plasticizer Industrial Association, Tokyo (1998).
348. BASF AG, unpublished results, Project No. 07R0491/97143, (1999).

349. K. Shiota, M. J. Chou, H. Nishimura: "Embryotoxic Effects of Di-2-ethylhexyl Phthalate (DEHP) and Di-n-butyl Phthalate (DBP) in Mice," *Environ. Res.* **22** (1980) 245–253.
350. K. Shiota, H. Nishimura: "Teratogenicity of Di(2-ethylhexyl) Phthalate (DEHP) and Di-n-butyl Phthalate (DBP) in Mice," *EHP Environ. Health Perspect.* **45** (1982) 65–70.
351. K. Shiota, S. Mima: "Assessment of the Teratogenicity of Di(2-ethylhexyl) Phthalate and Mono-(2-ethylhexyl) Phthalate in Mice," *Arch. Toxicol.* **56** (1985) 263–266.
352. R. Wolowsky-Tyl, C. Jones-Price, M. C. Marr, C. A. Kimmel: "Teratologic Evaluation of Diethylhexyl Phthalate (DEHP) in CD-1 Mice," *Teratology* **27** (1983) 84 A–85 A.
353. Y. Nakamura, Y. Yagi, I. Tomita, K. Tsuchikawa: "Teratogenicity of Di(2-ethylhexyl) Phthalate in Mice," *Toxicol. Lett.* **4** (1979) 113–117.
354. J.C. Lamb IV: "Reproductive Effects of Four Phthalic Acid Esters in the Mouse", *Toxicol. Appl. Pharmacol.* **88** (1987) 255–269.
355. C. J. Price et al.: "Developmental Toxicity of Butyl Benzyl Phthalate (BBP) in Mice and Rats," *Teratology* **41** (1990) 586; *Chem. Abstr.* P 51.
356. J. Hellwig, H. Freudenberger, R. Jackh: "Differential Prenatal Toxicity of Branched Phthalate Esters in Rats", *Food Chem. Toxicol.* **35** (1997) 501–512.
357. L. E. Gray, Jr., J. Ostby, J. Furr, M. Price, D. N. Veeramachaneni, L. Parks: "Perinatal Exposure to the Phthalates DEHP, BBP, and DINP, but not DEP, DMP, or DOTP, Alters Sexual Differentiation of the Male Rat", *Toxicol. Sci.* **58** (2000) 350–365.
358. H. Swenerton, L. S. Hurley: "Teratogenic Effects of a Chelating Agent and their Prevention by Zinc", *Science (Washington D. C.)* **173** (1971) 62–63.
359. E. A. Field et al.: "Developmental Toxicity Evaluation of Dimethyl Phthalate (DMP) in CD Rats," *Teratology* **39** (1989) 452, P 76 [A].
360. C. J. Price et al.: "Developmental Toxicity Evaluation of Diethyl Phthalate (DEP) in CD Rats," *Teratology* **39** (1989) 473, P 66 [A].
361. E. Hansen, O. Meyer: "No Embryotoxic or Teratogenic Effect of Dimethyl Phthalate in Rats after Epicutaneous Application," *Pharmacol*, Toxicol. (Amsterdam) 64 (1989) 237–238.
362. W. Kluwe, J. K. Haseman, J. F. Douglas, J. E. Huff: "The Carcinogenicity of Dietary di(2-Ethylhexyl) Phthalate (DEHP) in Fisher-344 Rats and B 6 C 3 F 1 Mice," *J. Toxicol. Environ. Health* **10** (1982) 797–815.
363. D. J. Benford et al.: "Species Differences in Response of Cultured Hepatocytes to Phthalate Esters," *Food Chem. Toxicol.* **24** (1986) 799–800.
364. J. V. Rodricks, D. Turnbull: "Interspecies Differences in Peroxisomes and Peroxisome Proliferation," *Toxicol. Ind. Health* **3** (1987) 197.
365. J. M. Ward et al.: "The Chronic Hepatic or Renal Toxicity of Di(2-ethylhexyl) Phthalate, Acetaminophen, Sodium Barbital, and Phenobarbital in Male B 6 C 3 F 1 Mice: Autoradiographic, Immunohistochemical, and Biochemical Evidence for Levels of DNA Synthesis not Associated with Carcinogenesis or Tumor Promotion," *Toxicol. Appl. Pharmacol.* **96** (1988) 494–506.
366. R. H. Hinton et al.: "Effects of Phthalic Acid Esters on the Liver and Thyroid," *EPH Environ. Health Perspect.* **70** (1986) 195–210.
367. K. N. Woodward: "Phthalate Esters, Cystic Kidney Disease in Animals and Possible Effects on Human Health: A Review," *Hum. Exp. Toxicol.* **9** (1990) 397–401.
368. J. F. S. Crocker, S. H. Safe, P. Acott: "Effects of Chronic Phthalate Exposure on the Kidney," *J. Toxicol. Environ. Health* **23** (1988) 433–444.
369. S. C. Price et al.: "Alterations in the Thyroids of Rats Treated for Long Periods with Di(2-ethylhexyl) Phthalate or with Hypolipidaemic Agents," *Toxicol. Lett.* **40** (1988) 37–46.
370. S. Khanna et al., *J. Environ. Biol.* **11** (1990) 22–34.
371. M. Kawano: "Toxicological Studies on Phthalate Esters: 1. Inhalation of Effects of Dibutyl Phthalate (DBP) on Rats. 2. Metabolism, Accumulation and Excretion of Phthalate Esters in Rats," *Nippon Eiseigaku Zaschi* **35** (1980/81) 693–701; *Nippon Eiseigaku Zaschi* **35** (1981) 687–701.
372. D. K. Agarwal, R. R. Maronpot, J. C. Lamb, W. M. Kluwe: "Adverse Effects of Butyl Benzyl Phthalate on the Reproductive and Hematopoietic Systems of Male Rats," *Toxicology* **35** (1985) 189–206.
373. A. B. DeAngelo, C. T. Garrett, L. A. Manolukas, T. Yario: "Di-n-octyl Phthalate (DOP), a Relatively Ineffective Peroxisome Inducing Straight Chain Isomer of the Environmental Contaminant Di(2-ethylhexyl)

Phthalate (DEHP), Enhances the Development of Putative Preneoplastic Lesions in Rat Liver," *Toxicology* **41** (1986) 279–288.
374. OECD (2004). Integrated HPV database. Available at http://cs3-hq.oecd/org/scripts/hpv.
375. European Chemical Bureau (ECB). http://ecb.jrc.it/existing-chemicals/.
376. International Agency for Research on Cancer (IARC). Some industrial chemicals. *IARC monographs on the evaluation of carcinogenic risks to humans*, vol. 77, IARC, Lyon 2002, pp. 41–148.
377. Umweltbundesamt: Phthalate — Die nützlichen Weichmacher mit den unerwünschten Eigenschaften, 2007. Available at http://www.umweltbundesamt.de.
378. H. M. Koch, H. M. Bolt, R. Preuss, J. Angerer: "New Metabolites of Di(2-ethylhexyl) Phthalate (DEHP) in Human Urine and Serum after Single Oral Doses of Deuterium-Labelled DEHP", *Arch. Toxicol.* **79** (2005) 367–376.
379. H. M. Koch, B. Rossbach, H. Drexler, J. Angerer: "Internal Exposure of the General Population to DEHP and Other Phthalates—Determination of Secondary and Primary Phthalate Monoester Metabolites in Urine", *Environ. Res.* **93** (2003) 177–185.
380. H. M. Koch, H. Drexler, J. Angerer: "An Estimation of the Daily Intake of Di(2-ethylhexyl) Phthalate (DEHP) and Other Phthalates in the General Population", *Int. J. Hyg. Environ. Health* **206** (2003) 77–83.
381. National Institutes of Health (NIH). Notice. FR document 05-2127. Federal Register, Vol. 70, No. 23, 43870-43871 (2005). Information available at http://cerhr.niehs.nih.gov/news/fedreg/fr2_4_05.pdf.
382. R. H. McKnee, J. H. Butala, R. M. David, G. Gans: "NTP center for the Evaluation of Risks to Human Reproduction Reports on Phthalates: Addressing the Data Gaps," *Reproductive Toxicol.* **18** (2004) 1–22.
383. B. C. Blount et al.: "Levels of Seven Urinary Phthalate Metabolites in a Human Reference Population", *Environ. Health Perspect.* **108** (2000) 979–982.
384. Centers for Disease Control and Prevention, Third National Report on Human Exposures to Environmental Chemicals, Atlanta: Department of Health and Human Services, Centers for Disease Control and Prevention, 2005.
385. S. Loff, U. Subotic, F. Reinicke, H. Wischmann, J. Brade; "Extraction of Diethylhexyl Phthalate from Perfusion Lines of Various Material, Length and Brand by Lipid Emulsions," *J. Ped. Gastroent. Nutrit.* **39** (2004) 341–345.
386. F. Welle, G. Wolz, R. Franz. "Migration von Weichmachern aus PVC Schläuchen in enterale Nahrungslösungen," *Pharma Int.* **3** (2005) 17–21.
387. G. Rock, R. S. Labow, M. Tocchi: "Distribution of Di(2-ethylhexyl) Phthalate and Products in Blood and Blood Components," *EPH Environ. Health Perspect.* **65** (1986) 309–316.
388. FDA. Safety Assessment of Di(2-ethylhexyl)phthalate (DEHP) Released from PVC Medical Devices. Govt Reports Announcements & Index 2004, 21.
389. EU Scientific Committee on Toxicity, Ecotoxicity and the Environment (CSTEE). Phthalate migration from soft PVC toys and child-care articles. Opinion expressed at the 3rd CSTEE plenary meeting. European Commission (1998).
390. P. Sjöberg, U. Bondesson, G. Sedin, J. Gustafsson: "Dispositions of Di- and Mono-(2-ethylhexyl) Phthalate in Newborn Infants Subjected to Exchange Transfusions," *Eur. J. Clin. Invest.* **15** (1985) 426–430.
391. Y. A. Barry, R. S. Labow, W. J. Keon, M. Tocchi: "Atropine Inhibition of the Cardiodepressive Effect of Mono(2-ethylhexyl) Phthalate on Human Myocardium," *Toxicol. Appl. Pharmacol.* **106** (1990) 48–52.
392. A. E. Elsisi, D. E. Carter, I. G. Sipes: "Dermal Absorption of Phthalate Diester in Rats", *Fundam. Appl. Toxicol.* **12** (1989) 70–77.
393. R. D. Short, E. C. Robinson, A. W. Lington, A. E. Chin: "Metabolic and Peroxisome Proliferation Studies with Di(2-ethylhexyl) Phthalate in Rats and Monkeys," *Toxicol. Ind. Health* **3** (1987) 185–196.
394. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA). Grenzwerte in der Luft am Arbeitsplatz. TRGS 900, August 2004. BArbBl 7/8 (2004). Available at http://www.baua.de/prax/ags/trgs900.pdf.
395. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA). Grenzwerte in der Luft am Arbeitsplatz. TRGS 900, January 2006. BArbBl **12** (2006). Available at http://www.baua.de/prax/ags/trgs900.pdf.
396. Deutsche Forschungsgemeinschaft (DFG), Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe: *MAK-*

und BAT-Werte-Liste 2006, Mitteilung 42, Wiley-VCH, Weinheim, 2006.

397. European Commission. Entscheidung der Kommission vom 7. Dezember 1999 über Maßnahmen zur Untersagung des Inverkehrbringens von Spielzeug- und Babyartikeln, die dazu bestimmt sind, von Kindern unter drei Jahren in den Mund genommen zu werden, und aus Weich-PVC bestehen, das einen oder mehrere der Stoffe Diisononylphthalat (DINP), Di-(2-ethylhexyl)phthalat (DEHP), Dibutylphthalat (DBP), Diisodecylphthalat (DIDP), Di-n-octylphthalat (DNOP) oder Benzylbutylphthalat (BBP) enthält. Decision 1999/815/EC, ABl. Nr. L 315: 46; 9. Dec. 1999.
398. European Commission. 22nd Amendment to Directive EC 76/769. 5. July 2005 (2005). europa.eu.int/comm/enterprise/chemicals/legislation/markrestr/preparation_en.htm.
399. European Commission. Directive 2005/84/EC of the European Parliament and of the Commission. 22nd Amendment to Council Directive 76/769/EEC. 14 December 2005. Official Journal of the European Union, L 344/40-43, 27 December 2005.
400. Umweltbundesamt (ed.): Substitution von PBT-Stoffen in Produkten und Prozessen. Leitfaden zur Anwendung umweltverträglicher Stoffe. Teil 5: Hinweise zur Substitution gefährlicher Stoffe. 5.1 Funktion: Weichmacher. Umweltbundesamt, FKZ 20128 231 (2003). Available at http://www.umweltbundesamt.de.
401. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA): *Verzeichnis krebserzeugender, erbgutverändernder oder fortpflanzungsgefährdender Stoffe*, Tätigkeiten und Verfahren nach Anhang I der Richtlinie 67/548/EWG, TRGS 905 und TRGS 906. July (2005). Available at: http://www.baua.de/prax/ags/cmr_liste.htm.

Phthalocyanines

GERD LÖBBERT, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 4
2.1. General Data 4
2.2. Particle Size 4
2.3. Color Properties 5
2.4. Crystal Structure and Polymorphism 5
2.5. Surface and Interfacial Properties . 6
2.6. Other Physical Investigations 7
3. Chemical Properties 7
3.1. Oxidation 7
3.2. Reduction 9
3.3. Catalytic Properties 9
4. Production 10
4.1. General Synthetic Methods 10
4.2. Reaction Mechanism 11
4.3. Industrial Production 12
4.3.1. Copper Phthalocyanine 12
4.3.2. Phthalocyanine 14
4.4. Pigment Finishing 14
4.4.1. Finishing by Treatment with Acid . . 15
4.4.2. Finishing by Grinding 16
4.5. Stabilization of Phthalocyanines . . 17
5. Phthalocyanine Derivatives 18
5.1. Halogenated Phthalocyanines 19
5.2. Phthalocyanine Sulfonic Acids and Sulfonyl Chlorides 20
5.3. Chloromethylphthalocyanines . . . 20
5.4. Phthalocyanines with Other Substituents 21
5.5. Polymeric Phthalocyanines 21
5.6. Phthalocyanine Analogues 21
6. Legal Aspects and Environmental Protection 22
7. Quality Specifications and Analysis 23
8. Uses 23
8.1. Phthalocyanine Pigments 23
8.2. Phthalocyanine Dyes 24
9. Economic Aspects 25
10. Toxicology and Occupational Health 26
11. References 27

1. Introduction

The term phthalocyanine was first used by R. P. LINSTEAD in 1933 [4] to describe a class of organic dyes, whose colors range from reddish blue to yellowish green. The name phthalocyanine originates from the Greek terms *naphtha* for mineral oil and *cyanine* for dark blue. In 1930–1940, LINSTEAD et al. elucidated the structure of phthalocyanine (H_2Pc) and its metal complexes [4–14]. The basic structure is represented by phthalocyanine (**1**) itself:

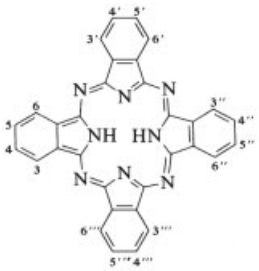

1

Phthalocyanine forms complexes with numerous metals of the periodic table [15–21]. Metal phthalocyanines MPc (**2**) and compounds with metalloids such as B, Si, Ge, and As or nonmetals such as P display a wide variety in their coordination chemistry.

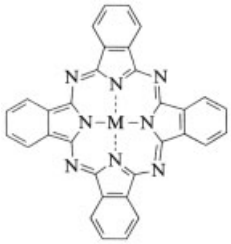

2

The coordination number of the square planar complexes of Cu, Ni, or Pt is 4. Higher coordination numbers of 5 or 6 with one or two additional ligands such as water or ammonia result in square-based pyramidal, tetrahedral, or octahedral structures (**3**) [22–25]. Complexes

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_213

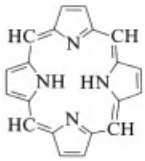

8

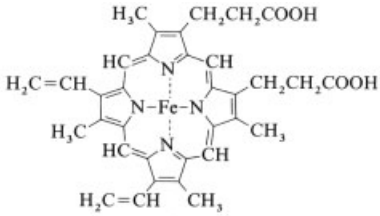

9

of lanthanides and actinides (e.g., $Nd^{III}HPc_2$ or $U^{IV}Pc_2$) have an unusual sandwich structure (**4**) with eight coordinated N atoms [26–28]. In contrast, reaction of 1,2-dicyanobenzene with uranium chloride in DMF gives a phthalocyanine ring with five isoindole units (**5**; super phthalocyanine) [29], [30]. A boron compound (**6**) with three former dinitrile units can be synthesized from boron halides and 1,2-dicyanobenzene [31].

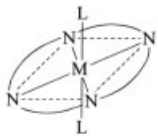

3

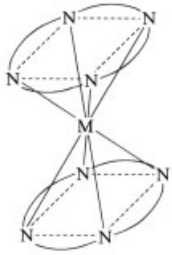

4

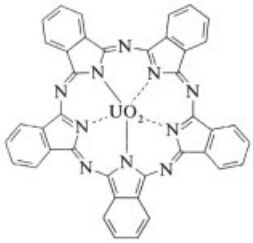

5

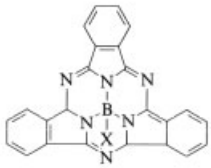

6

Phthalocyanines form polymers in which the central atom is a part of the polymeric chain. Elements such as Si^{IV}, Ge^{IV}, or Sn^{IV} yield covalently bonded polymers (**7**) [32–37]. The degree of polymerization n ranges from 65 to 140 for $(SiO_2Pc)_n$.

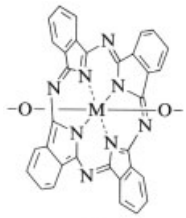

7

The phthalocyanines are structurally related to the macrocyclic ring system porphyrin (**8**). Formally, phthalocyanine can be regarded as tetrabenzotetraazaporphyrin and as the condensation product of four isoindole units (see above).

The phthalocyanines are structurally similar to naturally occurring porphyrins such as hemoglobin (**9**), chlorophyll a, vitamin B_{12} and turacin (a red dye from the wings of an African bird) [39]. Phthalocyanines themselves do not occur in nature.

Compounds with naphthalene or anthracene rings in place of the benzene nucleus also belong to the phthalocyanine family. The hydrogen atoms of benzene rings can be substituted in many ways (e.g., by halogens or by sulfonic acid, sulfonyl chlorides, or chloromethyl groups) (Chap. 5). Various polymers with interesting properties can be synthesized.

History. Braun and Tschernak [40] obtained phthalocyanine for the first time in 1907 as a byproduct of the preparation of *o*-cyanobenzamide from phthalimide and acetic anhydride. However, this discovery was of no special interest at the time. In 1927, de Diesbach and von der Weid prepared CuPc in 23 % yield by reacting *o*-dibromobenzene with copper cyanide in pyridine [41]. Instead of the colorless dinitriles, they obtained deep-blue CuPc and observed the exceptional stability of their product to sulfuric acid, alkalies, and heat. The third ob-

servation of a phthalocyanine was made at Scottish Dyes, in 1929 [42]. During the preparation of phthalimide from phthalic anhydride and ammonia in an enamel vessel, a greenish blue impurity appeared. DUNSWORTH and DRESCHER carried out a preliminary examination of the compound, which was analyzed as an iron complex. It was formed in a chipped region of the enamel with iron from the vessel. Further experiments yielded FePc, CuPc, and NiPc. It was soon realized that these products could be used as pigments or textile colorants. LINSTEAD et al. at the University of London discovered the structure of phthalocyanines and developed improved synthetic methods for several metal phthalocyanines from 1929 to 1934 [4–14]. For example, the reaction of dicyano aromatics and metal salts was patented in 1939 [43]. The important CuPc could not be protected by a patent, because it had been described earlier in the literature [41]. Based on LINSTEAD's work the structure of phthalocyanines was confirmed by several physicochemical measurements [44–50]. Methods such as X-ray or electron microscopy verified the planarity of this macrocyclic system. Properties such as polymorphism, absorption spectra, magnectic and catalytic characteristics, oxidation and reduction, photoconductivity and semiconductivity, solubility, and photochemical and dielectric properties were investigated from the 1930s to the 1950s.

Copper phthalocyanine was first manufactured by ICI in 1935, where its production from phthalic anhydride, urea, and metal salts was developed [51]. Use of catalysts such as ammonium molybdate improved the method substantially [52]. In 1936, I.G. Farbenindustrie began production of CuPc at Ludwigshafen, and in 1937 Du Pont followed in the United States. The most important of the phthalocyanines, CuPc, is now produced worldwide. The first phthalocyanine dye was a phthalocyanine polysulfonate [42]. Other derivatives, such as sulfonyl chlorides, ammonium salts of pyridyl phthalocyanine derivatives, sulfur and azo dyes, and chrome and triazine dyes, have been patented since 1930. At that time, the use of phthalocyanines as colorants for printing ink, paint, plastics, and textiles began. Of the industrial uses, the application of CuPc in printing inks is its most important use. The greenish blue CuPc shade is suitable for color printing. Other favorable properties such as light, heat, and solvent resistance led to the use of this blue pigment for paints and plastics. The chloro and bromo derivatives are important green organic pigments. Other derivatives are used in textile dyeing and printing or for the manufacture of high-quality inks (pastes for ballpoint pens, ink jets, etc.).

2. Physical Properties

2.1. General Data

The molecular mass of phthalocyanine, $C_{32}H_{18}N_8$, is 514.55. The densities of Pc compounds depend strongly on chemical composition:

β-H_2Pc	ϱ=1.43 g/cm^3
β-CuPc	1.61 g/cm^3
Polychloro CuPc	2.14 g/cm^3
β-CoPc	1.50 g/cm^3
β-NiPc	1.59 g/cm^3

The color of most Pc's ranges from blueblack to metallic bronze, depending on the manufacturing process. Ground powders exhibit colors from green to blue. Most compounds do not melt but sublime above 200 °C, which can be used for purification. An exception is $Si(OC_{18}H_{37})_2Pc$, which melts at 152 °C [53].

Solubility and Thermal Stability. H_2Pc, CuPc, and halogenated phthalocyanines have very poor solubility in organic solvents. Only in some high-boiling solvents such as quinoline, trichlorobenzene, and benzophenone is recrystallization possible at higher temperature. However, the solubilities have a maximum of several milligrams per liter. In common solvents such as alcohols, ethers, or ketones the solubility is considerable lower. One liter of benzene dissolves 0.15 mg of α- or γ-CuPc and 0.046 mg of β-CuPc at 40 °C [54]. The heats of solution are 34.3 kJ/mol for α-CuPc and 41.4 kJ/mol for β-CuPc [54]. The solubilities of other Pc complexes depend on the central atom.

Phthalocyanine and its unsubstituted metal derivatives dissolve in highly acidic media such as concentrated sulfuric acid, chlorosulfuric acid, or anhydrous hydrofluoric acid, presumably due to protonation of the bridging nitrogen atoms [55–57]. In the presence of strong bases, reversible deprotonation of the central imino

groups occurs [57]. The solubility in sulfuric acid depends on temperature and concentration [58]. In particular, the removal of copper from solid phthalocyanine by 25 – 65 % H_2SO_4 at 25 – 70 °C ranges from 0.6 to 9.2×10^{-6} mol/L after 5 to 100 h. The rate of decomposition of CuPc increases with increasing H_2SO_4 concentration, reaching a maximum at about 80 % H_2SO_4. The stability of metal phthalocyanines increases in the order: ZnPc<CuPc<CoPc <NiPc<$CuPcCl_{15}$ [59]. The color of phthalocyanine solutions in sulfuric acid depends on the degree of protonation (the N atoms of the ring systems are protonated by H_2SO_4; metals, e.g., Cu influence this protonation): H_2Pc gives a brownish yellow color; CuPc, greenish yellow to olive. The phthalocyanines can be precipitated from these solutions by addition of water. CuPc dissolves to the extent of 20 mg in 100 mL of liquid ammonia at − 33.5 °C [60]. Solubility can be improved in some cases by reversible oxidation with organic peroxides or hypochlorites; the Pc's are oxidized to substances soluble in organic solvents from which they can be regenerated by reduction [61].

Both H_2Pc and its derivatives exhibit high thermal stability. For example, CuPc can be sublimed without decomposition at 500 – 580 °C under inert gas and normal pressure [8]. In vacuum, stability up to 900 °C has been reported [62]. Polychloro CuPc is thermally stable up to 600 °C in vacuum. At higher temperature it decomposes without sublimation. CuPc decomposes vigorously at 405 – 420 °C in air. In nitrogen, sublimation and decomposition occur simultaneously at 460 – 630 °C [63], [64]. Generally all metal Pc's are more stable thermally in N_2 than in O_2. CuPc changes from the α- to the β-form at 250 – 430 °C [65]. Several MPc polyimides exhibit good thermal and thermooxidative stability at > 500 °C in air and nitrogen, and are therefore of interest for heat-stable films, coatings, and fibers [66].

2.2. Particle Size

Particle size and particle-size distribution are important criteria for Pc's used as pigments; they can be determined by the Debye – Scherrer-method [67], [68]; by adsorption measurements [67], [69], [70]; or by electron microscopy [67], [70–75]. The particle size calculated from adsorption data is ca. 0.044 µm for α-CuPc, 0.054 µm for β-CuPc, and 0.037 µm for polychloro CuPc [73].

The particle size of pigments dispersed, for example, in paint or printing ink binders can be determined by sedimentation analysis in centrifuges [69], [70], [75–79]. The particle diameters measured by this method are larger than those obtained by electron microscopy [69], [70]. This is due to incomplete dispersion of agglomerates from the pigment powder. These agglomerates can be dispersed totally only by prolonged grinding [69]. Lattice dimensions and crystal defects can be determined from the diffraction pattern that appears on high magnification under the electron microscope [80]. The particle size and particle form of Pc pigments are discussed in [70], [71], [73], [81].

2.3. Color Properties

Phthalocyanines absorb strongly in the visible range of 600 to 700 nm. They are thus blue to blue-green colorants. Their absorption spectra in solution are very different from those of the solids. The color of solid pigments is also influenced by the crystal modification [54], [82]. Color properties may depend to some extent on pleochroistic effects [70], [79] and on the nature of the central metal atom. In the series metal-free H_2Pc, NiPc, CuPc, CoPc with the same crystal modification, the color shifts from strongly greenish blue to strongly reddish blue, and the color purity decreases. The only metal phthalocyanine with a brilliant color suited for the cyan tone in color printing is CuPc. Color strength increases with decreasing particle size of the crystals, aggregates, and agglomerates. Figure 1 illustrates this for β-CuPc and polychloro CuPc [75].

Substitution of the benzene rings in phthalocyanine results in a color shift to greenish blue or green. Polychloro CuPc is bluish green. Partial substitution of the Cl atoms by Br atoms results in a yellowish green. Alkyl, aryl or other functional groups have a similar effect (Chap. 5.). A red shift can be attained by replacement of benzene rings by pyridine rings [83]. This compound has a more turbid color shade than CuPc, but it exhibits similar pigment properties.

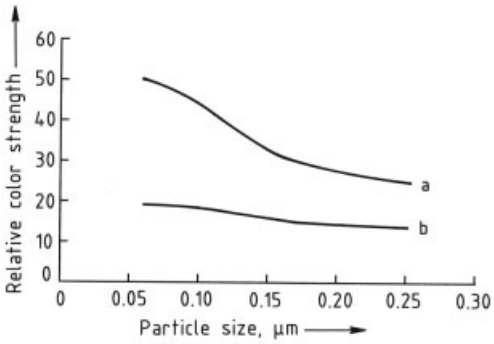

Figure 1. Relative color strength of β copper phthalocyanine (a) and polychloro copper phthalocyanine (b) as a function of particle size

2.4. Crystal Structure and Polymorphism

The crystal structure and polymorphism of phthalocyanines are reviewed in [84–87]. These are among the first organic solids whose crystal structures were determined by X-ray analysis of single crystals [45], [47], [48]. Phthalocyanine and its important metal derivatives (Fe, Co, Ni, Cu, Pt) have monoclinic lattices. Lattice constants and other X-ray data are tabulated in [88], [89]. The Pc ring system is always planar. Bond lengths and angles are influenced only slightly by the nature of the central metal atom [89]. Phthalocyanine and its metal derivatives occur in different crystal modifications [90], which has been shown by X-ray analysis and by IR and ESR spectroscopy. The modifications of H_2Pc, CuPc, and other derivatives differ in regard to color and physical properties. Five polymorphous modifications of CuPc are known; they are designated α, β, γ, δ, and ε [54]. Two more modifications have been described in the literature [27], [91–96].

According to their solubilities in benzene (good solubility implies instability), thermodynamic stability increases in the sequence $\alpha=\gamma<\delta<\varepsilon<\beta$ [54]. The enthalpy difference between α- and β-CuPC is 10.75 kJ/mol [97]. Three modifications of $H_2Pc-\alpha$, β [98], [99], and X [100] – have been found. Polychloro CuPc has only an α-modification [93], which is related to the α-modification of CuPc. All modifications exhibit one-dimensional stacking of the molecules (Fig. 2). However, the arrangement of the stacks with respect to one another differs, as illustrated by the structures of α- and β-CuPc (Fig. 2).

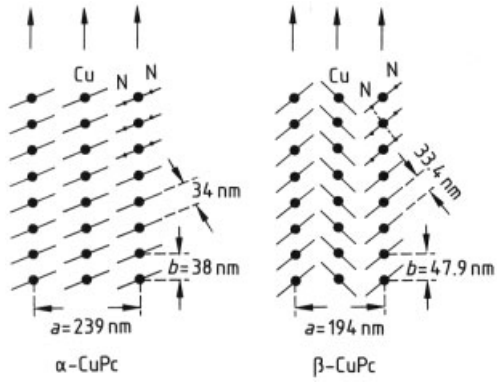

Figure 2. Arrangement of copper phthalocyanine molecules in the α- and β-modifications
The planar molecules appear as lines in this projection.

Phase transformations in crystals are of interest, impelled by the fact that the most desirable pigment in terms of color properties in a given medium may not be the most resistant to changes in crystal modification [2], [3], [54], [101]. During synthesis the most stable modification is formed; in the case of CuPc and H_2Pc this is the β-modification. Dissolving β-CuPc in concentrated sulfuric acid or treatment with 40 – 90 % sulfuric acid yields the finely divided α-form, after hydrolysis of the greenish yellow sulfate [2], [3], [99]. The β-modification is often converted to the α-modification by dry grinding in the presence of additives (e.g., salts). Figure 3 shows examples of the temperature-dependent $\beta \rightarrow \alpha$ transition by dry grinding. The rate of transformation decreases with increasing temperature.

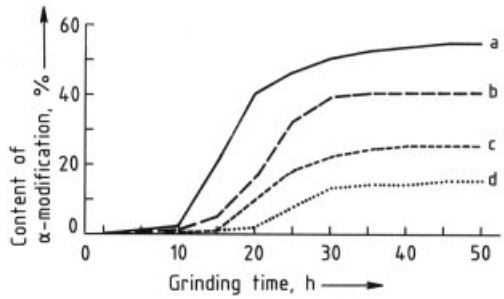

Figure 3. The $\beta \rightarrow \alpha$ transformation by dry grinding as a function of grinding time and temperature
a) 70 °C; b) 90 °C; c) 110 °C; d) 130 °C

Preparation of the γ- [102], δ- [103], and ε-modifications [104], [105] is described in the literature, and further modifications (π, x) are believed to exist [27], [91–96], [106]. All modifications can be transformed into the most stable form by heating in a high-boiling, inert solvent. In the absence of solvent, the transformation occurs at ca. 200 °C [2], [3]. Crystal growth has been studied by X-ray diffraction and electron microscopy [107], [108]. The transformation of CuPc in various organic media has been studied by electron microscopy and X-ray diffraction [109]. The preliminary growth of metastable α crystals precedes the transition to stable β crystals. Other studies of the $\alpha \rightarrow \beta$transition by visible spectroscopy, X-ray diffraction, electron microscopy, and BET determination of the specific surface area [110] indicate a three-stage process in terms of the classical theories of nucleation and crystal growth:

1) Ostwald ripening of the α-phase and nucleation of the β-phase
2) Disappearance of the α-phase and production of the β-phase
3) Final crystallization and ripening of the β-phase

The process is determined by nucleation of the stable phase [54], [111]. Treatment of β-CuPc with boiling solvents causes the β crystals to grow; treatment of α-CuPc with solvents causes the crystals to grow before transition to theβ-phase occurs, and the β crystals grow [112]. The behavior of metal phthalocyanines depends on the central metal ion [113]. Suspending α-CuPc with metal oxides such as alumina, magnesia, or zinc oxide in organic media results in an acceleration of the $\alpha \rightarrow \beta$ transformation [114]. Solvent-stable, flocculation-resistant α-, β-, and γ-modifications of CuPc and NiPc are prepared by combining phenyl-, chloro-, and sulfo-substituted tetraazaporphyrins with the MPc [115].

2.5. Surface and Interfacial Properties

A variety of surface and interfacial properties have been studied [2][3]. They include the effects of size reduction or milling, surface ionization energies [116], orientation overgrowth (epitaxial growth of crystals on a substrate surface) [117], the change of surface properties on crystal transformation [118], electric and photoelectric effects, interphase surface activity [119], and measurement of surface free energy and polarity [120]. Interfacial properties such as heats of adsorption and adsorption isotherms with organic liquids and water demonstrate the hydrophobic character of phthalocyanine pigments, which is in contrast to many organic and inorganic pigments [70], [73], [121]. In their properties, Pc's are very similar to graphite or diamond powder. Flocculation stability and rheological properties are influenced disavantageously. This effect, however, can be counteracted by addition of hydrophilic phthalocyanine derivatives [122–124]. In dispersions in organic media such as paints and binders, CuPc and polychloro CuPc are electrostatically charged [125], [126]. However, in contrast to many aqueous systems, this does not lead to stabilization of the dispersions. Adsorption of polymers improves flocculation stability. Both positive and negative charges are present on the same CuPc particles, which promotes flocculation, especially for rodlike structures [126]. Theoretical and experimental studies of the adsorption of a variety of organic compounds on phthalocyanines support the theory of privileged sites for adsorption on phthalocyanine at the central metal atom and along certain axes [127]. The stacking of phthalocyanine molecules results in two different surfaces. The surface parallel to the longitudinal axis of the stack is characterized primarily by benzene ring substituents (nonpolar hydrogen atoms). At the base of the stack, the dominant structural units are π-systems (nitrogen atoms and the central metal atom). Only the bases of the acicular crystals consist of easily polarizable atoms [128].

The specific surface area *s* of CuPc decreases on heating in vacuum at 300 °C from ca. 90 to 45 m^2/g, and its coloring power decreases by 10 % [2], [3], [121]. In vacuumα-CuPc , has *s*=30, 11, and 4 m^2/g at 20, 200, and 300 °C, respectively [121].

2.6. Other Physical Investigations

The physical properties of phthalocyanines have been investigated theoretically and in regard to application as semiconductors, catalysts, etc., by

a variety of physical methods. Absorption, fluorescence and phosphorescence, NMR, IR, Raman, and ESR spectra, as well as electric, photoelectric, magnetic, and semiconductor properties, are reviewed in [2], [3], [15], [129–133]. Some of these properties depend on the purity of the compounds [134]. The relationship to the catalytic properties of some Pc's is given in [135]. Mössbauer, ESCA, and inelastic tunneling spectroscopy, along with field ion and field electron microscopy, are reviewed in [2][3], [136], [137–140]. Although many of these physical properties suggest possible applications of phthalocyanines in semiconductors, catalysts, lubricants, lasers, or photography, they have not yet been exploited industrially.

3. Chemical Properties

The chemical properties of phthalocyanines are determined by the nature of the central atom, with each complex undergoing specific reactions. Nevertheless, some properties are universal.

3.1. Oxidation

The behavior of phthalocyanines upon electrochemical or chemical oxidation and reduction is important for determination of the redox potentials [141], [142] by cyclic voltametry or polarography, for the use of Pc's in preparative chemistry, and for application in photoredox processes [20]. Compared to porphyrins, phthalocyanines are easy to oxidize or reduce [20]. Oxidation can be reversible or irreversible, depending on the conditions. Phthalocyanines are stable to atmospheric oxygen up to ca. 100 °C. Stability depends on the central metal atom [143]. In aqueous solutions of strong oxidants, the phthalocyanine ring system is completely destroyed. Oxidation of one molecule of CuPc consumes one oxygen atom, which was confirmed by oxidation with cerium(IV) sulfate in sulfuric acid [10]:

$$[C_8H_4N_2]_4Cu + 3H_2SO_4 + 7H_2O + [O] \longrightarrow 4\,C_8H_5O_2N + CuSO_4 + 2(NH_4)_2SO_4$$

The phthalocyanine ring is oxidized to phthalimide. This reaction can be used for quantitative determination of CuPc content. Other possibilities are oxidation with sodium vanadate [144], nitric acid [145], or potassium dichromate in sulfuric acid [146]. During oxidation of CuPc, intermediates with an intense violet color are formed [8], [147–150]. Addition of a solution of CuPc in concentrated sulfuric acid to ice water containing 3 % nitric acid gives a violet precipitate, which can be converted back to the pigment by drying [151]. The stability of CuPc toward ozone indicates that this oxidant has little or no effect on its structure [152]. Adducts of $Fe^{II}Pc$, $Cr^{IV}Pc$, $Mn^{II}Pc$, and $Co^{II}Pc$ with nitrogen oxides are also known [151]. Well-defined oxidation products of phthalocyanines can be obtained in nonaqueous solution. Either oxidation occurs in the ring system, or the oxidation state of the central metal atom is changed. Oxidation of the ring system yield derivatives of tetracycloindolenines **10** and **11** [153], [154].

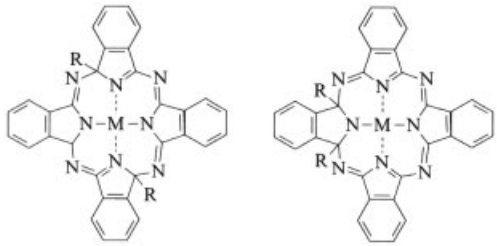

10 **11**

Oxidation with nitric acid in toluene, acetic acid, methanol, or pyridine yields **10** and **11**, with R=OH. Bromine in methanol or methanol – pyridine gives alkoxy or bromine derivatives (R=OCH_3 or Br) [61], [153], [154]. Acyl peroxides lead to acyl derivatives (R=OCOR) [154].

Oxidation products **10** and **11** are formed by metal-free phthalocyanine and MPc (M=Mg, Zn, Fe, Co, Ni, and Cu) [154]; polychloro CuPc does not react in this way. These oxidation products are much lighter than the blue dyes. For example, the nitrate of CuPc is yellow, and that of CoPc, yellow-orange [61], [153], [154]. Many of these complexes are crystalline and stable in the absence of reducing agents. They can be reduced to the corresponding Pc's in good yield [153], [154]. The same effect is observed on heating to 140 – 220 °C [61]. The good solubility of the oxidation products in organic solvents

such as chloroform, pyridine, DMF, and methanol is noteworthy. These phthalocyanine derivatives are of considerable industrial interest for textile dyeing. Their synthesis and applications are described in many patents (see Chap. 8).

Oxidation can also occur at the central metal atom of the phthalocyanine system. Transition metals, which occur in several oxidation states, exhibit this behavior. They usually form six-coordinate complexes on oxidation:

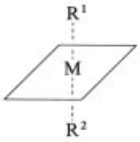

The chemistry of these metal phthalocyanines is reviewed in [15]. Manganese phthalocyanine, which occurs in five oxidation states from 0 to +4, is unique in this context [15]. Synthesis from manganese and phthalodinitrile yields $Mn^{II}Pc$ (**12**) [11]. Oxidation in organic solvents such as alcohol or chloroform or in the presence of a trace of acid yields $Mn^{III}Pc$ (**13**). In pyridine, dimethyl sulfoxide, diethylamine, or alkaline alcohol, this is converted to $Mn^{IV}Pc$ (**14**) [155–157]. The $Mn^{II}Pc$ reversibly chemisorbs oxygen in pyridine, which was first demonstrated by ELVIDGE and LEVER [156]. The oxygen absorbed at room temperature in pyridine solution can be removed by heating to boiling temperature. This process can be repeated several times, whereby small amounts of the phthalocyanine complex are decomposed. The mechanism of the process follows [157], [158]:

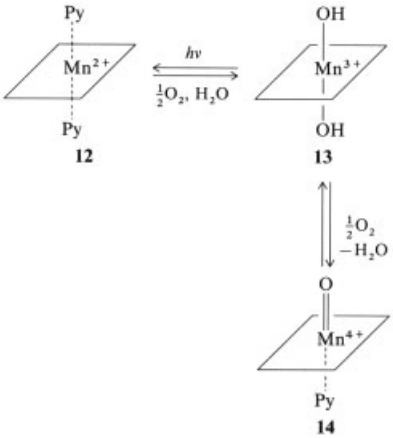

Complex **13** is not very stable; irradiation with light causes reduction to **12**. In the dark, **13** disproportionates to **12** and **14**. In the presence of oxygen, the oxidation of **13** continues to give **14**. In **14**, the manganese atoms are bridged by oxygen atoms [159]. This reversible uptake of oxygen seems to be analogous to oxygen transfer by heme, in which oxygen is fixed by a π-bond in the first step [160]. The catalytic properties of other phthalocyanines (e.g., FePc) can be interpreted in the same way [161].

Reduction of MnPc with lithium benzophenone in the presence of tetrahydrofuran results in formation of the lithium salt of $Mn^{I}Pc$ and the dilithium salt of $Mn^{0}Pc$ [162], [163].

The compound $Co^{I}Pc$ can be converted to $RCo^{III}Pc$ by reaction with alkyl halides [20], [164], [165]; $Co^{I}Pc$ can also be used as a selective reducing agent for aliphatic or aromatic nitro compounds [166] and as a reagent for cleaving β-halo alkyl groups, which are widely used to protect amino, carboxyl or phosphoryl groups in nucleotide or peptide chemistry [167]. Applications of MPc in ammonia synthesis [20], [168–170], Fischer–Tropsch synthesis [171], [172], or hydroformylation [173] are described in the literature.

3.2. Reduction

Either the central metal atoms or the carbon atoms of the phthalocyanine ring system can be reduced. The totally reduced form of the ring system, hexadecahydrophthalocyanine (**15**), can be produced from 3,4,5,6-tetrahydrophthalodinitrile [174].

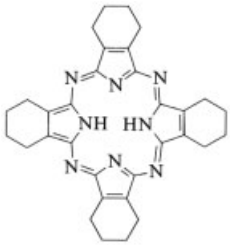

15

In reversible reduction, electrons are removed from the inner ring system or the central metal atom. In alkaline sodium dithionite solution, for example, conversion to a leuco form is

possible [42], [55], [175], [176]. The reduction products are weakly colored: greenish for CoPc and brownish yellow for FePc. With reducing agents such as alkali metals, multiple reductions ($MPc + n\,e^- \rightarrow MPc^{n^-}$; n=1 – 5 for Mn, n=1 – 4 for Ni) should be possible [20], [177]. For example, CuPc can be reduced with potassium in liquid ammonia [178], [179]. The solubility of the leuco form is usually too low for industrial purposes (e.g., CoPc). It can be increased by partial sulfonation. Indanthrene brilliant blue 4 G (C.I. Vat Blue 29, 74 140) is used for spin dyeing [180]. Although the structure of the leuco form has not been determined totally, a structure of type **16** can be assumed [181].

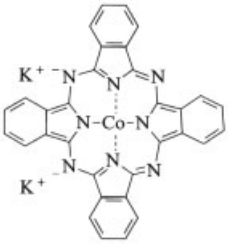

16

Reduction products with similar structures are known for other metal phthalocyanines [1, pp. 278 – 279]. Depending on the metal, reduction can take place either only in the ligand (Ni, Cu, Zn, Li) or first in the metal and then in the ligand (Cr, Mn, Fe, Co) [182].

3.3. Catalytic Properties

Many metal phthalocyanines, particularly the transition-metal complexes, can bind further ligands to form octahedral complexes. In some cases, this is accompanied by a change in oxidation state. This phenomenon is related to the catalytic properties exhibited by many members of this class of compounds [2, p. 61], [134], [135].

Two review articles deal with heterogeneous catalysis in dehydrogenation, oxidation, and electrocatalysis [183] and heterogeneous catalysis in gas-phase reactions by metal Pc's [184]. Catalysis by metal phthalocyanines was first discovered in 1936 in the exchange reaction between molecular hydrogen and water and the catalysis of water formation from oxygen and hydrogen [185]. Oxygen transfer reactions and catalysis of the decomposition of hydrogen peroxide were described later [186]. This suggested the similarity of metal phthalocyanines to heme and to enzymes such as hydrogenase [187]. Oxygen transfer reactions, particularly of iron phthalocyanines, were later observed in other reactions. Alkanes, olefins and aromatics, alcohols, aldehydes, alkyl aromatics, phenols, amines, thiols, cumenes, polymers, and sugars are oxidized by molecular oxygen in the presence of iron, copper, or cobalt phthalocyanine. In addition, hydrogenation, dehydrogenation, degradation, polymerization, isomerization, hydrogen-exchange reactions, reductive dehalogenation, hydrogenative thermal cracking, autooxidation, epoxidation, decarboxylation, and Fischer – Tropsch synthesis are catalyzed by phthalocyanines [2, p. 61], [3, I, p. 79 – 102], [15, p. 92], [134], [135], [183], [184].

A major interest in the reduction of molecular oxygen by Pc catalysts is the development of fuel cells [188]. The recognition of phthalocyanines as oxygen reduction catalysts is due to Jasinski, Jahnke, and Schoenborn, who studied catalysts for electrochemical reactions in fuel cells [189–192]. The phthalocyanines were the first platinum-free catalysts suitable for this purpose. The electrocatalytic activity of metal Pc's depends considerably on the central metal ion [161], [186], [193]. The order of decreasing electrocatalytic activity is Fe>Co>Ni>Cu>H, with H_2Pc having very low activity [3, I, p. 81]. The high activity and stability of FePc electrodes are reported in [194], [195]. Light is involved in the electrocatalytic reduction of oxygen by FePc, CuPc, and H_2Pc on a graphite carrier [2, I, p. 81]. Phthalocyanines also catalyze the reduction on carbon cathodes of substances other than oxygen [196].

The mechanism of these heterogeneous catalysts is basically similar to that of other solid catalysts. An advantage is the homogeneity of the active centers. In contrast to the usual catalysts, the metal atoms here are always within the same electronic surrounding of the macromolecule in the surface of the crystals. Depending on the structure of the electron shells of the metal atoms, adducts can be formed with the reactants [161], [197], [198]. The catalytic activity is therefore closely related to the interactions between substrate and catalyst [197]. Activity increases with increasing electrical con-

ductivity of the solid phthalocyanines [135] and is also influenced by the carrier material [199]. The stability of metal complexes in the presence of other chemicals is still a problem, and in fact the most active compounds are often the least stable [161]. Polymeric phthalocyanines exhibit higher stability [134].

Polymeric CoPc, for example, is an effective catalyst in secondary fuel cells. The oxygen formed in electrolysis is stored in the polymer matrix and reduced during electrochemical discharge [200]. Sulfonic acids of CoPc and CuPc are used in the oxidation of thiols in crude oil [2, p. 300], [201].

4. Production

4.1. General Synthetic Methods

Phthalocyanine complexes have been synthesized with nearly all the metals of the periodic table [1], [15], [21]. Despite the apparently complex structure of the Pc system, it is formed in a single-step reaction from readily available starting materials. The reaction is strongly exothermic. For example, the synthesis of CuPc from phthalodinitrile ($4\,C_8H_4N_2 + Cu \rightarrow C_{32}H_{16}N_8Cu$) has a reaction enthalpy of -829.9 kJ/mol. The low energy of the final product can be accounted for by resonance stabilization; this explains at least partially the relatively facile formation of the complex. The most important metal phthalocyanines are derived from phthalodinitrile, phthalic anhydride, Pc derivatives, or alkali-metal Pc salts.

From *o*-Phthalonitrile.

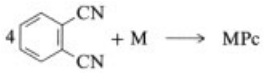

where M is a metal, a metal halide (MX_2), or a metal alkoxide [$M(OR)_2$]. The reaction is carried out in a solvent at ca. 180 °C or by heating a mixture of solid reactants to ca. 300 °C.

From Phthalic Anhydride.

$$\longrightarrow MPc + 4\,CO_2 + 8\,H_2O + X_2$$

This synthesis is carried out either in a solvent at 200 °C or without solvent at 300 °C.

From phthalimide derivatives, e.g., diimidophthalimide.

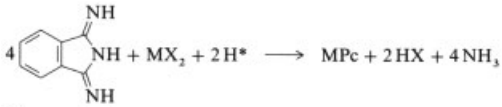

*from reducing agent

This synthesis is carried out in a solvent (e.g., formamide).

From Alkali-Metal Pc Salts. Dilithium phthalocyanine can be used as starting material for the production of other metal Pc's. The reaction is carried out in a solvent.

$$Li_2Pc + MX_2 \longrightarrow 2LiX + MPc$$

Metal-free phthalocyanine is obtained by the following procedures [1], [2], [3], [15]:

1) Decomposition of an unstable MPc with alcohol or acid

$$PcNa_2 + 2H_3O^+ \longrightarrow PcH_2 + 2Na^+ + 2H_2O$$

2) Direct synthesis (e.g., from phthalodinitrile)

Syntheses of MPc from phthalodinitrile or phthalic anhydride in the presence of urea are the two most important laboratory and industrial methods. They were also used originally by LINSTEAD and coworkers [11], [12]. This procedure allows the production of many phthalocyanine compounds [1], [2], [3]. Catalysts such as boric acid, molybdenum oxide, zirconium and titanium tetrachloride, or ammonium molybdate are used to accelerate the reaction and improve the yield [2][3]. Ammonium molybdate is especially effective. Reaction is carried out either in a solvent or by heating the solid components. When metal chlorides and phthalodinitrile are used as starting materials, the reaction products are partially chlorinated:

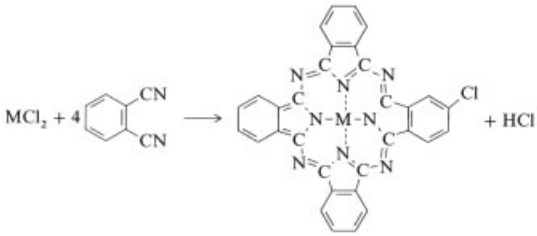

Lowering the reaction temperature or adding urea or basic solvents decreases the extent of chlorination. Solvents such as nitrobenzene, trichlorobenzene, alcohols, glycols, pyridine, and aliphatic hydrocarbons are employed. By using substituted phthalic acid compounds such as 4-chlorophthalic acid anhydride, 4-sulfophthalic acid anhydride or 4-nitrophthalimide, phthalocyanines with inner substitution can be produced. The products can often be purified by sublimation in vacuo at 300–400 °C. Soluble Pc's can be purified by recrystallization.

4.2. Reaction Mechanism

In the synthesis of phthalocyanines from phthalodinitrile or from pthalic anhydride and urea, intermediate isoindolenines (**17**) and polyisoindolenines are formed. According to [153], isoindolenines (**17**) are formed in the first step from dinitrile and polar compounds as follows:

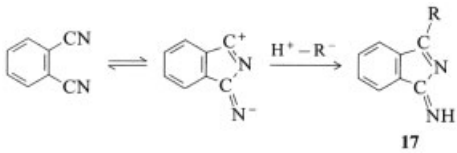

17

Metal salts, ammonia, primary or secondary amines, hydrogen sulfide, hydrogen bromide, or alkoxides are used as polar compounds. The isoindolenines condense to form the ring system around a central metal atom.

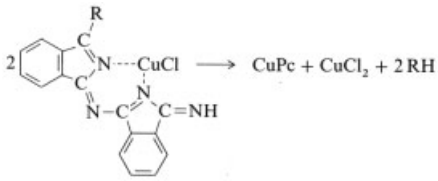

The intermediates cannot be isolated under normal synthesis conditions of ca. 200 °C. Some phthalocyanine developing dyes however use these intermediates, which react with metal salts on the fiber to give metal phthalocyanines (Chap. 8) [2], [3], [153], [202].

The synthesis of phthalocyanine from phthalic anhydride or similar phthalic acid derivatives also results in the formation of an intermediate compound [1], [2], [3]:

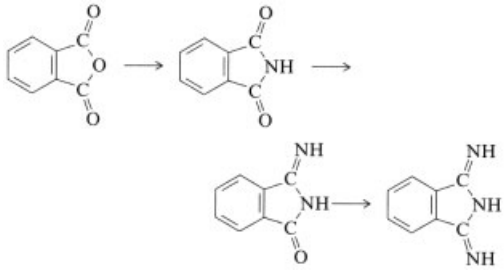

The 1,3-diiminoisoindolenine can be isolated as the nitrate [153].

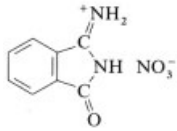

In a urea melt, the reaction involves formation of polyisoindolenines and their reaction with copper salts to form CuPc. Synthesis with ^{14}C-labeled urea or phthalimide showed that the urea carbon atom is not incorporated into the phthalocyanine framework but is released as ^{14}CO [203], [204].

4.3. Industrial Production

4.3.1. Copper Phthalocyanine

Two processes are commonly used for the production of copper phthalocyanine: the phthalic anhydride–urea process patented by ICI [51], [52] and the I.G. Farben dinitrile process [205]. Both can be carried out continuously or batchwise in a solvent or by melting the starting materials together (bake process). The type and

amount of catalyst used are crucial for the yield. Especially effective as catalysts are molybdenum(IV) oxide and ammonium molybdate. Copper salts or copper powder is used as the copper source [1], [2], [3]; use of copper(I) chloride results in a very smooth synthesis. Use of copper(I) chloride as starting material leads to the formation of small amounts of chloro CuPc. In the absence of base, especially in the bake process, up to 0.5 mol of chlorine can be introduced per mole of CuPc with CuCl, and up to 1 mol with $CuCl_2$. During pigment finishing (Section 4.4) these products can be processed to the stable α-modification, but not to the β-modification, of copper phthalocyanine.

The patent literature gives details of modifications and refinements of the original processes. A review of older processes is given in [1], [2], whereas examples of more modern production methods are described in [3], [206].

As apparatus for the batch process, an enamel or steel reactor with an agitator and pressure steam or oil heating suffices. Apparatuses used in the continuous synthesis in the presence of solvents and in the bake process are described in [207] and [208], [209], respectively. The choice of process depends on the availability and cost of the starting materials phthalodinitrile or phthalic anhydride. Although the phthalodinitrile process has certain advantages over the phthalic anhydride process, the latter is preferred worldwide because of the ready accessibility of phthalic anhydride. In this process the molar ratio of phthalic anhydride, urea, and copper(I) chloride is 4:16:1, with ammonium molybdate as catalyst. The mixture is heated in a high-boiling solvent such as trichlorobenzene, nitrobenzene, or kerosene. The solvent is removed after the formation of copper phthalocyanine. Frequently a purification step follows. A few typical examples are described below.

Phthalic Anhydride – Urea Process [52], [141, pp. 274 – 278]. Phthalic anhydride (100 kg), urea (135 kg), and trichlorobenzene (300 kg) (technical isomer mixture) are mixed in a 1000-L vessel, equipped with an oil bath and stirrer; 24 kg of copper(II) chloride and 0.5 kg of ammonium molybdate are then added. The mixture is heated up slowly within 1 h to 200 °C. Gas evolution (mostly CO_2 and some NH_3) with formation of phthalimide begins at 130 °C. Formation of CuPc begins at 160 – 170 °C, with simultaneous release of CO and NH_3. After stirring at 200 – 205 °C for 1 h, formation of the pigment is complete. The phthalocyanine is filtered off; washed with hot trichlorobenzene, trichlorobenzene at 50 °C, methanol, and hot water; and dried. The yield of CuPc is 87 kg (90 %).

Carrying out the reaction under pressure gives a high-purity CuPc pigment [210].

Several *dry processes* have also been described [3, II, p. 32].

In one of these [211], urea, copper(II) chloride, and ammonium molybdate are ground for 4 h in a ball mill, phthalic anhydride is added, and grinding is continued for 80 min. A small amount (10 to 60 % of the weight of phthalic anhydride) of solvent (nitrobenzene, naphthalene, or dodecylbenzene) is added, and the slurry is spread in stainless steel pans. These are heated in an oven at 195 °C for 8 h. The product is ground and slurried in dilute hydrochloric acid, filtered hot, washed with water, and dried. The yield is 84.5 – 86.6 % or 88 – 90 % when the reaction is carried out in a paddle drier [3].

The solvent can be replaced by ammonium chloride [212], a fourfold excess of phthalic anhydride, sodium chloride [213], or a 1:1 $NaCl-MgCl_2$ mixture [3], [214]. In the dry reaction, the ammonium molybdate catalyst can be replaced by a molybdenum or molybdenum alloy agitator [2], [3].

Another dry process is run continuously [215]. The dry, powdered reaction mixture is fed into a rotary furnace kept at 180 °C, and the dry product is discharged into a drum at a yield of 96 % [2], [3]. A vacuum method for the preparation of relatively pure CuPc is described in [216].

Phthalodinitrile Process. In the bake process [143, pp. 343 – 344], 100 kg of phthalodinitrile (ca. 95 % pure) is mixed with 24 kg of dry CuCl (1.25 mol) and 400 kg of anhydrous Na_2SO_4. The mixture is then slowly passed through a tunnel dryer within 1 h at 200 °C. The baked mixture is washed with hot water and 1 % hydrochloric acid, filtered off, and dried. The process can be carried out continuously.

One improvement of the process consists of grinding 100 kg of phthalodinitrile, 19 kg of anhydrous CuCl, and 20 kg of urea; mixing the

powder thoroughly (or grinding in a ball mill); and heating it to 150 °C. The temperature increases to 310 °C due to the heat of reaction, thus completing the reaction within a few minutes. After purification the yield is 97 %, and the product contains 0.3 % Cl [3, II, p. 30], [217]. Carrying out the reaction in the presence of a salt that decomposes at 30 – 200 °C to form ammonia improves the yield [218].

The reaction has also been carried out in solvents such as trichlorobenzene in the presence of pyridine [143]. Pyridine converts the insoluble copper(I) chloride into a soluble complex, which reacts more quickly.

A mixture of 300 kg of trichlorobenzene, 64 kg of dry CuCl, and 25 kg of pure pyridine is stirred at 80 °C for 30 min; 750 kg of trichlorobenzene and 150 kg of phthalodinitrile (95 % pure) are then added. The mixture is heated quickly to 200 – 205 °C and stirred for 6 h. The resulting melt can be purified as described above.

The reaction can be accelerated by the use of sodium hydroxide or sulfonic or carboxylic acids [219] instead of pyridine. Other high-boiling solvents such as nitrobenzene, benzophenone, or naphthalene can be used instead of trichlorobenzene [2], [3], [220–223].

The phthalodinitrile – solvent (pyridine) process can be used generally for the production of various metal phthalocyanines and is often suitable when other methods fail. A series of patents cover the low-temperature formation of copper phthalocyanine and continuous processes [3, II, pp. 31 – 32].

Dichlorobenzene Process. Until now, this synthesis [41] has found no use in industry. Heating *o*-dichlorobenzene (50 g) with quinoline (130 g), copper(I) bromide (49 g), and copper(I) cyanide (61 g) at 205 – 220 °C for 4.5 h gives 34.8 g of crude, well-crystallized CuPc (71 % yield) [143, pp. 274 – 278].

Diiminoisoindolenine Process. An alternative route is the formation of the isoindolenines, which are then treated with copper(II) salts [224–226]. 1,3-Diiminoisoindolenine is prepared by reaction of phthalonitrile with ammonia. The isoindolenine is then reacted with copper acetate in ethylene glycol and 2-chlorobenzonitrile at 60 – 70 °C for 1 h. The presence of benzyl cyanide can improve the solvent stability of the pigment [2], [3], [227], [228].

Copper phthalocyanine can also further be made by milling 145 g of diiminoisoindolenine, 24.8 g of copper(I) chloride, 700 g of anhydrous sodium sulfate, and 145 g of ethylene glycol for 3 h at 100 – 110 °C. After slurrying in water, filtration, reslurrying in 1 % hydrochloric acid, and dilution with sodium hydroxide, the product is suitable as a pigment [2][3], [229]. Other solvents and additives are also used for the reaction [230–233].

Production of γ-, δ-, ε-, π-, and X-CuPc. γ-CuPc is obtained by treating the α-modification with 30 % sulfuric acid and glycolmonobutylether at 110 °C [2, 3, 234–236].

δ-CuPc is formed when 25 g of CuPc is dissolved in 250 g of 98 % sulfuric acid and the resulting solution is added to 440 g of benzene, 800 mL of water, and 5 g of turkey red oil at 20 °C over 30 min. After emulsion separation and suspension of CuPc in the benzene layer for 48 h, it is washed with alcohol and water until acid free and dried at 70 °C [237]. Other examples are given in the literature[2], [3], [238].

ε-CuPc is made by mixing 350 g of urea and 295 g of 92 % sulfuric acid in a heavy-duty mixer at 80 °C; adding 400 g of 85 % CuPc made by the phthalic anhydride – urea process; and then adding 1500 g of sodium chloride and 100 g of water. After being mixed for 18 h at 50 – 60 °C the mixture is poured into 17 L of water, boiled, filtered, washed until acid and chloride free, and dried at 60 °C [2], [3]. Other variations are given in [182], [239–243].

π-CuPc is made by refluxing *o*-phthalodinitrile and copper(I) cyanide in 2-(dimethylamino)ethanol and ammonia [2], [3], [91], [92].

X-CuPc is prepared by synthesis of π-CuPc in the presence of a catalytic amount of X-CuPc [93], [94], by ball-milling of the α- and β-forms [95], or by vaporization methods [96].

For toxicological and ecological reasons, some of the solvents used in phthalocyanine synthesis, especially trichlorobenzene, must be replaced. High-boiling hydrocarbons [244–246] or esters [247] have been suggested as replacements. Other processes aim to produce the finished phthalocyanine pigment in order to avoid

the pigment finishing step (Section 4.4). In these processes, syntheses are carried out mostly in polar media (e.g., glycols) at low temperature and often lead to formation of the α-modification [248–253].

4.3.2. Phthalocyanine

Metal-free phthalocyanine has limited production for speciality uses. It is manufactured according to the original ICI process by reaction of phthalodinitrile with sodium amylate and alcoholysis of the resulting disodium phthalocyanine.

Under nitrogen, 20 kg of sodium is dissolved in 800 kg of amyl alcohol; 208 kg of phthalodinitrile is added, and the mixture is heated for 6 h to 125 °C. The Na_2Pc is filtered off, washed with amyl alcohol, and dried under vacuum. The dried and ground Na_2Pc is added to 850 kg of methanol at 5 – 10 °C, whereby metal-free Pc is formed. Stirring is continued until no variation of the sample is visible under the microscope. After dilution with water and filtration through a pressure funnel, washing with methanol and water, and drying under vacuum, H_2Pc is obtained as a powder in 80 % yield.

Phthalocyanine can also be obtained by melting phthalic anhydride and urea in the presence of catalytic amounts of ammonium molybdate [254], [255].

Anhydrous sodium polysulfide can be used instead of sodium alkoxides [256]. An electrochemical process that uses phthalonitrile as starting material in the presence of solvents is described in [257]. Other methods are given in [3, II, p. 29], [258–263], which also deal with production of the crystal modifications.

4.4. Pigment Finishing

Crude phthalocyanine or CuPc is obtained during production in the β-modification. The crystals thus obtained are too large to develop optimal pigment characteristics. They must therefore be subjected to a finishing or conditioning step in which the required modification, having optimal particle size with respect to properties such as color strength, color tone, color purity, grinding, flocculation, and rheology, is formed. Finishing methods are described in [2, p. 152], [264].

The particle size of crude phthalocyanine can be reduced by chemical or mechanical methods. The former involves dissolving phthalocyanine in concentrated sulfuric acid and precipitating it by addition to water. Phthalocyanines with a high halogen content must be treated with oleum or chlorosulfonic acid for purification. Phosphoric acid or sulfonic acids have been also proposed as solvents [2], [3]. The use of dilute sulfuric acid results in formation of a sulfate but leaves the pigment partially undissolved (soaking process). Depending on the sulfuric acid concentration, various sulfates are formed. Hydrolysis of the sulfate formed with 68 % sulfuric acid gives a very finely divided pigment [2], [3], [143]. Sulfates formed at other sulfuric acid concentrations must be ground before hydrolysis. The α-modification of the pigment is obtained by these methods. A variety of patents cover these and other processes [3, II, pp. 35 – 36].

The second finishing alternative involves grinding the dry or humid form, cold or warm, with or without addition of grinding agents such as sodium chloride. Organic solvents, especially aromatic hydrocarbons and their derivatives such as xylene, nitrobenzene or chlorobenzene, alcohols, ketones, or esters can also be used [2], [3]. Grinding is carried out in ball mills or kneaders.

In the absence of solvent, the unrefined pigment in the β-modification is converted to the α-form [2], [3], [264–266]. Figure 4 illustrates the change of particle-size distribution during milling (agglomeration of the primary crystals) and after finishing (deagglomeration and crystal ripening) of the crude pigment in an organic solvent.

Grinding additives such as sodium chloride are removed by boiling with water, and the pigment is isolated by filtration and drying. Little is known about the mechanism by which β-CuPc is ground and simultaneously transformed to the α-modification [264]. In the presence of solvent, this phase transformation is hindered, and a finely divided pigment in the β-modification is obtained.

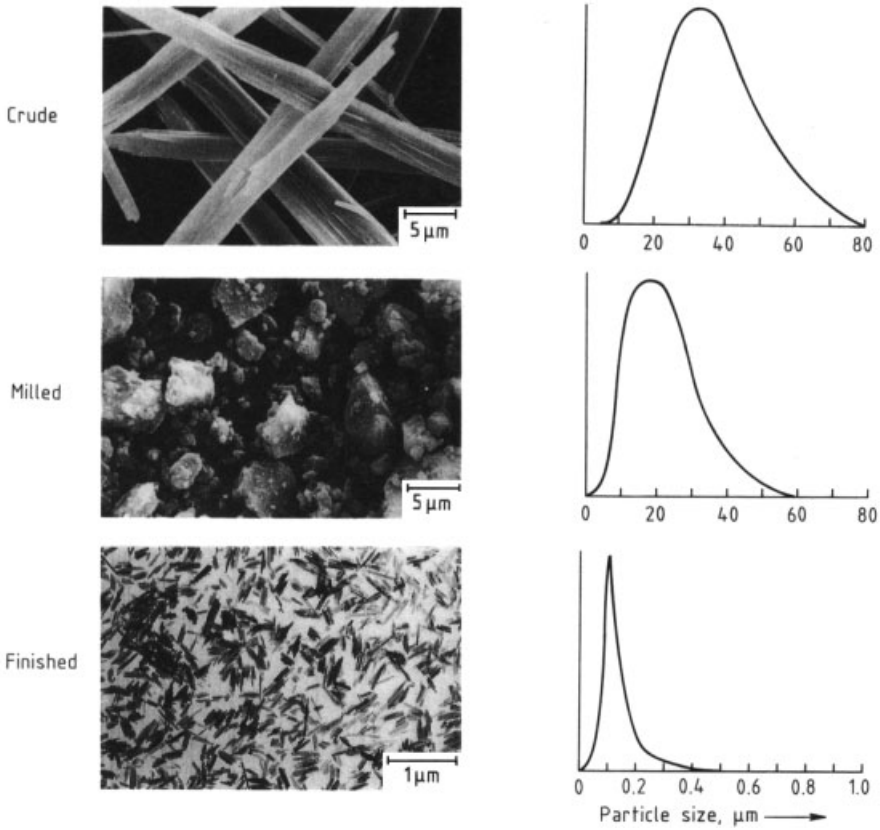

Figure 4. Particle-size distribution of crude copper phthalocyanine pigment after milling and finishing

4.4.1. Finishing by Treatment with Acid

Dissolving in Acid. Crude product is dissolved in concentrated sulfuric acid and precipitated with water. The degree of agitation and a correct rate of addition are important factors in precipitation [267]. The fineness of the suspension can be improved by addition of emulsifying agents [268], [2, p. 152] or simultaneous use of water-immiscible solvents [269], [2, p. 152], [270].

Hydrophilic agglomeration of pigment particles during drying must be avoided. Otherwise, the pigment may be difficult to disperse later. The tendency to agglomerate increases with decreasing particle size. Agglomeration can be reduced by various additives [2, p. 152] (e.g., ammonium naphthenate) [271], polyalkanolamines of fatty acids [272], nonaqueous solvents [273], or resins [274], [275]. Additives, particularly resins, coat the pigment particles, thereby impeding their agglomeration [276]. Pigments that are easy to disperse are obtained when they are mixed in water and the resin is deposited on the surface of the particles [277], [278].

Production of α-CuPc from Crude CuPc [143, p. 295], [198]. Crude CuPc (700 kg) is added quickly to concentrated sulfuric acid (96 %, 6300 kg), and the temperature is allowed to rise to 50 °C. The suspension is stirred overnight. When the phthalocyanine is completely dissolved, the suspension is sprayed into 35 000 L of water at 70 °C, and the temperature is allowed to rise to 85 – 90 °C. The precipitated pigment is filtered off in a filter press and washed until acid free. The press cake is then mixed with 25 000 L of water, and the remaining acid is neutralized with potassium carbonate and ammonia. After addition of 35 kg of lorol sulfonate the suspension is heated for 1 h until a clear filtrate is obtained. After filtration of the suspension, the press cake is washed with hot water and homog-

enized with sufficient water to give a 20 % paste, which is passed three to five times through a disintegrator (3000 rpm).

Pigment formation by soaking has several advantages compared to dissolution: less sulfuric acid is required, the unrefined phthalocyanine can be processed as a moist press cake, and the resulting pigment has improved color and dispersion properties. To obtain good results, the sulfuric acid concentration must be adjusted within exact limits. An industrial process in which the crude product is allowed to soak in 62 – 72 % sulfuric acid and then hydrolyzed with water [143, p. 294, 304] is carried out as follows:

At 35 °C, 200 kg of CuPc press cake (50 % water) is added to a mixture of 565 kg of 96 % sulfuric acid and 125 kg of ice. The relative density of a filtrate sample should correspond to a sulfuric acid concentration of 68 %. The suspension is stirred for 24 h. Conversion of the β-form to the α-form is monitored by microscopy or X-ray diffraction. Then 6000 L of water is pumped into the suspension. The acid-free washed pigment is again mixed with 2000 L of water. The mixture is made slightly alkaline with ammonia and boiled for 10 min, and the pigment is filtered off and dried under vacuum (14 h at 38 – 40 °C).

A similar soaking process can be used to form the green polychloro CuPc. Because the product is only sparingly soluble in sulfuric acid, it is stirred for 18 – 20 h at 35 – 40 °C in 95 % sulfuric acid, and the sulfate is decomposed at 70 – 93 °C with water.

Conditioning can also be carried out with polyphosphoric acid.

Process Example [279]. One part of crude CuPc is added to seven parts of polyphosphoric acid at 80 – 90 °C. Within 16 h the temperature rises to 195 °C; 30 parts of water (20 – 30 °C) is added, and the mixture is stirred for 2 h. The pigment is isolated by filtration.

For some uses, even a simple hot-water purification may be sufficient [280].

4.4.2. Finishing by Grinding

Finishing by grinding was first described in 1936 [281], [282]. In dry grinding, with salts such as sodium chloride or anhydrous sodium sulfate as grinding agents, the β-modification is converted to the α-modification. The coarse crystals of unrefined material are simultaneously ground to smaller size. Fine particles, high color strength, and good dispersive properties can be obtained when grinding agents are used [283]. Washing out the salts with water gives a color-intensive α-pigment. This procedure has been modified often [3, II, p. 37]. For example, pigment quality is improved by addition of surfactants, dispersants, or resins [2, p. 157], [3, II, p. 37].

The grinding process for β-CuPc is of greater importance than that for α-CuPc production [2, p. 157]. The simplest process involves grinding crude CuPc at 70 °C with a water-soluble salt and small amounts of aromatic or halogenated hydrocarbons or alcohols [2, p. 157], [3, II, p. 37], [284–288].

Process Example [288]. Iron rods (454 kg) of ca. 12.7-mm diameter and 24.4-mm length, crude CuPc (3.6 kg) (chlorine free), aluminum sulfate (22.7 kg), tetrachloroethylene (0.5 kg), and 0.15 kg of a surfactant containing cetyltrimethylammonium chloride are added to a 227-L ball mill. The mill is rotated for 12 h at 70 – 75 % of its critical velocity, and the temperature is allowed to increase to 60 – 70 °C. The charge is then removed and digested with hot, dilute sulfuric acid. A green CuPc pigment with high color intensity and good crystal stability is obtained. X-Ray analysis shows that the pigment is predominantly in the β-form.

Crude CuPc pigment can also be ground in the dry state without grinding aids. A surfactant is often added. This milled product exhibits satisfactory pigment properties only after further treatment in water immiscible organic solvents, for example, which are then removed by distillation after the finishing process is complete [289]. The β-modification of CuPc can also be produced by wet milling in the presence of surfactants [2, p. 157], [290]. Another widely used finishing process is the wet milling of crude pigments in mixers for highly viscous pastes (e.g., kneaders). Dry, crude pigment is kneaded with a water-soluble salt in an organic medium (polyalcohols or polyether alcohols).

Process Example [291]. In a kneader, 283 g of crude CuPc (chlorine free) is ground for 2 h with 1131 g of sodium chloride and 140 g of

poly(ethylene glycol) (M_r=400). Then, 30 g of this paste is boiled in 1 L of water and filtered to remove water-soluble substances. Washing is continued until no trace of water-soluble substance can be found. The moist paste is dried at 70 – 75 °C. A color-intensive β-CuPc is obtained.

A similar finishing process is used for partially or fully halogenated crude pigments. Another process for making an aqueous pigment dispersion involves activating the crude pigment in the presence of a phase-directing solvent (i.e., a hydrophilic agent that activates the pigment surface, facilitating its transfer to the aqueous phase), followed by fine milling in an emulsion containing a second phase-directing solvent [292].

Process Example [292]. A ball mill is charged with 906 g of chlorine-free crude CuPc and 45 g of turpentine and then rotated for 24 h at 70 % of the critical speed. An emulsion is produced by stirring 160 g of water, 12.5 g of turpentine, and 2.8 g of Aerosol OT (75 %) in a laboratory attritor for 2 – 3 min. At 375 rpm, 40 g of milled crude CuPc is added to the emulsion, and the water and turpentine are separated from the pigment slurry by filtration through a screen. The slurry is then added to 1 L of water containing sufficient hydrochloric acid to yield a pH of 1 – 2. After being heated to 75 – 90 °C for 30 min, the slurry is filtered, washed acid free, and dried.

4.5. Stabilization of Phthalocyanines

Unsubstituted and halogenated phthalocyanine and copper phthalocyanine possess a few principal disadvantages for industrial applications [2, p. 344], [125], [264]. In pure form they are extremely hydrophobic solids (Chap. 2.5) and therefore tend to flocculate in binders and paints [70], [73], which results in reduction of the color strength and poor flow properties of the dispersions. This effect is particularly apparent in combination with titanium dioxide [125]. The process can be observed by optical and electron microscopy [293]. A review with emphasis on CuPc and titanium dioxide in nonpolar media is given in [294]. A further disadvantage, especially of unchlorinated α-CuPc, is its instability in organic solvents, where conversion into the β-modification and formation of coarse crystals can occur [2, p. 165], [264].

These deficiencies are eliminated by various means. Unstable modifications, mainly the α-form of CuPc, can be stabilized by partial chlorination [2, p. 165]. Chlorine atoms in the 4-position are much more effective than those in the 3-position [170]. Partial sulfonation of CuPc increases flocculation stability. A combination of both methods leads to a high-quality pigment, particularly for solvent- containing coatings [2, p. 165]. Another method is to add phthalocyanines of tin, vanadium, aluminum, or magnesium [2, p. 165]. The addition of CuPc derivatives has also become important. These additives introduce new adsorption centers on the pigment surface. Monomeric and polymeric derivatives with basic substituents are, for example **18** [295] and **19** [296], [297].

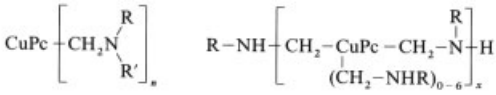

18, $n = 1-8$ **19**, $x \geq 2$

These molecules are very strongly adsorbed on the surface of the pigment, thus fixing basic centers that promote the adsorption of binder molecules. The colloid stability is due to the adsorbed layer of binder [122], [124], [276]. The addition of CuPc derivatives with long aliphatic chains gives pigments that also show high flocculation stability in the absence of polymers in aliphatic hydrocarbons [298]. Other additives are derived from the amides or salts of CuPc sulfonic acids [299–302].

These additives also strongly hinder recrystallization of the pigments and can therefore be used to stabilize unstable modifications such as the reddish, chlorine-free α-CuPc [303–306]. Other methods involve the use of surfactants [307–310] or adsorbed resin layers [311].

5. Phthalocyanine Derivatives

The copper phthalocyanine derivatives are of major industrial importance as green organic pigments (halogenated products) and dyes. The first pigments of this class were commercialized

in 1938. The first phthalocyanine dye was polysulfonated CuPc [42]. Since then, many patents describing various phthalocyanine compounds have been registered [1]. Substituted phthalocyanines are either accessible through synthesis from phthalocyanine derivatives (with the advantage of defined products) or through substitution of phthalocyanines [1, p. 255], [2, pp. 171, 192]. The latter method is favored in industry for economic reasons.

Synthesis. Usually 4 mol of a substituted phthalodinitrile or a substituted phthalic acid is used as starting material. A mixture of an unsubstituted and substituted starting material in approximate ratios, respectively, of 1:3, 2:2, or 3:1 can also be used. When reactivities of the two starting materials are approximately equal, Pc derivatives whose degree of substitution closely corresponds to the ratio of the starting materials are obtained. More often, however, a mixture of products results.

With the exception of tetrachlorophthalic acid, substituted phthalic acids, phthalimides, or phthalonitriles are industrially not readily accessible in pure form. Therefore, synthesis from phthalic acid derivatives is used only when the substitution reaction fails. For example, tetranitrophthalocyanines can be synthesized from 4-nitrophthalimide, whereas direct nitration of phthalocyanines gives only the oxidation products. Other substituted copper phthalocyanines such as tetra-4-phenyl copper phthalocyanine cannot be produced by direct introduction of the desired substituents.

Substitution. Copper phthalocyanine is preferred as starting material. Very little is known about the position of substitution. With the exception of hexadecachloro CuPc, all commercial Pc substitution products, as well as the tetrasubstituted derivatives synthesized from monosubstituted phthalic acids, are mixtures of isomers. Despite the 16 hydrogen atoms that can be substituted, only two different monosubstituted Pc's are possible. The number of disubstituted isomers is higher. Mono- to heptasubstituted Pc derivatives have not yet been isolated in isomerically pure form. In addition, only a limited number of isomers are accessible in pure form by synthesis. Only symmetrically substituted phthalic acids, phthalimides, or phthalodinitriles (3,6-di-, 4,5-di-, or 3,4,5,6-tetrasubstituted derivatives) yield pure isomers of octa- or hexadecasubstituted phthalocyanine derivatives. All other substituted phthalic acids give mixtures of isomers.

5.1. Halogenated Phthalocyanines

Perchloro and perchlorobromo copper phthalocyanines are of major industrial importance as green organic pigments with high lightfastness. Commercial development of these pigments began in 1935 [312], [313]. They are obtained by direct chlorination of copper phthalocyanine. The color changes only slightly from blue to green with the introduction of the first five to seven chlorine atoms, but a strong change in color occurs when the tenth to fourteenth chlorine atoms are introduced. Introduction of the last two chlorine atoms does not influence the shade. Replacement of chlorine by bromine displaces the color from blue to yellowish green [314]. These products were first marketed in 1959 [315]. Because the introduction of 15 to 16 bromine atoms is very difficult, commercial yellowish CuPc pigments thus contain 11 to 12 bromine atoms and 4 to 5 chlorine atoms per molecule. The halogenated derivatives are prepared by direct chlorination of copper phthalocyanine in a eutectic melt of aluminum and sodium chloride [143], [205], [313].

Hexadecachloro Copper Phthalocyanine [316]. In a 2.4-m^3 vessel equipped with a stirrer, 1800 kg of aluminum chloride and 420 kg of sodium chloride are heated to give a homogeneous melt; the temperature rises to 160 °C. The temperature is lowered to 100 – 110 °C in 1 – 2 h by external cooling; then 600 kg of CuPc is added over a period of 1 h. Chlorine is introduced initially at a rate of 30 kg/h, which is increased to 50 kg/h after 30 min and then to an average rate of 100 kg/h when addition of CuPc is complete. The temperature is kept below 150 °C by external cooling and regulation of the chlorine flow rate between 80 and 120 kg/h. The chlorine flow is stopped when the required degree of chlorination is attained (normally after 11 – 13 h). The melt is then heated to 180 °C, held at this temperature for 30 min, and

discharged into water. The polychlorophthalocyanine is isolated from the aqueous suspension. The yield is 1100 kg, with a chlorine content of 48 – 49 %. The aluminum content is < 0.05 %.

When CuPc is halogenated with bromine instead of chlorine in an $AlCl_3$ – NaCl melt in the presence of copper(II) chloride, a polybromo-CuPc is obtained that contains eleven bromine atoms and three chlorine atoms. A further development of the procedure includes the addition of sulfur trioxide or chlorosulfonic acid [317] and sulfur dichloride to the melt [318].

Low-melting complexes of aluminum trichloride with SO_3, $SOCl_2$, or SO_2Cl_2 can also be used as reaction media for the perchlorination of phthalocyanines, particularly CuPc, at 60 –200 °C. Especially significant is that on decomposition of this chlorination melt, perchlorinated CuPc precipitates in a form that does not require recrystallization from sulfuric acid [319].

Inorganic acids such as sulfuric or phosphoric acid are unsuitable solvents for chlorination because they lead to partial oxidative degradation of CuPc.

Organic solvents such as tri- or tetrachlorobenzene, acid chlorides, carbon tetrachloride, or nitrobenzene are also less suitable for halogenation of phthalocyanines [2, pp. 171, 192]. Only perchlorocyclopentene and perchloroindane have proved suitable [320]. Chlorination in molten phthalic anhydride or its chlorinated derivatives leads to complete chlorination of phthalocyanines. Chlorination of metal-free phthalocyanine (e.g., in phthalic anhydride) up to a content of about 12 chlorine atoms gives a yellowish green pigment [143]. Chlorination of CuPc in the absence of solvents is possible at 200 – 250 °C; the crude CuPc can be diluted with salts [321], [322].

Synthesis from chlorinated phthalic anhydride gives the α-modification of copper phthalocyanine derivatives directly. Either 4-chlorophthalic anhydride and urea or phthalodinitrile is used as starting material under conditions that lead to partial chlorination [323], [324]. The crude product often contains some of the β-modification, which is converted to the stable α-modification during pigment finishing.

Fluorinated phthalocyanines are also known [325], [326]. Hexadecafluorophthalocyanine can be prepared from tetrafluorophthalocyanine [203]. These products have surprising properties; for example, they are soluble in many organic solvents.

Halogenated phthalocyanines must be converted to useful pigments by finishing (Section 4.4). Solvents such as methanol, butanol, ethylene glycol, triethanolamine, esters of aliphatic and aromatic carboxylic acids [327], mixtures of aliphatic ketones and water, [328] or benzoic acid [329] are used as finishing agents.

5.2. Phthalocyanine Sulfonic Acids and Sulfonyl Chlorides

The sulfonic acids and sulfonyl chlorides, especially those of CuPc, are readily accessible. The sulfonic acids, $CuPc(SO_3H)_n$ with n=2, 3, or 4, were significant direct cotton dyes (C.I. Direct Blue 86, 74 186 and 87, 74 200). The sulfonyl chlorides are intermediates in the production of various copper phthalocyanine colorants [1, p. 259], [2, p. 192]. The watersoluble sulfonic acids are produced by heating copper phthalocyanines in oleum. By varying concentration, reaction temperature, and time, one to four sulfo groups can be introduced in the 4-position [330]. The products synthesized from 4-sulfophthalic acid exhibit slightly different properties. This is due to the different isomer distribution with respect to the 4- and 5-positions. Only one sulfo group is introduced in each benzene ring, as shown by the fact that only 4-sulfophthalimide is obtained on oxidative degradation. The most important dyes have two sulfonic acid groups per molecule and are prepared as follows:

Copper Phthalocyanine Disulfonic Acid [205]. Within 1 h, 150 kg of CuPc (97 – 98 % pure) is added to 1200 kg of 26 % oleum, with the temperature kept below 45 °C. The mixture is heated for ca. 12 h at 60 – 61 °C. The reaction is complete when a sample is fully soluble in dilute soda solution and the addition of sodium sulfate does not lead to any precipitation [198, p. 59]. The solution is then stirred into a mixture of 5000 L of saturated sodium chloride solution and 4000 kg of ice. The suspension is brought to 30 000 L with cold water and filtered through a filter press. The press cake is dried inside an air-circulating oven at 60 – 70 °C and then ground with soda.

Partially sulfonated cobalt phthalocyanine, which is an important dye (C.I. Vat Blue 29, 74 140), is prepared in a similar manner [331]. The corresponding disulfonic acid is used as an oxidation catalyst (Section 3.3). The cobalt [332], nickel [333], iron [334], chromium [335], manganese [336], and zinc [336] phthalocyanine sulfonic acids are synthesized by the urea process from 4-sulfophthalic acid.

The sulfonyl chlorides are prepared directly from chlorosulfonic acid and phthalocyanines [337]. Some of the sulfonyl chloride groups are hydrolyzed by the water formed in the reaction. Nearly pure phthalocyanine sulfonyl chlorides are obtained when phosphorus tri- or pentachloride or thionyl chloride is added. The positions of the sulfonyl chloride groups are the same as those of the sulfonic acids [330].

Copper Phthalocyanine Tetrasulfonyl Tetrachloride [338]. To 1500 parts of chlorosulfonic acid is added 193 parts of CuPc within 1 h. The mixture is heated 1 h to 70 – 75 °C, in 1.5 h to 75 – 130 °C, and for 4 h at 130 – 135 °C. The brown color of the starting material changes to green. After being cooled to 80 °C, 600 parts of thionyl chloride is slowly added and the mixture is stirred for 1 h at 75 – 80 °C. The solution is stirred into ice; the sulfonyl chloride is filtered off and washed with water. The tetrakis(3-sulfonyl chloride) is free of sulfonic acid groups.

A broad range of sulfonamide dyes are accessible by substitution of CuPc sulfonyl chlorides with amines. Dyes thus obtained can be used in nearly all fields of dyeing [1, p. 259], [2, pp. 171, 192].

5.3. Chloromethylphthalocyanines

After sulfonic acid chlorides, the chloromethyl derivatives constitute the second most important group of phthalocyanine derivatives. Like acid chlorides, they can be converted to a wide range of colorants by reaction with amines or phenols. In the original procedure [339] introduced by Haddock and Wood, the chloromethyl compounds are formed by substitution of phthalocyanines with bis(chloromethyl) ether in the presence of aluminum chloride and tertiary bases such as triethylamine or pyridine. Up to eight chloromethyl groups can be introduced. Another procedure involves the substitution of phthalocyanines with paraformaldehyde in sulfuric acid or chlorosulfonic acid [340]. Since bis-(chloromethyl) ether is carcinogenic [341], the reaction must be carried out in a hermetically sealed apparatus. Dichloromethyl ether is also produced in synthesis from paraformaldehyde in the presence of substances such as chlorosulfonic acid that produce hydrochloric acid.

5.4. Phthalocyanines with Other Substituents

Other phthalocyanine derivatives with various substituents, represented by formula **20**, are known for R=*tert*-butyl [342–346]; NR_2 [346–348]; OR [343], [344], [349–354]; SR [355], [356]; SiR_3 [343], [344]; CF_3 [357]; SCF_3 [358]; SO_2NR_2 [359–361]; COOH [362–366]; COOR [367], [368]; CN [347], [369–371]; NO_3 [372], [373]; and OH [374]. These derivatives are greener and more turbid than the copper phthalocyanines. More details and further references can be found in [1, p. 255], [2, pp. 192, 308].

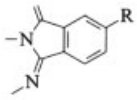

20

5.5. Polymeric Phthalocyanines

Combination of a phthalocyanine with a polymer or incorporation of a phthalocyanine into a polymer matrix is a powerful tool for designing new materials with special properties. Polymer phthalocyanines are of interest as catalysts, semiconductors, and lubricants because they are more stable than the corresponding monomers.

The following principles are used to form phthalocyanine polymers:

1) Linking through the central metal atom as part of the polymeric chain [30–35]
2) Linking of neighboring phenyl rings (diphenyl structure)
3) Linking through substitution (e.g., ether bridges, ethylene bridges)

4) Linking through phenyl rings that belong to two phthalocyanine systems
5) Linking of the phthalocyanines to a polymer chain
6) Linking the metal of a phthalocyanine with a polymer donor ligand, or electrostatic bonding of a charged phthalocyanine with a charged polymer chain

Details of the polymeric phthalocyanines are given in [1, pp. 274 – 275], [2, p. 328], [3, II, p. 91], [375, pp. 59 – 125].

5.6. Phthalocyanine Analogues

Phthalocyanine analogues include compounds such as porphyrins (**8**, Chap. 1), in which the four pyrrole rings are linked through one to four aza bridges and up to eight β-hydrogen atoms of the pyrrole are substituted [2, p. 308], [3, I, p. 195]. Like the phthalocyanines, many of these compounds exhibit very low solubility and are therefore potential pigments. Many such derivatives have been synthesized and studied. Because of their high cost and various application problems, none of these products have found use in industry.

Replacement of Benzo Rings by Other Ring Systems. The four benzo rings of phthalocyanines can be replaced by other aromatic ring systems such as naphthalene or aromatic heterocycles, for example, **21**–**27**.

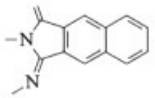

21

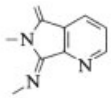

22

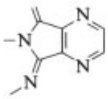

23

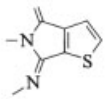

24

These compounds are derived from the corresponding *o*-dicarboxylic acids, their anhydrides, or *o*-dinitriles. The naphthalophthalocyanines (**21**) [376], their derivatives, [377], [378], and their metal complexes show reactions similar to phthalocyanines [13], [379]. The analogue **22** is redder than phthalocyanine [143, pp. 334 – 336], [380], and **23** is redder still but also more turbid [143, pp. 334 – 336], [381]. These are the reddest phthalocyanine analogues known. The tetrathiophenotetraazaporphyrin **24** has a greenish color [382]. Crown ether derivatives (**25**) [383–385] and cumyl phenoxy derivatives [386] have also been reported. Octaphenyl Pc's (**26**) and analogues **27** [387] have been prepared from the corresponding diphenyls and dicyanotriphenylenes. An anthracene analogue of Pc is prepared from 2,3-dicyano-9,10-diphenylanthracene [388].

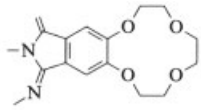

25

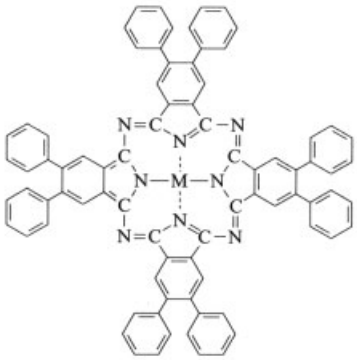

26

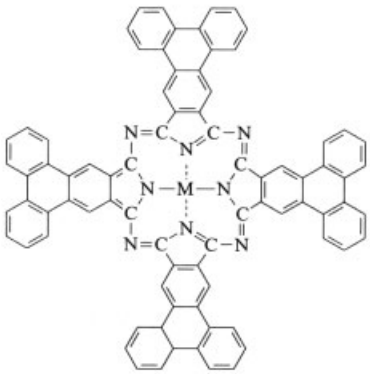

27

Derivatives of Tetraazaporphyrin. Alkyl and aryl derivatives of tetraazaporphyrin (**28**) such as tetramethyltetraazaporphyrin (**28**,

R=CH_3, R′=H) [389], octamethyltetraazaporphyrin (**28**, R=R′=CH_3) [390–392], and octaphenyltetraazaporphyrin (**27**, R=R′=C_6H_5) [393–395] belong to this group.

28

Hexadecahydrophthalocyanine can also be included in this group [177], [396]. The copper complexes of hexadecahydrophthalocyanine are among the reddest phthalocyanine analogues known. Starting materials for the latter two derivatives (the dialkylmaleic anhydride or dinitrile) [397] are not readily accessible; therefore these compounds are not in industrial use. Tetra- and octaphenyltetraazaporphyrins and their alkyl derivatives are potential oil-soluble dyes and siccatives [398–400].

Other Aza- and Thiaporphyrins. Metal complexes derived from mono-, di-, or triazatetrabenzoporphyrin or thiatetrabenzoporphyrins (**29**) are greener and more turbid than CuPc. They have no industrial importance. Tetrathio Pc's (**30**) are more strongly colored than the corresponding Pc's themselves [401–403].

29

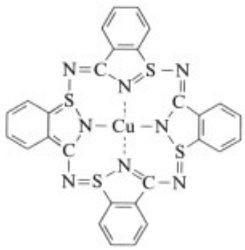

30

6. Legal Aspects and Environmental Protection

Although CuPc pigments exhibit very low toxicity, producers and users are faced with various regulatory changes, which are related to the presence of copper in these substances [404], [405].

Since the inhalation of dusts and mists of copper salts can result in irritation of nasal mucous membranes, OSHA has established PEL of 0.1 mg/m^3 for copper fumes, and 1 mg/m^3 for dusts and mists of copper salts [405]. The German TA Luft calls for 5 mg/m^3 Cu in off-gases. The MAK value for copper is 1 mg/m^3 [341]. The issue of copper in wastewater is more complex because concerns are based on the toxicity of soluble copper to various aquatic species. The EPA ambient water criteria for copper call for a 4-day average of 12 µg/L (1 day max. 18 µg/L), whereas the maximum contamination level of copper for public drinking water is 1 mg/L [406]. Thus, the 12 – 18-ppb copper in wastewater from CuPc production may be a problem and is handled by authorities on a case-by- case basis. Regulations in Germany call for 0.2 – 0.3 mg/L Cu in wastewater.

A method for removing copper(II) ions from wastewater in CuPc production is treatment with hydrosulfite to give a copper concentration of < 1 mg/L [407]. Another possibility is the addition of MoO_3, then the precipitation of molybdenum with a high molecular mass amine, and finally copper flotation with alkylxanthogenate [408]. The Resource Conservation and Recovery Act (RCRA; USA) of 1976 and the Hazardous and Solid Waste Amendments (HSWA; USA) of 1984 regulate the generation, transportation, treatment, storage, and disposal of hazardous and nonhazardous waste. Copper-containing waste is classified as nonhazardous under RCRA [405]; other regulations apply only to soluble or bioavailable copper, not to CuPc pigments. For the production of other metal phthalocyanine pigments, similar treatment of wastewater, off-gas, solid, and liquid waste; transportation; storage; and disposal involving the regulations must be taken into account, especially with respect to bioavailable metals.

Polychlorinated biphenyls (PCBs) have been detected in pigments manufactured in trichlorobenzene but not in those made with nonchlorinated solvents [409].

7. Quality Specifications and Analysis

Several procedures are applied to the testing of phthalocyanines. Tests for crystallization, flocculation, and applications in paints, plastics, and printing inks have been standardized [3, II, pp. 109 – 119].

Crystallization tests are carried out in xylene or in various paints [410], [411]. The flocculation test follows the ASTM standard specification for copper phthalocyanine [412]. This specification covers CuPc blue pigment in dry powder form for use in paint, printing ink, and related products. Tests include pigment composition, color, character of tint, oil absorption, reactions in identification tests, and dispersion and storage stability of the end product. Application tests for printing inks, paints, and plastics are described in [3, II, pp. 112 – 115], [412].

Identification methods consist of the trichloroacetic acid test [413], the N.P.I.R.I. standard test for pigment identification [414], and tests for various applications.

Quantitative determinations of the pigment are possible with cerium(IV) sulfate [10] or sodium vanadate [415]. Copper and chlorine determinations are often important [416].

8. Uses

Phthalocyanines are used predominantly (ca. 90 %) as pigments. These pigments are almost always copper phthalocyanines or their halogenated derivatives. This group of pigments accounts for 25 % of the total organic pigment capacity. Worldwide consumption of organic pigments in 1989 was estimated to be 250×10^3 t. Modification of the nearly insoluble phthalocyanines by introduction of various substituents has led to the use of Pc's as dyes for textiles, paper, and leather or as specialty dyes (e.g., for ballpoint pen inks). The colors of these dyes range from brilliant blue through turquoise to green. Phthalocyanine dyes are therefore used in almost all cotton dyes, synthetic-fiber dyes, sulfur and azo developing dyes, and reactive dyes.

8.1. Phthalocyanine Pigments

Because of their excellent color properties and high fastness to light, weathering, and various solvents, phthalocyanine pigments are used in nearly all fields. Furthermore, they are less expensive than other organic pigments. Phthalocyanine pigments are particularly important for gravure, offset, and flexo printing. The β-form of copper phthalocyanines is used predominantly in this field. Other important applications are in the paint, coating, plastics, textile, and spin-dyeing industries. Through special procedures during finishing of the crude pigment, optimal pigments are produced for each application. The use of additives leads to improved rheological properties and reduced flocculation. Apart from pure pigment powders, preparations are manufactured in which the pigment is already dispersed in a suitable binder system (e.g., polyethylene, nitrocellulose, polyamide, or liquid systems). The range of phthalocyanine products is now very large. For example, the Heliogen trademarks (powder pigments) of BASF include 48 products, of which 13 are the α-modification, 22 the β-modification, 2 the ε-modification, 1 is metal free, and 10 contain halogen.

Phthalocyanine pigment producers and trade names are as follows:

American Cyanamid, USA	Cyan Blue, Cyan Green
Aziende Colori Nazionali Affini, Italy	Turchese Segnale Luce, Verde Segnale Luce
BASF, Germany	Heliogen-Blau, Heliogen-Grün
Bayer, Germany	Helioecht-Blau, Helioecht-Grün
Ciba-Geigy, Switzerland	Irgalith-Blau, Irgalith-Grün
Dainichi Seika Colors, Japan	Chromofine Blue, Chromofine Green
Dainippon Ink and Chem. Co., Japan	Cyanine Blue, Cyanine Green, Fastogen Blue, Fastogen Green
Du Pont, USA	Cinquasia Blue, Cinquasia Green
Harmon Colors-National Aniline Division, USA	Phthalo-Blue, Phthalo-Green
Hilton Davis Chemical Co., USA	Peacoline Blue
Hoechst, Germany	Hostapermblau, Hostapermgrün
ICI, UK	Monastral Blue, Monastral Green
Kemisk Vaerk Koge, Denmark	Isophthalblau, Isophthalgrün
Produits Chimiques Ugine Kuhlmann, France	Cyanine Lutetia, Emerande Lutetia
Sandoz, Switzerland	Sandorin-Blau, Graphtol-Blau, Graphtol-Grün
Sumitomo Chemical Co., Japan	Sumitone Cyanine Blue, Sumitone Cyanine Green
Sun Chemical Co., USA	Fastolux Blue, Fastolux Green

Further information about manufacturers and trade names can be found under Pigment Blue 15, 74 160, in the Colour Index.

15 unstable α-CuPc
15:1 crystalline, stable α-CuPc
15:2 crystalline, stable, nonflocculating α-CuPc
15:3 β-CuPc
15:4 nonflocculating β-CuPc
15:6 ε-CuPc

C.I. Pigment Blue 16, 74 100 is metal free.

C.I. Pigment Green 7 is 74 260; C.I. Pigment Green 36, 74 265.

8.2. Phthalocyanine Dyes

The first phthalocyanine dye was a sulfonic acid derivative, synthesized in 1929 [42]. These dyes are still in use today to dye cotton. The sodium salts of copper phthalocyanine disulfonic acid (C.I. Direct Blue 86) and trisulfonic acid (C.I. Direct Blue 87) and their derivatives are also used. C.I. Acid Blue 185 is a wool dye. Finishing additives such as Solidogen (Hoechst) or Levogen (Bayer) improve the wetfastness of these CuPc sulfonic acids. Further applications are leather and paper dyeing.

Brilliant dyes that are readily soluble in alcohols and glycol ether can be obtained by reaction of Pc sulfonyl chlorides with amines. In addition to sulfonamide groups, these dyes contain ammonium sulfonate substituents. They can be used to color transparent paint, flexo and gravure printing ink, wood stains, plastics, and ballpoint ink. Commercial products include C.I. Solvent Blue 24, 74 380 and C.I. Solvent Blue 42, 44, 46, and 52. Substitution of copper phthalocyanine sulfonyl chlorides with primary amines that also contain a tertiary or secondary amino group leads to dyes that are soluble in acidic media. They are used mostly for printing ink, ballpoint ink, ink ribbons, copying ink, and carbon paper. Commercial products include Astrablau 6 GLL (Bayer) and Blaubase KG (BASF).

In addition to these widely applicable phthalocyanine dyes, a number of specialty dyes are used in the textile industry. In one case, functional groups are introduced into the phthalocyanine molecule, which solubilize the compound and are later cleaved on the fiber by further treatment. For example, the Inthion dyes (Hoechst) have thiosulfate esters in the side chains that are cleaved during the dyeing process [417].

The reaction of trichloromethyl copper phthalocyanine with tetramethylthiourea gives an isothiuronium salt, whose hydrophilic groups can be cleaved by weak bases [418]. Trade names can be found under C.I. Ingrain Blue 1, 74 240 (e.g., Alican Blue 8 G, ICI). Another possibility is linking a copper or nickel phthalocyanine derivative with a reactive group on the fiber (e.g., via sulfo groups) [419]. Commercial products of this type are listed in [420]. Examples are the copper phthalocyanine derivatives C.I. Reactive Blue 7, 74 460 (e.g., Cibacron Turquoise Blue, Ciba Geigy; Procion Turquoise MX-G, ICI; Remazol Turquoise Blue, Hoechst; Levafix Turquoise Blue E-G, Bayer; Primazin Turquoise Blue, BASF and nickel phthalocyanine derivatives such as Remazol Brilliant Green 6 B (Hoechst) and Cibacron Turquoise Blue (Ciba Geigy).

In contrast to the above-mentioned dyes, phthalocyanine developing dyes do not contain any substituents. They thus give purer, more genuine colors. Details of manufacturing, composition, and applications are given in [202], [421].

The phthalocyanine is generally synthesized from an isoindolenine or its derivative and a metal compound by drying and subsequent heating to 120 °C. The starting materials are impregnated in the fibers with water-soluble solvents such as glycol and diglycol monoethyl ether or diethanolamine. The solvents and additives act as reductants [421]. Important products are C.I. Ingrain Blue 2:1, 74 160 (e.g., Phthalogen Brilliant Blue IF 3 GM, Bayer and Sumilogen Brilliant Blue, Sumitomo). Other products include Phthalogen Turquoise 3 FBM and Phthalogen Marine Blue IRRM (Bayer).

Phthalocyanine dye producers and trade names are as follows:

American Cyanamid, USA	Calcomine Turquoise, Calcotone Green
Aziende Colori Nazionali Affini, Italy	Segnale Light Green, Eliama Light Turquoise
BASF, Germany	Lurantin Light Blue, Blaubase KG, Primazintürkisblau
Bayer, Germany	Astrablau, Sirius Supra Turquoise
Ciba Geigy, Switzerland	Chlorantine Fast Turquoise, Pergasol Turqoise Blue
Dainippon Ink and Chem. Co., Japan	Daivougen Blue, Symulon Turquoise, Fastogen Blue, Fastogen Green
Du Pont, USA	Pontamine Fast Turquoise, Brilliant Bond Blue
Hilton Davis Chemical Co., USA	Hiltamine Turquoise Blue
Hoechst, Germany	Remazol Turquoise Blue, Hostaperm Green
ICI, UK	Durazol Blue, Alican Blue, Monosol Blue
Sandoz, Switzerland	Pyrazol Turquoise, Cuprofix Blue, Finisol Blue Green
Sumitomo Chemical Co., Japan	Sumilight Supra Turquoise, Sumilogen Brilliant Blue

Further information can be found in the Colour Index:

C.I. Ingrain Blue 2:1, 74 160	developing dyes
C.I. Direct Blue 86, 74 180	CuPc disulfonate salt
C.I. Direct Blue 87, 74 200	CuPc trisulfonic acid
C.I. Ingrain Blue 3, 74 280	diiminoisoindolenine precursor of copper tetraphenylphthalocyanine
C.I. Solvent Blue 24, 74 380	CuPc, sulfonated
C.I. Reactive Blue 7, 74 460	CuPc tetrasulfonic acid containing 1 amide and 1 sulfophenylamide group

9. Economic Aspects

World capacity for crude phthalocyanine in 1987 was estimated at ca. 56 500 t [422]. Capacity in tonnes by geographic area is

North America	
United States	4 700
Mexico	750
South America	
Brazil	700
Argentina	250
Venezuela	250
Western Europe	
Germany	4 500
United Kingdom	3 100
France	2 000
Italy	3 700
Switzerland	200
Eastern Europe	3 000
Far East, Oceania	
Japan	18 000
India	3 400
South Korea	6 500
Taiwan	3 700
Thailand	300
China	1 000
Australia	400
Total	56 450

The major world producers are listed in Table 1. The producers were operating at ca. 80 % of capacity in 1989.

In the 1980s, only one new company in the United States began producing intermediates while many have dropped out. The greatest growth in intermediate production today is took place in India, Korea, and Taiwan, and these countries are likely to become even more significant producers [422].

Figure 5 shows the capacity distribution according to geographical regions; 59 % of the capacity is concentrated in the Far East. World capacity in the year 2000 is expected to reach 75 000 t.

Table 1. Major producers of crude phthalocyanine

Company	Location	Estimated 1989 capacity, t
Toyo/Cosmos	Japan, Taiwan	9000
DIC/Sudarshan	Japan, India	8000
Ciba Geigy	United Kingdom, Switzerland, Korea	4500
Phthalchem	United States	4500
BASF	Germany, Brazil	4500
Acna	Italy	3500
Woo Song	South Korea	3000
DNS	Japan	3000
ICI	United Kingdom, France	2000
Sigma	Taiwan	1500
Maintop	Taiwan	1000
Sanyo	Japan	1000

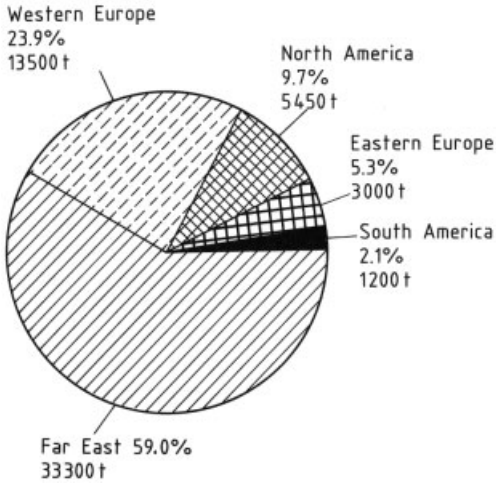

Figure 5. Production capacity for crude phthalocyanine according to geographical regions

Table 2. Consumption data for phthalocyanines (1989)

Application	Estimated consumption, t
Printing ink	19 500
Paint	9 500
Plastics	8 000
Textiles	3 000
Dyes	2 000
Specialty uses	2 000

Consumption data for phthalocyanines in paint, printing ink, plastics, textiles, dyes, and specialty uses (leather, paper, ballpoint ink) are listed in Table 2.

10. Toxicology and Occupational Health

Copper phthalocyanine pigments have found widespread acceptance because of their durability, excellent color, and good working properties. Adding to their attractiveness is the fact that Pc pigments are insoluble in water and lipids. No substantial scientific evidence exists to indicate that Pc pigments pose a significant risk to human health in the workplace or to the environment [404].

Acute Toxicity Tests. The CuPc pigments have LD_{50} values of > 10 g/kg in rats, which is more than 40 times less toxic than CuO or 70 times less toxic than $CuCl_2$ [404], [405]. The LD_{50} values were confirmed for the following CuPc pigments: phthalocyanine Blue 15 and Green 7 [423]. Doses of CuPc sulfonate up to 100 mg/kg in mice, guinea pigs, rabbits, cats, and dogs did not induce any toxic effects [424]. At concentrations of 5000 ppm, two commercial CuPc sulfonates were nontoxic to protozoans, small crustaceans, small fish, and, when injected, to rabbits, rats, and guinea pigs [425]. Skin and eye irritation studies [426] on CuPc and the Ames tests for mutagenicity were negative [427].

Chronic Toxicity Tests. No carcinogenic risk to humans was indicated in a study of laboratory mice, in which tumors failed to develop over an eight-month period [428]. A subchronic feeding study using doses up to 5000 mg/kg of body weight did not reveal any toxicity or pathological changes related to the Pc pigments administered [405]. The National Toxicity Program in the United States canceled a long-term bioassay on Pc's, because a 13-week feeding study of mice and rats did not indicate any toxicity [429]. Based on the very low solubility of the pigments in aqueous and nonaqueous media and the conclusion of nonbioavailability to humans, testing of a chemical fate and ecological effects of Pc green pigment was suspended by the TSCA Interagency Testing Committee. The FDA has approved the use of copper Pc in general and opthalmic surgery (e.g., for coloring polypropylene sutures, foils, and packaging), for contact lenses, and for food packaging [406]. The MAK value of copper in fumes is 0.1 mg/m^3 and in dust 1 mg/m^3. The TLV value for copper dust is 1 mg/m^3, which corresponds to 10 mg/m^3 CuPc [430] because the copper content of CuPc is usually <12 %. A Russian study recommends a maximum permissible concentration of 5 mg/m^3 in the workplace [431]. Workplace exposure can readily be reduced to meet these limits through engineering controls and protective equipment [404].

Studies of the uptake of sulfonated phthalocyanines (Ni, Zn, Cu, Co, Al) by cultured mammalian cells [432] and chick embryos [433] point to possible teratogenic effects. However, control experiments with copper(II) chloride suggest that at high concentrations copper is responsible for these findings rather than sulfonated phthalocyanine.

11. References

General References

1. G. Booth: "Phthalocyanines," in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. V, Academic Press, New York 1971, p. 241.
2. F. H. Moser, A. L. Thomas: *Phthalocyanine Compounds.* Reinhold Publ. Co., New York, Chapman and Hall, London 1963.
3. F. H. Moser, A. L. Thomas: *The Phthalocyanines,* vol. I and II, CRC Press, Boca Raton, Fla. 33431, 1983.

Specific References

4. R. P. Linstead, *Br. Assoc. Adv. Sci. Rep.* 1933, 465.
5. R. P. Linstead, *J. Chem. Soc.* 1934, 1016.
6. G. T. Byrne, R. P. Linstead, A. R. Lowe, *J. Chem. Soc.* 1934, 1017.
7. R. P. Linstead, A. R. Lowe, *J. Chem. Soc.* 1934, 1022.
8. C. E. Dent, R. P. Linstead, *J. Chem. Soc.* 1934, 1027.
9. R. P. Linstead, A. R. Lowe, *J. Chem. Soc.* 1934, 1031.
10. C. E. Dent, R. P. Linstead, A. P. Lowe, *J. Chem. Soc.* 1934, 1033.
11. P. A. Barret, C. E. Dent, R. P. Linstead, *J. Chem. Soc.* 1936, 1719.
12. P. A. Barret, D. A. Frye, R. P. Linstead, *J. Chem. Soc.* 1938, 1157.
13. J. S. Anderson, E. F. Bradbrook, A. H. Cook, R. P. Linstead, *J. Chem. Soc.* 1938, 1151.

14. R. P. Linstead, *Ber. Dtsch. Chem. Ges.* **A 72** (1939) 93.
15. A. B. P. Lever, *Adv. Inorg. Radiochem.* **7** (1965) 27.
16. L. J. Boucher in G. A. Melson (ed.): *Coordination Chemistry of Macrocyclic Compounds.* Plenum Press, New York 1979, chap. 7.
17. K. Kasuga, M. Tsutsui, *Coord. Chem. Rev.* **32** (1980) 67.
18. P. Sayer, M. Goutermann, C. R. Connell, *Acc. Chem. Res.* **15** (1982) 73.
19. C. Hamann et al. in P. Goerlich (ed.): *Organische Festkoerper und duenne Schichten,* Akademische Verlagsgesellschaft, Leipzig 1978, chap. 2.
20. D. Woehrle, G. Meyer: "Phthalocyanine – ein System ungewoehnlicher Struktur und Eigenschaften," *Kontakte* **3** (1985) 38; **1** (1986) 24.
21. F. Lux in C. J. Kevane, T. Moeller, *Proc. Rare Earth Res. Conf. 10th* 1973, 871.
22. M. S. Fischer, D. H. Templeton, A. Zalkin, M. Calvin, *J. Am. Chem. Soc.* **93** (1971) 2622.
23. T. Kobayashi et al., *Bull. Chem. Soc. Jpn.* **44** (1971) 2095.
24. F. Cariati, F. Morazzoni, M. Zocchi, *Inorg. Chim. Acta* **14** (1975) L 31.
25. J. R. Mooney, C. K. Choy, K. Knox, M. Kenney, *J. Am. Chem. Soc.* **97** (1975) 3033.
26. W. E. Bennet, D. E. Broberg, N. C. Baenziger, *Inorg. Chem.* **12** (1973) 930.
27. A. Gieren, W. Hoppe, *J. Chem. Soc., Chem. Commun.* 1971, 413.
28. K. Kasuga et al., *J. Am. Chem. Soc.* **102** (1980) 4835.
29. F. Lux et al., *Radiochim. Acta* **14** (1970) 57.
30. T. J. Marks, *J. Coat. Technol.* **48** (1976) 53.
31. A. Meller, A. Ossko, *Monatsh. Chem.* **103** (1972) 150.
32. D. Wöhrle, *Adv. Poly. Sci.* **50** (1983) 45.
33. K. Fischer, M. Hanack, *Chem. Ber.* **116** (1983) 1860.
34. J. Metz, G. Pawlowski, M. Hanack, *Z. Naturforsch.* **38 b** (1983) 378.
35. E. A. Orthmann, V. Enkelmann, G. Wegner, *Makromol. Chem. Rapid Commun.* **4** (1983) 687.
36. C. W. Dirk, T. Inabe, K. Schoch, T. J. Marks, *J. Am. Chem. Soc.* **105** (1983) 1539.
37. T. Inabe, J. W. Lyding, T. J. Marks, *J. Am. Chem. Soc., Chem. Commun.* 1983, 1084.
38. B. L. Wheeler et al., *J. Am. Chem. Soc.* **106** (1984) 7404.
39. O. Völker, *Naturwiss. Rundsch.* **8** (1955) 268.
40. A. Braun, J. Tscherniak, *Ber. Dtsch. Chem. Ges.* **40** (1907) 2711.
41. H. de Diesbach, E. von der Weid, *Helv. Chim. Acta* **10** (1927) 886.
42. Scottish Dyes, GB 322 169, 1929; DE 586 906, 1929.
43. ICI, GB 410 814, 1934; US 2 116 602, 1939 (I. M. Heilbron, F. Irving, R. P. Linstead).
44. R. P. Linstead, A. R. Lowe, *J. Chem. Soc.* 1934, 1031.
45. I. M. Robertson, *J. Chem. Soc.* 1935, 615;1936, 1195.
46. I. M. Robertson, R. P. Linstead, C. E. Dent, *Nature (London)* **135** (1935) 506.
47. I. M. Robertson, I. Woodword, *J. Chem. Soc.* 1937, 219;1940, 36.
48. R. P. Linstead, I. M. Robertson, *J. Chem. Soc.* 1936, 1195;1936, 1736.
49. E. W. Mueller, *Naturwissenschaften* **37** (1950) 933.
50. A. J. Melmed, E. W. Mueller, *J. Chem. Phys.* **29** (1958) 1037.
51. ICI, GB 464 126, 1935 (M. Wyler).
52. ICI, GB 476 243, 1936 (A. Riley).
53. P. C. Krueger, M. E. Kennedy, *J. Org. Chem.* **28** (1963) 3379.
54. D. Horn, B. Honigmann, *XII. Fatipec Kongreßbuch,* Verlag Chemie, Weinheim 1974, p. 181.
55. M. A. Dahlen, *Ind. Eng. Chem.* **31** (1939) 839.
56. D. L. Ledson, M. V. Twigg, *Chem. Ind.* **3** (1975) 129.
57. D. L. Ledson, M. V. Twigg, *Inorg. Chim. Acta* **13** (1975) no. 1, 43.
58. B. D. Berezin, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Technol.* **2** (1959) no. 10, 169.
59. B. D. Berezin, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Technol.* **6** (1963) no. 5, 841; **7** (1964) no. 1, 111.
60. G. W. Watt, J. W. Dawes, *J. Inorg. Nucl. Chem.* **14** (1960) 32.
61. C. J. Petersen, *J. Org. Chem.* **22** (1957) 127.
62. E. A. Lawton, *J. Phys. Chem.* **62** (1958) 384.
63. Y. Taru, K. Takoaka, *Shikizai Kyokaishi* **55** (1982) no. 1, 2.
64. M. Radulescu, R. Vîlceanu, *J. Therm. Analysis* **7** (1975) no. 1, 209.
65. M. Zhan, B. Zheng, X. Gu, *Xiamen Daxue Xuebao Ziran Kexueban* **25** (1986) no. 2, 192.
66. B. N. Archor, G. M. Fohlen, J. A. Parker, *J. Polym. Sci., Polym. Chem. Ed.* **21** (1983) no. 4, 1025; **21** (1983) no. 11, 3063; **22** (1984) no. 2, 319.

67. B. Honigmann, J. Stabenow, *VI. Fatipec Kongreßbuch,* Verlag Chemie, Weinheim 1962, p. 89.
68. F. M. Smith, *Br. Ink Maker* **7** (1964) 21.
69. M. Herrmann, B. Honigmann, *Farbe+Lack* **75** (1969) 337.
70. R. Sappok, B. Honigmann: "Organic Pigments," in G. D. Parfitt, K. S. W. Sing (eds): *Characterization of Powder Surfaces.* Academic Press, New York 1976, p. 231.
71. B. Honigmann, *Ber. Bunsen-Ges. Phys. Chem.* **71** (1967) 239.
72. W. Jettmar, *VII. Fatipec Kongreßbuch,* Verlag Chemie, Weinheim 1964, p. 343.
73. R. Sappok, *J. Oil Colour Chem. Assoc.* **61** (1978) 299.
74. W. Jettmar in J. Grehn (ed.): *Handbuch der Mikroskopie in der Technik,* **vol. 7,** Umschau Verlag, Frankfurt 1975, p. 369.
75. P. Hauser, M. Herrmann, B. Honigmann, *Farbe + Lack* **76** (1970) 545; **77** (1971) 1097.
76. J. Beresford, *J. Oil Colour Chem. Assoc.* **50** (1967) 594.
77. W. Carr, *J. Paint Technol.* **42** (1970) 696.
78. W. Carr, *J. Oil Colour Chem. Assoc.* **54** (1971) 155.
79. P. Hauser, D. Horn, R. Sappok, *XII. Fatipec Kongreßbuch,* Verlag Chemie, Weinheim 1974, p. 191.
80. J. Stabenow, *Ber. Bunsen-Ges. Phys. Chem.* **72** (1968) 374.
81. B. Honigmann, *Farbe+Lack* **82** (1976) 815.
82. F. M. Smith, J. D. Easton, *J. Oil Colour Chem. Assoc.* **49** (1966) 614.
83. IG Farbenind., DE 696 590, 1937 (B. Bienert); FIAT 1313, 3, 324 (1948).
84. M. Suito, N. Uyeda, M. Ashida, *Senryo to Yakuhin* **12** (1967) no. 2, 41.
85. M. Tomita, *Senshoku Kogyo* **21** (1973) no. 12, 725.
86. E. Suito, *Kotai Butsuri* **1** (1976) no. 3, 151.
87. E. Suito, *J. Electronmicrosc.* **18** (1969) no. 4, 341.
88. B. Honigmann, H. U. Lenne, R. Schroedel, *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **122** (1965) 185.
89. C. J. Brown, *J. Chem. Soc.* **A 1968,** 2488; **1968,** 2494.
90. G. V. Susisch, *Anal. Chem.* **22** (1950) 425; FIAT 1313, III, p. 412 (1948).
91. P. J. Brach, M. A. Lardon, DE-OS 2 218 788, 1972; GB 1 395 769, 1972; US 3 708 293, 1972; FR 2 138 865, 1972; CA 995 211, 1972.
92. P. J. Brach, H. A. Six, DE-OS 2 218 767, 1972; GB 1 395 615, 1972; US 3 708 292, 1972; FR 2 138 730, 1972; CA 996 931, 1972.
93. Xerox Corp., DE-AS 2 026 057, 1970 (P. J. Brach, H. A. Six).
94. P. J. Brach, H. A. Six, DE-OS 2 026 057, 1970; GB 1 312 946, 1975; US 3 927 026, 1975.
95. J. F. Byrne, P. F. Kurz, US 3 357 989, 1967; GB 1 169 901, 1967; DE 1 619 654, 1967.
96. J. H. Sharp, R. L. Miller, M. A. Lardon, DE-OS 1 944 021, 1970; GB 1 268 422, 1972; US 3 862 127, 1972; FR 2 016 641, 1972.
97. J. H. Beynon, A. R. Humphreys, *Trans. Faraday Soc.* **51** (1955) 1065.
98. M. Shigemitsu, *Bull. Chem. Soc. Jpn.* **32** (1959) 607.
99. B. Honigmann, *Farbe+Lack* **70** (1964) 787.
100. Xerox Corp., US 3 357 989, 1966 (J. F. Byrne, P. F. Kurz).
101. T. Silina, B. G. Aristov, *Zh. Vses. Khim. Ova. im. D. I. Mendeleeva* **19** (1974) no. 1, 77.
102. American Cyanamid Comp., US 2 770 629, 1952 (J. W. Eastes).
103. ICI, GB 912 526, 1962 (B. P. Brand).
104. Kemisk Vaerk Koge, GB 981 364, 1961; US 3 160 635, 1960.
105. BASF, DE-OS 2 210 072, 1972 (B. Honigmann, I. Kram).
106. Nippon Shokubai Kagaku Kogyo, DE-OS 2 659 211, 1977 (A. Komai et al.).
107. C. Hamann, H. Wagner, *Krist. Tech.* **6** (1971) 307.
108. M. Ashida, N. Uyeda, E. Suito, *J. Cryst. Growth* **8** (1971) 45.
109. E. Suito, N. Uyeda, *Kolloid Z., Z. Polym.* **19** (1974) no. 1, 77; **193** (1963) no. 2, 97.
110. B. Honigmann, D. Horn, *SCI Monogr.* **38** (1973) 283.
111. B. Honigmann, D. Horn: *Particle Growth in Suspensions,* Academic Press, New York 1973.
112. B. G. Aristov et al., *Congr. FATIPEC* **13** (1976) 112.
113. N. Uyeda, *Prog. Org. Coat.* **2** (1973) no. 2, 131.
114. N. Kawashima, T. Suzuki, K. Meguro, *Bull. Chem. Soc. Jpn.* **49** (1976) no. 8, 2029.
115. Bayer, DE-OS 2 104 200, 1972 (H. Leister, K. Wolf).
116. M. Pope, *J. Chem. Phys.* **36** (1962) 2810.
117. M. Ashida, *Bull. Chem. Soc. Jpn.* **39** (1966) no. 12, 2625; **39** (1966) no. 12, 2632.
118. K. Meguro et al., *Kogyo Kagaku Zasshi* **69** (1966) no. 6, 1724.
119. O. J. Schmitz, P. J. Sell, K. Hamann, *Farbe+Lack* **79** (1973) no. 11, 1049.

120. S. Wu, K. J. Brzozowski, *J. Colloid Interface Sci.* **33** (1971) no. 3, 686.
121. V. Y. Davidov, A. V. Kiselev, T. V. Silina, *Kolloidn. Zh.* **34** (1972) no. 1, 30; **36** (1974) no. 2, 359; **36** (1974) no. 4, 762; **36** (1974) no. 5, 945.
122. W. Black, F. T. Hesselink, A. Topham, *Kolloid Z. Z. Polym.* **213** (1966) 150.
123. K. Merkle, H. Schaefer, in T. C. Patton (ed.): *Pigment Handbook,* vol, 3, J. Wiley & Sons, New York 1973, p. 157.
124. A. Topham, *Prog. Org. Coat.* **5** (1977) 237.
125. V. T. Crowl, *J. Oil Colour Chem. Assoc.* **50** (1967) 1023.
126. W. Ditter, D. Horn, *Proc. Int. Conf. Org. Coat. Sci. Technol. 4th* 1978.
127. A. V. Kuznetov, C. Vidal-Madjar, G. Guiochon, *Bull. Soc. Chim. Fr.* **5** (1960) 1440.
128. U. Kaluza in: *Physical/Chemical Fundamentals of Pigment Processing for Paints and Printing Inks,* E. Moeller, Filderstadt 1981.
129. J. M. Assour, S. E. Harrison, *J. Am. Chem. Soc.* **87** (1965) no. 3, 651.
130. N. M. Ksenofontana et al., *Zh. Prikl. Spektrosk.* **205** (1974) no. 5, 834.
131. G. B. Birell, C. Burke, P. Dehlinger, O. H. Griffith, *Biophys. J.* **13** (1973) no. 5, 462.
132. H. F. Shurvell, L. Pinzuti, *Can. J. Chem.* **44** (1965) no. 2, 125.
133. M. Starke, H. Wagner, *Z. Chem.* **9** (1969) no. 5, 193.
134. K. Beales, D. D. Eley, D. J. Hazeldine, T. F. Palmer in H. Kropf, F. Steinbach (eds.): *Katalyse an Phthalocyaninen,* Thieme, Stuttgart 1973, p. 1.
135. H. Meier: *Organic Semiconductors,* Verlag Chemie, Weinheim 1974.
136. H. Fluck in [134], p. 37.
137. H. Rupp, U. Weser, *Biochim. Biophys. Acta* **446** (1976) no. 1, 151.
138. C. K. Joergensen, H. Berthov, *Mat.-Fys. Medd.-K. Dan. Vidensk. Selsk.* **38** (1972) no. 15, 15.
139. C. H. Kleint, K. Moeckel, *Surf. Sci.* **40** (1973) no. 2, 343.
140. S. de Cheveigne, A. Leger, J. Klein, *Proc. Int. Conf. Low Temp. Phys. 14th,* **vol. 3,** 1975, p. 491.
141. A. B. P. Lever, P. C. Minor, *Inorg. Chem.* **20** (1981) 4015.
142. A. B. P. Lever et al., *Adv. Chem. Ser.* **201** (1982) 237.
143. FIAT 1313, vol. III (1948).
144. G. G. Rao, T. P. Sastri, *Fresenius Z. Anal. Chem.* **169** (1959) 11.
145. N. P. Kanyaev, A. A. Spryskov, *Zh. Prikl. Khim. (Leningrad)* **25** (1952) 1220.
146. A. M. Islam, A. Waser, A. A. El-Mariah, A. A. Salman, *J. Oil Colour Chem. Assoc.* **57** (1974) no. 4, 134.
147. P. A. Barett, R. P. Linstead, G. A. P. Tuey, *J. Chem. Soc.* 1939, 1809.
148. R. P. Linstead, F. T. Weis, *J. Chem. Soc.* 1950, 2981.
149. A. E. Cahill, H. Taube, *J. Am. Chem. Soc.* **73** (1951) 2487.
150. P. George, D. J. E. Ingram, J. E. Bennett, *J. Am. Chem. Soc.* **79** (1957) 1870.
151. C. Ercolani, C. Neri, *J. Chem. Soc.* **A 1967,** 1715; **1968,** 2123.
152. C. W. Frank et al., *Electrophotogr. Int. Conf. 2nd,* 1974, p. 52.
153. F. Baumann et al. *Angew. Chem.* **68** (1956) 133.
154. Du Pont, US 26 628 957, 1951 (C. I. Pedersen).
155. H. A. Rutter, J. D. McQueen, *J. Inorg. Nucl. Chem.* **12** (1960) 361.
156. J. A. E. Elvidge, A. B. P. Lever, *Proc. Chem. Soc., London* 1959, 195.
157. G. Engelsma, A. Yamamtuoto, E. Markham, M. Calvin, *J. Phys. Chem.* **66** (1962) 2517.
158. A. Yamatuoto, L. K. Philips, M. Calvin, *Inorg. Chem.* **7** (1968) 847.
159. L. H. Vogt, A. Zalkin, D. H. Templeton, *Inorg. Chem.* **6** (1967) 1725.
160. J. S. Griffiths, *Proc. R. Soc. London, A,* **235** (1956) 23.
161. F. Beck et al., *Z. Naturforsch. A: Phys. Phys. Chem. Kosmophys.* **28 A** (1973) 1009.
162. R. Taube, H. Munke, *Angew. Chem.* **75** (1963) 639; *Angew. Chem. Int. Ed. Engl.* **2** (1963) 477.
163. R. Taube, M. Zach, K. A. Stanske, T. Heidrich, *Z. Chem.* **3** (1963) 392.
164. R. Taube, *Pure Appl. Chem.* **38** (1974) 427.
165. R. Taube, H. Drevs, D. Steinborn, *Z. Chem.* **6** (1978) 8.
166. H. Eckert, *Angew. Chem.* **93** (1981) 216; *Angew. Chem. Int. Ed. Engl.* **20** (1981) 208.
167. H. Eckert, Y. Kiesel, *Synthesis* 1980, 947.
168. S. Naito, K. Tamuro, *Z. Phys. Chem.* **94** (1975) 150.
169. M. Ichikawa, M. Soma, T. Onishi, K. Tamaru, *Bull. Chem. Soc. Jpn.* **41** (1968) 1739.
170. S. Mizuo et al., *J. Phys. Chem.* **73** (1969) 1174.
171. M. Ichikawa et al., *J. Am. Chem. Soc.* **91** (1969) 1538.

172. K. Watanabe et al., *Proc. R. Soc. London, A:* **333** (1971) 51.
173. E. H. Homeier, US 3 984 478, 1975.
174. G. E. Ficken, R. P. Linstead, *J. Chem. Soc.* 1952, 4847.
175. M. Shigemitsu, *Bull. Chem. Soc. Jpn.* **32** (1959) 544.
176. M. Shigemitsu, *Bull. Chem. Soc. Jpn.* **32** (1959) 502.
177. N. M. Bigelow, M. A. Perkins in H. A. Lubs (ed): *The Chemistry of Synthetic Dyes and Pigments,* Rheinhold Publ., New York 1955, p. 577.
178. G. W. Watt, J. W. Daws, *J. Inorg. Nucl. Chem.* **14** (1960) 32.
179. W. A. Alexander, P. L. Pauson, *J. Inorg. Nucl. Chem.* **17** (1961) 186.
180. Bayer, US 2 613 128, 1949 (B. Bienert, F. Baumann); GB 704 310, 1949. M. Kunz, *Text. Rundsch.* **6** (1951) 546.
181. F. Grund, *J. Soc. Dyers Colour.* **69** (1953) 671.
182. B. Honigmann, J. Kranz, DE-OS 2 210 072, 1973;GB 1 411 880, 1973;FR 2 174 089, 1973.
183. S. Manassen, *Catal. Rev.* **9** (1974) no. 2, 223.
184. F. Steinbach, H. H. Schmidt, M. Zobel, *Catal. Proc. Int. Symp.,* 1975, 417.
185. M. Calvin, E. G. Cockbain, M. Polanyi, *Trans. Far. Soc.* **32** (1936) 1436. M. Polanyi, *Trans. Far. Soc.* **34** (1938) 1191.
186. A. Cook, *J. Chem. Soc.* 1938, 1761, 1768, 1774, 1845.
187. D. Rittenberg, A. I. Krasna, *Discuss. Faraday Soc.* **20** (1955) 185.
188. J. H. Wood, C. W. Keenan, W. E. Bull in *Fundamentals of College Chemistry,* 3rd ed., Harper and Row, New York 1972.
189. R. Jasinski, *J. Electrochem. Soc.* **112** (1965) 526.
190. H. Jahnke, *Ber. Bunsen-Ges. Phys. Chem.* **72** (1968) 1053.
191. H. Jahnke, M. Schoenborn, *19th CITCE Meeting,* Detroit 1968; *Journ. Int. Etude Piles Combust. C. R. 3rd,* 1969, 60.
192. H. Jahnke, M. Schoenborn, G. Zimmermann, *Proc. Symp. Electrocatal.* 1974, 303.
193. H. Kropf, *Angew. Chem.* **84** (1972) 219; *Angew. Chem. Int. Ed. Engl.* **11** (1972) 239.
194. J. Yamaki, *J. Appl. Electrochem.* **15** (1985) 441.
195. J. Yamaki, *J. Electrochem. Soc.* **132** (1985) 2125.
196. S. Meshitsuka, M. Ichikawa, K. Tamaru, *J. Chem. Soc., Chem. Commun.* **5** (1974) 168.
197. H. J. Joswig, H. H. Schmidt, F. Steinbach, R. Stritzel, *Proc. Int. Congr. Catal. 6th,* 1976, 9583.
198. F. Steinbach, K. Mittner in [134]p. 122.
199. H. Meier, E. Zimmerhackl, W. Albrecht, M. Tschirwitz in [134]p. 104.
200. O. Hirabaru et al., *J. Chem. Soc., Chem. Commun.* 1983, 481.
201. Universal Oil Products, GB 849 998, 1959.
202. H. Vollmann: "Phthalogen Dyestuffs," in K. Venkatamaran (ed.): *The Chemistry of Synthetic Dyes,* vol. V., Academic Press, New York 1971, p. 283.
203. S. N. Brumfield, V. W. Folz, C. M. McGhee, A. L. Thomas, *J. Org. Chem.* **27** (1962) 2266.
204. S. N. Brumfield, B. C. Mays, A. L. Thomas, *J. Org. Chem.* **29** (1964) 2484.
205. BIOS Final Report 960.
206. Y. L. Meltzer: "Phthalocyanine Technology," *Chem. Process. Rev.* **42** (1970).
207. BASF, US 3 412 102, 1968 (G. Schulz, R. Polster).
208. Hoechst, DE 2 256 170, 1972 (W. Deucker, E. Spietschka, D. Steide).
209. E. F. Klenke Jr., US 2 964 532, 1960.
210. Dainichiseika Color Chem., JP 02 9851, 1981 (Y. Abe).
211. Ciba Geigy, GB 909 375, 1962; DE 1 202 419, 1967; CH 428 046, 1967 (D. Razavi, W. Fioroni, V. Meister).
212. Allied Chemical Corp., GB 991 419, 1965.
213. Y. Abe, M. Muto, JP 70 07661, 1970.
214. S. Ohira, M. Muto, JP 70 07662, 1970.
215. D. E. Mack, US 3 188 318, 1965.
216. T. P. Prassad, *Res. Ind.* **33** (1988) no. 2, 144.
217. E. Spieltschka, W. Deucker, DE-OS 1 644 679, 1970; GB 1 228 997, 1970.
218. J. C. R. Nicaise, L. A. Cabut, US 3 985 767, 1977.
219. H. Hiller, W. Kirshenlohr, DE-OS 2 045 908, 1972; GB 1 353 049, 1972;FR 2 107 595, 1972.
220. Y. Kuwabara, JP 65 27186, 1965.
221. Y. Abe, S. Horiguchi, JP-Kokai 77 36130, 1977.
222. R. Polster, R. Schroedel, D. v. Pigenot, GB 1 073 348, 1967; FR 1 410 814, 1967.
223. K. Hoelzle, CH 352 436, 1961.
224. S. Susuki, Y. Bansho, Y. Sakashita, K. Ohara, *Nippon Kagaku Kaishi* **51** (1976) 1460.
225. BASF, DE 2 250 938, 1972 (M. Gaeng).
226. Dainichiseika Colour Chem., JP 48 022 117, 1973.
227. M. Gaeng, DE-OS 2 250 938, 1974.
228. H. Leister, K. H. Wolf, R. Hoernle, DE-OS 2 012 507, 1971;GB 1 273 968, 1971.
229. T. Akamatsu, Y. Tezuka, DE-OS 2 136 767, 1972;GB 1 306 055, 1972;FR 2 103 195, 1972.

230. T. Akamatsu, Y. Tezuka, JP-Kokai 73 10128, 1973.
231. A. Komai, N. Shirane, JP-Kokai 75 127 928, 1975.
232. A. Komai, M. Ninomiya, N. Shirane, JP-Kokai 75 128 720, 1975.
233. T. Kasiwagi, Y. Kuwahara, JP-Kokai 76 22720, 1976.
234. Dainichiseika Colour Chem., JA 67 0970, 1967 (Y. Abe).
235. S. Suzuki, Y. Bansho, Y. Tanabe, *Kogyo Kagaku Zasshi* **72** (1969) no. 3, 720.
236. S. Sekiguchi, Y. Bansho, O. Kaneko, *Kogyo Kagaku Zasshi* **70** (1967) no. 4, 499.
237. B. P. Brand, US 3 150 150, 1964; GB 912 526, 1964.
238. Dainichiseika Colour Chem., JP 67 0870, 1967 (Y. Abe).
239. Ciba Geigy, DE 2 742 066, 1977 (I. R. Wheeler).
240. M. Miyatake, S. Tamura, S. Ishizuka, JP-Kokai 73 76 925, 1973.
241. I. Kumano, M. Miyatake, S. Tamura, S. Ishizuka, JP-Kokai 74 59 136, 1974.
242. Dainippon Ink Chem., JP 71 68950, 1981 (N. Fukada).
243. Dainippon Ink Chem., JP 71 49358, 1981 (N. Fukada).
244. Kawasaki Chem. Ind. Chem., JP 49 116 121, 1974 (R. Matsuura).
245. Dainichiseika Colour Chem., JP-Kokai 52 036130, 1977 (Y. Abe).
246. Toyo Ink MFG KK, DE-OS 2 642 416, 1976 (A. Sato et al.).
247. Toyo Ink MFG KK, JP-Kokai 52 010 326, 1977 (T. Funazo).
248. Ten Horn Pigmentchem., DE-OS 2 618 429, 1976 (J. J. Einerhand et al.).
249. Ciba-Geigy, DE-OS 2 637 202, 1976 (R. Barraclough, R. Laugley).
250. ICI, GB 1 374 793, 1972 (F. Hauxwell, H. R. Murton).
251. Japan Cataly. Chem. Ind. KK, JP-Kokai 50 124 927, 1975 (S. Kitano).
252. ICI, DE-OS 2 224 117, 1972 (F. Hauxwell, H. R. Murton).
253. Dainichiseika Colour Chem., JP 73 38332, 1971; JP 49 131 224, 1974 (Y. Abe).
254. Du Pont, US 2 820 796, 1954 (F. F. Ehrich).
255. ICI, US 200 005 152, 1932 (J. F. Thorpe, J. Thomas).
256. Ciba-Geigy, CH 579 131, 1976 (Z. Seha).
257. BASF, DE 2 711 005, 1977 (P. Thoma, W. Habermann, J. Kranz).
258. E. Stocker, A. Pugin, US 3 060 189, 1962; BE 609 166, 1962; GB 952 775, 1962; DE 1 218 639, 1962.
259. Z. Seha, DE-OS 2 364 689, 1974; GB 1 410 310, 1974; FR 2 212 399, 1974; JP-Kokai 75 157 417, 1974; CH 579 131, 1974; US 3 998 839, 1976.
260. BASF, GB 998 255, 1965; NL 6 405 836, 1965 (H. Mueller).
261. L. Weinberger, P. J. Brach, S. J. Grammatica, US 3 509 146, 1970; GB 1 232 241, 1970;DE 1 770 779, 1970.
262. Toyo Ink MFG KK., JP-Kokai 57 164 158, 1982 (R. Ohshima).
263. American Cyanamid Corp., US 4 320 059, 1980 (B. Casas).
264. F. M. Smith, J. D. Easton, *J. Oil Colour Chem. Assoc.* **49** (1966) 614.
265. Ciba Geigy, US 2 791 589, 1957 (A. Pugin).
266. Bayer, FR 1 265 104, 1960 (H. Raab, R. Hoernle).
267. Du Pont, US 2 334 812, 1940 (S. R. Detrick, C. R. Brandt).
268. Interchem. Corp., US 2 262 229, 1941 (V. Giambaivo).
269. American Cyanamid Corp., US 2 359 737, 1941 (H. T. Lacey, H. Z. Lecher).
270. BASF, EP 63 321, 1981; DE-OS 114 928, 1981 (T. Leary).
271. Du Pont, US 2 294 381, 1940 (H. E. Burdick).
272. Du Pont, US 2 305 379, 1941 (S. R. Detrick, J. W. Lang).
273. Du Pont, US 2 540 775, 1948 (S. R. Detrick).
274. Du Pont, US 2 173 699, 1936 (A. Siegel).
275. GAF Corp., DE-AS 1 003 885, 1955 (R. Brouillard, T. D. Mutaffis).
276. J. Moilliet, D. A. Plant, *J. Oil Colour Chem. Assoc.* **52** (1969) 289.
277. ICI, GB 978 242, 1964 (A. E. Ambler, R. W. Tomlinson).
278. ICI, GB 957 440, 957 984, 1964 (A. E. Ambler, R. W. Tomlinson).
279. Mobay Chem. Corp., US 304 127, 1981.
280. BASF Corp., CA 1 211 431 A, 1982; US 425 784, 1986.
281. IG Farbenind., GB 470 079, 1937 (G. W. Johnson).
282. General Aniline Works Inc., GB 457 426, 1936 (J. Y. Johnson).
283. Du Pont, US 24 021 670, 1942 (F. W. Lane).
284. American Cyanamid Corp., US 2 486 304, 1946 (S. A. Loukomsky).
285. Du Pont, US 255 672 630, 1949 (F. W. Lane).
286. Ciba, DE 846 757, 1950; 913 216, 1951 (W. Wettstein).
287. Geigy, FR 1 113 323, 1954.

288. Du Pont, DE-AS 1 161 533, 1960 (J. Jackson).
289. Du Pont, US 3 017 414, 1959 (J. W. Minnick, R. L. Sweet).
290. Bayer, DE-OS 2 160 208, 1971 (H. J. Moells et al.).
291. California Ink Co., US 2 982 666, 1958 (M. Chun, A. M. Enskine).
292. BASF Corp., US 123 670, 1980 (T. E. Donegan, J. H. Bantjes, T. G. Leary).
293. W. Jettmar, *Dtsch. Farben Z.* **24** (1970) 55.
294. R. B. McKay, *Surfactant Sci. Ser.* **21** (1987) 361.
295. BASF, GB 949 739, 1964; GB 985 620, 1965 (A. Schoellig, R. Schroedel, H. J. Sasse).
296. ICI, GB 1 082 945, 1964; GB 972 805, 1964 (W. Black, J. Mitchell, A. Topham).
297. B. G. Aristov, R. G. Feizulova, *Zh. Fiz. Khim.* **53** (1979) no. 7, 1859.
298. ICI, GB 1 343 606, 1971 (H. P. Dryhurst-Paget, L. R. Rogers, A. Topham, J. K. D. Royle, J. F. Stansfield).
299. Ciba-Geigy, DE-OS 2 720 464, 1977 (I. R. Wheeler, G. H. Robertson).
300. Ciba-Geigy, DE-OS 2 800 181, 1977 (I. R. Wheeler, G. H. Robertson); DE-OS 2 753 008, 1977.
301. ICI, DE-OS 2 461 567, 1974 (G. Hull, A. Topham).
302. V. A. Smrchek, S. G. Vilner, V. P. Gorin, G. A. Burgrov, *Lakokras. Mater. Ikh. Primen.* **2** (1984) 11.
303. BASF, DE 2 519 753, 1975; DE 2 516 054, 1975 (W. Jettmar, B. Honigmann, R. Sappok).
304. Produits Chimiques Ugine Kuhlmann, US 3 985 570, 1975; FR 7 041 758, 1970; FR 7 041 759, 1970 (L. A. Cabut, J. C. H. R. Hardouin, M. E. K. Huille, D. F. X. Pigasse).
305. Bayer, GB 1 422 834, 1973 (H. Leister).
306. L. Cabuf, J. C. Hardouin, H. Ciceron, A. Chapelle, *Double Liaison Chim. Peint.* **21** (1974) no. 224; **159** (1974) no. 225, 212.
307. A. Lorinc, C. Szterjopulosz, *Abh. Akad. Wiss. DDR* 1976, 529.
308. A. Lorinc, C. Szterjopulosz, T. Mezhdunar, *Kongr. Poverkhin.-Akt. Veshchestvam, 7th,* **3** (1976) 669.
309. B. G. Aristov, R. K. Feizulova, Y. G. Frolov, *Izv. Vyssh. Uchebn. Zaved., Khim. Khim. Tekhnol.* **23** (1980) no. 10, 1269.
310. V. A. Moroz, V. L. Tikhonov, *Lakrokas. Mater. Ikh. Primen.* **5** (1981) 17.
311. B. G. Aristov et al., *Lakrokas. Mater. Ikh. Primen.* **4** (1974) 10.
312. ICI, GB 461 268, 1937 (R. P. Linstead, C. E. Dent).
313. IG Farbenind., DE 717 164, 1935 (H. Tietjens).
314. Du Pont, US 2 247 752, 1937 (A. L. Fox).
315. E. A. Wich, *Am. Ink Maker* **37** (1959) 26.
316. BASF, DE 2 504 150, 1975 (H. Geeren, W. Fabian).
317. Ciba, US 2 862 929, 1958 (A. Caliezi, W. Kern, T. Holbro).
318. Du Pont, US 2 833 784, 1958 (F. F. Ehrich).
319. GAF Corp., US 2 873 279, 1959 (D. I. Randall, J. Taras).
320. Bayer, DE 928 344, 1952 (H. Vollmann).
321. Du Pont, US 2 586 598, 1952 (G. Barnhart, R. W. Grimble).
322. IG Farbenind., FR 838 009, 1939.
323. Du Pont, US 2 933 505, 1957 (J. Jackson).
324. BASF, GB 1 073 348, 1964 (R. Polster, R. Schroedel, D. V. Pigenot).
325. Du Pont, US 3 006 921, 1957 (V. Weinmayr).
326. Imperial Smelting Corp., GB 1 037 657, 1964 (D. E. M. Wotton).
327. Mobay Corp., US 236 955, 1981; US 275 488, 1981; EP-A 58 888, 1986 (J. F. Santimauro).
328. BASF, DE 3 636 428, 1986 (J. Kranz, K. Schmeidl).
329. BASF, EP-A 0 182 207, 1986 (J. Kranz).
330. R. P. Linstead, F. T. Weiss, *J. Chem. Soc.* 1950, 2977.
331. Bayer, US 2 613 128, 1949 (F. Baumann, B. Bienert).
332. N. Fukada, *Nippon Kagaku Zasshi* **75** (1954) 1141; *Chem. Abstr.* **51** (1957) 12 729.
333. F. Fukada, *Nippon Kagaku Zasshi* **75** (1954) 586; *Chem. Abstr.* **51** (1957) 11 154.
334. F. Fukada, *Nippon Kagaku Zasshi* **76** (1955) 1378; *Chem. Abstr.* **51** (1957) 17 945.
335. F. Fukada, *Nippon Kagaku Zasshi* **77** (1956) 1421.
336. F. Fukada, *Nippon Kagaku Zasshi* **79** (1958) 396, 980; *Chem. Abstr.* **54** (1960) 4612.
337. IG Farbenind., DE 891 121, 1937 (F. Nadler, H. Hoyer, O. Bayer).
338. Ciba, BE 565 782, 1957 (E. Parette, F. Maes).
339. ICI, GB 586 340, 1944 (N. H. Haddock, C. Wood).
340. BASF, DE 843 726, 1950 (A. Tartter).
341. Berufsgenossenschaft der chemischen Industrie: *MAK-Werte 1989,* Nov. 30, 1989.
342. M. J. Camenzind, C. L. Hill, *J. Heterocycl. Chem.* **22** (1985) 2.
343. M. Hanack, J. Metz, G. Pawloski, *Chem. Ber.* **115** (1982) 2836.
344. J. Metz, O. Schneider, M. Hanack, *Inorg. Chem.* **23** (1984) 1065.
345. A. B. P. Lever et al., *J. Am. Chem. Soc.* **103** (1981) 6800.

346. S. A. Mikhalenko, V. M. Derkacheva, E. A. Luk'-yanets, *Zh. Obshch. Khim.* **51** (1981) 1650; *Chem. Abstr.* **95** (1981) 196 503 v.
347. D. Woehrle, B. Schulte, *Makromol. Chem.* **186** (1985) no. 11, 2229.
348. S. A. Mikhalenko, E. A. Luk'yanets, *Zh. Obshch. Khim.* **46** (1976) no.108, 9, 2156.
349. N. O. Sigl, *J. Heterocycl. Chem.* **18** (1981) 1613.
350. T. M. Keller, J. R. Griffith, *J. Flourine Chem.* **12** (1978) 73.
351. V. M. Derkacheva, S. S. Jodko, O. L. Kaliya, E. A. Luk'yanets, *Zh. Obshch. Khim.* **51** (1981) 2319; *Chem. Abstr.* **96** (1982) 21 207 k.
352. C. L. Leznoff, T. W. Hall, *Tetrahedron Lett.* 1982, 3023.
353. G. Pawlowski, M. Hanack, *Synthesis* 1980, 287.
354. C. Piechocki et al., *J. Am. Chem. Soc.* **104** (1982) 5245.
355. A. W. Snow, N. L. Jarvis, *J. Am. Chem. Soc.* **106** (1984) 4707.
356. A. W. Snow, J. R. Griffith, N. P. Marullo, *Macromolecules* **17** (1984) 1614.
357. G. Pawlowski, M. Hanack, *Synth. Commun.* **11** (1981) 351.
358. I. G. Oksengendler, N. V. Kondratenko, E. A. Luk'-yanets, L. M. Yagupolskii, *Zh. Org. Khim.* **14** (1978) no. 5, 1046.
359. L. I. Solovèva, S. Mikhailenko, E. V. Chernykh, E. A. Luk'yanets, *Zh. Obshch. Khim.* **52** (1982) 90; *Chem. Abstr.* **96** (1982) 219 256 d.
360. BASF, DE 3 622 363, 1986 (M. J. Degen, H. Kraus).
361. K. H. Yahya, A. A. Dawood, *J. Chem. Eng. Data.* **33** (1988) no. 4, 529.
362. H. Shirai, A. Marayuma, K. Kobayashi, N. Hoyo, *Macromol. Chem.* **181** (1980) 575.
363. S. Higaki, K. Hanabusa, H. Shirai, N. Hojo, *Macromol. Chem.* **184** (1983) 691.
364. J. H. Schutten, J. Zwart, *J. Mol. Catal.* **5** (1979) 109.
365. B. N. Achar, G. M. Fohlen, J. A. Parker, J. Keshavayya, *Indian J. Chem. Sect. A.* **27 A** (1988) no. 5, 411.
366. L. I. Solov'eva, S. A. Luk'yanets, *Zh. Obshch. Khim.* **50** (1980) no. 112, 1122.
367. H. Shirai et al., *Makromol. Chem.* **185** (1984) no. 12, 2537.
368. Nippon Teleg. & Teleph., JP 02 3788, 1985 (A. I. Korakawa).
369. D. Woehrle, G. Meyer, B. Wahl, *Macromol. Chem.* **181** (1980) 2127.
370. D. Woehrle et al., *Ber. Bunsen-Ges. Phys. Chem.* **91** (1987) no. 9, 975.
371. D. Woehrle, B. Wahl, *Tetrahedron Lett.* **3** (1979) 227.
372. V. N. Klyuev, L. S. Shiryaeva, SU 260 044, 1960.
373. S. Horiguchi, *Shikizai Kyokaishi* **38** (1965) no. 3, 99.
374. V. F. Borodkin, R. D. Komarov, SU 419 539, 1974.
375. D. Woehrle in C. C. Leznoff, A. B. P. Lever (eds.): *Phthalocyanines: Properties and Applications,* VCH Verlagsgesellschaft, Weinheim 1989, pp. 59 ff.
376. S. A. Mikhalenko, E. A. Luk'yanets, *Zh. Obshch. Khim.* **39** (1969) no. 11, 2554.
377. T. A. Shatskaya, M. G. Gal'pern, V. R. Skvarchenko, E. A. Luk'yanets, *Zh. Obshch. Khim.* **57** (1987) no. 10, 2364.
378. S. A. Mikhalenko, L. A. Yagodina, E. A. Luk'yanets, *Zh. Obshch. Khim.* **46** (1976) no. 7, 1598.
379. IG Farbenind., GB 457 526, 1936 (A. V. Aerden).
380. IG Farbenind., DE 696 590, 1940 (R. J. Recidon).
381. IG Farbenind., GB 471 418, 1937 (W. W. Groves).
382. R. P. Linstead, E. G. Noble, J. M. Wright, *J. Chem. Soc.* 1937, 911.
383. A. R. Koray, V. Ahsen, O. Bekaroglu, *J. Chem. Soc., Chem. Commun.* **12** (1986) 932.
384. R. Hendriks, Ot E. Sielcken, W. Drenth, R. J. M. Roeland, *J. Chem. Soc., Chem. Commun.* **19** (1986) 1464.
385. N. Kobayashi, Y. Nishiyama, *J. Chem. Soc., Chem. Commun.* **19** (1986) 1462.
386. A. W. Snow, N. L. Jarvis, *J. Am. Chem. Soc.* **106** (1984) 4706.
387. S. A. Mikhalenko, L. A. Yagodina, E. A. Luk'yanets, *Zh. Obshch. Khim.* **46** (1976) no. 7, 1598.
388. V. N. Kopranenkov, E. A. Luk'yanets, *Zh. Obshch. Khim.* **41** (1971) no. 10, 2341.
389. M. Whalley, P. M. Brown, D. B. Spiers, *J. Chem. Soc.* 1957, 2882.
390. M. E. Baguley, H. France, R. P. Linstead, M. Whalley, *J. Chem. Soc.* 1955, 3521.
391. ICI, GB 750 240, 1954; US 2 744 913, 1951; US 2 681 341, 1951 (W. F. Beech et al.).
392. Sumitomo Chem. Co., JA 633, 1959.
393. H. P. Kaufmann, DE 920 666, 1951.
394. IG Farbenind., DE 663 552, 1935.
395. A. H. Cook, R. P. Linstead, *J. Chem. Soc.* 1937, 929.

396. ICI, US 2 681 344, 1954; US 2 681 345, 1954 (H. France).
397. Bayer, DE 964 324, 1954; DE 1 001 785, 1954 (G. Roesch, H. Klappert, W. Wolff).
398. Hoechst, DE-AS 1 017 720, 1954 (F. Erbe).
399. American Cyanamid, US 2 850 505, 1954 (D. W. Hein).
400. H. P. Kaufmann, DE-AS 1 034 794, 1954.
401. V. N. Klyuev, F. P. Snegireva, *Izv. Vyssh. Ucheb. Zaved. Khim. Khim. Tekhnol.* **14** (1971) no. 2, 258.
402. V. N. Klyuev, F. P. Snegireva, SU 196 856, 1967.
403. V. N. Klyuev, F. P. Snegireva, SU 293 024, 1971.
404. P. G. Webb, *Am. Ink Maker* **65** (1987) no. 2, 11.
405. R. E. Gosselin, H. C. Hodge, R. P. Smith, M. N. Gleason: *Chemical Toxicology of Commercial Products,* 4th ed., Williams and Wilkins, Baltimore 1976.
406. USEPA: "Ambient Water Quality Criteria for Copper," PB 85 227 023 (1984).
407. E. Kwiatkowska, E. Zawadzka, A. Stach, *Biul. Inf. Barwniki Srodki Pomocniczc* **29** (1985) no. 1, 17.
408. J. Formanek, H. Kutna, CS 218 295, 1980.
409. R. C. Buchta et al., *J. Chromatography* **325** (1985) no. 2, 456.
410. E. C. Botti, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **305** (1950) 408.
411. M. J. H. Turk, *Verfkroniek* **32** (1959) 494.
412. A.S.T.M., *Standards,* part 28, American Society of Testing Materials, Philadelphia, Pa., 1978, p. 211.
413. R. L. P. R. Hepsworth, *Chem. Ind.* 1952, 272.
414. C. E. Moore, *Paint Manuf.* **27** (1957) 377.
415. G. G. Rao, T. P. Sastri, *Z. Anal. Chem.* 1959, 11.
416. D. Kruh, *Anal. Chem.* **42** (1970) no. 14, 1849.
417. Hoechst, BE 590 125, 1960 (J. Gevers).
418. ICI, GB 576 270, 1944; GB 587 636, 1944 (N. H. Haddock, C. Wood); GB 633 160, 1947.
419. K. H. Schuendehuette: "Chromophore Systems," in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. VI, Academic Press, New York 1974, p. 312.
420. P. Rys, H. Zollinger, *Text. Chem. Color.* **6** (1974) 62.
421. J. Eibl, *Melliand Textilber. Int.* **56** (1975) 398.
422. A. P. Hopmeier, *Am. Ink Maker* **67** (1989) no. 9, 58.
423. *Kirk-Othmer,* **17,** 787.
424. F. R. Wrenn Jr., M. L. Good, P. Handler, *Science* **113** (1951) 525.
425. E. Neuzil, J. Ballenger, *Compt. Rend. Soc. Biol.* **146** (1952) 1108.
426. *NPIRI Raw Materials Data Handbook,* **vol. 4,** 1984.
427. P. Miloy, K. Kay, *J. Toxicol. Environ. Health* **4** (1978) 31.
428. A. Haddow, E. Hornung, *J. Nat. Cancer Inst.* **24** (1960) 109.
429. *NTP Technical Bulletin* **1** (1981) no. 5, 8.
430. American Conference of Governmental Industrial Hygienists: *Documentation of the Threshold Limit Values and Biological Exposure Indices,* 5th ed., Cincinnati 1986.
431. B. A. Kurlyandski et al., *Gig. Sanit.* **1** (1975) 92.
432. E. Ben-Hur, A. Newman, S. W. Crane, I. Rosenthal, *Cancer Lett.* **38** (1987) no. 1/2, 215.
433. S. Sandor, O. Prelipceanu, I. Checiu, *Rev. Roum. Morphol. Embryol. Physiol. Morphol. Embryol.* **31** (1985) no. 3, 173.

Physical Properties, Estimation of → Estimation of Physical Properties
Pickling → Metals, Surface Treatment
Piezoelectric Materials → Ceramics, Electronic

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Pigments, Inorganic

Standard Article

Hans G. Völz[1], Jürgen Kischkewitz[2], Peter Woditsch[2], Axel Westerhaus[2], Wolf-Dieter Griebler[3], Marcel De Liedekerke[4], Gunter Buxbaum[5], Helmut Printzen[2], Manfred Mansmann[2], Dieter Räde[2], Gerhard Trenczek[2], Volker Wilhelm[2], Stefanie Schwarz[6], Henning Wienand[6], Jörg Adel[6], Gerhard Adrian[7], Karl Brandt[8], William B. Cork[9], Heinrich Winkeler[10], Wielfried Mayer[10], Klaus Schneider[11], Lutz Leitner[2], Hendrik Kathrein[2], Ekkehard Schwab[6], Helmut Jakusch[6], Manfred Ohlinger[6], Ronald Veitch[6], Günter Etzrodt[12], Gerhard Pfaff[13], Klaus-Dieter Franz[13], Ralf Emmert[14], Katsuhisa Nitta[15], Robert Besold[16], Harald Gaedcke[17]

[1]Krefeld, Germany
[2]Bayer AG, Krefeld, Germany
[3]Sachtleben Chemie GmbH, Duisburg, Germany
[4]Zinkwit Nederland bv, Eijsden, The Netherlands
[5]Bayer AG, Leverkusen, Germany
[6]BASF Aktiengesellschaft, Ludwigshafen, Germany
[7]Langelsheim, Germany
[8]Dr. Hans Heubach GmbH & Co. KG, Langelsheim, Germany
[9]Reckitt's Colours Ltd., Hull, United Kingdom
[10]Degussa AG, Hanau, Germany
[11]Degussa AG, Frankfurt/Main, Germany
[12]BASF Lacke+Farben AG, Besigheim, Germany
[13]E. Merck KGaA, Darmstadt, Germany
[14]E. M. Industries, Hawthorne, New York, United States
[15]Merck Japan, 970/04 Onahama, Japan
[16]Eckart-Werke, Fürth, Germany
[17]BASF Lacke+Farben AG, Stuttgart, Germany

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a20_243.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. General Aspects
1.2. General Chemical and Physical Properties
1.2.1. Fundamental Aspects
1.2.2. Methods of Determination
1.3. Color Properties
1.3.1. Fundamental Aspects
1.3.2. Color Measurement
1.3.3. Tinting Strength, Lightening Power, and Scattering Power
1.3.4. Hiding Power and Transparency
1.4. Stability Towards Light, Weather, Heat, and Chemicals
1.4.1. Fundamental Aspects
1.4.2. Test Methods
1.4.2.1. Light Stability
1.4.2.2. Weather Resistance
1.4.2.3. Heat Stability
1.4.2.4. Fastness to Chemicals
1.5. Behavior of Pigments in Binders
1.5.1. Fundamental Aspects
1.5.2. Test Methods
1.5.2.1. Pigment - Binder Interaction
1.5.2.2. Dispersing Behavior in Paint Systems
1.5.2.3. Miscellaneous Pigment - Binder Systems
2. White Pigments
2.1. Titanium Dioxide
2.1.1. Properties
2.1.2. Raw Materials
2.1.2.1. Natural Raw Materials
2.1.2.2. Synthetic Raw Materials
2.1.3. Production
2.1.3.1. Sulfate Method
2.1.3.2. The Chloride Process
2.1.3.3. Pigment Quality
2.1.3.4. Aftertreatment
2.1.3.5. Problems with Aqueous and Gaseous Waste
2.1.4. Economic Aspects
2.1.5. Pigment Properties
2.1.6. Analysis
2.1.7. Uses of Pigmentary TiO_2
2.1.8. Uses of Nonpigmentary TiO_2
2.1.9. Toxicology
2.2. Zinc Sulfide Pigments
2.2.1. Properties
2.2.2. Production
2.2.3. Commercial Products

2.2.4. Uses
2.2.5. Economic Aspects
2.2.6. Toxicology
2.3. Zinc Oxide (Zinc White)
2.3.1. Introduction
2.3.2. Properties
2.3.3. Production
2.3.4. Quality Specifications
2.3.5. Uses
2.3.6. Economic Aspects
2.3.7. Toxicology and Occupational Health
3. Colored Pigments
3.1. Oxides and Hydroxides
3.1.1. Iron Oxide Pigments
3.1.1.1. Natural Iron Oxide Pigments
3.1.1.2. Synthetic Iron Oxide Pigments
3.1.1.3. Toxicology and Environmental Aspects
3.1.1.4. Quality
3.1.1.5. Uses
3.1.1.6. Economic Aspects
3.1.2. Chromium Oxide Pigments
3.1.2.1. Properties
3.1.2.2. Production
3.1.2.3. Quality Specifications and Analysis
3.1.2.4. Storage and Transportation
3.1.2.5. Uses
3.1.2.6. Economic Aspects
3.1.2.7. Toxicology and Occupational Health
3.1.3. Mixed Metal Oxide Pigments
3.1.3.1. Properties
3.1.3.2. Production
3.1.3.3. Quality Specifications and Analysis
3.1.3.4. Storage and Transportation
3.1.3.5. Legal Aspects
3.1.3.6. Uses
3.1.3.7. Economic Aspects
3.1.3.8. Toxicology and Occupational Health
3.2. Cadmium Pigments
3.2.1. Cadmium Sulfide
3.2.2. Cadmium Yellow
3.2.3. Cadmium Sulfoselenide (Cadmium Red)
3.2.4. Cadmium Mercury Sulfide (Cadmium Cinnabar)
3.2.5. Properties and Uses
3.2.6. Quality Specifications
3.2.7. Economic Aspects
3.2.8. Toxicology and Environmental Protection
3.3. Bismuth Pigments
3.3.1. Properties
3.3.2. Production
3.3.3. Uses
3.3.4. Toxicology
3.4. Chromate Pigments
3.4.1. Chrome Yellow
3.4.2. Molybdate Red and Molybdate Orange
3.4.3. Chrome Orange
3.4.4. Chrome Green and Fast Chrome Green
3.4.5. Toxicology and Occupational Health
3.5. Ultramarine Pigments
3.5.1. Chemical Structure
3.5.2. Properties
3.5.3. Production
3.5.4. Uses
3.5.5. Toxicity and Environmental Aspects
3.5.6. Economic Aspects
3.6. Iron Blue Pigments
3.6.1. Structure
3.6.2. Production
3.6.3. Properties
3.6.4. Uses
3.6.5. Toxicology and Environmental Aspects
4. Specialty Pigments
4.1. Magnetic Pigments
4.1.1. Iron Oxide Pigments
4.1.2. Cobalt-Containing Iron Oxide Pigments
4.1.3. Chromium Dioxide
4.1.4. Metallic Iron Pigments
4.1.5. Barium Ferrite Pigments
4.2. Anticorrosive Pigments
4.2.1. Principles
4.2.2. Phosphate Pigments
4.2.2.1. Zinc Phosphate
4.2.2.2. Aluminum Phosphate

4.2.2.3. Chromium Phosphate
4.2.2.4. New Pigments Based on Metal Phosphates
4.2.2.5. Multiphase Phosphate Pigments
4.2.3. Other Phosphorus-Containing Pigments
4.2.4. Borosilicate Pigments
4.2.5. Borate Pigments
4.2.6. Chromate Pigments
4.2.7. Molybdate Pigments
4.2.8. Lead and Zinc Cyanamides
4.2.9. Ion-Exchange Pigments
4.2.10. Metal Oxide Pigments
4.2.10.1. Red Lead
4.2.10.2. Calcium Plumbate
4.2.10.3. Zinc and Calcium Ferrites
4.2.10.4. Zinc Oxide
4.2.11. Powdered Metal Pigments
4.2.11.1. Zinc Dust
4.2.11.2. Lead Powder
4.2.12. Flake Pigments
4.2.13. Organic Pigments
4.2.14. Toxicology
4.3. Luster Pigments
4.3.1. Nacreous and Interference Pigments
4.3.1.1. Optical Principles
4.3.1.2. Natural Pearl Essence
4.3.1.3. Basic Lead Carbonate
4.3.1.4. Bismuth Oxychloride
4.3.1.5. Metal Oxide - Mica Pigments
4.3.1.6. New Developments
4.3.1.7. Uses
4.3.2. Metal Effect Pigments
4.4. Transparent Pigments
4.4.1. Transparent Iron Oxides
4.4.2. Transparent Iron Blue
4.4.3. Transparent Cobalt Blue and Green
4.4.4. Transparent Titanium Dioxide
4.4.5. Transparent Zinc Oxide

[Top of Page]

1. Introduction

Hans G. Völz

1.1. General Aspects

Definition. The word “pigment” is of Latin origin (pigmentum) and originally denoted a color in the sense of a coloring matter, but was later extended to indicate colored decoration (e.g., makeup). In the late Middle Ages, the word was also used for all kinds of plant and vegetable extracts, especially those used for coloring. The word pigment is still used in this sense in biological terminology; it is taken to mean dyestuffs of plant or animal organisms that occur as very small grains inside the cells or cell membranes, as deposits in tissues, or suspended in body fluids.

The modern meaning associated with the word pigment originated in this century. According to accepted standards (Table 1, “Coloring materials: Terms and definitions”), the word pigment means a substance consisting of small particles that is practically insoluble in the applied medium and is used on account of its coloring, protective, or magnetic properties. Both pigments and dyes are included in the general term “coloring materials”, which denotes all materials used for their coloring properties. The characteristic that distinguishes pigments from soluble organic dyes is their low solubility in solvents and binders. Pigments can be characterized by their chemical composition, and by their optical or technical properties. In this introductory chapter, only inorganic pigments used as coloring materials are discussed.

Table 1. Listing of standards for pigments

Keywords	ISO	EN	ASTM	DIN
Acidity/alkalinity	787–4		D 1208	ISO 787–4
Aluminum pigments and pastes:				
Sampling and testing			D 480	55 923
Specifications	1247		D 962	55 923
Barium chromate pigments:				
Specification	2068			
Bleeding	787–22	14463–4	D 279	EN 14463–4
Carbon black pigments (see also lampblack):				
Ash Content				53 586
Black value				55 979
Solvent-extractable material			D 305	55 968
Specification			D 561	55 968
Cadmium pigments:				
Specification	4620			
Chalking degree:				
Adhesive tape method	4628–6	ISO 4628–6		EN ISO 4628–6
KEMPF method			D 4214	53 159
Photographic method			D 659	
Velvet method	4628–7	ISO 4628–7		EN ISO 4628–7
Change in strength (see ease of dispersion and				

PVC)				
Chemical resistance	2812–1 to 5	ISO 2812–1 to 5		EN ISO 2812–1 to 5
Chlorides, water-soluble (see matter soluble)				
Chromium oxide pigments:				
Specification	4621		D 263	ISO 4261
Climates:				
Containing evaporated water	6270–2	ISO 6270–2		EN 6270–2, EN 23270
Standardized	554			50 014
Open air				50 019 -1
SO_2 atmosphere	6988	ISO 6988		EN ISO 6988
Coating materials:				
Terms and definitions	4618	ISO 4618		55 945
				EN ISO 4618
				EN 971–1
Color depth				53235–1 to 2
Color differences:				
CIELAB	7724–3		D 1729	6174
			D 2244	
CIE94			E 308	
CMC	7724–4			
DIN93				6176
Conditions/evaluation of measurements	7724–2			53 236
Significance				55 600
Color in full-shade systems				
Black pigments	787–25		D 3022	55 985–2
Colored pigments	787–25		D 3022	55 985
White pigments	787–25		D 2805	55 983
Coloration of building materials			EN 12878	EN 12878
Colorimetry	7724–1 to 3		E 259	5033–1 – 9
			E 308	6174
Coloring materials:				
Classification				55 944
Terms and definitions	4618–1	971–1		55 943
				EN ISO 4618
Corrosion testing:				
NaCl	9227		B 117	50 021, EN 9227
SO_2	6988	ISO 6988		EN ISO 6988
Density of pigments:				
Centrifuge method	787–23	ISO 787–23		EN ISO 787–23
Pyknometer method	787–10	ISO 787–10	D 153	EN ISO 787 -10
Dispersion, ease of:				
Alkyd resin and alkyd melamine system:				53 238–30
Drying by oxidation:				53 238–33
Stoving type				53 238–31
Automatic muller	8780–5	ISO 8780–5	D 387	EN ISO 8780–5
Bead mill	8780–4	ISO 8780–4		EN ISO 8780–4
Change in gloss	8781–3	ISO 8781–3		EN ISO 8781–3
Change in tinting strength	8781–1	ISO 8781–1		EN ISO 8781–1
Fineness of grind (see below)				
High speed impeller mill	8780–3	ISO 8780–3		EN ISO 8780–3
Introduction	8780–1	ISO 8780–1		EN ISO 8780–1
Oscillatory shaking machine	8780–2	ISO 8780–2		EN ISO 8780–2
Triple roll mill	8780–6	ISO 8780–6		EN ISO 8780–6
Fineness of grind	1524		D 1210	EN 21524
	8781–2	ISO 8781–2		EN ISO 8781–2
Gloss, measurement	2813	ISO 2813	D 523	67 530
			D 5307	EN ISO 2813
Heat stability (see also PVC)	787–21		D 2485	53 774–5
Hiding power:				
Color difference method				
Contrast ratio	6504–3			ISO 6504–3
General introduction				
Pigmented media	6504–1	ISO 6504–1	D 2805	55 987, EN 6504–1
Wedge-shaped layer				55 601
White and light gray media			D 2805	55 984
Hue of near white specimens				55 980
Hue relative of near white specimens	7724–5			55 981
Iron blue pigments:				
Methods of analysis	2495		D 1135	

Specification	2495		D 261	55 906
Iron, manganese oxide pigments:				
Methods of analysis	1248		D 50	ISO 1248
Natural, specification	1248		D 3722	ISO 1248
Siena, specification	1248		D 765	ISO 1248
Umber, specification	1248		D 763	ISO 1248
Iron oxide pigments:				
Black, specification	1248		D 769	ISO 1248
Brown, specification	1248		D 3722	ISO 1248
FeO content			D 3724 D 3872	
Methods of analysis	1248		D 50	55 913–2
Red, specification	1248		D 3721	ISO 1248 55 913–1
Yellow, specification	1248		D 768	ISO 1248 ISO 1248
Lampblack pigments:				
Specification			D 209	55 968
Lead chromate pigments:				
Method of analysis	3711		D 126	ISO 3711
Specification	3711		D 211	ISO 3711
Lead chromate/phthalocyanine blue pigments:				
Methods of analysis			D 126	
Specification			D 212	
Lead chromate green pigments:				
Methods of analysis	3710		D 126	
Specification	3710		D 212	
Lead red (see red lead)				
Lead silicochromate pigments (basic):				
Methods of analysis			D 1844	
Specification			D 1648	
Lead white (see white lead)				
Light stability (see also resistance to light):				
Short test	11341, 11507		ISO 11341, ISO 11507	EN ISO 11341, EN ISO 11507
Lightening power of white pigments	787–24	ISO 787–24	D 2745	55 982
Lightness:				
White pigment powders				53 163
Lithopone pigments:				
Specification	473		D 3280	55 910
Matter soluble in HCl:				
Content of As, Ba, Cd, Co, Cr, Cr(VI), Cu,	3856–1		D 3718	53 770–1
Hg, Mn, Ni, Pb, Sb, Se, Zn	to 3856–7		D 3618	to 53 770–15
			D 3624 D 3717	
Preparation of extract	6713			53 770–1
Matter soluble in water:				
Chlorides	787–13		ISO 787–13	EN ISO 787–13
Cold extraction	787–8	ISO 787–8	D 2448	EN ISO 787–8
Cr(VI) content				53 780
Hot extraction	787–3	ISO 787–3	D 2448	EN ISO 787–3
Nitrates				
Nessler reagent	787–13		ISO 787–13	EN ISO 787–13
Salicyclic acid method	787–19	ISO 787–19		ISO 787–19
Sulfates	787–13		ISO 787–13	EN ISO 787–13
Matter volatile	787–2	ISO 787–2	D 280	EN ISO 787–2
Molybdenum orange pigments:				
Methods of analysis	3711		D 2218	ISO 3711
Nitrates, water soluble (see matter soluble)				
Oil absorption	787–5	ISO 787–5	D 281 D 1483	EN ISO 787–5
Opacity: paper, cardboard	2471			53 146
Particle size analysis:				
Basic terms:	13320–1	13320–1		
		3406–1 to 9		
Electrical sensing zone	13319	13319		
Image analysis method	13322			
Photon correlation spectroscopy	13321	ISO 13321		
Representation:			D 1366	53 206–1
	9276–1			ISO 9276–1

Basic terms				
Logarithmic normal diagram	9276–2, 5			ISO 9276–2, -5
Power function grid	9276–4			ISO 9276–4
RRSB grid				66 145
Sedimentation method:				
Balance method				66 116–1
Basic standards			D 3360	66 111
Centrifugal liquid	13318–1 to 3			
Gravitational liquid	13317–1 to 3			
Pipette method				66 115
pH value	787–9	ISO 787–9	D 1208	EN ISO 787–9
Phthalocyanine pigments				
Methods of analysis			D 3256	
PVC, nonplasticized:				
Basic mixture				53 774–1
Heat stability				53 774–5
Test specimen preparation				53 774–2
PVC, plasticized:				
Basic mixture		EN 14469–1		EN 14469–1
Bleeding		EN 14469–4		EN 14469–4
Change in strength		EN 13900–2		EN 13900–2
Heat stability, in oven		EN 12877–1, -3		EN 12877–1, -3
Heat stability, mill ageing		EN 12877–1, -4		EN 12877–1, -3
Test specimen preparation		EN 14469–4		EN 12877–4
Red lead:				
Specifications	510		D 49	55 916
			D 83	
Reflectance factor; paper, cardboard:				
Fluorescent				53 145–2
Nonfluorescent	2469			53 145–1
Reflectometer (gloss assessment)	2813		D 5307	67 530
			D 523	EN ISO 2813
Residue on sieve:				
By water	787–7			53 195
Mechanical method	787–18	ISO 787–18		EN ISO 787–18
Resistance to light	787–15	ISO 787–15		EN ISO 787–15
Resistivity, aqueous extract	787–14	ISO 787–14	D 2448	ISO 787–14
Residual color difference (see Tinting strength)				
Sampling:	15528		D 3925	53 242–1
				EN ISO 15528
Scattering power, relative:				
Grey paste method	787–24	ISO 787–24		EN ISO 787–24
Black ground method				53 164
Specific surface area:				
BET method	9227			ISO 9227
N_2 adsorption				66 132
Permeability techniques				66 126–1
Spittle/sweat resistance				53 160
Standard depth of shade:				
Specimen adjustment				53 235–2
Standards				53 235–2
Strontium chromate pigments:				
Specification	2040		D 1845	55 903
Sulfates, water-soluble (see matter soluble)				
SO_2 resistance	3231		ISO 3231	53 771
				EN ISO 3231
Tamped volume	787–11	ISO 787–11		EN ISO 787–11
Test evaluation:				
Scheme	4628–1		ISO 4628–112877–1 to 4	53 230
Thermoplastics:				EN 12877–1 to -4
Basic mixtures				53 773–1
Heat stability			12877–1	EN 12877–1
Test specimen preparation				53 773–2
Tinting strength, relative:				
Change in	8781–1	ISO 8781–1		EN ISO 8781–1
FIAF method (see Color depth)				
Inorganic pigments				55 603
Lightness method				55 603

Organic pigments (see Color depth)				
Photometric methods			D 387	55 978
	787–24	ISO 787–24	D 3022 D 4838	EN ISO 787–24
Visual	787–16	ISO 787–1		ISO 787–16 1349
Weighted *K/S* value method				
Titanium dioxide pigments:				
Method of analysis	591–1	ISO 591–1	D 1394 D 3720 D 3946	EN ISO 591–1
Specification	591		D 476	EN ISO 591–1
Test methods	591		D 4563 D 4764 D 4797	EN ISO 591–1
Transparency:				
Paper, cardboard	2469			53 147
Pigmented/unpigmented systems				55 988
Ultramarine pigments:				
Methods of analysis			D 1135	
Specification	788		D 262	55 907
Viscosity	2884–1, -2		D 2196	2884–1, -2
Weathering in apparatus	4892–1 to -3	ISO 4892–1 to -3		EN ISO 4892–1 to -3
	11341	ISO 11341		EN ISO 11341 EN ISO 11507
White lead:				
Methods of analysis			D 1301	
Specification			D 81	
Zinc chromate pigments:				
Specification	1249			55 902
Zinc dust pigments:				
Methods of analysis	713		714	55969 55969
	3549	ISO 3549	D 521	EN ISO 3549
Specification	3549		D 520	55 969
Zinc oxide pigments:				
Methods of analysis			D 3280	55 908
Specification			D 79	
Zinc phosphate pigments:				
Methods of analysis	6745			ISO 6745
Specification	6745			ISO 6745

Extenders (fillers) are substances in powder form that are practically insoluble in the medium in which they are applied. They are usually white or slightly colored, and are used on account of their physical or chemical properties. The distinction between an extender and a pigment lies in the purpose for which it is used. An extender is not a colorant, it is employed to modify the properties or increase the bulk (volume) of a given material. For further details, see ⟶ Paints and Coatings; ⟶ Paper and Pulp – Fillers; and ⟶ Plastics, Additives – Fillers and Coupling Agents.

History. Natural inorganic pigments have been known since prehistoric times. Over 60 000 years ago, natural ocher was used in the Ice Age as a coloring material. The cave paintings of the Pleistocene peoples of southern France, northern Spain, and northern Africa were made with charcoal, ocher, manganese brown, and clays, and must have been produced over 30 000 years ago. About 2000 B.C., natural ocher was burnt, sometimes in mixtures with manganese ores, to produce red, violet, and black pigments for pottery. Arsenic sulfide and Naples yellow (a lead antimonate) were the first clear yellow pigments. Ultramarine (lapis lazuli) and artificial lapis lazuli (Egyptian blue and cobalt aluminum spinel) were the first blue pigments. Terra verte, malachite, and a synthetically prepared copper hydroxychloride were the first green pigments. Colored glazes for bricks (i.e., ceramic pigments) were widely used by the Chaldeans. Calcite, some phases of calcium sulfate, and kaolinite were the white pigments used at that time.

Painting, enamel, glass, and dyeing techniques reached an advanced state of development in Egypt and Babylon. A synthetic lapis lazuli (a silicate of copper and calcium) is still known as Egyptian blue. Antimony sulfide and galena (lead sulfide) were commonly used as black pigments, cinnabar as a red pigment, and ground cobalt glass and cobalt aluminum oxide as blue pigments. According to PLUTARCH, the Greeks and Romans did not regard the art of dyeing very highly, and made very little contribution to the development of new pigments. PLINY (23 – 79 A.D.) describes the pigments orpigment, realgar, massicot, red lead, white lead, verdigris, and pigments laked with alum, as well as the pigments already listed above. Certain types of chalk and clay were used as white pigments.

From the age of the migration of the peoples (fourth to sixth century A.D.) to the end of the late Middle Ages, there were no notable additions to the range of coloring materials. The reinvented pigment Naples yellow and certain dyestuffs for textiles from the orient were the only innovations. New developments in the field of pigments first occurred during the early Renaissance. Carmine was introduced from Mexico by the Spanish. Smalt, safflore, and cobalt-containing blue glasses were developed in Europe (⟶ Cobalt and Cobalt Compounds – Introduction).

The pigment industry started in the 18th century with products such as Berlin blue (1704), cobalt blue (1777), Scheele's green, and chrome yellow (1778).

In the 19th century, ultramarine, Guignet's green, cobalt pigments, iron oxide pigments, and cadmium pigments were developed in quick succession.

In the 20th century, pigments increasingly became a subject of scientific investigation. In the past few decades, the synthetic colored pigments cadmium red, manganese blue, molybdenum red, and mixed oxides with bismuth came onto the market. Titanium dioxide with anatase or rutile structures, and acicular zinc oxide were introduced as new synthetic white pigments and extenders, respectively. Luster pigments (metal effect, nacreous, and interference pigments) have assumed increasing importance.

Economic Aspects. World production of pigments in 2005 was ca. 7.4×10^6 t. Inorganic pigments accounted for ca. 97 % of this.

About one-third of this total is supplied by the United States, one-third by the European Community, and one-third by all the remaining countries. The German owned pigment industry supplied more than 40 % of the world consumption of inorganic colored pigments, including about 50 % of the iron oxides. Estimated world consumption of inorganic pigments in 2005 can be broken down as follows:

Titanium dioxide	68 %
Synthetic iron oxides	10 %
Carbon black pigments	12 %
Lithopone	4 %
Chromium oxide	1 %
Luster pigments	< 1 %
Mixed metal oxide pigments	< 0.5 %
Ultramarine	< 0.5 %

Pigment production is still a growth industry, but in the future the growth rate will decrease. The names of manufacturing companies are given in the corresponding.

Uses. The most important areas of use of pigments are paints, varnishes, plastics, artists' colors, printing inks for paper and textiles, leather decoration, building materials (cement, renderings, concrete bricks and tiles — mostly based on iron oxide and chromium oxide pigments), leather imitates, floor coverings, rubber, paper, cosmetics, ceramic glazes, and enamels.

The paint industry uses high-quality pigments almost exclusively. An optimal, uniform particle size is important because it influences gloss, hiding power, tinting strength, and lightening power. Paint films must not be too thick, therefore pigments with good tinting strength and hiding power combined with optimum dispersing properties are needed.

White pigments are used for white coloring and covering, but also for reducing (lightening) colored and black pigments. They must have a minimal intrinsic color tone.

When choosing a pigment for a particular application, several points normally have to be considered. The coloring properties (e.g., color, tinting strength or lightening power, hiding power, see Section Color Properties) are important in determining application efficiency and hence economics. The following properties are also important:

1. *General chemical and physical properties:* chemical composition, moisture and salt content, content of water-soluble and acid-soluble matter, particle size, density, and hardness (see Section General Chemical and Physical Properties)
2. *Stability properties:* resistance toward light, weather, heat, and chemicals, anticorrosive properties, retention of gloss (see Section Stability Towards Light, Weather, Heat, and Chemicals)
3. *Behavior in binders:* interaction with the binder properties, dispersibility, special properties in certain binders, compatibility, and solidifying effect (see Section Behavior of Pigments in Binders)

Important pigment properties and the methods for determining them are described later.

Classification. Inorganic pigments can be classified from various points of view. The classification given in Table 2 (for standards see Table 1, "Coloring materials, terms") follows a system recommended by ISO and DIN; it is based on coloristic and chemical considerations. As in many classification schemes, there are areas of overlap between groups so that sharp boundaries are often impossible. In this article white pigments are described in Chapter White Pigments, colored pigments in Chapter Colored Pigments, and specialty pigments in Chapter Specialty Pigments. For black pigments, see ⟶ Carbon – Carbon Black.

Table 2. Classification of inorganic pigments

Term	Definition
White pigments	the optical effect is caused by nonselective light scattering (examples: titanium dioxide and zinc sulfide pigments, lithopone, zinc white)
Colored pigments	the optical effect is caused by selective light absorption and also to a large extent by selective light scattering (examples: iron oxide red and yellow, cadmium pigments, ultramarine pigments, chrome yellow, cobalt blue)
Black pigments	the optical effect is caused by nonselective light absorption (examples: carbon black pigment, iron oxide black)
Luster pigments	the optical effect is caused by regular reflection or interference
Metal effect pigments	regular reflection takes place on mainly flat and parallel metallic pigment particles (example: aluminum flakes)
Nacreous pigments	regular reflection takes place on highly refractive parallel pigment platelets (example: titanium dioxide on mica)
Interference pigments	the optical effect of colored luster pigments is caused wholly or mainly by the phenomenon of interference (example: iron oxide on mica)
Luminescent pigments	the optical effect is caused by the capacity to absorb radiation and to emit it as light of a longer wavelength
Fluorescent pigments	the light of longer wavelength is emitted after excitation without a delay (example: silver-doped zinc sulfide)
Phosphorescent pigments	the light of longer wavelength is emitted within several hours after excitation (example: copper-doped zinc sulfide)

1.2. General Chemical and Physical Properties

1.2.1. *Fundamental Aspects [2]*

Chemical Composition. With few exceptions, inorganic pigments are oxides, sulfides, oxide hydroxides, silicates, sulfates, or carbonates (see Tables 3 and 4), and normally consist of single-component particles (e.g., red iron oxide, α-Fe_2O_3) with well-defined crystal structures. However, mixed and substrate pigments consist of nonuniform or multicomponent particles.

Table 3. Classification of white and black pigments

Chemical class	White pigments	Black pigments
Oxides	titanium dioxide	iron oxide black
	zinc white, zinc oxide	iron – manganese black
		spinel black
Sulfides	zinc sulfide	
	lithopone	
Carbon and carbonates	white lead	carbon black

Table 4. Classification of inorganic colored pigments

Chemical class	Green	Blue-green	Blue	Violet	Red	Orange	Yellow	Brown
Oxides and oxide – hydroxides								
Iron oxide pigments					iron oxide red	iron oxide orange	iron oxide yellow	iron oxide brown
Chromium	chromium	chromium						

oxide pigments	oxide	oxide hydrate green					
Mixed metal oxide pigments		cobaltgreen and blue			chromium rutile orange	nickel rutile yellow, chromium rutile yellow	zinc iron spinel, Mn – Fe brown
Sulfide and sulfoselenide pigments				cadmium sulfoselenide		cadmium sulfide, (Cd, Zn)S	
Chromate pigments	chrome green			molybdate red	chrome orange	chrome yellow, zinc yellow	
Ultramarine pigments		ultramarine green, blue, violet, and red					
Iron blue pigments		iron blue					
Others		manganese blue	cobalt manganese violet			Naples yellow, bismuth vanadate	

Mixed pigments are pigments that have been mixed or ground with pigments or extenders in the dry state (e.g., chrome green pigments are mixtures of chrome yellow and iron blue). If the components differ in particle size and shape, density, reactivity, or surface tension, they may segregate during use.

In the case of *substrate pigments*, at least one additional component (pigment or extender) is deposited onto a substrate (pigment or extender), preferably by a wet method. Weak, medium, or strong attractive forces develop between these pigment components during drying or calcining. These forces prevent segregation of the components during use.

Special substrate pigments include the aftertreated pigments and the core pigments. To produce *aftertreated pigments* the inorganic pigment particles are covered with a thin film of inorganic or organic substances to suppress undesirable properties (e.g., catalytic or photochemical reactivity) or to improve the dispersibility of the pigments and the hydrophilic or hydrophobic character of their surfaces. The particles can be coated by precipitation (e.g., aftertreated TiO_2 pigments, see Section Aftertreatment), by adsorption of suitable substances from solutions (usually aqueous), or by steam hydrolysis.

To produce *core pigments*, a pigment substance is deposited on an extender by precipitation or by wet mixing of the components. In the case of anticorrosive pigments (see Section Principles), whose protective effect is located on their surfaces, the use of core pigments can bring about a significant saving of expensive material. Extender particles are also treated by fixing water-insoluble organic dyes on their surfaces via lake formation.

Analysis. The industrial synthesis of inorganic pigments is strictly controlled by qualitative and quantitative chemical analysis in modern, well-equipped physico-chemical test laboratories. Quantitative chemical and X-ray analysis is carried out on raw materials, intermediates, and substances used for aftertreatment, but most importantly on the final products, byproducts, and waste products (wastewater and exhaust gas). This serves not only to fulfil quality requirements but also the demands of environmental protection. Quality control, carried out in specially equipped laboratories and supported by computers, includes testing of physical and technical application properties [3], [4]. Information on quality requirements for inorganic pigments is widely available in international (ISO), European (EN), and national standards (e.g., AFNOR, ASTM, BSI, DIN) [5], [6]. Standard analytical methods and conditions of delivery for the most important inorganic pigments are given in Table 1. Further information is given in later sections.

Crystallography and Spectra. The following are the most common crystal classes:

1. *Cubic*: zinc blende lattice (e.g., precipitated CdS), spinel lattice (e.g., Fe_3O_4, $CoAl_2O_4$)
2. *Tetragonal*: rutile lattice (e.g., TiO_2, SnO_2)
3. *Rhombic*: goethite lattice (e.g., α-FeOOH)
4. *Hexagonal*: corundum lattice (e.g., α-Fe_2O_3, α-Cr_2O_3)
5. *Monoclinic*: monazite lattice (e.g., $PbCrO_4$)

In ideal solid ionic compounds, the absorption spectrum is composed of the spectra of the individual ions, as is the case in ionic solutions. For metal ions with filled *s, p,* or *d* orbitals, the first excited energy level is so high that only ultraviolet light can be absorbed. Thus, when the ligands are oxygen or fluorine, white inorganic compounds result. The absorption spectra of the chalcogenides of transition elements with incompletely filled *d* and *f* orbitals are mainly determined by the charge-transfer spectrum of the chalcogenide ion which has a noble gas structure. For the transition metals, lanthanides, and actinides, the energy difference between the ground state and the first excited state is so small that wavelength-dependent excitations take place on absorption of visible light, leading to colored compounds [7].

X-Ray investigation of inorganic pigments yields information on the structure, fine structure, state of stress, and lattice defects of the smallest coherent regions that are capable of existence (i.e., crystallites) and on their size. This information cannot be obtained in any other way. Crystallite size need not be identical with particle size as measured by the electron microscope, and can, for example, be closely related to the magnetic properties of the pigment.

Particle Size. The important physical data for inorganic pigments comprise not only optical constants, but also geometric data: mean particle size, particle size distribution, and particle shape [8]. The standards used for the terms that are used in this section are listed in Table 1 ("Particle size analysis").

The concept of *particles and particle shape* corresponds to that used in the recommended and internationally accepted classification of pigment particles given in [9] (see Fig. 1 and Table 5).

Table 5. Definitions of particles and associated terms (see also Fig. 1)

Term	Definition
Particle	individual unit of a pigment that can have any shape or structure
Primary or individual particles	particles recognizable as such by appropriate physical methods (e.g., by optical or electron microscopy)
Aggregate	assembly of primary particles that have grown together and are aligned side by side; the total surface area is less than the sum of the surface areas of the primary particles
Agglomerate	assembly of primary particles (e.g., joined together at the corners and edges), and/or aggregates whose total surface area does not differ appreciably from the sum of the individual surface areas
Flocculate	agglomerate present in a suspension (e.g., in pigment–binder systems), which can be disintegrated by low shear forces

Figure 1. Primary particles, agglomerates, and aggregates

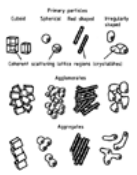

The term *particle size* must be used with care, as is borne out by the large number of different "particle diameters" and other possible terms used to denote size (see Table 6). In granulometry, so-called *shape factors* are often used to convert equivalent diameters to "true" diameters. However, the determination and use of shape factors is problematic.

Table 6. Particle size, particle size distribution, and characteristic quantities

Term	Definition
Particle size	geometrical value characterizing the spatial state of a particle
Particle diameter D_{eff}	diameter of a spherical particle or characteristic dimension of a regularly shaped particle
Equivalent diameter D	diameter of a particle that is considered as a sphere
Particle surface area S_T	surface area of a particle: a distinction is made between the internal and external surface areas
Particle volume V_T	volume of a particle: a distinction is made between effective volume (excluding cavities) and apparent volume (including cavities)
Particle mass m_T	mass of a particle
Particle density Q_T	density of a particle
Particle size distribution	statistical representation of the particle size of a particulate material
Distribution density	gives the relative amount of a particulate material in relation to a given particle size diameter. Density distribution functions must always be normalized
Cumulative distribution	normalized sum of particles that have a diameter less than a given particle size parameter
Fractions and class	a fraction is a group of particles that lies between two set values of the chosen particle size parameter that limits the class
Mean value and other similar parameters	the mean values of particle size parameters can be expressed in many ways, some values are used frequently in practice
Distribution spread	parameter for characterizing the nonuniformity of the particle size

In practice, empirically determined *particle size distributions* are represented by:

1. Tabulated results
2. Graphical representation in the form of a histogram (bar chart) or as a continuous curve
3. Approximation in the form of analytical functions

For standards see Table 1 "Particle size, representation".

Special distribution functions are specified in some standards (e.g., power distribution, logarithmic normal distribution, and RRSB distribution). The representation of particle size distributions and methods for their determination are described elsewhere (⟶ Particle Size Analysis and Characterization of a Classification Process). Methods of determination for pigments are rated in Section Methods of Determination.

The important parameters relating to particle size distribution are the *mean particle size* and the *spread of the distribution.* The way of expressing the mean particle size depends on the test method used or on which mean value best reflects the pigment property of interest. Depending on the spread of parameter, the various mean values for a given particulate material can differ considerably. The mean particle sizes of inorganic pigments lie in the range 0.01 – 10 µm, and are usually between 0.1 and 1 µm.

The *specific surface area* also represents a mean of the pigment particle size distribution. It can be used to calculate the mean diameter of the surface distribution. Care must be taken that the effect of the "internal surface area" is taken into account. If the product has an internal surface area which cannot be neglected in comparison to the external surface area, then the measured specific surface area no longer gives a true measure of the mean diameter. This applies, for instance, to aftertreated pigments because the treatment material is often very porous.

For anisometric particles (e.g., needle- or platelet-shaped particles) mathematical statistics may likewise be applied [10]. The two-dimensional logarithmic normal distribution of the length L and breadth B of the particles also allows the representation and calculation of the characteristic parameters and mean values. The eccentricity of the calculated standard deviation ellipse (Fig. 2) is a measure of the correlation between the length and breadth of the particle. By using more than two particle fineness parameters, this principle can be further extended in a similar manner.

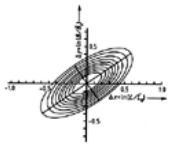

Figure 2. Standard deviation ellipses of a logarithmic normal distribution (yellow iron oxide pigment)

$\bar{L}_z$, $\bar{B}_z$ = median of L, B

The effect of particle size on optical properties of pigments is described in Section Color Properties.

1.2.2. *Methods of Determination*

General Methods. *Sampling* with a suitable device (beaker, sampling scoop) is applicable for solid materials, especially pigments, fillers, and resins in powder, granular, or lump form. For standards, see Table 1 ("Sampling").

Standard climates are constant artificial climatic conditions with a defined temperature and humidity and limited ranges of air pressure and flow velocity. They can be set up in "clima" cabinets, closets, or rooms. The standard climate 23/50 (23 °C and 50 % relative humidity) is recommended. For standards, see Table 1 ("Climates").

Evaluation Methods. In testing paint materials, paints, and other coating materials, the properties or variations in properties often cannot be described quantitatively but must be assessed subjectively. A uniform system of evaluation has been established in the form of a numerical scale to facilitate the assessment of results and mutual understanding. This system should only be used if a result cannot be obtained as a directly measured value. For standards, see Table 1 ("Test evaluation").

Matter Volatile and Loss on Ignition. The content of *matter volatile* in a pigment is determined by drying a sample in an oven at 105 ± 2 °C. This normally gives a measure of the moisture content. For standards, see Table 1 ("Matter volatile").

Loss on ignition is determined by various means depending on the pigment. The principle is the same in all the methods: a weighed sample is heated to a predetermined temperature, cooled in a desiccator, and reweighed. Special standard procedures are specified for iron oxide pigments and carbon black. For standards, see Table 1 ("Iron oxide pigments: Specifications" and "Carbon black pigments: Specifications").

Aqueous Extracts. The content of *matter soluble in water* in pigments is determined by hot or cold extraction of the pigments under prescribed testing conditions. For

standards, see Table 1 ("Matter soluble in water"). The decision whether to use hot or cold extraction depends on the properties of the pigment and should be agreed between the interested parties unless otherwise prescribed.

The *pH value* of a pigment denotes the pH of an aqueous suspension of the pigment prepared in a prescribed manner. For standards, see Table 1 ("pH value"). Apparatus: pH meter.

The *electrical conductivity* (or resistivity) of an aqueous pigment extract is determined from the electrical conductance (or resistance). For standards, see Table 1 ("Resistivity"). Apparatus: centrifuge with glass containers, resistance measurement bridge, electrolytic cell.

The *acidity* or *alkalinity* is measured as the quantity of cubic centimeters of 0.1 N alkali (NaOH or KOH) or acid solution that are required to neutralize an aqueous extract of 100 g pigment under prescribed conditions. Unless otherwise agreed, the pigment is extracted with hot water. If cold water is used, this must be specifically stated. For standards, see Table 1 ("Acidity/alkalinity"). Test reagents: 0.5 N HCl or H_2SO_4, 0.5 N NaOH or KOH, indicators (by agreement).

Water-Soluble Sulfates, Chlorides and Nitrates. The procedures are applicable to pigments and extenders. The choice of hot or cold extraction depends on the properties of the pigment and should be agreed. The anions are determined by the usual analytical methods. For standards, see Table 1 ("Sulfates", "Chlorides", and "Nitrates"). Apparatus: Nessler tubes or spectrophotometer, equipment for pH measurement.

Water-Soluble Chromium(VI). Chromium(VI) is determined photometrically in an aqueous extract obtained by hot or cold extraction using the color produced with diphenylcarbazide, or directly as chromate. For standards, see Table 1 ("Matter soluble in water"). Apparatus: equipment for shaking conical flasks, photometer with glass cells, pH meter.

Matter Soluble in HCl. The pigment is extracted with 0.1 mol/L hydrochloric acid under prescribed conditions to determine the content of As, Ba, Cd, Co, Cr, Cr(VI), Cu, Hg, Mn, Ni, Pb, Sb, Se, and Zn. For standards, see Table 1 ("Matter soluble in HCl"). Apparatus: shaker for conical flasks.

Particle Size Distribution. Methods for determining the particle size of pigments should provide not only the mean particle size but also the complete particle size distribution. Thus, it is preferable to use methods that give not only the parameters of a particular distribution function, but also allow direct measurement of the true particle size distribution. Methods of determination are given in ⟶ Particle Size Analysis and Characterization of a Classification Process. Closely specified standard methods of determination exist for particle sizes >1 µm. Several methods have been developed for the particle size region of particular relevance to pigments (<1 µm). The following methods are used:

1. Counting particles on an electron micrograph (⟶ Particle Size Analysis and Characterization of a Classification Process – Image Analysis)
2. Sedimentation analysis (⟶ Particle Size Analysis and Characterization of a Classification Process – Sedimentation Analysis)
3. Optical methods (⟶ Particle Size Analysis and Characterization of a Classification Process – Measurement by Light Scattering or Absorption).
4. Other methods that are based on physical volume effects (e.g., the Coulter counter, based on electrical conductivity, ⟶ Particle Size Analysis and Characterization of a Classification Process – Electrical Sensing Zone Method)

Particle counting of electron micrographs is performed on a computer screen with an automatic counting apparatus. In order to minimize errors, at least 2500 (but preferably 10 000) particles must be counted.

For pigments, counting is the most suitable method for several reasons. The counting operation is carried out in the binder medium of interest, whereas with sedimentation analysis or the Coulter counter the medium cannot be freely chosen. Furthermore, counting can be carried out under the dispersion conditions used for examination (in contrast to methods in which a specified binder medium must be used that can lead to a different state of dispersion). Problems associated with concentration are much less frequent than in other methods, which sometimes require extremely dilute suspensions (especially optical analysis and the Coulter counter). Another problem with optical analysis is that calibration is often difficult because the absolute scattering and absorption functions of real particles are not well known, and because isodisperse particulate matter is usually not available to draw up a calibration curve. The most important advantages of the counting technique based on electron micrographs lie in the direct observation of the particle shapes and in the ability to distinguish between primary particles, agglomerates, and aggregates.

Sedimentation analysis uses either the earth's gravitational field or a centrifugal field. Owing to the small size of pigment particles (<1 µm), the time taken for sedimentation in the gravitational field is considerable. This has led to developments in centrifugal sedimentation techniques, both as regards apparatus (centrifuges, disk-shaped rotors) and methods (two-layer technique). For standards, see Table 1 ("Particle size analysis: Sedimentation method"). The rate of fall of a single particle in a liquid in a gravitational or centrifugal field gives a measure of the particle size. Methods of measurement differ in the approach used to determine the different particle size fractions:

1. *Suspension Methods.* Analysis starts with a suspension in which a representative sample of the dispersed solid is homogeneously distributed over the whole of the sedimentation depth.
2. *Two-Layer Method.* A suspension is spread in a thin layer on the surface of a clear, solids-free liquid. The particles then fall through the liquid in order of decreasing sedimentation velocity and reach the measuring plane in succession.

In *optical methods*, extremely dilute aqueous or air-fluidized pigment suspensions are almost exclusively used. Instruments for the 0.1 – 10 µm particle size region mostly use Fraunhofer diffraction with additional correction for Mie scattering on small particles. Optical methods allow quick and reproducible measurements. Unfortunately, information is not usually available concerning the optical constants of the sample; this causes drawbacks as regards absolute particle size data.

Sieve Analysis. The sieving residue can be determined by two methods:

1. *Wet Sieving by Hand*. In the utilization of pigments, it is important to know the content of pigment particles that are appreciably larger than the mean particle size. This material can consist of coarse impurities, pigment aggregates (agglomerates), or large primary particles. The dried pigment is washed with water through a sieve of the appropriate mesh size, and the retained material is determined gravimetrically after drying. For standards, see Table 1 ("Residue on sieve: By water").
2. *Wet Sieving by a Mechanical Flushing Procedure.* The sieve residue is the portion of coarse particles that cannot be washed through a specified test sieve with water. The result depends on the mesh size of the sieve. For standards, see Table 1 ("Residue on sieve: Mechanical method"). Apparatus: Mocker's apparatus.

The *specific surface* is usually understood to mean the area per unit mass of the solid material, but it is sometimes useful to relate the surface area to the volume of the solid (see Section Fundamental Aspects). The specific surface area can only be determined indirectly owing to the small size of the pigment particles:

1. *Gas Adsorption by the Brunauer, Emmett, and Teller (BET) Method.* The specific surface area of porous or finely divided solids is measured. The method is limited to solids that do not react with the gas used (e.g., while the gas is adsorbed), and nonmicroporous materials. For standards, see Table 1 ("Specific Surface, BET Method" and "N_2 Adsorption").
2. *Carman's Gas Permeability Method*. A gas or a wetting liquid is made to flow through the porous material in a tube by applying vacuum or pressure. The pressure drop or flow rate is measured. For pigments, a modified procedure is used in which mainly nonlaminar flow takes place [11]. For standards, see Table 1 ("Specific surface: Permeability techniques").

Pigment Density. *Density* is determined by pyknometry at a standard temperature of 25 °C. For standards, see Table 1 ("Density"). Apparatus: pyknometer, vacuum pump or centrifuge.

The *apparent density* of a powdered or granulated material after tamping is the mass (g) of 1 cm^3 of the material after tamping in a tamping volumeter under prescribed conditions. The *tamped volume* is the volume (cm^3) of 1 g of the material. Tamped volume and apparent density after tamping depend mainly on the true density, shape, and size of the particles. A knowledge of these parameters allows decisions to be made regarding dimensions of packing materials and product uniformity. For standards, see Table 1 ("Tamped volume"). Apparatus: tamping volumeter.

Hardness and Abrasiveness. The abrasiveness of a pigment is not identical to its intrinsic hardness, i.e., the hardness of its primary particles. In practice the Mohs hardness is not therefore a useful indication of the abrasiveness of a pigment. Abrasion rather depends on pigment particle size and shape and is usually caused by the sharp edges of the particles. A standard test procedure for determining abrasiveness does not exist. A method based on the abrasion of steel balls by the pigment is described in [12].

1.3. Color Properties

1.3.1. *Fundamental Aspects [13-16]*

When a photon enters a pigmented film, one of three events may occur:

1. It may be absorbed by a pigment particle
2. It may be scattered by a pigment particle
3. It may simply pass through the film (the binder being assumed to be nonabsorbent)

The important physical – optical properties of pigments are therefore their light-absorption and light-scattering properties. If absorption is very small compared with scattering, the pigment is a white pigment. If absorption is much higher than scattering over the entire visible region, the pigment is a black pigment. In a colored pigment, absorption (and usually scattering) are selective (i.e., depend on wavelength).

Pigments and coatings may be unambiguously characterized by their spectral reflectance curves $\varrho(\lambda)$ or spectral reflectance factor curves $R(\lambda)$ (Fig. 3). The reflectance spectrum $\varrho(\lambda)$ or $R(\lambda)$ and hence the color properties can be almost completely derived from physical quantities [15] (Fig. 4):

1. *Colorimetry* relates the perceived color quality to the color stimulus, which in turn is based on the reflectance spectrum $\varrho(\lambda)$.
2. The *Kubelka – Munk theory* relates $\varrho(\lambda)$ to scattering, absorption, and film thickness (scattering coefficient *S*, absorption coefficient *K*, film thickness *h*).
3. The *theory of multiple scattering* (scattering interaction) relates the scattering coefficient *S* to the pigment volume concentration σ and to the scattering diameter Q_S of the individual particle. The absorption coefficient *K* is directly proportional to the absorption diameter Q_A and the concentration σ.
4. In *Mie's theory*, the scattering diameter Q_S and the absorption diameter Q_A are related to the particle size *D*, the wavelength λ, and the optical constants of the material (refractive index n and absorption index κ).

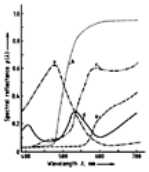

Figure 3. Spectral reflectance curves of some inorganic pigments in paints

a) Manganese blue; b) CdS; c) α-FeOOH; d) α-Cr_2O_3; e) α-Fe_2O_3

Figure 4. The relationships between the optical properties of pigments and their theoretical basis

Colorimetry [17-19]. The principles of colorimetry (⟶ Dyes, General Survey – Color Measurement and Colorant Formulation) are based on the fact that all color stimuli can be simulated by additively mixing only three selected color stimuli (trichromatic principle). A color stimulus can, however, also be produced by mixing the spectral colors. Thus, it has a spectral distribution, which in the case of nonluminous, perceived colors is called the spectral reflectance $\varrho(\lambda)$. After defining three reference stimuli, the trichromatic principle allows a three-dimensional color space to be built up in which the color coordinates (tristimulus values) can be interpreted as components of a vector (CIE system; for standards, see Table 1, "Colorimetry"; CIE = Commission Internationale de l'Éclairage). The three CIE tristimulus values depend on the spectral reflectance $\varrho(\lambda)$ and the spectrum of the iluminant $S(\lambda)$ as follows:

$$X = \int_{400}^{700} \bar{x}(\lambda)\varrho(\lambda)S(\lambda)\mathrm{d}\lambda$$

$$Y = \int_{400}^{700} \bar{y}(\lambda)\varrho(\lambda)S(\lambda)\mathrm{d}\lambda$$

$$Z = \int_{400}^{700} \bar{z}(\lambda)\varrho(\lambda)S(\lambda)\mathrm{d}\lambda$$

where $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are the CIE tristimulus values of the spectral colors and are called the CIE spectral tristimulus values (color matching function). The CIE chromaticity coordinates (*x, y*, and *z*) are given by

$$x = \frac{X}{X+Y+Z};$$

$$y = \frac{Y}{X+Y+Z};$$

$$z = 1 - x - y$$

They are represented as coordinates in a color plane. The chromaticity coordinates *x* and *y* are used to specify the saturation and hue of any color in the CIE chromaticity diagram (see ⟶ Dyes, General Survey). The CIE spectral tristimulus value $\bar{y}(\lambda)$ corresponds to the lightness sensitivity curve of the human eye. Therefore, a third color variable is specified in addition to *x* and *y*, namely the CIE tristimulus value *Y*, which is a measure of lightness.

This system allows exact measurement of color with worldwide agreement. For pigment testing, however, this is not sufficient because small color differences usually have to be determined and evaluated (e.g., between test and reference pigment). Using the CIE system, it is certainly possible to say which spectral distributions are visually identical, but this is not suitable for determining color differences. To establish color differences an "absolute color space" must be used. Here, colors are arranged three-dimensionally such that the distance between two colors in any direction in space corresponds to the perceived difference. Such a type of color space can be based on the color qualities lightness, hue, and saturation. Several such systems exist. The most widespread color system is probably the Munsell system, which is available in the form of an atlas.

For the quantitative determination of color differences, the transformation relationships between the CIE system (which has to be used for color measurement) and the physiologically equidistant color system must be established. Color differences can then be calculated in the latter system. A large number of color difference systems have been developed, mainly as a need of industrial color testing.

The Adams – Nickerson (AN) system, well-known for many decades and derived from the Munsell system, was recommended for pigment testing by DIN (German Standards Institute) and later worldwide by the CIE ("CIELAB"; for standards, see Table 1, "Color differences"). The three coordinates are denoted by *a** (red – green axis), *b** (the yellow – blue axis), and *L** (the lightness axis). To calculate the CIELAB coordinates, *X, Y*, and *Z* are first converted into the functions *X**, *Y**, and *Z** by using a relationship that approximately takes account of the physiologically equidistant lightness steps:

$$X^* = \sqrt[3]{X/X_n};$$

$$Y^* = \sqrt[3]{Y/Y_n};$$

$$Z^* = \sqrt[3]{Z/Z_n}$$

where X_n, Y_n, and Z_n are the CIE tristimulus values of the illuminant, especially a standard illuminant. For radicands < 0.008856, these equations become:

$$X^* = (7.787X/X_n) + 0.138$$

$$Y^* = (7.787Y/Y_n) + 0.138$$

$$Z^* = (7.787Z/Z_n) + 0.138$$

Values of *a**, *b**, and *L** are obtained from the values of *X **, *Y **, and *Z**:

$$a^* = 500\left(X^* - Y^*\right)$$

$$b^* = 200\left(Y^* - Z^*\right)$$

$$L^* = 116Y^* - 16$$

The components of the color difference are obtained as differences between the test sample (T) and the reference pigment (R):

$$\Delta a^* = a^*_T - a^*_R;\ \Delta b^* = b^*_T - b^*_R;\ \Delta L^* = L^*_T - L^*_R$$

The color difference is finally calculated as the geometrical distance between the two positions in the CIELAB color space:

$$\Delta E^*_{ab} = \sqrt{\Delta a^{*2} + \Delta b^{*2} + \Delta L^{*2}}$$

An important advantage of the CIELAB system is that the resulting color difference can be split up into component contributions, namely lightness, saturation, and hue, corresponding to the arrangement of the color space:

Lightness difference

$$\Delta L^* = L^*_{\mathrm{T}} - L^*_{\mathrm{R}}$$

Chroma difference (saturation difference)

$$\Delta C^*_{ab} = \sqrt{a^{*2}_{\mathrm{T}} + b^{*2}_{\mathrm{T}}} - \sqrt{a^{*2}_{\mathrm{R}} + b^{*2}_{\mathrm{R}}}$$

Hue difference

$$\Delta H^*_{ab} = \sqrt{\Delta E^{*2}_{ab} - \Delta L^{*2} - \Delta C^{*2}_{ab}}$$

The CIE has proposed a new color difference formula based on CIELAB. The new “CIE94” formula is [20] (for standards see Table 1, “Color Differences, CIE94”):

$$\Delta E^*_{94} = \sqrt{\left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2 + \left(\frac{\Delta L^*}{k_L S_L}\right)^2}$$

where $S_L = 1$, $S_C = 1 + 0.045\, C^*_{ab,\mathrm{R}}$, $S_H = 1 + 0.015\, C^*_{ab,\mathrm{R}}$, ($C^*_{ab,\mathrm{R}}$ is the chroma of the reference).

Here the k_i are the parameters of externally imposed conditions, such as conditions specific to a certain application or conditions differing from the standard conditions. The standard conditions are as follows: Illuminant D65, ca. 1000 1x, surroundings gray $L^* = 50$, ca. 100 cd/m^2, sample size over 4°, sample pair without gap between, homogeneous texture, $0 < \Delta E^*_{94} < 5$. Under these conditions, $k_L = k_C = k_H = 1$.

If neither of the two samples has been defined as the reference, C^*_{ab} should be taken equal to the geometric mean of $C^*_{ab,1}$ and $C^*_{ab,2}$.

For coating color testing, the weaknesses of the CIE system can be compensated by a suitable choice of parametric factors, for example $k_H/k_C = 1$ and $3 \le k_H/k_L \le 4$.

A number of color laboratories have recently had good experience with another color-difference system, the CMC system. (CMC: *C*olour *M*easurement *C*ommittee of the Society of Dyers and Colourists in the British Standards Institution (BSI)). For standards see “Color differences: CMC” in Table 1 and British Standard BS 6923.

The CIE has established a new color difference system: CIE DE 2000. The formulae, as taken from Technical Report CIE 142-2001, are given below:

$$\Delta E^2_{00} = \left(\frac{\Delta L'}{k_L S_L}\right)^2 + \left(\frac{\Delta C'}{k_C S_C}\right)^2 + \left(\frac{\Delta H'}{k_H S_H}\right)^2 + R_T \frac{\Delta C' \Delta H'}{k_C S_C k_H S_H}.$$

In the formula R_T is the rotation function, which is defined as

$$R_T = -\sin(2\Delta\theta) R_C \quad \text{with } \Delta\theta = 30 e^{-\left(\frac{R-275}{25}\right)^2} \quad \text{and}$$
$$R_c = 2\sqrt{\frac{\overline{C}'^7}{\overline{C}'^7 + 25^7}} \, and \, \overline{C} = \frac{1}{2}(C'_1 + C'_2).$$

The symbols used are defined as

$$L = L^*;\ a' = a^*(1+G);\ b' = b^*,$$
$$G = \frac{1}{2}\left(1 - \sqrt{\frac{C^{*7}_{ab}}{\overline{C}^7_{ab}{}^* + 25^7}}\right)$$
$$\text{with } \overline{C}^*_{ab} = \frac{1}{2}(C^*_{ab,1} + C^*_{ab,2})$$
$$C' = \sqrt{a'^2 + b'^2},$$
$$h' = \mathrm{Arctan}\frac{b'}{a'}$$

The color differences are formed according to the conventional method as the differences between the sample and the reference. The weighting functions are indicated below

$$S_L = 1 + \frac{q(\overline{L}' - 50)^2}{\sqrt{20 + (\overline{L}' - 50)^2}} \quad \text{with } q$$
$$= 0.015 \, and \, \overline{L}' = \frac{1}{2}(L'_1 + L'_2)$$
$$S_c = 1 + p\overline{C}' \text{ with } p = 0.045,$$
$$S_H = 1 + q\overline{C}' T \text{ with}$$
$$T = 1 - 0.17 \cos(\overline{h}' - 30) + 0.24 \cos(2\overline{h}')$$
$$+0.32 \cos(3\overline{h}') 0.2 \cos(4\overline{h}' - 63).$$
$$\text{and } \overline{h}' = \frac{1}{2}(h'_1 + h'_2).$$

The term containing R_T leads to a certain problem and is in contradiction to the fundamental equation for Euclidean color differences from STILES [21]:

$$ds^2 = dU_1^2 + dU_2^2 + dU_3^2 (\text{or } ds^2 = dr^2 + r^2 d\rho^2 + dU_3^2)$$

Further, as in practice, it is often necessary to break down the total color difference into its individual components ΔC, ΔH, ΔL, in order to assess its quality. In case of the first equation the last term on the right-hand side will remain, because it cannot be associated to either *C* or *H*. Therefore, the rotation must be incorporated directly into the components ΔC and ΔH.

Kubelka – Munk Theory. The Kubelka – Munk theory [22], [23] (see also ⟶ Dyes, General Survey – Colorant Formulation) is based on the fact that the optical properties of a film which absorbs and scatters light may be described by two constants: the absorption coefficient *K* and the scattering coefficient *S*. In a simplification, the flux of the diffuse incident light is represented by a single beam L^+, and the flux of the light scattered in the opposite direction by a beam L^-. Each beam is attenuated by absorption and scattering losses, but is reinforced by the scattering losses of the respectively opposite beam. The absorption and scattering losses are determined quantitatively by the two coefficients *K* and *S*. A simple system of two linked differential equations can be written. These can be integrated for the valid boundary conditions at the incident light side, and at the opposite side. Solutions for the transmittance τ and the reflectance ϱ are obtained from these integrals as a function of the absorption coefficient *K*, the scattering coefficient *S*, the film thickness *h*, and in special cases of the reflectance ϱ_o of a given substrate.

The most important and widely used quantity derived from the Kubelka – Munk theory is the reflectance of an opaque (infinitely thick) film that is described by a very simple equation:

$$K/S = (1 - \varrho_\infty)^2 / (2\varrho_\infty)$$

From this expression (Kubelka – Munk function) it follows that, within the range of validity of the theory, ϱ_∞ depends only on the ratio of the absorption coefficient to the scattering coefficient, and not on their individual values. The equation has been most useful where reflectance measurements are used to obtain information about

absorption and scattering (e.g., in textile dyeing, thin layer chromatography, and IR spectroscopy).

This theory is especially useful for computer color matching of pigmented systems [24-26]: absorption and scattering coefficients are combined additively using the specific coefficients of the components multiplied by their concentrations.

Multiple Scattering. The absorption coefficient *K* obeys Beer's law even at high pigment volume concentrations σ, and is therefore proportional to σ (Fig. 5). The relationship between the scattering coefficient *S* and the concentration gives rise to problems, however. The distance between the pigment particles decreases with increasing concentration; consequently there is interaction and hindrance between the light scattered by individual particles, and their scattering power usually falls. The scattering coefficient *S* is therefore linearly related to concentration only at low concentrations (the Beer's law region), at higher concentrations it remains below the linear value (Fig. 6). The concentration dependence of the scattering coefficient can be quantitatively represented by using empirical formulae [27], e.g., there is a linear relationship between S/σ and $\sqrt[3]{\sigma}$.

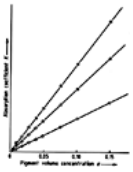

Figure 5. Absorption coefficient *K* for diffuse illumination as a function of the pigment volume concentration for three red iron oxide pigments

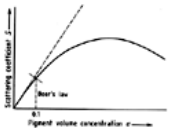

Figure 6. Scattering coefficients as a function of pigment volume concentration

Mie's Theory. MIE applied the Maxwell equations to a model in which a plane wave front meets an optically isotropic sphere with refractive index *n* and absorption index κ [28]. Integration gives the values of the absorption cross section Q_A and the scattering cross section Q_S; these dimensionless numbers relate the proportion of absorption and scattering to the geometric diameter of the particle. The theory has provided useful insights into the effect of particle size on the color properties of pigments.

Scattering is considered first. Here, the crucial parameter α in the Q_A and Q_S formulae is a relative measure of particle size because it is proportional to the particle diameter *D* and is inversely proportional to the wavelength λ. At a constant wavelength λ and for various relative refractive indices *n* (i.e., relative to the binder, $n = n_p/n_B$ where n_p and n_B are the refractive indices of the particle and binder, respectively), it gives the relationship between scattering and particle size (Fig. 7) [29]. If, on the other hand, the particle size *D* is kept constant, α denotes the relationship between the scattering and the wavelength (l/λ replaces *D* on the abscissa of Fig. 7). The well-marked maxima are typical, and their existence signifies (1) that an optimum particle size must exist with respect to lightening power [30], and (2) that for a given particle size, there must be a particular wavelength for maximum light scattering [31].

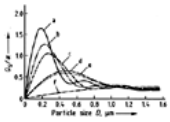

Figure 7. Scattering of white pigments as a function of particle size (λ = 550 nm)

a) Rutile; b) Anatase; c) Zinc sulfide; d) Zinc oxide; e) White lead; f) Barium sulfate

The first relationship can be used to predict the optimum particle size of white pigments (Table 7). The second relationship explains, for example, how white pigments in gray color mixtures can produce colored undertones as a result of selective light scattering (see Section Color Measurement).

Table 7. Refractive indices and optimal particle sizes of some white pigments (λ = 550 nm) calculated according to the van de Hulst formula [29]

Pigment	Formula	Mean refractive index		Optimal particle size (relative to binder) D_{opt}, µm
		Vacuum	Binder	
Rutile	TiO_2	2.80	1.89	0.19
Anatase	TiO_2	2.55	1.72	0.24
Zinc blende	α-ZnS	2.37	1.60	0.29
Baddeleyite	ZrO_2	2.17	1.47	0.37
Zincite	ZnO	2.01	1.36	0.48
Basic lead carbonate		2.01	1.36	0.48
Basic lead sulfate		1.93	1.30	0.57

The consequences of Mie's theory for *absorption* (i.e., for tinting strength) are now considered. Calculations from Mie's theory, using the relative refractive index *n* and the absorption index κ, are given in Figure 8 [32]. The parameter α on the abscissa can once more be taken as a relative measure of the particle size. The following conclusions may be drawn:

1. For very small particles, the absorption is independent of the particle size, and hence any further reduction in particle size does not produce additional absorption.
2. With increasing absorption index κ, the absorption of very small particles increases.
3. Absorption values for large particles are approximately equal for all relative refractive indices *n* and absorption indices κ, and decrease hyperbolically.
4. The top curve in Figure 8 applies to pigments with a high absorption index κ and low refractive index *n* (e.g., carbon black) and shows that the optimal particle size lies below a given limit.
5. The lowest curve applies to pigments with a small absorption index κ and high relative refractive index *n*, as is usually the case with inorganic pigments (e.g., red iron oxide). Here, there is a distinct maximum [11], [16].

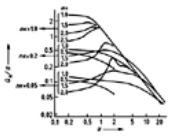

Figure 8. Absorption as a function of particle size

The above relationships (Fig. 8) show that the optical pigment properties depend on the particle size *D* and the complex refractive index $n^* = n\,(1 - i\,\kappa)$, which incorporates the real refractive index *n* and the absorption index κ. As a result, the reflectance spectrum and hence the color properties of a pigment can be calculated if its complex refractive index, concentration, and particle size distribution are known [33]. Unfortunately, reliable values for the necessary optical constants (refractive index *n* and absorption index κ) are lacking. These two parameters generally depend on the wavelength and, owing to the optical anisotropy of most pigments, on the illuminant and viewing direction. In pigments with a high absorption, the refractive index shows “anomalous dispersion”. Refractive indices of inorganic pigment materials are given at selected wavelengths in Tables 7 and 8. However, these values were measured on large crystals, and not on pigment particles. Unfortunately, no direct methods exist for

determining n and κ in colored inorganic pigments.

Table 8. Refractive indices of important colored inorganic pigments [86]

Mineral	Formula	Wavelength λ, nm	Refractive index		
			n_ω n_α	n_ε n_β	n_γ
Bismuth vanadate	$BiVO_4$	670	2.45 [a]		
Cobalt blue	$CoAl_2O_4$	(blue)	1.74		
Eskolaite	α-Cr_2O_3	671	2.5 [a]		
Greenockite	CdS	589	2.506	2.529	
Goethite	α-FeOOH	589	2.275	2.409	2.415
Hematite	α-Fe_2O_3	686	2.988	2.759	
Carbon [b]	C	578	1.97		
Crocoite	$PbCrO_4$	671	2.31	2.37	2.66
Magnetite	Fe_3O_4	589	2.42		
Red lead	Pb_3O_4	671	2.42		
Ultramarine [c]			1.50		

[a] Mean refractive index.
[b] For carbon arc lamps.
[c] Formula is $[AlSiO_4]_6(SO_4)(S,Cl)_2(Na,Ca)_8$.

1.3.2. *Color Measurement*

General. Two methods are used for color measurement: the tristimulus method using a three-filter colorimeter and the spectral method using a reflectance spectrophotometer (⟶ Dyes, General Survey – Color Measurement and Colorant Formulation), however, the three-filter colorimeters are gradually being displaced by spectrophotometers Manufacturers have developed three-filter reflectometers and recording spectrophotometers. Instruments are continuously being improved due to technological advances. Photometer illuminating/viewing geometries are standardized as methods A and B, and are designed to suit the individual application (see Section Test Methods). For standards, see Table 1 ("Color Differences, Conditions/Evaluation. . .").

Method A. The geometry d/8 (diffuse illumination, viewing from an angle of 8°) or 8/d (illumination from an angle of 8°, diffuse viewing) — including specular reflection — enables total surface reflection and reflection from the interior of a sample to be measured. An amount representing the surface reflection has to be subtracted from the measured value. Thus, measured color variations can be ascribed to differences or changes of the colorants in the interior of the sample.

Method B. Here, color differences in samples are evaluated in almost the same way as in visual evaluation by exclusion of gloss effects. Suitable geometries are d/8, d/0, 8/d, and 0/d with a gloss trap, and 45/0 and 0/45.

After the color of the sample and reference pigments has been measured, color differences are usually calculated by transformation of the X, Y, and Z values into the CIELAB system to calculate color differences (see ⟶ Dyes, General Survey – Chromaticity Diagram, ⟶ Dyes, General Survey, ⟶ Dyes, General Survey – Calculation of Color Differences). Color measurement results of black and white pigments can be expressed more simply because they only amount to a determination of the relative color undertone. For this, the environment of the reference pigment is divided into eight sectors, these being filled with color names from "red" to "violet". The octant in which the CIELAB color position of the sample is located is found by calculation.

When colored or black systems are reduced with white pigments, an undertone is observed, which is a particle-size effect of the white pigment (see Section Fundamental Aspects). These undertones can be conveniently expressed as CIELAB color differences. The effects can, however, also be measured by using the difference $R_z - R_x$ between the values obtained with the blue and red tristimulus filters. The undertone measured in this way depends on the lightness, and has a maximum at Y = 41.4. The lightness of a gray paste should therefore have this value to ensure that undertone differences between white pigments are comparable [34], [35].

Measured color differences are only true (i.e., significant) when they are not falsified by measuring errors. A *significance* test standard (see Table 1, "Color differences, significance") has been developed to check this [35]. The numerical value of a color difference must be higher than a critical value, which is statistically calculated using the standard deviation.

Problems concerning the *acceptance* (tolerance) of color differences (e.g., in production quality control or in computer color matching) should also be solved by mathematical statistics [16].

The *gloss* of pigmented coatings is not a true pigment property. The pigment can, however, influence the luster quality, mainly via its dispersing properties (see Section Test Methods). The degree of gloss of a coating can range from high gloss (specular reflection) to an ideally matt surface (complete scattering). *Gloss haze* is due to a disturbance of specular reflection: the reflected objects appear as seen through a veil, this is caused by halation effects. Gloss retention is discussed in Section Fundamental Aspects , a method of gloss measurement is described in Section Test Methods. Special problems arise in the measurement of black [36] and fluorescent pigments [37].

Methods of Determination. *Lightness*. The white pigment powder is compressed in a suitable powder press to give an even, matt surface. The CIE tristimulus value Y is measured with color measuring equipment. For standards, see Table 1 ("Lightness"). Apparatus: spectrophotometer or tristimulus colorimeter, powder press, white standard.

Full Shade. Full-shade systems are media that contain only a single pigment. The color of the full-shade system in an optically infinitely thick (opaque) coating is referred to as full shade. The *mass tone* denotes the color obtained when the pigmented medium is applied as a layer that does not hide the substrate completely (e.g., on a white substrate). Evaluation can be carried out visually or by color measurement. For standards, see Table 1 ("Color in full-shade systems"). Apparatus: spectrophotometer or tristimulus colorimeter.

Color Difference. The computer printout in Figure 9 shows the CIELAB color positions of a reference pigment and a test sample: the color difference ΔE^*_{ab}, and the derived differences (Section Fundamental Aspects), i.e., lightness difference ΔL^*, chroma difference ΔC^*_{ab}, and hue difference ΔH^*_{ab}. The color difference of colored pigments in reduction can be similarly determined, including the "color difference after color reduction" which arises when tinting strength matching is carried out (see Section Tinting Strength, Lightening Power, and Scattering Power). An improved color difference formula takes into account the fact that chroma and hue differences can themselves depend on the chroma [16].

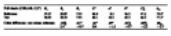

Figure 9. CIELAB color differences between two yellow iron oxide pigments

Undertone of Near-White Samples and Gray Undertones. The undertone of an almost uncolored sample is the small amount of color by which the color of a sample differs from ideal white or achromatic material. It is described by hue and chroma. The distance and direction of the CIELAB color position of the test sample (a^*_T, b^*_T) from the achromatic position (0, 0) are used to characterize the hue. The *relative undertone* is expressed by the distance and direction between the CIELAB color position of the test sample (a^*_T, b^*_T) and that of the reference pigment (a^*_R, b^*_R). In both cases, the distance is expressed by a figure and the direction by a color name. For standards, see Table 1 ("Hue relative of near white specimens"). Apparatus: spectrophotometer or tristimulus colorimeter for determining the CIE tristimulus values X, Y and Z with standard illuminants D 65 or C. With a computer, the color name can be printed out instead of the number of the octant (for example, see Fig. 10).

Figure 10. Undertone of two nearly white samples (TiO_2 pigments)

1.3.3. *Tinting Strength, Lightening Power, and Scattering Power*

The *tinting strength* is a measure of the ability of a colorant to confer color to a light-scattering material by virtue of its absorption properties. The *lightening power* can be considered as the tinting strength of a white pigment, and is a measure of its ability to increase the reflectance of an absorbing (black or colored) medium by virtue of its scattering power. Tinting strength is expressed as the mass ratio in which the reference pigment (mass, m_R) can be replaced by the test pigment (mass, m_T) to give the same color quality in a white system. Analogously, lightening power is the mass ratio in which the reference pigment can be replaced by the test pigment to give the same lightness in a colored system. Thus, the same equation defines tinting strength and lightening power:

$$P = (m_R/m_T)_{\Delta Q=\text{const}}$$

where ΔQ = const expresses the tinting strength matching. Both parameters are yield properties; if, for example, tinting strength is doubled, only half the weight of pigment is required. Optical properties can therefore provide information about the economic performance of a colored pigment ("value for money"). Testing of tinting strength and lightening power can be rationalized by means of the Kubelka – Munk theory (see Section Fundamental Aspects).

Tinting Strength. Within certain limits, relative tinting strength can be interpreted as the ratio of the absorption coefficients of equal masses of test and reference pigments. This procedure avoids visual matching of the test and reference pigments employed in a previously used method (tinting strength matching). For standards, see Table 1 ("Tinting strength, relative"). Materials and apparatus: white paste, spectrophotometer or tristimulus photometer.

In the determination of the tinting strength, specification of the matching between the test and the reference pigments is extremely important [38]. Since the criterion used for matching greatly influences the value of the tinting strength, it is not permissible to speak of tinting strength purely and simply. There are as many tinting strength values as there are matching criteria. In the German Standard, the following criteria are permitted:

1. The lowest of the three CIE tristimulus values *X, Y* and *Z*
2. The CIE tristimulus value *Y*
3. The depth of shade (see Section Light Stability).

When determining the tinting strength of inorganic pigments, the tristimulus value *Y* (lightness) is usually used [38].

Another standard method uses depth of shade as the matching criterion (FIAF method), but it is employed almost exclusively for testing organic pigments. This method can be applied by means of the "principle of spectral evaluation" (see Section Hiding Power and Transparency) [39], which uses the wavelength-dependent Kubelka – Munk coefficients $S(\lambda)$, $K(\lambda)$ to calculate the match [40]. The tinting strength for all other matching criteria can be determined by applying the principle of spectral evaluation. (For tinting strength in cement, see Section Miscellaneous Pigment – Binder Systems; for change in tinting strength, see Section Dispersing Behavior in Paint Systems).

Lightening Power. If the gray paste method is used to determine the lightening power (for standards, see Table 1, "Lightening power of white pigments"), the concentration at which the white pigment should be assessed has to be agreed between the intended parties in accordance with the type of application [41]. This process is especially recommended for the routine testing of pigments because the lightening power, the relative scattering power (see below), and the undertone can be determined with the same gray milling paste of the test pigment.

The method can be rationalized by means of the Kubelka – Munk theory. For standards, see Table 1 ("Tinting strength, relative: Photometric").

Relative Scattering Power. For standards, see Table 1 ("Scattering power, relative"). The relative scattering power *S* is the ratio of the scattering power S_T of the test white pigment to the scattering power of a white reference pigment S_R [42]. It can be determined in two ways:

1. *Black-Ground Method*. The relative scattering power is determined from the tristimulus values Y of the pigmented medium applied in various film thicknesses to black substrates. Compared with the gray paste method, the black-ground method has the advantage that it is not restricted to any particular test medium. Apparatus: spectrophotometer or tristimulus colorimeter.
2. *Gray Paste Method.* The relative scattering power is determined from the tristimulus values *Y* of gray pastes. The method has the advantage of being less time consuming than the black-ground method. The results of the two methods are not, however, generally in agreement. Materials and apparatus: black paste, spectrophotometer or tristimulus colorimeter.

1.3.4. *Hiding Power and Transparency*

The definition of hiding power is based on a black and white contrasting support upon which the film of coating is applied. The thickness *h* of the applied film is determined at the point at which the contrasting surface just disappears, as judged by eye. The film thickness (mm) which fulfils this condition is called the *hiding thickness*. Its reciprocal, the *hiding power* ($mm^{-1} = m^2/L$) is, like tinting strength, an indicator of yield because it gives the area (m^2) that can be covered up with 1 L of applied paint.

This traditional visual testing method has been improved by the use of a photometer and a colorimetric criterion to evaluate the hiding film thickness, e.g., $\Delta E^*_{ab} = 1$ for equally thick films applied to black and white backgrounds. The method can, however, be further rationalized. In this method only a single coating has to be prepared. The total reflectance spectrum is measured and after measuring the film thickness, the spectral scattering and absorption coefficients $S(\lambda)$ and $K(\lambda)$ are calculated at selected wavelengths (e.g., at intervals of 10 nm) by the Kubelka – Munk method. Once these values are known, the expected reflectances over black and white at the chosen wavelengths of the spectrum can be calculated for a given film thickness. Using the CIE and CIELAB systems, the CIELAB position of a color which would have this spectrum is calculated. This provides an iteration method for determining the film thickness at which the hiding power criterion $\Delta E^*_{ab} = 1$ is satisfied [39], [43], [44]. This principle of spectral evaluation [39] is the key to handling test methods based on the visual matching of two color samples (see also "Transparency" and "Tinting Strength", Section Tinting Strength, Lightening Power, and Scattering Power).

The *transparency* of a pigmented system denotes its ability to scatter light as little as possible. (see Table 1) The color change of a transparent pigmented system when applied to a black substrate has to be very small; the lower the color change, the higher is the transparency [45]. Measurement of transparency is important for assessing transparent varnishes and printing inks.

Hiding Power. The hiding criterion is an agreed color difference between two contrasting areas of a coated contrast substrate. The hiding power can be determined and expressed as D_V (m^2/L) or D_m (m^2/kg), which give the area (m^2) of the contrasting substrate that can be coated with 1 L or 1 kg, respectively, of the pigmented medium when the hiding criterion is satisfied. Layers of various thicknesses or mass-to-area ratios of the test material are applied to the contrast background. The color difference ΔE^*_{ab} between the two contrasting areas is determined by color measurement on each of the layers. The color difference is plotted against the reciprocal of the film thickness (mm^{-1}) or the reciprocal of the mass to area ratio (m^2/kg). The hiding power D_V (m^2/L) or D_m (m^2/kg) at the hiding criterion (e.g., $\Delta E^*_{ab} = 1$) is then obtained from the diagram (Fig. 11). For standards, see Table 1 ("Hiding power"). Apparatus: spectrophotometer or tristimulus colorimeter, film thickness measuring equipment.

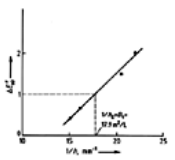

Figure 11. Determination of the hiding power of a yellow iron oxide pigment

A modern well equipped color measurement laboratory can use the "principle of spectral evaluation" described above [39] to simulate this procedure with a computer. The thickness of the hiding film can then be calculated in advance from the reflectance curves of a single film on a black/white substrate at a known film thickness.

Hiding Power of Achromatic Coatings. Equally thick coatings of the pigmented coating material are applied to a black and white background, and their reflectances measured with a photometer after drying. The hiding power is then calculated from the reflectances on black and white substrates and the film thickness. The hiding criterion is $\Delta L^* = 1$, in analogy with that for colored pigments [14]. The use of hiding power as a basis for assessing the economy and effectiveness of pigments is described in [46]. Apparatus: spectrophotometer or tristimulus colorimeter, black and white glass plates.

Transparency. Transparency is expressed quantitatively as the transparency number. This is defined as the reciprocal of the increase in color difference ΔE^*_{ab} on a black substrate obtained on increasing the film thickness *h* of the pigmented medium. The transparency number has the unit mm (= L/m^2). It indicates the number of liters of

pigmented medium needed to coat 1 m^2 of a black substrate in order to obtain a color difference of ΔE^*_{ab} =1 relative to this substrate. In a simplified method the transparency number can be determined by evaluating one or two points on the straight part of the $\Delta E^*_{ab}(h)$ curve. A computer method is more exact, furthermore calculations can be made using the spectral "principle of spectral evaluation." For standards, see Table 1 ("Transparency").

1.4. Stability Towards Light, Weather, Heat, and Chemicals

1.4.1. *Fundamental Aspects [47]*

Many pigmented systems show typical color or structural changes when subjected to intense radiation or weathering [48]. The best-known of these are yellowing [49], chalking, and loss of gloss. These processes involve photochemical reactions in which the pigment can act as a catalyst or in which the pigment itself undergoes chemical changes.

Inorganic pigments are chemically very stable and are classed as one of the most stable coloring matters. This is especially true for oxide pigments which often have a highly protective effect on the substrate [50]. Apart from the purely mechanical stabilization imparted by the pigment, this protective effect can give a coating an increased economic advantage. On the other hand, sulfide pigments can be oxidized by the atmosphere to form sulfates which can be washed away by rain [51].

The photochemical processes that take place when TiO_2-containing coatings undergo chalking have been elucidated [52]. The shorter wavelength radiation of sunlight acts on rainwater and atmospheric oxygen to form extremely reactive radicals (·OH, HO_2·) that cause deterioration of the coating matrix by oxidative attack. Titanium dioxide pigments may be stabilized by reducing the number of radical-producing hydroxy groups on the surface of the TiO_2 particles (e.g., by doping with zinc oxide). Alternatively, coatings of oxide hydrates are produced by aftertreatment of the pigment surface to give "reaction walls" on which the radicals are destroyed [53].

Continuous breakdown of the binder leads to the loss of gloss. Surface gloss gives an attractive appearance and is usually desired, it is obtained by ensuring that the pigment particles are well dispersed. The stability of gloss to weathering is therefore of great importance. If breakdown of the binder is so extensive that the pigment may be loosened, *chalking* takes place. Chalking is defined as loosening of pigment or extender particles following destruction of the binder at the surface. For standards, see Table 1 ("Coating materials").

Resistance to light and weather generally depend on the chemical composition, structure, defects, particle shape and size, and concentration of the pigment [54]. However, these properties also depend on the medium in which the pigment is used. Testing is carried out by open-air weathering, accelerated weathering, and chemical test methods.

Accelerated weathering is carried out in special weathering devices (e.g., Weatherometer, Xenotest) which simulate exposure to sunlight and periods of rain. The spectral composition of the light used (Fig. 12) should match the "global radiation" in temperate latitudes (global radiation denotes the sum of the direct solar radiation and indirect radiation from the sky). However, accelerated tests only allow limited conclusions to be drawn about light and weather resistances because it is not possible to accurately simulate all climatic conditions, their combinations and sequences, or sudden changes. Even so, natural weathering tests are not without problems: weather cannot be standardized, and a "uniform" weather therefore does not exist [55], [56]. Weathering results, even from successive years, can thus vary considerably [57]. Weather also varies with the geographical coordinates. For standards, see Table 1 ("Climates: Normalized"); the earth is classified into climatic regions and weathering adjustment scales exist [58].

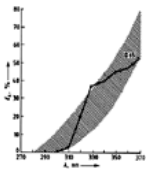

Figure 12. Spectral energy distribution of daylight phase D65 with permitted range of deviation (shading)

None of the chemical methods (e.g., mandelic acid or methanol test) gives a perfect indication of weather resistance. Most of them only measure partial qualities of the photochemical reactivity [59].

The oxide pigments have the highest *heat stability*, followed by the sulfide pigments which can even be used in enamels and glass melts (e.g., cadmium pigments). The oxide hydroxides and carbonates are less heat stable. Thermal stability is of great importance for modern coating techniques (e.g., for stoving finishes and other products requiring high temperatures). The heat resistance of a pigment primarily depends on the binder and the duration of heating. It is conveniently determined by observing hue changes in the pigmented coating system after a defined heat treatment, e.g., with white coatings yellowing occurs. For standards, see Table 1 ("PVC").

The *chemical resistance* of inorganic pigments, especially oxide pigments, is in general very high [60]. Other fastness properties such as overspray fastness, fastness to blooming, and fastness to plasticizers are of more relevance to organic pigments. For corrosion inhibiting (anticorrosive) pigments, see Section Anticorrosive Pigments.

1.4.2. *Test Methods*

1.4.2.1. *Light Stability*

Accelerated Tests. Light stability is a material property. It is defined as the resistance of coatings towards changes caused by the action of global radiation (daylight phase: standard illuminant D 65), possibly in the presence of moisture. In coatings, light and moisture can lead to chalking and changes in gloss and color. For standards, see Table 1 ("Light stability"). Apparatus: testing cabinets, equipment for wetting the test plates, equipment for producing air flow.

Evaluation of Color Changes (see Section Color Measurement). Color changes are differences in lightness, chroma, and hue such as exist between identical samples with different histories. Evaluation methods are particularly suitable for the determination and evaluation of color changes that occur following physical and chemical stresses on pigmented materials. The CIE tristimulus values *X, Y*, and *Z* of the samples or of different test locations on a sample are determined with a colorimeter . The CIELAB color difference is calculated for each sample or test location from the tristimulus values before and after the stress (see Section Color Measurement). Apparatus: spectrophotometer, tristimulus colorimeter.

Standard Depth of Shade. The depth of shade is a measure of the intensity of a color sensation. It increases with increasing chroma and generally decreases with increasing lightness. The colorfastness of colorations (e.g., textile dyeing) is dependent on depth of shade; the fastness properties of pigments or pigmented systems must therefore be tested at a predetermined depth of shade. This standard method therefore ensures that samples with standard depths of shade are used to determine the fastness properties of pigments or pigmented systems [61], [62]. For standards, see Table 1 ("Standard depth of shade"). Apparatus: spectrophotometer or tristimulus filter photometer with gloss trap. (For light stability in cement, see Section Miscellaneous Pigment – Binder Systems.)

1.4.2.2. *Weather Resistance*

The empirical tests described below determine the chalking and weathering differences between two pigments, but do not always give the true differences. Other test methods have therefore been developed, e.g., determination of mass losses on weathering (gravimetric test) [52].

Accelerated Tests. Weather resistance in an accelerated test is defined as the resistance of plastics towards changes caused by simulated open-air weathering (simulation of global radiation by means of filtered xenon arc radiation and periodic rain). After the weathering (measured by the product of intensity and duration), defined properties of the test sample are compared with those of an identical unweathered sample. Properties should be considered which are of practical importance, such as color or surface properties. For standards, see Table 1 ("Weathering in apparatus"). Apparatus: test chamber, rain and air humidification equipment, air flow equipment, radiation measuring equipment.

Degree of Chalking. The degree of chalking of a coating is measured by the quantity of loose pigment particles. It may be determined by three methods:

1. *Kempf Method.* For standards, see Table 1 ("Chalking degree: Kempf method"). Water-treated photographic paper is applied to the coating (with the humid gelatin side facing the coating surface) and pressed by means of a special stamp device. Loose pigment particles that are not attached to the binder, are picked up by the gelatin coating. The degree of chalking is judged visually (for scale of measurement see Table 1, "Test evaluation"). See also Section Methods of Determination. Apparatus: Kempf stamp equipment.
2. *Adhesive Tape Method.* For standards, see Table 1 ("Chalking degree; Adhesive tape method"). Loose pigment particles are picked up by transparent adhesive tape. The degree of chalking is determined from a reference scale (see Table 1, "Test Evaluation" and Section Methods of Determination) or by using a comparative scale that is agreed upon by the interested parties.
3. *Photographic Methods.* For standards, see Table 1 ("Chalking degree: Photographic method"). Chalking prints on paper are compared with photographic standards.

Assessment of Gloss. Gloss is not a purely physical quantity, it is also dependent on physiological and psychological factors. The usefulness of measured reflectance values lies in the possibility of detecting surface changes of the test sample and observing them over a long period. These changes can be produced by weathering, abrasion, and similar causes of surface wear, or by the method used to produce the surface. For standards, see Table 1 ("Reflectometer"). Apparatus: reflectometer with

incident light angles of 20°, 60°, or 85°. For standards see “Gloss” in Table 1.

Effects of Humid Climates. The specimen is exposed to warm air saturated with water vapor, with or without intermittent cooling to room temperature. The test is very suitable for assessing the corrosion resistance of metals, protective coatings, and composite materials in buildings. For standards, see Table 1 (“Climates: Containing evaporated water”). Apparatus: Kesternich condensation equipment.

1.4.2.3. *Heat Stability [63]*

Measured values of the heat stability of pigments depend on the type and duration of the heat applied, and also on the binder used. Low thermal stability in white coatings leads to yellowing, even when the unpigmented binder is resistant to yellowing. With colored coatings, changes in the hue of the pigmented binder can be measured colorimetrically (see Section Color Measurement). Standards only exist for special systems (e.g., hard and soft PVC; and cement, see Section Miscellaneous Pigment – Binder Systems). Thermal resistances of pigments are customarily quoted at two temperatures: (1) the maximum temperature at which the hue remains stable, and (2) the temperature at which a defined hue change is observed. For standards, see Table 1 (“Heat stability” and “Coloration of building materials”).

1.4.2.4. *Fastness to Chemicals [63]*

General chemical resistance can be classified as short-, medium-, or long-time resistances. Changes in the coating are assessed visually. For standards, see Table 1 (“Chemical resistance”). Apparatus: film thickness measuring equipment.

Resistance to Water in Sulfur-Dioxide-Containing Atmospheres. For standards, see Table 1 (“Corrosion testing; SO_2” and “SO_2 resistance”). General conditions are standardized for exposing samples to a varying climate where water condenses in the presence of sulfur dioxide, so that tests in different laboratories give comparable results. The test allows rapid detection of defects in corrosion-inhibiting systems.

Furthermore, a method for determining the colorfastness of pigments in binders in the presence of sulfur dioxide is also prescribed. The test is carried out simultaneously on ten identical samples and consists of three cycles. Apparatus: condensation equipment.

Salt Spray Fog Test with Sodium Chloride Solutions. The sample is sprayed continuously with a 5 % aqueous sodium chloride solution. There are three variations of the test: the salt spray fog test, the acetic acid salt spray test, and the copper chloride – acetic acid salt spray test.

For standards, see Table 1 (“Corrosion testing: NaCl”). Materials and apparatus: spray chamber, test solutions (5 % NaCl solutions).

Resistance to Spittle and Sweat. This test indicates whether a pigment on a child's colored toy is likely to be transferred to the mouth, mucous membranes, or skin during use. Strips of filter paper are wetted with $NaHCO_3$ and NaCl solutions and pressed against the test samples. The discoloration of the paper is judged visually. For standards, see Table 1 (“Spittle/sweat resistance”).

1.5. Behavior of Pigments in Binders

1.5.1. *Fundamental Aspects [64], [65]*

A pigment – binder dispersion is a suspension before it is dried; after drying it is a solid sol. In pigment – binder systems, the concepts and laws of colloid chemistry therefore apply. The dispersing of pigments and extenders in binders is an extremely complex process of a series of steps that can be interlinked [64], see also ⟶ Paints and Coatings – Production Technology. Dispersing involves the following steps:

1. *Wetting.* Removal of the air from the surface of the pigment particles and formation of a solvate layer [66], [67].
2. *Disintegration.* Breaking up of pigment agglomerates with external energy (e.g., by dispersing equipment).
3. *Stabilization.* Maintenance of the disperse state by creating repulsive forces between the particles (e.g., by coating them with solvate layers). These forces must be greater than the van der Waals attraction forces [68] that cause flocculation [69], [70].

The degree of dispersing usually has a large influence on the properties of the coating system [71], [72]. Some of these properties depend strongly on the degree of dispersing, and are therefore used to measure dispersibility. For standards, see Table 1 (“Ease of dispersion”). Examples are viscosity (pigment – binder dispersions always show non-Newtonian behavior properties before drying) and color properties, especially hue, tinting strength, gloss, and gloss haze [73] (see Sections Color Properties and Stability Towards Light, Weather, Heat, and Chemicals).

In many systems, tinting strength depends so highly on the degree of dispersing that the rate of tinting strength development can be taken as a direct measure of dispersibility [74-77] (see Section Dispersing Behavior in Paint Systems). The numerical values of the half-life times (characterizing the dispersibility) and of the increases in strength (characterizing the strength potential) depend on the binder and the dispersing equipment used.

In systems where different pigments are combined (e.g., white and colored pigments), segregation effects can occur which change the optical appearance (flooding [79]). Flooding can be counteracted by preflocculation of the pigments with suitable gelling agents, and in some cases by addition of extenders (see ⟶ Paints and Coatings – Wetting and Dispersing Additives).

In solvent-containing systems with several pigments, but also in full-shade systems, color changes (rub-out effect) can take place due to flocculation, especially if the systems are subjected to mechanical stress during application and drying. The flocculation tendency can be determined by means of the “rub-out test” (see ⟶ Paints and Coatings – Wetting and Dispersing Additives).

1.5.2. *Test Methods*

1.5.2.1. *Pigment – Binder Interaction*

Oil Absorption. The oil absorption gives the mass or volume of linseed oil required to form a coherent putty-like mass with 100 g pigment under specified conditions. The mixture should just not smear on a glass plate. For standards, see Table 1 (“Oil absorption”). Materials and apparatus: rough glass plate, spatula with steel blade, raw linseed oil.

Binder Absorption, Smear Point, and Yield Point [80]. Smear point and yield point are used to determine the binder needed to formulate a suitable millbase for grinding by dissolvers, roll mills, ball mills, attritors, sand mills, and pearl mills. The amount of binder is given in volume or mass units. Apparatus: rough glass plate, spatula with steel blade.

Viscosity. Viscosity is a useful parameter for coatings applied by brushing. It is determined by measuring the torque applied to a rotating cylinder or disk in or on the surface of a suspension. This method is applicable to Newtonian and non-Newtonian systems. In non-Newtonian suspensions, the apparent viscosity is obtained by dividing the shear stress by the shear rate. The viscosities provide information about the force required in the initial brushing phase of a coating. For standards, see Table 1 (“Viscosity”). Apparatus: cone/plate or rotating cylinder viscometer.

Fineness of Grind. The grind gauge (grindometer) consists of a steel block with a groove. The depth of the groove at one end is approximately twice the diameter of the largest pigment particle, and decreases continuously to zero at the other end. The sample is placed at the deep end of the groove and drawn to the other end with a scraper. The depth at which a large number of particles become visible on the surface of the pigment – binder system as pinholes or scratches is read off from a scale graduated in micrometers. For standards, see Table 1 (“Fineness of grind”). Apparatus: Hegman's grindometer gauge (block and scraper).

1.5.2.2. *Dispersing Behavior in Paint Systems*

Low-Viscosity Media. The time is measured that is required to produce a homogeneous suspension of particles in the dispersion medium using an oscillatory shaking machine equipped with several containers. Not only can small quantities of the millbase with the same composition be tested (as with other types of apparatus), but various millbases can also be tested under the same conditions. A low-viscosity alkyd resin system of the stoving or oxidatively drying type can be used as a test medium. For standards, see Table 1 (“Ease of dispersion: Oscillatory shaking machine”).

High-Viscosity Media (Pastes). The dispersion properties are determined with an automatic muller. The advantage is that small amounts of material can be simply tested under reproducible conditions. For standards, see Table 1 (“Ease of dispersion: Automatic muller“).

Development of Fineness of Grind. The quantity measured is the dispersing effect needed to achieve a given fineness of grind. Samples of the product are taken at various stages of the dispersion process and the fineness of grind is determined with a grindometer gauge. The method is used to compare diverse millbases material types of dispersing equipment, or dispersing methods used with the same pigment. For standards, see Table 1 (“Fineness of grind”).

Change in Tinting Strength, Half-Life Times, and Increase in Strength. For standards see Table 1 (“Change in strength”). Dispersibilities of pigments may be compared by means of their dispersing resistance, dispersing equipment may be characterized by means of dispersing effects. Measurement can be based on determination of half-life times in relation to the final tinting strength. The increase in strength, however, only gives the difference between the initial and final tinting strength [77]. Half-life times are functions of the dispersing rate. They are only valid for a given combination of dispersing equipment, dispersing process, and medium. Samples are taken from the millbase at specified stages in the dispersing process and mixed with a white paint based on the same binder material, or with a compatible white paste. The tinting strength is then

determined (see Section Tinting Strength, Lightening Power, and Scattering Power) and a graph is drawn; development of tinting strength is plotted against the dispersion stages (Fig. 13) [81].

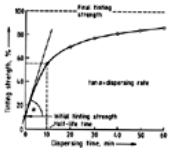

Figure 13. Change in tinting strength as a function of dispersing time

A new method allows to calculate the three dispersion velocities [78]:

- Size reduction vs. large particles
- Size reduction vs. small particles
- Size reduction vs. the vessel walls/grinding bodies

Surprisingly, the dispersion rate for the size reduction vs. small particles is negative, which indicates that no reduction but reagglomeration takes place. This new method requires two millbases with different pigment volume concentration and a computer.

Pigment Volume Concentration (PVC) and Critical Pigment Volume Concentration (CPVC). The pigment volume concentration (σ) is the fractional volume of pigment in the total solids volume of the dry paint film:

$$\sigma = \frac{V_p}{V_p + V_b}$$

where V_p is the pigment volume and V_b the binder volume. The *PVC* [82] is determined by separating the pigment fraction from a weighed paint sample by, for example, extracting the (liquid) binder with suitable organic solvents or solvent mixtures (sometimes using a centrifuge) or using combustion methods (sometimes including fuming with sulfuric acid) and analysis of the residue [83].

Under *CPVC* conditions, the pigment particles are at a maximum packing density, and the interstices are completely filled with binder. With smaller amounts of binder, the interstices are incompletely filled. The *CPVC* thus represents a pigment concentration boundary at which abrupt changes in the properties of the film occur.

The methods of determining the *CPVC* [84] are divided into two groups:

1. Methods based on the observation of a sudden change in properties in a series of *PVCs* (e.g., ion permeability, permeability to water vapor, color, gloss stability after coating with a silk-luster paint)
2. Methods based on the experimental production of the dense pigment packing typical of *CPVC*, e.g., by means of oil absorption, filtration, vacuum filtration, or by measuring the volume of the supercritical film

Measurement of Deposit. Common methods used to measure sediments are manual testing, sedimentation balance, radioactive methods, dipping of bodies, and probe tools.

The special measuring equipment known as the Bayer system [85], can test for deposits with great accuracy and good reproducibility. The force acting on a needle probe is measured which penetrates into the coating material from above.

1.5.2.3. *Miscellaneous Pigment – Binder Systems*

Plastics. Specifications for thermoplastics are given in Table 1 ("Thermoplastics"). Specifications for poly(vinyl chloride) are also listed ("PVC"). These specifications include methods for producing basic mixtures for testing pigments in PVC and for specimen preparation. A method is described for determining bleeding, i.e., the migration of coloring matter into a material in contact with the sampling equipment. Special procedures are also included for the determination of heat stability and increase in strength caused by cold rolling.

Building Materials. For standards, see Table 1 ("Coloration of building materials"). The following tests are prescribed for assessing the suitability of a pigment for coloring cement and lime-bonded building materials:

Relative tinting strength
Colorfastness in cement
Colorfastness in lime
Lightfastness
Heat stability
Influence of the pigment on hardness of the building material
Influence of the pigment on setting properties

Paper and Board. For standards, see Table 1 ("Opacity" and "Transparency"). Methods of measurement are specified for the reflectance of paper and board (nonfluorescent), and the opacity or transparency of paper (to measure the transmitted light).

[Top of Page]

2. White Pigments

Jürgen Kischkewitz, Peter Woditsch, Axel Westerhaus, Wolf-Dieter Griebler, Marcel De Liedekerke

White pigments include TiO_2, zinc white (ZnO), zinc sulfide, lithopone (a mixed pigment produced from zinc sulfide and barium sulfate), and white lead (basic lead carbonate). The optical properties of white pigments are a result of their low light absorption and their strong, mainly nonselective, scattering of light.

2.1. Titanium Dioxide

Titanium dioxide [13463-67-7], TiO_2, M_r 79.90, occurs in nature in the modifications rutile, anatase, and brookite. Rutile and anatase are produced industrially in large quantities and are used as pigments and catalysts, and in the production of ceramic materials.

Titanium dioxide is of outstanding importance as a white pigment because of its scattering properties (which are superior to those of all other white pigments), its chemical stability, and lack of toxicity. Titanium dioxide is the most important inorganic pigment in terms of quantity, 3.2×10^6 t were produced in 1995. World production of titanium dioxide pigment is shown in Table 9 [87-89].

Table 9. World production of TiO_2 pigment

Year	Sulfate process		Chloride process		Total
	10^3 t/a	%	10^3 t/a	%	10^3 t/a
1965	1254	90.3	135	9.7	1389
1970	1499	77.4	437	22.6	1936
1977	1873	72.3	716	27.7	2589
1988	1781	60.2	1178	39.8	2959
1995	1481	46	1739	54.0	3220
2000*	1540	40	2310	60.0	3850

* Estimated.

2.1.1. *Properties [90], [91]*

Physical Properties. Of the three modifications of TiO_2, rutile is the most thermodynamically stable. Nevertheless, the lattice energies of the other phases are similar and hence are stable over long periods. Above 700 °C, the monotropic conversion of anatase to rutile takes place rapidly. Brookite is difficult to produce, and therefore has no value in the TiO_2 pigment industry.

In all three TiO_2 modifications one titanium atom in the lattice is surrounded octahedrally by six oxygen atoms, and each oxygen atom is surrounded by three titanium atoms in a trigonal arrangement. The three modifications correspond to different ways of linking the octahedra at their corners and edges. Crystal lattice constants and densities are given in Table 10.

Table 10. Crystallographic data for TiO_2 modifications

Phase	CAS registry no.	Crystal system	Lattice constants, nm			Density, g/cm^3
			a	*b*	*c*	
Rutile	[1317-80-2]	tetragonal	0.4594		0.2958	4.21
Anatase	[1317-70-0]	tetragonal	0.3785		0.9514	4.06
Brookite	[12188-41-9]	rhombic	0.9184	0.5447	0.5145	4.13

Rutile and anatase crystallize in the tetragonal system, brookite in the rhombic system. The melting point of TiO_2 is ca. 1800 °C. Above 1000 °C, the oxygen partial pressure increases continuously as oxygen is liberated and lower oxides of titanium are formed. This is accompanied by changes in color and electrical conductivity. Above 400 °C, a significant yellow color develops, caused by thermal expansion of the lattice; this is reversible. Rutile has the highest density and the most compact atomic structure, and is thus the hardest modification (Mohs hardness 6.5 – 7.0). Anatase is considerably softer (Mohs hardness 5.5).

Titanium dioxide is a light-sensitive semiconductor, and absorbs electromagnetic radiation in the near UV region. The energy difference between the valence and the conductivity bands in the solid state is 3.05 eV for rutile and 3.29 eV for anatase, corresponding to an absorption band at < 415 nm for rutile and < 385 nm for anatase.

Absorption of light energy causes an electron to be excited from the valence band to the conductivity band. This electron and the electron hole are mobile, and can move on the surface of the solid where they take part in redox reactions.

Chemical Properties. Titanium dioxide is amphoteric with very weak acidic and basic character. Accordingly, alkali-metal titanates and free titanic acids are unstable in water, forming amorphous titanium oxide hydroxides on hydrolysis.

Titanium dioxide is chemically very stable, and is not attacked by most organic and inorganic reagents. It dissolves in concentrated sulfuric acid and in hydrofluoric acid, and is attacked and dissolved by alkaline and acidic molten materials.

At high temperature, TiO_2 reacts with reducing agents such as carbon monoxide, hydrogen, and ammonia to form titanium oxides of lower valency; metallic titanium is not formed. Titanium dioxide reacts with chlorine in the presence of carbon above 500 °C to form titanium tetrachloride.

Surface Properties of TiO_2 Pigments. The specific surface area of TiO_2 pigments can vary between 0.5 and 300 m^2/g depending on its use. The surface of TiO_2 is saturated by coordinatively bonded water, which then forms hydroxyl ions. Depending on the type of bonding of the hydroxyl groups to the titanium, these groups possess acidic or basic character [92], [93]. The surface of TiO_2 is thus always polar. The surface covering of hydroxyl groups has a decisive influence on pigment properties such as dispersibility and weather resistance.

The presence of the hydroxyl groups makes photochemically induced reactions possible, e.g., the decomposition of water into hydrogen and oxygen and the reduction of nitrogen to ammonia and hydrazine (see also Section Pigment Properties) [94].

2.1.2. *Raw Materials [95], [111]*

The raw materials for TiO_2 production include natural products such as ilmenite, leucoxene, and rutile, and some very important synthetic materials such as titanium slag and synthetic rutile. Production figures for the most important titanium-containing raw materials are listed in Table 11 [96]. Total production of titanoferrous raw materials, excluding that produced and consumed in the former Soviet Union and China, grew in 1994 to 3.69 × 10^6 t of contained TiO_2 [96]. Australia was the largest producer, followed by Africa, Canada, and Norway.

Table 11. Production of titanium-containing raw materials (1994)

Raw material	Location	Capacity, 10^3 t/a
Ilmenite*	Australia	1550
	Canada	1850
	Former Soviet Union	250
	India	280
	Norway	700
	Republic of South Africa	1500
	United States	320
	Others	450
	Total	7120
Rutile	Australia	216
	Former Soviet Union	30
	Republic of South Africa	90
	Sierra Leone	144
	Others	70
	Total	550
Synthetic rutile	Australia	417
	India	43
	Japan	10
	United States	140
	Total	625
Titanium slag	Canada	764
	Norway	170
	Republic of South Africa	773
	Total	625

* Includes ilmenite used in the production of titanium slag and synthetic rutile.

2.1.2.1. *Natural Raw Materials*

Titanium is the ninth most abundant element in the earth's crust, and always occurs in combination with oxygen. The more important titanium minerals are shown in Table 12. Of the natural titanium minerals, only ilmenite (leucoxene) and rutile are of economic importance. Leucoxene is a weathering product of ilmenite.

Table 12. Titanium minerals

Mineral	Formula	TiO_2 content, wt %
Rutile	TiO_2	92 – 98
Anatase	TiO_2	90 – 95
Brookite	TiO_2	90 – 100
Ilmenite	$FeTiO_3$	35 – 60
Leucoxene	$Fe_2O_3TiO_2$	60 – 90
Perovskite	$CaTiO_3$	40 – 60
Sphene (titanite)	$CaTiSiO_5$	30 – 42
Titanomagnetite	$Fe(Ti)Fe_2O_4$	2 – 20

The largest titanium reserves in the world are in the form of anatase and titanomagnetite, but these cannot be worked economically at the present time. About 95 % of the world's production of ilmenite and rutile is used to produce TiO_2 pigments, the remainder for the manufacture of titanium metal and in welding electrodes.

Ilmenite and Leucoxene. *Ilmenite* is found worldwide in primary massive ore deposits or as secondary alluvial deposits (sands) that contain heavy minerals. In the massive ores, the ilmenite is frequently associated with intermediary intrusions (Tellnes in Norway and Lake Allard in Canada). The concentrates obtained from these massive ores often have high iron contents in the form of segregated hematite or magnetite in the ilmenite. These reduce the TiO_2 content of the concentrates (see Table 13). Direct use of these ilmenites has decreased owing to their high iron content. A digestion process is employed to produce iron sulfate heptahydrate. In cases where iron sulfate is not required as a product, metallurgical recovery of iron from the iron-rich ilmenites and production of a titanium-rich slag are being increasingly used.

Table 13. Composition of ilmenite deposits (wt %)

Component	Tellnes (Norway)	Richard's Bay (Republic of South Africa)	Capel (Western Australia)	Quilon (India)
TiO_2	43.8	46.5	54.8	60.3
Fe_2O_3	14.0	11.4	16.0	24.8
FeO	34.4	34.2	23.8	9.7
Al_2O_3	0.6	1.3	1.0	1.0
SiO_2	2.2	1.6	0.8	1.4
MnO	0.3		1.5	0.4
Cr_2O_3		0.1	0.1	0.1
V_2O_5	0.3	0.3	0.2	0.2
MgO	3.7	0.9	0.15	0.9

The enrichment of ilmenite in beach sand in existing or fossil coastlines is important for TiO_2 production. The action of surf, currents, and/or wind results in concentration of the ilmenite and other heavy minerals such as rutile, zircon, monazite, and other silicates in the dunes or beaches. This concentration process frequently leads to layering of the minerals. Attack by seawater and air over geological periods of time leads to corrosion of the ilmenite. Iron is removed from the ilmenite lattice, resulting in enrichment of the TiO_2 in the remaining material. The lattice is stable with TiO_2 contents up to ca. 65 %, but further removal of iron leads to the formation of a submicroscopic mixture of minerals which may include anatase, rutile, and amorphous phases. Mixtures with TiO_2 contents as high as 90 % are referred to as *leucoxene*. Leucoxene is present in corroded ilmenite and in some deposits is recovered and treated separately. However, the quantities produced are small in comparison to those of ilmenite.

The concentrates obtained from ilmenite sand, being depleted in iron, are generally richer in TiO_2 than those from the massive deposits. Other elements in these concentrates include magnesium, manganese, and vanadium (present in the ilmenite) and aluminum, calcium, chromium, and silicon which originate from mineral intrusions.

Two-thirds of the known ilmenite reserves that could be economically worked are in China, Norway (both massive deposits), and the former Soviet Union (sands and massive deposits). On the basis of current production capacities, these countries could cover all requirements for ca. 150 years. However, the countries with the largest outputs are Australia (sands), Canada (massive ore), and the Republic of South Africa (sands). Other producers are the United States (sands, Florida), India (sands, Quilon), the Former Soviet Union (sands, massive ore), Sri Lanka (sands), and Brazil (rutilo e ilmenita do Brasil). In 1994 the production of ilmenite was ca. 1.2×10^6 t of contained TiO_2.

Rutile is formed primarily by the crystallization of magma with high titanium and low iron contents, or by the metamorphosis of titanium-bearing sediments or magmatites. The rutile concentrations in primary rocks are not workable. Therefore, only sands in which rutile is accompanied by zircon and/or ilmenite and other heavy minerals can be regarded as reserves. The world reserves of rutile are estimated to be 28×10^6 t, including the massive Piampaludo ore reserves in Italy, whose workability is in dispute.

As in the case of ilmenite, the largest producers are in Australia, the Republic of South Africa, and Sierra Leone. There is not enough natural rutile to meet demand, and it is therefore gradually being replaced by the synthetic variety. Compositions of typical rutile concentrates are given in Table 14.

Table 14. Composition of rutile deposits [97]

Rutile component	Content, wt %		
	Eastern Australia	Sierra Leone	Republic of South Africa
TiO_2	96.00	95.70	95.40
Fe_2O_3	0.70	0.90	0.70
Cr_2O_3	0.27	0.23	0.10
MnO	0.02		
Nb_2O_5	0.45	0.21	0.32
V_2O_5	0.50	1.00	0.65
ZrO_2	0.50	0.67	0.46

Al_2O_3	0.15	0.20	0.65
CaO	0.02		0.05
P_2O_5	0.02	0.04	0.02
SiO_2	1.00	0.70	1.75

Anatase, like rutile, is a modification of TiO_2. The largest reserves of this mineral are found in carboniferous intrusions in Brazil. Ore preparation techniques allow production of concentrates containing 80 % TiO_2, with possible further concentration to 90 % TiO_2 by treatment with hydrochloric acid [98]. The TiO_2 content of these mineral deposits is estimated at up to 10^8 t.

Ore Preparation. Most of the world's titanium ore production starts from heavy mineral sands. Figure 14 shows a schematic of the production process. The ilmenite is usually associated with rutile and zircon, so that ilmenite production is linked to the recovery of these minerals. Geological and hydrological conditions permitting, the raw sand (usually containing 3 – 10 % heavy minerals) is obtained by wet dredging (a). After a sieve test (b), the raw sand is subjected to gravity concentration in several stages with Reichert cones (d) and/or spirals (e) to give a product containing 90 – 98 % heavy minerals. This equipment separates the heavy from the light minerals (densities: 4.2 – 4.8 g/cm^3 and < 3 g/cm^3 respectively) [99].

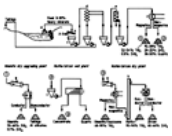

Figure 14. The processing of heavy mineral sands

a) Dredger; b) Sieve; c) Bunker; d) Reichert cones; e) Spirals; f) Magnetic separator; g) Dryer; h) Electrostatic separator; i) Shaking table; j) Dry magnetic separator; k) Vertical belt conveyer; l) Electrostatic separator

The magnetic minerals (ilmenite) are then separated from the nonmagnetic (rutile, zircon, and silicates) by dry or wet magnetic separation (f). If the ores are from unweathered deposits, the magnetite must first be removed. An electrostatic separation stage (h) allows separation of harmful nonconducting mineral impurities such as granite, silicates, and phosphates from the ilmenite, which is a good conductor. The nonmagnetic fraction (leucoxene, rutile, and zircon) then undergoes further hydromechanical processing (i) (shaking table, spirals) to remove the remaining low-density minerals (mostly quartz). Recovery of the weakly magnetic weathered ilmenites and leucoxenes is by high-intensity magnetic separation (j) in a final dry stage. The conducting rutile is then separated from the nonconducting zircon electrostatically in several stages (l). Residual quartz is removed by an air blast.

2.1.2.2. *Synthetic Raw Materials*

Increasing demand for raw materials with high TiO_2 contents has led to the development of synthetic TiO_2 raw materials. In all production processes, iron is removed from ilmenites or titanomagnetites.

Titanium Slag. The metallurgical process for removing iron from ilmenite is based on slag formation in which the iron is reduced by anthracite or coke to metal at 1200 – 1600 °C in an electric arc furnace and then separated. Titanium-free pig iron is produced together with slags containing 70 – 85 % TiO_2 (depending on the ore used) that can be digested with sulfuric acid because they are high in Ti^{3+} and low in carbon. Raw materials of this type are produced in Canada by the Quebec Iron and Titanium Corporation (QIT), in the Republic of South Africa by Richard's Bay Minerals (RBM), and to a smaller extent by Tinfos Titan and Iron K.S. (Tyssedal, Norway). Total slag production in 1994 was 1.4 × 10^6 t of contained TiO_2.

Synthetic Rutile. In contrast to ilmenite, only a small number of rutile deposits can be mined economically, and the price of natural rutile is therefore high. Consequently, many different processes have been developed to remove the iron from ilmenite concentrates without changing the grain size of the mineral because this is highly suitable for the subsequent fluidized-bed chlorination process. All industrial processes involve reduction of Fe^{3+} with carbon or hydrogen, sometimes after preliminary activation of the ilmenite by oxidation. Depending on the reducing conditions, either Fe^{2+} is formed in an activated ilmenite lattice, or metallic iron is produced.

The activated Fe^{2+}-containing ilmenite can be treated with hydrochloric or dilute sulfuric acid (preferably under pressure), and a “synthetic rutile” with a TiO_2 content of 85 – 96 % is obtained [100]. The solutions containing iron(II) salts are concentrated and then thermally decomposed to form iron oxide and the free acid, which can be used again in the digestion process [101].

Metallic iron can be removed in various ways. The following processes are described in the patent literature:

1. Size reduction followed by physical processes such as magnetic separation of flotation.
2. Dissolution in iron(III) chloride solutions [102], the resulting iron(II) salt is oxidized with air to give iron oxide hydroxides and iron(III) salts.
3. Dissolution in acid.
4. Oxidation with air in the presence of electrolytes. Various iron oxide or iron oxide hydroxide phases are formed depending on the electrolyte used. Possible electrolytes include iron(II) chloride solutions [103], ammonium chloride [104], or ammonium carbonate – carbonic acid [105].
5. Oxidation with the iron(III) sulfate from ilmenite digestion (see Crystallization.) [106], followed by crystallization of the iron(II) sulfate.
6. Chlorination to form iron(III) chloride [107].
7. Reaction with carbon monoxide to form iron carbonyls [108] which can be decomposed to give high-purity iron.

Another possible method of increasing the TiO_2 content of ilmenite is by partial chlorination of the iron in the presence of carbon. This is operated on a large scale by several companies [109], [110]. The most important companies producing synthetic rutile are located in Australia (Renison Goldfields Consolidated Tiwest, Westralien Sands), United States (Kerr-McGee Synthetic Rutile), India (Kerala Minerals and Metals, DCW, Bene-Chlor Chemicals), and Malaysia (Hitox). In 1994 production of synthetic rutile was ca. 669 × 10^3 t of contained TiO_2.

2.1.3. ***Production***

Titanium dioxide pigments are produced by two different processes. The older *sulfate process* depends on the breakdown of the titanium-containing raw material ilmenite or titanium slag with concentrated sulfuric acid at 150 – 220 °C. Relatively pure TiO_2 dihydrate is precipitated by hydrolysis of the sulfate solution, which contains colored heavy metal sulfates, sometimes in high concentration. The impurities are largely removed in further purification stages. The hydrate is then calcined, ground, and further treated.

In the *chloride process,* the titanium-containing raw materials ilmenite, leucoxene, natural and synthetic rutile, titanium slag, and anatase are chlorinated at 700 – 1200 °C. Titanium tetrachloride is separated from other chlorides by distillation. Vanadium tetrachloride (VCl_4) and vanadium oxychloride ($VOCl_3$) must, however, first be reduced to solid chlorides. The $TiCl_4$ is burnt at temperatures of 900 – 1400 °C to form TiO_2. This extremely pure pigment undergoes further treatment depending on the type of application.

2.1.3.1. *Sulfate Method*

The sulfate method is summarized in Figure 15.

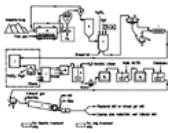

Figure 15. Production of TiO_2 by the sulfate process

a) Ball mill/dryer; b) Screen; c) Magnetic separator; d) Cyclone; e) Silo; f) Digestion vessel; g) Thickener; h) Rotary filter; i) Filter press; j) Crystallizer; k) Centrifuge; l) Vacuum evaporator; m) Preheater; n) Stirred tank for hydrolysis; o) Cooler; p) Moore filters; q) Stirred tank for bleaching; r) Stirred tank for doping; s) Rotary filter for dewatering; t) Rotary kiln; u) Cooler

Grinding. The titanium-bearing raw materials are dried to a moisture content of < 0.1 %. Drying is mainly intended to prevent heating and premature reaction on mixing with sulfuric acid. The raw materials are ground in ball mills to give a mean particle size of ca. 40 µm. The combination of grinding and drying shown in Figure 15 (a) is recommended. The small amount of metallic iron present in titanium slag is removed magnetically (c), which almost completely eliminates hydrogen evolution during subsequent digestion.

Digestion. *Batch digestion* is usually employed. The ground raw materials (ilmenite, titanium slag, or mixtures of the two) are mixed with 80 – 98 % H_2SO_4. The ratio of H_2SO_4 to raw material is chosen so that the weight ratio of free H_2SO_4 to TiO_2 in the suspension produced by the hydrolysis is between 1.8 and 2.2 (the so-called "acid number"). The reaction in the digestion vessel (f) is started by adding water, dilute sulfuric acid, oleum, or sometimes steam. The temperature initially increases to 50 – 70 °C due to the heat of hydration of the acid. The exothermic sulfate formation then increases the temperature to 170 – 220 °C. If dilute acid or sparingly soluble raw materials are used, external heating is required.

After the maximum temperature has been reached, the reaction mixture must be left to mature for 1 – 12 h, depending on the raw material, so that the titanium-containing components become as soluble as possible. Digestion can be accelerated by blowing air through the mass while the temperature is increasing, and also during the maturing period.

Several *continuous digestion processes* have been proposed [112]. A proven method is to continuously feed a mixture of ilmenite and water together with the acid into a double-paddle screw conveyor. After a relatively short dwell time (< 1 h), a crumbly cake is produced [113]. This process utilizes a more limited range of raw materials than the batch process because they need to be very reactive.

Dissolution and Reduction. The cake obtained by digestion is dissolved in cold water or in dilute acid recycled from the process. A low temperature must be maintained (< 85 °C) to avoid premature hydrolysis, especially with the product from ilmenite. Air is blown in to agitate the mixture during dissolution. With the ilmenite product, the TiO_2 concentration of the solution is 8 – 12 wt %, and with the slag product between 13 and 18 wt %.

The trivalent iron is hydrolyzed together with the titanium compounds, and adheres to the titanium oxide hydrate. Therefore, all the Fe^{3+} is reduced to Fe^{2+} by scrap iron during dissolution of the ilmenite product, or immediately afterwards. Reoxidation of the iron during subsequent processing is prevented with Ti^{3+} which is obtained by reducing a small part of the Ti^{4+}. Alternatively, reduction of Ti^{4+} to Ti^{3+} can be carried out in part of the solution under optimized conditions; this concentrated Ti^{3+} solution is then added in a controlled manner to the reaction solution [114]. In solutions obtained from titanium slag, the Ti^{3+} content of the solution must be decreased by oxidation with atmospheric oxygen so that no loss of yield occurs during hydrolysis.

With both ilmenite and titanium slag, mixed digestion can be carried out in which the Ti^{3+} content of the slag reduces all the Fe^{3+} to Fe^{2+}. The dissolved products obtained from the separate digestion of ilmenite and titanium slag can also be mixed [115], [116].

Clarification. All undissolved solid material must be removed as completely as possible from the solution. The most economical method is to employ preliminary settling in a thickener (g), followed by filtration of the sediment with a rotary vacuum filter (h). The filtrate and the supernatant from the thickener are passed through filter presses (i) to remove fines. Owing to the poor filtering properties of the solution, the rotary filter must be operated as a precoat filter. Preliminary separation in the thickener must be assisted by adding chemicals to promote sedimentation. Attempts to carry out the entire clarification process in a single stage using automated filter presses have been reported [117].

Crystallization. The solutions from slag digestion contain 5 – 6 wt % $FeSO_4$, and those from ilmenite digestion 16 – 20 wt % $FeSO_4$ after reduction of the Fe^{3+}. The solution is cooled under vacuum to crystallize out $FeSO_4 \cdot 7\ H_2O$ (j) and reduce the quantity of $FeSO_4$ discharged with the waste acid. The concentration of the TiO_2 in the solution is thereby increased by ca. 25 %. The salt is separated by filtration or centrifugation (k).

The iron sulfate is used in water purification, and as a raw material for the production of iron oxide pigments. Alternatively, it can be dehydrated and thermally decomposed to give iron(III) oxide and sulfur dioxide.

Hydrolysis. Titanium oxide hydrate is precipitated by hydrolysis at 94 – 110 °C. Other sulfuric-acid-soluble components of the raw material are precipitated simultaneously, mainly niobium as its oxide hydrate.

Hydrolysis is carried out in brick-lined, stirred tanks (n) into which steam is passed. The hydrolysate does not have any pigment properties, but these are strongly influenced by the particle size and degree of flocculation of the hydrolysate (mean particle size of hydrolysate is ca. 5 nm, and of TiO_2 pigments 200 – 300 nm).

The properties of the hydrolysate depend on several factors:

1. The hydrolysis of concentrated solutions of titanium sulfate (170 – 230 g TiO_2/L) proceeds very sluggishly and incompletely (even if boiled) unless suitable nuclei are added or formed to accelerate hydrolysis. The nuclei are usually produced by two methods. In the Mecklenburg method, colloidal titanium oxide hydrate is precipitated with sodium hydroxide at 100 °C; 1 % of this hydrate is sufficient. In the Blumenfeld method a small part of the sulfate solution is hydrolyzed in boiling water and then added to the bulk solution [118]. The particle size of the hydrolysate depends on the number of nuclei.
2. The particle size and degree of flocculation of the hydrolysate depend on the intensity of agitation during the nuclei formation by the Blumenfeld method and also during the initial stage of the hydrolysis.
3. The titanium sulfate concentration has a great influence on the flocculation of the hydrolysate. It is adjusted, if necessary by vacuum evaporation, to give a TiO_2 content of 170 – 230 g/L during hydrolysis. Lower concentrations result in a coarser particle size.
4. The acid number should be between 1.8 and 2.2. It has a considerable effect on the TiO_2 yield and on the particle size of the hydrolysate. For a normal hydrolysis period (3 – 6 h) the TiO_2 yield is 93 – 96 %.
5. The properties of the hydrolysate are affected by the concentrations of other salts present, especially $FeSO_4$. High concentrations lead to finely divided hydrolysates.
6. The temperature regime mainly affects the volume – time yield and hence the purity of the hydrolysate.

Purification of the Hydrolysate. After hydrolysis, the liquid phase of the titanium oxide hydrate suspension contains 20 – 28 % H_2SO_4 and various amounts of dissolved sulfates, depending on the raw material. The hydrate is filtered off from the solution (p_1) (weak acid), and washed with water or dilute acid. Even with acid washing, too many heavy metal ions are adsorbed on the hydrate for it to be directly usable in the production of white pigment. Most of the impurities can be removed by reduction (bleaching), whereby the filter cake is slurried with dilute acid (3 – 10 %) at 50 – 90 °C and mixed with zinc or aluminum powder (q). Bleaching can also be carried out with powerful nonmetallic reducing agents (e.g., $HOCH_2 – SO_2Na$). After a second filtration and washing process (p_2), the hydrate only has low concentrations (ppm) of colored impurities but still contains chemisorbed 5 – 10 % H_2SO_4. This cannot be removed by washing and is driven off by heating to a high temperature.

Doping of the Hydrate. When producing titanium dioxide of maximum purity, the hydrate is heated (calcined) without any further additions. This gives a fairly coarse grade of TiO_2 with a rutile content that depends on the heating temperature. However, to produce specific pigment grades, the hydrate must be treated with alkali-metal compounds and phosphoric acid as mineralizers (< 1 %) prior to calcination (r). Anatase pigments contain more phosphoric acid than rutile pigments. To produce rutile pigments, rutile nuclei (< 10 %) must be added; ZnO, Al_2O_3, and/or Sb_2O_3 (< 3 %) are sometimes also added to stabilize the crystal structure.

Nuclei are produced by converting the purified titanium oxide hydrate to sodium titanate, which is washed free of sulfate and then treated with hydrochloric acid to produce the rutile nuclei. Rutile nuclei can also be prepared by precipitation from titanium tetrachloride solutions with sodium hydroxide solution.

Calcination. The doped hydrate is filtered with rotary vacuum filters (s) to remove water until a TiO_2 content of ca. 30 – 40 % is reached. Pressure rotary filters or automatic filter presses can also be used to obtain a TiO_2 content of ca. 50 %. Some of the water-soluble dopants are lost in the filtrate and can be replaced by adding them to the filter cake before it is charged into the kiln. Calcination is performed in rotary kilns (t) directly heated with gas or oil in countercurrent flow. Approximately two-thirds of the residence time (7 – 20 h in total) is needed to dry the material. Above ca. 500 °C, sulfur trioxide is driven off which partially decomposes to sulfur dioxide and oxygen at higher temperatures. The product reaches a maximum temperature of 800 – 1100 °C depending on pigment type, throughput, and temperature profile of the kiln. Rutile content, particle size, size distribution, and aggregate formation are extremely dependent on the operating regime of the kiln. After leaving the kiln, the clinker can be indirectly cooled or directly air-cooled in drum coolers (u).

The exhaust gas must have a temperature of >300 °C at the exit of the kiln to prevent condensation of sulfuric acid in the ducting. Energy can be saved by recirculating some of the gas to the combustion chamber of the kiln and mixing it with the fuel gases as a partial replacement for air. Alternatively, it can be used for concentrating the dilute acid (see Section Problems with Aqueous and Gaseous Waste). The gas then goes to the waste-gas purification system.

Grinding. The agglomerates and aggregates in the clinker can be reduced to pigment fineness by wet or dry grinding. Coarse size reduction should be carried out in hammer mills prior to wet grinding in tube mills (with addition of wetting agents). The coarse fraction can be removed from the suspension by centrifugation, and recycled to the mills. Hammer mills, cross-beater mills, and particularly pendular and steam-jet mills are suitable for dry grinding. Special grinding additives can be used that act as wetting agents during subsequent pigment treatment or improve the dispersibility of untreated pigments.

2.1.3.2. *The Chloride Process*

The chloride process is summarized in Figure 16.

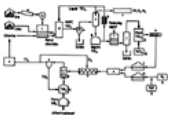

Figure 16. Flow diagram of TiO_2 production by the chloride process

a) Mill; b) Silo; c) Fluidized-bed reactor; d) Cooling tower; e) Separation of metal chlorides; f) $TiCl_4$ condensation; g) Tank; h) Cooler; i) Vanadium reduction; j) Distillation; k) Evaporator; l) $TiCl_4$ superheater; m) O_2 superheater; n) Burner; o) Cooling coil; p) Filter; q) TiO_2 purification; r) Silo; s) Gas purification; t) Waste-gas cleaning; u) Cl_2 liquefaction

Chlorination. The titanium in the raw material is converted to titanium tetrachloride in a reducing atmosphere. Calcined petroleum coke is used as the reducing agent because it has an extremely low ash content and, due to its low volatiles content, very little HCl is formed. The titanium dioxide reacts exothermically as follows:

$$TiO_2 + 2\,Cl_2 + C \rightarrow TiCl_4 + CO_2$$

As the temperature rises, an endothermic reaction also occurs to an increasing extent in which carbon monoxide is formed from the carbon dioxide and carbon. Therefore, oxygen must be blown in with the chlorine to maintain the reaction temperature between 800 and 1200 °C. The coke consumption per tonne of TiO_2 is 250 – 300 kg. If CO_2-containing chlorine from the combustion of $TiCl_4$ is used, the coke consumption increases to 350 – 450 kg.

The older *fixed-bed chlorination method* is hardly used today. In this process, the ground titanium-containing raw material is mixed with petroleum coke and a binder, and formed into briquettes. Chlorination is carried out at 700 – 900 °C in brick-lined reactors.

Fluidized-bed chlorination was started in 1950. The titanium raw material (with a particle size similar to that of sand) and petroleum coke (with a mean particle size ca. five times that of the TiO_2) are reacted with chlorine and oxygen in a brick-lined fluidized-bed reactor (Fig. 16, c) at 800 – 1200 °C. The raw materials must be as dry as possible to avoid HCl formation. Since the only losses are those due to dust entrainment the chlorine is 98 – 100 % reacted, and the titanium in the raw material is 95 – 100 % reacted, depending on the reactor design and the gas velocity. Magnesium chloride and calcium chloride can accumulate in the fluidized-bed reactor due to their low volatility. Zirconium silicate also accumulates because it is chlorinated only very slowly at the temperatures used. All the other constituents of the raw materials are volatilized as chlorides in the reaction gases.

The ceramic cladding of the fluidized-bed reactor is rather rapidly destroyed by abrasion and corrosion. If chlorination is interrupted, there is a further danger that the raw materials may sinter and eventually cannot be fluidized.

Gas Cooling. The reaction gases are cooled with liquid $TiCl_4$ either indirectly or directly (d). Crystallization of the chlorides of the other components causes problems because they tend to build up on the cooling surfaces, especially the large quantities of iron(II) and iron(III) chlorides formed on chlorination of ilmenite [119]. In this first stage, the reaction gases are cooled only down to a temperature (< 300 °C) at which the accompanying chlorides can be satisfactorily separated from the $TiCl_4$ by condensation or sublimation (e).

The gas then consists mainly of $TiCl_4$, and is cooled below 0 °C, causing most of the $TiCl_4$ to condense (f). The small amounts of $TiCl_4$ and Cl_2 remaining in the exhaust gas (CO_2, CO, and N_2) are removed by scrubbing with alkali (t).

Purification of $TiCl_4$. The chlorides that are solid at room temperature and the entrained dust can be separated from the $TiCl_4$ by simply evaporating (distilling) this off (j). Dissolved chlorine can be removed by heating or reduction with metal powders (Fe, Cu, or Sn).

Removal of vanadium tetrachloride (VCl_4) and vanadium oxychloride ($VOCl_3$) from the $TiCl_4$ by distillation is very difficult owing to the closeness of their boiling points. They are therefore reduced to form solid, low-value vanadium chlorides (i). An enormous number of reducing agents have been recommended; important examples are copper, titanium trichloride, hydrogen sulfide, hydrocarbons, soaps, fatty acids, and amines. After subsequent evaporation (j) the titanium chloride should contain < 5 ppm vanadium. If organic reducing agents are used, the residues may cause problems by baking onto the surfaces of the heat exchanger.

Phosgene and $SiCl_4$ can be removed by fractional distillation.

Combustion of $TiCl_4$ and Recovery of TiO_2. Titanium tetrachloride is combusted with oxygen at 900 – 1400 °C to form TiO_2 pigment and chlorine (n). The purified $TiCl_4$ is vaporized (k) and the vapor is indirectly heated to ca. 500 – 1000 °C (l). The reaction

$$TiCl_4 + O_2 \rightarrow 2\,Cl_2 + TiO_2$$

is weakly exothermic, and requires a high reaction temperature, so that the oxygen must also be heated to>1000 °C (m). This can be achieved with an electric plasma flame, by reacting part of the oxygen with carbon monoxide, or by indirect heating. Hot $TiCl_4$ and oxygen (110 – 150 % of the stoichiometric amount) are fed separately into a reaction chamber where they must be mixed as rapidly and completely as possible to give a high reaction rate. For this reason, and also because the TiO_2 has a strong tendency to cake onto the walls [120-122], many different reactor designs have been proposed and used. The same considerations apply to the cooling unit (o) where the pigment is very rapidly cooled to below 600 °C. Cooling zones of various geometries are used. If caking occurs, the material can be removed by introducing abrasive particles [123], [124].

The mixture of gases (Cl_2, O_2, CO_2) and pigment can be further cooled during dry separation of the pigment either indirectly or directly by solid particles, e.g., sand. The pigment-containing gas is then filtered (p). The gas stream is recycled to the cooling zone (o) of the combustion furnace and to the chlorination process as oxygen-containing chlorine via the liquefaction unit (u). The chlorine adsorbed on the pigment can be removed by heating or by flushing with nitrogen or air.

The wet separation process, in which the pigment-containing gas mixture (Cl_2, O_2, and CO_2) is quenched in water, has not become established.

2.1.3.3. *Pigment Quality*

The quality of the TiO_2 pigment is influenced by various factors. Reaction temperature, excess oxygen, and flow conditions in the reactor affect particle size and size distribution. Therefore, optimum conditions must be established for every reactor design. Caking of the TiO_2 on the walls of the reactor leads to impairment of quality.

The presence of water and a cesium compound during combustion of the $TiCl_4$ gives rise to nuclei which promote the formation of finely divided pigment particles with high scattering power [125]. It can be added directly to the oxygen or can be produced by the combustion of hydrogen-containing materials.

The presence of $AlCl_3$ promotes the formation of rutile and a more finely divided pigment. It is added in amounts of up to 5 mol %. Many methods have been proposed for rapidly generating and directly introducing the $AlCl_3$ vapor into the $TiCl_4$ vapor. Addition of PCl_3 and $SiCl_4$ suppresses rutile formation, so that anatase pigment is obtained [126]. However, pigments of this type have not appeared on the market.

Pigments produced by the chloride process (chloride pigments) have better lightness and a more neutral hue than pigments produced by the sulfate process (sulfate pigments). Pigments used in demanding applications are almost always subjected to inorganic aftertreatment.

2.1.3.4. *Aftertreatment*

Aftertreatment of the pigment particles improves the weather resistance and lightfastness of the pigmented organic matrix, and dispersibility in this matrix. The treatment consists of coating the individual pigment particles with colorless inorganic compounds of low solubility by precipitating them onto the surface. However, this reduces the optical performance of the pigment approximately in proportion to the decrease in the TiO_2 content. The surface coatings prevent direct contact between the binder matrix and the reactive surface of the TiO_2. The effectiveness of these coatings largely depends on their composition and method of application, which may give too porous or too dense a coating. The treatment process also affects the dispersibility of the pigment, and therefore a compromise often has to be made. High weather resistance and good dispersibility of the pigment in the binder or matrix are usually desired. These effects are controlled by using different coating densities and porosities. Other organic substances can be added during the final milling of the dried pigment.

Several types of treatment are used:

1. Deposition from the gas phase by hydrolysis or decomposition of volatile substances such as chlorides or organometallic compounds. Precipitation onto the pigment surface is brought about by adding water vapor. This method is especially applicable to chloride pigments, which are formed under dry conditions.
2. Addition of oxides, hydroxides, or substances that can be adsorbed onto the surface during pigment grinding. This can produce partial coating of the pigment surface.

3. Precipitation of the coating from aqueous solutions onto the suspended TiO_2 particles. Batch processes in stirred tanks are preferred; various compounds are deposited one after the other under optimum conditions. There is a very extensive patent literature on this subject. Continuous precipitation is sometimes used in mixing lines or cascades of stirred tanks. Coatings of widely differing compounds are produced in a variety of sequences. The most common are oxides, oxide hydrates, silicates, and/or phosphates of titanium, zirconium, silicon, and aluminum. For special applications, boron, tin, zinc, cerium, manganese, antimony, or vanadium compounds can be used [127], [128].

Three groups of pigments have very good lightfastness or weather resistance:

1. Pigments with dense surface coatings for paints or plastics formed by:
   a. Homogeneous precipitation of SiO_2 with precise control of temperature, pH, and precipitation rate [129]: ca. 88 % TiO_2
   b. Two complete aftertreatments, calcination is performed at 500 – 800 °C after the first or second aftertreatment [130]: ca. 91 % TiO_2
   c. Aftertreatment with Zr, Ti, Al, and Si compounds, sometimes followed by calcination at 700 – 800 °C [131]: ca. 95 % TiO_2
2. Pigments with porous coatings for use in emulsion paints obtained by simple treatment with Ti, Al, and Si compounds, giving a silica content of 10 % and a TiO_2 content of 80 – 85 %
3. Lightfast pigments with dense surface coatings for the paper industry that have a stabilized lattice and a surface coating based on silicates or phosphates of titanium, zirconium, and aluminum: ca. 90 % TiO_2

Coprecipitation of special cations such as antimony or cerium can improve lightfastness further [132]. After treatment in aqueous media, the pigments are washed on a rotary vacuum filter or filter press until they are free of salt, and then dried in, e.g., belt, spray, or fluidized-bed dryers.

Before micronizing the pigment in air-jet or steam-jet mills, and sometimes also before drying, the pigment surface is improved by adding substances to improve dispersibility and facilitate further processing. The choice of compounds used, which are mostly organic, depends on the intended use of the pigment. The final surface can be made either hydrophobic (e.g., using silicones, organophosphates, and alkyl phthalates) or hydrophilic (e.g., using alcohols, esters, ethers and their polymers, amines, organic acids). Combinations of hydrophobic and hydrophilic substances have proved especially useful for obtaining surface properties that give better dispersibility and longer shelf life [133].

2.1.3.5. *Problems with Aqueous and Gaseous Waste*

Aqueous Waste. In the *sulfate process*, 2.4 – 3.5 t concentrated H_2SO_4 are used per tonne of TiO_2 produced, depending on the raw material. During processing, some of this sulfuric acid is converted to sulfate, primarily iron(II) sulfate, the rest is obtained as free sulfuric acid (weak acid). Filtration of the hydrolysate suspension can be carried out to give 70 – 95 wt % of the SO_4^{2-} in a weak acid fraction containing ca. 20 – 25 % free acid, the remaining sulfate (5 – 30 %) is highly diluted with wash-water.

In the past it was common practice to discharge acid directly into the open sea or coastal waters. For a long time the weak-acid problem was the subject of controversy. As a result the European Community decided to stop the discharge of weak acid into open waters by 1993.

The European TiO_2 producers developed various effluent-treatment processes to meet the environmental requirements [134]. The most important processes are the precipitation of gypsum from the weak acid [135] and the concentration and recovery of the free and bound acid.

In the gypsum process, the acid effluent is treated in a first stage with finely divided $CaCO_3$ to precipitate white gypsum. After filtering off, washing, and drying, the white gypsum is used for the manufacturing of plasterboard. In a second stage, the residual metal sulfates in the filtrate are precipitated as metal hydroxides and further gypsum by adding calcium hydroxide. This mixture, the so-called red gypsum, can be used for landfill.

In the recycling process, both the free and the bound sulfuric acid (as metal sulfates) can be recovered from the weak acid in the calcination furnace (k, Fig. 17) and in metal sulfate calcination (Fig. 18). The process consists of two stages:

1. Concentration and recovery of the free acid by evaporation
2. Thermal decomposition of the metal sulfates and production of sulfuric acid from the resulting sulfur dioxide

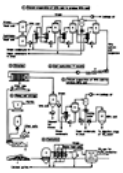

Figure 17. Weak acid recovery plant used by Sachtleben Chemie (based on know-how of Bayer AG)

a) Heat exchanger; b) Evaporator; c) Injection condenser; d) Stirred salt maturing vessels; e) Filter press; f) Bunker for pyrites; g) Coal silo; h) Bunker; i) Mixing screw unit; j) Covered store for mixed filter cake; k) Calcination furnace; l) Waste-heat boiler; m) Cyclone; n) Electrostatic precipitator; o) Stirred tank; p) Storage tank; q) Pump; r) Cooler

Figure 18. Waste heat recovery and sulfuric acid recycling during weak acid treatment (Bayer AG)

As a result of energy requirements only acid containing > 20 % H_2SO_4 can be economically recovered by evaporation. The weak acid is concentrated from ca. 20 – 25 % to ca. 28 % with minimum heat (i.e., energy) consumption, e.g., by using waste heat from sulfuric acid produced by the contact process [136], or from the waste gases from the calcination kilns used in TiO_2 production [137] (Fig. 18).

Following preliminary evaporation, further concentration is carried out in multi-effect vacuum evaporators. Since the water vapor pressure decreases strongly as the H_2SO_4 concentration increases, in general only two-stage evaporation can effectively exploit the water vapor as a heating medium. Evaporation produces a suspension of metal sulfates in 60 – 70 % sulfuric acid (stage 1 in Fig. 17). The suspension is cooled to 40 – 60 °C in a series of stirred tanks (stage 2, d) [138], giving a product with good filtering properties and an acid of suitable quality for recycling to the digestion process. Filtration (stage 3, e) is usually carried out with pressure filters [139] because they give a filter cake with an extremely low residual liquid content.

The concentration of the acid recycled to the digestion process depends on the quality of the titanium-containing raw material. For raw materials with a high titanium content, the 65 – 70 % sulfuric acid separated from the metal sulfates must be further concentrated to give 80 – 87 % acid (stage 5).

Concentration can be carried out in steam-heated vacuum evaporators, or by using the heat from the TiO_2 calcination kilns [140]. Cooling the acid obtained after this concentration process yields a suspension of metal sulfates that can be directly used for digestion of the raw material. The metal sulfates recovered from the sulfuric acid in stage 3 are moist because they contain 65 – 70 % sulfuric acid; they therefore have no direct use. They can be converted to a disposable material by reaction with calcium compounds [141]. Thermal decomposition of the metal sulfates to form the metal oxides, sulfur dioxide, water, and oxygen is energy intensive, but is advantageous from the ecological point of view. The energy requirement is ca. 4×10^9 J per tonne of filter cake. Thermal decomposition is carried out at 850 – 1100 °C in a fluidized-bed furnace (stage 6). The energy is supplied by coal, pyrites, or sulfur. The sulfur dioxide produced by the thermal decomposition is purified by the usual methods, dried, and converted into sulfuric acid or oleum. This pure acid or oleum is mixed with the recovered sulfuric acid and used in the digestion process.

The metal oxides produced by thermal decomposition contain all the elements initially present in the raw material apart from the titanium which has been converted into pigment. The mixture of metal oxides, mainly iron oxide, can be used as an additive in the building materials industry.

The continually increasing demand for environmentally friendly industrial processes has also led to the development of techniques for recycling the 5 – 30 % sulfate remaining in the the acidic wash water. [143]. In modern processes, up to 99 % of the sulfuric acid can be recovered and reused in production. In the *chloride process*, wastewater problems arise if the raw material contains < 90 % TiO_2. The metal chloride byproducts are sometimes disposed of in solution by the “deep well” method (e.g., at Du Pont). The metal chloride solutions are pumped via deep boreholes into porous geological strata. Special geological formations are necessary to avoid contamination of the groundwater by impurities.

Increasing restrictions also apply to the chloride process, so that efforts are continually being made to use the iron chloride byproduct, e.g., in water treatment and as a

flocculation agent. [144]. A process for treating metal chlorides with cement and alkaline compounds to produce rocklike aggregates for road building is described in [145].

Waste Gas Problems. The gases produced in the calcination kiln are cooled in a heat exchanger, and entrained pigment is removed, washed, and recycled to the process. The SO_2 and SO_3 formed during calcination are then scrubbed from the gases to form dilute sulfuric acid which is recycled.

2.1.4. *Economic Aspects*

The burning of $TiCl_4$ with oxygen or the calcination of TiO_2 hydrolysates produces either anatase or rutile pigments, depending on the doping and lattice stabilization. They are marketed directly or after being coated with oxides or hydroxides of various elements. Different treatments are necessary depending on the field of application, and all major pigment producers have a large number of pigment grades. Product groups are listed in Table 15. Pigments of all grades are available with or without organic treatment. Over 400 different TiO_2 pigment grades are on the market. Table 16 gives the capacities and processes of the most important pigment producers.

Table 15. Classification of TiO_2 pigments according to composition (DIN E 55 912, sheet 1, issue 07.85; ISO 591, 1977)

Pigment	Class	TiO_2 (min.), wt %	Water-soluble salts, wt %	Volatiles (max.), wt %
Anatase (Type A)	A 1	98	0.6	0.5
	A 2	92	0.5	0.8
Rutile (Type R)	R 1	97	0.6	0.5
	R 2	90	0.5	1.5
	R 3	80	*	*

* To be agreed.

Table 16. World TiO_2 pigment producers (1995)

Country	Company	Location	Capacity, 10^3 t/a		
			Chloride	Sulfate	Total
United States	Du Pont	New Johnsonville	305		305
		De Lisle	245		245
		Edge Moore	127		127
		Antioch	38		38
	SCM Corporation	Baltimore	50	66	116
		Ashtabula	165		165
	Kemira Oy	Savannah	91	54	145
	Kerr McGee Corporation	Hamilton	130		130
	Tioxide	Lake Charles	50		50
	NL Chemicals Incorporated	Lake Charles	50		50
Canada	NL Chemicals Incorporated	Varennes	50	36	86
Brazil	Titanio do Brasil/Bayer	Salvador		55	55
Mexico	Du Pont	Tampico	80		80
Total, America					*1592*
Germany	Kronos Titan	Leverkusen	80	35	115
	NL Chemicals	Nordenham		60	60
	Bayer	Krefeld – Uerdingen		105	105
	Sachtleben Chemie, Metallgesellschaft	Duisburg – Homberg		80	80
United Kingdom	Tioxide UK	Grimsby		100	100
		Greatham	80		80
	SCM Chemicals	Stallingborough	109	10	119
France	Thann et Mulhouse (Rhône-Poulenc)	Le Havre		95	95
		Thann		30	30
	Tioxide France	Calais		100	100
Finland	Kemira Oy	Pori		90	90
Italy	Tioxide Italia	Scarlino		80	80
Belgium	Kronos Titan	Langebrügge	50		50
	Bayer	Antwerpen		33	33
Spain	Tioxide Espan.	Huelva		80	80
Netherlands	Kemira	Botlek	45		45
Norway	Kronos Titan	Fredrikstad		32	32
Total, Western Europe					*1294*
Russia	Lakokraska	Chelyabinsk		10	10
Ukraine	Lakokraska	Amyansk		80	80
	Argokhim	Sumy		40	40
Poland	Zachem	Police		36	36
Yugoslavia	Cinkarna	Celje		25	25
Czech Republic	Precheza	Prerov		25	25
Total, Eastern Europe					*216*
Republic of South Africa	SA Tioxide	Umbogintwini		38	38
Total Africa					*38*
Australia	SCM Chemical	Kemerton	79		79
	Tioxide Australia	Burnie		35	35
	Kerr McGee Corporation	Kwinana	64		64
Total, Australia					*178*
Japan	Ishihara Sangyo Kaisha	Yokkaichi	55	100	155
	Tayca	Saidaiji		60	60
	Sakai Chemical	Onahama		43	43
	Furukawa Mining	Osaka		23	23
	Fuji Titanium	Kobe		16	16
	Titan Kogyo	Ube		17	17
	Tohoku Chemical	Akita		30	30
Total, Japan					*344*

India	Kerala Minerals & Metals	Kerala	22		22
	Travancore Titanium	Trivandrum		13	13
South Korea	Hankook Titanium	Incheon		36	36
Taiwan	China Metal & Chemicals	Chin Shin		10	10
	Du Pont	Kuan Yin	60		60
Malaysia	Tioxide	Terengganu		50	50
Saudi Arabia	Christal	Yanbu	55		55
Total, Far East (excluding Japan)					*246*
Total world capacity			*2080*	*1828*	*3908*

A considerable increase of capacity is planned by debottlenecking and building new plants. Pigment plant capacity is forecast to grow to 5.2×10^6 t in 2005 [146].

Powdered TiO_2 pigments are usually supplied in 25 kg sacks (50 lbs, USA) or in big bags containing 0.5 – 1 t pigment. Aqueous suspensions with solids contents of 68 – 75 % are also available and have great advantages as regards the distribution and metering of the pigment in aqueous systems. The dust formation that occurs with dry pigment is also avoided. With the development of products with improved flow properties and modern pneumatic delivery technology, supply in silo wagons is becoming increasingly important.

2.1.5. *Pigment Properties*

The pigment properties are extremely important when TiO_2 is used as a white pigment; they include lightening power, hiding power, lightness, hue, gloss formation, gloss haze, dispersibility, lightfastness, and weather resistance. These properties are a function of chemical purity, lattice stabilization, particle size and size distribution, and the coating produced by aftertreatment. They also depend on the medium and cannot generally be accurately described in scientific terms. Some of the important properties of TiO_2 pigments are described below.

Scattering Power. The refractive indices of rutile and anatase are very high (2.70 and 2.55, respectively). Even after incorporation in a wide range of binders, they lie in the range between 1.33 (water) and 1.73 (polyester fibers). The scattering power depends on the particle size, and for TiO_2 is at its maximum at a particle size of 0.2 µm (Mie's theory) [28]. The scattering power also depends on the wavelength; TiO_2 pigment particles with a size ≤0.2 µm scatter light of shorter wavelengths more strongly and therefore show a slight blue tinge, while larger particles have a yellow tone.

Hue. The whiteness (lightness and hue) of TiO_2 pigments depends primarily on the crystalline modification, the purity, and the particle size of the TiO_2 (see above). As the absorption band (385 nm) of anatase pigments is shifted into the UV region, compared with rutile pigments they have less yellow undertone. Any transition elements present in the crystal structure have an adverse effect on the whiteness, so manufacturing conditions are of the greatest importance. Thus, pigments produced by the chloride process (which includes distillative purification of $TiCl_4$ before the combustion stage) have a higher color purity and very high lightness values.

Dispersion. Good disintegration and dispersion of the TiO_2 pigments in the medium are necessary to obtain high gloss and low gloss haze. These requirements are satisfied by intensive grinding and by coating the pigment surface with organic compounds. The compounds used for this surface treatment depend on the field of application (see Section Aftertreatment).

Lightfastness and Weather Resistance. Weathering of paints and coatings containing TiO_2 leads to pigment chalking [147]. If weathering occurs in the absence of oxygen, or in binders with low permeability to oxygen (e.g., in melamine – formaldehyde resins), no chalking is observed, but graying takes place, which decreases on exposure to air. Graying is greatly reduced in the absence of water. Both effects are more severe with anatase pigments. Empirical stabilization processes have been developed by pigment producers, e.g., doping with zinc or aluminum prior to calcination.

According to modern theories, impairment of the lightfastness and weather resistance of TiO_2 pigments proceeds according to the following cycle [148]:

1. Molecules of water are bound to the TiO_2 surface, forming hydroxyl groups on the surface.
2. Absorption of light of short wavelength (anatase < 385 nm, rutile < 415 nm) occurs, producing an electron and an electron defect or “hole” (exciton) in the crystal lattice which migrate to the surface of the pigment.
3. At the surface of the pigment, an OH^- ion is oxidized to an $OH^\cdot$ radical by an electron “hole”. The OH^- ion is then desorbed and can oxidatively break down the binder. A Ti^{3+} ion is simultaneously produced by reduction of Ti^{4+} with the remaining electron of the exciton.
4. The Ti^{3+} ion can be oxidized by adsorbed oxygen with formation of an O_2^- ion. The latter reacts with H^+ and is converted into an $HO_2^\cdot$ radical.
5. The cycle ends with the binding of water to the regenerated TiO_2 surface.

The chalking process can be regarded as the reaction of water and oxygen to form $OH^\cdot$ and $HO_2^\cdot$ radicals under the influence of shortwave radiation and the catalytic activity of the TiO_2 surface:

$$H_2O + O_2 \xrightarrow[TiO_2]{h\nu} OH^\cdot + HO_2^\cdot$$

The enthalpy requirement for this reaction (312 kJ/mol) is provided by radiation of wavelength 385 nm. The cycle (1) – (5) is broken by excluding air or water. If oxygen is excluded or a binder is chosen in which the diffusion of oxygen is rate determining, a concentration of Ti^{3+} ions builds up. Graying then takes place, but this decreases with gradual exposure to oxygen. If water is excluded, rehydration and formation of surface hydroxyl groups do not take place; breakdown of the binder therefore ceases. Despite this photochemical breakdown of the binder, treated rutile pigments are used to stabilize many binders. This is because nonpigmented coatings are degraded by exposure to light and weathering; the added TiO_2 pigments prevent light from penetrating the deeper layers of the coating film and thus inhibit breakdown of the binder. High-quality TiO_2 pigments must satisfy stringent requirements with respect to weather resistance. They must withstand the severe climatic conditions of the Florida test, resisting a two-year exposure without appreciable chalking or deterioration of gloss.

2.1.6. *Analysis*

The crystal structure of the pigments is determined by X-ray analysis which is sensitive enough to determine 0.3 – 0.5 % anatase in the presence of 99.7 – 99.5 % rutile. For standards, see Table 1 (Titanium dioxide pigments; “Methods of analysis” and “Specification”).

A qualitative test for TiO_2 is a blue – violet coloration of beads of microcosmic salt ($NaNH_4HPO_4 \cdot 4\ H_2O$), or a yellow – orange coloration produced when hydrogen peroxide is added to a test solution in hot, concentrated sulfuric acid containing ammonium sulfate. For quantitative determination, the pigment is dissolved or digested in sulfuric acid and the solution is reduced to Ti^{3+} with cadmium, zinc, or aluminum. The Ti^{3+} ions are then usually titrated with a standard solution of iron(III) ammonium sulfate solution, with potassium rhodanide as an indicator, or using potentiometric end point determination.

Impurities can be determined by wet analysis, X-ray fluorescence, or spectrographic analysis (e.g., atomic absorption).

Typical analysis figures for an untreated rutile pigment are TiO_2 99.4 %, K_2O 0.24 %, P_2O_5 0.21 %, Fe_2O_3 40 ppm, Sb_2O_3 24 ppm, Al_2O_3 20 ppm, Mg 5 ppm, Zn 3 ppm, Cr 2 ppm, Mn, Cu, Hg, Cd, Co, Ni, Se, Sn, Ag < 1 ppm.

2.1.7. *Uses of Pigmentary TiO_2*

Titanium dioxide is used universally, having almost completely replaced other white pigments. Consumption figures for 1996 are given in Table 17 [149]. The greatest annual increase in use has been for coloring plastics (> 5.5 %), followed by the coloring of paper (ca. 3.0 %). Geographically, the increase in consumption of TiO_2 has been the greatest in Asia (see Table 18).

Table 17. World consumption of TiO_2 pigments in 1996 [149]

Use	Consumption, 10^3 t	Percentage

Coatings	1988	59
Paper	424	13
Plastics	686	20
Others	286	8
Total	3384	100

Table 18. Predicted percentage annual growth rates for use of TiO_2 (1993 – 2000) [149]

End use	United States	Middle and Eastern Europe, Africa	Asia and Pacific	World total
Coatings	3.4	2.0	5.0	2.5
Paper	2.0	4.0	4.0	3.0
Plastics	4.5	5.0	10.0	5.5
Total	3.0	2.5	6.5	3.3

Paints and coatings amount for the largest volume of TiO_2 production. The presence of the pigment enables the protective potential of the coating material to be fully exploited. As a result of continuing developments in TiO_2 pigments, coatings only a few micrometers thick fully cover the substrate. Commercially available pigments permit paint manufacture with simple dispersion equipment, such as disk dissolvers. Organic treatment (see Section Aftertreatment) prior to steam jet micronization yields pigments with improved gloss properties and reduced gloss haze for use in stoving enamels. Sedimentation does not occur when these products are stored, and they possess good lightfastness and weather resistance.

Printing Inks. Modern printing processes operate at coating thicknesses of < 10 µm, and therefore require the finest possible TiO_2 pigments. These very low film thicknesses are only possible with TiO_2 pigments that have a lightening (reducing) power seven times that of lithopone. Because of its neutral hue, TiO_2 is especially suitable for lightening (reducing) colored pigments.

Plastics. Titanium dioxide is used to color durable and nondurable goods such as toys, appliances, automobiles, furniture, and plastic packaging films. Furthermore, TiO_2 pigments absorb UV radiation with a wavelength < 415 nm and thus protect the pigmented goods from these harmful rays.

Fibers. Titanium dioxide pigments give a matt appearance to synthetic fibers, eliminating the greasy appearance caused by their translucent properties. Anatase pigments are used for this because their abrasive effect on the spinning operation is about one quarter that of the rutile pigments. The poor lightfastness of anatase pigments in polyamide fibers can be improved by treatment with manganese or vanadium phosphate.

Paper. In Europe, fillers such as kaolin, chalk, or talc are preferred as brightening agents and opacifiers in paper manufacture. Titanium dioxide pigments are suitable for very white paper that has to be opaque even when very thin (air mail or thin printing paper). The TiO_2 can be incorporated into the body of the paper or applied as a coating to give a superior quality ("art" paper).

Laminated papers are usually colored with extremely lightfast rutile pigments before being impregnated with melamine – urea resin for use as decorative layers or films.

Other areas of application for TiO_2 pigments include the enamel and ceramic industries, the manufacture of white cement, and the coloring of rubber and linoleum.

Titanium dioxide pigments are also used as UV absorbers in sunscreen products, soaps, cosmetic powders, creams, toothpaste, cigar wrappers, and in the cosmetics industry. Their most important properties are their lack of toxicity, compatibility with skin and mucous membranes, and good dispersibility in organic and inorganic solutions and binders.

Electrically conducting TiO_2 pigments have been produced by an aftertreatment to give a coating of mixed oxides of indium and tin, or antimony and tin [150]. These pigments are applied to fibers used in photosensitive papers for electrophotography, and for the production of antistatic plastics.

2.1.8. *Uses of Nonpigmentary TiO_2*

A number of industrial products require TiO_2 starting materials with well-defined properties for a specific application. Some of the most important of these grades of titanium dioxide are those with a high specific surface area, a small particle size, and very high reactivity. Stringent requirements often exist regarding purity and property consistency. The most important applications for nonpigmentary TiO_2 are vitreous enamels, glass and glass ceramics, electroceramics, catalysts and catalyst supports, welding fluxes, colored pigments, electrical conductors, chemical intermediates (e.g., potassium fluorotitanate), structural ceramics, UV absorbers, and refractory coatings.

Electroceramics. Titanates, for example, barium, strontium, calcium, and lead titanate, prepared from finely divided, high-purity TiO_2 hydrolysates, are used in capacitors, PTC resistors, and piezoelectric materials. The specifications of the TiO_2 starting materials with respect to purity, reactivity, and sintering properties are expected to become more stringent. The market is estimated to be several thousands of tonnes per annum as TiO_2. Strong annual growth is expected.

Catalysts. The most important application for TiO_2 is the removal of nitrogen oxides from waste gases from power stations and industry. The nitrogen oxides in the waste gas react with ammonia in the presence of oxygen over the catalyst to produce nitrogen and water (SCR, Selective Catalytic Reduction) [153]. The world market for SCR catalysts is believed to be several thousand tonnes per annum as TiO_2. In addition to TiO_2 the catalysts usually contain ca. 10 wt % tungsten oxide and 1 wt % V_2O_5 and are extruded into a honeycomb shape or supported in thin layers on metal sheets. The TiO_2 has high specifications with respect to purity, particle size, and porosity to ensure that the desired catalytic activity is obtained. An increasing demand is expected for the removal of nitrogen oxides from waste gases from stationary and nonstationary diesel engines. Another application is the combined removal of nitrogen oxides and dioxins from the waste gases of refuse incinerators [154].

Mixed Metal Oxide Pigments. The starting material is a TiO_2 hydrolysate, which is calcined with oxides of transition metals to form chromium rutile or nickel rutile pigments (see Section Properties). Depending on the particle size, pigments with various shades of yellow are produced. These pigments should have considerable growth potential on account of their quality and environmental safety.

UV Absorption. Nanostructured TiO_2 particles (particle size 5 – 50 nm) are used as sunscreens in the cosmetic industry. Nanosized TiO_2 is an effective absorber of UV-B (280 – 320 nm) and UV-A radiation (320 – 400 nm). Because of its small particle size, it appears transparent [155]. Intensive research work is in progress worldwide aimed at utilizing the photoactivity of TiO_2. Titanium dioxide catalyzes the decomposition of organic compounds in wastewater [156]; water is decomposed into hydrogen and oxygen in the presence of sunlight.

2.1.9. *Toxicology*

Titanium dioxide is highly stable and is regarded as completely nontoxic. Investigations on animals which have been fed TiO_2 over a long period give no indication of titanium uptake [157]. Absorption of finely divided TiO_2 pigments in the lungs does not have any carcinogenic effect [158].

2.2. Zinc Sulfide Pigments [159-161]

White pigments based on zinc sulfide were first developed and patented in 1850 in France. Although they are still of economic importance, they have continually lost market volume since the early 1950s when titanium dioxide was introduced. Only one modern production installation for zinc sulfide pigments still exists in the market-economy-oriented industrial nations (Sachtleben Chemie, Germany). There are other production plants in eastern Europe and in China.

The zinc-sulfide-containing white pigment with the largest sales volume is *lithopone* [1345-05-7], which is produced by coprecipitation and subsequent calcination of a mixture of zinc sulfide [1314-98-3], ZnS, M_r 97.43, and barium sulfate [7727-43-7], $BaSO_4$, M_r 233.40. For standards, see Table 1 (Lithopone pigments; "Specification"). Pure zinc sulfide is marketed as *Sachtolith.*

White zinc sulfide pigments maintain their market position in areas of use where not only their good light scattering ability but other properties such as low abrasion, low oil number, and low Mohs hardness are required.

They are often produced from many types of industrial waste. This recycling relieves pressure on the environment, as these materials would otherwise have to be disposed of.

2.2.1. *Properties*

Some physical and chemical properties of ZnS and $BaSO_4$ are given in Table 19.

Table 19. Properties of the components of zinc sulfide pigments

Property	Zinc sulfide	Barium sulfate *
Physical properties		
Refractive index *n*	2.37	1.64
Density, g/cm^3	4.08	4.48
Mohs hardness	3	3.5
Solubility in water (18 °C), wt %	1.8×10^{-4}	2.5×10^{-4}
Chemical properties		
Resistance to acids/bases	soluble in strong acids	insoluble
Resistance to organic solvents	insoluble	insoluble

* Component of lithopone.

A white pigment must not absorb light in the visible region (wavelength 400 – 800 nm), but should disperse incident radiation in this region as completely as possible. The spectral reflectance curves of zinc sulfide and barium sulfate (Fig. 19) fulfil these conditions to a large extent. The absorption maximum for ZnS at ca. 700 nm is a result of lattice stabilization with cobalt ions, whose function is explained in Section Lithopone.. The absorption edge in the UV-A region is responsible for the bluish-white tinge of zinc sulfide. Depending on the production process, zinc sulfide has a sphalerite or wurtzite lattice type.

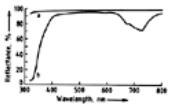

Figure 19. Spectral reflectance curves of barium sulfate and zinc sulfide

a) Barium sulfate; b) Zinc sulfide (Co-doped)

The refractive index *n* of ZnS, which determines its scattering properties, is 2.37 and is much greater than that of plastics and binders (*n* = 1.5 – 1.6). Spheroidal ZnS particles have their maximum scattering power at a diameter of 294 nm. Barium sulfate does not directly contribute to the light scattering due to its relatively low refractive index (*n* = 1.64), but acts as an extender, and increases the scattering efficiency of the ZnS.

The barium sulfate in lithopone can be identified thermoanalytically by a reversible endothermic transformation at 1150 °C. Both Sachtolith and lithopone are thermally stable up to ca. 550 °C in the presence of air. Due to their low Mohs hardness, they are less abrasive than other white pigments. Barium sulfate is practically inert toward acids, bases, and organic solvents. Zinc sulfide is stable in aqueous media between pH 4 and 10, and is largely inert toward organic media. In the presence of water and oxygen, it can be oxidatively decomposed by the action of UV radiation.

2.2.2. *Production*

Raw Materials. The source of zinc can be zinc oxide from a smelter, zinc dross or sweepings, ammonium chloride slag from hot dip galvanizing, or liquid waste such as pickle liquors from galvanizing plants. Variations in the price of zinc have a large effect on the economics of zinc sulfide pigments.

The starting material for water-soluble barium compounds is fused barium sulfide produced by coke reduction of naturally occurring barite with a low silica and strontium content. Suitable barite is readily available from many deposits worldwide.

Lithopone. The reaction of equimolar quantities of $ZnSO_4$ and BaS produces a white, water-insoluble coprecipitate with the theoretical composition 29.4 wt % ZnS and 70.6 wt % $BaSO_4$:

$$ZnSO_4 + BaS \rightarrow ZnS + BaSO_4 \quad (1)$$

By using a different molar ratio, this composition can be changed; for example, precipitation according to Equation (2) gives a product containing 62.5 wt % ZnS and 37.5 wt % $BaSO_4$:

$$ZnSO_4 + 3\,ZnCl_2 + 4\,BaS \rightarrow 4\,ZnS + BaSO_4 + 3\,BaCl_2 \quad (2)$$

Figure 20 is a flow diagram of lithopone production. The *solutions of zinc salts* contain impurities (e.g., salts of iron, nickel, chromium, manganese, silver, cadmium) that depend on their origins. The main sources of zinc sulfate solutions are zinc electrolyses and the reprocessing of zinc scrap and zinc oxide. The first stage of purification consists of chlorination. Iron and manganese are precipitated as oxide-hydroxides, and cobalt, nickel, and cadmium as hydroxides. The solutions are then mixed with zinc dust at 80 °C. All the elements more noble than zinc (cadmium, indium, thallium, nickel, cobalt, lead, iron, copper, and silver) are almost completely precipitated, while zinc goes into solution. The metal slime is filtered off and taken to copper smelters for recovery of the noble metal components. A small quantity of a water-soluble cobalt salt is added to the purified zinc salt solution. The cobalt (0.02 – 0.5 %) becomes incorporated into the ZnS lattice during subsequent calcination to stabilize the final product against light. Zinc sulfide that is not treated in this way becomes gray in sunlight.

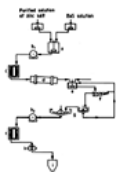

Figure 20. Flow diagram for lithopone production

a) Precipitation vessel; b) Rotary filter; c) Turbo dryer; d) Rotary kiln; e) Chilling vessel; f) Rake classifier; g) Thickener; h) Grinder; i) Silo

The *barium sulfide solution* is produced by dissolving fused barium sulfide in water. The barium sulfide is obtained by reducing an intimate mixture of crushed barite (ca. 1 cm lumps) with petroleum coke according to Equation (3) in a directly heated rotary kiln at 1200 – 1300 °C:

$$BaSO_4 + 2\,C \rightarrow BaS + 2\,CO_2 \quad (3)$$

The warm solution (60 °C) containing ca. 200 g/L barium sulfide is filtered and immediately pumped to the precipitation stage. Further purification is not necessary. Unreacted gangue and heavy metals are collected as insoluble sulfides in the filter cake. The almost clear solution can be stored only for a short period. Longer storage leads to undesirable polysulfide formation.

The zinc salt and BaS solutions are mixed thoroughly under controlled conditions (vessel geometry, temperature, pH, salt concentration, and stirring speed, see (a) in Fig. 20). The precipitated “raw lithopone” does not possess pigment properties. It is filtered off (b_1) and dried (c); ca. 2 cm lumps of the material are calcined in a rotary kiln (d) directly heated with natural gas at 650 – 700 °C. Crystal growth is controlled by adding 1 – 2 wt % NaCl, 2 wt % Na_2SO_4 and traces of Mg^{2+} (ca. 2000 ppm) and K^+ (ca. 100 – 200 ppm). The temperature profile and residence time in the kiln are controlled to obtain ZnS with an optimum particle size of ca. 300 nm.

The hot product from the kiln is quenched in water (e), passed via classifiers and hydroseparators (f) into thickeners (g), filtered on rotary filters (b_2), and washed until salt-free. The dried product is ground in high-intensity mills (g) and may undergo organic treatment (with a polyalcohol) depending on the application.

Figure 21 shows a scanning electron micrograph of lithopone. The ZnS and $BaSO_4$ particles can be distinguished by means of their size. The average particle diameter of $BaSO_4$ is 1 µm.

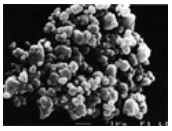

Figure 21. Scanning electron micrograph of lithopone. The larger particles are barium sulfate (mean size 1.0 µm) and the smaller particles are zinc sulfide (mean size 0.3 µm).

Sachtolith. Production is similar to that of lithopone. For process engineering reasons, a sodium sulfide solution is used as the sulfide component, and is formed according to Equation (4):

$$BaS + Na_2SO_4 \rightarrow Na_2S + BaSO_4 \quad (4)$$

The $BaSO_4$, which is produced as a byproduct, is filtered, washed, dried, and ground. It is a high-quality extender (Blanc fixe) used in the paint industry.

The Na_2S solution is mixed with a cobalt-treated zinc salt solution under precisely controlled conditions. The resulting zinc sulfide precipitate is calcined and processed to give the finished product.

Hydrothermal Process. Crystal growth of ZnS can be achieved by a hydrothermal process instead of by calcination. The raw lithopone is precipitated with a slight excess of sulfide at pH 8.5. The pH is then adjusted to 12 – 13 with sodium hydroxide solution, and 0.5 % sodium carbonate is added. The suspension is then autoclaved for ca. 15 – 20 min at 250 – 300 °C. In contrast to the wurtzite structure of the calcined product, the hydrothermal product has a sphalerite structure with a ca. 10 % greater scattering power. Although the product is of better quality, the hydrothermal process is less economic due to the high cost of the materials required for lining the autoclave (e.g., tantalum or a zirconium alloy).

Environmental Protection. During the reduction of barite and the calcination of Sachtolith and lithopone, sulfur dioxide is liberated. This is removed from the waste gas in a purification stage which is based on the reversible, temperature-dependent solubility of sulfur dioxide in polyglycol. The absorbed sulfur dioxide can be recovered as a liquid product or as a raw material for sulfuric acid. Any soluble barium in the residue from the dissolution of the fused BaS is removed by treatment with iron-containing waste hydrochloric acid. The gangue, in the form of a slime, is water-insoluble and can therefore be disposed of. The barium chloride solution is used as a raw material for the production of precipitated barium sulfate fillers.

2.2.3. *Commercial Products*

Commercial lithopone grades contain 30 % ZnS (red seal) and 60 % ZnS (silver seal). The ZnS content of Sachtolith is >97 %. Various chemical (e.g., polyalcohols, siloxanes, silanes) and mechanical treatments (e.g., jet milling) are used to obtain other products for special applications. The technical data for commercial red seal lithopone and Sachtolith are given in Table 20.

Table 20. Technical data for red seal lithopone and Sachtolith

Parameter	Standard	Sachtolith	Red seal lithopone
ZnS, wt %	DIN 55910	≥97	ca. 29
ZnO, wt %	DIN 55910	0.2	0.1
$BaSO_4$, wt %	DIN 55910	≤3	ca. 70
Brightness *		98	98 – 99
Lightening power	DIN 53192	400	120
Water-soluble salts, wt %	DIN 53197	< 0.2	< 0.2
Sieve residue **, wt %		0.02	< 0.1
Oil number	DIN 53192	12	8
pH	DIN 53200	6 – 8	7
Specific surface area, m^2/g		8	3

* $BaSO_4$ white standard=100.

** Test sieve 45 µm, DIN 4188, ISO DIS 3310/1.

2.2.4. *Uses*

Lithopone is mainly used in *coating materials* with relatively high pigment concentrations (Table 21). Examples are primers, plastic masses, putties and fillers, artists' colors, and emulsion paints. An important property of lithopone is its low binder requirement, giving paint products with good flow and application properties. It is suitable for almost all binder media, and has good wetting and dispersing properties. With optimum feed composition, good dispersion can be achieved simply by the action of a dissolver. It can be economically advantageous to use lithopone in combination with TiO_2 pigments; the good hiding power of the TiO_2 pigments is combined with the economic and technical advantages of lithopone. Due to the strong shift of the absorption band towards the blue, lithopone is especially useful as a white pigment for UV-cured paint systems. Zinc compounds have a fungicidal and algicidal action, and inclusion of lithopone or Sachtolith in paint formulations for exterior use therefore prevents attack by algae or fungi.

Table 21. Uses of Sachtolith and lithopone (as percentage of total consumption)

Use	Sachtolith	Lithopone	
		Western Europe	World
Paints	20	79	94
Plastics	64	17	2
Lubricants	6		
Others	10	4	4

The material advantages of lithopone are used in *plastics* (e.g., good lightfastness and clear bluish-white shade). The product also imparts very good extruding properties to the plastic resulting in high throughput rates and economic extruder operation. In fire-resistant systems, ca. 50 % of the flame retardant antimony trioxide can be replaced by nontoxic lithopone without any adverse effect.

Sachtolith is mainly used in plastics (Table 21). Functional properties such as lightening and hiding power are criteria for the use of Sachtolith. It has proved to be very useful for coloring many thermoplasts. During the dispersion process it does not cause abrasion of metallic production machinery or adversely effect the polymer, even at high operating temperatures or during multistage processing. Even ultrahigh molecular mass thermoplasts can be colored without problems. In glass-fiber-reinforced plastics, the soft texture of Sachtolith prevents mechanical fiber damage during extrusion. Sachtolith is also used as a dry lubricant during the fabrication of these materials.

The low abrasiveness of Sachtolith prolongs the operating life of stamping tools used in the manufacture of industrial rubber articles. The lightfastness and ageing resistance of many elastomers are improved by Sachtolith. It is also used as a dry lubricant for roller and plain bearings, and as a white pigment for greases and oils.

2.2.5. *Economic Aspects*

Total world production of lithopone in 1990 was 220×10^3 t. This was subdivided as follows (10^3 t):

Germany	30 (+7 Sachtolith)
Yugoslavia	5
Czechoslovakia	15
Soviet Union	50
People's Republic of China	120

Only estimated figures are available for the Soviet Union and the People's Republic of China.

A decrease in output is to be expected because replacement by TiO_2 is not yet at an end, especially in coating materials. In the long term, only the high-quality grades can maintain their place in the market, i.e., those in which technical properties are required in addition to light scattering.

2.2.6. *Toxicology*

The use of zinc sulfide and barium sulfate in contact with foods is permitted by the FDA (United States) and in most European countries. Some restrictions apply in France, Italy, the United Kingdom, and Czechoslovakia.

Soluble zinc is toxic in large amounts, but the human body requires small quantities (10 – 15 mg/d) for metabolism. Zinc sulfide is completely harmless in the human due to its low solubility. The acid concentration in the stomach and the rate of dissolution following ingestion are not sufficient to produce physiologically significant quantities of soluble zinc. LD_{50} values in the rat exceed 20 g/kg. No cases of poisoning or chronic damage to health have been observed in the manufacture of zinc sulfide pigments despite exposure to dust that occurs mainly during grinding and packing.

2.3. Zinc Oxide (Zinc White) [162-166]

2.3.1. *Introduction*

Zinc oxide [1314-13-2], ZnO, M_r 81.38, was formerly used only as a white pigment, and was named zinc white (C.I. Pigment White 4), Chinese white, or flowers of zinc. The term zinc white now denotes zinc oxide produced by the combustion of zinc metal according to the indirect or French process.

Historical Aspects. Zinc oxide has long been known as a byproduct of copper smelting. The Romans called it “cadmia” and used it as such in the production of brass. They also purified it for use in ointments by reduction followed by oxidation. In the Middle Ages, the alchemists thought that cadmia could be converted into gold.

In the mid-18th century, the German chemist CRAMER discovered that cadmia could be obtained by the combustion of metallic zinc. COURTOIS began to produce zinc white in 1781 in France, but it was not until 1840 that industrial production was started by LECLAIRE (indirect or French process). The use of this white pigment spread rapidly. Zinc oxide replaced white lead because it had the advantage of being nontoxic, of not darkening in the presence of sulfurous gases, and of having better hiding power.

Around 1850, S. WETHERILL of the New Jersey Zinc Company perfected a roasting furnace in which a grate was charged with coal and then covered with a mixture of zinc ore and coal. The zinc was reduced by the partial combustion of the coal and reoxidized at the furnace exit (direct or American process). These furnaces were subsequently improved but are now no longer used. During the second half of the nineteenth century, the use of ZnO in rubber was introduced to reduce the time needed for vulcanization. The discovery of the first organic accelerators for vulcanization in 1906 added to the importance of ZnO, which acts as an activator in these materials.

A third industrial production process exists but this wet process is less widely used.

2.3.2. *Properties*

Physical Properties. Zinc oxide is a fine white powder that turns yellow when heated above 300 °C. It absorbs UV light at wavelengths below 366 nm. Traces of monovalent or trivalent elements introduced into the crystal lattice impart semiconducting properties. The elementary particles of ZnO obtained by the thermal method may be granular or nodular (0.1 – 5 µm) or acicular (needle-shaped). Some physical properties are given below:

Density	5.65 – 5.68 g/cm^3
Refractive index	1.95 – 2.1
mp	1975 °C
Heat capacity	
25 °C	40.26 J mol^{-1} K^{-1}
100 °C	44.37 J mol^{-1} K^{-1}
1000 °C	54.95 J mol^{-1} K^{-1}
Thermal conductivity	25.2 W m^{-1} K^{-1}
Crystal structure	hexagonal, wurtzite
Mohs hardness	4 – 4.5

Chemical Properties. Zinc oxide is amphoteric; it reacts with organic and inorganic acids, and also dissolves in alkalis and ammonia solution to form zincates. It combines readily with acidic gases (e.g., CO_2, SO_2, and H_2S). It reacts at high temperatures with other oxides to form compounds such as zinc ferrites.

2.3.3. *Production*

About 1 – 2 % of zinc oxide is produced by the wet process, 10 – 20 % by the direct process, and the remainder by the indirect process.

Raw Materials. In the early days, the raw materials were mainly zinc ores or concentrates for the direct process, or metal from zinc producers for the indirect process. Nowadays, zinc oxide manufacturers mainly use residues and secondary zinc. This fact, combined with the demand for chemical purity imposed by the users, means that processes have had to be modified and a number of purification techniques are used.

Direct or American Process. The direct process is noted for its simplicity, low cost, and excellent thermal efficiency. It consists of an initial high-temperature reduction (1000 – 1200 °C) of a zinc-containing material (as oxide), the reducing agent being coal. Reduction takes place according to Boudouard's equations:

$$ZnO + C \rightleftharpoons Zn + CO$$

$$ZnO + CO \rightleftharpoons Zn + CO_2$$

$$C + O_2 \rightleftharpoons CO_2$$

$$CO_2 + C \rightleftharpoons 2CO$$

The zinc vapor and the CO gas are then oxidized to zinc oxide and carbon dioxide above the reaction bed or at the furnace exit. Various zinc-containing materials are used, e.g., zinc concentrates, metallization residues, byproduct zinc hydroxide, and above all zinc dross from casting furnaces or galvanizing. The dross must first be treated to remove chloride and lead by heating at ca. 1000 °C in rotary kilns.

Only rotary kilns are now used for the direct process; the use of static furnaces has been discontinued. The zinc content of raw materials is between 60 and 75 %. There are two types of rotary kiln:

1. One type is a long (ca. 30 m), fairly narrow (2.5 m diameter) kiln, heated by gas or oil. The raw material (a mixture of zinc-containing material and coal) is charged continuously either countercurrent or cocurrent to the combustion gases. The residues, which still contain some zinc and unburned coal, leave the furnace continuously at the end opposite to the feed end. The excess coal is sieved out and recycled. The combustion gases, containing zinc vapor, ZnO, and CO, pass into a chamber where oxidation is completed and large particles of impurities settle out. The gases are then cooled in a heat exchanger or by dilution with air. The zinc oxide is collected in bag filters.

2. The second type of rotary kiln is shorter (5 m) and has a larger diameter (ca. 3 m). Charging is continuous, but the dezincified residues are removed batchwise.

In both cases, operating conditions are controlled to obtain a high yield and to give the required particle shape and size. Provided no contamination is introduced, chemical purity is determined solely by the composition of the raw materials used.

Indirect or French Process. The zinc is boiled, and the resulting vapor is oxidized by combustion in air under defined conditions. The crystallographic and physical properties of the ZnO can be controlled by adjustment of the combustion conditions (e.g., flame turbulence and air excess). The chemical composition of the ZnO is solely a function of the composition of the zinc vapor.

Many types of furnace are available to produce vapor of the required purity from various raw materials and obtain a high yield of zinc. Pure zinc (super high grade, SHG; high grade, HG) or, to an increasing extent, metal residues (e.g., scrap zinc, die casting dross, or galvanizer's dross) are used as raw materials. Various liquid- or vapor-phase separation techniques are used for separating Cd, Pb, Fe, and Al from zinc metal before it is oxidized.

1. *Muffle Furnaces or Retorts of Graphite or Silicon Carbide.* The metal is fed into the furnace either batchwise as a solid or continuously as a liquid. The heat of vaporization is supplied by heating the outside of the retort with a burner. The nonvolatile residues (iron and lead in the case of dross from smelting) accumulate in the retort and must be removed at intervals. This is facilitated by tipping the retorts.
2. *Fractional Distillation.* The vapor, containing Cd, Pb, Fe, Al, and Cu, can be purified by fractional distillation in columns (New Jersey Zinc Co.) with silicon carbide plates. Oxidation takes place at the exit of the column.
3. *Furnaces with Two Separate Chambers.* The metallic raw material, which can be in large pieces, is fed into the first chamber where it melts. This is connected to the second, electrically heated chamber where distillation takes place in the absence of air. The first version of this type of furnace was constructed by LUNDEVALL [167]. The nonmetallic residues are removed at the surface of the melting chamber. Impurities, such as Fe, Al, and some of the Pb, accumulate in the distillation chamber and are periodically removed in the liquid state. The last traces of lead are then removed by fractional distillation.
4. *Smelting Process in a Rotary Kiln.* Indirect zinc oxide is also made by smelting in a rotary kiln, starting from the same raw materials. Melting, distillation, and part of the oxidation all take place in the same zone, allowing utilization of a large part of the heat of combustion of the zinc. By controlling the temperature and partial pressures of carbon dioxide and oxygen, the impurity content (Pb) can be limited and the shape and size of the ZnO particles can be adjusted, though to a lesser extent than in the other processes.

Wet Process. Zinc oxide is also produced industrially from purified solutions of zinc sulfate or chloride by precipitating the basic carbonate, which is then washed, filtered, and finally calcined. This method produces a grade of zinc oxide with a high specific surface area.

Products of this type are also obtained from waste hydroxides which are purified by a chemical route and then calcined.

Aftertreatment. Thermal treatment at temperatures up to 1000 °C improves the pigment properties of the ZnO and is mainly applied to oxide produced by the direct method. Controlled atmospheric calcination also improves the photoconducting properties of the high-purity oxide used in photocopying.

The ZnO surface is made more organophilic by coating it with oil and propionic acid. The ZnO is often deaerated and sometimes pelletized or granulated to improve handling properties.

2.3.4. *Quality Specifications*

Many standard specifications have been laid down for the more important uses of ZnO (rubber, paints, and the pharmaceutical industry). For standards, see Table 1 ("Zinc oxide pigments; Methods of analysis" and "Specification"). Various methods of classification are used, often based on the production process and the chemical composition. The most well-known are pharmacopeias, RAL 844 C3 (Reichs-Ausschuß für Lieferbedingungen), ASTM D79, BS 254, DAB 8, T 31 006 NF (French standard), and ISO R 275. Table 22 shows the classification of commercially available zinc oxide grades.

Table 22. Classification of commercially available grades of zinc oxide

Parameter	A	B	C	D
	Indirect process	Indirect process	Direct process	Wet process
ZnO (min.), wt %	99.5	99	98.5	93
Pb (max.), wt %	0.004	0.25	0.25	0.001
Cd (max.), wt %	0.001	0.05	0.03	0.001
Cu (max.), wt %	0.0005	0.003	0.005	0.001
Mn (max.), wt %	0.0005	0.001	0.005	0.001
Water-soluble salts (max.), wt %	0.02	0.1	0.65	1
Loss on ignition (max.), wt %	0.3	0.3	0.3	4
Acidity, g H_2SO_4/100 g	0.01	0.1	0.3	0.2
Specific surface area, m^2/g	3 – 8	3 – 10	1 – 5	25 (min.)

Classification based on color codes is common in Europe, but is of limited value. Manufacturers have their own standards. In general, the terms silver seal and white seal indicate category A, and red seal category B.

2.3.5. *Uses*

Zinc oxide has many uses. By far the most important is in the *rubber industry*. Almost half the world's ZnO is used as an activator for vulcanization accelerators in natural and synthetic rubber. The reactivity of the ZnO is a function of its specific surface area, but is also influenced by the presence of impurities such as lead and sulfates. The ZnO also ensures good durability of the vulcanized rubber, and increases its thermal conductivity. The ZnO content is usually 2 – 5 %.

In *paints and coatings*, zinc oxide is no longer the principal white pigment, although its superb white color is used by artists (⟶ Artists' Colors – Colored Inorganic Pigments). It is used as an additive in exterior paints for wood preservation. It is also utilized in antifouling and anticorrosion paints [168]. It improves film formation, durability, and resistance to mildew (having a synergistic effect with other fungicides) because it reacts with acidic products of oxidation and can absorb UV radiation.

The *pharmaceutical and cosmetic industries* use ZnO in powders and ointments because of its bactericidal properties. It is also used to form dental cements by its reaction with eugenol (⟶ Dental Materials).

In the field of *glass, ceramics, and enamels*, ZnO is used for its ability to reduce thermal expansion, to lower the melting point, and to increase chemical resistance. It can also be used to modify gloss or to improve opacity.

Zinc oxide is used as a *raw material* for many products: stearates, phosphates, chromates, bromates, organic dithiophosphates, and ferrites (ZnO, MnO, Fe_2O_3). It is used as a source of zinc in animal feeds and in electrogalvanization (⟶ Electrochemical and Chemical Deposition – Zinc Deposits). It is also used for desulfurizing gases.

Zinc oxide is used as a *catalyst* in organic syntheses (e.g., of methanol), often in conjunction with other oxides. It is present in some adhesive compositions.

The highest purity material is calcined with additives such as Bi_2O_3 and used in the manufacture of varistors (⟶ Ceramics, Electronic) [169]. The photoconducting properties of ZnO are used in photoreproduction processes. Doping with alumina causes a reduction in electrical resistance; hence, it can be used in the coatings on the master papers for offset reproduction [170].

2.3.6. *Economic Aspects*

The consumption of zinc oxide in Western Europe in 1990 was estimated to be 160 000 t. Annual world consumption is in the region of 500 000 t, representing ca. 10 % of the total world zinc production. The rubber industry consumes ca. 45 % of the total and the remainder is divided among a large number of industries.

2.3.7. *Toxicology and Occupational Health*

Unlike other heavy metals, zinc is not considered to be toxic or dangerous. It is an essential element for humans, animals, and plants. The human body contains ca. 2 g, and it is recommended that 10 – 20 mg should be ingested per day [171]. The oral LD_{50} value for rats is 630 mg/kg. The permitted concentration of the dust in air at the

workplace is 5 mg/m^3 (MAK), 10 mg/m^3 (TLV-TWA). Values for zinc oxide fumes are 5 mg/m^3(TLV-TWA) and 10 mg/m^3 (TLV-STEL).

If large quantities of ZnO are accidentally ingested or inhaled, fever, nausea, and irritation of the respiratory tract ensue after several hours. These symptoms rapidly disappear without long-term consequences.

[Top of Page]

3. Colored Pigments

Gunter Buxbaum, Helmut Printzen, Manfred Mansmann, Dieter Räde, Gerhard Trenczek, Volker Wilhelm, Stefanie Schwarz, Henning Wienand, Jörg Adel, Gerhard Adrian, Karl Brandt, William B. Cork, Heinrich Winkeler, Wielfried Mayer, Klaus Schneider

Colored pigments differ from black and white pigments in that their absorption and scattering coefficients are wavelength dependent with widely varying absolute values. The dependence of these coefficients on wavelength, particle size, particle shape, and their distributions determine the color, tinting strength, and hiding power of the pigments (see Section Fundamental Aspects).

3.1. Oxides and Hydroxides

In many inorganic pigments, lanthanides and transition elements are responsible for color. Metal oxides and oxide hydroxides are, however, also important as colored pigments because of their optical properties, low price, and ready availability. Colored pigments based on oxides and oxide hydroxides are either composed of a single component or mixed phases. In the latter, color is obtained by incorporation of appropriate cations.

3.1.1. *Iron Oxide Pigments*

The continually increasing importance of iron oxide pigments is based on their nontoxicity; chemical stability; wide variety of colors ranging from yellow, orange, red, brown, to black; and low price. Natural and synthetic iron oxide pigments consist of well-defined compounds with known crystal structures [172]:

1. α-FeOOH, goethite [1310-14-1], diaspore structure, color changes with increasing particle size from green-yellow to brown-yellow
2. ϒ-FeOOH, lepidocrocite [12022-37-6], boehmite structure, color changes with increasing particle size from yellow to orange
3. α-Fe_2O_3, hematite [1317-60-8], corundum structure, color changes with increasing particle size from light red to dark violet
4. ϒ-Fe_2O_3, maghemite [12134-66-6], spinel super structure, ferrimagnetic, color: brown
5. Fe_3O_4, magnetite [1309-38-2], spinel structure, ferrimagnetic, color: black

Mixed metal oxide pigments containing iron oxide are also used (see Section Mixed Metal Oxide Pigments). Magnetic iron oxide pigments are discussed in Sections Iron Oxide Pigments and Cobalt-Containing Iron Oxide Pigments. Transparent iron oxide pigments are described in Section Transparent Iron Oxides. Methods of analysis and specifications of iron oxide pigments are listed in the standards given in Table 1.

3.1.1.1. *Natural Iron Oxide Pigments*

Naturally occurring iron oxides and iron oxide hydroxides were used as pigments in prehistoric times (Altamira cave paintings) [173]. They were also used as coloring materials by the Egyptians, Greeks, and ancient Romans.

Hematite (α-Fe_2O_3) has attained economic importance as a red pigment, goethite (α-FeOOH) as yellow, and the umbers and siennas as brown pigments. Deposits with high iron oxide contents are exploited preferentially. Naturally occurring magnetite (Fe_3O_4) has poor tinting strength as a black pigment, and has found little application in the pigment industry.

Hematite is found in large quantities in the vicinity of Malaga in Spain (Spanish red) and near the Persian Gulf (Persian red). The Spanish reds have a brown undertone. Their water-soluble salt content is very low and their Fe_2O_3 content often exceeds 90 %. The Persian reds have a pure hue, but their water-soluble salt content is disadvantageous for some applications. Other natural hematite deposits are of only local importance. A special variety occurs in the form of platelets and is extracted in large quantities in Kärnten (Austria). This micaceous iron oxide, is mainly used in corrosion protection coatings.

Goethite is the colored component of yellow ocher; a weathering product mainly of siderite, sulfidic ores, and feldspar. It occurs in workable amounts mainly in the Republic of South Africa and France. The Fe_2O_3 content gives an indication of the iron oxide hydroxide content of the ocher, and is ca. 20 % in the French deposits and ca. 55 % in the South African.

Umbers are mainly found in Cyprus. In addition to Fe_2O_3 (45 – 70 %), they contain considerable amounts of manganese dioxide (5 – 20 %). In the raw state, they are deep brown to greenish brown and when calcined are dark brown with a red undertone (burnt umbers).

Siennas, mainly found in Tuscany, have an average Fe_2O_3 content of ca. 50 %, and contain < 1 % manganese dioxide. They are yellow-brown in the natural state and red-brown when calcined [174].

The processing of natural iron oxide pigments depends on their composition. They are either washed, slurried, dried, ground, or dried immediately and then ground in ball mills, or more often in disintegrators or impact mills (see ⟶ Size Reduction – Grinding Media Mills, ⟶ Size Reduction) [175].

Siennas and umbers are calcined in a directly fired furnace, and water is driven off. The hue of the products is determined by the calcination period, temperature, and raw material composition [176].

Natural iron oxide pigments are mostly used as inexpensive marine coatings or in coatings with a glue, oil, or lime base. They are also employed to color cement, artificial stone, and wallpaper. Ocher and sienna pigments are used in the production of crayons, drawing pastels, and chalks [177].

For standards, see Table 1 (Iron, manganese oxide pigments: “Methods of analysis”; “Natural, specification”; “Sienna, specification”; and “Umber, specification”).

The economic importance of the natural iron oxide pigments has decreased in recent years in comparison with the synthetic materials.

3.1.1.2. *Synthetic Iron Oxide Pigments*

Synthetic iron oxide pigments have become increasingly important due to their pure hue, consistent properties, and tinting strength. Single- component forms are mainly produced with red, yellow, orange, and black colors. Their composition corresponds to that of the minerals hematite, geothite, lepidocrocite, and magnetite. Brown pigments usually consist of mixtures of red and/or yellow and/or black; homogeneous brown phases are also produced, e.g., $(Fe, Mn)_2O_3$ and ϒ-Fe_2O_3, but quantities are small in comparison to the mixed materials. Ferrimagnetic ϒ-Fe_2O_3 is of great importance for magnetic recording materials (see Sections Iron Oxide Pigments and Cobalt-Containing Iron Oxide Pigments).

Several processes are available for producing high-quality iron oxide pigments with controlled mean particle size, particle size distribution, particle shape, etc. (Table 23):

1. Solid-state reactions (red, black, brown)
2. Precipitation and hydrolysis of solutions of iron salts (yellow, red, orange, black)
3. Laux process involving reduction of nitrobenzene (black, yellow, red)

Table 23. Reaction equations for the production of iron oxide pigments

Color	Reaction		Process
Red	$6\ FeSO_4 \cdot x\ H_2O + 1½\ O_2$	⟶ $Fe_2O_3 + 2\ Fe_2(SO_4)_3 + 6\ H_2O$	copperas process
	$2\ Fe_2(SO_4)_3$	⟶ $2\ Fe_2O_3 + 6\ SO_3$	
	$2\ Fe_3O_4 + ½\ O_2$	⟶ $3\ Fe_2O_3$	calcination
	2 FeOOH	⟶ $Fe_2O_3 + H_2O$	calcination
		⟶	Ruthner process

	$2\ FeCl_2+2\ H_2O+½\ O_2$	$Fe_2O_3+4\ HCl$	
	$2\ FeSO_4+½\ O_2+4\ NaOH$	$\longrightarrow Fe_2O_3+2\ Na_2SO_4+2\ H_2O$	precipitation
Yellow	$2\ FeSO_4+4\ NaOH+½\ O_2$	$\longrightarrow 2\ \alpha\text{-FeOOH}+2\ Na_2SO_4+H_2O$	precipitation
	$2\ Fe+2\ H_2SO_4$	$\longrightarrow 2\ FeSO_4+2\ H_2$	Penniman process
	$2\ FeSO_4+½\ O_2+3\ H_2O$	$\longrightarrow 2\ \alpha\text{-FeOOH}+2\ H_2SO_4$	
	$2\ Fe+½\ O_2+3\ H_2O$	$\longrightarrow 2\ \alpha\text{-FeOOH}+2\ H_2$	
	$2\ Fe+C_6H_5NO_2+2\ H_2O$	$\longrightarrow 2\ \alpha\text{-FeOOH}+C_6H_5NH_2$	Laux process
Orange	$2\ FeSO_4+4\ NaOH+½\ O_2$	$\longrightarrow 2\ \gamma\text{-FeOOH}+2\ Na_2SO_4+H_2O$	precipitation
Black	$3\ FeSO_4+6\ NaOH+½\ O_2$	$\longrightarrow Fe_3O_4+3\ Na_2SO_4+3\ H_2O$	1-step precipitation
	$2\ FeOOH+FeSO_4+2\ NaOH$	$\longrightarrow Fe_3O_4+Na_2SO_4+2\ H_2O$	2-step precipitation
	$9\ Fe+4\ C_6H_5NO_2+4\ H_2O$	$\longrightarrow 3\ Fe_3O_4+4\ C_6H_5NH_2$	Laux process
	$3\ Fe_2O_3+H_2$	$\longrightarrow 2\ Fe_3O_4+H_2O$	reduction
Brown	$2\ Fe_3O_4+½\ O_2$	$\longrightarrow 3\ \gamma\text{-}Fe_2O_3$	calcination
	$3\ Fe_3O_4+Fe_2O_3+MnO_2+½\ O_2$	$\longrightarrow (Fe_{11},Mn)O_{18}$	calcination

The raw materials are mainly byproducts from other industries: steel scrap obtained from deep drawing, grindings from cast iron, $FeSO_4 \cdot 7\ H_2O$ from TiO_2 production or from steel pickling, and $FeCl_2$ also from steel pickling.

Iron oxides obtained after flame spraying of spent hydrochloric acid pickle liquor, red mud from bauxite processing, and the product of pyrites combustion are no longer of importance. They yield pigments with inferior color properties that contain considerable amounts of water-soluble salts. They can therefore only be used in low-grade applications.

Solid-State Reactions of Iron Compounds. Black iron oxides obtained from the Laux process (see below) or other processes may be calcined in rotary kilns with an oxidizing atmosphere under countercurrent flow to produce a wide range of different red colors, depending on the starting material. The pigments are ground to the desired particle size in pendular mills, pin mills, or jet mills, depending on their hardness and intended use.

The calcination of yellow iron oxide produces pure red iron oxide pigments with a high tinting strength. Further processing is similar to that of calcined black pigments.

High-quality pigments called copperas reds are obtained by the thermal decomposition of $FeSO_4 \cdot 7\ H_2O$ in a multistage process (Fig. 22). If an alkaline-earth oxide or carbonate is included during calcination, the sulfate can be reduced with coal or carbon- containing compounds to produce sulfur dioxide, which is oxidized with air to give sulfuric acid [178-181]. The waste gases and the dissolved impurities that are leached out in the final stage present ecological problems, however.

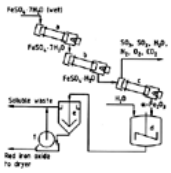

Figure 22. Production of copperas red

a) Dryer; b) Rotary kiln (dewatering); c) Rotary kiln; d) Tank; e) Thickener; f) Filter

Lower quality products can be obtained by single-stage calcination of iron(II) sulfate heptahydrate in an oxidizing atmosphere. The pigments have a relatively poor tinting strength and a blue tinge. Decomposition of iron(II) chloride monohydrate in air at high temperatures also yields a low-quality red iron oxide pigment [182].

In a new process, micaceous iron oxide is obtained in high yield by reacting iron(III) chloride and iron at 500 – 1000 °C in an oxidizing atmosphere in a tubular reactor [183].

Black Fe_3O_4 pigments with a high tinting strength can be prepared by calcining iron salts under reducing conditions [184]. This process is not used industrially because of the furnace gases produced.

Controlled oxidation of Fe_3O_4 at ca. 500 °C produces a single-phase brown $\gamma\text{-}Fe_2O_3$ with a neutral hue [185].

Calcination of α-FeOOH with small quantities of manganese compounds gives homogeneous brown pigments with the composition $(Fe, Mn)_2O_3$ [186]. Calcination of iron and chromium compounds that decompose at elevated temperatures yields corresponding pigments with the composition $(Fe, Cr)_2O_3$ [187].

Precipitation Processes. In principle, all iron oxide hydroxide phases can be prepared from aqueous solutions of iron salts (see Table 23). However, precipitation with alkali produces neutral salts (e.g., Na_2SO_4, NaCl) as byproducts which enter the wastewater.

Precipitation is especially suitable for producing soft pigments with a pure, bright hue. The manufacture of α-FeOOH yellow is described as an example. The raw materials are iron(II) sulfate ($FeSO_4 \cdot 7\ H_2O$) or liquors from the pickling of iron and steel (⟶ Iron Compounds), and alkali [NaOH, $Ca(OH)_2$, ammonia, or magnesite]. The pickle liquors usually contain appreciable quantities of free acid, and are therefore first optionally neutralized by reaction with scrap iron. Other metallic ions should not be present in large amounts, because they have an adverse effect on the hue of the iron oxide pigments.

The solutions of the iron salts are first mixed with alkali in open reaction vessels (Fig. 23 Route A) and oxidized, usually with air. The quantity of alkali used is such that the pH remains acidic. The reaction time (ca. 10 – 100 h) depends on the temperature (10 – 90 °C) and on the desired particle size of the pigment. This method yields yellow pigments (α-FeOOH) [188], [189]. If yellow nuclei are produced in a separate reaction (Route A in Fig. 23, tank c), highly consistent yellow iron oxide pigments with a pure color can be obtained [190].

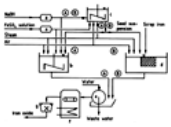

Figure 23. Production of yellow iron oxide by the precipitation (A) and Penniman (B) processes

a) Tank; b) Pigment reactor; c) Seed reactor; d) Pigment reactor with scrap basket; e) Filter; f) Dryer; g) Mill

If precipitation is carried out at ca. 90 °C while air is passed into the mixture at ca. pH≥7, black iron oxide pigments with a magnetite structure and a good tinting strength are obtained when the reaction is stopped at a FeO : Fe_2O_3 ratio of ca. 1:1. The process can be accelerated by operating at 150 °C under pressure; this technique also improves pigment quality [191]. Rapid heating of a suspension of iron oxide hydroxide with the necessary quantity of $Fe(OH)_2$ to ca. 90 °C also produces black iron oxide of

pigment quality [192], [193].

Orange iron oxide with the lepidocrocite structure (γ-FeOOH) is obtained if dilute solutions of the iron(II) salt are precipitated with sodium hydroxide solution or other alkalis until almost neutral. The suspension is then heated for a short period, rapidly cooled, and oxidized [194], [195].

Very soft iron oxide pigments with a pure red color may be obtained by first preparing α-Fe_2O_3 nuclei, and then continuously adding solutions of iron(II) salt with atmospheric oxidation at 80 °C. The hydrogen ions liberated by oxidation and hydrolysis are neutralized by adding alkali and keeping the pH constant [196]. Pigment-quality α-Fe_2O_3 is also obtained when solutions of an iron(II) salt, preferably in the presence of small amounts of other cations, are reacted at 60 – 95 °C with excess sodium hydroxide and oxidized with air [197].

Penniman process. The *Penniman process* is probably the most widely used production method for yellow iron oxide pigments [198], [199]. This method considerably reduces the quantity of neutral salts formed as byproducts. The raw materials are iron(II) sulfate, sodium hydroxide solution, and scrap iron. If the sulfate contains appreciable quantities of salt impurities, these must be removed by partial precipitation. The iron must be free of alloying components. The process usually consists of two stages (Route B in Fig. 23).

In the first stage, nuclei are prepared by precipitating iron(II) sulfate with alkali (e.g., sodium hydroxide solution) at 20 – 50 °C with aeration (c). Depending on the conditions, yellow, orange, or red nuclei may be obtained. The suspension of nuclei is pumped into vessels charged with scrap iron (d) and diluted with water. Here, the process is completed by growing the iron oxide hydroxide or oxide onto the nuclei. The residual iron(II) sulfate in the nuclei suspension is oxidized to iron(III) sulfate by blasting with air at 75 – 90 °C. The iron(III) sulfate is then hydrolyzed to form FeOOH or α-Fe_2O_3. The liberated sulfuric acid reacts with the scrap iron to form iron(II) sulfate, which is also oxidized with air. The reaction time can vary from ca. two days to several weeks, depending on the conditions chosen and the desired pigment. At the end of the reaction, metallic impurities and coarse particles are removed from the solid with sieves or hydrocyclones; water-soluble salts are removed by washing. Drying is carried out with band or spray dryers (f) and disintegrators or jet mills are used for grinding (g). The main advantage of this process over the precipitation process lies in the small quantity of alkali and iron(II) sulfate required. The bases are only used to form the nuclei and the relatively small amount of iron(II) sulfate required initially is continually renewed by dissolving the iron by reaction with the sulfuric acid liberated by hydrolysis. The process is thus considered environmentally friendly. The iron oxide pigments produced by the Penniman process are soft, have good wetting properties, and a very low flocculation tendency [198-206].

Under suitable conditions the Penniman process can also be used to produce reds directly. The residual scrap iron and coarse particles are removed from the pigment, which is then dried [207] and ground using disintegrators or jet mills. These pigments have unsurpassed softness. They usually have purer color than the harder red pigments produced by calcination.

The Laux Process. The Béchamp reaction (i.e., the reduction of aromatic nitro compounds with antimony or iron) which has been known since 1854, normally yields a black-gray iron oxide that is unsuitable as an inorganic pigment. By adding iron(II) chloride or aluminum chloride solutions, sulfuric acid, and phosphoric acid, LAUX modified the process to yield high-quality iron oxide pigments [208]. Many types of pigments can be obtained by varying the reaction conditions. The range extends from yellow to brown (mixtures of α-FeOOH and/or α-Fe_2O_3 and/or Fe_3O_4) and from red to black. If, for example, iron(II) chloride is added, a black pigment with very high tinting strength is produced [208]. However, if the nitro compounds are reduced in the presence of aluminum chloride, high-quality yellow pigments are obtained [209]. Addition of phosphoric acid leads to the formation of light to dark brown pigments with good tinting strength [210]. Calcination of these products (e.g., in rotary kilns) gives light red to dark violet pigments. The processes are illustrated in Figure 24.

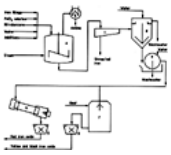

Figure 24. Production of iron oxide pigment by the Laux process

a) Reactor; b) Condenser; c) Classifier; d) Thickener; e) Filter; f) Dryer; g) Mill; h) Rotary kiln

The type and quality of the pigment are determined not only by the nature and concentration of the additives, but also by the reaction rate. The rate depends on the grades of iron used, their particle size, the rates of addition of the iron and nitrobenzene (or another nitro compound), and the pH value. No bases are required to precipitate the iron compounds. Only ca. 3 % of the theoretical amount of acid is required to dissolve all of the iron. The aromatic nitro compound oxidizes the Fe^{2+} to Fe^{3+} ions, acid is liberated during hydrolysis and pigment formation, and more metallic iron is dissolved by the liberated acid to form iron(II) salts; consequently, no additional acid is necessary.

The iron raw materials used are grindings from iron casting or forging that must be virtually free of oil and grease. The required fineness is obtained by size reduction in edge runner mills and classification with vibratory sieves. The iron and the nitro compound are added gradually via a metering device to a stirred tank (a) containing the other reactants (e.g., iron(II) chloride, aluminum chloride, sulfuric acid, and phosphoric acid). The system rapidly heats up to ca. 100 °C and remains at this temperature for the reaction period. The nitro compound is reduced to form an amine (e.g., aniline from nitrobenzene) which is removed by steam distillation. Unreacted iron is also removed (e.g., in shaking tables, c). The pigment slurry is diluted with water in settling tanks (d) and the pigment is washed to remove salts, and filtered on rotary filters (e). It may then be dried on band, pneumatic conveyor, or spray dryers to form yellow or black pigments, or calcined in rotary kilns (h) in an oxidizing atmosphere to give red or brown pigments. Calcination in a nonoxidizing atmosphere at 500 – 700 °C improves the tinting strength [211]. The pigments are then ground to the desired fineness in pendular mills, pin mills, or jet mills, depending on their hardness and application.

The Laux process is a very important method for producing iron oxide because of the coproduction of aniline; it does not generate byproducts that harm the environment.

Other Production Processes. The three processes already described are the only ones that are used on a large scale. The following processes are used on a small scale for special applications:

1. Thermal decomposition of $Fe(CO)_5$ to form transparent iron oxides (see Section Transparent Pigments) [212]
2. Hydrothermal crystallization for the production of α-Fe_2O_3 in platelet form [213]

3.1.1.3. *Toxicology and Environmental Aspects*

The Berufsgenossenschaft der Chemischen Industrie (Germany) has recommended that all iron oxide pigments should be classified as inert fine dusts with an MAK value of 6 mg/m^3. A substantial decrease is under consideration.

Iron oxide pigments produced from pure starting materials may be used as colorants for food and pharmaceutical products [214]. Synthetic iron oxides do not contain crystalline silica and therefore are not considered to be toxic, even under strict Californian regulations.

3.1.1.4. *Quality*

The red and black iron oxide pigments produced by the methods described have an Fe_2O_3 content of 92 – 96 wt %. For special applications (e.g., ferrites) analytically pure pigments with Fe_2O_3 contents of 99.5 – 99.8 wt % are produced. The Fe_2O_3 content of yellow and orange pigments lies between 85 and 87 wt % corresponding to FeOOH contents of 96 – 97 wt %. For standards, see Table 1 (Iron oxide pigments: “Black, specification”; “Red, specification”; “Yellow, specification”; and “FeO content”). Variations of 1 – 2 % are of no importance with respect to the quality of the pigments. Pigment quality is mainly determined by the quantity and nature of the water-soluble salts, the particle size distribution (hue and tinting strength are effected) and the average particle size of the ground product. The hue of red iron oxide is determined by the particle diameter, which is ca. 0.1 µm for red oxides with a yellow tinge and ca. 1.0 µm for violet hues.

The optical properties of the yellow, usually needle-shaped, iron oxide pigments depend not only on the particle size, but also on the length to width ratio (e.g., length = 0.3 – 0.8 µm, diameter = 0.05 – 0.2 µm, length: diameter ratio = ca. 1.5 – 8). In applications for which needle-shaped particles are unsuitable, spheroidal pigments are available [215]. Black iron oxide pigments (Fe_3O_4) have a particle diameter of ca. 0.1 – 0.6 µm.

Some iron oxide pigments have a limited stability on heating. Red iron oxide is stable up to 1200 °C in air. In the presence of oxygen, black iron oxide changes into brown γ-Fe_2O_3 at ca. 180 °C and then into red α-Fe_2O_3 above 350 °C. Yellow iron oxide decomposes above ca. 180 °C to form red α-Fe_2O_3 with liberation of water. This temperature limit can be increased to ca. 260 °C by stabilization with basic aluminum compounds. The thermal behavior of brown iron oxides produced by mixing depends on their composition.

3.1.1.5. *Uses*

All synthetic iron oxides possess good tinting strength and excellent hiding power. They are also lightfast and resistant to alkalis. These properties are responsible for their versatility. The principle areas of use are shown in Table 24.

Table 24. Main areas of use for natural and synthetic iron oxide pigments

Use	Amount, %		
	Europe	United States	World
Coloring construction materials	64	37	60
Paints and coatings	30	48	29
Plastics and rubber	4	14	6
Miscellaneous	2	1	5

Iron oxide pigments have long been used for coloring construction materials. Concrete roof tiles, paving bricks, fibrous cement, bitumen, mortar, rendering, etc. can be colored with small amounts of pigment that do not affect the setting time, compression strength, or tensile strength of the construction materials. Synthetic pigments are superior to the natural pigments due to their better tinting power and purer hue.

Natural rubber can only be colored with iron oxides that contain very low levels of copper and manganese (Cu < 0.005 %, Mn < 0.02 %). Synthetic rubber is less sensitive.

In the paint and coating industries, iron oxide pigments can be incorporated in many types of binders. Some reasons for their wide applicability in this sector are pure hue, good hiding power, good abrasion resistance, and low settling tendency. Their high temperature resistance allows them to be used in enamels.

The use of iron oxide as a polishing medium for plate glass manufacture has decreased now that other methods of glass production are available.

3.1.1.6. *Economic Aspects*

Accurate production figures for natural and synthetic iron oxide pigments are difficult to obtain, because statistics also include nonpigmentary oxides (e.g., red mud from bauxite treatment, intermediate products used in ferrite production). World production of synthetic iron oxides in 1995 was estimated to be 600 000 t; production of natural oxides was ca. 100 000 t. The most important producing countries for synthetic pigments are Germany, the United States, Italy, Brazil, China, and Japan. The natural oxides are mainly produced in France, Spain, Cyprus, India, Iran, Italy, and Austria.

The most important *manufacturers* are Bayer, Elementis, Laporte, and Toda.

3.1.2. *Chromium Oxide Pigments*

See also ⟶ Chromium Compounds – Chromium(III) Oxide and Chromium Hydroxide.

Chromium oxide pigments, also called chromium oxide green pigments, consist of chromium(III) oxide [1308-38-9], Cr_2O_3, M_r 151.99. Chromium oxide green is one of the few single- component pigments with green coloration. Chrome green is a blend of chrome yellow and iron blue pigments; phthalochrome green is a blend of chrome yellow and blue phthalocyanine pigments.

Natural, minable deposits of chromium oxide are not known. In addition to pigment grade, chromium oxide producers usually also offer a technical grade for applications based on properties other than coloration. These include:

1. Metallurgy: aluminothermic production of chromium metal (⟶ Chromium and Chromium Alloys)
2. Refractory industry: production of thermally and chemically resistant bricks and lining materials
3. Ceramic industry: coloring of porcelain enamels, ceramic frits, and glazes
4. Pigment industry: raw material for the production of chromium- containing stains and pigments based on mixed metal oxide phases (⟶ Ceramic Colorants, see also Section Mixed Metal Oxide Pigments)
5. Grinding and polishing agent: chromium (III) oxide is used in brake linings and polishing agents due to its high hardness

Chromium oxide hydroxide and hydrated chromium oxide pigments (Grignets Green) have a very attractive blue-green color. They are of low opacity, but provide excellent lightfastness and good chemical resistance. Loss of water on heating limits the application temperature. These pigments are no longer of industrial importance [216].

3.1.2.1. *Properties*

Chromium(III) oxide crystallizes in the rhombohedral structure of the corundum type; space group D^6_{3d}-$R\bar{3}$ c, ϱ 5.2 g/cm^3. Because of its high hardness (ca. 9 on the Mohs scale) the abrasive properties of the pigment must be taken into account in certain applications [217]. It melts at 2435 °C but starts to evaporate at 2000 °C. Depending on the manufacturing conditions, the particle sizes of chromium oxide pigments are in the range 0.1 – 3 µm with mean values of 0.3 – 0.6 µm. Most of the particles are isometric. Coarser chromium oxides are produced for special applications, e.g., for applications in the refractory area.

Chromium oxide has a refractive index of ca. 2.5. Chromium oxide green pigments have an olive green tint. Lighter greens with yellowish hues are obtained with finely divided pigments, and darker, bluish tints with larger particle diameters; the darker pigments are weaker colorants. The maximum of the reflectance curve lies in the green region of the spectrum at ca. 535 nm (Fig. 25, curve a). A weaker maximum in the violet region (ca. 410 nm) is caused by Cr – Cr interactions in the crystal lattice. Chromium oxide green pigments are used in IR-reflecting camouflage coatings because of their relatively high reflectance in the near infrared (Fig. 25, curve b).

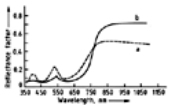

Figure 25. Dependence of the reflectance of chromium oxide on the wavelength

a) Regular pigment; b) Special product with larger particle size and high IR reflectance

Since chromium(III) oxide is virtually inert, chromium oxide green pigments are remarkably stable. They are insoluble in water, acid, and alkali and are thus extremely stable to sulfur dioxide and in concrete. They are light-, weather-, and temperature-resistant. A change of the tint only occurs above 1000 °C due to particle growth.

3.1.2.2. *Production*

Alkali dichromates are used as starting materials for the production of chromium(III) oxide pigments. They are available as bulk industrial products in the required purity. High impurity levels have an unfavorable effect on the hue.

Reduction of Alkali Dichromates. In industrial processes, solid alkali dichromates are reacted with reducing agents such as sulfur or carbon compounds. The reaction is strongly exothermic, and with sulfur proceeds as follows:

$$Na_2Cr_2O_7 + S \rightarrow Cr_2O_3 + Na_2SO_4$$

Sodium sulfate can then easily be separated by washing, because it is water soluble. The use of sulfur was first described in 1820 [218]. ROTH described the use of $K_2Cr_2O_7$ in 1927 [219]. If charcoal is used in place of sulfur, Na_2CO_3 is formed as byproduct [220].

Finely divided sodium dichromate (dihydrate or anhydrous) is mixed homogeneously with sulfur. This mixture is then reacted in a furnace lined with refractory bricks at 750 – 900 °C. An excess of sulfur is used to ensure completion of the reaction. The reaction mass is leached with water to remove water-soluble components such as sodium sulfate. The solid residue is then separated, dried, and ground.

If potassium dichromate is used instead of sodium dichromate, a green pigment with a more bluish hue is obtained.

If it is to be used as a pigment in paints and lacquers, chromium oxide green can be subjected to jet milling (micronization) to obtain the required properties (e.g., gloss).

Reduction of Ammonium Dichromate. Chromium(III) oxide can be obtained by thermal decomposition of ammonium dichromate. Above ca. 200 °C, a highly voluminous product is formed with elimination of nitrogen [221]. The pigment is obtained after addition of alkali salts (e.g., sodium sulfate) and subsequent calcination [222].

In the industrial process, a mixture of ammonium sulfate or chloride and sodium dichromate is calcined [223]:

$$Na_2Cr_2O_7 \cdot 2\,H_2O + (NH_4)_2SO_4 \rightarrow Cr_2O_3 + Na_2SO_4 + 6\,H_2O + N_2$$

The workup is then carried out as described above. A chromium oxide pigment obtained by this process typically contains (wt %):

Cr_2O_3	99.0 – 99.5
SiO_2	0.05 (max.)
Al_2O_3	0.1 (max.)
Fe_2O_3	0.05 (max.)
S	ca. 0.02
Water	ca. 0.3

Chromium oxides with a minimal sulfur content are preferred for metallurgical applications. These are obtained by reacting sodium dichromate with ammonium chloride or sulfate in a deficiency of 10 mol % [224]. Chromium (III) oxides with a low sulfur content can also be obtained by thermal aftertreatment [225]. Thermal decomposition of chromic acid anhydride (CrO_3) yields high-purity chromium(III) oxide [226].

The pigment properties of chromium oxides can be modified by precipitation of hydroxides (e.g., of titanium or aluminum), and subsequent calcining. This treatment changes the color to yellow-green, and decreases the flocculation tendency [227]. Aftertreatment with organic compounds (e.g., alkoxylated alkylsulfonamides) is also used [228].

Other Processes. Other production processes are suggested in the patent literature, but have not so far gained industrial importance. For instance, sodium dichromate can be mixed with heating oil and reacted at 300 °C. The soda formed must be washed out prior to calcining at 800 °C to avoid reoxidation in the alkaline melt [229].

In alkaline solution, sodium chromate can be reduced with sulfur at atmospheric pressure with formation of sodium thiosulfate. After neutralization, more sodium chromate is added to exhaust the reducing capacity of the thiosulfate. The mixture is calcined at 900 – 1070 °C [230].

Another process involves the shock heating of sodium dichromate in a flame at 900 – 1600 °C in the presence of excess hydrogen and chlorine to bind the alkali as sodium chloride [231]. This method is suitable for the preparation of pigment-grade chromium oxide of high purity, with an especially low sulfur content.

Environmental Protection. Since alkali dichromates or chromic acid anhydride are used as starting materials for the production of chromium(III) oxides, occupational health requirements for the handling of hexavalent chromium compounds must be observed [232]. The sulfur dioxide formed on reduction with excess sulfur must be removed from the flue gases according to national regulations, e.g., by oxidation to H_2SO_4.

Process wastewater may contain small amounts of unreacted chromates; recovery is uneconomical. Prior to release into drainage systems, the chromates in these wastewater streams must be reduced (e.g., with SO_2 or $NaHSO_3$) and precipitated as chromium hydroxide [233]. In Germany, for example, the minimum requirements for wastewater in the production of chromium oxide pigments are specified in [234].

3.1.2.3. *Quality Specifications and Analysis*

International, technical specifications for chromium oxide pigments are defined in ISO 4621 (1986), they must have a minimum Cr_2O_3 content of 96 wt %.

Various grades are defined according to their particle fineness as measured by the residue on a 45 µm sieve: grade 1, 0.01 % residue (max.); grade 2, 0.1 % (max.); and grade 3, 0.5 % (max.).

ISO 4621 (1986) also specifies analytical methods. Usually, analysis of chromium and the byproducts is preceded by melting with soda and sodium peroxide. The content of water-soluble or acid-soluble chromium is becoming important from the toxicological and ecological point of view. It is determined according to DIN 53 780 with water, or according to ISO 385 615 with 0.1 mol/L hydrochloric acid.

3.1.2.4. *Storage and Transportation*

Chromium(III) oxide pigments are thermally stable and insoluble in water. They are not classified as hazardous materials and are not subject to international transport regulations. As long as they are kept dry their utility as a pigment is practically unlimited.

3.1.2.5. *Uses*

The use of chromium(III) oxide as a pigment for toys, cosmetics, and in plastics and paints that come in contact with food is permitted in national and international regulations [235-242]. Maximum limits for heavy metals or their soluble fractions are usually a prerequisite. Because pure starting materials are used, these limits are satisfied by most types of chromium oxide.

Chromium oxide is equally important as a colorant and in its other industrial applications. As a pigment, it is used predominantly in the paint and coatings industry for high quality green paints with special requirements, especially for steel constructions (coil coating), facade coatings (emulsion paints), and automotive coatings.

A series of RAL tints (e.g., Nos. 6003, 6006, 6011, 6014, and 6015) can be formulated based on chromium oxide. As mentioned previously, chromium oxide is also an important pigment for the formulation of green camouflage coatings (e.g., RAL 6031-F 9, Natogreen 285, Stanag 2338, and Forestgreen MIL-C-46168 C).

Except for the expensive cobalt green, chromium oxide is the only green pigment that meets the high color stability requirements for building materials based on lime and cement [243]. In plastics, however, chromium oxide green is only of minor importance because of its dull tint.

The industrial significance of chromium oxide is due to its chemical and physical properties. Its high purity makes it suitable as a starting material for the aluminothermic production of very pure chromium metal.

Since the late 1970s chromium oxide has gained significance as a raw material in the refractory industry. The addition of chromium oxide to bricks and refractory concrete based on alumina significantly improves their stability against slag in the production and processing of pig iron. Chromium oxide bricks containing ca. 95 wt % Cr_2O_3 have become important in the production of E-glass fibers (see ⟶ Fibers, 5. Synthetic Inorganic – Chemical Composition and Properties) for lining melting tanks. These linings have substantially improved furnace stability (i.e., prolonged furnace life).

The high hardness of chromium oxide resulting from its crystal structure is exploited in polishing agents for metals and in brake linings. Addition of a small amount of chromium oxide to magnetic materials of audio and video tapes imparts a self- cleaning effect to the sound heads.

3.1.2.6. *Economic Aspects*

Important producers are American Chrome and Chemicals (USA), Bayer (Germany), British Chrome and Chemicals (UK), and Nihon Denko (Japan).

Statistical data on the consumption of chromium oxide have not been published recently. However, it can be assumed that the world capacity in 1996 was about 72 000 t/a; thereof 43 000 t/a in Eastern Europe and the GUS. The consumption as pigments in 1996 is estimated at 28 000 t

3.1.2.7. *Toxicology and Occupational Health*

Toxicological or carcinogenic effects have not been detected in rats receiving up to 5 % chromium(III) oxide in their feed [244] nor in medical studies performed in chemical plants producing chromium(III) oxide and chromium(III) sulfate [245]. The oral LD_{50} for chromium(III) oxide in the rat is > 10 000 mg/kg; it does not irritate the skin or mucous membranes.

Chromium(III) oxide is not included in the MAK list (Germany), the TLV list (USA), or in the list of hazardous occupational materials of the EC [246]. In practice, this means that chromium(III) oxide can be regarded as an inert fine dust with a MAK value of 6 mg/m^3.

3.1.3. ***Mixed Metal Oxide Pigments***

The term mixed metal oxide pigment denotes a pigment that crystallizes in a stable oxide lattice, and in which the color is due to the incorporation of colored cations in this lattice. Such compounds are regarded as solid solutions. The American “Dry Color Manufacturers' Association” has abandoned the term “mixed metal oxide inorganic colored pigments”, it denotes these colorants as “complex inorganic color pigments” [247].

According to V. M. GOLDSCHMITT, two substances with the same basic formula and crystal structure form solid solutions in a concentration range that depends on the degree of similarity of their ionic radii. A large range of solid solutions may be expected if the radius of the larger ion does not exceed that of the smaller by more than 15 % [248].

Solid solutions can also be formed to a limited extent with substances that have a similar formula but a different crystal structure. One of the substances imposes its structure upon the other. In lattices which are stable even if some sites in the unit cell are unoccupied, substances with different formulae can form solid solutions (anomalous mixed crystals). The incorporated ions must be of similar size, but do not need to have the same valency because electroneutrality is ensured by the presence of a compensating valency elsewhere in the lattice. Changes in the relative concentrations in the solid solutions lead to continuous variation of their properties.

In principle, the number of possible mixed metal oxide phases that can be produced from different substituent elements and oxide lattices is extremely large. Systematic investigations have greatly increased the knowledge about mixed metal oxide phases [249].

The only stable oxide lattices that have so far been of value as mixed metal oxide pigments are those with spinel, rutile, and hematite structures. These lattices possess not only good thermal and chemical stability, but also have a high refractive index which is important for good optical pigment properties.

3.1.3.1. *Properties*

Rutile Mixed Metal Oxide Pigments. The rutile (TiO_2) lattice offers considerable scope for variation (see Section Properties) [249]. In this lattice, titanium ions are surrounded by six neighboring oxygen atoms at the corners of a regular, but slightly distorted octahedron. Colored pigments may be obtained by substituting transition-metal ions for the titanium ions [250]. If the colored cation substituent is not tetravalent, another cation of appropriate valency must be substituted into the rutile lattice to maintain an average cation valency of four.

H. H. SCHAUMANN prepared the first mixed-phase rutile pigments by incorporating nickel oxide and antimony oxide into rutile to give a yellow pigment, and cobalt and antimony oxides to give an ocher pigment [251]. Other rutile mixed metal oxide phases were first used as ceramic colorants (e.g., with chromium – tungsten, brown [252]) or for coloring porcelain enamels (e.g., with copper – antimony, lemon yellow; manganese – antimony, dark brown; iron – antimony, gray [253]; and vanadium – antimony, dark gray). In commercial mixed-phase rutile pigments, between 10 and 20 mol % of the titanium ions are replaced by substituent metal ions.

Nickel and chromium mixed-phase rutile pigments are industrially important. When nickel and chromium are substituted into the rutile lattice, higher valency metals (e.g., antimony, niobium, or tungsten) must also be substituted to maintain an average valency of four.

Nickel rutile yellow [8007-18-9], C.I. Pigment Yellow 53:77788, is a light lemon yellow pigment with the approximate composition $(Ti_{0.85}Sb_{0.10}Ni_{0.05})O_2$. Antimony can be replaced by niobium without any appreciable change in color [254].

Chromium rutile yellow pigments [68186-90-3], C.I. Pigment Brown 24:77310, usually contain antimony to balance the valency. Their composition is approximately $(Ti_{0.90}Sb_{0.05}Cr_{0.05})O_2$. Depending on particle size, the color varies from light to medium ocher (buff). Stability towards plastics at high temperature is considerably improved by the incorporation of small quantities of lithium [255] or magnesium [256]. If antimony is replaced by tungsten, the products become darker in color, whereas replacement by niobium leaves the color unchanged.

Spinel Mixed Metal Oxide Pigments. The crystal structure of spinel, $MgAl_2O_4$, is made up of an approximately cubic closely packed array of oxygen ions (Fig. 26). The unit cell contains 32 oxygen ions. Of the 64 possible tetrahedral sites, 8 are occupied by magnesium ions; of the 32 possible octahedral sites, 16 are occupied by aluminum ions.

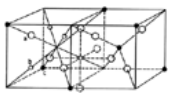

Figure 26. Structure of spinel showing two octants of the unit cell

Two oxygen ions of neighboring octants are included (dashed circles) to show the arrangement of oxygen around the octahedral sites.

a) Oxygen; b) Cations occupying octahedral sites; c) Cations occupying tetrahedral sites

The inverse spinel structure differs in that one type of cation occupies the tetrahedral and half of the octahedral sites of the spinel lattice, and the other cations occupy the remaining octahedral sites. This is indicated by writing the formula of, for example, zinc titanium spinel as $Zn(TiZn)O_4$. Spinels containing di- and tetravalent cations are mostly of the inverse type. Normal and inverse spinels should be regarded as idealized limiting structures, intermediate forms are often observed in practice. A large range of spinel phases can be formed due to the numerous combinations of cations in the oxidation states 1 – 6, especially from elements of the first transition-metal series. This range is increased if some of the cation sites remain unoccupied. Further details are given in [257].

Many spinel phases have been used as colorants for ceramics and porcelain enamels (⟶ Ceramic Colorants). The most important pigments based on spinel phases used in paints, plastics, and building materials are given in Table 25.

Table 25. Mixed-phase spinel pigments

Trivial name	Formula	CAS registry no.	C.I. name and no.	Color	Crystal structure
Cobalt blue	$CoAl_2O_4$	[1345-16-0]	Pigment Blue 28:77 346	reddish blue	spinel
	$Co(Al,Cr)_2O_4$	[68187-11-1]	Pigment Blue 36:77 343	greenish blue	spinel
Cobalt green	$(Co,Ni,Zn)_2TiO_4$	[68186-85-6]	Pigment Green 50:77 377	green	inverse spinel
Zinc iron brown	$ZnFe_2O_4$	[68187-51-9]	Pigment Yellow 119:77 496	light to medium brown	spinel
Spinel black	$Cu(Cr,Mn)_2O_4$	[68186-91-4]	Pigment Black 28:77 428	black	spinel
	$Cu(Fe,Cr)_2O_4$	[55353-02-1]	Pigment Black 22:77 429		intermediate between normal and inverse spinel

Cobalt Blue. This was formerly known as Thénard's blue, and is a cobalt aluminum spinel. It usually has a deficiency of cobalt (ca. 0.6 – 0.7 mol CoO/mol Al_2O_3) compared with the ideal formula $CoAl_2O_4$. It has a brilliant blue hue with a red tinge. The tetrahedral holes can be filled by incorporating zinc or magnesium ions in the lattice. If chromium is used in place of aluminum, the hue is shifted via neutral blue (molar ratio Al:Cr≈7:3) to a greenish blue (molar ratio Al:Cr≈3:7). Brilliance decreases with increasing chromium content, while the hiding power increases.

Cobalt Green. Incorporation of cobalt in the inverse titanium spinels Mg_2TiO_4 and Zn_2TiO_4 produces blue to green hues. The green color results when cobalt ions occupy the octahedral sites [258]. Both the blue and the green pigments have patent protection [259], and correspond to inverse spinels with the general formula $Mg_aCo_bZn_cTiO_4$ (where $a+b+c = 2$). A brilliant cobalt green is obtained if magnesium is replaced by nickel, giving the approximate stoichiometry $NiCo_{0.5}Zn_{0.5}TiO_4$ [260]. Turquoise spinels are obtained by including lithium, $Li_2(Co,Ni,Zn)Ti_3O_8$ [261].

Zinc Iron Brown. Variation of the zinc to iron ratio in red-brown iron spinel ($ZnFe_2O_4$), or replacement of some of the iron ions by aluminum and titanium ions, gives light to medium brown pigments. Inclusion of a small proportion of lithium ions considerably improves stability towards reducing agents, e.g., when the pigments are used to color plastics [262]. The partial replacement of iron ions by chromium(III) ions yields dark brown products [263].

Spinel Black. Iron – chrome spinels of composition $CuFe_{0.5}Cr_{1.5}O_4$ provide black pigments with good tinting strength. Other commercial black pigments are copper chromites, $CuCr_2O_4$, in which part of the chromium may be replaced by manganese.

Mixed Metal Oxide Pigments with Hematite and Bixbyite Structures. Manganese and chromium can be incorporated in the hematite lattice of α-Fe_2O_3 which has a higher refractive index than rutile. Brown, temperature-resistant iron oxide pigments with the hematite structure are described in Section Synthetic Iron Oxide Pigments, e.g., α-$(Fe,Mn)_2O_3$ [264] and α-$(Fe,Cr)_2O_3$ [265]. The bixbyite lattice of α-Mn_2O_3 can take up > 50 % Fe_2O_3 in solid solution while retaining its structure [266]. Heat-stable black pigments with good tinting strength are obtained.

3.1.3.2. *Production*

The mixed metal oxide pigments are usually produced by reaction of the components in the solid state at 800 – 1400 °C. The reactions proceed more readily if the components are reactive, finely divided, and intimately mixed. The reactants are therefore usually extremely finely divided oxides, hydroxides, carbonates, or nitrates that are mixed in the form of aqueous suspensions. As these raw materials are commercially available with great purity and fineness, coprecipitation of hydroxides or carbonates from aqueous salt solutions is generally not required. Rutile mixed metal oxide phases based on titanium dioxide are prepared from the reactive anatase modification or titanium dioxide hydrolysate, an intermediate in titanium dioxide production.

Wet mixes are usually dried before calcination. Calcination is performed continuously in rotary or tunnel kilns, or batchwise in directly fired drum or box furnaces. The temperature at which the mixed metal oxide pigments are formed can be reduced by adding mineralizing agents [250]. In the case of chromium rutile pigments, addition of magnesium compounds [256] or lithium compounds [255] before calcination improves thermal stability in plastics.

The formation of rutile mixed phases in the system Ti – Ni – Sb – O is described in detail in [267]. When nickel ions are to be included in the pigment and antimony trioxide is used to provide the antimony ions, oxygen must be present to ensure that nickel and antimony occupy the titanium sites in the lattice. If oxygen is excluded, only small amounts of nickel and antimony are taken up in solid solution, and no useful products are formed. The presence of oxidizing agents (e.g., nitrates) is necessary for the production of antimony- containing mixed metal oxide pigments.

After calcination, the products are ground to the desired fineness. Calcined products containing salts are subjected to wet grinding followed by washing, drying, and dry grinding. With salt-free calcination products, washing of the pigments can be omitted. and dry grinding is possible.

When mixed-phase rutile pigments are used in special paint systems, (e.g., stoving or acid-catalyzed lacquers), inorganic surface treatment in an aqueous medium can improve the gloss and flocculation properties. For example, an aqueous pigment suspension is first treated with a surfactant, and then coated with metal hydroxides or oxide hydrates [268].

Chromium rutile pigments with extremely high resistance to flocculation are obtained by applying a double-layer coating of oxides or oxide hydrates formed from tetravalent metals and oxides or oxide hydrates of aluminum [269].

3.1.3.3. *Quality Specifications and Analysis*

There are as yet no national or international (ISO) standards for mixed metal oxide pigments, and none are expected in the foreseeable future. However, the general tests of chemical, physical, and optical properties described in ASTM, DIN, or ISO standards for pigments may be utilized, such as DIN 6174, DIN 55 986, and ISO 787 Part 16 for the optical properties, and ISO 787 Parts 1 – 20 for the physical properties (see also Table 1).

There are no standard methods for the chemical analysis of mixed metal oxide pigments. They are first decomposed (e.g., with peroxide in the case of rutile mixed phases, or soda borax for the spinels) and the appropriate methods are then used to determine the elements.

3.1.3.4. *Storage and Transportation*

Mixed metal oxide pigments are thermally stable, water-insoluble materials. They are not classified as hazardous substances, and are therefore not subject to international transport regulations. When stored under dry conditions their pigment properties do not deteriorate.

3.1.3.5. *Legal Aspects*

Mixed metal oxide pigments are permitted in almost all countries for coloring paints and plastics in contact with foods, or for toy manufacture [270-274]. The legal limits for a number of elements (sometimes for the soluble fraction only) must be observed. Owing to the large number of raw materials and processes used, general purity requirements cannot be specified. Each manufacturer must supply relevant data.

In the EC guidelines for coloring agents for foods [275], cosmetics [276], or pharmaceuticals [277], there are no references to mixed metal oxide pigments, and these are therefore not permitted.

3.1.3.6. *Uses*

The outstanding properties of the mixed metal oxide pigments are their lightfastness and resistance to temperature, chemicals (including acid and alkali), and weathering.

Owing to their high scattering power, their hiding power is much better than that of organic pigments. They are therefore used in combination with organic pigments as base pigments for colored coatings, e.g., for lead-free and chromate-free automobile topcoats (nickel rutile yellow, chromium rutile yellow).

The stability of these pigments makes them suitable for all types of coatings, especially those which have to meet high standards of lightfastness and resistance to weathering or chemicals. They are especially suitable for baking enamels, coil coatings, powder coatings, as well as for coloring plastics, masonry coatings based on resin emulsions or waterglass, and building materials.

A very important use for rutile yellow is in coil coating of aluminum and steel, especially in the building industry, but also for containers, vehicles, and machinery.

Rutile yellow is used in rigid poly(vinyl chloride) for outdoor use (USA: vinyl siding); cobalt blue and cobalt green are employed to color high-density polyethylene (bottle crates). Rutile yellow is being used increasingly in engineering plastics to replace cadmium pigments, which have unfavorable toxicological and ecological properties. Rutile yellow pigments can be used cheaply and advantageously for this purpose instead of a combination of titanium dioxide and expensive organic colorants. Nickel and chromium rutiles can replace cadmium yellow as toning pigments. Rutile yellow pigments can be safely used in all plastics. However, chromium rutiles are only resistant up to 280 °C, though this disadvantage can be overcome by lattice stabilization (see Rutile Mixed Metal Oxide Pigments.).

Trade names of some mixed metal oxide pigments include Heucodur (Dr. H. Heubach, Germany); Irgacolor (Ciba-Geigy, Switzerland); Lichtechtpigmente (Bayer, Germany); Sicotan-Gelb, Sicopal-Blau (BASF, Germany); Tipaque Yellow (Ishihara, Japan); Kerafast (Blythe Colours, UK); and Meteor Colors (Engelhard, USA). Ferro (USA), another important producer of these pigments, does not use a specific trade name.

3.1.3.7. *Economic Aspects*

No statistical information concerning the use of mixed metal oxide pigments has been published. However, it is estimated that world consumption, excluding the former Eastern bloc and China, was ca. 15 000 t in 1996. There is no definite information available on whether or not production of mixed metal oxide pigments takes place in the former Eastern bloc countries or China.

3.1.3.8. *Toxicology and Occupational Health*

Mixed metal oxide pigments contain one or more heavy metals, whose oxides can have toxicological effects. However, these heavy metals in the mixed metal oxide phase occur as different compounds, even if the respective oxides are used as raw materials for pigment production. They have an extremely low solubility in water and dilute acids. Toxicological investigations with nickel and chromium rutile yellow have shown that, even at high dose rates (up to 1 % in animal feeds), these pigments are not toxic or biologically available [278]. Cobalt spinels, unlike cobalt oxide, do not have carcinogenic properties [279]. Also, the waste disposal of paints and plastics colored with mixed metal oxide pigments is regarded as safe [280].

Only nickel rutile yellow has so far been included in the MAK list in Class II b, edition 1996; toxicological data available for other pigments are inadequate. The mixed metal oxide pigments are not included in the EC guidelines for hazardous materials (67/548/EC) or in the TLV list for 1990/1995.

3.2. Cadmium Pigments

Among the inorganic pigments, cadmium pigments have particularly brilliant red and yellow colors, as well as a high durability. All cadmium pigments are based on cadmium sulfide and crystallize with a hexagonal wurtzite lattice (Fig. 27). The sulfur atoms have a hexagonal, tightly packed arrangement, in which the cadmium atoms occupy half of the tetrahedral holes. In this wurtzite lattice the cations and anions can be replaced within certain limits by chemically similar elements; nevertheless, the only substitutions which are used in practice are zinc and mercury for the cations, and selenium for the anions.

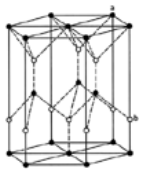

Figure 27. Crystal lattice of cadmium pigments (wurtzite structure)

a) Sulfur (selenium); b) Cadmium (zinc, mercury)

The use of zinc yields greenish yellow pigments due to the smaller lattice constants; mercury and selenium lead to expansion of the lattice. With an increasing content of selenium, or especially mercury, the shades of the pigments change to orange, red, and ultimately to deep red (bordeaux). The brilliant colors of cadmium pigments are primarily due to their almost ideal reflectance curves with a steep ascent (Fig. 28).

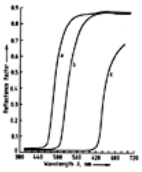

Figure 28. Reflectance curves of cadmium pigments (pigment volume concentration 10 %)

a) Cadmium yellow; b) Cadmium golden yellow; c) Cadmium red (bordeaux)

3.2.1. *Cadmium Sulfide*

Cadmium sulfide [1306-23-6], CdS, M_r 144.48, occurs as greenockite or cadmium blende in several natural deposits, which are, however, of no importance as pigments. The mineral crystallizes hexagonally in the wurtzite lattice (α-form).

When hydrogen sulfide is passed into cadmium salt solutions, cadmium sulfide is formed as a yellow precipitate with a zinc blende structure (cubic, β-form). The β-form can be converted into the α-form (e.g., by heating). α- Cadmium sulfide shows photoconductivity due to defects in the crystal lattice (usage in photovoltaic cells) [281]. The solubility in water at 25 °C is 1.46×10^{-10} mol/L [282]. Cadmium sulfide forms the basis for all cadmium pigments.

3.2.2. *Cadmium Yellow*

Cadmium yellow consists of pure cadmium sulfide (golden yellow color) or mixed crystals of zinc and cadmium sulfide [8048-07-5], (Cd, Zn)S, in which up to one-third of the cadmium can be replaced by zinc. The density of this pigment is 4.5 – 4.8 g/cm^3 and its refractive index is 2.4 – 2.5. The prevalent particle size is approx. 0.2 µm with cubic to spheroidal habits. Cadmium yellow is practically insoluble in water and alkali, and of low solubility in dilute mineral acid. It dissolves in concentrated mineral acid with generation of hydrogen sulfide.

Production. The raw material for the production of cadmium yellow pigments is high-purity cadmium metal (99.99 %), cadmium oxide, or cadmium carbonate. If the metal is used it is first dissolved in mineral acid. A zinc salt is then added to the solution in an amount that depends on the desired shade. This is followed by addition of sodium sulfide solution. An extremely finely divided cadmium sulfide or cadmium zinc sulfide precipitate is formed which does not possess any pigment properties. This intermediate product can also be obtained by mixing the cadmium or cadmium – zinc salt solution with sodium carbonate solution. An alkaline cadmium carbonate or cadmium zinc carbonate precipitate is formed which reacts in suspension with added sodium sulfide solution.

The crude precipitated cadmium yellow is washed and then calcined at approx. 600 °C, at which temperature the cubic crystal form changes to the hexagonal form. This process determines the particle size distribution, which is essential for the pigment properties.

If the starting material is cadmium oxide or cadmium carbonate, it is mixed with sulfur and calcined at approx. 600 °C. After calcination, the pigment product is washed with dilute mineral acid to remove any remaining soluble salts, dried, and ground.

3.2.3. *Cadmium Sulfoselenide (Cadmium Red)*

Pure cadmium selenide [1306-24-7], CdSe, is brownish black and has no pigment properties. Like cadmium sulfide, it is dimorphous and occurs in hexagonal and cubic modifications. Cadmium selenide is insoluble in dilute acid. It readily liberates hydrogen selenide in concentrated hydrochloric acid. It dissolves completely in fuming nitric acid, the Se^{2-} ions being converted to SeO_4^{2-} ions. Cadmium selenide is an *n*-type semiconductor.

Cadmium red consists of cadmium sulfoselenide [12656-57-4], [58339-34-7], Cd(S,Se), and is formed when sulfur is replaced by selenium in the cadmium sulfide lattice. With increasing selenium content, the color changes to orange, red, and finally dark red. The density of these pigments increases correspondingly from 4.6 to 5.6 g/cm^3 and the refractive index from 2.5 to 2.8. The crystals have cubic or spheroidal habits, and the prevalent particle size is 0.3 – 0.4 µm.

Production. Cadmium red pigments are produced in a similar way to the cadmium yellow pigments. The cadmium salt solution is prepared by dissolving the metal in mineral acid and then sodium sulfide is added. A certain amount of selenium powder is dissolved in the sodium sulfide solution to obtain the desired color shade. In an alternative procedure, the cadmium solution is mixed with sodium carbonate solution to precipitate cadmium carbonate, which is reacted with the selenium- containing sodium sulfide solution.

The cadmium red pigment intermediate is obtained as a precipitate and is filtered off, washed, and calcined at approx. 600 °C. As with cadmium yellow, calcination yields the red pigment and determines the particle size, particle size distribution, and color shade. Analogously to the cadmium yellow process, cadmium red can be produced by direct reaction of cadmium oxide or cadmium carbonate with sulfur and the required amount of selenium at ca. 600 °C.

Products in which selenium is totally or partially replaced by tellurium [283], [284] have not had any commercial application owing to their poor coloristic properties.

3.2.4. *Cadmium Mercury Sulfide (Cadmium Cinnabar)*

The cadmium in the wurtzite lattice of cadmium sulfide can be replaced by divalent mercury to give cadmium mercury sulfide (cadmium cinnabar). As the quantity of mercury increases, the lattice expands, and the color deepens from yellow to orange and finally dark red. The coloring properties of cadmium cinnabar resemble those of the selenium- containing cadmium red pigments. Use of these mercury- containing pigments can be justified for economic reasons. However, their use is no longer justifiable from the ecological point of view because adequate nontoxic substitutes are available.

3.2.5. *Properties and Uses*

Cadmium pigments are mainly used to color plastics. They have brilliant, pure shades (yellow, orange, red, and bordeaux), good hiding power, and moderate tinting strength.

Their process-relevant properties are determined by their high thermal stability and chemical resistance to aggressive additives or to molten plastics that have a reducing action (e.g., polyamides). Other advantages for use in plastics are good lightfastness, weather resistance, and migration resistance. The cadmium pigments also protect polyolefins against aging by absorbing UV light. Premature embrittlement is prevented and the polymer can be recycled [285].

A very useful feature of plastics colored with cadmium pigments is the dimensional stability of injection moldings with a large surface area. This combination of properties is not matched by any other class of colorant.

For economic reasons, cadmium pigments are no longer used for coloring plastics where the pigment property requirements are less stringent. In poly(vinyl chloride) and to a large extent in low-density polyethylene, they have been replaced by less expensive inorganic and organic pigments.

For ecological reasons, cadmium pigments are increasingly being replaced in plastics for which high demands are made on the colorants, sometimes with concessions as regards product quality. Plastics processed at high temperature are now colored with high-quality organic pigments (e.g., perylene, quinacridone, and high-quality azo pigments), often in combination with inorganic pigments, which provide the required covering power. With increasing processing temperature and for outdoor use, the amount of substitution product is lower. Some compromises in processability and product quality, including shade and brilliance have to be accepted.

A further use of cadmium pigments is in paints and coatings (powder paints, silicone resins, and automotive topcoats), but this is declining.

Cadmium pigments are irreplaceable for coloring enamels, ceramic glazes (⟶ Ceramic Colorants), and glass to the highest brilliance. Brilliant transparent glasses are obtained by adding a small amount of cadmium red, while 10 % cadmium red with a cadmium oxide stabilizer produces dark decorative glasses. Other areas of application are ceramics (wall and floor tiles, household and decorative ceramic ware). Very small amounts of cadmium pigments are used for artists' colors. The terms cadmium yellow and cadmium red have become synonyms for very brilliant red and yellow shades.

Cadmium pigments, especially cadmium red, are very sensitive to intensive grinding, which causes loss of brilliance due to an increase in the number of irregular lattice defects. A brilliant red shade may become a dirty brownish red.

The cadmium pigments are lightfast but, like all sulfide pigments, are slowly oxidized to soluble sulfates by UV light, air, and water. This photooxidation is more pronounced with cadmium yellow than with cadmium red and can still be detected in the powder pigment, which normally contains 0.1 % moisture.

Regulations Regarding Foods. The cadmium pigments fulfil the legal requirements of the EU countries for colorants used in plastics which come in contact with food.

Except for polyamide 6, only microgram quantities of cadmium can be extracted from colored plastics into simulants [286], [287]. Cadmium pigments are no longer listed in the French positive list.

To prevent even minimal intake of cadmium, the European Commission has passed a resolution which states that cadmium pigments should only be used for these purposes if adequate substitutes are not available [288].

The use of cadmium pigments in ceramics is controlled in the EU Guideline No. 84/500/EC. International standards [Part I (Methods) and Part II (Limit Values)] are as follows: ceramic surfaces ISO 6486, enamel ISO 4531, and glass ISO 7086.

3.2.6. *Quality Specifications*

The quality requirements and conditions for the supply of cadmium pigments given in ISO 4620 have lost their significance in practice. According to this standard, cadmium pigments must be free from organic colorants, inorganic colored pigments, and other additives. The acid-soluble components (0.1 mol/L HCl) should not exceed 0.2 % Cd (determined according to ISO 4620 and ISO 3856/4); in practice, they are usually < 0.01 % Cd.

Cadmium pigments are sold as homogeneous powders and as preparations mixed with barium sulfate to give the required tinting strength. To reduce the risk of inhalation (see Section Toxicology and Environmental Protection), they are supplied as low-dust powders and fine granules.

In dust-free and dispersed form, they are supplied as concentrated plastic granules (masterbatch pellets), as concentrated pastes, and as liquid colors. These products are added at different stages in the processing of plastics.

3.2.7. *Economic Aspects*

The output of cadmium pigments in the industrialized countries in 1995 was approximately 3000 t, of which about 2000 t was produced in Europe and 500 t in the United States. Consumption is distributed approximately as follows: plastics 90 %, ceramics 8 %, and others 2 %.

Manufacturers and *trade names* of cadmium pigments include Cerdec AG, Frankfurt, Germany; Orkem, Narbonne, France (Langdopec Pigments); J. M. Brown, Stoke-on-Trent, UK ; Cookson-Matthey, Kidsgrove, UK ; Ferro Corporation, Color Division, Cleveland, USA ; and Harshaw Chemical Co., Cleveland, USA.

3.2.8. *Toxicology and Environmental Protection*

Production. Since soluble cadmium compounds have a toxic effect on humans and the environment, all wastewater originating from the production of cadmium pigments must be treated to remove cadmium. This is carried out by methods outlined elsewhere (⟶ Cadmium and Cadmium Compounds – Waste Water). In the EC, the following limits for the pigment production must be observed: 0.2 mg Cd/L in wastewater and 0.3 kg Cd per tonne of used cadmium compounds, calculated as metal [289].

The exhaust air from the production plants and the ventilation equipment must be cleaned. Tubular filters and, more recently, absolute filters have been used to remove cadmium compounds in dust form. With these high-performance filters, it is possible to comply with the limits specified in Germany (0.2 mg Cd air, 1 g Cd/h, emission limit; 5 µg Cd m^{-2} d^{-1}, immission limit). During pigment production, occupational hygiene and safety measures for toxic materials must be observed. Exposure to cadmium can be determined by measurements of the workplace concentration and by examination of the cadmium levels in blood and urine. In Germany, the following maximum BAT values are permitted in the workplace: 15 µg Cd per liter urine and 1.5 µg Cd per 100 mL blood. The TRK value for cadmium pigments is 0.015 mg/m^3. In the United States, the TLV value for cadmium and its compounds is 0.01 mg Cd/m^3. The OSHA Cadmium Standard has a PEL value of 0.005 mg Cd/m^3.

Toxicity. The toxicity of soluble cadmium compounds is described elsewhere (⟶ Cadmium and Cadmium Compounds – Toxicology and Occupational Health). Cadmium pigments are cadmium compounds of low solubility. However, in dilute hydrochloric acid (concentration equivalent to stomach acid) small quantities of cadmium are dissolved, and long-term oral intake of cadmium pigments can lead to accumulation in the human body. On inhalation of subchronic amounts of cadmium pigments, a small proportion of cadmium is biologically available [290], [291]. Cadmium pigments have no acute toxic effect (oral LD_{50}, rat, > 10 g/kg). The pigments do not have any adverse effects on the skin and mucous membranes.

Genotoxic Effects and Carcinogenicity. Tests for DNA damage [292] and cell transformation caused by crystalline cadmium sulfide were positive [293]. Cadmium sulfide also proved to be carcinogenic by intraperitoneal and after intratracheal administration [294]. The significance of such animal studies is being controversially discussed by toxicologists.

Long-term animal feeding studies with various cadmium compounds show no carcinogenic risks. However, in inhalation studies with cadmium oxide, cadmium chloride, cadmium sulfate, and cadmium sulfide in rats, mice, and hamsters, all four compounds produced a significantly increased incidence of lung cancer in rats [295]. The results for mice were inconclusive, and no carcinogenicity was observed in hamsters [296]. Reinvestigation of the inhalation studies has shown that the test was not applicable to cadmium sulfide. The inhaled liquid cadmium sulfide suspension contained up to 40 % cadmium sulfate formed by light-induced oxidation.

In 1989 the German Senate Commission for the Assessment of Hazardous Substances classified cadmium and its compounds (including cadmium sulfide) among those substances whose carcinogenicity has been clearly proven in animal studies under conditions comparable to potential human exposure in the workplace (Group III A 2). According to the MAK classification, cadmium pigments and preparations that contain cadmium pigments in concentrations above 0.1 % must therefore be regarded as carcinogenic if they occur in an inhalable form. In the United States, the National Toxicology Program (NTP) has shown that cadmium and various cadmium compounds (including cadmium sulfide) have a carcinogenic effect in animal experiments. In 1991 the EU reported the above-mentioned error in the inhalation studies and classified cadmium sulfide in the EU Cancer List in Category III (suspected carcinogen). In the EU pure cadmium sulfide is labelled with the symbol T (Toxic) and the R phrases 22–40–48/23/25 (harmful if swallowed; possible risks of irreversible effects; toxic: danger of serious damage to health by prolonged exposure through inhalation and if swallowed). Cadmium compounds must be labelled with the symbol Xn (Harmful) and the Risk Phrases 20/21/22 (harmful by inhalation, in contact with skin, and if swallowed). The cadmium pigments (cadmium zinc sulfide and cadmium sulfide selenide) are exempted from labelling. The cadmium pigments are slightly soluble in dilute acids, and the dissolved cadmium is bioavailable. Therefore, some cadmium pigment producers label their cadmium zinc sulfide and cadmium sulfide selenide pigments like the cadmium compounds.

Limitations of Use. In 1981 Sweden prohibited the use of cadmium pigments (with some exceptions) for ecological reasons. In 1987 Switzerland prohibited the use of cadmium pigments in plastics. Exceptions are possible if valid reasons can be given.

In 1991 the EC passed a directive to prohibit the use of cadmium pigments as colorants for certain plastics that can easily be colored with other pigments [297]. A transition period was granted for further series of plastics, which may not be colored with cadmium pigments since January 1, 1996. The maximum cadmium content in plastics is limited to 0.01 %, a range which includes typical industrial impurity levels but in which technically convenient coloring is not practicable. Use in coating media is also prohibited since January 1, 1996. Use in artists' colors and ceramic products (including coatings used for ceramic products) is still permitted.

The directive determined that the prohibited lists shall be revised and possibly extended after five years. Revision is due in 1997, and it is expected that the limit value of 0.01 % will be lowered to 0.005 % Cd.

In 1989 the CONEG Model (Coalition of North-Eastern Governors) was passed by 14 U.S. States to reduce the environmental pollution by environmentally relevant heavy metals. They decided to limit the total content of the environmentally relevant heavy metals lead, cadmium, mercury, and hexavalent chromium. Since then, the other U.S. States have accepted the model. In the United States, the total content of the four elements must not exceed 100 ppm in packaging since 1994.

These four heavy metals have been included in the EU directive for packaging and packaging waste (94/62/EEC). The limitations on certain heavy metals becomes effective in 1998. The total content of the four elements will be reduced in three steps:

From July 1, 1998 600 ppm
From July 1, 1999 250 ppm
From July 1, 2001 100 ppm

Furthermore, production, handling, or putting into circulation of packaging colored with cadmium pigments will no longer be permitted in the EU. For cadmium-containing returnable crates in circulation, an exception rule passed in 1997.

3.3. Bismuth Pigments

Bismuth- containing special-effect pigments based on platelet-shaped crystals of bismuth oxide chloride (bismuth oxychloride, BiOCl) have been known for a long time (see Section Bismuth Oxychloride). More recently, greenish yellow pigments based on bismuth orthovanadate [14059-33-7], $BiVO_4$, M_r 324, have attracted increasing interest.

They represent a new class of pigments with interesting coloristic properties, extending the familiar range of yellow inorganic pigments (iron yellow, chrome yellow, cadmium yellow, nickel titanium yellow, and chromium titanium yellow).In particular they are able to substitute the greenish yellow shades of the lead chromate and cadmium sulfide pigments.

Historical Aspects. Bismuth vanadate occurs naturally as the brown mineral pucherite (orthorhombic), as clinobisvanite (monoclinic) and as deyerite (tetragonal). Its synthesis was first reported in 1924 in a pharmaceutical patent [298]. The development of pigments based on $BiVO_4$ began in the mid-1970s. In 1976, Du Pont described the preparation and properties of "brilliant primrose yellow" monoclinic bismuth vanadate [299]. Montedison developed numerous pigment combinations based on $BiVO_4$ [300]. Pigments containing other phases besides $BiVO_4$, e. g. Bi_2XO_6 (X= Mo or W), have been reported by BASF [301] and led to the first commercial product (trade name Sicopal Yellow L 1110). Since then Bayer [302], Ciba-Geigy [303], BASF [304] and others have published further methods for the manufacture of pigments based on $BiVO_4$. Pigments for use in paints and plastics are now commercially available from the following suppliers: BASF, Ciba-Geigy, Bayer, Capelle, Bruchsaler Farbenfabrik, and Heubach. Recently, BASF introduced a new reddish yellow bismuth vanadate pigment on the market (trade name Sicopal Yellow L 1600).

3.3.1. *Properties*

Most commercial bismuth vanadate pigments are now based on pure bismuth vanadate with monoclinic or tetragonal structure, though there are still pigments based on the two-phase system bismuth vanadate molybdate, $4\ BiVO_4 \cdot 3\ Bi_2MoO_6$. The pure bismuth vanadate pigments show a higher chroma than the two-phase systems. In the following, the physical and coloristic properties of a pure bismuth vanadate pigment are given (Sicopal Yellow L1100, BASF):

Density	6.1 g/cm^3
Refractive index *n*	2.45
Specific surface area (BET)	13 m^2/g
Oil absorption	33 g/100 g pigment
Composition:	
Bi	64.51 wt %
V	15.73 wt %
O	19.76 wt %

Bismuth vanadate, C.I. Pigment Yellow 184, is a pigment with a greenish yellow color, having high tinting strength, high chroma, and very good hiding power. When compared with other yellow inorganic pigments, it most closely resembles cadmium yellow and chrome yellow in its coloristic properties (Fig. 29). Bismuth vanadate shows a sharp increase in reflectance at 450 nm and considerably higher chroma than iron yellow or nickel titanium yellow. It has very good weather resistance both in full shade and in combination with TiO_2. Pigment properties follow:

Hue angle H° (HGD, hue grade; CIELAB)	93.7
Chroma *C*ab* (CIELAB)	95.5
Hiding power over black/white (42.5 wt % in dry film)	ca. 70 µm
Weather resistance (DIN 54 002)	
Full shade	4 – 5
Mixed with TiO_2, 1 : 10	4 – 5
Chemical resistance in cross-linked paint films	
Acid	5 (2 % HCl)
Alkali	5 (2 % NaOH)
Heat stability	> 200 °C

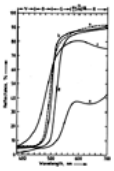

Figure 29. Reflectance curves of yellow pigments

a) $BiVO_4$; b) CdS; c) $(Ti,Ni,Sb)O_2$; d) $PbCrO_4 – PbSO_4$; e) FeOOH

3.3.2. *Production*

The production process for bismuth vanadate pigments consists usually of a precipitation reaction followed by a calcination step. The calcination step varies from producer to producer and may even be omitted, depending on the precipitation process and the desired product properties.

$$\underset{\text{in } HNO_3 \text{ solution}}{Bi(NO_3)_3 + NaVO_3 + Na_2MoO_4} \xrightarrow{NaOH} \underset{\text{Gel particle}}{2\ Bi_2O_3 \cdot V_2O_5 \cdot MoO_3 \cdot x\ H_2O} + NaNO_3 \qquad (4)$$

$$\underset{\text{Gel particles}}{2\ Bi_2O_3 \cdot V_2O_5 \cdot MoO_3 \cdot x\ H_2O} \xrightarrow{600\,^\circ C} \underset{\text{Polycrystalline particles}}{BiVO_4 \cdot Bi_2MoO_6} + H_2O \qquad (5)$$

In this process a fine precipitate is first formed by adding an alkaline solution of sodium or ammonium vanadate to an acidic bismuth nitrate solution in the presence of a considerable amount of phosphate, or in the inverse process by adding the bismuth nitrate solution to a sodium or ammonium vanadate solution (Fig. 30). Thereafter, the pH is set between 5 and 8 with sodium hydroxide and the precipitate crystallizes to monoclinic or tetragonal bismuth orthovanadate, usually by heating to reflux. After crystallization the pigment is often coated with calcium or zinc phosphate or aluminum oxide to improve the lightfastness and weather resistance. The final step is usually calcination followed by wet milling and spray drying (Fig. 31). The spray drying process produces a finely granulated, dust-free pigment. For use in plastics the pigment is additionally coated with silica and other components to increase its heat stability in certain polymers, e.g. polyamide, up to 320°C.

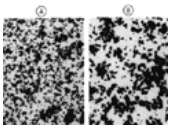

Figure 30. Particle size after precipitation (A) and subsequent heating (B) (magnification×6000)

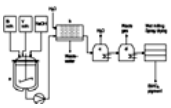

Figure 31. Flow diagram for the production of bismuth vanadate pigment

a) Reaction vessel; b) Filter; c) Dryer; d) Furnace treatment

3.3.3. *Uses*

Bismuth vanadate pigments are used in the manufacture of lead-free, weather resistant, brilliant yellow colors for automotive finishes and industrial paints. They are suitable for the pigmentation of solvent-containing paints, water-based paints, powder coatings, and coil-coating systems. They can be mixed with other pigments to produce brilliant colors in the orange, red and green regions, such as the German standard colors RAL 1021, 1028, 2004, 3020, 6018, and 6029. Various producers have developed bismuth vanadate pigments for use in plastics, especially polyethylene and technical plastics. The annual production of bismuth vanadate pigments is ca. 500 t. In the near future an increasing market volume of 1000 to 2000 t/a is expected.

3.3.4. *Toxicology*

Bismuth vanadate pigments do not exhibit acute toxicity upon inhalative or oral incorporation. They show chronic toxicity on inhalation due to the vanadium content. The NOEL for rats is 0.1 mg/m^3 (exposure: 3 months, 6 h/d, 5 d/week). The critical factor for the inhalation toxicity is the true amount of pigment in the lungs, and not just the concentration in air. Because of the high density of the pigment and because the supply form is a fine granulate, the risk of inhalation is very low. Therefore the dust-free pigments can be handled under the usual hygienic working conditions [305].

3.4. Chromate Pigments

The most important chromate pigments include the lead chromate (chrome yellow) and lead molybdate pigments (molybdate orange and molybdate red) whose colors range from light lemon yellow to reds with a blue hue. Chrome yellow, molybdate orange, and molybdate red are used in the production of paints, coatings, and plastics, and are characterized by brilliant hues, good tinting strength, and good hiding power. Special treatment of the pigments has allowed continual improvement of their resistance to light, weathering, chemicals, and temperature.

The chromate pigments are also combined with blue pigments (e.g., iron blue or phthalocyanine blue) to obtain high-quality chrome green and fast chrome green pigments. Molybdate orange and molybdate red pigments are often combined with red organic pigments, giving a considerable extension of the color range.

Lead chromates, lead molybdates, chrome greens, and fast chrome greens are supplied as pigment powders, low-dust or dust-free preparations, or as pastes. For standard specifications, see Table 1.

The anticorrosive pigments zinc chromate, zinc tetraoxychromate, and strontium chromate are described in Section Chromate Pigments

3.4.1. *Chrome Yellow*

The chrome yellow pigments [1344-37-2], C.I. Pigment Yellow 34:77600 and 77603, are pure lead chromate or mixed-phase pigments with the general formula $Pb(Cr,S)O_4$ [306] (refractive index 2.3 – 2.65, density ca. 6 g/cm^3). Chrome yellow is insoluble in water. Solubility in acids and alkalis and discoloration by hydrogen sulfide and sulfur dioxide can be reduced to a minimum by precipitating inert metal oxides on the pigment particles.

Both lead chromate and lead sulfochromate (the latter is a mixed-phase pigment) can be orthorhombic or monoclinic; the monoclinic structure is the more stable [307]. The greenish yellow orthorhombic modification of lead chromate is metastable at room temperature, and is readily transformed to the monoclinic modification under certain conditions (e.g., concentration, pH, temperature). The latter modification occurs naturally as crocoite.

Partial replacement of chromate by sulfate in the mixed-phase crystals causes a gradual reduction of tinting strength and hiding power, but allows production of the important chrome yellows with a greenish yellow hue.

Production. In large-scale production, lead or lead oxide is reacted with nitric acid to give lead nitrate solutions, which are then mixed with sodium dichromate solution. If the precipitation solutions contain sulfate, lead sulfochromate is formed as a mixed-phase pigment. After stabilization the pigment is filtered off, washed until free of electrolyte, dried, and ground.

The color of the pigment depends on the ratio of the precipitating components and other factors during and after precipitation (e.g., concentration, pH, temperature, and time). According to WAGNER [308], the precipitated crystals are orthorhombic, but change very readily to the monoclinic form on standing; higher temperatures accelerate this conversion. Almost isometric particles that do not show any dichroism can be obtained by appropriate control of the process conditions. Needle-shaped monoclinic crystals should be avoided because they lead to disadvantages such as low bulk density, high oil absorption, and iridescence in the coating film.

Unstabilized chrome yellow pigments have poor lightfastness, and darken due to redox reactions. Further developments have led to improvements in the fastness properties of chrome yellow pigments, especially toward sulfur dioxide and temperature. This has been achieved by coating the pigment particles with compounds of titanium, cerium, aluminum, antimony, and silicon [309-317].

Carefully controlled precipitation and stabilization provide chrome yellow pigments with exceptional fastness to light and weathering, and very high resistance to chemical attack and temperature, enabling them to be used in a wide field of applications. The following qualities are commercially available:

1. Unstabilized chrome yellows (limited importance)
2. Stabilized chrome yellows with high color brilliance, stable to light and weathering
3. Highly stabilized chrome yellow pigments
   - very stable to light and weathering
   - very stable to light and weathering, and resistant to sulfur dioxide
   - very stable to high temperature, light, and weathering
   - very stable to high temperature, sulfur dioxide, light, and weathering
4. Low-dust products (pastes or powders)

Lightfast chrome yellow pigments that are coated with metal oxides (e.g., of aluminum, titanium, manganese) are produced by Du Pont [310].

A chrome yellow that is coated with large amounts of silicate and alumina and which shows improved stability to temperature, light, and chemicals is also produced by Du Pont [311].

Bayer describes pigments containing lead chromate stabilized in aqueous slurry with silicate- containing solutions and antimony(III), tin(II), or zinc compounds [312].

ICI produces light- and weatherfast chrome yellow pigments stabilized with antimony compounds and silicates in the presence of polyhydric alcohols and hydroxyalkylamines [313].

Ten Horn describes a process for the production of lead sulfochromate containing at least 50 % lead chromate [314]. This has a low acid-soluble lead content (< 5 % expressed as PbO, by BS 3900, Part B3, 1965).

BASF produces temperature-stable lead chromate pigments with a silicate coating obtained by hydrolysis of magnesium silicofluoride [315]

Heubach has developed a process for the alternate precipitation of metal oxides and silicates [316], [317]. A homogenizer is used to disperse the pigment particles during stabilization. Products obtained have a very good temperature resistance and very low lead solubility in acid (< 1 % Pb by DIN 55 770, 1986 or DIN/ISO 6713, 1985).

Continuous processes for the production of chromate pigments have been developed in the United States and Hungary [318], [319].

Uses. Chrome yellow pigments are mainly used for paints, coil coatings, and plastics. They have a low binder demand and good dispersibility, hiding power, tinting strength, gloss, and gloss stability. Chrome yellows are used in a wide range of applications not only for economic reasons, but also on account of their valuable pigment properties. They are important base pigments for yellow colors in the production of automotive and industrial paints.

Chrome yellow pigments stabilized with a large amount of silicate play a major role in the production of colored plastics (e.g., PVC, polyethylene, or polyesters) with high temperature resistance. Incorporation into plastics also improves their chemical resistance to alkali, acid, sulfur dioxide, and hydrogen sulfide.

Chrome green and fast chrome green mixed pigments are produced by combining chrome yellow with iron blue or phthalocyanine blue (see Section Chrome Green and Fast Chrome Green).

World production of chrome yellow in 1996 was 37 000 t.

3.4.2. *Molybdate Red and Molybdate Orange*

Molybdate red and molybdate orange [12656-85-8], C.I. Pigment Red 104:77605, are mixed-phase pigments with the general formula $Pb(Cr,Mo,S)O_4$ [306]. Most commercial products have a MoO_3 content of 4 – 6 % (refractive index 2.3 – 2.65, density 5.4 – 6.3 g/cm^3). Their hue depends on the proportion of molybdate, crystal form,

and particle size.

Pure tetragonal lead molybdate, which is colorless, forms orange to red tetragonal mixed-phase pigments with lead sulfochromate. The composition of molybdate red and molybdate orange pigments can be varied to give the required coloristic properties; commercial products usually contain ca. 10 % lead molybdate. Lead molybdate pigments have a thermodynamically unstable tetragonal crystal modification that can be transformed into the undesirable stable yellow modification merely by dispersing [320]. This is especially true of the bluish varieties of molybdate red which have larger particles whose color can be changed to yellow by shear forces. The tetragonal modification of the lead molybdate pigments must therefore be stabilized after precipitation [321], [322].

The fastness properties of the molybdate orange and molybdate red pigments are comparable with those of the chrome yellows. As with the chrome yellows, the pigment particles can be coated with metal oxides, metal phosphates, silicates, etc., to give stabilized pigments with high color brilliance and good fastness properties, as well as highly stabilized grades with very good resistance to light, weathering, sulfur dioxide, and temperature, and with a very low content of acid-soluble lead (DIN 55 770, 1986 or DIN/ ISO 6713, 1985).

The colors of lead molybdate pigments vary from red with a yellow hue to red with a blue hue. Since chrome orange is no longer available (see Section Chrome Orange), molybdate orange has become much more important.

Production. In the Sherwin – Williams process, a lead nitrate solution is reacted with a solution of sodium dichromate, ammonium molybdate, and sulfuric acid [323]. Instead of ammonium molybdate, the corresponding tungsten salt can be used, giving a pigment based on lead tungstate. The pigment is stabilized by adding sodium silicate (25 % SiO_2) and aluminum sulfate ($Al_2(SO_4)_3 \cdot 18\ H_2O$) to the suspension, which is then neutralized with sodium hydroxide or sodium carbonate. The pigment is filtered off, washed until free of electrolyte, dried, and ground. Treatment with silicate increases the oil absorption; it also improves light fastness and working properties.

In the Bayer process molybdate red is formed from lead nitrate, potassium chromate, sodium sulfate, and ammonium molybdate [312]. The pigment is then stabilized by adding water glass (28 % SiO_2, 8.3 % Na_2O) to the suspension with stirring, followed by solid antimony trifluoride, stirring for 10 min, and further addition of water glass. The pH is adjusted to 7 with dilute sulfuric acid and the pigment is filtered off, washed free of electrolyte, dried, and ground.

To give the lead molybdate pigments very good stability to light, weathering, chemical attack, and temperature, the same methods are used as those for the stabilization of chrome yellow pigments (see Section Chrome Yellow) [309-317].

Uses. Molybdate orange and molybdate red are mainly used in paints, coil coatings, and for coloring plastics (e.g., polyethylene, polyesters, polystyrene). The temperature-stable grades are the most suitable for coil coatings and plastics.

Molybdate orange and molybdate red have a low binder demand; good dispersibility, hiding power, and tinting strength; and very high lightfastness and weather resistance. Stabilization (see “Production”) also gives high-grade pigments with good resistance to sulfur dioxide and high temperature.

Like the chrome yellows, the molybdate reds are used to produce mixed pigments. Combinations with organic red pigments give a considerably extended color range. Such combinations have very good stability properties because the lightfastness and weather resistance of many organic red pigments are not adversely affected by molybdate pigments.

Total world production of molybdate orange and molybdate red in 1996 was 13 000 t.

3.4.3. *Chrome Orange*

Chrome orange [1344-38-3], C.I. Pigment Orange 21:77601, is a basic lead chromate with the composition $PbCrO_4 \cdot PbO$ but is no longer of technical or economic importance.

This product was obtained by precipitating lead salts with alkali chromates in the alkaline pH range. By controlling the pH and temperature, the particle size and thus the hue could be varied between orange and red.

3.4.4. *Chrome Green and Fast Chrome Green*

Chrome greens, C.I.Pigment Green 15:77510 and 77600, are combined or mixed pigments of chrome yellow (see Section Chrome Yellow) and iron blue (see Section Iron Blue Pigments) with the formula

$$Pb(S,Cr)O_4/Fe_4^{III}[Fe^{II}(CN)_6]_3 \cdot xH_2O$$

Fast chrome greens, C.I. Pigment Green 48:77600, 74160, and 74260, are combinations of chrome yellow (see Section Chrome Yellow) and phthalocyanine blue or phthalocyanine green (⟶ Phthalocyanines). For high-grade fast chrome greens, stabilized and highly stabilized chrome yellows are usually used.

The density and refractive index of the chrome greens and fast chrome greens depend on the ratio of the components of the mixture. Their hues vary from light green to dark blue-green, again depending on the ratio of the components.

Production. Chrome green and fast chrome green pigments can be prepared by dry or wet mixing.

Dry Mixing. The yellow and blue or green pigments are mixed and ground in edge runner mills, high-performance mixers, or mills giving intimate contact of the pigment particles. Excessive increase of temperature must be avoided, because this can lead to spontaneous combustion [324]. Differences in the density and particle size of the components can lead to segregation and floating of the pigment components in the coating. Wetting agents are therefore added to avoid these effects [325].

Wet Mixing. Pigments with brilliant colors, high color stability, very good hiding power, and good resistance to floating and flocculation are obtained by precipitating one component onto the other. Solutions of sodium silicate and aluminum sulfate or magnesium sulfate are then added for further stabilization [323].

Alternatively, the components are wet milled or mixed in suspension and then filtered. The pigment slurry is dried, and the pigment is ground.

Uses. Chrome greens have very good dispersibility, resistance to flocculation, bleeding, and floating and very good fastness properties. This is especially true of the fast chrome greens that are based on high-grade phthalocyanine and highly stabilized chrome yellows. They are therefore used in the same applications as chrome yellow and molybdate red pigments (i.e., for the pigmentation of coating media and plastics).

Pigments consisting of zinc potassium chromates combined with blue pigments are no longer of importance.

3.4.5. *Toxicology and Occupational Health*

Occupational Health. Precautions have to be taken and workplace concentration limits have to be observed when handling lead- and lead chromate-containing pigments. General regulations exist for all lead- containing materials [326]. Concentration limits are as follows:

MAK value (lead)	< 0.1 mg/m^3
BAT values	
Lead (blood)	< 70 µg/dL
Lead (blood – women < 45 years)	< 30 µg/dL
δ-Aminolevulinic acid	
(urine, Davies method)	< 15 mg/L
(women < 45 years)	< 6 mg/L
TLV-TWA value (lead)	< 0.15 mg/m^3

It is accepted that the BAT limit has been complied with if the blood lead level does not exceed 50 µg/dL (or for women of < 45 years, 30 µg/dL).

The EEC Directive EEC 82/605 specifies maximum lead concentrations in the air of < 150 µg/m^3 and permitted blood lead levels of 70 – 80 µg/dL, with δ-aminolevulinic acid values of 20 mg/g creatinine [327].

MAK limits for lead chromates and lead chromate pigments are not given. They are classified as substances suspected of having carcinogenic potential (MAK: Group III B; TLV-TWA: 0.05 mg Cr/m^3, A 2). However, extensive epidemiological investigations have given no indication that the practically insoluble lead chromate pigments have any carcinogenic properties [328], [329]. Such properties have been reported for the more soluble zinc chromate and strontium chromate pigments.

These chromate pigments can be safely handled if the various rules and regulations regarding concentration limits, safe working practices, hygiene and industrial medicine are adhered to.

Environmental Aspects. Dust emissions from approved manufacturing plants must not exceed 5 mg/m^3 for lead and chromium with a total mass flow exceeding 25 g/h (TA-Luft) [330].

According to latest German wastewater legislation [332] for inorganic pigment manufacturing processes discharging directly into public stretches of water, mass limits for lead and chromium related to tonnes of average output (t_{prod}) are:

Lead	0.04 kg/t_{prod}
Chromium (total)	0.03 kg/t_{prod}

These requirements replaced earlier legislation dating from 1984 [331]. Lower limits might be set by local or regional authorities, even for “nondirect” discharges into municipal sewer systems.

Waste containing lead and lead chromate that cannot be recycled must be taken to a special waste disposal site under proper control.

Labeling. In the EC lead chromate and lead chromate pigments must be appropriately labeled. Such substances must be marked with a skull and crossbones (T) [333-336]. Additionally, the following risk (R) and safety phrases (S) must be used:

R 61 May cause harm to the unborn child
R 62 Possible risk of impaired fertility
R 33 Danger of cumulative effects
R 40 Possible risk of irreversible effects
S 53 Avoid exposure — obtain special instructions before use
S 45 In case of accident or if you feel unwell, seek medical advice immediately (show the label where possible)

In the 21st adaption of EEC Council Directive 67/548, the lead chromate pigments C.I. Pigment Yellow 34 and C.I. Pigment Red 104 have been added individually [337]. These pigments are classified in the same manner as lead chromate and must be labeled with a skull and crossbones (T) and the above-mentioned risk (R) and safety phrases (S).

According to the EEC Council Directive for the labeling of preparations [338] in conjunction with the 21st adaption to EEC Council Directive 67/548, Nota 1 [337], such materials containing more than 0.5 % lead are labeled in the same way as the pure lead pigment with a skull and crossbones (T) and the corresponding R and S phrases.

With respect to improved protection of public health, special restrictions on carcinogenic and teratogenic substances and their corresponding preparations have been established by the 14th amendment of EEC Council Directive EEC 76/769 [339].

In accordance with the 14th amendment of EEC Council Directive EEC 76/769 [339] and revised ChemVerbV [340], lead chromate based pigments and preparations are no longer permitted to be used by private consumers and have to be labeled with the phrase: “Only for industrial purposes”.

Lead-containing coatings and paints with a total lead content exceeding 0.15 % of the total weight of the preparation must carry the phrase: “Contains lead. Should not be used on surfaces liable to be chewed or sucked by children” in accordance with EEC Council Directive 89/178 [341] and German GefStoffV [333]

Lead chromate pigments are not permitted for use as coloring materials for plastics for consumer goods [342], or for coatings for toys, according to the European Standard EN 71 part 3 [343].

For transportation, the labeling to be used (GGVS/GGVE, ADR/RID) for the pigment as well as for its preparations is Class 6.1, No. 62 c, hazard symbol 6.1 A, and a skull and crossbones, if the lead content soluble in hydrochloric acid [c(HCl) = 0.07 mol/L] exceeds 5 %.

3.5. Ultramarine Pigments

Blue ultramarine — blue from over the sea — is the name which European artists of the Middle Ages gave to the pigment derived from lapis lazuli a semiprecious stone imported mainly from Afghanistan. Ultramarine was the supreme blue of medieval times, but eventually it became scarce and very expensive.

In 1828, J. B. GUIMET in France and CHRISTIAN GMELIN in Germany independently devised similar processes for synthetic preparation. Relatively abundant supplies soon became available, the price fell dramatically, and ultramarine was adopted as a general-purpose color.

Synthetic ultramarines are inorganic powder pigments, commercially available in three colors:

1. Reddish blue, C.I. Pigment Blue 29:77007 [57455-37-5]
2. Violet, C.I. Pigment Violet 15:77007 [12769-96-9]
3. Pink, C.I. Pigment Red 259:77007 [12769-96-9]

A green variety, once produced in small quantities, is no longer available.

Within limits set by stability considerations, the proportions of the chemical constituents can vary. The typical lattice repeat unit of a blue ultramarine is $Na_{6.9}Al_{5.6}Si_{6.4}O_{24}S_{4.2}$. The violet and pink variants differ from the blue mainly in the oxidation state of the sulfur groups. This is reflected in somewhat lower sodium and sulfur contents.

3.5.1. *Chemical Structure*

Reviews of work on the structure of ultramarine are given in [344], [345].

Ultramarine is essentially a three-dimensional aluminosilicate lattice with entrapped sodium ions and ionic sulfur groups (Fig. 32). The lattice has the sodalite structure, with a cubic unit cell dimension of ca. 0.9 nm. In synthetic ultramarine derived from china clay by calcination (see Section Production), the lattice distribution of silicon and aluminum ions is disordered. This contrasts with the ordered array in natural ultramarines.

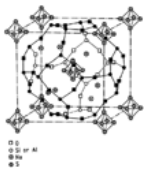

Figure 32. Schematic drawing of the basic structure of standard ultramarine showing the available sites for sulfur and sodium

In the simplest ultramarine structure, equal numbers of silicon and aluminum ions are present and the basic lattice unit is $Na_6Al_6Si_6O_{24}$ or $(Na^+)_6(Al^{3+})_6(Si^{4+})_6(O^{2-})_{24}$ with a net ionic charge of zero as required for structural stability.

The nature of the sulfur groups responsible for the color is reviewed in [346-348]. There are two types of sulfur group in blue ultramarine, S_3^- and S_2^-, both being free radicals stabilized by lattice entrapment. In the predominant S_3^- species, the spacing between the three sulfurs is 0.2 nm and the angle between them is 103°. S_3^- absorbs a broad

energy band in the visible green – yellow – orange region centered at 600 nm, while S_2^- absorbs in the violet – ultraviolet region at 380 nm (Fig. 33).

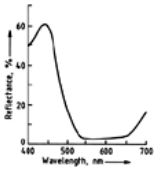

Figure 33. Spectral reflectance distribution of blue ultramarine

The basic lattice $(Na^+)_6(Al^{3+})_6(Si^{4+})_6\ (O^{2-})_{24}$ is derived from $Si_{12}O_{24}$ by substituting six of the silicon ions by aluminum. Every Al^{3+} must be accompanied by a Na^+ so that the overall ionic charge for the structure is zero. Hence, six of the eight sodium sites are always filled by sodiums required for lattice stability and the remaining two sites are filled with sodiums associated with ionic sulfur groups. This means that only one S_3^{2-} polysulfide ion can be inserted into the lattice (as Na_2S_3) even though subsequent oxidation to S_3^- (Section Production) leads to loss of one of the accompanying sodium ions. This gives a basic ultramarine lattice formula of $Na_7Al_6Si_6O_{24}S_3$.

To increase the sulfur content and thereby improve color quality, the lattice aluminum content can be decreased by including a high-silicon feldspar in the manufacturing recipe (Section Production). This reduces the number of sodium ions needed for lattice stabilization and leaves more for sulfur group equivalence. A typical product would be $Na_{6.9}Al_{5.6}Si_{6.4}O_{24}S_{4.2}$ with a stronger, redder shade of blue than the simpler type.

In violet (Fig. 34) and pink (Fig. 35) ultramarines, the lattice structure is little changed, but the sulfur chromophores are further oxidized, possibly to S_3Cl^-, S_4, or S_4^-.

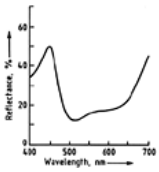

Figure 34. Spectral reflectance distribution of violet ultramarine

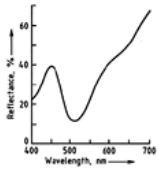

Figure 35. Spectral reflectance distribution of pink ultramarine

Ultramarines are zeolites, though lattice paths are restricted by 0.4 nm diameter channels. The sodium ions can be exchanged for other metal ions (e.g., silver, potassium, lithium, copper). Although this produces marked color change, none of the products have commercial value.

3.5.2. *Properties*

The basic ultramarine color is a rich, bright reddish blue, the red-green tone varying with chemical composition. The violet and pink derivatives have weaker, less saturated colors (see Figs. 33-35, for reflectance spectra).

The color quality of commercial pigments is developed by grinding to reduce particle size and thus enhance tinting strength. Mean particle size ranges typically from 0.7 to 5.0 µm. Figure 36 shows an electron micrograph of a 1.0 µm grade. Fine pigments are lighter in shade and rather greener than coarser grades, but when reduced with white, their color is brighter and more saturated.

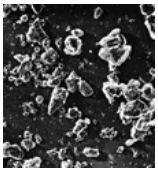

Figure 36. Electron micrograph of blue ultramarine with a mean particle size of 1.0 µm (magnification×5 800)

With a refractive index close to 1.5, similar to that of paint and plastics media, ultramarine gives a transparent blue in gloss paints and clear plastics. Opacity is obtained by adding a small quantity of a white pigment. Increasing quantities of white give paler shades, and a trace of ultramarine added to a white enhances the whiteness and acceptability.

In many applications ultramarine blue is stable to around 400 °C, violet to 250 °C, and pink to 200 °C. All have excellent lightfastness with a 7 – 8 rating (full and reduced shades) on the International Blue Wool Scale. Color fade attributed to light exposure or moderate heat is almost always caused by acid attack. Ultramarines react with all acids, and if there is sufficient acid, the pigment is completely decomposed losing all color to form silica, sodium and aluminum salts, sulfur, and hydrogen sulfide. Evolution of hydrogen sulfide with acids is a useful test for ultramarine. Grades resistant to transient acidity are available, in which the pigment particles are protected by a coating of impervious silica. Blue and violet grades are stable in mildly alkaline conditions, but pink tends to revert to a violet shade.

Ultramarines are insoluble in water and organic solvents, so the color does not bleed or migrate from paints or polymers.

Being macromolecular, the fine ultramarine particles have a high surface energy and are cohesive. The finer grades, with their greater surface area, are therefore less easy to disperse than the coarser types, and some are available with pigment surfaces treated to reduce energy and improve dispersability.

Ultramarines adsorb moisture on the external particle surfaces and at the internal surfaces of the zeolite structure. External surface moisture (1 – 2 % according to particle size) is driven off at 100 – 105 °C, but the additional 1 % of internal moisture needs 235 °C for complete removal.

Ultramarine particles are hard and can cause abrasion in equipment handling either dry or slurried pigment.

Specific gravity is 2.35, but bulk density of the pigment powder is much lower, varying with particle size between 0.5 and 0.9 g/cm^3.

Oil absorption also varies with particle size (usually 30 – 40 g %). The pH lies between 6 and 9. Ultramarine pigments are largely odorless, nonflammable, and do not support combustion.

3.5.3. *Production*

Ultramarine is made from simple, relatively cheap materials, typically china clay, feldspar, anhydrous sodium carbonate, sulfur, and a reducing agent (oil, pitch, coal etc.). Use of a synthetic zeolite has been proposed [349], [350], but this method is not known to be in commercial use.

Clay Activation. The clay is heated to about 700 °C to destabilize the kaolinite structure by removing hydroxyl ions as water. This can be either a batch process with the clay in crucibles in a directly fired kiln, or a continuous process in a tunnel kiln, rotary kiln, or other furnace.

Blending and Heating Raw Materials. The activated clay is blended with the other raw materials and dry-ground, usually in batch or continuous ball mills, to a mean size

approaching 15 µm. Typical recipes (in wt %) are:

	Green-tone	Red-tone
Calcined clay	32.0	30.0
Feldspar		7.0
Sodium carbonate	29.0	27.0
Sulfur	34.5	33.0
Reducing agent	4.5	3.0

The ground mixture is heated to about 750 °C under reducing conditions, normally in a batch process. This can be done in directly fired kilns with the blend in lidded crucibles of controlled porosity, or muffle kilns. The heating medium can be solid fuel, oil, or gas. The sodium carbonate reacts with the sulfur and reducing agent at 300 °C to form sodium polysulfide. At higher temperatures the clay lattice reforms into a three-dimensional framework, which at 700 °C is transformed to the sodalite structure, with entrapped sodium and polysulfide ions.

Oxidation. The furnace is allowed to cool to ca. 500 °C when air is admitted in controlled amounts. The oxygen reacts with excess sulfur to form sulfur dioxide, which exothermically oxidizes the di- and triatomic polysulfide ions to S_2^- and S_3^- free radicals, leaving sodium sulfoxides and sulfur as byproducts. When oxidation is complete, the furnace cools and is unloaded — a full kiln cycle can take several weeks. The “raw” ultramarine product typically contains 75 wt % blue ultramarine, 23 wt % sodium sulfoxides, and 2 wt % free (uncombined) sulfur with some iron sulfide.

Purification and Refinement. The purification and refinement operations can be batch or continuous. The raw blue is crushed and ground, slurried in warm water, then filtered and washed to remove the sulfoxides. Reslurrying and wet grinding release the sulfurous impurities and reduce the ultramarine particle size, often to 0.1 – 10.0 µm. The impurities are floated off by boiling or cold froth flotation.

The liquor is then separated into discrete particle size fractions by gravity or centrifugal separation; residual fine particles are reclaimed by flocculation and filtration. The separated fractions are dried and disintegrated to give pigment grades differentiated by particle size. These are blended to sales-grade standard, adjusting hue, brightness, and strength to achieve specified color tolerance.

Violet ultramarine can be prepared by heating a mid-range blue grade with ammonium chloride at ca. 240 °C in the presence of air. Treating the violet with hydrogen chloride gas at 140 °C gives the pink derivative.

A good ultramarine pigment would meet the following specification:

Tinting strength/standard	± 2 % max.
Reduced shade/standard	1 CIELAB unit max.
Free sulfur	0.05 % max.
Water-soluble matter	1.0 % max.
Sieve residue (45 µm)	0.1 % max.
Moisture	
coarse grade	1.0 % max.
fine grade	2.0 % max.
Heavy metals	traces

For standards see Table 1 (“Ultramarine pigments: specification”).

3.5.4. *Uses*

The stability and safety of ultramarine pigments are the basis of their wide range of applications which include the following:

- Plastics
- Paints and powder coatings
- Printing inks
- Paper and paper coating
- Rubber and thermoplastic elastomers
- Latex products
- Detergents
- Cosmetics and soaps
- Artists' colors
- Toys and educational equipment
- Leather finishes
- Powder markers
- Roofing granules
- Synthetic fibers
- Theatrical paints and blue mattes
- Cattle salt licks
- White enhancement

Plastics. Blue ultramarine can be used in any polymer; violet ultramarine has a maximum processing temperature of 250 °C, and pink ultramarine has a maximum processing temperature of 200 °C. With PVC, acid-resistant grades must be used if color fades during processing. Surface-treated grades are available for enhanced dispersibility. Ultramarines do not cause shrinkage or warping in polyolefins. Ultramarine pigments are permitted worldwide for coloring food- contact plastics.

Paints. Ultramarine pigments are used in decorative paints, stoving finishes, transparent lacquers, industrial paints, and powder coatings. They are not recommended for colored, air-drying paints for outdoor use in urban atmospheres.

Printing Inks. Ultramarine pigments can be used in inks for most printing processes including hot-foil stamping. Letterpress, flexography, and gravure need high-strength grades; lithography needs water-repellent grades; any grade is suitable for screen inks, fabric printing, and hot-foil stamping.

Paper and Paper Coating. Ultramarine pigments are used to enhance the hue of white paper or for colored paper. They can be added directly to the paper pulp, or used in applied coatings. When added to the pulp, acid-resistant grades must be used if acid-sizing is employed. They are particularly suitable for colored paper for children's use.

Detergents. Ultramarine pigments are widely used to enhance the effects of optical brightening agents in improving whiteness of laundered fabrics [351]. They do not stain or build up with repeated use.

Cosmetics and Soaps. Ultramarine pigments are widely used in cosmetics. Pink is not recommended for toilet soaps because of the color shift to violet. Advantages are complete safety, nonstaining, and conformance to all major regulations.

Artists' Colors. This traditional use for ultramarine in all types of media is still an important application. Unique color properties, stability, and safety are highly prized.

Toys and other Articles/Materials for Children's Use. Ultramarine pigments are widely used in plastics and surface coatings for toys, children's paints and finger paints, modeling compositions, colored paper, crayons, etc. They comply with major regulations and standards.

3.5.5. *Toxicity and Environmental Aspects*

Ultramarines are safe in both manufacture and use. Their only known hazard is the evolution of hydrogen sulfide on contact with acid. Massive exposure of workers in well over a century of manufacture, worldwide use for whitening clothes, and use in a number of countries for whitening sugar have all been without reported ill-effects.

Tests sponsored by Reckitt's Colours confirm that acute oral toxicity in rats and mice (LD_{50}) is greater than 10 000 mg/kg. Fish toxicity (LC_{50} in rainbow trout) exceeds

32 000 mg/L. Ultramarine is nonmutagenic, nonirritant, and nonsensitizing to skin.

There is no listed threshold limit value or maximum exposure limit for the pigment. Normal practice is to consider it a nuisance dust with TLV 10 mg/m^3. The pigment is not listed as a dangerous substance in the EC nor in any similar national or international classification; neither is it classified as hazardous for disposal.

The production process evolves close to 1 t of gaseous sulfur dioxide and 0.3 t of water-soluble sodium sulfoxides for every tonne of pigment produced. These must be disposed of in an environmentally acceptable manner. If the soluble salts are fully oxidized, they can be discharged safely into tidal waters. Future legislation in all producing countries may require removal of sulfur dioxide from the effluent gases before discharge to the atmosphere.

3.5.6. *Economic Aspects*

Ultramarines can be categorized as either laundry grades, which are low-strength and sometimes low-purity materials, or as industrial grades, which are high-strength, high-purity pigments.

Factories in several countries produce laundry-grade materials, including several in the People's Republic of China, India, Eastern European countries, and one in Pakistan. Neither the numbers of these units nor their outputs are accurately known.

There are only three major producers of high-grade ultramarine pigments — Dainichi Seika (Japan), Nubiola (Spain), Prayon (Belgium), and Reckitt's Colours (France, United Kingdom) — with two smaller producers in Austria and Colombia. In 1990 total worldwide production was ca. 20 000 t/a.

3.6. Iron Blue Pigments

The term iron blue pigments as defined in ISO 2495 has largely replaced a variety of older names which were either related to the place where the compound was produced or represented particular optical properties (e.g., Berlin Blue, Bronze Blue, Chinese Blue, Milori Blue, Non-bronze Blue, Paris Blue, Prussian Blue, Toning Blue, and Turnbull's Blue). These names usually stood for insoluble pigments based on microcrystalline Fe(II)Fe(III) cyano complexes which do not differ significantly in their composition; many were associated with specific hues. A standardized naming system has been demanded by users and welcomed by manufacturers, and has led to a reduction in the number of names [352].

Iron blue [14038-43-8], C.I. Pigment Blue 27:77510 (soluble blue is C.I. Pigment Blue 27 : 77520), was discovered in 1704 by DIESBACH in Berlin by a precipitation reaction, and can be regarded as the oldest synthetic coordination compound. MILORI was the first to produce it as a pigment on an industrial scale in the early 1800s [353].

3.6.1. *Structure*

X-Ray and infrared spectroscopy show that iron blue pigments have the formula $M^{I}Fe^{II}Fe^{III}(CN)_6 \cdot H_2O$ [354]. M^{I} represents potassium, ammonium, or sodium, of which the potassium and ammonium ions are preferred because they produce excellent hues in industrial manufacture.

The crystal structure of the $Fe^{II}Fe^{III}(CN)_6$ grouping is shown in Figure 37. A face-centered cubic lattice of Fe^{2+} is interlocked with another face- centered cubic lattice of Fe^{3+} to give a cubic lattice with the corners occupied by iron ions. The CN^- ions are located at the edges of the cubes between each Fe^{2+} ion and the neighboring Fe^{3+}; the carbon atom of the cyanide is bonded to the Fe^{2+} ion and the nitrogen atom is coordinatively bonded to the Fe^{3+} ion. According to [355] the deep blue color is the result of electron transfer from iron(II) to iron(III) with the absorption of light energy (charge-transfer complex).

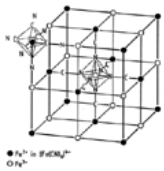

Figure 37. Crystal structure of iron blue [356]

The wide-mesh lattice of the crystal contains relatively large spaces which can be occupied alternately by alkali ions and water molecules. In order to retain the crystal structure — and also the optical properties — water molecules must be present. Loss of water beyond a certain limit brings about fundamental changes in the pigment properties. A bonding mechanism for the coordinated water molecules assumes that both zeolitic and adsorptive states are possible.

Many investigations helped to elucidate the structure of iron blue [357-360].

3.6.2. *Production*

Iron blue pigments are produced by the precipitation of complex iron(II) cyanides by iron(II) salts in aqueous solution. The product is a whitish precipitate of iron(II) hexacyanoferrate(II) $M^{I}_2Fe^{II}[Fe^{II}(CN)_6]$ or $M^{II}Fe^{II}[Fe^{II}(CN)_6]$, (Berlin white), which is aged and then oxidized to the blue pigment [356].

Potassium hexacyanoferrate(II), sodium hexacyanoferrate(II), or mixtures thereof are usually used (⟶ Cyano Compounds, Inorganic). When the pure sodium salt is used, the pigment properties are obtained by adding a potassium or ammonium salt during the precipitation of the white paste product or prior to the oxidation stage.

The iron(II) salt used is crystalline iron(II) sulfate or iron(II) chloride solution. The oxidizing agent can be hydrogen peroxide, alkali metal chlorates, or alkali metal dichromates. Industrial precipitation is carried out batchwise in large stirred tanks by simultaneous or sequential addition of aqueous solutions of potassium hexacyanoferrate(II) and iron(II) sulfate to a dilute acid. The filtrate from the white paste product must contain a slight excess of iron. Temperature and concentration of the starting solutions have a decisive influence on the size and shape of the precipitated particles. The suspension of white paste is aged by heating. The ageing period varies in length and temperature depending on the required properties of the finished pigment. This is followed by oxidation to form the blue pigment by adding hydrochloric acid and sodium or potassium chlorate [361]. Finally, the suspension of the blue pigment is pumped into filter presses either immediately or after having been washed with cold water and decanted. After filtering, it is washed until free of acid and salt.

Depending on the pigment type, the filter press cakes (35 – 60 % solids) are either molded into cylindrical pellets and dried to a finished dust-free product; liquefied and spray dried; or dried and then ground to form a powder pigment.

Dispersibility can be improved by adding organic compounds to the pigment suspension before filtering to prevent the particles from agglomerating too strongly on drying [362], [363]. In another method (the Flushing process), the water in the wet pigment paste is replaced by a hydrophobic binder [364]. Although these and other methods of pigment preparation produce fully dispersible products consisting mainly of iron blue and a binder [365-367], they have not become commercially established.

A “water-soluble” blue can be manufactured by adding peptizing agents, which improve the water solubility by an emulsifying action. This gives a transparent colloidal solution in water without the use of high shear forces [356].

For quality standards, see Table 1 (“Iron blue pigments: Methods of analysis” and “Specification”).

3.6.3. *Properties*

Hue, relative tinting strength, dispersibility, and rheological behavior are the properties of iron blue pigments with the most practical significance. Other important properties are the volatiles content at 60 °C, the water-soluble fraction, and acidity (ISO 2495). Pure blue pigments are usually used alone (e.g., in printing inks) and do not need any additives to improve them. Finely divided iron blue pigments impart a pure black tone to printing ink.

Owing to their small particle size (see Table 26 and Figs. 38 and 39), iron blue pigments are very difficult to disperse. A graph of cumulative particle size distribution is given in Figure 40 for a commercial quality iron blue and for a micronized grade with similar primary particle size. The micronized grade gives greater tinting strength in dry mixtures than the blues obtained from standard grinding. The average size of the aggregates in the micronized material is ca. 10 µm compared with around 35µm for the normal quality product.

Table 26. Physical and chemical properties of iron blue pigments (Vossen Blau and Manox Blue grades)

Type	Vossen Blau 705	Vossen Blau 705 LS [g]	Vossen Blau 724	Manox Blue 460 D	Manox Easisperse HSB 2
Colour Index Number	77 510	77 510	77 510	77 510	77 510

Colour Index Pigment	27	27	27	27	27
Tinting strength [a]	100	100	100	115	95[h]
Hue	pure blue	pure blue	pure blue	pure blue	pure blue
Oil absorption [b], g/100 g	36 – 42	40 – 50	36 – 42	53 – 63	22 – 28
Weight loss on drying [c], %	2 – 6	2 – 6	2 – 6	2 – 6	2 – 6
Tamped density [d], g/L	500	200	500	500	550
Density [e], g/cm^3	1.9	1.9	1.9	1.8	1.8
Mean diameter of primary particles, nm	70	70	70	40	80
Specific surface area [f], m^2/g	35	35	35	80	30
Thermal stability, °C	150	150	150	150	150
Resistance to acids	very good	very good	very good	very good	very good
Resistance to alkalis	poor	poor	poor	poor	poor
Resistance to solvents	very good	very good	very good	very good	very good
Resistance to bleeding	very good	very good	very good	very good	very good

[a] DIN ISO 787/XVI and DIN ISO 787/XXIV.
[b] DIN ISO 787/V, ASTM D 281, or JIS K 5101/19 (JIS: Japanese Industrial Standard).
[c] DIN ISO 787/II, ASTM D 280, or JIS K 5101/21.
[d] DIN ISO 787/XI or JIS K 5101/18.
[e] DIN ISO 787/X or JIS K 5101/17.
[f] DIN 66 131.
[g] LS=Luftstrahlmühle (air jet mill).
[h] Surface-treated, easily dispersible type.

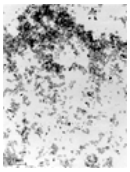

Figure 38. Electron micrograph of an iron blue pigment of small particle size (Manox Blue 460 D)

Figure 39. Electron micrograph of an iron blue pigment of normal particle size (Vossen Blau 705)

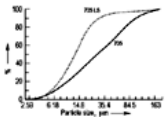

Figure 40. Cumulative particle size distribution curve of a normal (705) and a micronized (705 LS) iron blue pigment of equal primary particle size

LS = Luftstrahlmühle (air jet mill)

Iron blue pigments are thermally stable for short periods at temperatures up to 180 °C, and therefore can be used in stoving finishes. The powdered material presents an explosion hazard; the ignition point is 600 – 625 °C (ASTM D 93–52). The pigments are combustible in powder form; ignition in air is possible above 140 °C [356].

Iron blue pigments have excellent light- and weatherfastness. When mixed with white pigments, these properties can disappear [368]. A clearcoat (as commonly applied in automobile manufacture) overcomes this problem [369]. Figure 41 shows changes in residual gloss and color after a short weathering period. The pigments are resistant to dilute acid and oxidizing agents, and do not bleed. They are decomposed by hot, concentrated acid and alkali. Other properties are listed in Table 26.

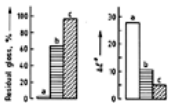

Figure 41. Residual gloss and ΔE^*_{ab} values for isocyanate-cross-linked polyacrylate resins that contain 15 wt % Vossen Blau 2000 relative to the binder and 15 wt % TiO_2 (rutile) relative to the iron blue pigment after 1000 h fast exposure to UV [369]

a) Without clearcoat; b) With clearcoat but without UV protection; c) With clearcoat and UV protection

Vossen Blau 2000 is an older pigment type that has been replaced by Manox Blue 460 D

3.6.4. *Uses*

Total production of iron blue in 1975 was ca. 25 000 t/a, but in 1995 it was only ca. 13 000 – 15 000 t/a. The main consumer in Europe and the United States is the printing ink industry. The second largest use in Europe, especially of micronized iron blue pigments, is for coloring fungicides, but use in the paint industry is decreasing.

Printing Ink Industry. Iron blue pigments are very important in printing, especially rotogravure, because of their deep hue, good hiding power, and low cost. Iron blue is often mixed with phthalocyanine pigments for multicolor printing. Another important use is in controlling the shade of black printing inks (⟶ Printing Inks). Typical amounts used are 5 – 8 % for full-shade rotogravure inks and 2 – 8 % for toning black gravure and offset inks.

Iron blue pigments are used in the manufacture of single- and multiple-use carbon papers and blue copying papers, both for toning the carbon black and as blue pigments in their own right [356].

Toning of Black Gravure Inks. For the toning of black gravure inks, 2 – 6 % of Vossen-Blau 705 is used together with 6 – 12 % of carbon black, for example. Combinations with red pigments having a blue undertone are also common. Because of the poor dispersibility of iron blue compared with carbon black it is both economical and practical to disperse the blue pigment in a separate step. While the visual judgement of black is influenced by the individual ability of the observer to distinguish small color differences in deep black, it is possible, with the help of photometric measurements, to graphically interpret objective evaluations by means of physical data [370].

Figure 42 illustrates the color changes of a low-structure LCF carbon black (LCF = low color furnace) on addition of various toners. A mixture of asphalt resin, calcium – zinc resinate and phenolic resin was used as binder. The pigment concentration was 13.2 %, and the toner was added in 2.2 % steps up to 6.6 % with a corresponding reduction of the carbon black content from 13.2 to 6.6 %.

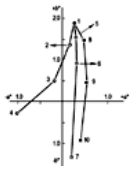

Figure 42. Color coordinates of black gravure inks with various toners

Toning of Black Offset Printing Inks. The basic requirements for the successful use of iron blue as a toning agent in offset printing inks are good dispersibility and resistance to fountain solutions. Resistance is understood here as the hydrophobicity of the pigment. This property prevents wetting of the pigment by water and therefore its peptization. Nonresistant iron blue can render the ink useless by adsorbing water to above the normal content. A negative side-effect of peptization is the “dissolution” of the blue pigment from the printing ink and the resulting blue coloration of the fountain solution, with the familiar problems of printing-plate contamination, also known as scumming or toning. The combined dispersion of pigments is only practical with colorants of similar dispersibility. Toning agents with a considerably higher resistance to dispersion than carbon black are delivered by the manufacturer as a predispersed paste or must be ground separately by the user. A new generation of pigments has been developed which covers the demand for sufficient resistance to fountain solutions combined with good dispersibility. Easily dispersible iron blue is preferred for combined dispersion with carbon black in the so-called co-grinding process. The coloristic effects of these iron blue pigments as toners for an LCF-type carbon black are compared with Pigment Blue 15 : 3 (β-copper phthalocyanine) and Pigment Blue 61 (Alkali Blue) in Figure 43. The pigment concentration of the inks is 24 %; i.e., the amount of carbon black was reduced correspondingly with the addition of 3, 6, or 9 % of blue pigment.

Carbon black used as a self-color pigment has a tendency towards black with a blue untertone at higher pigment concentration, even without the addition of toning agent (1 – 4 in Fig. 43: 15, 18, 21, and 24 % carbon black, resp.). However, without the addition of a blue pigment, the achromatic point cannot be achieved. The desired color hue can be achieved by adding various toning agents, albeit with varying red / green spread. Iron blue (Manox Easisperse 154) brings about a clear shift towards green with a relatively small shift in the blue direction (5 – 7). The target is more successfully achieved by using a mixture of iron blue and Pigment Red 57 : 1 in the ratio 4 : 1 (8 – 10) The addition of Pigment Blue 61 results in a significant step towards blue-red (11 – 13). In addition, a bronze effect occurs and intensifies with increasing distance from the achromatic point. With Pigment Blue 15 : 3, the hue of the black ink moves towards blue-green (14 – 16).

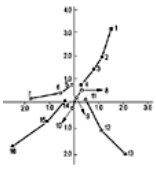

Figure 43. Color coordinates of black offset printing inks with various toners

Agriculture. Since ca. 1935, and especially in Mediterranean countries, blue inorganic fungicides based on copper and used for treating vines have largely been replaced by colorless organic compounds. Micronized iron blue pigments are used to color these fungicides (normally at a concentration of 3 – 8 wt %), so that even small amounts become visible due to the high color intensity, and precise control is possible. The fungicide is usually milled or mixed with a micronized iron blue pigment [371].

A welcome side effect of treating fungi (e.g., *Peronospora plasmopara viticola*) with iron blue is the fertilization of vines in soils that give rise to chlorosis. Leaf color is intensified, aging of the leaves is retarded, and wood quality (“ripeness”) is also improved [372-374]. Iron is necessary for chlorophyll synthesis, which improves grape quality and yield. Other iron salts do not have this effect [373].

Paints and Coatings. Iron blue pigments are used in the paint industry, especially for full, dark blue colors for automotive finishes. A full shade with good hiding power is produced by 4 – 8 % iron blue pigments.

Paper. Blue paper can be produced by adding “water-soluble” iron blue pigment directly to the aqueous phase. Alternatively, a suitable iron blue pigment can be ground together with a water-soluble binder, applied to the paper, dried, and glazed (quantity applied: ca. 10 % in the dispersion).

Pigment Industry. The importance of iron blue in the production of chrome green and zinc green pigments has greatly increased worldwide (see Section Chrome Green and Fast Chrome Green).

Medical Applications. Iron blue has become important as an agent for decontaminating humans and cattle that have ingested radioactive material. The isotope ^{137}Cs which would otherwise be freely absorbed via the human or animal digestive tract exchanges with the iron(II) of the iron blue [375], [376] and is then excreted in the feces [377]. Gelatin capsules containing 500 mg iron blue are marketed as Radiogardase-Cs (Heyl). Thallium ions have been found to behave similarly [378-380]. The gelatin capsules for this purpose are sold as Antidotum Thalii (Heyl) [381].

3.6.5. *Toxicology and Environmental Aspects*

Iron blue pigments show no toxicity in animal studies and are therefore not expected to cause adverse effects on human health. No toxic effects have been reported in humans for blue pigments used experimentally or therapeutically.

Toxikokinetic studies have shown that the adsorption of iron blue pigments is very low. Following intravenous injection of a ^{59}Fe radio-labelled iron blue pigment, the $[^{59}Fe(CN)_6]^{4-}$ion was rapidly and virtually completely excreted with the urine. After oral administration of iron(III) cyanoferrate (^{59}Fe) ca. 2 % of the labelled hexacyanoferrate ion was absorbed by the gastro-intestinal tract [382]. Most of the substance is excreted with the feces [383] and there was no evidence of its decomposition.

The decomposition of iron blue pigment salts to toxic cyanide in aqueous systems is very low. The 5 h CN-release of $KFe[Fe(CN)_6]$ in artificial gastric and intestinal juices was 141 or 26 µg/g respectively, and in water 37 µg/g. The corresponding figures for $Fe_4[Fe(CN)_6]_3$ were 64, 15, and 22 µg/g [384].

In the breath of rats after i.p. injection of ^{14}C-labelled $KFe[Fe(CN)_6]$, less than 0.01 % (detection limit)of the administered radioactivity attributable to ^{14}C-labelled $KFe[Fe(CN)_6]$ was found, whereas in another study 0.04 – 0.08 % of the orally administered dose was found in the exhaled air [385]. It can be concluded that the hexacyanoferrate (II) complex disintegrates only to a small extent in the intestinal tract after oral administration. This is confirmed by the results of acute oral toxicity studies at high doses, which show no clinical symptoms or lethality. The LD_{50} values are above 5000 – 15 000 mg/kg (limit tests) [386-388].

In primary irritation tests on rabbits, no or only slight effects were observed on the skin and in the eyes, respectively [388], [389]. No skin sensitization occurred in a Guinea pig maximization test [386].

The subchronic (90 – 120 d) consumption of iron blue pigments at concentrations of 1 – 2 % in the food or drinking water of rats influenced body weight gain slightly, but no other clinical signs or histopathological changes were observed [390-393]. After the administration of daily doses of 200 or 400 mg/kg for 10 d to dogs, body weight increased and the general condition remained unaffected [394].

In a bacterial test system (Ames test) no increase of mutagenicity was detected with or without the presence of a metabolic system [388].

In human volunteers who received 1.5 or 3.0 g iron(III) hexacyanoferrate(II) for up to 22 d no effects were reported, apart from slight obstipation [394], [395].

Since $Fe_4[Fe(CN)_6]_3$ can bind cesium, iron blue pigments are used in clinical practice as an antidote for the treatment of humans contaminated with radioactive cesium (see also Section Uses). Clinical use of iron(III) hexacyanoferrate(II) in doses up to 20 g/d for decontamination of persons exposed to radioactive cesium has not been associated with any reported toxicity [384].

Iron blue pigment salts are also used as an effective antidote for thallium poisoning. Iron(III) hexacyanoferrate(II) interferes with the enterosystemic circulation of thallium ions and enhances their fecal excretion [397].

In a semistatic acute freshwater fish toxicity test (*Leuciscus idus melanotus*) with a saturated solution of different iron blue pigment compounds (with undissolved material at the bottom or filtered solutions) no death occurred within 96 h. Based on the quantity weighed the NOEL is > 1000 mg/L (nominal concentrations) [398].

The bacterial toxicity has been measured according to DEV, DIN 38 412, L3 [TTC (2,3,5-triphenyl-2*H*-tetrazolium chloride) test]. The result gives an EC_{50} varying between 2290 and 14 700 mg/L, and estimated NOEC values in the range of < 10 to 100 mg/L [399]. There are no harmful effects on fish, but the toxic effects on bacteria constitute a slight hazard when iron blue pigments are present in water.

[Top of Page]

4. Specialty Pigments

Lutz Leitner, Hendrik Kathrein, Ekkehard Schwab, Helmut Jakusch, Manfred Ohlinger, Ronald Veitch, Günter Etzrodt, Gerhard Pfaff, Klaus-Dieter Franz, Ralf Emmert, Katsuhisa Nitta, Robert Besold, Harald Gaedcke

This chapter deals with pigments that produce special effects: magnetic pigments (Section Magnetic Pigments), anticorrosive pigments (Section Anticorrosive Pigments), luster pigments (Section Luster Pigments), and transparent pigments (Section Transparent Pigments). Luminescent pigments are described elsewhere, see ⟶ Luminescent Materials.

4.1. Magnetic Pigments

For a complete description of magnetic recording media, see ⟶ Information Storage Materials – Magnetic Recording.

4.1.1. *Iron Oxide Pigments*

Ferrimagnetic iron oxide pigments are used in magnetic information storage systems such as audio and video cassettes, floppy disks, hard disks, and computer tapes (⟶ Information Storage Materials). Cobalt-free iron(III) oxide and nonstoichiometric mixed-phase pigments have been used since the early days of magnetic tape technology. Currently, γ-Fe_2O_3 [1309-37-1] and Fe_3O_4 [1317-61-9] (the latter in small amounts) are mainly used in the production of low-bias audio cassettes [iron oxide operating point IEC I standard (International Electrotechnical Commission)], and studio, broadcasting, and computer tapes.

Production. The shape of the pigment particle is extremely important for ensuring good magnetic properties. Isometric iron oxide pigments produced by direct precipitation are seldom used. Since 1947, needle-shaped γ-Fe_2O_3 pigments have been prepared with a length to width ratio of ca. 5:1 to 20:1 and a crystal length of 0.1 – 1 µm [400].

Anisometric forms of Fe_3O_4 with the spinel structure or γ-Fe_2O_3 with a tetragonal superlattice structure do not crystallize directly. They are obtained from iron compounds that form needle-shaped crystals (usually α- and γ-FeOOH, see Section Iron Oxide Pigments) [401-403]. The oxyhydroxides are converted to Fe_3O_4 by dehydration and reduction. Reducing agents may be gases (hydrogen, carbon monoxide) or organic compounds (e.g., fatty acids). The particle geometry is retained during this process.

Since the pigments are subjected to considerable thermal stress during this conversion, the FeOOH particles are stabilized with a protective coating of sintered material (usually silicates [404], phosphates [405], chromates [406], or organic compounds such as fatty acids [407]).

Finely divided stoichiometric Fe_3O_4 pigments are not stable to atmospheric oxidation. They are therefore stabilized by partial oxidation or by complete oxidation to γ-Fe_2O_3 below 500 °C.

In an alternative process, the starting material consists of needle-shaped particles of α-Fe_2O_3 instead of FeOOH pigments [408], [409]. The synthesis is carried out in a hydrothermal reactor, starting from a suspension of $Fe(OH)_3$, and crystal growth is controlled by means of organic modifiers.

Properties. Magnetic pigments with very different morphological and magnetic properties that depend on the field of application and quality of the recording medium, are used. The largest particles (length ca 0.6 µm) are used in computer tapes. The noise level of the magnetic tape decreases with decreasing particle size. Fine pigments are therefore being used increasingly for better quality compact cassettes.

The magnetic properties may be determined by measurement of hysteresis curves on the powder or magnetic tape (⟶ Magnetic Materials – Parameters of Soft Magnetic Materials).

Table 27 shows some quality requirements for the most important applications of magnetic pigments (⟶ Information Storage Materials – Introduction, ⟶ Information Storage Materials – Principles, ⟶ Information Storage Materials – Magnetic Recording Media). Column 4 gives the coercive field strength (H_c) required for information storage materials. The coercive field is the magnetic field required to demagnetize the sample.

Table 27. Some quality requirements for iron oxide and metallic iron magnetic pigments

Field of application	Pigment type	Approximate particle length, µm	Specific surface area, m^2/g	Coercive field strength H_c, kA/m	Saturation magnetization, M_s, $\mu T \cdot m^3/kg$	M_r/M_s
Computer tapes	γ-Fe_2O_3	0.60	13 – 17	23 – 25	86 – 90	0.80 – 0.85
Studio radio tapes	γ-Fe_2O_3	0.40	17 – 20	23 – 27	85 – 92	0.80 – 0.85
IEC I compact cassettes						
standard (iron oxide operating point)	γ-Fe_2O_3	0.35	20 – 25	27 – 30	87 – 92	0.80 – 0.90
high grade	Co – γ-Fe_2O_3	0.30	25 – 37	29 – 32	92 – 98	0.80 – 0.90
IEC II compact cassettes (CrO_2 operating point)	Co – γ-Fe_2O_3, Co – Fe_3O_4	0.30	30 – 40	52 – 57	94 – 98	0.85 – 0.92
IEC IV compact cassettes (metal operating point)	metallic iron	0.35	35 – 40	88 – 95	130 – 160	0.85 – 0.90
Digital audio (R-DAT) *	metallic iron	0.25	50 – 60	115 – 127	130 – 160	0.85 – 0.90
½″ Video	Co – γ-Fe_2O_3 Co – Fe_3O_4	0.30	25 – 40	52 – 57	94 – 98	0.80 – 0.90
Super-VHS video	Co – γ-Fe_2O_3	0.20	45 – 50	64 – 72	94 – 96	0.80 – 0.85
8-mm video	metallic iron	0.25	50 – 60	115 – 127	130 – 160	0.85 – 0.90

* R-DAT: rotary digital audio tape.

The saturation magnetization M_s is a specific constant for the material and for magnetic iron oxides is principally determined by the Fe^{2+} ion content. The ratio of remanent magnetization to saturation magnetization (M_r/M_s) for the tape depends mainly on the orientation of the pigment needles with respect to the longitudinal direction of the tape, and should approach the theoretical maximum value of unity as closely as possible.

Apart from the morphological and magnetic properties, usual pigment properties such as pH value, tap density, soluble salt content, oil absorption, dispersibility, and chemical stability are of great importance for the manufacture of magnetic recording materials.

Producers of magnetic iron oxides include BASF and Bayer (Germany); Ishihara, Sakai, Showa Denko, Titan K., and Toda K (Japan); 3 M, Mitsui, Magnox, ISK Magnetics (USA); Saehan Media (Korea); and Herdilla (India).

World production of cobalt-free magnetic iron oxides in 1995 was ca. 13 000 t, of which ca. 96 % was used in compact cassettes and audio tapes, and ca. 4 % in computer tapes.

4.1.2. *Cobalt-Containing Iron Oxide Pigments*

Cobalt-containing iron oxides form the largest proportion (ca. 75 %) of magnetic pigments produced today. Due to their high coercivity they can be used as an alternative to chromium dioxide for the production of video tapes, high-bias audio tapes (CrO_2 operating point), and high-density floppy disks.

Production. The iron oxide pigments described in Section Iron Oxide Pigments are either doped or coated with cobalt:

1. Body-doped pigments contain 2 – 5 % cobalt that is uniformly distributed throughout the bulk of the pigment particles. It is either incorporated during production of the FeOOH precursor or precipitated as the hydroxide onto one of the intermediate products [410] using cobalt(II) salts as the cobalt source.
2. Cobalt- coated pigment particles (2 – 4 wt % Co) consist of a core of γ-Fe_2O_3 or nonstoichiometric iron oxide phase, and a 1 – 2 nm coating of cobalt ferrite with a high coercivity [411]. The coating can be produced by adsorption of cobalt hydroxide, or epitaxial precipitation of cobalt ferrite in a strongly alkaline medium [412], [413]. Surface- coated pigments show better magnetic stability than doped pigments.

Properties. Pigments with a coercive field strength of 50 – 56 kA/m are used in video cassettes, high-bias audio cassettes (chromium dioxide operating point IEC II), and high-density floppy disks. Depending on the quality of the tape, the particle size varies between 0.2 and 0.4 µm (see Section Iron Oxide Pigments, Table 27).

Pigments with a higher coercive field strength (ca. 70 kA/m) and smaller particle size (particle length ca. 0.15 – 0.2 µm) are used for super VHS cassettes.

Pigments treated with only small amounts of cobalt (0.5 – 1 wt % Co, coercive field strength ca. 31 kA/m) are used as an alternative to cobalt-free γ-Fe_2O_3 pigments for high-quality low-bias audio cassettes. The other parameters described in Section Iron Oxide Pigments are also of importance for cobalt-treated pigments.

Cobalt- containing pigments are mainly produced by the magnetic iron oxide producers listed in Section Iron Oxide Pigments. World production for 1995 was 40 000 t, of which the highest proportion (ca. 85 %) was used for video tapes.

4.1.3. *Chromium Dioxide*

See also ⟶ Chromium Compounds – Chromium(IV) Oxide (Chromium Dioxide), ⟶ Chromium Compounds – Transportation, Storage, and Handling, ⟶ Chromium Compounds – Environmental Protection, ⟶ Chromium Compounds – Ecotoxicology, ⟶ Chromium Compounds – Toxicology and Occupational Health; ⟶ Information Storage Materials.

In the course of the development of pigments for magnetic information storage, CrO_2 was the first pigment material that gave a higher recording density than γ-Fe_2O_3. In the field of audio recording this led to the IEC II standard or “chrome position”.

Physical Properties. Chromium dioxide [12018-01-8], chromium(IV) oxide, CrO_2, is a ferromagnetic material with a specific saturation magnetization M_s/ϱ of 132 A m^2/kg at 0 K corresponding to the spin of two unpaired electrons per Cr^{4+} ion. The M_s/ϱ value of CrO_2 at room temperature is ca. 100 A m^2/kg [414]; CrO_2 magnetic pigments reach values of 77 – 92 A m^2/ kg. The material crystallizes with a tetragonal rutile lattice in the form of small needles which have the desired magnetic shape anisotropy. The morphology of the particles can be varied with several dopants, particularly antimony and tellurium [415]. The coercive field strength (in addition to shape) can be controlled by doping with transition metal ions which modify the magnetocrystalline anisotropy of the material; the Fe^{3+} ion being industrially important [416]. Depending on the iron content up to ca. 3 mol %, H_c values of CrO_2 may vary between 30 and 75 kA/m [417], [418]. Iron doping also increases the Curie temperature from 115 to ca. 170 °C. Because the Fe^{3+} ions are coupled antiferromagnetically with Cr^{4+} ions in the crystal lattice, the saturation magnetization (M_s/ϱ) of CrO_2 decreases with increasing dopant levels [419].

Other important properties of CrO_2 when used as a magnetic pigment are its black color, electrical conductivity (2.5 – 400 Ω^{-1} cm^{-1} [420]) and relatively high crystal hardness (Mohs hardness 8 – 9 [424]). Therefore, coating formulations based on CrO_2 require less or even no additives such as carbon black (good conductivity, black color) or refractory oxides such as alumina.

Chemical Properties. Pure CrO_2 slowly disproportionates in the presence of water. The CrO_2 crystal surface of commercial pigments is therefore topotactically converted to β-CrOOH which serves as a protection layer [421]. The conditions for this reaction have a great influence on the chemical properties of the pigment surface [422], [423]. Thus, the isoelectric point in water can vary between pH 3 and 7 [424], which affects dispersibility. In the absence of moisture, CrO_2 is stable up to ca. 400 °C; above this temperature it decomposes to form Cr_2O_3 and oxygen.

Uses and Economic Aspects. Chromium dioxide is used exclusively for magnetic recording media, e.g., tapes for audio, video, and computer applications. An application of particular interest depends on its relatively low Curie temperature. This allows thermomagnetic duplication at temperatures low enough for the base polymer of magnetic tapes, and is exploited in a commercial, high-speed copying process (⟶ Information Storage Materials – Video Recording) [425]. Chromium dioxide may also be used in combination with cobalt-modified iron oxides (see Section Cobalt-Containing Iron Oxide Pigments) in the production of magnetic recording media. The world production of CrO_2 in 1995 amounted to 8000 t, ca. 11 % of the total consumption of magnetic pigments. Producers are BASF and Du Pont.

4.1.4. *Metallic Iron Pigments*

The magnetization of iron is more than three times higher than that of iron oxides. Metallic iron pigments can have a coercive field strength as high as 150 kA/m, depending on particle size. These properties are highly suitable for high-density recording media (⟶ Information Storage Materials – Metal Particles). Oxidation-resistant products based on metallic pigments first became available in the late 1970s.

Production. Metallic iron pigments are commercially produced by the reduction of acicular (needle-shaped) iron compounds [426]. As in the production of magnetic iron oxide pigments, the starting materials are iron oxide hydroxides (see Section Iron Oxide Pigments) or iron oxalates, which are reduced to iron in a stream of hydrogen either directly or via oxidic intermediates.

Due to their high specific surface area, metallic pigments are pyrophoric, so that passivation is necessary. This can be achieved by slow, controlled oxidation of the particle surface [427].

Properties. The coercive field strength of metallic iron pigments is primarily determined by their particle shape and size, and can be varied between 30 and 150 kA/m. Pigments for analog music cassettes (H_c≈90 kA/m) usually have a particle length of 0.35 µm (see Section Iron Oxide Pigments, Table 27). The length to width ratio of the pigment needles is ca. 10:1. Finely divided pigments (particle length ca. 0.25 µm) with a coercive field strength of 120 kA/m are used for 8-mm video and digital audio cassettes (R-DAT), tapes used by television organizations (ED Beta, Betacam SP MP, M II, Digital Video D 2), and for master video cassettes (mirror master tapes). In the field of data storage, a small quantity is used in micro floppy disks.

Metallic pigments have a higher specific surface area (up to 60 m^2/g) and a higher saturation magnetization than oxidic magnetic pigments. Their capacity for particle alignment corresponds to that of the oxides (see Section Iron Oxide Pigments, Table 27).

Economic Aspects. The largest producers of metallic iron pigments are Dowa Mining, Kanto Denka K., and Mitsui Toatzu.

World consumption in 1995 was ca. 1800 t, of which ca. 75 % was used in the manufacture of video tapes and ca. 14 % for audio tapes, and 11 % for data storage applications. Consumption is expected to increase.

4.1.5. *Barium Ferrite Pigments*

Barium ferrite pigments have been considered for several years for high-density digital storage media [428], [429]. They are very suitable for preparing unoriented (e.g., floppy disks), longitudinally oriented (conventional tapes), and perpendicularly oriented media. In the latter the magnetization is oriented perpendicular to the coating surface. They are required for perpendicular recording systems which promise extremely high data densities, especially on floppy disks (⟶ Information Storage Materials – Perpendicular Recording). Barium and strontium ferrites are also used to prevent forgery of magnetic stripes, e.g., in cheque and identity cards.

Properties. Hexagonal ferrites have a wide range of structures distinguished by different stacking arrangements of three basic elements known as M, S, and Y blocks [430]. For magnetic pigments, the M-type structure (barium hexaferrite [12047-11-9] $BaFe_{12}O_{19}$) is the most important. The magnetic properties of M-ferrite can be controlled over a fairly wide range by partial substitution of the Fe^{3+} ions, usually with combinations of di- and tetravalent ions such as Co and Ti. Barium ferrite crystallizes in the form of small hexagonal platelets. The preferred direction of magnetization is parallel to the *c*-axis and is therefore perpendicular to the surface of the platelet. The specific

saturation magnetization of the undoped material is ca. 72 A m^2/kg and is therefore somewhat lower than that of other magnetic oxide pigments. In barium ferrite the coercive field strength is primarily determined by the magnetocrystalline anisotropy and only to a limited extent by particle morphology. This is the reason why barium ferrite can be obtained with extremely uniform magnetic properties. Barium ferrite pigments have a brown color and chemical properties similar to those of the iron oxides.

Production. There are three important methods for manufacturing barium ferrite on an industrial scale: the ceramic, hydrothermal, and glass crystallization methods. The main producers are Toshiba and Toda.

Ceramic Method. Mixtures of barium carbonate and iron oxide are reacted at 1200 – 1350 °C to produce crystalline agglomerates which are ground to a particle size of ca. 1 µm. This method is only suitable for the high- coercivity pigments required for magnetic strips [431].

Hydrothermal Method. Iron [Fe(III)], barium, and the dopants are precipitated as their hydroxides and reacted with an excess of sodium hydroxide solution (up to 6 mol/L) at 250 –350 °C in an autoclave. This is generally followed by an annealing treatment at 750 – 800 °C to obtain products with the desired magnetic properties. Many variations of the process have been described [432-436], the earliest report being from 1969 [437]. In later processes, hydrothermal synthesis is followed by coating with cubic ferrites, a process resembling the cobalt modification of iron oxides (see Section Cobalt-Containing Iron Oxide Pigments). The object is to increase the saturation magnetization of the material [438-440].

Glass Crystallization Method. This process was developed by Toshiba [441]. The starting materials for barium ferrite production are dissolved in a borate glass melt. The molten material at ca. 1200 °C is quenched by pouring it onto rotating cold copper wheels to produce glass flakes. The flakes are then annealed to crystallize the ferrite in the glass matrix. In the final stage the glass matrix is dissolved in acid. In a variation of this process, the glass matrix is produced by spray drying [442].

Magnetic Recording Properties. Typical values of physical properties of barium ferrite pigments used in magnetic recording are given in Table 28.

Table 28. Typical properties of barium ferrite pigments

Application	Specific surface area, m^2/g	Platelet diameter, nm	Platelet thickness, nm	H_c, kA/m	M_s/ϱ, A m^2/kg
Unoriented (floppy disk)	25 – 40	40 – 70	15 – 30	50 – 65	50 – 65
Oriented	25 – 60	40 – 120	10 – 30	55 – 100	50 – 65
Magnetic strips	12 – 15	100 – 300	50 – 100	220 – 440	60 – 70

Barium ferrite is highly suitable for high-density digital recording mainly because of its very small particle size and its very narrow switching field distribution. It also has a high anhysteretic susceptibility and is difficult to overwrite [443]. This is partly explained by positive interaction fields between particles in the coating layer [444]. The high anhysteretic susceptibility makes barium ferrite media particularly suitable for the anhysteretic (bias field) duplicating process (⟶ Information Storage Materials – Media Requirements for Digital Recording) [445]. Unlike many other magnetic materials used in high-density recording, barium ferrite, being an oxide, is not affected by corrosion [446]. Processing of the pigment can be problematic, e.g., applying orienting fields can easily lead to unwanted stacking of the particles which has adverse effects on the noise level and the coercive field strength of the magnetic tape. A marked temperature dependence of the magnetic properties was a problem in the early days, but this can be overcome by appropriate doping [447].

4.2. Anticorrosive Pigments

4.2.1. *Principles*

Metal corrosion and the relevant electrochemical theory are described elsewhere ⟶ Corrosion. This chapter deals with the protection of metal surfaces against corrosion by means of coatings that contain anticorrosive pigments. The degree of corrosion protection depends not only on the pigment, but also on the binder, and these must complement each other chemically. The uses of anticorrosive pigments in paint binders are summarized in Table 29. For standards, see Table 1 ("Corrosion testing").

Table 29. Possible uses of anticorrosive pigments in various primers

Anticorrosive pigment	Air-drying alkyd primers	Alkyd–melamine primers	2-Component epoxy primers	2-Component acrylic isocyanate primers	Chlorinated rubber primer	Aqueous polymer dispersions	Poly (vinyl-butyral) primers	Electro-deposition coatings	Mirror coatings **	References
Zinc phosphate	+	+	+	+	+	+	o	o	–	[448-452]
Chromium phosphate	+	o	–	–	–	–	–	o	–	[453]
Aluminum phosphate	+	+	+	–	+	+	+	–	–	[449], [451], [454], [455]
Zinc phosphomolybdate	+	+	+	–	+	+/o	+	o	–	[449], [452], [456]
Zinc calcium phospho-molybdate	+	–	+	–	–	+/o	–	+/o	–	[456]
Phosphate multiphase pigments (see Table 33)										
Product 1	+	+	o	+	+	+	o	o	–	[448]
Product 2	+	+	+	+	+	+	o	o	–	[448]
Product 3	+	+	+	–	+	+	+	o	–	[449]
Zinc hydroxyphosphite	+	–	+	–	+	+	–	–	–	[457]
Lead cyanamide	–	+	–	–	–	–	–	+	+	[448]
Zinc cyanamide	–	+	–	–	–	–	–	o	+	[448], [458]
Barium metaborate	+	+	–	–	–	+	–	–	–	[459], [460]
Zinc borophosphate	o	–	o	–	–	+	+	–	–	[448]
Zinc potassium chromate	+	+	+	+	+	+	+	–	–	[451]
Zinc tetrahydroxy-chromate	+	+	–	–	–	+	+	–	–	[451]
Strontium chromate *	+	+	+	–	–	–	+	+	–	[451]
Red lead	+	+	+	–	+	–	+	–	o	
Zinc salt of 5-nitro-isophthalic acid	+	+	+	+	+	o	+	o	–	[448]

[a] + Commonly used (good anticorrosive properties);
[b] o Anticorrosive properties depend on the formulation;
[c] – Seldom used.
[*] Mainly on aluminum.
[**] Mirror coatings are used to protect the thin metal surface (silver, copper) of metal-coated glass against oxidation or reaction with H_2S or SO_2.

Anticorrosive pigments may be divided into three types [461]:

1. *Pigments with a physical protective action* are chemically inert and are termed *inactive* or *passive*. An example is micaceous iron ore [462], [463]. These lamellar pigments are packed in layers; they lengthen the pathways and obstruct the penetration of ions. They improve adhesion between the substrate and the coating, absorb UV radiation, and protect the underlying binder (Fig. 42), [464-468].
2. *Pigments with chemical protective action* contain soluble components and can maintain a constant pH value in the coating. They are termed *active* and their action depends on reactions in the interfacial areas between the pigment and substrate, pigment and binder, or between pigment and ions that penetrate into the coating [469], [470]. An example is red lead. Redox reactions can occur to form protective compounds (oxides or oxide hydrates that may contain pigment cations). Saponification of the binder or neutralization of the acidic decomposition products enable a definite pH value to be maintained in the coating.
3. *Pigments with an electrochemical protective action* passivate the metallic surface [471], [472]. Those that prevent corrosion of the iron by forming a protective coating (e.g., phosphate pigments) are regarded as being active in the anodic region of the metal surface (anodic protection) [473]. Pigments that prevent rust formation due to their high oxidation potential (e.g., chromates) are said to be active in the cathodic region (cathodic protection) [474].

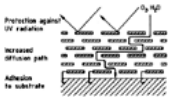

Figure 44. Schematic representation of the physical anticorrosion effects of flake pigments

Corrosion of iron is explained by the position of iron in the electrochemical series of the elements (Fe/Fe^{2+}: – 0.44 V). In steel, local anode and cathode areas are found due to the presence of phases containing, for example, carbon, carbides, and oxides. These latent local cells are activated by moisture, oxygen, and current- carrying electrolytes and the following reactions occur between the anode areas consisting of iron, and the cathode areas containing carbides or oxides.

Anode:

$$Fe \rightarrow Fe^{2+} + 2e^-$$

$$2Fe^{2+} + O_2 + 2H_2O + 2OH^- \rightarrow 2Fe(OH)_3 \rightarrow 2FeO(OH)[Rust] + 2H_2O$$

Cathode:

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^-$$

The rust can promote further corrosion. The OH^- ions formed at the cathode produce local high alkalinity and cause hydrolysis of, for example, esterified binders, resulting in detachment of the primer from the substrate [465].

Active anticorrosive pigments inhibit one or both of the two electrochemical partial reactions. The protective action is located at the interface between the substrate and the primer. Water that has diffused into the binder dissolves soluble anticorrosive components (e.g., phosphate, borate, or organic anions) out of the pigments and transports them to the metal surface where they react and stop corrosion. The oxide film already present on the iron is thereby strengthened and sometimes chemically modified. Any damaged areas are repaired with the aid of the active substance. Inhibition by formation of a protective film is the most important mode of action of the commoner anticorrosive pigments.

4.2.2. *Phosphate Pigments*

The most important phosphate- containing pigments are zinc phosphate [448-450], [452]; basic zinc phosphate [451], [452]; chromium phosphate [453]; aluminum triphosphate [454]; barium phosphate [461]; aluminum zinc phosphate [449]; and zinc/iron phosphate [452]. Phosphate ions form protective coatings of basic iron(III) phosphate on an iron surface. Their composition is described in [472]. The phosphating of metal surfaces is described elsewhere, ⟶ Metals, Surface Treatment – Phosphating.

The reactions of phosphate pigments in binders and on metal surfaces were described by G. MEYER [475], and J. RUF [476]. They postulate that basic complexes are formed from phosphate pigments containing water of crystallization and they react with inorganic ions or with carboxyl groups of the binder. Reaction of the pigment with the binder results in physical changes such as deterioration of adhesion or gelling. In paints that contain hydroxy- and carboxy-functional binders, gelling during storage can be prevented by using an excess of alcohol as the solvent.

Toxicological data for phosphate pigments are described in Section Toxicology.

4.2.2.1. *Zinc Phosphate [448-452]*

The most important phosphate- containing pigment is zinc phosphate [7779-90-0], $Zn_3(PO_4)_2 \cdot 4H_2O$, M_r 458.1. It can be used with a large number of binders (see Table 29) and has a very wide range of uses [448], [449], [451], [452], [477].

Zinc phosphate is usually produced on an industrial scale from zinc oxide and phosphoric acid, or from zinc salts and phosphates [478]. Composition and properties are given in Table 30.

Table 30. Composition and properties of zinc and chromium phosphate pigments

Property (standard)	Commercial products		
	Zinc phosphate $Zn_3(PO_4)_2 \cdot 2 – 4 H_2O$	Micronized zinc phosphate $Zn_3(PO_4)_2 \cdot 2 – 4 H_2O$	Chromium phosphate $CrPO_4 \cdot 3 H_2O$
Metal content, wt % (ISO 787, part 2)	51	51	ca. 27
Phosphate content, wt % (ISO 787, part 2)	49	< 49	ca. 47
Chloride content, wt % (ISO 787, part 13)	0.01	0.05	< 0.1
Sulfate content, wt % (ISO 787, part 13)	< 0.1	0.05	0.04
Water-soluble content, wt % (ISO 787, part 3)	< 0.1	0.1	0.4
Sieve residue, wt % (ISO 787, part 18)	0.01	0.01	max. 0.3
Density, g/cm^3 (ISO 787, part 10)	3.2	3.4	2.5

Specific surface area, m^2/g (DIN 66 131/66 132)	1.0	4.3	22
Loss on ignition, wt % (ISO 787, part 2)	10 – 11	13	26
Oil absorption value, g/100 g (ISO 787, part 5)	20	24	65
Conductivity, µS (ISO 787, part 14)	100	300	ca. 500
pH (ISO 787, part 9)	6.2 – 7	7	6.5
Color	white-beige	white-beige	green

The mechanism of the action of zinc phosphate is shown in Figure 43. Zinc phosphate dihydrate pigment is hydrated to the tetrahydrate in an alkyd resin binder [479]. The tetrahydrate is then hydrolyzed to form zinc hydroxide and secondary phosphate ions which form a protective film of basic iron(III) phosphate on the iron surface [475]. The anticorrosive action of zinc phosphate depends on its particle size distribution. Micronization improves the anticorrosive properties [480-482].

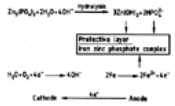

Figure 45. Passivation of iron by zinc phosphate [475]

The effect of corrosion-promoting ions on the anticorrosive properties of zinc phosphate is described in [483], [484].

Trade names are as follows:

Zinc phosphate: Sicor ZNP/M, ZNP/S (BASF, Germany); Heucophos ZP 10 (Dr. H. Heubach, Germany); Halox Zinc Phosphate (Halox Pigments, USA); Phosphinox PZ 20 (SNCZ, France); Hispafos N 2 (Colores Hispania, Spain).

Basic zinc phosphate: Heucophos ZPO (Dr. H. Heubach, Germany); Hispafos SP (Colores Hispania, Spain).

4.2.2.2. *Aluminum Phosphate*

Commercial aluminum phosphate anticorrosive pigments consist of aluminum zinc phosphate hydrates [66905-65-5], [449], [451], or zinc- containing aluminum triphosphate [13939-25-8], [454]. Their compositions and properties are listed in Table 31. Their applications in commercial binders are listed in Table 29. Aluminum zinc phosphate hydrate pigments are produced by reacting acid solutions of aluminum hydrogen phosphate with zinc oxide and alkali aluminate. The precipitated pigment is filtered off from the mother liquor, washed, dried, and ground [485]. Commercial aluminum triphosphate pigments contain ions of trimeric phosphoric acid which form stable aluminum-containing iron phosphate complexes [454].

Table 31. Composition and properties of aluminum phosphate anticorrosive pigments

Property	Aluminum triphosphate (commercial product 1; K-White 82) [454]	Aluminum triphosphate (commercial product 2; K-White 84) [454]	Aluminum zinc phosphate hydrate (commercial product 3; Heucophos ZPA, Phosphinal PZ 04) [449], [451]
Al content, wt %	5.5 – 7.7	4.7 – 6.9	4 – 5
Zn content, wt %	11.6 – 14.9	21.3 – 24.5	35 – 39
P_2O_5 content, wt %	42.0 – 46.0	36.0 – 40.0	30/37
SiO_2 content, wt %	13.0 – 17.0	11.0 – 15.0	
Loss on ignition (600 °C), wt %	ca. 10	ca. 8	10 – 16
Density, g/cm^3	3.0	3.1	3.2
Oil absorption value, g/100 g	35	30	30/35
Particle size, µm	0.5 – 10	0.5 – 10	0.5 – 10
pH value	6 – 7	6 – 7	6.5
Water-soluble content, wt %	max. 1.0	max. 1.0	0.1
Color	white	white	white

The aluminum phosphate pigments give good adhesion of the paint film to the metallic substrate.

Trade names are as follows:

Zinc aluminum phosphate: K-White 105 (Teikoku Kako, Japan); Heucophos ZPA (Dr. H. Heubach, Germany); Phosphinal PZ 04 (SNCZ, France).

Aluminum triphosphate: K-White 82, -84 (Teikoku Kako, Japan).

4.2.2.3. *Chromium Phosphate [453], [461]*

Chromium phosphate [16453-74-0], $CrPO_4 \cdot 3\,H_2O$, M_r 201.0, is produced from chromium(III) salts and alkali phosphates. The physical and chemical properties are listed in Table 30; important areas of use are given in Table 29 [453].

Chromium phosphate has a low solubility. It is therefore nearly always used in combination with other anticorrosive pigments. It is an extremely good long-term inhibitor, but is less effective during the initial phase of corrosion protection.

4.2.2.4. *New Pigments Based on Metal Phosphates*

Developments in the field of phosphate pigments are aimed at correcting the deficiencies of phosphate (such as the moderate corrosion protection against accelerated weathering in the salt spray test) during the initial phase of corrosion protection. Improvements include crystallization to form new structures, replacement of cations (e.g., of zinc by calcium), modification of anions (phosphovanadates and phosphocarbonates), optimization of the particle size distribution, providing the zinc ions and phosphate ions in separate compounds, and use of synergistic effects. New pigments are listed in Table 32. Some of them are commercially available, e.g., the zinc molybdatophosphates and zinc calcium molybdatophosphates. Trade names include Heucophos ZMP [449], Moly-White ZNP and MZA [456], and Actirox 102 and 106 [452], [497].

Table 32. New anticorrosive metal phosphate pigments

Composition	Binder	Metal substrate Reference

$CaZn_2(PO_4)_2 \cdot 2\ H_2O$	alkyd resin	iron	[486]
$CaHPO_4 \cdot 2\ H_2O/ZnO$	chlorinated rubber polyvinyl butyral	aluminum	[487]
$CaHPO_4 \cdot 2\ H_2O/ZnO$	phenolic resin alkyd resin	iron	[488]
$MgHPO_4 \cdot 3\ H_2O/ZnO$	polyvinyl butyral	aluminum	[487]
$MgHPO_4 \cdot 3\ H_2O/ZnO$	phenolic resin alkyd resin	iron	[488]
$CaHPO_4 \cdot 2\ H_2O/MgHPO_4 \cdot 3\ H_2O$	alkyd resin	iron	[489]
$[0.5\ MgHPO_4 \cdot 0.5\ MgCO_3] \cdot x\ H_2O$	epoxy polyamide	aluminum	[490]
$[0.4\ SrHPO_4 \cdot 0.6\ SrCO_3] \cdot x\ H_2O$+0.1 – 3 % fluoride	epoxy polyamide	aluminum	[490]
$TiO_2/ZrO_2/SiO_2 \cdot x\ P_2O_5$	alkyd – melamine	iron	[491]
$(Fe_x,Cr_y,Na_z)PO_4 \cdot x\ H_2O$	epoxy resin	iron	[492]
$Al(H_2PO_4)_3 \cdot x\ H_2O/ZnO$	alkyd resin		[493]
Phosphate glasses containing			
ZnO/Al_2O_3	alkyd resin		[494]
MgO, SiO_2, Al_2O_3, CaO, B_2O_3	alkyd resin		[495]
Mg/Ca phosphovanadate	epoxy, alkyd – melamine	iron	[496]
$[Zn_3(PO_4)_2 \cdot 2 – 4\ H_2O] \cdot [x\ ZnMoO_4]$	alkyd	iron	[449]
$(Zn,Ca)_3(PO_4,MoO_4)_2 \cdot x\ H_2O$	alkyd, epoxy	iron	[456]

4.2.2.5. *Multiphase Phosphate Pigments*

Multiphase anticorrosive phosphate pigments are based on appropriate combinations of inorganic phosphates with sparingly soluble, electrochemically active organic corrosion inhibitors [448]. Examples include Sicor SPO (zinc phosphate with zinc salt of 5-nitroisophthalic acid, BASF) [448]; Sicor NOP (zinc phosphate with calcium salt of 5-nitroisophthalic acid, BASF) [448]; and Heucophos ZPZ (basic zinc phosphate with 0.3 % organic components, Heubach) [449]. The multiphase pigments are produced in one combined synthesis. A slurry of precipitated zinc phosphate is mixed intensively with the zinc or calcium salt of 5-nitroisophthalic acid (Section Organic Pigments), filtered, and washed and ground until the specified conductivity is reached. They are more effective than physical mixtures of pigments produced in separate processes. Multiphase phosphate pigments are used in place of zinc chromate if the anticorrosive effect of pure zinc phosphate is insufficient in the initial weathering phase (e.g., in the salt spray test). The very good anticorrosive effect of multiphase pigments in air-drying alkyd resins or in alkyd – melamine stoving enamels is comparable to that of zinc chromate, even in open air weathering. Commercial multiphase pigments differ mainly in the solubility of the organic corrosion inhibitor. Systems with more soluble inhibitors are used in the binders with a low swelling capacity (e.g., epoxy resins or chlorinated rubber).

The compositions and properties of some commercial multiphase anticorrosive pigments are listed in Table 33. Their anticorrosive properties in various binders are shown in Table 29.

Table 33. Multiphase anticorrosive phosphate pigments (commercial products)

Property	Commercial product 1 (Sicor SPO, BASF) [448]	Commercial product 2 (Sicor NOP, BASF) [448]	Commercial product 3 (Heucophos ZPZ, Heubach) [449]
Zinc content, wt %	63	62	58
Phosphate content, wt %	28	30	39
Organic inhibitor content, wt %	2.5	0.5	0.3
Loss on ignition, wt %	9	9	10 – 12
Water-soluble content, wt %	0.9	0.5	0.8
Chloride/sulfate content, wt %	0.02	0.02	<0.05
pH	7	7	7
Conductivity, µS	260	240	650
Density, g/cm^3	3.7	3.8	3.7
Specific surface area, m^2/g	7	5	
Oil absorption value, g/100 g	ca. 20	ca. 20	16
Color	white	white	white
CAS registry no.	[92308-56-0]	[92308-55-9]	[34807-26-6]

Trade names include Sicor SPO, -NOP (BASF, Germany); Heucophos ZPZ (Dr. H. Heubach, Germany); and Novinox PZ 02 (SNCZ, France).

4.2.3. ***Other Phosphorus-Containing Pigments***

Iron Phosphide. Commercial iron phosphide anticorrosive pigments usually consist of Fe_2P [12751-22-3], M_r 142.7, with traces of FeP and SiO_2. The pigment is a powder with a metallic gray color and contains 70 % Fe, 24 % P, 2.5 % Si, and 3.0 % Mn. The density is 6.53 g/cm^3 and the mean particle size is ca. 3 – 5 µm.

Iron phosphide anticorrosive pigments are recommended by manufacturers as replacement materials for zinc dust to reduce the price of zinc-rich paints [498]. For toxicological data, see Section Toxicology.

A *trade name* for iron phosphide is Ferrophos (Hooker Chemicals & Plastics Corp., USA).

Zinc Hydroxyphosphite. Commercial zinc hydroxyphosphite [64539-51-1] is a white, nontoxic pigment with basic character.

The pigment has the following physical properties [457]:

Oil absorption value, g/100 g	15 – 20
Density, g/cm^3	3.96
pH, 2 % suspension	6.5 – 7.5
Mean particle size, µm	2 – 3

Water-soluble content, g/100 cm^3	< 0.01
Specific conductivity, Ω/cm	9700

Its anticorrosive properties in various binders are given in Table 29.

A *trade name* for zinc hydroxyphosphite is Nalcin 2 (National Lead Chemicals, USA; Kronos Titan, Germany).

4.2.4. *Borosilicate Pigments*

Borosilicate pigments [59794-15-19] usually contain calcium or zinc ions in a matrix of silicon dioxide and boron trioxide, x (Ca,Zn) · y SiO_4 · z BO_3. Aqueous slurries of this white pigment are alkaline (pH > 9). Borosilicate pigments are recommended as nontoxic alternatives to basic lead silicate. Their main field of application is in waterborne binders and electrodeposition coatings [450], [499], [500].

A *trade name* for calcium borosilicate is Halox CW 291, -2230 (Halox Pigments, USA).

4.2.5. *Borate Pigments*

Barium metaborate [13701-59-2], BaO · B_2O_3 · H_2O, [459] and zinc borophosphate [84012-98-6], ZnO · x B_2O_3· y P_2O_5 · 2 H_2O, are colorless pigments. Their properties are listed in Table 34; their uses in various binders are listed in Table 29. They both have a relatively high water solubility. Barium metaborate is coated with silica to reduce this.

Table 34. Properties of metal borate pigments

Property	Barium metaborate	Zinc borophosphate
Density, g/cm^3	ca. 3.3	2.8
pH	9 – 10	8
Water-soluble content, wt %	0.4	1 – 1.5
Use	primer, topcoat	primer

The anticorrosive effects of these pigments depend mainly on their ability to maintain high pH values in the coating. They are most effective in the initial phase of corrosion protection. The borate ion neutralizes acidic foreign ions and binder decomposition products, but is also thought to act as an anodic passivator, forming a protective film. The effectiveness of metal borate pigments in aqueous air-dried anticorrosive coatings is described in [501].

The metal borate pigments are classified as having a relatively low toxicity [461], [502]. For toxicological data, see Section Toxicology.

Examples of *trade names* are Butrol (Buckmann Laboratories, USA) for barium metaborate and Sicor BZN (BASF, Germany) for zinc borophosphate.

4.2.6. *Chromate Pigments*

Anticorrosive chromate pigments are summarized in Table 35. Their possible combinations with various binders are listed in Table 29. Standards for specifications of chromate pigments are listed in Table 1.

Table 35. Metallic anticorrosive chromate pigments

Name	Formula	CAS registry no.	Synonym	Appearance	Solubility in water, g/L
Zinc chromate	$ZnCrO_4$	[13530-65-9]	chromic acid – zinc salt (1:1) zinc yellow C.I. Pigment Yellow 36	lemon yellow powder	insoluble
Zinc tetraoxychromate	$ZnCrO_4$ · 4 Zn $(OH)_2$	[15930-94-6]	basic zinc chromate ZTO chromate zinc tetrahydroxychromate C.I. Pigment Yellow 36	yellow powder	ca. 0.04
Basic zinc potassium chromate	4 ZnO · K_2O ·4 CrO_3 · 3 H_2O	[37300-23-5]	basic zinc chromate zinc chromate pigment lemon yellow C.I. Pigment Yellow 36	lemon yellow triclinic flakes	2.5 – 5
Zinc potassium chromate	$KZn_2(CrO_4)_2OH$	[11103-86-9]	chromic acid – potassium zinc salt (2 : 2 : 1) zinc yellow C.I. Pigment Yellow 36	yellow powder	sparingly soluble
Strontium chromate	$SrCrO_4$	[7789-06-2]	strontium chromate A strontium yellow C.I. Pigment Yellow 32	yellow powder	ca. 2

The anticorrosive action of the chromate pigments is based both on chemical and electrochemical reactions [474], [503-508]. Electrochemical passivation and chemical reaction are illustrated in Figure 46 [509], [510]. Passivation is based on electrochemical processes in the cathodic region. In addition, a protective film is also formed by reaction of chromate ions with metal ions at the surface of the substrate to form metal oxide hydrates.

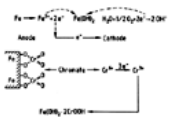

Figure 46. Passivation of iron by chromate pigments [509], [510]

The anticorrosive properties of this class of pigments depend on:

1. The content of water-soluble chromate ions
2. The ratio of water-soluble chromate ions to water-soluble corrosion-promoting ions (chloride and sulfate ions)
3. The active pigment surface in the coating (i.e., particle size distribution and dispersibility)

Chromate- containing pigments are classified as toxic; their use is therefore very limited and they must be appropriately labeled [511], [512]. For a more detailed discussion

of the toxicology of chromates, see Section Toxicology and Occupational Health and ⟶ Chromium Compounds – Toxicology and Occupational Health.

Zinc-Containing Chromate Pigments. Zinc chromate (⟶ Chromium Compounds), is produced by reacting an aqueous slurry of zinc oxide or hydroxide with dissolved chromate ions followed by neutralization, or by precipitation of dissolved zinc salts with dissolved chromate salts. Zinc tetraoxychromate is produced from zinc oxide and chromic acid in an aqueous medium. Basic zinc potassium chromate is obtained by reacting an aqueous slurry of zinc oxide with potassium dichromate and sulfuric acid. The pigments are washed, filtered, dried, and ground.

Trade names include Zinkchromat CZ 20 – CZ 40 (SNCZ, France); Zinkchromat 1W (BASF, Germany); and Zinktetraoxichromat TC 20 – TC 40, LOW DUST (SNCZ, France).

Strontium Chromate. Strontium chromate is precipitated from solutions of sodium dichromate and strontium chloride, followed by filtration, washing, drying, and grinding. A primer composition based on strontium and calcium chromates is described in [513], [451].

A *trade name* for strontium chromate pigments is Strontiumchromat L 203 S – L 203 E (SNCZ, France).

Lead Silicochromate. [514]. Lead silicochromate [11097-70-4], 4 ($PbCrO_4 \cdot PbO$) + 3 ($SiO_2 \cdot 4\ PbO$), is an orange powder. This pigment is a core pigment, in which the active pigment substance ($PbCrO_4$) is precipitated onto an inert core (SiO_2).

In the production process used by National Lead Industries, a solution of chromic acid is added to an aqueous slurry of lead oxide, finely ground silica, and a small amount of basic lead acetate:

$$2\,PbO + CrO_3 \rightarrow PbCrO_4 \cdot PbO$$

The product is heated in a rotary kiln causing the excess lead oxide to react with the basic lead chromate, forming tetrabasic lead chromate:

$$3\,PbO + PbCrO_4 \cdot PbO \rightarrow PbCrO_4 \cdot 4\,PbO$$

On raising the temperature the following reaction takes place on the surface of the silica:

$$4(PbCrO_4 \cdot 4\,PbO) + 3\,SiO_2 \rightarrow 4(PbCrO_4 \cdot PbO) + 3\,(SiO_2 \cdot 4\,PbO)$$

The pigment contains 47.0 wt % PbO, 5.4 wt % CrO_3, and 47.6 wt % SiO_2. Properties are as follows: density 4.2 g/cm^3, oil number 15 g/100 g, tamped volume 163 cm^3 /100 g, sieve residue (> 42 µm) 0.004 wt %, moisture content 0.1 wt %, and pH ca. 7.

A *trade name* for lead silicochromate is Onor (National Lead Chemicals, USA) [457].

4.2.7. *Molybdate Pigments*

Molybdenum-based anticorrosive pigments offer a nontoxic alternative to the zinc chromate pigments [515]. They all have a neutral color (white) but the pure compounds are very expensive. To produce economically competitive pigments molybdate and phosphate pigments are combined, or molybdate compounds are applied to inorganic fillers (e.g., zinc oxide, alkaline-earth carbonates, or talc) [456], [516-518].

Commercially important pigments include basic zinc molybdates ($ZnMoO_4$ [13767-32-3], ZnO [1314-13-2], sodium zinc molybdates ($Na_2Zn(MoO_4)_2$ [11139-48-3]), basic zinc calcium carbonate molybdate, and basic calcium zinc carbonate phosphate molybdate pigments ($CaMoO_4$ [7789-82-4], $CaCO_3$ [471-34-1], ZnO [1314-13-2], $Zn_3(PO_4)_2$ [7779-90-0]). Properties are given in Table 36.

Table 36. Properties of molybdenum-containing pigments

Property	Zinc molybdate pigment	Calcium zinc molybdate pigment	Basic sodium zinc molybdate pigment	Basic calcium zinc phosphomolybdate pigment
Density, g/cm^3	5.06	3.00	4.00	3.00
Oil absorption value, g/100 g	14	18	14	16
Mean particle size, µm (Fisher subsieve sizer)	0.65	1.88	1.02	2.2
pH	6.5	8.5		8

Phosphate- containing molybdate pigments are especially suitable for water-thinnable or latex-based binders, because they improve adhesion to iron substrates. The other molybdate pigments are mainly used in solventborne binder systems.

Unlike chromate ions in chromate pigments, the MoO_4^{2-} ions in molybdate pigments are not chemically reduced in most coatings. Hence, they are ineffective for cathodic protection. Their protective action is assumed to be due to activity in the anodic region, similar to that of phosphate ions. As with the protective phosphate films, molybdate films are very resistant to chloride and sulfate [519]. The duration of maximum activity depends on the metal ions used in the pigment and is probably due to differences in solubility.

Trade names are as follows:

Molybdenum-based: Moly-White 101, 212, 331 (The Sherwin-Williams Company, USA).

Zinc molybdenum phosphate: Actirox 102, -106 (Colores Hispania, Spain); Moly-White ZNP, -MZA (The Sherwin-Williams Company, USA); Heucophos ZMP (Dr. H. Heubach, Germany).

4.2.8. *Lead and Zinc Cyanamides*

Lead cyanamide [20837-86-9], $PbCN_2$, M_r 247.23, is a lemon yellow powder. Zinc cyanamide [20654-08-4], $ZnCN_2$, M_r 105.41, is a white to beige powder. Their properties are listed in Table 37 [520].

Table 37. Properties of lead and zinc cyanamides

Property	Lead cyanamide	Zinc cyanamide
Density, g/cm^3	ca. 6.8	ca. 3.1
Apparent density, g/cm^3	550 – 750	ca. 250
Metal content, wt %	ca. 83	ca. 60
CN_2 content, wt %	ca. 16	ca. 34
Specific surface area, m^2/g	2	ca. 50
pH	9 – 11	8.5
Conductivity, µS	< 300	< 1000

The cyanamides are active anticorrosive pigments which have a passivating action under alkaline conditions. The action of lead in anticorrosive pigments is discussed in Section Red Lead. Heavy metal oxides, especially of iron and manganese, catalyze the conversion of cyanamide to urea even below 20 °C [521]:

$$H_2N-C\equiv N + H_2O \rightarrow H_2N-CO-NH_2$$

At alkaline pH, however, the dimerization of cyanamide to dicyanodiamide is favored. Dicyanodiamide acts as a corrosion inhibitor:

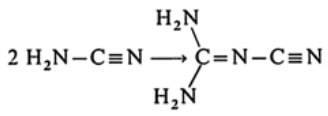

In strongly alkaline solution, hydrolysis to ammonia takes place:

$$H_2N-C\equiv N + H_2O + 2\,OH^- \rightarrow 2\,NH_3 + CO_3^{2-}$$

The main fields of application are in mirror coatings [458], [520], electrodeposition coatings, and primers (see Table 29).

Cyanamide pigments are produced from industrial-grade calcium cyanamide which is first dissolved. Sulfide and phosphide impurities are precipitated as iron or lead salts [522-527] or oxidized [528-530] and filtered off together with graphite impurities. The pure calcium cyanamide is reacted in an aqueous medium with soluble lead or zinc salts or with a slurry of lead oxide or zinc oxide [522], [524], [528], [531-533]. The pigments are filtered, washed, dried, and ground. Zinc cyanamide [534] and pure lead cyanamide are not explosive. An explosion reported during the production of lead cyanamide was caused by contamination with small amounts of acid or nitrates [535].

Zinc cyanamide is nontoxic, but the toxicological classification of lead cyanamide has to take its lead content into account. For toxicological data, see Section Toxicology.

Trade names include Bleicyanamid LY 80, Zinkcyanamid (BASF, Germany).

4.2.9. *Ion-Exchange Pigments*

Ion-exchange pigments were developed as nontoxic alternatives to the chromate pigments. They consist of a silicate carrier (zeolite [536] or amorphous silica gel [537]) to which calcium ions are bound. Commercial ion-exchange pigments have the following properties [538-540]:

pH	ca. 9
Conductivity, µS	ca. 300
Ca content, wt %	ca. 5
Water-soluble fraction, %	ca. 0.9
Oil absorption value, g/100 g	40 – 60
Density, g/cm^3	1.8
Mean particle size, µm	3 – 5
Color	white
Apparent density, g/L	500 – 600

Ion-exchange pigments should be classified as active anticorrosive pigments. They act by exchanging their calcium ions for hydrogen ions in the paint film. In this way acidic chemical substances are neutralized. The calcium ions then bind to the metal oxide surface. The pigments have good anticorrosive properties owing to their high pH value, but the danger of pinholing (depending on the binder) should be pointed out.

A *trade name* for ion-exchange pigments is Shieldex AC3, AC5 (W. R. Grace, UK).

4.2.10. *Metal Oxide Pigments*

4.2.10.1. *Red Lead*

Red lead [1314-41-6], Pb_3O_4, M_r 685.57, crystallizes in the tetragonal system and is a red powder with a density of 9.1 g/cm^3. It decomposes at ca. 500 °C at atmospheric pressure. For standard specifications, see Table 1.

Red lead should be regarded as the lead salt of orthoplumbic acid, H_4PbO_4, i.e., it is lead(II) orthoplumbate, Pb_2PbO_4, in which PbO_6 octahedra are linked by Pb(II) ions [541].

Red lead is produced industrially by oxidizing lead monoxide (PbO) at ca. 460 – 480 °C with agitation in a stream of air for 15 – 24 h. Most red lead is used in the glass, ceramic, and accumulator industries where an apparent density of < 2 g/mL is adequate. For the paint industry, however, highly dispersed red lead is normally necessary, with a sieve residue of < 0.1 % on a 0.063 mm sieve (ISO 787, Part 18) and an apparent density of 1.3 – 2.0 g/mL (ISO 510, DIN 55 516).

The electrochemical action of red lead results from the fact that lead has valencies of 2 and 4 in lead orthoplumbate: Pb(IV) compounds are reduced to Pb(II) in the cathodic region [542]. The chemical anticorrosive effect is a result of lead soaps that are formed when fatty acids in the binder react with the red lead. The lead soaps permeate the paint film as lamellae, and give good mechanical strength, water resistance, and adhesion to the steel surface. Furthermore, the corrosion-promoting chloride and sulfate ions are precipitated by lead(II) ions [543].

Possible combinations of red lead with various binder systems are listed in Table 29 [509]. Red lead is still used for heavy-duty anticorrosion applications, especially for surfaces bearing residual traces of rust. In waterborne paints, red lead has no advantages over zinc phosphate [544].

A *trade name* for red lead is Bleimennige, hochdispers (Heubach & Lindgens, Germany).

4.2.10.2. *Calcium Plumbate*

Calcium plumbate [12013-69-3], Ca_2PbO_4, density 5.7 g/cm^3, is a beige powder formed from lead monoxide and calcium oxide at ca. 750 °C in a stream of air [545].

The anticorrosive properties of calcium plumbate are inferior to those of red lead [509]. Calcium hydroxide is formed as a hydrolysis product when water penetrates through a primer that contains calcium plumbate. The pH at the metal surface then increases to ca. 11 – 12 which inhibits corrosion.

The most important use is in primers for zinc- coated substrates. The pH change occurring on hydrolysis of the calcium plumbate etches the zinc surface which improves adhesion of primers, especially on hot-dip galvanized steel [546].

4.2.10.3. *Zinc and Calcium Ferrites*

Many paint formulations contain iron oxide as an extender. It is a physically protective anticorrosive pigment (only to a small extent). In order to obtain a chemically protective anticorrosive pigment with active constituents the iron oxide is heated with oxides or carbonates of alkaline earths (CaO, $CaCO_3$) or zinc (ZnO) to form pigments of the ferrite type [547], [548]. The following systems have been reported for alkyd resin primers: 2 CaO · Fe_2O_3, CaO · Fe_2O_3 [11119-52-9] and Zn(Mg)O · Fe_2O_3 [68187-51-9]. In the coating these pigments are hydrolyzed with water to form alkaline-earth hydroxides or zinc hydroxide which prevent corrosion by increasing the pH. Alkaline-earth soaps are also formed in certain binder media [549]. However, the pigment volume concentration must be high to ensure good results [550].

Only one zinc ferrite pigment has attained economic significance. Its properties are as follows [551]:

Water-soluble salts, wt %	max. 0.6
Oil absorption value, g/100 g	ca. 22
pH	9 – 11
Density, g/cm^3	5.0

A *trade name* for zinc ferrite is Anticor 70 (Bayer, Germany; Mobay Chemical Corporation, USA).

4.2.10.4. *Zinc Oxide*

Zinc oxide [1314-13-2], ZnO, M_r 81.37, is a white powder that is usually used in combination with active anticorrosive pigments. For a detailed description, see Section Zinc Oxide (Zinc White) . It has the following physical properties:

Oil absorption value, g/100 g	20 – 24
Density, g/cm^3	5.6
Apparent density, cm^3/100 g	100
Mean particle size, µm	0.11 – 0.22
BET surface area, m^2/g	3 – 10

The inhibiting action of zinc oxide is based on its ability to react with corrosive substances and to maintain an alkaline pH in the coating. It also reacts with acidic components of the binder to form soaps and absorbs UV light.

The lead content of commercial zinc oxide depends on the manufacturer and is in the range 0.002 – 1.5 %. For a zinc oxide coating to be considered lead-free, the lead content must be less than 1.5 %.

4.2.11. ***Powdered Metal Pigments***

4.2.11.1. *Zinc Dust*

Zinc dust [7440-66-6], Zn, M_r 65.37, *mp* 419.4 °C, density 7.14 g/cm^3, is a free-flowing blue-gray powder composed of spheroidal particles. It is produced by melting zinc in a crucible, vaporizing it at ca. 900 – 950 °C, and condensing and sifting the product. Alternatively, molten zinc is atomized with a nozzle to produce dust, which is then sifted (⟶ Zinc). Properties of commercial zinc dust pigments are listed in Table 38 [552], [553]. For standard specifications, see Table 1 ("Zinc dust pigments").

Table 38. Properties of commercial zinc dust pigments

Property	Zinc dust Ultra 25	Zinc dust Ultra 35
Total zinc content, wt %	>99.0	>99.0
Metallic zinc content [a], wt %	94 – 96	94 – 96
Lead content, wt %	< 0.005	< 0.005
Cadmium content, wt %	< 0.005	< 0.005
Iron content, wt %	0.003	0.003
Copper content, wt %	0.001	0.001
Acid-insoluble material, wt %	< 0.1	< 0.1
Sieve residue [b], wt %	< 0.01	< 0.01
Mean particle size [c], µm	2.8 – 3.2	3.3 – 3.8

[a] $KMnO_4$ method.
[b] Sieve residue on 16 900 sieve (DIN 4188, aperture: 45 µm).
[c] Air permeability method.

The action of zinc dust in primers with organic binders is based on sealing effects and electrochemical processes. The zinc reacts with water and atmospheric oxygen that diffuse into the binder, forming zinc hydroxide which is then neutralized by sulfuric acid (from SO_2 in the air) and hydrochloric acid (from Cl-containing substances in the air, e.g., NH_4Cl). This causes an increase in volume and decreases permeability. The corrosion products of zinc also have an anticorrosive action [554]. Cathodic protection takes place when the zinc and iron come into contact; the zinc content in the primer must be at least 94 – 96 % [555-558]. Zinc dust coatings are used in large quantities for structural steel, including underwater steel construction and shipbuilding. Zinc dust is also used in inorganic binder systems (alkali silicates or alkyl silicates) in the form of two- component systems [559] (⟶ Paints and Coatings – Alkyl Silicates).

A *trade name* is Zinc Dust Ultra 25 and 35 (Lindgens & Söhne, Germany).

4.2.11.2. *Lead Powder*

Lead powder [7439-92-1], Pb, M_r 207.19, *mp* 327.5 °C, density 11 g/cm^3, is a dark gray powder containing 99 % metallic lead and ca. 0.5 % lead(II) oxide. It is produced by spraying molten lead at ca. 0.5 MPa and sifting. Owing to its high surface area (particle size 1 – 15 µm), it is liable to oxidize, and is therefore supplied in airtight packages or as a paste.

Lead powder can be combined with many binders [560], [561]. It does not affect the stability or viscosity of the paint. Binders that absorb only small amounts of water are particularly suitable (e.g., epoxy resins, chlorinated rubber). When formulating paints based on lead powder, care must be taken not to dilute it with other pigments and extenders by more than 5 vol %.

Lead powder coatings are mainly used for protecting against aggressive chemicals. They have a high UV reflection and extremely good elasticity. Lead powder pigments and pastes are also used in radiological protection.

4.2.12. ***Flake Pigments***

Anticorrosive flake pigments are summarized in Table 39. They increase the barrier resistance of a coating towards water and aggressive gases by increasing the length of the diffusion path in the coating [465-467] (see Section Principles). The interaction between pigment and binder should be as water-resistant as possible to prevent the diffusion of water from the surface of the flakes through the coating.

Table 39. Anticorrosive flake pigments

Pigment	CAS registry no.	Formula	Density, g/cm^3	Color
Micaceous iron ore	[1317-60-8]	Fe_2O_3 (specular hematite)	4.6 – 4.8	dark gray
Talc	[14807-96-6]	$Mg_3Si_4O_{10}(OH)_2$	2.6 – 2.8	white, light green to brown
Muscovite	[114755-46-3]	$KAl_2Si_3AlO_{10}(OH,F)_2$	2.8	gray-white
Phlogopite	[114733-61-8]	$KMg_3AlSi_3O_{10}(OH,F)_2$	2.8 – 2.9	brown
Flake aluminum	[7429-90-5]	Al	2.7	silver
Flake zinc	[7440-66-6]	Zn	7.1	gray

In principle, all lamellar minerals may be used as barrier pigments, e.g., micaceous iron ore [562-564], layer silicates (mica), linear polymeric silicates (wollastonite), and talc [565]. However, untreated mica and talc are not very suitable because they are highly permeable to water [465]. The surface can be modified with, for example, silanes or titanates, to reduce water permeability and improve adhesion [566]. If the chemistry of the treated pigment surface is adapted to the functional groups of the binder, the pigment can be used in aqueous anticorrosive dispersions [567-569].

Flake aluminum pigments with varying platelet thicknesses and shapes are used for corrosion protection [570]. For standard specifications, see Table 1 ("Aluminum pastes and pigments"). They are coated with a water-repellent, fatty film and are therefore particularly suitable for conventional solventborne coating systems. They have outstandingly good weather resistance [570-572].

Flake zinc pigments have a barrier effect and also act by a cathodic anticorrosive mechanism. Compared with zinc dust coatings, flake zinc pigments are formulated with lower pigment volume concentrations [573].

4.2.13. *Organic Pigments*

Unlike soluble inhibitors, organic anticorrosive pigments are sparingly soluble organic compounds or metallic salts of organic acids and are used in binders instead of, or in addition to, inorganic anticorrosive pigments. Inhibitors are not viable substitutes for active anticorrosive pigments in coatings [574]. Organic anticorrossive pigments were developed to replace toxic, chromate-based anticorrosive pigments.

Trade names include Sicorin RZ (BASF, Germany); Alkophor 827 (Henkel, Germany); and Irgakor 252 (Ciba-Geigy, Switzerland).

The zinc salt of 5-nitroisophthalic acid [60580-61-2] [448] is produced as an easily dispersible pigment by a wet chemical process from 5-nitroisophthalic acid and zinc oxide [575]. It has the following properties:

Color	colorless
Density, g/cm^3	2.5
Oil absorption value, g/100 g	40
Bulkiness, cm^3/100 g	650
Specific surface area, m^2/g	16
Heat resistance, °C	ca. 300
Zinc content, wt %	ca. 44
Organic content, wt %	ca. 49
Chloride content, wt %	ca. 0.02
Sulfate content, wt %	ca. 0.03
Nitrate content, wt %	ca. 0.01
Volatile content, wt %	ca. 5
Water-soluble content, %	ca. 1.6
pH	7.2
Conductivity, µS	ca. 500

A less expensive replacement for zinc chromate is obtained by combining the zinc salt of 5-nitroisophthalic acid with zinc phosphate pigments (Section Multiphase Phosphate Pigments) [448], [473], [576]. A concentration of 0.5 – 2.0 % based on the liquid coating is recommended.

The electrochemical anticorrosive action of the zinc salt of 5-nitroisophthalic acid is comparable with that of zinc potassium chromate [577].

Table 29 gives possible combinations with binders. The compound is nontoxic (see Section Toxicology).

Other Organic Anticorrosive Pigments. The following organic anticorrosive pigments are either commercial products or have been reported in the literature:

1. (2-Benzothiazolylthio)succinic acid [578], [579]
2. Zinc mercaptobenzothiazole [580], [581]
3. Basic zinc salt of *N*-benzosulfonylanthranilic acid [448], [582]

They are used in the binder at concentrations of 0.5 – 2.0 %, if possible in combination with zinc phosphate or other suitable anticorrosive pigments. For toxicological data, see Section Toxicology.

4.2.14. *Toxicology*

The change in the toxicological classification of lead- and chromate-based anticorrosive pigments has resulted in reappraisal of the formulation of anticorrosive coatings.

Acute oral toxicities (LD_{50}) of anticorrosive pigments are given in Table 40. For a detailed description of the toxicology of lead, see ⟶ Lead – Toxicology and Occupational Health; for chromates, see ⟶ Chromium Compounds – Toxicology and Occupational Health.

Table 40. Acute oral toxicity of anticorrosive pigments

Pigment	LD_{50} (rat), mg/kg	Reference
Zinc phosphate	> 5000	[448]
Aluminum triphosphate	> 11 000	[454]
Multiphase phosphate pigment	> 5000	[448]
Iron phosphide	> 21 500	[494]
Barium metaborate	850	[459]
Zinc potassium chromate *	> 640	[448]
Zinc tetraoxychromate *	> 5000	[448]
Calcium, zinc carbonate/molybdate	> 5000	[456]
Zinc molybdate	11 800	[456]
Lead cyanamide	3 500	[448]
Zinc cyanamide	< 1 800	[448]
Red lead	> 15 000	[583]
Zinc oxide	> 10 000	[584]
Zinc salt of 5-nitroisophthalic acid	> 10 000	[448]
Basic zinc salt of *N*-benzosulfonylanthranilic acid	> 4 000	[448]
Zinc mercaptobenzothiazole	> 5 000	[580]
2-Benzothiazolylthiosuccinic acid	> 5 000	[579]

* Classified as a carcinogenic substance in Section III A1 of the list of MAK values.

4.3. Luster Pigments

As shown in Figure 47 A, *conventional pigments* interact with light by absorption and/or diffuse scattering (see Section Color Properties). Luster pigments comprise nacreous pigments (Section Nacreous and Interference Pigments) and metal effect pigments (Section Metal Effect Pigments).

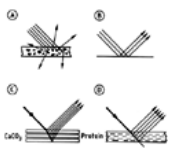

Figure 47. Optical principles of conventional and luster pigments

A) Conventional pigment that absorbs and scatters light; B) Metal effect pigment with complete regular reflection; C) Natural pearl composed of alternating layers of protein and $CaCO_3$; D) Nacreous pigment: the pearl is simulated by parallel orientation of the pigment platelets

Nacreous pigments simulate the nacreous luster of natural pearls (Fig. 47 C), which consist of alternating, transparent layers of materials with a high ($CaCO_3$) and a low (protein) refractive index. Small platelets of a high-refractive-index material (i.e., the nacreous pigment) are oriented in parallel alignment in a matrix of lower refractive index, e.g., a paint binder or plastic (Fig. 47 D). Synthetic nacreous pigments are transparent or light-absorbing platelet crystals. They may also have a multilayer structure, the layers having different refractive indices and light absorption properties.

Metal effect pigments (Fig. 47 B) consist of small metallic flakes (mainly aluminum or Cu/Zn bronze) which act as small mirrors that reflect almost all of the incident light.

4.3.1. *Nacreous and Interference Pigments*

Synthetic or natural pigments used to achieve lustrous, brilliant, or iridescent color effects by interference on thin optical films are called nacreous or pearlescent pigments [585-589]. They were originally used to simulate the appearance of natural pearls. The visual effects produced by reflection and transmittance of light by thin multilayer films are not restricted to pearls and shells but are widespread in nature [590]. Extending GREENSTEIN's definition [586], nacreous (pearlescent) pigments are "thin platelets of high refractive index which partially reflect and partially transmit light," or contain layers of this kind. DIN 55 943 (April 1990) proposes the comprehensive term "interference pigments" for nacreous pigments consisting of materials that do not absorb visible light (e.g., TiO_2) or that contain materials with a high refractive index and additional absorption characteristics (e.g., Fe_2O_3). However, in this article the term "nacreous pigments" is used.

The use of pearls and nacreous shells for decorative purposes goes back to ancient times (e.g., in Chinese wood intarsia). The history of pearl pigments dates back to 1656, when the French rosary maker JAQUIN isolated a silky lustrous suspension from fish scales (pearl essence) and applied this to small beads to create artificial pearls [591]. It took more than 250 years to isolate the pearl essence material (guanine platelets) and understand the pearl effect [592]. Attempts were made to create synthetic pearl colors as organic or inorganic, transparent, highly refractive coatings and pearl pigments as crystalline platelets [593]. From 1920 onwards hydroxides, halides, phosphates, carbonates, and arsenates of zinc, calcium, barium, mercury, bismuth, lead, and other cations were produced for this purpose. Only the traditional natural pearl essence (Section Natural Pearl Essence), basic lead carbonate (Section Basic Lead Carbonate), and bismuth oxychloride (Section Bismuth Oxychloride) are still of importance.

The strong demand for pearl effects came from the growing coatings and plastics industries who wanted to improve the acceptance and popularity of their products. Furthermore, nacreous pigments also allowed artists and designers to create new visual effects similar to those found in nature [594], [595]. The breakthrough for nacreous pigments came with the invention of mica coated with metal oxides (Section Metal Oxide – Mica Pigments) [596]. Mica-based nacreous pigments now account for > 85 % of the world market.

Important nacreous pigments and producers are listed below:

- Natural pearl
  - Engelhard Corp./Mearl, NJ, USA
- Basic lead carbonate
  - J. Jaeger GmbH, Viechtach, Germany
  - Polipearl S.A., Argentina
  - Semo Ltd., Seoul, South Korea
- Bismuth oxychloride
  - Engelhard Corp./Mearl , NJ, USA
  - Rona, N.J., USA (subsidiary of Merck KGaA)
- Metal oxide – mica
  - BASF, Ludwigshafen, Germany
  - Kemira OY, Pori, Finland
  - Engelhard Corp./Mearl , NJ, USA
  - Merck KGaA, Darmstadt, Germany (and overseas subsidiaries: E.M. Industries, NY, USA; Merck Japan Ltd., Onahama, Japan)

Nacreous pigments are used to obtain pearl, iridescent (rainbow), or metallic effects, and in transparent color formulations to obtain brilliance or two-tone color and luster flops (change with viewing angle). The most important applications are plastics, coatings, printing inks, cosmetics, and automotive paints. The two major nacreous pigment producers, Merck KGaA and Engelhard Corp./Mearl, do not publish production data.

Table 41 shows an overview of pigments with luster effects. Effect pigments can be classified as metal platelets, oxide-coated metal platelets, oxide-coated mica platelets, platelet-like monocrystals, and comminuted PVD films (PVD = physical vapor deposition). Aims of new developments are new effects and colors, improvement of hiding power, more intense interference colors, increased of light and weather stability, and improved dispersibility. Of special interest are pigments which are toxicologically safe and which can be produced by ecologically acceptable processes.

Table 41. Overview of effect pigments [597]

Pigment type	Examples
Metallic platelets	Al, Zn/Cu, Cu, Ni, Au, Ag, Fe (steel), C (graphite)
Oxide-coated metallic platelets	surface-oxidized Cu and Zn/Cu platelets, Fe_2O_3-coated Al platelets
Coated mica platelets	
nonabsorbing coating	TiO_2 (rutile), TiO_2 (anatase), ZrO_2, SnO_2, SiO_2
selectively absorbing coating	FeOOH, Fe_2O_3, Cr_2O_3, TiO_{2-x}, TiO_xN_y,$CrPO_4$, $KFe[Fe(CN)_6]$
totally absorbing coating	Fe_3O_4, TiO, TiN, $FeTiO_3$, C, Ag, Au, Fe, Mo, Cr, W
Platelet-like monocrystals	BiOCl, $Pb(OH)_2 \cdot 2\ PbCO_3$, α-$Fe_2O_3 \cdot n\ SiO_2$, $Al_xFe_{2-x}O_3$, $Mn_yFe_{2-y}O_3$, $Al_xMn_yFe_{2-x-y}O_3$, Fe_3O_4, reduced mixed phases, Cu phthalocyanine
Comminuted PVD films	Al, Al (semitransparent)/SiO_2/Al/SiO_2/Al (semitransparent)

4.3.1.1. *Optical Principles*

Nacreous and Interference Pigments. The optical principles of nacreous (interference) pigments are shown in Figure 48 for a simplified case of nearly normal incidence without multiple reflection and absorption. At the interface P_1 between two materials with refractive indices n_1 and n_2, part of the beam of light L_1 is reflected (L_1') and part is transmitted (i.e., refracted) (L_2). The intensity ratios depend on n_1 and n_2 (⟶ Optical Materials – Optical Properties of Optical Materials). In a multilayer arrangement as found in pearl or pearlescent and iridescent materials (Fig. 47 C) [587], each interface produces partial reflection. After penetration through several layers, depending on the size and difference between n_1 and n_2, virtually complete reflection is obtained, provided that the materials are sufficiently transparent [595].

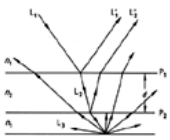

Figure 48. Simplified diagram showing nearly normal incidence of a beam of light (L_1) from an optical medium with refractive index n_1 through a thin solid film of thickness d with refractive index n_2

L'_1 and L'_2 are regular reflections from phase boundaries P_1 and P_2. L_3 represents diffuse scattered reflection from the transmitted light.

In pigments that simulate natural pearl effects, the simplest case is a platelet-shaped particle with two phase boundaries P_1 and P_2 at the upper and lower surfaces of the particles, i.e., a single, thin, transparent layer of a material with a higher refractive index than its surroundings. For small flakes with a thickness of ca. 100 nm, the physical laws of thin, solid, optical films apply [598].

Multiple reflection of light on a thin solid film with a high refractive index (Fig. 48) causes interference effects in the reflected light and in the complementary transmitted light. For the simple case of nearly perpendicular incidence, the intensity of the reflectance (I) depends on the refractive indices (n_1, n_2), the layer thickness (d), and the wavelength (λ) [599], [600]:

$$I = \frac{A^2 + B^2 + 2AB\cos\theta}{1 + A^2B^2 + 2AB\cos\theta}$$

$$A = \frac{n_1 - n_2}{n_2 + n_1} \quad B = \frac{n_2 - n_1}{n_2 + n_1} \quad \theta = 4\pi\frac{n_2 d}{\lambda}$$

With given n_1 and n_2 the maximum and minimum intensities of the reflected light — seen as interference colors — can be calculated and agree well with experimental results [601]. Refractive indices of materials commonly used in nacreous pigments follow:

Vacuum/air	1.0	$Pb(OH)_2 \cdot 2\,PbCO_3$	2.0
Water	1.33	BiOCl	2.15
Proteins	1.4	Carbon (diamond)	2.4
Plastics	1.4 – 1.7	Fe_3O_4	2.4
Mica	1.5	TiO_2 (anatase)	2.5
$CaCO_3$ (aragonite)	1.68	TiO_2 (rutile)	2.7
Natural pearl (guanine, hypoxanthine)	1.85	Fe_2O_3 (hematite)	2.9

In practice platelet crystals are synthesized with a layer thickness *d* calculated to produce the desired interference colors (iridescence) [601], [602]. Most nacreous pigments now consist of at least three layers of two materials with different refractive indices (Fig. 49). Thin flakes (thickness ca. 500 nm) of a material with a low refractive index (mica) are coated with a highly refractive metal oxide (e.g., TiO_2, layer thickness ca. 50 – 150 nm). This results in particles with four interfaces that constitute a more complicated, but still predictable thin film system. The behavior of more complex multilayer pigments containing additional, thin, light-absorbing films can also be calculated if appropriate optical parameters are known [601].

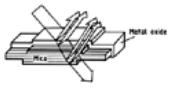

Figure 49. Simplified scheme of light reflection at the phase boundaries of a metal oxide – mica pigment

Color effects depend on the viewing angle [602-606]. Nacreous pigment platelets split white light into two complementary colors that depend on the platelet thickness. The reflected (interference) color dominates under regular (maximum) reflection, i.e., when the object is observed at the angle of regular reflection (Fig. 50 B). The transmitted part dominates at other viewing angles under diffuse viewing conditions provided there is a nonabsorbing (white) or reflecting background (Fig. 50 A). Variation of the viewing angle therefore produces a sharp gloss (reflectance) peak, and the color changes between two extreme complementary colors [602], [605]. The resulting complex interplay of luster and color is measured goniophotometrically in reflection and at different angles (⟶ Dyes, General Survey) [601]. No general standard measurement geometries have yet been specified. However, colorimetric analysis always includes measurements close to regular conditions and under diffuse conditions. This can be done, for example, by tilting a standard pigmented film on a drawdown card (Fig. 50). The colorimetric data are interpreted according to CIE $L^*a^*b^*$ data. A nacreous pigment is characterized by a minimum of three $L^*a^*b^*$ data sets measured under different conditions (e.g., 0°/45° black background, 22.5°/22.5° black background, 0°/45° white background). An analysis of these data specifies a pigment on the basis of its hiding power, luster, and hue [604]. The influence of other measurement geometries, especially of different commercially available instruments, is discussed in [601], [602].

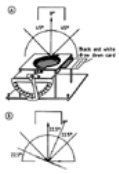

Figure 50. Colorimetric determination of nacreous pigments

The platelet particles are incorporated in a lacquer of defined thickness on a standard black and white drawdown card. The card is mounted on a rotatable sample holder and measured under (A) 0°/45° (diffuse reflection) and (B) 22.5°/22.5° (regular reflection, luster).

Against a black background or in a blend with carbon black, the transmitted light is absorbed and the reflected interference color is seen as the mass tone (i.e., overall color) of the material. In blends of nacreous pigments with absorbing colorants, the particle size of the latter must be well below the scattering limit, i.e., they must be transparent (see Section Transparent Pigments). The nacreous effect or iridescent reflection is otherwise quenched by the hiding pigments. This also applies to blends with strongly reflecting metal effect pigments (e.g., aluminum). Blends of pigments with different interference colors obey an additive mixing law, (e.g., blue + yellow = white) instead of the subtractive color mixing of pure absorption pigments (e.g., blue + yellow = green) [602].

The behavior of combinations of interference pigments with absorption and/or metal effect pigments is too complicated to predict and requires thorough goniophotometric analysis. The nacreous pigment effects generally dominate under regular (gloss) viewing conditions, and the absorbing colorant under diffuse viewing conditions [606].

4.3.1.2. *Natural Pearl Essence [589], [591], [606-611]*

Natural pearl essence (Essence d'Oriente, Fish Silver) — a pigment suspension derived from fish scales, skin, or bladder — is the oldest commercial pearl luster pigment. It consists mainly of a mixture of the purines guanine [73-40-5] (75 – 97 %) and hypoxanthine [68-94-0] (25 – 3 %). The ratio of these two purines depends on the fish species (mainly herrings, sardines, and other white fish) and their geographical origin (Japan, Norway, and the northeast U.S. – Canadian coastal border).

The pigments are formed in the fish scales as platelet-shaped crystals (0.05 µm × 1 – 10 µm × 20 – 50 µm). A commercial synthetic process for producing purines with this crystal shape has not been found. An aqueous suspension of fish scales is therefore extracted with organic solvents to dissolve and remove the proteins. The remaining dispersion contains purine crystals and scale, which are separated from one other by a complicated washing and phase-transfer process [611].

Due to its tendency to agglomerate in dry form, natural pearl essence is handled as a 22 – 25 % dispersion in various media (e.g., nitrocellulose lacquer for nail enamel, aqueous or alcoholic media for lotions or shampoos) It is used almost exclusively in cosmetic applications [611]. Despite the high price, natural pearl essence has several advantages over synthetic pearlescent pigments: it is less fragile; has a low density (1.6 g/cm^3), which reduces settling in liquid formulations; and has a very high, but soft luster (n = 1.79 – 1.91, the highest value being along the shortest axis). The world market for natural pearl essence in 1995 was estimated to be less than 50 t.

4.3.1.3. *Basic Lead Carbonate [585-589], [612], [613]*

The first commercially successful synthetic nacreous pigments were hexagonal platelet crystals of lead salts: thiosulfate, hydrogenphosphate, hydrogenarsenate, and most important nowadays, basic carbonate. Basic lead carbonate [1319-46-6] $Pb(OH_2) \cdot 2\,PbCO_3$, M_r 775.7, is precipitated from aqueous lead acetate or lead propionate solutions with carbon dioxide.

$$3\,Pb(OCOCH_3)_2 + 2\,CO_2 + 4\,H_2O \rightarrow 2\,Pb(OH)_2 \cdot 2\,PbCO_3 + 6\,CH_3COOH$$

Under appropriate reaction conditions regular hexagonal platelets (ca. 50 nm thick and 20 µm in diameter) can be obtained. The high refractive index (n = 2.0), high aspect ratio (> 200), and the extremely even surface of basic lead carbonate make it an optical match to natural pearl essence.

The platelet thickness can be adjusted to produce interference colors by modifying the reaction conditions. When aligned with its plane orthogonal to the incident light, the platelet crystal behaves as a thin, solid, optical film (see Fig. 48) with two phase-shifted reflections from the upper and lower crystal planes (the phase boundaries).

The crystals are mechanically sensitive, and their high density (6.4 g/cm^3) results in fast sedimentation. In view of their agglomeration tendency and occupational health (toxicity) risks they are not produced in powder form, but are flushed from the aqueous phase into suitable organic solvents or resins and handled as stabilized dispersions.

Currently, use of basic lead carbonate is limited to artificial pearls, buttons, and bijouterie. Due to the low chemical stability of this pigment and toxicity problems, it is being increasingly replaced by bismuth oxychloride and mica-based pigments. Worldwide production of basic lead carbonate pigment in 1995 was ca. 1000 t.

4.3.1.4. *Bismuth Oxychloride [585-589], [614]*

Powders containing bismuth compounds have long been used for decorative purposes to generate a shiny luster or lustrous colors (e.g., facial cosmetic powders in ancient

Egypt, imitation pearls made by coating glass and ceramic beads). Bismuth oxychloride [7787-59-9], bismuth oxide chloride, BiOCl, M_r 260.4, was the first synthetic nontoxic nacreous pigment. It is produced by hydrolysis of acidic bismuth solutions in the presence of chloride ions. Precipitation conditions may be varied (concentration, temperature, pH, pressure) or surfactants added to obtain the desired crystal quality. The virtually tetragonal bipyramidal structure is thereby “squashed” into a flat platelet.

Pure BiOCl is available in three grades with different nacreous effects that depend on the aspect ratios and crystal size:

1. Low- or medium-luster powder (aspect ratio 1 : 10 to 1 : 15), mainly used as a highly compressible, white, lustrous filler with excellent skin feel
2. Dispersion of high luster quality (aspect ratio 1 : 20 to 1 : 40) consisting of square or octagonal platelets in nitrocellulose lacquers (nail polish) or castor oil (lipsticks)
3. Dispersion of very high luster quality (aspect ratio > 1 : 50) consisting of lens-shaped platelets in nitrocellulose lacquer, castor oil, or butyl acetate

Pigments consisting of BiOCl- coated mica or talc and blends of BiOCl with other organic or inorganic colorants are also available.

The dominant market for BiOCl is still the cosmetic industry. Its low light stability (it turns from silver white to metallic gray in sunlight), relatively high price, fast settling (high density, 7.73 g/cm^3), and mechanical sensitivity limit its use in technical applications. Although the darkening reaction is not yet understood, low-luster grades with improved light stability are available. Some manufacturers promote the combination with UV stabilizers for technical purposes. Uses are in the button industry, bijouterie, printing, and for X-ray contrast in catheters. The world market is ca. 500 t/a.

4.3.1.5. *Metal Oxide – Mica Pigments*

The dominant class of nacreous pigments is based on platelets of natural mica coated with thin films of transparent metal oxides (see Fig. 49). Mica minerals are sheet layer silicates and are described in detail elsewhere (⟶ Mica). Nacreous pigments are usually based on transparent muscovite [99401-63-5] but some are based on natural or synthetic phlogopite [110710-26-4]. Although muscovite occurs worldwide, few deposits are suitable for pigments; it is biologically inert and approved for use as a filler and colorant [615], [616].

Selection and pre-processing of the mica substrate is one of the key factors which determine the quality and appearance of nacreous pigments. The aspect ratio of the final pigment depends on the particle size distribution of the mica platelets, which have a thickness of 300 – 600 nm and various diameter ranges (e.g., 5 – 25, 10 – 50, 30 – 110 µm). Since light is regularly reflected from the planes of the metal-oxide- coated mica and scattered from the edges, brilliance and hiding power are inversely related to each other.

A mica pigment coated with a metal oxide has three layers with different refractive indices and four phase boundaries $P_1 – P_4$: (P_1) TiO_2 (P_2) mica (P_3) TiO_2 (P_4). Interference of light is generated by reflections of all six combinations of phase boundaries, some of which are equal: $P_1P_2 = P_3P_4$, $P_1P_3 = P_2P_4$, P_1P_4, and P_2P_3. The thickness of the mica platelets varies in accordance with a statistical distribution. Consequently, interference effects involving the phase boundaries between the mica substrate and the oxide coating add together to give a white background reflectance. The interference color of a large number of particles therefore depends only on the thickness of the upper and lower metal oxide coating layers [586], [599], [600].

The development of the mica-based pigments started with pearlescent colors (Fig. 51 A, TiO_2 – mica). It was followed by brilliant, mass-tone- colored combination pigments (i.e., mica, TiO_2, and another metal oxide) with one color (interference color same as mass tone) or two colors (interference and mass tone different) that depend on composition and viewing angle (Fig. 51 B). In the 1980s further development was made by coating mica particles with transparent layers of iron(III) oxide (Fig. 51 C) [617].

Figure 51. Upper half of metal oxide – mica pigments

Increasing layer thickness of metal oxide causes different interference colors in reflection. Combination with absorption colorants (e.g., Fe_2O_3) produces metallic effects.

A) Interference colors; B) Combination pigments; C) Metallic colors

Titanium Dioxide – Mica. The first multilayer pigments were marketed in the 1960s as TiO_2- coated muscovite micas [588], [596]. Two different processes are used for coating mica in aqueous suspension on a commercial basis:

1. Homogeneous hydrolysis (at 100 °C)

$$TiOSO_4 + \text{mica} + H_2O \rightarrow TiO_2 - \text{mica} + H_2SO_4$$

2. Titration

$$TiOCl_2 + 2\,NaOH + \text{mica} \rightarrow TiO_2 - \text{mica} + 2\,NaCl + H_2O$$

The pigments are then dried and calcined at 700 – 900 °C. The titration (chloride) process is preferred for interference pigments with thick TiO layers because it is easier to control.

Chemical vapor deposition in a fluidized bed has also been proposed [618] (> 100 °C):

$$TiCl_4 + 2\,H_2O + \text{mica} \rightarrow TiO_2 - \text{mica} + 4\,HCl$$

Only the TiO_2 anatase crystal modification is formed on the mica surface. Small amounts of SnO_2 are therefore used to catalyze conversion to the rutile structure with its higher refractive index, brilliance, color intensity, and superior weather and light resistance [587], [588], [589].

The sequence of interference colors obtained with increasing TiO_2 layer thickness agrees with theoretical calculations in the color space [601], [619], [620]. An experimental development of $L^*a^*b^*$ values is given in [587] (see also ⟶ Dyes, General Survey).

TiO_2 – mica pigments are used in all color formulations of conventional pigments where brilliance and luster are required in addition to color, i.e., in plastics, coatings, printing, and cosmetics. The major market for silver white pigments (pearl pigments, “white metallic”) is the plastics industry.

Table 42 contains a comparative overview of TiO_2 – mica, basic lead carbonate, bismuth oxychloride, and natural fish silver pigments. Some further physical data are summarized in Table 43.

Table 42. Properties and application of pearlescent pigments [621]

Pearlescent pigment	Advantages	Disadvantages	Main fields of application
Natural fish silver	very low density, high luster, nontoxic, light stable	high price, low hiding power, limited availability	nail lacquers
Basic lead carbonate	very high luster, good hiding power, low price, light stable	high density, limited chemical and thermal stability, toxic	buttons, bijouterie
Bismuth oxychloride	very high luster, good hiding power, nontoxic	limited light stability, high density	decorative cosmetics, buttons, bijouterie
Titanium dioxide – mica	high luster, good hiding power (dependent on particle size), highest chemical, mechanical, and thermal stability, nontoxic, low price, low density	inferior luster compared to top qualities of basic lead carbonate and bismuth oxychloride	plastics, lacquers, cosmetics, printing inks, ceramic products

Table 43. Technical data of pearlescent pigments

Pearlescent pigment	Shape	Particle size, µm	Thickness, nm	Density, g/cm^3
Natural fish silver	needles, longish platelets	10 – 40	40 – 50	1.6
Basic lead carbonate	hexagonal crystals	4 – 20	40 – 70	6.4
Bismuth oxychloride	flat tetragonal-bipyramidal crystals	5 – 30	100 – 700	7.7
Titanium dioxide – mica	platelets	1 – 200	200 – 500	3

Iron Oxide – Mica. Like titanium dioxide, iron(III) oxide is suitable for coating mica platelets. It combines a high refractive index (metallic luster) with good hiding power and excellent weather resistance.

Commercial Fe_2O_3 – mica pigments are produced by precipitation of iron(II) or iron(III) ions in aqueous mica suspensions and calcination of the resulting coated particles at 700 – 900 °C [585], [587-589]:

$$2\,FeCl_3 + \text{mica} + 3\,H_2O \rightarrow Fe_2O_3 - \text{mica} + 6\,HCl$$

It is also possible to produce iron oxide – mica pigments by a direct CVD fluidized bed process in which iron pentacarbonyl is oxidized and Fe_2O_3 is deposited on the mica surface (⟶ Iron Compounds – Carbonyl Iron Oxide.) [622].

Independent of the synthesis route, iron(III) oxide crystallizes in the α-modification (hematite) after calcination. Brilliant, intense colors are obtained with 50 – 150 nm layers of Fe_2O_3 (hematite) on muscovite (see Fig. 51 C). Absorption and interference colors are produced simultaneously and vary with layer thickness. The red shades are especially intense because interference and absorption enhance each other. An intense green – red flop with different viewing angles is possible at a Fe_2O_3 layer thickness producing green interference (Section Optical Principles) [623].

Combination Mica-Based Pigments. Simple blending of transparent absorption pigments with pearlescent pigments is only one way to attain new coloristic effects. It is possible to produce nacreous pigments coated with a layer of transparent absorption colorant to realize more pronounced brilliant colors with a sharper color flop. An additional advantage of such pigments is the elimination of dispersion problems associated with transparent absorption pigments due to their small particle size and high surface area.

One possibility for attractive combination pigments is the coating of TiO_2 – mica pigments with an additional layer of an inorganic or organic colorant. The thickness of the TiO_2 layer is decisive for the brilliance or interference effect under regular viewing conditions, whereas the transparent colorant dominates at all other viewing angles (Fig. 52). A deep, rich color with a luster flop at all angles is attained when the colorant and interference color are matched. If the interference color and the mass tone of the colorant are different, a color flop (two-tone pigments) is seen in addition to the luster flop.

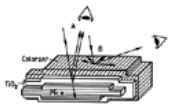

Figure 52. Combination pigment consisting of TiO_2 – mica coated with an additional layer of an absorption colorant

Under regular conditions (A), a brilliant color effect dominated by thin film reflection is visible. All other viewing angles (B) show the color of the transparent absorption colorant (Section Transparent Pigments).

Iron(III) oxide is the most important metal oxide for combination with titanium dioxide on mica flakes. Brilliant golden pigments result which can be applied for several purposes. Two routes are used to synthesize these pigments, and different structures are formed [585], [587-589]. In the first, a thin layer of Fe_2O_3 is coated onto the surface of a TiO_2 – mica pigment. The overall interference color is the result of both metal oxide layers. The mass tone is determined by the Fe_2O_3 layer, and interesting gold pigments (e.g., reddish gold) are possible. In the second case, co-precipitation of iron and titanium oxide hydroxides on mica particles followed by calcination leads to greenish gold pigments. Interference and mass tone can be explained as above. The mass tone in this case is, however, further modified due to the additional formation of the highly refractive yellowish iron titanate phase Fe_2TiO_5 (pseudo-brookite) [624].

Other inorganic colorants used instead of iron oxide for combination pigments are Cr_2O_3 (green), iron blue, cobalt blue, Fe_3O_4 (black), and carbon black. In the case of the two black colorants, the interference color is seen as the mass tone. There is an analogy to blends with black pigments in a color formulation, where the transmitted part of the light is absorbed.

Coating of TiO_2 – mica with an organic colorant for a mass-tone or two-tone pigment is performed by precipitation or deposition on the mica pigment surface in aqueous suspension, assisted by complexing agents or surfactants. Another method is to fix the colorant as a mechanically stable layer by using proprietary additives.

4.3.1.6. *New Developments*

Coloristic Variations on the Basis of Mica. Mica platelets can be coated with a variety of compounds to produce novel pigments. Solid-state reactions and CVD processes extend the possibilities for the synthesis of mica pigments. In addition, the calcination of the materials in the presence of inert (e.g., N_2, Ar) or reactive gases (e.g., NH_3, H_2, hydrocarbons) allows the formation of phases which are not producible by working in air. Table 44 contains a summary of nacreous mica pigments with special coloristic properties.

Table 44. Examples for mica-based pearlescent pigments with special coloristic properties

Pigment composition	Preparation	Remarks	Refs.
TiO_{2-x} – TiO_2 – mica	TiO_2 – mica + H_2 (Ti, Si), T > 900 °C (solid-state reaction)	gray, blue-gray	[625]
TiO_xN_y – TiO_2 – mica	TiO_2 – mica + NH_3, T > 900 °C (solid-state reaction)	gray, blue-gray	[622]
$FeTiO_3$ – TiO_2 – mica	Fe_2O_3 – mica + H_2 ,T > 600 °C (solid-state reaction)	gray (ilmenite pigments)	[626]
Fe_3O_4 – mica	Fe_2O_3 – mica + H_2 $T \approx 400$ °C (solid-state reaction)	black	[627]
	Fe^{2+} + O_2 + mica (precipitation)	black	[627]
	$[Fe(CO)_5]$ + O_2 + mica (CVD process)	black	[622]
TiN – mica	$TiCl_4$ + NH_3 + mica (CVD process)	gold	[597]
TiO_2 – C – mica	$TiOCl_2$ + C + mica (precipitation), calcination under N_2	silver-gray, interference colors (carbon inclusion pigments)	[628]
$BaSO_4$ – TiO_2 – mica	Ba^{2+} + SO_4^{2-} + $TiOCl_2$ + mica (precipitation)	low-luster pigments	[629]
Fe_3O_4 – mica (mica surface only partially coated)	Fe^{2+} + O_2 + mica (precipitation)	transparent colors	[604]

Reduction of suitable aqueous metal salts by electroless plating yields metal-coated mica pigments. They are less expensive than noble metal flakes (Au, Ag) and their brilliant metallic appearance is comparable with that of metallic flakes [587], [588].

Functional Pigments Initially, metal oxide – mica pigments were developed purely for their excellent coloristic properties. Since then, they have also become of interest for functional uses. In coatings with a high content of platelet fillers, an advantageous overlapping roof-tile arrangement is possible that provides close interparticle contact,

barrier effects, and dense covering. The composition and thickness of the oxide layer on the mica surface are always responsible for the physical properties like electrical conductivity, magnetism, IR reflectivity, and laser markability. Table 45 lists data on some functional metal oxide – mica pigments.

Table 45. Functional metal oxide – mica pigments

Pigment composition	Property	Application	Refs.
$(Sn,Sb)O_2$ – mica, $Sn(O,F)_2$ – mica	electrically conductive	conductive flooring, antistatic packaging materials, light-colored primed plastic surfaces which can be electrostatically painted in further coating process, light-colored conducting surfaces for dust reduction under clean-room conditions	[630]
Fe_3O_4 – mica	magnetic	magnetic surfaces	[589], [627]
TiO_2 – mica	IR-reflective	IR-reflecting plastic sheets, e. g., for domed and continuous skylights	[631]
TiO_2 – mica, $(Sn,Sb)O_2$ – mica	laser sensitive	laser marking of plastics, coatings	[628]

New Developments Based on Non-Micaceous Systems The class of single crystal lustrous pigments is not limited to the nonabsorbing types like bismuth oxychloride and basic lead carbonate. Recent developments are absorbing pigments such as platelet-like graphite, laminar phthalocyanines and flaky iron oxides. These flakes consist of pure iron oxide or mixed phase pigments, e. g., $Al_xFe_{2-x}O_3$, $Mn_yFe_{2-y}O_3$, or $Al_xMn_yFe_{2-x-y}O_3$. Hexagonal platelet crystals with a diameter of 5 – 50 µm are grown under hydrothermal or flux conditions (melting) [632], [633]. They can be reduced to the corresponding isomorphous magnetites or used as substrates for additional coating with metal oxides [634].

The hematite platelets show a predominantly metallic effect. Very thin particles with a thickness of 50 – 400 nm display a pale copper gloss, which is indicative of interference. The shade can be varied and the properties of the platelets can be controlled by doping. Al or Mn is incorporated by substitution of Fe in the hematite lattice, and Si is incorporated interstitially [635]. Laminar iron oxide pigments are interesting because of their excellent fastness to light and outdoor exposure and their good mechanical stability. Their main applications up to now are in automotive paints and cosmetics.

Novel developments in nacreous and special effect pigments include the group of comminuted PVD thin films This group includes ground single-layer or multilayer film particles. The production of the film is the first step of the synthesis. In the second step, the film is ground to small platelets. Examples are pigments derived from amorphous or nanocrystalline aluminum or pigments consisting of a sequence of layers. These layers act as transmission filters towards each side; a typical sequence is as follows: a specular metal layer (e. g., Al, Cr), a dielectric layer with low refractive index (e. g. MgF_2), and a semi-transparent metal layer. Such compositions are known as Fabry – Perot structures [636]. The optical effect of these pigments is characterized by a very high gloss with brilliant colors. The pigments show a strong color flop when viewed from different angles. PVD film pigment can be used in the graphics industry, particularly as thin-film security device pigments in security printing [637].

Several development programs are concentrated on the search for substrate- containing systems in which mica is replaced by other platelet-like materials. For example, it is possible to replace mica by kaolin or talc and to produce bright conductive pigments by coating with $(Sn,Sb)O_2$ [638].

New pigments based on transparent silica flakes show extremely strong optical effects different to those of mica pigments [639]. Angle-dependent colors and other effects, achieved by the combination of these SiO_2 flakes with thin titania and/or iron oxide coating layers, have led to a new generation of pearl-luster pigments. They add a new dimension to the existing possibilities of color styling with luster pigments.

Other platelet crystals with a high aspect ratio have been reported for phthalocyanine [640], graphite [641], and 1,4-diketopyrrolopyrroles [642].

The CVD process in a fluidized bed (see Section Metal Oxide – Mica Pigments) is also used for depositing interference Fe_2O_3 layers on aluminum flakes [622] as well as for the formation of titanium nitride on platelet substrates [643] or TiO_2 [644], [645].

Regular scattering of spherical particles (opalescence) of TiO_2 is also now used to generate eye-catching color effects [646], [647].

4.3.1.7. *Uses*

Nacreous and interference pigments are used as colorants or part of color formulations for all applications where traditional pigments are used, but where additional color depth, brilliance, iridescence, color shift (flop), and other spectacular effects are desired [648]. Mica-based pigments dominate; their combination of pearl and interference effects, brilliance, stability, and weather resistance is unsurpassed. Furthermore, they are nontoxic [616].

Nacreous pigments require transparent or at least translucent binders or other carriers. Formulations with other pigments have to take their transparency and color mixing rules into account. Producers specify certain product lines for specific applications on the basis of national regulations and technical considerations. They also provide handling guidelines and base formulas.

Color systems with nacreous pigments can be formulated in four ways [606]:

1. Blending nacreous pigments with other pigments in a single-color formulation
2. Using a two-layer coating system in which the upper coat contains the nacreous pigment and the lower coat contains the hiding absorption or metal effect pigments
3. Using a two-layer system with a transparent absorption color formulation applied on top of a layer containing the nacreous pigment
4. Using a multilayer nacreous pigment consisting of an absorbing and/or iridescent coating applied in a single coat to the substrate in the form of platelets (see Section Metal Oxide – Mica Pigments)

Special care must be taken to ensure parallel alignment of the platelet pigments during application. In plastics formed by injection molding, parallel alignment is hindered by the high viscosity of the polymer melts. Flow lines can be avoided by appropriate die/mold design and surface patterning. In coatings [649], [650] and printing inks [651], proper formulation, rheology control, settling and shrinking of the films on drying result in parallel, stacked alignment of the platelets. In automotive coatings [652] additional surface modification of the pigments is used to increase long-term weather resistance [653-655].

For ceramic applications the pigment is coated with an additional SnO_2 layer to stabilize it against the aggressiveness of the frits at high temperature [656].

Cosmetic applications require specific nacreous and interference pigments that are approved for use by national regulations [657], [658]. Nacreous pigments are not used in cosmetics solely for their optical effects. TiO_2 – mica can be used as a sunscreen because TiO_2 strongly absorbs light in the near UV region [659]. A large variety of surface modifications and additives (e.g., monodisperse, submicron spheres of SiO_2) are used to improve processability, control oil absorption [660], and give a softer skin feel [661].

4.3.2. *Metal Effect Pigments*

Definition. Metal effect pigments consist of flakes or lamellae of aluminum (aluminum bronzes), copper and copper – zinc alloys (“gold bronzes”), zinc, and other metals (DIN 55 943, 55 944; FSO 4618–1).

Physical Principles. The metallic effect is caused by the reflection of light at the pigment surface (see Fig. 47); luster (remission) decreases when the proportion of light scattered by the edges and corners of the pigment particles increases. As the pigment becomes coarser, the proportion of reflected light becomes larger; brilliance, brightness and, in the case of “nonleafing” aluminum pigments, the flop (color change with viewing angle) therefore increase. With finer pigments, the proportion of scattered light is greater; brilliance, brightness, and flop are therefore lower but the hiding power is better.

The metal effect depends not only on particle size and particle size distribution, but also on the orientation of the metal flakes within the coating film, particle shape, the transparency of the binder matrix, and the presence of colored pigments or dyes.

Particle Size Distribution. The particle size distribution is determined by conventional sieving or other methods such as laser granulometry (see Section Fundamental

Aspects : Particle Size).

The required particle size of the pigments depends on the intended use and can vary from a few micrometers (e.g., in offset printing inks), to medium grades of < 25 µm or < 45 µm (e.g., in "metallic" automotive topcoats, gravure and flexographic printing inks), to coarser grades (e.g., in corrosion-inhibiting systems, colorants for plastics), and even up to "spangle" size of several millimeters.

The thickness of the flakes can vary from a hundredth of a micrometer up to approximately one micrometer.

The term shape factor denotes the ratio of the platelet thickness to the platelet diameter.

Production. Metallic pigments were first produced in the Middle Ages when gold leaf was made by hand beating. This was replaced by stamping machines. Metal effect pigments are now usually produced in ball mills, using dry milling (Hametag process) and wet milling (Hall process). Other special production processes are used on a small scale (e.g., production of aluminum foil by vaporization, followed by size reduction).

During the ball milling process, a lubricant is added to prevent cold fusion; it also determines the wetting properties of the metallic pigment, i.e., "leafing" or "nonleafing".

Leafing and Nonleafing.
Leafing pigments float on the surface of paint or printing ink films as a result of high interfacial tension. They form a coherent surface film whose reflective properties depend on the particle fineness. Stearic acid is the lubricant usually used in the manufacture of these products.

Nonleafing pigments are fully wetted by the medium and are distributed uniformly throughout the coating. Colored metallic effects are produced by combining the pigments with transparent pigments or dyes. The nonleafing properties are obtained by using lubricants that consist of branched- chain or unsaturated fatty acids (e.g., oleic acid) or polar substances (e.g., fatty amines).

Commercial Products. Depending on the production method and area of application, the pigments are supplied commercially in powder form or as solvent- containing pastes. Pastes for the paint industry are normally made with white spirit or a mixture of white spirit and solvent naphtha. Special pastes containing mineral oils are produced for offset printing inks, polar solvents are used for gravure and flexographic inks, and plasticizers are used to make pastes for coloring plastics. Stabilized aluminum pastes with water or water-miscible solvents are available for waterborne coating systems [662]. Pigments specially coated with organic (e.g., acrylics) or inorganic materials (e.g., silica) are available for powder coatings.

Uses. Metallic pigments are used to produce metallic or chrome effects and also as functional pigments (e.g., for corrosion protection, as conductive paints, or for protection against electromagnetic radiation).

They are employed in paints and coatings; graphics and printing inks (offset, gravure, flexographic and screen printing, and bronzing); and in plastics (coloration by, e.g., masterbatches, fillers, gel coating techniques, and other coatings).

Other industrial uses of aluminum flake include the building industry (production of aerated concrete) and the chemical industry (e.g., production of titanium dioxide, pyrotechnics, and explosives). Copper powder in flake form is used in the chemical industry (e.g., for phthalocyanine production, for lubricants).

Aluminum pigments are produced by grinding aluminum powder, usually obtained by atomizing molten aluminum. The starting material is mainly aluminum ingots with a purity of 99.5 % (DIN 1712), or pure aluminum (> 99.95 %) for special outdoor applications (acid-resistant grades). For standard specifications, see Table 1 ("Aluminum pigments and pastes").

Aluminum powder forms explosive mixtures with air; the lower explosive limit being ca. 35 – 50 g/cm^3. Dust-free aluminum paste is therefore used in most applications.

Aluminum pigments are mainly produced by wet milling in white spirit. The resulting pigment suspension is usually fractionated, sieved, and filtered on a filter press. The filter cake is mixed with solvents to give the usual commercial consistency with 65 wt % solids. Other solvents may be used instead of white spirit, the choice being governed by the intended application. The white spirit can be removed by vacuum drying and replaced by another solvent if necessary (e.g., for printing inks or coloring plastics).

Aluminum behaves as an amphoteric metal, liberating hydrogen from both alkaline and acid aqueous media. In waterborne paints and coloring systems, special stabilized aluminum pigments are therefore required [662]. Aluminum pigments must not be used with halogen- containing solvents because they react very violently to release hydrogen halides (Friedel – Crafts reaction).

The leafing properties of aluminum pigments can be adversely affected by polar solvents or binders. "Leafing-stabilized" aluminum pigments are available for these applications.

Along with special pigments for modern environmentally-friendly coating systems (e.g., waterborne paints and powder coatings) described above, aluminum pigments coated with colored metal oxides (e.g., iron oxide) have appeared on the market which produce not only reflection but also interference effects [663]. Novel color effects are obtained by combining the metal-oxide- coated pigments with transparent colored pigments.

Aluminum has a very low toxicity (⟶ Aluminum – Toxicology and Occupational Health) and is permitted as a coloring agent (EEC no. E 173, C.I. 77 000). It is also approved by the FDA, § 175 300 (USA).

Copper and Gold Bronze Pigments. Copper and gold bronze pigments (powdered copper – zinc alloys) are usually produced by dry milling. Depending on the alloy composition, the following "natural shades" are produced:

Copper	100 % copper
Pale gold	ca. 90 % copper and 10 % zinc
Rich pale gold	ca. 85 % copper and 15 % zinc
Rich gold	ca. 75 % copper and 25 % zinc

Copper – zinc alloys with a higher zinc content (brass) cannot be ground or formed into flake pigments on account of their brittleness. Controlled oxidation of "natural" bronze powders converts them into "fired" bronze powders. These shades (e.g., English green, lemon, ducat gold, fire red) are produced as a result of interference effects that depend on the thickness of the oxide coating.

Copper and gold bronze pigments are not as colorfast as aluminum pigments because they decompose to produce colored oxides and corrosion products. However, stabilized pigments (e.g., with a silica coating) are also available for critical applications in binders with high acid values or that react with copper or zinc.

The classical fields of application for the gold bronze pigments are the graphics industry (bronzing); printing (offset, gravure, flexographic, and screen printing); and coloring plastics. They are also used in the paint industry for decorative finishes (e.g., dip coatings for candles).

Gold bronze pigments are nontoxic. In Germany they are permitted in food packaging materials provided that they do not lead to contamination of the food. They are also allowed for printing on cigarette paper and filter tips according to the tobacco regulations if the zinc content does not exceed 15 %. Similar regulations apply in the United States and other countries.

Flake Zinc Pigments. Flake zinc pigments are used mainly as high-quality anticorrosive pigments in powder or paste form. Owing to their platelet structure they have a considerably higher surface area than spherical zinc dust particles. They can therefore take up much more binder which produces a more flexible coating film than that obtained with zinc dust. Other advantages include a lower settling tendency, good remixing, and problem-free application onto the precoated surface.

The flake zinc pigments give a considerably brighter, better metallic effect than zinc dust. Their appearance can be further improved by combining them with aluminum pigments or by applying a topcoat based on aluminum pigment.

Another widespread application for flake zinc pigments is the coating of small articles with complex shapes (e.g., screws, steel springs, bolts, rivets) and for special anticorrosive paints (Section Zinc Dust).

4.4. Transparent Pigments

Pigments become transparent in binders when the difference between the refractive index of the pigment (which depends on wavelength) and that of the binder is low, and when the particle size of the pigment is in the range 2 – 15 nm. The decisive factor for transparency is, however, the thickness of the pigment particles in the direction in which the light travels.

In practice, pigment transparency is measured by determining the color difference between a sample of the pigmented system on a black background and the black background itself. The transparency measurement method and the definition of the transparency index are described in DIN 55 988.

In a pigment of given chemical composition, transparency can only be influenced by particle size (i.e., by the size distribution of the primary particles) and is thus dependent on the synthesis conditions. Owing to the small size of the primary particles, the pigments tend to agglomerate. Since agglomeration can only be remedied with great difficulty by mechanical grinding, attempts are made to suppress agglomeration by using additives (e.g., sodium salts of higher fatty acids or alkylamines) during synthesis prior to the separation of the pigment from the suspension. In spite of this, almost all transparent pigments are difficult to disperse. They all have a specific surface area (BET) of > 100 m^2/g.

4.4.1. *Transparent Iron Oxides*

Transparent yellow iron oxide has the α-FeO(OH) (goethite) structure; on heating it is converted into transparent red iron oxide with the α-Fe_2O_3 (hematite) structure. Differential thermogravimetric analysis shows a weight loss at 275 °C. Orange hues develop after brief thermal treatment of yellow iron oxide and can also be obtained by blending directly the yellow and red iron oxide powders.

Production.
Transparent yellow iron oxide [51274-00-1], C.I. Pigment Yellow 42:77 492, is obtained by the precipitation of iron(II) hydroxide or carbonate with alkali from iron(II) salt solutions and subsequent oxidation to FeO(OH). On an industrial scale, oxidation is usually carried out by introducing atmospheric oxygen into the reaction vessel. Important factors for good transparency are high dilution during precipitation, a temperature of < 25 °C during oxidation, and short oxidation times (< 5 h). Oxidation can be carried out under acidic or basic conditions [664], [665]. The best results are obtained by using 6-wt % iron(II) sulfate solutions and precipitation with a 10 wt % solution of sodium carbonate, whereby ca. 85 % of the amount required for complete precipitation is added. The starting material is usually crystalline $FeSO_4 \cdot 7\,H_2O$, obtained as a byproduct from the pickling of iron or in the production of titanium dioxide from ilmenite by the sulfate method (see Section Sulfate Method). In order to improve pigment properties, the suspension is matured for about a day before filtration. It is then filtered, dried, and carefully ground to a powder. The primary particles are needle-shaped and have an average length of 50 – 100 nm, a width of 10 – 20 nm, and a thickness of 2 – 5 nm.

Transparent red iron oxide, C.I. Pigment Red 101:77 491, is obtained by heating the yellow pigment (e.g., in a cylindrical rotary kiln) at 400 – 500 °C (Fig. 53).

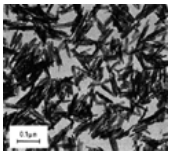

Figure 53. Electron micrograph of a transparent red iron oxide pigment (Sicotrans Red 2815)

Transparent red iron oxides containing iron oxide hydrate can also be produced directly by precipitation. A hematite content of > 85 % can be obtained when iron(II) hydroxide or iron(II) carbonate is precipitated from iron(II) salt solutions at ca. 30 °C and when oxidation is carried out to completion with aeration and seeding additives (e.g., chlorides of magnesium, calcium, or aluminum) [666]. Transparent iron oxides can also be synthesized by heating finely atomized liquid pentacarbonyl iron in the presence of excess air at 580 – 800 °C [667], [668]. The products have a primary particle size of ca. 10 nm, are X-ray amorphous, and have an isometric particle form. Hues ranging from red to orange can be obtained with this procedure, but it is not suitable for yellow hues.

Transparent brown iron oxides are produced by precipitating iron(II) salt solutions with dilute alkali (sodium hydroxide or sodium carbonate) and oxidizing with air. Only two-thirds of the precipitated iron hydroxide, oxide hydrate, or carbonate is oxidized. Alternatively, the iron oxides can be produced by complete oxidation of the precipitated iron compounds and subsequent addition of half the amount of the initial iron salt solution and precipitation as hydroxide [669], [670]. However, these products have the composition $FeO \cdot Fe_2O_3$ and are ferromagnetic; they are not of practical importance.

Properties and Uses. As far as resistance to light, weather, and chemicals is concerned, transparent iron oxides behave in a similar manner to the opaque iron oxides (see Section Iron Oxide Pigments). In addition, they show a high UV absorption, which is exploited in applications such as the coloring of plastic bottles and films used in the packaging of UV-sensitive foods [671], [672].

Worldwide consumption of transparent iron oxides is 2000 t/a. They are mainly used in the production of metallic paint in combination with flaky aluminum pigments and in the coloring of plastics for bottles and fibers.

Toxicology. Special toxicological studies on transparent iron oxides have not yet been carried out. The results of opaque iron oxides are applicable, see Section Toxicology and Environmental Aspects.

Trade names and producers include Capelle (Gebroeders Cappelle N.V., Belgium), Fastona Transparent Iron Oxide (Blythe Colours Ltd., UK), Sicotrans (BASF and BASF Lacke+Farben, Germany), and Trans Oxide (Hilton Davis, USA).

4.4.2. *Transparent Iron Blue*

Iron blue [14038-43-8], C.I. Pigment Blue 27:77 520 (Milori blue), also occurs in finely dispersed forms (primary particle size < 20 nm, specific surface area 100 m^2/g) that are more transparent than the conventional iron blue pigments. They are generally produced by the same procedure as the less transparent iron blue pigments (see Section Iron Blue Pigments) but a higher dilution factor is used. The transparency of these pigments is exploited solely in the production of printing inks (illustration gravure). All other properties are described in Section Iron Blue Pigments.

Trade names include Manox Iron Blue (Manox, UK), and Vossen-Blue 2000 (Degussa, Germany). Transparent iron blue pigments are also produced by Dainichiseika (Japan).

4.4.3. *Transparent Cobalt Blue and Green*

Transparent Cobalt Blue. Cobalt blue [1345-16-0] C.I. Pigment Blue 28 : 77 346, is also produced as a transparent pigment by precipitating cobalt and then aluminum as hydroxides or carbonates from salt solutions using alkali. The hydroxides or carbonates are then filtered, washed, dried, and calcined at 1000 °C [673]. It is important to carry out the precipitation with high dilution and to distribute the alkali uniformly throughout the entire reaction volume.

Transparent cobalt blue pigments have the form of small tiles (primary particle size 20 – 100 nm, thickness ca. 5 nm, specific surface area (BET) ca. 100 m^2/g). They are resistant to light and weather and are used in metallic paint, but are of little importance.

Transparent Cobalt Green. Analogous to cobalt blue, a transparent cobalt green spinel, C.I. Pigment Green 19:77 335 can be manufactured. The green and blue transparent cobalt pigments are used as filter materials for cathode ray tubes (CRTs) [674].

4.4.4. *Transparent Titanium Dioxide*

Titanium dioxide can also be prepared with a primary particle size of 10 – 30 nm, and hence shows transparent properties. Microcrystalline titanium dioxide was first mentioned in Japanese patents in 1978.

The use of micro titanium dioxide as a white pigment is limited because its light scattering is very low due to the very small crystal size, which means that the coloring effect of conventional TiO_2 is lost. The physical properties are changed significantly; there is strong UV absorption. Consequently, micro TiO_2 is used as a thermally stable, nontoxic, nonmigratory UV -absorbing additive:

- For cosmetics, mainly for sunscreen formulations because of its effective protection over the UV - A, UV -B, and UV -C spectrum [675]
- For automotive paints, especially in combination with aluminum flakes, giving a pearlescent-like appearance; the color flop depends on the concentration of micro TiO_2 in the formulation [677], [678]
- For clear coats and wood varnishes, where it protects the sub strate by virtue of its UV -absorbing properties
- For improving the UV durability of plastics films, and for protecting UV -sensitive foods in plastics wrapping [676]
- In silicon rubber as a heat stabilizer, as a catalyst for hydrogenation [679] and oxidation [680], and in surface protective films for furniture and optical materials [681]

Microfine TiO_2 is produced mainly in the rutile form by various manufacturing routes:

$$TiCl_4 + 4\,NaOH \rightarrow TiO_2 + 4\,NaCl + 2\,H_2O$$

$$Na_2TiO_3 + 2\,HCl \rightarrow TiO_2 + 2\,NaCl + H_2O$$

$$Ti(OC_3H_7)_4 + 2\,H_2O \rightarrow TiO_2 + 4\,C_3H_7OH$$

The process steps include precipitation, neutralization, filtration, and washing, drying, and micronization. Due its small particle size, transparent titanium dioxide has high photoactivity. To reduce this and to give better weatherfastness, the particles are coated with various combinations of inorganic oxides (e.g., silica, alumina, zirconia, iron oxide) before drying in a similar way to that used for conventional titanium dioxide.

Microfine TiO_2 with predominant anatase structure can be manufactured by flame hydrolysis of $TiCl_4$ at < 700 °C [682]:

$$TiCl_4 + 2\,H_2 + O_2 \rightarrow TiO_2 + 4\,HCl$$

World consumption, estimated at 1300 t/a, is increasing.

Trade names include Titanium Dioxide P 25 (Degussa, Germany), Hombitec RM Series (Sachtleben Chemie, Germany), and Micro Titanium Dioxide MT-Series (Tayca, Japan). Transparent titanium dioxide is also produced by Ishihara (Japan) and Kemira Oy (Finland).

4.4.5. *Transparent Zinc Oxide*

Similar to transparent titanium dioxide, Microparticles of zinc oxide are manufactured by sol – gel processes or precipitation in the presence of protective colloids to limit particle growth [682].

An industrial process operates with solutions of zinc sulfate and zinc chloride in the ratio 1 : 2 [684]. Basic zinc carbonate is precipitated by feeding simultaneously the zinc salt solution and a mixed solution of sodium hydroxide and sodium carbonate into a reactor charged with water. The precipitated product is washed intensively several times and then spray dried. It is used as a UV -shielding additive for cosmetics and paints.

Trade names include Sachtotech Micro-Zinkoxid (Sachtleben, Germany) and Z-cote HP 1 (SunSmart, United States).

[Top of Page]

References

1. *Colour Index International*: Society of Dyers and Colourists, 3rd ed., **vol. 5**, Bradford 1987.
2. *Kirk-Othmer*, 3rd ed., **17**, 788 – 838. H. G. Völz in H. Kittel (ed.): *Lehrbuch der Lacke und Beschichtungen*, **vol. II**, Verlag W. A. Colomb, Berlin 1974, pp. 27 – 47.
3. H. Rechmann, G. Sutter, *Farbe+Lack* **82** (1976) 793 – 796.
4. L. Gall, *Farbe+Lack* **81** (1975) 1015 – 1024.
5. E. Fritzsche, *Farbe+Lack* **80** (1974) 731 – 740.
6. H. G. Völz, *DEFAZET – Dtsch. Farben Z.* **28** (1974) 559 – 563.
7. F. Hund, *Farbe+Lack* **73** (1967) 111 – 120.
8. G. Kämpf, *Farbe+Lack* **71** (1965) 353 – 365.
9. B. Honigmann, J. Stabenow in: *Congr. FATIPEC VI,* 1962, Verlag Chemie, Weinheim 1963, pp. 89 – 92.
10. H. G. Völz, *Farbe Lack* **90** (1984) 642, 752.
11. F. Kindervater in: *Congr. FATIPEC XII,* 1974, Verlag Chemie, Weinheim 1974, pp. 399 – 403.
12. S. Keifer, A. Wingen, *Farbe+Lack* **79** (1973) 886 – 873.
13. D. B. Judd: *Color in Business, Science and Industry*, J. Wiley & Sons, New York 1969.
14. H. G. Völz, *Progr. Org. Coat.* **1** (1973) 1.
15. H. G. Völz, *Angew. Chem., Int. Ed. Eng.* **14** (1975) 655.
16. H. G. Völz: *Industrial Color Testing*, 2. ed., Wiley-VCH, Weinheim, Germany 2001.
17. A. Berger, A. Brockes: *Bayer-Farben-Revue*, special issue 3/1, Leverkusen 1971.
18. F. W. Billmeyer, Jr., in: *Kirk-Othmer*, 3rd ed., **6**, p. 523 ffx.
19. M. Richter: *Einführung in die Farbmetrik*, Walter de Gruyter, Berlin-New York 1976.
20. *CIE Technical Report* **116** (1995).
21. W. S. Stiles, G. Wyszecki: *Color Science* 2nd ed., Wiley, New York 1982, Section 8.4.1, Eq. (1).
22. P. Kubelka, F. Munk, *Z. Tech. Phys.* **12** (1931) 539 – 601.
23. P. Kubelka, *J. Opt. Soc. Am.* **38** (1948) 448 – 457.
24. L. Gall, *Farbe+Lack* **80** (1974) 297 – 306.
25. K. Hoffmann, *Farbe+Lack* **80** (1974) 118 – 125, 221 – 229.
26. F. W. Billmeyer, Jr., R. L. Abrams, D. G. Phillips, *J. Paint Technol.* **45** (1973) 23 – 30, 31 – 38; *J. Paint Technol.* **46** (1974) 36 – 39; *J. Paint Technol.* **48** (1976) 30 – 36.
27. H. G. Völz, *Ber. Bunsen-Ges. Phys. Chem.* **71** (1967) 326 – 339.
28. G. Mie, *Ann. Phys.* **25** (1908) 377 – 444. Links
29. H. C. van de Hulst: *Light Scattering by Small Particles*, J. Wiley & Sons, New York 1957.
30. W. Jaenicke, *Z. Elektrochem.* **60** (1956) 163 – 174.
31. H. G. Völz, *DEFAZET – Dtsch. Farben Z.* **31** (1977) 469 – 479.
32. A. Brockes, *Optik (Stuttgart)* **21** (1964) 550 – 566.
33. R. Brossmann in: *Internationale Farbtagung "Color 69",* vol. II, Verlag Musterschmidt, Göttingen 1969, pp. 890 – 900.
34. H. G. Völz, *Farbe+Lack* **76** (1970) 9 – 14. S. Keifer, *Farbe+Lack* **81** (1975) 289 – 294.
35. H. G. Völz, *Farbe+Lack* **88** (1982) 264, 443.
36. W. Schumacher in: *Congr. FATIPEC XIII.* 1976, EREC, Paris 1977, pp. 582 – 587.
37. F. Finus, *DEFAZET – Dtsch. Farben Z.* **28** (1974) 494 – 500.
38. S. Keifer, H. G. Völz, *Farbe+Lack* **83** (1977) 180 – 185, 278 – 285.
39. H. G. Völz, *Progr. Org. Coat.* **15** (1987) 99.
40. L. Gall in: *Congr. FATIPEC IX,* 1968, Chimie des Peintures, Brussels 1969, pp. 34 – 39.
41. S. Keifer, *Farbe+Lack* **82** (1976) 811 – 814.
42. F. Vial in: *Congr. FATIPEC XII.* 1974, Verlag Chemie, Weinheim 1975, pp. 159 – 167.
43. L. Gall, *Farbe+Lack* **72** (1966) 955 – 965.
44. L. Gall, *Farbe+Lack* **72** (1966) 1073 – 1079.
45. H. G. Völz, *DEFAZET – Dtsch. Farben Z.* **30** (1976) 392 – 396, 454 – 457.
46. S. Keifer: "Deckvermögen," in: *Anorganische Pigmente für Anstrich- und Beschichtungsstoffe,* Bayer, Leverkusen 1973, Section 3.3.2.
47. H. G. Völz in H. Kittel (ed.): *Lehrbuch der Lacke und Beschichtungen*, **vol. II**, Verlag W. A. Colomb, Berlin 1974, pp. 22 – 27.
48. K. Brugger, N. Jehle in: *Congr. FATIPEC XIII.* 1976, EREC, Paris 1977, pp. 15 – 23.
49. R. Epple, A. Englisch in: *Congr. FATIPEC XII.* 1974, Verlag Chemie, Weinheim 1975, pp. 241 – 248.
50. G. Kämpf, W. Papenroth, *Farbe+Lack* **83** (1977) 18 – 29.
51. W. Grassmann, *Farbe+Lack* **66** (1960) 67 – 74.
52. H. G. Völz, G. Kämpf, H. G. Fitzky, A. Klaeren, 179th National Meeting, American Chemical Society, Houston, Texas, *ACS Symp. Ser.* **151** (1981) 163.
53. H. G. Völz, G. Kämpf, H. G. Fitzky, *Farbe+Lack* **78** (1972) 1037 – 1049.
54. Th. Bohmann in: *Congr. FATIPEC X* 1970, Verlag Chemie, Weinheim 1971, pp. 123 – 128.
55. W. Papenroth, P. Koxholt, *Farbe+Lack* **82** (1976) 1011 – 1016.
56. H. J. Freier, *Farbe+Lack* **72** (1967) 1127 – 1136.

57. H. Rechmann, *et al.*, in: *Congr. FATIPEC X*; 1970, Verlag Chemie, Weinheim 1971, pp. 3 – 15.
58. F. Sadowski, *Farbe+Lack* **80** (1974) 29 – 33.
59. S. Keifer in: *Congr. FATIPEC XII,* 1974, Verlag Chemie, Weinheim 1975, pp. 517 – 522.
60. P. Kresse in: *Congr. FATIPEC XIII,* 1976, EREC, Paris 1977, pp. 346 – 353.
61. L. Gall, *DEFAZET – Dtsch. Farben Z.* **30** (1976) 127 – 128.
62. F. Finus, W. Fleck, *DEFAZET – Dtsch. Farben Z.* **28** (1974) 153 – 158.
63. G. Lincke: *Einfärben Kunststoffe*, VDI-Verlag, Düsseldorf 1975, pp. 135 – 142.
64. D. Bialas in H. Kittel (ed.): *Lehrbuch der Lacke und Beschichtungen*, **vol. II**, Verlag W. A. Colomb, Berlin 1974, pp. 486 – 514.
65. G. Kämpf, *DEFAZET – Dtsch. Farben Z.* **31** (1977) 347 – 353.
66. G. R. Joppien, *Farbe+Lack* **81** (1975) 1102 – 1109.
67. G. Marwedel, *Farbe+Lack* **82** (1976) 789 – 792.
68. E. Herrmann, *DEFAZET – Dtsch. Farben Z.* **29** (1975) 116 – 118.
69. S. Keifer, *Farbe+Lack* **79** (1973) 1161 – 1166.
70. P. Kresse, *DEFAZET – Dtsch. Farben Z.* **28** (1974) 459 – 466.
71. M. Miele, *Farbe+Lack* **81** (1975) 495 – 504.
72. A. Zosel in: *Congr. FATIPEC XIII,* 1976, EREC, Paris 1977, pp. 653 – 660.
73. Th. Bohmann in: *Congr. FATIPEC IX,* 1968, Chimie des Peintures, Brussels 1969, pp. 47 – 50.
74. O. J. Schmitz in: *Congr. FATIPEC XII,* 1974, Verlag Chemie, Weinheim 1975, pp. 511 – 515.
75. L. Gall, U. Kaluza, *DEFAZET – Dtsch. Farben Z.* **29** (1975) 102 – 116.
76. A. Klaeren, H. G. Völz, *Farbe+Lack* **81** (1975) 709 – 716.
77. H. G. Völz, *Farbe+Lack* **96** (1990) 19.
78. H. G. Völz, *Progr. Org. Coat.* **33** (1998) 101.
79. U. Kaluza, *DEFAZET – Dtsch. Farben Z.* **27** (1973) 427 – 439; *DEFAZET – Dtsch. Farben Z.* **28** (1974) 449 – 459.
80. H. A. Gardner, G. G. Sward: *Paints, Varnishes, Lacquers and Colors*, Gardner Laboratory, Inc., 12th ed., Bethesda, USA 1962, pp. 251 – 253.
81. U. Kaluza, *Farbe+Lack* **80** (1974) 404 – 415.
82. H. A. Gardner, G. G. Sward in [80], pp. 501 – 509.
83. E. Stock: *Analyse der Körperfarben*, Wissenschaftl. Verlags GmbH, Stuttgart 1953.
84. P. Kresse, *Farbe+Lack* **71** (1965) 178 – 184.
85. W. Papenroth, K. Kox, *DEFAZET – Dtsch. Farben Z.* **28** (1974) 311 – 313.
86. *Landolt-Börnstein*, **part 8.** H. Wienand, W. Ostertag, *Bull. Bismuth Inst. (Brussels)* (1988) no. 5.
87. G. Clarke, *Ind. Miner. (London)* **251** (1988) Aug., 17 – 31.
88. D. Williams, *Eur. Chem. News* **67/68** (1996) 17. – 23. June, 36 – 39.
89. D. Rohe, *Chem. Ind. (Duesseldorf)* (1996) no. 10, 16 – 18.
90. R. Wyckoff: *Crystal Structure*, J. Wiley & Sons, New York 1965.
91. *Gmelin*, **41**, pp. 242.
92. H. P. Boehm, *Chem. Ing. Tech.* **46** (1974) 716.
93. A. R. v. Veen, *Z. Phys. Chem. Suppl.* **162** (1989) 215 – 219.
94. G. N. Schrauzer, T. D. Guth, *Am. Chem. Soc. Div. Org.* **99** (1977) 7189.
95. R. R. Towner, J. M. Gray, K. M. Porter, US-Geological, Survey Circular 930-G, United States Government Printing Office, Supt. of Docs. no.: I 19.4/2: 930-G (1988).
96. K. A. Taylor, *In. Miner. (London)* (1996) Apr., 47 – 63.
97. R. Adams, Environment Matters (GB), February 1990, EM 6/6 (1990).
98. Companhia Vale do Rio Doce, Projecto Titanio, CURO-reviste, vol. 7, no. 23, Mar. 86 (1986).
99. R. Leutz, *Erzmetall* **42** (1989) no. 9, 383.
100. British Titanium Products, DE-OS 2 038 244, 2 038 246 – 248, 1970 (F. R. Williams *et al.*); 2 038 245, 1970 (J. Whitehead *et al.*).
101. Ruthner, DE-OS 1 533 123, 1966.
102. Oceanic Process Corp., US 3 252 787, 1966 (C. D. Shiam).
103. Anglo American Corp. of South Africa, DE-AS 1 948 742, 1969 (T. J. Coyle, H. J. Bovey).
104. Laporte, DE-AS 1 184 292, 1961.
105. Du Pont, DE-AS 1 218 734, 1965 (J. W. Reeves).
106. American Cyanamid, DE-OS 2 744 805, 1977 (P. J. Preston *et al.*).
107. W. Dunn, DE-OS 2 528 894, 1975 (W. E. Dunn, Jr.,
108. Cochran, GB 1 368 564, 1974 (A. G. Starliper, A. A. Cochran).
109. Du Pont, US 3 926 614, 1974 (H. H. Glaeser).
110. Mitsubishi Chemical, US 3 950 489, 1974 (S. Fukushima).
111. G. Clarke, *Ind. Miner. (London)* **225** (1986) June, 47 – 55.
112. J. Barksdale: *Titanium*, 2nd ed., Ronald Press Comp., New York 1966, p. 240 ffx.
113. Du Pont, US 2 098 025, 1935, US 2 098 055, 1935 (J. E. Booge, J. J. Krchma, R. H. McKinney).
114. Bayer, DE 2 015 155, 1970 (G. Kienast, H. Stütgens, H. G. Zander).
115. FS Ishihara Sangyo Kaisha, Ltd., 30–5166, 1955.
116. Bayer, DE-OS 2 951 799, 1979 (P. Panek, W. Gutsche, P. Woditsch).
117. Bayer, DE 2 454 220, 1974 (R. Leiber, J. Leuriclan, J. Renier).
118. J. Barksdale, in [95], pp. 264, 276, 278.
119. Du Pont, US 2 446 181, 1946 (R. B. Kraus); DE 1 467 357, 1964 (E. L. Larins, O. Kleinfelder); US 3 628 913, 1969 (K. L. Uhland).
120. Kronos Titan, EP 0 265 551, 1990 (A. Hartmann, H. Thumm).
121. PPG Industries Inc., DE 1 592 960, 1967 (H. W. Rahn, K. W. Richardson).
122. Du Pont, DE 1 767 798, 1968 (J. R. Auld).
123. Cabot Corp., DE 1 908 747, 1969 (H. Weaver, R. B. Roaper II, Jr., US 3 607 049, 1970 (H. Weaver, R. B. Roaper II, Jr.,
124. Du Pont, US 5 266 108, 1992 (H. M. Hauck).
125. Du Pont, US 5 201 949, 1992 (A. Allen, G. R. Evers).
126. Du Pont, DE 2 342 889, 1973 (A. H. Angerman, C. G. Moore).
127. Du Pont, US 4 781 761, 1987 (H. W. Jacobson).
128. Ishihara Sangyo Kaisha, J 5 8134–158 A, 1982.
129. Du Pont, US 2 885 366, 1956 (R. K. Jeer).
130. Tioxide Group Ltd., GB 1 008 652, 1961 (A. W. Evans, C. Shon).
131. Kronos Titangesellschaft mbH, DE 1 208 438, 1960 (H. Rechmann, F. Vial, H. Weber). British Titanium Products, DE 1 467 412, 1965 (J. R. Moody, G. Lederer).
132. Du Pont, US 4 461 810, 1984 (H. G. Jacobson).
133. Bayer, DE 2 946 549, 1977 (K. Köhler, P. Woditsch, H. Rieck, F. Rodi).
134. U. Rothe, K. Velleman, H. Wagner, *Polym. Paint Colour J.* **182** (1992).
135. D. Ainley, *Chem. Rev.* **93** (1993) no. 11, 18 – 22.
136. Metallgesellschaft AG, Bayer, DE 2 529 708, 1977 (K. H. Dörr *et al.*).
137. Kronos, EP 313 715, 1989 (A. Kulling, A. Schinkitz, J. Mauer, J. Steinhausen).
138. Bayer, EP 133 505, 1985 (R. Gerken, G. Lailach, E. Bayer, W. Gutsche).

139. Bayer, EP 194 544, 1986 (R. Gerken, G. Lailach, A. van Fürden).
140. Bayer, EP 97 259, 1984 (R. Gerken, G. Lailach, K. H. Schultz).
141. Bayer, EP 132 820, 1985 (G. Lailach, R. Gerken, W. D. Müller, K. Brändle).
142. G. Mie, *Ann. Phys.* **25** (1908) 377. Links
143. Bayer, EP 393 430, 1990 (G. Lailach *et al.*).
144. W. Schäfer, *Korrespondenz Abwasser* **2** (1989) 143 – 147.
145. Du Pont, WO 95/31414, 1995 (P. Hill *et al.*).
146. J. Fischer, *Ind. Miner. (London)* (1996) Aug., 47 – 55.
147. G. Kämpf *et al.*, *Farbe+Lack* **79** (1973) 9 ffx.
148. H. G. Völz *et al.*, *Farbe+Lack* **82** (1976) no. 9, 805 ffx.
149. D. V. Boust, Intertech Conference, Vancouver, April 1997.
150. K. K. Ricoh, J 5 8025–363, 1981.
151. M. Kotter, H.-G. Lintz, F. Weyland, *Chem.-Ing.-Tech.* **58** (1986) no. 8, 617.
152. Mitchell Market Reports: Titania and the Titanates, 1990.
153. H. Krüger, *VGB Kraftwerkstechnik* **11** (1994) 371 – 395.
154. *Energie Spektrum* (1991) Apr., 20 – 21.
155. P. S. Indin, *Verfkroniek* **67** (1994) no. 11, 17 – 19.
156. Y. Parent, *Solar Energy* **56** (1996) no. 5, 429 – 437. Links
157. P. Fournier, *C. R. Hebd. Seances Acad. Sci.* **231** (1950) 1343.
158. H. Mühle, *Am. J. Ind. Med.* **15** (1989) 343 – 346. Links
159. H. Clausen: "Zinc-Based Pigments," in P. A. Lewis (ed.): *Pigment Handbook,* 2nd ed., vol. 1, John Wiley & Sons, New York 1988.
160. M. Issel, *Modern Paint and Coatings* **9** (1991) 35 – 42.
161. M. Cremer: "Non TiO_2 White Pigments with Special References to ZnS Pigments," Industrial Minerals, Pigment & Extenders Supplement, 1985.
162. H. Brown: *Zinc Oxide Rediscovered*, New Jersey Zinc Co., New York 1957.
163. H. Brown: *Zinc Oxide Properties and Applications*, International Lead Zinc Research Organ, New York 1976.
164. K. H. Ulbrich, W. Backhaus: "Zinkoxid in der Gummiindustrie," *Kautsch. Gummi, Kunstst.* **27** (1974) no. 7, 269 – 272.
165. G. Hänig, K. Ulbrich: "ZnO, Produkt zwischen Pigmentchemie und Hüttenwesen," *Erzmetall* **32** (1979) 140 – 146.
166. G. Heiland, E. Mollwo, F. Stockmann: *Electronic Processes in Zinc Oxide,* vol. 8, Solid State Physics, Academic Press, New York 1959.
167. Larvik, US 2 939 783, 1957 (G. Lundevall).
168. G. Meyer: *New Application for Zinc*, Zinc Inst. Inc. New York 1986.
169. E. Ziegler, R. Helbig, *Physik Unserer Zeit* **17** (1986) no. 6, 171 – 177.
170. Matsushita, DE 3 045 591, 1980 (Sonoda).
171. Kieffer, in R. Henklin (ed.): *Metalle als Lebensnotwendige Spurenelement für Pflanzen, Tiere und Menschen, Teil Zink*, Verlag Chemie, Weinheim, Germany 1984, pp. 117 – 123.
172. U. Schwertmann, R. M. Cornell: *IronOxides in the Laboratory, Preparation, and Characterization*, VCH Verlagsgesellschaft, Weinheim, Germany 1991. R. M. Cornell, U. Schwertmann: *The Iron Oxides*, VCH Verlagsgesellschaft, Weinheim, Germany 1996.
173. J. L. W. Jolly, C. T. Collins: "Natural Iron Oxide Pigments," *Iron Oxide Pigments*, part 2, Information Circular-Bureau of Mines 8813, Washington 1980.
174. H. Kittel: *Lehrbuch der Lacke und Beschichtungen*, **vol. II**, Verlag, W. A. Colomb Berlin 1974, p. 109.
175. E. Ack, *Farben Ztg.* **28** (1922/23) 493.
176. M. A. Bouchonnet, *Bull. Soc. Chim. Fr.* (1912) 9 345.
177. H. Wagner, R. Haug: "Gelbe Eisenoxydfarben," **8 d,** *Veröffentlichung des Fachausschusses für Anstrichtechnik bei VDI und VDCh,* VDI-Verlag, Düsseldorf 1934.
178. I. G. Farbenind., US 1 813 649, 1929 (P. Weise).
179. Minnessota Mining & Manuf. Co., US 2 634 193, 1947 (G. E. Noponen).
180. Minnessota Mining & Manuf. Co., US 2 452 608, 1941 (G. B. Smith).
181. Verein Österr. Eisen- und Stahlwerke, OE 176 206, 1952 (E. Petzel).
182. The Nitralloy Corp., US 2 592 580, 1945 (H. Loevenstein).
183. E. V. Carter, R. D. Laundon, *J. Oil Colour Chem. Ass.*, 1990 (1) 7 – 15.
184. Bayer, DE 2 653 765, 1976 (B. Stephan, G. Winter).
185. Bayer, DE 3 820 499, 1988 (B. Kröckert, G. Buxbaum, A. Westerhaus, H. Brunn).
186. Bayer, DE-AS 1 191 063, 1963 (F. Hund, H. Köller, D. Räde, H. Quast).
187. BASF, DE-OS 2 517 713, 1975 (W. Ostertag *et al.*).
188. Magnetic Pigment Co., US 1 424 635, 1919 (P. Fireman).
189. Ault & Wiberg Co., US 1 726 851, 1922 (E. H. McLeod).
190. Interchem. Corp., US 2 388 659, 1943 (L. W. Ryan, H. L. Sanders).
191. C. K. Williams & Co., US 3 133 267, 1934.
192. Reconstruction Finance Corp., US 2 631 085, 1947 (L. M. Bennetch).
193. Reymers Holms Gamla Ind., GB 668 929, 1950 (T. G. H. Holst, K. A. H. Björned).
194. Glemser, DE 704 295, 1937 (O. Glemser).
195. Pfizer, DE-OS 2 212 435, 1972 (L. M. Bennetch, H. S. Greiner, K. R. Hancock, M. Hoffman).
196. Hollnagel, Kühn, DL 26 901, 1960 (M. Hollnagel, E. Kühn).
197. C. K. Williams & Co., US 2 620 261, 1947 (T. Toxby).
198. West Coast Kalsomine Co., US 1 327 061, 1917 (R. S. Penniman, N. M. Zoph).
199. National Ferrite Co., US 1 368 748, 1920 (R. S. Penniman, N. M. Zoph).
200. Frazee, US 1 923 362, 1927 (V. Frazee).
201. Magnetic Pigment Co., US 2 090 476, 1936 (P. Fireman).
202. C. K. Williams & Co., US 2 111 726, 1932 (G. Plews); US 2 111 727, 1937 (G. Plews).
203. Magnetic Pigment Co., US 2 127 907, 1937 (P. Fireman).
204. Bayer, DE 902 163, 1951 (B. H. Marsh).
205. C. K. Williams & Co., US 2 785 991, 1952 (L. M. Bennetch).
206. Bayer, FR 1 085 635, 1953 (F. Hund); DE 1 040 155, 1954 (F. Hund).
207. Mineral Pigments, US 2 633 407, 1947 (D. W. Marsh).
208. I. G. Farbenind., DE 463 773, 1925 (J. Laux).
209. I. G. Farbenind., DE 515 758, 1925 (J. Laux).
210. I. G. Farbenind., DE 551 255, 1930 (U. Haberland).
211. Bayer, EP 0 249 843, 1987 (A. Westerhaus, K. W. Ganter, G. Buxbaum).
212. I. G. Farbenind., DE 466 463, 1926 (W. Schubardt, M. Grote).
213. Bayer, EP 0 014 382, 1980 (G. Franz, F. Hund).
214. R. C. Rowe, *Pharm. Int.* **9** 1984 221 – 224.
215. Bayer, DE 3 326 632, 1983 (W. Burow, H. Printzen, H. Brunn, K. Nollen).
216. "Properties and Economics," Pigment Handbook, 2nd ed., vol. I, John Wiley & Sons, New York 1988, pp. 309 – 310.
217. S. Keifer, A. Wingen, *Farbe+Lack* **79** (1973) 866–873.
218. J.-L. Lassaigne, *Ann. Chim. Phys., Ser. 2*, **14** (1820) 299 – 302.

219. H. C. Roth, US 1 728 510, 1927.
220. PPG Industries, US 4 127 643, 1977 (W. W. Carlin).
221. BASF, EP-A 0 238 713, 1986 (N. Müller *et al.*).
222. Pfizer Inc., EP 0 068 787, 1982 (V. P. Rao).
223. Bayer, US 4 040 860, 1976 (L. Mansmann, W. Rambold).
224. Bayer, DE 2 635 086, 1976 (W. Rambold, H. Heine, B. Raederscheidt, G. Trenczek).
225. British Chrome & Chemicals, US 4 296 076, 1981 (A. S. Dauvers, M. A. Marshall).
226. C. K. Williams & Co., US 2 250 789, 1940 (I. W. Ayers).
227. Bayer, DE 1 257 317, 1963 (F. Hund, D. Räde).
228. Bayer, DE-OS 2 358 016, 1973 (B. Knickel *et al.*).
229. Colores Hispania S. A., ES 438 129, 1975.
230. C. K. Williams & Co., US 2 560 338, 1950 (C. G. Frayne); C. K. Williams & Co., US 2 695 215, 1950 (W. A. Pollock).
231. Bayer, US 3 723 611, 1971 (G. Broja, K. Brändle, C. H. Elstermann).
232. Gefahrstoffverordnung – GefstoffV Dec. 16, 1987 (BG Bl. I, 2721).
233. Bayer, DE 3 123 361, 1981 (J. Rademachers *et al.*).
234. 37. Abwasser VWV 05. Sept. 1984, GMBL Nr. 22, 344; Water Pollution Control Law, Law No. 138, Dec. 25 1970 (Japan).
235. 178. Mitteilung BGesundhBl. 31 (1988) 363.
236. *Journal Officiel de la République Francaise* no. 1227 (1983) 75, 97, 109.
237. *Gazzetta Ufficiale della Republica Italiana*, Apr. 20, 1973, 3, 4.
238. *Belgisch Staatsblad* – Moniteur Belge Sept. 24, 1976, 12 030.
239. Verpakkingen – en Gebruiksartikelenbesluit (Warenwert) 01. Oct. 1980.
240. EG-Guidelines EEC 88/378, official gazette L 187 16. July 1988.
241. EEC Council Directive, EEC 76/768, on cosmetic products (18th Suppl.).
242. Japanese Pharmaceutical Affaires 1980.
243. E. Püttbach, *Betonwerk + Fertigteiltechnik* **2** (1987) 124 – 132.
244. S. Ivankovic, R. Preussmann, *Ed. Cosmet. Toxicol.* **13** (1975) 347 – 351.
245. U. Korallus, H. Ehrlicher, E. Wüstefeld, *ASP* **3** (1974) 51 ffx.
246. EG-Guidelines 67/548/EEC 1967 and Suppl.
247. DCMA Classification and Chemical Description of the Complex Inorganics Color Pigments, 3rd. ed., Dry Color Manufacturers' Association, Alexandria, Virginia, 1991.
248. V. M. Goldschmidt, *Skr. Nor. Vidensk. Akad. Kl 1: Mat. Naturvidensh. Kl* 1926 Sept., 17.
249. F. Hund, *Angew. Chem.* **74** (1962) 23 – 27. Links F. Hund, *Farbe + Lack* **73** (1967) 111 – 120.
250. Bayer, US 3 022 186, 1959 (F. Hund).
251. Du Pont, US 2 257 278, 1939 (H. H. Schaumann).
252. The Harshaw Chemical Comp., US 2 251 829, 1939 (C. J. Harbert).
253. Bayer, DE-AS 1 195 913, 1959 (H. Kyri, H. Weber).
254. Ferro Corp., US 3 832 205, 1973 (H. E. Lowery).
255. Bayer, DE-OS 3 740 635, 1987 (V. Wilhelm, M. Mansmann).
256. BASF Lacke + Farben AG, EP-A 0 233 601, 1987 (H. Knittel, R. Bauer, E. Liedek, G. Etzrodt).
257. G. Blasse, *Philips Res. Rep. Suppl.* **3** (1964) 1 – 139.
258. O. Schmitz-Dumont, B. Raederscheidt, *Naturwissenschaften* **41** (1964) 477 – 478.
259. Kewanee Oil Company, US 3 424 551, 1964 (I. E. Owen).
260. R. A. Eppler, *J. Am. Ceram. Soc.* **66** (1983) 794–801. Links
261. Deutsche Gold- und Silber-Scheideanstalt, formerly Roessler, US 3 876 441, 1974 (A. Broll, H. Mann) J. Alarcon, P. Escribano, R. M. Marin, *Br. Ceram. Trans. J.* **84** (1985) 172 – 175.
262. Bayer, DE-OS 3 819 626, 1988 (J. Rademachers, V. Wilhelm, S. Keifer, W. Burow).
263. S. H. Murdock, R. A. Eppler, *J. Am. Ceram. Soc.* **71** (1988) C-212 – C-214.
264. Bayer, DE-AS 1 191 063, 1963 (F. Hund, H. Köller, D. Räde, H. Quast).
265. BASF, DE-OS 2 517 713, 1975 (W. Ostertag *et al.*).
266. Bayer, DE-OS 1 767 868, 1968 (W. Holznagel, F. Hund, G. Gerlach).
267. H. B. Krause, H. W. Reamer, J. L. Martin, *Mater. Res. Bull.* **3** (1968) 233 – 240. Links
268. BASF Lacke+Farben AG, DE 2 936 746, 1979 (A. Seitz *et al.*).
269. Bayer, EP 0 075 197, 1982 (D. Messer, V. Wilhelm, R. Endres, H. Heine).
270. 178. Mitteilung, Bundesgesundheitsblatt 31 (1988) 363.
271. *Journal Officiel de la République Française* no. 1227 (1983) 75, 97, 109.
272. *Gazzetta Ufficiale della Republica Italiana* April 20, 1973, pp. 3, 4.
273. *Belgisch Staatsblad* – Moniteur Belge Sept. 24, 1976, p. 12 030.
274. Verpakkingen – en Gebruiksartikelenbesluit (Warenwet) Oct. 1, 1980.
275. EEC Council Directive, EEC 89/107, on coloring matters authorized for use in foodstaffs intended for human consumption.
276. EEC Council Directive, EEC 76/768, on cosmetic products.
277. EEC Council Directive, EEC 78/25, on coloring matters which may be added to medicinal products.
278. E. Bomhard *et al.*, *Toxicol. Lett.* **14** (1982) 189–194. Links
279. D. Steinhoff, U. Mohr, *Exp. Pathol.* **41** (1991) 169 – 174. Links
280. H. Endriß, D. Räde, *Kunststoffe* **79** (1989) 617–619.
281. K. Weiss, *Z. Naturforsch. A Astrophys. Phys. Phys. Chem.* **2A** (1947) 650 – 652.
282. S. F. Ravik, *J. Phys. Chem.* **40** (1936) 69.
283. Siegle & Co., DE 1 007 907, 1955.
284. Bayer, DE-OS 2 151 234, 1971.
285. H. Endriß, *Kunststoffe* **75** (1985) 10, 758 – 761.
286. D. Räde, A. Dornemann, *Proceedings of the Third International Cadmium Conference Miami, 1981*, Cadmium Association London, pp. 37 – 40.
287. H. Endriß, *Kunststoffe* **69** (1979) 39 – 43.
288. European Commission: “On the Use of Colourants in Plastic Materials Coming in Contact with Food,” Resolution AP (89) 1., Sept. 13, 1989.
289. EEC Council Directive, EEC 83/514, Sept 26, 1983, on the limits for cadmium in waste water.
290. G. Rusch, J. S. O'Grodnick, W. E. Rinehart, *Am. Ind. Hyg. Assoc. J.* **47** (1986) no. 12, 754. Links
291. H.-J. Klimisch, C. Gembardt, H.-P. Gelbke: *Lung Deposition and Clearance, Lung Pathology and Renal Accumulation of Inhaled Cadmium Chloride and Cadmium Sulfide in Rat*, BASF, Ludwigshafen 1991.
292. S. A. Robinson, O. Cantoni, M. Costa, *Carcinogenesis (London)* **3** (1982) no. 6, 657.
293. M. Costa, J. D. Heck, S. H. Robinson, *Cancer Res.* **42** (1982) 2757. Links
294. F. Pott *et al.*, *Exp. Pathol.* **32** (1987) 129. Links
295. H. Oldiges, D. Hochrainer, M. Glaser, *Toxicol. Environ. Chem.* **19** (1988) 217.
296. U. Heinrich *et al.*: “Investigation of the Carcinogenic Effects of Various Cadmium Compounds after Inhalation Exposure in Hamsters and Mice,” *Exp. Pathol.* **38**, in press. Links
297. EEC Council Directive, EEC 91/338, 1991.
298. E. Zintl, L. Vanino, DE 422 947, 1924.

299. Du Pont, US 4 026 722, 1976 (R. W. Hess); DE 2 727 864, 1977 (D. H. Piltingsrud); DE 2 727 863, 1977 (R. W. Hess); US 4 063 956, 1976 (J. F. Higgins).
300. Montedison, DE 2 933 778, 1979 (L. Balducci, M. Rustioni); DE 2 940 185, 1979 (L. Balducci, M. Rustioni); DE 3 004 083, 1980 (L. Balducci, M. Rustioni); DE 3 106 625, 1981 (L. Balducci, M. Rustioni).
301. BASF, EP 74 049, 1982 (H. Wienand, W. Ostertag, K. Bittler); BASF, EP 271 813, 1986 (H. Wienand, W. Ostertag, C. Schwidetzky, H. Knittel); BASF, DE 3 926 870, 1989 (H. Wienand, W. Ostertag, C. Schwidetzky).
302. Bayer, DE 3 315 850, 1983 (P. Köhler, P. Ringe); DE 3 315 851, 1983 (P. Köhler, P. Ringe, H. Heine); DE 4 119 668, 1991 (F. Schwochow, R. Hill); EP 492 244 1991 (F. Schwochow); EP 723 998 1996 (H. Schittenhelm, R. Hill)
303. Ciba-Geigy, EP 239 526, 1987 (F. Herren); EP 304 399, 1988 (R. Sullivan); EP 430 888, 1990 (L. Erkens, G. Schmitt, H. Geurts, W. Corvers); DE 4 037 878, 1990 (F. Herren, L. Erkens); US 5 399 335, 1993 (R. Sullivan).
304. BASF, EP551 637, 1992 (E. Liedek, H. Knittel, H. Reisacher, N. Mronga, H. Ochmann, H. Wienand); EP 640 566, 1994 (G. Etzrodt, H. Knittel, H. Reisacher).
305. H. Endriß, *Farbe + Lack* **100** (1994) 397 – 398. H. Endriß, M. Haid *Kunststoffe* **86** (1996) 538 – 540.
306. F. Hund, *Farbe + Lack* **73** (1967) 111 – 120.
307. L. J. H. Erkens *et al.*, *J. Oil Colour Chem. Assoc.* **71** (1988) no. 3, 71 – 77.
308. H. Wagner et al. *Z. Anorg. Allg. Chem.* **208** (1931) 249; *Farben-Ztg.* **38** (1933) 932.
309. J. F. Clay & Cromford Color, GB 730 176, 1951; H. Lesche, *Farbe + Lack* **65** (1959) 79, 80.
310. Du Pont, US 2 808 339, 1957 (J. J. Jackson).
311. Du Pont, DE-OS 1 807 891, 1969 (H. R. Linton).
312. Bayer, DE-OS 1 952 538, 1969 (C. H. Elstermann, F. Hund).
313. ICI, DE-OS 2 049 519, 1970 (Ch. H. Buckley, G. L. Collier, J. B. Mitchell).
314. Ten Horn Pigment, DE-OS 2 600 365, 1976 (J. J. Einerhand *et al.*).
315. BASF, DE 3 323 247 A1, 1983 (E. Liedek *et al.*).
316. Heubach, DE 3 806 214 A1, 1988 (I. Ressler, W. Horn, G. Adrian).
317. Heubach, DE 3 906 670 A1, 1989 (I. Ressler, W. Horn, G. Adrian).
318. R. Williams, Jr., "Continuous Chrome Yellow Process," *Chem. Eng. (N.Y.)* 1949, March.
319. Chemokomplex Vegyipari-Gep es Berendezes Export, Import Vallalat, DE-AS 1 592 848, 1971 (J. Scholtz et al.).
320. H. Schäfer, *Farbe + Lack* **77** (1971) no. 11, 1081 – 1089.
321. Hoechst, DE-OS 2 127 279, 1971 (R. Kohlhaas *et al.*).
322. Hoechst, DE-OS 2 062 775, 1970 (R. Kohlhaas *et al.*).
323. Sherwin-Williams, US 2 237 104, 1938 (N. F. Livingston).
324. E. Renkwitz, *Farben-Ztg.* **28** (1923) 1066; H. Levecke, *Farbe + Lack* **42** (1936) 41 – 43; H. Berger, *Arbeitsschutz* 1941, III/44.
325. *Paint Oil Chem. Rev.* **95** (1933), 86 – 92; A. E. Newkirk, S. C. Horning, *Ind. Eng. Chem. Ind. Ed.* **33** (1941) 1402 – 1407.
326. *Internationale MAK-Werte-Liste* (1990), Ecomed Verlag; TRGS 505 Blei (1988), Carl Heymanns Verlag.
327. EEC Council Directive, EEC 82/605, July 28, 1982.
328. J. M. Davies, *Lancet* **2** (1978) 18; Links *Br. J. Ind. Med.* **41** (1984) 158 – 169. Links
329. W. C. Cooper, Dry Colour Manufacturers' Association, Arlington 1983.
330. TA-Luft (Technische Anleitung zur Reinhaltung der Luft (1986), C. H. Beck'sche Verlagshandlung, München 1987.
331. 37. Abwasser VwV (05.09.1984) für Direkteinleiter (Herstellung Anorganischer Pigmente).
332. 37. Anhang zur Rahmen-Abwasser-Verwaltungsvorschrift vom 27. 08. 1991; § 7 a Wasserhaushaltsgesetz.
333. Gefahrstoffverordnung (GefStoffV), Okt. 26, 1993, Stand 10.94, und Chemikalien-Verbotsverordnung (ChemVerbV), Okt. 26, 1993, Stand 10.94, Deutscher Bundesverlag GmbH, Bonn.
334. EEC Council Directive, EEC 67/548, June 27, 1967.
335. EEC Council Directive, EEC 93/21, April 27, 1993 (18th adaption of EEC 67/548).
336. EEC Council Directive, EEC 93/72, Sept. 1, 1993 (19th adaption of EEC 67/548).
337. EEC Council Directive, EEC 94/96, Dec. 12, 1994 (21st adaption of EEC 67/548).
338. EEC Council Directive, EEC 93/18, April 5, 1993 (3rd adaption of EEC 88/379).
339. EEC Council Directive, EEC 94/60, Dec. 20, 1994 (14th ammendment of EEC 76/769).
340. Erste Verordnung zur Änderung chemikalienrechtlicher Verordnungen vom 12. Juni 1996, Deutscher Bundesverlag GmbH, Bonn 1996.
341. EEC Council Directive, EEC 89/178, Feb. 22, 1989.
342. Lebensmittel- und Bedarfsgegenständegesetz vom 15.08.1974 (BGBl S.1945).
343. Sicherheit von Spielzeug (Safety of toys) EN 71 Part 3, Dec. 1994.
344. S. E. Tarling, P. Barnes, A. L. Mackay, *J. Appl. Cryst.* **17** (1984) 96 – 99.
345. S. E. Tarling, P. Barnes, J. Klinowsky, *Acta Cryst.* **B44** (1988) 128 – 135.
346. R. J. H. Clark, M. L. Franks, *Chem. Phys. Lett.* **34** (1975) 69 – 72. Links
347. R. J. H. Clark, D. G. Cobbold, *Inorg. Chem.* **17** (1978) 3169 – 3174. Links
348. R. J. H. Clark, T. J. Dines, M. Kurmoo, *Inorg. Chem.* **22** (1983) 2766 – 2772. Links
349. Interchemical Corp., US 2 535 057, 1950 (A. E. Gessler, C. A. Kummins).
350. Toyo Soda, JP-KK 8 1 045 513, 1978.
351. Reckitt's Colours Ltd., *The Cost of Whiteness*, Hull, United Kingdom.
352. *Ullmann*, 4th ed., **18**, 623.
353. C. Clauss, E. Gratzfeld in H. Kittel (ed.): *Pigmente*, 3rd ed., Wissenschaftl. Verlags GmbH, Stuttgart 1960.
354. M. F. Dix, A. D. Rae, *J. Oil Colour Chem. Assoc.* **61** (1978) 69.
355. A. Ludi, "Berliner Blau", *Chem. Unserer Zeit* **4** (1988) 22.
356. Degussa *Vossen-Blau-Pigmente,* Frankfurt/M. 1973.
357. G. K. Wertheim *et al.*, *J. Chem. Phys.* **54** (1971) 3235. Links H. Buser *et al.*, *J. Chem. Soc. Chem. Commun.* **23** (1972) 1299.
358. M. L. Napijalo, V. Stefancic, *Fizika (Zagreb)* **8 Suppl.** (1976) 16.
359. R. J. Emrich *et al.*, *J. Vac. Sci. Techn. A* **5** (1987) 1307.
360. F. Herren *et al.*, *Inorg. Chem.* **19** (1980) 956. Links A. Ludi, *Chem. Unser Zeit* **22** (1988) 123.
361. Degussa AG, DE 1 188 232, 1964 (E. Gratzfeld).
362. Degussa AG, DE 976 599, 1952 (H. Verbeek, E. Gratzfeld).
363. Chem. Fabrik Wesseling AG, DE 1 061 935, 1955 (H. Verbeek, E. Gratzfeld).
364. Degussa AG, DE 233 669, 1962 (E. Gratzfeld).
365. Degussa AG, DE-OS 1 792 418, 1968 (E. Gratzfeld, E. Clausen, E. Ott).
366. Degussa AG, DE 1 937 832, 1969 (E. Gratzfeld).
367. Degussa AG, DE 1 949 720, 1969 (E. Gratzfeld, E. Kühn).
368. L. Müller-Fokken, *Farbe+Lack* **84** (1978) 489.
369. H. Ferch, H. Schäfer: *18th AFTPV-Kongreßbuch*, Nice 1989, p. 315.
370. "Fotometrische Messung tiefschwarzer Systeme". *Schriftenreihe Pigmente*, Nr. 24, Degussa AG, 192Frankfurt/M. 1989
371. "Vossen-Blau zur Färbung von Fungiziden." *Schriftenreihe Pigmente*, Nr. 50, Degussa AG, Frankfurt/M. 1985.
372. W. Koblet, *Schweiz. Z. Obst Weinbau* **1** (1965) 8.
373. H. Wiedmer et al.*: Agro-Dok* no. D 4341, Sandoz AG, Basel 1977.
374. R. Ciferri: "Le 4 Stagioni" (Montecatine) **4** (1963) no. 2, 2.
375. V. Nigrovic, *Int. J. Rad. Biol.* **7** (1963) 307.
376. V. Nigrovic, *Phys. Med. Biol.* **10** (1965) 81. Links

377. I. V. Tananayev, *Zh. Neorg. Khim.* **1** (1956) 66.
378. P. Dvorak, *Z. Gesamte Exp. Med.* **89** (1969) 151.
379. P. Dvorak, *Arzneim. Forsch.* **20** (1970) 1886.
380. P. Dvorak, *Z. Naturforsch. B* **26** (1971) 277.
381. V. Nigrovic, F. Bohne, K. Madshus, *Strahlentherapie* **130** (1966) 413.
382. P. Dvorak, *Z. ges. exp. Med.* **151** (1969) 89 – 91.
383. T. A. Shashina, *et al.*, *Gig. Tr. Prof. Zabol.*, **1** (1991) 35 – 36. Links
384. J. M. Verzijl *et al.*, *Clinical Toxicol.* **31** (1993) 553 – 562.
385. B. D. Nielsen *et al.*, *Z. Naturforsch.* **45** (1990) 681 – 690.
386. Degussa AG, unpublished report: Degussa AG US-IT-Nr. 84–0074-DKT (1984) and 88–0083-DKT, 88–0084-DKT (1988a).
387. *NPIRI (National Printing Ink Research Institute) Raw Materials Data Handbook*, Napim, New York, **4** (1983) 21.
388. Degussa AG, unpublished report: Degussa AG US-IT-Nr. 77–0069-FKT (1977).
389. Degussa AG, unpublished report: Degussa AG US-IT-Nr. 85–0081-DKT,85–0080-DKT (1985a), 87–0038-DKT, 87–0039-DKT, 87–0040-DKT (1987) and 88–0085-DKT (1988b).
390. F. Leuschner, H. Otto, Unpublished (1967), in: *Kosmetische Färbemittel*, Harald Boldt Verlag, Boppard 1977.
391. V. Nigrovic *et al.*, *Strahlentherapie* **130** (1966) 413 – 419.
392. M. Günther, Kernforschungszentrum Karlsruhe, KFK 1326, Gesellschaft für Kernforschung m.b.H., Karlsruhe, unpublished report US-IT-Nr. 70–0001-FKT (1970).
393. P. Dvorak *et al.*, *Naunyn-Schmiedebergs Arch. Pharmak.*, **269** (1971) 48 – 56.
394. K. Madshus *et al.*, *Int. J. Radiation Biol.*, **10** (1996) 519 – 520.
395. K. Madshus, A. Strömme, *Z. Naturforsch.* **A23** (1968) 391 – 392.
396. J. P. Mulkey *et al.*, *Vet. Human Toxicol.*, **35** (1993) 445 – 453.
397. V. Pai, *West Indian Med. J.*, **36** (1987) 256 – 258. Links
398. Degussa AG, unpublished report: Degussa AG US-IT-Nr. 85–0082-DGO, 85–0085-DGO, 85–0088-DGO, 85–0078-DGO (1985b)
399. Degussa AG, unpublished report: Degussa AG US-IT-Nr. 79–0046-DKO, 79–0047-DKO, 79–0088-DKO (1979) and 85–0079-DKO, 85–0083-DGO (1985c).
400. Armour Research Foundation, US 2 694 656, 1947 (M. Camras).
401. Bayer, DE 1 061 760, 1957 (F. Hund).
402. EMI, GB 765 464, 1953 (W. Soby).
403. BASF, DE 1 204 644, 1962 (W. Balz, K. G. Malle).
404. VEB Elektrochemisches Kombinat, Bitterfeld, DD 48 590, 1965 (W. Baronius, F. Henneberger, W. Geidel).
405. Agfa-Gevaert, DE 1 592 214, 1967 (W. Abeck, H. Kober, B. Seidel).
406. Bayer, DE 1 803 783, 1968 (F. Rodi, H. Zirngibl).
407. Pfizer, US 3 498 748, 1967 (H. S. Greiner).
408. Sakai Chemical Industries, US 4 202 871, 1980 (S. Matsumoto, T. Koga, K. Fukai, S. Nakatani).
409. A. R. Corradi *et al.*, *IEEE Trans. Magn.* **MAG-20** (1984) 33 – 38.
410. Bayer, DE 1 266 997, 1959 (W. Abeck, F. Hund).
411. Y. Imaoka, S. Umeki, Y. Kubota, Y. Tokuoka, *IEEE Trans. Magn.* **MAG-14** (1978) 649.
412. 3M, US 3 573 980, 1968 (W. D. Haller, R. M. Colline).
413. Hitachi Maxell, DE 2 235 383, 1972 (Okazoe, Akira).
414. M. Kryder *et al.*, *IEEE Trans. Magn.* **MAG-23** (1987) 45.
415. Du Pont, US 2 923 683, 1957 (J. N. Ingraham *et al.*). Bayer, DE-AS 1 467 328, 1963 (F. Hund *et al.*).
416. Du Pont, US 3 034 988, 1958 (J. N. Ingraham *et al.*).
417. E. Köster in C. D. Mee, E. D. Daniel (eds.): *Magnetic Recording,* vol. I, McGraw Hill, New York 1987.
418. H. Auweter *et al.*, *IEEE Trans. Magn.* **MAG-26** (1990) 66.
419. M. W. Müller *et al.*, *IEEE Trans. Magn.* **MAG-26** (1990) 1897.
420. *Ullmann*, **A7**, 77 ffx.
421. M. Essig *et al.*, *IEEE Trans. Magn.* **MAG-26** (1990) 69.
422. Du Pont, US 3 512 930, 1968 (W. G. Bottjer *et al.*).
423. BASF, EP-A 218 234, 1985 (W. Steck *et al.*).
424. BASF, unpublished.
425. G. R. Cole *et al.*, *IEEE Trans. Magn.* **MAG-20** (1984) 19.
426. G. Bate, *J. Appl. Phys.* **52** (1981) 2447. Links
427. M. Kishimoto, S. Kitahata, M. Amemiya, *IEEE Trans. Magn.* **MAG-22** (1986) 732 – 734.
428. T. Fujiwara, *IEEE Trans. Magn.* **MAG-23** (1987) 3125.
429. D. E. Speliotis, *IEEE Trans. Magn.* **MAG-25** (1988) 4048.
430. H. Hibst, *Angew. Chem.* **94** (1982) 263. Links
431. H. Stäblein in E. P. Wohlfarth (ed.): *Ferromagnetic Materials,* vol. 3, North Holland Publ., Amsterdam, Oxford, New York, Tokyo 1982.
432. Toshiba, EP-A 39 773, 1980 (H. Endo *et al.*).
433. Toda, EP-A 150 580, 1983 (N. Nagai *et al.*).
434. Dowa Mining, DE-OS 3 527 478, 1984 (K. Aoki).
435. Sakai Chemical, DE-OS 3 529 756, 1984 (S. Jwasaki *et al.*).
436. Ishihara, EP-A 299 332, 1987 (K. Nakata *et al.*).
437. Ugine Kuhlmann, DE-OS 2 003 438, 1969 (M. G. de Bellay).
438. Toda, EP-A 164 251, 1984 (N. Nagai *et al.*).
439. Toda, EP-A 232 131, 1986 (N. Nagai *et al.*).
440. Matsushita, EP-A 290 263, 1987 (H. Toril *et al.*).
441. Toshiba, DE 3 041 960, 1979 (O. Kubo *et al.*).
442. BASF, DE-OS 3 702 036, 1987 (G. Mair).
443. R. E. Fayling, *IEEE Trans. Magn.* **MAG-15** (1979) 1567.
444. R. J. Veitch, *IEEE Trans. Magn.* **MAG-26** (1990) 1876.
445. Y. Okazaki *et al.*, *IEEE Trans. Magn.* **MAG-24** (1989) 4057.
446. D. E. Speliotis, *IEEE Trans. Magn.* **MAG-26** (1990).
447. O. Kubo *et al.*, *IEEE Trans. Magn.* **MAG-24** (1988) 2859.
448. BASF: Sicor-Pigmente, company brochure, Ludwigshafen 1987.
449. Dr. Hans Heubach GmbH & Co. KG: Heucophos company information, Langelsheim 1989.
450. Halox-Pigments: *Halox-Pigments*, company information, Pittsburgh 1973.
451. Société nouvelle des Couleurs Zinciques, company information, Beauchamp 1990.
452. Colores Hispania: *Anticorrosive Pigments*, company information, Barcelona 1990.
453. Ten Horn Pigment bv., company information, Maastricht 1977.
454. Teikoku Kako Co. Ltd.: *K-White*, company information, Osaka 1987.
455. N. R. Whitehouse, *Polym. Paint Colour J.* **178** (1988) no. 4211, 239.
456. Sherwin-Williams Chemicals: *Moly White*, company information, Coffeyville, Kansas, 1986.
457. NL Chemicals: *Nalcin 2*, company information, Hightstown, N.J., 1983

458. PPG Industries Inc., US 4 707 405, 1987 (E. Evans *et al.*)
459. Buckman Laboratories Inc.; *Barium metaborate,* company information Memphis, Tennessee, 1989.
460. Alex R. Vleeshouwers, *Polym. Paint Colour J.* **178** (1988) no. 4224, 788, 790.
461. H. Wienand, W. Ostertag, *Farbe + Lack* **88** (1982) no. 3, 183 – 188.
462. S. Wiktorek, *Surface Coat. Australia* **24** (1987) no. 5, 11 – 16.
463. W. Funke, *Farbe + Lack* **86** (1980) no. 8, 730.
464. P. Kresse, *DEFAZET Dtsch. Farben Z.* **32** (1978) no. 5, 216, 217.
465. W. Funke, *Farbe + Lack* **89** (1983) no. 2, 86–91.
466. H. Haagen, R. Konzelmann, *Farbe + Lack* **94** (1988) no. 11, 893, 894.
467. H. Haagen, D. Martinovic, *Farbe + Lack* **95** (1989) no. 12, 892 – 895.
468. E. V. Carter, R. D. Laundon, *J. Oil Colour Chem. Assoc.* 1990 (1), 7 – 15.
469. K. A. van Oeteren, *Farbe + Lack* **79** (1973) no. 1, 45 – 48.
470. V. Cupr, *Maschinenmarkt IMM Ind. J.* **84** (1978) 439 – 442.
471. K. A. van Oeteren, *Fette Seifen Anstrichm.* **84** (1982) no. 6, 242 – 245.
472. E. Leitel *et al.*, *Werkst. Korros.* **39** (1988) 1 – 7.
473. Ch. Nitsche, *Farbe + Lack* **90** (1984) no. 6, 468 – 473.
474. M. Cohen, F. J. Beck, *Z. Elektrochem.* **61** (1958) 696.
475. G. Meyer, *Farbe + Lack* **68** (1962) no. 5, 315; *Dtsch. Farben-Z.* **20** (1966) 8; *Farbe + Lack* **79** (1967) no. 6, 529.
476. J. Ruf, *Werkst. Korros.* **20** (1969) 861 – 869.
477. J. A. Burkill, J. E. O. Mayne, *J. Oil Colour Chem. Assoc.* **71** (1988) no. 9, 273 – 275.
478. G. Sziklai, J. Szucs, *Hung. J. Ind. Chem.* **10** (1982) 215 – 221. Links
479. P. J. Gardner, I. W. McArn, V. Barton, G. M. Seydt, *J. Oil Colour Chem. Assoc.* **73** (1990) no. 1, 16.
480. G. Rasack, *Farbe + Lack* **84** (1978) no. 7, 497–500.
481. G. Adrian, A. Bittner, *J. Coat. Technol.* **58** (1986) 59 – 65.
482. P. Kresse, *Farbe + Lack* **83** (1977) no. 2, 85–95.
483. P. Reichle, W. Funke, *Farbe + Lack* **93** (1987) no. 7, 537 – 538.
484. M. Svoboda, *Farbe + Lack* **92** (1986) no. 8, 701–703.
485. Goslarer Farbenwerke Dr. Hans Heubach, DE 3 046 698, 1980 (W. Haacke *et al.*).
486. Bayer, DE-OS 2 656 779, 1976 (F. Hund, P. Kresse).
487. Herberts GmbH, DE-OS 3 720 779, 1989 (H. Becker, F. Sadowski, W. Stephan).
488. Hoechst AG, DE-OS 2 849 712, 1978 (G. Mietens *et al.*); Hoechst AG, EP 0 028 290, 1980 (K. Hestermann *et al.*).
489. Hoechst AG, DE-OS 2 916 029, 1979 (A. Maurer *et al.*).
490. Goslarer Farbenwerke Dr. Hans Heubach, EP 0 308 884, 1988 (A. Bittner); DE-OS 3 731 737, 1987 (A. Bittner).
491. Mizusawa Kagaku Koayo K.K., DE 2 159 342, 1971 (Y. Sugahara *et al.*).
492. Pfizer & Co., US 3 443 977, 1969 (L. M. Bennetch).
493. Teikoku Kako, JP 55 160–059, 1980 (M. Kinugasa).
494. Int. Standard Electric Corp., DE-OS 3 042 630, 1980 (C. F. Drake *et al.*).
495. Toy Soda, EP 35–798, 1981 (H. Minamide).
496. Nippon Paint Co, EP 0 259 748, 1987 (T. Okal *et al.*).
497. Colores Hispania, ES 8 605–015, 1985.
498. Hooker Chemicals & Plastics Corp.: *Ferrophos*, company information, Niagara Falls, N.Y., and D. E. de Jong bv. Apeldoorn, Holland, 1976.
499. Goslarer Farbenwerke Dr. Hans Heubach, DE 3 532 806.1, 1985 (M. Gawol, G. Adrian).
500. PPG Industries, US 4 837 253, 1989 (J. D. Mansell *et al.*).
501. H. Kossmann, *Farbe + Lack* **91** (1985) no. 7, 588–594.
502. S. Hellbordt, *Farbe + Lack* **85** (1979) no. 6, 485.
503. J. D'Ans, V. Groope, *Dtsch. Farben-Z.* **15** (1961) 51, 69.
504. U. R. Evans, *J. Chem. Soc.* 1927, 1020. Links
505. J. E. D. Mayne, *J. Chem. Soc.* 1948, 1932. Links
506. J. L. Rosenfeld, *Lakokras. Mater. Ikh. Primen.* 1961, 50.
507. H. H. Uhlig, *Chem. Eng. News* **24** (1946) 3154. Links
508. G. H. Cartledge, *Corrosion (Houston)* **18** (1962) 316.
509. J. Ruf: *Korrosionsschutz durch Lacke und Pigmente*, A. W. Colomb, H. Heenemann GmbH, Stuttgart – Berlin 1972.
510. J. Ruf, *Farbe + Lack* **75** (1969) no. 10, 943.
511. GeffStoffV, TRgA 602, part 2, 1987.
512. Verordnung über gefährliche Arbeitsstoffe (Arbeitsstoffverordnung – ArbStoffV) Feb. 11, 1982 (BGBl I, 144).
513. Nippon Paint KK, JP 6 1276–861, 1985.
514. G. Lincke, O. Schroers, K. D. Nowak, *Farbe + Lack* **82** (1976) no. 11, 1003.
515. G. Meyer, *DEFAZET Dtsch. Farben Z.* **5** (1963) 201 – 205.
516. W. J. Banke, *Mod. Paint Coat.* **70** (1980) Febr., 45–47.
517. American Metal Climax, CH 602 899, 1973 (D. R. Robitaille *et al.*); DE 2 334 541, 1973 (D. R. Robitaille *et al.*).
518. Noranda Miners Ltd., US 4 132 667, 1977 (D. Kerfoot); DE 2 814 454, 1978 (D. Kerfoot).
519. J. Ruf, *Farbe + Lack* **79** (1973) no. 1, 22 – 27.
520. BASF AG, Ludwigshafen, company information, 1990.
521. Horst Michaud *et al.*, *Chem. Ztg.* **112** (1988) 287–294.
522. Amer. Cyanamide Co., US 2 213 440, 1939; US 2 213 441, 1939.
523. Duisburger Kupferhütte, DT 956 219, 1956.
524. I. G. Farbenind., DT 651 684, 1934.
525. N. Lowicki, *Z. Erzbergbau Metallhüttenwes.* **2** (1949) 282, 283.
526. S. A. Miller, B. Bann, *J. Appl. Chem.* **6** (1956) 89–93.
527. V. Kontnik, V. Novak, *Chem. Prum.* **9** (1959) 18–22.
528. Süddeutsche Kalkstickstoffwerke, DT 803 351, 1949.
529. Lab. Chim. G. Guainazzi, IT 428 147, 1947.
530. Admin. Des. Mines Bouxwiller, FR 966 349, 1948.
531. E. Stock, H. Blum, *Dtsch. Farben-Z.* **5** (1951) 317–321.
532. Th. Goldschmidt, DT 940 291, 1952.
533. Duisburger Kupferhütte, DT 819 689, 1949.
534. BASF AG, Ludwigshafen, Company information, 1990.
535. Techn. Jahresbericht 1957 der Berufsgenossenschaft der chem. Industrie der Bundesrepublik Deutschland.
536. Int. Minerals and Chem. Corp., US 4 738 720, 1986 (P. Eckler, L. Ferrara).
537. The British Petroleum Company, EP 0 316 066, 1988 (T. E. Fletcher).
538. Grace & Co., London, company information, 1986.
539. B. P. F. Goldie, *J. Oil Colour Chem. Assoc.* **71** (1988) no. 9, 257 – 270.

540. Grace & Co., *Polym. Paint Colour J.* **177** (1987) no. 4189, 260.
541. A. F. Wells: *Structural Inorganic Chemistry*, Clarendon Press, Oxford 1962.
542. J. D'Ans, W. Breckheimer, H. J. Schuster, *Werkst. Korros.* **8** (1957) 677. J. D'Ans, H. J. Schuster, *Farbe + Lack* **61** (1955) no. 10, 453. J. E. O. Mayne, *Farbe + Lack*, **76** (1970) no. 3, 243.
543. G. Lincke, *Farbe + Lack* **77** (1971) no. 5, 443. G. Lincke, *Congr. FATIPEC* **12th** 563.
544. H. Berger, *Farbe + Lack* **88** (1982) no. 3, 180–182.
545. P. N. Martini, A. Bianchini, *J. Appl. Chem.* **19** (1969) 147.
546. J. F. H. van Eijnsbergen, *Farbe + Lack* **71** (1965) no. 10, 1005. G. Meyer, *Dtsch. Farben-Z.* **16** (1962) 347.
547. Bayer AG, DE 2 642 049, 1976 (F. Hund *et al.*). Bayer AG, DE 2 625 401, 1976 (F. Hund *et al.*).
548. Tada Kogyo, US 3 904 421, 1975 (S. Shimizu *et al.*).
549. P. Kresse, *Farbe + Lack* **84** (1978) no. 3, 156–159.
550. M. Svoboda, *Farbe + Lack* **96** (1990) no. 7, 506–508.
551. Bayer: Anticor 70, product information, Leverkusen 1989.
552. Lindgens, Köln-Mülheim, company information 1991.
553. M. Leclercq, *Farbe + Lack* **97** (1991) no. 3, 207–210.
554. T. Szauer, A. Miszczyk, *ACS Symp. Ser.* **322** (1986) 229 – 233.
555. K. M. Oesterle, R. Oberholzer, *Dtsch. Farben-Z.* **18** (1964) 151.
556. J. D'Ans, H. J. Schuster, *Farbe + Lack* **63** (1957) no. 9, 430.
557. G. Grillo, *Tech. Rundsch.* **60** (1968) 19.
558. A. Laberenz, *Farbe + Lack* **68** (1962) no. 7, 765.
559. E. V. Schmid, *Farbe + Lack* **84** (1978) no. 1, 16–19.
560. M. A. Newnham, *Paint Technol.* **42** (1968) 16.
561. R. Hawner, *Farbe + Lack* **69** (1963) no. 6, 611.
562. Kärntner Montanindustrie GmbH, company information, 1991.
563. E. Carter, *Paint Resin* **55** (1985) 4, 30.
564. E. V. Schmid, *Farbe + Lack* **90** (1984) no. 9, 759–765.
565. N. Wamser, *Farbe + Lack* **96** (1990) no. 6, 435–441.
566. R. D. Athey, *J. Water Borne Coat.* **8** (1985) 1, 7 –10. 23674
567. W. Grass, H. D. Merzbach, D. Skudelny, *Farbe + Lack* **94** (1988) no. 11, 895 – 897.
568. N. Wamser, E. Urbano, *Farbe + Lack* **95** (1989) no. 2, 109, 110.
569. R. D. Athey, *J. Water Borne Coat.* **8** (1985) 2, 10 – 20.
570. Eckhard, Fürth, company information, 1989.
571. G. Cinti, A. Papo, G. Torriano, *J. Oil Colour Chem. Assoc.* **68** (1985) 29 – 34.
572. C. H. Hare, M. G. Fernald, *Mod. Paint Coat.* **74** (1984) no. 10, 138 – 146.
573. R. Besold, *Farbe + Lack* **89** (1983) no. 3, 166–173.
574. W. Funke, *Farbe + Lack* **87** (1981) no. 9, 787, 788.
575. BASF Lacke+Farben AG, DE 2 502 781, 1975.
576. R. Pantzer, *Dtsch. Farben-Z.* **19** (1975) 13. R. Pantzer, *Oberfläche-Surf.* **17** (1976) 181.
577. J. Ruf, *Congr. FATIPEC* **12th** (1974) 571.
578. Ciba-Geigy AG, EP 0 128 862, 1984 (G. Berner *et al.*).
579. Ciba-Geigy, Basel, company information, 1986.
580. Henkel KGaA, DE 2 824 508, 1979 (H. Eschwey *et al.*).
581. Henkel KGaA, Düsseldorf, company information, 1983.
582. BASF Lacke+Farben AG, DE 3 616 721, 1986; EP 246 570, 1987 (E. Liedek, G. Hägele).
583. Laboratorium für Pharmakologie und Toxikologie, Hamburg-Hausbruch 1972.
584. Institut für Toxikologie, Bayer AG, Wuppertal 1972.
585. L. M. Greenstein in T. C. Patton (ed.): *Pigment Handbook*, 2nd ed., **vol. 1**, Wiley & Sons, New York 1988, p. 829.
586. W. Bäumer, *Farbe + Lack* **79** (1973) 747.
587. K. D. Franz, R. Emmert, K. Nitta, *Kontakte (Darmstadt)* 1992, no. 2, 3.
588. K. D. Franz, H. Härtner, R. Emmert, K. Nitta, in *Ullmann*, 5th ed. **vol. A20**, pp. 347 – 355.
589. R. Glausch, M. Kieser, R. Maisch, G. Pfaff, J. Weitzel: *Perlglanzpigmente*, Curt R. Vincentz Verlag, Hannover, 1996. G. Pfaff, *Chem. unserer Zeit* **31** (1997) 6.
590. H. Simon: *The Splendor of Iridescence*, Dodd, Mead and Co., New York 1971.
591. H. F. Taylor, *Drugs, Oils, Paints* **3** (1937) 106.
592. A. H. Pfund, *J. Franklin Inst.* **183** (1917) no. 4, 453.
593. A. v. Unruh, *Kunststoffe* **8** (1918) no. 5, 49.
594. S. J. Williamson, H. Z. Cummins: *Light and Color in Nature and Art*, J. Wiley & Sons, New York 1983.
595. K. Nassau: *The Physics and Chemistry of Color*, J. Wiley & Sons, New York 1983.
596. Du Pont, US 3 087 828, 1963 (H. R. Linton). Du Pont, US 3 087 829, 1963 (H. R. Linton). Mearl, US 3 437 515, 1969 (C. A. Quinn, G. J. Rieger, R. A. Bolomey). E. Merck, US 3 553 001, 1971 (H. Kohlschütter *et al.*).
597. W. Ostertag, *Nachr. Chem. Tech. Lab.* **42** (1994) 849.
598. O. S. Heavens: *Optical Properties of Thin Solid Films*, Dover Publ., New York 1965.
599. S. Hachisu, S. Okamoto, *Sci. Light* **11** (1962) no. 3, 157.
600. H. Becker, H. Rechmann, F. Rosendahl, *Farbe + Lack* **73** (1967) 17.
601. C. Schmidt, M. Fritz, *Kontakte (Darmstadt)* 1992, no. 2, 15.
602. F. Hofmeister, *Eur. Coat. J.* **1** (1987) no. 4, 400.
603. A. M. Gaudin, *J. Phys. Chem.* **41** (1937) 811.
604. R. Emmert, *Cosm. Toiletries* **104** (1989) no. 7, 57.
605. J. A. Dobrowolski, F. C. Ho, A. Waldorf, *Appl. Opt.* **28** (1989) 2702.
606. S. Teaney, I. Denne, *Kontakte (Darmstadt)* 1992, no. 2, 46.
607. S. Hachisu, *Sci. Light* **6** (1957) no. 1, 20.
608. L. M. Greenstein, *Proc. Sci. Sect. Toilet. Goods Assoc.* **45** (1966) May, 20.
609. A. J. Krajkeman, *Paint Manuf.* **22** (1952) no. 10, 371.
610. T. Marchmay, *Sci. Amer.* **185** (1921) 533.
611. P. Alexander, *Manuf. Chem.* **57** (1986) June, 60.
612. Y. Morita, *J. Chem. Educ.* **62** (1985) 1072. Links
613. S. Hachisu, *Sci. Light* **8** (1959) no. 1, 19 and 26. S. Hachisu, S. Okamoto, *Sci. Light* **11** (1962) no. 3, 157.
614. H. A. Miller, *Bull. Bismuth Inst. (Brussels)* **52** (1987) 1.
615. K. R. Skulberg *et al.*, *Scan. J. Work Environ. Health* **11** (1985) 65.
616. B. K. Bernard *et al.*, *J. Toxicol. Environ. Health* **29** (1985) 65. Links
617. R. Esselborn, K.-D. Franz: *XVII Congress AFTPV*, Nice 1987, p. 148.
618. BASF, US 4 552 593, 1985 (W. Ostertag).
619. S. Roesch, *Z. Wiss. Mikrosk. Mikrosk. Tech.* **64** (1959) 236and *Z. Wiss. Mikrosk. Mikrosk. Tech.* **66** (1966) 136.

620. H. Kubota, *Prog. Opt.* **1** (1961) 212.
621. G. Pfaff, R. Maisch, *Farbe + Lack* **101** (1995) 89.
622. W. Ostertag, N. Mronga, P. Hauser, *Farbe + Lack* **93** (1987) 973.
623. G. Gehrenkamper, F. Hofmeister, R. Maisch, *Eur. Coat. J.* **3** (1990) 80.
624. Merck KGaA, EP 307 747, 1989 (R. Emmert, M. Weigand).
625. Merck KGaA, WO 93/19131 (K. D. Franz, K. Ambrosius, S. Wilhelm, K. Nitta).
626. Merck KGaA, EP 354 374, 1990 (K. D. Franz, K. Ambrosius, C. Prengel).
627. Merck KGaA, US 4 867 793, 1989 (K. D. Franz, K. Ambrosius, A. Knapp, D. Brückner).
628. G. Pfaff, P. Reynders. *Chem. Rundschau Jahrbuch* (1993) 31.
629. Merck KGaA, US 4 603 047, 1986 (T. Watanabe, T. Noguchi).
630. D. Brückner, R. Glausch, R. Maisch, *Farbe + Lack*, **96** (1990) 411.
631. Merck KGaA/Hyplast, WO 94/ 05727, (T. Daponte, P. Verschaeren, M. Kieser, G. Edler)
632. Bayer, EP 14 382, 1982 (F. Hund, G. Franz).
633. E. V. Carter, R. D. Laundon, *J. Oil Colour Chem. Assoc.* (1990) no. 1, 7.
634. BASF, EP 283 852, 1988 (W. Ostertag, N. Mronga).
635. W. Ostertag, N. Mronga, *Macromol. Symp.* **100** (1995) 163.
636. Optical Coating Laboratory, US 4 434 010, 1981 (G. Ash).
637. R. W. Phillips, A. F. Bleikolm, *Appl. Opt.* **35** (1996) 5529.
638. R. Glausch, G. Pfaff, R. Maisch, *XXIst FATIPEC Congr.*, Amsterdam, Vol. 2, 1992, p.33.
639. Merck KGaA, WO 93/08237, G. Bauer, K. Osterried, C. Schmidt, R. Vogt, H. Kniess, M. Uhlig
640. W. Ostertag, *XIX FATIPEC Congr. Book*, **vol. 1**, Aachen 1988, p. 103.
641. Ciba-Geigy, US 4 517 320, 1985 (F. Bäbler, A. Weissmüller).
642. Ciba-Geigy, EP 408 498, 1991 (P. Bugnon, F. Herren, B. Medinger).
643. Mearl, EP 401 141, 1990 (B. R. Swallow).
644. P. Eskelinen: *Prepr. IFSCCJ Congress*, Yokohama, Vol. 1, 1992, p. 196.
645. V. P. S. Judin, *Polym. Paint Coat. J.* **182** (1992) 4312.
646. BASF, US 4 753 829, 1988 (S. Panush)
647. A. P. Philipse, *J. Mat. Sci. Lett.* **8** (1989) 1371.
648. O. Mayborodoff, *Double Liaison-Chim. Peint.* **308** (1981) p. 36.
649. R. Maisch, G. Houseman, K. D. Franz, *Polym. Paint Colour J.* **178** (1988) 100.
650. K. Dorfner, J. Ohngemach, *Plast. Eng.* **43** (1987) Jan., 33. Links
651. G. Houseman, *J. Oil Colour Chem. Assoc.* (1987) no. 11, 329.
652. S. Panush, *J. Paint Technol.* **45** (1981) 39. S. Panush, *J. Coat. Technol.* **57** (1985) 67.
653. Mearl, US 4 134 776, 1979 (C. J. Rieger, L. Armanini).
654. E. Merck, US 4 482 389, 1984 (K. D. Franz *et al.*).
655. E. Merck, US 4 828 623, 1989 (K. Nitta, T. Watanabe, I. Suzuki).
656. E. Merck, US 5 022 923, 1991 (A. Rau, K. D. Franz, K. Ambrosius).
657. A. Thurn-Müller, J. Hollenberg, J. Liston, *Kontakte (Darmstadt)* (1992) no. 2, 35.
658. G. Möschl, F. K. Soehngen, M. Kieser, *Seifen Öle Fette Wachse* **106** (1980) 93; *Seifen Öle Fette Wachse* **111** (1985) 643; *Seifen Öle Fette Wachse* **112** (1986) 45.
659. L. Armanini, *Cosm. Toiletries* **106** (1991) no. 3, 53.
660. M. L. Schlossman, *Cosm. Toiletries* **105** (1990) no. 2, 53.
661. E. Merck, EP 406 657, 1990 (A. Thurn-Müller, J. Dietz, C. Prengel).
662. R. Besold, R. Reißer, E. Roth, *Farbe + Lack* **97** (1991) 311.
663. W. Ostertag, *Congr. FATIPEC XIX*, **vol II/103**, 1988.
664. US 2 558 302, 1951 (G. C. Marcot *et al.*).
665. US 2 558 304, 1951 (G. C. Marcot *et al.*).
666. Bayer, DE-OS 2 508 932, 1975 (F. Hund, G. Linde).
667. Bayer, DE 2 210 279, 1973 (F. L. Ebenhöch, K.-P. Hansen, H. Stark).
668. BASF, DE 2 344 196, 1973 (F. L. Ebenhöch, D. Werner, G. Bock).
669. BASF F+F AG, DE-OS 2 228 555, 1972 (H. Gaedcke, R. Bauer).
670. Magnetic Pigment Comp., US 1 424 635, 1922 (P. Fireman). Reichard-Coulston Inc., US 2 574 459, 1947 (L. H. Bennetch).
671. F. Finus, *Farbe + Lack* **7** (1975) 604 – 607.
672. G. Narvuglio, R. F. Sharrock, R. J. Kennedy, *Oil J., Col. Chem. Assoc.* **61** (1978) 79 – 85.
673. BASF, DE 2 840 870, 1978 (A. Seitz).
674. Tokyo Shibaura Denkikbushiki Kaisha EP 0 019 710, 1980.
675. V. P. S. Judin, V. T. Salonen, *Seifen Oele Fette Wachse* **119** (1993) 491.
676. D. R. Robertson, F. Gaw, *Congr. Add. '95*, Paper 12.
677. W. H. Kettler, G. Richter, *Farbe + Lack* **98** (1992) no. 2, 93.
678. BASF, US 932 741, 1986 (S. Punush).
679. M. A. Vannice, R. L. Garten, *J. Catal.* **63** (1980) 255. Links
680. Huels, DE-OS 3 010 710, 1984 (K. Neubold, K. D. Gollner).
681. Mitsubishi Materials, JP 07 062 326.
682. Degussa, EP 609 533, 1994 (W. Hartmann, D. Kerner).
683. Nippon Shokubai, JP 07 232 919, 1994.
684. L. Brüggermann, DE 3 900 243, 1993 (G. Walde, A. Rudy).

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Pigments, Organic

Phthalocyanines is a separate keyword.

KLAUS HUNGER, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

WILLY HERBST, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Introduction 2
2. Physical and Technical Properties 3
2.1. General Aspects 3
2.2. Dispersibility 4
2.2.1. Pigment Powders 4
2.2.2. Pigment Preparations 5
2.3. Crystal Modifications 5
3. Uses 5
3.1. Printing Inks 5
3.2. Paints and Coatings 6
3.3. Plastics 7
3.4. Other Uses 8
4. Azo Pigments 8
4.1. Production 9
4.1.1. Diazotization 9
4.1.2. Coupling 10
4.1.3. Special Features of Azo Pigment Synthesis 11
4.2. Yellow and Orange Monoazo Pigments 13
4.2.1. Yellow Monoazo Pigments (Hansa Yellow Pigments) 13
4.2.2. Monoazopyrazolone Pigments 15
4.3. Disazo Pigments 15
4.3.1. Diarylide Pigments 15
4.3.2. Bis(*N*-acetoacetarylide) Pigments . 17
4.4. β-Naphthol Pigments 17
4.5. Azo Pigment Lakes 18
4.5.1. β-Naphthol Pigment Lakes 18
4.5.2. 2-Hydroxy-3-naphthoic Acid Pigment Lakes (BON Pigments) . . 19
4.5.3. Other Pigment Lakes 20
4.6. Naphtol AS Pigments 20
4.7. Benzimidazolone Pigments 22
4.8. Disazo Condensation Pigments . . 23
4.8.1. Yellow Pigments 23
4.8.2. Orange, Red, and Brown Pigments . 24
4.8.3. Properties 25
5. Metal-Complex Pigments 25
5.1. Azo Metal-Complex Pigments . . . 26
5.2. Azomethine Metal-Complex Pigments 27
5.3. Miscellaneous Metal-Complex Pigments 27
6. Isoindolinone and Isoindoline Pigments 28
6.1. Azomethine Type: Tetra chloroisoindolinone Pigments . . . 28
6.2. Methine Type: Isoindoline Pigments 29
7. Phthalocyanine Pigments 29
8. Quinacridone Pigments 29
8.1. Production 30
8.1.1. Synthesis 30
8.1.2. Aftertreatment 31
8.2. Properties 31
9. Perinone Pigments 32
10. Perylene Pigments 33
11. Anthraquinone Pigments 34
11.1. Hydroxyanthraquinone Pigments 34
11.2. Aminoanthraquinone Pigments . . 35
11.3. Heterocyclic Anthraquinone Pigments 36
11.3.1. Anthrapyrimidine Pigments 36
11.3.2. Indanthrone and Flavanthrone 36
11.4. Polycarbocyclic Anthraquinone Pigments 37
11.4.1. Pyranthrone Pigments 37
11.4.2. Anthanthrone Pigments 38
12. Diketopyrrolopyrrole (DPP) Pigments 39
13. Thioindigo Pigments 39
14. Dioxazine Pigments 40
15. Triphenylmethane Pigments 41
15.1. Free Sulfonic Acids ("Alkali Blue" Types) 41
15.2. Dye Salts with Complex Anions . . 41
16. Quinophthalone Pigments 42
17. Fluorescent Pigments 43
18. Environmental Aspects 44
18.1. Toxicology 44
18.2. Ecology 45
18.3. Impurities 45
19. References 45

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_371

1. Introduction [1–11]

The increased consumption of colored organic pigments in recent years far exceeds the increased turnover in the industrial sectors in which they are used. The most important reasons for the disproportionately large increase in pigment consumption include advertising and sales promotion, as well as the identification and differentiation of products in the industrial and domestic sectors. Reliable figures for the production of organic coloring agents are only available for certain countries. In the early 1980s, pigments were estimated to account for ca. 40 % of the overall production of organic coloring agents.

The largest area of use of organic pigments is in graphic printing inks, which consumes 50 % of organic pigment production. Almost 25 % of the organic pigments are used in paints and coatings. The coloring of plastics, including spin dyeing of synthetic fibers, consumes less than 20 % and is also an important area of use. The remaining organic pigments are used in textile printing and a number of smaller sectors, including contactless printing processes, office articles and accessories (e.g., colored pencils, crayons, chalks, and typewriter ribbons), and the coloring of wood, cosmetics, and paper.

Definition. Pigments are defined as inorganic or organic, chromatic or achromatic coloring agents that are practically insoluble in the application medium (DIN 55 943). Dyes are coloring agents that are soluble in the application medium. See also → Pigments, Inorganic

In coloring , the crystalline pigment is applied in the solid state, not in dissolved form, to the medium being colored. Both the chemical and the physical properties of pigments (e.g., particle size, particle-size distribution, special types of surface and specific surface area, crystal modification, and crystal form) are important for their industrial applications.

Pigment fractions may, however, also be soluble, depending on the composition of the system to be colored and processing conditions. Important properties of the pigmented system, especially migration, recrystallization, thermal stability, lightfastness, weather resistance, and coloristic properties, are often decisively influenced by the dissolved pigment fractions.

Many organic pigments and dyes have the same basic chemical structure. The insolubility required in pigments can be obtained by excluding solubilizing groups, by forming insoluble salts (lake formation) of carboxylic or sulfonic acids, by metal- complex formation in compounds without solubilizing groups, and particularly by incorporating groups that reduce solubility (e.g., amide groups).

Classification. Organic pigments can be classified in various ways, the two main criteria being chemical structure and coloristic features. A strict subdivision according to one of these criteria is not possible because of overlap. In this article, pigments are classified largely on the basis of chemical structure. They are broadly divided into azo pigments and nonazo or polycyclic pigments. Azo pigments are by far the largest group, not least because they are generally easy to produce and therefore more economical to use. Polycyclic pigments are synthesized according to many widely varying principles. Both groups of pigments can be further subdivided according to their chemical properties and uses. Characteristic properties and most important uses are given for each of the pigment classes described. Examples have been selected on the basis of technical and industrial importance.

Table 1 lists the trade names and producers of important ranges of organic pigments.

2. Physical and Technical Properties

2.1. General Aspects

Organic pigments have to satisfy requirements that may differ considerably depending on the area of use. These requirements relate not only to obvious coloristic properties such as hue, cleanness or brightness, and tinting strength, but also to many other technical properties that are important for trouble-free application of the pigment or necessary to ensure satisfactory use in the colored medium. Such properties include lightfastness, weather stability, and rheological properties. Pigments must also be resistant to organic solvents that may be present during processing and application. Other components of the colored medium (e.g., resins, plasticizers,

or stabilizers) must not adversely affect coloristic properties of the pigments; attention must also be paid to other fastness properties such as blooming, bleeding, or recrystallization.

Table 1. Producers and trade names of organic pigments

Producer	Trade name
BASF	Azalean, Fanal, Heliogen, Lithol, Macatawa, Paliogen, Paliotol, Saugatuck, Segnale, Sico, Sico Echt, Termosolido
Bayer	Bayplast, Fanchon, Helio, Indofast, Perrindo, Quindo, Thio Fast
Cappelle	Acetanil, Bonitol, Diacetanil
CDF-Chimie	Langdocyal
Ciba-Geigy	Cinquasia, Cromophtal, Horna, Irgalite, Irgazin
Cookson	Dalamar, Heucophthal, Monastral, Naphtanil
Dainichi Seika	Chromofine, Seika Fast
Dainippon Ink	Fastogen, Symuler, Symuler Fast
European Colour plc	Eljon
Hoechst	Hansa, Hostaperm, Novoperm, Permanent, PV, PV-Echt, Reflex, Vulcan, Vulcan-Echt
ICI	Lutetia, Monastral, Monolite, Polymon, Vynamon
Køge	Isol
Sandoz	Graphtol
Sansui Colour	Suimei
Sanyo	Sanyo, Sanyo Fast
Sudarshan	Sudarperm
Sumitomo	Sumicacoat, Sumitone, Sumitone Fast
SUN Chemical	Hancock, Montclair, Muskegon, Radiant, Rangoon
Tosoh Corp.	Rubicron
Toyo Ink	Lionogen, Lionol

The above features are determined not only by the chemical constitution of the pigment but also by its physical properties. The particle size and particle-size distribution of the pigment in the application medium play the most important role in this regard. Decreasing particle size (i.e., increasing comminution) generally leads to

1) Increase in the tinting strength in systems in which the intrinsic light scattering of the pigment can be ignored (e.g., in white reductions)
2) Change of shade and cleanness
3) Increase in transparency and thus a decrease in hiding power
4) Increase in gloss (e.g., in printing inks and coatings)
5) Increase in viscosity and deterioration of flow properties of the pigmented media
6) Reduction of lightfastness and weather stability
7) Deterioration of migration fastnesses
8) Lower resistance to solvents
9) Deterioration of recrystallization stability

Optimization of a specific pigment property is thereby connected with a change in other technical properties. Thus, a pigment developed for a specific field of application always represents a compromise between many coloristic and technical properties. For this reason, several commercial grades of a pigment with different particle-size distributions are generally marketed by producers.

The technical and fastness properties always refer to the complete pigment system, not to the pigment itself. For example, the lightfastness of a given pigment may be outstanding in one medium (corresponding to a figure of 8 on the eight-point wool scale), whereas in another medium it may correspond only to a figure of 2 on the same scale.

2.2. Dispersibility

2.2.1. Pigment Powders

During the production process, pigments associate to form agglomerates and aggregates. When pigments are incorporated in the application medium, these structures must be broken down (i.e., dispersed) as much as possible. Nearly all the above-mentioned, particle-size-dependent application and fastness properties and the economic effectiveness of the pigments are affected by the efficiency of dispersion. Unsatisfactory dispersion of the agglomerates may lead to processing problems, including pitting in the deep drawing of plastic sheets or thread breakage in melt-spun synthetic fibers.

Pigment agglomerates resist disintegration. The ease with which disintegration is effected is called the dispersibility and is one of the most important technical properties of pigments. The rate at which dispersion proceeds is a measure of dispersibility.

The dispersion process itself consists of a number of steps that generally occur in parallel.

These steps include breaking down agglomerates in the pigment powder into smaller particles (smaller agglomerates, aggregates, and individual, e.g., primary particles) [12]. The pigment surface is wetted by the binder constituents, and the particles are distributed uniformly in the entire medium. The resultant pigment dispersion must be stabilized (i.e., particles must not recombine by flocculation, reagglomeration, or other processes).

Depending on the dispersion system (i.e., type of pigment, binder, equipment, and processing conditions), one or another of these stages may occur with difficulty. Whereas wetting of the pigment surface is of only minor importance in coloring thermoplastic materials [e.g., rigid poly(vinyl chloride) PVC], it may be crucial when pigments are dispersed in plastisols (coating pastes) based on the same polymer.

Pigments for special areas of application can be optimized with regard to dispersibility. However, in practice a satisfactory degree of pigment dispersion is sometimes impossible or uneconomical to achieve.

2.2.2. Pigment Preparations

Preparations containing the pigment in an already dispersed form are an alternative to pigment powders. Dispersion is performed not by the pigment user but by the pigment producer or in special plants. Pigment preparations contain a carrier material that is identical to or compatible with the application medium, dispersing agents, etc. and depends on the intended use.

Treatment of pigments with a variety of compounds allows modification of dispersibility and other technical properties. One example is the "resination" of azo pigments. The presence of rosin or other hard resins during azo coupling increases the rate of nucleation and decreases the rate of crystal growth. These processes yield smaller pigment particles with a high transparency that are very important in, for example, multicolor printing.

Other agents produce the desired effect by chemical reaction with the base pigment. An example of this kind of surface- treated pigments is the reaction of diarylide yellow pigments with aliphatic amines to form ketimines and enolates. These derivatives are partially soluble in toluene and aliphatic hydrocarbons (the solvents used in publication gravure printing inks) and thus substantially reduce the viscosity of these media. The development of such pigments permitted the use of modern, high-speed gravure printing presses.

2.3. Crystal Modifications

In a large number of pigments, the molecules in the crystal can be arranged in different ways to form different crystal modifications. These modifications vary widely in coloristic and fastness properties. Unstable modifications are often transformed into thermodynamically stable forms, especially during heating or dispersion in the presence of certain solvents or other compounds. Even in pigment forms with similar coloristic properties, application and fastness properties (e.g., lightfastness and durability) can be considerably altered by these modifications. This should be taken into consideration when such pigments are used in practice.

Examples of the variety of requirements that pigments must fulfill for their most important areas of application are given in Chapter 3.

3. Uses

3.1. Printing Inks

Printing inks are discussed in detail elsewhere (→ Printing Inks). The layers formed with the printing inks are generally extremely thin (e.g., 0.7 – 1.2 μm thick in offset printing) and therefore necessitate the use of pigments with a very high tinting strength. A high pigment concentration is required in the ink, and this can lead to rheological and technical problems during printing, particularly for highly transparent pigments with a small particle size and high specific surface area. The latter is especially problematic in modern, high-speed printing presses. These problems can be reduced or eliminated by chemical modification or physical optimization of the pigments (e.g., with regard to particle-size distribution). The high operating speeds of modern printing presses also require fast-drying inks with good adhesion that allow the printed products to be stacked without deterioration of print

quality. As a result, the pigments and the compounds used for their surface coating must not adversely affect solvent release from the printed ink layers, a requirement that is not always easy to satisfy.

In both offset and letterpress printing, color scales are used in which narrow ranges of hues are specified on the basis of color measurement indices for the basic shades yellow, magenta, and blue [blue-green, e.g., in the European Standard (DIN 16 538, 16 539)]. This is intended to enable reprographic companies and printers to obtain printing inks that produce standard results under standard conditions.

In solvent-based printing inks, pigments must be extremely resistant to the solvents used in the ink during processing, particularly during dispersion, but also during transportation and storage of the inks. Toluene or aliphatic hydrocarbons are used in publication gravure printing inks, whereas ketones and esters are employed in packaging gravure printing inks. Unsatisfactory solvent resistance reduces the tinting strength and changes the shade due to recrystallization and also produces a generally undesirable, reduced transparency in the print.

In certain applications, pigments must also be resistant to the material being printed; one example is the fastness to plasticizers when plasticized poly(vinyl chloride) sheets are being printed. Other properties required in practice include fastness to overcoating, sterilization, and calendering; heat sealability; heat stability; and fastness to solvents or consumer goods (e.g., margarine, butter, or soap).

Fastness requirements are particularly stringent for pigments used in decorative papers that are printed by the gravure or flexographic methods and are used to produce laminated sheets. In addition to good lightfastness in the laminated sheet, the type of resin (melamine or polyester) is particularly important in the selection of pigment. For example, in sheets that are produced from polyesters dissolved in styrene monomer, the pigment must be highly resistant to polyesters and styrene.

3.2. Paints and Coatings

The desired coloristic properties and fastness properties of pigments in paints are governed primarily by requirements placed on the articles to which they are applied. Pigments must also satisfy requirements that are determined by the composition of the application medium, the type and concentration of solvents contained in the latter, and processing conditions.

Because air-drying, medium oil, or long oil alkyd resins generally contain only mineral spirit, aromatic hydrocarbons, and higher alcohols (i.e., solvents that are less aggressive toward pigments), the pigments for paints based on these binder systems can usually be chosen without any restrictions. The situation is different however in the case of stoving (baking) enamels (quite often systems based on melamine – formaldehyde or urea – formaldehyde resins), which undergo polycondensation with other resins (e.g., epoxy, short oil alkyd, or acrylic resins) and harden when baked in an oven. Such systems are used on a large scale in industrial finishes. The paints contain aromatic hydrocarbons (in particular xylene), mineral spirit, alcohols, glycols, esters, ketones, and many other solvents. The fastness properties that depend on pigment solubility (e.g., recrystallization or migration) are largely determined by the particle-size distribution, the processing temperature, and the pigment concentration. The solubility of pigments in the application medium increases with increasing temperature, as does the tendency to bleed into a coat of white paint or paint of another color applied on top of the paint film. For example, when a pigment is baked at 120 °C, bleeding into a white coating does not occur, whereas at 140 or 160 °C the coating is strongly colored. The composition of the paint, the type of solvent used, and the baking conditions are important in this connection.

Blooming of a pigment is another defect that depends on pigment solubility. Here, the pigment gradually crystallizes on the surface of a coating or plastic. The resulting layer of pigment crystals can be removed from the surface of the colored medium by rubbing. Blooming continues, however, and only ceases when a substantial proportion of the pigment has migrated from the pigmented system and crystallized on the surface. The prerequisite for blooming is a supersaturated solution of pigment in the application medium. Such a solution is obtained if

1) The pigmented system is subjected to relatively high baking temperature
2) The pigment dissolves partially or completely at the application temperature, and the amount of dissolved pigment exceeds its solubility at the storage or ambient temperature
3) At least partial supersaturation still exists on cooling.

The field of oven-drying systems has undergone extensive technical development, particularly because of environmental protection requirements and solvent emission. For example, thermoplastic acrylic resin systems contain 75 – 85 vol % solvent in the state of application. The high solvent emissions associated with this industrial paint sector have forced the paint industry to develop systems with lower solvent content. Low-solvent systems such as high-solids finishes (→ Paints and Coatings, Chap. 3.2.), nonaqueous dispersions (→ Paints and Coatings, Chap. 3.6.), waterborne systems (→ Paints and Coatings, Chap. 3.3.→ Paints and Coatings, Chap. 3.5.), powder coatings (→ Paints and Coatings, Chap. 3.4.), and systems cured with UV light or electron beams (→ Paints and Coatings, Chap. 3.7.) have been developed . Each of these systems places special requirements on the pigments used. This is especially true for powder coatings, where systems based on epoxy resins are becoming increasingly important (→ Paints and Coatings, Chap. 3.4.). Powder coatings are produced in a multistage process and are applied by electrostatic spraying or fluidized-bed procedures. On subsequent baking at ca. 180 – 200 °C the binder components melt and sinter, before cross-linking. The rheological properties of the melt and viscosity changes often lead to unsatisfactory wetting of the pigment surface. This results in technical problems (plate-out): after hardening, a surface deposit containing unwetted pigment forms on the surface of the colored paint film. The coating can be wiped off by vigorous rubbing and does not reform, in contrast to blooming. The shade of powder coatings often changes, in some cases considerably, as a result of chemical reactions between pigment and curing agent. These reactions depend on the conditions, temperature, and time used for processing and curing. Fastness properties (e.g., lightfastness) can also be affected. Changes of hue are not restricted to specific classes of pigments but occur even in pigments that otherwise have a high fastness level (e.g., quinacridones and dibromoanthanthrone). These difficulties can be reduced or eliminated by using appropriate combinations of pigment and curing systems.

In the paint sector a great deal of interest exists in pigments that may be regarded as extreme cases in relation to their physical properties, e.g., particle size. Such organic pigments include those with high transparency or high hiding power. The first group consists of very small particles and is used mainly for metallic finishes on automobiles. The pigments must have good rheological properties, high lightfastness, high durability, outstanding stability to solvents used in the paint systems, and thus also stability to bleeding and blooming. These requirements are often difficult to satisfy because of the relationships between technical properties and particle size.

Pigments with optimal hiding power are obtained by special finishing processes (e.g., by heating in solvents or water, if necessary under pressure). The pigments consist of very coarse particles: their specific surface area is 10 – 25 m^2/g, in contrast to conventionally treated grades, which have a specific surface area of 50 – 100 m^2/g. The mean particle size calculated from the particle-size distribution is ca. 0.2 – 0.5 µm, compared with 0.03 – 0.2 µm for conventional pigments. The considerably improved rheological behavior of paints produced with such opaque pigments enables the pigment concentration and thus the opacity to be increased further.

3.3. Plastics

In the plastics field, required pigment properties are largely determined by the type of plastic being colored. Organic pigments are difficult to disperse in plastics. The need for more economical plastics processing has led to the development of faster-output machines with higher processing temperatures and lower residence times. The conditions for pigment dispersion are thus particularly unfavorable. Although an unsatisfactory degree of dispersion rarely results in production faults or malfunctions, it does re-

duce process economy. For this reason, pigment preparations in which the pigment is already present largely in dispersed form are generally used to color plastics.

In some polymers (e.g., plasticized PVC), requirements concerning fastness to migration are of paramount importance and are concerned primarily with bleeding and blooming of the pigments. On bleeding, the pigment first dissolves in the application medium and then migrates to an identical or similar medium that is in contact with it. In the case of plasticized PVC, bleeding is attributed to solubility of the pigment in the plasticizer. Accordingly, the bleeding tendency and degree depend on the nature and concentration of the plasticizer and the pigment. Similar considerations also apply to blooming (see Section 3.2), which increases with processing temperature. Limiting concentrations for pigments in a given system are often known, above or below which blooming occurs. As with baking enamels, this concentration range also increases with increasing temperature.

Pigments that bloom always bleed. Pigments, that bleed do not, however, necessarily bloom. The chemical composition of pigments is also a decisive factor in migration (i.e., bleeding and blooming). Attempts are being made to produce migration-fast pigments by increasing molecular size, converting molecules containing acid groups into the corresponding metal salts (so-called pigment lakes or introducing functional groups (e.g., carbonamide groups).

Pigments must be extremely thermally stable in certain plastics. Plastics (e.g., polypropylene and polycarbonate) may be processed at temperatures as high as 320 °C for a few minutes. Lack of heat stability of a pigment or pigmented system is generally associated with a change in hue, but in many cases also with a change in fastness. Poor heat stability may be due to thermal decomposition of the pigment, dissolution of the pigment in the application medium, or changes in the physical properties of the pigment (e.g., particle-size distribution or crystal modification).

Migration does not occur in plastics whose glass transition temperature far exceeds room temperature (e.g., polystyrene or polymethacrylates). The pigments cannot crystallize or bloom even from supersaturated dye solutions. Pigments that are otherwise regarded as being only poorly fast to solvents and that migrate into other polymers may therefore be used for such plastics.

Requirements placed on pigments used for the spin dyeing of synthetic fibers are similar in many respects to those used for coloring plastics. In contrast to textile dyeing methods, in spin dyeing the material (polymer) is colored before the fibers are produced. Lightfastness, heat stability, and fastness to solvents are thus the most important requirements. In all spin dyeing methods the requirements for good dispersibility are particularly stringent. The particle or agglomerate size of the pigments must not exceed ca. 2 μm (i.e., one-tenth of the fiber diameter). Coarser particles decrease the tensile strength of the fiber and frequently cause the fiber to break, particularly during stretching. Because ensuring a sufficient degree of dispersion is very difficult in practice, the use of pigment preparations is often unavoidable in spin dyeing.

In dyeing polyamide fibers, chemical stability of the pigments with respect to the strongly reducing medium of the polyamide melt is also necessary during the polymerization process. Some polycyclic pigments such as phthalocyanine blue, phthalocyanine green, or quinacridones satisfy these requirements.

3.4. Other Uses

Organic pigments are used to color a large number of materials other than those described in Sections 3.13.23.3. The requirements are generally concerned with the solvent resistance necessary for production and processing, and the fastness properties, in particular lightfastness, needed in the end product.

Uses include coloring wood, paper, office articles and accessories, and artists' colors and drawing materials (e.g., colored pencils, wax crayons, chalk, and watercolors), as well as coloring cosmetics, mainly soap.

Textile printing is another important application of organic pigments and is normally considered separately from graphic printing. Pigment requirements in this sector are mainly governed by the end product – the printed textile – and its desired fastness properties (→ Textile Printing).

4. Azo Pigments

Azo pigments contain one or more azo groups −N=N− as the chromophore (**1**):

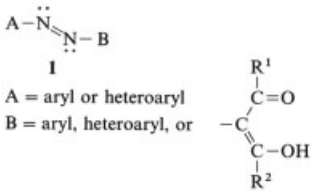

Residues A and B must not contain pigment-solubilizing groups, such as $-SO_3H$ or longer-chain alkyl groups (see Section 4.5), however.

The pigments are termed monoazo, disazo, trisazo, etc., depending on the number of azo groups per molecule. However, it is almost exclusively the monoazo and disazo pigments that are industrially important. The nomenclature of azo pigments is the same as that of azo dyes (→ Azo Dyes).

Although azo pigments can, in principle, be produced in all colors, yellow, orange, red, and brown shades are the most important in practice. Blue and green azo pigments are of little industrial importance, these shades being produced mainly with high-fastness and cheaper phthalocyanine pigments (see Chap. 7 and → Phthalocyanines).

4.1. Production

Although the azo group can be synthesized in many ways, very few methods apart from diazotization and coupling are important (→ Azo Dyes, Chap. 2.). They are used only where azo coupling is impossible (e.g., if the starting compounds are difficult to obtain).

Almost all azo pigments are therefore produced industrially by azo coupling in which an aromatic amine is diazotized and coupled with a suitable component. A relatively small number of diazo and coupling components are combined in a multiplicity of ways to form the basis for all industrially important azo pigments.

4.1.1. Diazotization

Diazotization, discovered by PETER GRIES in 1862, is the reaction of a primary aromatic amine with sodium nitrite in mineral acid solution, in which the amine is converted into the diazonium salt:

$$ArNH_2 + 2\,HY + NaNO_2 \longrightarrow Ar{-}N{\equiv}N^+Y^- + NaY + 2\,H_2O$$

Ar = aryl; Y = Cl, HSO_4

In industrial diazotization, one equivalent of an aromatic amine is generally dissolved in 2.5 – 3 equivalents of hydrochloric or sulfuric acid and then reacted at 0 – 5 °C with one equivalent of aqueous sodium nitrite solution. Weakly basic amines require larger amounts of more concentrated acid. Often such diazotizations can only be carried out with nitrosylsulfuric acid (HSO_4NO) in sulfuric acid.

Details of the mechanism and types of diazotization can be found elsewhere (→ Azo Dyes, Chap. 2.1.1.).

Aromatic diamines can be doubly diazotized (bisdiazotization). The term tetrazotization, although still used for this reaction, is incorrect and should not be employed.

Substituted anilines are generally the most suitable diazo components for azo pigments. The substituents confer many important properties on pigment fastness such as hue, solvent and migration lightfastness, and weather resistance. Important substituents include chlorine, methyl, methoxy, ethoxy, nitro, trifluoromethyl, carboxy, and carbalkoxy groups. Diazo components for producing azo pigments can be subdivided into five main groups.

1) Substituted aniline derivatives. Examples:

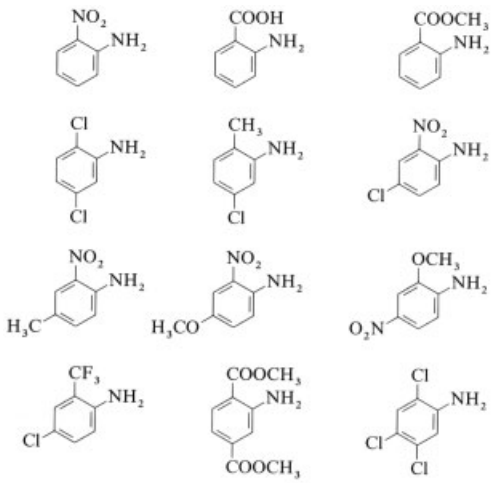

2) Aminobenzamides, aminoanilides, and aminobenzene sulfonamides. Examples:

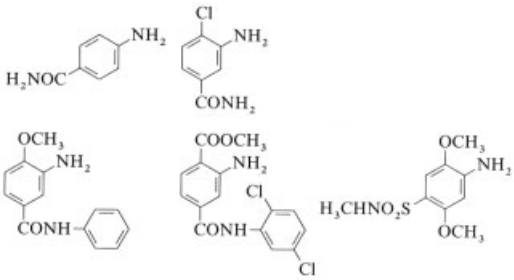

3) Aromatic diamines. Examples:

4) Polycyclic amines. Example:

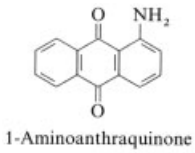

5) Heterocyclic amines and diamines. Examples:

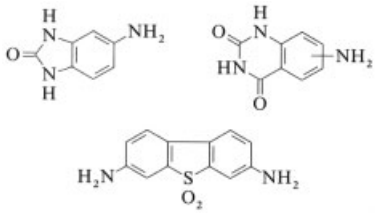

Diazonium compounds are often stable only in the cold. Heat, light, and heavy-metal ions accelerate the decomposition of diazonium compounds. Diazotizations are therefore generally performed in rubberized iron vessels, vessels lined with an acid-resistant cladding, or in stirred tanks made of glass-fiber-reinforced polyester, but they may also be carried out in wooden vats.

Because many diazonium compounds are explosive in the solid state, synthesis of the diazonium salt is followed as quickly as possible by coupling in an aqueous medium without intermediate separation of the product.

4.1.2. Coupling

Coupling is the electrophilic substitution of the diazo group for a hydrogen atom on the coupling component to form an azo bond:

$$Ar{-}N{\equiv}N^{+}\ Y^{-} + H{-}R \longrightarrow Ar{-}N{=}N{-}R + HY$$

HR = coupling component

Because free acid is formed in the coupling reaction, the optimum pH for coupling must be maintained by the addition of base. If coupling is performed with amines the optimum pH is 4 – 7, with phenols or naphthols 7 – 9.

General reaction conditions, along with the influences of substituents of the components in azo coupling, are described in detail elsewhere (→ Azo Dyes, Chap. 2.1.2.).

Coupling components for azo pigments are generally sparingly soluble in water and dissolve in strong bases as the naphtholates or enolates. The coupling components can be divided into *N*-acetoacetyl derivatives of aromatic or heterocyclic amines, hydroxynaphthalene derivatives, heterocyclic coupling components, and corresponding bifunctional components. In the following examples the coupling sites are identified by arrows.

***N*-Acetoacetyl derivatives of aromatic or heterocyclic monoamines** (**2**) are represented by the general formula:

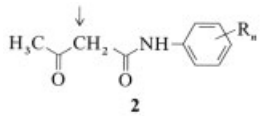

2

R = H, Cl, CH_3, OCH_3, OC_2H_5, NO_2, $NHCOCH_3$, or an annealed heterocyclic ring

$n = 1 - 3$

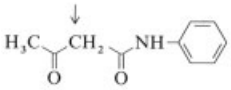

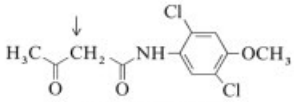

Naphtol AS-IRG

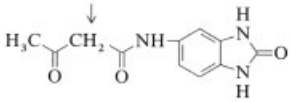

Examples: (see above)

2-Hydroxynaphthalene derivatives, particularly 2-hydroxy-3-naphthoic acid (BON), and their arylamides.
Examples:

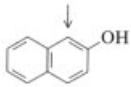

β-Naphthol

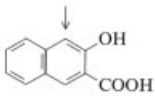

2-Hydroxynaphthoic acid (BON)

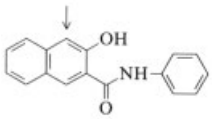

Naphtol AS

Naphtol AS is a trade name registered by Hoechst AG. Analogous pigments from other companies are designated Naphtol AS

Variation of the substituents in the arylamide residue and, to a lesser extent, in the 2-hydroxy-3-naphthoic acid residue of Naphtol AS gives a large number of derivatives that are used as coupling components in azo pigment synthesis, but also play an important role in Naphtol AS developing dyes (→ Azo Dyes, Chap. 6.).
Examples:

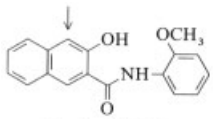

Naphtol AS-OL

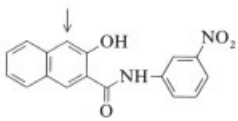

Naphtol AS-BS

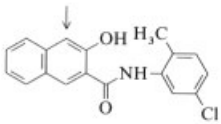

Naphtol AS-KB

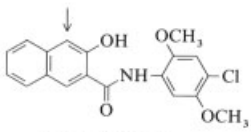

Naphtol AS-LC

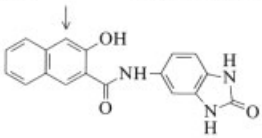

Heterocyclic Coupling Components Examples:

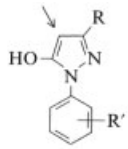

5-Pyrazolone derivatives
R = CH_3, $COOCH_3$, $COOC_2H_5$
R′ = H, Cl, CH_3

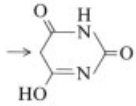

Barbituric acid

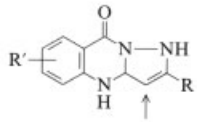

R = H, CH_3 R′ = H, Cl

Bifunctional Coupling Components Examples:

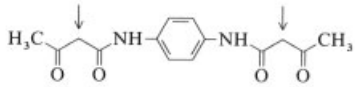

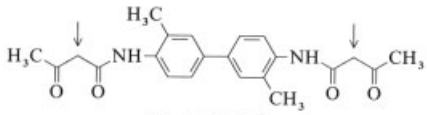

Naphtol AS-G

4.1.3. Special Features of Azo Pigment Synthesis

Whereas azo dyes generally must be salted out from the aqueous reaction solution after synthesis, azo pigments precipitate during synthesis as almost insoluble substances and are thus virtually impossible to purify subsequently. Particular attention must therefore be paid to the quality of the starting products used for the synthesis of azo pigments.

After diazotization, amines often form water-soluble diazonium salts, whereas the coupling components used for azo pigments are almost insoluble in water. Although they generally dissolve in alkali, the pH value is then too high for coupling. The alkali solution of the coupling component is therefore normally reprecipitated in the cold with acid (e.g., acetic or hydrochloric acid), if necessary with the addition of a dispersing agent. The coupling component thus becomes very finely suspended in water and can couple readily and smoothly with the dissolved diazonium salt.

Acid formed during azo coupling can be buffered by the sodium acetate formed on precipitation of the coupling component (see above). Otherwise dilute alkali must be added continuously during coupling to maintain a constant pH. Sodium phosphates, calcium carbonate, or magnesium oxide can be added as a buffer instead of sodium acetate.

The outcome of azo pigment synthesis may be influenced by the following parameters:

1) Type of coupling (i.e., the sequence and rate of addition of components)
2) Concentration of reactants
3) Temperature of the reaction mixture
4) Process parameters: shape and size of the reaction vessel, stirrer, design and stirring rate

Extremely small pigment particles usually remain after coupling. The drying of pressed pigment cakes after filtration and washing generally results in association of the pigment crystallites because of their high surface energy. In many cases the resulting agglomerates and aggregates cannot be satisfactorily comminuted even after intensive grinding. Optimum comminution is, however, a prerequisite for the use of a pigment, for both technical and economic reasons.

Azo coupling is therefore followed by "preparation" or thermal after treatment ("finishing") of the pigment. In the former, the pigment particles are coated with various substances (e.g., resins or plastic dispersions) during or after synthesis to prevent, or at least counteract, their reagglomeration.

With thermal after treatment, the pigment suspension may simply be heated for a few hours at 60 – 100 °C or higher. Often, however, the pressed filter cake is first washed, reformed into a paste, and then aftertreated at 60 – 100 °C. Very sparingly soluble pigments are treated in organic solvents (e.g., alcohols, glacial acetic acid, chlorobenzene, 1,2-dichlorobenzene, dimethylformamide). Preparation or finishing produces better-formed crystals and a narrower particle-size distribution, and therefore improves pigment dispersibility.

In conclusion, crystal physics are important in pigment synthesis. Once the insoluble pigment crystallites have been formed during synthesis, chemical influences no longer play a role. Only physical parameters such as crystal shape, crystal size, crystal quality, and particle-size distribution can subsequently be altered.

The apparatus used for the industrial synthesis of azo dyes (→ Azo Dyes, Chap. 2.5.) can also be used for the synthesis of azo pigments. For pigments requiring thermal after treatment, a pressure-resistant steel vessel with stirrer and connected filter press is necessary so that the crude pressed pigment cake can be made into a paste and heated with stirring (if necessary under pressure).

The industrial synthesis of azo pigments is now illustrated with some examples.

Production of C.I. Pigment Yellow 1, 11 680 [13] (see Section 4.2.1). *Diazotization.* In a diazotization vessel, 912 kg of 3-nitro-*p*-toluidine (4-methyl-2-nitroaniline) is stirred overnight with water. Then 3395 L of 5 mol/L hydrochloric acid is added, together with sufficient ice to lower the temperature to 0 °C. Sodium nitrite solution, 792 L of 40 wt %, is then run in underneath the surface. Any excess sodium nitrite is removed by addition of 3-nitro-*p*-toluidine. Kieselguhr or active charcoal is added as a clarifying agent to the diazonium salt solution, which is then filtered through clarifying filter equipment.

Dissolution of the Coupling Component. Acetanilide (1098 kg) is stirred with water for 1 h at 25 °C; 510 kg of calcium carbonate and 206 kg of 33 % sodium hydroxide solution is then added.

Coupling. The reaction mixture is heated to 33 °C, and the clarified diazonium salt solution is then run in. After 30 min hydrochloric acid is added to adjust the pH to 3 – 4. After an additional 20 min the pigment is separated by filtering the suspension through a filter press and

washed with water to remove salts. The filter cake is dried at 60 – 65 °C. The yield is 2042 kg.

Production of C.I. Pigment Red 112, 12 370 [14] (see Section 4.6). *Diazotization.* In a steel-enamel stirred tank, 500 kg of glacial acetic acid is heated to 65 °C, and 276 kg of 2,4,5-trichloroaniline is added. This solution is run into a mixture of 3500 L of water and 1120 L of 5 mol/L hydrochloric acid. The reaction mixture is cooled to 0 °C by addition of ice and rapidly diazotized with 185 L of 40 wt % sodium nitrite solution. The temperature must not exceed 0 °C. After 2 h, 107 kg of calcium carbonate suspended in water is added. The volume is made up to 8500 L with water; kieselguhr and active charcoal are added, and the solution is filtered through a clarifying filter.

Dissolution of the Coupling Component. At 30 °C, 420 kg of Naphtol AS-D is dissolved in 322 L of 33 % sodium hydroxide and 3000 L of water; 10 kg of kieselguhr is added, and the solution is filtered. The volume is made up to 14 000 L and temperature is 20 °C. Then, 84 g of Igepon T [the condensation product of oleyl chloride with sodium methyltaurine, $H_{33}C_{17}CO(CH_3)N(CH_2)_2SO_3Na$] in 200 L of water is added and the mixture is cooled to 3 °C with ice. The coupling component is precipitated by rapid addition of 235 L of glacial acetic acid.

Coupling. The reaction mixture is heated to 38 °C, and the filtered diazonium salt solution is run in under the surface of the suspension of the coupling component within 3.5 h. After completion of coupling, the pigment is isolated by filtering the suspension through a press, washing to remove salts, and drying at 50 – 55 °C. The yield is 700 kg.

Production of C.I. Pigment Yellow 13, 21100 [15] (see Section 4.3.1). *Bisdiazotization.* 151.8 kg of 3,3′-dichlorobenzidine dihydrochloride (100 %) is added to 2000 L of water and stirred overnight. The next morning 600 L of 5 mol/L hydrochloric acid is added, and the reaction mixture is cooled to −3 °C with ice. Bisdiazotization is performed by rapid addition of 157 L of 40 wt % sodium nitrite solution, whereby the temperature should not rise above 0 °C. Then, 10 kg of kieselguhr is added to the brownish-yellow solution, which is made up to 10 000 L and clarified.

Dissolution of the Coupling Component. Acetoacet-2,4-dimethylanilide, 258 kg, is dissolved in a mixture of 133 L of 33 % sodium hydroxide and 4000 L of water. Part of the acetoacetxylidide is precipitated by the addition of 150 L of 5 mol/L hydrochloric acid; 60 kg of sodium formate and 90 kg of calcium carbonate are then added, and the volume is made up to 8000 L. The temperature is 25 °C.

Coupling. The bisdiazonium salt solution is run in within 2.5 h underneath the surface of the suspension of the coupling component. Stirring is continued for another 30 min, and 200 L of 5 mol/L hydrochloric acid is added to adjust the pH to ca. 3.5. After an additional 30 min stirring, the reaction mixture is heated to 100 °C and held there for 1 h. The reaction mixture is then cooled to 70 °C by adding water; the pH should remain constant at 3.5. The suspension is filtered through a press and the pigment is washed to remove salts. After drying at 60 – 65 °C, 381 kg of pigment is obtained.

4.2. Yellow and Orange Monoazo Pigments

Yellow and orange monoazo pigments include compounds that are produced by coupling diazotized substituted anilines with linear or cyclic compounds containing an activated methylene group (e.g., acetoacetarylides or 1-aryl-5-pyrazolones).

4.2.1. Yellow Monoazo Pigments (Hansa Yellow Pigments)

The most important coupling components in the yellow monoazo pigment class are derivatives of acetoacetylaminobenzene (acetoacetarylides).

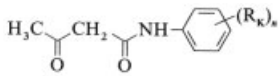

The preferred substituents R_K are CH_3, OCH_3, OC_2H_5, Cl, Br, NO_2; for industrially important products, *n* is 0 – 3. The acetoacetarylides are produced by reacting the corresponding aromatic amines with diketene or ethyl acetoacetate.

The yellow monoazo pigments are obtained by diazotizing substituted anilines and coupling them with acetoacetarylides. Their general synthesis is thus represented by **3**. (This designation is conventional in azo dyes and azo pigments. The horizontal arrow between the aromatic amine and the coupling component denotes "by coupling diazotized . . . with . . .").

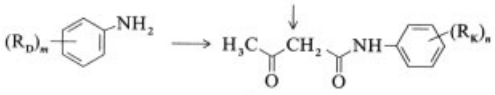

3

In the most important pigments, R_D and *m* have the same meaning as R_K and *n* (see above). Whereas the first red azo pigments based on β-naphthol as coupling component were marketed around 1900, the first yellow azo pigments with acetoacetarylides as coupling components were discovered by the firm Meister Lucius & Brüning in 1909 (Hansa Yellow). These pigments are still industrially important.

The basic structure of yellow monoazo pigments is generally described in the 2-oxoazo form (**4**),

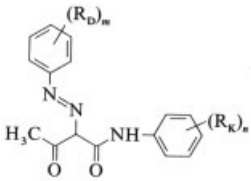

4

although the 2-oxohydrazone form (**5**), which has thus far been found exclusively in pigments investigated by three-dimensional X-ray structural analysis, probably occurs in reality [16].

C.I. Pigment Yellow 6, 11670 [*4106-76-7*]:

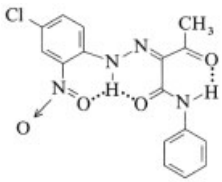

5

Yellow monoazo pigments are very lightfast and resistant to weathering, but only moderately resistant to solvents, overspraying, overpainting, and bleeding. Because they are relatively cheap to produce, some yellow monoazo pigments still are among the most important organic pigments even though they have been known for more than 80 years.

Yellow monoazo pigments are used mainly to pigment coating materials, especially air-drying paints and emulsion paints, as well as in the printing industry where a high recrystallization resistance is not required. These pigments are less suitable for pigmenting plastics because of their unsatisfactory migration fastness.

Examples of industrially important yellow monoazo pigments follow:

C.I. Pigment Yellow 1, 11680 [*2512-29-0*] (for industrial production, see Section 4.1.3)

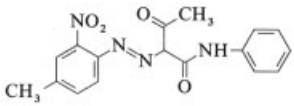

C.I. Pigment Yellow 3, 11710 [*6486-23-3*]:

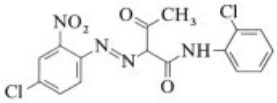

Both Pigment Yellow 1 and Pigment Yellow 3 are inexpensive. Pigment Yellow 1 has a medium yellow shade. Pigment Yellow 3 has a strong greenish-yellow shade. They are used in large amounts in almost full-shade ranges in paints.

C.I. Pigment Yellow 74, 11741 [*6358-31-2*] is also a pure greenish-yellow pigment, with a very high tinting strength (the order of magnitude of diaryl yellow pigments). This pigment is therefore also used in the printing ink industry when the better lightfastness of yellow monoazo pigments is required.

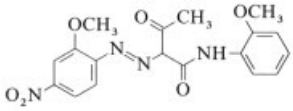

C.I. Pigment Yellow 97, 11767 [*12225-18-2*] has an extremely good lightfastness and weather resistance. It also has a much better solvent resistance and fastness to overpainting than the previously described yellow monoazo pigments:

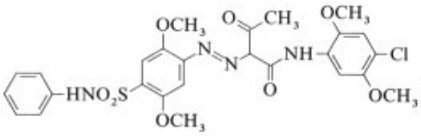

The main difference between Pigment Yellow 97 and other yellow monoazo pigments is the presence of an additional sulfonylaminophenyl group, which is primarily responsible for its better migration fastness. Because of its good fastness properties and clean shade, Pigment Yellow 97 is used on a large scale in artists' colors and industrial paints.

4.2.2. Monoazopyrazolone Pigments

The acetoacetylamino group of the coupling component in yellow monoazo pigments can be replaced by the 5-pyrazolone ring, in which the acetoacetyl group is incorporated into a heterocyclic ring system. When reacted with the same substituted anilines as diazo components, this large conjugated system (**6**) generally yields reddish-yellow pigments.

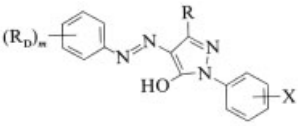

6

In **6**, R is usually CH_3, $COOCH_3$, or $COOC_2H_5$; X may be H, CH_3, or Cl; R_D is defined in the same way as R_K (see Section 4.2.1).

Despite their extremely good lightfastness in full shades and high tinting strength, monoazopyrazolone pigments are no longer important, primarily due to their unsatisfactory solvent resistance and migration fastness.

C.I. Pigment Yellow 10, 12710 [*6407-75-6*] still has some economic importance, however, particularly in the United States where it is used for road-marking paints:

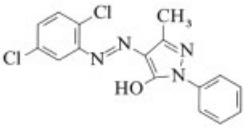

4.3. Disazo Pigments

Disazo pigments are divided into two structural types, depending on whether a bifunctional diazo component or a bifunctional coupling component is used.

Disazo pigments with *bifunctional diazo components* (diamines) are obtained by bisdiazotization of a diamine $H_2N-A-NH_2$ and subsequent coupling with two equivalents of coupling component. These pigments are also termed *diarylide pigments* and are represented by general formulas **7** and **8**:

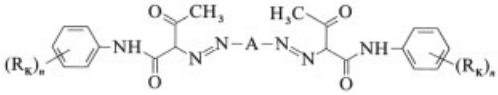

7
Yellow

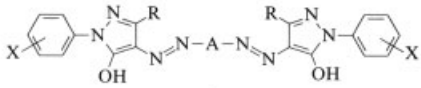

8
Orange, red
A = diphenyl residue

where R_K has the same meaning as in Section 4.2.1; X may be H, CH_3, or Cl.

Disazo pigments with *bifunctional coupling components* are obtained by coupling two equivalents of a diazotized amine $A-NH_2$ with one equivalent of a bisacetoacetylated diamine $H_2N-B-NH_2$. Disazo pigments are also called *bisacetoacetarylide pigments:*

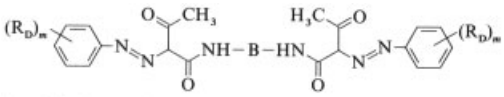

B = phenylene or diphenylene residue

4.3.1. Diarylide Pigments

Diarylide pigments are extremely important economically, particularly in the production of printing ink. They generally have structure **9**:

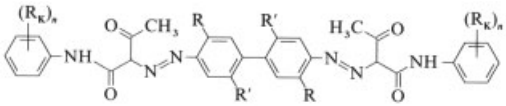

9

R = Cl, CH_3, OCH_3; R′ = H or R = R′ = Cl

3,3′-Dichlorobenzidine is by far the most important bisdiazo component.

The term "benzidine pigments" is still sometimes used for this group, but it is incorrect because benzidine has never been used to produce diarylide pigments.

Diarylide pigments are produced by the bisdiazotization of 3,3′-dichlorobenzidine, followed by coupling with two equivalents of an acetoacetic arylide (example in Section 4.1.3).

Because of their large molecular size, diarylide pigments are considerably more resistant to solvents and recrystallization than monoazo pigments. Because they contain more conjugated double bonds they have a substantially higher tinting strength but a lower lightfastness (although this is usually satisfactory for most applications).

Examples of industrially important diarylide pigments follow:

C.I. Pigment Yellow 12, 21090 [*6358-85-6*] (**10**) is an extremely important diarylide pigment with one of the largest organic pigment sales volumes.

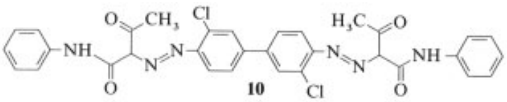

10

C.I. Pigment Yellow 13, 21100 [*5102-83-0*] (**11**) has an even higher tinting strength and better lightfastness than Pigment Yellow 12, good solvent and recrystallization resistance, and a shade complying with the European standard for three- and four- color printing. The production of Pigment Yellow 13 is described in Section 4.1.3.

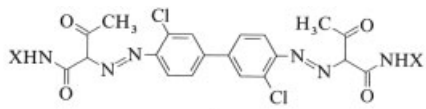

11

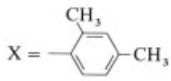

Both *C.I. Pigment Yellow 14, 21095* [*5468-75-7*] (**12**) and *C.I. Pigment Yellow 17, 21105* [*4531-49-1*] (**13**) are important greenish-yellow pigments.

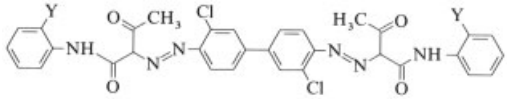

Y = CH_3: **12**
Y = OCH_3: **13**

These too may be regarded as typical printing ink pigments and are suitable for all printing processes. Pigment Yellow 13 and Pigment Yellow 17 are also used to color plastics.

If 2,2′,5,5′-tetrachlorobenzidine is used as the bisdiazo component instead of 3,3′-dichlorobenzidine, the hue is displaced toward greenish yellow. An important example is *C.I. Pigment Yellow 81, 21127* [*22094-93-5*] (**14**):

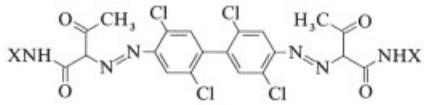

14

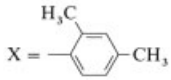

The solvent fastness of diarylide pigments and the resulting technical properties (e.g., fastness to blooming and fastness to overspray) are considerably superior to those of monoazo pigments. The lightfastness and weather resistance of most disazo pigments are, however, generally not sufficient for use in paints. Exceptions include Pigment Yellow 81 (**14**) and, in particular, *C.I. Pigment Yellow 83, 21180* [*5567-15-7*] (**15**):

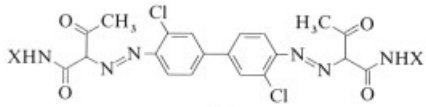

15

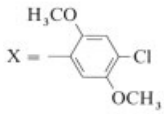

Pigment Yellow 83 is a reddish-yellow pigment of high tinting strength that is used because of its good technical properties in graphic

printing and in coloring coating materials and plastics.

Pigment Yellow 81 is also suitable for coloring plastics. Pigment Yellow 13 (**11**) and Pigment Yellow 17 (**13**) have good tinting strengths but are less lightfast; they are used as pigments in inexpensive plastic articles.

Two disazopyrazolone pigments are of particular industrial importance, namely, *C.I. Pigment Orange 13, 21110* [*3520-72-7*] (**16**) and *C.I. Pigment Orange 34, 21115* [*15793-73-4*] (**17**).

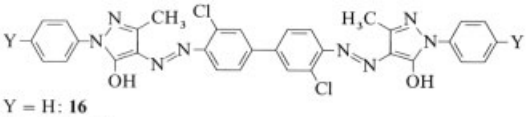

Y = H: **16**
Y = CH_3: **17**

Pigment Orange 13 (**16**) has a considerably more yellowish shade and is somewhat less lightfast than Pigment Orange 34 (**17**). Both pigments are used on a large scale in the graphics industry for all conventional printing methods. In addition, they are used in the plastics industry when less stringent fastness properties are required.

4.3.2. Bis(*N*-acetoacetarylide) Pigments

Of the bis(*N*-acetoacetarylide) pigments, *C.I. Pigment Yellow 16, 20040* [*5979-28-2*] (**18**) has the greatest industrial importance (see below).

Pigment Yellow 16 is produced by coupling two equivalents of diazotized 2,4-dichloroaniline with bisacetoacetylated *o*-toluidine (Naphtol AS-G). The pigment has a medium-yellow shade and is particularly suitable for paints because of its good fastness to blooming and overspray. It is also used in printing inks and plastics.

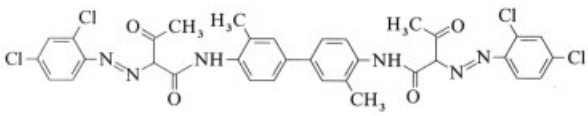

18

4.4. *β*-Naphthol Pigments

β-Naphthol pigments contain 2-hydroxynaphthalene (*β*-naphthol) as coupling component and anilines that are substituted mainly in the 2- or 4-positions as diazo components. Their general formula is:

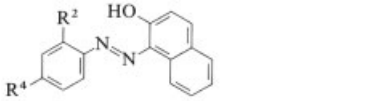

$R^2 = R^4$ = H, Cl, NO_2, CH_3, OCH_3, OC_2H_5

Like the Hansa Yellow pigments, *β*-naphthol pigments probably occur in the *o*-quinonehydrazone and not the hydroxyazo form (see Section 4.2.1). This was first demonstrated for *C.I. Pigment Red 1, 12070* [*6410-10-2*] (**19**) by three dimensional X-ray structural analysis [17]:

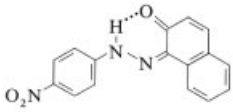

19

Crystal structure determinations subsequently carried out on Naphtol AS pigments gave similar results. All hitherto investigated pigment molecules are almost completely planar [18].

Azo compounds based on *β*-naphthol are among the oldest known synthetic dyes. However, *β*-naphthol pigments, such as C.I. Pigment Red 1 (**19**, Para Red), were first produced more than a century ago. Para Red was first used in "ice colors" as a developing dye for cotton (see → Azo Dyes, Chap. 6.1.1.1.). Subsequently, the compound was also produced by coupling in the presence of a carrier material and, finally, in the form of the pure substance as a pigment (Paratoner). The pigment is now used to only a limited extent, mainly for coloring inexpensive coating materials. Today only a few straight *β*-naphthol pigments (i.e., which have not been converted into lakes) are industrially important;

those that are used, however, are employed in large amounts.

The solvent resistance and migration fastness of these pigments are poor. Because of their clean hues, good lightfastness, weather resistance, and relatively low cost, *C.I. Pigment Red 3, 12120* [*2425-85-6*] (**20**, Toluidine Red) and *C.I. Pigment Orange 5, 12075* [*3468-63-1*] (**21**, Dinitraniline Orange) are among the most important organic pigments.

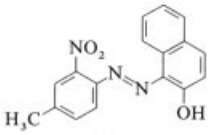

20

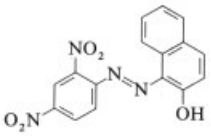

21

Both pigments are used mainly in the full-shade range to color coating materials. Grades of Pigment Red 3 (**20**) with a yellowish-red to bluish-red hue are commercially available. Pigment Orange 5 (**21**) is also used on a moderately large scale in the printing ink industry.

4.5. Azo Pigment Lakes

Azo pigment lakes are based on monoazo dyes having sulfonic acid groups. These soluble dyes are converted into insoluble pigments by salt formation (lake formation) with metal ions (e.g., from sodium, calcium, strontium, barium, or manganese salts). In the following formulas, the symbol M denotes metal ions. The lakes are divided into β-naphthol pigment lakes and 2-hydroxy-3-naphthoic acid (BON) pigment lakes.

4.5.1. β-Naphthol Pigment Lakes

Coupling a diazotized aniline sulfonic acid with β-naphthol yields a monoazo dye that is converted to the pigment by lake formation. The sodium salt of β-naphthol is coupled in an alkaline medium and then reacted with an alkaline-earth or manganese salt.

The first pigment of this series, a coupling product of 1-naphthylamine-2-sulfonic acid and β-naphthol (Lithol Red), was patented in 1899 by Julius (BASF). The sodium, barium, and calcium salts, which were initially precipitated on a carrier but were subsequently used as pure products, proved to be valuable pigments for the printing ink industry because of their insolubility.

Many β-naphthol pigment lakes were developed, but very few are still used today. Some of them are, however, of considerable industrial importance. Examples include the following:

C.I. Pigment Red 49, 15630 (**22**, Lithol Red, [*1103-38-4*] for M = Ba) and *C.I. Pigment Red 53, 15585* (**23**, Lake Red C, [*5160-02-1*] for M = Ba).

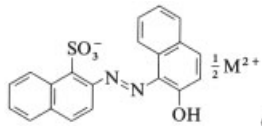

22

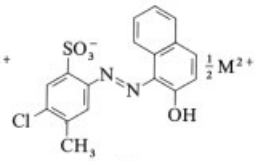

23

These pigments (M is usually barium) are particularly important for printing inks. Although Pigment Red 49 (**22**) is still produced on a large scale, particularly in the United States, its importance is decreasing. Pigment Red 53 (**23**) exists in both transparent and an opaque forms, depending on the method of production and the additives used (e.g., rosin soaps). The barium lake has a medium-red shade. Because of its high tinting strength and solvent fastness the pigment is used on a large scale in printing processes (i.e., for letterpress, offset, gravure, and flexographic printing).

C.I. Pigment Red 68, 15525 (Ca) [*5850-80-6*] is a clean yellowish-red pigment. Although only moderately lightfast, it is extremely heat stable, which makes it particularly suitable for pigmenting plastics (e.g., PVC or polyolefins).

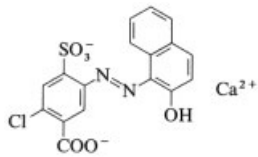

Production of C.I. Pigment Red 53:1, 15585:1 [*5160-02-1*] [19]. *Diazotization.* 487.3 kg of 4- chloro-3-methylaniline-6-sulfonic acid (CLT acid) is stirred overnight with a mixture of 3000 L of water and 114 kg of hydrochloric acid (calculated as 100 %). The next morning, the reaction mixture is cooled with ice

to 15 °C and the volume is made up to 10 000 L. Then, 288.6 L of 52.6 wt % sodium nitrite solution is run in beneath the surface within 3 h. The temperature is maintained at 15 – 18 °C; the final volume is 12 000 L.

Dissolution of the Coupling Component and Coupling. 158.4 kg of sodium hydroxide is dissolved in 4000 L of water at 20 °C; 326 kg of β-naphthol is then added. After ca. 20 min at 20 °C, all the naphthol has dissolved. The volume is made up to 9000 L, and the diazonium salt solution is run into the sodium naphtholate solution within 1 – 5 min. The reaction mixture is stirred for 1 h, and a mixture of ca. 50 – 55 kg of acetic acid and 300 L of water is added to neutralize the reaction mixture. The mixture is checked for a pH of 7 and for the absence of diazonium salt, and is stirred for 2 h. It is then filtered and washed with the last quarter of the mother liquor (if at all). About 440 kg of the pressed filter cake is stirred overnight with 4000 L of water. The suspension is then diluted with 7000 L of water and stirred overnight. The resulting suspension is allowed to flow into boiling barium chloride solution prepared from 14 000 L of water, 580 kg of sodium chloride, and 175 kg of barium chloride in a stirred tank. Addition lasts ca. 1.5 h. The temperature is then maintained for another hour at 102 – 104 °C. The suspension is filtered through a press, and the pigment is washed with mother liquor and dried at 70 °C. The yield is 550 kg.

4.5.2. 2-Hydroxy-3-naphthoic Acid Pigment Lakes (BON Pigments)

The BON group of pigments consists of monoazo pigment lakes in which 2-hydroxy-3-naphthoic acid (beta-oxynaphthoic acid, BON) is the coupling component. In this case too, aromatic amines, mainly those with a sulfonic acid group in the 2-position to the amino group, are used as diazo components. The BON pigments are more lightfast than β-naphthol pigments and have a more bluish-red shade (Rubine).

The pigments are preferably calcium or barium salts. Manganese lakes, which are generally more lightfast than the corresponding alkaline-earth salts, are most important in the paint industry but are also used in printing inks. They are seldom used for plastics, primarily because manganese catalyzes the oxidative decomposition of some plastics.

C.I. Pigment Red 57:1, 15850:1 [*5281-04-9*] (Lithol Rubine) is one of the most important organic pigments and an important printing ink pigment:

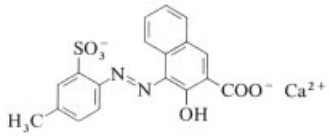

It is used in almost all printing ink sectors (in letterpress and offset printing, and in illustration and packaging gravure printing) for several reasons. It has a brilliant shade that corresponds to the standard "magenta" shade of the European scale for multicolor, letterpress, and offset printing, as well as high tinting strength, good lightfastness, and solvent resistance.

C.I. Pigment Red 48, 15865 (Permanent Red 2B Pigment) is used in all important areas. It is the calcium or manganese salt of the coupling product formed from 2-amino-4- chloro-5-methylbenzenesulfonic acid and 2-hydroxy-3-naphthoic acid:

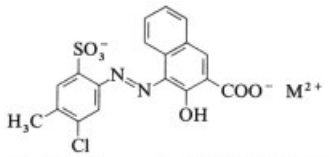

M = Ca: Pigment Red 48:2 [*7023-61-2*]
M = Mn: Pigment Red 48:4 [*5280-66-0*]

The manganese lake in particular is used for air-drying paints and stoving enamels because of its good lightfastness and fastness to overpainting.

C.I. Pigment Red 52, 15860 (BON Lake Red C) is also used in the graphics industry, mainly in the United States.

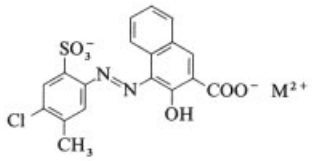

M = Ca: Pigment Red 52:1 [*17852-99-2*]
M = Mn: Pigment Red 52:2 [*12238-31-2*]

4.5.3. Other Pigment Lakes

In addition to the two groups of pigment lakes discussed in Sections 4.5.1 and 4.5.2, some acetoacetic arylide and Naphtol AS pigment lakes are also available that are used mainly to pigment plastics because of their good migration fastness. An example is *C.I. Pigment Yellow 61, 13880* (R = H) [*12286-65-6*]:

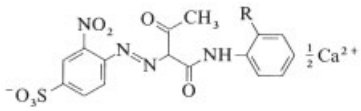

The pigment has a clean shade with a greenish-yellow hue, good lightfastness and weather resistance, and a particularly high thermal stability. It can be used to color PVC and polyolefins.

C.I. Pigment Yellow 62:1, 13940:1 (R = CH_3) [*12286-66-7*] has similar uses.

C.I. Pigment Red 151, 15892 [*61013-97-6*] is a bluish-red pigment that is used, among others, for pigmenting PVC and polyolefins. This pigment also has excellent heat stability:

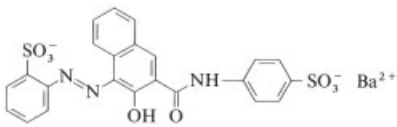

4.6. Naphtol AS Pigments

In Naphtol AS pigments, the anilide of 2-hydroxy-3-naphthoic acid (trade name: Naphtol AS) and its substitution products are used as coupling components. They have the following general structure:

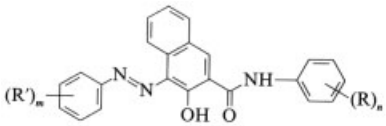

Some important substituents are R = CH_3, OCH_3, OC_2H_5, Cl, Br, NO_2, $NHCOCH_3$; R′ = R, $COOCH_3$, $CONHC_6H_5$; *m, n* = 0 – 3.

This type of pigment achieved initial popularity in the form of developing dyes for cotton (→ Azo Dyes, Chap. 6.1.1.2.). In this group of dyes, 2-hydroxy-3-naphthoylaniline (Naphtol AS) was first used as a coupling component instead of β-naphthol in 1911. The sodium salt of Naphtol AS is absorbed by cellulose fibers due to its substantivity. By "development" with diazo components (coupling), high colorfastness is obtained. Insoluble colorants (i.e., pigments) are thus formed on the fiber. The discovery of Naphtol AS pigments was patented as early as 1911. Naphtol AS and many derivatives with a substituted anilide ring (→ Azo Dyes; Colour Index) have been described. However, these derivatives only became important when the price of 2-hydroxy-3-naphthoic acid arylides fell as a result of the widespread use of Naphtol AS developing dyes.

Naphtol AS derivatives are produced by condensing 2-hydroxy-3-naphthoic acid with appropriately substituted anilines in the presence of phosphorus trichloride in aromatic hydrocarbons (e.g., toluene or xylene). The Naphtol AS derivatives (Red Naphtols) produce orange, red, maroon, or brown shades with diazotized aromatic amines, the red hues being particularly important. A wide range of Naphtol AS pigments is available commercially, some of which are produced in considerable amounts.

The pigments generally have high tinting strength, good resistance to chemicals, and good lightfastness, even in light shades. The somewhat poorer solvent fastness can be improved by incorporating amide groups.

C.I. Pigment Red 112, 12370 [*6535-46-2*] is the most important Naphtol AS pigment and is the only one included among the 20 most important organic pigments:

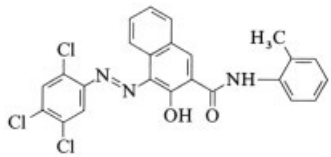

Because of its brilliant medium-red shade, high lightfastness, and good weather resistance, it is used in paints and printing inks in a wide variety of vehicles. Its unsatisfactory migration fastness results in the pigment not being used in plastics. Production is described in Section 4.1.3.

In some red Naphtol AS pigments, 2-methyl-5-nitroaniline is the diazo component. Exam-

ples are *C.I. Pigment Red 8, 12335* [*6410-30-6*] (**24**); *C.I. Pigment Red 17, 12390* [*6655-84-1*] (**25**); and *C.I. Pigment Red 22, 12315* [*6448-95-9*] (**26**):

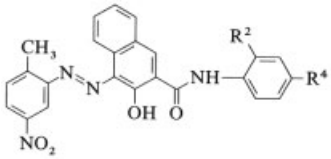

24: $R^2 = H$, $R^4 = Cl$
25: $R^2 = CH_3$, $R^4 = H$
26: $R^2 = R^4 = H$

In Europe, Pigment Red 8 is most important, whereas in the United States and Japan, Pigment Red 22 is used on a large scale. All three are used mainly in printing inks.

The color of Pigment Red 17 (**25**) is considerably shifted to the bathochromic side of the spectrum in its isomer *C.I. Pigment Red 12, 12385* [*6410-32-8*]:

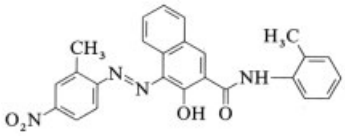

Displacement of the nitro group of the diazo component from the 5- to the 4-position changes the shade from a yellowish red to a bluish red (maroon). Pigment Red 12 is used primarily in paints.

In addition to 2,4,5-trichloroaniline, other chlorine-substituted anilines are also used as diazo components. An example is *C.I. Pigment Red 9, 12460* [*6410-38-4*] (**27**):

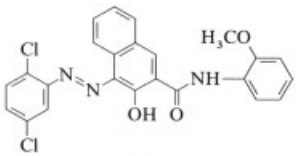

27

This compound is used mainly to pigment printing inks for letterpress and offset printing.

Pigments with a higher solvent resistance and bleeding fastness are obtained by introducing further carbonamide groups. Such pigments are also suitable for stoving enamels. *C.I. Pigment Orange 38, 12367* [*12236-64-5*] (**28**), for example, is a yellowish-red pigment that can also be used to pigment PVC.

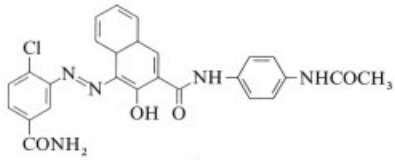

28

C.I. Pigment Red 146, 12485 [*5280-68-2*] (**29**) (Carmine) has a bluish-red shade. Because of its good solvent fastness it is widely used in paints and plastics, as well as special print finishes that can be oversprayed or laminated.

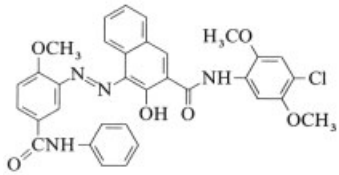

29

C.I. Pigment Red 170, 12475 [*2786-76-7*] (**30**) is a brilliant bluish-red pigment with good lightfastness and weather resistance. It is used mainly in industrial paints.

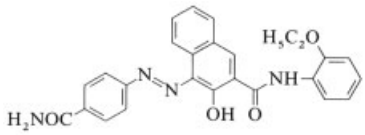

30

Another modification of the pigment with a greater hiding power and a more yellowish-red shade has even better fastness properties.

Pigment Red 146 (**29**) and Pigment Red 170 (**30**) are used in combination with inorganic red pigments (e.g., molybdate red) to produce particularly clean red shades with a high hiding power that are useful for industrial paints.

C.I. Pigment Red 187, 12486 [*59487-23-9*] contains three amide groups:

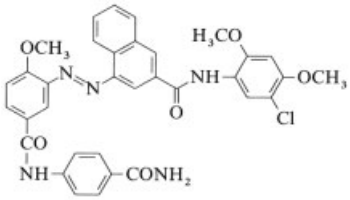

This bluish-red pigment is used to color PVC and polyolefins (including polypropylene for spin dyeing) because of its excellent migration fastness and thermal stability.

Naphtol AS pigments are not generally suitable for high-quality, exterior-use paints (automotive finishes). They are employed only in special cases to pigment plastics (e.g., PVC) because their migration fastness (e.g., with respect to overpainting, bleeding, and blooming) is generally unsatisfactory.

Migration-fast azo pigments that do not contain any lake-forming groups can be produced either by incorporating additional carbonamide groups or by substantial enlargement (doubling) of the molecule (see Section 4.8).

Incorporation of additional amide groups in the pigment molecule is particularly effective when these groups are part of a 5- or 6-membered, fused, heterocyclic ring.

Five-membered heterocyclic rings in the coupling component are extremely important technically in benzimidazolone pigments (Section 4.7).

4.7. Benzimidazolone Pigments [3, p. 523], [20]

The benzimidazoline pigments were developed by Hoechst and have been described in the patent literature since 1960 [21]. They contain a cyclic urea group (i.e., benzimidazolone ring) in the coupling component.

In yellow pigments the coupling component is 5-*N*-acetoacetylaminobenzimidazolone:

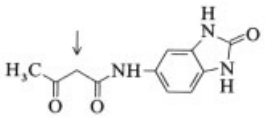

In red pigments the Naphtol AS derivative is used:

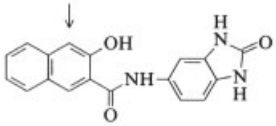

The coupling components are produced by reacting acetoacetic ester or diketene with 5-aminobenzimidazolone (yellow pigments) or by condensing 2-hydroxy-3-naphthoic acid with 5-aminobenzimidazolone in the presence of phosphorus trichloride as in the synthesis of Naphtol AS derivatives described in Section 4.6 (red pigments).

Diazotization of substituted anilines, followed by coupling with the above components, gives monoazo pigments of the Hansa Yellow type in the yellow range and of the Naphtol AS type in the red range.

The crude products are generally in an unsatisfactory crystalline state, which mostly results in pigments with hard particles that are difficult to disperse. The products are therefore usually thermally aftertreated in water, organic solvents, or mixtures of the two (Section 4.1.1).

These pigments can be produced with the same shade spectrum as azo pigments (i.e., ranging from greenish yellow through orange, red, carmine, maroon, and violet to brown). They have excellent solvent, overpainting, and bleeding fastnesses. These properties, as well as lightfastness and weather resistance, are superior to comparable Hansa Yellow and Naphtol AS pigments. Some benzimidazolone pigments are as lightfast and weather resistant as high-quality polycyclic pigments.

Many members of this class are used in high-quality industrial and exterior-use paints (e.g., automotive finishes, special metal-effect paints), plastics, and special printing inks. Some examples follow.

C.I. Pigment Yellow 120, 11783 [*29920-31-8*] is a yellow pigment with high tinting strength and good thermal stability. It is therefore used to particular advantage in plastics:

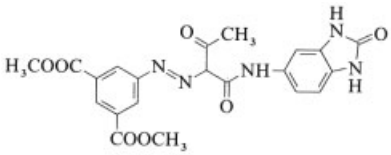

C.I. Pigment Yellow 151, 13980 [*61036-28-0*] is a pure greenish-yellow pigment that is used particularly in high-quality industrial paints, but also in plastics and lightfast prints:

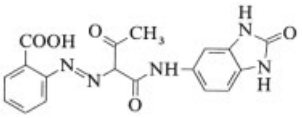

C.I. Pigment Yellow 154, 11781 [*63661-02-9*] is used particularly for high-quality paints because of its good lightfastness and weather resistance:

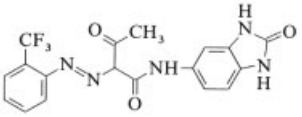

C.I. Pigment Orange 36, 11780 [*12236-62-3*] is also well suited to high-quality paints and coating materials (e.g., automotive finishes) and plastics:

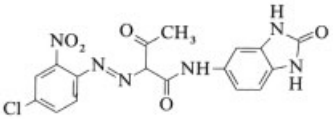

Pigments in which 2′-hydroxy-3′-naphthoyl-5-aminobenzimidazolone is used as coupling component include:

C.I. Pigment Red 175, 12519 [*6985-92-8*]

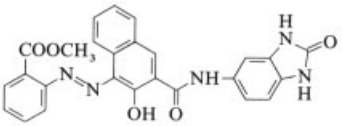

C.I. Pigment Red 176, 12515 [*12225-06-8*]

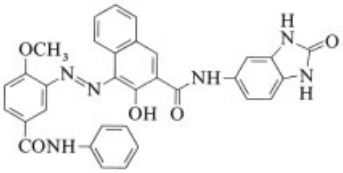

C.I. Pigment Violet 32, 12517 [*12225-08-0*]

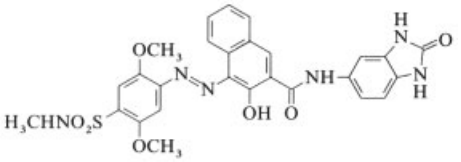

C.I. Pigment Brown 25, 12510 [*6992-11-6*]

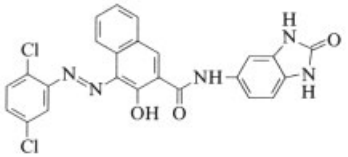

4.8. Disazo Condensation Pigments

[3, p. 561]

By use of bifunctional compounds, the molecules of conventional yellow and red monoazo pigments can be "doubled in size".

An attempt to produce such pigments was first made in the red range (Naphtol AS range) by double coupling of diazotized amines with bisnaphthols. However, in most cases, uniform pigments were not obtained. An industrially practicable method of synthesis was first discovered in 1951 by M. Schmid at CIBA, who condensed two monoazo pigment acid chlorides with one diamine. Pigments produced by this method are known as disazo condensation pigments.

4.8.1. Yellow Pigments

Yellow disazo condensation pigments have the two basic structures **31** and **32** shown in Figure 1.

Compounds of general formula **31** are produced by double coupling of diazotized aminobenzoic acids with the bisacetoacetylated coupling component, followed by conversion of the acid to the acid chloride and, finally, condensation with two equivalents of an aromatic amine.

In the synthesis of **32**, two equivalents of the coupling product formed from diazotized aminobenzoic acid and an *N*-acetoacetarylide (monoazo dye carboxylic acid) are converted to

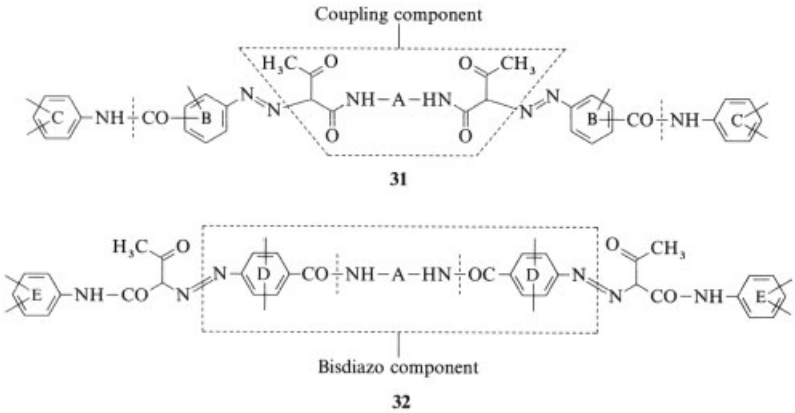

Figure 1. Basic structures of yellow disazo condensation pigments
A denotes the aromatic residue of a diamine. The aromatic rings B – E may include additional substituents (e.g., CH_3, OCH_3, Cl, Br, NO_2, OC_2H_5, $COOCH_3$).

the acid chloride and condensed with an aromatic diamine.

Theoretically, one-stage synthesis of disazo condensation pigments is also conceivable (i.e., in the case of **31** by double coupling of the diazotized aminobenzarylides with the bifunctional coupling component, or in the case of **32** by coupling the bisdiazo component with two equivalents of an *N*-acetoacetarylide). This is, however, impossible because, in both cases, the sparingly soluble monoazo pigment precipitates and is therefore only partially available for further reaction. The same applies to red disazo condensation pigments.

Examples of yellow disazo condensation pigments (mainly type **31**) are shown in Figure 2.

4.8.2. Orange, Red, and Brown Pigments

Orange, red, and brown disazo condensation pigments generally have the structure **33**:

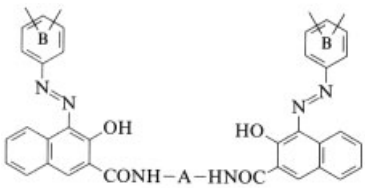

33

where A denotes the aromatic residue of a diamine; rings B may contain further substituents, including those with additional CONH groups. Like the yellow pigments, red pigments are produced by stepwise synthesis (see also Section 4.8.1):

1) Coupling of a diazotized amine with 2-hydroxy-3-naphthoic acid to form the monoazo dye acid
2) Conversion of the acid into the acid chloride with thionyl chloride, phosphorus oxychloride, or phosphorus pentachloride, if necessary in an organic solvent
3) Condensation of two equivalents of the monoazo dye acid chloride with one equivalent of an aromatic diamine in an organic solvent (e.g., 1,2-dichlorobenzene)

Examples of industrially important pigments follow.

C.I. Pigment Orange 31, 20050 [*12286-58-7*]:

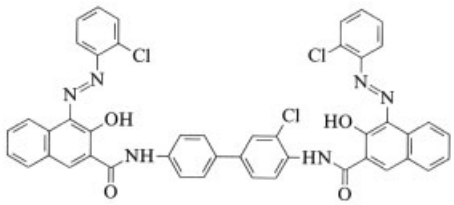

C.I. Pigment Yellow 93, 20710 [*5580-57-4*]:

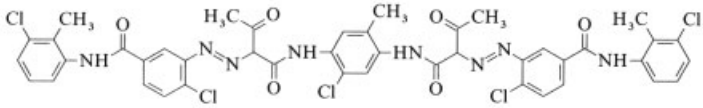

C.I. Pigment Yellow 94, 20038 [*5580-58-5*]:

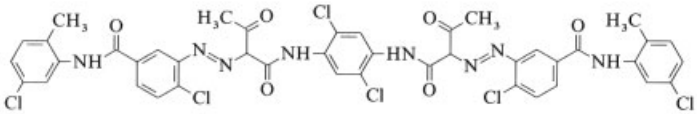

C.I. Pigment Yellow 95, 20034 [*5280-80-8*]:

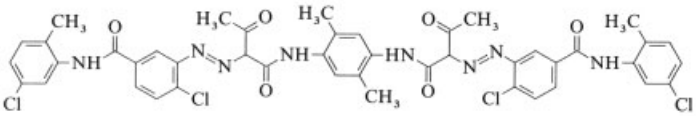

Figure 2. Examples of yellow disazo condensation pigments

C.I. Pigment Red 144, 20735 [*5280-78-4*]:

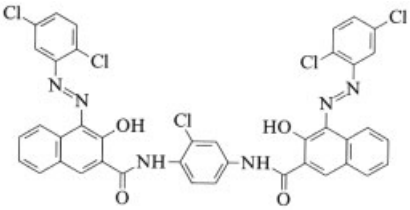

C.I. Pigment Red 166, 20730 [*12225-04-6*]:

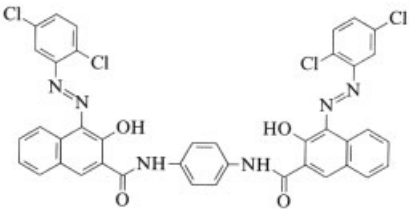

C.I. Pigment Brown 23, 20060 [*57972-00-6*]:

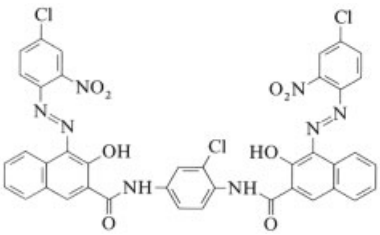

4.8.3. Properties

Due to their large molecular size, disazo condensation pigments have extremely good migration and solvent fastness, good thermal stability, high tinting strength (disazo pigments), and in some cases, better lightfastness and weather resistance than analogous monoazo pigments. These properties make them particularly suitable for the plastics and paint sectors.

5. Metal-Complex Pigments

Metal- complex pigments, with the exception of the large group of copper phthalocyanine pigments (→ Phthalocyanines, see also Chap. 7), are only of minor importance. Water-soluble metal complexes are, however, very important in textile dyes (Azo Dyes; → Metal-Complex Dyes, Chap. 5.1.).

One group of industrially important metal-complex pigments consists of 2,2′-dihydroxyazo compounds (**34**), which react with metals (primarily copper, nickel, cobalt, or zinc) to form 1 : 1 metal complexes (**35**) or 1 : 2 metal complexes (**36**).

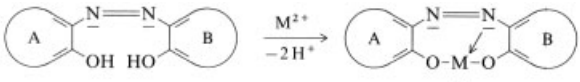

35a

or

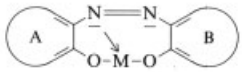

35b

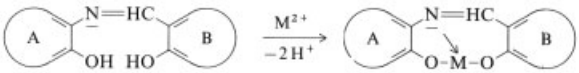

37

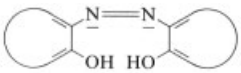

34

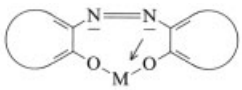

35

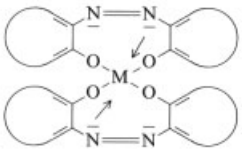

36

Azomethine metal complexes have also achieved a certain importance as pigments; here, the azo group −N=N− is replaced by an azomethine group −N=CH−.

In 1 : 1 azo metal complexes, specific coordination with the metal atom does not occur (**35a** and **35b**). Azomethine metal complexes, however, can only exist in one form (**37**). (see figure on top of this page)

Metalation of 2,2′-dihydroxyazo or azomethine compounds is generally performed with the appropriate metal salt in a variety of solvents.

The absence of solubilizing groups is a prerequisite for the pigment character of metal complexes. Metal- complex formation improves the lightfastness and the solvent, bleeding, and weather resistances of metal-free compounds. However, this is often offset by a dull shade.

5.1. Azo Metal-Complex Pigments

C.I. Pigment Green 10, 12775 [*51931-46-5*] (Nickel Azo Yellow) [3, p. 513] is the only azo metal- complex pigment of any industrial importance.

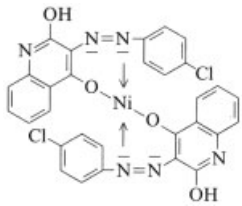

The yellowish-green pigment was developed by Du Pont in 1941 and came onto the market in 1946. It was long considered to be the azo pigment with the best lightfastness and weather resistance; however, it is only moderately resistant to acid and alkali, and has an inadequate bleeding fastness in stoving enamels. The pigment is used mainly in paints (e.g., in combination with phthalocyanine blue pigments) and, to a lesser extent, in plastics.

Interesting azo metal complexes, including the nickel complex of azobarbituric acid *C.I. Pigment Yellow 150, 12764* [*68511-62-6*] (**38**), have been described as pigments in the patent literature [22].

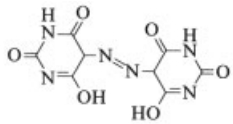

38

The exact structure of the nickel complex is unknown, though it is assumed to be a 1 : 1 layer complex [23]. The pigment is particularly suitable for paints because of its good lightfastness.

Azobarbituric acid can be obtained from 4-toluenesulfonyl azide and barbituric acid, which react to form diazobarbituric acid. The latter is then coupled with barbituric acid [24].

5.2. Azomethine Metal-Complex Pigments

Azomethine metal-complex pigments have been described by several companies. These predominantly yellow and orange pigments generally have outstanding lightfastness and weather resistance, and may therefore be used to particular advantage in paints subject to extremely stringent requirements [i.e., automotive finishes in single or metallic (aluminum) shades].

For example, the metal (mainly Cu) complexes of 2,2′-dihydroxyazomethines have interesting pigment properties. They are obtained from aromatic 2-hydroxyaldehydes and 2-aminophenols. Examples include *C.I. Pigment Yellow 117, 48043* [*21405-81-2*] (**39**) [25] and *C.I. Pigment Yellow 129, 48042* [*68859-61-0*] (**40**).

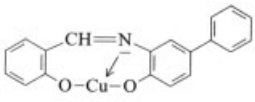

39

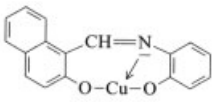

40

C.I. Pigment Orange 65, 48053 [*20437-10-9*] is obtained from *o*-phenylenediamine and 2-hydroxynaphthaldehyde after complexation with a nickel salt:

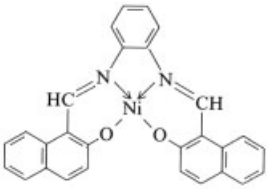

5.3. Miscellaneous Metal-Complex Pigments

Another basic structure, which however does not embody typical azomethine metal complexes, is represented by nickel complexes of the anilides of 2,3-dioximinobutyric acid [7], [26]:

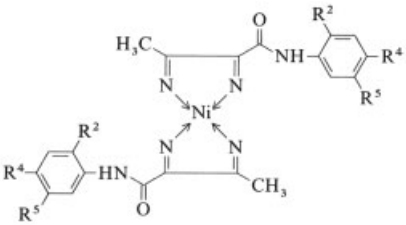

Different colors can be produced by varying the substituents R^2, R^4, and R^5:

C.I. Pigment Yellow 153, 48545 [*68859-51-8*]:
$R^2 = R^4 = R^5 = H$
Orange: $R^2 = OCH_3$, $R^4 = R^5 = H$
Red: $R^2 = R^4 = OCH_3$, $R^5 = H$

The pigments have good fastness properties and a high tinting strength [27]. They are produced by nitrosating the appropriately substituted *N*-acetoacetarylide with sodium nitrite in acetic acid, followed by formation of the oxime with a hydroxylamine salt:

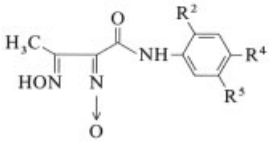

The metal complex is formed by reaction with a nickel(II) salt.

Isoindolinone metal-complex pigments (see also Chap. 6) are obtained by condensation of isoindoline with 2-aminobenzimidazole or 2-cyanomethylenebenzimidazole at ca. 130 °C, followed by complex formation with cobalt salts.

Examples include *C.I. Pigment Yellow 177, 48120* [*60109-88-8*] and *C.I. Pigment Yellow 179, 48125* [*63287-28-5*]:

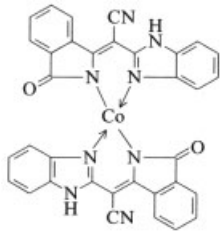

Pigment Yellow 177

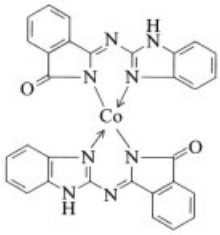

Pigment Yellow 179

6. Isoindolinone and Isoindoline Pigments [3, p. 703], [28]

6.1. Azomethine Type: Tetrachloroisoindolinone Pigments

Tetrachloroisoindolinone pigments contain two > C=N groups and must be regarded as a special type of disazomethine pigments. They consist of two tetrachloroisoindolinone rings joined by a diimine bridge:

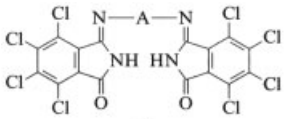

41

where A denotes an aromatic diamine residue.

The first nonchlorinated isoindolinone pigments were described in 1946 in an ICI patent. In 1952 and 1953, other authors published pigment syntheses based on the condensation of two equivalents of 3-amino-1-isoindolinone with diamines.

None of these "precursors" has the high tinting strength or the remarkable fastness (e.g., to solvents, light, and weathering) of the tetrachloroisoindolinone pigments, which were first described by GEIGY in 1964. Even pale shades of these pigments are extremely lightfast and weather resistant. The reasons for the high fastness are still unknown: it is assumed that intermolecular donor–acceptor complexes are formed between the phenyl or diamine residue and the chlorine-containing moiety of the molecule. The central aromatic ring A must be perpendicular to the two tetrachloroisoindolinone residues for steric reasons. Because of their good fastness properties the pigments are suitable for high-quality paints and plastics. The shade of the pigments is determined by the nature of the diamine residue: a larger conjugated system (diphenyldiamine residue) produces a bathochromic shift compared to a phenylenediamine residue.

The pigments are synthesized primarily by three methods:

1) Condensation of a diamine with two equivalents of a tetrachloroisoindolinone containing two chlorine atoms or methoxy groups as substituents in the 3-position:

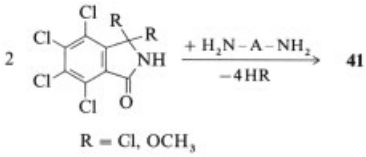

2) Condensation of a diamine with two equivalents of a tetrachloroisoindolinone containing an imino group in the 3-position:

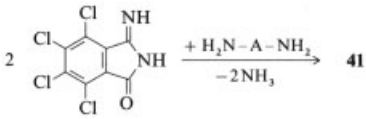

3) In a process developed by Dainippon Ink [29], two equivalents of tetrachlorophthalic anhydride are condensed with an aromatic diamine and ammonia in the presence of phosphorus pentachloride to form the bisacyl compound **42**, followed by ring closure with additional phosphorus pentachloride in trichlorobenzene to form the tetrachloroisoindolinone:

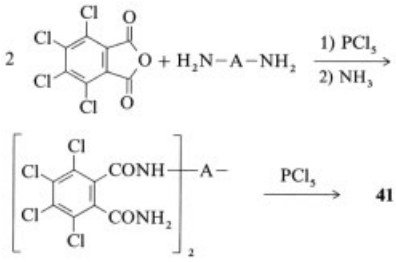

42

Examples of tetrachloroisoindolinone pigments include the greenish-yellow compound *C.I. Pigment Yellow 109, 56284* [*12769-01-6*] and the reddish compound *C.I. Pigment Yellow 110, 56280* [*5590-18-1*]. These pigments are used for paints, plastics, and high-quality printed products.

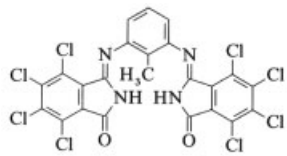

Pigment Yellow 109

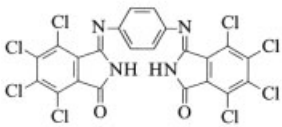

Pigment Yellow 110

6.2. Methine Type: Isoindoline Pigments

In methine-type pigments the isoindoline ring is coupled to two methine bridges:

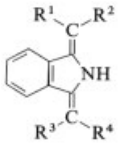

where $R^1 - R^4$ can denote CN, CONH-alkyl, CONH-aryl, CONH-heteroaryl; R^1 and R^2 and/or R^3 and R^4 may also be members of a heterocyclic ring.

An example is *C.I. Pigment Yellow 139, 56298* [*36888-99-0*] (**43**), a barbituric acid derivative of isoindoline [30]:

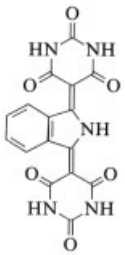

43

This reddish-yellow pigment has good light- and solvent fastness, and is produced by heating aminoiminoisoindolenine (diiminoisoindoline) with two equivalents of barbituric acid in glacial acetic acid or another solvent.

A red pigment of this type is *C.I. Pigment Red 260, 56259* [*71552-60-8*]; it has the following formula [31]:

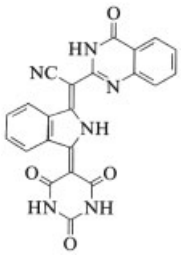

C.I. Pigment Orange 66, 48210 [*68808-69-5*] is a yellowish-orange pigment that is used particularly in paints:

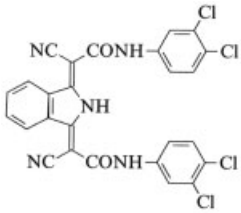

7. Phthalocyanine Pigments

Phthalocyanine pigments are the most important group of organic pigments worldwide and are described in detail elsewhere (→ Phthalocyanines).

By following the description given by De Diesbach (1927), copper phthalocyanine was first produced at about the same time by chemists at Scottish Dyes, Ltd. The first blue pigment of this type appeared on the market in 1935, and the first green pigment in 1938. Of historical interest is the fact that this pigment class was the first to be introduced industrially without previously having been used as a dye.

Five crystal modifications of copper phthalocyanine are known, of which the α- and β-types are of practical importance. The ε-modification is also of interest because of its reddish-blue shade. Phthalocyanines are the most important blue and green organic pigments because of their high tinting strength and outstanding stability with respect to light, weather, chemicals, and solvents, as well as their relatively low cost. Other blue and green organic pigments are only economically important for special applications.

Phthalocyanine pigments are suitable for all areas of use, such as paints and coating materials (including the shading of black and white hues), plastics, printing inks, and spin dyeing. The β-modification of phthalocyanine has been adopted as the standard blue (blue-green) shade of the European Scale for letterpress printing.

8. Quinacridone Pigments [3, p. 601], [32]

Quinacridone pigments are polycyclic and consist of five six-membered rings in a linear arrangement. The two peripheral rings are joined to the middle aromatic ring by two 4-pyridone rings:

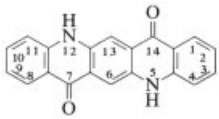

The two outer aromatic rings may be substituted.

Angular quinacridones were first mentioned in 1896; the linear ring system was described by LIEBERMANN in 1935 [33]. The significance of linear quinacridones as pigments was, however, only recognized in 1955 by W. STRUVE (Du Pont), who discovered a variation of the Liebermann process that enabled pure linear quinacridones to be produced on an industrial scale. He also recognized that unsubstituted quinacridone exists in different modifications, and described a process for their isolation.

In 1958, Du Pont introduced the first quinacridone pigments on the market. These pigments have since undergone significant development, particularly in Western industrialized countries, because they fill a gap in the range of bluish-red and red-violet pigments with outstanding fastness.

8.1. Production

The production of quinacridone pigments consists of synthesis and aftertreatment. If the pigment crystallizes in various forms (polymorphism), aftertreatment can be used to obtain a desired crystal modification (conditioning, finishing). However, crystal quality can also be improved by aftertreatment.

Unsubstituted quinacridones, dimethylquinacridones, dichloroquinacridones, and some mixed crystal phases, in some cases with quinacridonequinone, are available commercially.

8.1.1. Synthesis

Quinacridone [*1047-16-1*]. The process developed in 1955 by Du Pont is still industrially important for the synthesis of quinacridone. The succinosuccinate ester (**44**), which is easily obtained from dimethyl or diethyl succinate, is reacted with two equivalents of an aromatic amine (aniline, *p*- or *m*-toluidine, 4- or 3-chloroaniline) to form the 2,5-diarylamino-3,6-dihydroterephthalic acid diester (**45**). The latter can be converted directly to dihydroquinacridone (**46**) in the same reaction medium at 250 °C (thermal ring closure):

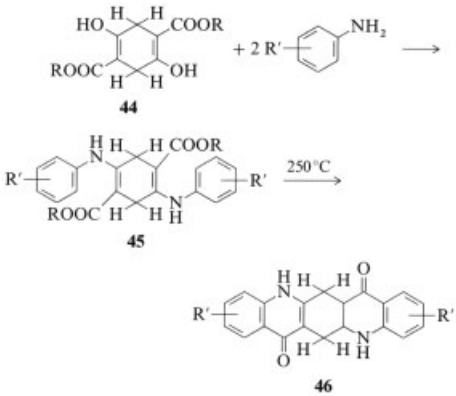

Dihydroquinacridone (**46**) can be oxidized to quinacridone with air or with sodium 3-nitrobenzenesulfonate in aqueous ethanol in the presence of sodium hydroxide. If a low sodium hydroxide concentration is used (ca. 2 % in the solvent mixture), the reaction mixture is oxidized heterogeneously (solid-state oxidation); with ca. 30 % sodium hydroxide in the solvent mixture, oxidation proceeds in homogeneous solution (solution oxidation). Different crystal modifications are formed depending on the nature of the reaction.

Another method is closely related to the synthesis described by LIEBERMANN. Here, the succinosuccinate ester (**44**) is condensed with an arylamine to form the ester **45**. The latter is oxidized to form the 2,5-diarylaminoterephthalic acid diester (**47**), from which the acid (**48**) is obtained by hydrolysis. The acid can be cyclized with polyphosphoric acid or other acid condensation agents to form the quinacridone (acid ring closure).

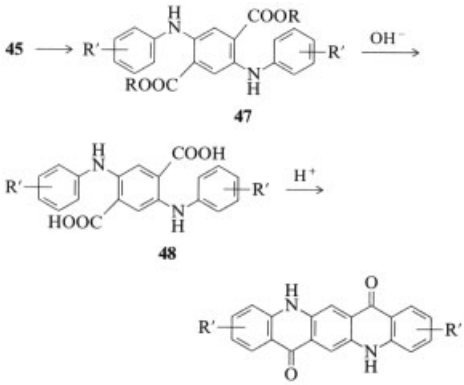

Another process permits the synthesis of asymmetrically substituted quinacridones, and starts from 2,5-dichloro- or 2,5-dibromoterephthalic acid. The halogen atoms can be replaced successively by arylamine residues (Sandoz). Cyclization is then performed as described above with acid condensation agents to form the quinacridone.

Another synthesis that proceeds via hydroquinone-2,5-dicarboxylic acid has been developed by BASF [32], [34].

Quinacridonequinone (C.I. 73920). A derivative that is of interest particularly for mixed crystal phases is the linear *trans*-quinacridonequinone (**50**). The compound can be prepared by cyclizing 2,5-bis(2′-carboxyanilino)-1,4-benzoquinone (**49**) with concentrated sulfuric acid or polyphosphoric acid at 150 – 200 °C. The starting material is obtained by condensing *p*-benzoquinone with anthranilic acid.

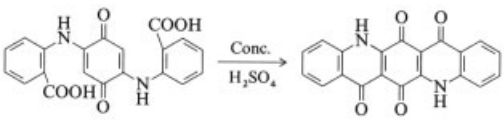

49 **50**

8.1.2. Aftertreatment

Quinacridone pigments can form different crystal modifications. Some of them are very significant industrially because of their outstanding fastness properties. Selective preparation of these modifications by aftertreatment of crude quinacridones is crucial.

Various crystal modifications of dihydroquinacridone may be formed in the synthesis of unsubstituted quinacridone by the Du Pont process, depending on the thermal ring closure method (see Section 8.1.1). The different forms of dihydroquinacridone yield various modifications of quinacridone, depending on the nature of the subsequent oxidation. For example, solid-state oxidation of α-dihydroquinacridone produces crude α-quinacridone, whereas solution oxidation produces crude β-quinacridone.

After synthesis, the crude pigment is usually finely divided by grinding it with salts in the presence of solvents that change or maintain the crystal modification. For example, with xylene the β-phase of the quinacridone pigment is obtained; dimethylformamide produces the γ-phase.

If quinacridone is synthesized by acid ring closure, the crude product occurs in a very finely divided form. The desired crystal modification can be obtained by suitable choice of solvent in subsequent thermal aftertreatment. For example, alkaline pretreatment followed by solvent treatment produces the β-phase of the quinacridone pigment, whereas direct solvent treatment yields the γ-phase.

The β-modification is usually formed by nonpolar solvents, whereas the γ-modification is formed by polar solvents.

8.2. Properties

Quinacridone pigments are insoluble in most organic solvents and all application media, and have high lightfastness and weather resistance as well as good general fastness. These properties are due to the formation of hydrogen bonds between the NH and CO groups of the molecules in the crystal lattice.

In solution, quinacridones have a pale yellow to orange color; the intense red color (with a yellowish-red to violet hue) is thus only formed by interactions in the crystal. Three-dimensional X-ray analyis revealed different crystal structures for the β- and γ-modifications: β-quinacridone forms crystal layers connected by intramolecular hydrogen bonds, the γ-phase has a "criss-cross" structure [35]. Color and properties can be influenced by altering particle size as well as in three ways:

1) Changing the crystal modification
2) Introducing substituents into the ring system
3) Forming mixed crystals of quinacridone with substituted quinacridones or with quinacridonequinone (**50**)

Quinacridone pigments, like the phthalocyanines, are among the organic pigments with the most outstanding fastness properties. They can comply with extremely stringent requirements (e.g., for industrial paints, particularly automotive finishes). They are, however, also used for the weatherfast coloring of plastics, for melt-spin dyeing, and for special printing inks. Their

good thermal stability means that they can also be used to color thermoplastics with a high softening point (e.g., polyolefins).

Examples of Commercial Products.

C.I. Pigment Violet 19, 46500 [*1047-16-1*], β-modification, has a violet shade and high tinting strength that makes it particularly suitable for adjusting red shades in automotive finishes by shading molybdate orange pigments; it can also be used for metallic finishes. Most of its important fastness properties are superior to those of the tetrachlorothioindigo pigments with similar shades (see Chap. 13).

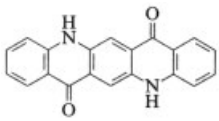

C.I. Pigment Violet 19, 46500, γ-modifications, is another industrially important modification of unsubstituted quinacridone. Depending on particle size, commercial grades range from bluish-red (more transparent) types to yellowish-red (more hiding) types. These grades are used especially in automotive finishes (e.g., to shade inorganic pigments).

C.I. Pigment Red 122, 73915 [*980-26-7*] (2,9-dimethylquinacridone, **51**) is obtained by ring closure of 2,5-bis-(4′-methylanilino)terephthalic acid (or esters) with acid condensation agents. It has a very pure bluish-red shade that lies between that of the β- and γ-modifications of Pigment Violet 19.

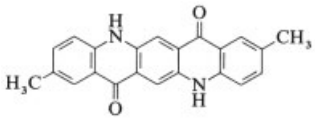

51

C.I. Pigment Red 202, 73907 [*68859-50-7*] (2,9-dichloroquinacridone, **52**) is formed from 2,9-dichlorodihydroquinacridone by solid-state oxidation (see Section 8.1.1) or from 2,5-bis-(4′-chloroanilino)terephthalic acid with sulfuric acid or polyphosphoric acid as condensation agent. The bluish-red modification of this pigment is of industrial importance.

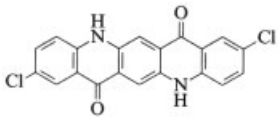

52

A yellowish-red dichloroquinacridone consists of a mixture of isomers, mainly 3,10-dichloroquinacridone, with an extremely clean shade.

Quinacridone mixed crystal phase pigments [34] include quinacridone (Pigment Violet 19) and quinacridonequinone (C.I. 73920). Whereas quinacridonequinone (**50**) alone is a pale yellow compound with no outstanding fastness properties (lightfastness), the mixed crystal phase with quinacridone has an interesting gold-orange shade with extremely good general fastness properties. The X-ray powder diffraction diagram shows a spectrum of reflection angles that are completely different from those of the individual pigments. To prepare the mixed crystal phase, a suspension of dihydroquinacridone (see Section 8.1.1) can be reacted with less than the molar amount of an oxidizing agent (e.g., chromate).

C.I. Pigment Red 207, a lightfast, scarlet pigment, consists of mixed crystals of quinacridone, Pigment Violet 19, and 4,11-dichloroquinacridone, C.I. 73906.

9. Perinone Pigments

The perinone pigments are synthesized from 1,4,5,8-naphthalenetetracarboxylic acid (**53**) or its monoanhydride (**53a**):

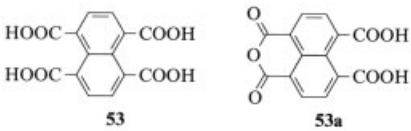

53 **53a**

The pigments are produced by reacting the anhydride with aromatic 1,2-diamines.

Naphthalenetetracarboxylic acid (**53**) is obtained by reacting acenaphthene (**54**) with malonodinitrile and aluminum chloride according to the Friedel–Crafts reaction. The condensation product (**55**) is then oxidized to di-

chloroacenaphthindandione (**56**) with sodium chlorate and hydrochloric acid. Finally, **56** is oxidized with sodium hypochlorite and potassium permanganate:

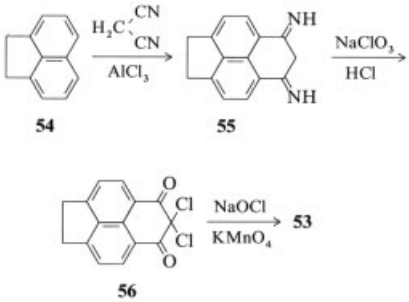

In another process, pyrene is halogenated to form the symmetric 1,3,6,8-tetrahalogen derivative, which is then hydrolyzed with sulfuric acid and oxidized to a diperinaphthindandione. The latter is finally oxidized in an alkaline medium to produce the tetrasodium salt of naphthalenetetracarboxylic acid [11].

A mixture of the two perinone isomers *Pigment Red 194* [*4216-02-8*] (**57**) and *Pigment Orange 43* [*4424-06-0*] (**58**) is obtained by condensing two equivalents of 1,2-diaminobenzene with one equivalent of naphthalenetetracarboxylic acid (e.g., in boiling glacial acetic acid).

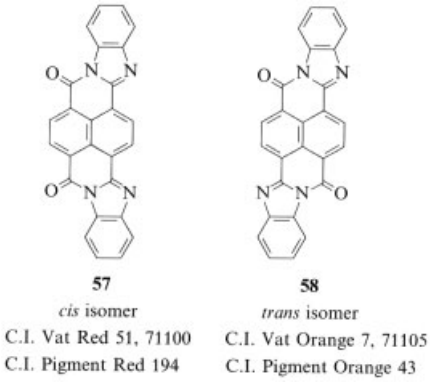

Originally only the isomer mixture was commercially available as a yellowish-red vat dye (*C.I. Vat Red 14, 71110*). The two isomers were subsequently separated by fractional precipitation with sulfuric acid or by heating in ethanol and potassium hydroxide, the *trans* isomer precipitating as the sparingly soluble potassium addition compound. The *cis* isomer Pigment Red 194 (**57**) has a dull red shade, whereas the *trans* isomer Pigment Orange 43 (**58**) has a pure orange shade.

Only when the importance of organic pigments increased was it recognized that some of the sparingly soluble vat dyes, including the perinones, could be used as pigments. Both isomers, but especially the orange *trans* compound, are now important organic pigments; they are marketed by several companies.

C.I. Pigment Orange 43, 71105 [*4424-06-0*] (**58**) is a brilliant orange pigment with a very high tinting strength, very good lightfastness, and weather resistance; good solvent and migration fastness; and outstanding thermal stability. It is particularly suitable for pigmenting plastics [e.g., PVC, polyolefins, and ABS (acrylonitrile – butadiene – styrene)], as well as for spin dyeing (e.g., of awnings and tents). In paints the pigment is used for powerful brightening effects and shading. Copper shades can be obtained by combining it with aluminum pigments.

C.I. Pigment Red 194, 71100 [*4216-02-8*] (**57**) is a red pigment with a high tinting strength but somewhat dull shade. Its lightfastness and weather resistance are outstanding, particularly in stoving enamels.

10. Perylene Pigments

The perylene pigments are based on 3,4,9,10-perylenetetracarboxylic acid diimides (**59**):

Like perinones (Chap. 9), the perylenetetracarboxylic acid diimides have long been known as vat dyes but have only recently become important pigments.

Production. Perylene pigments are generally prepared by heating perylenetetracarboxylic anhydride with primary aromatic or aliphatic amines in high-boiling solvents.

3,4,9,10-Perylenetetracarboxylic acid (**61**) is obtained by reacting 1,8-naphthalic acid imide

(**60**) with molten potassium hydroxide, followed by atmospheric oxidation to produce perylenetetracarboxylic acid diimide (**59**, R=H). The diimide is then hydrolyzed to give **61**. 1,8-Naphthalic acid imide (**60**) is synthesized by oxidizing acenaphthene (**54**) to 1,8-naphthalenedicarboxylic anhydride, which is then reacted with ammonia.

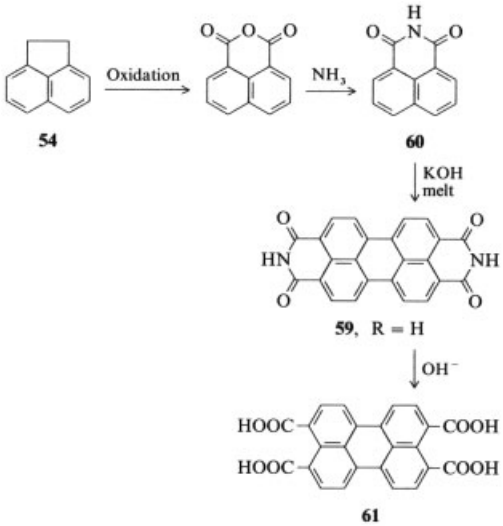

N,N′-dimethylperylimide is obtained from perylenetetracarboxylic acid diimide (**59**, R = H) by methylation.

Properties and Important Products. Perylene pigments have very good lightfastness, good to very good migration fastness, and high thermal stability. They are therefore used mainly to pigment plastics, but also as high-quality stoving enamels (e.g., for automobiles).

Commercial products are listed in Table 2. The most important of these are Pigment Red 149, 178, 179, and 190.

11. Anthraquinone Pigments

The term anthraquinone pigments denotes pigments that are anthraquinone derivatives or have a similar structure to anthraquinone (including higher condensed ring systems). With the exception of hydroxyanthraquinonesulfonic acid (see Section 11.1) and the anthraquinone azo pigments (see Section 11.2), all other compounds discussed in this chapter have long been used as vat dyes for dyeing textile fibers. Initial attempts to use these vat dyes as pigments generally produced dull shades and moderate tinting strengths. Vat dyes could only be used as pigments when methods became available for obtaining the required reproducible properties (e.g., higher chemical purity, optimum particle size and distribution, and appropriate crystal modifications).

The vat dyes were successfully converted into readily dispersible pigments by various aftertreatments including

1) Reduction to the leuco dye, removal of impurities, and reoxidation
2) Dissolution of the raw material (for example in sulfuric acid), precipitation in water or organic solvents, possibly with the addition of surfactants under carefully selected conditions
3) Thermal aftertreatment in water or organic solvents
4) Use of a wide range of dispersion and grinding processes (e.g., bead or ball mill grinding, comminution in the presence of salts or solvents)

Relatively few of the large number of vat dyes satisfy pigment requirements. Of these, only a small number have fastness standards approaching those of the phthalocyanine pigments. See also→ Anthraquinone Dyes and Intermediates, Chap. 3.1.4.

11.1. Hydroxyanthraquinone Pigments

Many derivatives of hydroxyanthraquinones and their sulfonic acids used to be commercially available as pigments, predominantly in the form of aluminum lakes. The number of these industrially important products has, however, fallen sharply. An example is *C.I. Pigment Violet 5:1, 58055:1* [*1328-04-7*]:

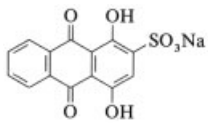

The commercially important form of this product is the 2:1 aluminum complex.

Table 2. Commercially available perylene pigments [36]

R (in formula **59**)	C.I. name, formula no.	CAS registry number	Shade
H	C.I. Pigment Violet 29, 71129	[*12236-71-4*]	red to maroon
H_3C	C.I. Pigment Red 179, 71130	[*5521-31-3*]	red to maroon
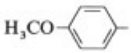	C.I. Pigment Red 190, 71140	[*6424-77-7*]	bluish red
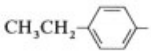	C.I. Pigment Red 123, 71145	[*24108-89-2*]	scarlet to red
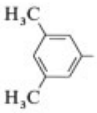	C.I. Pigment Red 149, 71137	[*4948-15-6*]	red
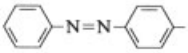	C.I. Pigment Red 178, 71155	[*3049-71-6*]	red
N – R = O	C.I. Pigment Red 224, 71127	[*128-69-8*]	bluish red

The hydroxyanthraquinones 1,2-dihydroxyanthraquinone (alizarine), 1,4-dihydroxyanthraquinone (quinizarin), and 1,2,4-trihydroxyanthraquinone (purpurin), some of which are metal-complex pigments. Alizarine, for example, was sold as the aluminum calcium salt of 1,2-dihydroxyanthraquinone [37].

11.2. Aminoanthraquinone Pigments

For the preparation of aminoanthraquinones, see → Anthraquinone Dyes and Intermediates, Chap. 2.5.

Several commercially available compounds were synthesized by acylation or benzoylation of monoamino- or diaminoanthraquinones or their hydroxy derivatives. Very few pigments in this range are still industrially important.

Anthraquinone azo pigments and anthraquinone pyrimidines are condensation products of aminoanthraquinone and are also discussed here.

Acylaminoanthraquinones. The reddish-yellow *C.I. Pigment Yellow 147, 60645* [*76168-75-7*] is obtained by condensation ("acylation") of 1-aminoanthraquinone with 2,4-dichloro-6-phenyl-1,3,5-triazine.

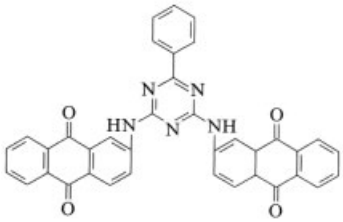

Aminoanthraquinones with Free Amino Groups. *C.I. Pigment Red 177, 65300* [*4051-63-2*] is used in high-quality paints, particularly for automobiles.

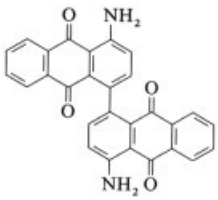

Anthraquinone Azo Pigments. Red to yellow pigments with, in some cases, extremely good general fastness are obtained by diazotization, especially of 1-aminoanthraquinone (→ Anthraquinone Dyes and Intermediates, Chap. 2.5.3.), followed by coupling with various components. Use of coupling components containing additional amide groups improves the migration fastness of the resulting

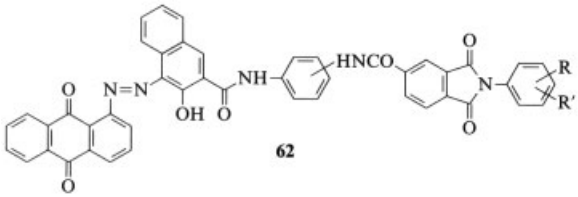

62

pigments. Red pigments (**62**) are obtained by coupling with Naphtol AS derivatives [38] (see above).

If barbituric acid is used as the coupling component, a yellow pigment (**63**) is obtained [39]:

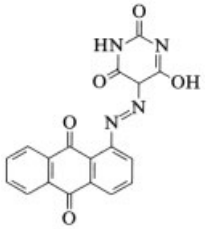

63

Orange or red pigments [40] with good fastness properties are synthesized by coupling diazotized 1-aminoanthraquinone with pyrazoloquinazolines of the formula:

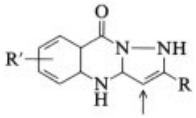

An example of such a pigment is *C.I. Pigment Red 251, 12925* [*71552-60-8*].

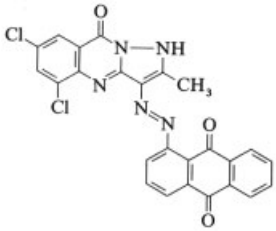

11.3. Heterocyclic Anthraquinone Pigments

11.3.1. Anthrapyrimidine Pigments

The anthrapyrimidine pigments are based on 1,9-anthrapyrimidine (→ Anthraquinone Dyes and Intermediates, Chap. 2.12.3.), which can be obtained from 1-aminoanthraquinone:

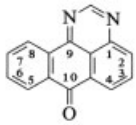

1,9-Anthrapyrimidine

C.I. Pigment Yellow 108, 68420 [*4216-01-7*], for example, is a greenish-yellow pigment with very good lightfastness and weather resistance. It is obtained by condensing 1-aminoanthraquinone with 1,9-anthrapyrimidine-2-carboxylic acid chloride. The pigment has long been known as a vat dye.

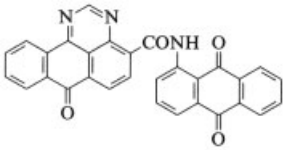

11.3.2. Indanthrone and Flavanthrone

Indanthrone (→ Anthraquinone Dyes and Intermediates). *Indanthrone, C.I. Pigment Blue 60, 69800* [*81-77-6*] was discovered in 1901 by R. Bohn and is one of the oldest vat dyes.

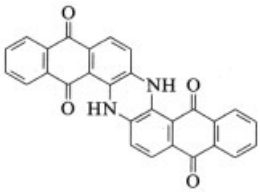

"Indanthren", the original compound name, was employed as a generic term for very high-quality

vat dyes. Subsequently, the compound was also used as a pigment.

Indanthrone is produced on an industrial scale mainly by treating 2-aminoanthraquinone in an alkali melt. Typically, the compound is heated to 210 – 225 °C in a mixture of potassium hydroxide and sodium hydroxide in the presence of sodium nitrate or potassium chlorate and potassium acetate or formate.

The pigment exists in four crystal modifications. The reddish-blue α-phase is of special interest; it has good thermal stability and very good light- and solvent fastnesses, approaching those of the phthalocyanines.

Because of its high cost, indanthrone is used mainly for high-quality paints and plastics, especially polyolefins. Its importance is limited, however, due to the availability of inexpensive phthalocyanine pigments with excellent fastness properties (→ Phthalocyanines).

Flavanthrone (→ Anthraquinone Dyes and Intermediates, Chap. 2.12.7.). *Flavanthrone yellow, C.I. Pigment Yellow 24, 70600* [*475-71-8*] (**64**) was first described in 1901 by R. Bohn.

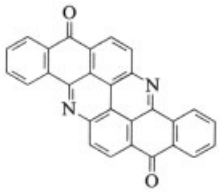

64

It attracted interest as a pigment when the demand for paints with higher lightfastness and weather resistance increased.

Flavanthrone is synthesized by reacting 1-chloro-2-aminoanthraquinone with phthalic anhydride to give 1-chloro-2-phthaloylaminoanthraquinone. The latter is cyclized to flavanthrone by heating with copper powder in trichlorobenzene at 180 – 200 °C, followed by treatment with excess sodium hydroxide.

The pigment can also be prepared in a one-stage process by heating 1-chloro-2-aminoanthraquinone with copper in a strongly polar aromatic solvent [41].

Various methods are available for purifying the crude pigment, for example:

1) Dissolution in concentrated sulfuric acid followed by hydrolysis
2) Conversion to the leuco form and reoxidation
3) Extraction with dipolar aprotic solvents

The pigment has a somewhat dull, reddish-yellow color, high tinting strength, and very good lightfastness and weather resistance, even in light shades.

11.4. Polycarbocyclic Anthraquinone Pigments

This section deals with polycyclic systems that do not contain hetero atoms (nitrogen). Very few systems are important as pigments.

11.4.1. Pyranthrone Pigments

The pyranthrone structure is related to that of flavanthrone (**64**), the only difference being that the nitrogen atoms of the latter are replaced by carbon atoms. Pyranthrone (pyranthronequinone, **65**) is classified as *C.I. Pigment Orange 40, 59700* [*128-70-1*]:

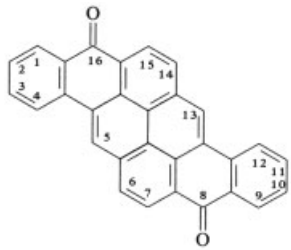

65

Pyranthrone is prepared from 1-chloro-2-methylanthraquinone (**66**), which is reacted with copper powder in boiling 1,2-dichlorobenzene to give 2,2′-dimethyl-1,1′-dianthraquinonyl (**67**). The latter is cyclized to pyranthrone by boiling for several hours with 50 % potassium hydroxide solution in isobutanol:

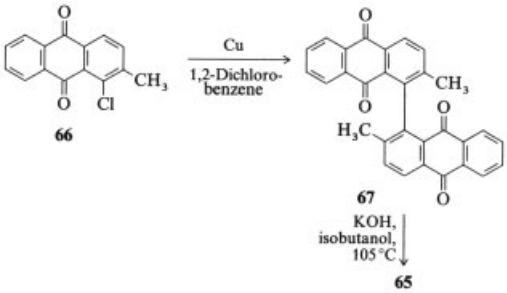

Unsubstituted pyranthrone and halogenated pyranthrone derivatives have yellowish-red to red hues with a high tinting strength. Lightfastness and weather resistance are improved by halogenation; tribromopyranthrone, for example, has substantially greater fastness than dibromopyranthrone. The halogenated pyranthrone pigments are mainly used in high-quality paints and coating materials, and also in combination with inorganic or aluminum pigments, especially for automotive finishes.

Halogenation is carried out with chlorine or bromine, preferably in chlorosulfonic acid in the presence of sulfur, iodine, or antimony as catalyst. The chlorine or bromine atoms are introduced in various positions of the pyranthrone ring. Pyranthrones with two to four halogen atoms per molecule can be obtained, depending on the quantity of halogen used [42], [43].

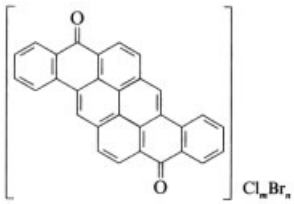

Cyclization of 3,3′-dichloro-2,2′-dimethyl-1,1′-dianthraquinonyl in polar organic solvents at 110 – 150 °C in the presence of alkali acetates yields 6,14-dichloropyranthrone [43]. Bromination of the latter yields bromodichloropyranthrones containing ca. 0.5 – 2.2 bromine atoms per molecule, depending on the quantity of bromine and the reaction conditions. Different shades are produced according to the degree of bromination [44].

Commercially available pigments include 2,10-dichloropyranthrone (*C.I. Pigment Orange 51* [*73309-48-5*]), dibromo-4,6-dichloropyranthrone (*C.I. Pigment Red 226* [*63589-04-8*], $m = 2$, $n = 2$); and tribromopyranthrone (*C.I. Pigment Red 216, 59710* [*1324-33-0*], $m = 0$, $n = 3$).

11.4.2. Anthanthrone Pigments

Anthanthrone (**68**) has the following structure:

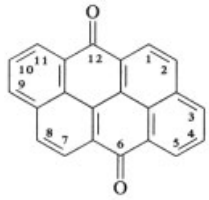

Only halogenated anthanthrones are used as pigments. Of particular interest is the yellowish-red (scarlet) pigment 4,10-dibromoanthanthrone, *C.I. Pigment Red 168, 59300* [*4378-61-4*]. This compound has a high tinting strength with extremely good lightfastness and weather resistance. It is mainly used to pigment high-quality industrial paints, but also to a lesser extent in the plastics sector.

Anthanthrone is synthesized from naphthostyryl (**69**), which is hydrolyzed to give 1-aminonaphthalene-8- carboxylic acid (**70**). The latter is diazotized and boiled in the presence of copper powder to yield 1,1′-dinaphthyl-8,8′-carboxylic acid (**71**), which is finally cyclized to anthanthrone (**68**) with sulfuric acid. Dibromoanthanthrone can be obtained directly by bromination of anthanthrone in the presence of iodine.

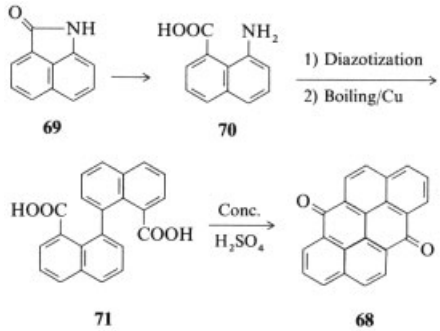

12. Diketopyrrolopyrrole (DPP) Pigments

A new type of heterocylic pigments based on a symmetrical chromophore, namely, the 1,4-diketopyrrolo[3,4-*c*]pyrrole system (**72**)

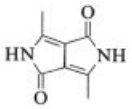

72

was discovered in the early 1980s and systematically developed into a group of red pigments (DPP pigments) with good fastness properties [45], [46].

The DPP pigments are synthesized by reacting succinate esters with benzonitriles in methanol in the presence of sodium methoxide:

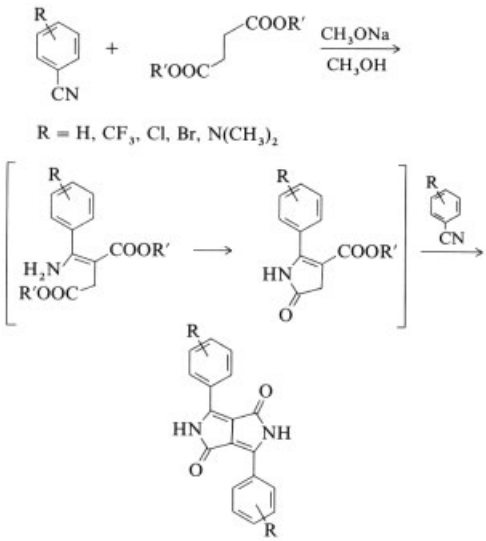

The particle size and associated properties can be controlled either by solvent aftertreatment or by synthesis conditions. A second benzonitrile derivative can be added during condensation; this mixed synthesis yields pigment types that are more transparent and have a higher tinting strength.

The pigments are extremely lightfast and weather resistant, with a good thermal stability and resistance to bleeding and chemicals. *C.I. Pigment Red 254* is commercially available.

13. Thioindigo Pigments

See also → Indigo and Indigo Colorants, Chap. 13..

Indigo was important at one time as a pigment for coloring rubber but is now seldom used for this purpose. Some thioindigo derivatives became important organic pigments in the 1950s. Thioindigo itself was of some importance as a vat dye for a brief period.

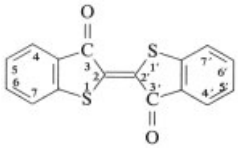

Thioindigo derivatives substituted in the aromatic ring with halogen (in particular chlorine) or methyl groups were first described about 80 years ago and were used up to the 1950s, mainly as vat dyes. Thioindigo derivatives became important as red and maroon pigments when suitable aftertreatment of the crude products had been developed.

Of the few thioindigo pigments still on the market, only the maroon *C.I. Pigment Red 88*, *73312* [*14295-43-3*] (4,7,4′,7′-tetrachlorothioindigo, **76**) is of any importance:

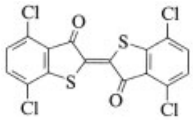

76

This compound has the best fastness properties of all known thioindigo pigments; it has a high lightfastness and weather resistance, even in reduced shades, and a very good migration fastness. The pigment is used in industrial paints, for example, in automotive finishes where it is often combined with inorganic red pigments (molybdate red, iron oxide), and in plastics.

C.I. Pigment Red 88 is obtained by reacting 2,5-dichlorothiophenol (**73**) with chloroacetic acid to form 2,5-dichlorophenylmercaptoacetic acid (**74**). The latter reacts with chlorosulfonic acid with ring closure to form thioindoxyl (**75**), which can be oxidized in an alkaline medium to form 4,7,4′,7′-tetrachlorothioindigo (**76**):

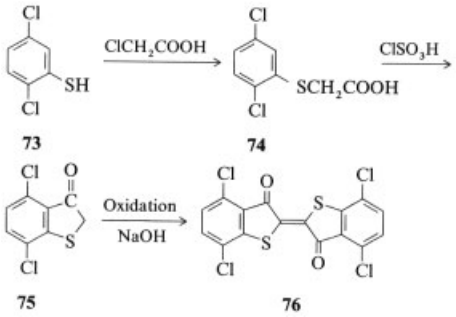

Various ring closure modifications have been described. For example, instead of (substituted) phenylmercaptoacetic acid, its 2- carboxy or 2-amino derivative can be cyclized to thioindoxyl. In the first case, cyclization is carried out in an alkali melt. In the second, cyclization is achieved by diazotization, Sandmeyer reaction with sodium cyanide/copper(I) cyanide, alkaline hydrolysis, and treatment with acid.

C.I. Pigment Red 181, 73360 [*2379-74-0*], 4,4′-dimethyl-6,6′-dichlorothioindigo, is of limited importance for coloring special plastics and cosmetics.

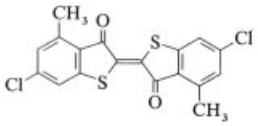

14. Dioxazine Pigments

See also → Azine Dyes, Chap. 5.2.

The parent compound of the dioxazine series is 9,10-dichlorotriphenedioxazine, which itself has no importance as a pigment:

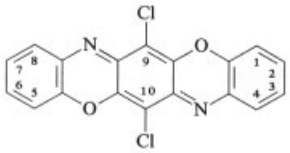

Derivatives with substituents in the two peripheral aromatic rings are of considerable technical interest.

To prepare dioxazine derivatives, chloranil (**77**) is first reacted with aromatic amines in the presence of acid-binding agents (e.g., sodium acetate) to form the 2,5-diarylamino-3,6-dichlorobenzoquinone (e.g., **78**). The corresponding dichlorodiphenedioxazine is obtained by oxidative cyclization of the latter in a high-boiling solvent in the presence of a catalyst (e.g., aluminum chloride or benzenesulfochloride) that lowers the cyclization temperature.

Ring closure is facilitated if 2-alkoxyanilines are used instead of anilines. The oxazine ring is then formed with the elimination of alcohol.

One compound of the dioxazine series has become especially important as a pigment, namely, *C.I. Pigment Violet 23, 51319* [*6358-30-1*] (**79**). The compound was first described as a pigment in 1952 by Hoechst [47]; it was originally used in the sulfonated form as a lightfast, substantive dye. Procedures such as precipitation with chloroacetic acid, phosphoric acid, or *p*-toluenesulfonic acid; salt grinding; or kneading in the presence of an organic solvent were necessary to obtain the required pigment properties.

Three-dimensional X-ray analysis revealed an angular structure (**79**). The pigment exists in two crystal modifications [35].

Pigment Violet 23 has a blue to reddish-violet color and a high tinting strength. It is extremely lightfast and resistant to weathering, and is used in nearly all media, especially to shade the color of phthalocyanine blue pigments.

The pigment is synthesized according to the reaction pathway described above:

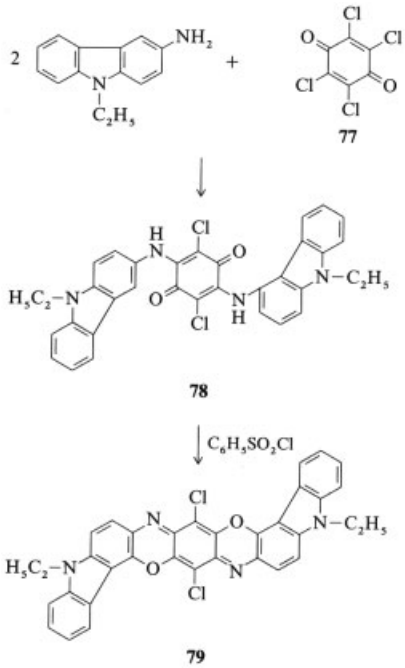

Condensation of chloranil with 3-amino-*N*-ethylcarbazole in 1,2-dichlorobenzene in the presence of sodium acetate at 55 – 70 °C yields the dichlorotriphenedioxazine derivative (**78**). The latter is then cyclized at 175 – 180 °C in a Friedel – Crafts reaction in the presence of benzenesulfochloride.

Other dioxazine pigments that are produced industrially on a limited scale contain acetyl or benzoylamino groups in the 2,6-positions and ethoxy groups in the 3,7-positions. The acylamino groups may also replace the chlorine atoms in the 9,10-positions, they are especially important for improving migration fastness.

An example of a dioxazine pigment is *C.I. Pigment Violet 37, 51345* [*57971-98-9*], which is prepared by Friedel – Crafts condensation of 2,5-dichloro-3,6-bisacetylamino-1,4-benzoquinone with 1,4-diethoxy-2-amino-5-benzoylaminobenzene in alcohol in the presence of sodium acetate:

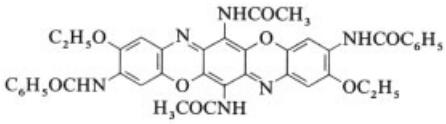

Subsequent ring closure occurs without oxidation by the elimination of alcohol.

15. Triphenylmethane Pigments

Two groups of triphenylmethane (triarylcarbonium) compounds are important as pigments. They are both salts of basic dyes: (1) inner salts of triphenylmethanesulfonic acids and (2) complex salts with inorganic heteropoly acids. Triarylmethane dyes are discussed in detail elsewhere (→ Triarylmethane and Diarylmethane Dyes). Only the two types most important as pigments are briefly discussed here. All interesting triphenylmethane pigments are blue, violet, or green. They are only moderately fast with respect to alkali, but have good fastness with respect to water, oil, acid, and hydrocarbons. Because of their high tinting strength and brilliance they are used for printing, mainly for shading black printing inks, especially sheet and roller offset inks.

15.1. Free Sulfonic Acids ("Alkali Blue" Types)

In contrast to the water-soluble sodium salts of the triphenylmethanesulfonic acids used for dyeing wool, the free sulfonic acids of 4,4′,4″-trisaminophenylmethane and of phenylated "pararosaniline" are generally insoluble in water. The reddish diphenyl derivatives (**80**) and the greenish triphenyl derivatives (**81**) are important pigments. Reddish-blue pigments (**82**) are also used.

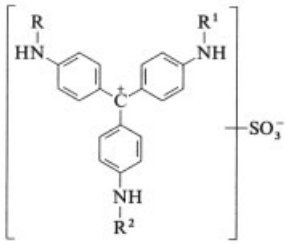

R, R¹, R² = H, C_6H_5, C_6H_5 **80**
= C_6H_5, C_6H_5, C_6H_5 **81**
= H, H, C_6H_5, **82**

The compounds are prepared from aniline and *p*-toluidine by acid-catalyzed phenylation with arylamines followed by sulfonation. Alternatively, 4-chlorobenzotrichloride is reacted with chlorobenzene in a Friedel – Crafts reaction to give 4,4′,4″-trichlorotriphenylmethyl chloride. The chlorine substituents of the latter are then replaced by arylamines, and the product is sulfonated [48]. The green shade increases with the number of phenyl groups in the molecule. One sulfonic acid group per molecule is the optimum number; at higher degrees of sulfonation tinting strength decreases and oilfastness deteriorates appreciably.

15.2. Dye Salts with Complex Anions

The triphenylmethane dyes for pigment formation were originally precipitated with tannin and potassium antimonyl tartrate on inorganic substrates (e.g., hydrated alumina, $CaSO_4$, or $BaSO_4$). These pigments had a poor lightfastness. To improve lightfastness and obtain satisfactory brilliance and tinting strength, complex salts were formed with inorganic het-

eropoly acids: phosphotungstic acid, and especially phosphomolybdic and phosphotungstomolybdic acids.

Two types of triphenylmethane dyes are still important for complex formation with the above-mentioned acids which are deprotonated to form anions (A^-), namely, **83**, which produces violet shades, and **84**, which produces blue shades.

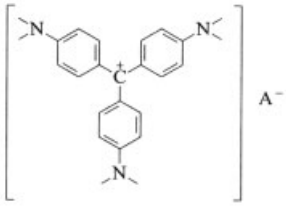

83

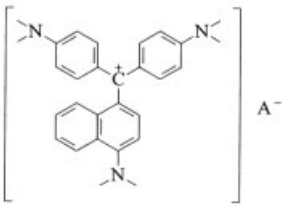

84

The nitrogen atoms carry either alkyl (e.g., methyl or ethyl) groups, aryl groups, or hydrogen and alkyl or aryl groups.

Xanthene derivatives (rhodamines) are derived from the triphenylmethanes and are also important pigments:

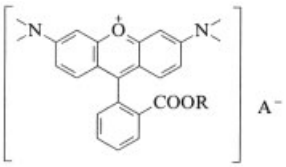

R = H, CH_3, C_2H_5

16. Quinophthalone Pigments [49], [50]

See also → Methine Dyes and Pigments.

Quinophthalone pigments are obtained by fusing quinaldine with phthalic anhydride:

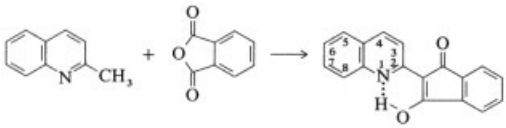

The structure of the enol form and the hydrogen bond were determined by spectroscopy [50].

Quinophthalones containing sulfonyl or carboxyl substituents are important anionic dyes. Quinophthalones with side chains containing basic quaternary amine groups are important as cationic dyes and especially as disperse dyes.

Lightfastness is considerably improved by introducing a hydroxyl group into the 3-position of the quinophthalone molecule (→ Methine Dyes and Pigments).

Attempts have been undertaken to make quinophthalones more solvent- and migration-fast by increasing the size of the quinophthalone molecule through introduction of appropriate substituents, condensation of quinaldines with pyromellitic dianhydride, or "dimerization" of the molecule via a diamide bridge, thus forming quinophthalone pigments. Suitable substituents include acylamino groups and, especially, chlorine or bromine atoms. Introduction of halogen atoms into the phthaloyl residue increases lightfastness, whereas halogen substitution in the phenyl moiety of the quinoline ring decreases lightfastness.

Quinophthalone pigments have yellow to red shades and are used for the bulk dyeing of plastics (e.g., PVC or polyolefins) and rubber, as well as for pigmenting coating materials and printing inks.

Examples. *C.I. Pigment Yellow 138, 56300* [*56731-19-2*], a clear greenish-yellow pigment, is obtained by reacting 8-aminoquinaldine with two equivalents of tetrachlorophthalic anhydride [51]:

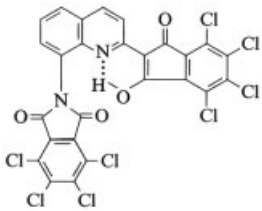

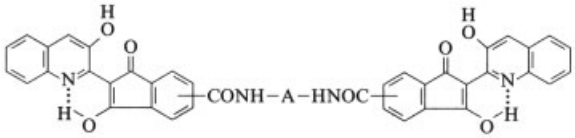

A red bisquinophthalone pigment is obtained from pyromellitic anhydride and two equivalents of 3-hydroxyquinaldine-4- carboxylic acid [52] (see above).

17. Fluorescent Pigments [53]

Fluorescent daylight pigments not only reflect light, but also emit a substantial proportion of the light they absorb (→ Fluorescent Dyes, Chap. 12. The wavelength of the emitted light is longer than that of the absorbed light. The term light in this context denotes electromagnetic radiation with a wavelength of ca. 300 – 700 nm (longer-wave UV and visible regions). The color of a fluorescing sample thus consists of a combination of reflected light (intrinsic color) and emitted light. Fluorescent pigments are therefore much more brilliant than normal pigments and have a much higher luminosity.

For fluorescence to occur in organic dyes (i.e., the release of a proportion of the absorbed energy as light), the following conditions must be satisfied:

1) Existence of a planar, delocalized π-electron system that requires only a small energy input to convert it to an excited state.
2) A molecular conformation with maximum rigidity to prevent vibrations that would otherwise lead to radiationless deactivation without emission of light.
3) Most industrially important substances only show fluorescence in the dissolved state. Therefore, the dyes must form dilute solutions because fluorescence is extinguished at high concentrations ($\geq$ ca. 10 %).

Colorants that are not dissolved in the application medium (i.e., pigments) cannot therefore generally be used in the pure state as fluorescing systems. In the mid-1940s, however, Switzer in the United States prepared solid solutions of fluorescent dyes in urea – formaldehyde resins [54] that were suitable as fluorescent pigments.

Of the numerous known fluorescing systems, only a few are used for the industrial production of fluorescent pigments. In the yellow range these include derivatives of 4-aminonaphthalic arylimide, benzoxanthene, benzothioxanthene, coumarin, and naphthaldazine. Rhodamines in particular are suitable for red fluorescent pigments. For more detailed descriptions of these compounds, see → Fluorescent Dyes, → Optical Brighteners.

4-Aminonaphthalic Arylimide Derivatives. An example is *C.I. Solvent Yellow 44, 56200* [*2478-20-8*]. Nitration and oxidation of acenaphthene yields 4-nitronaphthalic anhydride which is reduced and converted to Solvent Yellow 44 by heating with *m*-xylidine.

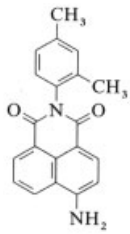

Benzoxanthene (85, A = O) and benzothioxanthene (85, A = S) derivatives are synthesized, for example, by ring closure of the diazonium salts of 4-(2′-aminophenoxy)naphthalic anhydride or of the corresponding imides in the presence of copper salts at 60 – 100 °C [55].

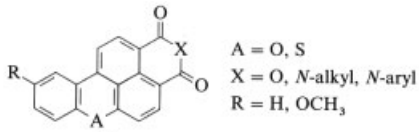

85

Coumarin Derivatives (→ Fluorescent Dyes, Chap. 3.). The fluorescent dye **86** is prepared by reacting 2- cyanomethylbenzimidazole with 4-diethylamino-2-hydroxybenzaldehyde in ethanol in the presence of piperidine, followed by boiling with 2 % hydrochloric acid [56].

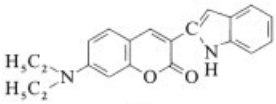

86

Naphthaldazine Derivatives. *C.I. Pigment Yellow 101, 48052* [*2387-03-3*] (**87**) is synthesized from 2-hydroxy-1-naphthaldehyde and hydrazine. It is the only known fluorescent pigment that can be used directly without being embedded in resin.

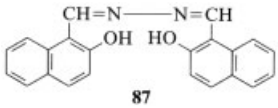

87

Rhodamines. *C.I. Basic Violet 10, 45170* [*81-88-9*] (Rhodamine B, **88**) is prepared by condensing phthalic anhydride with 1,3-diethylaminophenol at 180 °C in the presence of zinc (II) chloride.

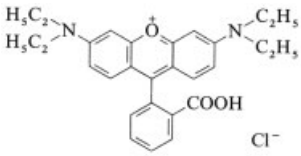

88

Production. Embedding fluorescent dyes in resins produces brittle, strongly fluorescent substances that can be comminuted by grinding but still have a relatively large particle-size distribution (ca. 1 – 30 µm). Fluorescent dyes are now generally embedded in clear, colorless matrix resins that are synthesized by polycondensation of aminotriazines (e.g., benzoguanamine, melamine, or urea) with aromatic sulfonamides and formaldehyde [57] (→ Amino Resins). Urethane resins have also been described, which are obtained by polyaddition of aliphatic diisocyanates with amides and amines [58]. Fluorescent pigments have a much better solvent resistance and lightfastness than fluorescent dyes used in solution.

General Synthesis. The aromatic sulfonamide is heated in a stirred vessel until it melts. Formaldehyde is then slowly added, and a clear, colorless resin is formed. The amine (e.g., benzoguanamine, melamine, urea) is simultaneously added for the cocondensation. The fluorescent dye is then dissolved in the molten resin. With yellow shades the dye concentration may reach ca. 10 %; with orange and red shades, only ca. 2 %. At 150 – 175 °C the hot liquid is poured onto metal sheets and cooled. The clear, brittle resin is comminuted to give an average particle size of 5 – 6 µm.

Uses. Daylight fluorescent pigments are used to color plastics, mainly nonpolar paint binder systems, and printing inks, as well as for textile printing. The pigments are transparent and therefore mainly applied to white substrates. To increase the range of shades, up to three fluorescent pigments (generally premixed as dyes) are often used. Important areas of use include the marking of warning and safety equipment and appliances (e.g., traffic signs; rescue, police, and fire department vehicles). The coloring of plastics and the printing of fluorescent pigments on paper, cardboard, or textiles are very important in the advertising sector.

18. Environmental Aspects

Organic pigments have an extremely low toxicological and ecological risk potential due to their physicochemical properties, particularly their extremely low solubility in water and application media [59]. As with all chemicals, human safety, removal of contamination from emissions and effluents, and proper disposal of waste are of paramount importance in the production and use of pigments.

18.1. Toxicology

Instructions on safe handling published in safety data sheets are based on the results of toxicological and ecological tests.

Testing of hundreds of organic pigments showed that they have extremely high oral LD_{50} values. Very few cause irritation of the skin or mucous membranes.

Numerous pigments have been screened for mutagenic effects in the Ames test [60]. This test was positive in only a few cases involving very small molecules that were slightly soluble in solvents (e.g., dimethyl sulfoxide) used in the Ames test.

To screen for carcinogenic potential, long-term animal feeding tests have to be performed (generally on rats and mice). No studies are known in which a significant carcinogenic potential has been detected for organic pigments. Animal experiments on dichlorobenzidine pigments, for example, showed that the compound was not reductively cleaved to the carcinogenic compound 3,3′-dichlorobenzidine. Furthermore, none of the epidemiological investigations revealed any toxic effects in humans [36]. Long-term feeding tests on Pigment Yellow 12, 13, 16, and 86 did not reveal a carcinogenic effect. The pigments did not exhibit any bioavailability (i.e., were not metabolized) [61].

18.2. Ecology

Organic pigments may be classified as inert dusts because of their sparing solubility in water and the solvents normally employed in their application media. The MAK value specified for fine dusts is 6 mg/m^3. Air, particularly at the workplace, must be kept substantially free from chemicals (pigments) by appropriate removal and separation procedures (filtration, adsorption). Possible exposure to pigment dusts must be prevented by protection equipment (respirators). Contamination of open water with organic pigments can also be avoided by filtration, sedimentation, adsorption, or possibly biological treatment of the effluent. Toxic effects on fish stocks are extremely unlikely because of the sparing solubility of organic pigments, and none have been described.

Organic pigments are not biologically degradable under aerobic conditions but probably decompose slowly under anaerobic conditions.

18.3. Impurities

Pigment impurities (e.g., traces of aromatic amines, heavy metals, or polychlorinated biphenyls, PCBs) may affect the results of toxicological investigations.

Only traces of the aromatic amines used for the synthesis of azo pigments should be present in the pigments. Limiting values for these amines have been laid down for certain areas of application; for example, a limit of 500 ppm has been specified for pigments in food packaging materials.

Traces of heavy metals in organic pigments are generally considerably below official limiting values. Appropriate limits of heavy metals in organic pigments must be obeyed in consumer goods, cosmetics, and toys [62].

In some countries (e.g., in the EC and the United States), strict regulations have been placed on PCBs because of their persistence in the environment. Permissible concentrations in the EC are 50 ppm, and in the United States 25 ppm.

Because of the insolubility and migration fastness of most organic pigments, no particular problems arise in handling and application. Regular tests should be carried out on colored materials and the pigments themselves to protect consumers and ensure that traces of pigments do not migrate from the printed or colored packaging (e.g., into food), and to confirm that the stringent purity requirements and other legal regulations are being observed [63]. Safety standards include EN 71.3, Safety of Toys (1988) for the EC and ASTM F963-86, Standard Consumer Safety Specifications on Toy Safety for the United States.

19. References

General References

1. *Colour Index,* 3rd ed. with supplements, Society of Dyers and Colourists and Americam Association of Textile Chemist and Colorists, Bradford 1971 – 1988.
2. W. Herbst, K. Hunger: *Industrielle Organische Pigmente,* VCH Verlagsgesellschaft, Weinheim 1987.
3. P. A. Lewis (ed.): *Pigment Handbook,* 2nd ed., vols. 1 – 3, J. Wiley & Sons, New York 1988.

4. J. Lenoir: "Organic Pigments," in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. V, Academic Press, New York 1971, pp. 313 ff.
5. D. Patterson: *Pigments,* Elsevier, Amsterdam 1967.
6. P. Rys, H. Zollinger: *Farbstoffchemie,* 3rd ed., Verlag Chemie, Weinheim 1982.
7. F. Kehrer, *Chimia* **28** (1974) no. 4, 173.
8. H. Gaertner, *J. Oil. Colour Chem. Assoc.* **46** (1963) 13.
9. V. C. Vesce, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **28** (1956) no. 377, part I.
10. V. C. Vesce, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **31** (1959) no. 419, part II.
11. N. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* 4th ed., Akademie-Verlag, Berlin 1966.

Specific References

12. DIN 53 206:Teilchengrößenanalyse, Grundbegriffe.
13. BIOS 1661/I, 52 (1946).
14. BIOS 1661/II, 134 (1946).
15. BIOS 1661/II, 164 (1946).
16. C. J. Brown, *J. Chem. Soc. A* 1967, 405. H.-C. Mez, *Ber. Bunsen-Ges. Phys. Chem.* **72** (1968) 389. K. Hunger, E. F. Paulus, D. Weber, *Farbe + Lack* **88** (1982)455 – 458. E. F. Paulus, W. Rieper, D. Weber, *Z. Kristallogr.* **165** (1983)137 – 149. A. Whitaker, *Z. Kristallogr.* **166** (1984) 177 – 188.
17. C. T. Grainger, J. F. McConnell, *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 25** (1969) 1962.
18. D. Kobelt, E. Paulus, W. Kunstmann, *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 28** (1972) 1319. D. Kobelt, E. Paulus, W. Kunstmann, *Z. Kristallogr.* **139** (1974) 15. A. Whitaker, *Z. Kristallogr.* **145** (1977) 271; **146** (1977) 173; **147** (1978) 99. E. F. Paulus, K. Hunger, *Farbe + Lack* **86** (1980) 116 – 120. E. F. Paulus, *Z. Kristallogr.* **160** (1982) 235 –243.
19. BIOS 1661/I, 87 (1946). BIOS 1661/II, 150 (1946).
20. E. Dietz, O. Fuchs, *Farbe + Lack* **79** (1973) no. 11, 1058.
21. Hoechst, DE 1 188 229, 1960 (K. Schilling, E. Dietz); DE 1 215 839, 1961 (K. Schilling, E. Dietz); DE 1 216 460, 1961 (G. Semler, K. Schilling); DE 1 217 008, 1961 (K. Schilling, E. Dietz).
22. Bayer, DE-OS 2 064 093, 1970 (K.-H. Schündehütte).
23. G. Stephan, *Int. Farbensymp. 6th,* Freudenstadt 1976.
24. M. Regitz, *Justus Liebigs Ann. Chem.* **676** (1964) 101. M. Regitz, *Angew. Chem.* **79** (1967) 786; *Angew. Chem. Int. Ed. Engl.* **6** (1967) 733. W. Ried, B. Peters, *Justus Liebigs Ann. Chem.* **729** (1969) 119.
25. BASF, DE-OS 1 544 404, 1966 (P. Dimroth, A. Haag).
26. G. Henning, *Int. Farbensymp. 6th,* Freudenstadt 1976.
27. BASF, DE-OS 1 252 341, 1966 (P. Dimroth); DE-OS 1 569 666, 1967 (P. Dimroth, H. Hiller).
28. A. Pugin, J. V. D. Crone, *Farbe + Lack* **72** (1966) 206.
29. *Off. Gaz. Jpn. Pat. Off.,* Nov. 17, 1977.
30. BASF, DE-AS 2 041 999, 1970 (G. Bock, W. Elser); DE-AS 2 628 409, 1976 (W. Lotzsch).
31. BASF, EP 132 818, 1983 (W. Lotzsch, G. Henning).
32. S. S. Labana, L. L. Labana, *Chem. Rev.* **67** (1967) no. 1, 1. P. Pollak, *Chimia* **30** (1976) 357. P. Pollak, *Prog. Org. Coat.* **5** (1977) 245.
33. H. Liebermann, *Justus Liebigs Ann. Chem.* **518** (1935) 245.
34. W. F. Spengemann, *Paint Varn. Prod.* **60** (1970) 37.
35. E. Dietz, *11th International Color Symposium,* Montreux 1991.
36. H. W. Gerarde, D. F. Gerarde, *JOM J. Occup. Med.* **16** (1974) 332 – 324. T. Gadian, *Chem. Ind.* **19** (1975) 821 – 830. I. Mac Intyre, *JOM J. Occup. Med.* **17** (1975) 23 – 26.
37. E. G. Kiel, P. M. Heertjes, *J. Soc. Dyers Colour.* **79** (1963) no. 1, 21.
38. BASF, DE-OS 2 059 677, 1970 (G. Ruider, P. Dimroth).
39. BASF, DE-AS 1 544 372, 1965 (W. Braun, I. Paetzke); DE-AS 1 544 374, 1965 (W. Braun, I. Paetzke).
40. BASF, DE-OS 2 843 873, 1978 (H. Kantar).
41. Ciba-Geigy, DE-OS 2 105 286, 1970 (M. Grelat).
42. IG-Farbenind., DE 563 997, 1928.
43. BASF, DE 2 007 848, 1970 (F. Graser, G. Bock).
44. BASF, DE-AS 2 017 450, 1970 (F. Graser).
45. A. Iqbal et al., *J. Coat. Technol.* **60** (1988) 37.
46. Ciba Geigy, US 4 415 685, 1983 (A. Iqbal, L. Cassar).
47. Hoechst, DE 946 560, 1952 (R. Sedlmayr).
48. Hoechst, DE 1 919 724, 1969 (T. Papenfuhs, E. Spietschka).

49. B. K. Manukian, A. Mangini, *Chimia* **24** (1970) 328.
50. A. Patil, G. Patkar, T. Vaidyanathan, *Bombay Technol.* **24** (1974) 26.
51. BASF, DE-AS 1 770 960, 1968 (J. Dehnert).
52. Sandoz, CH 463 667, 1965 (B. K. Manukian).
53. H.-P. Beyerlin, *DEFAZET Dtsch. Farben Z.* **28** (1974) no. 7, 302. C. D. Dane, *Chem. Ber.* **13** (1977) no. 9, 335.
54. Switzer Brothers, US 2 498 592, 1945.
55. Hoechst, DE 1 569 761, 1967 (O. Fuchs, H. Tröster).
56. Geigy, DE-AS 1 098 125, 1959 (H. Häusermann, J. Voltz). BASF, DE-OS 2 013 262, 1970 (J. Dehnert, G. Grau).
57. Switzer Brothers, US 2 809 954, 1954 (Z. Kazenas); US 2 938 873, 1958 (Z. Kazenas).
58. Siegle, DE 2 013 393, 1970 (H.-P. Beyerlin).
59. A. C. D. Cowley, *Polym. Paint Colour J.* (1985) 7 – 21.
60. B. N. Ames, J. McCann, E. Yamasaki, *Mutat. Res.* **31** (1975) 347 – 364.
61. F. Leuschner, *Toxicol. Lett.* **2** (1978) 253 – 260; "Bioassay of Diarylide Yellow for Possible Carcinogenity," *DHEW Publ. (NIH) (U.S.)* **77 – 830** (1977).
62. Ciba-Geigy: "Farbmittel für die Färbung von Lebensmittelverpackungen, Spielzeugen und Bedarfsgegenständen," technical report no. 3D.
63. Bundesgesundheitsamt (German Health Office), 178th report, Aug. 15, 1988.

Plant Growth Regulators

Weed Control is a separate keyword.

Hubert Sauter, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Bernd Zeeh, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Classes of Growth Regulators 3
2.1. Auxins and Auxin-Type Compounds . 3
2.2. Antiauxins 4
2.3. Gibberellins 4
2.4. Gibberellin Antagonists 5
2.5. Cytokinins 8
2.6. Abscisic Acid 8
2.7. Ethylene-Releasing Compounds 8
2.8. Other Growth Regulators 9
3. Defoliants and Desiccants 10
4. References 11

1. Introduction

Plant growth regulators, also known as bioregulators, are natural or synthetic substances that act in small amounts to influence developmental, metabolic, and growth processes in plants without having lethal effects. In agriculture and horticulture, they are employed to obtain specific advantages, which include increased yields, modification of growth forms, improved quality, modification of plant constituents, and facilitation of harvesting. In the classification of these substances, it is impossible to avoid a certain degree of overlap with the herbicides (→ Weed Control) because higher doses of some agents can also cause damage to the plant. In a broader sense, substances such as defoliants and desiccants are also classified as growth regulators or bioregulators.

Phytohormones. Plants synthesize a system of plant hormones (phytohormones) that regulate metabolic processes. The signal effects of this system on metabolic processes are comparable to those of animal hormone systems (→ Hormones, Chap. 1.1.1.), but differ as regards organ specificity. Animal hormones are synthesized at specific sites and produce an effect in a particular target organ, whereas phytohormones are formed at many sites and act on many different plant organs (multilocal); they therefore have a wide spectrum of activity. Phytohormones interact with one another, producing antagonistic, additive, or synergistic effects.

Five groups of substances are known to be involved in the transmission of external and internal signals from one part of the plant to another:

1) Auxins
2) Gibberellins
3) Cytokinins
4) Abscisic acid
5) Ethylene

Depending on the stage of plant development, varying amounts of hormones are formed in the plant organs in response to environmental stimuli. The stimulatory or inhibitory effect of phytohormones enables the plant to continually regulate vital processes, e.g., growth, differentiation, flowering, maturation, and the shedding (abscission) of leaves. It is often not the type but the concentration of the phytohormone that determines whether a given developmental process is stimulated or inhibited. Obviously, there is a chance that active substances can be used to specifically influence these complicated events, especially in cultivated plants.

History. Scientific findings in the field of plant physiology not only give insight into the molecular basis of plant development, but also provide technical applications that are becoming increasingly important for the cultivation of plants. The use of plant growth regulators in agriculture and horticulture began in the 1930s. They include both natural phytohormones, such as auxin (3-indolylacetic acid) and gibberellic acid (GA_3), as well as synthetic substances, such

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_415

as daminozide, chlormequat, and ethephon. Examples of these substances and their applications are listed in chronological order in Table 1. Some are of considerable economic importance (e.g., auxin-type substances, chlormequat), whereas others are no longer in use (e.g., ethylene, acetylene).

Table 1. Plant growth regulators

Year of introduction	Agent (trade name)	Application
1932	ethylene, acetylene	induction of flowering in pineapples
1936	auxin	production of seedless fruit
1943	1-naphthylacetic acid	thinning out of apples
1944	auxin-type substances e.g., 2,4-D ∗	selective herbicides
1948	maleic hydrazide	inhibition of the growth of grass
1959	gibberellin	improvement in the quality of grapes
1959	chlormequat (Cycocel)	strengthening the stalks of cereals
1962	daminozide	improvement in the quality of apples and cherries
1966	cytokinins	improvement in the fructification of grapes
1971	ethephon (Ethrel)	increase in the latex flow in rubber trees
1973	glyphosine (Polaris)	increase in the sugar yield from cane
1979	mepiquat (Pix)	prevention of excessive growth of cotton
1987	inabenfide (Seritard)	strengthening the stalks of rice
1989	triapenthenol (Baronet)	improvement in the stability of rape

∗ 2,4-Dichlorophenoxyacetic acid.

Uses. Irrespective of whether a plant is wild or cultivated, its development normally falls short of its capacity, i.e., its genetically defined potential is not fully exploited. Limiting factors are temperature; shortage of light, water, and nutrients; competing plants; insect damage; and microorganisms.

Increased yields can be achieved in plant cultivation by measures such as adequate fertilization and irrigation, use of high-yield varieties, optimal soil treatment, and reduction of diseases, pests, and weeds with agricultural pesticides. Additional increases in plant production can also be obtained by using growth regulators. Some regulators only change the outer form of plants (morphoregulators). Substances that influence other properties of cultivated plants are also being increasingly investigated. These properties include plant constituents (sugar, protein), resistance to stress (dryness, heat, frost, salty soil), and rhythm of plant development (fruit set, induction of flowering, acceleration of maturity, and shedding of fruit and leaves, i.e., abscission).

The development of a marketable growth regulator generally takes longer than the development of other plant treatment agents; determination of the dose rate and the best time for application is often critical and requires extensive outdoor experimentation under varying conditions.

Growth regulators have to meet the same requirements for approval and control as other plant treatment agents (see → Fungicides, Agricultural, → Insect Control, → Weed Control). Detailed information about the product properties of the active agents (e. g., toxicity, formulation, solubility, application, and trade names) are given in [17], [18].

Growth regulators are usually applied in agriculture by spraying them on foliage. Although seed treatment has environmental advantages, this method is rarely implemented because application has to take place at a defined time before sowing. In horticulture, the agent is often added during watering and is taken up through the roots.

Economic Aspects [19]. Compared with the three main groups of plant protection agents (herbicides, insecticides, and fungicides), growth regulators have a relatively small share of the world market (ca. 4 %). The growth rate in sales is, however, comparatively high. At the end of the 1980s the entire market volume for growth regulators was ca. US $ 750×10^6 (end-user level). The defoliants and desiccants (agents for facilitating harvesting, especially of cotton, tobacco, and potatoes) alone accounted for about 40 % of sales and were predominantly agents with commodity status. The application of growth-retarding agents in European cereal and rape cultivation and in American cotton is another economically important sector.

The most important agents (see Table 1) are maleic hydrazide, chlormequat, ethephon, daminozide, gibberellic acid, glyphosine, mepiquat, and thidiazuron.

2. Classes of Growth Regulators

Growth regulators belong to so many different chemical families that classification according to their chemistry does not seem very useful. The classification system used here is based on mechanisms of action; many growth regulators exert effects that are similar or antagonistic to those of native phytohormones, allowing a rough qualitative classification [1, p. 10].

The names used for growth regulators, like those for other agents in pharmaceutics and plant protection, are numerous and confusing. Manufacturers not only use trade names, but also apply to national and international authorities (e.g., ISO) for a common name. The common name is normally an abbreviation of the chemical name and is used to denote the pure active agent.

2.1. Auxins and Auxin-Type Compounds

Auxins and auxin-type growth substances exert the typical effects of the natural phytohormone auxin (3-indolylacetic acid): the promotion of longitudinal growth by cell elongation, stimulation of cell division in the cambium and roots, and control of enzyme activity.

Auxins have not gained much practical importance as growth-stimulating agents because of their narrow therapeutic dose range. A small overdose can cause the growth-regulating effect to become a herbicidal effect. Auxin-type substances are much more important as herbicides than as growth regulators (→ Weed Control). The following auxins are of significance as growth regulators.

3-Indolylacetic acid [*87-51-4*], $C_{10}H_9NO_2$, M_r 175.2, was isolated from human urine by Kögl in 1934 [20].

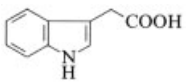

Tryptophan is the starting material for the biosynthesis of this substance [15, p. 39]. It is used to stimulate root growth in cuttings, in the production of seedless fruit, and in the induction of flowering in pineapples. However, indolylacetic acid has been largely replaced by other carboxylic acids.

3-Indolylbutyric acid [*133-32-4*], IBA, $C_{12}H_{13}NO_2$, M_r 203.2, *mp* 123 – 125 °C, colorless crystals.

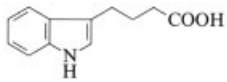

In the plant this acid is degraded to indolylacetic acid, presumably by β-oxidation. It is made from indole, γ-butyrolactone, and sodium hydroxide and is used in the same way as indolylacetic acid.
Trade names: Hormodin (MSD Agvet), Seradix (Rhône-Poulenc).

1-Naphthaleneacetic acid [*86-87-3*], $C_{12}H_{10}O_2$, M_r 186.2, *mp* 134 – 135 °C, colorless crystals, LD_{50} rat oral ≈ 2520 mg/kg.

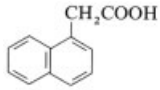

This substance is produced by the hydrolysis of 1-naphthaleneacetonitrile, which is obtained by the chloromethylation of naphthalene and subsequent reaction with sodium cyanide. It is used for the same purposes as indolylacetic acid and to thin out apples, olives, oranges etc. and to inhibit germination of potatoes during storage.
Trade names: Celmone (Excel Ind.); Thin'n Stop-Drop (Platte Chem), Fruitone-N, NAA 800, Tree-Hold (Rhône-Poulenc); Olive Stop (UAP Special Products).

1-Naphthaleneacetamide [*86-86-2*], $C_{12}H_{10}NO$, M_r 185.2, *mp* 182 – 184 °C, LD_{50} rat oral ≈ 1000 mg/kg. It is used in the same way as naphthaleneacetic acid.
Trade names: Amid-Thin W, Rootone (Rhône-Poulenc, Greenwood Chemical).

2.2. Antiauxins

In a narrower sense, antiauxins are defined as substances that compete with auxins for specific

receptors [21]. Only agents that are antiauxins in a broader sense have achieved practical importance; they inhibit the transport of auxins in the plant and, thus, strongly affect growth and morphology [22]. Frequently, these effects are exploited in herbicides, as in the case of the morphactins derived from fluorene-9- carboxylic acid (→ Weed Control). The following antiauxins are used as growth regulators.

***N-m*-Tolylphthalamic acid** [*85-72-3*], N-m-t, $C_{15}H_{13}NO_3$, M_r 255.3, $LD_{50} \approx 5230$ mg/kg (oral, male albino rats).

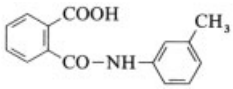

This substance is used to improve the yield of tomatoes, beans, and cherries by increasing fructification.
Trade names: Tomaset, Duraset (Uniroyal, Makhteshim-Agan).

2.3. Gibberellins

Natural gibberellins (GA) have a wide spectrum of phytohormonal effects, especially in the regulation of growth by cell elongation and cell division and in the induction of hydrolytic enzymes. More than 70 natural gibberellins, each containing 19 or 20 carbon atoms, have been found [23]. The nomenclature of these substances (GA_1, GA_2 . . .) follows the order of their discovery. Not all the gibberellins are biologically active, some are precursors in the biogenesis of active (usually C_{19}) gibberellins and others are inactivated metabolites.

The biosynthesis of these diterpene compounds involves mevalonate as the starting product which is converted to geranylgeranyl pyrophosphate (→ Terpenes) [24]. The latter is cyclized to give the key substance *ent*-kaurene. Stepwise oxidation then yields the acid, which is converted to the GA_{12} aldehyde by subsequent oxidation and ring contraction. Further biosynthetic steps give the gibberellins, which all have a carboxyl group at position 10 of the skeleton.

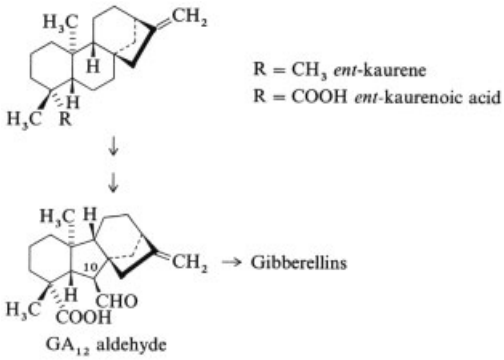

Gibberellic acid [*77-06-5*], GA_3, $C_{19}H_{22}O_6$, M_r 346.4, *mp* 233 – 235 °C. This substance is produced biotechnologically from the fungus *Gibberella fujikuroi* (*Fusarium moniliforme*).

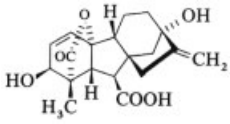

It is used in the United States at concentrations of 1 – 1000 ppm to produce seedless dessert grapes, to increase fruit and vegetable yields (celery and artichokes), and in the production of vegetable seeds.

Oral doses up to 1500 mg/kg produce no toxic symptoms in mice.
Trade names: Pro-Gibb, Pro Vide (Abbott Labs); Activol, Berelex, Grocel (ICI); Gibrel (MSD Agvet).

Promalin (Abbott), a mixture of GA_4/GA_7 with 6-benzyladenine, is also available for improving the quality and yield of apples.

2.4. Gibberellin Antagonists

Substances that inhibit gibberellin biosynthesis are of very great practical significance. Onium compounds of nitrogen, phosphorus, and sulfur form an important group. They inhibit gibberellin biosynthesis by blocking the cyclization steps leading to formation of *ent*-kaurene [25]. Macroscopic effects include growth retardation, stunted and compact growth, intense green coloring and, frequently, improved lodging resis-

tance of the plants [26]. The spectrum of activity in different plant varieties depends mainly on the individual structure of the onium compound [27].

Chlormequat [*999-81-5*], chlorocholine chloride (CCC), $C_5H_{13}Cl_2N$, M_r 158.1, *mp* ca. 245 °C (decomp.), LD_{50} rat oral ≈ 670 mg/kg.

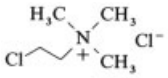

The biological effects of chlormequat were first described by Tolbert [28]. Chlormequat is the most important inhibitor of gibberellin biosynthesis. It is synthesized from trimethylamine and 1,2-dichloroethane.

It is used primarily in the cultivation of wheat in Europe. Application leads to strong, compact growth, increased chlorophyll content, and increased stalk diameter which prevents the cereal from falling over (especially wheat), before it is harvested [29], [30]. A mixture of chlormequat and ethephon has proved useful for barley. In the growing of ornamental plants, chlormequat is used to shorten the stems of poinsettias, azaleas, pelargoniums, and lilies.
Trade names: Cycocel, CCC (American Cyanamid, BASF) and as a mixture with ethephon: Terpal C (BASF).

Mepiquat [*24307-26-4*], *N,N*-dimethylpiperidinium chloride [31], [32], $C_7H_{16}NCl$, M_r 149.7, *mp* 350 °C, LD_{50} rat oral ≈ 1600 mg/kg.

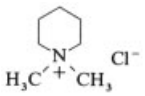

Mepiquat is produced from *N*-methylpiperidine and chloromethane. It prevents excessively lush vegetative growth of cotton [33–35] and therefore facilitates mechanical harvesting and results in earlier maturation, often in combination with a higher yield.

Mepiquat is used in combination with ethephon (see page 8) to stabilize barley stalks.
Trade names: Pix (BASF); mepiquat mixed with ethephon: Terpal (BASF).

Chlorphonium chloride [*115-78-6*], tri-(*n*-butyl)-2,4-dichlorobenzylphosphonium chloride [36], $C_{19}H_{32}Cl_3P$, M_r 397.8, *mp* 114 – 120 °C, LD_{50} rat oral ca. 180 mg/kg

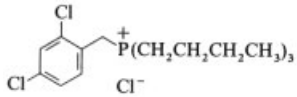

This substance is made from 2,4-dichlorobenzyl chloride and tributyl phosphine. It is used to obtain compact growth in ornamental flowers (lilies and chrysanthemums).
Trade name: Phosfon (Mobil Chemical) (discontinued).

Heterocyclic nitrogen compounds (pyrimidines, pyridines, diazetines, and triazoles) also inhibit gibberellin biosynthesis and are attaining increasing importance. These agents all have a sp^2-hybridized nitrogen atom which can complex the central iron atom in some cytochrome P450 enzymes. They block the oxidative steps of GA synthesis that lead to *ent*-kaurenoic acid [27].

Ancymidol [*12771-68-5*][37], $C_{15}H_{16}N_2O_2$, M_r 256.3, *mp* 110 – 111 °C, LD_{50} rat oral ≈ 4500 mg/kg.

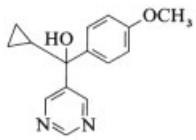

This substance is structurally related to pyrimidine fungicides, such as fenarimol (→ Fungicides, Agricultural, Chap. 8.6.4.). The fungicidal activity of these substances is based on the inhibition of cytochrome P450 enzymes in fungal ergosterol biosynthesis.

Ancymidol is produced by the metallation (butyllithium) of 5-bromopyrimidine and subsequent reaction with cyclopropyl 4-methoxyphenyl ketone [38].

Ancymidol reduces the length between internodes in many different plant varieties when applied to the leaves or the soil. It produces compact growth and shortened branches in ornamentals (e.g., chrysanthemums, poinsettias, tulips).
Trade names: Reducymol, A-Rest (Elanco).

Inabenfide [*82211-24-3*] [39], $C_{19}H_{15}ClN_2O_2$, M_r 338.8, *mp* 210 – 212 °C, LD_{50} rat and mouse > 15 000 mg/kg.

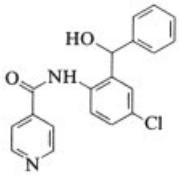

This substance is used to shorten and increase the lodging resistance of paddy rice. It is preferentially absorbed via the roots. For biology, see [40]; for mode of action, see [41].
Trade name: Seritard (Chugai).

Tetcyclacis [*65245-23-0*] [42], $C_{13}H_{12}ClN_5$, M_r 273.7, *mp* 190 °C, white crystals, LD_{50} rat oral ≈ 258 mg/kg.

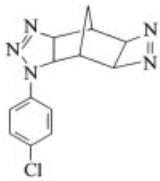

This substance is made from quadricyclane by the cycloaddition of azodicarboxylate and 4-chlorophenyl azide, followed by saponification, decarboxylation, and oxidation [43].

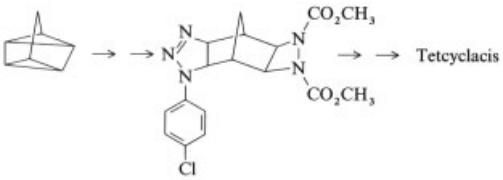

Tetcyclacis is highly active and has a wide spectrum of activity. However, it cannot be used under field conditions in the treatment of foliage because of its photolability. It is applied to rice by seed soaking or soil drenching. Tetcyclacis produces shorter, more compact seedlings which have better root development that are more suitable for mechanical planting.
Trade name: Kenbyo (BASF).

Pachlobutrazol [*76738-62-0*] [44], $C_{15}H_{20}ClN_3O$, M_r 293.8, *mp* 160 °C, LD_{50} rat oral 5346 mg/kg (Bonzi).

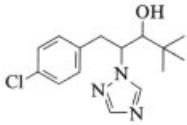

Pachlobutrazol is the diastereomer with a 2 *RS*, 3 *RS* configuration. It is used as a growth-retarding agent for ornamentals, amenity trees, apple and pear orchards, and for the reduction of lodging in grass seed. For a description of biology, see [45].
Trade names: Bonzi, Clipper, Cultar, Parley (ICI).

Uniconazole [*83657-22-1*] [46], [47], $C_{15}H_{18}ClN_3O$, M_r 291.8, *mp* 147 – 164 °C, LD_{50} rat oral 2020 mg/kg (male).

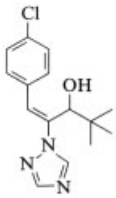

Uniconazole, like pachlobutrazol, is used in very low amounts (12 g/ha) in ornamentals, trees, shrubs, and rice.
Trade names: Romika, Sumiseven, Sumishort (Sumitomo); Ortho-Prunit, Ortho-Sumagic, Ortho-Prism (Valent U.S.A.).

Triapenthenol [*76608-88-3*] [48], [49], $C_{15}H_{25}H_3O$, M_r 263.4, *mp* 135.5 °C, LD_{50} rat oral > 5000 mg/kg.

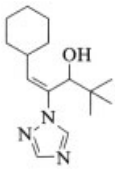

Triapenthenol is used to increase the stability of rape and in grass seed production. It also has a wide spectrum of activity as a growth-retarding agent, especially in dicotyledons (soybean, peanuts, potatoes, and orchard trees).

Trade name: Baronet (Bayer).

Dioxocyclohexane carboxylic acid derivatives with the following structure constitute another group of inhibitors of gibberellin biosynthesis [50], [51].

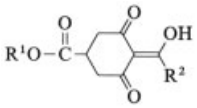

These compounds inhibit the oxidation steps after GA_{12} aldehyde [52].

Cimectacarb [*95266-40-3*] (proposed common name for CGA 163 935) [51], $C_{13}H_{16}O_5$, M_r 252.3, *mp* 36 °C, LD_{50} rat oral 4460 mg/kg.

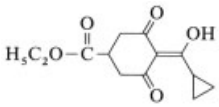

Cimectacarb is a growth-retarding agent for grain, rape, and grass [53].
Trade names: Omega, Vision (Ciba-Geigy).

Prohexadione- calcium [*127277-53-6*],[54], $C_{10}H_{10}O_5Ca$, M_r 250.3, *mp* 360 °C, LD_{50} rat oral $>$5000 mg/kg is a new growth retardant from Kumiai Chemical Industry. It is active after foliar application in cereals, turf grasses, rape and ornamental plants.

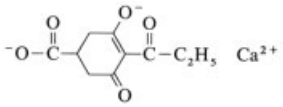

Daminozide [*1596-84-5*], *N,N*-dimethylaminosuccinamide [55], $C_6H_{12}N_2O_3$, M_r 160, *mp* 154 – 156 °C, LD_{50} rat oral ca. 8400 mg/kg.

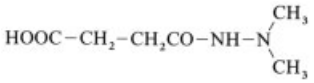

This substance is produced from succinic anhydride and 1,1-dimethylhydrazine. It is used in fruit farming to accelerate ripening (cherries) and improve both fructification (grapes) and fruit quality (about 30 % of the apples grown in the USA are treated with daminozide). It is also used to regulate linear growth of ornamentals.

Daminozide does not inhibit GA synthesis, but GA transport [56].
Trade names: Alar, B-Nine, Kylar (Uniroyal); Dazide (Fine Agrochemicals).

2.5. Cytokinins [57]

Cytokinins are a series of natural or synthetic adenine derivatives which have a side chain on the exocyclic nitrogen atom and exert certain phytohormonal effects. The typical activity profile of cytokinins includes the stimulation of cell division, slowing down of plant senescence, stimulation of bud formation and germination of seeds, and the promotion of assimilate transport.

The natural hormone zeatin [*1637-39-4*] and synthetic derivatives, e.g., 6-benzylaminopurine (benzyladenine) [*1214-39-7*] and 6-furfurylaminopurine (kinetin) [*525-79-1*], have been used in experimental work.

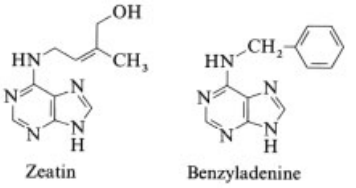

In spite of numerous patents for the application of cytokinins, these substances have little practical significance. A mixture of benzyladenine and gibberellic acids [see Section 2.3, Promalin (Abbott)] is commercially available. Some aryl and heteroaryl ureas also exhibit cytokinin-like activity [58]. Thidiazuron (see Chap. 3) which is used as a defoliant for cotton plants, is also a member of this group of compounds.

2.6. Abscisic Acid [58], [59]

Abscisic acid is a phytohormone which acts very much like a physiological antagonist of the cytokinins. It regulates the dormancy of seeds and buds, the fall of fruit and foliage, and the winter rest of plants. This acid also plays a role in stress

situations that are caused by environmental factors. In the case of water shortage, for example, the abscisic acid content of the plant increases, resulting in decreased transpiration due to the closing of the stomata.

d-(+)-Abscisic acid [*21293-29-8*], $C_{15}H_{20}O_4$, M_r 264.3, *mp* 160 – 162 °C, $[\alpha]_D$ +430°, is colorless and sensitive to light. The (±) acid melts at 191 °C.

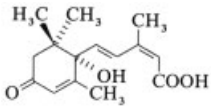

Experimental work with abscisic acid and its synthetic analogues is designed to exploit the use of these substances against stress caused by water shortage or cold [60]. These compounds are not yet, however, used in practice.

2.7. Ethylene-Releasing Compounds

Although the effects of ethylene on plant physiology have long been known, its classification as a phytohormone has been the subject of much controversy [61]. Endogenous ethylene is widely distributed in plant tissues (foliage, roots, flowers, seeds) and is synthesized during the ripening of fruit. However, unlike other phytohormones, it is not transported within the plant. Ethylene is probably only formed in the vicinity of its site of action. Its precursor, 1-aminocyclopropane-1- carboxylic acid, is the form that is transported. Ethylene is now defined as a plant hormone because it exerts many effects that regulate plant growth. Contrary effects are often observed, e.g., the inhibition and stimulation of metabolic processes or growth, depending on the type of plant. Thus, ethylene inhibits internodal growth in cereals, but stimulates the growth of sugar cane.

Ethylene is synthesized in the plant from carbon atoms 3 and 4 of methionine, $CH_3SCH_2CH_2CH(NH_2)COOH$, via the intermediates *S*-adenosylmethionine and 1-aminocyclopropane-1- carboxylic acid.

Ethylene gas cannot easily be used outdoors. Consequently, agricultural preparations have been developed that only release ethylene after they have been absorbed by the plant.

Ethephon [*16672-87-0*], 2- chloroethylphosphonic acid [62], $C_2H_6ClO_3P$, M_r 144.5, *mp* 74 – 75 °C (decomp.), LD_{50} rat oral ca. 4000 mg/kg, water soluble.

This acid is synthesized from phosphorus trichloride and ethylene oxide [63]. It decomposes to form ethylene in aqueous solutions at pH 4 [64].

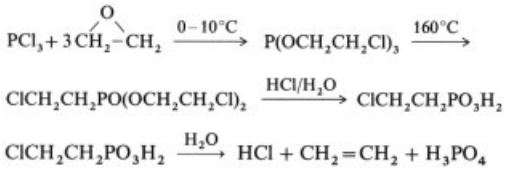

Ethephon is one of the most economically important growth regulators. It accelerates ripening (tomatoes, cherries, and citrus fruit), stabilizes stalks of barley and rice, induces flowering of pineapple plants, and increases yields in rubber plantations (stimulation of latex flow) [65].
Trade names: Ethrel, Florel, Cerone, Chipco, Prep (Rhône-Poulenc); Flordimex (Chemie AG Bitterfeld-Wolfen); Arvest Etheverse (C.F.P.I.); combination with mepiquat, Terpal (BASF); combination with chlormequat, Terpal C (BASF).

Etacelasil [*37894-46-5*] [66], $ClCH_2CH_2Si(OCH_2CH_2OCH_3)_3$, $C_{11}H_{25}ClO_6Si$, M_r 316.9, colorless liquid, *bp* 85 °C at 1.33×10^{-3} mbar, LD_{50} rat oral 2000 mg/kg.

It is used specially to facilitate the harvesting of olives (abscission agent). Depending on the type of olive, harvesting should be carried out 6 – 10 days after application (1 – 2 g of agent per liter).
Trade name: Alsol (Ciba-Geigy).

2.8. Other Growth Regulators

Maleic hydrazide [*123-33-1*], LD_{50} rat oral (as the sodium salt) ca. 6950 mg/kg. This substance was the first synthetic growth inhibitor and was introduced in 1948 under the code name MH-30 (US Rubber) [67], [68].

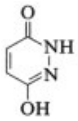

Maleic hydrazide inhibits cell division but does not influence cell elongation [69]. It is rapidly transported in the plant.

Maleic hydrazide is an economically important growth regulator. It is used to inhibit the growth of lateral shoots in tobacco (USA), to inhibit the growth of grass, and to inhibit the germination of potatoes and onions. When used on ornamental lawns, the grass frequently turns reddish brown because anthocyanin formation is apparently stimulated.
Trade names: Retard, Sucker Stuff, Super Sprout Stop, Super Sucker Stuff (Drexel Chemical); De-Cut, Fair-2, Fair-Plus, Super De-Sprout (Fair Products); Royal MH-30, Royal MH-30 SG, Royal Slo Gro (Uniroyal).

Mefluidide [*53780-34-0*] [70], $C_{11}H_{13}F_3N_2O_3S$, M_r 310.3, *mp* 183 – 185 °C, LD_{50} rat oral ca. 28 000 mg/kg, LD_{50} mouse oral ca. 1920 mg/kg.

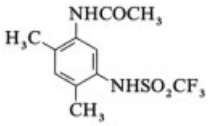

Mefluidide is used to inhibit the growth of grass (roadsides, playing fields, embankments, and green areas). This water-soluble agent is absorbed by grass mainly through its leaves. It inhibits growth due to cell elongation and largely suppresses seed head formation. Mefluidide also accelerates the ripening of sugar cane and is used in mixtures with herbicides.
Trade names: Embark (3 M), Compo grass regulator (BASF).

Amidochlor [*40164-67-8*], $C_{15}H_{21}ClN_2O_2$, M_r 296.8, *mp* 148 – 149 °C, LD_{50} rat oral 3100 mg/kg [71].

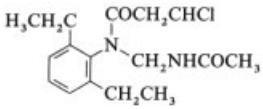

This substance is used as a growth inhibitor for turf grass.
Trade name: Limit (Monsanto)

Dikegulac sodium [*52508-35-7*], $C_{12}H_{17}NaO_7$, M_r292.6, *mp* > 300 °C, LD_{50} rat oral ca. 31 000 mg/kg (male), ca. 18 000 mg/kg (female). This substance is an intermediate in the production of ascorbic acid and was introduced as a herbicide and growth regulator in 1973 by Hoffmann–La Roche [71], [72].

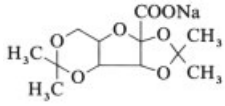

It is used to inhibit growth and promote the formation of lateral shoots in ornamentals (e.g., azaleas), and to inhibit the growth of shrubs and hedges (reduction of mechanical cutting).
Trade names: Atrinal (Dr. Maag), Atrimmec (PBI/Gordon).

Glyphosine [*2439-99-8*], *N,N*-bis(phosphonomethyl)glycine [73], $C_4H_{11}NO_8P_2$, M_r 263.1, colorless, *mp* 203 °C, LD_{50} rat oral ca. 3900 mg/kg.

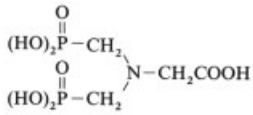

Glyphosine accelerates maturation and increases the sucrose content of sugar cane (2 – 5 kg/ha are applied about 4 – 8 weeks before harvesting) by reducing the inversion of sucrose [74]. It is produced from glycine and chloromethylphosphonic acid. Glyphosine is being gradually replaced by sodium glyphosate (Polado).
Trade name: Polaris (Monsanto)

Glyphosate [*1071-83-6*], *N*-(phosphonomethyl)glycine, $C_3H_8NO_5P$, M_r 169.1, *mp* 230 °C (decomp.), solubility in water 1.2 % at 25 °C. The sodium salt of glyphosate is available as " Polado" and is recommended for the same applications as glyphosine.
Trade name: Polado (Monsanto) [75].

Fosamine ammonium [*25954-13-6*], ammonium *O*-ethyl carbamoylphosphonate [76],

$C_3H_{12}N_2O_4P$, M_r 170, *mp* 175 °C, LD_{50} rat oral ca. 24 000 mg/kg (41.5 % product).

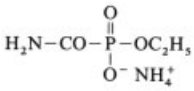

It is used to inhibit the growth of shrubs (e.g., in woods). It inhibits the formation of shoots if applied in the fall and causes growth inhibition or defoliation if sprayed on the leaves in spring.
Trade name: Krenite (Du Pont).

Folicysteine [*8064-47-9*] is a mixture of 5 % L-cysteine derivatives and 0.1 % folic acid in stabilized, buffered solution. It stimulates the growth and development of plants when used in low concentrations.
Trade name: Ergostim (Montedison).

Mixtures of long-chain fatty alcohols (e.g., octanol and decanol) are used in the United States to control the lateral shoots of tobacco plants by inhibiting the development of axillary buds. In ornamentals, this "chemical pruning" leads to more abundant flowering and improved branching (azaleas and rhododendrons).
Trade names: Antak, Sucker Plucker (Drexel); Off-Shoot (Cochran); Royaltac (Uniroyal).

3. Defoliants and Desiccants

The premature shedding of leaves is induced by defoliants to allow or facilitate harvesting. Defoliants are used primarily in American cotton cultivation and to a lesser extent in soybean, vines, and tomatoes.

Desiccants are substances that artificially accelerate the drying of parts of plants. This effect is exploited in the harvesting of cotton plants and in the destruction of potato foliage.

Some of these harvesting aids are also used as herbicides. Examples are paraquat and diquat, sodium chlorate, dimethylarsinic acid, arsenic acid (H_3AsO_4), and endothal (→ Weed Control). Thus, the classification of these substances as growth regulators is arbitrary. Other harvesting aids follow.

Butifos [*78-48-8*] [77], tributyl trithiophosphate, $(C_4H_9S)_3PO$, $C_{12}H_{27}OPS_3$, M_r 314.5, colorless or slightly yellow liquid, *bp* 150 °C at 0.4 mbar, LD_{50} rat oral 200 mg/kg. It is made from butanethiol and phosphorus oxychloride in the presence of a base. Butifos is a very effective defoliant for cotton. About half of the American cotton plantations are treated with organophosphorus defoliants about 1 – 2 weeks before harvesting.
Trade name: DEF (Mobay).

Merphos, tribufos [*150-50-5*], tributyl trithiophosphite [78], $(C_4H_9S)_3P$, $C_{12}H_{27}PS_3$, M_r 298.5, LD_{50} rat oral ca. 348 – 712 mg/kg, insoluble in water. It is made from phosphorus trichloride and tributanethiol and is used as a defoliant for cotton. The leaves fall in a relatively green state, leaving open cotton plants that can easily be mechanically harvested. In addition, fungal diseases (e.g., cotton boll rot) are suppressed by the better airing of the lower parts of the plant.
Trade name: Folex (Rhône-Poulenc).

Buminafos (proposed common name) [*51249-05-9*], dibutyl 1-butylaminocyclohexane phosphonate [79], $C_{18}H_{38}NO_3P$, M_r 343.5, liquid.

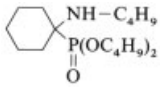

Cyclohexanone and butylamine are reacted to give the Schiff base, which is then reacted with tributyl phosphite. Buminafos is used to destroy potato foliage, to defoliate cotton plants, and as a contact herbicide (nonselective).
Trade name: Trakephon (Chemie AG Bitterfeld-Wolfen).

Dimethipin [*55290-64-7*] [80], $C_6H_{10}O_4S_2$, M_r 210, *mp* 166 – 168 °C, LD_{50} rat oral ca. 1150 mg/kg.

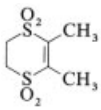

Dimethipin is produced from 3-chloro-2-butanone and ethane dithiol with subsequent oxidation. It is used as a defoliant for cotton plants

and as a potato vine desiccant. It enhances the maturation process and reduces seed moisture in the harvesting of rice, rape seed, flax, and sunflower.
Trade name: Harvade (Uniroyal).

Thidiazuron [*41118-83-6*][81], $C_9H_8N_4OS$, M_r 220.3, *mp* 217 °C (decomp.), LD_{50} rat oral ≈ 4000 mg/kg.

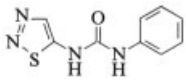

Thidiazuron is produced from 5-amino-1,2,3-thiadiazol and phenyl isocyanate. It is used to defoliate cotton plants and has cytokinin activity (see Section 2.5) [82].
Trade name: Dropp (Schering).

4. References

General References

1. W. Draber: "Natürliche und synthetische Pflanzenwachstumsregulatoren," in R. Wegler (ed.): *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* vol. 4, Springer Verlag, Berlin 1977, p. 1.
2. D. S. Letham, P. B. Goodwin, T. J. V. Higgins (eds.): *Phytohormones and Related Compounds: A Comprehensive Treatise,* vol. I und II, Elsevier, Amsterdam 1978.
3. W. Draber, J. Stetter: "Plant Growth Regulators," in *Chemistry and Agriculture.* Special Publication no. 36, The Chemical Society, Burlington House, London 1979.
4. N. B. Mandava (ed.): "Plant Growth Substances," *ACS Symp. Ser.* **III** (1979).
5. L. G. Nickell: *Plant Growth Regulators (Agricultural Uses),* Springer Verlag, Berlin 1982.
6. J. S. McLaren (ed.): *Chemical Manipulation of Crop Growth and Development,* Butterworths, London 1982.
7. T. H. Thomas (ed.): *Plant Growth Regulator Potential and Practice,* The British Plant Growth Regulator Group and The British Crop Protection Council, BCPC Publ., Croydon, U.K., 1982.
8. L. G. Nickell (ed.): *Plant Growth Regulating Chemicals,* vol. I and II, CRC Press, Boca Raton, Fla. 1983.
9. R. L. Ory, F. R. Rittig (eds.): "Bioregulators, Chemistry and Uses," *ACS Symp. Ser.* **257** (1984).
10. M. Bopp (ed.): *Plant Growth Substances 1985,* Springer Verlag, Berlin 1985.
11. N. Takahashi (ed.): *Chemistry of Plant Hormones,* CRC Press, Boca Raton, Fla., 1986.
12. A. F. Hawkins, A. D. Stead, N. J. Pinfield (eds.): *Plant Growth Regulators for Agricultural and Amenity Use,* BCPC Publ., Thornton Heath, U.K., 1987.
13. P. J. Davies (ed.): *Plant Hormones and their Role in Plant Growth and Development,* M. Nijhoff Publ., Dordrecht 1987.
14. J. A. Roberts, R. Hooley: *Plant Growth Regulators,* Blackie, Glasgow 1988.
15. T. C. Moore: *Biochemistry and Physiology of Plant Hormones,* 2nd ed. Springer Verlag, Berlin 1989.
16. M. Luib, P. E. Schott: "Einsatz von Bioregulatoren" in G. Haug, G. Schuhmann, G. Fischbeck (eds.): *Pflanzenproduktion im Wandel,* VCH Verlagsgesellschaft, Weinheim 1990, chap. 10.

Specific References

17. C. R. Worthing, S. B. Walker (eds.): *The Pesticide Manual – A World Compendium,* 7th ed., British Crop Protection Council, The Larenham Press, Larenham, U.K., 1983.
18. C. Sine (ed.): *Farm Chemicals Handbook 1990,* Meister Publ., Willoughby, Ohio.
19. BASF Aktiengesellschaft, Marketing Report: Crop Protection, Ludwigshafen.
20. F. Kögl, D. F. Kostermans, *Hoppe Seyler's Z. Physiol. Chem.* **228** (1934) 90.
21. L. J. Andus: *Plant Growth Substances,* **vol. I,** L. Hill, London 1972, 179 ff. .
22. C. C. McBready, *Annu. Rev. Plant Physiol.* **17** (1966) 283.
23. N. Takahashi, I. Yamaguchi, H. Yamane: "Gibberellins," in [11]p. 57.
24. J. E. Graebe, *Ann. Rev. Plant Physiol.* **38** (1987) 419.
25. W. Rademacher: "Gibberellins, Metabolic Pathways and Inhibitors of Biosynthesis," in P. Boeger, G. Sandmann (eds.): *Target Sites for Herbicide Action.* CRC Press, Boca Raton, Fla., 1989, p. 127.
26. W. Rademacher: "Inhibitors of Gibberellin Biosynthesis, Applications in Agriculture and Horticulture," in N. Takahashi et al. (eds.): *Gibberellins, Proceedings of the Gibberellin Symposium Tokyo 1989,* Springer Verlag, Berlin 1990.

27. H. Sauter, *ACS Symp. Ser.* **257** (1984) 9.
28. N. E. Tolbert, *Plant Physiol.* **35** (1960) 380; *J. Biol. Chem.* **235** (1960) 475.
29. J. Jung, *Arzneim. Forsch.* **30** (1980) 1974.
30. A. I. Zadocev, G. R. Pikus, A. L. Grincenko: *Chlorcholinchlorid in der Pflanzenproduktion,* VEB Dtsch. Landwirtschaftsverlag, Berlin 1977.
31. BASF, DE 2 207 575, 1972. (B. Zeeh et al.).
32. B. Zeeh, K. H. König, J. Jung, *Kem.-Kemi* **1** (1974) no. 9, 621.
33. J. Jung, B. Würzer, H. von Amsberg, 170. Meeting Am. Chem. Soc. Abstr. Paper 69, Pest. Div. 1975.
34. S. Behrendt et al., *Landwirtsch. Forsch. Sonderh.* **35** (1979) 277.
35. P. E. Schott, F. R. Rittig in J. S. McLaren (ed.): *Chemical Manipulation of Crop Growth and Development,* Butterworths, London 1982, p. 415.
36. Mobil Oil, US 3 268 323, 1961 (L. E. Goyette).
37. Eli Lilly, US 3 868 244, 1974 (H. M. Taylor, I. D. Davenport, R. E. Hackler).
38. Eli Lilly, DE 1 770 288, 1968 (J. D. Davenport, R. E. Hackler, H. M. Taylor).
39. Chugai, EP 38 998, 1980 (N. Shirakawa et al.).
40. K. Nakamma, *Jpn. Pestic. Inf.* **51** (1987) 23.
41. T. Miki et al., *Plant Cell. Physiol.* **31** (1990) 201.
42. J. Jung, H. Koch, N. Rieber, B. Würzer, *Z. Acker- Pflanzenbau* **149** (1980) 129 – 136.
43. BASF, DE 2 615 878, 1976 (R. Platz et al.). BASF, DE 3 001 580, 1980 (N. Rieber et al.).
44. ICI, EP 15 639, 1979 (B. Sugavanam).
45. B. G. Lever, S. J. Shearing, J. J. Batch: "PP 333 – A New Broad Spectrum Growth Retandant," in *British Crop Protection Conference – Weeds 1982,* vol. 1, BCPC Publ., Croydon, U.K., 1982, p. 3.
46. Sumitomo, EP 54 431, 1980 (Y. Funaki et al.).
47. H. Oshio, K. Izumi: "S-3307, A New Plant Growth Retardant – Its Biological Activities, Mechanism and Mode of Action," in P. Macgregor (ed.).: *Plant Growth Regulators in Agriculture,* FFTC Book Series no. 34, Taipei, Taiwan 1986, p. 202.
48. Bayer, EP 15 387, 1979 (W. Reiser et al.).
49. K. Luerssen, W. Reiser, *Pestic. Sci.* **19** (1987) 153.
50. Kumiai, JP 71 26483 (K. Motojima et al.).
51. Ciba-Geigy, EP 0 126 713, 1983 (H. G. Brunner).
52. I. Nakayama: "Mode of Action of Novel Plant Growth Regulators, Dioxocyclohexanecarboxylic Acid Derivatives," in N. Takahashi et al. (eds.): *Gibberellins, Proceedings of the Gibberellin Symposium,* Tokyo 1989, Springer Verlag, Berlin 1990.
53. E. Kerber, G. Leypoldt, A. Seiler: "CGA 163 935, A New Plant Growth Regulator for Small Grain Cereals, Rape and Turf," in *Brighton Crop Protection Conference – Weeds – 1989,* vol. 2, BCPC Publ., Croydon, U.K., 1989 p. 83.
54. T. Miyazawa et al.: "Prohexadione-calcium, a New Plant Growth Regulator for Cereals and Ornamental Plants," in *British Crop Protection Conference –Weeds 1991,* vol. 3, BCPC Publ., Farnham, United Kingdom, 1991, p. 967.
55. J. A. Ridell, H. A. Hageman, C. M. J'Anthony, W. L. Hubbard, *Science (Washington D.C.)* **136** (1962) 391. H. J. Brooks, *Nature (London)* **203** (1964) 1303.
56. K. Takeno, R. L. Legge, R. P. Pharis, *Plant Physiol.* **67 (suppl.)** (1981) 581.
57. R. Horgan, B. Jeffcoat (eds.): *Cytokinins: Plant Hormones in Search of a Role,* British Plant Growth Regulator Group, Bristol 1987.
58. T. Okamoto, K. Shudo, Y. Isogai: "Structural and Biological Links Between Urea and Purine Cytokinins," in J. Miyamoto, P. C. Kearney (eds.): *Pesticide Chemistry: Human Welfare and the Environment,* vol. I, Pergamon Press, Oxford 1983, p. 333.
59. In [4] p. 196.
60. W. Rademacher, R. Maisch, L. Liessegang, J. Jung in [12]p. 53. A. Flores, A. Gran, F. Laurich, K. Dörffling, *J. Plant Physiol.* **132** (1988) 362.
61. K. Lürssen, *Chem. Unserer Zeit* **15** (1981) 122. J. A. Roberts, G. A. Tucker (eds.): *Ethylene and Plant Development,* Butterworths, London 1985.
62. Amchem. Prods., DE 1 667 968, 1968 (C. D. Fritz, W. Evans, A. R. Cooke). GAF, DE 2 050 245, 1969 (D. I. Randall, C. Vogel, R. W. Wynn). DE 2 053 967, 1969 (D. I. Randall, C. Vogel).
63. Hoechst, DE 2 061 610, 1970 (G. Stähler).
64. A. R. Cooke, D. I. Randall, *Nature (London)* **218** (1968) 974.
65. J. R. Sterry, *Meded. Rijksfac. Landbouwwet. Gent* **34** (1969) 462. L. Ferte, J. Pecheur, *Phytiatr. Phytopharm.* **19** (1970) 113.
66. Ciba-Geigy, DE 2 149 680, 1971 (W. Föry, H.-P. Fischer).
67. D. L. Schoene, O. L. Hoffmann, *Science (Washington D.C.)* **109** (1949) 588.

68. US Rubber, US 2 614 916, 1949 (O. L. Hoffmann, D. L. Schoene).
69. T. Hoffmann, E. V. Parups, *Residue Rev.* **25** (1969) 59. L. D. Nooden, *Physiol. Plant.* **22** (1969) 260. J. Bonaly, *C. R. Hebd. Seances Acad. Sci. Ser. D.* **273** (1971) 150.
70. 3 M., US 3 639 474, 1969 (J. K. Harrington et al.).
71. Monsanto, US 3 901 685, 1971 (K. W. Ratts).
72. P. Bocion, W. H. De Silva, G. A. Hüppi, W. Szkrybalo, *Nature (London)* **258** (1975) 142.
73. Monsanto, US 3 556 762, 1968 (P. C. Hamm).
74. C. A. Porter, L. E. Ahlrichs, *Rep. Hawaii. Sugar Technol.* **30** (1971) 71. L. E. Ahlrichs, C. A. Porter, Proc. 11th Brit. Weed Contr. Conf., Brighton 1972, p. 1215. E. G. Jaworski, *J. Agric. Food Chem.* **20** (1972) 1195.
75. Monsanto, US 3 799 758, 1971 (J. E. Franz).
76. DuPont, US 3 619 166, 1969 (B. Quebedaux, jr.). DuPont, DE 1 923 273, 1969 (W. P. Langsdorf, jr., B. Quebedaux, jr.).
77. Chemargro, US 2 943 107, 1959 (K. H. Rattenbury, J. R. Costello). Pittsburgh Coke & Chem., 2 965 467, 1956 (F. X. Markley).
78. Virginia Carolina Chem., US 2 955 803, 1958 (L. E. Goyette). Chemagro, US 3 089 807, 1961 (L. Trademan, F. R. Yagelowich).
79. Chemiekombinat Bitterfeld, DD 94 280, 1971 (E. Günther et al.).
80. Uniroyal, US 3 920 438, 1973 (A. D. Brewer, R. W. Neidermyer, W. S. McIntire).
81. Schering, DE 2 214 632, 1972 (H. Schulz, F. Arndt).
82. M. C. Mok et al., *Phytochemistry* **21** (1982) 1509.

Plant Hormones → Plant Growth Regulators
Plant Protection → Acaricides
Plant Protection → Biological Control
Plant Protection → Crop Protection
Plant Protection → Fungicides, Agricultural
Plant Protection → Insect Control
Plant Protection → Molluscicides
Plant Protection → Nematocides
Plant Protection → Weed Control
Plasma (Blood) Substitutes → Blood
Plasma Chemical Vapor Deposition → Thin Films

Plasma Reactions

Harald Brachhold, Hüls AG, Marl, Federal Republic of Germany
Richard Müller, Hüls AG, Marl, Federal Republic of Germany
Günter Pross, Hüls AG, Marl, Federal Republic of Germany

1. Introduction 1
1.1. What Are Plasmas? 1
1.2. Thermal and Nonthermal Plasmas . . 2
1.3. Applications of Plasmas 2
2. Generation of Thermal Plasmas 3
2.1. Arc Plasma Generators 3
2.2. RF Plasma Generators 5
3. Plasma Reactors 6
4. Industrial Uses of Thermal Plasmas . 7
4.1. Production of Acetylene from Hydrocarbons 7
4.2. Production of Reduction Gases 8
4.3. Production and Processing of Powders 9
4.4. Diamond and Diamond-like Coatings 9
4.5. Pollution Control 10
5. Diagnostic Methods 10
6. References 10

1. Introduction

Plasmas are a versatile tool for the development of new industrial processes and products. They are usually generated by the ionization of gases by means of electrical energy. The properties of plasmas can be modified substantially by varying the process parameters and thus adapted to specific technical applications.

Systematic development toward the utilization of plasmas started in the early 1900s. The technology continues to advance steadily and to grow in technological importance.

A distinction is made between thermal and nonthermal plasmas:

Two features of *thermal plasmas* are their high temperature and high energy flux in the plasma volume as well as at the electrodes, if present. Thermal plasma reactors are therefore suitable for heating gases to high temperatures, for carrying out reactions that consume large amounts of energy or require high activation energies, and for heating and processing materials with high melting points.

Nonthermal plasmas have comparatively low gas temperatures and energy-conversion rates. Their industrial uses are based chiefly on the reactivity of ions or radicals generated in the plasma for gas-phase or surface reactions.

This article deals exclusively with thermal plasmas, describing methods and equipment used for plasma generation and giving examples of the ways in which plasma reactions are used.

1.1. What Are Plasmas?

Plasmas and the gases from which they are generated by the addition of energy differ fundamentally in their properties.

Plasmas are electrically conductive but, because they have equal concentrations of positive and negative charge carriers, are electrically quasineutral. In addition to electrically charged particles (electrons, singly and multiply charged ions), plasmas may contain neutral molecules, atoms, and radicals. The electrical discharge is generally accompanied by intense emission of electromagnetic radiation.

The energy content of a thermal plasma is much higher than that of a gas, because in addition to energy-storing mechanisms such as increased translational velocity, excitation of rotational, vibrational, and electronic states, and dissociation of molecules that are common to both states, plasmas have an additional mechanism: particle ionization.

The energy required to generate a plasma can be applied to the plasma gas by various methods. In industry, however, electrical energy is usually used for this purpose. A current is maintained in the discharge zone, whereby the majority of the charge transport is effected by the highly mo-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_427

bile electrons, which are accelerated in an electric field. Inelastic collisions of these electrons with neutral particles are the main source of new charge carriers, which replace the charge carriers lost by diffusion and recombination.

The properties of the plasma can be adjusted to suit the application by means of a number of process parameters, including gas species, gas pressure, current, voltage, external electric and magnetic fields, and electrode shape and material [1–4].

1.2. Thermal and Nonthermal Plasmas

Thermal and nonthermal plasmas differ markedly in both discharge characteristics and applications.

Thermal plasmas are close to thermodynamic equilibrium. These have, among other characteristics, a uniform temperature for all the particles and a very high temperature in the discharge region ("hot plasmas"). They include, in particular, arc discharges.

An essential condition for the formation of a thermal plasma is a sufficiently high working pressure (usually over 10 kPa). The resulting large number of collisions between particles (in particular between electrons and heavy positive ions or neutral particles) leads to rapid redistribution of energy so that equilibrium is reached.

Nonthermal plasmas, in contrast, are far from thermodynamic equilibrium. Typically, the electrons in these plasmas have a very much higher temperature than the heavy ions and neutral particles ("cold plasmas"). This group includes glow discharges as well as radio-frequency and microwave discharges at pressures below 10 kPa.

1.3. Applications of Plasmas

Plasmas have become industrially important by virtue of their special properties [5–8]. Well-known uses include welding, cutting, melting, etching, thin-film deposition, switching, and lighting.

In the context of this article, the term "plasma processes" refers to processes in which plasmas are employed to alter the physical properties and/or the chemical composition of substances or surfaces. Such processes include, in particular, the field of plasma chemistry [9–11].

The first large-scale industrial application of thermal plasmas was the Birkeland – Eyde process, which used an arc plasma to produce nitrogen(II) oxide from air [12].

Hüls' acetylene arc process, in operation since 1939, is one of the major industrial uses of thermal plasmas, with an installed capacity (1990) of 150 MW [13], [14].

Examples of newer thermal plasma processes include the production of synthesis gas by plasma reforming, the synthesis of ceramic powders, plasma spraying, chemical and physical vapor deposition (CVD, PVD), and the disposal of toxic wastes (see Chapter 5).

One established nonthermal plasma process is the industrial generation of ozone (→ Ozone, Chap. 4.) [15], [16]. Nonthermal plasma processes that have been introduced more recently include those used for the coating or treatment of surfaces [17]. Other economically important uses include the fabrication of microelectronic circuits and a process for the preparation of plastic components (e.g., auto bumpers) for solvent-free painting (see also → Plastics, Processing).

This article is limited to thermal plasma processes. It gives a survey of the basic types of plasma generators as well as special applications.

For detailed information on the full range of thermal plasma processes, especially those in the area of metallurgy, and on the plant-scale reactors employed by various firms, see [18–21].

2. Generation of Thermal Plasmas

Thermal plasmas used in industry are generated in two types of apparatus:

1) Arc plasma generators
2) Radio-frequency (RF) plasma generators

These devices operate on different physical principles and offer distinct levels of performance. While arc generators can be designed for several tens of megawatts, the power of RF generators is limited at present to ca. 1 MW.

2.1. Arc Plasma Generators

An arc is a d.c. or a.c. gas discharge with a high current density ($\geqq$100 A/cm^2) between at least two electrodes. The temperature reaches its maximum value (10 000 – 30 000 K) on the axis of the arc stream and declines rapidly away from the axis. The temperature profile depends on operating parameters such as gas pressure, current, and plasma medium [3], [22–24].

Arc and RF plasma generators have in common that a working gas, also called plasma gas, is continuously introduced into the generator in which the discharge takes place. If the purpose of the generator is the production of a plasma or hot gas stream to be used for special applications, the generator is sometimes referred to as a plasma torch. The hot stream leaving the plasma torch is often known as the tail flame of the discharge or as the plasma jet.

Nontransferred and Transferred Arcs. The arc in a plasma generator is said to be transferred or nontransferred, depending on the type of interaction between the discharge and the substance on which it acts.

This substance may be:

1) The plasma gas itself or a component of the plasma gas
2) A substance (e.g., in powder or droplet form) transported in the plasma gas entering the generator, where it interacts directly with the arc
3) A gas, liquid, or powder injected into the plasma or hot gas (tail flame) exiting the generator
4) A bulk solid or melt which is contacted by the plasma or tail flame exiting the plasma generator
5) A bulk solid or melt, serving as an electrode to which one end of the arc is attached

In the first four cases, the electrodes serve only to maintain the current (nontransferred arcs); the reaction takes place in the region between the electrodes or downstream of the generator. The electrodes, which do not take part in the plasma process itself, are designed for the longest possible service life and low erosion rates.

In the fifth case, one of the electrodes is also the object or reaction medium being processed (transferred arc). The large amounts of energy transferred to a relatively small area at the phase boundary are used to heat the object. In a d.c. arc, the object being heated is generally the anode.

Transferred arcs are employed, above all, in metal fabrication and processing. Another example is the synthesis of calcium carbide in a quicklime – coke melt (→ Calcium Carbide, Chap. 4.) [25], [26]. In general, transferred arcs are utilized for heating and processing bulk substances that do not readily vaporize.

In contrast, nontransferred arcs are used chiefly for the heating and/or conversion of gaseous or readily vaporized substances, or for the processing of fine powders and droplets.

High-Current and High-Voltage Arcs. A given generator power can be attained at various current/voltage (I/U) ratios. The desired I/U ratio depends in practice on how the generator is employed. For example, the current should be as high as possible and the voltage as low as possible when a transferred arc is used (Fig. 1) and the material being processed is heated directly by the arc spot (i.e., the striking point of the arc on the electrode surface).

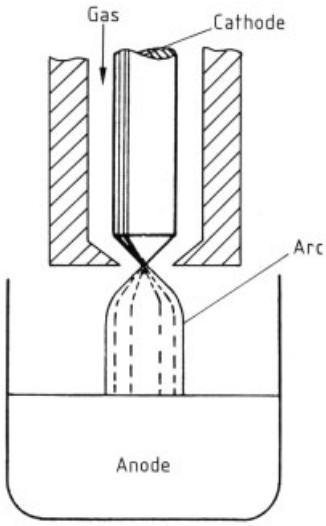

Figure 1. High-current generator with transferred arc

Short arcs with high I/U ratios (Fig. 2) are used as gas heaters when high energy densities must be imparted to small amounts of gas. Such generators are used, for example, in plasma spraying with high-melting powders, which are

introduced into the stream of plasma or hot gas (tail flame) downstream of the generator.

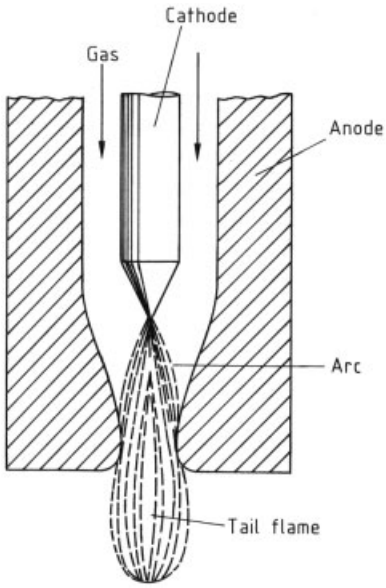

Figure 2. High-current generator with nontransferred arc

Long, high-voltage arcs are preferred when large quantities of gas are to be given moderate energy densities, that is, heated to a few thousand degrees (e.g., acetylene synthesis). The relatively large contact area between the arc and the flowing gas insures effective heat transfer to the medium and a high generator efficiency (80 – 90 %).

The length and thus the voltage of the arc can be influenced substantially by using a gas flow around the arc. This also reduces the spatial instability (flickering) typical of long arcs. A particularly effective way to stabilize an arc over a long discharge distance in a reactor is to place it in the center of a gas vortex (Fig. 3). Stabilization occurs because the arc avoids crossing streams of cold gas or penetrating into cold regions in order to minimize heat losses.

Figures 3 and 4 illustrate the designs of two frequently-used high-voltage d.c. discharge generators, which are equally suitable for gas heating and for chemical reactions in the gas phase.

In the design of Figure 3, the vortex-stabilized arc burns between two hollow electrodes made of steel or copper. A gas is fed tangentially into the generator between the electrodes. Features of this generator include:

1) A long arc, whose length depends on the flow conditions
2) Stable behavior of the arc over the whole discharge path
3) Rapid rotation of the arc spots over the inner surfaces of the electrodes, driven by the gas vortex

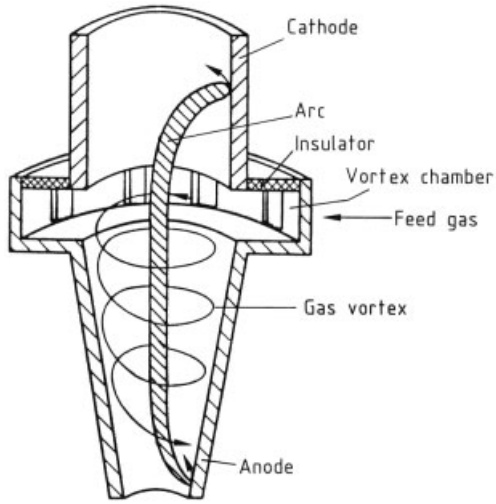

Figure 3. High-voltage generator with vortex-stabilized arc

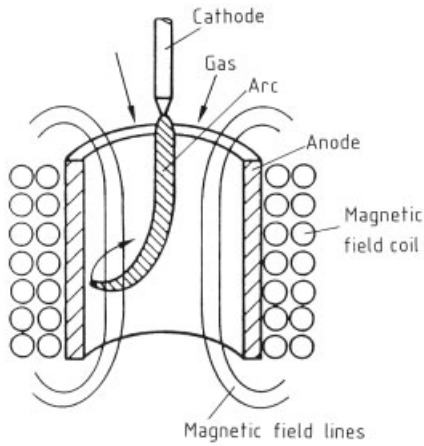

Figure 4. High-voltage generator with magnetically stabilized arc

In the magnetically stabilized arc (Fig. 4), the arc burns between a rod-shaped cathode, around which the gas flows, and a cylindrical anode, which is surrounded by a magnetic field coil that causes the anode arc spots to rotate.

The cathode in such a generator is made of a high-melting material (e.g., tungsten), which may be doped with certain oxides (ThO_2, CeO_2, Y_2O_3) to lower the work function. If very high currents (>15 kA) are used and the arc spots are stationary, severe material losses occur in the electrode region. In this case, graphite rod electrodes can be used with a feeding mechanism to maintain a constant distance between the electrodes.

Examples of the generator types shown in Figures 3 and 4, as used by Hüls and AVCO, are discussed in Chapter 5. Other generator manufacturers (e.g., Linde and Tioxide [19]) combine gas-flow stabilization with the use of a magnetic field to rotate the arc spots.

Electrode Shapes. The electrodes can vary widely in shape, depending on the power and generator type. Three shapes are widely used in plasma technology:

1) Rod electrodes, sometimes pointed to create a stationary arc spot on the tip
2) Flat electrodes with a stationary arc spot
3) Cylindrical or conical hollow electrodes such that the arc spots are stationary on the inner surfaces or, more commonly, rotate uniformly, driven by a gas vortex or by magnetic forces

Whether it is more desirable to have stationary or moving arc spots, or a combination of both, depends on the use of the generator.

Stationary arc spots are preferred:

1) When the material functioning as electrode is to be heated, melted, or evaporated by the arc spot (transferred arc)
2) When, in d.c. arc processes, the starting material at the cathode arc spot should be drawn directly into the hot center of the plasma (Maecker effect)

Moving arc spots, rotating about the inner surface of hollow electrodes, are preferred:

1) When, for the sake of long electrode life, the material must be protected against melting or vaporization
2) When the working gas flowing through the hollow electrodes is to be heated as uniformly as possible by the rapidly rotating arc

Intense electrode cooling is required in many cases if long electrode life is required.

2.2. RF Plasma Generators [27–29]

In contrast to arcs, RF discharges take place without electrodes. The plasma is sustained by capacitive or inductive coupling of an a.c. power supply. RF plasmas, unlike arc plasmas, are therefore free of contamination by vaporized electrode material. RF generators, however, have a comparatively low efficiency; typically only 40 to 60 % of the electrical energy delivered by the power supply is transferred to the plasma.

Capacitive coupling is seldom used. It is based on plasma generation by high-frequency alternating electric fields applied to the generator from outside.

Energy transfer by inductive coupling is employed far more often. The plasma generator is surrounded by an induction coil, which can be supplied with alternating current at a frequency ranging between 10^2 and 10^7 Hz, usually several megahertz. The oscillating magnetic field induces an eddy current in the reactor (Fig. 5). For this reason, the radial temperature distribution inside an RF plasma generator, in contrast to that in an arc-type generator, exhibits maxima near the wall (8000 – 10 000 K) and somewhat lower values near the axis of the generator.

The flow conditions in the reactor are complex and vary greatly with operating conditions. At the discharge end, the gas velocity is highest (tens of m/s) near the axis, declining toward the wall. Arc generators, in contrast, can have flow velocities of up to hundreds of meters per second.

3. Plasma Reactors

Single-Stage and Two-Stage Designs. When an electrical discharge is used for chemical and/or physical processes, either the generator itself or the component downstream of the generator performs the role of reactor. In the first case, the heating zone is identical with the reaction zone; the plasma generator and the reactor thus coincide.

In the second case, the plasma-generation zone and the reaction zone are spatially separated, so that one region can be unambiguously

identified as the generator and the other as the reactor. In the following, the reaction-engineering aspect of each arrangement is emphasized and the two designs are referred to as single-stage and two-stage reactors.

Chemical and/or Physical Reactions in Plasma Reactors. Chemical compounds cannot be formed in the extremely hot center of the discharge zone. The chemical reaction and the formation of stable molecules take place in the vicinity (one-stage reactor) or in the tail flame (two-stage reactor) of the discharge. These regions are still very hot and enriched with reactive particles diffusing out of the plasma zone. The extreme temperatures in the center of the discharge zone are sometimes used for the vaporization of powders.

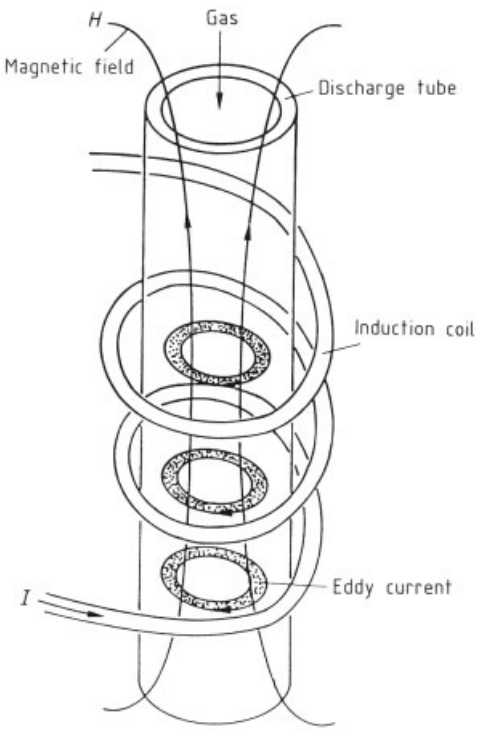

Figure 5. RF plasma generator

Many chemical and/or physical processes conducted in single-stage or two-stage thermal-discharge plasma reactors are indistinguishable, as far as the reaction mechanism is concerned, from high-temperature reactions in which energy is supplied to the reaction system in some other fashion. That is to say, the electrically charged particles (ions and electrons) that typify the plasma state are not necessarily crucial in determining how the reaction proceeds. For example, the gas-phase reactions in acetylene synthesis (Section 4.1) are essentially a sequence of very fast radical reactions.

There are, however, industrial applications in which charged particles diffusing from the discharge zone into the adjacent gas zone do play an important role, thus bringing about plasma-specific effects. For example, charged particles are good condensation nuclei for powder preparation from gaseous starting materials.

From the standpoint of reaction engineering, thermal plasmas are a suitable tool:

1) For supplying large amounts of energy very efficiently (e.g., for strongly endothermic reactions)
2) For gaining access to a high temperature range (1000 – 2500 K) in which certain desired products have their maximum thermodynamic stability
3) For generating reactive species at a high rate and thus achieving high space – time yields

4. Industrial Uses of Thermal Plasmas

Of the many important plasma processes now in use, five typical ones are described below.

4.1. Production of Acetylene from Hydrocarbons (see also → Acetylene, Chap. 4.3.)

Processes for obtaining acetylene from hydrocarbons, such as methane, require (1) large amounts of energy for the strongly endothermic reaction, (2) high reaction temperatures (above 1500 K) for reasons of thermodynamics and reaction kinetics, (3) short residence times (in the millisecond range), and (4) rapid cooling (quenching) of the reaction gas in order to prevent the unstable acetylene from decomposing to carbon black and hydrogen.

Hüls Arc Process [13], [14], [30], [31]. In the Hüls single-stage arc process (→ Acetylene, Chap. 4.3.1.), the first two of the above requirements are met by the arc discharge itself and the third by a high flow velocity of the reaction medium in the reactor; quenching at the end of the process is accomplished by the addition of water. This is followed by removal of carbon black from the product gas, separation of the gas components, and recycling of the unreacted starting materials.

For an annual output of 120 000 t acetylene (1990) Hüls now has available 19 arc reactors rated at up to 10 MW.

The heart of this single-stage arc process is a high-voltage arc burning in a vortex of the gaseous feed hydrocarbons. Downstream of the acetylene-forming main reaction zone, in the bottom part of the reactor, the reaction gas still has a temperature and enthalpy high enough to allow the formation of ethylene from liquid hydrocarbons. Liquid hydrocarbons that are particularly suitable as starting materials for ethylene synthesis are therefore sprayed into the reactor below the arc zone (prequench).

The thermal efficiency of the reactor is ca. 90 %, and about 50 % of the electrical energy is transformed into chemical energy. The specific energy requirement (for natural gas feedstock) is ca. 11 MWh per tonne of acetylene, and the acetylene yield is 0.5 t per tonne of feedstock.

A serious drawback to the Hüls single-stage arc process is the high yield of carbon black, which results from local overheating of the feed hydrocarbons in the immediate vicinity of the arc discharge. This problem was a major reason for the development of the two-stage process.

In the two-stage plasma arc process (→ Acetylene, Chap. 4.3.2.), hydrogen is heated by a vortex-stabilized arc discharge. The feed hydrocarbons are injected into the hot stream of hydrogen exiting the generator, and the process continues in the downstream reactor.

In comparison with the single-stage process, this two-stage version has the following advantages:

1) The more homogeneous spatial temperature distribution at the point where the starting materials are introduced, downstream of the generator, reduces local overheating of the hydrocarbons
2) Hydrogen, used as heat-transport medium, promotes the cracking reaction by supplying highly reactive hydrogen atoms
3) The possibility of introducing hydrocarbons in gaseous, liquid, or even solid form expands the range of possible feedstocks

Disadvantages include:

1) Additional energy consumption for recycling hydrogen from the refining stages
2) Low thermal efficiency (ca. 80 %)

As expected, the 8.5 MW prototype of a two-stage plasma reactor, operated on the same feedstocks, exhibited a much lower yield of carbon black than the single-stage process.

The two-stage reactor was also successful in the conversion of crude oil to acetylene.

Hoechst-Knapsack Process [32]. A variant of the Hüls two-stage process has been developed by Hoechst-Knapsack in collaboration with Hüls for the conversion of naphtha to acetylene (→ Acetylene, Chap. 4.3.2.). This is one of the few examples where a high-current arc furnace is used for heating gases.

In a 10 MW three-phase arc reactor with continuously advancing graphite electrodes, hydrogen, used as heat-transfer medium, is heated and then mixed with the feedstock. At present no full-scale plant has been constructed.

Acetylene Generation from Coal [33–36]. Both Hüls and AVCO have studied the conversion of powdered coal in arc reactors (→ Acetylene, Chap. 4.3.3.). Both firms used W/ThO_2 rod cathodes and cylindrical copper anodes, with a magnetic field to rotate the anode arc spot over the inner surface of the anode (cf. Fig. 4). The powdered coal was inlet to the reactor from the side of the cathode, carried in a stream of hydrogen; rapid and efficient heating took place in the immediate vicinity of the arc.

The experimental reactors were operated at powers from a few hundred kilowatts up to 1.5 MW. For a feedstock consisting of high-volatile gas-flame coals, the specific energy requirement varied from 10 to 15 MWh per tonne of acetylene, depending on process conditions.

Brown coal has also been used successfully in plasma reactors, but the high water and oxygen contents of this feedstock led to formation of

synthesis gas ($H_2 + CO$), which partly or completely replaces acetylene as reaction product.

4.2. Production of Reduction Gases

The pyrolytic processes leading to the formation of acetylene from hydrocarbons require that oxygen and substances containing oxygen be excluded. If steam, carbon dioxide, or oxygen is added to the hydrocarbon in a suitable proportion at the inlet to the plasma reactor, the product is a mixture of CO and H_2 (reduction or synthesis gas). This reforming reaction, like acetylene formation, occurs only at high temperatures.

The reduction gas can be used, still at a high temperature, for the direct reduction of iron ores, or it can be utilized for hydrogen production after conversion of the carbon monoxide.

On the basis of its single-stage arc process (see Section 4.1), Hüls has developed a natural-gas reforming process, which has now been licensed to USCO in the Republic of South Africa [37], [38]. The process uses four plasma reactors with a power of up to 8 MW, which are fed with a methane-containing gas and steam. In the plasma reactor the gas is converted to CO and H_2, then led directly to the reduction furnace. To overcome the hydraulic resistance of the ore bed, the plasma reactors operate at a pressure of 0.4 MPa.

The plasma processes for acetylene and reduction gas production described above have in common the advantage that the consumption of fossil feedstocks is reduced to just the amounts needed to form the products. The energy needed for heating the reactants and for the reaction enthalpy can be supplied via the plasma in the form of electrical energy, which can be generated from nonfossil sources.

In conventional autothermal processes, the energy needed to sustain the reaction is generated by burning part of the feed hydrocarbons.

SKF [39] has developed a process in which a mixture of oxygen, carbon dioxide, and steam is heated in one or several plasma generators. Finely divided coal is injected into the plasma stream, and the mixture is led directly to a coke-filled shaft furnace. The reduction gas formed in the vertical furnace at ca. 1000 °C is then employed for the reduction of oxides.

The optimum CO/H_2 ratio for the reduction can be obtained by the selection of the $O_2/H_2O/CO_2$ ratio. The SKF plasma generators have vortex stabilization and supplementary magnetic stabilization. High energy densities are obtained through the inclusion of electrically isolated spacers between cathode and anode, so that a long arc and thus a high power can be sustained with little gas. The plasma generators are operated at 6 – 7 MW.

In Sweden, along with a demonstration plant for iron ore reduction (Plasmared process), two commercial facilities for the reduction of ferromanganese and waste dust from the steel industry (Plasmadust process [40]) are now in operation.

4.3. Production and Processing of Powders [6], [41–45]

Production of Ceramic Powders. Very intensive work is being done on the preparation and processing of metallic and ceramic powders with plasma technology. Volatile halides are frequently used as starting materials. Under suitable conditions, these react to give:

1) Oxides in O_2 plasmas
2) Nitrides in N_2/NH_3 plasmas
3) Carbides in hydrocarbon plasmas
4) The elements in H_2 plasmas

In many cases, the products crystallize from the gas phase as ultrafine powders with mean grain diameters in the micrometer range and narrow particle size distributions. Examples of the many ceramic compounds prepared, mostly in small laboratory reactors, include the carbides and nitrides of silicon, aluminum, boron, zirconium, tungsten, hafnium, and titanium; and the oxides of silicon, titanium, aluminum, zirconium, and yttrium.

Because the halides are readily pyrolyzed, either arc or RF plasma reactors can be employed, both as single- or two-stage devices. Combined reactors (preheating with arc plasma, followed by an RF plasma) have also been recommended for complete conversion of the starting material [46].

In the area of ceramic powder synthesis, only the production of TiO_2 for pigment use has become established industrially. Tioxide [19], [20]

reacts the starting material titanium tetrachloride in 2 MW reactors operated with oxygen as plasma gas. The products are titanium dioxide and chlorine. The capacity for pigment-grade powder made in plasma reactors is 50 000 t/a.

Not only gaseous, but also liquid and solid feedstocks can be used for production of powders.

Plasma Processing of Ceramic and Metallic Powders. The partial or complete melting of powder grains with the aid of plasma reactors is used chiefly in the following three areas:

In *spheroidizing* [47], [48], irregularly shaped powder grains are transformed to a spherical shape by melting and resolidification. This operation markedly improves the free-flowing qualities of the powder. Good transport properties are desired, for example, in ceramic powders used for plasma spraying.

The *GTE process* [49] for the preparation of mixed powders and alloys comprises two steps. In the first, agglomerate powder particles are made from two starting powders with the addition of a binder in a spray drying operation. In the second step, the agglomerates are melted in the plasma to form compact particles. Under suitable conditions, the two components can also be homogenized in this step.

In *surface coating* [50], ceramic or metal particles are injected into a plasma or hot gas stream. The molten particles impact on the surface at a high velocity (up to 500 m/s for arc plasmas) and, after cooling, form highly adherent protective coatings. Depending on the type of spray powder used and the substrate, the resulting surface coatings enhance the mechanical durability of the workpiece, improve its corrosion resistance, or afford good electrical or thermal insulation.

Both arc and RF plasma generators with powers of roughly 10 to 100 kW are used for plasma spraying.

For more details on these processes, see → Metals, Surface Treatment, Chap. 5.).

4.4. Diamond and Diamond-like Coatings [51–53]

Plasma technology is being increasingly used to coat surfaces with amorphous carbon films (a-C : H) and diamond films. Low- and high-pressure plasmas are employed for this process. High carbon deposition rates can be achieved, especially with thermal plasmas.

A conceivable apparatus could, for instance, feature a carbon-containing gaseous mixture being added to the tail flame (cf. Fig. 2) of an arc discharge. The plasma gas can be, for example, argon and/or hydrogen. The plasma jet, enriched by carbon-containing reactive molecules or fragments of molecules (e.g., radicals and atoms) is led to the substrate which is to be coated. The reactive particles generated in the plasma impinge on the surface where they react to form the diamond film.

4.5. Pollution Control [19], [54], [55]

A relatively new application of plasma technology is the disposal of problem waste in a process utilizing the high temperatures available in the plasma.

The type and quantity of materials to be disposed of dictate the power and design of the reactor. Arc plasma reactors are generally used. Typical examples of projects either in development or already tested on a plant scale are:

1) Destruction of chlorine-containing substances, especially those containing PCBs
2) Melting of toxic and radioactive materials into glass-like masses or slags
3) Treatment of dusts containing heavy metals from the metal-refining industry
4) Disposal of medical waste

Oxygen plasmas are usually used to destroy organic compounds.

5. Diagnostic Methods

Development of plasma processes up to now has been mainly empirical. The optimal use of such processes, however, demands that diagnostic methods be increasingly employed as an aid in development. Information needed for the construction of process models, including the time and space dependences of the temperatures, concentrations, and velocities of the reacting particles, is provided by such methods.

The preferred techniques are optical ones, which involve no contact and hence no interference with the process. Emission spectroscopy, a passive method, uses the radiation emitted by the plasma to measure data such as temperature and density [56]. In contrast, absorption spectroscopy and scattering methods are active techniques in which laser beams are directed into the plasma. While the absorption methods often fail because of inadequate spatial resolution, scattering methods [57] are becoming more and more important, especially where complex reaction-gas mixtures are involved whose properties vary over time and space.

Coherent anti-Stokes Raman spectroscopy (CARS), in particular, offers important advantages [58]. These include directional emission of the measuring signal and the ability to discriminate this signal from the high intrinsic background emission of the plasma source. As a result, the technique can be used even in industrial processes that are not readily accessible and which produce high levels of contamination. The method is already used in industrial combustion systems, such as motors and turbine engines, and is now being adopted for the examination of plasma-chemical processes [59].

In contrast, laser-induced fluorescence spectroscopy (LIF) [58] has limited applicability when collisional quenching of the laser-excited electronic energy levels of molecules leads to invalid signals and when the high background radiation level makes the measuring signal hard to detect.

Only in simple systems can Rayleigh scattering be used to determine density and temperature. In industrial plasma processes with high contamination levels, the dominant Mie scattering from particles makes it difficult or impossible to use this technique [57].

Mie scattering does, however, find practical use in laser Doppler anemometry for the measurement of flow velocities [60].

6. References

1. A. von Engel: *Electric Plasmas – Their Nature and Uses,* Taylor and Francis Ltd., London – New York 1983.
2. M. N. Hirsch, H. J. Oskam (eds.): *Gaseous Electronics,* Academic Press, New York 1978.
3. W. Finkelnburg, H. Maecker: "Elektrische Bögen und thermisches Plasma" in: *Handbuch der Physik,* vol. XXII, Springer Verlag, Berlin – Göttingen – Heidelberg 1956, p. 254.
4. V. Dembrovsky: "Plasma Metallurgy," *Mater. Sci. Monogr.* **23** (1985).
5. G. Ecker, K. U. Riemann, A. Schuhmacher: *Zum Technologiepotential der Plasmaphysik,* VDI Verlag, Düsseldorf 1985.
6. P. Fauchais: "Utilisation Industrielle, Actuelle et Potentielle des Plasma: Synthesis, Traîtement des Poudres, Traîtements Métallurques, Traîtement de Surface," *Rev. Phys. Appl.* **15** (1980) 128.
7. S. M. Rossnagel, J. J. Cuomo, W. D. Westwood (eds.): *"Handbook of Plasma Processing Technology," Noyes Data Corp.,* Park Ridge, N.J. 1990.
8. L. A. Ettlinger, T. D. Nainan, R. P. Quellete, P. N. Cheremisinoff: "High-Temperature Plasma Technology Applications," *Electrotechnology (Ann Arbor Mich.)* **6** (1980).
9. H. Drost: *Plasmachemie,* Akademie-Verlag, Berlin 1978.
10. S. Vepřek: "Plasmachemie heterogener Systeme," *Nachr. Chem. Tech. Lab.* **36** (1988) no. 3, 248.
11. R. F. Baddour, R. S. Timmins: *The Application of Plasmas to Chemical Processing,* M.I.T. Press, Cambridge, Mass. 1976.
12. K. Birkeland, S. Eyde, *Electrotech. Ind.* **2** (1904) 399.
13. H. Gladisch: "How Hüls Makes Acetylene by DC Arc," *Hydrocarbon Process. Petr. Refin.* **41** (1962) 195.
14. R. Müller, G. Kaske: "The Use of Plasma Chemical Processes for Chemical Reactions," *Erdöl Kohle Erdgas Petrochem.* **37** (1984) no. 4, 149.
15. U. Kogelschatz: "Ozonherstellung," *Elektrowärme Int. Ed. B* **45** (1987) B 3/4.
16. R. C. Rice, A. Netzer: *Handbook of Ozone Technology and Applications,* **vols. 1 and 2,** Ann Arbor Sci. Publ., Ann Arbor, Michigan 1982/1984.
17. E. Broszeit et al. (eds.): *Plasma Surface Engineering,* vols. 1 und 2, DGM Informationsgesellschaft, Oberursel 1989.
18. J. Feinmann (ed.): *"Plasma Technology in Metallurgical Processing,"* Iron and Steel Society, Warrendale 1987.
19. M. Orfeuil et al.: "Arc Plasma Processes – A Maturing Technology in Industry," *UIE Arc Plasma Review 1988,* International Union for Electroheat, Paris 1988.

20. Battelle – Geneva Research Centre: *Assessment of Prospects for Thermal Plasma Furnace Technology,* Multiclient Study, Carouge – Genève 1984.
21. D. R. MacRae: "Plasma Arc Process System Reactors and Applications," *Plasma Chem. Plasma Process.* **9** (1989) no. 1, 85 pp. (Supplement).
22. E. Pfender: "Electric Arcs and Arc Gas Heaters," in: *Gaseous Electronics,* vol. I, Chap. 5, Academic Press, New York 1978, p. 291.
23. E. Pfender, M. Boulos, P. Fauchais: "Methods and Principles of Plasma Generation," in: *"Plasma Technology in Metallurgical Processing,"* Iron and Steel Society, Warrendale 1987, Chap. 4, p. 27 ff.
24. P. Mogensen, J. Thörnblom: "Electrical and Mechanical Technology of Plasma Generation and Control," in: *"Plasma Technology in Metallurgical Processing,* Iron and Steel Society, Warrendale 1987, Chap. 6, p. 65 ff.
25. G. Strauß in: *Chemierohstoffe aus Kohle,* Thieme Verlag, Stuttgart 1977.
26. *Ullmann,* 4th ed., **9,** 70.
27. H. U. Eckert: "The Induction Arc: A State-of-Art Review," *High Temp. (Engl. Transl.)* **6** (1974) 99.
28. M. I. Boulos: "The Inductively Coupled RF Plasma," *Pure Appl. Chem.* **57** (1985) 1321.
29. P. N. Cheremisinoff, O. G. Farah, R. P. Quellette (eds.): *Radio Frequency / Radiation and Plasma Processing,* Chap. I, Technomic Publ. Co., Lancaster, PA 1985.
30. *Ullmann,* 5th ed., **A1,** 115.
31. H. Brachhold, R. Müller, G. Proß: "Status of Plasma Technology and Development Work of Hüls AG," *9th Int. Symp. Plasma Appl. Workshop Proc.* **2** (1989) 95.
32. K. Gehrmann, H. Schmidt, *Water Air Conserv. Pet. Ind.* 1971, 379.
33. C. Peuckert, R. Müller: "Acetylene from Coal by the Hüls Plasma Process – Different Operating Conditions and their Consequences for the Process," *Symp. Proc. Int. Symp. Plasma Chem. 7th,* **1** (1985) 274.
34. R. Müller, C. Peuckert: "Recent Developments for the Production of Acetylene from Coal by the Hüls Arc Process," *Symp. Proc. Int. Symp. Plasma Chem. 6th,* **1** (1983) 270.
35. R. Müller et al.: "Production of Acetylene from Coal by the Hüls Plasma Process," *Symp. Proc. Int. Symp. Plasma Chem. 8th,* **1** (1987) 707.
36. H. Regner, C. Peuckert, D. Bittner, H. Brachhold: "Production of Acetylene from Coal by the Hüls Plasma Process," *9th Int. Symp. Proc.* **III** (1989) L. 117.
37. L. Kerker, R. Müller: "Das Plasmareforming-Verfahren zur Erzeugung von Reduktionsgasen," *Stahl Eisen* **104** (1984) no. 22, 1137.
38. R. Müller, G. Kaske, L. Kerker: "Hüls Plasmareforming for the Production of Basic Chemicals," *Symp. Proc. Int. Symp. Plasma Chem. 8th,* **1** (1987) 701.
39. S. Santén, L. Bentell, B. Johansson, P. Westerlund in: *"Plasma Technology in Metallurgical Processing,"* Iron and Steel Society, Warrendale 1987, p. 112.
40. S. Santén, L. Bentell, B. Johansson, P. Westerlund in: *Plasma Technology in Metallurgical Processing,"* Iron and Steel Society, Warrendale 1987, p. 127.
41. P. Fauchais, E. Boudrin, J. F. Coudert, R. McPherson: "High Pressure Plasmas and Their Application to Ceramic Technology," *Top. Curr. Chem.* **107** (1983).
42. R. M. Young, E. Pfender: "Generation and Behaviour of Fine Particles in Thermal Plasmas – A Review," *Plasma Chem. Plasma Process.* **5** (1985) 1.
43. K. Akashi, *Pure Appl. Chem.* **57** (1985) 1197.
44. E. Pfender, M. Boulos, P. Fauchais: "Thermal Plasma Synthesis of Ceramic Powders," International Thermal Plasma Engineering, 1988, Report No. 5.
45. O. Schulz, H. Hausner: "Plasmasynthese keramischer Sinterpulver für Hochleistungs-Keramik," *Elektrowärme Int. Ed. B* **45** (1987) B 3/4 (June/Aug.) 174.
46. T. Yoshida, T. Tani, H. Nishimura, K. Akashi, *J. Appl. Phys.* **54** (1983) 640.
47. M. G. Fey, C. B. Wolf, F. J. Harwey: *Magnetite Spheroidization Using an AC Arc Heater,* Int. Round Table on Study and Application of Transport Phenomena in Thermal Plasma, Oelleilo 1975.
48. D. Gold et al.: "A 100 kW Three-Phase AC Plasma Furnace for Spheroidization of Aluminium Silicate Particles," *Plasma Chem. Plasma Process.* **1** (1981) no. 2, 161.
49. D. L. Houck: "Powders for Plasma Spray Coatings," *Prog. Powder Metall.* **34/35** (1978/79) 207.
50. D. A. Gardeman, N. L. Hecht: "Arc Plasma Technology in Material Science," *Applied Mineralogy,* **vol. 3,** Springer Verlag, Wien – New York 1972.

51. U. Ehlemann, H. G. Lergon, K. Wiesemann (eds.): *Int. Symp. Plasma Chemistry 10th,* **vol. 3,** chap. 3.1.
52. P. K. Bachmann, D. Leers, H. Lydtkin; "Towards a General Concept of Diamond Chemical Vapour Deposition," *Diamond and Related Materials* **1** (1991) no. 1, 1.
53. Z. P. Lu et al.: "Diamond Synthesis by DC Thermal Plasma CVD," *Plasma Chem. Plasma Process.* **11** (1991) no. 3, 387; **12** (1992) no. 1, 55.
54. R. C. Eschenbach: "Use of Plasma Torches for Melting Special Metals and for Destroying and Stabilizing Hazardous Wastes," *9th Int. Symp. Plasma Chem. Workshop Ind. Plasma Appl. Workshop Proc.* **2** (1989) 127.
55. J. V. R. Heberlein, W. J. Melilli, S. V. Digle, W. H. Reed: "Adaption of Non-Transferred Plasma Torches to the New Applications of Plasma Systems," *Workshop Ind. Plasma Appl. Workshop Proc.* **2** (1989) 1.
56. H. R. Griem: *Plasma Spectroscopy,* McGraw-Hill Book Co., New York 1964.
57. A. Leipertz: "Modular aufgebautes Laser System zur Bestimmung physikalischer Größen in Gasen," *Chem. Ind. (Düsseldorf)* **37** (1985) 802.
58. A. C. Eckbreth, P. A. Bonczyk, J. F. Verdieck: "Combustion Diagnostics by Laser Raman and Fluorescence Techniques," *Prog. Energy Combust. Sci.* **5** (1979) 253.
59. K. Müller-Dethlefs, M. Pealat, J.-P. E. Taran: "Temperature and Hydrogen Concentration Measurements by CARS in an Ethylen-Air-Bunsen Flame," *Ber. Bunsen Ges. Phys. Chem.* **85** (1981) 803.
60. F. Durst, A. Melling, J. H. Whitelaw: *Principle and Practice of Laser-Doppler-Anemometry,* Academic Press, New York 1976.

Plasma Spraying → Metals, Surface Treatment
Plasma Sputtering → Ceramics, Ceramic-Metal Systems
Plaster → Calcium Sulfate

Plasticizers

David F. Cadogan, ICI Chemicals and Polymers, Runcorn, Cheshire, United Kingdom

Christopher J. Howick, ICI Chemicals and Polymers, Runcorn, Cheshire, United Kingdom

1. Introduction 2
1.1. Definition 2
1.2. Types of Plasticization 2
1.3. Types of Plasticizer 2
2. Plasticizers in Common Use 2
2.1. Market Overview 2
2.2. Phthalate Esters 2
2.3. Adipate Esters 4
2.4. Trimellitate Esters 4
2.5. Phosphate Esters 5
2.6. Sebacate and Azelate Esters 5
2.7. Polyester Plasticizers 5
2.8. Sulfonate Esters 5
2.9. Specialty Phthalate Esters 5
2.10. Secondary Plasticizers 6
3. Mechanism of Plasticizer Action . . 6
3.1. The Lubricity Theory 6
3.2. The Gel Theory 6
3.3. The Free-Volume Theory 6
3.4. Solvation – Desolvation Equilibrium 7
3.5. Generalized Structure Theories . . . 7
3.6. Specific Interactions and Interaction Parameters 8
4. Plasticized PVC 9
4.1. Applications Technology 9
4.1.1. Suspension PVC 10
4.1.2. Emulsion PVC 10
4.2. Effect of Plasticizer Choice on the Properties of Flexible PVC 10
4.2.1. Plasticizer Efficiency 10
4.2.2. High-Temperature Performance 11
4.2.3. Low-Temperature Performance 12
4.2.4. Gelation Properties 12
4.2.5. Migration and Extraction 13
4.2.6. Plastisol Viscosity and Viscosity Ageing 13
4.2.7. Windscreen Fogging 14
4.2.8. Summmary 14
5. Plasticization of Polymers Other Than PVC 15
5.1. Requirements for Plasticization . . . 15
5.2. Plasticization of Other Polymers . . 15
6. Toxicology and Environmental Aspects 17
6.1. Toxicity Studies on Plasticizer Esters 17
6.2. Plasticizers in the Environment . . . 18
6.3. Storage and Handling 20
7. References 20

1. Introduction

1.1. Definition

A plasticizer is a substance incorporated into a material to increase its flexibility, workability, or distensibility. A plasticizer may reduce the melt viscosity, lower the temperature of the second-order transition, or lower the elastic modulus of the product. Plasticizers are inert, organic substances with low vapor pressures, predominantly esters, which react physically with high polymers to form a homogeneous physical unit, whether it be by means of swelling or dissolving or any other [1].

At present some 300 plasticizers are manufactured, of which at least 100 are of commercial importance. A list of some of the most widely used plasticizers is given in Table 1.

1.2. Types of Plasticization

Two principle ways of obtaining the plasticization effects described above are commonly utilized. Firstly, a rigid polymer may be *internally plasticized* by chemically modifying the polymer or the monomer so that the flexibility of the polymer is increased. A rigid polymer, however, can also be *externally plasticized* by addition of a plasticizer, which imparts the desired flexibility but is not chemically changed by reaction with the polymer. The latter process has the advantage of lower cost and permits the fabricator freedom in devising formulations to manufacture items of differing flexibility simply by adding different amounts of plasticizer, within the constraints of polymer – plasticizer compatibility.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_439

Table 1. Plasticizers in Common Use

Plasticizer	Abbreviation	Alcohol carbon number	CAS registry no.	M_r	Density at 20 °C, g/cm^3
Phthalates					
Diisobutyl phthalate	DIBP	4	[*84-69-5*]	278.3	1.039
Dibutyl phthalate	DBP	4	[*84-74-2*]	278.3	1.046
Diisoheptyl phthalate	DIHP	7	[*41451-28-9*] [*71888-89-6*]	362	0.991
L79 phthalate	L79P, 79P	7 – 9		380	0.985
L711 phthalate	L711P, 711P	7 – 11	[*68648-91-9*]		0.971
Dioctyl phthalate	DOP	8	[*117-81-7*]	390	0.984
Diisooctyl phthalate	DIOP	8	[*27554-26-3*]	390	0.983
Dinonyl phthalate	DNP	9	[*84-76-4*]	419	0.97
Diisononyl phthalate	DINP	9	[*28553-12-0*] [*68515-48-0*]	419	0.975
Diisodecyl phthalate	DIDP	10	[*68515-49-1*]	447	0.967
L911 phthalate	L911P, 911P	9 – 11	[*68515-43-5*]	454	0.96
Diundecyl phthalate	DUP	11	[*3648-20-2*]	474	0.953
Diisoundecyl phthalate	DIUP	11	[*85507-79-5*]	474	0.962
Undecyl dodecyl phthalate	UDP	11 – 12	[*68515-43-5*]	488	0.957
Diisotridecyl phthalate	DTDP	13	[*27253-26-5*]	531	0.952
Butyl benzyl phthalate	BBP	4 – 7	[*68515-47-9*]	312.3	1.119
Adipates					
Dioctyl adipate	DOA	8	[*103-23-1*]	370.6	0.929
Diisononyl adipate	DINA	9	[*33703-08-1*]	398	0.929
Diisodecyl adipate	DIDA	10	[*27178-16-1*]	427.1	0.915
Trimellitates					
Trioctyl trimellitate	TOTM	8	[*3319-31-1*]	530	0.991
L79 trimellitate	L79TM	7 – 9	[*68515-60-6*]	530	0.996
L810 trimellitate	L810TM	8 – 10	[*67989-23-5*]	592	0.973
Phosphate esters					
Tri-2-ethylhexyl phosphate	TOF	8	[*78-42-2*]	434	0.926
2-Ethylhexyl diphenyl phosphate	Santicizer 141	6 and 8	[*1241-94-7*]	362.4	1.091
Tricresyl phosphate	TCP	6	[*1330-78-5*]	368.2	1.165

1.3. Types of Plasticizer

Plasticizers can be divided into two principal groups: primary and secondary. *Primary plasticizers* lower the glass transition temperature T_g and increase the elongation and softness of the polymer. *Secondary plasticizers*, when added to the polymer alone, do not bring about such changes and may also have limited compatibility with the polymer. However, when added to the polymer in the presence of a primary plasticizer, secondary plasticizers enhance the plasticizing performance of the primary plasticizer.

2. Plasticizers in Common Use

2.1. Market Overview

The list of commercially available plasticizers, covering applications throughout polymer science, is extensive. This reflects the development of plasticizers to meet the increasing demand for polymers, especially PVC, for new and higher quality applications. This article discusses the main groups of plasticizers; for further details, see [1–4] and references therein.

The Western European plasticizer market had a capacity of 969×10^3 t/a in 1989 and at that time was growing at a rate of 3 % per annum [5]. In 1996 production of plasticizer in Western Europe amounted to 1253×10^3 t/a and in the United States to 636×10^3 t/a [6]. In terms of plasticizer types, the majority of this tonnage (>85 %) is standard phthalate (esters of phthalic anhydride with C_8-C_{10} alcohols). Reasons for this are the relatively low price and ready availability of feedstocks. The remainder of the market is taken up by phthalate esters of other alcohols, speciality phthalates, adipates, trimellitates, and other esters.

2.2. Phthalate Esters (see also → Phthalic Acid and Derivatives, Chap. 5.)

Diesters of phthalic anhydride are produced commercially from $C_1 - C_{13}$ alcohols, although the $C_1 - C_3$ esters are generally too volatile to find widespread use. Di-*n*-butyl phthalate (DBP) and diisobutyl phthalate (DIBP) (linear and branched C_4 esters) are used in many PVC formulations, principally for ease of gelation. Owing to their relatively high volatility, in comparison with other phthalates, they are often used in conjunction with higher molecular mass esters. Diisopentyl phthalate (DIPP) is generally used in a similar manner. Diisoheptyl phthalate (DIHP), an ester based on a C_7 iso-alcohol from the oxo alcohol process, shows good plasticizing action with PVC as well as with a number of other polymers [4]. Its principle use in PVC, when compared with higher molecular mass plasticizers, lies in its gelation characteristics and balance of volatility and viscosity, the volatility of DIHP being only slightly higher than the corresponding C_8 phthalate (DIOP).

Di-2-ethylhexyl phthalate (DEHP) (also known as dioctyl phthalate, DOP) has long been accepted as the industry standard general purpose primary plasticizer.

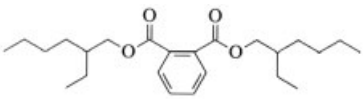

Di-2-ethylhexyl phtalate, DEHP
(dioctyl phthalate, DOP)

This is because its properties are adequate for many applications in the flexible vinyl industry. DEHP shows good gelation characteristics, good softening action, and adequate viscosity properties in emulsion PVC pastes. Diisooctyl phthalate, produced from C_8 oxo alcohols, shows performance in PVC which is generally similar to DEHP. Di-*n*-octyl phthalate (DnOP), a C_8 phthalate ester derived from a linear alcohol, is also produced, but in far smaller quantities than the branched phthalates. The linear structure imparts improved low-temperature performance and lower viscosity.

Diisononyl phthalate (DINP) is produced from C_9 oxo alcohols and has similar plasticizing properties to DEHP. Its slightly higher molecular mass results in a somewhat lower plasticizer efficiency which requires a greater addition of plasticizer to impart an equivalent softness (these effects are discussed in detail in Section 4.2.1). The higher molecular mass alcohol also imparts improved high-temperature performance and resistance to extraction. Other C_9 phthalates are also produced [e.g., DNP, DINP(S)] and vary in the linearity of the alcohol component in the ester, the more linear varieties giving improved properties.

Diisodecyl phthalate (DIDP), the phthalate ester of C_{10} oxo alcohols, exhibits superior high-temperature performance to DINP and DEHP due to its higher molecular mass (lower volatility).

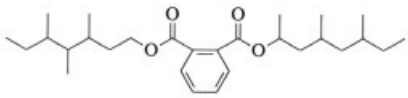

Diisodecyl phthalate, DIDP
(one of the possible isomers)

Its lower plasticizing efficiency compared to the lower phthalates requires greater plasticizer addition to impart the same softness, although this can, in some cases, be advantageous. Diundecyl phthalate (DUP), diisoundecyl phthalate (DIUP) and undecyl dodecyl phthalate (UDP) continue this trend. Ditridecyl phthalate (DTDP) is the highest molecular mass phthalate ester generally available in commercial quantities. Its high molecular mass results in low plasticizing efficiency, and some compatibility problems may occur at high plasticizer loadings. Its low volatility, however, also results in superior high-temperature performance and allows some strict electrical cable insulation and sheathing specifications to be met.

2.3. Adipate Esters

Alkyl esters of adipic acid have two principle advantages in the plasticization of PVC resins in comparison with the corresponding phthalate ester: (1) their lower intrinsic viscosity confers lower plastisol viscosity and good plastisol storage stability, enabling use in demanding applications and (2) they give superior low-temperature flexibility properties.

Of adipate esters in use in the flexible vinyl industry, di-2-ethylhexyl adipate, DEHA (also

known as dioctyl adipate, DOA) is the most common.

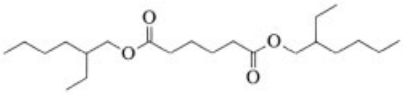

Di-2-ethyl hexyl adipate, DEHA
(dioctyl adipate, DOA)

It is widely used in flexible PVC food film (cling film). DEHA has good plasticizing efficiency and compatibility with PVC, and its very low viscosity enables it to be used in PVC plastisol applications where low viscosity is required, often in conjunction with a phthalate. Diisononyl adipate (DINA) is the ester of adipic acid with isononanol produced from oxo alcohols. It shows similar properties to DEHA. Its higher molecular mass does, however, increase its viscosity compared with DEHA, but this is accompanied by a reduction in volatility. The lower efficiency in comparison with DEHA, which requires a greater plasticizer addition, can result in lower plastisol viscosities. These effects are extended in moving to diisodecyl adipate (DIDA). Higher adipates (e.g., ditridecyl adipate, DTDA) are of limited use due to their lower compatibility with PVC.

2.4. Trimellitate Esters

Esters of trimellitic anhydride (TMA) have grown in popularity due to their excellent thermal properties and resistance to extraction. Tris-2-ethylhexyl trimellitate (TOTM), produced from trimellitic anhydride and 2-ethylhexanol, is the ester most widely used.

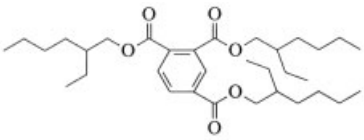

Tris-2-ethylhexyl trimellitate
(trioctyl trimellitate, TOTM)

Although of a lower plasticizing efficiency than the corresponding phthalate, the thermal stability and extraction resistance of TOTM are superior. These properties can be improved still further by using trimellitate esters produced from mixed linear alcohols, e.g., L79 trimellitate and L810 trimellitate. The improved properties brought about by these esters are discussed further in Chapter 4.

2.5. Phosphate Esters

The principle advantages of phosphate esters such as tricresyl phosphate as plasticizers for PVC is their low volatility and their ability to impart fire-retardant properties to a PVC formulation.

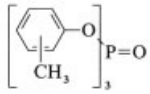

Tricresyl phospate

Tris(2-ethylhexyl) phosphate shows good compatibility with PVC and also imparts good low-temperature performance in addition to good fire retardancy.

2-Ethylhexyl diphenyl phosphate has widespread use due to its combination of plasticizing efficiency, low-temperature properties, migration resistance, and fire retardancy.

2.6. Sebacate and Azelate Esters

Esters from the linear sebacic and azelatic acids are used in PVC applications for the same reasons as the corresponding adipates, i.e., low viscosity and good low-temperature performance. Di-2-ethylhexyl sebacate (DOS) and di-2-ethylhexyl azelate (DOZ) are the most widely used.

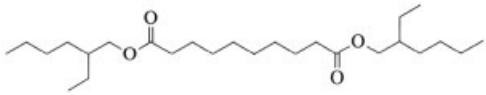

Di-2-ethylhexyl sebacate, DOS

In comparison to adipates, however, these esters tend to have a significantly higher cost and consequently only find use in extreme low-temperature applications.

2.7. Polyester Plasticizers

Polyester plasticizers, or polymeric plasticizers as they are more generally known, have found

use due to their low volatility and high extraction resistance.

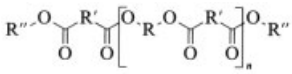

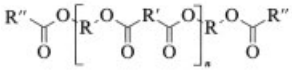

Polymeric plasticisers

They are typically based on condensation products of propane- or butanediols with adipic acid or phthalic anhydride. The growing polymer chain may then be end-capped with an alcohol or a monobasic acid, although non-end-capped polyesters can be produced by strict control of the reaction stoichiometry. Due to their higher molecular mass compared to other plasticizers they have much higher viscosities (3 – 10 Pa · s), which in some cases can make processing with PVC difficult.

2.8. Sulfonate Esters

Sufonate ester plasticizers for PVC are typically aryl esters of a $C_{13} - C_{15}$ alkane sulfonic acid. They are relatively efficient and easily processable plasticizers with good extraction properties.

2.9. Specialty Phthalate Esters

In addition to the phthalate plasticizers described in Section 2.2, other phthalate plasticizers are also commercially available. These are speciality products which find use as high-quality plasticizers in specific applications. Their quality tends to be reflected in their price since they require either a further manufacturing stage (e.g., transesterification) or the use of alcohol feedstocks of greater linearity than the standard phthalates.

Benzyl butyl phthalate (BBP) shows excellent compatibility and fusion characteristics with PVC and is widely used in vinyl flooring for these reasons and also due to its superior stain resistance.

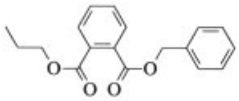

Benzyl butyl phthalate, BBP

Linear and semilinear phthalates (e.g., DnOP, 911P, DUP) have lower viscosities than the corresponding branched ester. This results in lower plastisol viscosity and also lower plasticizer volatility. 911P, a phthalate ester based on linear and semilinear $C_9 - C_{11}$ alcohols, is the premium plasticizer for flexible PVC in automotive applications due to its low fogging (the condensation of volatile material on a car windscreen, causing impaired visibility); see Section 4.2.7.

2.10. Secondary Plasticizers

Secondary plasticizers (also referred to as extenders) continue to play a major role in flexible PVC technology. They do not have the same mode of softening action as a primary plasticizer but do impart softening behavior when used in combination with a primary plasticizer. In some very hard PVC formulations, however, a secondary plasticizer alone may be of use. The major secondary plasticizers are chlorinated hydrocarbons. These are hydrocarbons chlorinated to varying amounts, typically 30 – 70 %. For a given hydrocarbon, plasticizer viscosity increases with increasing chlorine content. Chlorinated hydrocarbons of the same chlorine content may, however, have different volatilities and viscosities if they are based on different hydrocarbons. The principal advantages of such materials are: (1) they improve fire retardancy due to their chlorine content and (2) formulations containing a secondary plasticizer are often cheaper. Precise knowledge of compatibility between standard plasticizers and chlorinated paraffins is required since some mixtures become incompatible with each other and the PVC resins at certain concentrations. Phthalate – chlorinated hydrocarbon compatibility decreases as the molecular mass of the phthalate and the plasticizer loading in the PVC increases; for specific examples, see [1].

3. Mechanism of Plasticizer Action

Several theories of plasticizer action have been proposed, ranging in detail and complexity. Whilst each theory is not exhaustive, an understanding of the plasticization process can be gained by combining ideas from each theory. The theories are discussed in detail in [1].

3.1. The Lubricity Theory

The lubricity theory is based on the assumption that the rigidity of the resin arises from intermolecular friction binding the chains together in a rigid network. On heating, these frictional forces are weakened so as to allow the plasticizer molecules between the chains. In the resulting mixture at room temperature the plasticizer molecules act as lubricants for the polymer chains.

3.2. The Gel Theory

The gel theory extends the lubricity theory in that it deals with the idea of the plasticizer acting by breaking the resin – resin attachments and interactions and by masking these centers of attachment from each other, preventing their reformation. Such a process may be regarded as being necessary but by itself insufficient to explain a completely plasticized system since while a certain concentration of plasticizer molecules will provide plasticization by this process the remainder will act more in accordance with the lubricity theory, with unattached plasticizer molecules swelling the gel and facilitating the movement of plasticizer molecules, thus imparting flexibility. Molecules acting by this latter action may in fact constitute the bulk of plasticizer molecules.

3.3. The Free-Volume Theory

The free-volume theory extends the above ideas and also allows a quantitative assessment of the plasticization process. The free volume of a polymer is described by Equation 1 and Figure 1

$$V_f = V_t - V^0 \quad (1)$$

where

V_f = the free volume of the polymer
V_t = specific volume at temperature t
V^0 = specific volume at arbitrary reference point, usually 0K

and V_G^0 and V_L^0 are the specific volumes of the glass and liquid extrapolated to 0K.

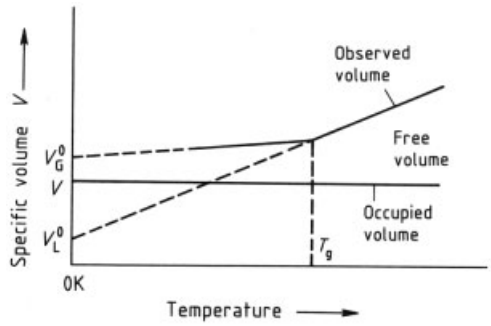

Figure 1. The free-volume concept [4]

Free volume is a measure of the internal space available in a polymer for the movement of the polymer chain, which imparts flexibility to the resin. Figure 1 shows how the amount of free volume in the polymer increases sharply at the glass transition temperature T_g. Consequently, the study of plasticization is a study of how to lower T_g, thus creating a polymer that is flexible at ambient temperature. Plasticizers therefore increase the free volume of the resin and also ensure that free volume is maintained as the resin – plasticizer mixture is cooled from the melt. Combining these ideas with the gel and lubricity theories implies that plasticizer molecules that do not interact with the polymer chain must simply fill free volume created by those molecules that do. These molecules may also be envisaged as providing a screening effect that prevents interactions between neighboring polymer chains, thus preventing reformation of the rigid polymer network on cooling.

For the plasticized resin, free volume can arise from:

1) Motion of the chain ends
2) Motion of the side chains
3) Motion of the main chain

These motions can be increased in a variety of ways, including:

1) Increasing the number of end groups
2) Increasing the length of the side chains
3) Increasing the possibility of main chain movement by the inclusion of segments of low steric hindrance and low intermolecular attraction
4) Introduction of a lower molecular mass compound which imparts the above properties
5) Raising the temperature

The introduction of a plasticizer, which has a lower molecular mass than the resin, can impart a greater free volume per volume of material since (1) there is an increase in the proportion of end groups and (2) it has a lower glass transition temperature T_g than the resin. A detailed mathematical treatment of this [7] can be carried out to explain the success of some plasticizers and the failure of others. Clearly, the use of a given plasticizer in a certain application will be a compromise between the above ideas and physical properties such as volatility, compatibility, high- and low-temperature performance, and viscosity. This choice will be application dependent, i.e., there is no ideal plasticizer for every application.

3.4. Solvation – Desolvation Equilibrium

Since plasticizer migrates from plasticized polymers, it is clear that plasticizer molecules are not bound permanently to the polymer as in an internally plasticized resin, but that an exchange or equilibrium mechanism is present. This implies that there is no stoichiometric relationship between polymer and plasticizer, although some quasistoichiometric relationships appear to exist [8], [9]. This idea is extended in Section 3.6.

3.5. Generalized Structure Theories

In their simplest form generalized structure theories attempt to produce a visual representation of the mechanism of plasticizer action [1]. The theories are based on the concept that if a small amount of plasticizer is incorporated into the polymer mass it imparts slightly more free volume and gives more opportunity for the movement of macromolecules. Many resins tend to become more ordered and compact as existing crystallites grow or new crystallites form at the expense of the more fluid amorphous material. For small additions of plasticizer, the plasticizer molecules may be totally immobilised by attachment to the resin by various forces. These tend to restrict the freedom of motion of small portions of the polymer molecule necessary for the absorption of mechanical energy. Therefore it results in a more rigid resin with a higher tensile strength and modulus than the base polymer itself. This phenomenon is therefore termed antiplasticization. A study by Horsley [10] used X-ray diffraction to show that small amounts of dioctyl phthalate (DOP) progressively increase the order in the PVC. Above these concentrations the order decreases and the polymer becomes plasticized (see also Section 3.6).

3.6. Specific Interactions and Interaction Parameters

Early attempts to describe PVC – plasticizer compatibility were based on the same principles as used to describe solvation, i.e., "like dissolves like" [11]. To obtain a quantitative measure of PVC – plasticizer compatibility a number of different parameters have been used. These are briefly described below.

The Hildebrand solubility parameter δ gives a measure of compatibility between plasticizer and a resin based on the solvating power of the plasticizer for the resin under investigation. For complex molecules this can be estimated [12] by using a set of additive constants for the more common groups in plasticizer molecules. These constants are designated F and the overall value for the plasticizer is given by the equation

$$\delta = \sum F/V \tag{2}$$

i.e., a summation of each solubility constant for each molecular portion of the plasicizer molecule divided by the molecular volume. Data so calculated for different plasticizers can be compared, with the higher values suggesting greater compatibility.

The polarity parameters Φ were evaluated by VAN VEERSEN and MEULENBERG [13] and despite their apparent simplicity they show a good correlation with plasticizer activity for nonpolymeric plasticizers. The parameter is defined as:

$$\Phi = \frac{M\,(A_p/P_o)}{1000} \tag{3}$$

where

M = molecular mass of plasticizer
A_p = number of carbon atoms in the plasticizer, excluding aromatic and carboxylic acid carbon atoms
P_o = number of polar (e.g., carbonyl) groups present

The factor of 1000 is used to produce values of a convenient size.

The Solid-Gel Transition Temperature T_m is a measure of plasticizer activity. It is the temperature at which a single grain of PVC dissolves in excess plasticizer. More efficient plasticizers have lower values of T_m due to their higher solvating power.

The Flory – Huggins Interaction Parameter χ. These ideas, based on a study of polymer miscibility, were applied to plasticizers by ANAGNOSTOPOULOS [14] according to the equation:

$$1/T_m = 0.002226 + 0.1351\,(1-\chi)\,/V_1 \tag{4}$$

where V_1 is the molar volume of the plasticizer, calculated from the molecular mass and the density at T_m.

The activity parameter α is based on work by BIGG [15]. It is another measure of plasticizer activity and is defined as

$$\alpha = 1000\,\frac{(1-\chi)}{M} \tag{5}$$

Summaries. Over the past few years these methods have been assessed and extended by many workers, in particular the Loughborough group [16], [17]. It was shown that solubility parameters are capable of classifying plasticizers of a given family in terms of their compatibility with PVC but that they are of limited use for comparing plasticizers of different families (e.g., phthalates with adipates). Polarity parameters provided useful predictions of the activity of monomeric plasticizers but again were not able to compare activity of plasticizers from different families. In all cases it was not possible to predict the behavior of polymeric plasticizers.

Specific Interactions. Ideas on the subject of specific interactions between PVC and a plasticizer molecule, as a basis of plasticization, can be considered a more detailed form of some of the ideas discussed above. Clearly some mechanism of attraction and interaction between PVC and plasticizer must exist for plasticizer to be retained in the polymer after processing.

The role of specific interactions in the plasticization of PVC has been proposed from work on specific interactions of esters in solvents (e.g., hydrogenated chlorocarbons) [18], work on blends of polyesters with PVC [19–25], and work on plasticized PVC itself [26–28]. Modes of interaction between the carbonyl functionality of the plasticizer ester or polyester were proposed, mostly on the basis of results from Fourier transform IR spectroscopy (FTIR). Shifts in the absorption frequency of the carbonyl group of the plasticizer ester to lower wavenumber, indicative of a reduction in polarity (i.e., some interaction between this functionality and the polymer) were reported [25–27]. Work performed with dibutyl phthalate [27] suggested an optimum concentration at which such interactions were maximized. Spectral shifts were in the range 3 – 8 cm^{-1}. Similar shifts were also reported in blends of PVC with polyesters [19–25], again showing a concentration dependence of the shift to lower wavenumber of the ester carbonyl absorption frequency.

Recent Studies. Some recent work has extended these ideas by using new analytical techniques, in particular molecular modeling and solid-state NMR spectroscopy.

Molecular Modeling. The computer modeling of molecules is a rapidly growing branch of chemistry (→ Molecular Modeling) [29–31]. High-resolution graphics and fast computers allow the operator to build molecules in minimum energy configurations and view them in real time. This model can be constructed from crystallographic coordinates available from data

bases or by simple intervention from the operator. Molecular mechanics or quantum mechanics programs are then used to arrive at a likely structure.

A range of plasticizer molecule models and a model for PVC were generated and energy minimized to observe their most stable conformations. Such models highlighted the free volume increase caused by the mobility of the plasticizer alkyl chains. More detailed models were also produced to concentrate on the polar region of the plasticizer and its possible mode of interaction with the polymer. These showed the expected repulsion between areas on the polymer and plasticizer of like charge as well as attraction between the negatively charged portions of the plasticizer and positively charged portions of the PVC.

Solid-State NMR Spectroscopy. Recent advances in technology have made the study of solids by NMR techniques of considerably greater ease than in previous years. For the accumulation of solid state ^{13}C NMR spectra, cross polarisation magic angle spinning (CP-MAS) can be utilized to significantly reduce signal broadening effects present in the solid state but not in the liquid state. The technique was used to study the molecular effects of plasticization by comparing spectral shifts of PVC and plasticizer under various degrees of processing. For PVC plasticized with DIDP two processing temperatures, 130 °C and 170 °C, were used, representing a low degree and high degree of plasticization, respectively. The comparison of the spectra showed no shift in the resonance frequency of the carbonyl group with processing temperature. The most significant difference in the two spectra was in the aliphatic carbon resonances. The spectra of the more plasticized sample showed resonance shifts and increased resolution for these carbon atoms. This again shows a strong dependence of successful plasticization on the conformation of the alkyl chains of the plasticizer ester (linked to the increased free volume).

4. Plasticized PVC

4.1. Applications Technology

Over 90 % of plasticizer sales by volume are into the PVC industry. The reason is that the benefits imparted by the plasticization of PVC are far greater than those imparted to other polymers. PVC stands alone among polymers in its ability both to accept and retain large concentrations of plasticizer. This is due to a morphological form consisting of highly amorphous, semicrystalline, and highly crystalline regions. Without the wide range of additives available (e.g., plasticizers, stabilizers, fillers, lubricants, pigments) PVC would be of little use. The development of PVC as a commodity polymer is fundamentally linked to the development of its additives.

The two principle types of PVC are suspension and emulsion PVC. The plasticizer applications technology associated with these two forms are different and are discussed separately here. Details of the polymerization techniques giving rise to these two distinct polymer types are discussed in [2], [32]; see also → Poly(Vinyl Chloride).

4.1.1. Suspension PVC

PVCs produced by suspension polymerization have a relatively large particle size (typically 100 – 150 µm). Additionally, these particles are highly porous and are therefore able to absorb large amounts of liquid plasticizer during a formulation mixing cycle. Flexible PVC formulations based on a suspension polymer are typically processed by a dry-blend process, during which all formulation ingredients are heated (typically 70 – 110 °C) and intimately mixed to form a dry powder (PVC dry-blend) which contains all formulation ingredients. This dry-blend can be either stored or processed immediately. Processing of suspension resin formulations is performed by a variety of techniques such as extrusion, injection molding, and calendering to totally fuse the formulation ingredients and form the desired product.

4.1.2. Emulsion PVC

Emulsion PVC resins differ from those manufactured by the suspension process principally in terms of particle size but also in the presence of emulsifiers and surfactants not used in suspension polymerization. The particle size of an

emulsion polymer (typically 1 – 2 µm) is much lower than that of a suspension resin. This low particle size imparts a lack of porosity to the resin and thus the mixing of formulation ingredients in a dry-blending process is not possible. A typical flexible PVC formulation prepared with an emulsion polymer will be a liquid or paste (plastisol). This consequently leads to a different applications technology area. PVC plastisols can be fused by techniques such as spreading, rotational molding, and slush molding. Plastisols may also be semi-gelled (i.e., enough heat is imparted to convert the plastisol into a solid but without the full development of tensile properties brought about by complete fusion) for storage.

As the formulation ingredients in a plastisol are in the liquid form, the viscosity of the plastisol is of major importance, and the intrinsic viscosity of the plasticizer has a major contribution to the plastisol viscosity, as do the polymerization conditions of the resin. The desired plastisol viscosity can be obtained by careful selection of polymer, plasticizer, and other formulation ingredients but also by shear rate. PVC plastisols are either (1) pseudoplastic or shear-thinning (viscosity decreases with shear), (2) near-Newtonian (viscosity remains nearly constant with shear), or (3) dilatant (viscosity increases with shear). Choice of plastisol ingredients to give desired viscosity effects will depend on the application requirements of the plastisol.

4.2. Effect of Plasticizer Choice on the Properties of Flexible PVC

This section discusses the effect of plasticizer choice on a range of properties of flexible PVC articles. Certain properties are more important for some applications than others, and hence some plasticizers find more extensive use in some application areas than others. The PVC technologist must ascertain the most important properties for the application in question and then make the correct choice of plasticizer.

4.2.1. Plasticizer Efficiency

Plasticizer efficiency is a measure of the concentration of plasticizer required to impart a specified softness to PVC. The softness may be measured as a British Standard Softness (BSS) or a Shore hardness (BSS 35 is equivalent to a Shore A hardness of 80 – test method BS 2782). For a given acid constituent of plasticizer ester, plasticizer efficiency decreases as the carbon number of the alcohol chain increases; e.g., for phthalate esters efficiency decreases in the order DBP>DIHP>DOP>DINP > DIDP>DTDP. For example, an additional six parts per hundred of PVC (phr) of DIDP than DOP is required to give a hardness of Shore 80 when all other formulation ingredients remain constant. In addition to the size of the alcohol chain, the amount of branching is also significant, with the more linear isomers being of greater efficiency. Choice of the acid constituent can also be significant. For the same alcohol constituent, phthalate and adipate esters are approximately equivalent but both are considerably more efficient than the corresponding trimellitate.

Figure 2 compares various plasticizers with respect to their ability to impart flexibility to a PVC resin. This shows the lower plasticizing efficiency of a C_{10} phthalate relative to a C_8 phthalate and also the lower plasticizing efficiency of the trimellitate esters compared to the phthalates.

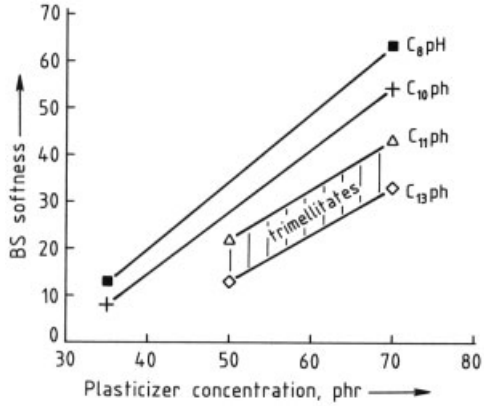

Figure 2. Relative efficiency of plasticizers
ph = phthalate

4.2.2. High-Temperature Performance

High-temperature performance in flexible PVC is related to (1) plasticizer volatilization and (2)

plasticizer degradation. Plasticizer volatilization, both from the finished article during use at elevated temperatures (e.g., electrical cable insulation) and also during processing (i.e., release of plasticizer fume) is directly related to the volatility of the plasticizer in use. Hence the higher molecular mass plasticizers give superior performance in this area. Figure 3 shows the percentage mass loss from plasticized PVC (50 phr of plasticizer) for various plasticizers at 100 °C. The superior performance of the trimellitate esters and inferior performance of adipate esters, relative to the phthalate, is illustrated.

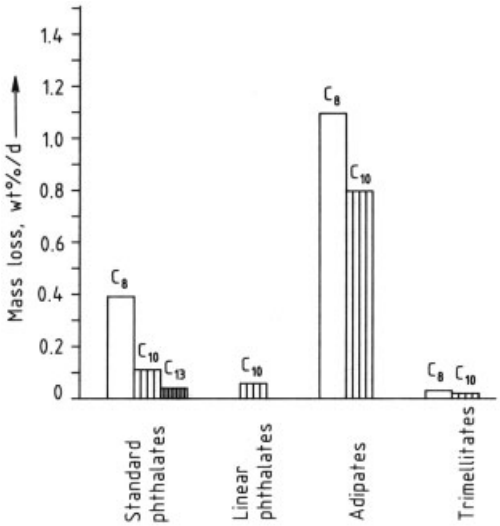

Figure 3. Volatile loss in use from plasticized PVC (50 phr plasticizer) for various plasticizers

For phthalate plasticizers thermal stability decreases in the order DTDP > DIDP > DINP > DOP > DIHP > DBP. Higher molecular mass esters such as trimellitates improve this still further and are used in demanding cable specifications which have strict mass loss requirements. Polyester plasticizers give the best performance in this area, with performance improving with increasing molecular mass. Additionally, branched esters have somewhat higher volatilities than their linear equivalents.

High-temperature performance also includes the generation of fume in the workplace, for which the same structure relationships apply. Not only does excessive plasticizer volatilization have environmental consequences but inaccuracies in formulation can be incurred since not all the plasticizer enters the PVC resin, resulting in a harder material than calculated.

Plasticizer molecules can undergo thermal degradation at high temperature. Esters based on the more branched alcohol isomers are more susceptible to such degradation. This can, however, be offset by the incorporation of an antioxidant, and plasticizer esters for cable applications frequently contain a small amount of antioxidant.

4.2.3. Low-Temperature Performance

The ability of plasticized PVC to remain flexible at low temperature is of great importance in certain applications (e.g., external tarpaulins, underground cables). For this property the choice of the acid constituent of the plasticizer ester is of major importance, with the linear aliphatic adipic, sebacic, and azeleic acids giving excellent low-temperature flexibility compared to the corresponding phthalates and trimellitates.

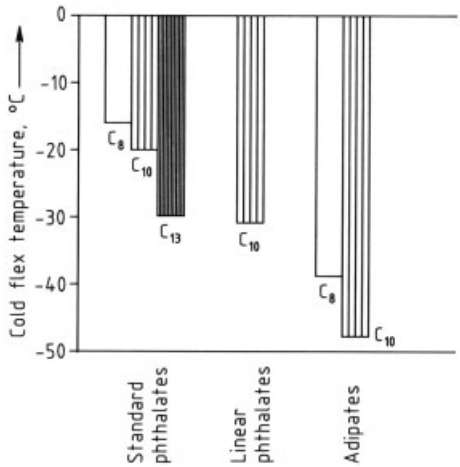

Figure 4. Cold flex temperature for various PVC – plasticizer combinations (50 phr plasticizer)

There is also a major contribution to low-temperature performance from the alcohol portion of the ester; the greater the linearity of the plasticizer the greater the low-temperature flexibility. Figure 4 shows the cold flex temperature (the lowest temperature at which a sample of flexible PVC of standard dimensions can be twisted by 200° in the Clash and Berg test) for various PVC – plasticizer combinations. The increasing low-temperature performance with increasing

alkyl chain length is illustrated, as is the superior low-temperature performance of the adipate esters.

4.2.4. Gelation Properties

The gelation characteristics of a plasticizer are related to its efficiency, and both properties are often discussed together. The gelation characteristics are a measure of the ability of a plasticizer to fuse with the polymer to give a product of maximum elongation and softness (i.e., maximum plasticization). Gelation properties are often measured either (1) as a processing temperature – the temperature to which the plasticizer and polymer must be heated in order to obtain these properties or (2) as a solution temperature, the temperature at which a grain of polymer dissolves in excess plasticizer, giving a measure of the solvating power of the plasticizer. Ease of gelation is related to plasticizer polarity and molecular size. The greater the polarity of a plasticizer molecule the greater the attraction it has for the PVC polymer chain and the less additional energy, in the form of heat, is required to cause maximum plasticizer – PVC interaction. The most active plasticizers are able to bring about these effects soon after the T_g of the polymer (70 – 80 °C) is reached, whereas the less active plasticizers require temperatures of ca. 180 °C for maximum elongation properties to be obtained. The polarity of the plasticizer is determined by both acid type and alcohol chain length. Aromatic acids, being of greater polarity, tend to show greater ease of gelation than aliphatic acid esters.

Molecular size also has a key contribution and explains why molecules of similar polarity can show different gelation properties. The smaller the plasticizer molecule the easier it is for it to enter the PVC matrix; larger molecules require more thermal energy to establish the desired interaction with the polymer. Since branching influences molecular size, this too has a contribution to gelation properties, with the more branched isomers showing greater activity. Thus for the phthalate esters ease of gelation decreases in the order BBP > DBP > DIHP > DOP > DINP > DIDP > DTDP. Figure 5 shows the greater ability of the more molecularly compact plasticizers to impart full gelation properties (i.e., maximum tensile strength).

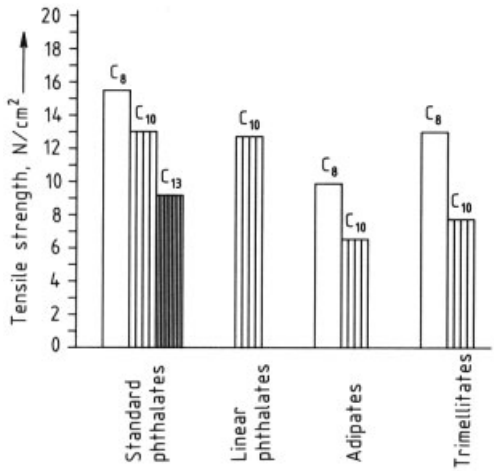

Figure 5. Tensile strength of various PVC – plasticizer combinations at 160 °C (50 phr plasticizer)

4.2.5. Migration and Extraction

When plasticized PVC comes into contact with other materials, plasticizer may migrate from the plasticized PVC into the other material. The rate of migration depends not only on the plasticizer employed but also on the nature of the contact material.

Plasticizer can also be extracted from PVC by a range of solvents, including water. How aggressive a particular solvent is depends on its molecular size and its compatibility with both the plasticizer and PVC. Water extracts plasticizer very slowly, oils are slightly more aggressive, and low molecular mass solvents are the most aggressive.

The key characteristics for migration and extraction resistance is molecular size. The larger the plasticizer the less it tends to migrate or be extracted. The extreme case is seen by the use of polymeric plasticizers in applications where excellent migration and extraction resistance is required. There is also a contribution from the linearity of the alcohol component of the plasticizer ester. The greater the linearity of the ester the greater its migration and extraction rate in comparison to the more branched isomers. Figure 6 shows the results of a standard test for the migration of plasticizer into rigid (unplasticized

PVC). Equivalent masses of plaques containing different plasticizers were sandwiched between two rigid PVC discs and the plasticizer migration into the rigid PVC after the test was measured. The improved performance of the longer-chain esters and the trimellitate esters is illustrated.

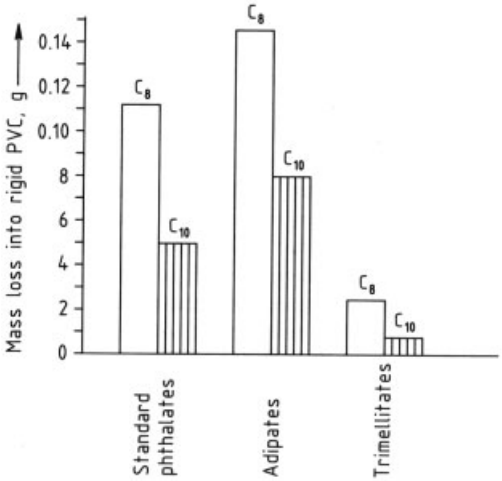

Figure 6. Mass loss into rigid (unplasticized) PVC (11d at 70 °C)

4.2.6. Plastisol Viscosity and Viscosity Ageing

Plastisol viscosity and viscosity ageing are of great importance in emulsion PVC applications. Plastisol viscosity, for a given emulsion PVC resin, has a strong dependence on plasticizer viscosity, with the lower molecular mass and more linear esters showing the lowest viscosity and hence the lowest plastisol viscosity (i.e., plastisol viscosity, for a common set of other formulation ingredients, increases in the sequence DBP < DIHP < DOP < DINP < DIDP). In spite of these viscosity differences, however, if plastisols are formulated to equal softness, i.e., taking into account the efficiency of the plasticizers involved, more of the less efficient plasticizer must be employed to impart the same softness to the product being manufactured. The addition of this extra liquid to the plastisol may produce an equivalent viscosity to that of the plastisol with the less viscous plasticizer. Esters based on aliphatic acids, being of lower viscosity than the corresponding aromatic acids, show lower plastisol viscosities, and adipate esters have found widespread use in plastisol applications although to meet other requirements (e.g., volatility, gelation characteristics) they are often employed in a blend with other esters.

Plastisols are often mixed and then stored rather than processed immediately. It is of great importance in this case for the plasticizer to show little or no paste thickening action at the storage temperature, and clearly in this instance it is not advisable to use a plasticizer of too great an activity, since grain swelling, leading to plastisol viscosity increase, can occur at low temperatures for some active plasticizer systems. Figure 7 compares the ageing of plastisols based on various plasticizers. The molecularly compact plasticizers such as BBP and DIBP can enter the PVC matrix easily and cause a rapid increase in viscosity with time. The ageing characteristics improve as the polarity of the plasticizer is reduced. Greater linearity of the plasticizer (e.g., 911 phthalates) also improve this further, with adipates also showing very good viscosity stability.

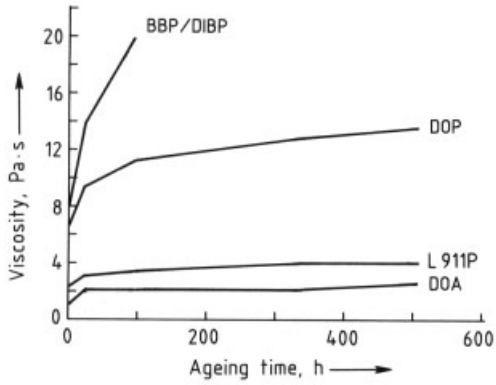

Figure 7. Viscosity ageing of PVC plastisols at 23 °C

4.2.7. Windscreen Fogging

The phenomenon of car windscreen fogging has been known for some time. Fogging is the condensation of volatile material on the car windscreen causing a decrease in visibility to the driver. Although this volatile material may arise from a variety of sources (e.g., exhaust fume sucked in through the ventilation system), material from inside the car (e.g., crash pads, rear

shelves) may also contribute to windscreen fogging on account of the high temperatures which can arise in a car standing in sunlight. In the case of flexible PVC such a contribution may arise from emulsifiers in the polymer, stabilizers, and plasticizers. In each case manufacturers have studied their products in detail and now recommend low-fogging polymers, stabilizers, and plasticizers. Tests have been designed (e.g., DIN 75 201) to assess the fogging performance of both the PVC sheet and the raw materials used in its production. These tests involve heating a sheet of raw material for a specified period at a set temperature in an enclosed apparatus with a cooled glass plate above the sheet or raw material. The reflectance of the glass plate before and after the test is then compared to ascertain the degree of fogging. Figure 8 shows the relationship between plasticizer molecular mass and linearity with fogging performance. The test was carried out at 90 °C for a 6 h duration (a variety of tests, each with their own specified temperature and time, are in use). In such a test, the fogging performance of a plasticizer is related to its (1) volatility, (2) refractive index, and (3) surface tension. The precise nature of (2) and (3) is somewhat complex [33] and attempts to improve the test are currently in progress. In the case of plasticizer volatility it is clear that a higher concentration of plasticizer leads to increased fogging in the test. The higher molecular mass and more linear plasticizers give superior performance, and trimellitate esters (in suspension PVC applications) and L911 phthalates, with their high degree of linearity, and consequently low viscosity for plastisol applications, have achieved widespread use as low-fogging plasticizers for these applications.

4.2.8. Summmary

Table 2 gives a summary of the properties described above. For each criterion the effect of (1) increasing plasticizer concentration, (2) increasing the size of the plasticizer molecule, (3) increasing the linearity of the plasticizer molecule, and (4) changing the acid constituent of the ester is discussed. A tick indicates improved performance, a cross indicates poorer performance, and a dash indicate no effect; ticks and crosses in parentheses indicate that any changes tend to be marginal.

It can be seen from this section that there is clearly no perfect plasticizer for every application: choice will depend on the performance requirements of the article being manufactured and additionally price.

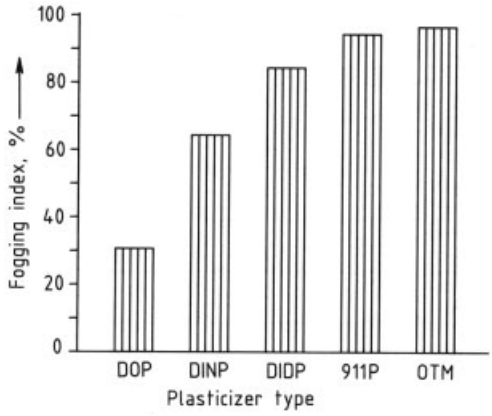

Figure 8. Fogging index of various plasticizers (6 h at 90 °C)

5. Plasticization of Polymers Other Than PVC

5.1. Requirements for Plasticization

The plasticization of PVC accounts for the vast majority of plasticizer sales. However, significant amounts of plasticizers are used in non-PVC polymers and this may become increasingly important in the future. Although PVC is unique in its ability to accept and retain large quantities of plasticizer, effective plasticization of other resins with slightly modified plasticizers may be possible if certain conditions, specific to the polymer, are met.

The factors which must be considered for the plasticization of a polymer are

1) The need – even though some polymers may be compatible with large concentrations of plasticizer the resultant softening may be of little use
2) Short- and long-term compatibility – the ability of a polymer to accept and retain the plasticizer

Table 2. Summary of the effect of plasticizer choice on the properties of flexible PVC

Property (at equal softness)	Increased concentration	Increased size	Increased linearity	Acid used*
Efficiency (higher)	—	×	✓	ph=ad>tr
High-temperature properties	×	✓	✓	tr>ph>ad
Low-temperature properties	✓	✓	✓	ad>ph=tr
Gelation temperature	(✓)	×	(×)	ph>ad≥tr
Migration/ extraction	×	✓	×	tr>ph>ad
Plastisol viscosity	✓	(✓)	✓	ad>ph
Plastisol ageing	—	✓	(✓)	ad>ph
Fogging	×	✓	✓	tr>ph>ad
Price	✓	✓	×	ph>ad>tr

*ph=phthalate; ad=adipate; tr=trimellitate.

For a plasticizer to enter a polymer structure the polymer should be highly amorphous (e.g., crystalline nylon will retain only a small quantity of plasticizer if it retains its crystallinity). Once it has penetrated the polymer the plasticizer fills free volume and provides polymer chain lubrication, increasing rotation and movement.

The plasticizer content of a polymer may be increased by suppressing crystallization in the polymer, but if crystallization subsequently occurs the plasticizer is exuded. For highly crystalline resins, the small amounts of plasticizer allowable can change the nature of the small amorphous regions with a consequent change in properties.

Plasticizer compatibility has been studied in many resins and a summary of these studies is given in [4].

5.2. Plasticization of Other Polymers

Acrylic Resins. Although considerable information of the plasticization of acrylic resins is available, the subject is complicated by the fact that acrylic resins constitute a large family of polymers rather than a single polymeric species. An infinite variation in physical properties may be obtained through copolymerization of two or more acrylic monomers selected from the available esters of acrylic and methacrylic acid [34].

Plasticizers are used in the acrylics industry to produce tough, flexible coatings. Compatabilities with common plasticizers are up to 10 wt %, although in some cases (for low molecular mass plasticizers) it can be higher (e.g., a formulation of 100 phr PMMA, 150 phr DBP, 225 phr chalk, 25 phr resorcinol has been used [4]). PMMA is used in small amounts with PVC. Cast acrylics, however, require a high glass transition temperature and high rigidity, hence no plasticizer is required.

Listed plasticizers include all common phthalates and adipates.

There has been interest in the development of acrylic plastisols similar to those encountered with PVC. Clearly the same aspects of both plastisol viscosity and viscosity stability are important. Patents appear in the literature [35] indicating that the number of available plasticizers which show both good compatibility with arcrylic resins and satisfactory long term plastisol stability may be fewer than those showing equivalent properties with emulsion PVC resins.

Polyamides. The high degree of crystallinity of polyamides means that plasticization can oc-

cur only at very low levels. Plasticizers are used in nylon but are usually sulfonamide based since these are generally more compatible than phthalates (DEHP is 25 phr compatible, other phthalates less so; sulfonamides are compatible up to 50 phr).

Poly(ethylene terephthalate) (PETP) is crystalline and hence difficult to plasticize. Additionally, since PETP is used as high-strength film and textile fiber, plasticization is not usually required, although esters showing plasticizing properties with PVC may be used in small amounts as processing aids and external lubricants. Plasticizers have also been used to aid the injection molding of PETP, but only at low concentrations.

The main interest in plasticizers for PETP is in dyeing. Due to its lack of hydrogen bonds PETP is relatively difficult to dye. Plasticizers can increase the speed and intensity of dyeing. The compounds used, however, tend to be of low molecular mass since high volatility is required to enable rapid removal of plasticizer from the product.

Polyolefins. Interest has been shown in the plasticization of polyolefins [4], but their use generally results in a reduction of physical properties [16], and compatibility may be only 2 wt %. Most polyolefins give adequate physical properties without plasticization.

Plasticizers have been used with polypropylene to improve its elongation at break [15], although the addition of plasticizer can lower the glass transition temperature, room-temperature strength, and flow temperature. This can be overcome by simultaneous plasticization (ca. 15 wt %) and cross-linking. Plasticizers used include DOA.

Polystyrene is compatible with common plasticizers, but the modification of properties produced is of little value. Small amounts of plasticizer (e.g., DBP) are used as a processing aid.

Fluoroplastics. Conventional plasticizers are used as processing aids for fluoroplastics up to a concentration of 25 %. However, certain grades of Kel-F (polychlorotrifluoroethylene) contain up to 25 wt % plasticizer to improve elongation and increase softness. The plasticizers used are usually low molecular mass oily chloroethylene polymers [4].

Rubbers. Plasticizers have been used in rubber processing and formulations for many years [36], although phthalic and adipic esters have found little use since cheaper alternatives, e.g., heavy petroleum oils, coal tars, and other predominantly hydrocarbon products, are available for many types of rubber. Esters (e.g., DOA, DOP, DOS) can be used with latex rubber to produce large reductions in glass transition temperature. More polar elastomers such as nitrile rubber and chloroprene rubber are insufficiently compatible with hydrocarbons and require a more specialized type of plasticizer, e.g., a phthalate or adipate ester. Approximately 50 % of nitrile rubber used in Western Europe is plasticized at 10 – 15 phr (a total of $5-6\times10^3$ t/a), and 25 % of chloroprene at ca. 10 phr (ca. 2×10^3 t/a). Usage in other elastomers is very low but may increase due to toxicological concerns over polynuclear aromatic compounds.

Work on the use of high molecular mass esters in nitrile rubber has compared DINP with DBP and showed that at the 10 phr the nitrile rubber was effectively plasticized with DINP [37].

6. Toxicology and Environmental Aspects

6.1. Toxicity Studies on Plasticizer Esters

Acute Toxicity. The acute toxicity (LD_{50}) of plasticizers is extremely low (>2000 mg/kg). For example, DOP has an oral LD_{50} of >30 000 mg/kg.

In addition to their low acute toxicity many years of practical use coupled with animal tests show that plasticizers do not irritate the skin or mucous membranes.

Numerous long-term feeding trials on plasticizers have been carried out. The majority of these studies have been on di-2-ethylhexyl phthalate (DEHP, commonly called DOP) because it accounts for around 50 % of the plasticizer usage in Europe and has been considered

as a model for the other phthalates. See also → Phthalic Acid and Derivatives, Chap. 6.5.).

Prior to 1980 such studies showed no adverse effects. However, in 1980 the results of a two-year feeding study carried out as part of the National Toxicology Program/National Cancer Institute (NTP/NCI) Bioassay Program in the United States indicated that DOP causes increased incidence of liver tumors in rats and mice and that di-2-ethylhexyl adipate (DEHA, commonly called DOA) had a similar effect in mice but not rats [38], [39]. In these studies the levels of plasticizers fed were very high, this being only possible because of their low acute toxicity. Their diets contained up to 12 000 ppm of DOP and 25 000 ppm of DOA (this is comparable with a human consuming 0.5 L of DOA every day).

A large number of more recent investigations [40–43] carried out in both Europe and the United States on a variety of plasticizers and different types of animals clearly show:

1) Plasticizers are not genotoxic.
2) Oral administration of plasticizers, fats, and other chemicals (such as hypolipidaemic drugs) to rodents causes a large increase of microbodies in the liver called peroxisomes. This peroxisome proliferation is considered by some authors to be linked to the formation of liver tumors.
3) The administration of plasticizers, fats, or hypolipidaemic drugs to nonrodent species such as the marmoset (a primate considered to be metabolically closer to humans) does not lead to peroxisome proliferation and liver damage. Some commonly used hypolipidaemic drugs (e.g., clofibrate), which cause peroxisome proliferation in rodents, have been used by humans for many years with no ill effects.
4) These species differences have also been observed in in vitro studies on the liver cells of rats, mice, guinea pigs, marmosets, and humans. Peroxisome proliferation was observed in the rat and mouse cells but not in those of humans, marmosets, or guinea pigs.
5) On the basis of these differences in species response (especially if the extremely high exposure in the animal tests is also taken into consideration) it is concluded that plasticizers pose no significant hazard to humans.

Human Exposure to Plasticizers. Experimental investigations and assessments in the United States and Europe have shown that the average human intake of plasticizers amounts to 2 g per person per year. This is mainly due to traces of DOA migrating from food packaging. The level of exposure is considerably below the dose at which toxic effects are to be expected. The no observed effect level (NOEL) for DOA and DOP is ca. 40 mg per kilogram body weight per day, or higher depending on the kind of effect, i.e., at least 1000 g per year for an average adult. There is thus a 500 fold safety factor between the estimated plasticizer intake of 2 g per person per year and the NOEL. Taking into consideration the difference in response between rodents and primates, the safety factor is in fact even greater.

The Opinion of Official Experts and Legislative Authorities. The IARC of the WHO has classified DOP as "an agent possibly carcinogenic to humans," largely because of the NTP/NCI rodent studies. However, IARC has not classified DOP as a carcinogen for regulatory purposes [44].

The EC Working Group "Classification and Labeling of Dangerous Substances" has examined all the toxicological data and has concluded that DOP should not be labeled as a human carcinogen [45].

In Germany the Advisory Body for Environmentally Relevant Existing Substances (BUA) has summed up the situation in their report No. 4 (1986) as follows [46]:

"Chronic damages have been noticed in rodents exposed to high DOP concentrations (rat: over 40 – 70 mg per kg bodyweight per day). The rat because of a characteristic feature in its DOP metabolism – different from the human metabolism – is especially sensitive to this substance. Thus, data obtained via tests with these animal species have no relevance for man. Tests with primates – having a metabolism comparable to that of humans – showed that DOP was rather ineffective. According to these findings there is no basis to suspect chronic damages due to DOP exposure at environmentally relevant concentrations."

Similarly the German Health Authority (BGA) has concluded that the carcinogenic ef-

fect of DOP in rats and mice is specific to this species and is not relevant to humans.

In the United States the FDA has not introduced any new regulations following the results of the NTP/NCI studies.

6.2. Plasticizers in the Environment

The release of plasticizers to the environment may occur during their production and distribution, during incorporation into PVC, and by loss from the finished article during use or after disposal.

The controlled nature of modern manufacturing processes makes it unlikely that significant loss of plasticizer to the environment occurs during its production (measurements of total emission from a modern plasticizer plant have given values of less than 0.01 %).

The extremely low solubility of plasticizers in water means that little is lost by leaching from flexible PVC either during use or after disposal. The increasing use of modern incineration techniques to dispose of domestic waste results in complete combustion of plasticizers to carbon dioxide and water.

The main way in which plasticizers enter the environment is therefore by evaporation during processing with PVC. The extent of such losses depends on the process and plasticizer used and has been estimated to vary from 0.02 % for injection molding up to 1 % for coating processes. However, these levels are being continually reduced by the installation of incineration, scrubbing, and filtration systems in processing plants.

Plasticizers can be detected in the environment in the ultratrace (parts per billion) range.

There are no indications of an accumulation of plasticizers in water, soil, or air because they are biologically and photochemically degraded. Degradation is particularly rapid under aerobic conditions to produce carbon dioxide and water as for other organic substances.

The occurrence and effects of phthalates in the environment has been reviewed in detail by ECETOC (European Chemical Industry Ecology and Toxicology Centre) [47]. It is clear from this report that in laboratory tests microorganisms and fish can take up dialkyl phthalates; however, they are rapidly excreted and no toxic effects have been seen. The report concludes that at the levels corresponding to current use and disposal practices, phthalates do not represent a hazard to the environment.

Health Aspects of Flexible PVC. Small quantities of plasticizers may be lost from PVC by evaporation and extraction or by migration into other materials.

Plasticizer Evaporation From Flexible PVC. The use of plasticizers in the production of vinyl wall and floor covering can lead to plasticizer vapor being present in room air. However, the plasticizers used have extremely low vapor pressures and thus at 25 °C the maximum concentration of DOP is only ca. 0.01 mg/m^3. The concentration of the higher molecular mass plasticizers are even lower. Also, given the comparatively slow plasticizer release rate and normal ventilation, then the plasticizer concentration in reality is extremely low. The high concentrations sometimes reported must therefore be attributed to analytical errors. This is demonstrated by the fact that recent measurements of the air in an emission chamber containing 1 m^2 of PVC flooring for 96 h showed no detectable levels of plasticizer; the detection limit was 4 parts per billion.

Flexible PVC Toys. In the United States the Consumer Products Safety Commission (CPSC) has expressed concern that children might ingest plasticizers by sucking teething rings and pacifiers. The estimated intake of plasticizers by infants in this way varies over several orders of magnitude up to a maximum of 0.07 mg kg^{-1} d^{-1}. On the basis of the toxicological evidence this does not represent a health risk. For other toys, such as balls and dolls, the intake of plasticizers is negligible. The use of plasticizers in PVC toys for infants is regulated in some countries.

Medical Uses of Plasticized PVC. Flexible PVC has been established for many years as one of the most important materials used in the manufacture of a wide variety of medical products. The main advantages of PVC are its low toxicity and its flexibility, clarity, and sealing properties. These permit the manufacture of lightweight, breakage-resistant medical devices that can be supplied ready for use in sterile packages.

It is known that low levels of DOP migrate from flexible PVC blood bags into blood. There is published evidence [48] showing that this has

a beneficial effect in that it significantly reduces the destruction of red cell walls compared to that which occurs on storage in other plastic or glass systems. DOP is the only plasticizer authorized by the European Pharmacopoeia for use in the manufacture of PVC blood bags.

Plasticized PVC in Food Packaging. Flexible PVC is used to produce food wrap film, cap seals for bottles and jars, and tubing for beverages. DOP, DINP, and DIDP are used in caps and seals, whereas DOA and polymeric adipates are used in the production of food wrap film. DOA is used in this application because it gives a film with a high degree of cling which remains flexible at low temperatures. It also provides the high permeability to oxygen and water vapor that is necessary to ensure good food storage.

The migration of plasticizer from these packaging materials into foodstuffs has been measured in the UK by the Ministry of Agriculture Fisheries and Food (MAFF) both for domestic and retail packaging of food. These measurements together with a knowledge of the typical diet has enabled MAFF to assess the maximum possible intake of plasticizers as 0.02 mg per person per day of phthalates and 8 mg per person per day of DOA [49]. These estimates are based on a worst case situation, which assumes that all solid foodstuffs are purchased wrapped in cling film and that 50 % of these are subsequently rewrapped in the home.

The UK Committee on Toxicity of Chemicals in Food, Consumer Products, and the Environment (COT) has examined these data together with the relevant toxicological information and has concluded [50]:

"Toxic effects have been demonstrated for some plasticizers, but the safety margins between toxic doses and human intakes are large and we consider it very unlikely that there are any adverse health effects from the current or past use of plasticizers in food packaging materials."

Recent CEFIC studies [51], [52] involving the administration to humans of deuterium-labeled DOA have established the major human urinary metabolites and their excretion rates. This, together with analysis of 24 h urine samples from a limited survey of the UK population, has shown the average DOA intake to be 3 mg per person per day. This is well below the tolerable daily intake of 18 mg per person proposed by the EC Scientific Committee for Food and hence again confirms the suitability of plasticized PVC film for all food contact applications.

Regulations and Provisions. In terms of handling and storage, plasticizers are in general classified as nonhazardous.

In different countries the recommended 8 h TWA limit for occupational exposure varies between 2 and 10 mg/m^3.

These recommended exposure limits have been reviewed by the Health and Safety Authorities in the United Kingdom and Germany, and in the light of recent toxicological findings they have confirmed that the current limits of 5 and 10 mg/m^3, respectively, are satisfactory.

The regulations controlling the use of plasticizers in food contact applications differ from country to country. However, a common feature of the regulations is that plasticizer that migrates into foodstuffs should not have any adverse effects on their taste or smell and in general must not exceed 60 ppm. Because of the large volume of toxicological data indicating that phthalates and adipates pose no health risk to humans, regulatory authorities in most countries allow their use in caps, seals, tubing, and wrapping films.

6.3. Storage and Handling

Plasticizer esters are relatively inert, thermally stable liquids with high flash points and low volatility. Consequently they can be stored safely in mild steel storage tanks or drums for extended periods of time. Exposure to high temperatures for extended periods, as encountered in drums in hot climates, is not recommended since it may lead to a deterioration in product quality with respect to color, odor, and electrical resistance.

7. References

1. J. K. Sears, J. R. Darby: *The Technology of Plasticizers,* Wiley and Sons, New York 1982.
2. D. L. Buszard in W. V. Titow (ed.): *PVC Technology,* 4th ed., Elsevier, Amsterdam 1984.
3. *Ullmann,* 4th ed., **24,** 349 – 380.
4. *Kirk-Othmer,* 3rd ed., **18,** 111.

5. M. Kampmann, *Kunststoffe* **79** (1989) no. 10, 897.
6. K. Weissermel, H.-J. Arpe, *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998, p. 428.
7. A. K. Doolittle, *J. Poly. Sci.* **2** (1947) no. 1, 121.
8. A. Hartmann, *Kolloid Z.* **142** (1955) 123.
9. R. S. Barshtein, G. A. Kotylarevski, *Sov. Plast. (Engl. Transl.)* **7** (1966) 18.
10. R. A. Horsley in P. Morgan (ed.): *Progress in Plastics,* Iliffe and Sons, New York 1957, pp. 77 – 88.
11. A. Kirkpatrick, *J. Appl. Phys.* **11** (1940) 255.
12. P. A. Small, *J. Appl. Chem.* **3** (1953) 76.
13. G. J. Van Veersen, A. J. Meulenberg, *SPE Tech. Pap.* **18** (1972) 314.
14. C. E. Anagnostopoulos, A. Y. Coran, W. R. Gamrath, *J. Appl. Polym. Sci.* **4** (1960) 181.
15. D. C. H. Bigg, *J. Appl. Polym. Sci.* **19** (1975) 3119.
16. S. V. Patel, M. Gilbert, *Plast. Rubber Processing Appl.* **6** (1986) 321.
17. L. Ramos de Valle, M. Gilbert, *Plast Rubber Processing Appl.* **13** (1990) 151.
18. A. Garton, P. Cousin, R. E. Prud'Homme, *J. Polym. Sci., Polym. Phys. Ed.* **21** (1983) 2275.
19. M. M. Coleman, J. Zarian, *J. Polym. Sci. Polym. Phys. Ed.* **12** (1979) 837.
20. M. M. Coleman, D. F. Varnell, *J. Polym. Sci. Polym. Phys. Ed.* **18** (1980) 1403.
21. J. J. Schmidt, J. A. Gardella, L. Salvati, *Macromolecules* **22** (1989) 4489.
22. M. Aubin, Y. Bedard, M. F. Morrissette, R. E. Prud'Homme, *J. Polym. Sci. Polym. Phys. Ed.* **21** (1983) 233.
23. M. B. Clark, C. A. Burkhardt, J. A. Gardella, *Macromolecules* **22** (1989) 4495.
24. D. J. Walsh, S. Rostami, *Adv. Polym. Sci. Rev.* **70** (1985) 119.
25. D. F. Varnell, M. M. Coleman, *Polymer* **22** (1981) 1324.
26. D. L. Tabb, J. L. Koenig, *Macromolecules* **8** (1975) no. 6, 929.
27. M. Theodorou, B. Jasse, *J. Polym. Sci. Polym. Phys. Ed.* **21** (1983) 2263.
28. E. Benedetti et al., *J. Polym. Sci. Polym. Phys. Ed.* **23** (1985) 1187.
29. *Chemistry in Britain,* Special Issue on Computational Chemistry, Nov. 1990.
30. *Chemistry and Industry,* Dec. 1990.
31. J. P. Sibilia: *A Guide to Materials Characterisation and Chemical Analysis.* VCH, Weinheim 1988.
32. M. J. Bunten, M. W. Newman, P. V. Smallwood, R. C. Stephenson: *Encyclopedia of Polymer Science and Engineering,* **vol. 17 and suppl.**, John Wiley and Sons, New York 1989, pp. 241 – 392.
33. Fogging: A Problem that won't Go Away, *Proceedings of PRI Miniconference,* Oct. 1991, Plastics and Rubber Institute, London.
34. General Motors, US 3 178 386 (R. J. Hickman).
35. Teroson, GB 1 516 510; US 4 210 567.
36. S. S. Kurtz, J. S. Sweely, W. J. Stout in P. F. Bruins (ed.): *Plasticizer Technology,* Reinhold Publishing, New York 1965.
37. N. J. Clatden, C. J. Howick, unpublished work.
38. National Toxicology Programme, publication no. NIH-1768 (1980).
39. National Toxicology Programme, publication no. 82-1773 (1982).
40. P. Reddy et al., *CRC Crit. Rev. Toxicol.* **13** (1983).
41. P. Reddy et al., *Nature* **283** (1983) 397.
42. P. Schmezer et al., *Carcinogenesis* **9** (1988) 37.
43. *Di-2-Ethylhexyl Phthalate: A Critical Review of the Available Toxicological Information,* CEFIC, Brussels 1985.
44. *IARC Monogr.* **29** (1982) 281.
45. *Official Journal of the European Communities* **L 222** (17/8/1990) 49.
46. *Di-2-Ethylhexyl Phthalate,* BUA substance report no. 4, 1986.
47. *An Assessment of the Occurrence and Effects of Dialkyl Ortho-Phthalates in the Environment,* ECETOC technical report no. 19, Brussels 1985.
48. T. N. Estep et al., *Blood* **64** (1964) 1270.
49. Ministry of Agriculture, Fisheries and Food (UK), "Plasticisers: Continuing Surveillance," *Food Surveillance Paper* no. 30, HMSO, London 1990.
50. Ministry of Agriculture, Fisheries and Food (UK), *Survey of Plasticizer Levels in Food Contact Materials in Food,* Surveillance Paper no. 21, HMSO, London 1987.
51. N. J. Loftus, W. J. Laird, J. E. Leeser, G. T. Stool, B. H. Woollen, *Hum. Exp. Toxicol.* **9** (1990) 326.
52. N. J. Loftus, W. J. Laird, M. F. Wilks, G. T. Steel, B. H. Woollen, British Toxicology Society Meeting, Edinburgh, September 1991.

Plastics, Additives

Plasticizers and Release Agents are separate keywords

RAINER WOLF, Sandoz Huningue S.A., Huningue, France (Chaps. 1 – 7, 9)

BANSI LAL KAUL, Sandoz Huningue S.A., Huningue, France (Chap. 8)

1. Introduction 2
2. Antioxidants 3
2.1. Purpose and Requirements 3
2.2. Chemical Classes 4
2.3. Mechanism of Action 6
2.4. Test Methods 7
2.5. Uses 8
2.6. Legal Aspects 8
2.7. Economic Aspects 8
3. Light Stabilizers 10
3.1. Purpose and Requirements 10
3.2. Chemical Classes 11
3.3. Mechanism of Action 11
3.4. Test Methods 17
3.5. Uses 18
3.6. Economic Aspects 19
4. Heat Stabilizers for Poly(Vinyl Chloride) 19
4.1. Purpose and Requirements 19
4.2. Chemical Classes 19
4.2.1. Metal-Containing Stabilizers 19
4.2.2. Metal-Free Stabilizers 20
4.3. Mechanism of Action 21
4.4. Test Methods 22
4.5. Uses 22
4.6. Economic Aspects 23
5. Lubricants, Slip, Antiblocking, and Mold-Release Agents 23
5.1. Purpose and Requirements 23
5.2. Chemical Classes 24
5.3. Mechanism of Action 25
5.4. Test Methods 25
5.5. Uses 26
5.6. Economic Aspects 26
6. Flame Retardants 27
6.1. Purpose and Requirements 27
6.2. Chemical Classes 28
6.3. Mechanism of Action 29
6.4. Test Methods and Standards 35
6.5. Uses 35
6.6. Toxicology 37
6.7. Economic Aspects 38
7. Fillers and Coupling Agents 38
7.1. Purpose and Requirements 38
7.2. Forms and Chemical Structures . . 38
7.3. Mechanism of Action 39
7.4. Test Methods 39
7.5. Uses 40
7.6. Economic Aspects 41
8. Dyes and Pigments 41
8.1. Requirements 42
8.2. Pigments 42
8.2.1. Inorganic Pigments 42
8.2.2. Organic Pigments 43
8.2.3. Dyes 44
9. Miscellaneous Additives 44
9.1. Nucleating Agents 44
9.2. Antistatic Agents 46
9.3. Impact Modifiers 47
9.4. Chemical Blowing Agents 48
9.5. Optical Brighteners 49
10. References 50

Abbreviations:

ABS	acrylonitrile – butadiene – styrene polymer
HIPS	high-impact polystyrene
PBT	poly(butylene terephthalate)
HDPE	high-density polyethylene
LDPE	low-density polyethylene
LLDPE	linear low-density polyethylene
PETP	poly(ethylene terephthalate)
PVC	poly(vinyl chloride)
SAN	styrene – acrylonitrile copolymer

1. Introduction

As plastics find new applications, they face increasingly stringent requirements regarding their service life, durability, and many other properties. A variety of approaches are taken to meet these requirements. One approach is to modify the chemistry of the macromolecules

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_459

from which plastics are made (e.g., by using new monomers or copolymerization). Polymer blends also offer a great deal of scope for innovation. Another important way of improving the properties of plastics is to employ additives. These agents have made a decisive contribution to the widespread use of plastics and promise to be useful tools for solving future problems.

Plastics additives can be classified into three main groups:

1) Additives that stabilize plastics against degradation and aging during processing or in use. Degradation usually involves chain cleavage of the macromolecules and can proceed through the addition of energy (e.g., shear forces, heat, UV light) or chemical attack (e.g., oxidation, hydrolysis). These additives are called antioxidants, light stabilizers, or heat stabilizers.
2) Additives that facilitate or control processing (e.g., lubricants, mold-release agents, or blowing agents).
3) Additives that impart new, desirable qualities to plastics, such as resistance to burning, transparency or color, improved mechanical or electrical properties, dimensional stability, and degradability. Such additives include flame retardants, fillers, dyes, pigments, antistatic agents, nucleating agents, optical brighteners, impact modifiers, and plasticizers. Plasticizers are discussed elsewhere, → Plasticizers.

Surveys of additive groups are given in [1], [2]. New developments are published monthly in [3]. The total amount of additives (excluding fillers and pigments) used worldwide by the plastics industry in 1990 is estimated at ca. 1.7×10^6 t (Fig. 1).

The concentration of additives varies greatly, ranging from a few parts per million for some stabilizers to more than 50 % for certain flame retardants or fillers.

Although additives can be added to the monomer prior to polymerization, they are usually introduced immediately after polymerization, blended, and extruded to form granular (pelletized) products and compounds. Many additives are not introduced until the granules are processed into moldings, films, or fibers.

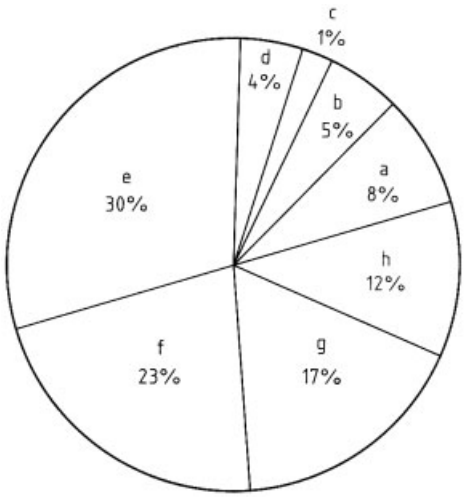

Figure 1. World consumption of plastics additives excluding fillers and pigments in 1990 (100 % = 1.7×10^6 t)
a) Lubricants and related compounds; b) Antioxidants; c) Light stabilizers; d) Miscellaneous additives; e) Flame retardants; f) Plasticizers; g) Poly(vinyl chloride) heat stabilizers; h) Impact modifiers

Additives with monomer functions can be incorporated chemically into the plastic if they are introduced before polymerization. In the case of agents added before processing, the use of "master batches" (i.e., granular products containing the concentrated additive in a polymer vehicle) is often advisable.

Plastics generally contain many additives. Preblended additive systems and combination master batches are commercially available; they contain optimal proportions of additives that are mutually compatible or have a synergistic action.

2. Antioxidants

See also → Antioxidants.

2.1. Purpose and Requirements

Antioxidants prevent or retard the autoxidation of polymers and minimize associated damage (e.g., discoloration, reduction in gloss, cracking, embrittlement); i.e., they stabilize the physical properties of plastics. Oxidation reactions generally proceed via different mechanisms that depend on the structure of the polymer. They are

often catalyzed by catalyst residues and contaminants. They are also accelerated by the addition of thermal or mechanical energy during plastics production and processing.

Three forms of stabilization are used: prestabilization, stabilization during processing, and long-term stabilization.

Most antioxidants are themselves oxidized and consumed in performing their function, so that the oxidation behavior of the additive in a given polymer is crucial for its effectiveness [4]. A number of other requirements apply to antioxidants:

1) They must be thermally stable and nonvolatile at processing temperatures
2) They must be soluble in polymers and no chalking should occur at service temperatures
3) They must not have an intrinsic color, and their oxidation products must have minimal color
4) Any acidic hydrolysis products must not corrode machinery
5) They must resist extraction
6) They must be odorless and tasteless
7) They must not create toxicity problems (many must be approved as indirect food additives)

2.2. Chemical Classes

Antioxidants are divided into two classes on the basis of their mode of action (see Section 2.3):

1) *Primary antioxidants*: sterically hindered phenols, secondary aromatic amines, and sterically hindered amines (HALS)
2) *Secondary antioxidants*: phosphites, phosphonites, thioethers (sulfides), and metal salts

Primary antioxidants can be used by themselves for prestabilization and long-term stabilization. Secondary antioxidants are used in combination with primary antioxidants, especially for stabilization during processing and for long-term stabilization under severe thermal conditions. Synergistic effects often occur in such systems (e.g., phenol and phosph(on)ites or phenol and thioethers). Antagonistic effects are also observed, however (e.g., between thioethers and HALS) [5].

The stabilizer system often includes other components that complex or neutralize degradation-promoting catalysts or acid traces. These may be metal deactivators (oxamides, hydrazones, and hydrazides) or acid scavengers [calcium or magnesium stearate, hydrotalcite (basic Mg/Al carbonate), and epoxy compounds].

Sterically Hindered Phenols. Industrial products are mostly derived from 2,6-di-*tert*-butylphenol (**1**; $n = 1$, R = H); or, less often, from2-*tert*-butyl-6-methylphenol, 2-*tert*-butyl-5-methylphenol, or other hindered phenols. Tabulations are given in the literature [1], [6].

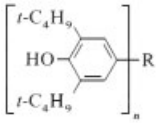

1

The simplest representative and oldest product of this class is 2,6-di-*tert*-butyl-4-methylphenol [*128-37-0*] (butylated hydroxytoluene, BHT) with $n = 1$ and R = CH_3. Variation of n from 1 to 4 and of R results in a range of hindered phenols with high molecular mass and thus low volatility, good solubility in certain polymers, and only slightly colored oxidation products. Products **2**–**4** are commonly used.

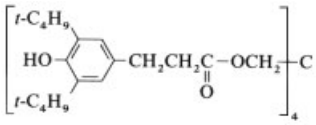

2 [*6683-19-8*]

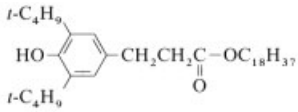

3 [*2082-79-3*]

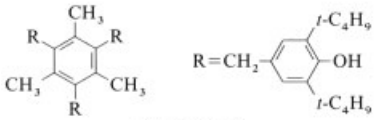

4 [*1709-70-2*]

The natural compound α-tocopherol [*59-02-9*] or vitamin E (**5**) is also used as an antioxidant [7], [8].

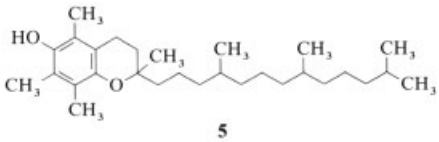

5

Secondary Aromatic Amines and Dihydroquinolines. Industrial products are derived chiefly from diphenylamine or *p*-phenylenediamine. A widely used representative of this class is 4-isopropylaminodiphenylamine [*101-72-4*] (**6**):

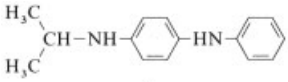

6

2,2,4-Trimethyl-1,2-dihydroquinoline [*147-47-7*] (**7**) should also be mentioned:

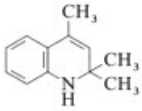

7

Sterically Hindered Amines (HALS). The abbreviation HALS stands for " hindered amine light stabilizers." They are described in more detail in Chapter 3 and are based on the 2,2,6,6-tetramethylpiperidine structure.

Phosphites and Phosphonites. The simplest industrial phosphite antioxidants are tris(nonylphenyl) phosphite [*26523-78-4*] (TNPP) (**8**) and tris(2,4-di-*tert*-butyl) phosphite [*31570-04-4*] (**9**).

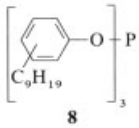

8

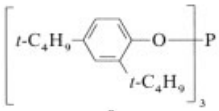

9

The cyclic phosphites **10**, **11** [9], and **12** [10] are also available:

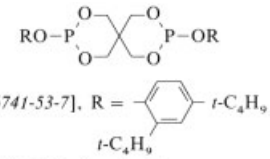

10a [*26741-53-7*], R = t-C_4H_9

10b [*3806-34-6*], R = $C_{18}H_{37}$

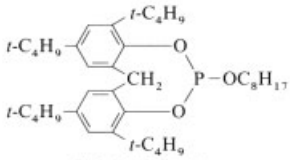

11 [*126050-54-2*]

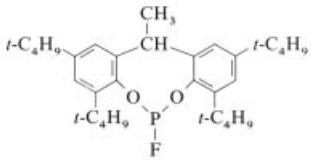

12 [*118337-09-0*]

Compound **13** is an industrially used phosphonite antioxidant [11]:

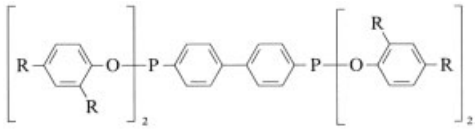

13 [*38613-77-3*], R = t-C_4H_9

The substituents influence not only the volatility and the solubility in polymers, but also the stability of the product to hydrolysis.

Organosulfur Compounds. Organosulfur antioxidants in practical use are the dilauryl [*123-28-4*] (**14**, $n = 12$), dimyristyl [*16545-54-3*] (**14**, $n = 14$), and distearyl [*693-36-7*] (**14**, $n = 18$) esters of thiodipropionic acid (**14**) which are abbreviated as DLTDP, DMTDP, and DSTDP, respectively.

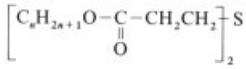

14

Disulfides such as dioctadecanyl disulfide [*2500-88-1*] or higher molecular mass products such as **15** are used less often.

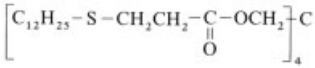

15 [*29598-76-3*]

Metal Compounds. Zinc and nickel dithiocarbamates or mercaptobenzimidazoles are sometimes used.

Metal Deactivators. The oxamides or hydrazine derivatives used as metal deactivators often have hindered phenol functions in the molecule as well, so that they also have a primary antioxidant action [1]. A typical structure is:

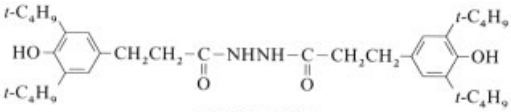

16 [*32687-78-8*]

2.3. Mechanism of Action

Thermooxidative degradation is an autocatalytic, free-radical chain reaction that involves initiation, propagation, branching, and termination stages. Primary radical sites $R^{\cdot}$ are initially formed by the abstraction of labile hydrogen atoms (e.g., tertiary or allylic hydrogens or those in the α-position to ether oxygen or amide nitrogen atoms). In the presence of oxygen, the primary radicals are oxidized to peroxy radicals $ROO^{\cdot}$. The peroxy radicals then take part in intermolecular (Eq. 1) and intramolecular (Eq. 2) chain propagation reactions to form new radical sites and hydroperoxides:

$$ROO^{\cdot} + RH \rightarrow ROOH + R^{\cdot} \quad (1)$$

(2)

The chain propagation reactions are rate determining for thermooxidative degradation. Tertiary hydrogen atoms are about three times more reactive than secondary hydrogens. The hydroperoxides undergo thermal decomposition or participate in reactions catalyzed by trace metals to form oxy ($RO^{\cdot}$, $HO^{\cdot}$) or peroxy radicals ($ROO^{\cdot}$), which cleave additional bonds in the polymer and generate new radical sites in the chain or break the polymer chain to give carbonyl groups:

The extent of thermooxidative degradation in many polymers can be estimated via the concentration of peroxy radicals, hydroperoxide groups, and carbonyl groups. Macroradicals can also react with C=C bonds in the polymer, especially with vinyl groups [12], or can recombine and result in undesirable cross-linking (gel formation).

Antioxidants act by trapping free radicals formed in the polymer (primary antioxidants), reduce hydroperoxides to hydroxyl groups (secondary antioxidants), or form complexes with trace metals that would otherwise catalyze hydroperoxide decomposition (metal deactivators). Antioxidant mechanisms are described in [13–18]; see also → Antioxidants, Chap. 3.2.→ Antioxidants, Chap. 3.3.

Hindered Phenols. Hindered phenols are radical-trapping antioxidants for oxy and especially peroxy radicals:

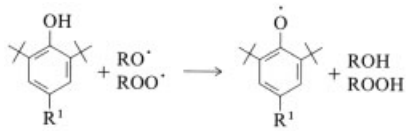

The phenoxy radicals with their bulky substituents are stabilized by steric hindrance and cannot attack the polymer. These radicals are transformed to quinonoid structures that can recombine with other peroxy radicals:

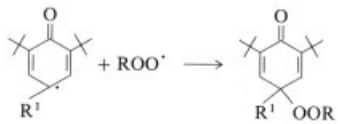

The formation of yellow quinonoids is a drawback of this class. The intensity of the color depends on the structure of R^1 ; it is particularly high in the case of BHT due to the formation of stilbenequinones.

Secondary aromatic amines also act by trapping peroxy radicals. Since the products derived from secondary aromatic amines are generally dark in color, their use is restricted to colored elastomers. Their stabilizing action against ozone degradation should be noted.

HALS are used primarily as light stabilizers. They can trap both carbon radicals and peroxy radicals at ambient temperatures. The active species is a hindered nitroxyl radical, which is regenerated (see Chap. 3). HALS are effective long-term stabilizers [19], [20]; nonvolatile, high molecular mass derivatives are especially useful [21].

Phosphites and Phosphonites. The secondary phosphite and phosphonite antioxidants reduce hydroperoxides and are themselves oxidized to phosphates and phosphonates:

$$P^{III} + ROOH \rightarrow P^{V} + ROH$$

Peroxy radicals are analogously reduced to oxy radicals. Phosphites and phosphonites may also complex and thus deactivate trace metals (e.g., polymerization catalyst residues). Their oxidation products are colorless, so they are particularly successful as color stabilizers. Their action is especially useful in the processing of polymer melts, where they also protect phenolic antioxidants against premature oxidation.

Organosulfur compounds decompose hydroperoxides. Sulfides, for example, are converted to sulfoxides, sulfenic acids, and sulfur dioxide, which in turn can reduce further hydroperoxides. Sulfur-containing stabilizers can also regenerate phenolic antioxidants from their quinonoid structures. Their action is most pronounced at higher temperatures in solid plastics. The odor of the reaction products can be a drawback.

Metal deactivators act by chelating metal ions [22], particularly metals that have a number of oxidation states and catalyze hydroperoxide decomposition even at low temperatures. These metals include copper (from cablewires), iron, manganese (from fillers), and titanium (from polymerization catalysts).

For the effects of metals and metal deactivators, see [23].

2.4. Test Methods

The two most important test methods for antioxidants are the multiple extrusion test and the oven test (DIN 53 383, ASTM D 1870-68).

The *multiple extrusion test* examines degradation in the polymer melt (i.e., stabilization during processing). The antioxidant is blended with the polymer, and the resulting granules are regranulated up to ten times in succession in an extruder under defined temperature conditions. The yellowness index (ASTM D 1925-70) and the melt viscosity (melt flow index, MFI, ASTM D 1238-70) or solution viscosity are measured. The melt viscosity can also be determined online in an extruder equipped with a bypass [24].

In the *oven test*, polymer degradation is studied in the solid phase (i.e., long-term stabilization is determined). Test plates of the polymer are placed on rotating trays in special furnaces with air recirculation. Their appearance (browning, reduction in gloss and transparency), mechanical properties (brittleness, tensile strength, impact strength), and analytical data (carbonyl bands in the IR) are evaluated after various times. To shorten the test period, the oven test is often run at 100 – 150 °C, but correlation of the results with room temperature is problematic especially in the case of metal deactivators.

Other test methods include

1) Thermogravimetric analysis, especially for plastics such as polyoxymethylene that tend to undergo thermal depolymerization
2) Torque measurements on polymer melts in the Brabender, especially for plastics that tend to cross-link (e.g., polyethylene)
3) Transparency measurements on extruded films
4) Determination of the induction period in an oxygen absorption test at elevated temperature (OIT test)

2.5. Uses

Polyolefins. The optimal antioxidants for polyolefins depend on the type of resin, production process (catalyst; solution, suspension, or bulk polymerization), comonomer, filler, application (moldings, films, fibers), and other factors. Modern bulk polymerization processes, in which purification steps are omitted or polymers are produced directly as spherical granules, have led to reformulation of stabilization systems.

The primary antioxidant BHT, with its high volatility and yellowing tendency, is losing ground to higher molecular mass hindered phenols (e.g., **2**–**4**). For prestabilization up to 250 ppm of **3** is usually added at the polymerization stage; ca. 500 – 1000 ppm of a phenolic antioxidant is subsequently added for stabilization during processing and service. Higher concentrations are seldom used because of the risk of yellowing.

The main secondary antioxidants used during processing are the phosphites **8**–**10** and the phosphonite **13**, whereby especially **13** not only stabilizes the melt viscosity but also offers much better color stabilization than phosphites such as **8** and **9**. The phosph(on)ites are used at concentrations of 500 – 2000 ppm.

Figures 2 and 3 illustrate the stabilization obtained with phosphites and phosphonites during the processing of polypropylene and linear low-density polyethylene (LLDPE), respectively. Whereas polypropylene exhibits an increase in melt flow index (degradation by chain cleavage), LLDPE shows a decline (degradation dominated by cross-linking).

The selection of antioxidants also depends on the solubility in the polymer, which is especially critical for LLDPE. If the polymerization process already yields a granular product in the reactor, antioxidants can be applied to the reactor granules by spraying on the surface. Low-melting products such as **13** are especially suitable for this purpose.

Thioethers are preferentially used as synergists when severe thermal requirements apply to long-term stabilization. They are employed at concentrations of ca. 1000 – 3000 ppm.

Metal deactivators are used at the 500 – 5000 ppm level (higher concentrations in talc-filled polypropylene).

Other formulation guidelines are given in [1].

Styrene Polymers. Phenolic antioxidants (up to 2000 ppm) are added to high-impact polystyrene modified with polybutadiene. They are usually introduced before polymerization. Phosphites and phosphonites are sometimes employed as color stabilizers.

In acrylonitrile – butadiene – styrene copolymers (ABS), the antioxidants are usually introduced into the polybutadiene latex phase by emulsification before it is blended with polystyrene – polyacrylonitrile. Along with 1000 – 2000 ppm of phenolic antioxidants, larger amounts of phosphite **8** or DLTDP are used.

Other Plastics [1]. Antioxidants play a less important role in plastics other than polyolefins and the styrene polymers described above. Secondary aromatic amines, BHT, TNPP, as well as high molecular mass phenols are used in elastomers and polyurethanes.

Although engineering thermoplastics (e.g., polycarbonates, polyesters, polyamides) are less sensitive to thermooxidative degradation, hindered phenols and phosph(on)ites are used for color stabilization or long-term stabilization.

2.6. Legal Aspects

Because plastics are widely used in food packaging and agriculture, antioxidants must be approved as indirect food additives by national agencies such as the FDA (USA) and the Bundesgesundheitsamt (Federal Health Office, Germany). Unified approvals are expected to apply throughout the EC as of 1993. Approval is based on toxicity and extraction tests (extraction of antioxidants from plastics by food simulants). Most well-established antioxidants have received broad approval for food use.

2.7. Economic Aspects

Table 1 presents 1990 market volumes of antioxidants. Polyolefins, ABS, and polystyrene account for ca. 80 – 90 % of the antioxidant market. Prices of antioxidants are given in [25].

Antioxidants are often marketed in the form of powder blends. The most important producers and trade names are listed in Table 2.

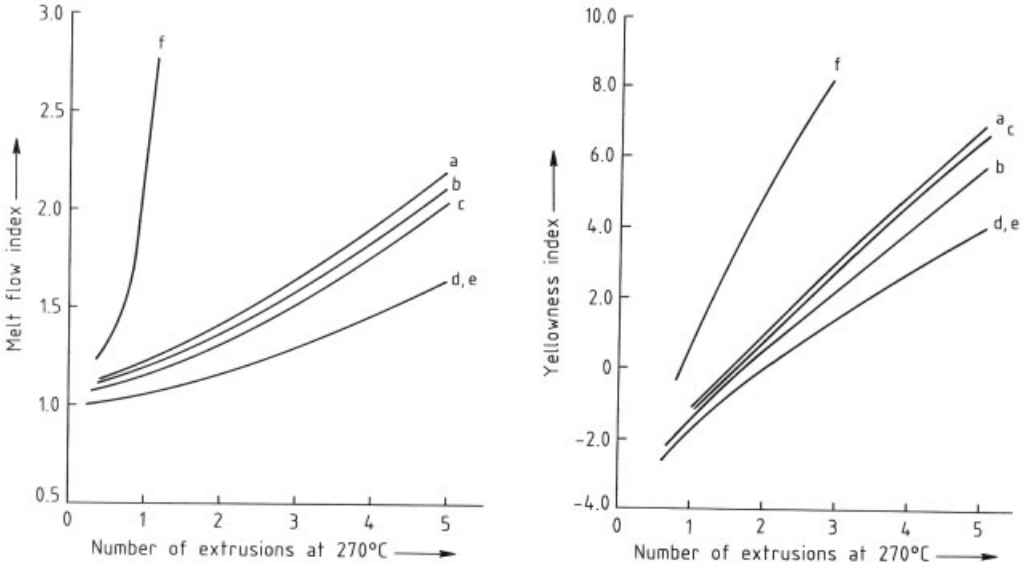

Figure 2. Stabilization of polypropylene in a multiple extrusion test with 0.1 % phenolic antioxidant (**2**) and 0.1 % calcium stearate as basic stabilization and 0.1 % phosph(on)ite process stabilizer
a) Phosphite (**9**); b) Phosphite (**10a**); c) Phosphite (**12**); d) Phosphonite (**13**); e) Phosphite (**11**); f) No process stabilizer

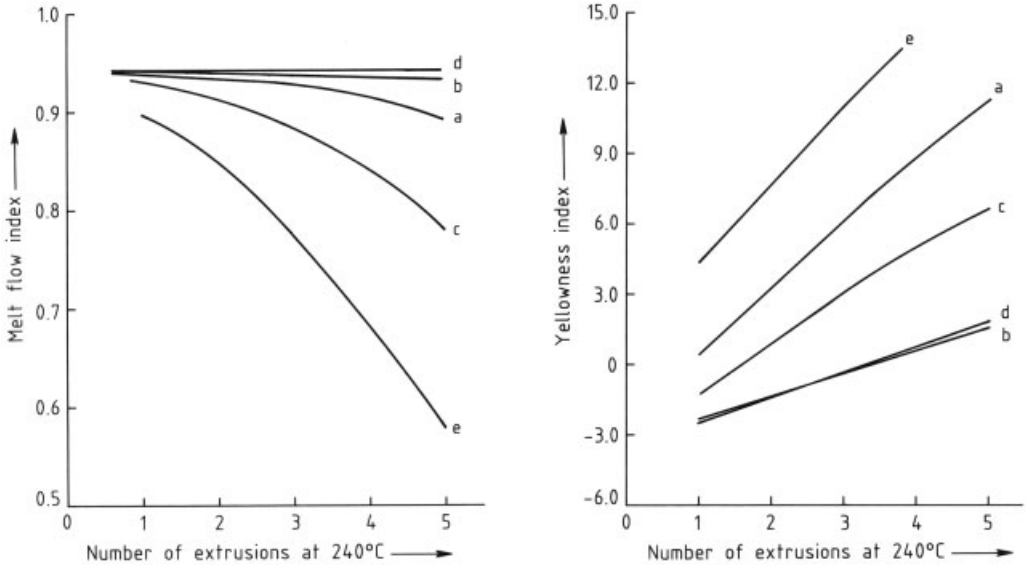

Figure 3. Stabilization of linear low-density polyethylene in a multiple extrusion test with 0.1 % phenolic antioxidant (**3**) and 0.1 % calcium stearate as basic stabilization and 0.15 % phosph(on)ite process stabilizer
a) Phosphite (**9**); b) Phosphite (**10a**); c) Phosphite (**10b**); d) Phosphonite (**13**); e) No process stabilizer

Table 1. Estimated market volume of antioxidants in 1990 (10^3 t)

Region	Phenols	Thioethers	Phosph(on)ites	Other	Total
United States	14	3	6	2	25
Western Europe	15	3	7	3	28
Japan	8	2	3	2	15
Other	10	2	3	2	17
Total	47	10	19	9	85

Table 2. Manufacturers and trade names of antioxidants

Manufacturer	Trade names	Phenols	Phosph(on)ites	Thioethers	Aromatic amines	Metal deactivators	Blends
Adeka Argus	Mark	X	X	X		X	X
Akzo Chemie	Perkanox	X					
American Cyanamid	Cyanox	X		X			X
BASF	Sicostab	X					X
Bayer	Vulkanox	X			X		
B. F. Goodrich Chem.	Goodrite	X			X		
Chem. Werke Lowi	Lowinox	X	X				
Ciba-Geigy	Irganox, Irgafos	X	X	X		X	X
Daiichi Kogyo Seiyaku	Lasumit			X			
Dover Chemicals	Doverphos		X				
Eastman Chemicals	Tenox	X				X	
Enichem Synthesis	Anox, Alkanox	X	X				X
Ethyl Corporation	Ethanox	X	X				
Ferro Corporation	Oxi-chek	X					
General Electric	Ultranox, Weston	X	X		X		X
W. R. Grace				X			
Hitachi	Cunox					X	
Hoechst	Hostanox	X	X	X		X	
ICI	Nonox, Topanol	X			X		X
Jokoku Chemicals			X				
Mitsubishi Petrochem.	Seenox	X		X			
Monsanto	Santonox, Santowhite Flectol	X			X		
Morton Thiokol	Carstab			X			
Olin Chemicals	Wytox	X	X				
Pennwalt	Anoxyn			X			
PMC Specialities Group	Prodox	X					
Raschig	Ralox	X					
Rhône-Poulenc	Rhodianox, Garbafix	X	X				
R. T. Vanderbilt	AgeRite, Vanox	X	X	X	X		
Sandoz Chemicals	Sandostab		X				
Schenectady Chemicals	Isonox	X					
Shell	Ionol, Ionox	X					
Shipro Kasei	Seenox	X		X			
Sumitomo	Sumilizer	X	X	X			
Ube Industries						X	
Uniroyal	Naugard, Aranox	X	X		X	X	X
Witco	Mark	X	X	X			X
Yoshitomo				X			

3. Light Stabilizers

3.1. Purpose and Requirements

Synthetic and natural polymers vary in their sensitivity to environmental influences. Ultraviolet radiation in sunlight plays a critical role because it has sufficient energy to break chemical bonds. The cleavage sites react with atmospheric oxygen and accelerate degradation (free-radical chain reactions; see Section 2.3). Many plastics therefore suffer yellowing, surface cracking, embrittlement, reduction of gloss, or chalking

after a short time in outdoor service, and ultimately disintegrate. Light stabilizers are used to maximize protection against photodegradation.

Ultraviolet radiation is subdivided into UV-A ($\lambda = 320 - 400$ nm), UV-B ($\lambda = 280 - 320$ nm), and UV-C ($\lambda < 280$ nm). Only UV-A and UV-B reach the Earth's surface in sunlight, and UV-B is largely responsible for the degradation of plastics.

The difference in damage (yellowing) caused by UV-A and UV-B in poly(vinyl chloride) (PVC) is shown in Figure 4. Plastics are generally fairly stable in the long-wavelength UV-A (350 – 400 nm). Most plastics are sensitive to wavelengths of 290 – 320 nm.

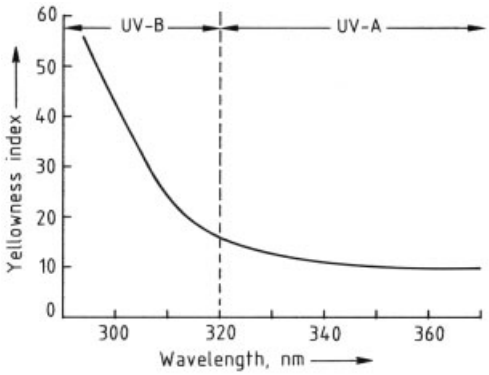

Figure 4. Damage in PVC as a function of the wavelength of UV light (300-h radiation)

Two methods are used to protect plastics against photodegradation: (1) incorporation of light stabilizers in the bulk plastic, and (2) coating with a light-stable or light-stabilized material that is largely opaque to the dangerous UV range.

Light stabilizers must have high light absorption; must resist extraction with water, hydrolysis, and thermal volatilization; and must themselves be stable toward UV radiation. Ideally, a light stabilizer should not be consumed while carrying out its function; it should operate in a closed cycle so that it still exists in active form even after a long period of weathering or use.

Other requirements include solubility in the polymer, stability under processing conditions, compatibility with other additives, and colorlessness. In certain applications, light stabilizers must be approved as indirect food additives.

The service life of a plastic is more than doubled by light stabilizers. Modern systems extend service life by a factor of 10 – 20 in many plastics, thus permitting their outdoor use.

3.2. Chemical Classes

Light stabilizers are classified in three groups according to their mode of action:

1) UV absorbers: 2-hydroxybenzophenones (BP-UVA), 2-hydroxyphenylbenzotriazoles (BT-UVA), oxalanilides (OA-UVA), 2-hydroxyphenyltriazines (TA-UVA), cinnamates (CA-UVA), salicylates (SA-UVA), and formamidines (FA-UVA)
2) Energy quenchers: nickel complexes (NIC)
3) Radical scavengers: hindered amines (HALS) and 4-hydroxybenzoates (HB)

Table 3 lists some of the most important light stabilizers now in use, along with their trade names and producers.

Stabilization can also be achieved by incorporation of carbon black, titanium dioxide, or zinc oxide. The effect of colored pigments varies: phthalocyanine blue pigments have a stabilizing action; organic red and yellow pigments sensitize photodegradation.

3.3. Mechanism of Action

Photodegradation of Polymers. Photodegradation is initiated when radiation is absorbed by chromophores in the polymer structure itself (e.g., C=C, C=O) or by impurities or defects. Energy can be dissipated from excited states by emission of radiation (fluorescence from short-lived singlet states or phosphorescence from long-lived triplet states), release of heat, or energy transfer to nearby chromophores. If none of these mechanisms is available, photodegradation takes place with bond cleavage and formation of radicals. These radical sites react with oxygen and initiate autoxidative chain reactions. Excited molecular segments can also transfer energy to oxygen, converting it to higher-energy singlet oxygen, which accelerates oxidative degradation.

Table 3. Light stabilizers

Structure, CAS registry no., code	Trade name (producer)
2-Hydroxybenzophenones	
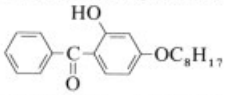 [*1843-05-6*] BP-UVA 1	Cyasorb UV 531 (American Cyanamid) Mark 1413 (Adeka Argus) Chimassorb 81 (Ciba-Geigy) UV-Chek AM 300 (Ferro) Hostavin ARO 8 (Hoechst) Rhodialux P (Rhône-Poulenc) Uvasorb 3 C (Sigma) Seesorb 102 (Shipro Kasei) Aduvex 248 (Shell) Lowilite 22 (Chem. Werke Lowi) Sumisorb 130 (Sumitomo) Viosorb 130 (Kyodo Yakuhin) Uvinul 408 (BASF)
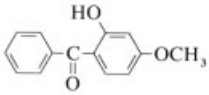 [*131-57-7*] BP-UVA 2	Cyasorb UV 9 (American Cyanamid) Uvinul M-40 (BASF) UVA Bayer 325 (Bayer) Chimassorb 90 (Ciba-Geigy) Gafsorb 2H 4M (GAF) Rhodialux A (Rhône-Poulenc) Uvasorb MET (Sigma) Seesorb 101 (Shipro Kasei) Viosorb 110 (Kyodo Yakuhin) Sumisorb 110 (Sumitomo)
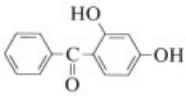 [*131-56-6*] BP-UVA 3	Uvinul 400 (BASF) Aduvex 12 (Shell) Gafsorb 24 DH (GAF) DHB (Riedel de Haen) Rhodialux D (Rhône-Poulenc) Seesorb 100 (Shipro Kasei) Viosorb 100 (Kyodo Yakuhin)
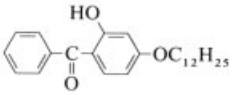 [*2985-59-3*] BP-UVA 4	Chimassorb 125 (Ciba-Geigy) Eastman Inhibitor DOBP (Eastman Chem.) Gafsorb 2H 4DD (GAF) Rhodialux 1200 (Rhône-Poulenc) Seesorb 103 (Shipro Kasei)
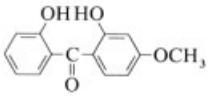 [131-53-3] BP-UVA 5	Cyasorb UV 24 (American Cyanamid) Aduvex 24 (Shell) Sumisorb 140 (Sumitomo)
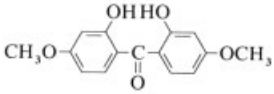 [*131-54-4*] BP-UVA 6	Univul D-49 (BASF) Aduvex 424 (Shell)

Table 3. (Continued)

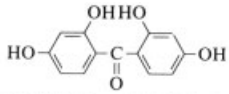 [*131-55-5*] BP-UVA 7	Aduvex 412 (Shell) Uvinul D 50 (BASF)
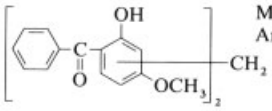 [69119-80-8] BP-UVA 8	Mark LA 51 (Adeka Argus)
2-Hydroxyphenylbenzotriazoles	
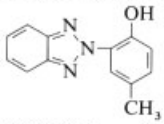 [2440-22-4] BT-UVA 1	Tinuvin P (Ciba-Geigy) Mark LA 32 (Adeka Argus) Uvasorb SV (Sigma) Seesorb 701 (Shipro Kasei) Lowilite 55 (Chem. Werke Lowi) Viosorb 520 (Kyodo Yakuhin) Sumisorb 200 (Sumitomo)
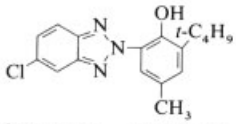 [*3896-11-5*] BT-UVA 2	Tinuvin 326 (Ciba-Geigy) Mark LA 36 (Adeka Argus) Seesorb 703 (Shipro Kasei) Viosorb 550 (Kyodo Yakuhin) Sumisorb 300 (Sumitomo)
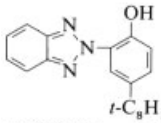 [*3147-75-9*] BT-UVA 3	Cyasorb UV 5411 (American Cyanamid) Sumisorb 340 (Sumitomo) Viosorb 583 (Kyodo Yakuhin) Seesorb 709 (Shipro Kasei)
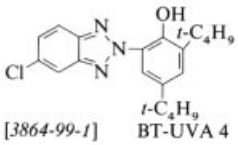 [*3864-99-1*] BT-UVA 4	Tinuvin 327 (Ciba-Geigy) Mark LA 34 (Adeka Argus) Seesorb 702 (Shipro Kasei) Viosorb 580 (Kyodo Yakuhin)
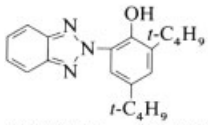 [*3846-71-7*] BT-UVA 5	Tinuvin 320 (Ciba-Geigy) Seesorb 705 (Shipro Kasei) Viosorb 582 (Kyodo Yakuhin) Sumisorb 320 (Sumitomo)
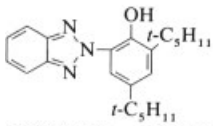 [*25973-55-1*] BT-UVA 6	Cyasorb 2337 (American Cyanamid) Tinuvin 328 (Ciba-Geigy) Seesorb 704 (Shipro Kasei) Viosorb 591 (Kyodo Yakuhin) Sumisorb 350 (Sumitomo)

Table 3. (Continued)

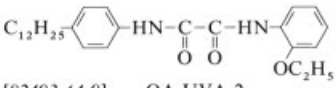

Sanduvor 3206 (Sandoz)

[*82493-14-9*] OA-UVA-2

Salicylates

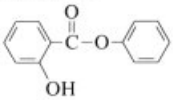

Eastman Inhibitor OPS (Eastman Chem.)
Seesorb 201 (Shipro Kasei)

[*118-55-8*] SA-UVA-2

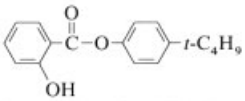

Seesorb 202 (Shipro Kasei)
Rhodialux K (Rhône-Poulenc)
Viosorb 90 (Kyodo Yakuhin)

[*87-18-3*] SA-UVA-2

Formamidines

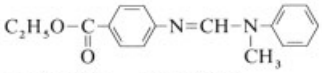

Givsorb UV-1 (Givaudan)

[*57834-33-0*] FA-UVA-1

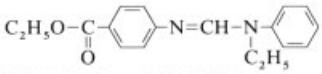

Givsorb UV-2 (Givaudan)

[*65816-20-8*] FA-UVA-2

4-Hydroxybenzoates

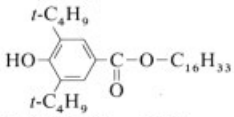

Cyasorb UV 2908 (American Cyanamid)

[*67845-93-6*] HB-1

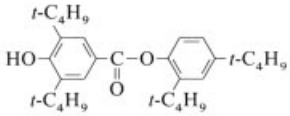

Tinuvin 120 (Ciba-Geigy)
Seesorb 712 (Shipro Kasei)
UV-Chek AM 340 (Ferro)
Viosorb 80 (Kyodo Yakuhin)
Sumisorb 400 (Sumitomo)

[*4221-80-1*] HB-2

Nickel complexes

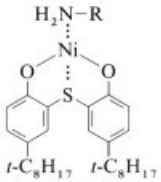

Cyasorb UV 1084 (American Cyanamid)
Chimassorb N 705 (Ciba-Geigy)
Rhodialux Q 84 (Rhône-Poulenc)
Uvasorb Ni (Sigma)

R = C_4H_9 [*14516-71-3*] NIC-1

same,
R = C_8H_{17} [*67668-65-9*] NIC-2

Seesorb 612 NH (Shipro Kasei)

Table 3. (Continued)

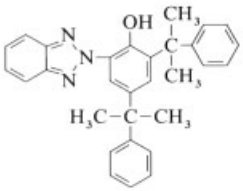

Tinuvin 234 (Ciba-Geigy)
Tinuvin 900 (Ciba-Geigy)

[*70321-86-7*] BT-UVA 7

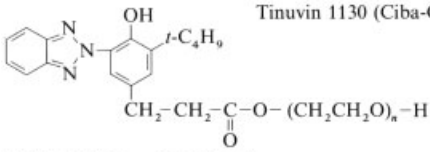

Tinuvin 1130 (Ciba-Geigy)

[*104810-48-2*] BT-UVA 8

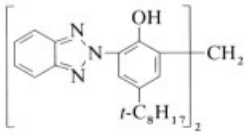

Mark LA 31 (Adeka Argus)

[*103597-45-1*] BT-UVA 9

2-Hydroxyphenyltriazines

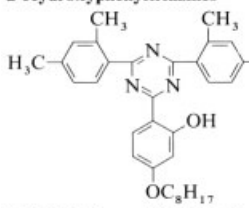

Cyasorb 1164 (American Cyanamid)

[*2725-22-6*] TA-UVA-1

Cinnamates

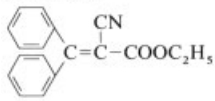

Uvinul N-35 (BASF)
Seesorb 501 (Shipro Kasei)
Viosorb 910 (Kyodo Yakuhin)

[*5232-99-5*] CA-UVA-1

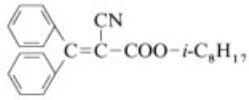

Uvinul 539 (BASF)

[*6197-30-4*] CA-UVA-2

Oxalanilides

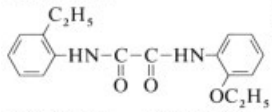

Sanduvor VSU (Sandoz)
Tinuvin 312 (Ciba-Geigy)

[*23949-66-8*] OA-UVA-1

Table 3. (Continued)

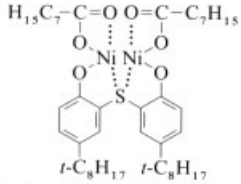 [*89073-38-1*] NIC-3	UV-Chek AM 205 (Ferro)
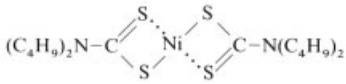 [*13927-77-0*] NIC-4	UV-Chek AM 104 (Ferro) Antigene NBC (Sumitomo) Vanox NBC (R. T. Vanderbilt)
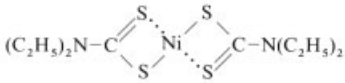 [*14267-17-5*] NIC-5	Robac Ni PP (Robinson Brothers)
Hindered amines	
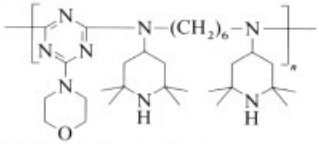 [*90751-07-8*] HALS 1	Cyasorb UV 3346 (American Cyanamid)
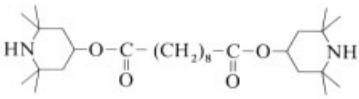 [*52829-07-9*] HALS 2	Tinuvin 770 (Ciba-Geigy) Mark LA 77 (Adeka Argus) Sanol LK 770 (Sankyo) Lowilite 77 (Chem. Werke Lowi)
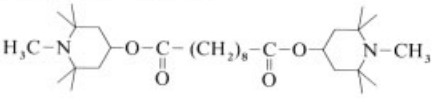 [*41556-26-7*] HALS 3	Tinuvin 765 (Ciba-Geigy) Tinuvin 292 (Ciba-Geigy) Sanol 292 (Sankyo)
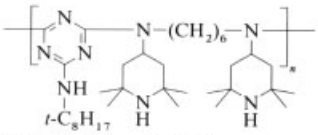 [*71878-19-8*] HALS 4	Chimassorb 944 (Ciba-Geigy)
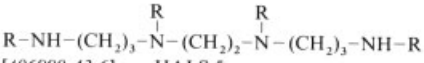 [*106990-43-6*] HALS 5 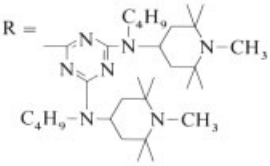	Chimassorb 119 (Ciba-Geigy)

Table 3. (Continued)

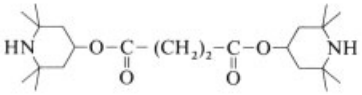

Tinuvin 780 (Ciba-Geigy)

[62782-03-0] HALS 6

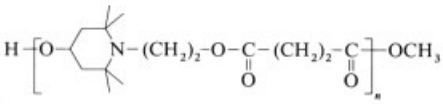

Tinuvin 622 (Ciba-Geigy)

[*70198-29-7*] HALS 7

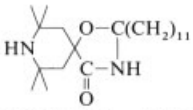

Hostavin N-20 (Hoechst)

[*64338-16-5*] HALS 8

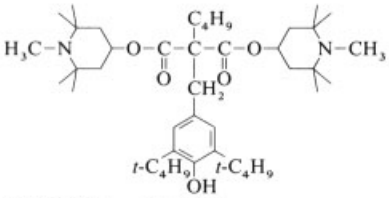

Tinuvin 144 (Ciba-Geigy)

[*63843-89-0*] HALS 10

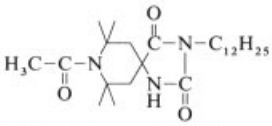

Tinuvin 440 (Ciba-Geigy)

[82537-67-5] HALS 10

Tinuvin 123 (Ciba-Geigy)

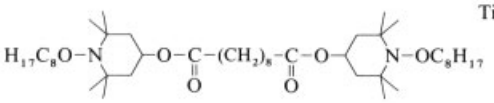

[*122586-52-1*] HALS 11

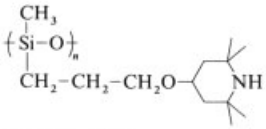

Uvasil 299 (Enichem)

M_r 1100 – 2500
[*102089-33-8*] HALS 12

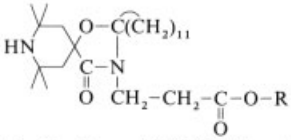

Sanduvor 3050 (Sandoz)

R = $C_{12}H_{25}$ [*85099-51-0*] HALS 13
R = $C_{14}H_{29}$ [*85099-50-9*]

Table 3. (Continued)

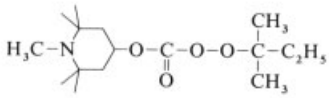

Lupersol HA 505 (Atochem)

[*116214-18-7*] HALS 14

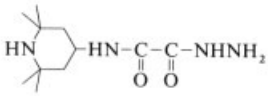

Luchem HA-R 100 (Atochem)

[*122035-71-6*] HALS 15

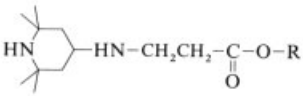

Sanduvor 3052 (Sandoz)

R = $C_{12}H_{25}$ [*119530-69-7*] HALS 16
R = $C_{14}H_{29}$ [*119530-70-0*]

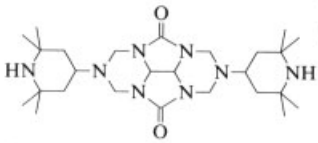

Uvinul 4049 H (BASF)

[*109423-00-9*] HALS 17

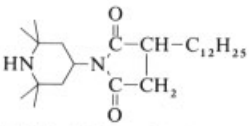

Cyasorb UV 3581 (American Cyanamid)
Sanduvor 3055 (Sandoz)

[*79720-19-7*] HALS 18

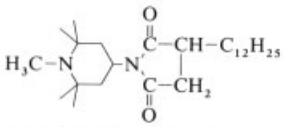

Cyasorb UV 3604 (American Cyanamid)
Sanduvor 3056 (Sandoz)

[*106917-30-0*] HALS 19

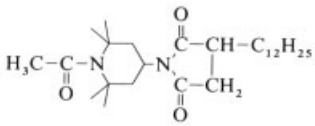

Cyasorb UV 3668 (American Cyanamid)
Sanduvor 3058 (Sandoz)

[*106917-31-1*] HALS 20

Oxidation to carbonyl groups in the macromolecules results in chain cleavage via the Norrish reactions:

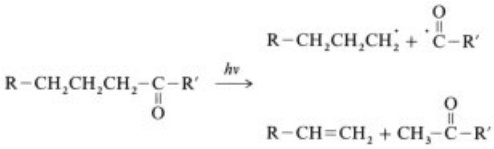

Many plastics with aromatic residues (e.g., polycarbonates, polyesters, polyamides) tend to participate in "photo-Fries reactions", whose products may be colored or may be precursors of colored products:

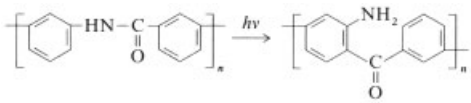

In semicrystalline plastics, UV radiation is scattered by the crystallites and increases the quantum efficiency of polymer degradation. Photodegradation occurs mainly in the amorphous, oxygen-permeable regions.

Ultraviolet Absorbers. The UV absorbers listed in Table 3 have intense absorption bands in the UV-B and some of them in the UV-A as well. They therefore filter out harmful radiation before it reaches the chromophores in the plastic.

The UV absorbers must dissipate the absorbed energy quickly as heat and not as radiation. Mechanisms proposed for this action include proton transitions in the excited singlet state [26–28]. Other quenching mechanisms have also been postulated [29].

The position and intensity of absorption can be modified by substituents on the aromatic rings. For example, alkoxy substituents in the 4- and 4′-positions of 2-hydroxybenzophenones bring about bathochromic shifts in the absorption maxima, as do alkyl substituents in the phenyl ring of 2-hydroxyphenylbenzotriazoles. The steepness of the UV absorption curve on the long-wavelength side is important with respect to intrinsic color–even slight residual absorption around 400 nm causes yellow coloration.

The UV absorbers hinder the photolysis of polymers by competing for UV absorption. They are thus particularly effective when degradation is due primarily to photolysis and not to

oxidation. Since the intensity of radiation depends on the thickness of the absorbing layer (Lambert – Beer law), these substances are more effective within the plastic than at its surface. They are therefore only marginally effective in thin films and fibers.

Energy Quenchers. Light stabilizers based on nickel complexes, like some UV absorbers, are energy quenchers. They absorb energy from excited polymer regions or from singlet oxygen, and dissipate it as heat. The action of quenchers depends on concentration, not on layer thickness.

Radical Scavengers. Sterically hindered 4-hydroxybenzoates and amines act chiefly as radical scavengers. 4-Hydroxybenzoates can be regarded as photostable phenolic antioxidants [30]. The hindered amines (HALS) are much more important, however; they all have a 2,2,6,6-tetramethylpiperidine ring [31]. The cyclic mode of action of HALS is shown in Figure 5 [32–43]. The stabilizing action is due not to the HALS compound itself ($N-R^1$) but to the nitroxyl radical ($N-O^{\cdot}$) formed by photooxidation during weathering. The nitroxyl radical reacts with the polymer radicals formed by photolysis to give ethers ($N-O-R$), which then react with peroxy radicals to regenerate the nitroxyl radical. Other mechanisms are based on the trapping of acyl peroxide radicals [44] or the decomposition of hydroperoxides [45].

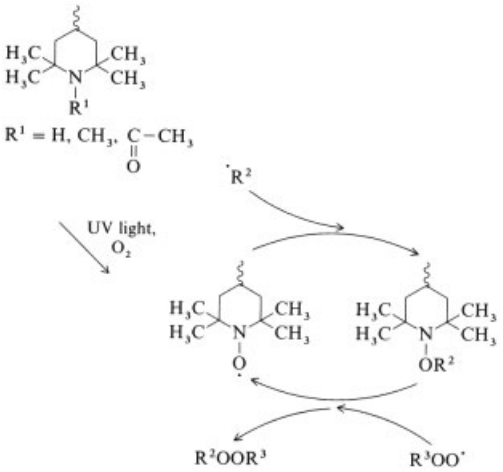

Figure 5. Stabilization mechanism of HALS

The action of HALS depends on their concentration and not on layer thickness. They are therefore equally as effective at the surfaces of plastics in films, fibers, coatings, and thick layers. They are, however, sensitive to certain chemicals, such as organic thioether costabilizers [5], halogen flame retardants [46], kaolin fillers, pesticides [47], and acids. Synergistic effects between HALS and UV absorbers are widely used in industry, and synergies have also been observed with nickel quenchers [48]. A combination of low molecular mass HALS (which migrate to the surface) and high molecular mass compounds (which are more strongly anchored in the bulk polymer) is also recommended [21]. Developments include 2,2,6,6-tetramethylpiperazinones [49] and HALS that are chemically bound to polyolefins [50].

3.4. Test Methods

Light stabilizers are incorporated into polymers; processed into specimens in plate, film, fiber, or coating form; and then exposed to radiation in the presence or absence of moisture to determine their weather resistance and lightfastness, respectively.

Weathering may be natural or artificial (i.e., accelerated). Changes in color (DIN 53 236), yellowness index (ASTM D 1925-70), gloss (DIN 67 530), cracking (ISO 4628 Part 4), tensile strength, elongation at break (ASTM D 638, ISO IR 527), or impact resistance (ASTM D 256, ISO R 179) are used to assess lightfastness and weather resistance. The intensity of the IR carbonyl bands is also very often used as a criterion of photodegradation. See also → Paints and Coatings, Chap. 9.2.5.; → Plastics, Properties and Testing, Chap. 2.7.3.

Open-air weathering tests are performed on a commercial basis in all climatic zones. Important sites are located in Central Europe (temperate climate), southern Florida (subtropical), Arizona (desert climate), Okinawa, Thailand, Australia, South Africa, Israel, and Singapore.

In *accelerated sunlight tests*, Fresnel reflectors are used to focus sunlight on the specimens (ASTM E 838-81). Examples of such tests include the Equatorial Mount with Mirrors for Acceleration (EMMA) test and the Equatorial Mount with Mirrors for Acceleration plus Water

Spray (EMMAQUA) test, which are performed in Arizona.

In *artificial weathering* the specimens are subjected to precisely defined climatic conditions (temperature, irradiation, rainfall) in specially designed apparatus. Xenon arcs, carbon arcs, and fluorescent lamps are used as illumination sources (see Table 4); filtered xenon light best simulates global irradiation [51], [52].

Table 4. Devices for accelerated weathering tests

Producer	Device	Light source
W.C. Heraeus GmbH, Hanau, Germany	Xenontest	xenon lamps
Atlas Electric Devices Company, Chicago, USA	Suntest Weather-Ometer	xenon lamps
	Fade-Ometer	carbon-arc lamps
	Sunchex	UV fluorescent tubes
Q-Panel Company, Cleveland, USA	UV-CON Q-U-V	UV fluorescent tubes
Weiss Technik AG, Reiskirchen, Germany	Global-UV-Testgerät	special UV lamps
Dainippon Plastics, Tokyo, Japan	Bauart BAM Eye Super UV Tester	special UV lamps

Standards for artificial weathering tests include DIN 53 384, DIN 53 231, and ANSI/ASTM G 53-57.

3.5. Uses

Light stabilizers are employed mainly in plastics for outdoor service (e.g., garden furniture, stadium seating, window frames, floor coverings) films for protective coverage and agriculture; light fixtures and illuminated signs; automotive parts such as bumpers and body paints; and sports equipment and clothing. Light stabilizers extend the service lives of these articles by years.

The approximate breakdown of light stabilizer uses according to type of plastic is as follows:

Polyolefins (polyethylene, polypropylene)	55 %
Acrylic and polyurethane coatings	20 %
PVC	10 %
ABS and polystyrene	5 %
Unsaturated polyester resins	4 %
Miscellaneous substrates	6 %

The stabilizers are generally used in concentrations of 0.05 – 0.5 %, reaching 3 % in especially critical applications. Recommended formulations can be found in [1].

Polypropylene is particularly sensitive to photodegradation. The photostabilization of polypropylene is reviewed in [53]. Thick moldings are stabilized with 0.2 – 0.6 % of UV absorbers (e.g., BP-UVA 1, BT-UVA 2) and/or radical scavengers (e.g., HB-1 and 2, HALS 1 – 8). HALS (chiefly of high molecular mass, such as HALS 1, 4, or 5) are employed to stabilize films and fibers.

Polyethylene. Although polyethylene is more stable to light than polypropylene, it requires good stabilization for outdoor service. The selection of light stabilizers is especially critical because of the danger of surface chalking due to migration of incompatible additives. Low-density and linear low-density polyethylene (LDPE and LLDPE) films contain up to 0.4 % of a nickel quencher and UV absorber combination (e.g., NIC-1 and BP-UVA 1) or of a polymeric HALS (e.g., HALS 1, 4, or 7). In filled grades, higher concentrations (ca. 1.5 %) can also produce more than a tenfold increase in service life in open-air weathering. In high-density polyethylene (HDPE) the best results are obtained with HALS (e.g., HALS 2, 4, or 7); 0.1 % is often sufficient.

Styrene Polymers. The UV absorbers and UV absorbers – HALS combinations (e.g., BT-UVA 1 with HALS 2) are recommended for polystyrene, ABS, and styrene – acrylonitrile copolymers (SAN) at a total concentration of ca. 1 % [54].

Poly(Vinyl Chloride). The lightfastness of PVC depends heavily on its heat stabilization or fillers. Barium and cadmium stabilizers, tin carboxylates, and titanium dioxide substantially improve lightfastness. Tin mercaptides are unsuitable for outdoor service or necessitate high levels of titanium dioxide. HALS are not very effective. For transparent PVC in particular, the use of 0.1 – 0.3 % UV absorbers (e.g., BP-UVA 1 or 2, BT-UVA 1 or 5, OA-UVA 1) is recommended.

Miscellaneous Plastics. Polyurethanes, especially those derived from aromatic isocyanates and polyether polyols, tend to yellow and show a deterioration of mechanical properties when exposed to light. Ternary combinations of HALS, UV absorbers, and phenolic antioxidants have a good stabilizing action, especially in foamed plastics [55].

The UV absorbers (e.g., benzotriazoles such as BT-UVA 1 and 2) and HALS are used to stabilize poly(methyl methacrylate) [56]; when colorlessness is important, oxalanilides (OA-UVA 1) are employed. Oxalanilides are also used in unsaturated polyesters and polycarbonates, but a satisfactory solution still has to be found to the problem of light-induced yellowing in polycarbonates.

3.6. Economic Aspects

Market volumes for light stabilizers are given in Table 5. The most important producers and trade names are listed in Table 3.

Roughly 50 % of the light stabilizers are sold to plastics processors in the form of master batches (concentrates containing $\leq 20\,\%$ light stabilizers in appropriate polymer vehicles). Combination master batches supplemented with pigments or flame retardants are being developed.

4. Heat Stabilizers for Poly(Vinyl Chloride)

4.1. Purpose and Requirements

Poly(vinyl chloride) (PVC) is thermally unstable. Heating results in elimination of HCl, the formation of polyene sequences, and rapid discoloration. This autocatalytic reaction begins at ca. 100 °C; at 180 °C a marked brown color occurs after a few minutes. Special heat stabilizers are therefore required for the processing and use of this high-tonnage plastic. These additives are rated according to several criteria besides their efficiency as heat stabilizers:

1) Versatility (use with mass, suspension, and emulsion PVC)
2) Effect on melt rheology of PVC
3) Lubricant action
4) Migration, plate-out
5) Compatibility with other additives and pigments
6) Effect on transparency
7) Effect on lightfastness
8) Effect on electrical insulation
9) Fogging (especially for automotive interior parts)
10) Occupational hygiene and approval as indirect food additives
11) Easy handling and costs

Combinations of stabilizers and costabilizers are normally used to obtain optimal effects. Many systems are employed in industry; PVC stabilization is discussed in detail in [57], [58].

4.2. Chemical Classes

Heat stabilizers for PVC are divided into metal-containing and metal-free stabilizers (costabilizers).

4.2.1. Metal-Containing Stabilizers

The strengths and weaknesses of the main metal-containing heat stabilizer classes are summarized in Table 6.

Lead stabilizers include basic lead carbonates (e.g., $x\,PbO \cdot PbCO_3$), sulfates, phosphites, stearates, maleates, phthalates, salicylates, and silicates, as well as neutral lead stearate $Pb(OOC{-}C_{17}H_{35})_2$. They have an excellent cost – performance ratio.

Mixed metal stabilizers include systems of barium – cadmium, barium – zinc, and calcium – zinc compounds, mainly carboxylates (e.g., stearates, laurates, oleates, or alkylbenzoates) naphthenates, or phenolates. The Ba – Cd types are most effective, followed by Ba – Zn, and Ca – Zn. They must always be reinforced with costabilizers. Developments include a zinc glycinate $H_2NCH_2COOZnOOCCH_3$ [59], which is also used in combination with barium or calcium carboxylates.

Table 5. Estimated market volume of light stabilizers in 1990 (10^3 t)

Region	BP-UVA	BT-UVA	HALS	Others	Total
United States	1.1	1.7	2.4	0.6	5.8
Western Europe	1.0	1.4	2.2	0.6	5.2
Japan	0.4	1.3	1.3	0.3	3.3
Other	0.4	0.9	1.1	0.3	2.7
Total	2.9	5.3	7.0	1.8	17.0

Table 6. Strengths and weaknesses of metal-containing heat stabilizers for PVC ∗

Property	Lead stabilizers	Mixed metal stabilizers		Tin mercaptides	Tin carboxylates
		Ba – Cd	Ba – Zn, Ca – Zn		
Versatility	–	–	–	+	+
Lubricant action	– (+ for stearates)	+	+	–	–
Migration/plate-out	+	–	+/–	+	+
Transparency of PVC	–	–	–	+	+
Light stability	+	+	–	–	+
Sulfur staining	–	–	+	–	+
Odor	+	–	+	–	+
Electrical insulating properties, water absorption	+	–	–	–	–
Toxicology, food approval in some cases	–	–	+	+	–

∗ + No problems, often beneficial effect; – Problematic

Organotin stabilizers are characterized by their tin content (4 – 25 %). They are classified as those containing sulfur (tin mercaptides, i.e., thiolates) and those containing none (tin carboxylates). The first group are the most efficient heat stabilizers. They can be employed without costabilizers and include mainly mono- and dialkyltin mercaptides, mercaptoacetates, mercaptopropionates, and thioglycolates with the following formula:

$$(R^1)_n Sn(SR^2)_{4-n}$$

where $n = 1$ or 2; R^1 = methyl, *n*-butyl, *n*-octyl, or *n*-dodecyl; and R^2 is, for example, $-CH_2COO$-alkyl, $-CH_2CH_2COO$-alkyl, or $-CH_2CH_2OOC$-alkyl. The alkyl group is often isooctyl.

"Methyltin stabilizers" (R^1 = methyl) offer better thermal stability than products containing higher alkyl groups, because the latter suffer thermal β-elimination. Cyclic dialkyltin sulfides are also employed, especially in multicomponent additive systems.

Most of the sulfur-free organotin stabilizers are dibutyltin or dioctyltin carboxylates, especially cyclic maleates, semiesters of maleic acid, or laurates.

Antimony stabilizers include antimony thiolates [e.g., $Sb(SCH_2COOR)_3$].

4.2.2. Metal-Free Stabilizers

These costabilizers always have to be used in conjunction with metal-containing stabilizers. The most important costabilizers follow.

Phosphites. Tris(nonylphenyl) phosphite and mixed aromatic – aliphatic phosphites, such as diphenyldecyl phosphite, are often used.

Epoxy Compounds. The principal epoxy stabilizers are epoxidized plant oils, chiefly epoxidized soybean oil.

Polyols include pentaerythritol, trismethylolpropane, and tris(hydroxyethyl) cyanurate.

Other costabilizers include β-aminocrotonates, dihydropyridines, and 1,3-diketones.

4.3. Mechanism of Action

The primary step in the thermal degradation of PVC is free-radical formation by the elimination of activated chlorine atoms, such as those on allyl groups or in tertiary positions. The chlorine atoms attack the polymer, resulting in hydrogen abstraction (formation of HCl) and double-bond formation [60].

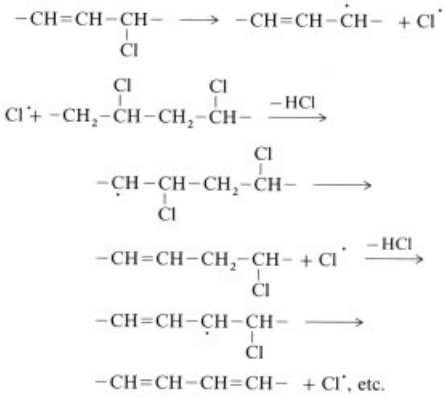

Double-bond formation occurs in preference to oxidation of the radical sites to peroxy radicals in the presence of oxygen. Elimination of HCl with the formation of polyene structures continues. Products containing $\geq 6-7$ conjugated double bonds are colored.

Ionic mechanisms have also been proposed, but these are less prominent at higher temperature. The formation of carbenium chlorides in the polymer backbone results in deepening of the color (" halochromy"). Thermooxidative processes accelerate degradation.

Stabilizers neutralize labile chlorine atoms by reacting with them to form more stable compounds:

The exchange reactions (Eq. 3 and 4) (Frye – Horst theory [61]) form the basis for many mechanistic analyses. The metal stabilizers are finally converted to chlorides: R_2SnCl_2, $ZnCl_2$, $CdCl_2$, and $PbCl_2$. The metal stabilizers also react with and neutralize hydrogen chloride, which would catalyze further PVC degradation.

The metal chlorides $CdCl_2$ and especially $ZnCl_2$ are, however, Lewis acids and have a destabilizing influence because they catalyze further elimination of HCl. Consequently, although initial stabilization is excellent, further heating leads to a sudden deep coloration of the PVC (" zinc burning"). The remedy is to exploit the synergism of mixed metal stabilizers. Cadmium and zinc carboxylates are always used in combination with barium or calcium carboxylates, which exchange anions with zinc chloride and cadmium chloride. The zinc and cadmium stabilizers are therefore regenerated, with the formation of barium and calcium chlorides which do not enhance PVC degradation. Barium and calcium carboxylates also react preferentially with HCl.

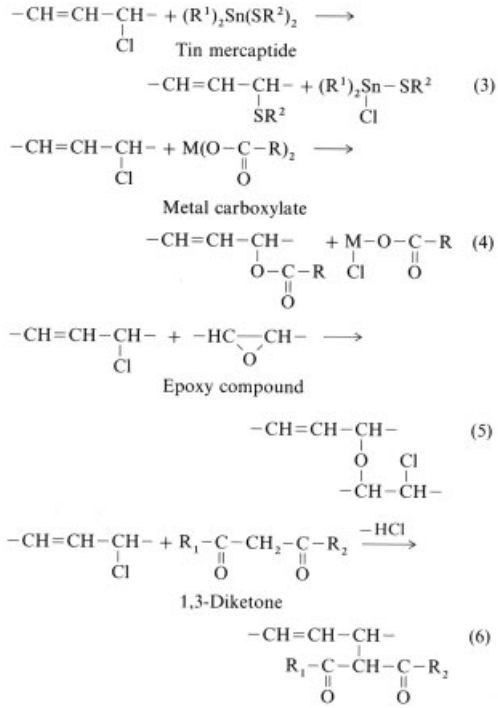

No catalytic effect on degradation is observed with lead chloride and (di)alkyltin chlorides; the alkyl groups in the latter compounds have a deactivating action. For the effects of metal chlorides, see [62], [63]; for the mechanism of organotin stabilizers, see [64].

Costabilizers, such as polyols [63] and epoxy compounds, also neutralize harmful metal chlorides and improve long-term stabilization with calcium – zinc stabilizers. β-Diketones [65], [66] and dihydropyridines, in contrast, improve the initial color stabilization; the dihydropyridines hydrogenate polyene sequences

and thereby shorten their length. Tin mercaptides and the thiols formed by their reaction with HCl can also react with polyene double bonds. The double bond in tin maleates can react with polyenes (Diels – Alder addition):

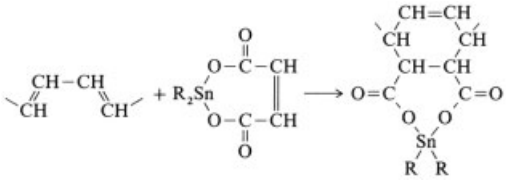

Finally, alkyltin mercaptides and phosphite costabilizers act as hydroperoxide decomposers, thus protecting against oxidative degradation.

4.4. Test Methods

The most important test methods are based on static and dynamic heating. Analytical methods such as the determination of dehydrochlorination at 180 – 200 °C are also common.

In the *static heating test*, films produced in a two-roll mill or plastisols applied to a substrate are heated in a furnace (e.g., to 180 – 200 °C), and the specimen color is observed as a function of time. In the *dynamic test*, PVC is treated in a two-roll mill, Brabender plastograph, or extruder at specified temperatures. Samples are withdrawn after various times or numbers of extrusion passes and pressed into films, whose color is then determined (usually via the yellowness index).

Because color development in PVC is fairly rapid, these tests (i.e., heat treatment) seldom require more than a few hours. The profile of the yellowness index curves reflects the effects of stabilizers, particularly with regard to short-term and long-term stabilization, and improvements due to costabilizers.

The effect of heat stabilizers on the processing and service qualities of PVC is the concern of extensive testing programs [67].

4.5. Uses

The choice of stabilizer system depends on the processing methods used (extrusion, injection molding, calendering, plastisol application), the intended uses of the PVC, and compatibility with other additives. Technologies and regulations vary greatly. The following survey can therefore give only general indications and trends. For further details, see the producers' literature.

Rigid PVC. Solid stabilizers or stabilizer blends are generally preferred for rigid PVC. Organotin stabilizers have found wide use because processing conditions are more severe than those for flexible PVC.

Window Frames and Sidings. Tin mercaptides are used in the United States; in Europe a shift from Ba–Cd stabilizers has been observed into Ca–Zn costabilizer systems via Ba–Cd–Pd and Pb formulas. Tin carboxylates are also used, especially for transparent or colored frames, because of their good weather resistance.

Pipes, Including Water-Supply Pipes. Tin mercaptides are used in the United States. Lead stabilizers predominate in Europe (except France) and Japan.

Beverage Bottles and Bottle Caps. The PVC is stabilized with food-approved tin mercaptides or Ca–Zn costabilizer systems containing epoxidized soybean oil, 1,3-diketones, or polyols as costabilizers.

Flexible PVC. Most of the stabilizers used for flexible (plasticized) PVC are in liquid form; Ba–Cd and Ba–Zn systems occupy the leading position, with a trend toward cadmium-free systems including Ca–Zn products, e.g., in foils, films, leather imitates (shoes, automotive interiors), hoses, and floor coverings. Lead stabilizers are used in electrical cables.

Stabilizer Content. Heat stabilizer contents of 0.3 – 5 % are common. Levels of organotin stabilizers are at the low end of the range ($\leq$ 2.5 %), mixed metal stabilizers in the middle, and lead stabilizers for cable insulation at the high end. Phosphite or 1,3-diketone costabilizers are generally used below 0.5 %, whereas epoxidized soybean oil is added in concentrations of 5 % or higher.

Legal Aspects. Some countries place strict limits on the use of certain stabilizer classes; for example Sweden and Japan restrict the use of cadmium stabilizers. Similar provisions are under discussion for the EC. Other broad limits

apply to lead stabilizers in drinking-water pipes in France and the United States.

Stabilizers for products that come in contact with food (e.g., beverage bottles, packaging films) must be officially approved as indirect food additives. Many octyltin and dodecyltin mercaptides [e.g., di-*n*-octyltin bis(isooctylthioglycolate)], most of the Ca–Zn stabilizers, and many costabilizers (especially stearylbenzoylmethane, dihydropyridine and epoxidized oils) have such approval.

4.6. Economic Aspects

Table 7 lists estimated 1990 market volumes for metal-based heat stabilizers that are used as solids, pastes, or liquids. Dust-free solid forms such as granules, flakes, or "small bags" are preferred. "One-pack" systems, which contain primary and secondary stabilizers preblended with lubricants and other additives, are widely used. Most producers offer broad assortments, with individual products that are continually being modified and tailored to specific applications. Among the nonmetallic costabilizers, epoxidized soybean oil holds a dominant position with more than 100 000 t/a used worldwide in PVC.

Table 7. Estimated market volume of PVC heat stabilizers in 1990 (10^3 t)

Region	Tin	Lead	Mixed metal	Total
United States	20	10	25	55
Western Europe	10	50	40	100
Japan	10	30	15	55
Other	10	30	15	55
Total	50	120	95	265

Table 8 lists the most important producers and trade names of PVC heat stabilizers.

Table 8. Producers of PVC heat stabilizers

Producer	Trade name	Type ∗
Acima	Metastab	Sn, Me
Adeka Argus	Mark	Me, Pb, C, Sn
Akishima Chemicals		Me, Sn
Akzo	Interstab, Stanclere, Estabex	Me, Sn, Pb, C
BASF	Sicostab	Me
Bleiberger Bergwerks Union	Austrostab	Pb, Sn, Me, C
Cardinal Chemicals	CC	Sn
Chemson	Naftovin, Strandex	Pb, Me, Sn
Ciba-Geigy	Irgastab, Reoplast	Sn, Me, C
Commer	Prosper	Sn, Me, Pb, C
CWM O. Bärlocher	Bärostab, Okstan	Me, Sn, Pb, C
Ferro Corporation	Therm Chek	Me, Sn, C
Hammond Lead Products	Halstab	Pb
Hebron	Vinstab	Sn, Me, Pb, C
Henkel	Edenol, Stabilox, Stabiol	C
Hoechst	Hostastab	Sn, C
Katsuta Kako	Advastab	Me, C, Sn
Kemichrom	Kemistab	Sn, Me, Pb, C
Kyodo Chemicals		Sn, Me
Lagor	Lastab, Synesol	Me, Sn, C
Meister	Meister	Sn, Me, C
Morton Thiokol	Advastab	Sn
M & T Chemicals	Thermolite	Sn, Me, Pb, C
Nissan Ferro Organic Chem.	Therm Chek	Me
Nitto Kasei		Sn
Plastics Specialities & Technologies		Me
Reagens	Reatinor, Reablend	Sn, Me, Pb, C
Rhône-Poulenc	Rhodiastab	C
Sakai Kagaku		Me, C, Sn
Sankyo Organic Gosei	Stann	Sn
Swedstab	Swedstab	Sn, Me, C
Synthetic Products	Synpron	Sn, Me
Tinstab SA	Tinstab	Sn, Me, C
Tokyo Fine Chemicals	Embilizer	Sn
R. T. Vanderbilt	Vanstray, Barostab	Me
Witco Corporation	Mark	Me, Sn

∗ C=costabilizer; Me=mixed metal; Pb=lead; Sn=organotin.

5. Lubricants, Slip, Antiblocking, and Mold-Release Agents

5.1. Purpose and Requirements [68], [69]

Lubricants (→ Lubricants and Related Products), slip, antiblocking, and mold-release agents (→ Release Agents) control the frictional and adhesive properties of plastics during processing and in service.

Lubricants also improve the dispersion of pigments and fillers in plastics. Uniform colors can be obtained by breaking up agglomerates in pigment preparations. Better dispersion of fillers (e.g., talc) improves flow limits and material properties.

Internal lubricants reduce friction between polymer particles and molecules during the melting (plastication) of plastics and transport

of the melt. They thus reduce energy consumption on plastication, lower melt viscosity, improve flow properties, increase output of processing machinery, and allow processing under less stringent conditions.

External lubricants reduce the friction and adhesion of polymer melts on hot metal surfaces of processing machinery (e.g., extruder screws and cylinders, rolls). This reduces abrasion between the polymer melt and the metal, and improves melt flow. It also improves the gloss, smoothness, and regularity of the surface of the plastic.

Mold-release agents prevent the adhesion of molded plastics to metal cooling rolls and molds. *Slip agents* improve the frictional qualities of finished plastic surfaces (e.g., bearings and gears) and prevent adhesion between stacked films. *Antiblocking agents* prevent adhesion between film layers on rolls or in stacks when pressure and heat are applied.

Antislip agents (adhesion promoters) impart limited tack to finished articles to reduce slip (e.g., between filled bags made of plastic films).

It is not always possible to assign a particular product to one of the above classes. For example, the boundary between internal and external lubricants depends on the type of polymer, the additive concentration, the presence of other lubricants, and the processing temperature. Many products also perform multiple functions: external lubricants are often effective as mold-release or slip agents. Carefully matched combinations of additives are often required. This group of additives must also comply with standards of odor and taste, migration behavior, product stabilization [70], and thermal stability.

Lubricants and related compounds may have disadvantageous effects on some product properties. Internal lubricants can act as plasticizers and thus reduce the heat distortion temperature. External lubricants can reduce transparency and impair the printability, cementability, and weldability of plastics. These additives also affect mechanical properties, especially in case of overlubrication, which causes inhomogeneities due to inadequate fusion of the material.

5.2. Chemical Classes

Many lubricants are amphoteric compounds containing both a long hydrocarbon chain and polar groups (e.g., hydroxyl, ester, or acid groups). Lubricants are characterized by their molecular mass, acid or hydroxyl number, saponification number, metal content, melting or dropping point, cold-setting temperature, thermal stability, volatility, density, viscosity, and (for solid agents) particle size. The action of a lubricant depends on its polarity. Polar molecules act as internal lubricants in polar polymers such as PVC and as external lubricants in nonpolar polymers such as polyolefins, and vice versa. The polarity is strongly influenced by the length of the hydrocarbon chain. Lubricants can be classified as fatty acids and their derivatives ($C_{12}-C_{22}$ chains), montanic acids and their derivatives ($C_{28}-C_{32}$), paraffins ($C_{20}-C_{70}$), and polyolefin waxes (molecular mass 2000 – 10 000).

The substances of greatest practical importance are listed below; for further details, see → Lubricants and Related Products.

Fatty acids, fatty alcohols, and their derivatives comprise the following groups:

1) Stearic, 12-hydroxystearic, palmitic, and behenic acids
2) Cetyl and stearyl alcohols
3) Metal stearates (Zn, Ca, Pb, Al, Mg, and Na)
4) Butyl stearate, tridecyl stearate, glycerol monostearate (GMS), glycerol monoricinoleate, glycerol monooleate, glycerol tristearate, glycerol tri-12-hydroxystearate (hydrogenated castor oil), trimethylolpropane tristearate, and pentaerythritol tetrastearate
5) Distearyl adipate, distearyl phthalate, cetyl palmitate, cetyl stearate, stearyl stearate, and behenyl behenate
6) Isostearamide, oleamide, erucamide, ethylene bisstearylamide (EDS or "amide wax")
7) Oligomeric fatty acid esters ("complex fatty acid esters") are also used. These are obtained by polycondensation of dicarboxylic acids (e.g., adipic acid), polyols (e.g., pentaerythritol), and monocarboxylic acids (e.g., stearic acid). They may be partially saponified or may contain calcium.

Montanic Acids and their Derivatives (Montan Waxes). Montanic acids are obtained from

lignite and consist of mixtures of $C_{28}-C_{32}$ acids. Derivatives include calcium montanates, sodium montanates, and esters with ethylene glycol, butanediol, glycerol, or pentaerythritol.

Paraffins may be derived from petroleum (straight-chain members are solids, branched-chain members are liquids) or are synthetic (Fischer – Tropsch) paraffins.

Polyolefin Waxes. The most common polyolefin waxes are polyethylene waxes with branched and linear structures; partially crystalline polypropylene waxes are less common. Polar polyethylene waxes are obtained by oxidation; they have acid numbers of ca. 20 – 70.

Other Chemical Classes. Fluorinated polymers, chlorinated paraffins, silicones, graphites, and other substances are used as lubricants and related compounds. Silica gel, chalk, and other minerals are employed as antiblocking agents and slip aids.

This article does not cover externally applied mold-release agents (e.g., waxes, fluorinated polymers, silicones, and metallic soaps), which are sprayed or brushed onto mold surfaces. These methods are used chiefly with thermosetting resins (e.g., polyurethane, unsaturated polyester, epoxy).

5.3. Mechanism of Action

The friction-lowering property of *lubricants* is due to their long hydrocarbon chains ($> C_{12}$). The lubricating effects depend on interactions between the lubricant and the polymer matrix (solubility of the lubricant in the polymer) and on the association between the lubricant and the metal surfaces. Unlike plasticizers, which are completely soluble in the polymer, lubricants should have only limited solubility; this is ensured by their amphoteric character.

Internal lubricants [71] have a relatively high affinity for polymers. Polar lubricants in, for example, PVC are either dissolved in or strongly bound to the surface of the associations of polymer molecules that occur in the plastic state. Typical effects are a lowering of the softening temperature due to swelling, a reduction in melt viscosity, and frictional damping between polymer particles. Good internal lubricants have high saturation limits and haze points.

External lubricants are strongly associated to the metal surfaces of processing equipment or form discrete phases within the polymer melt. They only have a low affinity for the polymers. Typical effects are a reduction in adherence to the metal and in friction between polymer particles, which leads to slower plastication. External lubricants often cause haze. Overlubrication must be avoided because some adhesion to the metal is necessary for conveyance of the melt in the extruder and also in calendering.

Mold-release agents form a film between the metal surface and the cooled plastic. *Slip agents* migrate to the plastic surface, where their hydrocarbon chains act as a lubricant. *Antiblocking agents* slightly roughen the film surface so that air is trapped between stacked layers. *Antislip agents* artificially create surface roughness where nonlubricating particles incorporated in the melt protrude from the film.

5.4. Test Methods

Methods based on practical applications are applied to lubricant testing. One of the most important is the plastograph (e.g., the Brabender Plasticorder). The formulated plastic powder is placed in a kneading chamber at a preset temperature (e.g., 140 – 160 °C) and plasticated by a twin-shaft device running at constant speed. The torque is plotted as a function of time (Fig. 6). The curve passes through a maximum torque value T_{max}, marking the onset of plastication in the powder mixture. It then declines to a minimum T_{min}, followed by an increase to a second maximum T_e as a result of growing friction. Plastication is complete at T_e. The time between T_{max} and T_e is called the plastication time t_g. The test is used to establish suitable ratios of internal and external lubricants. For example, a shorter t_g or an increase in the ratio $(T_{max} - T_{min})/t_g$ indicates internal lubrication. As the hydrocarbon chain in the lubricant becomes longer, the maximum kneading resistance declines.

Other tests employ extrusiometers (measurement of polymer melt temperature and pressure in the extruder cylinder and at the die), capillary viscometers (quantity of material discharged per unit time at constant pressure and temperature) [72], or roller mills (time that elapses between

charging and adhesion of the polymer sheet on the rolls). In the spiral test, material is injected into a spiral mold and the distance the material travels in the runner is measured.

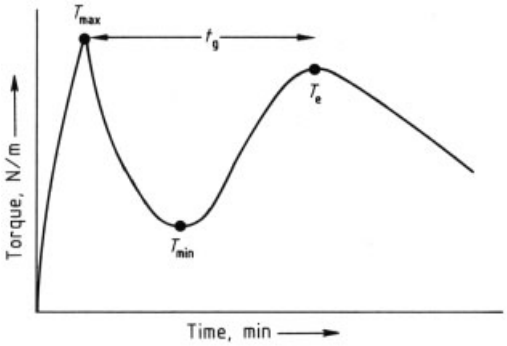

Figure 6. Schematic plastogram of PVC with lubricant (Brabender Plasticorder)

Slip and blocking effects are studied in skid and adhesion tests with films (e.g., ASTM D 1893 or D 3354). Haze and surface gloss measurements are an important component of all lubricant tests.

5.5. Uses

The main application of lubricants is in PVC (ca. 50 %), followed by styrene polymers, polyolefins, engineering thermoplastics, and thermoplastic elastomers.

Rigid PVC always requires lubricants for processing (usual content 1 – 4 %), and lubricant combinations (preformulated systems) are often employed. Selection of lubricants is dictated by the type of heat stabilizer used: stearic acid with lead stabilizers; 12-hydroxystearic acid with barium – cadmium stabilizers (to reduce plate-out); glycerol monostearate, montan waxes, or oxidized polyethylene waxes with tin stabilizers; and complex fatty acid esters with calcium – zinc stabilizers.

Fatty acid esters, calcium stearate, and fatty alcohols are typical internal lubricants. Montan waxes, amide waxes, paraffins, oxidized polyethylene waxes, and fatty acids act as external lubricants. Heavy-metal soaps and complex fatty acid esters fall between the two types. Lubricants allow processing under gentler conditions and thereby contribute to stabilization. A combination of calcium stearate as internal lubricant, paraffin wax as external lubricant, and oxidized polyethylene wax as mold-release agent has been used with success for PVC pipes.

Flexible PVC requires less lubricant (ca. 0.5 %). Liquid types,such as glycerol monooleate or paraffin oils, are used. Amide waxes serve as antiblocking agents.

The flow behavior of *ABS, SAN, and polystyrene* can be improved with butyl stearate, glycerol monostearate, fatty acid esters, paraffin oils, oxidized polyethylene waxes, and montan waxes. Zinc stearate and amide wax (0.1 – 2 %) are used as mold-release agents.

The calcium stearate used as an acid radical scavenger is often sufficient for *polyolefins*. Neutral polyethylene and polypropylene waxes are added to pigment preparations for polyethylene and polypropylene. Talc-filled polypropylene can contain ca. 0.5 % of partially saponified montan wax. A suitable mold-release and slip agent for films is 0.1 – 0.3 % of a fatty acid amide in polyethylene or of erucamide in polypropylene. Finally, 0.05 % of fluorinated polymers can be added to LLDPE films to combat melt fracture (a surface defect occurring at high viscosity and resulting from the alternating wall adhesion and slip of the extrudate). Highly effective slip aids for polyolefin films include micronized synthetic silica gels.

For *engineering thermoplastics* [polycarbonates, polyamides, poly(ethylene terephthalate) (PETP), poly(butylene terephthalate) (PBT)], thermally stable pentaerythritol esters of fatty acids are preferred; they have a better transparency than montanate esters. Amide wax can also serve as a mold-release agent. Filled thermoplastics represent a broad field for lubricants.

Because most lubricants are derived from natural sources, they are often approved for food packaging products.

5.6. Economic Aspects

Worldwide consumption (ca. 1990) of lubricants and related compounds is ca. 130 000 – 140 000 t, with Western Europe, the United States, and the rest of the world accounting for a third each. Consumption is equally divided between the five following groups: fatty acids and esters, metallic soaps, amide waxes, paraffin and polyethylene waxes, and miscellaneous types.

Table 9. Producers and trade names of lubricants and related compounds

Manufacturer	Trade name
Abril Industrial Waxes	Abril Wax
Akzo Chemicals	Dopral, Interstab, Armoslip, Aramid
Allied Signal	AC Polyethylene
Ashland Chem.	
Astor Wax	Comboloob
Axel Plastics Res. Labs.	Mold Wiz
CWM O. Bärlocher	Bärolub
BASF	Sicolub, Wachs BASF
Dr. T. Böhme	Tebestat
Chemson	Naftolube, Naftozin, Listab
Ciba-Geigy	Irgawax
Corriel	Lubriol, Synthewax, Glycmonos
Commer	Vinlub
Croda Universal	Crodamid, Crodacid
Deer Polymer	
Du Pont	Zonyl, Viton
Durham Chemicals	Durolube
Eastman	Epolene, Myverol, Myvaplex
Th. Goldschmidt	Tego, Tegin
Grace	Sylobloc
C. P. Hall	
Harcros Chemicals	Harochem, Harowax, Harogel, Lankroplast
Harwick Chemicals	Lubrox, Lubrex
Henkel	Loxiol, Loxamid
Hoechst	Hostalub, Hostamont, Hoechst Wachse
Hüls	Vestowax
Humko Chemicals (division of Witco)	Kemamide, Kemester, Mold Pro, Micro-ken
ICI	Atmer
Kao Corporation	
Kemichrom	Kemfluid
Lion Akzo	
Lonza	Glycolube
3 M	Dynamar
Marabushi Oil Chemicals	
Mathe Co.	Coad
Mazer Chemicals (division of PPG)	
Morton Thiokol	Advawax
M & T Chemicals	Stavinor
New Japan Chemical Co.	
Nippon Oil & Fats	
Nitto Kasei	
Reagens	Realube
Rohm & Haas	
Rousselot	
Sakai	
Speciality Products	Zn-Stearate
Struktol Co.	Polydis TR
Swedstab	Swedlub, Swedstab
Tonnan Kagaku Kogyo	
Unichema Chemie	Unislip, Uniwax, Estol, Priolube, Prifrac, Pristerene

Table 9 lists some important producers and trade names of lubricants and related compounds. The number of products on the market is relatively large because new formulations and easy-to-handle forms are continually appearing. Most products for PVC are sold to converters; those for other polymers are often sold to the polymer producers.

6. Flame Retardants [73–78] See also → Flame Retardants.

6.1. Purpose and Requirements

The replacement of classical materials by combustible plastics in buildings, equipment, and vehicles creates a fire risk. Flame retardants are used to combat this danger. A useful measure of flammability for plastics is the limiting oxygen index (LOI), which denotes the volume fraction of oxygen in an oxygen – nitrogen mixture that just supports combustion of a well-defined specimen under standard conditions. A high LOI indicates a low flammability. The LOI values for the most important plastics follow:

Polyoxymethylene	15.5
Polyethylene	17.3
Poly(methyl methacrylate)	17.5
Polypropylene	17.6
Polyurethane	17 – 19
Polystyrene	18.0
Unsaturated polyester	18.5
ABS	18 – 19
Epoxy resin	19.0
PBT	21.5
Polyamide 12	21.6
PETP	22.0
Polycarbonate	23 – 25
Polyamide 6	23 – 26
PVC, flexible	24 – 25
Poly(phenylene oxide) – polystyrene	24 – 25
Polyamide 66	24 – 26
PVC, rigid	40 – 45
Polytetrafluoroethylene	95.0

Flame retardants in plastics affect ignitability, combustion rate, heat release, fuel contribution, smoke evolution, and the formation of toxic or corrosive gases. The significance of these parameters depends on the situation. Although flame retardants can be especially effective in the early stages of a fire and can reduce flammability, plastics modified with these additives are still combustible.

In many cases and especially at high concentrations, flame retardants lower the quality and

increase the price of a plastic. They are therefore used only when mandated. Selection is determined by the effectiveness and price of the flame retardant, as well as the processibility of the plastic, and influences on its mechanical,optical,and electrical properties.Interactions with other additives (e.g., light stabilizers) must also be considered.

6.2. Chemical Classes

Flame retardants may be inorganic substances, halogenated organic compounds, organophosphorus compounds, or other organic substances. They are further classified as additive or reactive. Reactive flame retardants are incorporated into the polymer molecule by copolymerization, cross-linking, or grafting.

Inorganic Flame Retardants. Producers and trade names are summarized in Table 10.

Aluminum trihydroxide, $Al(OH)_3$, and *magnesium dihydroxide,* $Mg(OH)_2$, [79] are the largest group of flame retardants in terms of consumption. Particle size varies from 1 to 30 µm. High concentrations (40 – 60 wt %) of these additives are required and are associated with a severe impairment of mechanical properties. To overcome this, products coated with coupling agents (e.g., silanes, fatty acid esters) are recommended. The coating may, for example, improve the impact strength of the modified plastic or reduce the viscosity of prepolymers. Aluminum and magnesium hydroxides often suppress smoke evolution. Use is limited by their thermal stability (dehydration); $Al(OH)_3$ decomposes above ca. 200 °C, and $Mg(OH)_2$ above 300 – 340 °C.

Antimony Compounds. Antimony trioxide, Sb_2O_3; antimony pentoxide, Sb_2O_5; and sodium antimonates (e.g., $Na_2O \cdot Sb_2O_5 \cdot x\,H_2O$) are largely used as synergists with halogenated organic flame retardants or in PVC. In the absence of halogen compounds they have little effect. Antimony trioxide is sold in various purities, whitenesses, and grain sizes (mainly 1 – 3 µm, down to <1 µm for special grades). Antimony pentoxide is available in colloidal form (0.03 µm). Because of occupational health issues relating to antimony trioxide dust, nondusting preparations are preferred (i.e., concentrated master batches containing ca. 80 % of active agent, granules, or wetted grades).

Zinc borates [80] often serve as a partial replacement for antimony trioxide, giving some advantages in suppression of smoke evolution, afterglow, weaker pigmentation, and lower cost. Combinations with aluminum trihydroxide are also employed. Negative effects on the stability of halogen-containing plastics must be considered.

Ammonium polyphosphate, $(NH_4PO_3)_n$, is included in intumescent systems [81], [82] that contain polyols and melamine derivatives (→ Flame Retardants). The products are also marketed in coated or microencapsulated form to reduce water absorption, for example.

Microencapsulated or inertized *red phosphorus* is a very effective flame retardant that can be processed at up to 280 °C [82],[83]. Its drawbacks are a dark color and potential handling hazards.

Other inorganic flame retardants include iron oxides (Fe_2O_3 and Fe_3O_4); hydrated calcium and magnesium carbonates; bismuth oxides and carbonates [84]; boric acid and borates; zinc hydroxystannate [$ZnSn(OH)_6$], zinc stannate ($ZnSnO_3$), and other zinc compounds [85]; and mineral substances that decompose endothermically at certain temperatures to release water or noncombustible gases. Smoke-inhibiting additives for PVC, unsaturated polyesters, or polyurethanes include calcium carbonate, tin oxides, copper oxide, molybdenum trioxide, and ammonium molybdate.

Halogenated Organic Flame Retardants. Halogenated compounds are effective flame retardants, especially in conjunction with antimony trioxide. They can therefore be used in low concentration so that they do not greatly affect product properties. A drawback is the fact that corrosive gases (HBr, HCl) are always produced in the event of fire and can cause secondary damage. Light stability of the plastic is also reduced, especially by brominated compounds.

The most important members of this group are brominated aromatics, above all decabromodiphenyl oxide and polymeric compounds. Some important producers and trade names are listed in Table 11. Aliphatic bromine compounds are often more effective but less thermally stable. This disadvantage is partially off-

Table 10. Producers and trade names of inorganic flame retardants

Compound	CAS registry no.	Producer (trade name)
Aluminum trihydroxide, $Al(OH)_3$	[*21645-51-2*]	Aluchem; BA Chemicals, Alcan Aluminium (Baco Superfine, Ultrafine); Alcoa (Hydral, Flameguard); Atochem; Kaiser Aluminium; Martinswerke, Lonza (Martinal); Reynolds Metals; Solem Industries, J. M. Huber (Micral, Zerogen); Vereinigte Aluminium Werke (Apyral); Showa Aluminium Industries; Sumitomo Aluminium Smelting
Magnesium dihydroxide, $Mg(OH)_2$	[*1309-48-4*]	Dead Sea Bromine (FR-20); Kyowa Chemicals; Lonza (Magnifin); Morton Thiokol (Versamag); Solem Industries, J. M. Huber (Zerogen); Veitscher Magnesitwerke
Antimony trioxide, Sb_2O_3 (some as concentrate or in wetted form)	[*1309-64-4*]	Anzon (Firebloc, Envirostrand, Oncor); Asarco; Laurel Industries (Fire Shield); M & T Chemicals (Thermoguard); Amspec Chemical; BA Chemicals (Timonox)
Antimony pentoxide, Sb_2O_5	[*1314-60-9*]	Nyacol Products (Nyacol)
Sodium antimonates, e.g., $Na_2Sb_2O_6 \cdot x\,H_2O$		M & T Chemicals (Thermoguard)
Zinc borates, e.g., $4\,ZnO \cdot 6\,B_2O_3 \cdot 7\,H_2O$	[*12536-65-1*]	Borax Holdings; Climax Performance Materials (ZB-223); US Borax & Chemicals (Firebrake ZB)
Barium metaborate, e.g., $BaB_2O_4 \cdot x\,H_2O$		Buckman Laboratories (Busan)
Ammonium polyphosphate, $(NH_4PO_3)_n$ (some coated or encapsulated, some in systems with flame retardants containing N)	[*68333-79-9*]	Albright & Wilson (Amgard MC); Hoechst (Exolit); Monsanto (Phos-Chek); Great Lakes (Char Guard 329, CN 1197)
Red phosphorus	[*7723-14-0*]	Albright & Wilson (Amgard CRP); Hoechst (Exolit 405)

set in brominated neopentyl structures that do not have hydrogen atoms in the α-position to the bromine. Hexabromocyclododecane is an example of a brominated cycloaliphatic compound; tetrabromobisphenol A is the most important reactive type.

Chlorinated flame retardants (Table 12) are only about half as effective as brominated ones because of the higher bond energy of chlorine. Chlorinated paraffins and chlorinated cycloaliphatics are mainly used [86]. Fluorine and iodine compounds are not employed as flame retardants.

Organophosphorus Flame Retardants. Important products and trade names are listed in Table 13.

The most important organophosphorus flame retardants are halogen-free triaryl phosphates. Phosphonates (e.g., dimethyl methylphosphonate), phosphine oxides, and halogenated phosphate esters and phosphonates are also used. Advantages include low corrosivity of the combustion gases, lack of effect on polymer transparency, and suppression of afterglow. Drawbacks include volatility, sensitivity to hydrolysis, and negative effects on the heat distortion temperatures of plastics. Cyclic organophosphorus compounds (phosphorinanes, phosphazenes) overcome some of these problems.

Miscellaneous Organic Compounds. Organic additives occasionally used in flame retardant systems or for special types of plastics include the following:

1) Polyols (e.g., pentaerythritol) for intumescent systems [81]
2) Nitrogen-containing compounds (e.g., melamine or melamine cyanurate), often combined with $(NH_4PO_3)_n$
3) Silicones [87] as smoke inhibitors and impact modifiers when high concentrations of $Al(OH)_3$, $Mg(OH)_2$, or $(NH_4PO_3)_n$ are used as flame retardants
4) Free-radical formers (e.g., dicumyl peroxide or 2,3-dimethyl-2,3-diphenylbutane) together with brominated compounds, especially in expandable polystyrene [88]
5) Alkali-metal salts of organic sulfonates for polycarbonates [89]
6) Ferrocene as a smoke inhibitor in PVC

6.3. Mechanism of Action

Three elements are essential for a fire: heat, fuel, and air (oxygen) (Fig. 7). When heated, plastics undergo pyrolysis, forming characteristic degradation products, some of which appear as smoke whereas others serve as fuels that can be ignited

Table 11. Producers and trade names of brominated organic flame retardants

Structure	Name, CAS registry no.	Producer (trade name)
; $x + y = 10$	decabromodiphenyl oxide [*1163-19-5*]	Dead Sea Bromine (FR 1210) Ethyl Corporation (Saytex 102) Great Lakes Chem. (DE-83)
; $x + y = 8$	octabromodiphenyl oxide [*32536-52-0*]	Dead Sea Bromine (FR 1208) Ethyl Corporation (Saytex 111) Great Lakes Chem. (DE-79)
; $x + y = 5$	pentabromodiphenyl oxide	Dead Sea Bromine (FR 1205) Ethyl Corporation (Saytex 115) Great Lakes Chem. (DE-71)
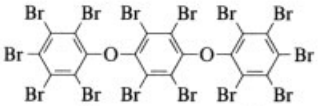	tetradecabromodiphenoxybenzene [*58965-66-5*]	Ethyl Corporation (Saytex 120)
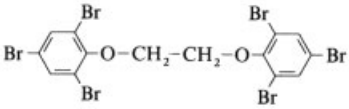	bis(tribromophenoxy)ethane [*37853-59-1*]	Great Lakes Chem. (FF 680)
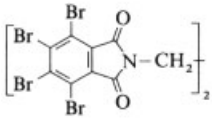	ethylene-bis(tetrabromophthalimide) [*32588-76-4*]	Ethyl Corporation (Saytex BT 93)
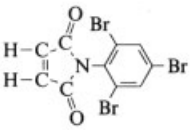	*N*-(2,4,6-tribromophenyl)maleinimide [*59789-51-4*]	Dead Sea Bromine (FR 1033)
	pentabromotoluene [*87-83-2*]	Dead Sea Bromine (FR 705)
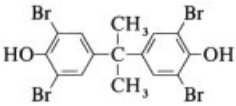	tetrabromobisphenol A [*79-94-7*]	Dead Sea Bromine (FR 1524) Ethyl Corporation (Saytex RB 100) Great Lakes Chem. (BA 59)
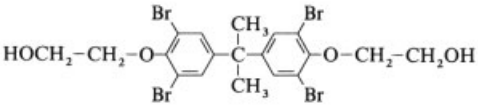	tetrabromobisphenol A-bis(2-hydroxyethyl ether) [*4162-45-2*]	Dead Sea Bromine (FR 1525) Great Lakes Chem. (BA 50)
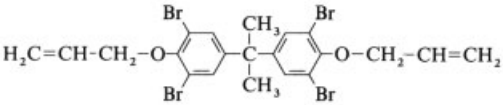	tetrabromobisphenol A-bis(allyl ether) [*25327-89-3*]	Dead Sea Bromine (FR 2124) Great Lakes Chem. (BE 51)

Table 11. (Continued)

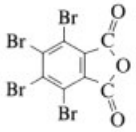	tetrabromophthalic anhydride [*632-79-1*]	Ethyl Corporation (Saytex RB 49) Great Lakes Chem. (PHT 4)
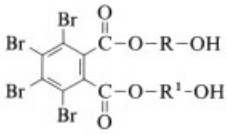	mixed esters of tetrabromophthalic acid [*77098-07-8*] with diethylene glycol and propylene glycol [*75790-69-1*]	Ethyl Corporation (Saytex RB 79) Great Lakes Chem. (PHT 4-diol)
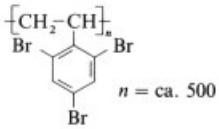 n = ca. 500	poly(2,4,6-tribromostyrene) [*57137-10-7*]	Ferro Corporation (Pyro-chek 68, Pyro-chek LM)
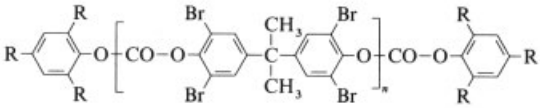 $n = 3-5$; R = H or Br	poly(tetrabromobisphenol-A)-polycarbonate [*28906-13-0*], [*71342-77-3*]	Great Lakes Chem. (BC 52, 58)
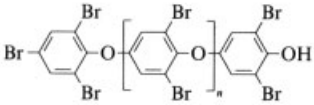	poly(dibromophenylene oxide) [*26023-27-8*]	Great Lakes Chem. (PO 64)
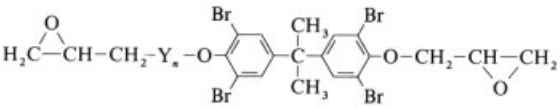 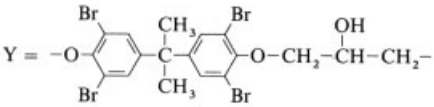 $n = 1$ to >50	brominated epoxy resins	Makhteshim Chemical Works (F 2000 Series) M & T Chemicals (Thermoguard)
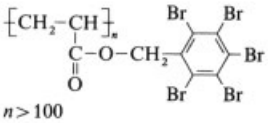 $n > 100$	poly(pentabromobenzyl acrylate) [*59447-57-3*]	Dead Sea Bromine (FR 1025)
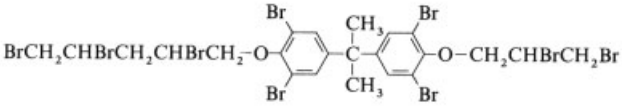	tetrabromobisphenol A-bis-(2,3-dibromopropyl ether) [*21850-44-2*]	Great Lakes Chem. (PE 68) Riedel de Haen (OBPE)
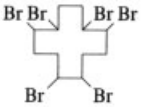	hexabromocyclododecane [*3194-55-6*]	Dead Sea Bromine (FR 1206) Ethyl Corporation (Saytex HBCD) Great Lakes Chem. (CD 75-P) Riedel de Haen (HBCD)

Table 11. (Continued)

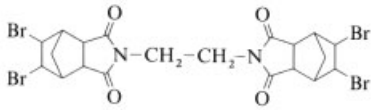	ethylene-bis(dibromonorbornane dicarboximide) [*41291-34-3*]	Ethyl Corporation (Saytex BN 451)
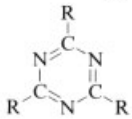 R = $OCH_2CHBrCH_2Br$	1,3,5-tris(2,3-dibromopropoxy)-2,4,6-triazine [*52434-59-0*]	Akzo Chem. (Interstab FR 930)
$(HOCH_2)_2C(CH_2Br)_2$	2,2-bis(bromomethyl)-1,3-propanediol [*3296-90-0*]	Dead Sea Bromine (FR 522) Ethyl Corporation (Saytex FR 1138)
$HOCH_2-C(CH_2Br)_3$	2,2-bis(bromomethyl)-3-bromo-1-propanol [*1522-92-5*]	Dead Sea Bromine (FR 513)
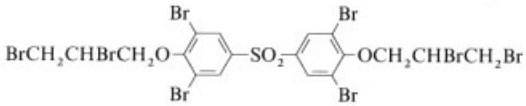	3,5,3′,5′-tetrabromo-4,4′-dihydroxydiphenylsulfone-bis-(2,3-dibromopropyl ether) [*42757-55-1*]	Marubeni (Non Nen 52)
Brominated polyols, some also containing chlorine		Dead Sea Bromine Solvay (IXOL B 251)

once they are mixed with air. Oxidative combustion reactions in the gas phase (the "fire") are strongly exothermic. Thus, large amounts of thermal energy are returned to the plastic and to the surroundings. Pyrolysis and the associated fuel generation are thereby accelerated, and the fire propagates. Flame retardants interfere with these processes by one or more chemical or physical mechanisms [90].

Fuel Dilution. Noncombustible products (CO_2, H_2O) are generated by thermal decomposition of the flame retardant and released into the gas phase. Examples of such flame retardants are $Al(OH)_3$, $Mg(OH)_2$, hydrated carbonates, boric acid, and borates. The decomposition temperature of the flame retardant must be compatible with the processing temperature and pyrolysis temperature of the plastic.

Endothermic Action. If the pyrolytic decomposition of the flame retardant is endothermic, it can absorb part of the energy released during combustion. This mechanism often accompanies fuel dilution.

Protective Coating. The flame retardant forms a protective coating on the surface of the pyrolyzing plastic. It acts as thermal insulation and a barrier to oxygen, and also prevents combustible pyrolysis products and smoke particles from escaping into the gas phase. Most of the inorganic and organophosphorus flame retardants function by this mechanism. The protective layer can be glassy or have a porous carbon structure (intumescent char). In the latter case, the flame retardant system includes components that are responsible for char formation (e.g., polyols and phosphates) and expansion (e.g., melamine derivatives) [81], [91].

Pyrolysis Modification. Pyrolysis of the plastic can be chemically modified, for example, by dehydration in cellulosic polymers and polyurethanes or by inhibition of thermal depolymerization [e.g., in poly(methyl methacrylate)] to reduce the amount of combustible products generated. The proportion of solid carbon residue (char) is increased, and a noncombustible product (e.g., water) is formed. Organophosphorus compounds employ this mechanism; it is particularly effective in plastics containing oxygen. This mechanism may, however, increase smoke density.

Table 12. Producers and trade names of chlorinated organic flame retardants

Structure	Name	CAS registry no.	Producer (trade name)
Chlorinated paraffins, 30–70% Cl, some also containing bromine			Atochem (Electrofine), Dover Chemical (Chlorez), Ferro, Hoechst, Hüls, ICI (Cerechlor), Occidental Chemical (Chlorowax)
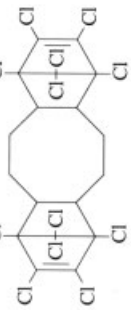	1,2,3,4,7,8,9,10,13,13,14,14-dodecachloro-1,4,4a,5,6,6a,7,10,10a,11,12,12a-dodecahydro-1,4:7,10-dimethanodibenzo(a,e)cyclooctene	[13560-89-9]	Occidental Chemical (Dechlorane Plus)
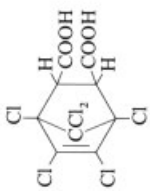	1,4,5,6,7,7-hexachlorobicyclo[2.2.1]-5-heptene-2,3-di-carboxylic acid	[115-28-6]	Occidental Chemical, Velsicol
	corresponding anhydride	[115-27-5]	

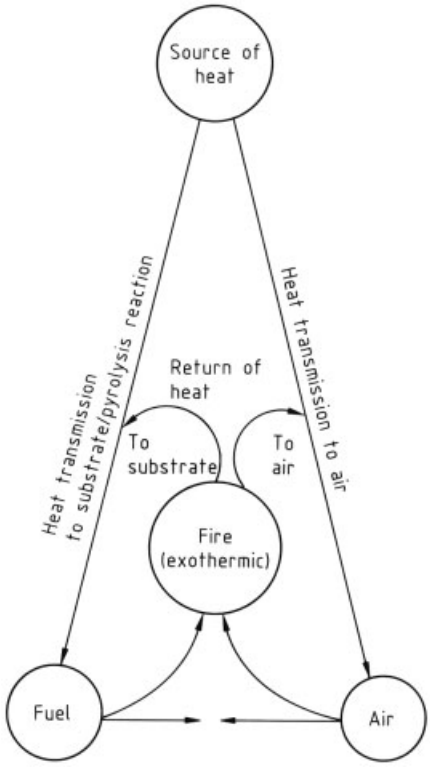

Figure 7. The Emmons fire triangle

Gas-Phase Quenching. Combustion processes in the gas phase are highly exothermic, free-radical chain reactions in which the fuel is oxidized by oxygen or hydroxyl radicals, for example:

$$CO + HO^{\bullet} \longrightarrow CO_2 + H^{\bullet} + \text{heat}$$

$$H^{\bullet} + O_2 \longrightarrow HO^{\bullet} + O^{\bullet} + \text{heat}$$

$$RH + O^{\bullet} \longrightarrow HO^{\bullet} + R^{\bullet} + \text{heat}$$

Flame retardants that are active in the gas phase must neutralize the highest-energy radicals, such as $HO^{\cdot}$ or $H^{\cdot}$. Halogenated compounds are especially effective in this respect. Pyrolysis produces hydrogen halides or, if antimony trioxide is present [92], antimony halides, which react as follows:

$$HBr + HO^{\bullet} \longrightarrow H_2O + Br^{\bullet}$$

$$HBr + H^{\bullet} \longrightarrow H_2 + Br^{\bullet}$$

$$SbBr_3 + H^{\bullet} \longrightarrow HBr + Br_2Sb^{\bullet}$$

Table 13. Organophosphorus flame retardants

Type or structure	Name, CAS registry no.	Producer (trade name)
Triaryl phosphates		
Triphenyl phosphate	[*115-86-6*]	Akzo (Phosflex)
Tricresyl phosphate	[*1330-78-5*]	Ciba-Geigy (Reofos)
Isopropylated triphenyl phosphate		FMC (Kronitex)
		Monsanto (Santicizer)
Other phosphates, phosphonates, or phosphinates		
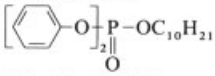	diphenyl decyl phosphate [*14167-87-4*]	Monsanto (Santicizer)
$CH_3-\underset{\underset{O}{\|}}{P}-(OCH_3)_2$	dimethyl methylphosphonate [*756-79-6*]	Akzo (Fyrol DMMP) Albright & Wilson (Amgard DMMP) Ciba-Geigy (Reoflam DMMP)
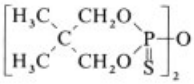	1,3,2-dioxaphosphorinane-2,2′-oxybis-(5,5-dimethyl-2,2′-disulfide) [*4090-51-1*]	Sandoz (Sandoflam 5060)
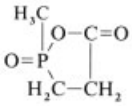	2-methyl-2,5-dioxo-1,2-oxaphospholane [*15171-48-9*]	Hoechst
$(C_2H_5O)_2\underset{\underset{O}{\|}}{P}-CH_2-N(CH_2CH_2OH)_2$	diethyl *N,N*-bis(2-hydroxyethyl)amino-methylphosphonate [*2781-11-5*]	Akzo (Fyrol 6)
Polyols containing phosphorus		Akzo (Fyrol 35) Hoechst (Exolit 413)
$O=P(OCH_2CH_2Cl)_3$	tris(2-chloroethyl) phosphate [*115-96-8*]	Akzo (Fyrol CEF) Albright & Wilson (Amgard TCEP) Bayer (Disflamoll) Hoechst (Genomoll P)
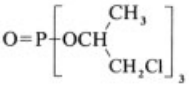	tris(2-chloroisopropyl) phosphate [*13674-84-5*]	Akzo (Fyrol PCF) Albright & Wilson (Amgard TMPC)
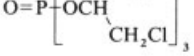	tris(2,2′-dichloroisopropyl) phosphate [*13674-87-8*]	Akzo (Fyrol FR-2) Albright & Wilson (Amgard TDCP)
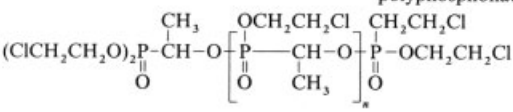	polyphosphonate [*59036-37-2*]	Sandoz (Sandoflam 5087)
Halogenated phosphate esters and other proprietary phosphorus compounds		Akzo (Fyrol) Albright & Wilson (Amgard) Ciba-Geigy (Reoflam) FMC (PB-460, PB-528) Great Lakes Chem. Sandoz (Sandoflam)
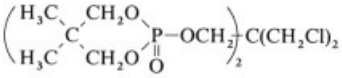	1,3,2-dioxaphosphorinane-2,2′{[2,2-bis (chloromethyl)-1,3-propanediyl]bis(oxy)}-bis(5,5-dimethyl-2,2′-dioxide) [*83044-97-7*]	Sandoz (Sandoflam 5085)

Reactions of this type are less exothermic and block free-radical chain reactions. A drawback of this mechanism is that the gas phase becomes more corrosive. The smoke density may also increase due to inhibition of the combustion reactions.

6.4. Test Methods and Standards

The use of flame-retardant materials (especially in building materials, furnishings, electrical or electronics equipment, mining, and transportation) is covered by application-specific regulations. Many test methods and standards have been created to classify the burning behavior of materials but there is little correlation between them. The small-scale tests performed in materials development can only simulate actual burning behavior within limits; full-scale tests must therefore be performed.

Test methods and standards are specified by (inter)national standards organizations, industry groups or organizations, and government agencies. Numerous test methods are described in [78]. See also, → Flame Retardants, Chap. 3. The methods encountered most frequently in development and testing laboratories are listed in Table 14. The relatively simple UL 94 test is most widely used and recognized. A standard test rod is clamped in a vertical position, and an igniting flame is applied to the lower end twice for 10 s. The second ignition takes place as soon as the flame from the first ignition is extinguished. The rating criteria are the afterburning time and the ignition of cotton wadding placed under the specimen caused by molten drops of material. The test gives a rating of V-0, V-1, or V-2:

V-0: afterburning time of single specimen <10 s, of five specimens with two ignitions each <50 s; no ignition of wadding
V-1: afterburning time of single specimen <30 s, of five specimens with two ignitions each <250 s; no ignition of wadding
V-2: same as V-1, but with ignition of wadding

The specimens must not be completely burned, otherwise no such classification is given.

6.5. Uses

About 10 % of all plastics (chiefly PVC, ABS, polystyrene, unsaturated polyesters, polypropylene, polyethylene, and polyurethanes) contain flame retardants. The main applications are in building materials and furnishings (structural elements, roofing films, pipes, foamed plastics for insulation and furniture, wall and floor coverings); transportation (equipment and fittings for aircraft, ships, automobiles, and railroad cars); and in the electrical industry (cables; housings and components for television sets, office machines, and household appliances; lamination of printed circuits).

More than 90 % of the flame retardants used in thermoplastics are of the additive type. They are added before, during, or after polymerization, but usually when the polymers are processed into compounds or finished products. In the latter case, the retardant is often used as a highly concentrated masterbatch containing 50 – 80 % of the agent.

Reactive flame retardants are added to thermosetting resins before the cross-linking reaction.

Poly(Vinyl Chloride). Up to half of the dioctyl phthalate plasticizers in *flexible PVC* can be replaced by chlorinated paraffins or phosphate esters (e.g., triphenyl or tricresyl phosphate) [93]. Other common additives are Sb_2O_3 (2 – 4 %) and $Al(OH)_3$ (40 %).

In *rigid PVC*, Sb_2O_3 (1 – 4 %) and $Al(OH)_3$ (10 – 20 %) are employed and in some cases zinc borate and zinc stannate. MoO_3, ferrocene, or $CaCO_3$ can be used as smoke inhibitor.

Acrylonitrile – Butadiene – Styrene. A variety of halogenated flame retardants are used for ABS in conjunction with Sb_2O_3; most common for V-0 formulations is a combination of octabromodiphenyl oxide (ca. 20 %) and Sb_2O_3 (ca. 6 %). Halogen-free formulations, for example with $Mg(OH)_2$, are still under development.

Expandable Polystyrene. Hexabromocyclododecane is used at the 3 – 5 % level to comply with building codes; 0.5 % dicumyl peroxide or another free-radical former is added as synergist [88].

Table 14. Test methods and standards (selected)

Test	Standard	Criteria	Applications
LOI test	ASTM D 2863	flammability	development laboratory
Vertical test	UL 94, UL 94-5 V	ignitability dripping	development laboratory electrical industry
Vertical test	DIN 54 332	ignitability	development laboratory floor coverings
Horizontal test	ASTM D 635	burning rate	development laboratory
Miscellaneous tests	FAR 25	ignitability burning rate smoke density	aircraft interiors
Horizontal test	FMVSS 302	burning rate	automotive interiors
Cone colorimeter	ISO DP 5660	ignitability heat release smoke toxic gases	development laboratory materials in buildings
Epiradiator	NF P 92-501	ignitability burning rate	materials in buildings
Surface spread of flame test	BS 476, Parts 5 – 7	ignitability burning rate	materials in buildings
NBS smoke chamber	ASTM E 662	smoke density	development laboratory aircraft interiors
Toxicity test	DIN 53 436	toxicity of pyrolysis gases	development laboratory materials in buildings
Brandschachttest	DIN 4102	flammability heat release	materials in buildings
Room corner test	NT Fire 025	flame spread smoke toxic gases	materials in buildings
Steiner tunnel test	ASTM E 84	flame spread smoke	materials in buildings

High-Impact Polystyrene. The V-0 grades contain ca. 12 % decabromodiphenyl oxide or tetrabromobisphenol A together with 3 – 4 % Sb_2O_3. Ethylene-bis(tetrabromophthalimide) and low molecular mass brominated polystyrenes are also suitable. For V-2 grades, brominated aliphatics are used. Red phosphorus (15 %) has been proposed for halogen-free V-0 grades.

Polyester Resins. The use of 50 – 60 % $Al(OH)_3$, preferably coated, is widespread. Halogenated reactive retardants, such as tetrabromophthalic anhydride and (for more light-stable grades) 2,2-bis(bromomethyl)-1,3-propanediol are also employed.

Low molecular mass brominated epoxy polymers of the F – 2000 series, together with Sb_2O_3, offer another alternative. Zinc borate or dimethyl methylphosphonate is also used.

Polyolefins. In polypropylene, 5 – 20 % tetrabromobisphenol A-bis(2,3-dibromopropyl ether) is used, along with 3 – 8 % Sb_2O_3 [94]. Other halogenated compounds are also used in combination with Sb_2O_3: ethylene-bis(tetrabromophthalimide), ethylene-bis(dibromonorbornane dicarboximide) or Dechlorane Plus. Halogen-free V-0 grades are obtained by using 25 – 35 % of an intumescent system based on ammonium polyphosphate – pentaerythritol – melamine derivatives or 50 – 60 % $Al(OH)_3$ or $Mg(OH)_2$ [95].

Chlorinated paraffins combined with Sb_2O_3 are used in LDPE films. Brominated aromatics or Dechlorane Plus with Sb_2O_3 are used for HDPE and cross-linked polyethylene. Levels of 30 % halogenated organics and 10 % Sb_2O_3 are required for V-0 grades. Halogen-free V-0 grades (e.g., cable sheathing compounds) must contain at least 60 % $Al(OH)_3$ or $Mg(OH)_2$; addition of zinc borate, zinc stannate, or silicones is recommended. Intumescent systems have also been proposed.

Polyurethanes. Flexible polyurethane foams (e.g., for upholstery) contain 20 – 30 %

melamine together with a few percent of halogenated phosphate [tris(chloroethyl) phosphate, tris(2-chloroisopropyl) phosphate] or chlorinated paraffins. A combination of graphite and tris(2-chloroisopropyl) phosphate has also been successful. Other formulations include $Al(OH)_3$ or phosphorus-containing products such as dimethyl methylphosphonate, tris(2-chloroethyl) phosphate, tris(2-chloroisopropyl) phosphate, and ammonium polyphosphate.

Rigid foams employ phosphorus-containing or (preferably) halogenated diols and polyols [e.g., diethyl-*N,N*-bis(2-hydroxyethyl)aminomethyl phosphonate, 2,2-bis(bromomethyl)-1,3-propanediol or IXOL B 251] in combination with dimethyl methylphosphonate or halogenated phosphates.

Thermoplastic polyurethane can be rendered flame resistant with brominated compounds and Sb_2O_3 [96] or intumescent systems.

Polyamides. A variety of formulations are used for V-0 grades of polyamide 6 and polyamide 66 [97]. In systems based on brominated aromatics (15 – 20 %) and Sb_2O_3 (4 – 5 %), ethylene-bis(tetrabromophthalimide) and polymeric flame retardants such as poly(tribromostyrene), poly(dibromophenylene ether), and F – 2000 products dominate. Dechlorane Plus is also employed with synergists (e.g., Sb_2O_3, zinc borate, or Fe_2O_3 and Fe_3O_4).

In halogen-free systems, polyamides contain 7 – 10 % red phosphorus or, in the absence of glass-fiber fillers, melamine cyanurate. Tests with $Mg(OH)_2$ are under way.

Polyesters (PBT and PETP). The products used for polyesters are similar to those for polyamides, but combinations of brominated compounds (10 – 20 %) with Sb_2O_3 (3 – 5 %) clearly dominate. A cyclic methylphosphinic acid derivative 2-methyl-2,5-dioxo-1,2-oxaphospholane is chemically incorporated into polyester fibers at the condensation step.

Polycarbonates. Tetrabromobisphenol A is incorporated into polycarbonates as a comonomer; very small amounts of the alkali salts of aromatic sulfonic acids improve flame retardance [89]. Opaque V-0 polycarbonate grades are obtained by using 2 – 4 % brominated aromatics and 1 – 2 % Sb_2O_3 as additive retardants.

Poly(Methyl Methacrylate). The cast plastic can be treated with 13 – 20 % of a halogenated polyphosphonate (Sandoflam 5087) [98].

Poly(Phenylene Oxide) – Polystyrene. V-0 grades of blends of polystyrene and poly(phenylene oxide) can be obtained with ca. 15 % halogen-free triaryl phosphates.

Epoxy Resins. Reactive flame retardants (tetrabromobisphenol A and derivatives, glycidyl ethers of brominated aromatics) are employed for epoxy resins (minimum bromine content 20 %). For printed-circuit laminates, the use of Sb_2O_3 must be avoided. Tests on epoxy resins are being carried out with red phosphorus, $Al(OH)_3$, and zinc borate.

6.6. Toxicology

The question arises whether flame retardants, especially those with chemical action, increase the toxicity of combustion gases. Animal studies have shown that pyrolysis gases emitted by flame-retarded plastics are often less toxic than those from untreated plastics [99]. Standards governing toxicity tests on combustion gases include DIN 53 436 [100] and ISO/IEC/TR 9122-1. Carbon monoxide is the most significant factor in combustion-gas toxicity.

The inherent toxicity of some flame retardants and potential hazards during their processing and use in plastics have been discussed. Of special concern are the polybrominated diphenyl ethers because polybrominated dibenzodioxins and dibenzofurans have been detected in their decomposition products. The Verband der Chemischen Industrie (Chemical Industry Association) in Germany has decided to gradually phase out these retardants. In the United States, an initiative of the Brominated Fire Retardant Industry Panels (BFRIP) has led to in-depth studies on the relevance of the findings [101].

Chlorinated paraffins must be labeled in the United States if their residual carbon tetrachloride content exceeds 0.1 %.

6.7. Economic Aspects

The estimated world market (1990) of flame retardants is 500 000 t, which is subdivided as follows:

Inorganic	290 000 t
Brominated organic	80 000 t
Chlorinated organic	50 000 t
Organophosphorus	60 000 t

The dominant inorganic product is $Al(OH)_3$, which is also the cheapest flame retardant available.

Producers and trade names of flame retardants are listed in Tables 10, 11, 12, 13.

7. Fillers and Coupling Agents

7.1. Purpose and Requirements

Filled (or reinforced) plastics contain large amounts of fillers (20 – 50 wt % referred to the polymer or even occasionally higher). Fillers are divided into *extenders*, which merely provide bulk and lower product cost, and *reinforcements*, which improve several of the polymer properties. *Coupling agents* improve the adhesion between filler and polymer, preferably via chemical bonds; their use confers reinforcing properties on inexpensive extenders, improves the performance of reinforcements, and allows the filler content to be increased. The reinforcing effect of fillers depends on their chemistry, shape (fibers, flakes, spheres), and size (fiber length, particle size). The most important reinforcement effects are

1) Increased strength, ultimate elongation, rigidity (modulus of elasticity), and in some cases impact strength
2) Improved heat distortion temperature and dimensional stability, reduced shrinkage, and improved stability of mechanical values at high temperatures and over extended times (fatigue)
3) Modification of density (usually increased)
4) Improved chemical resistance and lower water absorption
5) Better surface quality and surface hardness

Special fillers are used to obtain compounds with the following properties:

1) Low flammability
2) Electrical conductivity and electromagnetic shielding
3) Radiation and UV shielding
4) Biodegradability
5) Noise suppression

Fillers are characterized by a number of parameters that include geographic origin (minerals); particle size and size distribution or fiber length; purity, especially heavy-metal content; water content; whiteness; oil adsorption value or plasticizer adsorption value (ASTM D 281-31, for PVC fillers); density; specific heat; and thermal conductivity. Consistency, especially with respect to purity and particle size, is an important criterion of quality.

Occupational health aspects (e.g., dust) and processing qualities (e.g., abrasive behavior toward machinery, effects on viscosity, and wettability – dispersability in the melt) are also important factors influencing the selection of fillers.

7.2. Forms and Chemical Structures

Spherical Fillers and Powders. Calcium carbonate is quantitatively the most important filler. Natural forms (e.g., chalk, marble, or limestone) are ground to powder in dry or wet form. Finer synthetic or precipitated forms [often abbreviated as CCPs (calcium carbonate precipitated) are obtained by passing carbon dioxide through milk of lime or by reacting calcium chloride and sodium carbonate.

Other mineral fillers include ground dolomite [$CaMg(CO_3)_2$], calcium sulfate (gypsum), and barium sulfate (ground barite or precipitated forms, which have a very high density of 4.5 g/cm^3).

Silica is used as a natural crystalline product (sand, quartz powder) or amorphous synthetic product. The latter is obtained by precipitation from a waterglass solution with acid or by flame hydrolysis from silicon tetrachloride (pyrogenic silica gel).

Solid and hollow glass microspheres of various sizes are employed. Hollow microspheres have the lowest density of all fillers (0.2 – 0.4 g/cm^3).

Other powders include wood flour, cork flour, starch, metal oxides (e.g., Al_2O_3, MgO, ZnO, TiO_2, iron oxides), and metal powders (e.g., Al, Cu, Ni, bronze, Zn, Ag-coated silicate spheres).

Carbon black is obtained by incomplete combustion of gaseous and liquid hydrocarbons. The quality and potential applications (black pigment, conductivity additive, or reinforcement) depend on the production process, purity, and particle size.

Flake fillers include talc, kaolin (China clay), mica, silicon carbide, aluminum trihydroxide, graphite, and metals (e.g., aluminum).

Fibrous Fillers and Reinforcements. A wide variety of inorganic and organic fibrous fillers and reinforcements are used. New ceramic and mineral fibers are being developed.

Carbon fibers are made primarily by oxidative thermal cyclization of polyacrylonitrile fibers. They are classified by their strength. Short fibers (e.g., 3 or 6 mm) are cut from filament yarns.

The starting material for the production of *glass fibers* is glass filament, which is cut into textile glass a few millimeters long. This product is milled to short fibers (e.g., 0.2 – 0.6 mm) when the reinforcement is blended with the polymer in the extruder.

Aramid fibers are obtained, for example, by the polycondensation of terephthaloyl chloride and *p*-phenylenediamine. They are cut to various lengths and used as reinforcement.

Single-crystal fibers (i.e., *whiskers*) can be based on metals, metal oxides, or metal carbides.

Coupling Agents. Typical silane finishes have the general structure $(X)_3Si(CH_2)_3$-Y, in which X is a short-chain alkoxy group (OCH_3, OC_2H_5), an acyloxy group, OH, or Cl; Y is a functional group such as mercapto,amino,azido,methacrylate, epoxy, an anhydride function, or a longer hydrocarbon chain.

Functional groups (e.g., vinyl) can also be attached directly to the silicon. Polymeric silanes [102] containing alkoxy groups and functional groups, as well as bis(trimethoxy) silyl, are also used. New developments include blends of silanes having different functions [103].

The *titanate and zirconate systems* have the general formula $X_mM(O\text{-}X\text{-}R\text{-}Y)_n$, where M denotes Ti or Zr; X denotes an alkoxy group; Z may be a carboxyl, sulfonyl, phosphate, or pyrophosphate group; R can be a long hydrocarbon chain; and Y is a functional group such as amino or methacryloxy. Not all the groups Z, R, and Y need be present. A survey of recommended types is given in [104].

Miscellaneous Coupling Agents. These include aluminates [104], zirconium aluminates [105], stearic acid and stearates, chlorinated paraffins, polyolefin waxes, and carboxylated polybutadienes, or polypropylenes modified with acrylic or maleic acid.

7.3. Mechanism of Action

Reinforcement depends on the transmission of mechanical energy (impact energy, tensile stress, etc.) from the polymer matrix to the filler. It increases with the strength of (chemical) bonding between the fillers and the matrix. Bonding also limits the mobility of the polymer chain and stiffens the material. Fibrous and flake fillers and reinforcements give rise to anisotropic properties.

Chemical bonding seldom occurs directly between the matrix and the filler; it is usually mediated by coupling agents. In silanes, zirconates, and titanates, for example, the X groups react directly, or after hydrolysis, with hydroxyl or carboxyl groups on the filler surface. The finish materials, particularly silanes, also undergo condensation to form higher molecular mass units:

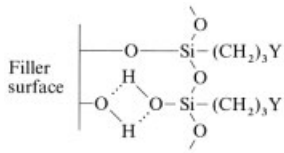

The Y functional groups react with the polymer (especially in thermosets) or are anchored to the polymer by physical mechanisms, for example, by penetration into the coiled structure in thermoplastics.

7.4. Test Methods

Fillers are tested on receipt, particularly for particle size (screen or sedimentation analysis),

acidity, whiteness, density, oil adsorption, and plasticizer adsorption value [dioctyl phthalate (DOP) number]. Tests can be performed by standard methods [106]. Testing the adhesion between the filler and the polymer matrix and correlating it with important physical properties is especially difficult. Publications in this area deal chiefly with fiber-reinforced polymers [107].

7.5. Uses

Nearly all thermosetting and thermoplastic resins are filled or reinforced for special applications. The main applications are in PVC, followed by polyesters, polypropylene, and polyamides. The following examples give a brief summary of the types of fillers and reinforcements used and their effects. Further information can be found in the literature (e.g., [108]).

Calcium carbonate is important in PVC. Natural forms of calcium carbonate are preferred for flexible PVC because they have lower plasticizer values than precipitated forms. Filler contents of 30 – 60 % or higher are common for cables, film, and floor covering, leading to substantial reductions in cost and additional savings of lubricants, plasticizers, white pigments, or stabilizers.

Precipitated calcium carbonates with surface treatments (coated CCPs) are preferred for rigid PVC. Contents of 10 – 20 %, occasionally up to 50 %, are used for window frames and other profiles, foam, and pipes. Important considerations, in addition to lower cost, are improved impact strength and stiffness, higher production rate, and improved surface appearance. $CaCO_3$ is valued for its low abrasivity in processing.

Other areas for $CaCO_3$ use include polyolefins. Coated types are used for polypropylene in industrial applications (increased heat distortion temperature, dimensional stability, rigidity, and impact strength). Films of LDPE and LLDPE can also be filled with 30 – 40 % $CaCO_3$, which provides slip and antiblocking effects (e.g., trash bags). Calcium carbonate is also used in polyamides, PETP, PBT, and semiopaque grades of other thermoplastics, such as poly(methyl methacrylate).

Barium sulfate is employed in thermoplastics and provides acoustic damping, X-ray absorption, and increased density.

Kaolin is used in elastomers and flexible PVC for electric cables because of its high water resistance and excellent insulating properties. Kaolin has a high whiteness but also a higher plasticizer adsorption value than, for example, $CaCO_3$.

Mica is employed, for example, in automotive parts made of polypropylene.

Talc is a widely used filler for PVC and polyolefins. Polypropylene components gain in heat distortion temperature and rigidity but lose impact strength and thermooxidative stability.

Aluminum trihydroxide and magnesium dihydroxide are used as flame retardants in polymers such as polypropylene, polyethylene (cable), PVC (carpet backing), and unsaturated polyesters. Filler contents are high (>50 %).

Zinc oxide serves as an UV absorber in polyolefins.

Wood Flour and Starch. Wood flour is employed in acoustical components made of, for example, HDPE or PVC. The addition of starch, preferably in conjunction with oxidation promoters, is reported to render polyethylene films biodegradable.

Metal powders and carbon black are added to many polymers to improve thermal or electric conductivity.

Silica gel has many applications in thermoplastics (PVC, polyamides) and thermosets (polyester resins, epoxy resins). It is also used as antiblocking and slip agent in LDPE.

Glass spheres are used in polyolefins (including foams) and engineering thermoplastics (polyamides, PETP, PBT). Hollow glass spheres lower the density and result in high transparency.

Glass fibers are by far the most important reinforcement for plastics. For further details, see → Fibers, 5. Synthetic Inorganic, Chap. 2.1. They are generally used at levels of 30 – 40 %, and improve tensile and flexural strength, elastic modulus, and heat distortion temperature. The reinforcing effect increases with fiber length and coupling agent. Fiber length must be matched to the elastic modulus of the matrix. Engineering thermoplastics [e.g., polyamides, poly(phenylene oxide), PETP, PBT] and ABS are usually supplied with glass-fiber reinforcement (typical length 0.2 mm) and used for housings or injection moldings in electrical and electronics equipment.

Polypropylene with glass-fiber reinforcement (typical length 0.66 mm) is also important, especially for automotive parts in the United States. Reinforcement results in polypropylene with properties similar to those of engineering thermoplastics. In the thermoset area, long fibers are used for reinforcement. Glass-fiber reinforcement is less common in polyethylene, PVC, and polystyrene.

Carbon fibers are used mainly as short fibers processed into granular compounds. They provide dimensional stability, electrical conductivity, and (especially metal-coated grades) electromagnetic shielding. They are used in engineering thermoplastics, and epoxy resins. For further details, see → Fibers, 5. Synthetic Inorganic, Chap. 5.

Aramid fibers (see also, → Fibers, 2. Structure, Chap. 4.2.) are used chiefly as long fibers (in thermoplastics, their length corresponds to that of the compound granules). They lend rigidity with a low density. Preferred substrates are polyamides, PETP, and epoxy resins.

Coupling agents are employed at levels of 0.1 – 4 % referred to the filler. They are usually applied to the filler (coated fillers). Aqueous solutions, emulsions, and liquid products are applied by spraying. Coupling agents can, however, be also incorporated directly into the polymer or supplied as powdered concentrates in inert fillers and blended in at the compounding stage.

Silane coupling agents are applied to glass fibers, glass beads, and silicates such as mica, talc, wollastonite, and kaolin. They are also used in polymers treated with the flame retardants $Al(OH)_3$ and $Mg(OH)_2$. Silanes (except for special grades) are not, however, suitable for $CaCO_3$, carbon fibers, or metal oxides. They are useful for service in a humid atmosphere because they afford hydrolysis-resistant adhesion.

Titanates are especially good for $CaCO_3$, $BaSO_4$, carbon fibers, aramid fibers, metals, and metal oxides. They tend to become discolored in the presence of phenolic antioxidants. *Zirconates* are used for both $CaCO_3$ and silicates; they do not discolor with phenolic antioxidants.

Stearic acid and stearates are mainly used in $CaCO_3$, especially CCP grades. *Chemically modified polypropylene and polybutadiene* are used on glass fibers, mica, $CaCO_3$, and other fillers, chiefly in polypropylene.

Functional groups in the finish should be selected in accordance with the type of polymer. Amino groups are suitable for polyamides, polycarbonates, PETP, PBT, PVC, and many thermosets. Methacrylate groups are used with polystyrene, ABS, SAN, and thermosets, as are epoxy and mercapto groups. Vinyl groups are employed for polystyrene, polyolefins, and PVC. Azido groups are recommended for polyolefins; inert, long-chain hydrocarbon groups can also be used instead of functional groups.

7.6. Economic Aspects

The market volume for fillers and fiber reinforcements is difficult to estimate but is probably on the order of at least 5×10^6 t/a. Mineral fillers are most important, followed by synthetic fillers and glass fibers. Coupling agents have reached a market volume of ca. 10 000 t/a. Because of the large number of suppliers, reference must be made to the literature [108], [109]. Table 15 lists some important producers and trade names of coupling agents.

8. Dyes and Pigments [110–112]

Although the main function of a colorant (i.e., a dye or a pigment) is to impart decorative character and sales appeal, it may also influence the

functional properties of the substrate. The selection of colorants from a complex group of minerals and chemical substances has been cited as one of the higher forms of the "black art" of plastics compounding. The problem of colorant selection has become more acute because of the growing complexity of legislation dealing with product safety, occupational safety, and environmental protection. Thus the coloration of plastics is becoming an increasingly precise science.

Table 15. Producers and trade names of coupling agents

Producer	Trade name	Type
Akzo		titanates
BP Performance Polymers	Polybond	functionalized polymers
Cavedon	Cavco	zirconates, zirconium aluminates
Degussa		silanes
Dow Corning		silanes
Du Pont		titanates
Eastman	Epolene	functionalized polymers
Hercules	Az-cup	silanes
Hüls	Dynasilan, Polyvest	silanes, titanates, functionalized polymers
Kenrich Petrochemicals	Ken-React	titanates, zirconates, zirconium aluminates
Nippon Soda		titanates
PRC	Prosil	silanes
Titanium Intermediates		titanates
Union Carbide	Ucarsil PC	silanes
Wacker Chemie		silanes

8.1. Requirements [113], [114]

Modern color-delivery systems (particularly master batches) provide overall economy and ease of application. The main requirements for a colorant used in plastics include

1) Inertness under processing conditions (visible and invisible thermal fatigue)
2) Physical and chemical inertness toward the polymer
3) Compatibility with other additives (antagonism/synergism)
4) Required performance during the intended service life of the plastic product
5) Nontoxicity
6) Ability to be disposed of in discarded plastic items without causing extensive environmental nuisance (ecological safety)

Colorants are chosen from a wide range of organic and inorganic compounds. Pigments are particulate organic and inorganic solids that are virtually insoluble in the medium into which they are incorporated. The pigment particles must be dispersed in this medium. Dyes are organic compounds that are usually dissolved in the substrate. Incompatibility is manifested by "crocking", the migration of colorant to the free surface of the plastic from which it can be rubbed off.

Colorants are introduced into plastics by several methods [115]. The dry colorant may be blended with the polymer before processing (dry blending); dispersion or dissolution is less efficient than with color concentrates (master batches). Master batches are predispersed colorants in a physical form that is easy to handle and use (usually pellets, but also liquid or paste). They offer numerous advantages over powdered colorants, particularly more hygienic working conditions.

8.2. Pigments

8.2.1. Inorganic Pigments [116], [117]

See also → Pigments, Inorganic.

Inorganic pigments are insoluble in plastics and hence do not bleed or migrate; they are therefore the colorants of choice for plastics. They generally exhibit good opacity and hiding power. Colors range from dull earth tones (iron oxides) to bright yellows and reds (pigments based on lead and cadmium).

Titanium dioxide is the most widely used pigment; it imparts opacity and whiteness to all plastics. Rutile grades are preferred. They are produced by the chloride process, which gives a brighter product, or the sulfate process, which gives a product with a slight "bone" undertone but a lower abrasiveness. To enhance dispersibility and lower the photo- and antioxidant reactivity toward polymer resins, the pigments are usually coated with inorganic materials (e.g., silica or alumina). They are also aftertreated with organic compounds to render them hydrophobic and make them easier to disperse. Plastic-grade titanium dioxide pigments of high hiding power usually have a much finer particle size, in the 0.17 – 0.24 μm range.

Zinc Pigments. Although not as important as titanium dioxide, white zinc-based pigments do have technical advantages, particularly where wear of processing equipment and damage to the polymer during processing or in service are crucial.

Iron oxide pigments constitute a very important class of plastics colorants, especially because of their relatively low cost and because of toxicology and environmental considerations. The major shortcomings of iron oxide pigments are their rather dirty shades, ranging from buff yellow to brick red, brown, and black; their abrasiveness with respect to processing equipment; and the fact that they promote the degradation of sensitive polymers (e.g., PVC, polyolefins, and polyamides) by light and weathering. Moreover, the yellow, brown, and black iron oxides are stable only up to 180 °C.

Chromium Oxide Green. The dull green chromium oxide pigments have a poor tinting strength but outstanding heat, light, and chemical stability, making them suitable for all polymers. They are, however, difficult to disperse.

Lead Chromate Molybdate Pigments. The versatile group of lead chromate molybdate pigments ranges in color from primrose to lemon yellow, orange, scarlet, and even green. Silica-coated chromate and molybdate pigments have been developed to provide heat and weather resistance in plastics. Although they offer a combination of bright shades, high hiding power, excellent lightfastness, and low cost, the trend now is toward pigments that do not contain heavy metals.

Cadmium pigments are ideal for coloring plastics and are the only bright colors available for many industrial polymers. They have excellent heat stability, even in highly aggressive industrial polymers such as nylons; a distinct hue ranging from yellow to orange, deep red, and maroon; and very good lightfastness. Stability under wet conditions is rather limited, however. Cadmium pigments do not cause dimensional instability in colored plastics. As with lead chromate pigments, however, stringent regulations and greater environmental consciousness are forcing plastics colorant users to phase out pigments based on cadmium.

Mixed Metal Oxide Pigments. In the search for lead- and cadmium-free formulations, metal oxide pigments (especially nickel antimony titanate and chromium antimony titanate) have attained prominence in both coating and plastics applications. Although their color intensity is low, they are usually combined with color-intense organic colorants to produce the required shades. Titanate pigments have exceptional heat, weather, and chemical stability. Cobalt aluminate pigments are used mainly as shading components.

Ultramarine Blue. Apart from their poor acid resistance, ultramarine blues exhibit excellent all-round fastness. Unlike the blue phthalocyanine pigments, ultramarine blues do not induce distortion in polyolefins. With their excellent heat stability, ultramarines are ideally suited for color correction of white and clear polymers.

Special-Effect Pigments [118]. Pearlescent (nacreous), iridescent, and metallic colorants are increasingly used in plastics. Synthetic nacreous pigments consist of TiO_2-coated mica. Colored pearlescent pigments (e.g., Iriodin, Merck) have an additional coat of iron oxide or chromium oxide. These pigments provide a pearly luster and are used, for example, in cosmetic packaging and fashion articles.

A metallic appearance can be imparted by metal flakes (e.g., aluminum).

8.2.2. Organic Pigments [116], [117], [119], [120]

See also → Phthalocyanines, → Pigments, Organic.

Organic pigments have cleaner and brighter shades and a much higher tinting strength than inorganic pigments. They tend to dissolve in polymers, however, particularly at very low concentration. The dissolved pigment can undergo thermal degradation [121]. Organic pigments generally produce translucent or semitransparent coloring. Many color formulations are derived from a combination of inorganic and organic pigments whose properties complement one another.

In contrast to inorganic pigments, it is difficult to indicate a particular class of organic pigments for a particular application.

The newer types of monoazo yellows and oranges (e.g., C.I. Pigment Yellow 180, 181, and 182; Orange 64) and monoazo yellow toners (C.I. Pigment Yellow 183, 190, and 191) possess improved heat stability up to 280 °C in polyolefins and even ABS, similar to those of disazo condensation pigments. Among monoazo red toners, C.I. Pigment Red 151 and 247 show good heat stability but rather poor lightfastness.

Diarylide yellows, oranges, and reds tend to undergo thermal decomposition in polyolefins above 200 °C, particularly at low concentration, generating potentially toxic 3,3′-dichlorobenzidine. Their use has recently been limited to PVC.

Azo condensation pigments have adequate heat stability for use in polyolefins and other resins processed at moderate temperature. They generally tend to be dirty in mass tones, however, especially reds. An exception is C.I. Pigment Red 242, a brilliant scarlet, which is of great interest as a heavy-metal replacement.

Except for their tendency to induce distortion in polyolefins, the phthalocyanine pigments, especially the blues, represent the nearest approaches to ideal pigments on the market. Not only are they thermally and photochemically stable, but they also exhibit excellent chemical stability (unlike many inorganic pigments).

Metal-complex pigments are generally rather dirty in shade, particularly in white reductions. C.I. Pigment Orange 68, a heterocyclic metal-complex pigment, is one of the substitutes for heavy-metal formulations in nylons and polycarbonates.

Polycyclic and some heterocyclic pigments (including diketopyrrolopyrrole pigments [122]) generally have very good heat stability in nonpolar polymers. In polar resins, especially nylons, they tend to poor heat and light stability. Of particular importance is the novel heterocyclic perinone pigment C.I. Pigment Yellow 192, which is an ideal cadmium substitute in nylons and polyesters.

Other organic pigments used to color plastics include isoindolines, isoindolinones, dioxazines, anthraquinonoids including indanthones, perylenes, perinones, thioindigoids, quinacridones, and quinophthalones.

8.2.3. Dyes [123]

Because of economy and ease of application, polymer-soluble dyes have been used to color polymers in brilliant and transparent shades. They show reasonably good lightfastness in transparent colorations. In combination with opacifying inorganic pigments, especially titanium dioxide, their lightfastness is rather poor. Polymer-soluble dyes are usually suitable for transparent plastics but not for polyolefins.

Dyes for plastics have a variety of chemical structures. Azo dyes are generally heat- and light-sensitive. Metal-complex azo dyes, especially the chromium complexes, exhibit good stability in nylons. Anthraquinone, quinophthalone, methine, naphthazine, perinone, cumarin, thioindigo, and thioxanthene dyes are usually employed.

Fluorescent pigments are used to produce eye-catching colorations, but tend to plate out in plastics [118]. They consist of finely divided resin particles that contain fluorescent dyes; see → Fluorescent Dyes, Chap. 12..

9. Miscellaneous Additives

9.1. Nucleating Agents

The properties of thermoplastics depend on their degree of crystallization and the morphology of their crystalline phases. Plastics may be amorphous (e.g., PVC, polystyrene, LDPE), of moderate crystallinity (e.g., LLDPE, polypropylene, polyamide 6, PETP), or highly crystalline (e.g., HDPE or polyamide 66). Chain mobility, tacticity, and side-chain volume determine the degree of crystallinity, as do the temperature profile, pressure, and orientation during polymer processing.

During cooling of polymer melts the disordered coiled structure is transformed into an ordered state with the formation of lamellar crystallites. The crystal nuclei can be residual, unmelted polymer crystallites already present in the melt (self-nucleation) or added agents (nucleating agents). The latter make crystallization energetically more favorable and increase

the number of nuclei [124–127]. The resulting crystallites form a morphological superlattice structure (e.g., spherulites), which have a central primary nucleus and a radially symmetric structure with amorphous domains occurring between the spherulites. Scattering of light by coarse spherulites (> 0.5 µm) reduces the transparency of plastics.

Nucleating agents control the crystallization behavior of moderately crystalline plastics in the following ways:

1) Increasing the crystallization temperature by ca. 10 – 20 °C
2) Accelerating crystallization rate
3) Decreasing the spherulite diameter and narrowing the size distribution

Classification. Nucleating agents can be divided into

1) Inorganic fillers and pigments: talc, kaolin, silica gel, TiO_2
2) Salts of carboxylic acids [128]: sodium benzoate, aluminum 4-*tert*-butyl benzoate, sodium montanate, sodium β-naphthenate
3) Sodium organophosphates [129]
4) Dibenzylidene sorbitols [130], [131]
5) Waxes [132] and ionomers [133]
6) Systems of carboxylic acids and amines [134]

Some producers and trade names are listed in Table 16.

Properties. Nucleating agents offer some or all of the following advantages. Cycle times in injection molding are reduced by as much as 30 %. They improve optical properties (particularly transparency and gloss), mechanical properties, dimensional stability, and oxidation resistance(less oxygen diffusion in crystalline phases). They also decrease water absorption (polyamides).

Preferred nucleating agents have a high melting point, insolubility in polymer melts, good wettability and dispersability in melts, small particle size (e.g., 1 µm), adequate thermal stability and no odor. They are also approved as indirect food additives (desirable).

Table 16. Producers and trade names of nucleating agents

Manufacturer	Trade name	Type
Adeka Argus	Mark NA	organic sodium phosphates
Allied Signal	Aclyn	ionomer
Du Pont	Surlyn	ionomer
EC Chemical Co.	EC-1	dibenzylidene sorbitol
Hoechst	Hostalub, Hostamont	sodium montanate
ICI Speciality Chemicals	SCS, Clarifex	proprietary
Milliken Chemical Co.	Millad	dibenzylidene sorbitol
Mitsui Toatsu	NC	dibenzylidene sorbitol
New Japan Chem.	Gellal	dibenzylidene sorbitol
Sandoz Chemicals	Sandostab	aluminum *tert*-butyl benzoate
Schering	Geniset MD	dibenzylidene sorbitol
Shell	AL-PTBBA	aluminum *tert*-butyl benzoate
Witco Corp.	Mark	organic sodium phosphates

Test Methods. Several methods are employed to measure crystallinity.

FT – IR spectroscopy is the simplest way of determining the degree of crystallinity, (i.e., the mass fraction of the crystalline phase) [135].

Differential scanning calorimetry is used to measure the crystallization rate and the temperature at which the maximum crystallization rate occurs [136].

Polarization microscopy is useful for measuring spherulite size. Wide-angle X-ray diffraction is used for more detailed studies, for example, of allomorphic forms.

Uses. Nucleating agents are used mainly in polyamide 6, PETP, polyethylene, and polypropylene. Product applications include packaging films, beverage containers, video cassettes, and automotive parts. They are added at a level of 1000 – 5000 ppm and incorporated by dry blending, at the polycondensation stage, or in masterbatch form.

Shortening the injection-molding cycle time is a primary consideration for PETP; inorganic additives, waxes, or sodium salts are used. The situation is similar for polyamide 6, in which many mechanical properties can be improved. Sodium benzoate, aluminum benzoate, phenyl phosphates, and dibenzylidene sorbitols (clarifiers) are employed with polyolefins.

9.2. Antistatic Agents [137]

Plastic surfaces can become electrostatically charged through friction. Charge can build up during processing or service, for example, by the friction of air against the plastic surface. This has several disadvantages:

1) Dirty surface due to attracted dust (a problem with furniture, records, packaging, bottles, and cassettes)
2) Electric shock from walking on plastic floor coverings
3) Adhesion problems in processing films, fibers, foams, and powders
4) Functional disorders in electronic devices and computers
5) Spark generation and dust explosions

Surface-active antistatic agents form a conductive surface layer, in the simplest case a thin film of water, that allows the charge to dissipate quickly or prevents it from building up. In some cases, other additives must be used to impart volume conductivity throughout the bulk of the plastic; these are referred to as *conductivity additives*.

Most plastics have surface resistances of $10^{14}-10^{16}\,\Omega$, which can be lowered to $10^{8}-10^{10}\,\Omega$ by antistatic agents utilizing the water-film mechanism. Certain applications, such as explosion hazard areas or electromagnetic shielding, call for even lower surface resistance or lower volume resistivity; conductivity additives can be used to obtain volume resistivities of $10^{1}-10^{-2}\,\Omega\cdot\mathrm{cm}$.

Quality criteria for antistatic agents include the permanence of their effect, thermal stability, nonvolatility, color, product form, interaction with stabilizers (especially in PVC), and surface effects (tack, weldability, printability). Approval for use as an indirect food additive is often required.

Surface-active antistatic agents are applied by solution spraying (external antistatic agents, nonpermanent) or incorporated into the plastic mass (internal antistatic agents). They are used at levels of 0.1 – 2.5 %, reaching 5 – 7 % in flexible PVC. Internal antistatic agents are amphoteric compounds with limited polymer compatibility (i.e., they tend to migrate to the surface, where their polar moieties "attract" and build up a water-film). Their rate of action depends on their diffusion rate, which is determined by their solubility, molecular mass, and structure, as well as on polymer type and morphology and the presence of other additives and fillers. Internal antistatic agents have a degree of permanence because when they are removed from the surface they are replaced by migration from the reservoir in the bulk plastic.

Surface-active antistatic agents may be cationic, anionic, and nonionic, or organometallic.

Cationic antistatic agents include quaternary ammonium salts and sulfonium or phosphonium salts, mainly chlorides, nitrates, hydrogen phosphates, or 4-toluenesulfonates with long hydrocarbon chains. Products derived from imidazoline and pyrazoline are also used.

These agents are employed in rigid PVC (although they may impair heat stabilization) and in polystyrene. Cationic antistatic agents are generally not approved for food contact and are relatively expensive.

Anionic antistatic agents include sodium alkyl sulfonates, alkyl phosphonates, and alkyl dithiocarbamates. They are used in rigid PVC (with less detrimental effect on heat stabilization).

Nonionic antistatic agents include ethoxylated (or propoxylated) fatty alcohols, fatty amines, or fatty acid amides: polyethylene glycol esters of fatty acids and alkylphenols; glyceryl esters of fatty acids; and sorbitol esters. They are used for many polymers, in particular polyolefins, flexible PVC, polystyrene, and ABS. Most of these agents have a liquid or waxy consistency.

Organometalic compounds include neoalkyl titanates and zirconates [138] that have a high thermal stability and low migration; their effectiveness is less influenced by humidity. They are especially recommended for polyolefins, but also for PVC and polystyrene.

Conductivity additives reduce the resistivity in a threshold fashion: when their concentration reaches a certain level (the percolation point), conductivity increases abruptly [139]. They are used at levels of 5 – 10 % and occasionally up to 20 %. Examples are carbon black, metal powders or fibers (e.g., Al or Cu), silver-coated silicates, nickel-coated chopped carbon

or glass fibers, or highly conjugated organic polymers (e.g., polyacetylenes, polypyrroles, polythiophenes, and polyanilines), which are oxidized or reduced (doped) to form semiconducting polymers. A survey of conductive additives is given in [140].

Test Methods. The most important test methods measure the surface or volume resistivity (e.g., ASTM D 257 or DIN 53 482, respectively). The frictional charging tendency and the half-life of the accumulated charge (in seconds) can also be measured (DIN 53 486 E).

Empirical soiling tests (e.g., with cigarette ash) should also be mentioned. All testing should be performed under a controlled atmosphere (e.g., 22 °C and 50 % R.H.).

Economic Aspects. The annual worldwide consumption of surface-active antistatic agents is ca. 10 000 t; ethoxylated fatty amines are the leading group in terms of volume. Polyolefins, PVC, and polystyrene use more than 90 % of these products.

Important producers (and trade names) include Aceto Chem. (Catafor), Akzo (Armostat, Interstat), American Cyanamid (Cyastat), CWM Bärlocher (Bärostat), Bayer (Statexan), Dr. Th. Böhme (Tebestat), Chemax Inc. (Chemstat), Dai-ichi Kogyo Seiyaku (Resistat), Dow Chemical (Stature), Eastman (Myvaplex) B. F. Goodrich (StatRite), Henkel (Dehydat), Hoechst (Hostastat), ICI (Atmer), Kao Chem., Kenrich Petrochem. (Ken-Stat), Lonza (Glycolube), Marubishi Oil, Mazer Chemical Division of PPG (Larostat), Nippon Oil & Fats, Sandoz (Sandin), Sherex (Varstat), Swedstab (Swedstat), and Witco (Markstat, Kenamine).

Many producers are in the market with antistatic master batches that contain 10 – 50 % of the active ingredient in polymer vehicles, making the usually hygroscopic compounds easier to process.

9.3. Impact Modifiers

High-tonnage plastics such as PVC, polyolefins, or polystyrene have high rigidity but are brittle. Additives (impact modifiers) must therefore often be used to improve impact strength, especially at low temperature [141], [142]. In contrast to plasticizers, impact modifiers must not reduce but increase the heat distortion temperature. Other important selection criteria are their effect on weather resistance and transparency.

Impact modifiers are elastomeric copolymers with low glass transition temperatures. They are dispersed as discrete soft phases in the thermoplastic. Transfer of impact energy to the elastomer phases requires not only a good distribution but also adhesion at the interface between the two phases by chemical bonding or physical cross-linking. Elastomeric properties and adhesion dictate the structural principle of impact modifiers, which are naturally very substrate specific. Most such agents become active above a certain concentration (transition point), between 5 and 15 %.

Classification. Impact modifiers are frequently incorporated into the polymer during compounding; in PVC, they are blended in by the converter. The most important impact modifier classes follow.

Methyl Methacrylate – Butadiene – Styrene Copolymers (MBS). The MBS agents are prepared by graft polymerization of styrene and methyl methacrylate on polybutadiene. They are used chiefly in PVC, where they offer advantages in clear products (bottles) and, to a lesser extent, in polycarbonates (automotive parts). Because of their sensitivity to light, they are less suitable for outdoor service.

Acrylates include weatherfast polyacrylate esters based on poly(butyl acrylate) and poly(2-ethylhexyl acrylate), as well as skin – core graft copolymers of butyl acrylate and methyl methacrylate. In addition to pure polyacrylates ("all acrylics"), butyl acrylate – styrene – methyl methacrylate terpolymers are used which have good transparency and weather resistance. Acrylate impact modifiers are employed mainly in PVC (e.g., pipes, window frames, sidings). Polyacrylates functionalized with carboxyl groups are recommended for polyamides.

Graft Copolymers of Acrylonitrile and Styrene on Polybutadiene (ABS). ABS is a classical impact modifier for PVC (profiles, sheets, pipes, housings, bottles). The transparency of articles produced depends on the composition. ABS is less suitable for outdoor applications.

Chlorinated Polyethylene. Polyethylene containing 30 – 40 % chlorine is employed in PVC, lending high impact strength especially at low temperature.

Ethylene – propylene – diene copolymers (EPDM) are used in polypropylene, for example, in automotive parts (bumpers, instrument panels). Functionalized derivatives have been developed for engineering thermoplastics (polyamides, PBT, PETP), for example, by grafting on acrylic acid, maleic acid, or maleic anhydride. The carboxyl groups react with the end groups of the polymers.

Styrene – Butadiene Rubber. Finely divided styrene – butadiene rubber in polystyrene gives high-impact products with a better transparency than high-impact polystyrene (HIPS) produced by graft polymerization of styrene on polybutadiene.

Other impact modifiers include ethylene – vinyl acetate, acrylonitrile – styrene – acrylate, and styrene – maleic anhydride copolymers. New functionalized products are being developed for special applications, including those in the thermoset area.

Test Methods. Impact strength tests are specified in the following standards: ASTM D 256, ISO R 179, DIN 53 453. In PVC, for example, modifiers increase the impact value by a factor between 10 and 30.

Economic Aspects. Annual demand for impact modifiers is ca. 150 000 t; ca. 60 % of this is used in PVC. The most important class is MBS, followed by ethylene – propylene – diene copolymers, acrylates, and ABS.

Important producers (and trade names) include Arco Chem. (Arvyl), BASF (Vinuran), Bayer (Baymod, Novadur, Levapren), B. F. Goodrich, Dow Chemical, DSM (Keltan), Du Pont (Elvaloy), Exxon Chemical (Vistalon), General Electric Specialty Chemicals (Blendex), Hoechst (Hostapren), Hüls (Vestolit), Kanegafuchi Chem./Kaneka Texas/Kaneka Belgium (Kane Ace), Kureha Chem., M.A. Industries (MA),Metco America/Metablen b.V.(Metablen, Durastrength), Mitsubishi Rayon, Mobay (Baymod), Monsanto (Elix, Cadon), Nippon Gohsei (Soarblen), Nova Polymers (Novalar, Novalene), Polysar (Taktene), Rohm & Haas (Acryloid, Paraloid), Shell (Kraton), and Wacker Chemie (Vinnol).

9.4. Chemical Blowing Agents

Plastics are often converted into foamed products with physical or chemical blowing agents to obtain substantial savings in weight and improve insulating properties (thermal and acoustic). See also → Foamed Plastics. Physical agents include permanent gases (N_2, CO_2, air) and substances that are gases at the processing temperature (e.g., chlorofluorocarbons, dichloromethane, $C_5 - C_7$ hydrocarbons). Chemical blowing agents are usually insoluble in the polymer and undergo thermal decomposition or chemical reactions during processing to give gaseous products [143]. While there are technological limits to the use of permanent gases, chlorofluorocarbons have come under severe criticism because of their potential for depleting the ozone layer (Montreal Protocol, October 1987).

The most important criteria for chemical blowing agents are

1) Their decomposition temperature must be matched to the processing temperature of the plastic (150 – 250 °C). Decomposition should not take place spontaneously, should be only slightly exo- or endothermic, and should occur within a relatively narrow temperature window (5 – 15 °C).
2) The gas should be produced with a high yield (usually 100 – 225 mL/g) and should not be toxic, flammable, or corrosive.
3) The solid decomposition products should not interfere with processing (plate-out, migration, or discoloration).
4) They should have a small particle size, a narrow particle-size distribution, and be easily dispersed in the polymer.
5) Approval as an indirect food additive is often required.

Classification. The following are the most important classes of chemical blowing agents.

Azodicarbonamide [*123-77-3*], thermal decomposition 205 – 215 °C, gas yield 220 mL/g, gases produced N_2, CO (ca. 2 : 1), some NH_3, CO_2. The decomposition temperature can be lowered to 155 °C with "kickers", which are either metal compounds (e.g.,

ZnO, zinc stearate, Ba–Zn and K–Zn systems, and lead salts) or organic substances (e.g., acids, bases, urea). Azodicarbonamide is used in PVC, polyolefins, polystyrene, ABS, polyamides, poly(phenylene oxide), acrylates, and other resins.

4,4′-Oxybis(benzenesulfohydrazide) [*80-51-3*], thermal decomposition 150–160 °C, gas yield 125 mL/g, gases produced N_2, H_2O. This compound is suitable for plastics that are foamed at relatively low temperature: PVC, LDPE, ethylene–vinyl acetate copolymer, and polystyrene. In comparison with azodicarbonamide it has the advantage of colorlessness.

Cyanuric trihydrazide [*10105-42-7*], trihydrazinotriazine, thermal decomposition 275 °C, gas yield 225 mL/g, gases produced N_2, NH_3. This compound is used in ABS, polypropylene, and polyamides.

p-Toluenesulfonyl semicarbazide [*10396-10-8*], thermal decomposition 220–245 °C, gas yield 140 mL/g, gases produced N_2, CO_2. This compound is used in ABS, polypropylene, HDPE, polyamides, polystyrene, rigid PVC, and PBT.

Isatoic anhydride [*118-48-9*], thermal decomposition 210–225 °C, gas yield 115 mL/g, gas produced CO_2. This compound is used in polystyrene, ABS, poly(phenylene oxide), PBT, polycarbonates, and polyamides.

5-Phenyltetrazole [*18039-42-4*], thermal decomposition 240–280 °C, gas yield 200 mL/g, gas produced N_2. This compound is used in high-temperature processing, especially in polycarbonates, PBT, poly(phenylene oxide), and polyamides.

Citric acid and sodium bicarbonate, thermal decomposition 150–200 °C, gas produced CO_2. In contrast to the compounds listed above, this system generates blowing gases by an endothermic reaction. It can be used in polystyrene, polypropylene, ABS, PVC, and other plastics.

Inorganic carbonates decompose to form CO_2 and are used in PVC, polystyrene, and polyolefins.

Most chemical blowing agents (ca. 60 %) are used in PVC, followed by polyolefins, including cross-linked polyethylene (ca. 30 %), polystyrene, and engineering thermoplastics. Polyurethanes and polystyrene are foamed mainly with physical blowing agents.

Chemical blowing agents are used at a concentration of 1–3 %, higher in cross-linked polyethylene. They are employed as powders or in other forms [144], including nondusting granules, master batches, or pastes in combination with plasticizers, kickers, and other additives. The blowing agent can be injected into polymer melts or preblended with the polymer granules and the compound fed to extrusion or injection molding.

Economic Aspects. The annual consumption of chemical blowing agents is ca. 15 000 t; azodicarbonamide is by far the most important product.

Important producers (and trade names) include Bayer/Mobay (Porofor), Boehringer Ingelheim (Hydrocerol), Dong Jim Chem. (Unicell), Eiwa Chem. (Neocellborn, Vinyfor), Fairmount Chem. (Azocel), Hebron (Vinstab, Hebron), Hoechst (Hostatron), J. M. Huber (Activex), Kum Yang (Cellcom), Lucidol/Pennwalt (Luperfoam), M & T (Azubol), Otsuka Chem. (Unifoam), Sankyo Kasei (Cellmic), Schering/Sherex (Ficel, Genitron), Toyo Hydrazine Ind. (Azobis), and Uniroyal/Olin (Celogen, Expandex, Kempore).

9.5. Optical Brighteners

Optical brighteners [145] enhance the whiteness and brilliance of plastics that have a slightly yellow intrinsic color (→ Optical Brighteners). They absorb UV light and emit part of it as fluoresence in the blue–violet region of the spectrum after 10^{-7}–10^{-9} s.

Important criteria for optical brighteners are the hue of the emitted light (bluish, greenish, or reddish) and their lightfastness. They must also be soluble in the polymer substrate, thermally stable during processing, and resistant to migration.

Optical brighteners have conjugated systems based on stilbene, benzoxazole, coumarin, thiophene, oxadiazole, naphthalene, or triazine building blocks.

They are used in many plastics, at levels between 0.005 and 0.1 %; formulations with TiO_2, especially rutile, require the higher levels. Overdosing produces a greenish tinge. Optical bright-

eners are also used as concentrates in fillers or plasticizers, or as master batches.

Producers (and trade names) of optical brighteners suitable for use as plastics additives include BASF (Ultraphor), Bayer (Blankophor), Ciba-Geigy (Uvitex, Tinopal), Eastman Chemicals (Eastobrite, Kodel), Hoechst (Hostalux), ICI (Fluolite), Sandoz (Leucopur), Sigma (Optiblanc), and Sumitomo (Whitex, Whitefluor).

10. References

1. R. Gächter, H. Müller: *Kunststoffadditive,* 3rd ed., Carl Hanser Verlag, München – Wien 1989.
2. J. T. Lutz: *Thermoplastic Polymer Additives,* Marcel Dekker, New York 1989.
3. *Additives for Polymers,* Elsevier, Oxford, published monthly.
4. E. L. Shanina, V. A. Belyakov, G. E. Zaikov, *Polym. Degr. Stab.* **27** (1990) 309.
5. K. Kikkawa, Y. Nakahara, Y. Ohkatsu, *Polym. Degr. Stab.* **18** (1987) 237.
6. T. J. Henman: *World Index of Polyolefin Stabilizers,* Kogan Page, London 1982.
7. BASF, US 4 806 580, 1987 (G. Bock, H. Trauth, W. Weber, P. Lechtken).
8. F. Nabholz, M. D. Castle: *International Conference on Advances in the Stabilization and Controlled Degradation of Polymers,* Luzern 1989.
9. Adeka Argus Chemical Co., EP 336 606, 1989 (K. Nishikawa, T. Haruna, M. Hamajima).
10. Ethyl Corp., EP 280 938, 1988 (L. P. J. Burton).
11. Sandoz AG, US 4 075 163, 1976 (K. Hofer, G. Tscheulin).
12. S. Moss, H. Zweifel, *Polym. Degr. Stab.* **25** (1989) 217.
13. G. Scott (ed.): *Developments in Polymer Stabilisation,* **vols. 1 – 8,** Elsevier Applied Science, London – New York 1979 – 1987.
14. N. S. Allen (ed.): *Degradation and Stabilisation of Polyolefins,* Elsevier Applied Science Publishers, London – New York 1983.
15. N. Grassie, G. Scott: *Polymer Degradation and Stabilisation,* Cambridge University Press, Cambridge 1985.
16. P. P. Klemchuk, P. L. Horng, *Polym. Degr. Stab.* **7** (1984) 131.
17. G. Scott, *Chem. Ind. (London)* 1987, 841.
18. J. Pospisil, *Polym. Degr. Stab.* **20** (1988) 181.
19. N. S. Allen, J. T. Kotecha, J. L. Gardette, J. Lemaire, *Polym. Degr. Stab.* **11** (1985) 181.
20. D. Vyprachticky, J. Pospisil, J. Sedlar, *Polym. Degr. Stab.* **27** (1990) 227.
21. F. Gugumus, *Polym. Degr. Stab.* **24** (1989) 289.
22. E. Blatt, H. J. Griesser, J. H. Hodgkin, A. W. H. Mau, *Polym. Degr. Stab.* **25** (1989) 19.
23. Z. Osawa, *Polym. Degr. Stab.* **20** (1988) 203.
24. G. Menges, W. Michaeli, C. Schwenzer, L. Czyborra, *Plastverarbeiter* **40** (1989) 207.
25. E. M. Lewis, *Mod. Plast. Encyclopedia* **67** (1990) 165.
26. J. G. Calvert, J. N. Pitts: *Photochemistry,* Wiley, New York 1967, p. 534.
27. H. E. A. Kramer, *Farbe+Lack* **92** (1986) 919.
28. M. Allan, et al., *Polym. Degr. Stab.* **15** (1986) 311.
29. N. S. Allen, *Chem. Soc. Rev.* **15** (1986) 373.
30. N. S. Allen, A. Parkinson, F. F. Loffelmann, P. V. Susi, *Angew. Makromol. Chem.* **116** (1983) 203.
31. Sankyo KK, JP 7 1 31 734, 1968 (K. Murayama).
32. D. J. Carlsson, D. W. Grattan, T. Suprunchuk, D. M. Wiles, *J. Appl. Polym. Sci.* **22** (1978) 2217.
33. E. T. Denisov in [13]vol. 3 (1980) p. 1; vol. 5 (1982) p. 37.
34. D. M. Wiles, D. J. Carlsson, *Polym. Degr. Stab.* **3** (1980) 61; **6** (1984) 1.
35. N. S. Allen, *Polym. Degr. Stab.* **2** (1980) 129.
36. N. S. Allen, J. L. Gardette, J. Lemaire, *Polym. Degr. Stab.* **3** (1981) 199.
37. R. Bagheri, K. B. Chakraboty, G. Scott, *Polym. Degr. Stab.* **4** (1982) 1.
38. S. Al-Malaika, G. Scott in [14] p. 283.
39. N. S. Allen, J. L. Gardette, J. Lemaire, *Polym. Degr. Stab.* **8** (1984) 133.
40. G. Scott, *Polym. Degr. Stab.* **10** (1985) 97.
41. P. P. Klemchuk, M. E. Gande, *Polym. Degr. Stab.* **22** (1988) 241.
42. E. T. Denisov, *Polym. Degr. Stab.* **25** (1989) 209.
43. A. J. Chirinos Padron, *Rev. Macromol. Chem. Phys.* **C 30** (1990) 107.
44. P. P. Klemchuk, M. E. Gande, *Makromol. Chem. Makromol. Symp.* **28** (1989) 117.
45. K. H. Chan, D. J. Carlsson, M. D. Wiles, *J. Polym. Sci. Polym. Lett. Ed.* **18** (1980) 607.
46. F. K. Meyer, F. Gugumus, E. Pedrazzetti, *Chemiefasern/Textilind.* **35** (1985) 840.
47. F. J. Barahona, J. M. G. Vasquez, *Plasticulture* **65** (1985) 3.
48. A. J. Chirinos-Padron, *Polym. Degr. Stab.* **25** (1989) 101.

49. B. F. Goodrich, EP 258 598, 1987 (J. T.-Y. Lai, P. N. Son).
50. G. Scott, *Makromol. Chem. Makromol. Symp.* **28** (1989) 59.
51. F. Gugumus in [13], vol. 8 (1987).
52. D. Kockott, *Polym. Degr. Stab.* **25** (1989) 181.
53. N. S. Allen, A. Chirinos-Padron, T. J. Henman, *Polym. Degr. Stab.* **13** (1985) 31.
54. T. Kurumada, *Polym. Degr. Stab.* **19** (1987) 263.
55. F. Stohler, K. Berger, *Angew. Makromol. Chem.* **158/159** (1988) 233.
56. Röhm GmbH, EP 16 870, 1979 (L. Hosch).
57. K. S. Minsker, S. V. Kolesov, G. E. Zaikov: *Degradation and Stabilisation of Vinyl Chloride Based Polymers,* Pergamon Press, Oxford 1988.
58. E. D. Owen: *Degradation and Stabilisation of PVC,* Elsevier Applied Science Publishers, London 1984.
59. B. Y. K. Ho, *J. Vinyl Technol.* **6** (1984) 162.
60. R. Schlimper, *Plaste Kautsch.* **14** (1967) 657.
61. A. F. Frye, R. W. Horst, *J. Polym. Sci.* **40** (1959) 419; **45** (1960) 1; *J. Polym. Sci.,* **Part A 2** (1964) 1765, 1785, 1801.
62. H. O. Wirth, H. A. Müller, W. Wehner, *J. Vinyl Technol.* **1** (1979) 51.
63. D. Braun, D. Hepp, *Angew. Makromol. Chem.* **66** (1976) 23.
64. T. van Hoang, A. Michel, A. Guyot, *Polym. Degr. Stab.* **9** (1984) 73, 89.
65. A. Michel, T. van Hoang, B. Perrin, M. France-Llauro, *Polym. Degr. Stab.* **3** (1981) 107.
66. K. S. Minsker, S. V. Kolesov, V. M. Yanborisov, G. E. Zaikov, *Polym. Degr. Stab.* **15** (1986) 305.
67. J. Büssing in: *Kunststoff Handbuch 2/1, PVC,* 2nd ed., Carl Hanser Verlag, München – Wien 1986, p. 551.
68. T. Riedel in [1], p. 443.
69. K. Worschech in *Kunststoff Handbuch 2/1, PVC,* Carl Hanser Verlag, München 1986.
70. T. A. Skowronski, J. F. Rabek, B. Ranby, *Polym. Degr. Stab.* **12** (1985) 229.
71. G. Pfahler, T. Riedel, *Kunststoffe* **66** (1976) 694.
72. G. Plato, G. Schröter, *Kunststoffe* **50** (1960) 163.
73. J. W. Lyons: *The Chemistry and Uses of Fire Retardants,* Wiley-Interscience, New York 1970.
74. W. C. Kuryla, A. J. Papa: *Flame Retardancy of Polymeric Materials,* **vols. 1 – 5,** Marcel Dekker, New York 1973 – 1979.
75. M. Lewin, S. M. Atlas, E. M. Pearce: *Flame Retardant Polymeric Materials,* **vols. 1 – 2,** Plenum Press, New York – London 1975 – 1978.
76. A. H. Landrock: *Handbook of Plastics Flammability and Combustion Toxicology,* Noyes Publications, Park Ridge 1983.
77. C. F. Cullis, M. M. Hirschler: *The Combustion of Organic Materials,* Clarendon Press, Oxford 1981.
78. J. Troitzsch: *International Plastics Flammability Handbook,* Hanser Publishers, Munich – Vienna – New York 1990.
79. G. Kirschbaum, *Kunststoffe* **79** (1989) 1205.
80. W. J. Kennelly in: *Flame Retardants 90,* Elsevier Applied Science, London – New York 1990, p. 9.
81. G. Camino, L. Costa, G. Martinasso, *Polym. Degr. Stab.* **23** (1989) 359; **28** (1990) 17.
82. H. Staendeke, D. J. Scharf, *Kunststoffe* **79** (1989) 1200.
83. E. N. Peters in [74], vol. 5.
84. L. Costa, G. Camino, M. P. L. di Cortemiglia, *Polym. Degr. Stab.* **14** (1986) 113, 115, 165.
85. P. A. Cusack, A. W. Mark, J. A. Pearce, S. J. Reynolds, *Fire Mater.* **14** (1989) 23.
86. R. L. Markezich, C. S. Ilardo, R. F. Mundhenke in: *Flame Retardants 90,* Elsevier Applied Science, London – New York 1990, p. 88.
87. R. B. Bush, *Plast. Eng.* **42** (1986) April, 29.
88. J. Eichhorn, *J. Appl. Polym. Sci.* **8** (1964) 2497.
89. V. Mark, *Org. Coat. Plast. Chem.* **43** (1980) 71.
90. L. Costa, G. Camino, *Polym. Degr. Stab.* **20** (1988) 271.
91. R. Delobel, M. LeBras, N. Ouassou, F. Alistiqsa, *J. Fire Sci.* **8** (1990) 85.
92. J. J. Pitts, P. H. Scott, D. G. Powell, *J. Cell. Plast.* **6** (1970) 35.
93. D. L. Buzzard in: *Fire Retardants 83,* The Plastics and Rubber Institute, London 1983, p. 21.
94. N. K. Jha, A. C. Misra, P. Bajaj, *Rev. Macromol. Chem. Phys.* **24** (1984) 69.
95. S. Miyata, T. Imahashi, H. Anabuki, *J. Appl. Polym. Sci.* **25** (1980) 415.
96. J. Sutker, *J. Fire Sci.* **1** (1983) 66.
97. J. G. Williams in: *Flame Retardants 83,* The Plastics and Rubber Institute, London 1983, p. 91.
98. R. Wolf, *Kunststoffe* **76** (1986) 943.
99. C. J. Hilado, *Mod. Plast.* **54** (1977) July, 64; **55** (1978) April, 92.

100. F. H. Prager et al., *J. Fire Sci.* **5** (1987) 308.
101. R. C. Kidder in *Flame Retardants 90,* Elsevier Applied Science, London – New York 1990, p. 9.
102. B. Akles, J. Steinmetz, J. Hogan, *Mod. Plast. Int.* **17** (1987) 49.
103. Union Carbide, US 4 481 322, 1984 (R. E. Godlewski, F. D. Osterholtz).
104. S. J. Monte, G. Sugerman; *Ken-React Reference Manual,* Kenrich Petrochemicals, Bayonne, N.J., 1985.
105. L. B. Cohen, *Plast. Eng.* **39** (1983) 29.
106. E. E. Lang in: *Kunststoff Handbuch 2/1, PVC,* 2nd ed., Carl Hanser Verlag, München – Wien 1986, p. 678.
107. D. L. Caldwell, F. M. Cortez, *Mod. Plast.* **65** (1988) Sept., 132.
108. H. P. Schlumpf in *Kunststoffadditive,* 3rd ed., Carl Hanser Verlag, München – Wien 1989, p. 549; R. Kleinholz, G. Heyn, R. Stolze in *Kunststoffadditive,* 3rd ed., Carl Hanser Verlag, München – Wien 1989, p. 617.
109. H. S. Katz, J. V. Milewski: *Handbook of Fillers and Reinforcements of Plastics,* Van Nostrand Reinhold, New York 1978.
110. B. L. Kaul, *J. Oil Colour Chem. Assoc.* **70** (1987) 349.
111. B. F. Greek, *Chem. Eng. News* **66** (1988) 35.
112. D. Spencer, *Mod. Plast.* **67** (1990) April, 19 (suppl. Waste Solution).
113. B. G. Murray in C. T. Patton (ed.): *Pigment Handbook,* **vol. 2,** Wiley-Interscience, New York 1973, p. 277.
114. G. Sonn, *Am. Ink Maker* **65** (1987) June, 54.
115. W. Damm in R. Gächter, H. Müller (eds.): *Plastics Additives Handbook,* 2nd ed., Hanser Publishers, Wien 1987, p. 506.
116. A. P. Hopmeier: *Encyclopedia of Polymer Science and Technology,* **vol. 10,** Wiley-Interscience, New York 1969, p. 157.
117. P. E. Lewis (ed.): *Pigment Handbook,* 2nd ed., vol. 2, Wiley-Interscience, New York 1987.
118. A. S. Wood, *Mod. Plast.* **66** (1989) Sept., 58.
119. W. Herbst, K. Hunger: *Industrial Organic Pigments,* VCH Verlagsgesellschaft, Weinheim 1992.
120. J. D. Sanders: *Pigments for Inkmakers,* SITA Technology, London 1989.
121. *Eur. Plast. News* **17** (1990) June, 44.
122. F. Baebler, Proc. of RETEC Meeting in Huron (USA), Sept. 1989, 205.
123. *Colour Index,* 3rd ed. and suppl., Society of Dyers and Colourists, Bradford 1982.
124. J. Jansen in [1]p. 893.
125. F. L. Binsbergen, *Polymer* **11** (1970) 253, 309.
126. F. L. Binsbergen, *J. Polym. Sci. Polym. Phys. Ed.* **11** (1973) 117.
127. A. M. Chatterjee, F. P. Price, S. Newman, *J. Polym. Sci. Polym. Phys. Ed.* **13** (1975) 2369, 2385, 2391.
128. J. M. Dekoninck, R. Legras, J. P. Mercier, *Polymer* **30** (1989) 910.
129. Adeka Argus, US 4 258 142, 1981 (T. Ohzeki, M. Ahutsu, J. Kawai); US 4 463 113, 1984 (Y. Nakahara, M. Ahutsu, T. Haruna, M. Takahashi).
130. C. C. Carroll, *Mod. Plast.* **61** (1984) Sept., 108.
131. New Japan Chem. Co., JP 53 117 044, 1977 (T. Kobayashi, T. Nobe).
132. L. Bourland, *Plast. Eng.* **43** (1987) July, 39.
133. V. Caldas, G. R. Brown, J. M. Willis, *Macromolecules* **23** (1990) 338.
134. ICI, EP 267 695, 1987 (J. P. Trotoir, C. Bath).
135. S. L. Hsu in: *Comprehensive Polymer Science,* vol. 1, Pergamon Press, Oxford 1989, p. 429.
136. J. Kaiser, *Kunststoffe* **80** (1990) 330.
137. G. Pfahler in [1]p. 779.
138. S. J. Monte, G. Sugerman in [104], suppl., p. 35.
139. B. Wessling, *Kunstst. J.* **21** (1987) 50.
140. P. Mapleston, *Mod. Plast. Int.* **19** (1989) no. 10, 66.
141. D. Hepp in [1], p. 525.
142. G. Menzel in: *Kunststoff Handbuch 2/1, PVC,* Carl Hanser Verlag, München – Wien 1986, p. 712.
143. H. Hurnik in [1], p. 843.
144. R. R. Puri in Abstracts 1st European Conference on High Performance Additives, British Plastics Federation, London 1988, p. 83.
145. K. Berger in [1], p. 807.

Plastics, Analysis

Plastics, Properties and Testing is a separate keyword.

KARL-FRIEDRICH ELGERT, Institut für Textilchemie, Denkendorf, Federal Republic of Germany

1. Introduction 1
2. Polymer Isolation 3
3. Preliminary Tests for Polymer Identification 6
4. Chemical Analysis 7
4.1. Heteroelements and Functional Groups 7
4.2. Separation and Identification Schemes 9
4.3. Spectroscopic Methods 9
5. Molecular Mass Determination . . . 13
5.1. Absolute Methods 14
5.1.1. Membrane Osmometry 14
5.1.2. Lowering Vapor Pressure 15
5.1.3. Ultracentrifugation 16
5.1.4. Light Scattering 17
5.2. Relative Methods 19
5.3. Determination of Molecular Mass Distribution 20
6. Determination of Sequential Structure 26
6.1. NMR Spectroscopy 27
6.2. Pyrolysis Gas Chromatography . . . 30
7. Determination of Chemical Heterogeneity 30
8. References 33

1. Introduction

The polymeric chains of most industrially and economically important plastics are built up from only a small number of elements: mainly C, H, O, N, Cl, F, and Si. Nevertheless, plastics analysis constitutes an extremely complicated and wide-ranging field that requires the application of many chemical and physical analytical methods.

The subject of plastics analysis is treated here in terms specifically of the analysis of polymer chains. Most plastics – that is to say, materials formulated on the basis of synthetic polymers – contain significant amounts of additives such as pigments, plasticizers, stabilizers, and fillers, but analysis of these secondary components constitutes a separate field of study. In certain cases the desirable properties of a particular plastic arise only during the course of subsequent thermal or mechanical treatment associated with shaping or molding. There has also been increasing interest in the determination of polymer degradation products and monomer residues.

The fundamental problem in plastics analysis is the heterogeneity of macromolecular materials with respect to molecular mass and the numerous isomeric possibilities open to a polymeric chain.

For example, vinyl chloride is a gas under normal conditions, but its lower oligomers are liquids, and poly(vinyl chloride) is a solid. Thus the physical, physicochemical, and applications characteristics of a particular batch of poly(vinyl chloride), such as solubility, crystallinity, strength, and elasticity, are a function not only of the structure of the polymeric chain, but also of the degree of polymerization.

Isomerism. The empirical formula obtained by elemental analysis (see Section 4) is generally of little help in the precise identification of a plastic or a polymer. This is due to the constitutional, configurational, and conformational isomerism characteristic of a polymer chain, which is reflected in a structure that can be described at primary through quaternary levels. Thus, an empirical formula such as $(C_4H_8)_n$ is equally valid for achiral polyisobutene and its positional isomer poly(1-butene), which contains a series of pseudoasymmetric carbon atoms (C∗).

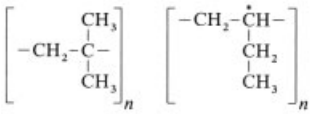

Poly(isobutene) Poly(1-butene)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_509

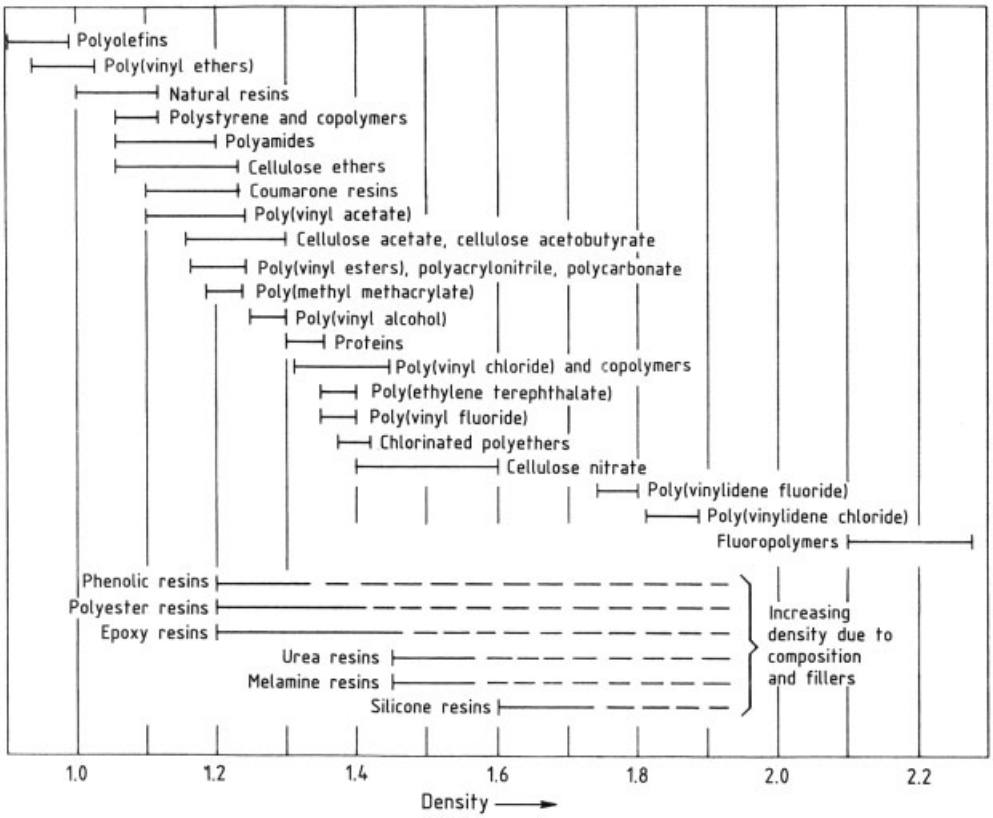

Figure 1. Density of plastics [62]

Polymolecularity in a polymer is another form of constitutional isomerism, which results from the stochastic character of the polymerization reaction but can also arise from decomposition, aging, or cross-linking reactions that occur during processing or utilization. The determination of polymolecularity is discussed in Chapter 5.

Another source of constitutional isomerism may be the way in which monomer units are linked, although head-to-tail linkage usually predominates. An example is poly(1-butene):

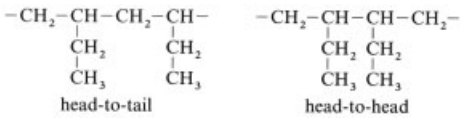

Apart from the nature of polymer end groups and the extent of branching, the sequence of structural units within the polymer chain is also of major importance. Analysis of sequential structure is treated in Section 4.3 and Chapter 6.

Plastics that contain pseudoasymmetric carbon atoms in their polymer chains, such as poly(1-butene), can also exhibit *configurational isomerism*. The sequence of such pseudoasymmetric C atoms has a major effect on properties such as melting point, glass transition temperature, and crystallinity. There is also isomerism of the *cis – trans* type to be considered in the case of unsaturated polydienes, whose structural units include C=C double bonds. The number of possibilities for isomerism – and thus the analytical challenge – increases dramatically on moving from homopolymers to copolymers. Analysis of configurational isomers is the subject of Section 6.1.

As a result of the wide diversity in types of plastic materials and the range of potential problems, it is impossible to describe a single general scheme for polymer analysis. Both the number

of questions that might be raised regarding the nature of a completely unknown polymer and the catalog of applicable analytical methods are very extensive. However, it is often possible to monitor product quality by determining a single structural property by means of a standardized analytical method. The scope and complexity of an analytical investigation is thus a direct function of the nature of the question.

The most superficial description of an unknown plastic must include some characterization of the *primary structure :* chemical composition, molecular mass and molecular mass distribution, and the sequential arrangement of subunits. This leads to the following series of analyses for identifying an unknown plastic:

1) Determination of solubility (Chap. 2)
2) Isolation of macromolecular components (Chap. 2)
3) Preliminary tests for polymer identification (Chap. 3)
4) Chemical analysis (Chap. 4)
5) Molecular mass determination (Chap. 5)
6) Sequence analysis (Chap. 6)
7) In the case of a copolymer, analysis of chemical heterogeneity (Chap. 7)

For a discussion of structure at the secondary through quaternary levels (e.g., conformation, orientation, and crystallinity), see → Plastics, Properties and Testing.

Literature. Plastics analysis is treated in detail in various standard works [1–26]. There also exists an extensive specialized literature [27–63] devoted to specific problems (e.g., analysis of silicones [51], or fiber-reinforced plastics [52], which is further supplemented by tabular works [53] and data from handbooks [54–57]. The review literature should be consulted for more recent developments in plastics analysis, especially with respect to highly specialized questions [58], [59]. For the analysis of selected additives, see [60–63].

2. Polymer Isolation

Solubility is the basis for preliminary separation of a plastic into polymer components and additives. Insoluble fillers and pigments can be removed from a polymer solution by filtration or centrifugation. A plastic that is received in the form of a solution, emulsion, or suspension must first be dried by evaporation of low molecular mass solvents; nonvolatile additives remain in the residue.

A polymeric material that is to be extracted or brought into solution for analytical investigation should be first converted into the most finely divided state possible. Prior freezing of plastic or elastic samples in liquid nitrogen often facilitates size reduction.

Solubility. Polymers and solvents are miscible only if the corresponding free energy of mixing is negative [64], [65]:

$$\Delta G_{\mathrm{cm}^3} = \frac{RT}{V_\mathrm{s}} \left(\Phi_\mathrm{s} \ln \Phi_\mathrm{s} + \frac{\Phi_\mathrm{p}}{m} \ln \Phi_\mathrm{p} + \mu \Phi_\mathrm{s} \Phi_\mathrm{p} \right) \quad (1)$$

G_{cm^3} Gibbs free energy of mixing per cm^3 solution
V_s Molar volume of solvent
Φ Volume fraction (s = solvent; p = polymer)
m Ratio of molar volumes of polymer and solvent
μ Constant

The interaction parameter μ is a function of temperature that takes into account the enthalpy of mixing. Miscibility can be anticipated for any polymer – solvent system in which $\mu < 0.5$. In the case of an amorphous polymer it is possible to ignore, to a first approximation, the change in entropy, which is the justification for focusing here strictly on the enthalpy of mixing. In the expression:

$$\Delta H_{\mathrm{cm}^3} = \Phi_\mathrm{s} \Phi_\mathrm{p} \left[(E_\mathrm{s}/V_\mathrm{s})^{1/2} - (E_\mathrm{p}/V_\mathrm{p})^{1/2} \right]^2 \quad (2)$$

ΔH_{cm^3} Enthalpy of mixing for 1 cm^3 solution
Φ Volume fraction
E Molar cohesion energy
V Molar volume

the term E/V is referred to as the cohesion energy density, from which can be derived the solubility parameter $\delta = \sqrt{E/V}$. The following list gives δ values for various polymers and solvents.

Polymers	
Polyethylene	7.9
Polyisoprene	7.9
Polyisobutylene	8.05
Styrene rubber	8.60
Polybutadiene	8.6
Polystyrene	8.7
Neoprene	9.2
Poly(dimethyl siloxane)	9.4
Nitrile rubber	9.5
Poly(vinyl chloride)	9.7
Poly(methyl methacrylate)	10.2
Cellulose acetate	10.9
Nylon 66	14.5
Poly(vinyl alcohol)	23.4
Solvents	
Heptane	7.43
Diethyl ether	7.70
Cyclohexane	8.25
Carbon tetrachloride	8.62
Benzene	9.21
Chloroform	9.40
Chlorobenzene	9.70
Acetonitrile	9.75
Acetone	9.89
Dichloromethane	10.04
Carbon disulfide	10.10
Dioxane	10.15
Methanol	14.48
Water	23.41

Solubility can be expected when polymer and solvent have comparable solubility parameters. Such δ values provide useful orientation in a search for a suitable solvent, although many exceptions are known, attributable to such factors as molecular mass, degree of branching and cross-linking, configuration, and crystallinity.

Precipitation. A considerable degree of polymer purification can often be achieved by precipitation with a precipitating agent, a process that may need to be repeated several times. However, some oligomers may remain in solution, as a result of which precipitation may cause a change in the molecular mass distribution.

If a plastic and a solvent have different δ values the plastic will be insoluble in that solvent. Therefore, addition of such a solvent to a polymer already dissolved in another solvent should lead to precipitation. Methanol is a poor solvent for most polymers but miscible with many other organic solvents, so it is frequently used as a precipitating agent. Generally, the lower the molecular mass of a plastic, and the more atactic or amorphous it is, the higher its solubility. Raising the temperature also leads to an increase in solubility. Appropriate solvents and precipitating agents have been established empirically for a wide range of plastics; for a tabular summary see [53].

If precipitation proves unsuccessful for the removal of additives it is usually necessary to resort to preparative chromatography [66–68].

Extraction. Plastics can be extracted with organic solvents either in a Soxhlet apparatus or in a high-pressure extractor with supercritical solvents such as CO_2. Extraction is the method of choice for pretreatment of an insoluble plastic. A prerequisite to success is finding an extractant that will dissolve the additives at the temperature of operation but act as a nonsolvent for polymeric components. If necessary, several extractants can be utilized in sequence, although inorganic fillers or pigments that are insoluble in organic solvents will always remain in the sample after extraction (see Table 1).

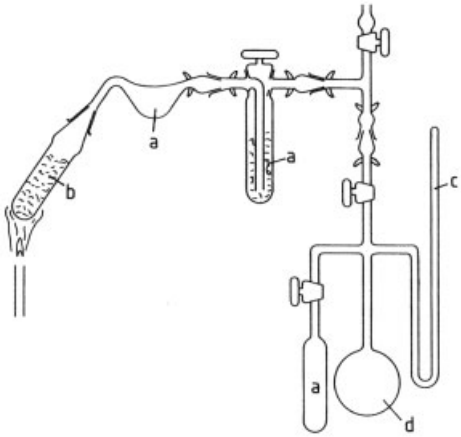

Figure 2. Pyrolysis of plastics by vacuum distillation [3]
a) Condensate traps; b) Sample; c) Manometer; d) IR sample cell

3. Preliminary Tests for Polymer Identification

Once a set of macromolecular components has been isolated, characterization can begin on the basis of physical methods. Powerful and rapid spectroscopic procedures are now available for establishing chemical structure, although their successful utilization presupposes access to the required instrumentation and an appropriate data

Table 1. Solubility of common plastics and polymers [14] ∗

Plastic/polymer	Soluble in	Insoluble in
Alkyd resin	chlorinated hydrocarbons, lower alcohols, esters	hydrocarbons
Amine – formaldehyde resin, cured	benzylamine (60 °C), ammonia	
Methyl cellulose	water, dilute sodium hydroxide, dichloromethane, methanol	acetone, ethanol
Cellulose esters	ketones, esters	aliphatic hydrocarbons, water
Cellulose nitrate	lower alcohols, acetic esters, ketones	ether, benzene, chlorinated hydrocarbons
Chlorinated rubber	esters, ketones, carbon tetrachloride, tetrahydrofuran	aliphatic hydrocarbons
Chloroprene rubber	toluene, chlorinated hydrocarbons	alcohols
Polychlorotrifluoroethylene	hot fluorinated solvents	all common solvents
Poly(vinyl chloride)	dimethyl formamide, tetrahydrofuran, cyclohexanone	alcohols, butyl acetate, dioxane, hydrocarbons
Chlorinated poly(vinyl chloride)	dichloromethane, cyclohexane, benzene, tetrachloroethylene	
Poly(vinylidene chloride)	tetrahydrofuran, ketones, butyl acetate, dimethylformamide (hot), chlorobenzene	alcohols, hydrocarbons
Acrylonitrile – butadiene – styrene copolymer	dichloromethane	alcohols, water, aliphatic hydrocarbons
Styrene – butadiene copolymer	ethyl acetate, benzene, dichloromethane	alcohols, water
Rubber	chlorinated and aromatic hydrocarbons	oxygen-containing solvents
Polytetrafluoroethylene	fluorocarbon oil (hot)	all solvents
Poly(vinyl fluoride)	above 110 °C: cyclohexanone, propylene carbonate, dimethyl sulfoxide, dimethylformamide	
Poly(vinylidene fluoride)	dimethyl sulfoxide, dioxane	
Phenolic resin	alcohol, ketones	chlorinated hydrocarbons, aliphatic hydrocarbons
Polyacrylamide	water	alcohols, esters, hydrocarbons
Polyacrylonitrile	dimethylformamide, butyrolactone nitrophenol, dimethyl sulfoxide, mineral acids	alcohols, esters, ketones, formic acid, hydrocarbons
Polyacrylates	aromatic hydrocarbons, esters, chlorinated hydrocarbons, acetone, tetrahydrofuran	aliphatic hydrocarbons
Polymethacrylates	aromatic hydrocarbons, dioxane, chlorinated hydrocarbons, esters, ketones	ether, alcohols, aliphatic hydrocarbons
Polyamides	phenols, formic acid, conc. mineral acids	alcohols, esters, hydrocarbons
Polybutadiene	aromatic hydrocarbons, cyclohexane, dibutyl ether	alcohols, esters
Polycarbonates	chlorinated hydrocarbons, dioxane, cyclohexanone	alcohols, aliphatic hydrocarbons, water
Polyesters, unsaturated, uncured	ketones, styrene, acrylic esters	aliphatic hydrocarbons
Poly(ethylene terephthalate)	cresol, conc. sulfuric acid, chlorophenol	
Polyethylene	dichloroethylene, tetralin, hot hydrocarbons	polar solvents, alcohols, esters
Poly(ethylene glycol)	chlorinated hydrocarbons, alcohols, water	aliphatic hydrocarbons
Polyformaldehyde	hot solvents, phenols, benzyl alcohol, dimethylformamide	alcohols, ketones, esters, hydrocarbons
Polyisoprene	benzene	alcohols, ketones, esters, hydrocarbons
Polypropylene	at high temperature: aromatic and chlorinated hydrocarbons, tetralin	alcohols, esters, cyclohexanone
Polystyrene	aromatic and chlorinated hyrocarbons, pyridine, ethyl acetate, methyl ethyl ketone, dioxane, tetralin	alcohols, water, aliphatic hydrocarbons
Polyurethane	tetrahydrofuran, pyridine, dimethylformamide, formic acid, dimethyl sulfoxide	ether, alcohols, benzene, water, hydrogen chloride (6 N)
Poly(vinyl acetal)	esters, ketones, tetrahydrofuran	methanol, aliphatic hydrocarbons
Poly(vinyl formal)	dichloroethane, dioxane, glacial acetic acid, phenols	aliphatic hydrocarbons
Poly(vinyl acetate)	aromatic and chlorinated hydrocarbons, acetone, methanol, esters	aliphatic hydrocarbons
Poly(vinyl alcohol)	formamide, water	ether, alcohols, esters, ketones, aliphatic and aromatic hydrocarbons

∗ For comprehensive table, see [14].

Table 2. Flammability of plastics and polymers [14]

Nonflammable	Self-extinguishing	Flammable
Polychlorotrifluoroethylene	poly(vinyl chloride)	polyolefins
Polytetrafluoroethylene	vinyl chloride copolymers	polystyrene
Chlorinated poly(vinyl chloride)	poly(vinyl fluoride)	copolymers of styrene
Urea resins	polyvinylcarbazole	poly(vinyl ethers)
Melamine resins	polyvinylpyrrolidone	poly(vinyl esters)
Silicone resins	polyacrylonitrile	poly(vinyl alcohol)
	filled phenolic, epoxy, and polyester resins	polyacrylates
		polyamides
		polycarbonates
		poly(ethylene terephthalate)
		polyurethanes
		cellulose
		alkyd, coumarone, phenolic, epoxy resins

base. In the absence of such information it may still be possible to narrow the range of structural possibilities in relatively simple ways.

Analysis is most straightforward for soluble polymers. A material consisting of a mixture of various polymers must first be separated into its macromolecular components, preferably by preparative chromatography [66–71] or precipitation. Such a separation may prove problematic in the case of a copolymer, since separability then becomes a function of heterogeneity with respect to both molecular mass and chemical composition of the polymer molecules (see Chap. 7).

Density. Density measurements are conducted either with a pycnometer or on the basis of buoyancy. A rough determination of density can be made by introducing an unknown sample into a series of nonsolvents of known density. By using Figure 1, the nature of the material can then be established with a reasonably high degree of certainty. Crystalline polymers are generally denser than amorphous materials. Density can also be affected by the presence of fillers. Density measurements on plastic materials that contain entrapped gases (e.g., foams) must be preceded by melting or precipitation.

Combustion Characteristics. The combustibility of a plastic can be investigated in the flame of a Bunsen burner. Apart from distinguishing between flammable and inflammable materials, it is also important to observe whether a substance burns only while it is in the flame or if burning persists outside the flame as well. Table 2 provides an overview of flammability characteristics for various plastics.

Thermal Degradation. Pyrolysis behavior and the characteristics of a pyrolysate give further information on the nature of an unknown plastic. The sample is heated to redness in a pyrolyzer (Fig. 2). Volatile products are either condensed or collected in a cotton wool filter. Approximately 100 mg of sample is sufficient for this analysis.

The nature and odor of a pyrolysate can be characteristic for a particular polymer. Highly volatile pyrolysates are produced by polystyrene, polyisobutene, polybutadiene, and polyacrylates. Oily, resinous, or waxy products arise from polyethylene, polypropylene, and poly(1-butene) as well as from copolymers of styrene with butadiene and acrylonitrile. Black carbonaceous residues result from the pyrolysis of cellulose and its derivatives, as well as from poly(vinyl chloride) and resins.

Formation of ammonia suggests a polyamide or a resin based on urea, melamine, aniline, or casein. Pyrolysates with a pungent odor are released by poly(vinyl chloride), polyacrylates, and poly(vinyl formates).

The combination of pyrolysis with hydrogenation of the resulting fragments, followed by gas chromatographic separation and identification,can provide considerable information on the sequential structure of a polymer (see Section 6.2).

Table 3. Chemical analysis of common polymers [14]

Elements (other than C and H)	Polymers
None	polymerized hydrocarbons (e.g., polyethylene, polypropylene, polybutadienes, polystyrene, poly(methyl styrene), poly(vinyl benzene), butadiene – styrene copolymer)
O	Carbohydrates (e.g., celluloses, starches), phenol – formaldehyde resins, phenol – ether resins, phenol – furfural resins, phthalic esters and related compounds, alkyds, poly(vinyl esters), poly(allyl esters), polyacrylic and related esters, polycarbonates, aliphatic polyesters, epoxy resins, aldehyde resins, poly(vinyl alcohol)
N (O)	cellulose nitrate, polyacrylonitrile and copolymers, polyacrylamide, nylons and other polyamides, poly(ethylene imines), urea – formaldehyde resins, melamine – formaldehyde resins, polyurethane, polyester – urethane
Cl (O)	Chlorinated polyhydrocarbons, poly(vinyl chloride), poly(vinylidene chloride), poly(vinyl chloride acetate), polychloroacrylates, chlorinated alkyds
F (O)	polytetrafluoroethylene, poly(vinyl fluoride)
F, Cl	Polychlorotrifluoroethylene
S (O)	alkyl polysulfides, polysulfones, thiophenol condensates
N, S	thiourea – formaldehyde resins
Cl, S	thioplastics, vulcanized neoprenes
N, S, P	casein, protein
Si	silicones
Ti	poly(titanic esters)
B	poly(boric esters)
N, P, Cl	poly(phosphonitrilic halides)

4. Chemical Analysis

4.1. Heteroelements and Functional Groups

Heteroelements. In conjunction with quantitative analysis for the principal elements C and H, the potassium fusion test for heteroelements remains a very important analytical technique [3] (see Table 3). Roughly 100 mg of a finely divided plastic is fused in a test tube with a pea-sized portion of potassium, heated to a red glow, and then quenched in 10 cm^3 of water. The resulting solution is filtered and divided into six equal portions for the identification of Cl, F, N, S, P, and Si.

Chlorine. The filtrate is acidified with HNO_3 and treated with $AgNO_3$ solution; a white precipitate soluble in ammonia indicates chlorine.

Fluorine. After acidification with HCl the solution is treated with $CaCl_2$ solution; a white precipitate of CaF_2 is characteristic of fluorine.

Nitrogen. The solution is treated with a spatula-full of $FeSO_4$, boiled, cooled, and a few drops of $FeCl_3$ solution are added. A precipitate of Prussian blue upon acidification with HCl indicates nitrogen.

Table 4. Saponification number of plastics, resins, and polymers [6]

< 20	100 – 200	> 200
Polyolefins	natural resins	polyacrylates
Methyl cellulose	alkyd resins	polymethacrylate
Polystyrene		poly(vinyl acetate)
Poly(vinyl alcohol)		cellulose esters
Phenol – formaldehyde resins		poly(butylene terephthalate)
Natural and synthetic rubber		polycarbonates
Poly(vinyl chloride)		
Chlorinated poly(vinyl chloride)		
Polyamides		
Natural waxes		

Sulfur. The cold solution is treated with a freshly prepared solution of sodium prussiate. A deep violet color represents a positive test for sulfur.

Phosphorus. Acidification of the solution with HNO_3 and addition of a few drops of ammonium molybdate solution results in a yellow precipitate in the presence of phosphorus.

Silicon. The polymer sample is boiled with a mixture of sulfuric and nitric acids and then evaporated. The residue is pyrolyzed to SiO_2, which is fumed with a mixture of sulfuric and hydrofluoric acids. Appearance of a white precipitate on a moist piece of black paper is evidence of silicon.

Functional Groups. Chemical tests for heteroelements are supplemented by tests for specific functional groups. In addition to establishing the presence of C=C double bonds by Br_2 ad-

dition, determination of the saponification number is useful. For this purpose a known amount of plastic is dissolved in alcohol, benzene, or toluene, and refluxed for 2 – 3 h with a known excess of 1 N KOH. The excess alkali is then determined by back-titration with 1 N HCl. The saponification number is defined as the number of milligrams of KOH consumed per gram of polymer. Many polymers containing no heteroelements other than oxygen can be distinguished on the basis of their saponification numbers, as indicated in Table 4.

A specific color reaction between a plastic and dichloroacetic acid sometimes serves as a rapid and simple means of identification [3].

Table 5. Pyrolysis of polymers [14]

Observations	Tests	Conclusions
Smell of burnt paper	vapors burn with characteristic flame – yellow with little smoke	cellulose
+ smell of acetic acid		cellulose acetate
+ smell of butyrates		cellulose butyrate
Smell of acetic acid	fumes; acidic to indicator paper	acetate polymers
Cyanide		polyacrylonitrile
Smell of burning vegetation	test vapor with damp universal indicator paper; alkaline vapors	polyamides (nylon), polyurethanes, protein-type resins
Smell of formaldehyde		urea – formaldehyde
+ amine-like smell		melamine – formaldehyde
+ phenolic smell		phenol – formaldehyde
Sweetish odor		methacrylate and related compounds
Acrid odor	allow vapors to impinge on a glass rod moistened with silver nitrate solution; cloudy precipitate	poly(vinyl chloride) and its copolymers; natural and synthetic chlorinated rubbers, chloroprene and chlorinated polyolefins
+ heavy black ash	allow vapors to impinge on a glass rod moistened with silver nitrate solution; cloudy precipitate	probably poly(vinylidene chloride)
Sulfurous	allow vapors to impinge on a glass rod moistened with barium chloride solution; cloudy precipitate	sulfur-containing rubbers, etc.

4.2. Separation and Identification Schemes

A simple identification scheme based on polymer pyrolysis is illustrated in Table 5.

Figure 3 shows a more elaborate scheme that uses a set of solubility tests, and combines pyrolysis with both thin-layer chromatography and chemical identification. Specific tests applied to the pyrolyzate result in identification of a particular plastic. In the case of a copolymer, various isomeric possibilities with respect to the principal chain require that chemical analysis be supplemented by physical methods of structure determination. Purely chemical methods may also fail to distinguish clearly between a copolymer and a blend of two homopolymers.

If detailed and unambiguous structural analysis is required spectroscopic methods should be used. Not only do they often allow nondestructive testing of the sample, they also give considerably more information than elemental analysis, particularly with regard to structural subunits and their sequence within the polymer chain (see Sections 4.3 and 6.1).

Table 6. Positions of characteristic IR bands [3]

Polymer	Position of band, μm
Homopolymers	
cis-1,4-Polybutadiene	3.31 – 6.06 – 7.12 – 7.64 – 8.07 – 9.82 – 12.8. . . .12.9 – 13.5 – 14.5
trans-1,4-Polybutadiene	7.48 – 8.09 – 9.49 – 10.3. . .10.4 (double band) – 12.93
1,2-Polybutadiene	5.47 – 6.10 – 7.32 – 8.10 – 10.07 – 11.0 – 12.3. . . .12.4
Copolymers	
Butadiene – styrene	6.12 – 6.7 – 10.8 – 10.35 – 10.99 – 13.2 – 14.3
Butadiene – methyl methacrylate	5.8 – 7.25 – 8.38 – 8.75 – 10.3 – 10.95 – 13.1
Butadiene – acrylonitrile	4.48 – 6.1 – 10.3 – 10.9

4.3. Spectroscopic Methods

Infrared spectroscopy (IR) [72–77] is now a very important tool in the routine analysis of plastics, both for establishing macromolecular composition and for identifying additives.

The IR portion of the electromagnetic spectrum can be divided into three regions:

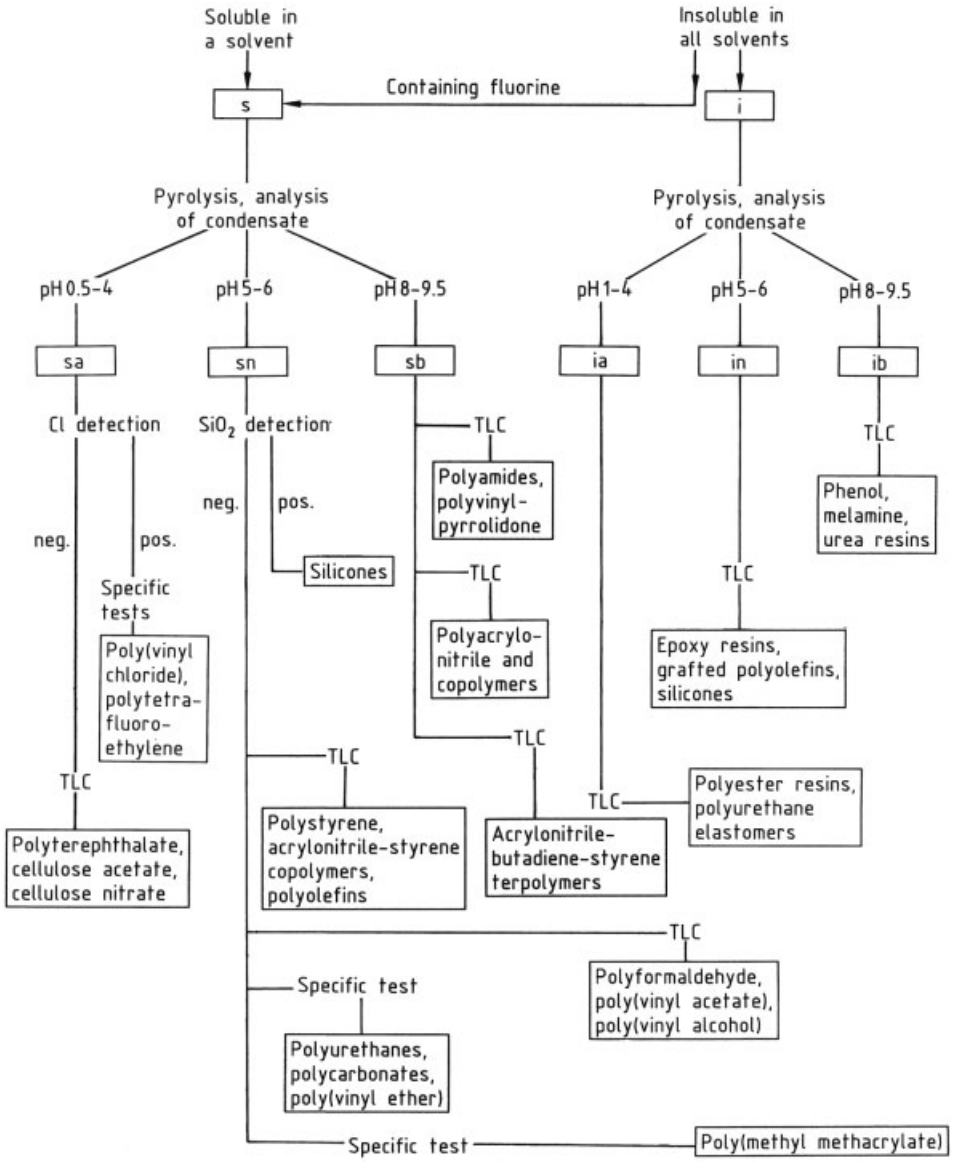

Figure 3. Identification of plastics (simplified; for further details, see [70], [71])
Solubility test with the solvents: water, tetrahydrofuran, dimethylformamide, xylene, formic acid, nitrobenzene
s = soluble; i = insoluble; a = acid; n = neutral; b = basic

Near IR (NIR)	0.76 – 2.5 µm, 13 200 – 4000 cm^{-1}
Mid IR (MIR)	2.5 – 50 µm, 4000 – 200 cm^{-1}
Far IR (FIR)	50 – 1000 µm, 200 – 10 cm^{-1}

Plastics analysis can be performed in all three regions, but MIR is preferred for routine work. Each structural element in a polymer chain has associated with it several absorption lines at various positions in the spectrum, corresponding to the number and type of IR-active vibrational modes (Table 6). For a macromolecule with a degree of polymerization P and a structural unit that consists of N atoms, $3NP - 5 \cong 3PN$ vibrations can be anticipated. However, the number of vibrations actually observed is considerably

Table 7. Chain structure of ethylene – propylene copolymers from IR spectra [3]

Wavelength, µm	Structure	Sequence of monomer units*
12.25	$-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-$	polypropylene units in ht polymerization
13.3	$-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-$	polypropylene units in tt polymerization equivalent to ethylene unit between two propylene units in hh polymerization
13.6	$-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-CH_2-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-$	ethylene unit between two propylene units in ht polymerization
13.7	$-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-CH_2-CH_2-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-$	ethylene unit between two propylene units in tt polymerization two ethylene units between two propylene units in hh polymerization
13.83	$-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-CH_2-CH_2-CH_2-CH_2-\underset{\underset{CH_3}{\vert}}{CH}-CH_2-$	two ethylene units between two propylene units in ht polymerization

*hh = head-to-head; tt = tail-to-tail; ht = head-to-tail.

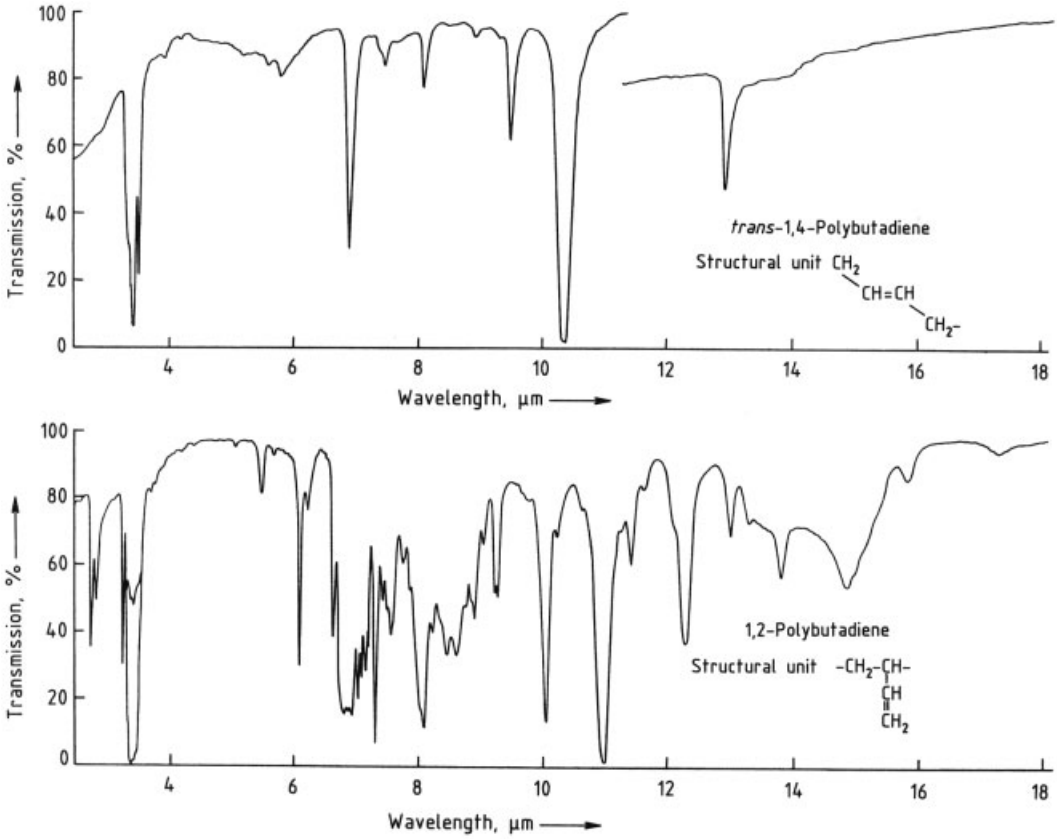

Figure 4. IR spectra of polybutadiene [3]

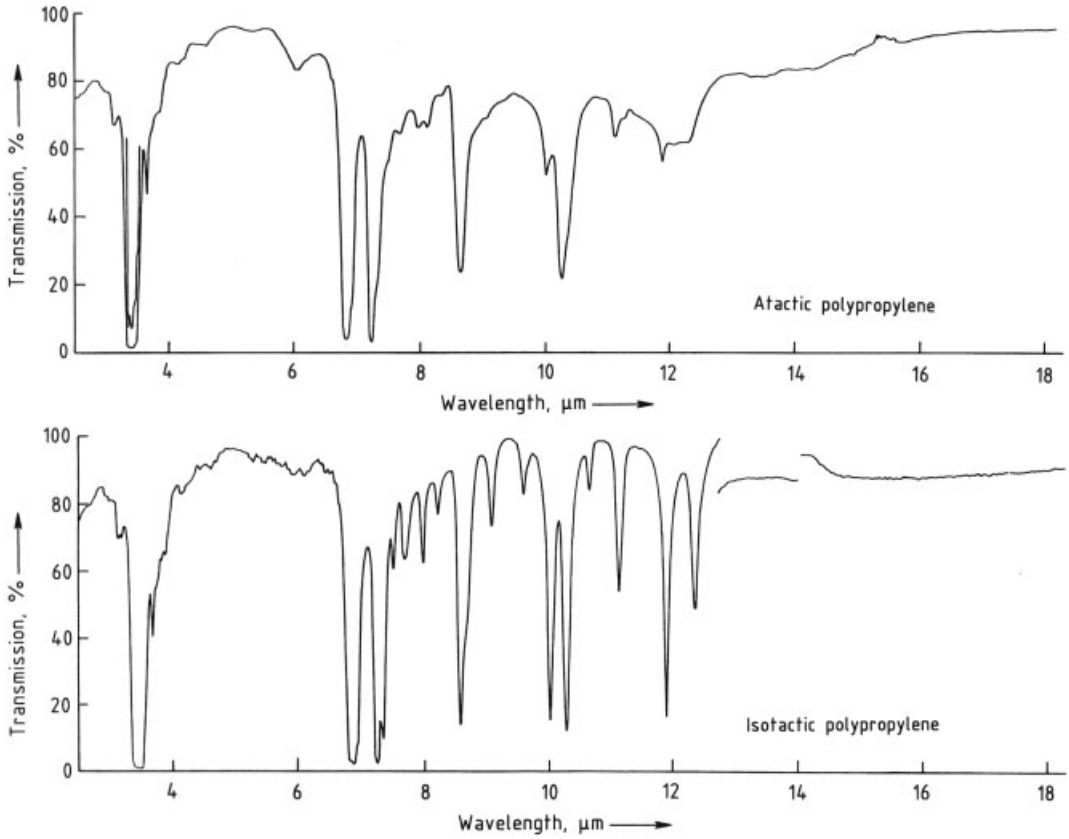

Figure 5. IR spectra of polypropylene [3]

smaller, so that, contrary to what might be expected, polymers are not completely opaque to infrared radiation. By using group theory together with symmetry considerations, it can be shown that $3N - 4$ fundamental vibrations are expected for a linear polymer chain.

In practice, direct assignment of the large number of observable vibrations that still remain is not always feasible. The coupling of molecular vibrations is generally so weak that a spectrum of an unknown polymer can usually be identified most easily by comparing it with spec-tra of low molecular mass model compounds, or with spectra of polymers of known structure. In those cases in which inter- or intramolecular coupling is evident, however, the IR spectrum can provide important additional structural information (Figs. 4 and 5) and the sequence of structural units (Table 7) [3].

Extensive spectral catalogs are now available containing tables of group frequencies [3]. Nevertheless, one should never rely too heavily on an analysis based solely on spectral comparisons. Confirmatory tests should be conducted as well. Elemental analysis (Section 4.1) and other physical methods such as NMR spectroscopy (Section 6.1) are especially recommended.

Quantitative Spectroscopy. IR spectra of polymers are preferably recorded with highly automated Fourier transform instruments (FT-IR).

Spectra of soluble plastics can be recorded in largely IR-transparent solvents such as carbon disulfide, carbon tetrachloride, and dichloromethane. However, the preferred method is to examine a polymer film, prepared either as a thin section or cast directly as a film. A film can also be prepared from a solution evaporated

on a sodium chloride plate. A film thickness of ca. 10 μm is usually adequate for a reliable measurement.

An insoluble plastic can also be analyzed in the form of a fine powder. In this case the sample is first mixed with a powdered alkali metal halide (potassium bromide) and then compressed into a pellet.

The ATR (attenuated total reflection) method [77] allows IR analysis to be carried out on the surface of a plastic sample. The method is based on examining IR radiation that has been reflected from a plastic surface in contact with an IR-transparent substance of lower refractive index. Selective absorption results in a spectrum analogous to a transmission spectrum. Multiple total reflection techniques increase the sensitivity of the analysis [3].

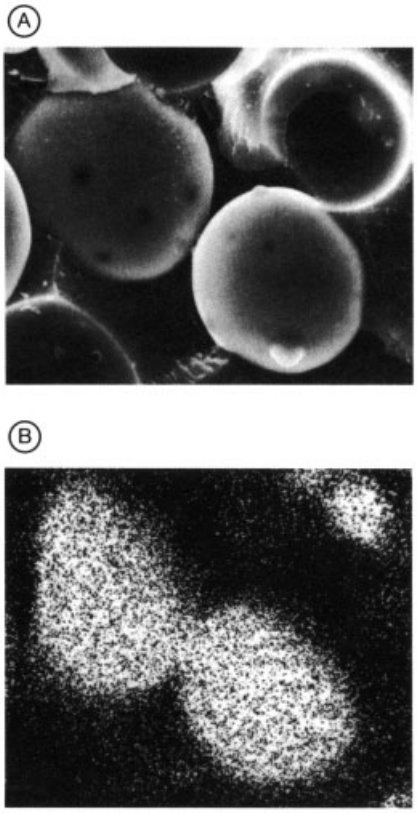

Figure 6. Scanning electron micrograph of the fracture plane of a phenol–formaldehyde resin filled with glass beads (A) and identification of silicon by X-ray fluorescence
(B) Magnification × 300.

Other Spectroscopic Methods. Comprehensive reviews are provided in [9] and [33]. Methods for detecting traces of metals in plastics include flame photometry [1] and atomic absorption spectroscopy, applicable in both the visible and ultraviolet regions of the spectrum.

Elemental analysis and the spectroscopic methods so far discussed provide results that correspond to averages over an entire plastic sample. In the case of plastics containing inorganic pigments (titanium white, cadmium yellow) or fillers (glass fibers, carbon black) information on the distribution of the additives is often required. Appropriate analytical methods here include X-ray fluorescence spectroscopy and especially the electron beam microprobe. The latter method employs an electron beam to probe systematically the nature of a plastic surface. Adjusting the analyzer so that it corresponds to X-ray lines of a particular element (e.g., silicon in the glass beads of a filler) gives a picture of the local concentration distribution of that particular element (Fig. 6). The resolution of an electron microprobe is ca. 1 μm. Elements of atomic number > 5 can be detected with a sensitivity up to 100 ppm (ca. 10^{-14} g/cm^3).

5. Molecular Mass Determination

Introduction. Molecular masses for industrially important polymers extend to values as high as 10^6. Methods of molecular mass determination can be divided into two categories: absolute and relative [78–104]. Evaluation based on an absolute method involves in addition to experimental data (e.g., osmotic pressure, intensity of dispersed light) only such universal factors as the gas constant. The most important absolute methods are membrane osmosis, vapor pressure osmosis, combined sedimentation and diffusion, sedimentation equilibrium, and light scattering [78], [79], [83], [87].

Membrane osmosis is based on the equilibrium established by a membrane separating a polymer solution from pure solvent. Molecular mass determination by combined sedimentation and diffusion requires analysis of two types of material transport within a polymer solution: (1) sedimentation under the influence of a powerful

centrifugal field, and (2) free diffusion. Alternatively, establishment of a stationary state in an analytical cell under the influence of a weak centrifugal field and consideration of the resulting equilibrium between sedimentation and diffusion also permits calculation of an absolute molecular mass. The method of light scattering requires the analysis of the intensity of Tyndall light as a function of the angle of observation.

Relative methods also involve polymer properties that are dependent upon molecular mass, but ones that do not lend themselves to direct conversion into unique molecular mass values. Essential to this approach is a calibration curve, prepared by using an absolute method. The most important relative methods utilize intrinsic viscosity (Staudinger index) or gel permeation chromatography (GPC). The experimental results in both cases depend not only on molecular mass but also on the conformation of polymer molecules in solution.

Since polymerization reactions are stochastic processes, most polymer samples are polymolecular. Standard separation methods do not allow the separation of macromolecules differing in degree of polymerization by only a single monomeric unit. For this reason, all absolute methods for molecular mass determination give characteristic mean values. Investigation of the molecular heterogeneity of a polymer must therefore include two discrete steps: (1) separation of the polymer components, and (2) determination of molecular mass for the separated components. Both ultracentrifugation and gel chromatography are effective for separating homopolymers of differing molecular mass, so these methods are widely used in the determination of molecular mass distribution.

Average Molecular Masses. Each molecule of solute in a solution makes an equal contribution to the osmotic pressure regardless of molecular mass. Osmotic analysis of a polymer thus gives the number-average molecular mass $\bar{M}_{\mathrm{n}}$. If n_i is the number of moles of molecules with molecular mass M_i, the corresponding mass concentration will be $n_i = c_i/M_i$. It therefore follows that:

$$\bar{M}_{\mathrm{n}} = \frac{\sum_i n_i M_i}{\sum_i n_i} = \frac{\sum_i c_i}{\sum_i c_i M_i} \tag{3}$$

Dispersion of light depends on the molecular mass, and therefore gives the weight-average molecular mass $\bar{M}_{\mathrm{w}}$:

$$\bar{M}_{\mathrm{w}} = \frac{\sum_i n_i M_i^2}{\sum_i n_i M_i} = \frac{\sum_i c_i M_i}{\sum_i c_i} \tag{4}$$

In a molecularly uniform polymer $\bar{M}_{\mathrm{w}} = \bar{M}_{\mathrm{n}}$, whereas for a polymolecular sample $\bar{M}_{\mathrm{w}} > \bar{M}_{\mathrm{n}}$. The discrepancy becomes greater the broader the mass distribution. The molecular nonuniformity U

$$U = \left(\bar{M}_{\mathrm{w}}/\bar{M}_{\mathrm{n}}\right) - 1 \tag{5}$$

is a measure of the molecular heterogeneity so long as the samples compared have similar molecular mass distributions. In some cases, samples with differing molecular mass distributions may actually have identical values of $\bar{M}_{\mathrm{w}}$ and $\bar{M}_{\mathrm{n}}$, a consequence of the fact that Equations (3) and (4) are based on summations.

Not all analytical methods lead to simple mean values such as $\bar{M}_{\mathrm{w}}$ and $\bar{M}_{\mathrm{n}}$. A more general definition of an average $\bar{M}$ is

$$\bar{M}_{\beta} = \frac{\sum_i c_i M_i^{\beta}}{\sum_i c_i M_i^{\beta-1}} \tag{6}$$

For $\bar{M}_{\mathrm{n}}$, $\beta = 0$, whereas $\beta = 1$ for $\bar{M}_{\mathrm{w}}$. Depending upon the experimental method, the mean that is determined may have $\beta > 1$ or an exponent β with a fractional value. For example, the method that combines sedimentation velocity with diffusion leads to a mean somewhat smaller than $\bar{M}_{\mathrm{w}}$, and the same is true for the intrinsic viscosity $[\eta]$.

5.1. Absolute Methods

5.1.1. Membrane Osmometry

The osmotic pressure Π of a solution is a function of the chemical potential μ_1 and partial molar volume V_1 of the solvent:

$$\Pi = -\Delta\mu_1/V_1 \tag{7}$$

The molecular mass is derived from the Vant'Hoff equation, which applies to ideal solutions

$$\Pi = \frac{RTc}{\bar{M}_n} \tag{8}$$

The assumption of ideality for a polymer solution is valid only in exceptional cases. Nonideality of a solution results in an osmotic pressure and, therefore, an apparent molecular mass that depends upon the concentration of the solute, a consequence of the differing spatial demands of solvent and polymer molecules. However, the additional enthalpy and entropy terms that determine μ_1 may cancel out at the so- called Θ temperature, resulting in a pseudoideal solution. The Θ temperature is of major significance in the determination of molecular masses and molecular mass distributions of polymers. Such Θ systems are in fact known for a large number of polymers [53]. At the Θ temperature, plotting Π/c versus c gives a line parallel to the c axis. The advantage of acquiring data at the Θ temperature is that this permits measurement of the molar mass distribution at a single concentration.

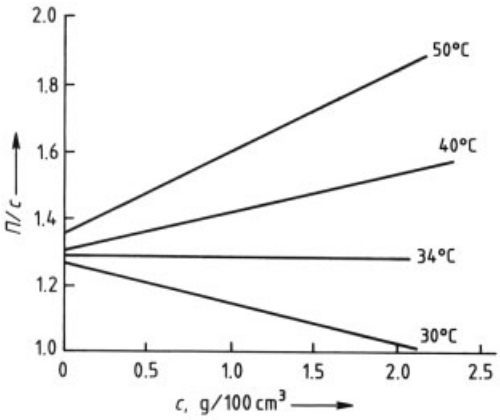

Figure 7. The dependence of osmotic pressure on temperature [21]

An example is given in Figure 7 for the case of polystyrene dissolved in cyclohexane. The molecular-mass-independent Θ temperature in this case is 34 °C. Increasing the temperature leads to an additional term for the osmotic pressure, reflecting the effect of the chemical potential. Lowering the temperature makes this term negative. The nonideal concentration dependence observed at temperatures other than the Θ temperature can be represented by a power series in c:

$$\frac{\Pi}{c} = \frac{RT}{M_n} + A_2 c + \ldots.. \tag{9}$$

As shown in Figure 7, the molecular mass $\bar{M}_n$ can be determined by extrapolation to $c \to 0$.

Experimental Technique. In an osmometer, such as that illustrated schematically in Figure 8, a polymer solution and a pure solvent are separated by a semipermeable membrane, ideally one that permits passage only of the solvent. Each of the chambers is surmounted by a capillary. The difference in chemical potential between the two chambers creates an osmotic pressure Π, which can be quantified in terms of a difference in height Δh of the liquid in the two capillaries. Equilibration of the osmotic pressure may require from several hours to several days. Automatic osmometers eliminate the need for waiting until equilibrium has been achieved. Instead, a servo-activated device applies an external pressure continuously to the solution chamber, and the magnitude of the pressure required to maintain equivalence is taken as a quantitative measure of the osmotic pressure Π. Another type of osmometer utilizes chambers with fixed volumes, permitting direct evaluation of Π by means of anelectronic pressure sensor.

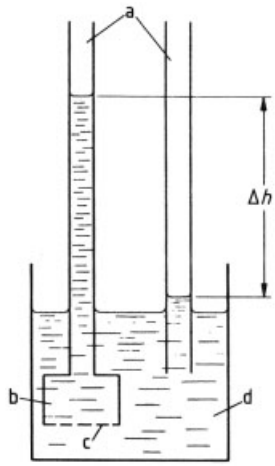

Figure 8. Schematic of a membrane osmometer [21]
a) Capillaries; b) Polymer solution; c) Membrane; d) Pure solvent

Since the osmotic pressure decreases with increasing molecular mass, the scope of membrane osmometry is limited by the minimum

detectable pressure, corresponding to a molecular mass of at most 10^6. The reliability of an osmotic determination is largely a function of the permeability of the membrane. If molecules of low molecular mass (e.g., oligomers, foreign solvent introduced during precipitation) can enter the solvent chamber then the apparent value for the molecular mass of the polymer will lie above $\bar{M}_n$. The extent of such error is a function of membrane permeability and the proportion of foreign molecules.

Osmosis is not limited to molecular mass determination on homopolymers. It is also the method of choice for copolymers regardless of the sequence of structural units or the extent of chemical heterogeneity, because osmotic pressure is a function only of the number of molecules present, irrespective of their chemical structures.

5.1.2. Lowering Vapor Pressure

Vapor Pressure Osmometry. According to Raoult's law, the ratio of the vapor pressure of a pure solvent p_1^0 to that of a solution p_1 is equal to the mole fraction of the solvent x_1:

$$p_1/p_1^0 = x_1 = 1 - x_2 \tag{10}$$

x_2 mole fraction of solute and the relative decrease in vapor pressure is:

$$\Delta p_1/p_1^0 = x_2 = \frac{n_2}{n_1} \tag{11}$$

n_1 number of moles of solvent
n_2 number of moles of solute

In a vapor-pressure osmometer, solvent distills, a consequence of the difference in vapor pressure exerted by solvent and solution, onto droplets of solution. The heat of condensation is measured, which for an ideal system is proportional to the number of moles of dissolved polymer (i.e., with a constant sample size, inversely proportional to the molecular mass). The range of applicability of vapor-pressure osmometry is thus a function of the precision and sensitivity with which temperature differences can be measured. Determinations of this type are possible up to $\bar{M}_n = 10^4$. The results must be extrapolated to infinite dilution, as in the case of membrane osmometry, and for the same reasons.

Freezing Point Depression, Boiling Point Elevation. The relative decrease in vapor pressure of a polymer solution can also be determined by measuring the depression of freezing point ΔT_f or elevation of boiling point ΔT_b. The corresponding cryoscopic and ebullioscopic methods are based on the relationship that exists between vapor pressure and temperature, as given by the Clausius – Clapeyron equation:

$$dp/dT = \Delta H/\Delta V \cdot T \tag{12}$$

ΔH heat of vaporization
ΔV volume change on evaporation

The ebullioscopic method is the more important in practice. Molecular masses are derived from the equation:

$$\frac{\Delta T_b}{c} = \frac{RT^2}{\bar{M}_n} \cdot \frac{\bar{V}_1}{\Delta H_s} \tag{13}$$

$\bar{V}_1$ specific volume of solvent
H_s specific heat of vaporization of solvent

The term $RT^2 \times \bar{V}_1/\Delta H_s$ is the ebullioscopic constant K for a particular solvent.

5.1.3. Ultracentrifugation

Combined Sedimentation and Diffusion. Sedimentation rates are measured by subjecting a polymer solution to a high centrifugal acceleration in an ultracentrifuge, depicted schematically in Figure 9. The centrifugal acceleration is expressed as $\omega^2 r$, where ω is angular velocity and r is the distance from the axis of rotation [83–85]. Centrifugal force causes dissolved molecules to be sedimented at the bottom of the sample cell as a function of molecular mass and conformation in solution. Directed material transport is opposed by forces of friction. With F as the molar coefficient of friction, $\bar{V}_2$ the partial specific volume of the solute, and ϱ_1 the density of the solvent, the following equation of motion is applicable:

$$M\left(1 - \bar{V}_2 \varrho_1\right) = F\frac{t}{\omega^2 r} = F \cdot s \tag{14}$$

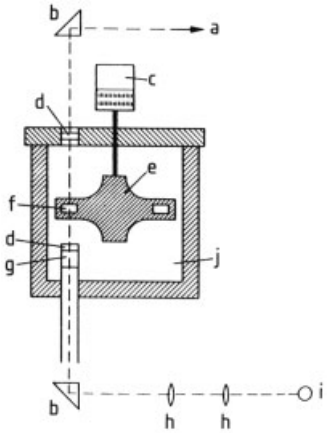

Figure 9. Schematic of an ultracentrifuge [21]
a) Camera; b) Prism; c) Motor; d) Window; e) Rotor; f) Sample cell; g) Filter; h) Condenser; i) Light source; j) Vacuum chamber

The sedimentation constant s thus represents a rate of sedimentation through a uniform field. At infinite dilution, Einstein's law is applicable, providing a relationship between the coeffficient of friction F and the diffusion constant D:

$$F = RT/D \tag{15}$$

Equations (14) and (15) can be combined to give the Svedberg equation:

$$M = \frac{s_0}{D_0} \cdot \frac{RT}{\left(1 - \tilde{V}_2 \varrho_1\right)} \tag{16}$$

By combining the results obtained in separate sedimentation and diffusion experiments it is possible to eliminate the effect of the coefficient of friction on both transport processes. In this way, determination of molecular mass via the Svedberg equation becomes an absolute method. The procedure requires optimization of the rotational speed of the centrifuge as well as the choice of a suitable solvent [53]. For sedimentation measurements, the density of the polymer must be greater than that of the solvent. The transport velocity for the polymer in a centrifuge cell is measured by detecting either the UV spectrum or the refractive index of the polymer.

As in all experiments involving molecular mass determination, concentration dependencies must also be considered. If measurements are not made at the Θ temperature, then a series of s and D at various concentrations must be measured. The corresponding concentration dependencies can be expressed as:

$$\frac{1}{s} = \frac{1}{s_0}\left(1 + k_s c\right) \tag{17}$$

$$D = D_0\left(1 + k_D c\right) \tag{18}$$

The resulting molecular mass is a mean that arises from averaging the two measured variables; its value is close to $\bar{M}_w$. The method is applicable over the molecular mass range 10^4–10^7 [83].

Sedimentation Equilibrium. If the rotational velocity of a centrifuge is so adjusted that a mass-transport equilibrium is established between sedimentation and diffusion, then during the time interval dt the same number of molecules will migrate through a cross-section of the cell in the direction of the cell floor (due to sedimentation) as toward the axis of rotation (due to diffusion). This equilibrium follows the relationship:

$$\frac{\partial c}{c} = \frac{M\left(1 - \tilde{V}_2 \varrho_1\right)\omega^2 r \mathrm{d}r}{RT} \tag{19}$$

Integration between r_1 and r_2 leads to an equation for determining the molecular mass:

$$M = \frac{2RT \ln\left(c_1/c_2\right)}{\left(1 - \tilde{V}_2 \varrho_1\right)\omega^2\left(r_2^2 - r_1^2\right)} \tag{20}$$

Equation 20 is applicable only to an ideal system and requires extrapolation to $c \to 0$ for a real system.

The nature of the resulting mean is dependent upon the method used for interpreting the experiment. In the simplest case the z-mean, corresponding to $\beta = 2$ in Equation (6) is obtained. The range of applicability is extremely large, extending from monomers to molecular masses of 10^6. Analyses involving high molecular mass require long periods of time for the establishment of equilibrium. The time factor can be reduced by utilizing cells with low fill heights together with a multicell rotor. Solvents are selected on the same basis as in sedimentation transport.

5.1.4. Light Scattering

Introduction. The electric field associated with a light wave causes the electron shells of a polymer molecule to vibrate with a frequency corresponding to that of the monochromatic excitation beam. Excited molecules then emit scattered light in all directions at a wavelength λ_0 identical to that of the primary irradiation (the Tyndall effect) [86–88].

The intensity I of a light beam passing through a nonabsorbing polymer solution is reduced relative to the initial intensity I_0 by an amount equivalent to the light that is scattered. The overall scattering intensity attributable to N molecules within a unit volume (1 cm^3) is expressed by the turbidity number τ:

$$\tau = \frac{I}{I_0} = \frac{128}{3}\pi^5 N \lambda_0^4 \alpha^2 \tag{21}$$

The polarizability α of a substance is a function of its refractive index n:

$$\alpha = \frac{1}{N}\frac{n^2-1}{4\pi} \tag{22}$$

Combining Equations (21) and (22) leads to:

$$\tau = \frac{I}{I_0}\frac{32}{3}\pi^3 \frac{(n^2-1)^2}{N\lambda_0^4} \tag{23}$$

The above equations apply strictly only to gases but can also be used for dilute solutions of polymers. For an ideal solution, the scattering intensity is a function of the difference between the refractive index of the solvent n_1 and that of the dissolved polymer n, introduced as the refractive index increment $\partial n/\partial c$ [53]. Replacing the number of macromolecules by the mass concentration c and the molecular mass $\bar{M}_w$ leads to the expression:

$$\frac{\tau}{c} = \bar{M}_w \frac{32}{3}\pi^3 \frac{n_1^2}{N_A \lambda_0^4}(\partial n/\partial c) \tag{24}$$

N_A is the Avogadro number from which the weight-average molecular mass $\bar{M}_w$ of an unknown polymer can be calculated. Equation (24) contains no term for scattering due to the solvent, so this must be subtracted from the overall scattering τ of the polymer solution. The value of τ is determined in practice by measuring the intensity of laterally scattered radiation as a function of the angle Θ. All measurements and their interpretation are influenced by the dimensions and shapes of the dissolved macromolecules relative to λ_0.

Macromolecules Smaller than $\lambda/20$. If the solute molecules responsible for scattering are small relative to the wavelength of the source they can be treated as point sources of radiation. This is the case for flexible chain molecules up to the dimension $\lambda/20$. If the irradiating light is polarized in a plane perpendicular to the plane of observation, the resulting intensity I_θ of light scattered through an angle Θ is independent of the angle of observation Θ. In this case the relative scattering intensity R_θ per unit volume and distance r for any angle of observation Θ is a function of the turbidity number τ according to:

$$R_\Theta = \frac{r^2 I_\Theta}{I_0} = \frac{3}{8\pi}\tau \tag{25}$$

where I_0 is the intensity of the incident light. Introducing the optical constant K,

$$K = 4\pi^2 \frac{n_1^2}{N_A \lambda_0^4}(\partial n/\partial c) \tag{26}$$

which is characteristic of a particular system and wavelength leads to an equation applicable to a dilute ideal solution. This defines the molecular mass as a function of the amount of light scattered through an angle Θ, an expression which is also appropriate for pseudoideal systems (see Section 5.1.1):

$$K\frac{c}{R_\Theta} = \frac{1}{\bar{M}_w} \tag{27}$$

Analogous to the situation with osmotic pressure, the concentration dependence of the reduced scattering intensity for real polymer solutions is taken into account by the Debye relationship:

$$K\frac{c}{R_\Theta} = \frac{1}{\bar{M}_w} + 2A_2 c + \ldots.. \tag{28}$$

where A_2 is the second virial coefficient of the osmotic pressure. In this case $\bar{M}_w$ is established by measurement of a concentration series and then extrapolating to infinite dilution.

Macromolecules larger than $\lambda/20$ can no longer be regarded as point oscillators because of interference with the laterally scattered light.

Path differences and, therefore, the amount of interference quenching of the scattered radiation increase with increasing angle of observation and with increasing size of the macromolecules. However, at an angle of 0° – where direct measurement is of course impractical – no interference occurs, so Equation (27) remains valid even for large molecules. The angular dependence of scattering intensity for large macromolecules is a function not only of molecular size but also of shape and molecular heterogeneity. These factors are accounted for in the equation for $\bar{M}_w$ by the scattering function P_θ:

$$K\frac{c}{R_\Theta}=\frac{1}{P_\Theta \bar{M}_w}+\frac{2A_2c}{P_\Theta}+..... \quad (29)$$

For statistically entangled, uniform macromolecules with a mean distance of $\sqrt{h^2}$, the following expression is applicable to P_θ:

$$P_\Theta^{-1}=1+\left(8\pi^2/9\lambda_0^2\right)\cdot h^{-2}\cdot\sin^2(\Theta/2) \quad (30)$$

Scattering functions have been computed for a wide range of molecular shapes [86], [87]. Their utility extends beyond providing a correction factor for use in determining $\bar{M}_w$. Quantitative evaluation of the angular dependence of the scattering function also provides additional information on macroscopic conformation (size and shape) of the dissolved macromolecules. It follows from Equations (29) and (30) that:

$$K\frac{c}{R_\Theta}=\frac{1}{\bar{M}_w}+D\cdot\sin^2(\Theta/2)+2A_2c+..... \quad (31)$$

where

$$D=\frac{1}{\bar{M}_w}\cdot\frac{8\pi^2}{9\lambda_0^2}\cdot h^{-2} \quad (32)$$

The determination of the molecular mass of a large macromolecule by means of Equation (31) involves extrapolation to $\Theta \to 0$ and $c \to 0$ with the aid of a Zimm diagram, as shown by the example in Figure 10. Here the expression Kc/R_θ is plotted simultaneously against the concentration c and the angular function $\sin^2(\Theta/2)$. This facilitates both of the required extrapolations, which are linear and intersect the ordinate at $1/\bar{M}_w$. In a similar way one can also evaluate the concentration dependence of the virial coefficient A_2 and the angular dependence of the molecular dimension $\sqrt{h^2}$. The method of light scattering is applicable up to a molecular mass of 10^8.

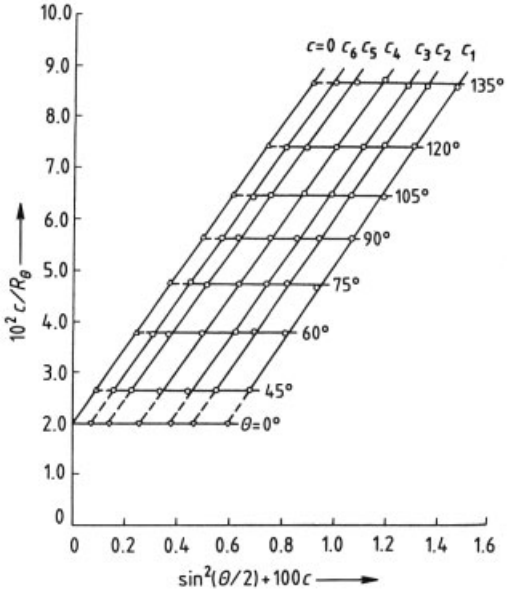

Figure 10. Zimm diagram of polytetrahydrofuran ($\bar{M}_w = 2.26 \times 10^6$) in 2-propanol ($T = 317$ K) [87]

Molecular Masses of Copolymers. Equation (28) for $\bar{M}_w$ assumes a macromolecular system with a uniform index of refraction; in other words, it is strictly valid only for homopolymers. A copolymer, on the other hand, consists of structural units that differ chemically, and thus would be expected to differ with respect to refractive index as well. Each component makes its own characteristic contribution to the overall scattering through the term $(\partial n/\partial c)^2$.

The refractive index increment for a copolymer is a function of the mass fractions m_A and m_B of the individual components and the refractive index increments of the corresponding homopolymers:

$$(\partial n/\partial c)_{AB}=m_A(\partial n/\partial c)_A+m_B(\partial n/\partial c)_B \quad (33)$$

The measured scattering intensity for a copolymer therefore depends not only on molecular mass but also on chemical composition.

The molecular mass of a copolymer that is heterogeneous both chemically and in degree of polymerization can be determined by conducting light-scattering experiments in at least three solvents [88]. Considerable experimental effort is required, but it is possible to establish in this

way a mean mass $\bar{M}_w$ and thereby the nonuniformity U.

The equations used in studying homopolymers can also be applied to chemically heteregeneous copolymers provided structural units in the latter are isorefractive. This is the case to a good approximation, for example, with ethylene – propylene copolymers.

Experimental Technique. Figure 11 illustrates the design of a light-scattering apparatus. The measurement requires total absence of dust from the solutions and the cuvettes. Solutions can be purified by filtration or ultracentrifugation. The refractive index increment is determined by means of a differential refractometer. In order to achieve a high scattering intensity it is important that the refractive indices of solvent and solute differ as much as possible.

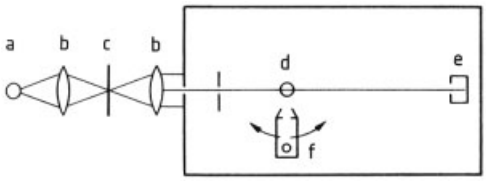

Figure 11. Schematic of a light-scattering apparatus
a) Light source; b) Collimator; c) Filter and polarizer; d) Sample cell; e) Light trap; f) Photomultiplier

5.2. Relative Methods

Intrinsic Viscosity. Even a low concentration of macromolecules leads to a significant increase in the viscosity of a solvent. The increase observed for a given amount of a particular polymer depends not only upon molecular mass but also upon the dimensions of the molecules in solution. Macroconformation for a statistically entangled macromolecule is a function of both solvent and temperature. Use of intrinsic viscosity for characterizing a molecular mass requires that the experiment be conducted at constant temperature [53], [89–91]. The intrinsic viscosity $[\eta]$ (Staudinger index) is defined as:

$$[\eta] = \lim_{\substack{c \to 0 \\ G \to 0}} \left[\frac{\eta_{sp}}{c} \right] \tag{34}$$

This represents a limiting value for the reduced specific viscosity η_{sp} for a concentration $c \to 0$ and a shear rate $G \to 0$. Extrapolation to $c \to 0$ can be accomplished graphically by plotting η_{sp}/c against η_{sp} or c. In most cases a linear relationship results. The reduced viscosity is determined by measuring the flow times t for the solution and t_1 for the solvent, leading to:

$$\eta_{sp} = \frac{t\varrho - t_1 \varrho_1}{t_1 \varrho_1} \tag{35}$$

ϱ, ϱ_1 Density of solution and solvent, respectively

At low concentration $\varrho_1 = \varrho$, so this can be simplified to:

$$\eta_{sp} = \frac{t - t_1}{t_1} \tag{36}$$

Extrapolation to $G \to 0$ can be dispensed with for the low molecular mass range and by employing an Ubbelohde viscometer with a low shear rate. Large macromolecules tend to become partially oriented upon passage through a capillary as a consequence of a reduction in frictional resistance by solvent shearing. This structural viscosity is taken into account by measuring the viscosity of a solution as a function of the applied pressure p. An Ostwald viscometer is convenient for the purpose. Extrapolation to $G \to 0$ can be accomplished graphically by plotting η_{sp} against p. Linear extrapolation is possible here as well. Automatic viscometers have also been developed. These devices take into account the role played by shear rate as well as measuring the flow time.

The relationship between intrinsic viscosity $[\eta]$ and molecular mass for a linear, statistically entangled thread-like chain molecule is given by the Kuhn expression:

$$[\eta] = K M^{\alpha} \tag{37}$$

A log – log plot of $[\eta]$ versus $\bar{M}_{w,n}$ gives a straight line with slope α. The value of the exponent depends upon the flexibility of the entanglement. It approaches 1 in a thermodynamically favorable solvent. The minimum value of 0.5 is obtained with Θ systems (cf. Section 5.1.1). Representative values of α lie in the range 0.6 – 0.9. The intrinsic viscosity $[\eta]$ leads to a mean of the form:

$$M_\eta = \left[\frac{\sum_i c_i M_i^\alpha}{\sum_i c_i} \right]^{1/\alpha} \tag{38}$$

This is identical to $\bar{M}_w$ when $\alpha = 1$, but it becomes increasingly smaller as α decreases. For this reason it is advantageous to calibrate the intrinsic viscosity $[\eta]$ by means of an absolute method such as light scattering. Figure 12 illustrates $[\eta] - M$ calibration curves for various temperatures and solvents in the case of polystyrene. For additional systems see [53].

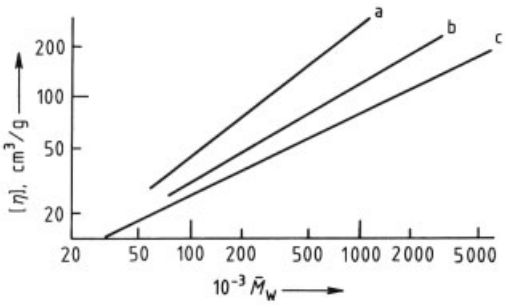

Figure 12. Logarithmic plot of viscosity versus molecular mass of polystyrene
a) In benzene ($T = 290$ K); b) In methyl ethyl ketone ($T = 290$ K); c) In cyclohexane ($T = 305$ K)

Relative molecular mass determination on the basis of viscometry is of major technical importance. Absolute methods are preferable for reasons of time and expense only with polymers for which an expression corresponding to Equation (35) has yet to be established. Relative methods provide reliable results whenever there is reasonable agreement between the molecular mass distributions of the reference and the sample.

5.3. Determination of Molecular Mass Distribution

Introduction. The nonuniformity U is incapable of providing an unambiguous characterization of polymolecularity. Samples with identical values of $\bar{M}_w$ and $\bar{M}_n$ may differ in their molecular mass distributions. The polymolecularity of a polymer sample can be established fully only by fractionation. In a batch separation procedure the total mass of polymer molecules in the system is represented in terms of an integral distribution over a discrete set of molecular masses between M_0 and M_i:

$$J_w(M) = \sum_{M_0}^{M_i} w_i \Delta M \tag{39}$$

Equation (39) describes a step function which progresses in steps of the monomer mass dM. A corresponding experimental measurement assumes prior molecular mass evaluation of the various fractions with molecular mass M_i. However, changes in physical characteristics for similar macromolecules differing only in size are so small that isolation of a fraction containing only molecules of a single degree of polymerization is essentially impossible. In a continuous separation method the step function of Equation (39) is replaced by a continuous integral distribution function:

$$J_w(M) = \int_{M_0}^{M_i} w(M)\,\mathrm{d}M \tag{40}$$

with the normalization:

$$J_w(M) = \int_0^\infty w(M)\,\mathrm{d}M = 1 \tag{41}$$

The first derivative of this expression is the differential molecular mass distribution, which provides a mass fraction for macromolecules in the increment M+dM:

$$w(M) = \frac{\mathrm{d}J_w(M)}{\mathrm{d}M} \tag{42}$$

Experimental determination of a molecular mass distribution thus always entails use of a method of separation combined with molecular mass analysis of the individual fractions.

Polymers can be separated by taking advantage of macromolecular properties that depend exclusively on molecular mass, including solubility (precipitation fractionation) and sedimentation rate. Chromatographic procedures have also assumed considerable practical importance.

Precipitation Fractionation. The solubility of a polymer depends not only on the nature of the structural units present, but also on molecular mass. This characteristic of polymer solutions serves as the basis for a simple and effective approach to establishing a molecular mass

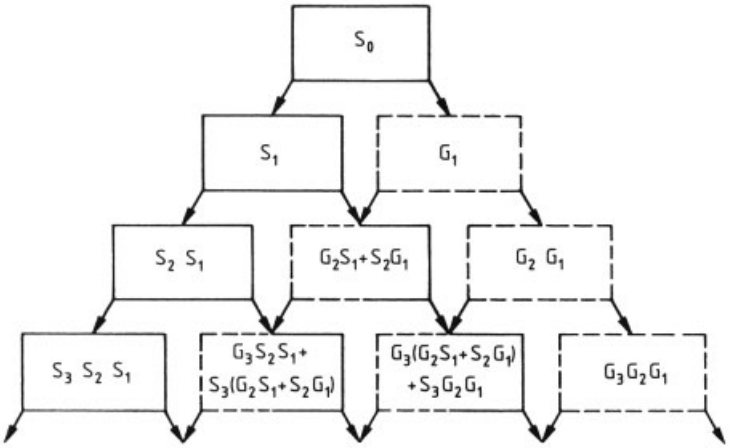

Figure 13. Schematic fractionation of polymers by dissolution and precipitation [9]
S = Solution; G = Gel

distribution. The process begins with a ca. 1 % solution of a polymer, to which is slowly added dropwise with stirring a precipitating agent, usually methanol. Turbidity is observed after the addition of a certain amount of the precipitating agent. The mixture is then warmed slightly to restore complete dissolution. Subsequent lowering of the temperature again leads to precipitation. After some period of time the system separates into a polymer-rich gel phase (G_1) and a polymer-poor sol phase (S_1). In most cases the polymer is the more dense component, so the gel phase collects at the bottom of the precipitating vessel. Evaporation of this phase gives the polymer fraction with the highest molecular mass. The sol S_1 is then treated with additional precipitating agent leading to a second stage of separation, etc. This procedure allows the isolation of up to 10 fractions from a single polymer sample. Fractions do not consist exclusively of molecules with uniform molecular masses, however, but instead contain components with both higher and lower masses as well. Even repetition of the process will not alter this situation [92], [93].

In order to improve the fractionation the gel phase G_1 can be treated again with precipitating agent, producing another sol phase that can be combined with S_1 subfractionation procedure. The triangular fractionation method illustrated schematically in Figure 13 represents another effective approach to fractionation.

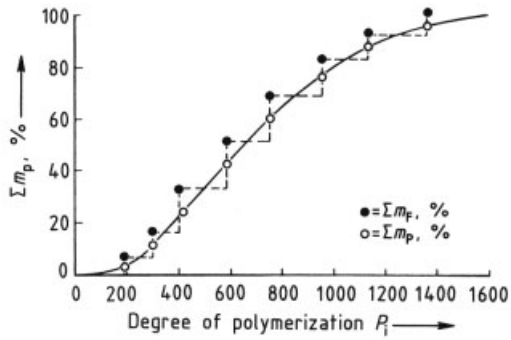

Analysis of typical fractionation data

Fraction	Degree of polymerization P_i	wt % of fraction m_F	Σm_F, %	Σm_p, %
VIII	190	6.4	6.4	3.2
VII	300	9.0	15.4	10.9
VI	420	17.2	32.6	24.0
V	585	17.6	50.2	41.4
IV	755	18.5	68.7	59.4
III	955	13.5	82.2	75.5
II	1140	9.7	92.0	87.0
I	1370	8.1	100.0	95.0

Figure 14. Molecular mass distribution by dissolution fractionation [21]
m_F = mass fraction of a fraction; m_P = mass fraction with degree of polymerization P_i

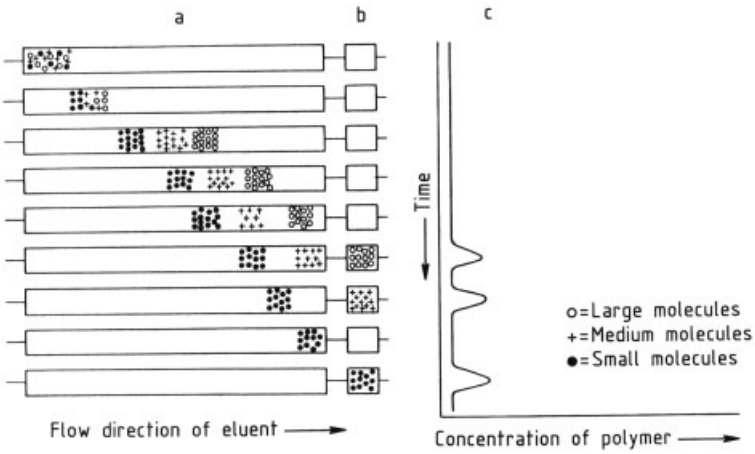

Figure 15. Principle of polymer fractionation by size exclusion chromatography (SEC) [31]
a) Column; b) Detector; c) Detector signal

Once a set of fractions has been isolated the corresponding masses must be determined, along with mass fractions relative to the total sample. Molecular masses are determined by one of the methods described in Sections 5.1 and 5.2. The results permit one to construct a molecular mass distribution consistent with Equation (39) (Fig. 14).

Precipitation fractionation is a very simple procedure. Its chief limitation lies in the small number of fractions generated, which might, for example, fail to reveal a bimodal character of the molecular mass distribution. The time required for separating the gel phases can also be very great, especially since the sol phase increases in volume at each stage in the process, requiring the use of large amounts of solvent. Nevertheless, precipitation has the advantage compared to preparative chromatography that it represents an economical source of gram-quantity fractions, which can in turn serve as starting points for analyzing a polymer by other methods (e.g., configurational analysis with NMR; see Section 6.1).

Gel permeation chromatography (GPC) is a form of liquid column chromatography in which carrier solvent is passed at a constant rate through a column consisting of cross-linked polymer beads previously allowed to swell in the same solvent. The total volume V_t of the column includes the volume occupied by both mobile (V_0) and stationary (V_x) phases. Only a portion of V_x is actually utilized in the separation process: the pore volume V_p [94–103].

Figure 15 shows a simplified model of the separation process. Separation occurs on the basis of the hydrodynamic volume of a given macromolecule, which is in turn a function of molecular mass. It is this volume that governs the extent to which the macromolecule can permeate into the pore structure of a suitably dimensioned microporous gel.

The sample that is to be separated is introduced onto the top of the column as a solution in the mobile phase. Polymer molecules then distribute themselves between the mobile phase V_0 and the pore volume of the stationary phase V_p on the basis of a diffusion-driven equilibrium defined by their hydrodynamic volumes and the pore diameter of the gel. As solvent flows through the column, large molecules are unable to penetrate into the gel matrix due to their diameters, so these are eluted first within the elution volume V_0. By contrast, molecules with diameters smaller than the pore openings diffuse into the interiors of the pores. The smaller the macromolecule the more deeply it is able to penetrate into the stationary phase, and the more its progress is retarded relative to solvent flow outside the pores. In the absence of special interactions between the polymer and the surface of the stationary phase, the observed order of elution and the corresponding elution volume V_e will be a function only of molecular size.

Small macromolecules capable of passing freely between V_p and V_0 are eluted last. The difference between the elution volumes for totally excluded molecules and the smallest macromolecules present corresponds to V_p.

The *gel chromatogram* is generated by measuring the polymer concentration in the eluate as a function of elution volume. The polymer concentration can be measured by monitoring the refractive index and/or ultraviolet absorption of the eluate. Larger columns can be used for preparative GPC, providing an effective alternative to precipitation fractionation.

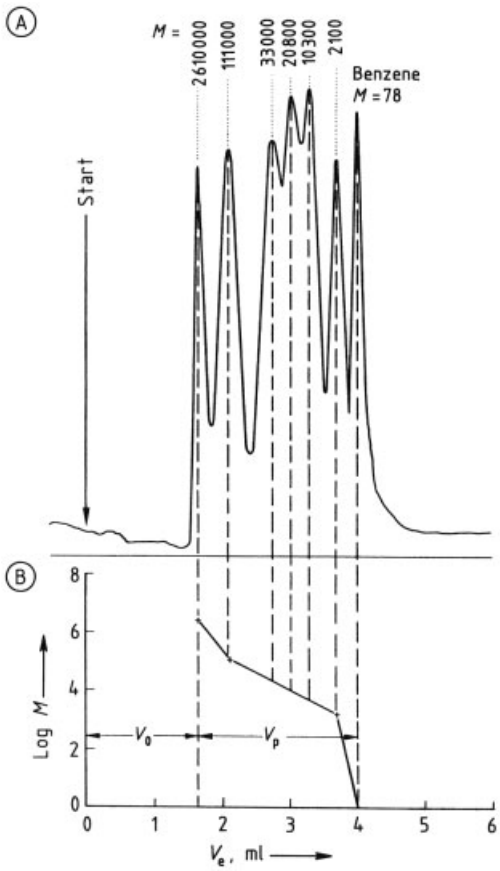

Figure 16. Chromatogram (A) and calibration curve (B) for SEC [96]
V_0 = Volume of eluent in column; V_p = pore volume; V_e = volume of eluate

Even for a completely homogeneous fraction, the quality of the column packing and effects of axial diffusion during the separation process cause peak broadening. An elution curve can be calibrated by chromatography of a set of uniform polymers with known masses determined by an absolute method, as illustrated in Figure 16. The linear portion of the elution curve is described by the function:

$$\log M = \log M_a - A V_e \tag{43}$$

where M_a is the molecular mass above which separation is impossible and A is a constant specific to a particular system and apparatus. This method is applicable within the molecular mass range $10^2 - 10^6$.

The relationship shown in Figure 16 between elution volume V_e and molecular mass M is strictly valid only for a particular homologous series of polymers, and it must be established with the aid of calibration samples. If the intrinsic viscosity $[\eta]$ is used to take into account hydrodynamic volumes of polymers that differ chemically by plotting $[\eta]M$ vs. V_e, the result is a universal elution curve applicable to multiple polymer systems. Within its range of linearity GPC can be used as a relative method for the determination of molecular mass [98].

Evaluation. Gel chromatograms often consist of several broad, unsymmetrical, and overlapping peaks, as illustrated by the example of polypropylene adipate in Figure 17. If the relationship between detector signal and polymer concentration is known, mass fractions $w(M)$ can be computed from the corresponding surface areas in the chromatogram of a series of oligomers. Equation (39) is then used to construct a step distribution function for the various molecular masses. It is not possible in the high molecular mass range to isolate fractions with uniform molecular mass. The typical gel chromatogram must instead be regarded as a superposition of uniform fractions characterized by differing molecular masses. The problem in interpretation is therefore to distinguish the contribution to peak broadening that can be attributed directly to molecular mass distribution in the sample from peak-broadening effects inherent in the system.

The chromatogram of a single fraction can be described as a Gaussian curve of the form:

$$f(V) = A\left(h/\sqrt{\pi}\right)\exp\left[-h^2\left(V - V_{max}\right)^2\right] \tag{44}$$

where V_{max} is the elution volume at the maximum of the curve, h represents the breadth of

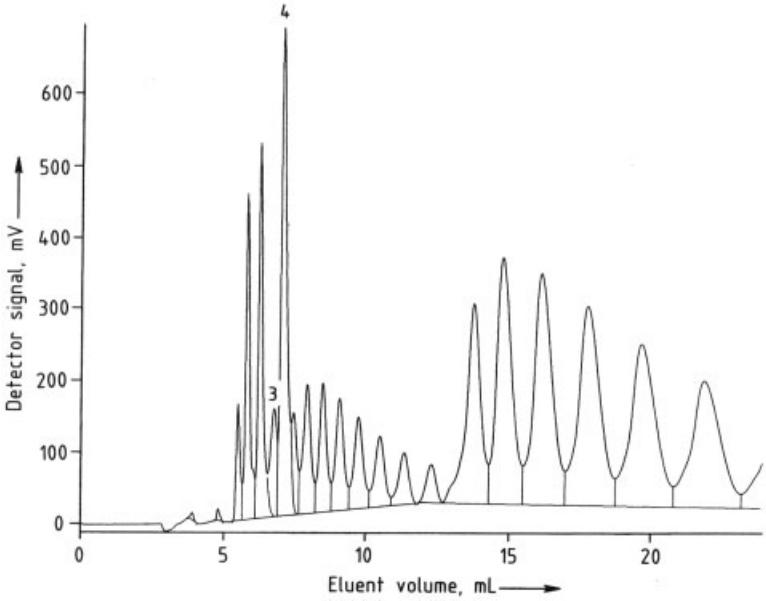

Figure 17. Fractionation of oligomers of different size and composition [101]
Chromatographic separation for coupling with NMR detection (see Fig. 28). Sample: polypropylene adipate; Eluent: D_2O/methanol-d_4; Column: Chromegabond 5 µm

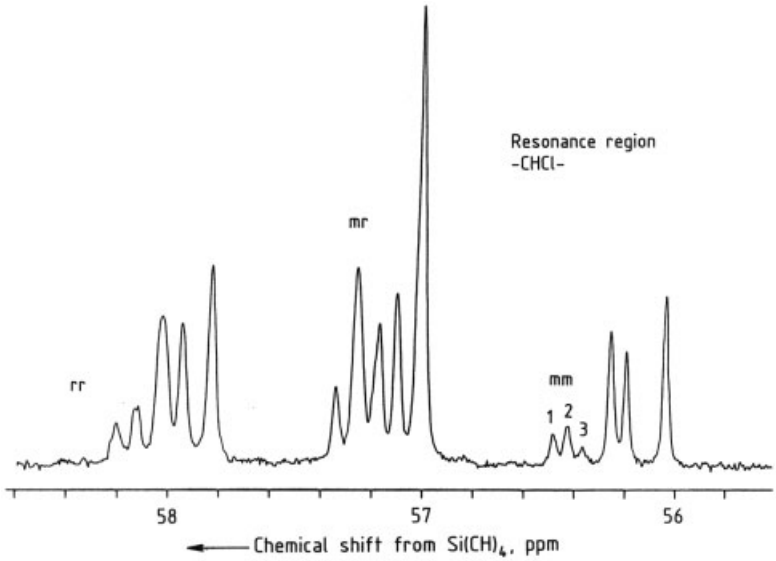

Figure 18. 67.88 MHz ^{13}C NMR spectrum of poly(vinyl chloride)
The signals 1 – 3 indicate configurational sequences (see Table 8) of up to 7 monomer units, rmmmmr, mmmmmr, and mmmmmm, respectively. Solvent: tetrahydrofuran-d_8; Temperature: 373 K; Broadband decoupling

the Gaussian curve, and A is the area under the chromatogram. For a system of i fractions each with area A_i:

$$f(V) = \sum_i A_i \left(h_i/\sqrt{\pi}\right) \exp\left[-h_i^2 (V - V_{\max,i})^2\right] \quad (45)$$

If the number of fractions is very large, or if the chromatogram cannot be resolved into uniform fractions in the high molecular mass range, the summation A_i can be replaced by the continuous function $w(Y)$. For h=constant it follows that:

$$f(V) = \int w(Y) \left(h/\sqrt{\pi}\right) \exp\left[-h^2 (V+Y)^2\right] \quad (46)$$

where $w(Y)$ represents a chromatogram from which diffusion broadening has been removed. This can in turn be transformed into the molecular mass distribution by using a calibrated elution curve; various methods are available for this [99].

A second and very effective approach for determining a molecular mass distribution entails greater experimental effort: extrapolation to zero of the data from a gel chromatogram of a polymer in terms of both concentration and elution velocity [100].

GPC is a very flexible analytical method as a result of the availability of a wide variety of column packings and solvents. Calibration curves permit coverage of a molecular mass range extending from oligomers up to 10^6. GPC coupled with GC has become the most important method for the identification and molecular mass determination of oligomers.

Ultracentrifugation. The sedimentation constant

$$s_0 = \left(\frac{\dot{r}}{\omega^2 r}\right)_{\tau \to 0, M, T, S} \quad (47)$$

where $\dot{r} = dr/dt$ describes the rate of sedimentation (extrapolated to $c \to 0$) in a uniform field for a macromolecule of molecular mass M in the solvent S at temperature T. The dependency upon molecular mass is expressed by:

$$s_0 = k M^{\alpha} \quad (48)$$

A log – log plot of the experimental results gives a straight line. The observed value of the exponent is a function of the thermodynamic quality of the solvent. The smallest values are associated with sedimentation systems in which $\alpha = 0.5$. Numerous relationships based on Equation (48), calibrated by using absolute methods, have been tabulated [53]. For this reason the determination of molecular mass by sedimentation is also important as a relative method (for the absolute method, see Section 5.1.3).

Recording the polymer concentration during sedimentation as a function of time and location gives a set of curves in which the information relevant to molecular mass distribution is initially masked by other parameters influencing the separation process. Methods for determining molecular mass distribution by ultracentrifugation are described in detail in [78], [79], [83–85].

Ultracentrifugation is capable of producing forces up to 400 000 g. The lower limit for establishing a molecular mass distribution is $M = 10^4$. The unmatched flexibility of ultracentrifugation with respect to relative and absolute mass determinations as well as sample separation comes at the expense of high investment and operating costs, factors which must be taken into account in any comparison with GPC.

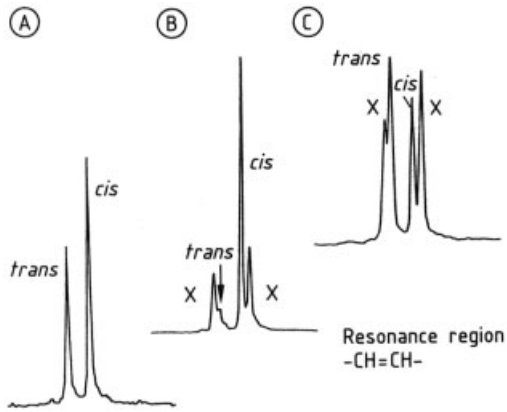

Figure 19. 67.88 MHz ^{13}C NMR spectrum of 1,4-polybutadiene
x = transition from *cis* to *trans* sequences; Solvent: deuterochloroform; Temperature: 300 K; All samples with random sequence distribution
A) Mixture of *cis*- and *trans*-1,4-polybutadiene; B) *cis*- with 20 % *trans*-1,4-polybutadiene sequences; C) 1,4-Polybutadiene with *cis* and *trans* sequences (1: 1)

6. Determination of Sequential Structure

Introduction. The properties and applications of plastics are determined largely by the

primary structures of their mainly linear macromolecules, which are formed by the sequential coupling of a single monomer or several chemically distinct species (see Chap. 7). One important type of problem in plastics analysis derives from the numerous types of isomerism, a situation that becomes increasingly complex in the case of co- and multipolymers [1], [104–109]. The sequence of two structural characteristics A and B can be block-like, alternating, or random. Structural units comprising *configurational sequences* of a single monomer contain pseudoasymmetric carbon atoms. Such polymers can be classed as isotactic (meso, m), syndiotactic (racemic, r), or atactic (random distribution of m and r linkages). Table 8 provides a summary of possible configurational sequences for discrete groups of three (triads) and five (pentads) structural units. The number of possible sequences N increases dramatically with the sequence length n:

Sequence length n:	2	3	4	5	6
Sequences $N(n)$:	2	3	6	10	20

Table 8. Tactical sequences in poly-α-olefins [3].

The monomer unit $-CH_2-CHR-$ is symbolized by .

	Sequence	Symbol	Probability
Triads	mm		P_m^2
	mr		$2P_m(1-P_m)$
	rr		$(1-P_m)^2$
Pentads	mmmm		P_m^4
	mmmr		$2P_m^3(1-P_m)$
	rmmr		$P_m^2(1-P_m)^2$
	mmrm		$2P_m^3(1-P_m)$
	mmrr		$2P_m^2(1-P_m)^2$
	rmrm		$2P_m^2(1-P_m)^2$
	rmrr	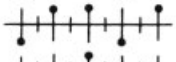	$2P_m(1-P_m)^3$
	mrrm	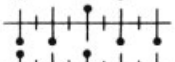	$P_m^2(1-P_m)^2$
	rrrm		$2P_m(1-P_m)^3$
	rrrr	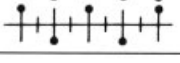	$(1-P_m)^4$

If addition of a new monomer unit occurs independently of the structure of the growing chain, then the probability P for a particular sequence can be shown to be $P_r = 1 - P_m$. One speaks in this case of Bernoulli statistics [107].

Configurational sequences also arise in the 1,4 addition of 1,3-dienes such as butadiene. The resulting polymer chain contains one C=C double bond per structural unit, which may have either the *cis* or *trans* configuration. The term *constitutional sequences* refers to the situation in which identical monomer units are capable of bonding in different ways (see Table 8). *Compositional sequences* are characteristic of copolymers into which two different comonomers have been incorporated [108].

6.1. NMR Spectroscopy

The most useful technique for the analysis of sequences in polymer chains is NMR spectroscopy. It often facilitates quantitative evaluation of even the smallest structural details of a polymer. For a comprehensive treatment of the fundamentals of NMR, see [109–119], [121]; general principles are treated in [106]. Nuclei that are important in the analysis of polymers by NMR are listed in Table 9; they suffice for characterizing all the important types of polymers.

Nuclear magnetic resonance spectroscopy of polymers exploits the fact that the resonance frequencies of the investigated nuclei depend strongly on the constitutional, configurational, and conformational structure of the polymer. Each structural feature plays a role, so that each of the sequences listed in Table 8 is associated with specific resonance signals. Primary structures have been assigned for a large number of polymers in solution on the basis of ^1H and ^{13}C NMR spectroscopy [3], [104], [113], and NMR spectroscopy of heteroelements is gaining in importance [115].

Poly(Vinyl Chloride). The ^{13}C NMR spectrum of poly(vinyl chloride) is an example of a simple atactic polymer.

In Figure 18 configurative sequences corresponding to those listed in Table 8 can be identified. Configurational differentiation is possi-

Table 9. NMR properties of nuclei important in the analysis of plastics [115]

Nucleus	Natural abundance	NMR frequency in 1 T magnetic field, MHz	Sensitivity relative to proton	Magnetogyric moment	Spin
^{1}H	99.9844	42.577	1.000	2.79270	1/2
^{2}H (D)	0.0156	6.536	0.00964	0.85738	1
^{13}C	1.108	10.705	0.0159	0.70216	1/2
^{14}N	99.635	3.076	0.00101	0.40357	1
^{15}N	0.365	4.315	0.00104	− 0.28304	1/2
^{17}O	0.037	5.772	0.0291	− 1.8930	5/2
^{19}F	100	40.055	0.834	2.6273	1/2
^{29}Si	4.70	8.460	0.0785	− 0.55477	1/2
^{31}P	100	17.235	0.0664	1.1305	1/2
^{35}Cl	75.4	4.172	0.00471	0.82089	3/2
^{37}Cl	24.6	3.472	0.00272	0.68329	3/2

ble for sequences of up to seven successive monomer units.

Polybutadiene. The ^{13}C NMR spectrum of polybutadiene is particularly instructive with respect to the effectiveness of the NMR method. Butadiene molecules can be linked in either a 1,4 or a 1,2 configuration

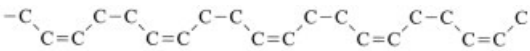

cis-1,4-block

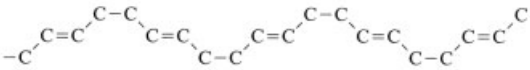

trans-1,4-block

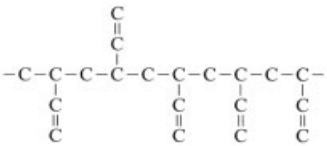

1,2-block

Figure 19 A shows a ^{13}C NMR spectrum of a mixture of *cis*-1,4- and *trans*-1,4-polybutadiene. Separate resonance signals are observed in the C=C double bond region for each of the possible structures. The transition from *cis* to *trans* and vice versa during isomerization of the polymers can also be observed by means of NMR spectroscopy (Fig. 19 B and C).

Depending on the polymerization conditions, polybutadiene can contain both 1,4 and 1,2 linkages. In the NMR spectrum (Fig. 20) it is possible to identify all eight of the possible triad sequences for combinations of *cis* (c), *trans* (t), and 1,2 (v) linkage between monomer units. Evaluation of the spectrum reveals that monomer addition in this sample was independent of the structure of the previously formed chain.

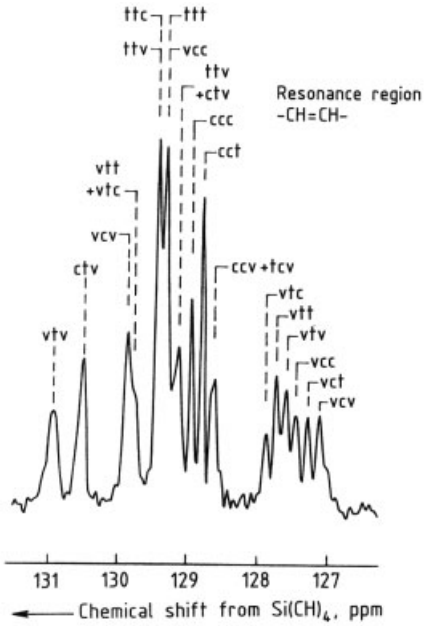

Figure 20. ^{13}C NMR spectrum of polybutadiene with 1,4- and 1,2-sequences
Conditions and sequence distribution as in Figure 19.

The 1,2 linkage of two butadiene units corresponds formally to an α-olefin, and it is in this way that the ^{13}C NMR spectrum in Figure 21 must be interpreted with respect to the configurative sequences listed in Table 8. The results are consistent with a Bernoulli distribution of the

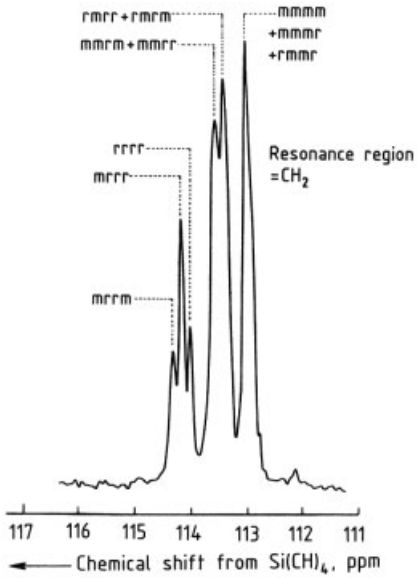

Figure 21. ^{13}C NMR spectrum of 1,2-polybutadiene at 67.88 MHz
Conditions and sequence distribution as in Figure 19.

sequences. The observed distribution depends upon the polymerization conditions.

2D NMR Spectroscopy. Two-dimensional NMR spectroscopy allows the investigation of the nature of magnetic interactions between adjacent nuclides. Verifying the presence of a bond between particular nuclides significantly simplifies the interpretation of an NMR spectrum. Figure 22 illustrates a typical coupling experiment [111], [114] conducted on a fluorine- containing polymer [119].

3D NMR spectroscopy is a tomographic procedure for solid polymers. In addition to providing primary structural information on insoluble polymers it can also be used to test the quality of finished products made of plastic. From a methodological standpoint 3D NMR bridges the gap between structural analysis and qual-

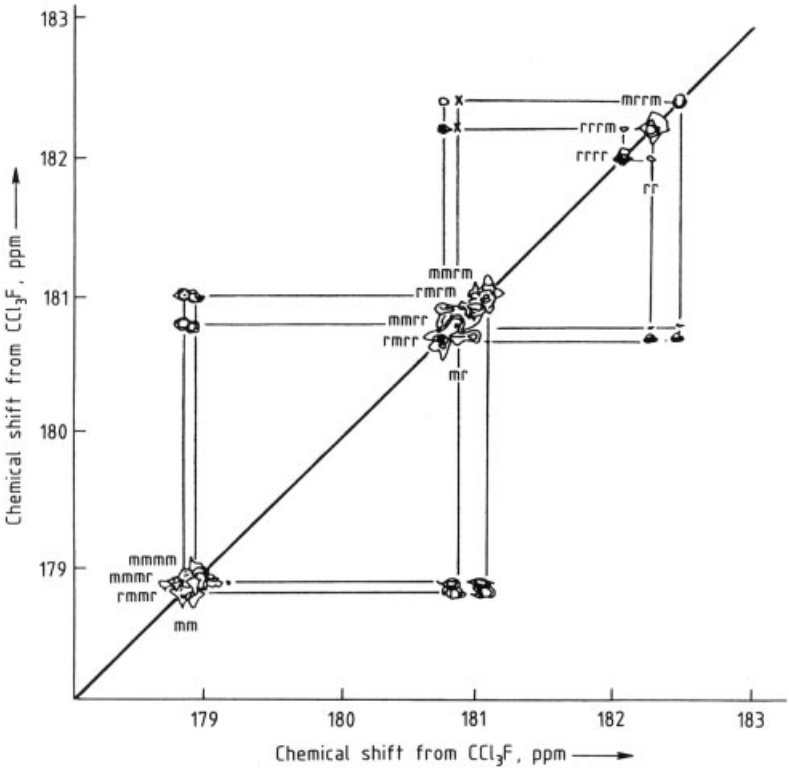

Figure 22. Two-dimensional ^{19}F NMR spectrum of poly(vinyl fluoride) [119]
Contour map of two-dimensional (refers to two different observation frequencies) J-correlated ^{19}F NMR spectrum. Sequences of five monomer units are assigned unambiguously.

ity control. Figure 23 shows the structure of a polybutadiene molding [120].

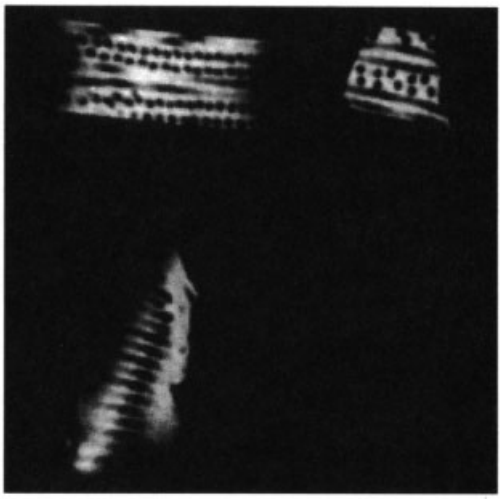

Figure 23. NMR micrograph of filled synthetic rubber [120] Three-dimensional (refers to space coordinates) ^{1}H NMR spectrum of a sample of reinforced tire rubber. Position of inhomogeneously distributed cord can be located without degradation of sample.

6.2. Pyrolysis Gas Chromatography

The pyrolytic methods described in Chapter 3 can be combined with gas chromatography [122–125] and mass spectrometry [126] to provide an extremely sensitive analysis of the chain structure of a polymer. Pyrolysis conditions are chosen such that the chain fragments released contain up to 30 monomer units, requiring temperatures of 500 – 800°C. The nature and number of observed fragments depends not only on the empirically determined experimental conditions but also on the composition and structure of the polymer (structural units, branching, tacticity, and sequence in the case of copolymers).

The pyrolysis of polyethylene proceeds as shown in Scheme 1. Simultaneous hydrogenation of the fragments often greatly simplifies the resulting pyrogram. The pyrogram of polyethylene consists exclusively of *n*-alkanes (Fig. 24). The presence of isoalkanes in a pyrogram is evidence for branching of the polymer chain. Under constant experimental conditions, each polymer gives a characteristic pyrogram, permitting structural assignment by comparison with standard pyrograms from structurally defined polymers. In favorable cases it is sometimes possible to identify stereoisomeric fragments [126] (Fig. 25).

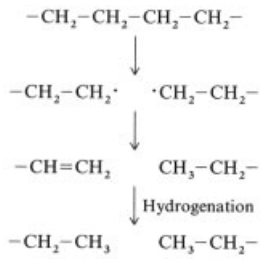

Scheme 1. Pyrolysis of polyethylene

Use of isotopically labeled polymers permits the application of pyrolysis chromatography in kinetic investigations (depolymerization, sequences in copolymers).

7. Determination of Chemical Heterogeneity

Methods that combine the online detection of chemical structure and molecular mass distribution are of growing importance.

Molecular Heterogeneity of Copolymers. Copolymerization of the monomers A and B can be described in terms of four different growth reactions:

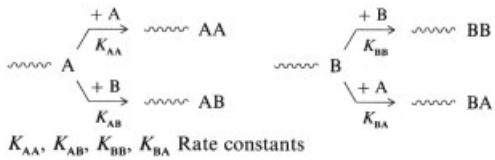

Monomers typically differ in their reactivity towards a growing chain. The reactivity relationship can be expressed in terms of two copolymerization parameters:

$$r_A = \frac{K_{AA}}{K_{AB}}\,;\; r_B = \frac{K_{BB}}{K_{BA}} \tag{49}$$

It follows for a radical copolymerization that the structure of emerging copolymer will vary as a function of the extent of conversion:

$$\frac{m_A}{m_B} \sim \frac{A}{B} \cdot \frac{r_A A + B}{r_B B + A} \tag{50}$$

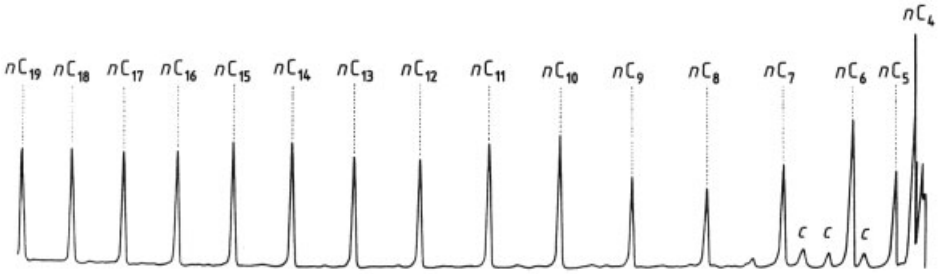

Figure 24. Pyrogram of polyethylene in a hydrogen atmosphere [124]
n = Linear olefins; c = Cyclic olefins

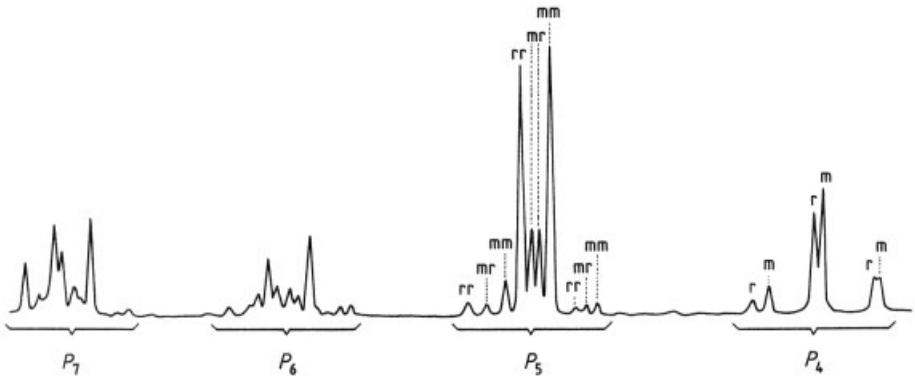

Figure 25. Pyrogram of isotactic polypropylene in a hydrogen atmosphere [124]
$P_4 - P_7$ = Degree of polymerization. For assignment of configurational sequences, see Table 8

A, B mole fractions of monomers A and B in initial monomer mixture
m_A, m_B mole fractions of monomers A and B in copolymer

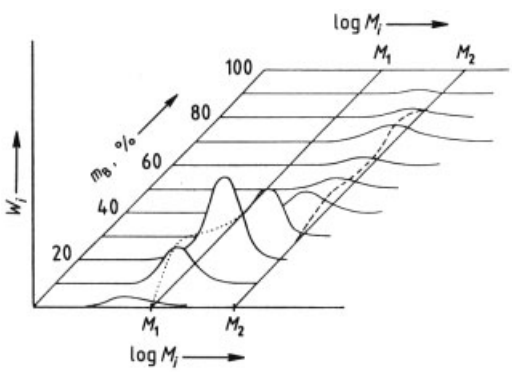

Figure 26. Heterogeneity of copolymer [9]
W_i = Frequency of the macromolecules with molecular mass M_i and chemical composition m_B

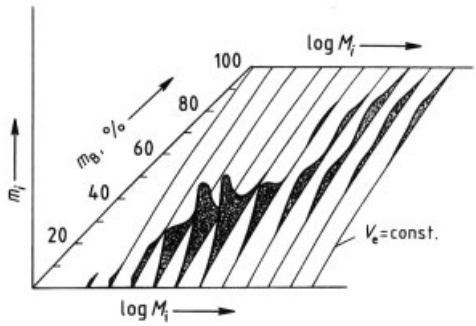

Figure 27. Distribution function of copolymer [9]
Methods: SEC and turbidity titration
m_i = Fraction of copolymer AB with molecular mass M_i and composition m_B; V_e = GPC elution volume

Generally, the macromolecules of a copolymer are heterogeneous with respect to molecular mass, chemical composition, and sequential structure. The differential distribution for a typical copolymer with respect to molecular mass and chemical composition is illustrated

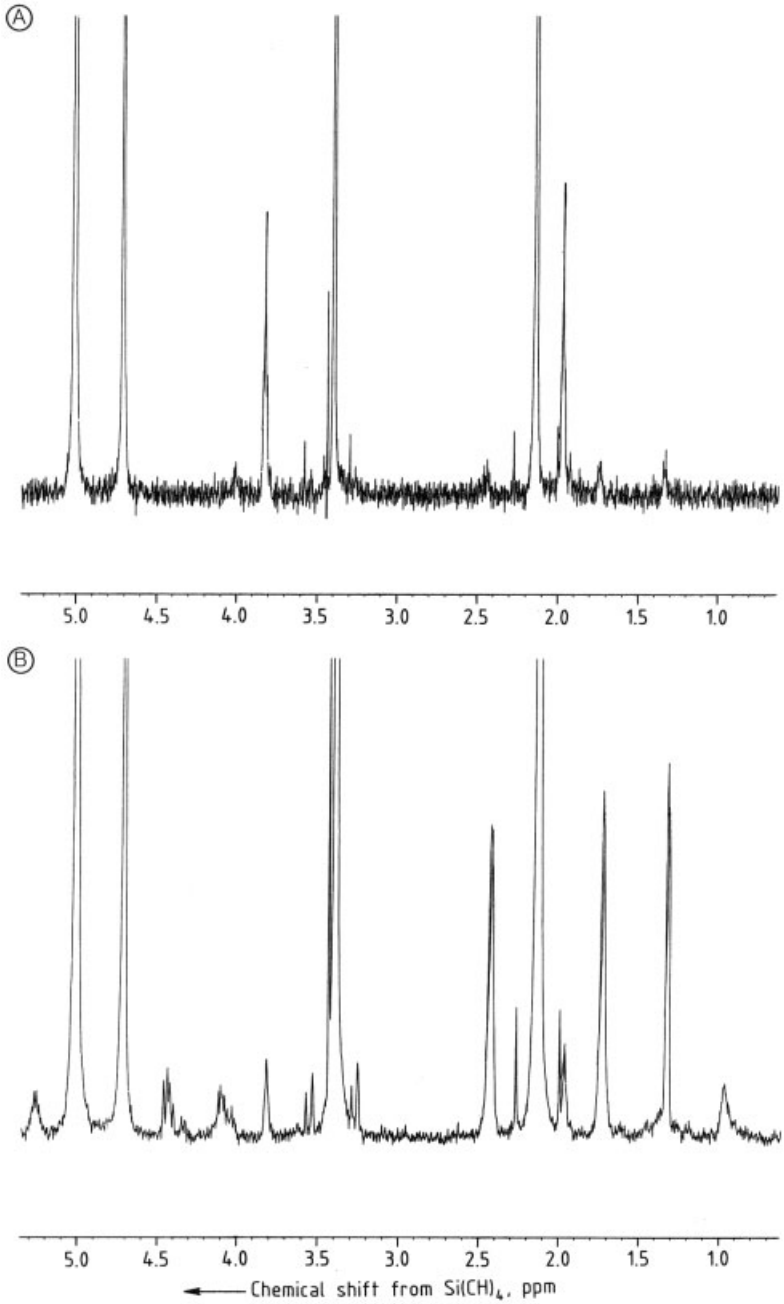

Figure 28. 500 MHz ^{1}H NMR spectrum of fractions from LC-NMR [127]
Online NMR spectra of the fractions 3 (A) and 4 (B) from Figure 17. The chemical structure of different components can be identified by analysis of the NMR spectrum.

in Figure 26. Macromolecules with a particular molecular mass M_1 vary in their composition m_B. By the same token, molecules of a particular composition m_B exhibit a molecular mass distribution. Such a distribution function for a copolymer cannot be established by a single method of separation, since the experimental parameter that distinguishes one molecule from another is no longer determined only by molecular mass, but also by chemical composition and sequential structure. As a result, several structure-specific separation techniques must be used.

This condition can be met to a first approximation by the application of turbidity titration to eluate fractions obtained by gel permeation chromatography (see page 23). A copolymer sample is first separated by gel chromatography to give a series of fractions containing molecules of comparable hydrodynamic volume, but not necessarily with the same molecular mass. Under certain conditions subsequent turbidity titration – essentially a precipitation fractionation monitored on the basis of light scattered laterally at an angle of 90°C (see Section 5.1.4) – makes it possible to determine the chemical composition of the polymer. Here both the amount of precipitant and the rate of precipitant addition are characteristic of the composition of the copolymer, whereas these factors are largely insensitive to molecular mass. The result is a differential distribution function for the copolymer, as shown in Figure 27 [9].

Structural Heterogeneity of Oligomers. Coupling of gradient HPLC and NMR spectroscopy opens up new ways for the online analysis of polymerization reactions. This method is applicable to polymers containing heteroelements such as ^{19}F, ^{31}P etc. The minimal amount of substance for detection of a fraction is ca. 30 µg. The spectra in Figure 28 demonstrate the considerable potential of the method [127].

8. References

1. J. L. Koenig: *Spectroscopy of Polymers,* American Chemical Society, Washington 1991.
2. C. D. Craver (ed.): "Polymer Characterization," *Advances in Chemistry Series,* vol. 203, American Chemical Society, Washington 1983.
3. D. O. Hummel, F. Scholl: *Atlas of Polymer and Plastic Analysis,* vols. 1 – 3, VCH, Weinheim 1991.
4. D. McIntyre (ed.): *Characterization of Macromolecular Structure,* National Academy of Sciences, Washington 1968.
5. J. V. Dawkins (ed.): *Developments in Polymer Characterization,* Appl. Science Publishers, London 1982.
6. G. M. Kline: "Analysis of Monomers and Polymer Materials," *Analytical Chemistry of Polymers,* **part 1,** Interscience Publ., New York 1959.
7. G. M. Kline: "Analysis of Molecular Structure," *Analytical Chemistry of Polymers,* **part 2,** Interscience Publ., New York 1960.
8. G. M. Kline: "Identification Procedures and Chemical Analysis," *Analytical Chemistry of Polymers,* **part 3,** Interscience Publ., New York 1962.
9. M. Hoffmann, H. Krömer, R. Kuhn: *Polymeranalytik,* vols. 1 and 2, Thieme Verlag, Stuttgart 1976.
10. R. Nitsche, K. A. Wolf: *Kunststoffe, Struktur, physikalisches Verhalten und Prüfung,* Springer Verlag, Berlin 1961.
11. H. Batzer (ed.): *Polymere Werkstoffe,* vols. 1 – 3, Thieme Verlag, Stuttgart 1985.
12. A. Cooper (ed.): *Determination of Molecular Weight,* John Wiley and Sons, New York 1989.
13. R. A. Komoroski: *NMR Spectroscopy of Synthetic Polymers in Bulk,* VCH, Weinheim 1986.
14. L. S. Bark, N. S. Allen: *Analysis of Polymer Systems,* Applied Science Publishers, London 1983.
15. F. Rodriguez: *Principles of Polymer Systems,* Hemishere Publishing Corp., New York 1989.
16. L. Mandelkern: *An Introduction to Macromolecules,* Springer-Verlag, New York 1983.
17. S. L. Rosen: *Fundamental Principles of Polymeric Material,* Wiley-Interscience, New York 1982.
18. D. McIntyre: *Characterization of Macromolecular Structure,* Nat. Acad. Sciences, Publ. 1573, Washington 1968.
19. J. L. Koenig: *Chemical Microstructure of Polymer Chains,* Wiley-Interscience, New York 1980.
20. J. F. Rabek: *Experimental Methods in Polymer Chemistry,* Wiley-Interscience, New York 1980.

21. B. H. Vollmert, *Polymer Chemistry,* Springer-Verlag, New York 1973.
22. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca 1953.
23. P. E. M. Allen, C. R. Patrick: *Kinetics and Mechanisms of Polymerisation Reactions,* Wiley-Interscience, New York 1974.
24. C. Tanford: *Physical Chemistry of Macromolecules,* Wiley-Interscience, New York 1961.
25. G. M. Bartenev, J. V. Zelenev: *Physik der Hochpolymere,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1979.
26. B. Wunderlich: *Macromolecular Physics,* vols. 1 and 2, Academic Press, New York 1973.
27. C. D. Craver in T. Provder (ed.): *Polymer Characterization: Physical Property, Spectroscopic and Chromatographic Methods,* American Chemical Society, Washington 1990.
28. P. J. Flory: *Statistical Mechanics of Chain Molecules,* Interscience Publ., Wiley 1969.
29. F. A. Bovey, F. H. Winslow: *Macromolecules,* Academic Press, New York 1979.
30. A. J. Hopfinger: *Conformational Properties of Macromolecules,* Academic Press, New York 1973.
31. F. Scholl, R. Vieweg, D. Braun: *Kunststoffhandbuch,* vol. 1, Hanser Verlag, München 1975.
32. G. Bandel, R. Houwink: *Chemie und Technologie der Kunststoffe,* vol. 1, Akademische Verlagsges., Leipzig 1954.
33. B. Carroll: *Physical Methods in Polymer Chemistry,* vol. 1, Marcel Dekker, New York 1969.
34. N. W. Johnston, H. J. Harwood, *Macromolecules* **3** (1970) 20.
35. H. Domininghaus: *Die Kunststoffe und ihre Eigenschaften,* VDI Verlag, Düsseldorf 1976.
36. G. C. Ives, J. A. Mead, M. M. Riley: *Handbook of Plastic Test Methods,* Iliffe Books, London 1971.
37. D. McIntire: *Characterization of Macromolecular Structure,* Nat. Acad. Sci., Washington, D.C., 1968.
38. P. W. Allen: *Techniques of Polymer Characterization,* Butterworth, London 1959.
39. M. Stratmann: *Erkennen und Identifizierung der Faserstoffe,* Spohr Verlag, Stuttgart 1973.
40. G. J. Domsch, *Kunststoffe* **61** (1971) 669.
41. E. A. Turi (ed.): *Thermal Analysis in Polymer Characterization,* Heyden, Philadelphia 1981.
42. H. G. Kilian in A. Weiss (ed.): *Progress in Colloid & Polymer Science,* vols. 1 – 86, Steinkopf Verlag, Darmstadt 1991.
43. A. Abe, H. Benoit in H. J. Cantow (ed.): *Progress in Polymer Science,* vols. 1 – 101, Springer Verlag, New York 1991.
44. S. R. Palit, B. M. Mandal, *J. Macromol. Sci. Rev. Macromol. Chem.* **2** (1968) 225.
45. J. Mitchell, Jen Chiu, *Anal. Chem.* **43** (1971) 267.
46. J. M. Zeigler, F. W. G. Fearon: *Silicon-based Polymer Science: A Comprehensive Resource,* American Chemical Society, Washington 1989.
47. R. A. Dickie, S. S. Labana, R. S. Bauer: *Cross-Linked Polymers: Chemistry, Properties and Applications,* American Chemical Society, Washington 1989.
48. T. R. Crompton: *The Analysis of Plastics,* Pergamon Press, Oxford 1984.
49. W. C. Wake, M. J. Loadman in B. K. Tidd (ed.): *Analysis of Rubber and Rubber-like Polymers,* Elsevier, Amsterdam 1983.
50. E. Schröder, G. Müller, K.-F. Arndt: *Polymer Characterization,* Oxford University Press, New York 1989.
51. A. L. Smith: *Analysis of Silicones,* Wiley-Interscience, New York 1974.
52. R. Taprogge: *Faserverstärkte Hochleistungsverbundstoffe,* Vogel Verlag, Würzburg 1974.
53. J. Brandrup, E. H. Immergut: *Polymer Handbook,* 3rd. ed., Wiley-Interscience, New York 1984.
54. J. Urbanski, W. Czerwinski, F. Majewska: *Handbook of Analysis of Synthetic Polymers and Plastics,* John Wiley and Sons, New York 1977.
55. H. Mark, N. G. Gaylord, N. M. Bikales: *Encyclopedia of Polymer Science and Technology,* vols. 1 – 16, Interscience Publ., New York 1964.
56. E. Müller (ed.): "Makromolekulare Stoffe," *Methoden der Organischen Chemie,* **vol. 14, part 1/2,** Georg Thieme Verlag, Stuttgart 1961.
57. R. Vieweg, D. Braun: *Kunststoffhandbuch,* vols. 1 –9, Hanser Verlag, München 1963.
58. American Chemical Society: *Chemical Abstracts and CAS Online,* Washington.
59. DKI, Deutsches Kunststoff Institut: *DKI Datenbank,* Darmstadt.
60. T. R. Crompton: *Chemical Analysis of Additives in Plastics,* Pergamon Press, Oxford 1971.
61. M. Wandel, H. Tengler, H. Ostramow: *Die Analyse von Weichmachern,* Springer-Verlag, Berlin 1967.

62. A. Krause, A. Lange, M. Ezrin: *Plastics Analysis Guide,* Hanser Verlag, München 1983.
63. J. Mitchell, Jr. (ed.): *Applied Polymer Analysis and Characterization,* Hanser Verlag, München 1987.
64. H. Morawetz: *Macromolecules in Solution,* Interscience, New York 1965.
65. A. F. M. Barton (ed.): *CRC Handbook of Solubility Parameters and other Cohesion Parameters,* CRC Press Inc., Ohio, 1983.
66. P. A. Bristow: *Liquid Chromatography in Practise,* hept, Wilmslow 1976.
67. A. Zlatkis, R. E. Kaiser, *J. Chromatogr. Libr.* **9** (1976).
68. J. Cazes: *Liquid Chromatography of Polymers and Related Materials,* Marcel Dekker, New York 1977.
69. A. Krause, A. Lange: *Kunststoff-Bestimmungsmöglichkeiten,* Hanser Verlag, München 1970.
70. D. Braun: *Simple Methods for Identification of Plastics,* Macmillan, New York 1982.
71. D. Braun, J. C. Jung, *Asbest Kunstst.* **23** (1970) 651.
72. W. Brügel: *Einführung in die Ultrarotspektroskopie,* 4th ed., Steinkopf Verlag, Darmstadt 1969.
73. S. Krimm, *Fortschr. Hochpolym. Forsch.* **2** (1960) 51.
74. A. Elliott: *Infrared Spectra and Structure of Organic Long Chain Polymers,* Arnold Verlag, London 1969.
75. R. Zbinden: *Infrared Spectroscopy of High Polymers,* Academic Press, New York 1964.
76. G. Schnell, *Ber. Bunsenges. Phys. Chem.* **70** (1966) 297.
77. N. J. Harrick: *Internal Reflectance Spectrometry,* Interscience Publ., New York 1967.
78. P. E. Slade Jr.: *Polymer Molecular Weights,* **vols. 1,** 2, Marcel Dekker, New York 1975.
79. L. H. Peebles Jr.: *Molecular Weight Distributions in Polymers,* Interscience Publ., New York 1971.
80. J. S. Mattson, H. B. Mark, H. C. MacDonald Jr.: *Computers in Polymer Science,* Marcel Dekker, New York 1977.
81. T. Provder (ed.): *Computer Applications in Applied Polymer Science II,* American Chemical Society, Washington 1989.
82. G. C. Lowry (ed.): *Markov Chains and Monte Carlo Calculations in Polymer Science,* Marcel Dekker, New York 1970.
83. H. Fujita: *Mathematical Theory of Sedimentation Analysis,* Academic Press, New York 1962.
84. J. W. Williams (ed.): *Ultracentrifugal Analysis,* Academic Press, New York 1963.
85. H. K. Schachmann: *Ultracentrifugation in Biochemistry,* Academic Press, New York 1959.
86. H. A. Stuart: *Die Physik der Hochpolymeren,* vol. 1, Springer-Verlag, Berlin 1952.
87. M. B. Huglin: *Light Scattering from Polymer Solutions,* Academic Press, London 1972.
88. H. Benoit, *Ber. Bunsenges. Phys. Chem.* **70** (1966) 286.
89. J. Schurz: *Viskositätsmessungen an Hochpolymeren,* Kohlhammer Verlag, Stuttgart 1972.
90. G. Meyerhoff, *Fortschr. Hochpolym. Forsch.* **7** (1970) 1.
91. G. V. Schulz, H. A. Stuart: *Physik der Hochpolymeren,* vol. 2, Springer-Verlag, Berlin 1953.
92. B. Vollmert: *Polymer Chemistry,* Springer-Verlag, Berlin 1973.
93. M. J. R. Cantow: *Polymer Fractionation,* Academic Press, New York 1969.
94. K. H. Altgelt, L. Segal: *Gel Permeation Chromatography,* Marcel Dekker, New York 1971.
95. H. Determann: *Gel-Chromatographie,* Springer-Verlag, Berlin 1967.
96. H. Engelhardt: *Hochdruck-Flüssigkeits-Chromatographie,* Springer-Verlag, Berlin 1975.
97. T. Provder (ed.): *Detection and Analysis in Size Exclusion Chromatography,* American Chemical Society, Washington 1987.
98. T. Provder (ed.): "Size Exclusion Chromatography," ACS Symp. Ser. **254** (1984).
99. H. L. Tung: *Fractionation of Synthetic Polymers,* Marcel Dekker, New York 1977.
100. J. Cazes (ed.): "Liquid Chromatography of Polymers," *Chromatographic Science,* **vol. 19,** Marcel Dekker, New York 1981.
101. W. Heitz, *Angew. Chem.* **82** (1970) 675.
102. K. C. Berger, G. V. Schulz, *Makromol. Chem.* **136** (1970) 221.
103. H.-J. Cantow, *Makromol. Chem.* **30** (1959) 169.
104. F. A. Bovey: *Polymer Conformation and Configuration,* Academic Press, New York 1969.
105. H. Günther: *NMR-Spektroskopie,* Thieme Verlag, Stuttgart 1973.

106. T. C. Farrar, E. D. Becker: *Pulse and Fourier Transform NMR,* Academic Press, New York 1971.
107. F. A. Bovey: *High Resolution NMR of Macromolecules,* Academic Press, New York 1972.
108. E. Klesper, G. Sielaff, D. O. Hummel: *Polymer Spectroscopy,* Verlag Chemie, Weinheim 1974.
109. J. Schaefer, G. C. Levy: *Carbon-13 NMR Spectroscopy,* Wiley-Interscience, New York 1974.
110. Q. T. Pham, R. Petiaud, L. M. France: *Proton and Carbon NMR Spectra of Polymers,* John Wiley and Sons, Chichester 1984.
111. W. R. Croasmun, R. M. K. Carlson: *Two-Dimensional NMR Spectroscopy,* VCH Publishers, New York 1987.
112. W. M. Pasika: *Carbon-13 in Polymer Science,* American Chemical Society, Washinton 1979.
113. A. E. Tonelli: *NMR-Spectroscopy and Polymer Microstructure,* VCH Publishers, New York 1989.
114. G. E. Martin, A. S. Zektzer: *Two-Dimensional NMR Methods for Establishing Molecular Connectivity,* VCH Publishers, New York 1988.
115. R. K. Harris, B. E. Mann: *NMR and the Periodic Table,* Academic Press, London 1978.
116. G. C. Levy (ed.): *Topics in Carbon-13-NMR Spectroscopy,* vols. 1 – 4, Wiley-Interscience, New York 1974.
117. J. B. Strothers: *Carbon-13 NMR Spectroscopy,* Academic Press, New York 1972.
118. R. R. Ernst, G. Bodenhausen, A. Wokau: *Principles of Nuclear Magnetic Resonance in One and Two Dimensions,* Clarendon Press, Oxford 1991.
119. M. Bruch, F. A. Bovey, *Macromolecules* **17** (1984) 2547.
120. D. Groß, V. Lehmann, W. E. Hull: *NMR Microscopy, Applications in Biology, Biomedicine and Material Science,* Bruker Analytische Meßtechnik, Rheinstetten 1991.
121. R. K. Harris (ed.): *Nuclear Magnetic Resonance (A Specialist Periodical Report),* vols. 1 – 18, The Chemical Society, London 1972.
122. M. P. Stevens: *Characterization and Analysis of Polymers by Gas Chromatography,* Marcel Dekker, New York 1969.
123. O. Hummel: *Polymer Spectroscopy,* Verlag Chemie, Weinheim 1974.
124. M. Seeger, H.-J. Cantow, *Makromol. Chem.* **176** (1975) 2059.
125. S. A. Liebmann, E. J. Levy: "Pyrolysis and GC in Polymer Analysis," *Chromatographic Science Series,* **vol. 29,** Marcel Dekker, New York 1985.
126. A. L. Yergey, C. G. Edmonds, I. A. S. Lewis: *Liquid Chromatography/Mass Spectrometry,* Plenum Press, New York 1990.
127. M. Spraul, M. Hofmann, H. Glauner: *LC-NMR – New Life for an Old Technique,* Bruker Report 2/1990, Bruker Analytische Meßtechnik, Rheinstetten1990.

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Plastics, General Survey

Standard Article

Hans-Georg Elias[1]
[1]Michigan Molecular Institute, Midland, MI 48 640, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a20_543
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Polymers
1.1.1. Fundamental Terms
1.1.2. Nomenclature
1.2. Plastics
1.2.1. Fundamental Terms
1.2.2. Designations
1.3. History of Plastics
1.4. Economic Importance
2. Molecular Structure of Polymers
2.1. Constitution
2.1.1. Homopolymers
2.1.2. Copolymers
2.1.3. Branched Polymers
2.1.4. Ordered Chain Assemblies
2.1.5. Unordered Networks
2.2. Molar Masses and Molar Mass Distributions
2.2.1. Molar Mass Average
2.2.2. Determination of Molar Mass
2.2.3. Molar Mass Distributions
2.2.4. Determination of Molar Mass Distributions
2.3. Configurations
2.4. Conformations
2.4.1. Microconformations
2.4.2. Conformations in Ideal Polymer Crystals
2.4.3. Conformations in Polymer Solutions
2.4.4. Unperturbed Coils
2.4.5. Perturbed Coils
2.4.6. Wormlike Chains
3. Polymer Manufacture
3.1. Raw Materials
3.1.1. Wood
3.1.2. Coal
3.1.3. Natural Gas
3.1.4. Petroleum
3.1.5. Other Natural Raw Materials
3.2. Polymer Syntheses: Overview
3.2.1. Classifications
3.2.2. Functionality
3.2.3. Cyclopolymerizations
3.3. Chain Polymerizations and Polyeliminations
3.3.1. Thermodynamics
3.3.2. Free-Radical Polymerizations
3.3.3. Anionic Polymerizations
3.3.4. Cationic Polymerizations
3.3.5. Ziegler–Natta Polymerizations
3.3.6. Copolymerizations
3.4. Polycondensations and Polyadditions
3.4.1. Bifunctional Polycondensations
3.4.2. Multifunctional Polycondensations
3.4.3. Polyaddition
3.5. Polymer Reactions
3.5.1. Polymer Transformations
3.5.2. Block and Graft Formations
3.5.3. Cross-Linking Reactions
3.5.4. Degradation Reactions
4. Plastics Manufacture
4.1. Homogenization and Compounding
4.2. Additives
4.2.1. Overview
4.2.2. Chemofunctional Additives
4.2.3. Processing Aids

4.2.4. Extending and Reinforcing Fillers
4.2.5. Plasticizers
4.2.6. Colorants
4.2.7. Blowing Agents
4.3. Processing
5. Supermolecular Structures
5.1. Noncrystalline States
5.1.1. Structure
5.1.2. Orientation
5.2. Polymer Crystals
5.2.1. Introduction
5.2.2. Crystal Structures
5.2.3. Crystallinity
5.2.4. Morphology
5.3. Mesophases
5.3.1. Introduction
5.3.2. Types of Liquid Crystals
5.3.3. Mesogens
5.3.4. Lyotropic Liquid Crystalline Polymers
5.3.5. Thermotropic Liquid Crystal Polymers
5.3.6. Block Copolymers
5.3.7. Ionomers
5.4. Gels
5.5. Polymer Surfaces
5.5.1. Structure
5.5.2. Interfacial Tension
5.5.3. Adsorption
6. Thermal Properties
6.1. Molecular Motion
6.1.1. Thermal Expansion
6.1.2. Heat Capacity
6.1.3. Heat Conductivity
6.2. Thermal Transitions and Relaxations
6.2.1. Overview
6.2.2. Crystallization
6.2.3. Melting
6.2.4. Liquid Crystal Transitions
6.2.5. Glass Transitions
6.2.6. Other Transitions and Relaxations
6.2.7. Technical Methods
6.3. Transport
6.3.1. Self-Diffusion
6.3.2. Permeation
7. Rheological Properties
7.1. Introduction
7.2. Shear Viscosities at Rest
7.2.1. Fundamentals
7.2.2. Molar Mass Dependence
7.2.3. Concentrated Solutions
7.3. Non-Newtonian Shear Viscosities
7.3.1. Overview
7.3.2. Flow Curves
7.3.3. Melt Viscosities
7.4. Extensional Viscosities
7.4.1. Fundamentals
7.4.2. Melts
7.4.3. Solutions
8. Mechanical Properties
8.1. Introduction
8.1.1. Deformation of Polymers
8.1.2. Tensile Tests
8.1.3. Moduli and Poisson Ratios
8.2. Energy Elasticity
8.2.1. Theoretical Moduli
8.2.2. Real Moduli
8.2.3. Temperature Dependence
8.3. Entropy Elasticity
8.4. Viscoelasticity
8.4.1. Fundamentals
8.4.2. Time–Temperature Superposition
8.5. Dynamic Behavior
8.5.1. Fundamentals
8.5.2. Molecular Interpretations
8.6. Fracture
8.6.1. Overview
8.6.2. Theoretical Fracture Strength
8.6.3. Real Fracture Strength
8.6.4. Impact Resistance
8.6.5. Stress Cracking
8.6.6. Fatigue
8.7. Surface Mechanics
8.7.1. Hardness

8.7.2. Friction
8.7.3. Abrasion and Wear
9. Electric Properties
9.1. Dielectric Properties
9.1.1. Relative Permittivity
9.1.2. Dielectric Loss
9.1.3. Dielectric Strength and Tracking Resistance
9.1.4. Electrostatic Charging
9.2. Electrical Conductivity
9.3. Photoconductivity
10. Optical Properties
11. Polymer Composites
11.1. Introduction
11.1.1. Overview
11.1.2. Mixture Rules
11.2. Filled Polymers
11.2.1. Microcomposites
11.2.2. Molecular Composites
11.2.3. Macrocomposites
11.3. Homogeneous Blends
11.3.1. Thermodynamics
11.3.2. Plastification
11.3.3. Rubber Blends
11.3.4. Plastic Blends
11.4. Heterogeneous Blends
11.4.1. Compatible Polymers
11.4.2. Blend Formation
11.4.3. Toughened Plastics
11.4.4. Thermoplastic Elastomers
11.5. Expanded Plastics
12. Waste Disposal

[Top of Page]

1. Introduction

Plastics are commercially used materials that are based on polymers or prepolymers. The name plastics refers to their easy processibility and shaping (Greek: *plastein* = to form, to shape).

Plastics and polymers are not synonyms. Polymers or prepolymers are raw materials for plastics; they become plastics only after physical compounding and/or chemical hardening. The same polymers may be used not as plastics but as fibers, in paints, or in other applications. Other polymers may be utilized as fibers, elastomers, thickeners, ion-exchange resins, etc., but not as plastics.

1.1. Polymers

1.1.1. *Fundamental Terms [1-5, 8-14]*

Polymers are chemical substances (chemical compounds) composed of polymer molecules. The term polymer refers to molecules composed of many units (Greek: *poly* = many, *meros* = parts). Polymer molecules may thus consist of many atoms, usually a thousand or more, thereby having high molar masses (“molecular weights”). Benzene can, for example, be “polymerized” from three acetylene molecules and was originally called a polymer. What is now called a polymer consists of molecules with hundreds and thousands of such units; it was therefore termed “high polymer” in the older literature.

The term polymer carries with it the connotation of molecules composed of many equal “mers”,such as the ethylene units in polyethylene, $R'(CH_2CH_2)_nR''$. The number *n* of mers in a polymer molecule is called the degree of polymerization *X*. There are, however, many polymer molecules (especially biopolymer molecules) with very different types of mers per molecule, such as protein molecules $H(NH – CHR – CO)_nOH$ with up to 16 different R substituents in irregular sequence. A less constraining and more general term for polymer molecule is thus macromolecule. No sharp dividing line with respect to the number of units per molecule exists, however, between macromolecules and low molar mass compounds (micromolecules).

Macromolecules constitute the simplest individual chemical constituents of a polymer. A macromolecule may exist as an individual entity (in linear and branched polymers) or may be thought of as the primary molecule before chemical cross-linking.

Macromolecules exist in nature; examples are nucleic acids, proteins, polysaccharides, polyprenes, and lignins. Some of these naturally occurring polymers are used by man without further chemical transformation (e.g., cellulose for paper and cardboard). The chemical transformation of natural polymers with retention of their chain structures leads to semisynthetic materials, for example, cellulose acetates from cellulose. Chains of other natural polymers are cross-linked before commercial use. Examples are the hardening of casein (a protein) by formaldehyde to galalith (plastic) or the vulcanization of *cis*-1,4-polyisoprene (natural rubber) to an elastomer.

Most polymers are, however, synthesized chemically from molecules with low molar masses, the so- called monomers. Examples are the preparations of polyethylene from ethylene, poly(vinyl chloride) from vinyl chloride, nylon 6 from ∈- caprolactam, or nylon 66 from adipic acid and hexamethylenediamine. Some industrial polymers result from the chemical conversion of other synthetic polymers, for example, poly(vinyl alcohol) from poly(vinyl acetate).

1.1.2. *Nomenclature [1], [5], [15]*

The nomenclature of individual polymers and plastics is as confusing as their classification according to properties. Various systems of nomenclature are used simultaneously, often by the same author. Abbreviations and acronyms abound, sometimes with different meanings for the same letter combinations and other times without explanation. In addition, about 25 000 trade names are used worldwide for plastics, fibers, and elastomers. Furthermore, a polymer from a certain company may come in many different grades depending on the processibility and application, sometimes up to 100 per polymer type. Some of these grades may even bear different trade names for various applications. On the other hand, the same trade name is occasionally used for plastics based on very different polymer types. An example is Bexloy of Du Pont, which may be an amorphous polyamide (Bexloy C), a high-impact polyester (Bexloy J), an ionomer (Bexloy W), or a thermoplastic elastomer (Bexloy V).

The following nomenclature systems are commonly used for polymers:

Long-known Natural Polymers often have Trivial Names. Examples are cellulose, the polymeric sugar (-ose) of the plant cell; casein, the most important protein of milk and cheese (Latin: *caseus* = cheese); nucleic acids, the acids found in the cell nucleus; catalase, a catalyzing enzyme.

Synthetic Polymers are often Named after Their Monomers. Polymers of ethylene thus lead to polyethylene, those of styrene to polystyrene, those of vinyl chloride to poly(vinyl chloride), and those of a lactam to a polylactam. This “polymonomer” nomenclature has the disadvantage that the constitution of monomeric units of the polymer molecules is not identical with the constitution of the monomers themselves. For example, the polymerization of ethylene, $CH_2=CH_2$, leads to $\sim(CH_2 – CH_2)_n\sim$, a saturated compound and thus a polyalkane, not an unsaturated “ene” as the name polyethylene may suggest. The polymerization of lactams (cyclic amides) does not give macromolecules with intact lactam rings in the polymer chains but gives open- chain polyamides, etc. This polymonomer scheme is also ambiguous if a monomer can lead to more than one characteristic unit in a polymer. An example is acrolein, $CH_2=CH(CHO)$, which can polymerize via the ethylenic double bond to give $\sim[CH_2 – CH(CHO)]_n\sim$, via the aldehyde group to $\sim[O – CH(CH=CH_2)]_n$, or via both to give six-membered rings in polymer chains.

For trade purposes, certain polymer names may denote not only homopolymers but also copolymers, contrary to what the “chemical” names imply. For example, the copolymers of ethylene with up to 10 % butene-1, hexene-1, or octene-1 are known as linear low-density polyethylenes (LLDPEs). The commonly used chemical names of

plastics thus often do not indicate the true chemical structure of the monomeric units of the polymers on which they are based.

Polymers are often named after Characteristic Groups in Their Repeating Units. Polyamides are thus polymers with amide groups – NHCO – in their repeating units; for example ~$[NHCO(CH_2)_5]_n$~ = polyamide 6 = nylon 6 = poly(∈- caprolactam). Other examples are polyesters with ester groups – COO – or polyurethanes with urethane groups – NH – CO – O – in the chains. A disadvantage is that this naming scheme is identical with that of organic chemistry where a polyisocyanate denotes a low molar mass compound with more than one isocyanate group per molecule [e.g., $C_6H_3(NCO)_3$]. A macromolecular polyisocyanate would thus be a polymer with many intact isocyanate groups per chain, for example, poly(vinyl isocyanate) ~$[CH_2 – CH(NCO)]_n$~. The poly(isocyanate)s of polymer chemistry, on the other hand, possess polymerized isocyanate groups as, for example, in ~$(NR – CO)_n$~. Such compounds are unfortunately also often called polyisocyanates.

IUPAC Names. IUPAC recommends the use of constitutive names, similar to those used in inorganic and organic chemistry. The nomenclature of low and high molar mass inorganic molecules follows the additivity principle; those of low molar mass organic molecules the substitution principle. The nomenclature of organic macromolecules is a hybrid of both principles: the smallest repeating units are thought of as biradicals according to the substitution principle; then their names are added according to the additivity principle, put in parentheses, and prefixed with "poly." Names of repeating units are written without spaces between words. The polymer ~$[O – CH_2]_n$~ from formaldehyde, $H_2C=O$, is thus called poly(oxymethylene). The polycondensation of ethylene glycol $HO – CH_2 – CH_2 – OH$ with terephthalic acid $HOOC – (p\text{-}C_6H_4) – COOH$ leads to a polymer ~$[O – CH_2CH_2 – O – OC(p\text{-}C_6H_4)CO]_n$~ with the systematic name poly(oxyethyleneoxyterephthaloyl). The trivial names of this polymer are polyethylene terephthalate, and poly(ethylene glycol terephthalate), poly(ethylene terephthalate). It is also known as "saturated polyester" (although it is only one of the many possible saturated polyesters), as PET (an acronym) or PETP (an abbreviation) in the plastics literature, by the acronym PES in the fiber literature, and as PETE for recycling purposes.

IUPAC names are rarely used in the plastics literature. They are however important for systematic searches in *Chemical Abstracts* and other literature services.

1.2. Plastics

1.2.1. *Fundamental Terms [16-39, 41-53]*

Early plastics resembled natural resins. These natural resins are organic solids that break with a conchoidal fracture in contrast to the planar surfaces created upon the fracture of crystalline materials or the drawn-out zones formed upon the breaking of gums and waxes. Natural resin refers mainly to oleoresins from tree sap but is also used for shellac, insect exudations, and mineral hydrocarbons (⟶ Resins, Natural).

Early plastics were thus sometimes called synthetic resins. The word resin is today occasionally used for any organic chemical compound with medium to high molar mass that serves as a raw material for plastics (for a definition of the term resin according to current standards, see ⟶ Resins, Synthetic). Resin is not to be confused with rosin, which refers to mixtures of C_{20} fused-ring monocarboxylic acids such as pine oil, tall oil, and kauri resin. Rosin is the main component of naval stores (⟶ Resins, Natural).

Plastics are usually divided into two groups according to their hardening processes. Those that yield solid materials by simple cooling of a polymer melt (a physical process) and soften while being heated are called *thermoplastics*. *Thermosets*, on the other hand, harden through chemical cross-linking reactions between polymer molecules; when heated, they do not soften but decompose chemically (⟶ Thermosets). The shaping of a thermoplastic is thus a reversible process: the same material can be melted and processed again. A thermoset cannot be remelted and reshaped; its formation is irreversible.

Thermoplastics are normally composed of fairly high molar mass molecules because many physical properties effectively become molar mass independent only above a certain molar mass. Examples are melting temperatures and the moduli of elasticity. Other properties, however, increase with increasing molar mass (e.g., melt viscosities).

Thermosets are usually generated from fairly low molar mass polymers, called oligomers (science) or prepolymers (industry). High molar masses are unnecessary here because chemical reactions between prepolymer molecules lead to an interconnection of these molecules (cross-linking) and thus to a giant molecule, with 100 % conversion of the prepolymer. The properties of these giant polymer molecules are obviously independent (or nearly so) of the molar mass of the primary molecules. Prepolymers are thus thermosetting materials and become true thermosets only after the hardening reaction.

Plastics are usually divided into four groups: commodity plastics (also called standard plastics or bulk plastics), engineering plastics (sometimes referred to as technical plastics or technoplastics), high-performance plastics, and functional plastics (or specialty plastics). A commodity, engineering, or high-performance plastic may have many different applications, whereas a functional plastic has one very specific application. Polyethylene, a commodity plastic, may according to its type or grade be used for containers, as packaging film, as agricultural mulch, etc. Poly(ethylene – co – vinyl alcohol) with a high content of vinyl alcohol units, on the other hand, is a functional plastic that is used only as a barrier resin. Other functional plastics are employed in optoelectronics, as resists, as piezoelectric materials, etc. Functional plastic is not synonymous with functional polymer or functionalized polymer, because the latter terms refer to polymers with functional chemical groups (i.e., groups with specific chemical reactivities).

Commodity plastics are manufactured in great amounts, hence the terms bulk plastics or standard plastics. Engineering plastics possess improved mechanical properties compared to commodity plastics. Such improved properties may be higher moduli of elasticity, smaller cold flows, higher impact strengths, etc. Engineering plastics are also often defined as those thermoplastics that maintain dimensional stability and most mechanical properties above 100 °C or below 0 °C. High-performance plastics, on the other hand, are engineering plastics with even more improved mechanical properties.

No sharp dividing lines exist between commodity plastics and engineering plastics on the one hand or engineering plastics and high-performance plastics on the other; nor are there generally agreed upon property levels beyond which polymers are designated as engineering or high-performance plastics. The three groups of plastics, in general, include the following:

Commodity Plastics. Poly(vinyl chloride), polyethylenes (high density, low density, very low density), isotactic polypropylene, standard polystyrene.

Engineering Plastics. Poly(ethylene terephthalate), poly(butylene terephthalate), polyamides (aliphatic, amorphous, aromatic), polycarbonates, polyoxymethylenes, poly(methyl methacrylate), some modified polystyrenes such as styrene – acrylonitrile (SAN) and acrylonitrile – butadiene – styrene (ABS) copolymers, and high-impact polystyrenes (SB), as well as various blends such as poly(phenylene oxide) – polystyrene and polycarbonate – ABS.

High-Performance Plastics. Liquid crystalline polymers (LCPs), polyetheretherketone (PEEK), various polysulfones, polyimides, etc.

Some *thermosets* are also classified as engineering or high-performance plastics. In general, thermosets are, however, considered a separate group of plastics. They comprise alkyd (⟶ Alkyd Resins), phenolic (⟶ Phenolic Resins), and amino resins (⟶ Amino Resins) (melamine and urea resins), epoxides (epoxies) (⟶ Epoxy Resins), unsaturated polyesters (including so- called vinyl esters) (⟶ Polyester Resins, Unsaturated), polyurethanes (⟶ Polyurethanes), and allylics (⟶ Allyl Compounds – Allyl Esters).

1.2.2. *Designations [54-58]*

The plastics trade and technical literature indulges in trivial names, abbreviations, and acronyms. Acetal polymer is, for example, the name given to polymers with $– OCH_2 –$ as the major repeating unit. Polymers of methacryl esters ~$[CH_2 – C(CH_3)(COOR)]_n$~ are known as methacrylics (or often only acrylics), but polyacrylonitrile ~$[CH_2 – CH(CN)]_n$~ is not considered an "acrylic" in the plastics literature (the textile literature calls it an acrylic fiber). Styrenics comprise homopolymers of styrene as well as copolymers of styrene with other monomers. Nylon, originally a protected trade name, is now a generic term for polyamides, especially aliphatic ones.

Abbreviations often have more than one meaning (Tables 1, 2, 3, 4): PBT may, for example, denote poly(butylene terephthalate), but it is also used for poly(*p*-phenylenebenzbisthiazole). Exhaustive lists of commonly used abbreviations of polymers and additives as recommended by various organizations have been compiled [58].

Tables 1, 2, 3, and 4 give the recommendations for abbreviations of names of thermoplastics and thermosets by ASTM/ANSI (American Society for the Testing of Materials/American National Standards Institute [D 1600–86 a]), DIN (German Industrial Standards [7728]), ISO (International Standardization Organization [1043 – 1986 (∈)]), IUPAC (International Union of Pure and Applied Chemistry [Pure Appl. Chem. 59 (1987) 691]). "Other" means other abbreviations used in the field but not recommended by ASTM, DIN, ISO, or IUPAC. The recommended abbreviations are often neither identical with those of fibers of the same chemial structures nor with those of elastomers containing the same monomer units; some of these also deviate from those recommended for recycleable plastics. Several abbreviations have different meanings in either two or more of the ASTM, DIN, ISO and IUPAC systems and/or the technical literature, resp.; symbols with more than one meaning are characterized by # in Tables 1-4.

The abbreviations are in part based on the poly(monomer) nomenclature, i.e., on the names of the monomers used in the manufacture of polymers, sometimes, however, without a prefix "P" for "poly" (Table 1). The names of monomers for copolymers are given in alphabetical order without regard to their prevalence.

Table 1. Polymers by chain polymerization (addition polymerization), copolymerization, polycondensation (condensation polymerization), and polyaddition which are characterized by abbreviations based on the poly(monomer) nomenclature

Polymers of	ASTM	DIN	ISO	IUPAC	Other
Acrylester+acrylonitrile+butadiene	ABA	A/B/A			

Acrylester+acrylonitrile+styrene	ASA	A/S/A	A/S/A		
Acrylester+ethylene	EEA#		E/EA		EAA
Acrylic acid	PAA				PAS#
Acrylonitrile	PAN	PAN		PAN	
Acrylonitrile+butadiene	PBAN				
Acrylonitrile+butadiene+methyl methacrylate+styrene					MABS
Acrylonitrile+butadiene+styrene	ABS	ABS	ABS	ABS	ABS
Acrylonitrile+ethylene-propylene-diene + styrene		A/EPDM/S		AES	
Acrylonitrile+methyl methacrylate	AMMA	A/MMA	A/MMA		
Acrylonitrile+styrene	SAN	SAN	SAN	SAN	PSAN
Acrylonitrile+styrene+chlorinated polyethylene		A/PE-C/S			
Adipic acid+hexamethylenediamine	PA 6.6	PA 66	PA 6.6	PA 6.6	
Adipic acid+tetramethylene glycol					PTMA
Allyldiglycol carbonate	ADC				
p-Aminobenzoic acid					PAB
Aminotriazol					PAT
11-Aminoundecanoic acid		PA 11			
Azelaic anhydride					PAPA
Bisphenol A+phosgene	PC#	PC#	PC#	PC#	
Bitumen+ethylene				ECB	
Butadiene+methyl methacrylate + styrene		MBS			
Butadiene+styrene (thermoplastic)	SB	S/B	S/B	S/B	PASB
Butene-1	PB	PB	PB		BT
Butyl acrylate		PBA		PBA	
Butylene glycol+terephthalic acid	PBT	PBT	PBT		PTMT
Butyl methacrylate					PBMA
∈-Caprolactam	PA 6	PA 6	PA 6	PA 6	
Chlorotrifluoroethylene	PCTFE	PCTFE	PCTFE	PCTFE	
Chlorotrifluoroethylene+ethylene					ECTFE
Cresol+formaldehyde	CF	CF	CF	CF	
1,4-Cyclohexanedimethylol+terephthalic acid					PCDT
Diallyl chloroendate (=diallyl ester of 4,5,6,7,7-hexachlorobicyclo-[2,2,1]5-heptane-2,3-dicarbonic acid)	PDAC				
Diallyl fumarate	PDAF				
Diallyl isophthalate	PDAIP				
Diallyl maleate	PDAM				
Diallyl phthalate	PDAP	PDAP	PDAP	PDAP	DAP
1,4-Dichlorobenzene+disodium sulfide	PPS	PPS			
2,6-Dimethylphenol+oxygen	POP				PPO™
Dodecanolactam (laurolactam)	PA 12	PA 12	PA 12	PA 12	
Ethyl acrylate					PEA
Ethyl acrylate+ethylene		E/EA			EEA#
Ethylene	PE	PE	PE	PE	PL
Cross-linked polymer		VPE			XLPE
High-density polymer	HDPE	PE-HD	PE-HD		
Low-density polymer	LDPE	PE-LD	PE-LD		
Low-density polymer, linear	LLDPE				
Medium density polymer	MDPE				
Ultra-high molar mass polymer	UHMW-PE				
Ethylene+methyl acrylate+vinyl chloride		VC/E/MA	VC/E/MA		
Ethylene+methyl methacrylate	EMA	E/MA			
Ethylene+propylene		E/P	E/P		PEP
Ethylene+propylene (+diene)		EPDM			
Ethylene+tetrafluoroethylene	ETFE	E/TFE			
Ethylene+vinyl acetate	EVA	E/VA	EVA		VAE
		E/VAC	E/VAC		
Ethylene+vinyl acetate+vinyl chloride		VC/E/VAC	VC/E/VAC		
Ethylene+vinyl chloride		VC/E	VC/E		
Ethylene glycol					PEG
Ethylene glycol+maleic anhydride	UP#				
Ethylene glycol+terephthalic acid (ester)	PET	PET	PETP	PETP	PETE
Fast crystallizing polymer					CPET
Oriented polymer					OPET
Ethylene oxide	PEO	PEOX	PEOX	PEO	
Formaldehyde (or trioxane)	POM#	POM#	POM#	POM#	
Formaldehyde+furan	FF				
Formaldehyde+melamine	MF	MF	MF	MF	
Formaldehyde+melamine+phenol		MPF	MPF		
Formaldehyde+phenol	PF	PF	PF	PF	
Formaldehyde+urea	UF	UF	UF	UF	
Furfural+Phenol	PFF				
Hexafluoropropylene+tetrafluoroethylene	FEP	FEP	FEP		
Hexamethylenediamine+sebacic acid		PA 610			PA 6.10
p-Hydroxybenzoic acid	POB				PHB#
3-Hydroxybutyric acid					PHB#

2-Hydroxyethylmethacrylate					PHEMA
Isobutylene	PIB	PIB	PIB	PIB	IM
Laurolactam (dodecanolactam)	PA 12	PA 12	PA 12	PA 12	
Linseed oil, epoxidized		ELO	ELO		
Maleic anhydride+ styrene	SMA	S/MA			
Methacrylimide		PMI			
Methyl acrylate					PMA
Methyl acrylate+vinyl chloride		VC/MA	VC/MA		
Methyl α-chloromethacrylate	PMCA				
Methyl methacrylate	PMMA	PMMA	PMMA	PMMA	
Methyl methacrylate+vinyl chloride			VC/MMA		
1,4-Methylpentene	PMP	PMP	PMP		TPX
α-Methylstyrene		PMS			PAMS
α-Methylstyrene+styrene	SMS	S/MS	S/MS		
Octyl acrylate+vinyl chloride		VC/OA	VC/OA		
Perfluoroalkoxyalkane	PFA	PFA			
Phosphoric acid					PPA#
Propylene	PP	PP	PP	PP	
Oriented polymer					OPP
Oriented biaxially					BOPP
Propylene+tetrafluoroethylene (alt.)					TFE-P
Propylene oxide	PPOX	PPOX	PPOX		
Soybean oil, epoxidized		ESO	ESO		
Styrene	PS	PS	PS	PS	
Expanded (foamed) polymer					EPS
High-impact polymer					HIPS
Impact resistant polymer	SRP	IPS			
Oriented polymer					OPS
Tetrafluoroethylene	PTFE	PTFE	PTFE	PTFE	
Tetrahydrofuran					PTHF
Polymer with hydroxy endgroups					PTMEG
Triallyl cyanurate	PTAC				
Trifluoroethylene				P 3 FE	
Trioxane (+comonomers)	POM#	POM#	POM#	POM#	
Vinyl acetate	PVAC	PVAC	PVAC	PVAC	PVA
Vinyl acetate+vinyl chloride	PVCA	VC/VAC	VC/VAC		
N-Vinylcarbazole	PVK	PVK	PVK		
Vinyl chloride	PVC	PVC	PVC	PVC	PCU, V
Polymerized in bulk					M-PVC
Polymerized in emulsion					E-PVC
Polymerized in suspension					S-PVC
Polymer as flexible film					FPVC
Polymer as oriented film					OPVC
Polymer as rigid film					RPVC
Vinyl chloride and vinylidene chloride		VC/VDC	VC/VDC		
Vinyl fluoride	PVF	PVF	PVF	PVF	
Vinylidene chloride	PVDC	PVDC	PVDC	PVDC	
Vinylidene fluoride	PVDF	PVDF	PVDF	PVDF	
Vinyl methyl ether					PVME
N-Vinylpyrrolidone	PVP	PVP	PVP		

Other abbreviations are based on characteristic groups of polymers (Table 2) or indicate polymers synthesized by chemical transformation of base polymers (Table 4). Special abbreviations apply to blends, reinforced polymers, etc. (Table 4).

Table 2. Abbreviations for polymers named after a characteristic polymer group

Characteristic group	ASTM	DIN	ISO	IUPAC	Other
Amide	PA	PA	PA	PA	
Amide, aromatic	PARA				
Metal coated films	PA**				
Saran coated films	PA*				
Amide – imide	PAI	PAI			
Aryl amide	PARA				
Aryl sulfone					PAS#
Benzimidazole					PBI
Carbodiimide					PCD
Carbonate, aromatic	PC#	PC#	PC#	PC#	
Epoxide (epoxy)	EP	EP	EP	EP	
Glass-fiber reinforced					GEP
Epoxide – ester			EPE		
Ester, saturated		SP			
Ester, thermoplastic	TPES				
Metallized polymer film					MPE
Ester, unsaturated	UP#	UP#	UP#	UP#	

Glass fiber reinforced					FRP
Ester – alkyd	PAK				
Ester – ether (thermoplastic elastomer)	TEEE				
Ester – imide	PEI				
Ether – block-amide	PEBA	PEBA			
Etheretherketone	PEEK				
Ether – imide		PEI			
Ether sulfone	PES	PES			
	PESU				
Imide	PI	PI			
Isocyanurate		PIR			
Parabanic acid					PPA#
Phenylene ether	PPE	PPE			
Phenylene sulfone	PPSU	PPSU	PPSU		PSU
Silicone	SI	SI	SI		
Sulfone	PSUL	PSU			
Urethane	PUR	PUR	PUR	PUR	
Thermoplastic	TPUR				
Thermoset	TSUR				

Table 3. Abbreviations for polymers produced by chemical transformation of other polymers

Resulting polymer	ASTM	DIN	ISO	IUPAC	Other
Carboxymethyl cellulose	CMC	CMC	CMC	CMC	
Carboxymethyl hydroxyethyl cellulose					CMHEC
Casein, cross-linked with formaldehyde	CS	CS	CS	CS	
Cellulose, as Cellophan	C				
Ditto, coated with Saran	C*				
Cellulose acetate	CA	CA	CA	CA	
Cellulose acetobutyrate	CAB	CAB	CAB	CAB	
Cellulose acetopropionate	CAP	CAP	CAP	CAP	
Cellulose nitrate	CN	CN	CN	CN	NC
Cellulose plastics, unspecified	CE				
Cellulose propionate	CP	CP	CP	CP	
Cellulose triacetate	CTA	CTA			TA
Ethyl cellulose	EC	EC	EC	EC	
Hydroxyethyl cellulose					HEC
Hydroxypropyl cellulose					HPC
Hydroxypropyl methyl cellulose					HPMC
Methyl cellulose		MC			
Polyethylene, chlorinated	CPE	PE-C	PEC		
Chlorosulfonated	CSM				CSR
Poly(ethylene-co-vinyl alcohol)		E/VAL			EVAL
					EVOH
Polypropylene, chlorinated		PPC			
Polyvinyl alcohol	PVAL	PVAL	PVAL	PVAL	PVOH
Polyvinyl butyral	PVB	PVB	PVB		
Polyvinyl chloride, chlorinated	CPVC	PVC-C	PVCC		PC#, PeCe
Polyvinyl formal	PVFM	PVFM	PVFM	PVFM	

Table 4. Abbreviations for blends, reinforced polymers, etc.

Resulting polymer	ASTM	DIN	ISO	IUPAC	Other
Elastomers, thermoplastic					
Containing ester and ether groups	TEEE				
Olefin-based	TEO				
Styrene-based	TES				
Epoxide, glass-fiber reinforced					GEP
Plastic, carbon-fiber reinforced				KEK	CFP
Glass-fiber reinforced					GFK
Man-made fiber reinforced					CFK
Metal fiber reinforced					MFK
Poly(acrylonitrile-co-styrene)+chlorinated polyethylene					ACS
Sheet molding compound					SMC
with high glass fiber content					HMC

Special properties of polymers are symbolized by up to four additional letters. These letters are arranged after the symbol for the base polymer according to ISO and DIN; in the technical literature, they are, however, commonly placed in front of the base symbol, following ASTM. DIN recommends the following letters for special properties:
C = chlorinated; D = density; E = expanded, expandable; F = flexible, liquid; H = high; I = impact resistant; L = linear, low; M = molar mass, average (mean), molecular; N = normal, novolac; P = plasticized; R = raised (enhanced), resol; U = ultra, plasticizer-free (unplasticized); V = very; W = weight (mass); X = cross-linked. The technical literature also uses BO = biaxially oriented and O = oriented (usually in one direction). In the ASTM scheme, an asterisk (*) denotes a saran coated polymer film, two asterisks (**) a metallized polymer film.

Companies have long used internal classification systems, and the military has issued plastics specifications. An industry-wide classification system for thermoplastic materials

introduced by ISO is based on an alphanumeric “line call-out,” a special code (see below) [54], [55]. A material is characterized by several data blocks that indicate its composition and certain property data or ranges. These property data refer to essential criteria. Since different thermoplastics have different applications, property data are restricted to different “leading properties” for each type of thermoplastic.

The property data covered by this classification scheme are thus not comprehensive. Rather, various plastics types are characterized by one to three sets of leading criteria, selected from

1. Chemical structure data such as content of vinyl acetate (VAC) or acrylonitrile units (AN), isotacticity (IT), or density (D) as a measure of branching
2. Molar mass data such as intrinsic viscosity (IV) or Fikentscher's K value (FK)
3. Bulk density (BD)
4. Rheological data such as melt flow rate (MFR) (melt flow index)
5. Thermal data such as Vicat temperature (VT) or torsional stiffness temperature (TST)
6. Mechanical properties such as modulus of elasticity (*E*), tensile stress at 100 % strain (TS), Shore hardness (SH), or impact strength (notched) (ISN)

The following leading criteria are used:

Polyethylene	D, MFR
Polypropylene	I, MFR
Polystyrene and acrylonitrile – styrene copolymers	VT, MFR
Styrene – butadiene, acrylonitrile – butadiene – styrene, and acrylonitrile – styrene – acrylic acid copolymers	VT, MFR, ISN
Ethylene – vinyl acetate copolymers	VAC, MFR
Poly(vinyl chloride)	IV, FK, BD
Poly(vinyl chloride), unplasticized	VT, E, ISN
Poly(vinyl chloride), plasticized	TS, SH, TST
Polyamides	IV, E
Polycarbonates	IV, MFR, E
Poly(methyl methacrylate)	IV, VT
Poly(ethylene terephthalate)	IV

The standard designation of a thermoplastic material consists of a description block, a standard number block, and a series of data blocks.

Designation Block. The designation block gives the type of material (e.g., thermoplastic, molding, etc.).

Standard Number Block. The standard number block consists of the number of the ISO (or DIN or other national) standard, followed by a hyphen and *data block 1*. This data block contains the abbreviation of the chemical name of the material [e.g., PE = polyethylene, PVC = poly(vinyl chloride)]. This may be followed by analytical data such as the vinyl acetate content of ethylene – vinyl acetate copolymers (EVAC). However, these are not the exact analytical data but rather code numbers for the range (called “cell”) in which this material can substitute for a similar one. Separated by a hyphen, supplemental information on this material may be given (e.g., H = homopolymer, P = plasticized, U = unplasticized, E = polymerized in emulsion).

Data block 2 may contain up to four letters that give qualitative information. The first letter denotes the intended application (e.g., B = blow molding, G = general purpose, P = paste resin, Y = textile yarn). The following one to three letters can code up to three essential additives or supplemental information, for example, A = antioxidant, D = powder (dry blend), L = light and weather stabilizer, S = slip agent.

Data Block 3. Quantitative information about the designated properties is contained in data block 3. The encoding of these data is different for each plastic material and each testing standard. For example, the code 20 – D 050 in data block 3 for a polyethylene tested according to DIN 16 776 means a material with a density of 0.918 g/cm^3 (cell 020) and a melt flow rate of 4.2 g/min (cell 050) measured at 190 °C under a load of 2.16 kg (D). On the other hand, a third data block 22 – 030 for a polyamide 12 means a polymer with an intrinsic viscosity of 210 cm^3/g (cell 22) and a modulus of elasticity of 280 MPa (cell 030). The definition of the cells can be found in special cell tables.

Data block 4 gives information about the type and content of fillers or reinforcing materials; these codes are independent of the types of polymers. The first letter gives the type of filler, for example, C = carbon or G = glass. The second letter indicates the shape of the filler (e.g., F = fiber, S = sphere). The letters are followed by a code for the mass content of the filler; 15 for the range 12.5 – 17.5 wt %, 60 for the range 57.5 – 62.5 wt %.

Data block 5 is reserved for specifications based on individual agreements between supplier and customer. It may code additional requirements, restrictions, or supplemental information.

A thermoplastic may thus carry the designation “Molding material DIN 7744-PC, XF, 55 – 045, GF 30,” which indicates a polycarbonate (PC) with no special indication of its application (X) but with special burning characteristics (F), an intrinsic viscosity in a coded range of 55 (here for [ϒ] = 56 cm^3/g), and a melt flow rate of 5.5 g/(10 min) (coded as 045), which contains 30 % (30) of glass (G) fibers (F).

A designation system similar to the above ISO – DIN system has been recommended by ASTM [56], [57]. A typical ASTM designation, for example, would read “Molding material ASTM D-4000 PI 000 G 42 360”; this material is a polyimide (PI). Since no individual property table has been specified for polyimides, the cell table G of ASTM D-4000 has been used to characterize the properties of PI. This particular cell table identifies five different properties by designating cell limits to the five digits following the letter G. The first digit characterizes the minimum tensile strength, here “4”, which according to cell table G means at least 85 MPa. The second digit gives the cell value for the flexural modulus (here “2” for 3500 MPa). The third digit denotes the Izod impact strength (50 J/m ⟶“3”), and the fourth digit the heat deflection temperature (300 °C ⟶“6”). The fifth digit is undetermined; the “0” indicates an unspecified property.

1.3. History of Plastics [59-64]

The first plastics were prepared long before their macromolecular nature was discovered. Decorative coatings based on polymers such as egg white or blood proteins were used in the cave paintings of Altamira, Spain, as early as 15 000 B.C. Later painting methods utilized gelatin and the polymers resulting from naturally drying vegetable oils. The use of these materials resulted from the eternal desire of artists for easy-to-handle materials that give an optimal artistic effect and have infinite stability.

These requirements were not restricted to surface coatings, which are two dimensionally applied plastics. Cow horns were used in the Middle Ages to prepare windows for lanterns and intarsia in wood. To this purpose, the horns had to be flattened by steaming, a very difficult process. The flattened horns also tended to curl up after a while. An early, easy-to-process substitute for natural horn was reported by the Bavarian monk WOLFGANG SEIDEL (1492 – 1562); it undoubtedly derived from far older recipes. This imitation horn was based on the protein casein, the white material from skim milk. Casein was extracted from skim milk with hot water, treated with warm lye, and shaped while being warmed; the desired shape was then fixed by immersion in cold water. Unknown inventors later discovered that the addition of inorganic fillers increased the mechanical stability of this early thermoplastic. The same material was used later for children's building blocks by the German aviation pioneer OTTO LILIENTHAL (1848 – 1896) and his brother GUSTAV. In 1885 a patent based on the same physical process was granted to the American EMERY EDWARD CHILDS. The properties of these plastics were improved markedly by the chemical reaction of casein with formaldehyde, as described in 1897 by the German inventors WILHELM KRISCHE and ADOLF SPITTELER. The resulting thermoset was called galalith (Greek: *gala* = milk, *lithos* = stone); it is still used today for haberdashery.

Another early thermoset used natural rubber as the raw material. In 1839 the American CHARLES GOODYEAR discovered the cross-linking (vulcanization) of natural rubber to an elastomer by sulfur under the action of “white lead” (basic lead carbonate) and heat. His brother NELSON used larger amounts of sulfur and in 1851 invented ebonite, a hard, black thermoset.

The first fully synthetic thermoset was invented in 1906 by the Belgian-born American chemist LEO H. BAEKELAND, who heated various phenols with formaldehyde under pressure and produced insoluble hard masses. These “bakelites” were recognized in 1909 as excellent electrical insulators and thus became one of the foundations of the modern electrical industry. BAEKELAND was not the first to prepare phenolic resins though; they were first observed by ADOLF VON BAEYER in 1872. BAEYER's substances were merely resinous materials, and BAEKELAND's “heat and pressure” process was necessary before commercially useful materials could be produced.

The first *semisynthetic thermoplastics* originated from cotton. Cotton fibers consist of cellulose; they have been used by man since prehistoric times. Since cotton is relatively easily grown, many attempts have been made to improve its textile properties. The Englishman JOHN MERCER discovered in 1844 that the treatment of cotton with aqueous solutions of caustic soda (sodium hydroxide) leads to fibers with increased strength, higher luster, and improved dyeability. The Frenchman L. FIGUIER demonstrated in 1846 that cellulose paper was strengthened in a manner similar to cotton when it was treated with sulfuric acid. In 1853 W. E. GAINE received an English patent for the same process, which delivered a parchment-like material. An English patent by THOMAS TAYLOR described the formation of very resistant materials from layers of paper sheets by the combined action of zinc chloride and pressure. These products were called “vulcanfiber” since they resembled the vulcanization products of natural rubber.

Mercerized cotton, artificial parchment, and vulcanized paper result from physical transformations of cellulose. If, however, cellulose materials are treated not by sulfuric acid alone but by a mixture of sulfuric and nitric acids, a chemical reaction to cellulose nitrate (nitrocellulose, gun cotton) occurs. This chance discovery by the German CHRISTIAN FRIEDRICH SCHÖNBEIN in 1846 paved the way for the invention of the first semisynthetic thermoplastic. In 1862 ALEXANDER PARKES found an easy method for processing cellulose nitrate into thermoplastic masses by the addition of castor oil, camphor, and dyes. The manufacture of the resulting "parkesine" was difficult, however, and production ceased in 1867. The use of alcoholic solutions of camphor by DANIEL W. SPILL was equally unsuccesful. In 1869 a patent was granted to the American JOHN WESLEY HYATT for the use of camphor without camphor oil or alcohol. The resulting "celluloid" is generally considered the pioneering thermoplastic [63].

Other early, fully synthetic thermoplastics have a far longer history, albeit not as industrial materials. Polymers of formaldehyde were found by JUSTUS VON LIEBIG in 1839; yet extensive scientific investigations by HERMANN STAUDINGER in the 1920s and 1930s and major industrial development work by Du Pont were necessary before it became an engineering plastic in 1956.

Vinyl chloride was synthesized in 1838 by H. V. REGNAULD, who also observed the formation of resins. Vinyl chloride was polymerized in 1912 by KLATTE in Germany and OSTROMUISLENSKY in England, but not until 1931 was the first poly(vinyl chloride) produced commercially in Germany by I. G. Farben.

Styrene was discovered by a chemist called NEUMAN as reported in W. NICHOLSON's 1786 *Dictionary of Practical and Theoretical Chemistry*. The conversion of styrene to a solid mass was discovered by E. SIMON in 1839, who considered it a styrene oxide. A. W. VON HOFMANN and J. BLYTH showed in 1845 that the alleged styrene oxide was an isomer of styrene; they called it metastyrene. It was however only in 1920 that the polymeric nature of polystyrene was recognized by H. STAUDINGER; it was first produced commercially by I. G. Farben (Germany) in 1930.

A highly branched polyethylene is found in nature as the "mineral" elaterite. Lightly branched synthetic polyethylene was first obtained in 1933 by high-pressure polymerization of ethylene at ICI (England); its commercial production began in 1939. Linear polyethylenes were first synthesized in 1953 by low-pressure polymerization using catalysts based on transition metals (Ziegler), chromium oxide (Phillips Petroleum), and molybdenum oxide (Standard Oil of Indiana). For the first time, Ziegler – Natta catalysts also enabled the polymerization of α-olefins such as propylene to solid, stereoregular polymers; the commercial production of isotactic polypropylene (PP) began in Italy, the Federal Republic of Germany, and the United States in 1957. The use of these catalysts and the availability of inexpensive feedstocks from petroleum refining led to rapid growth of the polymer industry (Table 5).

Table 5. World production of polymers (in 10^6 t/a)

Type	1940	1950	1960	1970	1980	1990
Plastics	0.36	1.62	6.7	31.0	59.0	100
Fibers						
Synthetic	0.005	0.069	0.70	5.0	11.5	15.7
Semisynthetic	1.1	1.6	2.6	3.4	3.3	3.2
Natural **	8.7	8.0	12.8	14.0	17.7	18 *
Elastomers						
Synthetic	0.043	0.54	1.94	5.9	8.7	9.1 *
Natural	1.44	1.9	2.02	3.1	3.9	4.9 *

* Estimated.
** Textile fibers only.

Plastics are the leader in this growth, followed by synthetic fibers and synthetic elastomers. The world production of semisynthetic fibers (rayon and cellulose acetate) (⟶ Cellulose Esters) has, however, declined, mainly because of cost and environmental concerns in the developed countries and despite increased production of semisynthetic fibers in the developing countries.

1.4. Economic Importance [65-69]

The rapid growth of plastics production is a result of three factors: growth of world population, average increase in standard of living, and replacement of older materials by plastics. The growth pattern may be economy driven [determined by the price of materials (substitutes); shortages ("ersatz"), or government intervention (environmental concerns)], or technology driven (determined by technological requirements, internal knowhow, accidental discoveries).

The world population grew from 2.532×10^9 in 1950 to 5.32×10^9 in 1980, while the world plastics production (approximately equal to consumption) climbed from 0.6 kg per capita (1950) to ca. 18.8 kg per capita (1990) (Table 6).

Table 6. Production of plastics, world population, and consumption per capita

Region/country	Production in 10^6 t/a				
	1950	1960	1970	1980	1990
European Community	0.3	2.3	11.5	19.7	30
United States	1.1	2.8	9.1	16.1	30
Japan	0.04	0.7	5.3	7.5	13
Eastern Europe	0.1	0.5	4.0	6.9	13
Other	0.07	0.3	1.1	8.8	14
Total	1.6	6.7	31.0	59.0	100
World population, $\times10^{-6}$	2532	3062	3730	4498	5320
Per capita consumption in kg	0.6	2.2	8.3	13.1	18.8

These numbers are approximate because statistics are incomplete and may differ considerably from source to source. For example, 1987 world plastics production is given as either 37.5×10^6 t ([65], excluding South Africa and the People's Republic of China) or 55.4×10^6 t [66].

Great differences in plastics consumption exist between various countries. In 1988, the apparent annual per capita consumption as calculated from production plus import minus export (i.e., including stockpiling) was 128 kg in the Federal Republic of Germany; 104 kg in the United States of America, 95 kg in Switzerland; and 81 kg in Japan and the United Kingdom but in 1978 only 1.2 kg in Black Africa, 1.1 kg in Bolivia, and 0.4 kg in India [67] and has not much changed since. Since world population continues to grow and people in less developed countries justifiably aspire to higher living standards, the world plastics production is most likely to increase further despite some environmental concerns in highly industrialized countries (see below). Although all regions will share in this growth, it will be most pronounced in "other" countries (e.g., Saudi Arabia, countries in Central and South America, Southeast Asia, and the Indian subcontinent) and, if planned economies are replaced by free-market types, also in the former Comecon countries.

The replacement of other materials by plastics is more difficult to judge. Since developing countries possess a great and unsatisfied need for goods that have long been taken for granted in more affluent societies, an analysis cannot be based on world production of materials but only on that in highly developed countries such as the United States (Table 7).

Table 7. Per capita consumption of selected materials in the United States in 1983 (estimated population: 234×10^6) [68]

Materials	Source Consumption in kg/a Imports in % of consumption

Nonmetallic minerals			
Sand, gravel	N [a]	2576	0
Stone	N [a]	3322	0
Cement	MN [b]	287	6
Clay	MN [b]	270	0
Gypsum	N [a]	81	39
Metals			
Raw steel	S [c]	421	15
Aluminum, primary	S [c]	15.0	6
Copper	S [c]	6.4	31
Lead	S [c]	4.8	10
Zinc	S [c]	3.8	39
Polymers			
Wood, wood products			
Lumber	N [a]	466	22
Plywood	MN [b]	29.6	4
Paper	MN [b]	277	6
Other	?	35	?
Fibers (excluding export – import of fabrics and garments)			
Cotton	N [a]	5.5	(50) [e]
Rayon, acetate	MN [b]	3.6	66
Wool, silk	N [a]	1.6	92
Synthetic fibers	S [c]	16.9	?
Elastomers (excluding exported rubber goods)			
Natural rubber	MN [b]	2.6	100
Synthetic rubber	S [c]	6.7	(12) [e]
Plastics (excluding manufactured goods)			
Plastics, resins	S [c]	71	(9) [e]
	MN [b]	0.2 [d]	?
Leather	MN [b]	6.4	?

[a] Natural product.
[b] MN = physically or chemically modified natural product.
[c] S = synthetic product.
[d] Estimated.
[e] Net exports.

Table 7 gives the apparent consumption of selected important materials, calculated from production plus import minus export i.e., including stockpiling but excluding the import – export of manufactured goods such as machinery (metals), garments (fibers), tires (cars), newsprint (paper), etc. The highest per capita consumption is in construction materials (nonmetallic minerals, lumber, partly steel and aluminum), followed by packaging materials (partly steel, aluminum, paper).

Plastics constitute a fairly small percentage of all consumed materials (Table 7). However, they have the highest growth rate (Fig. 1). The data in Figure 1 are given on a volume basis because materials are used by volume although they are sold by mass. Although the share of plastics in terms of total materials produced (and mainly disposed) is quite low, it causes environmental concern because of its very visible proportion in household refuse (see Chap. Waste Disposal).

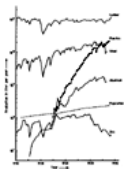

Figure 1. Production of important materials (in liters per year) in the United States during 1910 – 1990.

All metals are on smelter basis; - - - U.S. population (numbers). Reported commercial data were converted as follows: 1 board foot lumber = 2.36 L, 1 short ton ≈ 907 kg, 1 lb ≈ 0.454 kg; densities of steel (7.85 g/cm^3), aluminum (2.70 g/cm^3), zinc (7.13 g/cm^3), and plastics (1.1 g/cm^3) (average, weighted according to market shares of major plastics). Data from [69].

Since construction and packaging are major applications for lumber, steel, and aluminum, it can be expected that plastics are important replacements for these materials. Packaging and building – construction are indeed the most important applications of plastics (Table 8); they also show the highest growth rates (ca. 15 % per year).

Table 8. U.S. plastics sales and applications in 1988 (in 1000 t)

Polymer	Domestic sales	Domestic applications [a]								
		A	B	E	F	H	P	T	Y	U
PE-HD	3 743	9	251	53	5	143	1 692	19	77	1 494
PE-LD	4 479		127	168	4	191	1 092	20	74	2 803
PP	3 316	67	18	21	30	92	470	170	18	2 430
PS	2 329	77	127	186	55	115	108		108	1 553
SAN	62	10		2		10	8			32
ABS	562	91	69	53	2		2	95	17	233
Other styrenics	553									553
PPO alloys	82	20	2	1						59
PVC	3 779	64	2 301	232	41	29	199	101	15	809
Other vinyls [b]	435		2				53			380
PMMA	316	4	110					25		177
POM	58	5	7	2				12		32
PA	253	8		36			25	69		115

PC	195	16	52	25				28		74
PETP, PBT	911	3	3	20			450	20		415
Fluoropolymers	13									13
Cellulosics	41	1	2	2			3		1	32
Thermoplastic elastomers	225									225
UP [c]	623	30	132	17	9			64		371
EP	213	4	19	25			15	45		105
PF [d]	1 376	15	411	45	77	18	6	10		794
UF, MF [d]	688		121	14	10	6	3			534
PUR [e]	1 319	73	309		257				331	349
Alkyds	145									145
Other [f]	500	16		21	3	29	86	52	41	252
Total	26 216	513	4 063	923	493	633	4 212	730	682	13 979
% of accounted		4.2	33.2	7.6	4.0	5.1	34.4	6.0	5.5	
% of U.S. total	100	2.0	15.5	3.5	1.9	2.4	16.0	2.8	2.6	53.3
% of FRG [g]	100		21.0	15.0	5.0	2.5	25.0	7.0		10.5

[a] A=appliances, B=building/construction, E=electrical/electronic, F=furniture, H=household goods, P=packaging, T=transportation, Y=toys, U=unaccounted; 1982 FRG percentages (bottom line) are given for comparison.
[b] Poly(vinyl acetate), poly(vinyl butyral), ethylene – vinyl alcohol copolymers, etc.
[c] Without glass fibers.
[d] Resin in plywood assumed as 20 %.
[e] Raw materials.
[f] Coumarones, high-performance resins, etc.
[g] Plus paints etc. (10 %) and agricultural (4.0 %).

Plastics are consumed in far lesser amounts in electric – electronics applications, transportation, toys, housewares, appliances, and furniture; the U.S. use in agriculture and paints and coatings has not been reported. Plastics are critical and irreplaceable components in certain fields such as the electric – electronics industry. Consumption patterns are similar for nations with comparable affluence; the United States and West German applications of plastics do not differ significantly when the fact is considered that ca. 53 % of U.S. applications are not accounted for.

[Top of Page]

2. Molecular Structure of Polymers [1-5, 8-14]

2.1. Constitution

2.1.1. *Homopolymers*

A homopolymer is defined as a polymer derived from one type of monomer. The term homopolymer thus refers to the origin of the monomers in a polymer and not to the actual constitutional units. The polymerization of a monomer CH_2=CHR leads to constitutional units – CH_2CHR –. These units may be connected to each other not only in head-to-tail positions but also in head-to-head and tail-to-tail arrangements:

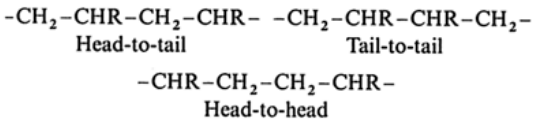

The proportion of head-to-head regioisomeric units increases with decreasing size of substituents R and with reduced resonance stabilization of the growing species. Free-radical polymerization of vinyl acetate [CH_2=CH(OOCCH_3)] leads to 1 – 2 %, of vinyl fluoride (CH_2=CHF) to 6 –10 %, and of vinylidene fluoride (CH_2=CF_2) to 10 – 12 % tail-to-tail connections.

Monomer molecules can also polymerize to a lesser or greater extent via “wrong” groups. The amount of such isomeric units depends on the constitution of the monomer and the mode of polymerization. Acrolein (**1**), for example, polymerizes free radically via the carbon – carbon double bond (**2**) and the aldehyde group (**3**); it forms branched (**4**) or cross-linked units and even intramolecular rings (**5**) from two units of (**4**):

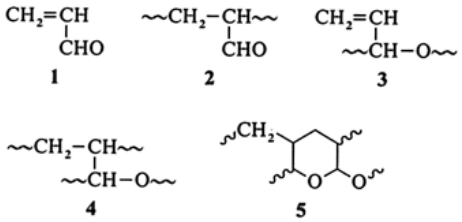

Monomers may also isomerize during polymerization, especially cationic polymerization. 4,4-Dimethyl-pentene-1 (**6**) polymerizes at low temperature (≈–130 °C) via the carbon – carbon double bond (**7**), but at higher temperature (≈0 °C) leads to a “phantom polymer” (**8**) with three carbon atoms per monomeric unit:

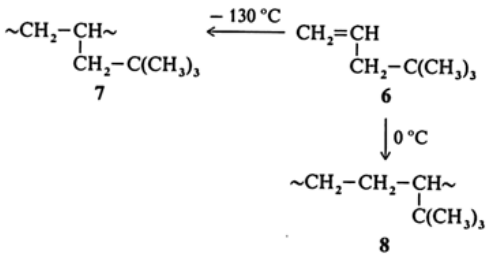

Irregular structures may also be created by chain-transfer reactions to polymer molecules (see Chap. Polymer Manufacture); these reactions lead to branched polymers.

The presence of such isomeric units can often be neither proved nor disproved. Polymer structures are therefore generally depicted by their most common constitutional repeating units; they are idealized structures.

The groups at the ends of polymer chains are likewise not shown, in part because they are often unknown and in part because their structure does not influence most of the polymer properties. End groups may be initiator or catalyst residues or groups that are generated by transfer reactions. These groups rarely affect mechanical properties; they may, however, negatively influence the thermal or photochemical stability of polymers.

2.1.2. *Copolymers*

Copolymers are generated from more than one type of monomer; they are called bipolymers, terpolymers, quaterpolymers, etc., according to the number of monomer types. The copolymerization of ethylene ($CH_2{=}CH_2$) and propylene [$CH_2{=}CH(CH_3)$] thus leads to the bipolymer poly(ethylene- *co*-propylene). Polymers of butadiene ($CH_2{=}CH{-}CH{=}CH_2$) with 1,4-units $-CH_2-CH{=}CH-CH_2-$ and 1,2-units $-CH_2-CH(CH{=}CH_2)-$ are not called copolymers because they are generated from only one monomer type. Rather, they are pseudocopolymers such as the ones that result from partial saponification of poly(vinyl acetate), $\sim[CH_2{-}CH(OOCCH_3)]_n\sim$, to a pseudocopolymer with vinyl acetate units and vinyl alcohol units $-CH_2-CH(OH)-$. The term copolymer thus does not refer to the chemical structure of the resulting polymer, but to the monomers from which it derives.

The succession of monomeric units in copolymer chains is known as their sequence. Monomer units "a" and "b" alternate in *alternating copolymers* (Table 9), which are limiting cases of periodic copolymers $\sim(abb)_n\sim$, $\sim(aabb)_n\sim$, $\sim(abc)_n\sim$, etc.

Table 9. Types of copolymers from monomers A, B, and C (monomeric units are characterized by lower-case letters)

Name	Structure	Shorthand name
Copolymer without specified sequence	~ (a/b) ~	poly(A-co-B)
Statistical copolymer (example)	~ a-b-b-b-a-a-b-a-a-a-a-b ~	poly(A-stat-B)
Random copolymer (with Bernoulli statistics of sequence)	ditto	poly(A-ran-B)
Alternating copolymer	~ a-b-a-b-a-b-a-b-a-b-a-b ~	poly(A-alt-B)
Periodic copolymer	~ a-b-b-a-b-b-a-b-b-a-b-b ~	poly(A-per-B-per-B)
	~ a-b-c-a-b-c-a-b-c-a-b-c ~	poly(A-per-B-per-C)
Diblock copolymer	~ a.....a-b.....b ~	poly(A)-block-poly(B)
Triblock copolymer	~ a.....a-b.....b-c.....c ~	poly(A)-block-poly(B)-block-poly(C)
Graft copolymer	~ a-a-a.....a-a-a.....a ~ (b$_m$, b$_n$)	poly(A)-graft-poly(B)

The sequence of monomeric units in *statistical copolymers* is determined by the statistics of copolymerization (Markov statistics of zeroth, first, etc., order). *Random copolymers* are special cases of statistical copolymers: the sequence follows Bernoulli statistics (i.e., zeroth-order Markov statistics).

Graded copolymers (i.e., tapered copolymers) exhibit a compositional gradient along the chain: one chain end is enriched in "a" units, the other in "b" units. *Block copolymers* are extreme cases of such graded copolymers; they consist of blocks of homosequences that are joined via their ends. Linear multiblock copolymers with short blocks are called *segment(ed) copolymers*. In *graft copolymers*, "b" blocks are connected to "a" chains via center monomeric units.

Copolymers poly(A- co-B) of monomers A and B possess properties quite different from blends of homopolymers A and B.

2.1.3. *Branched Polymers*

Open chains of the type $R-m_n-R$ exhibit the simplest structures. They posses *n* monomeric "m" units and two R end groups. They are also called linear chains (Fig. 2, L) because of their one-dimensional connectivity or their unbranched structures and because they were originally (and wrongly) assumed to be completely extended (stretched out). Spiropolymers and ladder polymers (see below) are also linear polymers because they are not "branched" in the macromolecular sense.

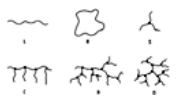

Figure 2. Schematic representation of linear and branched marcomolecules

L = Linear (unbranched) chain in almost extended state; R = ring molecule; S = starlike molecule with three subchains, C = comblike molecule with four branching points of functionality 3; B = randomly branched chain with subsequent branching; D = dendrimer with three-functional branch points and three generations of subchains

Cyclic polymers (ring polymers; Fig. 2, R) consist of linear polymer molecules that are joined via their ends. They do not have end groups and are not called macrocycles because that term is traditionally reserved for "big rings" of ca. 15 – 20 chain atoms.

Branched polymers contain branch points that connect three or more subchains (sometimes called subunits). The following types of branched polymers are distinguished:

1. *Star polymers* possess one branching point from which three or more subchains radiate (Fig. 2, S).
2. *Dendrimers* (dendritic polymers) are star polymers in which the subchains are themselves starlike branched (Fig. 2, D). They are also called cascade polymers (because of the cascade-like sequence of branching points), isotropically branched polymers, or isobranched polymers (since each generation of newly added monomer molecules is "isotropically" arranged around the central core), or much more poetically, starburst polymers.
3. *Comb polymers* have "long" side chains, whose chemical structures usually differ from those of the main chains (Fig. 2, C). The term *graft copolymer* refers to comb polymers into which branches have been introduced by subsequent grafting onto a primary chain. Comb polymers whose side chains exhibit liquid crystalline (LC) behavior are called LC side- chain polymers.
4. *Randomly branched polymers* have a random distribution of branching points (Fig. 2, B); the subchains may be further branched (Christmas tree branching) or not. Short- chain and long- chain branching are distinguished. *Short- chain branches* are often generated by intramolecular transfer reactions, for example, the butyl side chain of polyethylene is formed during free-radical polymerization of ethylene:

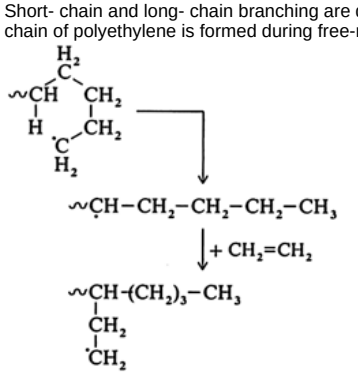

Long- chain branches are usually generated by intermolecular transfer reactions. In the polymerization of vinyl acetate, side- chain radicals formed by transfer reactions may initiate the polymerization of additional vinyl acetate monomers.

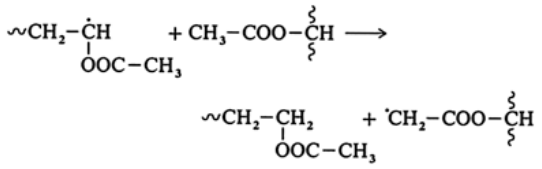

Such side- chain radicals are also formed by radical transfer from certain initiator radicals.

Side chains introduced by the inherent structure of a (co)monomer are not called branches in polymer science. Neither linear poly(vinyl acetate) nor the low- crystallinity, low-density polymers formed by copolymerization of ethylene with small amounts of α-olefins (butene-1, hexene-1, octene-1) are considered branched; the latter are known as linear low-density polyethylenes.

2.1.4. *Ordered Chain Assemblies*

Two or more chains may also be joined at more than two points by chemical bonds in an ordered way. The chains are thought to be fully extended for classification purposes; a distinction is made among one-, two-, and three-dimensional assemblies or, according to IUPAC, among catena (Fig. 3 –1), phyllo (Fig. 3 –2) and tecto (Fig. 3 –3) compounds.

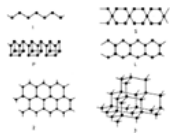

Figure 3. Schematic representation of some ordered chain assemblies

1 = Catena polymer in all-*trans* conformation; S = Spiro polymer; P = Pearl-stringlike polymer; L = Ladder polymer; 2 = Phyllo polymer; 3 = Tecto polymer

Linear polymers are catena polymers in this nomenclature since the fully extended chains extend in only one direction (Fig. 3 –1). Spiro polymers possess two chains with connecting atoms shared by both of them (Fig. 3-S).

Ladder polymers consist of two chains joined by bonds in regular intervals (Fig. 3-L); they are also called double-strand polymers. Such ladder polymers are generated by first forming a linear chain with pendant side groups that are subsequently polymerized. An example is the polymerization of 1,3-butadiene to 1,2-polybutadiene with subsequent cyclization to cyclopolybutadiene

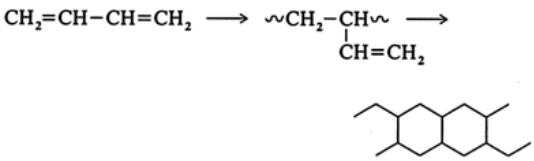

Ladder polymers may also be synthesized by two-step polycondensation or polyaddition reactions; in general, however, only fairly short ladderlike units are formed in this way (see below). One-step ladder syntheses are very rare. Polymerization of the silicon analogue of cubane, $(C_6H_5)_8Si_8O_{12}$, does not lead to a ladder polymer of type L in Figure 3 but to a pearl-string-like polymer P (Fig. 3-P). Many silicates are ladder polymers or double-ladder polymers.

Phyllo polymers are also called layer or parquet polymers (Fig. 3). Graphite and the silicates montmorillonite, bentonite, and mica are well-known examples. Cell walls of bacteria consist of bag-shaped polymers (i.e., layer polymers). Diamond and quartz are examples of tecto polymers (Fig. 3).

2.1.5. *Unordered Networks*

Multifunctional oligomer and polymer molecules can be interconnected via chain groups or end groups to cross-linked polymers (networks). The cross-linking points must be at least trifunctional; that is, the cross-linkable groups must be at least bifunctional. The cross-linking bridges between chains can be short or long; the distribution of cross-linking points, at random or in clusters. Both chemical and physical bonds can be employed for cross-linking; chemically and physically cross-linked polymers are thus distinguished.

Chemical Networks. The structure of chemical networks depends on the relative amount of cross-linking sites and the state of the reactants during the cross-linking reaction. The network structure is to a first approximation independent of the chemical structure of the cross-linking sites. The relative amount of cross-linking sites determines the cross-link density. Light cross-linking does not change the mobility of chain segments between cross-links, and the resulting networks behave as elastomers if the chain segments are above the glass transition temperature. Networks with strong cross-linking or chain segments below the glass transition temperature are employed for thermosets.

Networks are fairly *homogeneous* with respect to the distribution of cross-links if they are prepared by polymerization in bulk or from homogeneous solutions and if the corresponding linear chains are soluble in their own monomers or in the applied solvent (Fig. 4, HN). Inhomogeneous networks are formed if phase separation occurs during polymerization; for example, if the low molar mass polymers generated early are insoluble in the applied solvent or the remaining monomer. The resulting *macroreticular networks* exhibit cavities (Fig. 4, MN).

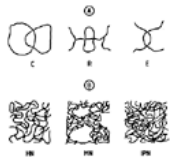

Figure 4. Schematic representation of cross-linked polymers

A) Physical cross-links; B) Chemical networks

C = catenane; E = entanglement; HN = homogeneous network; IPN = interpenetrating network; MN = macroreticular network; R = rotaxane

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

In principle, two chemically different networks poly (a) and poly (b) can coexist independently of each other in interpenetrating networks (IPNs). In reality, these networks are not molecularly interdispersed; rather, each network forms interconnected domains with higher cross-link densities. Semi-interpenetrating networks consist of cross-linked poly (a) in uncross-linked poly (b).

Chemical cross-links are in most cases irreversible. These cross-linked polymers do not dissolve in solvents. With an increase in temperature, subchains decompose rather than revert to the original monomers. Chemically reversible cross-linked networks are being investigated because they would be of value in the recycling of thermosets.

Physical Networks. Physically cross-linked polymers, on the other hand, are in most cases reversibly cross-linked. Examples are microdomains in block copolymers, ion clusters in ionomers, or even crystallites in partly crystallized polymers. Physically cross-linked polymers often dissolve in suitable solvents.

Entanglements (Fig. 4, E) behave under strain as physical cross-links because the two chains cannot diffuse away rapidly enough; the chains disentangle, however, if sufficient time is allowed. Catenanes (Fig. 4, C) and rotaxanes (Fig. 4, R) have irreversible physical cross-links. Catenanes consist of two intertwined rings. Such structures are well known for deoxyribonucleic acids; they have also been observed for some synthetic polymers. Rotaxanes are rings on chains that are sealed on both ends by branched or cross-linked structures in such a way that the ring cannot slip out. Such structures are suspected to exist in some cross-linked polydimethylsiloxanes.

2.2. Molar Masses and Molar Mass Distributions

Molar masses of polymers can be varied over a wide range; they may extend from several hundreds to several millions, albeit not for every polymer type. The molar mass ranges used industrially depend on the limitations of the available polymerization procedures, the restrictions imposed by processing methods, and the desired properties of the plastics.

Presently available polymerization methods all employ statistical approaches to initiation and propagation of growing chains, as well as to termination and transfer, where applicable. The resulting polymer molecules thus show a distribution of molar masses; they are not molecularly uniform, and the experimentally determined molar masses represent averages. Synthetic copolymers are, in addition, nonhomogeneous with respect to the distribution of different types of monomer units; they are polydisperse with respect to both constitution and molar mass.

In contrast, biopolymers are synthesized via matrix polymerization. Many (if not all) biopolymers are monodisperse in their biological environments: each macromolecule of a polymer has the same constitution and the same molar mass. Examples are nucleic acids and enzymes. Upon isolation from living cells and further processing, some degradation may occur. Industrially employed celluloses thus exhibit molar mass distributions.

The recently discovered starburst dendrimers are synthetic polymers that are molecularly uniform. Their molecular homogeneity approaches that of isotopic distributions in atoms and molecules. It is caused by a self-limitation of molecular growth due to surface crowding after a certain number of growth generation steps [70].

2.2.1. *Molar Mass Average*

Many polymer properties depend on the type and range of molar mass distributions. In many cases, molar mass averages are determined instead, in part because it is

experimentally easier and in part because a number is simpler to grasp than a function.

The averages employed in polymer science are usually arithmetic averages. The polymer molecules of size i (degree of polymerization X_i, molar mass M_i) are counted according to their statistical weight g_i, which may be their mole fraction $x_i \equiv n_i / \Sigma_i n_i$, their mass fraction $w_i \equiv m_i / \Sigma_i m_i = x_i M_i / \bar{M}_n$, or their z-fraction defined as $Z_i \equiv z_i / \Sigma_i z_i = w_i M_i / \bar{M}_w = x_i M^2_i / (\bar{M}_n M_w)$. The number-, mass-, and z-average molar masses are defined as

$$\bar{M}_n \equiv \Sigma_i x_i M_i = \frac{\Sigma_i w_i}{\Sigma_i (w_i / M_i)} \quad (1)$$

$$\bar{M}_w \equiv \Sigma_i w_i M_i = \frac{\Sigma_i x_i M_i^2}{\Sigma_i x_i M_i} \quad (2)$$

$$\bar{M}_z \equiv \Sigma_i Z_i M_i = \frac{\Sigma_i w_i M_i^2}{\Sigma_i w_i M_i} = \frac{\Sigma_i x_i M_i^3}{\Sigma_i x_i M_i^2} \quad (3)$$

Each of these three average moments is essentially a single moment of the distribution of molar masses present. There are, however, molar mass averages which are composed of more than one moment, for example, molar masses from sedimentation and diffusion coefficients for certain types of molecular shapes and solvent interactions.

2.2.2. *Determination of Molar Mass*

Many methods exist for the determination of molar masses. Most important for the *number-average molar mass* is *osmometry*, which measures the osmotic pressure Π of polymer solutions at various polymer concentrations c against the solvent. Solution and solvent are separated by a semipermeable membrane that is permeable to the solvent but not to the polymer molecules.

Older osmometers determined osmotic pressure in true thermodynamic equilibrium, which was reached only after days and weeks because large volumes of solvent had to be moved through the membrane. Modern osmometers compensate for the increase in pressure difference between solution and solvent caused by the flow of solvent into the solution chamber via a servomechanism by a change of solvent pressure; equilibrium pressures are established after 10 – 30 min. Reduced osmotic pressures, Π/c, determined at various concentrations, are extrapolated to zero concentration to give the number-average molar mass according to the van't Hoff equation, $\bar{M}_n = RT/[\lim_{c \to 0} (\Pi/c)]$. Number-average molar masses may be determined by osmometry in the range of 10^4 g/mol (lower limit of semipermeability of membranes) to ca. 10^6 g/mol (upper limit of sensitivity).

Mass-average molar masses are determined mainly via static *light scattering* of dust-free dilute solutions. Modern instruments measure the so- called Rayleigh ratio $R_o = K \cdot c \cdot \bar{M}_{w,\,app}$, at very low angles relative to the incident light beam, where K is an optical constant and $M_{w,\,app}$ an apparent mass average molar mass which is calculated from R_o, c, and K. Values of $K \cdot c/R_o$ are measured at various polymer concentrations and extrapolated to zero concentration. The limiting value $\lim_{c \to 0} (K \cdot c/R_o)$ delivers the inverse mass-average molar mass, $1/\bar{M}_w$. Molar masses from several hundred to several million grams per mole can be determined by light scattering, the upper limit being given by experimental problems such as multiple scattering. If measurements are performed over a wide range of scattering angles, important additional information about the radii of gyration of the molecules can be gained from the angular dependence of the Rayleigh ratio.

The *viscometric method* is fast and simple and therefore the most often applied method. From solvent viscosity η_o and viscosities η of dilute solutions of various concentrations c (measured at low shear rates), reduced viscosities (i.e., viscosity numbers) are calculated via $\eta_{red} = (\eta - \eta_o)/(\eta_o \cdot c)$. The reduced viscosities are extrapolated to zero concentration to give the *intrinsic viscosity* (limiting viscosity number) $[\eta]$ (DIN recommends the symbol J_o). This quantity has the physical unit of a specific volume since the concentration c is measured as mass per volume. The recommended units are milliliters per gram (cubic centimeters per gram), but data in 100 mL/g are still found in the literature; in the older literature, $[\eta]$ was given as Z_η in liters per gram.

Intrinsic viscosities measure the volume occupied by 1 g polymer at $c \to 0$, that is, the specific volumes (inverse densities) of isolated polymer molecules. Since molecular density varies with size for most molecule shapes, intrinsic viscosities can be used to measure molar masses for a homologous series of polymer molecules.

The viscosity increment η_i of unsolvated, compact spheres is, according to EINSTEIN:

$$\eta_i = (\eta - \eta_o)/\eta_o = (5/2)\phi = (5/2)(N_A V_H c/M) \quad (4)$$

where N_A is the Avogadro number, Φ the volume fraction, $V_H = \rho_H/m$ the hydrodynamic volume, ρ_H the density, and m the mass of a sphere. Hard spheres from the same material have the same density. Since $m \sim M$ and $\eta_i/c \approx [\eta]$, intrinsic viscosities of spheres are independent of their molar masses (i.e., $[\eta] = K \cdot M^0$).

The hydrodynamic densities of other particle shapes such as rods and coils change, however, with the molar mass

$$[\eta] = K \cdot M^a \quad (5)$$

The exponent $a = 2$ for rigid rods of infinite length. Theory and experiment indicate a value of $a = 1/2$ for random coils in the unperturbed state (see below); this is a limiting case for very flexible chains in theta solvents. A theta solvent is a solvent being at a temperature at which a polymer adopts the theta state, i.e., where coil molecules are in their unperturbed state with a Gaussian distribution of segments (no excluded volumes). A theta solvent is a thermodynamically bad solvent. Theory predicts $a = 0.764$ for nondraining coils of high molar mass in thermodynamically good solvents (Fig. 5). Experimentally, values of $0.5 < a < 0.764$ are often found for flexible chains (such as most thermoplastics in solution) because their molar masses are not high enough for the assumptions of the theory to be fulfilled.

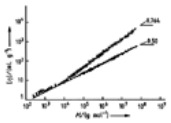

Figure 5. Molar mass dependence of intrinsic viscosities of flexible, coillike molecules: polystyrene in the theta solvent cyclohexane at 34.5 °C (•) and in the thermodynamically good solvent benzene at 25 °C (o)

The slopes at high molar masses adopt the theoretical values of 0.764 (good solvent) and 0.500 (theta solvent)

Wormlike molecules may adopt values of $0 < a < 2$ because they resemble spheres and ellipsoids at low molar masses, rigid rods at medium ones, and random coils at very high molar masses. This behavior is found for stiff molecules such as the double helices of deoxyribonucleic acids in aqueous solutions or certain poly(α-amino acids) in helicogenic solvents (see Chap. Supermolecular Structures).

Viscometry delivers a peculiar molar mass average, such as can be derived from equation (5):

$$\bar{M}_\eta = ([\eta]/K)^{1/a} = (K^{-1} \cdot \Sigma_i w_i [\eta]_i)^{1/a} = (\Sigma_i w_i M_i^a)^{1/a} \quad (6)$$

Numerical values of viscosity-average molar masses $\bar{M}_\eta$ lie between number and mass averages for $0 < a < 1$ and become identical with mass averages if $a = 1$.

2.2.3. *Molar Mass Distributions*

The types and characteristic parameters of molar mass distributions of polymers are determined thermodynamically or kinetically by the synthesis conditions and, to a smaller extent, by the processing of plastics. The molar mass distributions are mathematically described by distribution functions, which may be discontinuous (discrete) or continuous and differential or integral (cumulative) (Fig. 6). In these functions, polymer molecules of size i (molar mass M_i,degree of polymerization X_i) are considered according to their statistical weights, which may be mole fractions x_i, mass fractions w_i, etc.

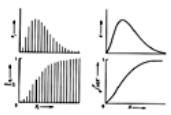

Figure 6. Distributions of mole fractions x of degrees of polymerization X

Left: discontinuous (discrete); right: continuous; top: differential; bottom: integral (cumulative)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

All molar mass distributions are discrete because successive polymer molecules of a homologous series differ from each other by a degree of polymerization of 1 and the molar mass M_u of one monomeric unit. The discrete distribution functions can normally be replaced by continuous ones since so many different degrees of polymerization exist and the difference in molar masses between molecules with M_i and M_{i+1} is small compared to the average molar mass.

The various types of distribution functions are usually named after their discoverers. The *Schulz – Zimm distribution* function derives from processes in which reactive chains add other monomer or polymer molecules until the chains are deactivated. Such deactivation processes may be chain termination reactions (chain polymerization, polyelimination) or the decrease of reactivity by lowering of temperature (polycondensation, polyaddition). The mass fraction of polymer molecules with degree of polymerization X_i is given by

$$w_i = \frac{(k/\bar{X}_n)^k \cdot X_i^k \cdot \exp(-kX_i/\bar{X}_n)}{\Gamma(k+1)} \tag{7}$$

where k is the degree of coupling of chains (e.g., $k = 2$ for recombination of two growing chains to one dead chain and $k = 1$ if two growing chains react by disproportionation to two dead chains) and Γ ($k = 1$) the gamma function of ($k = 1$). This function is called *Schulz – Flory* or *Flory distribution* for the special case of $k = 1$ (polycondensation, radical chain polymerization with termination by disproportionation). Equation (34 reduces for $k = 1$ and high degrees of polymerization to

$$w_i = (1/\bar{X}_n)^2 \cdot X_i \cdot [1 - (1/\bar{X}_n)]^{x_i} \tag{8}$$

The various average degrees of polymerization are connected via

$$\bar{X}_n/k = \bar{X}_w/(k+1) = \bar{X}_z/(k+2) \tag{9}$$

Poisson distributions are formed if a constant number of chains starts to grow simultaneously and if monomer molecules are added to the chains at random and independent of previous additions. The mass fraction of molecules of degree of polymerization X_i is given by

$$w_i = \frac{X_i \cdot (\bar{X}_n - 1)^{X_i - 1} \cdot [\exp(1 - \bar{X}_n)]}{(X_i - 1)!(\bar{X}_n)} \tag{10}$$

and the interrelationship between the various averages by

$$\bar{X}_w/\bar{X}_n = 1 + (1/\bar{X}_n) - (1/\bar{X}_n)^2 \tag{11}$$

The ratio of mass to number average approaches 1 at infinitely high molar masses. The Poisson distribution is thus a very narrow distribution, in contrast to the Schulz – Zimm distribution where the ratio $\bar{X}_w/\bar{X}_n$ equals 2 for $k = 1$ and 3/2 for $k = 2$.

The *Schulz – Zimm distribution* is a special case of the *Kubin distribution*, an empirical, generalized exponential distribution (GEX distribution) with three empirical, adjustable parameters γ, $\in$, and β:

$$w_i = \frac{\gamma \cdot \beta^{(\varepsilon+1)/\gamma} \cdot X_i^{\varepsilon} \cdot \exp(-\beta X_i^{\gamma})}{\Gamma[(\varepsilon+1)/\gamma]} \tag{12}$$

This expression converts to the Schulz – Zimm distribution if $\gamma = 1$, $\in = k$, and $\beta = k/\bar{X}_n$. The Kubin distribution also includes the *Tung distribution* ($\in = \gamma - 1$) and the logarithmic normal distribution ($\gamma = 0$; $\in = \infty$), both of which frequently describe molar mass distributions of polymers from Ziegler – Natta polymerizations. The Kubin distribution is a very adaptable distribution because it contains a “stretched exponential” (i.e., a variable X with an exponent γ in the exponential term).

Some of these distribution functions are compared in Figure 7. Note that only the logarithmic normal distribution shows a maximum in the distribution of mole fractions and that none of the maxima in the distributions of the mass fractions correspond to simple molar mass averages.

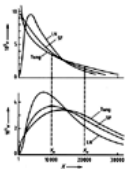

Figure 7. Continuous differential distributions of degrees of polymerization X for a) the logarithmic normal distribution (LN); b) the Schulz – Flory distribution (SF); and c) the Tung distribution (Tung), shown as distributions of mole fractions (top) and corresponding mass fractions (bottom) for a polymer with $\bar{X}_n$ = 10 000 and $\bar{X}_w$ = 20 000

The Poisson distribution is so narrow that it is practically identical with the vertical line for $\bar{X}_n$ = 10 000.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

2.2.4. *Determination of Molar Mass Distributions*

Molar mass distributions can be determined by preparative *fractionation* of polymers from solutions since the various species of a polymer-homologous series exhibit small differences in solubility. Fractionation occurs upon change of temperature or addition of a nonsolvent. The molar masses of the resulting fractions are determined in separate experiments. Fractionations are inexpensive but time consuming.

The method of choice for the fast determination of molar mass distribution is *size exclusion chromatography* (SEC). Dilute polymer solutions are placed on the top of a column filled with a porous carrier. Low molar mass molecules can enter the pores, but high molar mass molecules cannot. Medium-sized molecules enter with difficulty and remain for shorter times than low molar mass molecules. Higher molar masses are thus eluted first, and elution curves are observed that give the concentration of eluted molecules as a function of the eluted volume V_e. The maximum of the elution curve is called the retention volume.

The retention volumes depend on the SEC system (carrier, solvent, temperature) and the polymer being investigated. The carrier may consist of rigid porous materials (e.g., porous glass beads) or swollen, cross-linked polymers (e.g., cross-linked polystyrenes or dextrans). In the latter case, the method is called gel permeation chromatography (synthetic polymers) or gel filtration (biopolymers).

Retention volumes are constant below and above certain molar masses of the polymers (Fig. 8). The volume V_o at lower molar masses gives the total volume available for the flow of solvent; the volume V_i at upper molar masses indicates the interstitial volume between carrier particles. Polymer molecules can thus be separated only in the range $V_i < V_e < V_o$, where V_e is the elution volume.

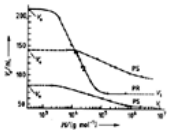

Figure 8. Molar mass dependence of elution volumes V_e

a) Coil-like linear polystyrenes PS with two different SEC columns from cross-linked polystyrenes; b) Various spheroidal proteins PR with cross-linked dextran

V_o = Total volume; V_i = Interstitial volume

Elution volumes depend to a first approximation on the logarithms of molar masses (Fig. 8). A function $V_e = K \cdot \ln M$ is thus often used to determine unknown molar masses with the help of a constant K derived from calibrations with narrow-distribution polystyrene standards. This procedure does not give absolute molar masses because polystyrenes and test polymers with the same molar masses possess different chromatographically effective volumes. A frequently used “universal” calibration method thus employs a function $V_e = f(\log [\eta] \cdot M)$ since $[\eta]$ measures the specific hydrodynamic volume of the solute molecules and $[\eta] \cdot M$ has the physical unit of a molar volume.

2.3. Configurations

Polymers with symmetric repeating units, such as $-CH_2-CH_2-(-CH_2-CH_2-)-CH_2-CF_2-$, and $-NH-CO-(CH_2)_5-$, do not possess configurational isomers. Such isomers do exist, however, for polymer molecules with nonsymmetric repeating units, such as $-CH(CH_3)-$ (polymethylmethylene), $-CH(CH_3)-CH_2-$ (polypropylene), and $-NH-CO-CH(CH_3)-$(polyalanine). Polypropylene, for example, has two configurational repeating units **9** and **10** with two different monomeric units each:

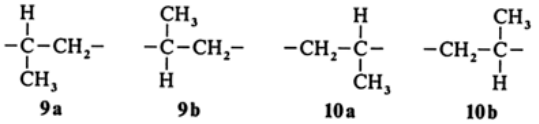

9a **9b** **10a** **10b**

The configurational units **9a** and **9b** are enantiomeric; they belong to the same constitutional unit – $CH(CH_3)$ – CH_2 –. The configurational units **9a** and **10a**, on the other hand, are based on two different constitutional units. Several configurational units may be joined to give steric repeating units; the three simplest steric repeating units for polypropylene are

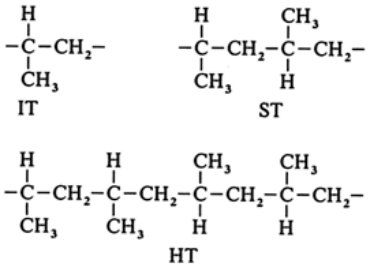

IT ST HT

The repetition of these units leads to polymer chains that are called isotactic (IT), syndiotactic (ST), and heterotactic (HT). A *heterotactic unit* ht in a polymer chain consists of three monomeric units (the first or last three of HT), but the repetition of such ht-triads does not lead to a completely heterotactic polymer HT because it would consist of alternating heterotactic and syndiotactic triads. Whether **9a** or **10a** is used as the simplest configurational repeating unit is immaterial since infinitely long polypropylene chains of **9a** differ from those of **10a** only by the orientation of these units.

The term *tacticity* thus refers to the relative arrangements of configurational units in a chain. Relative configurations are classified by starting at one end of the polymer chain and considering the configuration around a central atom relative to the preceding one. This classification is different from that of the absolute configuration of organic chemistry where the configuration of each central atom is determined relative to the ligand with the lowest seniority.

The ligands R of *isotactic polymer* molecules are always "on the same side" if these molecules are shown in Fischer projections or in other stereo formulas with hypothetical *cis* conformations of the chains. In stereo formulas based on *trans* conformations, ligands are only on the same side for isotactic molecules consisting of two chain atoms per monomeric unit (Fig. 9), that is ~(CHR – CH_2)~.

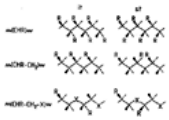

Figure 9. Isotactic (it) and syndiotactic (st) polymers with monomeric units – CHR –, – CHR – CH_2 –, and – CHR –CH_2–X– in *trans* conformations

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Real polymer chains of the types ~CHR~, ~CHR – CH_2~, ~X – CHR – CH_2~, etc., contain configurational mistakes; they are neither 100 % isotactic nor 100 % syndiotactic. The tacticity of such chains is expressed by the fraction X_J of their isotactic or syndiotactic diads, triads, etc. A diad consists of two monomeric units, a triad, of three; etc. Each monomeric unit of a polymer chain belongs to two tactic diads, three tactic triads, four tactic tetrads, and so on. Figure 9 thus shows five diads, four triads, three tetrads, . . . , etc., for ~ CHR ~. The sum of the mole fractions of all J-ads of given J must equal 1 ($x_i + x_s \equiv 1$, $x_{ii} + x_{is} + x_{si} + x_{ss} \equiv 1$, etc.), where i=isotactic diad, s=syndiotactic diad, ii=isotactic triad of two isotactic diads, etc. The mole fraction of isotactic diads is given by the mole fraction of isotactic triads plus 1/2 of the mole fraction of the sum of the two heterotactic triads, that is, $x_i = x_{ii} + (1/2)(x_{is} + x_{si})$, etc. The number-average sequence length of isotactic sequences is thus given by $(\bar{X}_I)_n = 2\, x_i/(x_{is} + x_{si})$.

The presence and/or fraction of tactic J-ads can be investigated by a number of experimental methods. 2 D-NMR spectroscopy allows absolute determination of the types of J-ads and their amounts up to pentads, whereas conventional high-resolution NMR spectroscopy requires prior knowledge of the J-ad type by X-ray crystallography or other methods. Infrared spectroscopy detects only diads, whereas crystallinity, solubility, glass and melt temperatures, and chemical reactions may or may not indicate the presence of shorter or longer tactic sequences.

Highly isotactic polymers can be produced by Ziegler – Natta polymerization of α-olefins (e.g., it-polypropylene or it-polybutylene-1). Special Ziegler – Natta catalysts lead to st-polypropylene and st-polystyrene; these syndiotactic polymers have no commercial use thus far. Highly syndiotactic polymers are also generated by most very low-temperature free-radical polymerizations of vinyl and acryl monomers. Conventional vinyl and acryl polymers are, however, synthesized at ambient temperature or above, and consist, at best, of slightly syndiotactic polymers; in general, they are considered atactic polymers.

Polymers with different geometric isomerism are also tactic. Depending on the configuration of the chain segments relative to the double bonds in the chain *cis*-tactic (ct) and *trans*-tactic (tt) structures are distinguished. The ct structure of1,4-polyprenes[~ $(CH_2 – CR{=}CH – CH_2)_n$ ~] from the polymerization of 1,3-dienes $CH_2{=}CR – CH{=}CH_2$ via both double bonds corresponds to an E isomer; the tt structure, to a Z isomer. If the polymerization proceeds via one double bond only, 1,2-polydienes are formed that may be isotactic, syndiotactic, atactic, etc.

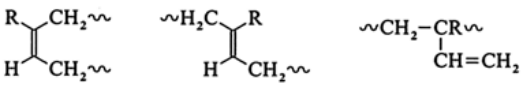

cis-1,4 (E) *trans*-1,4 (Z) 1,2 (it or st)

2.4. Conformations

2.4.1. *Microconformations*

Rotations of atoms or groups of atoms around single bonds create spatial arrangements called conformations (organic chemistry) or microconformations (macromolecular chemistry). The sequence of these microconformations determines the shape of the macromolecule — the macroconformation (in statistical mechanics, the configuration).

In principle, an infinite number of conformations is possible around each single bond. In practice, certain positions are energetically preferred; only these are called (micro) conformations. In two joined tetrahedrons, such as ethane ($H_3C – CH_3$), two extremes of energetically different positions are possible: the staggered position corresponds to a minimum of energy and the eclipsed position to a maximum of energy if repulsive forces are prevalent (Fig. 10). On rotation by 360° around the C – C axis, three energetically equivalent eclipsed and three energetically equivalent staggered positions may be occupied since the three H atoms bound to the same C atom ("bonded atoms") are equivalent. Small molecules, such as ethane, can be considered as definite species in conformations with energy minima; they are called conformers, rotamers, or rotational isomers.

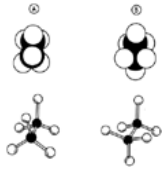

Figure 10. Microconformations of ethane

A) Eclipsed position (*cis* or synperiplanar); B) Staggered position (*trans* or antiperiplanar)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The number of types of conformers increases if the three bonded atoms are not equivalent, for example, in butane ($CH_3 – CH_2 – CH_2 – CH_3$) (one CH_3 group and two H atoms

bonded to each carbon atom participating in the central C – C bond). In polymer chains, one of the bonded atoms is never equivalent because it is part of the chain; in polyethylene (~ CH_2 – CH_2 ~), the bonded atoms are two H atoms and the chain. There are two energetically different eclipsed positions (*cis* and *anti*) and two energetically different staggered ones (*trans* and *gauche*). Each *gauche* and each *anti* position can occur in two spatially different positions that are energetically equivalent (plus and minus) if two of the three bonded atoms are equivalent. These (micro)conformers have different names in macromolecular and in organic chemistry (Fig. 11).

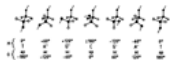

Figure 11. Chain conformations and their names and symbols in macromolecular (M) and organic (O) chemistry

• Chain atoms; ○ substituents; T = *trans*, A = *anti*; G = *gauche*; C = *cis*; ap = antiperiplanar; ac = anticlinal; sc=synclinal; sp = synperiplanar

All eclipsed positions are sterically hindered in polymer chains and therefore only *trans* and *gauche* positions must be considered. The conformational energy is the energy difference between the energies of the *trans* and *gauche* conformations. An activation energy (potential energy) is necessary to overcome the rotational barrier between *trans* and *gauche* conformations. This rotational barrier increases with decreasing length of the central bond and with increasing number and increasing size of bonded atoms. Its value is, for example, 12.1 kJ/(mol bond) for the C – C bonds in polyethylene chains ~$(CH_2 – CH_2)_n$~ but only 2.1 kJ/mol for the CH_2 – CO bonds in polyester chains with units ~CH_2 – CO – O – CH_2~. Since these bond energies can easily be overcome by thermal energy, (hindered) rotations are possible around the chain bonds and the molecules may adopt many chain conformations (Section Conformations in Polymer Solutions). Isolated aliphatic polyester chains ~$OOC(CH_2)_x$~ (with x>3) are thus more flexible than carbon chains; such polymers are used as polymeric plasticizers.

Two different types of flexibility are distinguished. A chain molecule is said to be *statically flexible* if it possesses many accessible conformational minima. *Dynamic flexibilities* are characterized by low barriers between conformational minima.

2.4.2. *Conformations in Ideal Polymer Crystals*

A regular sequence of microconformations leads to regular macroconformations of polymer chains. The chains are linearly extended; the end-to-end distance is commonly called the contour length r_{cont} of the chain (this term originally referred to the contour of the chain along the individual chain bonds). Chains in all-*trans* microconformations are said to be fully extended, for example, the chain in Figure 3-1.

Polyethylene crystallizes ideally in an all-*trans* conformation because the shortest distance between nonbonded hydrogen atoms (0.254 nm) is greater than the sum of the van der Waals radii of H atoms (0.24 nm). The size of substituents R in isotactic poly(α-olefins) (~$(CH_2$ – CHR$)_n$ ~), however, forces the microconformation around each second chain bond to adopt a *gauche* position (see Fig. 11). All *gauche* conformations must be alike for steric reasons; conformational diads G^+G^- and G^-G^+ are forbidden (Fig. 11). The chain thus adopts either a . . . $TG^+TG^+TG^+$ . . . or a $TG^-TG^-TG^-$ . . . macroconformation; that is, it becomes helical. The number of monomeric units per complete turn is determined mainly by the size of the immediate substituents. it-Polypropylene ~ $(CH_2 – CH(CH_3))_n$ ~ has three propylene units per one turn (3_1 helix, Fig. 12 A), it-poly(4-methylpentene-1) (P4MP) ~ $(CH_2 –CH(CH_2CH(CH_3)_2))_n$ ~ has seven units per two turns (7_2 helix = 3.5 helix, Fig. 12 B), and poly(3-methylbutene-1) (P3MB) has four units per one turn (4_1 helix, Fig. 12 C). The conformational angles are not necessarily the ideal ones of 0° for *trans* and 120° for *gauche* as found for it-polypropylene; they are rather –13°/110° for P4MP and –24°/96° for P3MB. Conformational positions with deviations up to ± 30° from the ideal conformational angles are still named after the ideal microconformations.

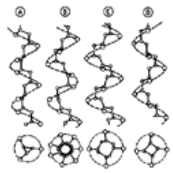

Figure 12. Helix types of poly(α-olefins) ~ CH_2– CHR~

A) 3_1 Helix; B) 7_2 Helix; C), D) Different types of 4_1 helices

Reprinted with permission by Societa Italiana di Fisica, Bologna, from [71]

The R substituents of syndiotactic vinyl polymers [~$(CH_2$ – CHR$)_n$~] are farther apart than those of their isotactic counterparts (see Fig. 9). In general, *trans* conformations thus have the lowest energy in those st-polymers where only repulsive forces operate. Attractive forces, such as intramolecular hydrogen bonds between neighboring OH groups in poly(vinyl alcohol) (~$[CH_2$ – CH(OH)$]_n$~, PVAL), lead to different conformational sequences: isotactic PVAL exists in all-*trans* conformations, whereas syndiotactic PVAL forms helices.

Chain atoms with free electron pairs lead to *gauche* effects. The chains of crystallized polyoxymethylene [~(O – $CH_2)_n$~] exist in all-*gauche* conformation (9_5 helix). Polyoxyethylene [~(O – CH_2 – $CH_2)_n$~] and polyglycine [~(NH – CO – $CH_2)_n$~], on the other hand, possess the conformational sequence TTG. The resulting 7_2 helices of polyglycine are stabilized by intramolecular hydrogen bonds between the first, fourth, seventh, etc., peptide bonds of a chain.

Polypeptides [~(NH – CO – CHR$)_n$~] are based on chiral monomeric units. An L-polymer may thus form two different helices: a right-handed and a left-handed one. These helices are diastereomers with different energy contents. In general, one “handedness” is preferred over the other in polymers with chiral base units. Poly (L- α-amino acids) generally form right-handed helices, whereas poly([*S*]- α-olefins) and most polysaccharides from D-sugars exist as left-handed ones.

2.4.3. *Conformations in Polymer Solutions*

Helices can survive melting and dissolution processes only if the helical structure is stabilized by intramolecular attractive forces. Examples are the hydrogen bonds in helices of poly α-amino acids or hydrogen bonds plus base stacking in double helices of deoxyribonucleic acids. These forces may be so strong that the molecules decompose rather than melt on heating. A delicate balance between intramolecular bonding and solvation of substituents is necessary to preserve helical structures in solution; examples are poly (α-amino acids) in helicogenic solvents.

Helices that are generated in crystals by packing of chains whose microconformations originate from repulsive forces do not survive the melting process intact. Each macromolecule can form many macroconformers that equilibrate rapidly. Only very short helical sequences in very low concentrations may thus exist in melts.

Two extreme cases must be considered for the dissolution of polymers. Only weak interactions (or none at all) exist between monomeric units of apolar polymers and apolar solvent molecules. Conformational changes are thus entropy driven; the sequence of microconformations is irregular, and the polymer molecule adopts the macroconformation of a coil.

In apolar enantiomeric polymers in apolar solvents, long conformational sequences are conserved, although fast conformational transformations from left- to right-handed helices and vice versa may occur in enantiomeric polymers. On average, few microconformations are converted into other ones. These helical sequences may be stabilized by association processes.

Polar solvents, on the other hand, interact strongly with polar polymers; these solvents cause strong changes in microconformations. Since the ligands around each chain bond can adopt various microconformations and these microconformations can change rapidly, a given macromolecule may exist in time in many macroconformations, similar to those of simple apolar polymers in apolar solvents:

A polymethylene [~$(CH_2)_n$~] with degree of polymerization n = 20 001 possesses 20 000 chain bonds, each of which can adopt three microconformations: *trans*, *gauche* (plus), and *gauche* (minus). According to statistics, such chain can exist in $3^n = 3^{20\,000} \approx 10^{9542}$ different macroconformations. Conversely, a collection of chains may have many macroconformations at any given time. None of these macroconformations adopts a simple geometric shape, not even instantaneously. Rather, the chains form rapidly changing coil structures that can be made visible by electron microscopy as two-dimensional projections of the three-dimensional coil shape for chains of sufficiently long chain diameter (Fig. 13).

Figure 13. Electron micrograph of the double helix molecules of deoxyribonucleic acid chains showing two-dimensional random coils

Reprinted with permission by Academic Press, London, from [72]

2.4.4. *Unperturbed Coils*

The instantaneous shapes of coils cannot be determined by presently known methods. Their radius of gyration can be measured, however, by light scattering, small-angle X-ray scattering, and small-angle neutron scattering. The end-to-end distances of coiled linear chains can be modeled for various types of chains and calculated for individual chains via the rotational isomeric state method.

The simplest model is that of a freely jointed chain (no fixed bond angles) that assumes infinitely thin segments without interactions. The mean square end-to-end distance, $\langle r_{oo}^2 \rangle$, of such a coil is given by the number *N* of its unspecified segments with segment lengths *b*:

$$\langle r_{oo}^2 \rangle = N \cdot b^2 \tag{13}$$

Real chains posses fixed bond angles τ between chain atoms and finite torsion (or dihedral) angles θ (θ = 0° for *trans* conformations). The mean square end-to-end distance for such a chain in the unperturbed state (zero net polymer – polymer and polymer – solvent interactions) is given by

$$\langle r_o^2 \rangle = N \cdot b^2 \cdot \left(\frac{1 - \cos\tau}{1 + \cos\tau}\right) \cdot \left(\frac{1 + \cos\theta}{1 - \cos\theta}\right) \tag{14}$$

$$\langle r_o^2 \rangle = N \cdot b^2 \cdot \left(\frac{1 - \cos\tau}{1 + \cos\tau}\right) \cdot \sigma^2 \tag{15}$$

The factor σ is called the *hindrance parameter* or steric factor. Often the hindrance parameter and the bond angle are combined to give the *characteristic ratio* C_N:

$$\begin{aligned} C_N &= \langle r_o^2 \rangle / (N^2 \cdot b^2) \\ &= (1 - \cos\tau) \cdot (1 + \cos\tau)^{-1} \cdot \sigma^2 \end{aligned} \tag{16}$$

The characteristic ratio becomes approximately constant for chains with more than 100 atoms. A chain can also be characterized by its *Kuhn length* L_K, which can be calculated from its unperturbed mean square end-to-end distance, its degree of polymerization, and the effective length b_{eff} of monomeric units (i.e., length of units in chain direction projected on a plane; $r_{cont} = X \cdot b_{eff} = N_K \cdot L_K$)

$$\langle r_o^2 \rangle = N_K \cdot L_K^2 \tag{17}$$

End-to-end distances can rarely be measured directly. They are however related to the experimentally accessible radii of gyration *s* via

$$\langle s_o^2 \rangle = \langle r_o^2 \rangle / 6 \tag{18}$$

This relationship applies to all chains with random flight statistics (e.g., unperturbed chains and freely jointed chains). It is not valid for perturbed chains (Section Perturbed Coils).

Coil densities are very low (cf. Fig. 13). The volume fraction of monomeric units is, for example, only 0.012 at the center of gyration for a polyethylene chain with a molar mass of 1.19×10^6 g/mol; 99.8 % of the space of these polyethylene coils is occupied by solvent molecules (in very dilute solutions) or by units of other polymer chains (in melts).

A chain in a melt cannot distinguish between its own segments and segments of other chains. Because the interactions between the segments of various chains are also the same, such chains adopt their unperturbed dimensions, which has been confirmed by small-angle neutron scattering. Most of the coil volume is filled by segments of other chains to avoid empty space. Chains of sufficient length may thus become entangled, which manifests itself in properties such as diffusion and melt viscosity.

2.4.5. *Perturbed Coils*

Coils adopt their unperturbed dimensions in melts or in certain solvents at certain temperatures (so- called theta solvents). In such theta solvents, polymer – solvent and polymer – polymer interactions cancel each other and the chain behaves as if it is infinitely thin. In thermodynamically good solvents, polymer – solvent interactions dominate and the coil is swollen. The volume requirements of such coils must thus lead to interpenetrations by other coils even at very low concentrations. Because segments are not infinitely thin, however, a part of the total space is excluded for segments of other chains (and also for other segments of the same chain).

2.4.6. *Wormlike Chains*

Real polymer chains are not totally flexible. Their finite thickness and the partially hindered rotation around chain bonds prevent them from adopting all possible positions in space. Such chains can be described by the model of the wormlike chain (Kratky – Porod model).

The characteristic parameter of this model is the *persistence length a*. This parameter is defined as the average of the projection of the end-to-end-distance of an infinitely long and infinitely thin chain in the direction of the first segment. It can be calculated from the radius of gyration and the conventional contour length via

$$\langle s_o^2 \rangle = a^2 \{ (y/3) - 1 + (2 - y) - (2/y^2)[1 - \exp(-y)] \} \tag{19}$$

where $y = r_{cont}/a$. For flexible chains, it is related to the unperturbed radius of gyration via

$$\langle r_o^2 \rangle / 6 = \langle s_o^2 \rangle = a \cdot r_{cont}/3 = a \cdot N_K \cdot L_K/3 \tag{20}$$

For infinitely stiff chains,

$$\langle s_o^2 \rangle = a^2 \cdot y^2/12 = (r_{cont})^2/12 \tag{21}$$

that is an infinitely stiff chain behaves like an infinitely thin rod. Wormlike chains thus describe the whole transition from rodlike molecules (small *y*) to random coils (large *y*). The model is strictly valid for infinitely thin chains, but the error produced by this assumption is negligible if the persistence length is much greater than the chain diameter.

[Top of Page]

3. Polymer Manufacture

3.1. Raw Materials

The raw materials for most industrially used plastics are synthetic polymers and prepolymers; only a few are derived from naturally occurring polymers or monomers. The overwhelming majority of polymers for plastics are organic materials; very few are semiorganics (i.e., inorganic polymer chains with organic substituents). Raw materials for polymers include wood, coal, petroleum, natural gas, and certain natural oils.

3.1.1. *Wood*

Wood (⟶ Wood) is a naturally occurring composite of oriented cellulose fibers in a continuous matrix of cross-linked lignin; it is plasticized by water and "foamed" by air (vacuoles). The water content varies from 40 – 60 % for green wood to 10 – 20 % for air-dried wood. Solids include cellulose (42 %); hemicelluloses (28 – 38 %); lignin (19 – 28 %); and proteins, resins, and waxes (2 – 3 %), depending on the type and age of the plant. Most of the wood is used directly as fuel or for construction purposes, but one-sixth is converted into wood pulp or cellulose pulp for the manufacture of paper, cardboard, or rayon. A very small amount of wood is filled with monomers that are subsequently polymerized to give polymer wood.

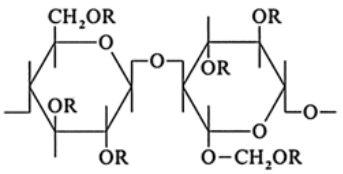

Repeating unit (cellobiose) of cellulose (R=H) and cellulose derivates

Cellulose pulp is chemically transformed into cellulose derivatives (⟶ Cellulose; ⟶ Cellulose Esters; ⟶ Cellulose Ethers):

1. 2,5-Acetate with 2,5 acetate groups, R=OOC–CH_3, per glucose residue (fibers, cigarette filters)
2. Cellulose triacetate with three acetate groups per glucose residue (fibers)
3. Cellulose acetobutyrate with R = 29 – 6 mol% acetyl and 17 – 48 mol % butyryl groups (a thermoplastic for tubes, packaging)

4. Cellulose nitrate (R = NO_2; various degrees of substitution; gun cotton, films, lacquers, celluloid)
5. Cellulose ethers with R = CH_3, C_2H_5, $CH_2CH_2OR^1$, or $CH_2 – CHOR^2 – CH_3$, where R^1 may be $CH_2CH_2OR^1$ or CH_2CH_2OH and $R^2 = CH_2–CHOR^2–CH_3$ or $CH_2 – CHOH–CH_3$ (thickeners, binders, suspension agents, injection molding materials, etc.)

Huge amounts of *lignin sulfonates* (⟶ Lignin – Commercial Lignins) are generated during wood pulp production, most of which are burned. Small amounts are used for street pavement, as binders for foundry sand, or as drilling agents, and even smaller amounts as raw materials for organic intermediates (⟶ Lignin – Uses). Degradation products of lignin have been utilized for the synthesis of ion-exchange resins and other polymers.

3.1.2. *Coal*

Coals (⟶ Coal) are polymerized and partly cross-linked hydrocarbons with average compositions between $C_{75}H_{140}O_{56}N_2S$ (peat) and $C_{240}H_{90}O_4NS$ (anthracite). They were the primary raw materials for acetylene-based monomers during the first half of this century but have since been replaced in most countries by oil and natural gas as monomer sources.

Coals deliver coke and coal tar. A number of aliphatic monomers are produced from coke via acetylene, such as chloroprene, various vinyl monomers, acrylonitrile, and hexamethylenediamine. Coal tar delivers aromatics such as benzene, xylene, phenol, and phthalic anhydride, which lead subsequently to polystyrene, phenolic resins, and glyptal resins, to name a few.

3.1.3. *Natural Gas*

Natural gas (⟶ Natural Gas) is a gas with a high proportion of aliphatic hydrocarbons. European gas is rich in methane, whereas American and Saudi Arabian gases are relatively rich in higher hydrocarbons. Natural gas is processed to synthesis gas, ethylene, and acetylene. These gases are used to produce a variety of monomers for polymers.

3.1.4. *Petroleum*

Petroleum (crude oil) (⟶ Oil Refining) is the main feedstock for monomers. It consists of 95 – 98 % hydrocarbons and 2 – 5 % oxygen, nitrogen, and sulfur compounds. The hydrocarbons are largely aliphatic, partly naphthenic, and to a small extent aromatic, depending on the source. Petroleum distillation leads to saturated hydrocarbons (gas, naphtha, gasoline, kerosene, diesel fuel, heating oil, etc.). For monomer production, these hydrocarbons are subsequently cracked to mixtures of olefins that are fractionated by distillation. The resulting compounds are used as such or are converted further into the desired monomers.

The polymer industry is a major petrochemical customer: about 60 % of the cracking and subsequent products of naphtha are used to produce polymers (ca. 50 % for plastics – elastomers, ca. 10 % for fibers). However, the final yields of polymers and converted plastics, respectively, are fairly small: 1000 kg of petroleum delivers only 30 kg of polyethylene and subsequently 16 kg of polyethylene film, in addition to 360 kg of useful byproducts [liquid gas, gasoline, propylene, etc.].

3.1.5. *Other Natural Raw Materials*

Vegetable oils are a relatively large source of raw materials. They consist of mixtures of fatty acid triglycerides and are usually subdivided into drying oils with high linolenic and linoleic acid content, semidrying oils with high linoleic and oleic acid content, and nondrying oils with high oleic acid content (⟶ Fats and Fatty Oils). Some oils are used directly for paints; others are chemically converted into monomers.

Castor oil is converted into methyl ricinoleate, which is thermally cracked into methyl undecenate and heptanal. Methyl undecenate is the raw material for 11-aminoundecanoic acid, which is subsequently polycondensed to form polyamide 11. Alkali scission of castor oil or ricinoleic acid produces sebacic acid, a monomer used to prepare polyamide (PA) 610.

Soybean oil is converted by a series of chemical reactions into the methyl ester of the C_{10} amino acid $H_2N(CH_2)_9COOCH_3$, which is the monomer for PA 10. The plant *Crambe abyssinica* contains about 55 % of erucic acid, which leads to the C_{13} amino acid and subsequently to PA 13 (~$[NH(CH_2)_{12}CO]_n$~).

Corn husks and other agricultural waste are rich in *pentosans*, which can be converted into tetrahydrofuran and further into polytetrahydrofuran $[HO(CH_2CH_2CH_2CH_2O)_nH]$, which serves as a soft block in certain polyurethanes.

The supply of raw materials from vegetable sources is relatively small. It is also fairly uncertain because quality and quantity may depend on weather conditions or be influenced by political turmoil. Most monomers for the polymer industry are thus obtained from fossil feedstocks.

3.2. Polymer Syntheses: Overview

3.2.1. *Classifications*

Polymers may be synthesized from monomers by polymerization or from other polymers by chemical transformations. The nomenclature and classification of these reactions have grown and changed with the times; they are not very systematic and are sometimes confusing.

All reactions of small molecules to macromolecules are called *polymerizations*. Polymerizations are classified according to the

1. Origin of polymers (biopolymerizations, synthetic polymerizations)
2. Chemical structure of monomers (vinyl polymerizations, diene polymerizations, ring-opening polymerizations, etc.)
3. Chemical structure of resulting polymers (linear, branching, cross-linking, isomerizing, ring-forming, etc., polymerizations)
4. Relative composition of monomers and monomer units (addition polymerization, condensation polymerization, etc.)
5. Formation of low molar mass byproducts (polycondensation vs. polyaddition, etc.)
6. Type of polymerization initiation (thermal, catalytic, photochemical, enzymatic, etc.)
7. Type of propagating species (free radical, anionic, cationic, etc.)
8. Type of mechanism (equilibrium, addition, insertion, living, stepwise, etc.)
9. Reaction media (bulk, solution, emulsion, suspension, crystal, mesophase, etc.)
10. State of matter (gaseous, homogeneous, heterogeneous, etc.)

The most common classification is according to (4) the relative chemical composition of monomer vs. monomeric unit and (5) the type of molecules that may react with the growing polymer chain. The resulting four types of polymerization can be depicted schematically as

$$P_n + P_m \rightarrow P_{n+m} \quad \text{Polyaddition}$$

$$P_n + M \rightarrow P_{n+1} \quad \text{Chain polymerization}$$

$$P_n + P_m \rightarrow P_{n+m} + L \quad \text{Polycondensation}$$

$$P_n + M \rightarrow P_{n+1} + L \quad \text{Polyelimination}$$

where M denotes monomer; P_n, P_m, P_{n+m}, P_{n+}, denote polymers and L leaving molecules (e.g., water in polyesterifications of diacids and diols).

The names of three of these four types have been used with different meanings by different researchers over the past 70 years. The following discussion thus centers around the present nomenclature recommendations by IUPAC.

Chain Polymerization. Polymer chains grow by repeated addition of monomer molecules M to polymer molecules P_i, consisting of *i* monomeric units m, an initiator end group R, and an "active center *," without the formation of leaving molecules:

$$R(m_n)m^* + M \rightarrow R(m_{n+1})m^* \tag{23}$$

An example is the polymerization of vinyl monomers (CH_2=CHR) by chain carriers Y*, which may be free radicals $Y^{\cdot}$, anions Y^-, or cations Y^+ :

$$Y-CH_2-CHR^* + CH_2=CHR \rightarrow Y-CH_2-CHR-CH_2-CHR^* \tag{24}$$

A *living polymerization* is a chain polymerization consisting only of fast initiating reactions ($Y^*+CH_2=CHR \longrightarrow Y – CH_2 – CHR^*$) and propagation reactions [$Y –(CH_2 – CHR)_n^* + CH_2=CHR \longrightarrow Y –(CH_2 – CHR)_{n+1}^*$] without termination of the growing chains and without side reactions. In such reactions, the degree of polymerization is directly proportional to the conversion *u* of monomer molecules or *p* of functional groups, respectively (Fig. 14, curve CP – L). An example is the polymerization of styrene CH_2=CH(C

$_6H_5$) by the butyl anion $C_4H_9^-$ from lithiumbutyl (butyllithium) LiC_4H_9.

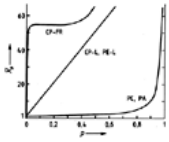

Figure 14. Change of number-average degree of polymerization $\bar{X}_n$ with the extent of reaction of functional monomer groups *p* for living chain polymerizations CP-L and living polyeliminations PE-L (with initial monomer – initiator ratio 100 : 1 mol/mol), free-radical chain polymerizations CP-FR with gel effect (schematic for high initiator concentrations), and equilibrium polycondensations PC and polyadditions PA (numerically exact)

The variation of the degree of polymerization with monomer conversion is quite different for *radical polymerizations* such as the polymerization of styrene by benzoyloxy radicals $C_6H_5COO^{\cdot}$ from the thermal decomposition of dibenzoyl peroxide, $C_6H_5COO–OOCC_6H_5$. Radicals are here formed successively and not "at once" as are the butyl anions in living polymerizations. Growing chains are terminated by reaction with other radicals. Many side reactions may occur because of the high and fairly indiscriminate reactivity of radicals. Depending on reaction conditions, the degree of polymerization may remain constant (primarily at low conversions) or increase with monomer conversion (so- called gel effect) (Fig. 14, curve CP – FR).

Chain polymerizations have generally been called *addition polymerizations* in the literature. Addition polymerization is not to be confused with polyaddition (see below).

Condensative Chain Polymerization. This is the name recommended by IUPAC for a polymerization by repeated addition of monomer molecules with elimination of low molar mass molecules L; this polymerization has thus also been called *polyelimination* [5]:

$$R(m_n)m^* + M \rightarrow R(m_{n+1})m^* + L \qquad (25)$$

Biological polymerizations to polysaccharides seem to proceed exclusively through polyeliminations. Cellulose of cotton, for example, is formed by polymerization of guanosine diphosphate D-glucose attached to a lipid matrix. The reaction proceeds via insertion of the glucose moiety into the chain with elimination of guanosine diphosphate.

Synthetic polymers are rarely formed by polyelimination. Examples are the polymerizations of α-amino acid *N*- carboxyanhydrides (Leuchs anhydrides) and the "oxidative polymerization" of 2,6-dimethylphenol to polyphenylene oxide (PPO, PPE):

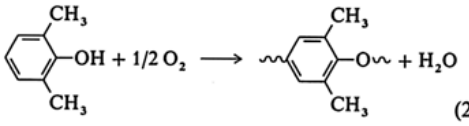

(26)

Polycondensations. In polycondensation, polymer growth proceeds by reaction between molecules of all degrees of polymerization (not just between a polymer molecule and a monomer molecule as in chain polymerization and polyelimination), with the formation of molecules having low molar mass. Many polymers with hetero atoms in the chains are generated by polycondensation. An example is the polycondensation of 11-aminoundecanoic acid to polyamide 11:

$$n\,H_2N(CH_2)_{10}COOH \rightarrow H[NH(CH_2)_{10}CO]_nOH + (n-1)H_2O \qquad (27)$$

In these reactions, dimers (*n* = 2) are formed first, which may then react with monomers to form trimers (*n* = 3) or with other dimers to form tetramers (*n* = 4), etc. The reaction is thus statistical with respect to the size of the molecules to which the functional groups are attached. The degree of polymerization of the polymer is a number average $\bar{X}_n$ over all degrees of polymerizations of the polymer molecules.

All amino groups have the same chance to react if their reactivity is independent of molecule size. The reaction of amino groups attached to molecules with higher molar mass adds, however, considerably more to the increase of the degree of polymerization than the reaction of lower molar mass molecules: the degree of polymerization snowballs with increasing extent of reaction of functional groups (Fig. 14, curve PC).

The term polycondensation was used in the past as a synonym for condensation polymerization. The latter term included both the current concepts of polycondensation and polyelimination (condensative chain polymerization).

Polyaddition. This is a polymerization by reaction among molecules of all degrees of polymerization but without the formation of leaving molecules. The number-average degree of polymerization of the resulting polymers varies with the extent of reaction of functional groups in the same way as it does in polycondensation. An example of polyaddition is the polymerization of a diisocyanate with a diol to form a polyurethane:

$$n\,OCN-R-NCO + n\,HO-R'-OH \rightarrow OCN(R-NH-CO-O-R'O)_nH \qquad (28)$$

These polymerizations have also been called "polyadduct formations." Polyaddition is not to be confused with addition polymerization, the term used in the literature for chain polymerization and occasionally also for polyaddition.

3.2.2. *Functionality*

Polymerizable monomers must be at least bifunctional under reaction conditions so that they can connect to the growing chains on one hand and generate new coupling sites on the other. Ethylene and other monomers with ethylenic double bonds are such bifunctional monomers under all polymerization conditions (e.g., see Eq. 24). Divinyl compounds, such as the divinylbenzenes $CH_2=CH – C_6H_4 – CH=CH_2$, are thus tetrafunctional compounds in free-radical polymerizations, leading first to branched and later to cross-linked polymers. The second vinyl group of a divinylbenzene possesses, however, a different reactivity after polymerization of the first one. The two reactivities may differ so strongly in certain ionic polymerizations of divinylbenzenes that linear polymers, and not cross-linked ones, result (e.g., in the cationic polymerization of 1,4-divinylbenzene by acetyl perchlorate in dichloromethane).

The isocyanate group – N=C=O is a monofunctional group under the conditions of polyaddition reactions (Eq. 28). It may, however, undergo a chain polymerization to polyisocyanates via the N=C double bond

$$n\,R-N=C=O \rightarrow \sim(NR-CO)_n\sim \qquad (29)$$

and it thus reacts as a bifunctional group under these conditions. It may even behave as a half-functional group because urethane groups formed according to Equation (28) may react further to allophanate groups by the addition of isocyanate groups

$$\sim NH-CO-O\sim + O=C=N\sim \longrightarrow \sim NH-CO-\underset{\wr}{N}-CO-O\sim \qquad (30)$$

The functionality *f* of polymer molecules is given by the functionality f_0 of the monomer molecules and the number *X* of mers if no intramolecular rings are formed between functional groups.

$$f = 2(f_o - 1) + (X-2)(f_o - 2) = 2 + X(f_o - 2) \qquad (31)$$

A monomer with functionality f_0 = 3 thus gives a dimer (*X* = 2) with a functionality *f* = 4 and a 100-mer (*X* = 100) with a functionality *f* = 102. Since the molecule needs to be bifunctional only for the formation of linear molecules, the "extra" functionalities lead to interconnections between various molecules (branching) and thus finally to cross-linked, "infinitely large" molecules.

3.2.3. *Cyclopolymerizations*

Polymerizations of multifunctional molecules may proceed not only intermolecularly to produce branched and cross-linked molecules but also intra/intermolecularly with the formation of rings in chains. These polymerizations are called cyclopolymerizations. They occur especially with monomers having carbon – carbon double bonds in the 1,5- or 1,6-position. An example is acrylic anhydride (**11**), $CH_2=CH –CO – O – OC – CH=CH_2$, which polymerizes free-radically even at high monomer concentrations to soluble polymers with 90 – 100 % six-membered (**12**) and five-membered (**13**) ring structures, and only 0 – 10 % "linear" monomeric units **14** (R symbolizes a polymer chain and * an active chain end):

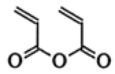

11

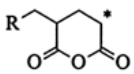

12

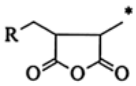

13

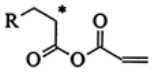

14

3.3. Chain Polymerizations and Polyeliminations

Chain polymerization and polyelimination (condensative chain polymerization) agree in their polymerization kinetics, molar mass distributions, etc., if similar polymerization mechanisms are compared. They differ in the formation of low molar mass byproducts and thus in the design of industrial polymerization reactors.

3.3.1. *Thermodynamics*

Polymerizations occur only if their molar standard Gibbs energies ΔG_p^o are negative. Since the Gibbs energy

$$\Delta G_p^o = \Delta H_p^o - T \cdot \Delta S_p^o = -RT \cdot \ln K \quad (32)$$

is determined by both the polymerization enthalpy ΔH_p^o and the polymerization entropy ΔS_p^o, four different cases are possible:

Case 1. Both enthalpy and entropy are negative; the entropy term $-T \cdot \Delta S_p^o$ becomes more positive with increasing temperature. At a certain temperature, the enthalpy and entropy terms balance each other; no polymerization to high molar mass molecules is possible above this *ceiling temperature*. Oligomers may be formed at the ceiling temperature, however, since polymerization equilibria consist of a series of consecutive equilibria between monomer and growing chains of different degrees of polymerization.

Case 1 is the most common case. It is found for monomers with polymerizable double and triple bonds, such as C=C, C=O, C=S, or C=N. The polymerization entropies of such monomers are practically determined only by the loss of translational entropy; they are thus almost independent of the chemical structure of the monomers. The polymerization enthalpy of α-olefins CH_2=CHR is influenced little by resonance stabilization and is practically independent of the size of the substituent R. Steric hindrance, however, strongly decreases the polymerization enthalpy of 1,1-disubstituted compounds CH_2=CRR′. Styrene and α-methylstyrene, have practically the same polymerization entropies (– 105 J K^{-1} mol^{-1}) but very different polymerization enthalpies (– 71 vs. –35 kJ/mol, respectively) for the polymerization of liquid monomers to condensed polymers. The ceiling temperatures in bulk are 303 °C (styrene) and 60 °C (α-methylstyrene), (i.e., the latter cannot be homopolymerized to high molar mass compounds above 60 °C).

σ-Bonds are opened and formed again in ring-opening polymerization of cyclic monomers. Differences of bond energies in monomers and polymers are thus practically zero, and the polymerization enthalpy is determined by delocalization and strain energies. Polymerization enthalpies are very negative for ring-opening polymerization of three-membered rings, increase to about zero for five- or six-membered rings, drop again to negative values, and return almost to zero for very large rings. The polymerization entropies of very small rings are very negative due to the release of rotational entropy upon polymerization. They become less negative with increasing ring size.

Case 2. The polymerization enthalpy is negative (or zero) and the polymerization entropy positive. Polymerization is possible at all temperatures in this rare case, the polymerization of cyclooctamethyltetrasiloxane being an example.

Case 3. The polymerization enthalpy is positive (or zero) and the polymerization entropy is negative. No polymerization is possible at any temperature. This case seems to apply to the polymerization of acetone.

Case 4. Both polymerization enthalpy and polymerization entropy are positive. A *floor temperature* exists below which no polymerization is possible. This rare case is fulfilled for the polymerization of tightly packed cyclic monomers such as cyclooctasulfur or oxacycloheptane (oxepane) since the number of rotatory degrees of freedom increases strongly upon polymerization to open chains.

Polymerization enthalpies and entropies are influenced by the state of monomers and polymers (gaseous, liquid, crystalline, etc.); the physical interaction between reactants themselves or reactants and solvents; and pressure. Especially important for industrial polymerizations is the fact that most (but not all, see above) chain polymerizations are exothermic. The heat of polymerization can be considerable: adiabatic polymerization of gaseous ethylene to crystalline polyethylene would result in a temperature increase of 1800 K for complete monomer conversion. Polymerization reactors must be designed in such a way that the heat of polymerization can be removed rapidly because, otherwise, inhomogeneous charges may result or the reactor may even explode.

3.3.2. *Free-Radical Polymerizations*

Free-radical polymerizations are initiated and propagated by free radicals. Such polymerizations are relatively insensitive to impurities in the reaction mixture and fairly easy to control. They are thus the methods of choice for the industrial production of plastics if the monomers can be subjected to free-radical polymerizations. This is true for ethylene and many ethylene derivatives: most commodity plastics are thus obtained from free-radical polymerizations (Table 10).

Table 10. Plastics by free-radical chain polymerizations [*]

Monomers	Polymerization in						Monomeric units
	Gas phase	Bulk	Suspension	Emulsion	Solution	Precipitation	
Living polymerization							
p-Xylylene	+						– CH_2 – (*p*-C_6H_4) – CH_2 –
Irreversible polymerization to linear or slightly branched polymers							
Ethylene		+	+	(+)	(+)	(+)	– CH_2 – CH_2 –
Styrene		+	+	(+)	(+)		– CH_2 – CH(C_6H_5) –
p-Methylstyrene			+	+			– CH_2 – CH(*p*-$CH_3C_6H_4$) –
Vinyl acetate		(+)	(+)	+	(+)		– CH_2 – CH(OOCCH$_3$) –
Vinyl chloride	+	(+)	(+)	+	(+)		– CH_2 – CHCl –
Vinyl fluoride		+					– CH_2 – CHF –
Vinylidene fluoride			+	+			– CH_2 – CF_2 –
Trifluorochloroethylene			+				– CF_2 – CFCl–
Tetrafluoroethylene			+				– CF_2 – CF_2 –
Methyl methacrylate		+	+	+	+		– CH_2 – C(CH$_3$) ($COOCH_3$) –
Cross-linking polymerization							
Diallyl phthalate		+					see Equation (32)

[*] + = major processes; (+) = minor processes. Other industrial polymers by free radical homopolymerizations include poly(vinyl acetate) (for coatings and adhesives),

poly(acryl ester) (for adhesives), polyacrylamide (for thickeners), polyacrylonitrile (for fibers), polychloroprene (for elastomers), poly(acrylic acid) (for thickeners), and poly(*N*-vinyl pyrrolidone) (for various applications).

Other industrial polymers formed by free-radical homopolymerizations include poly(vinyl acetate) (for coatings and adhesives), poly(acryl ester) (for adhesives), polyacrylamide (for thickeners), polyacrylonitrile (for fibers), polychloroprene (for elastomers), poly(acryl acid) (for thickeners), and poly(*N*-vinylpyrrolidone) (for various applications).

Polymerizations may be carried out in the gas phase, bulk, suspension, emulsion, solution, or under precipitation. Most monomers are bifunctional; they lead to linear or slightly branched polymers and thus to thermoplastics. The only thermosets obtained from free-radical homopolymerizations are diallyl and triallyl compounds that cross-link under polymerization conditions:

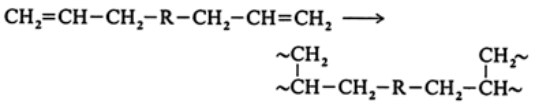

(33)

where R may be $COO(CH_2)_2O(CH_2)_2OOC$ [diethylene glycol bis(allyl carbonate], *o*-$C_6H_4(COO)_2$ (diallyl phthalate), or *m*-$C_6H_4(COO)_2$ (diallyl isophthalate). Other allyl compounds are used as cross-linking comonomers (diallyl fumarate, diallyl maleate, triallyl cyanurate).

Most free-radical polymerizations follow the kinetics of true chain reactions. In initiating (start) reactions (Eq. 35), radicals $R^{\bullet}$ formed from the decomposition (Eq. 34) of initiator molecules I react with monomer molecules M. The resulting "monomer radicals" $R-M^{\bullet}$ add further monomer molecules in the propagation reaction (Eq. 36) until the growing polymer radicals $R{\sim}{\sim}{\sim}M^{\bullet}$ are terminated by recombination (Eq. 37) or disproportionation (Eq. 38) with other polymer radicals or by addition of initiator radicals (Eq. 39):

$$I \rightarrow 2R^{\bullet} \tag{34}$$

$$R^{\bullet} + M \rightarrow R-M^{\bullet} \tag{35}$$

$$R-M^{\bullet} + M \rightarrow R-M-M^{\bullet}, \text{etc.} \rightarrow R-M_{n-1}-M^{\bullet} \tag{36}$$

$$R-M_{n-1}-M^{\bullet} + {}^{\bullet}M-M_{m-1}-R \rightarrow R-M_{n+m}-R \tag{37}$$

$$R-M_{n-1}-M^{\bullet} + {}^{\bullet}M-M_{m-1}-R \rightarrow R-M_{n-2}-M{=}M + X-M_m-R \tag{38}$$

$$R-M_{n-1}-M^{\bullet} + R^{\bullet} \rightarrow R-M_n-R \tag{39}$$

In Equation (38), X is an atom or a group that is transferred from the ultimate monomeric unit of one polymer radical to another polymer radical, for example, the H atom in the reaction

$$\sim CH_2-C^{\bullet}HR + {}^{\bullet}CHR-CH_2\sim \rightarrow \sim CH_2-CH_2R + CHR{=}CH_2\sim$$

These reactions are kinetically classified as termination reactions and not as transfers because they lead to dead chains. The term "transfer" is rather reserved for chemical transfers that generate new propagating radicals, such as

$$\sim CH_2-C^{\bullet}HR + AX \rightarrow \sim CH_2-CHRX + A^{\bullet} \tag{40}$$

Initiation. Free-radical polymerizations can be initiated thermally by thermal initiators, by redox initiators, by photoinitiators, or electrolytically.

Thermal initiation is utilized in the polymerization of *p*- cyclophane to poly(*p*-xylylene):

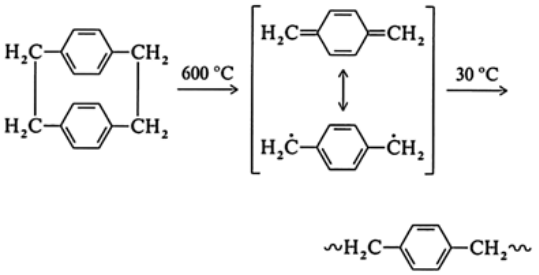

Propagation proceeds by addition of biradicals of various sizes (i.e., by polyaddition and not by chain polymerization). The resulting polymer is cross-linked since radicals also attack methylene groups, which generates more than two radical sites per reactant and thus cross-linking sites. The reaction is utilized for the formation of thin coatings.

Styrene is the only monomer that is polymerized commercially by thermal initiation through monoradicals. The primary Diels – Alder adduct of two styrene molecules reacts with another styrene molecule to give two different starter radicals:

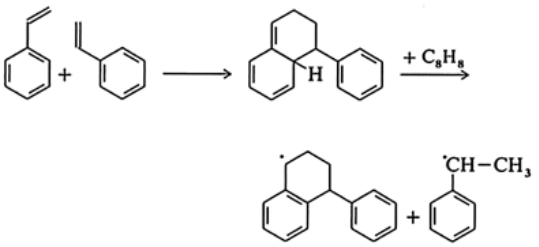

Only relatively small amounts of styrene are, however, polymerized commercially by this method. Most styrene polymerizations, like many other commercial free-radical polymerizations, are started by radicals from *thermal initiators*. These thermal initiators dissociate homolytically into two radicals at elevated temperature, usually 60 – 80 °C. For the bulk polymerization of styrene, high-temperature initiators such as 1,2-dimethyl-1,2-diethyl-1,1-diphenyl-ethane or vinylsilane triacetate are preferred. For styrene polymerization in suspension, dibenzoyl peroxide (BPO) ($C_6H_5CO-O-O-OCC_6H_5$) and *tert*-butyl perbenzoate [$C_6H_5CO-O-O-C(CH_3)_3$] are used.

Many other bulk polymerizations employ diisopropyl peroxydicarbonate (IPP), which decomposes according to

$$(CH_3)_2CHO-COO-OOC-OCH(CH_3)_2 \rightarrow 2(CH_3)_2CHO-COO^{\bullet} \tag{41}$$

Water-soluble initiators such as dipotassium persulfate ($K_2S_2O_8$) and certain redox initiators are used in emulsion polymerizations. *Redox initiators* generate radicals by reaction of a reducing agent with an oxidizing agent (e.g., $Fe^{2+} + H_2O_2$). These systems need far less activation energy for the formation of radicals, and polymerization can thus be initiated at low temperature.

Photochemical initiations are utilized for the production of lithographic plates (⟶ Imaging Technology) and for hardening lacquers but not in the manufacture of plastics. *Electrolytic polymerizations* find applications in the coating of metal sheets by plastics.

Only initiator radicals are formed at the early stages of free-radical polymerizations. These radicals are successively converted into monomer radicals by addition of initiator radicals to monomer molecules and into polymer radicals by further addition of monomers. At the same time, some polymer radicals are removed by termination reactions.

Finally, a steady state is established in which as many radicals are formed as are removed by termination. This state of constant total radical concentration (ca. 10^{-8} mol/L) is in general observed within seconds at monomer conversions of 10^{-2} – 10^{-4} %.

Propagation. Monomer molecules are usually added head-to-tail to growing polymer radicals:

$$\sim CH_2-CHR^{\bullet} + CH_2=CHR \rightarrow \sim CH_2-CHR-CH_2-CHR^{\bullet} \quad (42)$$

but tail-to-tail additions do occur (see Section Constitution).

Free-radical polymerizations lead in general to atactic or predominantly syndiotactic polymers. The syndiotacticity increases with decreasing polymerization temperature.

Termination. Termination reactions by recombination, disproportionation, and (at higher initiator concentrations) addition of initiator radicals, cannot be avoided. Since initiator radicals and thus also polymer radicals are formed successively and are deactivated at random by different termination reactions, distributions of molar masses are formed. The width of these distributions is given by the relative proportion of the different types of termination reactions.

Chain Transfer. The transfers of radicals to monomers, solvents, and initiators terminate polymer chains and generate others (Eq. 40). The ratios of rate constants k_{tr} of transfer to rate constants k_p of propagation are called transfer constants, $C_{tr} = k_{tr}/k_p$; at 60 °C they possess values between ca. 2×10^{-6} for the transfer of polystyryl radicals to benzene and 5700 for the transfer of poly(vinyl acetate) radicals to carbon tetrabromide.

Chain transfers often lead to new radicals that have approximately the same reactivity as the disappearing ones. In this case, the rate of polymerization is not decreased by chain transfer, but the degree of polymerization is lowered. Certain transfer agents with high transfer constants (such as thiols) are thus often employed to regulate molar mass.

Retardation occurs if the new radical is more sluggish in adding monomers than the transferring polymer radical; *inhibition*, if a new chain does not start at all. Chain transfers to polymers lead to branched polymers; most radically polymerized polymers are thus slightly branched.

The chain transfer to nonionic allyl monomers generates resonance-stabilized allyl radicals. The addition of monomer molecules to these radicals produces radicals

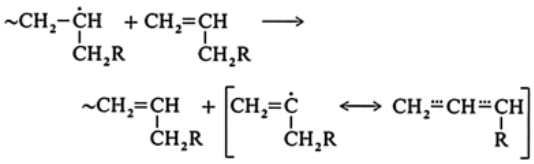

(43)

that are more stabilized by resonance than their precursors and thus do not start a polymer chain. The polymerization of noncharged monoallyl monomers thus leads to oligomers with degrees of polymerization of ca. 10 – 20.

Polymerization Kinetics. Initiator radicals are generated by initiator decomposition (Eq. 34) with a rate of $2 f \cdot k_d \cdot [I]$ where f is the radical yield (i.e., the fraction of radicals that start the polymerization). Initiator radicals disappear in the start reaction with a rate of $k_{st}[I^{\bullet}][M]$. In the steady state, the rate of the start reaction equals the rate of the termination reaction(s), thus $R_{st} = R_t$, and for terminations through deactivation by other polymer radicals,

$$d[P]/dt = R_{st} - R_t = 2 f \cdot k_d \cdot [I] - k_t[P]^2 = 0 \quad (44)$$

Since monomer is practically consumed only by the propagation reaction with a rate of $R_p = -d[M]/dt = k_p[P^{\bullet}][M]$, the propagation rate at diminishing initiator consumption is obtained from Equation (44)

$$R_p = -d[M]/dt = k_p(2 f \cdot k_d/k_t)^{1/2} \cdot [I]_o^{1/2}[M] \quad (45)$$

The polymerization rate is directly proportional to the monomer concentration; bulk polymerization is thus always faster than solution polymerization. The monomer concentration decreases with increasing monomer conversion as does the rate R_p.

A “self-acceleration” of the polymerization and a concomitant increase in the degree of polymerization are observed at higher monomer conversion, especially in bulk and in concentrated solutions. This *gel effect* or *Trommsdorff – Norrish* effect is caused by a changing diffusion control of the termination by mutual reaction of two polymer radicals (i.e., a decrease of termination reactions), which in turn comes from an increasing entanglement of polymer chains. Because neither the rate of the initiator decomposition nor the rate of monomer addition is affected by increasing entanglements, both overall polymerization rates and molar masses increase. The gel effect broadens molar mass distributions.

Cross-Linking Polymerizations. Free-radical chain polymerizations are employed commercially in the manufacture of diallyl thermosets and in styrene – divinylbenzene copolymerizations for polymer beads. Primary macromolecules produced at very low monomer conversions are almost linear and carry pendant polymerizable groups. The likelihood of an attack at these groups increases with increasing mass average molar mass. Branched molecules are formed that convert to cross-linked polymers at the gel point. Monomer conversion at the gel point cannot at present be predicted theoretically for such free-radical polymerizations.

Process Engineering. Free-radical polymerizations may be performed in the gas phase, in bulk, solution, emulsion, suspension, or under precipitation (Table 10). Each method has advantages and disadvantages.

Bulk Polymerizations. Industrial bulk polymerizations are performed in *liquid monomers*, occasionally with the addition of 5 – 15 vol % solvent as a polymerization aid. The process leads to very pure products. However, much heat is generated per unit volume and, in addition, by the gel effect. Local overheating may cause multimodal molar mass distributions, branching, polymer degradation, discoloring, or explosion. Bulk polymerizations are thus often terminated at 40 – 60 % monomer conversion. The remaining monomer is distilled and recovered. Alternatively, polymerizations may be performed in two steps: first in large batch reactors, then in thin layers. Residual monomer is usually removed by steam.

Suspension polymerization is a “water- cooled” bulk polymerization. Water-insoluble monomers are suspended as small droplets of 0.001 – 1- cm diameter with the help of suspending agents. The polymerization is initiated by oil-soluble initiators. The droplets are converted by polymerization into pearls, and the process is also called *pearl polymerization*. Suspension polymerization allows easy control of the reaction and delivers the polymer as easy-to-handle beads. A disadvantage is the cost of water removal and cleaning, along with the sometimes deleterious effects of incorporated suspension agents.

Emulsion Polymerization. In emulsion polymerizations, water-insoluble monomers are solubilized in water by micelles of surfactant molecules (4 – 10-nm diameter); the monomers are also suspended in a few droplets (ca. 1000-nm diameter). Water-soluble initiators decompose into radicals that travel through the water to the micelles where polymerization begins. Diffusion of monomer from the droplets to the micelles replenishes the monomer that has been used up by reaction in the micelles. The micelles finally grow into latex particles of 500 – 5000-nm diameter. Since the reaction volume is very small in micelles (and in growing latex particles), termination reactions are rare and polymerization proceeds in a quasi-living manner, leading to much higher molar masses than bulk polymerization. In terms of reaction control, emulsion polymerizations have advantages similar to suspension polymerizations. In addition, they offer higher polymerization rates by redox initiators and higher molar masses (which can be regulated by transfer agents). The resulting latex can furthermore be used directly for paints, adhesives, and coatings.

Gas-Phase Polymerization. Gas-phase polymerizations can be initiated photochemically. Each growing particle contains only one radical and the polymerization is thus quasi-living. The polymerization rate is determined by the rate of monomer adsorption in the precipitating particles and later by monomer diffusion to the occluded polymer radicals. The resulting polymers are very clean.

Solution Polymerization. Solution polymerizations decrease polymerization rates by monomer dilution and allow easy removal of the heat of polymerization. Solvents are however costly to remove, and the process is used only if the resulting polymer solutions can be used directly or if monomers and polymers decompose in melts.

3.3.3. *Anionic Polymerizations*

Anionic polymerizations are initiated by bases or Lewis bases such as alkali metals, alkoxides, amines, phosphines, Grignard compounds, and sodium naphthalene, very often

in solution. The initiators dissociate “instantaneously” into the initiating species. These dissociations and the subsequent start reactions need little activation energy; anionic polymerizations thus often proceed with high speed even at –100 °C.

The initiating species and the growing macroions are rarely completely dissociated into anions and counterions. They are rather in equi-librium with various types of ion pairs (nondissociated species $\rightleftarrows$ contact ion pair $\rightleftarrows$ solvent-separated ion pair $\rightleftarrows$ free ions; plus associates of these species). The type and amount of initiating species strongly affect the polymerization rates and the tacticity of polymers.

The initiating species are present in their effective concentrations at the beginning of polymerization; they are not formed one after another as in free-radical polymerization. Termination reactions are fairly rare; anionic polymerizations are often “living.” Very low initiator concentrations are required for living polymerizations to high molar masses (see Section Classifications); however, high monomer – initiator ratios are difficult to control and may lead to low polymerization rates. In general, polymers from living polymerizations have very narrow molar mass distributions of the Poisson type if diffusion effects are avoided during the mixing of monomer and initiator solutions.

Very few anionic polymerizations are therefore employed for the production of thermoplastics, most notably those of formaldehyde, ∈- caprolactam (also for reaction injection molding), and laurolactam (Table 11). Anionic polymerizations are however the method of choice for the syntheses of block copolymers (e.g., thermoplastic elastomers of the styrene – butadiene – styrene type).

Table 11. Industrial anionic homopolymerizations

Monomer	Repeating unit	Application
Butadiene	$\sim CH_2 – CH=CH – CH_2\sim$	elastomer (1,4-*cis*)
Isoprene	$\sim CH_2 – C(CH_3)=CH – CH_2\sim$	elastomer (1,4-*cis*)
Methyl cyanoacrylate	$\sim CH_2C(CN)\ (COOCH_3)\sim$	adhesive
Formaldehyde	$\sim O – CH_2\sim$	engineering plastics
Ethylene oxide	$\sim O – CH_2 – CH_2\sim$	thickener
Glycolide	$\sim O – CO – CH_2\sim$	surgical threads
∈-Caprolactone	$\sim O – CO – (CH_2)_5\sim$	polymer plasticizer
∈-Caprolactam	$\sim NH – CO – (CH_2)_5\sim$	fiber, thermoplastic
Lauryllactam	$\sim NH – CO – (CH_2)_{11}\sim$	fiber, film
Hexamethylcyclotrisiloxane	$\sim O – Si(CH_3)_2\sim$	elastomer

3.3.4. *Cationic Polymerizations*

Three groups of monomers can be polymerized cationically: olefin derivatives CH_2=CHR with electron-rich substituents, monomers with double bonds containing heteroatoms or heterogroups Z (e.g., CH_2=Z), and rings with heteroatoms. Initiators are Brønsted acids (perchloric acid, trichloroacetic acid, trifluoromethanesulfonic acid, etc.); Lewis acids ($AlCl_3$, $TiCl_4$, etc.) with “coinitiators” such as water; and carbenium salts (acetyl perchlorate, tropylium hexachloroantimonate, etc.). The propagating macrocations are thermodynamically and kinetically unstable. They attempt to stabilize themselves by addition of nucleophilic species, which leads to very fast polymerization on one hand and to a host of transfer and termination reactions on the other. Very few monomers are thus polymerized cationically on an industrial scale (Table 12).

Table 12. Industrial cationic polymerizations

Monomer		Application [**]
Name	Structure	
Isobutene	$CH_2=C(CH_3)_2$	elastomers, adhesives, VI improvers
Vinyl ethers	CH_2=CHOR	adhesives, textile aids, plasticizers
Formaldehyde	CH_2=O	engineering plastic
Ethyleneimine	c-($NHCH_2CH_2$)	paper additive, flocculant
Tetrahydrofuran	c-[$O(CH_2)_4$]	soft segment for polyurethanes or polyether – ester elastomers

[*] c=cyclo.
[**] VI=viscosity index.

3.3.5. *Ziegler – Natta Polymerizations*

Ziegler – Natta polymerizations are chain polymerizations that are initiated and propagated by Ziegler catalysts, a vast group of coordination compounds from transition-metal compounds and metal alkyls. These polymerizations are also called coordination polymerizations or anionic – coordination polymerizations, although coordination of a monomer to an initiator need not lead to a Ziegler – Natta type of polymerization and no anions are involved at all.

Typical Ziegler catalysts for industrial polymerizations are shown in Table 13. The so- called third-generation catalysts for ethylene polymerizations are composed of $TiCl_3 + AlR_3 + MgCl_2$; their high catalyst yield of more than 50 000 g of polyethylene produced per 1 g of $TiCl_3$ makes removal of catalyst residues from the polymer unnecessary.

Table 13. Important industrial polymerizations with transition-metal catalysts

Monomer	Initiator	Application
Ethylene	$TiCl_3 – AlR_3$; Cr – silica	thermoplastic
Propene	$TiCl_3 – R_2AlCl$	thermoplastic
Ethylene+propene+nonconjugated diene	$VOCl_3 – R_2AlCl$	elastomer
Butadiene	$TiI_4 – AlR_3$; $R_3Al_2Cl_3 – Co(OOCR')_2$	elastomer
Isoprene	$TiCl_3 – AlR_3$	elastomer
Butene-1	$TiCl_3 – Et_2AlCl$	thermoplastic
4-Methylpentene-1	?	thermoplastic

Various complexes between transition-metal compounds and metal alkyls are known to be formed, but which of these are polymerization active and to what extent are rarely known. In any case, the transition-metal atoms of these complexes are bound to the growing chains; Ziegler catalysts are thus not catalysts but initiators.

Propagation proceeds via insertion of the monomer molecule between the already existing chain and the bound transition metal Mt:

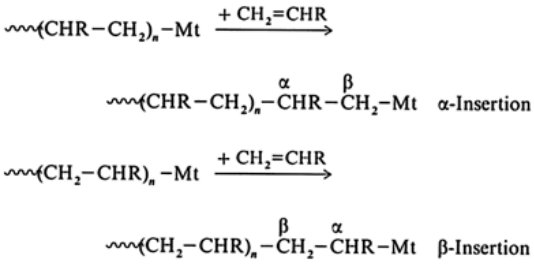

The isospecific polymerization of propene with $TiCl_4$ – AlR_3 is an α-insertion; the syndiospecific polymerization of the same monomer is very probably a β-insertion. The ability of Ziegler catalysts to regulate the stereocontrol of such polymerizations to a very high degree makes Ziegler – Natta polymerizations extremely useful industrially. Propene, for example, polymerizes free-radically or cationically to highly branched, atactic, low molar mass polymers that do not crystallize and have low glass transition temperatures. They can be used only for hotmelt adhesives. Ziegler – Natta polymerizations (Table 13) lead, however, to stereoregular polymers with high melting temperatures; isotactic polypropylenes find extensive use as thermoplastics.

Similar catalysts are used in *metathesis polymerizations*. Metatheses are exchange and disproportionation reactions of double bonds, mainly carbon – carbon double bonds in olefins and cycloolefins. A low molar mass example is the metathesis of pentene-2 to butene-2, pentene-2, and hexene-3 (in the ratio 1 : 2 : 1) by the catalyst system WCl_6 – $C_2H_5AlCl_2$ – C_2H_5OH. Industrially, cyclooctene, norbornene, and dicyclopentadiene are polymerized by metathesis polymerization. Cyclooctene gives polyoctenamer ~ CH=CH–$(CH_2)_6$ ~; both the *cis* and the *trans* isomers are elastomers (⟶ Rubber, 3. Synthetic). Norbornene polymerizes to a thermoplastic polymer

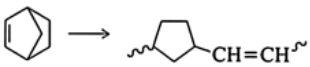

that is plasticized with mineral oil to give an elastomer (⟶ Rubber, 3. Synthetic). Dicyclopentadiene is considered as a monomer for reaction injection molding (RIM) processes.

Group transfer polymerizations are also insertion polymerizations because a monomer molecule is inserted into the bond between the active center and the remaining polymer chain. Depending on the monomer, these polymerizations may proceed by either α- or β-insertions and with the help of nucleophilic or electrophilic catalysts. Only a small-scale polymerization of alkyl methacrylates to low molar mass polymers by silylketene acetals and nucleophilic catalysts ($[(CH_3)_3SiF_2]^-$, $[HF_2]^-$, CN^-, etc.) is utilized industrially at present:

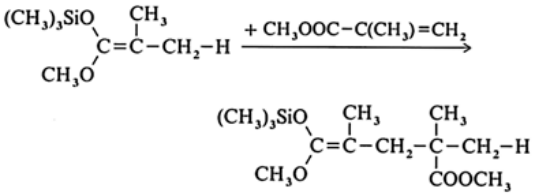

3.3.6. *Copolymerizations*

Copolymerizations are joint polymerizations of two or more monomers; they were also called interpolymerizations in the older literature. In the simplest case, both active chain ends ~a* and ~b* react irreversibly with the two monomers A and B:

$$\sim a^* + A \rightarrow \sim a - a^*;\ R_{aA} = k_{aA}[a^*][A]$$

$$\sim a^* + B \rightarrow \sim a - b^*;\ R_{aB} = k_{aB}[a^*][B]$$

$$\sim b^* + A \rightarrow \sim b - a^*;\ R_{bA} = k_{bA}[b^*][A]$$

$$\sim b^* + B \rightarrow \sim b - b^*;\ R_{bB} = k_{bB}[b^*][B] \qquad (46)$$

Four rates R_{ij} and four rate constants k_{ij} are to be considered in this terminal model, which corresponds to Markov first-order statistics. At high molar masses, monomers are consumed only by the four propagation reactions. The relative monomer conversion is given by

$$\frac{-d[A]/dt}{-d[B]/dt} = \frac{R_{aA} + R_{bA}}{R_{bB} + R_{aB}} = \left(\frac{k_{bA} + k_{aA}\left(\left[a^*\right]/\left[b^*\right]\right)}{k_{bB} + k_{aB}\left(\left[a^*\right]/\left[b^*\right]\right)}\right) \cdot \frac{[A]}{[B]} = \frac{d[A]}{d[B]} \qquad (47)$$

The ratios of rate constants of homopropagations and cross-propagations are called copolymerization parameters

$$r_A \equiv k_{aA}/k_{aB};\ r_B \equiv k_{bB}/k_{bA} \qquad (48)$$

Five different cases can be distinguished for each copolymerization parameter:

- *r = 0.* The rate constant of homopropagation is zero; the active center adds only the other monomer.
- *r < 1.* The other monomer is added preferentially.
- *r = 1.* Both monomers are added in the same amounts.
- *r > 1.* The own monomer is added preferentially but not exclusively.
- *r = ∞.* No copolymerization, only homopolymerization.

Depending on the relative numerical value of the two copolymerization parameters, five different types of copolymerization can be distinguished (Table 14).

Table 14. Types of copolymerization of monomers A and B

Types	Copolymerization parameters				
	Azeotropic		Nonazeotropic		
	r_A	r_B	r_A	r_B	$r_A r_B$
Alternating	0	0	0	> 0	0
Statistical	< 1	< 1	< $1/r_B$	> 1	< 1
Ideal	1	1	$1/r_B$	$1/r_A$	1
Block forming	> 1	> 1	> $1/r_B$	< 1	> 1
Blend forming	∞	∞	∞	< ∞	∞

In general, one of the monomers is consumed preferentially during copolymerization ($r_A \neq r_B$). This drift of copolymer composition can be avoided if the more reactive monomer is fed into the reactor according to its consumption. A conversion-independent polymer composition also results if copolymerization is performed under azeotropic conditions, which are defined by

$$d[A]/d[B] \equiv [A]/[B] \tag{49}$$

i.e., the relative rate of monomer consumption equals the ratio of monomer concentrations at any time: The monomer and the polymer compositions do not drift with the progress of polymerization.

This equation is a special case of the Lewis – Mayo equation, which describes the instantaneous composition of the copolymer as a function of the ratio of instantaneous monomer concentrations:

$$\frac{d[A]}{d[B]} = \frac{1 + r_A([A]/[B])}{1 + r_B([B]/[A])} = \frac{x_a}{x_b} \tag{50}$$

All types of copolymerization can in principle occur with the same monomer pair if various initiators are used. All copolymerizations of styrene and methyl methacrylate shown in Figure 15 are, for example, statistical. The free-radical-initiated copolymerization is, however, azeotropic, and the cationic and anionic ones are nonazeotropic. Initiation by $Et_3Al_2Cl_3$ and traces of oxygen leads to almost alternating copolymers, whereas cationic polymerization generates long styrene sequences and anionic polymerization produces long methyl methacrylate ones. The sequence length determines many polymer properties, most notably glass and melting temperatures.

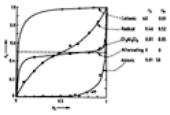

Figure 15. Instantaneous mole fraction x_s of styrene units in the polymer formed as a function of the instantaneous mole fraction x_s of styrene in the copolymerization of styrene with methyl methacrylate initiated by (o) cations; (•) free radicals; (⊕) $Et_3Al_2Cl_3$; (⊖) Anions

Solid lines: calculated with copolymerization parameters shown to the right; dotted line: theory for alternating copolymerizations.

Ionic copolymerizations very often lead to long block sequences, whereas free-radical copolymerizations tend in general to produce more random copolymers. Free-radical copolymerizations are thus the method of choice for industrial purposes if allowed by monomer structures (Table 15).

Table 15. Industrial copolymerizations

Monomers	Polymerization	Application
Free radical, linear (or slightly branched)		
Ethylene+10 % vinyl acetate	bulk	shrink films
Ethylene+10 – 35 % vinyl acetate	bulk	thermoplastics
Ethylene+35 – 40 % vinyl acetate	under precipitation [a]	films
Ethylene+>60 % vinyl acetate	emulsion	elastomers
Ethylene+<10 % methacrylic acid	bulk	extrusion coating
Ethylene+trifluorochloroethylene		thermoplastics
Butadiene+styrene	emulsion	multipurpose elastomers
Butadiene+37 % acrylonitrile	emulsion	oil-resistant elastomers
Vinyl chloride+3 – 20 % vinyl acetate	solution [b]	paints
Vinyl chloride+15 % vinyl acetate	solution [b]	Hi-fi records
Vinyl chloride+3 – 10 % propene	bulk	thermoplastics
Acryl esters+5 – 15 % acrylonitrile		oil-resistant elastomers
Acrylonitrile+4 % different monomers	under precipitation [c]	fibers with improved dyeability
Acrylonitrile+styrene		thermoplastics
Acrylonitrile+butadiene+styrene		thermoplastics
Tetrafluoroethylene+propene	emulsion	thermoplastics
Methacrylic acid+methacrylonitrile		hard foams (after cyclization)
Free radical, cross-linking		
Glycol methacrylate+2 – 4 % glycol dimethacrylate		contact lenses
Unsaturated polyesters+styrene or methyl methacrylate	bulk	glass-fiber-reinforced thermosets
Anionic		
Styrene+butadiene	solution	elastomers
Cationic		
Isobutene+2 % isoprene	solution	butyl rubber
Trioxane+ethylene oxide	solution	thermoplastics
Ethylene oxide+propylene oxide	solution	thickener, detergents
Ziegler – Natta polymerization		
Propylene oxide+nonconjugated dienes	solution	elastomers
Ethylene+propene+nonconjugated dienes		elastomers

[a] In *tert*-butanol.
[b] In acetone, 1,4-dioxane or hexane.
[c] In water.

3.4. Polycondensations and Polyadditions

3.4.1. *Bifunctional Polycondensations*

Polycondensations and polyadditions are generally divided into bifunctional reactions (two functional groups per reactant) and multifunctional ones (three or more functional groups per monomer molecule). Bifunctional reactions lead to linear macromolecules; multifunctional ones, to branched and finally cross-linked polymers. Bifunctional reactions are further subdivided into AB reactions (e.g., that of H_2N – R – COOH) and AA/BB reactions (e.g., H_2N – R – NH_2 + HOOC – R′ – COOH).

All reactants, from monomer to high polymer, may react to form macromolecules in bifunctional polycondensations and polyadditions. Only monomers are present at the beginning of the reaction; later, oligomers are present. If the reactivity of the functional groups is independent of molecule size, then the greater number of smaller molecules leads first to low molar mass polymers and only at high conversions to chemical compounds with high molar masses (see Fig. 14). In equilibrium, the number-average degree of polymerization of polymers from AB and stoichiometric AA – BB reactions is dictated by the extent of functional group reaction p:

$$\bar{X}_n = 1/(1 - p) \tag{51}$$

At 50 % conversion (p = 0.5), the number-average degree of polymerization of reactants is only 2 (this includes monomer molecules), and the number-average degree of polymerization of polymers ($2 < X_i < \infty$) is only 3. Industrial polycondensations rarely deliver number-average degrees of polymerization greater than 100 – 200 (i.e., p = 0.99 – 0.995 in equilibrium). The endgroups of these equilibrium polymers may condense further during processing by, e.g., extrusion or injection molding. The accompanying strong increase in viscosity is undesirable and these endgroups are therefore blocked against further polycondensation by “sealing” them with monofunctional reagents.

Equilibria are established not only by retroreactions (reversal of polycondensations and polyadditions) but also by catalyzed exchange reactions between chain segments. In these trans reactions, number-average molar masses remain constant, but mass-average molar masses increase if the initial ratio of mass to number-average molar mass was smaller than dictated by equilibrium conditions. An example is the transesterification of poly(ethylene terephthalate):

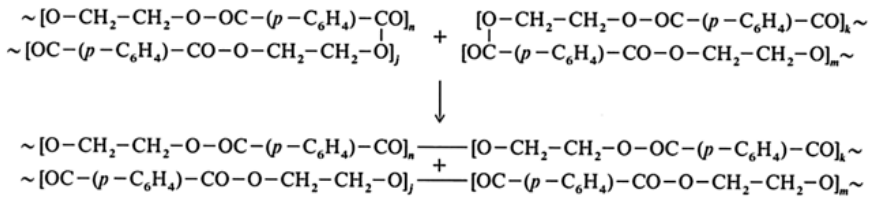

Higher degrees of polymerization at lower extents of reaction can be obtained by irreversible “activated reactions” and especially by heterogeneous reactions of fast reacting monomers. An example is the Schotten – Baumann reaction of diamines in water with diacid dichlorides in chloroform, for example. A polyamide film is formed at the interface of the two solutions. Functional groups of this polymer are buried in the interior of the film; they cannot react with each other. Exchange reactions and retroreactions are also absent. Only surface groups can react; the resulting irreversible reaction leads to high molar masses. Industrially, this reaction is utilized for one of the two industrial syntheses of polycarbonates, i.e., from the sodium salt of bisphenol A and phosgene (Table 16).

Table 16. Industrially important linear AA – BB polycondensations following Equation (52)

Polymer	Chemical structure				Remarks
	A	V	W	B	
Poly(ethylene	H	$O(CH_2)_iO$	CO – *p*-Ph – CO	OH	*i*=2 (ethylene) or
terephthalate)	H	$O(CH_2)_iO$	CO – *p*-Ph – CO	OCH_3	4 (butene)
Unsaturated polyesters	H	$O(CH_2)_2O$	CO – CH=CH – CO	$O_{1/2}$	from maleic anhydride (isomerizes mainly to fumaric acid residues)
Polycarbonate	H	O – *p*-Ph – $C(CH_3)_2$ – *p*-Ph – O	CO	OC_6H_5	
	Na	O – *p*-Ph – $C(CH_3)_2$ – *p*-Ph – O	CO	Cl	
Polyarylates	Na	O – Ar – O	CO – Ar – CO	Cl	different Ar
Poly(phenylene sulfide)	Na	S	*p*-Ph	Cl	
Polysulfides	Na	S	R	Cl	different R
Polysulfones	K	O – *p*-Ph – O	SO_2 – *p*-Ph – O – *p*-Ph – SO_2	Cl	different V and W units
Polyamides	H	$NH(CH_2)_iNH$	$CO(CH_2)_jCO$	OH	PA 66 (*i*=6; *j*=4) PA 610 (*i*=6; *j*=8) PA 46 (*i*=4; *j*=4)
Polyetheretherketone	K	O – *p*-Ph – CO – *p*-Ph – O	*p*-Ph – CO – *p*-Ph	F	

Ar=aromatic residue,
p-Ph=*para*-substituted phenylene group,
R=organic residue.

Most industrial polycondensations are of the AA – BB type (Table 16); they can be described by

$$nA-V-A+nB-W-B \rightarrow A-(V-W)_n-B+(2n-1)AB \quad (52)$$

where A– and B– are leaving groups, AB represents leaving molecules, –V– and –W– are monomeric units, and –V–W– are repeating units.

Industrial AB polycondensations are fairly rare. Examples are the formation of polyamide 11 (nylon 11) (Eq. 27), the polycondensation of carbodiimides

$$O=C=N-R-N=C=O \rightarrow \sim R-N=C=N\sim + CO_2 \quad (53)$$

and the synthesis of certain polysulfones

$$C_6H_5O(p\text{-}C_6H_4)SO_2Cl \rightarrow \sim(p\text{-}C_6H_4)-O-(p\text{-}C_6H_4)-SO_2\sim + HCl \quad (54)$$

3.4.2. *Multifunctional Polycondensations*

Multifunctional polycondensations are condensation reactions with the participation of monomers having higher functionality (f_o>2). At low conversions, branched molecules are formed with increasingly higher number of functional end groups per molecule (i.e., increased functionality *f*) :

$$f=2(f_o-1)+(X-2)(f_o-2)=2+X(f_o-2) \quad (55)$$

Only one such group is however needed to link two molecules; the probability of such a linkage thus increases dramatically with increasing degree of polymerization *X*. At a certain degree of conversion of functional groups, molecules are linked throughout the volume of the reactor. The viscosity increases strongly, and the mass-average molar mass of the cross-linked polymer approaches infinity at this “gel point.” Not all reactants are linked at the gel point; some are still soluble up to high conversions. The number-average degrees of polymerization of reactants are thus fairly low at the gel point, usually in the range 10<$\bar{X}_n$<50.

Knowledge of the gel point is technically very important. Polymers cannot, or can only with difficulty, be processed beyond the gel point and must thus be shaped before or during the cross-linking reaction. The gel point can be calculated from the functionality and relative amounts of reactive groups if intramolecular cyclizations are assumed absent. In practice, such cyclizations are often present and cross-linking occurs at higher extents of reaction than calculated. The theoretical calculations thus provide a safety margin.

Chemical cross-linking reactions are utilized for thermosets, certain leathers, and the preparation of elastomers from rubber. Thermosetting polycondensation reactions allow the hardening of phenol, amino, and alkyd resins. In the reaction of phenol(s) with formaldehyde, methylene, formal, and ether bridges are formed between the *ortho* and *para* positions of the phenols; the hardening of novolacs with hexamethylenetetramine (urotropin) also leads to imine structures (⟶ Phenolic Resins). Amino plastics result from the reaction of urea or melamine with formaldehyde (⟶ Amino Resins), and alkyd resins from multifunctional alcohols with bifunctional organic acids or their anhydrides (⟶ Alkyd Resins).

Cross-linking also occurs inadvertently during cyclopolycondensations. In such reactions, rings are formed during the polycondensation. An example is the formation of polyimides from pyromellitic acid dianhydride with aromatic diamines.

Step I is a polyaddition to a polyamic acid. Step II may occur as shown but also by intermolecular reaction. To prevent such cross-linking reactions, step I is performed at 10 – 15 % solids content and conversion is limited to 50 %. Polyimide films are formed from the cross-linking of polyamic films. Polybenzimidazoles and ladder polymers are similarly obtained through such cyclopolycondensations.

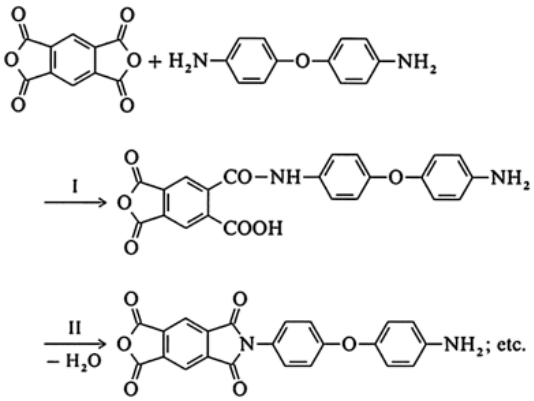

3.4.3. *Polyaddition*

No leaving molecules have to be removed during polyadditions, an important advantage for process engineering. Practically all industrially important polyadditions are thus performed with thermosetting resins. Polyurethanes are formed from diisocyanates (or triisocyanates) and polyols (see Eq. 28 for diols). Polyurethane formation may be so rapid that cross-linking and shaping can be performed in one step through reaction injection molding with short cycle times. Epoxides (epoxy resins) contain two or more epoxy groups per prepolymer molecule and are cured with either difunctional or multifunctional amines or with anhydrides of organic acids.

3.5. Polymer Reactions

3.5.1. *Polymer Transformations*

Several polymers are chemically transformed into industrially useful polymers by polymeranalogous reactions of their substituents (Table 17).

Table 17. Industrial polymer transformations (for ring formations see text)

Base units in primary polymer	Reagent	Conversion, %	Base units in final polymer	Application
$CH_2CH(OOCCH_3)$	ROH	98	$CH_2 – CHOH$	thickener
$CH_2CH_2CH_2CH(OOCCH_3)$	CH_3OH	99	$CH_2 – CHOH/CH_2 – CH_2$	engineering plastics
Cellulose $– NHCOCH_3$	H_2O	?	Cellulose NH_2	paper additive
Cellulose OH	CH_3COOH	83 – 100	Cellulose $OOCCH_3$	fibers, films
Cellulose OH	HNO_3	67 – 97	Cellulose $– ONO_2$	plastics, fibers
Cellulose ONa	oxirane	53 – 87	Cellulose $– OCH_2CHRCH_3$	thickener
Cellulose OH	oxirane	100 – 400 [a]	Cellulose $– OCH_2CH(OR')CH_3$	thickener
$CH_2 – CH_2$	Cl_2	25 – 40 [b]	$CH_2 – CHCl$	elastomer
		>40 [b]	$CH_2 – CHCl/CHCl – CHCl$	impact improver for PVC
$CH_2 – CHCl$	Cl_2	64 [b]	$CH_2 – CHCl/CHCl – CHCl$	adhesives, paints
$CH_2 – CH_2$	Cl_2/SO_2	<42 [b]; <2 [c]		coatings
$CH_2CH=CHCH_2/CH_2 – CHCN$	H_2	?	$(CH_2)_4/CH_2 – CHCN$	elastomer
$CH_2 – CH(C_6H_5)$	SO_3	low	$CH_2 – CH(C_6H_4SO_3H)$	ion exchangers
$N=PCl_2$	RONa	?	$N=P(OR)_2$	elastomer

[a] Further reaction of primary reaction products may lead to more than one monomeric unit of reagent per hydroxyl group of cellulose.
[b] Chlorine content in %.
[c] Ether groups per glucose unit.

These reactions per group are chemically identical to their low molar mass analogues. They differ in the effect of side reactions, which do not lead to removable byproducts as in micromolecules but to “wrong” structures in the polymer chain. In most cases, such structures decrease the desired polymer properties. Polymer transformations that are prone to side reactions are thus generally avoided. This is one reason why only five types of polymer-analogous reactions are utilized commercially: transesterification – saponification, chlorination – sulfochlorination, etherification, hydrogenation, and cyclization.

The reaction of poly(vinyl alcohol) with butyraldehyde leads to acetals. For statistical reasons, however, not all hydroxyl groups can be transformed:

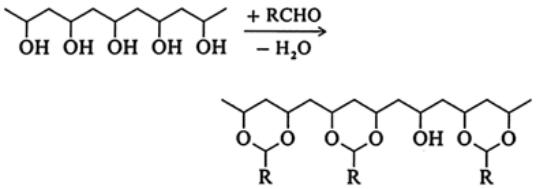

The resulting poly(vinyl butyral) is used as an intermediate layer in safety glass and for wash primers.

Heating of methacrylic acid – methacrylonitrile copolymers with ammonia produces polymethacrylimide structures that are hard foams:

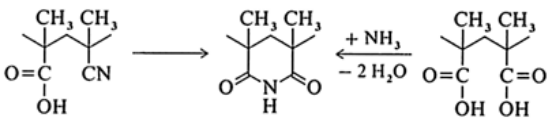

3.5.2. *Block and Graft Formations*

Reactive chain ends can either initiate the polymerization of other monomers or couple with other preformed macromolecules to give block polymers:

$$A_n{}^* + mB \rightarrow A_nB_m{}^* \quad (56)$$

$$A_n{}^* + {}^*B_m \rightarrow A_nB_m \quad (57)$$

Both methods are used industrially with a variety of strategies for the synthesis of diblock polymers A_nB_m as compatibilizing agents for polymer blends, triblock polymers $A_nB_mA_n$ as thermoplastic elastomers, and various multiblock polymers for different purposes, especially for thermoplastic elastomers. Anionic living polymerizations are used for the syntheses of poly(styrene – block-butadiene – block-styrene) and poly(styrene – block-isoprene – block-styrene). Polycondensations serve for the preparation of multi[{poly

(butylene terephthalate)} – block-poly{(polytetrahydrofuran) – terephthalate}], and various polyesteramide and polyetheramide segmented block polymers.

Graft copolymers result from the grafting of monomers on chemically different polymer chains. Grafts of isobutene – isoprene on polyethylenes, vinyl chloride on poly(ethylene-stat-vinyl acetate), ethylene – propylene on poly(vinyl chloride), or styrene – acrylonitrile on saturated acryl rubber also yield thermoplastic elastomers.

3.5.3. *Cross-Linking Reactions*

Cross-linking of preformed polymers is used for vulcanizing unsaturated rubbers with sulfur, vulcanizing saturated rubbers with peroxides, hardening unsaturated polyester chains with styrene or methyl methacrylate via copolymerization, or tanning collagen with certain reagents to leather.

3.5.4. *Degradation Reactions*

Degradation of polymers is undesirable for their application and often desirable for their disposal after use. The term degradation includes the decrease in degree of polymerization with preservation of the chemical structure of monomeric units (depolymerization), the unwanted chemical transformation of some monomeric units with preservation of the degree of polymerization (decomposition), and a combination of both. The terms degradation, depolymerization, and decomposition are sometimes used with the same meaning as in this article, but other times with broader or more restricted meanings; no accepted usage exists.

Depolymerizations involve chain scissions. They are retroreactions (i.e., the reverse of chain polymerizations, polyeliminations, polycondensations, or polyadditions). Chain scissions may be at random as in retropolycondensations and retropolyadditions or may be unzipping reactions as in retro- chain polymerizations and retropolyeliminations.

Unzipping can occur only with activated chains above their ceiling temperatures, that is, with living polymers, chain ends comprising certain functional groups, or radical ends generated by a prior chain scission.

Unzipping at higher temperature results in gaseous monomers that lead to voids in molded parts. This can be prevented or reduced by shorter sequence lengths of the unzippable polymer blocks. This strategy is utilized, for example, in the formation of acetal copolymers by copolymerization of trioxane (gives unzippable oxymethylene sequences ~OCH_2~) with ethylene oxide (oxyethylene sequences ~OCH_2CH_2~ do not unzip).

Decomposition can be caused by heat, light, oxygen, water, or other environmental agents as well as mechanical forces such as extensional flow, ultrasound, or a combination of these. It may lead to undesirable changes of mechanical, electrical, or optical properties. Unstabilized poly(vinyl chloride) ~$(CH_2CHCl)_n$~ discolors on decomposition because HCl is eliminated and colored polyene sequences ~$(CH{=}CH)_i$~ are formed. The polymer ultimately becomes brittle.

Decomposition is the primary degradation process in the pyrolysis of polymers, which may be desirable (rocket fuels, waste disposal) or not (fires), and often produces toxic fumes or smoke at lower flame temperatures. The risk of pyrolysis is enhanced for polymers with branching points, electron-accepting groups, long methylene sequences, and all groups capable of forming five- or six-membered rings. Thermostability is enhanced by aromatic rings, ladder structures, fluorine as a substituent, and low hydrogen content.

[Top of Page]

4. Plastics Manufacture

4.1. Homogenization and Compounding

Polymerization reactors do not deliver polymers that can be used directly as raw materials for plastics. Polymer melts must be filtered; melts, powders, beads, pellets, and granulates must be degassed; and all particulates must be dried or conditioned to environmental surroundings. Beads from suspension polymerization are washed and polymer solutions concentrated.

Batchwise polymerizations often lead to slightly different products, which are mixed (blended) to guarantee customers polymer grades according to specifications (microhomogenization). Granulates and pellets are sometimes similarly blended by macrohomogenization.

Compounding is the mixing of polymers and additives. This process can be performed either by adding single additives one at a time, by using additive systems, or by employing master batches. Additive systems are carefully adjusted mixtures of additives that are formulated to avoid mutually synergistic or antagonistic effects. Master batches are concentrates of additives in polymers; they facilitate the dosage of small amounts of additives.

The properties of compounds depend very much on the compounding process. Only heterogeneous compounds result when poly(vinyl chloride) is mixed with additives in regular mixers; these compounds cannot be processed directly to end-use products. High-performance mixers deliver free-flowing powders; these “dry blends” can be extruded and injection molded. Compounding is often carried out by specialized companies (compounders).

4.2. Additives [5], [73] (⟶ Plastics, Additives)

4.2.1. *Overview*

Polymers are rarely used directly as materials. They do not fulfill per se all technological requirements and become commercially useful only after they have been mixed with certain additives. These additives average ca. 23 % of the total weight of plastics (Table 18) but may range for individual plastics from 0 % (packaging films for food) to 90 % (barium ferrite-filled ethylene – vinyl acetate copolymers for magnetizable sealing strips). Their economic value approximates that of plastics because some additives command prices that are a hundred times those of the plastics themselves.

Table 18. Annual U.S. consumption of polymer additives for plastics, 1980 (plastics production: 16 076 000 t/a)

Additive	Consumption in t/a
Process additives	
Processing stabilizers	
Heat stabilizers (PVC)	43 000
Antioxidants	18 000
Processing aids	
Lubricants, slip agents, etc.	50 000
Polymeric processing aids	<1 000
Mold-release agents	<1 000
Nucleating agents	<1 000
Radical initiators for thermosets	15 000
Functional additives	
Stabilizing additives	
Antioxidants	see above
Metal deactivators	<1 000
Light stabilizers	2 400
Flame retardants	220 000
Biostabilizers	<1 000
Modifiers	
Colorants	150 000
Fillers and reinforcing agents	
Mineral-based	1730 000
Glass	462 000
Natural organic	133 000

Synthetic organic	2 000
Plasticizers	1 000 000
Blowing agents (without gases)	7 000
Impact modifiers	<5 000
Antistatic agents	2 000
Total	3 844 400

Polymer additives are to be distinguished from *polymer auxiliaries*; the former are added to polymers after the polymerization process, whereas the latter are used for the manufacture of polymers (e.g., polymerization catalysts, emulsifiers, initiators). The free-radical initiators used for the hardening of unsaturated polyesters are sometimes considered additives (added to the solution of unsaturated polyester molecules in monomers such as styrene) and sometimes auxiliaries (promoting the thermosetting reaction).

Additives are usually subdivided according to their application into process additives and functional additives. Process additives aid the processing of plastics by either stabilizing the chemical composition of polymers (processing stabilizers) or facilitating the processing itself (processing aids). Functional additives stabilize the chemical composition against attacks by environmental agents (stabilizing additives) or improve certain end-use properties (modifiers). Additives may also be classified according to their primary mode of action (chemical or physical). Some additives act in more than one way: a filler or colorant may also be a nucleating agent for crystallization, a pigment may enhance discoloration, etc.

4.2.2. *Chemofunctional Additives*

Polymers can be attacked by oxygen and ozone during processing and use. The attack generates radicals that cause chain reactions to produce even more radicals. Ultimately, changed chemical compositions, degradation, or cross-linking result. Oxidation is reduced if the accessibility of the oxidizable groups is limited or if radical formation is prevented. Oxygen diffusion into the polymer is slowed by certain coatings that act as mechanical barriers or by additives that diffuse into surface layers and are preferentially oxidized there.

Antioxidants (⟶ Antioxidants) prevent radical formation, at least during processing and for the targeted lifetime of the plastic. They are subdivided into deinitiators and chain terminators.

Deinitiators prevent the formation of radicals (therefore: preventive antioxidants) and are always used in combination with chain terminators (therefore also called secondary antioxidants). They are further subdivided into peroxide deactivators, metal deactivators, and UV absorbers. Peroxide deactivators (tertiary amines, tertiary phosphines, sulfides) convert hydroperoxides into harmless compounds before they can form radicals. Metal deactivators are chelating agents that form inactive complexes with catalytically active metal species (mainly from Ziegler – Natta polymerizations).

Chain terminators, such as hindered phenols, amines, and anellated hydrocarbons, react with already formed radicals and terminate the kinetic chain (chain-breaking or primary antioxidants). A combination of deinitiators and chain terminators often leads to synergistic effects. Antagonistic effects are also known (e.g., with carbon black as filler).

Light-induced degradation reactions can be prevented by the reduction of light absorption or by the addition of *UV absorbers* and quenchers. Less light is absorbed by the plastic if the surfaces are reflective or if certain pigments are added (e.g., carbon black). Ultraviolet absorbers either convert the incident light into harmless infrared radiation or are transformed into other chemical compounds. Quenchers deactivate excited states and themselves become excited.

Heat stabilizers prevent chemical transformations of plastics at higher temperature; their main application is for the stabilization of poly(vinyl chloride) against HCl elimination during processing.

Flame retardants either prevent oxygen access to burning plastics by formation of nonburning gases or by “poisoning” radicals generated by the burning. Some plastics are selfextinguishing because either CO_2 (polycarbonates) or water vapor and a protecting carbon layer are formed (vulcanized fiber: a $ZnCl_2$-treated cellulose) during burning. Chlorine and bromine compounds generate radicals during burning; these radicals combine with radicals from the degradation of plastics and stop the kinetic chain. Phosphorus compounds are oxidized during the burning to nonvolatile phosphor oxides, which either form a protective layer or are converted by water into phosphorus acids that catalyze the elimination of water.

The flammability of plastics is often characterized by the limiting oxygen index (LOI). The LOI value indicates the limiting value of the volume fraction of oxygen in an oxygen – nitrogen mixture that just allows the polymer to burn after ignition with a flame. Materials with LOI>0.225 are called flame retardant; those with LOI>0.27, self-extinguishing. Low LOI values (high flammabilities) are exhibited by polyoxymethylenes (0.14), polyolefins (0.17 – 0.18), saturated polyesters (0.20), and cellulose (0.20). High LOI values are exhibited by poly(vinyl chloride) (0.32), polybenzimidazole (0.48), and polytetrafluoroethylene (0.95). However, LOI values are not absolute measures of the flammability or combustibility of plastics because these properties also depend on flame temperatures, heat capacities, heat conductivities, melting temperatures, and melt viscosities. The hazard of burning plastics is also determined by smoke formation and the toxicity of the evolving gases.

4.2.3. *Processing Aids*

Processing aids facilitate the processing of plastics by enhancing either transport rates to the processing machines, flow behavior in these machines, achievement of final properties during processing, or removal of shaped articles from the machines or from each other.

Easy-flow grades of powdered polystyrenes, for example, often contain 3 – 4 % mineral oil, which forms a low-viscosity film on the surface of the particles and thus reduces friction in polar polymers; amphiphilic compounds such as metal stearates or fatty acid amides are used for this purpose. Such *external lubricants* also reduce the friction between polymer particles and the walls of the processing machine and the friction of polymer melts at such walls; they also prevent the cleavage of particles to smaller flow units. External lubricants are always incompatible with polymers and are thus found predominantly at polymer surfaces. They are related to *release agents*, which facilitate the separation of shaped articles from the tools (molds), and *slip agents*, which prevent the sticking together of shaped articles. Slip agents are thus sometimes also called lubricants.

Internal lubricants improve the flow behavior and homogeneity of polymer melts; they also reduce the Barus effect (see Section Flow Curves) and the melt fracture. Internal lubricants probably act by desegregating larger units (aggregates), which were probably formed during polymerization and are still present shortly after the melting of the polymer to a macroscopically homogeneous material. Typical internal lubricants are amphipolar compounds, e.g., modified esters of long- chain fatty acids.

Nucleating agents promote the crystallization of crystallizable polymers by generating many nuclei for crystallites. They prevent the formation of larger spherulites and thus improve the mechanical properties of plastics.

4.2.4. *Extending and Reinforcing Fillers*

Fillers are solid inorganic or organic materials. Some fillers are added mainly to improve the economics of expensive polymers; they are *extenders*. Extenders are usually particulate materials of corpuscular nature, such as chalk and glass spheres (aspect ratio ca. 1). Other fillers are *reinforcing agents* (“active fillers”); they improve certain mechanical properties. Reinforcing agents possess aspect ratios higher than 1; they may be short fibers, platelets (e.g., kaolin, talc, mica) (⟶ Reinforced Plastics), or long fibers (continuous filaments). Active fillers are sometimes subdivided into property enhancers (aspect ratio < 100) and true reinforcing fillers (aspect ratio > 100). No sharp dividing line exists between fillers and reinforcing agents, nor can the term “reinforcement” be unambiguously defined (e.g., it may denote an increase in breaking or impact strength or a decrease of brittleness).

A variety of fillers are used for plastics (Table 19).

Table 19. Fillers for thermoplastics (T), thermosets (D), and elastomers (E). For definition of other acronyms see Tables [1-4]

Filler	Application in	Concentration in %	Improved property
Inorganic fillers			
Chalk	PE, PVC, PPS, PB, UP	< 33 in PVC	price, gloss
Potassium titanate	PA	40	dimensional stability
Heavy spar	PVC, PUR	< 25	density
Talc	PUR, UP, PVC, EP, PE, PS, PP		white pigment, impact strength, plasticizer uptake

Mica	PUR, UP	< 25	dimensional stability, stiffness, hardness
Kaolin	UP, vinyls	< 60	demolding
Glass spheres	T, D	< 40	modulus of elasticity, shrinkage, compressive strength, surface properties
Glass fibers	T, D	< 40	fracture strength, impact strength
Fumed silica	T, D	< 3	tear strength, viscosity (increase)
Quartz	PE, PMMA, EP	< 45	heat stability, fracture
Sand	EP, UP, PF	< 60	shrinkage (decrease)
Al, Zn, Cu, Ni, etc.	PA, POM, PP	< 100	conductivity (heat and electricity)
MgO	UP	< 70	stiffness, hardness
ZnO	PP, PUR, UP, EP	< 70	UV stability, heat conductivity
Organic fillers			
Carbon black	PVC, HDPE, PUR, PI, PE, E	< 60	UV stability, pigmentation, cross-linking
Graphite	EP, MF, PB, PI, PPS, UP, PMMA, PTFE	< 50	stiffness, creep
Wood flour	PF, MF, UF, UP	< 5	shrinkage (decrease), impact strength
Starch	PVAL, PE	< 7	biological degradation

Glass-fiber-reinforced polymers often carry the abbreviation GRP (or FRP); those reinforced by synthetic organic fibers, CRP. Syntactic plastics are polymers reinforced with hollow glass spheres. The amounts of fillers added vary widely: in general, industry standards are about 30 wt % for thermoplastics and 60 wt % for thermosets.

Fillers act very differently in polymers. Some fillers form chemical bonds with polymers; an example is carbon black, which acts as a chemical cross-linker in elastomers. Other fillers can adsorb polymers on their surfaces: that is, physical bonds are introduced between fillers and polymers. On impact, adsorbed chain segments may take up energy and slip from the surface, which increases impact strength. Still other fillers act as nucleating agents in crystallizable polymers. Fillers furthermore constitute impenetrable walls to polymer coils. They restrict the number of conformational positions of chain segments near the filler surface; chains become less flexible, and tensile strengths and moduli of elasticity increase.

4.2.5. *Plasticizers (⟶ Plasticizers)*

Plasticizers are added to polymers to improve their flexibility, processibility, or foamability. About 500 different types of plasticizers are marketed. They are generally low molar mass liquids; polymer plasticizers are used in much smaller amounts. Of all plasticizers for thermoplastics, 80 – 85 % are used for poly(vinyl chloride); the most important plasticizer is di-2-ethylhexyl phthalate (“dioctyl phthalate,” DOP). The main plasticizer for elastomers is mineral oil; tires contain up to 40 %. Polymer plasticizers are mainly aliphatic polyesters and polyethers. The former are prepared by polycondensation; they thus possess fairly broad molar mass distributions and also contain monomeric and oligomeric molecules. Since their number-average molar mass is low (ca. 4000 g/mol), they are called oligomer plasticizers. Plasticization can also be achieved by copolymerization of the parent monomer with certain other monomers; this effect is called internal plasticization, in analogy to external plasticization by added high and low molar mass plasticizers.

External plasticizers are subdivided into primary and secondary ones. Primary plasticizers interact directly with chains by way of solvation. Secondary plasticizers are merely extenders; they can be used only in combination with a primary plasticizer. A certain plasticizer may thus be a primary or a secondary one, depending on the chemical constitution of the polymer. Mineral oils are, for example, primary plasticizers for polydienes but extenders for poly(vinyl chloride).

4.2.6. *Colorants*

Colorants are subdivided into dyes (soluble in polymer matrix) and pigments (insoluble). Textile fibers are dyed mainly with dyes. Pigments are preferred for plastics because they have a higher lightfastness and are more stable against migration than dyes. Colorants for plastics are dominated by titanium dioxide (60 – 65 %) and carbon black (20 %); only 2 % are dyes.

Pigment particles generally possess diameters of 0.3 – 0.8 µm, which allows the pigmentation of films and fibers of > 20-µm thickness. Very thin films and fibers are colored exclusively by organic pigments because these can be ground to much smaller diameters than inorganic ones. The hiding power of pigments increases with increasing difference between refractive indices of pigment and polymer.

Pigments need not have a special affinity for polymers. They must be wettable by the polymer melt, however, which can be achieved by treating them with surfactants. The aggregation of pigments is mainly the result of air inclusions; it can be removed by the application of vacuum. Pigments can be metered into plastics via master batches (in plastics), color concentrates (in plasticizers), or electric charging of the surface of polymer particles in granulate mixers. Up to 1 wt % of pigment can be mixed in by the last method.

4.2.7. *Blowing Agents*

Foamed plastics (plastic foams, cellular plastics, expanded plastics) are blends of polymers with gases (⟶ Foamed Plastics). They may be rigid (glass transition or melting temperature higher than use temperature) or flexible; their cell structure can be open or closed. The gases may be air, nitrogen, carbon dioxide, fluorinated hydrocarbons, etc.

Plastic foams can be produced by mechanical means (whipping, stirring), physical methods (shock volatilization of liquids, washing-out of solids), or chemical foaming either by internal foaming during the polymerization or by external foaming with chemical blowing agents. Chemical blowing agents are chemical compounds that decompose at elevated temperature with release of gases. The most widely used agent for natural rubber is *N,N′*-azobisisobutyronitrile $(CH_3)_2C(CN) – N{=}N – C(CN)(CH_3)_2$ (AIBN); for plasticized PVC, *N,N′*-dinitroso dimethyl terephthalamide $CH_3– N(NO) – CO – (p\text{-}C_6H_4) –CO – N(NO) – CH_3$ (NTA); and for other plastics, 1,1′-azobisformamide $H_2N – CO – N{=}N –CO–NH_2$ (ABFA). Blowing agents are used in amounts of ca. 0.1 % to eliminate sinks in injection molding; 0.2 – 0.8 % for injection-molded structural foams; 0.3 % for extended profiles; 1 – 15 % for vinyl plastisol foaming; and 5 – 15 % for compression-molded foam products.

4.3. Processing

Many different types of processes exist to convert monomers, prepolymers, or polymers into plastics (⟶ Plastics, Processing). In general, four procedures can be distinguished for processing raw materials into shaped plastics:

1. From monomers to polymers by direct polymerization with simultaneous shaping
2. From monomers by oligomerization to prepolymers, followed by simultaneous polymerization and shaping
3. From monomers by polymerization to polymers, followed by separate shaping
4. From monomers by polymerization to polymers, shaping to semifinished product, and finally to the end product

In all four cases, end products must sometimes be aftertreated (degated, polished, etc.).

Procedure 1 has only one process step and should be the most economical. It is widely used for wire coatings. However, technological problems are often encountered for shaped articles because the process involves the conversion of a low-viscosity liquid into a high-viscosity body. Since very high polymerization rates are required for this process to achieve high cycle times (phase sequences) in molding (i.e., number of articles produced per unit time), very few monomers, types of polymerization, and shaping methods are suitable. The method is utilized industrially in the RIM process, mainly for the polyaddition of diisocyanates with polyols. The anionic polymerization of lactams and the Diels – Alder polymerization of dicyclopentadienes and related compounds are also (in principle) suitable for RIM.

Procedure 2 is typical for thermosets. Oligomerizations are conducted to conversions shortly before the gel point. Additives are added, and the final processing step consists of a simultaneous cross-linking polymerization and shaping. Typical processing methods include molding, compression molding, cavity compression molding (hot pressing), injection – compression molding, injection molding, and extrusion.

Procedure 3 is generally chosen for thermoplastics. Polymers from the polymerization process are isolated before shaping and stored (e.g., as granulate) for a shorter or longer period of time. The stored polymers must be dried before further processing; otherwise, steam produced at higher processing temperatures may lead to voids and cavities in shaped articles. After compounding, the plastic raw material must be weighed out accurately for each cycle in discontinuous processing procedures. The filling factor (= raw material density/bulk density) must be known.

The choice of a processing technique is influenced technically by the rheological properties of the material and the form or shape of the desired article. The cost of processing machines and the cycle time are economically important. Commodity articles are produced by casting, centrifugal casting, hot pressing, injection – compression molding, coating, spraying, roll milling, extrusion, injection molding, cold forming, press forming, stretch forming, blowing, extrusion blowing, sintering, fluidized-bed sintering, flame

spraying, or hot blast sprinkling, and specialty articles by molding.

Procedure 4 is the method of choice for difficult-to-manufacture articles. Procedures such as welding, stamping, cutting, forging, sawing, boring, turning, milling (in the metalworking sense), sintering, and baking are employed for both thermoplastics and thermosets. The surfaces of plastics are sometimes treated further for technical (surface hardness, friction) or aesthetic reasons (gloss, color) by polishing, painting, metallizing, coating, etc.

[Top of Page]

5. Supermolecular Structures (⟶ Plastics, Properties and Testing)

Polymer properties are influenced not only by the chemical structure (constitution, molar mass, configuration, microconformation) but also by the physical structure of polymers. These structures may range from totally irregular arrangements of chain segments over shorter or longer parallelizations of chains, to voids and other defects in otherwise highly organized assemblies of polymer molecules. Two possible ideal structures exist in the solid state: perfect crystals and totally amorphous polymers. Polymer molecules are perfectly ordered in ideal crystals. They convert at the thermodynamic melting temperature into melts, which ideally are totally disordered. Amorphous polymers can be viewed as frozen-in polymer melts. They are polymer glasses that convert to melts at the glass transition temperature.

5.1. Noncrystalline States

5.1.1. *Structure*

Isolated polymer coils possess approximately a Gaussian distribution of chain segments; their segment density decreases with increasing chain length. However, the macroscopic densities of polymer melts do not change with chain lengths if end group effects on small molar mass molecules are neglected. Coils must therefore overlap considerably in polymer melts. At the glass transition temperature, cooperative segmental movements freeze in, and the physical structure of the melt is conserved. Small-angle neutron scattering studies have shown that the radii of gyration of amorphous polymers are indeed essentially identical for their melts, glasses, and solutions in theta solvents.

The absence of long-range order in melts and amorphous polymers does not exclude the presence of short-range order in these states. Because of the persistence of polymer chains, a parallelization of short segments seems probable, as is found, for example, for alkanes according to X-ray investigation. This local order does not exceed 1 nm in each direction.

The packing of chain segments cannot be perfect. Amorphous polymers thus possess “free volumes”, which are regions of approximately atomic diameters. The volume fraction of this free volume is ca. 2.5 % at the glass transition temperature and is independent of polymer constitution. Polymer segments in melts can move more freely than in the glassy state; the densities of melts are thus higher than the densities of glasses at the same temperature.

5.1.2. *Orientation*

Polymer segments, polymer molecules, and crystalline domains may be oriented along the machine direction by drawing or other mechanical processes. The orientation of chain segments need not necessarily lead to crystallization, however. An example is injection-molded polystyrene, which shows optical birefringence due to the orientation of segments but no X-ray crystallinity.

The degree of orientation of segments or crystallites can be measured by wide-angle X-ray or small-angle light scattering, infrared dichroism, optical birefringence, polarized fluorescence, and ultrasound velocity. The orientation is usually characterized for each of the three directions *a*, *b*, and *c* by a *Hermans orientation factor* f_i, where β is the angle between the draw direction and the principal axis of the segments:

$$f_i = (1/2)(3\langle \cos^2 \beta \rangle - 1) \quad (58)$$

The orientation factor becomes 1 for a complete orientation of that axis in the draw direction (β = 0°), –1/2 for a complete orientation perpendicular to the draw direction (β = 90°), and 0 for a random orientation.

Since methods to determine orientation factors are often expensive, time consuming, or difficult to perform, the easy-to- calculate *draw ratio* (length after drawing/length before drawing) is often used to characterize orientation. The draw ratio is, however, not a good measure of the degree of orientation because it depends on the history of the specimen. Drawing may also lead to viscous flow of the polymer without any orientation of segments and crystallites.

5.2. Polymer Crystals [74]

5.2.1. *Introduction*

Spheres and other geometrically simple entities can be arranged in crystal lattices irrespective of whether they are atoms, small molecules, large molecules such as enzyme, or latex particles. The centers of gravity of these entities occupy lattice sites that are regularly spaced in three dimensions. Lattices are called *superlattices* if the distances between lattice points significantly exceed atomic dimensions. The existence of three-dimensional order on the atomic level is called crystallinity.

Chain molecules can also be arranged in crystal lattices. The whole chain may be completely parallel to other chains as in the case of truly extended chain crystals; the microconformation of the chain can be all *trans* or helical (Fig. 16), and the chains may have a directional sense as in polyamide 6 or an inversion center as in polyamide 66 (Fig. 17). In other cases, only sections of the chains may be parallel to each other, either sections of a single chain as in folded chain crystals or those of neighboring chains as in fringed micelles. These ordered segments form crystalline regions in otherwise less ordered (though not necessarily completely disordered) polymers.

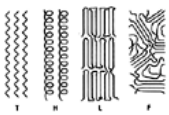

Figure 16. Schematic representation of crystalline polymers as extended chain crystals with chains in all-*trans* (T) or helical conformation (H), (L) lamellae of chains interconnected by tie molecules, and (F) fringed micelles. No bond angles are shown for H, L, and F.

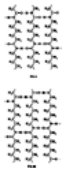

Figure 17. Pleated-sheet structures of polyamides 6 (anticlinal) and 66 (isoclinal)

Crystalline regions may be assembled into greater entities, which themselves may be more or less ordered. The types and degrees of order – disorder depend on external conditions such as temperature, pressure, cooling and heating rates, and presence of solvents. A variety of morphologies may thus exist for any crystallizable polymer [73], for example, spherical, platelet-like, fibrillar, and sheaflike structures of a polyamide 6 (Fig. 18).

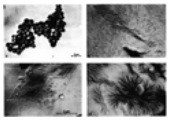

Figure 18. Electron micrographs of different morphologies of a polyamide 6

a) From a 260 °C solution in glycerol quenched into 20 °C glycerol; b) Same solution fast cooled (40 K/min); c) Same solution slowly cooled (1 – 2 K/min); d) Slow evaporation of formic acid solution at room temperature [75]

5.2.2. *Crystal Structures*

The unit cell describes the smallest, regularly repeating structure of a parallelepiped that generates a crystal by parallel displacements in three directions. The parallelepipeds are characterized by three lattice constants (distances between lattice points in three directions) and three lattice angles (Table 20).

Table 20. Crystal structure of some polymers

Lattice constants	Angles of unit cell

Polymer		N_u	*a*	*b*	*c*	α	β	γ	Helix	Crystal
			nm	nm	nm	°	°	°	type	system
PE	I [b]	2	0.742	0.495	0.254	90	90	90	2_1	R
	II [a, b]	2	0.809	0.479	0.253	90	90	107.9	2_1	M
PP, st		8	1.450	0.560	0.74	90	90	90	2_1	R
it	α [b]	12	0.665	2.09	0.495	90	99.6	90	3_1	M
it	β [b]	3	0.638	0.638	0.633	90	90	120	3_1	VI
it	γ [b]	12	0.647	2.140	6.50	89	100	99	3_1	T
PB, it	I [b]	18	1.769	1.769	0.650	90	90	90	3_1	VI
	II [b]	44	1.485	1.485	2.060	90	90	90	11_3	IV
	III [b]	8	1.238	0.892	0.745	90	90	90	4_1	R

[a] N_u=Number of monomeric units per unit cell.
[b] Crystal modification.
[c] R=(ortho)rhombic; M=monoclinic; VI=hexagonal; T=triclinic; IV=tetragonal.

Lattice constants reflect the microconformations of chains in crystal lattices. The *c*-direction is normally assigned to the chain direction; a value of *c* = 0.254 nm thus gives the distance between repeating units – CH_2 – CH_2 – in all-*trans* conformation (bond distance 0.154 nm, bond angle 112°). The *c*-values in carbon chains that are nonmultiples of 0.254 indicate the presence of non-all-*trans* conformations (e.g., helical structures in it-polypropylene and it-polybutylene-1). The *a* and *b* values reflect physical bonds (such as van der Waals or hydrogen bonds) that are longer than the chemical bonds in the *c*-direction. The anisotropy of bond lengths in the three spatial directions prevents the existence of cubic lattices for chain molecules. The other six lattice types are however found [hexagonal, tetragonal, trigonal, (ortho)rhombic, monoclinic, triclinic].

Most polymers form polymorphs (i.e., polymers of the same constitution and configuration can possess various energetically different crystal modifications). *Polymorphism* may be caused by different microconformations of polymer chains (e.g., polybutylene-1) or by different packing of chains with identical chain conformations (e.g., it-polypropylene). Such differences can be generated by small changes in processing conditions (e.g., different crystallization temperatures) or by stretching.

Copolymers can furthermore show *isomorphism* if different monomeric units can replace each other without change of lattice structure. Isomorphism of chains is also possible if the two corresponding homopolymers have analogous crystal modifications, similar lattice constants, and the same helix types. An example is a mixture of the γ-modification of it-polypropylene and the modification I of it-polybutylene-1 (Table 20). This phenomenon is sometimes also called allomerism. One company calls its crystalline copolymers from two or more olefinic monomers *polyallomers* because the monomeric units are isomorphic.

5.2.3. *Crystallinity*

Perfect crystal lattices should give sharp X-ray reflections and sharp melting points. Polymers do not exhibit these features, even if they show a crystallike appearance in electron micrographs. In addition, X-ray patterns of polymers frequently exhibit continuous, diffuse scattering. These findings can be interpreted by either a one-phase model (voids in crystals) or a two-phase model (coexistence of completely crystalline and completely amorphous regions). Crystallinity is defined as the presence of three-dimensional order on the level of atomic dimensions.

In reality, segments are ordered to various degrees. Since most experimental methods are sensitive to different types and degrees of order, various “degrees of crystallinities” are found depending on the experiment. The degree of crystallinity is the fractional amount of crystallinity. Stretched poly(ethylene terephthalate) has, for example, a degree of crystallinity of 59 % by IR spectroscopy, 20 % by density determination, and 2 % by X-ray diffraction.

5.2.4. *Morphology*

Earlier experimental findings were interpreted in terms of the two-phase model of partially crystalline polymers: the coexistence of crystalline reflexes and amorphous halos in X-ray diagrams, the independence of X-ray short periodicities on molar masses, the disappearance of X-ray long periodicities with increasing molar masses, lower macroscopic densities than predicted by X-ray densities of unit cells, broad melting ranges, optical birefringence of oriented polymers, and heterogeneity of partially crystalline polymers against chemical reactions.

In 1957, however, three different groups found that rhombic platelets are obtained if very dilute solutions of polyethylene are cooled. Similar platelets of 5 – 20-nm thickness are now known from many other crystallizable polymers (e.g., see Fig. 18). Electron diffraction showed sharp reflections typical of single crystals and chain directions perpendicular to the surface of the platelets. Since the contour lengths of the chains are greater than the thicknesses of the platelets, chains must be folded back as shown in Fig. 16 – L.

Crystallization from polyethylene melts similarly shows stacked single crystals. These lamellae are 75 – 80 % X-ray crystalline. After oxidation of the surface, a 100 % crystalline material remains. The folds or the interface between lamellae must thus be “amorphous.” Various models have been suggested for the surface layers: loose folds with adjacent reentry of the chains, a switchboard-like structure, adsorbed chains not involved in chain folding, loose or entangled cilia (tangling chain ends), and crystal bridges composed of tie molecules (Fig. 19). Electron microscopy has shown that crystal bridges exist in crystallized mixtures of high and low molar mass polyethylenes. Such interlamellar connections are responsible for some remarkable properties of partially crystalline polymers (see below). Highly crystalline polymers result if tie molecules and surface layers are removed (e.g., by grinding); these fairly low molar mass materials are called microcrystalline polymers.

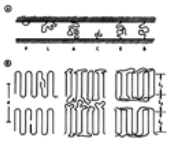

Figure 19. Models for the fine structure of lamellae of single crystals

A) Chains between two adjacent lamellae: F = sharp folds, L = loose loops, A = adsorbed chain, C = cilium, E = entangled cilium, B = crystal bridge or tie segment; B) Structures of lamellae: R=Regular lamellae with sharp folds, adjacent chain reentry, void from chain ends (below) and dislocation (above); L = lamellae with loose loops and adjacent reentry; S = switchboard model; *d* = crystallographic long period; L_c=thickness of lamella; L_a = thickness of “amorphous” surface layer ($d = L_c + L_a$)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Crystallization of melts sometimes delivers spherical, polycrystalline entities (Fig. 20). These spherulites exhibit radial symmetry. They consist of lamellae with tangential chain axes (Fig. 21 A); radial chain axes seem to be unknown (Fig. 21 B). The formation of spherulites can be controlled by the crystallization temperature, the cooling rate, and the addition of nucleating agents.

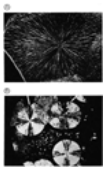

Figure 20. Spherulites of it-polypropylene under the phase contrast microscope (A) and the polarization microscope (B) [76]

Figure 21. Schematic representation of spherulite structures

A) Tangential chain axes; B) Radial chain axes (hypothetical)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Spherulites lead to opaque plastics if their diameters are greater than one-half the wavelength of incident light and their densities and refractive indices are different for crystalline and noncrystalline regions. The numbers and sizes of such spherulites are important for the fracture behavior of polymers.

Spherulites develop in a viscous environment if crystallization nuclei are present and if equal crystallization rates prevail in all three directions. Different crystallization rates exist in strongly stirred dilute solutions. Polymer chains then try to orient themselves along the flow gradients and crystallize in an extended (nonfolded) conformation. The resulting fibrils are organized in bundles with chains parallel to the fibril axes. These bundles act as nuclei – substrate for the epitaxial growth of the remaining chains. Since the shear rate is strongly reduced between bundles, crystallization leads to lamellae with folded chains whose axes are parallel to the fibril axes (Fig. 22). The resulting extended- chain structures are responsible for the remarkable properties of some ultraoriented polymers.

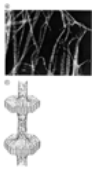

Figure 22. Shish kebab structures of linear polyethylene crystallized from rapidly stirred 5 % solutions in toluene [71]

A) Electron micrograph; B) Schematic drawing of chain arrangements

5.3. Mesophases

5.3.1. *Introduction*

Mesomorphic chemical compounds exhibit microscopic structures (mesophases) that are intermediate between crystals with long-range three-dimensional order and liquids without long-range order (Greek: *mesos* = middle, *morphe* = shape, form). Three types of mesophases are distinguished: liquid crystals, plastic crystals, and condis crystals. Since liquid crystals were discovered first and are furthermore the most common of the three classes, the term mesophase is often used as a synonym for liquid crystal.

Liquid crystals (LCs) (⟶ Liquid Crystals) possess a certain ordered structure like crystals, but they flow like liquids. *Thermotropic liquid crystals* exhibit this phenomenon in the pure state above a certain solid – mesophase transition temperature, where the solid may be a crystal or a glass. *Lyotropic liquid crystals* exhibit mesophase behavior in certain solvents above critical concentrations. Liquid crystalline behavior is present in both cases if the molecules contain “mesogens”, that are short rodlike sections that comprise the whole molecule as in low molar mass LCs, or segments thereof as in polymer liquid crystals (LCPs). Mesogens are anisotropic; they are oriented with respect to their mesogen axes but not with respect to their positions.

Plastic crystals exhibit order of positions but disorder of orientations of molecules. They are found for certain spherical low molar mass molecules packed in cubic lattices (i.e., they are isotropic). The absence of strong attractive forces between molecules and the presence of slip planes lead to easy deformabilities of plastic crystals, sometimes even under their own weight. No polymers are known that form plastic crystals.

Condis crystals are *con*formationally *dis*ordered crystals containing several conformational isomers in more or less ideal crystalline positions with respect to position and orientation of segments (conformational isomorphs). An example is the hexagonal high-pressure phase of polyethylene extended- chain crystals. It is often difficult to determine whether polymer crystals are normal or condis crystals.

On cooling, thermotropic liquid crystals either crystallize or solidify to “mesomorphic glasses” with retention of their liquid crystal structures. These glasses do not carry special names, but since “liquid crystalline glass” and similar names are oxymorons, the terms “LC glass,” “PC glass,” and “CD glass” have been suggested for the three classes of mesomorphic materials. The LCPs are used mainly as LC glasses not as liquid crystals per se, except in processing.

Condensed matter can thus exist in several of the five possible classes of pure phases: completely ordered crystal oC, mesomorphic glass mG, amorphous glass aG, ordered mesophase oM, and isotropic melt (liquid) iL. Seven types of biphasic materials exist: oC + mG, oC + aG, oC + oM, oC + iL, mG + aG, mG + iL, oM + iL, and three types of triphasic materials: oC+mG+aG, oC + mG+iL, oC + oM + iL. Since each class can be subdivided further, many physical structures (and thus properties) are possible for the same polymer.

Mesophasic polymers are characterized by domains of microscopic size (see below). “Microdomains” of similar size are also exhibited by certain block copolymers and, in much smaller sizes, by ionomers.

5.3.2. *Types of Liquid Crystals*

Liquid crystals and LC glasses are generated by rod- or disclike mesogens, which either comprise the whole molecule (as in low molar mass LCs) or sections of it (LCPs). Examples are

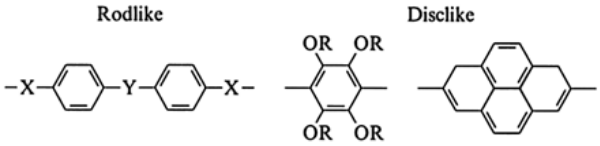

where X = O, OOC, COO, etc.; Y = COO, *p*-C_6H_4, CH=CH, N=N, etc.; R = CO $(CH_2)_nCH_3$, etc.; and phenylene residues may be replaced by 1,4- cyclohexane rings. The lengths of polymer mesogens are often comparable to those of repeating units; they seem to be identical with respect to persistence lengths.

Rodlike mesogens are also called calamitic (Greek: *calamos* = reed) and disclike ones discotic. Calamitic liquid crystals are further subdivided into smectic, nematic, and cholesteric mesophases. Smectic LCs show fan-shaped structures under the polarization microscope; they feel like soaps (Greek: *smegma* = soap). Nematic LCs form threadlike schlieren (Greek: *nema* = thread). Cholesteric LCs exhibit beautiful reflective colors, which were first discovered with cholesterol.

The birefringence of LCs originates from the anisotropy of mesogens, which are more or less aligned in microscopic domains with diameters in the micrometer range. These alignments lead to different refractive indices parallel and perpendicular to the direction of polarization of incident light. The domains themselves are arranged randomly. The presence of birefringence however is not proof for the existence of mesogens; it may also result from higher-melting crystallites in partially molten polymers or from crystallites in partially molten polymers or from crystallites in gels or concentrated solutions.

Liquid crystals are turbid because their domains are anisotropic structures whose dimensions are greater than the wavelength of incident light. At a certain “clearing temperature,” these domains melt to a transparent (isotropic) liquid.

The soaplike character of *smectic LCs* is due to two-dimensional, layerlike arrangements of mesogens (Fig. 23). At present, twelve different smectic types are known: four with the average direction of the long axes perpendicular to layer planes and seven with tilted axes. Examples are type S-A with mesogens perpendicular to the layer planes; type S-B, the same but with perfect hexagonal packing of mesogens; type S-C, with mesogen axes at an angle to the planes, etc.

Figure 23. Schematic representation of mesogens in different types of mesophases

S = smectic (types A and C); N = nematic; N–C = nematic cholesteric; C–D = columnar discotic; N–D = nematic discotic

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Nematic mesophases (Fig. 23, N) are one-dimensionally ordered. The principal axes of mesogens are more or less parallel to each other; the centers of mass are however distributed at random.

Cholesteric mesophases (Fig. 23, N – C) are observed only with chiral mesogens. They consist of layers of nematically ordered mesogens. These layers have a sense of rotation relative to each other due to the presence of chirality centers in mesogens. Cholesteric mesophases are thus screwlike nematics.

Discotic mesophases may exist in many different structures. The disclike mesogens are randomly arranged in nematic-discotic types (Fig. 23, N – D) but stacked like coins in columnar-discotic ones (Fig. 23, C – D).

5.3.3. *Mesogens*

Mesogens of polymers can be contained in main chains (MC-LCPs) or in side chains (SC-LCPs). Thermotropic MC-LCP glasses are used as engineering plastics; lyotropic MC-LCP glasses, for fibers. Thermotropic SC-LCPs are not used commercially at present but may find use in optical recording devices.

Mesogens usually constitute the whole molecule in low molar mass LC molecules. In macromolecules, this is true only for dissolved helices up to certain molar masses. Such helicogenic macromolecules are rigid rods at small molar masses but become coillike at very high degrees of polymerization (cf. the behavior of short and very long garden

hoses). This behavior is demonstrated by the molar mass dependence of their intrinsic viscosities, which are measures of the molecule volume per unit mass (Fig. 24). Rodlike molecules are characterized by their axial ratio (aspect ratio) $\mathbb{A}$ (length/diameter).

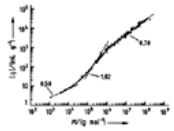

Figure 24. Molar mass dependence of intrinsic viscosities of wormlike chains

(○) Helices of poly(γ-benzyl-L-glutamate) (PLBG) in *N,N*-dimethylformamide at 25 °C; (•) double helices of deoxyribonucleic acids in dilute aqueous NaCl solutions at 20 °C. The two polymers are comparable because the helix diameters are similar (1.5 nm for PLBG, 2.0 nm for DNA). Numbers indicate exponents *a* in Equation (5)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

All other polymer chains form either flexible or semiflexible coils in dilute solution or in their isotropic melts (see examples in Fig. 25). Semiflexible (wormlike) chains SF allow rotations around some of their chain bonds. The rodlike character of their segments is however preserved in melts or concentrated solutions if the chain axes remain linear or if the segment linearity can be restored by a crankshaft motion. Conjugated bond structures are helpful but not necessarily required for mesogens since, for example, 1,4-phenylene rings may be replaced by 1,4- cyclohexylene units.

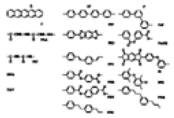

Figure 25. Examples of rodlike (R), wormlike or semiflexible (SF), and flexible (F) molecules and segments. See text for further explanation.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Linear chain axes are present in poly(*p*-phenylene) (PPP) or poly(*p*-phenylene benzoxazole) (PBO). Crankshaft motions around the chain axes are possible in the conjugated chains of poly(*p*-phenylene vinylene) (PPV), the partially conjugated chains of poly(*p*-hydroxybenzoic acid) (PHB), and even in the nonconjugated poly(*p*-phenylene alkylenes) as long as even numbers of methylene groups are present between phenylene rings as in poly(*p*-phenylene ethylene). Semiflexible molecules are characterized by their persistence lengths *a* or their Kuhn lengths $L_K = 2\,a$.

Rodlike molecules and semiflexible molecules with long persistence lengths form many physical bonds between parallel chains. The total energy of such assemblies is so high that individual intramolecular chemical bonds rather than all of the physical bonds are broken upon heating: these molecules decompose instead of producing thermotropic mesophases. They may however form lyotropic mesophases in suitable solvents. An example is poly(*p*-phenylene bisbenzoxazole).

The "stiffness" of the chain segments can be removed or reduced by several means (Fig. 26). Bulky substituents prevent the parallelization of chains; they "frustrate" crystallization. Nonlinear chain elements (such as *ortho* and *meta* substituents) work in the same way. Flexible chain segments reduce the persistence lengths of the chains.

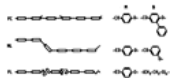

Figure 26. Flexibilization of rigid chain segments by frustrated crystallization (FC), nonlinear chains (NL), or flexible chain elements (FL); M = mesogen-forming monomeric units; B = stiffness-breaking units

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

5.3.4. *Lyotropic Liquid Crystalline Polymers*

Molecule axes of low molar mass rodlike molecules are disordered in isotropic melts, whereas mesogen axes are more or less parallel to each other in nematic and smectic mesophases. Very short rods resemble spheroids. They can arrange themselves in various stable ways, but the parallel ordering need not be much more stable than other arrangements. Therefore, a critical axial ratio exists above which the simple geometric anisotropy is sufficient to stabilize a mesophase.

This critical axial ratio has been calculated with the lattice theory of polymer solutions as $\mathbb{A}_{crit} \approx 6.42$. A geometric stabilization by repulsion is insufficient for $\mathbb{A} < \mathbb{A}_{crit}$ and the mesogens must exert additional orientation-dependent attractive forces.

In solution, repulsive forces between mesogens can act only at sufficiently high mesogen concentrations. At a critical volume fraction ϕ_p^*, phase separation occurs into a polymer-rich mesophase and a dilute isotropic phase. The dependence of ϕ_p^* on $\mathbb{A}$ can be described to a first approximation by [77]

$$\phi_p^* \approx 8\left(1 - 2L^{-1}\right)/L \tag{59}$$

The axial ratio $\mathbb{A}$ is the molar mass-dependent true axial ratio of rodlike molecules (i.e., helices) and the molar mass-independent Kuhn length of semiflexible molecules. Equation (59) is remarkably well fulfilled for both helical and semiflexible molecules (Fig. 27). An exception is the tobacco mosaic virus, probably because of additional charge effects that are not considered by the theory.

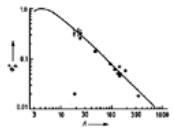

Figure 27. Critical volume fractions ϕ_p^* for phase separations of isotropic solution ⟶ nematic mesophase as function of $\mathbb{A}$ [i.e., the axial ratio of helical (rigid) molecules (circles) and the Kuhn length of semiflexible molecules (triangles, square)]

* Tobacco mosaic virus; solid line: prediction of Equation (59)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Lyotropic mesophases are formed by the helices of tobacco mosaic virus (*c*>2 %) and deoxyribonucleic acids (*c*>6 %) in dilute aqueous salt solutions as well as by poly(α-amino acids) in helicogenic solvents, e.g., poly(γ-methyl L-glutamate) in dichloromethane – ethyl acetate (12 : 5) (*c*>15 %). Hydroxypropyl cellulose (HPC) forms a cholesteric mesophase at room temperature in water at w_{HPC}>0.41.

On shearing, mesophase domains become oriented with respect to the mesogen axes (Fig. 28). This alignment becomes frozen-in on rapid quenching of thermotropic LCs or by precipitation of lyotropic ones into baths (half-lifes of orientation between a few seconds and a few hundredths seconds). The resulting LC glasses possess good orientations of mesogen axes in the shear direction (e.g., fiber direction), high tensile strengths, and low extensibilities.

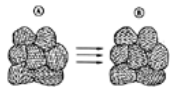

Figure 28. Alignment of mesogen axes upon shearing

A) Mesophase domains in liquid crystalline state at rest; B) During shearing or as LC glass with "frozen in sheared state"

A number of LC fibers are produced commercially. Poly(*p*-phenylene terephthalamide) (PPB – T) was originally spun from concentrated sulfuric acid solutions, whereas chlorinated *N*-methyl pyrrolidine is now used:

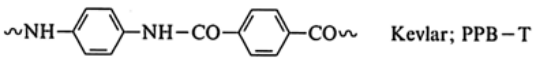

A similar copolyamide is also on the market

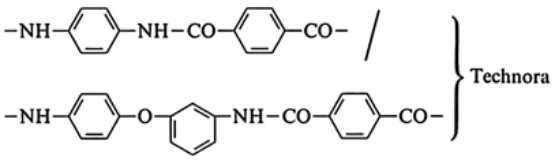

Even better LC fibers are produced from poly(*p*-phenylene-2,6-benzoxazoles) and poly(*p*-phenylene-2,6-benzthiazoles) (PBT), which are spun from poly(phosphoric acid) solutions, at present on a bench scale. Both "*cis*" (shown) and "*trans*" (X and N exchanged at one ring) structures are known:

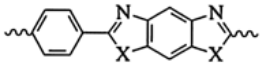

PBO (X = O); PBT (X = S)

5.3.5. *Thermotropic Liquid Crystal Polymers*

Non-Newtonian shear viscosities of liquid crystal polymers drop drastically with increasing shear rates due to the orientation of mesogen axes upon shearing. Less energy is thus needed for processing thermotropic LCPs than isotropic thermoplastics. The high orientation times permit a freezing-in of the mesogen orientation in injection molding, for example. Because of the higher than usual moduli of elasticity and tensile strengths, such LCPs are also called *self-reinforcing polymers*.

The first self-reinforcing polymer was the copolyester X7 G with 60 mol % *p*-hydroxybenzoyl and 40 mol % terephthaloyl glycol units. It is no longer produced because the less expensive glass-fiber-reinforced saturated polyesters possess similar properties. Xydar and Vectra are, however, industrially produced:

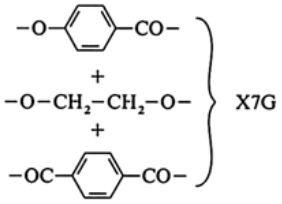

X7G

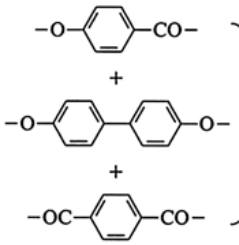

Xydar

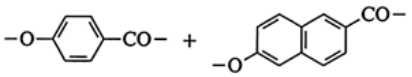

Vectra

Polymeric side- chain LCs can serve for the thermooptical storage of information. A guided laser beam increases the temperature locally, whereupon phase transformations occur. The change of order can then be frozen-in on cooling below the glass transition temperature. The resolution is ca. 0.3 µm.

5.3.6. *Block Copolymers*

Block copolymers consist of two and more blocks of constitutionally or configurationally different monomeric units (e.g., $A_n - B_m$, $A_n - B_m - A_n$). In most cases these blocks are thermodynamically or kinetically immiscible because their homopolymers A_n, B_m, etc., are not miscible. The demixing of block copolymers cannot proceed, however, to macroscopic phase separations because the blocks are chemically coupled. Similar blocks of different block copolymer molecules can thus only aggregate and form domains in a matrix of the other blocks. This phenomenon is called microphase separation.

Three different *applications* of block copolymers are known: polymeric detergents, thermoplastic elastomers, and compatibilizers for blends. *Polymeric detergents* consist of hydrophilic and hydrophobic blocks. Examples are the multiblock copolymers from watersoluble ethylene oxide blocks $[-(OCH_2CH_2)_n-]$ and water-insoluble propylene oxide blocks $-[OCH_2CH(CH_3)]_n-$. The hydrophobic blocks associate in water, which produces high viscosities.

A similar action is exerted by *compatibilizers* for nonmiscible blends of A_n and B_m polymers. These polymers are $A_p - B_q$ diblock copolymers: the A_p block resides in the A_n phases and the B_q block in the B_m ones. Compatibilizers thus anchor the two types of phases (Fig. 29).

Figure 29. Arrangement of A_n blocks (with A units •) and B_m blocks (with B units ○) in diblock (C, L) and triblock copolymers (S)

C: compatibilizers at an interface - - -; S: spherical A_n domains in continuous matrix B_m; L: lamellae of A_n and B_m

Thermoplastic elastomers consist of tri- or multiblock copolymers. Triblock copolymers $S_n - B_m - S_n$ possess a soft center polybutadiene block $- B_m -$ (glass transition temperature below use temperature) and two hard end polystyrene blocks $S_n -$ (glass transition temperature above use temperature). Below certain *m/n* ratios, styrene blocks form spherical microdomains in a continuous matrix of butadiene blocks. These spherical domains thus act as cross-linking units in an elastomer (Fig. 29, S). The hard domains "melt" above their glass transition temperatures, and the block copolymers can then be processed like thermoplastics.

The morphology of *diblock copolymers* is governed mainly by the relative spatial requirements of the blocks. Both blocks, if independent, would form unperturbed random coils. If the unperturbed volumes of both blocks are of equal size, all A_n blocks would line up in one layer and all B_m blocks in another because the blocks are (1) coupled to each other and (2) incompatible. The A_n layer faces another A_n layer, where the A_n blocks are coupled to B_m blocks, etc. (Fig. 29, L). The diblock polymers thus form lamellae with a thickness of two coil diameters. Similar structures result for triblock polymers $A_r - B_m - A_r$, where $r = n/2$.

If the volume of the A_n blocks is much lower than the volume of the B_m blocks, then the A_n blocks can no longer be packed into layers without violating the demand for tightest packing or deviating from the shape of unperturbed coils. Both possibilities are energetically unfavorable, and the smaller A_n blocks thus cluster together and form spherical domains in a continuous matrix of the larger B_m blocks. If the A_n blocks are somewhat larger (but not big enough to form lamellae), then cylinders would result (Fig. 30).

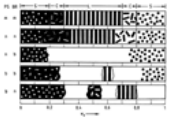

Figure 30. Morphology of styrene – butadiene diblock copolymers as a function of the mass fraction of the styrene units. White: polystyrene blocks (PS); black: polybutadiene blocks (BR). The block length distribution may be (m) molecularly homogeneous (theory), (n) narrow, or (b) broad

S = Spherical domains, C = cylinders (rods), L = lamellae (layers). Upper row: predictions of Meier theory; rows 2 – 5: schematic after experimental results of many authors

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

A great number of thermoplastic elastomers of both the triblock and the multiblock type is known [67]. Their chemical structures range from $(\text{styrene})_n - (\text{butadiene})_m - (\text{styrene})_n$ triblock copolymers to polyether – urethane multiblock polymers, ethylene – propylene copolymer – polyolefin blends, polyether – esters, and graft copolymers of butyl rubber on polyethylenes.

5.3.7. *Ionomers*

Ionomers are copolymers from primarily hydrophobic monomers with small amounts of ionic comonomers. The ionic groups may be in the main chain or in substituents. Four different types of ionomers are produced commercially.

$-CH_2-CH_2- + < 10\ mol\% - CH_2-\underset{\underset{COOH}{|}}{C}(CH_3)-$ Surlyn

$-CH_2-CH_2- + < 3.5-20\ mol\% - CH_2-\underset{\underset{COOH}{|}}{CH}-$ EEA Copolymer

$-CF_2-CF_2- + ?\ mol\% - CF_2-\underset{\underset{O-[CF_2-CF(CF_3)-O]_n-(CF_2)_2SO_2H}{|}}{CF}-$ Nafion

$-CH_2-CH_2- +$
$-CH_2-CH(CH_3)- +$ 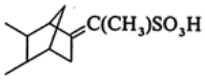 $C(CH_3)SO_3H$ Thionic

Surlyn, EEA, and Nafion are synthesized by direct copolymerization of the corresponding monomers. Thionic results from the postpolymerization sulfonation (< 5 %) of the diene units (usually 5-ethylidene-2-norbornene) of the primary ethylene – propylene – diene (EPDM) rubber. Both EEA and Nafion are used as acids, whereas Surlyn and Thionic are obtained by aftertreatment that leads to sodium or zinc salts.

The introduction of ionic groups into hydrophobic polymers leads to ion associations and subsequently to microphase separations. The ion association is controlled by the coordination number of the ions; that is, sodium ions with valence 1 and coordination number 6 are as good cross-linkers as zinc ions with valence 2. Cross-linking by ions does not occur via ion pairs but via clusters and domains of many ions. These ionic bonds dissociate at higher temperature, and the ionomers can then be processed like thermoplastics. At lower temperature, they behave either as thermoplastic elastomers (T_G of polymer segments lower than use temperature) or as "reversible" thermosets (T_G of segments higher than use temperature).

5.4. Gels

Chemically lightly cross-linked polymers swell upon addition of solvent to form gels. The maximum degree of swelling results from the attempt of segments located between cross-linking sites to attain their energetically most favored coil dimensions which in turn is counteracted by the elastic retraction due to cross-links. Physically cross-linked polymers behave in the same way if the physical cross-links (crystallites, ion clusters, etc.) survive the dissolution process.

The latter phenomenon is utilized in poly(vinyl chloride) pastes. The strong dipole – dipole interactions between C–Cl bonds leads to associations of chain segments, which at room temperature resist dissolution by plasticizers such as phthalates, adipates, sebacates, or citrates. Heating of PVC with these plasticizers and subsequent cooling leads to gelation by formation of a lightly physically cross-linked network. Plasticized PVC thus behaves like an elastomeric material. The physical cross-links dissociate at higher temperature, and the resulting PVC paste can be processed with extruders or roll mills.

5.5. Polymer Surfaces

5.5.1. *Structure*

Outer layers of solid or liquid polymers do not exhibit — against air, water, metal surfaces, etc. — the same average composition as the interiors of these polymers. Groups or segments with the lowest Gibbs interfacial energy will reside preferentially at the surface or interface. Surface structures may also be altered by chemical attack during processing or on prolonged use (e.g., by oxidation) or by physical processes such as transcrystallization.

This behavior can be studied by a wide variety of new spectroscopic methods, for example, the concentration of immediate surface groups by Fourier transform infrared spectroscopy (FT-IR); the composition of the upper 0.3 – 0.5-nm layer by ion scattering spectroscopy (ISS, LEIS); of the upper 1 – 10 nm by photoelectron spectroscopy [UPS, PE (S)], electron spectroscopy for chemical analysis (ESCA, XPES, XPS, IEE), and Auger spectroscopy (AES).

These methods have shown that poly-(2-vinylpyridine) and poly(4-vinylpyridine) have much more hydrophobic surfaces than their chemical structure indicates: CH_2 and CH groups face the surface; pyridine residues (C_5H_4N), the interior. Imide – carbonyl residues of polyimides, on the other hand, are preferentially found at the surface and aromatic rings in the interior. The surface is enriched by siloxane residues in polycarbonate – polydimethylsiloxane block copolymers but by styrene units in polystyrene – poly(ethylene oxide) block copolymers. Since the surface energy depends on the contact (air, water, metals, etc.) and also on kinetic effects (thermal history, solvents, etc.), one and the same polymer may possess various surface compositions and thus also surface properties.

5.5.2. *Interfacial Tension*

Interfacial tension is the force that acts at the interface between two phases; it is called surface tension for the interface condensed phase – gas phase.

Surface tensions γ_{lv} of liquid polymers against air decrease with the two-third power of the molar mass according to

$$\gamma_{lv} = \gamma_{lv}^{\infty} + K_e \cdot M^{-2/3} \quad (60)$$

They do not vary markedly with temperature. Typical values (at 150 °C) range from 13.6 mN/m (polydimethylsiloxane), 22.1 mN/m (it-polypropylene), 28.1 mN/m (polyethylene), and 33.0 mN/m poly(ethylene oxide). At present, no correlation of these surface tensions with chemical constitution is possible because the surface structures of liquid polymers are unknown.

Interfacial tensions between two liquid polymers are always lower than surface tensions. They are low for two apolar polymers but higher if one polymer or both of them are polar. For example, the interfacial tension of polyethylene – it-polypropylene is only 1.1 mN/m, but that of polyethylene – poly(ethylene oxide) is 9.5 mN/m and poly(ethylene oxide) – polydimethylsiloxane 9.8 mN/m.

The interfacial tensions between solid polymer and air are generally unknown. They do not differ markedly, however, from the critical surface tensions determined by the *Zisman method*. In this method, the liquid – air surface tensions of various liquids are plotted against the cosine of contact angles ϑ of these liquids against air and extrapolated to $\cos\vartheta \longrightarrow 0$. The resulting "critical" surface tensions of polymers are always lower than the surface tension of water: polytetrafluoroethylene, 18.5; polyethylene, 33; poly (vinyl alcohol), 37; polyamide 66, 46; and urea – formaldehyde resins, 61 mN/m. Fluorinated polymers have especially low critical surface tensions. They are not wetted by water (72 mN/m) or oils and fats (20 – 30 mN/m) and are therefore used as surface coatings to prevent sticking.

5.5.3. *Adsorption*

Macromolecules possess many adsorbable groups and segments. The type of adsorption depends on the adsorption energy per segment (group), the concentration of adsorbable macromolecules, and the duration of the experiment.

"Adsorption equilibria" are established in minutes to hours at smooth surfaces but may take days to attain at rough surfaces or powders. The adsorption time increases with the concentration and molar mass of the polymer.

Polymer coils tend to overlap even in fairly dilute solutions: the adsorbed polymer layers are almost always multilayers, except for the adsorption of oligomers (X<10 – 100) or from very dilute solutions (volume fractions <10^{-4} – 10^{-3}). At higher concentrations or molar masses, adsorbed amounts and structures of polymer layers are mainly kinetically controlled. Reorganization may occur with time: for example, the adsorption of polar polymers at polar surfaces can change from an initial loose loop structure to a more compact, flat covering. The layer of poly(ethylene oxide) adsorbed on chromium surfaces is only ca. 2 nm thick and so tightly packed that the refractive index of the adsorbed layer is identical with that of the crystalline polymer. The thickness of the surface layer of polystyrene of M = 176 000 g/mol adsorbed from a 5 mg/mL solution in the thermodynamically bad solvent cyclohexane is however about 27 nm (i.e., identical with the end-to-end distance in the unperturbed state). The physical structures of adsorbed polymers obviously play an important, yet widely unknown, role in many technological applications of polymers.

[Top of Page]

6. Thermal Properties

6.1. Molecular Motion

6.1.1. *Thermal Expansion*

Isotropic bodies expand upon heating equally in all three spatial directions because of the increasing thermal motions of atoms, groups, and molecules. The expansion is characterized by the cubic expansion coefficient $\beta = V^{-1} (\partial V / \partial T)_p$, which is usually converted by $\beta = 3\ \alpha$ into the linear expansion coefficient $\alpha = L^{-1} (\partial L / \partial T)_p$. Such isotropic materials are, for example, diamond ($\alpha = 1.06\times10^{-6}\ K^{-1}$), iron ($12\times10^{-6}\ K^{-1}$), water ($70\times10^{-6}\ K^{-1}$), and carbon disulfide ($380\times10^{-6}\ K^{-1}$) (all data at 25 °C). All of these materials exhibit the same types of bonds in the three directions: all covalent bonds between carbon atoms in diamond, all metallic bonds between iron atoms, all

hydrogen bonds between water molecules, and all dispersion forces between CS_2 molecules.

Polymer chains are however anisotropic: the intramolecular bonds along the chain are chemical (almost always covalent); the intermolecular bonds perpendicular to the chain, physical (dispersion forces, dipole – dipole interactions). On thermal expansion of polymer crystals, chains contract because of the increasing amplitude of the lateral motions. The thermal expansion coefficient in the chain direction is thus zero to negative, whereas the overall expansion coefficient is positive. The linear thermal expansion coefficients of polymers are thus averages over the three spatial directions; they lie between those of metals and liquids. Typical values are $\alpha = 60\times10^{-6}\ K^{-1}$ (polyamide 6) and $80\times10^{-6}\ K^{-1}$ [poly(vinyl chloride)].

Significant problems may thus arise due to different expansion coefficients for polymer –metal composites upon thermal stress. Another problem is the low dimensional stability of polymers on temperature change. This problem may be aggravated by a concomitant change of the water content of polymers or by recrystallization phenomena, both of which can lead to warping.

6.1.2. *Heat Capacity*

The molar heat capacity can be 3 *R* per atom according to the law of equal distribution of energy. In reality, degrees of freedom are always frozen in and the molar heat capacity is lowered. Empirically, a value of ca. 1 *R* has been found for solid polymers at room temperature. Poly(2,6-dimethylphenylene oxide) $[(C_8H_8O)_n]$ at 25 °C has a specific heat capacity of 1.22 J $K^{-1}\ g^{-1}$ and a molar heat capacity (per monomeric unit) of 146.4 J $K^{-1}\ mol^{-1}$. The molar heat capacity (per mole of atoms) is thus 146.4 J $K^{-1}\ mol^{-1}$/17 = 8.61 J $K^{-1}\ mol^{-1}$ (i.e., approximately 1 *R* = 8.314 J $K^{-1}\ mol^{-1}$).

Below the glass transition temperature T_G, heat capacities are not influenced by the degree of crystallinity of the polymer. At T_G, a stepwise increase is observed. The heat capacity passes through a maximum a few degrees below the macroscopic melting temperature; that is, the true melting temperature is given by the upper end of the melting range where the largest and most perfect crystals melt.

6.1.3. *Heat Conductivity*

Conventional polymers are electrical insulators. Heat is thus not transported by electrons but by elastic waves (phonones in the corpuscular model). The free path length of phonones is defined as the distance at which the intensity of elastic waves has decreased to 1/e. This free path length is about 0.7 nm for glasses, amorphous polymers, and liquids; it is practically independent of temperature. The slight decrease of heat conductivities (thermal conductivities) of amorphous plastics and elastomers below their glass transition temperatures must thus be caused by the decrease of heat capacities with decreasing temperature. For crystalline polymers, a strong decrease of heat capacities is observed at their melting points because packing densities decline drastically at these temperatures.

6.2. Thermal Transitions and Relaxations

6.2.1. *Overview*

Thermal transitions and relaxations are characterized by large changes of physical properties at the corresponding temperatures. In a true thermal transition, chemical compounds are in equilibrium on both sides of the transition temperature. An example is the melting transition.

Thermal relaxations, on the other hand, are kinetic effects. They depend on the frequency of the experimental method and thus on the time scale. Typical thermal relaxations are caused by the onset of translations and rotations of charges, dipoles, and chemical groups (i.e., by atomic motions).

Some experimental methods work at frequencies that such relaxations appear to be thermal transitions. The best-known example is the glass transition temperature at which hard, glassy polymers convert to soft, rubbery materials and vice versa. In many cases, a thermal effect cannot be unambiguously classified as either transition or relaxation.

Thermal transitions and relaxations can be detected and determined by many different experimental methods. The most commonly applied methods for the determination of *thermal transitions* are differential thermoanalysis (measures temperature differences between specimen and standard on heating or cooling with constant rate), differential scanning calorimetry (does the same for enthalpy differences), thermomechanical analysis (deformation of specimen under load), dynamical – mechanical analysis (either free or forced vibration of specimen), and torsional braid analysis (specimen on vibrating support). Figure 31 shows a typical thermogram.

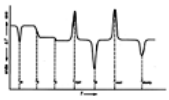

Figure 31. Idealized thermogram of a partially crystalline polymer with solid – solid transition T_{ss}, glass transition temperature T_G, liquid – liquid transition T_{ll}, maximum crystallization temperature T_{cryst}, melting temperature T_M, maximum temperature T_{react} of a chemical transformation, and maximum temperature T_{decomp} of chemical degradation

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Many methods are available for the study of molecular motions and thus *thermal relaxations*. These methods work with frequencies ν which correspond to correlation times t_c of 10^{-12} s < 1/ ν<10^6 s (11.5 d). Typical methods include quasi-elastic neutron scattering ($10^{-12}<t_c/s<10^{-8}$), NMR spin-lattice relaxation ($10^{-12} < t_c/s<10^{-5}$), dielectric relaxation ($10^{-10} < t_c/s<10^{-5}$), and photon correlation spectroscopy ($10^{-4}<t_c/s<10^2$). Thermal relaxations furthermore manifest themselves in sudden changes of mechanical properties, such as rebound elasticity ($t_c/s\approx10^{-5}$), penetrometry ($t_c/s\approx10^2$), mechanical loss ($10^3<t_c/s<10^7$), and thermal expansion ($t_c/s\approx10^4$). Slow methods (high t_c) are called “static” methods; fast ones, “dynamic.” Transition – relaxation phenomena are also detected by several empirical, standardized methods that measure the resistance of specimens against flow under various loads (Vicat temperature, heat distortion temperature, Martens temperature, etc.).

Various characteristic signals are observed at a fixed temperature for a given frequency (see insert in Fig. 32). They often cannot be correlated with molecular processes and are commonly indicated with descending temperature by letters in the sequence of the Greek alphabet, starting with the melting temperature (crystalline polymers, subscript c) or glass transition temperature (amorphous polymers, subscript a).

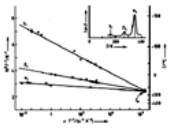

Figure 32. Dependence of inverse relaxation temperature 1/*T* on the logarithm of reduced frequency ν/T for various relaxation processes of a low-density polyethylene

Insert shows the mechanical relaxation spectrum at ν = 1000 Hz.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The various methods work at different frequencies and thus give different relaxation temperatures for the same molecular process. The frequency dependence of relaxation temperatures can be described by the Eyring equation for rate processes:

$$\nu = (k_B T/2\pi h)\ \exp\ (-\Delta H^{\neq}/RT)\ \exp\ (\Delta S^{\neq}/R) \tag{61}$$

where k_B is the Boltzmann constant, *h* the Planck constant, $\Delta H^{\neq}$ the activation enthalpy, and $\Delta S^{\neq}$ the activation entropy. Transformation of Equation (61) leads to

$$\frac{1}{T} = \frac{R}{\Delta H^{\neq}} \cdot \left[\left(\frac{\Delta S^{\neq}}{R} + \ln\frac{k_B}{2\pi h}\right) - \ln\frac{\nu}{T}\right] \tag{62}$$

In a $(1/T) = f\,[\ln(\nu/T)]$ plot, lines for various processes intersect at the melting temperature (T_M = 131 °C in Fig. 32). Such common intersects seem to be general for nonhelical polymers.

6.2.2. *Crystallization*

The crystallization of coillike polymers from dilute solutions leads to platelets (Fig. 18) in which folded polymer chains are arranged with their stems perpendicular to the fold surface (Fig. 19). From concentrated solutions and melts at rest, lamellae are organized into spherulites (Fig. 21), which may be transformed to row structures by shearing or drawing (Fig. 33).

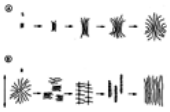

Figure 33. A) Formation of spherulites from lamellae on crystallization of melts at rest; B) Transformation of spherulites into row structures by shearing or drawing in ↕ direction.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Crystallization can be subdivided into two elementary processes: primary nucleation and crystal growth (secondary nucleation). Both processes determine the crystallization rates, which depend strongly on both temperature and polymer structure. At 30 K below the melt temperature, the linear crystallization rate may, for example, range between 5000 µm/min for polyethylene and 0.01 µm/min for poly(vinyl chloride). Symmetrically structured polymers usually crystallize rapidly; polymers with bulky groups or low tacticities, only slowly. Quenching of poly(ethylene terephthalate) melts leads, for example, to amorphous polymers, whereas quenching of polyethylene melts never gives amorphous polymers, even if liquid nitrogen is used.

Primary nucleation may be homogeneous (spontaneous, sporadic, thermal) or heterogeneous (simultaneous, athermal). Homogeneous nuclei are formed from segments of the crystallizing polymer molecules; they are very rare. Heterogeneous nuclei result from extraneous materials such as additives, dust particles, container walls, or specially added nucleation agents. Such nuclei must have minimum sizes of 2 – 10 nm. Their concentrations can range from ca. 1 nucleus per cubic centimeter (polyoxyethylene) to ca. 10^{12} nuclei per cubic centimeter (polyethylene).

Above the melting temperature, fragments of crystallites may survive for certain time periods. These fragments act as athermal primary nucleation agents on subsequent cooling. They are responsible for the “memory effect” (i.e., the reappearance of spherulites at the same locations after and before the melting), which occurs because of low diffusion rates at very high melt viscosities.

Growth of primary nuclei occurs by a *secondary nucleation process*. The growth rate is low just below the melting temperature because secondary nuclei are formed and dissolved rapidly. About 50 K below the glass transition temperature, on the other hand, motions of molecule segments are practically zero and the crystal growth rate is therefore low as well. Crystallization rates must thus exhibit a maximum between the melting and glass transition temperatures; the maximum crystallization rates are usually at (0.80 – 0.87) T_M (in kelvin).

In addition, the entire crystallization process can be subdivided into a primary and a secondary phase. The *primary phase* comprise the conversion of the total volume to a solid. At the end of this phase, the volume may be filled (e.g., with spherulites), but not all polymer segments between spherulites or between the lamellae of the spherulites may have crystallized. The polymer has not attained its maximal crystallinity (the crystallizability). Crystallization may thus continue during the *second phase*. In this aftercrystallization, lamellae may thicken, lattices become more perfect, etc.

Primary crystallization can be characterized by the Avrami equation:

$$\phi/\phi_\infty = 1 - \exp(-z \cdot t^n) \tag{63}$$

where Φ = fraction of crystallized volume, Φ_∞ = fraction of maximal attainable crystallinity for a given entity (e.g., spherulite), and z and n are constants that depend on both the nature of the nucleation process (homogeneous, heterogeneous) and the type of growing entity (rod, disk, sphere, sheaflet, etc.). The exponents n range between 1 and 7; empirically, they may assume fractional values.

6.2.3. *Melting*

Melting is defined as the thermal transition of a crystal to an isotropic melt. The melting temperature T_M (fusion temperature) is defined as the temperature at which crystallites are in equilibrium with the melt. Melting starts at the corners and edges of crystal surfaces; in contrast to crystallization, no nuclei are needed.

Segments of about 60 – 100 chain atoms participate in the melting process. During heating, segments are redistributed continuously between crystalline and noncrystalline regions; a melting range exists and no sharp melting point is observed. The melting temperature is defined as the upper end of the melting range because the biggest and most perfect crystals melt there. Published melting temperatures often refer to the maxima of $\Delta T = f(T)$ curves, however. Observed melting temperatures are in general lower than the thermodynamic melting temperatures of perfect crystals but may be occasionally higher because of overheating effects.

Melting temperatures increase with molar mass and become practically constant at molar masses of ca. 50 000 – 150 000 g/mol. The melt can be considered as a dilute solution of end groups in monomeric units, and the reduction of the melting temperature with decreasing molar mass (i.e., increasing concentration of end groups) can be described by the thermodynamic law for the lowering of freezing temperatures:

$$\frac{1}{T_M} = \frac{1}{T_M^0} + \left(\frac{2R}{\Delta H_{M,u}^m}\right) \cdot \frac{1}{\bar{X}_n} \tag{64}$$

where $\Delta H_{M,u}^m$ is the molar melt enthalpy per monomeric unit, $\bar{X}_n$ the number-average degree of polymerization, and T_M and T_M^o are the thermodynamic melting temperatures at finite and infinite molar masses. Similar depression of the melting point is caused by addition of low molar mass solvents and amorphous polymers or by statistical copolymerization; in all these cases, the term $2/\bar{X}_n$ must be replaced by

$$\left(*V_u^m / *V_1^m\right)\left[\left(1-\phi_2\right) - \chi(1-\phi_2)^2\right]\left(\text{solvent}\right) \tag{65}$$

$$\left(*V_u^m / *V_A^m\right) - \chi(1-\phi_2)^2\right]\left(\text{amorphous polymer}\right) \tag{66}$$

$$\left[-\ln x_u - \chi(1-\phi_2)^2\right]\left(\text{statistical copolymer}\right) \tag{67}$$

where $*V^m$ is the partial molar volume of solvent 1, amorphous polymer A, or monomeric units u of copolymer; x_u the mole fraction of units u; Φ_2 the volume fraction of crystallizable polymer 2; and χ an interaction parameter.

Melting temperatures $T_M = \Delta H_M^m / \Delta S_M^m$ are determined by the changes in molar melting enthalpies ΔH_M^m and molar melting entropies ΔS_M^m. The *melting entropy* results from conformational changes and volume changes upon melting. The melting entropy theoretically adds $R \cdot \ln 3$ = 9.12 J K^{-1} mol^{-1} for the formation of three conformers with equal energy and 7.41 J K^{-1} mol^{-1} for one *trans* and two *gauche* conformers in the case of polymethylene ~ $(CH_2)_n$ ~. The entropy change due to the volume change should add another 10.9 J K^{-1} mol^{-1} so that the theoretical melt entropy of polymethylene should be ca. 18.3 – 20.0 J K^{-1} mol^{-1}. Experimentally, only 9.9 J K^{-1} mol^{-1} is observed, which points toward either the existence of local order in melts or a high segment mobility below the melting temperature. The latter was found experimentally by broad-line NMR for *cis*-1,4-polyisoprene ($\Delta S_M^m = 4.8\ \mathrm{J\,K^{-1}\,mol^{-1}}$).

Melting enthalpies are usually between 1 and 5 kJ per mole of chain atom. Low values are to be expected for polymers with high chain mobilities below the melting temperature (*cis*-1,4-polyisoprene, aliphatic polyesters and polyethers). High values are found for polymers with strong interactions between chains and tight packing of chains in crystals (polyoxymethylene, it-polystyrene). Some of these strong interactions may survive the melting process; for example, most of the hydrogen bonds of polyamides are still detected by IR spectroscopy above the melting temperature.

Thus the primary factors for high melting temperatures are not intermolecular interactions (e.g., cohesive energies) but reduced flexibilities of chains. Low melting temperatures are found for polymers with low rotational barriers (ester, oxygen, sulfide groups in chains), high melting temperatures for tightly packedhelices [polyoxymethylene, it-poly(3-methylbutene)] and for ladder and ladderlike polymers [poly(*p*-phenylene), polybenzimidazole, etc.]. Such factors are responsible for the variation of melting temperatures with the number of methylene units in aliphatic polymer chains of the type ~X– $(CH_2)_n$~ (Fig. 34).

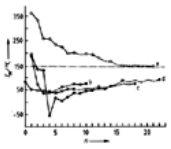

Figure 34. Melting temperatures of (○, ⊙, ⊕) aliphatic polymers ~ [X-$(CH_2)_n$]~ and (•) isotactic poly(α-olefins) ~ {CH_2–CH[$(CH_2)_n$H]}~ as function of the length n of methylene sequences

a) X = NHCO (polyamides); b) X = O (polyoxides); c) X = $COO(CH_2)_3OOC$ (aliphatic polyesters of trimethylene glycol)

– · – · – · theoretical melting temperature of polyethylene

6.2.4. *Liquid Crystal Transitions*

Thermal transitions of thermotropic LC polymers from their crystals to smectic (T_{cs}) or nematic phases (T_{cn}), from smectic to nematic mesophases (T_{sn}), and from nematic phases to isotropic melts (T_{ni}, clearing temperature) are thermodynamic first-order transitions. They exhibit steplike changes in volume, enthalpy, and entropy, just like the melting of three-dimensional crystals to isotropic melts. Thermodynamically stable phases can exist only between melting and clearing temperatures (T_{cs}, $T_{cn}<T_{ni}$); they are called *enantiotropic phases*.

Mesophases form dispersions in supercooled isotropic melts if clearing temperatures are lower than melting temperatures (T_{cs}, $T_{cn}>T_{ni}$). Such phases are thermodynamically

unstable compared to the crystalline; they are called *monotropic*.

Mesophases may be supercooled below $T_M = T_{cs}$ to smectic liquids sL* and below T_{sn} to nematic liquids nL* if crystallization can be suppressed. At even lower temperatures T_{gn} and T_{gs}, these supercooled liquids may yield anisotropic glasses nG and sG, respectively. Some of these transition temperatures cannot be measured directly, but their existence can be deduced from extrapolations of transition temperatures of copolymers to 100 % of the pure mesogenic compound (virtual transition temperatures).

The transition temperatures T_{trans} (i.e., T_{cs}, T_{gs}, T_{sn}, T_{cn}, T_{gn}, T_{ni}) depend on the degree of polymerization X in the same way the melting temperature does: $1/T_{trans} = f(1/X)$. The transition enthalpy ΔH_{ni} is always lower than the transition enthalpies ΔH_{gs} and ΔH_{sn} because the n $\longrightarrow$ i transition is from order to disorder, whereas the g $\longrightarrow$ s and s $\longrightarrow$ n transitions are from order to less order. Transition entropies of mesophases are lower than melt entropies; they usually have values of ca. 0.5 – 1.5 J K^{-1} mol^{-1}, with $\Delta S_{sn} < \Delta S_{ni}$.

6.2.5. *Glass Transitions*

Glass transitions are phenomenologically characterized by a change from a “hard,” noncrystalline, glasslike material to a rubbery to highly viscous “melt.” The viscosities at glass transitions are ca. 10^{12} Pa · s, independent of chemical structures. Glass transitions were thus thought to be “isoviscous” phenomena. Today, glass transitions are considered to occur at that physical state where all materials exhibit the same “free volume.”

Various free-volume fractions are discussed in the literature. The empirical Boyer – Simha rule relates a free-volume fraction f_{exp} to the cubic expansion coefficients β of liquid (L) and amorphous, glasslike (G) polymers and their glass transition temperatures T_G (see Table 21):

$$f_{exp} \approx (\beta_L - \beta_G) \cdot T_G \approx 0.11 \pm 0.02 \quad (68)$$

Table 21. Glass transition temperatures and free-volume fractions of polymers (for explanation of symbols, see text)

Polymer	T_G °C	f_{exp}	f_{WLF}	f_{fluc}
Polyethylene	– 80	0.098	0.025	
Polyisobutylene	– 73	0.079	0.026	0.0017
Poly(butyl methacrylate)	20	0.13	0.026	0.0010
Poly(vinyl acetate)	27	0.128	0.028	0.0023
Polystyrene	100	0.133	0.025	0.0035
Poly(methyl methacrylate)	105	0.118	0.025	0.0015

The Williams – Landel – Ferry (WLF) approach relates a free-volume fraction f_{WLF} to the probability of segment movements. Empirically, values of $f_{WLF} \approx 0.025 \pm 0.01$ were found; these values can be calculated, for example, from $K = f_{WLF}/(\beta_L - \beta_G)$ and $K' = \log e/f_{WLF}$ of the semiempirical WLF equation (Eq. 69) where $-\log a_t = \Delta(\log t)$ is a shift factor

$$T = T_G + \frac{K \cdot \log a_t}{K' - \log a_t} = T_G + \frac{51.6 \cdot \log a_t}{17.4 - \log a_t} \quad (69)$$

The Williams – Landel – Ferry equation applies to all relaxation processes; its use is restricted to temperatures in the range $T_G < T < (T_G + 100\ K)$.

The WLF equation allows calculation of the static glass transition from the various dynamic glass transition temperatures if the deformation times (inverse effective frequencies) of the methods are known. The glass transition temperatures of poly(methyl methacrylate) (PMMA) are given as 105 °C (thermal expansion, “static”), 120 °C (penetrometry), and 160 °C (rebound elasticity). The same polymer may thus exhibit very different mechanical properties if subjected to different stresses; at 140 °C, PMMA behaves as either a glass (rebound elasticity) or an elastomer (penetrometry).

The glass transition temperature indicates the onset of cooperative movements of chain segments of 25 – 50 chain atoms, which can be deduced from the ratio of molar activation energies and melt energies. These cooperative movements very probably involve *trans – gauche* transitions that proceed cooperatively along greater distances since only small changes of chain axes are involved according to deuterium NMR. The participation of segments of 25 – 50 chain atoms is also indicated by cross-linking experiments: as long as the average segment length N_{seg} between two cross-linking points is less than 25 – 50 chain atoms, no change of glass transition temperatures is observed. At $N_{seg} < 25 – 50$, T_G increases with the inverse molar mass of the segments.

Since both glass transition and melt temperatures depend on segmental motions, close relationship between these two temperatures can be expected. The empirical Beaman – Boyer rule states that $T_G \approx (2/3) T_M$, which holds reasonably well for many polymers except for chains such as polyethylene and polyoxyethylene for which $T_G/T_m \approx 1/2$.

A vast literature exists about effects of constitution on T_G. Cyclic macromolecules possess no end groups, and thus no free-volume effects from these. Small rings are furthermore strained (less possible microconformations). The greater the molar mass, the more microconformations can be adopted, the greater is the chain flexibility and the lower is T_G. The same is true for segment flexibilities of star-branched polymers and long side chains in comblike molecules (side- chain “crystallization”).

Linear relationships are found between the logarithms of glass transition temperatures T_G and the logarithms of cross-sectional areas A of either carbon, carbon – oxygen, or carbon –nitrogen chains (another measure of segment flexibilities). The three lines intersect at $A = 0.17\ nm^2$ and $T_G = 141$ K, which should be the lowest glass transition temperature possible. The lowest experimentally found glass transition temperature (150 K) is that of polydimethylsiloxane, ~ $(O – Si(CH_3)_2)_n$ ~.

Glass transitions can be decreased (or increased) by copolymerization with suitable monomers (internal plasticization) and by addition of external plasticizers (see Section Plasticizers).

6.2.6. *Other Transitions and Relaxations*

Experimentally, a number of other transition – relaxation temperatures are observed, mostly of unknown origin. Amorphous polymers exhibit weak “liquid – liquid” transitions at ca. $T_{II} \approx 1.2\ T_G$. Below the critical molar mass for entanglements, transition temperatures T_{II} equal flow temperatures T_F at which polymers start to flow under their own weight. At higher molar masses, $T_F > T_{II}$.

Another transition temperature $T_U \approx 1.2\ T_M$ seems to exist for crystalline polymers. This transition has been interpreted as the dissolution of smectic structures.

Few β-relaxations have been correlated with molecular phenomena. An example is the frequency-dependent boat – chair transition of cyclohexane rings, which occurs at e.g., – 125 °C (10^{-4} Hz) and + 80 °C (10^5 Hz).

6.2.7. *Technical Methods*

The technical testing on thermal transitions and relaxations of plastics is usually performed with simple methods under standardized conditions and always under load. *Martens numbers* measure temperatures at which the specimen has experienced a certain bend, *Vicat softening temperatures* give the temperatures for a certain penetration of a rod into the plastics, and the *heat distortion temperatures* (heat deflection temperatures) indicate the temperatures for a certain bending with a three-point method. The temperatures from these three methods do not only depend on transitions or relaxations but also on the elasticity of the specimen; Vicat and heat distortion temperatures are in addition affected by the surface hardness. The resulting softening temperatures are neither identical with glass transition nor with melting temperatures; they are often also not a good measure of the continuous service temperature of a plastic.

6.3. Transport

6.3.1. *Self-Diffusion*

Brownian movements cause molecules and their segments to interchange positions in fluid phases. If all entities are of the same type, such interchanges lead to a “self-diffusion”, which involves no net transport of polymers. Self-diffusions can be measured by pulsed field-gradient spin-echo NMR (segments) and radioactive tracers (molecules).

Coiled molecules behave in their melts as unperturbed coils with low coil densities, which are filled with segments of other coils. Self-diffusion of a segment must therefore occur by interchanging position with a segment of another molecule.

Self-diffusion coefficients decrease with the squares of molar masses (Fig. 35). According to *reptation theory*, this functionality is caused by temporary (but fairly long-lived) entanglements of polymer chains. Such entanglements cause the sudden change of molar mass dependences of melt viscosities above a critical molar mass M_c (Fig. 35).

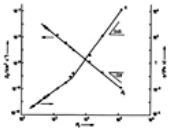

Figure 35. Segmental self-diffusion coefficients D_2 (○) and Newtonian melt viscosities η of alkanes and narrow-distribution polyethylenes (•) as function of the relative molecular mass M_r at 175 °C

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The resulting topological restraints make the polymer chain move through the maze of other segments like a reptile through brush. According to the Doi – Edwards theory, the test chain reptates in a tube of ca. 5-nm diameter, which is formed by other segments (Fig. 36). The theory predicts the experimentally found function $D_2=f(M^{-2})$ albeit only for $M>M_c$. Experimentally, the same function is also observed for $M<M_c$, probably because end groups cause successively greater free-volume fractions, which are assumed by the reptation theory to be molar mass independent.

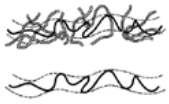

Figure 36. Reptation of a test chain (black) through the segments of a matrix (white). The “walls” of the tube are indicated by dotted lines.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The same dependence of self-diffusion coefficients of test chains on the squares of their molar masses is theoretically predicted and experimentally found for test chains in higher molar mass matrices of the same constitution and configuration ($M_{test} << M_{matrix}$); the molar mass of the matrix exerts no influence. If however $M_{test} >> M_{matrix}$, then a function $D_{test}=f[(M_{test})^{-1/2}\cdot(M_{matrix})^{-1}]$ is predicted and observed.

6.3.2. *Permeation*

The transport of extraneous material through polymers is called permeation. The resulting net flow of mass is caused by differences in chemical potentials, that is, concentration differences (at constant temperature) or thermal gradients (at constant concentration). Permeation may be desirable as in the dyeing of textiles or the controlled transdermal delivery of pharmaceuticals, or undesirable as in the loss of carbon dioxide from plastic bottles for carbonated soft drinks or in migration of plasticizers.

Permeation of chemical compounds through amorphous polymers below their glass transition temperatures or through crystalline polymers below their melting temperatures can occur either by flow through pores or by molecular transport. Pores have diameters much greater than the diameters of permeating substances (diameter of spheres, cross section of coillike chains); the interactions of permeants and pore walls are negligible. Molecular transport, on the other hand, depends on such interactions between permeant and matrix (i.e., on the solubility of the former in the latter). Both types of transport can be distinguished by their temperature dependence: permeation coefficients of gases decrease with temperature at pore membranes; they increase with temperature at solubility membranes.

The counteraction of the two types of permeation can be utilized in lamination. Oxygen permeates through pores in aluminum films but by molecular transport through polyethylene films. At 1 bar, permeation rates decrease from 5×10^{-5} cm^3/s for 0.025-mm-thick aluminum films to 5×10^{-13} cm^3/s for the same films laminated with 0.025-mm-thick PE films.

The permeation coefficient P is given by the product of the diffusion coefficient D and the solubility coefficient S of the permeant in the matrix and, in the steady state, by the expression to the right of the second equality sign:

$$P = D\cdot S = \frac{Q\cdot L}{A\cdot\Delta p\cdot t} \tag{70}$$

where Q = permeated amount, L = thickness of film (membrane, etc.), A = area, Δp = pressure difference, and t = time. The literature often uses different “practical units” for the various quantities of Equation (70), for example, Q in cm^3, L in mm, A in m^2, Δp in atm, and t in 24 h; P is then given in ($cm^3\cdot mm$)/(24 h m^2 atm). If like units are used, then the unit of P is of course $length^2\cdot time^{-1}\cdot pressure^{-1}$ and the unit of S is $pressure^{-1}$. Permeation coefficients P^* of liquids are usually measured without a pressure differential; their unit is cm^2/s and the unit of S is 1.

Permeation coefficients of gases in polymers vary widely; for example, oxygen permeates 10 million times faster through polydimethylsiloxane than through polyacrylonitrile (Table 22). Gases in general, have lower permeation coefficients in thermoplastics than in elastomers because segmental movements of the former are frozen-in below the glass transition temperature. Bulky substituents, orientation of polymer segments, crystalline domains, and added fillers all increase the pathway for a gas molecule through a polymer matrix; these tortuosity factors decrease permeation coefficients.

Table 22. Permeation coefficients of gases (P) and water vapor (P^*) through polymers at 30 °C [a]: $P=1\times10^{-14}$ $cm^2\ s^{-1}\ Pa^{-1}$ corresponds to $P^*=1\times10^{-9}$ cm^2/s at normal pressure ($p=1\times10^5$ Pa)

Polymer	$\frac{10^{14}P}{cm^2s^{-1}Pa^{-1}}$		$\frac{10^9P^*}{cm^2s^{-1}}$
	O_2	CO_2	H_2O
Polydimethylsiloxane	25 000	85 000	40
cis-1,4-Polyisoprene	2 000	10 000	0.3
Butyl rubber	100	500	0.1
Polystyrene			
Regular	200	1 000	1
Biaxially oriented	0.1	100	0.5
Poly(ethylene terephthalate)			
Regular	4	20	0.2
Biaxially oriented	0.2	1	0.2
Poly(vinylidene chloride)	0.05	0.15	0.02
Cellulose	0.03	0.1	10
Polyacrylonitrile	0.002	0.02	0.02
Required for bottles			
Cola	1	0.5	0.14

Beer	0.05	0.5	0.14

[a]Literature values vary widely because polymers are often not identical with respect to crystallinity, orientation, water absorption, etc.

The *permeation of nondissolving liquids* through a polymer is proportional to t^n. The exponent *n* depends on the ratio of the relaxation time of the polymer – solvent system to the diffusion time of the solvent (i.e., on the Deborah number *DB*).

Three different regimes are normally considered. In the regime denoted as *Case I*, the mobility of the permeant is much smaller than the relaxation of the polymer segments (*DB*<0.1). The movement of the permeant causes "instantaneous" conformational changes of the polymer segments. Both permeant and polymer behave as viscous liquids: the system can be described by Fickian diffusion laws (i.e., *n*=1/2).

In the *Case II* regime, the mobility of the permeant is much higher than the relaxation of the polymer segments (*DB*>10). The physical structure of the polymer does not change during the permeation; the polymer appears to the permeant as an elastic body. Case II is characterized by a sharp demarcation line between the glassy inner polymer core and the swollen zone advancing with constant speed. The permeating amount is directly proportional to time (i.e., *n*=1).

Relaxation and permeation become comparable for 0.1<*DB*<10. This *third regime* is usually called anomalous diffusion or viscoelastic diffusion (1/2<*n*<1).

[Top of Page]

7. Rheological Properties

7.1. Introduction

Materials exhibit two limiting types of behavior against deformation. Typicial liquids such as water flow under their own weight and are irreversibly deformed (viscous behavior). Typical solids such as iron resist deformation; they return from small deformations to their former states after removal of loads (elastic behavior). Polymers commonly combine both types of behavior: they are viscoelastic materials. Their melts exhibit viscous behavior at small deformations and elastic properties at larger ones. Polymer solids respond elastically at small deformations but begin to flow at larger ones.

Polymer melts and solutions have extremely high viscosities which may, in addition, be dependent on deformation rates and duration of the experiment. Air has, for example, a viscosity of 10^{-5} Pa · s; water, 10^{-3} Pa · s; and glycerol, 1 Pa · s, whereas polymer melts exhibit viscosities of ca. 10^2–10^8 Pa · s.

Three types of viscosity are usually distinguished:

1. Shear viscosity describes the rate of shear flow as function of the applied stress.
2. Extensional viscosity measures the rate of extensional (elongational) flow as function of tensile stress.
3. Bulk viscosity relates the rate of deformation of volume to the applied hydrostatic pressure.

Shear viscosities are the most often studied rheological properties of polymers; they are of great importance for the processing of plastics by extrusion or injection molding, for example. Much less is known about *extensional viscosities* (important for blow forming and fiber spinning) and practically nothing about the *bulk viscosities* of polymers.

7.2. Shear Viscosities at Rest

7.2.1. *Fundamentals*

Nine different shear stresses may be assigned to a three-dimensional body: one parallel to each of the three spatial directions (σ_{11}, σ_{22}, σ_{33}) and six perpendicular to these (σ_{12}, σ_{13}, etc.). A body is by definition sheared in the 2 – 1 direction. The ratio of shearing force *K* to contact area is called the shear(ing) stress $\sigma_{21}=K/A$; it produces a shear strain γ. The ratio $G=\sigma_{21}/\gamma$ is the shear modulus. Between layers of distance *y* moving parallel to each other with different rates *v*, a shear gradient $\dot{\gamma}=dv/dy$ thus exists. The ratio of shear stress to shear rate is the (dynamic) viscosity (Newton's law);

$$\eta = \sigma_{21}/\dot{\gamma} \tag{71}$$

its inverse $1/\eta$ is called fluidity.

Viscosities are independent of shear rates for Newtonian liquids ($\eta=\eta_o$); these viscosities are also called zero-shear viscosities, viscosities at rest, or stationary viscosities. For Newtonian liquids, the shear modulus G_o is independent of the extent of deformation. η_o and G_o are true material constants, whereas η and *G* depend on shear rates and sometimes also on shearing time.

Shear stresses and shear gradients (and thus viscosities) can be measured with a variety of instruments that usually belong to one of three groups: capillary, rotatory, and cone – plate viscometers. A number of industrially used instruments provide viscometric indicators (but neither shear stress, shear gradient, nor viscosity); this group includes Höppler viscometers, Cochius tubes, Ford beakers, and instruments that measure melt flow indices or Mooney values (Mooney viscosities).

Thermoplastics are usually characterized by their *melt flow indices* $MFI_{T/F}$. The melt flow index measures the mass of polymer extruded in 10 min by a standard load *F* from a standard plastometer at temperature *T*. It is a measure of the (usually non-Newtonian) fluidity. The higher the melt flow indices, the lower are the molar masses of polymers of the same constitution.

The Mooney "viscosity" is really a measure of elasticity; it is used mainly for elastomers but also for polymer melts. A polymer is deformed in a standardized cone – plate viscometer at constant rotational speed and constant temperature *T*; after *t* minutes, the force to recover is read.

7.2.2. *Molar Mass Dependence*

Newtonian viscosities exhibit two different regimes for their (mass-average) molar mass dependencies $\eta_o=K \cdot M^a$: a weaker dependency (*a*=1/2 – 1) for low molar masses and a stronger one (*a*≈3.4) at higher molar masses (Fig. 35). The transition from one to the other is thought to be caused by the onset of molecular entanglements that cause the molecules to behave as physically cross-linked networks.

The number of entanglements can be assumed to be constant at low shear rates. The elasticity of such a network is described by its shear modulus *G* (unit of pressure). Because shear viscosities have the unit pressure · time, $\eta_o=G_o \cdot t$. The reptation model identifies the time *t* as the time required for a chain to leave the tube. This time is proportional to the third power of the number of segments per molecule (i.e., the viscosity should be proportional to the third power of the molar mass). The deviation between the theoretically predicted α=3 and the experimental value of α=3.4 is thought to be due to "breathing" of the tube. Breathing pushes nonentangled chain loops back into the surrounding matrix; chain ends cause additional relaxations and the tube length decreases.

7.2.3. *Concentrated Solutions*

Newtonian viscosities of concentrated solutions increase with both solute concentrations *c* and molar masses *M* or, since $[\eta]=K \cdot M^a$, with intrinsic viscosity $[\eta]$ as well. The concentration *c* measures the mass of polymer per unit volume of solution; the intrinsic viscosity, the volume of polymer molecules per unit mass of polymer. The product $c \cdot [\eta]$ is thus a measure of the volume fraction of polymer molecules that would be occupied by isolated polymer coils. At higher concentrations, coils start to overlap and the total occupied volume is smaller than the one demanded by isolated coils (i.e. $c \cdot [\eta]>1$).

At low values of $c \cdot [\eta]$, η_i (the "specific viscosity") and $c \cdot [\eta]$ are proportional (Fig. 37), where $\eta_i=(\eta/\eta_1)-1$, η is the viscosity of the polymer solution at rest, and η_1 the viscosity of the solvent. At higher $c \cdot [\eta]$ values, $\eta_i \sim (c \cdot [\eta])^q$ or, since $[\eta] \sim M^a$, $\eta_i \sim c^q \cdot M^{aq}$. Because at very high concentrations $\eta_i \approx \eta/\eta_1$ and η approaches the Newtonian melt viscosity $\eta_o \sim M^{3.4}$ for high molar masses, $a \cdot q$=3.4 is obtained. In theta solvents and melts, coils are unperturbed (*a* = 1/2) and *q*=6.8 (Fig. 37). In good solvents, *a* ⟶ 0.764 and *q*=4.55. These relationships are independent of the chemical nature of polymers and solvents.

Figure 37. Relative viscosity increment ("specific viscosity") as function of $c \cdot [\eta]$ for (○) polystyrenes in *trans*-decalin (theta solvent) or toluene (good solvent G) at 25 °C; (•) *cis*-1,4-polyisoprene in toluene at 34 °C; and (△) hyaluronates in water at 25 °C

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

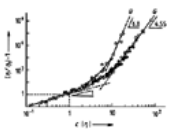

7.3. Non-Newtonian Shear Viscosities

7.3.1. *Overview*

The shear stress σ_{21} is directly proportional to the shear gradient $\dot{\gamma}$ for Newtonian liquids; these relationships and thus the viscosities η are moreover, independent of the duration of the experiments. Non-Newtonian viscosities (apparent viscosities), on the other hand, vary with shear rate and sometimes even with time.

Various dependencies of (apparent) viscosity on shear rate are found for time-independent non-Newtonian liquids (Fig. 38). *Plastic bodies* (Bingham bodies) exhibit a yield value *y*, i.e., shear stresses have finite values σ_0 at $\dot{\gamma} \longrightarrow 0$. Above σ_0, such bodies may behave in a Newtonian (ideal Bingham bodies) or non-Newtonian (pseudoplastic Bingham bodies) manner. This behavior seems to be due to the disappearance of aggregates. An example for a Bingham body is tomato ketchup.

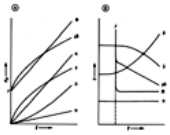

Figure 38. A) Shear stress σ_{21} and B) shear viscosity η as a function of shear rate $\dot{\gamma}$ for Newtonian (N), dilatant (D), and pseudoplastic (P) liquids and for ideal (iB) and pseudoplastic (pB) Bingham bodies (*y*=yield)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

In *dilatant liquids*, the shear stress increase is more than linearly proportional to shear rate; viscosities increase with shear rate (shear thickening). This behavior occurs frequently in polymer dispersions.

A decrease of apparent viscosity with shear rate is most common for polymer melts. Because such behavior resembles that of pseudoplastic Bingham bodies, it is called *pseudoplasticity* in the English literature ("structural viscosity" in German), although such "pseudoplastic" materials do not possess a yield value. Pseudoplastic behavior eases polymer processing from melts and reduces the energy required (e.g., lower pressures can be applied in injection molding). It is usually accompanied by an orientation of chain segments that is most pronounced in the processing of liquid crystalline polymers with rigid mesogens and the ultradrawing of flexible polymers.

Dilatant and pseudoplastic liquids are characterized by "instantaneous" adoption of shear rates on application of shear stresses (i.e., by time-independent apparent viscosities). *Thixotropic materials* exhibit a decrease of apparent viscosity with time at constant shear rate; examples are dispersions of bentonite and other platelet-like silicates. *Rheopectic* or *antithixotropic* materials show an increase of apparent viscosity with time at constant shear rate. The flow behavior may be further complicated by wall effects. Certain dispersions and gels exude liquid on application of a shear stress. The liquid acts as an external lubricant and a pluglike flow results (e.g., tooth pastes). An additional complication may result from the onset of turbulence, which usually occurs at much lower Reynolds numbers in non-Newtonian than in Newtonian liquids.

7.3.2. *Flow Curves*

A plot of log $\dot{\gamma}$=*f* (σ_{21}) or vice versa is usually called a flow curve. A generalized flow curve may contain an initial Newtonian region, followed by pseudoplasticity and a second Newtonian regime. Since experiments are difficult to conduct at high shear rates, many rheologists doubt the existence of true second Newtonian regions. Dilatancy may or may not be present. Finally, turbulence sets in and melt fracture occurs (Fig. 39).

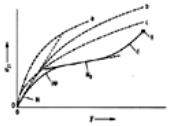

Figure 39. Generalized flow curve with first Newtonian region (N), pseudoplasticity (pp), second Newtonian region (N_2), dilatancy (d), and onset of turbulence (t)

- - - Schematic representations of various viscosity "laws": a) Prandtl – Eyring; b) Rabinowitsch – Weissenberg; c) Ostwald – de Waele

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Several empirical laws have been suggested for the description of flow curves, all of which usually apply only to limited ranges of shear rates. Examples include

$$\dot{\gamma} = a \cdot (\sigma_{21})^m \text{ Ostwald – de Waele} \quad (72)$$

$$\dot{\gamma} = b \cdot \sinh(\sigma_{21}/d) \text{ Prandtl – Eyring} \quad (73)$$

$$\dot{\gamma} = f \cdot \sigma_{21} + g \cdot (\sigma_{21})^3 \text{ Rabinowitsch – Weissenberg} \quad (74)$$

where *a, b, d, f, g*, and *m* are empirical constants. The exponent *m* is known as the "flow exponent"; it takes a value of 1 (Newtonian liquid) and < 1 (pseudoplasticity). No constitutive equations are known that cover the entire range of rheological phenomena.

7.3.3. *Melt Viscosities*

Polymers below the critical molar masses M_c for entanglements show extended Newtonian ranges. For $M>M_c$, non-Newtonian behavior appears at lower shear rates as molar mass increases (lower melt flow index) (Fig. 40). At high shear rates, a constant exponent *q* is approached for the function η=*f* ($\dot{\gamma}^q$). The non-Newtonian viscosities are no longer proportional to the mass-average molar masses; the broader the molar mass distribution, the more the number average seems to be the correct corresponding quantity.

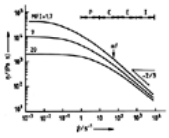

Figure 40. Dependence of melt viscosities of polyethylenes with different melt flow indices (MFI, inversely proportional to molar mass) on shear rates (mf=onset of melt fracture for the highest molar mass polyethylene)

The ranges of the various processing regimes overlap: P=compression molding, C= calendering, E=extrusion, I=injection molding. Shear rates refer to those at orifices and are much lower in the mold (tool, die).

Reprinted with permission by BASF AG, Ludwigshafen [78]

The decrease of apparent viscosity with shear rate is very important for plastics processing. Viscosities describe a frictional behavior: the higher the viscosity, the higher is the internal friction of the melt and the greater is the proportion of energy provided that is converted into heat. Strong non-Newtonian behavior thus saves energy. Polymer melts are therefore processed at the highest possible shear rates. The upper range is given by the processing method (calendering, extrusion, etc.), and the polymer properties (thermal degradation, melt fracture, etc.).

Surface roughnesses of barrel walls, diameter changes, etc., create additional rate components in extrusion that are dampened by the viscosity of the liquid. These disturbances become stronger with increasing flow rate and can no longer be dampened at high shear rate. Finally, turbulence sets in. In entangled polymer melts, additional elastic vibrations occur due to the presence of physical cross-links. The resulting elastic turbulences lead to rough surfaces of the extrudate, which are subsequently frozen-in upon exit from the orifice. The polymer surface appears "fractured"; "melt fracture" thus does not refer to a breakage of the extrudate strand.

Molecular coils are deformed if their melt is pressed through an orifice (Fig. 41). For $M>M_c$, segments can no longer slip from each other at high stresses and short times because of entanglements; a "normal" stress builds up perpendicular to the stress direction. At the die exit, this stress is relieved and the coil returns to the thermodynamically more favorable shape of an unperturbed coil. The melt expands perpendicular to the flow direction. This phenomenon is known as Barus effect or memory effect (melts), the Weissenberg effect (solutions), parison swell (extrusion), swelling (blow molding), etc. It is especially pronounced for high molar mass tails in polymers with $M>M_c$ because of the 3.4 power dependence of η on $\bar{M}_w$.

Figure 41. Parison swell upon extrusion (Barus effect)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Negative Barus effects are known for solutions of rodlike molecules or LCPs with rodlike mesogens (diameter of strand smaller than diameter of orifice). If such molecules crystallize after exiting from the die, the strand contracts perpendicular to the extrusion direction and the strand diameter becomes smaller than the diameter of the orifice.

The effect of normal stress can be determined by the Bagley equation. A force $F_f = \pi R^2 p$ is exerted on a liquid during the flow through a capillary with radius R and length L under a pressure p. It is counteracted by a frictional force $F_r = 2\pi RL\sigma_{21}$. In steady state, $F_f = F_r$; thus $p = 2\sigma_{21}(L/R) = 2\eta \cdot \dot{\gamma} \cdot (L/R)$ since $\sigma_{21} = \eta \cdot \dot{\gamma}$. In the Bagley diagram, pressure p is accordingly plotted at constant shear rate against die geometry L/R (Fig. 42). For non-Newtonian liquids, a relationship

$$p = p_o + K \cdot (L/R) \tag{75}$$

is found. The intercept p_o at $L/R \longrightarrow 0$ is identified with the pressure loss caused by the elastically stored energy of the flowing melt and the formation of a steady-state flow profile at both ends of the capillary (die). The higher L/R is, the higher must the applied pressure be. Die lengths should thus be as short as possible.

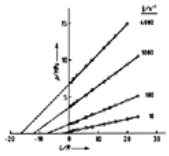

Figure 42. Bagley diagram for a high-impact polystyrene at 189 °C and different shear rates

The intercept at $L/R \longrightarrow 0$ is the pressure correction; the intercept at $p \longrightarrow 0$, the Bagley correction.a) $\dot{\gamma}$=4000 s^{-1} ; b) $\dot{\gamma}$=1000 s^{-1} ; c) $\dot{\gamma}$=100 s^{-1} ; d) $\dot{\gamma}$ = 10 s^{-1}

Reprinted with permission by BASF AG, Ludwigshafen [78]

7.4. Extensional Viscosities

7.4.1. *Fundamentals*

Polymer liquids can be elongated considerably without being broken. This extensibility allows fiber spinning from melts and solutions, blow molding of hollow bodies, vacuum forming of parts, etc.

The extensional viscosity η_e (elongational viscosity) is given by the ratio of tensile stress σ_{11} in draw direction 11 to the elongational rate $\dot{\varepsilon}$

$$\eta_e = \sigma_{11}/\dot{\varepsilon} \tag{76}$$

The type of deformation must always be indicated for extensional viscosities, contrary to shear viscosities. The three principal deformation rates can be defined in such a way that $\dot{\varepsilon}_{11} \geq \dot{\varepsilon}_{22} \geq \dot{\varepsilon}_{33}$. The ratio $m = \dot{\varepsilon}_{22}/\dot{\varepsilon}_{11}$ characterizes the special type of elongational flow: theory gives $m = -1/2$ for uniaxial elongation, $m = 1$ for equal-biaxial, and $m = 0$ for planar (pure shear). Uniaxial elongational viscosities are important for fiber spinning; they are the only ones used to characterize fluids. Biaxial elongational viscosities play an important role in blow and vacuum forming; very little is known about them.

The elongation in Equation (76) is the true strain $\in' = \ln(L/L_o)$ (Hencky strain) and not the nominal strain $\in = (L - L_o)/L_o$ (Cauchy strain, engineering strain). The rate of elongation is thus $\dot{\varepsilon} = d\in/dt = d \ln L/dt = L^{-1}(dL/dt)$.

Extensional viscosities are very difficult to measure. Elastomers and melts of entangled polymer coils can be stretched between rotating rollers. Extensional viscosities of solutions can be determined if two liquid jets streaming toward each other are redirected by rollers or siphoned off.

7.4.2. *Melts*

Extensional and shear viscosities depend very differently on deformation rates. At low rates, extensional viscosities are independent of the rate of extension (Fig. 43). The uniaxial extensional viscosity at rest (historically called "Trouton viscosity") is three times the Newton viscosity $[(\eta_e)_o = 3(\eta_s)_o]$, whereas the biaxial extensional viscosity at rest is six times its shear counterpart.

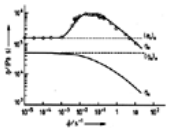

Figure 43. Shear viscosity η_s as function of shear rate $\dot{\gamma}$, and extensional viscosity η_e as function of the uniaxial extensional rate $\dot{\varepsilon}$ of a polyethylene at 150 °C

Reprinted with permission by Steinkopff Verlag, Darmstadt [79]

Above a critical deformation rate, shear thinning is observed for the shearing and extension of melts of linear coil molecules. At the same rate, melts of branched coil molecules become dilatant on extension. The apparent extensional viscosities then pass through a maximum and decrease with further increase of extension rates. The maximum of the extensional viscosity increases with broader molar mass distribution and increased long chain branching, which indicates the strong influence of entanglements on extensional viscosities.

7.4.3. *Solutions*

Rodlike molecules and segments are increasingly oriented in the extensional direction with increasing elongational rates. The molecular axes are no longer distributed at random, and the solution becomes anisotropic and thus birefringent. A limiting value is asymptotically reached if all molecular axes are completely aligned in the flow direction.

Flexible molecules in solution are only slightly deformed and oriented at comparable extension rates because elastic (entropic) forces cause the chains to return to the thermodynamically favored coil shape. At high critical extension rates, these retraction forces can be overcome and the chain axes become oriented in the flow direction. Only an incremental increase of extension rates is needed to orient the chains completely. Molecules are stressed more and more above the critical extension rates until they finally break. This fracture occurs primarily in the middle of the molecule so that degradation products possess 1/2, 1/4, 1/8, etc., of the initial molar mass. Such degradation occurs under very low deformation rates for rigid macromolecules, for example, during pipetting of dilute solutions of high molar mass deoxyribonucleic acids.

These degradations by extensional flow are not caused by turbulence because they happen at lower Reynolds numbers than those of pure solvents. Chain degradations by turbulence do occur on shearing of very dilute solutions of flexible coil molecules (e.g., 10^{-4} g/mL aqueous solutions of polyoxyethylenes). The degradation reduces the frictional resistance of liquids up to 75 %. This "Toms effect" by added small amounts of such polymers eases the flow of crude oil through pipelines and increases the distance and height to which water can be directed at fires.

[Top of Page]

8. Mechanical Properties

8.1. Introduction

8.1.1. *Deformation of Polymers*

Mechanical properties of a polymer include the deformation of bulk polymers or their surfaces, the resistance to such deformation, and the fracture under static or dynamic loads. Deformations may be reversible or irreversible; they can be caused by drawing, shearing, compression, bending, and torsion, as well as by combinations of these.

Reversible deformations are due to the presence of elasticity. Irreversible deformations are also called inelastic; they are further subdivided into deformation by viscous flow, plasticity, phase transformations, craze formation, cracking, viscoelasticity, creep, etc. An inelastic deformation of metals by viscoelasticity is known as anelasticity; in polymer physics, anelasticity denotes a reversible elasticity with retardation, which does not lead to energy dissipation. Deformation of the upper layers of a polymer body is characterized by its "hardness," which influences friction and abrasion.

The term elasticity may refer to either energy or entropy elasticity. These elasticities differ in their molecular mechanisms and the resulting phenomenological behavior. In *energy-elastic bodies* (steel, plastics, and elastomers at low strains), torsion and bond angles are changed and bond distances are enlarged on deformation, whereas the

macroconformations of chains, for example, remain basically the same. A deformation of *entropy-elastic bodies* (elastomers at high strains), on the other hand, leads to entropically unfavorable positions of chain segments which, however, cannot slip irreversibly from each other because of their cross-linking ("rubber elasticity"). The deformation thus changes the macroconformation (molecular picture), decreases entropy (thermodynamics), and creates normal stresses (mechanics).

These molecular changes are reflected in the properties of energy- and entropy-elastic materials:

	Energy-elastic	Entropy-elastic
Elastic moduli	large	small
Reversible deformation	small	large
Temperature change on deformation	cooling	warming
Length change on heating	expansion	contraction

The deformation behavior of plastics depends on the molar mass of their constituting polymers and on the testing temperature: That is, whether the molar mass is greater than the one needed for the establishment of entanglements between chains ($M>M_{ent}$) and whether the testing temperature is lower than the glass transition temperature of the polymers ($T<T_G$). All mechanical deformations recover for $M>M_{ent}$ and $T<T_G$ since chain entanglements cannot reorganize below the glass transition temperature (memory effect). Yielded plastics and crazed plastics recover on (not too long) heating above the glass transition temperature. These plastics thus do not show true plastic flow.

8.1.2. *Tensile Tests*

Mechanical properties of polymers are most commonly evaluated by tensile tests in which a specimen is drawn with constant speed. The tensile stress σ_{11} is recorded as a function of time t, draw ratio (strain ratio) $\lambda=L/L_o$, or tensile strain (elongation) $\epsilon=(L-L_o)/L_o=\lambda-1$. If a specimen is extended to $L=2.5\ L_o$ of the original length L_o, then it is said to have been drawn by 150 %.

The tensile stress of *elastomers* increases continuously with increasing tensile strain until the polymer finally ruptures at σ_R and ϵ_R (Fig. 44). Typical *thermoplastics* (and most fibers) follow Hooke's law

$$\sigma_{11} = (F/A_o)\cdot\varepsilon = E\cdot\varepsilon \tag{77}$$

for small strains (up to point I), where A_o is the original cross section of the specimen, F is the force, and E is the tensile modulus (Young's modulus). Point I is thus called the proportionality limit or elastic limit; it is defined for a remaining strain of 0.1 % after removal of the stress.

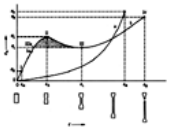

Figure 44. Stress – strain curves of an elastomer (a) and a partially crystalline thermoplastic (b) (schematic) (for further explanation, see text)

The necking effect shown below is characteristic for conventional thermoplastics; it is not found for elastomers and hard-elastic thermoplastics.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The maximum of the stress – strain curve is called the upper yield (point II), and the subsequent minimum is the lower yield (point III). The ratio of upper yield stress σ_S to tensile modulus E is practically constant for all polymers; the numerical value of $\sigma_S/E\approx0.025$ suggests that van der Waals bonds are broken and molecule segments begin to move more freely.

Brittle polymers break at the upper yield. Tough polymers continue to extend and the stress either remains constant (see below) or decreases. The latter phenomenon is called stress softening. It is typical for polymers with neck formation (telescope effect) and is nominal since it disappears if the stress is given relative to the actual cross section instead of the initial one.

The region between points III a – II – III is known as the ductile region; its area describes the absorbed energy and thus the toughness of the specimen. The subsequent increase of stress with strain is called stress hardening. At point IV, the ultimate strength (tensile strength, tenacity at break) σ_B and ultimate elongation (elongation at break) ϵ_B occur.

Stress – strain curves may differ considerably for plastics. Polymers with high Young's moduli (steep initial slopes) are called hard polymers; those with low moduli, soft; this hardness should not be confused with surface hardness. Typical hard polymers are phenolics [phenol – formaldehyde (PF)], polyacetals [polyoxymethylene (POM)], polycarbonates (PC), and poly(ethylene terephthalates) (Fig. 45). Polymers are further characterized by their stress – strain behavior between upper yield and failure. Polymers without yield cannot absorb energy and thus break easily; they are hard-brittle polymers (e.g., PF). Polycarbonates, on the other hand, show an extended ductile region and a fairly high fracture strain; they are called hard-tough. Polyethylene is similar with respect to ductile behavior and strain hardening; the modulus is however much lower, and PE is considered soft-tough.

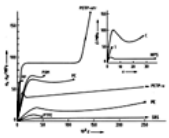

Figure 45. Stress – strain curves of polymers at room temperature (see text for explanation)

The stress – strain behavior described above is typical for tensile experiments. Tensile stresses lead to strong deformations in neck zones and cause microscopic voids at which fracture originates. Atactic polystyrene (PS) is such a hard-brittle polymer under tension T. No voids can be formed, however, under compression C, and PS appears as a hard-tough material (Fig. 45, insert). Rubber modification of PS leads to high-impact polystyrene (HIPS), which behaves quite differently from PS under tension (see Section Impact Resistance).

No yield value is found if drawn polymers are further subjected to tensile tests. In biaxially stretched poly(ethylene terephthalate) (PETP-str) some chain segments are already oriented, whereas other remain in their original positions. Biaxially stretched films are thus under stress, which is utilized in shrink films. Such films are used for the packaging of goods. On heating semicrystalline polymers above the glass transition temperature and below the melting temperature, chain segments become more mobile. Molecules attempt to attain their unperturbed dimensions and the films shrink, covering the goods tightly.

8.1.3. *Moduli and Poisson Ratios*

Elasticities can be described by three elastic moduli: tensile modulus or Young's modulus E, shear modulus G, and bulk modulus or compressive modulus K. Their values are inversely proportional to the corresponding compliances for static deformations (but not for dynamic ones):

Moduli	Compliances
$E=\sigma_{11}/\epsilon$	$D=1/E$
$G=\sigma_{21}/\gamma$	$J=1/G$
$K=p/(-\Delta V/V_o)$	$B=1/K$

The three simple moduli are related to each other for small deformations of simple isotropic bodies:

$$E = 2G\cdot(1+\mu) = 3K\cdot(1-2\mu) \tag{78}$$

where $\mu=(\Delta d/d_0)/(\Delta L/L_0)$ is the Poisson ratio (Poisson number), d is the diameter, and L is the length of the specimen. Poisson ratios can only adopt values $0<\mu<1/2$ for isotropic bodies but may assume $\mu>1/2$ for anisotropic ones. Equation (78) is invalid for anisotropic bodies and viscoelastic materials; $E/3<G<E/2$ and $0<K<E/3$, are however, still valid for these materials.

Common polymers behave with respect to μ, E, G, and K more like liquids than like metals (Table 23). Ultradrawn and self-reinforcing polymers may however exhibit tensile moduli that exceed those of steel (see below).

Table 23. Poisson ratios and elastic constants of various materials

Material	μ	*E*/GPa	*G*/GPa	*K*/GPa
Water	0.50	≈ 0	≈ 0	≈ 2
Gelatin (80 % water)	0.50	0.002		
Natural rubber	0.495	0.0009	0.0003	≈ 2
Polyethylene, LD	0.49	0.20	0.070	3.3
Polystyrene	0.38	3.4	1.2	5.0
Granite	0.30	30	12	25
Steel	0.28	211	80	160
Glass	0.23	60	25	37
Quartz	0.07	101	47	39
Aluminum oxide fibers	0	2000	1000	667

8.2. Energy Elasticity

8.2.1. *Theoretical Moduli*

The tensile moduli of common polymers are far lower than the theoretical moduli deduced from their chemical and physical structures (Table 24). Such theoretical moduli can be calculated from bond lengths, valence angles, and force constants for the deformation of these quantities. The theoretical moduli agree well with the microscopic lattice constants, which are determined experimentally by X-ray diffraction (change of Bragg reflexes), Raman spectroscopy, or inelastic neutron scattering under load. The longitudinal theoretical moduli $E_{\parallel}$ are far greater than the transverse moduli $E_{\perp}$ because the former are controlled by covalent bonds and the latter by van der Waals forces.

Table 24. Modulus of elasticity

Polymer	$E_{\parallel}$/GPa			$E_{\perp}$/GPa
	Theory	Lattice	Tensile	Lattice
Polyethylene	340	325	< 1	3.4
Polypropylene, it	50	42	< 3	2.9
Polyoxymethylene				
Orthorhombic	220	189	< 2	7.8
Trigonal	48	54	2	
Poly(*p*-phenylene terephthalamide)	182	200	132	10
Poly(4-methylpentene)	6.7		1	2.9

The larger the cross-section of the single polymer chain, the larger is the force distributed to fewer chains per unit area and the smaller are the moduli (Fig. 46). Chains in the all-*trans* conformation always exhibit higher theoretical moduli than do helical chains since the elongation of the former changes bond angles, whereas the extension of the latter involves only lower-energy torsion angles.

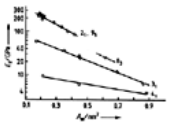

Figure 46. Longitudinal lattice moduli $E_{\parallel}$ as function of cross-sectional area A_m of polymer chains in all-*trans* (2_1) or helical conformations (9_5, 8_3, 3_1, 4_1) of their main chains

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Polyethylene has the highest theoretical modulus of all one-dimensional chains ($E_{\parallel}$=340 GPa), about one-third that of diamond ($E_{\parallel}$=1160 GPa) with its "naked" carbon chains. Polypropylene has the highest theoretical modulus of helical chains ($E_{\parallel}$=50 GPa).

Moduli approaching these theoretical values have been realized by ultradrawing of mats of polyethylene single crystals, giving polymers with $E_{\parallel}$=240 GPa. Industrially, high-modulus polyethylene fibers are manufactured by gel spinning of ultra-high molecular polyethylene ($E_{\parallel}$=97 GPa ≙90 N/tex, 1 tex=1 g per 1000 m).

8.2.2. *Real Moduli*

The much lower tensile moduli of conventionally processed polymers result from their disordered physical structures. In amorphous polymers, chain segments are oriented at random. Even in partially crystallized polymers, amorphous layers exist and chain axes (stems of lamellae) are distributed at random. The moduli of flexible polymers can be increased somewhat by processing under external force fields, for example, by partial orientation of chain segments during fiber spinning (Table 25, column L). Extrusion of solid polymers is a particularly effective method: the longitudinal modulus of polyoxymethylene increased to 24 GPa from 2 GPa on hydrostatic extrusion.

Table 25. Moduli and fracture strengths of conventional polymers as isotropic molding masses (I) or in draw direction of fibers (L) and of thermotropic (TT) and lyotropic (LT) glasses longitudinal (L) and transverse (T) to draw direction compared to isotropic polymers (I)

Polymer	*E*/GPa			σ_B/MPa		
	L	T	I	L	T	I
PE-LD	?	?	0.15	?	?	23
PA 6.6	13	?	2.5	1000	?	74
PETP	19	?	0.13	1400	?	54
TT X 7 G [a]	54	1.4	2.2	151	10	63
TT Vectra [b]	11	2.6	5.0	144	54	97
LT Kevlar [c]	138	7	?	2800	?	?
LT PPBT [d]	120	17	62	1500	680	700

[a] X 7 G=poly(*p*-hydroxybenzoate-*co*-ethylene terephthalate).
[b] Vectra=poly(*p*-hydroxybenzoate-*co*-2-hydroxy-6-naphthalate).
[c] Kevlar 49=poly(*p*-phenylene terephthalamide).
[d] PPBT=30 % poly(*p*-phenylene benzbisthiazole) in poly(2,5-benzimidazole).

Rodlike mesogens of liquid semiflexible polymers align in mesophase domains. The domains orient themselves in shear fields; the orientations can be frozen-in to LCP glasses. Such self-orienting polymers possess much higher moduli in the longitudinal direction than conventionally processed flexible polymers (Table 24). Their transverse moduli are also higher because of intermolecular dipole – dipole interactions.

8.2.3. *Temperature Dependence*

Young's moduli change characteristically with temperature for the various classes of polymers (Fig. 47). Five characteristic ranges can be distinguished: glassy (GL), leatherlike (LE), rubbery (RE), viscoelastic (RF), and viscous (VF).

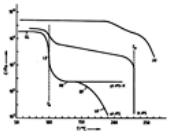

Figure 47. Temperature dependence of Young's moduli for conventional polymers

at-PS=Amorphous (atactic) polystyrene ($M>M_c$); at-PS-X=Its slightly cross-linked product; it-PS=Partially crystalline (isotactic) polystyrene; PF=Hardened phenol – formaldehyde resin; GL=glasslike; LE=leatherlike; RE=rubber (entangled); RF=viscoelastic; VF=viscous flow

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Amorphous Polymers. Moduli (0.1 – 1 GPa) are practically independent of temperature below the glass transition temperature T_G but drop to 10^5–10^6 Pa at $T=T_G$; the polymer appears leathery around T_G and rubbery above the glass transition temperature. The rubbery region is maintained for entangled (high molar mass) polymers but is nonexistent for low molar masses ($M<M_c$). Moduli decrease on further temperature increase until the polymers behave like viscous liquids ($E\approx10^3$ Pa).

Elastomers. At use temperatures, these polymers are above their T_G's and thus behave like cross-linked thermoplastics above the T_G's of the latter. The decrease in modulus at T_G is limited by (and a measure of) the degree of cross-linking. At higher temperatures, elastomers decompose and the moduli decrease drastically.

Crystalline polymers have only a weak leathery region around T_G since their amorphicity is slight and their crystallites act as physical cross-linkers. The modulus at the beginning of the rubbery plateau is a measure of the degree of crystallinity. The moduli decrease slowly during the rubbery "plateau" because more and more crystallites are molten. The remaining crystallites practically disappear at the melt temperature: the moduli drop drastically and the polymers behave like viscous liquids.

Thermosets. Thermosets are highly cross-linked polymers. They have high Young's moduli below T_G, usually one decade higher than thermoplastics. A very weak glass transition is accompanied by a somewhat leathery behavior. The subsequent rubbery region is not very marked because the cross-linking density is high.

8.3. Entropy Elasticity

Elastomers show pronounced entropy elasticities. They exhibit simultaneously some characteristics of solids, liquids, and gases. Like solids, they display Hookean behavior at not too high deformations (i.e., they show no permanent deformation after removal of the load). Moduli and expansion coefficients, on the other hand, resemble those of liquids. Like compressed gases, stresses increase with increasing temperature for elastomers at $T>T_G$.

The elastic behavior can be modeled with various theories. In the simplest case, dislocations of network junctions are assumed to be affine to the macroscopic deformation of the network. The tensile stress σ_{11} varies with the elongation $\lambda=L/L_o$ according to

$$\sigma_{11} = RT\cdot[M_c]\cdot(V_o/V)^{-1/3}(\lambda-\lambda^{-2}) \tag{79}$$

It depends on the volumes before (V_o) and after (V) deformation and on the molar concentration $[M_c]$ of network junctions. True Young's moduli can be obtained from the limiting value of $\sigma_{11}/(\lambda-\lambda^{-2})$ at $\lambda \longrightarrow 1$, for example, for volume- constant deformations ($V_o/V=1$):

$$\begin{aligned} RT\cdot[M_c] &= \lim_{\lambda\to1}[\sigma_{11}/(\lambda-\lambda^{-2})] \\ &= \sigma_{11}/[3(\lambda-1)] = \sigma_{11}/(3\varepsilon) = E_o \end{aligned} \tag{80}$$

The Young's modulus is thus directly proportional to the molar concentration of network junctions (cross-linking sites), independent of the nature of the latter. On shearing, elastomers behave like Hookean bodies because

$$RT\cdot[M_c] = \sigma_{21}/\gamma = G \tag{81}$$

They are, however, non-Hookean for elongations because $E_o\neq\sigma_{11}/\in$ (Eq. 80).

8.4. Viscoelasticity (⟶ Plastics, Properties and Testing)

8.4.1. *Fundamentals*

Most polymers do not revert "instantaneously" to their initial states after the removal of loads; they are neither ideal energy elastic (Section Energy Elasticity) nor ideal entropy elastic (Section Entropy Elasticity). These processes take certain times, (i.e., time-independent elastic and time-dependent viscous properties work together to produce a viscoelastic behavior). If stress, strain, and strain rate can be combined linearly, then the process is said to be linear elastic. In addition, some polymers may be irreversibly deformed.

The two ideal cases of response to deformations can be well described by mechanical models: a spring for a Hookean body (instantaneous response) and a dashpot for a Newtonian liquid (linear time dependence of response) (Fig. 48). The Maxwell element combines spring and dashpot in a series; the Voigt – Kelvin element, in a parallel manner.

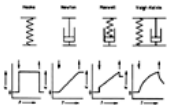

Figure 48. Time dependence of deformations according to various models. Loads are added at ↓ and removed at ⇑

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The Maxwell element describes a *relaxation* (i.e., the decrease of stress at constant deformation; Table 26). Linear combination of elastic deformation rates $d\gamma_e/dt=(1/G)(d\sigma/dt)$ and viscous deformation rates $d\gamma_\eta/dt=\sigma/\eta$ yields, after integration of the resulting expression and indexing the time for this particular behavior for $d\gamma/dt=0$,

$$\sigma = \sigma_0\cdot\exp(-G\cdot t_e/\eta) = \sigma_o\cdot\exp(-t_e/\tau) \tag{82}$$

Table 26. Simple mechanical models for the deformation of polymers

Model	Function	Behavior [*]	
		Initial	Final
Newtonian liquid	$\sigma = \eta\cdot\dot\gamma$	L	L
Voigt – Kelvin element	$\sigma = G\cdot\gamma + \eta\cdot\dot\gamma$	L	S
Maxwell element	$\sigma + (\eta/G)\cdot\dot\sigma = \eta\cdot\dot\gamma$	S	L

Hookean body	$=G\cdot\gamma$	S	S
Jeffrey's body	$\sigma+(\eta/G)\cdot\dot{\sigma}=G\cdot\gamma+\eta\cdot\dot{\gamma}$	S	S

* S=solid-like behavior; L=liquid-like.

The relaxation time $\tau=\eta/G$ indicates the time after which the stress has fallen to the e_{-1}th fraction of its initial value. The ratio τ/t_e of the relaxation time τ to the time scale t_e of the experiment is called the Deborah number ($DB=\tau/t_e$); it is 0 for liquids, ∞ for ideal-elastic solids, and approximately 1 for polymers near their glass transition temperatures.

Retardation is defined as the increase of deformation with time at constant stress. It is characterized by a "creep" of the material. Since this phenomenon was first observed on seemingly solid polymers at room temperature, it is also called "cold flow." In principle, retardation phenomena can be described by a Maxwell element. Because of mathematical difficulties in the solution of the equations, a special model is prefered (Voigt – Kelvin element) (Fig. 48), which yields for the deformation γ_r at constant stress σ_o after indexation for retardations r

$$\gamma_r=(\sigma_o/G_r)\cdot[1-\exp(-G_r\cdot t_r/\eta)] \tag{83}$$

where G_r is the retardation modulus, also often called the relaxation modulus. The retardation time t_r indicates the time at which the deformation has reached (1–1/e)=0.632 of the final deformation σ_o/G_r. Retardation times and relaxation times are of the same magnitude, but not equal because they rely on different models.

The term "viscoelasticity" is sometimes used to describe the reversible deformation according to Equation (83). However, it is often applied to the total deformation, which is composed of the contributions by Equation (83), a Hookean body with $\gamma_e=\sigma_o/G_o$, and a Newtonian liquid [$\gamma_\eta=(\sigma_o/\eta_o)\cdot t$]:

$$\begin{aligned}\gamma_{tot}&=\{(1-G_o)+(t/\eta_o)+(1/G_r)\cdot[1-\exp(-t/\tau)]\}\cdot\sigma_o\\&=\gamma_e+\gamma_\eta+\gamma_r\end{aligned} \tag{84}$$

The three deformation terms γ_e (elastic), γ_η (viscous), and γ_r (viscoelastic) are often not explicitly evaluated. The time-dependent viscous and viscoelastic parts are rather combined into a new parameter $\gamma_c=\gamma\cdot t^n$ (Findlay law). The resulting function for the creep curve allows an extrapolation to long-time behavior from short-time experiments:

$$\gamma_{tot}=\gamma_e+\gamma_c=\gamma_e+\gamma\cdot t^n \tag{85}$$

8.4.2. *Time – Temperature Superposition*

Deformations, shear moduli, and shear compliances are time and temperature dependent (Fig. 49). The moduli vary about one decade for six decades in time at constant temperature, and up to one decade for each 10 K at constant frequency. Since no single experimental method can cover the 15 – 20 decades of frequency that are required for good characterization of a polymer, time – temperature data from various techniques are usually combined with the help of the Boltzmann superposition principle.

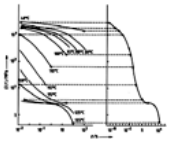

Figure 49. Time dependence of the shear modulus from measurements of the stress relaxation of a poly(methyl methacrylate) ($\bar{M}_v$=3 600 000 g/mol) at various temperatures (left) and the resulting time – temperature superpositions for a reference temperature of 115 °C (right) (circles represent equivalent positions)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

This principle states that the deformation (or recovery) caused by an additional load (or removal thereof) is independent of previous loads or their removal. The $G=f(t, T)$ curves can be combined if (1) the relaxation time spectrum is temperature independent and (2) the thermal activation is the same over the entire time and temperature range (no transition or relaxation temperatures). A reference temperature is chosen close to the static glass transition temperature (115 vs. 105 °C in Fig. 49) and the G values are shifted horizontally with the help of a shift factor from the WLF equation (Eq. 69).

8.5. Dynamic Behavior

8.5.1. *Fundamentals*

Dynamic – mechanical methods expose the specimen to periodic stresses. The polymer either is put under torsion once and then oscillates freely (torsion pendulum) or is subjected continuously to forced oscillations (e.g., Rheovibron). In addition, ultrasound, dielectric, and NMR methods can be used to study the dynamic properties of polymers.

In the simplest case, the applied stress is sinusoidal with a frequency ω ($\sigma_t=\sigma_o\cdot\sin\omega\cdot t$). The deformation of ideal-elastic bodies follows the stress instantaneously ($\gamma_t=\gamma_o\cdot\sin\omega\cdot t$) but that of viscoelastic polymers experiences a delay ($\gamma_t=\gamma_o\cdot\sin(\omega\cdot t-\vartheta)$). The stress vector is assumed to be a sum of two components: one component is in phase with the deformation ($\sigma'=\sigma_o\cdot\cos\vartheta$); the other is not ($\sigma''=\sigma_o\cdot\sin\vartheta$).

Each of these two components possesses a modulus. The *real modulus* (*shear storage modulus*) G' measures the stiffness and shape stability of the specimen:

$$G'=\sigma'/\gamma_o=(\sigma_o/\gamma_o)\cdot\cos\vartheta=G^*\cdot\cos\vartheta \tag{86}$$

whereas the *imaginary modulus* (*shear loss modulus*) G'' describes the loss of usable mechanical energy by dissipation into heat:

$$G''=\sigma''/\gamma_o=G^*\cdot\sin\vartheta \tag{87}$$

The same quantities can also be derived if complex variables are introduced

$$G^*=G'+i\cdot G''=\left[(G')^2+(G'')^2\right]^{-1/2} \tag{88}$$

The loss factor δ is the ratio of imaginary to real modulus. It is the same for shear and Young's moduli but not for compression moduli

$$\Delta=\tan\vartheta=G''/G'=E''/E'<K''/K' \tag{89}$$

8.5.2. *Molecular Interpretations*

The *shear storage moduli* of low molar mass polymer melts with narrow molar mass distributions increase continually with increasing frequency (Fig. 50). At high normalized frequencies $\alpha_T\omega$, all storage moduli asymptotically approach a limiting line, regardless of molar mass. This part of the relaxation spectrum thus originates from the mobility of chain segments; it is called the transition range.

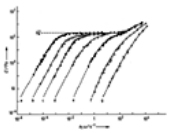

Figure 50. Frequency dependence of the shear storage modulus G' for narrow-distribution polystyrene melts of different molar masses M. A shift factor α_T was used to convert to 160 °C data measured at various temperatures

a) M=581 000 g/mol; b) M=351 000 g/mol; c) M=215 000 g/mol; d) M=113 000 g/mol; e) M=46 900 g/mol; f) M=14 800 g/mol; g) M=8900 g/mol

Reprinted with permission by Hüthig and Wepf Publ., Basel [5] after data of [80]

High molar mass polymers show a corresponding frequency dependence of *loss moduli* at very low frequencies (called end range), followed by a plateau at higher frequencies (plateau modulus G_N^o), and finally the transition range. The end ranges of these spectra are molar mass dependent; this behavior must come from long-range conformational changes. Since the transition range characterizes viscous behavior, and the end range viscoelastic behavior, the plateau range must reflect rubbery behavior (see Fig. 47 for the temperature dependence).

The *rubbery behavior* of polymer melts can be described by the theories of entropy elasticity according to which the shear modulus of chemically cross-linked polymers depends on the molar concentration of network junctions. The plateau modulus G_N^o of melts thus indicates the concentration of temporary junctions (entanglements). The molar

mass M_e of segments between such junctions of polymers with volume fractions Φ_p in solution (Φ_p=1 for melts) and polymer melt densities ρ_p is given by

$$M_e = RT \cdot r_p \cdot \phi_p / G_N^o \quad (90)$$

These dynamic entanglement molar masses M_e are a factor 2.0±0.2 lower than the corresponding molar masses M_c from rest viscosities (Table 27).

Table 27. Critical molar masses for entanglement from shear moduli (M_e) and rest viscosities (M_c)

Polymer	T	M_c	M_e	M_c
	°C	g/mol	g/mol	M_e
Polyethylene	190	3 800	1 790	2.1
Polypropylene, it	190	7 000		
Polyisobutylene	25	15 200	8 800	1.7
Polydimethylsiloxane	25	24 500	10 500	2.4
Poly(vinyl acetate), at	57	24 500	12 000	2.0
Poly(α-methylstyrene), at	100	28 000	13 500	2.1
Polystyrene, at	190	35 000	18 100	1.9

The plateau is not well developed or may even be absent for polymers with broad molar mass distributions; a complicated dependence of shear compliances on higher molar mass averages has been predicted by reptation theory.

8.6. Fracture

8.6.1. *Overview*

Polymers break very differently depending on their chemical and physical structure; environment (humidity, solvents, temperature); and the type, duration, and frequency of deformation. Some polymers break immediately; others are unchanged even after months. The fracture surface can be smooth or splintery; the elongation at fracture, less than 1 % or greater than 1000 %.

Two fracture modes can be distinguished: brittle and tough (ductile). Brittle polymers fracture perpendicularly to the stress direction, tough polymers longitudinally (Fig. 51). A polymer is defined as brittle if its elongation at break is less than 20 %.

Figure 51. Failure modes of polymers (matrix M), fibers (F), and fiber-reinforced polymer composites (C) by brittle failure (br), plastic flow (pl), shear band formation (sb), shearing (s), kink formation (k), bending (b), longitudinal splicing (sp), formation of kink bands (kb), and step formation by compression (cp)

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

Brittle fractures are rare for ideal solids since many bonds must be severed simultaneously. Real polymers however contain many small imperfect regions that act as “nuclei” for the formation of microcracks. Brittle polymers usually possess “natural” microvoids, which may also appear in drawn amorphous polymers or through separation of crystal lamellae in hard-tough polymers.

Tough failures (ductile fractures) are caused by viscous flow (“plastic flow”). This process may involve the slipping of chain segments past each other (amorphous polymers) or the movement of crystalline domains (partly crystalline polymers). Polymer chains may also de-entangle at long times and under small stresses. The same processes and additional ones may occur on failure of composites (Fig. 51).

Polymers are subjected to very different stress conditions in typical applications; they thus experience different failure modes. Test methods try to simulate complex real-life situations by standardized procedures. They include long-term experiments such as static deformations under constant load by tension, compression, or bending; short-term methods such as tensile tests under various speeds or impact tests with unnotched or notched specimens; dynamic testing with variation of the number of loadings – unloadings, impacts, vibrations, etc.

8.6.2. *Theoretical Fracture Strength*

The fracture of brittle polymers generates free radicals. Since the probabilities of such homolyses depend on bond strengths, which also determine tensile moduli, relationships must exist between the theoretical moduli and the theoretical fracture strengths of polymers.

Bonds are severed if atoms are separated from each other by certain distances L_b greater than their equilibrium distances L_o. The necessary theoretical strength $\sigma^o_{\parallel}$ is given by [5].

$$\sigma_{\parallel}^o = \frac{E_{\parallel}^o \cdot (L_b - L_o)}{\pi L_o} = K \cdot E_{\parallel}^o \quad (91)$$

Polymer main- chain bonds break at approximately the same relative distance (L_b≈1.3 L_o) because bond lengths and strengths are not too different for bonds such as C–C, C–O, and C–N. Thus, K≈0.095 and the theoretical fracture strength $\sigma^o_{\parallel}$ should be ca. one-tenth of the theoretical tensile modulus $E^o_{\parallel}$ in the chain direction, regardless of the chemical nature of the polymer. Polyethylene with a theoretical modulus of 340 GPa (Table 24) should thus have a theoretical fracture strength of ca. 32 GPa [i.e., much higher than the theoretical strength of steel (ca. 20 GPa)]. Industrially manufactured ultradrawn polyethylene fibers have higher experimental fracture strengths than steel (2.9 GPa=2.7 N/tex vs. 2.5 GPa).

The theoretical fracture strength – tensile modulus relationship of Equation (91) has been realized for certain ultradrawn polyethylenes in which both the predicted proportionality constant K=0.095 and the first power of E were found (Fig. 52). The type of same relationship is also observed for molecular composites of poly (*p*-phenylene-2,6-bisbenzthiazole) (PPBT) with rodlike mesogens in coillike polybenzimide (ABPBI), albeit with a lower proportionality constant. In other ultradrawing experiments, a power dependence $\sigma = K' \cdot E^n$ was found, however.

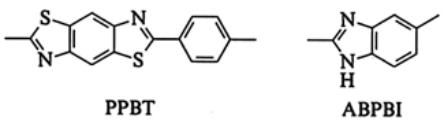

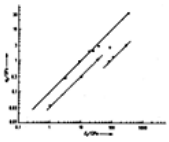

Figure 52. Tensile strength at break as function of tensile modulus: (•) experimentally ultradrawn ultra-high-modulus polyethylenes; (◯) industrially manufactured ultradrawn polyethylene (Dyneema); (◯) theory for perfectly aligned polyethylene; (♦) heterogeneous molecular composites of ABPBI fibers or PBT fibers or films in ABPBI matrix; (♦) homogeneous molecular composites of PBT fibers or PBT fibers for films in ABPBI. Solid line corresponds to σ_B=0.095 · $E_{\parallel}$

8.6.3. *Real Fracture Strength*

The lower than theoretical fracture strengths of most polymers (see Tables 24 and 25) are caused by many factors. Theoretical fracture strengths relate to infinitely long, completely aligned, immobile polymer chains. End groups and chain folds act as disturbances: fracture strengths of conventional polymers increase with increasing molar mass and become practically constant above a “critical” molar mass.

At the latter molar mass range, chain segments of amorphous polymers are distributed at random. A brittle fracture across such polymers will create two new surfaces with a

total surface energy of 2 γ_{lv}. The theoretical fracture strength for brittle, energy-elastic bodies

$$\sigma_a^o = (E \cdot \gamma_{lv} / L_o)^{1/2} \quad (92)$$

now depends on the product of modulus and surface energy (Ingles theory). The Ingles theory works well for silicate glasses.

The experimentally found fracture strengths σ_b of plastics are however much lower than the strengths predicted by Equation (92). The ratios $\sigma_a^o / \sigma_{exp}$ of molded (unoriented) plastics decrease with increasing elongation ϵ_b at break (Fig. 53), whereas those of drawn fibers (partially oriented chain segments) increase.

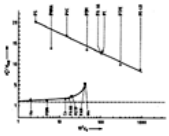

Figure 53. Ratio $\sigma_a^o / \sigma_{exp}$ of theoretical [Eq. (92)] and experimental fracture strengths as function of elongation at break ϵ_b for (○) nonoriented (molded plastics) and (•) oriented (fibers) polymers

He=hemp, Co=cotton, Wo=wool (for other abbreviations see Tables 1-4)

The reason for the lower than expected fracture strengths of amorphous polymers is the presence of microvoids, which act on drawing as nuclei for cracks. According to the Griffith theory, a crack can grow only if the energy required for the fracture of chemical bonds is just surpassed by the stored elastic energy. This theory predicts a dependence of the fracture strength on crack length *L*:

$$\sigma_B = [(2E \cdot \gamma_{lv}) / (\pi \cdot L)]^{1/2} \quad (93)$$

This functionality is indeed observed for artificially introduced long cracks. The predicted fracture strengths are however much lower than those found by experiment. Furthermore, deviations from Equation (93) occur at small crack lengths because the fracture behavior of plastics is not dominated by the cleavage of chemical bonds but by other types of energy absorption (crazing; shear flow).

On drawing, stresses are imposed on microvoids. The polymer reaches its upper yield stress at sufficiently high stress concentrations at the tip of such a void and relieves the stress by stress softening (Fig. 44). The induced cooperative movements of chain segments cause long-range changes of macroconformations. In partially crystalline polymers, these changes can occur only in amorphous domains; spherulitic polymers break accordingly either between spherulites or in the radial spherulite direction.

The cooperative movements of segments lead to either shear or normal stress yielding. On shearing, the whole specimen yields either homogeneously or heterogeneously (localized). In the latter case, shear bands are formed at angles of 38 – 45° to the stress direction (Fig. 51). Chain segments are arranged at angles between shear bands and stress directions.

All polymers with upper yield values form crazes upon stress softening (Fig. 54), regardless of whether they are amorphous, crystalline, linear, or cross-linked. Crazes can be up to 100 µm long and 10 µm wide; their long axes are parallel to the stress direction. They are not voids since their interior is filled with amorphous microfibrils of 0.6 – 30 nm diameter; these microfibrils are oriented in the stress direction (i.e., perpendicular to the craze long axes). On further deformation, microvoids are formed.

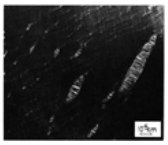

Figure 54. Crazes in a polystyrene drawn to 25 % [81]

The formation of crazes is the primary mechanism for the dissipation of stress energy. It is utilized in the rubber reinforcement of polystyrene. Rubber-modified polypropylene, on the other hand, deforms mainly by shear flow.

8.6.4. *Impact Resistance*

Impact strength is the resistance of a material to impact. It is one of the many quantities used to characterize the strength of a material under (the usually complex) use conditions; all test methods are thus standardized. Most test methods measure the energy required to break a notched or unnotched specimen (Izod, Charpy, high-speed tensile). Impact speeds range from 10^{-5}–10^{-1} m/s in conventional tensile tests to 20 – 240 m/s for high-speed tensiles; elongation speeds are usually from 10^{-3} to 10^{4} s^{-1}.

Impact strengths depend on experimental conditions. The smaller the radius of the notch, the higher is the stress concentration at the tip and the lower is the impact strength. At very low temperature, all polymers are brittle. The mobility of chain segments increases with increasing temperature, allowing stresses to be relieved by shear-band or craze formation: impact strengths increase with temperature, especially near the glass transition temperature. Polymers with additional transition temperatures below the glass transition temperatures are for the same reason almost always more impact resistant than polymers without such transitions. Nonentangled polymers exhibit very low impact strengths because no crazes can be formed. The impact behavior of polymers can be improved considerably by modification with rubber (see Section Toughened Plastics).

8.6.5. *Stress Cracking*

Stress cracking (stress corrosion, stress crazing) is the formation of crazes under the physical action of chemicals, especially surfactants. Stress corrosion starts at polymer surfaces and proceeds into the interior until the polymer finally cracks. The appearance and the extent of stress cracking depend on the polymer – reagent interaction and the magnitude of the stress.

Effects are weak in nonwetting liquids but strong in polymer – liquid systems with solubility parameters of polymers and liquids matching each other and even more dramatic under tension in the presence of surfactants. Stress cracking decreases with increasing molar mass of the polymer since entanglements allow stresses to relax elastically. Cross-linked polymers are less prone to stress cracking for the same reason. Stress cracking is also reduced if polymer plasticizers are present in plastics because these additives increase the mobility of chain segments and thus the ability to relieve stresses. The same action is responsible for the fact that no stress corrosions are observed above glass transition temperatures.

8.6.6. *Fatigue*

Materials may be damaged not only “instantaneously” (i.e., on impact) but also by static or periodic loads after certain times or number of loadings. This fatigue is characterized by the *fatigue limit* (endurance) at which the plastics are not damaged even after infinite time and the *fatigue strength*, which indicates the load at which damage sets in after a certain time.

Plastics may be subjected to static loads for certain times *t*, after which their fracture strengths σ_B are measured by tensile tests. The logarithms of strength of amorphous polymers usually decrease linearly with logarithms of time due to viscous flow (Fig. 55). Partially crystalline polymers show a bend in these lines after certain times, which indicates a change from tough fracture (short times) to brittle fracture (long times), probably caused by recrystallization phenomena.

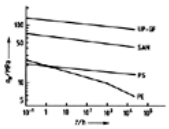

Figure 55. Time dependence of tensile fracture strengths after static loading during time *t*. UP-GF=Glass-fiber-reinforced unsaturated polyester; SAN=Styrene – acrylonitrile copolymer (impact polystyrene); PS=Polystyrene; PE=Polyethylene

Reprinted with permission by BASF AG, Ludwigshafen [78]

8.7. Surface Mechanics

8.7.1. *Hardness*

The hardness of a material is its resistance to penetration by another body. Hardness is a very complex quantity; it depends on Young's modulus, yield stress, and stress hardening. A general definition of hardness, applicable to all materials, does not exist; neither does a universally applicable testing method. The various technical test methods thus emphasize one or another factor that contributes to the hardness of a specific class of materials.

The hardness of *hard plastics* is normally characterized by various Rockwell (ISO, ASTM) or ball indentation hardnesses (DIN, ISO). These methods measure the indentation of

a polymer by a steel sphere under load and thus the compression set and the recoverable deformation. The plastic deformations of polymers increase with time (creep), whereas those of metals are time independent. Because of the short duration of the hardness test polymers exhibit relatively high Rockwell hardnesses.

The hardness of *soft plastics* is characterized by their durometer (ASTM) or various Shore hardnesses (ISO, DIN). These methods measure the resistance to penetration by a truncated cone (static methods). Hardness properties of metals and hard plastics are evaluated by another Shore hardness that uses the rebound of a small steel sphere (dynamic method).

All methods measure the hardness of surfaces, not of the interior of the specimen. The surface may, for example, be plasticized by the humidity of the air. Crystallizable polymers may have lower surface hardnesses than interior hardnesses (if the plastic had been injected into a cold mold) or the reverse may be true (if transcrystallization occurred).

8.7.2. *Friction*

Friction, the resistance against the relative movement of two bodies contacting each other is measured by the friction coefficient $\mu = R/L$ (i.e., the ratio of friction R to total load L). Friction depends in a complex and not understood way on both the surface roughness of the specimen and its mechanical properties.

The rolling of *hard bodies on soft materials* is determined almost exclusively by the deformation of the soft base (i.e., its viscoelastic properties). Elastomers thus have fairly high friction coefficients of $0.5 < \mu < 3.0$, depending on the contacting body and its type of movement (rolling, sliding).

The sliding of *hard bodies on other hard bodies* occurs on the tops of the microscopic surfaces: the true contact area is much smaller than the geometric one. The applied load thus acts on very small effective areas. Local stresses are high and the tops are leveled. Large adhesion forces exist between chemical groups of the resulting effective contact areas of both bodies, which must be overcome by breaking the bonds or by shearing one of the materials. The adhesive friction $R = A_w \cdot \sigma_b$ is given by the effective contact surface A_w and the shear strength σ_b. Soft materials possess high effective surfaces (large A_w) and are easily sheared (small σ_b), whereas the opposite is true for hard materials. Plastics, metals, and ceramics therefore often exhibit very similar friction coefficients (Table 28).

Table 28. Friction coefficients of various sliding bodies

Plastic	Friction coefficient of		
	Plastic on plastic	Plastic on steel	Steel on plastic
Poly(methyl methacrylate)	0.8	0.5	0.45
Polystyrene	0.5	0.3	0.35
Polyethylene, high density	0.1	0.15	0.20
Polyethylene, low density	0.3		0.80
Polytetrafluoroethylene	0.04	0.04	0.10

8.7.3. *Abrasion and Wear*

Abrasion is the loss of material from surfaces by friction. It is thus affected by both friction properties and hardnesses of the specimens. The abrasion coefficient K is given by the applied force F, the linear speed v of the contacting body, the total time t, and the volume loss ΔV of the abraded material: $K = \Delta V/(F \cdot v \cdot t)$.

Abrasion coefficients vary widely with polymer type and state (resting, mobile) (Table 29).

Table 29. Abrasion coefficients *K* for moving plastics against resting materials

Resting material	10^{10} *K*	Moving polymer	10^{10} *K*
	MPa^{-1}		MPa^{-1}
Polycarbonate A	200 000	polyamide 66	11 000
Polyamide 66	250	polycarbonate A	9 800
Polyamide 66	220	polyamide 66	510
Polyamide 66	10	polyacetal	12
Polyacetal	11	polyamide 66	15
Steel		polyamide 66	8 600
Steel		polyamide 66 with 30 % glass fibers	1

The best resistance against abrasion is shown by polyureas, followed by polyamides and polyacetals. It can be enhanced greatly by addition of certain fillers (e.g., short fibers).

[Top of Page]

9. Electric Properties

Matter is subdivided according to its specific electrical conductivity σ into insulators ($\sigma = 10^{-14}$–10^{-22} S/cm), semiconductors (10^{2} –10^{-9} S/cm), conductors (> 10^{3} S/cm), and superconductors (≈ 10^{20} S/cm). Most plastics are insulators (Section Dielectric Properties), but certain polymers are intrinsic semiconductors and some may even be conductors after doping (Section Electrical Conductivity).

9.1. Dielectric Properties

9.1.1. *Relative Permittivity*

Groups within a molecule and entire molecules of an insulator are polarized by applied electric fields. Polarization is usually measured by the ratio of capacitances of a condensor in vacuo and in the specimen (i.e., the relative permittivity ϵ_r of the specimen; formerly called the dielectric constant). Relative permittivities are low for apolar polymers [polytetrafluoroethylene (PTFE); PE], higher for polymers with polarizable groups (PS, PC), and still higher for polar materials [dry polyamide (PA)] (Table 30).

Table 30. Electrical properties of plastics

Polymer	Δw [a]		ρ_v [c]	ρ_s [d]		S [f]	
	%	ϵ_r [b]	Ω · cm	Ω	tan δ [e]	$kV^{-1}mm^{-1}$	U [g]/V
PTFE	0	2.15	10^{18}		0.0001	40	> 600
PE	0.05	2.3	10^{17}	10^{13}	0.0007	70	600
PS	0.1	2.5	10^{18}	10^{15}	0.0002	140	500
SAN		3			0.0070	100	
ABS		3.2	10^{15}	10^{13}	0.02	15	600

PVC	< 1.8 [h]	< 3.7 [h]	10^{15}	10^{13}	0.015	< 50	< 600
PA 6							
Dry	0	3.7	10^{15}		0.03	< 150	600
Conditioned	9.5	7	10^{12}		0.3	80	600
CA	4.7	5.8	10^{13}		0.03	35	
UP		3.4	10^{13}	10^{12}	0.01	50	500
PUR		4	10^{13}		0.05	30	
PF		8	10^{14}		0.05	12	
With inorganic fillers			10^{10}		< 0.5	14	< 150
Without inorganic fillers			10^{9}		< 0.5	10	125
UF		6	10^{11}		0.1	10	

[a] Δw=water absorption (at 50 % relative humidity.
[b] ϵ_r=relative permittivity ("dielectric constant").
[c] ρ_v=volume resistivity (inverse specific electrical conductivity).
[d] ρ_S=surface resistivity (inverse surface conductivity).
[e] tan δ=dissipation factor (loss tangent, at 1 MHz).
[f] *S*=dielectric strength.
[g] *U*=tracking resistance (method KC).
[h] Depends on impurities from polymerization (e.g., emulsifier residues).

The relative permittivity thus increases with increasing water content of plastics (ϵ_r=81 for water). It also increases with increased segmental mobility (ϵ_r=13.0 for *cis*-1,4-polyisoprene); rubber-modified plastics have higher relative permittivities than conventional plastics. Expanded plastics and elastomers are composites with air (ϵ_r=1.00058) and subsequently have low relative permittivities of ca. 1 – 2. Fillers increase the relative permittivities to values up to 170 (filled thermoplastics) and up to 18 000 (filled elastomers).

9.1.2. *Dielectric Loss*

Dipoles try to follow the direction of an electric field if an alternating current is applied. The required adjustment times correspond to the orientation times of groups and molecules. The faster the alternation, the longer the orientation lags behind the field and the greater is the electrical energy consumed. Available output power is decreased because electric power is lost by conversion into thermal energy.

The ratio of power loss N_v to total power output N_b is called the dielectric dissipation factor or loss tangent tan δ, which can also be expressed as the ratio of imaginary relative permittivity (ϵ'') to real relative permittivity (ϵ'):

$$N_v/N_b = \tan\delta = \sin\delta/\cos\delta = \varepsilon''/\varepsilon' \quad (94)$$

Sometimes power factors sin δ are given instead of loss tangents tan δ.

Polymers with high loss factors $\epsilon \cdot \tan\delta$ can be heated and thus welded by high-frequency fields; PVC is an example. Polymers with low loss factors [PE, PS, polyisobutylene (PIB)], on the other hand, are excellent insulators for high-frequency conductors.

Real and imaginary relative permittivities depend on the frequencies *v* of the alternating current (Fig. 56). The function $\epsilon'=f(v)$ corresponds to a dispersion and the function $\epsilon''=f(v)$ to an absorption of energy. Transitions and relaxations consume energy and therefore inflection points (ϵ'- curves) and maxima (ϵ''- curves) are found at appropriate temperatures and frequencies.

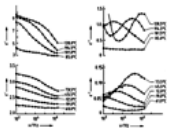

Figure 56. Frequency dependence of real (ϵ^+) and imaginary relative permittivities (ϵ'') of a poly(vinyl chloride) at various temperatures

Upper: β-dispersion; lower: γ-dispersion.

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

9.1.3. *Dielectric Strength and Tracking Resistance*

The imaginary part of the relative permittivity is caused by the dissociation of polar groups that may be either inherent to the polymer or introduced by extraneous impurities. These polar groups must be of ionic nature since the electrical conductivities of conventional plastics are strongly temperature dependent (electronic conductivities are far less temperature dependent).

Heat is caused to develop by the imaginary part of the relative permittivity. The low thermal conductivities of plastics do not allow this heat to dissipate, and the temperature increases. Ionic conductivities are thus increased until a breakdown (arcthrough) finally occurs. The resistance against such a breakdown is measured by the electric strengths *S* of a plastic ("dielectric strength") (Table 30).

A breakdown can also occur through tracking on the surface of a plastic. The tracking resistance is difficult to measure because surface resistivities are 2 – 3 decades lower than volume resistivities (Table 30). The tracking resistance is thus measured by standardized methods, such as the maximal voltage that does not cause tracking if 50 drops of an aqueous 0.1 wt % NH_4Cl solution are applied between two platinum electrodes that are under an alternating current and 4 mm apart on the specimen surface. A polymer has a good tracking resistance if it forms volatile products and no carbon upon degradation (volatile monomer by depolymerization of PMMA, volatile oligomers by degradation of PE or PA). Poly(*N*-vinyl carbazole) does not form volatile products and thus has a poor tracking resistance although it is a good insulator.

9.1.4. *Electrostatic Charging*

Static electricity originates from an excess or a deficiency of electrons on isolated or ungrounded surfaces. It can be created by rubbing two surfaces against each other (triboelectric charging) or by contact of a surface with ionized air. Matter is charged electrostatically if specific conductivities are lower than ca. 10^{-8} S/cm and relative humidities lower than ca. 70 %. All conventional plastics can thus be electrostatically charged (Table 30). Charge densities may vary between, e. g., 8.2 C/g for polychlorotrifluoroethylene (PCTFE) and – 13.9 C/g for phenolic resins, and charges may vary between 3000 V/cm (POM vs. PA 6) and – 1700 V/cm (ABS vs. PA 6).

Static charging can be reduced by incorporation of conducting fillers into the plastics, such as carbon black or metal powders (internal antistatics). External antistatics reduce surface resistivities by increasing the polarity of the surface via application of humidity-absorbing additives or by reducing friction through lubricants or coating with PTFE.

9.2. Electrical Conductivity (→ Polymers, Electrically Conducting)

Electrical conductivities in metals are caused by $N/V=10^{21}$–10^{22} (quasi) free electrons per cubic centimeter with electric charges $e \equiv 1.6\times 10^{-19}$ C and mobilities $\mu=10$ – 10^6 cm^2 V^{-1} s^{-1}. According to

$$\sigma = (N/V)\cdot\mu\cdot e \quad (95)$$

specific electrical conductivities of metals are therefore between ca. 630 000 (Ag) and 10 400 (Hg) S/cm. Graphite exhibits $\sigma=10^4$ S/cm in plane direction and 1 S/cm perpendicular to it. Semiconductors possess carrier mobilities similar to metals but at far lower carrier concentrations.

Charge transfer occurs fairly easily in metals and semimetals because atoms are tightly packed. Chain atoms of polymer molecules are also close in the chain direction because of covalent bonds. Only van der Waals and/or dipole forces act between chains, however, resulting in large intermolecular atomic distances and very difficult charge transfers. All electrons are furthermore localized in covalent polymer chains. Conventional polymers are thus insulators.

Polymers with conjugated chains [e.g., *trans*-polyacetylene ~$(CH{=}CH)_n$~] are, for these reasons, merely semiconductors and insulators even if the chains are planar. These low

conductivities can be increased substantially, however, if the polymers are doped with substances such as I_2, AsF_5, BF_3, etc.: doping with AsF_5 increases the specific electrical conductivity of *trans*-polyacetylene from 10^{-9} to 1200 S/cm and of poly(*p*-phenylene) from 10^{-15} to 500 S/cm. Such doped polymers can be processed like thermoplastics to any shape desired, which together with their light weight makes them attractive for many applications.

The action of these dopants is quite different from those of small amounts of dopants in inorganic semiconductors. The doping of inorganic semiconductors such as GaP, InSb, or Ge generates quasi-free electrons (*n*- carriers) or defect electrons (*p*- carriers), whereas the doping of suitable organic polymers leads to oxidation (*p*-doping) or reduction (*n*-doping) reactions. Sizable effects in organics are thus achieved only if large amounts of dopants are used, often up to 1 : 1 molar ratios of dopants to repeating units.

Doped polymers exhibit neither Curie paramagnetism (localized charge carriers) nor Pauli paramagnetism (electrons delocalized over the entire system). Thus the electrical conductivity of such systems is assumed to be due to solitons or polarons.

Double and single bonds alternate in *trans*-polyacetylene. Since these bonds are exchangeable, two low-energy states A and B with equal energy must exist. A *soliton* is a kind of topological kink that separates the A state from the B state with opposite bond alternation. *Polarons* are similar kinks between aromatic and quinoid structures:

Soliton

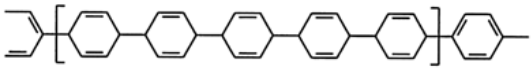

Polaron

The kink is small (about 14 chain atoms in *trans*-polyacetylene) and thus very mobile. Two types of solitons and polarons exist. Neutral solitons and polarons possess radicals that are produced as defects during isomerization of the polymers. Charged solitons and polarons are created by doping, either as carbonium ions or as carbanions.

Solitons can move only along the chain; there is no tunnel effect between chains. Because of this anisotropic behavior, doped polymers (and conducting low molar mass organic molecules) are called low-dimensional conductors, synthetic metals, or organic metals.

The most important condition for the existence of electrical conductivity in organic polymers seems to be the ability to form overlapping orbitals. The planar structure of *trans*-polyacetylene promotes the overlapping of its π and *p* orbitals. In poly(*p*-phenylene sulfide) (PPS) ~[S – (*p*-C_6H_4)]$_n$~, *p* and *d* orbitals of sulfur atoms probably overlap with the π systems of phenylene groups; PPS – AsF_5 shows an electrical conductivity of 10 S/cm, although the chain is not planar and the phenylene residues are arranged at angles of 45 °C to the planar zigzag chain of the sulfur atoms.

Electrically conducting polymers are presently used in small batteries, for example, a complex of poly(2-vinylpyridine)/I_2 as cathode in Li/I_2 batteries for pacemakers (σ=10 $^{-3}$ S/cm). Other possible applications of electrically conducting polymers are in solar cells, electrolysis membranes, microwave shielding, and integrated circuits.

9.3. Photoconductivity

Light generates radical ions and thus photoconductivities in certain systems. This effect is used in xerography to generate pictures of objects (e.g., copies of documents). The early photoconducting material As_2Se_3 has since been replaced by poly(*N*-vinylcarbazole), which absorbs UV light and forms an exciton, which is ionized by an electric field. The polymer is nonconducting in visible light; upon sensibilization by certain electron donors, charge-transfer complexes are formed, however. Another photoconducting system consists of polycarbonate A and triphenylamine.

[Top of Page]

10. Optical Properties

Many optical properties of plastics depend on their *refractive index n*, for example, reflection, gloss, transparency, and hiding power. Refractive indices are in turn determined by the polarizabilities *Q* according to the Lorenz – Lorentz relationship

$$(n^2-1)/(n^2-2) = (4/3)\pi Q = (4/3)\pi(N/V)\cdot\alpha \qquad (96)$$

where *N/V* is the number concentration of molecules with polarization α. The polarization is a function of the dipole moments of all groups in a molecule (i.e., the mobility and the number of electrons per molecule). Contributions to the refractive index are thus much higher for carbon atoms than for hydrogen atoms. Since the contributions of the latter can be neglected and since carbon atoms dominate the structures of polymers, all polymers possess approximately the same refractive index of 1.5 (e.g., PMMA 1.492, PP-it 1.53, PS 1.59). Deviations from this rule exist for strongly polarizable polymers (PTFE 1.37), polymers with bulky conjugated substituents (PVK 1.69), or polymers with a high content of noncarbon atoms (PDMS 1.40). According to the molecular structure of all known polymers, their refractive indices should be between ca. 1.33 and 1.73.

Molecules are more tightly packed in crystalline polymers than in amorphous ones. Refractive indices thus increase with increasing crystallinity. Because crystalline polymers are always anisotropic, different polarizabilities and refractive indices are exhibited in the chain direction and perpendicular to it.

A part of the light falling on a homogeneous, transparent body is reflected. *Reflectivity* is defined as the ratio of the intensities of reflected and incident light $R=I_r/I_0$, which according to Fresnel's law depends on the angles of incidence α and refraction β:

$$R = \frac{I_r}{I_0} = \frac{1}{2}\left[\frac{\sin^2(\alpha-\beta)}{\sin^2(\alpha+\beta)} + \frac{\tan^2(\alpha-\beta)}{\tan^2(\alpha+\beta)}\right] \qquad (97)$$

The reflectivity is low at small incident angles (R=0.040 for n=1.5 at 10°) and rises sharply at higher ones (R=0.388 for n=1.5 at 80°).

Gloss is the ratio of the reflection of the specimen to that of a standard, for example, a body with n_D=1.567 in the paint industry. The maximum theoretical gloss R_o is given by Equation (97). It increases with increasing refractive index of the specimen and with increasing angle of incidence (Fig. 57). The maximum theoretical gloss is almost never achieved because of surface roughnesses.

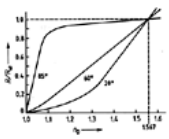

Figure 57. Maximum theoretical gloss as a function of the refractive index of specimen for various angles of incidence α and a standard with n_D=1.567

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

The maximum *transparencies* $\tau_i=1-R_o$ can be calculated from $R_o=(n-1)^2/(n+1)^2$ by Fresnel's law for $\alpha \longrightarrow 0$ and $\beta \longrightarrow 0$. They can have values between 98.0 % (n=1.33) and 92.8 % (n=1.73) for polymers, (i.e., between 2 and 7.2 % of the indicent light is maximally reflected at the polymer – air interface). These ideal transparencies are rarely achieved for polymers because a small portion of the light is always absorbed and/or scattered. Poly(methyl methacrylate), one of the most transparent plastics, never exceeds a transparency of ca. 93 % between wavelengths of 430 and 1100 nm (theory 96.1 %). Transparencies decrease above wavelengths of ca. 1150 nm and (rapidly) below 380 nm. All polymers except halogenated polyethylenes absorb infrared radiation.

A distinction is usually made between transparent and translucent polymers. *Transparent* polymers have transparencies > 90 %; they look clear even at greater thicknesses. *Translucent* polymers (τ_i<0.90) appear clear only as thin films. They are also called contact clear because films look clear in contact with packaged goods but turbid when viewed alone.

An additional loss of clarity is caused by light scattering. Electromagnetic waves lose part of their energy by scattering in inhomogeneous systems. The loss of contrast by forward scattering is called *haze*. The combined loss by forward and backward scattering makes a specimen milky. A body appears opaque if local fluctuations of refractive indices or orientations of anisotropic volume elements are present. The different volume elements must also be larger than the wavelength of incident light. The clarity of a material can thus be increased considerably if the size of the different volume elements (e.g., microdomains) is decreased. Diminished differences in refractive indices improve

the clarity to only a small extent. Lamellar structures are optically less heterogeneous than spherulites of approximately the same diameter. Under certain conditions, clear polyethylene films can be produced by quenching and orientation, although crystalline lamellae with dimensions greater than the wavelength of light are present in these films.

[Top of Page]

11. Polymer Composites (→ Composite Materials)

11.1. Introduction

11.1.1. *Overview*

The term composite has various meanings. In general, it denotes a complex material in which two or more distinct substances combine to produce some properties not present in any individual component. In biomaterials, it means any two joined materials (e.g., polymer- coated titanium parts). In engineering, composites are more narrowly defined as physical admixtures of various materials that are present as distinct phases. The term “phase” is used here in the descriptive sense, not in the thermodynamic one. The engineering term “composite” thus includes only heterogeneous composites. Heterocomposites can be further subdivided into microcomposites (dispersed phase in the micrometer to millimeter range) and macrocomposites. Typical nonpolymer heteromicrocomposites are metal alloys; a typical heteromacrocomposite is steel-reinforced concrete.

Polymer composites are defined as composites in which at least one component is of a polymeric nature. In accordance with the general definition of a composite and in contrast to the engineering use of the word, the term “polymer composite” is often used to cover not only heterogeneous mixtures of a polymer and another material (minerals, fibers, other plastics, elastomers, etc.), but also homogeneous (singlephase) materials of two polymers (homogeneous polymer blends). Such polymer composites are a subgroup of “multicomponent polymer systems”, which also include copolymers (two or more types of monomeric units chemically bound together).

Polymer composites occur naturally (wood, bamboo, lobster shells, bone, muscle tissue, etc.) or are manufactured synthetically. Synthetic polymer composites can also be subdivided into single-phase (homogeneous) and multiphase (heterogeneous) composites. Further subdivisions may be according to the second component's size (microcomposites, macrocomposites), chemical nature (air, low molar mass plasticizer, elastomer, plastic, mineral), geometry (particulate, fiber, platelet, fabric), and macroconformation (if polymeric: coil, rod), etc. (see Fig. 58). Most of these subclasses have different technical names depending on whether the glass transition temperature of the continuous phase is above (plastics) or below (rubbers) the use temperature. The same name is often given to various subtypes.

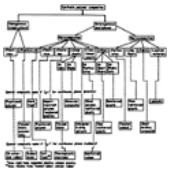

Figure 58. Classification of polymer composites of a base polymer (matrix polymer) with other materials (plasticizer, plastic, air, long fiber, layered materials). Some of these materials have their molecules in different shapes (coils, rods) or have various outer shapes (particulates, short fibers, platelets). Adopted from [1].

A mixture of two (or more) different polymers is often called a *blend* when both polymers are either above (rubber) or below (plastics) their glass transition temperatures. Blends are distinguished from composites of two polymers such as polymer – fiber-reinforced or rubber-toughened plastics. The term blend is sometimes restricted to mean only “incompatible polymer mixtures”, whereas compatible polymer mixtures are called polymer alloys. The latter term is, however, occasionally more narrowly used for mixtures of two crystallizable polymers.

The term *miscibility* refers to admixtures on the molecular level (nanometer range), that is true thermodynamic solubility. It is sometimes used more loosely for any polymer mixture that exhibits only one glass transition temperature. *Compatibility* denotes the ability of admixtures to be blended to heterogeneous microcomposites that do not separate into macroscopic phases.

11.1.2. *Mixture Rules*

The composition dependence of properties Q of composites can often be described by mixture rules. The *generalized simple mixing law* relates Q to the fractions f_A=1–f_B of components A and B with properties Q_A and Q_B:

$$Q^n = Q_A^n \cdot f_A + Q_B^n \cdot f_B \tag{98}$$

The generalized simple mixing law is usually applicable to single-phase composites without specific interactions of its components and to two-phase systems with regularly dispersed discrete phases of “infinite” dimensions. It includes three special cases that are known by different names in different fields (Table 31) and are illustrated in Figure 59.

Table 31. Names of simple mixing rules

Field	Property exponent in generalized mixing law		
	***n*=1**	***n*=0**	***n*=– 1**
Mathematics	arithmetic mean	harmonic mean	geometric mean
Chemical engineering	rule of mixtures	logarithmic mixture rule	inverse rule of mixtures
Mechanical engineering	Voigt model		Reuss model
Electrical engineering	parallel		series
Materials science	upper bound		lower bound

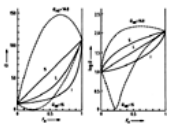

Figure 59. Mixing laws for composites with Q_A=110 and Q_B=10: left: Q=f (f_A); right: log Q=f (f_A)

Simple mixing laws with interaction parameter

Q_{AB}=0 (S=simple mixing law; L=logarithmic mixing law; I=inverse mixing law

- - - Composites with interaction parameters Q_{AB}=140 and 15, as indicated

The *rule of mixtures* (n=1) applies, for example, to various moduli if both components behave elastically. An example is Young's modulus parallel to the fiber direction for epoxide resins reinforced with long glass fibers (see below). The same composites follow the *inverse rule of mixtures* (n=–1) if the modulus is measured perpendicular to the fiber direction and no slippage occurs. The *logarithmic mixture rule*

$$\log Q = f_A \cdot \log Q_A + f_B \cdot \log Q_B \tag{99}$$

is obtained from Equation (98) for n=0 after some mathematical manipulation. The only known case concerns the permeation of oxygen through ABS resins. The proper fraction f_i in Equations (98) and (99) is always the volume fraction. If weight fractions are used instead, deviations from the curves shown in Figure 59 occur, which may be confused with either deviations from additivity or specific interactions.

True deviations from the upper bound S and lower bound I (Fig. 59) may be caused by irregularities in phase distributions (discrete phases) or phase continuities (continuous phases), phase alignments (anisotropic phases), bonding (adhesion), or nonuniform stresses. The resulting functions always lie between the upper and the lower bound if specific interactions are absent. Many empirical, semiempirical, and theoretical expressions have been proposed for such composites. A widely used function is the Halpin – Tsai equation

$$Q = Q_B \cdot \frac{1 + K_1 K_2 f_A}{1 - K_2 f_A};\ K_2 = \frac{Q_A - Q_B}{Q_A + K_1 Q_B} \tag{100}$$

where K_1 is an adjustable constant. The Halpin – Tsai equation gives the lower bound for $K_1 \rightarrow 0$ and the upper bound for $K_1 \rightarrow \infty$. Particulate fillers often have $K_1 \approx 0$, and

fiber-reinforced plastics $K_1 \approx 2\ L/d$, where 2 L/d is the aspect ratio of the fiber.

Specific interactions between components can be considered by an additional interaction term. This term is differently defined, for example:

$$Q = Q_A f_A + Q_B f_B + K_I f_A f_B; \quad K_I = 2(2Q_{AB} - Q_A - Q_B) \tag{101}$$

where the property term Q_{AB} is taken at $f_A = f_B = 1/2$. The value Q_{AB} may be an interaction parameter between components A and B in homogeneous mixtures or the property of the interphase between A and B phases in heterogeneous mixtures. The upper bound for absent interactions or interphases is retrieved for the condition $Q_{AB} = (Q_A + Q_B)/2$. The functions for $Q_{AB} \neq (Q_A + Q_B)/2$ may lie fully or partially outside the range between upper and lower bounds (Fig. 59); that is synergistic effects (above the upper bound) or antagonistic effects (below the lower bound) may be present. They may also show maxima and minima, depending on the relative magnitudes of Q_{AB}, Q_A, and Q_B. Maxima or minima that exceed the highest property value (either Q_A or Q_B) are sometimes considered the true criteria for synergistic and antagonistic effects, respectively.

11.2. Filled Polymers

11.2.1. *Microcomposites*

The addition of fillers (low thermal expansion) to polymers (high thermal expansion) reduces the proportion of the latter and thus the shrinkage of plastics on processing (Table 32).

Table 32. Effect of fillers on some properties of amorphous (A) and crystalline (C) thermoplastics [(⇑) weak increase, ⇑ increase, ⇑ ⇑ strong increase, (⇓) weak decrease, ⇓ decrease, ⇓ ⇓ strong decrease]

Property	Extenders		Reinforcing agents	
	A	C	A	C
Shrinkage	⇓	⇓	⇓	⇓
Heat deflection temperature		(⇑ ⇓)		⇑
Melt flow	⇓	⇓	⇓	⇓
Young's modulus	(⇑)	(⇑)	⇑	⇑
Flexural modulus	⇑	⇑	⇑	⇑ ⇑
Fracture strength	⇓	⇓	⇑	⇑
Brittleness	⇑	⇑	⇑ ⇓ *	⇑ ⇓ *

* Tough plastics become more brittle and brittle plastics become tough.

The melt flow is also lower.

Heat deflection temperatures (HDT) of amorphous polymers are either the same as those of unfilled plastics or no more than ca. 10 K higher (Fig. 60), group a), probably because of the increased viscosity of the filled plastic. The HDTs of amorphous thermoplastics are almost identical with glass transition temperatures.

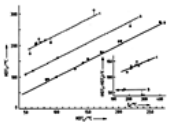

Figure 60. Heat deflection temperatures HDT_F of thermoplastics filled with 30 wt % short glass fibers as a function of heat deflection temperatures HDT_u of the corresponding unfilled polymers, both at 1.85 N/mm^2 (ISO/R 75 A)

(•) Amorphous polymers; (○, ○) crystalline polymers; solid line: $HDT_F = HDT_u$; broken lines: empirical

a) Amorphous polymers; b) Apolar crystalline polymers; c) Polar crystalline polymers

Insert: difference $HDT_F - HDT_u$ as function of melting temperatures T_M of unfilled crystalline polymers

The HDTs of crystalline polymers depend on a number of factors. Unfilled or particulate-filled crystalline polymers show HDTs that are often considerably lower than their melting temperature. Upon fiber-filling, HDTs of crystalline polymers increase strongly; they may reach values near the melting temperatures. This behavior is thought to be due to additional polymer crystallization near, or polymer adsorption on, the fiber surface, epitaxial layers have been found to extend from the filler up to 150 nm into the matrix. The surface layers overlap; the resulting physical network reduces creep.

Two groups of HDTs of glass-fiber-filled crystalline polymers can be distinguished (Fig. 60). The HDTs of group b are about 50 –60 K above those of their unfilled counterparts. This group comprises apolar polymers such as polyethylene, polypropylene, and poly(phenylene sulfide); polyoxymethylene also belongs to group b because its oxygen groups are buried deep inside compact helical structures, which are thus apolar on their surfaces. The difference in HDTs of filled and unfilled group b polymers does not vary with melting temperature. Group c includes polar crystalline polymers (polyesters, polyamides, etc.). For this group, the difference in HDTs of filled and unfilled polymers increases with the melting temperature of the unfilled polymer.

Young's moduli of thermoplastics increase with filler content (Fig. 61). The increase is small for spheroidal fillers (glass spheres, chalk), stronger for platelets (kaolin, talc), and strongest for fibers (glass). Short fibers are usually added in amounts of ca. 30 wt %, which corresponds to ca. 17 vol % for glass fibers (ρ=2.55 g/cm^3) in an average polymer (ρ≈1.2 g/cm^3). Higher fiber content often leads to packing problems since the maximum volume fraction of three-dimensional randomly packed rods with L/d=20 is already ca. 25 % (glass fibers typically have diameters d=10 – 20 µm and lengths of ca. 0.2 mm after injection molding). Particulate fillers can be added in far greater amounts without sacrificing properties (see Table 19) because maximum packing densities of spheres are, for example, 0.745 for the hexagonally most dense and 0.637 for the randomly most dense packing.

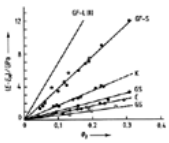

Figure 61. Influence of the volume fraction Φ_F of fillers on the difference in Young's moduli of filled thermoplastics (E) and their polymer matrices (E_M)

GS=glass spheres; C=chalk; K=kaolin; GF=glass fibers; (S=short; L=long)

○=polyethylene; ○=polypropylene; ⊗=polyoxymethylene; ⊕=poly(butylene terephthalate); •=polyamide 6

The solid line for GF-L (||) corresponds to the simple mixing law. From data of [82]

Young's moduli E of filled materials can be described by a modified mixing law

$$E = E_M \phi_M + f E_F \phi_F = E_M + (f E_F - E_M) \cdot \phi_F \tag{102}$$

where f is an adjustable parameter. The data in Figure 61 can be described with f=0.39±0.03 (chalk), 0.66±0.04 (kaolin), and 1.61 (talc). The effect of 30 wt % of short glass fiber was tested in 14 thermoplastics, and f was found to be 0.57±0.04. These numbers increase with the specific area of the fillers, which are given as 0.3 – 2.2 (ground limestone), 0.5 – 1 (glass fibers), 6 – 22 (kaolin), and 6 – 17 m^2/g (talc) [83]. Composites obviously increase in stiffness as the contact between filler and polymer segments increases. Good adhesion is not necessary, since E is measured for strains approaching zero.

The adjustable parameter f of Equation (102) adopts a physical meaning for short fiber reinforced plastics. In such composites, forces are transferred to the hard fiber from the soft matrix if (1) the Young's modulus of the fiber is greater than the Young's modulus of the matrix ($E_F > E_M$), (2) the fiber length L_F exceeds a certain critical value $L_{F,\,crit}$, and (3) the shear strength in the matrix is smaller than the shear strength between fiber and matrix (the fiber would be otherwise pulled out of the matrix unless the fiber breaks). In this case, $f = 1 - [L_{F,\,crit}/(2\ L_F)]$.

Fracture strengths are, however, measured for maximum strains. The failure mode of filled thermoplastics is matrix dominated. Electron micrographs show that at low fiber contents, planar cracks spread through the matrices and debonded fibers are pulled out of either surface; the matrix must thus adhere well to the filler.

The tensile strength σ_B at break is given by the rule of mixtures ($\sigma_B = \sigma_F\ \Phi_F + \sigma'_{M'}\ \Phi_M$) where $\sigma'_{M'}$ is the tensile stress of the matrix upon application of a load σ_B to the composite. This relationship applies if (1) the tensile strength σ_B is greater than the contribution $\sigma_M\ \Phi_M$ of the matrix component to the total tensile strength σ_B (that is $\sigma_B > \sigma_M\ \Phi_M$), and (2) the volume fraction Φ_F of the fiber exceeds a certain minimum value ($\Phi_F > \Phi_{F,\,min}$).

Particulate fillers usually strongly reduce the fracture strengths of composites of amorphous thermoplastics, whereas those of crystalline polymers are reduced less or not at all. *Short fibers* generally increase the fracture strengths by factors of 1.5 – 2 for amorphous and crystalline polymers alike. Rigid PVC is an exception because its fracture strength is reduced on fiber-filling. The moduli and strengths can be further enhanced if coupling agents are used, which improves the bonding between fiber and matrix (e.g., silanes for glass-fiber-filled polymers).

11.2.2. *Molecular Composites*

The problem of insufficient bonding between fiber and matrix can be dramatically reduced if so- called molecular composites are used. These materials are dispersions of semiflexible (rodlike) polymer molecules with mesogenic units in chemically similar polymer matrices, e.g., poly (*p*-phenylenebenzobisthiazoles) in polybenzimides (see Fig. 52). The mesogens are not truly molecularly dissolved; rather, they form microdomains in the matrix. The segment axes are randomly distributed in these domains if the molecular composites are formed from polymer solutions with total polymer concentrations *c* below the critical isotrope phase – mesophase concentration $c_{i/n}$. Microdomains show mesophase behavior if films or fiber are formed at $c > c_i/_n$. Such microcomposites have excellent moduli and fracture strengths (Fig. 52).

11.2.3. *Macrocomposites*

Polymer macrocomposites are heterogeneous composites of polymers and “macro”-sized additives. Examples are polymer concrete, composition board, long fiber-reinforced polymers, and plywood.

Polymer concrete is composed of polymer [usually polyester or poly(methyl methacrylate)], aggregate (sand, ground limestone, etc.), and other additives (dyes, etc.); it does not contain cement. *Polymer- concrete hybrids* include polymer – cement hybrids (cement is partially replaced by polymers) and polymer-impregnated concrete (conventional concrete saturated with a monomer that is polymerized in place). Polymer concretes have higher moduli and higher tensile and compressive strengths than ordinary concretes. They are not vulnerable to damage by freeze-thaw cycles.

Composition boards are panel products. They include particleboard, fiberboard, hard board from flakes or shearings, etc. The fillers are mainly wood products; the resins, urea –formaldehyde (interior applications), and polyurethanes or phenolic resins (exterior).

Long-fiber-filled polymer macrocomposites range in structure from polymers filled with “long” fibers 6 – 12 mm in length to fiber strands and mats filled with polymers. The filler content can often reach 65 wt %. Upon application of tensile stress, local stress concentrations are transferred by shear forces onto the plastic – fiber interface and distributed over the much greater fiber surface area. The fibers must therefore bond well to the plastic (use of coupling agents) and must have a certain length to avoid slipping.

Young's moduli of long glass-fiber-reinforced epoxide resins follow the simple mixing law if the tensile stress is in the fiber direction (Fig. 62). The inverse mixing law is not observed as well for stresses perpendicular to the fiber direction since some fibers bend under stress.

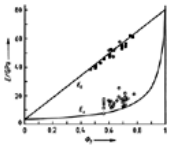

Figure 62. Young's moduli of epoxy – glass-fiber composites parallel (||) and perpendicular (⊥) to the fiber direction.

The simple mixing law corresponds to *f*=1 in Equation (102). For long fibers, *f* has the meaning of an orientation factor. It becomes 1/6 for fibers distributed three-dimensionally at random, 1/2 for 90° cross-plies measured in either fiber direction, and 3/8 for a uniform, planar distribution of fibers.

This group of macrocomposites also includes prepregs (fiber mats soaked with unsaturated polyester resins) and the composites generated by filament winding (wound filaments saturated with epoxies or unsaturated polyesters), both of which are subsequently polymerized.

11.3. Homogeneous Blends

Mixtures of polymers with other polymers are called polymer blends (⟶ Polymer Blends). Such blends may be composed of two thermoplastics (plastic blends), two elastomers (rubber blends), a plastic filled with an elastomer as the dispersed phase (rubber-modified plastics), an elastomer with a plastic as the dispersed phase (polymer-filled elastomer), or a plastic filled with a polymer melt or a low molar mass liquid (plasticized polymers). Whether or not these blends are truly miscible on a molecular level depends on the thermodynamics of the systems.

11.3.1. *Thermodynamics*

If two components 1 and 2 are mixed, entropy *S*, enthalpy *H*, and volume *V* change. In many cases, volume changes can be neglected and the Gibbs energy of mixing is given by the changes in enthalpy and entropy: $\Delta G_{mix} = \Delta H_{mix} - T \cdot \Delta S_{mix}$. These changes can be calculated by the Flory – Huggins theory, which assumes that the mixture can be modeled by a three-dimensional lattice on which the monomeric units (or solvent molecules) can be placed. Each monomeric unit experiences the same force field; the theory is thus a mean-field theory. The assumption of a constant force field is fairly well fulfilled for polymer blends and concentrated polymer solutions but not for dilute solutions.

The entropy and enthalpy of mixing are calculated separately. *Entropies* are calculated for ideal solutions (zero mixing enthalpy); all environment-dependent entropy changes are thus zero (translational, vibrational, and inner rotational entropies do not change). Units can however be arranged relative to each other in many ways; however, there is a combinatorial entropy that is also called “configurational entropy” in statistical mechanics.

The mixing of two components is modeled as a quasi- chemical reaction. The *mixing enthalpy* is thus determined by the change in interaction energies and the number and type of nearest neighbors. Enthalpy and entropy terms are added to give the molar Gibbs energy of mixing:

$$(\Delta G_{mix})^m / RT = \chi \cdot \phi_1 \phi_2 + \phi_1 \cdot \ln \phi_1 + (X_1/X_2) \cdot \phi_2 \cdot \ln \phi_2 \qquad (103)$$

which, on differentiation with respect to the molar amounts of the components, yields the chemical potentials, for example, of component 1

$$\Delta\mu_1 = RT\{\chi \cdot \phi_2^2 + \ln(1-\phi_2) + [1-(X_1/X_2)] \cdot \phi_2^2\} \qquad (104)$$

where χ is the parameter for the interaction between component 1 and component 2 (usually with numerical values between 0 and 2), also called the Flory – Huggins parameter. All three terms in braces must be considered if a polymer 2 is mixed with a solvent 1, for example, a plasticizer ($X_1 << X_2$). If both components are polymers however, then $X_1 \approx X_2$ and the entire entropy contribution comes from the relatively small logarithmic term. Thus little combinatorial entropy is gained on mixing two polymers, and the miscibility is determined mainly by the mixing enthalpy, that is, by the term $\chi\,\phi_2^2$.

Phase separation occurs if the curve ($\Delta G_{mix})^m = f(\Phi_2)$ (at *T*=const.) has a shape that can be touched at two points by a tangent. The compositions at these points are Φ_2' and Φ_2'' (with $\Phi_2' < \Phi_2''$). Two phases exist for the unstable range $\Phi_2' < \Phi_2 < \Phi_2''$, which is separated from the stable ranges $\Phi_2 < \Phi_2'$ and $\Phi_2'' > \Phi_2$ by the so- called binodal. The unstable range itself is subdivided by the spinodal into two metastable ranges and one unstable one. Spinodals are characterized by inflection points in the Gibbs energy – composition functions and extremal values in the chemical potential – composition functions. Maximum, minimum, and inflection point become identical at the critical point, which is defined by a zero value of the second derivative of the chemical potential

$$\partial^2 \Delta\mu_1 / \partial\phi_2^2 = RT \cdot \left[2\chi - (1-\phi_2)^{-2}\right] = 0 \qquad (105)$$

This critical point is characterized by a critical volume fraction $\phi_{2,crit} = 1/\left(1 + X_2^{1/2}\right)$ and a critical interaction parameter $\chi_{crit} \approx (1/2) + (1/X_2)^{1/2}$.

The temperature dependence of χ is approximately

$$\chi \approx A + (B/T) \qquad (106)$$

where *A* and *B* are system-dependent constants; *B* is positive for endothermal mixtures [i.e., χ increases with increasing temperature and a homogeneous solution exists above an "upper critical solution temperature" (UCST)]. Other systems are homogeneous only below a "lower critical solution temperature" (LCST). The demixing is correspondingly either enthalpy (UCST) or entropy (LCST) induced.

The terms UCST and LCST do not indicate the absolute position of demixing temperatures since some systems show LCST>UCST (Fig. 63). Depending on the system, either an hourglass-type diagram or a closed miscibility gap is shown.

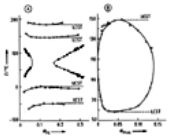

Figure 63. Phase separation temperature as function of polymer – solvent composition

A) Polystyrenes with various molar masses in acetone; B) Poly(vinyl alcohol- co-vinyl acetate) (93 : 7) in water

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

11.3.2. *Plastification*

Plastification of a plastic is the flexibilization of the material either by added plasticizers (external plastification) or by incorporated flexibilizing comonomer units (internal plastification).

The molecular action of plasticizers consists of a flexibilization of chain segments, which can have various causes. *Polar plasticizers* increase the proportion of *gauche* conformations in polar chains, thus decreasing rotational barriers. *Primary plasticizers* can dissolve helical structures and crystalline regions if they are thermodynamically good solvents. Furthermore both *primary and secondary plasticizers* are diluents; their addition increases the distance between polymer segments and thus decreases the activation energy needed for cooperative movement of segments. Solvation does not per se increase chain flexibility, however, because solvent shells act as substituents and increase the rotational barriers.

The molecular effect of an increase in chain flexibility by plasticization is measured as a decrease of glass transition temperatures (Fig. 64). Small molecules are better plasticizers than bigger ones of similar chemical constitution. The plasticizer efficiency decreases with increasing thermodynamic "goodness" of the plasticizer for the polymer (greater interaction, i.e., bigger solvation). Good plasticizers with respect to lowering glass transition temperatures are thus small molecules that act as theta solvents. These plasticizers are however bad for technological applications because they bleed (transport of plasticizer to the surface of the plastic), migrate (transport of plasticizer into another contacting material), or are extractable (e.g., transport into packaged foods). These transport phenomena can be reduced if polymer plasticizers are used. Increasing molar masses lead to decreasing thermodynamic compatibility, however (see below). Technical plasticizers are therefore in most cases a compromise between thermodynamic plasticizer efficiency and kinetically hindered plasticizer transport.

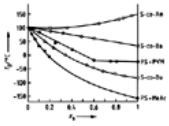

Figure 64. Glass transition temperatures of (•) polystyrene plasticized with methyl acetate (MeAc) or poly(vinyl methyl ether) (PVM) and (○) copolymers of styrene with butadiene (Bu), butyl acrylate (Ba), or acryl amide (Am)

Plastification can also be followed by the changes of certain mechanical properties such as lowering of fracture strength, increase of elongation, and decrease of tensile moduli. Such methods sometimes show "antiplasticizations" at low plasticizer concentrations (Fig. 65) (e.g., increases of moduli). Glass transition temperatures are not increased, however. The effect thus cannot be the result of increased segment mobility. It may be due to a healing of microvoids in amorphous polymers such as polystyrene or to additional crystallization of slightly crystalline polymers such as poly(vinyl chloride).

Figure 65. A) Glass transition temperature T_G and B) fracture strength σ_B and elongations ϵ_B at break of a poly(vinyl chloride) plasticized with tricresyl phosphate (TCP).

Reprinted with permission by Hüthig and Wepf Publ., Basel [7]

11.3.3. *Rubber Blends*

About 75 % of all elastomers are employed as blends. Heterogeneous blends of natural (rubber – styrene) – butadiene rubber or (*cis*-butadiene rubber – styrene) – butadiene rubber are used mainy for tire treads since they reduce abrasion. Homogeneous blends are more rare. They possess only one glass transition temperature which, for 50 – 50 blends, is an average of the glass transition temperatures of the two parent polymers. Heterogeneous blends exhibit two glass transition temperatures, both practically identical to those of the parent polymers. Homogeneous blends are usually formed if the difference in solubility parameters is smaller than 0.7 $cal^{1/2}\ cm^{-3/2}$ (1.44 $J^{1/2}\ cm^{-3/2}$).

11.3.4. *Plastic Blends*

Plastic blends are usually immiscible because the combinatorial entropy of mixing is too small (high molar masses) and the enthalpy changes on mixing are often positive. Miscibility is observed for three types of systems:

1. Chemically similar polymers, for example, polystyrene – poly(*o*- chlorostyrene), showing both LCST and UCST
2. Systems having specific interactions between different components, for example, polystyrene – poly(vinyl methyl ether); these systems show only LCSTs
3. Systems consisting of oligomers, for example, oligo(ethylene oxide) – oligo(propylene oxide); these systems possess only UCSTs

Homogeneous polymer blends show only single glass transition temperatures; the composition dependence of these T_G's can be described by the simple rule of mixtures. Two-phase systems always show two glass transition temperatures if the phase diameters are greater than ca. 3 nm. Transparency, on the other hand, is no indication of a single-phase blend since opacity can be observed only if the refractive indices of the two phases are sufficiently different and if the phase diameters are greater than ca. one-half the wavelength of incident light. The presence of two phases in clear specimens can often be detected with electron microscopy by special staining techniques.

Modified PPO, a blend of poly(2,6-dimethylphenylene oxide) and polystyrene, is the industrially leading *homogeneous blend*. Blends of poly(vinyl chloride) with poly(methyl methacrylate), poly(ethylene-co-vinyl acetate) or chlorinated polyethylene are also said to be one-phase systems.

11.4. Heterogeneous Blends

11.4.1. *Compatible Polymers*

Two polymers may be thermodynamically immiscible but still mechanically compatible. These compatible blends are two- or multiphase systems that show multiple glass transition temperatures, which do not change with composition. The components do not separate, however, under mechanical stress during the expected life of the product. Such mechanical compatibilities can be improved or introduced by the addition of diblock polymers. One block of these diblock polymers is compatible with one phase type of the heterogeneous blend and the other one with the other type (see Section Block Copolymers). The two blocks need not necessarily be chemically identical with the two parent polymers, that is, a compatibility of poly (A) and poly (B) can be achieved not only by addition of poly (A)-*block*-poly (B) but also by poly (A)-*block*-poly (C) if the C block is either miscible with poly (B) or can form mixed crystals with poly (B).

11.4.2. *Blend Formation*

Heterogeneous polymer blends can be produced by mixing together two polymers (as melts, lattices, or in solution) or by in situ polymerization of a monomer in the presence of a dissolved polymer.

The energy taken up during *melt mixing* is used for flow processes and for the generation of surfaces of new microdomains. After some time, a steady state is established and the domain size becomes constant. No macroscopic demixing occurs because of the low diffusion coefficients resulting from the high viscosities.

Latex blending consists of the mixing of aqueous dispersions of two polymers. Far lower temperatures and lower shear fields can be employed compared to melt blending. The good mixing of the latex particles remains after coagulation. The domain size is, however, restricted to the size of the latex particles themselves; it is not altered by subsequent melting of the coagulate.

Solution blending involves the mixing of two polymer solutions. Miscible polymers can be blended to domains of molecular size. Solutions of immiscible polymers demix, however, at very low concentrations, sometimes under fractionation with respect to molar masses. Domains grow further on solvent removal by distillation or freeze drying.

In situ polymerization involves solutions or gels of polymers in monomers, which are subsequently polymerized. The in situ polymerization of styrene in a styrenic polydiene solution, which leads to rubber-toughened polystyrene (HIPS) is most important industrially. The polymerization of a cross-linkable monomer in a gel of a cross-linked rubber in the very same monomer results in interpenetrating networks.

Various blending processes produce blends with very different properties (Fig. 66). The high notched impact strengths of the polymerization-blended materials (line P) are caused by a strong anchoring of phases due to the formation of graft copolymers of styrene on polybutadiene and cross-linking within the rubber domains, both caused by free-radical initiators. Such processes are less prevalent during melt blending (radical formation by high shearing at elevated temperature) (line M) and latex blending (low shearing at ambient temperature) (line L). In situ polymerization is thus the method of choice for unsaturated rubber (easy cross-linking and grafting by chain transfer) and monomer – rubber pairs with favorable *Q*, *e*-values for copolymerizations. In all other cases, blends are formed by melt mixing.

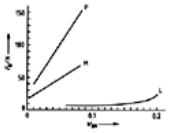

Figure 66. Impact strength with notch F_B as function of the mass concentration w_{BR} of *cis*-1,4-polybutadiene in its blends with polystyrene. P=In situ polymerization of styrene; M=Melt blending; L=Latex blending

Reprinted with permission by Plenum, New York [84]

11.4.3. *Toughened Plastics*

Rubber-modified plastics (toughened plastics, high-impact plastics) consist of rubber domains dispersed in plastic matrices. The domain size varies with the blending process. Typical values are ca. 0.1 µm for melt-blended poly(vinyl chloride) – acryl rubber and 1 µm for polystyrene – polybutadiene. The domains are often multiphased; small plastic domains are imbedded in the rubber domains (Fig. 67). The morphologies depend strongly on the blending process. During stirred in situ polymerizations, phase inversions often occur if the amount of newly formed plastic approaches that of the incipient rubber. During phase inversion, already present plastic particles may be embedded in the newly formed rubber domains.

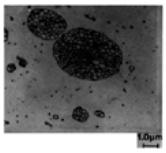

Figure 67. Rubber domains in high-impact polystyrene by in situ free-radical polymerization of styrenic polybutadiene solutions

Reprinted with permission by Society of Plastics Engineers [85]

Rubber-toughened plastics are valued because of their improved impact strength. On impact, very many crazes are formed near the equators of the rubber particles. The crazes propagate until they encounter an obstacle (e.g., rubber particle or shear band) or until the stress concentration at the tip of the craze becomes very low. Many small crazes result, and the stress is evenly distributed if the rubber phase is cross-linked and binds well to the thermoplast phase [e.g., by in situ formed graft copolymers (HIPS)]. In contrast, stress peaks concentrate at a few defect points in normal thermoplasts.

11.4.4. *Thermoplastic Elastomers (* ⟶ Thermoplastic Elastomers*;* ⟶ Rubber, 3. Synthetic*)*

Thermoplastic elastomers (elastoplastics, thermoplastics, plastomers) are processed like thermoplastics but applied like elastomers. Their unique properties follow from their molecular structures; their chains consist of “soft” and “hard” segments in block, graft, or segmented copolymers composed of monomeric units A and B. The A and B segments are mutually incompatible and form locally separated regions. With a well-designed molecular architecture, domains of the “hard” A segments (transition temperature>service temperature) act as physical cross-links in the continuous matrix of the “soft” B segments (transition temperature<service temperature) (Section Block Copolymers). Such transition temperatures may be glass transition temperatures in amorphous polymers or melting temperatures in partially crystalline polymers. The A segments attain mobility above these transition temperatures and the elastoplasts become processible.

Thermoplastic elastomers comprise linear triblock polymers of the polystyrene – *block*-polydiene – *block*-polystyrene type; radial, star, or “teleblock” polymers of the same monomeric units; urethane segment or block copolymers with polyester or polyether soft segments; polyesteramides and polyesteretheramides; graft copolymers of butyl rubber on polyethylene, vinyl chloride on poly(ethylene-co-vinyl acetate), styrene – acrylonitrile on saturated acryl rubbers, and various ionomers.

Thermoplastic elastomers can also be produced by “dynamic vulcanization.” The mastication of blends of conventional rubbers and crystalline poly(α-olefins) leads to chain scissions. The resulting macroradicals cross-link the rubber domains.

The mechanical properties of thermoplastic elastomers are determined mainly by their morphologies. In styrene – butadiene – styrene triblock polymers, for example, morphologies are governed by the spatial requirements of the various blocks (Section Block Copolymers). With a low content of hard styrene segments, small spherical polystyrene microdomains are formed in the soft polybutadiene matrix. These microdomains act as physical cross-linkers. The distances between the domains are large, and the polymer behaves as a weakly linked elastomer with correspondingly high extension (Fig. 68). The domains are larger and their distances are shorter at 28 % styrene units and the polymer strengthens (stiffens). The stiffening becomes stronger for rodlike styrene domains (39 % S). Lamellar morphologies (53 % S) show an “unruly” behavior initially because of reorientation of lamellae. At even higher styrene content, the polymer behaves as a plasticized tough thermoplast (65 % S) or almost like polystyrene itself (80 % S).

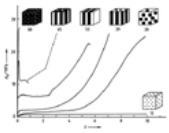

Figure 68. Stress – strain curves and morphologies of styrene – butadiene – styrene triblock polymers with various styrene contents

Reprinted with permission by Hüthig and Wepf Publ., Basel [5]

11.5. Expanded Plastics (⟶ Foamed Plastics)

Expanded plastics (foams, foamed plastics, cellular plastics) are blends of plastics with air. They are subdivided according to their rigidity, their cell structure, and the nature of the parent plastics. Rigid expanded plastics are used mainly for thermal insulation, and flexible foams for damping and cushioning materials.

The rigidity (“hardness”) of expanded plastics follows the properties of the parent polymers. Phenolic and urea resins thus yield brittle-rigid plastics; polystyrene and hard poly (vinyl chloride), tough-rigid; and polyethylenes, polyurethanes, and plasticized PVC, semirigid to flexible foams. The elastic moduli of expanded plastics decrease approximately proportional to their polymer content (simple mixing law). Since, however, the stiffness of an article increases with the third power of the wall thickness, the gases in expanded plastics work as enhancers that reduce material costs and weights.

Tensile strengths of foamed plastics also follow the simple mixing law. Because the tensile strengths of gases are diminishingly small, tensile strengths σ_B of foamed plastics are determined by the volume fraction Φ_p and strengths $\sigma_{B,p}$ of the plastics themselves [i.e., $\sigma_B \approx \sigma_{B,p} \cdot \Phi_p = \sigma_{B,p} \cdot (\rho / \rho_p)$] (Fig. 69). Compression strengths follow a similar simple mixing law.

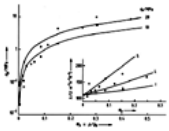

Figure 69. Tensile strengths of expanded plastics as a function of the volume fraction of the rigid (•) or flexible (○) parent polymers P

Solid lines indicate a) Simple mixing law for σ_p=20 MPa; b) Simple mixing law for σ_p=10 MPa;Insert: Thermal conductivities (for λ_p=700 and λ_{air}=110 J (m^{-1} h^{-1} K^{-1}) of rigid (•) and flexible (o) plastic foams S: Simple mixing law; L: Logarithmic mixing law; I: Inverse mixing law

The cell structure can be open, closed, or mixed. Open- celled foams are always air filled, regardless of the blowing or expanding gases used for foam manufacture. The trapped gases in closed- cell structures can be exchanged with the surrounding air only by slow diffusion through the polymer matrix. Since thermal conductivities of low-density foams (ρ_p<0.3) obey, to a first approximation, a logarithmic mixing law, they are considerably affected by the thermal conductivities of trapped gases [(λ=93 J m^{-1} h^{-1} K^{-1}

(nitrogen), 56 J m^{-1} h^{-1} K^{-1} (CO_2), 26.4 J m^{-1} h^{-1} K^{-1} (CCl_3F)]. The gas exchange is considerably reduced in integral skin foams (dense skin, low-density core) and in syntactic foams (plastics filled with hollow spheres; spheres may be made of glass, ceramics, or plastics and contain either gases or a vacuum).

[Top of Page]

12. Waste Disposal (→ Plastics, Recycling)

Four different, although frequently interconnected, aspects must be considered in the disposal of plastic waste: (1) technological demands, (2) economic considerations, (3) energy consumption, and (4) environmental concerns.

Modern technology demands certain material properties for certain applications that can be delivered only by polymers and not by other materials such as metals or ceramics. Examples range from rubber tires to high-tech applications such as polymer resists in electronic chip production. The desirability of polymers for technology-driven demands is usually undisputed; the discussion centers rather around the desirability of commodity plastics, especially for packaging purposes.

It is also not much disputed, at least among people that are well-versed in technology, that polymeric materials are very economical compared to other materials in terms of production and use. However, the ecological costs are questioned: the hidden costs of the “free” use of air, water, and land in the production and disposal of polymers and the suspected high energy consumption in the use of polymers.

Energy consumption has been studied by consumer agencies (e.g., NATO Science Committee), industry (e.g., BASF), and government (e.g., Swiss Environmental Protection Agency). Although differences exist, newer data all agree that the energy consumed in plastics production agrees favorably with that of other materials (Table 33) if use (on a volume basis) rather than sale (on a weight basis) is considered.

Table 33. Energy consumption for the production of various materials according to various sources

Material	Consumption, kJ/cm^3		
	NATO	BASF	SEPA *
Lumber	2.0		
PE-HD		63	68
PVC	11	73	84
ABS		89	87
PET		113	
PS, expanded		1.4	
PUR, expanded		3.0	
Paper (kraft, unbleached)	40		63
Tin plate			210
Aluminum	< 460		750
Glass	< 125		> 26

* Swiss Environmental Protection Agency.

Energy considerations must however be based on specific applications of materials. For example, these considerations must take into account for a certain good the current possible use of different materials for its manufacture, its use (including cleaning), possible recycling, and/or disposal by landfill or incineration (with partial use of heat). This is shown for the example of 1000 1-L milk packages (in Table 34).

Table 34. Energy consumption and critical environmental data for 1000 1-L packages of milk [86]

	PE		Paper	Glass
	Tube	Bottle	carton	bottle
Use cycle (life)	1	1	1	40
Weight per life in kg	7.0	22.0	25.0	40.0
Energy consumption in MJ *	400	1510	1770	650
Environmental impacts				
Air, 10^6 m^3	5.9	20.5	30.9	5.4
Water, m^3	3.0	17.2	154	9.3
Solids (landfill), L	2.3	9.8	8.3	6.8
Solids, (average Swiss disposal), L **	0.35	3.0	1.3	3.3

* For average Swiss disposal: 23 % landfill and 77 % incineration (55 % with use of energy).
** Solid waste from PE (ash from production processes), paper (ash from production and paper), glass (with 60 % recycling assumed), aluminum and tin plate (100 %, no recycling).

Energy data are again favorable for plastics, as are the critical amounts of air and water employed in their production, use, and disposal (i.e., volumes to which air and water can be maximally charged without exceeding Swiss environmental laws). It must be mentioned in this context that only ca. 5 % of the world petroleum production is used for the manufacture of polymers; the remainder is burned (as gasoline, heating oil, etc.), often with very low energy efficiency (ca. 30 % in cars) and large amounts of harmful byproducts (exhaust gases).

Environmental concerns include possible harmful effects by plastic articles, plastics themselves, polymers, incineration products, and byproducts of polymer and plastics manufacture. Environmentalists estimate the amount of plastics discarded annually into the high seas as 1×10^6 t. Certain plastics articles are undoubtedly harmful to wildlife, for example, six-pack rings and old fishing nets to fish and other animals. Since man is notoriously careless, the use of biodegradable plastics for such articles may be justified.

Biodegradable plastics either contain biodegradable polymers themselves or are compounds of conventional polymers with biodegradable polymers such as starch. Biodegradation may occur via hydrolysis (starch, polyesters) or UV radiation (polymers containing carbonyl groups) or both. The degradation however is slow, taking from several months under laboratory conditions to several years in landfills. Biodegraded products may no longer be visible, but whether biodegradation proceeds to small molecules and whether these molecules are harmful to living organisms are still not known. Harmful polymer additives may be released during biodegradation. Conventional polymer molecules, on the other hand, are not harmful to living beings, at least not to higher ones, because they can be neither resorbed nor digested. The energy content of plastics is of course wasted during biodegradation.

Biodegradable polymers thus do not seem to be the general solution to the plastic waste problem. Contrary to popular opinion, the amount of plastics in household refuse is relatively small, ca. 7 wt % vs. ca. 50 % paper, and almost constant over at least 25 years according to archaeological studies of landfills. Weight fractions and not volume fractions of plastics must be used for comparison with other materials because hollow plastic containers and expanded plastics are compressed by the weight of the landfill (glass bottles are not compressed). Biodegradation processes are very slow in landfills (absence of light, air, and even microorganisms); some buried papers could be read after 25 years in a landfill.

This leaves recycling and incineration as possible ways for disposing of plastics. Most industrial plastic waste is already recycled or incinerated. The problem is the collection, cleaning, and recycling of household refuse. Only ca. 2 % of all plastics consumed were recycled in the United States (Table 35) and the Federal Republic of Germany in 1988.

Table 35. Recycling of plastics in the United States (1988)

Polymer	Mass in 1000 t		% Recycled
	Consumed	Recycled	
Polyolefins	11 538	343	3.0
PVC	3 779	34	0.9
Styrenics	3 516	39	1.1
PETP (bottles)	911	54	5.9
PA	253	27	10.7
Other	6 219	< 1	< 0.01
Total	26 216	498	1.9

Since unsorted recycled polymers represent mixtures of different grades and even different polymers, they can be used only for low-value articles. For comparison: the gross discards recovered in the United States were 25 % of aluminum, 23 % of paper and paperboard, and 9 % of glass.

The sorting of discarded plastics articles is made more easy by recently introduced codes which are stamped on the bottom of containers or other inconspicuous places. The codes consist of three broken arrows, arranged in a triangle, and a number, sometimes also combined with letters and/or bar code (Fig. 70). The numbers and letter combinations indicate the type of plastics. The letter combinations are not always identical with the acronyms and abbreviations recommended by ISO, ASTM, DIN, IUPAC, etc. for the same polymers. The following number and letter codes are used: 1=PETE [for poly(ethylene terephthalate)], 2=HDPE (for high-density polyethylene), 3=V (for vinyl chloride polymers), 4=LDPE (for low-density polyethylene), 5=PP (for polypropylene), 6=PS (for polystyrene), 7 (for all other plastics, including multilayered materials).

Figure 70. Symbols for recyclable plastics. Left: U.S. symbol with code number and code letters. All others: various symbols used by European manufacturers.

Incineration is the present method of choice in many European communities: 77 % of all household refuse is incinerated in densely populated Switzerland and 30 % in West Germany, but only 10 % in the United States where 80 % of the refuse is accumulating rapidly in landfills. Furthermore, concerns exist about the emission of HCl from the burning of PVC, the generation of dioxins, and the disposal of heavy-metal- containing ash. In the Federal Republic of Germany, only 18 % of the HCl comes from the incineration of refuse (and not more than half of it from PVC); the primary sources of HCl are power plants (1978 data). This HCl is removed by scrubbing, etc., and may be used to bind heavy metals in compost from incineration. Generation of dioxins can be reduced significantly by high-temperature incinerators, which of course command higher investment costs.

[Top of Page]

References

1. H.-G. Elias: “Technologie,” *Makromoleküle*, **vol. 2**, Hüthig and Wepf, Basel 1992.
2. F. W. Billmeyer, Jr.: *Textbook of Polymer Science*, 3rd ed., Wiley, New York 1984.
3. F. Rodriguez: *Principles of Polymer Systems*, 3rd ed., Hemisphere Publ., New York 1989.
4. G. Champetier, R. Buvet, J. Néel, P. Sigwalt, (eds.): *Chimie Macromolèculaire*, **2 vols.**, Hermann, Paris 1970 – 1972.
5. H.-G. Elias: “Grundlagen,” *Makromoleküle*, 5th ed., vol. **1**, Hüthig and Wepf, Basel 1989.
6. *Macromolecules*, 2nd ed., **2 vols.**, Plenum, New York 1984.
7. *Makromoleküle*, 4th ed., Hüthig and Wepf, Basel 1981.
8. L. Mandelkern: *An Introduction to Macromolecules*, 2nd ed., Springer, New York 1972, 1982.
9. J. M. G. Cowie: *Polymers: Chemistry and Physics of Modern Materials*, Intext Educational Publ., New York 1974; *Polymers: Chemistry and Physics of Modern Materials*, 2nd ed., Routledge, Chapman and Hall, New York 1991.
10. F. A. Bovey, F. H. Winslow (eds.): *Macromolecules*, Academic Press, New York 1979.
11. R. B. Seymour, C. E. Carraher, Jr. *Polymer Chemistry*, 2nd ed., Dekker, New York, 1989.
12. A. Rudin: *The Elements of Polymer Science and Engineering*, Academic Press, New York 1982.
13. P. C. Hiemenz: *Polymer Chemistry*, Dekker, New York 1984.
14. P. Munk: *Introduction to Macromolecular Science*, Wiley, New York 1989.
15. International Union of Pure and Applied Chemistry: *Compendium of Macromolecular Nomenclature (Purple Book)*, Blackwell Sci. Publ., Oxford 1991.
16. J. Schultz: *Polymer Materials Science*, Prentice-Hall, Englewood Cliffs, N.J., 1974.
17. J. A. Brydson: *Plastics Materials*, 3rd ed., Butterworth, London 1975.
18. S. Rosen: *Fundamental Principles of Polymeric Materials for Practicing Engineers*, Cahners Books, Boston 1973.
19. J. H. DuBois, F. W. John: *Plastics*, 5th ed., Van Nostrand-Reinhold, New York 1976.
20. W. E. Driver: *Plastics Chemistry and Technology*, Van Nostrand-Reinhold, Cincinnati 1974.
21. *Houben-Weyl*, 5th ed., **20.**
22. G. W. Becker, D. Braun et al. (eds.): *Kunststoff-Handbuch*, 2nd ed., Hanser, München, 1983 ff. (many vols.).
23. H. Mark, C. Overberger, G. Menges, N. M. Bikales (eds.): *Encyclopedia of Polymer Science and Engineering*, 2nd ed., **19 vols.**, Wiley, New York 1985 – 1990.
24. W. A. Kargin (ed.): *Enciklopedia Polimerov*, **3 vols.**, Sovietskaya Enciklopedia Publ., Moscow 1972.
25. G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science*, **7 vols.**, Pergamon, Oxford 1989.
26. *Chemiefasern auf dem Weltmarkt*, 7th ed., Deutsche Rhodiaceta , Freiburg/Br., 1969, Supplement 1976 (trade names of fibers).
27. The International Plastics Selector, *Commercial Names and Sources*, Cordura Publ., San Diego 1978.
28. *Parat (Fachinformationszentrum Chemie*, Berlin), Index of Polymer Trade Names, VCH, Weinheim, Germany 1987.
29. J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook*, 3rd ed., Wiley, New York 1989.
30. O. Griffin Lewis: *Physical Constants of Linear Homopolymers*, Springer, Berlin 1968.
31. R. E. Schramm, A. F. Clark, R. P. Reeds (eds.): *A Compilation and Evaluation of Mechanical, Thermal, and Electrical Properties of Selected Polymers*, U.S. National Bureau of Standards , Washington, D.C., 1973.
32. P. A. Schweitzer: *Corrosion Resistance Tables (Metals*, Plastics, Nonmetallics, Rubbers), Dekker, New York 1976.
33. The International Plastics Selector, San Diego 1977.
34. H. Saechtling: *Kunststoff-Taschenbuch*, 24th ed., Hanser, Munich 1989; International Plastics Handbook, 2nd ed., Hanser, München 1989.
35. *Modern Plastics Encyclopedia*, McGraw-Hill, New York (annually in October as issue 10 A).
36. *Encyclope* die Francaise des Matieres Plastiques, Les Publicateurs Techniques Association , Paris, annually.
37. J. E. Williams (ed.): *Computer-Readable Databases*, American Library Association, Chicago 1985.
38. German Plastics Institute, Plastics Databank POLYMAT, Darmstadt; Fachinformationszentrum Chemie, Berlin (6000 polymers from 80 manufacturers).
39. Plastics Databank CAMPUS (= Computer Aided Material Preselection by Uniform Standards), (5000 thermoplastics with 50 most important properties each), issued for their company products by 17 European companies.
40. Polymer Materials Database, Fachinformationszentrum Berlin (continuously updated).
41. E. R. Yescombe: *Plastics and Rubbers: World Sources of Information*, 2nd ed., Appl. Sci. Publ., Barking 1976.
42. G. J. Patterson: *Plastics Book List*, Technomic Publ., Westport 1975.

43. P. Eyerer: *Informationsführer Kunststoffe*, VDI-Verlag, Düsseldorf 1976.
44. J. Schrade: *Kunststoffe (Hochpolymere), Bibliographie aus dem deutschen Sprachgebiet, Erste Folge* 1911 – 1969, I. Schrade, Schweiz. Aluminium AG., Zürich 1976, Zweite Folge, Zürich 1980.
45. O. A. Battista, *The Polymer Index*, McGraw Hill, New York 1976.
46. S. M. Kaback: "Literature of Polymers," *Encycl. Polym. Sci. Technol.* **8** (1968) 273; J. T. Lee, *Encycl. Polym. Sci. Eng.*, 2nd ed., **9** (1987) 62.
47. W. J. Roff, J. R. Scott: *Handbook of Common Polymers*, Butterworths, London 1971.
48. M. Ash, I. Ash: *Encyclopedia of Plastics, Polymers and Resins*, **4 vols.**, Chem. Publ. Co., New York 1982 –1988.
49. M. S. M. Alger: *Polymer Science Dictionary*, Elsevier Applied Sci. Publ., Barking 1989.
50. American Society for Testing Materials, ASTM D-1600–86 a; D-1418–67; D-4020–81.
51. German Industrial Standards DIN 7723; 7728 (1988); 16 913; 55 950; 60 001.
52. European Textile Characterization Law.
53. International Organization for Standardization, ISO 1043 – 1978, ISO 1629 – 1980.
54. Guide for the Harmonisation of Designations of Thermoplastic Materials, ISO/TC 61/SC 9 N 435, April 1981; ASTM D 4000.
55. K. Wiebusch, *Kunststoffe-German Plastics* **72** (1982) 22, 167.
56. F. O. Swanson, *Plastics Engng*. (1983) Jan. 31.
57. R. H. Wehrenberg, II, *Mat. Engng.* (1984) Feb. 48.
58. H.-Elias *et al.*, *Polymer News* **9** (1983) 101; **10** (1985) 169.
59. J. H. DuBois: *Plastics History – U.S.A.*, Cahners, Boston 1971.
60. J. K. Craver, R. W. Tess (eds.): *Applied Polymer Science*, Am. Chem. Soc., Washington, D.C., 1975.
61. F. M. McMillan: *The Chain Straigtheners: Fruitful Innovation. The Discovery of Linear and Stereoregular Polymers*, MacMillan, London 1981.
62. R. B. Seymour (ed.): *History of Polymer Science and Technology*, Dekker, New York 1982 (Reprint of articles in *J. Macromol. Sci.-Chem.* **A 17** (1982) 1065 –1460.
63. R. Friedel: *Pioneer Plastic. The Making and Selling of Celluloid*, University of Wisconsin Press, Madison, 1983.
64. W. Glenz (ed.): *Kunststoffe – Ein Werkstoff macht Karriere*, Hanser, München 1985.
65. United Nations, Statistical Yearbooks.
66. Calculated from data in *Kunststoffe* **73** (1983) no. 10.
67. H.-G. Elias, F. Vohwinkel: *New Commercial Polymers*, Gordon and Breach, New York 1986; Neue polymere Werkstoffe, Hanser, München 1983.
68. H.-G. Elias, *Grosse Moleküle*, Springer, Berlin 1985 *Mega Molecules*, Springer, Berlin 1987. *Megamolekuly*, Khimiya, Leningrad 1990.
69. United States, Statistical Abstracts (annually).
70. D. Tomalia, A. M. Naylor, W. A. Goddard III, *Angew. Chem.* **102** (1990) 119; *Angew. Chem. Intern. Ed.* **29** (1990) 138.
71. G. Natta, P. Corradini, *Nuovo Cimento, Suppl.* **15** (1960) 11.
72. D. Lang, H. Bujard, B. Wolff, D. Russell, *J. Mol. Biol.* **23** (1967) 163. Links
73. R. Gächter, H. Müller (eds.): *Plastics Additives Handbook* 3rd ed., Hanser, München 1989.
74. A. E. Woodward, *Atlas of Polymer Morphology*, Hanser, München 1988.
75. C. Ruscher, E. Schulz, private communication.
76. R. J. Samuels, private communication.
77. P. J. Flory, *Adv. Polym. Sci.* **59** (1984) 1. Links
78. *Kunststoff-Physik im Gespräch*, 8th ed., BASF, Ludwigshafen 1988.
79. H. M. Laun, H. Münstedt, *Rheol. Acta* **17** (1978) 415. Links
80. S. Onogi, T. Matsuda, K. Kitagawa, *Macromolecules* **3** (1970) 109. Links
81. S. Wellinghoff, E. Baer, private communication.
82. G. E. Ehrenstein, R. Wurmb, *Angew. Makromol. Chem.* **60/61** (1977) 157.
83. T. H. Ferrigno, in S. Katz, J. V. Milewski (eds.): *Handbook of Fillers and Reinforcements for Plastics*, Van Nonstrand, New York 1978, p. 11.
84. J. A. Manson, L. H. Sperling: *Polymer Blends and Composites*, Plenum, New York 1976.
85. S. L. Aggarwal, R. L. Livigni, *Polym. Eng. Sci.* **17** (1977) 498.
86. Bundesamt für Umweltschutz, Oekobilanzen von Packstoffen (Schriftenreihe Umweltschutz), no. 24, Bern 1984.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Plastics, Processing

Gert Burkhardt, Institut für Kunststoffverarbeitung, Aachen, Federal Republic of Germany
Ulrich Hüsgen, Institut für Kunststoffverarbeitung, Aachen, Federal Republic of Germany
Matthias Kalwa, Institut für Kunststoffverarbeitung, Aachen, Federal Republic of Germany
Gerhard Pötsch, Institut für Kunststoffverarbeitung, Aachen, Federal Republic of Germany
Claus Schwenzer, Institut für Kunststoffverarbeitung, Aachen, Federal Republic of Germany

1. Processing of Thermoplastics . . . 2
1.1. Introduction . . . 3
1.2. Compounding Processes . . . 3
1.2.1. Mixing . . . 4
1.2.2. Rolling and Kneading . . . 5
1.2.3. Pelletizing . . . 6
1.2.4. Shredding and Grinding . . . 6
1.2.5. Storage and Transportation . . . 7
1.3. Pressureless Processing Techniques . . . 8
1.3.1. Casting . . . 8
1.3.2. Dipping . . . 10
1.3.3. Coating . . . 11
1.3.4. Foaming . . . 13
1.4. Processing under Pressure . . . 14
1.4.1. Compression Molding . . . 14
1.4.2. Rolling and Calendering . . . 14
1.4.3. Extrusion . . . 16
1.4.4. Blow Molding . . . 23
1.4.5. Injection Molding . . . 28
1.5. Forming Processes . . . 36
1.5.1. Drawing . . . 36
1.5.2. Thermoforming . . . 37
2. Processing of Thermosets . . . 39
2.1. Vinyl Ester Resins, and Phenol–Formaldehyde Resins . . . 39
2.1.1. Raw Materials . . . 43
2.1.2. Mold Materials . . . 45
2.1.3. Preparation of Resins . . . 45
2.1.4. Pressureless Processing without Fiber Reinforcement . . . 46
2.1.5. Pressureless Processing with Glass-Fiber Reinforcement . . . 47
2.1.6. Low-Pressure Processes . . . 48
2.1.7. Press Molding . . . 49
2.1.8. Semicontinuous Processing Techniques . . . 53
2.1.9. Continuous Processes . . . 54
2.2. Curable Molding Compounds and Laminates . . . 56
2.2.1. Processing Techniques . . . 56
2.2.1.1. Processing Curable Molding Compounds . . . 56
2.2.1.2. Manufacture of Laminated Plates . 60
2.2.2. Processing Properties . . . 61
2.2.3. Processing Conditions . . . 62
2.2.4. Influence of Processing on Molding Properties . . . 65
2.2.5. Molds . . . 66
2.2.6. Machines . . . 68
3. Further Processing . . . 70
3.1. Machining . . . 70
3.1.1. Estimating Machinability . . . 70
3.1.2. Shaving Formation . . . 70
3.1.3. Surface Quality . . . 71
3.1.4. Cutting Forces . . . 72
3.1.5. Cutting Temperature and Tool Wear 73
3.1.6. Guidelines for Machining Polymers 73
3.1.7. Processing of Fiber-Reinforced Plastics . . . 74
3.2. Bonding . . . 77
3.2.1. Welding . . . 77
3.2.1.1. Heated-Tool Welding . . . 78
3.2.1.2. Hot-Gas Welding . . . 80
3.2.1.3. Radiation Welding . . . 81
3.2.1.4. Friction Welding . . . 81
3.2.1.5. Induction Welding . . . 82
3.2.2. Adhesion . . . 83
3.2.3. Mechanical Bonding . . . 85
3.3. Surface Treatment . . . 85
3.3.1. Mechanical Processes . . . 85
3.3.1.1. Hot Embossing . . . 85
3.3.1.2. Flocking . . . 86
3.3.2. Painting and Printing . . . 87
3.3.2.1. Painting . . . 87
3.3.2.2. Printing . . . 90
3.3.3. Wet Chemical Deposition Processes 90
3.3.3.1. Chemical Deposition . . . 90
3.3.3.2. Electroplating of Plastics . . . 91
3.3.4. Physical Deposition Processes . . . 93
3.3.5. Dry Reactive Processes . . . 95
3.3.5.1. Surface-Modifying Processes . . . 95
3.3.5.2. Coating Processes . . . 96
4. References . . . 97

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a20_663

1. Processing of Thermoplastics

This article considers the processing of thermoplastics to semifinished and finished products, as well as the processing stages that occur prior to the final manufacturing step. The complexity and diversity of the subject limit the depth of the discussion here to the more fundamental aspects.

1.1. Introduction

Thermoplastics consist of macromolecules that are formed by addition polymerization [e.g., poly(vinyl)chloride], polycondensation (e.g., polycarbonate), polyaddition (e.g., polyurethane), or polymer transformation [e.g., transformation of poly(vinyl acetate) to poly(vinyl alcohol)]. The macromolecules can be linear or branched and are entangled. Additionally, they are held together by van der Waals forces. These can be classified according to their origin as dispersive, dipole, or induction forces and hydrogen bonds. Depending on their molecular structure, thermoplastics can be amorphous or semicrystalline. The thermal and rheological behavior is decisive for processing and application characteristics. Below the softening temperature (glass transition temperature), thermoplastics are rigid and brittle. The intramolecular forces decrease with increasing temperature, which results in an increase in the mobility of the molecular chains. Above the glass transition temperature, amorphous polymers exhibit elastic behavior caused by chain entanglements. With semicrystalline materials, dimensional stability is maintained by the unmelted crystalline regions; the simultaneously existing "melted" amorphous regions give high impact strength. With increasing temperature, the chain entanglements loosen even more due to increased thermal motion; with semicrystalline materials this occurs even in crystalline zones at temperatures above the crystalline melting point. The result is the plastic melt required for processing. The transition from solid to melt is reversible; on cooling, the processes described above occur in the reverse order. A characteristic of the melt is that the material still possesses elastic properties due to the residual chain entanglements. Additionally, flowing of the melt leads to alignment of the molecules. The degree of orientation in a preferential direction depends on the temperature and the length of the molecular chains. When the material is cooled, this orientation is to some extent frozen in, a situation that may be desirable. For example, this leads to an increase in the mechanical properties in the direction of orientation. However, internal stresses are also frozen in, which relax when the temperature is increased and can lead to distortion of the finished product. To suppress this phenomenon, during molding, the melt must be held at the maximum temperature possible for a long residence time, to allow relaxation to occur. This is not always possible, because thermal damage can result.

The rheological properties of polymer melts are of considerable importance for processing. The strong dependence of viscosity on both temperature and shear velocity plays a significant role in the selection and design of processing machines and molds, as well as the choice of operating parameters. During processing a rapid input of energy is desirable to give a processible melt with sufficiently low viscosity as quickly as possible. Because of the poor thermal conductivity of polymers, conductive heating is unsuitable since it is too slow, and large temperature gradients could result. Convection also plays a minor role because high-molecular melts exhibit laminar flow at extremely low Reynolds numbers. Hence in most cases, the polymer is melted by internal friction (dissipation), generated by the introduction of mechanical energy and the resulting friction and shearing forces to which the molecules are exposed. Polar polymers can also be heated in a high-frequency electromagnetic field. Radiation heating, generally employing IR radiation, is typically used to warm surfaces and planar articles. However, the most important method is melting by dissipation in screw machines, combined with conductive heating via the contact surfaces.

Removal of heat from the melt is more difficult because the only mechanism available is thermal conduction. This can lead to major problems, especially with thick-walled moldings. Hence, the low thermal conductivity of polymers, and the resulting problems regarding heating and cooling, are decisive in the selection of processing technology and have a significant influence on process economics.

1.2. Compounding Processes

Compounding includes all the processing stages between the manufacture of the crude polymer and the molding step. The transition from compounding to processing can be a gradual change. Thus, storage, transportation, metering, mixing, plastication and granulation are regarded as compounding.

Compounding is necessary because either as-polymerized plastics cannot be processed directly or their properties are inadequate. Consequently, additives must be introduced to the crude polymer to allow the production of articles on an industrial scale.

Heat stabilizing additives enable the polymer to withstand processing temperatures without thermal degradation. Additionally, processing aids are employed which, for example, reduce melting time, influence melt viscosity, or decrease the material's residence time in the processing machine by preventing adhesion of the polymer to metallic machine components.

By using plasticizers, rigid and brittle polymers can be rendered flexible and ductile, opening up new applications.

Similarly, optical properties can also be varied by employing additives. By using pigments, polymer products can be manufactured in all colors imaginable, and gloss can be specifically influenced.

Fillers are increasing in importance not only because of their volume-enlarging effects but also because of their influence on processing and properties.

Another significant aspect of compounding is that many polymers are produced in a form unsuitable for further processing (e.g., as a fine powder). In such cases the material must be reshaped in an intermediate processing step known as pelletizing, which, for example, improves the feeding behavior in processing equipment.

Generally, polymer compounding is carried out by the polymer manufacturer, so the processor can purchase a compound that is ready for use. On an industrial scale, the only thermoplastic compounded by the processor is poly(vinyl chloride).

1.2.1. Mixing

The production of a homogeneous mixture of polymer and additives places high demands on machine and process. The objective of mixing is to distribute the additive throughout the crude polymer as evenly as possible without excessively stressing the polymer itself. The two basic processes are cold mixing at room temperature, in which the components are simply mixed, and hot mixing, during which absorption and diffusion occur.

Examples of cold mixers are the tumble mixer, in which mixing occurs simply under the action of gravitational forces, and the ribbon mixer, which mixes the materials with a spiral in a horizontal barrel.

Hot mixers are classified according to the method of introducing heat. The polymer can be heated by contact with heated metal components, by irradiation, by convection (e.g., in a plowshare mixer), or by dissipation of mechanical work in the product in the form of heat of friction. Because this form of energy introduction requires high mixer revolutions, these mixers are known as high-speed or turbo mixers. This type of mixing allows rapid heating of the materials and is therefore widely used throughout the plastics processing industry. It is described in more detail below, by using poly(vinyl chloride) (PVC) compounding as an example.

PVC is compounded in mixing vessels equipped with ring- or propeller-shaped mixing elements (Fig. 1). The vessel is usually double walled to allow better temperature control. The peripheral velocity of the mixing elements is generally 20 – 50 m/s. These simultaneously mix and heat the components by throwing them against the container walls and causing them to rise. This results in a whirling motion, with the material being circulated between the center of the container and the base of the vessel. The powder is heated by friction between particles and by particle collisions.

During the initial phase of mixing, the components are merely distributed homogeneously. Actual hot mixing begins above 50 °C. On further heating, during which the plastication temperature of PVC (ca. 150 °C) must not be exceeded, the additives are ground together, and meltable additives are melted. The PVC powder, which is not ground in this process, absorb the

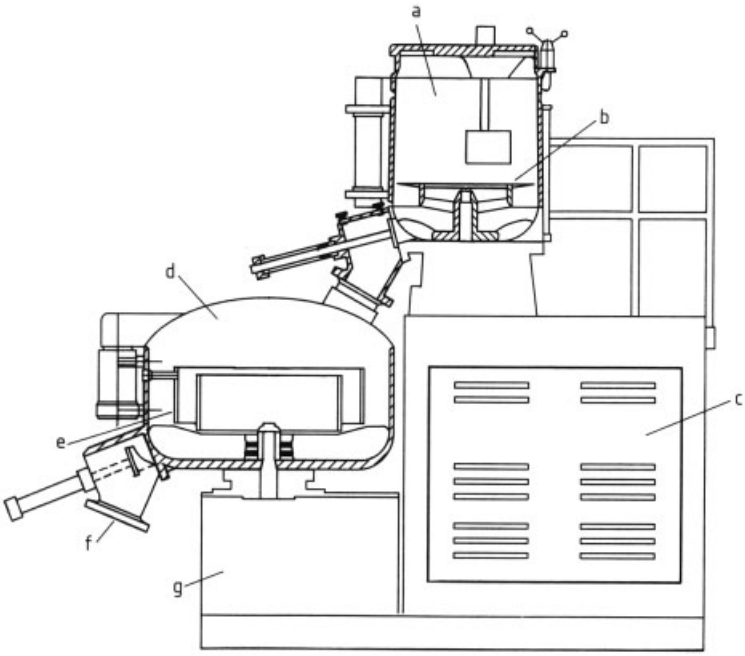

Figure 1. Heating/cooling mixer combination for compounding PVC
a) Heating mixer; b) Mixing element (propeller); c) High-speed drive; d) Cooling mixer; e) Stirrer; f) Outlet; g) Low-speed drive

molten additives so that at the end of hot mixing a dry powder, the so-called dry blend, results. This material has a final temperature between 110 and 130 °C, and must be cooled in a cool mixer before it can be stored. It is allowed to fall into a water-cooled mixer situated beneath the hot mixer, where it is cooled to ca. 40 °C with constant, slow stirring to prevent fusion of the pellets.

The dry blend produced in this way can be processed without further compounding steps, as is typical for rigid PVC by use of twin screw extruders, or it can be pelletized, as is normal for plasticized PVC.

1.2.2. Rolling and Kneading

Thermoplastics are often plastified prior to processing. The material is compacted, melted, degassed (if necessary), homogenized, and modified with additives. The oldest process is *rolling*, which involves passing the plastic between two counterrotating, heated rolls where it is plastified and pressed to a thin layer. The plastic is heated by contact with the rolls and by friction in the buildup of material before the roll gap and during passage through the narrow gap between rolls. The major advantage of this system is that it is an open process which, in addition to facilitating addition of additives, also allows scrap material such as edge trimmings to be reintroduced without any difficulty. Because of the high surface-to-volume ratio, the material temperature can be controlled precisely. This is of considerable significance with thermally sensitive materials such as PVC. Furthermore, cleaning the machines is straightforward, and the process can be monitored visually.

A disadvantage is that rolling is a batch process and difficult to automate. More recently, a roll unit has been developed which can be operated in a continuous mode thanks to spiral-shaped grooves that are worked into the roll surfaces. The polymer is fed to one side of the rolls and, passing through the roll gap several times, is driven by the spiral grooves to the other side where it is taken off.

Internal mixers also process polymers batchwise. The material is kneaded between blades rotating in a double barrel and discharged when plastication is complete.

In continuous kneaders, the kneading blades are in the form of rotors with multiple screw channels on the feed side that transport the polymer to the actual kneading elements. The kneading zone is followed by a pump zone, which ejects the plasticated material.

The planetary roller extruder is based on the principle of a single-screw extruder, where the screw is designed as a spindle around which several planetary spindles rotate (Fig. 2A). This design enables intensive shearing and mixing of the material due to repeated renewal of the material surface. Because of the thin layers of material and the large surfaces of the central and planetary spindles, precise temperature control is possible.

In the *plasticator* a feed screw transports the polymer to a conical rotor in a conical cylinder housing (Fig. 2B). The conical gap can be adjusted by moving the cone relative to the housing. The excellent homogeneity of the plasticated product can be attributed to the fact that the powdered material is initially sintered and subsequently plastified. The material is spread out by the cone and repeatedly separated and relayered by spiral grooves or threads. Excellent temperature control is possible due to the large surface area. The plasticated material finally passes to further processing units via a discharge screw.

Additionally, a range of other compounding equipment is available, including cokneaders and co- or counterrotating twin-screw extruders. All possess various advantages and drawbacks; hence their selection is not necessarily a question of material-specific or process aspects alone. Plant conditions and economics also play major roles.

1.2.3. Pelletizing

Pelletizing is the manufacture of pellets (granules) of equal shape and size, to give optimum feed behavior on the processing equipment. The material to be granulated is melted in a single- or twin-screw extruder and fed to a pelletizer. Both hot and cold pelletizing processes are used. Depending on the manufacturing procedure, bead, cylinder, or cube granulate is obtained.

In cold pelletizing, strands, ribbons, or sheets are formed by a pelletizing die; after solidification, they are chopped to the desired shape by using rotating knives.

In hot pelletizing, the plasticated material is pressed through a die, and the exiting strand is chopped by a blade that rotates on the surface of the die plate. Cooling of the melt occurs after pelletizing, either with air (dry pelletizing) or water (wet pelletizing), although in the latter case, the water must be subsequently removed. With materials of very low viscosity and materials that tend to adhere, the melt passes through the die plate directly into a water bath (underwater pelletizing) where it is cut by blades rotating under the water surface. The resulting pellets must then be dried.

1.2.4. Shredding and Grinding

During the processing of thermoplastics, a variety of off-specification semifinished and finished products are generated, such as pipes, profiles, injection moldings, and hollow articles, along with scrap such as flash, die leakages, cutting die residues, and film and sheet edge trimmings. For reasons of economy and ecology, these materials are fed to reclaim systems (recycling of single-product manufacturing scrap).

These materials must be cut to produce a pellet form suitable for processing (chip granulate). In doing so however, a range of particle sizes is always obtained.

Cutting mills are generally employed for regranulating. Scrap material is added through a feed system: original-length pipes and profiles are fed horizontally; grinders for film edge trimmings can be incorporated directly in suction pipes. The rotor revolves with high momentum and is equipped with multiple cutting heads that operate against one or more stationary cutting edges mounted in the housing.

In some models, to increase efficiency the rotating cutting heads are divided and displaced with respect to each other. The feed material is cut in the cross direction and – with repeated reorientation – further chopped in subsequent cutting operations. When the desired particle size is reached, the material is ejected through a sieve whose pores determine the particle size. Heat generated in the cutting process can be removed by water cooling of the housing and rotor.

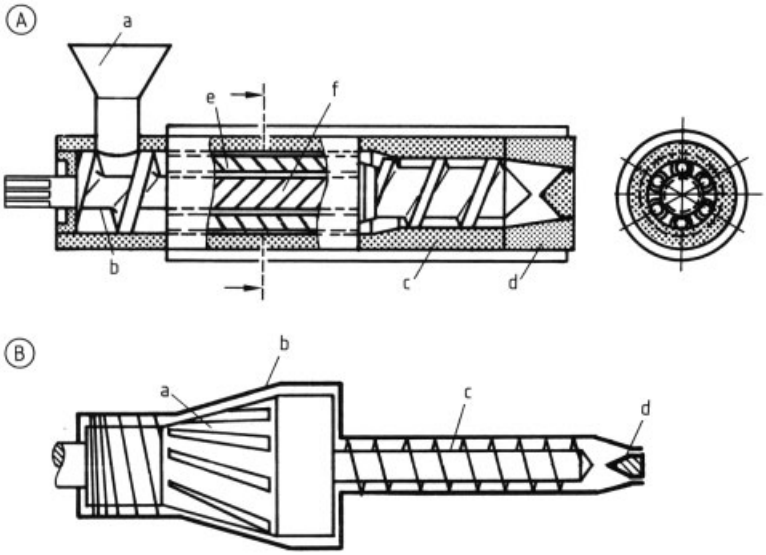

Figure 2. A) Planetary screw extruder: a) Hopper; b) Feeding screw; c) Housing; d) Die; e) Planetary spindle; f) Central spindle
B) Plasticator: a) Rotor; b) Housing; c) Discharge screw; d) Die

For processing larger plastic articles, a two-stage grinding process has proved effective. The first stage involves coarse chopping with a guillotine (for film rolls and start-up chunks, i.e., material formed before melt calibration commences) or by a preliminary cutting rotor (for bulky materials). The second stage normally involves the cutting mills described above.

To process thermoplastics in pressureless processes (see Section 1.3), they must generally be available as a flowable, finely divided powder. In many cases, they are not produced in this form during polymer syntheses, or compounding with additives in a mixing and granulation process may be necessary. Sinter powders are therefore often produced by grinding granular materials.

In the impact disk mill, the product passes from the feed hopper over a vibration-metering rim. A rotating wheel catapults the material with high velocity against a ribbed conical disk rotating in the opposite direction. The adjustable gap between the rotating rear disk and the conical front cover determines the fineness of the ground material. Powerful suction equipment then removes the powder from below. Heating of the plastic is prevented by the high air throughput and by water cooling the housing. The discharged powder is classified in a cyclone and then sieved. Coarse material is returned to the grinding mill.

1.2.5. Storage and Transportation

Automatic equipment has become well established in the polymer processing industry for storing and transporting base polymers and intermediate products. Compared to manual operations, these provide improved working conditions, reduction in material losses, lower workplace and machine contamination levels, and a reduction of accident risks.

For the storage of polymers, generally in powder or pellet form, large external silos are suitable. These can be freighted by road or rail. The outlet cone of the silo is generally fitted with a discharge mechanism. The formation of granule bridges is prevented by ventilation of the silo base or by using vibrating frames, bunker cushions, slotted shelves, or similar techniques.

Reserve materials or smaller charges and additives are shipped in bulk containers, drums, or sacks. These are emptied by special discharge equipment, tipper mechanisms, or sack shakers. Within a plant, small intermediate silos, (homogenizing) vessels, and containers are used to store raw materials and prepare polymer mixtures. Containers are preferred when frequent

material changes are involved because they are portable and can be stacked. When fitted with a sliding floor, they can also be used as feeding vessels for the hoppers on processing equipment.

Powders and pellets are generally conveyed with pneumatic equipment. High-pressure conveyors (operating pressure 0.2 – 0.4 MPa) are used for high outputs when filling storage silos from pressure-resistant rail or truck tanks. Medium-pressure conveyors operating at ca. 0.15 MPa are generally used for transport from the storage silos to day or intermediate containers. Low-pressure conveyors with radial blowers are used for short transport distances. Suction conveyors, operating at ca. 50 KPa are most suitable for feeding machine hoppers and gravimetric equipment. The transport interval can be controlled by using fill-level sensors, and air filters are automatically cleaned by back flushing. Mechanical conveyors include screw conveyors (also available as flexible versions) or oscillation conveyors in the form of tubes, channels, and spirals with electromagnetic vibrators or motor drives.

Oscillating displacement pumps, particularly piston, membrane, and bellows pumps, are used for metering liquids such as plasticizers and other additives for compounding. Solids are volumetrically metered by using single- or twin-screw equipment, vibration channels, rotating plates, or cellular wheel sluices. Gravimetric metering equipment can be used for all types of solids, regardless of particle size, homogeneity, flow properties, or adhesion. Weighing vessels, as used in batch processes, are also applicable for continuous metering by discharging with a screw mechanism whose rate of revolution is adjusted to match the weight of material required in a given time. With metering belts, either the belt speed or the rate of revolution of the feeding mechanism is adjusted.

The material flow in a processing plant, including compounding equipment, is shown schematically in Figure 3.

Some thermoplastics, such as polycarbonate (PC), polyamide (PA), and poly(methyl methacrylate) (PMMA), must be dried prior to processing; otherwise the mechanical or optical properties of the resulting products would be considerably impaired. Drying is carried out by heating for several hours in an oven at 80 to 120 °C, depending upon the material concerned.

1.3. Pressureless Processing Techniques

Pressureless proceses are used for starting materials that flow and can be formed under the action of gravity without application of external forces. These include monomers, polymer solutions and dispersions, heat-gelable PVC – plasticizer mixture (pastes), low-viscosity melts, and finely divided, meltable powders. The solid end product is formed by reaction, evaporation of the solvent or dispersant, or cooling.

1.3.1. Casting

Monomer Casting. Thick-walled articles, embeddings, sheets, blocks, and rods, preferentially of polystyrene, polyamides, and poly(methyl methacrylate), can be produced by polymerization of monomers in molds. Control of the heat of reaction necessitates prolonged reaction times. Adequate temperature regulation is required to avoid thermal stresses. The choice of mold material depends on the extent of use and the desired surface quality of the molding. Apart from metal molds with cooling channels, it is often possible to use simpler molds made of sheet metal, wood, ceramic, or glass. To guarantee ease of release and ensure smooth molding surfaces, even with porous materials, those components of the mold that come in contact with the monomers must be sealed by coating or impregnating.

Film Casting. Filtered polymer solutions, dispersions, or melts are cast onto a substrate (e.g., paper, textiles), directly onto a metal drum or rotating belt, or in the case of cellulose films, from a slot die into a precipitation bath.

The solvents are recovered by evaporation in heating and separation stages. The only major application of this process today is manufacture of cellulose acetate films, mainly for the photographic industry.

Slush Molding. Hollow articles made from PVC pastes are manufactured in a pressureless process using two-part metal molds. The molds are heated to the gelation temperature and filled

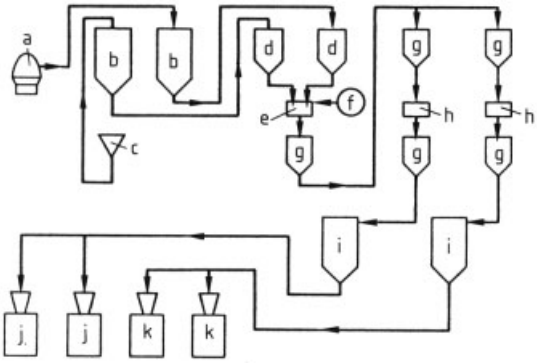

Figure 3. Schematic of a processing plant
a) Rail or truck container; b) Storage silo; c) Sack chute; d) Intermediate silo; e) Weighing; f) Additives; g) Interim storage; h) Mixer; i) Silo for one day's supply; j) Granulator; k) Processing machine

with paste. The paste gels at the outermost edges to form a skin, whose thickness depends on the temperature and the composition of the paste. After excess paste is poured off, the layer remaining in the mold is completely gelled in an oven. After cooling, the hollow molding is removed from the mold.

Rotational Casting. A further development of slush molding is the manufacture of hollow articles from plasticized PVC by rotational casting of PVC pastes or powders. The method is also applicable to other thermoplastics or liquid monomers containing activators and polymerization catalysts.

In this case, only the amount of paste required for the manufacture of the article is metered to the mold, so that removal of excess paste and postgelation are not necessary. During the heating stage, the mold is rotated about two perpendicular axes (Fig. 4), causing uniform wetting of the cavity surface. To achieve optimum distribution of mold contents, the two axes are rotated at (usually adjustable) different speeds (max. 30 rpm). The individual process steps (filling and closing the mold, rotating in an oven, cooling, and removal of the molding) are performed at separate stations. Either the molding is transported to the appropriate station on rails, or the heating oven or cooling air system can be brought to the rotational casting machine.

Numerous modifications of the process that permit shorter cycle times are known. In modern equipment, an oven is no longer used; instead the double-walled aluminum molds are heated with a circulating heat-transfer fluid. The direct-contact heating allows higher production rates. The gelation temperatures are 180–220 °C. Filling levels and cycle times are determined in pretests.

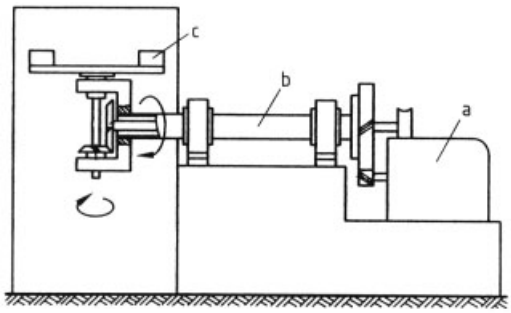

Figure 4. Rotational casting machine
a) Motor; b) Drive shaft; c) Molds

Advantages of this process include the low investment costs and the ability to manufacture moldings of various dimensions without pinch or weld seams, virtually free of stresses and with constant wall thicknesses.

Centrifugal casting is used in the manufacture of thick-walled, rotationally symmetrical moldings (e.g., pipes, sockets). In a variation of monomer casting – in this case, mainly limited to polyamide – polymerization occurs in the partially filled mold during rapid rotation about the central axis. The material in the mold is pressed outward and compacted by centrifugal forces. High-quality internal coatings (e.g., for pipes) can be applied directly by centrifugal

casting, preferably using meltable powders (see Section 1.3.4).

1.3.2. Dipping

A mold or article may be dipped into solutions, dispersions, PVC pastes, melts, or meltable powders; after its removal, the material adhering to the surface is brought to its end-use state by heating or solidification.

Paste Dipping. For the manufacture of articles in which one side remains open (gloves, boots, etc.), a positive model is immersed and the coating is gelled. The process was adapted from the rubber industry; procedures and equipment are practically the same. The positive models are made of metal, glass, or porcelain. For mass production, series of identical molds mounted on dipping frames are used (Fig. 5). Degassed PVC paste is placed in a vessel that is raised and lowered during the dipping process. In some cases, the vessel is stationary and the dipping frame is moved. The immersion and removal stages must be carried out relatively slowly to prevent the introduction of air into the paste and to allow it to flow homogeneously from the mold. The vessel or dipping frame must be moved smoothly without vibration to avoid imperfections or marks on the molded article. This is normally accomplished by using hydraulic drives.

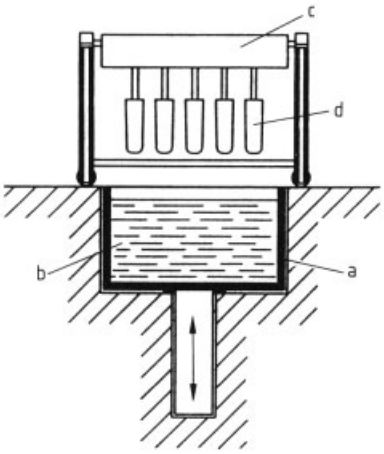

Figure 5. Dipping unit
a) Vessel; b) PVC paste; c) Dipping frame; d) Molds

During gelation by warming, the pastes temporarily become fluid. To prevent loss of material due to increased flow, it is necessary to heat by using a thermal shock, with rotation, or in an oil bath. Gelation is normally completed in a circulating air oven. The molds carrying the gelled coating are subsequently cooled to ca. 50 °C by immersion in water. At this temperature, even harder moldings can be removed easily from the molds, which are generally treated with talc before use.

Other items (usually made of metal, e.g., grids or handles for tools) can similarly be coated by using a dipping process. Cables and wires can be continually coated in this way. Thicker coatings can be obtained by repeatedly coating a form or object after gelation of the previous layer.

Powder Immersion (Fluidized-Bed Sintering). The dipping of articles directly into polymer melts has not gained major significance. In contrast, immersion of preheated and degreased (metallic) objects in plastic powders is widely used in the manufacture of plastic coatings. In the simplest case, the heated articles are immersed directly in the powder; a continuous coating is formed as a result of melting on the surface.

A more elegant and economical process, which is also applicable for large, irregularly shaped articles, is powder immersion using the fluidized-bed sinter technique. In this case, an immersion bath with fluidized powder is employed (Fig. 6). The upper floor of the immersion vessel has many small holes; a porous ceramic plate is usually employed.

The layer of powder at the base of the vessel is agitated by introducing a gas stream, usually an inert gas (e.g., nitrogen) to avoid oxidation reactions, filling the vessel with a fluidized powder. The preheated articles are immersed in the open bath. Thicker coatings can be obtained by repeating the heating and immersion stages. The powder must have good flow properties and give melts of low viscosity. To ensure pore-free coatings, the articles must usually be subjected to posttreatment, either in an oven or with a naked flame. This process is employed in the production of anticorrosion polyolefin or polyamide coatings. Fluidized-bed sintering is carried out both manually on individual articles or small

production runs and on a fully automated, mass production scale.

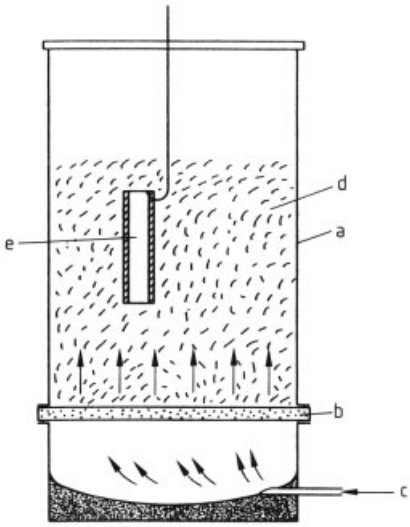

Figure 6. Fluidized-bed sintering unit
a) Immersion bath; b) Filter plate; c) Air inlet; d) Fluidized powder; e) Heated article

1.3.3. Coating

The coating process is used to coat sheet materials such as paper and textiles and, in combination with release papers, in the production of self-supporting films. All products that can be cast maybe used for coating, such as polymer solutions, dispersions, and melts. PVC pastes, which are used in the production of coatings, floor tiles, and imitation leathers, are the most important.

Equipment. A coating machine consists of an unwinding station, an expander, coating head, gelling channel, cooling system, and a rewind station (Fig. 7). The unwind station is slowed by using disk brakes, whereas the rewind station is driven so that the product is held under tension but stretched as little as possible when running through the system. The expanders provide a smooth crease-free sheet. The PVC paste is applied to the moving sheet manually or mechanically by the coating head and then spread to the required thickness by a doctor blade.

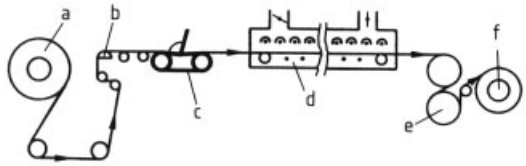

Figure 7. Coating unit
a) Unwinding station; b) Expander; c) Coating head; d) Gelling tunnel; e) Cooling roll; f) Rewind station

The doctor blades are housed in a revolving holder so that they can easily be changed and adjusted in height and inclination. For thick coatings, doctor blades with a wide base (5 – 20 mm) are employed. These are designed with a sharp back edge and an undercut; any rising paste gathers here and can be removed periodically so that it does not drip onto the coating (Fig. 8). For thinner coatings, doctor blades with rounded edges 1 – 2 mm wide are used so that the paste is not forced too deeply into the fabric.

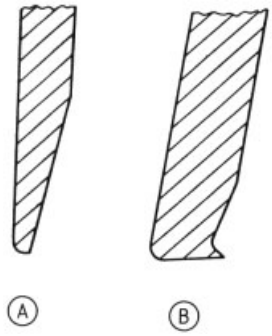

Figure 8. Doctor blades
A) Narrow base with rounded edges; B) Wide base with undercut

Various doctor blade systems are employed, depending on the composition of the paste, the coating material, and the intended use of the product (Fig. 9). With an air knife, the web runs freely between two rollers or metal bars; air knives are generally used with low-viscosity pastes. For many fabrics the knife-on-blanket doctor, in which the fabric lies on an endless rubber belt driven by two rollers, is suitable. This system requires intensive maintenance because of the ease of contamination and the high likelihood of swelling of the rubber belt caused by the plasticizers in the coating. With knife roll coaters, the sheet is supported by a revolving roller, which results in an exactly defined con-

tact area. With steel rolls, complications caused by loose and irregular fabrics can occur, particularly with thin coatings (punctures, disturbances in the coating process); this can be prevented by coating the rollers with soft rubber. The reverse roll coater is employed for coated articles for technical applications in which a defined coating thickness and high operating speeds are important. Two rolls are positioned above a single lower roll (typically steel) over which the web is transported. These two rolls distribute the escaping paste and form it into a film. The separation between the rolls is adjustable and the rotation speeds can be regulated.

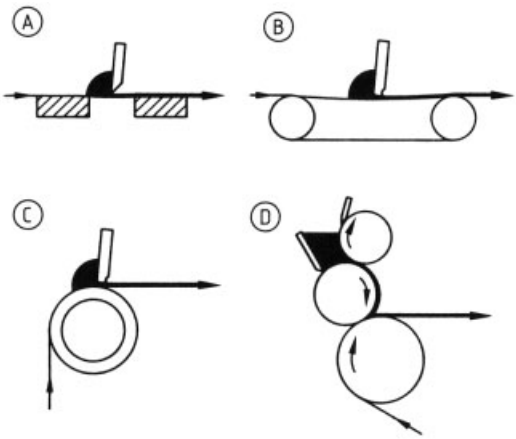

Figure 9. Doctor blade systems
A) Air knife; B) Knife-on-blanket doctor; C) Knife roll coater; D) Reverse roll coater

The length of the gelling tunnel depends on the required operating speed; it can be as long as 20 m. The heat required for gelation of the paste is supplied by either hot air or infrared dryers. With circulating air heat, the air is warmed by steam or oil heating or in heat exchangers using oil burners. Alternatively, the tunnel can be heated by the combustion gases of an oil heater. In modern equipment, air is directed at an oblique angle toward the web through nozzles arranged in the direction counter to the machine direction. With infrared heating the elements are combined in groups that can be slid into position and individually activated. In this way, heating – one of the more expensive aspects in a large-scale operation – can be adapted to suit the process in question. In some cases, circulating air heating is combined with infrared heaters.

The cooling system generally consists of two water-cooled, corrosion-resistant hollow drums. The cooling effect must be high enough that no adhesion between individual layers occurs during winding, nor is any subsequent material deformation (impression from the substrate web) possible. Continuous winding is performed with multistation winders.

Process. In a coating process, depending on the article being manufactured, various thicknesses can be applied in a single- or multiple-step system. Loosely woven fabrics are first smoothed on a textile calender (steel against paper rolls) at 80 – 120 °C. The coating weight of the paste depends on its composition, the water content of the filler system, and the heating system used in the gelling tunnel. With coating weights $>500\,g/m^2$, the danger of blister formation during gelation exists; hence, the desired coating thickness must be applied in several layers.

The primer layer must have excellent adhesion to the fabric; in general, compositions with high plasticizer content and low filler content are used. After gelation of the primer at ca. 130 °C, the filler or intermediate coating can be applied, in one or more stages. Pastes with high filter content or expanded pastes are often used. Each coating stage is followed by a preliminary gelation stage; if insufficient heating capacity is available in the gelling tunnel, thicker coatings must be exposed to an intermediate full gelation stage. The topcoat has the function of providing the product with an appropriately dry and durable surface. For particularly dry surfaces (e.g., necessary for upholstery) a finishing methacrylic coating is applied to seal the surface. This is usually carried out in a roll coating system with an immersion roll (Fig. 10). A mat-steel roll is immersed in a bath, and excess coating is removed by using a doctor blade. The sheet takes up the coating from the steel roll under pressure exerted by the rubber roll mounted above.

When required, a structured surface (e.g., grained) can also be generated in an embossing stage. The product is transported from an unwinding station and passed over a teflon-coated heating drum. After additional heating with in-

frared radiation, the sheet is embossed at 150–160 °C by a patterned rubber roll. The rubber roll is cooled constantly by spraying with water, which is subsequently removed by a nip roll. Before rewinding, the web is cooled by passing it over a series of cooling drums. Because the coating process normally involves multiple layers and generally higher working speeds (30 m/min), the surface is usually embossed in a separate process (embossing speed 1–5 m/min), particularly when frequent design changes are commonplace.

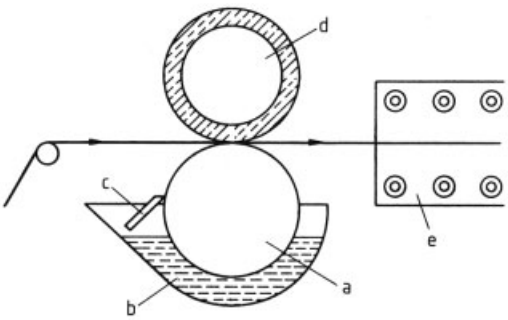

Figure 10. Immersion roll coater
a) Steel roll; b) Trough; c) Doctor blade; d) Rubber roll; e) Drying oven

Foamed products can be produced by adding blowing agents or by employing mechanically foamed pastes (whipped foam; air is emulsified by rapid mixing). To attain specific foaming characteristics, precise temperature control and residence time in the gelling tunnel are required. For the manufacture of unsupported foamed sheets, or when light fabrics or nonwoven materials are being coated, indirect coating methods (transfer or countercoat techniques) are used. Either the coatings are applied in reverse order (topcoat first) onto endless metal or sieve belts, or carrier fabrics or release papers are employed as transport substrates; after being passed over the cooling drums, they are removed and wound individually.

1.3.4. Foaming (see also → Foamed Plastics)

In principle, all polymers can be foamed. To date however, only a limited number of materials have been employed in the manufacture of foams. These include the thermoplastics polystyrene, poly(vinyl chloride), and low-density polyethylene (LDPE), as well as phenolic and polyurethane resins.

Foams are generated by adding physical (gases and low-boiling compounds) or chemical blowing agents (gas-releasing compounds) to the polymer.

A basic requirement for foaming plastics is that the foaming process begin while the polymer is in a flowable form that allows bubbles to develop. Once the bubbles have reached the optimum size for the desired material properties, this condition must be fixed.

The flowability of the polymer to be foamed can be achieved in several ways. With thermoplastics, the polymer can be melted or used as a solution. With thermosets, flowability is obtained by using an incompletely cross-linked polymerization mixture or prepolymer.

The required bubbles can also be generated in various ways. With chemical blowing agents, the gas is evolved either as a byproduct of the polymerization reaction (polycondensation), as a product of a reaction between the polymer and the blowing agent, or as a result of chemical decomposition of the blowing agent.

Physical blowing agents include the mechanical whipping in of air or other gases and the evaporation of low-boiling liquids.

Polyurethane Foam (see also → Polyurethanes). Polyurethanes manufactured by reaction injection molding (RIM) techniques are processed exclusively by using the so-called one-shot process, which involves direct reaction of the starting materials to form the polymer; (i.e., the reaction begins immediately after mixing of individual components). The starting materials permit wide variations in density and hardness to be achieved. By modifying the chemical composition, soft, flexible, or hard articles can be produced. Accordingly, the range of applications for products made from polyurethane is broad and includes, for example,

1) Soft foams for matresses and upholstery
2) Hard foams for thermal insulation (construction industry, refrigerators)
3) Integral foams for automobile body components, where a variation in density over the cross section is required

4) Microcellular (compact) RIM components (e.g., computer housings)

Figure 11 illustrates various applications as a function of material density and stiffness. The incorporation of fabrics, long and short fibers, and fillers opens up additional areas of application.

Thermoplastic Foams. To foam a thermoplastic melt, a component must be added that is gaseous at predetermined temperature and pressure. Nitrogen is added directly to the melt by special gassing units; low-boiling hydrocarbon compounds can be added to the melt or during polymer synthesis. Chemical propellants are usually powdered substances that decompose at a certain temperature, generally releasing nitrogen.

The advantage of chemical blowing agents over direct gassing is that structural foam components can be manufactured on conventional injection molding equipment, and also that the raw material can be prepared simply by mixing the blowing agent with the granulate.

To guarantee optimum foaming of the melt, the gas formation reaction must go to completion within the prescrew zone. The reaction must not begin in the area around the feed opening of the plastication unit, however, because this would allow part of the gas to escape. To set the barrel temperature and obtain conditions that meet these requirements, the decomposition temperature must be known.

In conventional thermoplastic foam casting processes, the mold cavity is only partially filled, and complete filling is obtained as a result of the foaming reaction.

1.4. Processing under Pressure

This section considers processes in which solid thermoplastic materials are melted, molded to shape, and subsequently cooled. Because of the extremely high melt viscosities that can be involved, the molding process may require high pressure; for example, several tens of megapascals for extrusion and >100 MPa for injection molding.

1.4.1. Compression Molding

Compression molding of polymers to semifinished or finished products dates back to the onset of polymer processing but has meanwhile been almost totally displaced by injection molding thermoplastics. Compression molding is now used only in the manufacture of thick-walled components and components made of materials that can not be plastified in extruders [e.g., ultra-high molecular mass polyethylene (PE)]. However, compression molding still plays a significant role in the processing of elastomers and thermosets.

In its simplest form, a press consists of fixed and moving platens. The plunger and cavity components of the compression mold are mounted on these plates, with the parting plane horizontal. Closing and exertion of pressure are usually carried out hydraulically. In compression molding, the cold plastic is placed in the cavity and melted by heating the mold under slight pressure (compression). After melting, the mold is closed and pressure is applied. The molding cools via the mold surface, during which time the pressure is increased further to prevent void formation. Once the ejection temperature has been reached the press is released and the molding removed. The drawback of this process is the long cycle time, which is due to the material being heated and cooled by conduction via the mold surface. For this reason, compression molding is highly uneconomical.

One method of improving the economics is to employ multiplaten presses. In this case, multiple (up to 20) plane-parallel heatable and coolable press tables are loaded simultaneously with plastic, lying on press platens, by using special handling equipment and then pressed simultaneously.

In transfer molding, the starting materials are melted in a cylinder prior to injection into the mold cavity. The mold need only be cooled in this technique. Transfer molding is thus an intermediate between injection molding and compression molding.

1.4.2. Rolling and Calendering

Rolling mills consisting of only two rolls are employed exclusively in the compounding of thermoplastics. These so-called roll mills are used

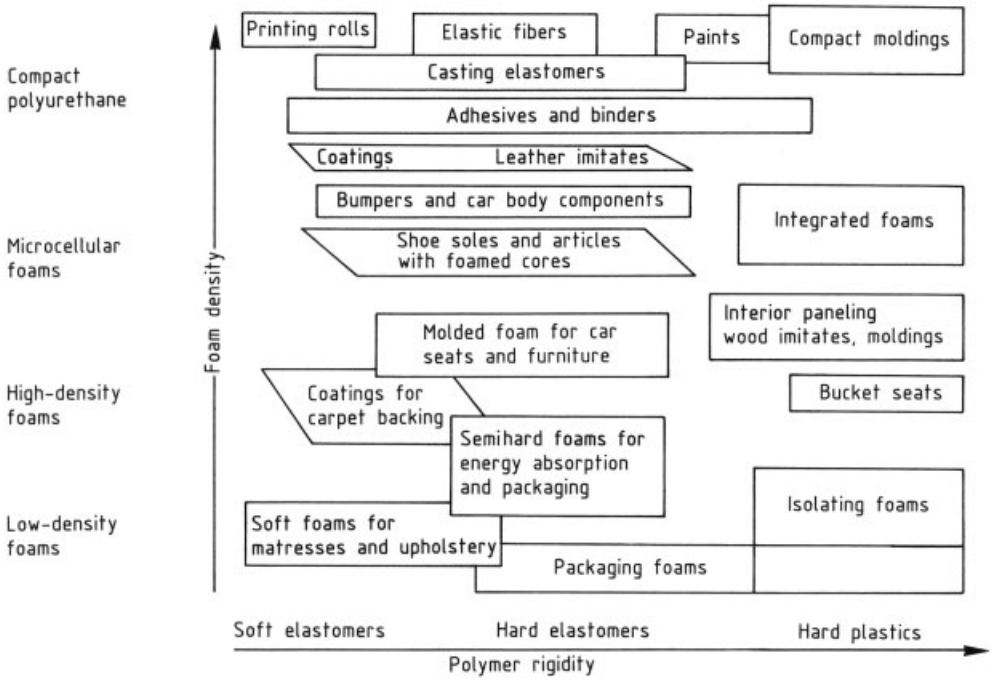

Figure 11. Overview of polyurethane applications

for additive compounding, plastication, and homogenization (see Section 1.2.2). The resulting product is a sheet that can be either granulated after cooling or fed in plasticated form directly to processing equipment. The main application is processing of PVC and rubber, typically in a batch process.

If a polymer is to be shaped by using rolls, then more than two rolls are necessary. Such machines are known as calenders and can be equipped with up to seven rolls, but more commonly come with four. Depending on the arrangement of the rolls, they are referred to as I, F, L, or Z calenders (Fig. 12). The rolls are generally made of highly polished cast iron and are individually driven and heated. Oil or water can be used as heating medium. A calender is generally operated with increasing temperature and speed from one roll to the next. The roll nips can be individually adjusted. Because of the high pressure in the nips, as well as the weight of the roll itself, large rolls tend to sag and hence give an irregular thickness distribution across the width of the final product. This is counteracted by using convex-ground rolls, by roll bending using additional bearings, or by offsetting the roll axes with respect to each other. The roll mill itself must be extremely rigid and stable to withstand the forces to which it is exposed. For this reason, calenders are among the most expensive types of polymer processing equipment. Since they can be economically operated only under conditions of extremely high throughput, they are used for a few special products such as floor coverings and films made of plasticized PVC. Typical calenders have roll lengths (i.e., machine widths) of ca. 1.5 m, but calenders with widths exceeding 3 m are also used. Sheet thicknesses that can be manufactured range from 0.1 to 1 mm. Thinner gauges can be obtained by drawing down using faster speeds for the takeoff rolls. Takeoff speeds are typically 40 – 100 m/min, with throughputs up to several tonnes per hour.

Figure 13 illustrates a calendering line with a four-roll F calender. The roll arrangement shown is employed for processing plasticized PVC, since in this configuration plasticizer vapor that escapes can be removed easily by suction and thus prevented from condensing on the other rolls. The first roll nip is fed continuously from an extruder and conveyor belt. In the first nip a kneading stock is built up from material

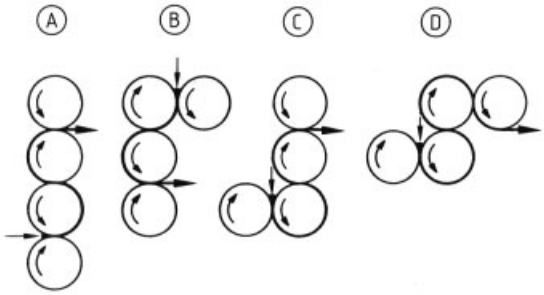

Figure 12. Configuration of calender rolls in four-roll calenders
A) I type; B) F type; C) L type; D) Z type

that is forced back by pressure in the nip. Figure 14 shows a cross section through this stock, illustrating the various currents that are formed. Whereas the lower part of the sheet can pass through the nip essentially unaffected, the remaining material is subject to the currents, which lead to continuous creation of new surfaces and increasing homogenization of the melt. Thus the formation of this kneading stock is of considerable significance for product quality. After preforming in the first nip, final shaping takes place at the next two nips, and the surface quality is established.

A process related to calendering is the roll melting process. In this case, cold polymer is fed to the initial roll nip from a hopper, or is preplasticated in an additional pair of rolls or an extruder. In this case the first roll nip must complete melting of the polymer. The material is then rolled and drawn to the desired thickness in subsequent roll nips. In the manufacture of coatings and laminates, preheated substrate is fed to the last roll nip so that the bond is formed under pressure. After this, both materials are removed simultaneously. Surface effects can be included by embossing with additional rolls.

1.4.3. Extrusion

Extrusion is the term used to describe the continuous manufacture of a semifinished plastic product (film, sheet, pipe, profile). An extrusion line consists of the following:

1) Extruder
2) Mold or die
3) Calibration and cooling
4) Takeoff
5) Finishing

Extruders. The extruder is the heart of every extrusion line; its task is to feed a homogeneous melt of predetermined temperature and pressure into the mold or die. Figure 15 provides an overview of the various types of extruder. An extruder normally consists of:

1) Machine frame
2) Drive (motor, gears, back pressure bearing)
3) Plasticating unit [screw(s), barrel, heating system]
4) Control panel (regulation, control mechanisms, electrical supplies)

Three types of plasticating extruder are most widely used:

1) Single-screw extruders with smooth feed bushes and three-zone screws
2) Single-screw extruders with grooved feed bushes and screws with mixing and sometimes shear zones
3) Twin-screw extruders with closely intermeshing, co- or counterrotating screws

Single-Screw Extruders. A distinction is drawn between melt and plasticating extruders. The basic concepts of both types are essentially identical. The different feed materials – melt extruders are fed with molten plastics, and plasticating extruders with free-flowing solid particles – require specific screw geometries.

The functions of a melt extruder are limited to melt transportation and homogenization. For this reason, additional details of this type of extruder are not considered here. All further discussions refer to plasticating extruders. An extruder is normally classified by its screw diameter and effective screw length, which is given

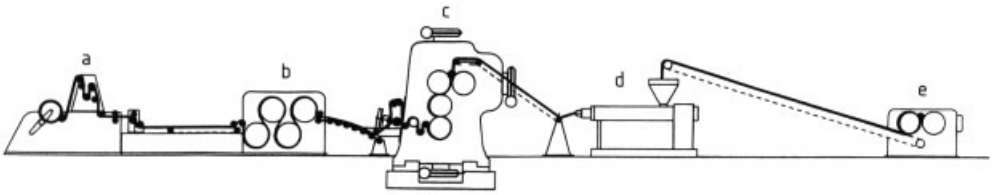

Figure 13. Calendering line
a) Winder and edge cutter; b) Cooling rolls; c) Four-roll calender, F type; d) Extruder; e) Mixing roll mill

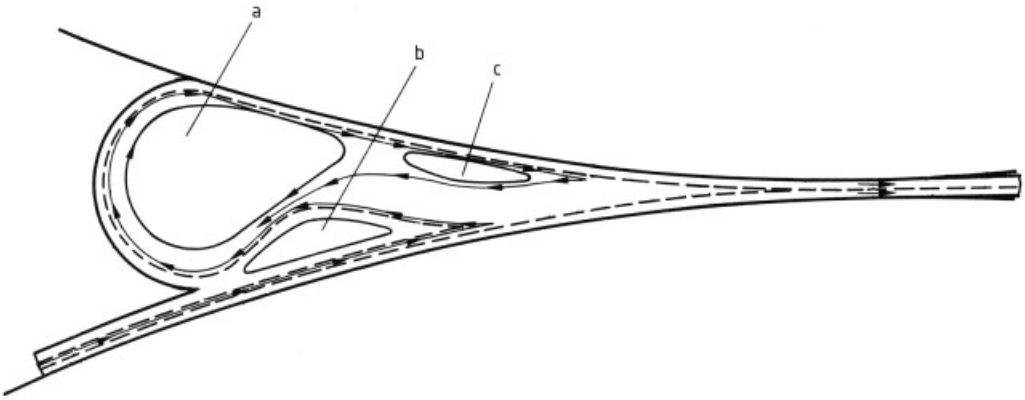

Figure 14. Flow relations in the roll gap
a) Kneading current; b) Feeding current; c) Outlet current

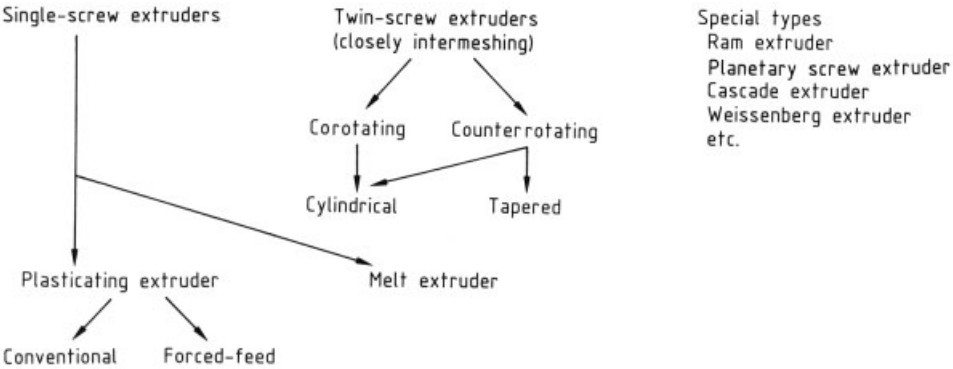

Figure 15. Classification of extruders

as a multiple of the screw diameter D (e.g., 50 mm/25 D).

The functions of a plasticating extruder are summarized in Figure 16. Transportation of the material in the extruder results from adhesion of the melt to the barrel and the screw, which move relative to one another. This mechanism is known as drag transportation. The resulting drag flow is superimposed by the pressure flow, whose magnitude and direction depend on the pressure drop along the length of the screw.

A third flow component is leakage flow, which is the flow that exists between the screw flight and the barrel wall. Because of the narrow gap width, the material is exposed to particularly high shear forces in this region.

Extruder Screws. The screw is the most important element of an extruder. Because of the multitude of demands (feeding, transportation, plastication, homogenization), the geometry varies along the screw length. Figure 17 shows typical screw geometries. The three-zone

Function		Characteristics of the process technology	Engineering solution
Transport	Solid	Low bulk density Low drag forces	High flight depth Grooved feed zone
	Melt	Higher melt density, usually barrel wall adhering	Lower flight depth
Plastication		Solid-melt mixture with low thermal conductivity and increasing density	Decreasing flight depth Shear elements
Homogenization		Inhomogeneous material and temperature distribution	Shallow-flighted mixing zone Mixing elements

Figure 16. The functions of a plasticating extruder

screw is the most widely used; it is made up of the following sections:

1) The feed zone, which draws the solid plastic from the hopper into the extruder and transports it
2) The compression zone, which compresses and plasticates material from the feed zone
3) The metering zone, which homogenizes the melt and heats it to the required temperature

Most thermoplastics can be processed with three-zone screws, although some polymers require additional shear and mixing zones.

Nevertheless, other screws have become established for special applications. The short-compression-zone is used for semicrystalline thermoplastics. The short-compression-zone results from the decrease in channel depth from the deep-flighted feed zone to the shallow-flighted metering zone.

The long-compression-zone screw, which is used for PVC, has a constantly increasing root diameter toward the tip (see Fig. 17C). Extruders or screws designed for specific applications can be regarded as technically suitable when they fulfill the following conditions:

1) Constant, pulsation-free transport
2) Thermally and materially homogeneous melt
3) Ability to remain within the thermal, mechanical, and chemical material degradation limits

These demands must be met in all cases; additionally, for economic reasons, they must be combined with high mass throughputs and low specific operating costs.

Modern extruders have screw diameters D ranging from 19 mm (laboratory extruder) to special designs of 300 mm. The length has increased from the 6 D of melt extruders for rubber processing to ca. 25 D for typical industrial plastication extruders. Plastication extruders with lengths up to 30 D are also available, and degassing extruders have lengths >35 D.

In addition to diameter and length, screws are classified according to the distribution of the functional zones and the compression ratio. The latter is defined as the ratio of the channel depth of the metering zone h_2 to that of the feed zone h_1. To ensure complete filling of the screw, the compression ratio must at least equal the ratio of the bulk density of the solid in the feed zone to the density of the melt in the metering zone.

The screws described here are employed in two different extruder concepts, which differ in the shape of the extruder barrel in the feed zone, the so-called feed bushes.

Conventional single-screw extruders have a smooth barrel surface in the feed zone; the metering zone is regarded as decisive for output. Molten material from the previous zones is homogenized in the metering zone and subsequently forced through the attached extrusion die under pressure.

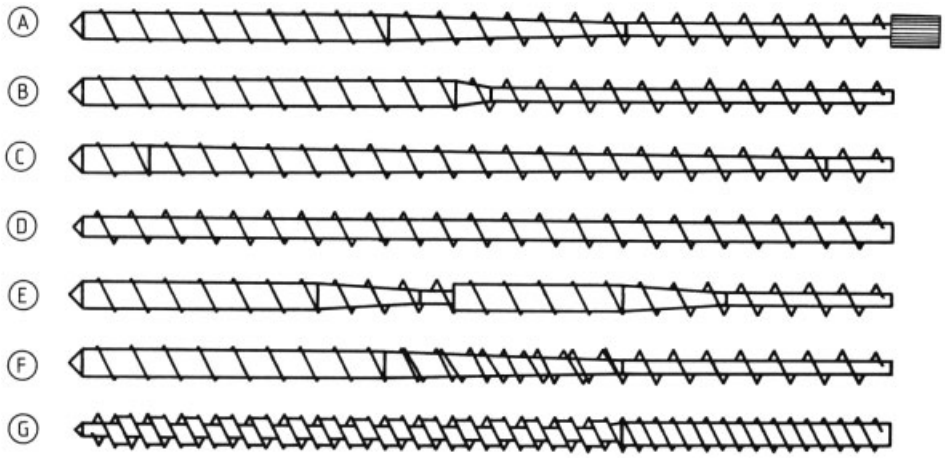

Figure 17. Plasticating screws in common use
A) Three-zone screw; B) Short-compression-zone screw; C) Long-compression-zone screw; D) Conveying screw (only in combination with grooved bushes); E) Degassing screw; F) Maillefer screw; G) Double-flighted wave screw (rarely used)

In this design, the metering zone acts as a pressure generator. The pressure along the length of the screw depends not only on screw geometry, but also on the die back pressure.

Transport of the solid plastic is governed by coulombic friction between the barrel walls and the granules and between the granules themselves.

The result is that a strong correlation exists between the throughput behavior of the extruder and other operational parameters (temperature, back pressure, etc.). In particular, the temperature of the feed zone plays an important role. For this reason, the feed zone is generally cooled.

Fixed Transport Single-Screw Extruder. Feed bushes with axial grooves were developed to attain large, back-pressure-independent throughputs. The so-called forced-feed bush is part of the cylinder which forms the feeding zone; it is usually removable. The interruptions in the otherwise smooth barrel wall caused by the regularly spaced axial grooves prevent the cold feed polymer from rotating with the screw. Thus they increase the coefficient of friction of the barrel wall. This leads to increased throughput and generation in the feed zone of the pressure required for flow through the following zones.

The feed bush zone must be cooled to prevent melting of the plastic. Figure 18 shows a typical grooved bushing with conically formed grooves. This eliminates the possibility of mixing due to back flow caused by the back pressure. Pressure flow, drag flow, and leakage flow contributions are oriented in the same direction. For this reason, mixing zones must be used to obtain satisfactory melt homogeneity.

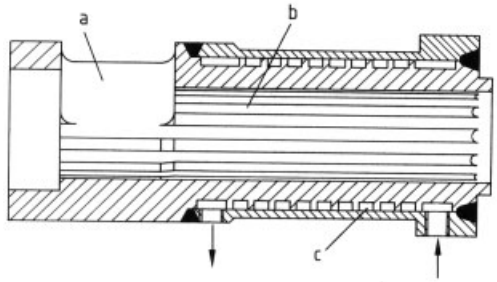

Figure 18. Forced-feed bush
a) Feed opening; b) Axial grooves; c) Cooling channels

Plastication Process. The heat required to melt the plastic is introduced by dissipation and by heat-transfer mechanisms. This includes heating by conduction from the heated barrel to the cold material and the generation of heat by internal and external friction (dissipation) in and between the material.

Twin-Screw Extruders. Twin-screw extruders exist in co- and counterrotating versions; the latter are also available with conically tapered screws. The screws are non-intermeshing or closely intermeshing.

Corotating extruders are used in compounding, whereas counterrotating, intermeshing systems with cylindrical or tapered screws are employed for processing rigid PVC (dry blend).

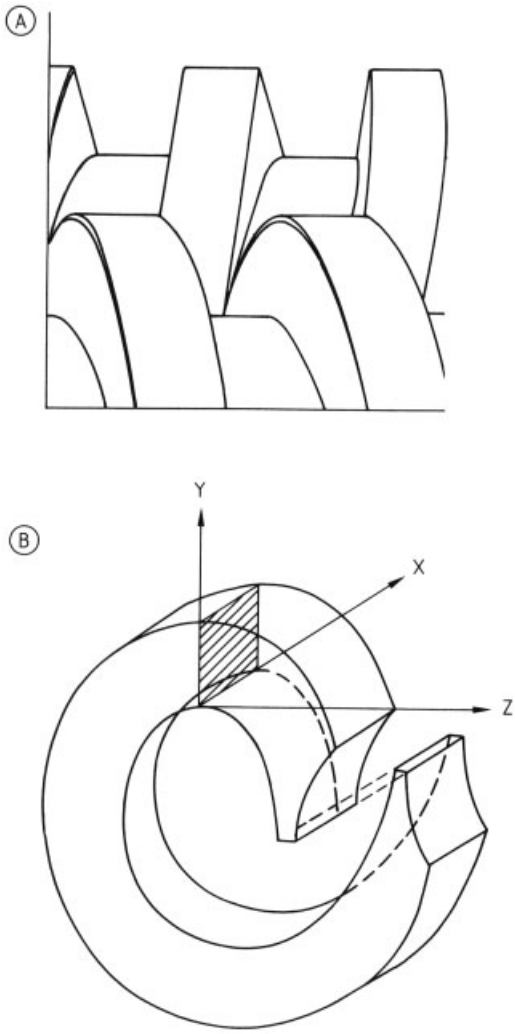

Figure 19. A) Intermeshing counterrotating twin screw; B) Strip of melt removed from one of the C-shaped chambers

Corotating Twin-Screw Extruders. With corotating twin-screw extruders, the melt contained in one screw channel is transferred to the other channel with each rotation. The transport mechanism (drag forces) is comparable to that of a single-screw extruder. The melt however is exposed to a greater shear stress due to the increased path length through the extruder.

This type of construction enables integration of closely intermeshing kneading and shear elements into the screw without any problem. Such extruders are almost exclusively employed for compounding.

Counterrotating Twin-Screw Extruders. The most widespread equipment for the extrusion of rigid PVC profiles and pipes consists of closely intermeshing, counterrotating twin-screw extruders. This type of extruder has a different transport mechanism. Each screw segment forms a sealed chamber, which means that during transport, virtually no exchange of material between neighboring chambers occurs over the full length of the screw from hopper to screw tip. Drag forces are not required for this forced conveying; hence, dissipation heating occurs to a considerably lesser extent.

Material heating occurs essentially by barrel heating, which can be controlled precisely and allows a mild treatment of sensitive products. The closely intermeshing screws form C-shaped chambers that are sealed by the screw flights. The C-shaped chamber is formed because the flight of one screw runs in the channel of the other and thus seals it (Fig. 19).

Generally, the counterrotating twin-screw extruder is operated with partially filled screw chambers. This is necessary to ensure a constant flux of material and an optimum balance of extruder, material, and mold. If the extruder were operated with full chambers, the volume of the individual chambers would have to be matched to the volume of polymer, which varies during heating and melting. Otherwise, very high pressures would be generated which lead to increased wear of the cylinder and screws.

The material being transported melts extremely slowly. Only in the final screw channels before the die does sintering occur, caused by the pressure buildup, and material is forced through the chamber gaps (leakage flow). This portion of material melts spontaneously under the short-term shear stress.

Although this transportation and melt mechanism guarantees well-controlled, mild treatment of the plastic, the product is often nonhomogeneous and insufficiently plastified.

The twin-screw extruder is a relatively expensive piece equipment due to the following factors:

1) Expensive bearings that are difficult to house
2) Complex drives
3) Barrel with a figure-eight bore

The intermeshing of the screws and the pressure buildup in the last screw channels lead to extensive wear of screw and barrel. Extruders with diameters up to 160 mm are commonly employed.

Degassing Extruders. During the melting process, volatile components, such as moisture, decomposition products, and residual monomers, may degas. Since this can lead to defects in the extrudate, these gases must be removed from the system. This is carried out with degassing screws in which, after the compression zone, the channel depth is increased (decompression) so that the pressure drops to ambient levels. A degassing opening in the barrel is incorporated at this position, through which the material is degassed by vacuum. Ideally, at this point in the extruder the material should not be fully melted and should have as large a surface area as possible. The degassing zone is followed by further compression and metering zones (Fig. 17E). A degassing screw can thus be regarded as a series connection for two screws, whereby the balance between the two is of importance in preventing the melt from leaking out of the degassing orifice.

Degassing extruders can be of both single- and twin-screw types. With counterrotating twin-screw extruders for PVC processing purposes, a degassing zone is standard.

Special Extruders. Apart from the extruders mentioned above, a range of extruders (Fig. 15) exist for special applications, all of which have individual advantages and drawbacks. The most significant of these are the cascade extruders in which two separate extruders are operated in series so as to divide the feeding and melting stages of the extrusion process from the homogenization and output stages, allowing better control and regulation of the process.

Ram extruders are piston plastication units employed in the processing of materials that are not suitable for processing with screws (ultra-high-molecular weight polyethylene, polytetrafluoroethylene).

Extruder Temperature Control. An extruder is heated by barrel heaters. Several heating zones are used to establish a temperature gradient along the barrel.

Commonly employed heating systems include the following:

1) Electrical heating with band heaters
2) Aluminum jackets with cast in heating wires and cooling pipes or ribs
3) Heating and cooling with a heat-transfer medium (oil, steam, pressurized water) that is in direct contact with the barrel walls (double-walled barrel with spiral ribs in the cavity)

The first two methods are sometimes combined with electric cooling fans. The third method is seldom used.

A thermocouple is located in the barrel wall of each heating zone, from which a temperature-dependent electrical signal is input to a control unit.

Additionally, during the processing of thermally sensitive materials such as PVC, the temperature of the screw is regulated with a heat-transfer oil.

Extrusion dies mounted on the extruder give the melt, the desired cross-sectional form. These forms include

1) Solid strand profiles (ribbon, quadratic, etc.)
2) Hollow profiles (shutters, window profiles, etc.)
3) Open profiles (U-shaped, etc.)
4) Pipes
5) Films
6) Sheets
7) Fibers

Furthermore, sheathing dies exist for cables, as well as a range of other constructions. The dies are heated, usually electrically.

All the dies contain a flow channel – the so-called runner system (melt distributor) through which the cylindrical melt strand leaving the extruder must flow. The following melt distributors are of importance.

The *spider* or *displacer* (*torpedo*) is the simplest runner system. It has a streamlined torpedo-shaped mandrel connected to the outer wall of the flow channel by so-called spider legs. The side facing the extruder is conical, but

this changes to the desired outlet cross-sectional form as it approaches the die outlet (Fig. 20).

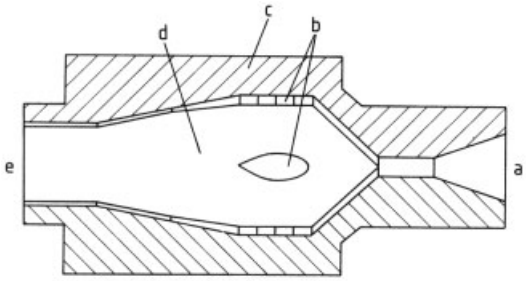

Figure 20. Spider-type manifold
a) Inlet; b) Spider legs; c) Die ring; d) Torpedo; e) Outlet

Applications: pipes, profiles, etc.

The *coat hanger distributor* (Fig. 21) spreads the melt over a required width. Melt enters the feed manifold, where the cross section tapers according to the decrease in mass flow, and flows over the so-called island region. This area contains choking elements (restrictor bars, adjustable lips), which can be used to correct irregular melt outputs across the die width.

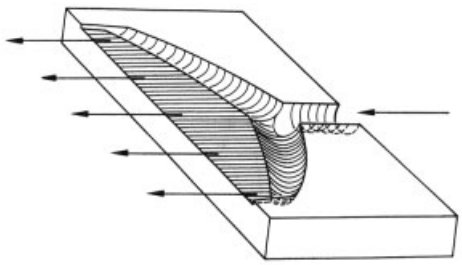

Figure 21. Coat hanger distributor

A variation of the coat hanger distributor is the fishtail distributor in which the feed manifold has a geometrically simplified form.
Applications: films, sheets, etc.

The *pinole* (side-fed die) is a cylindrical or conical coat hanger distributor (Fig. 22), which forms a tubular melt. What is characteristic, however, is that the incoming flow direction is not parallel to the outgoing cylinder, but perpendicular to it.
Applications: pipes, tubular film, cable sheathing, etc.

The *spiral mandrel distributor* is a core with multiple spiral-flow channels of progressively decreasing depth in the direction of the die outlet (Fig. 23). This causes separation of the melt flow into axial and circumferential components. The result is a melt that is homogeneously distributed around the complete circumference and is free from the flow lines and seams that can occur, for example, at flow junctions following the spider legs and with pinole dies.

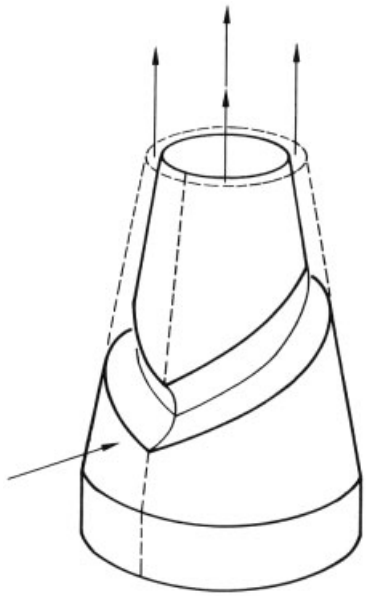

Figure 22. Pinole

Application: blown films (tubes).

Subsequent Equipment. The melt leaving the extrusion die must be fixed in shape and dimension; this occurs in the calibration stage. The extrudate is pressed against the walls of the calibrator by compressed air or a vacuum and is cooled. After leaving the calibrator, it must be sufficiently solidified that during subsequent cooling stages no further distortion can occur and the dimensions are fully retained. Thus the length of the calibration zone must be adapted to the throughput and the geometry of the extrudate.

Transport of the extrudate through the calibrator requires high takeoff tensions due to the friction between the extrudate and the surface of the calibrator. Doing this without deformation by using only the cooling capacity of the cali-

brator is rarely possible, and it is usual to include a subsequent cooling zone.

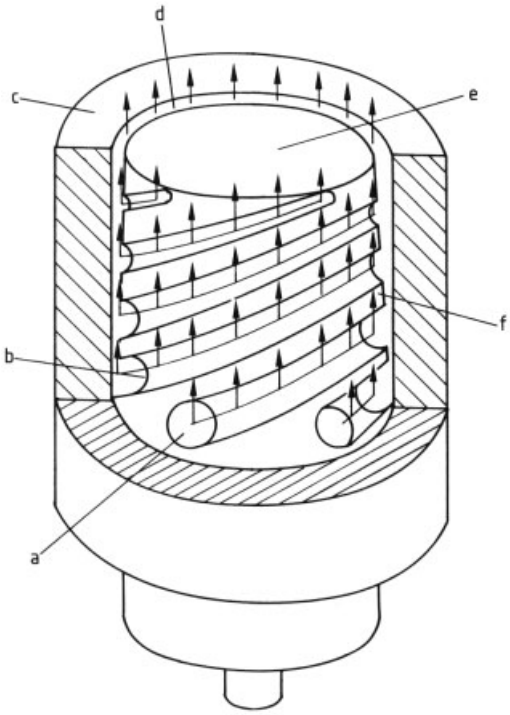

Figure 23. Spiral mandrel distributor
a) Feeding; b) Spiral channel; c) Casing; d) Annular gap; e) Mandrel; f) Steps

For profiles, pipes, cables, and similar products, water baths through which the respective extrudate is passed are also employed. Flat extrudates are cooled with rolls. It is also possible to employ water sprays and air cooling mechanisms.

The cooling zone is followed by the takeoff unit whose objective is to remove the extrudate from the die (via calibrator and cooling zone) at a constant speed. Various takeoff systems are available such as band, roll, or caterpillar takeoffs. Elastic extrudates such as cables, fibers, films, and tubes are wound on reels following the takeoff. Rigid extrudates (pipes, sheets, profiles, etc.) are cut to length by saws or guillotines and then stacked.

Examples of Equipment.

Profile Extrusion Line. Figure 24 shows a typical profile extrusion line. It consists of an extruder and a die followed by calibration, cooling, caterpillar takeoff, and cutting.

Flat film (*sheet*) *extrusion lines* (Fig. 25) consist of an extruder and a coat hanger die. These are followed by a roll mill, normally with three rolls, which calibrate and cool the film (sheet). Subsequent to this is a roller conveyor, which allows air cooling of the sheet and the takeoff roll. Finally, the film is wound, or the sheet is cut and stacked.

Blown film lines are employed to manufacture very wide films (Fig. 26). They consist of an extruder with a blown film head normally equipped with a spiral distributor. The vertically extruded tubular film is passed through a cooling ring where it is cooled by air. The tube is inflated to the desired circumference by blowing air into it. At the upper part of the tower, the tube is folded together on a collapsing board and the air is removed by nip rolls. Finally, the film is wound on reels. A calibration unit to maintain constant film bubble width can be mounted between the cooling ring and the collapsing board.

Coextrusion. The multitude of demands placed on modern extruded products can sometimes not be met by a single material. A solution to this problem is coextrusion, in which two or more molten materials are combined in a die under pressure. In adaptor coextrusion dies, the individual melt streams are combined in a so-called adaptor and subsequently shaped together in the melt distributor. In multiple manifold dies, each melt stream is individually shaped prior to being combined. Possible adhesion problems between various materials are addressed by incorporating coupling agents.

Well-known examples of the use of coextrusion include cable insulation and packaging films, where composite films with up to seven layers are used.

With coextrusion lines, the individual extruders and subsequent units are essentially identical to the corresponding units in the conventional lines.

1.4.4. Blow Molding

The most commonly employed techniques for manufacturing hollow articles from thermoplastic materials are extrusion blow molding and re-

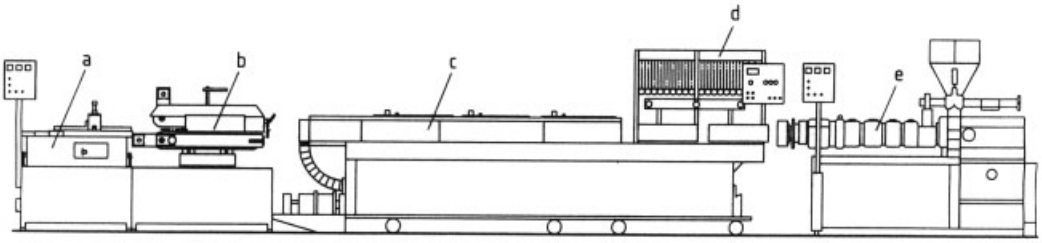

Figure 24. Profile extrusion line
a) Saw; b) Takeoff unit; c) Cooling system; d) Calibrator; e) Extruder

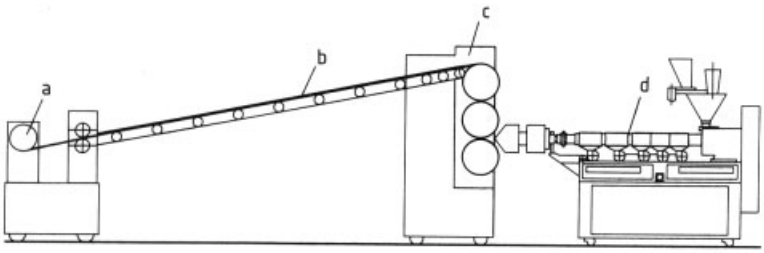

Figure 25. Flat film extrusion line
a) Winder; b) Roller belt; c) Roll mill; d) Extruder

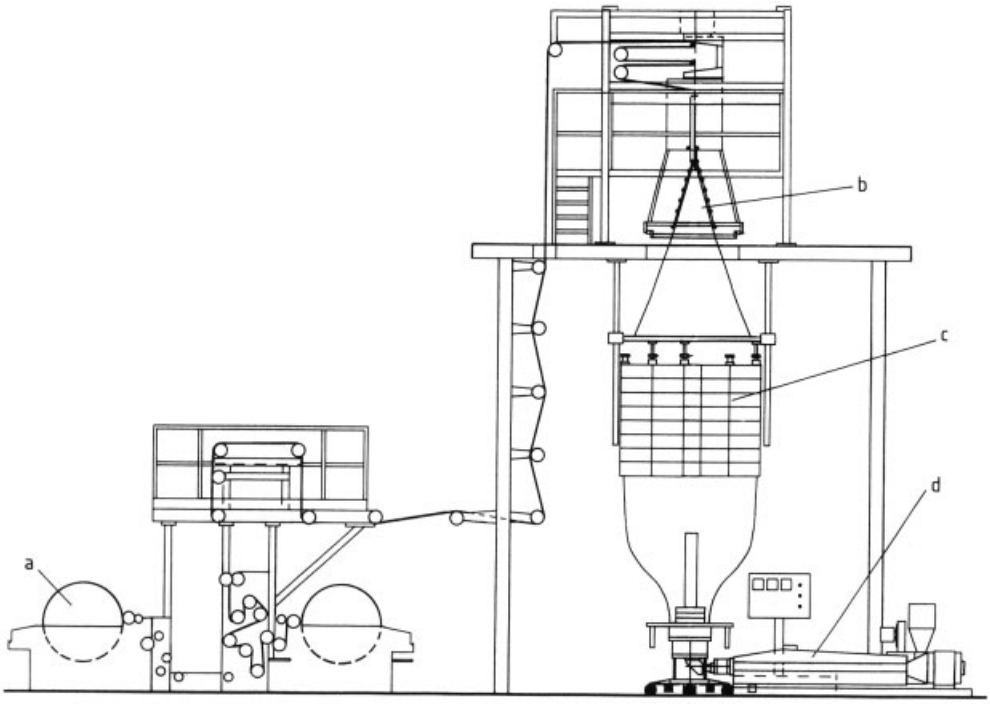

Figure 26. Blown film line
a) Twin winder; b) Guide rolls and collapsing boards; c) Calibrating basket; d) Extruder

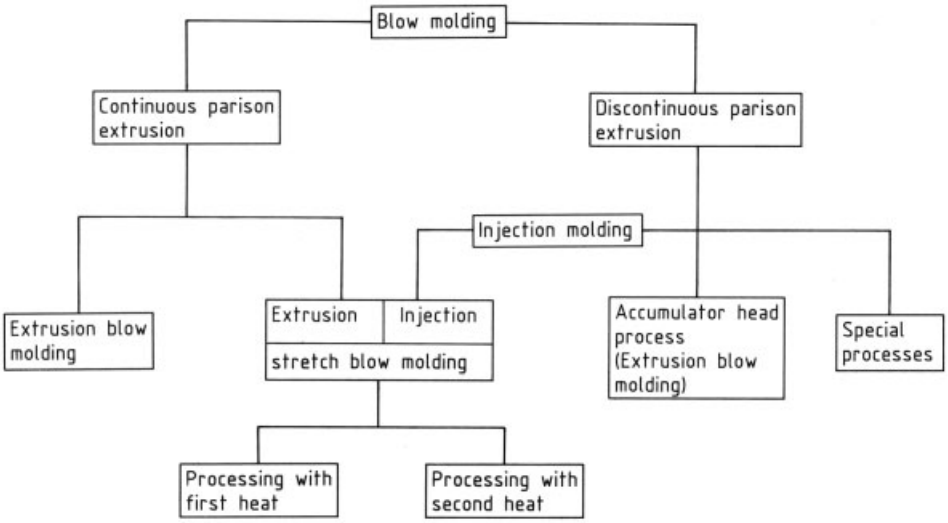

Figure 27. Blow molding processes

lated processes such as stretch blow molding. Figure 27 gives an overview.

Blow-molded articles include simple products such as bottles and more complex components such as ventilation ducts, surfboards, suit cases, roof racks, or fuel tanks for automobiles. The volume of such articles can vary from a few milliliters (packaging for medical products) to ca. 13 000 L (tanks for heating oil).

Extrusion blow molding is made up of two parallel processes:

1) Continuous extrusion of a preform (parison)
2) Batchwise shaping of the preform in a mold by using compressed air

The individual steps of the process are shown in Figure 28. The first step begins when the extruded tube (parison) has reached the required length. The mold closes around the parison, which is then cut by a blade. The mold is then conveyed to the blowing station, the blow mandrel is inserted into the mold, and the actual blow process is carried out. After cooling, the mold is opened and the finished molding removed. The open mold then returns to the position below the extrusion die so that it can accept a new parison. Large and heavy molds are not moved; instead, the parison is cut by a gripper that transports it to the mold.

An extrusion blow molding machine has four major components (Fig. 29):

1) Machine frame with closing mechanism and blowing station
2) Mold
3) Extruder
4) Crosshead and die

The machine frame is the basic framework of the machine and contains the closing mechanism as an integral component. The latter has the task of opening and closing the mold halves (similar to an injection molding machine), as well as transporting the mold between the die and the blowing station. The closing unit is normally of lighter construction than the counterparts involved in injection molding since the mold closing forces are relatively low because of the low mold internal pressure (ca. 0.8 MPa). The closing mechanism is driven hydraulically.

The blowing station is that part of the equipment in which the mold is located when the blowing mandrel is inserted and the actual blow process occurs.

The blowing mandrel introduces inflating air and shapes and calibrates the neck of the finished article. Figure 30 shows a closed mold containing a parison and the inserted blow mandrel.

The mold consists in general of two halves in which the negative contours of the molding have been incorporated. This has the function of shaping the parison, cooling the molding, and

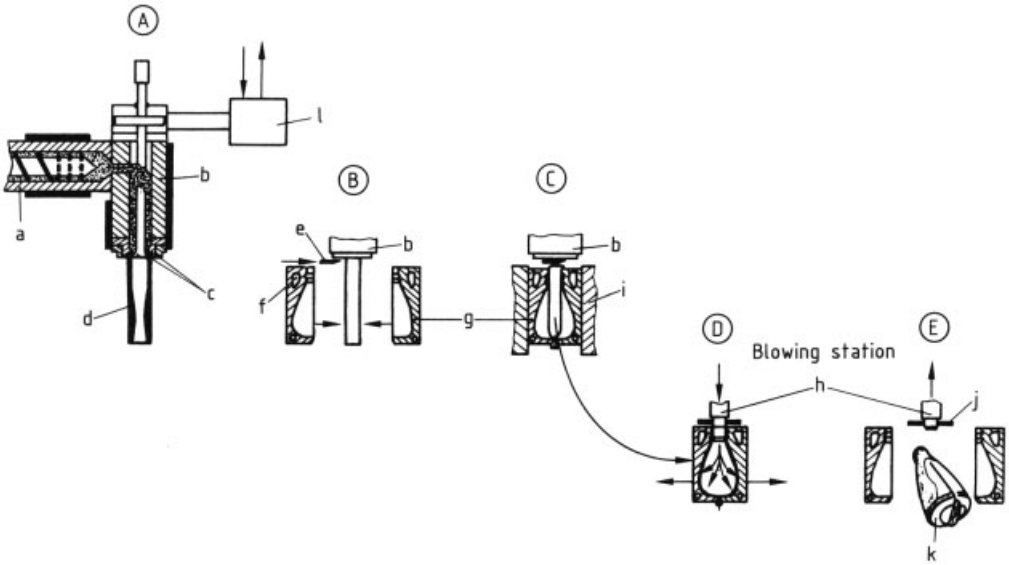

Figure 28. Extrusion blow molding process
A) Parison extrusion; B) Positioning of blow mold; C) Collecting and cutting of parison; D) Shaping and cooling; E) Demolding and flash trimming
a) Extruder; b) Crosshead die; c) Nozzle and mandrel; d) Parison; e) Cutter; f) Cooling channel; g) Blow mold; h) Blow mandrel; i) Closing unit; j) Stripper ring; k) Molding; l) Servohydraulic

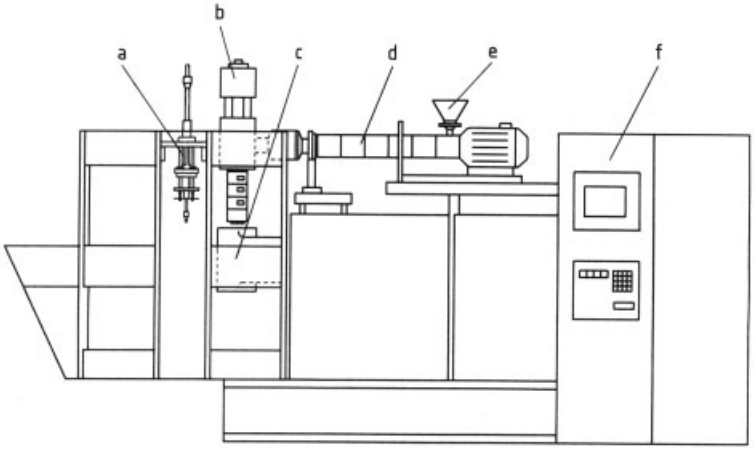

Figure 29. Extrusion blow molding machine (Krupp – Kautex type)
a) Blowing station; b) Blowing head; c) Mold; d) Extruder; e) Metering station; f) Control unit

forming the welds. The latter are necessary because the parison is a portion of an extruded tube, which can only be blow molded when the ends have been sealed, preventing the escape of compressed air. In the case shown in Figure 30, one of the ends is sealed by the blow mandrel. At the other end, the tube is compressed during the closing motion by the pinch-off edges of the mold, and the molten material is welded together. These weld seams are a telltale characteristic of extrusion blow molded products. Thus, materials for blow molding must be weldable.

Cooling is possible on only one surface of the molding – the outer surface – by a water-cooled mold. Hence, cooling times are longer than those associated with injection molding, in which heat can also be removed through the core of the mold. This problem is counteracted by incorpo-

rating an exhaust air bore in the blowing mandrel. Compressed air flows through this and over a throttle (to maintain the required pressure); hence a circulating exchange of air takes place within the blow section, which also has some heat-transport effects (purging air technique).

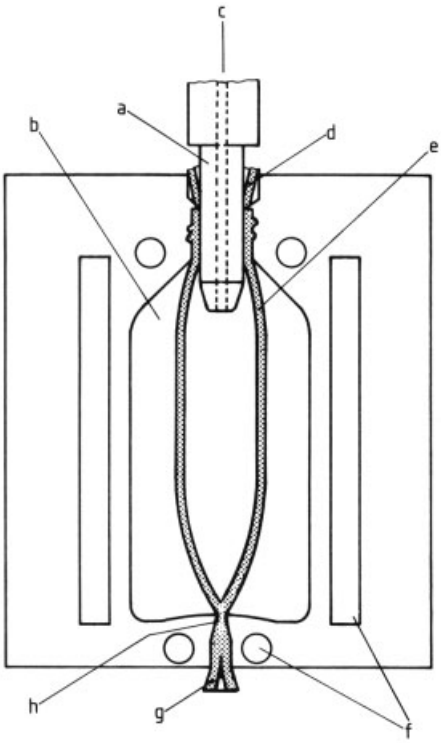

Figure 30. Blow mold
a) Blow mandrel; b) Mold cavity; c) Blowing air inlet; d) Top flash; e) Parison; f) Cooling channels; g) Bottom flash; h) Pinch-off blade

Cooling times can be reduced further by using highly cooled nitrogen (N_2 process) as a blowing medium or adding liquid carbon dioxide (CO_2 process) to the compressed air.

Because compressed air seldom has a pressure greater than 0.4 – 0.6 MPa, no great demands are placed on the mechanical design of the mold. Due to its good thermal conductivity and workability, aluminum is widely used.

Like other processing techniques, the extruder must provide a thermally homogeneous melt stream of predetermined flow rate and pressure. For blow molding, single-screw extruders with a screw length of 20 – 25 D are generally used. A grooved and cooled feed zone is also used for some plastics (see Section 1.4.3).

In general, the screws are equipped with mixing zones to attain adequate melt homogeneity.

The so-called crosshead or parison die is flanged onto the extruder. This has the task of changing the melt flow direction from horizontal to vertical and forming a tube.

Near the outlet region, the flow channel is generally conical, and either the inner or the outer cone is movable, allowing continuous adjustment of parison wall thickness during extrusion. In this way, uneven blow ratios, which can occur with articles of complex geometry, can be accommodated and products with almost constant wall thicknesses can be manufactured (Fig. 31). This technique is referred to as wall thickness programming.

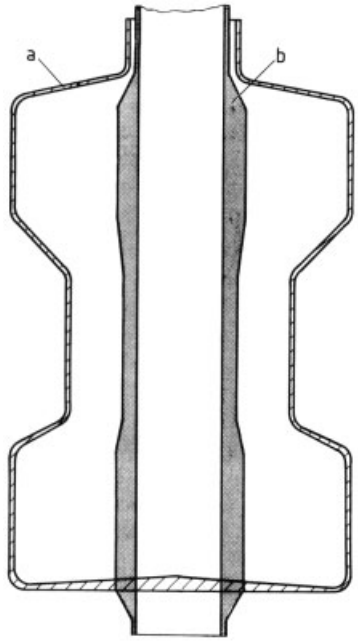

Figure 31. Wall-thickness profiled parison
a) Blown product; b) Parison

A distinction is made between blow heads that function in a continuous mode and so-called accumulator heads. In the former, a continuously extruded tube is cut and passed to the blow mold in a cyclic operation. With large articles (i.e., molding volume greater than ca. 30 L), the parison itself is so heavy that it may deform under its

own weight during extrusion. In such cases, accumulator heads are employed in which the melt is collected in a storage chamber and rapidly ejected by a piston to form the parison (Fig. 32). Modern storage heads are equipped with ring-shaped storage chambers, which are continuously fed with melt by one or more extruders according to the FIFO (first-in, first-out) principle.

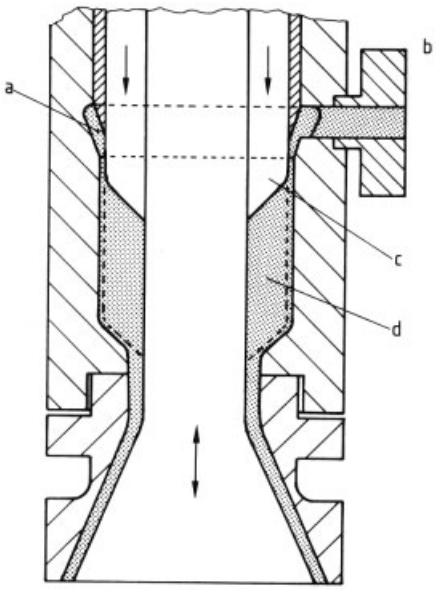

Figure 32. Accumulator head
a) Ring channel; b) Extruder; c) Annular piston; d) Accumulator chamber

Use of parison wall thickness regulators is also possible with storage heads.

Stretch blow molding is a special form of blow molding. The process exploits the fact that high degrees of orientation can be introduced in plastics by drawing at temperatures near the glass transition temperature or crystalline melting point. Mechanical properties are considerably improved in this way. In this process, the preform is drawn not only in the circumferential direction (as with extrusion blow molding) but also in the machine direction. The longitudinal drawing is brought about mechanically by using a ram. Because of the relatively low temperature at which this process is carried out, high deformation forces are required (air pressure up to 2 MPa).

Figure 33 illustrates the principal steps involved in stretch blow molding, beginning with an injection molded preform.

Preferred materials for use with this process are PVC, polypropylene (PP), and poly(ethylene terephthalate) (PETP).

The preforms for stretch blown moldings are prepared mostly by injection molding and are subsequently either oriented immediately in an injection stretch blow molding machine or first cooled and then fed to a separate stretch blow molding machine.

Preforms made of PVC for example, which are not suited for manufacture by injection molding, are prepared in an extrusion blow molding process and then processed on extrusion blow molding equipment.

Injection-molded preforms have the advantage of exact wall thickness distributions and a precisely calibrated neck. Additionally, they are free of weld seams.

1.4.5. Injection Molding

Injection molding is the most important process for producing moldings from thermoplastics, elastomers, and thermosets. The significance of injection molding is due to its ability to manufacture complex molding geometries in a single stage with high levels of reproducibility. A finishing operation (removing the runner, deflashing) is seldom necessary. Because of the high degree of automation possible, this molding process is also suitable for mass production operations.

Apart from its use for commodity articles, this technique is increasingly being used to produce technical moldings (e.g., compact discs, automobile bumpers).

The size of the molding can vary from a few tenths of a gram to several kilograms.

Injection molding machines consist of several individual units (Fig. 34). The most important components are the plastication unit for melting the plastic pellets, the clamping unit for holding and closing the injection mold, the control unit for coordinating the process, and a temperature control unit for cooling or heating the mold.

Injection molding is a discontinuous process in which plastic pellets, granules or powder is

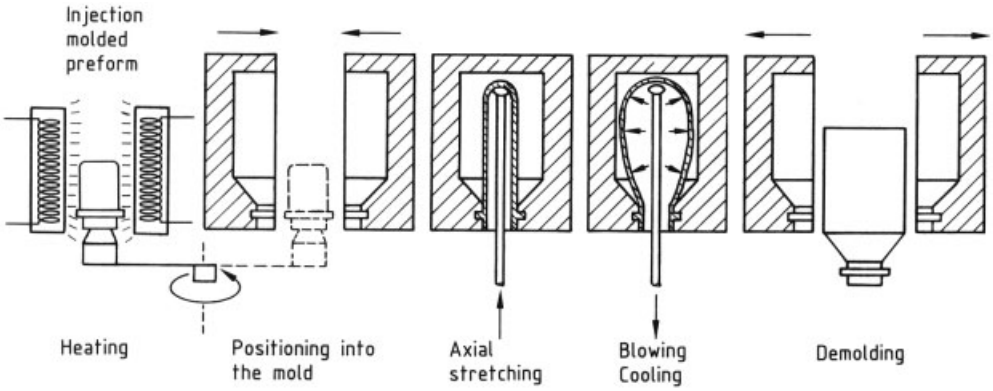

Figure 33. Stretch blow molding process

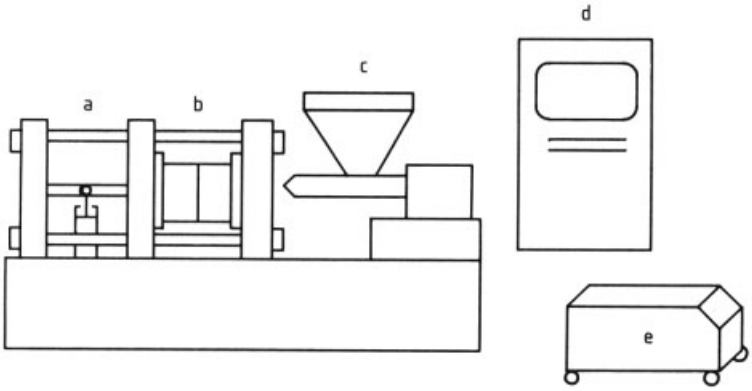

Figure 34. Schematic of an injection molding machine
a) Clamping unit; b) Mold; c) Plasticating unit; d) Control unit; e) Temperature control unit

melted and injected under pressure into the cavity of a mold, where it is solidified by cooling or thermally cross-linked.

Process Stages. Production of an injection molding involves the following steps, which make up the injection molding cycle (Fig. 35):

1) Closing the mold
2) Injection
3) Holding pressure stage
4) Cooling stage
5) Metering
6) Opening the mold
7) Ejection/removal of the molding

This cycle is repeated for each molding.

The cycle begins by closing the mold after ejection of the molding from the previous cycle (Fig. 36). The hot melt prepared in the front part of the injection cylinder is injected into the mold cavity through the runner system. Because of the increasing length of the flow path and the cooling of the melt in the mold, the pressure required to move the cylinder and polymer melt increases. Toward the end of the filling stage, the maximum injection pressure must be exerted. Depending on the material, molding, and process control, this can vary from several bar to 2000 bar. Injection times vary from several hundredths of a second to minutes.

The holding pressure phase begins at the end of the injection phase. Its objective is to counteract volume losses caused by cooling of the melt in the mold and hence avoid sink marks and voids in the molding. At the beginning of the injection process, the melt, which is at 200–300 °C, begins to cool in the mold. After the holding pressure stage, the molding must be cooled until its dimensional stability is sufficient for ejection. The cooling time thus extends from the beginning of injection until the mold is opened.

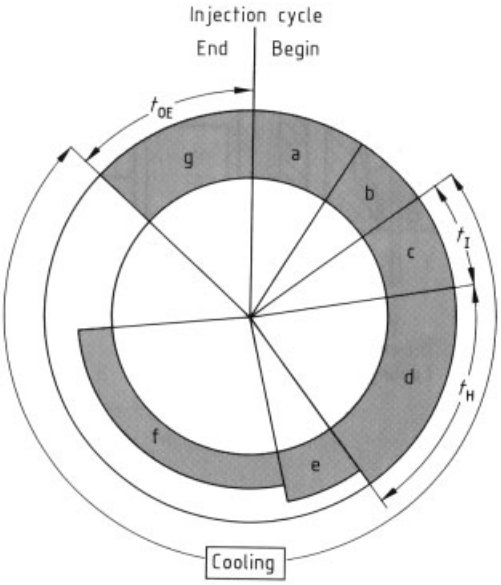

Figure 35. Injection molding cycle
a) Closing the mold; b) Moving the plasticating unit forwards; c) Injection; d) Holding pressure phase; e) Moving the plasticating unit backward; f) Metering; g) Opening the mold and ejection of molding

At the end of the holding pressure phase, the plastication unit begins to prepare the plastic for the next cycle by rotating the screw. The screw moves helically backward in the plastication cylinder against a back pressure and thus transports melt to the area in front of the screw.

After the cooling stage, the mold is opened and the molding is removed mechanically, pneumatically, or by handling equipment.

The total cycle time depends on the maximum wall thickness and the geometrical complexity of the molding. With simple, thin-walled moldings, cycle times of ca. 1 s are possible, whereas with bulky, thick-walled moldings, cycle times of several minutes may be necessary. The process itself is coordinated by a control unit and is fully automated.

Machine Components.

Plastication Unit. The plastication unit of an injection molding machine plastifies, meters, stores, and injects the plastic.

To meet these requirements, general-purpose screws are used (Fig. 37).

With these so-called reciprocating screws, the polymer is plasticated by the rotary motion of the screw as in an extruder (see Section 1.4.3) (Fig. 38). Simultaneously with the turning movement, the screw moves backward from the forward end position against a back pressure. This translatory motion results from transport of the melt into the area in front of the screw.

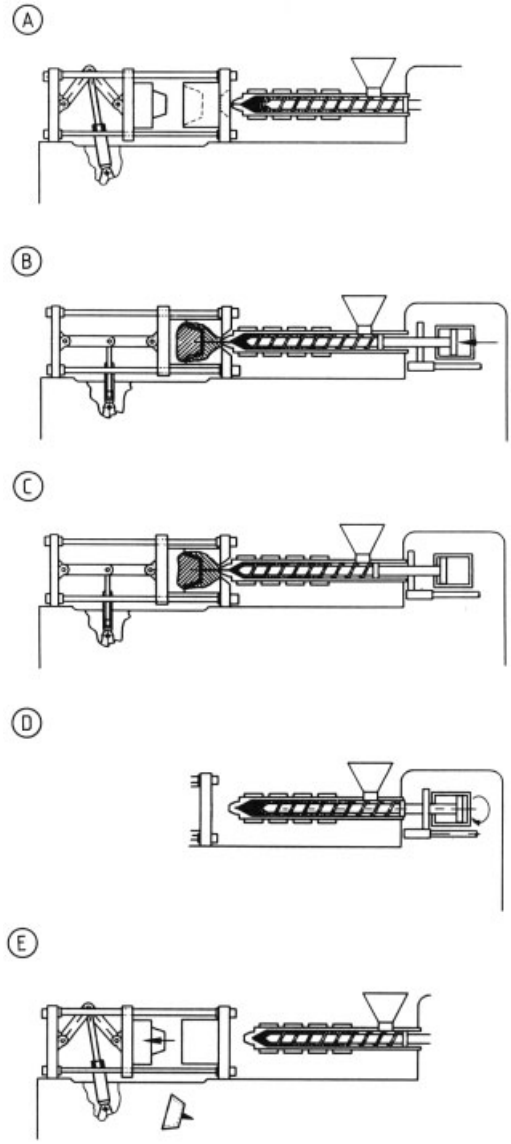

Figure 36. Machine status in various stages of injection molding cycle
A) Closing the mold; B) Injection; C) Holding pressure and cooling stage; D) Metering; E) Mold opening and ejection of molding

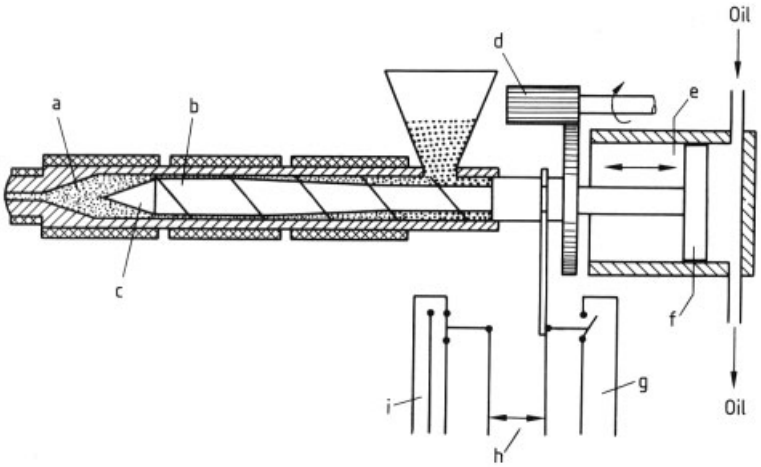

Figure 37. Conventional plasticating unit
a) Shot chamber; b) Screw; c) Screw tip; d) Screw drive; e) Hydraulic cylinder; f) Hydraulic piston; g) Back limit switch (stops rotation of screw); h) Injection stroke; i) Front limit switch (switching from injection pressure to holding pressure)

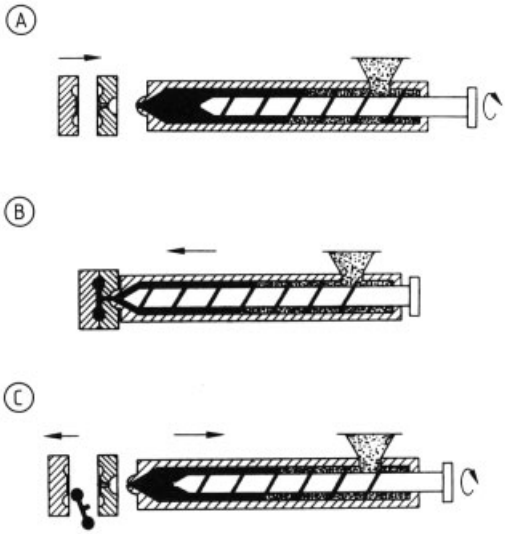

Figure 38. Reciprocating screw injection molding machine
A) Melt in front of the screw; B) Injection by screw stroke; C) Demolding and plasticating by screw rotation

At the end of the metering step, material is injected into the mold by a purely axial motion of the screw. Both the rotational and the translational movements of the screw are generally carried out hydraulically. However, using electric drives for the rotational motion is becoming more common, because they consume less energy than hydraulic systems.

The injection pressure for axial displacement of the screw is provided by a hydraulic cylinder.

The general-purpose screw has three zones: (1) feed zone, (2) compression zone, and (3) metering zone.

The feed zone draws the polymer granules from a hopper into the screw. To achieve high mass throughputs, despite the low bulk density, the channel depth of the screw is relatively high.

In the compression zone, material transported from the feed zone is compressed.

In the output or metering zone, the melt is homogenized and heated to processing temperature.

Screw diameters range from 14 to 200 mm, and screw lengths, between 18 D and 24 D. To adapt the injection volume to the volume of the molding, usually the stroke of the screw (i.e., the distance between the front position after injection and the back position after plastication) is varied. The screw stroke cannot be increased to an unlimited extent by varying the back position, because when the screw is moved back the remaining screw length decreases. Therefore the effective length of the screw for melting and homogenizing the plastic is decreased, which results in a lower temperature for the plasticated material.

It is also possible to exchange the plastication unit and, therefore, the screw diameter and maximum stroke on a given machine.

During plastication, the machine nozzle that joins with the mold is sealed either by the runner of the molding or by a shutoff nozzle (Fig. 39) to

prevent leakage of the melt caused by the pressure under which the plasticated material is held.

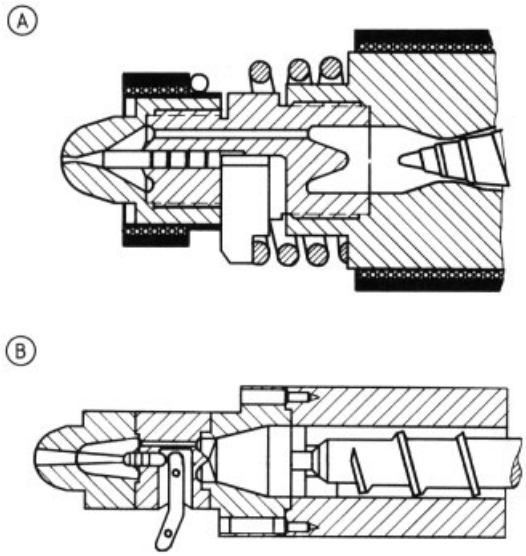

Figure 39. Shutoff nozzles
A) Needle valve nozzle (spring loaded); B) Needle valve nozzle (separately controlled)

To prevent back flow of the melt during injection, a nonreturn valve is sometimes mounted at the tip of the screw (Fig. 40). The locking ring is pushed forward during plastication, which allows the melt to advance; during injection, the ring is pushed back, sealing the area in front of the screw.

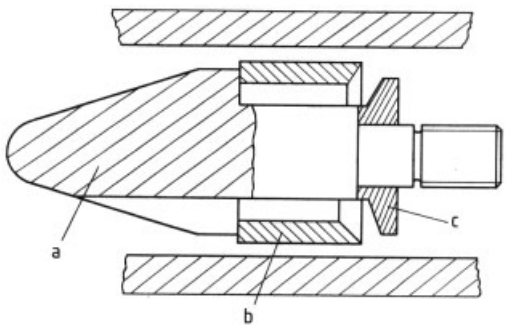

Figure 40. Nonreturn valve with a mobile locking ring
a) Screw tip; b) Sliding ring; c) Thrust ring

Besides the general-purpose screw, screw geometries also exist that are specially adapted to suit the polymer to be processed. These include thermoset screws, rubber screws, and special screws. With degassing screws (Fig. 41), any moisture or residual monomer present in the melt is removed through an opening in the cylinder wall.

The most significant advantages of the reciprocating screw system include the mild plastication that allows the processing of thermally sensitive products, the lower cycle time, and compact construction compared to ram plasticators. Apart from the horizontal construction illustrated in Figure 34, the plastication unit can also be mounted vertically.

Although such reciprocating screws are most frequently employed, specialty constructions also exist, for example ram extruders with screw preplastication units (Fig. 42). The polymer is plasticated in an extruder, and the resulting melt flows through a nonreturn valve into the reservoir chamber in front of the injection ram. Such systems are used only in special cases.

Clamping Unit. The clamping unit contains the halves of the mold, that are mounted on platens, and opens and closes the mold. Additionally, in most cases, mechanisms for ejecting the molding (e.g., cylinders for activating ejector pins) are also housed in the clamping unit.

During opening or closing of the mold, only one platen is moved. The platen on the side nearest the plastication unit is fixed to the machine frame and, hence, is immobile. The movement of the closing-side platen is guided by four tie rods (Fig. 43).

After closing, the full clamping force is exerted to prevent the mold from opening as a result of internal pressure during the injection and holding pressure phases. Otherwise, melt leaks in the parting surface (two mold halves) and which leads to formation of flash and possibly damage to the mold.

Clamping forces of injection molding equipment range from 0.1 to 100 MN. Two main types of clamping mechanism are used:

1) Toggle clamping units (Fig. 43)
2) Fully hydraulic clamping units (Fig. 44)

In the toggle clamping unit, the closing force is generated by extending a toggle driven by a hydraulic cylinder. The advantage of the toggle clamping unit is the favorable force – speed characteristic. High manufacturing costs are the main drawback.

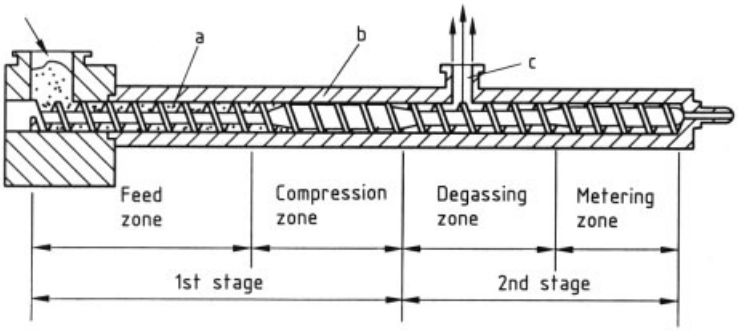

Figure 41. The degassing – plasticating unit of an injection molding machine
a) Melting; b) Cylinder; c) Degassing outlet

With a fully hydraulic clamping unit, the clamping force is applied by a hydraulic cylinder alone. The advantage of this system is cheaper design; its disadvantage is the generally slower opening and closing motion.

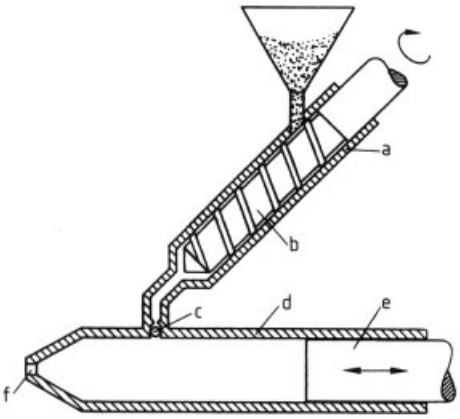

Figure 42. Injection molding with screw preplastication
a) Extruder barrel; b) Screw; c) Nonreturn valve; d) Injection cyclinder; e) Piston; f) Nozzle

Injection Molds. The central component in the injection molding process is the mold itself (Fig. 45). The injection mold does not form part of the machine, but must be constructed or modified for each molding. The number of design variations corresponds to the multitude of different injection-molded components. All molds, however, fulfill the following basic tasks. From a technological point of view, the mold must

1) Distribute the melt
2) Shape the melt to the desired molding geometry
3) Cool the melt (thermoplastics) or heat the contents of the mold (elastomers, thermosets)
4) Eject the molding

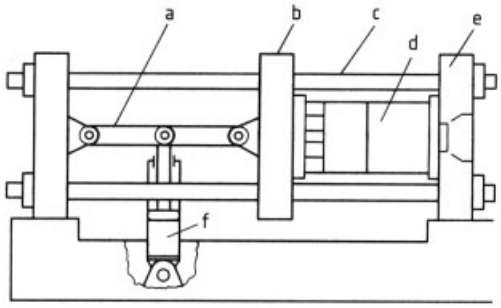

Figure 43. Toggle clamping unit
a) Toggle system; b) Moving platen; c) Tie rod; d) Mold; e) Fixed platen; f) Driving cylinder

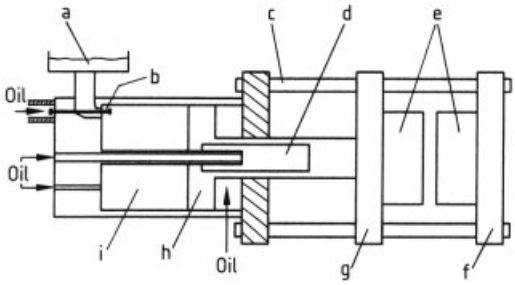

Figure 44. Hydraulic clamping unit
a) Oil tank; b) Suction valve; c) Tie rods; d) High-speed injection cylinder; e) Mold; f) Fixed platen; g) Moving platen; h) Main piston; i) Main cylinder

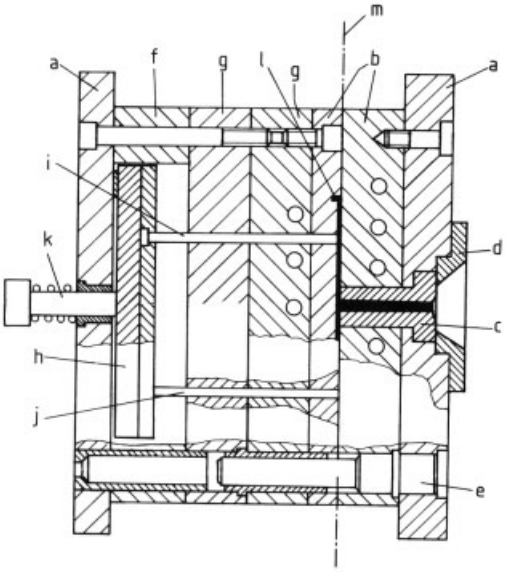

Figure 45. Injection mold
a) Base plate; b) Mold plate; c) Sprue bush; d) Locating ring; e) Guide pin; f) Supporting strip; g) Support plate; h) Ejector plate; i) Ejector pin; j) Ejector plate return pin; k) Ejector bolt; l) Molding; m) Parting line

Additionally, the mold must meet design objectives, such as

1) Accomodate all forces
2) Transfer all motion
3) Guide mold components

To allow ejection of the molding, the mold consists of at least two parts. The front half, which contains the gate system and possibly also the flow channels, is mounted on the fixed, machine platen; the other half including the ejector mechanism is mounted on the moving platen of the clamping unit. Centering devices are used for precise adjustment.

In some cases, the molds have exchangeable inserts, which allow a basic unit to manufacture various moldings.

The mold is joined to the nozzle of the injection unit by the runner channel. The radius of the sprue bushing must always be greater than the radius of the nozzle, and similarly, the diameter of the sprue bushing bore must be larger than the nozzle bore. With voluminous moldings, a conical sprue gate is often used (Fig. 46). When the mold is opened, this is broken off at the nozzle. It remains as a stalk on the molding and must subsequently be removed. With a pin gate, the melt is injected through short, narrow holes into the mold cavity. During demolding, the gate breaks at the junction with the molding; no finishing process is required. Disk and film gates are used, for example, in the production of gear wheels or cartridges to avoid weld lines during filling. The film gate is advantageous with flat moldings because in such cases he linear filling avoids warpage and distortion of the molding.

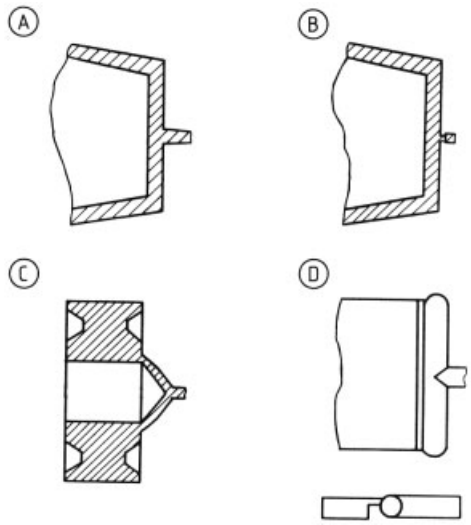

Figure 46. Gate types
A) Sprue gate; B) Pin gate; C) Disk gate; D) Film gate

Multicavity molds, which are used for simultaneous production of several moldings, are equipped with mold cavities connected to the gate channel by the manifold system. As with multiple gates on an individual mold, the gate and the runner system (hot runner mold) can be heated to reduce material requirements. Material in the hot runner system remains molten and is thus available for use in the next filling stage.

The runner system should be designed so that the melt fills each cavity uniformly and simultaneously at the same temperature and pressure. The position of the gate is chosen so as to avoid as much as possible the formation of weld lines that result when melt substreams combine. The position of the gate also determines the flow direction and hence the orientation on which mechanical strength, shrinkage, and distortion of the molding depend.

Special Processes. In addition to the classical injection molding process described above, several variations exist.

Injection – Compression Molding. With thin-walled, large-area moldings (i.e., when long flow lengths are involved), mold resistance is often so large that extremely high injection pressures must be exerted to fill the mold. With injection – compression molding, the mold is filled at only partial clamping pressure, causing the mold halves to open slightly (Fig. 47). To complete filling, the full clamping pressure is applied, whereby the clamping unit essentially operates as a press. The holding pressure exerted by the injection unit prevents the flow of the melt out of the mold cavity. Due to the homogeneous pressure over the complete injection area, moldings without warpage are obtained.

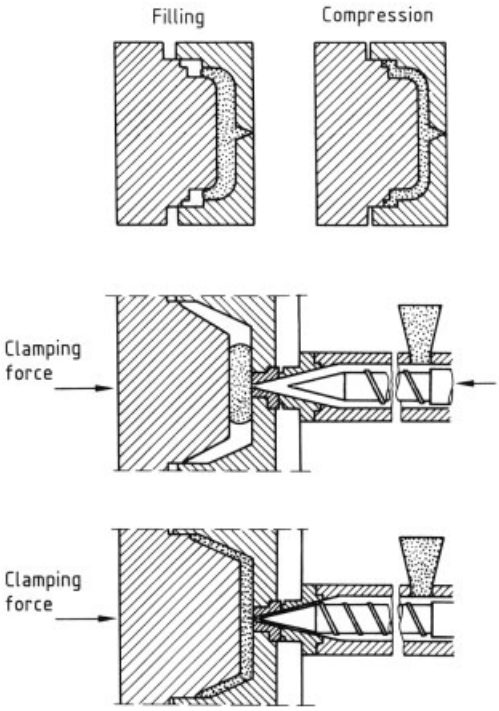

Figure 47. Injection – compression molding process

Intrusion Injection Molding. With thick-walled moldings, the injection pressure required for mold filling is usually so low that the pumping pressure from the rotating screw is sufficient. In the flow molding process (intrusion), the screw near the front end position plasticates the material and pumps the melt relatively slowly into the mold cavity under extrusion pressure alone. The holding pressure that counters shrinkage of the cooling molded component is applied by means of a short translational movement of the screw after the filling stage is completed. This process is used in the manufacture of thick-walled containers made of polyethylene and of PVC fittings whose volume exceeds the injection volume of the equipment by a factor of three to five.

Two-Component Injection Molding. The principal steps involved in two-component injection molding are shown in Figure 48.

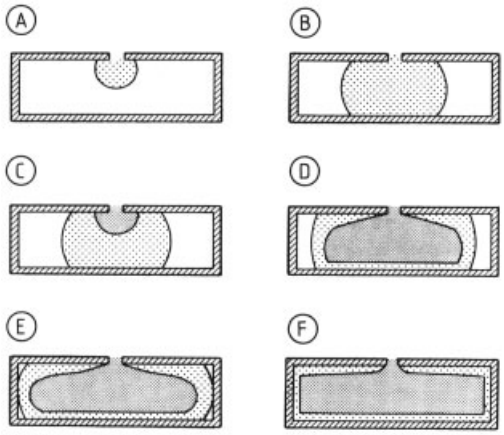

Figure 48. Two-component injection molding

The mold is first partially filled with material A. After the changeover point has been reached, complete filling takes place by injection of component B. During this stage, material A is displaced by material B and forced against the walls of the mold. A multilayer molding is obtained in which the surface component A completely encloses the core component B. The external appearance of the molding and other surface properties are hence determined exclusively by material A.

The possible combinations of surface and core material are essentially unlimited, if adhesion between the two components is adequate and both materials exhibit similar shrinkage behavior. Two-component injection molding also offers economical and environmental advantages. The use of polymer waste and re-

granulate as the core material allows the molding to be manufactured at lower cost.

A second plastication cylinder is required for two-component injection molding.

Multicolor Injection Molding. In contrast to the two-component process, a sandwich-structured molding does not result with multicolor injection molding. In this case, all of the components form an integral part of the molding surface. This is illustrated in Figure 49 by using a typewriter key as an example. Another application is in the covers of automobile taillights; these are also injection molded with several colors.

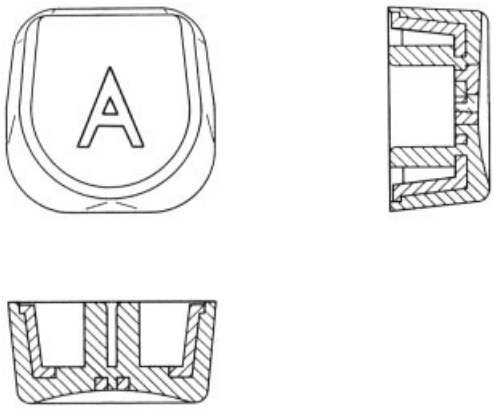

Figure 49. Two-color injection molding of a typewriter key

Such combinations of different colors are achieved by rotating one-half of the mold (Fig. 50). First, one component is injected into a defined part of the mold cavity, while those regions of the cavity intended for components of another color are sealed. By rotating one-half of the mold, these regions are made available for filling with additional components.

Thermoplastic Foam Casting. Structured foam moldings are manufactured from thermoplastics (usually polystyrene, polystyrene copolymers, and polyolefins) containing a blowing agent. Moldings with a compact surface and a foamed core are obtained, which are employed mainly in the audio and furniture industries, as well as technical applications. Ram injection molding machines with screw preplastication units are generally used. The process requires lower injection pressures and closing forces than conventional injection molding techniques. The mold is only partially filled (50 – 80 %). Complete filling of the mold results subsequently from the foaming process. A holding pressure phase is hence unnecessary.

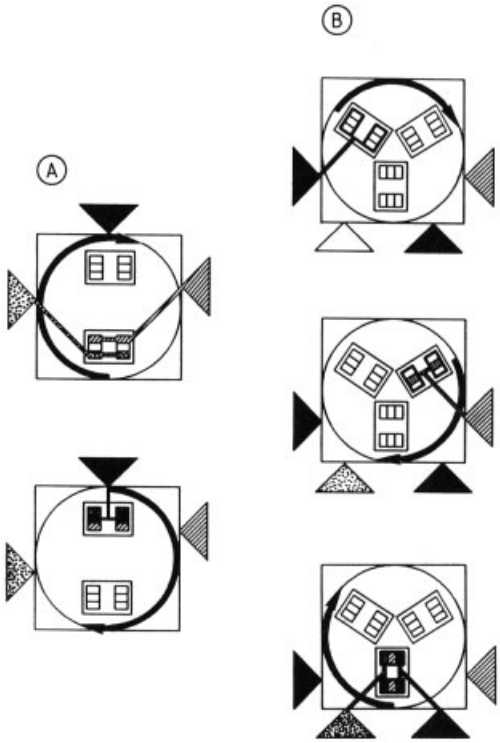

Figure 50. Multicolor injection molding
A) Three-color injection molding; B) Four-color injection molding

1.5. Forming Processes

The term forming refers to processes in which semifinished plastic products are heated and shaped under the influence of external forces. Apart from modifying the shape of an article, an objective of the process is often to increase material strength by introducing molecular orientation.

1.5.1. Drawing

To obtain an anisotropic material behavior in thermoplastics (usually an increase in mechanical stiffness and strength), the molecular chains

must be oriented and the orientation frozen in. One way of achieving this is by stretching or drawing. The original semifinished product is heated to the drawing temperature and then mono- or biaxially drawn, followed by fixing the oriented state by cooling while the final shape is maintained. With amorphous thermoplastics, drawing is normally done at or around the glass transition temperature, and with semicrystalline materials, slightly below the crystalline melting temperature. When a drawn product is reheated, the forming process is reversed due to elastic deformations frozen into the article. If the draw ratio is not too high, the article may even revert to its original form. This ability to return to the original form can be desirable in some cases (e.g., shrink wrapping films). Generally, however, the product must be tempered (heat set) after drawing to relax any frozen-in elastic stresses.

The drawing of monofilaments and films is of major industrial importance.

Monofilaments are drawn in so-called godets (driven draw rolls). The fibers are fed through a preheating zone (oven or tempering bath) and taken off at high speed. The draw ratio is determined by the adjustable speed difference between individual godets. Typical draw ratios range from 1 : 3 to 1 : 12. Finally, drawn fibers pass through a tempering zone in which elastic deformations can relax while the orientation that has been introduced is retained. To compensate for shrinkage, the draw rolls run at slightly reduced speeds in this area. The whole process takes place simultaneously on many fibers, which are subsequently wound individually on spools. At this stage, a constant filament tension must be maintained. With multifilament machines, several hundred filaments are drawn simultaneously. Cold drawing is in principle a similar process, in which preheating is omitted. The fibers are heated by the internal shear to which the material is exposed on constriction.

Films are also drawn to increase their mechanical strength. Another reason, however, can be to reduce thickness. The film is heated by heated rolls and then passed over two pairs of rolls, with the second pair rotating at a higher speed than the first. This leads to monoaxial orientation in the machine direction. During this step, the film necks. If the film is to be biaxially oriented, it is held at both edges by a series of clips mounted on a chain that draw the web in the transverse direction (Fig. 51). This process can be carried out simultaneously with longitudinal drawing or separately in a second process step. Once the drawing process has been completed, the film is cooled and wound in a roll. As with calenders and fiber spinning equipment, biaxial drawing units with integrated film extrusion are expensive. The most important oriented films include polyethylene, polypropylene, polystyrene, and polyester. Major products include films for magnetic tapes and for capacitors.

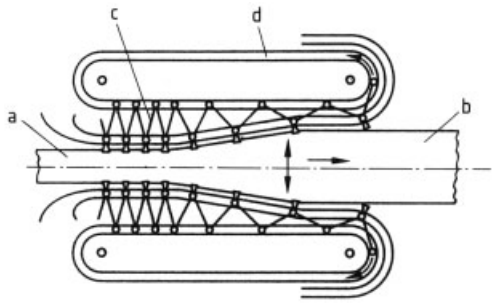

Figure 51. Biaxial simultaneous drawing
a) Initial sheet; b) Drawn sheet; c) Tenter chain; d) Guide frame

1.5.2. Thermoforming

The manufacture of three-dimensional moldings from flat plastic preforms such as films or sheets, under the influence of heat and pressure or vacuum, is a process referred to as thermoforming. In principle, the process can be carried out by a variety of techniques. Industrially, the most important method is the molding of a heated preform using a vacuum or compressed air, as well as mechanical stretching. Forming is carried out in the rubber-elastic temperature range, although in contrast to drawing , the introduction and freezing-in of orientation must be avoided as much as possible. For this reason, forming temperatures as high as possible are selected, which are essentially limited only by the onset of flow of the material, so that sufficient opportunity to relax is available. Depending on the material and temperature, geometric draw ratios of several hundred percent can be attained in thermoforming.

The range of products that can be manufactured by this process varies from packag-

ing containers (up to 100 000 articles per hour) to bulky moldings such as swimming pools 8 m × 4 m × 1.5 m (1.5 articles per hour). Furthermore, this is also an important process for the production of automobile components such as dashboards and passenger compartment panels.

The preferred materials for processing include amorphous thermoplastics [PVC, polystyrene (PS), acrylonitrile – butadiene – styrene (ABS), styrene – acrylonitrile (SAN), cellulose acetate butyrate (CAB), PMMA, and PC], as well as semicrystalline materials (PP, PE, PETP) and special laminates that economically combine the favorable properties of several polymers in one film.

Thermoforming is a purely stretching technique that is carried out as a continuous process. The article to be formed is secured at the edges, and the geometric increase in area is thus accompanied by a reduction in thickness.

Two types of industrial thermoforming machines have become established:

1) Single-station machines
2) Multistation machines

With *single-station machines* (Fig. 52) the heating and forming steps take place in the same station. The heating system is generally conveyed to and from the article; occasionally, however, the secured preform is transported to the heating zone. With multistation machines (Fig. 53), heating and forming take place in separate stations.

The technique of thermoforming polymers can be broken down into three basic process steps:

1) Heating
2) The transfer and forming process
3) Cooling

Heating, the first stage in all thermoforming techniques, is of decisive importance for the final quality of the molding. The process involves heating the sheet from its original temperature to the forming temperature.

Because films and sheets in thicknesses ranging from ca. 0.1 to 12 mm are thermoformed, a variety of preheating methods have evolved:

1) Convection heating
2) Contact heating
3) Heating with infrared radiation

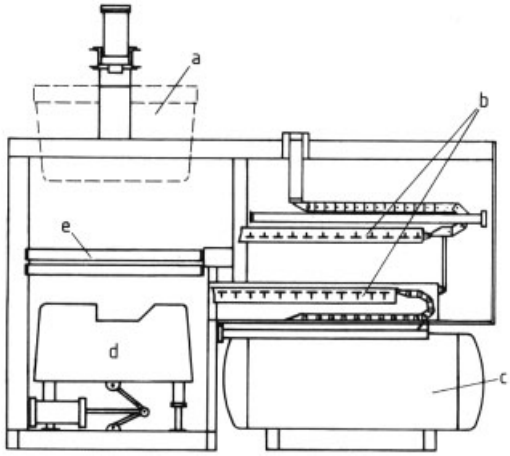

Figure 52. Single-station thermoforming machine
a) Top mold; b) Movable heaters; c) Vacuum tank; d) Bottom mold; e) Molding station

In all the techniques, heating can be either single or double sided. With the exception of extremely thin films, two-sided heating is generally employed.

Heating with infrared radiation is used most widely today. Its major advantage is that because a portion of the energy penetrates directly into the bulk of material, high heat flow densities can be used to decrease the heating time without causing thermal damage to the polymer surface. Ceramic and quartz elements are the primary sources of radiation. The heating stations in modern thermoforming equipment consist of numerous, individually controlled radiator elements in a modular structure. In all thermoforming processes the heated preform is rapidly drawn biaxially. This generally takes place in half a second and thus can be assumed to be isothermal as long as it is a free draw (i.e., without contact to the mold). On contact with the walls of the mold, however, the material cools immediately. Particularly with semicrystalline thermoplastics, which are formed at or slightly below the crystalline melting point, this leads to a spontaneous increase in the resistance to deformation. To exert the necessary force, mechanical drawing mechanisms, compressed air, and vacuum are utilized.

With thermoforming, a distinction is drawn between male and female forming processes

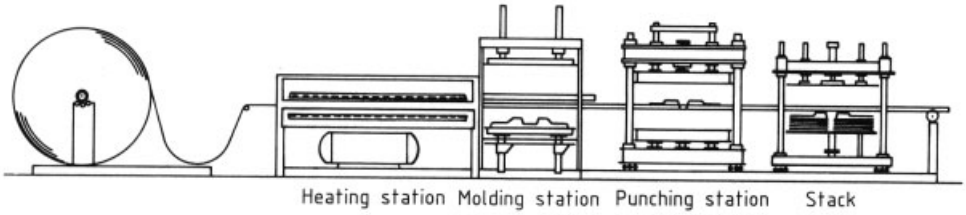

Figure 53. Multistation thermoforming machine

(Fig. 54). The decision to employ a male or a female process depends primarily on which side of the molding is to be replicated exactly. Only the surface in contact with the mold is dimensionally precise, whereas the other surface exhibits inherent inaccuracies, particularly in the unavoidable rounding of corners. The cooling process begins, by definition, once the film comes in contact with the mold and ends at the onset of the unmolding step.

For mass production purposes, the mold is equipped with temperature regulation. Mold temperatures vary from approximately 10 °C (PP packaging materials), to 70 °C [polystyrene (PS) inner panels for refrigerators], up to 100 °C [high-density polyethylene (HDPE) moldings]. Depending on the numbers of articles to be manufactured, the machines are fitted with single- or multicavity molds.

In most cases, one side of the material is cooled by contact with the mold, and the other surface is cooled by convection. Convective cooling occurs either via ambient air (free convection) or by use of a ventilator (forced convection). To intensify the cooling effect, atomized moisture is also added to the ventilated air (spray nozzle cooling).

The main advantages of thermoforming compared to alternative manufacturing techniques are the low mold costs, the short processing times, and when multicavity molds are used, the high throughputs. Additionally, reduced wall thicknesses are possible compared to injection molding. Thermoformed articles generally have about half the weight of equivalent injection-molded products.

In contrast to injection molding processes in which pressure acts only on the surface perpendicular to the flow direction during molding, in thermoforming the total mold surface is used. This is the reason thermoforming pressures are 20 – 30 times lower, and also the reason large moldings can be manufactured with lighter equipment, lower forming pressure, and lower energy consumption.

2. Processing of Thermosets

2.1. Vinyl Ester Resins, and Phenol – Formaldehyde Resins

In curable polymers, the liquid or meltable resins as received from the supplier are rendered solid, insoluble, and unmeltable during processing by addition of a curing agent. This reaction can occur independently or be initiated by heat or ultraviolet radiation.

This two-phase formation of the resin has the following advantages: a multitude of processing and design possibilities are available when starting materials that are liquid or that can be melted are employed. The resins can be dyed and processed with or without filters. By embedding reinforcing fibers, composite materials are obtained whose mechanical properties can be modified to suit the demands placed on the completed product simply by selecting the type, amount, and orientation of the fibers introduced, along with the processing technique employed. The properties of the finished article are governed not only by the choice of material, but also by the processing technology and the curing method.

Various processing techniques differ mainly in the complexity of the equipment used to provide the shape and to impregnate the reinforcing fibers. Customized moldings, small series, and bulky molding, as well as paneling and protective layers, can be manufactured manually with simple equipment. For large-scale operations, mechanized processes with a low manual cost contribution but correspondingly complex plants are available.

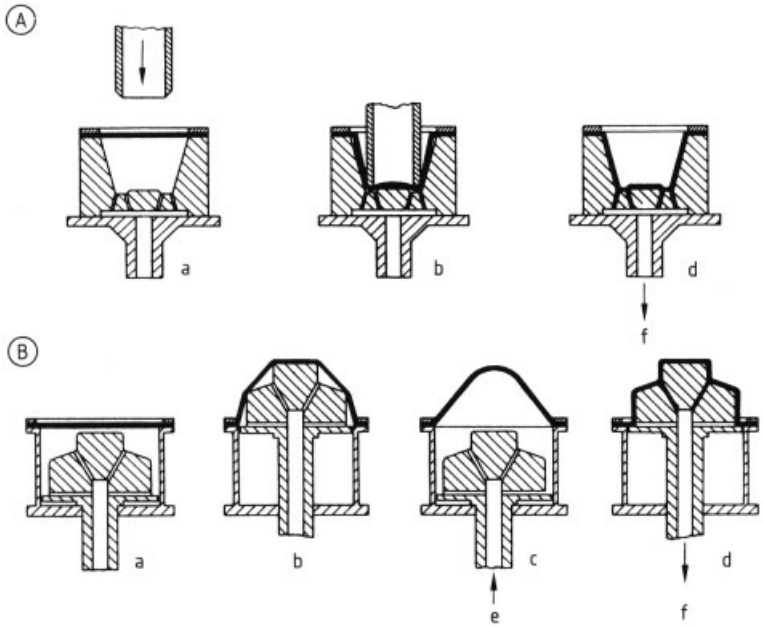

Figure 54. Vacuum forming
A) Female forming; B) Male forming
a) Original state; b) Mechanical prestretching; c) Pneumatic prestretching; d) End state; e) Compressed air; f) Vacuum

Table 1 gives an overview of the processing techniques for glass-fiber-reinforced composites. The use of more sophisticated processing machinery leads to improved reproducibility of properties, greater dimensional precision, and savings in terms of finishing operations. When selecting a process, one should consider the fact that the strength of the finished product depends strongly on fiber content. This is influenced, among other things, by the starting material and the processing pressure.

Table 2 lists a selection of mechanical characteristics of glass-fiber-reinforced unsaturated polyester (UP) resins with various types of reinforcing material and fiber content. Figure 55 shows the market development for glass-fiber-reinforced polymers from 1986 to 1988.

Table 3 lists the main areas of application, and Table 4 the processing techniques that are employed.

Table 3. European glass-fiber-reinforced plastics market in 1988 according to end use

End use	Consumption, %
Construction	14.2
Sports and leisure	7.3
Industry and agriculture	18.0
Transportation	25.8
Electricity and electronics	18.1
Consumer goods	8.6
Military and other uses	8.0

Table 4. European glass-fiber-reinforced plastics market in 1988 divided according to process

Process	Use, %
Manual lamination	20.4
Fiber spraying	7.4
Press molding	29.2
Granulation (thermoplastic)	23.9
Injection molding (thermosets)	2.7
Continuous lamination	6.8
Pultrusion	1.5
Filament and rotational processes	4.5
Others	3.6

Table 1. Applications and characteristics of different processing methods for fiber–plastic composites *

	Manual processing	Fiber spray-up process	Vacuum bag process	Resin transfer molding	Cold liquid resin press molding	Hot liquid resin press molding	Press molding of molding compounds	Filament winding	Rotation molding	Pultrusion	Continuous lamination	Autoclave process
Application	single pieces, small series, large-area pieces	small to medium series, large-area pieces, linings	small series, large composite material	medium series	medium series	large series	large series	wound articles	rotationally symmetrical hollow moldings	solid and hollow profile	smooth and curved plates	high loaded components small series
Resins used	UP, EP	UP	UP, EP	UP	UP, EP	UP, EP	UP, EP, PF	UP, EP, VE	UP, EP, VE	UP, (EP)	UP	EP
Suitable reinforcing materials (glass content, wt%)	mat (20–30) fabric (35–50) rovings (50–70)	chopped strands (20–30)	mat (20–35) fabric (40–60)	mat (20–25) fabric (40–60)	mat (25–40) preform (25–40) fabric (50–60)	mat (25–45) preform (25–45) fabric (50–60)	mat (20–40) chopped strands (20–40)	rovings fabric (40–70) mat	mat (20–35) chopped strands (20–35) fabric (30–40)	rovings fabric (50–70)	mat (20–25)	fabric
Labor costs	high	high	high	medium	medium to high	medium	low	low medium to high	low	low	low	high
Investment	low	medium	low	medium	medium to high	high	high	high	high	high	very high	medium
Molds	open, one or multipart mold made of wood, metal, glass-fiber-reinforced plastic	open, one or multipart mold, made of wood, metal, glass-fiber-reinforced plastic	one part or with rigid counter-mold, made of wood, glass-fiber-reinforced plastic	closed, split mold made of glass-fiber-reinforced plastic	closed, split mold made of glass-fiber-reinforced plastic	closed, split mold made of steel	closed, split or multipart mold made of steel	one or multipart, steel mold core	one or multipart, rotor made of steel, aluminum	profile mold, sometimes with steel core		open, one part with vacuum bag
Economically feasible molding size, m^2			0.5–30	0.5–20	0.3–5	0.3–5	0.05–3	0.1–100	1–100		width ≤5 m any length	
Cycle time	30 min to several days	30 min to several days	30 min to several hours	30 min to several hours	5–30 min	2–10 min	2–5 min	depends on molding size	10 min to several hours			several hours
Processing temperature, °C	room temperature	room temperature	room temperature	room temperature	20–60	80–120	120–160	room temperature	room temperature	≤160	room temperature and higher	80–200
Pressure, MPa			≤1	≤4	≤10	10–30	2–50	filament tension	centrifugal force			≤2

Table 1. (Continued)

Wall thickness, mm	2–10 usual	2–10 usual	1–5 usual	1–5	1.5–5	1–5	1–10	1–10	3–10	3–20	1–3	0.5–10
Wall thickness tolerances, %	mat ≤50 fabric ≤10	≤50 unavoidable	≤20	≤20	≤20	≤20	≤10	depends on reinforcement ≤20	≤20	≤10	≤10	≤5
Different wall thickness	possible	possible	limitedly possible	possible	barely possible	not possible	possible	possible	possible	usual in pulling direction	not possible	possible
Undercuts	possible with multipart mold	possible with multipart mold	limitedly possible	possible with multipart mold	not possible	not possible	possible	not possible	not possible	perpendicular to the pulling direction possible	perpendicular to the pulling direction possible	possible with multipart mold
Stiffening layers	possible	possible	possible	possible	limitedly possible	limitedly possible	possible	possible	not possible	possible	not possible	not usual
Gel coat	usual one side	usual one side	one-sided; with rigid countermold both sides possible	both sides possible	limitedly possible	not possible	not possible	possible	possible	not possible	possible	not usual
Surface	one side smooth	one side smooth	one side smooth; with rigid countermold both sides smooth	both sides smooth possible	both sides smooth, minor fiber structure	both sides smooth, minor fiber structure	both sides smooth	one side smooth	both sides smooth	all sides smooth	both sides smooth	one side smooth
Aftertreatment	edge trimming	edge trimming	edge trimming	deflashing, edge trimming	edge trimming	deflashing	deflashing	separate	separate	separate	separate, edge trimming	edge trimming

* UP = unsaturated polyester; EP = ethylene–propylene; PF = phenol–formaldehyde; VE = vinyl ester.

Table 2. Dependence of the properties of glass-fiber-reinforced UP moldings on reinforcing material and glass content

Property	Glass mat		Woven glass fabric		Glass rovings
Glass content					
wt %	20 – 30	40 – 50	40 – 50	50 – 60	70 – 80
vol %	11 – 17	25 – 32	25 – 32	32 – 42	53 – 66
Density, g/cm^3	1.3 – 1.5	1.5 – 1.75	1.5 – 1.75	1.6 – 1.85	1.9 – 2.1
Thermal expansion coefficient, 10^{-6} K^{-1}	40 – 30	24 – 20	22 – 18	18 – 16	14 – 12
Thermal conductivity, W K^{-1} m^{-1}	0.14 – 0.19	0.23 – 0.31	0.19 – 0.25	0.35 – 0.31	0.37 – 0.41
Tensile strength, N/mm^2	65 – 90	130 – 170	200 – 240	240 – 275	700 – 800
Breaking elongation, %	2	2	2	2	2
Modulus of elasticity (tensile), N/mm^2	5000 – 7000	9000 – 10 000	10 000 – 14 000	14 000 – 17 500	21 000 – 26 000
Bending strength, N/mm^2	115 – 145	180 – 220	220 – 260	260 – 300	400 – 500
Modulus of elasticity (bending), N/mm^2	5000 – 7000	9000 – 11 000	10 000 – 14 000	14 000 – 17 500	
Compressive strength, N/mm^2	110 – 135	165 – 200	150 – 180	180 – 200	
Shear modulus, N/mm^2	2000 – 3000	4000 – 5000	2000 – 3000	3000 – 4000	
Impact strength, kJ/m^2	25 – 55	75 – 105	130 – 160	160 – 190	

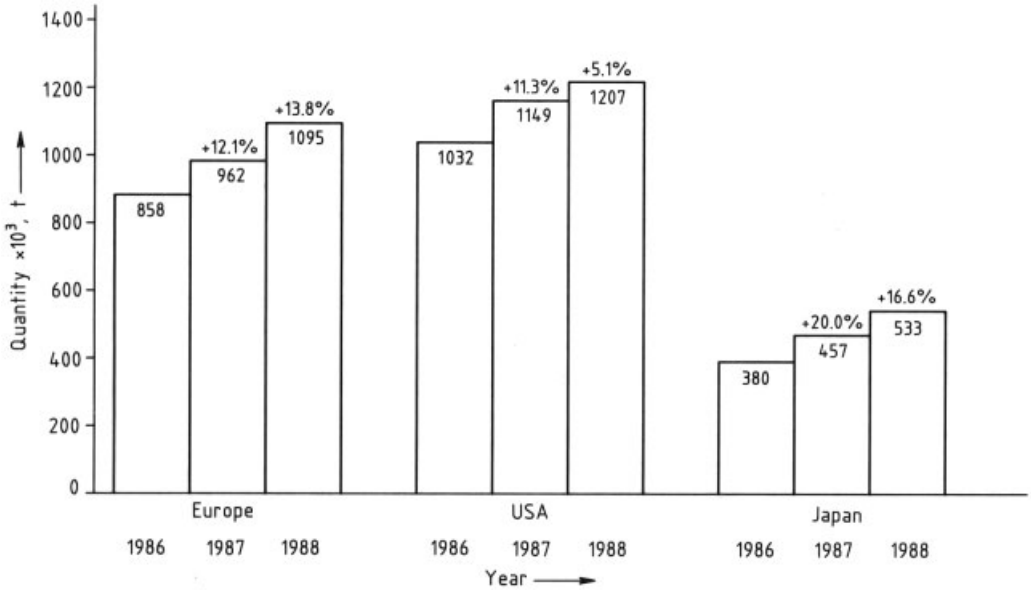

Figure 55. Development of glass-fiber-reinforced plastics market

2.1.1. Raw Materials

Selection. The most widely used reinforcing fibers are glass fibers (→ Fibers, 5. Synthetic Inorganic, Chap. 2.1.) in the form of rovings, nonwoven fiber matting, and glass fabrics. Carbon and aramid fibers, which were initially applied in aviation and space, are increasingly being used in industrial applications. Their significantly higher price compared with glass fibers is often justified by their much higher stiffness and strength and their more favorable thermal expansion behavior. All reinforcing fibers must be coated with a coupling agent to achieve satisfactory adhesion to the polymer matrix. Additionally, the coupling agents protect the fibers from abrasion during processing.

The resins are selected primarily for their final properties, corresponding to the demands placed on the finished product. With few exceptions, unsaturated polyester (UP), epoxy (EP), and vinyl ester (VE) resins can be processed by using all the techniques described below, although differing processing properties must be considered in selecting the resin. This applies particularly for phenol – formaldehyde (PF) resins, which cure via a condensation re-

action and hence can be processed only under pressure.

The advantages of *EP resins* are air-drying cure; high filler levels possible; wide variation due to different combinations of resin and hardener; low volume shrinkage during curing (2 – 3 %), which begins before gelation and hence results in extremely good retention of dimensions and freedom from stress in the end product; no deformation during curing; high adhesive strength (i.e., good adhesion to the substrate and the fibers) and hence high mechanical strength of fiber-reinforced components, especially during dynamic loading.

In the latter case, the favorable mechanical properties of EP resins, including high fracture strain and tensile strength, also play an important role. High application temperatures may necessitate the use of EP resins. Certain types permit long-term service temperatures exceeding 200 °C.

Disadvantages of EP resins include high price; long curing times and hence higher processing costs; extreme care needed to meter resin and hardener; high viscosity, which makes degassing the resin and impregnating glass fibers difficult; sometimes difficulty in demolding as a result of low shrinkage.

The advantages of *UP resins* are low price; low curing times (i.e., rapid demolding is possible); easy handling due to low viscosity; low susceptibility to erroneous mixing of the final formulation.

Disadvantages of UP resins are higher volume shrinkage on curing (6 – 8 %), which makes it difficult to manufacture stress-free components and leads to a tendency to deform. In addition, during selection of the type of resin, attention must be paid to the reactivity. Low-reactivity resins cure with relatively low stresses, and the end products have good elasticity but generally exhibit poor resistance to corrosion and aging. High-reactivity products can be cured easily without heating, but removal of the heat of reaction must be effective to minimize stresses in the end product, shrinkage, and separation of the resin from the fibers, as well as possible thermal damage to the material itself. The long-term service temperature of components made from UP and VE resins is limited to 70 – 120 °C.

The advantages of *VE resins* compared to the chemically related UP resins include improved mechanical properties and, particularly, greater chemical stability, especially toward acidic media. Their disadvantage is the high price.

The PF resins represent one of the oldest and cheapest classes of polymer. Their advantages include very low manufacturing costs, excellent chemical and thermal stability, favorable combustion behavior, and low toxicity of combustion products. This has led to considerable importance in applications where safety is a prime consideration (aircraft inner paneling, transportation). The drawbacks of PF resins are poor mechanical properties (low fracture strain) along with problematic processing. Because the latter can be carried out only under pressure, techniques involving open molds cannot be employed. Only processes involving presses and autoclaves are of any importance.

Because of their high price, EP resins are employed where higher material costs are justified by improved adhesion to the substrate or improved mechanical properties (e.g., for surface protection and coatings).

Because they are easy to process, UP resins are employed preferentially for glass-fiber-reinforced products, with the exception of components that must withstand high dynamic loads, such as helicopter rotor blades; components for aircraft, rockets, and spacecraft; and skis.

Where corrosion problems are likely to be encountered, resins are selected according to their resistance to chemical attack.

Transportation, Storage, and Safety. The inflammability of resin systems depends on their chemical composition. During storage, transport, and processing, the regulations and safety information provided by the supplier must be followed.

Reactive resins must be stored below 20 °C, with the exclusion of atmospheric moisture and light. Resins supplied as a final mixture ready for direct reaction and including the curing agent – as is the case, for example, with preimpregnated fiber materials known as prepregs – must be stored under refrigerated conditions at −18 °C.

During handling and processing of UP and VE resins, the skin and respiratory irritations that can result from the styrene component must be considered (MAK 20 ppm), along with the strong skin irritations that can be caused by sev-

eral cross-linking agents. Particular care is necessary when dealing with peroxides, which are often used as hardeners (explosion hazard).

During the processing of EP and PF resins, safety gloves, clothes and glasses must be worn.

Several cross-linking agents for EP resins (e.g., amines) are strong skin irritants or can lead to allergic reactions. With PF resins, adequate ventilation must be ensured to remove reaction products.

2.1.2. Mold Materials

The choice of mold material depends on the processing technique and the number of articles to be manufactured. For casting, molds can be manufactured from flexible materials such as plasticized PVC or silicone rubber. These allow the production of moldings with undercuts. The number of castings is limited by the low solvent resistance.

Wood and plaster are used for making molding patterns; wood is also employed in manufacturing single-component molds for individual moldings or small production runs using pressureless processing techniques. The surface must be sealed with a paint that is resistant to aromatic compounds.

For the manufacture of glass-fiber-reinforced moldings in pressureless processes and the production of small series by cold press molding, a processor generally prepares his own molds from glass-fiber-reinforced UP or EP. The EP resins have the advantage of greater dimensional stability, but they must be coated to protect them from aromatic compounds. By casting a supporting substrate of polymer concrete or concrete on reverse surface, the molds can be made rigid.

For large production runs, metallic molds are most suitable. For hot press molding, heated molds made of tempered steel are employed. The sealing edges of molds for liquid resin press molding are formed as nip, shear, or pinch ridges. Molds for processing molding compounds and sheet molding compounds are described in Section 2.2.5.

To ensure faultless demolding, mold surfaces are treated with a release agent [wax, silicone grease, poly(vinyl alcohol)]. Prior to application of the release agent, the remains of any previous molding processes must be removed thoroughly.

2.1.3. Preparation of Resins

The resin components are either premixed by the supplier (e.g., prepregs) or metered and mixed by the processor. This is performed by automated mixers fitted directly to the processing machine (e.g., injection techniques such as resin transfer molding) or by hand mixing.

In all cases, the equipment (mixing vessels, stirrers, metering units, etc.) must be clean, dry, and free of residual deposits from previous batches. All components, including glass fibers must be thoroughly dried and, if necessary, warmed to room temperature before use. Mixing is best performed with a high-speed stirrer. The speed should be adjustable to avoid the inclusion of air bubbles. Highly viscous batches (e.g., EP resins) are prepared and held under vacuum until completely degassed.

The components are added in the following order: resin; styrene (with UP resins); plasticizer and diluent (with EP resins); hardener; additives such as stabilizers, lubricants, dyes, film formers, conductivity additives, thickeners, thixotropic agents, flame retardants, fillers, and accelerators. For properties and applications of additives, see → Plastics, Additives.

Materials that are difficult to disperse or additives that are not divided finely enough are made into a slurry before mixing by using small amounts of UP resin or EP thinner.

With UP resins, the curing reaction begins with addition of the second reactive component; initially the resin gels and then solidifies rapidly. Processing of the resin mixture must therefore be completed within the narrowly limited pot life dictated by formulation and processing temperature (i.e., before the onset of gelation). With resins that already contain an accelerator, the hardener is therefore always added as the final component. To avoid exceeding the pot life of UP resins prior to processing, the following procedure can be followed: all of the additives are mixed into the resin and the resulting batch is divided into two equal volumes. The hardener is added to one half, and the accelerator to the other. Immediately before processing, equal volumes of the two batches are mixed together.

With EP resins, the rate of hardening decreases with increasing viscosity. The pot life

therefore does not end so sharply with the onset of gelation as with UP resins.

2.1.4. Pressureless Processing without Fiber Reinforcement

Casting and Embedding. Because of their light colors and low viscosity, UP resins are particularly suitable for casting figures and embedding specimens. The resin mixture can be adequately degassed by allowing it to stand for a short period prior to casting. For embedding two or more layers are usually cast. After the first layer has begun to cure, the item being embedded is laid on the surface and covered with further resin mixture. Because of shrinkage during curing, the casting must generally be finished by grinding and polishing.

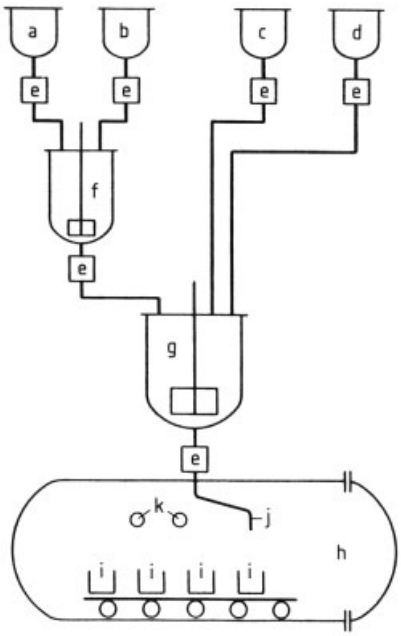

Figure 56. Schematic of a casting plant for epoxy resins
a) Heated storage tank for resin; b) Storage tank for fillers; c) Heated storage tank for hardener; d) Storage tank for other additives; e) Metering units; f) Heated, evacuable mixer; g) Heated and evacuated mixer for final mixing; h) Evacuable casting vessel; i) Casting mold; j) Swivel arm; k) Sight glasses

EP resins are preferred for casting and embedding technical articles, where the slightly darker color is unimportant but high demands are placed on dimensional stability (e.g., for models) as well as embedding of electronic components. Figure 56 shows a schematic of a casting plant for EP resins. Resin and hardener are firstly melted or heated to lower the viscosity. Fillers and additives are then added. To degas the viscous resin mixture completely, mixing and casting must generally be done under vacuum.

When embedding electronic components, penetration of the resin mixture into joints and voids can be promoted by removing the vacuum once the casting stage has been completed.

In dipping processes, components are immersed in the resin mixture while under vacuum and exposed to atmospheric pressure after their removal.

In the drip process, the cold resin mixture drips onto a rotating, preheated substrate. During this process, the viscosity decreases and the resin penetrates any voids. This process is used among other things to bond axially parallel windings of wound articles (see Section 2.1.8).

Coating. Another important area for which EP resins are particularly well suited due to their excellent adhesion properties is in surface coatings. Coating is carried out with a liquid, low-viscosity resin – hardener system or with solid components dissolved in a solvent. Resin powders can be applied by fluidized-bed sintering, flame spraying, or electrostatic coating.

These processes and the continuous coating of printed paper or fabric webs with UP resins are described in Section 1.3.

Polymer Concrete. Both EP and UP resins are used in the manufacture of polymer plaster, mortar, and concrete. Because of their extremely low shrinkage and low viscosity, resins based on poly(methyl methacrylate) are also employed. These resins replace the conventional cement and water binder used for concrete. Gravel, pebbles, sand, and ground stone are used as additives. Polymer concretes have much shorter setting times and can be subjected to loading after about one day. Preparation and processing correspond with the techniques already in use with cement-based concrete.

The mechanical strength of polymer concrete is higher than that of cement concrete, and polymer concrete adheres well to set concrete. It is thus widely used to cast joints, for repair pur-

poses, and above all as a coating for concrete surfaces when increased resistance toward acid, alkali, and salt solutions is required.

Another important application is the manufacture of bases for machines or machine housings. The advantage over the more common gray cast iron or cast alloy construction is the vibrational damping exhibited by polymer concrete.

The choice of resin depends mainly on the required thermal stability and chemical resistance, as well as the maximum allowable shrinkage.

For outdoor use, the sensitivity of the hardening reaction to water (soil moisture, rain) and temperature must be considered.

Foamed polymer concrete with low density, good mechanical properties, and low thermal conductivity can be produced by adding blowing agents to the resin mixture.

2.1.5. Pressureless Processing with Glass-Fiber Reinforcement

Gel Coating. The appearance of moldings made from glass-fiber-reinforced plastics (GRP) can be impaired by the presence of glass-fiber on the surface. Gel coatings are used to cover the glass fibers, giving a smoother surface and improving the resistance to weathering and chemicals. Such layers adhere better to the composite surface than subsequently applied paint coatings because they are anchored by chemical bonds. The gel coating process can be combined with most processing techniques.

To produce a gel coating, a layer of fiber-free resin is sprayed or painted onto the surface of the mold in a thickness of 300 – 600 µm. The coating can be dyed or filled to improve abrasion resistance. Before the following composite layer can be applied, the gel coating must be cured sufficiently so it is not swelled by styrene or solvents and cannot be penetrated by glass fibers during subsequent processing stages. The curing reaction must, however, not be so complete that bonding to the composite layer is reduced.

Manual Production of GRP Components. The manual process (Fig. 57) is the simplest method of producing GRP moldings and requires only basic equipment. It is also the oldest processing technique. Production of GRP moldings began in the 1940s in the United States and, despite the development of machine processes, has retained a considerable portion of the market. The manual process is especially suitable for manufacturing large-area moldings; for small series using single-part, light-construction molds; and for reinforcing components made of thermoplastics. The moldings can be designed freely and are not subject to constraints regarding size.

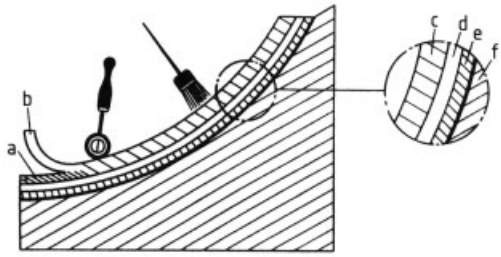

Figure 57. Manual production of GRP moldings
a) Resin; b) Fiber reinforcement; c) Laminate layer; d) Gel coat; e) Release agent; f) Mold

A gel coat is first applied to the mold whose surfaces have been coated with a release agent. After initial curing, the first resin layer is applied and a layer of glass fibers, nonwoven glass-fiber matting, or glass fabric is applied by using a brush or fleece roll and a structured roll. Any air bubbles that form between the fine layer and the laminate must be removed. This primary composite layer is first cured slightly to prevent the ingress of air bubbles and movement of the layer during further processing. After this, additional composite layers are quickly applied wet-in-wet. The number of reinforcing layers depends on the strength required of the finished molding. The embedding of subcomponents made from other materials, such as stiffening ribs and screw thread sockets, is also possible. If required, a finishing surface coating, similar to the gel coating can be applied to cover the glass fibers on the surface. A film former (paraffin dissolved in styrene) is usually added to the topcoat when UP resins are used, to protect the surface of the composite from atmospheric oxygen. This also gives a tack-free surface even on curing at room temperature.

Before the molding is removed the curing reaction must be so far advanced that the article is not damaged by tearing or by separation of

the resin and fibers (stress whitening). Subsequently, the product is stored in simple support equipment until curing is complete to avoid deformation, because upon hardening, all deformations become fixed. Moldings that are manufactured in single-part molds have a smooth surface only on the side that was in contact with the mold.

Attainable glass contents with glass matting range between 20 and 30 wt %, and with glass fabrics between 40 and 50 wt %. Suitable matrix materials include UP, VE, and EP resins.

Fiber Spray-up Process. The fiber spray-up process is a partially mechanized manual process. It is particularly useful for linings and in the manufacture of complex moldings. Only UP resins are used in this process.

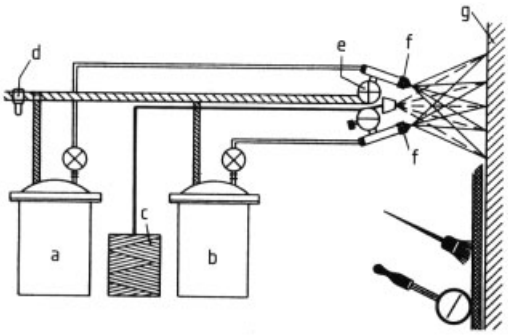

Figure 58. Fiber spray-up process
a) Resin with hardener; b) Resin with accelerator; c) Roving reel; d) Compressed air inlet; e) Fiber cutting device; f) Spray gun; g) Mold

After application of a gel coating, which for linings also serves to seal any pores, the resin mixture and chopped glass fibers are sprayed simultaneously onto the mold and then, in a manner similar to the fully manual process, are compressed with a roller and degassed with a structured roll. The fiber spray-up equipment illustrated in Figure 58 operates with two resin feeds, one of which contains the hardener, and the other the accelerator. The two components are fed to a double pistol and sprayed with compressed air. They are mixed immediately prior to contact with the mold. The glass rovings are fed to a chopping unit located between the two nozzles, where they are chopped to staple fibers and accelerated along with the two streams of resin onto the mold. With high-pressure spray equipment, accelerator-containing resin and hardener are pumped from storage tanks to the spray pistol via a mixing and metering unit.

With the fiber spray-up process, the layering of glass fibers and the glass content are comparable to those of manual processes. Maintaining constant wall thickness, however, requires skill and experience in spraying, compacting, and degassing the fibers, which initially lie very loosely. An advantage over the manual process is that large batches of resin with prolonged pot life can be processed, while very rapid curing times are also tolerated. Furthermore, the labor-intensive cutting of the glass matting to size is avoided, along with losses incurred during trimming. Start-up and cleaning procedures for fiber spray equipment require considerable time. Usage is thus recommended only when extended production times without long interruptions can be expected.

2.1.6. Low-Pressure Processes

Low-pressure processes allow the production of articles of large surface area in small to medium production runs. Manufacturing times are usually shorter than with manual processes but longer than with compression molding. Low-pressure processes can be used to improve laminates manufactured with pressureless techniques (vacuum film process) and permit direct impregnation of fiber materials in a closed mold.

Vacuum bag process:

1) With single-component rigid mold and elastic sheet as countermold (Fig. 59): The resin mixture is applied to the mold; the glass-fiber reinforcement is added and, after an airtight seal is generated, impregnated with the resin by applying a vacuum.
2) Molding with two elastic sheets: A laminate that has been impregnated and degassed by hand is deep drawn by vacuum between two elastic films. This technique is suitable for the manufacture of domes.
3) Form and counter form are rigid: After the form has been coated with resin mixture, the fiber reinforcement and, when appropriate, the core material are applied, and the counterform, which is also coated with resin mixture,

is used to form an airtight seal. The necessary compression is generated by applying a vacuum. Two-sided topcoats are possible. This process is suitable for items of large surface area and is used primarily to coat plywood sheets and rigid foam cores for refrigerated trucks and containers.

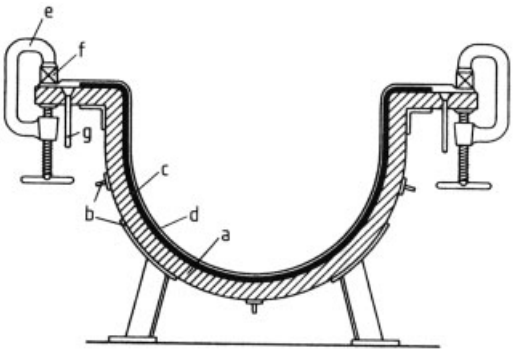

Figure 59. Vacuum bag molding
a) Mold; b) Reinforcing profiles; c) Laminate; d) Flexible sheet as countermold; e) Clamp; f) Sealing frame; g) Suction vents

Autoclave Processes. The low-pressure autoclave process is a variation of the vacuum bag technique. It is used mostly for the manufacture of high-quality components for the aviation and aerospace industries. The costs are high due to the use of a heated, pressurized autoclave and high-quality prepregs. Furthermore, the structure of the laminate is complex, containing release films and fabrics, suction, pressure distribution, and surface layers (Fig. 60). Cycle times can be several hours.

Resin Transfer Molding. In the resin transfer molding (RTM) process (Fig. 61), two-component or multicomponent rigid molds made from metal or GRP are employed, which are opened and closed by hydraulic mechanisms. The fiber reinforcement is cut to shape and placed in the mold. The mold is closed, and the resin mixture is injected from the bottom under pressure. Air release valves allow control of the resin flow. Application of a vacuum assists the impregnation stage. Large-surface moldings with topcoat matts on both surfaces can be manufactured with this technique.

Apart from moldings made from preshaped glass-fiber matting and UP resins (fiber content 20 – 40 vol %), which can withstand moderate loadings, the RTM process is equally suited for producing high-quality components from fiber fabrics and EP resins (fiber content up to 65 vol %). The injected resin must be of low viscosity and highly reactive. However, for modern resins, these requirements are associated with low softening temperatures and hence low use temperatures of the final molding.

2.1.7. Press Molding

Press molding allows the production of medium to large series of GRP moldings with smooth surfaces on both sides, good dimensional stability, and high glass content. The presses are generally hydraulic; two- or multi-part molds are used. Labor costs are considerably lower than with pressureless processing.

Liquid Resin Press Molding. In liquid resin press molding the glass-fiber reinforcement is placed in the mold, and the resin mixture is poured into the center so that the flow paths to all sides of the mold are equally long. When the press is closed, the resin mixture is distributed evenly throughout the reinforcing material, and any air remaining in the mold is forced out of the cavity by the flow front (Fig. 62). The mold closing speed must be reduced significantly over the final 5 – 10 mm of the closing movement to avoid shifting or tearing of the fiber insert due to rapid flow of the resin and to prevent entrapment of air pockets.

Generally, glass matting is used as fiber reinforcement; glass fabrics or preforms are also used. In the production of preforms, glass fibers cut to a length of 20 – 50 mm are drawn onto a punched screen that exactly matches the shape of the item to be pressed. The fibers are bonded by spraying with a binder; after the binder has been dried, the preform can be processed. With moldings containing severe bends this process avoids the complex cutting and laying of glass matting, and overlap areas are avoided.

Topcoats can also be incorporated with liquid resin molding. However, this requires an extendable press table with two interchangeable lower mold halves to achieve acceptable cycle

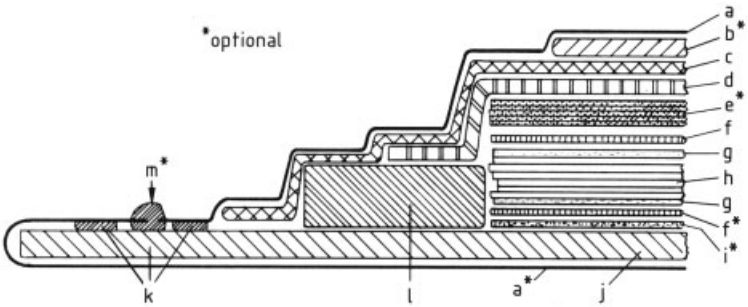

Figure 60. Structure of a laminate produced in the autoclave process
a) Vacuum bag; b) Compression plate; c) Pressure-distribution layer; d) Release film, perforated; e) Absorbent material; f) Release film or fabric; g) Peeling film; h) Fiber composite laminate; i) Release film; j) Mold; k) Sealing compound; l) Mold boundary; m) Double gasket

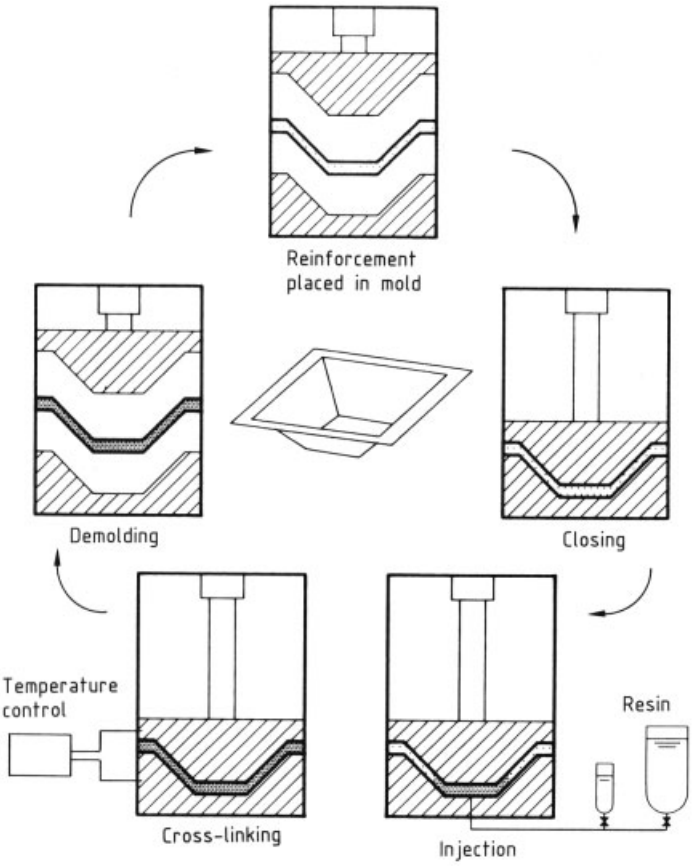

Figure 61. Resin transfer molding process

times. To obtain smooth surfaces a polymer or glass-fiber nonwoven fabric is best used as the first and last fiber layer. After being demolded, the pressed molding must be stored on support equipment to prevent distortions.

Generally UP resins are used; EP resins are less suited for liquid resin molding because of their long curing times. Filler contents up to 40–60 parts per 100 parts resin improve the surface quality by suppressing the fiber structure. Articles with varying wall thickness cannot be manu-

factured with wet pressing techniques because a constant glass content cannot be achieved at the junctions. The use of ejectors to assist in removing the molding is not recommended because of the possibility of malfunction caused by ingressing resin. They are usually not necessary because the mold must be slightly conical to avoid shifting or tearing of the glass-fiber reinforcement on closing.

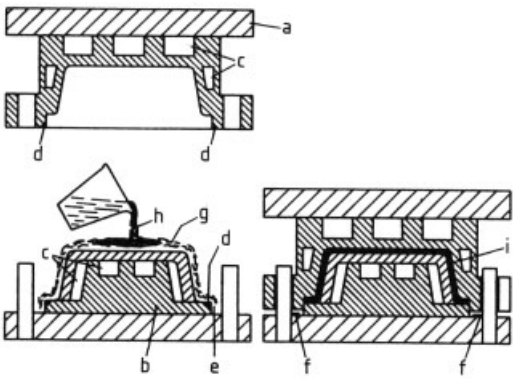

Figure 62. Liquid resin press molding
a) Female mold; b) Male mold; c) Heating channels; d) Cutting edges; e) Overflow; f) Excess; g) Fiber reinforcement; h) Resin; i) Molding

Cold liquid resin press molding uses highly reactive UP resins containing hardener and accelerator. Pressing temperatures range from 20 to 60 °C. The amount of resin mixture is adjusted so that the resin begins to gel a few seconds after the mold closes. The mold surfaces are warmed by the heat of reaction. Additional heating of the mold is thus not necessary but does help to maintain a constant cycle time, particularly for the initial six to eight press cycles at the beginning of a shift or following a machine stoppage until the molds have reached their final temperature. Typical press pressures range from 0.1 to 1 MPa. When textile glass matting or preforms are used, glass contents of 25 – 35 wt % can be achieved.

Low pressures permit the use of molds made of GRP, supported by a polymer concrete or concrete backfilling for rigidity (Fig. 63). These molds are considerably cheaper than steel molds. Their lifetime is shorter; however, they are well suited for small- to medium-sized runs in which the required cycle times of 5 – 15 min are acceptable. The mold seal forms a nipping edge. By reducing the pressing layer thickness around this nipping edge to approximately one-third of the original thickness, the glass fibers are so compressed that they prevent leakage of the resin and permit the necessary pressure buildup.

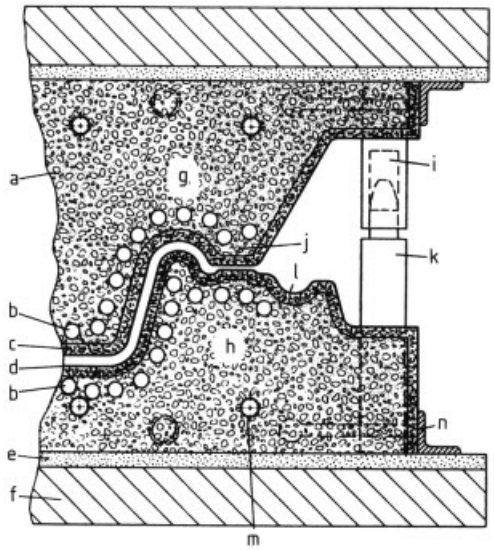

Figure 63. Compression mold made from glass-fiber-reinforced plastic with pinch-off edges for cold press molding
a) Backfilling of polymer concrete or cement concrete; b) Heating tube; c) Mold surface of glass-fiber-reinforced plastic; d) Gel coat; e) Rubber plate; f) Press table; g) Male mold; h) Female mold; i) Guide sleeve; j) Pinch-off edge; k) Guide pin; l) Overflow channel; m) Reinforcement of back filling; n) Frame for fixing mold

Hot liquid resin press molding is suitable for the mass production of moldings. The presses and molds are described in Section 1.4.1. When processing UP resins, the cavity plate and core plate should be heated individually; temperature differences of 5 – 10 °C assist demolding, and the quality of the surface in contact with the warmer plate is improved. Liquid heating of the molds is preferable to electric heating because it allows removal of the heat of reaction. Press pressures of 0.5 – 5 MPa and temperatures of 80 – 120 °C are used.

In hot press molding, curing is initiated by external heat. UP resins can thus be used without the addition of accelerators. The pot life of the resin mixture ranges from several days to

weeks at room temperature. The required press time is ca. 1 min per millimeter of wall thickness. Since the curing reaction begins shortly after the resin comes in contact with the mold surface, the filling and closing phases must be carried out rapidly. The mold seals form shear or immersion edges (Fig. 64).

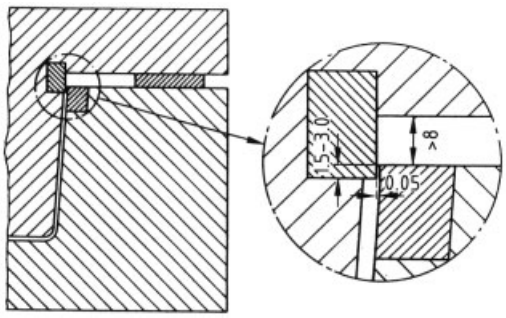

Figure 64. Shear edges on a steel compression mold

Press Molding of Preimpregnated Fiber Articles. Preimpregnated textile fibers (prepregs) contain all the necessary components for manufacturing a pressed molding. The mixing of resins and the combining of resins with reinforcing material, which are necessary with wet pressing, are avoided during such processing. A distinction is made among preimpregnated fabrics, preimpregnated rovings, sheet molding compounds (DIN 16913), and fiber-containing molding compounds (DIN 16911 and 16912). The processor can prepare these molding mixtures or purchase them ready for use. See Section 1.4.1 for processing molds and presses.

Prepregs in the form of fabrics or parallel glass, carbon, or aramid fibers are manufactured by impregnating the fibers with epoxy or phenolic resin. The resin is essentially solid, but still reactive and slightly tacky, and must be dissolved in a solvent or melted to penetrate the fibers completely. The resulting prepreg must be cut to fit the mold exactly because it cannot flow. Prepregs are employed mainly for highly loaded, thin-walled components (e.g., in aircraft). Besides press molding, processing with autoclave techniques is also common (see Section 2.1.6).

Sheet molding compounds are manufactured exclusively from UP resins. The resin mixture is composed of approximately equal weights of resin and filler (chalk, talc, clay), sometimes with addition of a lubricant. Curing agents with high initiation temperatures are used to give long pot lives at room temperature. To increase viscosity, thickening agents such as magnesium oxide are added, which cause a sudden thickening of the liquid to give a dry, leathery consistency. The thickening process takes from several hours to several days.

Impregnation is carried out immediately after the resin is mixed. Either glass is treated on equipment operating in a continuous process, or the resin mixture is applied on a film, cut glass fibers are added, and another film coated with resin is applied as a cover. After processing the fiber – resin mixture between several pairs of rolls, the sheet including the protective films is wound in a roll (Fig. 65). Because the thickening reaction continues during storage, the pot life is limited.

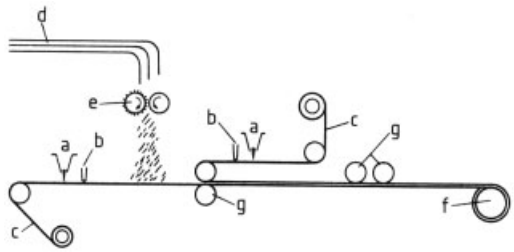

Figure 65. Production of sheet molding compounds
a) Feed tank for resin; b) Doctor blade; c) Cover sheet; d) Roving strands; e) Fiber cutting device; f) Winding of sheet molding compound; g) Impregnation of fibers with resin

When working with sheet molding compounds preparation of the resin mixture is no longer necessary, and processing does not require the use of exactly cut sections. Because of the high viscosity, all components flow at the same rate under the molding pressure at the mold temperature. Thus, moldings with varying wall thickness can be manufactured, although in such cases, higher pressures are required than with liquid resin press molding.

To avoid the surface subsidence of areas with thick walls caused by the volume shrinkage of UP resins, low-shrinkage resin systems [low-profile (LP) resins] have been developed. These contain solutions of thermoplastics dissolved in

styrene. On curing, part of the styrene contained in the thermoplastic particles evaporates prior to polymerization and causes foaming of the resin mixture, which in turn compensates for shrinkage.

Fiber-containing molding compounds are obtained by kneading chopped fibers into the resin mixture or by cutting continuous, impregnated roving strands to the desired length. These molding compounds can also be processed by injection press molding and injection molding (see Section 1.4.5). However, abrasion of metallic components by the hard glass fibers is a problem.

2.1.8. Semicontinuous Processing Techniques

Filament winding is a highly reproducible process for the manufacture of hollow articles with high mechanical strength. The equipment and the manufacturing technology are complex. Resins are selected according to their mechanical, thermal, and chemical loading; UP, EP, and VE resins are generally used. The fiber content of the product depends on the laminate structure, the process parameters, and the viscosity of the resin. Normal values range between 50 and 60 %.

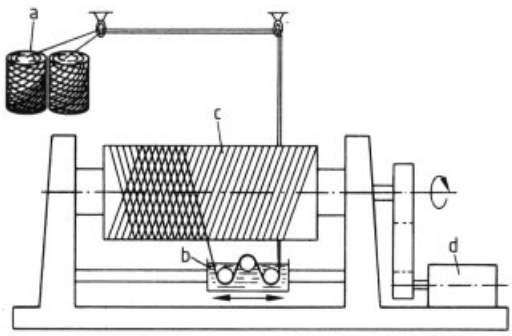

Figure 66. Filament winding (lathe system)
a) Roving reels; b) Support with impregnating bath and fiber guide; c) Winding mandrel with laminate; d) Drive

The so-called lathe system is used for winding containers, container jackets, and pipes (Fig. 66). The winding equipment consists of a rotating winding mandrel and an oscillating fiber guide, the drives of which can both be regulated. Rovings are generally used for reinforcement, although woven tapes can also be employed. The rovings are fed through a resin impregnation bath, which is frequently combined with the fiber guide system to form a single unit. The impregnated fibers are then wound onto the core under constant tension and with an exact geometrical pattern governed by the ratio of the rotation speed of the core to the feeding speed of the fiber guide.

Modern winding equipment has up to six axes or degrees of freedom for the fiber guides and is usually computer controlled. This control system calculates data for the motion of individual axes, corresponding to the geometry of the winding mandrel and the desired winding pattern.

In parallel winding, individual roving strands are wound side by side without a gap, at a winding angle of almost 90°. To obtain sufficient strength in the axial direction, either warp-knitted fabrics are incorporated between the individually wound layers, or dry or impregnated glass fibers are sprayed between the winding layers.

With diagonal or cross winding, a winding angle of 30 – 70° is employed. The strength of the wound article can be adapted to the expected loading in the axial and radial directions by varying the winding angle.

Depending on the size of the molding, cold or hot curing resin systems are used for winding. Smaller moldings a few meters in length, which fit into hardening ovens or whose mandrels can be heated, are cured thermally to obtain superior material properties and extended resin pot lives. In contrast, large containers and pipes – diameters up to 12 m can be manufactured at present – must be cured in the cold or by using ultraviolet radiation.

Because the wound molding shrinks on the mandrel during curing, retractable winding mandrels are used to facilitate removal. Container bases, flanges, and supports are manufactured batchwise and are bonded to the container jacket or pipe by adhesion or lamination.

Feed silos; containers for drinks, chemicals, and heating oil; and pipes for the chemical industry (e.g., pipes for acid, wastewater, and exhaust air) are manufactured via winding processes. Mechanically loaded components including drive shafts and centrifuge drums can also be wound.

For the manufacture of highly stressed pressure vessels for the aerospace industry, equip-

ment is used in which the winding mandrel rotates about two axes, or the fiber feed and the mandrel both rotate. Single fibers are used and wound in complex patterns.

A continuously operating winder has been designed for the mass production of pipes or container jackets in which an endless steel band is employed as the winding mandrel. This band runs spirally around a support using transport rolls, and is unwound at the outlet of machine and fed back through the inside of the core. The continually wound laminate comprising radially oriented rovings and axial fabric or matting reinforcement is cured by passage through a heated section after winding.

Rotation Molding. Cylindrical hollow articles with smooth internal and external surfaces can be manufactured by rotation molding. Glass matting or chopped glass fibers, sometimes in combination with textile fabrics, are employed as reinforcing materials. The choice of resin depends on the anticipated exposure to corrosive influences. The rotation equipment consists of a balanced cylindrical steel drum of adjustable rotation speed. With small-diameter pipes, the reinforcing matting or fabric is rolled in the stationary drum. On rotating, the strands are unrolled and pressed smoothly against the wall of the drum. The resin mixture is then distributed evenly across the complete length by use of a metering device. Due to the influence of centrifugal force, the reinforcing material is pressed against the wall and the resin mixture impregnates the fiber layers and displaces the air. Finishing coats can also be applied in this way. Pipes with smooth inner surfaces can be obtained by using a slight excess of resin. Generally, cold curing resin systems are employed. However, heating of the drum is advantageous, particularly with EP resins.

Larger vessels with diameters of several meters can be manufactured from UP resins by combining the rotation process with a modified fiber-spraying technique (Fig. 67). The mold is an aluminum rotor strengthened by external ribs and mounted on driven rolls. The feed vessel for the resin mixture and the fiber cutter are mounted in the head of a telescopic arm that is driven back and forth during coating in the rotating rotor. The glass fibers leaving the cutter are oriented by air blowing so that the ratio of longitudinal to transverse orientation matches the expected mechanical loading of the container. Vertical containers, feed silos, and similar articles can be made with progressively increasing wall thickness from top to bottom. The roof of the container is placed in the rotor prior to rotation and is immediately bonded onto the container wall. The attainable glass content is 25 – 30 wt %.

2.1.9. Continuous Processes

Pultrusion. The pultrusion process is particularly suitable for mass production of uniaxially loaded solid and hollow profiles of any length. Labor costs are low, and equipment costs depend on the profile cross section. Glass rovings are used as reinforcement.

Figure 68 shows a vertical pultrusion machine. The rovings are wet with the resin compound in an immersion bath and then drawn into a heated profile mold. An inlet nozzle at the entrance of the mold retains any excess resin and entrapped air. The reactivity of the resin, curing temperature, temperature of the mold, and takeoff speed must be matched so that the resin mixture is dimensionally stable when it leaves the profile mold. Premature gelation causes adhesion to the mold, and delayed gelation can lead to deformation of the profile cross section. The solidified profile subsequently passes through a heating channel and is cured completely. The profile is then cut to length by a flying saw. Attainable glass contents vary between 50 and 75 wt %. The profiles can withstand high tensile forces in the fiber direction, but the strength in the transverse direction corresponds to that of the resin alone. This can be improved by employing textile tapes or looped rovings.

Continuous lamination is used for the manufacture of plane and longitudinally or laterally corrugated endless sheets of constant thickness. Glass matting or chopped glass fibers are used for reinforcement. Lightly colored, light-stabilized UP resins are employed. The production of corrugated sheets was the first large-scale processing technique for UP resins. It requires careful matching of resin – hardener system, curing temperature, and production speed to obtain an economical throughput via rapid curing times and to avoid a sudden curing reaction. This could

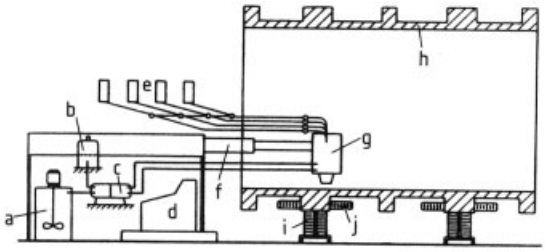

Figure 67. Centrifugal casting plant for large vessels
a) Vessel for resin and accelerator; b) Vessel for hardener; c) Mixing and metering pump; d) Control unit; e) Roving reels; f) Telescopic arm; g) Application unit with cutting device, compressed air inlet, and resin feed; h) Rotor; i) Driving rolls; j) Guide rolls

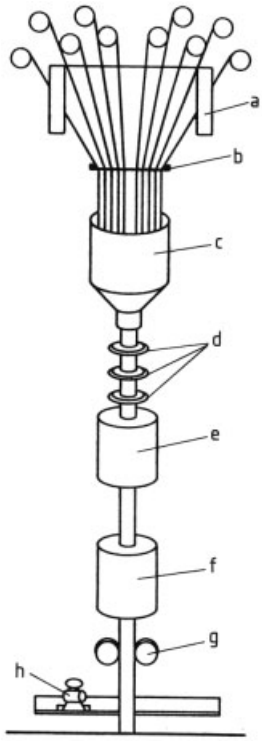

Figure 68. Pultrusion
a) Roving frame; b) Roving guide; c) Impregnating bath; d) Inlet nozzles; e) Profile mold; f) Heating channel; g) Take-off unit; h) Cutting device

lead to shrinkage of the resin from the fibers, which results in loss of transparency and embrittlement. Two processes have been developed for the production of corrugated sheets, which differ in the method of shaping. Figure 69 shows a machine for the production of transversely corrugated sheets. The resin is cast onto a cellulose or polyester release film and spread homogeneously by a doctor blade. The glass matting is then added, or glass fibers are distributed from a cutter unit fed with rovings positioned above the substrate. An open zone follows in which the resin is allowed to impregnate the fibers and air is forced out. Subsequently, a second release film is added from above, and a pair of calibrating rolls presses excess resin and air bubbles out of the laminate. The forming unit consists of a forming roller with displaced upper and lower rolls. The sheet passes between the shaping rolls in the curing oven, which consists of gelation and curing zones. The release films are then removed, the edges are trimmed, and the completed corrugated sheet is wound.

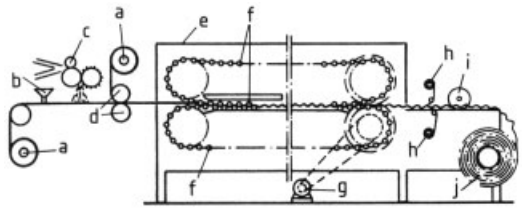

Figure 69. Continuous lamination of transversely corrugated sheets
a) Release film; b) Application of the resin; c) Fiber cutting device; d) Calibrating rolls; e) Hardening oven; f) Forming roller; g) Drive; h) Removal of release films; i) Trimming; j) Windup

With longitudinal corrugating equipment, the impregnated laminate layer, calibrated between two films, is first shaped on a special forming element to a narrower-profile band and then passed through the curing oven on a longitudinally profiled slide system.

A significant improvement of the continuous process can be achieved by use of UV-curing UP

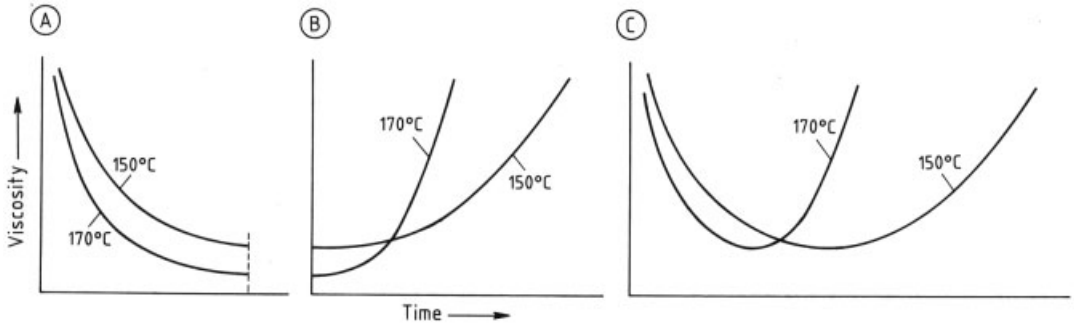

Figure 70. Flow – cure behavior of curable molding compounds
A) Molding compound without hardener, heated in the mold; B) Molding compound with hardener, heated immediately to mold temperature; C) Actual behavior of curable molding compound in press molding

resins because the hardening process is easier to control.

2.2. Curable Molding Compounds and Laminates

Moldings are manufactured from curable compounds under pressure and heat in closed molds; the molding compound is melted (or softened), shaped, and cured.

Terminology. Molding compounds are unshaped products that can be permanently molded to articles or semifinished products under pressure and temperature. Curable (thermoset) and uncurable (thermoplastic) molding compounds are available. Curable molding compounds always contain ca. 50 – 60 % filler and reinforcing materials. Press compounds are molding compounds that are suitable for press molding, injection press molding, or extrusion. Injection molding compounds are used for injection molding. The processing of curable molding compounds and the manufacture of laminates are described below; for low-pressure processes, see Section 2.1.6.

2.2.1. Processing Techniques

2.2.1.1. Processing Curable Molding Compounds

Curable molding compounds are formed and cured under the influences of temperature and pressure. On heating, the curable resin (which is present in the mass as a binder) softens, and the molding compound flows, filling the mold cavity. At the same time the cross-linking reaction begins, in which the binding agent and hence the molding compound harden. From this moment the material can no longer be melted; the curing (cross-linking) is irreversible. The molding is removed while still hot.

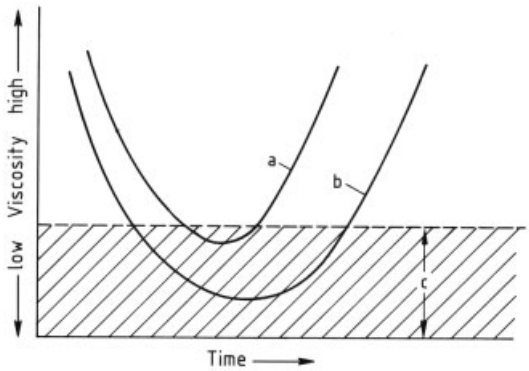

Figure 71. Flow – cure behavior for hard and soft flow
a) Hard flow; b) Soft flow; c) Possible processing range

Flow – Cure Behavior. Figure 70 illustrates the flow – cure behavior of curable molding compounds. Figure 70A shows how the molding compound is warmed and softens in the heated compression mold when no curing reaction takes place (this can be brought about by omitting the curing agent during preparation of the molding compound, which then behaves like a thermoplastic). Figure 70B shows how the viscosity of the molding mass increases and the mass hardens if the curing agent is added suddenly and the

mold is heated to 150 – 170 °C. Figure 70C, the superposition of Figures 70A and 70 B, shows the changes in viscosity observed when the mold is heated to 150 or 170 °C, filled with cold molding compound, and closed. Figure 71 shows the difference in behavior between hard and soft flow molding compounds. A hard-flowing compound generally cures faster than a similar compound with soft flow. The required viscosity of the molding compound depends on the design of the molding, the processing method, and the available forming pressure. In all cases, it must lie within the shaded zone in Figure 71; otherwise, the compound is too hard and the mold cavity will not be correctly filled.

Preparation Stages. Curable molding compounds are prepared by the manufacturers to be processed without further additives; the molding compound supplied to the processor contains all the necessary fillers, curing agents, stabilizers, lubricants, colorants, and flame retardants. The processor adds other materials only in exceptional cases. Occasionally, small amounts of lubricant can be added (e.g., 0.1 – 0.2 % zinc stearate) when particular processing conditions make this necessary. Molding compounds for screw injection molding machines can be colored by the processor, although it is difficult to manufacture uniformly colored molding in this way and the technique is seldomly used.

Release agents prevent adhesion of the cured resin to the surfaces of the mold that come in contact with the molding compound (mold cavity, gate, runner system, and all parts that could lead to the formation of flash). Waxes, metal soaps, fats, oils, and silicones, which are applied as solutions, sprays, or aqueous emulsions, or are melted and spread with a brush or compressed air, are available.

In the production of moldings that are to be painted, only release agents that do not impair adhesion of the paint should be used; otherwise, the molding would have to be degreased in an extra process stage.

Metering of the molding compound in a preliminary step is necessary for press molding, injection press molding, and pelletizing. Injection molding machines meter the material themselves (in the compressed state), as do plastication units of all types.

Automatic metering systems require satisfactory flow properties of the molding compound. Finely powdered compounds or those with high dust content require special accessories to permit automatic metering. Fibrous or doughy compounds can be metered only – if at all – by using special units.

Molding compounds are cut to the desired length and reweighed after the release film is removed. Exact cutting is not usually necessary.

Pelletizing of the molding mass is usually carried out for the following reasons:

1) Because the mold cavity is too small for the molding compound
2) To rationalize the feeding process
3) To permit high-frequency prewarming

Automated pelletizing machines operate with either constant pellet height or constant pelletizing force. Typical pelletizing pressures are 100 MPa and above. Circular pellets are generally used, with heights amounting to 20 – 60 % of the diameter. In special cases, square, rectangular, or ring-shaped pellets are produced.

Prewarming the molding compound simplifies processing in the case of press molding and injection press molding. Prewarmed compounds flow better and hence require lower press pressures and shorter closing times for press molding, as well as reduced injection pressure and shorter injection times for injection press molding; furthermore, the curing time is also reduced, particularly with press molding.

Phenolic and aminoplastic molding compounds evolve moisture and other volatile components during prewarming, which has a favorable effect on the properties of moldings. They exhibit reduced shrinkage, better electrical insulation, improved surface gloss, and reduced moisture absorption.

Polyester molding compounds incorporating styrene as a curing agent should be prewarmed only slightly or not at all.

Widely used prewarming processes are

1) High-frequency prewarming
2) Screw plastication
3) Oven warming (only applicable to phenolic and aminoplastic molding compounds, usually to remove moisture and other volatile components)
4) Infrared prewarming, where a thin layer of the compound is exposed to radiation

Molding compounds can be prewarmed quickly and homogeneously in a high-frequency field; pelletized materials are best suited. A prerequisite for a homogeneous temperature is constant thickness of the molding compound or constant pellet height and density. Typical prewarming times range between 0.5 and 2 min; temperatures of 100 – 140 °C are used.

Press Molding. A metered amount of the molding mass – either loose or in pellet form; as-received, prewarmed, or plasticated – is fed into the open mold, which has already been heated to curing temperature (Fig. 72). When the mold is closed, the mass softens as a result of heat flux from the mold walls and flows into the mold contours. Curing then takes place while the mold is closed. The hot molding is removed when sufficient curing has occurred and is then deflashed. Advantages of the process are the simplicity of the molds and presses; the low levels of filler orientation in the molding and, therefore, the almost isotropic properties; and the low molding shrinkage. Disadvantages include long curing times (especially when prewarming or preplastication is not employed), pronounced flash formation because the mold is closed only when the compound has begun to flow out of the mold (displacement of air), and the large deviations in mold-independent dimensions (see DIN 16 901).

Large, high-quality moldings (e.g., for automobile bodies) can be manufactured by press molding of molding compounds. Such moldings are processed by using computer-controlled, parallel-regulated hydraulic presses. The investment costs are high.

The following stages are involved in press molding:

1) Filling the mold with the molding compound
2) Closing the press until the mass is squeezed out of the mold
3) Opening the press briefly ("breathing", optional)
4) Closing the press until curing is complete
5) Opening the press
6) Removal of the molding
7) Cleaning (blowing out) the mold

When metallic inserts are to be incorporated in the molding, they are placed in the mold prior to filling with the molding compound.

Injection Press Molding. In injection press molding (Fig. 73) a metered quantity of the molding compound, usually pelletized and prewarmed or preplasticated, is introduced into the injection cylinder of the mold. After the mold is closed, the compound is compressed by a piston, further plasticated, and injected through the gate into the mold cavities. The internal friction that results from this action leads to further heating, which accelerates the curing reaction. After curing, the molding and gate are either simultaneously or separately ejected from the open mold.

The advantage of injection press molding over press molding result from the fact that the mass is injected into the closed mold while hot and, hence, of relatively low viscosity:

1) Less flash is formed
2) The mold-independent dimensions are more exact
3) The manufacture of complicated contours and variable wall thicknesses is possible
4) Embedding of sensitive metallic components is possible
5) Curing times are reduced, particularly with thick-walled moldings

The disadvantages include the high consumption of molding compound since the material that remains in front of the injector piston and in the runner system also cures and cannot be reused. Additionally, depending on the gate geometry, variations in filler orientation can result, which in turn can lead to anisotropy of the strength and dimensional stability.

Injection Molding. In injection molding of curable molding compounds, the material is metered, plasticated, and injected into the heated mold by the screw of the injection cylinder (Section 1.4.5). The cylinder is heated moderately, while the mold is heated to the final curing temperature. Due to shear heating of the compound (dissipation) during plastication in the cylinder and passage through the nozzle and gate, it is well plasticated and prewarmed. This leads to shorter curing times in the mold because only small amounts of heat need to be supplied conductively. The cycle times are short. More information on injection molding equipment is given in Section 1.4.5.

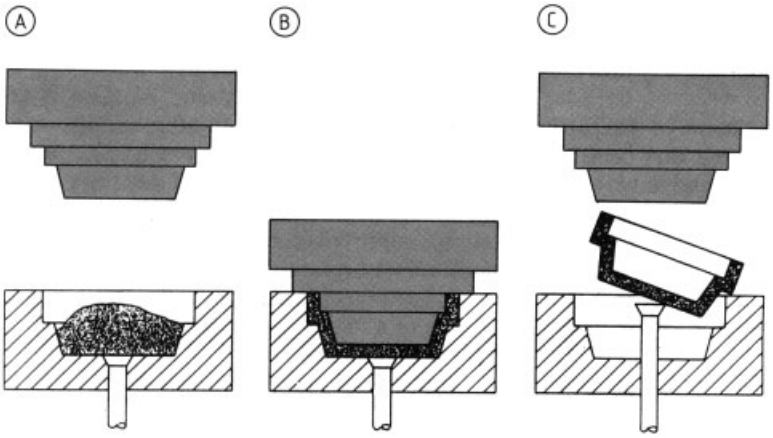

Figure 72. Press molding
A) Filling; B) Pressing; C) Ejection

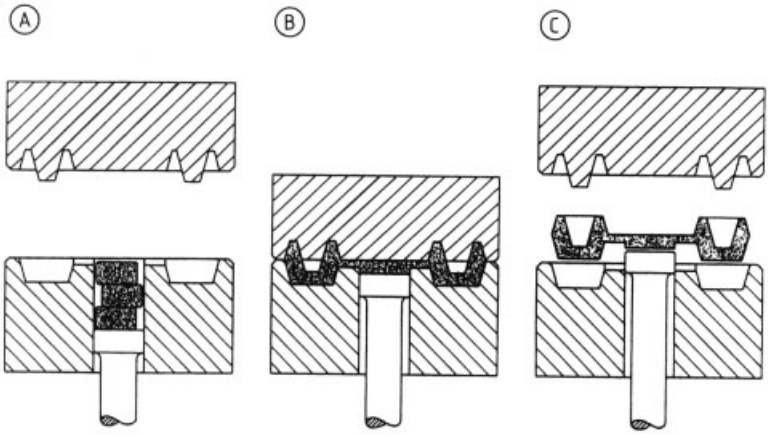

Figure 73. Injection press molding
A) Filling; B) Injection and pressing; C) Ejection

The advantages and disadvantages of injection molding compared to press molding are similar to those for injection press molding. The process cycle is generally shorter than with injection press molding, and pelletizing and high-frequency warming stages are not required. Additionally, a higher degree of automation is possible with injection molding.

Injection – compression molding (Fig. 74) is an adaptation of injection molding. With injection – compression molding, the mold is initially closed without pressure, and the plasticated molding mass is injected, whereby the mold opens up a few millimeters. The mold is then closed under pressure, and the compound is formed (see also Section 1.4.5).

Injection – compression molding offers the following advantages:

1) Filler orientation similar to that of press molding, which improves with increasing flow of the compound during pressing (the favorable filler orientation leads to more isotropic strength and improved dimensional precision)
2) Relatively low closing force required for pressing
3) Relatively low injection pressure and therefore lower back pressure at the screw and higher shot volume
4) Higher injection speeds, which can partially compensate for time losses caused by pressing

5) Disappearance of the gate during embossing (given appropriate mold construction)
6) Improved degassing
7) More uniform molding surface characteristics

Posttreatment. The molding is generally subjected to posttreatment after removal from the mold.

Setting. In certain cases, particularly when high dimensional precision is required, thermoset moldings are placed in a cooling template after being demolded.

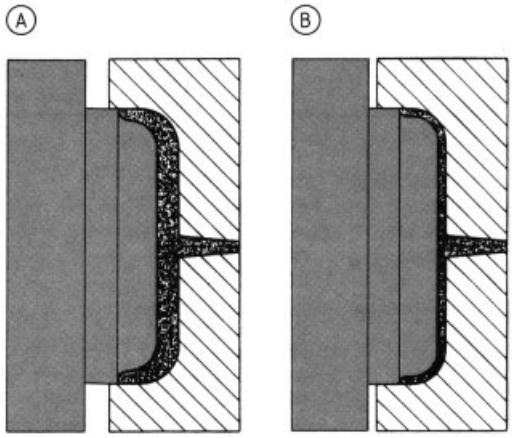

Figure 74. Injection – compression molding
A) Injection; B) Compression

Deflashing. Moldings produced by press, injection press, and injection molding have varying amounts of flash that must be removed. For reasons of economy, flash should be removed as quickly and simply as possible. For this, the flash must be as thin as feasible. Flash formation depends on the molding compound, processing technique, processing conditions, and above all on the quality of the mold. The finer the filler in the molding compound, the better the molding can be deflashed. Automatic deflashing is not possible with long-fiber fillers. Several of the more common deflashing processes are compared in Table 5.

Air-jet deflashing is the technically most developed and economical technique. With particle-jet deflashing, particles should be softer than the material from which the flash is to be removed. Organic materials are typically employed for jet particles (e.g., ground apricot stones or polyamide granules). The minimum particle diameter is ca. 0.5 mm; otherwise, the intensity of the collision is too low. Modern deflashing equipment has multiple nozzles, along with dust removal units and feed mechanisms for antistatic agents.

Polishing. If the mold has perfect surface quality, the resulting molding will have a smooth, glossy surface. This applies only to a limited extent to compounds with coarse fillers. The surface can be further improved by grinding and polishing.

Small moldings that are deflashed in drums can be simultaneously polished by addition of polishing agents.

Annealing. Postcuring is not usually necessary for thermoset moldings. However, heat treatment of the molding is often advantageous if it is to be thermally loaded during use. With polycondensation resins, annealing results in the slow discharge of volatile components, which in turn results in the advanced removal of postshrinkage, improved electrical properties, and increased application temperature.

General statements regarding duration and temperature of annealing are not possible; they depend on the case in question and must be determined experimentally. An extended treatment at lower temperature is always better for the material than a short period of annealing at higher temperature, even though the shrinkage is identical in both cases. This is particularly true for thermally sensitive aminoplastic materials.

2.2.1.2. Manufacture of Laminated Plates

Laminated plates are semifinished products manufactured from layered resin substrates with curable resins as binders. Phenolic, aminoplastic, unsaturated polyester, epoxy, and silicone resins are used as binders. Suitable resin substrates include paper (for paper-base laminate); cotton and glass fabrics (for fabric-based laminates); glass matting (for glass-mat-based laminates); and wood veneers (for synthetic resin laminated wood).

In the manufacture of laminated sheets, webs of the resin substrate are coated or impregnated with resin, dried (when necessary), cut, layered,

Table 5. Deflashing of thermoset moldings

Deflashing	Application	Performance (number of articles per unit time)	Deflashing quality	Dust formation	Influence on surface quality
Mechanical treatment (drilling, filing)	flash thickness >0.4 mm or fiber-filled molding	low	high	low [c]	none
Abrasive belt treatment	flash in one or two planes (max.)	low	high	low [c]	damage possible
Drum deflashing	only with loose flash; no brittle parts and no sensitive metal inserts	high	moderate to good [a]	low [b]	low, not recommended for larger moldings with high surface quality
Air jet	flash thickness: 0.1 mm (max.)	high (several nozzles)	sufficient to good [a]	low [c]	none
Particle jet	flash thickness: 0.1 – 0.4 mm	high	good	low [c]	none

[a] Depending on the shape of the molding.
[b] With additional aids such as wet sawdust.
[c] Removal by suction is recommended.

and pressed between plane parallel plates. Phenolic and aminoplastic laminated papers and fabrics are usually pressed in multiplate presses at ca. 160 °C and 10 – 13 MPa (with synthetic resin laminated board at 25 MPa or higher).

Additionally, twin-belt presses are increasingly being employed because of their ability to function in a continuous process.

Laminates based on other resins are often pressed at considerably lower pressure (several bar).

Depending on binder, temperature, and laminate thickness, curing times vary from a few minutes to several hours. For some applications, laminated sheets are coated on one or both sides with special surface layers, for example:

1) With copper foil (laminated paper or fabric laminated to copper foil) for the production of printed circuits
2) Fine wood veneers for laminated board
3) With colored or printed, aminoplastic resin-impregnated papers for decorative laminated boards (e.g., for the furniture industry)

Aminoplastic-impregnated, decorative paper laminates can also be laminated directly to wood sheets (e.g., chip board).

Laminate moldings can also be pressed by using closed press molds once they have been appropriately cut and layered. Examples of this include pressed plywood moldings (for chairs) and pressed moldings made of polyester mat-based laminates.

2.2.2. Processing Properties

The processibility of a curable molding compound depends on structure, flow characteristics, and curing behavior.

The *structure* plays an important roll in automatic metering; consistent, rounded particles with good flow properties are preferred. The dust content should be low. Molding compounds containing fibrous reinforcing materials (e.g., cellulose, glass fibers) are granulated in many cases; glass-fiber compounds are also available in rod form. Impact-resistant molding compounds are employed in randomly oriented fiber structures; however, in such cases, the molding compound must be weighed and added by hand, and the mold must have a sufficiently large cavity.

Flow characteristics are adjusted by the manufacturer of the molding compound. With condensation resin molding compounds this is carried out by varying the extent of condensation of the binder during production. The flow level of a molding compound is selected according to the processing technique and the complexity of the molding. It also exerts an influence on the quality of the molding. Hard-flowing compounds generally give lower molding shrinkage, better surface gloss, and improved electrical insulation. However, they require higher processing pressure, often cannot be injected, and often cure more quickly.

The *curing behavior* of a molding compound can also be varied by the manufacture within certain limits. As a rule, a processor desires high curing speeds to attain high production throughputs. The higher the processing temperature, the

faster is the curing reaction. However, the temperature can be increased only to a level at which the mold can still be filled without difficulty.

2.2.3. Processing Conditions

Processing guidelines can be established and described only for individual cases; they differ according to molding compound, processing technique, mold, and molding. Furthermore, processing difficulties or molding faults that may arise can be dealt with only as individual cases. The causes may lie in the molding compound, the processing equipment, or processing conditions so that a knowledge of all possible circumstances is necessary to make any evaluation or correction.

Table 6 gives an overview of processing conditions for the most commonly employed curable molding compounds.

Phenolic resin molding compounds are the thermosets that are easiest to process, regardless of whether press molding, injection press molding, or injection molding is used. This applies particularly to the standard compound of phenolic resin with sawdust (DIN 7708, type 3), but also to most other phenolic molding compounds.

Some fiber and macerate compounds, because of their structure, cannot be processed automatically (or only by using special equipment).

The most important operating conditions in press molding processes include temperature, pressing force, and curing time. When pressing phenolic molding compounds, the mold temperature is typically 160 – 170 °C. Higher temperature leads to quicker curing; hence the highest temperature possible is always used. The upper temperature limit is

1) For compounds containing sawdust, ca. 180 °C
2) For compounds with organic fibers or fabric cuttings, ca. 170 °C
3) For compounds with inorganic resin substrates, ca. 190 °C

In all cases, care must be taken to ensure that the temperature is not so high as to partially cure the material while the mold is closing and thus lead to faulty moldings.

The harder the flow characteristics, the larger the resin substrate, or the less intense the preheating, the higher is the pressing force required. Additionally, design, wall thickness, and height all play a role in determining the required pressure.

With *press molding*, curing time depends on the thickness of the pressed component, the mold temperature, the curing characteristics of the molding compound, and the prewarming. For phenolic resins without prewarming, a mold temperature of 160 °C, and a molding thickness of 2 – 6 mm, a curing time of 30 – 60 s is estimated per millimeter of wall thickness, whereby the thickest part of the molding must be considered. Curing time can be decreased by increasing the mold temperature and particularly by pre-heating. In this way, the curing time even of thick-walled moldings can be reduced to 1 – 2 min.

Additionally, the curing time can be shortened by ventilation (i.e., temporary opening of the mold shortly after the initial closing stage, which allows any built-up gases to escape).

With *injection press molding*, the most important operating conditions are mold temperature, injection pressure, closing pressure, and curing time. The same applies in this case regarding mold temperature as reported above for pressing techniques.

In practice, injection pressures up to ca. 200 MPa, and in certain cases even higher, are employed. Often independent adjustment of the injection velocity and the injection force is desirable; too high an injection velocity can cause local scorching of the molding due to adiabatic overheating if air present in the mold cannot escape quickly enough.

The required injection pressure depends on the mold (molding design, gating), the molding compound (flow properties, resin substrate), and preheating. The curing time is generally only slightly dependent on wall thickness in injection press molding. When the gate is correctly dimensioned, the molding compound enters the cavity at the full mold temperature, and after the injection phase, only curing takes place. In practice, curing times in the injection press molding of phenolic compounds lie in the range of 0.7 – 2 min.

With *injection molding,* preparation stages (pelletizing, prewarming) are not required.

Table 6. Processing conditions for curable molding compounds ∗

Process	Phenolic resin molding compound	Urethane resin molding compound	Melamine resin molding compound	Melamine–phenol resin molding compound	Polyester resin molding compound
Press molding					
Mold temperature, °C	160 – 190	130 – 160	150 – 170	150 – 170	130 – 190
Molding pressure, MPa	15 – 80	15 – 80	15 – 80	15 – 80	10 – 40
Approximate hardening time per millimeter of wall thickness, s	30 – 60	20 – 40	20 – 40	20 – 40	10 – 20
Injection press molding					
Mold temperature, °C	160 – 190	130 – 160	150 – 170	150 – 170	140 – 190
Outer injection pressure, MPa	50 – 200	50 – 200	50 – 200	50 – 200	30 – 80
Hardening time, s	40 – 120	30 – 120	30 – 120	30 – 120	10 – 60
Injection molding					
Temperature, °C					
Cylinder (feed zone)	65 – 85	70 – 80	65 – 80	65 – 80	40 – 60
Nozzle	85 – 120	95 – 125	90 – 120	90 – 110	60 – 90
Screw	65 – 85	70 – 80	65 – 120	65 – 80	40 – 60
Mold	110 – 140	120 – 140	130 – 140	125 – 135	80 – 110
Compound	170 – 190	140 – 160	160 – 180	160 – 180	150 – 190
Outer injection pressure, MPa	80 – 250	100 – 250	100 – 250	100 – 250	30 – 100
Hardening time ∗∗, s	20 – 80	15 – 80	15 – 80	15 – 80	10 – 60

∗ Optimal conditions for individual cases must be determined experimentally. Some cases require conditions that do not lie in the range described here.
∗∗ Independent of wall thickness for thickness up to 20 mm.

However, several specific operating conditions in addition to those for injection press molding must be considered: the temperature of the plastication and injection units, screw rotation speed during plastication, back pressure during plastication, holding pressure, plastication time, and cycle time. In many cases, the mold temperature exceeds that used in press or injection press molding.

The injection cylinder is generally cooled in the feed zone. The next zone, the metering zone, is generally heated to 65 – 85 °C. The temperature of the front zone of the cylinder and the nozzle should be 85 – 120 °C. Up to a diameter of ca. 50 mm, control of the temperature of the screw is not normally necessary.

Fluid heating systems give good temperature consistency and permit removal of the excess heat of friction that can be generated under certain circumstances during plastication.

The temperature of the compound, measured on the plasticated material injected through the nozzle, is often considerably higher (20 – 40 K) than the temperature of the cylinder. Hotter molding compound can be injected more easily and hardens faster, but the risk of operational interruption is also greater.

Adjustment of the back pressure to counter the reverse movement of the screw during plastication allows control of the plastication stage. At low back pressure (3 – 5 MPa) plastication is still possible while the injection unit is withdrawn; high back pressure (8 – 15 MPa) leads to high compound temperature. Computer control of the back pressure opens up new possibilities for decreasing curing time.

The injection pressure in injection molding corresponds roughly with that of injection press molding with thoroughly preheated compounds. Depending on the flow rating and resin substrate of the molding compound, as well as the complexity of the mold, the injection pressure typically lies in the range 80 – 200 MPa, with an injection time of 3 – 10 s. After injection, a holding pressure is applied. The holding pressure should be 50 – 80 % of the injection pressure and be maintained until the material in the gate has hardened.

The curing time is generally lower with injection molding than with injection press molding. If the layout of the mold (gate) is correct, the curing times for phenolic molding compounds range from 0.3 to 1 min. They are independent of molding thickness over a wide range of values.

The cycle times are determined mainly by the curing time. Only in cases of very long plastication times or very short curing times does the cycle depend on plastication time rather than curing time. Whenever possible, the plastication

time in such cases is reduced by processing with a faster screw.

Aminoplastic molding compounds (i.e., molding compounds with urea, melamine, or melamine – phenol resins) have similar processing characteristics to phenolic molding compounds. Hence, the processing information provided above for phenolic compounds is largely valid for aminoplastics. Deviations exist only for the processing temperature and curing time (see Table 6).

Urea resin molding compounds are particularly sensitive to overcuring, which can occur with standard curing times at 170 °C mold temperature, or with prolonged curing times at 160 °C. Melamine and melamine – phenolic resin molding compounds are less sensitive to overcuring, although excessive mold temperatures are not recommended.

With thin-walled moldings, the temperature can be increased to the upper limits of the regions mentioned above; with thick-walled components, however, lower temperatures are used. When press and injection press processes are employed, extensive high-frequency preheating is normal.

The curing time of aminoplastic molding compounds is usually 30 – 50 % shorter than that of corresponding phenolic compounds, given identical processing techniques, although mold temperatures with urea resin molding masses are as a rule 10 K lower.

Polyester resin molding compounds behave differently from phenolic and aminoplastic resins, the latter being condensation products that cure relatively mildly. In contrast, the polymerization reaction involved in curing polyester resins is usually strongly exothermic.

The mold temperature influences flow properties. If the mold is too hot, the resin cures before the mold cavity is completely filled; hence, the mold temperature has an upper limit. At the lower temperature limit the molding sticks in the mold. Between these two extremes, sufficient room for maneuverability exists to allow for malfunction-free operation. Additionally, the working temperature range is determined by the composition of the molding compound, particularly by the type of curing catalyst employed.

Except at low temperature, curing time is usually relatively short; in some cases, a few seconds suffice.

Air that is entrapped in the mold and unable to escape can easily lead to faults in the molding, particularly with injection press molding injection molding. This involves not only scorching but also insufficiently cured areas because atmospheric oxgen can disturb the curing reaction.

Contamination of the molding mass by traces of phenolic resins can impair curing; processing of phenoplastic and polyester resin molding masses should be done separately.

In mold construction, undercuts must be avoided for polyester resin molding masses, since moldings with undercuts cannot be demolded.

Ventilation of the mold is generally not recommended for processing polyester resin molding compounds, because the resulting decrease in the mold temperature is undesirable.

Polyester resin molding compounds are available as granules, long fibers (premix), and resin matting (sheet molding compounds). Granules are used for press, injection press, and injection molding.

Processing conditions are listed in Table 6.

Polyester resin molding compounds with randomly distributed long glass fibers – usually with styrene as cross-linking agent and moist doughy to semidry, strawlike consistency – can be processed at relatively low pressure (ca. 10 MPa for press molding and $\geq$30 MPa for injection press molding) and also injection molded by using special equipment. Additionally, polyester resin premix is frequently processed with low-pressure press molding techniques (see Section 2.1.6).

Styrene-free, dry polyester molding compounds are also available; they usually contain diallyl phthalate as cross-linking agent. These are also supplied as rods (chopped strands with parallel glass fibers) that require higher processing pressures.

Polyester resin sheet molding compounds are usually processed by press molding to give large-area moldings. Similarly, polyester sheet molding compounds are often processed by low-pressure press molding techniques (see Section 2.1.6).

2.2.4. Influence of Processing on Molding Properties

The properties of a plastic molding are determined not simply by material properties, but also determined by other factors, particularly the shape of the molding and the processing itself. For this reason, tabulated values determined on standard samples are valid only for those samples and cannot be transposed to any arbitrarily designed molding. The most important factors from a processing viewpoint are degree of curing and filler orientation.

The *degree of curing* depends on the original state of the molding compound and the processing procedure, in particular the variation of temperature with time and the length of the curing process. The degree of curing may not be constant throughout the molding. A thick-walled press molding, for instance, may be fully cured in the outer layers but still uncured in the center. With moldings made from urea resin compounds overcured, undercured, and well-cured regions may be found within one and the same article.

Overcuring can be recognized by the lack of surface gloss. In addition, bubbles can occur, and the molding may adhere to the steel walls of the mold. Overcuring of aminoplastics can lead to bubbles and, in some cases, cracks.

Variations in curing conditions can lead to internal stresses that might prove to be a drawback if the molding gives the appearance of being correctly cured. Undercuring of the complete molding is naturally a more significant problem; with aminoplastics this is also true of overcuring. Many properties depend on the degree of curing for phenolics and aminoplastics, these include surface gloss, stiffness, dimensions, water uptake, and electrical insulation properties.

With all processing techniques, *filler orientation* depends strongly on the flow process involved. Since, when processing thermosetting molding compounds, the tensile deformations exceed the shear deformations, the filler and reinforcing additives are oriented mainly perpendicular to the flow direction.

Filler orientation particularly determines the dimensional stability and strength of the molding. In injection press molding and injection molding, orientation can be influenced mainly by the gate position and less so by processing conditions. With press molding, the flow and hence the orientation can be influenced by the type of feeding, (e.g., concentrating the material in one position). Orientation has a greater effect with long-fiber reinforcement than with, for example, sawdust.

Molding shrinkage is defined as the difference between the dimensions of the mold and the corresponding dimensions of the molding. It is influenced significantly by the molding compound and the filler orientation, and additionally by the processing conditions, in particular the mold temperature and the pressure in the mold cavity. The influence of filler orientation on shrinkage arises because shrinkage is impeded in the direction parallel to the reinforcing fibers but not in the direction perpendicular to them.

The following applies to the relationship between shrinkage and processing conditions:

1) The higher the mold temperature, the greater is the molding shrinkage
2) The higher the pressure on the molding, the smaller is the process shrinkage
3) Low curing leads to increased molding shrinkage

Distortions arising from local differences in shrinkage are also influenced by filler orientation. The higher the content of inorganic components in a resin substrate, the less pronounced is the tendency to distort. In contrast, the longer the fibers in impact-resistant compounds, the greater is the distortion to be expected. Distortion can also result from local differences in the curing reaction (mold temperatures) or from inhomogeneous cooling of the molding.

The molding shrinkage can be reduced or even completely compensated by adding suitable fillers (e.g., thermoplastic powders), especially in the case of polyester resin molding compounds (low-profile systems).

Mechanical properties are often influenced strongly by filler orientation. With long-fiber reinforcement (glass fibers) and pronounced orientation, the strength in the direction of highest orientation can be several times greater than that in the least oriented direction. Materials containing short-fiber fillers such as sawdust exhibit similar differences in strength in the flow direction and perpendicular to it; however, these differences are small (typically 10 – 30 %). In contrast, curing conditions play only a secondary

role in mechanical strength, if the molding has not been strongly over- or undercured.

The electrical insulation properties of phenolics and aminoplastics depend on the degree of curing. Undercuring has a negative effect, particularly on surface resistance.

To attain the best possible insulation characteristics:

1) As hard flowing a molding mass as possible should be used
2) The molding mass should be dried in an oven prior to processing
3) The molding must be perfectly cured

With aminoplastics, overcuring can be as detrimental as undercuring. Phenolic resins are less sensitive to overcuring.

For polyester molding masses, the relationship between degree of curing and electrical properties has minor practical significance. Undercured polyester moldings can be recognized visually and are not suitable for use even though the electrical characteristics are adequate.

2.2.5. Molds

The quality of the mold has a considerable influence on the manufacturing costs of the molding. With small production runs, the mold should be as simple as possible even if this necessitates use of a finishing process. For mass production, use of a more precise and possibly more complex mold is worthwhile if it avoids finishing.

Most molds for curable molding compounds are manufactured by machining or spark erosion of steel blocks. High-quality steels are employed, particularly for mold cavities. The forming surfaces are polished, hardened, and generally hard chromed to protect them against corrosion and wear. When cavities are manufactured by spark erosion, completely hardened steel is used. Nitriding is rarely employed in molds for processing thermosets. Mold components that are subject to the greatest wear should be easily replaceable or made from hard metals (e.g., gate inserts). With multicavity molds, the cavities themselves are often produced by hollowing with a stamp. Several systems of standard mold units are available with which mold manufacture can be rationalized. They include mold plates made of high-quality steel out of which the mold cavities can be cut, platens, guiding elements, and ejector systems that can be combined in a modular fashion.

Approximately 30 – 40 W/kg is sufficient to heat molds weighing 50 kg or more. Smaller molds and flat molds require up to ca. 80 W/kg. At least 20 mm of steel should separate the heating element from the mold cavity. The temperature sensor should be located between the heating element and the cavity. During operation, the mold temperature should not vary by more than 5 K. Ideally, each component of the mold should be equipped with its own heating and temperature control unit. Especially with small molds, insulation against heat loss is advantageous. Insulation by a thermal insulation sheet attached to the mounting platens is recommended in all cases to reduce heat losses.

In its simplest form, a *press mold* consists of a punch and a cavity (Fig. 72), generally electrically heated and mounted on the moving and fixed tables of a press. If required, an ejector system is built into the punch, cavity, or both, which releases the molding after the mold has been opened. The ejector pins are so distributed that the molding is ejected without indentation or deformation; the pressing surfaces are sufficiently large that they do not penetrate the molding even when it is still hot.

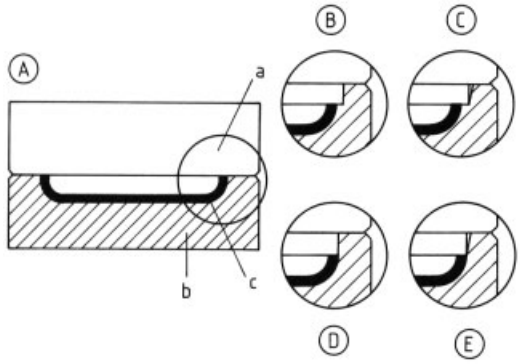

Figure 75. Shaping of the flash
A) Flash mold; B) Flash mold with charge cavity; C) Flash mold, charge cavity with inclined wall; D) Positive mold; E) Positive mold with inclined wall
a) Plunger; b) Punch; c) Mold

The formation of the flash surface plays a major role in a press mold (Fig. 75). One version or another is preferred, depending on whether a

vertical or horizontal flash is desired (a function of the preferred deflashing method), the available molding pressure, how exact the metering is, whether much air or molding mass is driven out via the flash, and the degree of precision demanded for the thickness of the base.

For press moldings with undercuts, side openings, screw threads, metal implants, and other specialties, multicomponent molds, split molds, and molds with movable inserts, side slides, or supplements are employed. Multicavity molds have several identical or different cavities, with separate or common filling volumes.

Injection press molds usually consist of a top part, a bottom part with injection cylinder, and an injection plunger activated by the weaker plunger of the injection press (transfer press, two-plunger press) (Fig. 73). Channels lead from the free end of the injection cylinder to the mold cavities, which are located in the parting surface between the upper and lower parts of the mold. The narrowest part of the channel, the gate, is located at the junction with the mold cavity.

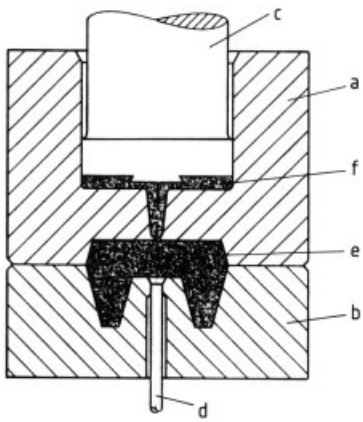

Figure 76. Injection press mold for single-plunger press
a) Top part of mold with injection cylinder; b) Bottom part of mold; c) Injection plunger; d) Ejector pin; e) Mold; f) Injection mass with gate

In another type of injection press mold (Fig. 76), the injection cylinder is incorporated in the top part of the mold, and the molding compound is injected through a gate in the top part into the mold cavity. These molds are used on single-plunger presses. Other mold constructions are also available that are suitable for injection pressing on single-plunger presses.

For *injection molding*, mold constructions typical for the injection molding of thermoplastics are generally preferred; such molds (Fig. 77) have two mold plates, one containing the sprue bush and the other the ejector system.

The cavities and runner systems are incorporated in the parting surface between the two plates.

Molds for injection press molding and injection molding are also available as multicomponent molds, split molds, molds with side slides, or screw thread inserts, etc. Molds for injection press and injection molding require venting channels if air contained in the cavity cannot escape quickly enough via leakages in the parting surface or ejector system during injection of the molding compound into the cavity. The channels are generally best positioned opposite the gate and should be approximately 0.1 – 0.2 mm deep and 2 – 8 mm wide. Alternatively, the closed mold can be evacuated prior to injection of the molding compound.

In molds for injection press and injection molding the gate is the narrowest point in the injection channel and lies directly in front of the junction with the mold cavity. The gate determines the flow of the molding compound during injection into the mold cavity and thus also the filler orientation and structure of the cured material.

With optimum dimensioning of the gate, the molding compound is subject to further heating as it passes through this final constriction so that it enters the mold cavity at a temperature just slightly below the mold temperature. If the gate is too narrow, the mold cavity is not completely filled; if it is too wide, the compound enters the cavity at too low a temperature and takes too long to cure. However, in injection press and injection molding, the conditions do not depend only on the gate, but also on the flow and curing properties of the molding compound along with the pressure and temperature.

As a rule of thumb for dimensioning the gate, with curable molding masses, for each cubic centimeter of cavity volume V, a gate cross section F of 1 mm^2 should be allowed:

$F/mm^2 = V/cm^3 \pm 50\,\%$

The margin of $\pm 50\,\%$ permits special cases to be taken into account. Molding compounds

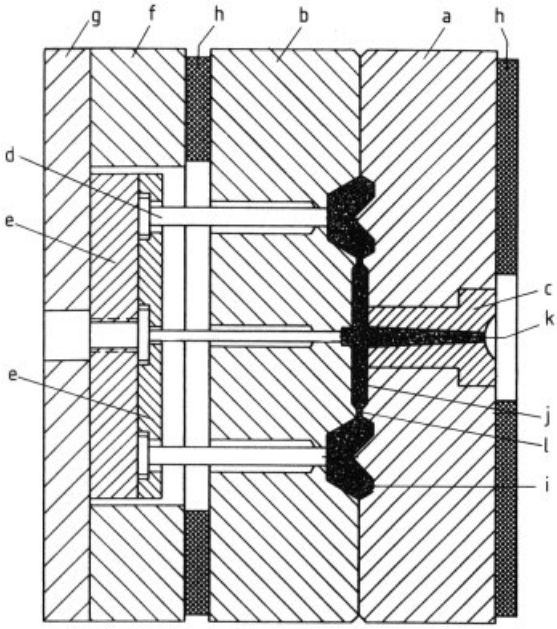

Figure 77. Injection mold
a), b) Mold plates; c) Sprue bush; d) Ejector pin; e) Ejector plates; f) Spacer; g) Adapter plate; h) Heat insulating plate; i) Mold; j) Runner; k) Sprue; l) Gate

with coarse and fibrous reinforcing materials require larger gates than those containing sawdust; furthermore, the flow properties of the compound, injection pressure, temperature, and design of the molding must also be considered.

The above rule is valid primarily for sheet or film gates (i.e., for rectangular cross sections with height-to-width ratios ranging from 1 : 10 to 1 : 200). Cylindrical pin gates can be made considerably narrower than estimated by the formula. Large amounts of material can generally pass through a pin gate without difficulty; cases are known in which the throughput is greater than 20 cm^3/mm^2. The pin gate is also easier to machine; it can be removed more easily from the molding and normally does not require finishing. It does, however, cause pronounced filler orientation in the immediate vicinity of the gate.

The most homogeneous orientation is achieved with a film gate over the complete width or length of the molding. The film gate should not be thinner than 0.2 mm, and with larger moldings or with impact-resistant masses, no thinner than ca. 0.4 mm.

The runner system should have a much larger cross section, and thus offer less resistance to flow than the gate; depending on the gate type and the position of the runner system, the latter should be three to ten times larger in cross section than the gate. It should approach the mold cavity with as full a cross section as possible so that the actual gate is short.

2.2.6. Machines

Presses and Injection Presses. Motor-driven *toggle presses* permit pressing forces up to ca. 1 MN. The toggle is so designed that the press initially closes rapidly and with little force, but toward the end of the motion becomes slower and reaches full force only over the final few millimeters. Overloading is prevented by incorporating a slip clutch between the drive and the toggle. The movement and the development of force of the toggle press are in most cases well suited for the conventional processing of curable molding compounds by press molding, however they are unsuitable when the full molding pressure is required over a displacement of several centimeters (e.g., the motion of an injection plunger in injection press molding).

Hydraulic presses dominate for press forces >1 MN; they are widely employed, however,

in laboratory presses down to 100 kN and are available up to ca. 100 MN. Depending on construction, distinctions are made between frame presses and column presses; downstroke presses (lower press table fixed, upper table moved by a piston) and upstroke presses; and presses with single drives (usually oil) or with group drives (usually water).

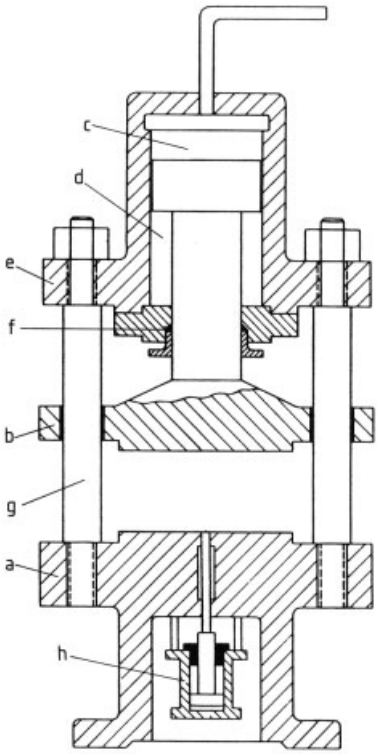

Figure 78. Hydraulic downstroke column press with hydraulic ejector mechanism
a) Fixed lower table; b) Movable upper table with differential piston for drawback; c) Pressure chamber for molding; d) Pressure chamber for drawback; e) Main cylinder; f) Concentric gasket; g) Pillars; h) Hydraulic ejector

For high-quality moldings, electronically parallel controlled presses are often used. In addition to the main cylinders that activate the closing and pressing mechanism, additional cylinders are present to ensure the parallelism of the press table even under asymmetrical loading.

Two-piston presses, which are often employed in injection press molding, are usually equipped with a powerful upper piston to close the mold and a weaker lower piston to operate an injection cylinder or an ejector. Figure 78 is a schematic of a two-piston press (column form). The force ratio between the pistons of a two-piston press is generally between 2 : 1 and 8 : 1. Apart from those mentioned above, alternative constructions are available (e.g., with injection cylinders positioned above).

Large-surface presses, with pressing forces that are relatively small compared to the area of the plate, are used primarily for processing resins with glass-fiber reinforcement, and also for low-viscosity polyester resin molding compounds with long glass fibers [e.g., DIN 16 911, type 801 (premix) or DIN 16 913, type 830].

Multiplaten presses with 2 – 20 parallel press platens are used to simultaneously press several identical items. Multiplaten presses, usually with steam-heated and, when required, water-cooled platens are used to press laminates (see Section 2.2.1.2). The platen size is up to 10 m^2, and the press force up to ca. 100 MN.

Semi- or fully automatic operation of all the press types outlined above is possible. In semiautomatic operation the process stages decisive for processing consistency – closing the press, controlling the curing time, opening the press, the injection stage in injection press molding, and ejection – are automated. The remaining steps – feeding the mold, removing the molding after completion of the program, and cleaning the mold – are carried out manually. With fully automatic plants, manual operations are limited to filling the molding compound hopper and transporting the completed moldings.

Injection Molding Equipment. Injection molding equipment for curable molding compounds (see Section 1.4.5) is similar to the screw injection molding machines used for processing thermoplastics. The major components are the clamping unit and the injection unit.

The clamping unit is a toggle or hydraulic press that closes and opens the mold and often activates the ejectors.

The injector unit, which consists of a cylinder (with cylinder head and nozzle), preferably fluidheated, and a screw, meters and plasticates the molding compound and injects it into the mold. For transportation, metering, and plastication, the screw rotates, usually driven by an electric or hydraulic motor. For injection, it moves lon-

gitudinally (normally without rotation); in this case the drive is almost always hydraulic. The injection unit can be moved in the longitudinal direction so that the nozzle can either contact the mold sprue bush or be removed from it.

With the majority of machines, the closing and injection units are positioned horizontally one after the other. However, other constructions also exist, for instance, where the closing and injection unit can be rotated into place when needed.

Injection molding of curable molding compounds requires higher injection pressure than for thermoplastics (at least 150 MPa, or better still, 200 – 250 MPa). Many thermoplastic injection molding machines can be modified for use with curable molding compounds.

Injection molding machines for curable molding compounds generally have a capacity of ca. 50 – 1000 cm^3 and a closing force of 5 MN; larger machines are, however, also in operation.

3. Further Processing

3.1. Machining

3.1.1. Estimating Machinability

The special properties and low cost of thermoformed plastics have led to their widespread use in the construction of machines and other equipment. Machining of components from plastics is significant because of the increasing number of design variations of manufactured components; the demand for greater stength and for cheap production of samples, individual articles, and small series; and the fabrication of thermoformed components (e.g., holes in PVC window frames for hinges). An important prerequisite for economical machining is the selection of optimum cutting conditions. This requires knowledge of the machinability of polymers and the wear resistance of the tools. A prerequisite for the economical utilization of machining techniques is a knowledge of the parameters that influence the machinability of the material. Experience from the metalworking industry cannot be applied directly to polymers. Even a comparison between the physical properties of metals and those of polymers shows that certain factors exerting a large influence on the machinability of metals are only slightly or not at all significant for plastics. Furthermore other factors can arise in polymer processing, such as the generation of dust, evolution of gaseous compounds, and melting of the component or of the shavings removed from it, that do not occur with metals.

The following criteria, which are also used in the metal industry, are used for judging machinability:

1) Effective machining forces
2) Resulting machining temperatures
3) Attainable surface quality, shaving formation, and wear of machining tools

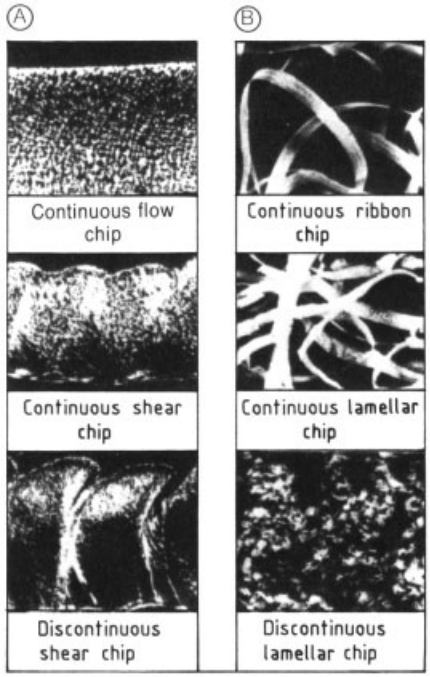

Figure 79. Classification of shaving types (A) and forms (B)

3.1.2. Shaving Formation

With thermoplastics, shaving formation influences the attainable surface quality and the static and dynamic process controlling the cutting action and advancing force. The tendency to form long, ribbonlike shavings often hinders machining, and the poor thermal conductivity causes

welding of the shavings, as well as distortion, leveling, melting, and smearing of the surface.

In turning polyamides, depending on machine conditions and material properties, three basic types of shaving may result, which can generally be broken down into continuous and discontinuous forms (Fig. 79). This figure shows that the shape of the shaving is a result of the type of shaving, which depends on water content, cutting geometry, and cutting speed, and its cross section.

The shape of the shaving is assessed by using the so-called shaving volume number, which is the ratio of the volume required by an amount of disordered shavings to the volume of material from which they were produced. It is most favorable for discontinuous shavings.

Drilling low-density polyethylene (LDPE), polyacetal, and polyamides with spiral drills gives similar shaving shapes, with three characteristic types (folded shavings, helical shavings, lamellar shavings) [4]. Milling of polyamides gives shaving typical of those obtained from machining metals (ribbon shavings, snarl shavings, spiral shavings, helical shavings, crumble shavings, needle shavings, and melted shavings).

3.1.3. Surface Quality

In the qualitative evaluation of the turned surfaces of thermoplastic polymers, depending on shaving conditions, three different characteristic surface structures can arise (Fig. 80): (1) a groove profile, (2) a smeared groove profile, and (3) a groove profile with breakouts. The occurrence of surface breakouts under certain machining conditions during turning was first reported by Spur and Zug [6–8] for polyamides and polyacetal, and for poly(vinyl chloride) and polyacetal resins by Ferichmann [9]. With several materials, the tendency to form surface breakouts is particularly pronounced (e.g., rigid PVC and bulk-polymerized polyamide [9], [10]).

As with metals, the surface roughness value T_t as defined by DIN 4760 can be used for a quantitative characterization of surface quality. It is the maximum vertical separation between the highest and the lowest points of a surface structure. Experimental determinations during turning of polyamide illustrate that the maximum roughness increases with increasing feed, whereby water-containing materials exhibit higher roughnesses than dry materials [6], particularly under conditions of large feed. The surface roughness decreases with increasing cutting angle. For polyamide 6 containing 15 wt % glass fiber and polyamide 6 with 5 wt % graphite, increasing machining time causes an increase in surface roughness, whereas with the extruded materials polyamide 6 and polyamide 6 containing 1 wt % MoS_2, the surface roughness remains unchanged. The time dependence of the surface roughness with glass- and graphite-filled polyamides results from increased tool wear caused by the fillers.

In milling and grinding polyamides, surface melting generally occurs with tooth feeds $f < 0.1$ mm. For $f > 0.02$ mm, the surface roughness increases with increasing feed, whereby the best surface quality is attained with a rake angle of $\alpha = -6°$ to $+6°$. With short machining times of <30 s and cutting speeds $V < 500$ m/min, the best surface quality is obtained with a small clearance angle ($\gamma < 5°$). Longer cutting times and higher cutting speeds ($V > 500$ m/min) require larger clearance angles due to increased warming of the cutting tool, particularly with small tooth feeds. Large stressed cross sections can lead to increased *breakouts* from the surface, proportional to the cutting speed. Surface *breakouts* occur more frequently with greater depth of cut and positive rake angle.

As a consequence of restricted shaving removal and the continuously changing cutting geometry of spiral drills, large surface roughness values result from drilling. The surface quality is influenced mainly by the heat generated during the machining process.

Drilling of polyamide and polyacetal showed that with high levels of feed, lamellar shavings occur that are accompanied by surface breakouts [10]. In contrast, in drilling LDPE, no lamellar shavings occur even at greater feeds, and the surface remains free of breakouts. With bulk-polymerized polyamide 6, polyamide 6 containg 0.5 wt % MoS_2, polyamide 6 containing 25 wt % glass fibers, polyamide 6 containing 35 wt % glass fibers, polyamide 66 containing 25 wt % glass fibers, polyamide 66 containing 35 wt % MoS_2 and polyacetal, the surface roughness increases once a threshold feed value has been exceeded ($f = 0.4$ mm; with poly-

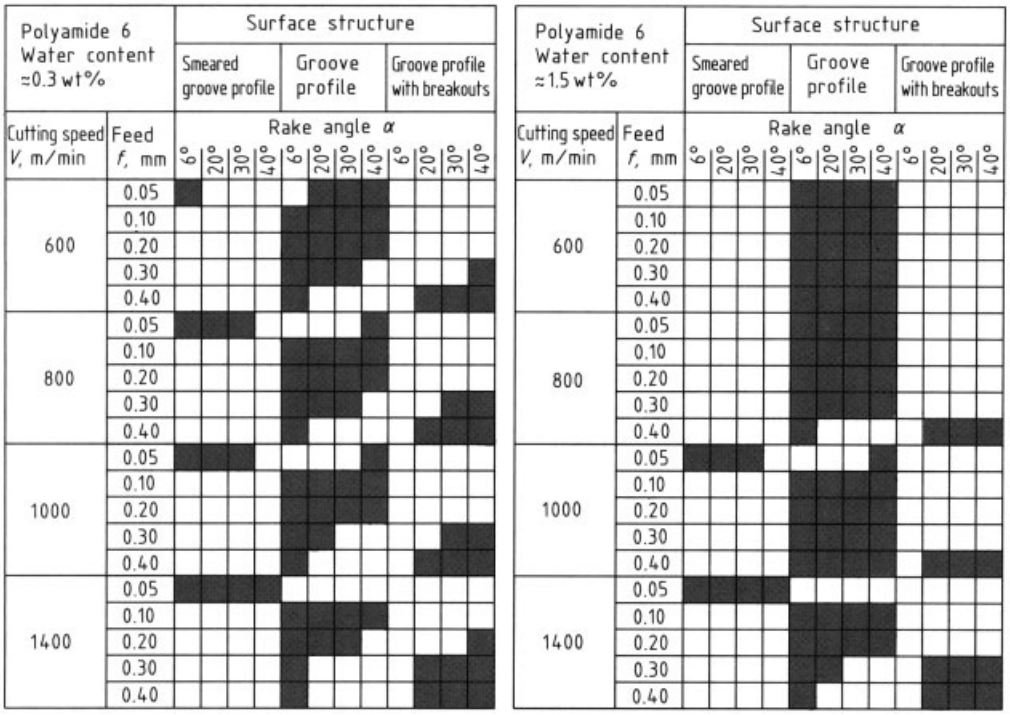

Figure 80. Occurrence of certain surface structures as a function of cutting speed, effective cutting angle, feed, and water content of Ultramid KR 1337

acetal, $f = 0.8$ mm). An exception to this tendency is polyamide 6 containing 5 wt % MoS_2 at a cutting speed V of 40 m/min, and also LDPE. With polyamide the surface roughness decreases with increasing feed, whereas for LDPE the surface roughness is almost independent of feed. Since surface quality is reduced by melting of the bore wall, particularly at small feeds and high cutting speeds, it can be improved by cooling. Although external application of a drilling oil emulsion causes little or no improvement in surface quality, surface roughness can be reduced, especially at high cutting speeds, by cooling with compressed air via a series of internal cooling channels. The best surface quality is obtained at feeds of $f < 0.8$ mm by circulating drilling oil through internal cooling channels.

3.1.4. Cutting Forces

The effective cutting and feed forces involved in machining thermoplastics are generally far lower than those required for metals. The cutting forces that occur on turning polyamides increase almost linearly with increasing feed and depth of cut, if shaving shape is constant. Another significant factor affecting cutting force is the machining angle. Thus, for the same shaving cross section, the cutting force during turning of polyamide at a rake angle of 50° can be 50 % lower than at a rake angle of 0°. This applies only when the moisture content is constant [6]. The correlation is particularly significant for the widely used polyamides (PA 66, PA 6) because the high moisture absorption of these materials leads to considerable changes in mechanical properties. In turning polyamides, the cutting forces are approximately 30 % lower for a material with a moisture content of 3.5 wt % than for a dry sample.

When drilling polycarbonate, polyethylene, poly(vinyl chloride), poly(methyl methacrylate), polystyrene, acrylonitrile – butadiene – styrene (ABS) polymers, polytetrafluoroethylene, polyacetal, polypropylene, and polyamide, the torque and feed force increase linearly or slightly degressively with increasing feed. The speed of the spiral drill has little influence on shaving force [11]. Hoffmann investigated the

influence of cooling lubricants on feed force and torque in drilling bulk-polymerized polyamide 6 and showed that when cooling lubricants are employed, the advancement force increases slightly because the strength of the polymer is extremely temperature dependent [4]. Whereas the torque increases significantly when internal compressed air cooling is employed, the lubricating effect that occurs when drilling oils are used causes it to increase to a considerably lesser extent due to the reduction of the fiber friction at the bore walls. The lowest values are obtained without cooling by using drills equipped with internal cooling channels.

3.1.5. Cutting Temperature and Tool Wear

The temperatures that arise during the machining of unfilled thermoplastics are so low that they do not influence the lifetime of the machining tools. Measurements carried out during drilling and turning have shown that the lowest temperatures occur in the machining of polyethylene, and the highest with polyamide. Under certain machining conditions, temperatures can exceed the melting point of the material being processed. However, due to the low thermal conductivity of polymers, this affects only the surface areas. The cutting temperature is particularly critical when gaseous decomposition products can form, which is especially applicable to nitrogen-containing polymers, phenol – formaldehyde resins, and fluoropolymers [12].

Tool wear is considerably lower in the machining of unfilled polymers than it is with metals [13]. In machining filled polymers, a significantly greater amount of wear occurs compared to unfilled plastics. This is especially true for graphite-filled materials, although glass-fiber-filled materials also cause increased tool wear [4, 6, 14, 15].

3.1.6. Guidelines for Machining Polymers

Tables 7, 8, 9 list guidelines for turning, milling, and drilling thermoplastics and thermosets with hard metal tools. Standardized tools are used for machining polymers, with cutting edges made of high-speed steel, hard metal, or ceramic. For the production of small series, bits with cutting edges made of high-speed steel can be used, whereas for mass production, hard metal is preferred because less wear occurs.

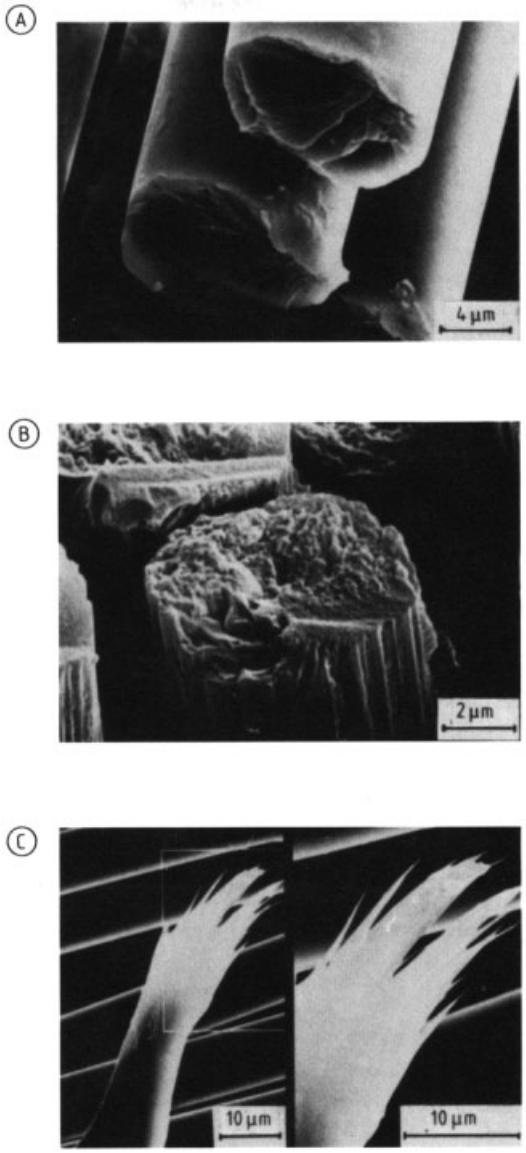

Figure 81. Properties and fracture characteristics of glass (A), carbon (B), and aramid (C) fibers

Table 7. Guidelines for turning plastics

Material	Type of turning	Depth of cut, mm	Cutting speed, m/min	Feed, mm	Cutting geometry, degrees		Attainable surface quality, µm
					Clearance angle γ	Rake angle α	
Thermoplastics							
Polyolefins	rough	3 – 5	250 – 400	0.5 – 1	5 – 15	≦0	6 – 10
	smooth	1 – 3	350 – 600	0.1 – 0.6			
Polystyrene	rough	1.5 – 4	80 – 120	0.1 – 0.2	8 – 12	≦0	3 – 6
	smooth	0.5 – 1.5	100 – 200	0.02 – 0.08			
PVC, rigid	rough	2 – 6	100 – 250	0.4 – 0.5	12 – 18	−5 to 0	3 – 10
	smooth	0.5 – 1	250 – 350	0.1 – 0.2			
Poly(methyl methacrylate)	rough	1.5 – 3	100 – 150	0.1 – 0.2	10 – 15	0 – 5	6 – 10
	smooth	0.5 – 1.5	200 – 350	0.05 – 0.1			
Polyamide	rough	2 – 4	100 – 200	0.2 – 0.5	5 – 15	35 – 45	3 – 6
	smooth	1 – 2	200 – 250	0.1 – 0.2			
Polyacetal	rough	2 – 5	120 – 200	0.2 – 0.3	10 – 15	−5 to 5	6 – 10
	smooth	0.4 – 2	180 – 300	0.1 – 0.25			
Thermosets							
Phenolics	rough	2 – 4	200 – 300	0.2 – 0.3	5 – 10	20 – 25	10 – 20
	smooth	0.5 – 1.5	350 – 450	0.08 – 0.15			
Aminoplastics	rough	1.5 – 3.0	200 – 350	0.1 – 0.15	5 – 10	20 – 25	10 – 20
	smooth	0.5 – 1.0	400 – 500	0.05 – 0.1			
Paper- and fabric-based laminate	rough	2	100 – 180	0.3 – 0.5	12 – 18	15 – 25	6 – 20
	smooth	0.5 – 2	200 – 250	0.1 – 0.2			
UP – GF∗	rough	1.5 – 3	90 – 120	0.1 – 0.2	8 – 10	−5 to −2	10 – 20
	smooth	0.5 – 1.5	120 – 150	0.05 – 0.1			

∗ UP – GF=glass-fiber-reinforced unsaturated polyester resin.

Table 8. Guidelines for milling plastics

Material	Depth of cut, mm	Cutting speed		Feed per tooth		Cutting geometry, degrees		Attainable surface quality, µm
		Rough, m/min	Smooth, m/min	Rough, mm	Smooth, mm	Clearance angle γ	Rake angle α	
Thermoplastics								
Polyolefins	1 – 5	250 – 350	300 – 500	0.2 – 0.3	0.05 – 0.2	15 – 20	−5 to 0	6 – 10
Polystyrene	1 – 5	150 – 200	250 – 400	0.3 – 0.5	0.04 – 0.1	20 – 25	10 – 20	6 – 20
PVC, rigid	3 – 8	250 – 350	350 – 550	0.4 – 0.6	0.05 – 0.3	25 – 30	20 – 25	6 – 20
Poly(methyl methacrylate)	1 – 5	130 – 150	180 – 300	0.1 – 0.2	0.03 – 0.1	25 – 30	20 – 25	4 – 8
Polyamides	1 – 5	100 – 140	150 – 180	0.2 – 0.5	0.05 – 0.1	25 – 30	20 – 25	6 – 10
Polyacetals	1 – 4	120 – 160	150 – 200	0.1 – 0.3	0.03 – 0.1	6 – 14	0 – 8	6 – 10
Thermosets								
Phenolics	3 – 7	150 – 250	200 – 300	0.3 – 0.6	0.08 – 0.4	20 – 30	20 – 25	6 – 10
Aminoplastics	3 – 7	150 – 250	200 – 300	0.4 – 0.8	0.08 – 0.3	20 – 30	20 – 25	6 – 10
Paper- and fabric-based laminate	1 – 4	80 – 150	150 – 250	0.2 – 0.4	0.08 – 0.1	20 – 25	20 – 30	6 – 10
UP – GF ∗	1 – 5	50 – 150	200 – 600	0.1 – 0.2	0.05 – 0.1	10 – 15	5 – 10	10 – 20

∗ UP – GF=glass-fiber-reinforced unsaturated polyester resin.

3.1.7. Processing of Fiber-Reinforced Plastics

The processing of fiber-reinforced plastics (see Section 2.1.5) leads to more problems than the processing of unreinforced or filled polymers.

During the manufacture of a fiber composite component, several processing operations are generally required, in which the properties of the fiber material must be taken into account. Figure 81 gives an overview of the relevant properties and the fracture behavior of glass, carbon,

Table 9. Guidelines of drilling plastics [35]

Material	Average drill speed for a given drill diameter, min^{-1}			Feed, mm	Clearance angle γ, degrees	Rake angle α, degrees	Point angle σ, degrees	Recommended drill oversize, mm	Coolant
	4 mm	10 mm	20 mm						
Thermoplastics									
Polyolefins	4000	1800	1100	0.1 – 0.5	10 – 20	0 – 10	110 – 120	0.15 – 0.2	none
Polystyrene	800	550	180	0.1 – 0.3	6 – 10	5 – 8	49 – 90	0.05 – 0.08	water
PVC, hard	2000	950	630	0.2 – 0.6	10 – 30	0 – 5	110 – 130	0.06 – 0.09	compressed air
Poly(methyl methacrylate)	1000	600	300	0.2 – 0.3	8 – 12	10 – 15	60 – 90	0.05 – 0.08	water
Polyamides	7500	2500	1000	0.1 – 0.4	10 – 12	8 – 10	100 – 120	0.08 – 0.12	compressed air
Polyacetal	4800	2200	1300	0.1 – 0.25	10 – 15		60 – 100	0.08 – 0.10	none
Thermosets									
Phenolics	4000	1600	800	0.1 – 0.3	6 – 8	12 – 24	60 – 100	0.02 – 0.05	compressed air
Aminoplastics	4800	2000	850	0.05 – 0.15	5 – 10	10 – 15	60 – 120	0.06 – 0.08	compressed air
Paper- and fabric-based laminate	3500	1200	650	0.2 – 0.4	15 – 20	10 – 15	90 – 110	0.05 – 0.08	compressed air
UP – GF ∗	9000	5400	2800	0.05 – 0.1	10 – 12	14 – 16	140 – 150	0.03 – 0.05	water

∗ UP – GF=glass-fiber-reinforced unsaturated polyester resin.

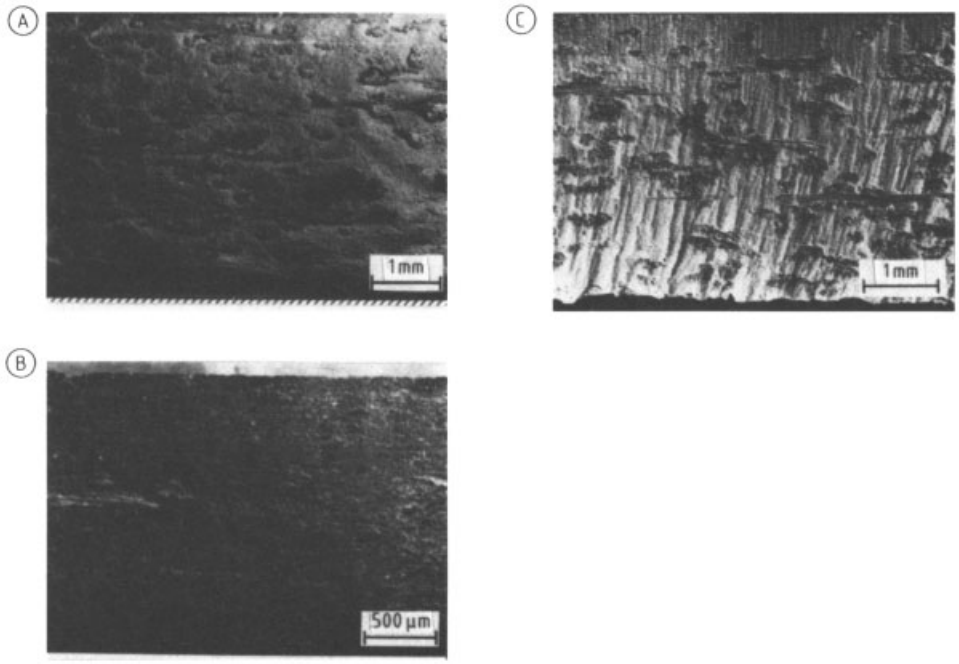

Figure 82. Cut surface of a polyester – glass fiber SMC component after laser cutting (A), milling (B), and water jet cutting (C)

and aramid fibers. The brittle fracture behavior of glass and carbon fibers contrasts with the ductile failure of aramid fibers, which tend to split in the fiber direction because of their molecular structure. In addition to the physical properties of individual fibers, the laminate structure, fiber orientation, fiber length, and fiber content of the product also influence processing characteristics. In general, machining operations are required at the beginning and the end of the manufacturing sequence: to cut the semifinished product to size and to produce the desired final contours. Continuous fiber sheet molding compounds are a widely used starting material. Trimming is generally performed manually with knives, scissors, or special equipment. Even this

processing stage exhibits a range of problems and requires the use of tools adapted to suit the type of fiber employed. Difficulties also arise from adhesion of resin and fiber fragments to the equipment, which prevents a clean cut. Frequently, fibers are pulled out of the matrix, and ragged cut edges result. Prepreg trimming can be improved by using ultrasonically vibrating blades.

After shaping and curing, a finishing operation is generally necessary. This can involve edge trimming and deflashing, as well as generation of dimensionally precise functional surfaces. Finishing can be carried out by classical techniques such as milling, turning, and drilling, or by newer methods such as laser, water-jet, or abrasive water-jet cutting. None of these methods is suitable for all finishing operations. For example, laser and water-jet cutting are not well suited for generating high-quality functional surfaces, but have advantages in edge trimming and deflashing or in introducing *cutouts*.

Figure 82 shows the topography of cut surfaces generated by laser cutting, water-jet cutting, and milling of a sheet molding compound (SMC) material. The condensation of decomposed material and melt residues on surfaces cut by laser are typical of this method. The porous, relatively smooth surface is characteristic. Cracking of the surface indicates that the sample has suffered some thermal damage. Cracks are also apparent in the surface cut by a water jet. These are caused by local mechanical overloading of the fiber – matrix composite as a result of pressure buildup in the sample. Damage occurs mainly to areas in the vicinity of the jet outlet because, in this position, support by underlying material is either partially or totally absent. Characteristic surface features include the almost parallel ridges oriented against the direction of advancement, whose course and size can be influenced over a wide range by the choice of process parameters. The best optical appearance is given by the milled surface. In this case, only slight unevenness is apparent, and the surface is macroscopically free from machining damage. Milling also permits the highest advancement speed. The brittle fracture characteristic of glass fibers is an advantage for mechanical methods such as water-jet cutting or milling, and the large difference between the melting points and decomposition temperatures of the fiber and matrix materials causes problems with processing by use of thermal energy.

Additionally, the high abrasiveness of glass and carbon fibers also causes problems. Milling thus necessitates the use of hard metal tools. Often polycrystalline diamond (PCD) cutting edges are preferred. Despite the considerably higher cost compared to hard metal tools, the use of PCD is often justified by improved processing quality and reduced machine down time due to longer lifetime of the tool.

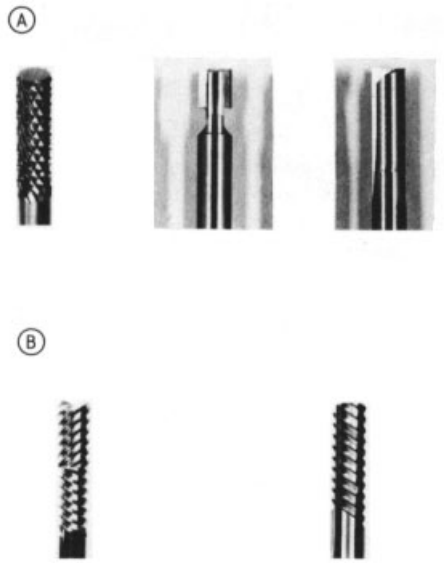

Figure 83. Milling cutters for glass and carbon-fiber-reinforced plastics (A) and aramid-fiber-reinforced plastics (B)

In processing aramid-reinforced materials, the abrasive effects do not cause as many problems as the high toughness of the fibers and hence the ragged appearance of the cut edge. For this reason, opposing screw formed ground hard metal bits are preferentially employed to grind aramid composites in contrast to one- or two-bit heads to grind GRP and carbon-reinforced plastic (CRP) (Fig. 83).

A comparison of the various methods for finishing fiber-reinforced plastics is given in Table 10.

Apart from technological and economic aspects, environmetal aspects of the processes will become more important in the future, including the emission of pollutants and noise.

Table 10. Criteria for evaluating the processing of fiber-reinforced plastics

	Process			
Criterion	Water-jet cutting	Abrasive water-jet cutting	Laser cutting	Contour milling
Mechanical stress	low	low	very low	medium
Thermal stress	very low	very low	very high	low – medium
Pollution	none	none	very high	high (dust)
Noise	high	high	medium	very high
Width of cut	low	medium	low	tool diameter
Shape of cut	slightly conical	conical	slightly conical	any
Minimum radius, mm	0.5 – 1.0	1.5 – 2.0	0.2 – 0.5	tool diameter
Symmetry of cut edge	parallel to jet direction	parallel to jet direction	parallel to jet direction	perpendicular to tool axis
Tool wear	low	medium – high	very low	medium – high
Acessibility of the workpiece	both sides	both sides	both sides	one side

3.2. Bonding

Bonding processes for polymers can be categorized into three groups:

1) Welding
2) Adhesion
3) Mechanical bonding

A designer employing polymers has a limited choice of bonding process because fully cross-linked polymers cannot be welded, and several non-cross-linked polymers can be adhered to only with extreme difficulty or to a certain degree. All polymers can be mechanically bonded to each other or to other materials, (e.g., by riveting, bolting, or clips). However, with fiber-reinforced polymers, because of anisotropy, clip and plug bonds are problematic, and the sensitivity to notch effects places constraints upon riveting and bolting.

Welding and adhesion yield permanent bonds; this also applies to riveting. All other mechanical bonds can be regarded as reversible. The choice of bonding technique is determined by processing factors, particularly bonding time. Welding and mechanical bonding require little time, whereas adhesion necessitates longer periods to allow for surface pretreatment and hardening of the adhesive.

3.2.1. Welding

Only polymers that can be softened or melted by heating can be welded (i.e., non-cross-linked polymers) [16, 17]. Exceptions to this include poly tetrafluoroethylene, heat-resistant polyaromatics such as polyimides and polybenzimidazoles, and ultrahigh-molecular-mass thermoplastics (e.g., UHMW-PE and cast PMMA), whose softening or melting temperature lies above their thermal decomposition temperature.

During welding, the polymers are heated in the region surrounding the weld seam until they pass into the plastic state. They can then be joined under pressure, whereby macromolecules of the two substrates undergo mutual diffusion processes.

Apart from possible thermal chain degradation, chemical reactions do not take place during polymer welding. To obtain an acceptable, materially intact joint, the material flow process in the plane of the seam must take place at a temperature comparable to that of the original forming process (injection molding, extrusion, etc.).

Various welding techniques are generally classified according to the way in which energy is introduced:

1) Thermal conduction
2) Convection
3) Radiation
4) Friction
5) Induction

The following welding techniques are commonly used today (the energy source involved is given in parentheses):

1) Heated-tool (HT) welding (thermal conduction)
   a) Direct heating of seam area
      - HT butt welding
      - Heated wedge welding
      - Welding by bending
      - HT socket welding
      - Electric socket welding

b) Indirect heating of seam area
   Thermal sealing
   Heat impulse welding
2) Hot-gas (HG) welding (convection)
   HG area welding ∗
   Draw welding ∗
   HG extrusion welding ∗
3) Radiation welding (thermal radiation)
   HT radiation welding
   Laser welding
   Laser extrusion welding ∗
4) Friction welding processes (friction)
   a) Heating via internal friction within the material
      High-frequency welding
      Ultrasonic welding
   b) Heating via external friction
      Rotation friction welding
      Vibration welding
5) Induction or electromagnetic welding (induction) ∗

(Asterisked processes require the use of welding rods)

3.2.1.1. Heated-Tool Welding

In HT techniques the welding zone is heated conductively by electrically heated metallic elements. In direct HT welding, heat flows directly from the heating element to the weld zone; with indirect HT welding, heat transport proceeds from an external source through the welding substrate to the welding zone. Because of the poor thermal conductivity of polymers, indirect HT welding is used only for thin materials (films).

In *direct HT welding*, both welding partners are heated by the same heating element. For this, the components are mounted in their respective welding positions on two mechanisms that can be closed against each other. The heating element is introduced between the welding substrates, and the sections to be heated are pressed against the front and back surfaces.

Once the desired temperature has been reached, both welding partners are lifted briefly from the heating element, which is then removed, and the welding process is completed by contacting the welding partners under pressure.

In all HT welding processes, the following parameters influence the quality of the weld seam:

1) Melt length (reduction in dimensions of the welding substrate during heating phase)
2) Heating time
3) Time required to remove heating element
4) Seam pressure
5) Holding time

Heated-tool butt welding is simple to automate, which is essential for reproducible welding seam quality. It requires short cycle times and is particularly suitable for welding partners with large weld seam areas (solid preforms, pipes, window frame profiles, transport pallets) and also for fine welds (e.g., taillights for automobiles). Due to melting of the bonding partners, a so-called weld bead is usually formed, which in some cases must be removed in a separate step. In the most favorable cases, the strength of the weld can reach the strength of the bonded materials.

Heated-wedge welding is a continuous bonding process for polymer films and plastic-coated fabrics (Fig. 84A). In the manual process, the heated metal wedge and the pressure roll are operated by hand. In the mechanized version, the heated wedge and the transport and pressure roll are fixed, and sheets to be welded are drawn over the wedge. The thickness of overlapping films can reach 0.4 mm. Welding speeds up to 15 m/min can be reached.

Welding by bending is a manual technique used for sheets and plates (Fig. 84B). The wedge-shaped tip of the heated tool is inserted three-quarters of the way into the welding substrate and is removed after a period of heating (essentially pressureless). The sheets are welded together by bending. To ensure that sufficient welding pressure can be introduced, the angle at the tip of the heating tool should be approximately 20 % smaller than that at the opposing end. To reduce the heating period for thick sheets, a wedge-shaped groove can be milled into the sheet.

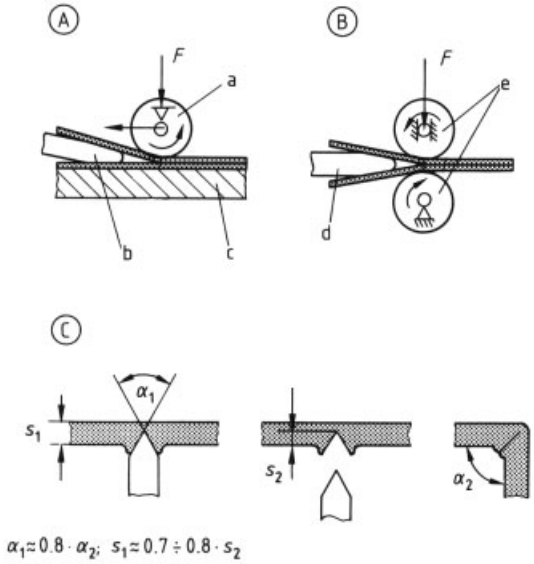

Figure 84. Manual (A) and mechanized (B) heated wedge welding, and welding by bending (C)
a) Pressure roll (manually operated); b) Heated wedge (movable); c) Support; d) Heated wedge (fixed); e) Transport and pressure rolls

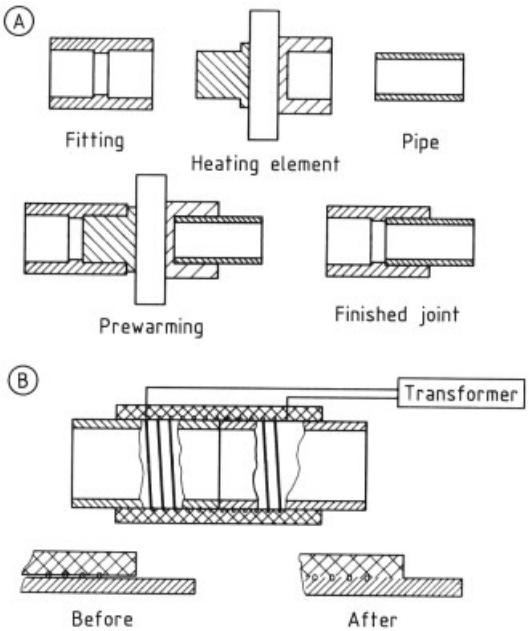

Figure 85. Heated-tool socket welding (A) and electric socket welding (B)

Heated-tool socket welding is a process for bonding pipes with sockets by means of an overlapping weld seam (Fig. 85A).

The heating element has the inner contours of the socket on one side and the outer contours of the pipe on the other. The dimensions of the heating element, pipe, and fitting are such that after heating, removal of the heating element, and repositioning, pressure is built up at the weld seam during welding. The technique is employed almost exclusively for pipes made of PE and PP.

Electric socket welding is another technique for bonding pipes. The injection-molded socket has electrical heating coils incorporated in the inner surface, which, after positioning at the end of the pipe, are connected to a welding transformer (Fig. 85B). The inner surface of the sleeve and the outer surface of the pipe are plasticated. Because the sleeves are generally injection molded with intrinsic orientation, shrinkage occurs on heating, which generates the radial pressure required for welding. The coils remain within the bond.

An *indirect heated-tool* welding technique is thermal contact welding, which finds widespread application for moldings and films of low wall thickness (<1 mm) (Fig. 86). The surface of the continually heated element can be structured to permit combination of the welding process with embossing. Areas of application include packaging technology (heat sealing, welding on lids, closing plastic sacks, etc.). With heat impulse welding, the heating element is not heated continually (Fig. 86B). At the onset of each welding phase, thin metal rods coated with an antiadhesive layer [generally polytetrafluoroethylene (PTFE)] are subjected to a short but powerful current impulse and are thus heated. The heat passes through the film to the weld zone by conduction. In the two-sided process, the temperature profile in the weld zone is symmetrical because of the arrangement of the heating bars, but it is asymmetrical in the one-side process. The main application is in packaging technology, although this technique is not suitable for welding large areas because of the heaters used (narrow bus bars). Because the welds cool via the ram and counterram, cycle times are shorter than in heat contact welding.

In all indirect HT welding techniques, a temperature profile across the thickness of the weld

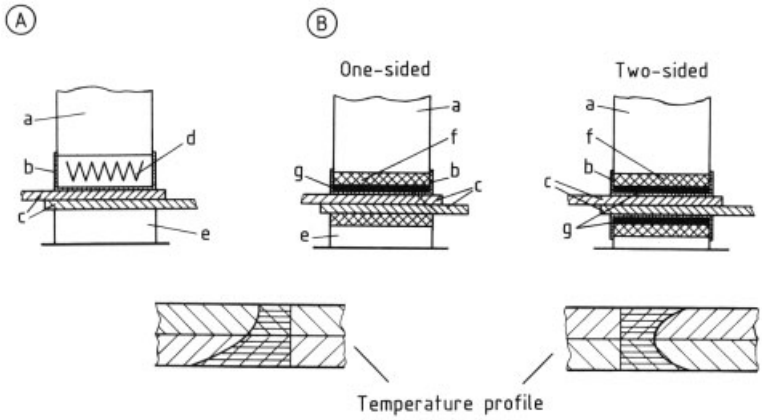

Figure 86. Thermal sealing (A) and heat impulse welding (B)
a) Ram; b) Antiadhesive coating; c) Films; d) Heating element, continuously heated; e) Counterram; f) Thermal insulation; g) Impulse heating bar, discontinuously heated

seam is formed during heating, with the highest temperature on the surface nearest the heating element. The welding parameters (HT temperature and heating time) must therefore be adjusted so that a sufficiently high temperature is reached in the welding plane, without overheating the polymer on the surface.

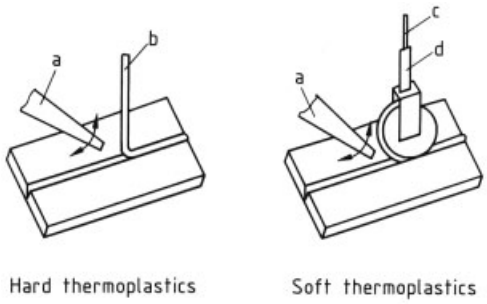

Figure 87. Hot-gas welding
a) Hot gas; b) Welding rod (manually guided); c) Welding rod; d) Pressure roll (manually guided)

3.2.1.2. Hot-Gas Welding

Hot-gas welding is usually carried out manually.

The heat-transfer medium is generally hot, clean compressed air; for oxidation-sensitive materials, inert gases such as nitrogen or carbon dioxide are also used. The surfaces to be joined are heated via the hot gas stream and welded under pressure, usually with welding rods (Fig. 87). This method is particularly suitable for assembly and repair, as well as machine and container construction. Use of hot-gas welding is recommended when access to the weld seams is difficult or when the materials concerned are difficult to weld.

The most important requirements for high weld quality are constant welding pressure, gas temperature, and gas flow rate, as well as precise guiding of the gas nozzle and welding rod.

Three variations of hot-gas welding exist.

In swirl welding, the welding substrate and welding rod are heated by a swirling motion of the circular hot-gas die. The welding rod (a few millimeters in diameter) is fed perpendicular to the surface to be welded (Fig. 87). With soft weld rods, a pressure roll is used to avoid folding of the welding rod.

In draw welding, the welding rod is warmed separately in a specially formed hot-gas nozzle. The metal housing removes heat from the gas stream and transfers it to the welding rod by conduction. The welding rod leaves the die in a plasticated state and is pressed onto the basic material by a pressure lip (Fig. 88B). Higher welding speeds are possible than with swirl welding.

Extrusion welding is mainly employed to bond thick-walled moldings because deep weld grooves can be filled in a single operation. The substrate is warmed with hot gas during which time the welding rod is fed as a plasticated ribbon directly from an extruder and pressed into the groove by a welding lip (Fig. 88A).

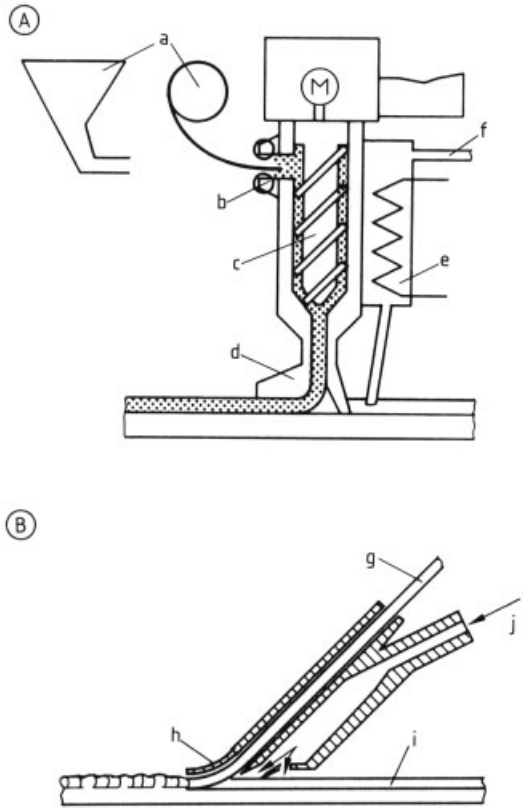

Figure 88. Extrusion (A) and draw (B) welding
a) Pellet hopper or wire reel; b) Wire feed; c) Extruder; d) Welding shoe; e) Heater; f) Air supply; g) Welding rod; h) Pressure lip; i) Part to be welded; j) Hot air

3.2.1.3. Radiation Welding

Radiation welding techniques can generally be classified into two groups: those in which the thermal radiation from a heating element is utilized and those that employ lasers.

With HT radiation welding, the welding surfaces are not brought in contact with the heating element. Instead, heat transport from the heating element to the weld region takes place across a 0.5 – 1 mm air gap, mainly by radiation and partly by convection. The process is particularly advantageous for polymers that tend to adhere to the heating element and be drawn into fibers. To accelerate the heating process, radiation heating elements are generally operated at higher temperature than those employed in conductive methods. Care must be taken not to overheat the surfaces to be welded.

In laser welding, heating of nontransparent polymers is performed with a laser. This process is employed only in special cases for films and in winding fiber-reinforced polymer components with a thermoplastic matrix.

In laser extrusion welding, the substrate is heated by a laser instead of the hot gas used in conventional extrusion welding.

3.2.1.4. Friction Welding

Friction welding utilizes the heat of friction to plasticate the welding surface. A distinction is made between external and internal friction. Whereas the relative motions of the welding components cause heating via external friction, the electrical and mechanical damping characteristics of the material and the molecular vibrations lead to heating by internal friction.

Spin welding, a technique employing external friction, is suitable only for bonding components that are rotationally symmetrical. One of the welding substrates is rotated while the other is mounted on a bearing capable of rotating about the same axis, but initially prevented from doing so by a brake, and is lightly pressed against the rotating partner (1 – 2 N/mm^2). The relative motion (0.8 – 2.5 m/s) causes heating of the welding surfaces and finally melting. Once a large enough melt reservoir has been generated, the brake is released, and both components rotate at the same speed. If one component is particularly large or does not have rotational symmetry, rotation of the substrate can be stopped for welding. The weld must be cooled while still under compression. The technique can be carried out straightforwardly on a lathe. A disadvantage is the long setup time.

In vibration welding, direct heating of the weld plane is brought about by relative motion of the two partners while under pressure. An oscillating rotational (with rotationally symmetrical welding surfaces) or translational motion (frequencies of 100 – 300 Hz) is used. Important process parameters that must be adjusted to suit both the geometry of the articles to be welded and the material itself are

1) Friction speed or frequency and amplitude (up to 2.5 mm)

2) Friction time
3) Friction pressure (up to 4 N/mm^2)
4) Weld seam pressure
5) Holding time

Both processes are simple to automate and have short cycle times (ca. 5 s), so they are ideally suited for mass production operations. Additionally, vibration welding techniques can be used in the manufacture of hollow articles (e.g., brake fluid reservoirs and pressurized containers for automobile coolants) and also for bonding larger items (e.g., automobile bumpers) even when relatively soft thermoplastics are used.

A technique that utilizes the internal friction in a material is *ultrasonic welding* at a frequency of 20 – 60 kHz. The mechanical oscillation of a piezoelectric vibrator is introduced into one of the two components to be welded via a booster and a sonotrode. The components are mounted between the sonotrode and a stationary anvil. The sonotrode is shaped so that a standing mechanical wave is generated with its maximum amplitude at the contact surfaces of the two components. With appropriate shaping of the contact surfaces (energy directing), heating of the material by internal friction is rapid and concentrated in the contact area. Hence, short cycle times can be attained (0.2 – 1.5 s).

Many variations of this welding technique exist. The characteristic of near-field welding is the close proximity of the sonotrode contact surface and the welding surface (<6 mm). The method is employed to bond films and sheets of almost all thermoplastics. Only rigid polymers [such as styrene – acrylonitrile (SAN), ABS, PMMA, polyoxymethylene (POM), PC, and polystyrene (PS)] are suitable for distant field welding, where the energy losses are not too great across the longer ultrasonic pathway between the sonotrode contact and the welding surface. The main area of application of distant field welding involves the bonding of moldings (sometimes made from different polymers). The major process parameters that must be adjusted to suit the specific material and geometry of the welding components include

1) Sonotrode shape
2) Contact pressure
3) Amplitude of oscillation and ultrasonic power
4) Irradiation and holding time

A related process involves embedding metal inserts in thermoplastics (e.g., screw thread sockets) by mounting the metallic component on the sonotrode and pressing it into the polymer under the influence of ultrasonic irradiation. Ultrasonic equipment is also employed for riveting; a shaped sonotrode is used to form the head on a thermoplastic rivet.

Ultrasonic welding is suitable for use in mass production bonding operations and gives highly reproducible weld seam quality. By appropriate design of the welding surface, weld seam bead can be concealed behind undercuts and hence need not be removed in a finishing stage.

In high-frequency welding a polar material is heated by the molecular dipole oscillations generated by the alternating electric field of a capacitor (frequency ca. 27 MHz). By suitable shaping of the capacitor electrodes, heating can be limited to the immediate vicinity of the welding surface. Cycle times are then short. Electrodes are also employed to apply the bonding force to the welding partners. Specially shaped electrodes can be used to emboss the weld seam or to generate a weld seam with a cutting edge.

Suitable materials are polymers with dielectric loss factors greater than 0.01 [e.g., cellulose acetate (CA), PA, PMMA, polyurethane (PUR), and PVC]. This technique is used mainly for large-area welding of sheet materials. Product examples include PVC rainwear and inflatable life rafts.

High-frequency welding can be automated and used in mass production.

3.2.1.5. Induction Welding

In induction welding, weld fillers that contain electrically conducting particles are heated by eddy currents induced in the particles by an alternating electromagnetic field (frequency range 3 – 10 MHz) (Fig. 89). The weld filler is a (profiled) preform placed in a groove situated in the bonding surface. Welding takes place in an induction field with the components held under slight pressure. The composition and conducting particle content of the filler must be selected to suit the substrate to be welded. The filler can also be applied to the substrate during the manufacturing process by using two-component injec-

tion molding (see Section 1.4.5). Because heating of the substrate takes place by conduction, cycle times are longer than with high-frequency or ultrasonic welding.

The induction process is also used to open welded seams by placing the welded article in an induction field (e.g., opening containers for transporting corrosive materials).

3.2.2. Adhesion

In contrast to welding, all polymers can be bonded to themselves, each other, and other materials, such as wood, textiles, glass, and metals, by using adhesives [18–20], with the exception of solvent adhesion, which strictly speaking should be regarded as a welding process. During adhesion the substrates are not brought directly in contact with each other. In this case, a layer of similar or dissimilar polymer that can partially diffuse into the substrate surface is responsible for the bond. No or very little heat is necessary for adhesion, so that the structure and properties of the bonding substrates remain unchanged. In contrast to welding, adhesion can also be used for large-area bonding, a fact that is of considerable importance in the packaging industry. Adhesive bonds can be optically superior to welded bonds, which is important with transparent polymers. However, the manufacture of strong adhesive bonds often requires multistage processes and can take several days. The strength of polymer adhesive bonds can be nondestructively tested only to a very limited extent.

A wide range of adhesives with various adhesion and curing mechanisms are used for bonding plastics (see → Adhesives).

Some polymers can be softened by applying solvents and then fused together. Strictly speaking, this technique should be regarded as diffusion welding. Solvent-based adhesives that harden by evaporation of the solvents are also frequently employed. The solvents can also cause the surface of the bonding substrates to swell and hence can facilitate mutual diffusion of adhesive molecules and the molecules of the substrate. This is possible only with thermoplastics. Although adhesion with solvents or solvent-based adhesives is simple and economical, problems often arise with this process due to aging of the bond because the solvent cannot be removed completely from the joint after bonding.

Heat sealing and melt adhesives have also gained importance in bonding plastics. The highest adhesive bond strengths with polymers are generally obtained with chemically hardening one- or two-component adhesives. In most cases, single-component adhesives require heating for the curing reaction, whereas two-component systems also cure at room temperature.

Adhesion processes can now be extensively automated. Continuous adhesion process and short cycle times can be realized, particularly when heat sealing and melt adhesives are used.

Polymers can be classified by their adhesive bonding behavior as

1) Easily bondable polymers
2) Limitedly bondable polymers
3) Difficulty bondable polymers

The differences in bondability are due to variations in the solubility of the polymers and the polarity of the macromolecules (Table 11).

Table 11. Polarity and solubility of plastics

Plastic	Polarity	Solubility	Adhesive capacity
Polyethylene	nonpolar	very slightly soluble	poor
Polypropylene	nonpolar	slightly soluble	medium
Polytetrafluoroethylene	nonpolar	insoluble	very poor
Polyisobutylene	nonpolar	readily soluble	good
Polystyrene	nonpolar	soluble	good
Poly(vinyl chloride)	polar	soluble	good
Polyterephthalates	very polar	insoluble	medium
Poly(methyl methacrylate)	polar	soluble	good
Polyamide 66 – 610	polar	slightly soluble	medium
Polyamide 6 – 11	polar	slightly soluble to insoluble	medium to poor

Further influences are the surface tension and hence wettability. Table 12 lists surface tensions for some polymers and adhesives. A prerequisite for adhesion is that the surface tension of the adhesive be lower than or equal to that of the polymer. Only when this applies does the binder adapt itself to the contours of the solid surface during the wetting stage so that the physical and chemical interfacial reactions required for good

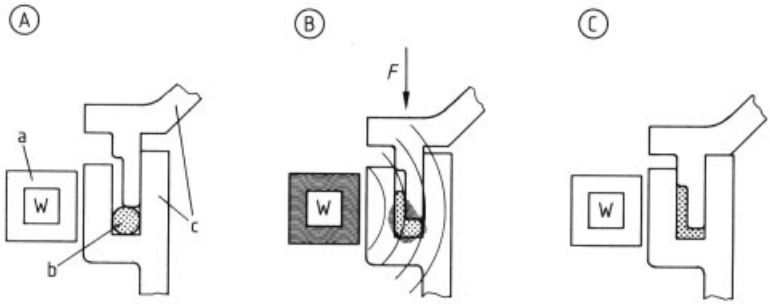

Figure 89. Induction welding
A) Before welding; B) During welding; C) After welding
a) Induction coil; b) Weld filler; c) Parts to be joined

adhesion can occur. Only in solvent-based adhesives containing strongly dissolving components can the wetting criterion be neglected to a certain extent. If the surface tension of the plastic is lower than that of the adhesive, the surface must be mechanically prepared or, in the case of difficultly bondable polymers, chemically pretreated.

Table 12. Surface tension of plastics and adhesives

	Surface tension, 10^{-3} N/m
Plastics	
Polytetrafluoroethylene	18
Polyethylene	31
Polystyrene	33
Poly(vinyl chloride)	40
Polyester	43
Cellulose film	45
Polyamide 66	46
Adhesives	
Acid-curing phenolic resins	78
Urea – formaldehyde	71
Phenol – resorcinol resins	48
Casein resins	47
Epoxy resins	30 – 47
Poly(vinyl acetate) rubber	38
Nitrocellulose	26

Easily bondable polymers are: poly(vinyl chloride) with low plasticizer content, polystyrene, polyacrylates, polycarbonates, and polyurethanes, as well as polyester, epoxy, and phenolic resins in unreinforced and reinforced state. Whereas polystyrene can be bonded most easily with solvent-based adhesives, the remaining members of the group can be bonded reliably by a variety of adhesives without pretreatment or after mechanical roughening. The surfaces of the bonding substrates should be cleaned with solvents or mechanically even when easily bondable polymers are involved, to remove contaminants and residual release agents.

The limited bonding polymers include PVC with high plasticizer content, which must be thoroughly cleaned with organic solvents prior to application of the adhesive. The pronounced shrinkage and swelling behavior of PVC can lead to difficulties in using solvent-based adhesives. Plasticized PVC can be bonded quickly and reliably by using solvents alone (e.g., tetrahydrofuran). Greater problems are encountered with adhesion to synthetic rubbers and polyamide. The latter can be diffusion bonded by using formic acid; otherwise, two-component adhesives are required, which however, give strong bonds only if the surface of the polyamide substrate has been mechanically roughened. Furthermore, the adhesion to polyamide is strongly dependent on the chemical structure of the polymer.

Difficult-to-bond polymers include all polyolefins, fluoropolymers, polyacetals, and silicone resins. These polymers cannot be bonded in the untreated state. Polyolefins can be chemically etched and treated electrostatically by using corona discharge equipment or treated with an oxidizing flame (see Section 3.3.5.1). Care must be taken that the basic structure of the plastic is not altered by surface pretreatment. The greatest difficulties are encountered in adhesive bonding fluoroethylene – propylene copolymers and polytetrafluoroethylene, which can be bonded with adhesives only after etching with sodium in liquid ammonia or sodium naphthalenide. Although both processes are complicated, they give bonds of high mechanical strength and good aging resistance.

Developments in adhesives are concerned with improving bond strength, resistance and stability, and simplifying processibility. An example involves single-component reactive adhesives based on cyanoacrylates, which are ideally suited for bonding rubber and whose hardening takes only seconds or minutes. High manufacturing speeds can also be attained with heat sealing and melt adhesives, whose major advantage is their freedom from solvents. Melt adhesives based on polyamides, linear polyesters, or aromatic polysulfones exhibit good thermal stability and have processing times of a fraction of a second. Of the two-component reaction adhesives, flexible polyurethanes have a broad range of uses for bonding polymers, including construction bonding and the manufacture of laminates such as packaging materials, light core sheets, film-laminated sheet metals, and metal-coated polymers.

3.2.3. Mechanical Bonding

Mechanical bonding processes include the conventional methods of bolting and riveting. The bolts and rivets can be made of the same or different materials with respect to the bonding substrates [21, 22]. The use of self-tapping screws avoids cutting threads in a previous process stage. Screw bonds that are loosened frequently and have a diameter of >4 mm should be equipped with a sunken or embedded metallic thread socket. Rivet bonds can be prepared by using the techniques employed in the metalworking industry. The rivet heads can be formed either by hot forming or cold pressing. Cycle times for riveting processes are short, and the mechanical strength of riveted bonds is satisfactory. Snap-in rivets are inserted in rivet holes and fixed to form a solid bond by spreading of the rivet shaft.

A particularly elegant bonding variation for polymers involves the so-called clamp or snap-in connectors in which protrusions or hooks become interlocked in correspondingly shaped undercuts when the two components are brought in contact. Bonds of this type can be rendered air- or watertight by use of gaskets or sealants. Mechanical bonds of this form can also be formed between polymers and other materials.

3.3. Surface Treatment

The surfaces of plastic components are treated to meet functional or decorative requirements that cannot be met by the base material alone [23, 24]. Surface treatment processes can be classified as follows:

1) Mechanical processes such as hot embossing and flocking
2) Painting and printing with solvent-based or chemically binding coatings
3) Wet chemical deposition processes such as electroplating or electroless metal deposition
4) Physical deposition of inorganic materials
5) Dry chemical processes such as fluorination, sulfonation, and plasma polymerization

3.3.1. Mechanical Processes

3.3.1.1. Hot Embossing

In hot embossing, a multilayer hot embossing sheet is applied to a substrate under pressure by a heated stamp. The term "embossing" is misleading in that it suggests the remolding of the substrate surface, which is not the case. Descriptions such as (color) film printing, (color) hot film printing, or dry printing are more apt, since, because of the way in which the (multilayer) decoration is applied, the initial surface topology remains unchanged, with the only exception being the gloss. Hot embossing can be used to coat the entire surface of an article (equivalent to painting), along with applying letters or a fine pattern of a different color to the substrate (e.g., scales and labeling on dials).

The hot stamp itself has the pattern to be embossed in relief on its surface (Fig. 90). The raised areas activate the hotmelt adhesive layer, which is located nearest the substrate during the stamping process; hence adhesion of the stamping laminate results in these areas. The deeper-cut areas of the stamp do not come in contact with the film, so no heat transfer and hence no activation of the adhesive occur, and the decorative and adhesive layers of the film are removed when the stamp is removed from the substrate.

A relief structure already present in the substrate surface can be highlighted by color by

stamping with a flat stamp (relief stamping). With counterstamping, transparent materials are embossed on the reserve surface.

In the displacement stamping process, the stamp is generally planar or can be adapted to the slightly convex or concave shape of the article. In roll stamping, which is especially suited for large-area hot embossing and relief stamping, the stamp has the form of a roll that proceeds across the surface of the substrate with the stamping film positioned between them (Fig. 91).

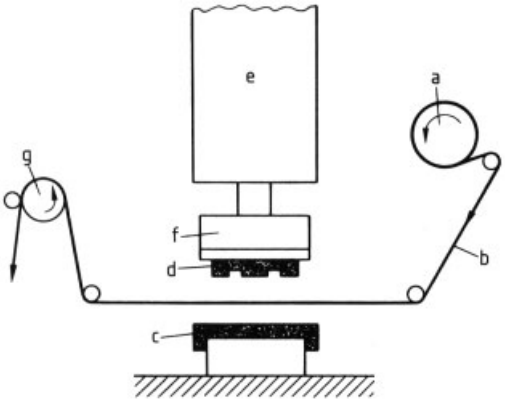

Figure 90. Hot embossing
a) Film roll; b) Film; c) Workpiece; d) Embossing stamp; e) Press; f) Heater; g) Feed roll

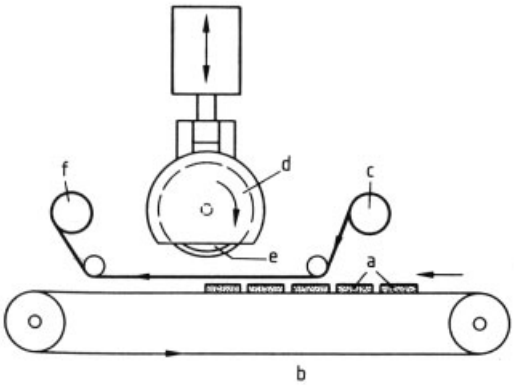

Figure 91. Roll stamping process
a) Workpieces; b) Conveyor belt; c) Roll of film; d) Heater; e) Embossing roll; f) Film windup

As a rule, a thermoplastic substrate surface is necessary to obtain a satisfactory bond with the embossing film. The required thermoplastic surface can be produced on thermosets, metals, ceramics, or wood by applying paint coating. Embossing films are available in a multitude of decors, as mat or gloss films, in wood colors, single or multicolored, metallized (Fig. 92), etc.

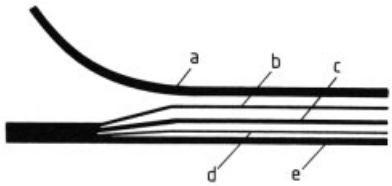

Figure 92. Construction of a metallized film
Layers c – e make up the embossing film.
a) Substrate sheet; b) Release film; c) Protective coating; d) Metallic layer; e) Adhesive layer

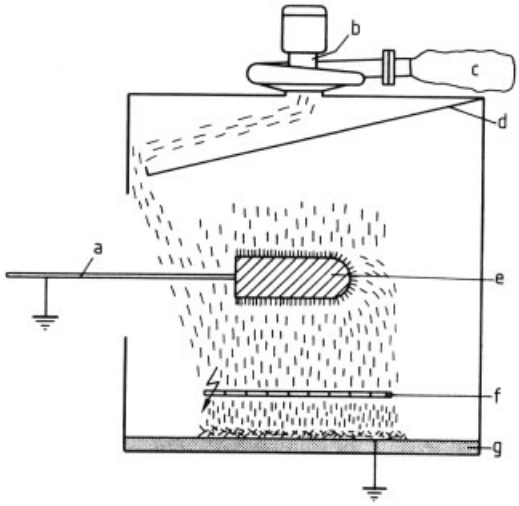

Figure 93. Electrostatic flocking
a) Rod; b) Extractor fan; c) Filter bag; d) Air-guiding plate; e) Molding; f) Electrode; g) Bottom tray

3.3.1.2. Flocking

Flocking is used to produce a velvetlike surface. Colored fibers 0.3 – 0.5 mm in length are used, made from polyamide or viscose, for example. In mechanical flocking processes, which use vibration or compressed air, the flocks form an unoriented and nonuniform layer on the surface. For this reason, electrostatic processes, some-

times supplemented by mechanical techniques, are almost exclusively employed (Fig. 93). To ensure alignment in the electrostatic field, the flocks are produced in a free-flowing, conducting form. Prior to flocking, the substrate surface is coated with an adhesive in which the flocks become permanently anchored. Sometimes, surface pretreatment is necessary (see Section 3.3.5.1). Equipment is available for flocking films, sheets, moldings, and hollow articles employed in the automobile industry, the packaging sector, toys, fashion jewelry, etc.

3.3.2. Painting and Printing

Whereas printing of plastic articles is carried out almost exclusively for decorative purposes, the reasons for painting polymers can be divided into three categories:

1) To meet functional requirements
2) For decorative purposes
3) As an intermediate step prior to further surface modification techniques, such as printing, film embossing, flocking, and metallizing

3.3.2.1. Painting

Previously, it was believed that, in contrast to metals, polymers did not have to be painted because they were not prone to corrosion. However, this situation has changed considerably. Many plastic components can meet their respective specifications regarding optics, light, and weather resistance only when painted.

The thickness of the paint coatings varies between 5 and several hundred micrometers, depending on intended application.

When painting properties are considered, distinctions are drawn between thermosets and thermoplastics, between solvent-sensitive and solvent-resistant plastics, between solid and foamed polymers, and between foams with smooth and those with open-pored surfaces. The polymer and the coating must be matched with regard to their chemical compatibility, adhesion, and elasticity.

Surface State, Pretreatment. Critical demands are placed on the surface of the substrate to be painted and also on the design of the molding itself. The molding must be designed to suit the painting process (i.e., without sharp corners, edges, angles, or deep recesses and undercuts). Due to surface tension effects, paints recede from sharp edges (edge recession) and form a thin coating at these critical wear positions [26]. Undercuts and deep recesses are insufficiently coated in most spray painting processes. With larger planar areas, problems are caused by dust particles, especially with high-gloss coatings and metallization.

The surface to be coated must be clean (i.e., free from grease, dirt, sweat, and residual release agents, particularly those containing silicon). Contamination decreases adhesion and causes surface defects such as spots, dull or glossy areas, and flow disturbances. When necessary, cleaning procedures must be applied: cleaning with trichloroethylene vapor with compact thermosets, mixtures of organic solvents with degreasing and antistatic action, possibly supplemented by ultrasonic techniques or aqueous baths containing surfactants. Electrostatic charging, which causes wetting and flow problems and dust contamination, can be neutralized by blowing with ionized air.

With injection-molded articles, release agents and lubricants may concentrate in and around the weld lines and thus lead to surface defects in the coating. Stresses are often frozen into the molding due to poor selection of injection conditions and inappropriate choice of gate type or gate position [27]. In particular, in solvent-sensitive materials regions of increased internal tensile stress can lead to stress corrosion, deformation, and even loss of mechanical strength. Similarly, the type and amount of pigment in a polymer as well as choice of filler can influence results. Successful surface coating therefore requires knowledge of the type of polymer material, the process parameters for producing the molding, the quality of the coating material, and its processing conditions.

Besides these generally applicable aspects, several polymers require further specific measures prior to coating [26, 27]. Polyethylene, polypropylene, and epoxy resins are nonpolar and have very low surface tensions. To render them wettable and achieve satisfactory adhesion

of a coating, they must be oxidized (see also Section 3.3.5.1) immediately before the coating is applied by

1) Low-pressure methods
2) Plasma treatment
3) Electrical methods at ambient pressure, such as corona treatment
4) Thermal processes with a strongly oxidizing flame (Kreidl – Traver process)
5) Immersion in chromic acid – sulfuric acid

However, antioxidants and stabilizers can reduce the effectiveness of such measures. With polypropylene and ethylene – propylene elastomers, the required paint adhesion can be attained by using special primers. Plasticized PVC and cellulose esters often contain plasticizers that can migrate into coatings and soften them. They must therefore be coated with paints that are resistant to migration of plasticizer.

Application Methods and Paint Systems. All nonelectrostatic coating methods (see Section 1.3) are in principle applicable for use with polymers: high-pressure spraying at 200 – 600 kPa with air atomization, high-pressure airless spraying at 10 – 20 MPa, dipping, flow coating, roller application, pouring, centrifugation, drum application, and brushing (see also → Paints and Coatings, Chap. 8.3.). Electrostatic spraying can be used with nonconducting polymers by coating them with a conducting solution [26, 28].

Apart from economic aspects, the form of the molding and the requirements placed on the surface, as well as the temperature and solvent sensitivity of various polymers, must be considered in selection of the coating method. Curing of the coating is generally accelerated by heating. Manufacturers' information on the heat resistance of the plastics is of limited value here. Heat resistance is determined on standard test specimens, whereas commercially manufactured moldings are seldomly produced under optimal conditions and have internal stresses and variations in wall thickness, and the coating and curing process often takes place under conditions of mechanical loading. This can cause deformation, losses in dimensional stability, and surface defects due to excessive penetration of solvents. Table 13 lists empirical heat resistances for fabricated plastic components. Bubble formation in the coating is possible at elevated temperature as a result of the evolution of formaldehyde from solid thermosets and from residual blowing agents in foamed thermosets.

Table 13. Empirical short-term heat resistance of fabricated plastic components

Plastic	Heat resistance, °C
Polystyrene – butadiene	50
Polystyrene	60
Poly(vinyl chloride)	60
Cellulose acetate and others	60
Styrene – acrylonitrile	70
Polyethylene	70
Acrylonitrile – butadiene – styrene	80
Poly(methyl methacrylate)	80
Poly(phenylene oxide)	110
Polyamide	110 – 120
Polypropylene	110
Polycarbonate	120
Polysulfone	150
Polyoxymethylene	150
Poly(ethylene terephthalate)	150
Polyurethane (foamed)	60
Urea – formaldehyde resin, melamine – formaldehyde resin	110
Phenol – formaldehyde resin	120 – 150
Unsaturated polyester	150

Paint systems are selected according to the demands placed on the surface of the finished article and the coating characteristics of the substrate itself. These include physically drying paints (i.e., those that dry simply by evaporation of the solvent, e.g., nitrocellulose – resin lacquers and acrylic and vinyl paints); oxidative drying fatty-acid-containing resin paints; chemically drying multicomponent paints based on epoxy resins, polyurethanes, or unsaturated polyester resins; and resin stoving paints (see Section 3.3.2). For further information on various paint systems, see → Paints and Coatings.

The solvent sensitivity of many thermoplastics must be taken into account in formulating the paint (Table 14). Generally, solvent sensitivity increases with increasing temperature in the painting and drying process, as well as with increasing contact with the paint (e.g., in dipping). In particular, involatile solvents can diffuse deeper into the substrate surface and thus cause changes in molecular structure that lead to surface roughness and loss of mechanical strength. Highly solvent-sensitive polymers (polystyrene, polystyrene – butadiene, polycar-

Table 14. Short-term resistance of plastics to solvents during painting ∗

Plastic	Alcohol and glycol ether	Esters	Ketones	Gasoline	Benzene	Chlorinated hydrocarbons
Polystyrene	+	–	–	+	–	–
Polystyrene – butadiene	+	–	–	+	–	–
Styrene – acrylonitrile	+	–	–	+	–	–
Acrylonitrile – butadiene – styrene	+	○	–	+	○	○
Poly(phenylene oxide)	+	○	–	+	○	–
Poly(methyl methacrylate)	–	○	○	+	+	+
Polycarbonate	+	–	–	+	○	–
Poly(vinyl chloride)	+	+	○	+	+	○
Cellulose acetate and others	+	○	○	+	+	+
Polyamide	+	+	+	+	+	+
Polysulfone	+	○	○	+	○	–
Polyoxymethylene	+	+	+	+	+	+
Polyethylene, polypropylene,epoxy resins	+	+	+	+	+	+
Poly(ethylene terephthalate)	+	+	+	+	+	+
Thermosets	+	+	+	+	+	+

∗ + resistant; ○ partially resistant; – nonresistant.

bonate) allow only a limited choice of paints [29].

Polyurethane integral foams, thermosets, and elastomers have a closed, smooth surface. Differences in flexibility and foam weight influence the surface integrity and hence the coating characteristics. Single-layer coatings with smooth surfaces can be achieved if the surface quality is high. The best results are obtained with two-pack polyurethane paints based on aliphatic isocyanates. Apart from self-releasing polyurethane foams, the choice of release agent and its subsequent removal are particularly important. Thermoplastic cast foams exhibit an open-pore surface, which gives good adhesion to coatings and is less sensitive to solvent effects than unfoamed plastics. Smooth surfaces can be obtained only by multilayer coating with intermediate polishing. For this reason, paints with structured surfaces are generally employed.

Areas of Application. A precise distinction between functional and decorative coatings cannot generally be made [26]. Varnishes and paints that absorb UV radiation prevent photochemical degradation of ABS, PC, POM, and poly(propylene oxide) (PPO), which would otherwise result in discoloration, spots, loss of gloss, and embrittlement. Coatings increase the resistance to many aggressive media such as solvents, water, chemical cleaning baths, atmospheric influences, adhesives, and plasticizers. Furthermore, by using colored paints, direct coloration of the polymer, which requires expensive high-quality pigments, can be avoided. Matt metallic coatings are obtained by using aluminum or bronze paints, of either the leafing or the nonleafing type. In leafing products the metal flakes float on the surface, are lighter in color, and are contact sensitive, whereas in nonleafing products they remain embedded in the coating and exhibit greater scratch and corrosion resistance.

The fire resistance of plastics containing flame retardants can be improved further by application of flame-retarding paints.

Antistatic coatings prevent electrostatic charging and consequent collection of dust on nonconducting polymer surfaces. Conductive coatings filled with graphite or metal powders (copper, silver, or their alloys) are used as internal coatings (e.g., in television sets and computers). Nonconducting metallic paints, which are used primarily for decorative purposes, prevent creep currents in electronic equipment.

During the electroplating of polymers (see Section 3.3.3.2), cover coatings are used for partial chromium plating. The coatings protect the substrate from damage caused by chemical baths; no metal is deposited on the coatings themselves. A sharp interface with the metal-plated areas is, however, practically impossible to achieve. If this is required, the entire molding should be chromium plated, and selected areas should be lacquered (e.g., automobile radiator grills).

Abrasion-resistant coatings consisting of polysiloxane resin are used to protect plastic

lenses and mechanically loaded polycarbonate components from scratching.

Molds for polyurethane components can be sprayed with mold-release paints instead of release agents so that a coated molding is formed in the mold (in-mold coating).

Paints also serve as a base layer and adhesion promoter for printing and film embossing with simultaneous coloration.

In the manufacture of consistently colored or mat surfaces, and also to cover faults such as joints, colorless, mat, or pigmented scratch-resistant paints are utilized. Thermoplastic foam castings and glass fiber reinforced plastic components are coated with strongly filling paints to achieve optically satisfactory surfaces.

Solvents and other volatile components released during drying pass into the exhaust air in gaseous form; the exhaust must therefore be treated to comply with emission limits. Because vast amounts of air are involved in paint drying, treatment equipment is large and expensive. A trend toward the use of fewer solvent-based paints in favor of waterborne systems is apparent, particularly in the automobile industry.

3.3.2.2. Printing

In principle, all printing techniques are applicable to polymers [23,30,31]. The most suitable process for a given task depends on the material thickness and the substrate form (film, sheet, molding). Prior to printing, the material may have to be pretreated (see Section 3.3.5.1). For several applications (e.g., wood grain printing), a primer layer is applied before the actual print layer, and a topcoat protects the print from damage. Because the substrate is generally hard and nonabsorbent, indirect or transfer printing techniques are used, which employ an elastic intermediate substrate, generally made of rubber (e.g., gravure printing with horizontal or vertical rolls). The printed pattern is transferred, via a rubber roll, from a deeply etched, ink-filled steel applicator roll to the planar printing substrate.

Films are generally printed by using rotary techniques (flexo or gravure printing). Sheets that are to be vacuum remolded are printed in an offset process. The distorted printed pattern is straightened during remolding.

The most common technique – silk screen printing – is used for both sheets and moldings. The tightly woven fabric (usually made of polyester or polyamide) is mounted tautly in a frame and partly coated with a UV-cured solvent-resistant layer [30]. The sieve is positioned close to the article to be printed, and the ink is pressed through the screen onto the substrate by using an elastic doctor blade. The tension in the sieve ensures its withdrawal from the surface once the process is complete.

Moldings with complex shapes can be printed by indirect silk screen printing with an intermediate rubber substrate. With the tampo-print technique, also an indirect printing method, a negatively etched printing block is automatically filled with ink and doctored by a steel blade. The ink remaining in the etched areas is then transferred to the substrate by a tampon. The highly flexible printing tampon made of silicone rubber can be manufactured in a variety of forms so that irregularities and curvatures in the surface of the article to be printed do not play a significant role. Even multicolor wet-in-wet printing is possible.

3.3.3. Wet Chemical Deposition Processes

Wet chemical processes are used to deposit metallic layers on polymers. Because polymers are generally electrical insulators, components that are to be electroplated must first be coated with a metallic, conducting layer in an electroless process.

3.3.3.1. Chemical Deposition

Chemical deposition processes are based on the techniques used to manufacture mirrors: The metal (in the case of mirrors, usually silver) is deposited from an aqueous salt solution by a chemical reducing agent such as formaldehyde, glyoxal, or hydrazine.

Similar methods have been developed for depositing copper and nickel on polymer surfaces [23,32]; these are generally immersion processes. The clean, wettable polymer surface is first "activated" or "seeded" with silver or palladium. For example, to obtain silver nucleation sites, the surface is immersed in a solution of tin (II) chloride, rinsed, and dipped in a basic silver nitrate solution. For palladium seeding a

palladium salt solution is employed, followed by a reducing agent. A comparable effect can be achieved by immersion in a solution of colloidal palladium. Nucleation with 0.1 – 1.0 mg/dm^2 of noble metal is a prerequisite for specific metal deposition by chemical metallizing techniques. The most common chemical metallizing baths contain the metal (Cu or Ni) as an aqueous solution of a complex salt along with the reducing agent (e.g., sodium hypophosphite or diethylaminoborane for nickel baths and formaldehyde for copper baths). The composition of the bath is adjusted so that metal deposition begins when the nucleated article is immersed and ceases when it has been removed. A coherent metal layer ca. 0.3 µm thick is deposited on the polymer typically within 10 min. This thin metallic layer can subsequently be thickened by electroplating.

With small production runs (e.g., jewelry), a spray process can also be employed for chemical metallization. A solution of a silver salt and a solution of a reducing agent are sprayed simultaneously onto an activated polymer surface by using a two-component spray gun. The components react to form a continuous layer of silver.

3.3.3.2. Electroplating of Plastics

The thin metal layers deposited by electroless methods are extremely sensitive to abrasion and corrosion. When demands other than decoration arise, such as hardness, abrasion resistance, corrosion resistance, and electrical conductivity, the thin layers must be reinforced electrolytically [23, 32]. Total metallic behavior of the layer is reached only with layer thicknesses of ca. 20 µm. However, because the Young's modulus of metals is approximately two orders of magnitude greater than that of the polymer, and the coefficient of thermal expansion is approximately an order of magnitude lower, with thick metal deposits, forces occur at the interface that cannot be withstood by the adsorptive forces arising between polymer and metal as a result of even the smallest mechanical or thermal loading.

Although electrolytic reinforcement normally leads merely to sheathing of the molding with metal, it is suitable for making small spherical articles such as typewriter heads. The poor adhesion of the metal is in fact desirable for electroforming [33] (e.g., for the manufacture of press matrices or injection molds from plastic models). The range of applications of metallizing techniques with polymeric moldings is limited by the poor adhesion of the metal, which cannot be improved by mechanical treatment of the plastic surface.

Polymer electroplating differs from the methods discussed above in that metal adhesion approximately two orders of magnitude greater can be obtained. Development of this process began in 1965 when it first became possible to produce a strong bond between a metal and a specially treated ABS graft polymer [34, 35]. Within a few years, a fully developed technology resulted that was adapted to suit ABS polymers [32] and that is still used to electroplate polymer moldings on a mass production scale with automated equipment.

The special significance of ABS graft polymers is a result of their structure. They have a styrene – acrylonitrile copolymer matrix in which fine spheres of an elastomer phase are dispersed (Fig. 94). This elastomer phase can be removed to a depth of ca. 1 µm by oxidation in a chromic acid – sulfuric acid etching bath [32, 36], without damaging the matrix material, in 5 – 10 min at 60 °C. A submicroscopic pore system with dimensions corresponding to the diameter of the elastomer particles is generated in the surface (Fig. 95).

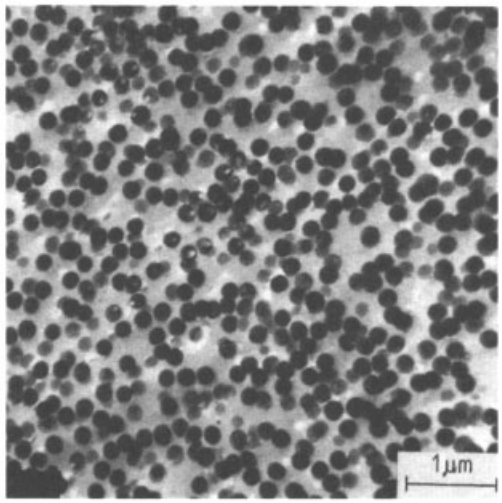

Figure 94. ABS graft polymer

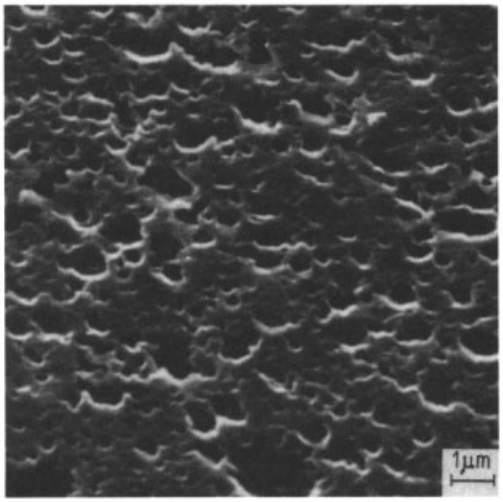

Figure 95. Surface of an ABS graft polymer after treatment in an etching bath

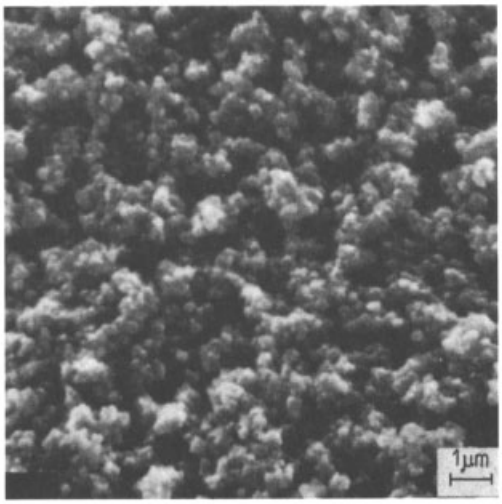

Figure 96. Underside of conducting metal layer (plastic removed by solvent)

After etching, the surface is activated, and nickel or copper can be deposited chemically (see Section 3.3.3.1). Initially the pores are filled, and subsequently a continuous, strongly anchored conductive metal layer is then formed that can be electrolytically reinforced, for example, with 20 µm of copper, 10 µm of nickel, or 0.2 – 0.8 µm of chromium by standard processes (see → Electrochemical and Chemical Deposition). The lower surface of the metal layer (Fig. 96) forms a positive replica of the pore system generated by etching. A polymer – metal composite is thus formed that is bonded so strongly that removal by force of the metal leads to cohesive fracture of the polymer in the vicinity of the metal layer. The adhesion of the metal layer is hence determined by the tensile strength of the outermost polymer layers [32, 37].

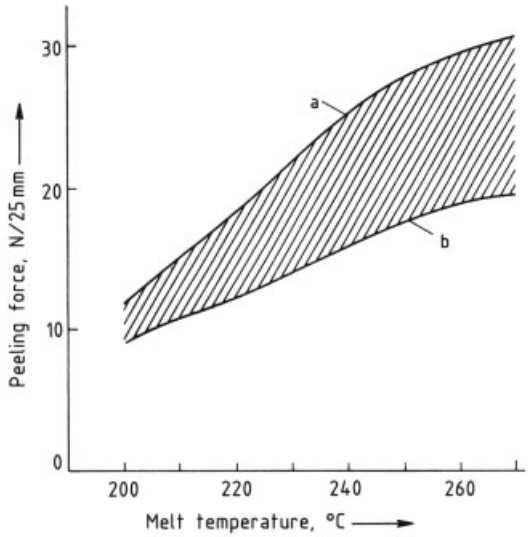

Figure 97. Peeling force as a function of melt temperature in injection molding for a plastic disk of 50 mm diameter and 2 mm thickness
a) Slow injection; b) Fast injection

During the forming of thermoplastics by injection molding, orientation is frozen into the outermost layers, in particular, which can lead to a significant anisotropy in the mechanical properties. The processing conditions for injection molding (melt temperature, injection speed, mold temperature, wall thickness, position and design of the gate, etc.) often exert a decisive influence on the adhesion and the quality of electroplated moldings [32, 37]. Figure 97 shows the influence of melt temperature on peeling force (see below). Whereas with thick-walled, orientation-free (pressed) moldings, a peeeling force of >100 N per 25 mm can be reached, under less favorable processing conditions (high orientation) this value can fall below 10 N per 25 mm. Peeling forces >15 N per 25 mm are adequate to meet the majority of technical requirements.

Electroplated ABS is a composite material with a specific property profile [32, 37, 38]. The surface properties (gloss, hardness, abrasion resistance, corrosion resistance) are determined by the thickness and the structure of the metal layer.

Due to the reinforcing influence of the metal, the molding stiffness and strength increase proportionally to the ratio of the thickness of the metal to that of the polymer, up to approximately three to five times that of the unplated material. The creep behavior typical of polymers under loading is suppressed. The heat resistance of the plated molding is considerably superior to that of unplated materials.

The ABS electroplating process can also be applied to other polymers with only minor modifications. These include ABS blends with other polymers (e.g., PC) as well as polystyrene and poly(phenylene oxide) modified with butadiene rubber particles.

Polypropylene can be rendered electroplatable by using specific additives. In these PP products, the amorphous and crystalline regions are attacked by the etching bath to various extents. This leads to extensive fissured roughening of the polymer surface, which gives good anchorage and adhesion of the metal layer [23].

Most polymers can now be electroplated by employing other etching or pretreatment techniques [23, 39].

Quality Test. For planar surfaces, the peel strength (DIN 53 494) is used as a measure of adhesion. The quality of moldings is evaluated by means of alternating temperature tests (DIN 53 496).

Applications. Electroplated ABS moldings are employed where their low manufacturing costs offer advantages over metal components, or where their advantages relative to unplated material due to increased hardness, stiffness, dimensional stability, and reduced weight are apparent [23]. Particularly in the automobile, radio broadcasting, audio, electronic, lining, writing instrument, and watchmaking industries, electroplated polymers find application as fittings, covers, handles, buttons, levers, keys, frames, and housings. Uses range from radiator grills to dials for wristwatches. The emphasis has shifted from decorative effects (e.g., for fashion articles such as buckles, buttons, fasteners, and chains) to functional components such as chassis, shields, waveguides, and switches in the electrical industry.

3.3.4. Physical Deposition Processes

Various techniques are employed to produce transparent (layer thickness $<0.2\,\mu m$) or high-gloss (layer thickness $>0.2\,\mu m$) metallic coatings and abrasion- or scratch-resistant ceramic protective coatings.

Atmospheric Pressure Techniques. Metal spraying permits the rapid buildup of a relatively thick layer.

Metal powder (powder process) or wire (wire process) is melted (e.g., in a plasma burner) and sprayed onto the polymer surface by a gas jet. Rough metal layers with pores are formed rapidly, which exhibit low adhesion to the base material. For more details on thermal spraying of metals, see → Metals, Surface Treatment, Chap. 5.

Vacuum Techniques. The deposition of metals in a high vacuum is of greater importance than atmospheric pressure methods. The metal is evaporated and condenses on the cold surface of the substrate in a vacuum chamber ($10^{-2} - 10^{-3}$ Pa). It forms a dense film with high reflectivity. The following methods are used:

1) Thermal evaporation, whereby heat is generated by electrical heating, induction, or current flowing directly through the molten metal to be evaporated.
2) Electron beam evaporation, in which a high-energy electron beam is focused on the metal to be evaporated.
3) Cathodic sputtering in which a gas discharge (generally in an argon atmosphere) is ignited in the chamber: the chamber walls form the anode, while the metal is located on the bar or plate cathode. The energetic gas ions that bombard the metal surface dislodge particles, some of atomic dimensions, that deposit on the substrate surface. Magnetic fields are frequently employed to increase the efficiency of evaporation and reduce the thermal stress to which the substrate is exposed.

Thermal evaporation techniques do not permit deposition of alloys of defined composition because fractionation occurs as a result of differences in the boiling points of the alloy components. Deposition of alloys is possible by using electron beam evaporation and cathodic sputtering.

All evaporators operate with a directed jet of metal vapor so the vacuum chamber is not completely filled with an atmosphere of metal vapor, which explains why only the substrate surface that is in direct line with the evaporator is coated. For this reason, the molding surface should not be too rough and should have no undercuts.

Because of the time required to generate a high vacuum, substrates are introduced into the chamber in batches and attached to a turntable so that they can be coated successively.

The substrate surfaces are often activated prior to evaporation. A corona discharge is ignited in the chamber at ca. 10 Pa. The ions impinging on the substrate surface cause cleavage of chemical bonds near the surface and thus a roughening of the surface and the formation of polar structures. In this way, nucleation sites are generated for the condensing metal, and good adhesion of the coating is attained.

Apart from aluminum, which is used in most cases because of the high reflectivity that can be attained, silver, gold, copper, chromium, zinc, nickel, cobalt, tin, titanium, and indium are employed, as well as several nonmetallic materials such as borosilicate glass and silicon monoxide.

The thin metal films are generally extremely sensitive to mechanical damage and are prone to rapid attack by moisture and environmental chemicals (exceptions include aluminum, which on exposure to air immediately forms a hard protective oxide layer, and the noble metals). Therefore, in most cases the metal layers are covered with a protective coating.

All polymers are suitable as substrates providing they do not evolve any materials that could lead to disturbances in the high vacuum (e.g., plasticizer from PVC or moisture from PA 6, PA 66, and cellulose hydrate). Such substrates must be precoated with a primer that hinders degassing.

The metallic deposit reproduces the finest structures and defects of the substrate, and adhesion is seldom adequate. Hence, for these reasons as well, use of a primer can be beneficial.

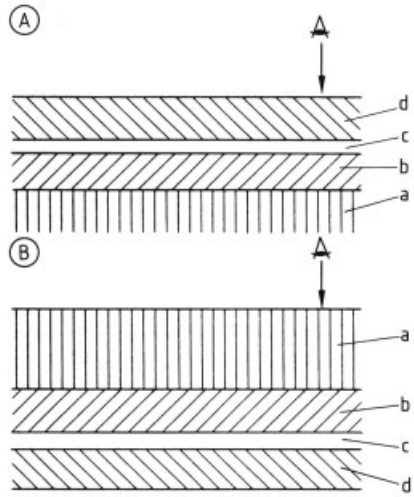

Figure 98. First-surface (A) and second-surface (B) metallization
a) Transparent substrate; b) Primer; c) Metal coating; d) Protective coating

This primer layer must form a smooth, defect-free, high-gloss surface and exhibit excellent adhesion to the substrate. After drying – almost always at elevated temperature to drive out any residual solvents – no disturbing substances may be released into the vacuum. Furthermore, the primer coating should exhibit good adhesion to the metal film and should not be colored. It should be unaffected by the metal film and resistant to any solvents in the protective coating.

Similarly, the protective coating must exhibit excellent adhesion to the metal and must not have any detrimental effects on the substrate, the primer coat, or the metallic layer.

The primer and protective coatings are colorless or colored with translucent, soluble, lightfast dyes. They can be used to imitate other metals such as chromium, copper, gold, and brass. Highly homogeneous but less lightfast colors are obtained by immersing the colorless protective coatings in dye solutions.

Transparent polymers such as poly(methyl methacrylate) and polycarbonate (Fig. 98) can be first-surface or second-surface metallized. With second-surface metallizing, the primer coating is often unnecessary, and the protective coating is generally highly pigmented and opaque.

Typical areas of application for polymer metallization include components for automo-

bile interiors, taillight and headlight reflectors, sanitary fittings, signs, flashlight reflectors, lamp components, Christmas tree decorations, toys, fashion jewelry, caps for cosmetic bottles, frames for pictures and mirrors, and wristwatch dials.

3.3.5. Dry Reactive Processes

Dry reactive processes are employed to functionally modify the surface of a polymer or as a pretreatment for further coating processes.

Differences are drawn between:

1) Processes during which the substrate is itself chemically altered
2) Application of additional layers to the surface

3.3.5.1. Surface-Modifying Processes

Corona treatment is an electrical pretreatment process. The substrate is located at ambient pressure in a discharge field between two electrodes to which an a.c. voltage of 10 – 20 kV with a frequency of 20 – 40 kHz is applied. This technique is used mostly to activate the surface of sheets and films by passing them between two roll electrodes with a gap of ca. 1 mm. The attainable surface energies are high despite the relatively short treatment times. The treatment times depend on the electrical performance of the equipment although this is limited by possible electrical breakdown of the substrate. Whereas regularly shaped hollow moldings are relatively amenable to corona treatment by placing them on appropriately formed counterelectrodes, the technique is difficult to apply to irregularly shaped articles. In such cases, counterelectrode-free techniques are used in which, by applying very high voltages, the charge carriers can be sprayed from wire bundles onto the plastic component.

With substrates containing low molecular mass additives (e.g., lubricants, plasticizers), a decrease in surface energy with storage time following treatment can be observed. This is caused by migration of the low molecular mass components.

The toxic substances, such as nitrogen oxides and ozone, generated in large amounts in the discharge field must be removed.

Fluorination is employed not only to activate polymer surfaces for successive coating processes (adhesive bonding, painting, printing, etc.) but also to permanently modify surface properties such as wettability and permeability [42, 44]. The substrate is exposed to a nitrogen atmosphere containing up to 10 % fluorine at slightly reduced pressures and at temperatures up to 110 °C. In *oxyfluorination* [45], oxygen is added to the gas mixture. Apart from the increased surface energy, fluorinated substrate surfaces are characterized by a pronounced diffusion barrier effect toward organic solvents, for example. For this reason, the technique is employed to treat the inner surface of plastic fuel tanks to reduce permeation of the fuel. With oxyfluorination, good barrier effects are achieved even after short treatment times (ca. 10 min). With fluorination, better barrier effects can be achieved but only after extended treatment times. As a process for activating the substrate surface, fluorination has an advantage over corona treatment because it gives higher surface energies. Furthermore, the activating effect remains longer with substrates containing low molecular mass components because the rate of diffusion of these components in the fluorinated substrate layers near the surface (1 – 2 μm) is also reduced. However, longer treatment times are required, and equipment for the continuous fluorination of sheets is more expensive than corona machines. Fluorination has proved relatively problem free for the surface activation of irregularly shaped components.

Because the gaseous fluorinated products that are formed (e.g., hydrogen fluoride) and fluorine gas are severe pollutants, measures to guarantee compliance with permitted workplace and exhaust gas concentrations are necessary (operation at slightly reduced pressure, exhaust gas cleaning).

Flame treatment of plastics is a surface activation process that is operated at atmospheric pressure. The substrate surface is briefly exposed to the flame from a gas burner (propane, butane, town gas, or natural gas) adjusted to give strongly oxidizing conditions (excess oxygen).

The substrate is oxidized at the surface, and the polar groups that are formed lead to improved wettability. Significant process parameters include the gas – air ratio, flame temperature, and duration of treatment, all of which must be adjusted so that the substrate is not heated too strongly and the surface is not visibly altered. The process is applicable to films, sheets, and moldings with large smooth surfaces. Flame treatment is employed primarily with polyolefins (e.g., PP automobile bumper linings). Because the activating effect is lost rapidly, flame-treated articles should be coated immediately after treatment.

Plasma treatment of plastics is also a surface activation process. It is carried out in vacuum chambers at 10 – 100 Pa with a nonthermal plasma (gas discharge) in a gas atmosphere consisting of an inert gas or reactive gases (e.g., oxygen). In the plasma, gas atoms and molecules are ionized and accelerated toward the substrate surface, which they impact with high energy, causing cleavage of chemical bonds and formation of polar groups. In *glow discharge* processes the plasma is fed with a d.c. or an a.c. voltage field (frequency range 50 Hz up to several tens of megahertz). The plasma burns between two electrodes located in the chamber (resistive coupling), whereby one of the electrodes can be an electrically conducting substrate. For treating sheets and films, the separation between electrodes at 200 – 400 V is considerably larger than with corona treatment and amounts to several tens of centimeters. When a glow discharge is used as a pretreatment (e.g., for metallizing) for moldings, the separation between the electrodes can be considerably larger and voltages up to 5 kV are applied.

When energy is introduced into the plasma by using *microwaves* (gigahertz range), significantly shorter treatment times (<1 s) are possible due to the increased charge density in the plasma.

These pretreatment techniques are used, for example, in the physical metallization or coating of films and moldings (automobile bumper covers [46]), and also in the treatment of carbon, aramid, and high-strength PE fibers to improve fiber-to-matrix adhesion when they are employed in fiber – polymer composites [47, 48].

Because plasma techniques require considerable expenditures for equipment due to the low pressures required, they are usually applied only when a subsequent vacuum coating process is intended.

3.3.5.2. Coating Processes

Gasplating is normally carried out at atmospheric or slightly elevated pressure. It is used in metallizing, where volatile metal compounds (e.g., nickel tetracarbonyl) are employed, which decompose on the preheated substrate surface and form a pore-free, homogeneous metal film.

Plasma polymerization is a vacuum coating process that is suitable for metallic substrates and also for thermally sensitive polymers due to the low thermal loading of the substrate [47, 49–56]. Monomer gases (e.g., ethylene, vinyl chloride, hexamethyldisiloxane) are vented into the vacuum chamber containing the substrate at 1 – 1000 Pa. A nonthermal plasma is ignited in which the monomer gas is activated. The plasma can be fed by an electrical d.c. or a.c. voltage field with frequencies of 50 Hz to several megahertz between two electrodes located within the chamber (resistive coupling) or outside it (capacitive coupling – not applicable to d.c. plasmas). Energy input using microwaves (generally 2.45 GHz) gives the highest attainable coating rate due to the high charge carrier density in the plasma. A distinction is made between processes in which the plasma burns directly on the surface to be coated and those in which it is remote from the surface (so-called downstream processes). In the latter, the monomers and (optional) propellant gas particles that are activated in the plasma (plasma reactor) must flow through a plasma-free zone to the substrate surface, whereby the first recombination reactions occur in the gas phase. Indirect ionization is employed with particularly sensitive monomers. Here, a propellant gas (e.g., a noble gas) is activated in the plasma, and the monomer gas is activated by ionized propellant ejected from the plasma. This prevents fragmentation of the monomer by direct contact with the plasma.

The activated particles and recombined intermediate products meet at the substrate surface and, in the course of further chemical reactions,

form layers of highly cross-linked polymer. Low molecular mass, gaseous products that are not incorporated in the layers are removed by the vacuum pump. Apart from the composition of the feed gas, which can contain other inert or reactive gases in addition to the monomers, process parameters that influence the rate of deposition and the properties of the layer include reactor geometry, separation between plasma and substrate in downstream processes, substrate material and temperature, flow of feed gas, process pressure, excitation energy, excitation frequency, and coating time.

Because of the high degree of cross-linking, plasma-polymerized layers exhibit high density, mechanical stability, and chemical resistance. They also have electrical breakdown voltages and good heat resistance. Furthermore, even at very low thicknesses (10 nm) they are free of micropores, so that they are ideally suited as diffusion barrier layers. Due to the high energy of the particles formed during the process, the layers generally adhere well to the substrate.

As in metallizing in high vacuum, the properties of the layer can be varied across its thickness by varying the operating parameters or the composition of the process gas during coating.

The process is used in the coating of reflectors for headlights [46, 57]. These are mounted in charges on a rotating carousel in a high vacuum, where they are pretreated (glow discharge), coated with aluminum, and subsequently coated with a protective layer 20 – 30 nm thick by plasma polymerization with a process gas containing hexamethyldisiloxane and oxygen. Because of their extremely low thickness, the protective layers have no optical influence but are effective diffusion barriers (condensed water resistance is 1000 times greater than without the coating) and also have a high surface energy so that condensing materials (moisture or degassed products from other headlight components) spread across the surface to form thin films and hence have no optical effects. Other applications include the hydrophilic coating of contact lenses and the coating of medical materials to improve biocompatibility [58].

On a laboratory scale, a variety of layer properties have been investigated. For instance, with ethylene as a monomer, it has been possible to apply a layer to the inside of PE screw top bottles (1.2 L) that reduces the permeation of gasoline to $< 2\,\%$ of that of uncoated bottles [59, 60]. The highly cross-linked polyethylene layer was coated onto the substrate within 40 s by using a microwave plasma; it had a high density ($>1.5\,\mathrm{kg/dm^3}$) and a thickness of 100 –200 nm. Because the coating also exhibits excellent adhesion and does not suffer appreciable loss of barrier properties even after severe deformation (drop test), it represents a viable alternative to fluorination in the manufacture of plastic fuel tanks.

An advantage of plasma polymerization is that a very high proportion of the monomers employed is incorporated into the film, and only small amounts are present in the evacuated residual gases. The material costs of the process are thus very low. Additionally, environmentally harmless materials are used in many cases so that the exhaust gases need not be cleaned. Because vacuum processes are involved, however, considerable equipment expenditure is necessary. Especially with microwave plasma polymerization, the process cycle time is influenced primarily by the capacity of the vacuum pumps (evacuation time), whereas the actual coating time is low (generally <1 min). For this reason, chambers for rough evacuation of the components to be coated through which the substrate passes to and from the actual process chamber, are useful.

4. References

General References

1. O. Lauer: *Aufbereiten von Kunststoffen,* Hanser Verlag, München 1971. *Aufbereiten von PVC,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1976. *PVC-Heißmischen (Dryblend),* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1968. *Speichern, Fördern und Dosieren von Kunststoffen,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1971. *Granulieren von thermoplastischen Kunststoffen,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1974. W. Dalhoff: *Systematische Extruder-Konstruktion,* Krauskopf Verlag, Mainz 1974. H. Domininghaus: *Fortschrittliche Extrudertechnik,* VDI-Verlag, Düsseldorf 1970. *Der Extruder im Extrusionsprozeß,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1989. H. Herrmann:

Schneckenmaschinen in der Verfahrenstechnik, Springer Verlag, Berlin 1972. M. Jacobi: *Grundlagen der Extrudertechnik,* Hanser Verlag, München 1960. W. Mink: *Grundzüge der Extrudertechnik,* Zechner & Hüthig Verlag, Speyer 1973. H. Potente: *Modellgesetze für Ein- und Zweischneckenmaschinen,* Hanser Verlag, München 1981. C. Rauwendaal: *Polymer Extrusion,* Hanser Verlag, München 1986. G. Schenkel: *Kunststoff-Extrudertechnik,* Hanser Verlag, München 1963. E. G. Fischer: *Extrusion of Plastics,* Butterworths, London 1976. *Extrudieren von Schlauchfolien,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1973. *Extrudieren von Rohren und Profilen,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1974. *Extrudierte Feinfolien und Verbundfolien,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1976. G. Menges, H. Recker: *Automatisierung in der Kunststoffverarbeitung,* Hanser Verlag, München 1986. *Rechnergesteuerte Extrusion,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1976. Z. Tadmor: *Engineering Principles of Plasticating Extrusion I,* Klein Reinhold Book Corporation, New York 1970. W. Michaeli: *Extrusionswerkzeuge für Kunststoffe und Kautschuke,* Hanser Verlag, München 1990. F. Hensen, W. Knappe, H. Potente: *Handbuch der Kunststoffextrusionstechnik,* **vols. I and II,** Hanser Verlag, München 1989. W. Mink: *Grundzüge der Hohlkörperblastechnik,* Zechner & Hüthig Verlag, Speyer 1969. D. v. Rosato: *Blow Molding Handbook Dominick,* Hanser Verlag, München 1989. R. Holzmann: *Gestalten von Blasformteilen,* VDI-Verlag, Düsseldorf 1971. O. Plajer: *Werkzeuge für das Blasformen,* Zechner & Hüthig Verlag, Speyer 1969. *Spritzblasen,* VDI-Gesellschaft Kunststofftechnik, Düsseldorf 1976. J. L. Throne: *Thermoforming,* Hanser Verlag, München 1987. A. Thiel: *Grundzüge der Vakuumformung,* Zechner & Hüthig Verlag, Speyer 1969. K. Stoeckhert: *Kunststoff-Lexikon,* Hanser Verlag, München 1975. F. Johannaber, K. Stoeckhert: *Kunststoffmaschinenführer,* Hanser Verlag, München 1984. O. Scharz, F.-W. Ebeling, G. Lüpke, W. Schelter: *Kunststoffverarbeitung,* Vogel, Würzburg 1988. D. H. Morton-Jones: *Polymer Processing,* Chapman and Hall, New York 1989. G. Menges: *Einführung in die Kunststoffverarbeitung,* Hanser Verlag, München-Wien 1979. F. Johannaber: "Spritzgießmaschinen," in: *Kunststoffmaschinenführer,* Hanser Verlag, München – Wien 1979. W. Elbe: "Untersuchungen zum Plastifizierverhalten von Schneckenspritzgießmaschinen," Dissertation, RWTH Aachen 1973. R. v. Hooven, A. J. Kaminski: "Entgasungsspritzgießen, ein Laborkonzept wird in die Praxis umgesetzt," *Plastverarbeiter* **31** (1990) no. 8, 441 – 446. Th. Copetti, R. Krebser: "Vollhydraulische und Kniehebelschließsysteme," *Kunststoffe* **70** (1980) no. 2, 821 – 825. G. Menges, P. Mohren: *Anleitung zum Bau von Spritzgießwerkzeugen,* Hanser Verlag, München – Wien 1983. S. Schäper: "Nukleierung thermodynamisch getriebener Polyurethan-Reaktionsschäume," Dissertation, RWTH Aachen 1977.

General References

2. W. Becker, D. Braun, B. Carlowitz: *Kunststoff-Handbuch,* **vol. 1 "Die Kunststoffe",** Hanser Verlag, München 1990. W. Becker, D. Braun, W. Woebcken: *Kunststoff-Handbuch,* **vol. 10 "Duroplaste,"** Hanser Verlag, München 1988. H. Saechtling: *Kunststoff-Taschenbuch,* 24th. ed., Hanser Verlag, München 1989. K. Stoeckhert: *Kunststoff-Lexikon,* 6th. ed., Hanser Verlag, München 1975. *DIN-Taschenbuch 21,* "Kunststoffnormen," 8th. ed., Beuth-Verlag, Berlin 1979. W. Schönthaler: *Verarbeiten härtbarer Kunststoffe,* VDI-Verlag, Düsseldorf 1973. W. Bauer, W. Woebcken: *Verarbeitung duroplastischer Formmassen,* Hanser Verlag, München 1973. W. Bucksch, H. Briefs: *Preßwerkzeuge in der Kunststofftechnik,* 2nd. ed., Springer Verlag, Berlin 1962. K. Stoeckhert: *Formenbau für die Kunststoff-Verarbeitung,* Hanser Verlag, München 1969. Berufsgenossenschaft der chem. Industrie, *Merkblatt für das Verarbeiten von Polyester- und Epoxidharzen,* Verlag Chemie, Weinheim 1966. VDI-Richtlinie 2007, Epoxidharze im Fertigungsmittelbau, VDI-Verlag, Düsseldorf 1966. VDI-Richtlinie 2011, Faserverstärkte Reaktionsharzformstoffe, VDI-Verlag, Düsseldorf 1973. H. Jahn: *Epoxidharze,* Verlag für die Grundstoffindustrie, Leipzig 1969. F. Breckner: *Glasfaserverstärkte*

Kunststoffe; Anwendungen, Eigenschaften, 2nd. ed., Verlag Fraunhofer-Gesellschaft, 1989. H. Domininghaus: *Die Kunststoffe und ihre Eigenschaften,* VDI-Verlag, Düsseldorf 1988. O. Schwarz: *Kunststoffkunde,* Vogel, Würzburg 1988. W. Michaeli, M. Wegener: *Einführung in die Technologie der Faserverbundwerkstoffe,* Hanser Verlag, München 1989. R. W. Meyer: *Handbook of Polyester Molding Compounds & Molding Technology,* Chapman and Hall, New York 1987. S. M. Lee: *International Encyclopedia of Composites,* VCH Publishers, New York 1990.

General References

3. W. Knig, L. Neder: "Schneiden mit ultraschallerregtem Messer," *Ind. Anz.* **107** (1985) 49. W. König, L. Neder: "Bessere Schnittflächen mit Ultraschall," *Plastverabeiter* **37** (1986) no. 11. "Neue Werkstoffe und hohe Präzision – eine Herausforderung für die Fertigstellungstechnik" in AWK Aachener Werkzeugmaschinen-Kolloquium, (ed.): Produktionstechnik auf dem Weg zu integrierten Systemen, VDI-Verlag, Düsseldorf 1987. W. König, P. Graß, Ch. Wulf: "Machining of Fibre Reinforced Plastics," *CIRP Ann.* **34** (1985) no. 2, 537 – 548. F.-J. Trasser, M. Wehner, "Schnittflächen beim Laserstrahlschneiden von faserverstärkten Kunststoffen CO2-, Excimer-Laser – ein Verfahrensvergleich," *Ind. Anz.* **109** (1987) no. 43/44, 57 – 58. "Bearbeitungstechnik für Verbundwerkstoffe – eine vielschichtige Aufgabenstellung," *VDI-Fachtagung METAV,* June, 8 – 9, 1988. H. K. Tönshoff, V. Hohensee, "Bearbeitung faserverstärkter Kunststoffe," *ZWF, Z. Wirtsch. Fertigung* **81** (1986) no. 2, 106 – 111.

Specific References

4. V. Hoffmann: "Beitrag zur Untersuchung der Zerspanbarkeit thermoplastischer Kunststoffe beim Bohren," Dissertation, Techn. Univ. Berlin 1971.
5. E. Lemke: "Beitrag zur Untersuchung der Zerspanbarkeit von Plastomeren," Dissertation, Techn. Univ. Berlin 1972.
6. G. Zug: "Beitrag zur Untersuchung des Zerspanverhaltens von thermoplastischen Kunststoffen beim Drehvorgang," Dissertation, Techn. Univ. Berlin 1970.
7. G. Spur, G. Zug, *WT-Z. Ind. Fertigung* **59** (1969) no. 10, 485 – 490.
8. G. Spur, G. Zug, *ZWF Wirtsch. Fertigung* **66** (1971) no. 9, 441 – 445.
9. B. Ferichmann: "Ermittlung fertigungsgerechter Arbeitsbedingungen und Untersuchung des Zerspanungsverhaltens beim Drehen thermoplastischer Kunststoffe," Dissertation, RWTH Aachen 1966.
10. G. Spur, G. Zug, *Werkstatt Betr.* **101** (1968) no. 6, 325 – 328.
11. A. Kobayashi, *Machining of Plastics,* McGraw-Hill, New York 1967.
12. H. G. Schmidt, *ZWF Wirtsch. Fertigung* **60** (1965) no. 6, 286 – 287.
13. M. Masuko, S. Ammi, *CIRP Ann.* **13** (1966) no. 4, 389 – 397.
14. G. Spur, G. Zug, *ZWF Wirtsch. Fertigung* **66** (1971) no. 2, 47 – 51.
15. G. Spur, V. Hoffmann, *ZWF Wirtsch. Fertigung* **67** (1972) no. 2, 61 – 64.
16. M. N. Watson, *Joining Plastics in Production,* The Welding Institute, Cambridge 1989.
17. A. Neumann et al.: *Grundlagen der Schweißtechnik – Plastfügetechnik,* VEB Verlag Technik, Berlin 1985.
18. G. Habenicht: *Kleben,* Springer Verlag, Berlin 1986.
19. W. Brockmann, L. Dorn, H. Käufer: *Kleben von Kunststoff mit Metall,* Springer Verlag, Berlin 1989.
20. G. Menges, G. Stockhausen, M. Reinke: *Kleben – ein Nachschlagewerk über die Verwendung von Klebstoffen in der Kunststoffverarbeitung,* Herausgeb. IKV, Aachen 1981, Verlag TüV Rheinland, Köln 1991.
21. D. Wimmer: *Kunststoffgerecht konstruieren,* Hoppenstedt Verlag, Darmstadt 1989.
22. U. Delpy: *Schnappverbindungen aus Kunststoff;* Expert-Verlag, Ehningen 1989.
23. U. Zorll, E.-C. Schütze: *Kunststoffe in der Oberflächentechnik,* Kohlhammer, Stuttgart 1986.
24. H. Saechtling, W. Zebrowski: *Kunststoff-Taschenbuch,* 24th ed., Hanser Verlag, München–Wien 1989.
25. W. A. Zisman: *J. Paint Technol.* **44** (1972) 564.
26. P. Berns: *Ind. Lackierbetr.* **44** (1976) 125.
27. P. O. Damm: *Plastverarbeiter* **25** (1974) 683 – 687.
28. F. Gudehus: "Lackieren von Kunststoffteilen," *Kunststoffe* **78** (1988) 6.
29. P. Berns: *Kunststoffhandbuch,* **vol. 5, Polystyrol,** Hanser Verlag, München 1969, pp. 369 – 376.

30. V. Bartelmäs: *Kunststoffhandbuch,* **vol. 5, Polystyrol,** Hanser Verlag, München 1969, pp. 814 – 849.
31. W. Dominiinghaus: *Dekorieren von Kunststofformteilen,* VDI-Verlag, Düsseldorf 1971.
32. *Kunststoffgalvanisierung – Handbuch für Theorie und Praxis,* Leuze Verl., Saalgau/Württ. 1973.
33. P. Spiro: *Electroforming,* R. Draper Ltd., Teddington, England, 1968.
34. K. Wiebusch, H. Hendus, E. Zahn: *Metallische überzüge auf Kunststoffen,* Hanser Verlag, München 1966, pp. 9 – 28.
35. K. Heymann, W. Riedel, G. Woldt: *Metallische überzüge auf Kunststoffen,* Hanser Verlag, München 1966, pp. 48 – 63.
36. K. Wiebusch: *Galvanotechnik* **59** (1968) no. 8, 640 –652.
37. K. Wiebusch: *Galvanotechnik* **61** (1970) no. 12, 984 –993.
38. K. Wiebusch: *Konstruktion, Fertigung und Anwendung feinwerktechnischer Kunststoffteile,* VDI-Verlag, Düsseldorf 1976, pp. 109 – 123.
39. W. Goldie: *Metallic Coating of Plastics Electrochem.,* Publ. Ltd., Ayr, UK, 1968.
40. K. Wiebusch: *Galvanotechnik* **61** (1970) no. 9, 704 –714.
41. R. P. Tison: *Plating* **1** (1973) 47 – 52.
42. C. Bliefert: "Fluorierung von Kunststoff-Kraftstoffbehältern," *Kunststoffberater* **6** (1986) 14 – 16.
43. Verein Dt. Ingenieure (ed.): *Sperrschichtbildung bei Kunststoff-Hohlkörpern,* VDI-Verlag, Düsseldorf 1986.
44. T. Volkmann, H. Widdecke: "Fluorination of Polyethylene Films," *Macromol. Chem. Macromol. Symp.* **25** (1989) 243 – 248.
45. T. Volkmann, H. Widdecke: "Oxifluorierung von Polyethylen," *Kunststoffe* **79** (1989) no. 8, 743 – 744.
46. J. Kieser: "Plasmapolymerisation als industriell einsetzbare Technik," VDI-Kolloquium *Dünnschichttechnologien,* 6./7. Nov., Hagen 1986.
47. R. H. Ludwig: "Plasmapolymerisation – ein Verfahren zur Erzeugung dünner Schichten," Dissertation, RWTH Aachen 1989.
48. H. Yasuda: "Tandem Plasma-Polymerization Apparatus for Continuous Coating of Fibers and Films," *ACS Symp. Ser.* **108,** (1979) 277 – 286.
49. G. Menges, P. Plein: "Plasmapolymerisation – maßgeschneiderte Beschichtungen für Kunststoffteile," *Kunststoffe* **78** (1988) no. 10, 1015 – 1018.
50. M. R. Wertheimer, M. Moisan: "Comparison of Microwave and Lower Frequency Plasmas for Film Deposition and Etching," *J. Vac. Sci. Technol. A* **3** (1985) no. 6, 2643 – 2649.
51. G. Menges et al.: "Plasmapolymerisation – ein zukunftsträchtiges Beschichtungsverfahren," *Coating* **1** (1989) 2 – 7.
52. R. Ludwig: "Plasmapolymerization – A New Technology for Surface Modification," *Annu. Tech. Conf. Soc. Plast. Eng.* 1989, 915 – 917.
53. H. V. Boening: *Plasma Science and Technology,* Hanser Verlag, München 1982.
54. G. Menges, P. Plein: "Die besonderen Schichten – Plasmapolymerisation steht vor breiter industrieller Nutzung," *Ind. Anz.* **109** (1987) no. 16, 23 – 24.
55. J. Kieser, M. Sellschopp: "Hochvernetze Polymerschichten mit einstellbaren Eigenschaften," *VDI-Nachr.* **41** (1984) 28.
56. J. Musil: "Microwave Plasma: Its Characteristics and Applications in Thin Film Technology," *Vacuum* **36** (1986) no. 1 – 3, 161 – 169.
57. G. Benz: "Schutzschichten durch Plasmapolymerisation," *Bosch Tech. Ber.* **8** (1986/1987).
58. Deutsche auf dünnem Eis, Industriemagazin, Jan., 1987, 124 – 129.
59. W. Michaeli et al: "Plasmabehandelte PE-Fasern – eine Verstärkungsfaser ohne Haftungsprobleme," *J. Coating* (1991) 1.
60. W. Michaeli, M. Londschien, R. Ludwig: "Plasmapolymerisierte Sperrschichten," *Kunststoffe-Plastics* **8** (1990), 13 – 16.

Plastics, Properties and Testing

MANFRED STAMM, Max-Planck-Institut für Polymerforschung, Mainz, Federal Republic of Germany (Chap. 1)

BODO CARLOWITZ, Königstein im Taunus, Federal Republic of Germany (Chap. 2)

1. States of Order of Polymers 2
1.1. Glassy State and Melt 3
1.2. Crystalline State 6
1.3. Liquid Crystals 7
1.4. Blends and Block Copolymers . . . 8
2. Properties and Testing 8
2.1. Mechanical Properties and Their Testing 9
2.1.1. Classification of Plastics Based on Mechanical Properties 9
2.1.2. Viscoelastic Behavior of Polymers . . 11
2.1.3. Mechanics of Deformation 14
2.1.4. Friction and Wear 21
2.1.5. Fracture Behavior 22
2.1.6. Control of Mechanical Properties through Orientation, Internal Stresses, and Morphology 26
2.1.7. Reinforced and Filled Plastics 28
2.2. Acoustic Properties and Their Testing 28
2.3. Thermal Properties and Their Testing 29
2.3.1. Specific Heat 29
2.3.2. Thermal Conductivity 30
2.3.3. Thermal Expansion 31
2.4. Electrical Properties and Their Testing 32
2.4.1. Insulating Properties 33
2.4.2. Dielectric Properties 34
2.4.3. Dielectric Strength 35
2.4.4. Electrets; Semiconducting and Conducting Polymers 37
2.5. Optical Properties and Their Testing 37
2.5.1. Refraction and Birefringence 37
2.5.2. Transmittance and Gloss 38
2.5.3. Color and Colorimetry 39
2.6. Density and Physical-Chemical Properties and Their Testing 41
2.6.1. Density 41
2.6.2. Liquid Absorption and Swelling; Desorption of Plasticizers 41
2.6.3. Permeation of Gases and Vapors . . . 41
2.7. Stability Properties and Their Testing 43
2.7.1. Water Absorption 43
2.7.2. Effect of Chemicals 43
2.7.3. Effects of Heat, Light, and Weathering 44
2.7.4. Effect of High-Energy Radiation . . . 45
2.7.5. Stability against Fire 46
2.7.6. Effect of Microorganisms 47
2.8. Processing Properties and Their Testing 47
2.8.1. Molding Compounds 48
2.8.2. Conventional Methods for Characterizing Flow Behavior 48
2.8.3. Rheology of Plastics 49
3. References 52

Abbreviations

a	constant
a	thermal diffusivity
a_k	impact strength (notched)
a_k	stress concentration factor
A	amplitude
A	flow area
b	constant
b	heat penetration coefficient
B	static load
c_p	specific heat
C	elastic constants
$C(t)$	compliance
d	mechanical loss factor
d	thickness of layer
D	diffusion constant
D_r	"dispersive power"
e	strain
$\dot{e}$	strain rate
E	modulus of elasticity
$E_c(t)$	creep modulus
E_d	dielectric strength
$E(t), E_r(t)$	time-dependent modulus of elasticity
f	oscillating frequency
F	friction force
F_d	factor for damping effects
F_g	factor for dimensional effects

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_001

G', G shear modulus
$h(\tau)$ distribution function of the relaxation time or relaxation spectrum
$H(\ln\tau)$ relaxation spectra
I isotropic stress invariant
I moment of inertia
I_0 incident light intensity
K bulk modulus
l length
L thickness of composite
M complex modulus
M' dynamic modulus
M'' loss modulus
$M(t)$ relaxation modulus
M_i spring constant
n_C refractive index at C-line
n_F refractive index at F-line
N cycles of failure
N normal force
P permeability
P differential operator
P_z polarization direction
Q differential operator
t loading time
T_c contact temperature
ν Abbé number
α linear coefficient of thermal expansion
α molecular polarizability
β volume coefficient of thermal expansion
γ shear
$\dot{\gamma}$ shear rate
$\dot{\gamma}_\mathrm{a}$ apparent shear rate
$\dot{\gamma}_\mathrm{t}$ true shear rate
δ phase shift
$\tan\delta$ loss factor
ε total strain
$\dot{\varepsilon}$ strain rate
ε_0 constant strain during relaxation test
ε dielectric constant
ε extensional deformation
$\dot{\varepsilon}$ extension rate
$\varepsilon'(\omega)$ dielectric constant (permittivity)
$\varepsilon''(\omega)$ dielectric loss factor
$\varepsilon*$ complex dielectric constant
$\varepsilon', \varepsilon'_\mathrm{r}$ relative permittivity
ε'' dielectric loss factor
η viscosity
η_0 zero shear viscosity
η_F extensional viscosity
η_S structure viscosity
λ thermal conductivity
λ logarithm of the degree of stretching
Λ logarithmic decrement
μ friction coefficient
ξ integration variable
ϱ density
ϱ_0 constant
ϱ_D volume resistivity
σ stress
$\dot{\sigma}$ stress rate
σ_0 constant stress during creep test
σ_a initial stress; stress amplitude
$\sigma_{B/t}$ creep strength
σ_crit stress exeding a critical value
$\sigma_{\varepsilon/t}$ "time yield limit"
σ_m mean stress
σ_max maximal stress
σ_n nominal stress
τ relaxation time
τ shear stress
τ_a apparent shear stress
τ_t true shear stress
ω angular frequency

1. States of Order of Polymers

The state of order in plastics is, despite some peculiarities, very similar to that of low molecular mass compounds. Amorphous, crystalline, and liquid crystalline phases are observed, which exhibit short- and long-range order in one, two, or three dimensions. Therefore, the *characterization techniques* used to determine the state of order are also very similar. The most common techniques are summarized in Table 1, which also lists the typical dimensions and the principal information that can be obtained. One generally distinguishes among the following typical scales for the state of order:

1) The submolecular scale (typically 0.1 – 10 nm) where the state of order of chain segments and monomer units is considered
2) The molecular scale (typically 5 – 20 nm) where the spatial arrangement of one chain is investigated
3) The morphological scale (10 nm to several micrometers) where the superstructure formed by many polymer molecules is envisaged

The crystal structure, crystallinity, and orientation of polymers are determined by X-ray methods [1], whereas the crystal morphology

Table 1. Most common techniques for the determination of order in polymers

Technique	Information	Typical dimensions
Wide-angle X-ray scattering (WAXS)	crystallinity, orientation, crystal structure, near range order	0.1 – 1 nm
Small-angle X-ray scattering (SAXS)	morphology, long spacing, density fluctuations	1 – 100 nm
Small-angle neutron scattering (SANS)	chain conformation	1 – 100 nm
Electron microscopy (EM)	morphology,	
Electron diffraction (ED)	long spacing,	10 – 1000 nm
Electron energy loss spectroscopy (EELS)	crystal structure, element distribution	
Light scattering (LS)	larger structures,	
Microscopy	different phases,	1000 nm – 1 mm
Birefringence	orientation	

or phase structure is obtained from electron microscopy [2] and light scattering [3]. The use of neutron scattering [4] offers the unique possibility of "seeing" the spatial arrangement or conformation of single molecules in the bulk by selective deuteration (Fig. 1).

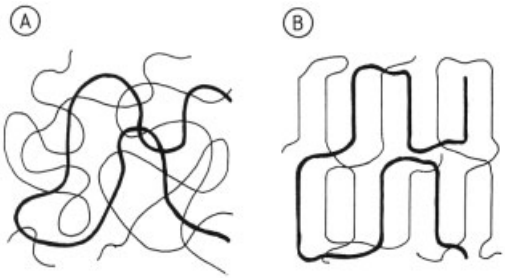

Figure 1. Schematic of chain conformations in the (A) amorphous and (B) semicrystalline state
A single chain is marked by the solid line and may represent a labeled (deuterated) chain among otherwise unlabeled molecules. In the amorphous state (melt or glass), molecules adopt a Gaussian coil conformation, whereas in the semicrystalline state, both amorphous and crystalline regions are present. During rapid crystallization from the melt, the overall chain conformation does not change significantly and chain segments are incorporated into crystalline lamellae.

The two *principal states of order* in polymers are the amorphous and the semicrystalline states. A peculiarity of polymers is that, in general, amorphous regions are also present in the crystalline state and the crystallinity can be varied over a wide range by means of the crystallization and annealing conditions. It is evident from Figure 1 that mechanical properties, such as the elasticity and the toughness of plastics, are determined largely by the connectivity and entanglement of the molecules. In particular, in the crystalline state the crystallites are interconnected by tie molecules, giving the material its mechanical strength. Hence the physical properties of polymers are determined not only by their chemical nature and composition but also by the state of order–the chain conformation, morphology, phase structure, and orientation of molecules. Therefore, the control and determination of the state of order are important areas of polymer science. They can determine the final use of products and also offer possibilities for modification and new areas of application. Examples include the stress-induced crystallization of rubber, which is undesirable because it makes the material brittle and unelastic, and the production of ultra-high-modulus polyethylene fibers by extreme orientation from a gel state to obtain a material that can be much tougher than steel in terms of the specific modulus.

Special morphologies and superstructures are observed in polymer blends and block copolymers, while the state of order of liquid crystalline polymers is between that of the crystalline and the amorphous materials.

1.1. Glassy State and Melt

Polymers in the glassy state and in the isotropic melt generally have a low degree of order: they are *amorphous.* Only short-range order is present, similar to that observed in simple liquids. Because polymers are large molecules, a distinction can be made between the intra- and intermolecular order of monomer units. Very stiff macromolecules have a high orientation correlation along the chain, while the center-of-mass positions of molecules with respect to one another are distributed statistically in space (i.e., a large intramolecular and no intermolecular order). However, in many cases, stiff molecules tend to form liquid crystalline phases (see Section 1.3).

Polymer Melt. Flexible chains form a Gaussian coil-like conformation [5], [6] with little intramolecular order. The *random-coil conformation* of the melt (Fig. 1 A) is also observed in dilute solution under theta conditions. In the ideal situation, no effective interaction occurs between a chain and its surroundings. The center-of-mass positions of the molecules are random, and the long-range order observed in the crystalline state is completely absent. This picture can be tested by small-angle neutron scattering in the melt [7], in which the radius of gyration R_g – a measure of the mean dimension of the chain and the chain conformation – can be obtained. Contrast for neutrons is generated by the deuteration of one chain in a matrix of otherwise undeuterated chains (Fig. 1). The deuterated chains are then "visible" against the others for neutrons. The assumption is that deuteration does not change the thermodynamics, and particularly the chain conformation, which to a good approximation is true in most cases [8], [9]. Investigations indicate that the chain conformation in the melt and under theta conditions is identical and, in many cases, can be described by an unperturbed coil. In the random- coil model with a random distribution of chain segments, the radius of gyration R_g should be proportional to the square root of the molecular mass M_w ($R_g = \alpha M_w^{0.5}$) [22], which to a good approximation is true for many polymers (Table 2).

Table 2. Molecular dimensions of some polymers in different states of order as determined by small-angle neutron scattering

Polymer	$\alpha = R_g/M_w^{0.5}$, nm g$^{-0.5}$	Reference
Melt		
Polyethylene	0.045	[7]
Polystyrene, atactic	0.028	[10]
Polybutadiene	0.035	[11]
Poly(ethylene oxide)	0.042	[12]
Polypropylene, isotactic	0.035	[13]
Glass		
Poly(methyl methacrylate), atactic	0.027	[14–16]
Polystyrene, atactic	0.028	[17]
Poly(vinyl chloride)	0.040	[18]
Poly(ethylene terephthalate)	0.039	[19]
Semicrystalline state		
Polyethylene	0.045	[20]
Polypropylene, isotactic	0.034 – 0.038	[13]
Polystyrene, isotactic	0.024 – 0.029	[21]
Poly(ethylene oxide)	0.052	[12]

The pair correlation function $h(r)$ of monomer units, irrespective of their connectivity to different chains, is a measure of the *near range order* of the monomer units (Fig. 2 A) [23]. In starting from one monomer unit in the melt, at a certain distance r_0 a peak in $h(r)$ is observed, corresponding to other monomer units at the nearest neighbor position, which form a first shell around the reference monomer. At larger distances, no spatial correlations exist and $h(r)$ becomes constant. It goes to zero at r=0 since a double occupancy is excluded.

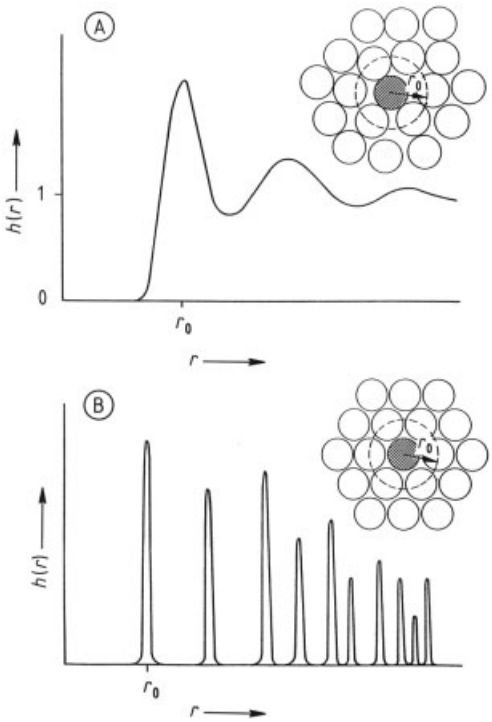

Figure 2. Distribution of monomer units $h(r)$ as a function of distance r in the (A) amorphous and (B) crystalline state for a simplified model in which monomer units are replaced by spheres
Connecting bonds to neighboring monomer units are not shown. The distance distribution $h(r)$ between centers of the spheres is continuous in the amorphous state and discrete in the crystalline state. Figure 2 A exhibits only a short-range, and Figure 2 B a long-range, positional order. The first coordination shell of radius r_0 is also indicated in the insert (dashed line).

Because of interactions of various kinds this simple picture must be modified. Even in the case of largely noninteracting particles such as in a melt of *n*-alkanes or polyethylene indications can be found of additional near range order effects because of steric reasons [24], [25]. There is a slight tendency for neighboring *trans*

sequences to align themselves parallel to one another giving rise to a small intermolecular interaction which can be observed by magnetic birefringence or depolarized light scattering [24]. These interactions do not, however, influence the conformations of the chains significantly as can be seen from small-angle neutron scattering, which is consistent with the unperturbed- coil picture of the single chains. Scattering data of isotactic, syndiotactic, and atactic poly(methyl methacrylate) [16], [26–28] and polycarbonate [29] over a wide angular range have been compared with model calculations [30], [31] based on the rotational isomeric state model [22] in which the rotational configurations (isomers) are taken into account with their proper statistical weight. The agreement is very good in those regimes where first at small angles the overall chain conformation and second at intermediate angles the segmental configuration dominate the scattering. The unperturbed- coil model thus seems to be a good approximation for the spatial arrangement of flexible polymer molecules in the amorphous state.

Glassy State. The glassy state is of great importance for practical applications of plastics. Many atactic polymers do not crystallize when they are cooled down from the melt but form an amorphous glassy state. This can be described as a frozen melt in which the mobility of the chains is frozen in and the state of order of the melt at the glass transition temperature T_g is retained. The detailed nature of this transition is still not clear, but various aspects have been investigated and analyzed [32]. Many properties, such as mechanical modulus, dielectric constant, and specific heat, change at T_g (Fig. 3). A change in slope occurs at T_g for both volume V and enthalpy H, whereas a jump is observed for the specific heat c_p and the coefficient of thermal expansion α. The smooth transition in V and H and the jump in the second derivatives c_p and α suggests a second-order phase transition, which, however, cannot be demonstrated rigorously.

The glass transition temperature depends on the rate of cooling or heating, the frequency of the measuring technique (e.g., in dielectric or dynamic mechanical investigations), and the annealing conditions [32], [33]. The glassy state is thus not an equilibrium state. Since an equilibrium thermodynamic description of the glassy state [34] is questionable [35], a nonequilibrium free-volume theory [36] is mostly used, which assumes that the volume of the melt consists of the van der Waals volume of the molecules V_0 and an additional free volume V_f. For the movement of chains in the melt a certain number of voids or empty sites are required and are taken into account by the free volume. The free volume of the melt decreases with decreasing temperature (Fig. 3) and is frozen in at T_g. At T_g the free volume is estimated to be 2.5 % of the total volume.

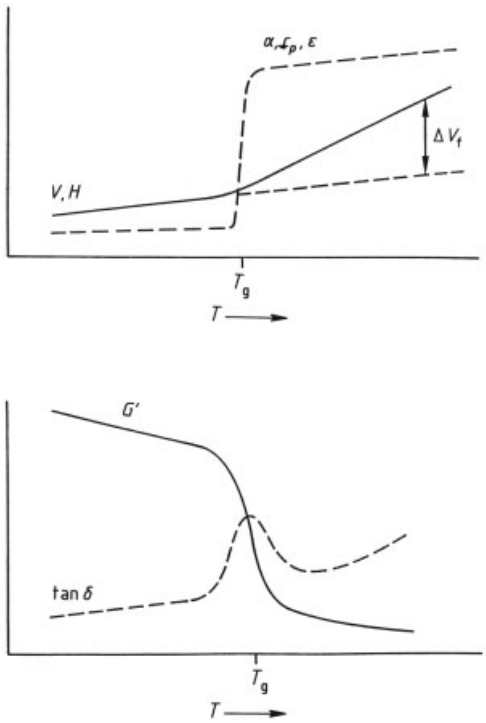

Figure 3. Schematic of the change of various properties of plastics at the glass transition temperature T_g: volume V, enthalpy H, linear coefficient of thermal expansion α, specific heat c_p, dielectric constant ε, shear modulus G', and mechanical loss factor $\tan\delta$
Also indicated in the upper figure is the increase of free volume ΔV_f with temperature according to the free-volume theory of the glass transition.

The empirical Williams – Landel – Ferry (WLF) approach [37] can be used to calculate the time – frequency – temperature behavior of the viscosity, diffusion constant, dielectric or mechanical relaxation, etc., and is an expression of the dynamic nature of the glassy state. Glass transition temperatures can be estimated

from the chemical structure [38]. A list of the glass transition temperatures of many polymers is found in [39].

Blends; Polymer Orientation. In *blends* of glassy materials a single glass transition is observed when miscibility on a molecular level is achieved [40]. For blends of immiscible plastics the glass transition temperatures of the pure components are measured, which indicate phase segregation on the microscopic level. The addition of a low molecular mass component to plastics may lower T_g (plasticizer effect) [32], but can also cause crystallization or phase separation.

The order in amorphous polymers can be increased by *orientation* of the molecules, which is achieved, for instance, by mechanical deformation [41]. With some materials, stress-induced crystallization is observed (e.g., rubber) [42], which strongly affects mechanical properties. Similarly, on annealing above T_g some plastics [e.g., poly(vinyl chloride), polycarbonate, poly(ethylene terephthalate] undergo partial crystallization [32]. In most crystalline plastics a large portion of the material is amorphous, and crystalline lamellae are formed between the amorphous regions.

1.2. Crystalline State

Some plastics can crystallize and form a crystalline state that depends largely on sample history and preparation. Only polymers with well-defined tacticity can crystallize, whereas atactic materials are generally amorphous. In contrast to low molecular mass materials, plastics form only a *semicrystalline state*, and generally both crystalline and amorphous regions are present in the polymeric crystalline state (Fig. 1 B). In the crystalline regions a long-range spatial order is present (Fig. 2), and monomer units occupy well-defined locations within the crystal. Crystal structures of polymers have been determined by X-ray and electron diffraction and are known for most crystalline polymers [39], [43]. The structural parameters depend, however, on sample treatment, and a different crystalline order may result from heating, quenching, annealing, orientation, or the application of pressure. The degree of crystallinity and the morphology (i.e., the superstructure of the crystallites) also depend on sample treatment.

Melt-Crystallized Material. During *crystallization from the melt* a lamellar morphology (Fig. 4) develops in which amorphous regions and crystalline lamellae form alternating layers with a well-defined periodicity L, typically in the range 5 – 30 nm. L is thus significantly smaller than the contour length (total length of the chain along its contour) of a polymer chain, and the molecules fold several times within the lamellar morphology (Fig. 1 B). The detailed chain conformation has long been under discussion [44], and various regular and statistical folding models have been introduced. This question has been investigated [4] in small-angle neutron scattering (SANS) experiments. One of the striking results from melt-quenched polyethylene is that the radius of gyration R_g does not change during crystallization (see Table 2), and the chain conformations in melt and quenched crystal are quite similar [7], [20]. Due to entanglements of the chains in the melt, long-range diffusion of segments is not possible, and only segments of the chain straighten during crystallization [45]. The overall chain conformation remains, however, essentially constant. The statistical folding model of Flory [22], [46] for melt- crystallized polymers has been modified to include the true density of chains in the amorphous regions [47] and tie molecules between lamellae [48].

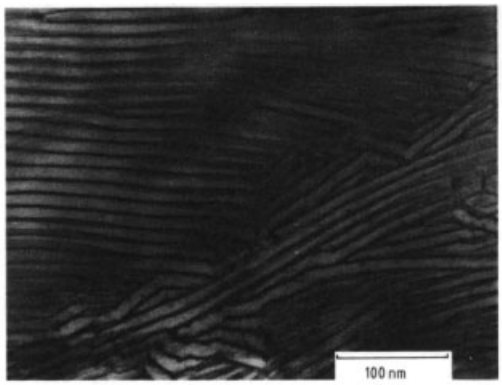

Figure 4. Transmission electron micrograph of polyethylene crystallized from the melt [2]
A thin cut of material has been stained to achieve a contrast between amorphous and crystalline regions. A side view through the crystalline lamellae is shown.

Solution-Crystallized Material. In polyethylene samples crystallized from solution a much more regular folding structure and regular morphology is observed (Fig. 5) [44]. Conformational models include "stacked-sheet" [45] or "superfolding" [49] structures in which crystalline stems of one molecule fold into stacks of an ordered arrangement along the (110) crystallographic direction, which is the main crystallization plane during crystallization from solution. Allowing the molecules during crystallization more time for rearrangement with respect to typical segmental relaxation times results generally in more regular chain folding structures; however, few experiments are performed on slowly crystallized or annealed samples [50]. Good agreement is obtained between SANS and IR experiments [51], which are both sensitive to chain conformation.

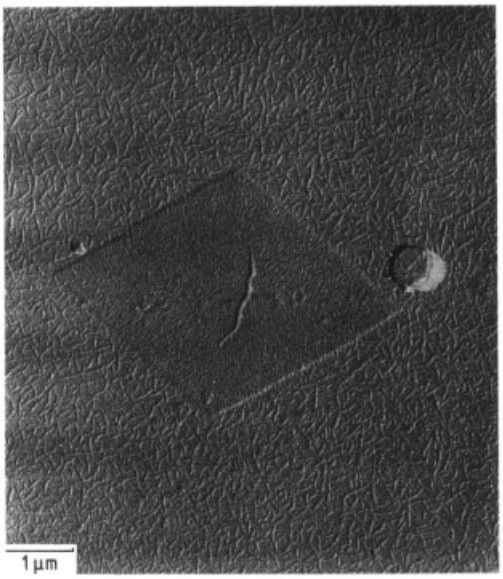

Figure 5. Surface replica [54] of a polyethylene single crystal crystallized from dilute solution
A top view of a crystal lamella is shown. As a result of decoration with low molecular mass polyethylene (evaporation of PE and epitaxial growth on the surface), four quadrants and pleating due to pyramidal shape can be seen.

Degree of Crystallinity and Morphology. The *degree of crystallinity* depends on the material and on sample preparation. For rapidly melt-quenched materials it is near zero, and for some highly tactic or symmetric materials it can reach values of ca. 90 %. Crystallinities of polyethylene and polytetrafluoroethylene for instance may be 80 – 90 %, of highly isotactic polystyrene and isotactic polypropylene 60 – 70 %, and for most other semicrystalline polymers < 50 %.

The crystallinity is determined by X-ray diffraction, whereby the intensity of the amorphous halo is compared to the Bragg peak intensity [52]. The unstructured broad amorphous halo reflects the near range order in the amorphous regions (Fig. 2 A), while the crystalline long-range order gives rise to sharp Bragg peaks. To transform this ratio into absolute values a fully amorphous and fully crystalline sample must be investigated or other techniques (e.g., density measurements) have to be used for comparison. Depending on preparation conditions, different crystalline structures may be observed [43], [52]. Isotactic polypropylene for instance exhibits monoclinic, pseudohexagonal, and triclinic modifications [53].

The *morphology* of crystalline materials is determined by optical or electron microscopy [2], [52]. Whereas single crystals are observed in solution- crystallized material (Fig. 5), various morphologies occur in melt- crystallized materials. Crystalline lamellae (Fig. 4) may be arranged in spherulitic, dendritic, or other supramolecular structures [55]. Specific morphologies (e.g., fibrillar and "shish-kebab" structures) are observed in stretched plastic materials [56–58].

1.3. Liquid Crystals

Plastics can also form liquid crystalline ordered phases in which, like in low molecular mass compounds, smectic, nematic, and cholesteric phases are formed, depending on the composition and temperature [59], [60]. The state of order is between crystals and amorphous materials. In a nematic phase, for example, the center-of-mass position of the rigid mesogenic groups is random (like in an amorphous liquid), whereas mesogenic groups are preferentially oriented in one direction (as in a crystal). Long-range orientational order is present, but only short-range positional order.

Depending on the location of the mesogenic groups, liquid crystalline polymers are classified as side- chain and main- chain polymers (Fig. 6). A potential advantage of polymeric liquid crystals over the corresponding low molecular mass

compounds is the possibility of freezing the ordered state by quenching the polymer below its crystallization or glass transition temperature. A major area of application is thus the production of ordered, high-modulus fibers from liquid crystalline phases [61].

Many main-chain liquid crystalline polymers are believed to be rigid, but they become more difficult to process with increasing rigidity. Various methods have been discussed [60], [62] for making them soluble and processible, including the grafting of flexible side chains (Fig. 6 C) or the introduction of flexible spacers between the mesogenic groups (Fig. 6 B). A potential application is the blending of rigid chains with flexible ones to form reinforced plastics on a molecular level (molecular composites) [63]. The field of liquid crystalline polymers is still growing, and quite different developments and applications may be envisaged.

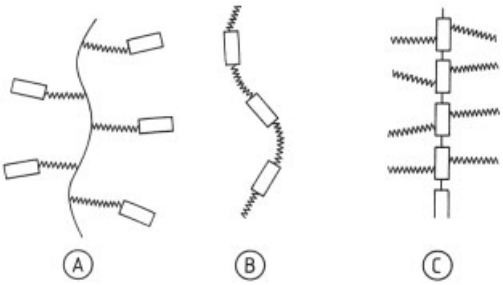

Figure 6. Different types of liquid crystalline polymers Rigid mesogenic groups (rectangle), separated by flexible spacers (wiggly line), are incorporated into the side chain (A) and the main chain (B). For better processibility, flexible groups may be attached to a rigid backbone (C).

1.4. Blends and Block Copolymers

Phase-separated structures are observed in blends and block copolymers in which the components are incompatible [64], [65]. The components are mostly amorphous but can also be crystalline. Observed morphologies can be very regular if, for instance, butadiene – styrene diblock copolymers with narrow molecular mass distributions are used. Depending on the molecular composition, cylindrical or lamellar ordered structures are formed [66]. In some cases [67], ordered crystalline (hexagonal) arrangements of phase-separated domains with macroscopic dimensions are observed. Within the domains, polymers may be amorphous or crystalline, while the superstructure of the domains forms a lattice.

The state of order of blends and block copolymers is very important for their applications since many plastics are blended to improve their properties. The structures can be quite complex, and the compatibility effect of additives acting at the interface between phases is not always easy to understand [68].

2. Properties and Testing

The term high polymers or plastics refers to compounds whose molecules contain more than 1000 – 1500 atoms. These substances display a number of special properties that are unknown in low molecular mass compounds. Such features occur in nearly all realms of physical and engineering properties, as described below.

Typically, the properties of plastics depend strongly on the structure and morphology, which result from the manufacturing process, as well as on loading parameters.

Over the course of plastics development, this specific behavior has dictated the methods used to investigate their properties. Plastics testing today is therefore an independent branch of engineering. Its procedures are discussed in parallel with the description of properties.

The standardization of testing methods (ISO, DIN, and other national standards) is also important for plastic materials. Standardization of plastics types based on their profile of properties is no longer customary except for a few groups of thermosets.

For thermoplastics a classification and nomenclature system has been developed that is consistent for ISO, DIN, and several other standards. For compatibility with modern data bank systems, as well as rationalization, a selection has been made from the large number of test methods available, and the profile of properties has been published as a “table of reference values” [69]. The table contains information on specimen preparation and data on properties related to processing, as well as mechanical, thermal, and optical properties. It is highly suitable for data collections because the properties listed can be compared readily.

2.1. Mechanical Properties and Their Testing

2.1.1. Classification of Plastics Based on Mechanical Properties

The mechanical properties of all groups of plastics cannot be discussed within a single context. Rather, three major classes of polymers exist that differ in their behavior under mechanical loading over a wide temperature range: thermoplastics, elastomers, and thermosets. Their differences in behavior result essentially from differences in physical as well as chemical structure.

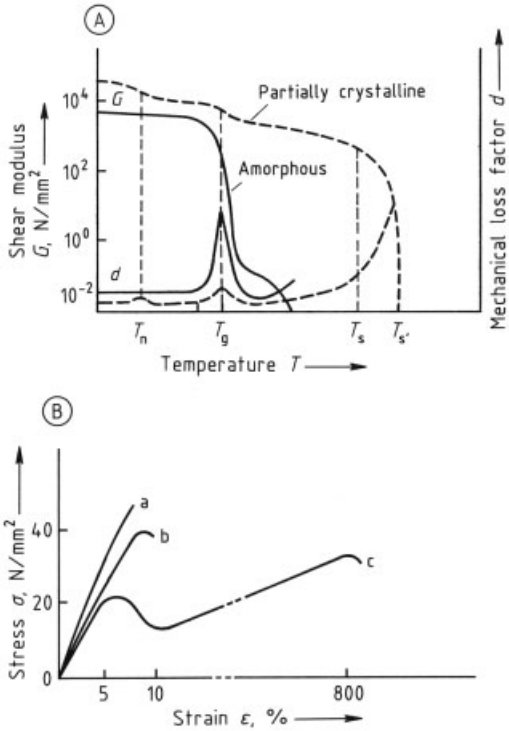

Figure 7. A) Temperature dependence of shear modulus G and mechanical loss factor d for amorphous and partially crystalline thermoplastics (schematic)
T_g = glass-transition temperature; T_n = temperature of a secondary relaxation mechanism; T_s = onset of melting of crystalline regions; $T_{s'}$ = end of melting of crystalline regions
B) Stress – strain curve for thermoplastics (schematic)
a) Amorphous material at moderate temperature; amorphous and partially crystalline material at low temperature or high strain rate; b) Amorphous and partially crystalline material at moderate temperature and strain rate; c) Partially crystalline material at moderate temperature and strain rate

Thermoplastics. Figures 7 A and 7 B plot the shear modulus versus temperature and the stress – strain curve for thermoplastics. This class includes un-cross-linked amorphous or partially crystalline plastics. At low temperature they are energy elastic, but above a certain temperature they exhibit viscous flow. Softening takes place in the glass transition range for amorphous thermoplastics (see page 16) and in the crystallite melting range for partially crystalline thermoplastics. Between the elastic and viscous-flow ranges there may be temperature range in which the material behaves in a rubberlike (entropy-elastic) manner.

In un-cross-linked substances such as thermoplastics, or in only partially cross-linked substances, entropy-elastic and viscous properties are superimposed. These substances are said to exhibit viscoelastic behavior (see Section 2.1.2).

Thermosets. Figures 8 A and 8 B present typical curves of shear modulus and loss factor versus temperature and of stress versus strain for thermosets (based on the curves for phenolic resins differing in degree of cross-linking). Thermosets are closely cross-linked polymers that exhibit energy-elastic or energy/entropy-elastic behavior at low temperature. At or above 60 °C, they show entropy-elastic behavior with very limited deformability. Figure 8 shows how, as cross-linking increases, the glass transition temperature increases and the main transition step becomes smaller, while the curve becomes flatter and the damping maximum becomes lower and broader. The steps disappear altogether upon curing and no main softening occurs in the closely-cross-linked network [70].

Elastomers. Figures 9 A and 9 B show curves of shear modulus and loss factor versus temperature for natural rubber, along with a schematic stress – strain curve. Elastomers are loosely cross-linked polymers that exhibit energy-elastic behavior at low temperature but do not exhibit viscous flow even at high temperature.

Above the glass transition temperature (20 °C or lower), elastomers exhibit rubberlike behavior. In this range, most have shear moduli between ca. 0.1 and 100 N/mm^2 and a high reversible deformability. The stress–strain curve has a characteristic S shape.

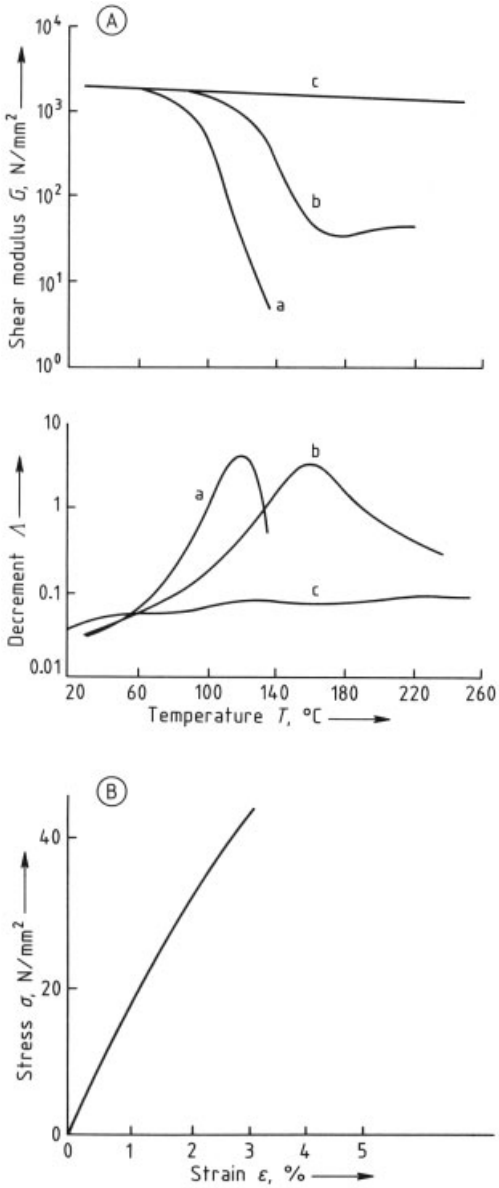

Figure 8. A) Temperature dependences of shear modulus G and logarithmic decrement Λ (as a measure of the loss factor d) for thermosets (phenol – formaldehyde resin)
Cross-linking increases in the order a<b<c.
B) Stress – strain curve for thermosets (schematic)

The glass transition temperature T_g is the mean temperature of the freezing range of a polymer, in which the microscopic Brownian motion of molecular chain segments becomes frozen and the polymer is converted to the glassy state. The temperature dependences of many physical properties change at the glass transition temperature (see Fig. 3). Some properties, such as the modulus, undergo a step-function change.

Energy Elasticity, Entropy Elasticity, and Viscous Behavior. Strictly speaking, *energy-elastic* (perfectly elastic), mechanically and thermodynamically reversible behavior, in the sense of classical elasticity theory, does not occur in any material, particularly not in plastics. The main process in deformation is a distortion of the bond angles of the valence bonds, or else work is done against intra- or intermolecular forces (potential forces).

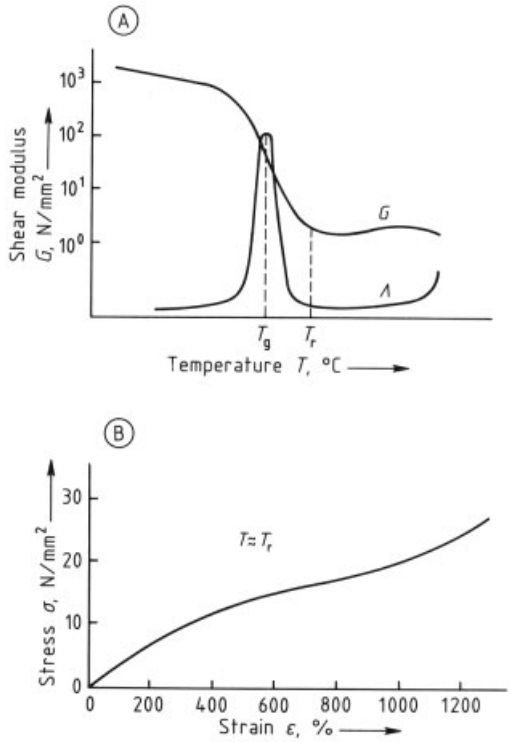

Figure 9. A) Temperature dependences of shear modulus G and logarithmic decrement Λ (as a measure of the loss factor d) for elastomers
T_g = glass transition temperature; T_r = room temperature
B) Stress – strain curve for elastomers (schematic)

In *entropy-elastic* (rubberlike) substances, elasticity does not result from attractive forces but is a kinetic phenomenon like the volume elasticity of a gas. In the ideal case, the internal energy does not change when an amorphous, loosely cross-linked polymeric material network is deformed. Instead, the entropy decreases as

the coiled chains are brought into a more highly ordered state. The hypothesis of constant internal energy implies that the work of deformation must be entirely transformed to heat; that is, a rubberlike substance (elastomer) must become heated when deformed. The assumption of decreasing entropy leads to the conclusion that the stress in a specimen must increase proportionally to the absolute temperature if the strain is held constant, or that if the tensile stress is held constant, the specimen must contract when heated.

Both effects have been the subject of extensive experimental and theoretical studies, initially in thermodynamic terms. However, complete understanding was achieved only with the application of the kinetic theory. This theory cannot be described in detail here [71], but chemical cross-linking of the chains is not absolutely necessary for rubberlike behavior. In partially crystalline polymers, for example, the crystallites act like cross-linking sites. In amorphous polymers under brief loads, entanglements of the chains can act as cross-links. When phase separation occurs in block copolymers, the regions composed of one of the two components play the role of cross-links (physical cross-linking).

Viscous behavior can manifest itself in mechanically reversible creep (i.e., when stress is removed, the specimen returns to the starting position, not spontaneously but after a delay) or in flow (i.e., slipping of molecular chains past one another). Pure flow occurs (even if only as an approximation) in the melt. It is idealized as Newtonian flow, in which the deformation is mechanically and thermodynamically irreversible. However, this idealized state does not exist in polymers (see Section 2.8.3).

2.1.2. Viscoelastic Behavior of Polymers

In ideally elastic solids, the relationship between stresses and the corresponding strains is described by Hooke's law. The stresses and strains can be represented by diadics (second-order tensors), between whose coefficients linear relationships exist. This statement is the general form of Hooke's law. The mathematical expression is $\sigma_{kl} = c_{klmn} e_{mn}$, where σ_{kl} is the stress and e_{mn} is the strain. The 21 constants c_{klmn} required to describe this relationship in the linear range are known as elastic constants. They can be time dependent. In isotropic materials, the number of constants is reduced to two (Lamé constants), both of which can also be time dependent. In practice, other elastic constants or moduli, which themselves are functions of the Lamé constants, are also used. In particular, they are used as parameters in the attempts to mathematically describe the mechanical properties of elastic materials (see below).

In the nonlinear range, the number of elastic constants increases greatly, especially for anisotropic materials. Those of the 21 constants which vanish in the linear range for symmetry reasons have a finite, time-dependent value in the nonlinear range. The number of constants is so large that the measurement of all required constants and their time dependencies, although theoretically possible, is impractical. Thus, in this respect, there are limits to the exact description of the mechanical behavior. The problem is then to make an appropriate selection of constants so that the values measured adequately describe the mechanical behavior in practice.

For more than a century ago, many materials have been known that do not fit into this scheme because their behavior depends on their history. Such effects, not predicted by classical theory, were called "memory effects" (elastic aftereffects) [72], [73]. For many years, the study of this phenomenon was of only academic interest, but the situation changed with the development of plastics.

Plastics were found to have properties exhibited by both elastic and viscous substances; hence, the term viscoelasticity.

Mechanical behavior was studied by means of static and dynamic measurements involving static and fluctuating loads and deformations. The resulting coefficients, however, applied to only a limited experiment. The experimental results were systematized by using a heuristic principle, the superposition principle, which has since been recognized as the mathematical formulation of the fundamental assumption in the theory – assumption of the linearity of viscoelastic effects. Within the limits of this hypothesis, a coherent phenomenological theory could be constructed that left nothing unexplained and provided a complete picture of viscoelastic behavior. The theory cannot be derived and ex-

plained here; for a comprehensive account, see [74].

The effects and functions of the theory can be classed in two groups:

Group I, which relates to creep experiments or measurements with predetermined stress.

Group II, which relates to relaxation effects or measurements with predetermined strain.

With the superposition principle a linear viscoelastic theory could be constructed but stresses and strains can also be interrelated by linear differential equations of the general form $\mathbf{P}\sigma = \mathbf{Q}\varepsilon$, where **P** *and Q* are differential operators with respect to time.

This representation permits an especially advantageous mathematical description of linear viscoelastic behavior in relation to deformation problems with viscoelastic solids. The general explicit form of such differential equations is

$$a_0\sigma + a_1\frac{\mathrm{d}\sigma}{\mathrm{d}t} + a_2\frac{\mathrm{d}^2\sigma}{\mathrm{d}t^2} + \cdots = b_0\varepsilon + b_1\frac{\mathrm{d}\varepsilon}{\mathrm{d}t} + b_2\frac{\mathrm{d}^2\varepsilon}{\mathrm{d}t^2} + \cdots \quad (1)$$

Equations of this type arise when viscoelastic behavior is described in terms of mechanical models consisting of elastic springs subject to Hooke's law, combined with viscous dampers obeying Newton's law of viscosity. Such models are wholly fictitious, except that electrical networks described in a similar way are realistic (see Section 2.4.2).

Mechanical Models and Relaxation – Time Spectra. The basic models are the Maxwell and Voigt – Kelvin models (see Fig. 10).

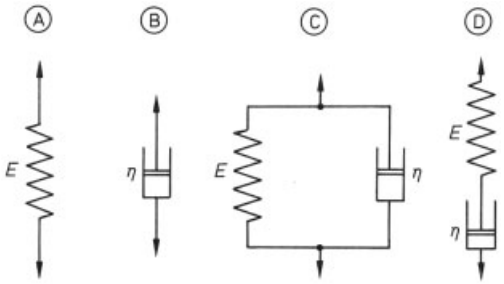

Figure 10. Simple models for describing linear mechanical behavior
A) Spring; B) Damper; C) Voigt – Kelvin model [72], [73]; D) Maxwell model
E = modulus of elasticity; η = viscosity

Maxwell model. The total strain consists of two parts, $\varepsilon = \varepsilon_{\mathrm{spring}} + \varepsilon_{\mathrm{damper}}$, where the stress is the same in damper and spring: $\sigma = E\,\varepsilon_{\mathrm{spring}}$; $\sigma = E\tau \cdot \dot{\varepsilon}_{\mathrm{damper}}$ where $E\tau = \eta$. Appropriate manipulations lead to the differential equation $\dot{\sigma} = E\dot{\varepsilon} - \sigma/\tau$ and hence to the following expressions: for the stress-relaxation test ($\varepsilon = \varepsilon_0$),

$$\sigma = \sigma_0 \mathrm{e}^{-t/\tau} \quad (2)$$

and for the deformation (creep) test ($\sigma = \sigma_0$),

$$\varepsilon = (\sigma_0/E\tau)\,t + \sigma_0/E \quad (3)$$

Similar arguments for the *Voigt model* lead to the differential equation

$$\dot{\varepsilon} = (\sigma/E\tau) - \varepsilon/\tau \quad (4)$$

and to $\sigma = E\,\varepsilon_0$ for the relaxation test and

$$\varepsilon = (\sigma_0/E)\left(1 - \mathrm{e}^{-t/\tau}\right) \quad (5)$$

for the creep test.

These simple models are not, however, suitable for "correctly" describing creep or relaxation behavior, because the Maxwell model represents relaxation alone (stress-relaxation test) and the Voigt model, retardation behavior alone (creep test). The behavior of real polymers requires a combination of at least three or four basic models. A given viscoelastic behavior can be characterized by two arrangements of basic models: (1) a number of Voigt models in series plus a single spring and a single damper, also in series; and (2) a number of Maxwell elements in parallel.

The constants of the two arrangements are linked to one another, and either model can be calculated from the other. The mathematical proof of equivalence is laborious. The model selected depends on convenience. The Voigt type is preferable when the stress is given and the strain must be calculated, and the converse is true for the Maxwell type. Expanded types are referred to as "generalized" Voigt or Maxwell models.

The final step in the analysis is the transition from a generalized to a continuous mechanical model. Such infinite networks are characterized not by a finite number of constants but by a continuous function of an independent variable. These functions, which define continuous models, are called distribution functions. The extension of the theory is not simply a mathematically

elegant solution. The restriction to a finite number of models, each having a well-defined characteristic time (retardation or relaxation time), would imply that processes in real viscoelastic polymers take place with very well-defined characteristic times, but this is true for only a few crystalline materials, not for polymers.

To clarify the situation, consider a generalized Maxwell model with relaxation times $\tau_1 < \tau_2 < \tau_3 < \cdots$. The model is determined if the corresponding spring constants M_i are given. The subdivision into different τ_i is made finer and finer, so that in the limit, the variable τ becomes continuous, taking on all values between 0 and ∞.

The "spring constant" associated with relaxation times between τ and $\tau + \mathrm{d}\tau$ has a value $h(\tau)\,\mathrm{d}\tau$, $0 < \tau < \infty$, where $h(\tau)$ is a continuous function of the relaxation time called the "distribution function of relaxation times" or, for brevity, the relaxation spectrum. If the relaxation spectrum $h(\tau)$ is known for all times, the mechanical behavior of the model is completely determined. In particular, the stress-relaxation modulus $M(t)$ is found by integrating over relaxation time (here with the use of a logarithmic time scale):

$$M(t) = \int_0^\infty H(\ln\tau)\,\mathrm{e}^{-t/\tau}\,\mathrm{d}\ln\tau \qquad (6)$$

The relaxation spectrum cannot, however, be measured directly; it must be calculated from measured phenomenological functions by means of special equations. The calculation is not simple in every case and can be done to only a rough approximation; the spectra are therefore of little importance for purposes of phenomenological description. Their value lies, instead, in the thermodynamic and structural interpretation of the relaxation process. Figure 11 shows relaxation spectra $H(\ln\tau)$ of polyethylene and polypropylene at various temperatures [75]. Inclusion of the temperature dependence leads to a threedimensional representation of the relaxation spectrum. Special reduction methods can be used to extend the spectra to shorter or longer times, after which the plots for polypropylene and polyethylene, for example, exhibit distinct maxima, whose interpretation is discussed in [75]. For example, in polypropylene, the principal maximum at short times is a criterion for mechanisms in amorphous domains, whereas the maximum at long times is a criterion for mechanisms in crystalline domains. Figure 12 shows the relaxation spectrum of polypropylene at 23 °C (after extension to short times) and the temperature traces of the two maxima [76].

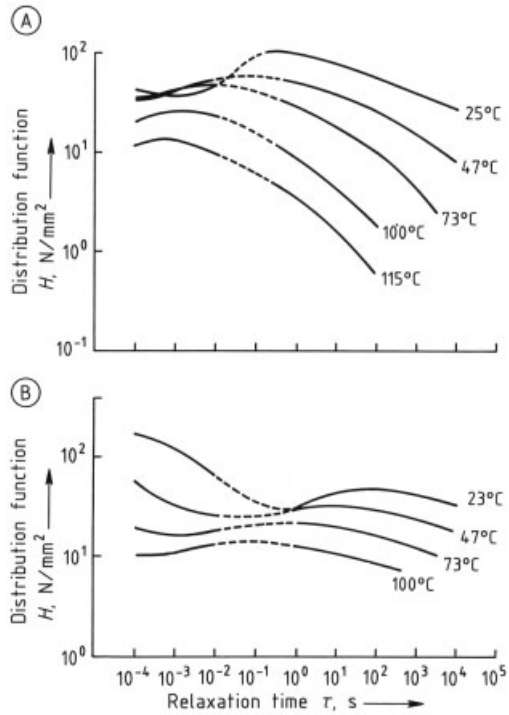

Figure 11. Relaxation spectra at various temperatures
A) Polyethylene; B) Polypropylene

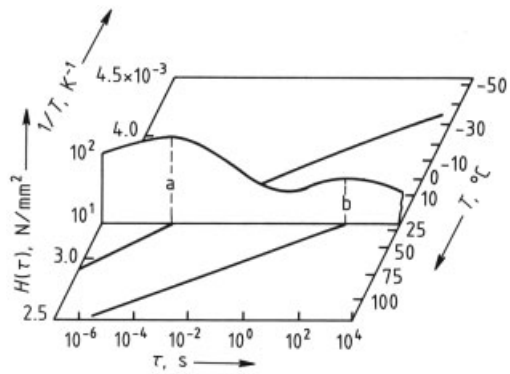

Figure 12. Relaxation spectra of polypropylene as a function of time and temperature [75], [76]
a) Principal maximum; b) Crystalline maximum

As has been mentioned, relaxation spectra can be obtained from creep and relaxation tests only by indirect means; for immediate practical purposes, the results of these "phenomenological" experiments are therefore used directly. They cover loading times $t \geq 1$ s, with no upper bound on the time. These experiments are also referred to as static tests. The model theory of viscoelasticity is thus supplemented by a phenomenological treatment of viscoelasticity.

2.1.3. Mechanics of Deformation

Creep Modulus and Relaxation Modulus. The strongly time-dependent behavior of plastics (especially thermoplastics) under mechanical deformation, resulting from viscoelastic properties, is characterized by the time-dependent creep modulus (retardation modulus) and the time-dependent relaxation modulus. General laws for these can be derived from the viscoelasticity theory (see Section 2.1.2). For the relaxation test, this principle leads to an integral representation in the form

$$\sigma(t) = \int_0^t M(t-\xi)\,\dot{\varepsilon}(\xi)\,\mathrm{d}\xi \tag{7}$$

with variable of integration ξ [77].

In Equation (7), $M(t)$ is the relaxation modulus, which replaces the modulus of elasticity E, the shear modulus G, and the bulk modulus K; $\dot{\varepsilon}$ is the strain rate. The more specific notation $E_\mathrm{r}(t)$ should be used for the modulus $E(t)$ derived from the relaxation test, to distinguish it from the reciprocal compliance $1/C(t) = E_\mathrm{c}(t)$ discussed below, which is determined from the creep test. The superposition principle applied to the creep test leads to an integral representation

$$\varepsilon(t) = \int_0^t C(t-\xi)\,\dot{\sigma}(\xi)\,\mathrm{d}\xi \tag{8}$$

which is analogous to Equation (7). As in the representation of $\sigma(t)$, $C(t-\xi)$ is the time-dependent elastic compliance, which is referred to as the creep compliance and takes the place of the elastic compliance $C_E(t)$, the shear compliance $C_G(t)$, and the bulk compliance $C_K(t)$. In practice, $C_E(t)$ is also written as $1/E_\mathrm{c}(t)$.

The relaxation modulus $M(t)$ and the compliance $C(t)$ are linked by the general relation

$$\int M(t-\xi)\,C(\xi)\,\mathrm{d}\xi = t$$

which implies the condition $M(t)\,C(t) \leq 1$. The equality holds if the time dependence is weak, so that $E\mathrm{r}(t) = E\mathrm{c}(t)$; otherwise, $E\mathrm{r}(t) < E\mathrm{c}(t)$.

If the time dependence is marked, the elastic modulus measured in the relaxation test is lower than that from the creep test. However, this holds only with no stress dependence in the creep test and no strain dependence in the relaxation test. This is not, however, the case in general, so in practice, $E\mathrm{r}(t)$ can also be greater than $E\mathrm{c}(t)$.

The results of creep tests play an important role in construction with plastics. In the creep test, a stress σ is applied to the specimen at time $t = 0$ and held constant thereafter (relative to the initial cross section). The increase in strain with time (creep) $\varepsilon(t)$ is measured. Figure 13 shows how σ and ε vary with time in the creep test. The ε versus t curve is commonly referred to as the creep curve. For sufficiently low stress, Hooke's law in the form $E\mathrm{c}(t) = \sigma/\varepsilon(t)$ gives the time-dependent modulus $E(t)$ from the creep and the constant (initial) stress.

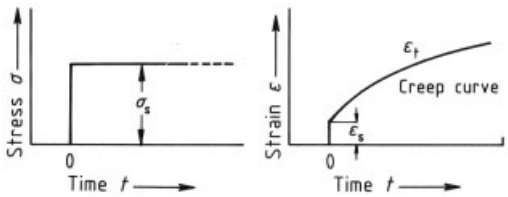

Figure 13. Variation of σ and ε with time in creep test
εs = instantaneous strain (immediately after the application of stress)

The creep modulus $E\mathrm{c}(t)$ is independent of the stress only if the stress lies in the Hooke's law range, where the stress and the strain measured after a given time are proportional to one another. Outside the linear range, the creep modulus decreases with increasing stress. When values are cited for the creep modulus, the stresses at which they were determined must be specified. In practice, uniaxial tensile loads are usually applied in creep testing, chiefly because this loading mode gives an identical state of stress at all points in the specimen. The performance of tensile creep tests is governed by DIN 53 444 [78]. The specimens used are the same as for the

tensile test of DIN 53 455 [79]. Tensile specimens are loaded by the impactless application of weights, and strain is measured at fixed time intervals. A plot of the results gives the above-mentioned creep curve. The test involves three variables: stress, determined by load and (initial) specimen cross section; strain; and loading time. Two-dimensional plots, each with two explicit variables and one as parameter, yield the family of curves shown in Figure 14. For clarity's sake, only one stress – time curve (yield limit curve) and one isochron are plotted.

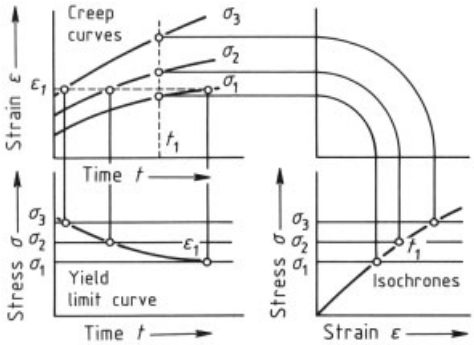

Figure 14. Derivation of yield limit curves and isochrons from creep curves

The creep modulus $Ec\,(t)$ is derived from a family of creep curves with stress as parameter; the result can be plotted, for example, versus time, as shown in Figure 15. Plotting points for the respective stress – strain values from the creep curve, with loading time as parameter, gives so-called isochronous stress – strain curves, which show the time and stress dependence clearly. Fields of isochrones for a number of temperatures convey much information to the designer (see Fig. 16).

The stress σ that leads to a total strain ε after time t is called the time yield limit $\sigma_{\varepsilon/t}$.

The curve connecting creep stresses for equal strains is now known as a "stress versus time curve" (formerly termed yield limit curve; see Fig. 14). The set of creep curves forms the creep diagram, which may also contain the test stress σ and the time t to fracture (or to a certain sign of damage). These give the creep rupture curve or a certain "damage curve." The term creep strength curve is also commonly used for the creep rupture curve, because ultimately it is the curve joining the creep strengths. The creep strength $\sigma_{B/t}$ is the static load B, relative to the initial specimen cross section, that causes fracture of the specimen after a certain loading time t. The creep diagram can be used to estimate the time-dependent deformation occurring under uniaxial tensile loading and is a useful design aid for simple loading modes (Fig. 17). The time-dependent modulus of elasticity, however, is more universal in its application because, in many cases, it can also be used in the equations for materials engineering.

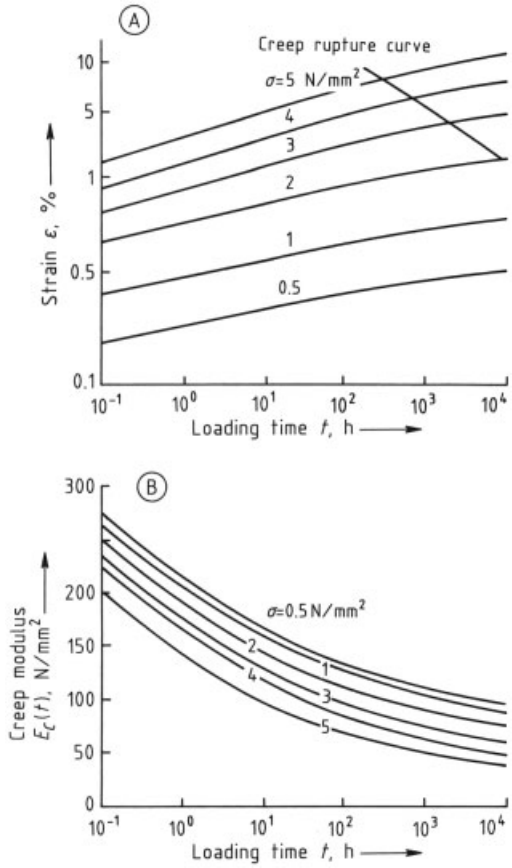

Figure 15. Creep curves and creep-modulus curves
A) Creep curves at 65 °C for high-density polyethylene with creep rupture curve from a long-term pressure test on high-density polyethylene pipes
B) Creep-modulus curves derived from creep curves

The relaxation test is experimentally far more difficult to perform than the creep test. The deformation must be held constant over time, and the mechanical stress must be measured as a function of time. For details, see [81].

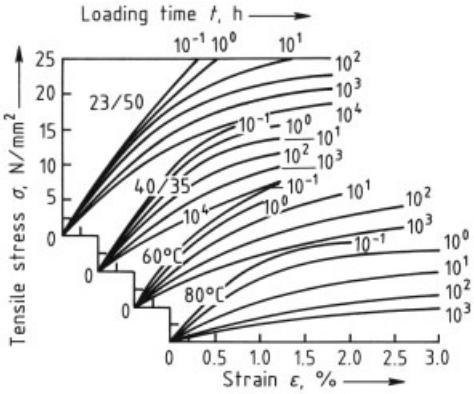

Figure 16. Isochronous stress – strain curves for acrylonitrile – butadiene – styrene; $ak = 12$ kJ/m2 [80] 23/50 = 23 °C/50 % relative humidity; 40/35 = 40 °C/35 % relative humidity

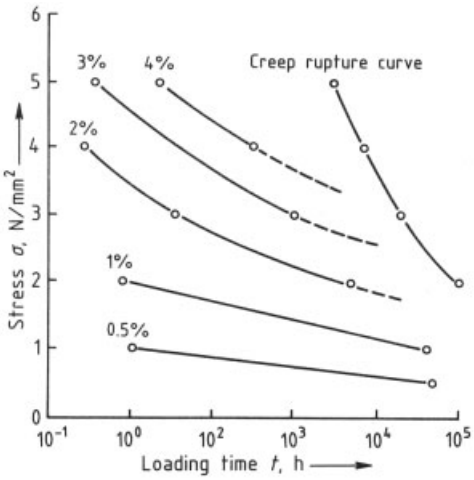

Figure 17. Creep diagram for high-density polyethylene with 0.5, 1, 2, 3, and 4 % yield limit curves and the creep-rupture curve

Vibration Tests. Static tests involve loading times longer than ca. 1 s. If information about the millisecond range is needed, vibration tests must be performed. These usually cover two orders of magnitude in frequency; the ranges in conventional vibration tests are roughly between 10^{-1} and 10^4 Hz, corresponding to time t (if $t = 1/\omega$) in the range of 10^{-5} to 1 s.

As is customary in treatment of vibration, complex quantities are employed to describe the phenomena. The theory leads to the definition of a complex modulus (see Fig. 18):

$$\boldsymbol{M} = |\boldsymbol{M}|\,\mathrm{e}^{\mathrm{i}\delta} = M' + \mathrm{i}M''$$

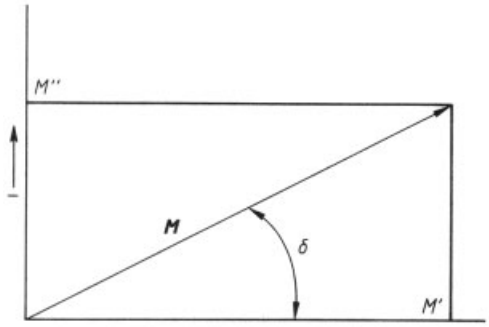

Figure 18. Representation of M in the complex plane

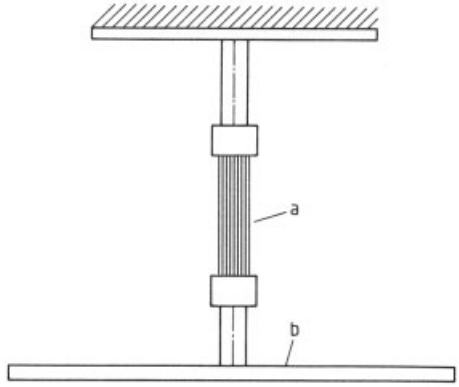

Figure 19. Schmieder – Wolf torsional pendulum a) Specimen; b) Rotating mass

where M', the real part of $\boldsymbol{M}$, is called the dynamic modulus (as distinct from the time-dependent modulus $M(t)$; see page 14) or the storage modulus. It is a measure of the recoverable energy transformed in the deformation occurring during one vibration. The parameter M'', the imaginary part of $\boldsymbol{M}$, is called the loss modulus. It is a measure of the vibrational energy unrecoverably transformed to heat (dissipated). The quotient of the real and imaginary parts of $\boldsymbol{M}$, which is the tangent of the phase angle δ of the complex modulus, is known as the mechanical loss factor $d = M''/M' = \tan\delta$. It is a relative measure of vibrational energy losses compared to recoverable energy.

The chief methods used in plastics testing are the torsional vibration test and the bending vibration test. The various techniques for exciting torsional vibrations have in common that a specimen acting as a torsion spring is connected to an oscillating mass. Figure 19 shows one possible arrangement.

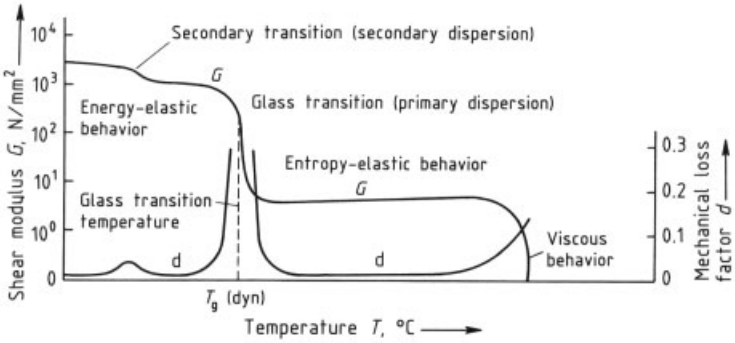

Figure 20. Definition of state regions based on the curve of shear modulus G versus temperature T (DIN 7724)

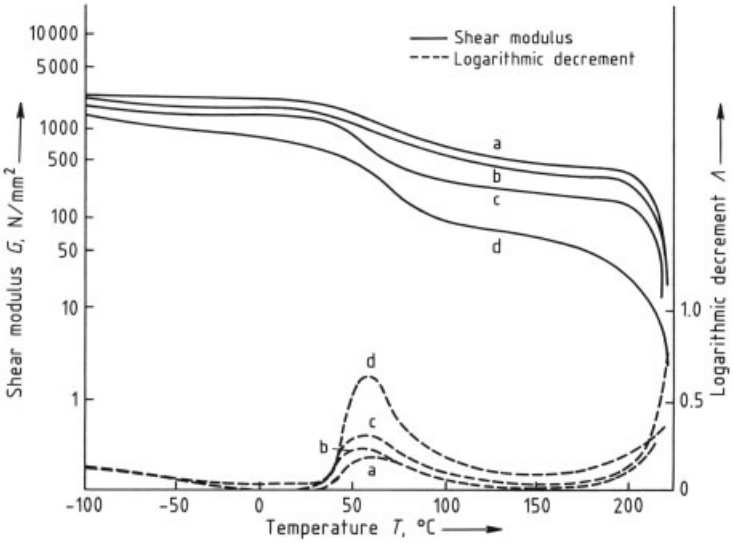

Figure 21. Temperature dependences of shear modulus and logarithmic decrement for poly(butylene terephthalate)
a) 30 % Glass fiber; b) 20 % Glass fiber; c) 10 % Glass fiber; d) Unreinforced

If the upper specimen clamp is moved slightly, this torsional pendulum is set into damped natural oscillation. The logarithmic decrement $\Lambda = \ln(A_n/A_{n+1})$ and the mechanical loss factor $d \approx \Lambda/\pi$ (for $\Lambda \leq 2$) are calculated from the amplitudes A of successive oscillations. The shear modulus G is determined from the moment of intertia $\boldsymbol{I}$ of the oscillating mass, the square of the oscillation frequency f, and factors F_g and F_d relating to the dimensions and the damping effect, respectively:

$$G = \boldsymbol{I} f^2 F_g F_d$$

The test method is described in detail in DIN 53 445 [82]. Figure 20 shows how the curve of shear modulus G versus temperature T can be used to separate the various states of high polymers. The glassy range is characterized by high modulus and low damping. The diagram also shows clearly the temperature below which the substance described by this G versus T curve behaves as a solid. Temperature ranges in which the modulus is strongly temperature dependent, while the damping has a maximum, represent softening ranges. The measured results also show whether flow sets in as soon as softening begins or whether cohesive forces are still present. Figure 21 gives examples of shear modulus curves over a wide temperature range.

Because of the amount of information it provides, the torsional vibration test has also been incorporated into the table of reference values mentioned earlier (see page 9).

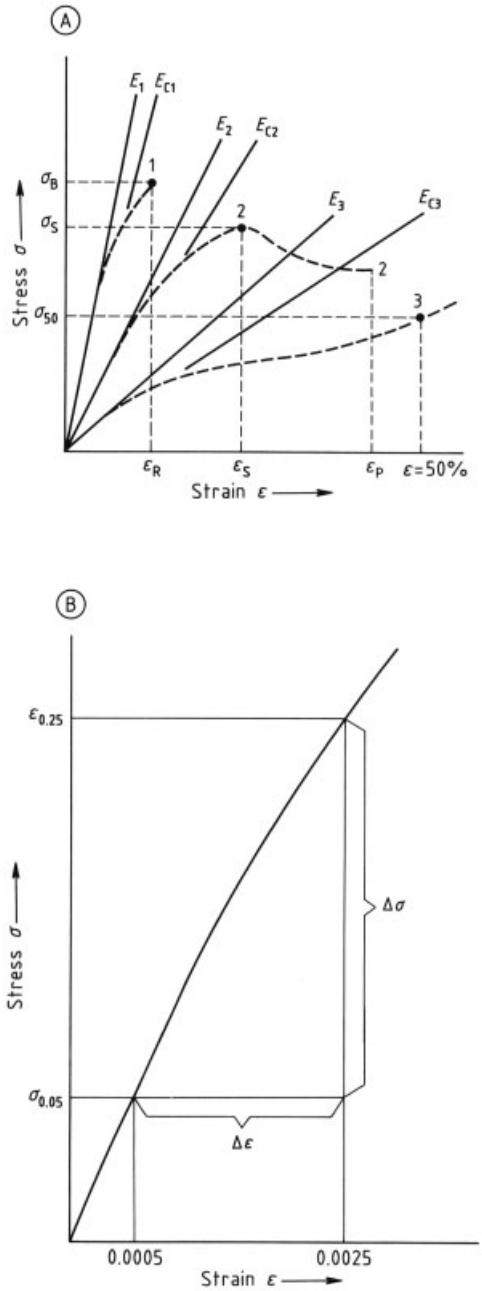

Figure 22. Evaluation of tests with steady load
A) Values from tensile test, for three typical shapes of stress – strain curves [86]
E = elastic modulus; E_c = creep modulus after 10^3 h; ε_R = elongation at rupture, ε_S = extension; σ_P = tensile strength; σ_S = yield stress; σ_{50} = stress at 50 % strain
B) Determination of elastic modulus (DIN 53 457/1987)
$\sigma_{0.25}$ = stress at 0.25 % strain; $\sigma_{0.05}$ = stress at 0.05 % strain; $\Delta\sigma = \sigma_{0.25} - \sigma_{0.05}$, $\Delta\varepsilon = 0.002$ (= 0.2 %)

In the bending vibration test [83], the top end of a rod-shaped specimen is clamped in a fixed position and the rod is excited to steady-state resonance bending vibrations by an electromagnetic transducer at the bottom. The amplitude of the vibration is measured by a transducer near the clamping point. The modulus of elasticity is calculated from the resonant frequency, and the loss factor from the width of the resonance curve. The wide frequency range employed in this method (1 – 10 000 Hz) is an advantage of the bending vibration test. Because frequencies lie in the audible range, the test is particularly important for studies relating to acoustic applications of plastics for example for noise suppression [84]. The torsional vibration test gives the shear modulus G, whereas the bending vibration test yields the elastic modulus E. To a good approximation, the conversion formulas are

$$E = 2G\,(1+\mu) \text{ or}$$
$$E = \frac{3G}{1+G/\,(3K)}$$

where μ is Poisson's ratio and K is the bulk modulus; see also [85].

Behavior under Steady and Fluctuating Loads. To characterize the time-dependent load-deformation behavior, the laborious long-term and vibration tests are supplemented by steady-load and fluctuating-load tests. These techniques are important for product development, inspection, and preliminary materials selection. Tests with steady load include tensile, bending, and compressive tests, as well as hardness measurement; those with fluctuating load include flexural fatigue tests. Studies of friction and wear properties complete the series of mechanical measurements. For the designer, however, by far the most informative results are the temperature and time functions describing the long-term behavior under tensile loading together with the curves of shear modulus versus temperature. The predictive power of all other methods is limited by comparison [86]. The tensile test [87] is, however, largely exempt from the limitations mentioned. The elongation and strength values from this test, given uniform strain and a homogeneous stress distribution, allow a good assessment of the load deformation behavior. Figure 22 A presents three typical curve shapes [86], along with the stress and

strain values listed in the table of reference values for plastics.

Outside the linear range of the stress – strain curve, the bending test [88] gives unrealistically high values. This test is therefore no longer included in the table of fundamental values [89].

Compression tests [90] are important in the evaluation of some thermoset molding compounds.

Static Modulus of Elasticity. The elastic modulus derived from stress – strain curves (or force –deformation curves) recorded in steady-load tests, like the values from tensile, bending, and compressive tests, is a single-point value with respect to strain rate. It has adequate predictive power for product development, inspection, and materials preselection in most cases. Derivation of the modulus for a tensile load is illustrated in Figure 16 B.

The applicable standard [91] specifies a uniform rate of extension (crosshead speed) of 1 mm/min and restriction of the evaluation to the stress – strain curve between 0.05 and 0.25 % strain. Because the stress-strain relationship for rigid plastics is quasi linear in this range, what is obtained is an approximate tangent modulus. A secant modulus is found for soft plastics with a nonlinear stress-strain relationship.

The standard does not, however, distinguish between tangent and secant moduli in the listing of values. Thus, the elastic modulus in this standard is given by

$$E = \frac{\sigma_{0.25} - \sigma_{0.05}}{0.002} \left(\text{units of } E \text{ and } \sigma\text{: N/mm}^2\right)$$

where $\sigma_{0.25}$ is the stress at 0.25 % strain and $\sigma_{0.05}$ is the stress at 0.05 % strain. Table 3 lists tensile moduli of elasticity for several reinforced and unreinforced plastics.

Multiaxial Loading. The strengths measured under steady load, such as the tensile strength, apply to a one-dimensional load system. Application of these values to multiaxial stresses requires that the multiaxial states be linked with the individual strengths through fracture criteria that reduce the multiaxial state to a single reference stress. Definition of this reference stress depends on the fracture hypothesis selected. Finally, the reference stress is compared with the individual (usually tensile) strength. The fracture criteria used for metals cannot as a rule be applied to plastics, because isotropic fracture behavior – the precondition for most of the classical criteria – is generally not exhibited by plastics [92].

Table 3. Static modulus of elasticity for unreinforced and reinforced plastics (maximum value in each case)

Product (abbreviations from DIN 7728) ∗	Elastic modulus E, N/mm^2 ∗∗
ABS, moderate impact strength	2 500 (T)
	2 300 (B)
ABS+30 % GF	6 000 (T)
	4 500 (B)
PS	3 500 (T)
SB	3 000 (T)
SAN	3 800 (T)
ASA	2 900 (T)
PTFE	750 (T)
FEP	700 (T)
PFA	700 (T)
PE-HD (HDPE)	1 400 (T)
PP	1 800 (T)
PP+40 % T	4 500 (T)
POM (Copolymer)	3 200 (T)
POM+30 % GF	10 000 (T)
PBT	2 700 (T)
PBT+30 % GS	4 000 (T)
PBT+30 % GF	10 000 (T)
PBT+30 % CF	16 000 (B)
PA 6 (d)	2 500 (T)
PA 6+30 % GF (d)	8 300 (B)
PA 6+30 % CF (d)	16 000 (B)
PA 66 (d)	3 000 (T)
PA 66+30 % GF (d)	10 000 (B)
PC	2 200 (T)
PC+30 % GF	7 000 (T)
PPO	2 000 (T)
PPO+30 % GF	9 000 (T)
PEEK	3 800 (B)
PEEK+30 % CF	13 500 (B)
PPS	4 200 (B)
PPS+40 % CF	19 600 (B)
PMMA	3 300 (T)
PVC	3 300 (T)
PVC-HI	2 600 (T)
PES	2 600 (B)
PES+30 % CF	14 300 (B)

∗ GF = glass fiber; T = talc; GS = glass spheres; CF = carbon fibers; d = dry;
∗∗ T = tensile modulus; B = bending modulus.

Classical failure criteria must therefore be modified to describe more accurately the failure behavior of polymers. A detailed discussion of these modifications can be found in [93].

Hardness. Hardness is not a fundamental property of plastics; the concept is ultimately subjective. Probably the most widely acknowl-

edged definition is resistance to the penetration of another, harder body.

As in methods for metals, hardness testing on plastics employs a standard penetrator, often a hardened steel ball or steel pin, that is pressed by a defined force into a plane surface of the material being tested, with the depth of penetration being measured [94–97].

The viscoelastic behavior of plastics results in two basic difficulties: (1) the depth of penetration depends on the rate of loading, and (2) measuring the diameter of the impression is difficult. Because the impression springs back after the load is removed, it would have to be measured immediately. Thus, either the depth of penetration is measured after a given time under load, or a transparent penetrator is employed.

Not only does the hardness depend on time, temperature, and the nonlinear stress – strain relationship, it also depends on the magnitude of the forces applied, the modulus, and the shape of the penetrator. Accordingly, hardness measured by a given method cannot be compared with that obtained by another method. The hardness of elastomers and other "soft" materials is characterized by the Shore hardness value [97], while that of rigid thermoplastics and all thermosets is expressed as the ball indentation hardness [95].

Behavior under Fluctuating Load (Fatigue). A large fluctuating load applied many times can lead to material damage that results in failure (fatigue failure) at a load value far below the strengths measured in steady-load tests (tensile strength, tensile creep strength). Test methods devised for metals have been extended to plastics, with allowance for the fact that metals, in contrast to plastics, have low damping and high thermal conductivity. The loading frequencies for plastics are therefore markedly lower than for metals, lying between ca. 600 and 3000 min^{-1}.

Fluctuating loads are classified as loads with constant stress and those with constant strain. The many possible forms of loading cannot be discussed individually here. As Figure 23 shows, three loading ranges are distinguished: fluctuating compressive, alternating, and fluctuating tensile. In the repeated flexural fatigue test (Wöhler method), several (e.g., six to ten) equivalent specimens are subjected to appropriately graduated oscillating loads, and the number of cycles to failure is recorded for each. The same mean stress σ_m is chosen for all specimens, and the initial stress amplitudes are graded. The mean stresses and stress amplitudes, which vary within a test because of relaxation and heating, are then plotted against the common logarithm of the number of cycles. The curves obtained are known as run curves, and the complete set of curves obtained in the test is the family of run curves. In addition, the Wöhler curve proper (σ_a–N or stress-life curve) is plotted; its points represent pairs of initial stress amplitude σ_a and cycles to failure N, plotted with a logarithmic horizontal (N) axis and a linear vertical (σ_a) axis. Figure 24 presents Wöhler curves for polystyrene, styrene – butadiene, and acrylonitrile – butadiene – styrene (ABS) [80].

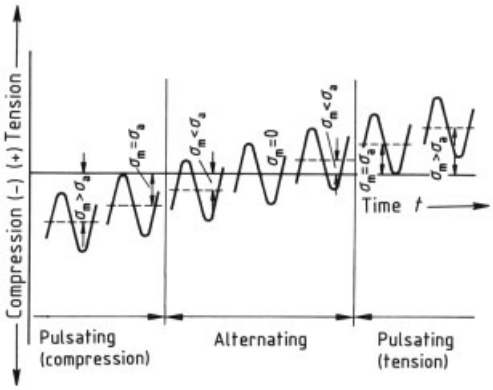

Figure 23. Load ranges in vibrational test
σ_0 = maximum stress; σ_u = minimum stress; σ_m = mean stress $[\sigma_\mathrm{m} = 0.5\,(\sigma_0 + \sigma_\mathrm{u})]$; σ_a = stressexcursion $[\sigma_\mathrm{a} = \pm 0.5\,(\sigma_0 - \sigma_\mathrm{u})]$

Service Temperature Range. The best way of assessing the service temperature range is to use curves showing the properties of interest as a function of temperature. In many cases, functional tests on a finished part can also be employed. Nonetheless, evaluating the service range – in relation to, say, mechanical behavior – by performing standard tests on specimens is often useful. This is successful, with some reservations, for determining the upper limit service temperature (heat stability) but not for determining brittleness temperatures. Attempts to devise a satisfactory technique for measuring these have not been successful.

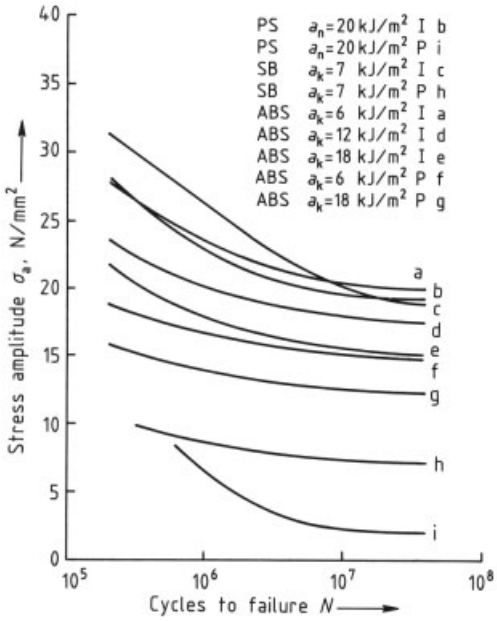

Figure 24. Wöhler curves for polystyrene, styrene – butadiene, and ABS, after Oberbach [71] a_n = impact toughness; a_k = notch impact toughness, S = injection-molded specimen; P = molded specimen

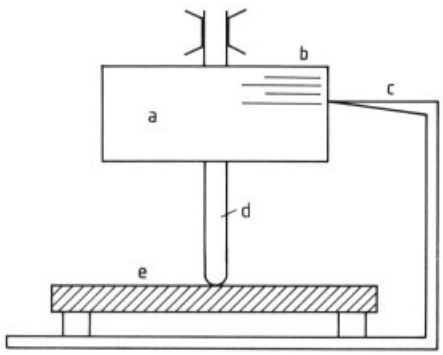

Figure 25. Setup for the HDT method (schematic) a) Weight; b) Scale; c) Indicator; d) Bending mandrel; e) Specimen

The DIN heat stability tests are the Martens test [98] and the heat deflection temperature (HDT) test among [99]. The Vicat softening temperature (VST) indicates the softening behavior at the upper limit of the service temperature range [100].

Figure 25 shows a schematic of the HDT test. The heat deflection temperature is the temperature at which the bending test specimen, loaded in the middle, deforms by a given amount as its temperature is increased. Various test methods differ in the magnitude of the bending stresses applied. In the Martens test, the bending moment acts in a four-point loading scheme. The VST is the temperature at which a 1-mm^2 circular steel pin under a given force penetrates 1 mm vertically into the specimen as the temperature is increased.

2.1.4. Friction and Wear

Friction is defined in general as the resistance that must be overcome to set or maintain two contacting bodies in relative motion. The coefficient of friction μ is calculated from the normal force N between the contacting surfaces and the frictional force F opposing their motion: $\mu = F/N$. In other words, the frictional force is proportional to the normal force (first law of friction). This formula holds approximately for metals; the coefficient of friction for plastics decreases with increasing normal force. Friction measurements on plastics should therefore employ a graded series of normal forces.

The concept of friction is often broken down into static and dynamic friction, even though perfectly static friction is impossible. The term static friction normally refers to the amount of force required to induce a relative motion between two surfaces in contact; that is, what is measured is the frictional force with very slow movement.

Dynamic friction means the force that will maintain a constant, perceptible, relative motion between contacting surfaces. Velocity thus plays a role in determining the frictional force, even though for many materials the frictional force changes almost negligibly as relative velocity changes by several powers of ten. This finding corresponds to the third law of friction. For plastics, however, it does not hold perfectly. The frictional force may change by a factor of three or even four as the velocity varies over several powers of ten. Unfortunately, the velocity effect cannot be wholly separated from the temperature effect, but the velocity dependence must still be determined for plastics. The heat produced in a sliding contact is approximately proportional to the coefficient of friction, the sliding velocity, and the normal force.

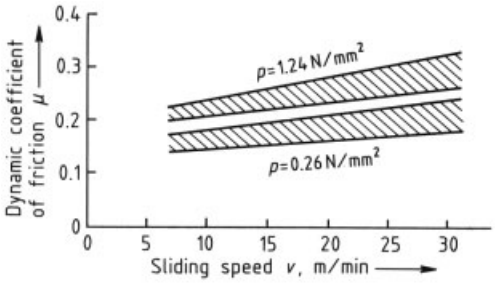

Figure 26. Dynamic coefficient of friction μ as a function of sliding velocity v for ultrahigh molecular mass polyethylene at various surface pressures
p = mean surface pressure

Table 4. Wear (reduction in specimen diameter) of some plastics in dry sliding against 16 MnCr 5 steel

Product (abbreviations from DIN 7728)	Wear, µm per km of sliding distance
PA 66	0.09
PA 6	0.23
PA 610	0.32
PA 11	0.8
PET (PETP)	0.5
POM (copolymer)	8.9
PE-HD (HDPE)	1.0
PE-LD (LDPE)	7.4
PP	11
PTFE	21
PS	115
SAN	23
SB	51
ABS	8.4
PMMA	4.8
PVC	5.6
PC	22

Figure 26 shows how the coefficient of dynamic friction depends on the sliding velocity for ultrahigh molecular mass polyethylene under various pressures.

In general terms, wear refers to all abrasive processes involving mechanical energy. The mechanisms governing wear are complex and intimately interconnected (see also → Abrasion and Erosion). For plastics, relatively little experience has been gathered with regard to the effects of various parameters. The following methods are considered proven: The wear in dry sliding against a granular material is determined by the friction wheel technique [101]; a similar method is described in [102]. Another method measures the wear of plastics in dry sliding against a rotating steel shaft. This technique is used to test plastics for sliding bearings, especially those designed to run dry [102]. Table 4 lists values measured in a setup of this type.

2.1.5. Fracture Behavior

Plastics exhibit three types of deformation and fracture behavior:

1) Quasi-brittle fracture
2) Fracture after partially or completely irreversible shear deformation
3) Fracture by crazing after irreversible deformation of the specimen

This classification applies mainly to thermoplastics; fracture in thermosets is nearly always brittle. The fracture behavior of elastomers is described by the theory of Bueche and Halpin [103].

Brittle Fracture. The term brittle fracture denotes fracture with virtually no deformation, characterized by a high propagation velocity. If the latter is placed in the foreground, brittle fracture can be said to take place even if some plastic deformation occurs first. Brittleness and toughness as material properties are influenced by test conditions and depend on the temperature, the rate of loading, and the stress state. In this sense, any plastic experiences low-deformation fracture in some range or ranges.

Despite their brittle behavior, plastics have a relatively high specific fracture energy (energy to produce 1 cm^2 of fracture surface), roughly a hundred times that of inorganic glasses. The reasons for this are that (1) the edges of an advancing fracture undergo substantial irreversible deformation, even in the glassy state; and, (2) the numerous crazes formed in the material act not as energy absorbers but as starting points for new fractures (see page 22).

Fracture Mechanics. A few words should be said about the applicability of fracture mechanics to plastics. The fundamental idea of fracture mechanics is to analyze the mechanical behavior of an individual crack in a linearly elastic, homogeneous, and isotropic continuum. This basic concept is, however, valid only if the solid shows linearly elastic behavior even in its smallest regions, that is, at the tip of the crack. In other words, the material must be perfectly brittle. The energy transformation is then simply the conversion of supplied elastic energy to

pure surface energy. Even for perfectly brittle glasses, however, the energy balance is not this simple; the "supply side" includes elastic energy derived from internal stresses, whereas on the "demand side" the liberation of plastic deformation energy plays a dominant part, especially for crack propagation in plastics. The specific surface energy is orders of magnitude lower than this deformation energy.

To describe this ductile behavior, the concept of crack resistance has been introduced, which describes ductile behavior without the necessity of knowing or postulating details about the shape, size, and time behavior of the plastic zone itself. Fracture mechanics becomes less and less valid as the crack-tip plasticity increases. A number of models of this plastic behavior exist, including that of DUGDALE.

In the area of *quasi-brittle fracture,* however, linearly elastic fracture mechanics (LEFM) offers a fairly good description even of viscoelastic materials. If certain fitting parameters are introduced, the framework of LEFM can also be extended to materials with small plastic deformation components. As the deformation behavior becomes more strongly plastic to completely plastic, the corrections are no longer sufficient and the linearly elastic solution ceases to be meaningful. For such cases, however, viscous fracture mechanics is based on a different set of principles and may provide a fracturemechanical description. The application of fracture mechanics to polymers is reviewed in [104].

Fracture after Partial or Complete Shear Deformation. A ductile fracture, in contrast to a brittle one, is characterized by large amounts of macroscopic deformation. The deformation may arise through uniform elongation or, in true ductile fracture, by yielding or necking. In metals and ceramics, direct inspection reveals lattice defects (dislocations) in this yielding region. Movements of the dislocations are responsible for observed plastic or permanent deformation; such motion does not occur below the tensile yield stress. This behavior provides a basis for clearly distinguishing between elastic and plastic (irreversible) strain in metals and ceramics. The curves plotted for plastics are very similar to those for metals (Fig. 27). The analogous behavior of metals and some plastics is underlined further by the occurrence of necking. In thermoplastics, for example, one or several such necking zones migrate over the specimen until the whole has experienced yielding (unless it has fractured first). The behavior can be described by the Eyring yield equation; detailed discussion and calculations based on a kink model are presented in [105].

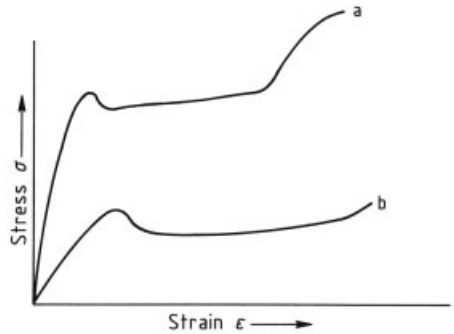

Figure 27. Stress – strain behavior of materials with yielding
a) Metals; b) Thermoplastics

In metals, the planes of maximal shear stress (e.g., a plane at 45° to the direction of uniaxial loading) exhibit so-called Lüders lines [106]. These are tracks left by slip lines along which the parts of the specimen are displaced. The lines are regarded as evidence of a dislocation-dependent deformation process. Similar lines are observed in both crystalline and amorphous plastics. Their occurrence in crystalline plastics is interpreted in terms of a dislocation-dependent slip mechanism, but this interpretation does not apply to amorphous materials. ROSEN advanced a theory based on the observed birefringence [107]. According to this hypothesis, birefringence results from a new, noncrystalline state of aggregation. The lines are a domain of oriented material having a refractive index different from that of the surrounding material.

Fracture after Deformation with the Formation of Crazes (*Flow Zones*). Crazes or flow zones are planar inhomogeneities in a high polymer, produced by plastic deformation in the direction of an applied tensile stress [108–110]. The planar surface of the craze forms perpendicular to the maximum principal tensile stress (Fig. 28). A craze can be as large as 1 cm^2 in area but no more than 4 μm thick.

Surface crazes display arched crack edges or dislocations. In contrast to a crack, a craze contains material in a highly stretched, unthinned

state and behaves as an optically homogeneous medium. The craze stands out sharply against the surrounding normal polymer material.

STERNSTEIN et al. have investigated the conditions under which deformation through craze formation (normal yielding) occurs [111], [112]. The results show that crazes can form only in the vicinity of a positive isotropic stress component (stress invariant I_1): $I_1 = \sigma_1 + \sigma_2 + \sigma_3 > 0$. However, even when this condition is fulfilled, crazes only occur when the maximum principal normal stress exceeds a time- and temperature-dependent critical value σ_{crit}.

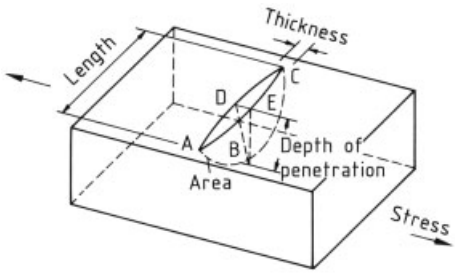

Figure 28. Schematic illustration of a craze
AC = length; DE = thickness; EB = depth of penetration

The formation of flow zones in impact-resistant two-phase or multiphase systems is of great engineering importance. When flow zones develop in these systems, they act as energy absorbers. All analyses are based on the assumption that elastomer particles in such a system act as stress concentrators under mechanical load, because their elastic modulus is low compared to that of the matrix. BOHN has further shown that the soft phase must have a sufficiently uniform distribution and must be sufficiently cross-linked and grafted in order to produce and maintain the special dilatation stress state in the hard matrix on cooling [113].

Fracture Morphology of Plastics. So many factors affect fracture morphology in plastics that only a few examples can be discussed here. Brittle plastics exhibit a behavior similar to that studied in detail by SMEKAL for inorganic glasses. According to those findings, the fracture front advances radially in all directions from a primary notch site. The primary notch is surrounded by a smooth surface, the "mirror," which is bounded by a semicircle. SMEKAL states that the transport of heat in the initial phase (i.e., in the "mirror region" of the fracture) is an essential feature of the fracture phenomenon. This region of the fracture, which is traversed at a low velocity, is therefore referred to as the thermal phase of the fracture. The remainder of the process occurs at a very high fracture velocity (i.e., athermally), and this phase is called the athermal phase of the fracture. The fracture mirror and the athermal phase are also observed in brittle plastics, especially in fracturing after high strain rates (Fig. 29) [114]. The fracture mirror often displays periodic fine structures that can yield information about the rate of loading. In ductile materials, fissured morphologies are observed; these depend on, among other factors, orientation, stress distribution, and rate of loading.

A point of special interest is the fracture morphology of fiber-reinforced plastics. Fracture surfaces are observed by optical microscopy or by scanning electron microscopy at a magnification of 60 – 24 000.

Measurement of Fracture Behavior. Instrumental investigation of fracture behavior takes place in the context of steady-load and fluctuating-load mechanical tests, both short- and long-term, with tensile, compressive and bending loads. These tests are performed at moderate to low strain rates.

A special test procedure has been developed to study deformation and fracture behavior at high strain rates (impact behavior). To a large extent, plastics owe their great versatility to high impact strength combined with other qualities. Hence, many studies have dealt with the property of impact strength. The translation from values measured on test specimens to in-service performance (the behavior of moldings and semifinished products in a variety of applications) is more difficult for impact strength than for the mechanical properties already discussed.

For example, the influence of viscoelastic behavior can be estimated only in the quasi-linear range. At the extreme loads, required for evaluating performance under impact loading, which usually far exceed the linear range (especially for unreinforced materials), other tests must be used. These include impact flexural and impact tensile tests, both performed with notched-bar

specimens. Studies are also carried out with multiaxial loading modes.

As in low-strain-rate tests, the stress – strain diagram or the deformation energy absorbed up to a certain damage level (usually to fracture) is determined.

Numerous standardized tests have been developed for this purpose. In general, however, they cannot be assessed in terms of how well they correlate with in-service performance under impact and shock loads.

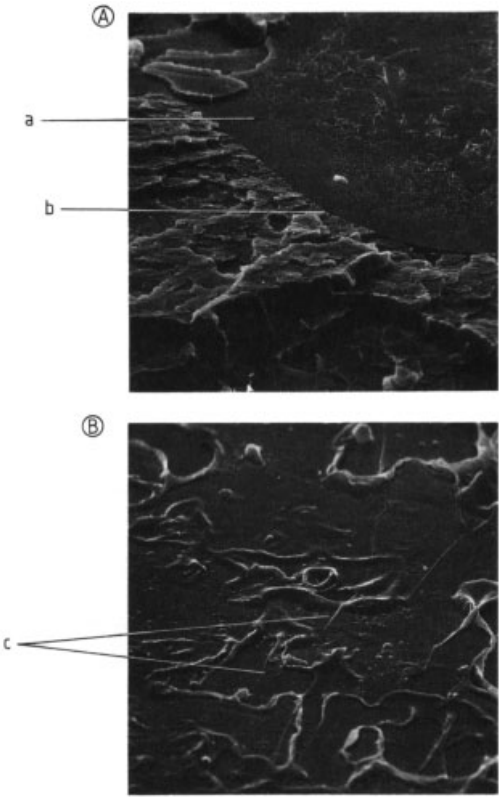

Figure 29. A) Boundary between thermal (mirror) and athermal fracture phases after impact tensile loading of polystyrene, [114] (scanning electron micrograph, ×180); B) Spickle structure in the mirror, with typical "thermal threads" (×1800)
a) Mirror; b) Thermal – athermal phase boundary; c) Thermal threads

The methods used to obtain impact values for inclusion in the table of reference values were chosen to satisfy a number of criteria, one being that the test should show a distinct difference in energy absorption between ductile fracture (in which the specimen is merely cracked) and brittle fracture.

This condition is met, for example, by the Izod notched-bar toughness test or the Charpy impact bending test; see Figure 30 [115–117]. Notched and unnotched specimens can be tested by either method. Both belong to the class of flexural impact tests in which the impact energy absorbed up to failure is measured. For reasons of measurement technique, no determination of the stress – strain relationship is included in this test. The V-shaped notches used have a 45° angle between sides and various notch root radii. Impact tensile tests are performed on plastics along with the flexural impact test, especially when the specimens in the flexural impact or notched-bar flexural impact test do not fracture. Notched and unnotched specimens are used in the impact tensile test as well.

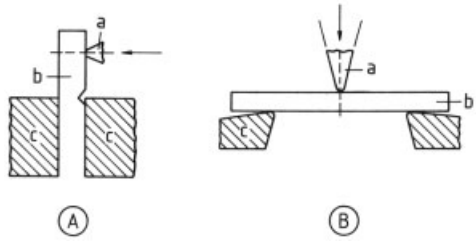

Figure 30. Setups for impact bending test
A) Izod test, notched specimen; B) Charpy test, unnotched specimen
a) Striker; b) Specimen; c) Clamp or support

The notch action at the root of the notch is described by the theoretical stress concentration factor $a_k = \sigma_{max}/\sigma_n$, where σ_n is the nominal stress with no notch and σ_{max} is the maximum stress in the direction of loading at the root. For a notched flexural impact specimen (0.25-mm radius at notch root), a_k is ca. 2.4, and for a double-notched impact tensile specimen (DIN 53 448, 0.25-mm radius at notch roots), ca. 2.9 (values determined by photoelastic examination [114]). Thus the notch action is roughly comparable in both specimens even though the notch radii are very different. Stress distribution outside the notch region, however, are quite dissimilar.

In specimens with marked partly plastic behavior, the decrease in a_k from blunter to sharper notches is not very great. This is due to the fact that despite the large stress concentration

factors, the partly plastic behavior prevents the maximum stress at the notch root reaching the high value predicted by elasticity theory [118].

Flexural impact and impact tensile tests are supplemented by a technique employing biaxial impact loading: the falling-weight impact test. A circular or square flat specimen is laid on a ring-shaped support and loaded in the center by the impact of a falling weight. Experimental setups of this kind are described in [119–121]; they are also used for investigating the impact behavior of films and finished or semifinished parts. Given a well-defined criterion of failure, the damaging energy at which a given percentage of specimens fail can be determined. Results are evaluated either statistically (probit method [122]) or by a staircase technique (up-and-down method [123]).

Methods that allow the plot of applied force versus deformation to be continued up to failure have greater predictive power than the impact and falling-weight tests described here, which yield the damaging energy or the proportion of damaged specimens.

In addition to the damaging force and deformation, these tests also measure the damaging energy for tensile loading or for biaxial loading by impact on the centrally loaded plate [124]. The test apparatus, which features automated performance and evaluation, is complicated. The devices for tensile loading are called high-speed tensile testers. Combined machines for high-speed tensile tests and impact penetration tests at high rates of deformation are also available. These devices have a maximum rated tensile force of 100 kN, with an impact velocity up to 20 m/s. The force is measured with piezoelectric pickups, and the strain is measured with inductive strain gauges or high-resolution precision laser systems [125].

Interpretation of results from the biaxial plate-loading test is difficult, chiefly because the state of orientation in the specimen varies from one plane layer to another. Apart from taking extreme orientation distributions into account, apparent anomalies in impact penetration test results can be explained only by analyses based on elasticity theory (e.g., the theory of sheets) [126].

Because of these difficulties, the results of impact tests are not generally regarded as material values.

If specimen fabrication and testing conditions are controlled very carefully, however, the tests are suitable for preliminary materials selection and especially for development in the field of impact-resistant plastics. Table 5 lists impact penetration test results for several plastics.

Table 5. Damaging work W_s per unit thickness d, damaging force F_s, and damaging deformation l_s for several plastics, after DIN 53 443 Sheet 2

Product (abbreviations from DIN 7728)	W_s/d, Nm/mm	F_s, N	l_s, mm
PC	5500	9000	23
ABS	1450	3600	16
PS	30	300	3
PE	1550	2600	20
PVC-HI	3300	5400	18
PF Type 31	30	440	3

2.1.6. Control of Mechanical Properties through Orientation, Internal Stresses, and Morphology

The term orientation refers to the alignment of structural elements (i.e., polymer chains or chain segments, crystalline regions, or domains). Orientation takes place in the initial production process or results from subsequent drawing and is frozen in on transformation to the solid state, with simultaneous occurrence of disorientation processes. A material usually exhibits greater strength and less elongation in the orientation direction. Multiaxial states of orientation in molded parts can often influence mechanical behavior in an unforeseen way. Orientation is used in the production of high-strength monofilaments (fibers and wires) and biaxially drawn (and thus biaxially oriented) films with high impact penetration strength.

Orientation is of particular importance in the fabrication of film ribbons, where drawing ratios up to 1:15 are used. The anisotropy of the strength in the longitudinal and transverse directions, resulting from the orientation, can be so great that the ribbons split when loaded transversely. This splitting tendency is desirable in string and yarn. If splitting is to be avoided, the molecules must be oriented slightly transverse to the drawing direction during processing [127].

Molecular orientation in polymers can be determined by wide-angle X-ray diffraction (lim-

ited to polymers that are more than 10 % crystalline); birefringence (the second moment of the distribution is measured); linear dichroism (suitable for determining the amorphous and crystalline orientations separately); acoustic velocity (simple method for fibers); or nuclear resonance. An approximate measure of orientation is the shrinkage at high temperature, which results mainly from the prior molecular alignments and is thus an entropy-elastic effect. In contrast, energy-elastic effects play little part in shrinkage. Figure 31 shows how orientation (characterized by birefringence) influences the notched-bar impact bending strength.

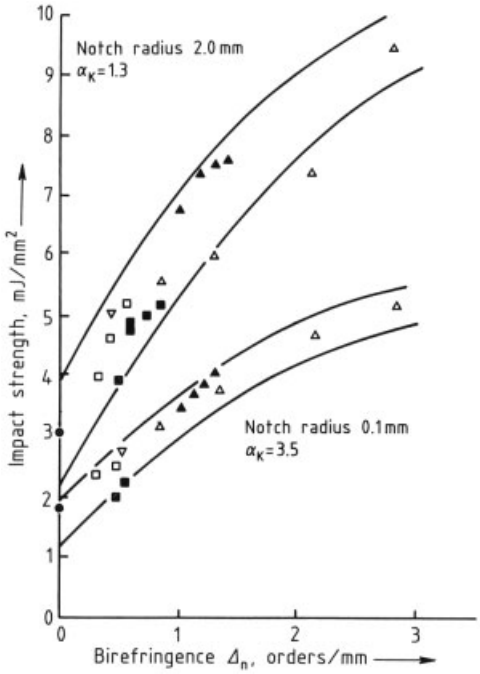

Figure 31. Notched-bar bending strength of polystyrene (specimen with double V notches) as a function of birefringence and notch geometry

Molded plastic parts can exhibit internal stresses, which may significantly affect mechanical properties. According to Wübken [128], the three main causes of internal stresses in amorphous moldings are cooling stresses, stresses due to the expansion of parts ejected from the mold under pressure, and flow-related Hooke's law stresses. Deformations resulting from internal stresses are energy elastic and therefore very small in comparison to those associated with orientation. The internal stresses in a molding are at equilibrium. If, however, the stresses are not at equilibrium (e.g., on demolding), the injection-molded part deforms until equilibrium is reached. Internal stresses in components subject to compressive loads, such as pressure piping and pressure vessels, require special attention. The addition of internal stresses to the multiaxial stresses produced by external forces can, under unfavorable conditions, lead to a reduction in the design creep strength. Internal stresses can be determined by deformation measurements on components [128]. A quasi-nondestructive test based on hardness measurements has been described [129] (see Section 2.7.2 for stress cracking).

Of the morphological states that influence mechanical properties, the only ones discussed here are crystallinity and, for amorphous plastics, particle structure. The partially crystalline state and the associated superlattice structures are discussed above (see Section 1.2). Higher crystallinity means higher rigidity and yield stress. Because crystallinity is correlated with density, measuring the density provides a simple criterion for evaluating the crystalline fraction.

Partially crystalline superlattice structures called spherulites (for their spherical shape) are formed when crystallization takes place from solution and especially from the melt. The spherulite diameter depends on the chemical structure of the macromolecules and the thermal prehistory. Polyethylene, for example, crystallizes mainly in fine spherulites having diameters of 5 – 200 µm (depending on the rate of cooling), whereas polypropylene and poly-1-butylene cooled at similar rates may form spherulites up to several millimeters in diameter.

In polymerization processes, spherulite size is often controlled by addition of nucleating agents in order to achieve desired properties, such as high rigidity or impact strength.

Rigidity (modulus, yield stress) generally increases with increasing spherulite radius, whereas toughness decreases, although exceptions can be found to this rule. Domain structures, which are especially marked in polymer blends and graft and block copolymers, also exhibit surprising engineering properties. The flow behavior, rigidity, and impact strength are improved.

Impact-resistant thermoplastics, which include, among others, grades of polystyrene and

poly(vinyl chloride) (PVC), consist essentially of a continuous hard component (polystyrene, PVC) in which a soft component [e.g., polybutadiene, poly(vinyl acetate)] is dispersed in the form of domains. The number, shape, and distribution of the domains, as well as the structure of the matrix – domain phase boundaries, govern the properties of the blend. Other domain structures of engineering interest are found in block copolymers, such as those of butadiene and styrene [130].

If only a reduction in rigidity is required, and no special demands are imposed on the energy absorption in impact loading, internal or external plasticization can be used. Internal plasticization is the copolymerization of two compatible monomers whose glass transition temperatures differ greatly.

External plasticization is a solvation process in which low molecular mass plasticizers are bound to polymer macromolecules.

2.1.7. Reinforced and Filled Plastics

A reinforced plastic is a composite material with a plastic matrix. Such a composite has an elastic modulus and strength, in at least one direction, that far exceed those of the unreinforced matrix. Particle-reinforced composites (e.g., with talc, powdered chalk, powdered minerals, short fibers, or glass spheres) are distinguished from fiber-reinforced composites (e.g., with glass, aramid, or carbon fibers). Particle-reinforced thermoplastics have the advantage that they can be injection molded. The composites exhibit greatly improved rigidity, strength, and heat deflection temperature. Moldings exhibit reduced shrinkage and thermal expansion. Curable molding compounds also include particle-reinforced composites, some of which are free-flowing materials that can be injection molded.

Fiber-reinforced materials most commonly have thermoset matrices of unsaturated polyester (UP) or epoxy (EP) resin. An important class is glass-fiber-reinforced plastics. The various reinforcements (rovings, filaments, mats) have been standardized [131]. For further information and properties, see [132].

Reinforced thermosets tested in the longitudinal direction have moduli of ca. 30 000 N/mm^2 and tensile strengths of 700 N/mm^2; thermoplastics with carbon-fiber reinforcement (50 – 60 wt % carbon fiber) exhibit tensile strengths of 270 – 280 N/mm^2. The high carbon fiber content, however, lowers the volume resistivity to 10 – 100 $\Omega \cdot$ cm.

Additives and fillers are also used to modify plastics properties. Only a few examples can be mentioned here. Carbon black lowers resistivity (affording safety against electrostatic charging) and also reduces the coefficient of friction. Graphite, molybdenum disulfide, and carbon fibers lower the dynamic coefficient of friction. Metal powders (bronze, chromium – nickel steel) are used in bearing materials, valves, piston rings, and gaskets. Aluminum flakes provide electromagnetic shielding.

2.2. Acoustic Properties and Their Testing

Because most plastics have relatively high mechanical loss factors (at least two orders of magnitude higher than most metals), they can be used to attenuate solid-borne sound. This application is undoubtedly the most interesting in the field of acoustics, especially for noise suppression. Dimensionally stable plastics with high internal damping (i.e., high mechanical loss factor) include polypropylene and polyamide. Examples of acoustic damping applications are covers, gears, and rollers made of these materials. Unfortunately, glass-fiber reinforcement greatly lowers the loss factor, so that glass-fiber-reinforced plastics are not well suited to sound attenuation. The hum generated by metal plates can be suppressed by viscoelastic coatings. These are special amorphous thermoplastics, in particular vinyl acetate copolymers, or two-component systems based on polyurethane or epoxy resins. These coatings are improved further by active fillers such as vermiculite, a micalike clay material, or graphite.

To assess the damping properties of such coatings, the dynamic flexural rigidity and the bending wave rate must be known. These quantities are derived from the dynamic modulus of elasticity, which in turn is determined in flexural vibration tests performed at audio frequencies [83], [84].

Figure 32 shows curves of loss factor versus temperature for a 1-mm steel sheet with a

2-mm sound-deadening material and also for a steel – plastic – steel laminate (i.e., a symmetrical composite). Undamped steel sheet has a loss factor $d \approx 0.0001$; the maximum with heavy edge damping is 0.001. The laminates exhibit d values up to 1.0.

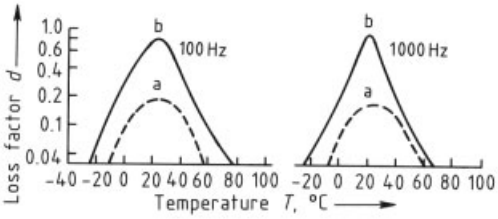

Figure 32. Mechanical loss factor d as a function of temperature T at 100 Hz and 1000 Hz for steel – plastic composite sheets
a) 1-mm steel sheet with 2-mm sound-deadening material;
b) Steel – plastic – steel composite (0.5 – 0.2 – 0.5 mm)

Homogeneous solid plastics do not attenuate airborne sound because even the lightest solids do not satisfy the requirement of wave impedance matching between air and absorbing medium. Plastics have a wave impedance several orders of magnitude greater than that of air. To suppress airborne sound, therefore, open-pored foams or fibrous materials are employed. Absorption takes place by friction between air particles and components of the skeleton. Open-pored foamed plastics used in this absorbing role include foamed urea – formaldehyde resin and nonrigid PVC foam. Foamed polystyrene slabs, even though they have closed pores, can be used for acoustic purposes if the surface is perforated.

If solid-borne sound is to be blocked (as opposed to attenuated), soft materials of low dynamic flexural rigidity are required. Low-frequency vibrations from machinery cannot, however, be blocked with plastics; instead, rubberized metal elements, steel springs, or air springs are employed.

Another application represents a far more important use of plastics in blocking solid-borne sound: the suppression of footfall noise by so-called floating floors. Acoustic requirements for buildings [133] are met by insulating slabs based on expanded-bead polystyrene foam, whose resilience is enhanced by postprocessing.

Plastics, even filled materials (e.g., with expanded glass or clay), are not very suitable for blocking airborne sound. They are not considered for this purpose by builders, because this type of soundproofing always requires heavy masses with the lowest possible flexural rigidity. Plastics with heavy fillers (PbO or $BaSO_4$) do meet the mass requirement and also have the high coincidence cutoff frequency needed for soundproofing [84]. Heavy, nonrigid plastic films are used in automobiles and machinery.

2.3. Thermal Properties and Their Testing

A knowledge of the thermodynamic properties of plastics – and, especially, of their temperature dependence – is important for the economics of production and for the quality and functioning of products. Relevant data include specific heat, specific enthalpy, thermal conductivity, thermal diffusivity, specific volume, compressibility, and coefficients of thermal volume and linear expansion. With this information, the processor and the user can calculate the heats needed for processing, mixing temperatures, volume changes on cooling, shrinkages during and after processing, and temperature-related dimensional changes in semifinished products and moldings. Furthermore, several of these quantities must be known if the reactions and internal changes taking place in plastics are to be understood.

2.3.1. Specific Heat

The specific heat of a plastic is not always an unequivocal function of temperature. Crystalline and amorphous polymers do have comparable specific heats at very low temperature, but sizable differences occur in the vicinity of transformations (e.g., glass transition, crystallization, and melting). The specific heat in these ranges depends on the state (e.g., the degree of crystallization) of the polymer and thus on its thermal prehistory [134], [135]. If the specific heat at constant pressure c_p is known, the enthalpy differences ΔH between the initial T_0 and final temperatures T can be calculated by integration:

$$\Delta H\,(T)_{T_0} = H_T - H_{T_0} = \int_{T_0}^{T} c_p\,(T)\,\mathrm{d}T.$$

Enthalpy depends on pressure, and because of the high pressures that frequently occur in plastics processing this dependence must be taken into consideration. Figure 33 A shows schematically the specific heat as a function of temperature for amorphous thermoplastics. The steplike rise in the glass transition range is characteristic. Partially crystalline thermoplastics are distinguished by the fact that an additional enthalpy, the heat of fusion, must be supplied to melt crystalline regions. The heat of fusion results in a maximum in the curve of specific heat versus temperature for partially crystalline plastics, as shown schematically in Figure 33 B. With regard to the temperature dependence of specific heat and enthalpy, see [136], [137].

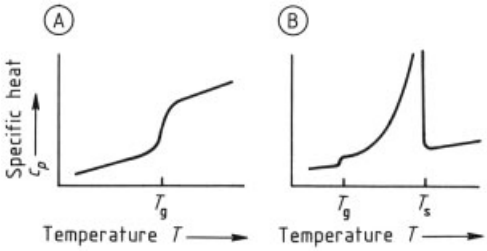

Figure 33. Temperature dependence of specific heat
A) Amorphous thermoplastics; B) Partially crystalline plastics
T_s = crystallite melting point

The enthalpy of the uncured starting product is of interest for the energy balance in the processing of thermosets.

Measurements of the specific heat of plastics from nearly 0 K to ca. 600 K can be performed only with adiabatic calorimeters. A very simple, quick, and sufficiently accurate measurement can, however, be carried out with dynamic differential calorimeters [138]; the technique is called DSC (differential scanning calorimetry). A review of differential scanning calorimeters can be found in [139].

Not only does DSC yield the same results as adiabatic calorimetry, it also makes new types of measurements possible. Examples are the detection of glass-transition steps, exothermic crystallization, melting, and determination of stabilizer content [140].

Differential scanning calorimetry also permits indirect determination of, for example, crystalline content, degree of cross-linking, degree of curing, thermal stability, and plasticizer content. Studies of reaction kinetics are possible as well [141–143]. Figure 34 shows enthalpy curves for several commercially important thermoplastics.

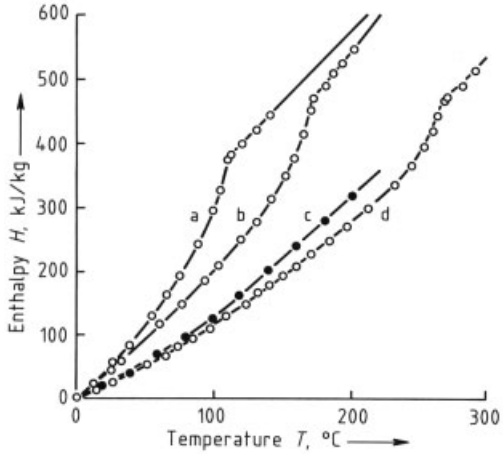

Figure 34. Enthalpy curves for some thermoplastics
a) Low-density polyethylene; b) Polypropylene; c) Polystyrene; d) Poly(ethylene terephthalate)

2.3.2. Thermal Conductivity

Plastics are poor heat conductors, because they have virtually no free electrons available for conduction mechanisms such as those in metals.

For amorphous thermoplastics at 0 – 200 °C, the thermal conductivity lies between 0.125 and 0.2 $W\,m^{-1}\,K^{-1}$. Partially crystalline thermoplastics, in contrast, have a higher thermal conductivity, because the ordered crystalline regions conduct better than the amorphous regions. As melting of the crystalline regions progresses, the thermal conductivity approaches that of the amorphous regions. Figure 35 plots the thermal conductivities of branched polyethylene, polypropylene, and polystyrene as functions of temperature and pressure [144]. Branching has a marked effect on thermal conductivity; molecular mass and molecular-mass distribution have no influence.

As cross-linking increases (thermosets, elastomers), thermal conductivity rises, because van der Waals bonds are progressively replaced by valence bonds with their greater molecular thermal conductivity. Moreover, the changes and transformations characteristic of polymers are less strongly marked in curves of thermal con-

ductivity (and other thermal values) at higher degrees of cross-linking.

At content below 5 %, additives (air or even metals) do not greatly influence thermal conductivity. For higher filler contents, numerous mixture rules have been developed. They are, however, not universally applicable. The most reliable is still the geometric mean model of Pringelhof and Throne, which was rediscovered by Agari and Uno [145] and confirmed for many mixtures. According to this model, the logarithmus of the thermal conductivity increases linearly with increasing volumetric filler content:

$$\log\lambda = V \cdot \log\lambda_f + (1-V) \cdot \log\lambda_m$$

where λ is the thermal conductivity of the mixture, λ_f the thermal conductivity of the filler, λ_m the thermal conductivity of the matrix, and V the filler content in vol %. Oriented plastics and plastics with oriented fillers, whose thermal conductivities differ from that of the matrix, exhibit anisotropic thermal conductivity.

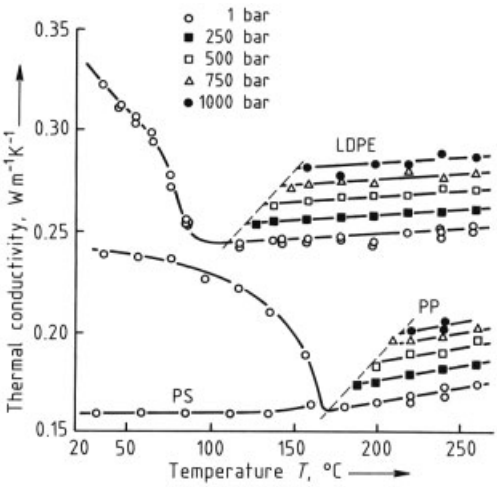

Figure 35. Thermal conductivity of low-density polyethylene, polypropylene, and polystyrene as a function of temperature [144]

For commercial closed-cell foamed plastics [polystyrene, poly(vinyl chloride), and phenolic resin], the thermal conductivity lies between approximately 0.035 and 0.042 $W\,m^{-1}\,K^{-1}$.

The thermal conductivities of solids are usually measured under steady-state conditions with the Poensgen two-plate apparatus [146]. A setup described by Fritz and Küster is suitable for the investigation of foamed materials [147]. A very simple device for determining the thermal conductivity of foamed materials with an accuracy sufficient for practical purposes is described in [148].

Two quantities derived from thermal conductivity are important in plastics processing. Cooling takes up a large part of the cycle time in injection molding, and a quantity called the thermal diffusivity a is used in calculating this process: $a = \lambda/c\varrho$, where λ is thermal conductivity, c is specific heat, and ϱ is density. Thermal diffusivity has the units m^2/s or m^2/h.

The second important quantity is the heat penetration coefficient $b = (\lambda\, c\, \varrho)^{1/2}$, expressed in $J\,m^{-2}\,s^{1/2}\,K$. This coefficient determines the initial contact temperature T_c (in Kelvin) when two bodies A and B come in contact: $T_c = (b_A T_A + b_B T_B)/(b_A + b_B)$. The formula indicates, for example, the temperature that prevails at the mold wall when a hot plastic mass flows into a cooler mold and before heat removal by the cooling system becomes effective [149].

2.3.3. Thermal Expansion

The linear coefficient of thermal expansion lies between ca. 0.6×10^{-4} and $2.3\times10^{-4}\,K^{-1}$ for thermoplastics or 0.2×10^{-4} and $0.6\times10^{-4}\,K^{-1}$ for thermosets (in the service temperature range for each case).

The linear coefficient α can be measured directly with a dilatometer (e.g., a quartz-tube dilatometer as specified in ASTM D 696) or, more accurately, against a reference material with a differential dilatometer.

This coefficient can also be derived from measurements of the specific volume or of the volume coefficient of thermal expansion β. The approximate formula is $\alpha \approx \beta/3$.

The specific volume V is also measured in dilatometers [150–153].

If the specific volume is plotted against temperature for amorphous thermoplastics, the curve exhibits a kink near the glass-transition temperature. The expansion coefficient, which is the first derivative with respect to temperature, thus has a discontinuity (a step). This phenomenon can be interpreted with Eyring and Hirai's vacancy theory of liquids: As the temperature rises, the equilibrium vacancy concentration begins to increase at the glass transition. This in-

crease manifests itself as an increase either in the free specific volume or in the slope of the specific volume–temperature curve. These effects are more marked for amorphous polymers than for partially crystalline ones, because the thermal behavior of partially crystalline materials is governed chiefly by the melting of crystallites.

At the onset of melting (i.e., as the structure begins to break down), the curve of specific volume versus temperature rises faster than linearly. After melting is complete, the curve becomes linear again but has a steeper slope than before melting; see Figure 36 [151].

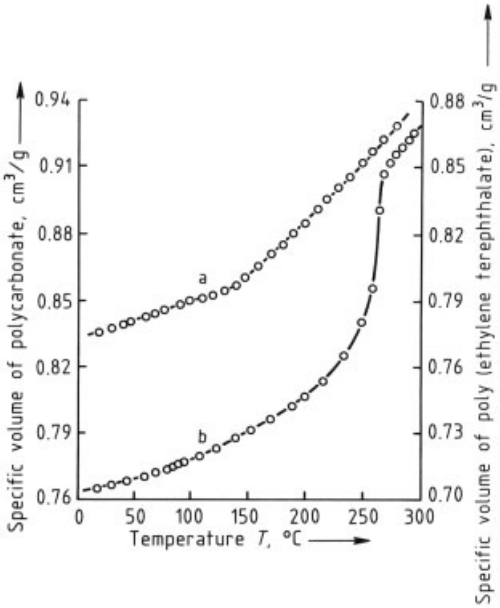

Figure 36. Specific volume as a function of temperature
a) Polycarbonate, $\varrho_{20} = 1.2\,g/cm^3$ (example of an amorphous product); b) Poly(ethylene terephthalate), $\varrho_{20} = 1.4\,g/cm^3$ (example of a partially crystalline product) [154]

On cooling from the melt, however, these discontinuities in the curves of specific volume versus temperature show a dependence on cooling rate. Figure 37 illustrates schematically what happens with amorphous materials. The vacancy concentration is somewhat lower in slow cooling than in fast cooling. The various branches correspond to states differing in vacancy concentration. Among the consequences of this behavior is that V versus T curves are measured only with increasing temperature because the heating curve cannot be reproduced on cooling.

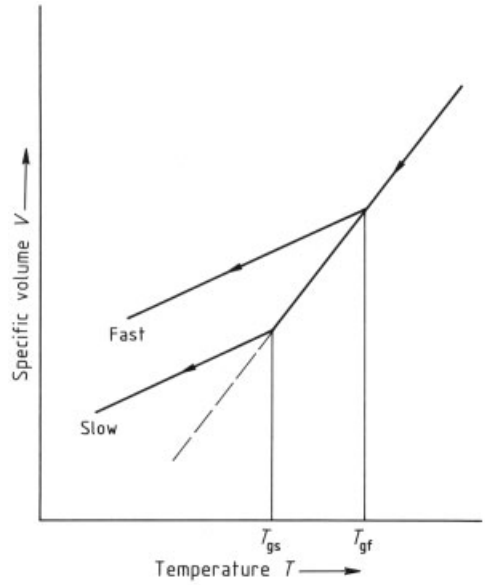

Figure 37. Temperature curves of specific volume for fast and slow cooling of an amorphous plastic
T_{gs} = glass transition temperature for slow cooling; T_{gf} = glass transition temperature for fast cooling

***p – V – T* Diagrams.** Because plastics are processed over a wide range of temperatures and pressures, the variation of the specific volume not only with temperature but also with pressure must be known. The literature describes several methods for investigating the relationship of p, V, and T [154–159].

In some methods, the specimen is placed in a cylinder and compressed from one or both ends by pistons. In another, the specimen is under hydrostatic pressure in a mercury-filled piezometer located in a pressure vessel filled with silicone oil. Figure 38 shows the $p-V-T$ diagram of an amorphous plastic (polystyrene) and of a partially crystalline plastic (high-density polyethylene). As in the V versus T plots, the shape of the curve depends on the cooling rate.

2.4. Electrical Properties and Their Testing

Plastics owe their great variety of uses in electrical and electronic devices to their combination of electrical properties, rational fabrication methods, and widely variable engineering properties.

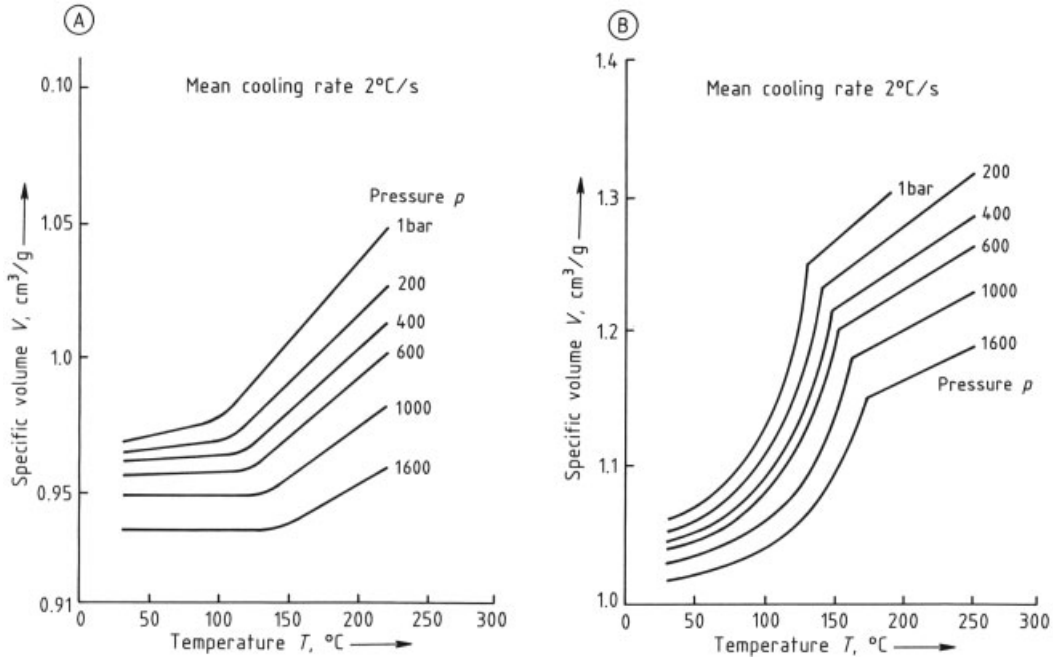

Figure 38. p–V–T diagrams
A) Polystyrene; B) Polyethylene

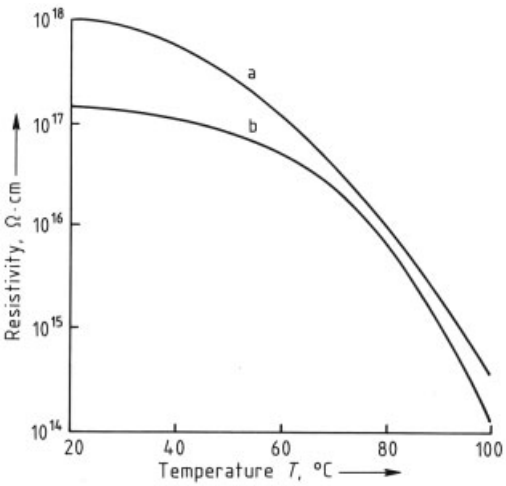

Figure 39. Resistivity of polypropylene films as a function of temperature
a) Polypropylene; b) Biaxially drawn polypropylene

2.4.1. Insulating Properties

Volume Resistivity. Plastics used as electrical insulators have volume resistivities of 10^{10}–$10^{19}\,\Omega\,\mathrm{cm}$. Substantially lower values can be achieved with additives such as carbon black, graphite, carbon fibers, and metal powders or flakes, making possible a reduction in static charging due to high resistance, which can lead to safety problems in some instances (e.g., when flammable liquids are transferred from plastic vessels, when plastic flooring is employed in spaces where an explosion hazard exists, and in mining). In general, no hazardous charging occurs at resistivities of less than $10^{10}\,\Omega\,\mathrm{cm}$.

Charge transport in plastics occurs mainly by ions and to only a minor extent by electrons. Whereas Ohm's law holds rigorously for metals, insulators not only have a charge carrier current but also an internal charge current resulting from polarization of the plastic. This internal charge current builds up a space charge, which decays on a scale of hours. The volume resistivity ϱ_D is a function of temperature, having the form $\varrho_D = \varrho_0 \exp(b/T)$, where ϱ_0 and b are constants and T is absolute temperature. Figure 39 shows how the volume resistivity of polypropylene films depends on temperature.

Surface Resistivity and Resistivity between Plugs. Because the insulation surface usually makes a large contribution to leakage, sur-

face resistivity is also defined. This quantity does not, however, have the predictive power of volume resistivity, which is a material-specific value.

In the case of inhomogeneous insulators such as laminated composites or plastics containing electrically conducting fillers such as carbon black, the insulator resistance is often determined between two metal pins inserted into the material 25 mm apart (resistivity between plugs). This measurement includes an unknown contribution from surface resistivity.

Resistivity measurements are current–voltage measurements, preferably carried out at 100 or 1000 V d.c. with standard electrodes [158].

2.4.2. Dielectric Properties

Certain formal analogies exist between the relaxation behavior of viscoelastic bodies under mechanical loading, which can be described in terms of a complex modulus, and the dielectric behavior of plastics, which is characterized by a complex dielectric constant (permittivity) $\varepsilon* = \varepsilon'(\omega) - i\,\varepsilon''(\omega)$, where ω is angular frequency, $\varepsilon'(\omega)$ the dielectric constant (permittivity), and $\varepsilon''(\omega)$ the dielectric loss factor. The ratio $\varepsilon''/\varepsilon'$ is equal to the tangent of the phase angle between the voltage and the current: $\tan\delta = \varepsilon''/\varepsilon'$.

Just as viscoelastic behavior in the mechanical case can be simulated by spring–damper models, dielectric behavior can be described in terms of electrical networks. The analogy is, as already mentioned, merely formal; it has to do with the fact that both phenomena come under a common theory, the theory of the elastic aftereffect (memory effect), advanced as early as 1893 by Wiechert and extended by Wagner in 1913 to the theory of the dielectric aftereffect.

The mathematics of viscoelastic and dielectric behavior is the same as that of the magnetic aftereffect, linear network theory, and linear integral equations. Figure 40 illustrates schematically how the dielectric values for high-polymer electrical insulators depend on the frequency of the applied alternating field. Here ε' is the relative permittivity ε'_r, which is a measure of the polarization attainable at any frequency: At low frequencies, the molecules have enough time to orient themselves in the field, so that ε' takes on its largest value. At higher frequencies, the field changes so rapidly that the dipoles can no longer follow and ε' declines to smaller values. The value of ε' falls off substantially when the period of the field is of the order of the dipole relaxation time. The value ε'' is a measure of the energy transformed in one oscillation period; this energy is extracted from the field and converted to heat as the dipoles orient themselves. This quantity passes through a maximum when the period of the field is equal to the relaxation time; it approaches zero at very low and very high frequencies.

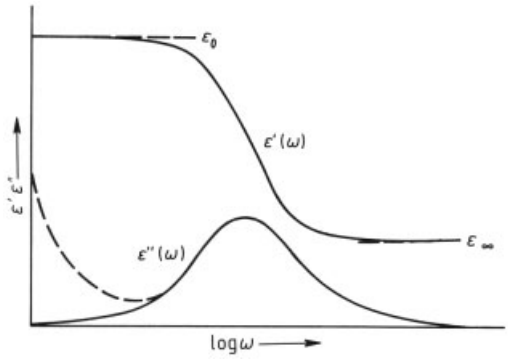

Figure 40. Schematic curves of ε'_r and $\varepsilon''_r = \varepsilon'_r \tan\delta$ versus frequency for high-polymer electrical insulators

Because $\varepsilon''_r/\varepsilon'_r = \tan\delta$, $\varepsilon''_r = \varepsilon'_r \tan\delta$. The product $\varepsilon'_r \tan\delta$ is also called the dielectric loss factor.

The measured ε''_r at low frequencies can be distorted by a metrological effect not related to polarization: the d.c. conductance, whose contribution to the total loss decreases with frequency (dashed curve in Fig. 40).

If plastics are to be heated in a high-frequency (HF) field, the loss factor must be made as large as possible (HF welding of polar plastics). In insulation, small $\tan\delta$ values are desirable to minimize the evolution of heat. When polar plastics (e.g., PVC blends) are used in cable manufacturing, the loss factor maximum must be shifted to the highest possible temperature so that thermal breakdown will not be promoted at service voltages.

In nonpolar substances, an external a.c. field brings about polarization only by deforming the electron distribution. This displacement of the charge carriers is, however, independent of frequency, so that ε_r remains virtually constant over

a wide frequency range. At very high frequencies (in the infrared), ε_r approaches the square of the optical refractive index.

The Clausius – Mosotti equation holds for nonpolar substances under certain conditions. This equation

$$\frac{\varepsilon-1}{\varepsilon+2}\frac{M_r}{\varrho}=\frac{4\pi}{3}N\alpha$$

describes the relationship between the dielectric constant and the density, where ε is the dielectric constant, N is Avogadro's number, M_r is molecular mass, ϱ is density, and α is molecular polarizability.

The temperature curve of $\tan\delta$ for nonpolar straight- chain polymers generally shows two maxima of the loss factor: the α absorption (e.g., due to motion of the CH_2 dipoles in polyethylene) and the γ absorption (e.g., due to relaxation in crystalline regions in PE). Branching gives rise to a third absorption region, β absorption, which lies between the α and γ absorptions in the $\tan\delta$ curves; see Figures 41 A and B.

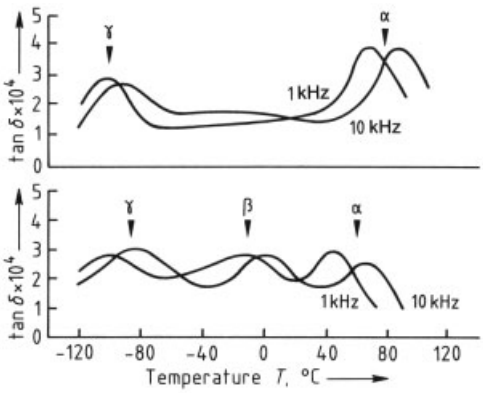

Figure 41. Plot of $\tan\delta$ versus temperature at 1 and 10 kHz for polyethylene
A) High-density polyethylene; B) Low-density (branched) polyethylene

Often, the loss factor and dielectric constant must be measured over a wide range of temperature and frequency (from a few hertz to more than 10 GHz, although most measurements are made between 50 and 100 MHz). The dielectric constant ranges from somewhat over 1 for foamed plastics up to 20 or more. Loss factors range from 3×10^{-5} for polyethylene with no residual catalyst to 0.1 or higher.

Materials with anisotropic structure may display anisotropic dielectric properties. Measurement is complicated by the variety of methods needed to cover the wide frequency range. Up to about 10^8 Hz, however, only two measuring principles are needed. In general, the circuit used must contain coils, capacitors, and resistors. From the audio frequency range up to 100 kHz, bridge circuits are used without exception. In a.c. measurements, at least one branch of the bridge, not containing the specimen, must have a blind component, either an inductance or a capacitance. Examples of such bridges are the Schering type and the transformer measuring bridge. Automatic, self-balancing transformer bridges, which are unexcelled for quick routine measurements, are now available.

The range from 10 kHz to 100 MHz corresponds roughly to radio frequencies. Measurements in this range are performed by tuned-circuit methods (resonance techniques with tank circuits and a modified capacitance standard).

Above 10^8 Hz, different principles are employed. They include measuring-line methods and techniques with capacitively loaded coaxial resonators.

For further information on measuring methods see [159–162]. Table 6 lists dielectric values at 50 Hz and 20 °C for a number of plastics.

Table 6. Dielectric properties of some plastics at 20 °C and 50 Hz (mean values)

Product (abbreviations from DIN 7728)	Dielectric constant ε	Loss factor $\tan\delta$
PF 31.5	6 – 8	0.4
UP resins	3 – 3.8	0.003 – 0.01
EP resins	3.2 – 4.8	0.002 – 0.006
PE	2.28 – 2.34	0.0005
PP	2.27	0.0005
PVC, rigid (E)	3.8 – 4.3	0.03
PVC, rigid (S)	3.4 – 3.7	0.02
PVC/DOP 60:40	7.5 – 8.0	0.08
PA 6	4.3	0.03
PS	2.5	0.0003
PTFE	2.1	0.0001

2.4.3. Dielectric Strength

Current theories of breakdown in a solid dielectric start with the assumption that breakdown is linked with the free electrons that are always present, which are set in motion by the electric field. Part of the energy obtained from the field is given up to the dielectric, and this process is decisive for the dielectric: If the field is strong

enough, the energy absorbed from the field by the electrons exceeds the energy lost to the dielectric, and the electron energy increases without limit. This energy is transformed to heat and ultimately destroys the dielectric.

The theory of collisional ionization assumes that the electrons, after exceeding the maximum loss to the dielectric, are accelerated further until they have sufficient energy to liberate more electrons by collisional ionization. The result is an avalanche-like increase in the number of electrons.

In the theory of collective breakdown, in contrast, the energy excess is assumed to increase the thermal energy of the collective of free electrons. The resulting temperature rise is then transferred to the dielectric [163].

The breakdown voltage is measured with a 100- or 200-kV a.c., 50-Hz power supply, usually on plate or film specimens. The breakdown voltage per unit specimen thickness is the dielectric strength E_d, in kV/mm. This value is not, however, a specific material property; it depends on a number of factors, especially specimen thickness [164], [165].

A quantity of special engineering interest is the long-term dielectric strength – the breakdown strength as a function of the time over which the voltage is applied, often under service conditions [166]. The service life of an insulator is influenced by a variety of aging processes, including the buildup of space charge due to electron enrichment, electrical aging due to microcorona discharges, electrochemical aging due to mobile contaminants (e.g., water), and thermal aging [167].

Tracking Resistance. If contaminants and moisture on the surface of an insulating material form a conducting bridge between two contacts between which a voltage is imposed, surface leakage currents can damage the surface. This process is simulated in testing to assess tracking. Two standard methods are available, one with voltages up to 600 V (low-voltage tracking) [168] and the other up to 6 kV (high-voltage tracking) [169].

In the low-voltage method, contamination is simulated by a test solution dripped onto the surface. At a predetermined voltage selected from a series, five specimens must survive after 50 drops; failure is signaled by the presence of a given current. Variants of the test are described in [168].

The high-voltage method is intended for insulating materials for open-air service at high voltages. The specimen is tilted 45° from the horizontal, and a stream of a test solution is allowed to flow down the bottom of the specimen between two stainless steel electrodes. Two failure criteria can be applied: the current flowing over the specimen surface or the presence of tracks [169].

Other, application-oriented tests on insulating materials should be mentioned briefly.

The ability of insulating materials to withstand surface glow discharges is important for the service life of electrical equipment. The glow discharge is intimately related to breakdown behavior. Such discharges attack the surface both mechanically and chemically. In the test of this behavior, graduated voltages are imposed on a specimen with a point – plate electrode arrangement, and the time to breakdown is measured. The service life curve obtained in this way for the insulating material makes it possible to rank insulators with respect to glow discharge behavior [170], [171].

If insulators are in contact with conductors to which d.c. voltage is applied, corrosion of the conductors can result. This corrosion depends on the insulator – metal combination and is affected by temperature, atmospheric humidity, and exposure time. Electrolytic corrosion can take place on any base metal. This phenomenon is tested by imposing a d.c. voltage in a humid atmosphere; two brass electrodes are employed. Results are assessed by comparing the colors of the brass foils with a standard color scale [172], [173].

Plastics filled with conductive carbon black, aluminum flakes, stainless steel fibers, or metallized carbon fibers are suitable for shielding electromagnetic fields. These compounds, used for the housings of electronic devices, are supposed to prevent electromagnetic interaction of the devices.

Measurements of dielectric strength and surface resistance are not adequate for the evaluation of shielding qualities, and special methods must be used. Shielding effectiveness is measured in dB. A value of 30 – 40 dB is considered sufficient for plastics. For details, see [174], [175], and literature cited therein.

2.4.4. Electrets; Semiconducting and Conducting Polymers

Electrets are dielectrics bearing a quasi-permanent charge, in which positive and negative charges, either isolated or created by polarization, are kept apart for an extended time. The best-known electrets are fluorinated ethylene–propylene copolymers (EP-F), polycarbonate (PC), polytetrafluoroethylene (PTFE), polypropylene (PP), poly(ethylene terephthalate) (PETP), and polyphenylene ether (PPE).

Electrets are used in electroacoustic transducers (e.g., microphones), radiation dosimeters, and air filters (electrostatic dust collection).

Several polymers now under development exhibit piezoelectric and pyroelectric effects. The most important product in this area is poly(vinylidene fluoride) (PVDF). To achieve the strongest possible piezoelectric effect, the material is drawn and given a permanent dipole orientation in a strong electric field.

The true semiconductive and conductive polymers include polyacetylene, polypyrrole, poly(*p*-phenylene), and poly(*p*-phenylene sulfide). Detailed reviews of this area have been published by Roth [176] and Mair and Roth [177]. Numerous literature references can be found in [178]. See also → Polymers, Electrically Conducting.

2.5. Optical Properties and Their Testing

Optical elements made of plastics include lenses, prisms, rear reflectors, glazings, lighting enclosures, optical-fiber light guides, compact discs, and many others. The most important optical properties for these applications are refractive index and lighting qualities (as defined, for example, in lighting standards), especially transmittance. In many other components for household and industry, aesthetic qualities such as color, gloss, and opacity are more important, but these properties are difficult to correlate with visual impressions.

2.5.1. Refraction and Birefringence

When a ray of light passes from one isotropic medium to another, the sine of the angle of incidence is in a constant ratio to the sine of the angle of refraction (both measured from the normal to the interface). If the light passes from a vacuum (or, less precisely, from air) into a denser medium, this ratio is called the refractive index n. The refractive index is dimensionless and depends on the wavelength of the light and on the temperature.

General methods for measuring refractive indices (i.e., not necessarily restricted to plastics) can be found in optics texts; standards also exist for measurements on plastics [179], [180]. Table 7 lists refractive indices for a number of plastics. Wavelength dependence is characterized by the dispersion. In practical optics, either the "dispersive power" $D_r = (n_F - n_C)/(n_D - 1)$ or its reciprocal, the Abbé number $\nu = 1/D_r$, is employed. Subscripts refer to wavelengths at which the indices are measured.

Table 7. Refractive indices n_D^{20} of plastics

Product (abbreviations from DIN 7728)	Refractive index n_D^{20}
CA	1.49 – 1.51
CP	1.47 – 1.48
CAB	1.45 – 1.47
PTFE	1.35
FEP	1.338
PFA	1.35
ETFE	1.4028
POM	1.48
PA 6	1.53
PC	1.586
PE-(LDPE)	1.51
PE-(HDPE)	1.545
PMMA	1.4893 – 1.491
PP	1.5030
PAN	1.5187
PS	1.590 – 1.592
SAN	1.565 – 1.569
PVDC	1.60 – 1.63
PSU	1.633
PVC	1.52 – 1.55
PU	1.5 – 1.6
PVA	1.4665
PVDF	1.42
SB (30 % PS)	1.53
EP	1.55 – 1.60
UP (50 % styrene)	1.523 – 1.540
PVC+DOP 60/40	1.52
PF	1.7
UF	1.54 – 1.56

Birefringence is a measure of optical anisotropy. It is defined as the maximum difference between two refractive indices determined in two perpendicular directions. Some materials show isotropic behavior until they are de-

formed while in the glassy state. The deformation, which manifests itself in small changes in bond angles and interatomic distances, leads to the phenomenon *stress birefringence,* which is proportional to the applied stress. This behavior is utilized for, among other purposes, photoelastic examination, which enables the investigation of plane and even triaxial states of stress in metal structures by using plastic models. The preferred material for this application is phenol – formaldehyde resin because it has a relatively high stress birefringence.

The other cause of birefringence in plastics is orientation of the polymer chains brought about in processing or by drawing. This *orientation birefringence* is greater than the stress birefringence. Birefringence measurement is one of the techniques used to determine the state of orientation in polymers.

Birefringence is measured by the transmission method (Fig. 42) [181]. A ray of monochromatic light is polarized so that the oscillation direction of the electric vector lies in the $x-z$ plane and makes a 45° angle with the z direction. The ray next passes through the specimen, in which the polymer molecules should be aligned preferably with the z axis, and then through an analyzer whose polarization direction P_z is perpendicular to that of the polarizer. The intensity of light transmitted by the analyzer is measured with a photomultiplier. If the polarization in the specimen does not change, the transmitted light has zero intensity; if the specimen is birefringent, the intensity will be

$$I_b = I_0 \sin^2 \left[\pi \frac{d}{\lambda} (n_x - n_z) \right]$$

where I_0 is the incident light intensity. From the intensity ratio I_t/I_0, the specimen thickness along the path of the light, and the wavelength of light employed, the birefringence $n_x - n_z$ can be calculated as

$$n_x - n_z = \frac{\lambda}{\pi \cdot d} \arcsin \sqrt{I/I_0}$$

This method is used when the degree of drawing and thus the degree of orientation are high.

A second method for measuring birefringence uses the Babinet compensator, which replaces the analyzer and the photomultiplier. Essentially, it consists of two quartz wedges that are adjusted so as to offset the phase shift δ between the x and z components of the emerging light vector.

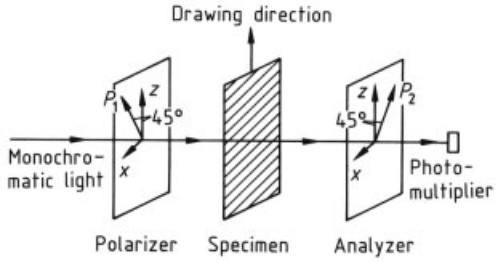

Figure 42. Setup for birefringence measurement [181]

2.5.2. Transmittance and Gloss

The light-transmitting qualities of plastics are described in terms of the visual impression that the specimen gives under precisely defined illumination. In the usual order, these terms are clear, almost clear, transparent, translucent, semiopaque, and opaque.

The customary definitions in lighting technology are as follows [182]:

Transmittance is the percentage of light transmitted without deflection. Without deflection means that only the components propagated in the direction of the incident ray are considered. Plastics with particularly high transmittances include poly(methyl methacrylate) (92 % at 380 – 780 nm), polycarbonate (88 %), and cellulose esters (86 % at 420 nm, 90 % at 660 nm for cellulose acetate). For some applications, the ultraviolet transmittance of these clear products can be reduced by addition of UV absorbers. Modified and filled polymers are clear or transparent only if their components have similar refractive indices or if the particle size of the incoherent phase is smaller than the wavelength of the light.

Haze refers to the portion (intensity) of light transmitted by a specimen that deviates from the direction of the incident light because of forward scattering in the material. Materials with high haze components are said to be semiopaque or translucent.

Both transmittance τ and haze vary with the geometry of the optical measuring system, which must therefore be precisely defined. The

same holds for the type of light employed. Examples are Colorimetric Standard Illuminants A (incandescent lamp), C (same as A with Davis – Gibson filter), and D 65 (daylight with a color temperature of ca. 6500 K).

Once the spectral characteristics of the light and the spectral sensitivity of the human eye (characterized by "standard spectral curves") are known, mathematical formulas are applied to relate the wavelength dependence to the measured energy spectrum of the transmitted light. Finally, transmittance is determined.

The most common way of determining absorption spectra is by use of spectrophotometers. If the monochromator is placed in front of the specimen, the light, whose wavelength is continually varied, passes through the specimen, and the transmitted intensity of each wavelength is measured successively by a photocell. Alternatively, the light is dispersed only after it has passed through the specimen, and the intensities of the individual wavelength components are then led to the photocell. With this scanning arrangement, however, the long measuring time can be shortened considerably if all wavelength components are led simultaneously to a series of photocells. The result (spectral curve) is then determined in a few microseconds. Some applications of transmittance measurements in the visible spectrum include determination of the spectral or mean transmittances of colored and colorless plastics; color characterization (tristimulus values or total transmittance) for signal lights in air and marine navigation, railroads, and road transport; color matching; aging tests; damping measurements in light guides; and testing the lightfastness and weather resistance of colored plastics.

The haze of thin specimens with low scattering is measured in a standardized setup [183]. For reasons having to do with experimental technique and to bring the results into accord with subjectively assessed scattering (haze) properties, the scattered light measured is the component lying between two coaxial cones (about the axis of incidence) with opening angles of 12° and 80°.

For specimens with higher scattering (up to 30 %), the measuring system of ASTM 1003-61 is employed. The setup features an integrating sphere and can be reconfigured to measure the incident flux as well as the directly transmitted, scattered, or total transmitted flux. The parameters total transmittance, diffuse transmittance, and haze are then calculated with allowance for the scattering of the integrating sphere. High-quality transparent acrylates have haze values of the order of 0.15 %.

Gloss. Hunter defines gloss as the ratio of the regularly (specularly) reflected light intensity to that of a high-quality optical mirror. Gloss can be defined in a number of other ways. Most glossmeters measure the specular gloss defined by Hunter. Most instruments employ fixed incidence and measurement angles (e.g., 45° in the Lange glossmeter).

Highly polished, optically flat black glass plates are often used as reference standards. Their reflectivity can be calculated from the refractive index.

Measurements of the spectral reflectance curve (i.e., the dependence of reflectivity on incidence angle) give a more meaningful description of gloss. The instrument used for these measurements is a goniophotometer [184–187].

The values obtained by gloss measurement techniques, however, do not correlate perfectly with the visual impression of gloss. This becomes obvious when an inspector assessing gloss turns an object one way and another, thus scanning through many incidence and reflection angles. Polarization also plays a role in the impression of gloss.

2.5.3. Color and Colorimetry

Color is not an absolute property but depends on the surface condition of the specimen as well as the conditions under which it is examined. Nor does the term "artificial light" or "daylight" provide enough information; for example, the color of sunlight varies with the position of the sun in the sky. Two colors that the observer subjectively judges to be the same under given illumination may not match under other conditions. These problems are, of course, not specific to the plastics industry. The concepts of colorimetry, measuring methods, and test conditions are specified in DIN 5033 [188]. American and British standards also exist for this field [189], [190]. The CIE international vocabulary for lighting technology defines terms and nomenclature; prob-

lems involved in color differentiation are discussed in ISO 2573 [191], [192].

Very simple methods for assessing small color changes or departures from colorlessness in the direction of yellow (yellowing) are described in [191]. The objective determination of color with scales and formulas is a difficult undertaking but is still important in plastics production and applications.

Transparent, translucent, and opaque colored plastics must generally exhibit narrowly defined hues, whether they are to be used in signaling, advertising, or illumination. Consumer products and industrial articles must meet similar requirements, especially when colored plastics have to match other components in perceived color.

Problems of colorimetry can be mentioned only briefly here. The human eye interprets incident radiation with three distinct sensitivity functions at once. These are called the standard spectral curves and relate to the three primary colors red, green, and blue. One of these (green) is taken as the basis for the spectral curve describing the human perception of brightness. The standard spectral curves are now mathematically related to the energy distribution of the radiation giving rise to a color sensation and to the reflectance curve of the specimen in the visible spectrum (see also Section 2.5.2). The relations between the incident energy distribution (e.g., the Colorimetric Standard Illuminants) and the standard spectral curves are tabulated, so all that is necessary is to link them with the reflectances. The resulting tristimulus values X, Y, and Z are normalized and can be used to generate a two-dimensional representation. The normalized values x, y, and z satisfy the condition $x+y+z=1$; thus the color location in the $x-y$ system, which formally characterizes the color, can be unambiguously determined. The colors of the spectrum, plotted in this way, lie on a horseshoe-shaped curve (Fig. 43). Together with the brightness reference value Y (which cannot, however, be presented in the color triangle), the point (x, y) completely describes a color. All natural colors lie inside the curve. The location of the "white point" depends on the type of illumination. The saturation decreases toward this point. Colors lying on a straight line radiating from the white point have the same hue. In Figure 37, for example, the square and the triangle are color locations with the same hue, and the square is more saturated than the triangle. It is important here that colors with the same color location are the same only for the illumination conditions illustrated, not necessarily under other illumination conditions. Such colors are said to be conditionally the same, or metameric colors. Colors that must appear the same under all illuminations must have identical reflectance curves.

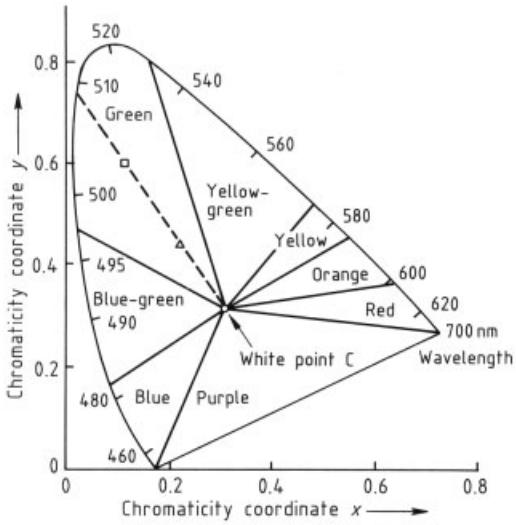

Figure 43. The color triangle

In principle, any spectrophotometer that permits measurements in the visible spectrum and has a reflector attachment is suitable for colorimetry.

An important class of instruments, especially for routine color measurement, is the three-filter (tristimulus) type, which does not yield complete reflectance curves but facilitates the determination of X, Y, and Z values. These devices use three calibrated filters (tristimulus filters) to evaluate the fraction of incident light reflected from a specimen. The filters, in combination with the light source and the system of photoelectric cells, give the three standard spectral curves. Because the curves are not reproduced perfectly, the measurements do not rank as absolute values. However, these instruments are suitable for relative measurements (e.g., comparing the color of a plastic product with that of a color standard). However, only metameric color matching can be realized. As mentioned above, the geometry of the setup is also important for reflectance measurement. Geometries

for colorimetry include "direct" illumination at 45° with observation at 0° and "diffuse" illumination at 0° with the integrating sphere used for reflected-light measurement. The appropriate geometry depends on the surface structure of the specimen. The most commonly used Colorimetric Standard Illuminant is D 65, which resembles the formerly customary type C, because the behavior of nonluminous colors in daylight is of particular interest. For color-matching purposes, fluorescent lamps or, even better, xenon arc lamps are used to provide artificial daylight.

The locations (in the color triangle) of standard and specimen must lie within a certain tolerance if they are to appear "the same." A number of workers have attempted to establish these tolerance ranges. Formulations to match standard colors can be found by computer calculations based on colorimetric measurements (Kubelka – Munk and other methods).

2.6. Density and Physical-Chemical Properties and Their Testing

2.6.1. Density

High-precision density measurements can be performed with relatively simple apparatus. Because some physical properties of plastics are single-valued functions of density, changes in these properties can be characterized indirectly on the basis of density measurements. An example of such a property is crystallinity, which in turn affects certain mechanical properties.

Density often occupies an important place in specifications; in the case of polyethylene, for example, it is essential for classification and identification [192].

The determination of density for solid plastics (except foamed plastics) is described in ISO 1183 [193]. The standard covers four methods, corresponding to the type of plastic in question. These methods are also contained in DIN 53 479 [194]; DIN 53 420 describes density measurements on foamed plastics. Compact, unfilled, unreinforced plastics have densities between ca. 900 kg/m^3 (polypropylene) and 2300 kg/m^3 (polytetrafluoroethylene).

2.6.2. Liquid Absorption and Swelling; Desorption of Plasticizers

Many polymers can absorb liquids (often up to an equilibrium concentration). The processes involved are adsorption, diffusion, and dissolution or swelling. Un-cross-linked amorphous polymers swell without limit in some liquid media and ultimately dissolve entirely (e.g., polystyrene in benzene). A table identifying good and poor solvents for some 60 important polymers appears in [195].

If the polymer is cross-linked or partially crystalline (with the crystallites acting like cross-link sites), swelling may be limited because the entropy-elastic restraint of the network prevents further liquid absorption.

Thermodynamic equilibrium is reached when the free swelling energy $\Delta F = \Delta H - T\,\Delta S$ vanishes or when the heat of swelling ΔH is equal to the product of the temperature T and the entropy of swelling ΔS. The time to equilibrium depends not only on the swelling agent and its concentration, but also on the geometry of the sample and, above all, the diffusion coefficient of the swelling agent in the polymer. Figure 44 shows the equilibrium water uptake of several plastics as a function of relative humidity.

The opposite process, liquid loss, occurs when low molecular mass plasticizers diffuse out of plastics and become desorbed.

2.6.3. Permeation of Gases and Vapors

The permeation of gases and vapors in plastics can be described as a two-stage process. First, the permeating substance dissolves in the surface of the plastic; second, the molecules diffuse through the material, forming a concentration gradient, and then evaporate again. The concentration c of the permeating gas at the surface depends on the gas pressure p or the volumetric concentration in the gas space. For permanent gases (H_2, O_2, N_2, CO_2, noble gases, etc.), the concentration is directly proportional to the partial pressure p (Henry's law: $c = kp$). This simple relationship does not hold for the diffusion of vapors.

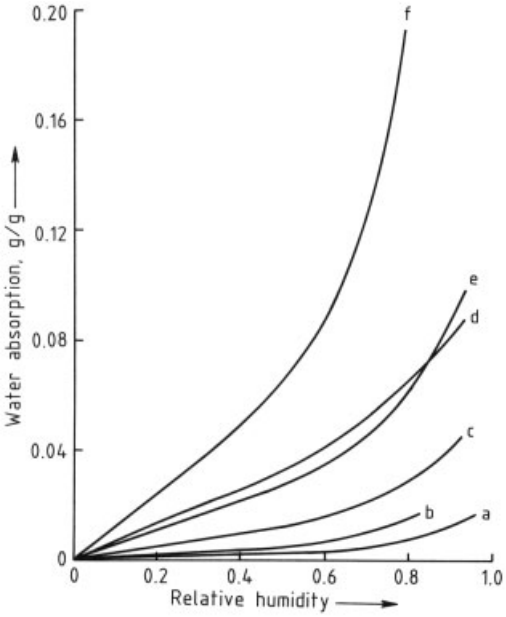

Figure 44. Equilibrium water absorption of some plastics as a function of relative humidity
a) Poly(vinylidene chloride); b) Chlorinated rubber; c) Poly(vinyl butyral); d) Polyamide 66; e) Cellulose acetate; f) Poly(vinyl alcohol)

The steady-state diffusion of a gas or vapor is described by Fick's first law:

$$\frac{\mathrm{d}Q}{\mathrm{d}t}\frac{1}{A} = -D\frac{\mathrm{d}c}{\mathrm{d}x}$$

where $\mathrm{d}Q/\mathrm{d}t$ is the flow velocity of the medium, A the flow area, D the diffusion constant, and $\mathrm{d}c/\mathrm{d}x$ the concentration gradient. Transient diffusion is described by Fick's second law.

Consider a layer of finite thickness d and assume that D is independent of concentration. Integrating the above equation over the thickness gives

$$\frac{\mathrm{d}Q}{\mathrm{d}t}\cdot\frac{1}{A} = \frac{D\,(c_1 - c_2)}{d}$$

where c_1 and c_2 are the concentrations at the two surfaces. If concentration is expressed in terms of pressure, the result is

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = D\cdot S\cdot A\frac{p_1 - p_2}{d}$$

where $D \cdot S$ is the permeability P; $\mathrm{d}Q/\mathrm{d}t$ is constant in the steady state. In experiments, $p_1 - p_2 = \Delta p$ is held constant, so that

$$P = \frac{Qd}{\Delta t A \Delta p}$$

in units of $\mathrm{cm^3\,cm^{-1}\,s^{-}\,mbar^{-1}}$.

The practical quantity gas permeability is easier to work with. It is the volume of gas in cubic centimeters that permeates a 1-$\mathrm{m^2}$ area of the product being tested in 24 h at a given temperature and a given pressure difference (e.g., 1 bar). The corresponding unit is $\mathrm{cm^3\,m^{-2}\,d^{-1}\,bar^{-1}}$. The unit employed for vapor permeability, especially water-vapor permeability, is $\mathrm{g\,m^{-2}\,d^{-1}}$. The thickness of the test specimen must be stated, along with the gas and water-vapor permeabilities, because the units do not allow for reduction to unit thickness. Furthermore, the pressure difference must be known in the case of water-vapor permeability; in practice, this is accomplished by establishing a given atmospheric humidity difference.

The steady-state transport equation holds only for gases and vapors that are not very soluble in the polymer and do not react chemically with it. Vapors can often act as plasticizers, thus increasing diffusion.

The permeability is temperature dependent in all permeant – polymer systems. The diffusion coefficient for both gases and vapors increases with increasing temperature. With increasing temperature, the solubility factor S increases for gases but decreases for vapors.

Permeation is also influenced by the molecular size and polarity of the permeating substance. The morphology of the polymer (crystallinity, branching, orientation, etc.) plays an important role in the process.

Multilayer films (laminates) are commonly used in packaging. The P coefficient of such a composite can be calculated from the P values of the individual plies:

$$\frac{1}{P_\mathrm{L}} = \frac{1}{L}\sum_{i=1}^{n}\frac{d_i}{P_i}$$

where L is the thickness of the composite, P_L is its permeability, and d_i and P_i are the thickness and the permeability of the ith ply.

This equation, however, applies only when permeability does not depend on pressure. The order of the plies has no effect in this case; otherwise, the overall permeability depends on the order of the plies.

Methods of measuring gas permeability include those employing pressure gauges and those involving carrier gases. In the first group, the quantity of gas permeating the specimen in a given time is found by measuring pressure and volume changes [196], [197]. In carrier-gas methods, gas flows through the specimen while the pressure remains the same on both sides. The test gas flows at a constant rate through a chamber on one side of the specimen, while a second (carrier) gas flows, also at a constant rate, through the chamber on the other side. The carrier gas delivers the permeating test gas to a detector, which can operate on the absorption or thermal conductivity principle [198]. A development in the carrier-gas technique was the insertion of a gas chromatograph column between the measuring cell and the detector [199].

To determine water-vapor permeability, an aluminum test cell filled with calcium chloride, for example, is employed as absorption chamber. The cell is closed by the specimen and sealed with wax at the edges. This apparatus is placed in an atmosphere having a well-defined temperature and humidity, and the weight gain of the absorption cell is measured at regular intervals [200], [201].

2.7. Stability Properties and Their Testing

The usefulness of plastics in many fields depends on their resistance to environmental factors. These complicated effects cannot always be simulated, and testing must often be accelerated. The loads applied are thus greater than would occur in the normal loading cycle. Accordingly, for short-term tests in particular, the correlation with in-service performance must be examined carefully and interpreted critically.

2.7.1. Water Absorption

Like other swelling agents (such as alcohols), water has a plasticizing effect on some plastics (polyamides, polyurethane). This action is based on the cleavage of hydrogen bonds between polar groups of adjacent macromolecules [202]. The result is a decrease in the glass-transition temperature of the amorphous regions in the plastic. The material, which is brittle and fragile at normal service temperatures and completely dry, becomes tough and flexible. Apart from the desired plasticization at room temperature, the dissolution of hydrogen bonds also leads to an increase in density, and thus modulus, at low temperature, although this effect generally has no practical importance.

Polymers whose main chain contains hydrolyzable groups (e.g., polyesters, polyamides) can be irreversibly damaged by water. In addition, water that has diffused into the material acts as a radical donor, thus promoting photochemical aging (see below) and, through the loosening of the polymer structure (hydrogen bond cleavage), facilitates the diffusion of oxygen and free radicals.

A number of standard test methods have been devised to measure water absorption quantitatively: water absorption in cold water [203], [204]; water absorption in boiling water [205], [206]; and behavior in humid air [207]. In the standards, water absorption is measured relative to a well-defined dry condition, to the dry condition with allowance for constituents lost to the water, or to the as-delivered condition of a specimen. Water absorption is stated as a percent weight change or as the absolute weight change in milligrams.

2.7.2. Effect of Chemicals

Water uptake and solubility in water are just one aspect of the more general concept of chemical resistance. The reason for treating the action of water on plastics separately is the ubiquity of this medium.

Plastics are resistant to the action of many chemicals, with the exception of certain classes of compounds. Organic solvents are among the most corrosive compounds towards plastics, since the plastics are themselves organic materials. Generally, nonpolar plastics are sensitive to nonpolar solvents, and polar plastics to polar solvents.

The standards for determining chemical resistance are ISO 175 and DIN 53 476 [208], [209]. For unstressed specimens, all that is necessary is to investigate how factors such as absorption, extraction, and chemical attack alter

physical properties. Detailed tables of chemical resistance can be found in [210], [211].

Plastics in service are often subject to external mechanical stresses or residual stresses from production. A plastic may be stable against a given reagent when unstressed but fail in contact with the same reagent under stress. This phenomenon, called stress cracking, results from purely physical processes, in contrast to the stress corrosion cracking of metals (see also → Corrosion). The three most important processes are wetting, diffusion, and swelling. Wetting and swelling can increase the notch stress at weak points and microcracks that are always present in the material; diffusion is crucial for the transport of the medium to such weak points.

Methods for studying stress cracking are not completely analogous to practical experience. For example, the Bell test, specially conceived for polyethylene (PE), is very controversial. The Plax test described in ASTM 2561 has been devised for the evaluation of hollow articles made of PE. The setup enables hollow articles to be tested under constant internal pressure. The crack-opening medium is a 10 % aqueous solution of polyoxyethylnonylphenol. Cracking is detected with a simple electrical instrument.

2.7.3. Effects of Heat, Light, and Weathering

In DIN 50 035, Sheet 1, aging is defined as the totality of chemical and physical processes occurring irreversibly in a material over time. In this sense, the consequences of heat, light, and weathering are aging processes. The definition also implies that time is a significant variable in testing and, hence, that aging takes place in not spontaneous processes but relatively slow ones whose effects are seen only over a time span comparable to the service life. Thus the above definition excludes, for example, irreversible processes that are associated with processing operations [212].

Oxidation always accompanies aging due to heat and light. The most important apparatus for tests of thermal aging is the furnace. It can be operated with air recirculation and no fresh air, with fresh air circulation, or with fresh air metered into the recirculating air. Specimens can also be thermally aged by heating in a metal block, heating bath, or thermal analysis apparatus.

Light and Weathering. Weathering includes the action of many components: heat, water vapor, visible light, ultraviolet radiation, and the effects of oxygen and ozone. The synergism among these processes is often difficult to predict. In addition, the varying intensities of the components can have a more severe effect than a steady high level. The variability of the weather at any location and the difference between climatic regions render reliable correlations between an accelerated weathering test and natural weathering doubtful. Only when specimens are subjected to natural processes do weathering tests give reliable results. The possibility of accelerating the test consists in selecting a suitable weathering location. The three most important climatic regions for open-air weathering are temperate (e.g., Germany, with a mean global radiation of 400 MJ m^{-2} a^{-1}), warm-dry (Phoenix, Arizona, ca. 7900 MJ m^{-2} a^{-1}), and warm-humid (Bombay, 6300 MJ m^{-2} a^{-1}). The part of the spectrum responsible for photochemical aging of plastics ($\lambda \leq 380$ nm) accounts for only 4 – 6 % of global radiant energy; the remainder is transformed to heat when it is absorbed by the specimen. Open-air weathering tests are described in [213–215]. Resistance to natural weathering is expressed as the change in properties tested as a function of irradiation. Weatherfastness in the natural weathering test is expressed as the color change.

In many cases, open-air testing is done to determine not the weathering resistance or weatherfastness, but merely the stability to light or lightfastness (change in color only). These tests are governed by ISO 877 [216]. Color changes can be evaluated with the Woll lightfastness scale described in DIN 5003 or ISO 105 [217] or by the (more laborious) spectral method of DIN 5033 Sheet 4 or Sheet 6 (seeSection 2.5.3).

Artificial weathering and artificial illumination allow a fairly constant radiation flux to be obtained, so that irradiation is roughly proportional to time. Various types of lamps are employed. High-pressure mercury-vapor lamps are suitable for testing if short wavelengths are filtered out. Carbon arc lamps are still used, even though their spectrum bears no resemblance to sunlight. Xenon arc lamps have more recently

found wide use in the plastics industry. With suitable filters, sunlight can be modeled well by this light source, although the acceleration over sunlight is not very great. To take one example, the action of xenon light on ABS is about fives times that of sunlight.

Although accelerated tests (i.e., simulated or artificial weathering and illumination) do not yield absolute results as to weather- or lightfastness, a number of standard methods have been in use for years [218–220]. Figure 45 shows how the impact strength of impact-resistant polystyrene decreases with time under xenon arc light. A comprehensive survey of the measurement of aging criteria, the extrapolation to determine service life, and the problem of reference standards is given in [212].

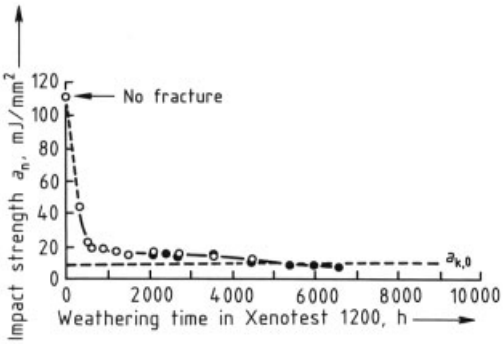

Figure 45. Impact toughness of high-impact-strength polystyrene (SB) as a function of illumination time in the Xenotest 1200 apparatus [212]
$a_{k,0}$ is the notched impact strength of the unweathered probe

2.7.4. Effect of High-Energy Radiation

The service properties of all organic polymers are impaired when they are exposed to sufficient doses of high-energy radiation. The type of radiation (e.g., electrons, gamma rays, or reactor radiation) has no influence; only the absorbed dose of energy is important.

The dose rate may, however, be significant in the presence of atmospheric oxygen. In polymers that mainly undergo radiation cross-linking in the absence of oxygen, chain degradation occurs by radiative oxidation, leading to a loss of stability. Oxygen has little effect on polymers that undergo primarily radiative degradation. Degradation by radiative oxidation depends on time, and thus on dose rate, because it is determined on the one hand by oxygen diffusion into the material and on the other by decay of the peroxide radicals formed [221], [222]. The density of the polymer also plays a role in diffusion.

Plastics can be ranked in terms of their stability to high-energy radiation by means of half-value doses. Measurement of the half-value dose for the elongation at rupture has proved useful. Thus, the half-value dose is the dosage, in grays (1 Gy = 1 J/kg = 100 rad), at which the ultimate elongation falls to half the starting value (before irradiation). Figure 46 illustrates the situation with respect to tensile strength and elongation at rupture. At a certain dose rate (e.g., 2 kGy/h), the elongation reaches half of its starting value at a dose of 32 kGy. The tensile strength declines in a similar way, but drops to half the starting value only at 60 kGy.

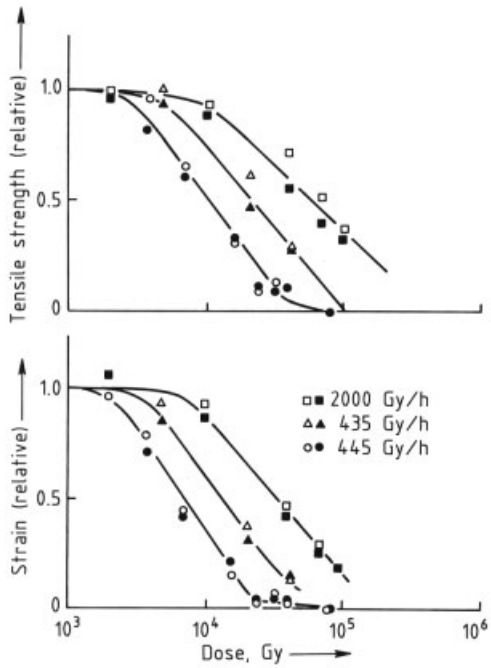

Figure 46. Tensile strength and elongation at rupture of polypropylene monofilaments (0.4-mm diameter) irradiated in air at various dose rates
Initial tensile strength: 507 N/mm² (solid circles); 397 N/mm² (open circles) [223]

The stability to high-energy radiation often depends strongly on the stabilizing system of the polymer; see Table 8 [221], which lists half-

value doses (elongation) for polypropylene with various stabilizers. WILSKI points out that all the values in this table are "correct" in themselves, but only the value for the product with no stabilizer represents the true value applicable to this product type.

Table 8. Half value dose (extension) of Hostalen PPN polypropylene [223] ∗

Stabilizer system	Half-value dose, 10^3 Gy
No stabilizer	5
N,N'-Di-β-naphthyl-*p*-phenylenediamine (DPPD)	6
Trilauryl phosphite	6
Irganox 1076	14
Irganox 1010	15
Ionox 330	17
Phenothiazine	22
Ionol	26

∗ Dose rate 24 Gy/h in air, ^{60}Co radiation, molded sheets 0.5 mm thick. All specimens contain 0.2 % calcium stearate, stabilizer concentration 0.5 %.

WÜNDRICH [222] and WILSKI [224] have provided a good introduction to this field, presenting detailed summaries of results with extensive references to other publications and values.

2.7.5. Stability against Fire

The combustion of plastics cannot be described quantitatively. It includes a number of steps, some of which have not yet been studied. For combustion to begin, three events are necessary: heating, decomposition, and ignition of the polymer. Table 9 lists the decomposition, flame initiation, and self ignition temperatures, along with the heats of combustion, for several plastics. Thermoplastics tend to soften and melt in the presence of an ignition source. The intensity of burning, smoke evolution, and the formation of residues depend on the chemical structure.

Because of their three-dimensional cross-linked structure, thermosets burn without first softening and flowing. They also generate a smaller amount of decomposition products than thermoplastics. The action of heat on their surfaces results in carbonization. In some cases, ignition does not occur.

The requirements imposed on the fire performance of plastics, especially when they are employed in construction, electrical equipment, transportation, furniture, and appliances, mean that flame retardants must be used in many cases. Flame retardants fall into two classes, reactive and additive. Reactive flame retardants are chemically incorporated in the polymer molecule and are used chiefly in polyester and epoxy resins and in polyurethanes. They have a permanent action. Additive flame retardants are usually added only after polymerization. The use of high molecular mass flame retardants also has a permanent effect.

The most common flame retardants include chlorinated and brominated compounds, as well as phosphorus compounds. The literature contains discussions of synergistic effects between individual additives.

Fire-protection tests, also called test methods for evaluating fire performance, are so numerous – and the associated problems so complex – that only a few general points can be discussed here.

All authors who have dealt with this topic state that fire performance is not a material property but depends on flammability, flame propagation, heat evolution, and several associated phenomena: fume density, toxicity, and corrosiveness of the combustion gases [225]. A particularly important point is that general statements regarding flammability and combustion – such as flammable, flame resistant, combustible, self-extinguishing, difficultly flammable, and the like – without reference to a specific test method, can lead to serious misinterpretations. A numerical value referred to a specific test method or to an unambiguous code representing flammability test results must always be given.

Tests are performed on bench, laboratory, and full scales. Most of the standards relate to specific applications such as mining, electrotechnology, transportation, furniture and appliances, and construction.

Some important testing standards are listed below.

On a bench scale, the processes in incipient combustion are studied in *reaction to fire* tests. One such is described in ISO 1210.2 [226]: The flame from a Bunsen burner is tilted at 45° and applied to horizontal specimen. The plastic is assigned to one of three categories, depending on the flammability and the duration of burning, if any. The tests DIN 53 438 [227] and ASTM D 635-77 [228] are similar.

Table 9. Decomposition temperature ranges T_d, flame initiation temperatures (FIT), self ignition temperatures (SIT), and heats of combustion ΔH for some plastics [225]

Product (abbreviations from DIN 7728)	T_d, °C	FIT, °C	SIT, °C	ΔH, kJ/kg
PE	340 – 440	340	350	46 500
PP	320 – 400	320	350	46 000
PS	300 – 400	350	490	42 000
PVC	200 – 300	390	450	20 000
ABS		390	480	36 000
PMMA	180 – 280	300	430	26 000
PA 6	300 – 350	420	450	32 000
PA 66	320 – 400	490	530	32 000
PTFE	500 – 550	560	580	
PAN	250 – 300	480	560	
Cellulose	280 – 380			17 500
Cotton		210	400	17 000

A somewhat controversial method is measurement of the *oxygen index*. The oxygen concentration at which the specimen just burns for 3 min is determined [229]. However, the oxygen content in the combustion mixture is unrealistic.

In electrotechnology, flammability measurements are material tests and product tests. The VDE standards include, in the first category, the Schramm–Zebrowski glow-bar test [230] and, in the second, the glowing-wire test [231].

Material-specific tests devised by Underwriters Laboratories (UL) in the United States have achieved worldwide recognition. Especially important test specifications include UL 94, which includes several fire safety tests on materials [232]. The UL 94 test is used in all plastics application areas except building construction.

Tests leading to the ratings 94 HB, 94 V-0, 94 V-1, and 94 V-2 are important. The letter "H" denotes horizontal clamping of the specimen, whereas "V" stands for vertical clamping. In the HB test, the flame from a Bunsen burner inclined at 45° is applied to the free end of the specimen. The evaluation criteria are burning rate or extinction of the specimen before a standard mark is reached.

In the V test, the specimen is suspended vertically and the flame from the upright burner is applied to its bottom end. The response is evaluated from continued burning, afterglow, and ignition of cotton wool by burning droplets. The best rating is 94 V-0. The test result depends on specimen thickness, which must therefore be stated.

The fire performance of plastics used in construction is evaluated in accordance with DIN 4102 [233]. This standard provides for rating as nonflammable (class A) or flammable (class B); flammable materials are further classified as difficultly flammable, normally flammable, and easily flammable. Most plastics, because of their organic nature, achieve only class B. In some cases, such as composites with inorganic materials, they are rated in class A.

2.7.6. Effect of Microorganisms

Synthetic polymers are generally not affected by microorganisms (bacteria, algae, and fungi). The picture is complicated by the presence of additives, reinforcements, plasticizers, and so forth. General instructions for the investigation of this problem in ISO 846 are set forth [234]. The result of any such action is characterized by visual description of the growth of the organisms. Two methods of study are available. In method A, determination is made of the extent to which plastics can serve as a nutrient substrate for microorganisms; in method B, the fungicidal properties of plastics are examined. See also [235].

2.8. Processing Properties and Their Testing

The processing of plastics requires a knowledge of the as-delivered properties of molding compounds, the flow behavior of thermoplastics, and the simultaneous flow and curing behavior during forming for thermosets. Molding compounds take the form of powders, granules, preformed pellets, liquids, or masses of doughy

consistency; chips, sheets, and strips are sometimes encountered.

Important processing qualities of thermoplastics, curable molding compounds, and casting resins are identified and briefly described in the following material.

2.8.1. Molding Compounds

An essential condition for trouble-free automatic processing of thermoplastic and curable molding compounds is that the material flows freely. This quality is determined by measuring the time required for the compound to run through a funnel of a given size [236]. The discharge time of a powder can be influenced, for example, by the addition of aerosols.

The particle size and particle-size distribution govern the packing density and thus the filling of a completely positive mold or, in the case of a polymer in paste form [e.g., poly(vinyl chloride) in plasticizers], the formation of a gel. Screen analyses and visual counting methods are generally employed; microscopic examination is also used. Plastics in powder form are commonly analyzed on screens having openings of 63 – 2000 µm; the series usually has six screens with mesh sizes in a geometric series [237]. Often only two fractions are specified; in the case of vinyl chloride homopolymers and copolymers, for example, the oversize fractions on a 250-µm and a 63-µm screen are stated as the R (250) and R (63) values, respectively.

Particle-counting techniques based on optical measurements are important in the determination of particle-size distribution. In the laser granulometer, the beam from a helium–neon laser is guided through a cuvette that contains the specimen in dispersion form. The particles diffract the coherent beam, forming diffraction rings whose diameters are inversely proportional to the grain size. The size of the diffraction ring determines whether a particle is counted or not, depending on its spherical volume. The number of particles counted in a fraction (a given volume range) is converted to their volume and then to mass. In this way, the mass fraction is obtained as a function of the mean particle size (similar to screen analysis) [238].

The bulk density is informative in relation to metering. This value is the ratio of mass to volume when the molding compound is treated in a particular way (compacted, if apparent density is being measured). For long-strand fibers and chips, an apparent density is measured. The bulk density is easily determined with a filling funnel and a graduated flask.

The essential processing criterion for reaction resins is the viscosity as delivered; the viscosity after warm storage is also used for epoxy resins.

For reaction molding compounds, the time required to reach a given viscosity limit or the gelling time (open time, pot life) is measured under standard conditions [239], and the curing behavior (curing time) is determined.

The curing time is the time after which certain properties (e.g., heat deflection temperature, acetone-soluble matter, or styrene monomer content in the case of unsaturated polyester resins) no longer vary within certain limits.

2.8.2. Conventional Methods for Characterizing Flow Behavior

The only source of adequate information about the flow behavior of polymers is the flow curve (see Section 2.8.3). However, experimentally plotting a family of flow curves at a sufficient number of temperatures is laborious, and only in exceptional cases can this be done in the context of process monitoring. Attempts have therefore been made to characterize the flow behavior by simple, conventional measuring methods. This is especially appropriate for thermosetting molding compounds because curing takes place during processing and makes impossible the analysis of rheological phenomena in a simple way [240].

One of the most frequently used conventional test methods for the flow behavior of thermoplastics is determination of the melt flow index (MFI) or the melt volume index (MVI) [241]. For MFI, the quantity measured is the amount of polymer melt extruded from a nozzle in 10 min when a specified force is applied with a plunger (units of g/10 min). In automatic instruments, the plunger motion is measured as a function of time to give an MVI; the specimen volume expelled through the nozzle in a specified time under the stated conditions is determined (units of cm^3/10 min).

When the melt index values do not correlate with in-service behavior (chiefly for injec-

tion molding), the flow behavior is assessed on the basis of the spiral flow test. A specimen in the form of a spiral is made in an injection-molding machine; the flow qualities are evaluated in terms of the length of the spiral.

In production monitoring and control, the melt viscosity or melt index is measured continuously with bypass rheometers (Fig. 47) [242], [243].

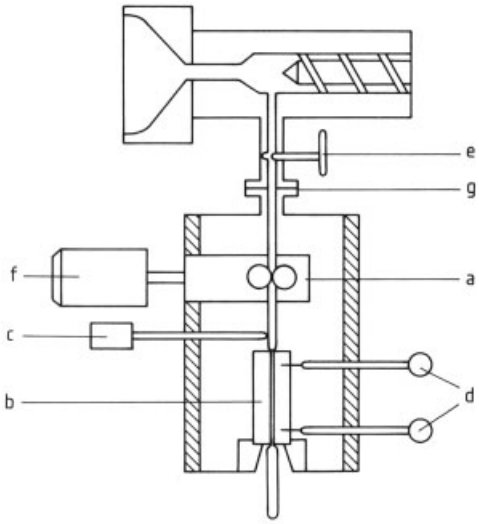

Figure 47. Göttfert bypass rheograph, open one-pump system with independent pressure and temperature controls
a) Melt pump; b) Nozzle; c) Pressure pickup; d) Temperature pickup; e) Shutoff valve; f) Pump drive motor; g) Flange

The flow behavior of curable molding compounds is far more difficult to characterize than that of thermoplastics, because in thermosets a chemical curing reaction proceeds while flow continues during the production process. The names of the test methods for these resins reflect the shapes of the specimens molded in the test [244].

If information beyond that offered by these simple methods is required, the measurements must conform to well-defined and easily understood techniques, as in the case of thermoplastics. Such measurements have been described by Kanavec and others [245], [246]. The variation of viscosity with time is determined in a rotational viscosimeter. Use of this method means that processing conditions can be varied to an extent comparable with plant practice. The Bauer and Eichler kneading-chamber method [247] and the Philips method [248] cannot be described in detail here.

2.8.3. Rheology of Plastics

Earlier (Section 2.8.2), the flow behavior of plastics melts was said to be describable only in terms of flow curves, that is, plots of shear stress versus shearing rate. The measurements are performed with simple shear deformation. Because deformations in processing nearly always involve extension, the extensional properties of the melt must often be investigated as well . For thermoplastics, the term melt refers to the condition above the glass-transition temperature for polymers that solidify in glassy form (e.g., polystyrene or polycarbonate) and to the condition above the maximum melting point of the crystalline regions for polymers that solidify in crystalline form.

The essential features in the rheology of thermoplastic melts are easily observable when a melt is extruded from a nozzle; see Figure 48 [249]. The behaviors illustrated at a, b, and c can be described as follows:

a) *Structural* (*non-Newtonian*) *viscosity.* The flow rate m increases faster than proportionally to the extrusion pressure p; that is, viscosity decreases as the mechanical load applied to the melt increases.

b) *Entropy elasticity.*The molecules become oriented by the flow in the nozzle. This corresponds to a rubberlike elastic deformation, which is canceled as soon as the force is released outside the nozzle. The extrudate has a much larger diameter than the nozzle. This phenomenon is called extrudate swelling.

c) *Viscoelasticity.*The coupling between viscous and elastic processes is time dependent.

This behavior implies that melts (and also solutions) of polymers are non-Newtonian, rubberlike, viscoelastic liquids.

Shear Viscosity. The flow behavior during shear deformation is described by the shear stress τ and the time derivative of the shear γ, that is, the shear rate $\dot{\gamma}$. In addition to the nonlinear formulas $\tau = \mathbf{F}(\dot{\gamma})$ and $\dot{\gamma} = \mathbf{f}(\tau)$, a viscosity η is also defined for polymers: $\eta = \tau/\dot{\gamma}$.

This viscosity is independent of the shear rate only at very low values of the latter; that is, only in this region does a polymer melt behave like a Newtonian liquid (linear flow region). At the high shear rates used in plastics processing, viscosity decreases rapidly. The high rate of deformation obviously changes the tangled-chain structure in the melt. This concept gives rise to the term "structural" viscosity. Thus, quite generally, $\tau = \eta_s(\dot{\gamma})\,\dot{\gamma}$, where η_s is the structural viscosity.

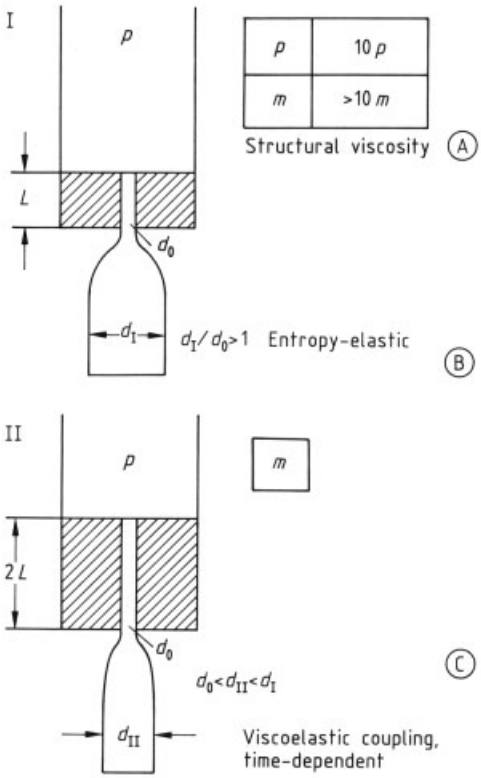

Figure 48. Rheological properties of plastic melts [249]

The shear-rate-dependent viscosity at very low shear rates is called the zero-shear viscosity:

$$\eta_0 = \lim_{\dot{\gamma}\to 0} \eta_s(\dot{\gamma})\,.$$

Figure 49 shows curves of viscosity versus shear rate for a Polyamide 6 melt at various temperatures.

In both the linear and the nonlinear regions, viscosity decreases with increasing temperature; see also [250].

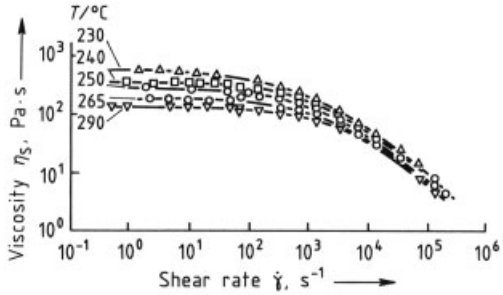

Figure 49. Viscosity of a polyamide 6 melt at various temperatures

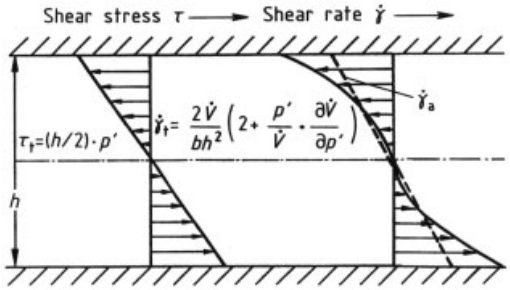

Figure 50. Steady-state shear flow in a flat-slot capillary [251]
τ_t = true shear stress at wall; $\dot{\gamma}_t$ = true shear rate at wall; $\dot{\gamma}_a$ = apparent shear rate at wall; $\dot{V}$ = volume flow rate; p' = axial pressure gradient; b = capillary width
Dashed curve: Newtonian liquid; solid curve: structurally viscous liquid

The zero-shear viscosity η_0 is strongly dependent on the molecular mass above a critical value $M_r = M_c$. For commercial polymers, which always have $M_r > M_c$, this strong dependence of η_0 on M_r also manifests itself in the viscosity function $\eta(\dot{\gamma})$. The molecular-mass distribution, on the other hand, influences the elastic phenomena associated with viscous flow. Extrudate swelling becomes increasingly marked as the distribution becomes broader. The form of the viscosity function stays the same as molecular mass varies but changes with molecular-mass distribution.

Flow curves are usually determined with capillary viscosimeters. A disadvantage of the generally very convenient capillary viscosimeter is that the shear rate in the melt is not constant from point to point. Shear stress and shear rate increase from the center of the capillary toward the wall. The shear rate is zero at the capillary axis (see Fig. 50). In contrast, all particles of the liquid can be subjected to the same shear rate

in rotational viscosimeters. This uniformity of shear rate cannot be achieved in a capillary.

Because it is easy to use, however, the capillary viscosimeter is employed widely in the roudetermination of flow curves. A standard method has been devised so that the results obtained can be compared [252]. The melt is forced through a capillary of specified size; either the test pressure at a specified volume flow rate or the volume flow rate at a specified test pressure is determined. In general, one charge in the cylinder yields several pairs of volume flow rate and test pressure values. The apparent shear rate $\dot{\gamma}_a$ is calculated from the flow rate and the capillary diameter, and the apparent shear stress τ_a is calculated from the capillary dimensions and the test pressure. The form $\dot{\gamma}_a = f(\tau_a)$ is recommended for flow curves used in comparisons (e.g., in quality control). The capillaries must have identical dimensions. The use of apparent flow curves and viscosity curves for the design of flow channels should be avoided. The Weissenberg and Rabinowitsch correction method [253], which cannot be discussed further here, gives the shear stress at the capillary wall (true shear stress τ_t) and the true shear rate $\dot{\gamma}_t$. The geometry-independent flow curve $\dot{\gamma}_t = f(\tau_t)$ and the viscosity function $\eta(\dot{\gamma}) = \tau_t/\dot{\gamma}_t$ are obtained from these values. The functions cited are indeed measured for the wall layer only, but they are valid in the bulk of the liquid as well, so that the subscript t can be omitted in the notation (units are s^{-1} for shear rate, Pa for shear stress, and $Pa \cdot s$ for viscosity). If the measured flow curves or viscosity functions are to be used for calculating steady-state shear flows in plastics-processing machinery, it is desirable to derive a mathematical approximation to the measured curves. Two approaches have proved useful: the Ostwald/de Waele power law approximation and the Carreau approximation [254], [255]. A large number of flow curves for commercial thermoplastics have been collected in [256].

Extensional Viscosity. During many processes employed in the plastics industry, such as melt spinning, film blowing, blow molding, coating at high takeoff speed, and injection molding, the melt is subjected not only to shearing but also to extension. Often the last deformation process performed on the melt before it hardens is an extension process that establishes the molecular orientation in the finished article. In contrast to low-viscosity Newtonian liquids with low tensile strength, plastic melts can be extended greatly without rupturing. The total extension consists of an entropy-elastic flow component and a viscous flow component; the ratio of the two components depends strongly on the external conditions and the product. The extensional deformation ε is characterized by the length change dl relative to the instantaneous length l: $d\varepsilon = dl/l$ (Hencky's measure).

Thus the Hencky extension ε of a rod with initial length l_0 is given by $\varepsilon = \int_{l_0}^{l} dl/l = \ln(l/l_0) = \ln\lambda$. The Hencky extension is thus equal to the natural logarithm of the degree of stretching λ, and the rate of extension is $\dot{\varepsilon} = \dot{l}/l_0$.

The measurement of stress – strain or tensile creep curves of melts is described in [256], [257]. An arrangement described by Münstedt is shown in Figure 51, whereas Figure 52 shows stress – strain curves of a low-density polyethylene specimen at various constant extension rates [258].

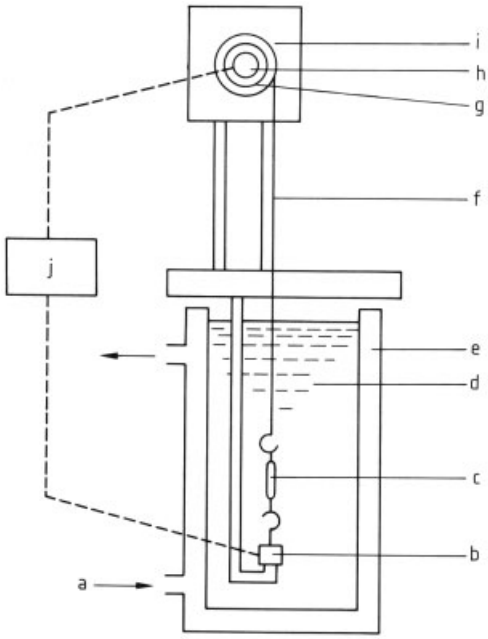

Figure 51. Münstedt extensional viscosimeter [256]
a) Heating liquid; b) Load-measuring cell; c) Specimen; d) Oil bath; e) Temperature-controlled vessel; f) Belt; g) D.c. motor; h) Strain gauge; i) Takeup spool; k) Monitoring and control unit

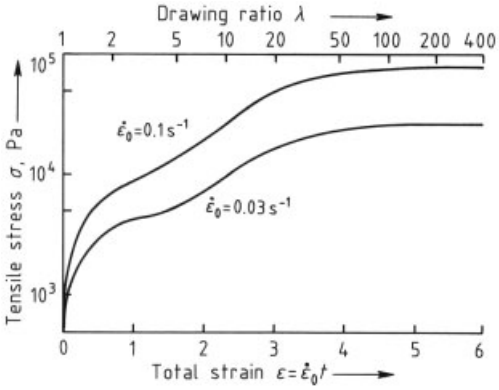

Figure 52. Stress – strain curves for low-density polyethylene melt at 150 °C for various constant rates of extension

The extensional viscosities derived from stress-strain curves are plotted as a function of either extension at constant tensile stress σ_0 or extension at constant extension rate $\dot{\varepsilon}$. The extensional viscosity $\eta_F(t)$ is given by $\eta_F(t) = \sigma(t)/\dot{\varepsilon}_0$, where $\sigma(t)$ is the measured tensile stress relative to the true cross-sectional area at time t, and $\dot{\varepsilon}_0$ is the extension rate determined in an extensional rheometer, which is constant over time; see also [259].

Finally, some flow anomalies of importance in plastics processing should be mentioned [259]. When a critical flow rate or wall shear stress value is exceeded, rough areas, termed melt fractures, are seen on the extrudate surface. These phenomena cannot be attributed to the critical turbulence of a flow, determined by the Reynolds number; they are due to the viscoelasticity of the melt (elastic turbulence). The onset of elastic turbulence is associated with a critical state of stress. This critical stress decreases as the molecular-mass distribution of the melt broadens. Modifications in the nozzle inlet reduce the occurrence of melt fracture [260].

If the melt is subjected to extensional loading at the nozzle outlet, periodic rupturing of the extrudate surface can be induced. The scaly surface is referred to as sharkskin. A periodic alternation of wall adhesion and wall slipping is observed at a critical wall stress in the nozzle, particular in high-density polyethylene melts. This phenomenon can also lead to a surface structure resembling melt fracture.

3. References

1. H. P. Klug, L. E. Alexander: *X-ray Diffraction Procedures,* J. Wiley, New York 1974.
2. I. Voigt-Martin, *Adv. Polym. Sci.* **67** (1985) 194.
3. B. J. Berne, R. Pecora: *Dynamic Light Scattering,* J. Wiley, New York 1976.
4. G. D. Wignall: "Neutron Scattering," *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vol. 10,** J. Wiley, New York 1987.
5. P. J. Flory, *J. Chem. Phys.* **17** (1949) 303.
6. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca 1953.
7. G. Lieser, E. W. Fischer, K. Ibel, *J. Pol. Sci.* **13** (1975) 29.
8. F. S. Bates, S. B. Dierker, G. D. Wignall, *Macromolecules* **19** (1986) 1938.
9. F. S. Bates, G. D. Wignall, *Phys. Rev. Let.* **57** (1986) 1429.
10. G. D. Wignall, D. G. H. Ballard, J. Schelten, *Eur. Polym. J.* **10** (1974) 861.
11. A. M. Fernandez, L. H. Sperling, G. D. Wignall, *Macromolecules* **19** (1986) 2572.
12. E. W. Fischer, M. Stamm, M. Dettenmaier, P. Herschenröder, *ACS Polymer Preprints* **20** (1) (1979) 219.
13. D. G. H. Ballard, P. Cheshire, G. W. Longman, J. Schelten, *Polymer* **19** (1978) 379.
14. R. G. Kirste, W. A. Kruse, J. Schelten, *Makromol. Chem.* **162** (1972) 299.
15. R. G. Kirste, W. A. Kruse, J. Schelten, *Kolloid Z. Z. Polym.* **251** (1973) 919.
16. R. G. Kirste, W. A. Kruse, K. Ibel, *Polymer* **16** (1975) 120.
17. D. G. H. Ballard, G. D. Wignall, J. Schelten, *Eur. Polym. J.* **9** (1973) 965.
18. P. Herschenröder, Thesis, University of Mainz 1978.
19. K. P. McAlea, J. M. Schultz, K. H. Gardner, G. D. Wignall, *Macromolecules* **18** (1985) 477.
20. J. Schelten et al., *Polymer* **17** (1976) 751.
21. J. M. Guenet, *Polymer* **22** (1981) 313.
22. P. J. Flory: *Statistical Mechanics of Polymer Molecules,* J. Wiley, New York 1969.
23. A. Guinier: *X-ray Diffraction,* W. H. Freeman, San Francisco 1963.
24. E. W. Fischer et al., *Faraday Discuss. Chem. Soc.* **68** (1979) 26.
25. J. T. Bendler, *Macromolecules* **10** (1977) 162.
26. M. Dettenmaier et al., *Macromolecules* **19** (1986) 773.
27. H. Ito, T. P. Russell, G. D. Wignall, *Macromolecules* **20** (1987) 2213.

28. J. M. O'Reilley, D. M. Teegarden, G. D. Wignall, *Macromolecules* **18** (1985) 2747.
29. W. Gawrisch, M. G. Brereton, E. W. Fischer, *Polym. Bull.* **4** (1981) 687.
30. M. Vacatello, D. Y. Yoon, P. J. Flory, *Macromolecules* **23** (1990) 1993.
31. B. C. Laskowski, D. Y. Yoon, D. McLeand, R. J. Jaffe, *Macromolecules* **21** (1988) 1629.
32. E. R. Haward: *The Physics of Glassy Polymers, Materials Science Series,* Appl. Sci. Publ., London 1973. S. E. Keinath, R. L. Miller, J. K. Rieke (eds.): *Order in the Amorphous "State" of Polymers,* Plenum Press, New York 1987.
33. L. C. E. Struik: *Physical Aging in Amorphous Polymers and Other Materials,* Elsevier Publ., Amsterdam 1978.
34. E. A. Di Marzio, J. H. Gibbs, *J. Polym. Sci.* **A 1** (1963) 1417. E. A. Di Marzio, J. H. Gibbs, *Macromolecules* **9** (1976) 763.
35. J. Jäckle, *Rep. Prog. Phys.* **49** (1986) 171.
36. A. J. Kovacs, *Fortschr. Hochpolym. Forsch.* **3** (1966) 394.
37. J. D. Ferry: *Viscoelastic Properties of Polymers,* 3rd ed., J. Wiley, New York 1980.
38. D. W. van Krevelen: *Properties of Polymers, Their Estimation and Correlation with Chemical Structure,* 2nd ed., Elsevier, Amsterdam 1976.
39. J. Bandrup, E. H. Immergut: *Polymer Handbook,* 3rd ed., J. Wiley, New York 1989.
40. D. R. Paul, S. Newman: *Polymer Blends,* vols. 1 and 2, Academic Press, New York 1978.
41. I. M. Ward: *Mechanical Properties of Solid Polymers,* 2nd ed., J. Wiley, Chichester 1983.
42. U. Eisele, *Progr. Colloid Polym. Sci.* **66** (1979) 59.
43. H. Tadokaro: *Structure of Crystalline Polymers,* J. Wiley, New York 1979.
44. *Faraday Discuss. Chem. Soc.* **68** (1979), various papers and references therein.
45. M. Stamm, E. W. Fischer, M. Dettenmaier, P. Convert, *Faraday Discuss. Chem. Soc.* **68** (1979) 263.
46. P. J. Flory, *J. Am. Chem. Soc.* **84** (1962) 2857.
47. D. Y. Yoon, P. J. Flory, *Macromolecules* **17** (1984) 868.
48. E. W. Fischer et al., *J. Polym. Sci. Polym. Phys. Ed.* **22** (1984) 1491.
49. S. J. Spell, D. M. Sadler, *Polymer* **25** (1984) 739.
50. J. M. Guenet, *Polymer* **22** (1981) 313. J. P. Guenet, C. Picot, *Macromolecules* **16** (1983) 205.
51. S. J. Spells, A. Keller, D. M. Sadler, *Polymer* **25** (1984) 749.
52. B. Wunderlich: "Crystal Structure, Morphology, Defects," *Macromolecular Physics,* vol. 1, Academic Press, New York 1973.
53. A. Turner Jones, A. J. Cobbold, *J. Polym. Sci. Polym. Lett. Ed.* **6** (1968) 539. D. R. Morrow, *J. Macromol. Sci. Phys.* **3** (1969) 53. A. Turner Jones, *Polymer* **12** (1971) 487.
54. I. G. Voigt-Martin, private communication.
55. D. C. Bassett: *Principles of Polymer Morphology,* Cambridge Univ. Press 1981.
56. I. M. Ward: "The Preparation, Structure and Properties of Ultra-high Molecular Flexible Polymers," *Adv. Polym. Sci.* **70** (1985) 1.
57. H. H. Kausch: *Polymer Fracture,* 2nd ed., Springer Verlag, Berlin 1987.
58. U. Eisele: *Introduction to Polymer Physics,* Springer Verlag, Berlin 1990.
59. A. Blumstein (ed.): *Polymeric Liquid Crystals,* Plenum Press, New York 1985.
60. R. A. Weiss, C. K. Ober (eds.): "Liquid-crystalline Polymers," *ACS Symp. Ser.* **435,** Washington 1990.
61. J. M. G. Dobb, J. E. McIntyre (eds.): *Adv. Polym. Sci.* **60/61** (1984), various papers in this volume.
62. M. Ballauff, *Angew. Chem. Int. Ed. Engl.* **101** (1989) 261.
63. D. Dutta, H. Fruitwale, A. Kohli, R. A. Weiss, *Polym. Eng. Sci.* **30** (1990) 1005.
64. O. Olabisi, L. M. Roberson, M. T. Shaw: *Polymer-Polymer Miscibility,* Academic Press, New York 1979.
65. L. A. Utracki, R. A. Weiss (eds.): "Multiphase Polymers: Blends and Ionomers," *ACS Symp. Ser.* **395,** Washington 1989.
66. G. E. Molan (ed.): *Colloidal and Morphological Behavior of Blockcopolymers,* Plenum Press, New York 1971.
67. F. S. Bates, J. H. Rosedale, G. H. Frederickson, *J. Chem. Phys.* **92** (1990) 6255.
68. R. Fayt, R. Jerome, P. Teyssie, in [65] Chap. 2, pp. 38 – 66. References for Chapter 2.
69. J. Schmitz, E. Bornschlegl, G. Dupp, G. Erhard, *Plastverarbeiter* **39** (1988) 50.
70. H. Oberst in G. Schreyer (ed.): *Konstruieren mit Kunststoffen,* Hanser Verlag, München 1972, p. 444.
71. L. R. G. Treloar in H. A. Stuart (ed.): *Die Physik der Hochpolymeren,* **vol. 4,** Springer Verlag, Berlin – Göttingen – Heidelberg 1956, p. 294 and p. 311.
72. Lord Kelvin: *Encyclopedia Britannica,* London 1875.

73. W. Voigt, *Abh. Ges. Wiss. Göttingen, Math. Phys. Kl.* **36** (1890).
74. A. J. Staverman, F. Schwarzl in [71] p. 1 – 95.
75. H. Oberst, L. Bohn, *Rheol. Acta* **1** (1961) 608 – 617.
76. W. Retting, *J. Polym. Sci. Polym. Symp.* **42** (1973) 605 – 615.
77. A. J. Staverman, F. Schwarzl in [71] p. 11 – 29.
78. DIN 53 444, Prüfung von Kunststoffen, Zeitstandzugversuch, Entwurf 1987.
79. DIN 53 455, Prüfung von Kunststoffen, Zugversuch, 1988.
80. K. Oberbach: *Kunststoff-Kennwerte für Konstrukteure,* Hanser Verlag, München 1975.
81. W. Kaufmann, D. Hoffmann, *Kunststoffe* **59** (1969) 173.
82. DIN 53 445, Prüfung von Kunststoffen, Torsionsschwingungsversuch, 1986.
83. DIN 53 440, Biegeschwingungsversuch, Blatt 1 – 3, 1973.
84. B. Carlowitz, *Plastverarbeiter* **31** (1980) 84.
85. H. Oberst in G. W. Becker, J. Meißner, H. Oberst, H. Thurn (eds.): *Elastische und viskose Eigenschaften von Werkstoffen,* Beuth-Vertrieb, Berlin – Köln – Frankfurt 1963, p. 19.
86. K. Oberbach, *Kunststoffe* **79** (1989) 713.
87. DIN 53 455, Prüfung von Kunststoffen, Zugversuch, 1988.
88. DIN 53 452, Prüfung von Kunststoffen, Biegeversuch, 1988.
89. K. Oberbach, L. Rupprecht: "Wege zur Reduzierung des Aufwandes bei der Kunststoffprüfung und zur Förderung der Vergleichbarkeit der ermittelten Kennwerte," *Vortrag auf der Fachtagung "Möglichkeiten der Rationalisierung bei Kunststoff-Prüfungen",* Süddeutsches Kunststoff-Zentrum, Würzburg, 2.–3. 12. 1986.
90. DIN 53 454, Prüfung von Kunststoffen, Druckversuch, Entwurf 1988.
91. DIN 53 457, Prüfung von Kunststoffen, Bestimmung des Elastizitätsmoduls im Zug-, Druck- und Biegeversuch, 1987.
92. W. Schneider, R. Bardenheier, *Z. Werkstofftech.* **6** (1975) 269 – 280.
93. R. Bardenheier, *Z. Werkstofftech.* **8** (1977) 379 –388.
94. ISO 2039, Plastics and Ebonit, Determination of Hardness by Ball Indentation Method, 1973.
95. DIN 53 456, Prüfung von Kunststoffen, Härteprüfung durch Eindruckversuch, 1973.
96. ISO 868, Plastics, Determination of Indentation Hardness by Means of a Durometer (Shore Hardness), 1978.
97. DIN 53 505, Härteprüfung nach Shore A und D, 1973.
98. DIN 53 462, Prüfung von Kunststoffen, Bestimmung der Formbeständigkeit in der Wärme nach Martens, 1987.
99. DIN 53 461, Prüfung von Kunststoffen, Bestimmung der Formbeständigkeitstemperatur, 1987.
100. DIN 53 460, Prüfung von Kunststoffen, Bestimmung der Vicat-Erweichungstemperatur.
101. DIN 53 754, Prüfung von Kunststoffen, Bestimmung des Abriebs nach dem Reibradverfahren.
102. H. Hachmann, E. Strickle, *Kunststoffe* **59** (1969) 47.
103. J. C. Halpin, *Rubber Chem. Technol.* **38** (1965) 1007 – 1038.
104. J. C. Williams: *Fracture Mechanics of Polymers,* Horwood, Chichester 1984.
105. A. S. Argon, *Philos. Mag.* **28** (1973) 839 – 865.
106. G. Jacoby, Ch. Cramer, *Rheol. Act* a 7 (1968) 23 –51.
107. B. Rosen: "Phenomenological Aspects of Progressive Weakening in the Highly Viscoelastic Glasses" in H. Rosen (ed.): *Fracture Processes in Polymeric Solids,* J. Wiley&Sons, New York 1964.
108. P. Beardmore, S. Rabinowitz, *J. Mater. Sci.* **10** (1975) 1763 – 1770.
109. R. P. Kambour, *J. Polym. Sci. Part A* **2** (1964) 4159 – 4163.
110. R. P. Kambour, *Macromol. Rev.* **7** (1973) 1 – 154.
111. S. S. Sternstein, L. Ongchin, A. Silverman, *Appl. Polym. Symp.* **7** (1968) 175 – 199.
112. S. S. Sternstein, F. A. Myers, *J. Macromol. Sci. Phys.* **B 8** (1973) no. 3 and 4, 539 – 571.
113. L. Bohn, *Angew. Makromol. Chem.* **20** (1971) 129.
114. B. Carlowitz: *Untersuchungen zum Bruchverhalten ausgewählter thermoplastischer Kunststoffe bei Schlagbeanspruchungen,* Technisch-wissenschaftlicher Bericht des IKV Aachen (ed.), Aachen 1979.
115. ISO 180, Determination of the IZOD-Impact Resistance of Rigid Plastics, 1982.
116. ISO 179, Determination of the CHARPY-Impact Resistance of Rigid Plastics, 1982.
117. DIN 53 453, Prüfung von Kunststoffen, Schlagbiegeversuch, 1982.

118. H. Dietmann, *Arch. Eisenhüttenwes.* **40** (1969) 1011 – 1022.
119. H. Grimminger, U. Troltenier, *Materialprüfung* **4** (1962) 4.
120. W. Grobusch, H. Jesse, *Kunststoffe* **58** (1968) 648.
121. DIN 53 443 Blatt 1, Prüfung von Kunststoffen, Stoßversuch Fallbolzenversuch, 1975.
122. J. Cornfield, N. Mantel, *J. Am. Stat. Assoc.* **45** (1950) 181.
123. M. S. Bartlett, *J. Statistical Society (Supplement)* **VIII** (1946) 113.
124. DIN 53 443 Blatt 2, Prüfung von Kunststoffen, Stoßversuch Durchstoßversuch mit elektronischer Meßwerterfassung.
125. Schenk & Co., Darmstadt, personal communication.
126. A. S. Wolmir: *Biegsame Platten und Schalen,* VEB-Verlag für Bauwesen, Berlin 1962.
127. E. Hensen in E. Hensen, W. Knappe, H. Potente (eds.): *Handbuch der Kunststoff-Extrusionsanlagen,* **vol. 2,** Hanser Verlag, München – Wien 1986, p. 277.
128. G. Wübken, *Plastverarbeiter* **26** (1975) 17 – 23.
129. H. H. Racke, Th. Fett, *Instn. Mech. Engnrs., Experimental Stress Analysis,* Paper 39, London.
130. G. Kämpf, *Angew. Makromol. Chem.* **60/61** (1977) 292 – 346.
131. DIN 61 853, 61 854, 61 855, Textilglas für die Kunststoffverstärkung, 1975.
132. R. Taprogge, R. Scharwächter, P. Hähnel: *Faserverstärkte Hochleistungs-Verbundwerkstoffe,* Vogel-Verlag, Würzburg 1975.
133. DIN 4109 Teil 1 – 5, Schallschutz im Hochbau, 1962 – 1979.
134. Th. Grewer, H. Wilski, *Kolloid Z. Z. Polym.* **226** (1968) 46.
135. H. Wilski, *Kolloid Z. Z. Polym.* **238** (1970) 426.
136. G. Nachtrab, *Kunststoffe* **60** (1970) 261.
137. W. Knappe, G. Nachtrab, G. Weber, *Kunststoffe* **62** (1972) 455.
138. DIN 51 005, Thermische Analyse, Begriffe.
139. W. Hemminger, H. J. Seifert, *Nachr. Chem. Tech. Lab.* **34** (1986) M 1.
140. H. Wilski, D. Strobl, Unveröffentlichte Ergebnisse.
141. B. Wunderlich, H. Baur: "Heat Capacities of Linear High Polymers," *Adv. Polymer Sci.* **7** (1970) 151.
142. K.-H. Illers, *Makromol. Chem.* **127** (1969) 1.
143. B. Wunderlich in E. A. Turi (ed.): *Thermal Characterization of Polymeric Materials,* "The Basis of Thermal Analysis," Academic Press, New York 1981.
144. W. Dietz, *Tech. Mess. ATM* V 9213-6 (1976) 123, 153, 195.
145. Y. Agari, T. Uno, *J. Appl. Polym. Sci.* **32** (1986) 5705.
146. DIN 52 612, Teil 1, Bestimmung der Wärmeleitfähigkeit mit dem Plattengerät.
147. W. Fritz, W. Küster in *Wärme- und Stoffübertragung,* **vol. 3,** Springer Verlag, Heidelberg 1979, p. 156.
148. S. Goldfein, J. Calderon, *J. Appl. Polym. Sci.* **9** (1965) 2985.
149. I. Catie: , Dissertation, Aachen 1972.
150. K. Ueberreiter, W. Klein, *Chem. Tech.* **15** (1942) 5.
151. K. H. Hellwege, H. Hennig, W. Knappe, *Kolloid Z. Z. Polym.* **183** (1962) 110.
152. K. Ueberreiter, G. Kanig, *Glas Hochvak. Tech.* **2** (1953) 218.
153. H. Wilski, *Kunststoffe* **54** (1964) 10.
154. V.-H. Karl, F. Asmussen, K. Ueberreiter, *Angew. Makromol. Chem.* **62** (1977) 145.
155. V.-H. Karl, F. Asmussen, K. Ueberreiter, *Makromol. Chem.* **178** (1977) 2649.
156. V.-H. Karl, F. Asmussen, K. Ueberreiter, *Progr. Colloid Polym. Sci.* **64** (1978) 97.
157. G. Menges, P. Thienel, *Kunststoffe* **65** (1975) 696.
158. DIN 53 482/VDE 0303, Teil 3/05, Prüfung von Werkstoffen für die Elektrotechnik; Messung des elektrischen Widerstandes von nichtmetallenen Werkstoffen, 1983.
159. DIN VDE 0303, Teil 4, Bestimmungen für elektrische Prüfungen von Isolierstoffen; Bestimmung der dielektrischen Eigenschaften, 1969.
160. DIN VDE 0303, Teil 13, Prüfung von Isolierstoffen, Dielektrische Eigenschaften fester Isolierstoffe im Frequenzbereich von 8,2 bis 12,5 GHz, 1986.
161. Rohde & Schwarz-Mitteilung, *Meßgeräte, Meß-systeme,* München 1988.
162. S. Roberts, A. v. Hippel, *J. Appl. Phys.* **17** (1946) 610.
163. D. Kind, H. Kärner: *Hochspannungs-Isoliertechnik (für Elektrotechniker),* Vieweg & Sohn, Braunschweig 1982.
164. DIN 53 481/VDE 0303, Teil 2, VDE-Bestimmungen für elektrische Prüfungen von Isolierstoffen; Durchschlagspannung, Durchschlagfestigkeit, 1974.

165. E DIN VDE 0303, Teil 2/IEC 15 A (CO) 52, Prüfung von Isolierstoffen, Durchschlagspannung und Durchschlagfestigkeit bei technischen Frequenzen, 1987.
166. H. J. Mair: *Isolier- und Mantelwerkstoffe aus thermoplastischen Kunststoffen in Kabel und isolierten Leitungen,* VDI-Verlag, Düsseldorf 1984.
167. H. J. Mair in G. Braun, G. W. Becker, B. Carlowitz (eds.): "Die Kunststoffe, Chemie, Physik, Technologie," *Kunststoff-Handbuch,* vol. 1, 2nd ed., Hanser Verlag, München 1990, p. 760.
168. DIN VDE 0303, Teil 1/IEC 112, Verfahren zur Bestimmung der Vergleichszahl und Prüfzahl der Kriechwegbildung auf festen Isolierstoffen unter feuchten Bedingungen, 1984.
169. E DIN VDE 0303, Teil 10/IEC 587, Prüfung von Isolierstoffen; Hochspannungskriechwegbildung, 1987.
170. DIN 83 485, Teil 1/VDE 0303 Teil 7, VDE-Bestimmungen für elektrische Prüfungen von Isolierstoffen; Verhalten unter Einwirkung von Oberflächen-Glimmentladungen, 1974.
171. E DIN VDE 0303, Teil 7 IEC 15 B (CO) 65, Prüfung von Isolierstoffen; Oberflächen-Glimmentladungen, 1987.
172. DIN VDE 0303, Teil 6, Bestimmung der elektrolytischen Korrosion, 1968.
173. E DIN VDE 0303, Teil 6/IEC 426, Bestimmung der korrosionsbegünstigenden Wirkung durch isolierende Werkstoffe, 1987.
174. K.-H. Möbius, *Kunststoffe* **78** (1988) 53.
175. K.-H. Möbius, *Kunststoffe* **78** (1988) 345.
176. S. Roth, *Phys. Bl.* **40** (1984) 321.
177. H. J. Mair, S. Roth (eds.): *Elektrisch leitende Kunststoffe,* Hanser Verlag, München – Wien 1986.
178. H.-G. Elias, F. Vowinkel: *Neue polymere Werkstoffe für die industrielle Anwendung,* Kapitel 14, Elektrisch leitfähige Polymere, Hanser Verlag, München – Wien 1983.
179. ISO R 489, Plastics, Determination of the Refractive Index of Transparent Plastics, 1966.
180. DIN 53 491, Prüfung von Kunststoffen, Bestimmung der Brechungszahl und Dispersion, 1955.
181. G. L. Wilkes, *Adv. Polym. Sci.* **8** (1971) 91 – 136.
182. DIN 5036, Teil 1 bis Teil 4, Strahlungsphysikalische und lichttechnische Eigenschaften von Materialien, 1977, 1978, 1979.
183. DIN 53 490, Prüfung von Kunststoffen, Bestimmung der Trübung von durchsichtigen Kunststoffschichten.
184. DIN 67 530, Reflektometer als Hilfsmittel zur Glanzbeurteilung an ebenen Anstrich- und Kunststoffoberflächen.
185. H. Haussühl, K. Hamann, *Farbe Lack* **64** (1958) 642.
186. Th. Kosbahn, *Farbe Lack* **70** (1964) 693.
187. J. S. Christie, R. S. Hunter, *SPE Coloring and Decorating of Plastics* **9** (1975) 33.
188. DIN 5033, Blatt 1 – 8, Farbmessung, 1966 – 1980.
189. ASTM E 308-66, Recommended Practice for Spectrophotometry and Description of Color in CIE 1931 System.
190. BS 4727, Part 4, Terms Particular to Light and Color, 1971.
191. ISO 3558, Plastics, Assessment of the Color of Near White or Near Colorless Materials.
192. DIN 16 776, Kunststoff-Formmassen, Polyethylen (PE)-Formmassen Einteilung und Bezeichnung, 1984.
193. ISO 1183, Methods for Determining the Density of Plastics excluding Cellular Plastics.
194. DIN 53 479, Prüfung von Kunststoffen und Elastomeren, Bestimmung der Dichte, 1976.
195. H. Descheimer, O. Fuchs in R. Nitsche, K. A. Wolf (eds.): *Kunststoffe,* **vol. 1,** Springer Verlag, Berlin – Göttingen – Heidelberg 1962, pp. 733 – 740.
196. ISO 2056, Plastics, Determination of the Gas Transmission Rate of Films and Thin Sheets under Atmospheric Pressure, Manometric Method, 1974.
197. DIN 53 380, Bestimmung der Gasdurchlässigkeit, 1969.
198. R. A. Pasternak et al., *J. Polym. Sci. Polym. Phys. Ed.* **8** (1970) 467.
199. D. G. Pye et al., *J. Appl. Polym. Sci.* **20** (1976) 287.
200. ISO 1195, Plastics, Determination of the Water Vapour Transmission Rate of Plastics Films and Thin Sheets, Dish Method, 1970.
201. DIN 53 122, Blatt 1, Bestimmung der Wasserdampfdurchlässigkeit, Gravimetrisches Verfahren, 1974.
202. K.-H. Illers, *Makromol. Chem.* **38** (1968) 168.
203. ISO 62 and Appendix 1, Plastics, Determination of Water Absorption.
204. DIN 53 495, Prüfung von Kunststoffen, Bestimmung der Wasseraufnahme nach Lagerung in kaltem Wasser, 1973.
205. ISO 117, Plastics, Determination of Boiling Water Absorption, 1959.

206. DIN 53 471, Prüfung von Kunststoffen, Bestimmung der Wasseraufnahme nach Lagerung in kochendem Wasser, 1976.
207. DIN 53 473, Prüfung von Kunststoffen, Bestimmung der Wasseraufnahme in feuchter Luft, 1973.
208. ISO 175, Plastics, Determination of the Effecting of Liquid Chemicals, 1981.
209. DIN 53 476, Prüfung von Kunststoffen, Bestimmung des Verhaltens gegen Flüssigkeiten, 1979.
210. B. Dolezel: *Die Beständigkeit von Kunststoffen und Gummi,* Hanser Verlag, München – Wien 1978, pp. 373 – 625.
211. B. Carlowitz: *Kunststoff-Tabellen,* 3rd ed., Hanser Verlag, München – Wien 1986.
212. J. Voigt, Untersuchung der Wärme- und Lichtbeständigkeit in I 101 I, p. 937.
213. DIN 53 386, Prüfung der Wetterbeständigkeit im Naturversuch (Freibewitterung), 1982.
214. ISO 4607, Methods of Exposure to Natural Weathering, 1978.
215. ASTM D 1435-75, Outdoor Weathering of Plastics, 1979.
216. ISO 877, Determination of Resistance to Change upon Exposure under Glass to Daylight, 1976.
217. ISO 105, Textiles, Tests for Color Fastness, 1978.
218. ISO 4892, Plastics, Methods of Exposure to Laboratory Light Sources, 1981.
219. DIN 53 387, Prüfung von Kunststoffen, Kurzprü-fung der Wetterbeständigkeit, 1982.
220. DIN 53 389, Prüfung von Kunststoffen, Kurzprü-fung der Lichtbeständigkeit, 1974.
221. K. Wündrich, unpublished results, 1979.
222. K. Wündrich, *Radiat. Phys. Chem.* **24** (1985) 503.
223. H. Wilski, *Radiat. Phys. Chem.* **29** (1987) 1 – 14.
224. H. Wilski in [167] p. 954.
225. J. Troitzsch: *Brandverhalten von Kunststoffen,* Hanser Verlag, München – Wien 1981.
226. ISO 1210.2, Determination of Flammability of Plastics in the Form of Bars, 1970.
227. DIN 53 438, Teil 1 – 3, Verhalten beim Beflammen mit einem Kleinbrenner, 1977.
228. ASTM D 635-77, Rate of Burning and/or Extent and Time of Burning of Self-Supporting Plastics in a Horizontal Position, 1977.
229. C. P. Fenimore, F. J. Martin, *Mod. Plast.* **43** (1966) 144.
230. VDE 0304, Teil 3/5.70, Bestimmungen für Prüfverfahren zur Beurteilung des thermischen Verhaltens fester Isolierstoffe; Teil 3 Brennverhalten, 1970.
231. VDE 0471, Teil 2, VDE-Bestimmungen für die feuersicherheitliche Prüfung von elektrotechnischen Erzeugnissen, ihren Baugruppen und Teilen; Glühdrahtprüfung, 1975.
232. Underwriter Laboratories: UL 94, 2nd ed. 1973: Tests for flammability of plastics materials for parts in devices and appliances.
233. DIN 4102, Teile 1 – 8, Brandverhalten von Baustoffen und Bauteilen, 1977, 1978.
234. ISO 846, Determination of Behaviour under the Action of Fungi and Bacteria, Evaluation by Visual Examination or Measurement of Change in Mass or Physical Properties, 1978.
235. H. M. Haldenwanger: *Biologische Zerstörung der makromolekularen Werkstoffe,* Springer Verlag, Berlin – Heidelberg – New York 1970.
236. DIN 53 492, Bestimmung der Rieselfähigkeit von körnigen Kunststoffen, 1977.
237. DIN 4188, Drahtsiebböden, 1977.
238. Deutsche Solvay-Werke GmbH, Solvay-Information, *Bestimmung der Korngrößenverteilung von Vinylchlorid (VC)-Polymerisaten,* Rheinberg 1987.
239. DIN 16 945, Reaktionsharze, Reaktionsmittel, Reaktionsharzmassen; Prüfverfahren, 1976.
240. E. P. Weißler in [169]p. 613.
241. DIN 53 735, Bestimmung des Schmelzindex von Thermoplasten, 1983.
242. K.-H. Moos, *Kunststoffe* **75** (1985) 3.
243. A. Göttfert, *Kunststoffe* **76** (1986) 1200.
244. W. Schönthaler: *Verarbeiten härtbarer Kunststoffe,* VDI-Verlag, Düsseldorf 1973.
245. J. F. Kanavec: Inst. f. techn.-ökon. Information, Thema 7 Nr. V 56 – 66 Moskau 1956.
246. P. Ehrentraut, W. Dalhoff, *Kunststoffe* **57** (1967) 439.
247. W. Bauer, K. Eichler, *Ind. Anz.* **90** (1968) 187.
248. G. J. P. Dujardin, *Kunststoffe* **61** (1971) 177.
249. J. Meißner: "Rheologie von Polymer-Schmelzen," vorgetragen beim *Rheologie-Fortbildungskursus* vom 12. – 16. 10. 1981 an der ETH Zürich.
250. G. V. Vinogradov, A. Y. Malkin, *J. Polym. Sci. Polym. Phys. Ed.* **4** (1966) 135.
251. VDMA (ed.): *Kenndaten für die Verarbeitung thermoplastischer Kunststoffe,* Part 2 Rheologie, Part 4 Rheologie 2, Hanser Verlag, München – Wien 1982, 1986.
252. DIN 54 811, Bestimmung des Fließverhaltens von Kunststoffschmelzen mit einem Kapillar-Rheometer, 1984.
253. Z. Rabinowitsch, *Phys. Chem.* **145 A** (1929) 1.

254. K. Geiger, H. Kühnle, *Rheol. Acta* **23** (1984) 355.
255. P. J. Carreau: Ph. D. Thesis, University of Wisconsin, 1968.
256. H. Münstedt, *J. Rheol. (N.Y.)* **23** (1979) 421.
257. J. Meißner, *Chem. Eng. Commun.* **33** (1985).
258. H. M. Laun, H. Münstedt, *Rheol. Acta* **15** (1976) 517.
259. H. M. Laun, in [169] p. 247.
260. F. Ramsteiner, *Kunststoffe* **62** (1972) 766.

Plastics, Recycling

WALTER KAMINSKY, Institut für Technische und Makromolekulare Chemie, Universität Hamburg, Federal Republic of Germany

1. **Introduction** 1
2. **Recycling of Single-Grade Wastes (Primary Recycling)** 3
3. **Recycling of Mixed Wastes** 4
4. **Chemical Products** 6
4.1. **Hydrolysis and Alcoholysis** 6
4.2. **Hydrogenation** 7
4.3. **Pyrolysis** 8
5. **Combustion** 12
6. **Dumping and Degradation** 13
6.1. **Degradation by UV Radiation** 13
6.2. **Biodegradation** 13
7. **References** 15

1. Introduction

In recent decades, plastics have undergone a spectacular rate of growth. Whereas, initially, they were mostly regarded as substitutes for other materials, they have now been developed into high-grade products. They are used in all branches of industry and everyday life and represent useful modern technical materials with versatile and, to some extent, irreplaceable applications [1–3]. Because of their high resistance to physical and chemical attack and their very light weight, plastics have made possible many new developments, for example, in the clothing and packaging sectors, the manufacture of machinery and motor vehicles, in agriculture, electronics, and communication technology. New high-strength composite materials find application in air- and space travel, in the construction of motor vehicles, and in sport und recreation. The use of plastics as a proportion of total material consumed will increase further in the future because plastics can be produced to fulfill almost any user requirement. They also have lower energy costs per unit volume of production (Fig. 1) and lower energy costs for the production of nonreturnable containers (Table 1) [4–6].

World production of plastics in 1990 amounted to ca. 100×10^6 t, in 2000 to ca. 170×10^6 t, and an annual increase of 7 % is expected [98]. Inevitably, a continually increasing quantity of used plastics is arising, and the shorter the lifetime of the products, the greater the amount of waste. At present, ca. 35 % of the plastics produced in the industrial countries of the West occurs in waste form [7]. The length of time in use of plastics breaks down as follows:

< 1	year	20 %
1 – 8	years	15 %
> 8	years	65 %

i.e., most plastics are used in products with a long service life. Many of these products have not yet appeared as waste (e.g., plastics used in the building industry and plastic pipes for water, wastewater, and gas).

A large proportion of scrap plastics (40 %) is mixed up with household garbage. Figure 2 shows the composition of household garbage in the Federal Republic of Germany [8]. The plastics content is ca. 5.4 wt %, which corresponds to ca. 2.5×10^6 t/a of plastic waste.

Approximately 0.5×10^6 t of this is directly recycled, 0.7×10^6 t is disposed of in combustion plants, and 1.3×10^6 t is dumped. The plastic waste consists of:

60 – 65 wt %	polyolefins (polyethylene, PE; polypropylene, PP)
10 – 20 wt %	polystyrene (PS)
12 – 17 wt %	poly(vinyl chloride) (PVC)
10 – 15 wt %	other plastics
3 – 10 wt %	impurities

In future, the amount of plastic waste is expected to increase and slowly approach the quantity of plastics manufactured.

The numerous applications of plastics have also clearly revealed the necessity for a sensible disposal procedure for scrap materials. Nonreturnable packaging material is often disposed of carelessly which results in a conspicuous litter problem, especially in the countryside. Further-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_057

Table 1. Comparison of energy requirements for the manufacture of nonreturnable containers [6]

Container	Glass		Plastic			
	Weight, g	Energy requirement, kJ	Weight, g		Energy requirement, kJ	Energy saved by use of plastic, %
Milk bottle (1 L)	370	3850	(PE)	20	1600	58
Cooking oil bottle (1 L)	390	4050	(PVC)	28	1770	56
Yoghurt container (180 g)	112	1160	(PS)	7.5	630	45

more, most polymers are manufactured from petroleum, which is available only in limited supply, so that efforts are being made to minimize polymer wastes or to provide a suitable means of disposal or reuse. Methods include regranulation, alcoholysis, hydrolysis, pyrolysis, use of degradable plastics, combustion, and well-managed dumping sites [9–19].

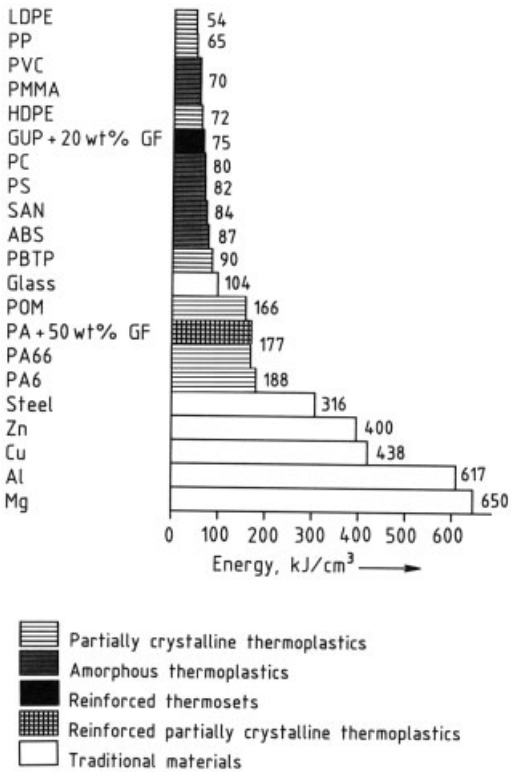

Figure 1. Energy requirements per unit for the manufacture and processing of various materials [4]
GUP = glass fiber reinforced unsaturated polyester resin

The object of recycling is to decrease the amount of waste by recirculation of raw materials and to improve material utilization. Recycling is an established procedure for scrap metals, glass, and paper.

Recycling is not, however, always worthwhile [20]. It is only worthwhile if the emissions, energy consumption, and environmental pollution caused by collection and reprocessing are less than those resulting from the manufacture of new products from raw materials and the dumping or combustion of the wastes. Hence 100 % recycling of plastic wastes is not realizable, because, e.g., in the case of finely divided, mixed, or strongly bonded materials, the cost of collection and separation is too high.

Waste economy programs have the following objectives:

1) Reduction of waste levels by producer and consumer
2) Increased recycling in manufacturing and processing
3) Safe disposal of wastes
4) Allocation of costs in accordance with the principle that "the polluter pays"

Together with more economical use of raw materials and energy as well as the improvement and optimization of plants for the production of monomers and polymers, this has already lessened the ecological problems caused by production and disposal of plastics.

2. Recycling of Single-Grade Wastes (Primary Recycling)

To optimize a recycling operation, it is best to reuse the plastic waste directly after separation and cleaning if possible [21–23]. Only then can the high energy input required for synthesis be retained. Single-grade wastes arising in industries producing and processing plastics have long been recycled internally for economic reasons.

Recycling of single-grade wastes has now reached a high level of technical and economic

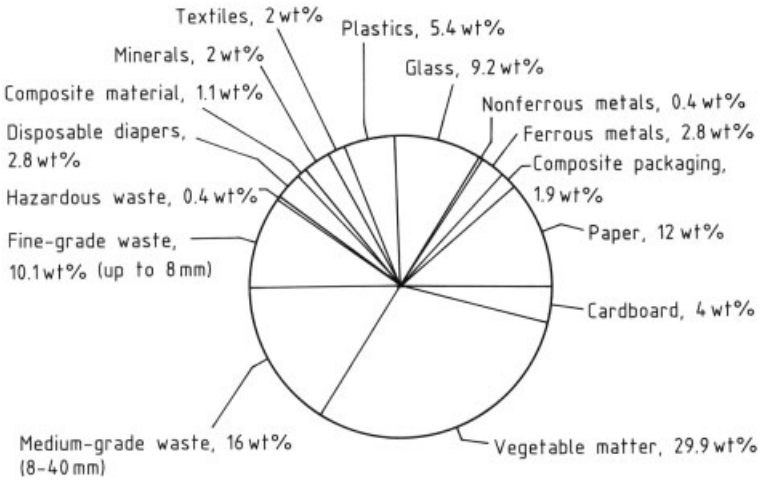

Figure 2. Composition of domestic waste in the Federal Republic of Germany in 1985 (total quantity 45×10^6 t) [8]

performance. Clean regenerated material commands a price of up to 70 % of that of new material. This allows greater expenditure on processing measures, which include size reduction, washing, drying, metal separation, sorting, compounding, restabilization, homogenizing, degassing, and granulation.

Only thermoplastic waste can be reprocessed directly. It can be used alone or, more often, added to virgin polymer in various proportions [24–26]. The main technical problems encountered in primary recycling are:

1) Degradation of the material due to repeated processing, resulting in deterioration of properties such as appearance, chemical resistance, processability, and mechanical characteristics
2) Contamination of the reprocessed plastic
3) The handling of scrap with a low bulk density, for example, thin films, foam, or small plastic articles such as cups, tubes, and pens

Degradation of the material occurs in all stages of recycling, but it is usually possible to complete four to five full cycles without needing to add new material [27]. Plastics vary in their sensitivity to degradation during processing. Oxidation causes breakdown of high- and low-density polyethylene, which results in reduction of the melt flow index. Recycling of polypropylene increases the melt flow index and decreases the impact strength. The recycled material is also sensitive to contamination by metals, metal compounds, pigments, and fillers, which promotes degradation that is characterized by an increase in brittleness, dark streaks, and odor. Polystyrene is sensitive to contamination during reprocessing, which causes deterioration of some physical properties and a color shift.

Recycling of PVC is simple and is practiced widely [28–31]. Here, the stabilizers become depleted and the material begins to yellow and emit odors. Reprocessing of acrylics causes a color shift but, for less critical applications, up to 100 % reprocessed material may be used. Polyamides can easily be reprocessed, although they are sensitive to contamination and tend to brown. The properties are generally maintained over a wide range of processing pressures and temperatures.

Clean reground acrylonitrile – butadiene – styrene copolymer can be mixed with virgin material in any proportion. Degradation of the recycled terpolymer results in deterioration of the physical properties, particularly the impact strength, and darkening of the material.

To reprocess waste plastics, they must be ground to a particle size close to that of the virgin polymer. Size reduction is usually carried out using a granulator. Various granulators are used, depending on the size of the pieces, the required throughput, and the shape of the plastic waste (e.g., sprues, sheet, pipe, film). A granulator consists of a hopper, a cutting chamber with knives, a screen, and a drive [32]. The hopper is designed to handle specific material shapes. Granulators may be equipped with a feeding de-

vice, such as an anger feeder or counterrotating rollers for films. Most hoppers have bends, baffles, or doors to prevent fly-back of the plastic pieces.

In some cases, single-grade plastic waste from the plastics processing industries can be reused directly, together with new material, without the need for further treatment. The same extruders or injection-molding machines can be used as for the processing of new material.

Some plastic wastes are difficult to grind with standard granulators because of their physical properties (e.g., rubbery or low-melting materials, and those of low bulk-density such as foam, films, and fibers). Such materials must be granulated by cryogenic grinding or in a plast agglomerator [33].

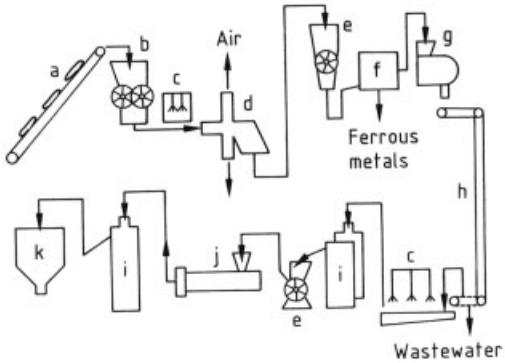

Figure 3. Flow sheet for separation of garbage (Funabashi, Japan)
a) Conveyor; b) Shredder; c) Washer; d) Air classifier; e) Mill; f) Magnetic separation; g) Rotary scrubber; h) Belt screen; i) Centrifugal dryer; j) Pelletizing extruder; k) Silo

3. Recycling of Mixed Wastes

The processing of mixed, soiled, or composite plastic waste is more difficult than the recycling of unmixed waste [34]. Waste material generated by industrial, craft, and trade industries contains a wide mixture of different plastics. The number of different materials present and the degree of soiling is most pronounced in household garbage and in the plastics fraction resulting from the scrapping of automobiles. These contain a variety of plastics used for packaging and trim including polyethylene, PVC, polyesters, polymer blends, foamed polystyrene, polyurethanes, and thermosets [35–37].

Foamed polystyrene and other foam packaging materials can be sorted relatively easily, but processed and colored PVC, polyethylene, polypropylene, polyester, or polystyrenes can only be recognized and distinguished from one another by specialists. Many plastics also contain up to 50 wt % fillers, plasticizers, stabilizers, and other additives which may alter the physical properties such as density, transparency, and flexibility considerably.

Therefore, few of these waste materials are used and even then, only when they are available in sufficiently large quantities and with a constant composition. In this way, plastic shrink packaging, different varieties of sack materials, imitation leather wastes, and bottle and transport containers enter the material recycling system.

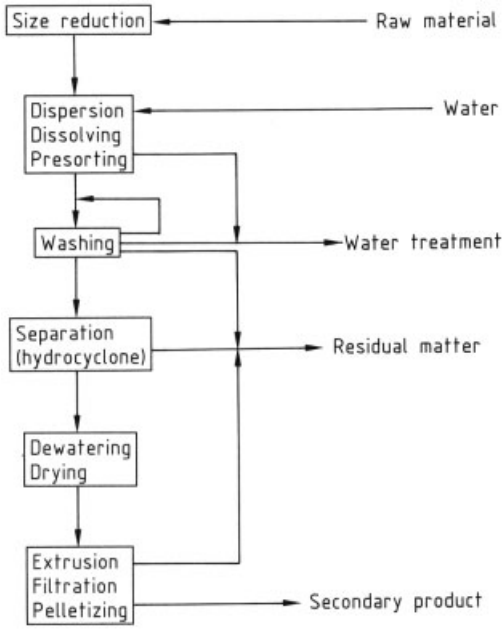

Figure 4. Scheme for processing and separating mixed plastic wastes [9]

In principle, reprocessing of mixed plastic wastes by melting them together is possible for thermoplastics. However, this is only carried out occasionally because different types of plastics are usually incompatible, i.e., not miscible at the boundary region of the polymer phases (see Table 2) [38]. Because of the resulting low strength mixed polymers are mostly used for

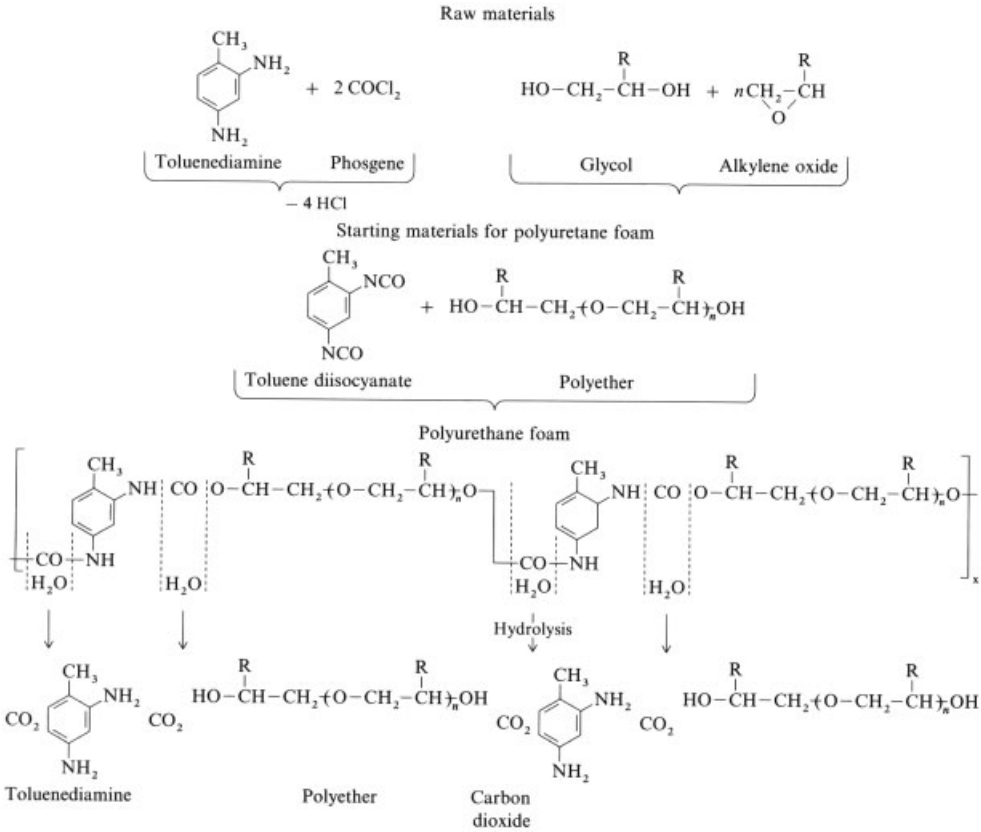

Figure 5. Hydrolysis of polyurethane foam [48]

thick-walled products or slabs which must compete with more favorable concrete or wood products and can only be marketed in special cases [39].

Plastic wastes from segregated collection are reduced in size, cleaned, washed, and pelletized as shown in Figure 3 [40].

Short-screw extruders or roll extruders are used for processing. Roll extruders are more flexible in operation because the gap width is adjustable. Wastepaper baskets, plant pots, garden fencing, fence posts, and park benches can be produced in this way [41].

Mitsubishi Petrochemical has developed a special extruder screw, known as a reverser, in which mixed plastic scrap is melted and pressed into moldings with diameters up to 15 cm. For a throughput of 500 kg/h, the power requirements are 75 kW for the motor and 54 kW for heating [42].

Higher-quality products can be produced if the plastic waste is mechanically sorted. Wet and dry procedures exist. Wet procedures can be classified as float – sink processes either under gravity or under centrifugal force (hydrocyclone). Hydrocyclones are commonly used for the separation of polyolefins (especially polyethylenes) from other plastics [43]. Dry processes are separation in cyclones and air classification. Figure 4 shows the processing sequence for dealing with films and hollow moldings made of polyethylene, cable waste from production and installation, soft-drink bottles made of poly(ethylene terephthalate), and battery scrap made of polypropylene.

If the densities of the polymers differ sufficiently, fraction purities of up to 99 % can be attained with hydrocyclones [43]. However, even in the case of polyethylene, there are more than one hundred different grades which will now be

Table 2. Miscibility of thermoplastics ∗ [38]

	Standard polystyrene	Impact-resistant polystyrene	Styrene – acrylonitrile copolymer	ABS	Polyamide	Polycarbonate	Poly(methyl methacrylate)	Polyformalde-hyde	Poly(vinyl chloride)	Polypropylene	Low-density polyethylene	High-density polyethylene	Poly(butylene terephthal-ate)
Shock - resistant polystyrene	1												
Styrene – acrylonitrile copolymer	6	6											
Acrylonitrile – butadiene – styrene copolymer (ABS)	6	6	1										
Polyamide	5	4	6	6									
Polycarbonate	6	5	2	2	6								
Poly(methyl methacrylate)	4	4	1	1	6	1							
Polyformaldehyde	6	6	6	5	6	6	5						
Poly(vinyl chloride)	6	6	2	3	6	5	1	6					
Polypropylene	6	6	6	6	6	6	6	6	6				
Low-density polyethylene	6	6	6	6	6	6	6	6	6	6			
High-density polyethylene	6	6	6	6	6	6	6	6	6	6	1		
Poly(butylene terephthalate)	6	6	6	5	5	1	6	6	6	6	6	6	
Poly(ethylene terephthalate)	5	5	6	5	5	1	6	6	6	6	6	6	6

∗ Degree of miscibility: 1 = very good; 6 = incompatible.

mixed together and thus alter the physical properties adversely.

After separation the shredded and washed plastic must be dewatered and dried to $< 0.4\,\%$ of residual water. The material is then temporarily stored and pelletized. Inclusions such as water and air are removed by vacuum degassing. A filter located at the end of the extruder removes residual impurities from the melt. In many extruders, which have throughput of up to 1000 kg/h, the filter is exchanged automatically. The pelletized product is stored, homogenized, and packed using pneumatic equipment. Between 5 and 10 L of water are required per kilogram of regranulated material [44–47].

PVC wastes can be separated due to their high density ($1.119 - 1.39\,g/cm^3$) and can be processed into e.g., floor tiles or used to produce drainage and irrigation pipes.

4. Chemical Products

Plastic wastes may contain so many varieties, be so badly contaminated, or have such a modified molecular structure due to degradation that direct recycling is not worthwhile. Such wastes include composite materials, filled plastics, as well as thermosets and elastomers, which cannot be remelted. Recycling of these materials can only be achieved by breaking down the macromolecules into smaller fragments by hydrolysis, alcoholysis, aminolysis, hydrogenation, or pyrolysis.

4.1. Hydrolysis and Alcoholysis

Hydrolysis leads to direct recovery of the original raw materials by targeted reaction of water molecules at the linkage points of the starting materials (thus reversing the polymerization process). All hydrolyzable plastics such as polyamides, polyesters, polycarbonates, polyureas, and polyurethanes are resistant to hydrolysis under normal conditions of use. It is therefore necessary to use extreme conditions for hydrolysis [48–51].

Hydrolysis of polyurethane foams is particularly interesting since they have a very low density (30 kg/m^3) and thus take up considerable storage space. Furthermore, valuable products can be recovered by hydrolysis (Fig. 5) [48].

Figure 6 shows a continuous hydrolysis process using a specially equipped twin-screw extruder in which the operating pressure of 15–30 bar is maintained by the high-pressure steam generated at the reaction temperature of 200–230 °C. The shredded polyurethane foam is hydrolyzed back into the basic raw materials of the foam (polyethers and amines). Depending on the throughput rate, the residence time of the material varies between 5 and 30 min.

Product yields are outstanding: almost 100 % of the polyether and ca. 90 % of the amine can be recovered. The regenerated materials can be reused directly, together with fresh starting material, for the manufacture of the same foam material. Raw materials in the waste can thus be fed back again to the same production process and the environment is thus not burdened by that quantity of waste material.

Alcoholysis. Chemical degradation of polyurethanes can also be achieved by alcoholysis to give a polyhydroxy alcohol and small urethane fragments formed by transesterification (Scheme 1) [52, 53]. Carbon dioxide is not formed in this reaction. If a diol is used as the alcohol, then the urethane fragments also contain terminal hydroxyl groups. These polyhydroxy alcohols can be converted directly to polyurethane foam following the addition of isocyanates and varying proportions of new polyhydroxy alcohols.

Polyesters and polyamides can also be alcoholized in the same manner as polyurethanes. Polyamides break down to give amino end groups instead of hydroxyl groups. Since all these reactions are equilibrium reactions, complete monomerization can only be achieved with a large excess of diol: oligomers are formed when only a slight excess of glycol is used.

Since mass-produced plastics such as polyethylene, polypropylene, polystyrene, and diene elastomers cannot by hydrolyzed, hydrogenation or pyrolysis must be employed.

4.2. Hydrogenation

In hydrogenation the C–C bonds of polyolefins and polystyrene are broken by the addition of hydrogen. Alkanes are by far the largest group of products resulting from this mild decomposition process [54]. Since other bonds (e.g., ester bonds) can also be cleaved by hydrogenation, most types of plastic can be degraded in this way and therefore the plastic waste does not require sorting. Since the reaction proceeds very slowly, hydrogenation is usually carried out at elevated temperature (500 °C) and pressure (400 bar) in the presence of cobalt–molybdenum catalysts. Continuous or batch high-pressure operation and the cost of hydrogen increases processing costs. Hydrogenation does, however, produce high yields (60–90 %) of liquid hydrocarbons. Hydrogenation of PVC-containing wastes splits off HCl, and the product contains ca. 20 ppm of organic chlorine.

A pilot plant for hydrogenation of waste plastics has been built by the Union Rheinische Braunkohlen-Kraftstoff AG in Wesseling, Germany.

4.3. Pyrolysis

Pyrolysis is thermal cleavage in complete or partial absence of air, accompanied by simultaneous generation of pyrolysis oils and gases suitable for chemical utilization or energy generation.

Thermoplastics, thermosets, elastomers, composite materials, and heavily soiled or damaged plastics can be recycled by pyrolysis. The wide range of input materials results in a broad spectrum of different decomposition products which are not always easy to separate. Therefore, the products are usually less economically attractive than those from hydrolysis.

The advantage of pyrolysis over combustion is the 5- to 20-fold reduction in the volume of product gases which leads to considerable savings in gas purification. Furthermore, the pollutants are concentrated in a cokelike residue which acts as a matrix. In addition, it is possible to obtain hydrocarbons, and some processes even recover valuable crude chemicals.

Pyrolysis is complicated by the fact that plastics, rubbers, and biopolymers exhibit poor thermal conductivity and the degradation of

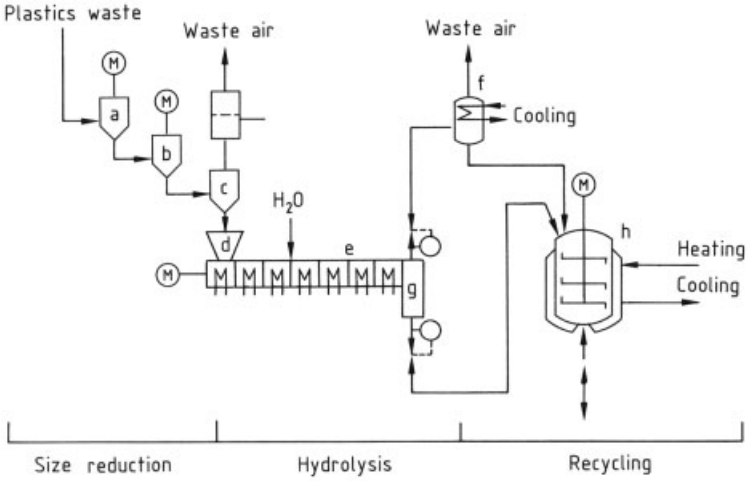

Figure 6. Continuous hydrolysis apparatus
a) – c) Stirred hoppers to remove air from foam (evacuation); d) Feeder; e) Twin-screw extruder, heated and cooled in different segments; f) Cooler for low-boiling products; g) Separator for product gas and liquids; h) Crystallizer for separation of diamines and polyesters

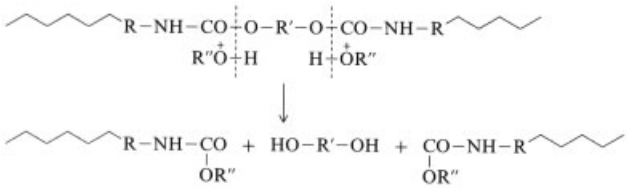

Scheme 1. Alcoholysis of polyurethanes

macromolecules requires large amounts of energy. Pyrolysis of plastic wastes, used tires, and biopolymers has been studied in melting vessels, blast furnaces, autoclaves, tube reactors, rotary kilns, coking chambers, and fluidized-bed reactors [55–64]. Table 3 gives a summary of processes for the pyrolysis of plastics. Some processes pyrolyze plastics combined with household garbage.

Rotary kiln processes are widely used for pyrolysis [58–62]. The wastes have relatively long residence times in rotary kilns (≥ 20 minutes), whereas residence times in fluidized-bed reactors rarely exceed a few seconds, with a maximum of 1.5 min. Due to the large temperature gradient in the rotary kiln, the product spectrum is very diverse. For this reason, the gases and oils resulting from the process are normally used for the direct generation of energy (heat).

In fluidized-bed reactors, the fluidized bed is generated by a flow of air or an inert gas directed from below through a layer of fine-grained material such as sand or carbon black, the flow rate being sufficient to create a swirl in the bed. In this state, the fluidized bed behaves as a liquid.

The advantages of fluidized beds for pyrolysis of plastics are excellent heat and mass transfer as well as constant temperature throughout the reactor. This results in largely uniform product spectra, the absence of moving parts in the hot zone of the reactor, and a completely closed system, i.e., the reactor can easily be sealed.

In Japan, fluidized-bed pyrolysis plants are operated with air or oxygen as fluidizing gas [59]. Partial oxidation generates part of the necessary fission energy and part of the product is burnt. The product oils are partially oxidized and their energy content is ca. 10 % below that of pure hydrocarbons. Tsukishima in Japan and the researchers at West Virginia University use two fluidized beds, one for combustion and the other for pyrolysis [24, 64].

In recent years a number of projects were conducted at the Universität Hamburg to determine whether plastic wastes, used tires, and waste oil

Table 3. Processes for the pyrolysis of plastics

Process/developer	Procedure	Products	Development stage	Reference
Mitsubishi	melting vessel, indirectly heated, 500 °C	oil, some gas	100 kg/h, Tokyo, Japan	[60]
Union Carbide	extruder barrel, indirectly heated, 420 – 600 °C	waxes	pilot plant, USA	[24]
Kobe Steel	rotary kiln, indirectly heated, 500 – 700 °C	oil, gas, energy	1 t/h, Kobe, Japan	[61]
Vöest-Alpine ∗	shaft kiln, directly heated	gas, steam	650 kg/h, pilot plant, Linz, Austria	[62]
KWU process ∗	rotary kiln, indirectly heated, 450 – 500 °C	energy	3 t/h, plant, Ulm, Germany	[58]
Noell, Dr. Otto process	rotary kiln, indirectly heated, 650 – 700 °C	pyrolysis oils, gas	6 t/h, plant, Salzgitter, Germany	[58]
Waterloo University	short-dwell pyrolysis, 500 – 700 °C	oil, gas	laboratory plant, Waterloo, Canada	[63]
West Virginia University ∗	two fluidized beds, one of which is oxidizing	gas	laboratory plant, West Virginia University, USA	[24]
Tsukishima, Ebara ∗	two fluidized beds, one of which is oxidizing	energy	4 t/h, Yokohama, Japan	[64]
Hamburg process	fluidized bed, indirectly heated, 600 – 900 °C	pyrolysis oils, gas, carbon black	20 – 60 kg/h, University of Hamburg; 0.5 t/h, ABB plant, Ebenhausen/Ingolstadt, Germany	[59]

∗ Handles plastics mixed with household garbage.

residues would be suitable as a source of raw materials when pyrolyzed in an indirectly heated fluidized-bed reactor [65–68]. Laboratory-scale plants with capacities of 60 – 3000 g/h and two pilot plants with capacities of 10 – 40 kg/h of plastic and 120 kg/h of used tires were installed. Figure 7 shows the scheme of a pilot plant for the pyrolysis of plastics. The core of the plant is a fluidized-bed reactor (c) with an inner diameter of 450 mm. A fluidized bed of quartz sand with a temperature of 600 – 900 °C is used for the pyrolysis of plastics, which are fed into the reactor through a double-flap gate or a screw. Pyrolysis gas, preheated to 400 °C, is used to fluidize the bed. Heat input takes place indirectly through jacketed jet pipes, which are heated by pyrolysis gas.

The exhaust gases are then directed through a heat exchanger. The product gases emerging from the fluidized bed are separated from residual carbon and fine dust in a cyclone (d) and then cooled to room temperature.

The condensed oil fractions are distilled in two distillation columns (j) using the bottom product of the second distillation column (mostly xylenes) as a quenching medium. High-boiling tar is also produced as well as two fractions, one rich in toluene and the other in benzene.

The gas, largely stripped of liquid products, is compressed in five membrane compressors (r), connected in parallel, and stored in gas tanks. Part of the gas is used as fuel for the radiation heating tubes while the remainder, preheated by hot fuel gases in the heat exchanger (p), is used for fluidizing the sand bed. The excess gas is burnt in a flare (s).

Up to 50 % of the input material can be recovered in liquid form, which corresponds to a mixture of gasoline and bituminous coal tar with about 95 % aromatics. The oil can be processed into chemical products by the usual petrochemical methods. Reaction management aims to maximize the yield of aromatics. Comparative raw material value analysis showed that the chemical processing of high-aromatic oil to give raw materials is more favorable economically than direct use as a fuel.

The product gases are high-energy fuel gases with a calorific value of ca. 50 MJ/m^3 (STP). Only about 40 vol % is required as fuel for the jacketed jet pipes.

Plastic waste can be pyrolyzed even if it is soiled. Table 4 summarizes the various types of plastic and rubber which have been used as input material.

The main gaseous compounds are methane, ethane, ethylene, and propene; the oil mainly consists of benzene, toluene, and naphthalene (Table 5).

It is generally difficult to recover monomers by pyrolysis, and ethylene and propene yields from polyolefins do not exceed 60 wt %. The pyrolysis of polystyrene gives ca. 64 wt % of the styrene monomer. The liquid must be purified by an expensive process to give polymerization-grade styrene monomer. In contrast fluidized-bed pyrolysis of poly(methyl methacrylate) produces 97 wt % of methyl methacrylate [69, 70].

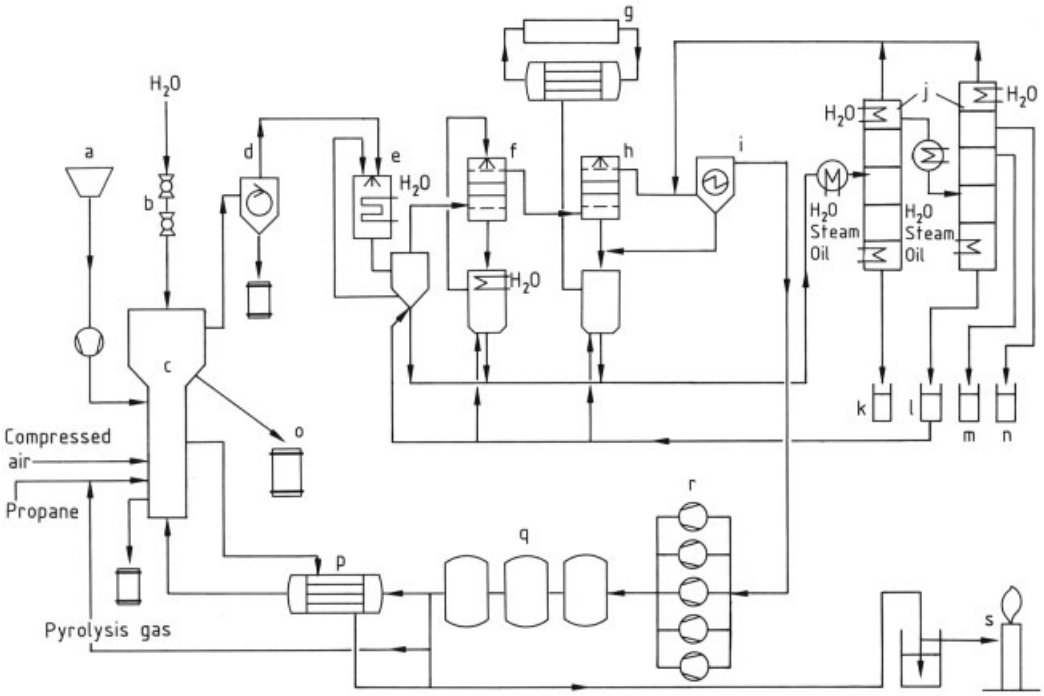

Figure 7. Schematic of a fluidized-bed process for pyrolysis of polymers (throughput 20 – 60 kg/h)
a) Supply silo; b) Sluice; c) Fluidized-bed reactor; d) Cyclone; e) Quencher; f) Washer (room temperature); g) Cryostat; h) Washer (low temperature); i) Electrostatic precipitator; j) Distillation columns; k) High-boiling fraction; l) Xylene fraction; m) Toluene fraction; n) Benzene fraction; o) Sand overflow; p) Heat exchanger; q) Gasometer; r) Compressors; s) Flare

Table 4. Products from pyrolysis of plastic waste in a fluidized bed

Feed	Pyrolysis temperature, °C	Gas, wt %	Oil, wt %	Residue, wt %	Others, wt %
Polyethylene (PE)	760	55.8	42.4	1.8 C	
Polypropylene (PP)	740	49.6	48.8	1.6 C	
Polystyrene (PS)	580	9.9	24.6	0.6	64.9 styrene
Mixture PE – PP – PS	750	52.0	46.6	1.4	
Polyester (PETP)	768	50.8	40.0	7.1	2.1 H_2O
Polyurethane (PUR)	760	37.9	56.3	0.5	5.0 H_2O
					0.3 HCN
Polyamide (PA-G)	760	39.2	56.8	0.6	3.4 HCN
Polycarbonate (PC)	710	26.5	46.4	24.6	2.5 H_2O
Poly(methyl methacrylate) (PMMA)	450	1.25	1.4	0.15 C	97.2 MMA
Poly(vinyl chloride) (PVC)	740	6.8	28.1	8.8	56.3 HCl
Polytetrafluoroethylene (PTFE)	760	89.3	10.4	0.3	
Medical syringes	720	56.3	36.4	5.8	1.5 steel
Plastic from household separation	787	43.6	26.4	25.4	4.6 H_2O
Plastic from car shredding	733	29.9	26.7	27.6	14.0 metals
					1.8 H_2O
Ethylene – propylene – diene rubber (EPDM rubber)	700	32.3	19.2	47.5	1.0 H_2O
Styrene – butadiene rubber (SB rubber)	740	25.1	31.9	42.8	0.2 H_2S
Scrap tires	700	22.4	27.1	39.0	11.5 steel
Lignin	500	3.4	29.9	49.3	17.4 H_2O
Cellulose (wood)	700	47.1	23.0	18.6 C	11.3 H_2O
Sewage sludge	600	34.3	27.7	33.2	4.8 H_2O

Pyrolysis of poly(vinyl chloride) produces hydrogen chloride (56 wt %) and carbon black, while the composition of the other pyrolysis products is similar to those obtained from polyethylene. Carbon dioxide and carbon monoxide are formed from polyesters, polyurethanes, polyamides, and cellulose-containing materials. The plastic fraction from the separation of household garbage contains up to 70 % of plastics, of which about 15 % is PVC.

Table 5. Components from the pyrolysis of different plastic wastes in a fluidized-bed process

Pyrolysis products	Feed			
	Polyethylene, wt % (740 °C)	Plastics from household separation, wt % (787 °C)	Polyester, wt % (768 °C)	Polyurethanes, wt % (760 °C)
Hydrogen	0.75	0.70	0.30	0.23
Methane	23.6	17.5	3.8	3.5
Ethane	6.7	2.9	0.24	0.76
Ethylene	19.8	9.8	1.5	5.4
Propane	0.08	0.08	0.01	0.01
Propene	5.5	1.3	0.08	1.7
Butene	0.54	0.13	0.01	
Butadiene	1.6	0.44	0.04	1.2
CO_2		2.3	17.2	14.4
CO		1.2	27.4	8.7
HCN				0.31
HCl ($CaCl_2$)		7.2		
Hexene	0.02	0.002	0.02	0.2
Benzene	19.1	13.8	18.3	5.0
Toluene	3.9	5.6	2.5	1.8
Styrene	0.5	3.1	2.1	0.01
Indene	0.25	1.0	0.41	1.1
Naphthalene	2.8	3.8	1.4	1.5
Methylnaphthalene	0.63	0.84	0.25	0.69
Diphenyl	0.30	0.41	1.47	0.15
Fluorene	0.20	0.22	0.18	0.27
Phenanthrene/anthracene	0.52	0.66	0.08	0.62
Phenylnaphthalene	0.08	0.36	0.21	0.01
Fluoranthene	0.08	0.26	0.02	0.01
Acetonitrile				0.58
Acrylonitrile				0.95
Benzonitrile				1.8
Ketones, acetales			0.15	9.2
Acetophenone			1.4	0.3
Other compounds	11.3	9.0	11.7	34.1
Carbon black	1.75	6.4	7.1	0.5
Water		2.6	2.1	5.0
Inorganic compounds		5.3		

Addition of lime to the fluidized bed enables the hydrogen chloride from the decomposition of PVC to be bound chemically. However, the calcium chloride that is formed in the process tends to clog the fluidized bed at higher concentrations. Therefore, attempts are being made to eliminate HCl gas in a preliminary step at 300 – 400 °C before more extensive cracking reactions take place. The dry HCl gas would then be available for further use.

The residual content of organic chlorine is important for the value of the pyrolysis oils. For processing in existing petrochemical plants, the chlorine content should not exceed 10 ppm. Oils obtained from the pyrolysis of mixed plastics contain 50 – 200 ppm of organically bound chlorine. Chlorinated dibenzodioxines were not found amongst these chloroorganic compounds [56]. Further dehalogenation of the pyrolysis oil can be achieved by adding sodium vapor to the pyrolysis gas stream.

The process has been successful for the pyrolysis of scrap tires and rubber. Even whole tires with a maximum weight of 20 kg were completely pyrolyzed within 1.5 – 4 min at temperatures above 650 °C. The amount of pyrolysis gas produced in the process was sufficient to heat the reactor without an additional energy supply [66]. The pyrolysis products consisted of the following five fractions (wt %):

Gas		15 – 20
Oil		20 – 30
Water		5 – 10
Carbon black +	filling materials	30 – 40
Steel cord		5 – 20

The main components, apart from the solids (carbon black and steel cord), are methane, ethylene, ethane, benzene, and toluene.

Although ca. 2 wt % of sulfur is incorporated into the rubber, the hydrogen sulfide content was < 0.3 wt %. After pyrolysis, most of the sulfur is found in the carbon black fraction as zinc or calcium sulfide. During pyrolysis, reaction takes place between the sulfur-containing fission products and the alkaline filling materials. The car-bon black that is obtained is suitable for recycling as a filling material.

Several pyrolysis plants have been tested over a period of years. Oil and gas are the most valuable materials produced so that the profitability of the process depends largely on the price of petroleum. Up to now, pyrolysis has not been an economical process. This could, however, change rapidly due to increased oil prices and greater recognition of the need to recycle whenever possible.

5. Combustion

If it is not practical or economic to reuse plastic waste, it is still possible to make use of its energy content. This means the loss of the energy that was required to manufacture the plastic and also of the hydrocarbon structure. However, the energy content (calorific value) of plastics is higher than that of most other materials. The calorific value of 1 kg of plastic waste is roughly that of 1 L of fuel oil and higher than that of coal. The plastic waste in household garbage contributes up to 30 % of its calorific value, which allows garbage-based power stations to produce electricity, steam, or remote heating [7–73].

As long as ca. 94 % of oil products are used for heating purposes (only ca. 6 % is processed in the chemical industry), there would be no objections to this form of energy recycling since it allows the substitution of fuel oil and reduces the volume of garbage. However, halogen-containing plastics such as poly(vinyl chloride), poly(vinylidene chloride) and polytetrafluoroethylene create problems as they release substantial volumes of hydrogen chloride or fluoride on combustion. Nitrogen-containing plastics (e.g., polyamides, polynitriles, polyurethanes) increase NO_x emissions. Furthermore, combustion releases metals and other organic compounds from dyes, pigments, stabilizers, and plasticizers present in the plastics. Halogen-containing plastics, in particular, are also suspected of contributing to the formation of highly toxic dioxins.

The use of modern but elaborate and expensive flue-gas washing facilities makes it possible to reduce the level of these harmful substances to the legal limits [74]. In particular, combustion of household garbage containing PVC produces HCl in the flue-gases which requires considerable expenditure for neutralization, anticorrosion treatment, and disposal. The possibility of using acidic neutralization water (hydrochloric acid) for the extraction of sludges containing heavy metals is being discussed. Following neutralization, the salt (sodium chloride) formed can be used in chlor-alkali electrolysis [75].

The desire for material recycling and the costs associated with the combustion of some plastics wastes are, however, reinforcing efforts to reuse plastics fractions in a targeted manner.

6. Dumping and Degradation

In most countries, waste material is generally dumped; in Germany this amounts to 70 % of the total. Large quantities of plastic waste are dumped with household garbage. Since plastics are usually insoluble and harmful substances are not washed out, they do not cause any increased environmental burden in a well-managed dumping site. As the proportion of plastics in household garbage increases, however, it will become necessary to precompact the garbage both to reduce its elasticity and slip potential and also to use the increasingly scarce dumping sites more efficiently. However, the plastic waste is unlikely to decompose under dump site conditions and will persist over a long period. Dumping also makes no use of either the material value or the energy content of plastics produced from petroleum.

Because of the highly visible litter problem caused by, e.g., plastic bags, bottles, and containers and the need to reduce space demands in dumping sites, it is often demanded that plastics should be degradable. Plastics which are biodegradable or are degraded by UV radiation have been developed and are in use for special applications in medicine, cosmetics, and agriculture. Polymers which are soluble in water also exist.

Degradation of plastics to give materials of lower molecular mass can be effected by heat, light, and both chemical and biological processes [76–80]. One possibility is to manufacture plastics which, after use, are degraded by the influence of light or microorganisms. However, the degradation products may have a greater adverse impact on the environment than the plastics in their original form. Although materials that break down in this way would lessen the visible aspect of litter, the degradation products could penetrate into the groundwater. In the case of microbial degradation, it would be difficult to ensure that long-life products (e.g., drain pipes, plastic components used in construction) would not also be attacked.

6.1. Degradation by UV Radiation

Stabilizers are usually added to plastics to protect them against the effect of light. To make

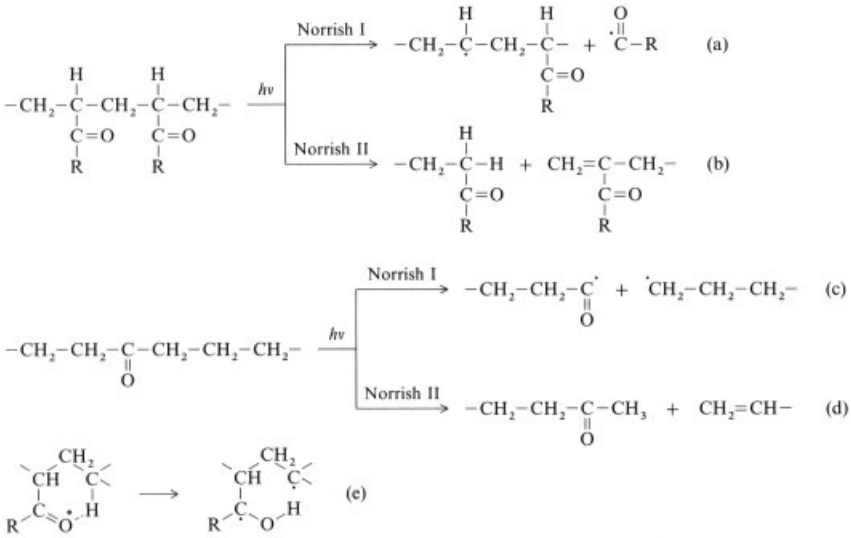

Figure 8. Photochemical reactions in polymers with ketone groups in the side chain (a), (b); in the main chain (c), (d); and radical mechanism (e)

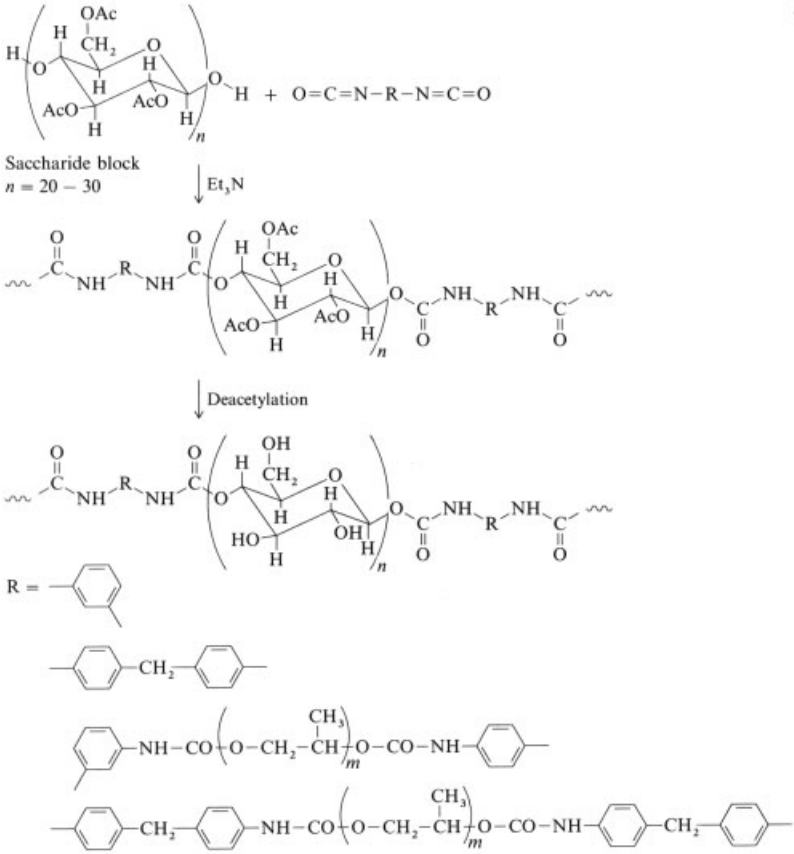

Figure 9. Synthesis of biodegradable block copolymers containing cellulose units

them more sensitive to light, UV-absorbing monomers can be added. This is best achieved by incorporation of carbonyl- containing materials, the chains of which break down under the influence of light to give lower molecular mass compounds [81, 82]. A simple example of this is the copolymerization of ethylene and carbon monoxide (1 – 5 wt %) [83]. Incorporation of carbon monoxide into the polymer chain facilitates radical degradation. Figure 8 shows the reaction mechanism for this type of degradation.

The reaction can proceed by two different mechanisms (Norrish I and Norrish II). In the case of Norrish I, which is favored, degradation results primarily from chain scission next to the carbonyl group to release two radicals, whereas the Norrish II reaction gives a saturated ketone and a product with a terminal vinyl group.

Other possibilities include adding sensitizers to the polymerization mixture. The advantage of the copolymerization technique is that the polymers do not contain any extractable, potentially toxic components. Additives which have been suggested include benzophenone, anthraquinone, diketone, and carboxyaldehyde derivatives [85] and iron, cobalt, manganese, or copper salts [86].

6.2. Biodegradation

With the exception of some cellulose derivatives and aliphatic polyesters, most synthetic polymers are not attacked by microorganisms [87]. Straight-chain hydrocarbons are effectively only attacked up to a molecular mass of 500 (approximately C_{36}), and above a molecular mass of 620 virtually no degradation occurs. Terpolymers of ethylene, carbon monoxide, and vinyl acetate undergo slow biological degradation [88]. Another, more elaborate route consists of incorporating cellulose or starch units into the polymer chain (Fig. 9) [89]. The cellulose or starch units are then degraded by microorganisms, leaving the synthetic part as material with a lower molecular mass.

Polymers could be made vulnerable to biological attack by incorporation of additives. It is known that additives such as plasticizers or lubricants are attacked by microorganisms and that thereby plastics containing these additives become brittle [90, 91].

In addition, thermoplastic polymers based on poly(hydroxycarboxylic acids) are also available, which are produced synthetically or by microorganisms and which are used in the packaging industry and in medicine [92–95]. These polymers are attractive because of their biodegradability and their compatibility with the environment and within the human body. However, their biodegradation in the environment depends on suitable conditions, including that of sufficient moisture. Furthermore, only the β-hydroxy acids can be polymerized biologically and degraded enzymatically (by microorganisms) within a period of months. So far, the main application has been the poly(hydroxybutyric acid), PHB, manufactured by ICI using a biotechnological process. The material is, however, still about twice as expensive as polyethylene.

More rapid biological degradation is obtained with polylactones [96, 97] from β-propiolactone, γ-butyrolactone, σ-valerolactone, or caprolactone, and especially copolymers which contain glycolic acid, lactic acid, or β-hydroxybutyric acid. Polylactides are rapidly degradable and are therefore used for medical purposes in the form of threads, tissue, or films. Depending on the alkyl group, the polylactides are hydrolyzed within a few days and the monomeric components are then degraded biologically.

7. References

1. A. M. Thayer, *Chem. Eng. News* **68** (1990) no. 30, 37.
2. *Modern Plastics Encyclopedia 1978 – 79,* McGraw-Hill, New York 1978.
3. *Chem. Rundsch.* **43** (1990) no. 38.
4. H. Käufer in K. J. Thomé-Kozmiensky (ed.): *Recycling von Abfällen,* **vol. 1,** EF-Verlag, Berlin 1989, p. 203.
5. H. Kindler, H. Nikles, *Kunststoffe* **70** (1980) 802.
6. D. Pautz, H. J. Pietrzeniuk: "Abfall und Energie," *Abfallwirtsch. Forsch. Prax.* **13** (1984).
7. K. W. Steiner, *Kunststoffe* **77** (1987) 954.
8. Umweltbundesamt, Daten zur Umwelt, Berlin 1986.
9. H. Sinn, *Chem. Ing. Tech.* **46** (1974) 579.

10. H. Collin, G. Grigoleit, G.-P. Bracker, *Chem. Ing. Tech.* **50** (1978) 836.
11. M. Doerr, I. Luederwald, H. R. Schulten, *J. Anal. Appl. Pyrolysis* **8** (1985) 109.
12. H. Sueyoshi, Y. Kitaoka, *Hydrocarbon Process.* **51** (1972) no. 10, 161.
13. J. L. Jones, S. B. Radding: "Solid Wastes and Residues, Conversion by Advanced Thermal Processes," *ACS Symp. Ser.* **76** (1978).
14. J. A. Beckmann, G. Crane, E. L. Kay, J. R. Laman, *Rubber Chem. Technol.* **47** (1974) 597.
15. W. Hofmann, *Gummi Asbest Kunstst.* **28** (1975) 38.
16. H. W. Schnecko, *Chem. Int. Tech.* **48** (1976) 443.
17. H. Emminger, *Kunststoffe German Plastics* **68** (1978) 5.
18. W. Kaminsky, H. Sinn, J. Janning, *Chem. Ing. Tech.* **51** (1979) 419.
19. W. Kaminsky, *Umschau* (1984) no. 22, 674.
20. R. Braun: "Abfallwirtschaft in der Schweiz" in *Technik für den Menschen,* Proceedings, 9/10, Basel April 1976, p. 97.
21. J. L. Pavoni, J. E. Heer, D. J. Hagerty: *Handbook of Solid Waste Disposal,* Van Nostrand Reinhold, New York 1975, p. 294.
22. K. O. Tiltman: *Recycling betrieblicher Abfälle,* WEKA, Kissing 1990.
23. Horizons US 3 912 664, 1975 (E. Wainer); *Chem. Abstr.* **84** (1976) 45 489.
24. J. Leidner: *Plastic Waste,* Marcel Dekker, New York 1981.
25. G. Scott, *Resour Recovery Conserv.* **1** (1976) 381.
26. H. W. Yang, R. Farris, J. C. Chien, *J. Appl. Polym. Sci.* **23** (1979) 11.
27. K. B. Abbas, A. B. Knutsson, S. H. Berglund, *CHEMTECH* 1978, Aug., 502.
28. W. Tötsch, M. Sordo in K. J. Thomé-Kozmiensky (ed.): *Recycling International,* **vol. 1,** EF-Press, Berlin 1989, p. 421.
29. J. Herrler, *Müll Abfall* **6** (1986) 238.
30. E. M. Kampouris, D. C. Diakoulaki, C. D. Papaspyrides, *J. Vinyl Technol.* **8** (1986) 79.
31. G. Härdtke, K. Marek, B. Bilitewski, K. Kijewski: "Recycling von Kunststoffabfällen", *Beih. Müll Abfall* **27** (1988) 64.
32. M. H. Naitove, *Plast. Technol.* **21** (1975) 5.
33. N. B. Frable, *Plast. Eng.* **32** (1976) 5.
34. H. Mylenbusch in K. J. Thomé-Kozmiensky (ed.): *Recycling International,* **vol. 1,** EF-Press, Berlin 1989, p. 361.
35. I. Peters, *VDI Nachr.* **43** (1989) no. 26, 3.
36. G. Menges, *Kunststoffe* **79** (1989) 195.
37. G. Eder in K. J. Thomé-Kozmiensky (ed.): *Recycling International, Recycling of Plastics,* EF-Press, Berlin 1987, p. 78.
38. K. W. Steiner, *Kunststoffe* **74** (1984) 186.
39. H. Menges, VKE, Frankfurt 1980.
40. W. Kaminsky, J. Menzel, H. Sinn, *Conserv. Recycl.* **1** (1976) 91.
41. H. V. Smith, *Polym. Plast. Technol. Eng.* **12** (1979) 122.
42. M. G. Lever, R. S. Parker, *Plast. Rubber Process* **3** (1978) 3.
43. H. Trawinski in R. Pickford (ed.): *Pap. Int. Conf.,* 1984, p. 393, BHRA Fluid Eng. Cranfield, U.K.
44. M. R. Paske, *Polym. Age* **5** (1974) 9.
45. J. Leidner, *Polym. Plast. Technol. Eng.* **10** (1978) 21.
46. W. Fillman, *Plast. Rubber Process* **3** (1978) 3.
47. H. Frederix, *Kunststoffe German Plastics* **68** (1978) 19.
48. E. Grigat, *Kunststoffe German Plastics* **68** (1978) 12.
49. Bayer, DE-OS 2 542 022, 1977 (G. Schneider, K. Wagner, H. W. Patzelt); *Chem. Abstr.* **87** (1977) 6852 x.
50. Ford Motor Co., US 4 316 992, 1982 (J. L. Gerlock, J. Braslaw, J. Albright) *Chem. Abstr.* **96** (1982) 218 707 g.
51. Wyandotte Chem. Corp., US 3 441 616, 1966 (L. C. Pizzini, J. T. Patton); *Chem. Abstr.* **71** (1969) 4107 d.
52. G. Bauer in K. J. Thomé-Kozmiensky (ed.): *Recycling International,* **vol. 1,** EF-Press, Berlin 1989, p. 415.
53. F. Simioni, S. Bisello, *Cell. Polym.* **2** (1983) 281.
54. J. Korff, K.-H. Keim, *Erdöl Erdgas* **105** (1989) 223.
55. L. F. Albright, B. L. Crynes, W. H. Corcoran: *Pyrolysis Theory and Industrial Practice,* Academic Press, New York 1983.
56. G. L. Ferrero, K. Maniatis, A. Buekens, A. V. Bridgewater: *Pyrolysis and Gasification,* Elsevier Appl. Sci., London 1989.
57. *Ullmann,* 4th ed., **6,** 553.
58. K. J. Thomé-Kozmiensky: *Recycling International, Pyrolysis of Wastes,* EF-Press, Berlin 1985.
59. H. Sinn, W. Kaminsky, J. Janning, *Angew. Chem.* **88** (1976) 737; *Angew. Chem. Int. Ed. Engl.* **15** (1976) 660.
60. K. Matsumoto, S. Kurizu, T. Oyamoto in: *Conversion of Refuse to Energy,* Montreux, Nov. 3 – 5, 1975, conf. papers IEEE Catalogue no. 75, Eiger AG, Zürich, p. 538.

61. Kobe Steel, JP 74 111 986, 1974 (K. Inoue, T. Sakai, Y. Kuno); *Chem. Abstr.* **83** (1975) 120 448.
62. Vergasungsverfahren, Voest-Alpine, Prospect, 1990.
63. J. Piskorz, D. Radlein, D. S. Scott, *J. Anal. Appl. Pyrolysis* **9** (1986) 121.
64. J. L. Jones, S. B. Radding: "Thermal Conversion of Solid Wastes and Biomass," *ACS Symp. Ser.* **130** (1980).
65. W. Kaminsky, *Ressour. Recovery Conserv.* **5** (1980) 205.
66. W. Kaminsky, H. Sinn, *Hydrocarbon Process.* **59** (1980) no. 10, 187.
67. W. Kaminsky, *J. Anal. Appl. Pyrolysis* **8** (1985) 439.
68. W. Kaminsky, A. B. Kummer, *J. Anal. Appl. Pyrolysis* **16** (1989) 27.
69. W. Kaminsky, J. Franck, *J. Anal. Appl. Pyrolysis* **19** (1991) 311.
70. A. Buekens, F. De Wolf, J. Schoeters in G. L. Ferrero, K. Maniatis, A. Buekens, A. V. Bridgwater (eds.): *Pyrolysis and Gasification,* Elsevier Appl. Sci., London 1989, p. 580.
71. J. L. Pavoni, J. E. Heer, D. J. Hagerty: *Handbook of Solid Waste Disposal,* Van Nostrand Reinhold, New York 1975.
72. C. H. Huang, C. Dalton: "Energy Recovery from Solid Waste," vol. 2, *NASA CR* **2526** (1974) 243.
73. K. J. Thomé-Kozmiensky: *Recycling International,* EF-Press, Berlin 1982.
74. C. F. Cullis, M. M. Hirschler: *The Combustion of Organic Polymers,* Clarendon Press, Oxford 1981.
75. "Sodium Chloride from PVC," *Plast. Univers.* **12** (1986) 29.
76. W. L. Hawkins: *Polymer Degradation and Stabilization,* Springer Verlag, Berlin 1984.
77. W. Schnabel: *Polymer Degradation,* Akademie Verlag, Berlin 1981.
78. N. Grassie: *Developments in Polymer Degradation,* 7th ed., Elsevier Appl. Sci., London 1987.
79. I. C. McNeill, A. Rincon, *Polym. Degrad. Stab.* **31** (1991) 163.
80. L. Costa, G. Camino, M. P. Luda di Cortemiglia, *Polym. Degrad. Stab.* **14** (1986) 159.
81. M. Day, D. M. Wiles, *J. Appl. Polym. Sci.* **16** (1972) 203.
82. J. E. Guillet, R. G. W. Norrish, *Proc. R. Soc. London A* **233** (1955) 153.
83. Ethylene Plastique S.A., DE-OS 1 918 562, 1969 (P. Le Brasseur, R. Coquelin) Chem. Abstr. 72 (1970) 13 512.
84. E. Dan, J. E. Guillet, *Macromolecules* **6** (1973) 230.
85. Mitsubishi Petro-Chem. Co., JP 38 688 71, 1968 (O. Akiyoshi) *Chem. Abstr.* **77** (1972) 76 131.
86. Eastman Kodak Co., US 454 510, 1966 (G. Newland, G. Cirear, J. Tamblyn) *Chem. Abstr.* **71** (1969) 71 597.
87. F. Rodriguez, *Chem. Tech.,* July (1971) 409.
88. Ruhrchemie DE-OS 23 018 895, 1973 (H. Birnkraut, W. Klug, R. Steiner, W. Wicke) *Chem. Abstr.* **82** (1975) 58 974.
89. S. Kim, V. T. Stannett, R. D. Gilbert, *J. Macromol. Sci. Chem.* **A 10** (1976) 671; *J. Polym. Sci. Polym. Lett. Ed.* **11** (1973) 731.
90. G. H. Booth, A. W. Cooper, *J. Appl. Bacteriol.* **31** (1968) no. 3, 305.
91. G. J. L. Griffin: "Biodegradable Fillers in Thermoplastics," *Conf. on Degradability of Polymers and Plastics,* IEE, London 1973.
92. C. G. Gebelein, C. E. Carraher: *Polymeric Materials in Medication,* vol. 32, Plenum Press, New York 1984.
93. Z. Jedlinski, M. Kowalczuk, *Macromolecules* **22** (1989) 3242.
94. M. Endo, T. Aida, S. Inoue, *Macromolecules* **20** (1987) 2982.
95. H. R. Kricheldorf, N. Scharnagl, *J. Macromol. Sci. Chem.* **A 26** (1989) 951.
96. H. R. Kricheldorf, I. Kreiser-Saunders, N. Scharnagl, *Macromol. Chem. Macromol. Symp.* **32** (1990) 285.
97. *Chem. Rundsch.* **43** (1990) no. 45, 1.
98. www.kvs.ch/bulletin2000/bulletin2_2000/kap2.htm

Plastics, Reinforced → Reinforced Plastics
Plastics Technology → Polymerization Processes
Platelet-Affecting Drugs → Cardiovascular Drugs

Platinum Group Metals and Compounds

Hermann Renner, Degussa AG, (retired), Hanau, Federal Republic of Germany (Chaps. 1 – 5, Section 10.8)

Günther Schlamp, Demetron GmbH, (retired), Hanau, Federal Republic of Germany (Chap. 2, Sections 7.2, 8.1, 8.2, 10.7, 10.9, 11.1 – 11.3)

Ingo Kleinwächter, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Chap 6, Section 7.1)

Ernst Drost, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Sections 8.3, 10.2)

Hans Martin Lüschow, Degussa-Hüls AG, (retired), Hanau, Federal Republic of Germany (Chap. 9)

Peter Tews, Allgem. Gold und Silberscheideanstalt (AGOSI), Pforzheim, Federal Republic of Germany (Section 10.1)

Peter Panster, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Section 10.3)

Manfred Diehl, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Section 10.4)

Jutta Lang, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Section 10.5)

Thomas Kreuzer, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Section 10.6)

Alfons Knödler, Forschungsinstitut für Edelmetalle und Metallchemie, (retired), Schwäbisch Gmünd, Federal Republic of Germany (Section 10.10.1)

Karl Anton Starz, dmc2 Degussa Metals Catalysts Cerdec AG, Hanau, Federal Republic of Germany (Sections 10.10.2, 10.10.3)

Klaus Dermann, Ducera Dental GmbH und Co. KG, Hanau, Federal Republic of Germany (Section 10.11)

Josef Rothaut, Ducera Dental GmbH und Co. KG, Hanau, Federal Republic of Germany (Section 10.11)

Ralf Drieselmann, dmc2 Degussa Metals Catalysts Cerdec AG, Frankfurt/M., Federal Republic of Germany (Section 11.4)

Catrin Peter, Klinikum der Friedrich Schiller Universität, Jena, Federal Republic of Germany (Chap. 12)

Rainer Schiele, Klinikum der Friedrich Schiller Universität, Jena, Federal Republic of Germany (Chap. 12)

1. History 3
2. Properties 6
3. Occurrence 8
3.1. Abundance 8
3.2. Ores and Their Origin 10
3.3. Primary Deposits 10
3.4. Secondary Deposits 12
3.5. Recovery of Secondary Platinum Group Metals 13
3.6. Reserves and Resources 15
4. Mineral Dressing and Beneficiation 15
4.1. Treatment of Alluvial Platinum Deposits 15
4.2. Treatment of Primary Deposits . 16
4.3. Treatment of Nickel Ores 17
4.4. Treatment of Metal Scrap 17
4.5. Treatment of Dross 18
4.6. Treatment of Supported Catalysts 18
4.7. Treatment of Solutions 19
5. Dissolution Methods 19
5.1. Dissolution in Aqua Regia 20
5.2. Dissolution in Hydrochloric Acid – Chlorine 20
5.3. Dissolution in Hydrochloric Acid – Bromine 20
5.4. Other Dissolution Processes . . . 21
5.5. Dissolution by Salt Fusion 21
6. Separation of Platinum Group Metals 21

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_075

6.1. Chemistry of Platinum Group Metal Separation 22
6.2. Older Separation Processes 25
6.3. Current Separation Processes . . 26
6.4. Processes Used in Coarse Separation 28
6.5. Purification 31
6.6. Conversion of Salts into Metals . 33
6.7. Partial Purification 35
6.8. Treatment of Internally Recycled Material 35
6.9. Construction Materials 35
7. Platinum Group Metal Compounds 36
7.1. Inorganic Compounds 36
7.1.1. Platinum Compounds 36
7.1.2. Palladium Compounds 37
7.1.3. Rhodium Compounds 38
7.1.4. Iridium Compounds 38
7.1.5. Ruthenium Compounds 38
7.1.6. Osmium Compounds 39
7.2. Organic Compounds 39
8. Alloys 39
8.1. Alloy Systems 39
8.2. Special Alloys 40
8.3. Methods of Treatment 49
9. Quality Specifications and Analysis 49
9.1. Quality Specifications 49
9.2. Qualitative Analysis 50
9.3. Quantitative Analysis 51
9.4. Purity Analysis 52
9.5. Trace Analysis 52
10. Uses 52
10.1. Jewelry, Coinage, Investment . . 52
10.2. Apparatus 53
10.3. Heterogeneous Catalysts 53
10.4. Fuel Cells 56
10.5. Homogeneous Catalysts 56
10.6. Automotive Emission Control Catalysts 57
10.7. Sensors 58
10.8. Electrical Technology 60
10.9. Electronics 60
10.10. Coatings 62
10.10.1. Coatings Produced by Electrolysis 62
10.10.2. Coatings Produced by Chemical Reaction 64
10.10.3. Coatings Produced by Physical Methods 64
10.11. Dental Materials 65
11. Economic Aspects 65
11.1. Supply 65
11.2. Demand 66
11.3. Prices 66
11.4. Commercial Aspects 66
12. Toxicology 68
13. References 69

1. History [1–10], [166], [167]

Figure 1 gives an historical survey of important platinum group metal discoveries and platinum group metal technology.

Early Times. The earliest evidence of platinum is provided by a gold etui covered with hieroglyphic inscriptions, dating from the 7th century B.C. Around 1900, BERTHELOT (1827 – 1907) investigated the etui, which was kept in the Louvre, and found that some of the inlays hitherto thought to be silver were in fact platinum.

In ca. 1900, jewelry made of native platinum was discovered in Ecuador, which was part of the Inca empire in pre-Columbian times. This probably dates from the first five centuries A.D.

In ancient times, the technique of washing river sands and fusing together the grains of platinum was undoubtedly known, as was the making of alloys by heating alluvial platinum and gold with a blowpipe, these methods being similar to those used in gold extraction and working. Pre-Columbian Indians were familiar with a powder metallurgy technique.

16th–18th Century [11], [12]. In their search for gold in the New World, particularly in the area of present-day Colombia, the Spanish often found alluvial ("placer") platinum. The earliest written report of the metal was made in 1557 by J. C. SCALIGER (1484 – 1558). However, platinum was worked to produce jewelry and utensils only after man had learned that the metal, initially believed to be infusible, could be melted and cast by first adding other metals to lower its melting point. Platinum was essentially regarded as a troublesome material, accompanying gold and silver, which lowered their workability. The Spanish named the metal *platina*, the diminutive form of *plata* (silver), as a derogatory term because it was found only in small quantities or as small granules. The

Year	Discovery	Development	Period
1990			
		Paraplatin in cancer therapy	Technical evolution
1985			
		Automobile exhaust catalysts Oxo synthesis with homogeneous Rh catalysts	
1980	Stillwater, U.S.A.	Ru thick film pastes	
	UG-2 Reef, R.S.A.		
1970			
		Cisplatin in cancer therapy	
		Spark plugs with Pt-Ir electrodes	
1960		Homogeneous catalysis of ethylene oxidation with Pd/Cu Oil refining with supported Pt catalysts	
1950			
1940			
		Processing of sulfidic PGM ores in R.S.A.	Modern times
1930	Merensky Reef, R.S.A.	NH_3 oxidation on Pt-Rh catalysts	
1920			
1910		Fountain pen nib alloys	Industrial revolution
1900			
		Electrical industry (1870: resistance thermometer; 1880: Pt-Rh thermoelement) jewelry, dentistry	
1850			
	Ru (1844) Rh, Ir, Os (1804) Pd (1803)	Platinum coins in Russia Dobereiner's lighter Pt purification via (NH_4) $[PtCl_6]$	
1800			Scientific discoveries
		Beginnings of platinum technology (purification, working)	
1750	Pt (1750)		
		Scientific study of platinum begins	
1700			
1600			
	Native Pt (1559)	Platinum finds in New Grenada (Columbia)	
1500			
1400			Middle ages: Natural philosophy, alchemy
1300			
		Platinum jewelry in Inca Empire (Ecuador, Peru)	
1200			
1000			
500			Ancient times: Iron age
A.D. 0			
B.C.			
500			
		Gold etui with platinum inlays (Thebes, Egypt)	
1000			Bronze age
2000			

Figure 1. Historical survey of important platinum group metal discoveries and platinum group metal technology

terms "white gold" and "heavy silver" were also sometimes used. When the metal became well known in Europe in 1748, thanks to the Spanish mathematician A. DE ULLOA (1716 – 1795), the demand for platinum increased. Its high density enabled it to be used to adulterate gold, and its introduction into Europe was therefore prohibited. Until 1908, the price of platinum was lower than that of gold. Around 1750, the scientific investigation of platinum began, initiated largely by C. WOOD (1702 – 1774), who learned of the metal in Jamaica in 1741. He can be regarded as the true discoverer of platinum, having presented a paper in 1750 to the Royal Society entitled "The New Semi-Metal Called Platina." In this initial period, fundamental investigations into the chemistry of platinum (solubility in aqua regia, precipitation by addition of ammonium chloride, fusion by addition of arsenic, lead cupellation, etc.) were carried out by W. WATSON (1715 – 1787), W. LEWIS (1708 – 1781), and others. These workers, including WOOD, often collaborated. Research at this time had as its primary aim, distinguishing and separating platinum and gold, although platinum was at first widely believed to consist of gold contaminated with other elements.

An important development in platinum technology was the technique of converting the alluvial (placer) deposits into platinum sponge, which could be satisfactorily formed by heat into compact platinum or platinum artifacts. F. C. ACHARD (1753 – 1821) discovered the method of oxidizing an easily fusible platinum – arsenic alloy to remove arsenic. As early as 1784, he produced the first platinum crucible from platinum sponge made in this way.

19th Century. Around 1800, the accompanying metals in native platinum, which generally contains up to 80 % Pt, were discovered. W. H. WOLLASTON (1766 – 1828) discovered palladium (initially also known as "new silver") and rhodium; S. TENNANT (1761 – 1815) discovered iridium and osmium. At the same time, the existence of platinum as a true element was established. C. CLAUS (1796 – 1864) discovered ruthenium in 1844.

In 1823, J. W. DOEBREREINER (1780 – 1849) first used the catalytic action of platinum in the gas lighter named after him. In about 1810, the process invented by W. H. WOLLASTON became established: dissolving the raw material in aqua regia, purifying it by precipitation of ammonium hexachloroplatinate, $(NH_4)_2[PtCl_6]$, and heating this strongly to form platinum sponge. In 1817, this led to the formation of Johnson, Matthey & Co., London [13], which laid the foundation of modern platinum technology.

Until the beginning of the 19th century, almost all platinum was obtained from the area now known as Colombia, which from 1739 to 1819 belonged to the Spanish Crown Dependency of New Granada. At this time, ca. 1 t/a of platinum was extracted. ALEXANDER VON HUMBOLDT was the most accomplished prospector and developer of noble-metal deposits in New Granada (1819) and of the newly discovered platinum deposit in the Urals. In 1819, platinum was extracted from alluvial gold, but soon after this, the main production was switched to platinum placers with low gold content. In 1825, Russia became the primary producer of platinum. The minting of platinum coins in Russia in 1828 – 1845 necessitated an increase in production, which reached 3.5 t in 1843 [14–16]. Almost all Russian platinum ore was used for coinage (15 t total). Technology for the extraction and treatment of placer deposits had reached a high standard by this time.

After the closing of the refinery in St. Petersburg, which was associated with the local mint and produced a metal of ca. 97 % platinum, 1.2 % iridium, 0.5 % ruthenium, 0.25 % palladium, 1.5 % iron, and 0.4 % copper, the refining and working of Russian platinum ore were carried out almost exclusively in Western Europe. New companies for the refining and working of platinum were founded, including Desmoutis, Paris (1822); Baker & Co., New York (from 1904: Engelhard, Newark, New Jersey); Heraeus, Hanau (1851); Siebert, Hanau (1881; from 1930 part of Degussa, Frankfurt); J. Bishop & Co., Malverne, Pennsylvania and others.

20th Century [17–20]. The platinum industry grew vigorously after 1880 due to increasing demands of the electrical industry, dentistry, and chemical technology. The primary consumer was the United States. In 1913, annual production of crude platinum reached 7 t. At this time, Russia began to make itself less dependent on other countries by constructing a modern platinum refinery and smelting point at

Ekaterinburg (now Sverdlovsk). However, at the result of an ideologically negative attitude toward noble metals and a false estimate of their economic impor-tance, platinum production was abandoned after the Russian Revolution.

The result was that, after World War I, Colombia once again became the largest platinum producer. The raw materials were exclusively alluvial deposits, which were extracted by panning, as they are even today. Almost all of the Colombian crude platinum was refined in the United States.

In 1925, production in the Urals was restarted. New alluvial deposits were discovered in Siberia. Also, mining of primary platinum became important. Today, a large proportion of the platinum group metals supplied by the CIS comes from sulfidic nickel deposits. For some decades, the Soviet Union has been one of the largest producers of platinum and especially palladium. Production figures are not available thus far. Estimates are approximate and are based on export figures.

During World War I, the demand for nickel increased, and Canada became an important producer of palladium and platinum, because nickel ores also contained platinum group metals. Since 1890, these ores have been extracted by the Mond Nickel Co. (since 1961, INCO, London), later allied with the International Nickel Co. of Canada. From 1925, platinum metals were produced in their own refinery in Acton in London. Another important producer of platinum is Falconbridge Nickel Mines in Toronto.

After World War I, worldwide demand could no longer be satisfied by Russia and other producers, and new sources were developed. In South Africa, the search was particularly well rewarded. Platinum and osmiridium were discovered at Black Reef (1888), Witwatersrand (1892), Great Dyk (Rhodesia, 1918), and later at Waterberg. The award of prospecting rights to private persons led to prospecting on a wide scale. A systematic geological survey of the Bushveld by J. Merensky was highly successful, leading to the discovery of platinum in the dunite pipes (1924) and subsequently in the stratiform platinum deposits known as the Merensky Reef (1925), the largest platinum deposit in the world.

This discovery brought about the greatest upheaval yet seen in the platinum market. It occurred during a period of both high demand (mainly for catalysts for ammonia oxidation) and high prices, the result being the foundation of about 50 producing companies. The primary platinum-bearing rock typical of South African deposits presented the ore treatment and smelting technologies with completely new problems. Technical difficulties combined with a fall in the price of platinum led to a slump. The largest of the surviving companies formed Rustenburg Platinum Mines Ltd. in 1931, whose output reached 3 t/a during World War II.

After World War II, the Union of South Africa (Republic of South Africa) became the main producer of platinum. From 1969, all of the crude platinum produced by South Africa was refined by Johnson Matthey in England, although Matthey Rustenburg Refiners now refines considerable quantities of material produced in Rustenburg. Recently, some new producers have appeared in South Africa [e.g., Impala Platinum, Western Platinum (a subsidiary of Lonrho and Falconbridge), and Anglo-Transvaal Consolidated Investment (Anglovaal) and its subsidiary Atok Platinum Mines].

Figure 2 shows the history of world platinum output, and Table 1 lists the development of PGM production according to region.

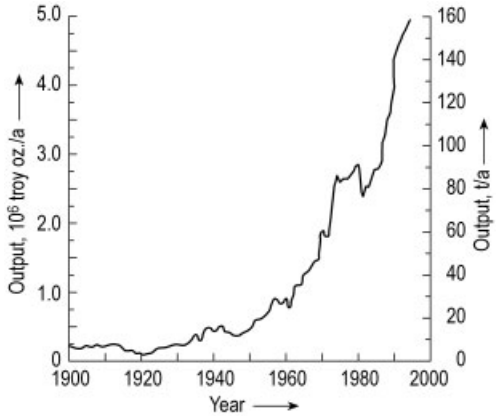

Figure 2. History of world platinum output (including GUS sales to Western World)

2. Properties [166], [167]

Atomic Properties [21], [22]. The electronic structures of platinum group metals start

Table 1. Historical development of the production of platinum group metals according to region∗ (in t/a)

	1800	1850	1900	1910	1920	1930	1940	1950	1960	1970	1975	1980	1985	1990	1995	1997
Colombia	1	< 1	< 1	< 1	2	2	2	1	1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
Russia ∗∗	1	2	12	11	1	3	10	6	12	30	70	70	70	70		191
Canada			≪ 1	≪ 1	1	2	6	7	12	12	12	12	12	12		
South Africa						2	2	4	20	50	90	120	120	130		189
United States			≪ 1	≪ 1	≪ 1	< 1	2	2	1	1	1	1	1	10		33
World	1	2	12	12	4	10	22	20	45	93	173	203	203	222	378	413

∗ Other countries account for only 1 – 2 % of world production.
∗∗ Estimated.

from the inert gas structures of krypton and xenon:

Ru	[Kr]	$4d^7$	$5s^1$	
Os	[Xe]	$4f^{14}$	$5d^6$	$6s^2$
Rh	[Kr]	$4d^8$	$5s^1$	
Ir	[Xe]	$4f^{14}$	$5d^7$	$6s^2$
Pd	[Kr]	$4d^{10}$		
Pt	[Xe]	$4f^{14}$	$5d^9$	$6s^1$

Two valence orbitals exist: an *s* shell and an inner *d* shell. The small energy difference between them means that the electron shells are filled in inconsistently, so that with some elements, one or both electrons from the outer 5 *s* or 6 *s* shells are taken up by the 4 *d* or 5 *d* shells.

The Platinum Group Metals show close similarities in general physical and chemical properties. The very small differences in the atomic sizes of the elements of the second and the third row of the periodic system are explained by the full occupation of the $4f^{14}$ electron orbitals of the "heavy" elements Os, Ir, Pt (see "Lanthanide contraction"). Due to the electronic structure of these elements, which is also determined by relativistic phenomena, causing splitting of p- and d-orbitals to altered energy levels, they show, compared to the "light" platinum group elements (Ru, Rh, Pd), marked differences in their catalytic activities both in homogeneous and heterogeneous catalysis [23] (see also Sections 10.3 and 10.5).

Because of the small energy differences between the valence shells, a number of oxidation states occur. The following oxidation states are known in the compounds of platinum group metals (principal oxidation states in bold print):

Ru:	−2,	0,	+2,	**+3**,	**+4**,	+5,	+6,	+7, +8
Rh:	−1,	0, +1, +2,		**+3**,	+4,	+5,	+6	
Pd:	0,		**+2**,	+3,	+4			
Os:	−2,	0, +1, +2, +3,			**+4**,	+5,	+6,	+8
Ir:	−1,	0, +1, +2, +3,			**+4**,	+5,	+6	
Pt:	0,		**+2**,		**+4**,	+5		

The other properties typical of transition metals are very marked; for example, catalytic activity due to their readiness to change valence, formation of intermediate compounds with different reagents, color, paramagnetism due to unpaired electrons, and strong tendency to form complexes.

Comparisons within the group of platinum metals, also including neighboring elements, often give an insight into the relationship between electronic configuration and chemical properties.

Many properties show marked similarities along the two horizontal rows:

Ru ↔ Rh ↔ Pd ↔ (Ag)
Os ↔ Ir ↔ Pt ↔ (Au)

Also, the tendency toward complex formation and higher oxidation states is more marked in the row of heavy elements than in the row of light elements.

Vertical similarities between the elements also occur (e.g., in their behavior toward acids; (see Table 3):

Ru	Rh	Pd	(Ag)
↕	↕	↕	↕
Os	Ir	Pt	(Au)

Table 2. Atomic and physical properties of the platinum group metals , p. 267[35]

	Ru	Rh	Pd	Os	Ir	Pt
Atomic properties						
Atomic number	44	45	46	76	77	78
Atomic weight	101.07	102.905	106.42	190.2	192.22	195.08
Crystal structure	hcp	fcc	fcc	hcp	fcc	fcc
Lattice constant, Å	$a = 2.701$ $c = 4.275$ $c/a = 1.583$	3.803	3.887	$a = 2.735$ $c = 4.319$ $c/a = 1.579$	3.839	3.924
Atomic radius, Å	1.325	1.345	1.376	1.377	1.357	1.377
Mechanical properties						
Density (20 °C), g/cm^3	12.45	12.41	12.02	22.59	22.56	21.45
Density liquid at melting point, g/cm^3	10.90	10.70	10.49	20.1	19.39	18.91
Hardness (HV), Kg/mm^2	250 – 500[a]	130	50	300 – 680[a]	200	48
Young's modulus, GPa	485	386	124	570	538	173
Modulus of rigidity, MPa	172	153	51	220	214	67
Poisson's ratio	0.29	0.26	0.39	0.25	0.26	0.39
Coefficient of compressibility	0.31	0.36	0.52	0.28	0.28	0.36
Thermal properties						
Melting point, °C	2310	1966	1554	3045	2410	1772
Vapor pressure at *mp*, mbar	5.5	3.5	20	20	4	0.1
Boiling point, °C	3900	3700	2970	5000	4130	3827
Thermal conductivity (20 °C), W m^{-1} K^{-1}	117	150	75	87	147	73
Coeff. of linear expansion (0 – 100 °C), 10^{-6} K	9.5	8.5	11.1	6.5	6.8	9.0
Electrical properties						
Specific electrical resistance (0 °C), μΩ cm	6.7	4.34	9.725	8.2	4.7	9.825
Temperature coefficient of electrical resistance (0 – 100 °C), 10^{-3} K	4.0	4.6	3.8	4.2	4.3	3.92
Transition temperature of superconductivity	0.47	0.9		0.71	0.11	
Thermoelectric voltage against Pt (0 – 100 °C), mV	0.684	0.70	− 0.570		0.660	0
Work function	4.52	4.8	4.97	4.55	5.40	5.27

Similarities also exist in the following diagonal sequence, sometimes with ruthenium and iridium changing places:

```
Ru   Rh   Pd   (Ag)
 ↕↘  ↕↘  ↕↘   ↕
Os   Ir   Pt   (Au)
```

The chemical properties of the platinum group metals and the chemistry of their compounds fit less well into such a scheme than their physical properties.

All platinum group metals have several naturally occurring stable isotopes, with the exception of rhodium, which has only one.

Physical Properties [22, 24–28], [167]. The outstandingphysical properties of the platinum group metals are of great importance for their industrial use. These include high melting point, low vapor pressure, high temperature coefficient of electrical resistivity, and low coefficient of thermal expansion. Table 2 lists atomic and crystal data and physical properties of the platinum group metals.

Osmium is the densest metal known. The density has been recalculated on the basis of experimental data and the atomic weight of 190.23 +/− 0.03 to be 22.587 +/− 0.009 g/cm^3, while for iridium the corresponding values are 192.217 +/− 0.003 for the atomic weight and a density of 22.562 +/− 0.009 g/cm .

Chemical Properties [29–34]. Table 3 shows the chemical behavior of platinum group metals toward various reagents. Resemblances are most clear within the two groups of elements. The aim is to quantify the reactions, so as to provide useful information when choosing construction materials or dissolving platinum group metals. Some combinations of reagents that can be used to dissolve these metals are also detailed in Chapter 5.

Table 3. Chemical resistance of the platinum group metals

Reagent	Conditions	Temperature, °C	Pd	Pt	Rh	Ir	Ru	Os
Hydrochloric acid	36 %	20	–	–	–	–	–	–
Hydrochloric acid	36 %	100	•	–	–	–	–	●
Nitric acid	65 %	20	●	–	–	–	–	●
Nitric acid	65 %	100	●	–	–	–	–	●
Sulfuric acid	96 %	20	–	–	–	–	–	–
Sulfuric acid	96 %	100	●	–	•	–	–	–
Sulfuric acid	96 %	300		•	●			
Hydrobromic acid	60 %	20	●	•	•	–	–	–
Hydrobromic acid	60 %	100	●		●	–	–	●
Hydroiodic acid	57 %	20	●	–	–	–	–	•
Hydroiodic acid	57 %	100	●	●	–	–	–	●
Hydrofluoric acid	40 %	20	–	–	–	–	–	–
Phosphoric acid		100	•	•	–	–	–	●
Acetic acid	99 %	100	–	–	–	–	–	
Hydrochloric acid/chlorine	20 %/saturated	20	●	•				
Hydrochloric acid/chlorine	20 %/saturated	80	●	●				
Hydrochloric acid/chlorine	20 %/saturated	100	●	●	•	•		
Hydrochloric acid/bromine		20	●	•	•			
Hydrochloric acid/bromine		100	●	•				
Aqua regia		20	●	●	–	–		●
Aqua regia		100	●	●	–	–		●
Aqua regia		150	●	●		●		
Hydrochloric acid/H_2O_2		20	●					
Hydrochloric acid/H_2O_2		100	●					
Hydrobromic acid/bromine	60 %	100			●			
Water/bromine		20	•	–	–	–	–	–
Ethanol/iodine		20	•	–	•	–	–	
Sodium hypochlorite solution		20	●	–	•		●	●
Sodium hypochlorite solution		100	●	–	•	•	●	●
Sodium cyanide solution		20	●	–				
Sodium cyanide solution		100	●	●				
Copper(II) chloride solution		100	•	–				
NaOH melt	+air	500	•	•	•		●	●
KOH melt	+air	500	•	●	•		●	●
NaOH melt	+air	800	●		•	–	●	●
KOH melt	+air	800	●		•		●	●
$KHSO_4$ melt	+air	440	•	–	●	–		
NaCN melt	+air	700	●	●	●	●	●	
KCN melt	+air	700	●	●	●	●	●	
NaCN/KCN melt (2 : 1)	+air	550	●	●	●	●	●	
Chlorine, gaseous	dry	20	●	•	–	–	–	–
Chlorine, gaseous	moist	20	●	•	–	–	–	●
Bromine liquid	dry	20	●	●	–	–	–	●
Bromine liquid	moist	20	●	●	–	–	–	●
Iodine, solid	dry	20	–	–	–	–	–	•
Iodine, solid	moist	20	•	–	•	–	–	–
Fluorine, gaseous		20		•				
Hydrogen sulfide, gaseous	moist	20	–	–	–	–	–	–

– Mass loss $<0.01\,mg\,cm^{-2}\,h^{-1}$; ideal as construction material
• Mass loss ca. $0.1\,mg\,cm^{-2}\,h^{-1}$; limited use as construction material
● Mass loss ca. $1\,mg\,cm^{-2}\,h^{-1}$; limited use for dissolution processes
● Mass loss $10\,mg\,cm^{-2}\,h^{-1}$; suitable for dissolution processes.

3. Occurrence

3.1. Abundance [10], [31], [36–39], [166], [167]

The abundance of platinum group metals (PGMs), which occupy an intermediate position based on their atomic number and atomic weight, would be expected to be 10^{-4} ppm, based on the mode of formation of atomic nuclei [40]. They are concentrated in planetary regions, reaching ca. 30 ppm in the earth. Considerable fractionation has taken place in the earth's interior, due mainly to the siderophilic chemical character of PGMs, so that virtually the entire mass of PGMs is in the earth's metallic core. The siliceous lithosphere is estimated to contain 0.05 – 0.5 ppm.

On theoretical grounds, the PGMs in the earth are assumed to contain ca. 20 % each of plat inum, palladium, ruthenium, and osmium, and ca. 6 % each of rhodium, and iridium. In the case of ruthenium and especially osmium, these values are not reflected in the deposits mined to date.

3.2. Ores and Their Origin [10], [31], [32], [36], [41–48], [166], [167]

The PGMs in the lithosphere have been transferred from the earth's interior. Tectonic movements of the earth's crust, followed by the eruption of magma, have led to their presence in regions close to the surface. Solidification processes, differences in melting point and density, gas emissions, convection due to heat, and the flow and eruption of magma, have all produced concentration and separation effects, mainly in marginal zones. Chemical interaction with high-temperature silicate layers, especially their sulfide, arsenide, antimonide, selenide, and telluride components, has also played a major role. Almost invariably, the platinum group elements have separated from ultrabasic magmas. Norite (Mg – Fe – Ca – Al silicate) contains mainly sulfidic intrusions, whereas those in dunite (Mg – Fe silicate) are mainly sulfide free. Platinum and palladium (often with nickel, copper, chromium, etc.) sometimes undergo hydrothermal reactions with chlorides in the earth's interior. All of these processes have led to the formation of the primary deposits of platinum-bearing rock. Workability depends on many factors–concentration of platinum metals, accessibility, size of deposit, value and potential uses of accompanying materials–and is economical in only a few cases.

When primary deposits are altered and transported by the natural action of the hydrosphere and atmosphere, secondary deposits, also known as placers or alluvial deposits, are formed. Mechanical concentration of the heavy constituents by flowing water takes place, together with chemical dissolution and reprecipitation of the platinum metals. Recently, hydrothermal processes have been shown to be considerably more important than was at first thought. Alluvial deposits usually originate from dunite.

The platinum metals occur in a large number of minerals. Workable ore deposits contain mainly sperrylite ($PtAs_2$), cooperite (PtS), stibiopalladinite (Pd_3Sb), laurite (RuS_2), ferroplatinum (Fe – Pt), polyxene (Fe – Pt – other platinum metals), osmiridium (Os – Ir), and iridium platinum (Ir – Pt). These minerals are associated with particular carrier materials, which are often valuable themselves (e.g., iron pyrites, nickel iron pyrites, or chrome iron ore).

The minerals are seldom present in an exact stoichiometric ratio. This is true of the platinum group metals themselves,which are nearly always present in varying ratios, and of the accompanying elements with which they form compounds or alloys. Isomorphism opens up the possibility of further variations. Isomorphic intercalations make up a larger part of the economically workable reserves. Electron probe microanalysis (EPM) has enabled a large number of definite compounds and intermetallic phases to be identified where mixtures or homogeneous solid solutions had formerly been assumed to be present.

Outside of true deposits, platinum group metals are widely distributed in very high dilution as isomorphous combinations with various metals such as nickel, cobalt, and copper, mainly in their sulfides.

3.3. Primary Deposits [49–65], [166], [167]

The dunite bodies in the Urals are the most important of the old Russian platinum deposits, ei-

ther as primary deposits or as material for the formation of secondary deposits. The only important primary deposit is at Nishnij-Tagil, which is mined in many locations. It is a projectile-like intrusion of dunite through the earth's crust, which appears to reach a depth of more than 100 km and has an area on the earth's surface of 25 km^2. Platinum metals are concentrated in dunite in the form of striae, lenses, nests, and pillars. They are present mainly as polyxene (Fe – Pt), iridium-rich platinum, and osmiridium, and are often associated with serpentine or chromite. Most of the platinum has been obtained from ores with a platinum metal content of 10 – 20 ppm. However, ores containing 400 ppm are found in some places. In the total dunite body, the average platinum content is ca. 0.1 ppm. Other dunite bodies in this region are not economically important either for their metal content or for the extent of the deposit.

The dunite pipes at the eastern border of the South African Bushveld have a similar origin to the Russian dunite bodies and are of the same base material. The platinum-bearing core of this narrow intrusion often has a diameter of 20 m or less. Veins are rarely present. The excavations at Driekop and Onverwacht are the most well known. Mining is difficult. Since the exploitation of the Merensky Reef, they can be mined economically only in certain cases. Platinum group metals are sometimes present in the metallic state and sometimes as sperrylite in association with chromite. Platinum concentrations are 1 – 200 ppm, and locally higher.

The richest and scientifically most interesting deposits, although small in extent, are the quartz lodes of Rietfontein on the Waterberg in the Western Cape Province. Owing to their hydrothermal origin, the platinum metals are present in native form, accompanied by hematite. The PGM content of the ore often reaches almost 5000 ppm. Deposits of similar structure have not been found elsewhere.

The largest known primary deposit of PGMs is the South African Bushveld Complex, with a total extent of ca. 250 km from north to south and 480 km from east to west. This oval-shaped zone was apparently produced by an outflow of PGM-bearing norite magma into a flat basin. At the bottom of this, the minerals sperrylite ($PtAs_2$) and cooperite (PtS) have separated along with iron pyrites, nickel pyrites, copper pyrites, and chromite. The relatively thin PGM-bearing layers at the edges of the outflow near the earth's surface are mined. These are in both the oxidation and the sulfidic zones, and begin at a depth of < 100 m. The working depth extends to almost 1000 m. The platinum content of the mined ore is 3 – 20 ppm. Palladium is less important than platinum; all the platinum group metals are present. The deposit has great economic importance owing to its consistent quality, ease of extraction, and large extent. The Bushveld is mined at the western and southwestern edges (Rustenburg region) and the eastern and northeastern edges (Lydenburg region) (Fig. 3) in two wide areas. The Bushveld is also the world's largest chromium and vanadium deposit, and has large deposits of nickel, cobalt, copper, iron, tin, fluorspar, and alusite, magnesite, and asbestos.

These deposits were discovered in 1925 by the geologist Hans Merensky (1871 – 1952) [17]. He laid the foundations for the most important phase in the development of platinum extraction by his wide-ranging prospecting work and fundamental observations on the geology of the Bushveld. The most important areas of PGM deposits in the western and eastern part of the Bushveld are known as the Merensky Reef. Not far from these deposits lie the less important Platreef and the UG 2 Reef (Upper Group Reef), which has recently increased in importance due to its very high rhodium content in many places. This metal is in great demand and therefore very expensive.

The UG 2 reef is a chromite layer running parallel to the Merensky Reef, but somewhat deeper [66], [67].

Some examples of small deposits include the Black Reef in the Bushveld and, a few hundred kilometers north, the Great Dyk Mine in Zimbabwe.

The Stillwater deposit in Montana, United States is similar in origin and structure to the Bushveld deposit. However, PGM separation occurred at the boundary zone of a subterranean magma intrusion resembling a sloping lens. This deposit extends for a length of ca. 80 km. The total PGM content, with high levels of palladium and rhodium, is higher than that of the Merensky Reef deposit.

The nickel sulfide deposits of the Sudbury, Ontario district of Canada, yielding iron pyrites (FeS), pentlandite (F – NiS), and copper pyrites

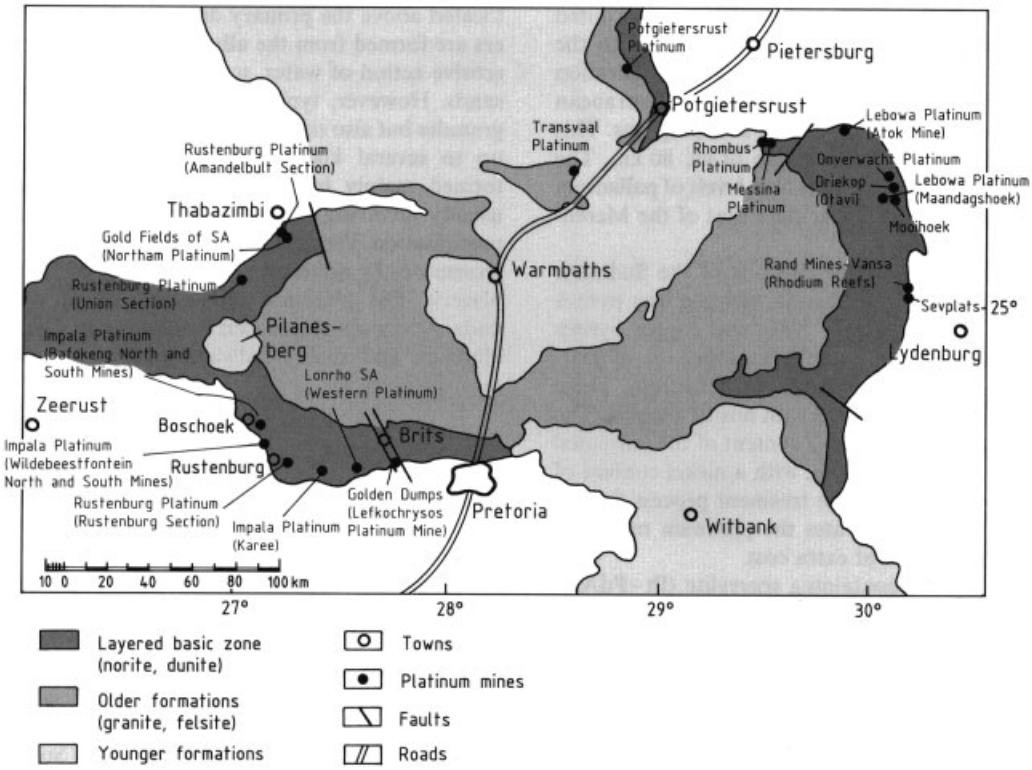

Figure 3. Geology and platinum mines of the Bushveld Complex

($CuFeS_2$), are today important sources of platinum, and more especially of palladium. These deposits are associated with noritic magma. The average platinum metal content of the untreated ore is only ca. 0.3 ppm, with a nickel content of ca. 2 %. However, the treatment process for the nickel ore concentrates the platinum metals to > 50 ppm without extra cost.

Palladium-containing sperrylite ($Pt-PdAs_2$) and stibiopalladinite (Pd_3Sb) are present in the ores, and the platinum group metals also form isomorphic mixtures with the heavy-metal sulfides. During formation of these deposits, the sulfides of nickel and copper acted as collectors. The large reserves in Noril'sk in Central Siberia, which have a high palladium content, are similar to the Canadian deposits with regard to composition and ore treatment methods. Some lead – zinc ores [e.g., in the Rammelsberg (Harz) mountains], contain small amounts of platinum and palladium.

Comparisons with other metals indicate that PGM reserves may exist in the Antarctic [68]. The magma flow in the South African Bushveld occurred in an early geological epoch (Cambrian – more than 2×10^9 years ago), when the African and Antarctic continents were still part of Gondwanaland.

3.4. Secondary Deposits [166], [167]

Alluvial deposits are typical reserves of this type. They were produced in recent geological epochs (Quaternary – Holocene) by weathering and washing of primary deposits, mainly dunitic, which resulted in concentration of the more resistant and heavier components (i.e., the metallic and arsenidic platinum minerals, gold, magnetite, chromite, cassiterite, zircon, and granite) in clay and sand. These oxidation zones are often located above the primary deposits. River placers are formed from the alluvial deposits by the erosive action of water, to yield so-called black sands. However, typical nuggets (mainly small granules but also sizable metallic lumps

weighing up to several kilograms) seem to have been formed mainly by mechanical agglomeration, usually involving chloridic dissolution and reprecipitation. Very old deposits that formed conglomerates by adhesive action are termed fossil placers. The platinum metal content of secondary deposits varies over a wide range. Deep alluvium and river headwater placers are the most productive.

For easily washable sand, a PGM content of 0.05 ppm is economic. Before World War I, the PGM content of economic deposits was ca. 2 ppm, but today it is much lower. More platinum metals are recovered from river placers than from alluvial deposits.

Secondary deposits occur mainly in the Urals, Siberia, Colombia, and Ethiopia. Often, the recovery of gold and platinum metals from such deposits is so interdependent that sometimes one metal, and sometimes the other, is the main product. Osmiridium is obtained mainly in Alaska and in the Witwatersrand of South Africa, along with secondary deposits of gold.

3.5. Recovery of Secondary Platinum Group Metals [69–72]

Platinum metals not only are extracted from ore but,due to their high value, are also recovered from a wide range of industrial residues. These residues are of variable composition and quality, and recovery plants must be very flexible.

Often, the recovery operation is included in the sales contract for semifinished and finished goods, so materials are sent directly for recovery and do not appear on the raw materials market. For regular customers, accounts are kept of the weights of noble metals involved. The supply of material is facilitated, and risks due to price variation are minimized.

Metallic Materials. Large quantities of metallic materials in the form of used platinum – rhodium gauze catalysts result from the oxidation of ammonia (→ Nitric Acid, Nitrous Acid and Nitrogen Oxides). These catalysts must be reprocessed chemically after 3 – 18 months of use. At present, ca. 50 t of Pt–Rh is bound up in these gauzes. The Pd – Au gauzes used to recover Pt – Rh vaporized in ammonia oxidation plants are also recycled.

The glass industry generates large quantities of defective components for chemical recovery (e.g., from melting vessels and other equipment).

Spinnerets from textile fiber manufacture must normally be replaced after about one year of operation owing to erosion of the holes.

Defective laboratory equipment, mainly crucibles and dishes, makes a considerable contribution, but chemical apparatus components are of little significance.

Considerable quantities of platinum – rhodium scrap are provided by the electrical measurement industry, mainly in thermocouple components. Other residues from the electrical and electronic sectors include electrical contacts, heater elements, and electronic components.

Manufacture of fountain pen nibs yields residues in the form of small spheres or dust that can contain ruthenium, osmium, iridium, rhenium, tungsten, molybdenum, tantalum, nickel, and cobalt as alloy components. Treatment of these materials is among the most difficult of all separation techniques.

Dross. Waste materials include slag, ash, furnace residue, corrosion residue from equipment, and precipitation residue. Catalyst residues of poorly defined composition are also included. The PGM content of these materials is usually low; they are extremely variable and usually nonmetallic.

Supported Catalysts [70]. Large quantities of platinum group metals must be recovered from spent catalysts. The most important of these, both in quantity and in value, are the heterogeneous catalysts used in the petroleum industry, especially in reforming processes, where >50 t of platinum is bound up worldwide. The lifetime of these catalysts is 4 – 8 years. Residues usually contain 0.3 – 0.7 % platinum on γ-Al_2O_3, and usually also include palladium, rhodium, iridium, and rhenium, which must also be recovered. The other large-scale processes in the petroleum industry–hydrofining and hydrocracking–yield palladium and platinum catalysts on aluminum silicate carriers.

Another source of material for recovery of platinum metals is spent automobile catalytic converters [69–71]. However, the PGM content is low (2 g per unit), and the units are enclosed in

Table 4. Relative proportions of platinum group metals in selected deposits, and their grades

	Bushveld complex		South Africa,	Sudbury,	Noril'sk,	Colombia	Stillwater,	Average
	Merensky Reef	UG 2 Reef	Plat Reef	Canada	CIS		United States	
Platinum, %	59	42	42	38	25	93	19	45
Palladium, %	25	35	46	40	71	1	66.5	30
Ruthenium, %	8	12	4	2.9	1		4.0	5
Rhodium, %	3	8	3	3.3	3	2	7.6	4
Iridium, %	1	2.3	0.8	1.2		3	2.4	1
Osmium, %	0.8		0.6	1.2		1		<1
Gold, %	3.2	0.7	3.4	13.5			0.5	
Grade, g/t	8.1	8.71	7–27	0.9	3.8		22.3	

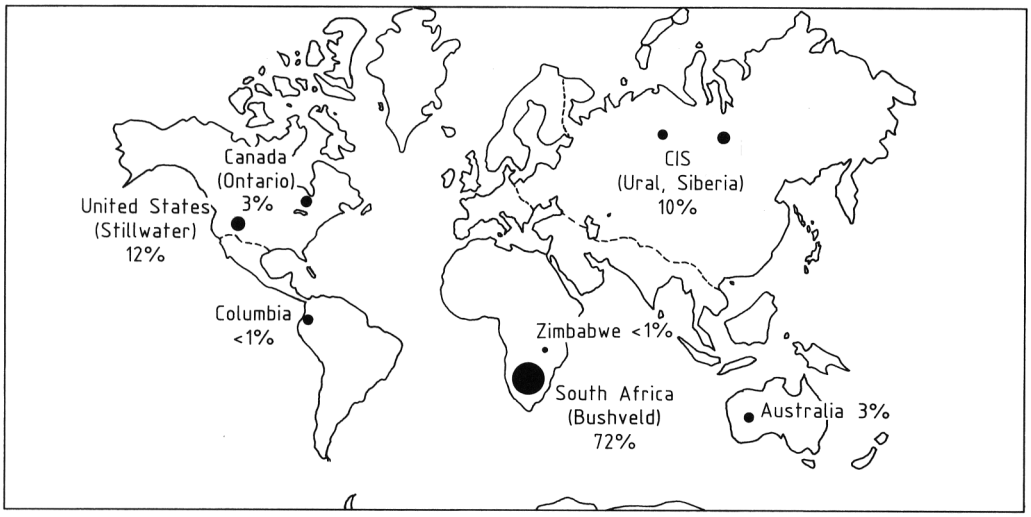

Figure 4. World platinum group metal reserves (total 70 000 t)

steel sheet and widely scattered. Collection and treatment are therefore difficult. A satisfactory and economical solution to this problem has yet to be found. In Germany, collection logistics are coupled to the recycling of automobiles.

The chemical industry produces considerable amounts of palladium catalysts on carbon carriers, often in a moist state that presents sampling problems. Similar residues come from platinum–carbon, rhodium–carbon, and PtO_2 catalysts, although these are sometimes unsupported. Platinum asbestos, which was formerly used widely in sulfuric acid production, is now of very little importance.

Solutions [69]. Amounts of liquid residues from homogeneous catalysts used in the oxo process (hydroformylation) have increased relatively rapidly. The rhodium content of the organic solvents or oily process residues is between 50 and 1000 ppm. Sometimes, these organic solutions contain iridium, ruthenium, or p alladium. Aqueous residues, especially homogeneous catalysts containing rhodium, are currently being produced in increasing quantities.

The electroplating industry yields exhausted electrolytes that cannot be regenerated. The most important of these contain tetranitroplatinate(II), rhodium(III) sulfate, and rhodium(III) phosphate.

Finally, the processes used for separating the platinum group metals also produce waste solutions that must be reclaimed (e.g., mother liquor from crystallization). These operations form part of the separation process.

Radioactive Residues. An as-yet unsolved problem is the treatment of radioactively con-

taminated platinum equipment from chemical laboratories and processes.

Fission of ^{235}U in nuclear power stations produces considerable amounts of platinum group metals; one tonne of spent reactor fuel contains 1.2 kg of palladium, 0.5 kg of rhodium, and 2.3 kg of ruthenium [72].

However, the radioactivity of the material has not permitted the commercial use of this PGM source until now [73]. The most important PGM isotopes in spent reactor fuel are: ^{107}Pd ($t_{1/2}$ 7×10^6 a), ^{102}Rh ($t_{1/2}$ 3 a), and ^{106}Ru ($t_{1/2}$ 1 a). ^{107}Pd is a very low-energy β-emitter, which would not exclude its use in many major applications; alternatively, it may be removed by isotope separation. The active isotopes of rhodium and ruthenium will have decayed to background levels after intermediate storage of ca. 30 years.

3.6. Reserves and Resources

Natural Reserves. Data concerning reserves (discovered by prospecting and having assessable economic value) and resources (which include additional supposed deposits and those with no current economic value) are very dependent on the time prospecting was carried out, and also on technical and economic parameters.

For platinum group metals, the current estimate of workable deposits is 70 000 t [46], [75], [76], 20 years ago, a figure of about one-third of this was assumed [77].

World reserves of the individual metals can best be estimated from the observed compositions of the deposits (Table 4) and the total amounts of PGMs that they contain (Fig. 4). This does not include osmiridium, which usually occurs with gold.

Industrial Residues. Industrial residues have considerable potential for the supply of platinum group metals, in addition to their extraction from ore. In most sectors, the possibilities are now fully exploited. In others, recovery is difficult because of the low PGM contents of the waste materials (e.g., certain catalysts). Economic recovery is also difficult when small PGM-containing components are widely distributed (e.g., in electronics).

In particular, no satisfactory solution has been found to the problem of collecting used automobile exhaust catalysts. In 1991, more than 7 t of platinum was recovered from this source in the United States and Europe and about 0.5 t of rhodium in the United States [248].

A total of ca. 1 t/a of rhodium is currently produced worldwide in nuclear power stations. Some of this is placed in intermediate storage, and some in final repositories.

4. Mineral Dressing and Beneficiation

4.1. Treatment of Alluvial Platinum Deposits [5], [10], [19], [30–32], [36], [41], [43–48], [78–86], [166], [167]

Gangue materials must normally be removed from the platinum-bearing placer deposits. Some deposits are extracted by subsurface mining.

Hydraulic classification can be carried out by hand washing with simple equipment, such as shovels, sieves, and troughs. This was the most common method of treating the Colombian deposits and is still used today. More modern methods involve rotary sieves, troughs, and perforated boxes with running water or water jets.

The most economical process is dredging river placers. This is a simple earth-moving technique linked with a natural inexhaustible water supply. Large installations have capacities of several thousand tonnes of sedimentary material per day, with a water consumption of ca. 10 times this figure. Aboard the dredger, deposits are treated mainly by gravity concentration with sieves and sedimentation equipment, using stirrers, thickeners, washing boxes, etc.

The last stage of concentration is often magnetic separation, with various field strengths used to separate magnetite, ferroplatinum-containing minerals such as chromite, and nonmagnetic components into fractions. Often, a final hand washing is carried out. In some types of deposit, gold particles or gold-containing platinum particles can be separated as amalgam from the platinum concentrate. Another process involves concentrating the platinum metals chemically by dissolving the other components in nitric acid. The concentrates so produced can contain up to

90 % platinum group metals. These can be used directly by refineries.

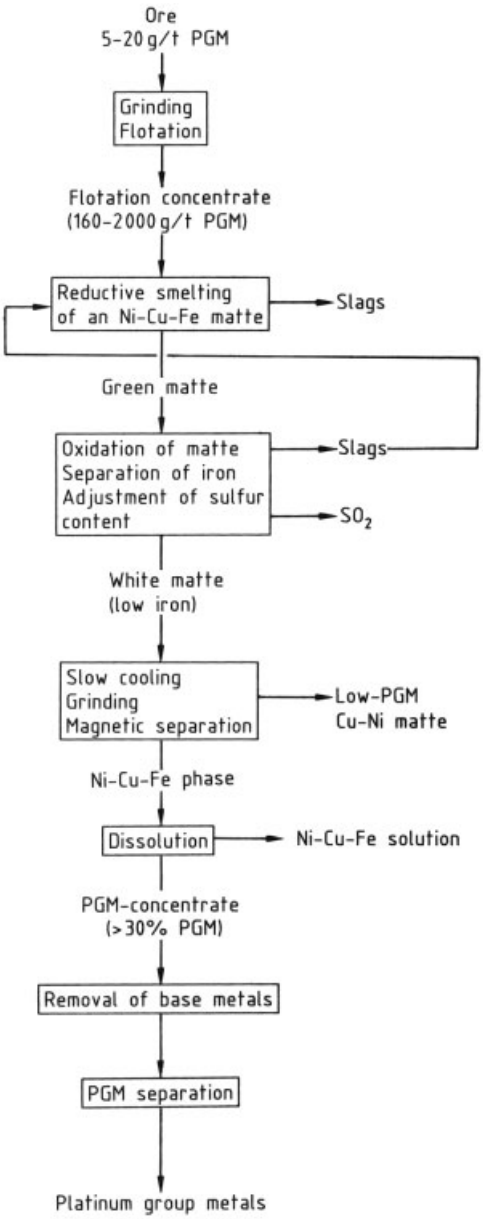

Figure 5. Winning of platinum group metals from sulfide ores

A problem may occur with high losses of platinum carried out as very fine metal from the gravity separation process, often exceeding the amount of the product itself. Losses can be reduced to some extent by recycling the lighter fraction during hydraulic classification of the platinum deposits.

4.2. Treatment of Primary Deposits [46], [56], [66], [67], [86], [87], [166], [167]

The treatment of platinum-bearing rock, which is always supplied in lump form from primary deposits, consists of an initial size reduction by crushing and grinding, usually wet grinding.

After discovery of the South African primary deposits in the Merensky Reef, attempts were initially made to treat this platinum-bearing rock by the methods used for alluvial deposits. Many processes were investigated including gravity concentration, flotation, and metallurgical and chemical processes, such as chloride formation by calcining the powdered ore at 500 – 600 °C in the presence of sodium chloride. The results were unsatisfactory.

The modern process for the winning of PGMs from sulfide ores is shown in Figure 5. Ground ores from the workable oxidation zone, which contain the platinum metals in native form, are first subjected to gravity concentration on cordyroy and James tables or by hydrocyclones, to separate the metallic particles from the platinum-bearing minerals and give a concentrate with a high PGM content that can be processed quickly with low losses.

Flotation is then carried out to remove the gangue from the sulfidic minerals, which are also associated with arsenidic and sulfidic platinum metal compounds and very finely divided elemental platinum metals. This concentrates the platinum group elements by a factor of 10 – 50. After filtration with a rotary filter, the platinum metals in the flotation concentrate are present at a total concentration of several hundred parts per million, along with a small percentage of sulfur, copper, nickel, and iron.

This sequence of process steps is not suitable for all types of deposits and production equipment. Magnetic or electrostatic separation can sometimes be carried out before the smelting operations.

The pelletized material is smelted in a shaft furnace to form a copper – nickel matte. Oxygen is then blown into the converter to oxidize the iron sulfide selectively to iron oxide,

which forms a slag. These two processes concentrate the platinum metals in the copper – nickel matte to > 0.1 %. The blowing operation is controlled so as to give the correct sulfur content for the next concentration stage. In this recently developed, slow-cooling matte separation process, a finely crystalline, homogeneously distributed Ni – Cu –Fe phase, in which the PGMs are concentrated in high yield, is formed in the almost iron-free copper sulfide matte phase (see → Nickel, Chap. 4.4.1.). The product is ground, and the PGM-containing magnetic Ni – Cu – Fe phase is recovered from the PGM-free matte by magnetic separation. Base metals are then removed by treatment with sulfuric acid and oxygen. The concentrate obtained contains 50 – 90 % PGMs. The next stage is to separate the platinum group metals from one another.

The older pyrometallurgical process is time consuming and does not give such good separation. After air blowing the iron, the converter matte produced is smelted with sodium sulfide and separated into copper-containing and nickel-containing layers (tops and bottoms process). The nickel-containing material is roasted and then reduced in a reverberatory furnace to give PGM-containing impure nickel. The platinum group metals are recovered from the anode slime produced during electrorefining. A method of this kind is apparently still used to treat sulfidic PGM deposits in the Urals.

Problems arise during treatment of the high-chromite ores of the UG 2 reef by the sulfidic route; these are therefore often mixed with ore from the Merensky deposit. A process that shows great economic promise for the future is to smelt a metallic concentrate directly from the ore in a plasma furnace [66]. Both types of ore are suitable, although the chromite-containing ore is better.

4.3. Treatment of Nickel Ores [63], [64]

In the processing of sulfide nickel ores, which always contain copper, the platinum group metals follow the nickel in the smelting process. When the crude copper is electrorefined, the amount of PGMs obtained in the anode slime is very small.

When nickel is electrorefined, the platinum group metals remain behind in the slime formed at the nickel anode. The base metals and the silver are dissolved by acid treatment, giving a concentrate that contains ca. 70 % platinum group metals and differs from the nickel anode slime obtained from platinum ores of the Merensky reef in having an appreciably higher palladium content. Also, the concentration of platinum group elements in the unrefined nickel is about 100 times lower than that in the nickel obtained from platinum ore from the Merensky or UG 2 reef (see Section 4.2).

When crude nickel is refined by the carbonyl process (see → Nickel, Chap. 6.3.), residues with a lower PGM content are obtained. These are suitable for concentration by smelting under reducing conditions with lead(II) oxide and sodium carbonate. The lead is driven off, and the silver is dissolved and removed. These concentrates are of high enough quality for separation into individual metals.

The treatment of PGM-containing copper anode slime obtained from nickeliferous pyrrhotite (→ Silver, Silver Compounds, and Silver Alloys) is a long and costly process. The metals copper, selenium, tellurium, arsenic and antimony must first be removed (e.g., by forming their sulfates in a high-temperature process), followed by dissolution or by producing slags via smelting in the presence of potassium nitrate with an air blast. This yields so-called Doré metal. The platinum group metals are then concentrated in the anode slime produced in the subsequent silver electrorefining process (→ Silver, Silver Compounds, and Silver Alloys).

Copper ores can also contain very small amounts of platinum group metals. These appear in the copper anode slime during copper electrorefining.

4.4. Treatment of Metal Scrap [166], [167]

A high proportion of the metallic waste from used equipment and from semifinished products can simply by dissolved without any prior treatment (see Chap. 5). These types of material include crucibles, dishes, thermocouple elements, gauze catalysts, and fiber spinneret nozzles.

Massive materials such as heavy-gauge sheets or bars should be size reduced by crushing or machining to produce swarf. For the highly

refractory metals of the platinum group (Rh, Ir, Ru, and Os) and their alloys, mechanical methods are usually not sufficient to produce a surface susceptible to dissolution. This is also true of platinum alloys with > 30 % rhodium or 20 % iridium. In these cases, alloys of platinum or palladium, which can be dissolved more easily, are preferable.

Very highly dispersed noble-metal black, which is often more soluble, is obtained by alloying the highly refractory platinum metals with base metals and then dissolving the latter out. The noble-metal black must not be heated, or the optimum surface properties for the solution process would be adversely affected. Suitable alloying elements include copper, lead, nickel, zinc, aluminum, bismuth, and silver.

Commercial powdered rhodium and iridium also cannot be dissolved by direct chemical means, but they can be treated with chlorine at ca. 500 – 600 °C to form chlorides (also insoluble), and these can be reduced at low temperature (e.g., by hydrogen or by hydrazine in aqueous suspension) to produce finely divided blacks that can be dissolved in hydrochloric acid – chlorine.

Concentrates of platinum metals often contain oxides that are less soluble in oxidizing acid mixtures than the metals. In these cases, the material must first be reduced by heating in a hydrogen atmosphere or by treating with aqueous hydrazine hydrate at ca. 80 °C. For rhodium oxides, solutions must be highly alkaline and at their boiling point.

4.5. Treatment of Dross [88, 93–96]

If platinum metals cannot be separated by chemical or mechanical methods from accompanying nonmetallic materials, as is usually the case with low-grade waste, pyrometallurgical processes must be used (as in ore treatment).

The most convenient and long-established pyrometallurgical process for low-grade waste is smelting with lead in a shaft furnace to produce slag. Materials containing silver and gold (→ Silver, Silver Compounds, and Silver Alloys; → Gold, Gold Alloys, and Gold Compounds, Chap. 6.2.) are treated in this way. The lead acts as a collector for the platinum group metals. The presence of gold and silver also considerably affects the distribution equilibrium of the platinum group metals in the melt. In the lead shaft furnace, some of the high-melting noble metals rhodium, iridium, and ruthenium separate as so-called furnace shows. When the lead is oxidized and removed as litharge, most of the Rh, Ir, and Ru (so-called bottom metals) precipitates from the increasingly silver-rich alloys. Considerable amounts of rhodium and iridium pass into the shaft furnace slag and are lost. Ruthenium and particularly osmium are lost in large amounts in flue dust and waste gases.

When the gold – silver alloy from the above process is electrorefined to obtain silver, gold and the platinum group metals remain in the anode slime. If this consists mainly of gold, it is converted to pure gold by Wohlwill electrolysis (→ Gold, Gold Alloys, and Gold Compounds, Chap. 5.3.), in which platinum and palladium are concentrated in the electrolyte, and silver chloride and the remaining platinum group metals in the anode slime. Alternatively, the silver can be dissolved from the gold – silver alloy by nitric acid to form silver nitrate, which is purified by thermal decomposition of the accompanying nitrates; the platinum group metals remain in the water-insoluble oxide residue.

All the concentrates mentioned above can be dissolved, and the individual platinum group metals obtained from these solutions.

4.6. Treatment of Supported Catalysts [70], [71], [97–100]

Spent, inactive catalysts consisting of platinum metals supported on active carbon, or carrier-free noble-metal catalysts that have become coated with organic residues, are concentrated by combustion. These materials are sometimes spontaneously flammable. When they are being burned, strong air currents are suppressed to prevent dust losses. If the resulting ash contains platinum metal oxides, these are reduced to the metal.

Catalysts with incombustible carriers insoluble in acid and alkali (e.g., γ-Al_2O_3, silica gel, asbestos, and zeolites) can often be treated with oxidizing acid, but the noble metals dissolve completely only in the absence of organic residues (especially tarry matter) and if the carriers are very porous. Otherwise, the platinum

metals must be concentrated by the lead shaft furnace process (see Section 4.5).

Reforming catalysts consist of γ-Al_2O_3 impregnated with platinum, platinum – rhodium, platinum – iridium, or platinum – rhenium. This carrier material is soluble in acid and alkali, and is dissolved (e.g., in hot sulfuric acid or hot caustic soda solution) leaving the noble metal as an insoluble residue. However, small, but not negligible, amounts of noble metal also go into solution. Moreover, sulfidic impurities in the catalyst can lead to the release of toxic hydrogen sulfide. Other practical processes are dissolving the catalyst in sodium carbonate solution (usually at 220 °C in a pressurized reactor), or sintering with sodium hydroxide or sodium carbonate. Carbon and other products of the breakdown of mineral oil in the spent catalyst must be burned off before treatment with acid or alkali, because these materials interfere with the filtering of the platinum metal concentrate.

Spent automobile exhaust catalysts can be processed at high temperature in a plasma or a submerged arc furnace. In the former, a plasma is produced between an electrode and the molten feed material. The energy of recombination of the plasma is released into the melt. In the resistance furnace, slag serves as the electrical resistance in which heat is produced. In both processes, the oxide carrier is melted with or without addition of a flux of lower the melting point. Iron or copper is added as a collector, forming a metallic melt that takes up the platinum group metals. The concentration of PGMs can reach 20 %.

4.7. Treatment of Solutions [69], [101], [102]

In homogeneous catalysis, high-boiling distillation residues are usually produced that contain no valuable materials apart from rhodium and sometimes ruthenium. These residues can be carefully burned, and the ash treated by wet chemical processes. Other techniques for recovering rhodium have been suggested (i.e., liquid – liquid extraction, reductive precipitation of the metal, and pyrolytic hydrogenation). A process used in industry, especially for the treatment of rhodium-containing oily residues from oxo synthesis, is precipitation of acid-soluble rhodium telluride by reacting the organically bound rhodium with tellurium. This process is notable for the high efficiency of rhodium recovery.

The methods used to produce concentrates from a variety of aqueous wastes are to a large extent the same as those used for the internal recycling of platinum metals in the solutions produced in metal winning processes (see Section 6.8).

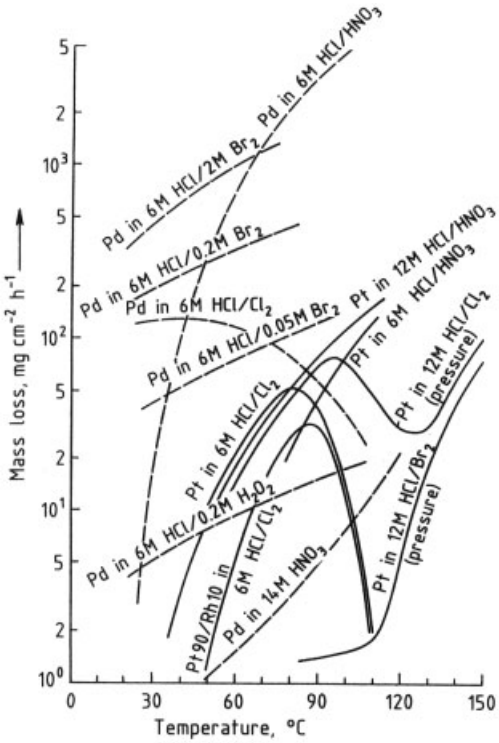

Figure 6. Rate of dissolution of platinum and palladium in oxidizing acids

5. Dissolution Methods [36], [48], [166], [167]

Whether separating platinum group metals or producing compounds and catalysts, the usual starting point is an aqueous solution.

Most raw materials can be dissolved in oxidizing acids. Dissolution can be carried out at atmospheric pressure, or at elevated pressure to obtain higher temperature. Convective or microwave heating can be used.

Highly refractory raw materials, such as rhodium and iridium powder, can be converted into a highly dispersed form, by chlorination at ca. 500 °C, followed by reduction with hydrogen at 100 – 200 °C or hydrazine in aqueous suspension. Alternatively, rhodium can be oxidized to Rh_2O_3 and then reduced.

In some cases, melt processes at higher reaction temperature must be used. A review of the dissolution properties of platinum group metals is given in Figure 6 and Table 3.

The cost of the dissolution step is an appreciable fraction of the cost of the entire process if many small amounts must be dissolved individually for the determination of value; therefore, the choice of method is important.

5.1. Dissolution in Aqua Regia

The following reaction takes place in aqua regia:

$$8\,HCl + 2\,HNO_3 + Pt \rightarrow H_2PtCl_6 + 4\,H_2O + 2\,NOCl$$

as does the following decomposition:

$$3\,HCl + HNO_3 \rightarrow Cl_2 + 2\,H_2O + NOCl$$

The highest rates of dissolution occur at the boiling point of aqua regia. The rate depends only slightly on acid concentration between 6 and 12 M, and is virtually unaffected by the presence of dissolved platinum group metals.

The aqua regia method is preferred for compact metallic platinum and high-platinum alloys (sheet, wire, turnings, etc.). The reactions must be carried out in a sealed apparatus to prevent loss of material by splashing and mist formation. Before the next stage of treatment, nitric acid must usually be removed by concentration with the addition of hydrochloric acid.

To suppress side reactions, the reactants must not be allowed to boil too violently, although maintaining them at the boiling point and use of 6 M acid offers the most convenient method of sustaining consistent reaction conditions. To avoid boiling over, especially during heating, the initial charge of nitric acid should be added slowly and evenly to boiling hydrochloric acid. This method should also be used when aqua regia is added to replace spent acid.

Finely powdered material also tends to cause boiling over. To avoid a violent reaction when dissolving palladium-rich alloys, dilute acids should be used.

5.2. Dissolution in Hydrochloric Acid – Chlorine

In the reaction between platinum and a solution of chlorine in hydrochloric acid

$$Pt + 2\,HCl + 2\,Cl_2 \rightarrow H_2PtCl_6$$

the dissolution rates for platinum and platinum-rich alloys have a maximum in the range 80 – 90 °C (see Fig. 6). As the boiling point of hydrochloric acid (110 °C) is approached, the rate of reaction decreases rapidly, increasing again at higher temperature. For palladium, the rate of dissolution is considerably higher.

For these dissolution reactions, 6 – 8 M hydrochloric acid is preferred since, in this concentration range, both the amount of acid consumed and the amount of hydrogen chloride in the waste gas are lower than if more concentrated hydrochloric acid is used. The presence of dissolved platinum metals does not decrease the rate of dissolution. By careful control of addition rates, a very slight excess of chlorine can be used, so that this method leads to lower levels of waste gas contamination than the aqua regia method. However, the rates of dissolution decrease considerably in the absence of agitation, and the process is most suitable for easily stirred powders, slurries, or concentrates and unsuitable for bulk material.

Metal dissolves more slowly in the vapor phase or in refluxing hydrochloric acid containing dissolved chlorine (see Fig. 6).

Hydrogen peroxide can be used in place of chlorine as oxidizing agent. Optimum concentrations are difficult to maintain owing to the decomposition of hydrogen peroxide into oxygen and water; therefore, reaction rates are very low compared with those for $HCl - Cl_2$. The $HCl - H_2O_2$ system has some limited use for palladium.

5.3. Dissolution in Hydrochloric Acid – Bromine

Palladium dissolves more rapidly in hydrochloric acid – bromine mixtures than in hydrochloric acid – chlorine (see Fig. 6). Bromine is more

soluble than chlorine in hydrochloric acid, enabling high halogen concentrations to be used. If the temperature is kept well below the boiling point of bromine (59 °C), losses in the waste gases become very small. Bromide formed in the dissolution reaction can be oxidized to bromine by chlorine in an easily controllable reaction, recovered by distillation, and recycled. Since the conversion of bromide to bromine

$$2\,Br^- + Cl_2 \rightarrow Br_2 + 2\,Cl^-$$

or

$$H_2PtBr_6 + 3\,Cl_2 \rightarrow H_2PtCl_6 + 3\,Br_2$$

can take place in parallel with the dissolution process, bromine can be present in substoichiometric amounts.

Platinum and its alloys with rhodium or iridium are more resistant to hydrochloric acid – bromine than to hydrochloric acid – chlorine.

5.4. Other Dissolution Processes

Concentrated nitric acid is suitable for dissolving palladium (see Fig. 6).

Concentrated sulfuric acid dissolves finely divided rhodium at ca. 300 °C, but strongly heated rhodium powder is largely insoluble in sulfuric acid.

Concentrated hydrobromic acid is the only reagent that can directly dissolve oxides of the platinum group metals such as PdO, PtO_2, Rh_2O_3, and IrO_2. Platinum can be dissolved in a mixture of hydrobromic acid and bromine above 100 °C in a pressurized reactor. Metallic rhodium is quite soluble in concentrated hydrobromic acid near its boiling point.

Powdered ruthenium or osmium can be treated with an alkaline solution of potassium peroxodisulfate ($K_2S_2O_8$) to form solutions of ruthenate or osmate. In acidic media, osmium is oxidized by peroxodisulfate or chromic acid to form OsO_4.

Perchloric acid has been reported to be a solvent for platinum and its alloys, ruthenium, and osmium.

5.5. Dissolution by Salt Fusion [34]

Aqueous dissolution processes are not successful for all types of concentrates and raw materials. Often, reactions in molten salts are more effective. For many raw materials, molten salt methods are selective and therefore also suitable for separation.

Rhodium can be converted into water-soluble rhodium(III) sulfate at ca. 600 °C by melting with potassium or sodium hydrogensulfate, which is converted to the pyrosulfate with loss of water. Platinum, iridium, and ruthenium are not attacked.

Ruthenium is converted into water-soluble $K_2[RuO_4]$ when reacted with $KOH - KNO_3$. Osmium reacts similarly. This reaction can be used to treat finely divided osmiridium, especially after preliminary treatment with molten zinc, in which the iridium remains undissolved (see also Section 4.4). Potassium salts are more effective than sodium salts. Alkali-metal carbonates can be used instead of alkali-metal hydroxides.

Ruthenium and osmium are very readily attacked by molten Na_2O_2. The reaction can be prevented from becoming too violent by addition of NaOH or Na_2CO_3 to reduce the melt temperature. Iridium can be oxidized to acid-soluble iridate by fusion with Na_2O_2.

Iridium can also be converted to acid-soluble iridate by treatment with fused BaO_2 or BaO_2 – $Ba(NO_3)_2$. This has the advantage of enabling the cation to subsequently be removed simply by precipitation as $BaSO_4$. The process is also used to treat osmiridium, although osmium is converted to OsO_4 and must be recovered from the waste gas.

When mixtures of powdered platinum group metals are heated at 700 °C in a chlorine atmosphere, acid-soluble chlorides of palladium and platinum, and acid-insoluble $RhCl_3$, are formed. Chlorination of iridium and ruthenium produces substoichiometric acid-insoluble products.

All of the platinum group metals form water-soluble complex chlorides when treated with sodium chloride in a chlorine atmosphere.

6. Separation of Platinum Group Metals [31], [103], [263]

After the platinum group metals have been dissolved, the individual metals must be recovered. Depending on the raw material, the solution may contain all or some of the PGMs, together with gold, silver, and base metals.

Over the past 200 years, numerous separation processes have been developed and used [31], [32], [36], [41]. Many of the individual steps are still very important in modern processes. Others are of historical interest, but may be used in special cases to treat very small quantities. In addition to the large-scale separations carried out by mining companies and refineries a number of small companies exist.

In general, coarse separation is followed by a purification stage. The process used for coarse separation is determined largely by the composition of the starting solution, and the purification process depends on the particular PGM. The purification stage is necessary because, with few exceptions, an individual platinum group metal of commercially acceptable purity cannot be isolated from complex solutions in a single step.

In practice, the composition of the raw material changes frequently and profoundly, especially in the recovery of secondary metal. Here, the crude separation process must be very flexible. In separation plants that are parts of mining operations, the raw material is usually fairly consistent over a long period of time.

The most important separation processes today make use of various combinations of precipitation, crystallization, solvent extraction, and distillation.

To achieve the necessary performance economically, a high selectivity, a good yield of noble metal, and (to minimize loss of the interest-earning potential of high-value materials) a short treatment time are required.

6.1. Chemistry of Platinum Group Metal Separation [104–115]

Of the many known reactions in PGM chemistry, certain ones have been developed to separate the metals or have been shown to be especially useful for this purpose. This is true both for crude separation and for purification; in many cases the same reaction is equally important in both stages.

Dissolution Properties. Differences in reactivity toward chemical reagents, especially oxidizing acids and molten salts, can often be utilized for coarse separation of platinum group metals. Good separation is possible only if platinum and palladium do not form mixed crystals with the refractory PGMs.

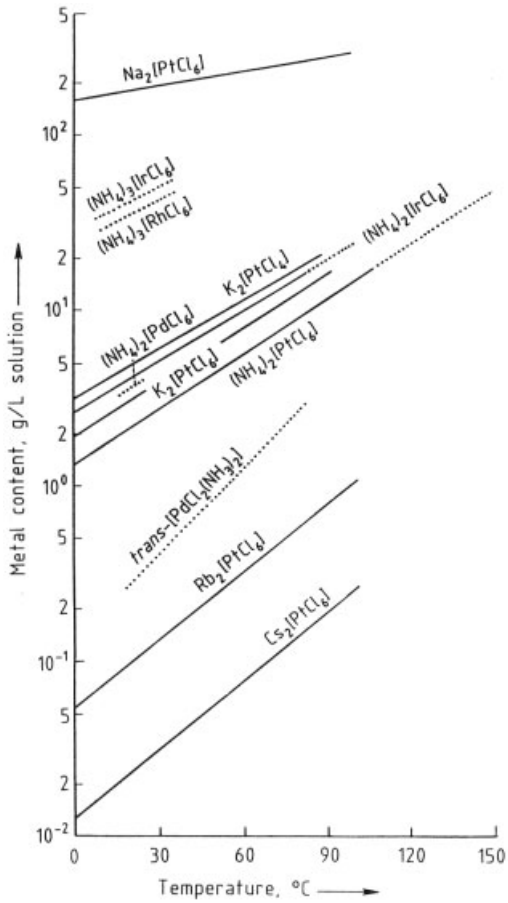

Figure 7. Solubility of platinum group metal chloro complexes in water as a function of temperature
. . . . unstable solution equilibrium; —- stable solution equilibrium

Solution Equilibria. In the crystallization process (by either evaporation or precipitation), the solubilities of chloro complexes of the platinum group metals are very important, as is the possibility of altering these solubilities. The

temperature dependence of these solubilities is generally quite large and can be utilized or must be considered, in many stages of the process. In the range between room temperature and 100 °C, solubilities differ by about a factor of 10 (Fig. 7). An even more useful phenomenon in separation technology is the effect on solubility of the addition of a common ion, which is utilized in precipitation crystallization (Sections 6.3 and 6.4). Solubility can usually be varied over two orders of magnitude (Fig. 8). The solubility of AgCl is affected not only by the concentration of the acid, but also by that of H_2PtCl_6.

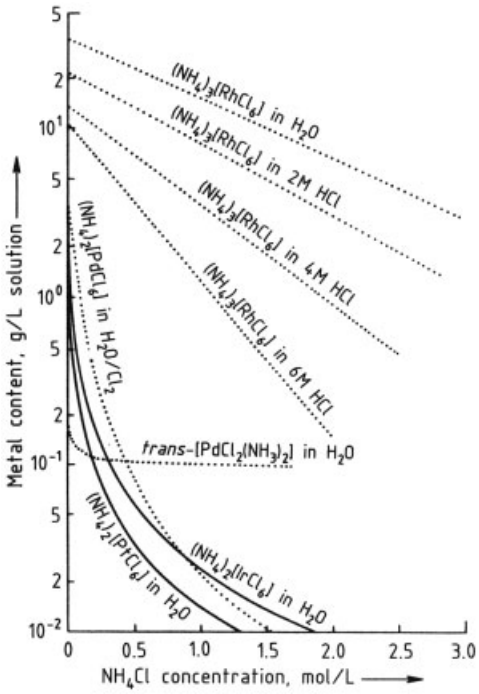

Figure 8. Dependence of the solubility of platinum group metal chloro complexes on the concentration of NH_4Cl and HCl at 20 °C
. . . . unstable solution equilibrium; —- stable solution equilibrium

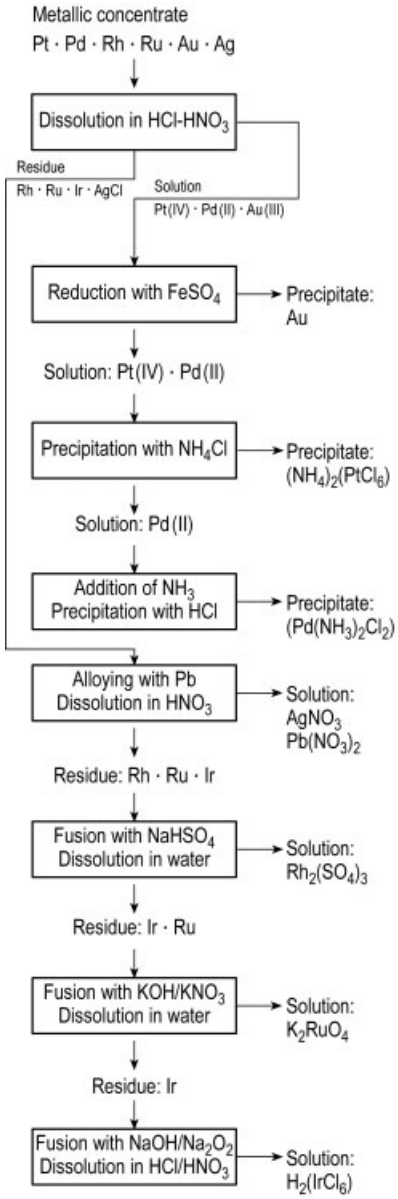

Figure 9. Older separation process A

Redox Reactions and Stability of Complexes [88], [116]. Many separations are made possible by valence changes in redox systems. This applies mainly to precipitation crystallization, but also to solvent extraction and distillation. In practice, the kinetic and thermal stabilities of the valences in complexes are important, as is the redox stability (Table 5).

Hydrolysis [96], [104], [117]. Hydrolytic effects can frequently reduce the completeness of a desired reaction, thereby adversely affecting separation. Most chloro complexes of PGMs are susceptible to hydrolysis at high temperature, thereby reducing their redox stability. The possible stepwise hydrolytic conversion of $[RhCl_6]^{3-}$ to $[Rh(H_2O)_6]^{3+}$, and of $[PtCl_6]^{2-}$

Table 5. Stability of platinum group metal chloro complexes

Metal	Oxidation state	Complex	Redox stability	Kinetic stability	Thermal stability
Pt	IV	$[PtCl_6]^{2-}$	stable	very stable	very stable
Pt	II	$[PtCl_4]^{2-}$	unstable	unstable	
Pd	IV	$[PdCl_6]^{2-}$	unstable	stable	unstable
Pd	II	$[PdCl_4]^{2-}$	stable	very stable	
Ir	IV	$[IrCl_6]^{2-}$	stable	stable	
Ir	III	$[IrCl_6]^{3-}$	stable	stable	
Rh	III	$[RhCl_6]^{3-}$	stable	stable	unstable
Ru	IV	$[RuCl_6]^{2-}$	unstable		unstable

to $[Pt(OH)_6]^{2-}$, must be taken into account. Similar hydrolytic effects are also common with ruthenium and iridium. The reactions outlined here generally take place with widely varying rates. This is true for reaction in both directions (e.g., for hydrolysis and for the reverse reaction in the presence of acid). For example, when a suspension of $(NH_4)_2[PtCl_6]$ is heated, the concentration of dissolved platinum (after being cooled to 20 °C) increases in 24 h from 2.2 g/L to twice this amount, and after 96 h it increases fourfold. This is due to partial hydrolysis to give soluble compounds. The reverse reaction, brought about by heating with hydrochloric acid at its azeotropic concentration, occurs at similar rates.

Ion Exchange Chromatography. The use of ion exchange is limited mainly to the separation of PGMs from base metals [118]. For example, selective adsorption of $[PtCl_6]^{2-}$ on strongly basic anion-exchange resins is possible, followed by elution with sodium hydroxide solution, yielding $[Pt(OH)_6]^{2-}$. The separation of a PGM mixture into individual PGMs by chromatography is described in [119], [120].

Organic Precipitation Agents [121], [122]. Palladium can be selectively precipitated by dimethylglyoxime, but the precipitate is difficult to filter. Separation systems have also been reported that depend on the formation of complexes of the PGM with substituted formazanes and substituted benzoylthiourea. Precipitable organometallic compounds can often be recovered by solvent extraction instead of precipitation.

Distillation and Sublimation. The only industrial use of distillation or sublimation at present is the separation of RuO_4 and OsO_4, usually by steam distillation. These substances have high vapor pressures even at room temperature, so a noncondensable carrier gas cannot be used because of the high carryover of material. A possible future process is the separation of PGM by using sublimation of PGM fluorides [123].

Reduction to Metal. Since the PGMs are all electrochemically noble metals, selective reduction and cementation by base metals are not possible. In the past, collective cementation from aqueous solution by zinc was often an important step, both for separation and for the recovery of PGMs from recycled solutions. However, the use of zinc as a cementation agent is now ruled out for environmental reasons. Where cementation cannot be avoided, iron, aluminum, or Fe – Al alloys can be used instead. In aqueous media, hydrazine, formate, or boranate can be used to reductively precipitate elemental PGMs.

Solvent Extraction [124–133]. Many liquid – liquid extraction systems have been described that can be used for solvent extraction of metals. Much research into the platinum group metals has been carried out, mainly aimed at their separation. The extraction processes do not usually involve a true Nernst distribution of the extractable compounds between the organic and the aqueous phases. More often, the extractable compounds (generally organometallic) are formed in the extraction system itself. The reactants are dissolved in an inert organic phase or, if they are themselves liquids, can be used as such. The system is generally diluted with an inert solvent to lower the viscosity.

The extractants and extraction mechanisms are classified in the following groups:

1) Compound formation

2) Anion exchange
3) Cation exchange
4) Solvation
5) Solvent extraction without reaction

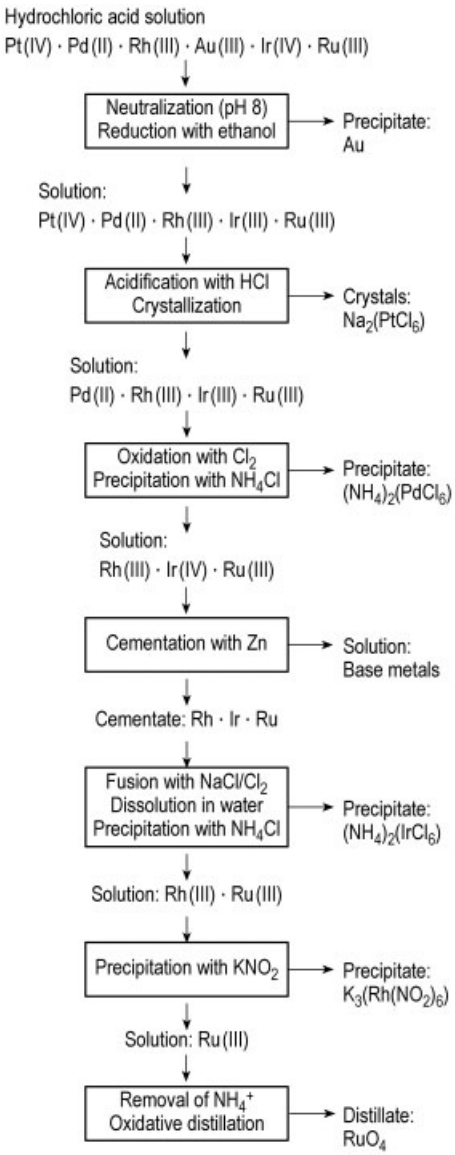

Figure 10. Older separation process B (Siebert)

Of the many systems investigated, very few are of practical use.

The organic phase often increases significantly in viscosity as metal content increases and is usually diluted by a hydrocarbon mixture in the 150 – 200 °C *bp* range.

The extracted metal is stripped from the organic phase by an aqueous phase.

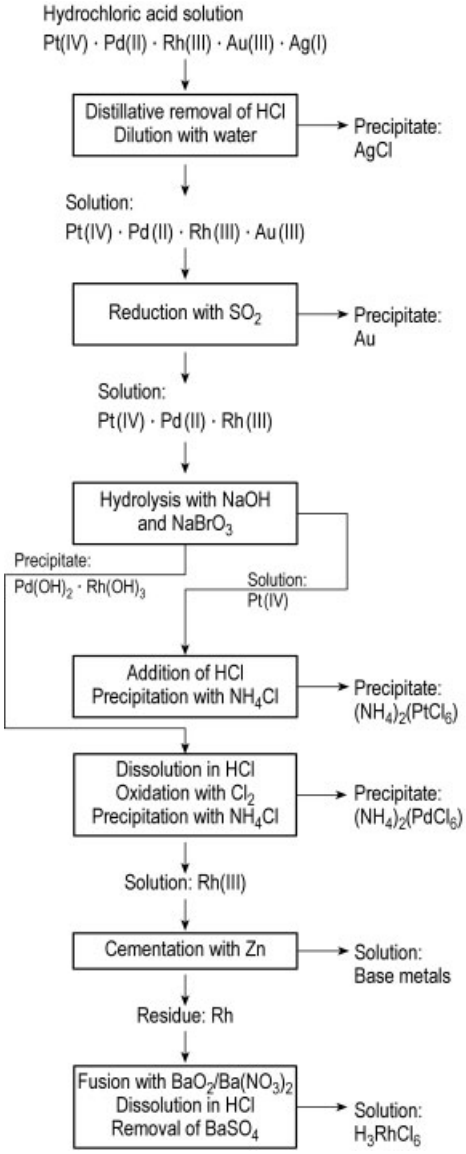

Figure 11. Older separation process C (Gilchrist)

6.2. Older Separation Processes [31]

Some separation methods, shown in Figures 9, 10, 11, were commonly used for separating PGMs in earlier times. Modifications of these may still be used today or included with more modern individual stages in an integrated pro-

cess. In special cases, they can contribute to the solution of separation problems.

Separation Process A (Fig. 9) [36], [43]. The process used by INCO in the Acton refinery in England has been known for a long time. It is especially suitable for treatment of the anode slime from the electrolysis of nickel. Here, selective dissolution of Pt – Pd can be achieved since the refractory PGMs are concentrated in separate particles. Some special process steps are used, including alloying with lead, and selective dissolution of rhodium, ruthenium and iridium by salt fusion.

Separation Process B (Fig. 10). A process formerly used by the platinum smelting company Siebert/Degussa is characterized by the recycling of unacceptably large amounts of platinum due to the relatively high solubility of sodium chloroplatinate. Recovery of the rarer platinum group metals by cementation led to serious pollution of the wastewater. The cemented metals were usually dissolved with $NaCl - Cl_2$. An advantage of the process was the efficient separation of iridium.

Separation Process C (Fig. 11) [36], [134], [135] is suitable only for raw materials with low palladium, rhodium, and iridium content, because of the difficulty of handling their hydroxide precipitates, which have adsorptive properties. It has been used only in small separation plants for secondary metal and yields very pure platinum.

6.3. Current Separation Processes

Modern separation processes (see Figs. 12, 13, 14) are designed for high separation efficiency, minimum recycle, reduced holdup times of expensive PGMs, and minimization of losses of noble metals. Ecological considerations are often crucial, even when costs must be kept low. Thus, cementation on zinc, an extremely expensive process even if the cost of necessary wastewater treatment is excluded, has to a very large extent been eliminated from PGM separation processes for environmental reasons.

Separation Process D (Fig. 12) [31], [88]. All process steps can be carried out on the same solution by using simple, similar process operations, which is economically beneficial. The process is very flexible, which is an advantage, particularly for the recovery of secondary metal when the ratio of PGMs changes continuously. If the ruthenium content is high, RuO_4 should be removed by distillation as the first step of the separation process.

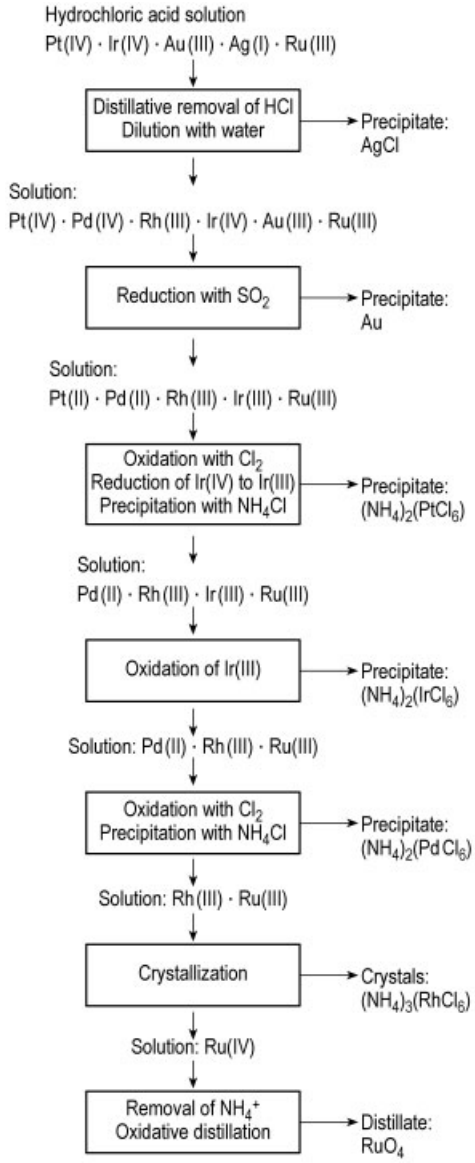

Figure 12. Modern separation process D

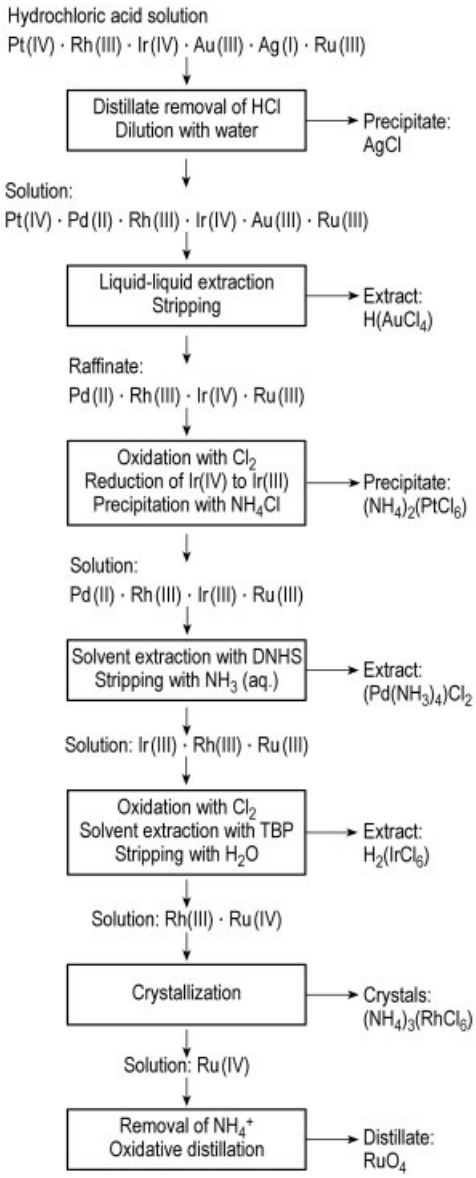

Figure 13. Modern separation process E

Separation Process E (Fig. 13). Solvent extraction with dialkyl sulfide is used. Unlike process D, this provides a practically quantitative and sharp separation of palladium, which yields favorable conditions for the purification of palladium and platinum. Solvent extraction of iridium can be carried out almost quantitatively and is very useful when iridium content is low. Under these conditions, use of precipitation crystallization would result in excessive amounts of iridium remaining as soluble $(NH_4)_2[IrCl_6]$ in the mother liquor.

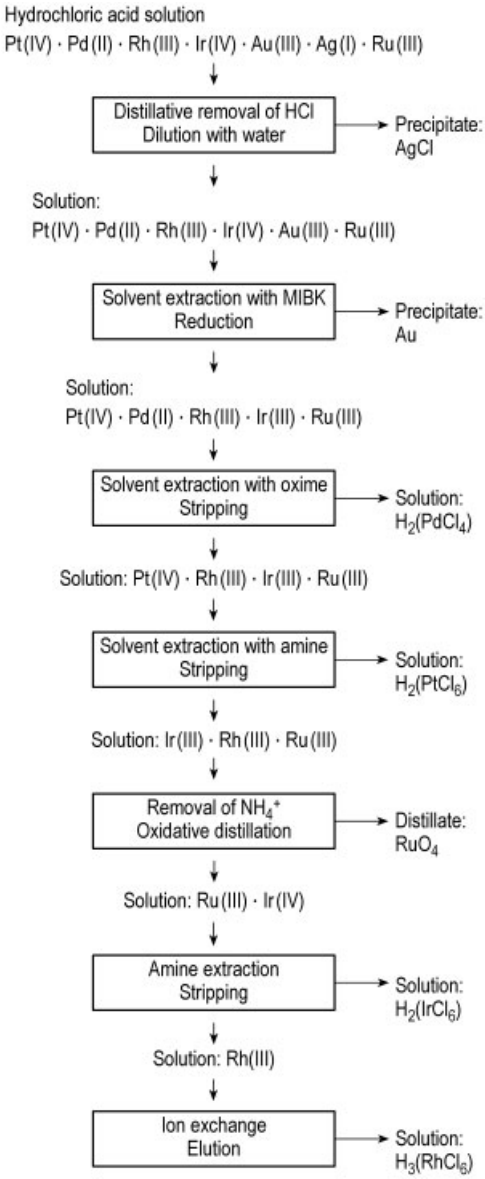

Figure 14. Modern separation process F (Matthey Rustenburg refiners)

Separation Process F (Fig. 14) [136]. A process used mainly in mining operations (i.e., with primary raw materials) has, as its most notable feature, solvent extraction of platinum, which is the metal present in highest concen-

tration. Oximes are better than thioethers for solvent extraction; the higher rate of complex formation enables column technology to be used. Also, the separation process can be set up as an integrated unit. Ion exchange, which is also carried out on columns, results only in separation of base metals.

Similar separation processes based on solvent extraction are in industrial use or development [129–133], [137–143].

6.4. Processes Used in Coarse Separation [31]

Separation of Silver. Insoluble silver chloride is formed in the dissolution process and is removed by careful filtration from the starting solution in concentrated hydrochloric acid. However, this acid at its usual concentration leads to unacceptably high levels of silver remaining in solution due to $H[AgCl_2]$ formation. Lowering the free hydrochloric acid concentration, preferably by evaporation and dilution, considerably reduces the solubility of silver chloride. The settling rate and ease of filtration are improved by flocculating agents.

Traces of silver (usually together with Cu and Pd) can be extracted from chloride solutions of platinum group metals at pH 5.5 by dithizone in chloroform.

Separation of Gold. The usual process for gold separation involves reducing Au (III) to elemental gold in acid solution. Reducing agents that can be used include iron(II) salts, oxalic acid, sulfur dioxide, and ascorbic acid. Sodium nitrite, hydrogen peroxide, sodium formate, and ethyl alcohol are also used [133]. The separation process is improved if reducing agents and reaction conditions are chosen such that Pt (IV) is not reduced to Pt (II). Otherwise, Pt (II) must be reoxidized to Pt (IV). The separation of gold from PGMs is becoming increasingly important [143].

Selective Dissolution Reactions and Melt Reactions. Dissolution processes usually do not give very sharp separation of platinum group metals.

Problems with salt fusion lead to poor diffusion conditions. In dry chlorination of platinum group metals, the chlorides are thermodynamically stable over only a small range. Also, in reaction of the metals with sodium chloride and chlorine, temperatures must be chosen such that unreacted metal particles are not blocked by molten salt. In practice, achieving complete reaction in a single stage is usually impossible.

Separation of Platinum, Iridium, and Palladium by Precipitation Crystallization [144–146]. The same basic process has been used for the separation of the largest quantities of platinum group metals since the beginning of separation technology–precipitation crystallization of ammonium hexachloro complexes $(NH_4)_2[MCl_6]$. This process can be optimized in many ways by influencing solubilities (e.g., by valence changes, addition of a common ion, or changing the temperature and rate of precipitation).

The first step in separation is usually to precipitate platinum as $(NH_4)_2[PtCl_6]$. If the dissolved platinum is present as Pt^{II}, it must first be oxidized to Pt^{IV} (e.g., by chlorine). Excess dissolved chlorine is driven off by boiling, which also causes the palladium that had been oxidized to palladium(IV) to be reduced to palladium(II). Any iridium(IV) present must be selectively reduced to iridium(III), for example, with iron(II) salts or ascorbic acid. The ammonium hexachloroplatinate(IV) is precipitated preferably by slow addition of concentrated ammonium chloride solution at room temperature with stirring. The hydrochloric acid concentration should be 1 M, and the platinum content should be 50 – 200 g/L. After precipitation, the excess of NH_4Cl in the solution should be at least 0.5 mol/L. The mother liquor is removed from the salt by washing with semisaturated cold ammonium chloride solution. The deep-yellow ammonium hexachloroplatinate(IV) is recovered with a purity of 99.0 – 99.5 % in $>$ 99 % yield.

Attempts have been made to precipitate the $(NH_4)_2[PtCl_6]$ slowly by gradual formation of NH_4^+ ions in the reaction medium (e.g., by hydrolysis of urea or hexamethylenetetramine). In this way, an almost ideally homogeneous precipitation medium is produced. However, the purities and yields achieved do not justify the cost.

To precipitate iridium as $(NH_4)_2[IrCl_6]$, iridium(III) is converted to iridium(IV) by treatment with an oxidizing agent at ca. 100 °C, while

palladium remains in the divalent state at this temperature. Black $(NH_4)_2[IrCl_6]$ is filtered off from the cold solution. It is more readily soluble than the corresponding platinum salt, and therefore sufficient excess of NH_4Cl (0.5 – 1.0 mol/L) and good cooling (20 °C) must be used, and the minimum amount of concentrated NH_4Cl solution must be used for washing. Chlorine is then passed into the solution at room temperature, the excess NH_4Cl in solution being > 1 mol/L, and palladium is precipitated in the pure state in the form of the sparingly soluble, brick-red ammonium hexachloropalladate(IV), $(NH_4)_2[PdCl_6]$. The separation from rhodium(III) is, however, not sharp.

The main problem encountered in this precipitation is the difficulty of producing sufficiently large crystals for filtration, which can be favored by adding the palladium solution to the NH_4Cl solution. High temperature favors crystal growth but must be avoided to prevent the following reaction from occurring:

$$[PdCl_4]^{2-} + Cl_2 \underset{60\,^\circ C}{\overset{20\,^\circ C}{\rightleftharpoons}} [PdCl_6]^{2-}$$

A virtually quantitative yield is obtained if the crystals are isolated quickly and washed with a cold concentration NH_4Cl solution saturated with Cl_2.

The same basic separation process can be carried out by using potassium chloride instead of ammonium chloride. Using potassium chloride has advantages and disadvantages:

- As the solubilities of potassium hexa chloro salts are lower than those of corresponding ammonium salts, separation yields are better
- Precipitation of $(NH_4)_2$ $[PdCl_6]$ by passing chlorine through an NH_4^+-containing solution is dangerous, because explosive nitrogen trichloride is formed. This can be oxided by using potassium chloride [147]
- Potassium ions are introduced into the purification processes (see below) and are more difficult to remove than immonium ions

Separation of Rhodium. Ammonium hexachlororhodate, $(NH_4)_3[RhCl_6]$, can be isolated from rhodium solutions in hydrochloric acid such as those obtained after removal of the other platinum group metals. Ammonium chloride is added, followed by concentration and crystallization. Much of the remaining $(NH_4)_2[PtCl_6]$ and $(NH_4)_2[IrCl_6]$ is precipitated at the same time. Chlorides of the base metals mostly remain in solution. The $(NH_4)_3[RhCl_6]$ can be selectively redissolved in water at room temperature to form an almost saturated solution. This gives a purified solution of rhodium. Both $(NH_4)_2[PtCl_6]$ and $(NH_4)_2[IrCl_6]$ are recovered by filtration.

The rhodium can also be isolated as chloropentamminerhodium(III) chloride, $[RhCl(NH_3)_5]Cl_2$, or potassium hexanitrorhodate(III), K_3 $[Rh(NO_2)_6]$, but the precipitations must be carried out in weakly alkaline solution. These methods are unsuitable for solutions that contain large quantities of base metals because, under reaction conditions, these form insoluble hydroxides that are difficult to filter.

Cementation of rhodium by other metals (see Section 6.3) can also be carried out. However, all of the noble metals and the copper remain with the rhodium. An advantage is that precipitation of rhodium is complete in this process.

Distillation of Ruthenium(VIII) Oxide. The most important industrial process for the isolation of ruthenium is distillation of the very volatile ruthenium(VIII) oxide, RuO_4, from aqueous solution. This compound is formed in solutions of potassium ruthenate(VI), $K_2[RuO_4]$, and hexachlororuthenate(III), K_3 $[RuCl_6]$, by oxidation with chlorine at around neutral pH or with sodium chlorate at pH 0 – 3. Oxidation can also be carried out in dilute sulfuric acid solution with potassium permanganate, or with potassium chlorate in sulfuric acid solution, but these methods are less satisfactory for safety reasons. The volatile RuO_4 is absorbed by dilute hydrochloric acid, and is converted into water-soluble chlororuthenate complexes.

Even in the absence of potassium permanganate, chlorate, or their reaction products, the distillation of RuO_4 is hazardous. The presence of NH_4^+ salts can lead to the formation of explosive chlorides of nitrogen. Therefore, removing ruthenium at the start of a separation process is often preferable. The formation of chlorine dioxide, chlorate, and perchlorate is also possible. Moreover, RuO_4 can explode by spontaneous decomposition, especially at > 100 °C and on contact with organic substances. To im-

prove safety, work should be carried out under an inert gas and by avoiding the presence of large quantities (e.g., using a continuous process).

Distillation of Osmium(VIII) Oxide. As a rule, osmium is present in only a few primary and secondary raw materials. Osmium should, if possible, be isolated as the first step of a separation process, so that later operations will not lead to loss of volatile OsO_4.

Solutions obtained from digestion of ores normally contain osmates. Treatment of solutions in dilute sulfuric acid with oxidizing agents such as CrO_3, $K_2S_2O_8$, or HNO_3 causes osmium(VIII) oxide to be formed and to distill off. From this distillate, OsO_4 can be isolated directly as a water-insoluble oil (*mp* 40 °C). Alternatively, it can be absorbed in potassium hydroxide solution, forming K_2OsO_4, or can be converted to metallic osmium by reaction with formaldehyde solution. Owing to the toxicity of OsO_4, its escape into the atmosphere must be prevented by use of the best possible absorption equipment. No danger of explosion exists.

Also, OsO_4 can be formed and distilled off by heating fine osmium powder in a stream of oxygen.

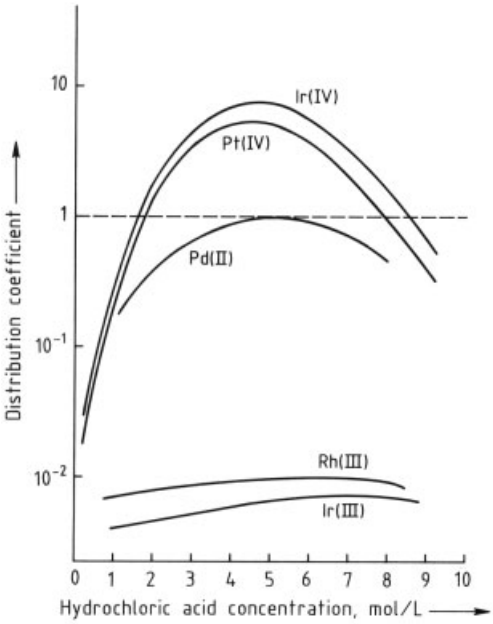

Figure 15. Extraction of platinum group metal chloro complexes by tributyl phosphate

Solvent Extraction. Liquid – liquid extraction is often characterized by distribution coefficients that vary greatly from element to element. These can usually be modified by chemical methods to give conditions favorable for the separation of PGMs. However, the separation factors resulting from the ratios of the distribution coefficients rarely enable a single separation stage to provide purities that fulfill modern requirements for metal quality. Although separations are generally better than those effected by precipitation crystallization, coarse separation by solvent extraction must be followed by purification.

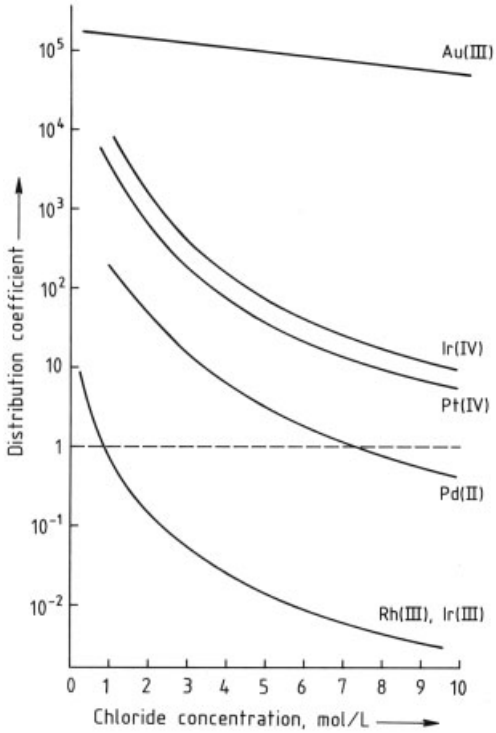

Figure 16. Extraction of platinum group metal and gold chloro complexes by trioctylamine

In practical industrial separation processes, mainly four extractants are used [109], [130–133], [143–146]. The most long-standing process is the separation of Pt(IV) and Ir(IV) from Rh(III) with tributyl phosphate (TBP) [148], [149]. The extraction of $H_2[PtCl_6]$ and $H_2[IrCl_6]$ must be carried out in 4 – 6 M HCl (Fig. 15). If the usual mixture of 1 part TBP with 3 parts hydrocarbon solvent is used, the capac-

ity of the organic phase is limited in practice to ca. 10 g of Pt – Ir and it is therefore less suitable for extraction treatment of solutions whose main constituent is platinum (e.g., normal refinery feed solutions).

Trioctylamine is used widely for the solvent extraction of large amounts of platinum from refinery solutions [150–152]. It is also used for the solvent extraction of Ir (IV) (Fig. 16). Stripping can be facilitated by changing the stripping solution or the valence of the metal. Diluent solvents are also employed when carrying out solvent extraction with amines.

For the solvent extraction of palladium, thioesters are currently used (e.g., di-*n*-hexyl sulfide or di-*n*-octyl sulfide) [110], [151–155]. They produce good separation from other PGMs and have a high capacity for the metal. A disadvantage is the poor kinetics of complex formation, the reaction times required being 30 mins or more. Hence, mixer – settler equipment must be used rather than column equipment. Stripping with aqueous ammonia produces $[Pd(NH_3)_4]Cl_2$; this can be converted easily to $[Pd(NH_3)_2Cl_2]$, which is then purified.

The method used for high-tonnage production of copper by solvent extraction with oximes is also successfully used for palladium (Fig. 17). However, high concentrations of palladium in the organic phase cannot be achieved.

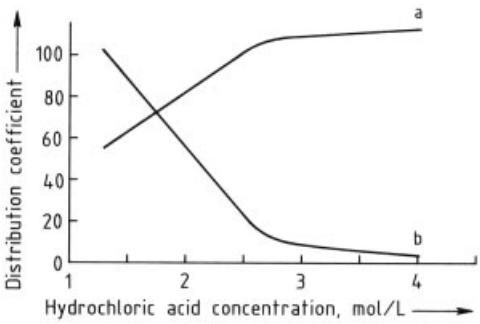

Figure 17. Extraction of palladium(II) and platinum(IV) chloro complexes by an aldoxime (LIX TN 1911)
a) Platinum(IV); b) Palladium(II)

Numerous other solvent extraction systems have been investigated, some of which are undergoing industrial development [114], [115], [149], [156–159].

6.5. Purification

Methods used for purification of the products from the separation process employ the same principles (i.e., valence change, temperature change, addition of common ions, and modification of complexes by chemical reaction).

Purification of Platinum. Recrystallization of sodium hexachloroplatinate(IV), $Na_2[PtCl_6]$, from hot water is effective, but large amounts of platinum must be recycled.

Very effective purification is achieved by the oxidative hydrolysis of the metallic impurities palladium, rhodium, iridium, ruthenium, and base heavy metals, which are removed as insoluble hydroxides (see Fig. 11). Precipitation from hot aqueous solution is carried out by addition of sodium bromate at pH 6.5 or by passing through chlorine at slightly alkaline pH. The oxide hydrates adsorb considerable amounts of platinum. The yield of platinum therefore depends on the purity of the feed material.

The classical method of purification is by repeated precipitation of $(NH_4)_2[PtCl_6]$ or $K_2[PtCl_6]$. The crude salt is converted into metallic platinum, dissolved to form $H_2[PtCl_6]$ (see Chap. 5), and reprecipitated. By repeating this many times, very high purities are obtained, although the cost is also high. Also, $(NH_4)_2[PtCl_6]$ can be converted directly to H_2PtCl_6 by oxidation of the NH_4^+ or by cation exchange. In the oxidative decomposition of NH_4^+ (e.g., by heating the aqueous solution with addition of Cl_2 or HNO_3), care must be taken to use a carrier gas or distillative conditions to prevent accumulation of explosive nitrogen – halogen compounds.

Conventional crystallizations from water are possible, but the solubility of $(NH_4)_2[PtCl_6]$ is relatively low, so that despite favorable temperature coefficients (see Fig. 7), large volumes must be used. Modified crystallization methods are more economic (e.g., utilization of the considerably higher solubilities in water above 100 °C at higher pressure). The purification effected by crystallization is enhanced by hydrolysis. In neutral or very slightly acidic media above 90 °C, the conversion of the chloro complexes of the platinum group metals (especially metals other than platinum) into hydroxo complexes and oxide hydrates begins. These cannot

form mixed crystals with $(NH_4)_2[PtCl_6]$. The extent of hydrolysis of hexachloroplatinate(IV) increases with increasing temperature and reaction time. Therefore, high temperature and long reaction time lead to a lower yield of crystalline product. The reaction is reversible in the presence of hydrochloric acid.

When the products of solvent extraction by substituted amines are stripped, solutions of platinum compounds in hydrochloric acid are obtained. From these, pure $(NH_4)_2[PtCl_6]$ is precipitated by addition of NH_4Cl.

Purification of Palladium [160]. Impure ammonium hexachloropalladate(IV), $(NH_4)_2[PdCl_6]$, can be dissolved directly for purification. This is achieved by the following reaction:

$$3\,(NH_4)_2[PdCl_6] + 20\,NH_3 \rightarrow$$
$$3\,[Pd(NH_3)_4]Cl_2 + 12\,NH_4Cl + N_2$$

Acidification of a solution of $[Pd(NH_3)_4]Cl_2$ precipitates the sparingly soluble, pale yellow *trans*-diamminedichloropalladium(II):

$$[Pd(NH_3)_4]Cl_2 + 2\,HCl \rightarrow [PdCl_2(NH_3)_2] + 2\,NH_4Cl$$

The impure salt should be dissolved quickly at room temperature with agitation to prevent appreciable quantities of the $(NH_4)_2[PtCl_6]$ present as an impurity from dissolving. Undissolved material, mainly hydroxides and $(NH_4)_2[PdCl_6]$, is filtered off. Hydrochloric acid is added to the solution until a pH of 1 is reached, causing $[PdCl_2(NH_3)_2]$ to precipitate as easily filterable crystals. Possible causes of poor purification are too high a concentration of $(NH_4)_2[PdCl_6]$ in the impure material, too prolonged a reaction with ammonia, too high a concentration of ammonia, or problems with filtration of the hydroxidic precipitate before precipitating $[PdCl_2(NH_3)_2]$. High temperature is also detrimental. If the acid concentration is too high, the yield is reduced due to further reaction to form soluble $(NH_4)_2[PdCl_4]$.

Other possible methods for dissolving $(NH_4)_2[PdCl_6]$ are

$$2\,(NH_4)_2[PdCl_6] + N_2H_4 \rightarrow 2\,(NH_4)_2[PdCl_4] + 4\,HCl + N_2$$

or

$$(NH_4)_2[PdCl_6] \rightarrow (NH_4)_2[PdCl_4] + Cl_2$$

Pure $(NH_4)_2[PdCl_6]$ can be obtained in almost quantitative yield by passing chlorine through the solution and adding NH_4Cl.

The purification of $(NH_4)_2[PdCl_6]$ by reprecipitation is useful in the presence of silver and most base metals. In the presence of rhodium, platinum, iridium, and copper, however, better purification is achieved by precipitating $[PdCl_2(NH_3)_2]$, although higher losses of palladium to the mother liquor must be tolerated.

Dichlorodiamminepalladium(II), $[PdCl_2(NH_3)_2]$, can be produced by direct precipitation from an impure solution of palladium resulting from crude separation or inadequate purification. It is dissolved in aqueous ammonia solution to form $[Pd(NH_3)_4]Cl_2$ and reprecipitated by hydrochloric acid.

An ammoniacal stripping solution is used after solvent extraction of palladium by dialkyl sulfide. Palladium can be precipitated from this by acidifying to pH 1, thus producing $[Pd(NH_3)_2Cl_2]$ directly in a highly pure form.

Purification of Iridium. The purification of crude ammonium hexachloroiridate(IV), $(NH_4)_2[IrCl_6]$, is carried out by precipitation of the unwanted elements as sulfides; these consist mainly of other platinum group metals and heavy base metals. This converts hexachloroiridate(IV) to hexachloroiridate(III), which must be reoxidized to precipitate $(NH_4)_2[IrCl_6]$. The sulfides are precipitated from slightly acidic solution by hydrogen sulfide. To ensure optimum precipitation conditions for all elements present, the sulfide precipitation must usually be carried out repeatedly. Iridium can be purified by liquid – liquid extraction [161]. It is extracted relatively easily in the tetravalent state, but not in the trivalent state (cf. Fig. 15). Reduction by Fe^{2+}, As^{3+}, or ascorbic acid at room temperature causes selective reduction of Ir (IV) or Ir (III); Pt (IV) is unaffected.

Purification of Rhodium [162]. Ammonium hexachlororhodate(III), $(NH_4)_3[RhCl_6]$, can be crystallized from water or hydrochloric acid. However, the yields are poor, as is the degree of separation from platinum and iridium. Precipitation of $[Rh(NH_3)_5Cl]Cl_2$ or $(NH_4)_3[Rh(NO_2)_6]$ is more effective, although yields are not good,

and treatment of the mother liquor is problematic.

Good results are obtained when $H_2[PtCl_6]$ and $H_2[IrCl_6]$ are removed by liquid – liquid extraction, and base metal cations by strongly acidic ion-exchange resins. At one time, treatment of crude rhodium with chlorine at 700 °C was an important process. The reaction product consisted mainly of insoluble $RhCl_3$ from which the chlorides of platinum, palladium, and base metals were dissolved by treatment with hydrochloric acid. Satisfactory purification could usually be obtained only by repeating the process.

Crude potassium hexanitritorhodate(III), $K_3[Rh(NO_2)_6]$, can be converted to hexachlororhodate(III) by heating with hydrochloric acid. This can be followed by reprecipitating $K_3[Rh(NO_2)_6]$ or precipitating $[RhCl(NH_3)_5]Cl_2$.

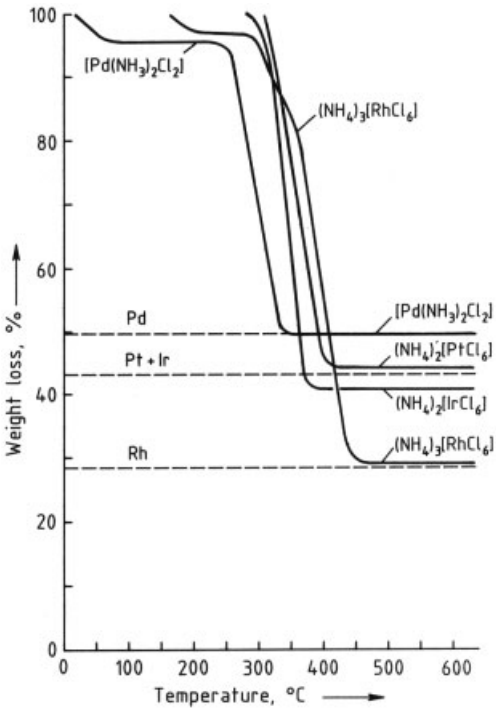

Figure 18. Thermal decomposition of platinum group metal complexes

6.6. Conversion of Salts into Metals [31], [36], [41], [108], [118]

The end products of PGM separation are metal compounds, but all platinum group metals are marketed almost exclusively in the form of metallic sponge or powder.

Conversion to the metals is carried out mainly by thermal decomposition of the ammonium chloro complexes (calcination). These compounds have decomposition temperatures (Fig. 18), of 200 – 500 °C. The size and geometry of the charge affect the reaction course. The low thermal conductivity of the charge means that the reaction proceeds slowly, but smooth production of the desired decomposition products is ensured. Also, the negative heat of reaction of the decomposition tends to slow down the process. In practice, favorable conditions are achieved by external heating at 900 – 1000 °C.

In earlier days, the thermal decomposition of PGM compounds was carried out in gas-heated muffle furnaces that were not very leaktight. The best of these were provided with a collection chamber for NH_4Cl (sublimation temperature 340 °C) and included a waste gas system [43]. Today, tubular fused silica reactors, enclosed in an electrically heated furnace, are used. A stream of inert gas can be passed through at a controlled rate, and absorption equipment is present for NH_4C and HCl. Automated plants are in operation [163], [164].

Thermal treatment of PGM compounds can also be carried out in the presence of hydrogen. For safety reasons, a mixture of hydrogen and nitrogen below the ignition temperature is preferable.

Another possibility to convert PGM compounds into metals is chemical reduction with, e.g., hydrazine in aqueous media. This process must be used if the final products of purification are salts other than the ammonium salts (e.g., $K_2[PtCl_6]$).

Platinum. The compound $(NH_4)_2PtCl_6$ can be converted to platinum sponge by thermal decomposition at 800 °C, according to

$$3\,(NH_4)_2[PtCl_6] \rightarrow 3\,Pt + 2\,NH_4Cl + 16\,HCl + 2\,N_2$$

Excessive carryover of platinum can be avoided by introducing ammonium hexachloroplatinate(IV) into the reaction zone at ca. 300 °C

and heating it rapidly. Waste gas is washed with water to extract ammonium chloride, both for environmental reasons and to prevent the loss of traces of platinum.

If the reaction is carried out in a hydrogen atmosphere, elemental platinum is formed below the sublimation temperature of NH_4Cl (ca. 350 °C). Platinum carryover is then appreciably less, and more ammonium chloride is formed:

$$(NH_4)_2[PtCl_6] + 2\,H_2 \rightarrow Pt + 2\,NH_4Cl + 4\,HCl$$

If NH_4Cl is not sublimed but is removed by dissolving in water, finely divided platinum powder can be produced by this process. Calcination of the hexachloroplatinate(IV) of an alkali metal is not recommended, because the alkali metal chloride produced is difficult to remove by aqueous dissolution from the platinum sponge.

Hexachloroplatinates can usually be reduced in aqueous solution or suspension according to

$$(NH_4)_2[PtCl_6] + N_2H_4 + 6\,NaOH \rightarrow Pt + 6\,NaCl + 2\,NH_3 + N_2 + 6\,H_2O$$

The reaction proceeds best above 80 °C, with excess NaOH maintained at 2 M and the reaction controlled by adjusting the rate of addition of N_2H_4. The reaction must be flushed with nitro-gen, because hydrogen can be formed in a side reaction. The platinum black formed is washed to remove salts, and either dried to produce powder or heated strongly to produce sponge. Overall, wet chemical reduction is more expensive than calcination but should be used if alkali metal hexachlorometallates are present, because their decomposition products cannot be sublimed.

Palladium. The thermal decomposition of dichlorodiaminepalladium(II), $[Pd(NH_3)_2Cl_2]$, begins at ca. 290 °C in an inert atmosphere. Process conditions are the same as for $(NH_4)_2[PtCl_6]$. The reaction is

$$3\,[Pd(NH_3)_2Cl_2] \rightarrow 3\,Pd + 4\,NH_4Cl + 2\,HCl + N_2$$

For the calcination of ammonium hexachloropalladate(IV), which loses chlorine at ca. 280 °C, the temperature should be increased extremely slowly over the range 100 – 300 °C. If calcination is carried out in air, the palladium sponge formed contains large amounts of PdO, which must be reduced by strong heating in a hydrogen atmosphere.

The wet chemical reduction

$$2\,[Pd(NH_3)_2Cl_2] + N_2H_4 \rightarrow 2\,Pd + 4\,NH_4Cl + N_2$$

and the reduction of $(NH_4)_2[PdCl_6]$ proceed quantitatively in strongly ammoniacal solution above 50 °C. Sodium formate in alkaline solution can also be used as reducing agent. A nitrogen atmosphere must be used in both cases to prevent formation of explosive gas mixtures. Chemical reduction often produces pyrophoric palladium black, which must be dried in an inert gas atmosphere.

Iridium. The best method of producing iridium is to reduce ammonium hexachloroiridate(IV), $(NH_4)_2[IrCl_6]$, in a hydrogen atmosphere ca. 800 °C, yielding the metal powder. Calcination in air gives an oxide-containing material. Wet chemical reduction with hydrazine is possible only in strongly alkaline sodium hydroxide solution and does not always go to completion.

Rhodium. Ammonium hexachlororhodate(III) is subjected to the same process of thermal reduction as used for $(NH_4)_2[IrCl_6]$.

Wet chemical reduction with hydrazine according to

$$4\,(NH_4)_3RhCl_6 + 3\,N_2H_4 + 24\,NaOH \rightarrow 4\,Rh + 3\,N_2 + 12\,NH_3 + 24\,NaCl + 24\,H_2O$$

must be carried out with a large excess of hydrazine and NaOH at ca. 100 °C to ensure complete reaction.

Ruthenium. Ammonium hexachlororuthenate(III – IV) is reduced, preferably in a hydrogen atmosphere, at ca. 800 °C. Aqueous solutions of ruthenates(IV) (e.g., produced by adding RuO_4 to an alkali-metal hydroxide solution) are treated with ethanol to precipitate hydrated RuO_2, which is reduced in a hydrogen atmosphere at 800 °C.

Osmium. Osmium tetroxide that has been absorbed in potassium hydroxide solution is precipitated as potassium osmate(IV) by adding ethanol. This is reduced by strong heating in a hydrogen atmosphere or reacted with $NaBH_4$ in aqueous solution [165]. The product must be washed with water in both cases.

6.7. Partial Purification

Platinum – Rhodium Mixtures. For industrial use, platinum and rhodium are often alloyed together, and separating these elements during the recovery process may be unnecessary. The solutions obtained by dissolving metallic platinum – rhodium scrap can be used. The neutral solutions are treated with a sodium-loaded ion exchanger to replace all unwanted cations with sodium. Sometimes, Na^+ is replaced by H^+, giving a solution of $H_2[PtCl_6]$ and $H_3[RhCl_6]$ suitable for further treatment (e.g., reduction by hydrazine to give Pt – Rh powder).

Similar processes can be used for treating some single metal catalysts (e.g., material from the calcination of carbon – palladium).

Surface Cleaning. Some returned metallic materials with only surface contamination require merely surface cleaning, thereby avoiding a wet separation process. They can be treated with acids, such as hydrofluoric acid, or with fused salt (e.g., $NaHSO_4$), although some of the outer layer of platinum dissolves.

6.8. Treatment of Internally Recycled Material

A buildup of base metals occurs in recycled material such as mother liquor from precipitation and crystallization. These base metals originate from the raw materials and from salts formed in the reactions, and also include small quantities of platinum group metals from the separation process. The quantity or recycled noble metals should normally not exceed 1 % of the metal in the starting material. The extent of recycle must be limited, because losses are relatively high when noble metals are recovered from this type of material.

Solutions containing platinum metals with only small amounts of other materials should be concentrated by evaporation in the presence of hydrochloric acid. They can be combined with fresh raw material for separation.

If the solutions contain only ammonium chloride and platinum group metals, they can often be reused directly as a precipitating agent after concentration in acid conditions.

If the solutions contain large amounts of salts of alkali metals, alkaline earths, or other metals along with platinum and palladium residues, it is advantageous to separate the ions $[PtCl_6]^{2-}$ and $[PdCl_4]^{2-}$ from the other materials by using strongly basic ion exchangers. Good separation is achieved if only the two chloro complexes are present. For rhodium and iridium this condition is often not achieved, because of hydrolysis to give hydroxo complexes, so this method cannot be used unless special precautions are taken.

In earlier times, noble metals were recovered from such solutions by cementation with zinc in a hydrochloric acid solution. Aluminum is now the preferred reducing agent because of the high cost of zinc, which must be present in large excess, and the resultant environmental problems. Noble metals can also be recovered with sodium borohydride. Alternatively, they can be reduced by hydrazine in alkaline solution. If no precautions are taken, hydroxides of base metals are also precipitated, but these can easily be redissolved. Copper is also precipitated in this reduction process. After redissolving the metals, the latter can be reprecipitated as copper(II) oxalate or another copper(I) or copper(II) salt that is sparingly soluble in slightly acidic solution.

All materials from which noble metals have been recovered must be monitored carefully for possible noble-metal content before disposal, so that no irrevocable losses occur.

6.9. Construction Materials

Many of the operations described above pose special problems in the choice of materials for equipment and vessels.

Plastics resistant to mineral acids can often be used. Vessels and reactors lined with hard rubber are resistant to acids and saturated chlorine solutions at 90 °C. Unsuitable metals for commonly occurring solutions of platinum and other platinum group metals include stainless steel, titanium, zirconium, the Hastelloys, and silver, because these materials cause cementation of platinum group metals. Tantalum is an exception. It is resistant to dissolved halogens and does not cement platinum group metals from solution. It is a suitable material for steam-heated heat exchangers. Equipment constructed of borosilicate glass and glass-lined steel reactors are chemi-

cally resistant to all of these materials. For solutions containing large amounts of ruthenium or osmium, other construction materials are recommended, including graphite for some items.

For treating raw materials with acidic molten salts or salts that have a chlorinating action, and for calcination, the use of fused silica equipment is best. Treatment with molten oxidizing alkaline salts is carried out in silver or nickel crucibles.

7. Platinum Group Metal Compounds [5], [6], [30], [32], [36], [41], [88], [105–107], [166], [263–270]

Because of the many possible oxidation states of platinum group metals and their tendency to form complexes, the range of compounds is extremely wide.

7.1. Inorganic Compounds

7.1.1. Platinum Compounds

Hexachloroplatinic(IV) acid, $H_2[PtCl_6] \cdot 6\,H_2O$, M_r 517.92, theoretical platinum content 37.68 %, platinum content of commercial product ca. 40 % [commercial names: chloroplatinic acid, hydrogen hexachloroplatinate(IV)], forms reddish brown crystals (*mp* ca. 150 °C), which deliquesce in moist air and are readily soluble in water or alcohol. The usual commercial product is an aqueous solution with a platinum content of 25 %.

Production (see Chap. 5). Platinum sponge is treated with moderately concentrated hydrochloric acid saturated with chlorine in a stirred vessel at ca. 80 °C. The solution obtained is evaporated until it reaches 150 °C. When cooled, it changes into solid $H_2PtCl_6 \cdot 4.5\,H_2O$, which contains ca. 40 % platinum, no mother liquor is produced. If the platinum is dissolved in aqua regia instead of hydrochloric acid – chlorine, the nitric acid and NOCl formed must be removed completely, preferably by repeated evaporation to a syrupy consistency and redissolution in hydrochloric acid.

Uses. Hexachloroplatinic(IV) acid is the most industrially important platinum compound. It is used in the production of most other platinum compounds and preparations. It is used primarily to make catalysts by impregnating support materials and as a precursor hydrosilylation catalysts.

Hexachloroplatinates(IV). Ammonium hexachloroplatinate(IV), $(NH_4)_2[PtCl_6]$, and, to a limited extent, the orange-red salt sodium hexachloroplatinate(IV), $Na_2[PtCl_6]$, are important in platinum separation processes (see Chap. 6).

Platinum Chlorides. Platinum(IV) chloride, $PtCl_4$, is produced by careful dehydration of $H_2PtCl_6 \cdot 6\,H_2O$ at ca. 300 °C in a stream of chlorine. It is a red-brown, crystalline, hygroscopic powder with a relatively narrow range of thermal stability.

Above 380 °C, $PtCl_4$ liberates chlorine and forms platinum(II) chloride, $PtCl_2$, which is stable between 435 and 580 °C. Platinum(III) chloride, $PtCl_3$, is probably formed as an intermediate in this reaction. Above 580 °C, further decomposition occurs to yield metallic platinum. Platinum(IV) chloride is the only chloride of platinum that is soluble in water.

Platinum(II) chloride can also be obtained from a solution of $H_2[PtCl_4]$ by careful evaporation under vacuum. Chlorides of platinum can be produced by reacting chlorine with finely divided platinum. The chloride obtained will be the one that is stable at the reaction temperature.

Oxides of Platinum. The most industrially important oxide of platinum is platinum(IV) oxide hydrate, $PtO_2 \cdot H_2O$, a hydrogenation catalyst. To prepare this substance, a solution of 9 parts sodium nitrate and 1 part platinum as $H_2[PtCl_6]$ is evaporated to dryness. The product mixture is finely powdered and added to molten $NaNO_3$ (10 parts) at 520 °C with agitation. The product is dissolved in water, washed, and carefully dried to give a brown powder, insoluble in aqua regia, with poorly defined stoichiometry.

Platinum(II) oxide, PtO, and platinum(II) oxide hydrate, $PtO \cdot H_2O$, are also known, but have little or no industrial importance.

Other Platinum Compounds. The following compounds are used in aqueous electrochemistry: potassium tetranitroplatinate(II), $K_2[Pt(NO_2)_4]$; dinitrodiammineplatinum(II),

$[Pt(NH_3)_2(NO_2)_2]$; and sodium hexahydroxoplatinate(IV), $Na_2[Pt(OH)_6]$.

Potassium tetracyanoplatinate(II), $K_2[Pt(CN)_4]$, can be used in electroplating by molten salt electrolysis. Barium tetracyanoplatinate(II), $Ba[Pt(CN)_4]$, is used in the manufacture of fluorescent screens. Many other cyano complexes are known. Simple cyanides such as $Pt(CN)_2$ and $Pd(CN)_2$ also exist.

Potassium tetrachloroplatinate(II), $K_2[PtCl_4]$, is a starting material for the synthesis of most Pt (II) compounds. Platinum(II) acetylacetonate is used in the pyrolytic production of platinum surface coatings (see Section 10.10.2) and is superior to other platinum compounds in this application.

Dichlorodiammineplatinum(II), $[PtCl_2(NH_3)_2]$, is of historical interest and was known in two forms: "Peyrone's chloride" and "Reiset's second chloride." In 1893, WERNER recognized that these were *cis* and *trans* isomers. *cis*-Diammine dichloroplatinum, also known as cisplatin or cis-platinum, is used in cancer therapy. Carboplatin (*cis*-diammine-1,1-cyclobutanedicarboxylatoplatinum), a second-generation cancer drug shows fewer side effects than cis-Platinum. Drugs for oral application like JM 216 are still in clinical trials [271].

7.1.2. Palladium Compounds [168]

Tetrachloropalladic(II) acid, $H_2[PdCl_4]$, is stable only in solution. Commercial solutions in hydrochloric acid contain 20 % palladium and are dark brown.

Production. The method of producing $H_2[PdCl_4]$ solution is similar to that for $H_2[PtCl_4]$ solution. The metal is dissolved in $HCl-Cl_2$ or $HCl-HNO_3$, the rate of dissolution being higher than with platinum (see Fig. 6). If dissolution occurs below ca. 50 °C, hexachloropalladic(IV) acid is formed first.

Uses. The solution of tetrachloropalladic(II) acid is the most industrially important palladium preparation. It is the starting material for almost all other palladium compounds, particularly catalysts.

Palladium(II) chloride, $PdCl_2$, M_r 177.31, theoretical palladium content 60.0 %, palladium content of commercial product 59.9 %, is a brown to brownish violet powder, insoluble in water, but readily soluble in hydrochloric acid and solutions of alkali-metal chlorides. It sublimes at 590 °C. Decomposition begins at 600 °C and is complete at 740 °C.

Production. The best method of production of $PdCl_2$ is careful evaporation of a solution of $H_2[PdCl_4]$ in hydrochloric acid, preferably in a rotary evaporator.

Uses. Anhydrous $PdCl_2$ is the starting material for a number of palladium compounds.

Palladium Oxides. Poorly defined oxide hydrates are obtained by adding alkali to aqueous solutions of Pd (II) compounds. The Pd (IV) oxide hydrates obtained from Pd (IV) solutions release oxygen. A catalytically active palladium preparation analogous to $PtO_2 \cdot x\,H_2O$ can be obtained by evaporating a solution of $H_2[PdCl_4]$ and $NaNO_3$, and fusing the product.

Stoichiometric palladium(II) oxide, PdO, which is crystallographically well defined, is obtained by reaction of palladium black with oxygen or air at 750 °C. Decomposition occurs at 850 °C.

Ammine Complexes of Palladium. The addition of ammonia to solutions of palladium chloride first causes the formation of a pink precipitate of the binuclear complex $[Pd(NH_3)_4PdCl_4]$, known as Vauquelin's salt, which is converted to soluble tetramminepalladium(II) chloride, $[Pd(NH_3)_4]Cl_2$, by further addition of NH_3. Acidification of this solution yields the sparingly soluble light-yellow *trans*-diamminedichloropalladium(II), $[PdCl_2(NH_3)_2]$. These compounds and reactions are important in the industrial separation of palladium (see Chap. 6) and in electroplating processes.

Other Palladium Compounds. Ammonium hexachloropalladate(IV), $(NH_4)_2[PdCl_6]$, is important in separation technology. It is an oxidation product of tetrachloropalladate(II) (see Chap. 6). Palladium(II) sulfate, palladium(II) nitrate, and palladium(II) acetate are of some importance in preparative chemistry. The ready formation of palladium hydride is exploited industrially for example for the purification of hydrogen.

7.1.3. Rhodium Compounds [169]

Hexachlororhodic(III) Acid, H_3RhCl_6. Wet separation processes and molten salt dissolution usually produce $(NH_4)_3[RhCl_6]$ or $Na_3[RhCl_6] \cdot 12\,H_2O$ in crystalline form. Both compounds form deep-red crystals, which are very soluble in water. Free hexachlororhodic(III) acid, $H_3[RhCl_6]$, can be obtained by cation exchange, by oxidative decomposition of the cation in the case of the NH_4^+ salt (see Section 6.2) or by dissolution of finely divided rhodium powder in hydrochlorine acid and chlorine gas. The composition of the chloro complexes of rhodium depends strongly on the concentration of hydrochloric acid, the temperature, the duration of reaction, and the previous treatment of the solution. By starting with a strongly acidic medium, the following transformations are possible:

$$[RhCl_6]^{3-} \rightleftharpoons [RhCl_5(H_2O)]^{2-} \rightleftharpoons [RhCl_4(H_2O)_2]^{-} \rightleftharpoons [RhCl_3(H_2O)_3] \rightleftharpoons [RhCl_2(H_2O)_4]^{+} \rightleftharpoons [RhCl(H_2O)_5]^{2+} \rightleftharpoons [Rh(H_2O)_6]^{3+}$$

The color varies from raspberry red to brown.

Rhodium Chlorides. Anhydrous, brownish red rhodium(III) chloride, $RhCl_3$, which is insoluble in water and mineral acids, is obtained by heating rhodium powder to 700 °C in a chlorine atmosphere.

So-called soluble rhodium chloride, with the approximate composition $RhCl_3 \cdot 2.5\,H_2O$, is obtained by evaporating a solution of $H_3[RhCl_6]$ in hydrochloric acid, preferably in a rotary evaporator. This compound is the most important commercial rhodium product, and is used as the starting material for other rhodium compounds and catalysts.

Rhodium Sulfate. Rhodium(III) sulfate, $Rh_2(SO_4)_3$, can crystallize with varying amounts of water; it can be obtained by dissolving $Rh(OH)_3 \cdot x\,H_2O$ in sulfuric acid or by the action of hot sulfuric acid on freshly precipitated rhodium black. It is used in rhodium plating baths.

Chlorotris(triphenylphosphine)rhodium(I), $RhCl[P(C_6H_5)_3]_3$, is formed by the reaction of triphenylphosphine with soluble rhodium(III) chloride in ethanol solution under reflux. It separates as deep-red crystals.

This complex, and rhodium compounds with similar structures, are important homogeneous catalysts (see Section 8.1).

7.1.4. Iridium Compounds

Hexachloroiridic(IV) acid, $H_2[IrCl_6]$, is produced in aqueous solution by the methods used to obtain $H_3[RhCl_6]$ (i.e., from the ammonium or sodium salts). It is by far the most important iridium compound. The oxidation state is relatively easily changed (in either direction), the very intense dark brown $[Ir^{IV}Cl_6]^{2-}$ changing to the lighter colored $[Ir^{III}Cl_6]^{3-}$ or vice versa. The compounds are used as precursors in the preparation of coated anodes for the production of chlorine and sodium hydoxide. The ammonium salts of both anions are important in separation technology (see Chap. 6).

Iridium(III) Chloride, $IrCl_3$. A product stoichiometrically deficient in chlorine is obtained by heating iridium powder in a stream of chlorine at 650 °C. The chlorination reaction is assisted by the presence of carbon monoxide.

7.1.5. Ruthenium Compounds

Ruthenium(VIII) oxide, RuO_4 is formed by the oxidation of aqueous solutions of ruthenates. It has a melting point of 25.5 °C and is therefore a liquid at room temperature. It is both toxic and explosive (see Section 10.1).

Soluble Ruthenium(IV) Chloride. So-called soluble ruthenium chloride consists mainly of $Ru(OH)Cl_3$ with some $RuCl_3 \cdot H_2O$, and is prepared by reacting RuO_4 with hydrochloric acid and evaporating the solution. The compound $(NH_4)_2[RuCl_6]$ can be precipitated by adding NH_4Cl. Solutions of ruthenium chloride in hydrochloric acid are used in the production of coated anodes.

Potassium ruthenate(VI) , $K_2[RuO_4]$, is formed by fusing ruthenium with $KOH - KNO_3$ or by dissolving RuO_4 in potassium hydroxide solution.

7.1.6. Osmium Compounds

Osmium(VIII) oxide, OsO_4, is the most important osmium compound. It is formed even at room temperature by the oxidation of osmium powder by air. It is prepared by strongly heating osmium in a stream of oxygen at 500 – 800 °C. The melting point is 40.6 °C, and the boiling point 131.2 °C. Sublimation takes place even at room temperature. OsO_4 used as an oxidation catalyst or as an aqueous solution to stain organic tissue for microscopic analysis.

Potassium osmate(VI), $K_2[OsO_2(OH)_4]$, is formed by the reaction of OsO_4 with potassium hydroxide solution and alcohol.

7.2. Organic Compounds [170], [171]

The PGMs form numerous compounds with direct metal-carbon links both as carbonyls and as organometallic compounds. Many of these are of industrial importance in homogeneous catalysis (see Section 10.5). Complex organic compounds of ruthenium and osmium can facilitate electron-transfer mechanisms in molecular structures and also serve as sensitizer materials for photoinduced hydrogen production (see Section 10.9).

Osmium bipyridyl redox polymers combined with a redox enzyme can be used as miniature in vivo sensors for implantation into diabetes patients for continuous monitoring of their blood glucose levels (wired enzyme biosensor, [271]). The divalent and trivalent complexed osmium ions can exchange electrons at high rates. The osmium polymer connects electrically the redox centres of the enzymes by introducing electron relays close to the enzyme active site. A sol – gel based amperometric biosensor incorporates an osmium redox polymer as mediator for detection of lactate [272].

Ruthenium phthalocyanine complex compounds have proven effective as drugs for photodynamic cancer therapy. It is administered intraperitoneally, followed by laser irradiation of the tumor tissue. The compound acts in the presence of light as sensitizer to induce the photochemical destruction of the unwanted tissue [273].

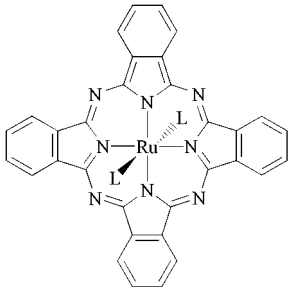

Ruthenium complexes with polyaminocarboxylate ligands act as effective nitric oxide scavengers. The binding property for nitric oxide enables macrophage cells to kill tumor cells and may also form the base for drugs against other nitric oxide influenced diseases including hypertension, epilepsy, diabetes, arthritis, and septic shock [274].

Giant, dendritic organic coordination complexes of **ruthenium and osmium** (dendrimers, [276–279]) with up to 22 ruthenium atoms have been synthesized, which may be used for the production of thin films.

Ruthenium forms catalytic active complex compounds with C_{60} molecules [280–282].

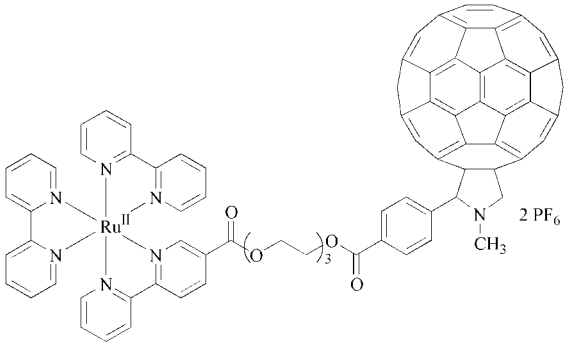

8. Alloys [28], [30–32], [88–91], [166], [167], [172–184]

8.1. Alloy Systems

All the platinum group metals are infinitely miscible with one another in the liquid state. When

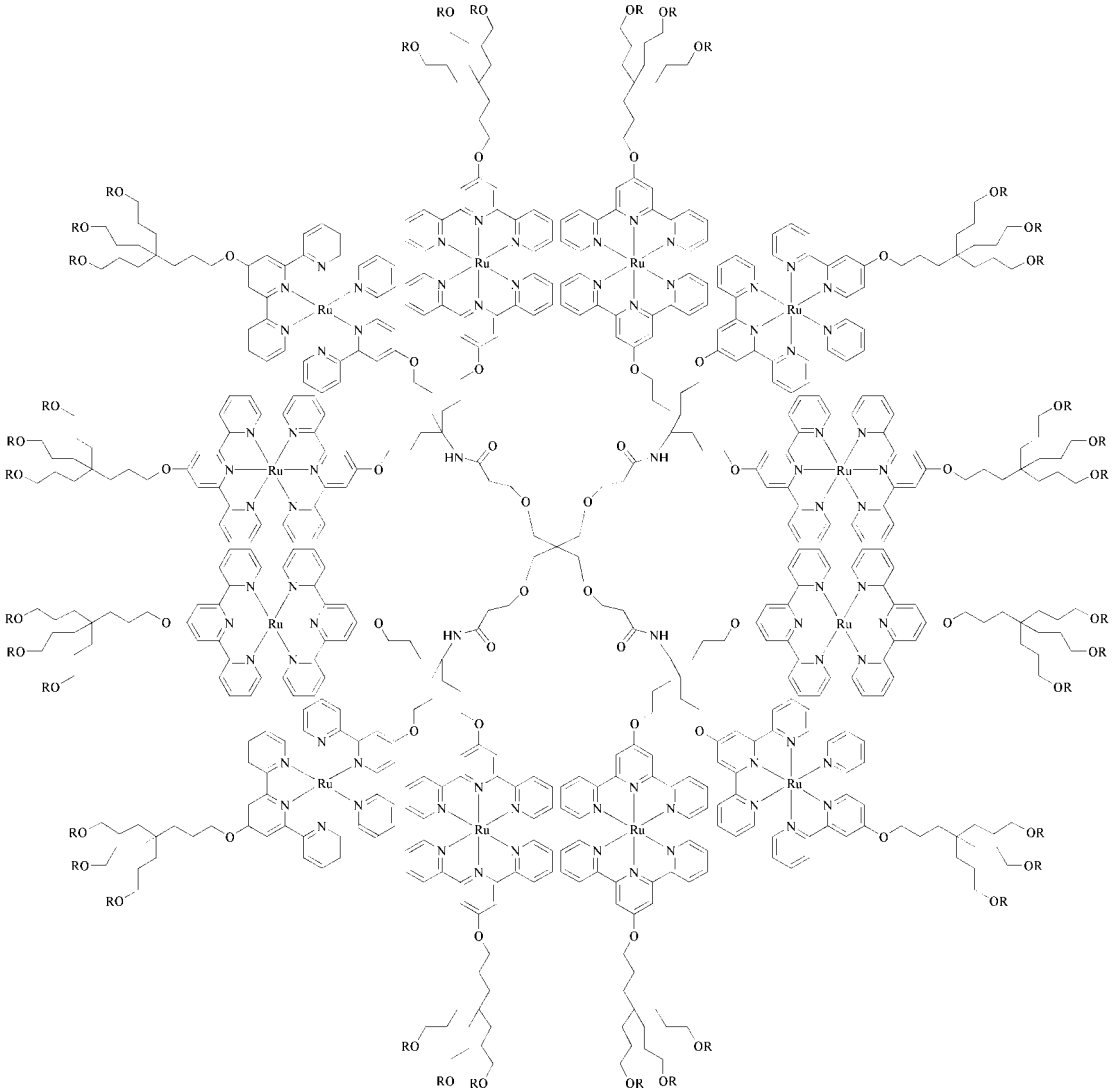

the molten alloys solidify, or when a mixture of the powdered metals is sintered, platinum, iridium, palladium, and rhodium, which crystallize in the face-centered cubic system, form mixed crystals with each other. At lower temperature, miscibility gaps are observed in many cases.

Some phase diagrams of the binary alloys of platinum according to Massalski [90] are shown in Figures 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33; of industrial importance are mainly the binary alloys Pt – Rh, Pt – Ir, Pt – Pd, Pt – Au, but Pt alloyed with base metals (Ni, Co, W) and a few ternary or higher alloys are also used.

8.2. Special Alloys

Platinum and palladium are the most important base components of industrially utilized PGM alloys. Special properties of interest concern corrosion resistance, mechanical strength, high temperature resistance, electrical conductivity, thermoelectric power, and magnetic behavior.

Platinum – palladium alloys are used in jewelry, for electrical contacts, and for equip-

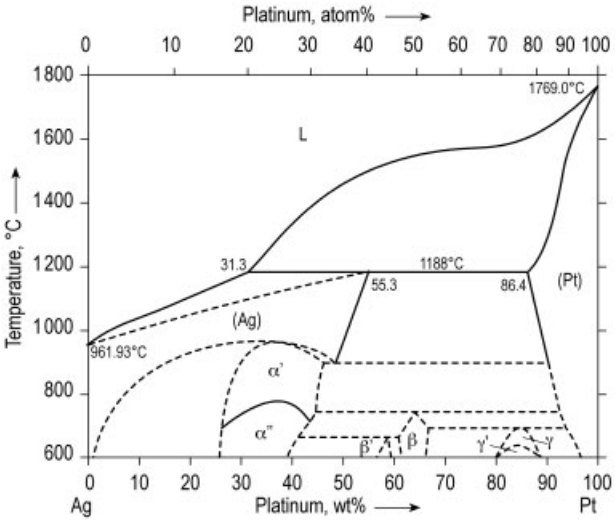

Figure 19. Phase diagram of binary platinum group metal alloys: Ag – Pt

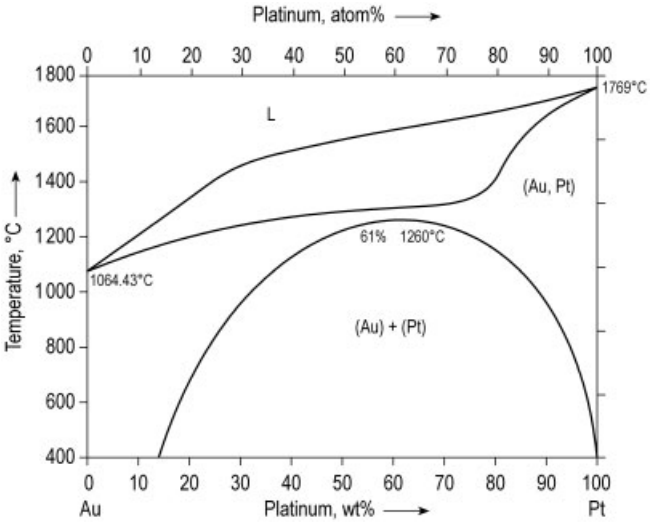

Figure 20. Phase diagram of binary platinum group metal alloys: Au – Pt

ment parts for glass production and construction elements in highly corrosive environments.

Platinum and palladium alloys with W, Re, Ni and other metals of Groups 6 and 8 – 10 are suited for use in high-temperature strain gauges [283]. Pt alloys are used in spark plugs for automotive engines.

Dispersion-strengthened platinum and platinum alloys are produced by coprecipitation of the metals and refractory oxides, for instance, with ZrO_2 (zirconium oxide grain-stabilized platinum, ZGS, Johnson Matthey). Dispersion-strengthened **platinum** alloys with Zr and additional Y are superior in temperature stability and have higher recrystallization temperatures [285].

Platinum can also be effectively dispersion hardened by incorporating very small amounts (0.04 – 0.08 wt %) of refractory carbides (TiC, ZrC). The strengthening is due to a partial decomposition and reaction of the carbide with the matrix. The low carbide concentration offers the advantage that ductility, electrical conduc-

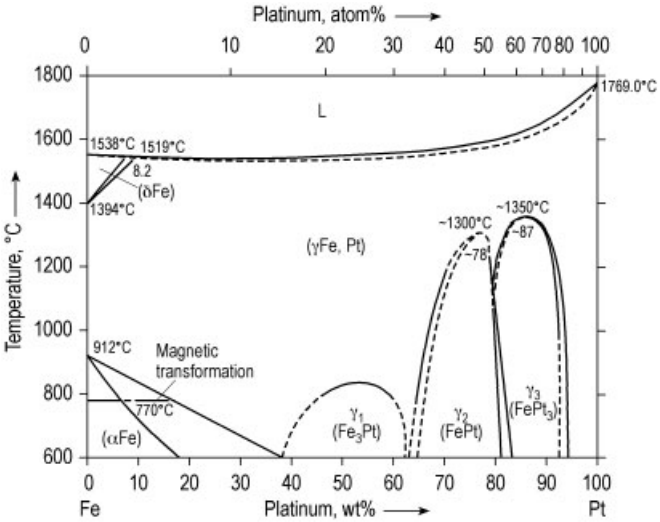

Figure 21. Phase diagram of binary platinum group metal alloys: Fe – Pt

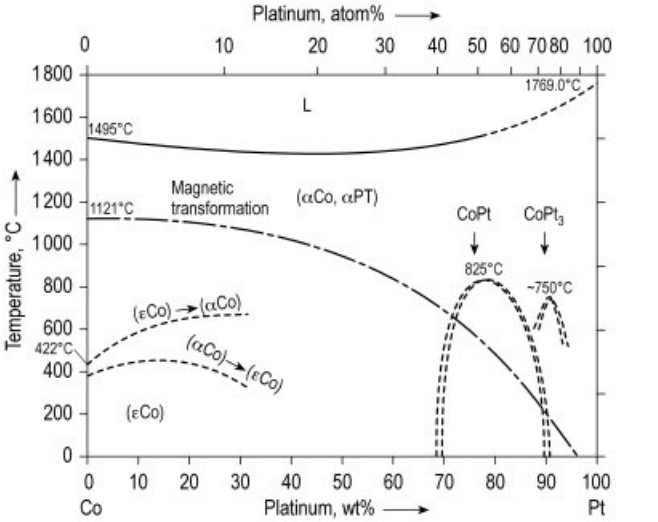

Figure 22. Phase diagram of binary platinum group metal alloys: Co – Pt

tivity, and the recrystallization behavior are not seriously impaired.

Platinum – gold and **platinum – gold – rhodium** alloys have higher strength than pure platinum at 1000 °C, finer grain structure, and better resistance to wetting by molten glass.

Platinum – iridium alloys find different applications in electrochemistry, chemical technology, medicine, and jewelry. Iridium contents (in wt %) are for laboratory ware 0.4 – 0.6 %, jewelry 5 – 15 %, medical applications 10 %, electrical contacts 10 – 25 %, electrodes 10 %, pens, needles, springs 25 – 30 %. Maximum hardness (ca. 360 HB) is reached with 30 % iridium.

W – Th – Al-doped **platinum** – 26 wt % **iridium** and platinum – 30 wt % **rhodium** are used to encapsulate plutonia fuel pellets for spacecraft heat sources [290].

Platinum – rhodium alloys with 10 % rhodium are used as catalyst for the oxidation of NH_3, as the standard element in thermocouples (Pt – Rh of different compositions or Pt), and as heater elements in electrically heated high-tem-

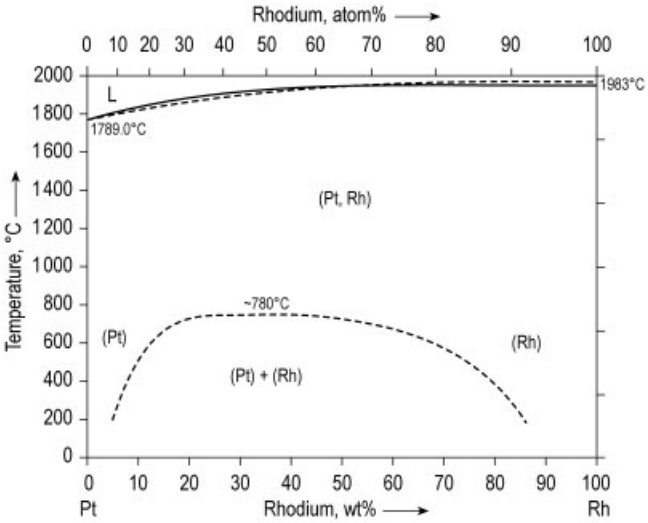

Figure 23. Phase diagram of binary platinum group metal alloys: Pt – Rh

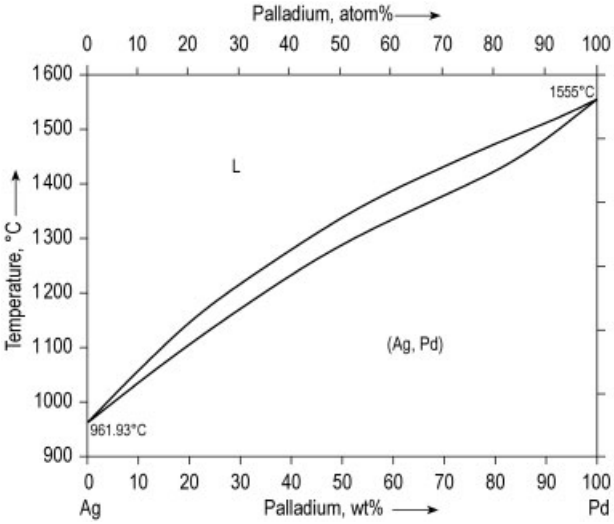

Figure 24. Phase diagram of binary platinum group metal alloys: Ag – Pd

perature furnace. Pt – Rh alloys with additions of Au are especially resistant to wetting by molten glass. They are used in machine parts for the production of high-tech glasses for computer monitors, cathode ray tubes, television screens, liquid crystal displays, and plasma panels; glass fibers for telecommunication; and for use in glass-reinforced construction materials.

Platinum – ruthenium alloys. Ruthenium strengthens platinum very effectively; 10 % ruthenium increases the hardness to 168 HV. The alloy PtRu5 has the same strength, hardness, and corrosion resistance as the alloy PtIr10 ("hard platinum"). Applications are jewelry items, electrodes for electrochemistry, spark plug electrodes, electrical contacts, and wear-resistant wires in potentiometers. Higher alloy compositions up to 14 % ruthenium and ternary types with 15 % rhodium and 6 % ruthenium are used for electrical contacts and potentiometer wires for heavy-duty applications.

Platinum – cobalt multilayers on various substrates are used for ultrahigh-capacity hard disks, for magneto-optical storage, and perpen-

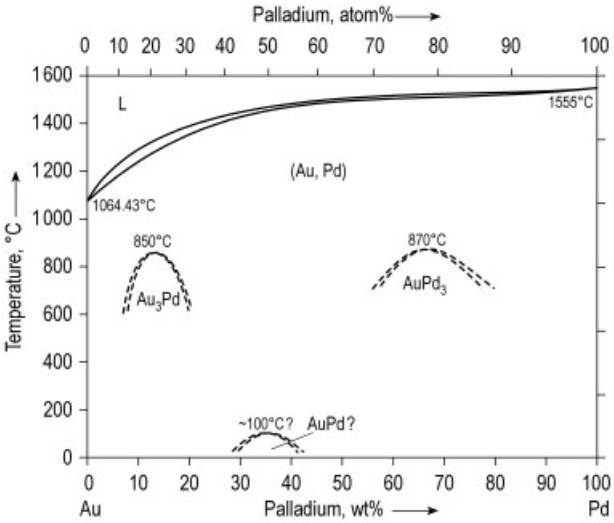

Figure 25. Phase diagram of binary platinum group metal alloys: Au – Pd

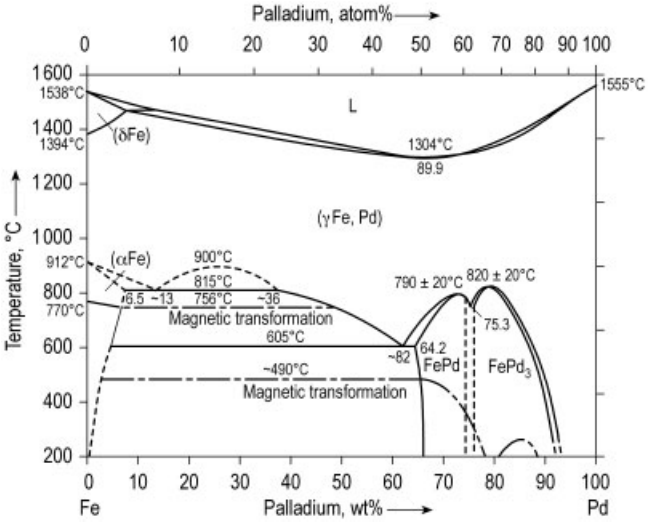

Figure 26. Phase diagram of binary platinum group metal alloys: Fe – Pd

dicular recording applications. Special processing enhances domain wall pinning effects and leads to higher coercivity [286]. Pt – Fe alloys have maximum coercivity at 39.5 wt % platinum [287].

A new class of magnetic behavior, random quantum spin chain paramagnetism, is exhibited in one-dimensional compound $Sr_3CuPt_{1-x}Ir_xO_6$ prepared as solid solution of the antiferromagnetic Sr_3CuPtO_6 and the ferromagnetic Sr_3CuIrO_6. The platinum : iridium ratio determines the magnetic behavior of the solid solution [288]. Alloys of the type $Cr_{1-x}Ir_xO_2$ ($0 < x < 0.3$) are ferromagnetic materials with high coercivity [289].

Platinum – nickel alloys have been proven in electrical switches and as high-strength and highly temperature resistant construction materials in the chemical industry. Compositions range from trace amounts up to 20 % nickel with maximum tensile strength of about 1725 MPa after 90 % cold reduction.

Platinum – tungsten alloys, commonly used in compositions with 5 – 8 wt % Co, have excel-

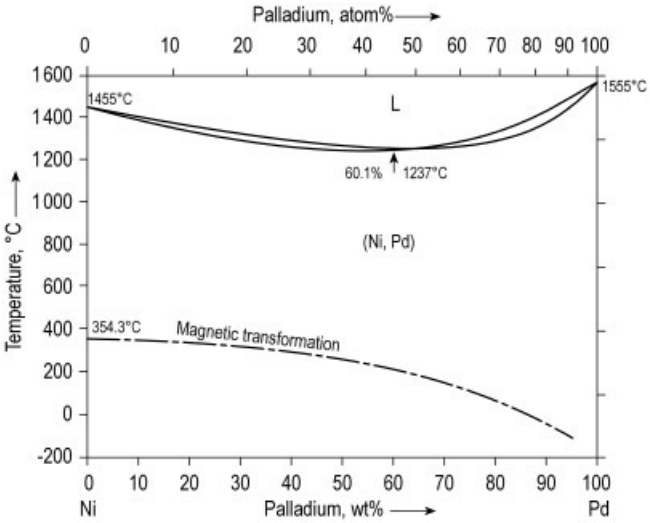

Figure 27. Phase diagram of binary platinum group metal alloys: Ni – Pd

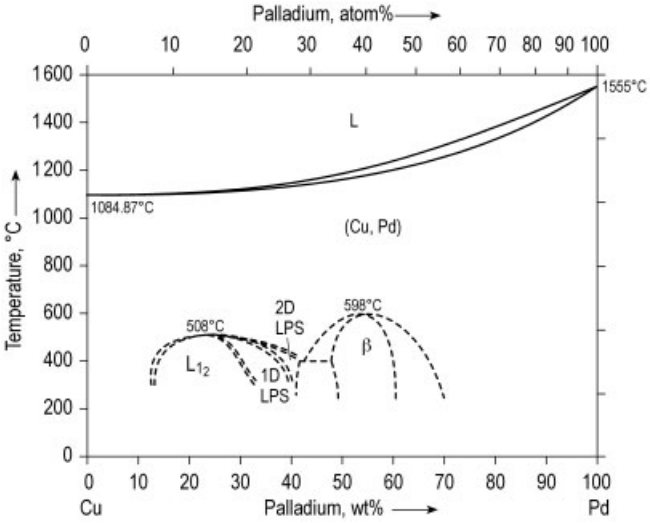

Figure 28. Phase diagram of binary platinum group metal alloys: Cu – Pd

lent wear resistance. They are used for potentiometers, glow wires in copiers, diode switches, heater elements and cladding on rods for fuel elements in nuclear reactors.

Platinum electrodes are hardened with **iridium** for use in spark plugs and as corrosion resistant electrodes.

In equiatomic Ti – Ni alloys, (known for the shape-memory effect), substitution of nickel by 2 atom % **ruthenium** stabilizes the high temperature B2-type phase at room temperature [292]. Partial replacement of Ni by **palladium** in the equiatomic TiNi alloy improves the shape memory properties [293].

Platinum alloys are components in catheters for pacemakers, in dental gold alloys, in marker bands, in guidewires and in temporary and permanent implants.

Palladium – gold alloys are the basis of precision resistance alloys. Addition of iron accentuates the formation of long range ordering. The alloy **Au50Pd45Mo5** has a resistivity of 100 μΩcm and a temperature coefficient of re-

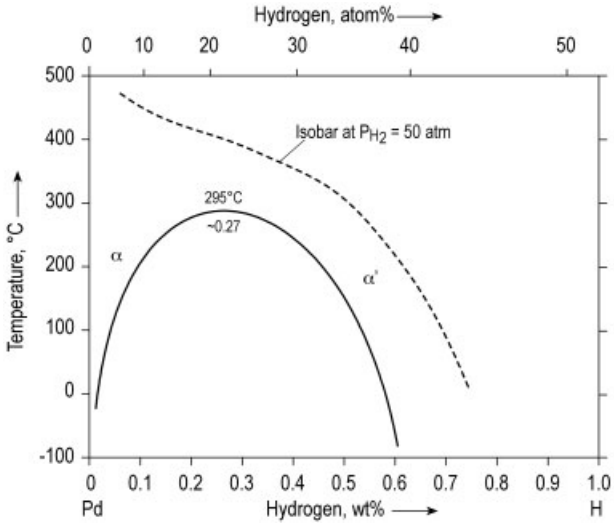

Figure 29. Phase diagram of binary platinum group metal alloys: Pd – H

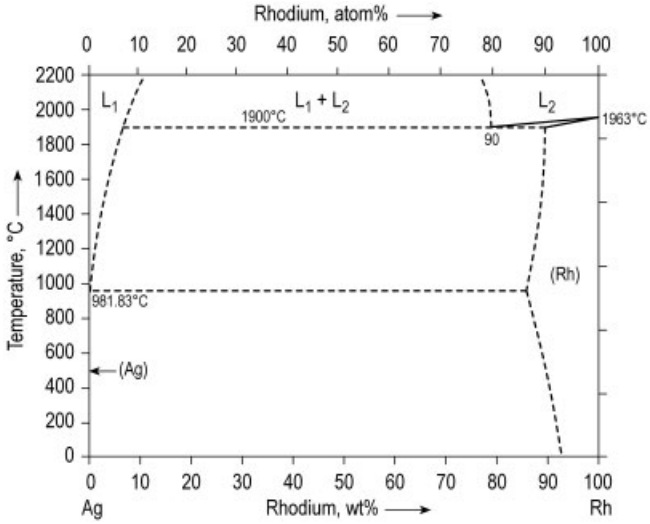

Figure 30. Phase diagram of binary platinum group metal alloys: Ag – Rh

sistance (TCR) of 0.00012 K^{-1} between 0 and 100 °C.

The tensile strength reaches values of about 1060 N/mm^2.

Palladium – silver alloys are used as capacitor electrode layers, as silver conductor paths, and as resistor elements in thick-film circuits. The palladium – silver alloy with 40 atom % Ag has the highest electrical resistivity at room temperature. It finds use for precision resistors in electrical technology. The resistivity between 20 and 100 °C is nearly constant and amounts to 42.5 $\mu\Omega \cdot$ cm; between 20 °C and 800 °C the resistivity changes from 42.5 to 44 $\mu\Omega \cdot$ cm.

Ternary alloys containing silver and copper with up to 10 to 25 % **palladium** are hard solders for high-strength.

Palladium – ruthenium with 4.5 wt % ruthenium is a standard jewelry alloy in the United States. Alloys with ruthenium contents up to 12 % serve as material for electrical contacts and as catalysts for the reduction of nitric oxide.

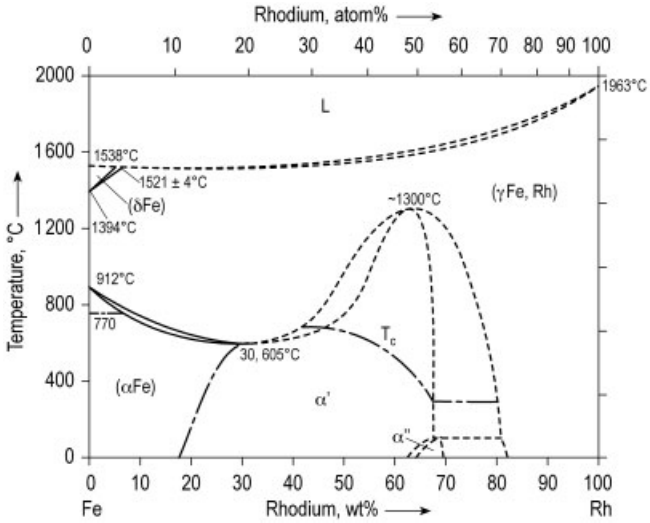

Figure 31. Phase diagram of binary platinum group metal alloys: Fe – Rh

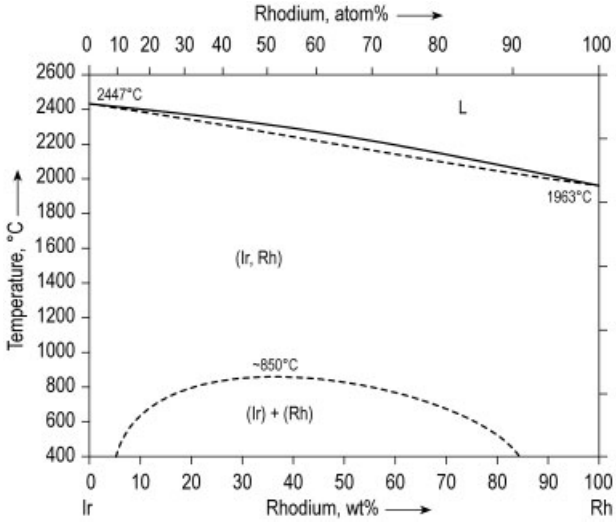

Figure 32. Phase diagram of binary platinum group metal alloys: Ir – Rh

Palladium alloys with 60 – 80 wt % Pd for use in dentistry contain copper and gallium as hardening elements and tin to lower the melting temperature. Best corrosion resistance is attained with a single-phase structure in the composition range of Cu 4 – 6 wt %, Ga 5 – 7 wt %, and Sn 6 wt %. Addition of 1 wt % palladium to silver – tin – copper amalgam alloys improves heterogeneous phase formation and the thermal stability [294].

Palladium alloys with iron and with cobalt show ferromagnetism.

By alloying palladium with 0.1 wt % cobalt, the palladium atoms surrounding a cobalt atom, are strongly magnetized to magnetic moments in the order of 10 μB, which is about six times larger than that of a single cobalt atom.

Palladium and certain palladium alloys are unique in their capability to dissolve large amounts of hydrogen up to about 2800 times their own volume [295–297]. Hydrogen enters palladium interstitially. The hydrogen atoms are located in the octahedral holes of the fcc lattice. Above 295 °C, palladium and hydrogen

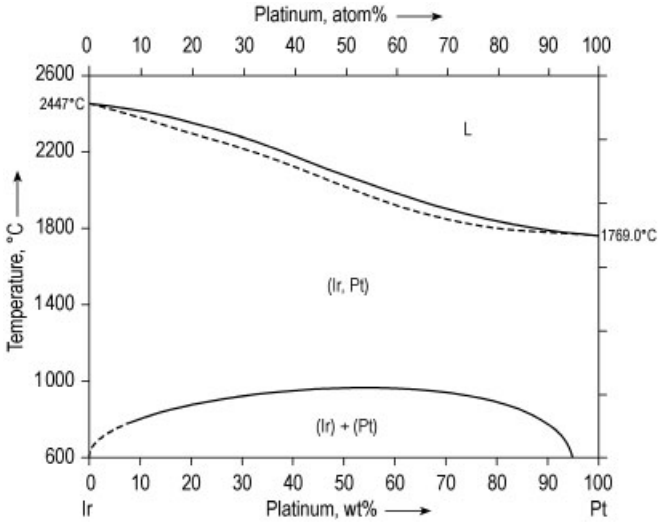

Figure 33. Phase diagram of binary platinum group metal alloys: Ir – Pt

form a continuous serious of solid solutions with fcc structure. Below that temperature, the phase splits into a fcc palladium-rich phase and an fcc hydrogen-rich phase, and the result is a miscibility gap which broadens with decreasing temperature. The occupation of interstitial sites by hydrogen atoms causes expansion of the palladium lattice. The lattice parameters increase with increasing hydrogen content from 3.891 Å for pure palladium up to 4.06 Å at 75 atom % hydrogen [298]. Thermal cycling of Pd – H alloys in the duplex phase region causes high mechanical stresses due to changes of the lattice dimensions with changing amounts of dissolved hydrogen. Pure palladium, when repeatedly heated and cooled in hydrogen-containing atmosphere, becomes brittle and cracks.

Palladium – silver alloys with 20 – 25 wt % silver dissolve larger amounts of hydrogen than pure palladium and are resistant to thermal cycling [299]. **Pd – Ag23** exhibits extremely high hydrogen permeability and is used for semipermeable membranes for the extraction and purification of hydrogen up to 99.99999 % purity. Palladium – silver permeation membranes are provided in the form of bundled tubes with wall thickness of ca. 75 μm, sealed at one end. Hydrogen is pressed from outside through the tubes, dissociating and diffusing through the alloy to the inner side of the tubes, leaving most impurities outside the membrane [300], [301].

The solubilities of the hydrogen isotopes in palladium and palladium – silver alloys differ to an appreciable degree. Diffusion membranes of palladium or **palladium – silver** alloys can therefore be used for separation and concentration processes of different hydrogen isotopes [302]. The equilibrium hydrogen/deuterium separation factors vary from ca. 2.4 at 25 °C to 3.7 at − 80 °C [303]. For the hydrogen/tritium ratio, a value of 2.8 was measured at 25 °C [304].

The ternary alloy **Pd – Ti – Al** is ferromagnetic, in contrast to the antiferromagnetic alloy Pd – Ti [25].

Pt and Pt alloys are electrocatalysts in low-temperature fuel cells. The *phosphoric acid fuel cell* (PAFC) contains a **Pt – Co – Cr alloy** on the cathode side and straight platinum on carbon as the anode. The higher activity of alloys compared to pure platinum on the cathode side is due to the fact that the non-noble metal components are leached from the alloy crystallite surfaces, leaving a highly reticulated and active surface [306], [307]. The catalytic activity is effected by the preparation method and the crystal structure of the three-component alloy [308]. The *proton exchange membrane fuel cell* (PEMFC) contains a **platinum – ruthenium**-loaded graphite anode and pure-platinum-loaded graphite cathode [309]. The ruthenium alloy protects the catalyst against poisoning by the carbon monox-

ide which may be present in the hydrogen supply. Platinum – ruthenium alloys are also used as the anode catalyst in the *direct methanol fuel cell* (DMFC [310]). They prevent the formation of linearly bonded carbon monoxide (Pt–C≡O) during the electrosorption of methanol on the platinum.

Iridium alloys are used in microelectrodes for testing water quality by square-wave anodic stripping voltammetry [311].

Alloying of titanium and high-strength titanium alloys (e.g., Ti – 6Al – 4V) with 0.1 % **ruthenium** effects improvements in the corrosion resistance against hydrochloric acid solutions and can supersede corresponding Ti-Pd alloys. Selective dissolution generates on the surface cathodic sites of low hydrogen overvoltage. This accelerates hydrogen ion reduction and thus causes a substantial shift in the corrosion potential towards higher nobity, where the protective surface oxide film is stable. Applications are apparatus parts for processing corrosive chemicals [291].

Platinum metals based intermetallics in alloyed or composite microstructures show interesting properties in high temperature strength and oxidation resistance as needed for materials in high temperature service. Prototypes are Ir – IrN-composites with refractory metals. Pt – Al – Ru-alloys, ferritic alloys with Ru, and eutectic Ru – RuAl and Ir – IrAl alloys [312].

8.3. Methods of Treatment [88], [107], [166], [167]

The starting material for the manufacture of PGM alloys from platinum and palladium is generally metal sponge. The remaining elements – rhodium, iridium, ruthenium, and osmium – are supplied as powders. Platinum and palladium are also available in ingot form.

Owing to their high melting points, the platinum group metals are almost invariably melted in an induction furnace, either under vacuum or with argon as a protective gas. The vigorous mixing action produced by inductive heating is an additional advantage in producing alloys. Crucible materials can be aluminum oxide; zirconium oxide; or magnesium oxide. Pure calcined calcium oxide was also used in the past. The molten metal is poured into water-cooled copper molds. The refractory platinum group metals, ruthenium and osmium, are seldom needed as pure, solid metals, but may be compacted by sintering. Electric-arc melting and electron-beam melting are sometimes used. Electron-beam purification has recently become an important method of treating iridium. This melting process causes selective evaporation of trace elements and impurities at ca. 2700 °C.

Platinum, palladium, and most of their alloys can be mechanically shaped by forging, rolling, drawing, and by deep drawing, etc., at room temperature, with intermittent heating. Formability decreases with increasing proportions of the other platinum group metals in the ascending series palladium, rhodium, iridium, ruthenium, and osmium. Special techniques are used for weaving and knitting Pt – Rh catalyst gauzes and for drilling extremely small holes in the platinum alloy spinneret nozzles used in the production of fibers.

Platinum and the other metals can also be processed by powder metallurgy. The process of inert gas spraying is used to produce powders. Particle sizes are around 50 μm.

Platinum and palladium can be fusion welded without special precautions by TIG, plasma, and laser welding. Platinum metal can also be diffusion welded by hammering at red heat. A protective gas must be used with rhodium and iridium.

9. Quality Specifications and Analysis [5], [36], [185–198]

9.1. Quality Specifications [185], [200], [201]

In industry, the purity and fineness of platinum and platinum metals are even today often described by the classifications "technically pure" (99 %), "chemically pure" (99.9 %), "physically pure" (99.99 %), and "spectroscopically pure" (99.999 %). These are taken from the recommendations of the Physikalisch-Technische Reichsanstalt [29].

Present practice is to use the following qualities:

99 %	(now hardly used)
99.9 %	
99.95 %	
99.98 %	
99.99 %	
99.999 %	
99.9999 %	(only in exceptional cases)

Other specifications are used for particular applications (e.g., physical properties if used in sensors or microelectronics [202], [203]).

Apart from osmium ASTM standard specifications are available for the chemical composition of some grades of the other refined PGMs (Table 6).

Table 6. ASTM standard specifications for platinum group metals

Metal	Standard	Grades
Pt	B 561 - 94	99.95, 99.99
Pd	B 589 - 94	99.95
Rh	B 616 - 96	99.80, 99.90, 99.95
Ir	B 671 - 81 (Reappr. 1993)	99.80, 99.90
Ru	B 717 - 96	99.80, 99.90

Most commercially available platinum and palladium have purities of 99.95 – 99.98 %. Higher purities usually sell at higher prices. Platinum and palladium are generally supplied as ingots or bares (stamped with the purity). Other platinum metals are supplied mainly as powder.

Platinum and palladium ingots of 500 – 1000 g are preferred for investment purposes in Japan [1].

9.2. Qualitative Analysis

The older wet chemical qualitative reactions have not entirely lost their usefulness.

Platinum can be identified by the precipitation of Pt^{IV} as yellow crystalline $K_2[PtCl_6]$ or $(NH_4)_2[PtCl_6]$ from dilute hydrochloric acid solution by the addition of a KCl or NH_4Cl solution. Other quadrivalent platinum group ions interfere. Analogous to Cassius gold purple, platinum forms platinum purple in weakly acid solutions by reduction with $SnCl_2$, Zn, or Al. The detection limit is ca. 1 mg/L, but Pd, Rh and Au can interfere.

The chemical detection of palladium uses precipitation of Pd^{II} as the yellow, ammonia-soluble dimethyldioxime complex. In contrast to the corresponding Pt^{II} compound, the Pd complex precipitates at low temperatures.

There is no simple specific detection reaction for the mainly trivalent rhodium. If other colored ions are absent, it can be recognized by its raspberry pink color in solutions of rhodium(III) chloride or hexachlororhodate(III) complexes. If KOH is added to a Rh^{III} chloride solution, no precipitation occurs at first; the brown hydroxide precipitates only when ethanol is added. This distinguishes between Rh and other PGMs.

Iridium can be detected by heating a solution in concentrated sulfuric acid to fuming and, after cooling, adding ammonium nitrate. Renewed warming of the solution produces an intense blue color by formation of anionic complexes of Ir^{IV}. Chemically, even microgram quantities of ruthenium can be detected in hydrochloric acid solution by the formation of a blue complex of Ru^{III} with thiourea.

Osmium interferes with Ru detection with thiourea by formation of a red complex of Os^{III}, which can be used for identification of Os. It can be separated from Ru by distillation from nitric acid solution. The stinging odor of (the toxic) osmium tetroxide is also characteristic.

In the PGM industry spectrochemical methods are used almost exclusively for qualitative analysis. The detection sensitivity for platinum in solutions by using the 265.85 nm emission line is only slightly higher with inductively coupled plasma optical emission spectroscopy (ICP-OES) than with flame atomic absorption spectroscopy (FAAS). X-ray fluorescence (XRF) offers quick detection or identification facilities for PGMs in metallic or nonmetallic materials. The 244.8 nm, 340.46 nm, and 363.47 nm lines are used amongst others for the spectro-analytical detection of palladium by classical spectrography, FAAS, or ICP-OES. The surest way of detecting rhodium spectroanalytically is by using the 343.49 nm line with the above-mentioned methods. Frequently used lines of Ir are at 212.68 nm, at 215.27 nm, and at 224.27 nm. For Ir, the detection limit of ICP-OES is two orders of magnitude better than with FAAS. The spectroanalytical detection or identification of Ru by means of FAAS or ICP-OES offers no problems, as a series of lines, e.g., at 240.27 nm or 349.89 nm, are available. Among all PGMs, Os has the lowest detection limit when ICP is used.

9.3. Quantitative Analysis [313]

The methods of analysis used for economic control reasons when buying and selling crude products, alloys, compounds, etc., must reflect the high value of PGMs.

Chemical Assay Methods. For commonly occurring combinations of elements, mandatory methods of analysis and sampling have been developed. The classical methods of separation and gravimetrical determination outlined briefly below are today of minor importance. PGMs present in concentrations from μg/g to a few percent are often subjected to preliminary concentration by *fire assay* with lead (see → Gold, Chap. 9.3.), nickel sulfide, or copper as collector prior to the determination.

Platinum – Base Metal Alloys. Platinum is precipitated as $(NH_4)_2[PtCl_6]$ from a nitrate-free solution of $H_2[PtCl_6]$ by addition of NH_4Cl, the precipitate is heated strongly, which converts it to platinum, and then weighed [314]. Another possibility to determine platinum gravimetrically is the reduction with mercury(I) chloride [315]. For Pt on α-Al_2O_3 catalysts used in the petroleum industry, the possible method of determining platinum is by the coloration produced with tin(II) chloride, which is determined photometrically after extraction with ethyl acetate or after preliminary separation of platinum by reduction with formic acid.

Platinum – Rhodium Alloys. A finely divided black metal is obtained by fusion with zinc followed by treatment with hydrochloric acid. Platinum is dissolved in aqua regia, $(NH_4)_2[PtCl_4]$ is precipitated and heated strongly to yield platinum. The remaining noble metal is reductively precipitated from the mother liquor, combined with the residue remaining after aqua regia treatment, and treated with chlorine at 650 – 700 °C. The $RhCl_3$ obtained, which is insoluble in aqua regia, is heated strongly in a hydrogen atmosphere to produce the metal, which is weighed.

Platinum – Palladium Alloys. Palladium is quantitatively precipitated with dimethylglyoxime from a dilute solution in aqua regia. It is filtered off, washed thoroughly, calcined, reduced to the metal by hydrogen, and weighed. The dimethylglyoxime in solution is decomposed by aqua regia, and platinum is converted to $(NH_4)_2[PtCl_6]$, precipitated as sulfide, isolated as the metal by reduction with hydrogen, and weighed.

Platinum – Iridium Alloys. Platinum – iridium alloys are converted into a finely divided state by fusion with zinc, and removing the zinc with HCl, and the platinum is dissolved with aqua regia. Iridium remains undissolved and, after reduction with hydrogen, is weighed.

The classical dissolution for alloys like Pt – Ir and Pt – Rh beginning with a fusion with zinc can be replaced advantageously by dissolving the alloys in a mixture of HCl and HNO_3 under pressure in a sealed glass tube which is heated in a drying furnace, or by dissolution in a PFA or quartz vessel by using a microwave furnace.

The PGM content of commercial compounds and preparations is usually determined gravimetrically by reduction in solution with hydrazine or by reduction of the solid sample with hydrogen. Reductive precipitation from solution in strong hydrochloric acid by treatment with zinc turnings is an exact method if carefully carried out.

Physical Methods. Physical methods, especially spectroscopic ones, are increasingly being used, because they are inexpensive and, with improvements in the equipment, have sufficient accuracy for economic purposes. If brought into solution by suitable acid treatment and/or fusion, such as Na_2O_2 fusion, and after separation from interfering base metals, the PGMs can be determined simultaneously by X-ray fluorescence, or ICP-OES. Application of the principle of internal standardization (→ Gold, Gold Alloys, and Gold Compounds, Chap. 9.) is mandatory for precise work. The measurement procedures outlined in [316] for Pt and in [317] for Pd can be applied also to the ICP-OES of Rh, Ir, and Ru. The difficult and time-consuming separation of the PGMs from one another is no longer necessary. Therefore, analysis of complex materials containing precious metals, such as PGM ore concentrates containing gold, offers no problems. XRF can also be used for direct anlysis of metallic materials with sufficient accuracy for production control in the manufacturing of defined alloys. Calibration of this method may however be difficult.

9.4. Purity Analysis

The precious metal content of commercial grades of PGMs can be determined more precisely by measuring the total content of impurities by physical methods than by direct determination of the main constituent. The most important of these methods are emission spectroscopy, e.g., with spark or glow discharge excitation. Recently, DC arc spectrometry became important again, because the application of a solid-state detector [318] enables simultaneous background corrections. Spark ablation ICP-OES offers an alternative method. When PGMs have been brought into solution (if necessary by dissolution in a mixture of HCl and HNO_3 under pressure by conventional heating in a sealed glass tube or by using a microwave furnace), the metallic impurities can be determined simultaneously with a ICP-OE spectrometer with a solid-state detector [charge injection device (CID) or segmented-array charge-coupled device (SCD)]. Another solution technique of interest here is ICP-MS. Atomic absorption spectroscopy is of minor importance. Mass spectrometry with glow discharge ablation (GD-MS) is used in special cases, e.g., for 99.995 % or purer metals.

9.5. Trace Analysis

The determination of traces of precious metals is of great importance in mineral prospecting. Here metal contents of only a few μg/g are of interest. Methods such as graphite tube AAS, neutron activation analysis, total reflection X-ray fluorescence analysis, and ICP mass spectrometry (ICP-MS) are used after appropriate enrichment procedures (e.g., by fire assay with nickel sulfide as collector).

The analysis of traces of precious metals in waste materials from the precious metal processing industry, especially refineries, has always been very important in monitoring the economics of the process.

Trace analysis of platinum, palladium, and rhodium has become important in recent years. Whereas both the occurrence and the use of PGMs have hitherto been located, automobile exhaust-gas catalysts now constitutes a widespread use. Although no toxic effects on the environment have thus been reported, methods of analysis must be continually improved as permitted levels in the soil and the atmosphere become increasingly strict. The collection of the PGMs in a nickel sulfide phase by fire assay followed by GFAAS or ICP-MS is suitable for analysis in the ng/g range.

10. Uses [3], [5], [30], [31], [204], [167]

10.1. Jewelry, Coinage, Investment [166]

In recent decades, platinum has been in such demand for jewelry and high-quality watchmaking that more than one-third of all the platinum consumed has been used for this purpose. Of this amount, Japan uses about 90 %, with the United States taking the next largest share. Most of the alloys employed contain 95 % platinum and are stamped PLAT 950. The mechanically more stable alloys Pt 95 – Cu 5 (Pt96/Cu 4), the harder alloys PtW 95/5, Pt 96 – Pd 4, and Pt 90 – Ir 10 are used for mounting precious stones. Pt/Co is used for jewelry casting. A relatively new alloy has a Pt content 99.9 %; the hardness for jewelry is obtained with a small amount of silicon. Forming costs for platinum are considerably higher than those for gold because the metal is more difficult to work, and recycling costs are high. These factors markedly affect total cost.

Very good resistance to corrosion and erosion is required of alloys used for the nib tips of high-quality fountain pens [166]. These alloys are based on ruthenium, iridium, and osmium.

In the Russian Empire in 1828 – 1845, a total of 15 t of platinum was formed into 3-, 6-, and 12-rouble pieces by sintering, minted, and brought into circulation.

The fineness was 99 % platinum or 99.2 % PGM [15]. Between 1977 and 1980, platinum coins were again minted for the Moscow Olympic Games. These were 150-rouble pieces, fineness: 9999, $\frac{1}{2}$ ounce, in five designs. Since then, platinum coins have been officially minted in other states (e.g., a Platinum Noble in the UK, a Platinum Maple Leaf in Canada, and a Platinum Koala in Australia). The standard issue of all these coins contained one troy ounce (31.1035 g) of fine platinum. Small ingots, usually containing 99.995 % platinum, are also minted, intended mainly for small investors. The

use of palladium for this purpose has recently begun on a small scale.

10.2. Apparatus [205–208]

Laboratory Technology. Platinum crucibles, dishes, boats, and electrodes (for electrogravimetry) have long been basic items of equipment in chemical laboratories. In earlier times, the purity of platinum was variable. Moreover, equipment made of pure platinum is not dimensionally stable, so today the more stable alloys Pt 97 – Ir 3, Pt 95 – Au 5, and occasionally Pt 99.7 – Ir 0.3, are mainly used [209].

Another type of dimensionally stable platinum is dispersion-hardened platinum, which contains 0.1 to 0.2 wt % of high-melting oxide (e.g., ZrO_2, Y_2O_3) in finely divided form [88]. The presence of this material reduces the mobility of crystal imperfections. It is produced by precipitation, internal oxidation, or mechanical mixing followed by sintering of the powder. The metal can be unalloyed platinum, or alloys of platinum with rhodium or gold. Rhodium and iridium crucibles can resist extremely aggressive conditions in nonoxidizing atmospheres.

Chemical and Electrochemical Apparatus [210], [319]. The platinum metals and their compounds are of limited use in chemical process technology owing to their high cost. However, platinum components are essential in fluorine chemistry. Bursting disks for protection against excessive pressure are occasionally made of platinum or palladium. Single crystals for the optical industry and for laser technology are grown in crucibles of platinum, rhodium, or (most commonly) iridium [210]. For the electrochemical production of peroxo compounds, such as hydrogen peroxide and peroxosulfate, massive platinum anodes or hollow platinum wires are used, although such use has decreased [319].

Glass Technology [210–213]. The corrosion and erosion resistance of platinum metals toward molten glass is not matched by any other material [88]. High-purity optical glass is best melted in crucibles of unalloyed platinum. Special iron-free glass for sodium vapor lamps, X-ray windows, and cathode-ray tubes is melted in Pt – Rh vessels.

In the automated production of bottles by large holding furnaces, the use of platinum components ensures problem-free operation over long periods. These components include complete feeder systems, stirrers, plungers, skimmers, pouring funnels, thermocouple protection tubes, and level-measuring devices. Nowadays dispersion-hardened platinum in increasing quantities is used in glassmaking technology.

In the manufacture of glass fibers and glass wool, the spinnerets are made of Pt – Rh. Rock wool and slag wool are made by a centrifugal spinning process with platinum centrifuges.

Textile Technology. Nozzles made of platinum alloys such as Au 70 – Pt 30 and Pt 92 – Ir 8 are used for spinning viscose fibers [214].

Soldering Materials. Gold and palladium have been used in some special solders.

10.3. Heterogeneous Catalysts [92], [215–220]

All platinum group metals have strong catalytic activity, especially for hydrogenation reactions. The catalytic properties of the platinum metals group elements can be regarded in relation to those of the elements of group 8 – 10 of the first row of the periodic system. A clear correspondence to the differences in their catalytic behavior has been found in their behavior in coordination formation, in complex formation, and in chemisorption behavior. The bond dissociation energies of the related compounds, complexes, and adsorbed atoms or molecules are weaker for the elements of the first two rows than for those of the third row. Insofar, closer relation exists between the catalytic characteristics of the elements Fe, Co, Ni and Ru, Rh, Pd compared with the elements of the third row. This special energetic situation allows easier intermolecular exchange of π- and $\pi\sigma$ electrons, which takes place in numerous heterocyclic and homogeneous catalytic reaction. Examples are the reaction of carbon monoxide with hydrogen, hydroformylation, reactions with methanol, and the catalysis of reactions with unsaturated hydrocarbons. The heterogeneous catalysts, including supported catalysts, are of major importance in

Table 7. Modification of the catalytic activity of the platinum group metals by additional metals

Base metal	Additional metal	Reaction	Effect of additive
Pt	5 – 20 % Rh	ammonia oxidation	increased NO yield lower Pt losses
Pt	Ge, Sn, In, Ga	dehydrogenation and hydrocracking of alkanes	longer catalyst life due to fewer deposits
Pt	Sn + Re	dehydrocyclization and aromatization of alkanes	increased catalyst activity and stability
Pt	Pb, Cu	dehydrocyclization and aromatization of alkanes	effective aromatization
Pt, Pd, Ir	Au	oxidative dehydrogenation of alkanes, *n*-butene to butadiene, methanol to formaldehyde	better selectivity
Pd	Sn, Pb	selective hydrogenation of alkynes to alkenes	
Pd	Ni, Rh, Ag	alkane deyhdrogenation and dehydrocyclization	hinders coking
Pd	Au, Cd	ethylene and acetic acid to vinyl acetate	

chemical technology. As of 1998, 220 – 240 t of platinum group metals are in use worldwide in the form of supported catalysts. Next in order of importance are bulk metallic catalysts in the form of wire gauzes, highly dispersed carrier-free metals, and oxides. Platinum and palladium are technically much more important than rhodium, ruthenium, iridium, and osmium. Often, two or more platinum group metals are combined (Table 7).

Supported Catalysts. The properties of supported catalysts are very dependent on the interaction between noble-metal catalysts and support material. Classical support materials were asbestos and pumice stone. Important support materials today include aluminum oxides, activated carbon, silicates (e.g., zeolites), and silica, also in combinations and as washcoats. Powdered support materials are used as suspension catalysts; porous pressed shapes, pellets, and monolithic structures in newer applications are used as fixed-bed catalysts. Reactions catalyzed by platinum group metals are listed in Table 8 [319].

The largest demand for platinum catalysts comes from the petroleum industry [221], [222]. These are used in reforming high-boiling fractions from the distillation of crude oil at atmospheric pressure. The most important reactions are the dehydrogenation of alkylcyclohexanes, the isomerization and dehydrogenation of alkylcyclopentanes, and the dehydrogenation and cyclization of alkanes, all of which form aromatics, as well as hydrogenative cleavage of alkanes and naphthenes, and dealkylation of alkylaromatics. The resulting product mixtures are gasolines with high octane numbers. The most important catalysts are Pt – γ-Al_2O_3 Pt, Re/γ-Al_2O_3, and Pt, Sn/γ-Al_2O_3. Other catalysts such as Pt, Ir/γ-Al_2O_3, Pt, Pd/γ-Al_2O_3 or Pt, Ge/γ-Al_2O_3 may also be used.

Another use for PGM catalysts in the petroleum industry is in hydrotreating processes. In addition to base-metal catalysts, platinum and palladium catalysts supported on γ-Al_2O_3 or aluminum silicates (zeolites) are also common. Supported palladium catalysts are used in hydrogen peroxide production.

In the pharmaceutical industry, palladium/active carbon hydrogenation catalysts are widely used, mainly with a palladium content of 5 %. Other catalysts used include rhodium/activated carbon, platinum/activated carbon, palladium – platinum/activated carbon, and iridium/activated carbon, sometimes promoted by other elements, e.g., vanadium, iron, or bismuth. Lindlar catalysts (Pd, Pd/$CaCO_3$) play an important role in pharmaceutical industry. Platinum on electrically conducting carbon black is used as catalyst in fuel cells. Furthermore precious metal formulations are used for exhaust gas catalysts, e.g., to control automotive emissions (→ Automobile Exhaust Control).

Solid Metal Catalysts [223–226]. Platinum – rhodium gauzes are used as catalysts for the large-scale oxidation of ammonia to nitric oxide (Ostwald and Brauer process) in the manufacture of nitric acid (→ Nitric Acid, Nitrous Acid, and Nitrogen Oxides).

The oxidation reaction

$$4\,NH_3 + 5\,O_2 \xrightarrow{900\,^\circ C} 4\,NO + 6\,H_2O$$

takes place in reactors in which many fine-mesh gauzes are stacked above one another. The rate of loss of noble metal varies between 0.2 and

Table 8. Catalysis by platinum group metals

Reaction	Noble metal catalyst		
	Powder	Compound	Support
a) Oxidation and other reactions on noble metal catalysts			
Methanol → formaldehyde	Ag	-	-
Ethylene → ethylene oxide	-	-	Ag/Al_2O_3
Ethylene → acetaldehyde	-	$PdCl_2$ in solution	-
Amines → nitriles	Ag	-	-
Complete oxidation of organic compounds	-	-	Pd/Al_2O_3
Disulfides to sulfoxides	-	RuO_4 in CCl_4	-
Alcohols to aldehydes, ketones	-	RuO_4 in CCl_4	-
Olefins, oxidative cleavage	-	RuO_4 in CCl_4	-
Synthesis or decomposition of organic compounds			
$CO + H_2$ to methane	-	-	Ru/Al_2O_3
to paraffin hydrocarbons	-	RuO_2	$Ru + K_2CO_3$
Addition of CO and H_2 to			
olefins to give alcohols	-	RhO_3	-
olefins to give ketones	-	$RhCl_3$	-
olefins to give esters	-	Rh oxide	-
Addition of silane to olefins	-	-	Pt/carbon Al_2O_3
Cleavage of formic acid	Rh, Os	Ru oxide	Pt/asbestos
Leavage of formaldehyde	Ru, Os	OsO_4	$Pd/BaSO_4$
Inorganic catalysis			
$H_2O_2 \rightarrow H_2O + 1/2\,O_2$	Ru, Pd	-	-
NO → hydroxylamine	-	-	Pt/C PtAu5/C PtIr1/C Rh/C
$NH_3 \rightarrow NO$	PtRh10		
b) Reduction and isomerization on noble metal catalysis			
H_2 addition to			
acetylene and homologs	Rh, Pd, Pt	-	Rh, Pd/Al_2O_3 $Ag/Pd70/30/Al_2O_3$ Pd and $Pt/BaSO_4$
ethylene and homologs	-	-	Ru/C
aromatic hydrocarbons	-	-	Rh/Al_2O_3
aromatic compounds, hydrogenation of the aromatic ring	-	RuO_2	Rh/Al_2O_3
heterocycles, hydrogenation of the aromatic ring	-	-	Rh/Al_2O_3
Reforming of crude oil			
dehydration (cracking) to aromatics			Pd/Al_2O_3
isomerization/Cyclization	-	-	Pt/Al_2O_3 $PtIr/Al_2O_3$
cleavage to alkanes	-	-	Ru/SiO_2
Reduction			
of ketones to alcohols	-	-	Ru/C Pd/C Pt/C
of carboxylic acids to alcohols	-	RuO_2	Ru/C Rh/C
nitrogen compounds to amines	-	PtO_2	Pd/C Pt/C
of disulfides to mercaptans	-	RuO_2	-
of mercaptans to alcohols + H_2S	-	RuO_2	-

2.0 g per tonne of nitrogen reacted, depending on process conditions. If a rhodium – platinum alloy is used, the rate of loss of noble metal is reduced; powdering of the gauze takes place more slowly, and its lifetime is increased.

In the production of hydrogen cyanide (Andrussow process),

$$CH_4 + NH_3 + 3/2\,O_2 \rightarrow HCN + 3\,H_2O$$

platinum – rhodium or platinum – iridium gauzes are used as catalysts (→ Cyano Compounds, Inorganic, Chap. 1.2.1.).

The Degussa process

$$CH_4 + NH_3 \rightarrow HCN + 3\,H_2$$

works with α-Al_2O_3 tubes coated on the inside with a platinum film, forming a platinum – aluminum alloy.

Finely Divided Metallic Catalysts. The so-called platinum metal blacks are well-established catalysts for hydrogenation, dehydrogenation, and oxidation, mainly on a laboratory scale. Their high degree of dispersion is shown by their large specific surface area ($> 20\,m^2/g$). These catalysts are very active even at low temperature, so reactions with labile molecules are possible. In the production of hydrogen peroxide by the anthraquinone process (→ Hydrogen Peroxide, Chap. 4.1.2.1.), palladium black is used to a considerable extent in suspension in a nonaqueous medium or as a fixed bed to catalyze the reduction of alkylanthraquinone to the corresponding hydroquinone. The platinum oxide $PtO_2 \cdot H_2O$ (Adam's catalyst) is used widely as an unsupported catalyst in the pharmaceutical industry for the hydrogenation of alkenes and carbonyl groups. Nishimura's catalyst contains rhodium oxide in addition to platinum oxide.

Catalyst Poisoning and Regeneration [227]. All catalysts are liable to be deactivated (poisoned) by chemical, mechanical, and thermal means. Typical contact poisons for platinum group metals include sulfur, phosphorus, carbon monoxide, hydrogen cyanide, lead, and mercury.

Reactivation is often possible, especially with supported catalysts, by heating, steam treatment, reaction with oxidizing or reducing gases, or washing.

10.4. Fuel Cells

Precious metals are used as anode and cathode catalysts in low-temperature fuel cells such as the AFC (Alkaline Fuel Cell), PAFC (Phosphoric Acid Fuel Cell), PEMFC (Protone Exchange Membrane Fuel Cell) and DMFC (Direct Methanol Fuel Cell). In the interesting PEMFC, which has a significant potential for replacing automobile internal combustion engines, for the cathode side platinum catalysts supported on carbon black are used. For the anode side, mainly because of the required CO tolerance when working with reformed gas platinum/ruthenium catalysts are preferred. For AFC platinum/gold associated with nickel is applied. For PAFC quaternary catalysts with platinum are used; platinum/rhodium/iron was reported as cathode catalysts; on the anode side normally platinum is used. The DMFC, similar to the PEM, works with platinum/ruthenium on the anode side, sometimes additionally containing tungsten, and with platinum on the cathode side.

10.5. Homogeneous Catalysts [30], [88], [228–230], [322]

Homogeneous catalysis by platinum group metals, especially rhodium, has made a major impact in industrial chemistry. Important characteristics of these catalysts include high activity and, therefore, low concentration; high selectivity leading to absence of side products; and mild reaction conditions, enabling the use of low temperature and pressure and facilitating the control of reaction conditions. These factors are of great importance and allow expensive PGMs to be used economically. The discovery of chlorotris(triphenylphosphine)rhodium(I), $[RhCl\{P(C_6H_5)_3\}_3]$, a highly effective hydrogenation catalyst, by Wilkinson [231] opened a new chapter in homogeneous catalysis and in the complex chemistry of rhodium. Today, a number of modified complexes of this and similar types are known.

Homogeneous catalysis by rhodium complexes in oxo synthesis (hydroformylation) has attained great importance and made possible the synthesis of industrially important aldehydes from alkenes (see → Oxo Synthesis). This process was formerly carried out with homogeneous cobalt catalysts.

By using $[RhH(CO)\{P(C_6H_5)_3\}_3]$ as the catalyst, reaction temperatures between 60 and 120 °C and pressures of less than 50 bar could be realized, in contrast to the cobalt catalysts, for which temperatures up to 200 °C and pressures as high as 300 bar are necessary. Often

other compounds are used as precursors from which the actual catalyst is formed during the reaction. These include $[Rh(CO)_2(acac)]$ and $[Rh(CO)\{P(C_6H_5\}_3(acac)]$ (acac = acetylacetonate).

Another approach to oxo synthesis is to work in a biphasic system. In this case a watersoluble ligand like $P(C_6H_4SO_3Na)_3$ and a rhodium salt (e.g., nitrate, acetate, or sulfate) is used as catalyst precursor [325].

The Monsanto acetic acid process uses Rh^{III} iodide as the catalyst precursor. In the cativa process [323] BP uses Ir acetate as complex precursor.

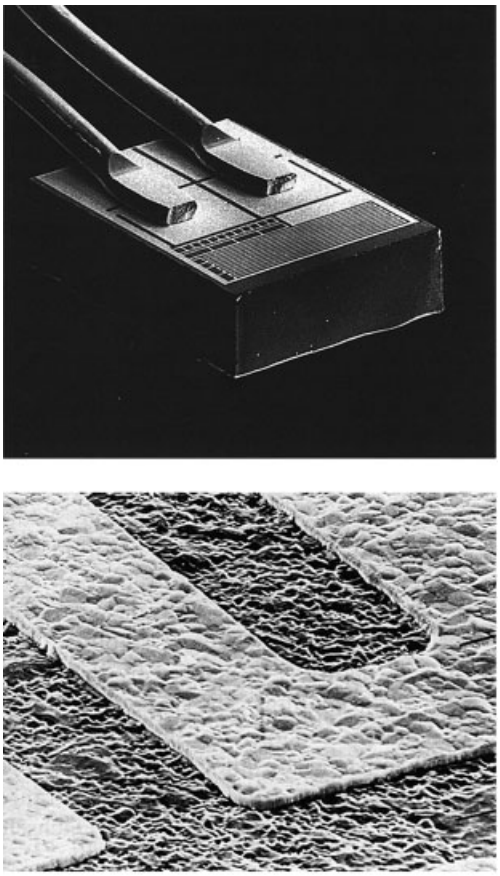

Figure 34. a) Platinum layer resistor element (dim 2×2.3 mm) with welded connecting wires, layer resistance adjusted by laser cuts. b) enlarged view of platinum layer paths on the ceramic substrate. Path width ca. 15 μm (Diehl et al. [167])

Table 9. Commonly used noble metal thermocouple pairs

Positive leg	Negative leg	Operating temperature, °C
Cu	Au 99.4 Co 0.6	−240 to 0
Pt 90 Ir 10	Au 60 Pd 40	0 to 700
Pt 90 Rh 10	Pt	850 to 1600
Pt 70 Rh 30	Pt 94 Rh 6	1000 to 1800
Rh 40 Ir 60	Ir	1000 to 2200

The most important application of platinum in homogeneous catalysis is the vulcanization of silicon rubber. In addition, the synthesis of selected organofunctional silanes can be realized by homogeneous precious metal catalysis [326].

A new era in homogeneous catalysis started with the discovery of asymmetric hydrogenation to produce enantiomerically pure compounds by using metal complexes with chiral ligands [225], [328]. Many other known reactions [322], [324] are now also performed under asymmetric conditions like the isomerization of allylamines [327], [328], allylsubstitution [329] and Heck reactions [327–329]. The products are mainly intermediates like amino acids, amines, and steroids for production of pharmaceutical and agrochemicals. Examples of such ligands are DiAMP [331], BINAP [332], [333], Deguphos [334], Josiphos [335], DuPhos [336], [337]; products include L-Dopa (drug for the treatment of Parkinson's disease), L-Menthol (aroma), and Metolachlor (herbizide). The catalysts are formed in situ by reaction a precursor like $[Rh(cod)_2]BF_a$ or [Rh(cod)acac] (cod = 1,5-cyclooctadiene) with the chiral ligand.

10.6. Automotive Emission Control Catalysts [32], [233–236], [338–345]

The use of supported noble-metal catalysts for the treatment of exhaust gases from automobiles has been introduced beginning in the 1970s, around the world (see also → Automobile Exhaust Control). Regulations controlling the quality of exhaust gases can be complied with only by the use of such catalysts. In particular, carbon monoxide (CO), hydrocarbons (CH_x), and nitrogen oxides (NO_x) must be reduced to a small percentage of their initial concentrations. Metals that have a catalytic action on exhaust gases include Pt, Pd, and Rh and all combinations thereof.

Formulations that can reduce all three harmful components are known as three-way catalysts. The three-way technique also includes control of the air – fuel ratio by means of a λ-sensor (see Section 10.7). Exhaust control catalysts have their optimum activity only if the engine and the catalytic unit are matched. The initially used pellet catalysts have since been replaced by monolith-type catalysts. The exhaust gas passes through numerous parallel channels of extruded bodies (monoliths) that are coated with a catalytically active layer, the so-called wash-coat. The wash-coat consists of the PGM on γ-Al_2O_3, rare earth oxides, and other metal oxides. The oxide mixture is designed to provide high thermal stability and the ability to compensate fluctuations of the air fuel ratio through oxygen storage. In some applications monoliths made from foils are also used.

Each catalytic converter contains approximately 2 g of PGM. The demand for these materials worldwide amounts to more than one-third of the annual production of platinum and more than four-fifths of the annual production of rhodium. In the early 1990s the use of Pd has increased significantly. The PGM catalysts are the only practicable materials for this purpose. The required lifetime of an exhaust-gas catalyst, as laid down by the widely accepted U.S. standard, corresponds to 80 000 or 160 000 km.

Tighter standards in the future and the development of engines with higher fuel efficiency (lean-burn engines, new diesel engines) have spurred the further development of these technologies, e.g., catalysts with the ability to temporarely store hydrocarbons or nitrogen oxides, electrically heated catalysts, and catalysts that also reduce soot emissions from diesel engines. All these catalyst types use PGMs.

Catalysts similar to those used for the treatment of automobile exhaust gases are also used to treat emissions from thermal power plants and industrial waste gases contaminated with harmful organic materials.

10.7. Sensors [30], [88], [166], [167]

Thermocouples. Platinum metals and their alloys are widely used in thermocouples (Table 9), because these have longer lifetimes than those made of base metals and can be used at higher temperature. The most important thermocouple element pair used in industry is Pt 90/Rh 10 – Pt (wire thickness: ca. 0.5 mm), which gives accurate temperature measurements over the range 800 – 1600 °C.

Pt-Pt/Rh thermocouples have improved thermal stability and mechanical strength due to dispersion hardening by alloying the platinum with yttrium [346].

Resistance Thermometry. Platinum resistance thermometers are based on the temperature dependence of the electrical resistivity. The temperature-coefficient of electrical resistance (TCR) of high-purity platinum is 3.927×10^{-3}/K. Basic values of the changes of resistance over temperature ranges (e.g., − 200 °C and + 850 °C) of platinum resistance thermometers are specified in international standards (DIN IEC 751).

At temperatures below about 20 °K the TCR of pure metals becomes too low in comparison to the absolute value of the resistance and cannot be measured. Alloys of rhodium with 0.5 atom % iron or alloys of gold with 1.15 atom % manganese show due to magnetic effects an increasing TCR with decreasing temperature (Kondo effect [347]) to sufficient high, measurable values.

Platinum wire-wound resistors, whose resistance is determined by the length and cross section of the wire, have been almost completely displaced by so-called thin-film resistors, manufactured by thin-film technology. A film of platinum ca. 1 μm thick is deposited on an insulating ceramic substrate by vacuum deposition or cathodic sputtering. The required resistance value is adjusted by laser cutting. A meandrous resistance path is formed in the film of noble metal by cutting with a laser beam or by ionic etching through a masking coating produced by photolithography (Fig. 34). The thickness of the film and the total path length determine the resistance at 0 °C. Path widths are in the region of 20 μm. The connecting leads, which are joined to the thin film by special bonding processes, also consist of noble metals or their compounds. Modern resistance thermometers based on thin-film technology have the advantage over wire-wound resistance thermometers because they can be manufactured by highly automated methods and

require only small amounts of PGM (<1 mg per device).

Iridium-layers have values of thermal expansion which are comparable to those of the ceramic. They are therefore applied on devices for temperature measurements at temperatures down to about 200 °C.

Resistance thermometers are used today in a wide range of technologies, from household appliances to chemical equipment, and from road transportation to air and space travel.

Special Sensor Devices [348]. Oxygen selective sensors ("lambda-probes"), use the high electrical conductivity imparted to solid electrolytes (e.g., ZrO_2) by the presence of oxygen ions, compared with their normal conductivity due to electrons. The ionic conductor is provided on both sides with platinum electrodes and it acts as a separator between two gas spaces. Any difference in the oxygen partial pressure can be determined by measuring the electromotive force (Nernst electromotive force) between the electrodes. Lambda probes are used in automotive cars to maintain an accurate air/fuel ratio and to monitor the performance of the emissions control system (automobile exhaust-gas catalysts in Otto-engines).

Noble-metal catalysts are also used in sensors for detecting combustible gases in air or explosive gas mixtures. Heat evolution is normally detected by sensors known as Pellistors, which consist of platinum wire in a small sphere of Al_2O_3 (diameter: 1.5 mm). The sphere is saturated with a noble-metal catalyst. If a chemical reaction takes place on the catalyst, a temperature increase occurs that can be measured by the change in resistance of the platinum wire. These heat evolution sensors are used in explosion protection equipment.

Platinum and palladium porphyrin complexes and ruthenium(II) diimines are used in optical oxygen sensors. These sensors respond specifically to molecular oxygen by a change in the intensity of luminescence which is emitted from a probe molecule. The luminescence is reversibly quenched by molecular oxygen. Optical sensors have been developed for oxygen detection in hydrogen, carbon dioxide, and as biosensor for determine the level of glucose [349].

Thin palladium layers act as hydrogen filters in optical switching devices. These systems are based on the hydrogen-dependent transition from the metallic state of yttrium hydride $YH_{1.8}$ to the transparent state $YH_{2.9}$ (Fig. 35). Thin yttrium layers are deposited by molecular-beam epitaxy on transparent glass plates covered with a protective niobium layer and an upper palladium layer. The palladium layer impedes oxidation of the yttrium hydride and provides a catalytic surface hydrogen dissociation and diffusion into or out of the yttrium hydride layer. Depending on the hydrogen concentration, the cell switches from mirror reflection ($YH_{1.8}$) to transparency ($YH_{2.9}$). Switching occurs in less than 0.1 s. Applications are expected for automotive and construction parts [351], [352]. An other hydrogen-concentration-dependent device consists of an inner NbH layer, covered on both sides by iron plates. The system behaves ferromagnetically at low H concentration; higher H concentration effect the transition to the antiferromagnetic state. The system is suited as magnetic-field sensor, for instance, in read and write heads of computer hard disks.

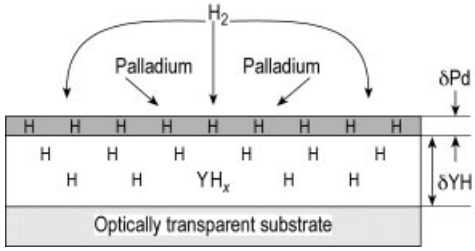

Figure 35. Optical switching device with a palladium thin layer [350]

Osmium -bipyridil-redox polymers combined with a redox enzyme can be used as miniature in vivo sensor for implantation into diabetes patients for continuous monitoring of their blood glucose levels ("Wired enzyme biosensor", [353]). The divalent and the trivalent complexed osmium ions can exchange electrons at high rates. The osmium polymer electrically connects the redox centres of the enzymes by introducing electron relays close to the enzyme active site.

A sol-gel based amperometric biosensor incorporates an osmium redox polymer as mediator for detection of lactate [354].

Table 10. Main applications of platinum group metals in electronics , p. 557[92]

Element	Form					
	Bulk	Bimetal	Surface coat			
			electroplated	sputtered or evaporated	pastes, lacquers	solder connection
Ru			reed contacts	Ru – O for cathodes in special tubes	thick-film circuits (resistor pastes for hybrid circuits)	Mo – Ru high-temperature solders for electron tubes
Rh			slide contacts, coating of Cu disks for thyristors, surface hardening of Ag coatings on pressure contacts	hard contact layers for pressure contacts on Ag surfaces, corrosion protection	metallizing pastes for sensor surfaces	
Pd	bonding wires for transistors, diodes	plug-in contacts	plug-in contacts	multilayer capacitors	conductor and electrode paste for thick-film circuits, multilayer capacitors	Ag – Cu – Pd solders for electron tubes package joining
Pt	test resistors			test resistors, diffusion barriers, IC technology		

10.8. Electrical Technology [205], [237]

In low-current technology, electrical contacts of gold, silver, palladium, rhodium, platinum, and their alloys are used. Contacts made of bulk noble metals are becoming less common than those coated by electrolytic and mechanical processes.

For a long time, large numbers of so-called reed contacts have been used in telephone relays. These are Ag – Pd-, rhodium-, or ruthenium-coated contacts operating under a protective gas. Such contacts must be resistant to sparking, corrosion, and erosion, and must not become welded together. During a lifetime of 30 years, they must carry out millions of switching operations without damage. Furthermore, their contact resistance must be low, and their nominal resistance must be constant.

Platinum sealing wires are also now used only for specialty light bulbs and transmitter tubes. Heating elements made of platinum metals or their compounds are used in resistance-heated furnaces above 1200 °C. Windings of Pt 70 – Rh 30 can be used up to 1500 °C, and ribbon windings of rhodium up to 1800 °C. Furnace elements made of platinum metals can be used without a protective gas. Platinum, rhodium, iridium, and ruthenium are used as susceptor materials in induction furnaces.

Platinum – iridium alloys are useful for spark plug contacts for high-power aeroplane and automobile engines.

10.9. Electronics

The platinum group metals and their alloys are key materials in electronic devices and components (Table 10). Those include "active" discrete semiconductor devices (e.g., diodes, transistors, semiconductor memories), single-crystal integrated circuits (ICs), automotive electronic devices for engine-management systems, automotive oxygen sensors (lambda sensors), and safety devices (ABS, airbag), and so-called passive components (e.g., resistors, capacitors). The metals are mainly applied as surface layers, but also in form of bulk material (sintered from powder) in discrete capacitor and resistor devices.

Important applications of platinum metals in electronics are layer arrays of conductor and resistor paths deposited on substrate plates to form electronic circuits. The main types of electronic circuits are printed circuits (PC) and thick-film and thin-film circuits. They are usually additionally loaded with discrete active and passive electronic components (hybrid technology).

On printed circuits boards, metallizing is primarily carried out by electroplating. In thick-

film circuity, the noble metal layers are produced by the use of paste preparations, which consist of noble metal powders, glass frits, and organic binders. They are applied by screen printing onto ceramic substrates, followed either by drying or by firing at elevated temperatures. The glass frits act as adhesives to the ceramic substrates.

In thin-film circuits the noble metal layers are precipitated on glass or special ceramic substrates by sputtering, by evaporation, or by thermal decomposition of noble metal resinates.

The powders used in the pastes are generally produced by reduction of noble metal salts [355]. The reaction conditions determine shape, size, and specific surface area of the powder. Bimetallic (alloyed) powders are produced by mechanical mixing to give composite blends or by coprecipitation, followed by thermal treatment of the dried powders. Their specific surface area is usually between 1.5 and 10 m^2/g. The particle shapes can be spheroidal, microcrystalline, or flake.

Platinum finds application as barrier layers and as contact pads on semiconductor chips. Platinum interlayers on RuO_2 electrodes of ferroelectric capacitors (lead zirconate titanate PZT) improve their stability and reduce leakage losses [356]. Platinum additions in silver-based metallizing preparations and thick-film conductor pastes for screen printing of hybrid circuits serve as stabilizing component against silver migration. In Ag-, AgPd-, or Au-containing thick-film pastes, Pt powder increases the resistance to scavenging during soldering operations with Sn and Sn – Pb solder alloys. Palladium, pure or in combination with silver, is the most important component in pastes used for the production of the inner electrodes of multilayer capacitors. Special fine-grain powders are needed to obtain sufficiently thin and homogeneous layer structures (Fig. 36). It is produced by chemical reduction of aqueous $Pd(NO_3)_2$ or $PdCl_2$ solution. It is used in pure form or alloyed or mixed with silver powder (mainly of composition Ag – Pd30 wt %). Corresponding electrode pastes are applied to $BaTiO_3$ substrates and fired at temperatures depending between 1100 and 1300 °C, depending on grain size and substrate material. Palladium is also used as component in silver-based conductor pastes and for resistor elements in thick-film circuity. As an additive in silver metallizing pastes, Pd improves the resistance to silver ion migration.

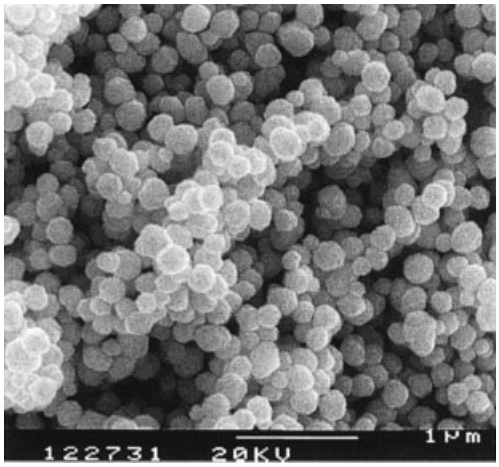

Figure 36. Fine-grain palladium powder for the production of electrodes in multilayer ceramic capacitors [167]

Electroless (by chemical reduction) deposited, thin palladium coatings form the first metallization step for metallizing bare polyimide surfaces, followed by chemical and galvanic deposition of thicker copper layers. This procedure is especially applied for the plating of holes in the polyimide substrates to connect electrical functions in multilayer PCs.

Ruthenium is the most important base element of screen-printed resistor layers in thick-film circuits. These layers are printed with paste preparations consisting of RuO_2 with additions of PbO and Bi_2O_3. They form by firing complex ruthenium compounds of high thermal stability and with very low temperature coefficients of electrical resistivity. Exact resistance values are adjusted by applying small cuts by means of laser beam into the resistor layer ("laser trimming").

Platinum resinates are used for the production of conductive coatings on ceramic substrates onto which further layers can be deposited electrolytically. Thin-film cobalt – platinum alloys and multilayers play an important role as magnetic media on hard disks for magnetic and magneto-optic storage and recording devices. Various organo platinum compounds, combined with special organic dyes, act as sensitizers to effect the transformation of light energy into charge separation in molecular systems. Applications are the light-induced charge transfer in

photovoltaic equipment to produce electric current, also for the photochemical production of hydrogen from water, from hydrogen sulfide, or a nonaqueous source. Organoruthenium compounds, deposited on nanocrystalline titanium dioxide particles, serve as charge-transfer sensitizers in photovoltaic devices [357]. The particles are placed as a thin transparent film between two glass plates which have been coated with transparent conducting oxide layers (Fig. 37). The sensitizer part is separated from the counterelectrode by an triiodide/iodide electrolyte. A platinum coat on the transparent conducting oxide layer catalyzes the cathodic reduction of triiodide to iodide. The light-induced charge transfer from the excited sensitizer to the semiconductor takes the route via interlocking organic groups. The charge separation is accomplished by the fact that the forward electron transfer from the metal via the ligand into the TiO_2 semiconductor membrane is orders of magnitude faster (< 1 ps) than the back diffusion because of the less favorable electronic coupling between the semiconductor and the oxidized ruthenium(III) ion. The overall efficiency for solar light conversion to electricity in this arrangement reaches 10 – 11 %. Higher efficiency is expected to be achieved by replacing the liquid electrolyte by a solid electron-hole conductor in which conduction does not involve mass transfer and which allows the redox level to be matched to the ground state oxidation potential of the sensitizer [358]. Applications are expected in the low-powder range (e.g., solar watches with the photovoltaic cell incorporated in the cover glass), followed by development of interconnected modules, and production of solar tiles.

Organometallic platinum and ruthenium(II) compounds, anchored to a wide-band semiconductor, are promising materials for performing photoinduced hydrogen production [359–363]. A tandem device that achieves the direct cleavage of water into hydrogen and oxygen by visible light works by connection of two photosystems: a first film of nanocrystalline WO_3 and a second, dye-sensitized nanocrystalline TiO_2 film. The blue part of the spectrum is absorbed by the first layer, creating by band gap excitation valence band holes which serve to oxidize water to oxygen. The second film captures the green and red part, transmitted through the top electrode, thereby producing photovoltage, which enables the formation of hydrogen by the conducting electrons [359], [360].

Organic polypyridyl complexes of ruthenium and osmium are able to effect photoinduced electron transfer in molecular structures ("molecular wires" [365]). Such components are needed to link molecular-sized electronic devices (e.g., optical switches, fast recording devices, miniaturized sensors) electronically to the surrounding assembly. Their function is to provide electronic connection between spatially remote redox-active subunits. They must maintain strict stereochemical integrity of the final assembly. Unidirectional electron flow is promoted by external stimulation (e.g., absorption of a photon, application of a potential at an attached electrode, or by chemical means).

RuO_2/TiN storage electrodes serve for the production of dynamic random access memory capacitors [366]. Surface layers of osmium and ruthenium on dispenser cathodes lower the electronic work function [367].

10.10. Coatings [30], [31], [93], [166]

Coatings of platinum metals may be applied to a substrate by various methods.

10.10.1. Coatings Produced by Electrolysis [238], [368], [369]

Platinum. Lustrous thin films for decorative purposes are deposited from an electrolyte obtained by treating potassium tetranitritoplatinate(II), $K_2[Pt(NO_2)_4]$, with sulfuric acid. The active complex is $H_2[Pt(NO_2)_2SO_4]$. This is more stable than other platinum plating baths.

Relatively thick coatings can be deposited from alkaline solutions of sodium hexahydroxy platinate(IV), $Na_2[Pt(OH)_6]$.

A typical bath is sodium hexahydroxoplatinate 20 g/L, sodium hydroxide 10 g/L, pH 13, 75 °C, current density 0.8 A/dm^2. A disadvantage is the low stability due to decomposition of the hydroxy salt. The use of the corresponding potassium salts eliminates this difficulty [370], [371].

Platinum(II) diamminedinitrite, $[Pt(NH_3)_2(NO_2)_2]$ ("P salt"), was formerly

very widely used and today along with $Na_2[Pt(OH)_6]$, is used for coating titanium by aqueous electrolysis. These coatings are of lower quality than those produced by high-temperature electrolysis. Also, extensive pretreatment of the substrate metal is necessary, mainly to prevent cathodic absorption of hydrogen.

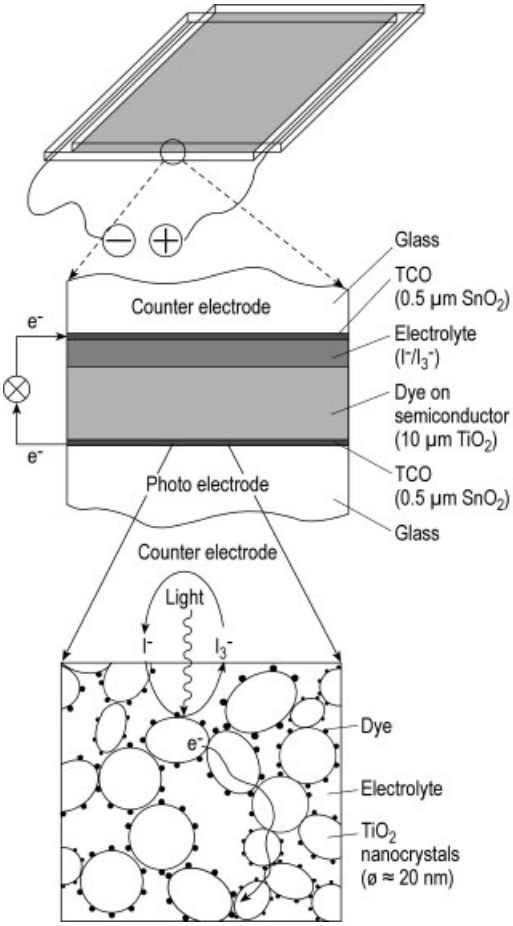

Figure 37. Layout of a nanocrystalline TiO_2 solar cell. The transparent dye derivatized titanium dioxide film is sandwiched between two conducting glass electrodes. It harvest light from both front and back sides. TCO is a transparent conducting oxide coating layer with a highly active, stable platinum electrocatalyst [357] (courtesy Platinum Metals Review)

The Q bath is based on a solution of $[Pt(NH_3)_4]^{2+}$ (Pt concentration 5 – 20 g/L) in a dilute phosphate buffer, pH 10.3 – 10.5. At operating temperatures above 90 °C, high-quality, thick Pt deposites are obtained which are extensively used in the engineering industry when components are subjected to high-temperature corrosion [372], [373].

Rhodium [169], [374–376]. Rhodium coatings are superior to all other PGM coatings, having better hardness, mechanical and chemical stability, and reflectivity. For these reasons, rhodium is more widely used than the other platinum group metals (see also → Electrochemical and Chemical Deposition, Chap. 8.16.).

Electrolytes consisting of rhodium sulfate (10 – 20 g/L Rh) in sulfuric acid can give comparatively thick coatings (10 – 12 μm). Crack formation can be countered by addition of sulfite, selenic acid, magnesium sulfate, or magnesium sulfamate. These baths are used to produce rhodium coatings for heavy-duty slip ring contacts and optical reflectors, and were at one time used for reed contacts in telephone technology.

Rhodium sulfate or phosphate baths containing 2 g/L of rhodium are used for decorative purposes and give strongly reflective thin films (0.1 – 0.5 μm) that prevent silver or white gold from tarnishing. They are also used for plating eyeglass frames.

Palladium [168]. Palladium can sometimes replace gold in coatings for electronics. The plating baths contain palladium(II) diamminedinitrite, $[Pd(NH_3)_2(NO_2)_2]$ (palladium P-salt), with additions of sodium and ammonium nitrates, or sulfamates [377], [378].

Successful results have recently been obtained with plating baths based on organopalladium complexes. Concentrated hydrochloric acid plating baths with dissolving palladium anodes are rarely used now.

Palladium – nickel alloys with 70 – 75 % Pd can be electrodeposited from ammoniacal solutions. Such deposits, overflashed with gold are used as a replacement for gold in the electronics industry [379], [380].

Ruthenium [381–383] is deposited electrolytically from solutions of ruthenium complexes with a bridging nitrido ligand, for example, nitridooctachlorodiaquodiruthenate(IV), $(NH_4)_3[H_2O(Cl_4Ru)–N–(RuCl_4)H_2O]$, the basis of the so-called RuNC baths. The corresponding complex with hydroxo ligands instead of chloro ligands can also be used.

Ruthenium baths based on ruthenium nitrosylsulfamate are also used.

High-Temperature Electrolysis (HTE) [30], [88], [239–241], [384], [385]. Platinum group metals can be deposited electrolytically from cyanide-containing melts in which a soluble anode provides the noble metal that is deposited on the cathode. The metal most widely used in this process is platinum, although iridium is sometimes used.

The molten salt, an eutectic mixture of KCN and NaCN, is maintained at 500 – 600 °C in an electrically heated crucible. The platinum is deposited from the melt onto the substrate, which forms the cathode in the electric circuit.

Process parameters are as follows:

Electrolyte	$K_2[Pt(CN)_4]$ in KCN – NaCN
Operating temperature	500 – 600 °C
Deposition voltage	0.1 – 2 V
Cathodic current density	1 – 5 A/dm^2
Cathodic current yield	> 60 %
Mean rate of deposition	10 – 20 μm/h

Oxidation of the melt and of the metal is prevented by a protective atmosphere of high-purity argon. Suitable substrate materials for HTE platinum deposition include refractory metals such as titanium, niobium, tantalum, molybdenum, tungsten, zirconium; steels; and other metals such as copper. They must generally be pretreated by sandblasting and degreasing.

Advantages of HTE platinum coatings, compared with those produced by aqueous electrolysis, include low porosity; a wide range of film thicknesses (1 – 400 μm), excellent adhesion; good resistance to mechanical impact, thermal shock, and sliding wear; high thermal conductivity; and high ductility.

The largest use is in producing anodes for cathodic corrosion protection [242] (often in the form of expanded metal) or insoluble anodes for electrolytic technology. Platinum-coated molybdenum wire is used in the construction of thermionic tubes and for fusion with glass. High-temperature electrolysis is also used for electroforming.

10.10.2. Coatings Produced by Chemical Reaction [31]

Surface coatings of platinum metals can be deposited onto metallic or nonmetallic substrates by chemical reduction or thermal decomposition of PGM compounds.

Platinum acetylacetonate is used for platinization of titanium. The titanium articles are painted with a solution of this compound, dried, and heated. Repeated application is necessary. The process is not widely used due to allergic reaction and the low quality of the coatings.

Platinum coatings are produced on ceramic tubes by a new technique using a lacquer [243], [386]. Such tubes are used as catalysts for the synthesis of hydrogen cyanide.

Platinum metal resinates are used as ceramic colorants [244], [245]. These can be employed to produce conductive platinum coatings on ceramic substrates, onto which further layers can be deposited electrolytically. Platinum coatings are also used for decorative purposes.

Hexachlororuthenic(III) acid, $H_3[RuCl_6]$, dissolved in alcohol, is painted onto the substrate, dried, and heated strongly in air to produce an extremely adherent coating of RuO_2 [246], [247]. Titanium sheet, coated in this way, has outstanding electrochemical properties and is being used in chloralkali electrolysis as dimensionally stable anodes. The RuO_2 coatings are also used in the production of circuit boards by screen printing.

10.10.3. Coatings Produced by Physical Methods [88], [167]

Platinum can be deposited as a thin film on substrates by evaporating the metal in a high vacuum. Modern thin-film technology is based mainly on the generation of fine particles from a cathode (cathodic sputtering). Due to the high kinetic energy of these metal particles, extremely good adhesion is obtained. In the case of platinum group metals, sputtering targets are used for microelectronics and for resistance thermometers. Ion plating is also used to produce high-quality coatings.

Roll-bonded PGM coatings are no longer important.

10.11. Dental Materials [88]

Platinum and palladium have major importance in dentistry (→ Dental Materials, Chap. 2.3.).

Up to ca. 1900, dental protheses such as gold fillings, crowns, and bridges were made mainly of alloys with the same composition as those used in coinage and jewelry, and later of 22-carat gold (83.3 %). Platinum was used only to make fixing pins.

Important quality requirements such as oral stability, biological compatibility, color, strength, and good functional properties are achieved only by modern dental alloys. These complex alloys, unlike the earlier gold-containing ones, have a very fine-grained, homogeneous crystalline structure, giving much better corrosion resistance. Also, their hardness is controllable over a wide range and can be adapted to particular requirements.

Standard Alloys. Modern alloys for making dentures or prostheses have compositions mainly in the range 65 – 85 % gold, 1 – 10 % platinum, 1 – 4 % palladium, 10 – 15 % silver, 0 – 10 % copper, and 0 – 1 % zinc.

In low-gold alloys, which are cheaper, the gold is replaced partially by silver or palladium. These alloys contain 40 – 60 % gold, 0 – 2 % platinum, 3 – 10 % palladium, 23 – 35 % silver, 0 –12 % copper, and 0 – 5 % tin, indium, and zinc.

Silver-base alloys have very low gold content. Adequate oral stability can be achieved by including at least 25 % platinum group metals or gold. Inexpensive alloys in this category contain at least 20 – 25 % palladium, 0 – 3 % gold, ca. 55 –70 % silver, and copper and zinc, and are therefore no longer yellow, but white.

Alloys with ceramic veneering have been developed in the early 1960s. With the help of these alloys ceramics can be fused on a cast metal framework. The main properties of the alloys necessary for this task are a thermal expansion coefficient similar to that of the ceramic and a solidus temperature sufficiently above the fusing temperature of the ceramic. With early ceramics firing was carried out at 980 °C. Alloys for these ceramics needed a solidus temperature above 1050 °C. To achieve this goal the content of platinum and/or palladium in the alloy had to be increased, which caused whitening of the alloy beside an increase of the melting range. In recent years, low fusing ceramics became available, which allow to use alloys with compositions closer to the standard alloys and alloys with a more yellow color.

Alloys for high fusing ceramics with a high gold content have compositions in the range 75 – 90 % gold, 5 – 10 % platinum, 1 – 10 % palladium, 0 – 3 % silver, with small additions of iron, indium, tin, zinc, and other elements.

Attempts have been made to reduce the cost of these alloys by replacing gold by palladium and silver. Typical Au – Pd – Ag alloys have compositions in the following range: 42 – 62 % gold, 25 – 43 % palladium, 5 – 20 % silver, and 4 – 10 % base metals such as copper, tin, indium, gallium, and zinc. Typical silver-free Au – Pd alloys have the following composition: 45 – 60 % gold, 32 –40 % palladium, and 0 – 14 % of the base metals tin, indium, zinc, or gallium.

Palladium-base dental alloys were introduced in Germany in the early 1980s. These were developed to reduce cost. Silver-free and silver-containing alloys are available. Both contain very small amounts, if any, of gold and platinum. The range of compositions for the silver-containing palladium-based alloys is 53 – 61 % palladium, 28 – 38 % silver, 0 – 12 % tin, indium, zinc, and gallium. For the silver-free palladium-based alloys, the composition range is 73 – 88 % palladium, 0 – 2 % gold, and 12 – 26 % of the base metals copper, cobalt, tin, indium, and gallium.

11. Economic Aspects [5], [10], [24], [30], [46], [48], [57–63],[76], [235], [248–255]

11.1. Supply

The most important PGM-producing mines are located in

- South Africa (Amplats, Lonmin Platinum, Impala, Northham, Kroondal Platinum, Trojan Platinum)
- Russia (Noril'sk Nickel, Koryak and Kondyor mines)
- Canada (Inco, Falconbridge, North American Palladium)
- USA (Stillwater)

– Zimbabwe (Hartley PLatinum, Zimbabwe Platinum, Anglo-American Zimbabwe).

The world total output 1998 of platinum, palladium and rhodium is listed in Table 11 and shown in Fig. 38.

Table 11. World supply of Pt, Pd, and Rh in 1998 by region (in 10^6 oz)

	Pt	Pd	Rh
South Africa	3680	1820	400
Russia	1300	5800	110
North America	285	660	16
Others	135	120	4
Total	5400	8400	530

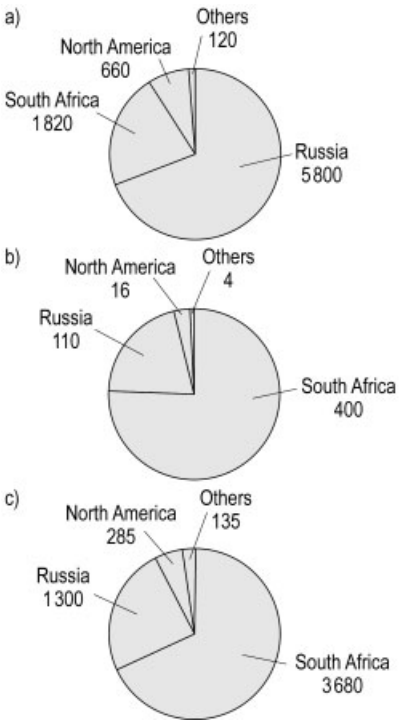

Figure 38. World palladium, rhodium, and platinum supply in 1998 by region (in 10^6 oz)
a) palladium (total 8400); b) rhodium (total 530); c) platinum (total 5400)

In addition to PGMs obtained from ores, a considerable amount of secondary material exists. The extent of this is very variable and fluctuates with time.

Averaged over all the PGMs, it amounts at present to ca. 20 % of primary production, with the greatest emphasis on platinum and especially palladium. The present effort directed toward recovery from spent automobile catalysts will considerably increase the proportional secondary production of platinum and rhodium.

A high proportion of the PGMs is recycled by the users or, in the case of contract suppliers of semifinished products (e.g., catalysts), by the producers. These materials do not appear on the market either on the supply or on the demand side, so supply figures essentially reflect only mined products.

11.2. Demand

Table 12 and Figure 39 show the demand for platinum, palladium, rhodium, iridium and ruthenium by application.

11.3. Prices

The prices of the PGMs are strongly influenced by factors of demand for new technical applications, by availability of supply, and by speculative stock building, leading to considerable differences between monthly and yearly high and low price notations (see Table 13). Rising demand for technical applications, especially in the autocatalyst sector has caused strongly rising prices for platinum, palladium, and rhodium.

11.4. Commercial Aspects

Since 1975, a so-called London free market price for platinum is calculated and published twice daily. Since 1976, a London palladium quotation has also been available. Both of these were only price indications, not to be compared with the fixed quotations for gold and silver. The London Platinum and Palladium Market (LPPM) has since been established and publishes fixed quotations twice daily.

Platinum and palladium are also traded on the New York Mercantile exchange (NYMEX) and on the TOCOM in Tokyo.

Trading in platinum on the New York exchange is in the form of contracts for units of 50 troy ounces. The purity of platinum supplied in fulfillment of such a contract must be at least 99.9 %, and the platinum must originate from

Table 12. PGM demand in 1998 by application (in 10^6 oz)

	Pt	Pd	Rh	Ir	Ru
Autocatalyst	1830	4470	490	36	
Chemical	265	230	31		
Dental		1230			
Electrical	320		8		
Electronics		2070			184
Electrochemical				30	74
Glass	220		34		
Investment	315				
Jewelry	2370	250			
Petrochemistry	130				
Process catalysts				3	87
Others	305	115	10	36	30
Total	5755	8365	573	105	375

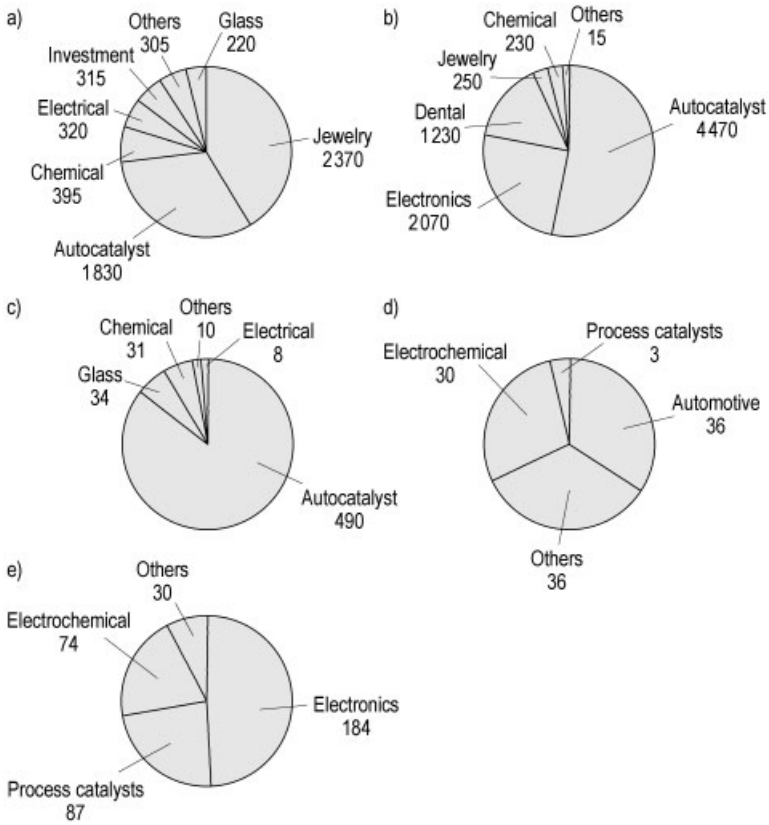

Figure 39. World PGM demand in 1998 by application (in 10^6 oz). Total quantities include recycled material (PA 405×10^6 oz, Pd 175×10^6 oz, Rh 57×10^6 oz)
a) platinum (total 5755); b) palladium (total 8365); c) rhodium (total 573); d) iridium (total 105); e) ruthenium (total 375)

Table 13. PGM price lows and highs (1996 – 1998, in $/oz)

	1996		1997		1998	
	low	high	low	high	low	high
Pt	367	433	342	497	334	429
Pd	114	143	117	245	201	417
Rh	320	345	200	370	370	780
Ir	60	10	110	290	270	575
Ru	29	55	38	41	39	65

a producer or assayer registered with the exchange. For palladium, the contract amount is 100 troy ounces, and the purity must be 99.9 %.

Spot transactions (immediate delivery and payment) are guided by the London quotation and the New York exchange quotation.

Palladium in recent years has found increasing utilization in the production of automotive catalysts. Due to very unstable supply from the main palladium producer of the world, Russia, palladium prices increased more than threefold in the period from December 1996 to May 1998.

12. Toxicology

Platinum is used in therapy; cisplatin is currently used for cancer treatment. Known side-effects are sickness, vomiting, hypersensitivity reactions, ototoxicity and nephrotoxicity [387].

The tetroxides of osmium and ruthenium strongly irritate and corrode the conjunctiva and cornea of the eye and the mucous membranes of the respiratory tract. Temporary toxic effects can occur in bone marrow, liver, and kidneys. The salts of hexachloroplatinic acid $H_2[PtCl_6]$ [*16941-12-1*] have a strong allergenic potential (see below); platinum metal is not allergenic.

The MAK value for platinum compounds has been set as low as 2 μg/m^3 (as Pt), not because of their toxicity but due to allergenic potential. For osmium tetroxide [*20816-12-0*] the MAK value is even lower because of its strong irritating effect: 0.2 μg/m^3. In the United States, TLVs for water-soluble rhodium salts are 1 – 10 μg/m^3. For other platinum group metals toxicological limits have not yet been established.

The median concentrations of platinum in blood, plasma/serum, and urine are about 1 ng/L [388]. The limit for platinum and osmium in industrial effluents is 3 mg/L [389]. In the United States, most platinum group metals are classified as "poisonous" or "very poisonous" in wastewater treatment plants [390].

Automotive Emission Control Catalysts. Toxicological effects of platinum-group metals have been studied increasingly because of their widespread use as automotive emission control catalysts. According to a recent investigation, an engine that is equipped with a modern three-way catalyst emits approximately 15 ng of platinum per cubic meter of exhaust gas in an automobile travelling at a constant speed of 100 km/h; platinum metal is emitted, as well as platinum compounds that are yet to be analyzed [391]. The study could not detect any health risk arising from the use of automotive emission control catalysts, but they undoubtedly have an environmental benefit because they decrease the emission of nitrous oxides, hydrocarbons, and carbon monoxide. Even the most sensitive methods could not detect a correlation between the uptake of platinum from the environment and traffic density [392].

Hexachloroplatinate Allergy. Salts of hexachloroplatinic acid $H_2[PtCl_6]$ are among the strongest synthetic allergens. Symptoms range from irritation of skin and mucous membranes to life-threatening attacks of asthma. Cigarette smokers run a higher risk of sensitization to platinum salts. The symptoms of allergy usually start after several weeks of exposure.

Hypersensitivity to hexachloroplatinate is irreversible, and further exposure must be avoided immediately.

For prophylaxis, exposure to dust and aerosols must be avoided by use of appropriate equipment. Special workplace hygiene is required when hot solutions of hexachloroplatinates, hexachloroplatinic acid, or the products

of their thermal decomposition are handled. Exposed persons must undertake skin tests (prick test) at regular intervals. In this way, sensitization is detected at an early phase and appropriate measures can be taken.

Other platinum and platinum group metal compounds are not known to have an allergenic potential. Recently, skin allergies to the platinum-base alloys used in dental materials have been reported but are still under discussion. The German government has recommended to test the corrosion resistance and biocompatibility of palladium – copper alloys prior to their use in dental restoration [393].

13. References

1. D. McDonald: *A History of Platinum,* J. Matthey Co., London, 1960.
2. D. McDonald, L. B. Hunt: *A History of Platinum and its Allied Metals,* J. Matthey Co., London 1982.
3. *Platinum Metals Review,* **vol. 1 (1957) – vol. 44 (2000),** J. Matthey Co., London.
4. A. Binz: *Edelmetalle – Ihr Fluch und Segen,* W. Limpert-Verlag, Berlin 1943.
5. Fonds der Chemischen Industrie: "Edelmetalle – Gewinnung, Verarbeitung, Anwendung," *Folienserie,* **no. 12,** Frankfurt 1989.
6. *Römpp Lexikon Chemie,* 10th ed., Thieme, Stuttgart 1996 – 1999.
7. H. Römpp: *Chemie der Metalle,* Franckh'sche Verlagshandlung, Stuttgart 1941.
8. P. Walden: *Chronologische übersichtstabellen zur Geschichte der Chemie (Altertum bis 1950),* Springer-Verlag, Berlin 1952.
9. S. Neufeld: *Chronologie Chemie 1800 – 1980,* VCH, Weinheim 1987.
10. H. Quiring: "Platinmetalle", *Die metallischen Rohstoffe,* vol. 16, Enke Verlag, Stuttgart 1962.
11. J. Morin: *La Platine, l'Or Blanc ou le Huitieme Mètal,* Paris 1758.
12. A. Galán, R. Moreno: "Platinum in Eighteenth Century – A Further Spanish Contribution to an Understanding of Its Discovery and early Metallurgy," *Platinum Met. Rev.* **36** (1992) 40 – 47.
13. D. McDonald: "One Hundred and Fifty Years – An Anniversary Review of Johnson Matthey's Role in the Economic History of Platinum", *Plat. Met. Rev.* **11** (1967) 18 – 29.
14. H. M. Severin: *Gold and Platinum Coinage of Imperial Russia from 1701 – 1911,* Crown and Taler Publ. Co., New York 1958.
15. H.-G. Bachmann, H. Renner: "Nineteenth Century Platinum Coins," *Platinum Met. Rev.* **28** (1984) 126 – 131.
16. A. L. von Erengreuth: *Allgemeine Münzkunde und Geldgeschichte des Mittelalters und der neueren Zeit,* R. Oldenbourg, München 1971.
17. L. B. Hunt: "Dr. Hans Merensky – Centenary of the Discoverer of the Platinum Reef," *Platinum Met. Rev.* **15** (1971) 102 – 107.
18. J. Levin: *The Story of Mintek 1934 – 1984,* Council for Mineral Technology (Mintek), Randburg 1985.
19. A. M. Edwards, M. H. Silk: *Platinum in South Africa,* Council for Mineral Technology (Mintek), Randburg 1987.
20. H. Houben: *Festschrift zum fünfzigjährigen Bestehen der Platinschmelze G. Siebert GmbH, Hanau,* Alberti's Hofbuchhandlung, Hanau 1931.
21. D. G. Cooper: *Das Periodensystem der Elemente,* VCH, Weinheim 1972.
22. E. Fluck, K. G. Heumann, *Periodensystem der Elemente, IUPAC-Empfehlungen,* VCH, Weinheim 1985.
23. G. C. Bond, *Plat. Met. Rev.* **44** (2000) 146 – 155.
24. H. P. Münster, *Taschenbuch des Metallhandels,* 8th ed., Metallverlag, Berlin 1989.
25. G. H. Aylward, I. J Findlay: *Datensammlung Chemie,* VCH, Weinheim 1975.
26. National Bureau of Standards Circulars.
27. R. W. Douglass, F. C. Holden, R. J. Jaffee: *High-Temperature Properties and Alloying Behavior of the Refractory Platinum-group Metals,* Battelle Memorial Institute, Columbus 1959.
28. E. M. Savitskii, V. P. Polyakova, M. A. Tylkina: *Palladium Alloys,* Primary Sources/Publ., New York 1968.
29. Doduco: *Datenbuch – Handbuch für Techniker und Kaufleute,* 2nd ed., Pforzheim 1977.
30. The Platinum Asociation: *Platinum, Production, Properties and Applications, a Short Course,* Hill & Knowlten, Frankfurt 1989.
31. *Ullmann* 5th ed., **A21,** 75 ff.
32. *Kirk-Othmer,* 4th ed., **19,** 349.
33. Degussa: *Platingeräte – Beständigkeit und Handhabung,* Hanau 1976.
34. R. Bock: *Aufschlußmethoden der anorganischen und organischen Chemie,* VCH, Weinheim 1972.

35. M. Fieberg, H. Renner, G. Schlamp in G. L. Trigg, E. H. Immergut (eds.): *Encyclopedia of Applied Physics,* **vol. 18,** Wiley– VCH, Weinheim 1997, pp. 235 – 279.
36. Gmelin 8th ed., System no. **68** (Platin), **65** (Palladium), **64** (Rhodium), **67** (Iridium), **63** (Ruthenium), **66** (Osmium).
37. E. Merian: *Metals and their Compounds in the Environment – Occurrence, Analysis, and Biological Relevance,* 2nd ed., VCH, Weinheim 1991.
38. E. Merian (ed.): *Metalle in der Umwelt,* VCH, Weinheim 1984.
39. R. G. Schwab, "Was wissen wir über die tieferen Schichten der Erde ?", *Angew. Chem.* **86** (1974) 612 – 624.
40. C. Bresch, *Zwischenstufe Leben,* Piper-Verlag, München 1977.
41. W. Schreiter: *Seltene Metalle,* vol. 2, Dtsch. Verlag f. Grundstoffind., Leipzig 1961.
42. H. Schneiderhöhn: *Erzlagerstätten,* Fischer Verlag, Stuttgart 1962.
43. V. Tafel: *Lehrbuch der Metallhüttenkunde,* vol. 1, Hirzel Vlg., Leipzig 1951.
44. A. Cissarz: "Verbreitung der Platinmetalle, ihre lagerstättenkundliche Stellung und ihre Anteile an der Weltproduktion, *Erzmetall* **25** (1972) 7 – 16.
45. Gesellschaft Deutscher Metallhütten- und Bergleute (ed.): "Edelmetalle – Exploration und Gewinnung," *Schriftreihe der GDMB,* no. 44, VCH, Weinheim 1986.
46. W. Gocht: *Handbuch der Metallmärkte-Erzvorkommen, Metallgewinnung, Metallverwendung, Preisbildung, Handelsregelungen,* 2nd ed., Springer-Verlag, Berlin 1985.
47. F. Pawlek, *Metallhüttenkunde,* Walter de Gruyter Verlag, Berlin 1983.
48. *Winnacker-Küchler,* 4th ed., **4,** pp. 541 – 572.
49. R. L. Bates, J. A. Jackson (eds.): *Glossary of Geology,* 2nd ed., American Geological Institute, 1980.
50. J. R. Loebenstein: *Mineral Commodity Profiles, 1984,* Bureau of Mines, US Department of the Interior, Washington, annually.
51. V. I. Smirnov (ed.): *Ore Deposits of the USSR,* vol. II, Pitman Publishing, 1977.
52. C. B. Coetzee: *Mineral Resources of the Republic of South Africa,* Department of Mines, Pretoria 1976.
53. J. Lurie: *South African Geology for Mining, Metallurgical, Hydrological and Civil Engineering,* McGraw-Hill, Johannesburg 1981.
54. W. Herzberg: "Der Bergbau Südafrikas," *Metall* **37** (1984) 407 – 411.
55. R. C. Hochreiter, D. C. Kennedy, W. Muir, A. I. Woods: "Platinum in South Africa", *J. S. Afr. Inst. Min & Metall.* **85** (1985) no. 6, 165.
56. A. M. Edwards, M. H. Silk: "Platinum in South Africa," *Mintek Special Publication,* no. 12, Randburg 1987.
57. H. Renner, U. Tröbs: "Rohstoffprofil Ruthenium", *Metall* **42** (1988) 714 – 716.
58. H. Renner, U. Tröbs: "Rohstoffprofil Osmium", *Metall* **44** (1990) 687–690.
59. H. Renner, U. Tröbs: "Rohstoffprofil Iridium", *Metall***40** (1986) 726 – 730.
60. H. Renner, U. Tröbs: "Rohstoffprofil Rhodium", *Metall* **38** (1984) 1002 – 1005.
61. O. Stoppinski: *Platinum Group Metals,* National Research Council, Washington 1977.
62. Hargreaves, Williamson (eds.): *The Platinum Industry, Prospects in Recovery,* Shearson/American Express, London 1984.
63. L. B. Cabril: *Platinum-Group Elements,* The Canadian Institute of Mining and Metallurgy, Ottawa 1981.
64. "Metalle aus dem Meteoritenkrater – Nickel, Kupfer, und Platin in Sudburry", *Neue Zürcher Zeitung* **1990,** no. 258, Nov. 7.
65. T. P. Mohide: *Platinum Group Metals – Ontario and the World,* Ministry of Natural Resources, Ontario 1979.
66. Mintek: "The Sucessful Development of an Industrial Process for the Recovery of Platinum-group Metals from the UG-2 Reef, *Application Report,* no. 1, Randburg, 1987.
67. Mintek: "Recovery of PGMs from the UG-2 Reef", *Research Digest,* no. 29, Randburg, 1988.
68. G. Hempel: "Wozu Polarforschung ?", *Erzmetall* **37** (1984) 577 – 584.
69. H. Renner: *Behandlung und Verwertung von verbrauchten edelmetallhaltigen Katalysatoren, Müll- und Abfallbeseitigung (Müllhandbuch),* 2nd ed., no. 8595, E. Schmidt-Verlag, Berlin 1987.
70. W. Hasenpusch: "Aufarbeitung von Autoabgaskatalysatoren – Die oberirdische Platin-Mine", *EFCE Publ. Ser.* **10** (1989) no. 79.
71. H. Giegerich, C. Hensel: "Autokat-Recycling – ein Beitrag zur künftigen Edelmetallversorgung", *Metall* **44** (1990) 684 – 687.

72. G. A. Jensen: "Platinum-group Metals as a Possible Byproduct of Nuclear Reactions," *Precious Met. Proc. Int. Precious Met. Inst. Conf., 9th Meeting,* Allentown, Pa. 1985, 235 – 252.
73. G. Buckow et al.: *Verwertung von Reststoffen aus einer Wiederaufbereitungsanlage,* Gesellschaft für Reaktorsicherheit (GRS), Köln 1986.
74. H. Röthemeyer: *Endlagerung radioaktiver Abfälle,* VCH-Verlagsges., Weinheim 1991.
75. R. Saager: *Metallische Rohstoffe von Antimon bis Zirkonium,* Bank Vontobel, Zürich 1984.
76. V. E. McKelvey, E. N. Cameron: *The Mineral Potential of United States 1975 – 2000,* University Wisconsin Press, Madison 1973.
77. D. Meadows, E. Zahn, P. Milling: *Die Grenzen des Wachstums,* 15th ed., Deutsche Verlagsanstalt, Stuttgart 1990.
78. L. Trueb: "Das Platin und seine fünf Geschwister einer Geschichte der Platinmetalle", *Neue Zürcher Zeitung* **1989,** no. 298 (Dec. 21).
79. L. Trueb: "Platinmetalle – Vom Bergwerk zum Katalysator", *Neue Zürcher Zeitung* **1989,** no. 194 (Jun. 14).
80. O. S. North: *Mineral Exploration, Mining and Processing Patents,* McGraw Hill, New York, annually.
81. E. Henglein: *Lexikon Chemische Technik,* VCH, Weinheim 1988.
82. G. Agricola: *Vom Berg- und Hüttenwesen,* DTV-Verlag, München 1980.
83. W. Truthe: *Hundert Jahre Gold- und Silber-Scheidung nebst Gewinnung der Platinmetalle,* Degussa 1943, unpublished.
84. E. Jackson: *Hydrometallurgical Extraction and Reclamation,* Ellis Horwood, Chichester 1986.
85. E. M. Savitskii: *Handbook of precious metals,* Hemisphere Publishing Corp., New York 1989.
86. J. C. Mostert, P. N. Roberts: "Electric Smelting at Rustenburg Platinum Mines Ltd. of Nickel-Copper Concentrates containing Platinum Group Metals", *J. S. Afr. Inst. Min. Metall.* **73** (1973) no. 9, 290.
87. C. F. Brugman, D. G. Kerfoot, "Treatment of Nickel-Copper Matte at Western Platinum by the Sherritt Acid Leach Process," *International Nickel Conference,* Toronto, Canada, 1985.
88. Degussa: *Metall – Forschung und Entwicklung im Degussa-Forschungszentrum Wolfgang,* Frankfurt 1991.
89. G. Schlamp: "Noble Metals and Their Alloys," in R. Cahn (ed.): *Materials Science and Technology,* vols. 8/9, Structure and Properties of Nonferrous Metals, VCH, Weinheim 1995, pp. 467 – 587.
90. T. B. Massalski (ed.): *Binary Alloys Phase Diagrams,* Vols. 1 and 2, ASM Metals Park, OH 1986.
91. T. B. Massalski et al. (ed.): *Noble Metal Alloys Phase Diagrams,* The Metallurgical Soc. USA Warrendale, PA 1986.
92. G. Petzow, G. Effenberg: *Ternary Alloys,* VCH, Weinheim 1988.
93. J. E. Hoffmann, *J. Met.* **40** (1988) no. 6, 40 – 44.
94. E. J. Kohlmeyer, K. von Sprenger, *Z. Anorg. Chem.* **257** (1948) 199 – 214.
95. V. Jung, *J. Met.* **33** (1981) no. 10, 42 – 44.
96. Degussa, DE 2 059 235, 1970 (V. Jung).
97. V. Jung: "The Treatment of Automotive Exhaust Catalysts", *TMS/GDMB-Conference,* Productivity and Technology in the Metallurgical Industries, Köln, Sept. 1989.
98. D. Meredith et al., *Stahl + Eisen* **17** (1988) 796 – 800.
99. S. E. Stenkvist, *Steel Times,* 1985, no. 10, 480 – 483.
100. Iron and Steel Society, *Plasma Technology in Metallurgical Processing,* Grand Junction, CO/USA, 1987.
101. Degussa, DE 3 223 501, 1983 (H. Renner, K. Kleis, R. Schlodder).
102. W. C. Heraeus GmbH, DE 2 911 193, 1979.
103. International Precious Metals Institute (ed.): *Current Publications and Proceedings of Conferences and Seminars,* Dave Schneller, Boulder 1979, pp. 92 ff.
104. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 4th ed., John Wiley & Sons, New York 1980, pp. 950 – 966.
105. F. A. Cotton, G. Wilkinson: *Anorganische Chemie,* 4th ed., VCH, Weinheim 1985.
106. G. Wilkinson, R. D. Gillard, J. A. McCleaverty: *Comprehensive Coordination Chemistry,* vol. 5, Pergamon Press, Oxford 1987.
107. F. R. Hartley: *The Chemistry of Platinum and Palladium,* John Wiley & Sons, New York 1973.
108. S. E. Livingstone, in J. C. Bailar et al. (eds.): *Comprehensive Inorganic Chemistry,* **vol. III.** Pergamon Press, New York 1973, pp. 163 – 1370.
109. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* VCH, Weinheim 1988.

110. S. Daamach, G. Cote, D. Bauer: "Separation of Platinum-group Metals in Hydrochloric media: Solvent Extraction of Palladium(II) with Dialkyl Sulfides", IPMI 1977, No. 22.
111. D. W. Agers, E. R. de Ment: "The Evaluation of New LIX Reagents for the Extraction of Copper and Suggestions for the Design of Commercial Mixer-Settler Plants," The Metallurgical Society of AIME, Paper no. A 72 – 82, 1972.
112. Matthey Rustenburg Ref., GB 2 013 644, 1978/1979 (K. J. Shanton, R. A. Grant).
113. Matthey Rustenburg Ref., DE-OS 2 459 099, 1973/1975 (J. B. Payne).
114. L. Manziek (ed.): "Precious Metals Recovery and Refining", *Proceedings of the Precious Metals Recovery and Refining Seminar,* Nov. 12 – 14, 1989, Scottsdale, Arizona.
115. V. H. Apprahamian, G. P. Demopoulos, G. B. Harris: "The Behavior of Impurities during the Solvent Extraction of Platinum Metals with TN 1911 Extractants" in *Precious Metals 1991,* IPMI, Allentown, Pa. 1991, p. 143.
116. W. Hasenpusch: "Die Trennung der Platinmetalle", *CLB, Chem. Labor Betr.* **38** (1987) 454.
117. Ch. Baes, R. E. Mesmer: *The Hydrolysis of Cations,* J. Wiley and Sons, New York 1976.
118. M. Knothe: "Untersuchungen zum Verhalten des Rh, Pd, Ir und Pt im Chloridmedium an Anionenaustauschern", *Z. Anorg. Allgem. Chem.* **463** (1980) 204 – 212.
119. A. P. Evers, R. J. Edwards, M. Fieberg, DE 27 26 558 B 2, 1976/1981.
120. G. Schmuckler, US 4 885 143, 1989.
121. W. Prodinger: *Organische Fällungsmittel in der quantitativen Analyse,* Enke Verlag, Stuttgart 1954.
122. J. Fries, H. Getrost, *Organische Reagentien für die Spurenanalyse,* E. Merck, Darmstadt 1977.
123. D. A. Boyd: *A Novel Approach to PGM Refining,* IPMI, 1992.
124. R. Marr, H.-J. Bart: "Metallsalz-Extraktion", *Chem.-Ing.-Tech.* **54** (1979) 119 – 129.
125. L. Alders: *Liquid-Liquid Extraction – Theory and Laboratory Practice,* Elsevier Publishing, Amsterdam 1959.
126. G. M. Ritcey, A. W. Ashbrook: *Solvent Extraction: Principles and Application to Process Metallurgy,* Elsevier, New York 1984.
127. F. Habashi: "Hydrometallurgy", *Principles of Extractive Metallurgy,* vol. 2, Gordon and Breach, New York 1970.
128. G. H. Morrison, H. Freiser: *Solvent Extraction in Analytical Chemistry,* J. Wiley & Sons, New York 1957.
129. E. J. Barnes, J. D. Edwards: "Solvent Extractions at Inco's Acton Precious Metal Refinery", *Chem. & Ind. (London)* **1982** no. 3, 151 – 164.
130. R. J. Edwards: "Refining of the Platinum-group Metals", *J. Met.* **28** (1976) no. 8, 4 – 9.
131. R. J. Edwards: "Selective Solvent Extractants for the Refining of Platinum Metals", *Proceedings of the International Solvent Extraction Conference ISEC* 7, Vol. 1, CIM, Special Vol. 21, 24 (1979).
132. D. S. Flett: "Solvent Extraction in Precious Metal Refining", *IMPT Seminar,* London 1982.
133. E. M. Elkin, P. W. Bennett: "A new Technique for the Recovery of Palladium and Platinum from Gold Electrolyte", *J. Met.* **17** (1965) no. 3, 252 –254.
134. G. Brauer: *Handbuch der präparativen anorganischen Chemie,* vol. 3, Enke, Stuttgart 1980.
135. R. Gilchrist: "The Platinum Metals", *Chem. Rev.* **32** (1943) 277 – 372.
136. Matthey Rustenburg Ref., GB 2 065 092, 1980 (R. A. Grant).
137. Matthey Rustenburg Refiners: "The Separation Chemistry of Rhodium and Iridium", *Proceeding of a Seminar of the International Precious Metals Institute,* IMPI, Scottsdale, Arizona, 1989.
138. P. Charlesworth: "Separating the Platinum Group Metals by Liquid-Liquid Extraction", *Platinum Met. Rev.* **25** (1981) 106 – 112.
139. L. R. P. Reavill, P. Charlesworth: "The Application of Solvent Extraction to Platinum Group Metals Refining", *Proceedings of the International Solvent Extraction Conference,* ISEC '80 vol. 3, Liège 1980.
140. M. J. Clearne, P. Charlesworth, D. J. Bryson: "Solent Extraction in Platinum Group Metal Processing", *J. Chem. Technol. Biotechnol.* **29** (1979) 210 –214.
141. Matthey Rustenburg Ref., US 3 979 207, 1974 (J. J. McGregor).
142. INCO, US 4 390 366, 1983 (R. K. Lea, J. D. Edwards, D. F. Colton).
143. M. Fieberg, R. J. Edwards: "The Extraction of Gold from Chloride Solutions", *Rep. Natl. Inst. Metall. (S. Afr.)* 1978, no. 1996.
144. A. S. Myerson, K. Toyokura (eds.): "Crystallisation as a Separations Process," *ACS Symp. Ser.* **438** (1990).

145. W. Gösele et al.: "Feststoffbildung durch Kristallisation und Fällung", *Chem.-Ing.-Tech.* **62** (1990) 544 – 552.
146. W. Hasenpusch, K. Bonin: "Kristallisation von Ammonium-Hexachloroplatinat(IV)," *Chem.-Ztg.* 1991, no. 5, 129 – 133.
147. M. Knothe, W. Hosenpusch, *Inorg. Chem.* **35** (1996) 4529 – 4530.
148. R. B. Wilson, W. D. Jacobs: "Separation of Iridium from Rhodium by Extraction with Tributyl Phosphate", *Anal. Chem.* **33** (1961) 1650 – 1652.
149. S. J. Tanaka: "The Recovery and Purification of Rhodium using Superlig™ Technology from a Platinum-group Metals Stream," *IPMI 1992.*
150. Matthey Rustenburg Ref., US 4 382 067, 1982/1983 (R. A. Grant).
151. Matthey Rustenburg Ref., DE-OS 2 457 672, GB 5 682 673, 1973/1979 (J. J. McGregor).
152. PGM Industries, DE-OS 27 26 390, 1977/1979 (J. Baltz, E. Coltrinari).
153. National Institute for Metallurgy (NIM), GB 1 490 815, 1974/1977 (R. J. Edwards).
154. AT & T Technologies, US 4 479 922, 1983 (R. Haynes, A. Jackson).
155. H. Renner: "The Selective Solvent Extraction of Palladium by the Use of Di-Normal-Hexylsulfide", *Rep.-MINTEK,* 1985, no. 217.
156. G. P. Demopoulos: "Solvent Extraction in Precious Metals Refining", *J. Met.* **1986** no. 6, 13 – 17.
157. G. P. Demopoulos: "A Radically New Solvent Extraction Process for Rhodium Recovery", in *Precious Metals 1992,* IPMI, 1992.
158. M. Grote, U. Hüppe, A. Kettrop: "Solvent Extraction of Noble Metals by Formazons(II)," *Hydrometallurgy* **19** (1987) 51 – 68.
159. K.-H. König: "Zur Solventextraktion von Platinmetallen," Achema, Frankfurt 1985.
160. M. Knothe: "Die Fällung des Pd^{II} als $Pd(NH_3)_2Cl_2$ sowie des Verhaltens verschiedener Begleitelemente," *Z. Anorg. Allg. Chem.* **503** (1983) 213 – 223.
161. A. P. Evers, R. J. Edwards, M. M. Fieberg, SA 7 63 681, 1976.
162. R. J. Edwards, M. M. Fieberg, W. te Riele, B. C. Want, US 4 155 750, 1979.
163. A. Prior, K. H. Ohrbach, A. Kettrup, G. Matuschek: "Calcination of Platinum Group Metal Complexes to Form Pure Metals", *IPMI Seminar,* 1988.
164. A. Prior, B. Lerwill: "Cost Management Strategies in Refineries in the 1990s", *IPMI Seminar,* Tempe, Arizona, 1991.
165. V. S. Khain, E. V. Fomina: "Reduction of OsO_4 by Sodium Tetrahydroborate in Acidic Medium", *Russ. J. Inorg. Chem. (Engl. Transl.)* **25** (1980) 307 – 308.
166. Degussa: *Edelmetall-Taschenbuch,* Frankfurt 1967.
167. Degussa: *Edelmetall-Taschenbuch,* 2. Aufl., Hüthig-Verlag, Heidelberg 1995.
168. E. M. Wise: *Palladium,* Academic Press, New York 1968.
169. J. A. Federov: *Rhodium,* Nauka Vlg., Moskau 1966.
170. A. F. Hollemann, E. Wiberg, N. Wiberg: *Lehrbuch der Anorganischen Chemie,* 101. Aufl., W. de Gruyter, Berlin – New York 1995, pp. 1546 (Ru, Os), 1572 (Rh, Ir), 1602 (Pd, Pt), 1628 – 1716 (Carbonyle).
171. J. Huheey, E. Keiter, R. Keiter (eds.): *Anorganische Chemie,* 2. Aufl., W. de Gruyter, Berlin – New York 1995, pp. 735 – 864.
172. A. S. Darling: "Some Properties and Applications of the Platinum Group Metals", *Int. Met. Rev.* **18** (1973) 91 – 122.
173. E. Raub: *Die Edelmetalle und ihre Legierungen,* Springer-Verlag, Berlin 1940.
174. E. Jäneke: *Kurzgefaßtes Handbuch aller Legierungen,* C. Winter Verlag, Heidelberg 1949.
175. J. L. Haughton, A. Prince: *The Constitution Diagrams of Alloys, A Bibliography,* Institute of Metals, London 1956.
176. A. Prince: *Multicomponent Alloy Constitution, Bibliography 1955 – 1973,* Metals Society, London 1978.
177. A. Prince: *Multicomponent Alloy Constitution, Bibliography 1974 – 1977,* Metals Society, London 1981.
178. W. Home-Rothery, G. V. Raynor: *The Structure of Metals and Alloys,* Institute of Metals, London 1954.
179. M. Hansen: *Der Aufbau der Zweistoff-Legierungen,* Springer-Verlag, Berlin 1936.
180. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* 2nd ed., McGraw-Hill, New York 1958.
181. R. P. Eliott: *Constitution of Binary Alloys,* 1st suppl., McGraw-Hill, New York 1965.
182. F. A. Shunk: *Constitution of Binary Alloys,* 2nd suppl., McGraw-Hill, New York 1969.
183. W. G. Moffatt: *The Handbook of Binary Phase Diagrams,* General Electric Company, Schenectady, N.Y., 1976/1977.

184. G. Petzow, G. Effenberg: *Ternary Alloys,* VCH, Weinheim 1988 onwards.
185. J. C. Chaston: "The Purity of Platinum", *Platinum Met. Rev.* **15** (1971) 122 – 128.
186. A. Wogrinz: *Analytische Chemie der Edelmetalle,* Enke Verlag, Stuttgart 1936.
187. F. Ensslin: *Edelmetall-Analyse,* Springer-Verlag, Berlin 1964.
188. Chemikerausschuß der GDMB (ed.): *Analyse der Metalle,* 2nd ed., Springer-Verlag, Berlin 1961.
189. F. E. Beamish, J. C. van Loon: *Analysis of Noble Metals,* Academic Press, New York 1977.
190. S. Kallmann: "A Survey of the Determination of the Platinum Group Elements", *Talanta* **34** (1987) no. 8, 677 – 698.
191. H. G. Sigel, A. Sigel (eds.): *Handbook on Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988.
192. E. D. Goldberg et al.: "Determination of Platinum Sediments," *Anal. Chem.* **58** (1986) 616 – 620.
193. H.-M. Lüschow: "Probenahme, Theorie und Praxis", *Schriftenreihe der GDMB,* VCH, Weinheim 1980.
194. H.-M. Lüschow: "Zur Probenahme von Edelmetallen, Teil 1: Probenahme von metallischen Materialien", *Erzmetall* **42** (1989) 153 – 159.
195. H.-G. Bachmann, E. Koberstein, R. Straub: *Chemie-Technik* **7** (1978) 441 – 446.
196. J. Suchomel, *Fresenius' Z. Anal. Chem.* **300** (1980) 257 – 266; *Fresenius' Z. Anal. Chem.* **307** (1981) 14 – 18.
197. B. Welz: *Atomic Absorption Spectrometry,* VCH, Weinheim 1985.
198. G. Jäger: *Über die Platinbestimmung in Platforming-Katalysatoren,* Dissertation, Mainz 1957.
199. W. Diehl: "Die quantitative Bestimmung der Verunreinigungen der AG, Au, Pt und Rh mit einem 3,4-m-Gitterspektrographen", *Metall (Berlin)* **23** (1969) 587 – 589.
200. F. Mylius: "Reinheitsgrade von Metallen des Handels", *Z. Anorg. Chem.* **74** (1912) 407.
201. F. Mylius, A. Mazzuchelli: "über die Platinanalyse," *Z. Anorg. Allg. Chem.* **89** (1914) 1 – 38.
202. DIN 43 760, 1968.
203. DIN 43 710, 1977.
204. *Gmelin,* **Platin, suppl. vol. A 1,** 8.
205. H. Wolf: "Die Edelmetalle in Forschung und Industrie", *Metall* **12** (1958) no. 7, 3 – 16.
206. J. Sagoschen: "Platin-Laborgeräte", *Chem. Ztg.* **88** (1964) 420 – 429.
207. G. Reinacher: "Platin-Geräte im Labor," *Chem. Anlagen + Verfahren* 1970, no. 1/2, 61 – 63.
208. Degussa: *Platin-Geräte für Labor und Praxis,* no. 1 ff., Frankfurt 1977 ff.
209. G. Reinacher: "Über die Gewichtskonstanz von Geräten aus der Legierung Platin/3 % Iridium bei analytischen Operationen", *Werkst. Korros.* **1964,** no. 1, 84 – 88.
210. G. Reinacher, H. Roters: "Neuere Anwendungen von Platinwerkstoffen in Chemie und Glasindustrie", *Metall* **23** (1969), 570 – 575.
211. G. Reinacher: "Anwendung von Platin-Werkstoffen in Glashütten", *Chem. Anlagen + Verfahren* **1975,** no. 10, 27 – 34.
212. D. Böttger: "The Use of Platinum in the Glass Industry," *Glass* **62** (1985) no. 5, 177 – 178; *Metall* **39** (1985) 748 – 750.
213. M. Rowe: "Noble Metals in the Glass Industry," *Ceram. Eng. Sci. Proc.* **6** (1985) 256 – 268.
214. W. Funk, R. Schumm: "Spinndüsen – Bauteile für die Chemiefaserindustrie", *Chemiefasern + Text. Anwendungstech./Text.-Ind.* **22** (1972) 518 – 521.
215. J. R. Anderson, M. Baudart: *Catalysis, Science and Technology,* vols. 12, Springer-Verlag, Berlin 1981.
216. Fonds der Chemischen Industrie: "Katalyse," *Folienserie,* no. 19, Frankfurt 1986.
217. K. Weissermel, H. J. Arpe: *Industrial Organic Chemistry,* VCH, Weinheim 1978.
218. P. N. Rylander: *Catalysis Hydrogenation over Platinum Metals,* Academic Press, New York, 1967.
219. W. Hölderich, M. Schearzmann, W. D. Mross: "Heterogene Katalysatoren in der chemischen Industrie," *Erzmetall* **39** (1986) 292 – 298.
220. A. J. Bird, A. B. Stiles: *Catalyst Supports and Supported Catalysts,* Butterworth, London 1987.
221. J. Falbe, U. Haserodt: *Katalysatoren, Tenside und Mineralöladditive,* Thieme Verlag, Stuttgart 1978.
222. D. M. Little: *Catalytic Reforming,* Pennwall Pub., Tulsa 1985.
223. W. Ostwald, E. Brauer, in H. Huben (ed.): *Festschrift zum fünfzigjährigen Bestehen der Platinschmelze G. Siebert GmbH, Hanau,* Hanau 1931, pp. 240 – 256.
224. H. Holzmann: "über die katalytische Oxidation von Ammoniak bei der industriellen Salpetersäure-Herstellung", *Chem.-Ing.-Tech.* **39** (1967) 89 – 95.

225. F. Sperner, W. Hohmann: "Rhodium-Platinum Gauzes for Ammonia Oxidation," *Platinum Met. Rev.* **20** (1976) 12 – 20.
226. H. Connor: "The Role of Platinum Alloys Gauzes in the Ammonia Oxidation Process," *Platinum Met. Rev.* **11** (1967) 2 – 9, 60 – 69.
227. E. G. Schlosser: "Die Katalysator-Desaktivierung und -Vergiftung in verfahrenstechnischer Sicht," *Chem.-Ing.-Tech.* **47** (1975) 997 – 1005.
228. R. Ugo: "Aspects of Homogeneous Catalysis," D. Reibel Publ, Dordrecht 1970.
229. F. J. Smith: "New Technology for Industrial Hydroformulation", *Platinum Met. Rev.* **19** (1975) 93 –95.
230. K.-H. Schmidt: "Neuentwicklungen in der homogenen Katalyse", *Chem. Ind.* **37** (1985) 762 – 765.
231. J. Osborn, G. Wilkinson, *Chem. Commun.* 1965, 131.
232. H. Brunner: "Rhodium-Katalysatoren für die enantioselektive Hydrosylierung – ein neues Konzept zur Entwicklung asymmetrischer Katalysatoren", *Angew. Chem.* **95** (1985) 921 – 931.
233. G. J. K. Acres, B. J. Cooper: "Automobile Emission Control Systems", *Platinum Met. Rev.* **16** (1972) 74 –86.
234. W. Weigert, E. Koberstein, E. Lakatos: "Katalysatoren zur Reinigung von Autoabgasen", *Chem.-Ztg.* **97** (1973) 469 – 478.
235. O. Stopinski: *Platinum-group Metals,* National Research Council, Washington 1977.
236. E. Koberstein: "Katalysatoren zur Reinigung von Autoabgasen", *Chemie in unserer Zeit* **18** (1984) no. 2, 37 – 45.
237. A. Keil: *Werkstoffe für elektrische Kontakte,* Springer, Berlin 1960.
238. A. von Krusenstjern: *Edelmetall-Galvanotechnik,* Leuze-Vlg., Saulgau 1970.
239. D. Schlain et al.: "Elektrodesposition of Platinum Metals from Molten Cyanides," *Platinum Met. Rev.* **21** (1977) 38 – 42.
240. J. H. Notton: "Fused Salt Platinum Plating for Industrial Applications," *Platinum Met. Rev.* **21** (1977) 122 – 128.
241. H.-H. Beyer, F. Simon, *Metall* **34** (1980) 1016 – 1018.
242. J. Wurm, S. Sari: "Verchromen mit platinierten Titananoden," *Oberfläche-Surf.* 1989 no. 12, 19 – 21.
243. F. Endter: "Die technische Synthese von Cyanwasserstoff aus Methan und Ammoniak ohne Zusatz von Sauerstoff," *Chem.-Ing.-Tech.* **30** (1958) 305 –310.
244. L. B. Hunt: "The Platinum in the Decoration of Porcelain and Pottery," *Platinum Met. Rev.* **22** (1978) 138 – 148.
245. Degussa: *Glanzgoldfibel,* Frankfurt 1979.
246. V. de Nora, J.-W. Kühn-von Burgsdorff: "Der Beitrag der dimensionsstabilen Anoden (DSA) zur Chlortechnologie," *Chem.-Ing.-Tech.* **47** (1975) 125 –128.
247. J.-W. Kühn-von Burgsdorff: "Anwendung dimensionsstabiler Elektroden in der Elektrochemie," *Chem. Ind. (Düsseldorf)* **28** (1976) 406 – 408.
248. Johnson Mattey: *Platinum,* London, anually.
249. IMPT (ed.): *Precious Metals – New and Review,* **vol. 1 ff,** Allentown, 1976 ff.
250. Degussa, *Geschäftsbericht,* Frankfurt, annually.
251. Knies: "Die Marktsituation der Edelmetalle", *Metall,* annually.
252. J. H. Jolly: *Platinum-group Metals,* US Dep. of the Interior, Bureau of Mines, Washington 1968 ff.
253. K. Bolz, U. Harms, P. Pissulla, H. Schmidt: *Gold, Platinmetalle und Diamanten in der sowjetischen Handelspolitik,* Verlag Weltarchiv, Hamburg 1985.
254. H. Kutzer: "Metalle – zwischen Mangel und Überfluß", *Handelsblatt-Schriftreihe,* Düsseldorf, 1981.
255. F. H. Kemper: "Metalle – Belastung für den Menschen?" *Erzmetall* **40** (1987) 541 – 549.
256. J. Pepys, C. A. C. Pickering, E. G. Hughes: "Asthma due to Inhaled Chemical Agents – Complex Salts of Platinum," *Clin. Allergy* **2** (1972) 391 – 396.
257. M. C. Cleare, E. Hughes, B. Jacoby, J. Pepys: "Immediate (typ I) Allergic Responses to Platinum Compounds," *Clin. Allergy* **6** (1976) 183 – 195.
258. Bundesverband der Pharmazeutischen Industrie: *Rote Liste – Verzeichnis von Fertigarzneimitteln,* Edition Cantor, Aulendorf, annually.
259. A. I. G. McLaughlin et al.: "Toxic Manifestations of Osmium Tetroxide," *Brit. J. Ind. Med.* **2** (1946) 183.
260. G. Rosner, R. F. Hertel: "Gefährdungspotential von Platinemissionen aus Autoabgas-Katalysatoren", *Staub-Reinhalt. Luft* **46** (1986) 281 – 285.
261. A. Bärtsch, C. Schlatter: "Platinemissionen aus Automobil-Katalysatoren," *Schriftenreihe Umweltschutz,* no. 55, Bundesamt für Umweltschutz, Bern 1988.

262. W. Hasenpusch: "Explosive Reactions with Platinum-group metals," *Chem.-Ztg.* **111** (1987) 57 – 60.
263. W. Beck, W. Weigand: "Chemistry of Platinum Metals," *Prax. Naturwiss. Chem.* **40** (1991) 9 – 14.
264. S. D. Robinson, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **85** (1988) 209 – 44.
265. S. D. Robinson, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **87** (1990) 105 – 29.
266. S. D. Robinson, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **88** (1991) 183 – 209.
267. D. T. Thompson, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **89** (1992) 209 – 42.
268. G. Reid, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **90** (1993) 227 – 49.
269. M. A. Halcrow, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **93** (1996) 241 – 267.
270. G. Reid, *Annu. Rep. Prog. Chem., Sect. A: Inorg. Chem.* **91** (1994) 261 – 94.
271. T. J. Ohara, *Plat. Met. Rev.* **39** (1995) 54 – 62.
272. T.-M. Park et al., *Talanta* **44** (1197) no. 6, 973 – 978.
273. M. J. Abrams, *Plat. Met. Rev.* **39** (1995) 14 – 18.
274. S. P. Fricker (ed.): *Metal compounds in Cancer Therapy; Metals in Health and Diseases Series,* **vol. 1,** Chapman and Hall, London 1994.
275. S. P. Fricker, Ruthenium, Nitric Oxide and Disease, *Plat. Met. Rev.* **39** (1995) no. 4, 150 – 154.
276. M. J. Moss, *Plat. Met. Rev.* **39** (1995) 33 – 36.
277. G. R. Newcome et al., *J. Chem. Soc., Chem. Commun.* (1993) 925.
278. G. R. Newcome et al., *Macromol. Symp.* **98** (1995) 467.
279. G. R. Newcome et al., *Angew. Chem. Int. Ed. Engl.* **34** (1995) 2023.
280. D. T. Thompson, *Plat. Met. Rev.* **40** (1996) 23 – 25.
281. D. T. Thompson, *Plat. Met. Rev.* **39** (1995) 108 – 11.
282. V. I. Sokolov, V. Bashilov, *Plat. Met. Rev.* **42** (1998) 18 – 24.
283. T. Lizhen, G. Jinxing, *Plat. Met. Rev.* **38** (1994) 98 – 108.
284. G. L. Selman, J. G. Day, A. A. Bourne, *Plat. Met. Rev.* **18** (1974) no. 2, 46 – 75.
285. Q. Zhang, D. Zhang, S. Jia, *Plat. Met. Rev.* **39** (1995) 167 – 171.
286. J. H. Kim, S.-C. Shin, *Jpn. J. Appl. Phys.* **35** (1996) 342 – 345.
287. Y. Tanaka et al., *J. Magn. Magn. Mater.* **170** (1997) 289 – 297.
288. T. N. Nguyen, P. A. Lee, H.-C. zur Loye, *Science* **271** (1996) 489 – 491.
289. G. Demazeau et al., *J. Magn. Magn. Mater.* **140 – 144** (1995) 165 – 166.
290. E. A. Franco-Ferreira, *Plat. Met. Rev.* **41** (1997) no. 4, 154 – 163.
291. R. W. Schutz, *Plat. Met. Rev.* **40** (1996) 54 – 61.
292. E. L. Semenova, N. Yu. Rusetskaya, V. M. Petyukh, V. Ye. Listovnichiy, *Plat. Met. Rev.* **39** (1995) 174 – 179.
293. D. Goldberg, Y. Xu et al., *Mater. Lett.* **22** (1995) nos. 5, 6, 241 – 248.
294. E. Herda, R. Higuchi-Rusli, Parangtopo, *Mater. Lett.* **30** (1997) nos. 5, 6, pp. 347 – 350.
295. G. A. Moore, *Trans. Electrochem. Soc.* **75** (1939) 237 – 269.
296. F. A. Lewis: *The Palladium – Hydrogen System,* Academic Press, New York 1967.
297. F. A. Lewis, *Plat. Met. Rev.* **38** (1994) 112 – 118.
298. G. Nelin, *Phys. Stat. Solidi* **45** (1971) no. 2, 527 – 536.
299. F. A. Lewis, *Plat. Met. Rev.* **26** (1982) 20 – 27, 70 – 78, 121 – 128.
300. J. B. Hunter, *Plat. Met. Rev.* **4** (1960) no. 4, 130 – 131.
301. A. G. Knapton, *Plat. Met. Rev.* **21** (1977) 44 – 50.
302. J. C. S. Booth, 11th World Hydrogen Energy Conference, Stuttgart, Germany, June 23 – 28, 1996.
303. E. Wicke, G. H. Nernst, *Ber. Bunsenges.* (1964) 224 – 235.
304. G. Sicking, Dissertation, Münster, Germany, 1970.
305. K. U. Neumann et al., *J. Magn. Magn. Mat.* **137** (1994) no. 3, 264 – 268.
306. A. Freund et al., *Catalysis Today* **27** (1996) 279 – 283.
307. K.-A. Starz, Third European Precious Metals Conference, Florence, Sept. 17 – 19, 1997.
308. M. Watanabe, K. Tsurumi, T. Mizukami, T. Nakamura, P. Stonehart, *J. Electrochem. Soc.* **141** (1994) no. 10, pp. 2659 – 2668.
309. T. R. Ralph, *Plat. Met. Rev.* **41** (1997) 102 – 113.
310. M. P. Hoghart, G. A. Hards, *Plat. Met. Rev.* **40** (1996) 150 – 159.
311. M.-L. Tercier, N. Parthasarathy, J. Buffle, *Electroanalysis* **7** (1995) no. 1, 55 – 63.
312. I. M. Wolff, P. J. Hill, *Plat. Met. Rev.* **44** (2000) 158 – 166.

313. H.-M. Lüschow: *Edelmetallanalytik, Eine Übersicht über bewährte klassische und moderne Methoden,* Heft 69 der Schriftenreihe der GDMB, Clausthal-Zellerfeld, 1993.
314. ISO Standard No. 11210, "Determination of Platinum in Platinum Jewellery Alloys – Gravimetric method after precipitation of diammonium hexachlorolatinate".
315. ISO Standard No. 11489, "Determination of Platinum in Platinum Jewellery Alloys – Gravimetric Method after Reduction with Mercury(I) Chloride".
316. ISO Standard No. 11494, "Determination of Palladium in Palladium Jewellery Alloys – ICP-Solution-Spectrometric Method using Yttrium as Internal Standard Element".
317. ISO Standard No. 11495, "Determination of Platinum in Platinum Jewellery Alloys – ICP-Solution-Spectrometric Method using Yttrium as Internal Standard Element".
318. J. Roberts et al.: "Detection limits for trace elements in high purity gold using a DC arc spectrograph with solid state CID detection," Appl. Note No. 60, Thermo Jarrell Ash Corp. Franklin, MA 1995.
319. G. Ertl, H. Knözinger, J. Weitkamp (eds.): *Handbook of Heterogeneous Catalysis,* VCH, Weinheim 1997. E. Drost et al., "Platinwerkstoffe für Hochtemperatur-Einsatz," *Metall* **50** (1996) nos. 7/8, 492.
320. E. Auer, A. Freund, J. Pietsch, T. Tacke, *Applied Catalysis A* **173** (1998) 259 – 271.
321. H. Lindlar, *Helvetica Chimica Acta* **35** (1952) 446 – 450.
322. B. Cornils, W. A. Hermann (eds.): *Applied Homogeneous Catalysis with Organometallic Compounds,* VCH, Weinheim 1996.
323. J. H. Jones, *Plat. Met. Rev.* **44** (2000) 94 – 105.
324. G. W. Parshall, S. D. Ittel: *Homogeneous Catalysis,* Wiley Interscience, New York 1992.
325. B. Cornils, W. A. Hermann (eds.): *Aqueous-Phase Organometallic Catalysis,* Wiley-VCH, Weinheim 1998.
326. B. Marciniec (ed.): *Comprehensive Handbook on Hydrosilylation,* Pergamon Press, Oxford 1992.
327. I. Ojima (ed.): *Catalytic Asymmetric Synthesis,* Wiley, New York 1993.
328. R. Noyori (ed.): *Asymmetric Catalysis in Organic Synthesis,* Wiley, New York 1994.
329. B. M. Trost, D. L. Van Vranken, *Angew. Chem. Int. Ed. Engl.* **31** (1992) 228.
330. O. Loiseleur, M. Hayashi, N. Schmees, A. Pfaltz, *Synthesis* (1997) 1338.
331. D. B. Vineyard, W. S. Knowles, M. J. Sabacky, G. L. Bachmann, D. J. Weinkauf, *J. Am. Chem. Soc.* **99** (1977) 5946.
332. A. Miyashita et al., *J. Am. Chem. Soc.* **102** (1980) 7932.
333. H. Takaya et al., *J. Org. Chem.* **51** (1986) 629.
334. U. Nagel, E. Kinzle, J. Andrade, G. Prescher, *Chem. Ber.* **119** (1986) 3326.
335. A. Togni, *Angew. Chem.* **108** (1996) 1581.
336. M. Burk, J. E. Feaster, R. L. Harlow, *Organometallics* **9** (1990) 2653.
337. M. J. Burk, *J. Am. Chem. Soc.* **113** (1991) 8518.
338. E. Lox, K. Ostgathe, P. Zelenka: "Reduction of Diesel Exhaust Emission by Using Oxidation Catalysis," *SAE Tech. Pap. Ser.* **1990,** 902 111.
339. B. H. Engler, E. Lox, K. Ostgathe, W. Cartellieri, P. Zelenka: "Diesel Oxidation with Low Sulfate Formation for HD-Diesel Engine Application," *SAE Tech. Pap. Ser.* **1991,** 910 607.
340. B. H. Engler, E. Lox, K. Osthgathe, T. Ohata, K. Tsuchitani, S. Ichihara, G. Garr, D. Psaras: "Recent Trends in the Application of Tri-Metal Emission Control Catalysts," *SAE Tech. Pap. Ser.* **1994,** 940928.
341. J. Leyrer, E. Lox, W. Strehlau: "Design Aspects of Lean NO_x-Catalysts for gasoline and Diesel Engine Applications," *SAE Tech. Pap. Ser.* **1995,** 952 495.
342. D. Lindner, E. Lox, R. v. Yperen, K. Ostgathe, T. Kreuzer: "Reduction for Exhaust Gas Emissions by Using Pd-based Three-way Catalysts," *SAE Tech. Pap. Ser.* **1996,** 960 802.
343. W. Strehlau, J. Leyrer, E. Lox, T. Kreuzer, M. Hori, M. Hoffmann: "New Developments in Lean-NO_x Catalysis for Gasoline Fueled Passenger Cars in Europe," *SAE Tech. Pap. Ser.* **1996,** 962 047.
344. W. Strehlau, J. v. d. Tillaart, J. Leyrer, E. Lox, W. Müller: "Evaluation of NO_x-Storage Catalysis for Lean Burn Gasoline Fueled Passenger Cars," *SAE Tech Pap. Ser.* **1997,** 970 465.
345. L. Mußmann, D. Lindner, E. Lox, R. v. Yperen, T. Kreuzer, I. Mitsushima: "The Role of Zirconium in Novel Three-way Catalysts," *SAE Tech. Pap. Ser.* **1997,** 970 746.
346. B. Wu, G. Liu, *Plat. Met. Rev.* **41** (1997) 81 – 85.
347. J. Kondo, *Progr. Theor. Physics* **34** (1965) 372 – 382.
348. W. Diehl, *Degussa Edelmetalltaschenbuch,* Hüthig Verlag, Heidelberg 1995, p. 468.

349. A. Mills, *Plat. Met. Rev.* **41** (1997) 115 – 127.
350. J. N. Huiberts et al., *Nature* **380** (1996) 231 – 234.
351. R. Giessen, *Phys. Bl.* **53** (1997) 1207 – 1212.
352. A. Remhof, *Phys. Rev.* **56** (1997) 2897 – 2899.
353. T. J. Ohara, *Plat. Met. Rev.* **39** (1995) 54 – 62.
354. T.-M. Park et al., *Talanta* **44** (1997) no. 6, 973 – 978.
355. K. Lutz, Hagener Symp. Pulvermet. November 19th/20th Hagen, 1998.
356. H. N. Shareef, K. R. Bellur, A. I. Kingon, O. Auciello, *Appl. Phys. Lett.* **66** (1995) 239 – 241.
357. M. Grätzel, *Plat. Met. Rev.* **38** (1994) 151 – 159.
358. M. Grätzel in R. D. McDonell (ed.): AIP Conf. Proc. 404: Future Generation Photovoltaic Technologies: First NREL Conference, Colorado October 1997, 1-56396-704-9, pp. 119 – 126.
359. K. Sayama, H. Arakawa, *J. Chem. Soc. Faraday Trans.* **93** (1997) no. 8, 1647 – 1654.
360. K. Gurunathan, P. Maruthamuthu, M. V. C. Sastri, *Int. J. Hydrogen Energy* **22** (1997) no. 1, 57 – 62.
361. M. Misara, R. N. Pandey, O. N. Srivastava, *Int. J. Hydrogen Energy* **22** (1997) no. 5, 501 – 508.
362. L. Shivalingappa, J. Sheng, T. Fukami, *Vacuum* **48** (1997) no. 5, 413 – 416.
363. T. Maruyama, T. Yamamoto, *J. Phys. Chem. B.* **101** (1997) no. 19, 3806.
364. R. Argazzi et al., *J. Phys. Chem. B* **101** (1997) no. 14, 2591 – 2597.
365. V. Grosshenny, A. Harriman, M. Hissler, R. Ziessel, *Plat. Met. Rev.* **40** (1996) 26 – 35, 72 – 77.
366. K. Takemura et al., *Jpn. J. Appl. Phys.* **34** (1995) 5224 – 5229.
367. P. Zalm, J. A. Van Stratum, *Philips techn. Rev.* **27** (1966) nos. 3/4, 69 – 75.
368. F. H. Reid: "Electrodeposition of the Platinum Group Metals," *Metallurg. Rev.* **8** (1963) 167.
369. F. H. Reid: "Platinum Metal Plating – A Process and Application Survey," *Trans. Inst. Met. Finish.* **48** (1970) 115.
370. Johnson Matthey & Co., GB 363 569, 1931 (A. R. Powell, A. W. Scott).
371. Engelhard Ind. Inc., FR 1 273 663, 1960 (H. W. Robinson).
372. Johnson Matthey, EP 358 375, 1996 (J. M. Albon, W. J. Davis, P. E. Skinner, S. G. Warren).
373. P. E. Skinner: *Platinum Met. Rev.* **33** (1989) 102.
374. F. H. Reid: "Some Experimental and Practical Aspects of Heavy Rhodium PLating," *Trans. Inst. Met. Finish.* **33** (1956) 105.
375. M. Danemark, Y. Rusconi: "Rhodium – Eigenschaften und elektrolytische Abscheidung," *Galvanotechn. Oberflächensch.* **3** (1062) 287.
376. M. Branik, F. Kummer: "Die galvanische Rhodiumabscheidung," *Galvanotechnik* **72** (1981) 1175.
377. F. H. Reid: "Palladium Plating – Processes and Applications in the United Kingdom," *Plating* **52** (1965) 531.
378. J. A. Abys: "Palladium Plating," *Plat. Surf. Finish.* **82** (1995) no. 8, 67.
379. P.Wilkinson: "The Replacement of Gold by Palladium – Nickel on Connectors," *Trans. Inst. Met. Finish.* **65** (1987) 33.
380. K. J. Whitlaw: "Gold-Flashed Palladium – Nickel for Electronic Contacts," *Trans. Inst. Met. Finish.* **64** (1986) 62.
381. F. H. Reid, J. C. Blake: "Electrodeposition of Ruthenium," *Trans. Inst. Met. Finish.* **38** (1961) 45.
382. G. S. Reddy, P. Taimsalu: "Electrodeposition of Ruthenium," *Trans. Inst. Met. Finish.* **47** (1969) 187; US 3 576 724, 1971.
383. A. F. Bogenschütz, J. L. Jostan, W. Mussinger: "Galvanische Abscheidungen von Rutheniumschichten," *Galvanotechnik* **67** (1976) 98.
384. G.-B. Dick: "Galvanische Abscheidungen von Platin durch Hochtemperaturelektrolyse," *Galvanotechnik* **79** (1988) 4066.
385. G.-B. Dick: "Galvanische Abscheidungen von Platin, Tantal und Iridium durch Hochtemperaturelektrolyse," *Galvanotechnik* **86** (1995) 1427.
386. Degussa, DE 196 17 040, 1998 (L. van Hippel, C. Bussek, J. Sauer, M. Sauer, D. Arntz).
387. H. Marquardt, G. Schäfer (eds.): *Lehrbuch der Toxikologie,* Wiss.-Verl., Mannheim 1994.
388. J. Messerschmidt et al., *Fresenius J. Anal. Chem.* **343** (1992) 391.
389. E. Merian (ed.): *Metals and Their Compounds in the Environment,* VCH, Weinheim 1991.
390. U. Förstner: *Umweltschutztechnik,* 4. Aufl., Springer Verlag, Berlin 1993.
391. Gesellschaft für Strahlen- und Umweltforschung (GSF) und Forschungszentrum Jülich (eds.): Edelmetallemissionen, Zwischenbericht GSF, München 1990.

392. J. Angerer, J. Schaller: "Akute und chronische Toxizität von Spurenelementen," Schriftenreihe der Gesellschaft für Mineralstoffe und Spurenelemente e.V., XIV, Wissenschaftliche Verlagsgesellschaft, Stuttgart 1993, p. 214.
393. Bundesgesundheitsamt (BGA; ed.): Legierungen in der zahnärztlichen Praxis, Informationsschrift des Bundesgesundheitsamtes, BGA Berlin 1993.

Plutonium and Plutonium Compounds

LOTHAR KOCH, Commission of the European Communities, Joint Research Centre, Institute for Transuranium Elements, Karlsruhe, Federal Republic of Germany

1. **Isotopes** 1
2. **Properties** 2
2.1. **Physical Properties** 2
2.2. **Chemical Properties** 4
3. **Production** 5
4. **Alloys** 5
5. **Compounds** 7
6. **Toxicology** 9
7. **References** 11

1. Isotopes

The first isotope of plutonium (Pu) to be discovered was ^{238}Pu, which was generated by SEABORG, MCMILLAN, KENNEDY, and WAHL in 1940 by bombarding uranium with deuterons:

$$^{238}\mathrm{U}\,(\mathrm{d},\,2\mathrm{n})\,^{238}\mathrm{Np} \xrightarrow[2.1\ \mathrm{d}]{\beta^-} {}^{238}\mathrm{Pu}$$

The natural occurrence of plutonium was discovered later. In 1942, SEABORG, PEPPARD, GARNER, and BONNER isolated ^{239}Pu from Canadian pitchblende and Colorado carnotite, in which traces of plutonium (U : Pu=1 : 10^{-12}) are formed continuously by neutron capture. Primordial ^{244}Pu was isolated by HOFFMANN et al. from Precambrian bastnasite in 1971. In 1991, COWAN and ECONOMON discovered the double-beta decay of ^{238}U to give ^{238}Pu.

Larger quantities (ca. 10 t) of anthropogenic plutonium are distributed on earth as a result of military activities (bomb tests, satellite accidents, bombers, rockets, etc.) and releases from nuclear facilities. Thus far, ca. 400 t of plutonium has been produced in nuclear power stations outside centrally planned economies (i.e., excluding the former Eastern-bloc countries). The ^{239}Pu stockpiled in weapons is estimated to exceed 100 t.

Fifteen isotopes of plutonium have been synthesized; long-lived isotopes are listed in Table 1. They can be produced by multiple neutron capture by ^{238}U and hence occur in spent nuclear fuel with varying abundances. Both ^{239}Pu and ^{241}Pu are readily fissile with thermal neutrons, for which the absorption cross sections

Table 1. Nuclear properties of long-lived plutonium isotopes

Mass number	Half-life, a	Decay *	Main α-radiation, MeV	Thermal cross section, barn		Critical mass, kg
				Capture	Fission	
238	87.7	α	5.499 (70.9 %)	500	17	7.8 (metal sphere reflected by 10 cm iron)
	4.80×10^{10}	s.f.	5.457 (29.0 %)			
239	2.41×10^{4}	α	5.155 (73.3 %)	271	742	0.531 (in aqueous solution, fully H_2O reflected)
	5.50×10^{15}	s.f.	5.143 (15.1 %)			
						5.425 (metal sphere, fully H_2O reflected)
240	6.56×10^{3}	α	5.168 (72.8 %)	290	< 0.08	
	1.34×10^{11}	s.f.	5.123 (27.1 %)			
241	14.4	$\beta > 99\,\%$	4.896 (83.2 %)	370	1011	0.260 (in aqueous solution, fully H_2O reflected)
		$\alpha = 2.41 \times 10^{-3}\,\%$	4.853 (21.1 %)			
242	3.76×10^{5}	α	4.901 (74 %)	19	0.2	
	6.80×10^{10}	s.f.	4.857 (16 %)			
244	8.26×10^{7}	α	4.589 (81 %)	1.7		
	6.60×10^{10}	s.f.	4.546 (19 %)			

* $\alpha = \alpha$-decay; $\beta = \beta$-decay; s.f. = spontaneous fission.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_133

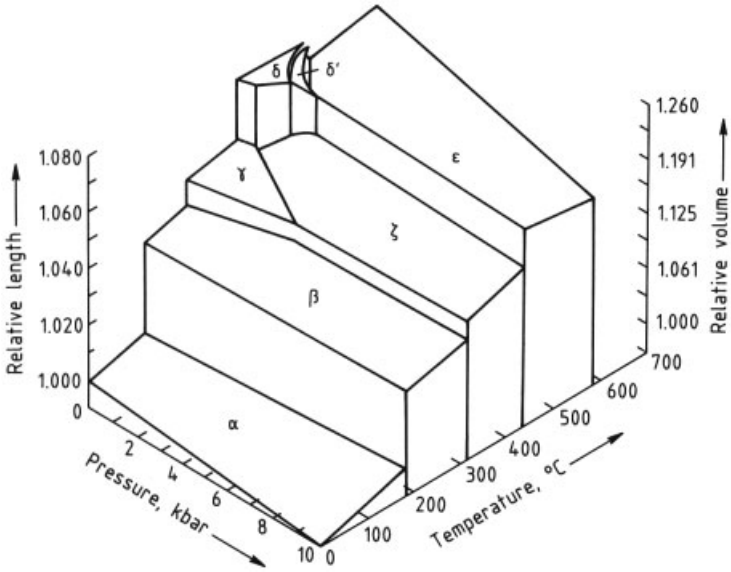

Figure 1. Pressure – temperature – volume diagram of plutonium

are higher by about two orders of magnitude. Consequently they have lower critical masses in aqueous solution since water moderates the velocity of neutrons. For ^{239}Pu the critical mass given for metal spheres refers to H_2O reflectors only (Table 1).

2. Properties

2.1. Physical Properties

Plutonium is a silvery lustrous metal that exhibits five phase transitions (Table 2) between ambient temperature and its melting point (640 ± 2 °C). Under pressure an additional phase occurs.

Table 2. Structural data for plutonium (see also Fig.)

Phase	Symmetry	Unit cell dimensions, pm	X-ray density, g/cm^3	Transformation temperature at STP, °C
α	simple monoclinic	at ca. 21 °C	19.86	
		$a = 618.3 \pm 0.1$ $b = 482.2 \pm 0.1$ $c = 1096.3 \pm 0.1$ $\beta = 101.79° \pm 0.01°$		
				122 ± 4
β	body-centered monoclinic	at ca. 190 °C	17.70	
		$a = 928.4 \pm 0.3$ $b = 1046.3 \pm 0.4$ $c = 785.9 \pm 0.3$ $\beta = 93.13° \pm 0.03°$		
				207 ± 5
γ	face-centered orthorhombic	at ca. 235 °C	17.14	
		$a = 315.9 \pm 0.1$ $b = 576.8 \pm 0.1$ $c = 1016.2 \pm 0.2$		
				315 ± 3
δ	face-centered cubic	at ca. 320 °C	15.92	
		$a = 463.71 \pm 0.04$		
				457 ± 2
δ'	body-centered tetragonal	at ca. 465 °C	16.00	
		$a = 334 \pm 1$ $c = 444 \pm 4$		
				479 ± 4
ε	body-centered cubic	at ca. 490 °C	16.51	
		$a = 363.61 \pm 0.04$		

Because of the different densities of each phase, large changes in expansion are observed during phase transitions, which complicates the use of plutonium for military purposes and as a fuel in nuclear power stations (Fig. 1). Table 3 lists expansion coefficients, Figure 2 shows the heat capacity, and Figure 3 the magnetic susceptibility of the different phases. For further properties, see [1].

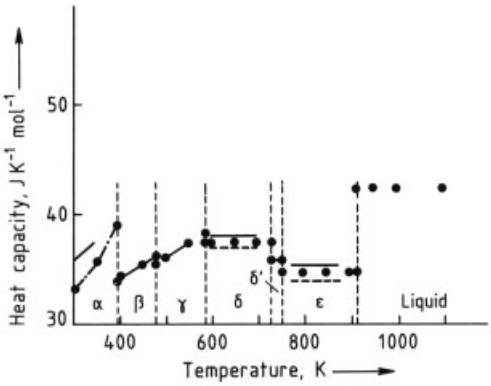

Figure 2. High-temperature heat capacity of plutonium metal
• = Selected values

Table 3. Linear thermal expansion of plutonium [1]

Phase	Principal coefficient	Temperature range, °C	Mean coefficient, 10^{-6} °C^{-1}
α	α_p	− 186 to +101	42.3
	α_1	21 – 104	60
	$\alpha_2 = \alpha_b$		75
	α_3		29
	α_p		54
β	α_1	93 – 190	94
	$\alpha_2 = \alpha_b$		14
	α_3		19
	α_p		42
γ	α_a	210 – 310	-19.7 ± 1.0
	α_b		39.5 ± 0.6
	α_c		34.3 ± 1.6
	α_p		34.6 ± 0.7
δ	α	320 – 440	-8.6 ± 0.3
δ'	α_a	452 – 480	444.8 ± 12.1
	α_c		-1063.5 ± 18.2
	α_p		-65.6 ± 10.1
ε	α	490 – 550	36.5 ± 1.1
Liquid	α_v	664 – 788	93

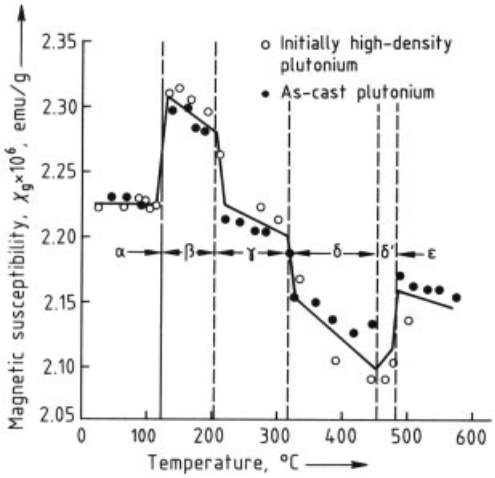

Figure 3. Magnetic susceptibility of plutonium as a function of temperature

Measurements of the thermal conductivity of the different plutonium phases are not in agreement. For α-Pu at ambient temperature it ranges from 0.084 to 0.041 W cm^{-1} K^{-1}.

2.2. Chemical Properties

Plutonium metal readily reacts with oxygen and moisture at ambient temperature, especially in bulk quantities due to self-heating. The surface turns from grayish to dark blue and finally green oxide layers. In aqueous sodium chloride solution, it is converted to Pu(O)H. Plutonium dissolves rapidly in hydrochloric or hydrobromic acid but is only slowly attacked by hydrofluoric acid. It undergoes passivation in nitric acid but dissolves in the presence of fluoride ions. Sulfuric acid converts plutonium slowly to its sulfate.

The behavior of plutonium in aqueous solution is unique. Five oxidation states exist: Pu^{3+}, Pu^{4+}, PuO_2^+, PuO_2^{2+}, and PuO_5^{3-} (Table 4) of which up to four can be present simultaneously in significant quantities. The ions possess characteristic absorption bands that are used in spectrophotometric analysis (Table 5). Standard potential schemes for plutonium in 1 mol/L HCl, neutral solution, and 1 mol/L OH^- at 25 °C are shown (see top of the following page):

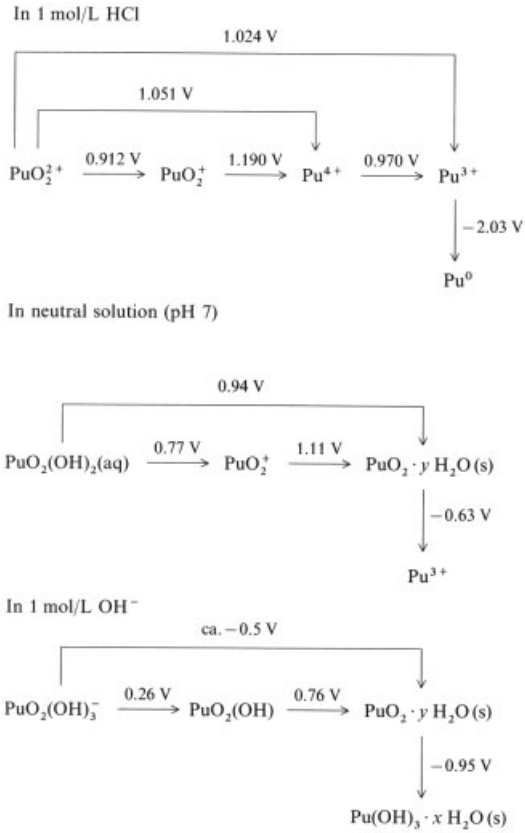

Table 4. Plutonium ions in aqueous solution

Oxidation state	Form	Color
III	Pu^{3+}	blue
IV	Pu^{4+}	tan
IV	$PuO_2 \cdot nH_2O$ (polymer)	bright green, dark green
V	PuO_2^+	red-violet or pink
VI	PuO_2^{2+}	tan, orange (green)
VII	PuO_5^{3-}	blue

Table 5. Predominant absorption bands of plutonium ions

Absorption bands, nm	Molal extinction coefficients			
	$\varepsilon_{Pu(III)}$	$\varepsilon_{Pu(IV)}$	$\varepsilon_{Pu(V)}$	$\varepsilon_{Pu(VI)}$
470	3.46	49.60	1.82	11.25
569	34.30	5.60	17.10	1.75
600	35.30	0.91	0.50	1.35
603	35.40	0.96	0.60	1.20
655	3.10	34.40	1.15	0.90
700	0.75	10.88	0.44	0.25
775	12.40	11.90	9.87	2.90
833	5.25	15.50	4.00	550
953	1.20	0.40	1.76	19.10

The plutonium cation Pu^{4+} is initially hydrolyzed to a monomer before a colloidal polymer forms, preferably in slightly acidic solution. Depolymerization is very slow and requires heating in strong acid or oxidation.

The PuO_2^+ cation disproportionates readily at low acidity (Fig. 4). The Pu^{4+} cation forms complexes with anions of inorganic acids. The nitrate complexes $Pu(NO_3)_x^{(4-x)+}$ (with $x = 1$–4) are employed for the separation of plutonium by solvent extraction and anion exchange. With organic compounds, various complexes have been observed. Ethylenediaminetetraacetic acid forms complexes with plutonium cations of all valences and is often used to decontaminate surfaces or to wash out ingested plutonium.

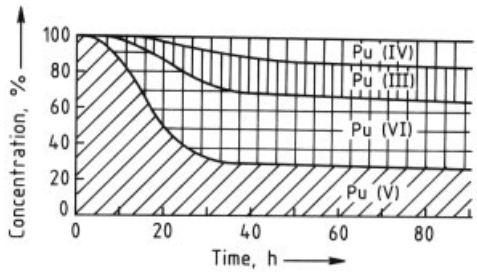

Figure 4. Change in composition of 0.1 mol/L HNO_3 Pu(V) solution by disproportionation

3. Production

Bulk quantities of plutonium are produced through neutron capture by ^{238}U in nuclear reactors. Thus plutonium is a byproduct of nuclear energy generation. The ^{238}Pu used as a heat source in nuclear batteries (for powering devices in space, pacemakers, etc.) is produced by the reaction

$$^{237}Np\,(n,\gamma)\,^{238}Np \xrightarrow[2.1\ d]{\beta^-} {}^{238}Pu$$

together with traces of ^{239}Pu, or by the reaction

$$^{241}Am\,(n,\gamma)\,^{242}Am \xrightarrow[16.0\ h]{\beta^-} {}^{242}Cm \xrightarrow[163\ d]{\alpha} {}^{238}Pu$$

together with ^{242}Pu (17 %), formed by branching decay of ^{242}Am.

Plutonium from spent nuclear fuel is recovered by the PUREX process (liquid – liquid extraction of Pu with tri-*n*-butyl phosphate; see → Nuclear Technology, Chap. 5.4.1.6.,→ Nuclear Technology, Chap. 5.4.1.7.,→ Nuclear Technology, Chap. 5.4.1.8.,→ Nuclear Technology, Chap. 5.4.1.9.). Nonaqueous recovery by electrorefining from molten alkali-metal and alkaline earth metal chlorides is an alternative for metallic fuels. Plutonium metal is prepared from its oxide by reduction with calcium in molten $CaCl_2$ or $CaCl_2$ –CaF_2.

4. Alloys [2,3]

The formation of different phases with varying thermal expansion limits the use of plutonium metal. Therefore, alloys have been extensively studied for military applications and for use as nuclear fuels.

The occurrence of intermetallic compounds can be summarized as follows:

1) Group 3, lanthanides, actinides, and group 4: Either no intermetallic compounds are observed, or plutonium-rich phases with a broad range of existence are found.
2) Groups 5 and 6 do not form intermetallics.
3) Group 7: Only a Laves-phase-type compound PuM_2 exists; PuM_2 compounds are observed with most metallic elements.
4) Groups 8 – 12: With increasing valence, the occurrence of intermetallic compounds in the binary phase diagrams shifts from the plutonium-rich side toward lower plutonium concentrations. Exceptions are the systems Pu – Pd, Pu – Pt, and Pu – Au on the plutonium-rich side, and the occurrence of PuM_2 phases in group 8.
5) Groups 13 and 14: Although the system Pu – B can be fitted into the systematics of the previous category, intermetallics in groups 13 and 14 spread across the whole concentration range. This is due to the occurrence of *p*-electrons in the base metals M. Compounds very rich in plutonium are absent; however, compounds of the type Pu_3M and PuM_3 have been observed.
6) Group 15: These compounds form the borderline of intermetallic compounds. Beyond this, toward groups 16 and 17, saltlike compounds with covalent and ionic structures are observed.

A compilation of phase diagrams and thermodynamic properties of plutonium alloys can be found in [2,3].

The plutonium – molybdenum system (Fig. 5) is an example of immiscibility in the solid state. On solidification, the six normal-pressure modifications of plutonium occur as separate phases besides molybdenum. The same general appearance is shown by the phase diagrams of the systems Pu – Cr, Pu – Nb, Pu – Ta, Pu – V, and Pu – W. Some base metals of this category are useful as containers for liquid plutonium.

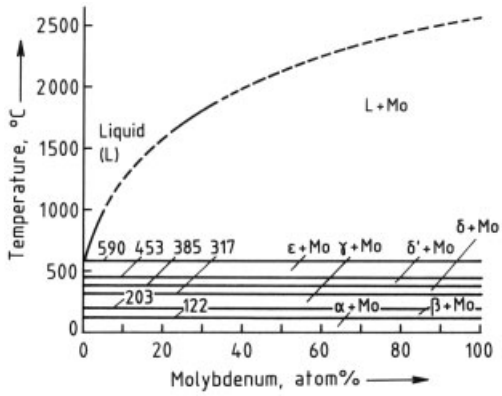

Figure 5. Plutonium – molybdenum phase diagram

The plutonium – lanthanum system (Fig. 6) is an example of systems that have a miscibility gap in the melt. In this system, no compound formation has been observed. Similar systems are Pu – Pr, Pu – Nd, Pu – Sm, and Pu – Y.

The plutonium – aluminum system (Fig. 7) and the plutonium – gallium system are typical examples of δ-retainer systems. In both systems, the δ-phase is stabilized over a wide temperature range (from room temperature to > 600 °C) and a high plutonium concentration (3 – 10 atom % Al and 3 – 8 atom % Ga). Thus, these systems are useful for practical applications of self-heating bulk quantities as in nuclear warheads, for example, where pure plutonium metal would change its geometry due to the differing expansion coefficients of the various phases.

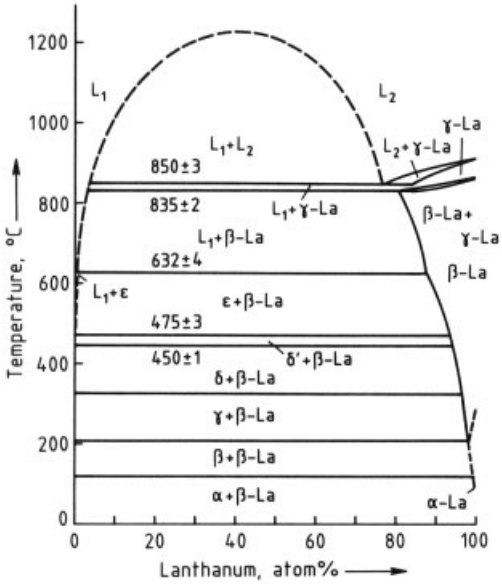

Figure 6. Plutonium – lanthanum phase diagram

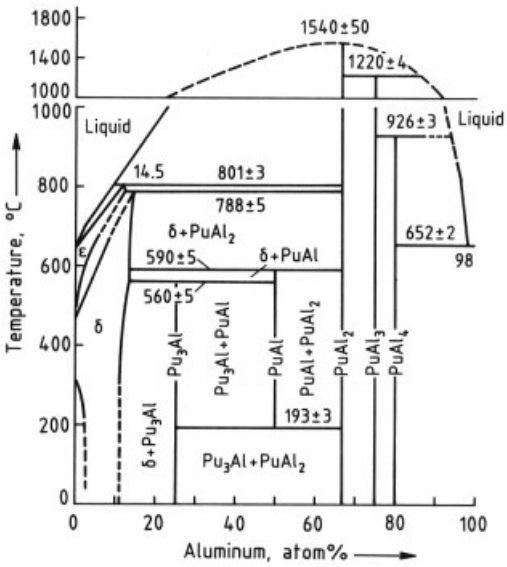

Figure 7. Plutonium – aluminum phase diagram

The plutonium – nickel system (Fig. 8) is a typical example of a system with an element of groups 8 – 10. Such systems are characterized by a large number of intermetallic compounds. The systems Pu – Fe, Pu – Co, Pu – Ru, Pu – Rh, Pu –Pd, Pu – Os, and Pu – Pt are similar.

As an example of high solubility in plutonium, the Pu – U system is shown in Figure 9. Even though complete miscibility is observed in this case, the system is unsuitable for practical applications because the phase transformations leading from one allotrope to the next occur throughout the whole concentration range, preventing the use of such alloys over a wide temperature range at high plutonium concentrations (> 70 atom % Pu). At lower plutonium concentrations, in the range of 35 – 70 atom % U, ζ-stabilization takes place, and the ζ-phase is stable up to 600 °C near 65 atom % U.

The Pu – U phase diagram exhibits many phases, and the low melting point of the alloy limits its use as a nuclear fuel. The addition of zirconium (ca. 10 %) raises the melting point and stabilizes the swelling behavior of the fuel under irradiation.

5. Compounds [2,3]

Like the other actinides, plutonium forms hard, black, metallic hydrides of the type PuH_{2+x} and PuH_3. They are inert to cold water but react slowly at 90 °C.

Four carbides of Pu exist: Pu_3C_2, PuC_{1-x}, Pu_2C_3, and PuC_2 (Fig. 10). They decompose at high temperature with peritectic reactions, are oxidized when heated, and dissolve in oxidizing acids with the formation of CO_2 and carboxylic acids (e.g., malonic acid).

Several silicides have been observed: PuSi, Pu_3Si_5, $PuSi_2$, Pu_5Si_3, and Pu_3Si_2. They have a metallic appearance and are hard, brittle, and pyrophoric. They are oxidized in air and are attacked by water.

Pnictides exists for the elements N, P, As, Sb, and Bi: PuN is black (*mp* 2570 ± 30 °C, under 100 kPa N_2), reacts slowly with cold but readily with hot water, and dissolves in mineral acids; PuP is a dark-gray compound that melts at 2600 °C under argon; PuAs has a gray metallic appearance; PuSb melts at 1980 ± 30 °C under 300 kPa Ar.

Chalcogenides exist for the elements O, S, Se, and Te: PuO_2 is a constituent of nuclear fuels. The plutonium – oxygen phase diagram exhibits the existence of several oxides; their crystal data are listed in Table 6. The stoichiometric, metallically lustrous $PuO_{1.5}$ melts at 2085 ± 25 °C

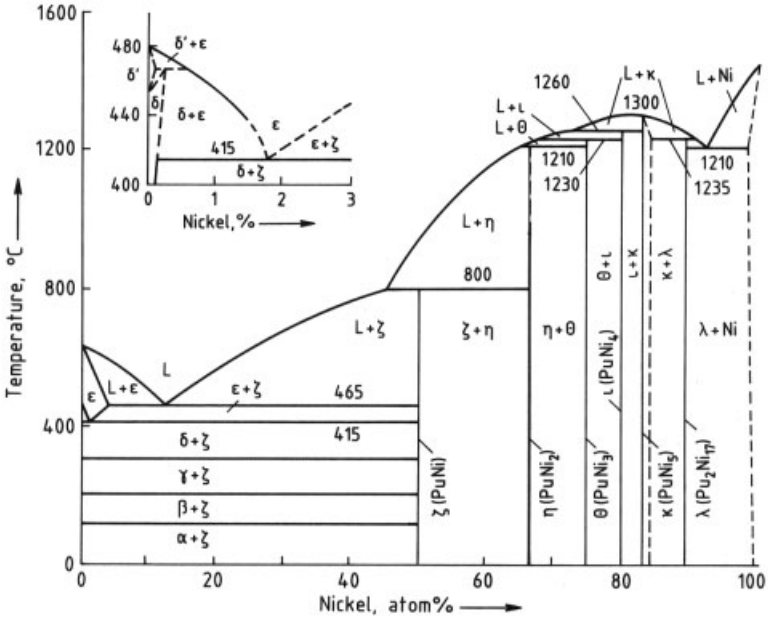

Figure 8. Plutonium – nickel phase diagram

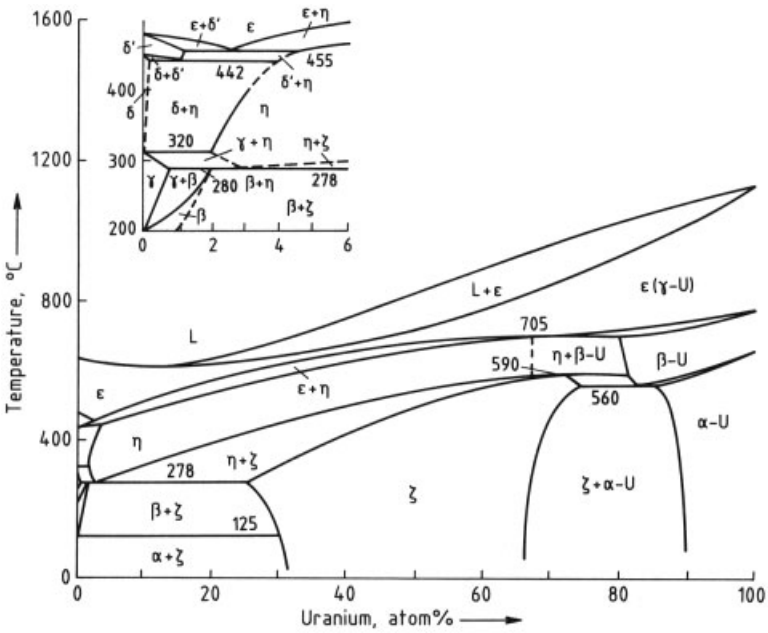

Figure 9. Plutonium – uranium phase diagram

and PuO_2 at $2390 \pm 20\,°C$. Of technical relevance is the substoichiometric dioxide PuO_{2-x} ($x = 1.61 - 1.98$), which acts as an oxygen sink in nuclear fuels, absorbing oxygen liberated during the fission process (and not bound by fission products). It is an olive-green compound like the stoichiometric PuO_2, which ranges in color from dull yellow to green and finally to black depending on the starting material that is converted to the oxide. Depending on the Pu: O ratio, PuO_{2-x} has a melting point ranging from 2430 to 2510 °C.

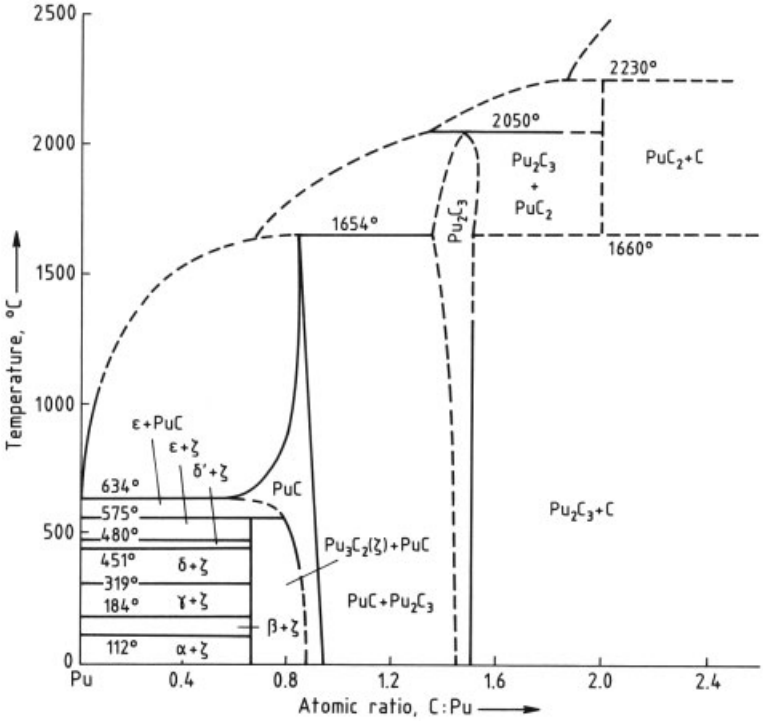

Figure 10. Phase diagram of plutonium – carbon system [4]

Table 6. Crystal structure data for plutonium oxides

Compound	Space group	Unit cell dimensions, pm	X-ray density, g/cm^3	*mp*, °C
$PuO_{1.50}$	*P*3*m*1	$a = 384.1 \pm 0.6$ $c = 595.8 \pm 0.5$	11.47	2085 ± 25
$PuO_{1.51}$	*Ia*3	$a = 1104 \pm 2.0$	10.20	
$PuO_{1.61}$	*Ia*3 or *Fm*3*m*	$a = 1095 - 1101$ or $a = 540.9$		
PuO_2	*Fm*3*m*	$a = 539.6 \pm 0.03$	11.46	2390 ± 20

Plutonium hydroxides and hydrated oxides are obtained by precipitation from aqueous solution. With excess alkali-metal hydroxides or NH_4OH, Pu^{3+} precipitates as blue or pale purple $Pu(OH)_3 \cdot xH_2O$, Pu^{4+} as olive-green $Pu(OH)_4 \cdot xH_2O$, and PuO_2^{2+} as $PuO_2(OH)_2 \cdot H_2O$. Addition of H_2O_2 to solutions of Pu^{4+} in dilute acid leads to the formation of plutonium peroxide precipitates that contain about three peroxide groups per atom, together with varying amounts of anions (e.g., SO_4^{2-}, NO_3^-).

Several plutonium sulfides exist, including PuS, Pu_3S_4, Pu_2S_3, PuS_2, and various nonstoichiometric sulfides. All are black powders except PuS, which is brown. The following oxysulfides are known: PuOS, $Pu_4O_4S_3$, Pu_2O_2S, and $Pu_2O_2S_3$. Three selenides have been identified: PuSe (*mp* 2075 ± 30 °C, 300 kPa Ar); Pu_2Se_3 (η and γ phase); and $PuSe_{2-x}$, which decomposes above 600 °C.

In the Pu – Te system the following compounds have been observed: PuTe (*mp* 1870 ± 30 °C, 300 kPa Ar); $PuTe_2$, which decomposes when heated to give $PuTe_{2-x}$; and two modifications of Pu_2Te_3. The compound $PuTe_3$ is obtained by reacting plutonium hydride with excess tellurium.

The following halides and oxyhalides of plutonium are known:

PuF_3	violet-blue
PuF_4	pale brown
$PuF_4 \cdot nH_2O$	pink-white
$PuF_4 \cdot 2.5H_2O$	pink
PuF_6	reddish brown

$PuCl_3$	emerald green
$PuCl_3 \cdot 6H_2O$	blue
$PuBr_3$	green
$PuBr_3 \cdot 6H_2O$	blue
PuI_3	bright green
PuOF	metallic
PuOCl	blue-green
PuOBr	dark green
PuOI	green
PuO_2F_2	white
$PuO_2Cl_2 \cdot 6H_2O$	greenish yellow
$PuOF_4$	chocolate brown

The trifluoride and tetrafluoride hydrates are obtained by precipitation from aqueous solution. They undergo dehydration when heated in HF atmosphere to give PuF_3 and PuF_4. Like its homologues, PuF_6 is volatile (*mp* 52 °C, *bp* 62.15 °C).

6. Toxicology

Because of its high radiotoxicity, plutonium is now handled exclusively in airtight glove boxes (Fig. 11), which are under reduced pressure to prevent leakage of aerosols or volatile plutonium compounds in case of a containment rupture. Only very small quantities (in the range of microcuries) can be handled outside glove boxes with precautions, since the limits for ingestion and inhalation of plutonium are extremely low (Table 7). These limits [5] are based on the recommendations of the International Commission on Radiological Protection. Compared to the radiological risks by incorporation, extracorporeal radiation by plutonium is relatively low. Here the same precautions are taken as for pure beta or gamma emitters.

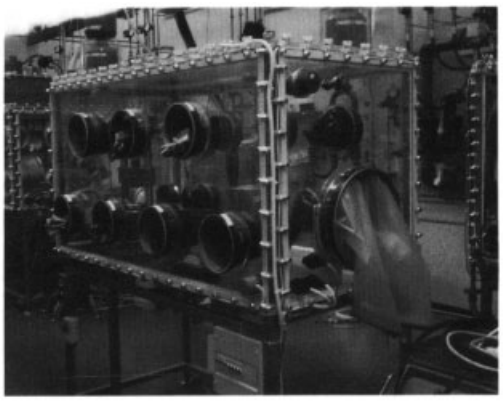

Figure 11. Glove box

Control of the Risk by Incorporation. In laboratories where plutonium is handled, personnel are protected against incorporation by

1) Continuous monitoring of air
2) Routine checks by body counters
3) Routine urine analysis

For operations (e.g., bagging in and out of plutonium in glove boxes) where the risk of contamination exists, gas masks must be worn by the operator. Rooms housing glove boxes have reduced pressure and are vented through absolute filters. When leaving controlled laboratories, operators must check their clothing, hands, and feet for possible contamination.

Release of Plutonium to the Environment. In addition to stratospheric injection from nuclear weapons testing, plutonium may be released into the troposphere during routine or accidental release from nuclear reactor operation or fuel reprocessing. Simulation tests under typical meteorological conditions have found that 99 % of 0.3-µm particles are removed from the atmosphere in 3 – 60 d, whereas larger particles deposit more quickly. For plutonium deposited in the terrestrial ecosystem the transfer to humans via the food chain is reduced by a factor of at least 10^8 because of metabolic discrimination [6].

Characteristic discrimination factors (DF) for plutonium movement in the food chain are as follows:

Soil → plant and root uptake	$DF_{plant/soil}$ $10^{-4} - 10^{-6}$
Crop plant → humans	$DF_{human/plant}$ 10^{-4}
Forage plants → grazing animals	$DF_{animal/plant}$ 10^{-4}
Plant → grazing animal → human beings	$DF = 10^{-4} \times 10^{-4} = 10^{-8}$

where, for example, $DF_{plant/soil}$ is the ratio of the plutonium concentration in the plant to the plutonium concentration in the soil.

Discrimination against plutonium also occurs at each level of the food chain for the uptake from aquatic systems. A discrimination factor of $10^{-4} - 10^{-5}$ was found between plankton and pelagic fish, depending on dietary habits. The ingestion of plutonium is different if it stems from the beginning of the food chain, such as algae.

Table 7. Annual limit of plutonium uptake ∗

Radionuclide	Form	Exposure of operators		Exposure of public	
		Annual limits of incorporation by inhalation, Ci	Derived concentration limits in air for an exposure of 2000 h/a, Ci/m^3	Annual limits of incorporation by inhalation, Ci	Annual limits of incorporation by ingestion, Ci
^{238}Pu	W	5.4×10^{-9}	2.4×10^{-12}	5.4×10^{-10}	(a) 8.1×10^{-7}
	Y	1.6×10^{-8}	8.1×10^{-12}	1.6×10^{-9}	(b) 8.1×10^{-6}
^{239}Pu	W	5.4×10^{-9}	2.2×10^{-12}	5.4×10^{-10}	(a) 5.4×10^{-7}
	Y	1.4×10^{-8}	5.4×10^{-12}	1.4×10^{-9}	(b) 5.4×10^{-6}
^{240}Pu	W	5.4×10^{-9}	2.2×10^{-12}	5.4×10^{-10}	(a) 5.4×10^{-7}
	Y	1.4×10^{-8}	5.4×10^{-12}	1.4×10^{-9}	(b) 5.4×10^{-6}
^{241}Pu	W	2.7×10^{-7}	1.1×10^{-10}	2.7×10^{-8}	(a) 2.7×10^{-5}
	Y	5.4×10^{-7}	2.7×10^{-10}	5.4×10^{-8}	(b) 2.7×10^{-4}
^{242}Pu	W	5.4×10^{-9}	2.4×10^{-12}	5.4×10^{-10}	(a) 8.1×10^{-7}
	Y	1.6×10^{-8}	5.4×10^{-12}	1.6×10^{-9}	(b) 8.1×10^{-6}

∗ W = week; Y = year; (a) = all compounds, except oxides and hydroxides; (b) = oxides and hydroxides.

7. References

1. M. E. Hasbrouck: *Plutonium Metallurgy Notebook,* BNWL-37, Battelle Northwest Laboratory, Richland, Wash. 1965.
2. J. J. Katz, G. T. Seaborg, L. R. Morss: *The Chemistry of the Actinide Elements,* 2nd ed., vol. 1, Chapman and Hall, London 1986.
3. *Gmelin,* vol. 7 B.
4. H. Matzke: *Science of Advanced LMFBR Fuels,* Elsevier, Amsterdam 1986.
5. Official Journal of the European Communities, L265, vol. 27, 5th Oct. 1984.
6. *The Environmental and Biological Behaviour of Plutonium and some other Transuranium Elements,* OECD report, Paris 1981.

Plywood → Wood
Polonium → Radionuclides
Polyacetals → Polyoxymethylenes
Polyacetylenes → Polymers, Electrically Conducting

Polyacrylamides and Poly(Acrylic Acids)

Polyacrylates and Polymethacrylates are separate keywords. General aspects of polymers, properties and testing, processing, additives, states of order, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

FREDRIC L. BUCHHOLZ, The Dow Chemical Company, Midland, Michigan 48667, United States

1. Introduction 1
2. Raw Materials 2
3. Production 2
3.1. Free-Radical Polymerization 2
3.2. Polymerization Processes 4
3.2.1. Solution Polymerization 4
3.2.2. Inverse Emulsion Polymerization . . . 5
3.2.3. Suspension Polymerization 6
3.2.4. Precipitation Polymerization 6
3.2.5. Polymer-Analogous Reactions 6
3.3. Finishing 7
4. Properties 7
5. Uses 9
5.1. Poly(Acrylic Acids) and Their Salts 10
5.2. Polyacrylamides and Hydrolyzed Polyacrylamides 10
5.3. Other Polymers 11
6. Economic Aspects 11
7. Toxicology, Occupational Health, and Environmental Aspects 12
8. References 13

1. Introduction

It is difficult to imagine an aspect of our modern world not touched by polyacrylamides and poly(acrylic acids). These hydrophilic polymers, their copolymers with other hydrophilic monomers, and their metallic salts are utilized in many applications that include mining, water treatment, sewage treatment, boiler-water additives, superabsorbent diapers, paints, detergents, industrial coatings, cosmetics, ceramics, textiles, ion-exchange resins, papermaking, oil well drilling, agriculture, industrial adhesives, cements, drug delivery, and oil production.

The utility of these polymers is directly related to their chemical structure, functionality, and molecular mass. The acid and amide functions are highly solvated by water, and are responsible for the solubility of the polymers in aqueous solutions. When cross-linked, the polymers swell in water without dissolving. The properties of polymers containing acrylic or methacrylic acid are dependent on the pH of the aqueous solution in contact with the polymer. Because the metal ion salt interacts with more water molecules than the unneutralized acid group, higher viscosity or higher swelling is observed at elevated pH, which is useful in thickeners or absorbent polymers. The proximity of the carboxyl groups to one another in poly(acrylic acid) gives rise to chelating properties with di- and trivalent metal ions which is exploited in boiler-scale control and detergent builders. The high polymerizability of the monomers allows the preparation of high molecular mass polymers which are useful as flocculants. When the molecular mass is reduced, the polymers can be used as dispersants. The reactivity of the amide and acid groups allows simple chemical modifications which change the polymer properties and amplify the number of applications.

2. Raw Materials

Some useful physical properties of the water-soluble monomers used in the production of polyacrylamides, poly(acrylic acid), and poly (methacrylic acid) are listed in Table 1. All the monomers are toxic and should be handled with great care. Contact with the skin, eyes, and mucous membranes should be avoided. Breathing of vapors should also be avoided; operations should be conducted in vessels vented to scrubbers or in laboratory fume hoods. The monomers are quite reactive; their contamination can lead to uncontrolled polymerization, and must be

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_143

avoided. Dedicated storage vessels are used for the monomer. Manufacturers' data should be consulted for the safe handling and use of the products, e.g., [1–3].

Acrylic acid is described in detail elsewhere (→ Acrylic Acid and Derivatives). It is manufactured via oxidation of propene and has the largest production volume of the hydrophilic monomers. Acrylic acid is very corrosive to the skin and eyes, and its vapors are irritating to the eyes and lungs. Personal protective equipment such as rubber gloves, goggles, impervious body coverings, and an organic vapor respirator should be used in the case of possible contact with acrylic acid. Commercial acrylic acid is normally inhibited against polymerization with 4-methoxyphenol (methyl ether hydroquinone, MEHQ). As this inhibitor requires oxygen for long-term effectiveness, air should be added to storage containers. The monomer should not be allowed to freeze because this causes the inhibitor to separate from the monomer by fractional crystallization. The resulting uninhibited acrylic acid is harzardous due to the significantly increased potential for uncontrolled polymerization.

Runaway polymerizations have occurred. A rail car of acrylic acid exploded in Teesside (UK) in 1976 as a result of too rapid thawing of frozen acid. The highly exothermic, adiabatic vinyl polymerization can become so hot that decarboxylation of the polymer occurs, with concomitant production of large volumes of carbon dioxide. A less violent, unplanned, runaway polymerization has been described [4]. Here the major product of the vigorous reaction was polyester, which is formed in a Michael-type reaction between molecules of acrylic acid.

Methacrylic acid is similar to acrylic acid in most regards and should be handled accordingly. For a detailed description of this compound, see → Methacrylic Acid and Derivatives.

Acrylamide is described elsewhere (→ Acrylic Acid and Derivatives, Chap. 3.). It has toxic effects on the nervous system. A low-dose, long-term exposure study in rats showed an increased incidence of tumors. At higher doses, laboratory animals experienced reproductive deficiencies [2]. Contact with the monomer should be strictly avoided. Solid acrylamide sublimes readily so that contact with the vapor is a likely route for exposure. Therefore, acrylamide is supplied commercially as an aqueous solution to make handling simpler and to reduce potential exposure to vapors. As with acrylic acid, oxygen is critical for inhibition of polymerization and dissolved oxygen should be present during storage because MEHQ is used as a free-radical scavenger. Acrylamide monomer is readily absorbed through the skin so that rubber gloves and impervious body clothing are required when contact with the aqueous solution is likely.

N-hydroxymethylacrylamide, *N*-methyl-acrylamide, and *N,N*-diethylacrylamide showed peripheral neuropathy in rats. All three acrylamide derivatives were more toxic than acrylamide itselft in this respect [5]. All amide monomers should be handled with great care.

3. Production

3.1. Free-Radical Polymerization

Kinetics. An understanding of the polymerization kinetics of the monomers is essential for the control of polymer production. The molecular mass and molecular mass distribution of the polymers are key determinants of their physical properties, and are influenced strongly by polymerization kinetics. Polymerization kinetics have been studied extensively for acrylic acid, its salts, and acrylamide. Less extensive data are available for the other monomers but the general characteristics are similar.

The kinetics of free-radical polymerization of acrylic acid, methacrylic acid, acrylamide, and methacrylamide have been investigated in water, dimethylsulfoxide, tetrahydrofuran, and formamide at 30 – 60 °C with azobis(isobutyronitrile) as initiator [6]. The polymerization rate is proportional to the monomer concentration and to the square root of the initiator concentration. The acrylics react faster than the methacrylics under identical conditions. The polymerization rate depends on the nature of the solvent; values of the kinetic constants for each monomer in each solvent were determined [6].

Table 1. Properties of monomers used in the production of polyacrylamides, poly(acrylic acids), and their copolymers

Name	CAS registry number	Molecular formula	M_r	*mp*, °C	*bp*, °C	Density, g/cm^3	Flash point, °C	Heat of polymerization, kJ/mol	Oral (rat) LD_{50}, mg/kg
Acrylamide	[*79-06-1*]	C_3H_5NO	71.08	84.5	136 (3.33 kPa)	1.122 (30 °C)		81.5	124
2-Acrylamido-2-methylpropanesulfonic acid	[*15214-89-8*]	$C_7H_{13}NSO_4$	207.21	185			160	92.0	1410
Acrylic acid	[*79-10-7*]	$C_3H_4O_2$	72.06	13.5	141 (101.3 kPa)	1.040 (30 °C)	50	77.5	340
N,N-Dimethylacrylamide	[*2680-03-7*]	C_5H_9NO	99.15	−40	81 (2.80 kPa)	0.926 (25 °C)	70		700
N-Isopropylacrylamide	[*2210-25-5*]	$C_6H_{11}NO$	113.16	64					
Methacrylic acid	[*79-41-4*]	$C_4H_6O_2$	86.10	16	163	1.015	76	56.3	1600 (mouse)
Methacrylamide	[*79-39-0*]	C_4H_7NO	85.12	110				56.0	459
N,N′-Methylenebisacrylamide	[*110-26-9*]	$C_7H_{10}N_2O_2$	154.17	300		1.24 (30 °C)		82.7	390

The polymerization kinetics for the acid monomers become complicated as the pH of the system changes. At constant overall monomer concentrations, the polymerization rate initially falls as the pH is increased from 2 to 7, then rises slowly from pH 7 to 12 [7], [8]. This has been attributed to a change in the propagation rate constant as the degree of ionization increases [9]. At pH 7 to 12, rates increase with an increase in the ionic strength of the medium [10]. This is accomplished either by adding simple monovalent salts or by increasing the ionized monomer concentration.

Initiators play a complex role in polymerization kinetics, and initiator fragments can be important in the subsequent use of the polymer: for example, solution properties can deteriorate with time, depending on impurities [11]. Thermal initiators (e.g., azo compounds and peroxides) produce free radicals in a homolytic scission reaction. The reaction, producing a pair of radicals, is characterized by the temperature at which the specific initiator is half consumed in 10 h. With those initiators the polymerization rate generally depends on the monomer concentration and the square root of the initiator concentration [12], [13]. The monomer may react with the initiator prior to the reaction of the free radical with monomer, and induce the decomposition of the initiator. In such cases (e.g., acrylic acid and sodium persulfate), the polymerization rate depends on the 1.5 power of monomer concentration and the square root of the initiator concentration [9], [10], [14]. Redox initiators produce free radicals in a reaction between a reducing agent and an oxidizing agent. Examples include the persulfate – thiosulfate reaction [15] and the manganic ion (Mn^{3+}) – isobutyric acid reaction [16]. For redox initiators that do not involve metal ions, the polymerization rate usually depends on the first power of the monomer concentration. The rate also depends on the concentration of oxidant and reductant, the stoichiometry of their reaction, and the extent to which the initiators (usually the reductant) participate in radical termination reactions [15], [17–19]. When metal ions are involved in the redox reaction, the kinetics are more complex because the metal ion also participates in the termination step of the polymerization [16], [20], [21].

Copolymerization. See also → Polymerization Processes, Chap. 2.3. The monomers listed in Table 1 are frequently copolymerized with one another and with other monomers. The relative reactivities of the monomers affect the microstructure and thus the physical properties of the polymer. When a comonomer is more reactive than its partner, it tends to be used up earlier in the reaction. The final product then has a distribution of composition as well as a distribution of molecular masses.

Reactivity ratios are derived from measurements of polymer composition as a function

Table 2. Reactivity ratio as a function of pH [23]

pH	Acrylic acid r_1	Acrylamide r_2	Methacrylic acid r_1	Acrylamide r_2
4	0.32±0.046	0.57±0.067	2.80±0.44	0.20±0.069
6	0.33±0.20	0.85±0.62	0.19±0.08	0.55±0.078
8	0.63±0.004	0.12±0.004	0.32±0.004	0.53±0.003

of conversion. Data from the entire conversion range of the copolymerization of acrylamide (AAm) and acrylic acid (AA) were utilized to calculate the following reactivity ratios: r_1(AA)=1.45±0.33, r_2(AAm)=0.57±0.04 [22]. Reactivity ratios for acrylic or methacrylic acid copolymerized with acrylamide have also been determined as a function of pH [23]. Selected data are given in Table 2. Reactivity ratios for these and many other potential comonomers have been tabulated [24].

3.2. Polymerization Processes

The techniques described in this section are described in detail elsewhere (→ Polymerization Processes).

3.2.1. Solution Polymerization

Batch Polymerization. The batch polymerization of the water-soluble monomers in solution is straightforward. The monomers are dissolved in water (or another solvent for both the monomer and polymer) at a desired concentration (usually 10 – 70 %). Although acrylamide and acrylic acid have been polymerized in formamide or dimethyl sulfoxide solution [6], water is used in industrial processes. The monomer solution is deoxygenated by bubbling an inert gas through the solution or by a series of evacuations and repressurization with an inert gas. The desired free-radical initiator is then added and the temperature, usually 40 – 80 °C depending on the specific initiator used, is brought to the appropriate point for polymerization to begin.

The initial monomer concentration is limited since polymerization of the undiluted monomers is extremely dangerous due to the high heat of polymerization and rapid polymerization kinetics [25]. The concentration is also limited by the capacity of the equipment for removing the heat of polymerization and by the molecular mass of the desired polymer. If no heat can be removed from the reactor, the adiabatic temperature rise can be calculated for a given monomer concentration [26]. A suitable monomer concentration can then be calculated, based on the desired initial and final temperatures [27]. The viscosity of the reaction mass is also important with regard to the equipment available for pumping the product. Higher molecular mass products generally dictate a lower monomer concentration because of their high viscosities, whereas low molecular mass products allow higher monomer concentrations.

Water is the most common solvent for solution polymerization. An undesirable, insoluble, "popcorn" polymer (PCP) of acrylic acid can be formed in polymerization mixtures containing 0.53 – 0.88 mol fraction water [28]. However, acrylic acid can be polymerized within this range without evidence of such insoluble products. For example, a water-absorbing resin has been synthesized by polymerization of a high concentration of acrylic acid or its salts and a cross-linker in aqueous solution [29]. The concentration of the monomers was ≥ 65 wt %, corresponding to a mole fraction of water of 0.68. This resin was intentionally cross-linked to yield a highly swellable polymer (not PCP).

Low molecular mass, water-soluble polymers are often desired, and several processes have been developed to produce them industrially. In one such process, sodium hydrogensulfite is added continuously to an aqueous monomer solution in a stirred tank while introducing fine air bubbles; polymerization occurs below 80 °C [30]. The sulfite acts as a chain-transfer agent to lower the molecular mass. In another industrial process, poly(acrylic acid) of molecular mass 1000 – 20 000 is prepared at 55 wt % concentration by polymerization with hydrogen peroxide and transition-metal activators such as copper, iron, or manganese [31].

Continuous Polymerization. Polymers can often be produced more efficiently by using a continuous polymerizer. An aqueous solution of monomers, initiators, and other additives is fed into one end of the polymerizer and the polymeric product removed from the other. Copolymers of acrylic acid and acrylamide of controlled molecular mass were produced in a packed-tube reactor operating at elevated pressure [32]. Monomer concentration was 5 – 50 wt %. The solution contained free-radical initiators at concentrations of 0.5 – 30 wt %; the reaction was adiabatic, beginning at 19 – 121 °C, depending on the desired molecular mass of the polymer. In another process, 30 – 40 wt % aqueous monomer solution is pumped onto a continuous, moving belt. Angled rollers at the edges of the belt cause a trough to form over a section of the belt. The trough holds the liquid monomers. As polymerization proceeds, a gel is formed from the reagents. The belt section flattens after it passes the final set of rollers, and the gel is released from the belt and is finally discharged from the end of the conveyor. Superabsorbent polyacrylates are prepared industrially in this manner [33].

3.2.2. Inverse Emulsion Polymerization

In inverse emulsion polymerization, tiny droplets of aqueous monomer solution are dispersed in a continuous organic phase (usually aromatic or aliphatic hydrocarbons), prior to polymerization. The droplets are stabilized by surfactants. Free-radical polymerization of the monomers is conducted in a manner similar to that described for solution polymerization (Section 3.2.1) [34], [35].

An obvious reason for using this technique instead of solution polymerization is to simplify the handling of the polymer product. The low-viscosity dispersion of monomer and polymer droplets is easy to agitate and pump, and allows more efficient heat removal during polymerization. In addition, the particle size of the product is more easily controlled because it depends on the nature and amount of the emulsifier used to form the droplets prior to polymerization. Droplet size is typically 0.1 – 10 μm. However, the process is more complicated; several more ingredients need to be controlled and recycled, especially the hydrocarbon phase (unless the product is sold in emulsion form). It is usually less efficient on a volumetric basis, due to the presence of the continuous organic phase. For example, such processes typically employ 30 wt % of aqueous monomer phase that itself has a monomer concentration of 30 – 60 wt % [35]. The overall content of monomer in the emulsion is therefore 12 – 18 wt %, compared to 25 – 50 wt % typical of solution processes. Experimental, high-concentration emulsion processes have been developed. Acrylamide has been polymerized in an inverse emulsion of decane, consisting of 94 wt % aqueous (internal) phase [36]. The emulsions were most stable when the monomer concentration in the aqueous phase was 20 wt %, although limited stability was observed from 20 – 40 wt % aqueous phase. Therefore, overall monomer concentration in the stable emulsion was 19 wt %, and as high as 38 wt % in emulsions of limited stability. A significantly better result was obtained when a bimodal droplet size distribution was formed [37]. In this case, a total polymer content of 35.5 wt % and a relatively low emulsion viscosity were reported. Copolymers of acrylamide and methacrylic acid have also been produced by inverse emulsion polymerization [38].

3.2.3. Suspension Polymerization

In suspension polymerization, small droplets of aqueous monomer solution are dispersed in a hydrocarbon phase, generally with the aid of other polymers but also with inorganic materials (e.g., derivatized clays [39]). Free-radical polymerization of the monomers is then conducted in a manner similar to that described for emulsion systems (see Section 3.2.2). The droplet size is controlled by the amount of dispersant used; more dispersant generally yields smaller particle size. Larger polymer particles can be obtained than with emulsion processes. Typically, 15 – 500 μm spherical beads are formed. Cellulose esters or cellulose ethers are used as protective colloids in the suspension polymerization of sodium acrylate, acrylamide, 2-acrylamido-2-methylpropanesulfonic acid, or sodium styrene sulfonate [40]. Oil-soluble, cationic copolymers of alkyl-substituted styrene and quaternary dialkylaminoalkyl acrylate salts have also been used [41]. A mixed suspending agent for

water-soluble monomers consisted of hydrophobic fumed silica and a copolymer of acrylic acid and lauryl methacrylate and yielded a controlled particle size distribution [42].

3.2.4. Precipitation Polymerization

In precipitation polymerization,the monomer is polymerized in a medium which is a solvent for the monomer but a nonsolvent for the polymer. As the polymer forms, it precipitates from the monomer phase. With the proper choice of medium, polymer isolation and purification is simplified. For example, cross-linked polymers can be produced by copolymerizing acrylic or methacrylic acid with small amounts of an allyl phosphate or phosphite with free-radical initiators in a solvent such as benzene, hexane, acetone, or chloroethane at ca. 20 – 90 °C [43]. This technology is the basis for the process used industrially to produce Carbopol poly(acrylic acid) resins. The precipitating polymerization of poly(acrylic acid) has been accomplished in methyl ethyl ketone [44].

3.2.5. Polymer-Analogous Reactions

The term polymer-analogous reactions denotes reactions that are used to modify the functional groups of polymers. Such reactions are complex from a practical perspective because of the high viscosities of polymer solutions and the changing solubilities of the polymers as they are converted to a different functionality. Polymers of acrylate esters can be hydrolyzed to the corresponding acid polymers and their salts. Polyacrylate esters can also be reacted with amines to yield polyacrylamide or polymethacrylamide derivatives. Polyacrylonitrile can be saponified to a copolymer of acrylic acid salt and acrylamide.

Poly(acrylic acid) undergoes reactions typical of other organic carboxylic acids. It has been reacted with diazomethane to yield poly(methyl acrylate) and with thionyl chloride to yield the acid chloride, which has increased reactivity in acylation reactions such as amidation. Poly(acrylic acid) can also be dehydrated above 200 °C to yield a polyanhydride. The polyacids are generally reacted with neutralizing agents, e.g., alkali-metal hydroxides, carbonates, or amines, to yield the salt.

Polyacrylamides can also be chemically derivatized [45]. Polyacrylamide reacts with formaldehyde in aqueous solution to yield a partially hydroxymethylated polymer. It is also readily hydrolyzed with aqueous sodium hydroxide or sodium carbonate solution to give a polymer with a maximum of ca. 86 % acrylic acid moieties. Amide groups are converted to amine groups via the Hofmann reaction; the amide is reacted with sodium hydroxide and hypochlorite, yielding the amine. Amine groups can also be introduced into polyacrylamide by means of the Mannich reaction. Polyacrylamide is reacted with formaldehyde and an amine to produce a *N*-aminomethylated polymer. Intramolecular imidization occurs when polyacrylamide is heated with a strong acid catalyst.

Due to the processing difficulties encountered in conducting polymer-analogous reactions, such reactions are not typically used to convert a polymer to one with different functional groups. Nevertheless, the technique is used widely for cross-linking polyacrylamides and poly(acrylic acids). Polyacrylamide is cross-linked by reaction of its amido groups with glyoxal (→ Glyoxal). Alternatively, a reactive monomer can be synthesized from acrylamide and glyoxylic acid, copolymerized with acrylamide, and used as a reactive site for subsequent cross-linking (→ Glyoxylic Acid). Poly(acrylic acid) is cross-linked by reaction of its carboxylic acid groups with glycols; this technique is used industrially to cross-link the surface of superabsorbent polymers [46].

3.3. Finishing

The finishing operations used depend on the polymerization process employed (Section 3.2). In *solution processes*, the water is first evaporated using hot, rotating drum dryers or continuousbelt, hot-air ovens. The dry flakes or lumps are then milled to the desired particle size and packaged in bags ranging in size from 22 to 1000 kg.

In *inverse emulsion processes*, water is usually boiled from the emulsion and the dried emulsion treated with surfactants that promote

dissolution of the product in water. The polymer can be coagulated and separated from the organic phase by treating the emulsion with another organic solvent such as acetone. However, typically the dried emulsion is sold in liquid form and packaged in drums or tank trucks.

In *suspension processes*, the polymer is dried by boiling off the water. The beads are filtered from the organic phase using a centrifugal or other industrial filter. The residual organic phase is then evaporated from the beads using an explosion-proof dryer. The product is packaged in bags.

The most important aspect of the finishing of these polymers is the reduction in the amount of residual monomers. Many techniques have been developed. Conversion of monomer to polymer can be improved by efficient reaction with initiators [47]. The monomer can also be reacted with an additive to yield a presumably less toxic, saturated derivative. Sodium sulfite [48], gaseous sulfur dioxide [49], and amines or other active hydrogen compounds such as malonate esters [50] have been used for this purpose. Simple distillation [51] and solvent extraction [52] have also been employed.

4. Properties

The properties of the polymers as solids and in solution (particularly aqueous solution) determine their applications. Their hydrophilic nature and molecular mass are key determinants of the usefulness of the polymers for specific applications. The molecular mass and the molecular mass distribution can be measured by light scattering, osmometry, and ultracentrifugation. Intrinsic viscosity is the most useful and widely used method of correlation between molecular mass and viscosity: appropriate equations are given below for individual polymers. Molecular masses are also correlated to the radius of gyration of the molecules in solution by means of gel permeation chromatography (size-exclusion chromatography) or hydrodynamic chromatography. The very high molecular mass polymers are susceptible to degradation by high shear [53], oxidation [54], and conformational changes [55], which can be induced by impurities introduced as polymerization modifiers. The above methods are useful for measuring the extent of such changes.

Analysis of the polymers and determination of their content of comonomer residues are of practical importance. These can be determined by infrared spectroscopy [56] and nuclear magnetic resonance (NMR) spectrometry [57]. The tacticity of the polymers is also measured by NMR spectrometry [58]. Thermal transitions and polymer chemistry are investigated with differential scanning calorimetry, thermogravimetric analysis (TGA), and TGA – mass spectrometry.

Poly(acrylic acid) [*9003-01-4*] is soluble in many polar and usually hydrogen-bonded solvents such as water, methanol, ethanol, dioxane, dimethylformamide, 2-propanol, formamide, ethylene glycol, and acetic acid. The value of the glass transition temperature (T_g) is still in dispute because it is difficult to obtain a solventfree sample. T_g has been reported at 106 °C [59], at 118 °C [60], and at 126 °C [61]. Mark – Houwink – Sakurada relationships for determining the weight-average molecular mass M_W from the intrinsic viscosity $[\eta]$ have been determined in many solvents but the following two are widely used:

$[\eta]=0.085\,M_W^{0.5}$ mL/g dioxane, 30 °C [62]

$[\eta]=0.1062\,M_W^{0.5}$ mL/g 0.2 mol/L HCl, 14 °C [63]

The exponents of 0.5 indicate theta conditions for the polymer. Under theta conditions, a given polymer solvent pair behaves as if it were an ideal solution. This is because the attraction between polymer chain segments just equals the loss of entropy brought about by interpenetration of polymer molecules. If conditions for solvation are made any poorer than those at the theta condition, the polymer precipitates. For poly(acrylic acid), lowering the temperature from the upper critical solution temperature, UCST, will cause phase separation. Poly(acrylic acid) is stable in acid or base solution. The solid polymer slowly liberates water upon heating above 100 °C, forming the anhydride. The polymer decomposes slowly above 250 °C, liberating carbon dioxide.

The simple alkali-metal salts have been characterized extensively, especially poly(sodium acrylate) [*9003-04-7*]. Poly(potassium acrylate) [*25608-12-2*] and poly(ammonium acrylate) [*9003-03-6*] are also known. These poly-

mers are soluble in water and in aqueous salt solutions. Small amounts of organic solvents dissolved in water are sometimes tolerated. The fully neutralized sodium salt has a glass transition temperature at about 250 °C [60]. Obtaining an accurate value is difficult because of the hygroscopic nature of the polymer. Mark – Houwink – Sakurada relationships have been determined in many aqueous solutions; for example

$[\eta]=1.24\times10^{-3}\,M_W^{0.5}$ mL/g 1.5 mol/L NaBr, 15 °C[64]

$[\eta]=0.146\times10^{-3}\,M_V^{0.80}$ mL/g 0.1 mol/L NaCl, 25 °C [65]

Poly(acrylic acid) salts are more stable to heat than the acid polymer. Anhydride formation is not observed above 50 mol % neutralization of acid groups and up to 250 °C, where carbon dioxide evolution is observed.

Polyacrylamide [*9003-05-8*] is soluble in polar, hydrogen-bonded solvents such as water, aqueous salt solutions, ethylene glycol, formamide, ethanolamine, dimethyl sulfoxide, and morpholine. The glass transition temperature is again difficult to determine since it is difficult to prepare a solvent-free sample; T_g is reported as 170 °C [66] and 188 °C [61]. The density of the polymer is 1.302 g/cm^3 [67]. The diffusion coefficient D and the sedimentation coefficient S_0 have been determined as a function of molecular mass [68]:

$D=1.24\times10^{-4}\,M_W^{-0.53}$ cm^2/s

$S_0=1.07\times10^{-15}\,M_W^{0.48}$ svedberg

Mark – Houwink – Sakurada relationships have been determined in many solvents [69]:

$[\eta]=9.33\times10^{-3}\,M_W^{0.75}$ mL/g water

$[\eta]=7.19\times10^{-3}\,M_W^{0.77}$ mL/g 0.5 mol/L NaCl

$[\eta]=1.36\times10^{-1}\,M_W^{0.51}$ mL/g ethylene glycol

A theta condition has been established in a 2:3 ratio of methanol:water at 20 °C [70], and has been estimated at −38 °C in aqueous solution.

Solutions of polyacrylamide in strong acid gradually become turbid due to imide formation from amide groups. The amide groups hydrolyze to ammonia and the carboxylic acid above pH 9. The solid polymer is stable below 210 °C, but the amide groups decompose to ammonia above 210 °C.

Poly(methacrylic acid) [*25087-26-7*] is soluble in ethanol, methanol, water, dimethylformamide, dioxane, and 2-ethoxyethanol. The polymer and its sodium salt [*54193-36-1*] have been characterized because of curiosity about the conformational change which occurs upon neutralization of the acid polymer to about 30 mol %. The glass transition temperature occurs at 185 °C [71]. Mark – Houwink – Sakurada relationships are:

$[\eta]=22.4\times10^{-3}\,M_W^{0.679}$ mL/g methanol, 25 °C[72]

$[\eta]=6.6\times10^{-3}\,M_W^{0.5}$ mL/g 0.002 mol/L HCl, 30 °C [73]

The above theta condition at 30 °C in dilute hydrochloric acid is questionable in light of experiments that show phase separation in poly(methacrylic acid) solutions when the temperature is raised to 56 °C [63].

The stability of poly(methacrylic acid) and its salts to acids and bases is similar to that of poly-(acrylic acid). The solid polymer evolves water above 180 °C, forming an anhydride.

Polymethacrylamide [*25014-12-4*] is soluble in methanol, water, ethylene, glycol, and mixtures of acetone and water. A Mark – Houwink – Sakurada relationship has been established as

$[\eta]=2.1\times10^{-3}\,M_W^{0.76}$ mL/g water[74]

Stability of polymethacrylamide, either in solution or as the dry solid, has not been reported.

The major *copolymers* are those of acrylic acid and acrylamide, prepared either by partial hydrolysis of polyacrylamide [75] (hydrolyzed, sodium salt [*25987-30-8*] and hydrolyzed, potassium salt [*31212-13-2*]) or by copolymerization of the respective monomers [61]. The properties of the copolymers depend on the mole fractions of the two monomer residues [76] and on the distribution of molecular masses of the sample [77]. The intrinsic viscosity (and the second virial coefficient) of the copolymers shows a maximum at 33 mol % acrylic acid residues [76]. Flocculation rate increases with molecular mass; maximum activity was observed with a 70 mol % acrylate copolymer [78]. Solution characteristics of heterodisperse copolymers were determined at various compositions in salt solutions by viscometry, light scattering, and gel permeation chromatography. Mark – Houwink – Sakurada relationships have been determined over the entire composition range with corrections for heterodispersity [77]. Copolymers of methacrylic acid and acrylamide have been similarly characterized as a function of pH [79]. The sudden

increase in intrinsic viscosity observed at pH 4 – 5.5 has been attributed to a sudden unfolding of the chains when the electrostatic repulsion of the ionic groups is just sufficient to overcome the hydrophobic bonding of the α-methyl groups.

Poly(N-isopropylacrylamide) [*25189-55-3*] is of interest because of a lower critical solution temperature (LCST) of just 31 °C in aqueous solution [80]. Dissolved poly(*N*-isopropylacrylamide) precipitates and its cross-linked gels deswell when the temperature is raised above the LCST. It is soluble in water, methanol, and tetrahydrofuran. The glass transition temperature has been measured at 130 °C [81]. A Mark – Houwink – Sakurada relationship has been determined:

$[\eta]=0.112\,M_W^{0.51}$ mL/g water, 20 °C [82]

Poly(N,N-dimethylacrylamide) [*26793-34-0*] is soluble in water, methanol, 2-methoxyethanol, methyl ethyl ketone, diethyl ether, toluene, chloroform, and dioxane. It has a low glass transition temperature (89 °C) [83]. The polymer has been extensively characterized only in methanol and water. The Mark – Houwink – Sakurada relationships are [84]:

$[\eta]=17.5\times10^{-3}\,M_W^{0.68}$ mL/g methanol, 25 °C

$[\eta]=23.2\times10^{-3}\,M_W^{0.81}$ mL/g water, 25 °C

Poly(2-acrylamido-2-methylpropanesulfonic acid) [*27119-07-9*] is soluble in water, aqueous salt solutions, and dimethylformamide. The sodium salt is soluble in aqueous solution. Mark – Houwink – Sakurada relationships for the homopolymer and its sodium salt are [85]:

$[\eta]=2.11\times10^{-3}\,M_W^{0.8}$ mL/g H^+ form, 5 mol/L NaCl

$[\eta]=1.95\times10^{-3}\,M_W^{0.85}$ mL/g Na^+ salt in 0.5 mol/L NaCl

Solution properties of carboxylated and sulfonated copolymers have been compared [86].

5. Uses

The molecular mass of a polymer is a key factor in determining its usefulness for a particular application. Polymers of molecular mass less than 20 000 are used as sequestrants and scale inhibitors. Polymers of molecular mass between 20 000 and 80 000 are useful as pigment dispersing agents. Polymers with molecular mass between 10^5 and 10^6 are used as textile finishing agents and retention aids for papermaking. Molecular masses exceeding 10^6 are necessary for flocculants and thickening agents [87]. High molecular mass, cross-linked polymers are used as fluid absorbents.

The ionic character of the polymer is another important factor in determining its applications. Choices range from nonionic polyacrylamides to partially hydrolyzed polyacrylamides or copolymers of acrylic acid and acrylamide and finally to the salts of poly(acrylic acid). The appropriate choice of ionic content depends on the pH, ionic strength, surface charge of any substrate with which the polymer interacts, and other conditions specific to the application.

The polymers are supplied as solutions, emulsions, and powders. Low molecular mass polymers are typically supplied as viscous, aqueous solutions containing 20 – 50 wt % polymer. Above molecular mass 2×10^5, the high viscosity of such solutions makes them too difficult to meter into processes and mix well; they are supplied as powders packaged in 22 kg bags. The powders are difficult to dissolve in water because the particles hydrate exceedingly fast, forming gelatinous clumps. Special mechanical dispersers are available from the suppliers to aid dissolution [88]. High molecular mass polymers are also supplied as inverse emulsions; micrometer-sized polymer particles are dispersed in a liquid hydrocarbon with the aid of surfactants. These emulsions are of relatively low viscosity and readily pumped and metered. When the emulsion is added to water the emulsion inverts and the small polymer particles quickly dissolve. Emulsion products are supplied in drums or tank trucks. Cross-linked polymers such as superabsorbents are supplied as granules in 1-t sacks that empty from the bottom to make emptying the polymer into hoppers easier.

5.1. Poly(Acrylic Acids) and Their Salts

Water-soluble poly(acrylic acid) and its neutralized salts having a molecular mass of 2000 – 5000 are used as scale inhibitors, sludge (mud, silt, clay) dispersants in cooling water systems, and as dispersants for preparing $CaCO_3$ or clay slurries as fillers, pigments, or paper-coating materials [89].

Homo- or copolymers of acrylic acid and methacrylic acid and mixtures with up to 10 wt % alkyl acrylate comonomers prevent soil redeposition in liquid detergent formulations [90]. The polymer, neutralized or not, has a molecular mass of $10^5 - 10^6$. Relatively small amounts of polymer prevent soil redeposition, improve cleaning performance and viscosity control.

Copolymers having small amounts of pendant hydrophobic groups are also useful as hydraulic fracturing fluids for oil well drilling [91]. The fluids are readily formulated to give an initial viscosity which is retained for long periods at high temperatures and under high shear rates; this is due to hydrophobic bonding between the pendant groups. The polymer molecular mass can be below 10^6 and yet provide high viscosity in solution because the associating hydrophobic groups create a high pseudomolecular mass.

Cross-linked poly(sodium acrylate) is used as an absorbent in baby diapers [92–94]. Other applications include adult incontinent products [95], feminine hygiene products [96], absorbents in transmission cables [97], and agricultural mulches [98]. Cross-linked polymers of acrylic or methacrylic acid, neutralized to greater than 50 mol %, may also be used to provide high water retention capacity and high viscosity in plaster [99]. A cross-linked, water-absorbent acrylic acid polymer can be used in formulations for sustained-release tablets for oral administration [100]. A blend of water-swellable particles of anionic polymer of acrylic acid or methacrylic acid and their soluble salts and cationic polymers of amino acrylate and aminoalkyl acrylamide is used in prepasted wall-covering adhesive [101] to reduce water absorption. Cross-linked poly(acrylic acid) is also used as a weakly acidic, cation-exchange resin (→ Ion Exchangers, Chap. 2.2.1.).

5.2. Polyacrylamides and Hydrolyzed Polyacrylamides

Polyacrylamides are typically not available with molecular masses below 10^5 because of the very small termination constant affecting the polymerization of acrylamide. Slightly hydrolyzed polyacrylamide with molecular mass $(1-5)\times 10^5$ is useful as a strengthening agent in finished paper. Polymers with molecular masses up to 10^6 add wet strength in paper production. Incorporation of about 5 wt % hydrophobic vinyl linkages into the polymer increases the adsorption of the polymer by cellulose fibers [102].

Polyacrylamides with molecular masses of $(1-2)\times 10^6$ are used as retention aids in papermaking. The polymer binds the smaller fibers and any filler particles to the larger fibers.

Polyacrylamides and hydrolyzed polyacrylamides having molecular masses of $(2-20)\times 10^6$ are used in applications involving flocculation or to increase the viscosity of aqueous systems. Flocculation is the major use for polyacrylamides. Flocculant applications include clarification of water used in industrial processes (e.g., as pretreatment of boiler water), in treatment of effluent from paper mills, and in sewage treatment. In sewage treatment, polyacrylamides increase the removal of solids in the primary settling step, improve clarification of the discharged water, aid in sludge thickening, improve clarity in the supernatant from digestors, and increase water removal during sludge dewatering. The effectiveness of polyacrylamides in flocculation depends on molecular mass and on the mole fraction of ionic groups. Copolymers of acrylamide and acrylic acid of different compositions were prepared and their flocculation activity toward kaolin clay suspensions studied [78]. Flocculation rate increased with molecular mass, maximum flocculation activity being observed with 70 mol % acrylate copolymer. Applications for increased viscosity include drilling fluids used in oil production and flooding fluids used in secondary oil recovery.

Cross-linked polyacrylamides are produced by cross-linking during polymerization or by reacting the amide groups of the soluble polymer with another reagent, generally at the site of use. The permeability of subterranean formations in oil recovery is reduced by exposing them to a solution of high molecular mass, hydrolyzed polyacrylamide and a cross-linking solution containing zirconium ions [103]. The method is used to reduce permeability, improve control of fluid injection patterns in secondary or tertiary oil recovery, and reduce the quantity of water produced from the formation. Aqueous solutions of partially hydrolyzed polyacrylamide are gelled with alkali-metal dichromate and sodium hyposulfite to yield a plugging compo-

sition for underground strata [104]. The viscosity of the initial mixture builds up slowly with time so that it can be pumped into place before high viscosity develops. Non-cross-linked, hydrolyzed polyacrylamides are used as mobility control agents in secondary oil recovery [88].

A particulate, nonionic polyacrylamide, that is cross-linked during polymerization, is useful as a plant growth medium [105]. Inclusion of this swellable polymeric material improves water retention. An intimate mixture of finely divided polyacrylamide and polyacrylate is used as a hygroscopic seed coating [106]. The mixture attracts moisture to the seeds, improves germination rates, and accelerates seedling emergence and root development. Dust is controlled on roads, in mines, and particularly on surfaces of pulverized coal and mineral piles within open transit cars or trucks by applying oil-continuous emulsions of highly branched, cross-linked, water-swellable polymers of acrylamide and acrylic acid [107]. The polymers are in the form of microgelatinous particles (≤ 1 mm diameter) which swell dramatically in water and then bind dust particles.

5.3. Other Polymers

The polymers described in this section are generally experimental products, some of which are available in developmental quantities.

Lightly cross-linked poly(*N*-isopropylacrylamide) hydrogels have a phase transition at 31 °C. This property has been suggested to be useful in separations such as concentration of soy protein from aqueous extracts [108] or in controlled release of drugs [109].

The good balance of hydrophobic and hydrophilic groups in poly(*N,N*-dimethylacrylamide) and copolymers of *N,N*-dimethylacrylamide with other water-soluble monomers makes these materials soluble in a wide range of solvents. This suggests utility as thickeners in formulations with high concentrations of organic chemicals. The homopolymer is mutually soluble with poly(vinyl acetate), poly(methyl methacrylate), and polystyrene. Appropriate copolymers may serve as polymer compatibilizers [110]. The homopolymers of 2-acrylamido-2-methylpropanesulfonic acid and copolymers with acrylamide [111] and the corresponding alkali-metal salts have been used in applications where the chelation tendencies of carboxylate copolymers toward di- and trivalent ions are detrimental, and for thickening aqueous solutions of strong acids [112] (e.g., in the acidification of oil wells [113]).

6. Economic Aspects

The homo- and copolymers of acrylic acid, its salts, and acrylamide make up the majority of the production volume of these polymers. Manufacturers and trade names are listed in Table 3. In 1988, worldwide production of poly(acrylic acid) and polyacrylamide amounted to 230×10^6 kg and 144×10^6 kg, respectively.

The market for poly(acrylic acid) was divided as follows: 52.9 % as superabsorbent polymers, 18.4 % as detergent additives, 17.1 % as industrial dispersants, 7.8 % as water treatment chemicals (primarily sequestrants and scale inhibitors), and 3.8 % as coatings components and drilling muds. The polyacrylamide market was divided as follows: 47 % for water treatment (primarily flocculants), 33 % for additives in pulp and paper, 13 % for mineral processing, 6 % for oil recovery, and 1 % other. Geographically, 48.2 % of the poly(acrylic acid) was used in the United States and Canada, 31.3 % in Western Europe and 20.5 % in Japan. For polyacrylamide the breakdown was 40 % United States and Canada, 28 % Europe, and 32 % Japan.

Annual growth for all segments of the polyacrylamide market, except for enhanced oil recovery, is expected to be 4 – 5 % in the early 1990s. The poly(acrylic acid) market is expected to grow at about 6.5 % annually overall in the early 1990s, with strong increases in the superabsorbent and detergent additive segments [114].

7. Toxicology, Occupational Health, and Environmental Aspects

Polyacrylamides, hydrolyzed polyacrylamides, poly(acrylic acids) and their salts are relatively nontoxic materials. Care should be taken in the finishing of the polymers to strictly limit the content of any unreacted monomers because they

are toxic [115]. Handling of the monomers is described in Chapter 2.

Table 3. Manufacturers and trade names of poly(acrylic acids) and polyacrylamides

Company	Trade name	Type ∗
Poly(acrylic acids)		
Allied Colloids	Salsorb	a
	Antiprex	b
Alco Chemical	Alcosperse	b
Arakawa	Arasorb	a
BFGoodrich	Carbopol	d
Chemische Fabrik Stockhausen	Favor	a
Colloids, Inc.	Colloid	b
Dow Chemical	Drytech	a
Nippon Shokubai K.K.	Aqualic	a
Sanyo Chemical Industries	Sanwet	a
Seitetsu K.K.K.	Aquakeep	a
Rohm and Haas	Acrysol	b
Polyacrylamides		
Alco Chemical	Aquatreat	b
Allied Colloids	Magnafloc	c
	Percol	c,d
	Versicol	c,d
	Alcoflood	d
Callaway Chemical	Jayfloc	c
Chemische Fabrik Stockhausen	Praestol	b
CPS Chemical	Agesperse	b
Cyanamid	Cyanamer	b
	Cyanatrol	d
	Superfloc	c
Nalco Chemical	Nalcotrol	d
Röhm	Rohafloc	c
Rohm and Haas	Primafloc	c
Sanyo Chemical	Sanfloc	c
SNF Floerger	Floerger	c

∗ a = superabsorbent; b = dispersant, scale inhibitor; c = flocculant; d = thickener.

Poly(acrylic acid) showed low toxicity in acute screening tests indicated by an LD_{50} of 5000 mg/kg (rat, oral) and 3000 mg/kg (rabbit, dermal) [87], [116]. Screening tests for toxicity to aquatic wildlife showed that the LC_{50} was greater than 1000 ppm for daphnia, sunfish, and trout [117]. Neutralized products of high alkalinity can be irritating to the eyes and skin.

The toxicity of polyacrylamide and hydrolyzed (anionic) polyacrylamide containing only 500 – 800 ppm residual monomers has been evaluated in rats, dogs, and fish [118]. Low toxicity is indicated by the single oral dose LD_{50} which is greater than 4000 mg/kg in rats. The results from two-year feeding studies in rats showed no effect at 1 % dietary concentration and a slight retardation in growth at the 5 % dietary level. Dogs were unaffected by feeding diets containing up to 6 % of these polymers. Fathead minnows, bluegills, and rainbow trout were unaffected by up to 100 ppm of these polymers in the water for 90 days. Minnows were all killed by a solution of 2500 ppm of hydrolyzed polyacrylamide; the fish had extreme difficulty swimming due to the high viscosity of this test solution. Cationic derivatives of polyacrylamides can kill fish and should not be allowed into natural waterways. Polyacrylamide, cationic- and anionic-modified polyacrylamides, and poly(acrylic acid) and its salts strongly adsorb to many solids. They are thereby effectively removed from sewage in a waste treatment plant as a component of the sludge. Polyacrylamides showed no evidence of toxicity at a dosage of 10^4 mg/kg in rabbits [45].

8. References

1. *Acrylic Acid, A Summary of Safety and Handling,* Company brochure, Celanese Chemical, Dallas 1980.
2. *Aqueous Acrylamide Monomer Technical Book,* Dow Chemical, Midland 1988.
3. L. S. Kirch, J. A. Kargol, J. W. Magee, W. S. Stuper, *Plant Oper. Prog.* **7** (1988) 270.
4. L. B. Levy, J. D. Penrod, *Plant Oper. Prog.* **8** (1989) 105 – 108.
5. P. M. Edwards, *Br. J. Ind. Med.* **32** (1975) 31 – 38.
6. V. F. Gromov et al., *Eur. Polym. J.* **16** (1980) 529 –535.
7. H. Ito, S. Shimizu, S. Suzuki, *Kogyo Kagaku Zasshi* **58** (1955) 194 – 196.
8. G. Blauer, *Trans. Faraday Soc.* **56** (1960) 606 – 612.
9. V. A. Kabanov, D. A. Topchiev, T. M. Karaputdze, *J. Polym. Sci. Polym. Symp.* **42** (1973) 173 – 183.
10. S. P. Manickam, K. Venkatarao, N. R. Subbarat-nam, *Eur. Polym. J.* **15** (1979) 483 – 487.
11. H. Kheradmand, J. Francois, V. Plazanet, *J. Appl. Polym. Sci.* **39** (1990) 1847 – 1857.
12. S. P. Manickam, K. Venkatarao, U. C. Singh, N. R. Subbaratnam, *J. Polym. Sci. Polym Chem. Ed.* **16** (1978) 2701 – 2702.
13. S. P. Manickam, N. R. Subbaratnam, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 1679 – 1684.
14. J. P. Riggs, F. Rodriguez, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 3151 – 3165.

15. J. P. Riggs, F. Rodriguez, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 3167 – 3181.
16. P. Elayaperumal, T. Balakrishnan, M. Santappa, R. W. Lenz, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 2471 – 2479.
17. S. Lenka, P. L. Nayak, S. Ray, *J. Polym. Sci. Polym. Chem. Ed.* **22** (1984) 959 – 965.
18. K. Behari, K. Gupta, *Colloid Polym. Sci.* **262** (1984) 677 – 682.
19. J. S. Shukla, R. K. Tiwari, *J. Polym. Sci. Polym. Chem. Ed.* **19** (1981) 1517 – 1524.
20. S. Lenka, P. L. Nayak, *J. Polym. Sci. Polym. Chem. Ed.* **25** (1987) 1563 – 1568.
21. R. K. Samal, P. L. Nayak, T. R. Mohanty, *Macromolecules* **10** (1977) 489 – 492.
22. S. M. Shawki, A. E. Hamielec, *J. Appl. Polym. Sci.* **23** (1979) 3155 – 3166.
23. S. Ponratnam, S. L. Kapur, *Makromol. Chem.* **178** (1977) 1029 – 1038.
24. R. Z. Greenley in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* 3rd ed., John Wiley & Sons, New York 1989.
25. A. Katchalsky, *J. Polym. Sci.* **2** (1947) 432.
26. R. A. M. Thomson, C. K. Ong, C. M. Rosser, J. M. Holt, *Makromol. Chem.* **184** (1983) 1885-1892.
27. American Cyanamid, AU 8 818 380-A, 1989 (W. B. Davies).
28. J. H. Breitenbach, H. F. Kauffmann, *Makromol. Chem.* **175** (1974) 2597 – 2604.
29. Sanyo Chemical Industries, JP 63 275 607-A, 1988 (K. Tanaka, Y. Fujiura, T. Sumiya).
30. Nippon Shokubai Kagaku, JP 85 024 806-B, 1985 (T. Tsubakimoto, T. Shimomura, Y. Iric).
31. K. Hughes, G. Swift, *Adv. Chem. Ser.* **213** (1986) 145 – 151.
32. Nalco Chemical, US 4 143 222, 1979 (L. A. Goretta, R. R. Otremba).
33. Stockhausen, US 4 857 610, 1989 (M. Chmelir, J. Pauen).
34. Nalco Chemical, US 3 278 506, 1966 (W. M. Chamot, B. H. Robin).
35. Dow Chemical, US 3 284 393, 1966 (J. W. Vander-hoff, R. M. Wiley).
36. K.-J. Kim, E. Ruckenstein, *Makromol. Chem. Rapid Commun.* **9** (1988) 285 – 290.
37. Dow Chemical, US 4 565 836, 1986 (J. P. Easterly, R. R. Emerson, D. I. Lee).
38. V. Glukhikh, C. Graillat, C. Pichot, *J. Polym. Sci., Part A: Polym. Chem.* **25** (1987) 1127 – 1161.
39. Dow Chemical, US 2 982 749, 1961 (R. E. Friedrich, R. M. Wiley).
40. Kao Soap Co., US 4 446 261, 1984 (S. Harada, H. Yamasaki).
41. Kao Soap KK, JP 11 67 304-A, 1989 (Y. Sato, Y. Kono, Y. Yasuda, R. Tsushima).
42. Dow Chemical, US 4 708 997, 1988 (F. W. Stanley, Jr., J. C. Lamphere, Y. Chonde).
43. BF Goodrich, DE 1 595 727-B, 1975.
44. P. J. Kay, F. E. Treloar, *Makromol. Chem.* **175** (1974) 3207.
45. American Cyanamid, *Cyanamer, Polyacrylamides for the Processing Industries,* Wayne, N. J. 1982.
46. Nippon Shokubai Kagaku Kogyo, US 4 734 478, 1988 (T. Tsubakimoto et al.).
47. Nippon Shokubai Kagaku, JP 11 5 621-A, 1972 (T. Tsubakimoto).
48. Dow Chemical, US 2 960 486, 1960 (D. J. Pye).
49. Dow Chemical, US 3 780 006, 1973 (M. Zweigle).
50. Stockhausen, DE 3 724 709-A1, 1987 (M. Chmelir).
51. American Cyanamid, US 4 132 844, 1979 (M. W. C. Coville).
52. Casella, US 4 794 166, 1988 (F. Engelhardt et al.).
53. W. Nagashiro, T. Tsunoda, *J. Appl. Polym. Sci.* **21** (1977) 1149.
54. R. D. Shupe, *JPT J. Pet. Technol.* **33** (1981) 1513.
55. W.-M. Kulicke, R. Kniewski, *Makromol. Chem.* **181** (1980) 823.
56. W.-M. Kulicke, H.-H. Hörl, *Colloid Polym. Sci.* **263** (1985) 530 – 540.
57. C. Chen, D. D. Muccio, T. S. Pierre, *Macromolecules* **18** (1985) 2154 – 2157.
58. K. Hatada, T. Kitayama, K. Ute, *Polym. Bull. (Berlin)* **9** (1983) 241 – 244.
59. L. J. Thompson Hughes, D. B. Fordyce, *J. Polym. Sci.* **22** (1956) 509.
60. G. S. Haldankar, H. G. Spencer, *J. Appl. Polym. Sci.* **37** (1989) 3137 – 3146.
61. J. Klein, R. Heitzmann, *Makromol. Chem.* **179** (1978) 1895 – 1904.
62. S. Newman, W. R. Krigbaum, C. Laugier, P. J. Flory, *J. Polym. Sci.* **14** (1954) 451.
63. A. Silberberg, J. Eliassaf, A. Katchalsky, *J. Polym. Sci.* **23** (1957) 259.
64. A. Takahashi, M. Nagasawa, *J. Am. Chem. Soc.* **86** (1964) 543.
65. T. Kato, T. Tokuya, T. Nozaki, A. Takahashi, *Polymer* **25** (1984) 218 – 224.
66. H. K. Yuen, E. P. Tam, J. W. Bulock, *Anal. Calorim.* **5** (1984) 13 – 24.
67. American Cyanamid, *The Chemistry of Acrylamide,* Wayne, NJ, 1969.

68. T. Schwartz, J. Francois, G. Weill, *Polymer* **21** (1980) 247.
69. W.-M. Kulicke, R. Kniewski, J. Klein, *Prog. Polym. Sci.* **8** (1982) 373 – 478.
70. M. Bohdanecky, V. Petrus, B. Sedlacek, *Makromol. Chem.* **184** (1983) 2061.
71. W. E. Fitzgerald, L. E. Nielsen, *Proc. R. Soc. London, A* : 282 (1964) 137.
72. N. M. Wiederhorn, A. R. Brown, *J. Polym. Sci.* **8** (1952) 651.
73. A. Katchalsky, *J. Polym. Sci.* **6** (1951) 145.
74. P. Molyneux: *Water Soluble Synthetic Polymers,* **vol. 1,** CRC Press, Boca Raton, Florida, 1984, p. 104.
75. K. Nagase, K. Sakaguchi, *J. Polym. Sci., Part A* **3** (1965) 2475 – 2482.
76. W.-M. Kulicke, H.-H. Hörl, *Colloid Polym. Sci.* **263** (1985) 530 – 540.
77. K. J. MacCarthy, C. W. Burkhardt, D. P. Parazak, *J. Appl. Polym. Sci.* **33** (1987) 1683 – 1698.
78. W.-M. Kulicke, R. Kniewske, H.-H. Hörl, *Angew. Makromol. Chem.* **87** (1980) 195 – 204.
79. H. Rios, L. Gargallo, D. Radic, *Makromol. Chem.* **182** (1981) 665 – 668.
80. M. Heskins, J. E. Guillet, *J. Macromol. Sci. Chem.* **A2** (1968) 1441 – 1455.
81. O. Smidsrod, J. E. Guillet, *Macromolecules* **2** (1969) 272.
82. K. Kubota, S. Fujishige, I. Ando, *Polym. J.* **22** (1990) 15 – 20.
83. S. Krause et al., *J. Polym. Sci. Part, A* **3** (1965) 3573.
84. L. Trossarelli, M. Meirone, *J. Polym. Sci.* **57** (1962) 445 – 452.
85. L. W. Fisher, A. R. Sochor, J. S. Tan, *Macromolecules* **10** (1979) 949 – 959.
86. C. L. McCormick, D. L. Elliot, *Macromolecules* **19** (1986) 542 – 547.
87. Röhm, *Acrylic Resins: Polyelectrolytes,* Darmstadt 1985.
88. Dow Chemical, *Pusher Mobility Control Polymers: Enhanced Oil Recovery,* Houston 1980.
89. Rohm & Haas, EP 123 482-B, 1986 (W. M. Hann, J. Dupre, J. Natoli).
90. BF Goodrich Co., EP 213 500-A, 1987 (M. K. Nagarajan, F. J. Wherley, J. W. Frimel).
91. Dow Chemical, US 4 541 935, 1985 (V. G. Constein, M. T. King).
92. Dow Chemical, US 3 669 103, 1972 (B. G. Harper, R. N. Bashaw, B. L. Atkins).
93. Procter & Gamble, US 4 673 402, 1987 (D. A. Gellert, D. I. Houghton, P. T. Weisman).
94. Kao Soap Co., US 4 364 992, 1982 (O. Ito, N. Kazunori).
95. Kimberly-Clark, US 4 675 012, 1987 (J. A. Rooyakkers).
96. Kao Corp., US 4 519 799, 1985 (A. Sakurai, H. Mizutani).
97. Dow Chemical, EP 24 631-A, 1981 (K. E. Bow).
98. Unilever NV, EP 72 214-B, 1986 (J. A. Bosly, R. B. Dehnel, S. A. Symien).
99. Nippon Shokubai Kagaku, JP 88 046 724-B, 1988 (T. Tsubakimoto, T. Shimomura, Y. Iric, Y. Masuda).
100. BF Goodrich, US 3 330 729, 1967 (C. H. Johnson Jr.).
101. Allied Colloids, EP 77 618-B, 1986 (J. B. Clarke, J. F. Firth).
102. Amer. Cyanamid, US 3 840 489, 1974 (E. Strazdins).
103. Halliburton, US 4 524 829, 1985 (D. J. Hanlon, S. W. Almond).
104. S. Yuzhgeologiya, SU 1 239 140-A, 1986 (V. A. Evetskii, I. M. Davydov, O. K. Belkin).
105. Allied Colloids, EP 101 253-B, 1987 (J. B. Clarke).
106. Agrisorb Pty, EP 159 344-A, 1985 (P. R. Evans).
107. Nalco Chemical, US 4 417 992, 1983 (B. R. Bhattachar, W. J. Roe).
108. R. F. S. Freitas, E. L. Cussler, *Sep. Sci. Technol.* **22** (1987) 911.
109. A. S. Hoffman, A. Afrassiabi, L. C. Dong, *J. Controlled Release* **4** (1986) 213.
110. K. Ogino, *Yuki Gosei Kagaku Kyokaishi* **33** (1975) 504 – 506.
111. Dow Chemical, US 4 737 541, 1988 (D. L. Stavenger, J. W. Sanner, P. L. Slaber).
112. General Mills Chemical, US 3 931 089, 1976 (C. L. Karl).
113. Halliburton Co., US 4 107 057, 1978 (W. R. Dill, E. A. Elphingstone).
114. M. Fox, T. Gibson, R. Mulach, T. Sasano: "Acrylic Acid and Esters", *Chemical Economics Handbook,* SRI International, Menlo Park, Feb. 1990.
115. C. B. Shaffer, US Environmental Protection Agency, Office of Toxic Substances, document no.878 211 672, Washington 1983.
116. Rohm and Haas, *Polymeric Additives for Aqueous Systems,* Philadelphia 1986.
117. Rohm and Haas, *Acrysol Polymers for Household Laundry Detergents,* Philadelphia 1986.
118. D. D. McCollister, C. L. Hake, S. E. Sadek, V. K. Rowe, *Toxicol. Appl. Pharmacol.* **7** (1965) 639 – 651.

Polyacrylates

Polyacrylamides and Poly(Acrylic Acids) and Polymethacrylates are separate keywords. For polyacrylonitriles, see → Fibers, 4. Synthetic Organic, Chap. 5. General aspects of polymers, properties and testing, processing, additives, states of order, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Erich Penzel, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Raw Materials 2
3. Production 6
3.1. Emulsion Polymerization 6
3.1.1. Starting Materials 6
3.1.2. Production Processes 7
3.2. Solution Polymerization 8
3.3. Elimination of Residual Monomers 11
3.4. Toxicology and Environmental Aspects 11
4. Properties 12
4.1. Polymer Solutions and Dispersions 12
4.1.1. Polymer Solutions 12
4.1.2. Polymer Dispersions 12
4.2. Properties of the Polymers 14
5. Uses 17
5.1. Paints and Coatings 17
5.2. Paper Industry 18
5.3. Adhesives and Sealing Compounds 18
5.4. Textile Industry 19
5.5. Leather Industry 19
5.6. Miscellaneous Uses 19
6. Economic Aspects 19
7. References 20

1. Introduction

Acrylic acid and acrylate esters have been known since the middle of the nineteenth century. A systematic investigation of acrylate esters was published in 1901 by H. von Pechmann and O. Röhm [12]. A process for the industrial production of acrylate esters was developed in 1928 by H. Bauer [13] of Röhm & Haas. Solution polymers have been produced from methyl methacrylate since 1927 by Röhm & Haas (Darmstadt, Germany). Emulsion polymers were first developed on an industrial scale in 1929/30 by H. Fikentscher [14], [15], and were introduced onto the market by BASF as a polymer dispersion named " Corialgrund" for the surface finishing of leather.

The use of polyacrylates in many fields of application increased rapidly with the development of new methods for producing acrylic acid and acrylate esters (e.g., the process described by W. Reppe in 1953 [16]).

Acrylate esters are compounds of the general formula

where R is an alkyl group.

Methyl, ethyl, *n*-butyl, and 2-ethylhexyl acrylate are produced on a large scale. Other acrylate esters such as *tert*-butyl, isobutyl, or lauryl acrylate are also produced industrially, but have not yet become quantitatively important.

Homopolymers of acrylate esters are suitable only for a few areas of application, whereas their copolymers have numerous potential applications. Methacrylates (particularly methyl methacrylate), styrene, acrylonitrile, vinyl acetate, vinyl chloride, vinylidene chloride, and butadiene are used as comonomers to obtain the desired copolymer properties.

In addition, auxiliary monomers are frequently incorporated in polyacrylates to obtain specific technical properties in dispersions. Although these auxiliary monomers are only present in low concentration in the polymer, they have a substantial influence on the colloid chemistry and other properties of polymer dispersions. Commonly used auxiliary monomers

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_157

include acrylic acid, methacrylic acid, acrylamide, methacrylamide, and other monomers with functional groups.

Polyacrylates can be produced by various processes, emulsion polymerization being the most important industrial method. Solution polymerization is also of practical importance. Bulk polymerization is only performed on a small scale, but will presumably become more important in the future; printing plates are produced by this process. Radiation-curable paints also undergo bulk polymerization. In contrast to the methacrylates, suspension polymerization is of no technical importance.

2. Raw Materials

The production and properties of acrylic acid and its esters are described elsewhere, → Acrylic Acid and Derivatives [17]. Acrylates produced industrially by esterification of acrylic acid are available with a purity of 99 – 99.5 %. Special acrylates may have a somewhat lower purity. Acrylate esters are clear, colorless liquids, which are normally stabilized with 15 ± 5 ppm of hydroquinone monomethyl ether. These small amounts of stabilizer do not affect polymerization. Hydroquinone monomethyl ether does not cause discoloration during polymerization and is therefore preferred to hydroquinone. Since the stabilizer is effective only in the presence of oxygen, the monomers must be stored under air and not under an inert gas. In exceptional cases (e.g., overseas dispatch or in special acrylates), a higher amount of stabilizer is used ($\geq$ 100 ppm).

Unstabilized monomeric acrylates can be stored for only a few weeks at temperatures below 10 °C. The stabilized monomers are generally stored below 20 °C. Acrylic acid and methacrylic acid should not, however, be stored below 18 °C because they crystallize and spontaneously polymerize at this temperature.

Some physical properties of acrylates, comonomers, and auxiliary monomers are summarized in Table 1. Further data on the industrially most important acrylate esters are listed in Table 2. Guidelines for analyzing acrylates to determine their purity, acid number, water content, and stabilizer content [17] can be obtained from the technical data sheets of the manufacturers [18].

Acrylates with appropriate substituents can be copolymerized to obtain copolymers with special properties such as low flammability, antistatic behavior, cross-linkability, elasticity, solubility or insolubility in certain solvents, hydrophilicity, or hydrophobicity [22]. Such substituents include halogens, hydroxyl, epoxy, or *N*-methylol groups. Hydrophilicity is achieved by incorporating acrylic acid, acrylamide, or hydroxy-functional monomers (e.g., hydroxypropyl or hydroxyethyl acrylate). Incorporation of relatively large amounts of acrylic acid can increase hydrophilicity to such an extent that the emulsion polymers dissolve completely in water as a result of salt formation, to form a clear solution at pH 6 – 8 [23].

Compatibility with cationic auxiliaries is achieved by incorporating cationic monomers such as dimethylaminoethyl methacrylate or diethylaminoethyl acrylate.

Physical properties of selected acrylates and methacrylates with reactive groups are listed in Table 3 [17].

3. Production

Acrylates can be polymerized extremely easily because their carbonyl groups are adjacent to a vinyl group. Polyacrylates are produced almost exclusively by radical polymerization; conventional radical formers (e.g., peroxides and other per compounds) or azo starters are used as initiators. Polymerization can also be initiated photochemically, by γ-rays, or by electron beams. Although ionic (particularly anionic) polymerization [24] is possible, this process is not used industrially. Ordered (i.e., cis or trans) polymers can be produced by this polymerization process. The heat of reaction in the exothermic polymerization of acrylates is ca. 60 – 80 kJ/mol, and must be removed if the process is to be controlled effectively [25], [26]. The kinetics of radical polymerization and copolymerization are described in detail elsewhere (→ Polymerization Processes, Chap. 2.2., → Polymerization Processes, Chap. 2.3.). The reactivity ratios r_1 and r_2 for industrially important acrylates and comonomers are summarized in Table 4.

Table 1. Physical properties of monomeric acrylates and other comonomers [18]

Monomers	CAS registry no.	M_r	*fp*, °C	*bp* (101.3 kPa), °C	d_4^{25}	n_D^{25}	Solubility in water, g/100 cm^3	
							25 °C	80 °C
Acrylates								
Methyl acrylate	[*96-33-3*]	86	−74	80	0.951	1.4010	5.2	5.3
Ethyl acrylate	[*140-88-5*]	100	−69	99	0.917	1.4037	1.6	1.8
n-Propyl acrylate	[*925-60-0*]	114	−65	119	0.904	1.4100	1.5	1.6
n-Butyl acrylate	[*141-32-2*]	128	−61	148	0.895	1.4160	0.15	0.20
Isobutyl acrylate	[*106-63-8*]	128	−60	135	0.885	1.4122	0.18	0.22
sec-Butyl acrylate	[*2998-08-5*]	128	−63	131	0.887	1.4110	0.21	0.22
tert-Butyl acrylate	[*1663-39-4*]	128	−68	120	0.879	1.4080	0.15	0.15
n-Hexyl acrylate	[*2499-95-8*]	156	−46	89 [a]	0.879	1.4247	0.04	0.08
2-Ethylhexyl acrylate	[*103-11-7*]	184		217	0.880	1.4332	0.04	0.04
Lauryl acrylate	[*2156-97-0*]	240	−8	129 [b]	0.870	1.4437	<0.001	<0.001
Comonomers								
Methyl methacrylate	[*80-62-6*]	100	−48	101	0.939	1.4119	1.5	1.6
Ethyl methacrylate	[*97-63-2*]	114	−76	118	0.909	1.4138	0.54	0.70
n-Butyl methacrylate	[*97-88-1*]	142		163	0.889	1.4220	0.08	0.15
Isobutyl methacrylate	[*97-86-9*]	142	−35	155	0.879	1.4172	0.13	0.19
tert-Butyl methacrylate	[*585-07-9*]	142	−47	144	0.873	1.4120	0.10	0.16
n-Hexyl methacrylate	[*142-09-6*]	170	−65	207	0.894	1.4290	<0.03	0.03
Acrylonitrile	[*107-13-1*]	53	−82	78	0.805	1.3888	8.3	9.7
Methacrylonitrile	[*126-98-7*]	67	−36	90	0.795	1.3977	2.6	3.0
Styrene	[*100-42-5*]	104	−31	145	0.905	1.5455	0.02	0.06
1,3-Butadiene	[*106-99-0*]	54	−109	−4	0.615	1.4220 [c]	0.081	
1,2-Butadiene	[*590-19-2*]	54	−136	19	0.645	1.4200		
Vinyl acetate	[*108-05-4*]	86	−97	73	0.933	1.3959	2.9	2.9
Vinyl propionate	[*105-38-4*]	100	−81	95	0.917	1.4042	0.67	0.68
Vinyl chloride	[*75-01-4*]	62	−154	−13	0.911	1.398 [d]	0.11	
Vinylidene chloride	[*75-35-4*]	97	−123	32	1.208	1.4237	0.25	
Auxiliary monomers								
Acrylic acid	[*79-10-7*]	72	14	141	1.044	1.4185	∞	∞
Methacrylic acid	[*79-41-4*]	86	17	161	1.015	1.4288	∞	∞
Acrylamide	[*79-06-1*]	72	87	125 [a]	1.122		76	∞
Methacrylamide	[*79-39-0*]	85	112				27	

[a] 2.5 kPa.
[b] 3.8 kPa.
[c] −6 °C.
[d] Calculated.

Table 2. Further properties of industrially important acrylates

Property	Methyl acrylate	Ethyl acrylate	Butyl acrylate	2-Ethylhexyl acrylate
Vapor pressure, kPa				
0 °C	4.2	1.2	0.14	
25 °C	13	4.9	0.68	0.03
50 °C	36	17	2.8	0.16
75 °C	83	44	8.2	0.63
100 °C			22	2.1
125 °C			52	5.8
Heat of vaporization at boiling point, kJ/g	0.39	0.35	0.28	0.26
Specific heat (liquid), J g^{-1} °C^{-1}	2.00	1.967	1.925	1.88
Heat of polymerization, kJ/g	0.915	0.777	0.605	0.424
Rate constant of chain growth, L mol^{-1} s^{-1}	1000 (50 °C) [19]	800 (50 °C) [20]	679 (30 °C) [21]	155 (50 °C) [20]
Rate constant of chain termination $k_t \times 10^{-6}$, L mol^{-1} s^{-1}	3.55 (50 °C) [19]	1.76 (50 °C) [20]	6.4 (30 °C) [21]	0.233 (50 °C) [20]

Table 3. Acrylates and methacrylates with functional groups

Monomers	Formula	CAS registry no.	M_r	*fp*, °C	*bp*, °C (kPa)[a]	d_4^{25}	n_D^{25}	Solubility in water at 25 °C, g/100 cm³
Carboxyl								
Fumaric acid	HOOCCH=CHCOOH	[*110-17-8*]	116	286	165 (0.17)	1.625		0.7
Maleic acid	HOOCCH=CHCOOH	[*110-16-7*]	116	130		1.609	1.544[b]	78.8
Crotonic acid	$CH_3CH=CHCOOH$	[*3724-65-0*]	86	16	169 (101)	1.026	1.442[c]	
Itaconic acid	$CH_2=C(CH_2COOH)(COOH)$	[*97-65-4*]	130	175	decomp.	1.573	1.512[b]	9.1
Di- and triacrylates								
Ethylene glycol diacrylate	$CH_2=CHCOOCH_2CH_2OCOCH=CH_2$	[*2274-11-5*]	170		70 (0.13)		1.453[c]	
Diethylene glycol diacrylate	$H_2C=CHCOO(CH_2CH_2O)_2COCH=CH_2$	[*4074-88-8*]	224		94 (0.03)		1.457[c]	
Triethylene glycol diacrylate	$H_2C=CHCOO(CH_2CH_2O)_3COCH=CH_2$	[*1680-21-3*]	258		120 (0.02)	1.108	1.462[c]	5.0
Tetraethylene glycol diacrylate	$H_2C=CHCOO(CH_2CH_2O)_4COCH=CH_2$	[*17831-71-9*]	302	−53	328 (101)	1.078	1.463	9.1
Butanediol diacrylate	$CH_2=CHCOO(CH_2)_4OOCCH=CH_2$	[*1070-70-8*]	198	−5	120 (1.3)	1.049	1.455[c]	0.6
Hexanediol acrylate	$CH_2=CHCOO(CH_2)_6OOCCH=CH_2$	[*13048-33-4*]	226	9	107 (0.03)	1.015	1.458	0.12
Neopentyl glycol diacrylate	$CH_2=CHCOOCH_2C(CH_3)_2CH_2OOCCH=CH_2$	[*2223-82-7*]	212		101 (0.4)	1.024	1.451	0.9
Trimethylolpropane triacrylate	$(CH_2=CHCOOCH_2)_3CCH_2CH_3$	[*15625-89-5*]	296	−22	126 (0.03)	1.102	1.474	0.09
Tripropylene glycol diacrylate	$CH_2=CHCOO(CH(CH_3)CH_2O)_3OCCH=CH_2$	[*42978-66-5*]	300	−60	109 (0.03)	1.036		
Di-(meth)acrylamides								
N,N'-Methylene bisacrylamide	$CH_2=CHCONHCH_2HNCOCH=CH_2$	[*110-26-9*]	154	188				3.5
N,N'-Methylene bismethacrylamide	$CH_2=C(CH_3)CONHCH_2HNCOC(CH_3)=CH_2$	[*2359-15-1*]	182	164				
Hydroxyl								
2-Hydroxyethyl acrylate	$CH_2=CHCOOCH_2CH_2OH$	[*818-61-1*]	116	< −60	60 (0.13)	1.114[c]	1.451[c]	∞
2-Hydroxyethyl methacrylate	$CH_2=C(CH_3)COOCH_2CH_2OH$	[*868-77-9*]	130	−12	95 (1.3)	1.06[c]	1.450	∞
Hydroxypropyl acrylate	$CH_2=CHCOO(CH_2)_3OH$	[*25584-83-2*]	130		63 (0.13)	1.06[c]	1.448[c]	∞
Propylene glycol methacrylate	$CH_2=C(CH_3)COOCH_2CH(OH)CH_3$	[*2761-09-3*]	144		96 (1.3)	1.024	1.446	13.0
Butanediol monoacrylate	$CH_2=CHCOO(CH_2)_4OH$	[*2478-10-6*]	144		130 (2.7)	1.039	1.453[c]	∞
Ethers								
Ethyldiglycol acrylate	$CH_2=CHCO(OC_2H_4)_2OC_2H_5$	[*32002-24-7*]	188	−66	99 (0.5)	1.006	1.438[c]	14.0
Amines								
Dimethylaminoethyl acrylate	$CH_2=CHCOO(CH_2)_2N(CH_3)_2$	[*2439-35-2*]	143	−92	70 (2.9)	0.936[c]	1.437[c]	∞
Dimethylaminoethyl methacrylate	$CH_2=C(CH_3)COO(CH_2)_2N(CH_3)_2$	[*2867-47-2*]	157	−30	68 (1.3)	0.921	1.438	∞

Table 3. (Continued)

N-(3-Dimethylaminopropyl)-methacrylamide	$CH_2{=}C(CH_3)CONH(CH_2)_3N(CH_3)_2$	[*5205-93-6*]	182		92 (0.005)		1.479[c]	
Diethylaminoethyl acrylate	$CH_2{=}CHCOO(CH_2)_2N(C_2H_5)_2$	[*2426-54-2*]	171	−100	97 (3.6)	0.925[c]	1.443[c]	10.7
tert-Butylaminoethyl methacrylate	$CH_2{=}C(CH_3)COO(CH_2)_2NHC(CH_3)_3$	[*3775-90-4*]	185	< −70	105 (1.6)	0.914	1.440	11.5
Halogen								
2-Chloroacrylonitrile	$CH_2{=}CClCN$	[*920-37-6*]	88	−65	89 (101)	1.085	1.428	1
N-Hydroxymethyl								
N-Hydroxymethyl acrylamide	$CH_2{=}CHCONHCH_2OH$	[*924-42-5*]	101	79				∞
N-Hydroxymethyl methacrylamide	$CH_2{=}C(CH_3)CONHCH_2OH$	[*923-02-4*]	115	54		1.145		∞
Sulfonic acid								
2-Sulfoethyl methacrylate	$CH_2{=}C(CH_3)COOCH_2CH_2SO_3H$	[*10595-80-9*]	194		decomp.	1.325		∞
2-Acrylamido-2-methylpropane-sulfonic acid	$CH_2{=}CHCONHC(CH_3)_2CH_2SO_3H$	[*15214-89-8*]	207	185		1.401	1.560[b]	150
Polyfunctional								
Glycidyl methacrylate	$CH_2{=}C(CH_3)COOCH_2CH{-}CH_2$ (O)	[*106-91-2*]	142		75 (1.3)	1.073	1.448	3.0
Diacetone acrylamide	$CH_2{=}CHCONHC(CH_3)_2CH_2COCH_3$	[*2873-97-4*]	169	57	107 (0.03)			>100
Diacetone methacrylamide	$CH_2{=}C(CH_3)CONHC(CH_3)_2CH_2COCH_3$	[*22029-67-0*]	183		138 (1.5)			
Acrylamidoglycolic acid	$CH_2{=}CHCONHCH(OH)COOH$	[*6737-24-2*]	145	95				13
Methylacrylamidoglycol methyl ether	$CH_2{=}CHCONHCH(OCH_3)COOCH_3$	[*77402-03-0*]	173	73				18

[a] To convert kPa to mm Hg, multiply by 7.5; to convert to mbar, multiply by 10. [b] Calculated from refractive index. [c] 20 °C.

Table 4. Copolymerization parameters r_1 and r_2∗ [27]

Monomer 1	Monomer 2	r_1	r_2	Reference
Methyl acrylate	acrylonitrile	0.844	1.540	[28]
Q=0.45	butadiene	0.070	1.090	[29]
e=0.64 [46]	methyl methacrylate	0.400	2.150	[30]
	styrene	0.168	0.722	[28]
	vinyl acetate	6.300	0.030	[31]
	vinyl chloride	4.400	0.093	[28]
	vinylidene chloride	1.000	1.000	[32]
Ethyl acrylate	acrylonitrile	0.810	1.160	[33]
Q=0.41	methyl methacrylate	0.280	2.000	[34]
e=0.55 [46]	styrene	0.152	0.787	[35]
	vinylidene chloride	0.720	0.580	[36]
Butyl acrylate	acrylonitrile	0.820	1.080	[37]
Q=0.38	butadiene	0.074	1.040	[29]
e=0.85 [46]	methylmethacrylate	0.430	1.880	[38]
	styrene	0.180	0.840	[39]
	vinyl acetate	3.480	0.018	[40]
	vinyl chloride	4.400	0.070	[41]
	vinylidene chloride	0.873	0.934	[42]
2-Ethylhexyl acrylate	styrene	0.310	0.960	[43]
Q=0.37	vinyl acetate	7.500	0.040	[44]
e=0.24 [46]	vinyl chloride	4.150	0.160	[45]

∗ For definitions, see →Polymerization Processes.

3.1. Emulsion Polymerization

Emulsion polymerization is the most important industrial process for producing polyacrylates and their copolymers [47–50]. Polymers with high molecular masses are obtained at high polymerization rates; polymerization times are consequently short. Since the viscosity is generally low, the heat of polymerization can easily be removed via the aqueous phase. Emulsion polymers with solids contents of up to 60 wt % and molecular masses of 10^5 – 10^6 can be produced. These represent advantages compared with solution polymerization, and in addition there is no safety hazard due to flammable solvents.

3.1.1. Starting Materials

The most important starting materials for emulsion polymerization are water, monomers, emulsifiers, initiators, and, if necessary, modifiers.

Water. The water should preferably be distilled or completely demineralized because electrolytes can affect emulsion stability.

Monomers. The monomer composition is governed by the properties required by the relevant area of application.

Emulsifiers. Anionic, nonionic, and, less often, cationic or amphoteric compounds are normally used for emulsification. A combination of ionic and nonionic emulsifiers can considerably improve the stability of the dispersions.

Anionic emulsifiers include sodium, potassium, or ammonium salts of fatty acids and sulfonic acids; alkali salts of C_{12-16}-alkyl sulfates; ethoxylated and sulfated or sulfonated fatty alcohols; alkyl phenols; and sulfodicarboxylate esters. The emulsifiers are used at concentrations of 0.2 – 5 wt %, based on monomers.

Nonionic emulsifiers include ethoxylated fatty alcohols and alkyl phenols with 2 – 150 ethylene oxide units per molecule.

Cationic emulsifiers include ammonium, phosphonium, and sulfonium compounds with a hydrophobic moiety that contains at least one long aliphatic hydrocarbon chain.

Many companies offer very large ranges of emulsifiers and emulsifier mixtures that vary in terms of chemical structure. The emulsifier has several functions in emulsion polymerization. It is required for the preparation of the monomer emulsion, and later stabilizes the final dispersion. The particle size (diameter) of the dispersion can be adjusted in the range 50 – 1000 nm

by the choice of emulsifier, its amount, and the mode of addition.

Protective colloids such as cellulose derivatives, starch, or poly(vinyl alcohol) can also be used to stabilize the dispersions. Protective colloids are, however, less suitable for initiating particle formation.

Initiators. Water-soluble peroxo compounds (e.g., alkali persulfates, ammonium persulfate, or hydrogen peroxide) are used as initiators with a polymerization temperature between 50 and 85 °C. The initiators are used in amounts of 0.05 – 1 wt % (usually 0.2 – 0.5 wt %) based on monomer. Higher initiator concentrations may adversely affect the stability of the dispersion (electrolyte content) or improve it (additional stabilization); they also lower the molecular mass of the product.

If redox systems are used for initiation [51–53], polymerization can be performed below 50 °C. Oxidizing agents include hydrogen peroxide and alkali persulfates. Iron(II) sulfate, sodium bisulfite, sodium thiosulfate, or sodium formaldehyde sulfoxylate can be used as reducing agent. Iron(II) salts can also be added in small amounts as catalysts. The optimal quantitative ratio of oxidizing agent to reducing agent has to be determined experimentally.

Alkali persulfate cannot be used as an initiator in the preparation of cationic dispersions due to salt formation; water-soluble organic peroxides or hyrogen peroxide are more suitable. Polymerization is generally performed in the pH range 2 – 7.

Modifiers. The molecular mass of the polymer can be reduced by adding modifiers with the result that the polymers become more tacky. Common modifiers include halogen-containing compounds (e.g., carbon tetrachloride, carbon tetrabromide, bromoform, benzyl bromide, bromotrichloromethane), or thiols (e.g., butyl or dodecyl thiol).

3.1.2. Production Processes

Emulsion polymers are produced from acrylates in batch and semibatch (continuous emulsion or monomer feed) processes. All processes should be performed under nitrogen since oxygen inhibits polymerization.

Batch Process. The monomer emulsion is added to the cooled reaction vessel (<15 °C). The redox initiator is then added, generally as an aqueous solution. Polymerization begins and the temperature rises over the course of 0.5 – 1 h and then falls due to constant cooling and removal of the heat of reaction. The polymerization rate can sometimes be controlled by appropriate metering of the initiator.

The advantage of the batch procedure is the larger space – time yield. Disadvantages include the poor utilization of the cooling capacity, as well as poor reproducibility of the procedure and product properties.

If cooling is insufficient to prevent the reaction going out of control, the batch can be split into two or more partial batches. If monomers and initiator solution are added in an infinite number of small steps, this is referred to as a continuous feed (semibatch) process [54]. Accurate temperature control is easier with the semibatch (continuous feed) process. Mixed batch and semibatch procedures are used in practice.

Semibatch Procedure. In the *semibatch procedure with emulsion feed* [55], [56], the reaction vessel contains some water and is heated to the desired reaction temperature. Part (5 – 10 %) of the monomer emulsion and initiator solution is then added as an initial batch to the reactor.

Alternatively, some of the emulsifier and the water is added, the remaining emulsifier is then used to prepare the monomer emulsion. The initial batch is polymerized for 15 – 30 min. The number of particles and the particle size of the polymer dispersion are determined in this starting phase (batch).

Product properties can be modified by adjusting the proportion of emulsifier in the initial batch and the feed. Instead of an initial batch, a finely particulate dispersion can also be used as "seed". The monomer emulsion and initiator solution are then added from separate tanks over 2 – 3 h.

In the *semibatch procedure with monomer feed* the water and emulsifiers are first added to the reactor. The monomer mixture, initiator solution, and the solution of auxiliaries are then fed in continuously.

In semibatch processes specific product properties can be selectively adjusted by small changes (e.g., of temperature, pH, and delayed addition of certain monomers) [57]. If polymerization is performed under specified conditions, the properties of the resulting dispersions remain virtually constant from batch to batch. These advantages have to be offset, however, by increased complexity, for example as regards temperature control and feed regulation.

Many "specialty" chemicals are produced by batch or semibatch processes. Continuous emulsion polymerization is mainly used on an industrial scale in the production of styrene – butadiene rubber [58–61]. No bulk polyacrylates have so far been produced by a continuous process.

Equipment. Figure 1 shows a production plant for batch and semibatch operation. Water, monomers, emulsifiers, and additives are contained in storage tanks (a) and fed via metering devices (b) to the mixing tank (c), in which the monomer emulsion is produced. The mixing tank is made of stainless steel (V2A, V4A) and equipped with a stirrer and cooling jacket. The shape of the stirrer and stirring rate are chosen to ensure thorough emulsification of the monomers. The mixing tank is, however, unnecessary if the individual components are accurately metered and an in-line mixer is used to prepare the monomer emulsion, or a monomer feed procedure is employed.

The monomer emulsion is usually added to the polymerization reaction batchwise or continuously over a period of minutes or hours. The reactor may be connected to further feed vessels for initiator and other additives, and usually has connections for flushing with inert gas. The internal reactor temperature can be regulated by wall cooling or evaporative cooling.

Vessels, stirrers, and all parts that come into contact with the product are made of stainless steel. The interior wall of the reactor should be as smooth as possible to prevent the deposition of wall scale. The polymerization reactor may have a capacity of $30\,m^3$ or more. The design, size, and power of the stirrer depend on the batch size and viscosity of the dispersion. High stirring rates (shear rate) may lead to coagulation, while low stirring rates may result in insufficient mixing.

After the end of the feed stage the dispersion is maintained for a further 1 – 2 h at the reaction temperature to reduce the residual monomer content before being cooled. The dispersion is discharged through a filter (h) and pumped to a storage tank (i). In the storage tank the solids content and pH value are often adjusted and formulations are prepared. Stabilizers and microbiocides are also added.

The dispersions are transported to the consumer or customer in polyethylene-lined metal drums, polyethylene vessels, or tank cars. The purchase of dispersions in barrels is being superseded by delivery in large-capacity containers for economic reasons.

3.2. Solution Polymerization

See also → Polymerization Processes, Chap. 3.3.6.1.4.. Acrylates are polymerized as solutions in organic solvents [25], [62–64] if the user wishes to exploit specific properties of polymers in dissolved form (e.g., low molecular mass, good flow behavior, and homogeneous film formation after drying in paints or adhesives). As a result of growing environmental awareness (solvent disposal), solution polymerization of acrylates will probably not expand to the anticipated level.

Aromatic hydrocarbons such as benzene and toluene can be used as solvents for the polymerization of acrylates of long-chain alcohols; esters and ketones can be used for acrylates of short-chain alcohols.

If the solvent boils at the temperature used for polymerization, a large proportion of the heat of polymerization can be removed by evaporative cooling.

Solvents may act as modifiers in solution polymerization. This should be kept in mind when choosing the solvent [65–71]. Some transfer constants of solvents are summarized in Table 5. The lower the transfer constant, the higher will be the molecular mass of the polymer. The viscosity of the solution increases with the molecular mass of the polymer. This has an adverse effect on the mixing of the reactor contents and the removal of the heat of polymerization [72]; it its therefore essential to choose an optimum polymer concentration for a given molecular mass.

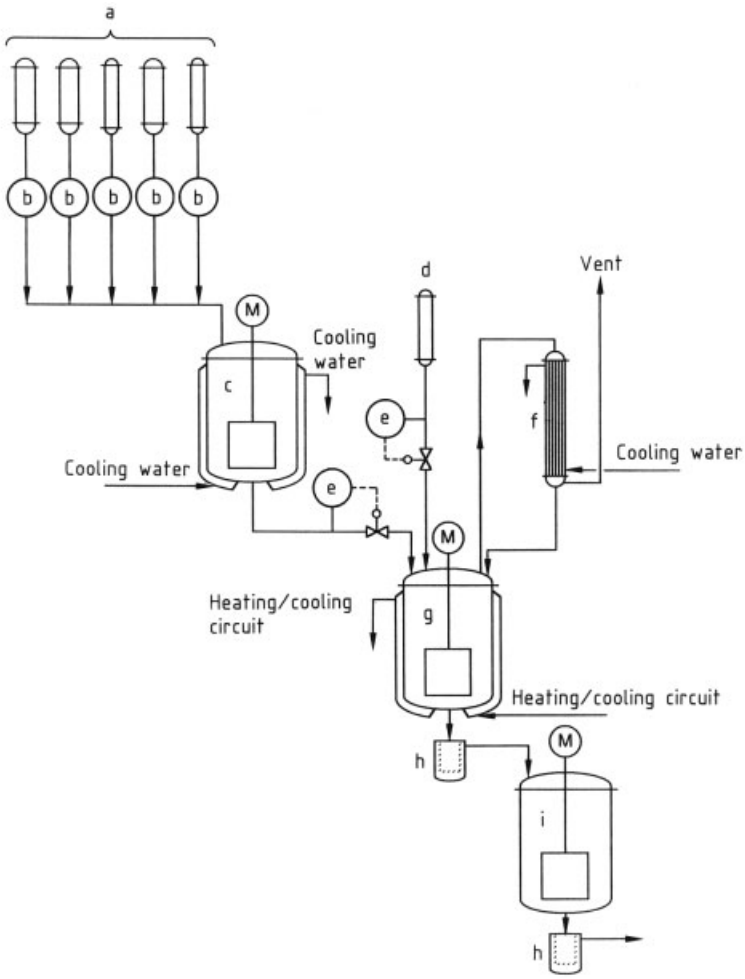

Figure 1. Production of polyacrylates by emulsion and solution polymerization
a) Storage tanks for monomers, distilled water, emulsifiers; b) Flow meter; c) Mixing tank; d) Storage tank for initiator; e) Flow registration control; f) Cooler; g) Polymerization reactor; h) Filter; i) Product storage tank

The molecular mass can be controlled by adding modifiers (e.g., halogenated aliphatic hydrocarbons and thiols [73]) or by adjusting the initiator concentration (increasing initiator concentrations lower the molecular mass).

Soluble azo compounds, peroxides, or hydroperoxides are used as initiators at a concentration of 0.01 – 2 wt % relative to the monomer. Some common peroxides and azo compounds together with their half-lifes are given in Table 6 [74], [75].

An initiator should be chosen with a half-life that guarantees a sufficient level of initiator throughout the feed time. Radical formation may be influenced by the solvent [76]. Azo compounds, e.g., 2,2′-azobis(isobutyronitrile), are often used as initiators in polymerization in organic solvents; 0.01 – 2 % of initiator is used depending on the desired product properties.

Solution polymerization can also be initiated with UV light (λ<360 nm) or γ-radiation (see Section 5.1). However, gels are often formed due to cross-linking reactions.

As with emulsion polymerization, solution polymers can be produced by batch or semibatch processes (see Section 3.1.2).

Semibatch processes can be efficiently controlled as regards temperature. Part of the sol-

Table 5. Transfer constants of various solvents

Monomer	Solvent	$k_{tr} \times 10^4$	Temperature, °C	Reference
Methyl acrylate	acetone	0.622	80	[65]
	benzene	0.326	80	[65]
	2-butanone	3.238	80	[65]
	n-butanol	2.747	80	[65]
	carbon tetrachloride	1.25	80	[66]
	chloroform	2.10	80	[67]
	cyclohexane	0.027	80	[65]
	2-butanol	2.496	80	[65]
	toluene	1.775	80	[65]
Ethyl acrylate	acetone	1.10	80	[68]
	benzene	0.525	80	[68]
	n-butanol	5.85	80	[68]
	sec-butanol	22.20	80	[68]
	carbon tetrachloride	1.55	80	[68]
	chloroform	1.49	80	[68]
	cyclohexane	1.22	80	[68]
	ethyl acetate	0.89	80	[68]
	hexane	0.97	80	[68]
	2-propanol	21		[69]
	methanol	0.32	60	[70]
	toluene	2.60	80	[68]
Butyl acrylate	ethanol	4.28	80	[71]
	methanol	0.47	80	[71]
	propanol	3.78	80	[71]
	2-propanol	14.12	80	[71]
2-Ethylhexyl acrylate	toluene	2.13	80	[20]

Table 6. Half-lives of organic peroxides and azo compounds used as initiators (0.1 mol/L in benzene)

Compound	Temperature for a half-life of 1 h, °C	Half-life, h					
		20 °C	40 °C	60 °C	80 °C	100 °C	120 °C
Acetylcyclohexanesulfonyl peroxide	46	57	2.4	0.1			
Diisopropylperoxy dicarbonate	61	400	18	1.2			
tert-Butylperoxy pivalate	74		92	5.7	0.5		
Dilauroyl peroxide	80			12	1.0		
Diacetyl peroxide	87			35	2.4		
Dibenzoyl peroxide	91			48	3.8	0.4	
tert-Butylperoxy 2-ethylhexanoate	92				6	0.4	
tert-Butylperoxy isobutyrate	96				6.7	0.6	
tert-Butylperoxy benzoate	124					18	1.6
Dicumyl peroxide	136						5.7
Di-*tert*-butyl peroxide	146						18
tert-Butyl hydroperoxide	200						
2,2′-Azobis(2,4-dimethylvaleronitrile)	64		25	1.8	0.17	0.021	
Azobis(isobutyronitrile)	82			15	1.2	0.12	

vent and monomer mixture (or all of the solvent with part or all of the initiator) is charged to the reactor. The remaining monomer – solvent mixture (or monomers) is added continuously over the course of 1 – 4 h. The monomers can also be added in portions. The batch is fully polymerized after 2 – 24 h.

The apparatus used in solution polymerization is similar to that for emulsion polymerization (Fig. 1). The monomers and solvents are fed into a stirred mixing tank or directly into a polymerization reactor via an in-line mixer.

In addition to the main feed line, the polymerization reactor has inlets for the initiator and modifier solutions. It is equipped with a heat-

ing/cooling jacket and an evaporative cooler. Most of the heat is generally removed by evaporative cooling. The initial batch is first heated by the heating jacket. Cooling is generally only necessary to control the reaction in the event of disorders.

The polymer solution should not be subjected to large temperature changes during storage. At elevated temperature there is a danger of solvent loss, while at low temperature the viscosity rises sharply and the polymer solution can no longer be pumped. Storage facilities (e.g., tank cars) should therefore be temperature controlled.

3.3. Elimination of Residual Monomers

Although polymerization proceeds almost quantitatively, residual monomers sometimes have to be removed.

After emulsion polymerization is complete, the dispersion is heated for a short period at 95 – 98 °C, if necessary under reduced pressure, and the residual monomers are expelled. Foaming can be suppressed by adding antifoaming agents.

Continuous removal of the residual monomers is, however, more effective. Steam is blown into the dispersion, which is under vacuum and begins to foam strongly. The rising dispersion foam is destroyed mechanically and the dispersion returned to the reactor; the monomers are condensed in a cooling device [77]. This process is used on an industrial scale.

In some cases the residual monomer content is lowered by adding small amounts of suitable initiators at the end of the polymerization.

3.4. Toxicology and Environmental Aspects

Polyacrylates, being high molecular mass compounds, are not toxic. Films and coatings produced from acrylate dispersions are permitted for food packaging in Germany [Recommendation XIV Plastics Dispersions (Empfehlung XIV Kunststoff Dispersionen) of the Federal Ministry of Health] and in the United States (21 CFR 176.189 Components of paper and paperboard in contact with dry food, FDA). Auxiliary substances also have to be taken into consideration in food legislation. The toxicology of polyacrylates is determined not by the polymers themselves, but by their auxiliaries (e.g., initiator residues, emulsifiers) and residual monomer content. Table 7 contains toxicity data for the most important acrylate esters. The residual monomer content of polymer dispersions is usually reduced to less than 300 ppm by the methods described in Section 3.3.

With solution polymers there are additional difficulties as regards toxicity, safety, and environmental pollution due to the use of organic solvents. This can be improved by choosing a suitable solvent. In most cases, however, solution polymers are not as environmentally safe as aqueous polymer dispersions; they are treated as combustible mixtures if the solvents are flammable.

4. Properties

4.1. Polymer Solutions and Dispersions

4.1.1. Polymer Solutions

The flow properties of polymer solutions and their formulations are of great importance as regards processability. The temperature dependence of these properties determines the temperature range in which the solution polymer can be handled.

The viscosity of polymer solutions depends on the molecular mass, molecular structure, temperature, and thermodynamic behavior of the solvent [72]. Information on polymer molecular mass, coil dimensions, interaction between solvents and polymers, and the thermodynamic behavior of a solvent are usually obtained from measurements on dilute solutions (light scattering, osmometry, sedimentation, viscosimetry), see also → Plastics, Analysis.

Interactions between the polymer molecules increase with rising concentration. Temporary polymer – polymer contacts are formed that finally lead to the formation of undesirable gels that do not flow and cannot be processed.

4.1.2. Polymer Dispersions

Dispersions consist of two phases, and are therefore more complicated systems than polymer so-

Table 7. Toxicity of the most important acrylate esters

Parameter	Methyl acrylate	Ethyl acrylate	Butyl acrylate	2-Ethylhexyl acrylate
OSHA				
PEL [a] (skin)	10 ppm or 35 mg/m^3	5 ppm or 20 mg/m^3	10 ppm or 55 mg/m^3	not regulated
STEL [b] (skin)		25 ppm or 100 mg/m^3		
Irritation data [78]				
rabbit eye	100 mg/24 h, moderate	1204 ppm/7 h	500 mg/24 h, mild	500 mg/24 h, mild
rabbit skin	10 mg/24 h open, mild	500 mg/open, mild	500 mg/open, mild	500 mg/open, mild
Toxicity data [78]				
LD_{50} rat, oral, mg/kg	277	800	900	5660
LC_{50} rat, inhalation, 4 h, ppm	1350	2180	2730	
LD_{50} rabbit, dermal, mg/kg	1243	1834	2000	8480
MAK	5 ppm or 18 mg/m^3	5 ppm or 20 mg/m^3	10 ppm or 55 mg/m^3	
IARC Group [c]	3	2	3	

[a] Permissible environmental limit (8 h TWA).
[b] Short term exposure limit (15 min TWA).
[c] International Agency for Research on Cancer (1987).

lutions; they often also contain polyelectrolytes. Almost all investigations are carried out on dilute systems.

The most important properties of polymer dispersions are appearance, concentration, pH, surface tension [79], residual monomer content, particle size, particle size distribution, viscosity, and stability.

Residual Monomer Content. Residual monomers are determined by chromatography. The volatile monomers can be quantitatively determined by gas chromatography using calibration and addition of an internal standard (e.g., toluene or methyl ethyl ketone [80]). An accuracy of ca. 1 ppm can be achieved.

The content of water-soluble, nonvolatile monomers (e.g., acrylic acid) is determined by high-pressure liquid chromatography [81].

Particle Size and Particle Size Distribution. The particle size and particle size distribution of emulsion polymers influence other properties (e.g., the viscosity of dispersions with a high solids content). At the same concentration and with the same auxiliary system, coarse dispersions generally have a lower viscosity than fine dispersions due to electrostatic effects or, in the case of dispersions with protective colloids, to steric effects.

Particle size and particle size distribution can be determined by *electron microscopy* (measurement range: 1 – 1000 nm) [82].

A simple method for determining the mean particle size is to measure the *light transmission* of a highly dilute sample. The relationship between the turbidity and particle radius of spheres is described by Mie's theory [83]. The particle diameter is shown as a function of the light transmission in Figure 2. An important parameter is m, the ratio of the refractive index of the polymer to that of water. The approximate refractive index of copolymers can be calculated with the following formula:

$$n = w_1 n_1 + w_2 n_2 + w_3 n_3 + ..$$

where the mass fractions $w_1 + w_2 + w_3 .. . = 1$ and n_1, n_2 are the refractive indices of the homopolymers.

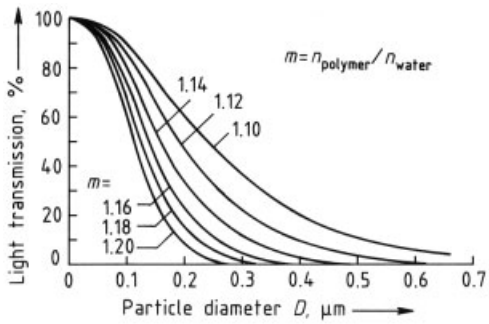

Figure 2. Relationship between light transmission and particle size for 0.01 wt % polymer dispersion (path length 2.5 cm, λ 546 nm)

A mean particle size can also be determined from *static light scattering* (measurement

range: 300 – 1500 nm). Hydrodynamic diameters can be accurately measured by *quasielastic light scattering* (QELS); measurement range: 10 – 3000 nm) [84–88].

The particle size distribution can be determined with the *ultracentrifuge* (measuring range: 10 – 20 000 nm in a gravity field of 20 – 200 000 g) [89]. The particles must be spherical.

The surface areas of latex (dispersion) particles are determined by *soap titration*, from which the particle size can be calculated [90–93]. Soap titration is simple and quick to perform, and yields reasonable values with homologous polymer dispersion series. Care should be taken, however, when evaluating latex surfaces of different hydrophilicity.

Viscosity. Dispersions with solids contents of 40 – 60 wt % exhibit a broad spectrum of rheological behavior [94], [95] that depends on the volume fraction, temperature, particle size, particle size distribution, and auxiliary system. Dispersions may be highly liquid to pasty, and may exhibit a Newtonian to viscoelastic flow behavior [96]. In dispersions structural viscosity ("shear thinning") is frequently observed; dilatancy is observed less often; thixotropy and rheopexy are rarely encountered.

Stability. Dispersions must have a certain degree of mechanical (shear) stability if coagulation is not to occur during pumping, stirring, or spraying. Mechanical stability is checked by determining the amount of coagulate after a stirring test; the stirrer, temperature, shear velocity, time, and other parameters should correspond to possible stresses occurring during use.

Thermal stability is checked by subjecting the dispersions to the anticipated service temperature conditions. A freeze – thaw test with five cycles is widely used. However, the results depend on the container size and temperature control so that the correlation between laboratory results and frost stability under relevant weathering conditions is generally unsatisfactory.

The polymer dispersion must be stable to electrolytes when diluted with hard water and when formulating products such as paints or adhesives. The stability is therefore tested by adding salts.

Minimum Film-Forming Temperature. Whether a dispersion forms a coherent film after evaporation of the water or whether a brittle, pulverizable layer is formed depends on the glass transition temperature of the polymer, residual water content, and temperature at which the film is formed. The lowest temperature at which a coherent film can still form from the copolymer dispersion is called the minimum film-forming temperature (MFT). The MFT is measured with a metal plate that has a temperature gradient [97], [98]. The temperature at which the film begins to crack or the so-called white point, at which the turbid film starts to become clear, is measured [99].

The MFT depends on the copolymer composition, as well as the particle size and polarity of the comonomers [100–102]. It is usually comparable to the glass transition temperature, but may be substantially lower, probably because the emulsifier and water act as a plasticizer.

4.2. Properties of the Polymers

Glass Transition Temperature. The glass transition temperatures (T_g) of polyacrylates and homopolymers of monomers that are frequently used in acrylate copolymers are summarized in Table 8. The values were determined by differential thermoanalysis on dispersion films (heating rate 20 °C/min, midpoint), the films being repeatedly heated until reproducible values were obtained. Other factors that influence the glass transition temperature (e.g., molecular mass, [103], [104]) are discussed elsewhere [105], [106].

The glass transition temperature of most copolymers can be accurately calculated with the nonlinear equation of Gordon and Taylor [107] given below. Other nonlinear equations are discussed elsewhere [108].

$$T_g = \frac{T_{gA} w_A + c T_{gB} w_B}{w_A + c w_B}$$

where w_A and w_B are the mass fractions of the monomers A and B, and T_{gA} and T_{gB} are the absolute glass transition temperatures of the corresponding homopolymers ($T_{gA} < T_{gB}$). The constant c takes into account the expansion coefficients of the melt and glass state of the two

Table 8. Properties of homopolymers

Monomer	CAS registry no. of polymer	T_g [a], °C	d_4^{25}	n_D^{25}
Methyl acrylate	[*9003-21-8*]	22	1.20	1.479
Ethyl acrylate	[*9003-32-1*]	− 8	1.13	1.464
n-Propyl acrylate	[*24979-82-6*]	−25	1.10	1.462 [b]
n-Butyl acrylate	[*9003-49-0*]	−43	1.06	1.474
Isobutyl acrylate	[*26335-74-0*]	−17	1.05	1.464 [b]
sec-Butyl acrylate	[*30347-35-4*]	− 6	1.06	1.462 [b]
tert-Butyl acrylate	[*25232-27-3*]	55	1.03	1.468
n-Hexyl acrylate	[*27103-47-5*]	−51	0.98	1.468 [b]
2-Ethylhexyl acrylate	[*9003-77-4*]	−58	0.99	1.433 [b]
Lauryl acrylate	[*26246-92-4*]	−17	0.99 [c]	1.468 [b]
Methyl methacrylate	[*9011-14-7*]	105	1.19	1.488
Ethyl methacrylate	[*9003-42-3*]	67	1.12	1.483
n-Butyl methacrylate	[*9003-63-8*]	32	1.06	1.483
Isobutyl methacrylate	[*9011-15-8*]	64	1.05	1.475
tert-Butyl methacrylate	[*25189-00-8*]	102	1.02	1.460
n-Hexyl methacrylate	[*25087-17-6*]	−14	1.01	1.479
Acrylonitrile	[*25014-41-9*]	105	1.18	1.519
Methacrylonitrile	[*25067-61-2*]	112	1.18	1.545 [b]
Styrene	[*9003-53-6*]	107	1.05 [d]	1.591 [d]
trans-1,4-Butadiene	[*40022-02-4*]	−107	0.97	1.518
cis-1,4-Butadiene	[*40022-03-5*]	−102	1.01	1.52
Vinyl acetate	[*9003-20-7*]	42	1.19	1.463
Vinyl propionate	[*25035-84-1*]	8	1.14	1.465 [b]
Vinyl chloride	[*9002-86-2*]	77 [e]	1.40	1.545
Vinylidene chloride	[*9002-85-1*]	−18 [f]	1.84 [g]	1.618 [b, g]
Acrylic acid	[*9003-01-4*]	130 [h]	1.54 [c]	1.544 [b]
Methacrylic acid	[*25087-26-7*]	162 [h]	1.45 [c]	1.563 [b]
Acrylamide	[*9003-05-8*]	220 [h]	1.44 [c]	1.603 [b]
Methacrylamide	[*25014-12-4*]	243 [h]	1.40 [c]	1.616 [b]

[a] Emulsion polymers or aqueous solution polymers, differential scanning calorimetry, 20 °C/min, midpoint. [b] From the refractive increment. [c] From the specific volume. [d] Atactic. [e] Polymerization temperature 90 °C. [f] Amorphous. [g] Crystalline. [h] After repeated heating.

homopolymers. Its value is generally between 0.5 and 2.0.

With the exception of *tert*-butyl acrylate, most acrylate homopolymers have extremely low glass transition temperatures. They are therefore too soft, too tacky, or have too high an elongation and insufficient strength for many areas of application. Polymers for specific applications can be produced by copolymerization with monomers whose homopolymers have high glass transition temperatures (e.g., styrene, acrylonitrile, methyl methacrylate, ethyl methacrylate).

Molecular Mass. See also → Plastics, Analysis, Chap. 5. The molecular mass can be calculated from the intrinsic viscosity $[\eta]$ and a knowledge of the two constants K and a:

$$[\eta] = K \cdot M^a$$

Values for K and a in various solvents are listed in Table 9.

Molecular mass determination does not generally present any difficulties with solution polymers. With acrylate emulsion polymers, however, cross-linking of the dispersed particles may cause problems. In this case it is impossible to produce a molecular-disperse solution. Before the molecular mass is determined it must therefore be checked whether a true dilute solution exists.

Mechanical Properties of Polymer Films. The temperature dependence of mechanical moduli [112], [113] provides detailed information about the mechanical behavior and structure of polymer films. Stress – strain diagrams

Table 9. Determination of the molecular mass according to the Staudinger equation (intrinsic viscosity in cm^3/g)

Polymer	Tetrahydrofuran, 25 °C [109]		Acetone, 20 °C		Chloroform, 20 °C [110]	
	$K \times 10^3$, cm^3/g	a	$K \times 10^3$, cm^3/g	a	$K \times 10^3$, cm^3/g	a
Poly(methyl acrylate)	10.0	0.73	5.6 [110]	0.75	3.0	0.85
Poly(ethyl acrylate)	8.9	0.75	8.8 [110]	0.72	7.3	0.78
Poly(*n*-propyl acrylate)	11.9	0.72				
Poly(*n*-butyl acrylate	7.4	0.75	5.2 [110]	0.73	13.1	0.69
Poly(isobutyl acrylate)	12.7	0.71				
Poly(*tert*-butyl acrylate)	43.4	0.60	4.7 (25 °C) [111]	0.75		
Poly(*n*-hexyl acrylate)	5.5	0.76				
Poly(2-ethylhexyl acrylate)	11.1	0.68				
Poly(*n*-lauryl acrylate)	27.3	0.58				

of polymer films [114] are plotted with conventional apparatus; rupture strength and elongation at break are thus obtained. The values exhibit a relatively high degree of scatter with dispersion films, however, due to film defects. It is therefore often more appropriate to use the force required for a specific elongation, or the elongation that occurs when a defined force is applied as a measure of the stress – strain behavior. Mechanical moduli can be accurately measured by performing tensile tests with small extensions.

Poly(methyl acrylate) films have a relatively low elongation at break and an extremely high rupture strength at room temperature. The elongation of poly(ethyl acrylate) films is greater and the rupture strength is considerably lower. Poly(ethyl acrylate) is elastic and almost rubberlike. The elongation of poly(butyl acrylate) is extremely high and the rupture strength very low; the films are not elastic, but exhibit viscous flow even at small elongations and cannot be freely handled. Poly(ethylhexyl acrylate) has an even higher elongation and the films have no strength.

Tackiness. Poly(methyl acrylate) is not tacky at room temperature, poly(ethyl acrylate) is only slightly tacky. Poly(butyl acrylate) and poly(ethylhexyl acrylate) are, however, extremely tacky.

Tackiness depends on the molecular mass: tacky films can be produced from poly(methyl acrylate) with a very low molecular mass, whereas tack-free films can be produced from poly(butyl acrylate) of very high molecular mass. Increasing chain length of the alcohol residue ($>C_8$) causes the tackiness to fall. The polyacrylates become increasingly "dry" and elongation at break decreases due to crystallization of the *n*-alkyl chain. For example, poly(*n*-hexadecyl acrylate) is waxy and rigid at room temperature.

Electrical and Optical Properties. The electrical and optical properties of acrylate homo- and copolymers are similar to those of the polymethacrylates. Polyacrylates hardly yellow under direct sunlight, because they only absorb UV radition below 290 nm. If UV stabilizers are added, their properties do not change at all under the action of light. This is an important advantage of polyacrylates.

Solubility. Polyacrylates with short side chains are relatively polar and thus soluble in polar solvents, aromatic hydrocarbons, and chlorinated hydrocarbons. Common solvents include

tetrahydrofuran, dimethylformamide, acetone, butanone, ethyl acetate, and chloroform. Precipitating agents include water, methanol, ethanol, diethyl ether, and aliphatic hydrocarbons.

The solubility parameters for different polyacrylates calculated according to SMALL [115], [116] vary only slightly. The values for polymethyl, polyethyl, and polybutyl acrylate and for the corresponding polymethacrylates are $21.3-18\,(\mathrm{MPa})^{0.5}$ and $10.4-8.8\,(\mathrm{cal/cm^3})^{0.5}$, respectively. Solubility in a certain solvent can be expected if the solubility parameters of the solvent and substance to be dissolved are comparable [117].

Chemical Stability. Polyacrylates are extremely resistant to oxygen, and only decompose very slowly under extreme conditions such as high temperature and in an oxygen-rich atmosphere. When heated, polyacrylates depolymerize to monomers much less readily than the corresponding polymethacrylates. Poly(methyl acrylate) decomposes at 292 – 399 °C to produce mainly methanol and carbon dioxide [118]. Volatile decomposition products of poly(ethyl acrylate) at 300 – 500 °C are ethylene, ethanol, and carbon dioxide. Butene, butanol, and carbon dioxide are formed from poly(butyl acrylate) in the same temperature range. Volatile decomposition products of poly(2-ethylhexyl acrylate) are 2-ethylhexene, ethylhexanol, and carbon dioxide [119]. Polyacrylates are largely resistant to acid and alkaline hydrolysis. The longer the alkyl radical, the greater the resistance to hydrolysis [120].

Cross-Linking of Polyacrylate Films. Solubility and tackiness are reduced by cross-linking, the rupture strength increases, whereas the elongation at break decreases. Emulsion polymers and, in special cases, solution polymers are cross-linked by incorporation (copolymerization) of monomers with groups that react with one another under certain conditions (e.g., change in pH or temperature) or in the presence of additives. Such monomers include *N*-methylolacrylamide, *N*-methylolmethacrylamide, and the corresponding ethers. The methylol groups react in the acid pH range at elevated temperature to form methylene bisacrylamide linkages with elimination of formaldehyde. Reviews of cross-linking systems are given in [121–128]. Environmentally friendly systems are preferred that cross-link at room temperature and do not lead to the release of toxic substances such as formaldehyde [128], [129]; an example is cross-linking of acrylic acid via formation of metal salts.

Cross-linking and cross-linking density can be determined by measuring the solvent uptake of the polyacrylate film. Suitable solvents are tetrahydrofuran or dimethylformamide, in which the noncross-linked polyacrylates are completely soluble. The relationship between swelling and cross-linking density can be derived from a model used to describe rubber elasticity [130–134]. Measurement of the cross-linking density from the temperature dependence of the modulus of elasticity [135–138] or modulus of torsion is somewhat more time-consuming.

5. Uses

Many important physical properties of polyacrylates and their copolymers can be varied within wide limits by a suitable choice of comonomers, auxiliaries, and polymerization process to give a large number of products for a wide variety of applications. The most important areas of application are concerned with protection, binding, and adhesion.

5.1. Paints and Coatings

See also → Paints and Coatings, Chap. 2.5.

Architectural Paints. Good pigmentability, processability, and film properties are important for the use of a polymer dispersion as a paint binder [139]. Vinyl ester copolymer dispersions mostly contain protective colloids, are coarsely particulate, and have favorable rheological properties for coating materials. However, vinyl acetate hydrolyzes on strongly basic substrates and the high water absorption of the dispersion films is also a disadvantage [140].

Acrylate copolymer dispersions usually contain emulsifiers and are finely particulate; they can therefore be highly extended with pigments [141], [142], but auxiliaries are necessary to adjust rheology. The low water absorption of the

dispersion films is also an advantage. Emulsion paints based on these dispersions are stable to hydrolysis and are largely weather-resistant [143–145].

Copolymers with very high styrene contents tend to yellow in UV light and are therefore only used for interior paints. Coats with a matt, silk, or high-gloss finish can be formulated. Acrylate ester – methacrylate ester copolymers ("pure acrylates") have outstanding light stability, even in UV light, and are resistant to hydrolysis. They are therefore commonly used as binders for transparent and low-pigment coatings [146]. Specially selected pure acrylate dispersions with wet adhesion on alkyd substrates are used in gloss paints.

Vinyl ester, vinyl ester – acrylate ester, and acrylate ester – methacrylate ester copolymers are widely available in the United States. Styrene acrylics and pure acrylates are becoming increasingly important in Europe.
Trade names include Acronal, Propiofan, Luhydran (BASF); Uramul (DSM); Crilat (Montedison); Mowilith (Hoechst); Walpol (Reichhold Chemie); Primal, Rhoplex (Rohm and Haas); Revacryl (Revertex); Plextol (Röhm); Rhodopas (Rhône-Poulenc); Synthacryl (Synthopol); Ucecryl (UCB); Vinacryl and Vinamul (Vinyl Products); Vinnapas (Wacker).

Coatings and Lacquers. Metals are often coated with a waterborne priming coat and a solvent-containing topcoat. The trend is toward topcoats with high polymer concentrations to reduce solvent emissions. In cases where stringent requirements do not have to be satisfied as regards weather resistance and corrosion protection, the trend is toward waterborne topcoats [147–151].

In the automobile industry three coats are generally applied:

1) A waterborne priming coat, which is mainly responsible for corrosion protection
2) An intermediate coat (filler) which compensates for any unevenness of the substrate
3) A pigmented topcoat

In metallic-effect coatings the topcoat consists of an aluminum-pigmented, colored metallic basecoat and a clearcoat.

The primers may be waterborne polyacrylate systems. Waterborne or low-solvent systems are increasingly used in fillers and metallic basecoats. Acrylic resins with hydroxyl groups, which are reacted with melamine resins or isocyanates, are used in the clearcoat; solution polymers offer advantages as regards application, gloss, and weather resistance.
Trade names include Acronal, Larodur, Luhydran, Lumitol, Luprenal (BASF); Desmophen A (Bayer); Synocryl (Cray Valley); Degalan (Degussa); Elvacite, Lucite (Du Pont); Macrynal, Synthacryl (Hoechst); Joncryl (Johnson); Neocryl (Polyvinyl Chemie); Acryloid, Paraloid (Rohm and Haas); Plexigum (Röhm); Setalux (Synthex); Synthalat (Synthopol Chemie).

Radiation-Curable Systems. Coating methods in which monomers and oligomers are cured (hardened) by exposure to radiation (usually UV light but also electron beams) are experiencing considerable growth (see → Paints and Coatings, Chap. 3.7.). Advantages of these methods include low emission of volatile organic compounds (VOCs), energy savings, high production rates, and excellent gloss of the coating. However, radiation curing can only be employed on flat surfaces, and cannot therefore be used in the automobile industry. Unsaturated polyester resins (mainly acrylate and methacrylate esters of epoxides), polyesters, oligoethers, urethanes, and polyhydric alcohols are used as binders. Reactive species (radicals) are formed after exposure of photoinitiators to UV light [152–158]. In electron-beam curing, free radicals are generated by interaction of accelerated electrons with organic compounds.

The main areas of use include the coating of wood (doors, furniture, parquet floors), paper (printing lacquers for record sleeves, postcards), and plastics (PVC coverings).

5.2. Paper Industry [159]

High-quality papers are coated with pigments to improve their printability, appearance, gloss, and other properties [160], [161]. Natural products (e.g., starch and casein) were first used as binders. One of the first acrylate-based dispersions, Acronal 500 D, was introduced on the market in the 1950s. Acrylate-based substitution products (cobinders) also became available for casein and starch [162]. The first fully synthetic

binders that could be used without cobinders were introduced in the 1960s [163]. Various dispersions with different polymer hardnesses are available for different types of paper and printing methods [164]; see also → Paper and Pulp, Chap. 1.1. Vinyl esters, acrylonitrile, or styrene are used as comonomers for acrylates; styrene gives good adhesion.

In the United States, acrylate dispersions are of minor importance as binders in coating compositions for paper, butadiene copolymer dispersions are used instead. These products are becoming increasingly important in Europe.

Polyacrylate cationic dispersions [165], and dispersions containing protective colloids are used as sizing agents in paper production.
Trade names include Acronal, Acrosol, Basoplast, Styronal (BASF); Primal, Rhoplex (Rohm and Haas).

5.3. Adhesives and Sealing Compounds [166–169]

Laminating adhesives, pressure-sensitive adhesives, building construction adhesives, and sealing compounds are produced from polyacrylates; see also → Adhesives. Homopolymers and copolymers with a low glass transition temperature (usually based on 2-ethylhexyl acrylate and butyl acrylate) are used to obtain the desired tackiness. The glass transition temperature can be adapted for individual applications by copolymerization with acrylonitrile, styrene, vinyl acetate, or methyl methacrylate. Products with a balanced adhesion – cohesion ratio can thus be obtained.

Tackiness is extremely important in emulsion adhesives and adhesives based on solution polymers. The higher the glass transition temperature, the lower has the molecular mass to be in order to obtain the necessary tackiness. The higher the molecular mass, the better is the cohesion of the film. Cohesion often has to be improved by additional cross-linking. Depending on the molecular mass, solution polymer concentrations of up to 50 % can be produced. Solvent naphtha, aromatic compounds, ethyl acetate, and alcohols are used as solvents. However, solution polymers are being replaced by aqueous dispersions.

Dispersions and solution polymers are used as industrial laminating adhesives (e.g., high-gloss foil laminating, automobile industry). Laminating adhesives for food and cosmetic packaging must be solvent free. Good adhesion to the substrate is important in this area of application, and can be achieved by using monomers with reactive groups (e.g., carboxyl or amide) or special additives.

Polyacrylates are replacing natural products in pressure-sensitive adhesives (e.g., label adhesives, pressure-sensitive sheets, adhesive tapes, adhesive plasters); glyoxal derivatives are used as cross-linking agents (→ Glyoxal, Chap. 5.). The advantages of polyacrylates compared with natural rubber, for example, are their better resistance to aging and temperature. Both solution polymers and dispersions are used [170–172]; solution polymers have a better water resistance than dispersion adhesives [173].

Structural adhesives (adhesives for floor coverings and tiles) are mainly dispersions [174]. Adhesives for floor coverings are formulated with tack-imparting resins and incorporate organic solvents. Developments in this field include the use of solvent-free systems, i.e., addition of solvent-free resins (e.g., aqueous tackifier dispersions). Tile adhesives are solvent free and are generally mixed with cement; dry powder adhesives will be of particular interest in the future.

Sealing compounds are based on dispersions and must be tack-free in the dry state. One-component polyacrylate systems offer economic advantages compared with silicones. Disadvantages are their water sensitivity and poor adhesion to glass and ceramics.
Trade names include Coldac (Allied Colloids); Airflex, Flexbond, Flexcryl (Air Products); Acronal, Lutonal, Propiofan (BASF); Lacstar, Voncoat (Dai Nippon); Emultex (Harco); Litex (Hüls); Revertex (Revertex); Rhodopas (Rhône – Poulenc); Texicryl (Scott Bader); Ucecryl, Ucefix (UCB); Vinnapas (Wacker).

5.4. Textile Industry

In the textile industry polyacrylates are used almost exclusively as emulsion polymers (e.g., as binders for pigment dyeing or printing) [175–180]. In almost all areas of application apart from bonding and laminating, soft, nontacky films are desired. The treated textiles generally

have to be resistant to washing and drycleaning; this is achieved by cross-linking the polyacrylates after film formation. Drycleaning resistance can be improved by copolymerization with polar monomers such as acrylonitrile, but the handle and feel are then hard.

Polyacrylates are also used as binders for textile flocking to produce velvety surfaces. In addition, fabrics may be coated with water-resistant dispersions (e.g., wax cloth and umbrella material).

Polyacrylates are of minor importance in high-grade finishes [181] that are used to improve textile properties such as handle, feel, abrasion resistance, and crease resistance. A further field of application of polyacrylate dispersions is the consolidation of nonwoven fabrics [182–186] for uses such as interlinings [187], filter materials, and carrier materials for leather imitates [188]. Thermosensitive dispersions that coagulate at a specific temperature during the drying process are also used in the production of nonwovens (→ Nonwoven Fabrics, Chap. 2.2.). *Trade names* include Acronal, Helizarinbinder, Perapret (BASF); Acramin (Bayer); Imperon (Hoechst); Primal (Rohm and Haas).

5.5. Leather Industry [189], [190]

The surface of leather is treated to make it hydrophobic while retaining its natural character, in particular appearance. Highly reversible stretching and good cohesion of the dispersion film are important for the use properties; the leather finish must not crack or tear after repeated folding and bending and should have a low tackiness [191–194]. See also → Leather, Chap. 8.4., → Leather, Chap. 11.1. The first synthetic polymer dispersion for industrial use was "Corialgrund" developed by BASF [193].
Trade names include Corialgrund, Corial Binder (BASF); Euderm (Bayer); Hycar (Goodrich); Fondocryl (Henkel); Melio Resin A (K. J. Quinn); Revertex (Revertex); Primal (Rohm and Haas); Resin RA (Stahl Chemical Ind. B.V.).

5.6. Miscellaneous Uses

Polyacrylates, mainly emulsion polymers, are used in many minor areas: for example, as sizes in the production of textile and glass fibers [195–197]; as polishes for floor coverings, automobiles and shoes; and as additives for hydraulic binders. Polyacrylates are used as additives in mortar and concrete, as well as in agriculture for seed protection and spraying with fertilizers (→ Fertilizers, Chap. 7.3.). They are also employed as matrices for ion exchangers cross-linked with divinylbenzene (→ Ion Exchangers, Chap. 2.1.).

6. Economic Aspects

Detailed production figures for acrylates are only available for the United States [198], and not for Western Europe and Japan. According to estimates for 1982 and 1984, the United States produced ca. 50 %, Western Europe ca. 35 %, and Japan ca. 15 % of all acrylate esters. These figures have probably not changed substantially. Production figures for acrylate esters from 1969 to 1987 are given in Table 10. Total production has increased in the United States and presumably in Europe, whereas in Japan it has remained constant. The percentage proportions of methyl and 2-ethylhexyl acrylates have remained virtually unchanged. Butyl acrylate showed average annual growth rates of 7.3 % between 1969 and 1984, while ethyl acrylate experienced a downturn of 0.6 %. The consumption of acrylate esters for 1989 and various fields of application is given in Table 11 [198].

Most polyacrylates are produced by emulsion polymerization. In the United States the proportion of solution polymers is 12 %; relevant figures are not available for Europe. Although the proportion of solution polymers (which are largely used for coatings and adhesives) may well fall, they will not be completely replaced by emulsion polymers. The use of polyacrylates in the adhesives and paper industries is substantially higher in Europe than in the United States. Methyl acrylate is used almost exclusively in the polymerization of acrylonitrile for fibers to improve dyeability. The mean annual growth rate in the consumption of polyacrylates for 1984 – 1989 is estimated to be 3.3 % for the United States and 2.8 % for Europe [198]; specialty fields have the highest growth rates.

Table 10. Production of acrylates (10^3 t/a) [198]

Year	Methyl acrylate	Ethyl acrylate	*n*-Butyl acrylate	2-Ethylhexyl acrylate	Other acrylates	Total
United States						
1969	14	91	29	15	4	153
1975	20	109	81	15	9	234
1980	29	122	157	31	17	356
1984	28	139	193	39	27	426
1987		147	234			
Western Europe [a]						
1984	60 – 75	60 – 75	105	45 – 60	6	300
Japan [b]						
1975						67
1980						111
1984						106

[a] Chemical Economics Handbook (CEH) estimated values.
[b] Excluding Asahi Chemical.

Table 11. Estimated consumption of polyacrylates in 1989 expressed in 10^3 t/a (%) [198]

Area of use	United States	Western Europe	Japan [a]
Dispersions			
Surface coatings	124 (32)	103 (36)	38 (33.4)
Adhesives	27 (7)	48 (17)	
Textiles	68 (17.5)	33 (11)	37 (34.9)
Leather	7 (2)	33 (11)	37 (34.9)
Paper	18 (4.5)	49 (17)	
Polishes	13 (3)		
Miscellaneous uses	18 (4.5)		13 (11.9)
Other polymers			
Acrylonitrile fibers	14 (3.5)	26 (9)	15 (13.9)
Solution polymers for surface coatings	45 (12)	28 (10)	
Miscellaneous uses	53 (14)		7 (6.7)
Total	387 (100)	287 (100)	110 (100)

[a] 1984 excluding Asahi Chemicals.

7. References

General References

1. D. R. Bassett, A. E. Hamielec (eds.): "Emulsion Polymers and Emulsion Polymerization," *ACS Symp. Ser.* **165** (1981).
2. D. C. Blackley: *Emulsion Polymerisation, Theory and Practice,* Applied Sci. Publ., London 1975.
3. D. C. Blackley: *High Polymer Lactices. Their Science and Technology,* **vols. I and II,** Maclaren & Sons, London 1966.
4. F. A. Bovey, I. M. Kolthoff, A. I. Medalia, E. J. Meehan: "Emulsion Polymerization," in *High Polymers,* vol. IX, Interscience, New York 1955.
5. F. Hölscher: "Dispersionen synthetischer Hochpolymerer I: Eigenschaften, Herstellung und Prüfung," in *Chemie, Physik und Technologie der Kunststoffe,* vol. 13, Springer Verlag, Berlin 1969.
6. M. B. Horn: *Acrylic Resins,* Reinhold Publ., New York 1960.
7. I. Piirma, J. L. Gardon (eds.): "Emulsion Polymerization," *ACS Symp. Ser.* **24** (1976).
8. H. Rauch-Puntigam, T. Voelker: "Acryl- und Methacrylverbindungen," in *Chemie, Physik und Technologie der Kunststoffe,* vol. 9, Springer Verlag, Berlin 1967.
9. H. Reinhard: "Dispersionen synthetischer Hochpolymerer II: Anwendung," in *Chemie, Physik und Technologie der Kunststoffe,* vol. 14, Springer Verlag, Berlin 1969.
10. E. H. Riddle: *Monomeric Acrylic Esters,* Reinhold Publ., New York 1954.
11. H. Warson: *The Applications of Synthetic Resin Emulsions,* Ernest Benn, London 1972.

Specific References

12. H. v. Pechmann, O. Röhm, *Ber. Dtsch. Chem. Ges.* **34** (1901) 427.

13. Röhm & Haas, DE 571 123, 1928, US 1 829 208 (W. Bauer).
14. IG Farbenind., DE 615 219, 1931, GB 387 736 (E. Scharf, H. Fikentscher).
15. IG Farbenind., DE 654 989, 1930, GB 358 534 (H. Fikentscher, C. Heuck).
16. W. Reppe, *Justus Liebigs Ann. Chem.* **582** (1953) 1.
17. L. S. Luskin: "Vinyl and Diene Monomers," in C. Leonhard (ed.): *High Polymers,* vol. 24, part I, Wiley Interscience, New York 1970, p. 105.
18. *Acrylate Esters – A Guide to Safety and Handling* compiled by Badische Corporation, Celanese Chemical, Rohm and Haas, Union Carbide Corporation, Williamsburg. *The Handling and Storage of Acrylic Esters,* BASF, Ludwigshafen 1989.
19. Z. A. Sinitsyna, Kh. S. Bagdasaryan, *Zh. Fiz. Khim.* **32** (1958) 1319.
20. M. Rätzsch, I. Zschach, *Plaste und Kautsch.* **21** (1974) 345.
21. D. J. Liaw, K. C. Chung, *J. Chin. Inst. Chem. Eng.* **13** (1982) 145.
22. H. Spoor, *Angew. Makromol. Chem.* **4/5** (1968) 142.
23. H. Wesslau, *Makromol. Chem.* **69** (1963) 220.
24. H. Yuki, K. Hatada, K. Ohta, Y. Okamoto, *Appl. Polym. Symp.* **26** (1975) 39.
25. E. H. Riddle: *Monomeric Acrylic Esters,* Reinhold Publ., New York 1954.
26. E. Trommsdorff et al. in E. Müller (ed.): *Methoden der organischen Chemie,* **vol. XIV/1,** Thieme Verlag, Stuttgart 1961, p. 1010.
27. R. Z. Greenley in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* 3rd ed., Wiley-Interscience, New York 1989, p. II/153.
28. C. S. Marvel, R. Schwen, *J. Am. Chem. Soc.* **79** (1957) 6003.
29. C. Walling, J. A. Davison, *J. Am. Chem. Soc.* **73** (1951) 5736.
30. V. P. Zubov, L. I. Valuev, V. A. Kabanov, V. A. Kargin, *J. Polym. Sci. Polym. Chem.* **9** (1971) 833.
31. N. G. Kulkarni, N. Krishnamurti, P. C. Chatterjee, M. A. Sivasamban, *Makromol. Chem.* **139** (1970) 165.
32. F. R. Mayo, F. M. Lewis, C. Walling, *J. Am. Chem. Soc.* **70** (1948) 1529.
33. W. M. Ritchey, L. E. Ball, *J. Polym. Sci. Polym. Lett. Ed.* **4** (1966) 557.
34. G. Markert, *Makromol. Chem.* **103** (1967) 109.
35. A. Fehervari, T. Földes-Berezsnich, F. Tüdös, *J. Macromol. Sci. Chem.* **A18** (1982) 337.
36. B. G. Elgood, B. J. Sauntson, *Chem. Ind. (London)* 1965, 1558.
37. T. Tamikado, Y. Iwakura, *J. Polym. Sci.* **36** (1959) 529.
38. N. Grassie, B. J. D. Torrance, J. D. Fortune, J. D. Gemmell, *Polymer* **6** (1965) 653.
39. J. H. Bradbury, H. W. Melville, *Proc. Royal Soc. London Ser. A* **222** (1954) 456.
40. A. F. Nikolaev et al., *Vysokomol. Soedin Ser. A* **11** (1969) 2418.
41. G. V. Tkachenko, L. V. Stupen, L. P. Kofman, L. Z. Frolova, *Zh. Fiz. Khim.* **31** (1957) 2676.
42. E. F. Jordan, K. M. Doughty, W. S. Port, *J. Appl. Polym. Sci.* **4** (1960) 203.
43. J. T. Khamis, *Polymer* **6** (1965) 98.
44. K. Moser, R. Signer, H. U. Stuber, *Chimia* **23** (1969) 393.
45. G. Talamini, G. Vidotto, C. Garbuglio, *Chim. Ind. (Milan)* **47** (1965) 955.
46. R. Z. Greenley in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* 3rd ed., Wiley-Interscience, New York 1989, p. II/267.
47. H. Fikentscher, H. Gerrens, H. Schuller, *Angew. Chem.* **72** (1960) 856.
48. A. E. Alexander, D. H. Napper, *Prog. Polym. Sci.* **3** (1971) 145.
49. G. Markert, *Angew. Makromol. Chem.* **123/124** (1984) 285.
50. *Houben-Weyl,* 4th ed., **E20/2,** 1150.
51. W. Kern, *Makromol. Chem.* **1** (1947/48) 199.
52. W. Kern, *Angew. Chem.* **61** (1949) 471.
53. C. F. Fryling, A. E. Follett, *Polym. Sci.* **6** (1951) 59.
54. K. W. Min, W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **C11** (1974) 177.
55. J. Snuparek Jr., *Acta Polym.* **32** (1981) 368.
56. J. Snuparek Jr., *Macromol. Chem. Phys. Suppl.* **10/11** (1985) 129.
57. N. Suetterlin, *Macromol. Chem. Phys. Suppl.* **10/11** (1985) 403.
58. G. Ley, H. Gerrens, *Makromol. Chem.* **175** (1974) 563.
59. J. F. Mac Gregor, A. Penlidis, A. E. Hamielec: *IFAC Instrumentation and Automation in the Paper, Rubber, Plastics and Polymerisation Industries,* Antwerpen 1983, p. 375.
60. M. Nomura in K.-H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig and Wepf Verlag, Basel 1986, p. 41.
61. B. W. Brooks in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Verlag Chemie, Weinheim 1989, p. 3.
62. D. H. Klein, *J. Paint Technol.* **42** (1970) 335.

63. W. H. Brown, T. J. Miranda, *J. Paint Technol.* **36, part 2,** (1964) 92.
64. *Houben-Weyl,* 4th ed., **E20/2,** 1156.
65. J. N. Sen. U. Nandi, S. R. Palit, *J. Indian. Chem. Soc.* **40** (1963) 729.
66. S. D. Gadkary, S. L. Kapur, *Makromol. Chem.* **17** (1955/56) 29.
67. S. L. Kapur, S. D. Gadkary, *J. Sci. Ind. Res. Sect. B* **17** (1958) 152.
68. P. V. T. Raghuram, U. S. Nandi, *J. Polym. Sci. Polym. Chem. Ed.* **7** (1969) 2379.
69. I. M. Likhterova, E. M. Ludina, *Zh. Obshch. Khim.* **42** (1972) 194.
70. Y. Hachihama, H. Sumitomo, *Technol. Rep. Osaka Univ.* **5** (1955) 497.
71. U. S. Nandi, M. Singh, P. V. T. Raghuram, *Makromol. Chem.* **183** (1982) 1467.
72. H. U. Moritz, *Chem. Eng. Technol.* **12** (1989) 71.
73. I. L. O'Brien, F. Gornick, *J. Am. Chem. Soc.* **77** (1955) 4757.
74. Peroxidchemie, Organische Peroxide, company brochure.
75. Akzo, Initiators for Polymer Production, company brochure.
76. I. Czajlik, T. Földes-Berezsnich, F. Tüdös, E. Vertes, *Eur. Polym. J.* **17** (1981) 131.
77. S. M. Englund, *Chem. Eng. Prog.* **77** (1981) no. 8, 55.
78. D. V. Sweet (ed.): *Registry of Toxic Effects of Chemical Substances,* U.S. Department of Health and Human Services, Washington 1988.
79. J. Hansmann, *Adhäsion* **21** (1977) 272.
80. G. Stoev, M. Angelova, *HRC CC J. High Resolut. Chromatogr. Chromatogr. Commun.* **10** (1987) 25.
81. T. R. Crompton: *Analysis of Polymers,* Pergamon Press, Oxford 1989, chap. 9.
82. E. B. Bradford, J. W. Vanderhoff, *J. Polym. Sci. Polym. Symp.* **3** (1963) 41.
83. M. Kerker: "The Scattering of Light and other Electromagnetic Radiation," in E. M. Loebl (ed.): *Physical Chemistry,* Academic Press, New York 1969.
84. R. Pecora (ed.): *Dynamic Light Scattering: Applications of Photon Correlation Spectroscopy,* Plenum Press, New York 1985.
85. R. S. Stock, W. H. Ray, *J. Polym. Sci. Polym. Phys. Ed.* **23** (1985) 1393.
86. P. G. Cummins, E. J. Staples, *Langmuir* **3** (1987) 1109.
87. S. E. Bott, *Polym. Mater. Sci. Eng.* **53** (1985) 68.
88. J. Wagner, *Chem. Ing. Tech.* **58** (1986) 578.
89. W. Mächtle, *Makromol. Chem.* **185** (1984) 1025.
90. T. R. Paxton, *J. Colloid Interface Sci.* **31** (1969) 19.
91. J. Kloubek, K. Friml, S. Petrikova, F. Krejci, *J. Polym. Sci. Polym. Phys. Ed.* **14** (1976) 1451.
92. J. G. Brodnyan, G. L. Brown, *J. Colloid Sci.* **15** (1960) 76.
93. H. Schuller, *Kolloid Z. Z. Polym.* **211** (1966) 113.
94. I. M. Krieger, *Adv. Colloid Interface Sci.* **3** (1972) 111.
95. J. Mewis, A. J. B. Spaull, *Adv. Colloid Interface Sci.* **6** (1976) 173.
96. F. B. Malihi, T. Provder, M. E. Koehler, *J. Coat. Technol.* **55** (1983) no. 702, 41.
97. T. F. Protzman, G. L. Brown, *J. Appl. Polym. Sci.* **4** (1960) 81.
98. G. Taschen, *Plaste Kautsch.* **24** (1977) 212.
99. DIN 53 787, 1974.
100. J. G. Brodnyan, T. Konen, *J. Appl. Polym. Sci.* **8** (1964) 687.
101. S. Eckersley, A. Rudin, *Polym. Mater. Sci. Eng.* **58** (1988) 1115.
102. DIN 53 765, 1991.
103. K. Ueberreiter, G. Kanig, *Z. Naturforsch.* **6A** (1951) 551.
104. K. Ueberreiter, G. Kanig, *J. Colloid Sci.* **7** (1952) 569.
105. M. C. Shen, A. Eisenberg, *Rubber Chem. Technol.* **43** (1970) 95.
106. A. Eisenberg, M. C. Shen, *Rubber Chem. Technol.* **43** (1970) 156.
107. M. Gordon, S. J. Taylor, *J. Appl. Chem.* **2** (1952) 493.
108. K. H. Illers, *Ber. Bunsenges. Phys. Chem.* **70** (1966) 353.
109. E. Penzel, N. Goetz, *Angew. Makromol. Chem.* **178** (1990) 191.
110. W. Wunderlich, *Angew. Makromol. Chem.* **11** (1970) 189.
111. R. Jerome, V. Desreux, *Eur. Polym. J.* **6** (1970) 411.
112. A. Zosel, *Farbe Lack* **94** (1988) 809.
113. D. J. Skrovanek, C. K. Schoff, *Prog. Org. Coat.* **16** (1988) 135.
114. A. Zosel, *Prog. Org. Coat.* **8** (1980) 47.
115. P. A. Small, *J. Appl. Chem.* **3** (1953) 71.
116. D. Mangaraj, S. Patra, S. B. Rath, *Makromol. Chem.* **67** (1963) 84.
117. H. Burrel, *Off. Digest* **27** (1955) part 2, 726.
118. S. L. Madorsky, *J. Polym. Sci.* **11** (1953) 491.
119. N. Grassie, J. G. Speakman, T. I. Davis, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 931.
120. R. F. B. Davies, G. E. J. Reynolds, *J. Appl. Polym. Sci.* **12** (1968) 47.

121. B. G. Bufkin, J. R. Grawe, *J. Coat. Technol.* **50** (1978) no. 641, 41.
122. J. R. Grawe, B. G. Bufkin, *J. Coat. Technol.* **50** (1978) no. 643, 67.
123. B. G. Bufkin, J. R. Grawe, *J. Coat. Technol.* **50** (1978) no. 644, 83.
124. J. R. Grawe, B. G. Bufkin, *J. Coat. Technol.* **50** (1978) no. 645, 70.
125. B. G. Bufkin, J. R. Grawe, *J. Coat. Technol.* **50** (1978) no. 647, 65.
126. J. R. Grawe, B. G. Bufkin, *J. Coat. Technol.* **51** (1979) no. 649, 34.
127. H. Warson, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **16** (1975) no. 1, 280.
128. H. R. Lucas, *J. Coat. Technol.* **57** (1985) no. 731, 49.
129. R. G. Lees, H. R. Lucas, G. A. Gelineau, *Polym. Mater. Sci. Eng.* **55** (1986) 315.
130. P. J. Flory, J. Rehner, *J. Chem. Phys.* **11** (1943) 512.
131. P. J. Flory, J. Rehner, *J. Chem. Phys.* **11** (1943) 521.
132. L. R. G. Treloar: *The Physics of Rubber Elasticity,* 3rd ed., Clarendon Press, Oxford 1975.
133. H. Gibmeier, K. Hummel, G. Kerrutt, *Kautsch. Gummi Kunstst.* **22** (1969) 537.
134. E. Penzel, A. Zosel, *Angew. Makromol. Chem.* **99** (1981) 23.
135. A. Zosel, *Farbe Lack* **82** (1976) 125.
136. A. Zosel, *Farbe Lack* **83** (1977) 804.
137. A. Zosel, *Prog. Org. Coat.* **8** (1980) 47.
138. A. Zosel, *Double Liaison Chim. Peint.* **34** (1987) no. 384, 305.
139. K. Weimann, *Farbe Lack* **93** (1987) 447.
140. E. Wistuba, *Congr. FATIPEC* **17** (1984) 209.
141. K. A. Safe, *J. Oil Colour Chem. Assoc.* **53** (1970) 599.
142. K. A. Safe, *Paint Manuf.* **30** (1960) 249.
143. E. V. Schmid, *Congr. FATIPEC* **14** (1978) 83.
144. R. Dhein, L. Fleiter, R. Küchenmeister, *Congr. FATIPEC* **14** (1978) 195.
145. L. A. Simpson, *Congr. FATIPEC* **14** (1978) 623.
146. K. Zimmerschied, *Congr. FATIPEC* **19** (1988) 105.
147. H. U. Schenck, H. Spoor, M. Marx, *Prog. Org. Coat.* **7** (1979) 1.
148. R. Zimmermann, *Farbe Lack* **82** (1976) 383.
149. W. Brushwell, *Farbe Lack* **86** (1980) 706.
150. W. Brushwell, *Farbe Lack* **90** (1984) 924.
151. G. Y. Tilak, *Prog. Org. Coat.* **13** (1985) 333.
152. G. A. Senich, R. E. Florin, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **24** (1984) 239.
153. K. Nitzl, *Coating* **13** (1980) 274.
154. K. Nitzl, *Papier (Darmstadt)* **35** (1981) 10A, V 80.
155. C. Decker, *J. Coat. Technol.* **59** (1987) no. 751, 97.
156. W. Brushwell, *Farbe Lack* **91** (1985) 812.
157. S. P. Pappas, *J. Radiat. Curing* **14** (1987) July, 6.
158. A. Vrancken, *J. Oil Colour Chem. Assoc.* **67** (1984) 118.
159. J. Reinbold, *Papier (Darmstadt)* **40** (1986) 10A, V 119.
160. J. Watanabe, P. Lepoutre, *J. Appl. Polym. Sci.* **27** (1982) 4207.
161. H. Wilfinger, *Wochenbl. Papierfabr.* **108** (1980) 153.
162. G. Hirsch, *Papier (Darmstadt)* **32** (1978) 10A, V 66.
163. M. Aschwanden, *Wochenbl. Papierfabr.* **104** (1976) 493.
164. D. Meck, A. Stephan, *Papier (Darmstadt)* **37** (1983) 137.
165. F. Reichel, *Wochenbl. Papierfabr.* **112** (1984) 373.
166. W. C. Wake: *Adhesion and the Formulation of Adhesives,* Applied Science, London 1976.
167. K. Eisenträger, W. Druschke in I. Skeist (ed.): *Handbook of Adhesives,* 2nd ed., Van Nostrand Reinhold, London 1977, chap. 32.
168. A. Zosel, *Colloid Polym. Sci.* **263** (1985) 541.
169. V. E. Basin, *Prog. Org. Coat.* **12** (1984) 213.
170. H. Jaeger, *Adhäsion* **29** (1985) no. 9, 32.
171. W. Druschke, *Adhäsion* **31** (1987) no. 5, 29 and no. 6, 26.
172. H. W. J. Müller, *FINAT News* (1984) no. 1, 25.
173. M. Toyama, T. Ito, *Polym. Plast. Technol. Eng.* **2** (1973) 161.
174. P. Fickeisen, R. Füßl, R. Hummerich, J. Neumann, *Adhäsion* **32** (1988) nos. 1 and 2, 24.
175. W. Schwindt, *Melliand Textilber.* **50** (1969) 670.
176. G. Frerker, *Melliand Textilber.* **50** (1969) 675.
177. U. Perkuhn, *Melliand Textilber.* **50** (1969) 678.
178. G. Faulhaber, *Melliand Textilber.* **50** (1969) 682.
179. W. Schwindt, G. Faulhaber, A. J. Moore, *Rev. Prog. Color. Relat. Top.* **2** (1970) 33.
180. D. Bechter, G. Kurz, *Melliand Textilber.* **58** (1977) 150.
181. *Ullmann,* 4th ed., **23,** 77.
182. A. Einwiller, *Allg. Vliesst. Rep.* 1976, 78.
183. H. Herlinger, D. Bechter, G. Kurz, *Chemiefasern Text. Ind.* **31/83** (1981) 936.
184. W. Loy, *Melliand Textilber. Int.* **69** (1988) 836.

185. K. Fischer, *Chemiefasern Text. Ind.* **36/88** (1986) 589.
186. D. Bechter, H. Herlinger, G. Kurz, *Adhäsion* **28** (1984) no. 9, 23.
187. H. C. Assent, *Melliand Textilber.* **65** (1984) 176.
188. R. Schmitt, *Chemiefasern Text. Ind.* **28/80** (1978) 911.
189. H. Kittel (ed.): *Lehrbuch der Lacke und Beschichtungen,* vol. 5, W. A. Colomb. Heenemann Verlagsges., Berlin 1977.
190. R. Schubert in H. Herfeld (ed.): *Bibliothek des Leders,* vol. 6, Umschau Verlag, Frankfurt/Main 1982.
191. L. Würtele, *Leder Häutemarkt Gerbereiwiss. + Prax.* **25** (1973) 224.
192. C. E. Cluthe, F. A. Desiderio, W. C. Prentiss, *J. Am. Leather Chem. Assoc.* **73** (1978) 22.
193. E. Penzel, G. Eckert, *Leder Häutemarkt Gerbereiwiss. + Prax.* **33** (1981) 418.
194. M. J. Osgood, *J. Oil Colour Chem. Assoc.* **70** (1987) 104.
195. *Ullmann,* 4th ed., **23,** 12.
196. P. Dürrbeck, H. Leitner, H. Schöpke, *Chemiefasern Text. Ind.* **37/89** (1987) 576.
197. H. Leitner, P. Dürrbeck, K. Stöhr, *Melliand Textilber. Int.* **70** (1989) 893.
198. *Chemical Economics Handbook,* SRI International, Menlo Park 1986.

Poly(Acrylic Acids) → Polyacrylamides and Poly(Acrylic Acids)
Polyacrylic Fibers → Fibers, 4.Synthetic Organic
Polyalkenamers → Rubber, 3.Synthetic
Poly(Alkylene Glycols) → Polyoxyalkylenes
Polyamide Fibers → Fibers, 4.Synthetic Organic

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Polyamides

Standard Article

Melvin I. Kohan[1], Steve A. Mestemacher[2], Rolando U. Pagilagan[3], Kate Redmond[4]

[1]MIK Associates, Wilmington, Delaware, United States
[2]E. I. Du Pont de Nemours & Co., Inc., Parkersburg, West Virginia, United States
[3]E. I. Du Pont de Nemours & Co., Inc., Parkersburg, West Virginia, United States
[4]E. I. Du Pont de Nemours & Co., Inc., Wilmington, Delaware, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a21_179.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Nomenclature
3. General Considerations in Polyamidation
3.1. Molecular Mass
3.2. Equilibrium and Rate Constants
3.3. Effects of Monomer Structure
3.4. Amide Interchange
3.5. Structure-Property Considerations
4. Other Polymerization Techniques
4.1. Variants of Hydrolytic Polymerization
4.2. Interfacial and Solution Polymerization
4.3. Ionic Polymerization
4.4. Solid-Phase Polymerization
5. Commercial Processes
5.1. PA 46
5.2. PA 66
5.2.1. Batch Production
5.2.2. Continuous Production
5.3. Other AABB-Polyamides
5.4. PA 6
5.5. PA 11
5.6. PA 12
6. Properties
6.1. Properties of Unmodified Polyamides
6.2. Copolymerization
6.3. Modification by Additives
6.3.1. Filled and Reinforced Polyamides
6.3.2. Blends and Alloys
6.3.3. Flame-Retardant Polyamides
6.3.4. Conductive Polyamides
6.3.5. Other Formulations
7. Processing
8. Uses

9. Ecological Aspects and Toxicology
9.1. Recycling
9.2. Polyamide Monomers
9.3. Other Aspects
10. Economic Aspects

[Top of Page]

1. Introduction

Polyamides are a versatile family of thermoplastics that have properties that vary broadly from relative flexibility to significant stiffness, strength, and toughness. The term “nylon” is generic and is equivalent to “polyamide”. A key characteristic of nylons is their resistance to oils, greases, lubricants, hydrocarbons, and most chemicals. Phenols, strong mineral acids, and oxidizing agents attack them.

Polyamides are synthetic polymers that contain multiple amide —CONH— groups as a recurring part of the chain. The most important commercial nylons are the semicrystalline materials, which account for over 90 % of global usage. There is another significant class of nylons that are referred to as being “amorphous”. These typically contain a ring structure. There is an almost unlimited range of nylon materials that can be made by using different monomers.

The first publications describing polycondensation [1], [2]—the basic principles of nylon synthesis were in 1929. Synthesis of poly(hexamethylene adipamide), the original “nylon”, was carried out by WALLACE H. CAROTHERS in a Du Pont laboratory in 1935 and provided a material whose properties were judged to be appropriate for apparel use. This invention culminated in the first patent for the production of synthetic polyamides in 1937 [3], and their subsequent commercial introduction in 1938. The other principal commercially important polyamide, nylon 6, based on polycaprolactam, was first made at IG Farbenindustrie in Germany by P. Schlack in 1938, who was issued a patent in 1941 [4]. The history of these nylons and those made from 11-aminoundecanoic acid and dodecanolactam were reviewed in 1986 [5-7].

The intellectual achievement of creating nylon was a major breakthrough in the science of synthetic polymer chemistry. This development provided incontrovertible evidence for the Staudinger thesis [8] that polymers are high molecular mass species and not aggregated entities. Materials with new types of physical properties were achievable, and could be tailored to various end uses by adjusting parameters such as molecular mass, viscosity, backbone chemistry, additives, etc. Carothers' classic 1931 review [9] of polymerization and his first definitions of the terminology of polycondensation are still valid today. Polycondensation polymerization is a type of step-growth or step polymerization [10] (see Chap. General Considerations in Polyamidation). The initial success of nylon led to the numerous scientific and engineering innovations that were required to commercially develop and extend polymer usage. Routes to starting materials were necessary, as was the development of polymerization equipment to manufacture the polymer, and analytical techniques to measure the materials.

Mechanical and physical properties, rate and equilibrium constants, thermal and oxidative stability, resistance to hydrolysis and solvents, and other chemical behavior depend on the specific nylon polymer, but the basic chemistry discussed herein applies to all of these polyamides. Even though nylon was initially used as a fiber, its application as a plastic for a variety of purposes such as brush filaments, wire coating, coil forms, gears, and a variety of parts occurred very soon thereafter. The focus of this article is on the polyamides used as plastics. Presently, the growth in engineering plastics applications is much greater than in fibers applications. The technology has evolved to the point where nylons, especially reinforced compositions, are widely used in place of metals, even in under-the-hood applications in automobiles (see Chap. Uses). For a discussion of polyamides used as fibers, see ⟶ Fibers, 4. Synthetic Organic – Polyamide Fibers.

[Top of Page]

2. Nomenclature

In current terminology the word “nylon” is a generic term reserved for aliphatic and semiaromatic polyamides. The fully aromatic polyamides are called “aramids” (see ⟶). The origin of the word nylon is uncertain but there is reasonable speculation that it is a mutation of the word “norun” from nylon's first use in hosiery [11, p. 3], [12]. Although nylons can be prepared from various reactive derivatives of carboxylic acids and amines, in general especially in commercial manufacture, nylons are prepared directly from (1) dicarboxylic acids and diamines, (2) ω-aminoacids, or (3) lactams:

$$n\,H_2N(CH_2)_xNH_2 + n\,HOOC(CH_2)_{y-2}COOH \rightleftharpoons H[NH(CH_2)_xNHCO(CH_2)_{y-2}CO]_nOH + (2n-1)H_2O \quad (1)$$

$$n\,H_2N(CH_2)_{x-1}COOH \rightleftharpoons H[NH(CH_2)_{x-1}CO]_nOH + (n-1)H_2O \quad (2)$$

$$n\,\underbrace{HN(CH_2)_{x-1}C}=O + H_2O \rightleftharpoons H[NH(CH_2)_{x-1}CO]_nOH \quad (3)$$

Equation (3) represents the hydrolytic polymerization of a lactam. It is described in detail in Section Equilibrium and Rate Constants. Two basic monomer types are possible. One where both the acid and amine functionalities are contained in the same molecule is called an AB type monomer and the case where one monomer contains the amine functionality and another the carboxylic moiety is referred to as an AABB type. In this case, two monomers are required for polymerization, one containing the amine and another the carboxylic moiety. A and B stand for the functional groups —NH_2 and —COOH, respectively. Lactams are considered AB type monomers as these are derivatives of amino acids, which, by virtue of their structures, have a strong driving force for cyclization (lactam formation). In current practice, AABB and AB may describe products that do not involve polymerization via amine and acid ends but can be visualized as if so made. This is realistic because under conditions of fabrication and processing, moisture is always present. Reaction of the polyamide with moisture under these conditions (hydrolysis), regardless of the starting monomers used in their preparation, would generate the corresponding amino acid or the acid and amine monomers. The AABB types are regarded as homopolymers because of the absolute requirement for alternation of both reactants to form a polymeric chain that includes both moieties in the repeating structure (see Chap. General Considerations in Polyamidation). This is consistent with IUPAC recommendations [13]. Copolymers result from simultaneous polymerization of AABB and AB types or use of more than one AA, BB, or AB type. They are common and were used as early as 1940 [6].

In the early development of nylons, the monomers used were most often linear aliphatic compounds. A simple system of nomenclature that is still used today is to designate these polyamides by the number of carbon atoms in the monomers. The polyamides (PA) or nylons are identified by numbers corresponding to the number of carbon atoms in the monomers (diamine first for the AABB type). Thus, two numbers represent an AABB polyamide; the first number refers to the diamine and the second number to the carboxyl monomer. An AB type is represented by a single number. Table 1 provides a list of the common names of the commercial nylon homopolymers, their numerical designations, their CAS registry numbers, and their CAS names. The monomers TMD in TMDT and PACM in PACMT are mixtures of isomers and TMDT and PACMT are in fact copolymers. The systematic IUPAC naming of polyamides [14] is awkward and is rarely encountered in commercial literature. The CAS names are largely in accord with IUPAC practice, but examination of Table 1 reveals some inconsistencies. Many of the IUPAC rules for nomenclature are conveniently assembled in a polymer handbook [15] wherein, however, older versions of the IUPAC names for PA 66 and 6 are used [16]: poly(iminoadipoyl-iminohexamethylene) and poly[imino(1-oxohexamethylene)].

Table 1. Commercial polyamides or nylons

Common name	*xy* or *x* *	CAS name, CAS registry no.
Poly(tetramethylene adipamide)	46	poly[imino(1,4-dioxo-1,4-butanediyl)imino-1,6-hexanediyl] [24936-71-8]
Poly(hexamethylene adipamide)	66	poly[imino(1,6-dioxo-1,6-hexanediyl)imino-1,6-hexanediyl] [32131-17-2]
Poly(hexamethylene azelaamide)	69	poly[imino-1,6-hexanediylimino(1,9-dioxo-1,9-nonanediyl)] [28757-63-3]
Poly(hexamethylene sebacamide)	610	poly[imino-1,6-hexanediylimino(1,10-dioxo-1,10-decanediyl)] [9008-66-6]
Poly(hexamethylene dodecanoamide)	612	poly[imino-1,6-hexanediylimino(1,12-dioxo-1,12-dodecanediyl)] [24936-74-1]
Poly(*m*-xylylene adipamide)	MXD 6	poly[iminomethylene-1,3-phenylene-methyleneimino(1,6-dioxo-1,6-hexanediyl)] [25805-74-7]
Poly(trimethylhexamethylene terephthalamide)	TMDT	1,4-benzenedicarboxylic acid, polymer with 2,2,4-trimethyl-1,6-hexanediamine and 2,4,4-trimethyl-1,6-hexanediamine [25497-66-9]
Poly(11-aminoundecanoamide)	11	poly[imino(1-oxo-1,11-undecanediyl)] [25035-04-5]
Polycaprolactam or polycaproamide	6	poly[imino(1-oxo-1,6-hexanediyl)] [25038-54-4]
Polydodecanolactam or polylaurolactam or polydodecanoamide	12	poly[imino(1-oxo-1,12-dodecanediyl)] [24937-16-4]
Hexamethyleneadipamide-hexamethyleneterephthalamide copolyamide	6T/66	1,4-benzenedicarboxylic acid, polymer with 1,6-hexanediamine and hexanedioic acid [25776-72-1]
ε-Caprolactam-hexamethyleneterephthalamide copolyamide	6T/6	1,4-benzenedicarboxylic acid, polymer with hexahydro-2*H*-axepine-2-one and 1,6-hexanediamine [25086-53-7]
Bis(4-aminocyclohexyl)methane-terephthalic acid copolymer	PACMT	1,4-Benzenedicarboxylic acid, polymer with 4,4′-methylenebis(cyclohexanamine) [117092-07-6]

* *xy* and *x* refer to the number of carbon atoms according to Equations (1) – (3).

Table 2. lists some of the most common monomers. As shown in the tables, numerical symbols are simple, convenient, and unambiguous only for linear aliphatic monomers. This system of nomenclature is inadequate and it is often necessary to use letters instead of numbers for monomers of more complex structures. The absence of unifying rules to deal with monomers other than linear, aliphatic monomers has led to different symbols for the same monomer. More recently ISO 1874-1 has been established and gives recommended symbols for some monomers; however, it fell short of establishing rules for assigning letter designation for complex monomers. Notice the new designation for the isomers of TMD wherein the two positional isomers are given designations of ND for the 2,2,4-trimethylhexamethylenediamine and IND for the 2,4,4-trimethylhexamethylenediamine. However, no distinct symbols were assigned to the *cis–cis, cis–trans*, and *trans–trans* isomers of PACM. Another aspect still not addressed by ISO 1874-1 is the configurational isomerism with polyamides of nonsymmetrical monomers such as ND, IND, and MPMD. These diamines can produce head-to-head, head-to-tail, and tail-to-tail isomerism in their polymers.

Table 2. Monomers for polyamides

Common and (CAS) name	CAS registry no.	*x* or *y* *	Source	Properties and synthesis
Adipic acid (hexanedioic acid)	[124-04-9]	6	benzene, toluene	⟶ Adipic Acid
Azelaic acid (nonanedioic acid)	[123-99-9]	9	oleic acid	⟶ Dicarboxylic Acids, Aliphatic – Azelaic Acid
Sebacic acid (decanedioic acid)	[111-20-6]	10	castor oil	⟶ Dicarboxylic Acids, Aliphatic – Sebacic Acid
Dodecanedioic acid	[693-23-2]	12	butadiene	⟶ Dicarboxylic Acids, Aliphatic
Dimer acid (fatty acids, dimers)	[61788-89-4]	36	oleic and linoleic acids	⟶ Dicarboxylic Acids, Aliphatic – Dimer Acids
Isophthalic acid (1,3-benzenedicarboxylic acid)	[121-91-5]	I	*m*-xylene	⟶ Carboxylic Acids, Aromatic ⟶ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid
Terephthalic acid (1,4-benzenedicarboxylic acid)	[100-21-0]	T	*p*-xylene	⟶ Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid
Tetramethylenediamine, 1,4-diaminobutane (1,4-butanediamine)	[110-60-1]	4	acrylonitrile and HCN	⟶ Amines, Aliphatic – Higher Diamines
Hexamethylenediamine, 1,6-diaminohexane (1,6-hexanediamine)	[124-09-4]	6	butadiene, propene	⟶
2-Methylpentamethylenediamine (1,5-pentanediamine, 2-methyl)	[15520-10-2]	MPMD	butadiene	
4,4′-Diaminodicyclohexylmethane (cyclohexaneamine, 4,4′-methylenebis-)	[1761-71-3]	PACM	aniline and formaldehyde	
m-Xylylenediamine (1,3-benzenedimethaneamine)	[1477-55-0]	MXD	*m*-xylene	
2,2,4-Trimethylenehexamethylenediamine	[3236-53-1]	ND	acetone	
2,4,4-Trimethylenehexamethylenediamine	[3236-54-1]	IND	acetone	
Dodecamethylenediamine, 1,12-diaminododecane	[2783-17-7]	12	butadiene	⟶ Amines, Aliphatic – Higher Diamines

(1,12-dodecanediamine)				
11-Aminoundecanoic acid (undecanoic acid, 11-amino-)	[2432-99-7]	11	castor oil	
ϵ-Caprolactam (2 *H*-azepin-2-one, hexahydro)	[105-60-2]	6	benzene, toluene	⟶ Caprolactam
Laurolactam, dodecanolactam, (azocyclotridecane-2-one)	[947-04-6]	12	butadiene	⟶

* *x, y* = number of carbon atoms according to Equations (1) – (3).

The ISO recommendation for polyamides is to use the symbol PA for polyamide followed by the monomer symbol. Pronunciation is in accord with the monomers so that the homopolymer from hexamethylenediamine and adipic acid, which is represented by PA 66, is pronounced PA six-six not sixty-six, PA 612 is PA six-twelve, PA 11 is PA eleven, etc. Often the space between PA and the monomer symbol is eliminated as in PA66, PA612, PA11. Copolymers are designated by "PA" followed by the polyamide symbols separated by a slash. For example, a copolymer of PA 6 and PA 66 is represented as PA 6/66, and a copolymer of PA 66 and PA 610 is PA 66/610. No provisions for polyamide blends have been established. KOHAN [11, p. 5] designates blends by separating the component polyamides with a slash. For example a blend of PA 6 and PA 66 is represented as PA 6/PA 66 to distinguish it from the copolymer PA 6/66.

[Top of Page]

3. General Considerations in Polyamidation

CAROTHERS [1] proposed the term "polycondensation" for polymerization reactions wherein two functional groups react to form a covalent bond with the elimination of a simple molecule as a byproduct in accordance with the definition of condensation reactions in organic chemistry. In this type of polymerization, chain growth involves the stepwise reaction of the functional groups. Step-growth polymerization, a broader term that encompasses all polycondensation polymerizations, was first used by LENZ [10]. Detailed coverage of step-growth polymerization is given in Polymerization Processes [17].

The basic concepts of equilibria and kinetics in polyamidation are in large measure due to FLORY and are summarized in his classic text [18]. Derivations of the following important expressions are to be found in many places, e.g., [17-20], and will not be developed here.

3.1. Molecular Mass

The extent of reaction, *p*, is the fraction of possible reaction that has occurred. The number-average degree of polymerization, $\bar{P}_n$ (IUPAC recommends $\bar{X}_n$), is the number of monomer units in the average polymer chain and is 2*n* for the AABB type nylon and *n* for the AB type where *n* is as shown in Equations (1) – (3). The typical polymer consists of molecules of varying molecular masses, so it is necessary to talk of average values. The number-average molecular mass, $\bar{M}_n$, is the product of $\bar{P}_n$ and the average molecular mass of the units in the polymer (see also ⟶ Plastics, Analysis – Molecular Mass Determination). This ignores the contribution of end groups, which is unimportant at usefully high molecular masses. $\bar{M}_n$ is, precisely, the sum over all species of the number fraction of each molecular mass species multiplied by its molecular mass. The relation between $\bar{P}_n$ and *p* is:

$$\bar{P}_n = \frac{1}{1-p} \tag{4}$$

This equation applies strictly to equilibrium conditions but is well suited to commercial polyamides. For PA 6 or PA 66, the average mass per amide is 113. Thus, $\bar{P}_n$ is 100 for a $\bar{M}_n$ of 11 300 and *p* is 0.990. This is at the low end of the range of useful $\bar{M}_n$ values. For a more typical value of about 20 000, *p* is 0.994. The need for pure monomers and clean, essentially complete reactions is clear. $\bar{M}_n$ for most commercial nylons is in the range 11 000 – 34 000. Physical properties depend on $\bar{M}_n$ which can be conveniently determined by end group analysis.

The weight-average molecular mass, $\bar{M}_w$, is given by $\bar{M}_w = \sum_{w_i}^{i} M_i + \sum_i (M_e)_i$ i.e., the sum of the products of the weight fraction w_i of each of the molecular mass species *i* and its molecular mass M_i plus the contributions of end groups. Melt viscosity, important in processing, varies with $\bar{M}_w^{3.4}$ [21-23]. $\bar{M}_w$ is available from light scattering or equilibrium ultracentrifugation. Solution viscosity is often used to approximate molecular mass and corresponds more nearly to $\bar{M}_w$. Both $\bar{M}_n$ and $\bar{M}_w$ can be obtained by chromatographic techniques [24-27]. $\bar{M}_w$ relates to *p* as shown:

$$\bar{M}_w = (1+p)/(1-p) \tag{5}$$

The distribution of molecular masses is important. A much-used index of the distribution is the ratio, $\bar{M}_w/\bar{M}_n$. Dividing Equation (5) by Equation (4) yields

$$\bar{M}_w/\bar{M}_n = 1+p \tag{6}$$

At useful molecular masses where p nearly equals 1, $\bar{M}_w/\bar{M}_n$ is almost exactly 2. This was defined by FLORY as “the most probable distribution” of molecular masses. In such a distribution, which applies closely to almost all commercial nylons, the number fraction of any species decreases with increasing molecular mass. The weight fraction versus molecular mass curve goes through a maximum, the position of which is equal to $\bar{P}_n$ [18], [28].

The most convenient analytical technique is solution viscometry, which delivers a so-called viscosity-average molecular mass, $\bar{M}_v$. Solution viscosity can be used to estimate both $\bar{M}_n$ and $\bar{M}_w$ as long as their ratio is known and remains the same. Formic acid is a convenient solvent for PA 6, PA 46, and PA 66, but *m*-cresol or sulfuric acid must be used for nylons with fewer amide groups. Details of the technique and expressions relating viscosity and molecular mass can be found in [19].

Added or adventitious reactants with more than two reacting groups per reactant molecule (or just two alike in the case of AB types) will lead to branching and, in turn, will change the equilibrium distribution of molecular masses [18], [29]:

$$\bar{M}_n = 2/[N_e - (f-2)b] \tag{7}$$

$$\bar{M}_w = 4/[N_e - f(f-2)b] \tag{8}$$

where

N_e = number of ends

f = functionality (number of reactive groups in a monomer molecule)

b = number of branch points

At onset of gelation, where $\bar{M}_w$ goes to infinity, the number of branch points is

$$b_{gel} = N_e/f(f-2) \tag{9}$$

Since the number of moles is $1/M_n$, b' the number of branch points per mole, is $b\,\bar{M}_n$. At the gel point

$$b' = 2\,b_{gel}/[N_e - (f-2)b_{gel}] \tag{10}$$

For $f = 3$, $b_{gel} = N_e/3$, and $b'_{gel} = 2\,b_{gel}/(N_e - b_{gel}) = 1$. That is, it takes only one branch point per polymer molecule to cause gelation if a trifunctional reactant such as trimesic acid [554-95-0] (1,3,5-benzenetricarboxylic acid), or bishexamethylenetriamine [143-23-7] [*N*-(6-aminohexyl)-1,6-hexanediamine] is present. The latter compound can be formed by intermolecular reaction of two amine ends of 1,6-hexamethylenediamine with loss of ammonia.

Star-type branching occurs when AB monomers react with a polyfunctional reactant that contains only A or B groups and not both kinds. In this special situation [30] gelation cannot occur. Broad distributions occur for low p values and narrow distributions for p approximating unity in which case $\bar{M}_w/\bar{M}_n = 1 + f^{-1}$. Branched polyamides were introduced commercially in the late 1990 [31-34] (see Section Control of Rheological Properties.).

The effect of mixing two polymers of different molecular masses is often questioned. In the case of $(\bar{M}_w)_1/(\bar{M}_n)_1 = (\bar{M}_w)_2/(\bar{M}_n)_2 \approx 2$, which most often applies to nylons [19]:

$$(\bar{M}_w/\bar{M}_n)_{blend} = 2 + \left[2\,F_1F_2(R-1)^2/R\right] \tag{11}$$

where F is the weight fraction of components 1 or 2 and R the ratio of number-average molecular masses of components 1 and 2. Thus, for $F_1 = F_2 = 0.5$ and $R = (\bar{M}_n)_2/(\bar{M}_n)_1 = 2$, an extreme case for nylons, $\bar{M}_w/\bar{M}_n = 2.25$. This type of blending, therefore, has little effect on the molecular mass ratio and the changes observed are due to the changes in $\bar{M}_w$ and $\bar{M}_n$.

The reaction of amine and acid to form an amide is described by the following equilibrium expression:

$$K_{eq} = \frac{[CONH][H_2O]}{[COOH][NH_2]} \tag{12}$$

At high molecular masses [CONH] is essentially constant and depends only on the average molecular mass per monomeric unit. For PA 6 or PA 66 this becomes 1/113. For numerical convenience the units of ends and other concentrations are often expressed as the number of moles (or equivalents) per 10^6 g so that [CONH] is 8850.

The product of end group concentrations, [COOH] [NH_2] = P, is determined by the temperature and moisture content. The molecular mass depends on the number of ends, which is minimized (and the molecular mass maximized) for a fixed P when the concentration of amine and acid ends is the same, i.e., balanced. An imbalance may be imposed on purpose to limit the molecular mass by adding an excess of one kind of monomer or a chain-terminating reactant that contains only one functional group.

Addition of AA or BB in the polymerization of AB yields a polymer that at high extents of reaction (p very close to unity) has essentially only one kind of end group and belongs to the class of star-type products mentioned above. In this case the distribution of molecular masses represented by $\bar{M}_w/\bar{M}_n$ no longer approximates to 2 and, because the functionality is 2, is

about 1.5. The addition of monofunctional acids or amines in the polymerization of AB or AABB and the addition of excess diacid or diamine in the polymerization of AABB does not alter the distribution and can be treated in a general way. If n is the amount of additive expressed as the mole fraction of AB or either AA or BB, and not the sum of AA and BB, then $\bar{P}_n$ is simply $1/n$ in the case of the AB polymer and $2/n$ for the AABB polymer. The factor of 2 is due to the fact that reaction of a difunctional monomer yields two amide groups. Including the contribution of the additive requires adding 1 to $\bar{P}_n$.

A problem in AABB polymerization lies in the fact that the imbalance of ends often results from loss of diamine during polymerization because its volatility is significantly higher than that of the diacid. It cannot, therefore, be anticipated but must be determined by analysis of the product. Assume N_A and N_B are the numbers of amine and acid ends found in the product in convenient units such as equivalents per 10^6 g. Assume also that a monofunctional acid has been added in an amount equal to n defined as in the above paragraph. Assume further that the average of the molecular masses of AA and BB less the 2 mol of water eliminated by polymerization is M so that the concentration of amide functions in the polymer, in the same units, is $10^6/M$. The total number of ends, N_T, in the polymer is $N_A + N_B + n\ 10^6/2\ M$. Because the polymer is linear, the number of amide groups per polymer molecule, that is $\bar{P}_n$, is $2 \times 10^6/MN_T$. By way of example, if N_A, N_B, and n are 70, 30, and 0.01, respectively for PA 66, N_T is $100 + 0.01 \times 4425$ and $\bar{P}_n$ is $2 \times 10^6/(113 \times 144) = 123$.

A related question is the implication of imbalanced ends for the equilibrium moisture content where it is desired to achieve or maintain a given molecular mass, i.e., have the same sum of ends. $P_{bal} = N_e^2/4$ if the ends are balanced, i.e., the amine and acid ends are equal in number and add up to N_e. For imbalanced ends and the same N_e, let $P_{imb} = N_A N_B$ where N_A and N_B are the new concentrations of amine and acid ends. For excess acid ends, define $r = (N_B - N_A)/N_e$. Since $N_e = N_B + N_A$, it is readily shown that $1 - r^2$ is $4\ N_A N_B / N_e^2$. Thus,

$$P_{imb} = (N_e^2/4)(1 - r^2) = P_{bal}(1 - r^2) \quad (13)$$

Because the equilibrium moisture content varies directly with the product of ends,

$$[H_2O]_{imb} = [H_2O]_{bal}(1 - r^2) \quad (14)$$

where $[H_2O]_{imb}$ is the equilibrium moisture content in the case of imbalanced ends. Thus, if $r = 0.5$, which corresponds to a ratio of unreacted amine ends to unreacted acid ends (or vice versa) of 3/1 and is rather large, $[H_2O]_{imb}$ is 0.75 $[H_2O]_{bal}$. Polymerization to a higher molecular mass or hydrolysis to a lower one occurs depending on whether the water content is above or below the equilibrium value. The significance for the drying of resin and melt processing is clear.

3.2. Equilibrium and Rate Constants

The equilibrium constant for the polycondensations corresponding to Equations (1) and (2) is defined in Equation (12). Few data exist for such polycondensations. Many equilibrium and rate studies have been concerned with the so-called hydrolytic polymerization of ε-caprolactam in the presence of water. A comprehensive coverage of lactam-based polyamides is given in [35]. This polymerization is not as simple as Equation (3) indicates. Three reactions in the hydrolytic polymerization of caprolactam are of prime concern:

1. Hydrolysis of the lactam by water to yield the amino acid (Eq. 3 with $n = 1$)
2. Condensation as in Equation (2)
3. Lactam addition to the amine end of the growing chain (Eq. 15):

$$n\ \underbrace{HN(CH_2)_5C}{=}O + H[NH(CH_2)_5CO]_mOH \rightleftharpoons H[NH(CH_2)_5CO]_{m+1}OH \quad (15)$$

Also contributing to the complexity is the formation of cyclic oligomers corresponding to the following formula [36]:

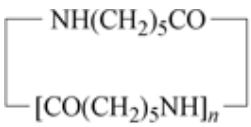

$n \geq 1$

Structures with n up to 7 have been reported [37]. The amount of lactam monomer ($n = 0$) at equlibrium depends on the temperature (Fig. 1), [19], [38], [39]; the monomer comprises about 75 % of the extractable material, i.e., monomer and cyclic oligomers [40]. Residual lactam monomer and its cyclic oligomers are relatively volatile and tend to vaporize and condense on cool surfaces during melt processing unless provision for their removal has been made. Cyclic oligomers exist not only in the higher AB-type nylons such as PA 11 and PA 12 but also in AABB types such as PA 66 and PA 610 [27], [28], [41],[42],[43]. However, the concentrations are much lower, and the extractable fractions are usually below 2 wt % [44-48] so that separation is not required as it is for PA 6.

Figure 1. Dependence of monomer on temperature in polycaprolactam (data from [19], [38], [39])

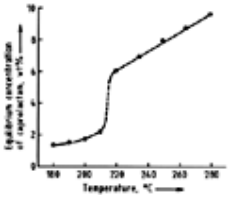

There are further complications. The amine and acid end groups may exist as ammonium and carboxylate ions; this raises the possibility of alternative reactions involving condensation reactions between ions or between an unionized end and an ion. Degree of hydration is another question. Also a concern is the use of concentrations instead of the theoretically correct activity coefficients. The dependence of conventionally calculated equilibrium constants K_{eq} on water content, which can affect ionization, hydration, and activities, has been observed many times [49-53]; these K_{eq} values are thus apparent equilibrium constants. GIORI and HAYES [54] showed for PA 6 that the apparent equilibrium constant for polycondensation increased with water content up to about 2 wt % and decreased thereafter (Fig. 2). They used vapor – liquid equilibria, which showed a negative deviation from Raoult's law, to obtain activities [55]. K_{eq} (their K_{III}) showed much less deviation with percent water when using activity coefficients instead of concentrations (Fig. 3). Their K_{eq} at 2 wt % water (using concentrations) is about 600 at 240 °C and 450 – 500 at 270 °C. Values of 519 and 377 for comparable conditions can be calculated from the data of TAI and coworkers [56]. These variations are typical of the data available on the hydrolytic polymerization of caprolactam, most of which have been generated at water contents much higher than practical for melt processing.

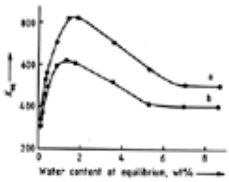

Figure 2. Apparent equilibrium constants for polycondensation versus equilibrium water content

a) 240 °C; b) 260 °C

Reproduced with permission of John Wiley and Sons from Figure 1 in [54].

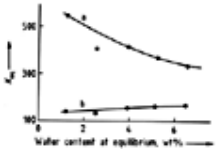

Figure 3. Equilibrium constants for nylon 6 polycondensation versus equilibrium water content at 270 °C

a) Apparent equilibrium constants; b) Equilibrium constants corrected for the water activity coefficient

Reproduced with permission of John Wiley and Sons from Figure 8 in [54].

Nylons are typically sold with about ≤ 0.2 wt % water. Extrapolation of data obtained at 0.2 wt % water at 240 and 260 °C by GIORI and HAYES [54] gives K_{eq} = 280 at 280 °C. TAI's data [56] yield K_{eq} = 410 at 280 °C at the lowest water content studied (0.42 mol/kg or 0.76 wt %). ZIMMERMAN [57] used a reported moisture content of 0.16 wt %, an end product of 3000, and an amide concentration of 8850 g equivalents/10^6 g at 101.3 kPa steam for PA 66 at 280 °C to calculate a K_{eq} of 260. The effect of moisture on processing requirements is best resolved by examining its significance for the stability of melt viscosity at the temperatures of concern as long as excessive holdup times and temperatures leading to thermal degradation are not involved [19].

Kinetic and thermodynamic constants for the polymerization of caprolactam have been summarized ([56], Table 3) and have been said to correlate with manufacturing experience with a deviation of less than 12 wt % [58]. The data in Table 3 are averaged for initial water contents of 0.82 and 1.18 mol/kg (1.48 and 2.12 wt %). Examination of the original data, which also included an initial water content of 0.42 mol/kg, indicates modest changes in the various constants except in the case of the temperature-independent constant, A^c_2 for catalyzed polycondensation which increased by a factor of 20.4 over the 2.8 increase in water content. Enhanced ionization and catalysis are thus indicated to be more important than reduced concentration of unionized species. Both second- and third-order kinetics (uncatalyzed and acid catalyzed) for all three processes (Eqs. 2, 3 with n = 1, and 15) have been observed. By way of illustration the appropriate expressions for the polycondensation reactions are shown below:

$$-\mathrm{d}[\mathrm{end}]/\mathrm{d}\,t = k_2^o[\mathrm{NH_2}][\mathrm{COOH}] \quad (16)$$

$$-\mathrm{d}[\mathrm{end}]/\mathrm{d}\,t = k_2^c[\mathrm{NH_2}][\mathrm{COOH}]^2 \quad (17)$$

Amidation is reversible so that the contribution of hydrolysis cannot be ignored. The rate constant for hydrolysis is k/K_{eq} where K_{eq} is defined as in Equation (12) and k is the rate constant for amidation. At low water contents and high extents of reaction, as in the case of commercial nylons, third-order kinetics apply. The optimum water concentration for polymerization varies with conversion—high initially to promote opening of the ring and low in the later stages to promote high conversion. Figures 4 and 5 illustrate this effect of water content.

Table 3. Kinetic and thermodynamic constants in the hydrolytic polymerization of caprolactam *

i	A^0_i, kg mol^{-1} h^{-1}	E^0_i, cal/mol	A^c_i, kg^2 mol^{-2} h^{-1}	E^c_i, cal/mol	ΔS_i, e.u.	ΔH_i, cal/mol
1	5.9874×10^5	1.9880×10^4	4.3075×10^7	1.8806×10^4	– 7.8846	1.9180×10^3
2	1.8942×10^{10}	2.3271×10^4	1.2114×10^{10}	2.0670×10^4	0.94374	-5.9458×10^3
3	2.8558×10^9	2.2845×10^4	1.6377×10^{10}	2.0107×10^4	– 6.9457	-4.0438×10^3

* Average of data obtained with initial water contents of 0.82 and 1.18 mol/kg (1.48 and 2.12 wt %, respectively). A_i = pre-exponential factor and E_i = activation energy from $k_i = A_i \exp(-E_i/RT)$; ΔS_i = entropy; ΔH_i = enthalpy; i = 1, ring opening (hydrolysis of lactam to aminocaproic acid); i = 2, polycondensation (of aminocaproic acid); i = 3, polyaddition (addition polymerization of ε-caprolactam). Reproduced with permission of John Wiley and Sons from Table II in [56].

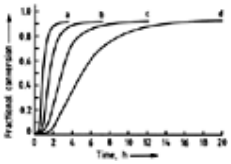

Figure 4. Effect of initial mass fraction of water (W_0) on rate and extent of conversion (T = 265 °C)

a) W_0 = 0.08; b) W_0 = 0.04; c) W_0 = 0.02; d) W_0 = 0.01

Reproduced with permission of John Wiley and Sons from Figure 8 in [51].

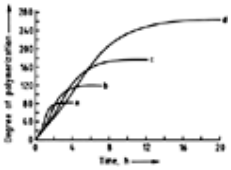

Figure 5. Effect of initial mass fraction of water (W_0) on rate and degree of polymerization

a) W_0 = 0.08; b) W_0 = 0.04; c) W_0 = 0.02; d) W_0 = 0.01

Reproduced with permission of John Wiley and Sons from Figure 10 in [51].

The amidation constants corresponding to Equations (1) and (2) are generally applicable because of the postulate that end group activity is independent of chain length [18], but overall reaction rates and ring-chain equilibria vary because of differences in ring stability. Rings with 5, 6, 12, or 13 members are more stable than the 7-membered caprolactam. The larger rings are also, however, more difficult to form. This accounts for the use of catalysts in synthesizing PA 12 and for its lower equilibrium concentration of extractables. An excellent review of the properties of lactams has been provided by SEKIGUCHI [52].

Catalysts for amidation are not used for PA 6 or PA 66, although a small amount of aminocaproic acid or diammonium carboxylate such as the salt used to make PA 66 can be added to the caprolactam – water mix to eliminate the induction time required for lactam hydrolysis. Reported catalysts for amidation include metal oxides and carbonates or acidic materials [59], strong acids [60], lead monoxide [61], terephthalate esters [62], acid mixtures [63], titanium alkoxides or carboxylates [64], and phosphorus compounds [65-67].

3.3. Effects of Monomer Structure

The ability of monomers to participate in polyamidation is markedly influenced by their tendency to form cyclic structures and the stability of those structures once formed. CAROTHERS [1] showed that it is difficult to form high molecular mass polymers when the formation of cyclic structures is favored. This principle affects the formation of both AB and AABB-type polymers, but in different ways. Lactams with five or six members are more stable than the seven-membered caprolactam rings, and their formation is favored over chain growth. The stability of the 5-membered ring lactam is so high that melt polymerization of pyrrolidone to PA 4 results only in low molecular mass polymers. Polymers of useful molecular mass have been obtained from these monomers only by using mild conditions such as solid-phase polymerization or anionic polymerization [68], and under these conditions the monomer will be regenerated when the polymer is heated (Section Equilibrium and Rate Constants). The 12- and 13- membered lactams are less reactive than caprolactam and require the use of catalysts for polymerization. Despite its enhanced reactivity, caprolactam has an equilibrium concentration of 8 % in the presence of its polymer. The equilibrium concentrations of 12- and 13-membered lactams in their polymerizations are less than 0.5 % because the reformation of these large rings is entropically unfavorable.

Ring formation can also lead to byproducts and chain-termination reactions in AABB polymerization. Succinic and glutaric acids, for example, produce only low molecular mass polymer under melt polyamidation conditions due to formation of the favored five- and six-membered cyclic imides derived from a reaction with the diamine comonomer. Although adipic acid has very little tendency to form its seven-membered cyclic imide, under typical polymerization conditions it does cyclize to the five-membered cyclopentanone by the elimination of carbon dioxide. The formation of pyrrolidine from 1,4-diaminobutane was reported by ROERDINK et al. [69], [70] in the manufacture of PA 46. The pyrrolidine formation and the instability of adipic acid preclude the attainment of high molecular mass in the melt polymerization of PA 46 (see Section PA 46).

3.4. Amide Interchange

Amide interchange (transamidation) is the interchange of amide groups between various polymer molecules or segments thereof. It is a factor when blending polyamides different molecular masses and/or different polymers. As noted above (Eq. 11), the effect of transamidation on molecular mass distribution is minor. Furthermore, there is rapid approach to an

equilibrium value [53], [71]. The degree of interchange necessary for producing a random copolymer is slow enough, however, that block polymers are feasible [53], [72], although high conversion of the mix results in block lengths significantly smaller than the original chain lengths [73]. A study of the interchange reaction involving polymerization of sebacamide and *N*,*N′*-diacetylhexamethylenediamine suggested that the rate increased with water concentration and was proportional to $[COOH]^{1/2}$ [53]. MILLER studied the reaction of model compounds with each other and with PA 66 [73]. The primary reaction was acidolysis; aminolysis was acid catalyzed and did not occur in the absence of acid.

3.5. Structure-Property Considerations

Polyamides are valued as a plastic due to their good mechanical properties, excellent chemical resistance, and their ability to resist high temperatures. These qualities are due to their crystallinity and high melting points (see Chap. Properties). Data that have been accumulated from the wide range of polyamides that have been studied have led to some generalizations on the effects of structure on their properties.

A large body of evidence has demonstrated that the underlying characteristic of polyamides that influences their properties is their ability to form strong hydrogen bonds. Any structural changes that affect the hydrogen bonding ability will have a very large effect on the properties. The dependence of the melting point, T_m, on the crystal heat of fusion, ΔH_u, and entropy of fusion, ΔS_u, is defined by:

$$T_m = \frac{\Delta H_u}{\Delta S_u}$$

The high melting points of polyamides are attributed to their low liquid-state entropy of fusion due to possible retention of hydrogen bonding in the melt and some stiffening of the amide group due to resonance [74]. The increase in melting point that goes with increased ability to form hydrogen bond is exemplified by PA 66 and PA 46. The latter has an increased ratio of CONH to CH_2 groups, which means more hydrogen bonding, and a increased melting point (see Chap. Properties). On the other hand, lateral substituents on either the carbon chain or amide nitrogen or the presence of comonomers result in lower crystallinity and melting points. More drastic effects are exhibited when the substituent is bulky and/or on the nitrogen atom. The T_m for PA 66 is 265 °C but the polymer from *N*,*N*-dimethylhexamethylenediamine and adipic acid is amorphous and rubbery. A more detailed discussion of this topic is covered in [75].

[Top of Page]

4. Other Polymerization Techniques

4.1. Variants of Hydrolytic Polymerization

This category includes the use of acid derivatives or precursors instead of the acid itself [76] in high-temperature reactions. Examples are esters [77], [78]; acid halides [79]; amides [80]; and nitriles [81-85]. Another variant is the use of high-boiling, stable liquids that may or may not be a solvent for the nylon, e.g., selected phenols or white oil [86], glycols [87], and polar solvents such as *N*-methyl-2-pyrrolidone or sulfolane [88]. The use of amine derivatives is limited. The reaction of an aromatic diisocyanate with a dibasic acid is an example [89]; an appropriate choice of solvent and catalyst is necessary to avoid side reactions. A special case involves the use of the relatively reactive oxalate esters that react with diamines at ordinary temperatures.

4.2. Interfacial and Solution Polymerization

These are low-temperature methods for the preparation of polyamides whose stability and/or viscosity prevent melt polymerization. The reaction of an amine and acid chloride to yield aramid-type products is typically involved. The classic text on the subject is that of MORGAN [90]. Reference to the technique can also be found in [91]. (See also, ⟶ High-Performance Fibers – Aramid Fibers. Aramid Fibers and ⟶ Polymerization Processes.)

4.3. Ionic Polymerization

Both anionic and cationic polymerization of anhydrous lactams have been demonstrated. Cationic methods have, however, failed to yield useful polymers because of side reactions that limit attainable molecular mass. Reviews of ionic polymerization have been provided [35], [51], [52].

The anhydrous, *anionic polymerization* of caprolactam has been of interest because the monomer is converted to polymer in minutes. This not only eliminates the need for costly polymerization facilities but also allows for use of inexpensive molds of complex design. It makes low-volume production feasible but is not economical for mass production of small parts. Unless modified, this method yields a polymer with properties somewhat different from hydrolytically polymerized PA 6 because of higher molecular mass and higher crystallinity [92].

Preferred techniques involve the use of the lactam anion, which is made in situ by reaction with a base (catalyst), and a lactam with an electrophilic substituent on the nitrogen atom (the cocatalyst). A very large number of publications and patents deal with the choice of catalyst and cocatalyst and their implication for advantages in process or properties. Examples of catalysts are metal hydrides, oxides, alkoxides, amides, carbonates, and Grignard reagents; cocatalysts are *N*-acyl caprolactams or compounds that form such derivatives in situ, e.g., anhydrides and isocyanates. Patents in the 1980s (e.g., [93-102]) have focused more on fillers, tougheners, applications, and process variants than on novel catalyst systems.

The mechanism of anionic polymerization is depicted below (M = metal, B = base, R = alkyl or aryl):

Anion formation:

$$\underbrace{HN(CH_2)_5C{=}O}_{\mathbf{1}} + MB \rightleftharpoons M^+\ {}^-\underbrace{N(CH_2)_5C{=}O}_{\mathbf{2}} + HB \qquad (18)$$

Initiation:

$$\mathbf{2} + RCO\underbrace{N(CH_2)_5C{=}O}_{} \rightleftharpoons$$

$$RCO\underbrace{N(CH_2)_5\overset{\overset{O^-}{|}}{C}}_{\mathbf{3}}{-}\underbrace{N(CH_2)_5C{=}O}_{} \qquad (19)$$

$$\mathbf{3} \rightleftharpoons RCON^-(CH_2)_5CO\underbrace{N(CH_2)_5C{=}O}_{\mathbf{4}} \qquad (20)$$

Anion regeneration:

$$\mathbf{1} + \mathbf{4} \rightleftharpoons \mathbf{2} + RCONH(CH_2)_5CO\underbrace{N(CH_2)_5C{=}O}_{\mathbf{5}} \qquad (21)$$

Propagation:

$$\mathbf{2} + \mathbf{5} \rightleftharpoons \text{as in Equations (19) and (20)} \rightleftharpoons$$
$$RCONH(CH_2)_5CON^-(CH_2)_5CO\underbrace{N(CH_2)_5C{=}O}_{} \qquad (22)$$

SEKIGUCHI [52] prefers a mechanism that requires reaction of **1** with **3** to yield **2** and **5**. As he points out, there are many possible reaction pathways and side reactions, some of which arise from the increased probability of mid-chain ion formation as polymerization proceeds and the ratio of polymeric chain to lactam monomer increases.

Sample temperature is plotted versus time and density versus temperature for cocatalyzed, anionic polymerization of caprolactam in Figures 6 and 7, respectively. The polymer crystallizes during the polymerization since the maximum temperature of the system is lower than the melting temperature of the polymer. The maximum temperature of the system is caused by the heat of polymerization and the heat of crystallization, and is thus higher than the polymerization temperature which can be calculated from the heat of polymerization under adiabatic conditions. Anionically polymerized polycaprolactams are more crystalline than hydrolytically polymerized ones, and thus have higher densities (1.16 versus 1.14 g/cm^3).

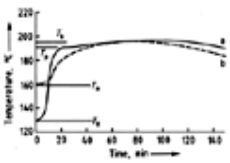

Figure 6. Temperature as a function of time in a typical cast polymerization of nylon 6 [92]

a) Temperature at center of large casting; b) Temperature of mold

T_o = initial temperature of polymerization (melt); T_p = final temperature of polymerization; T_k = maximum temperature; T_m = initial temperature of mold

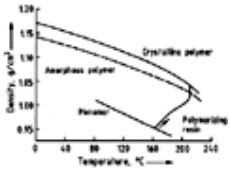

Figure 7. Density of caprolactam monomer and polymer as a function of temperature [92]

A toughened, anionically polymerized caprolactam adaptable to reaction injection molding (RIM) technology was developed by Monsanto. Reaction of a bislactam with a soft polyether terminated by hydroxyl groups, e.g., a poly (propylene oxide), in the presence of caprolactam and a magnesium Grignard reagent yields a bisacylated prepolymer which, on further reaction with lactam, yields the polymeric product [103-105].

PA 12 has been made by cocatalyzed, anionic polymerization [106]. Other nylons (e.g., PA 4, PA 5, and PA 7) have been similarly prepared but have not been commercialized. Successful polymerization of PA 4 and PA 5 has, however, been of considerable theoretical interest because these nylons are not otherwise available. PA 4 has received a great deal of attention and has been commercialized in fiber applications. However, its thermal instability makes it unsatisfactory for molding and extrusion applications [35, Vol. II, pp. 74–117]. The best technique for the synthesis of PA 4 involves the use of carbon dioxide in place of an electrophilically substituted lactam, but it has been argued that this is not a cocatalyzed

polymerization because the carbon dioxide cannot be found in the polymer [52].

4.4. Solid-Phase Polymerization

Solid-phase or solid-state polymerization (SPP or SSP) was first applied to low molecular mass polyamides by FLORY [107]. Applicability to a variety of polyamides with melting points ranging from about 190 to 290 °C was indicated. SPP of diamine – diacid salts has been discussed [108-110] but has not yet been commercialized. Highly crystalline PA 11 has been made in the laboratory by SPP of 11-aminoundecanoic acid [111]. MONROE [112] described a fluidized-bed process for increasing the molecular mass of polyamide particles not over 10 mm in diameter from an initial 1000 to 15 000 to over 15 000. BEATON [113] claimed a procedure which avoided agglomeration and permitted higher temperatures in a nonagitated bed of polymer. The rate of polymerization depends on particle size, temperature, and exposure to an inert gas maintained at a low relative humidity. As noted in [114], only the polycondensation reaction has been studied. SPP yielded higher molecular masses at equilibrium than expected for the water vapor pressure present [115]. This can be explained on the basis that the end groups and the absorbed water exist outside the crystalline regions. A broader than normal molecular mass distribution was also observed. Similarly, a transient distribution had been noted after the SPP of PA 12 wherein return to the equilibrium value ($\bar{M}_w/\bar{M}_n \approx 2$) occurred with a lowering of the weight-average molecular mass (but not of the degree of conversion) when the SPP product was melted [116]. The SPP of powdery 6-prepolymer was said to depend on the diffusion of acid ends and the initial molecular mass [117]. It is of interest that the number of bonds needed for the reported increases in molecular mass in both [112] and [117] is roughly proportional to the time required. Prior removal of oligomers facilitates SPP of PA 6 [118], [119].

The most significant application of SPP is in the synthesis of PA 46. A product with a narrow molecular mass distribution was obtained by SPP of 46-salt at 290 °C [120]. Prepolymerization in the presence of water followed by SPP yields a commercial quality product [121], [122]. Second-order kinetics are said to apply [121]. Special technique is required for PA 46 to control the tendency of the tetramethylenediamine to form a pyrrolidine ring which then acts as a monofunctional chain stopper. The utility of SPP in PA 46 is ascribed to the higher equilibrium constant for amidation at the lower temperature and to the higher activation energy for ring formation than for polymerization, 137 versus 83 kJ/mol [53], [69]. For a comprehensive coverage of SPP refer to [123].

[Top of Page]

5. Commercial Processes

The primary raw material sources of the monomers used to prepare commercial nylons have been listed in Table 2. Commentary on most monomer syntheses can be found in [5-7], [19]. The diamine for PA 46 is made via the addition of hydrogen cyanide to acrylonitrile [124]. For convenience, certain basic parameters of the commercial nylons have been assembled in Table 4.

Table 4. Basic parameters of commercial nylons

PA	CH_2/CONH [a]	M_u [b]	[CONH], $10^6/M_u$ [c]	T_m [d], °C	Density, g/cm^3
46	4	99.13	10 088	295	1.18
66	5	113.16	8 837	269	1.14
69	6.5	134.20	7 452	215	1.08
610	7	141.21	7 082	225	1.08
612	8	155.24	6 442	220	1.07
MXD 6		123.15	8 120	243	1.21
TMDT		144.20	6 935		1.12
11	10	183.30	5 456	190	1.05
6	5	113.16	8 837	228	1.14
12	11	197.30	5 068	180	1.02

[a] Ratio of methylene to amide groups per repeating unit (or average monomer unit).
[b] M_u is the average molecular mass per amide group.
[c] [CONH] is the concentration of amide groups.
[d] The melting temperatures (points) are as determined by the disappearance of the last vestige of birefringence on a hot stage and correspond to the minimum fabricating temperature. Estimates were made where such data were not available.

5.1. PA 46

The 46-salt can be converted into prepolymer at 180 – 240 °C under steam pressure [122], [125]. SPP is carried out at 220 – 275 °C. Variations include a continuous polymerization method involving short exposure at 293 °C [125], water extraction of prepolymer to improve its color [126], a two-stage heat treatment to make prepolymer [127], control of temperature and steam pressure to yield high molecular mass [128], and polymerization with some of the adipic acid replaced by terephthalic acid to give a PA 46/4T copolymer [129].

5.2. PA 66

The first step is preparation of pure, balanced salt in aqueous solution. Stoichiometric equivalence is determined by pH measurement. Some excess diamine is normally added to compensate for losses due to its relative volatility. The equivalence pH for the aliphatic salts approximates 7.6 [130], [131]. Charcoal decolorization of the salt solution is required unless the diamine has been carefully refined and, in particular, is free of *cis*-1,2-diaminocyclohexane [132]. Purging a charcoal bed with inert gas increases its life [133]. Treatment with diacid followed by removal of acid with diamine before use with salt solution avoids initial losses in molecular mass [134]. A two-stage, recycling method of salt preparation is used in a continuous process [135]. Two-step addition to diamine with intermediate evaporation of water yields a concentrated salt solution [136]. Ways of mixing molten acid and diamine have been patented [137-139]. To avoid precipitation when stored at about 25 °C a 50% aqueous salt solution is normally prepared.

5.2.1. *Batch Production*

The stored salt solution is concentrated under pressure to 65 – 80 % before charging to an autoclave. A typical cycle is shown in Figure 8 [140]. The essential features are heating to 210 °C under autogeneous pressure to reach a pressure of 1.75 MPa, gradually increasing the temperature to about 275 °C while releasing steam at a rate which maintains the pressure, reducing the pressure at a rate that avoids cooling and finally holding the batch at atmospheric or reduced pressure to obtain the target molecular mass before extruding the polymer under inert gas pressure. This procedure is designed to assure that there is enough water present to avoid freezing of the batch before the melting point has been reached. Water also minimizes excessive loss of diamine. Stirred autoclaves are used but are normally unnecessary. The extrudate is a wide ribbon that is quenched with water that is subsequently removed by jet blowers. The ribbon is cut into chips which are blended and packaged [141], [142]. Hermetically sealed containers are used to avoid absorption of moisture that can hydrolyze polymers to lower molecular masses during subsequent melt processing. The containers vary from 10 kg bags to 500 kg or larger cartons. Nylon molders and extruders have drying equipment to contend with situations where the moisture content of as-received or reprocessed material requires adjustment.

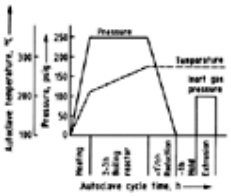

Figure 8. A typical autoclave cycle useful for batch preparation of nylon 66 (Du Pont Technical Laboratory, Seaford, Delaware)

Reproduced with permission of John Wiley and Sons from Figure 12.8 in [140]

5.2.2. *Continuous Production*

The same concerns for control of the rate of removal of water and loss of diamine exist as in batch polymerization, but the situation is complicated by the needs of a continuous process. Typically, a first stage involves evaporation/reaction with controlled loss of water to form a prepolymer and minimize loss of diamine (Fig. 9). Further reaction occurs in subsequent stages with controlled evaporation in devices known as “separators” and “flashers”. The desired molecular mass and water content are obtained in a “finisher”. A large patent literature exists that claims improvement in each of these devices or combinations thereof for process simplification (see the review [140]). An early process that includes an additive port for delusterant is shown in Figure 9 [143]. A modification thereof (Fig. 10) includes a separator after the flash tubes to provide an essentially steam-free melt to the finisher [144]. Further improvements include the addition of a cage structure at the top of the screw section for better flow and reduced gel formation [145], insertion of static mixers in the flash tubes [146], putting holes in the ribbon flight in the separator [147], and providing flow splitters in the flash tubes [148]. A modification between the flash tubes and the separator significantly increases steam release and achievable molecular mass [149]. Dimensions, time, temperature, and pressure vary somewhat with specific design, but the same basic needs outlined above always apply in order to obtain a molten, high molecular mass product of low water content that can be fed to spinners or granulators. Severe constraints such as holdup spots and exposure to oxygen must always be avoided in continuously operated, high-temperature equipment such as polymerizers, extruders, or injection molding machines.

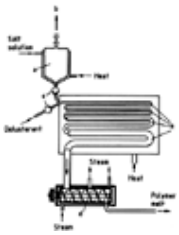

Figure 9. Continuous polymerizer for PA 66 [143]

a) Evaporator/reactor; b) Vent; c) Pump; d) Finisher; e) Flash tubes

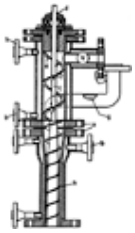

Figure 10. Two-part separator for polymerizer depicted in Figure 9 [144]

a) Flanges joining upper and lower members; b) Part of heating system; c) Shaft; d) Journal bearing; e) Ribbon flight; f) Spokes attaching ribbon to shaft; g) Feed from flash tubes; h) Transfer screw feeding melt to finisher

Another design (Fig. 11) involves essentially adiabatic evaporation in a second zone by directing the flow from an initial

evaporator/reactor through a perforated ring nozzle against a heat exchanger with a surface of high thermal conductivity [150]. Another device emphasizing adiabatic evaporation makes use of the observation that the presence of low molecular mass polymer in salt solution decreases the tendency of the salt to precipitate [151], and another resorts to a multiplicity of zones [152]. Other procedures include rectification columns in the first zone, e.g., [153], [154]; a one-stage process with a long, narrow, coiled tube emptying into a stirred melt pool [155]; modified screw finishers [156-158]; designs to provide thin films of melt [159-161]; an apparatus that uses a dry salt charge [162]; and a dispersion technique for high-viscosity polymers involving cyclic monomers [163].

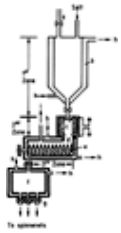

Figure 11. Polymerizer with essentially adiabatic evaporation in second zone [150]

a) Pressurized vent; b) Evaporator/reactor; c) Ring nozzle; d) Heat exchanger with surface of high thermal conductivity; e) Finisher; f) Heat exchanger; g) Gear pumps; h) Heat exchanger fluid; i) Exit tube to atmospheric vent or vacuum

5.3. Other AABB Polyamides

The polymerizations yielding PA 69, PA 610 and PA 612 are subject to the same concerns as that yielding PA 66. These polyamides are made in much the same way with differences due to different melt viscosities and melting points. The choice between continuous or batch process depends mostly on production volume. Because of smaller sales volume, the batch method is favored for these specialty nylons.

The AABB principles described above can easily be extended to monomers containing an aromatic or cycloaliphatic moiety in the AA, the BB, or both portions of the molecule. Because of the presence of these structures these polyamides present special problems due in part to their high viscosity and/or stability at the high processing temperatures. An example is PA MXD6, a polyamide prepared from *m*-xylylenediamine (MXD) and adipic acid. Careful control of steam pressure in the early stage of polymerization of MXD6 is required to minimize decomposition of the aromatic diamine [164]. In the production of TMDT polymer, the salt is precipitated from an aqueous solution with isopropyl alcohol, partially condensed in a stirred reactor, and finally polymerized to a number-average molecular mass of about 20 000 in an extruder [165].

An emerging area of interest is in the polyphthalamides (PPAs), also known as high temperature nylons (HTNs). These are high melting semiaromatic nylons resulting from the addition of an aromatic ring to the polymer backbone. The high polymerization temperatures required for their manufacture necessitates either a short residence time in the melt by using and extruder or polymerization below their melting points (see Section Solid-Phase Polymerization) [166], [167].

5.4. PA 6

Continuous processes are used by the major manufacturers of PA 6. The so-called VK tube (Vereinfacht Kontinuierlich = simplified continuous) was developed in Germany. It is a vertical tube operated at atmospheric pressure wherein heating and prepolymerization take place in the upper part and polymer is formed in the lower section [168]. Problems of initial stoichiometry and ingredient volatilization, inherent with AABB polymerization as already discussed, do not apply; however, the need to free the product of its equilibrium content of monomer and oligomers presents a different kind of problem. A schematic of the BASF process [6] shows a VK tube feeding a pelletizer followed by a water extraction unit (Fig. 12). A process [169] involving vacuum extraction is shown in Figure 13. Claim is made that controlling initial conversion to 45 % yields a product with less than 2 % of cyclic oligomers with a vacuum finish that does not require prolonged heating at less than 665 Pa [170]. Improved processes include a baffled hydrolyzer operating in plug flow and at a carefully controlled temperature [171] and a two-stage hydrolyzer [172].

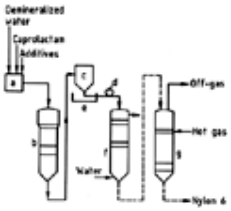

Figure 12. Production of nylon 6 by BASF

Reproduced with permission of Elsevier Science from Figure 5 in [6].

a) Feed tank; b) VK tube; c) Pourer; d) Pelletizer; e) Water bath; f) Extractor; g) Dryer

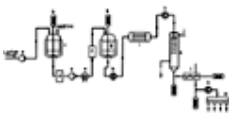

Figure 13. Production of nylon 6 by Allied Chemical [169]

a) Pump; b) Stirrer; c) Holding tank; d) Filter; e) Flowmeter; f) Preheater; g) Hydrolyzer; h) Metering pump; i) Polyaddition reactor; j) Vent; k) Vacuum flasher; l) Finisher; m) Spinning heads

The simulation and optimization of the hydrolytic polymerization of caprolactam have been attempted [58], [173], [174]. Lack of detailed manufacturing data has been indicated to prevent confirmation of the analysis.

Polymerization in an extruder has also been discussed [175]. Numerous publications and patents continue to appear that deal with the anionic polymerization of caprolactam. No large-scale applications exist as yet although modified systems in reaction injection molding (RIM) for the production of block copolyesteramides and copolyetheramides have been

developed by several companies [35]. Some details on recommended RIM procedures such as separation of feed streams, mold design, and temperature have been published [176], [177].

5.5. PA 11

PA 11 is made in a continuous process that is operated under inert gas at atmospheric pressure [178]. The monomer (11-aminoundecanoic acid, which is made from castor beans) is suspended in water and mixed with the desired additives (antioxidants, etc.) before charging to a three-zone, upright converter. The monomer is melted, water eliminated, and the polymerization begun in the top section. Near completion of the condensation occurs in the middle zone. Adjustment to the target molecular mass takes place in the bottom section which provides a melt pool for uniform discharge of product. The good thermal stability of the polymer and the low tendency of the monomer for cyclization simplify the process design.

Phosphoric acid catalyzes the polymerization as does hypophosphoric acid. The latter can, however, insolubilize the product if used in excess with prolonged heating [179].

5.6. PA 12

Polylaurolactam, PA 12, is made by heating dodecanolactam (which is made from butadiene) with an acid catalyst such as phosphoric acid in the presence of water to effect conversion to a prepolymer, then evaporating the water, and finally heating in an inert atmosphere to complete the polymerization [35], [180]. Temperatures up to 350 °C may be used to effect completion [181], [182]. Confirmation of the independence of reaction rate on molecular mass and of a most probable molecular mass distribution together with an expression relating degree of polymerization to viscosity in *m*-cresol are provided in [183].

[Top of Page]

6. Properties

The enduring utility of nylons rests upon their combination of properties and upon their susceptibility to modification. Key properties are resistance to oils and solvents; toughness; fatigue and abrasion resistance; low friction and creep; high strength; stability at elevated temperatures; fire resistance; drawability; good appearance; and good processibility.

Polyamides are readily modified in order to achieve desired changes in processing behavior or properties. This characteristic has allowed them to sustain their growth in spite of the advent of challenging new materials. However, modifications to enhance some specific properties can have negative effects on other properties. Basic modifications can be accomplished through changes in the polymer molecule itself and/or the incorporation of additives. Alternative techniques of modification include use of branching agents (see Section Control of Rheological Properties.) and post-polymerization treatments such as annealing, conditioning to a specific moisture level, dyeing, metallizing, painting, irradiation, or chemical reaction such as alkoxyalkylation.

6.1. Properties of Unmodified Polyamides

A problem common to most polymers is the lack of adequate and comparable data. Nylons are no exception; however, the increasing global acceptance of ISO protocols in measuring properties has alleviated the situation. A limited comparison of unmodified nylon resins is provided in Table 6. Even these data are not unambiguous because the molecular masses are not specified. The table is self-evident and reflects the differences in amide group concentration and crystallinity. Information on literally hundreds of nylon compositions can be obtained from various web sites maintained by the suppliers and in the Campus database [184]. Additional information is also available from the *Plastics Technology Handbook* and *Modern Plastics Encyclopedia*, which are both issued annually.

This introductory section, however, is concerned with those features of neat polyamides that determine their properties and account for differences between individual nylons. The most distinguishing feature is the amide group, which is responsible for strong hydrogen bonding between adjacent chains. Melting point increases with increasing ratio of CONH groups to CH_2 groups in the chain and the increased opportunity for hydrogen bonding, but not regularly because whether the number of CH_2 groups between the CONH groups is odd or even is important as well as the ratio of CH_2 to CONH. The odd number of CH_2 groups between the amide groups in PA 6 allows complete hydrogen bonding when the amides in adjacent chains have an opposed or antiparallel orientation but not when they have the same or parallel orientation (Fig. 14). Changing from a parallel to an antiparallel array requires inverting the entire molecular chain in this odd-numbered case, but only a one-segment lateral movement is needed if there is an even number of CH_2 groups, as in the case of PA 66 with its intervening number of 4 and 6 CH_2 groups. It is believed that this odd/even feature accounts for the lower melting point and percent crystallinity of PA 6 versus PA 66. It is the reason in general why PA-odd and PA-odd/even have lower melting points than comparable or similar PA-even/even, such as PA 6 versus PA 7.

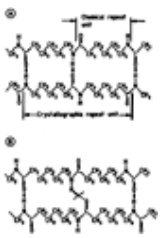

Figure 14. Schematic of a hydrogen bonded sheet of PA 6 with antiparallel (A) and parallel (B) orientation of amide groups

Increasing crystallinity results in higher stiffness, density, tensile and yield stress, chemical and abrasion resistance, and

better dimensional stability. It decreases elongation, impact resistance, thermal expansion, and permeability. Crystallinity is responsible for the development of microscopically observable structures, known as spherulites, that scatter light and make nylons opalescent unless very thin. The transparent nylons are amorphous and are made so by appropriate copolymerization involving at least in part aromatic or cycloaliphatic AB, AA, or BB monomers in order to have acceptable stiffness.

Orientation [5, p. 300] contributes importantly to properties not only in film, fiber, and strapping but also in injection-molded articles. The resulting anisotropy means that properties vary with the direction of stress and must be considered in rationalizing resin behavior.

Water absorption is characteristic of nylons. Unless compensated for by increased crystallinity, a higher proportion of amide groups leads to higher water adsorption. Increased water content has an effect analogous to that of increased temperature, i.e., enhanced segmental mobility with, for example, concomitant loss in stiffness and tensile strength, gain in toughness, and growth in dimensions (elongation). At very low temperatures, however, water stiffens the nylon. Thus, the brittleness temperature (ASTM D 746) of PA 66 is – 80 °C if dry and – 65 °C if conditioned to 50 % relative humidity. Properties are frequently reported in the “dry”, as-molded, condition corresponding to about 0.2 % water or less, and after equilibration to a specified relative humidity such as 50 or 65 %, and occasionally to 100 %. The greatest change occurs in the vicinity of the glass transition temperature (T_g) so that a useful aid in understanding behavior is a knowledge of the effect of humidity on T_g. This is complicated because of variation with the method of calculation or measurement (Table 5), but it is generally true that nylons with fewer CONH groups and lower water adsorption have a lower dry T_g but show less change of T_g with relative humidity [19], [121], [185-190].

Table 5. Glass transition temperatures (T_g, °C) of polyamides at various relative humidities (R.H.) [*]

	Dry				50 % R.H.		100 % R.H.	
PA	**A**	**B**	**C**	**D**	**C**	**D**	**C**	**D**
46	102	78	78					
66	82	65	80, 78, 66	48	35	15	– 15	– 37
6 (extracted)	56	65	75, 65	41	20	3	– 22	– 32
610	56	46	67, 70	42	40	10		
612	52	40	60	45	40	20	20	
11	36	29	53	43				
12	29	25	54	42			42	
MXD 6	71		68					
TMDT			150					

[*] A) Estimated from melting temperatures using T_g = (2/3) T_m [185]; B) Calculated from group contributions [186]; C) Dynamic measurements with torsion pendulum [121], [187-189]; D) Static measurements by differential thermal analysis or inflection point in curve of modulus of elasticity as a function of temperature [19], [190].

Table 6. Properties of unmodified polyamides [*]

Property	**ASTM**	**66**		**6**		**612**		**11**		**12**	
		Dry	**50 % R.H.**	**Dry**	**50 % R.H.**	**Dry**	**50 % R.H.**	**Dry**	**50 % R.H.**	**Dry**	**50 % R.H.**
Tensile strength, MPa	D 638	83	77	81	69	61	61	57	54	49	47
Yield strength, MPa	D 638	83	59	81	44	61	51			49	40
Ultimate elongation, %	D 638	60	≥ 300	200	300	15	30	120	300	250	250
Elongation at yield, %	D 638	5	25	9		7	40			10	20
Flexural modulus, MPa	D 790	2830	1210	2700	970	2030	1240	1170	1030	1410	1030
Izod impact strength, J/m	D 256	53	112	58	215	53	75	40		58	64
Water absorption, %	D 570										
		1.5		1.6		0.4		0.25		0.25	

24 h								
50 % R.H.		2.5		2.7	1.3		0.8	0.7
100 % R.H.		8.5		9.5	3.0		ca. 2.4	ca. 2.0
Heat deflection temperature, °C	D 638							
455 kPa		235		185	180		150	150
1820 kPa		90		65	90		55	52
Dielectric constant	D 150 (1 kHz)	3.9	7.0	3.8	4.0	5.3	3.7	3.6
Flammability	UL 94	V 2		HB	HB		HB	HB

* Data assembled from publications of various manufacturers for standard injection molding grades. Variations in molecular mass contribute to differences in properties.

Little change in sensitivity to aliphatic hydrocarbons occurs with a change in polyamide type, but aromatic hydrocarbons are more strongly absorbed by polyamides with fewer amide groups. For example, PA 66 and PA 6 absorb 1 % toluene; PA 610, PA 3, and PA 11 absorb 6.8 %.

With this background now in place, a comparison of the properties of various nylons can be made, moving from the least aliphatic, to the most aliphatic, and then to semiaromatic nylons.

PA 46 is structurally similar to PA 66, and is produced from 1,4-diaminobutane and adipic acid. This nylon has the highest amide-to-methylene group ratio and the highest crystallinity of the commercial nylons. The high amide concentration accounts for the high moisture absorption of the polymer. Its melting point is 290 °C.

PA 66, which has one amide group for every 4 or 6 (with an average of 5) methylenes, has a high melting point of 269 °C, good strength and creep resistance, and excellent gas barrier properties. These qualities have led to its widespread use in automotive mechanical and electrical and electronic applications.

PA 6 also has an amide to methylene ratio of 1:5, but the odd number of methylene units retards crystallization and results in a more amorphous polymer with a lower melting point of 228 °C. These characteristics can be desirable for extruded film and wire and cable jacketing, but make PA 6 somewhat more sensitive to moisture and chemicals and limits its resistance to temperature spikes.

PA 69, 610 and 612 are all produced from hexamethylenediamine plus azelaic, sebacic, or dodecanedioic acids, respectively. PA 612 is intermediate between 6/66 and 11/12 with an average amide to methylene group ratio of 1 : 8, and provides an excellent compromise between PA 66 and higher nylons, with its excellent dimensional stability, reasonably high melting point (212 °C), strength, creep resistance, and barrier properties. This balance leads to widespread use in toothbrush filaments, battery components, and parts that are exposed to hot aqueous environments.

Higher nylons, like PA 11 and 12, have much lower amide to methylene group ratios (1:10 and 1:11 respectively), which results in more polyethylene-like materials that have excellent dimensional stability, low-temperature toughness, and stress-crack resistance at the expense of lower melting point and strength. They are readily plasticized for use in flexible, chemically-resistant tubing that has high burst pressures. The long aliphatic chains impart a low affinity for moisture. In addition, PA 11 is available in powder grades that may be applied by electrostatic spray or in fluidized beds for coating metallic parts.

Semiaromatic nylons are generally copolymers that are commonly made using terephthalic acid and/or isophthalic acid as comonomers often with an aliphatic or alicyclic acid. Common semiaromatic nylons will contain units derived from hexamethylenediamine and adipic acid or caprolactam. They typically have high melting points (295–325 °C for commercial grades) and exhibit a significant decrease in moisture absorption when compared to purely aliphatic nylons. The aromatic groups in the polymer backbone make them less mobile than linear aliphatic nylons and contribute to their high stiffness and good strength, creep resistance, and dimensional stability. Biaxially oriented films made from semiaromatic polymers provide excellent gas and oxygen barrier properties allowing for use in packaging materials and bottles.

Control of Rheological Properties. Control of melt rheology is very important for nylon processing (see Chap. Processing). Low viscosity is desired for molding application for easy mold-filling. On the other hand, a high viscosity and melt strength are desired in extrusion and blow-molding applications. Control of molecular mass is the most convenient method for controlling rheological properties. Most polyamides have number-average molecular masses in the range of 11 000 to 34 000. Toughness increases with increasing molecular mass, but the effect is a moderate one. The major practical reason to change the molecular mass of a polyamide is to obtain a desired level of melt viscosity, which can range from less than 1000 Pa · s for injection molding to over 4000 Pa · s for shaped extrusion.

Another technique used to modify rheological properties is to change the molecular architecture. Introduction of low-level branching results in more non-Newtonian behavior, especially at low shear rates where an increase in melt strength and melt viscosity is obtained. Star nylons (see Section Molecular Mass), although known for a number of years, are a new class of commercially available semicrystalline polyamides that exhibit a quasi-Newtonian rheological behavior and

enhanced flow properties during processing [191]. The polymer is a modified PA 6 and has a star-like structure that is based on proprietary polymerization techniques. Polymerization catalysts and chemicals incorporated into the polymerization cycle produce a nylon with controlled branches. These materials can be compounded with large amounts of glass fibers or mineral fillers and molded with increased melt flow under normal processing conditions. The enhanced flow characteristics promote an excellent surface finish and aesthetic appearance. Commercial uses now include automotive engine covers, air-intake manifolds and trim parts, fishing reels, and appliance doors [192].

6.2. Copolymerization

Copolymerization to modify the properties of polyamides has long been practiced. This technique allows one to obtain varied properties without resorting to developing new polyamides from new monomers. Copolymerization increases the disorder in the polymer molecules and interferes with the ability of the molecules to pack and form hydrogen bonds. As a consequence, copolymers in general are characterized by a lower degree of crystallinity and a slower rate of crystallization, greater transparency, a lower modulus and strength, and greater solubility than the respective homopolymers. In a PA 66/6 copolyamide, for example, the melting point minimum is about 180 °C at 40 wt % of caprolactam units. PA 66/6 copolymers with about 40–65 wt % caprolactam units become soluble in hot methanol whereas the homopolymers PA 6 and PA 66 are insoluble in this solvent [19], [193], [194]. Terpolymers such as PA 66/6/610 are designed to further enhance solubility for use as binders [19], [195]. Transparency with good retention of properties is achieved by use of alicyclic or aromatic comonomers as has been noted for PA TMDT. A review of transparent polyamides that includes a comparison with other transparent engineering plastics is given in [196], [11, pp. 377–387].

There are quite a number of pairs of monomers, however, that have comparable distances between their functional groups and that yield copolymers in which the amide moieties can assume the proper orientation for hydrogen bonding. These molecular characteristics permit the comonomers to crystallize in the same morphology [197-199]. Instead of being depressed, the melting points of these isomorphous copolymers are weighted averages of the melting points of the component homopolymers. Some examples of isomorphous pairs of monomers are adipic and terephthalic acids; *p*-benzenedipropionic and sebacic acids; and *p*-xylylenediamine and *m*-xylylenediamine. For a more extensive discussion refer to [[11], pp. 370–372].

6.3. Modification by Additives

Additives are materials used in small amounts (less than 5 wt % and usually less than 1 wt %) to affect various aspects of polymer processing, polymer properties, and/or appearance. Additives used to alter properties or appearance include antioxidants, antistatic agents, biodegradative agents, biopreservatives, blowing agents, colorants, fragrances, and stabilizers against hydrolysis, thermal degradation, or UV degradation. Processing additives include coloration inhibitors, lubricants, mold-release agents, nucleating agents, and viscosity thickeners or reducers. They have little effect on the physical properties except in the instance of nucleating agents, which not only increase the rate of crystallization and the degree of crystallinity, but impart reduced mold shrinkage and increased tensile and flexural properties. Enhanced crystallization may be desired to shorten molding cycles, and also to modify mechanical properties.

Modifiers are used in larger amounts, usually more than 5 wt %, in order to achieve desired changes in properties. Modifiers include mineral fillers, glass or carbon fibers, lubricants to improve wear and friction, plasticizers, fire retardants, electrically conductive materials, and/or blends with other polymers to toughen or otherwise affect the nylon. A broad discussion of additives is given in [11, pp. 361–448], [200].

6.3.1. *Filled and Reinforced Polyamides*

Fillers and reinforcing agents have contributed significantly to the sustained growth in the use of nylons. Particulate mineral fillers (e.g., aluminum or calcium silicates) always increase the modulus of elasticity and often the tensile strength whereas elongation always decreases. Impact strength as measured by the notched Izod test may or may not improve; it depends on the choice of filler and matrix. Temperature resistance is increased.

Reinforcement with glass and carbon fibers has similar but larger effects with consistently improved Izod impact strengths because of the fiber acicularity. Warpage can be a factor in some applications. This again relates to fiber acicularity and is, accordingly, less in the typical mineral-filled compositions. Mixtures of fillers and reinforcing agents are used, therefore, to cope with such situations. Important factors are the efficacy of the bonding agent used to promote adhesion between the filler or the reinforcing agent and the polyamide matrix and the length to diameter ratio of the fiber in the final product. The properties in Table 7 were chosen to illustrate these effects but because of the number of variables involved they should not be construed as limiting.

Table 7. Selected properties of dry, filled/reinforced PA 66 [*]

Property	Neat	40 % Mineral[**]	33 % Glass fiber	Mineral and glass fiber	30 % Carbon fiber
Tensile strength, MPa (ASTM D 638)	83	89	186	130	241
Ultimate elongation, % (ASTM D 638)	60	17	3	12	
Flexural modulus, MPa (ASTM D 790)	2830	5240	8960	6900	20 000
Izod impact strength, J/m (ASTM D 256)	53	70	117	48	80
Heat deflection temperature,°C					

(ASTM D 638)					
at 455 kPa	235	230	260	257	
at 1820 kPa	90	185	255	239	257
Density, g/cm^3	1.14	1.51	1.38	1.45	1.28

* Data from publications of various manufacturers.
** "Mineral" as found in the nylon literature normally refers to various silicates and carbonates.

Long glass fibers are used in polyamide compositions with the attendant implications for properties and processing [201], [202]. Most of these compositions are produced by a pultrusion process [203], [204]. The continuous glass fibers are pulled through an impregnation line where they are coated with nylon resin. Upon cooling, the resin is chopped into long pellets that contain reinforcing fibers equal in length to the pellet. Traditional extrusion compounding of glass into nylon yields much shorter fibers, and hence less stiffness and worse transverse mechanical properties. The balance of properties makes long-fiber reinforced compositions ideal for the replacement of metals such as die-cast aluminum in many parts. In specialty applications, long fibers of carbon, aramids, and steel are used in addition to glass [205].

Nanocomposites are emerging with increased acceptance of the properties that they offer. They can improve strength, stiffness, heat deflection temperature, dimensional stability, gas barrier, flame retardancy, and electrical conductivity. Nanoparticle platelets are only 1 nm thick with an aspect ratio (*L/D*) sufficiently high that relatively low loadings (generally 2 to 6 wt %) provide significant property advantages. These clay-based additives have been used with PA 6, PA 612, and PA MXD 6 to improve stiffness heat resistance, oxygen barrier and gloss, but confer a disproportional loss of toughness. In 2000 these nylon compositions were first commercialized for container and packaging applications [206-208]. A more detailed discussion of the many aspects of nanotechnology is given in [209].

6.3.2. *Blends and Alloys*

Blends of nylon (and polymers in general) are mixtures that contain at least one other polymer or copolymer. The properties of the blends can be additive, averaged, synergistic, or antagonistic. An alloy is an immiscible polymer blend having a modified interface between both components and/or morphology. While positive synergism of the components is most desired, a weighted averaging of properties based on the volume component of resins is usually the rule. Nature rarely provides something for nothing so there are often tradeoffs of properties and for every advantage obtained by a resin blend, there are often disadvantages as well. The disadvantages are often so significant that the resulting material is unusable. The primary reason for the negative results is that most polymer pairs are immiscible because of high interfacial tension, which results in phase-separated components. A brief list of some common property tradeoffs is:

Stiffness, strength (e.g., glass, mineral)	vs. toughness, ductility
Melt flow	vs. toughness, elongation
Fire retardance	vs. environmental considerations, toughness, melt stability
Solvent resistance (i.e. crystallinity)	vs. mold shrinkage
Cycle time	vs. shrinkage
Softness	vs. creep, stiffness
Product complexity	vs. compounding yield, speed, cost

The successful use of blending to tailor-make materials for specific end use applications depends on the type of resins used, their morphologies, the relative levels of the resins, and the property tradeoffs that are acceptable [210-212].

Some of the polymers, that have been successfully blended with polyamides include polypropylene (PP), high impact polystyrene (HIPS), styrene-acrylonitrile (SAN) and acrylonitrile-butadiene- styrene (ABS) copolymers, poly(methyl methacrylate) (PMMA), styrene- maleic anhydride (SMA) copolymer, EPDM elastomers, polycarbonate (PC), and poly (phenylene ether) (PPE) [213].

For all blends, it is essential to develop processes that control the morphology and the interfaces of the phase-separated components. Interfaces can be modified by the addition of reactive and nonreactive compatiblizers, as well as by reactive extrusion to assure a controlled chemical reaction between the components and thus compatibilize the blend in situ [214].

The broad utility of nylons was restricted for many years by their notch sensitivity and loss of toughness at very low temperatures. For example, the impact strength of PA 66 at 23 °C and 50 % humidity is 37 J/m with a notch radius of 0.05 mm, 112 J/m with a 0.25 mm radius (the Izod specification of ASTM D 256), and 366 J/m with a 0.75 mm radius. At –40 °C with the standard radius of 0.25 mm, it is 27 J/m. Some improvements were made, but no major breakthrough occurred until suitably defined elastomeric modifiers were used [215]. This work yielded compositions that were ductile and exhibited Izod impact strengths over 534 J/m even when dry. Independence of notch radius down to 0.0025 mm; insensitivity to part thickness, orientation, proximity to the molding gate or rate of loading; and improvement down to about –20 °C were demonstrated [216]. One requirement in this work was a sufficiently small particle size of the dispersed elastomeric phase. Ductile behavior independent of part thickness has been shown for a PA 6–ABS blend (about 50/50 weight ratio) with a cocontinuous phase morphology [217]. Another variant of toughener involves a core – shell structure in the dispersed phase [218]. The effect of orientation, position in the molded part, rate of loading and temperature has not yet been shown for these alternative materials. Many studies on toughening nylons from both industrial laboratories and

academia have appeared. Morphology, type of toughener, testing with sophisticated instrumentation, and analysis of the mechanism of toughening have been studied. It is, however, beyond the scope of this report to delve into this broad and complex subject.

The number of impacts to failure of toughened PA 66 increases, as would be expected, as the impact energy is lowered below that required for failure in a single blow [219]. Noteworthy, however, is the resistance to repeated impact of neat PA 66 [219], [220].

Mineral-filled and glass-reinforced nylons have also been toughened. For example, a dry, mineral-filled product with a tensile strength (TS) of 89 MPa, a flexural modulus (FM) of 5300 MPa, and a notched Izod impact strength (Iz) of 69 J/m was toughened to yield one with corresponding values of 79 MPa, 4600 MPa, and 130 J/m [221]. Similarly, a dry, glass-reinforced PA 66 with a TS of 200 MPa, a FM of 8960 MPa, and an Iz of 100 J/m was toughened to yield a TS of 140 MPa, a FM of 7000 MPa and an Iz of 250 J/m [221]. These examples illustrate the feasability of the wide spread practice combining various additives and modifiers to meet the demands of specific applications.

6.3.3. *Flame-Retardant Polyamides*

Many nylons are rated UL 94 V-2, and fire-retardant grades with a UL 94 rating of V-0 in thin sections are available. For a discussion of UL ratings, see ⟶. Common flame retardants are highly-halogenated organic compounds and polymers such as brominated polystyrenes, poly(dibromostyrenes), or dechlorane [222], [223]. These are often used in combination with a synergist such as an antimony oxide. Red phosphorus is also used as a flame retardant [224]. Concerns about the toxic fumes released when these materials are heated [225] and the hazards involved in handling red phosphorus have led to the development of halogen- and elemental-phosphorus-free flame retardants based on melamine cyanurates that function by generating molecular nitrogen upon decomposition, which smothers the flame [226], [227]. More recently, flame retardants comprising a mixture of phosphinates and melamine cyanurates or melamine phosphates have been introduced [228].

Flame-retardant polyamides have many automotive, electrical, and electronics applications. One survey discusses control of tracking (formation of a conducting path on the polymer surface) [229]. Another review addresses properties of glass-reinforced thermoplastics including PA 6, PA 66, and PA 610 [230]. Colorants can increase flammability [231]. Safety is increasingly stressed and is generating even greater concern for smoke suppression and the toxicity of gases formed during combustion as well as the inhibition of ignition and flame support [225].

6.3.4. *Conductive Polyamides*

The burgeoning market for electronics has led to a great demand for electrically conductive polymers. Electronic devices generate radio frequency radiation during operation, and to prevent the consequent electromagnetic interference (EMI) with other devices, it is necessary to shield the source with an electrically conductive material. Since polyamides are not intrinsically conductive, EMI shielding of nylon parts can be accomplished in either a two-step process that involves coating the plastic with a conductive material, or, preferably, filling it with a conductive additive, such as carbon fibers, metal-coated carbon or glass fibers, metal fibers, carbon black, or graphitic carbon fibrils [232], [233].

It should be noted that neat and glass-reinforced nylons have also been successfully used for some time in electrical power engineering and electronics [234]. A variety of plastics containing metal fillers or carbon to make them electrically conductive, including PA 66, were exposed to an aqueous detergent solution at 25 and 80 °C and to isopropyl alcohol; the effect on EMI shielding was assessed [235]. The compositions with carbon generally outperformed those with metal fillers. The only PA 66 evaluated contained carbon fibers, and its retention of shielding was equivalent to that of the other materials tested.

Unadulterated polyamides are *thermal insulators*, but when combined with heat-conductive additives such as graphite fibers or ceramics like boron nitride and aluminum nitride, they can be used in heat-removal applications that include heat sinks for electronics and electrical applications [236].

6.3.5. *Other Formulations*

The amount of polyamide used and the nature of its dispersion in a blend with other polymers may be such that it is properly defined as the modifier rather than the base resin. One commercial example is a blend of nylon and poly(2,6-dimethyl-*p*-phenylene oxide). Allusion has been made to a blend with poly(phenylene sulfide) as the major ingredient [237]. Another composition presumably involves mostly ABS and some nylon [238]. A developmental material is a glass-reinforced blend of polycarbonate and nylon [239]. The literature on blends is voluminous. Brief reviews of all commercial blends including some commentary on their processing are given in [237-239]. A recent text focuses specifically on commercial blends and alloys [240]. It may reflect success in finding a compatibilizing agent for essentially incompatible polymers.

A different kind of product in which nylon is the principal component arises from reaction during molding of pellets that differ in chemical constitution; at least one component is caused to form a cross-linked structure. In one commercial case a pseudo- or semi-interpenetrating network (SIPN) results which physically binds a silicone modifier to a nylon matrix [241].

With the appropriate additives, polyamides can be used in a variety of specialized applications. For example, polyamides can be combined with magnetic materials such as neodymium-iron-boron alloy powders, permitting the injection molding of magnetic parts [242].

High-density polyamide formulations can be obtained by compounding with metal fillers and powders such as aluminum, zinc, steel, tungsten, or lead. The densities of the materials can be fine-tuned by using metals of different densities and varying their loadings. Addition levels of up to 95 wt % have been achieved. However, the polymer can still comprise 40–70 % of the volume, which permits these compositions to be processed using standard equipment. Wear and abrasion on extruder barrels, screws, and molds is an issue and hardened steel tooling is recommended. These polyamide metal

materials can serve as substitutes for metal in both practical (such as for X-ray and radiation shielding applications, fishing-line sinkers, and bullets) and aesthetic applications. In the former, the substitution of moldable plastics for metal parts permits a greater degree of design freedom. In the latter, the added heft and metallic feel given to the plastics creates a perception of greater value [243-245].

Rigid parts made from polyamides can be given soft coatings by overmolding them with thermoplastic elastomers. Suitable elastomers include polypropylene-EPDM thermoplastic vulcanizates that contain small amounts of polyamide block copolymer to serve as a compatibilizer. Such “soft-touch” materials have a variety of applications in consumer goods [246], [247]. Alternatively, block copolymers of polyethers and polyamides have elastomeric properties and their hardness can be tuned over a wide range. These materials can be overmolded onto other materials for a variety of consumer, industrial, automotive, and adhesive applications [248] (⟶Thermoplastic Elastomers).

The amount of polyamide used and the nature of its dispersion in a blend with other polymers may be such that it is properly defined as the modifier rather than the base resin. One commercial example is a blend of nylon and poly (2, 6-dimethyl-p-phenylene oxide). Allusion has been made to a blend with poly(phenylene sulfide) as the major ingredient [194]. Another composition presumably involves mostly ABS and some nylon [195]. A developmental material is a glass-reinforced blend of polycarbonate and nylon [196]. The literature on blends is voluminous. Brief reviews of commercial blends including some commentary on their processing are given in [194-196]. A recent text focuses specifically on commercial blends and alloys [197]. Blends in which there is some level of bonding between matrix and modifier, whether due to likeness of solubility parameters, hydrogen bonding, or some other factor are sometimes termed alloys.

[Top of Page]

7. Processing

Nylons are processed into useful articles by a variety of techniques—injection molding, extrusion, blow molding, monomer casting, solution coating, fluidized-bed or electrostatic coating, or forming [19]. The major fraction of the annual polyamide volume is processed by injection molding, extrusion, and blow molding. Each of these melt processes requires different melt viscosities for optimum processing. Whereas injection molding requires the lowest melt viscosity, for extrusion and blow molding a high melt strength is needed, which by necessity requires a high melt viscosity. Molecular mass can be altered to meet the viscosity required for the desired process. In addition, modifying molecular architecture is sometimes used to alter the rheological properties (see Section Control of Rheological Properties.).

Equation (12) governs the dynamics of the amidation/hydrolysis equilibrium in polyamides. The polyamide molecular mass will either increase or decrease, and consequently the melt viscosity, as well, depending on whether the moisture content is higher or lower, respectively, than the equilibrium moisture content. Thus, moisture control in polyamide melt processing is extremely important and because of the ubiquitous nature of moisture it is often the most difficult to control. Most problems in nylon processing could be attributed to poor control of moisture. These problems can manifest themselves in mechanical properties, appearance, and processibility.

Good flow and fast set-up make nylons relatively easy to process. Standard three-stage screws with or without mixing heads, corotating twin screws, and barrier screws with or without venting are all used. Most reciprocating screws in injection molding are of standard type. Moisture content should be somewhat lower for extrusion than for injection molding because the absence of high die pressure facilitates volatilization. Control of moisture to avoid hydrolysis or polymerization is always a concern. Processors use dehumidifiers to control the moisture present in molding materials especially when using regrind or while operating under hot, humid conditions. Failure to compensate for changes in molecular mass during measurement of melt viscosity because of failure to control moisture content adequately, e.g., by over drying has resulted in questionable estimates of the activation energy of viscous flow. Thermal degradation can also be a problem if the melt is held too long at too high a temperature. Thus, one study preferred to use a capillary mounted on an injection molding machine rather than an Instron capillary rheometer in part because of shorter hold-up times [249]. Values of 42–63 kJ/mol for the activation energy of viscous flow appear reasonable rather than those of 71–92 kJ/mol as have sometimes been reported [19], [250]. Master curves that are based on only the T_g and melt flow index (MFI) per ASTM D 1238 have been proposed for a number of polymers including nylons [251]. These curves plot the product of viscosity and MFI against shear rate divided by MFI. The sample curves in Figure 15, mostly at 280 °C, were taken from [252-255]. As shown in the figure, the apparent melt viscosity of the linear, aliphatic nylons starts to decrease rapidly at a shear stress of ca. 50 MPa.

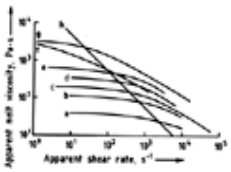

Figure 15. Dependence of apparent melt viscosity on apparent shear rate for various nylons at 280 °C (adapted from [19])

a) Nylon 610 [254]; b) Nylon copolymer [254]; c) Nylon 66, M_n ca. 18 000 [253], [254]; d) Nylon 6, M_n ca. 20 000 [253]; e) Nylon 6, M_n 19 100 at 203 °C [252]; f) TMDT at 250 °C [255]; g) Nylon 66, M_n ca. 34 000 [253], [254]; h) Line of constant shear stress of 50 MPa

Filled and reinforced nylons have higher melt viscosities than their neat analogues, but the differences become smaller as the level of shear increases [256], [257]. Processing variables affect the ultimate fiber size in the blend and can accordingly affect the properties of the molded product [258], [259]. Compounding studies indicated that the major fiber degradation occurs in the melting region [260]. A subsequent study came to the same conclusion and showed fiber fracture to be independent of fiber concentration and dispersion and to arise from fiber interaction with shear in the matrix [261].

Toughened nylons also tend to have higher melt viscosities than their neat counterparts. There is again diminution of the difference as shear increases. The lower modulus implies longer molding cycles, but in at least one case the toughener induced nucleation which made for fast molding [216].

[Top of Page]

8. Uses

Nylons are used in many and diverse ways. They are found in appliances, business equipment, consumer products, electrical/electronic devices, furniture, hardware, machinery, packaging, and transportation. This diversity makes classification and analysis difficult as shown in Table 8 which compares the pattern of consumption in the United States [200] with that in Western Europe [262].

Table 8. Pattern of nylon consumption in the North America and Western Europe in 2001 [262]

Class	Percentage use in	
	North America	Western Europe
Appliances and power tools	1.4	4.6
Consumer products	7.8	6.5
Electrical and electronic applications	7.6	19.3
Extrusion markets	13.1	24.1
Industrial equipment	9.0	6.5
Miscellaneous injection molding		3.4
Other	16.5	3.4
Transportation	44.5	32.2
Total percent	100	100
Total production, t/a	507 000	679 600

Transportation is the largest market for nylons. The good performance of nylon at high temperatures and pressures, as well as its chemical resistance and compatibility with many automotive fluids, along with the design flexibility available with plastics, paved the way for the adoption of polyamides in many under-the-hood applications. The introduction of high-performance mineral reinforced and glass reinforced versions of nylon greatly increased the ability to tailor properties for individual engine components. When government regulations required automobile manufactures to install pollution-control devices, designers turned to these proven materials for critical air and fuel system control components because of the polymer's performance and the relative ease of part manufacture. Utilizing nylon in the engine compartment reduces weight, which can improve fuel efficiency without sacrificing engine performance. Key developments were pioneered in Europe where smaller cars, higher gasoline costs, and emission control mandates drove innovation. Soon, commercial applications of nylons in air intake manifolds, brake fluid reservoirs, power steering reservoirs, and radiator end tanks were introduced. These developments gained footholds in Europe and spread to U.S. automobiles. Unreinforced resins are used in electrical connectors, wire jackets, windshield wiper and speedometer gears, and emission canisters. The softer nylons are used in fuel lines, air brake hoses, and spline shaft coatings. Glass reinforced nylons are used in engine fans, radiator headers and grilles, brake and power steering fluid reservoirs, valve covers, air intake manifolds, wheel caps, air brake contacts and head-rest shells. Applications for nylons combining tougheners and reinforcement include brackets, steering wheels, and accelerator and clutch pedals. Mineral-filled resins are used in wheel caps, radiator grilles, and mirror housings. Nylons containing both glass fibers and minerals are used in exterior parts such as fender extensions. Toughened nylons are found in stone shields and trim clips.

Electrical and electronic applications comprise a major market for nylons, albeit more so in Western Europe than in the United States. Flame-retardant materials are particularly important in this area. Uses include color-coded components, plugs, connectors, coil forms, wiring devices, terminal blocks, antenna mounting devices, and harness ties. A review of nylon in electrical power engineering and electronics has already been cited [234]. Wire and cable jacketing is used mostly over primary insulation because of the solvent, wear, and abrasion resistance of nylons. Relays, fittings, and contact makers constitute a partial list of applications in telecommunications.

Industrial applications are attracted to the excellent fatigue resistance and repeated impact strength of nylons. Examples are hammer handles and moving machine parts. The mechanical strength accounts for use in gears, bearings, antifriction parts, snap fits, and detents. Food- and textile-processing equipment, pumps, valves, agricultural and printing devices, and business and vending machines comprise a partial list of other industrial uses.

Consumer products exploit the toughness of nylons in ski boots, ice and roller skate supports, racquet equipment, and bicycle wheels. Kitchen utensils, toys, photographic equipment, brush bristles, fishing line, sewing thread, and lawn and garden equipment show that the breadth of utility of nylon also includes consumer products.

Appliances and power tools make use of the impact strength of nylons. Glass-reinforced resins, which combine stiffness at high temperatures with toughness and grease resistance, are used in handles, housings, and parts in contact with hot metal. Sewing machines, laundry equipment and dishwashers are examples of other appliances that have utilized nylons.

Film applications have grown in importance because of the use of nylons in coextruded, composite films for food

packaging. Polyamides are also used in cook-in bags and pouches. Another development involves dispersion of nylon platelets in a polyolefin matrix to combine both hydrocarbon and moisture barrier properties in blow molded containers [263], [264].

Heat Exchangers. A recent new application utilizing nylon resin in is heat exchangers for heating, ventilation, and air-conditioning (HVAC) applications (developed by Du Pont, trade name Caltrel®). Air-to-liquid exchangers provide heating, cooling and dehumidification for passenger cars and military vehicles. Air-to-air and liquid-to-liquid exchangers have been developed for industrial end uses such as in office buildings and restaurants.

Nylon provides excellent resistance to attack by chemicals, solvents, encrustation buildup, galvanic corrosion, and microbiological agents. These exchangers can endure in many environments, including areas near saltwater or exposed to aggressive vapors and fluids. Nylon has lower maintenance costs than metal because components have smooth internal surfaces that resist buildup of mineral deposits and debris, which often constrict fluid flow and decrease heat transfer performance of comparable metal heat exchangers. Any buildup is quickly and easily removed by flushing at moderate pressure.

Two significant advantages of nylon heat exchangers are their light weight and that they permit great design flexibility, allowing them to be adapted to many environments. Having lower weight is a continuing goal of the automotive industry, and the utilization of nylon makes the exchangers lighter and thus easier to install. The use of nylon also allows the units to be easily contoured to fit the available space, which is especially valuable in tight areas such as under the hood of an automobile [265].

Others Applications. The soft polymer-modified, anionically polymerized caprolactam has been used via RIM to make machine housings, grain buckets, and fuel tanks [266]. Blow molding compositions [267], PA 46 and MXD 6 continue to extend the utility in the market [268-270]. Products with improved processing characteristics, high- and low-temperature performance, and new property profiles obtained via the use of new modifications and combinations thereof are continually appearing [271], [272]. A PA 6 with reduced moisture sensitivity has been demonstrated [269], as has the use of block copolymers as thermoplastic elastomers [248].

The adaptability of nylon to be changed to even obtain diametrically opposite results is shown in Figure 16, which gives some examples of the applications involved. Nylons with qualified approval in the United States in 2002 for processing, handling, and/or packaging food are PA 66, 610, 66/610, 6/66, 11, 6, 6T/66, 612, 12, MXD6 and impact-modified MXD6, 12/AI (A = bis(4-amino-3-methylcyclohexyl)methane [6864-37-5]), 6I/6T, 6/69, 46, and TMDT [273].

Figure 16. Nylon readily tailored to market needs

[Top of Page]

9. Ecological Aspects and Toxicology

9.1. Recycling

Polyamides are quite stable, and are used primarily in durable applications that are not major contributors to the waste stream. Scrap and trim are generally reused during processing, further minimizing waste. Some consumer applications such as bristles, strapping, and composite film for packaging do involve routine disposal, but the total is well below 0.05 wt % of all collected refuse. As a result, commercial nylon recycling processes are not as well developed as they are for other thermoplastics. However, it is far more energy efficient to recycle nylon than to burn it [274], and European directives in 2001 that phase in a requirement that almost all plastics used in automobiles be recycled over the subsequent two decades will continue to spur development in the recycling of mixed thermoplastics. Several commercial initiatives to recycle readily-available PA 6 and PA 66 carpet fibers are also underway in the United States and Europe, but have met with mixed success at the outset due to low demand and the difficulties of efficiently and economically collecting raw materials [275-278].

Difficulties involved in recycling post-consumer polyamides include the fact that they are often used compounded with a wide variety of additives (see Section Modification by Additives), that need to be removed and the fact that the properties of polyamides generally suffer greatly upon multiple reheatings, as the polymers tend to degrade and lose molecular mass. Most of the current commercial recycling processes involve the depolymerization of the polymer chains and isolation of the monomers, which are then used as feedstocks to make near-virgin polyamides. PA 6 can be depolymerized to caprolactam by heating the polymer in the presence of superheated steam and an acid catalyst such as phosphoric acid and collecting the monomer by distillation [279], [280]. PA 66 or PA 6 can be depolymerized by ammonolysis, which involves heating the polymer with ammonia at elevated pressures in the presence of a Lewis acid catalyst to yield hexamethylenediamine (HMD) or caprolactam and a series of nitriles and amides, some of which can be hydrogenated to yield HMD or caprolactam precursors [281], [282]. Other methods exist that convert PA 66 or PA 6 to adipic acid by treating the polymers with an aliphatic acid at elevated temperatures and oxidizing the resulting alkylamides and bisalkyamides [283].

Depolymerization followed by repolymerization is an expensive process, however, and mechanical recycling of polyamides, which involves grinding or crushing the resin, followed by pelletizing, is also used, although the resulting resins are of lower quality and have limited uses. A solution-based polyamide recovery process has been developed in which the polymer is dissolved in an anhydrous polyol or aliphatic carboxylic acid at elevated temperatures, insoluble fillers are removed by filtration, and more solvent at a sufficiently low temperature to cause the polyamide to precipitate is added quickly [284]. As a result of the rapid quenching, this process leads to relatively little degradation of the polymer, which can be then be reused. A disadvantage of this process, however, is that the resulting solutions are quite viscous, which makes filtration through fine filters, and hence the removal of submicron particles, such as titanium dioxide (which is often used in polyamide fiber applications) difficult. To address this problem, a process was developed to partially depolymerize a polyamide dissolved in methanol or ethanol to achieve a viscosity suitable for filtration, separate the insoluble material, and then repolymerize the recovered low molecular polymer by solid-phase or melt polymerization [285].

9.2. Polyamide Monomers

Most polyamide monomers come from petroleum-based feedstocks, although two commercially important monomers, sebacic acid (decanedioic acid) and 11-aminoundecanoic acid, are derived from castor beans. The processes for preparing many common monomers often use ingredients that can generate byproducts and thus must be handled carefully. For example, adipic acid is generally produced by oxidizing a mixture of cyclohexanone and cyclohexanol with nitric acid, generating nitrous oxide, which is believed to be a greenhouse gas, as a byproduct. These byproducts are contained under rigorous manufacturing controls and not released into the environment. A variety of research projects have addressed these issues. A process has been developed that oxidizes benzene to phenol with nitrous oxide. The phenol is converted to cyclohexanone, which is oxidized to adipic acid with nitric oxide, and the resulting nitrous oxide byproduct is recycled back into the process [286], [287]. Recently a method has been reported to directly oxidize hexane to adipic acid in the presence of air using molecular sieves containing cobalt(III) ions [288]. A method has been developed to oxidize *p*-xylene to terephthalic acid using a heterogeneous catalyst in supercritical water [291].

9.3. Other Aspects

Polyamides degrade slowly in sunlight with the evolution of organic acids, aldehydes, and oxides of carbon. These are assimilated with other natural decomposition products. Resistance to attack by bacteria and fungi makes for very slow decomposition when buried in the soil. Nylons are normally inert, nonleaching, and well compacted in sanitary landfills so they pose no danger to subsurface water. Normal incineration in air at 800 °C yields the same gaseous products as do naturally occurring, proteinaceous materials such as wool, silk, hair, and leather, i.e., mainly carbon dioxide and carbon monoxide with small amounts of ammonia and hydrogen cyanide. Silk generates more of these products than does PA 66 under the indicated conditions of incineration.

Most polyamides are biologically inert and offer no toxicological problems except, possibly, in the instance of additives or some low molecular mass fragments. Caprolactam monomer has been reported to be a convulsant poison in sufficient doses although aminocaproic acid is not [290]. The same comment is applicable to the dermatological situation—no problem with the polymer but caution with respect to additives and caprolactam. Polyamides have been widely used as a suture material and many are approved for food uses. The use of wood fiber as a light-weight reinforcing agent in polyamides that comes from a renewable resource and leaves no ash upon incineration has also been investigated [292]. Polyamides can contribute to decreased emissions and fuel consumption when they replace heavier metals in automotive applications and can replace toxic metals in certain applications (see Section Other Formulations).

[Top of Page]

10. Economic Aspects

Modern Plastics publishes an annual review of new developments and market trends and usage data for most major thermoplastics in its February issue (the January issue prior to 2000) and many October issues of *Kunststoffe* publish reviews of new developments and market changes. *Plastics Technology* lists prices monthly. These have all contributed to the information provided in this article.

Consumption of polyamides in Western Europe has grown from 200 000 t/a in 1980 to 679 000 t/a in 2001, which corresponds to an average annual compounded growth rate of about 6 %. In the United States, consumption of polyamides has grown from 100 000 t/a in 1976 to 507 000 t/a in 2001, corresponding to an average annual compounded growth rate of 6–7 %. In Japan, consumption has grown from 186 t/a in 1994 to 267 t/a in 2000, before dropping back to 246 t in 2001, corresponding to an average annual compounded growth rate of 4–5 % between 1994 and 2001. Year to year changes are not uniform, and significant annual dips have occurred as a result of economic forces.

Major market segments worldwide for injection-molded polyamides are appliances/power tools, consumer products, electrical and electronics, and transportation. In the United States the average annual compounded growth for the first four was 4–5 % between 1988 and 2001, although the appliance segment experienced a sharp drop to 1988 levels between 1999 and 2001. The transportation/automotive segment is easily the largest and fastest growing, with an annual rate of 9–10 % between 1988 and 2001 and comprised about 44 % of the polyamide plastics market in 2001. A high of 254 t was consumed in 2000, but consumption dipped to 226 t in 2001. Between 1988 and 2001 the ratio of consumption for molding applications to extrusion applications ranged from about between two and three to one. Extrusion applications dipped sharply in 2000 and 2001 back to 1988 levels, however, and between five and six times more polyamide was consumed for molding applications than extrusion applications in those years.

In Europe, the largest and fastest growing area is also the transportation/automotive segment, which experienced an average annual compounded growth rate of 9–10 % between 1994 and 2001 and comprised about 32 % of the polyamide market in 2001. The other molding application segments grew at about 4–7 % during the same time period. Extrusion applications continued to grow at a rate of about 7–8 % during this time period. The ratio of consumption for molding applications to extrusion applications was fairly steady at about three to one during this time period.

Producers of polyamides are listed with their products and trade names in Table 9.

Table 9. Major producers of polyamides

Company	PA 6	PA 66	Other	Trade names
Asahi		x		Leona
Atofina	x	x	11, 12	Rilsan, Orgamid, Cristamid, Platamid
BASF	x	x	610	Ultramid
Bayer	x	x		Durethan
Degussa			12, amorphous	Vestamid, Trogamid
DSM	x	x	46, 610	Akulon, Stanyl
DuPont	x	x	612, phthalamides, amorphous	Zytel, Minlon
EMS-Grivory	x		12, amorphous	Grilon, Grilamid, Grivory
Enichem (Polimeri Europa)	x			Nivionplast
Honeywell	x	x		Capron
Mitsubishi Engineering-Plastics	x	x		Novamid
Mitsubishi Gas			MXD6	Reny
Mitsubishi Petrochemicals			phthalamides	Arlen
Rhône-Poulenc	x	x	610	Technyl
Snia Tecnopolimeri	x	x		Sniamid, Nailonplast
Solutia		x		Vydyne
Solvay			MXD6, phthalamides	Ixef, Amodel
Ticona	x	x		Celanese
Toray	x	x	610, 12	Amilan
Ube	x	x	12	Ube
Unitika	x			Unitika

[Top of Page]

References

1. W. H. Carothers, *J. Am. Chem. Soc.* **51** (1929) 2548–2559. Links
2. W. H. Carothers, J. A. Arvin, [1], 2560 – 2570.
3. DuPont, US 2 071 250, 1937 (W. H. Carothers).
4. IG Farbenindustrie, US 2 241 321, 1941 (P. Schlack).
5. M. I. Kohan: “The History and Development of Nylon-66” in R. B. Seymour, G. S. Kirshenbaum (eds.): *High Performance Polymers: Their Origin and Development*, Elsevier, New York 1986, pp. 19 – 37.
6. P. Matthies, W. F. Seydl: “History and Development of Nylon 6”, in [5] pp. 39 – 53.
7. G. B. Apgar, M. J. Koskoski: “The History of Development of Nylons 11 and 12,” in [5] pp. 55 – 65.
8. H. Staudinger, *Ber. Dtsch. Chem. Ges.* **53** (1920) 1073–1092.
9. W. H. Carothers, *Chem. Rev.* **8** (1931) 353–426. Links
10. R. W. Lenz: *Organic Chemistry of Synthetic High Polymers*, Interscience Publishers, New York 1967, pp. 1–9.
11. M. I. Kohan (ed.): *Nylon Plastics Handbook*, Hanser Verlag, Munich 1995, p. 3.
12. G. B. Kauffman, *Chem. Tech.* **18** (1988) 725–731.
13. Source-Based Nomenclature for Copolymers, IUPAC Macromolecular Division, Provisional Paper (1985) Rule 7.
14. Nomenclature of Regular Single-Strand Organic Polymers, IUPAC Macromolecular Division, Tentative, *Pol. Lett. Div.* **11** (1973) 389 – 414.
15. J. Brandrup: Nomenclature Rules in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook*, 3rd ed. Wiley-Interscience, New York 1989, pp. I/1-I/61.
16. R. Pfluger: “Physical Constants of Various Polyamides”, in [15] p. V/109.
17. G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (eds.): Comprehensive Polymer Science, vol. **5**, *Step Polymerization*, Pergamon Press, Oxford 1989.
18. P. J. Flory: *Principles of Polymer Chemistry*, Cornell Univ. Press, New York 1953.
19. M. I. Kohan (ed.): *Nylon Plastics*, Wiley-Interscience, New York 1973.
20. J. Zimmerman: *Polyamides in Encyclopedia of Polymer Science and Engineering*, 2nd ed., **vol. 11**, Wiley-Interscience, New York 1988, pp. 315 – 381.

21. F. Bueche, *J. Polym. Sci.* **43** (1960) 527 – 530.
22. T. G. Fox, S. Loshaek, *J. Appl. Phys.* **26** (1955) 1080–1082. Links
23. T. G. Fox, S. Gratch, S. Loshaek in F. R. Eirich (ed.): *Viscosity Relationships for Polymers in Bulk and in Concentrated Solution*, **vol. 1**, Rheology-Academic Press, New York 1956, pp. 431 – 493.
24. M. A. Dudley, *J. Appl. Polym. Sci.* **16** (1972) 493–504. Links
25. E. Jacobi, H. Schuttenberg, R. C. Schultz, *Makromol. Chem. Rapid Commun.* **1** (1980) 397 – 402.
26. G. Pastuska, U. Just, H. August, *Angew. Makromol. Chem.* **107** (1982) 173 – 184. Links
27. E. Biagini, E. Gattiglia, E. Pedemonte, S. Russo, *Makromol. Chem.* **184** (1983) 1213 – 1222.
28. P. J. Flory, *J. Am. Chem. Soc.* **58** (1939) 1877 – 1885.
29. W. H. Stockmayer, *J. Chem. Phys.* **11** (1943) 45 – 55. Links
30. J. R. Schaefgen, P. J. Flory, *J. Am. Chem. Soc.* **70** (1948) 2709 – 2718. Links
31. DSM, N. V., WO 96/35739, 1996 (A. Nijenhuis, M. Serne, E. van den Berg, R. Aberson).
32. Nyltech, WO 97/24388, 1997 (A. Cucinella, G. Di Silvestro, C. Guaita, F. Speroni, H. Zhang).
33. Nyltech, WO 99/64496, 1999 (G. Di Silvestro, F. Speroni, C. Yuan, H. Zhang).
34. R. Puffr, Z. Tuzar, K. Mrkvickova, J. Sebenda, *Makromol. Chem.* **184** (1983) 1957.
35. R. Puffr, V. Kubanek: *Lactam-Based Polyamides*, **Vols. I & II**, CRC Press, Boca Raton 1991.
36. H. Zahn, G. B. Gleitsman, *Angew. Chem. Int. Ed. Engl.* **2** (1963) 410 – 420.
37. P. Kusch, H. Zahn, *Angew. Chem. Int. Ed. Engl.* **4** (1965) 696.
38. T. M. Cawthon, E. C. Smith, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **1** (1960) Sep., 98.
39. H. K. Reimschuessel, *J. Polym. Sci.* **41** (1959) 457 – 466.
40. H. K. Reimschuessel, *Macromol. Rev.* **12** (1977) 65 – 139.
41. P. Feldmann, R. Feinauer, *Angew. Makromol. Chem.* **34** (1973) 9 – 18. Links
42. C. J. Brown, H. Hill, P. V. Youle, *Nature (London)* **177** (1956) 128.
43. H. Zahn, P. Miro, F. Schmidt, *Chem. Ber.* **90** (1957) 411 – 418.
44. S. Schaaf, *Faserforsch. Textiltech.* **10** (1959) 328–330.
45. R. Aelion, *Ind. Eng. Chem.* **53** (1961) 826 – 828.
46. K. Dachs, E. Schwartz, *Angew. Chem. Int. Ed. Engl.* **1** (1962) 430 – 435.
47. M. Genas, *Angew. Chem.* **74** (1962) 535 – 540.
48. C. F. Horn, B. T. Freure, H. Vineyard, H. J. Decker, *J. Appl. Polym. Sci.* **7** (1963) 887 – 896. Links
49. H. K. Reimschuessel: “Lactams,” in K. C. Frisch, S. L. Reegen (eds.): *Ring-Opening Polymerizations*, Dekker, New York 1969, pp. 303 – 326.
50. D. C. Jones, T. R. White: “Polyamides” in D. H. Solomon (ed.): *Step-Growth Polymerization*, Dekker, New York 1972, pp. 41 – 94.
51. H. K. Reimschuessel, *Macromol. Rev.* **12** (1977) 65–139.
52. H. Sekiguchi: “Lactams and Cyclic Imides” in K. J. Ivin, T. Saegusa (eds.): *Ring-opening Polymerization*, vol. **2**, Elsevier, London 1984, pp. 809 – 918.
53. R. J. Gaymans, D. J. Sikkema: “Aliphatic Polyamides” in G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (eds.): *Comprehensive Polymer Science*, vol. **5**, Step Polymerization, Pergamon Press, Oxford 1989, pp. 357 – 373.
54. C. Giori, B. T. Hayes, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 335 – 349.
55. C. Giori, B. T. Hayes, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 351 – 358.
56. K. Tai, H. Teranishi, Y. Arai, T. Tagawa, *J. Appl. Poly. Sci.* **25** (1980) 77 – 87.
57. J. Zimmerman: *Polyamides in Encyclopedia of Polymer Science and Engineering*, 2nd ed., **vol. 11**, Wiley-Interscience, New York 1988, pp. 315 – 381.
58. K. Tai, T. Tagawa, *Ind. Eng. Chem. Prod. Res. Dev.* **22** (1983) no. 2, 192 – 206.
59. Du Pont, US 2 130 523, 1935 (W. H. Carothers).
60. Du Pont, US 2 244 192, 1939 (P. J. Flory).
61. Du Pont, US 2 669 556, 1953 (C. A. Sperati).
62. SU 775 106, 1978 (B. S. Khaitin *et al.*); *Chem. Abstr.* **94** (1981) 48 033 f.
63. Firestone Tire and Rubber, US 3 705 881, 1969 (S. E. Schonfeld).
64. ICI, US 3 705 134, 1970 (D. James).
65. DuPont, WO 92/10526, 1992 (J. F. Buzinkai, M. R. DeWitt, R. C. Wheland).
66. DuPont, US 3 173 898, 1965 (W. M. Sum).
67. Allied Corp., US 4 471 081, 1984 (J. A. Bander, T. S. Brown, L. C. Lin, D. M. Vultaggio).
68. V. V. Korshak, T. M. Frunze: *Synthetic Hetero-chain Polyamides*, Jerusalem Program for Scientific Translations, Jerusalem 1964, pp. 17–21.
69. E. Roerdink, P. J. DeJong, J. Warnier, *Polym. Commun.* **25** (1984) 194–195.
70. E. Roerdink, J. M. M. Warnier, *Polymer* **26** (1985) 1582. Links
71. A. M. Kotliar, *J. Polym. Sci.: Polym. Chem. Ed.* **11** (1973) 1157 – 1165.
72. L. F. Beste, R. C. Houtz, *J. Polym. Sci.* **8** (1952) 395 – 407.
73. I. K. Miller, *J. Polym. Sci. Polym. Chem. Ed.* **14** (1976) 1403 – 1417.
74. F. W. Billmeyer, Jr.: *Textbook of Polymer Science*, 3rd ed., Wiley-Interscience, New York 1984, pp. 332–333.

75. J. Zimmerman: *Polyamides in Encyclopedia of Polymer Science and Engineering*, 2 ed., vol. **11**, Wiley-Interscience 1988, pp 340–353.
76. Du Pont, US 2 130 523, 1935 (W. H. Carothers).
77. Chemische Werke Witten, US 3 454 536, 1964 (G. Schade, F. Blaschke).
78. General Electric, US 4 567 249, 1983 (D. W. Fox, S. J. Shafer).
79. Du Pont, US 4 009 153, 1975 (H. Shin).
80. Standard Oil, US 4 719 285, 1986 (B. S. Curatolo, R. C. Sentman, G. P. Coffey).
81. Halcon International, US 3 847 876, 1972 (O. T. Onsager).
82. Standard Oil, US 4 520 190, 1983 (G. P. Coffey, B. S. Curatolo, R. L. Cepulla).
83. Standard Oil, US 4 542 205, 1984 (B. S. Curatolo, R. C. Sentman, G. P. Coffey).
84. Standard Oil, US 4 603 192, 1985 (G. P. Coffey, R. S. Sherrard, Jr., B. S. Curatolo, R. C. Sentman).
85. Standard Oil, US 4 725 666, 1986 (B. S. Curatolo, R. C. Seatman, G. P. Coffey).
86. Du Pont, US 2 190 770, 1936 (W. H. Carothers).
87. RED-Gesellschaft m.b.H., US 2 918 455, 1955 (E. Elod, R.-E. Dorr, H. Braun, J. Elian, M. Lepingle).
88. Sumitomo Chemical, US 4 721 772, 1982 (K. Ueno, K. Nagaoka, A. Miyashita).
89. W. J. Farrissey, B. K. Onder, P. S. Andrews, *Mod. Plast.* **58** (1981) March, 82, 84.
90. P. W. Morgan: *Condensation Polymers: By Interfacial and Solution Methods*, Interscience, New York 1965.
91. L. Vollbracht: “Aromatic Polyamides” in G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (eds.): *Comprehensive Polymer Science*, vol. 5, Step Polymerization-Pergamon, Oxford 1989, pp. 375 –386.
92. G. M. Carlyon: “Monomer Casting” in M. I. Kohan (ed.): *Nylon Plastics*, Wiley-Interscience, New York 1973, pp. 457 – 472.
93. Mitsubishi Monsanto, JP-Kokai 6 303 909, 1986 (T. Kato, H. Isekawa); *Chem. Abstr.* **108** (1988) 151 647 r.
94. Monsanto, US 4 595 746, 1984 (J. D. Gabbert, R. M. Hedrick).
95. BASF, DE-OS 3 439 461, 1984 (P. Horn, E. Baumann, M. Marx, P. Gerarts); *Chem. Abstr.* **105** (1988) 44 417 g.
96. Ube, JP-Kokai 60 223 822, 1984 (M. Ogasa, M. Matsumoto, A. Kawabata); *Chem. Abstr.* **104** (1986) 187 511 y.
97. PPG, US 4 501 821, 1983 (R. B. Hodeck, J. A. Seiner).
98. Istituto Chim, EP 130 360, 1983 (G. C. Alfonso, S. Russo, E. Pedemonte, A. Turturro, C. Puglisi); *Chem. Abstr.* **102** (1985) 132 965 c.
99. Ube, JP-Kokai 5 958 030, 1982; *Chem. Abstr.* **101** (1984) 111 589 x.
100. Aisin Seiki, JP-Kokai 58 210 924, 1982; *Chem. Abstr.* **100** (1984) 140 442 d.
101. VEB Leuna-Werke “Walter Ulbricht”, DD 202 885, 1983 (H. Schade *et al.*); *Chem. Abstr.* **100** (1984) 140 337 y.
102. Ural Inst. Chem., SU 937 475, 1980 (V. P. Begishev *et al.*); *Chem. Abstr.* **97** (1982) 199 022 q.
103. R. M. Hedrick, J. D. Gabbert, AJChE: A New RIM System from Nylon 6 Block Copolymers: Chemistry and Structure, National Mtg., Detroit, Summer, 1981.
104. J. D. Gabbert, R. M. Hedrick, *Polym. Process Eng.* **4** (1986) 359 – 373.
105. J. A. Lefelar, K. Udipi, *Polym. Commun.* **30** (1989) Feb., 38 – 40.
106. Harwe, US 4 223 433, 1980 (W. Hartmann).
107. Du Pont, US 2 172 374, 1937 (P. J. Flory).
108. BASF, GB 801 733, 1955.
109. California Research, US 2 987 507, 1957 (I. E. Levine).
110. E. M. Kampouris, *Polymer* **17** (1976) 409 – 412. Links
111. E. M. Macchi, A. A. Giorgi, *Makromol. Chem.* **180** (1979) 1603 – 1605.
112. Du Pont, US 3 031 433, 1958 (G. C. Monroe, Jr.).
113. Du Pont, US 3 821 171, 1972 (D. H. Beaton).
114. F. Pilati: “Solid-state Polymerization” in G. C. Eastmond, A. Ledwith, S. Russo, P. Sigwalt (eds.): *Comprehensive Polymer Science*, vol. 5, Step Polymerization, Pergamon Press, Oxford 1989, pp. 201 – 216.
115. J. Zimmerman, *J. Polym. Sci. Part B* **2** (1964) 955 – 958.
116. R. Feldmann, R. Feinauer, *Angew. Makromol. Chem.* **34** (1973) 1 – 7. Links
117. R. J. Gaymans, J. Amirtharaj, H. Kamp, *J. Appl. Poly. Sci.* **27** (1982) 2513 – 2526.
118. Du Pont, US 3 015 651, 1956 (E. W. Kjellmark, Jr.).
119. Unitika, JP 7 251 120, 1968 (Y. Nishida, K. Hirose); *Chem. Abstr.* **80** (1974) 15 507 w.
120. R. J. Gaymans, T. E. C. Van Utteren, J. W. A. Van Den Berg, J. Schuyer, *J. Polym. Sci. Polym. Chem. Ed.* **15** (1977) 537 – 545.
121. E. Roerding, J. M. M. Warnier, *Polymer* **26** (1985) 1582 – 1588.
122. Stamicarbon, NL-A 8 001 764, 1980 (R. J. Gaymans, E. H. J. P. Bour); *Chem. Abstr.* **96** (1982) 69 637 n.
123. J. M. Schultz, S. Fakirov: *Solid State Behavior of Linear Polyesters and Polyamides*, Prentice Hall, London 1990.
124. D. O'Sullivan, *Chem. Eng. News* (1984) May, 33–34.
125. Stamicarbon, EP-A 160 337, 1984 (R. J. Gaymans, A. J. P. Bongers); *Chem. Abstr.* **104** (1986) 89 240 q.
126. Asahi, JP-Kokai 61 188 421, 1985 (K. Higami); *Chem. Abstr.* **106** (1987) 67 885 t.
127. Bayer, DE-OS 3 526 931, 1985 (W. Nielinger *et al.*); *Chem. Abstr.* **106** (1987) 120 440 x.
128. Stamicarbon, US 4 722 997, 1985 (E. Roerdink, J. M. M. Warnier).

129. Stamicarbon, JP-Kokai 62 158 722, 1986; *Chem. Abstr.* **107** (1987) 237 562 a.
130. Du Pont, US 2 130 947, 1935 (W. H. Carothers).
131. Du Pont, US 2 163 584, 1936 (W. H. Carothers, G. D. Graves).
132. V. D. Luedeke: “Hexamethylenediamine”, in J. J. Mcketta (ed.): *Encyclopedia of Chemical Processing and Design*, Dekker, New York 1987, pp. 222 – 237.
133. Du Pont, US 3 337 612, 1964 (D. L. Sharps).
134. Du Pont, US 3 439 025, 1965 (R. Gallay, S. P. LaPenta).
135. BASF, US 4 233 234, 1978 (R-H. Rotzoll *et al.*).
136. Du Pont, US 4 442 260, 1983 (H. A. Larsen).
137. Du Pont, US 2 840 547, 1955 (W. L. Stump).
138. Monsanto, CA 715 476, 1962 (E. P. Brignac).
139. Monsanto, US 3 300 449, 1964 (E. P. Brignac).
140. D. B. Jacobs, J. Zimmerman: “Preparation of 6,6-Nylon and Related Polyamides” in C. E. Schildknecht, I. Skeist (eds.): *Polymerization Processes, High Polymers*, vol. 29, Wiley-Interscience, New York 1977, pp. 424 – 467.
141. Du Pont, US 2 289 774, 1939 (G. DeW. Graves).
142. Du Pont, US 3 491 177, 1967 (W. G. Johnson).
143. Du Pont, US 2 689 839, 1951 (W. W. Heckert).
144. Du Pont, US 3 113 843, 1959 (W. H. Li).
145. Du Pont, US 3 361 537, 1965 (N. A. Ferrante).
146. Du Pont, US 3 948 862, 1974 (J. M. Iwasyk).
147. Du Pont, US 4 134 736, 1978 (L. F. Hammond, Jr.).
148. Du Pont, US 4 299 498, 1979 (R. D. Sauerbrunn).
149. Du Pont, US 3 789 584, 1972 (J. M. Iwasyk, D. W. Rodeffer).
150. Vereinigte Glanzstoff-Fabriken, US 3 027 355, 1957 (H. Taul, F. Wiloth).
151. British Nylon Spinners, US 3 185 672, 1959 (P. C. Clemo, J. A. Briggs, W. Wilson, J. A. Carter).
152. BASF, US 4 060 517, 1974 (F. Mertes, H. Doerfel, E. Heil, C. Cordes).
153. Monsanto, US 3 296 217, 1964 (J. E. Tate).
154. Du Pont of Canada, US 3 900 450, 1973 (I. Jaswal, K. Pugi).
155. British Nylon Spinners, US 3 258 313, 1962 (G. D. Griffiths).
156. Du Pont, US 3 118 739, 1959 (W. G. Atkinson, J. L. Thomas, Jr.).
157. Du Pont, US 3 257 173, 1960 (J. E. Parnell).
158. Du Pont of Canada, US 3 960 820, 1974 (B. McD. Pinney).
159. American Enka Corp., US 3 044 993, 1957 (H. G. N. Tiemersma).
160. Monsanto, US 3 218 297, 1960 (G. W. Sovereign).
161. Monsanto, US 3 260 703, 1964 (A. C. Coggeshall).
162. Vereinigte Glanzstoff-Fabriken, US 3 130 180, 1958 (F. Wiloth).
163. Amoco, US 4 831 108, 1987 (J. A. Richardson, W. Poppe, B. A. Bolton, E. E. Paschke).
164. California Res. Corp., US 2 997 463, 1957 (F. G. Lum).
165. G. Bier: “New Polycondensation Polymers,” *Adv. Chem. Ser.* **91** (1969) 612 – 627.
166. BP Chemicals, US 5 079 307, 1992 (S. A. Taylor, M. B. Studholme, M. R. Orpin).
167. Mitsui Petrochemicals, US 6 130 312, 2000 (H. Murakami, S. Omari, K. Wakatsuru).
168. H. Jacobs, C. Schweigman: “Mathematical Model for the Polymerization of Caprolactam to Nylon-6 in Chemical Reaction Engineering,” *Proceedings of the Fifth European/Second International Symposium on Chemical Reaction Engineering*, Elsevier, Amsterdam 1972, pp. B 7–1 to-26.
169. Allied Chemical, US 3 578 640, 1968 (I. C. Twilley, G. J. Coll, Jr., D. W. H. Roth, Jr.).
170. Allied Chemical, CA 823 290, 1967 (I. C. Twilley, D. W. H. Roth, Jr., R. A. Lofquist).
171. Allied Chemical, US 3 813 366, 1973 (W. H. Wright, A. J. Bingham, W. A. Fox).
172. Allied Chemical, EP-A 0 038 473, 1980 (S. L. Yates, C. J. Cole, A. H. Wiesner, J. W. Wagner).
173. A. K. Ray, S. K. Gupta, *Polym. Eng. Sci.* **26** (1986) no. 15, 1033 – 1044. Links
174. A. Kumar, S. K. Gupta, *J. Macromol. Sci. JMS-Rev. Macromol. Chem. Phys.* **C 26** (1986) no. 2, 183–247.
175. G. Menges, T. Bartilla, *Polym. Eng. Sci.* **27** (1987) 1216 – 1220. Links
176. P. Wagner, *Kunststoffe* **73** (1983) 588 – 590.
177. C. R. Dupre, J. D. Gabbert, R. M. Hedrick, *Am. Chem. Soc. Polym. Chem. Div. Preprints* **25** (1984) no. 2, 296 – 297.
178. L. Notarbartolo, *Ind. Plast. Mod.* **10** (1958) 44 – 52.
179. Societe-Organico, FR 951 924, 1947 (M. Genas); *Chem. Abstr.* **46** (1952) 132 g.
180. N. V. Onderzoekingsinst. Res., BE 642 232, 1963.
181. ICI, GB 680 212, 1968 (H. McGrath).
182. BASF, DE-AS 1 495 149, 1963 (J. H. Kunde, J. H. Wilhelm, H. Metzger, H. Dorfel).
183. W. Griehl, D. Ruestem, *Ind. Eng. Chem.* **62** (1970) no. 3, 17 – 22.
184. For details go to www.campusplastics.com.

185. R. G. Beaman, *J. Polym. Sci.* **9** (1952) 470 – 472.
186. H. G. Weyland, P. J. Hoftyzer, D. W. Van Krevelen, *Polymer* **11** (1970) 79 – 87. Links
187. K-H. Illers, *Makromol. Chem.* **38** (1960) 168 – 188.
188. A. M. Thomas, *Nature (London)* **179** (1957) 862.
189. U. Rhode-Liebenau, *Kunststoffe* **55** (1965) 302 – 305.
190. G. A. Gordon, *J. Polym. Sci. Polym. Phys. Ed.* **9** (1971) 1693 – 1702.
191. R. Leaversuch, L. M. Sherman, M. H. Naitove, *Plast. Technol.* **47** (2001) Sept., 65–67.
192. For details go to www.rhodia-ep.com/index.htm.
193. F. Stastny, *Kunststoffe* **40** (1950) 273 – 304.
194. E. D. Harvey, F. J. Hybart, *J. Appl. Polym. Sci.* **14** (1970) 2133 – 2143. Links
195. W. E. Catlin, E. P. Czerwin, R. H. Wiley, *J. Polym. Sci.* **2** (1947) 412 – 419.
196. H-G. Elias, *Polymer News* **11** (1985) 40 – 43.
197. Imperial Chemical Industries, GB Pat App. 604/9, 1949 (H. Plimmer).
198. Inventa AG, CH 280 367, 1949 (R. Gabler).
199. O. B. Edgar, R. Hill, *J. Pol. Sci.* **8** (1952), pp. 1–22.
200. R. Gächter, H. Müller: *Plastics Additive Handbook*, 4th ed., Hanser, Munich 1993.
201. J. M. Crosby, T. R. Drye, *Mod. Plast.* **63** (1986) Nov., 74 – 84. Links
202. D. A. Cianelli, J. E. Travis, R. S. Bailey, *Plast. Technol.* **34** (1988) Apr., 83 – 89.
203. Libbey-Owens-Ford Glass, US 2 571 717, 1951 (A. M. Howard, L. S. Meyer).
204. Glastrustions, US 2 871 911, 1959 (W. B. Goldsworthy, F. Landgraf).
205. A. H. Tullo, *Chem. Eng. News* **80**(4) (2002) Jan. 28, 21.
206. M. Holmes, *Plastics Additives & Compounding*, January 2002, pp. 30 – 33.
207. R. Leaversuch, *Plast. Technol.* **47** (2001) October, 64–69.
208. P. S. Wilson: “Nanocomposite Market Opportunities” in *Proceedings of Nanocomposites 2001*, Executive Conference Management, Chicago 2001.
209. Proceedings of Nanocomposites 2001, Executive Conference Management, Chicago 2001.
210. D. R. Paul, S. Newman: *Polymer Blends*, Academic Press, New York 1978.
211. M. J. Folkes: *Polymer Blends and Alloys*, Blackie, London 1993.
212. L. A. Utracki: *Polymer Alloys and Blends*, Hanser Verlag, Munich 1989.
213. L. A. Utracki: *Polymer Alloys and Blends*, Hanser Verlag, Munich 1989, pp. 248–270.
214. W. E. Baker, C. Scott, G.-H. Hu: *Reactive Polymer Blending*, Hanser Verlag, Munich 2001.
215. Du Pont, US 4 174 358, 1977 (B. N. Epstein).
216. E. A. Flexman, Jr., *Polym. Eng. Sci.* **19** (1979) no. 8, 564 – 571. Links
217. R. E. Lavengood, F. M. Silver, *Soc. Plast. Eng. (Tech. Pap.)* **33** (1987) 1369 – 1374.
218. Rohm and Haas, US 3 796 771, 1972 (F. H. Owens, J. S. Clovis).
219. G. C. Adams: “Impact Fatigue of Polymers Using an Instrumented Drop Tower Device” in S. L. Kessler, G. C. Adams, S. B. Driscoll, D. R. Ireland (eds.): *Instrumented Impact Testing of Plastics and Composite Materials*, ASTM STP 936-Amer. Soc. Test. Matls., Philadelphia 1987, pp. 281 –301.
220. J. R. Heater, E. M. Lacey, *Mod. Plast.* **41** (1964) May, 123 – 126, 180 – 182. Links
221. B. Epstein: Soc. Plast. Industry, Reinf. Plast./Composites Institute, *34th Ann. Tech. Conf.* (1979) 10-A, pp. 1 – 6.
222. R. C. Nametz, *Plast. Comp.* **2** (1979) Jan./Feb., 31 – 40; **7** (1984) July/Aug., 26 – 39.
223. N. A. Burditt, *Plast. Comp.* **11** (1988) Sep./Oct., 51 – 56.
224. M. C. McMurrer, *Plast. Comp.* **11** (1988) July/Aug., 60 – 61.
225. V. Wigotsky, *Plast. Eng.* **43** (1987) Oct., 25–30. Links
226. Ems-Inventa, US 4 511 684, 1982 (E. Schmid, M. Hoppe).
227. Anon., *Plast. Eng.* **44** (1988) Oct., 53. Links
228. Clariant, US 6 255 371, 2001 (E. Schlosser, B. Nass, W. Wanzke).
229. C. S. Hardo, J. J. Duffy, *Plast. Eng.* **41** (1985) Sep., 51 – 54.
230. J. E. Theberge, J. M. Crosby, K. L. Talley, *Plast. Eng.* **44** (1988) Aug., 47 – 52. Links
231. K. J. Nangrani, R. Wenger, P. G. Daugherty, *Plast. Comp.* **11** (1988) Mar./Apr., 28 – 31.
232. Asahi, US 6 051 307, 2000 (T. Kido, S. Hasegawa).
233. Hyperion, US 5 651 922, 1997 (P. R. Nahass, S. O. Friend, R. W. Hausslein).
234. E. M. Doll, M. Koetting, *Soc. Plast. Eng. (Tech. Pap.)* **33** (1987) 406 – 409.
235. D. M. Bigg, *Polym. Compos.* **8** (1987) 1 – 7.
236. L. M. Sherman, *Plast. Technol.* **47** (2001) June, 52–53, 55–57.
237. M. T. Shaw, *Polym. Eng. Sci.* **22** (1982) no. 2, 115 – 121.
238. V. Wigotsky, *Plast. Eng.* **42** (1986) July, 19 – 29.
239. P. A. Toensmeier, *Mod. Plast.* **65** (1988) Oct., 62 – 67.
240. Polymer Blends and Alloys-Technomic, 1988 Lancaster.
241. B. Arkles, C. Carreno, *Soc. Plast. Eng. (Tech. Pap.)* **30** (1984) 486 – 487.

242. R. Leaversuch, *Plast. Technol.* **47** (2001) July, 43–45, 47.
243. R. D. Leaversuch, *Plast. Technol.* **47** (2001) June, 58–61, 63.
244. EP-A 0 641 836, 1995 (J. C. Gardner, P. J. Gardner, I. P. Oliver, T. Peake).
245. “Ideas to Market”, US 6 048 379, 2000 (A. V. Bray, B. A. Muskopf, M. L. Dingus).
246. J. De Gaspari, *Plast. Technol.* **44** (1998) June, 44–48.
247. J. A. Grande, *Mod. Plast.* **75** (1998) May, 74–79.
248. J. R. Flesher, Jr., *Mod. Plast.* **64** (1987) Sep., 100, 105 – 106, 110. Links
249. H. W. Cox, C. C. Mentzer, R. C. Custer, *Polym. Eng. Sci.* **24** (1984) 501 – 510. Links
250. P. Perrini, D. Romanini, G. P. Righi, *Polymer* **17** (1976) 377 – 381.
251. A. V. Shenoy, D. R. Saini, V. M. Nadkarni, *Rheol. Acta* **22** (1983) 209 – 222. Links
252. G. Pezzin, G. B. Gechele, *J. Appl. Polym. Sci.* **8** (1964) 2195 – 2212. Links
253. R. F. Westover: “Processing Properties” in E. C. Bernhardt (ed.): *Processing of Thermoplastic Materials*, Reinhold Publ. Co., New York 1959, pp. 547 – 679.
254. Plastics Department, Du Pont, *Tech. Bulletin: Molding Du Pont Zytel*^{R} *Nylon Resins . . . a Handbook for the 70's*, 1970.
255. R. Gabler *et al.*, *Chimia* **21** (1967) no. 2, 65 – 81. Links
256. J. C. L. Williams, D. W. Wood, I. F. Bodycot, B. N. Epstein, *Soc. Plast. Ind. Reinf. Plast. Div. Conf. Proceedings* **23** (1968) 2 C 1 – 2 C 8.
257. R. J. Crowson, M. J. Folkes, *Polym. Eng. Sci.* **20** (1980) 934 – 940. Links
258. W. C. Filbert, Jr., *SPE J.* **25** (1969) Jan., 65 – 69.
259. H. W. H. Yang, R. Farris, J. C. W. Chien, *J. Appl. Polym. Sci.* **23** (1979) 3375 – 3382. Links
260. J. M. Lunt, J. B. Shortall, *Plast. Rubber Process.* **5** (1980) June, 37 – 44.
261. R. v. Turkovich, L. Erwin, *Polym. Eng. Sci.* **23** (1983) 743 – 749.
262. *Mod. Plast.* **79** (2002) Feb., 27.
263. P. M. Subramanian, *Polym. Eng. Sci.* **25** (1985) June, 483 – 487. Links
264. P. M. Subramanian, V. Mehra, *Polym. Eng. Sci.* **27** (1987) May, 663 – 668. Links
265. R. Leaversuch, L. M. Sherman, M. H. Naitove, *Plast. Technol.* **47** (2001) Sept., 65–67.
266. L. K. English, *Mater. Eng. (Cleveland)* **106** (1989) Feb., 47 – 51.
267. W. B. Glenn, *Plast. Technol.* **31** (1985) June, 105 – 108.
268. R. Burns, *Plast. Technol.* **34** (1988) Feb., 60 – 65.
269. A. S. Wood, *Mod. Plast.* **66** (1989) Feb., 44 – 49.
270. M. Gabriele, *Plast. Technol.* **36** (1990) Mar., 55 – 57.
271. H. Frank, *Kunststoffe* **76** (1986) 853 – 855.
272. G. Blinne, E. Priebe, *Kunststoffe* **77** (1987) 988 – 993.
273. United States Code of Federal Regulations, Title 21, Part 177.1500 (Apr., 2002) 265–271.
274. J. H. Schut, *Plast. Technol.* **39** (1993) April, 22–24.
275. J. H. Schut, *Plast. Technol.* **48** (2002) May, 37–39.
276. M. Reisch, *Chem. Eng. News* **79**(38) (2001) Sept. 17, 22. Links
277. A. H. Tullo, *Chem. Eng. News* **78**(4) (2000) Jan. 24, 23–24.
278. R. F. Jones, M. H. Baumann, *Mod. Plast.* **75** (1998) May, 95–96, 98. Links
279. BASF, US 5 294 707, 1994 (R. Kotek).
280. J. H. Schut, *Plast. Technol.* **48** (2002) May, 37–39.
281. Du Pont, US 5 302 756, 1994 (R. J. McKinney).
282. Du Pont, EP-B1 0 740 648, 1998 (R. J. McKinney).
283. Du Pont, EP-B1 0 783 479, 1999 (E. F. Moran, Jr., R. J. McKinney).
284. Du Pont, US 5 430 068, 1995 (P. M. Subramanian).
285. Du Pont, WO-A2 01/94457, 2001 (M. S. McKinnon).
286. Solutia, WO 00/01654, 2000 (M. A. Rodkin et al.).
287. M. McCoy, *Chem. Eng. News* **78**(40) (2000) October 2, 32–34.
288. R. Raja, G. Sankar, J. M. Thomas, *Angew. Chem. Int. Ed.* **39** (2000), 2313–2316. Links
289. United States Code of Federal Regulations, Title 21, Part 177.1500 (Apr., 1989) 240 – 245.
290. M. W. Goldblatt, M. E. Farquharson, G. Bennett, B. M. Askew, *Br. J. Ind. Med.* **11** (1954) 1 – 10. Links
291. S. K. Ritter, *Chem. Eng. News* **79**(29) (2001) July 16, 27–34.
292. P. Mapleston, *Mod. Plast.* **79** (2002) Jan., 36.

[Top of Page]

Related Articles

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in ⟶ Plastics, General Survey; ⟶ Plastics, Properties and Testing; ⟶ Plastics, Processing; ⟶ Plastics, Additives; and ⟶ Plastics, Analysis, respectively. States of order are treated in ⟶ Plastics, Properties and Testing. Fundamental aspects of

polymerization reactions are treated in ⟶ Polymerization Processes

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Polycarbonates

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing ; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

VOLKER SERINI, Bayer AG, Krefeld-Uerdingen, Federal Republic of Germany

1. Introduction 1
2. Properties 2
2.1. Bisphenol A Polycarbonate 2
2.2. Modified Bisphenol A Polycarbonate 3
2.3. Bisphenol A Polycarbonate with Additives 4
3. Production 5
3.1. Interfacial Polycondensation 5
3.2. Melt Transesterification 6
4. Processing 6
5. Uses 7
6. Economic Aspects 8
7. Recycling 8
8. Toxicology and Environmental Aspects 9
9. References 9

1. Introduction

The economically most important polycarbonate is 2,2-bis(4-hydroxyphenyl)propane polycarbonate (**1**), also termed bisphenol A polycarbonate [*24936-68-3*] (BPA-PC).

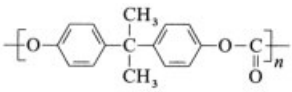

1

It was discovered by SCHNELL, BOTTENBRUCH, and KRIMM in 1953 at Farbenfabriken Bayer (now called Bayer), and independently by FOX at General Electric shortly afterwards [1], [2]. The polymer was first produced industrially and introduced to the market in Europe in 1958 by Farbenfabriken Bayer and in the United States in 1960 by Mobay Chemical Corporation (now called Miles Inc.) and General Electric.

The great commercial success of BPA-PC is due to its unique combination of properties: extreme toughness, outstanding transparency, excellent compatibility with several polymers, and high heat distortion resistance.

In addition to bisphenol A (BPA), many other bisphenols have been converted to polycarbonates [1–7]. However, very few of them have been used as monomers in industrial polycarbonate production (Section 2.2).

Examples include 1,1-bis(4-hydroxyphenyl)cyclohexane [*843-55-0*] (bisphenol C, BPC) (**2**), 2,2-bis(3,5-dibromo-4-hydroxyphenyl)propane [*79-94-7*] (tetrabromobisphenol A, TBBPA) (**3**), and 2,2-bis(3,5-dimethyl-4-hydroxyphenyl)propane [*5613-46-7*] (tetramethylbisphenol A, TMBPA) (**4**):

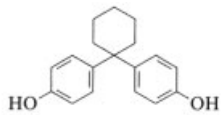

2

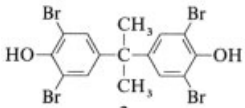

3

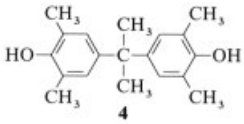

4

BPC is used as a comonomer for bisphenol A in special polycarbonates for cast film production to reduce the crystallization tendency of the polycarbonate during the cast-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_207

ing procedure. TBBPA is cocondensed with bisphenol A to increase the flame resistance of the polycarbonate. TMBPA polycarbonate [4] is highly heat resistant and has a low melt viscosity. It can be mixed with BPA-PC, polystyrene, styrene – acrylonitrile copolymers with a low acrylonitrile content (e.g., 10 wt %), or poly(vinyl chloride) to give a single phase [4], [6]. It was commercially available for several years from 1978 onwards. It is now used in small quantities and blended with BPA-PC, poly(butylene terephthalate), and impact modifiers to elevate the heat distortion resistance of polycarbonates [6]. More recently, 1,1-bis(4-hydroxyphenyl)-3,3,5-trimethylcyclohexane [*129188-99-4*] (BPTMC) (**5**) has become important for the production of highly heat resistant polycarbonates [8] which have been on the market since 1990.

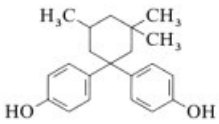

5

BPA-PC may be modified not only by cocondensation of bisphenol A with other bisphenol monomers, but also by cocondensation with blocks of other polymers. Aromatic polyester blocks [9], [10], aliphatic polyether blocks [11], and polysiloxane blocks [6], [12] have become industrially important in this respect (Section 2.2).

2. Properties

2.1. Bisphenol A Polycarbonate

BPA-PCs are soluble in chlorinated hydrocarbons (e.g., chloroform, dichloromethane, and dichloroethane) and nonhalogenated solvents (e.g., pyridine and *m*-cresol). They are insoluble in water, alcohols, and (cyclo)aliphatic hydrocarbons. BPA-PCs are generally amorphous and thus transparent when processed by injection molding or melt extrusion, or when cast from solution into films. They can, however, be crystallized by special treatment, for example, by heating at temperatures between the softening point and the crystalline melting point (ca. 260 °C), by treatment with solvents (e.g., acetone and toluene), or by stretching fibers or films [2].

Table 1. Properties of bisphenol A polycarbonate (Makrolon 2800, Bayer)

Property	Test standard	Value
Ultimate tensile strength, N/mm^2	ISO/R 527	72
Elongation at break, %	ISO/R 527	120
Modulus of elasticity (tensile test), N/mm^2	DIN 53 457-t	2400
Impact strength, kJ/m^2	ISO 179/2 D	unbroken
Notched impact strength, kJ/m^2	analogous to ISO 179/2 C radius 0.04 mm	35
Coefficient of thermal expansion (28 – 80 °C), 10^{-6} K^{-1}	analogous to DIN 53 752	70
Vicat softening temperature, VST B 120, °C	ISO 306	148
Heat deflection temperature, HDT A (1.80 N/mm), °C	ISO 75	131
Flammability (Underwriters Laboratory), 1.6 mm wall thickness	UL 94 (IEC 707)	V-2
Oxygen index	ASTM D 2863-77	26
Density, g/cm^3	ISO/R 1183	1.20
Refractive index n_D	ISO/R 489	1.586
Light transmission (1 mm thick), %	DIN 5036, part 3	88
Moisture absorption (23 °C, 50 % R.H.), %	analogous to DIN 53 495-1 L	0.15
Dielectric strength, transformer oil, 1 mm wall thickness, kV/mm	IEC 243	30
Surface resistivity, Ω	IEC 93	10^{15}
Volume resistivity, Ω · cm	IEC 93	10^{16}
Dielectric constant (dry) (50 Hz/1 kHz)	IEC 250	3.0/3.0
Dielectric loss factor (dry) tan $\delta \times 10^4$ (50 Hz/1 kHz)	IEC 250	9/10

Some properties of a standard BPA-PC are given in Table 1. Industrial BPA-PCs for injection molding and extrusion have weight-average molecular masses ($\bar{M}_W$) of 20 000 – 35 000. BPA-PC melts with molecular masses exceeding 60 000 cannot be processed due to their extremely high melt viscosity. These high molecular mass polycarbonates can, however, be cast from solution into films with desirable properties (Makrofol, Bayer). See also, →Films, Chap. 6.9.

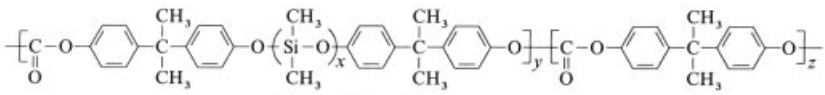

$x = 5-200$ **6**

The toughness and impact strength of BPA-PC are considerably higher than those of other amorphous, single-phase thermoplastics [e.g., polystyrene, poly(methyl methacrylate), and polysulfone]. The tough fracture of BPA-PC changes to a brittle fracture below − 10 °C or when the molding thickness exceeds ca. 6.5 mm [13].

The high transparency of BPA-PC and its high heat resistance (T_g ca. 150 °C which is 50 °C above that of polystyrene) are very important for its uses and applications. The long-term stability of BPA-PC against heat aging in air is ca. 130 °C depending on the test method; the compound starts to decompose at 400 °C. BPA-PC has a low flammability (oxygen index 26). Its flame resistance can be improved considerably by adding small amounts of flame retardants (see Section 2.3).

2.2. Modified Bisphenol A Polycarbonate

Injection molding is generally used to process BPA-PCs with weight-average molecular masses $\bar{M}_W$ of 20 000 – 35 000. At lower molecular masses the mechanical properties (e.g., toughness and elongation at break) deteriorate sharply; within the above range they increase with increasing molecular mass. As the mechanical properties continue to increase above $\bar{M}_W$ 35 000, the melt viscosity becomes very high and prevents injection molding. The mechanical properties of low molecular mass polycarbonates with a low melt viscosity can be improved by introducing branched alkylphenols instead of simple phenols as chain terminators [8], [14].

The introduction of branching points in the polymer chain by incorporating small amounts of polyfunctional compounds (e.g., trifunctional phenols) produces BPA-PCs whose melts have a high structural viscosity [6], [8], [15–18]. These polymers are particularly suitable for melt extrusion because they have a low viscosity under the shearing action in the extruder, and after leaving the extruder nozzle they are highly viscous and thus sufficiently stable. Complicated extrudates, such as twin-wall sheets, can be produced with high accuracy. The structural viscosity of melts of these polycarbonates is also important for the production of large hollow articles by extrusion blow molding (e.g., 20 L water bottles).

BPA-PC exhibits tough fracture and outstanding notched impact strength (ISO 180/4 A) above − 10 °C. Below − 10 °C however, the tough fracture changes to a brittle fracture and the notched impact strength is considerably reduced. The notched impact strength at low temperatures can be improved by incorporation of polysiloxane blocks [6], [12]; this shifts the tough – brittle transition to below − 40 °C. A siloxane content of 2 – 10 wt % produces a satisfactory effect.

Polydimethylsiloxane blocks are coupled to the polycarbonate via the bisphenol A residues (**6**) (see above).

Incorportation of aliphatic polyether blocks into BPA-PC produces hydrophilic polycarbonate – polyether block copolymers, which are of interest as hemodialysis membranes [11].

Partial replacement of the carbonate groups in BPA-PC by terephthalate or isophthalate groups leads to aromatic polyester carbonates with a higher heat distortion resistance and a higher stability against heat aging in air [9], [10].

Depending on the proportion of polyester and the terephthalate – isophthalate ratio, the glass transition temperature is between ca. 150 and 210 °C (i.e., the glass transition temperatures of BPA-PC and bisphenol A – terephthalic acid polyester, respectively). Aromatic polyester carbonates are, like BPA-PC, extremely tough and transparent. However, they have a significantly higher melt viscosity and a pale-yellow color.

The flame-resistance of BPA-PC can be improved by cocondensing small amounts of tetrabromobisphenol A (**3**) with bisphenol A. Partial or total replacement of bisphenol A in BPA-PC by tetramethylbisphenol A (**4**) increases the glass transition temperature from 150 to 203 °C [4] without reducing the flowability in injection molding.

The glass transition temperature of BPA-PC can be increased even further by partial or total replacement of bisphenol A by 1,1-bis(4-hydroxyphenyl)-3,3,5-trimethylcyclohexane [8]; glass transition temperatures of ca. 150 to 240 °C can be achieved in this way. These polycarbonates are transparent, tough, and colorless, even at high glass transition temperatures [8]. Their melt viscosities are favorable compared

with those of aromatic polyester carbonates, polysulfones, polyethersulfones, and polyetherimides having the same glass transition temperature.

BPA-PC is used in polymer blends (→ Polymer Blends, Chap. 4.6.). The most important blends are those with acrylonitrile – butadiene – styrene (ABS) graft polymers [6], [19], [20] and those with poly(butylene terephthalate) or poly(ethylene terephthalate), and impact resistance modifiers (e.g., ABS) [6], [19], [21]. Compared to ABS, BPA-PC – ABS blends have improved heat resistance. They also exhibit good processibility, light stability, surface quality, and high toughness, even at low temperature (tough fracture in the notched impact strength test). Blends of BPA-PC with poly(alkylene terephthalates) and impact strength modifiers have a high heat distortion resistance, and high toughness even at low temperature. Their resistance to gasoline and other solvents is better than that of BPA-PC.

2.3. Bisphenol A Polycarbonate with Additives

Most of the additives discussed in this section are described in detail elsewhere (→Plastics, Additives).

Mold-Release Agents. The addition of small amounts (< 1 wt %) of mold-release agents (e.g., fatty acid esters of higher alcohols, such as pentaerythritol or glycerol) to the polycarbonate improves the mold release of complicated parts [6]. See also → Release Agents.

Flame Retardants. The flame resistance of BPA-PC can be increased by incorporating tetrabromobisphenol (see Section 2.2) or by using additives or additive combinations [6]. For example, flame resistance is significantly improved by the addition of small amounts (< 1 wt %) of special salts, such as potassium perfluoro-*n*-butane-1-sulfonate, sodium 2,4,5-trichlorobenzenesulfonate, and potassium diphenylsulfone-3-sulfonate [22–25]. The tendency to form flaming drips after ignition can be reduced by using branched polycarbonates (see Section 2.2) or by adding small amounts (< 1 wt %) of polytetrafluoroethylene [26].

Light Stabilizers. Ultraviolet light, especially in combination with atmospheric oxygen and moisture, causes decomposition of BPA-PC. Decomposition starts with a photo-Fries rearrangement, which leads to the formation of salicylates, followed by benzophenones [6], [7]. These structures absorb visible light in the short-wavelength region and are thus responsible for the yellowing of polycarbonates. Stabilization against degradation by UV light (i.e., against yellowing) is important for exterior uses and applications and can be achieved by adding small amounts (< 1 wt %) of benzotriazole derivatives [7].

Reinforcement. Glass fibers (< ca. 35 wt %) are generally added to the polycarbonate for reinforcement. They improve rigidity and stiffness, and lower the coefficient of thermal expansion [6], [7], [27].

Miscellaneous Additives. *Heat stabilizers* (e.g., organic phosphites) can be added to the polycarbonate to prevent thermal oxidation during processing at high temperature [6]. *Blowing agents* (e.g., 5-phenyltetrazole and phenyldihydroxo-oxadiazonone) can be added during thermoplastic processing into moldings to obtain low-weight foamed structures (→ Foamed Plastics, Chap. 4.15.) [6]. Transparent or opaque polycarbonates can be obtained in a wide range of colors by using *organic dyes* and *inorganic pigments.*

3. Production

3.1. Interfacial Polycondensation

Interfacial polycondensation is currently used for the industrial production of polycarbonates [2], [5], [6]. Bisphenol A is reacted with phosgene at 20 – 40 °C in a two-phase mixture consisting of an aqueous, alkaline phase and an immiscible organic phase. The overall reaction is as follows:

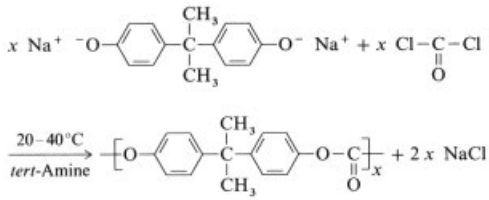

The bisphenol A is first dissolved in the aqueous phase as sodium bisphenolate and the phosgene is dissolved in the organic phase that generally consists of chlorinated hydrocarbons (e.g., dichloromethane). Reaction occurs at the phase interface to produce carbonate oligomers, which then enter the organic phase. A 10 – 20 mol % excess of phosgene (referred to the amount of bisphenol A) is generally used because a small proportion of the phosgene is hydrolyzed. The hydrolysis products (NaCl and Na_2CO_3) enter the aqueous phase. Relatively low pH values (9 – 11) are employed during phosgenation to minimize phosgene hydrolysis.

After the formation of carbonate oligomers, a catalyst is added. Tertiary aliphatic amines (e.g., triethylamine, tripropylamine, and tributylamine) are preferred as catalysts. They accelerate the polycondensation of oligomers to form high molecular mass products and catalyze the hydrolysis of excess terminal chlorocarbonate groups formed due to the excess of phosgene. This is important because the hydrolyzable chlorine contained in these groups would otherwise corrode the molds used during thermoplastic processing. During the polycondensation reaction the aqueous phase is kept at alkaline pH values (9 – 14).

Monofunctional phenols (e.g., 3 – 5 mol % phenol or 4-*tert*-butylphenol referred to bisphenol A) are added as chain terminators to the two-phase mixture before addition of the catalyst, to control the molecular mass. These compounds then form the terminal groups of the polycarbonate bound via carbonate groups.

Polycarbonate production can be carried out batchwise in stirred tanks or continuously in cascades of stirred tanks or in tubular reactors, in which the organic and aqueous phases are finely distributed by shear forces. After completion of polycondensation, the aqueous alkaline phase, which contains sodium chloride and sodium carbonate, is separated from the organic phase by centrifugation. The organic phase is then freed from residual traces of sodium bisphenolate by washing with dilute aqueous sodium hydroxide, followed by washing with dilute acid (e.g., hydrochloric or phosphoric acid) to neutralize the alkali. The organic phase is then washed free of electrolytes with demineralized water. All the wash stages are generally performed in separators.

Industrially, various processes are used to isolate the polycarbonate from solution. The most important are

1) Precipitation with nonsolvents for polycarbonate (e.g., heptane)
2) Precipitation by introduction of the solution into hot water and subsequent evaporation of the polycarbonate solvent
3) Evaporation of the solvent in an extruder

Polycarbonate is generally marketed in the form of granules, which are produced by melting the polymer in an extruder, extruding it as strands, cooling, and comminution. If the polycarbonate is isolated by evaporation of the solvent (chlorinated hydrocarbon) in an extruder, the granulate is obtained directly.

Stabilizing or modifying additives are generally incorporated into the polycarbonate during extrusion or in a subsequent compounding step.

3.2. Melt Transesterification

Diphenyl carbonate is transesterified in the melt with bisphenol A to form polycarbonate [2], [6]:

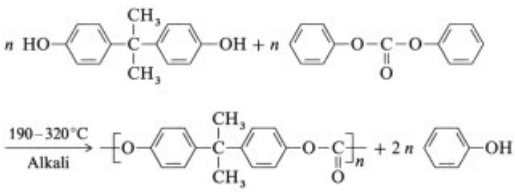

During this process phenol is split off and removed by distillation. Stirred tanks, helical reactors, or screw reactors (extruders) may be used. The reaction is usually catalyzed with very small amounts of alkali at 190 – 320 °C and reduced pressure (down to ca. 0.1 kPa). The temperature is initially low and the pressure is slightly reduced (e.g., 200 °C at 20 kPa). During the course of reaction the temperature is raised and the pressure reduced. Toward the end of the reaction the temperature must be high enough to maintain a sufficiently high melt flow, and the pressure must be as low as possible to remove all the eliminated phenol from the viscous melt. The polycarbonate is finally spun in the form of strands and granulated.

Although the melt transesterification process was the first industrial process for polycarbonate

production, it is no longer important. However, interest in it has revived due to the development of ecologically more favorable, phosgene-free production processes [8]. Reaction of carbon monoxide with alcohols in the presence of catalysts gives dialkyl carbonates that can, in turn, be reacted with phenol to form diphenyl carbonate [8]. The latter can then be reacted with bisphenol A, as described above, to produce polycarbonate by melt transesterification.

The melt transesterification process has also attracted renewed interest because of justifiable optimism that secondary reactions with deleterious effects on quality (e.g., discoloration) can be suppressed. This is achieved by using high-purity starting materials (diphenyl carbonate, bisphenol A), new catalyst systems, new high-viscosity reactors, and an alternative process route [8]. This route consists of the intermediate separation of an oligomeric carbonate produced by melt transesterification, which is then crystallized and subjected to solid-phase polycondensation.

4. Processing

Before thermoplastic processing, BPA-PC must be dried to a residual moisture content of < 0.02 wt % for the extrusion to < 0.01 wt %, usually at 120 °C. In modern, fast driers, drying times of 2 – 5 h are sufficient. It often makes sense to process the dried granulate in injection molding machines or extruders equipped with heated hoppers, so that it does not reabsorb moisture. Hoppers with circulating air heating (120 °C) are recommended. Moisture can also be removed during plasticization in the processing machine by means of a degassing unit. With very high temperatures or long residence times of the polymer melt the moldings should be checked for a decrease in molecular mass.

All modern injection molding machines are suitable for polycarbonate processing. Machines with three-section screws with a length – diameter ratio of 18 : 1 to 22 : 1 and a screw – channel depth ratio of 2 : 1 to 2.5 : 1 are recommended. The heating systems and regulating instruments must be designed for operation up to at least 350 °C: processing temperatures are generally 280 – 320 °C, the mold temperature should be 85 – 120 °C.

High-viscosity polycarbonates can be processed with extruders. Structurally viscous branched polycarbonates are particularly suitable for this (Section 2.2). Hollow articles, films, sheets, profiled sections, pipes, and strands are examples of extruded products.

Semifinished products can also be made from BPA-PC. Slow cooling in air or water from 80 °C without drawing the melt is required to obtain low-stress products. The temperature at the nozzle outlet is ca. 230 – 260 °C, depending on the melt strength. Films and sheets of BPA-PC can be subjected to thermoforming. BPA-PC can also be polished to a high gloss, coated, printed, embossed, or vacuum metallized, and bonded with solvents. Two-pack adhesives (e.g., epoxy, silicone, or polyurethane adhesives) are suitable for bonding BPA-PC parts to one another and to other materials. Parts made of BPA-PC may also be joined to each other by vibration, ultrasonic, friction, heated tool, and hot gas welding. The polymer can also be machined, sawed, and drilled.

High-quality films (2 – 200 μm thick) can be obtained by casting solutions of high molecular mass polycarbonates onto drums or bands (Section 2.1) and evaporating the solvent [7]. These films are particularly useful in the electrical sector (e.g., capacitors).

5. Uses

Electrical Sector. The electrical sector is the classical and major user of BPA-PC [8] because of the special combination of properties: outstanding electrical insulation, good flame retardance, high heat resistance, excellent toughness, dimensional stability, and transparency. The fact that these properties are largely independent of temperature and atmospheric humidity is also very important. Typical applications include housings for car telephones and distributor equipment, lamp sockets, multiple contact strips, relay parts, power plugs, safety switches, components for electronic calculators, fluorescent tube sockets, current meters, and winding supports. See also→Insulation, Electric.

Building and Construction. The building and construction industry has become the second most important user of polycarbonates [8].

The transparency, extreme toughness, and moderate price of polycarbonates make them suitable for window panes and roofing. Singly and multiply glazed panes are fabricated from solid polycarbonate sheets. Twin-wall and triple-wall sheets of polycarbonate combine optimum impact strength, good light transmission, light weight, and thermal insulation with minimal material consumption (low wall thickness) [28]. They are used as roofings for greenhouses, halls, and railway stations; as fences in ice hockey arenas; and as windows for telephone booths and gymnasiums.

Automobile Sector. In automobiles the high impact strength, heat resistance, transparency, and weather resistance of polycarbonates are especially important. Polycarbonate is used, for example, in headlight and taillight covers, reflectors, as well as fan and radiator grills. Polycarbonate – ABS blends have become very important in dashboards, where good heat resistance, high toughness, and tough fracture (splinter-free fracture) are required at low temperatures [6], [8], [29]. Blends of polycarbonate, poly(ethylene terephthalate) or poly(butylene terephthalate), and impact strength modifiers are also important [6], for example, in car bumpers, where low-temperature impact strength (down to $-40\,^{\circ}C$) and resistance to gasoline are necessary. The latter property is affected favorably by partially crystalline polyesters because they improve resistance to gasoline.

Optical Information Storage Systems. The use of BPA-PC for optical information storage systems is described in [6–8]. The best-known product in this area is the audio compact disc (CD), which is a read-only recording medium. For a complete description of optical recording, see→Information Storage Materials. Special polycarbonate grades fulfill the requirements of transparency, low birefringence, mechanical and dynamic stability, good heat distortion resistance and flow properties that are needed for CD production.

The information on a CD is stored digitally on the surface of the transparent polycarbonate in the form of small depressions (pits). The CD is produced by injection molding, and the pits are produced by using a mold with a stamper (→Information Storage Materials, Chap. 2.3.2.). The pits are scanned by a laser to retrieve the stored information (i.e., music). The same storage principle is employed in read-only memory (ROM) systems used to store data. In write-once read-many-times (WORM) memories the information can be "burnt" as a sequence of holes into a layer of rare-earth metal with a powerful laser beam. The sequence of holes can then be read as often as desired with weaker laser beams. This system is also termed direct read after write (DRAW).

The latest storage system, which has just come on the market, is the E-DRAW system (E = erasable). This system functions due to the effect of the different magnetizability of two very thin metal layers on top of one another. Laser beams of different strengths are used to write and read signals. These signals can also be deleted and rewritten as often as desired. As in the CD, ROM, and WORM (DRAW) systems, polycarbonate is used as a substrate for the E-DRAW data storage disc.

Miscellaneous Uses. Further important areas of use of polycarbonates include office equipment (casings, covers, keyboards), solution-cast films and extruded or coextruded films and sheets (cheque cards, cover sheets, capacitor sheets), foams (electrical distributor boxes, housings), medical applications (blood oxygenators, hemodialysis separating membranes, dialyzers, sterilizable equipment), packaging (baby bottles, water tanks, food containers), domestic applications (electric razors, sink units), and photographic or optical applications (camera cases, lenses, optical fibers).

6. Economic Aspects

Polycarbonate together with ABS (in 1989 about 2×10^6 t ABS were sold) and the thermoplastic polyamides (in 1989 about 8×10^5 t PA were sold, without fibers) belong to the top group of engineering thermoplastics [8]. Table 2 shows the volume of polycarbonate sold worldwide (including the PC proportion in blends) by manufacturers in the United States, Western Europe, and Japan (Triade). Production capacities for 1990 and those predicted for 1995 are also listed. Outside the Triade only Brazil had a low polycarbonate capacity (about 8000 t) in 1990. By

1995, however, important polycarbonate capacities (ca. 1×10^5 t) could exist in Korea, Brazil, and Taiwan [31].

Table 2. Worldwide polycarbonate sales volume (including the PC proportion in blends) and production capacities (10^3 t) [30]

Region	1990		1995 (predicted)	
	Sales	Capacity	Sales	Capacity
United States	260	353	356	427
Western Europe	186	266	250	363
Japan	110	153	162	238
Total	556	772	768	1028

The largest polycarbonate manufacturers are General Electric (1989: capacity 2.88×10^5 t, 45 % of world capacity of 6.36×10^5 t) and Bayer (1989: capacity 1.95×10^5 t, 31 % of world capacity) [31]. Trade names of General Electric products include Lexan (PC), Lexan PPC (aromatic polyester carbonate), Xenoy (PC – poly(butylene terephthalate) blends), and Cycoloy (PC – ABS blends). Trade names used by Bayer include Makrolon (PC), Makrofol (PC films), Makroblend (PC – polyester blends), Bayblend (PC – ABS blends), Apec (aromatic polyester carbonate), and Apec HT (BPTMC-based PC).

The amount of PC used in various sectors is subdivided as follows [8]:

Electrical/electronics, optical information storage media	49 %
Building and construction	24 %
Automobiles	8 %
Office equipment	6 %
Miscellaneous	13 %

7. Recycling

Polycarbonates exhibit a favorable behavior in both material recycling (reuse of polycarbonate after granulation of polycarbonate parts which are no longer required and are without chemical treatment) and chemical recycling. Chemical recycling is achieved by dissolving the polycarbonate in industrial solvents (e.g., chlorinated hydrocarbons), subsequent purification by filtration, and evaporation of the solvent [32]. In an alternative chemical recycling melted, the polycarbonate can be hydrolyzed, the bisphenol A is then recovered and reused in polycarbonate production [7], [33], [34].

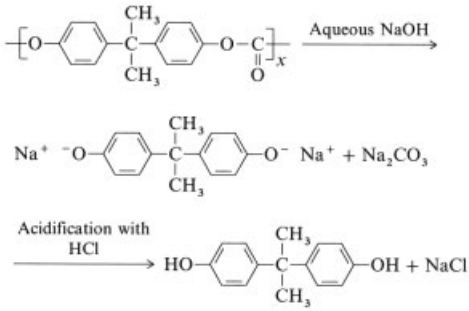

For effective recycling operations, old polycarbonate parts that are no longer required (e.g., in automobiles or electrical equipment) must be collected separately. This constitutes a logistics problem that is not always easy to solve.

8. Toxicology and Environmental Aspects

Harmful effects of polycarbonates are not known [35]. Polycarbonates have not proved toxic when included in rat feed at a concentration of 6 % [36].

The toxicity of the gases produced by combustion of flame-resistant polycarbonates (according to DIN E 5346/37) is similar to or lower than that of wood, wool, or cotton [37]. For safety precautions, exhaust gas and wastewater control, and toxicology of phosgene (Section 3.1), see →Phosgene.

Polycarbonates can be used in accordance with the relevant laws for the production of toys, medical apparatus, and articles that are in contact with food provided that their composition complies with detailed regulations of the respective countries (e.g., Recommendation XI of the BGA, the German Materials and Articles Ordinance, the EEC Directive (90/128) on materials and articles for food contact, and the U.S. FDA regulation 21 CFR § 177.1580).

9. References

1. W. F. Christopher, D. W. Fox: *Polycarbonates,* Reinhold Publishing, New York 1962.
2. H. Schnell: *Chemistry and Physics of Polycarbonates, Polymer Reviews,* vol. 9, Interscience Publishers, New York 1964.

3. G. Peilstöcker, H. Krimm: "Die Polycarbonate" in R. Vieweg, L. Goerden (eds.): *Kunststoff-Handbuch,* vol. 8; Polyester, Carl Hanser Verlag, München 1973, pp. 1 – 245.
4. V. Serini, D. Freitag, H. Vernaleken, *Angew. Makromol. Chem.* **55** (1976) 175 – 189.
5. H. Vernaleken: "Polycarbonates" in F. Millich, C. E. Carreher, Jr. (eds.): *Interfacial Synthesis,* vol. II, Marcel Dekker, New York 1977, pp. 65 – 124.
6. D. Freitag, U. Grigo, P. R. Müller, W. Nouvertné: "Polycarbonates" in H. F. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 11, John Wiley and Sons, New York 1988, pp. 648 – 718.
7. R. Pakull, U. Grigo, D. Freitag: "Polycarbonates" in R. Dolbey (ed.): *Rapra Review Reports,* Report 42, vol. 4, no. 6, Pergamon Press, Oxford 1991, pp. 1 – 30.
8. D. Freitag, G. Fengler, L. Morbitzer, *Angew. Chem. Int. Ed. Engl.* **30** (1991) 1598.
9. D. Rathmann, *Kunststoffe* **77** (1987) 10.
10. D. Rathmann, P. Tacke: "Polyarylate" in L. Bottenbruch: *Kunststoff-Handbuch,* 2nd ed., vol. 3/3: Technische Thermoplaste, Carl Hanser Verlag, München (in preparation).
11. R. Dhein et al., *Kunststoffe* **81** (1991) 972.
12. Bayer AG, DE-OS 3 334 782, 1983 (W. Paul, W. Nouvertné, H. Steinberger, W. Grape).
13. L. Morbitzer, U. Grigo, *Angew. Makromol. Chem.* **162** (1988) 87 – 107.
14. Bayer AG, DE-OS 2 842 005, 1980 (D. Freitag, W. Nouvertné, K. Burkhardt, F. Kleiner).
15. Farbenfabriken Bayer AG, GB 1 079 821, 1967 (H. Schnell et al.).
16. Farbenfabriken Bayer AG, GB 1 151 304, 1969 (H. Schnell et al.).
17. Bayer AG, DE-OS 2 113 347, 1972 (D. Freitag, U. Haberland, H. Krimm).
18. Bayer AG, DE-OS 2 500 092, 1976 (K. Idel, D. Freitag, W. Nouvertné).
19. W. Nouvertné, H. Peters, H. Beicher, *Plastverarbeiter* **33** (1982) 1070.
20. W. Witt, *Kunststoffe* **73** (1983) 498.
21. W. Witt, *Gummi Fasern Kunstst.* **38** (1985) 155.
22. V. Mark, *Org. Coat. Plast. Chem.* **43** (1980) 71.
23. J. L. Webb, *Org. Coat. Plast. Chem.* **43** (1980) 79.
24. Bayer AG, US 3 775 367, 1973 (W. Nouvertné).
25. L. S. Thomas, S. A. Ogoe, *Tech. Pap. Reg. Tech. Conf. Soc. Plast. Eng.* 1985, 100.
26. K. Kircher, *Kunststoffe* **74** (1984) 12.
27. K. Kircher, *Reinforced Plastics* **25** (1981) p. 382 and 25 (1981) p. 420.
28. K. Kircher, *Kunststoffe* **80** (1990) 1113.
29. L. Morbitzer, H. J. Kreß, C. Lindner, K. H. Ott, *Angew. Makromol. Chem.* **132** (1985) 19.
30. Chem. Systems Inc.: "Polycarbonates – Non-Phosgene Routes," New York, April 1991.
31. *Chemical Economics Handbook,* Polycarbonate Resins Report, SRI International, Menlo Park, Calif., Feb. 1990.
32. Bayer AG, DE-OS 4 022 011, 1990 (G. Fennhoff et al.).
33. Bayer AG, EP 0 000 904, 1980 (K. Idel).
34. General Electric, EP 0 355 319, 1989 (D. W. Fox, E. N. Peters).
35. G. Bornmann, A. Löser, *Arzneim. Forsch.* **9** (1959) 9 – 13.
36. R. R. Montgomery in: *Patty's Industrial Hygiene and Toxicology,* 3rd ed., vol. II c, Wiley-Interscience, New York 1982, p. 4441.
37. G. Kimmerle, F. H. Prager, *J. Comb. Tox.* **7** (1980) 42.

Polychlorotrifluoroethylene → Fluoropolymers, Organic
Polyester Fibers → Fibers, 4. Synthetic Organic

Polyester Resins, Unsaturated

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, Generals Survey; Plastics, Properties and Testing ; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing . Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Horst Krämer, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany

1. Introduction 1
2. Raw Materials 1
3. Production 3
4. Additives 4
5. Curing 5
6. Types of UP Resins 5
7. Testing and Properties 6
8. Uses and Economic Aspects 8
9. Occupational Health and Environmental Aspects 9
10. References 10

1. Introduction

Unsaturated polyester resins (abbreviated as UP resins according to DIN 7728, part 1) undergo cross-linkage to form thermosetting plastics (DIN 16 945). Most UP resins are medium- to low-viscosity, liquid solutions of unsaturated polyesters in copolymerizable monomers (usually styrene) as cross-linking agents.

Unsaturated polyesters are formed by polycondensation of unsaturated and saturated dicarboxylic acids with diols. UP resins are hardened (cured) or cross-linked by free-radical polymerization with initiators (e.g., peroxides) and accelerators at normal temperature (10–40 °C), or with peroxides alone above 80 °C. The unsaturated groups of the polyester chains react with the double bonds of the copolymerizable solvent monomer. Solid products are formed that are used as molding materials or in paints. They are also reinforced with fibers or filled, and used as putties or mortars.

Vinyl ester resins, which are also termed phenacrylate resins (DIN 16 946, part 2), are polyadducts of liquid epoxy resins and acrylic acid in styrene. Strictly speaking, they are not unsaturated polyester resins, but since they are processed, cured, and used similarly, they are also discussed.

2. Raw Materials

The properties of unsaturated polyesters depend on the nature and quantitative ratios of the raw materials (α-, β-unsaturated dicarboxylic acids, saturated dicarboxylic acids, and diols), as well as on the condensation conditions. Many different materials are quoted in the patent literature and in reference manuals [1–5]. However, only those described here are of practical importance. This limited number of raw materials has hardly changed since the early 1970s and permits wide variation as regards the processing and properties of UP resins.

A larger number of double bonds in the polyester chain increases the reactivity, the heat of reaction released during curing, and the thermal stability of the end product. The incorporation of long-chain dicarboxylic acids or diols increases the flexibility of the cured substance.

Unsaturated Dicarboxylic Acids. Maleic anhydride and fumaric acid are used almost exclusively as unsaturated acid components. They are described in detail elsewhere, see → Maleic and Fumaric Acids.

Maleic anhydride [*108-31-6*] is not as expensive as fumaric acid and contains one mole of water less than the free dicarboxylic acid, which is favorable for condensation. During condensation maleic acid largely rearranges into the isomeric form, fumaric acid. This rearrangement can also be promoted by catalysts.

Fumaric acid [*110-17-8*] is only used when the degree of rearrangement of maleic acid to fumaric acid is insufficient for the required properties of the UP resins. Use of fumaric acid instead of maleic anhydride in the condensation

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_217

improves the thermal stability and mechanical properties of the product. The disadvantages of fumaric acid compared with maleic anhydride are its higher cost and higher water content.

Saturated Dicarboxylic Acids. *Phthalic acid* [*88-99-3*] is used exclusively in the form of phthalic anhydride [*85-44-9*] (→ Phthalic Acid and Derivatives). On account of its favorable price and diverse potential uses, it is employed in all standard resins.

Isophthalic acid [*121-91-5*] (→ Terephthalic Acid and Dimethylterephthalates) has a better thermal stability, mechanical strength, and resistance to chemicals than phthalic acid.

Terephthalic acid [*100-21-0*] (→ Terephthalic Acid and Dimethylterephthalates) is used in resins where high impact strength, thermal stability, and weather resistance are required.

Tetrahydrophthalic anhydride [*935-79-5*] yields resins having a high impact strength but limited weather resistance. It is used primarily for resins employed in paints and automotive body fillers.

Adipic acid [*124-04-9*] (→ Adipic Acid) is used on account of its flexible carbon chain to produce soft resins and mixed resins that are employed to plasticize harder resins.

Hexachloroendomethylenetetrahydrophthalic (HET) acid [*115-28-6*] yields resins with low flammability and good chemical resistance.

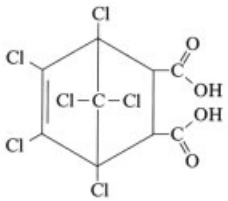

Tetrabromophthalic acid [*13810-83-8*] also lowers flammability, but tends to produce strong discoloration in sunlight due to the release of bromine.

Diols. *Ethylene glycol* [*107-21-1*] (→ Ethylene Glycol) is inexpensive, but it sharply reduces the styrene compatibility of UP resins. It is generally used in combination with other diols such as diethylene glycol, dipropylene glycol, or butylene glycol.

1,2-Propanediol [*57-55-6*], propylene glycol, is by far the most widely used diol for standard resins (→ Propanediols). It has good styrene compatibility, only slight tackiness after curing in air, and good mechanical properties.

Diethylene glycol [*111-46-6*] and *dipropylene glycol* [*110-98-5*] result in plasticization of the resins and increased water absorption. They are used mainly in combination with 1,2-propanediol to adjust the mechanical properties or refractive index of the cured resin.

1,3-Butanediol [*107-88-0*], butylene glycol, (→ Butanediol) improves weather resistance, but is not used as often as the aforementioned diols on account of its high price and limited availability.

Neopentyl glycol [*126-30-7*], 2,2-dimethyl-1,3-propanediol (→ Alcohols, Polyhydric, Chap. 2.2.1.) increases the thermal stability, chemical resistance, and weather resistance of UP resins. Particularly when combined with isophthalic acid, it yields high-quality products for use in chemical plant construction.

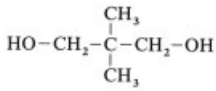

Ethoxylated or propoxylated bisphenol A [*80-05-7*] is very often used for resins with a high thermal stability and extremely good resistance to chemicals. With resins formulated from these diols it should be borne in mind that traces of phenolic groups can interfere with curing.

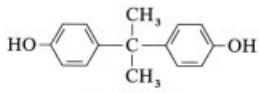

Bisphenol A

Although *bis(hydroxymethyl)tricyclodecane* [*26896-48-0*] is expensive, it imparts high thermal stability and chemical resistance, with tack-free hardening of uncovered surfaces.

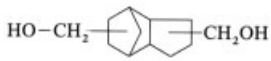

Dibromoneopentyl glycol [*3296-90-0*] is used to reduce combustibility. A disadvantage is that hydrogen bromide readily splits off during polycondensation and attacks the walls of

the reaction vessel. Dibromoneopentyl glycol is therefore seldom used.

N,N-Bis(2-hydroxyethyl)-p-toluidine [*3077-12-1*] is sometimes incorporated into the polyester chain by condensation to obtain resins that cure at normal temperatures with benzoyl peroxide but without an accelerator.

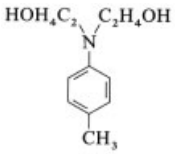

Monomers for Cross-Linking UP Resins. Vinyl, allyl, and acrylic compounds are suitable monomers for the cross-linking of UP resins. They are readily miscible with the resins and their homopolymerization rate does not exceed the copolymerization rate with the unsaturated polyester chain. In most cases low- to medium-viscosity UP resins are required. *Styrene* [*100-42-5*] is by far the most suitable monomer for these products (→ Styrene). The styrene content is normally between 25 and 45 wt %.

The following vinyl compounds are often recommended as alternatives, but are of little practical importance:

α-Methylstyrene [*98-83-9*] is insufficiently reactive. *4-Methylstyrene* [*622-97-9*] would be the most suitable alternative to styrene for technical reasons [6], but has an unpleasant odor and is not produced on an industrial scale.

Vinyltoluene, an isomeric mixture mainly of 4- and 3-methylstyrenes, is not very suitable on account of its low reactivity and unpleasant odor.

Divinylbenzene [*1321-74-0*] is not available in large amounts.

Methyl methacrylate [*80-62-6*] copolymerizes with difficulty with unsaturated polyesters. It is used in combination with styrene, however, to improve weather resistance and to lower the refractive index. The ratio of the refractive index of the resin to that of the glass fiber reinforcement determines the transparency of finished parts such as domelights and light panels.

Diallyl phthalate [*131-17-9*] is used as a cross-linking agent in the production of pourable, solid UP resin molding compositions. The viscosity of diallyl phthalate is too high for liquid UP resins and copolymerization is too slow for curing at room temperature.

Triallyl cyanurate [*101-37-1*] is also a conventional cross-linking agent for solid molding compositions.

3. Production

UP resins are usually produced by batchwise condensation of dicarboxylic acids or anhydrides with diols. Continuous esterification of acid anhydrides and epoxides may, however, also be employed.

Batchwise Condensation. Dicarboxylic acids are esterified with diols at 180 – 230 °C in a nitrogen atmosphere in heatable, coolable reactors while stirring. Esterification is generally performed with a slight excess (up to 10 %) of diols. The water formed in the reaction is distilled off. Passage of nitrogen through the reaction mixture promotes removal of water and increases the esterification rate considerably. The end point of the condensation is determined by measuring the acid number and, possibly, also the hydroxyl number and the viscosity. Commercial UP resins have a mean molecular mass of 2000 – 4000.

The still liquid melt is cooled and then dissolved, while stirring, in styrene in coolable mixers. The styrene contains inhibitors such as hydroquinone or *tert*-butylcatechol (see Chap. 4).

Even if the same raw materials are used, the resin products may differ as regards their processing properties and end properties depending on the amount of excess glycol, the temperature, and reaction conditions.

In the case of raw materials with widely differing esterification rates, two-stage condensation must be used to obtain uniform buildup of the polyester chain and thus favorable end properties. For example, if isophthalic acid and maleic acid are added simultaneously to the reaction vessel, maleic acid is incorporated more quickly into the polyester chain. In the case of isophthalic acid resins, isophthalic acid should therefore first be condensed with glycol in a molar ratio of 1 : 2 to form a bishydroxy ester. Esterification should be continued after adding maleic anhydride.

Azeotropic esterification is important only for certain paint resins. An entraining agent (e.g., toluene or xylene) is added and during condensation the water is distilled off together with the toluene. The toluene is then separated from the water with a dephlegmator and returned to the reactor. At the end of the reaction the toluene must be removed from the resin (e.g., with a thin-film evaporator).

Continuous esterification of acid anhydrides and epoxides is a very elegant, economic method. However, it is limited to the small number of raw materials that are available as anhydrides or epoxides. It cannot, therefore, meet the demands of the polyester processors for different resin types. Thus, for example, the extremely important isophthalic and terephthalic acid resins cannot be produced by this method because the dicarboxylic acids do not form anhydrides. The method is recommended only for producers with very large sales of standard resins.

The raw materials (e.g., maleic anhydride, phthalic anhydride, and propylene oxide) are converted into polyesters by reacting them for ca. 20 min and 200 °C, generally in the presence of catalysts. In contrast to batchwise condensation, no water of reaction is formed.

4. Additives

See also → Plastics, Additives

Inhibitors are used to prevent premature polymerization and to modify the reaction times. The most important are hydroquinone, 1,4-benzoquinone, *tert*-butylpyrocatechol (*tert*-butylcatechol), toluhydroquinone (methylhydroquinone), and 3,4-di-*tert*-butyl-*p*-cresol (3,4-di-*tert*-butyl-4-hydroxytoluene). All these inhibitors have specific effects and are used for different resin types. In some cases they are added prior to condensation, in other cases they are dissolved in styrene and added after condensation.

Optical brighteners are added to slightly yellowish resins for certain uses (e.g., transparent products or white buttons). Blue dyes, which become decolorized during curing, are commonly used as curing indicators.

Commercially available *light stabilizers* (e.g., substituted benzophenones and hydroxyphenyltriazoles) prevent premature yellowing of the hardened resins in sunlight.

Thixotropic agents have to be added to prevent sagging or dripping of the resins when applied to vertical surfaces. Finely dispersed silica is mainly used for this.

Most UP resins have a tacky surface when cured in air. Tackiness can be prevented by adding *skin-forming agents*, generally paraffins. The formation of a paraffin film also reduces styrene emission in pressureless processing. Resins containing such additives are termed "environmentally friendly" or "milieu polyesters".

The additives listed above are generally added to the resin by the resin manufacturer. The following substances are added to the resin by the processor.

Hardened UP resins have a surface resistance of $>10^{13}\,\Omega$. In order to prevent buildup of electrostatic charge, graphite and metal powders can be added as *antistatic agents* to reduce the surface resistance to ca. $10^4\,\Omega$.

Flame retardants include chlorinated paraffins and antimony trioxide. Aluminum hydroxide (hydrated alumina) is employed on a large scale for parts used in the electrical industry that have to be halogen-free (halogens can attack copper contacts and hydrogen chloride may be released during fires).

Both soluble *dyes* and insoluble *pigments* can be used for coloring. Pigments have to be made into a paste in resins or plasticizers before being added.

Fillers include chalk, quartz flour, and quartz sand. The type and amount of filler depend on the final use and desired properties of the resin. Filled resins serve as putties, compression molding compounds, casting compounds, mortars, and concrete. The fillers must be dry so that they do not to affect curing.

Blowing agents are used to produce polyester foams, various additives and systems are used for this (e.g., azodicarbonamide, *tert*-butylhydrazinium chloride and iron(III) chloride).

Thickening agents are required for the production of sheet molding compounds (SMC), and bulk molding compounds (BMC). Magnesium oxide and, in some cases, magnesium and calcium hydroxides are normally used. These

agents are mixed with fillers and used to impregnate glass fibers, thereby converting the molding compound into a highly viscous state.

Low-shrink and low-profile additives are solutions of thermoplastics in styrene and are added to reduce volume shrinkage occurring during curing. Polystyrene, poly(vinyl acetate), poly-(methyl methacrylate), saturated polyesters, cellulose acetobutyrate, and polycaprolactams are usually used. Polyethylene with a particularly low melting point is used to improve the properties of sheet molding compounds. On account of its insolubility it has to be added to the resin batch as a powder immediately before processing.

5. Curing

At normal temperatures UP resins can be cured with peroxides and accelerators. The most important systems include hydroperoxides (e.g., methyl ethyl ketone peroxide, cyclohexanone peroxide, and acetylacetone peroxide) with cobalt octanoate or, in exceptional cases, also vanadium octanoates or naphthenates, as well as benzoyl peroxide (dibenzoyl peroxide) with aromatic amines (e.g., *N,N*-dimethylaniline, *N,N*-diethylaniline, and *N,N*-dimethyl-*p*-toluidine).

Moisture slows down or prevents curing with hydroperoxides and cobalt salts, but does not affect reactions with benzoyl peroxide and amines. Complete curing generally requires heating above the glass transition temperature of the cross-linked polyester. Only completely cured products exhibit optimum mechanical properties and resistance to weathering and chemicals.

At elevated temperatures only peroxides are required for curing, accelerators are unnecessary. Examples include dibenzoyl peroxide above ca. 80 °C, *tert*-butyl 2-ethylperhexanoate above ca. 100 °C, and *tert*-butyl perbenzoate above ca. 120 °C.

UP resins can also be completely cured in a few minutes by exposing them to UV light from high-pressure mercury vapor lamps in the presence of photosensitizers, e.g., benzoin ether [*24650-42-8*] (ω,ω-dimethoxy-ω-phenylacetophenone).

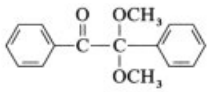

Some acyl phosphine oxide compounds [7] can be used in combination with the aforementioned UV sensitizers to allow curing in normal daylight.

Detailed information about optimum curing systems and reaction conditions for specific uses are given in the data sheets published by resin and reactant manufacturers.

During the transition from the liquid to the solid state, UP resins shrink by 5 – 8 % depending on their double bond content. Compensation of shrinkage by thermoplastics additives is discussed in Chapter 4.

6. Types of UP Resins

Many types of UP resins are available. Selection depends on processing properties and end properties. Processing properties include reactivity, viscosity, and thickening behavior with magnesium oxide. End properties include mechanical behavior, thermal stability, flammability, resistance to chemicals, and color.

Resins may be liquid (low-, medium- or high-viscosity) or solid. Their reactivity may be low, average, or high. A wide range of UP resins can be produced by appropriate choice of the following factors: type and quantity of raw materials, degree of condensation, acid and hydroxyl numbers (and thus the nature of the terminal groups), and the content of cross-linking monomers and other additives. Consequently, comparison or exchange of related resin types from different producers may be difficult.

Standard resins are formulated from inexpensive raw materials and have a wide variety of uses. They can be supplied with or without additives (light stabilizers, thixotropic agents, accelerators, or UV sensitizers).

Elastic grades have a high flexibility and elongation at break due to the incorporation of long-chain monomers, and are generally used in admixture with more brittle types. Some highly elastic grades are used for jointless roof coverings on account of their weather resistance and

high elongation at break of up to 200 % in tensile tests.

Flame-resistant resins generally contain chlorine or bromine in the polyester chain or in added compounds. Flammability can be adjusted to satisfy widely differing regulations and standards by using other additives such as phosphorus compounds, antimony compounds, and aluminum hydroxide.

Resins with a high chemical resistance are mainly synthesized from special raw materials such as isophthalic acid, terephthalic acid, neopentyl glycol, ethoxylated or propoxylated bisphenol A or bis(hydroxymethyl)tricyclodecane. They are used particularly in chemical apparatus (pipes and vessels).

Resins for *sheet molding compounds (SMC) and bulk molding compounds (BMC)* have specially adjusted ratios of acid and hydroxyl numbers and a constant water content. This ensures controlled thickening with magnesium oxide (MgO). Low-shrink and low-profile additives may also be added to the individual types, compatibility depends on the degree of unsaturation [8].

Resins for *fillers and putties* are preaccelerated with aromatic amines. This ensures accurate adjustment of curing with benzoyl peroxide and allows sanding within a short period of time. Rigid and plasticized resins are available for repair putties and light-colored products for marble putties.

Resins for free-flowing *compression molding compounds* are generally supplied monomer-free in powder form to the molding compound producers, who formulate the compounds with cross-linking agents (e.g., diallyl phthalate or triallyl cyanurate) and fillers.

Table 1. Properties of styrene-containing UP resins in the as-supplied state

Property	Test method	Test value
Acid number	DIN 53 402	10 – 30 mg KOH/g
Styrene content	DIN 16 945	25 – 45 %
Viscosity at 25 °C	DIN 53 015	0.2 – 4.5 Pa · s
Density at 20 °C	DIN 51 757	1.07 – 1.3 g/cm^3
Flash point	DIN 51 755	35 °C
Refractive index at 25 °C	DIN 53 491	1.524 – 1.559

Trade Names. Important trade names of UP resins in Europe and the United States include:

Alpolit	Hoechst
Altek	Alpha Corp.
Aropol	Ashland Chemical Corp.
Beetle	BIP
Civic	Neste Oy
Crystic	Scott Bader
Hetron	Ashland Chemical Corp.
Lamellon	DSM
Leguval	Bayer and DSM
Marco	US-Steel Corp.
Norpol	AS Jotun-Group
Norsodyne	Cray Valley
Palatal	BASF
Polylite	Reichhold Chemicals
Roskydal	Bayer
Selectron	PPG Industries
Sniatron	Alusuisse
Sirester	SIR
Torolithe	Routtand S. A.
Trusurf	OCF Fiberglas Corp.
Uralam	Synthetic Resins
Vestopal	Hüls
Viapal	Vianova Kunstharz

Due to company takeovers or mergers and sales of know-how and trade names, assignment of trade names to manufacturers may change.

7. Testing and Properties

A distinction should be made between the properties of UP resins in the as-supplied state (Table 1) and properties after curing (Table 2).

In order to determine the properties in the as-supplied state, simple, rapid tests are desired by producers and customers for quality assurance. Testing is generally carried out by agreement between suppliers and customers according to national standards, but also according to in-house procedures. The resins are often dispatched with test certificates describing the contents of the delivered batch.

Important properties used to characterize the liquid resins in the as-supplied state include the acid number, solids content, viscosity, refractive index, and reactivity. The acid number, solids content, and viscosity can be used to evaluate the degree of condensation and mean molecular mass. Further important properties of liquid resins are color, flash point, and density.

The acid number is measured by titration with potassium hydroxide solution. The solids content can be determined with automatic thermobalances by evaporating the volatile constituents (basically styrene). Viscosity is

checked with rotating-cylinder viscometers, but also by measuring efflux times from temperature-controlled flow cups. The viscosity can be lowered by dilution with styrene or other monomers. The refractive index is measured in refractometers, and is used to monitor the correct composition of the resin.

Table 2. Properties of cured UP resins without fillers or reinforcement fibers (without soft resins)

Property	Test method	Test value
Density at 23 °C	ISO/R 1183	1.20 – 1.40 g/cm^3
Flexural strength	ISO 178	80 – 160 N/mm^2
Outer fiber strain	ISO 178	2 – 7 %
Impact strength	ISO 179 small test bar	6 – 15 kJ/m^2
Tensile strength	ISO/R 527 test specimen 1	40 – 90 N/mm^2
Elongation at break	ISO/R 527 test specimen 1	1.5 – 5 %
Compressive strength	ISO 664	150 – 180 N/mm^2
Ball indentation hardness 10″	ISO 2039	200 – 245 N/mm^2
Barcol hardness (934-1)		35 – 48
Modulus of elasticity (bending)	DIN 53 457 Section 2.3.	3400 – 4300 N/mm^2
Mean specific heat		1.10 – 1.45 kJ kg^{-1} K^{-1}
Thermal stability (heat distortion)	ISO 75 method A	60 – 145 °C
Thermal conductivity	DIN 52 612	0.16 – 0.18 W m^{-1} K^{-1}
Coefficient of linear expansion	VDE 0304 Part 1 § 4	60 – 115 10^{-6} K^{-1}
Volume resistivity	IEC Publ. 93	$>10^{15}$ Ω · cm
Surface resistivity	IEC Publ. 167	$>10^{13}$ Ω
Dielectric dissipation factor		
at 800 Hz	IEC Publ. 250	0.003 – 0.005
at 1 MHz		0.015 – 0.035
Relative permittivity at 800 Hz	IEC Publ. 250	3.0 – 3.5
Comparative tracking indices	IEC Publ. 112	CTI>600
		CTI>600M
Dielectric strength	IEC Publ. 245	35 – 45 kV/mm

The reactivity of UP resins is generally determined by measuring the rise in temperature that occurs during curing. This method is more sensitive than differential scanning calorimetry. The initial temperatures chosen often depend on the subsequent processing temperatures (e.g., room temperature or the temperature of hot compression molds). Measurements differ widely as regards heating bath, vessels, thermocouples, amount of resin, and nature and amount of the reactants. Accurate comparisons of product data sheets published by various manufacturers cannot therefore be made. Examples of measurement setups are given in DIN 16 945.

The thickening of resins for SMC and BMC with magnesium oxide is an important property but is difficult to measure quickly. A future ISO standard will involve accelerated thickening at elevated temperature, and allows conclusions to be made about thickening behavior at room temperature. This test is satisfactory for resin processors, but resins with the same results may behave differently when tested under normal thickening conditions at 25 °C.

In most cases cured UP resins are relatively brittle, addition of fillers and reinforcing materials (e.g., glass fibers) is required to obtain technically useful properties. Nevertheless, cured UP resins must be tested in the pure state. Production of test specimens requires care and experience because factors such as internal stress arising during curing can falsify the results. Mechanical, thermal, and electrical properties are measured (Table 2). The tensile strength, elongation at break, and thermal stability are particularly important for comparing the performance of different resins.

8. Uses and Economic Aspects

UP resins are only used without fillers or reinforcing agents in small amounts for embedding decorative objects, in furniture lacquers, adhesives, paints (→ Paints and Coatings, Chap. 2.8.) and casting resins (impregnating resins) for insulating electrical equipment (→ Insulation, Electric).

The majority of UP resins are reinforced with glass fibers which impart high strength and rigidity to molded parts. Natural fibers (e.g., jute) and synthetic fibers are also used in elastic molded parts.

Casting compounds, putties (→ Paints and Coatings, Chap. 2.8.), and UP resin concrete are produced using fillers with different particle sizes and compositions.

The most important uses of UP resins follow:

Automobile Construction. Parts for special-order vehicles, caravans, and agricultural vehicles are produced by hand from glass-fiber-reinforced UP resins. Flat sheets and panels for truck superstructures and containers are pro-

duced industrially in continuous laminating processes. Body parts for automobiles and commercial vehicles are produced from preimpregnated SMC.

On account of their outstanding resistance and impact strength behavior, pressed parts made from UP resins reinforced with natural or synthetic fibers are employed in car interior fittings, engine compartments, and vehicle underbodies; they have gained a considerable market share. Filled UP resins are used as automobile repair putties.

Building Industry. The use of light panels made of glass-fiber-reinforced UP resins has greatly declined. UP resins with glass fiber reinforcement are used for domelight crowns, facade elements, cellar window frames, and light shafts made from SMC. Beacons, chimneys, and large aerials are wound from UP resins and glass fibers. Highly elastic casting compounds are used for flat roof coverings. Drainage systems with channels, catch basins, and gasoline separators are produced industrially from polyester concrete. Imitation marble and onyx can be produced from UP resins, colorants, and chalk (marble) or aluminum oxyhydrate (onyx); they are used to produce inexpensive, high-quality sanitary ware.

Electrical Industry. The electrical industry is still a main customer for molded parts made from SMC and BMC. Examples include cable distribution cupboards and switching cupboards, skirting boards for lamps, multiway connectors, switch covers, and clamping sleeves. Coils and windings are impregnated with UP resins.

Pipes, Vessels, and Apparatus. Cylindrical molded components reinforced with glass fibers for the chemical industry and oil transportation are fabricated in winding and centrifugal processes from chemically resistant resins whose chemical resistance is appropriate for the intended use. Glass-fiber-reinforced UP resins are also used for fertilizer and fodder silos and cooling towers.

Boat and Shipbuilding. Glass-fiber-reinforced UP resins have long been used for building sports, life-saving, and fishing boats. Larger vessels (e.g., 50-m long minesweepers with 30-mm thick sides) are now also being built from this material.

Miscellaneous Uses. Other molded parts made from glass-fiber-reinforced UP resins include machinery covers and casings, sports equipment, toys, garden furniture, and traffic signs. Important areas of use of nonreinforced UP resins include buttons, buckles, fashion accessories, and furniture finishings.

Economic Aspects. In 1988 ca. 1.3×10^6 t of UP resins were consumed worldwide.

Following a fall in consumption as a result of the oil crises at the end of the 1970s and subsequent stagnation, polyester resins experienced an increase in consumption. This was primarily due to increased use in the automobile industry, especially in compression molding and resin transfer molding.

Consumption according to areas of use is listed in Table 3. The apparent excess capacities in this table are somewhat misleading because they include alkyd resin capacities for the paint sector. The capacity of the UP resin manufacturers is fully utilized and an expansion in capacity can be expected in the next few years.

9. Occupational Health and Environmental Aspects [9]

Toxicological and environmental aspects of unsaturated polyester resins are principally determined by the properties of styrene (→ Styrene). Styrene is a liquid with a relatively high vapor pressure which means that there may be considerable emissions, particularly when polyesters are processed in open molds.

Irritation of the eyes and mucous membranes of the upper respiratory tract is observed and depends on the exposure period and concentration of styrene in the inspired air. Tiredness, loss of concentration, and similar symptoms have been reported at styrene exposure levels above 100 ppm at the workplace.

Styrene and styrene-containing UP resins were classified in 1988 by the EC as "slightly hazardous" with the danger symbol x_n, R-phrase 10 inflammatory, R-phrase 20 health hazard if inhaled, and R-phrase 36/38 eye and skin

Table 3. International distribution of the areas of use for UP resins (10^3 t)

	United States	Western Europe	Japan
Building and construction	241	125	74
Automobile construction	50	78	9
Pipes and vessels	62	61	67
Boats and shipbuilding	155	55	23
Electrical industry	13	28	9
Paints and other nonreinforced products	68	30	30
Miscellaneous uses	31	73	18
Total consumption	620	450	230
of which			
BMC/SMC ∗	64	58	58
Polymer concrete (in building and construction)	62	36	16
Production capacity	800	540	300

∗ BMC=bulk molding compounds; SMC=sheet molding compounds.

irritant. A grading with the R-phrase 45 "Can cause cancer" was specifically excluded.

In 1989 in the United States, styrene was classed by OSHA as not carcinogenic, and was included in the category "Avoidance of narcosis". The TLV-TWA value was reduced from 100 ppm to 50 ppm.

In Germany the MAK value was reduced from 100 ppm to 20 ppm in 1987 on account of reported neurotoxic effects. It was announced that the carcinogenic effects of styrene would be reexamined.

Maximum workplace concentrations for various countries follow (ppm):

Austria	200
Belgium	50
Commonwealth of Independent States	1.2
Former Czechoslovakia	47
Denmark	25
Finland	20
France	50
Germany	20
Italy	50
Japan	50
Netherlands	100
Norway	50
Poland	12
Sweden	20
Switzerland	50
United Kingdom	100
United States	50

In many countries these concentrations will probably be lowered in the near future.

The observance of low maximum workplace concentrations does not present any particular problem in the production of styrene, polystyrene, and styrene copolymers nor in the processing of polystyrene and styrene copolymers. In UP resin processing, however, it is often almost impossible to satisfy this requirement.

Styrene emissions can be considerably reduced by ventilation devices at the workplace, by using "environmentally friendly resins" that contain a skin-forming agent such as paraffins (Chap. 4), or by completely enclosing automatically operating plants. Statutory regulations (e.g., TA-Luft in Germany) impose limits on the discharge of evaporated styrene and waste-gas treatment plants are necessary if fairly high levels of emissions are reached (e.g., >100 mg/m^3 in TA-Luft).

Prolonged exposure of the skin to liquid UP resins often leads to irritation, and particular care must be taken when working with the peroxides used as curing agents on account of their corrosive effect. Goggles should always be worn to protect the eyes when handling peroxides.

UP resins are flammable and hazard classifications differ in individual countries. The content of volatile monomers (styrene), the flash point, and the viscosity of the resins are usually decisive factors. Details are given in the manufacturers' safety data sheets.

Liquid UP resins containing styrene as crosslinking agent and must not pass into the ground water. Relevant statutory regulations governing the storage and transportation of UP resins must be strictly observed. Storage tanks for UP resins should be installed in catchment basins to trap any leakages or escaping product.

10. References

1. P. H. Selden: *Glasfaserverstärkte Kunststoffe,* Springer Verlag, Berlin 1967, pp. 8 ff.
2. G. W. Becker, D. Braun (eds.): "Duroplaste," *Kunststoff-Handbuch,* vol. 10, Hanser Verlag, München 1988, pp. 94 ff.
3. H. V. Boenig: *Unsaturated Polyesters, Structure and Properties.* Elsevier, Amsterdam 1964.
4. P. F. Bruins: *Unsaturated Polyester Technology,* Gordon and Breach Science Publishers, New York 1976.
5. G. Pritchard: *Developments in Reinforced Plastics-1, Resin Matrix Aspects,* Applied Science Publishers, London 1980, pp. 59 ff.
6. Mobil Oil Corp., EP 0 003 406, 1979 (J. G. Murray).
7. BASF, EP 0 073 413, 1982 (A. Henne et al.).
8. Rohm and Haas, US 3 772 241, 1966 (C. H. Kroekel).
9. *AVK-Handbuch,* Arbeitsgemeinschaft Verstärkte Kunststoffe, Frankfurt 1988, chap. 6.

Polyesters

Alkyd Resins, and Polyester Resins, Unsaturated are separate keywords. For polyester fibers, see Fibers, 4. Synthetic Organic. General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Horst Köpnick, Bayer AG, Dormagen, Federal Republic of Germany (Chaps. 1, 2)

Manfred Schmidt, Bayer AG, Dormagen, Federal Republic of Germany (Chaps. 1, 2)

Wilhelm Brügging, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Sections 3.1 – 3.8)

Jörn Rüter, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Sections 3.1 – 3.8)

Walter Kaminsky, Institut für technische und makromolekulare Chemie, Universität Hamburg, Federal Republic of Germany (Section 3.9)

1. Introduction 2
2. Polyesters as Intermediates for Polyurethanes 2
2.1. Raw Materials 2
2.2. Production 3
2.2.1. General Information 3
2.2.2. Industrial Production 4
2.3. Quality Specifications and Testing . 5
2.4. Properties 5
2.5. Storage and Transportation 6
2.6. Uses and Economic Aspects 6
3. Thermoplastic Polyesters 7
3.1. Raw Materials 7
3.2. Production 9
3.2.1. General Principles 9
3.2.2. Production of Poly(Ethylene Terephthalate) from Dimethyl Terephthalate 10
3.2.3. Production of Poly(Ethylene Terephthalate) from Terephthalic Acid . . 12
3.2.4. Production of Poly(Butylene Terephthalate) 12
3.2.5. Production of Other Partially Aromatic Polyesters 13
3.2.6. Production of Fully Aromatic, Liquid Crystalline Polyesters 13
3.3. Analysis 14
3.4. Co-Components and Additives . . . 15
3.5. Properties 17
3.6. Uses 20
3.7. Toxicology 22
3.8. Economic Aspects 22
3.9. Recycling 23
4. References 25

1. Introduction

Classification. Polyesters are defined as all polymers which contain repeating units of the ester group

in the polymer chain. Polyesters are generally produced by condensation of polyfunctional carboxylic acids (or their derivatives) with polyfunctional alcohols. Other methods of production include polycondensation of hydroxycarboxylic acids and polymerization of cyclic esters (e.g., lactones).

Polyesters with widely varying properties can be produced depending on the starting materials:

1) High molecular mass linear polyesters (M_r >10 000) are produced from difunctional alcohols and dicarboxylic acids (or their derivatives, e.g., diesters or dichlorides) or from polymerizable lactones. They can be processed thermoplastically to molding materials (*thermoplastic polyesters*, Chap. 3). The term thermoplastic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_227

polyesters also includes polycarbonates, see → Polycarbonates.

2) Low molecular mass polyesters ($M_r \leq 10\,000$) from saturated aliphatic or aromatic dicarboxylic acids and difunctional or mixtures of di- and trifunctional alcohols are linear or slightly branched *intermediates for polyurethanes* (see Chap. 2).
3) Low molecular mass polyesters ($M_r \leq 10\,000$) from di-, tri-, and polyfunctional alcohols and polyfunctional, usually aromatic carboxylic acids or saturated or unsaturated fatty acids are classified as *alkyd resins*, see → Alkyd Resins.
4) *Unsaturated polyesters* are polyesters which can copolymerize with unsaturated compounds and are formed from polyfunctional alcohols and polyfunctional unsaturated carboxylic acids. After copolymerization with monomers such as styrene they can also be classified as thermosets, see →Polyester Resins, Unsaturated.

2. Polyesters as Intermediates for Polyurethanes

In 1937 O. Bayer developed an addition polymerization method for the synthesis of polyurethanes that was based on the reaction of diisocyanates with polyols [1]. Hydroxy-functional polyesters were soon introduced as reactants. The only highly branched polyesters known at that time were based on aromatic dicarboxylic acids, and they resulted in hard, highly cross-linked products that were processed in coatings and molding compounds [2]. In 1942 H. Pinten treated moderately branched aliphatic hydroxypolyesters with a diisocyanate and obtained a rubbery-elastic material ("I-Gummi") [3]. Systematic study of the reactions of linear and moderately branched hydroxypolyesters with excess diisocyanates, and subsequent treatment of these intermediates with water, diols, diamines, and other compounds containing reactive hydrogens, resulted in a wide range of polymers displaying advantageous properties (e.g., thermosets, thermoplastics, cast elastomers, hard and soft foams, molding compounds, hard and flexible coatings). Polyurethanes came to be known as "tailor-made polymers" [4], [5]. This diversity was partly a result of great flexibility with respect to starting materials, mean molecular mass, and mean functionality.

A broad spectrum of hydroxypolyesters with mean molecular masses between 400 and 6000, hydroxyl numbers of 28 – 300 mg KOH/g, and acid values generally under 1 mg KOH/g are employed as intermediates in the manufacture of polyurethanes. Highly branched polyesters, made largely from aromatic or hydroaromatic dicarboxylic acids, find use as binders for polyurethane coatings (DD coatings, Bayer), → Paints and Coatings[6].

The present review is limited to a discussion of linear and moderately branched hydroxypolyesters derived from aliphatic starting materials.

2.1. Raw Materials

Adipic acid is most commonly employed as the acidic component. Other dicarboxylic acids such as oxalic, succinic, and glutaric acid are of lesser importance because the resulting polyesters exhibit lower stability toward hydrolysis. Higher molecular mass aliphatic carboxylic acids (azelaic acid, sebacic acid) are economically unattractive. "Dimeric acids" [7], produced by the dimerization of unsaturated plant fatty acids and hydroxy-functional monocarboxylic acids, find application in special situations. The hydroxy-functional acid dimers afford the corresponding polyesters [8] by self-condensation onto a starter diol. Aromatic dicarboxylic acids such as phthalic acid (in the form of the anhydride), isophthalic acid, and terephthalic acid are sometimes added as modifiers. Typical diol components are ethylene glycol, diethylene glycol, 1,2-propanediol, 1,3-butanediol, 1,4-butanediol, neopentyl glycol, and 1,6-hexanediol. Glycerol, trimethylolethane, trimethylolpropane, and hexanetriol are employed to obtain branched polymers. Developments in the synthesis of hydroxypolyesters include the use of raw materials such as ε-caprolactone, which affords polycaprolactone [9] via ring-opening polymerization on a starter polyol. Oligo- or polyethers can also be incorporated into the polymer chain, leading to the corresponding etheresters [10].

2.2. Production

2.2.1. General Information

The only currently economic method for the production of polyesters is direct esterification of dicarboxylic acids with diols [11]. Since esterification is a reversible equilibrium reaction, the process can only be driven to completion if the resulting water of condensation is continuously removed. Polyesterification represents an example of *condensation polymerization*. Not only are monomeric units added to the growing chains, but individual chains also react with one other because they may have both carboxyl and hydroxyl terminal groups. The esterification conditions also permit constant transesterification within the chain itself. Consequently, polyesters exhibit relatively broad molar mass distributions, which can be calculated using the Schulz–Flory method [12], [13]. Transesterification also accounts for the fact that, if more than two starting materials are employed, these become incorporated into the polyesters in a statistical way regardless of their time of introduction.

The esterification of a dicarboxylic acid with a diol represents a higher-order reaction. In the absence of an added catalyst, and if the reverse reaction is prevented by removal of the resulting water, the kinetics of the reaction can be described as third order [14], [15] because the acidic groups themselves exert a catalytic effect on the reaction:

$$\text{Reaction rate} = k \cdot [\text{OH}] \cdot [\text{COOH}]^2$$

When an esterification catalyst is utilized, its influence becomes dominant and esterification exhibits second-order kinetics. Esterification is catalyzed by acids, bases, and transition-metal compounds. However, the commercial use of catalysts is limited because all catalysts tend to have undesirable effects on subsequent polyurethane reactions. In the case of polyesters that are used in reactions where traces of catalyst do not interfere, synthesis is catalyzed with low concentrations (ppm) of compounds of titanium, tin, or antimony. The appropriate starting ratios for the synthesis of a given hydroxypolyester can, in theory, be calculated from the desired number-average molecular mass (M_n), which can in turn be determined from the hydroxyl number.

Given the simplified formula

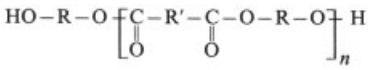

a linear polymer would result from n mol of dicarboxylic acid (HOOC−R′−COOH) and n+1 mol of diol (HO−R−OH). The degree of polycondensation, n, is calculated as follows:

$$n = \frac{M\,(\text{polyester}) - M\,(\text{diol})}{M\,(\text{dicarboxylic acid}) + M\,(\text{diol}) - 2M\,(\text{water})}$$

Losses of diol accompanying the removal of water of condensation and as a consequence of side reactions (formation of ethers and aldehydes) usually require the addition of an empirically determined excess of diol which depends upon the type of diol and the reaction conditions.

The first phase of the esterification is conducted under an inert atmosphere (nitrogen, carbon dioxide) and the water of condensation is removed rapidly by distillation through a column at >130 °C. Once an acid value of about 30 has been reached, the reaction must be accelerated in a second phase by facilitating the removal of water from the esterification equilibrium. This is done by application of a vacuum (vacuum-melt method), introduction of a carrier gas (usually carbon dioxide), or recirculation of an entrainer such as toluene or xylene (azeotropic distillation). The last approach is not applicable when the diol component is ethylene glycol or 1,2-propanediol, which themselves form azeotropic mixtures with the entrainers.

A low acid value (<1 mg KOH/g) is desired. This means that the reaction mixture tends to become highly depleted in one reactant toward the end of the reaction, greatly reducing the rate. This second phase of polyester synthesis therefore requires a large fraction of the overall reaction time (e.g., 20 of 30 h).

2.2.2. Industrial Production

Polyesters are produced industrially either batchwise or continuously according to a variety of methods. The vacuum-melt and the purge-gas-melt methods are preferred, the azeotropic method finds only limited application.

Batchwise Vacuum-Melt Method [16]. The apparatus constitutes a closed system, comprising a stirred-tank reactor equipped with an anchor agitator, a distillation column, a condensor, a receiver, and a vacuum pump. The starting materials are heated in the reactor, with stirring but in the absence of air. Heat is provided by electrical energy, high-pressure steam, or a heat-transfer medium. In the first phase of the reaction the temperature is held between 125 and 220 °C under ambient pressure and ca. 90 % of the resulting water is removed by distillation. A mixture of oligoesters is produced which is immediately condensed to polyester in the second reaction phase at 170 – 230 °C and with stepwise reduction of the pressure to 0.005 bar. Residual water of reaction is distilled from the mixture, together with excess diol and minor byproducts such as low molecular mass aldehydes and ketones (present in ppm amounts relative to the polyester). The process is monitored constantly by measuring the acid value, the hydroxyl number, and the viscosity.

Example of the preparation of a polyester with an average molecular mass of 2000 from ethylene glycol and adipic acid. A mixture of 86.8 kmol of adipic acid and 95.6 kmol of ethylene glycol, including an empirically determined excess of 1.1 kmol ethylene glycol, is heated with stirring under nitrogen to 130 °C. Once distillation begins the temperature of the reaction mixture is raised to 200 °C over the course of 10 h. During this period 156 kmol of water of reaction is distilled into the receiver. When distillation is complete, the acid value of the reaction mixture is ca. 55 mg KOH/g. The pressure in the apparatus is subsequently reduced in stages to 0.01 bar at constant temperature over a period of 15 h. During this period the remainder of the water of reaction is distilled off, along with excess ethylene glycol and byproducts. The reaction mixture has the following properties: hydroxyl number, 55.5; acid value, 0.7; viscosity, 0.515 Pa · s at 75 °C; water content, 0.02 wt %. After pressure equilization with nitrogen and cooling to 110 °C, the polyester product (7.7 kmol) is transferred to a storage tank.

Purge-Gas – Melt Method [17]. This process differs from the vacuum-melt method in that during the second phase of reaction the distillable products are removed by passage of an inert gas (N_2, CO_2) through the mixture rather than by lowering of the pressure. This method increases the loss of diol, a fact that must be taken into account in the raw material calculations. The apparatus must also be suitably adapted: the reactor is equipped with a gas inlet and the distillation column is replaced by a simple distillation tube. A vacuum pump is not required.

Azeotropic Method [18]. This process is conducted at atmospheric pressure in the presence of an inert solvent (e.g., toluene or xylene) which entrains and removes the water of reaction. A vacuum-melt apparatus can be adapted to the azeotropic method by including facilities for removal and continuous recycling of the entrainer solvent. Lower reaction temperatures (125 – 190 °C) are used and permit polyester to be produced under milder conditions. Solvent residues in the final product are removed under vacuum.

Wastewater and Exhaust Air. Both the water of reaction and the process exhaust air contain excess diols and reaction byproducts which are biodegradable in dilute aqueous solution. The processes do not therefore present environmental problems if an exhaust air purifier and a biological water treatment system are employed.

Construction Materials. The materials used for constructing the equipment must satisfy a variety of demands. High resistance to corrosion is essential with respect to the acidic reaction medium and the distillates to prevent contamination of the polyester with traces of metal. Moreover, stress may also arise in the reactor materials due to the high reaction temperatures. These requirements are best met by highly alloyed austenitic chromium – nickel – molybdenum steels, for example V 4 A steel (DIN 1.4571).

The material requirements posed by the raw materials and hydroxypolyester products are less severe. Stainless steel (V 4 A, V 2 A), pure aluminum, $AlMg_3$ alloy, and in special cases even unalloyed steel can therefore be used for pipes and containers for storage and transport.

2.3. Quality Specifications and Testing

Since hydroxypolyesters are subjected to further processing to polyurethanes, they must satisfy stringent quality criteria that can only be obtained through a high degree of product standardization. The most important input data for the polyurethane reaction include the hydroxyl number (DIN 53 240), acid value (DIN 53 402), water content (DIN 53 979), color number (iodine number, DIN 6162; Hazen/APHA, DIN 53 404), and viscosity (DIN 53 015), and are fixed within narrow tolerances for each type of polyester. Individual products often have additional quality criteria that depend on further processing.

The acid value is a measure of the content of terminal carboxyl groups in the polymer and is especially important. A minimum acid value is preferred because terminal carboxyl groups react with isocyanates much more slowly than hydroxyl groups, and can therefore lead to chain termination during polyaddition. Furthermore, the reaction of carboxyl groups with isocyanates produces amide functions and carbon dioxide, which can result in undesirable bubble formation in the polyurethane. Finally, residual acid in a polyester has negative effects on catalysis of the polyurethane reaction and the stability of the polyurethane product toward hydrolysis. Consequently, an acid value limit is mandatory for polyesters (generally <1 mg KOH/g, but often <0.5 mg KOH/g).

The presence of even trace amounts (ppm levels) of metals, acids or bases, fats, silicone oils, and other surface-active agents in polyesters must be avoided. Such impurities can originate from the starting materials or the processing equipment; their removal is extremely difficult, if not impossible. Consequently, both starting materials and apparatus must be subjected to rigid standards. The analytical data cited in Section 3.1 for the thermoplastic polyester raw materials must be regarded as minimum standards.

Manufacturers utilize the test methods already described, and other specially developed tests for product and process quality control. An applications test is often conducted for final quality assessment.

2.4. Properties

Physical Properties. The diverse applications of polyester-polyurethanes require a wide variety of starting materials. Pure polyadipates of ethylene glycol, 1,4-butanediol, and 1,6-hexanediol are waxy crystalline substances with softening points around 40 – 50 °C. Their crystallinity helps to provide the polyurethane products with outstanding physical characteristics, especially with regard to elasticity, tensile strength, and notch resistance. In some application areas, however, a tendency to harden at low temperatures or as a result of stress (crystallization of elongation) is problematic.

Polyesters made from mixtures of the aforementioned diols, alone or in combination with other diols, can be advantageous in such cases, and are either liquid or greasy at room temperature. 1,2-Propanediol, 1,3-butanediol, diethylene glycol, and neopentyl glycol afford liquid polyadipates.

Due to their broad molecular mass distributions (Schulz–Flory distribution) [12], [13], the polyesters have a relatively high viscosity (e.g., ca. 600 mPa · s at 75 °C and a mean molecular mass of 2000). Densities range between 1.1 and 1.5 g/cm^3. The flash points (DIN 51 758) lie above 200 °C, and the ignition points (DIN 51 794) above 400 °C. Polyesters are relatively soluble in common organic solvents, including acetone, ethyl acetate, and toluene.

Chemical Properties. Terminal hydroxyl groups are largely responsible for the chemical properties of the polyesters, which undergo reactions typical of alcohols. The most important reaction from an applications standpoint is the facile, quantitative addition of hydroxyl groups to the isocyanate groups of di- and polyisocyanates, which affords segmented, high molecular mass polyester-polyurethanes (→ Polyurethanes).

The polyesters exhibit a certain degree of sensitivity toward hydrolysis, and newly formed acidic groups have an autocatalytic effect. The rate of hydrolysis is also increased by the same compounds that catalyze esterification. Resistance to hydrolysis is a function of polyester structure and increases with increasing hydrophobic character and with decreasing ester-group content [16]. Polyesters derived from

long-chain diols and dicarboxylic acids are more stable than those based on short-chain starting materials. Diethylene glycol polyesters exhibit only marginal resistance to hydrolysis because of the hydrophilic effect of their ether groups. During further processing acid acceptors (e.g., carbodiimide, epoxides) are often added as hydrolysis stabilizers.

2.5. Storage and Transportation

During storage and transportation of polyesters, exposure to moisture and temperatures above 120 °C must be prevented to avoid undesirable effects during polyurethane production. The products are either stored in sealed, air-tight drums at ambient temperature or in heated, insulated containers under nitrogen at 70 – 110 °C. As liquids they are transported in barrels or in heatable, insulated bulk containers; as solids they are transported in drums.

The construction materials mentioned in Section 2.2.2 are generally utilized for storage and transport vessels; barrels are usually constructed from tin-free steel. Turbidity is sometimes observed in polyesters that are liquid at ambient temperature but have been stored for long periods of time at low temperature; this can be alleviated without affecting the quality of the product by heating the polyester briefly to 80 °C.

Recommendations regarding storage and transport conditions for polyesters are provided by manufacturers. Handling of polyesters presents no significant hazards – a consequence of their chemical characteristics. Spilled material is best treated by mixing it with sand and then sending it to an approved waste disposal site, or by burning it in an appropriate incinerator.

2.6. Uses and Economic Aspects

Hydroxypolyesters are sometimes reacted either alone or in combination with diols or diamines to give segmented polyurethanes, which are marketed as aqueous dispersions, solutions, or as powders that are thermoplastically processable or soluble in special solvents (→ Fibers, 4. Synthetic Organic, Chap. 3.; → Polyurethanes). In other cases they are combined on site with polyisocyanates and various additives to form soft to hard foams, elastomers, adhesives, flexible coatings, and joint fillers.

The polyesters described above are marketed by a number of firms; individual product types are further characterized by appended letters and/or numerical indices. Examples of trade names include:

Baycoll, Bayflex, Baygal, Baytec, Baytherm	Bayer
Crestapol	Scott Bader
Daltogel, Daltolac, Daltorez	ICI
Desmophen	Bayer
Diorez	Briggs and Townsend
Estolan	Lancro Chemicals
Fomrez	Witco Chem. Corp.
Lumitol/Elastophen	BASF
Multron/Multrathane	Miles Inc. (formerly Mobay Chem. Corp.)
Nippolan	Nippoly
Polyesteralkohol/Systol	BASF Schwarzheide GmbH
Vulkollan	Bayer

Estimated production capacities in 1983 were (in 10^3 t/a):

Federal Republic of Germany	89
German Democratic Republic and Poland	15
EC	188
United States	44
Japan	21
World	277

Worldwide consumption in 1983 was estimated at ca. 200 000 t.

3. Thermoplastic Polyesters

[19][20][21][22][23][24][25]

The history of thermoplastic polyesters (see also → Fibers, 4. Synthetic Organic, Chap. 2.) goes back to 1929 with the pioneering work of Carothers. The first thermoplastically processible polyesters synthesized from adipic acid and ethylene glycol were described by him in 1932. Polyesters only became of industrial interest in 1941, with the synthesis of high melting point products based on terephthalic acid.

The rapid industrial development of polyesters after World War II was initially restricted to polyester fibers based on poly(ethylene terephthalate) [*25038-59-9*] (PETP), poly(oxy-1,2-ethanediyloxycarbonyl-1,4-phenylenecarbonyl). A polyester based on poly(1,4-dimethylenecyclohexane terephthalate) [*25135-20-0*],

PDCT, occupies a special position (see also → Fibers, 4. Synthetic Organic, Chap. 2.4.2.).

Poly(ethylene terephthalate) was subsequently also used in the production of films (see also → Films, Chap. 6.8.).

Thermoplastic polyesters were first employed as construction materials in 1966 [26], and were initially based on poly(ethylene terephthalate). In 1970 the more readily processible poly(butylene terephthalate) [*24968-12-5*] (PBT), poly-(tetramethylene terephthalate), poly(oxy-1,4-butanediyloxycarbonyl-1,4-phenylenecarbonyl) was introduced onto the market.

A short time later these homopolyesters were supplemented by a series of thermoplastic copolyesters which are suitable for specific areas of application due to their special properties [27]. These copolyesters include polyetherester block copolymers as thermoplastic elastomers; copolyesters based on 1,4-cyclohexanedimethanol for hard, glass-clear injection molded articles; and other thermoplastic copolyesters of varying composition for powder coatings, paint binders (→ Paints and Coatings, Chap. 2.7.), and hotmelt adhesives.

A new class of fully aromatic thermoplastic polyesters has been developed – liquid crystalline polyesters [28], [29]. These polymers have aroused great interest on account of their outstanding mechanical and thermal properties. Two such products are commercially available, one being synthesized from 4-hydroxybenzoic acid and 6-hydroxy-2-naphthoic acid, and the other from 4-hydroxybenzoic acid, terephthalic acid, and 4,4′-dihydroxybiphenyl.

3.1. Raw Materials

Thermoplastic polyesters are generally produced from dicarboxylic acids and dihydric alcohols. The most common raw materials follow:

Dicarboxylic acids (and derivatives)
Dimethyl terephthalate [*120-61-6*]
Terephthalic acid [*100-21-0*]
Dimethyl isophthalate [*1459-93-4*]
Isophthalic acid [*121-91-5*]
Dimethyl adipate [*627-93-0*]
Adipic acid [*124-04-9*]
Azelaic acid [*123-99-9*]
Sebacic acid [*111-20-6*]
Dodecanoic diacid [*693-23-2*]
1,4-Cyclohexane dicarboxylic acid [*1076-97-7*]
Dimethyl-1,4-cyclohexane dicarboxylate ester [*94-60-0*]

Hydroxycarboxylic acids and lactones
4-Hydroxybenzoic acid [*99-96-7*]
6-Hydroxy-2-naphthoic acid [*16712-64-4*]
Pivalolactone [*1955-45-9*]
ε-Caprolactone [*502-44-3*]

Dihydric alcohols and bisphenols
Ethylene glycol [*107-21-1*]
1,2-Propanediol [*57-55-6*]
1,3-Propanediol [*504-63-2*]
1,4-Butanediol [*110-63-4*]
1,4-Cyclohexanedimethanol [*105-08-8*]
2,2-Dimethyl-1,3-propanediol [*126-30-7*]
1,6-Hexanediol [*629-11-8*]
Poly(tetrahydrofurandiol) [*9040-43-1*]
4,4′-Dihydroxy-1,1′-biphenyl [*92-88-6*]
1,4-Hydroquinone [*123-31-9*]

Terephthalic acid is by far the most important dicarboxylic acid. This compound was previously difficult to obtain in sufficiently pure form, dimethyl esters of dicarboxylic acids were therefore mainly used as raw materials. Lactones, hydroxycarboxylic acids, and bisphenols are of lesser importance.

Particularly stringent requirements are placed on the purity of the raw materials because impurities can either interfere with polycondensation via chain termination or branching (e.g., mono- or polyfunctional compounds) or can lead to undesirable secondary reactions and discoloration under the high reaction temperatures. These requirements apply to polyesters used as starting material for fibers, but particularly to the plastics sector where polyesters with substantially higher molecular masses are needed.

Analytical requirements for the four most important raw materials are as follows:

Dimethyl Terephthalate (DMT). For production and properties see → Terephthalic Acid and Dimethylterephthalates.

Purity (GC analysis)	>99.9 %
Solidification point (min.)	140.63 °C
Saponification value	577.8 mg KOH/g
Acid value (max.)	0.03 mg KOH/g
Acid value after 4 h at 175 °C (max.)	0.06 mg KOH/g
Color of the melt after 24 h at 170 °C (max.)	10 (Hazen number)
Color in H_2SO_4 (max.)	10 APHA
Fe content (X-ray fluorescence) (max.)	1 ppm
Nitro groups and nitroso groups should also be absent	

Terephthalic Acid (TA). For production and properties, see → Terephthalic Acid and Dimethylterephthalates. The purity of "polymer grade" terephthalic acid must satisfy appropriate requirements for the production of fibers, films, and molding compounds. As a rule terephthalic acid is available with the following specifications:

Acid value	675 ± 2 mg KOH/g
Ash content (max.)	15 ppm
Metals (atomic absorption) (max.)	10 ppm
of which Fe (max.)	2 ppm
Co, Mo, Ni, Ti, Mg (max.)	1 ppm
Ca, Al, Na, K (max.)	2 ppm
4-Carboxybenzaldehyde (polarography) (max.)	25 ppm
Water content (Karl Fischer) (max.)	0.5 %
Color in 5 % dimethylformamide (max.)	10 APHA

Under certain conditions (absence of fluorenone structures, maintenance of a defined color index) the 4-carboxybenzaldehyde content can reach 200 ppm without impairing the polyester quality [30].

Ethylene Glycol. For production and properties, see → Ethylene Glycol.

Relative density d_{20}^{20}	1.1151 – 1.1156
Boiling point range	196 – 200 °C
Solidification point	−11 to −13 °C
Acid content as acetic acid (max.)	0.005 %
Fe content (max.)	0.07 ppm
Chlorine compounds	none
Diethylene glycol content (max.)	0.08 %
Water content (max.)	0.08 %
Ash content (max.)	0.005 g/100 mL
Acetaldehyde content (max.)	30 ppm
Color index after 24 h at 170 °C	<10 (Hazen number)
UV transparency at m	
220 μ	70 %
275 µm	90 %
350 µm	95 %

1,4-Butanediol. For production and properties, see → Butanediols, Butenediol, and Butynediol, Chap. 1.3.

Purity (GC analysis) (min.)	99.3 %
Water content (max.)	0.05 %
Solidification point (min.)	19.5 °C
Color index (max.)	10 APHA
Ash content	0

3.2. Production

3.2.1. General Principles

Two-Stage Process. Generally, thermoplastic polyesters are produced in two stages. In the first stage a polyester precondensate (M_n 100 – 2000, depending on the molar ratio of the starting compounds) with the formula

$$HO-Y-O\,(OC-X-COO-Y-O)_x\,H,\ x \approx 1-10$$

is produced by transesterification of dicarboxylic diesters or by esterification of the dicarboxylic acid HOOC−X−COOH with excess dihydric alcohol HO−Y−OH. The precondensate then reacts in a second step (i.e., polycondensation) with the elimination of dihydric alcohol to form high molecular mass polyesters (M_n > 10 000) of the formula:

$$HO-Y-O\,(OC-X-COO-Y-O)_y\,H,\ y \geq 100$$

Transesterification and esterification are equilibrium reactions that are accelerated by catalysts. The reactions are continuously displaced towards the polyester precondensate by removing the more volatile alcohol or water.

Polycondensation is also an equilibrium reaction, which is shifted towards the high molecular mass polyester by constant removal of the eliminated diol or water (application of a vacuum or purging with an inert gas), and is accelerated by catalysts. The higher the desired molecular mass of the polyester, the more severe are the conditions required to remove the diol and water. Polycondensation is a step-growth reaction (of the type A – A+B – B=diol + dicarboxylic acid or A – B=hydroxycarboxylic acid), which leads to nonuniform end products. The molecular mass distribution can be calculated from the

molar ratio of diol to dicarboxylic acid and/or from the conversion p [31]. With a stoichiometric ratio of the A and B groups, the mass fraction W_x of the degree of polycondensation x can be specified as a function of the conversion p [32]:

$$W_x = \frac{(M_L + M_U x)}{[M_U + M_L (1-p)]} \, p^{x-1} (1-p)^2$$

M_L and M_U denote the molecular masses of the leaving molecules L and basic units U, respectively. For negligibly small molecular masses of the leaving molecules ($M_L/M_U \rightarrow 0$) or as an approximation for high conversion rates ($p \rightarrow 1$), the equation transforms to the Schulz–Flory distribution function

$$W_x = x p^{x-1} (1-p)^2$$

Introduction of the mean degree of polycondensation $\bar{P}_n = 1/(1-p)$ gives

$$W_x = \frac{x}{\bar{P}_n^2} \left(1 - \frac{1}{\bar{P}_n}\right)^{x-1}$$

The relationship $\bar{P}_n = 1/(1-p)$ clearly shows that a minimum degree of polymerization of 100 can only be achieved with conversions of $> 99\,\%$ (p=0.99).

The kinetics of esterification, transesterification, and polycondensation [33–40] have been investigated in great detail. See also → Polymerization Processes, Chap. 1., → Polymerization Processes, Chap. 2. In general, polycondensation is regarded as a second-order reaction [41], [42], with an activation energy that depends strongly on the catalyst system (PETP: uncatalyzed 96.6 kJ/mol [41], catalysis with manganese acetate 40.8 kJ/mol [42]).

Other Processes. In addition to the conventional two-stage process for producing linear thermoplastic polyesters, many other processes are used with other starting compounds. Addition of ethylene oxide to dicarboxylic acids gives bis(2-hydroxyethyl) esters. In analogy with the previously described esterification and transesterification products, these esters undergo polycondensation with the elimination of ethylene glycol [43].

Fully aromatic polyesters with a high melting point are produced by interfacial polycondensation of dicarboxylic dichlorides with bisphenols in the presence of bases and with elimination of hydrogen chloride [44]. Dicarboxylic acids also react with diesters (preferably acetates) of dihydric alcohols with the elimination of acid [45].

The esterification of an anhydride with a dihydric alcohol or the transesterification of anhydrides with epoxides yields linear polyesters [46], [47].

A special process for producing high molecular mass thermoplastic polyesters involves the polymerization of lactones in the presence of organometallic compounds [48].

Secondary Reactions [49–51]. In all methods for producing linear thermoplastic polyesters, secondary reactions that occur during polycondensation can alter the stoichiometric ratio and thus terminate polycondensation or confer undesirable properties on the end product.

The most important secondary reactions include:

1) Ether formation to produce diethylene glycol units (Y=C_2H_4)

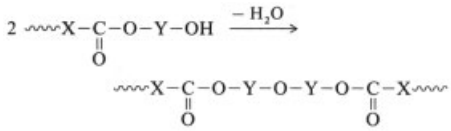

2) Dehydration of ethylene glycol to form acetaldehyde, or of 1,4-butanediol to form tetrahydrofuran

3) Ester pyrolysis which in the case of PETP, for example, produces carboxyl groups (poorer resistance to hydrolysis) and olefins (discoloration):

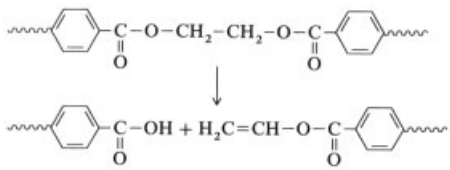

4) Ring formation between two adjacent carboxyl groups (particularly pronounced in adipic acid):

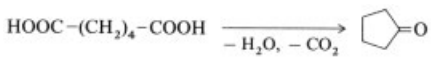

3.2.2. Production of Poly(Ethylene Terephthalate) from Dimethyl Terephthalate

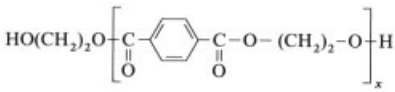

Poly(ethylene terephthalate) (PETP) is produced industrially both batchwise and continuously. The apparatus used for these processes is also suitable for the production of other thermoplastic polyesters. For the sake of simplicity, however, the batchwise production of PETP is discussed here. Separate vessels are often used to melt the starting compounds and for the transesterification and polycondensation reactions.

Melting. Dimethyl terephthalate (DMT) is melted at 150 – 160 °C in a stirred tank heated with steam, carrier oil, or electricity in an inert atmosphere (nitrogen).

Transesterification. The molten DMT and ethylene glycol are reacted in heated, stirred transesterification reactors at 150 – 200 °C. At the start of the reaction (at normal pressure in a nitrogen atmosphere) the lower temperature range is preferred to minimize sublimation of DMT. The methanol released during the transesterification is continuously distilled off from the reaction mixture. The excess of ethylene glycol is normally ca. 0.5 – 1 mol/mol DMT.

The use of catalysts is essential to achieve a reasonable transesterification rate at moderate temperatures. Although the catalysts listed in the patent literature cover practically the whole periodic system, weakly basic compounds such as amines, metal oxides, alkoxides, and acetates are used as transesterification catalysts in practice at a concentration of 0.01 – 0.1 wt % referred to DMT. Systematic investigations showed that the rate constant for the transesterification of DMT with ethylene glycol correlates with the electronegativity of the catalyst metal ion [52]. Metal compounds in which the electronegativity of the metal ion is in the range 1 – 1.7 are most effective (e.g., acetates of Ca, Mg, Zn, Cd, Pb, and Co). An almost identical order is found when the stability constants of the metal dibenzoylmethane complexes are related to the rate constants of the transesterification reaction [53].

The choice of a suitable catalyst is an important factor in the production of PETP as regards the economy of the production process and quality of the end product. Transesterification should proceed as quantitatively as possible because residual methyl ester groups can interfere with subsequent polycondensation [54]. This is particularly true if the activity of the transesterification catalysts is lowered by additives introduced at the polycondensation stage. The course of transesterification is therefore controlled by the amount of methanol released, whereby entrained ethylene glycol has to be taken into account (density measurement or ultrasonic monitoring).

Polycondensation. The transesterification product is added as a melt to the polycondensation reactor, which can be heated to over 300 °C and must be equipped with a very efficient stirrer. Some of the excess ethylene glycol is often first distilled off at normal pressure (protective gas, nitrogen) by gradually increasing the temperature to ca. 250 °C. Polycondensation then follows, with a reduction in pressure and a further increase in temperature. In order to obtain sufficiently high molecular masses, temperatures of ca. 270 – 280 °C at a final vacuum of <1 mbar are required.

Like the transesterification reaction, the polycondensation reaction has to be accelerated by catalysts. All the catalysts used for transesterification are, in principle, also suitable for polycondensation because the two reactions are extremely similar. The catalyst remaining in the reaction mixture after the transesterification can therefore be used further for polycondensation. This approach does not always provide satisfactory products, however, because active transesterification catalysts decrease the thermal stability of PETP [55]. In many cases the transesterification catalyst is therefore masked with suitable additives (mainly phosphorus compounds) and replaced by special polycondensation catalysts, usually antimony, germanium, titanium, or lead compounds at a concentration of 0.005 – 0.05 wt % referred to DMT.

Rapid polycondensation is achieved by quickly removing the eleminated ethylene glycol by means of intensive mixing and a maximum polymer melt surface area. In conventional

polycondensation plants this is achieved by intensive stirring. Distribution of the polymer melt in thin-film procedures as in thin-layer evaporators or in annular disk reactors is, however, more effective for obtaining very high molecular masses.

Polycondensation is stopped when a defined melt viscosity (measure of the molecular mass), is reached. The power consumption of the stirrer motor is often an adequate indicator of melt viscosity. In order to terminate polycondensation, the vacuum in the reaction vessel is removed with oxygen-free nitrogen and the melt is expelled from the reactor either under nitrogen pressure or, preferably, with spinning pumps. The melt is directly quenched with water and comminuted into chips or pellets to prevent oxidation by atmospheric oxygen.

Conventional batchwise and continuous polycondensation produces PETP that is processed into threads and fibers (solution viscosity ca. 50 – 80 cm^3/g measured in phenol/tetrachloroethane 60/40 wt % at a concentration of 0.0023 g/cm^3, see Section 3.3).

Before further processing the PETP must be carefully dried at 80 – 130 °C to reduce the water content to below 0.01 wt %. Drying is performed with the same apparatus used for solid-phase postcondensation. Heating during drying causes crystallization of PETP which is necessary to avoid agglomeration of the particles during prolonged storage.

A PETP with a fairly high intrinsic viscosity, such as can be produced with the thin-layer evaporator or annular disk reactor, is required for the plastics sector. Alternatively, however, the comminuted PETP (chips, pellets, etc.) is often subjected to solid-phase postcondensation for a fairly long time (~20 h) at temperatures of up to 250 °C in an inert gas stream or in vacuo, resulting in an increase in molecular mass. The rate-determining step in this reaction is diffusion of the eliminated ethylene glycol from the polymer particles, which depends on the particle size [56]. Postcondensation in the solid phase can be performed either batchwise (in vacuum dryers or in a fluidized bed) or continuously (in countercurrent with hot nitrogen).

Production Example for PETP [57]. 12.8 kg of DMT and 8.1 kg of ethylene glycol are transesterified in the presence of 3.32 g of manganese acetate and 0.83 g of cobalt acetate for 3 h at 195 °C under nitrogen in a 30-L steel reactor equipped with an anchor-type stirrer and a distillation column. The liberated methanol is distilled off. 10.76 g of triphenyl phosphate is then added and stirred in for 30 min at 195 °C; 2.75 g of germanium oxide and 0.5 g of tetraisopropyl titanate are added and stirring is continued for a further 30 min.

The temperature of the reaction mixture is then raised to 280 °C within 6 h and the pressure subsequently reduced to 0.3 mbar in 2 h. After a further 2 h under these conditions the vacuum is removed with nitrogen. The polyester melt is extruded in the form of strands, which are cooled in water and then pelletized. The solution viscosity of the polyester is 62 cm^3/g (in phenol/tetrachloroethane 60/40, concentration 0.0023 g/cm^3).

3.2.3. Production of Poly(Ethylene Terephthalate) from Terephthalic Acid

The direct esterification of terephthalic acid (TA) with ethylene glycol became important when economic processes were developed for producing fiber-grade TA [58]. An important advance was made when the reaction times were reduced by performing esterification under pressure (2.7 – 5.5 bar) at temperatures above the boiling point of ethylene glycol (198 °C) [59], [60].

The TA process now yields products that are qualitatively competitive with the polyesters produced in the DMT process (see Table 1) [61].

Table 1. Quality specifications for additive-free PETP used for film production [61]

Specification	Raw material	
	DMT	TA
Intrinsic viscosity, cm^3/g	65 ± 0.015	65 ± 0.015
Carboxyl terminal groups, mmol/kg	20 – 30	20 – 30
Ash content, wt %	<0.04	<0.02
Diethylene glycol content, wt %	0.5 – 1.2	0.7 – 1.2
Filter value (15 µm filter braid), bar cm^2/kg	<30	<30

In new plants direct esterification is the preferred method of PETP production. The advantages of direct esterification with TA compared with transesterification with DMT are

1) The higher reaction rate
2) The lower weight of TA compared with DMT (storage costs)
3) The use of water instead of methanol as condensation agent
4) No transesterification catalyst is required
5) Higher molecular masses are obtained

Temperatures of around 220 – 260 °C are generally maintained to overcome the sparing solubility of TA. The water formed is continuously removed by distillation. The molar ratio of ethylene glycol to TA should be < 2 (in most patents < 1.5 is specified) to suppress formation of diethylene glycol (see page 10). Direct esterification is possible without catalysts, but can be accelerated by basic catalysts (amines) or transesterification catalysts (usually Sb_2O_3).

Esterification is followed by gradual pressure release, the temperature is simultaneously increased and excess ethylene glycol is distilled off. The formation of PETP takes place in the polycondensation stage, which proceeds analogously to the DMT – ethylene glycol process as regards temperature, pressure, and catalysts (see Section 3.2.2, a detailed process description is given in [62], [63]).

Direct esterification at normal pressure can be achieved in a PETP precondensate as reaction medium, in which TA is readily soluble [64]. This process is particularly suitable for continuous direct esterification.

The secondary reactions occurring during polycondensation are of decisive importance in determining the quality of the PETP (see page 10). Ether formation results in diethylene glycol units which have adverse effects on dyeing behavior but also lower thermal and UV stability (for the mechanism of diethylene glycol formation, see [65]). Free carboxylic groups formed by ester pyrolysis have a negative influence on the postcondensation and hydrolysis stability of PETP. Ester pyrolysis also produces olefinic decomposition products that lead to discoloration of PETP.

3.2.4. Production of Poly(Butylene Terephthalate)

Poly(butylene terephthalate) (PBT) is produced analogously to PETP. The currently preferred process is still the transesterification of DMT with 1,4-butanediol followed by polycondensation. The main byproduct is tetrahydrofuran, formed from 1,4-butanediol by dehydration (see page 10). Although some alkyl tin compounds lower tetrahydrofuran formation, the resulting polymers have a lower thermal stability [66]. Although tetrahydrofuran formation is catalyzed by acids, TA is increasingly used as a raw material instead of DMT [67], [68]. For the advantages of the TA method, see Section 3.2.3.

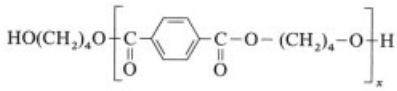

Transesterification of DMT is performed with a slight excess of 1,4-butanediol (molar ratio 1,4-butanediol: DMT < 2, preferably 1.3 – 1.7) at 160 – 220 °C and normal or excess pressure. Soluble titanium compounds are generally used as transesterification catalysts.

Polycondensation at ≤ 260 °C and < 1 mbar is usually also catalyzed by soluble titanium compounds (generally tetraalkyl titanates).

PBT can generally be subjected to polycondensation in the melt to give products with higher molecular masses than PETP (solution viscosity up to ca. 120 cm^3/g, measured in phenol – tetrachloroethane, weight ratio 3 : 2 at 25 °C, concentration 0.0023 g/cm^3). The product is dried in the same way as PETP, see Section 3.2.2. For some uses in the plastics sector (optical waveguide coverings, coatings for bowden wire, headlight dimmers) a higher molecular mass PBT is required, which can also be obtained by solid-phase condensation.

Production Example for PBT [69]. 8.0 kg of DMT and 5.9 kg of 1,4-butanediol (molar ratio with respect to DMT 1.6) are added to a 30-L reactor equipped with an anchor-type stirrer. After heating the mixture to 150 °C, 160 g of a 7 % solution of titanium tetrabutylate in 1,4-butanediol is added over 100 min while stirring to initiate transesterification. After gradual heating to 200 °C over 1.5 h, transesterification is continued for a further 3.5 h at 200 °C, and the methanol formed is distilled off.

The pressure is reduced from 1 bar to 1.3 mbar over 2 h, and the temperature is subsequently increased from 200 to 245 °C in 1 h.

Continued heating at 245 °C and 1.3 mbar for ca. 5 h gives a polymer with an intrinsic viscosity of 124 cm^3/g.

3.2.5. Production of Other Partially Aromatic Polyesters

Other high molecular mass thermoplastic polyesters, such as poly(1,4-cyclohexylenedimethylene terephthalate) [*25135-20-0*], PDCT, copolyesters or copolyetheresters (see → Fibers, 4. Synthetic Organic, Chap. 2.4.2.) are produced in a similar way to PETP and PBT. The different melting points of the end products and different reactivities of the monomers must, however, be taken into account. Depending on their chemical constitution (diol, diester, and diacid), the co-components (see Section 3.4) are added before, during, or after the transesterification phase, or before or during the esterification phase.

Production examples for polyesters based on 1,4-cyclohexanedimethanol [1,4-bis(hydroxymethyl)cyclohexane] and for polyetheresters are given in [70–72].

3.2.6. Production of Fully Aromatic, Liquid Crystalline Polyesters

Fully aromatic, liquid crystalline polyesters can be produced in the melt by three main processes. Direct esterification of phenols with aromatic carboxylic acids in the presence of catalysts such as titanium tetrabutyrate or dibutyl tin diacetate is described in [73]. On account of the high reaction temperatures (>300 °C) there is however the danger of decarboxylation of the carboxylic acids, and the method is therefore restricted to monomers that are stable at the melting point.

A second process involves reacting phenyl esters of aromatic carboxylic acids with relevant phenols. The phenyl esters can be produced by direct esterification or by reaction of diphenyl carbonate with dicarboxylic acids, with the formation of carbon dioxide and phenol.

The commercially available liquid crystalline polyesters Vectra [*81843-52-9*] and Xydar [*31072-56-7*] are produced by a third method: acidolysis of phenolic acetates with aromatic carboxylic acids [74], [75] (see top of the next page).

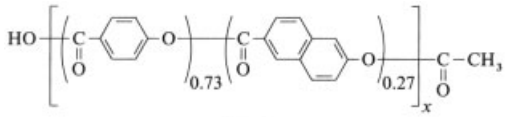

Vectra

Vectra is synthesized from 4-hydroxybenzoic acid and 6-hydroxy-2-naphthoic acid. Xydar is prepared from 4-hydroxybenzoic acid, 4,4′-dihydroxy-1,1′-biphenyl, and terephthalic acid. The basic structural formulae shown below can be modified by using other monomers such as isophthalic acid or 4-aminobenzoic acid, or by using a different monomer ratio [75–78]. In order to acetylate the phenolic hydroxyl groups, the monomers are added to a stainless steel stirred tank, followed by an excess of acetic anhydride under an inert gas atmosphere (nitrogen). Too large an excess of acetic anhydride affects the color of the polymer and leads to the formation of mixed anhydride structures. Most patents specify a molar ratio of acetic anhydride to phenolic OH groups of 1.02 – 1.18.

The acetylation reaction is performed at 120 – 150 °C under reflux and is complete after 1 to 5 h. The acetic acid that is formed and any acetic anhydride still present are distilled off. The temperature is then raised in stages to ca. 320 °C and a low molecular mass prepolymer is formed. Polycondensation is then usually carried out by gradual reduction of the pressure (<15 mbar) and distilling off further acetic acid until the desired molecular mass (10 000 – 30 000) is obtained. In contrast to Vectra, the maximum molecular mass of Xydar (*mp* 421 °C) in the melt is limited by the higher melting point and higher melt viscosity. Consequently, the Xydar prepolymer, which is still liquid at 320 °C, is discharged from the reactor under an inert gas, granulated, and postcondensed in the solid phase (inert gas or vacuum) at a maximum temperature of 365 °C.

Vectra is synthesized exclusively from high-purity hyroxycarboxylic acids and can, in theory, undergo polycondensation to yield products of infinitely high molecular mass. The molecular mass can be limited by adding small amounts of 4,4′-diacetoxybiphenyl [79]. Catalysts may, however, be added to accelerate acetylation and polycondensation (ca. 25 – 500 ppm referred to the polymer). Catalysts include alkali-metal acetates (e.g., sodium or potassium acetate) and

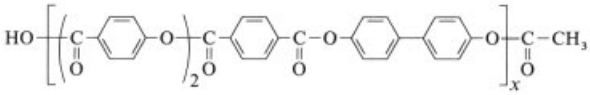

Xydar

acids (e.g., *p*-toluenesulfonic acid or 2-naphthalenesulfonic acid). Iron ions have a high catalytic activity in the production of Vectra [74]. The iron content of the polymer should not, however, exceed 50 ppm because this has an adverse effect on its thermal properties (see above).

Xydar is often synthesized in the presence of potassium sulfate (25 – 50 ppm referred to the polymer) to improve processibility and appearance of the polymer [75].

Production Example for Vectra [74]. 24.05 kg of 6-hydroxy-2-naphthoic acid and 47.85 kg of 4-hydroxybenzoic acid are added to a 151.4-L steel reactor (stainless steel 316) equipped with a helical stirrer, a solids and liquid trap, and a distillation column. The reactor is alternately evacuated (ca. 9 mbar) and flushed with nitrogen three times. 49.90 kg of acetic anhydride (technical quality) is then added. The reaction mixture is heated to 120 – 140 °C and kept at this temperature for 1 h. The temperature is then raised to 200 °C in 1.5 h. Polycondensation takes place at 200 –327 °C over a period of 3 h under nitrogen (2.8 m^3/h), followed by 1.25 h (at 327 °C) in vacuo (ca. 13 mbar). The hot polymer melt is then extruded and comminuted. The intrinsic viscosity of the polymer in pentafluorophenol at 60 °C (concentration 0.1 wt %) is 540 cm^3/g.

3.3. Analysis

Thermoplastic polyesters have to be analyzed before they are processed. The most important parameter to be determined is the molecular mass because this influences the processability of the molding compounds and the mechanical properties of the article produced from the polyester.

In most cases the molecular mass is determined indirectly via the reduced or specific viscosity (in cm^3/g) in a capillary viscometer (→ Plastics, Analysis). The intrinsic viscosity $[\eta]$ is thus obtained by extrapolation of the reduced viscosity to zero concentration and from this the molecular mass M [80] according to

$$[\eta] = K M^{a}$$

For polymers with a range of molecular masses, M becomes M_n (number average), M_w (weight average), or M_v (viscosity average) depending on the method of determination. See also, → Plastics, General Survey, Chap. 2.2.2.

The reduced viscosity depends on the solvent, concentration, and temperature. For quality control it is normally sufficient to specify the solution viscosity under defined conditions.

The constants K and a have been determined for PETP and PBT under various conditions (see Table 2). For PBT a correlation between the reduced viscosity and M_w is also described [86].

The melt viscosity η_m is just as important a quantity as the solution viscosity as regards the processing of thermoplastic polyesters. Measurement under defined conditions (temperature, viscometer, shear rate, etc.) is sufficient for monitoring quality. Like the solution viscosity, the melt viscosity is related to the molecular mass [87]:

$$\eta_{m,0} = K \bar{M}_w^{3.5}$$

where $\eta_{m,0}$ is the melt viscosity in Pa · s at a shear rate of 0. Values for K in Pa · s are

PETP [88]: $1.13 \times 10^{-14} \exp(-11.9755 + 6082.1/T)$

PBT [85]: 1.02×10^{-14} at 250 °C

Investigations on the melt viscosity of polyetheresters and fully aromatic, liquid crystalline polyesters are described in [89].

The molecular mass distribution of thermoplastic polyesters is determined only in special cases (for methods see [90]), for example to establish the influence of process or reactor technology on the polyesters. Most linear polyesters exhibit a Schulz – Flory distribution (see page 9). Further information on the quality of thermoplastic polyesters is obtained by measuring

Table 2. Constants K and α for PETP and PBT

Polyester	Solvent	K	α	Method of determination	Reference
PETP	phenol – tetrachloroethane 1 : 1, 20 °C	0.0755	0.685	terminal group determination (M_n)	[81]
PETP	2-chlorophenol, 25 °C	0.030	0.74	terminal group determination (M_n)	[82]
PETP	2-chlorophenol, 25 °C	0.0425	0.69	sedimentation, diffusion (M_w)	[82]
PETP	phenol-1,2-dichlorobenzene 1 : 1, 25 °C	0.0469	0.68	light scattering (M_w)	[83]
PETP	phenol – tetrachloroethane 3 : 2, 25 °C	0.0468	0.68	light scattering (M_w)	[84]
PBT	phenol – tetrachloroethane 3 : 2, 30 °C	0.0116	0.87	terminal group determination (M_n)	[85]
PBT	phenol-1,2-dichlorobenzene 1 : 1, 25 °C	0.0437	0.72	light scattering (M_w)	[83]

the melting point, glass transition temperature (both preferably by differential thermal analysis), the hydroxyl terminal groups [91], acid terminal groups [92], and the diglycol content [93].

The water content [94] is also of great importance for the processing of thermoplastic polyesters because water often causes hydrolytic decomposition at high processing temperatures. The equilibrium water content, for a given degree of polymerization $\bar{P}_n$ must not be exceeded on account of the danger of hydrolytic decomposition. This content can be calculated for high conversions from the approximate formula

$$\bar{P}_n = \sqrt{\frac{K}{n_{H_2O}}}$$

$\bar{P}$ where K is the equilibrium constant and n_{H_2O} denotes the number of moles of water per mole of monomer. For example, the maximum water content for PETP ($K \approx 4$) with a molecular mass $M_n \approx 10\,000$ ($\bar{P}_n = 100$) is 0.01 wt %.

A further criterion for the use of thermoplastic polyesters in the fiber or plastics sector is the intrinsic color. This is usually specified as a blue – yellow value (measured with an electric remission photometer) [95].

3.4. Co-Components and Additives [96]

Thermoplastic Copolyesters. The dyeability, elasticity, pilling behavior, shrinkage, hydrophilicity, and flame resistance of polyester fibers can be improved by condensation with appropriate comonomers [97].

The crystallinity, melting point, glass transition temperature, elasticity, flammability, solubility, and other properties of thermoplastic polyesters used in the plastics sector can be modified by selective co-condensation.

Examples of such copolyesters are glass-clear (amorphous) copolyesters based on 1,4-cyclohexanedimethanol, terephthalic acid, and isophthalic acid (i.e., poly(1,4-cyclohexylenedimethylene terephthalate-co-isophthalate) [*26124-27-6*] [98]); elastic copolyesters based on 1,4-butanediol, poly(tetrahydrofurandiol), and terephthalic acid (i.e., poly(butylene terephthalate-co[multibutylenoxy] terephthalate) [71], [72]; neopentyl-glycol-containing copolyesters with good solution properties in paint solvents [99]; and copolyesters based on 1,4-butanediol, terephthalic acid, and isophthalic acid with relatively low melting points for hotmelt adhesives [100]. Thermoplastic copolyesters with elevated glass transition temperatures (and thus im-proved heat resistance) can be obtained by incorporating heterocyclic diols during condensation [101]. For the production of copolyesters, see 3.2.5.

Additives. Raw polymers are seldom processed as such into semifinished or finished articles. Additives are usually added to the polymers in a process known as compounding or formulation. Compounding processes (e.g., mixing, kneading, extrusion, pelletization) are of great importance because the properties, processing, and use of thermoplastic polymers are substantially determined by the additives (e.g., nucleating agents, fillers, flame retardants, stabilizers, pigments).

Nucleating Agents [102], [103]. See also → Plastics, Additives, Chap. 9.1. The processibility of PETP in injection molding is mainly determined by its slow crystallization rate and by the associated danger of cold crystallization (tendency to distortion). High molding temperatures (ca. 140 °C) and addition of nucleating agents are therefore necessary to accelerate the nucleation rate. Inorganic substances such as

talc, magnesium oxide, or clay are added as nucleating agents in concentrations of 0.001 – 1 % with a maximum particle size of 2 – 3 µm. Organic compounds include salts of aromatic carboxylic acids, fatty acid salts, Montan waxes, and polymers such as polyethylene. The nucleating agents can be added at any stage during PETP production or as a powder to the granulate.

PBT and PBT copolyesters with polytetrahydrofuran generally crystallize so rapidly that nucleation is not necessary.

Glass Fibers [104]. See also → Plastics, Additives, Chap. 7. The heat resistance, rigidity, and hardness of thermoplastic polyesters can be considerably increased by adding glass fibers or beads. For example, the heat distortion temperature (according to Martens, DIN 53 458) of PBT reinforced with glass fiber varies as follows:

Heat distortion temperature, °C	50	75	120	160	190
Glass fiber content, wt %	0	10	20	30	50

The glass reinforcement may be added as beads, ground glass fiber, staple glass fiber, or long glass fiber. Maximum heat resistance, rigidity, and hardness are achieved by adding long glass fibers. These additives are mixed with the thermoplastic polyesters in special extruders in a separate compounding stage; amounts of up to 50 % are commonly employed.

Flame Retardants. See also → Flame Retardants, and → Plastics, Additives, Chap. 6. Thermoplastic polyesters (PETP and PBT) have a relatively low flammability because they escape the action of the flame due to melting and droplet formation (oxygen index of PBT, 23 %; UL 94 rating, HB); see also → Flame Retardants, Chap. 3.). Fillers and, in particular, glass fiber reinforcement prevent droplet formation, resulting in a sharp increase in flammability (oxygen index of PBT reinforced with 10 % glass fiber, 19 %). Legislation is increasingly promoting the use of self-extinguishing, nondrip plastics formulations.

Self-extinguishing thermoplastic polyesters are obtained by incorporating ca. 30 % of flame retardants, which are usually aromatic halogenated compounds combinedwith antimony oxide [105] or with phosphorus compounds [106] (see also → Flame Retardants). Examples are brominated polycarbonates [107] and polystyrenes [108], or ethylene(tetrabromophthalimide) in combination with antimony oxide. A combination of red phosphorus and antimony oxide is also used, particularly with PETP [109].

The flame retardants are not normally added during the production of thermoplastic polyesters, but in a subsequent compounding stage.

The polybrominated diphenyl ethers are suspected of producing polybrominated dibenzodioxins and dibenzofurans in the event of fire [110]; they are therefore no longer used in Germany, and interest in halogen-free flame retardants has greatly increased. Up to now, however, only developments and improvements of conventional systems, such as microencapsulated red phosphorus, ammonium polyphosphates, magnesium hydroxide, and magnesium carbonate have been described [111], [112]. The magnesium compounds have the disadvantage that they have to be added in substantially higher amounts and may therefore seriously affect the polymer properties.

Fully aromatic liquid crystalline polyesters have a low flammability and do not require flame retardants.

Stabilizers. See also → Plastics, Additives, Chap. 2., → Plastics, Additives, Chap. 3. In contrast to PETP and PBT, thermoplastic polyetheresters are extremely sensitive to photooxidative decomposition [113] and must be stabilized. Stabilizers include sterically hindered phenols or amines, which are added to the starting substances in an amount of 0.1 –0.5 wt % or are incorporated in the polymer melt [114].

The antioxidant 4,4′-bis(α,α'-dimethylbenzyl)diphenylamine has been approved by the FDA for use in phthalate, isophthalate, and terephthalate polyesters with 1,4-butanediol and polytetrahydrofuran.

Pigments [115], [116]. Pigments with an adequate heat resistance (e.g., titanium dioxide, carbon black, phthalocyanines, quinacridones, cadmium yellow) are suitable for coloring thermoplastic polyesters.

Admixture of Other Polymers. The properties of partially aromatic thermoplastic polyesters can be selectively improved by blending with other polymers. See also, → Polymer Blends, Chap. 4.7. PBTs whose impact properties have been improved with polybutadiene graft rubbers are used in car bumper systems

[117]. Impact strength can also be increased by adding polyethylene [118]. Higher flexibility is obtained by blending with polyether elastomers [119].

New alloying components such as styrene–maleic anhydride copolymers or acrylonitrile–styrene–acrylate copolymers also expand the application range, particularly of PBT [120]. Polymer blends of bisphenol-A-polycarbonate with PBT or PETP have good low-temperature strength and low shrinkage values, they have been commercially successful for several years [121].

Blends with polyurethanes and polyamides [122] or poly-ε-caprolactone [123] are also described.

3.5. Properties

The properties of thermoplastic polyesters depend primarily on the starting compounds (dicarboxylic acids, diols), their molecular mass, and added fillers. The properties of PETP, PBT, and PDCT as fiber raw materials are described in → Fibers, 4. Synthetic Organic, Chap. 2.2.1.,→ Fibers, 4. Synthetic Organic, Chap. 2.4.

PETP and PBT. The two most important thermoplastic polyesters for the plastics sector, namely PETP and PBT, can be described as partially crystalline polymers. They have a high hardness and rigidity, good creep strength, high dimensional stability, and very good slip and wear behavior [124], [125]. Whereas PBT occurs almost exclusively in the partially crystalline form, PETP can also be processed into amorphous molded bodies with high transparency; on heating to 70–100 °C this transparency is lost due to postcrystallization. Transparency can, however, be retained if the polyester chains are forcibly oriented, for example by biaxial stretching and heat setting (transparent films of PETP, → Films, Chap. 6.8.).

A special property of PETP is its slow crystallization, which necessitates molding temperatures of ca. 140 °C and the use of a nucleating agent and crystallization accelerator (see page 16).

PBT and PETP are resistant to water, weak acids and bases, alcohols, ketones, ethers, aliphatic hydrocarbons, and chlorinated aliphatic hydrocarbons at room temperature.

Solvents for PBT and PETP include hexafluoro-2-propanol, hexafluoroacetone, and 2-chlorophenol (mixed solvents, see Table 2).

Hydrolysis proceeds rapidly in the melt in the presence of moisture. The dependence of the hydrolysis rate on temperature and moisture content has been investigated [126], [127].

The most important mechanical, thermal, and electrical properties of PETP and PBT are summarized in Table 3.

Polyetheresters [128–130]. Thermoplastic polyetheresters are block copolymers synthesized from mutually incompatible rigid crystalline and soft amorphous segments. The rigid segments generally consist of an aromatic polyester (e.g., PBT), whereas the soft segments consist of an aliphatic polyether [e.g., poly(butylene glycol) or poly(ethylene glycol)]. On account of the physical cross-linking of the rigid and soft domains, polyetheresters behave like elastomers at the service temperature ($>T_g$, $<mp$). The cross-linking sites are, however, thermally labile and the polymers can therefore be thermoplastically processed above the melting point.

Important thermal, mechanical, and electrical properties of polyetheresters are summarized in Table 4.

The very low glass transition temperature of the polyether (ca. −50 °C) and the high melting point of the polyester (ca. 200 °C) result in polymers with a very broad service temperature range. Polyetheresters with Shore-D hardnesses of 35–72 can be formulated by varying the block length and the ratio of the rigid and soft components.

The special properties depend on the nature of the rigid and soft segments. Copolyesters based on butylene glycol terephthalate crystallize very rapidly from the melt, in contrast to polymers synthesized from ethylene glycol terephthalate blocks. Copolymers with poly-(propylene oxide glycol) and poly(tetramethylene ether glycol) exhibit similar swelling behaviors in water and oil. On the other hand, polymers based on hydrophilic poly(ethylene oxide glycol) swell markedly in hot water, but less in oil. The incorporation of poly(propylene oxide glycol) decreases the thermal stability.

Table 3. Properties of PETP and PBT [125]

Property	Test method	PETP			PBT		
		Partially crystalline	30 % glass fiber reinforced	Partially crystalline (standard grade)	Partially crystalline (tough grade)	30 % glass fiber reinforced	30 % glass fiber reinforced with flame retardant
Mechanical (23° C)							
Tensile strength and yield stress, N/mm^2	DIN 53 455	73	150 – 170	52	38	130 – 140	110 – 120
Elongation at break, %	DIN 53 455	300	2	>250	>250	2	2
Modulus of elasticity tensile test), N/mm^2	DIN 53 457	2850	10 000 – 12 000	2700	1100	8500 – 10 000	9000 – 11 000
Bending strength, N/mm^2	ASTM D 790	118	230 – 250	85	43	200 – 210	190 – 200
Impact strength, kJ/m^2	DIN 53 453		25 – 40			30 – 40	20 – 25
Notched bar impact strength, kJ/m^2	DIN 53 453	4	6 – 14	5	23	8 – 19 ∗	6 – 14 ∗
Ball indentation hardness, N/mm^2	DIN 53 456	170	200 – 230	139	86	200 – 230	210 – 240
Physical and thermal							
Density, g/cm^3	DIN 53 479	1.37	1.58	1.31	1.29	1.53	1.66
Moisture absorption (saturation 23 °C, 50 % relative humidity), %	ASTM D 570	0.35	0.25	0.20	0.20	0.13	0.10
Deflection temperature under load, °C	ISO/R 75, method A	82	225	50	50	205 – 225 ∗	205 – 225 ∗
Linear coefficient of expansion, 10^{-5} K^{-1}	ASTM D 696	7	3	8	13	3	3
Flammability	UL 94						V-0
Crystallite melting point, °C	differential thermal analysis	255 – 260	255 – 260	220 – 225	200 – 205	220 – 225	220 – 225
Electrical							
Volume resistivity, Ω · cm	VDE 0303, part 3	$>10^{14}$	$>10^{15}$	$>10^{15}$	$>10^{15}$	$>10^{15}$	$>10^{15}$
Surface resistance, Ω	VDE 0303, part 3	$>10^{13}$	$>10^{13}$	$>10^{13}$	$>10^{14}$	$>10^{13}$	$>10^{13}$
Dielectric loss factor tan δ,							
50 Hz	VDE 0303, part 4	0.0015	0.002	0.002	0.007	0.0024	0.0025
10^6 Hz		0.019	0.016	0.020	0.021	0.018	0.016
Dielectricity constant ε_r,							
50 Hz	VDE 0303, part 4	4.3	4.3	3.8	3.7	4.4	4.5
10^4 Hz		3.5	4.4	3.2	3.7	3.8	3.8
Dielectric strength, kV/cm	IEC Pub 234, (2 mm plate)	200	170	150	180	170	170

∗ Upper values refer to grades reinforced with long fibers.

Polyetheresters have a good notched bar impact value at low temperature, high strength, flexibility, and abrasion resistance. On account of their partially crystalline nature they are insoluble in most solvents, exceptions include *m*-cresol and tetrachloroethane which are used as solvents for viscosity measurements. The swelling resistance in oils, aliphatic and aromatic hydrocarbons, alcohols, ketones, and esters improves with increasing hardness. Polyetheresters are extremely impermeable to fuels. Their good hydrolysis resistance can be improved by adding polycarbodiimides [135].

The resistance of polyetheresters to light, heat, and oxidation is lower than that of PBT or PETP (for stabilization, see page 17).

Liquid Crystalline Polyesters [136–138]. The two partially crystalline polyesters Xydar and Vectra have only been on the market since the mid 1980s. These liquid crystalline polymers (LCPs) represent one of the most interesting developments in the technical plastics sector. See also → Liquid Crystals, Chap. 4.; → Plastics, General Survey, Chap. 5.3.

On account of their rod-shaped molecular structure, the polymer chains are partially or-

Table 4. Properties of polyetheresters [131], [132]; Xydar [133], [134]; and Vectra [134]

Property	Test method	Polyetherester	Vectra		Xydar	
			A 950 (base type)	A 130 (30 % glass fiber reinforced)	SRT 300 (base type, high-melting)	G 330 (30 % glass fiber reinforced)
Mechanical (23 °C)						
Tensile strength, N/mm^2	DIN 53 455	9 – 46	156	188	138	117
Elongation at break, %	DIN 53 455	170 – 700	2.6	2.1	4.9	2.6
Tensile modulus of elasticity, N/mm^2	ASTM D 638		9700	16 600	16 600	12 000
Bending modulus of elasticity, −40 °C, N/mm^2	ASTM D 790	160 – 2400				
23 °C N/mm^2	ASTM D 790	50 – 600	9000	15 000	13 800	11 600
100 °C N/mm^2	ASTM D 790	30 – 210				
Bending strengh, N/mm^2	ASTM D 790				131	150
Notched bar impact strength according to Izod, J/m −40 °C	ASTM D 256	without break				
23 °C	ASTM D 256	without break	520	150	130	176
Hardness, Durometer *D*, points	DIN 53 505	35 – 72				
Physical and thermal						
Density, g/cm^3	ASTM 0792	1.1 – 1.3	1.4	1.6	1.4	1.6
Moisture absorption, %	ASTM D 570	0.2 – 0.6	0.02	0.02	<0.1	<0.1
Heat distortion temperature (HDT/A, 1.82 N/mm^2), °C	ASTM D 648	<50	180	230	355	252
Linear coefficient of expansion, 10^{-6} K^{-1}						
Longitudinal	ASTM D 696	ca. 90 – 200	−5	5		5
Transverse		ca. 90 – 200	75	65		
Flammability	UL 94	HB	V-0	V-0	V-0	V-0
Melting point, °C	differential scanning calorimetry	148 – 219	280	280	421	
Vicat softening point, °C	DIN 53 460	80 – 207	145	177		
	ASTM D 1525				369	
Electrical						
Volume resistivity, Ω · cm	DIN 53 482	10^{10}–10^{14}	10^{16}	10^{16}		
Surface resistance, Ω	DIN 53 482	$>10^{11}$	4×10^{13}	8×10^{13}		
Dielectric loss factor tan δ, 10^6 Hz	ASTM D 150	0.036 – 0.06	0.02	0.017	0.041	0.033
Dielectric constant ε_r, 10^6 Hz	ASTM D 150		3.3 (kHz)	3.7	3.1	3.6
Dielectric strength, kV/mm	ASTM D 149		39	43	23	35

dered in the melt. The macromolecules adopt a parallel orientation in microscopically small regions, with the result that the melt exhibits properties of both a solid (order) and a liquid (mobility). This intermediate state is termed liquid crystalline. Xydar and Vectra have nematic phases, i.e., the longitudinal axes of the molecules are parallel, but the molecular centers of gravity are irregularly distributed (→ Liquid Crystals, Chap. 3.1.). Under the influence of a shear force, such as occurs in normal thermoplastic distribution, the microscopically ordered regions can be macroscopically oriented. This orientation is retained in the solid state and is responsible for the outstanding properties of liquid crystalline polyesters. These properties are, however, strongly orientation-dependent due to the anisotropy of the polymers. LCPs cannot therefore be compared with conventional thermoplastic polymers solely on the basis of values measured on uniaxial test bodies. Nevertheless, LCPs differ from "conventional" thermoplastics as regards the following properties (see also Table 4):

1) Their coefficient of thermal expansion is unusually low for polymers and can reach the values of steel; it can be adapted to other

materials by varying the processing conditions2) Good processibility: on account of their very low melt viscosity, even filigree structural parts can easily be filled
2) Low melt enthalpy (short cycle times)
3) Low flammability: V-0 rating according to UL 94 with wall thickness <0.8 mm
4) High heat resistance
5) Resistance over a wide temperature range to solvents, fuels, and most chemicals, except oxidizing acid and strong alkali
6) Resistant to radiation and transparent to short-wavelength light
7) Low water absorption
8) Very good mechanical properties (strength, impact resistance, modulus of elasticity)

The anisotropy of the mechanical properties and the thermal coefficient of expansion can be reduced by adding fillers (glass fibers, minerals), which simultaneously improve surface quality. LCPs have a poor weld strength, in unfilled LCPs it is about half the strength transverse to the flow direction. To overcome this, special molded parts are constructed in which the weld is located in such a way that it is not subsequently exposed to high stress.

The most important mechanical, thermal, and electrical properties of LCPs are summarized in Table 4.

3.6. Uses

Poly(Ethylene Terephthalate). More than 85 % of PETP is processed into fibers (→ Fibers, 4. Synthetic Organic, Chap. 2.3.3., → Fibers, 4. Synthetic Organic, Chap. 2.3.4., → Fibers, 4. Synthetic Organic, Chap. 2.3.5.). Poly-(1,4-dimethylenecyclohexane terephthalate) is used for special fibers (see → Fibers, 4. Synthetic Organic, Chap. 2.4.2.). The production of high-grade, heat-resistant PETP films is described in → Films, Chap. 6.8. A large proportion of PETP is used to produce gastight bottles for carbonated beverages (see Section 3.8). Highly stressed technical molded parts such as bearings, gearteeth, cam wheels, connectors, bolts, screws, and washers are produced from PETP by injection molding [139].

On account of the processing difficulties (see Section 3.4) fewer molded parts are made from PETP than from PBT. If new developments aimed at reducing the high molding temperatures prove to be successful, this proportion may rapidly increase [140]. Injection molding machines are mainly used for processing.

Trade names include Arnite, Akzo; Crastin, CibaGeigy; Grilpet, EMS-Chemie; Impet, Hoechst Celanese; Melinite, ICI; Rynite, Du Pont; Techster, Rhône-Poulenc; Valox, GE Plastics.

Poly(Butylene Terephthalate). PBT too can be spun into fibers (see → Fibers, 4. Synthetic Organic, Chap. 2.4.), which can be colored, without carriers, with disperse dyes and used in the carpets sector [141]. Trade names are sometimes the same as for PETP fibers. Examples include Trevira, Hoechst; Tergal, Rhône-Poulenc; Fortrell, Celanese.

Most PBT is processed by injection molding and extrusion, a relatively low-viscosity formulation is used for injection molding, and a medium-to high-viscosity formulation for extrusion.

Examples of injection-molded PBT products (unreinforced and reinforced) include housings for heated domestic appliances (e.g., coffee-making machines, egg boilers, toasters, hairdryers); bearing bushes, trip cams, gear wheels, fan wheels, ropesheaves, and pump impellers for the mechanical engineering sector; coil formers, capacitor covers, plug boards, and lamp sockets (often with flame retardants) for the electrical sector.

Protective covering for optical waveguides, films, sheets, monofilaments, and hydraulic lines are produced by extrusion. Paper and board can also be extrusion coated.

Trade names include B. Arnite, Akzo; Celanex, Hoechst Celanese; Crastin, Ciba-Geigy; Grilpet, EMS-Chemie; Orgater, Atochem; PBT, Toray; Pibiter, Montedison; Pocan, Bayer; Rynite, Du Pont; Techster, Rhône-Poulenc; Tenite, Eastman Kodak; Ultradur, BASF; Valox, GE Plastics; Vestodur, Hüls.

Copolyesters. Reference has already been made to possible uses of *thermoplastic copolyesters* (Section 3.4). Copolyetheresters [synthesized from dimethyl terephthalate, 1,4-butanediol, and poly(tetrahydrofuran)diol] are used on account of their high flexibility in the

cold (without plasticizers), particularly for winter sports articles (e.g., ski brakes, soles for cross-country ski shoes, disks for ski poles). A novel field of application is crawlers for snowmobiles, where high abrasion resistance is also required. Copolyetherester elastomers can be processed by rotational casting into tires for vehicles not intended for use on roads (lawnmowers, garden tractors, golf trolleys), and into hydraulic oil tanks. Highly flexible pipes (e.g., for hydraulics lines) can be produced by extrusion.
Trade names include Hytrel, Du Pont; Arnitel, Akzo; Lomod, General Electric Plastics.

Amorphous copolyesters [99] produced from dimethyl terephthalate, terephthalic acid, isophthalic acid, neopentyl glycol, ethylene glycol, and other monomers are soluble in conventional solvents and can be processed into highly elastic, sterilization-resistant coatings (coil coatings). See also → Paints and Coatings, Chap. 2.7.
Trade names include Dynapol L, Hüls; Vitel, Goodyear.

Partly crystalline copolymers based on poly(butylene terephthalate-*co*-isophthalate) are valuable hotmelt adhesives [100] for the shoe industry (shoe soles) or furniture industry.
Trade names include Dynapol S, Hüls; Vestamelt, Hüls.

Copolyesters based on similar starting components are used for powder coatings that produce highly flexible, weather-resistant coatings [142].
Trade names include Dynapol P, Hüls.

Liquid Crystalline Polyesters. Xydar and Vectra types cannot be used without reinforcement or fillers on account of their marked anisotropic properties and tendency to fibrillation. They are usually reinforced or filled with glass fibers, carbon fibers, or minerals and are processed by injection molding.

Areas of application exploit the properties described in Section 3.5. Examples include the automotive and aviation industries, where good thermal and mechanical properties, inherent flame resistance, and low fume formation are important for interior fixtures and fittings.

On account of the low thermal expansion, high heat resistance, and favorable electrical properties, LCPs are employed in the microelectronics and optoelectronics sectors, mainly for integrated circuits, connectors, coil formers, lighting fixtures, and sheathing for glass fiber cables.

On account of their good resistance to chemicals and solvents, LCPs are used as packings in distillation columns and for fuel lines. Vectra (Hoechst-Celanese) and Xydar (Amoco) are transparent to microwave radiation, and are therefore used for microwave cookware and structural parts in microwave ovens.

3.7. Toxicology

High molecular mass polycondensates are usually considered to be biologically inert. However, the biological effect of polymers can be influenced by soluble or volatile components of monomers, oligomers, additives, and degradation products.

The toxicology of PETP has been investigated, especially for its use as a packaging film and textile fiber. No toxic effects were observed [143]. Tests on bacteria showed that neither PETP polyester nor other substances contained in PETP bottles extracted with dichloromethane were mutagenic [144].

PETP, BPT, and various copolyesters with 1,4-cyclohexanedimethanol and/or poly(tetrahydrofurandiol) are permitted by the FDA. They can therefore be used in contact with foods under particular conditions (viscosity, extract content, temperature, etc.) in accordance with CFR from 1988: CFR 170.1240; CFR 177.1315; CFR 177.1590; CFR 177.1630; CFR 177.1660.

Polyterephthalate diol esters are approved by the German health authority (Bundesgesundheitsamt) for use according to the food and commodity law (Aug. 15, 1974). The starting materials which can be used are 1,4-butanediol, ethylene glycol, 1,4-cyclohexanedimethanol, poly(ethylene glycol), terephthalic acid, dimethyl terephthalate, isophthalic acid, adipic acid, azelaic acid, dimethyl azelate, sebacic acid, dimethyl sebacate, *n*-dodecanedioic acid.

3.8. Economic Aspects [140], [145–151]

PETP and PBT. The above-average growth in thermoplastic polyesters that occurred at the start of the 1980s has continued. Double-figure

Table 5. Development of demand for PETP and PBT molding materials (without blends) in 10^3 t

	1984	1986	1988	1989
PETP				
United States	9	13	18	20
Japan	6.5	8	10	12
Western Europe	3.8	5	6	8
PBT				
United States	36	51	60	66
Japan	29	39	49	55
Western Europe	19	23	29	31

growth rates (ca. 15 %) in the PBT and PETP molding materials sector have been achieved worldwide. Table 5 shows the development of demand for PETP and PBT between 1984 and 1989.

Most thermoplastic molding materials are processed in glass-fiber-reinforced form. PETP molding compounds are difficult to process and are therefore used in substantially smaller amounts than PBT molding compounds. The main areas of use are electrical engineering and electronics (25 – 30 % of overall production) and automobile construction (20 – 25 % without blends).

PBT – PETP blends, which have a good surface quality, as well as PBT – polycarbonate blends with high impact strength have become important. The amount of PBT used for blends in 1988 was ca. 10 000 t/a in Western Europe and ca. 20 000 t/a in the United States. The amount of PETP used for blends in 1988 was ca. 2000 t/a in Western Europe and 10 000 t/a in the United States.

PBT is largely used for construction materials, whereas PETP is mainly processed into fibers, sheets, films, and beverage containers. In 1989 15.6×10^6 t of synthetic fibers were produced worldwide; polyester fibers, primarily PETP, accounted for 8.4×10^6 t (54 %). The development of consumption for PETP sheets, films, and bottles is given in Table 6.

Films and sheets of PETP are chiefly used in the photographic sector, magnetic tapes, and packaging materials.

Polyetheresters. The worldwide demand for polyetheresters based on dimethyl terephthalate, 1,4-butanediol, and polytetrahydrofuran was estimated to be 38 000 t in 1991. Of this figure, the United States accounts for 22 000 t and Western Europe and Japan 8000 t each. The annual growth rates are 10 – 15 %. The most important areas of use are hoses, cables, and automobile parts.

Liquid Crystalline Polyesters. The plant capacities for LCPs are ca. 9000 t/a, of which Vectra (Hoechst-Celanese) accounts for ca. 2300 –3600 t/a [152] and Xydar (Amoco) for ca. 4500 t/a [153]. The demand in the United States (ca. 90 % of overall demand), Western Europe, and Japan is considerably lower, at ca. 2500 t/a. However, high average growth rates, particularly in the electrical engineering and electronics sector, of at least 15 % to more than 20 % per annum are anticipated [154].

The major proportion (ca. 60 %) of LCPs is used for electrical and electronics components as well as for microwave-transparent cook ware. The remainder is used in the industrial, telecommunications, automotive, and aviation sectors.

3.9. Recycling

Polyesters (PETP, PBT) play an important role in the production of beverage bottles, foils, films, and fibers. The local authorities, in particular, protest against the increase in the volume of waste caused by the use of nonreturnable polyester bottles. In countries in which these bottles have already been introduced (e.g., USA, Japan, France, Italy, and the UK) laws have been passed to enforce the return and reutilization of the bottles. Because of this, the recycling rate for PETP bottles in the United States, for example, is 90 % [155]. In Europe the proportion of collected PETP is ca. 4 %. Two methods are used for recycling polyester: (1) direct reuse after comminution and washing, and (2) alcoholysis.

Table 6. Development of consumption for PETP sheets/films and bottles in 10^3 t

Region	1984	1986	1988	1989
United States				
Sheets/films	209	200	232	237
Bottles	243	291	334	360
Western Europe				
Sheets/films	130	170	190	195
Bottles	70	110	175	195
Japan				
Sheets/films	93		180	
Bottles	27		65	

It is more difficult to directly reuse PETP foils and films of which the majority consist of composite materials. Such waste can only be processed by combustion or chemical recycling via pyrolysis.

Reuse. Prerequisites for good recycling of PETP materials are [156]:

1) There must be a sufficient amount of homogeneous recyclable material
2) An economical recycling process must be available for cleaning and separating the materials
3) The recovered plastic must have a sufficiently large market and a sufficiently high price to warrant economical recycling

This means that in designing and manufacturing a product, allowances have to be made for its subsequent recycling product. As a result of its high density (1.32 – 1.40 g/cm^3), PETP can be efficiently separated from most other plastics. Separation from poly(vinyl chloride) (PVC) is possible with the help of flotation. High standards are required during sorting for reuse as material for bottles. Due to the incompatability of PETP with the majority of other plastics the proportion of the latter should not exceed 100 ppm, i.e., only one polyethylene or PVC bottle should be among 10 000 PETP bottles.

In addition to collection, the following costs are estimated:

Sorting	0.15 €/kg
Grinding	0.10 €/kg
Washing, drying	0.20 €/kg

In the United States a new plant has been built by Du Pont and Waste Management of North America (WMNA) in Chicago (capacity 20 000 t/a); it separates the PETP chips and converts them to clean polymer. Du Pont then uses or sells the final product [155].

In every recycling step some cyclic and linear oligomers are formed. Especially the linear oligomers cause some problems during processing.

Hydrolysis and Alcoholysis. Another way to break down polyester is hydrolysis or alcoholysis which sometimes requires drastic reaction conditions and long reaction times. The hydrolysis of polyesters results in the formation of carboxylic acids and alcohols, from which new polyester can only be produced after separation and purification.

The breakdown of polyesters is more readily achieved by alcoholysis. The diols and dicarboxylates are formed according to the following reaction scheme:

$$\sim\!\sim R'-CO-O-R-O-CO-R'-CO-O-R\sim\!\sim \xrightarrow{R'OH}$$

$$\sim\!\sim R'COOR' + HOROH + R'OOC-R'-COOR' + HOR\sim\!\sim$$

If methanol is used for the alcoholysis of PETP, the monomers used for its polymerization are reformed. If multifunctional alcohols are employed, polyols can be produced, which can be modified and used for various purposes (e.g., for the production of polyurethane foam). Both processes are carried out on a commercial scale in the United States [157].

Glycolysis to monomer (dihydroxyethyl terephthalate)and low molecular mass oligomers represents a compromise between the regeneration of the starting materials by methanolysis or hydrolysis and direct remelt extrusion. Glycolysis signifies degradation with glycols [158]; it is less expensive than methanolysis or hydro-

lysis and more versatile than remelt extrusion. The resultant low-viscosity monomer can easily be filtered and is repolymerized to a useful molecular mass; it is used as a comonomer and ingredient for controlling luster, color, etc.

Glycolytic recycling can be a batch or continuous process, the rate depending on temperature, catalyst, state of subdivision of the feedstock and the glycolytic PETP ratio. In addition, the final monomer composition is controlled by the reaction time and holdup time after depolymerization. Low ratios of glycol to ester permit higher temperatures and faster reactions but result in higher molecular mass oligomers. Side reactions occurring during depolymerization are minimized by the addition of a buffer such as sodium acetate and by limiting the storage time at high temperature. PETP scrap suitable for glycolytic recycling includes production waste, fibers, films, flakes, and bottles; however, ingredients leading to side reactions, copolymers, end caps, or color must be minimized for critical products and satisfactory repolymerization rates.

Pyrolysis. If the polyester waste is highly soiled or combined with other materials, then in most cases it can only be chemically recycled by pyrolysis [159]. Pyrolysis denotes thermal decomposition under the exclusion of air. Mixtures of hydrocarbons are obtained, whose value is smaller than that of the glycolysis products but higher than that of the pure calorific value.

In the process developed by the University of Hamburg, polyester and other plastics are broken down in a fluidized bed that is indirectly heated by radial thermal pipes [160]. The fluidized bed consists of fine-grained sand (0.3 – 0.5 mm) and is heated to 650 – 850 °C. The pulverized polyester waste is transferred into the fluidized bed with a screw. After leaving the reactor the pyrolysis products are freed of entrained sand and fillers in a cyclone and cooled, whereby oils are condensed and expelled. The residual gas serves for the fluidization of the bed and as heating gas for the burner and other purposes. The plant has a capacity of 20 – 50 kg/h.

120.5 kg of chips consisting of PETP or exposed photographic films were fed with a screw (capacity 18.9 kg/h) into the fluidized bed and heated to 768 or 700 °C, respectively [161]. Table 7 summarizes the product groups obtained from PETP. The fraction of water, which is low (2.1 wt %), is the result of adhering moisture and formation during the decomposition of PETP. There are more water and solid residues in the pyrolysis products of film materials (Table 8). The main gaseous components are carbon monoxide and dioxide. Of the hydrocarbons, only methane and ethylene have notable percentages. The net calorific value (13.4 MJ/m^3) is ca. 40 % of the value for natural gas. As a result of its hydrogen content of 8.7 vol % the gas burns with a less vigorous flame.

Table 7. Products obtained after pyrolysis of 120.5 kg polyester at 768 °C in a fluidized bed

Products	Content	
	kg	wt %
Gas	61.2	50.8
Benzene- and toluene-boiling fractions	26.7	22.2
High-boiling fraction	12.8	10.6
Tar ($bp > 340$ °C)	8.7	7.2
Soot (organic solids)	8.6	7.1
Water	2.5	2.1

Table 8. Components (wt %) obtained after pyrolysis of PETP at 768 °C and film material (PETP, polyethylene, and gelatin) at 700 °C

Component	Pure PETP	Used film
Hydrogen	0.30	0.37
Methane	3.80	5.28
Ethane	0.24	1.70
Ethylene	1.50	3.44
Propene	0.08	1.73
Butene	0.01	0.41
Butadiene	0.04	0.27
Carbon dioxide	17.20	17.30
Carbon monoxide	24.70	16.39
Benzene	18.30	7.79
Toluene	2.50	1.48
Styrene, ethylbenzene	2.55	6.26
Indene	0.41	0.23
Naphthalene	1.40	2.04
Biphenyl	1.47	0.97
Fluorene	0.18	0.10
Ketones, acetates	0.15	0.12
Acetophenone	1.40	1.81
Other compounds	11.87	11.15
Carbon black, fillers	7.10	13.40
Water	2.10	8.76

The largest oil fraction is the benzene- and toluene-boiling fraction, in which the benzene, toluene, and xylene (BTX) aromatics account for over 97 %. This means that pyrolysis of 1 t of polyester waste yields 215 kg of the petrochemically valuable BTX aromatics. The concentra-

tions of the aliphatic and heteroatom-containing hydrocarbons are both < 0.2 wt %. These low concentrations make the processing of this pyrolysis oil fraction to BTX aromatics very much easier.

In contrast, the high-boiling components, constitute a multicomponent system. This fraction, including tars, accounts for 17.8 wt %. The main components are naphthalene, biphenyl, indene, and acetophenone. The high content of acetophenone (10 wt %) is characteristic of the pyrolysis of polyester (PETP). This compound is not formed during the pyrolysis of polyolefins. After pyrolysis the remainder of the oxygen bound to the polyester is distributed as follows: 62 % in carbon dioxide, 31.6 % in carbon monoxide, 5.5 % in water, 0.75 % in acetophenone, and 0.15 % in other keto-compounds and carboxylic acids.

The pyrolysis of polyester and composite materials such as films (PETP, polyethylene, gelatin) in the fluidized bed showed that oxygen-containing plastics can be largely defunctionalized and high proportions of mononuclear aromatics can be obtained.

4. References

1. O. Bayer, *Angew. Chem.* **59** (1947) no. 9, 257 – 272.
2. I. G. Farbenind., DE 756 058, 1940 (W. Bunge, O. Bayer, S. Petersen, G. Spielberger).
3. Dynamit Nobel, DRP-Anm. D 90 260, 1943 (H. Pinten).
4. A. Höchtlen, *Kunststoffe* **40** (1950) 221.
5. E. Müller et al., *Angew. Chem.* **64** (1952) no. 19/20, 523 – 531.
6. H. Wagner, H. F. Sarx: *Lack Kunstharze,* 5th ed., Hanser Verlag, München 1971.
7. R. D. Aylesworth et al., *Mod. Plast.* **35** (1958) no. 9, 145.
8. Bayer, EP 0 180 749, 1988 (G. Weber, K. König, M. Schmidt).
9. Union Carbide, US 2 890 208, 1956 (D. M. Young, F. Hostettler, C. Horn); US 3 284 417, 1966 (F. Hostettler, G. Magnus, H. Vineyard); EP 117 538, 1984 (L. Domeier, E. Hsi).
10. BASF, GB 2 147 595, 1983 (C. J. Reichel). Bayer AG, DE-OS 3 437 915, 1984 (M. Barnes, W. Betz). DE-OS 3 621 039, 1986 (M. Dietrich, M. Kapps, K. König, R. Nast).
11. J. H. Saunders, K. C. Frisch: *Polyurethanes, Chemistry and Technology,* Interscience, New York 1962.
12. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, New York 1953.
13. P. J. Flory, *Angew. Chem.* **86** (1974) 109.
14. W. Griehl, G. Schnock, *Faserforsch. Textiltech.* **8** (1957) no. 10, 408.
15. G. Henrici-Olivé, S. Olivé: *Polymerisation, Katalyse, Kinetik, Mechanismen,* Verlag Chemie, Weinheim 1969.
16. *Houben-Weyl* **14/2,** 2.
17. BASF, NL 6 505 683, 1965.
18. H. Batzer, *Makromol. Chem.* **7** (1951) 82.

General References

19. V. V. Korshak, S. V. Vinogradova: *Polyesters,* Pergamon Press, London 1965.
20. R. Vieweg, L. Goerden: "Polyester", *Kunststoff-Handbuch,* vol. VIII, Hanser Verlag, München 1973.
21. H. Ludewig: *Polyester-Fasern,* 2nd. ed., Akademie-Verlag, Berlin 1975.
22. *Process Economics Program,* Stanford Research Inst., Menlo Park, Ca., vol. 18, 1966, vol. 18 A, 1972; vol. 96 (1975); vol. 96 A, 1977.
23. B. M. Walker, C. P. Rader: *Handbook of Thermoplastic Elastomers,* Rheinhold Publ., New York 1988.
24. A. Blumstein: *Polymeric Liquid Crystals,* Plenum Press, New York 1985.
25. L. L. Chapoy: *Recent Advances in Liquid Crystalline Polymers,* Elsevier Applied Science Publishers, London 1985.

Specific References

26. D. W. van Krevelen, J. Bussink, F. J., Huntjens, J. L., Voigt: "Processing Polymers to Products," Internat. Congress 1966, Amsterdam Oct. 17 – 19, 1966, N.V.'t Raedthuys, Utrecht, p. 167.
27. H. G. Elias, F. Vohwinkel: *Neue Polymere Werkstoffe für die industrielle Anwendung,* Hanser Verlag, München 1988.
28. J. Economy, *Mol. Cryst. Liq. Cryst.* **169** (1989) 1.
29. W. J. Jackson, Jr., *Mol. Cryst. Liq. Cryst.* **169** (1989) 23.
30. M. Hizikata, *CEER Chem. Econ. Eng. Rev.* **9** (1977) no. 11, 32.
31. J. Dörffel, J. Rüter, W. Holtrup, R. Feinauer, *Farbe +Lack* **82** (1976) no. 9, 796.

32. P. J. Flory, *J. Amer. Chem. Soc.* **58** (1936) 1877. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca N.Y., 1953. H. G. Elias, *Makromoleküle,* **vol. 1,** Hüthig & Wepf Verlag, Basel 1990, p. 232.
33. J. Hsu, K. Y. Choi, *J. Appl. Polym. Sci.* **33** (1987) 329.
34. J. Otton, S. Ratton, *J. Polym. Sci. Part A* **27** (1989) 3535.
35. L. Yurramendi, M. J. Barandiaran, J. M. Asúa, *Polymer* **29** (1988) 871.
36. F. Pilati et al., *Polymer* **22** (1981) 799; **22** (1981) 1566.
37. F. Pilati et al., *Polymer* **24** (1983) 1479.
38. T.-Y. Ju et al., *Polymer* **27** (1986) 1111.
39. A. A. Khan, K. Y. Choi, *J. Appl. Polym. Sci.* **37** (1989) 707.
40. A. Fradet, E. Maréchal, *Adv. Polym. Sci.* **43** (1982) 52.
41. G. Challa, *Makromol. Chem.* **38** (1968) 105.
42. G. Reinisch, H. Zimmermann, G. Rafter, *Faserforsch. Textiltech.* **20** (1969) no. 5, 225.
43. T. Tani, K. Enok, *Hydrocarbon Process.* **49** (1970) no. 11, 146.
44. P. W. Morgan: *Condensation Polymers by Interfacial and Solution Methods,* Interscience, New York 1965.
45. M. Levin, S. S. Temin, *J. Polym. Sci.* **28** (1958) 79.
46. R. F. Fischer, *J. Polym. Sci.* **44** (1960) 155.
47. K. Hamann, *Makromol. Chem.* **51** (1962) 53.
48. UCC, US 3 021 313, 1959 (E. F. Cox, F. Hostettler).
49. J. M. Besnoin, K. Y. Choi, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 29** (1989) no. 1, 55.
50. B. Möller et al., *Acta Polym.* **33** (1982) no. 1, 38.
51. G. Rafler, J. Blaesche, *Acta Polym.* **33** (1982) no. 8, 472.
52. K. Yoda, K. Kimoto, T. Toda, *Nippon Kagaku Kaishi* **67** (1964) 909.
53. K. Tomita, H. Ida, *Polymer* **16** (1975) 185.
54. H. Zimmermann, E. Schaaf, *Faserforsch. Textiltech.* **20** (1969) no. 4, 185.
55. K. Tomita, *Polymer* **17** (1976) 221.
56. L.-C. Hsu, *J. Macromol. Sci. Phys.* **B 1** (1967) no. 4, 801.
57. Hüls, DE-OS 2 811 982, 1978 (G. Horlbeck, K. Burzin, R. Feinauer).
58. *Chem. Eng. News* **49** (1971) no. 1, 31.
59. Goodyear, GB 1 261 529, 1972 (R. V. Rinehart).
60. Goodyear, GB 988 548, 1965 (W. B. Pengilly).
61. U. Tiehle, *Kunststoffe* **79** (1989) 11.
62. Fiber Industries, GB 1 296 242, 1972 (D. J. Lowe).
63. Fiber Industries, CA 962 396, 1975 (P. Schaefer, W. H. Yates, P. A. Mason).
64. J. F. Kemkes, *J. Polym. Sci. Part C* **22** (1969) no. 2, 713.
65. S. G. Hovenkamp et al., *J. Polym. Sci. Polym. Chem. Ed.* **8** (1972) no. 3, 679.
66. Standard Oil, US 4 014 858, 1977 (G. R. Chipman).
67. H. D. Schumann, *Chemiefasern/Textilind.* **40** (1990) 1058.
68. Du Pont, EP 0 431 977, A 2, 1989 (D. J. Lowe).
69. Mitsubishi, DE-OS 2 539 249, 1974 (T. Kimura, K. Kohyama, T. Okada, K. Sakata).
70. Eastman Kodak, US 2 901 466, 1959 (C. J. Kibler, A. Bell, J. G. Smith).
71. Du Pont, DE 2 213 128 C 3, 1986 (W. K. Witsiepe).
72. Du Pont, DE 2 265 319 C 3, 1989 (W. K. Witsiepe).
73. Celanese Corporation, EP 0 088 546 B 1, 1983 (A. J. East).
74. Celanese Corporation, EP 0 117 086 B 1, 1984 (L. F. Charbonneau).
75. Dart Industries, DE 3 443 219 A 1, 1985 (A. B. Firestone).
76. Celanese Corporation, EP 0 183 433 B 1, 1989 (G. D. Kiss).
77. Celanese Corporation, DE 2 844 817 C 2, 1989 (G. W. Calundann).
78. P. Edwards, *Mod. Plast. Intern.* **19** (1989) March, no. 3, 32.
79. Celanese Corporation, US 4 567 247, 1984 (H. N. Yoon).
80. H.-G. Elias: *Makromoleküle,* 4th ed., Hüthig & Wepf Verlag, Basel 1981, p. 259.
81. H. M. Koepp, H. Werner, *Makromol. Chem.* **32** (1959) 79.
82. G. Meyerhoff, S. Shimotsuma, *Makromol. Chem.* **135** (1970) 195.
83. A. Horbach, R. Binsack, H. Müller, *Angew. Makromol. Chem.* **98** (1981) 35.
84. L. D. Moore, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **1** (1960) 234.
85. W. H. F. Borman, M. Kramer, *Org. Coat. Plast. Chem.* **34** (1974) 77.
86. D. Garske, W. Kaufmann, *Kunststoffe* **68** (1978) no. 2, 87.
87. F. Bueche, S. W. Harding, *J. Polym. Sci.* **32** (1958) 177.
88. D. R. Gregory, *J. Appl. Polym. Sci.* **16** (1972) 1479.

89. M. Brown, W. K. Witsiepe, *Rubber Age (N.Y.)* **104** (1972) no. 3, 35. K. F. Wissbrun, G. Kiss, F. N. Gogswell, *Chem. Eng. Commun.* **53** (1987) 149.
90. B. Vollmert: *Grundriß der Makromolekularen Chemie,* **vol. III,** E. Vollmert-Verlag, Karlsruhe 1982, p. 72.
91. A. Conix, *Makromol. Chem.* **26** (1958) 226. H. Zimmermann et al., *Faserforsch. Textiltech.* **18** (1966) 487, 536.
92. H. A. Pohl, *Anal. Chem.* **26** (1954) 1614.
93. L. H. Ponder, *Anal. Chem.* **40** (1968) 229. H.-D. Dinse, E. Tuček, *Faserforsch. Textiltech.* **21** (1970) no. 5, 205.
94. F. Salzer, H. Geldermann, *Kunststoffe* **60** (1970) no. 9, 688.
95. G. W. Becker, D. Braun: "Die Kunststoffe, Chemie, Physik, Technologie," *Kunststoffhandbuch,* Hanser Verlag, München 1990, p. 814 ff.
96. J. T. Lutz, Jr.: *Thermoplastic Polymer Additives,* Marcel Dekker, New York 1989.
97. P. Ehrler, G. Egbers, *Mitt. Text. Ind.* **79** (1972) no. 10, 361.
98. Eastman Kodak, Kodar Copolyester A 1 50, company brochure TR-50 (1975) and MB-58 B (1976).
99. Dynamit Nobel, DE-AS 1 807 776, 1970. Chem. Werke Hüls, DT 2 521 791, 1975 (F. Blaschke et al.).
100. Dynamit Nobel, DE-AS 1 912 117, 1969. K. Brühning, K. G. Sturm, *Adhesion* **3** (1973) 83.
101. Ciba-Geigy, DE-OS 2 436 109, 1973 (L. Buxbaum et al.).
102. R. Vieweg, L. Goerden: "Polyester," *Kunststoffhandbuch,* vol. VIII, Hanser Verlag, München 1973, p. 701.
103. J. M. Dekoninck, R. Legras, J. P. Mercier, *Polymer* **30** (1989) 910 – 913.
104. Celanese, DE-AS 2 042 447, 1970 (D. D. Zimmermann).
105. BASF, DE-OS 2 247 652, 1972 (W. Seydl). Celanese, DE-AS 2 042 450, 1970.
106. P. J. Koch et al., *J. Appl. Polym. Sci.* **19** (1975) 227.
107. Teijin, DE-OS 2 726 110, 1977 (N. Saiki, T. Kaneko).
108. BASF, DE-AS 2 515 473, 1975 (R. Wurmb et al.).
109. BASF, DE-OS 2 249 910, 1972 (W. Seydl).
110. D. Bieniek, M. Bahadir, F. Korte, *Heterocycles* **28** (1989) 719.
111. H. Staendeke, D. J. Scharf, *Kunststoffe* **79** (1989) no. 11, 1200.
112. G. Kirschbaum, *Kunststoffe* **79** (1989) no. 11, 1205.
113. A. V. Patsis (ed.): *Advances in the Stabilisation and Controlled Degradation of Polymers,* vol. I, Technomic Publishing Co., Lancaster – Basel 1986.
114. J. J. Zeilstra, *Angew. Makromol. Chem.* **137** (1985) 83 – 92.
115. S. B. Gordon, *Plast. Eng.* **39** (1983) no. 11, 37.
116. A. S. Wood, *Mod. Plast.* **15** (1985) no. 10, 60.
117. D. Rempel, *Kunststoffe* **76** (1986) 900.
118. Enka, DE 1 273 193, 1965 (W. Rein, E. Siggel).
119. Bayer, DE-OS 2 363 512, 1973 (H. Hespe, D. Rempel, L. Morbitzer).
120. A. S. Wood, *Mod. Plast. Intern.* **19** (1989) Aug., 32.
121. W. Witt, *Kunststoffe* **77** (1987) 1009.
122. BASF, DE-OS 2 350 852, 1973 (R. Wurmb, J. Kinde).
123. General Electric, DE-OS 2 331 826, 1973 (D. W. Fox, A. D. Wambach).
124. K. D. Asmus, *Kunststoffe* **62** (1972) 635.
125. F. Breitenfellner, J. Habermeier, *Kunststoffe* **66** (1976) 610.
126. K. Kishore, S. Sankaralingam, *Polym. Eng. Sci.* **24** (1984) 1043.
127. H. Zimmermann, N. Thalkim, *Polym. Eng. Sci.* **20** (1980) 680.
128. H. Krodel, *Kautsch.+Gummi, Kunstst.* **42** (1989) 304.
129. L. G. P. Dalmolen, *Kautsch.+Gummi, Kunstst.* **42** (1989) 205.
130. H. Dominninghaus, *Plastverarbeiter* **40** (1989) no. 2, 42.
131. Du Pont, Hytrel, technical information, 1989.
132. H. Saechtling: *Kunststoff Taschenbuch,* 24th ed., Hanser Verlag, München 1989.
133. Dartco Mfg., Xydar, technical information, 1985.
134. Hoechst, Vectra, technical information, Sept. 1989.
135. I. Whelan, J. Goff, *Br. Plast. Rubber* **2** (1985) 42.
136. P. D. Frayer, *Polym. Compos.* **8** (1987) no. 6, 379.
137. T.-S. Chung, *Polym. Eng. Sci.* **26** (1986) no. 26, 901.
138. J. R. Dole, *CHEMTECH.* **17** (1987) no. 4, 242.
139. *Gummi, Asbest, Kunstst.* **22** (1969) no. 4, 368. K. D. Asmus, D. Niedernberg, H. Schell, *Kunststoffe* **57** (1969) 266.
140. *Kunststoffe* **79** (1989) no. 10, 925.
141. L. Riehl, *Chemiefasern/Textilind.* **28** (1978) 636.

142. Chem. Werke Hüls, DE-AS 2 611 691, 1976 (H. J. Bax et al.).
143. T. Otaka et al., *Shokuhin Eiseigaku Zasshi* **19** (1978) no. 5, 431.
144. E. Nowak et al., *Tech. Pap. Reg. Tech. Conf. – Soc. Plast. Eng.* 1980, no. 5, 543.
145. Euwid Kunststoffdienst (1988) 4.
146. *Textil Prax. Int.* **45** (1990) June, 568.
147. H. Piechota, *Kunststoffe* **79** (1989) no. 10, 893.
148. H. M. Caesar, *Kunststoffe* **77** (1987) no. 10, 1004.
149. H. M. Caesar, F. J. M. M. Schiphorst, *Kunststoffe* **76** (1986) no. 10, 863.
150. *Mod. Plast. Internat.* **16** (1986) no. 1, 24; **17** (1987) no. 1, 20; **18** (1988) no. 1, 20; **19** (1989) no. 1, 26; **20** (1990) no. 1, 31.
151. *Kunststoffe* **79** (1989) no. 10, 896.
152. *Mod. Plast.* **4** (1990) 14.
153. *Chem. Mark. Rep.* (1989) Oct. 3, 35 (from data bank).
154. *High Technol.* (1989) June, 20 (from data bank).
155. R. A. Fleming, *Makromol. Chem. Macromol. Symp.* **57** (1992) 76.
156. J. van der Goorbergh, in *Recycling von Kunststoffen,* G. Menges, W. Michaeli, M. Bittner (eds.), Hanser Verlag, München 1992, p. 457.
157. D. Gintis, *Makromol. Chem. Macromol. Symp.* **57** (1992) 185.
158. G. Bauer in K. J. Thomè-Kozmiensky (ed.): *Recycling of Wastes,* EF-Press, Berlin 1989, p. 293.
159. H. Sinn, W. Kaminsky, J. Janning, *Angew. Chem. Int. Ed. Engl.* **15** (1976) 660.
160. W. Kaminsky, *J. Anal. Appl. Pyrolysis* **8** (1985) 439. W. Kaminsky, H. Sinn, C. Stiller, *Chem.-Ing.-Tech.* **57** (1985) 778.
161. W. Kaminsky, *Makromol. Chem. Macromol. Symp.* **48, 149** (1991) 381.

Polyether Rubber → Rubber, 3. Synthetic
Polyetherimides → Polymers, High-Temperature
Polyethers → Polyoxyalkylenes
Polyethers → Polyoxymethylenes
Polyethers → Poly(Phenylene Oxides)
Polyethylene → Polyolefins
Poly(Ethylene Glycols) → Polyoxyalkylenes
Polyethyleneimine → Aziridine
Poly(Ethylene Oxide) → Polyoxyalkylenes
Poly(Ethylene Terephthalate) → Polyesters
Polyformaldehyde → Polyoxymethylenes
Polyhydroxybenzenes → Phenol Derivatives

Polyimides

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes

Walter W. Wright, Surrey University, Guildford, United Kingdom (Chaps. 1 – 4)

Michael Hallden-Abberton, Rohm and Haas, Bristol, Pennsylvania 19907, United States (Chap. 5)

1. Introduction 1
2. Production 2
2.1. Condensation Polyimides 2
2.2. Thermoplastic Polyimides 4
2.3. Addition Polyimides 6
2.3.1. Norbornene End Groups 6
2.3.2. Maleimide End Groups 7
2.3.3. Acetylenic End Groups 8
2.3.4. Allylnadic End Groups 9
2.3.5. Benzocyclobutene End Groups 9
3. Processing 10
4. Properties and Uses 11
5. Polymethacrylimides 15
6. References 20

1. Introduction

The field of thermally stable heteroaromatic polymers is dominated by the polyimides. After several decades of research and development, the polyimides stand virtually alone as useful heat-resistant materials. They have been used as adhesives, fibers (→ High-Performance Fibers, Chap. 3.2.3.2.), films (→ Films, Chap. 6.11.), foams (→ Foamed Plastics, Chap. 4.9.), plastic moldings, resin matrices for composites, and wire-coating enamels. More recently their use as resin matrices has attracted the greatest attention [1] and synthetic activity.

The polyimides discussed in Chapters 2, 3 and 4 are synthesized from aromatic/heterocyclic monomers and can be loosely classified as condensation (Section 2.1), thermoplastic (Section 2.2), or addition (Section 2.3) types. Table 1 summarizes the trade names of these polyimides, their chemical structure, and classification. Strictly speaking, only those polyimides formed by addition reactions are thermosetting. The condensation type often behave in a similar way, when fully imidized many of them are almost completely insoluble and infusible.

Polymethacrylimides and their derivatives represent a special group of polyimides in which the imidization occurs after polymerization of unsaturated acyclic monomers. They are discussed in Chapter 5.

In 1988 the market for all types of polyimide was: United States ca. 4000 t, Japan ca. 850 t, and Western Europe ca. 800 t. All three markets were expected to grow annually by 10 – 20 % at least up to 1993.

The needs of the defense industry and the high-temperature requirements of the electronics industry have promoted the use of polyimide coatings, films, and laminates. Nevertheless molding resins dominate the market, amounting to 28 % of the total demand in Western Europe. Films are the next largest sector except in the United States where wire enamels take precedence. The films are used primarily in flexible printed circuit boards and for magnet wire insulation. Laminates are also used mainly in multilayer printed circuit boards.

2. Production

2.1. Condensation Polyimides

Most practical routes to polyimides up to the early 1970s involved condensation reactions [2]. Synthesis comprises low-temperature (ca. 50 °C) polycondensation of an aromatic dianhydride with a diamine in a polar solvent, (e.g., dimethylacetamide) to yield a soluble, fusible intermediate, i.e., polyamic acid [3–7]. Fibers, films, or impregnated fibrous reinforcements are

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_253

Table 1. Trade names, chemical structure, and classification of polyimides

Chemical structure	Trade name	Producer
Condensation		
BTDA (or BTDE)/MPD	FM-34/36	American Cyanamid (USA)
PMDA/DAPE	Kapton	Du Pont (USA)
	Toray	Du Pont (Japan)
BTDA (or BTDE)/MPD	Meldin-PI	Dixon (USA)
PMDA/DAPE	Pyre-ML	Du Pont (USA)
PMDA/DAPE	Pyralin	Du Pont (USA)
BTDA/DDM/propyltetramethyl disiloxane	SiPI	General Electric (USA)
BTDA (or BTDE)/MPD	Skybond	Monsanto (USA)
PMDA/BPTDA/DADMBP	Upilex	Ube Industries (Japan)
PMDA/DAPE	Vespel	Du Pont (USA)
Thermoplastic		
TMA/aromatic diamine	AI	Amoco (USA)
HFDE/aromatic diamine	Avimid/NR-150	Du Pont (USA)
polyesterimide	Isomid	Schenectady (USA)
TMA/aromatic diamine	Kerimid 500, Kermel	Rhône-Poulenc (France)
BTDA/3,3′-DABP	LARC-TPI	NASA-Langley (USA)
BTDA/mixed diisocyanates	PI-2080	Upjohn (USA)
TMA/aromatic diamine	Rhodeftal	Rhône-Poulenc (France)
polyesterimide	Terebec	Dr. Beck Co. (Germany)
TMA/aromatic diamine	Torlon	Amoco (USA)
polyetherimide	Ultem	General Electric (USA)
BTDA/aromatic cycloaliphatic diamine	XU-218	Ciba-Geigy (Switzerland)
Addition norbornene		
BTDA/NA/3,3′-DDM	LARC-13	NASA-Langley (USA)
BTDE/NE/AP 22	LARC-160	NASA-Langley (USA)
		US Polymeric (USA)
		Fiberite Corp. (USA)
		Hexcel (USA)
BTDA/NA/DDM	P 13 N	Ciba-Geigy (Switzerland)
HFDE/NE/PPD	PMR-11	NASA-Lewis (USA)
BTDE/NE/DDM	PMR-15	NASA-Lewis (USA)
		Fiberite Corp. (USA)
		Ferro (USA)
		Composites (USA)
		Hexcel (USA)
		BP (USA)
Addition bismaleimide		
Eutectic mixture of BMIs	Compimide 353	Technochemie (Germany)
BMI/diamine/monoamine	Compimide 795, 796	Technochemie (Germany)
BMI	F-178	Hexcel (USA)
Bismaleimidodiphenylmethane/DDM	Kerimid 601	Rhône-Poulenc (France)
BMI	Kinel 600	Rhône-Poulenc (France)
Bismaleimidodiphenylmethane/ diallylbisphenol A	Matrimid 5292	Ciba-Geigy (Switzerland)
Addition acetylenic		
Acetylene-terminated imide oligomers (USA)	Thermid 600	National Starch & Chemical Co.
Acetylene-terminated isoimide oligomers (USA)	Thermid IP 600	National Starch & Chemical Co.
Addition benzocyclobutene		
No commercial products yet available		

produced from this intermediate which is then converted to the fully cyclized polyimide, either by the action of heat (200 – 300 °C) [8–10], or chemically [11–14] using, for example, carboxylic acid anhydrides in the presence of tertiary amine catalysts as dehydrating agents (Fig. 1).

To obtain polyamic acid with a high molecular mass (necessary for films and fibers), strict attention must be paid to the purity, stoichiometry, and order of mixing of the reactants [8], [15]. High molecular mass is not a prerequisite for adhesives or for resin matrices for composites; here the main problems lie in the hydrolytic instability of the polyamic acid [8–10] and the volatile compounds (e.g., water, methanol, acetic acid) evolved during the cyclodehydration curing process which make processing more difficult. (The two stages of the reaction are covered in detail in [16]). The instability of the intermediate polyamic acids means that they must be stored

Table 1. (Continued)

Chemical structure	Trade name	Producer
Other trade names are Chem-Lon (Chem-Tronics, USA), LMB 1907 (Ciba-Geigy, Switzerland), Novac (Mitsubishi Chemical, Japan), and Solimide (International Harvester, USA).		
AP 22	=Jeffamine 22, $H_2NC_6H_4CH_2(C_6H_3NH_2CH_2)_nC_6H_4NH_2$	
BMI	=bismaleimide [*13676-54-5*]	
BPTDA	=biphenyl tetracarboxylic dianhydride	
BTDA	=benzophenone tetracarboxylic dianhydride [*2421-28-5*]	
BTDE	=diester of benzophenone tetracarboxylic acid	
3,3′-DABP	=3,3′-diaminobenzophenone	
DADMBP	=4,4′-diamino-3,3′-dimethyl biphenyl	
DAPE	=4,4′-diaminodiphenyl ether [*13174-32-8*]	
DDM	=4,4′-diaminodiphenylmethane [*101-77-9*]	
3,3′-DDM	=3,3′-diaminodiphenylmethane	
HFDE	=diester of 1,3-bis(3,4-carboxyphenyl)hexa-fluoropropene dianhydride	
MPD	=*m*-phenylenediamine [*108-45-2*]	
NA	=nadic anyhdride [*25134-21-8*]	
NE	=monoester of nadic acid	
PMDA	=pyromellitic dianhydride [*26265-89-4*]	
PPD	=*p*-phenylenediamine [*106-50-3*]	
TMA	=trimellitic anhydride [*552-30-7*]	

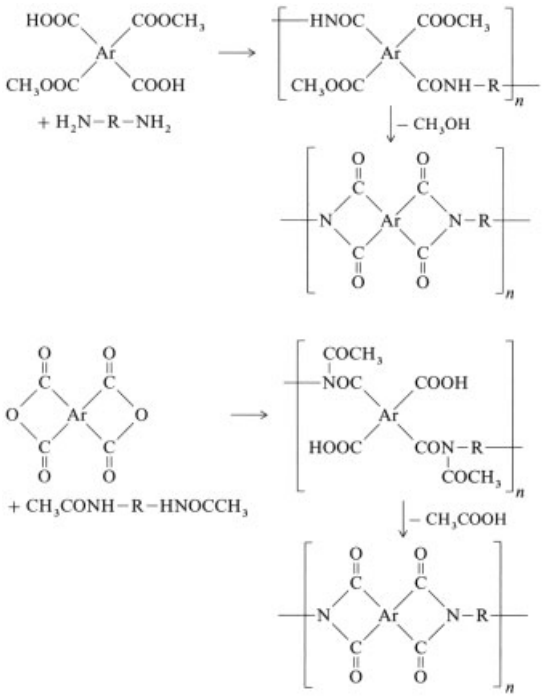

under rigorously dry, refrigerated conditions and even then shelf-life is strictly limited (several months). Water is evolved during imidization at elevated temperature leading to yet more hydrolysis. An additional difficulty, especially with the fabrication of thick samples, is that both softening and flow of the polyamic acid and cyclization to polyimide occur over similar temperature ranges. This necessitates careful control of processing conditions and frequently long, complex cure cycles.

Some of these problems have been alleviated by modifying the synthetic route illustrated in Figure 1. For example, use of semiesters of the anhydrides [17] or diacetylated diamines [18] means that an alcohol or acetic acid is eliminated during imidization rather than water (see figure on the left).

The intermediate polyamic acid formed from the diacetylated diamine is hydrolytically stable. The processing window between flow and cyclization is also increased. An additional advantage is that low-boiling, relatively nonpolar solvents (e.g., acetone) may be used for synthesis.

Volatiles are still produced during cyclization and this problem can only be overcome by synthesizing thermoplastic polyimides or variants which are cured by addition rather than condensation reactions.

2.2. Thermoplastic Polyimides

Thermoplastic polyimides are obtained by introducing large cyclic side groups into the polymer chain, by "tailoring" random sequences into the chain, or by the interposition of flexible, aliphatic linking groups between the aromatic and heterocyclic moieties. Russian workers have pursued the first approach and have produced "cardo" or "loop" polymers [19–21]:

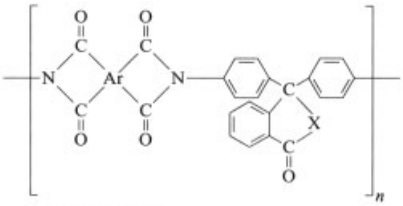

X = O, CO, NH

The solubility and fusibility of these polymers are governed both by the nature of the cyclic side group and the imide portion of the chain. If softening points (glass transition temperatures) are to remain below ca. 360 °C, then aliphatic units must also be introduced into the chain. One general drawback of thermoplastic polymers is that for use at elevated temperatures their softening points (glass transition temperatures) must be high (> 200 °C) and hence their processing temperatures must be even higher. This can be overcome by using intramolecular cyclization of active pendant groups which react at relatively low temperatures and increase the proportion of in-chain fused rings, thus raising the maximum use temperature [22], [23] (see top of right column).

Increased processibility may also be achieved by "tailoring" a random sequence into the polymer chain, the resultant dissymmetry accentuating the amorphous as opposed to the crystalline nature of the material [24], [25]:

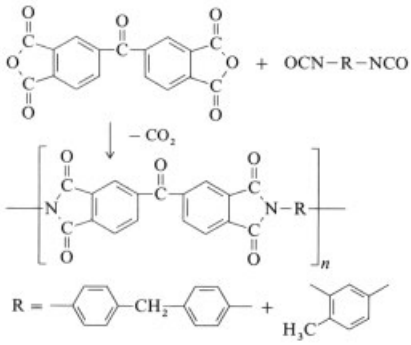

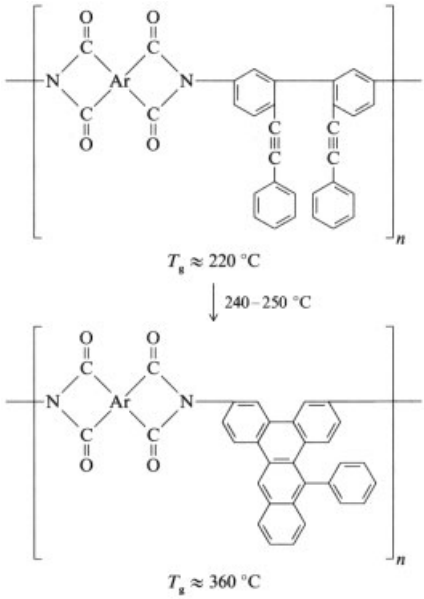

$T_g \approx 220$ °C

240–250 °C

$T_g \approx 360$ °C

A similar effect has been achieved by using 5[6]-amino-1(4-aminophenyl)-1,3,3-trimethylindane [26] or 3,3′-diaminobenzophenone [27] as the diamine component:

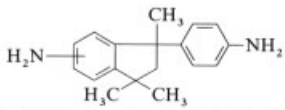

5[6]-Amino-1-(4-aminophenyl)-1,3,3-trimethylindane

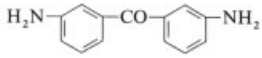

3,3′-Diaminobenzophenone

The incorporation of aliphatic linking groups normally lowers thermo-oxidative stability. An exception is the use of perfluoroalkylene groups [28], [29]:

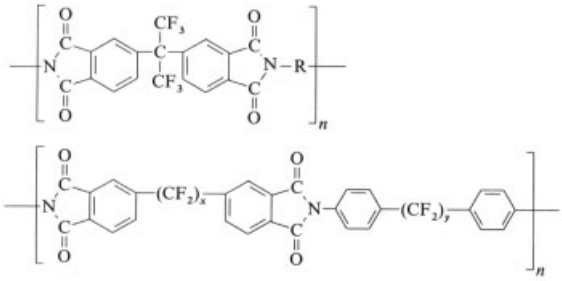

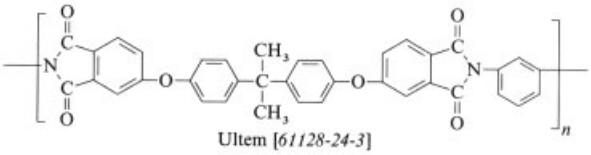

Ultem [*61128-24-3*]

Depending upon the values of x and y, polymers can be made with softening points from 150 °C upwards [30], [31]. They have been used as metal-to-metal adhesives and as laminating resins.

The most successful linking groups from the commercial point of view have been amide, ester, and ether. Amideimides have been prepared from precursors containing a preformed imide ring [32], but the most common route involves reaction of trimellitic anhydride with aromatic diamines or diisocyanates [33], [34]:

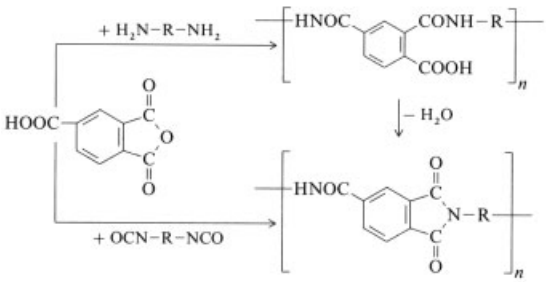

Esterimides, on the other hand, are normally made by reacting trimellitic anhydride with a diol, followed by reaction with a diamine [35]:

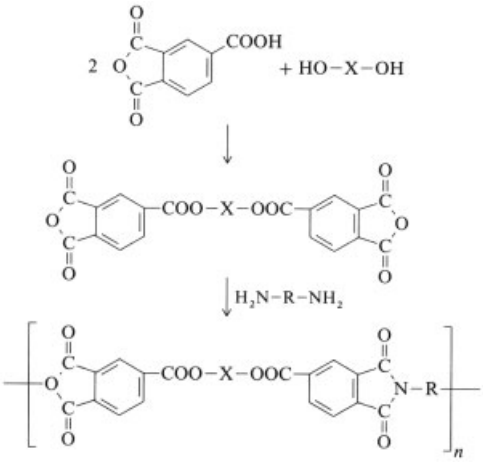

A polyetherimide (Ultem) was first marketed in 1982 and is produced by a nitro displacement polymerization process [36–39]. See → Polymers, High-Temperature, Chap. 2. (See figure on top of this page)

Although the amide, ester, and ether groups convey superior processing properties, they result in substantially lower thermal stability than that of the conventional aromatic – heterocyclic condensation polyimides. Considerable interest has therefore been centered on the addition polyimides where the aim is to achieve easier production whilst retaining heat resistance (Section 2.3).

2.3. Addition Polyimides

Addition polyimides usually consist of short-chain preimidized oligomers with reactive end groups that can react by addition polymerization. The end groups used have been norbornene (endomethylenetetrahydrophthalimide), maleimide, acetylenic (ethynyl), allylnadic, and benzocyclobutene.

2.3.1. Norbornene End Groups

The first example of this type of polyimide to become commerically available was based upon a low molecular mass (ca. 1300) amic acid prepolymer endcapped with reactive norbornene rings [40–43]:

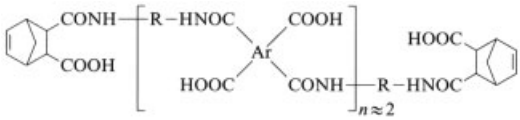

A solution of this prepolymer is used to impregnate fibrous reinforcement, the solvent (normally dimethylformamide) removed, and the polymer imidized by gradually heating from 65 to 204 °C:

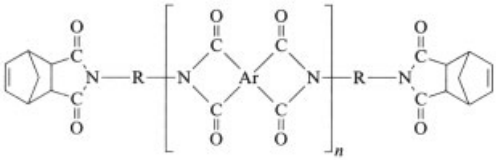

This polymer melts and flows above 260 °C and subsequently cross-links by "pyrolytic polymerization" of the unsaturated end groups at ca. 300 °C. Formerly this had been accepted to be a reverse Diels – Alder reaction with cyclopentadiene being released from the norbornene ring followed by its almost simultaneous polymerization with the maleimide ring also formed.

From a study of model compound reactions for the polymerization of monomeric reactants in situ (PMR) process an alternative mechanism has been suggested.

A variety of dianhydrides (Ar) and diamines (R) have been used to form the central unit of the chain, but the preferred reactants are mainly the same as those employed in condensation polyimides [44]. Although there is no elimination of volatiles due to imidization, considerable processing problems still arise from the high melt viscosity of the polymer, its short reaction time (which necessitates high heat buildup rates), and the difficulty of removing residual polar solvent. These difficulties have been overcome to a large extent by the PMR process. The norbornene chemistry involved is similar to that outlined above, but the semiesters of the anhydrides are normally used instead of the anhydrides and the impregnating solutions contain only monomeric reagents. All polymerization occurs in situ (e.g., on fibers). The principle of the PMR-15 system, which has received practical application, is shown in Figure 2 [45–48].

Other variants include the PMR-11 [49], [50], LARC-13 [51], and LARC-160 (Langley Research Center) [52] systems.

The PMR-15 system is most widely used as a resin matrix for composites. In common with many other high-temperature thermosets, it suffers from microcracking on thermal cycling. This tendency has been reduced by lowering the normal cure temperature (315 – 340 °C) to 280 – 290 °C [53]. In addition to lowering the thermal stresses induced in the material, a modified polymer network of reduced rigidity is also produced. The reactions occurring during thermal imidization of PMR-15 are very complex leading to prepolymers containing imide, amide, ester, anhydride, and salt functionalities [54]. Wide variations in prepolymer molecular mass and molecular mass distribution are also observed. The polymerization of *N*-phenylnadimide has been studied as a model for the cross-linking of PMR-15 [55].

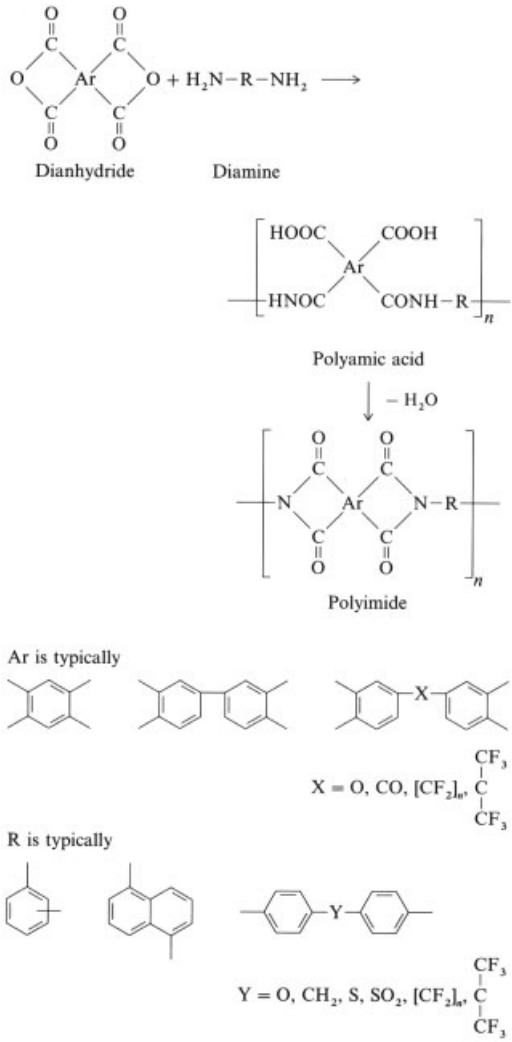

Figure 1. Synthesis of polyimides

2.3.2. Maleimide End Groups

The second type of addition polyimide is based on the bismaleimide structure, where cure oc-

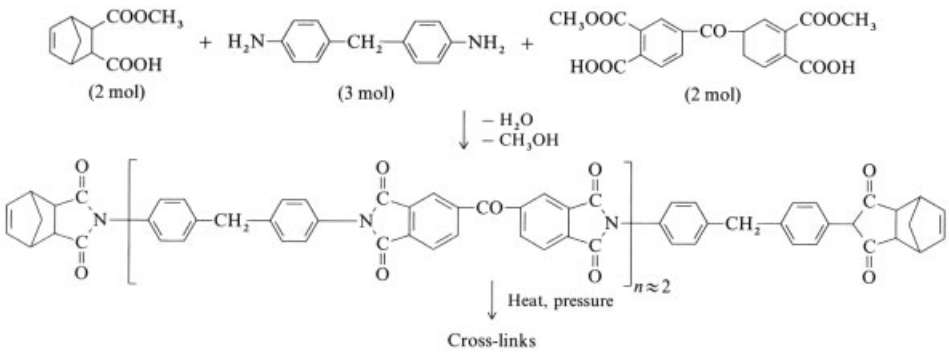

Figure 2. The principle of the PMR-15 system [45–48]

curs by a combination of addition reactions and Michael's reaction with amines [56–58]:

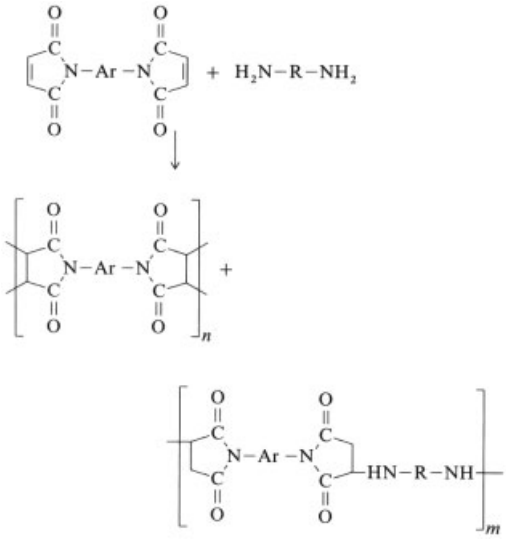

Use of a stoichiometric mixture of bismaleimide and diamine without an initiator results solely in linear chain extension and a polymer whose properties deteriorate rapidly at elevated temperatures. Free-radical cure, however, gives a very tightly cross-linked, brittle material. To obtain an optimum combination of properties, a careful balance must be maintained between the two reactions. This is achieved by using a judicious mixture of initiators and inhibitors [59]. Care must also be taken in the choice of solvent because the thermal stability of the product is very dependent upon residual solvent traces [60].

To obtain better processibility, toughness, and resistance to hot wet conditions than in bismaleimide homopolymers, attention has been devoted to blends of bismaleimides, either with themselves or with other comonomers. For example, the mixture of the three bismaleimides shown below forms a low melting point eutectic, which is specially adapted for filament winding applications [61]:

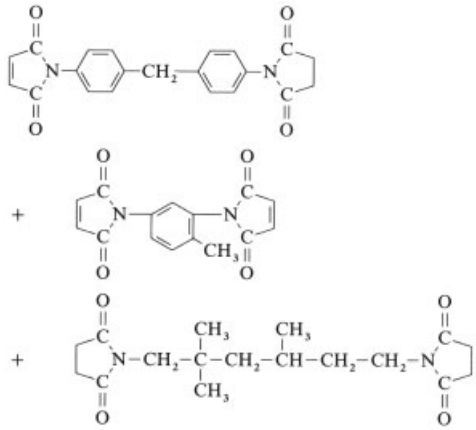

Other variants include blends of amide-linked maleimides and bismaleimides [62]:

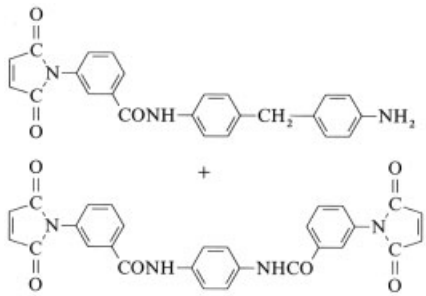

A number of comonomers have been developed as "curing agents" for bismaleimides.

Bis(propenylphenoxy)benzophenone is claimed to increase flexural strength and fracture toughness and to decrease moisture absorption [63]:

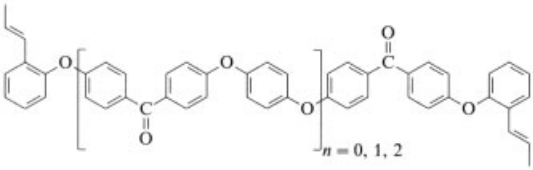

Cross-linking occurs between the maleimide and propenylphenyl groups.

Significant improvement in toughness is also achieved by the use of allylphenyl comonomers [64]. Benzocyclobutene-terminated imides can also be copolymerized with bismaleimides yielding other structures (see Section 2.3.5).

A drawback of introducing alicyclic structures into the polymer is that the thermal stability is reduced compared with a fully aromatic – heterocyclic structure. Use of acetylenic termination (Section 2.3.3) was initially hoped to overcome this.

2.3.3. Acetylenic End Groups

Relatively large imide prepolymers were synthesized with acetylenic end groups, which reacted thermally at 220 – 250 °C [65], [66].

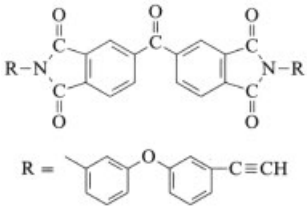

The hope was that the end groups would trimerize to yield phenylene rings and hence preserve heat resistance.

A complex mixture of products was obtained, however, which included a polyene structure containing 5 – 6 conjugated unsaturated bonds [67]. This has an adverse effect on thermo-oxidative stability. An acetylenic-terminated isoimide has also been synthesized [68]. The isoimide structure conveys better flow and solubility characteristics, but at elevated temperatures it rearranges to the imide.

Acetylenic-terminated polyimides have been synthesized from combinations of aromatic dianhydrides, aromatic diamines, and aromatic end-capping units [69], [70]. Polymers based upon 4,4′-(hexafluoroisopropylidene)bisphthalic anhydride are most attractive from the processing point of view since they are soluble in common solvents (e.g., acetone, tetrahydrofuran) [71].

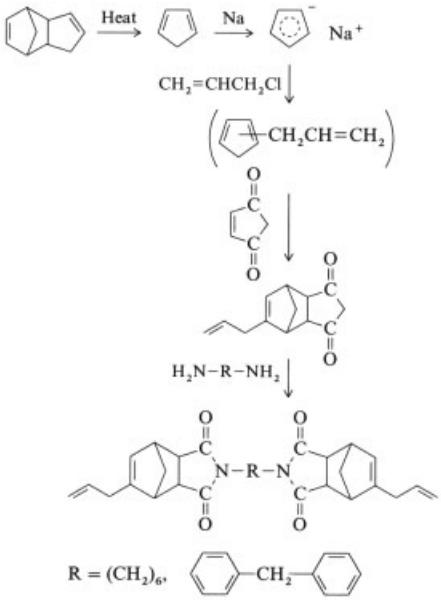

2.3.4. Allylnadic End Groups

A series of allylnadic imides, ranging from low-viscosity liquids to low-melting solids have been synthesized using dicyclopentadiene as starting material [72]. The dicyclopentadiene is thermally cracked and the resulting monocyclopentadiene is transformed to the desired products by a variety of routes.

These materials can be cured thermally and the cure can be accelerated by cationic additives (e.g., latent sulfonic acid). The cure mechanism probably involves a combination of cyclopolymerization and Diels – Alder rearrangements. The products have glass transition temperatures in excess of 300 °C.

2.3.5. Benzocyclobutene End Groups

Benzocyclobutene-terminated imides are easily synthesized by reaction of 4-aminobenzocyclobutene and aromatic tetracarboxylic acid dianhydrides:

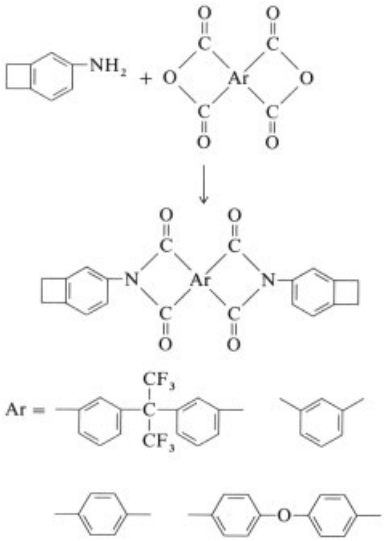

Under appropriate thermal conditions the strained four-membered ring of the benzocyclobutene opens to yield an extremely reactive diene, bis(*o*-quinodimethane), which then dimerizes:

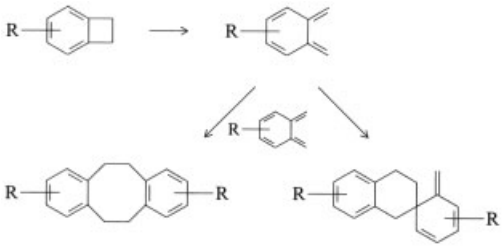

Another possible reaction is 1,4-addition analogous to the polymerization of *o*-xylylene [73], [74]:

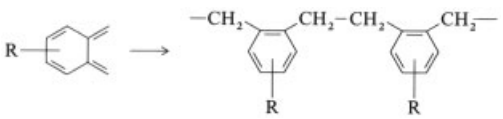

A benzocyclobutene-terminated imide can also be copolymerized with a conventional bismaleimide via a Diels – Alder reaction between the maleimide and *o*-quinodimethane groups [75]:

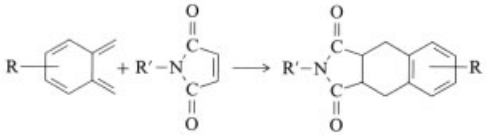

Little data on these systems are available to date.

3. Processing

Adhesives. Condensation, thermoplastic, and addition polyimides and polyamideimides have all been evaluated as metal-to-metal adhesives [76]. In the case of the condensation polymers, a solution of the intermediate polyamic acid is applied to a glass cloth carrier, and the bulk of the solvent removed by heating to 100 – 150 °C. The partially cured material is then placed between the adherends and cure is completed under pressure at high temperature (e.g., 0.5 MPa at 316 °C for 90 min). A postcure of some hours at 300 – 350 °C is necessary for optimum properties [77].

Fibers. See also, → High-Performance Fibers, Chap. 3.2.3.2. Fibers have been produced by wet spinning of thermoplastic polyimides [78] (e.g., Upjohns 2080 [79] or Upilex [80]). The alternative approach is to form polyamic acid fibers from solutions of prepolymer and to submit these to chemical, thermal, and mechanical treatment to produce the polyimide fibers. Russian authors have been especially active in this field and the main features of the two-step manufacturing process are described in [81]. Conventional wet and dry spinning equipment may be used.

Films. See also, → Films, Chap. 6.11.. The polyimide derived from pyromellitic dianhydride and 4,4′-diaminodiphenyl ether has dominated film production [2]. Films are cast from solutions of the intermediate polyamic acid and then converted to polyimide by heating at 300 °C. Commercial polyimide films are also available which are coated on one or both sides with fluorinated ethylene resin. That provides a

heat-sealable surface and also improves resistance to bases and concentrated acids.

Foams. See also, → Foamed Plastics, Chap. 4.9. Foams have mainly been produced from condensation polyimides. The powder obtained after spray drying the polyamic acid solution is foamed into blocks by heating in a microwave oven. The water and alcohol liberated during imidization act as the blowing agents. Syntactic foams have also been manufactured, for example, by mixing equal weights of polyimide and silica microballoons and processing at 300 °C under moderate pressure.

Plastic Moldings. Processing conditions depend on polyimide type. Some condensation polymers are subjected to powder metallurgy techniques to cold-form a part followed by sintering at 360 – 380 °C [82]. Direct or transfer compression molding at temperatures up to 350 °C and pressures up to 30 MPa has been used for thermoplastic and addition-type polyimides and for polyamideimides. Injection molding is carried out by placing the polymer in either pellet or powder form into the heated cavity (350 °C) of an injection molding machine, allowing softening to occur, and then injecting into a cooled cavity (230 °C), where the part remains until it is below its glass transition temperature. For high-quality moldings, prolonged annealing cycles are recommended, e.g., for a 12-mm thick article: 36 h at 150 °C, 36 h at 177 °C, 36 h at 204 °C, 36 h at 232 °C, and 48 h at 260 °C. Moldings are produced with a variety of fillers (e.g., short carbon or glass fibers, graphite, molybdenum disulfide, polytetrafluoroethylene).

Resin Matrices for Composites [1]. See also, → Composite Materials, Chap. 4.2. The processing of carbon fiber prepreg to produce laminates depends upon the type of polyimide used for impregnation, the equipment (autoclave or press), and the shape of the item and its thickness. Data for press cure of the different types of polyimide are summarized in Table 2. These are taken, wherever possible, from schedules recommended by the resin manufacturers. Some polyimides are not suitable for autoclave curing. The thermoplastic polyimides require the highest press temperatures, but no postcure cycle is necessary. The condensation and addition polyimides are fabricated at lower temperatures and, in general, at lower pressures, but require extended postcure cycles. For the condensation and acetylenic end-capped polymers a complicated stepwise postcure is essential for the development of optimum properties.

Wire-Coating Enamels. Polyesterimides find their major use in this area. Wire-coating formulations based upon polyamideimides and condensation polyimides are also available. The coated wires are produced with conventional equipment. The varnishes are deposited on the surface from solution, dried to remove solvent, and then cured in the case of the condensation polyimide by heating to ≥ 350 °C. For the other materials cure temperatures of ≥ 175 °C are required.

4. Properties and Uses

The properties of the different polyimide types in their diverse applications cannot all be listed here. Since the main justification for their use is their high temperature capability, this aspect of their behavior will be emphasized. Figure 3 compares the thermal stabilities of polyimides on the basis of weight loss [83]. The superior stability of the fully aromatic – heterocyclic polymers (e) compared with the addition polymers containing alicyclic rings (a) and (c) is evident. The relatively low thermo-oxidative stability of the polymer derived from the oligomers with acetylenic end groups (b) is a function of the polyene structure which results from polymerization of these materials [67].

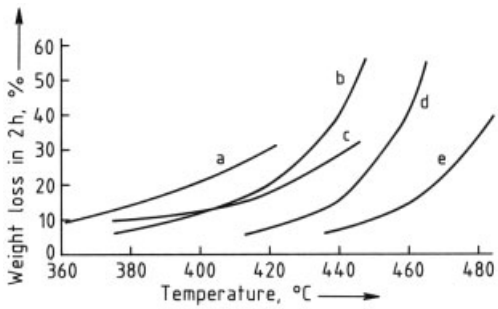

Figure 3. Comparison of thermal stability in air of different types of polyimide
a) Bismaleimide; b) Acetylenic; c) Norbornene; d) Thermoplastic; e) Condensation

Table 2. Comparison of press cure parameters

Polyimide type	Temperature, °C	Pressure, MPa	Time, min	Postcure	
				Temperature, °C	Time, h
Bismaleimide	120 – 180	1.4	60	250	24
Condensation	205	1.7	30, cooling	205 – 345	16
Acetylenic	205	1.4	60 – 120, cooling	230 – 400	22
PMR	232 – 315	6.9	90, cooling	343	16
Norbornene	315	1.4	60	not necessary	
Thermoplastic 2	250 – 350	2.0 – 3.5	60, cooling	none	
Thermoplastic 1	427	17.3	10, cooling	none	

Adhesives. Most commercial polyimide adhesives are supported on glass cloth carriers and contain antioxidants, generally arsenic compounds, to prolong service life at elevated temperatures. Aluminum powder is also often added to reduce thermal stress and to act as a reinforcing agent. Table 3 gives data for lap-shear joints (overlap of two materials bonded with adhe-sives) tested at elevated temperatures and made from condensation, thermoplastic, and addition polymers. The lap-shear strength is a standard test for adhesives and is a measure of the strength required to pull the joint apart. Attempts have been made to toughen the brittle LARC-13 adhesive matrix by the incorporation of elastomers [84]. Fluorosilicone or vinyl-terminated silicone elastomers were physically blended with the LARC-13 amic acid; alternatively, aromatic amine-terminated butadiene – acrylonitrile (ATBN) or silicone (ATS) elastomers were chemically reacted into the prepolymer backbone. The effect of these additions on elevated temperature properties is, however, generally deleterious.

The main applications of these heat-resistant adhesives are in the aerospace and electrical industry.

Fibers. Fibers have been made on a pilot-plant scale in the United States and the former Soviet Union from both condensation and thermoplastic polyimides. Figure 4 illustrates strength retention as a function of time and temperature in wet-spun fibers [79]. Table 4 gives heat aging data on a number of experimental polypyromellitimide fibers [16]. The fibers may be used for heat-resistant insulating materials, filter cloths, flameproof clothing, reinforcement of rubber articles, electrical insulation, and atomic reactors. The last of these uses is possible because of the high stability of polyimides to ionizing radiation.

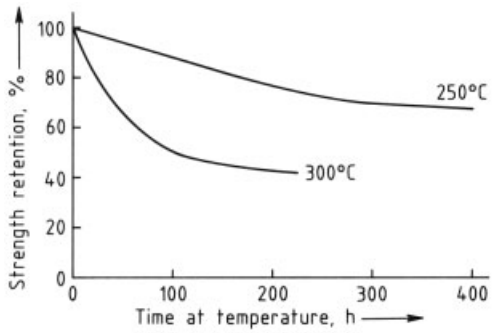

Figure 4. Strength retention of wet-spun polyimide fibers (Upjohn Polyimide 2080, thermoplastic type) as a function of time at 250 and 300 °C

Films. The first major application of a condensation polyimide was as a film – Du Pont's Kapton has been marketed since 1961. Table 5 gives typical properties of 0.0254-mm (0.001 inch) Kapton H film and also of Kapton H coated on one or both sides with fluorinated ethylene – propene resin. The coating provides heat-sealing capability and improves chemical resistance. Films have also been produced from the thermoplastic polyimide based upon diaminodiphenylindane [85], some mechanical properties at various temperatures are listed in Table 6.

The main area of use of polyimide films is electrical insulation. The aerospace industry employs them for wire, cable, slot, and winding insulation. They have also been specified for electric locomotive motors, subway trains, and motors of immersed petroleum pumps. Other applications include flexible printed circuits, magnetic tapes, and solar sails.

Table 3. Heat aging data for polyimide-bonded lap-shear joints

Adhesive type	Adherend	Aging conditions	Test temperature, °C	Lap-shear strength, MPa
Condensation (FM-34)	stainless steel	as made	25	29
			260	16
		192 h at 316 °C	25	22
			260	17
		500 h at 316 °C	25	18
			260	15
		1000 h at 316 °C	25	14
			260	12
	titanium	as made	25	26
			260	13
		192 h at 316 °C	25	12
			260	15
		500 h at 316 °C	25	14
			260	16
		1000 h at 316 °C	25	15
			260	15
		2000 h at 316 °C	25	7
			260	9
Thermoplastic (NR-150 A 2 G)	titanium	as made	23	34
			260	21
Thermoplastic (amideimide AI-1030)	stainless steel	as made	RT	25
		500 h at 260 °C	RT	19
	titanium	as made	RT	28
		500 h at 260 °C	RT	15
Addition (norbornene LARC-13)	titanium	as made	RT	20
			232	19
			288	9
		500 h at 232 °C	RT	20
			232	20
			288	15
Addition (bismaleimide Kerimid 500)	stainless steel	as made	RT	25
		1000 h at 250 °C	RT	17
		2500 h at 250 °C	RT	10
		4000 h at 250 °C	RT	10
		100 h at 300 °C	RT	10

∗ RT=Room temperature.

Foams. Conventional and syntactic foams have both been produced from condensation polyimides. Table 7 gives the compressive strengths at room and elevated temperatures of foams of differing density [86]. Figure 5 compares the temperature dependence of compressive strength of conventional and syntactic polyimide foams [87]. These foams have very low flame spread and smoke production. Their main applications are in acoustic and thermal insulation, e.g., vibration and heat protection for electronic equipment, light-weight fire barriers, and cryogenic insulation for fuel tanks. They have been employed in aircraft and rocket construction.

Plastic Moldings. Tables 8 and 9 list mechanical properties of unfilled and graphite-filled polyimide moldings, respectively, at elevated temperatures. Comparison of results is not simple because the moldings contain various amounts of filler and testing was done at different temperatures. Absolute figures may also depend upon whether testing is done perpendicular or parallel to the direction of molding. The former gives higher values (except for compressive strength) and hence only these are quoted.

The moldings have many applications (e.g., seals, gaskets, piston rings, pressure disks, sleeves, bearings, sliding and guide rods, valve shafts, and gear wheels). They are used in nuclear equipment, jet engines, data processing

Table 4. Heat aging data for polyimide fibers

Fiber type	Aging conditions	Test temperature, °C	Tensile strength, cN/tex
Arimide PM [a]	as made	100	43
	as made	200	33
	as made	300	26
	as made	400	19
	100 h at 300 °C	20	28 – 30
	216 h at 300 °C	20	22 – 25
	35 h at 400 °C	20	degraded
Arimide T [a]	as made	100	45
	as made	200	30
	as made	300	25
	as made	400	20
	100 h at 300 °C	20	36 – 48
	216 h at 300 °C	20	32 – 42
	25 h at 400 °C	20	18 – 25
PRD-14 [b]	as made	100	50
	as made	200	37
	as made	300	25
	as made	400	19
	100 h at 300 °C	20	30 – 33
	216 h at 300 °C	20	24 – 25
	25 h at 400 °C	20	degraded

[a] Produced in former Soviet Union.
[b] Produced by Du Pont (USA).

Table 5. Elevated temperature data for a condensation polyimide film (0.001 inch)

Property	Test temperature, °C	Kapton H	Kapton H coated with 0.0005 inch of fluorinated ethylene – propene resin	
			Coated one side	Coated both sides
Ultimate tensile strength, MPa	25	173	117	97
	200	117	76	62
Yield point at 3 %, MPa	25	69	50	41
	200	41	28	21
Stress to produce 5 % elongation, MPa	25	90	62	52
	200	59	38	58
Ultimate elongation, %	25	70	75	75
	200	90	85	95
Tensile modulus, GPa	25	2.97	2.21	1.73
	200	1.79	1.19	0.90
Dielectric constant at 1 kHz	25	3.5		
	200	3.0		
Dissipation factor at 1 kHz	25	0.003		
	200	0.002		
Volume resistivity, $\Omega \cdot$ cm	25	1×10^{18}		
	200	1×10^{14}		
Short-term dielectric strength at 60 Hz, V/mm	25	7000		
	200	5600		

equipment, soldering and welding equipment, brakes, self-lubricating bearings, rotary pumps, refrigeration plants, and also for printed circuit boards.

Table 6. Mechanical properties of athermoplastic diaminodiphenylindane-based polyimide film at elevated temperatures

Property	Value	Test temperature, °C
Ultimate tensile strength, MPa	77	25
	36	204
	27	260
Ultimate elongation, %	10	25
	39	204
	55	260
Tensile modulus, GPa	2.48	25
	3.38	204
	3.11	260
Time to brittleness in air, h	> 2000	200
	925	225
	250	250

Resin Matrices for Composites. This area has seen the marketing of the greatest number and variety of polyimides. A vast amount of data is available, especially on carbon-fiber-reinforced materials. Absolute strength values depend very much on the type of carbon fiber used for reinforcement. Tables 10 and 11 compare carbon fiber – polyimide resin unidirectional laminates based upon percentage retentions of interlaminar shear strength, flexural strength, and flexural modulus with time of aging at different temperatures. Table 10 also includes average figures for the initial room temperature values of these mechanical properties for the polyimides and carbon fiber. The data show that at least 80 % of the initial properties are retained at 200 °C for up to 1500 h. At higher temperatures the lifetimes of composites based upon 50 % retention of flexural strength are 1820 h at 232 °C, 1350 h at 300 °C and 470 h at 316 °C. These figures must be regarded as maximum values because, in service, the effects of absorbed moisture and applied stress must be considered in addition to thermo-oxidative aging.

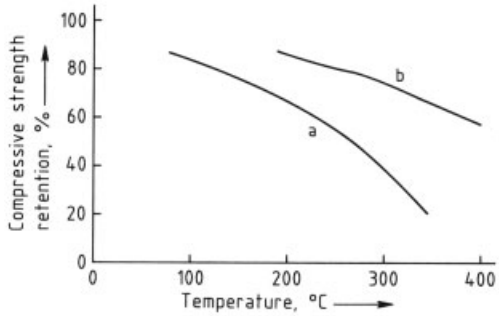

Figure 5. Compressive strength retention of polyimide foams as a function of temperature
a) Blown foam; b) Syntactic foam

The major application of these composites is as components of primary and secondary structures in the aerospace industry. They are also used in the electrical and electronics industry (e.g., for printed circuit boards).

Wire-Coating Enamels. Insulating coatings for electrical wiring have been made from polyimides (condensation type), polyamideimides, and polyesterimides. The polyimide coating has the best thermal stability with projected lifetimes of 100 000 h (>11 years) at 200 °C, 25 000 h (>2 years) at 240 °C, 10 000 h at 260 °C and 1000 h at 325 °C. Its drawbacks are low abrasion resistance, limited shelf-life of the coating solutions, and cost. The polyamideimides and polyesterimides overcome some of these drawbacks but at the expense of heat resistance. The polyimide coating also exhibits high elasticity at very low temperatures. A 0.625-mm diameter wire can be wrapped around a 3-mm diameter mandrel at 4 K without the coating cracking. With a polyester coating, cracking would occur on wrapping around an 18-mm diameter bar. Figure 6 compares the service lives of polyimide, polyamideimide, and polyester coatings [88].

The main applications of the coatings are in electrical equipment and machines that are operated at elevated temperatures.

5. Polymethacrylimides

Production. Commercial production of polymethacrylimides has been known since 1939 [89] and generally involved the use of

Table 7. Compressive strengths (pressure required for 10 % compression) of polyimide foams at room and elevated temperatures

Property	Temperature[a], °C	Foam designation[b] RI-7271-06	RI-7271-12	RI-7271-18
Density, kg/m^3	RT	107	218	328
Pressure, kPa	RT	145	932	4209
	93	138	828	3 726
	204	138	690	3312
	232	83	621	1656
	288	69	242	1518
	316	14	14	69

[a] RT=room temperature.
[b] Made by Monsanto.

Table 8. Mechanical properties of unfilled polyimide moldings at elevated temperatures

Property	Test temperature, °C	Polyimide type: Condensation (Vespel SP-1)	Thermoplastic (PI 2080)	Norbornene (P 13 N)	Amideimide (Torlon 4203) *
Tensile strength, MPa	23	100	118	50	186
	150				105
	260			39	52
	288		30		
	300	43			
Tensile modulus, GPa	23	3.35	2.07	3.84	
	260			2.39	
	288		0.97		
	300	2.07			
Elongation at break, %	23	8	10	1.4	12
	150				17
	260			1.8	22
	300	3			
Flexural strength, MPa	23	128	193	69 – 83	212
	150				156
	260			41 – 55	75
	288		35	35 – 41	
	300	72			
Flexural modulus, GPa	23	3.17	3.45	3.17 – 3.38	4.55
	150				3.59
	260			2.07 – 2.28	2.97
	288		1.11	1.93 – 207	
	300	1.90			
Compressive strength, MPa	23	253 **	206	258	276
	300	128 **			
Compressive modulus, GPa	23	3.93	2.04	2.89	2.84
	300	1.90			

* Contains 3 % TiO_2.
** Specimens did not fail, value is for 50 % strain.

a preformed polymeric resin, usually a methacrylic ester such as poly(methyl methacrylate) [*9011-14-7*], or a polymeric acid such as poly(methacrylic acid) [*79-41-4*], and an ammonia or amine source such as a free amine, ammonium salt, urea, or nitrile. These reactions, usually carried out in unpressurized vessels at about 140 – 210 °C for several hours, gave fairly high imidization levels, but the polymers contained significant amounts of methacrylic acid groups and associated ammoni-um or alkyl ammonium salts. This type of material was generally not suitable for thermoplastic uses because of the

Table 9. Mechanical properties of graphite-filled polyimide moldings at elevated temperatures

Property	Test temperature, °C	Polyimide type			
		Condensation (Vespel SP-21) [a]	Thermoplastic (PI 2080) [a]	Maleimide (Kinel 5505) [b]	Amideimide (Torlon 4301) [c]
Tensile strength, MPa	23	79	72	39	135
	150				73
	250			29	
	260				46
	300	41			
Tensile strength, GPa	23	4.49	1.84		
	300	3.11			
Elongation at break, %	23	4.5	7		6
	300	3.0			
Flexural strength, MPa	23	124	95	88	135
	150				73
	200			69	
	250			54	
	260				46
	288		28		
	300	68			
Flexural modulus, GPa	23	3.93	3.63	5.17	6.35
	150				5.04
	200			4.69	
	250			4.42	
	260				4.00
	288		1.42		
	300	2.48			
Compressive strength, MPa					
	23	200	144	153	207
	200			101	
	300	64			
Compressive modulus, GPa	33	4.35	1.87		4.69
	300	2.07			

[a] Contains 15 % graphite powder.
[b] Contains 25 % graphite powder.
[c] Contains 12 % graphite powder and 3 % polytetrafluoroethylene.

Table 10. Initial properties of carbon fiber – polyimide resin unidirectional laminates

Polyimide type	Carbon fiber type	ILSS ∗, MPa	Flexural strength, MPa	Flexural modulus, GPa
Condensation	high modulus	49	710	163
	high strength	72	1340	115
	type A	66	1420	108
Thermoplastic	high modulus	51	870	145
	high strength	61	1240	108
	type A	88	1390	104
Norbornene	high modulus	52		
	high strength	78	1430	100
Bismaleimide	high strength	85	1410	120
Acetylenic	high strength	83	1350	104
	type A	116	1190	108

∗ Interlaminar shear strength.

Table 11. Heat aging data for carbon fiber – polyimide resin unidirectional laminates (for initial properties, see Table)

Polyimide type	Aging time, h	Aging temperature, °C	Test temperature, °C	Retention of initial, value, %		
				ILSS, MPa	Flexural strength, MPa	Flexural modulus, GPa
Condensation	1000	200	25	81	100	99
	1000	300	25	61	76	82
	400	316	25		86	
	1500	200	200		89	
	2000	232	232	61	45	83
	1800	300	300		42	
	400	316	316		55	
Thermoplastic	5000	260	260	84	76	86
	1850	288	288	80	72	100
	3000	316	316	80	66	90
	500	343	343	78	60	81
Norbornene	2000	260	260	52		
	2000	288	288	21		
	500	316	316	41		
Bismaleimide	500	232	25	92	83	
	500	232	232	35	30	
Acetylenic	1000	316	25		45	
	125	316	316	100	83	90
	1000	316	316		56	

evolution of volatiles (methanol, water, amines) at thermoplastic processing temperatures.

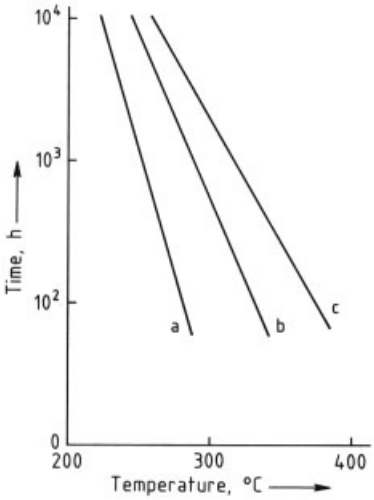

Figure 6. Comparison of service lives of wire-coating enamels
a) Polyesters; b) Polyamideimide; c) Polyimide

A later process [90] using a pressurized autoclave at higher temperatures and pressures for longer times, allowed a satisfactory degree of imidization (> 95 %) in the recovered (insoluble phase) polymer, so that the polymer was better suited for thermoplastic uses.

However, polymethacrylimides found their main commercial use as foamed polymers, using a variation of the above chemistry in which the nitrogen source was copolymerized into the polymer backbone, and pyrolytic cyclization – imidization reactions were carried out (→ Foamed Plastics, Chap. 4.10.) [91–94].

Other early methods for forming polymeric (meth)acrylimides included cyclopolymerization of imide-containing monomers [95–97], and treatment of methacrylic films with ammonia or amine [98], [99].

More recently, a continuous high-pressure reactive extrusion – devolatilization process has been developed [100]. A (meth)acrylate polymer is treated with substantially anhydrous ammonia or amine at high pressure and temperature in an extruder, followed by subatmospheric devolatilization. The advantage of this process is that nearly all cyclization reactions (imide and anhydride formation) are complete within the residence time of the extrusion, and residual volatiles levels are extremely low. This leads to a widely processible thermoplastic, which is suitable for injection or blow molding, thermoforming, etc.

Due to the fairly unusual chemistry that takes place in reactions of amines and polymethacrylates at high temperatures, formation of methacrylic acid, acid salts, and methacrylic anhydride is observed (Fig. 7). This is believed to occur via an unusual alkyl – oxygen cleavage mechanism. The consequent alkylation of amines required by this pathway has been observed [101–103], and a similar mechanism using dialkylamines and polymethacrylates to generate methyl methacrylate/methacrylic anhydride polymers has also been reported [104].

These thermoplastic polymethacrylimides can be improved further by a two-step continuous imidization – re-esterification process [17–19] in which the level of methacrylic acid (**1**) and anhydride (**2**) functionality in the final imide copolymer can be decreased or eliminated. This, in turn, substantially increases processibility, weatherability, flow, and miscibility of the imide copolymer with other polymers. Other imidization process modifications have also been reported [108–130].

Properties and Uses. The imidization of poly(methyl methacrylate) by most of the above processes can be accomplished with ammonia or primary alkyl amines. Imidization with ammonia gives rise to a mixed NH/NCH_3 polyimide functionality due to the incorporation of some of the alkylated amine into the polymer [130]. At high imidization levels (ca. 90 – 100 %), these materials, when prepared by the reactive extrusion process, have high glass transition temperatures (215 – 225 °C), good tensile strength (96 – 110 MPa, ASTM D-638) and modulus (4130 – 5170 MPa), high decomposition temperatures (1 %, thermogravimetric analysis, N_2, 390 – 405 °C), good oxidative stability, chemical resistance, and impact resistance similar to acrylic polymers. When methylamine is used as the imidizing agent, materials with a slightly lower glass transition temperature (ca. 168 – 180 °C at 90 – 100 % imidization) are obtained but the polymer is lower in color, better in weatherability, and more easily processed. A commercial grade of thermoplastic poly(*N*-methyl methacrylimide (KAMAX T-170, Rohm and Haas, USA) of ca. 90 wt % imidization and $\bar{M}_W$ 125 000 has the following properties [131]:

Refractive index	1.54
Density	1.22 g/cm^3
Coefficient of thermal expansion (23 °C)	4.7×10^{-5} K^{-1}
Tensile strength (23 °C)	79 MPa
Flexural strength (23 °C)	138 MPa
Flexural modulus (23 °C)	4270 MPa
Notched Izod impact strength (23 °C)	21 J/M
Heat distortion temperature under flexural load at 1.82 MPa, annealed at 130 °C for 4 h	156 °C
Glass transition temperature	168 °C
Total light transmission/haze (ASTM D 1003-61)	90 %/2 %

Because of their high modulus, tensile strength, heat distortion resistance, and optical properties, poly(*N*-methylmethacrylimides) are finding broad use in glazing and lighting applications. Automotive uses include taillight and headlight covers and bezels, where high hardness, surface gloss, optical transmittance, retroreflectivity, heat resistance, and weatherability are required.

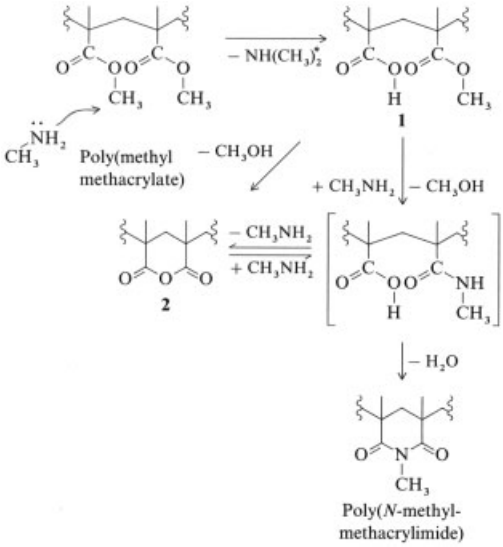

Figure 7. Reaction of poly(methyl methacrylate) with methylamine
∗ Dimethylamine formed in this reaction also participates in formation of methacrylic acid from methyl methacrylate (with generation of trimethylamine).

In siloxane-coated formulations, the high dimensional stability, rigidity, and low weight of these polymers make them useful for large glazing parts such as "targa" tops and "T" top automotive roofing units. High-wattage light piping and light guides are another use, where the superior optical clarity is an advantage over polycarbonate. Instrument clusters are also a growing

area of use. An area under development is high-intensity discharge (HID) lighting applications where very high heat and UV resistance over a long time period are needed.

Other uses of polymethacrylimides include photoresist resins [99], [132], packaging, medical devices, food and cosmetic containers, gas barriers, adhesives, and in polymer blends [133–138].

Polymethacrylimides containing some residual acid and anhydride functionality are useful as reactive compatibilizers and processing aids. In polyamide blends, for instance, polymer grafting reactions occur to give block copolymers [139].

6. References

1. J. P. Critchley, W. W. Wright, *Rev. High Temp. Mater.* **4** (1979) 107 – 138.
2. C. E. Sroog, *J. Polymer Sci. Macromol. Rev.* **11** (1976) 161 – 208.
3. Du Pont, CA 659 328, 1963 (A. L. Endrey).
4. Du Pont, US 3 179 361, 1965 (A. L. Endrey).
5. Du Pont, US 3 179 633, 1965 (A. L. Endrey).
6. Du Pont, US 3 179 614, 1965 (W. M. Edwards).
7. Du Pont, US 3 179 634, 1965 (W. M. Edwards).
8. G. M. Bower, L. W. Frost, *J. Polym. Sci.* **A 1** (1963) 3135.
9. L. W. Frost, I. Kesse, *J. Appl. Polym. Sci.* **8** (1964) 1039.
10. C. E. Sroog et al., *J. Polym. Sci.* **A 3** (1965) 1373.
11. Du Pont, US 3 179 630, 1965 (A. L. Endrey).
12. Du Pont, US 3 342 774, 1967 (E. R. Hoegger).
13. Du Pont, US 3 422 064, 1969 (W. G. Gall).
14. S. V. Vinogradova et al., *Polym. Sci. USSR* **16** (1974) 584.
15. R. A. Dine Hart, W. W. Wright, *J. Appl. Polym. Sci.* **11** (1967) 609.
16. M. I. Bessonov, M. M. Koton, V. V. Kudryavtsev, L. A. Laius: *Polyimides. Thermally Stable Polymers,* Consultants Bureau, New York 1987.
17. Du Pont, FR 1 360 488, 1963 (E. Lavin, A. H. Markhart, R. E. Cass).
18. ICI, US 3 832 330, 1974 (D. R. Dixon, J. B. Rose, C. N. Turton).
19. V. V. Korshak et al., *Izv. Akad. Nauk SSSR Ser. Khim.* **10** (1967) 2267.
20. S. V. Vinogradova et al., *Vysokomol Soedin.* **A 11** (1969) 2275.
21. S. V. Vinogradova, V. S. Vygodskii, *Russian Chem. Rev. (Engl. Transl.)* **42** (1973) 551.
22. F. L. Hedberg, F. E. Arnold, *J. Polym. Sci. Polym. Chem. Ed.* **14** (1976) 2607.
23. F. L. Hedberg, F. E. Arnold, AFML TR-76-198, 1977.
24. P. S. Carleton, W. J. Farrissey, J. S. Rose, *J. Appl. Polym. Sci.* **14** (1970) 1093.
25. W. J. Farrissey, L. M. Alberino, A. A. R. Sayigh, *J. Elastomers Plast.* **7** (1975) 285.
26. J. H. Bateman, W. Geresy, D. S. Neiditch, *ACS Org. Coat. Plast. Chem.* **35** (1975) no. 2, 77.
27. A. K. St. Clair, T. L. St. Clair, *28th Nat. SAMPE Symp.* **28** (1981) 165.
28. H. H. Gibbs, *7th Nat. SAMPE Tech. Conf.* **7** (1975) 244.
29. H. H. Gibbs, *21st Nat. SAMPE Symp.* **21** (1976) 592.
30. J. P. Critchley, P. A. Grattan, M. A. White, J. S. Pippet, *J. Polym. Sci.* **A 1** (1972) 1789.
31. J. P. Critchley, M. A. White, *J. Polym. Sci.* **A 1** (1972) 1809.
32. F. Hayano, K. Komoto, *J. Polym. Sci.* **A 1, 10** (1972) 1263.
33. Westinghouse, US 3 179 635, 1965 (L. W. Frost, G. M. Bower).
34. M. W. Alvino, L. W. Frost, *J. Polym. Sci.* **A 1** (1971) 2209.
35. R. E. Van Strian, J. R. Erzner, US 3 377 321, 1968.
36. F. J. Williams, P. E. Donahue, *J. Org. Chem.* **42** (1977) 2414.
37. F. J. Williams, H. M. Relles, P. E. Donahue, J. S. Manello, *J. Org. Chem.* **42** (1977) 3425.
38. F. J. Williams, H. M. Relles, J. S. Manello, P. E. Donahue, *J. Org. Chem.* **42** (1977) 3419.
39. I. W. Serfaty in K. L. Mittal (ed.): *Polyimides. Synthesis, Characterization, and Applications,* **vol. 1,** Plenum Press, New York 1984, pp. 149 – 161.
40. T. R. W. Systems, US 3 528 950, 1970 (H. R. Lubowitz).
41. T. R. W. Systems, US 3 647 529, 1972 (H. R. Lubowitz et al.).
42. T. R. W. Systems, US 3 781 240, 1973 (H. R. Lubowitz).
43. T. R. W. Systems, US 3 827 927, 1974 (H. R. Lubowitz).
44. S. L. Kaplan, S. S. Hirsch in R. D. Deanin (ed.): *New Industrial Polymers,* Marcel Dekker, New York 1975, chap. 8, p. 100.
45. T. T. Serafini, P. Delvigs, G. R. Lightsey, *J. Appl. Polym. Sci.* **16** (1972) 905.
46. T. T. Serafini, P. Delvigs, *Appl. Polym. Symp.* **22** (1973) 89.

47. M. P. Hanson, C. C. Chamis, *29th Ann. Tech. Conf. Reinf. Plast.* /Comp. Inst. Soc. Plast. Ind., Section 16-C, 1974.
48. T. T. Serafini, R. D. Vannucci, *30th Ann. Tech. Conf. Reinf. Plast.* /Comp. Inst. Soc. Plast. Ind., Section 14-E, 1975.
49. T. T. Serafini, R. D. Vannucci, W. B. Alston, report, NASA TM 71 894, 1976.
50. W. E. Winters, P. J. Cavano, *10th Nat. SAMPE Tech. Conf.* **10** (1978) 661.
51. T. L. St. Clair, D. J. Progar, *24th Nat. SAMPE Symp.* **24** (1979) 1081.
52. T. L. St. Clair, R. A. Jewell, *23rd Nat. SAMPE Symp.* **23** (1978) 520.
53. D. Wilson, et al., *SAMPE J.* (1987) May/June, 35.
54. J. N. Hay et al. in C. Feger, M. M. Khojasteh, J. E. McGrath (eds.): *Polyimides. Materials, Chemistry and Characterization,* Elsevier Science Publishers, Amsterdam 1989, pp. 305 – 320.
55. J. N. Hay, J. D. Boyle, S. F. Parker, D. Wilson, *Polymer* **30** (1989) 1032.
56. M. Bargain, A. Combet, P. Grosjean, UK 1 280 846, 1972.
57. M. Bargain, A. Combet, P. Grosjean, UK 1 280 847, 1972.
58. M. Gruffaz, DE 2 251 065, 1973.
59. M. A. J. Mallet, F. P. Darmory in R. D. Deanin (ed.): *New Industrial Polymers,* Marcel Dekker, New York 1975, chap. 9, p. 112.
60. F. P. Darmory, *18th Nat. SAMPE Symp.* **18** (1973) 317.
61. H. D. Stenzenberger, *Appl. Polym. Symp.* **22** (1973) 77.
62. H. D. Stenzenberger et al., *Br. Polym. J.* **15** (1983) 2.
63. H. D. Stenzenberger et al., *32rd Int. SAMPE Symp.* **32** (1987) 44.
64. H. D. Stenzenberger et al., *19th Int. SAMPE Tech. Conf.* **19** (1987) 372.
65. N. Bilow, A. L. Landis, T. J. Aponyi, *20th Nat. SAMPE Symp.* **20** (1975) 618.
66. R. H. Boschan, A. L. Landis, T. J. Aponyi, *21st Nat. SAMPE Symp.* **21** (1976) 356.
67. M. D. Sefcik, E. O. Stejskal, R. A. McKay, J. Schaefer, *Macromolecules* **12** (1979) 423.
68. A. L. Landis, A. B. Naselow, *14th Int. SAMPE Tech. Conf.* **14** (1982) 326.
69. N. Bilow, L. B. Keller, A. L. Landis, R. H. Boshan, *Soc. Adv. Mater. Proc. Eng. Ser.* **23** (1978) 791.
70. A. O. Hanky, T. L. St. Clair, *Soc. Adv. Mater. Proc. Eng. Ser.* **28** (1983) 711.
71. D. J. Capo, J. E. Schoenberg, *18th Int. SAMPE Tech. Conf.* **18** (1986) 710.
72. M. A. Chadhari, B. Lee, J. King, *32nd Int. SAMPE Symp.* **32** (1987) 597.
73. R. A. Kirchoff et al., *18th Int. SAMPE Tech. Conf.* **18** (1986) 478.
74. L. S. Tan, F. E. Arnold, *ACS Polym. Preprints* **26** (1985) no. 2, 178.
75. L. S. Tan, E. J. Soloski, F. E. Arnold, *ACS Polym. Mat. Sci. Eng.* **56** (1987) 650.
76. J. L. Cotter, *Rev. High Temp. Mater.* **3** (1978) 277.
77. R. A. Pike, R. C. Novak, M. A. DeCrescente, *Proc. 30th Ann. Tech. Conf. Reinf. Plast. Comp. Inst. Soc. Plast. Ind.,* **Section 18-G,** 1975.
78. A. K. St. Clair, T. L. St. Clair, *28th Nat. SAMPE Symp.* **28** (1981) 165.
79. Upjohn, Highlights, Polyimide fibers, report no. 3.
80. T. Kaneda, T. Katsura, K. Nakagowa, H. Makins, *J. Appl. Polym. Sci.* **32** (1986) 3133, 3151.
81. A. A. Konkin (ed.): *Thermally Stable and Noncombustible Fibers,* Khimiya, Moscow 1978.
82. J. J. Anderson, *SAE Technical Paper Series 810 968* (1981).
83. W. W. Wright in N. Grassie (ed.): *Application of Thermal Methods to the Study of the Degradation of Polyimides in Developments in Polymer Degradation* **vol. 3, chap. 1,** Applied Science Publishers, London 1981.
84. A. K. St. Clair, T. L. St. Clair in K. L. Mittal (ed.): *Polyimides, Synthesis, Characterization, and Applications,* **vol. 2,** Plenum Press, New York 1984, p. 977.
85. W. Collins, T. Villani, paper presented at *Conference on Thermosets versus High Performance Thermoplastics,* New Haven, Conn., 1977.
86. I. Serlin, A. H. Markart, E. Lavin, *Proc. 25th Ann. Tech. Conf. Reinf. Plast. Comp. Div. Soc. Plast. Ind., Section 19-A,* 1970.
87. P. B. Rand, *J. Cell. Plast.* **3** (1973) 130.
88. H. Lee, D. Stoffey, K. Neville: *New Linear Polymers,* McGraw-Hill, New York 1967, p. 194.
89. Du Pont de Nemours, US 2 146 209, 1939 (G. Graves).
90. Röhm GmbH, US 3 284 425, 1966 (G. Schroeder and K. Tessmar).
91. Röhm GmbH, DE 1 817 156, 1968 (W. Bitsch, W. Gaenzler, G. Schroeder).
92. Röhm GmbH, DE 2 726 259, 1977 (G. Schroeder).

93. Röhm GmbH, DE 2 726 260, 1977 (G. Schroeder).
94. W. Gaenzler et al., *Angew. Makromol. Chem.* **11** (1970) 91 – 108.
95. F. Gotzen, G. Schroeder, G. V. Schulz, *Makromol. Chem.* **88** (1965) 133 – 148.
96. T. L. Sokolova, G. D. Rudkovskaya, *J. Polym. Sci. Part C* **16** (1967) 1157 – 1166.
97. G. B. Butler, G. R. Myers, *J. Macromol. Sci. Chem. A 5* **1** (1971) 105 – 134.
98. H. Hiraoka, *Macromolecules* **10** (1977) 719.
99. IBM, US 3 964 908, 1976 (J. Bargon, E. Gipstein, H. Hiraoka).
100. Rohm and Haas, US 4 246 374, 1981 (R. M. Kopchik).
101. M. Hallden-Abberton, *ACS PMSE Proc.* **65** (1991) 361.
102. G. Schroeder, *Makromol. Chem.* **96** (1966) 227.
103. Röhm GmbH, DE 1 242 369, 1967 (G. Schroeder, K. Tessmar).
104. Rohm and Haas, US 4 874 824, 1989 (M. Hallden-Abberton, L. Cohen, R. Wood).
105. Rohm and Haas, US 4 727 117, 1988 (M. Hallden-Abberton, N. Bortnick, L. Cohen, W. Freed).
106. Rohm and Haas, US 4 954 574, 1990 (M. Hallden-Abberton et al.).
107. Rohm and Haas, US 5 004 777, 1991 (M. Hallden-Abberton et al.).
108. H. Hiraoka, E. Gipstein, J. Bargon, L. Welsh, *J. Appl. Polym. Sci.* **22** (1978) 3397.
109. Gulf, US 3 557 070, 1971 (H. D. Anspon, B. H. Clampitt, R. E. Gilbert).
110. Union Carbide, US 3 632 797, 1972 (F. Boardman).
111. K. Ziegenberg, E. Klesper, *Polym. Bull.* **2** (1980) 691.
112. K. Ziegenberg, E. Klesper, *Makromol. Chem.* **184** (1983) 1705 – 1726.
113. GAF, US 4 169 924, 1979 (E. Barabas, A. Klein).
114. Du Pont de Nemours, GB 2 101 139 A, 1982 (E. A. Flexman).
115. A. Strepikheev et al., *Volokna* **3** (1957) 74 (CA 54 : 1317 i).
116. SKY Polymers, US 4 331 783, 1982 (V. Stoy).
117. SKY Polymers, US 4 337 327, 1982 (V. Stoy).
118. SKY Polymers, US 4 370 451, 1983 (V. Stoy).
119. SKY Polymers, US 4 420 589, 1983 (V. Stoy).
120. N. Grassie, I. McNeill, J. Samson, *Eur. Polym. J.* **14** (1978) 931.
121. I. McNeill, M. Zulfiqar, *J. Polym. Sci.* **16** (1978) 2565.
122. J. Shepitka et al., *J. Appl. Polym. Sci.* **28** (1983) 3611.
123. Mitsubishi Rayon, US 4 745 159, 1988 (H. Anzai et al.).
124. Mitsubishi Rayon, US 4 816 524, 1989 (H. Anzai et al.).
125. Mitsubishi Rayon, US 5 073 606, 1991 (I. Sasaki, K. Nishida, H. Anzai).
126. Dow, US 5 045 604, 1990 (B. R. Maurer).
127. A. Jamieson, I. McNeill, *Eur. Polym. J.* **10** (1974) 217 – 225.
128. N. Grassie, I. McNeill, J. Samson, *Polym. Degrad. Stab.* **1** (1979) 17 – 35.
129. H. C. Haas, R. L. MacDonald, *J. Polym. Sci.* **9** (1971) 3583 – 3593.
130. Rohm and Haas, US 4 742 123, 1988 (R. M. Kopchik).
131. Rohm and Haas, KAMAX (TM) Resins Technical Notes, Philadelphia, Pa., March 1990.
132. Rohm and Haas, US 4 524 121, 1985 (R. D. Gleim, M. DeGrandpre).
133. Rohm and Haas, US 4 254 232, 1981 (D. Mueller).
134. Rohm and Haas, US 4 255 322, 1981 (R. M. Kopchik).
135. Rohm and Haas, US 4 415 706, 1983 (W. Staas).
136. Du Pont de Nemours, EPA 94 215, 1983 (L. Inskip, M. Waggoner).
137. Rohm and Haas, US 4 436 871, 1984 (W. Staas).
138. Rohm and Haas, patent pending (M. Hallden-Abberton, N. Bortnick, L. Cohen, W. Freed).
139. Rohm and Haas, US 5 084 517, 1992 (M. Hallden-Abberton).

Polyisobutene → Polyolefins
Polyisobutylene → Polyolefins
Polymer Alloys → Polymer Blends

Polymer Blends

Jan Bussink, GE Plastics, Bergen op Zoom, The Netherlands

Hendrik T. van de Grampel, GE Plastics, Bergen op Zoom, The Netherlands

1. Introduction 1
2. Market Requirements 4
3. Blend Technologies 5
3.1. Thermodynamics 6
3.2. Compatibility 7
3.3. Compounding 8
3.4. Reactive Compounding 10
3.5. Mechanical Properties 11
3.6. Additives 12
4. Property and Application Profiles of Selected Blends 13
4.1. Polyolefins 14
4.2. Poly(Vinyl Chloride) 15
4.3. Styrene Polymers and Copolymers . 16
4.4. Polyamides 18
4.5. Poly(2,6-Dimethyl-1,4-phenylene Ether) 19
4.6. Polycarbonates 21
4.7. Polyesters 23
4.8. Polyetherimides 23
4.9. Polysulfones 24
4.10. Polyoxymethylenes 24
4.11. Polyarylates and Copolyarylate – Carbonate Blends . 25
4.12. Poly(Phenylene Sulfide) 26
5. General Application Technology . . . 26
5.1. Processing of Blends 26
5.2. Design Technology 27
6. Future Developments 27
6.1. Recycling 27
6.2. New Developments 29
6.3. Outlook 31
7. References 31

Acknowledgement. The authors are deeply indebted to Veronique Hopmans for her contributions.

1. Introduction

A continuous need for new materials is created by the increasing complexity of products that result from technological breakthroughs and the desire to obtain improved, cheaper materials for existing applications. Alongside the development of new improved base polymers and their copolymers, blend technology is one of the most widely used routes to satisfy this need. Polymer blends (also known as alloys) are combinations of two polymers with one or more distinctive properties that contribute to or are required for the proper functioning of a specific application. Their enhanced, especially adapted property profiles provide advantages to the processor and end user such as extra ease of processing, better low-temperature impact performance, higher strength, higher or lower gloss, reduced flammability, and higher transparency, often in combination with a lower price.

The first generation of blends was developed in the 1950s to improve impact performance. This impact modification normally requires the utilization of rubbery components (materials with a low glass transition temperature, T_g, and generally based on polybutadiene structures) that need a stabilization package. The first impact-modified grades were either made by special polymerization techniques (e.g., copolymerization, block and graft polymerization) or by compounding techniques. Their properties are based upon those of the main matrix polymer and are consequently limited in their scope of application. Although impact properties were improved, a number of undesirable properties were introduced (e.g., delamination, poor surface quality, and inconsistent property profiles). Better materials were made by introducing core – shell impact modifiers. A polymer with a property profile that fulfills market requirements in all areas does not exist.

Introduced in the early 1960s, engineering polymers such as PC and PPE (Table 1) became available as new basic resins for further blend developments. These second-generation blends offer improved heat resistance and higher con-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_273

tinuous use temperatures compared to the older generation materials.

Table 1. Thermoplastics

Type of plastic	Abbreviation	Name
Commodity	PE	polyethylene
	PP	polypropylene
	PS	polystyrene
	PVC	poly(vinyl chloride)
Transitional	ABS	acrylonitrile – butadiene – styrene copolymer
	SAN	styrene – acrylonitrile copolymer
	SMA copolymer	styrene – maleic anhydride copolymer
		acrylics
Engineering	PA	polyamide
	PBT	poly(butylene terephthalate)
	PC	polycarbonate
	PETP	poly(ethylene terephthalate)
	POM	polyoxymethylene
	PPE	poly(2,6-dimethyl-1,4-phenylene ether)
	SMA terpolymer	styrene – maleic anhydride terpolymer
High-performance	PEI	polyetherimide
	PPS	poly(phenylene sulfide)
	PSU	polysulfone

Thermoplastics can be classified into commodity, transitional, engineering, and high-performance plastics (Table 1). Commodity or bulk plastics are made in large volumes, and generally have a low continuous use temperature. High-performance thermoplastics have a high temperature resistance but a low production volume and high price.

Engineering polymer blends are a special group of plastics, whose applications are related to and depend on their specially designed property profile. They all offer dimensional stability over a wide range of maximum temperatures in combination with a minimum continuous use temperature of 85 °C. In addition, they usually possess high impact resistance over a wide range of temperatures down to as low as − 40 °C. Their dimensional stability is determined by the component with the highest glass transition temperature or by the combination of a high crystallinity and high melting point of the highestmelting component. The latter case can be exploited best by reinforcement with inorganic fillers or fibers. The use of a polymer with a high glass transition temperature as a functional filler yields even more interesting systems because they are not restricted to a limiting volume fraction and orientation of the reinforcement. Figures 1 and 2 compare engineering blends with thermoplastics and composites in terms of continuous use temperatures and impact strengths.

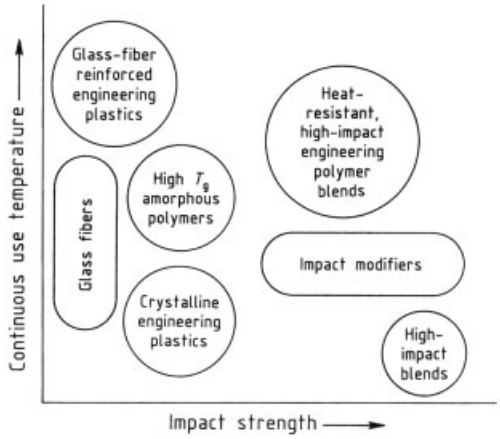

Figure 1. Second-generation polymer blends of engineering plastics

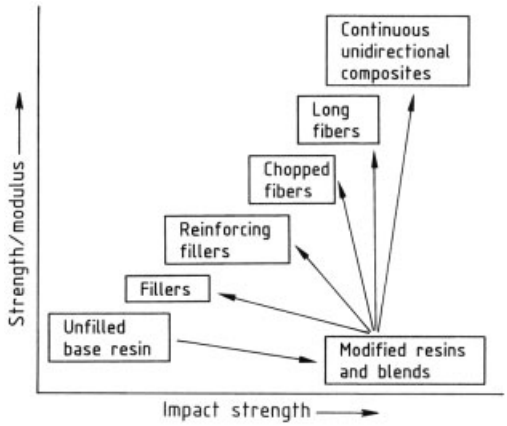

Figure 2. Relationship between mechanical properties and impact strength for various thermoplastic materials

In 1990 blends accounted for 10 % of total thermoplastics production [1–7]. Annual growth figures of 6 – 8 % are predicted versus 2 –3 % for other plastics. Upgrading products based on commodity plastics by reinforcing or blending with engineering polymers creates a very competitive situation for engineering plastics. It has

been estimated that the U.S. market for speciality polymeric blends was $\$900\times10^6$ in 1990, and was growing at an annual rate of 7 % [7]. Growth rates for individual engineering blends depend on:

1) Growth rate of major application area (e.g., automotive appliance) [8]
2) Continent of interest (U.S., Europe, Japan)
3) Position in product life cycle: e.g., PPE – HIPS blends are considered a mature blend compared with PC – ABS, and will therefore experience lower growth figure.

Companies that offer blends usually manufacture one or more of the blend components [9]. More polymer manufacturers and compounders are entering this business now because most basic patents for the production of thermoplastics have expired (e.g., PPE, PC) and blends allow facile expansion of the product range [4], [9–11].

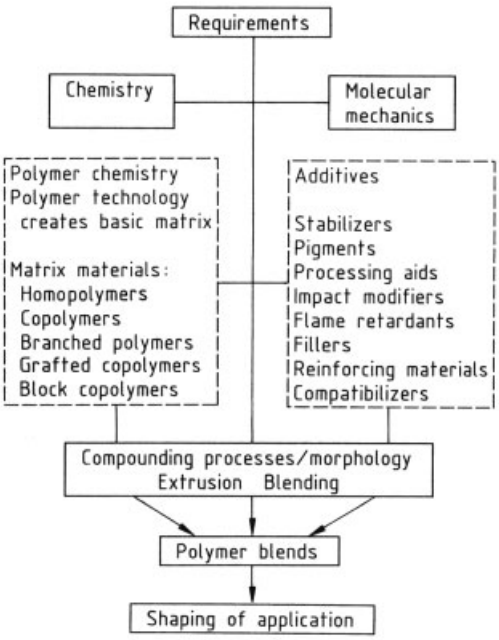

Figure 3. Schematic representation of blend development

From a technical point of view, it is important to translate the demands of a given application to the key physical and mechanical properties of the engineering plastics blends. Target values for heat and chemical resistance, high- and low-temperature impact strength, and stiffness should be assessed. The rationale used for choosing two (or more) polymers (the blend) for a specific application is important (Fig. 3). The dominant factor that determines whether a specific polymer blend is suitable for a certain application is the requirement profile (Chap. 2). The requirement can either be one demanded by the customer or end user, or one imposed by the standards of the specific industry segment (e.g., flammability standards in the aircraft industry).

Two polymers may be miscible, partially miscible, immiscible, or compatible. Consequently, the property profile resulting from blending two polymers is not a simple combination of the two individual property profiles. The relation between different blend classifications in terms of properties and morphology are discussed in Chapter 3. Techniques used to influence blend thermodynamics and ultimate property profile are also considered.

The general guidelines for the identification of a blend property profile are outlined in Chapter 4, and are subsequently applied to the most important blends. Once a blend of two polymers has been selected, qualification of the resin for a certain application depends on processing and design considerations (Chap. 5). Chapter 6 focusses on new developments in polymer blends.

2. Market Requirements

Specification of material requirements is determined by the needs of the end user and by the standards within an industry segment. All materials have to fulfill the requirements specified by industrial standards or mandated by law. These standards are controlled by standardized test methods and vary for different countries and industrial segments.

Blend development starts from the combination of a required property profile and an existing technology. The main industrial application areas for polymer blends are the automotive market, followed by the packaging industry, electronic and electrical industry, business equipment, and appliance sectors. Smaller application segments are construction, consumer goods, sheet manufacturers, and aerospace industries.

Automotive Applications. A number of different application areas can be discerned in an

automible. Horizontal exterior body parts demand a higher temperature resistance than vertical exterior body parts since they absorb more radiation from the sun. Both types require high surface quality and facile paintability, the coating technique being determined by the temperature resistance of the material. Mechanical properties are very important for structural body parts, but nonstructural body parts also require a high dimensional stability as characterized by a low coefficient of thermal expansion.

Bumper, or better bumper systems, have a crash-absorbing and an aesthetic function. The crash-absorbing function of bumper systems can be fulfilled by incorporating appropriate failure mechanisms. Decorative functions demand easy paintability, high UV resistance, and resistance to small impact loads. The trend is towards a separation of both functions. Facile dismantling is becoming important for recycling.

Wheel covers should have a good dimensional stability and be able to withstand the high temperatures caused by braking. Chemical resistance is usually provided by a coating.

Lighting systems consist of a light-reflecting component and a housing. The latter requires an extremely high dimensional stability and related mechanical properties. The reflector must be temperature resistant and possess sufficient reflecting capacity (surface quality and gloss).

The *passenger compartment* contains a high number of decorative applications designed for comfort. Resistance to low continuous loads (e.g., instruments in instrument panel) and glossy/nonglossy appearance are general considerations. The temperature requirement is related to exposure to the sun, but a heat distortion temperature of 120 °C is generally accepted. Key requirements for the dashboard material include a minimum heat distortion temperature of 120 °C, excellent accoustic properties (reduction of outside and engine noise), high impact resistance, and good scratch resistance. Paintability and compatibility with soft-touch overlay films are additional requirements.

Structural parts (e.g., chassis) demand such high mechanical properties that composites are the only feasible solution. *Under-the-hood applications* include cable and wiring, which demand electrical, chemical, and high-temperature resistance.

Fuel tanks place high demands on the barrier properties of the resin, in addition to its chemical resistance. PE in fuel tanks has been subjected to (sulfo)fluorination to increase surface polarity and improve barrier properties. Other techniques are employed in utilizing PA – PE or PA – PVAL and PVAL – PA blends [PVA = poly(vinyl alcohol)] for fuel tanks where a special layered morphology ensures impenetrability.

Applications in the *motor compartment* have to withstand harsh environmental conditions, and thus require a high-temperature and environmental (including chemical) resistance. The most suitable products are semicrystalline polymers.

Electronic and Electrical Applications. The dominating requirements in this industry are high heat and chemical resistance and high dimensional stability. A key requirement is a low, preferably temperature-independent, conductivity. The trend towards miniaturization demands better flowing resins coupled with a high heat resistance. Higher heat materials are also required for the introduction of the surface-mount circuit-board assembly [11] in which thermoplastics are combined with metals. The materials must be able to withstand chemical cleaning baths and temperatures over 260 °C for more than 1 min resulting from high-temperature soldering via wave, vapor-phase, IR, or laser methods. Dimensional stability should be maintained during these operations and under operating conditions. Electrical (office) equipment must fulfill stringent flammability requirements, particularly when used in restricted areas.

Appliance Sector. Basic requirements for household appliances are water resistance, chemical resistance, appearance (glossy etc.), and temperature resistance. Electrical properties (e.g., conductivity) and absence of toxicity risks must conform to consumer and government standards.

Building and Construction Applications. Material requirements differ from country to country. Public buildings are subject to the most severe rulings. The greatest drawback of plastics is their intrinsic flammability resulting from their organic nature. The fire tests mandated for plastics are thus difficult to pass. To reduce the

risk of fire, only relatively small amounts of plastics can be used in low-risk areas (e.g., outside sidings). The most important flammability characteristics are ignition and flame propagation behavior. Materials may have horizontal, vertical, or no fire propagation. Assessment of flammability characteristics is extremely complex, since the behavior of materials under fire conditions depends not only on the nature of the material, but also on its thickness and the configuration of the part.

Mechanical requirements are much easier to fulfill and depend largely upon the temperature resistance of the material.

Packaging Applications and Pipes. Basic requirements for packaging materials are barrier properties for water, oxygen, and carbon dioxide as well as excellent extrudability resulting from a high melt strength. Permeability requirements are usually satisfied by multilayer systems in which polymers are alternated with tie layers. Properties can be optimized by bidirectional orientation. These multilayer systems can be substituted by a single, special two-phase blend material.

Pipes for water distribution should have a long guaranteed life time (>100 years), and prevent migration of low molecular mass compounds into the water.

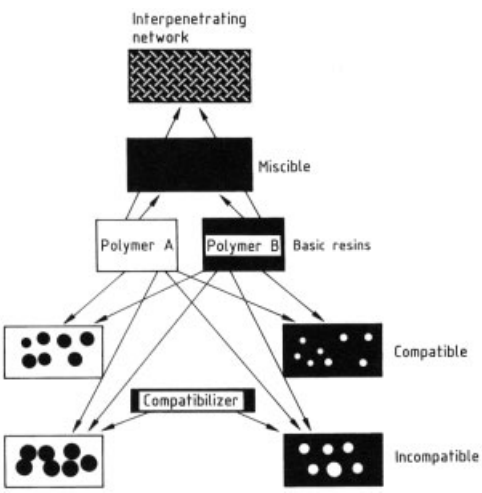

Figure 4. Possibilities of blend morphology

3. Blend Technologies

From a thermodynamic perspective, polymer blends may be completely miscible, partially miscible, or immiscible (Fig. 4). An almost fluid transition is possible between these systems; by modifying the base polymers each blend can be optimally positioned for the foreseen application.

3.1. Thermodynamics

The thermodynamic blend requirements have been reviewed elsewhere [12–15]. Blend polymers interact with each other in their amorphous state, either in a melt or dissolved in a common solvent. A thermodynamically stable mixture of two polymers is obtained if the free energy of the mixture is lower than the combined values of the individual components, i.e., the energy of mixing ΔG_{mix} should be negative. The free energy of mixing is composed of an enthalpy and entropy term [12–15]

$$\Delta G_{\mathrm{mix}} = \Delta H_{\mathrm{mix}} - T\Delta S_{\mathrm{mix}} \tag{1}$$

A homogeneous blend over the complete composition range of two components 1 and 2 is only obtained if

$$\left(\partial^2 G/\partial x^2\right) > 0 \tag{2}$$

in which x denotes the volume fraction of component 2. The entropy contribution to the overall free energy of the system is usually negligible, and is only important in special cases (e.g., blending with relatively low molecular mass components such as plasticizers).

The enthalpy resulting from the interaction energy between the two polymers determines whether they are completely or partially miscible. Complete mutual solubility results in a single-phase system. Partial solubility provides a two-phase system with a very low interfacial energy allowing the production of very finely dispersed blend systems without the need for a third component such as a compatibilizing block copolymer. Normally, even when very similar components, such as polyolefins are blended, the enthalpy term is positive and leads to nonmiscibility. Phase separation can only be prevented if interaction between the different chemical groups in the two polymers is possible (e.g.,

dipole – dipole interaction, charge transfer complex formation, hydrogen bonding). A negative enthalpy value leads to a completely or partially miscible blend system. The miscibility of a two- component system can be improved by increasing the interaction energy between the two components or reducing the interaction energy within each component. The latter effect is achieved by using random copolymers made from monomers that, by themselves, form immiscible homopolymers. The internal repulsion energy in these copolymers can thus improve miscibility [16]. The interaction energy between two polymers is the main factor that determines the final blend result.

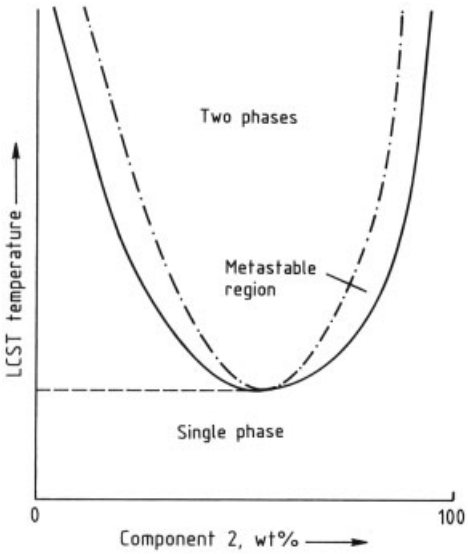

Figure 5. A schematic phase diagram for a polymer – polymer blend with a lower critical solution temperature (LCST)

The morphology of incompletely miscible systems can be structured. In a two-phase system with a dispersed phase and a continuous phase, the polymer constituting the more important continuous phase can be selected. The second possibility consists of an interpenetrating, co- continuous network formed from both polymers (Fig. 4). Morphological control depends on chemical interaction and rheological factors such as the melt viscosity ratio of the two polymers and the dependence of melt viscosity on shear rate. The compounding conditions used during melt blending are therefore of major importance for the final blend property profile (see Section 3.2 and Fig. 3).

The influence of temperature on blend thermodynamics is important because it determines phase separation behavior and morphology (see Section 5.1). A typical example is shown in Figure 5. Three regions can be discerned: (1) a thermodynamically stable region (miscibility), (2) a metastable region where the system is stable only towards small concentration fluctuations, and (3) a region where spinodal decomposition yields a two-phase system. The binodal (solid line) corresponds to $\Delta G_{\mathrm{mix}} = 0$, the spinodal (dashed-dotted line) is defined by $(\partial^2 G/\partial x^2) = 0$.

3.2. Compatibility

The term compatibility describes all useful (scientific) means of producing stable blend microstructures that provide technologically important properties for the performance of the part. In the absence of interactions between two blend components, the blend morphology is largely determined by the interfacial tension. The interface between two polymers in the melt is characterized by the interfacial tension and the domain adhesion [17]. The interfacial tension γ_{12} between polymers 1 and 2 is defined as the work required to increase the interface

$$\gamma_{12} = (\delta G/\delta A)_{P,T,n} \tag{3}$$

where A is the surface area between polymers 1 and 2. The magnitude of the interfacial tension is determined primarily by the polarity difference between the two phases [17]. It is also related to the surface tensions of the constituent polymers which are defined as the work required to increase the surface of a polymer by 1 cm^2. Increase in the surface area of a polymer involves the transport of polymer segments from the bulk to the surface which deprives them of the small forces acting in the bulk (e.g., van der Waals forces). In total,

$$\gamma_{12} = \gamma_1 + \gamma_2 - W_{\mathrm{a}} \tag{4}$$

where γ_1 and γ_2 are the surface tensions of polymers 1 and 2, and W_{a} is the work of adhesion (equal to the decrease in Gibbs free energy on forming an interface between polymers 1 and 2). In practical terms this means that the work

of adhesion increases as the interfacial attraction increases, corresponding to a decrease in interfacial tension. If two immiscible polymers are mixed, more interface is created and the free energy of the system increases. The system is thermodynamically unstable and strives toward a reduction of the total interface surface. Consequently, if two individually dispersed polymer droplets encounter each other, they agglomerate immediately. This leads to coarsening of the phase morphology which can only be prevented by the use of compatibilizers. A compatibilizer reduces the interfacial tension, improves the interfacial adhesion, permits finer dispersion during mixing, and stabilizes morphology during processing.

The most common types of compatibilizers are AB block copolymers for a blend of polymers A and B (Fig. 6) that are added prior to compounding or synthesized in situ [18], [19]. The block copolymer decreases the interfacial tension by dissolution of the copolymer block segments within their respective homopolymer phases, leading to covalent bonds being positioned across the interface. Physical anchoring of both phases contributes to the stability of the resulting morphology and mechanical strength. In commercial blends, 5 – 15 wt % compatibilizer is used.

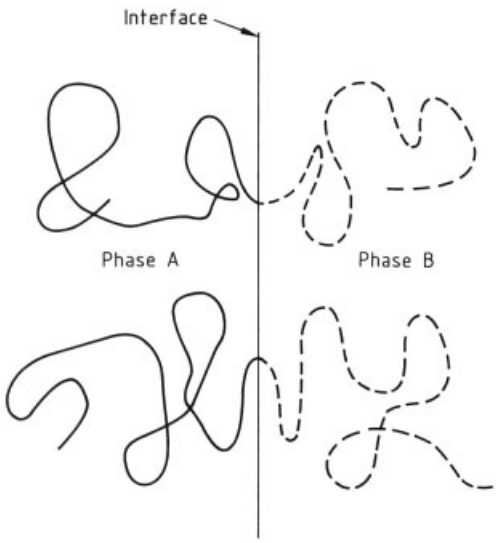

Figure 6. Ideal configuration of block copolymer at the interface between polymer phases A and B

Several other routes are available for compatibilizing phases in commercial multicomponent polymer blends. Large-scale phase separation can be prevented by promoting cocrystallization, by cross-linking the dispersed phase, or by the formation of interpenetrating networks. Ionic attraction between blend components can also be used to stabilize the phase structure. High stress shearing conditions may be applied to mechanically interlock both phases, but this technique is seldom practised.

3.3. Compounding

For scientific and screening purposes, limited amounts of blends are prepared by combining solutions of the blend components from the same solvent or mutually miscible solvents. Blends of miscible polymers can then be obtained by solvent evaporation or coprecipitation. Care must be taken not to heat a blend beyond a potential LCST temperature in a subsequent melt processing step. In dealing with partially miscible or compatible systems, coprecipitation followed by a melt cycle is usually employed [20], [21].

The main method of blend production is batch or continuous melt blending in kneaders or compounding extruders [20], [21]. The machines used are corotating twin-screw compounders with a mixing section based on the Erdmenger principle, Busch co-kneaders, and Banbury kneaders.

The technology involved in this method of manufacturing polymer blends is determined by the thermodynamics of the blend system (Section 3.1) and by the rheological conditions during melt blending. The final morphology of the blend in the compounding step is of prime importance for its resulting property profile.

Compounding machines are not simple extruders which merely melt and transport the resin [15], [22], [23]. Intensive mixing of the melt is achieved by the dispersive or distributive action of the screw element. Depending on the thermodynamics of the mixture of molten components, a dispersion of component 1 in 2 or 2 in 1 is obtained. The criteria determining the continuous phase are the volume fractions used in the melt-blending process, the ratio of the individual melt viscosities, and the surface tension (the polymer

with the lowest surface tension usually becomes the continuous phase).

Two basic microrheological processes can be discerned in the blending of polymers: distributive and dispersive mixing. Distributive mixing determines the homogeneous distribution of the second phase throughout the matrix, whereas dispersive mixing causes large particles to break up into smaller ones. During actual mixing both processes occur simultaneously.

Distributive Mixing. At the start of the blending process, the droplet size of the second phase is relatively large. Two counteracting stresses are in effect: the deforming shear stress τ and the interfacial stress σ/R. Their ratio is usually called the capillary number Ca.

$$Ca = \tau R/\sigma = \eta_m \dot{\gamma} R/\sigma \tag{5}$$

where

$\dot{\gamma}$ = shear rate (s^{-1})
η_m = viscosity of matrix (Pa · s)
σ = interfacial tension (N/m)
R = radius of particle (m)

Typical shear stresses in polymer melts are in the order of 10^4 N/m^2. The local radius of a dispersed droplet is approximately 10^{-3} m in the early stages of the mixing process, and σ is typically 10^{-2} N/m. An affine deformation is the result.

The total interface is a measure of the effect of distributive mixing and, in the absence of reorientation, is directly proportional to the total shear γ.

$$A = A_0 \gamma \tag{6}$$

where A is the interfacial area and A_0 the initial interfacial area.

If the layers are oriented along the shearing direction, such as in corotating twin-screw extruders, distributive mixing becomes more efficient.

When two polymers dissolve in each other, they do so by diffusion. The time required for complete dissolution is greatly reduced by decreasing the droplet size of the starting dispersion; this is facilitated by minimal surface tension between the two initially molten phases. The distributive mixing action is produced by the shearing and elongating forces that act upon the continuous phase and are transferred to the dispersed phase [15], [22], [23]. Under stagnant conditions, the dispersed phase consists of molten droplets which are deformed into thread-like structures. The length – diameter ratio of these structures depends on the interfacial tension constraints of the dispersed phase, and on the effectiveness of the transfer of the deformation forces from the continuous phase to the dispersed phase. The latter is maximized when the ratio of the apparent melt viscosities of the continuous and dispersed phases is close to one. Equation (7) describes the equilibrium conditions for droplet formation [24], [25]:

$$\eta_1/\eta_2 = \Phi_1/\Phi_2 \tag{7}$$

where η is the melt viscosity and Φ the partial volume.

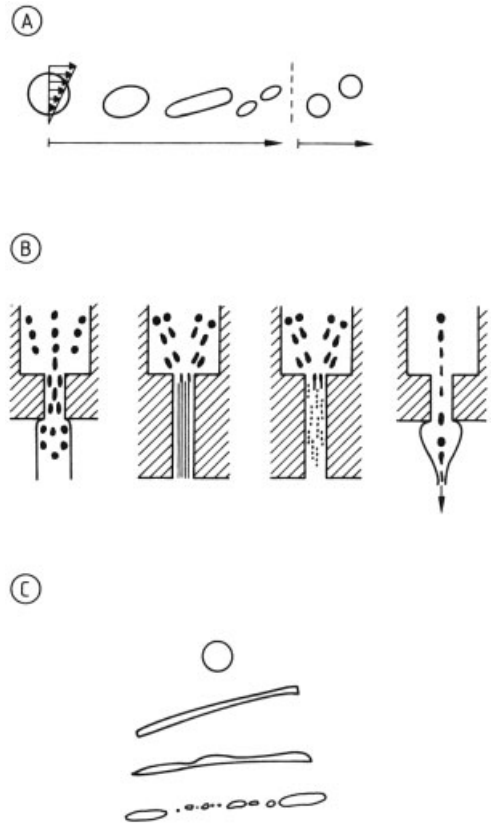

Figure 7. Droplet deformation
A) Deformation caused by shear flow; B) Deformation caused by elongational flow; C) Actual droplet formation (combination of A and B)

Dispersive Mixing. The dispersed phase deforms affinely with the fluid motion, resulting in elongated droplets. This process continues until the capillary number approaches unity. This limit is reached if the local radii are in the order of 10^{-6} m, resulting in $\sigma/R = 10^4$ N/m^2. The elongated droplets become unstable due to interfacial tension driven Rayleigh distortions, and droplets are formed, which are again subjected to shear stress.

This process of elongating and breaking up repeats itself until the droplet size reaches an equilibrium dimension determined by the above-mentioned factors. These process steps are shown in Figure 7.

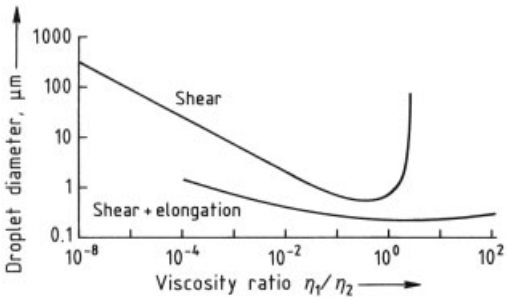

Figure 8. Influence of deformation forces on droplet diameter as a function of viscosity ratio η_1/η_2 with η_1 the dispersed and η_2 the matrix viscosity

Figure 8 illustrates the difference between the effect of shear and elongational flow on particle size. For high viscosity ratios ($\eta_1/\eta_2 > 4$), dispersed particles do not break up in shear flow because they rotate around themselves. Elongational flows are irrotational, and dispersive mixing is possible. The final quality of the blend dispersion depends entirely on this process. The interfacial tension has to be very low(< 5 mN/m) to obtain a dispersed phase with dimensions smaller than 1 µm. In cases when the two blended polymers have a higher interfacial tension, compatibilizers are necessary to obtain the necessary dispersion quality. These compatibilizers are generally block copolymers consisting of monomeric units that are (preferably) miscible or at least highly compatible with the two phases of the blend system. During the melt-mixing process, they position themselves at the interface between the two miscible components, and thus overcome the repulsive effects of interfacial tension by forming chemical bonds between the two blocks. In this way, addition of a compatibilizer leads to a decrease in interfacial tension, yielding smaller particle sizes at the same viscosity ratio and shear stresses. It also provides the necessary morphological stability in this thermodynamically unstable system. Furthermore, mechanical properties are optimized if compatibilizer blocks of sufficient molecular mass are entangled in the relevant polymer phase.

The basic blend of two polymers described above can be regarded as the base matrix, which has to be further adjusted to the needs of the application by addition of additives (e.g., impact modifiers). The location of the impact modifier can be predetermined by chemical bonding with the desired phase or by specific surface interfacial effects.

Predicting Morphology. Why does one of the dispersed resin phases encapsulate the other? Under equilibrium conditions such effects are probably due to interfacial energy differences between the blend components. The tendency of a liquid to spontaneously spread across a solid or liquid substrate can be expressed in terms of the surface and interfacial tensions of the components using Harkin's equation

$$\lambda_{ij} = \gamma_j - \gamma_i - \gamma_{ji} \tag{8}$$

where γ_j, γ_i, and γ_{ji} are the surface and interfacial tensions of the solid, liquid, and solid/liquid respectively, and λ_{ij} is defined as the spreading coefficient. Equation (8) can be approximated by

$$\lambda_{12} = \gamma_1 - \gamma_2 - \frac{4\gamma_1^d\gamma_2^d}{\gamma_1^d + \gamma_2^d} - \frac{4\gamma_1^p\gamma_2^p}{\gamma_1^p + \gamma_2^p} \tag{9}$$

where γ^d and γ^p constitute the distributions of the contribution to the surface tension caused by dispersive and dipole interactions, respectively.

Spreading is predicted to occur only for positive values of λ_{12}. Melt blending is usually carried out at elevated temperature, interfacial tensions should therefore be known at processing temperatures but variation in surface tension with temperature is comparable for most polymers (0.06 – 0.08 dyne · cm/ °C).

With three or more different components a number of interfacial tension combinations result. This can create a competitive situation lead-

ing to a number of options. If the spreading coefficient λ_{12} is positive, component 1 will encapsulate component 2 and as such will be dispersed in 3 (Fig. 9) [26]. A three-phase system develops in those cases where none of the possible three-phase combinations has a positive spreading coefficient. All these morphologies are thermodynamically controlled.

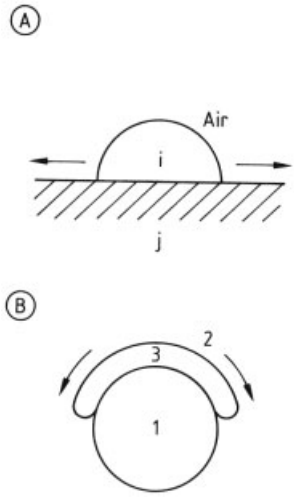

Figure 9. Schematic diagram showing spreading behavior of liquid on solid substrate (A) and one polymer phase on another within a third (B)

One of the most severe problems encountered in blends of two immiscible components with an insufficiently low interfacial tension is loss of morphology or dispersion through agglomerization. This leads to uncontrollable thermodynamically driven, phase separation and delamination with total loss of mechanical properties.

3.4. Reactive Compounding

Reactive compounding is used in those cases where the above-mentioned agglomeration problem has to be overcome by creating chemical bonds across the interface [27]. This technique can be applied with a single-phase matrix and a two-phase blend matrix. An example is phase-selective impact modification where the modifier is chemically bonded to one phase component of the blend matrix. This is possible through the use of specially functionalized components containing reactive groups. One component may be functionalized with a group that can react with an existing reactive site in the second component. Alternatively, both components may be functionalized with mutually reactive sites.

The most commonly used chemical reactions are those between carboxylic acids or anhydrides with amines or epoxy groups, and between amino and epoxy groups. Reactive extrusion involves the introduction of these reactive groups into a nonreactive polymer by means of radically induced graft reactions with unsaturated monomers or by replacing available reactive groups with more or different reactive groups.

The term in situ compatilization denotes those cases where a functional group is introduced to establish chemical bonds between the two phases or to generate a compatibilizer in situ. An example of this technology is found in the manufacture of the completely immiscible blend of PPE and PA with the assistance the addition of, for example, maleic anhydride (see Section 4.5, PPE – PA blends).

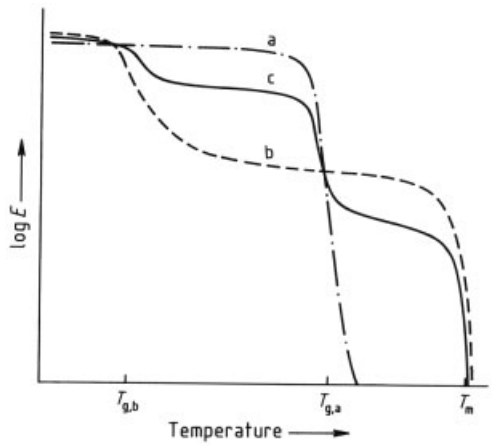

Figure 10. Generalized modulus – temperature diagram for two polymers and their blend
a) Amorphous polymer with high T_g; b) Crystalline polymer (e.g., polyamide, polyester) with low T_g; c) Blend of a and b

3.5. Mechanical Properties [28]

Nearly all thermoplastics are subjected to mechanical loading during their lifetime. Mechanical properties are therefore often used as the basis for material selection. Engineering

data sheets supplied by the manufacturer contain information on mechanical properties, thermal properties, electrical properties, flammability behavior, and chemical resistance. These properties are measured by standardized test procedures (e.g., ASTM, DIN, ISO), see → Plastics, Properties and Testing. Important mechanical properties for engineering blends are discussed briefly in this section [28].

Characterization of a thermoplastic (blend) usually starts with a curve of tensile modulus (E) plotted against temperature (Fig. 10).

The *glass transition temperature* T_g (1 Hz) is the temperature at which the larger chain segments either gain or lose rotational freedom. Its location shifts with the frequency of the test method, i.e., the performance of a material is difficult to predict based on merely T_g values. Secondary transitions also play an important role in determining impact resistance, especially if the T_g (1 Hz) lies close to the use (or room) temperature.

Impact strength is the ability of a material to withstand impulsive loadings. Tests are performed using notched and unnotched samples. The material fracture mode is important, being either ductile or brittle and depends on temperature (Fig. 11). The location of the ductile – brittle transition is usually reported.

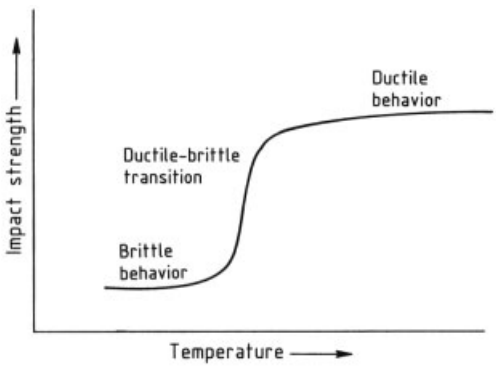

Figure 11. Typical impact behavior as a function of temperature

The *toughness* (ability to absorb mechanical energy without fracture) is measured as the area under standard (tensile) stress – strain curves. For many (engineering) thermoplastics it is usually improved by the incorporation of a low-modulus component (rubber) as a separate phase [29], [30]. When highly dispersed the rubber acts as an effective stress concentrator and initiates both crazing and shear-yielding phenomena in the matrix. Since both these processes can absorb large amounts of energy, impact-modified materials exhibit superior resistance towards crack propagation under impact conditions. The low-modulus phase may be incorporated by grafting, copolymerization, or melt blending. High stress concentration is reached when the ratio of the moduli of the rubber phase and the matrix is high. This value is increased when the rubber particle cavitates under the onset of the three-dimensional stress in the part.

Enhancement of the intrinsic toughness of the blend system increases impact resistance. The most important factor in optimizing the impact performance of a polymer resin is identifying the preferred deformation mechanism of the matrix polymer – crazing or shear band formation – because each mechanism requires a specific type of impact modification. In all cases the predominant deformation mechanism depends not only on the modified matrix system and its morphology, but also on variations in test conditions.

Two important phenomena are usually considered in *long-term behavior : creep* (the increase in strain for a material subjected to a constant load) and *stress relaxation* (reduction of stress induced in a material subjected to constant strain). Both processes provide estimates of the materials long-term dimensional stability and are important in selecting resins and design parameters for a certain application (see Section 5.2).

The *ultimate use temperature* is the maximum allowed temperature at which the material will provide the required dimensional stability under the foreseen conditions. The *continuous use temperature* is related to the oxidative stability or intrinsic stability of the blend system. This is usually determined by the weakest blend component, whereas the ultimate use temperature of the most thermally stable blend component determines the thermal resistance.

The two most commonly employed ultimate use temperatures are the Vicat softening temperature (Vicat A or B) and the heat distortion (deflection) temperature (HDT). Typical HDT values for completely amorphous polymers are 20 °C lower than the T_g values. The Vicat values are somewhere inbetween. For typical semicrys-

talline polymers, large differences are found between the high- and low-load HDT values. The Vicat values are usually close to the low-load HDT values.

3.6. Additives

The role of additives in modifying polymer properties should not be underestimated. See also, → Plastics, Additives.

The distribution of stabilizers between the various blend components (including impact modifiers) is important. Redistribution of stabilizers initially present in the weakest component, may yield surprising effects. For instance the UV resistance of PC – ABS blends is markedly higher than expected for an ABS system due to the presence of UV stabilizers in the continuous PC phase. Redistribution processes can be prevented by using chemically bonded additives.

Oxidation at moderate temperatures (thermal oxidation) and under the influence of UV radiation (photo-oxidation) can lead to cross-linking, lowering of molecular mass, and deterioration of mechanical properties. Unsaturated rubbery polymers, such as those used for impact modification, are particularly sensitive to attack by oxygen and heat. *Heat stabilizers* and *antioxidants* are added to suppress these reactions.

Flame retardants are applied to allow a polymer to conform to flammability requirements. A large variety of compounds (e.g., aluminum trihydrate, organic phosphates, antimony oxide in combination with various halogen compounds) are used for this purpose. They increase the amount of solid (nonvolatile) material by carbonization or cross-linking, and/or interfere in the oxidation processes taking place in the flame (flame poisoning).

Plasticizers are used to reduce the modulus of the resin (i.e., to increase flexibility) over the expected use temperature range. They are also effective as flow-improving additives and modify the continuous phase. They have to be selected on the basis of compatibility parameters. They are sometimes moderate molecular mass polymers with a low T_g, but more commonly oligomers with a molecular mass of 3000 – 1000. The use of polymeric plasticizers versus the traditional organic materials is growing due to environmental concerns (nonvolatile and less migrating) and also due to practical performance reasons.

Release agents used to provide easy mold release are selected according to their volatility and thermal stability in relation to the processing temperatures.

A number of types of *fillers* exist. Mica and lime are used to increase volume and reduce the material price. Other particulate fillers such as graphite, carbon black, and aluminum flakes are used to enhance the material's conductivity for a variety of purposes. Reinforcing fillers (e.g., glass fibers, continuous glass mats, or glass fabrics) are used chiefly to improve mechanical properties such as modulus and strength and increase the heat distortion temperature (Fig. 2).

Transesterification catalysts in blends of two polyesters may catalyze copolymer forming transesterification reactions during (melt) processing.

Catalyst complexing agents are added to prevent these unwanted side reactions. Compatibilizers have been discussed in Section 3.2. The role of *impact modifiers* has been discussed in Section 3.5.

4. Property and Application Profiles of Selected Blends

Thermoplastics are crystalline or amorphous. Crystalline resins (PA, POM, PP, PSU, polyesters) have superior resistance to organic solvents, a low melt viscosity, and are often not transparent. They also exhibit considerable mold shrinkage and warpage. High dimensional stability is only achieved below the glass transition temperature. Above this temperature, reinforcing fillers (e.g., glass fibers) are required to provide the desired dimensional stability and strength; they also enhance the tensile modulus. Furthermore, the crystallization rate and the history determine whether a material acts according to the characteristics described in Table 2.

Blending with engineering polymers is needed to improve their intrinsic property profile and expand their application range [31]. The amorphous polymers (ABS, PC) have excellent dimensional stability up to 20 °C below their T_g; maintain their properties at elevated temperature; and show good impact behavior, high trans-

parency, and low warpage. Amorphous polymers with a very high T_g are very difficult to process and/or very brittle. Also their application area is limited due to lack of chemical resistance. These issues had to be addressed before they were accepted by the market. Semicrystalline engineering polymers frequently exhibit a too slow or too rapid crystallization with a poorly controlled crystalline phase morphology. This causes a processing-dependent property profile, leading to unpredictable impact behavior. Furthermore their mechanical properties vary widely depending whether they are used below or above their glass transition temperature.

Table 2. Property profile of engineering thermoplastics

Resin	Advantages	Disadvantages
ABS	impact strength, processibility, surface appearance	HDT, flammability
ASA	weathering performance, ductility	HDT
HIPS	processibility, impact strength	chemical resistance, HDT
PA	processibility, impact strength if saturated with water, toughness, chemical resistance (crystalline PA)	water adsorption leads to general change in mechanical properties and dimensional stability, HDT
PBT	chemical and solvent resistance	processibility, shrinkage
PC	low-temperature toughness, HDT, constant modulus, high T_g	stress-crack sensitivity, solvent/chemical resistance, impact strength under certain conditions
PETP	chemical and solvent resistance	shrinkage, low-temperature toughness, crystallization speed
POM	tensile strength, modulus	impact strength
PP	chemical resistance, low price, high flammability	HDT, oxidation sensitive, brittle
PPE	HDT, rigidity, flame retardancy, impact strength	processibility, oxidative stability (UV)
PS	processibility	brittleness
PSU	excellent thermal and oxidative stability	expensive, notch sensitive

The advantages and disadvantages of the most important thermoplastics are listed in Table 2. The key to the development of a successful blend is to maintain the strengths of both polymers while removing their disadvantages. The heat distortion temperatures of these polymers and some engineering blends follow:

ABS	80 – 100
PBT	70
PC	135
PEI	170 – 240
PP	70
PPE	200
PS	80
SMA	105
ABS – PC	110
ABS – SMA	95
PA – PPE	170 – 190
PC – ABS	100 – 125
PC – PBT	120 – 140
PC – SMA	115
PEI – PC	170
PPE – HIPS	80 – 175

In the following sections selected commercially important blends and engineering blends are discussed. The treatment of the blends is based upon the main component (defined as being present in an amount of more than 50 wt %), and not the (for mechanical properties very important) continuous phase. Each section is concluded with a short list of applications that may be similar because different blends are targeted for the same application.

4.1. Polyolefins

Polyethylene. Polyethylene [*9002-88-4*] (PE) is classified according to density which is controlled by the attainable degree of crystallization and the density of the amorphous phase [32]. Densities range from 0.8 to 0.95 g/cm^3 and are primarily determined by their underlaying molecular structure which results from the way these polymers are synthesized. See also → Polyolefins, Chap. 1.

High-density PE (HDPE) is produced by comparatively low-pressure polymerization with Ziegler or complex oxide catalysts; density 0.95 g/cm^3, melting temperature 130 – 135 °C.

Low-density PE (LDPE) is traditionally made by a high-pressure route; density 0.92 g/cm^3, melting temperature 115 °C, but lower crystallinity than HDPE due to presence of branching structures.

Linear low-density PE (LLDPE) is comparable in structure and general properties with LDPE but lacks long- chain branching. It is produced by copolymerizing ethylene with butene, hexene, octene, or 4-methylpentene at low pressure.

Very low-density PE (VLDPE) is an extreme version of LLDPE; density may be as low as 0.88 g/cm^3.

Ethylene copolymers are typically polymers of ethylene with polar monomers such as acrylates.

Ultimate rheological properties can be adjusted by blending to suit the processing techniques used. Density can also be adjusted by blending because it simply follows linear additivity of the components used. From a practical point of view the polyolefin systems are completely compatible due to their very low interfacial tension.

Vinyl, acrylic, and even carbon monoxide comonomers are used to modify polyethylene. The possibilities of these modifications are, however, restricted because crystallinity and melt temperature rapidly decrease with comonomer content, resulting in products with a very low modulus of elasticity. Rubbery materials are produced in this way and are frequently used as impact modifiers for blends. Although chemical modification of polymers is preferred (primarily due to cost reasons) some property profiles can only be obtained through blending.

Polyethylenes are widely used as blend components because they improve a number of important properties such as flow, impact resistance, chemical resistance, and environmental stress cracking sensitivity. Cost – price reductions and volume – price improvements are an additional advantage.

One of the most important application areas of PE blends is the packaging industry. They display barrier properties in particular toward oxygen and water vapor, and chemical resistance toward aqueous systems, edible fats, and oils. In blends of polyamides with slightly functionalized PE, the desired blend morphology is best described by the term controlled delamination (i.e., agglomeration of the dispersed PE phase leads to a two-phase layer structure but does not reach an undesirably high level that would cause delamination). This can only be achieved through extrusion shaping [33]. The inherent incompatibility in PE – PA blends is overcome by chemical bond formation between carboxylic or maleic anhydride groups of the modified PE phase and amine end groups of the PA phase. The following types of compatibilizers have been used: maleic-anhydride or chlorine-modified polyolefins [34], ethylene – methacrylic acid copolymers, ionomers (maleic-anhydride-grafted PE), and epoxy compounds. Increasing the degree of PE modification with polar groups improves the adhesion of these intrinsically nonadhering polymers to more polar polymer blend components to give extremely tough materials [33], [35], [36]. Applications include automotive parts (mainly injection-molded compounds), power tools, farm and garden equipment, as well as sport and recreational markets.

PE – isobutene films have improved properties and toughness compared to HDPE films. Typical applications are single-layer extrusion films, and heavy-duty bags. HDPE, LDPE, and LLDPE polymers are blended in different ratios to obtain films in a customer-preferred density. Uses include films for bags.

Polypropylene. Polypropylene (PP) [*9003-07-0*] (→ Polyolefins, Chap. 2.) is used as homopolymers, random copolymers and block copolymers. Polypropylene is one of the most widely used commodity plastics. It has a good chemical resistance, excellent thermal stability, a low price, and high flowability. The polymer has three drawbacks: (1) it is sensitive to oxidation during processing and use, (2) it has a poor resistance to UV-induced degradation, and (3) it is rather brittle even though its glass transition temperature (− 10 °C) is below room temperature.

Thermal oxidation problems have been solved by use of specially developed additive packages. Impact performance can be improved by chemical modification (copolymerization) or blending. Comonomers are mainly ethylene and higher olefins. Cheap rubbers such as EPDM (ethylene – propylene – diene rubber) are frequently used for blending [37]. The miscibility between polyolefins is extremely limited in spite of their great chemical similarity (Fig. 7). However, due to the small interfacial tension the dispersibility is acceptable. Nevertheless, when blends of PP and other polyolefins are being made, they often require a compatibilizing block copolymer to provide the necessary degree of dispersion and to maintain morphological control during processing. The latter is essential for obtaining the desired mechanical properties of the product. Morphological stability can be fur-

ther enhanced by in situ cross-linking of the dispersed phase (dynamic vulcanization) [38]. The main applications of PP and its blends are in bumper systems and dashboards.

Blending with incompatible polymers can be achieved via utilization of a functionalized PP (e.g., by grafting with maleic anhydride). The inherent interfacial tension differences are thus overcome by means of chemical bonds instead of physical adhesion. Blends of PP and PA based on this technique are becoming available [37].

4.2. Poly(Vinyl Chloride)

Poly(vinyl chloride) [*9002-86-2*] (PVC) is the oldest and one of the most important plastics, see → Poly(Vinyl Chloride). Worldwide sales of PVC-based plastics were expected to exceed 13×10^6 t/a in 1991 [39].

PVC has a number of important properties. Its amorphous character and T_g of 75 °C provide an excellent basis for good dimensional stability under load up to 60 °C. Its high polarity gives chemical resistance towards a broad range of solvents and an excellent flame retardancy. However, pure PVC is rather brittle under notched impact conditions and has a limited thermal stability. This last property has successfully been addressed by stabilization techniques. In combination with extensive utilization of blend technology, all conventional processing techniques are now possible for PVC. The abundant presence of strong polar C – Cl bonds make PVC highly suitable for blending. Depending on the concentration of strongly interacting dipoles in the blend components, completely miscible or compatible blends can easily be created. The resulting two-phase morphologies either have a particulate or a controlled delaminated character which controls the final deformation mechanism and its performance in blends of PVC.

Chlorinated polyethylene (CPE) is being used as a blend component to improve the impact and flow behavior of PVC [40]. Compatibilizers are not necessary because the chlorine content can be adjusted to the required compatibility level (35 wt %).

Similar effects are obtained for copolymers of ethylene and vinyl acetate (EVA) and/or ethylene and ethyl acrylate (EEA) [40]. The ester groups interact with the H – C – Cl groups of the PVC and provide the required compatibility if the comonomer content is in the range 30 – 40 wt %. Afterchlorination of PVC yields CPVC which has a higher heat distortion than PVC. The T_g increase is directly related to the chlorine content. CPVC can be used directly or in completely miscible blends with PVC to obtain products having higher T_g s.

Heat-resistant blends can also be obtained by blending PVC with α-methylstyrene – acrylonitrile copolymer (see Section 4.3) which is completely miscible with PVC. Blends are also available with a low T_g. An important example is the blend of PVC with NBR rubber; depending on the nitrile content (critical level, 32 wt % acrylonitrile) a single-phase soft blend or a two-phase impact-modified PVC can be made.

Impact performance can be improved by blending with two-phase impact-modified materials such as ABS and ASA (acrylonitrile – styrene grafted onto acrylic rubber). ASA does not impair weatherability as is the case for ABS. Blends with poly(methyl methacrylate)(PMMA) also show good weatherability performance. Blends of PVC with ABS, ASA, SMA, and PMMA are mainly applied as structural panels in the building industry. The blends are extruded in sheets which are shaped by vacuum forming and press molding. The PVC provides mechanical strength and flame resistance. The second component should provide the required heat distortion temperature and surface properties. The impact modification of PVC and its blends has been increased by using shell – core impact modifiers [41]. These modifiers consist of a very small rubber core, usually a butadiene homo- or copolymer or a poly(butyl acrylate) copolymer, and a "hard shell" made of PMMA or SAN. The shell polymers should have optimal interaction with the matrix PVC. These impact modifiers are not only very effective but also allow high shear processing without deterioration of surface appearance.

Since 1980, PVC and its blends have been able to penetrate the injection-molding market. This was made possible by (1) improving the flowability of PVC by an order of magnitude via reduction of its molecular mass, (2) a greatly improved stabilization package, and (3) improved thermal control in modern molding machines to prevent overheating.

The construction and pipe markets are supplied by PVC – chlorinated polyethylene blends (preferred in the United States) and PVC/EVA blends (preferred in Europe). Both blends have better impact properties than PVC. EVA blends have a slightly better UV stability.

PVC – polyurethane blends are typically used in shoe soles, boots, oil-resistant articles, and calendered coatings by virtue of their UV stability, better thermal stability, and low temperature flexibility.

The additions of small amounts of methyl methacrylate (MMA) copolymers to PVC improves its hot melt strength and reduces melt fracture during processing. Toughness and impact performance can be further improved by blending with MMA- containing impact modifiers. Applications include clear films and window profiles.

4.3. Styrene Polymers and Copolymers

See also → Polystyrene and Styrene Copolymers.

Polystyrene. Polystyrene [*9003-53-6*] (PS) is a very brittle polymer (T_g 100 °C) due to a low entanglement density and lack of secondary transitions. Its usefulness has therefore been limited severely. Poor impact performance has been a major concern and was originally addressed by blending PS with styrene – butadiene (SBR) rubber and subsequently by the now generally established chemical processes for making high-impact PS (HIPS). The block copolymer styrene – butadiene (SB) and terpolymer styrene – butadiene – styrene (SBS) as well as their hydrogenated derivatives, are used as blend components to improve impact properties [36]. The resulting deterioration of the continuous use temperature (T_g reduction by at least 10 °C) can be solved by either making copolymers of styrene with stiffer or more polar comonomers such as α-methylstyrene, maleic anhydride, acrylonitrile, or by blending with a miscible high-T_g polymer such as poly(tetramethylbisphenol A carbonate) or PPE (see Section 4.5). Blending relatively small percentages of PPE ($< 30\,\%$) with PS can be important because it increases the heat distortion temperature by about 1 °C per percent PPE. The mechanical deformation behavior of the blends is also greatly improved by the strong increase in the overall entanglement density. The overall property profile of pure PS is not good enough to make it an interesting blend component for a two-phase system.

PS – polyolefin blends have been developed for packaging purposes bringing the combination of printability, fat resistance, and barrier properties. They again have a controlled delaminated morphology and are typically extrusion materials. However, the more polar and functionalized PS copolymers offer far better blend possibilities and are widely used and discussed below.

ABS. See also → Polystyrene and Styrene Copolymers, Chap. 5. Most commercial acrylonitrile – styrene – butadiene polymers [*9003-56-9*] (ABS) are blends of rubber-impact-modified styrene – acrylonitrile (SAN) copolymer with SAN or other SAN-miscible copolymers. Blending can take place in a suspension or emulsion by coagulation or melt blending during the compounding phase. ABS can be produced by three different techniques: emulsion, suspension, and mass (bulk) polymerization [42], [43]. Each technique delivers ABS with slightly different properties.

The main properties of ABS resins are their good to very high impact strength and low price, coupled with very acceptable mechanical (engineering) properties up to 90 °C. They have an excellent surface appearance and good chemical resistance. In addition, they possess the excellent injection-molding behavior of PS, good melt flow, low mold shrinkage, and an excellent surface finish. Their main disadvantage is their poor resistance toward prolonged exposure to light and heat due to the unsaturated bonds in the butadiene rubber impact modifier systems. UV resistance can be improved by incorporation of acrylic rubbers (ASA). Principal advantages of ASA (SAN grafted onto acrylic rubber) over ABS are high outdoor stability, high resistance to yellowing on exposure to UV light, and good resistance to environmental stress cracking [43]. Other essentials of ABS are its composition-dependent T_g (100 – 120 °C) and a continuous use temperature between 80 and 100 °C. The actual use temperatures for ABS can be en-

hanced up to 118 °C through the incorporation of α-methylstyrene. Its high flammability can be improved by addition of 10 – 15 % bromine-containing compounds and 3 – 5 % antimony trioxide. This results in a higher density, reduced color intensity, lower surface properties, reduced impact performance and processibility, and corrosivity. Blending is a preferred method used to increase the heat distortion temperature, to improve weatherability (unsaturated impact modifiers), and overcome flammability concerns.

ABS – PC and ASA – PC Blends. These blends offer a somewhat higher heat distortion temperature than ABS and good processing characteristics [43–45]. The Vicat softening temperature of 100 °C for ABS can be raised to 130 °C by blending with PC. The carbonate group of the PC and the nitrile group of the SAN matrix interact [46], providing just enough compatability for a useful blend. Optimum interaction between the polymers is obtained for a SAN containing 30 wt % acrylonitrile [46]. Intrinsic compatibility can be enhanced through addition of a third component containing ester groups (e.g., PMMA or PBT) [47], an effect that can be explained in terms of mutual miscibility.

ASA – PC blends offer a better resistance towards UV light and heat aging compared to ABS/PC (Fig. 12). Applications include instrument panels, distribution boxes, and kitchen exhaust hoods [48].

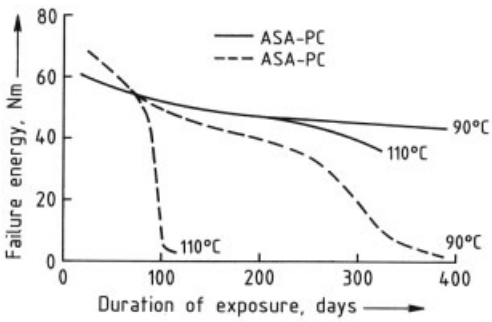

Figure 12. Failure energy of ASA – PC and ABS – PC as a function of the duration of exposure to heat

ABS – PVC Blends. ABS has excellent surface properties with high gloss, but burns readily and completely. PVC (T_g 85 – 95 °C) on the other hand lacks the good impact behavior and processibility of the styrenics. An excellent interaction exists between PVC and ABS, removing the need for compatibilizers. ABS – PVC blends are used in low-cost electrical panels such as covers and enclosures [7].

SMA – ABS Blends. Copolymers of styrene and maleic anhydride (SMA) have glass transition temperatures ranging from 100 to 140 °C, depending on the maleic anhydride content. The materials are rather brittle, but impact-modified resins are known. Impact performance can be improved by blending with ABS. Both resins are compatible and have a good processibility (styrenics). The miscibility window is remarkably small considering the similarity in structures [49]. Instrument panel frames is a large automotive application.

ABS – PA Blends. Both chemical resistance and heat distortion temperature can be improved by blending ABS with semicrystalline polyamides. Typical applications are automotive body panels, connectors, and under-the-hood components [7], [48]. See also Section 4.4.

4.4. Polyamides

Polyamides (PA) are either semicrystalline or amorphous. The polymer chain can consist of aliphatic, cycloaliphatic, and aromatic components. More than 60 different commercial polyamides are known. For further details, see → Polyamides.

Examples are included of the two most important general-purpose polyamides: (1) PA 6 [*25038-54-4*] made by ring opening polymerization of ε-caprolactam, and (2) PA 66 [*32131-17-2*] made from hexamethylenediamine and adipic acid. These aliphatic PAs are mostly used for their combination of wear resistance strength, chemical resistance, creep resistance, and low coefficient of friction. Good dimensional stability can only be obtained for (glass) filled materials which have the disadvantages of warpage and machine wear. Polyamides have a good strength and fatigue resistance. Tensile modulus values of the matrix can be increased by incorporation of aromatic monomers in the main chain. The fully aromatic PAs have a high intrinsic flame retardance.

Decomposition tem-peratures for most aliphatic and aliphatic – aromatic PAs, are 280 – 300 °C, limiting the processing temperature.

The properties of PAs are highly dependent on the interactions of the polar amide linkage. Higher-melting PAs can be obtained by increasing the frequency of amide linkages [50]. The amide linkage interaction between the polymer chains in the crystal structure via hydrogen bonding is particularly strong in the amorphous phase. It also causes water absorption which is related to the methylene – amide ratio. Water absorption decreases the dimensional stability of the polymer, reduces the T_g (plasticizing effect), but increases the toughness. The magnitude of the T_g reduction with respect to relative humidity changes depends on the amount of amide linkages and stiffness of the chain. A major drop in tensile modulus at room temperature results when the T_g drops from 50 to below 0 °C. Other physical properties (e.g., dielectric constant, water permeability) also change considerably.

There are three major reasons for blending PAs:

1) To increase their heat distortion temperature (in a conditioned state)
2) To reduce their moisture sensitivity and thus enhance dimensional stability
3) To improve their dry impact strength (at low temperatures)

Polyamides are very suitable for in situ compatibilization techniques (reactive extrusion) due to the presence of reactive amine or carboxyl end groups. Carbonamide groups are also available for hydrogen bonding.

PA – PA Blends. Melting points of semicrystalline PAs are sometimes very close to their degradation temperatures, creating either a very small or no processing window. Blending aromatic semicrystalline PAs with aliphatic PAs improves their processibility. Aliphatic PAs generally have lower T_gs, but a higher water absorption value which can lower the T_g in the dry state significantly. Aromatic PAs possess more free volume, and have a lower dimensional growth with the same water absorption.

Blends of amorphous transparent PAs have been reported [51]. Predictions of miscibility windows between PAs using binary interaction and intramolecular repulsion models proved that the amide concentration was more important than compositional differences [52–54].

PA – ABS Blends. Blends of PA 6 and PA 66 with ABS combine chemical and abrasion resistance with high impact resistance. The ABS phase provides good dimensional stability up to 90 °C that can be further improved by substitution of styrene in the ABS with α-methylstyrene [55]. These blends are also more warp resistant due to the reduced (semi)crystalline volume fraction. Compatibilization is achieved through hydrogen bonding between amide NH and nitrile CN groups. Anhydride-containing polymers such as SMA have been used to improve compatibilization by covalent bonding and to obtain a smaller particle size [56].

Triax 1000, (Monsanto) is an example of a PA – ABS blend. It is used for a variety of markets such as functional (mainly under-the-hood) automotive components, power tools, appliances, and sports goods. PA – ABS blends have excellent chemical resistance, stress resistance, and toughness due to their co- continuous network.

PA – PC Blends. Blends of PA 6 and PC have been investigated over the full range of compositions, and proven to be immiscible [57]. Chemical reactions are claimed to produce low molecular mass compounds that both plasticize and slightly compatibilize the blend. No commercial PA – PC blends have yet been developed.

PA – PP Blends. Blending PAs with PP raises impact strength, solvent and moisture resistance, and improves strength [34]. The addition of compatibilizers and/or glass fibers is required to obtain a blend with a stable morphology. Compatibilizers are formed in situ by using maleic anhydride-functionalized styrene – ethylene – butadiene – styrene (SEBS) triblock copolymers [58], [59], ethylene – propene copolymers, PP, HDPE, or glycidyl methacrylate copolymers [60] in concentrations of 1 – 30 wt %. Maleic-anhydride-modified compatibilizer intermediates yield a better dispersion (with particle diameters in the range 0.1 – 0.3 μm) and high impact values. Interestingly, material toughness is not enhanced by using smaller rubber particles [59].

PP is usually processed at 200 °C, but PA – PP blends are processed at 250 °C – the presence of PP leads to a readily flowing resin. These resins are used for decorative applications where dimensional stability is much more important than mechanical properties. Examples include automotive applications, sprinkler housings, and housings for industrial tools.

Impact-Modified PA. The notch sensitivity of most PAs can be reduced by blending with 20 – 30 % functionalized impact modifier such as maleic-anhydride-functionalized HDPE [61], [62], or maleic-acid-modified EPDM rubber. PA 66 has an intermediate level of ductility, characteristic of a semicrystalline resin. Chain mobility is reduced by the presence of crystalline regions and relaxation times of polymers are increased. Impact modifiers are used to create overlapping stress fields (optimum conditions at 30 vol % impact modifiers) which enhance ductility [61]. Good adhesion is provided by the maleic anhydride functionalities on the impact modifiers.

The materials are used for all applications where ordinary PAs fall short of required impact strength and toughness: oil pans, fan blades, connectors, and power tool housings.

4.5. Poly(2,6-Dimethyl-1,4-phenylene Ether)

Poly(2,6-dimethyl-1,4-phenylene ether) [25134-*01-4*] (PPE) and its blends have been extensively reviewed elsewhere, → Poly(Phenylene Oxides), but will be briefly summarized. PPE is an amorphous resin (T_g 215 °C) with a number of attractive properties [63]. It has a serious drawback however: its (melt) processing temperature is higher than temperatures at which severe side reactions take place, thus altering the polymer property profile.

PPE – HIPS Blends. The full miscibility of PPE with PS, discovered in the early 1960s [64], [65], led to the development of blends with lower (and thus acceptable) processing temperatures. Polystyrene has a low T_g and good melt processibility.

PPE – PS blends are fully miscible over the complete composition range as illustrated by a single T_g for each composition. This provides the end user with a tremendous flexibility in achieving required maximum use temperatures. Resulting heat distortion and Vicat B temperatures are discussed in [63], → Poly(Phenylene Oxides), Chap. 3.1.

The miscibility of PPE with styrene copolymers and polymers of styrene derivatives, e.g., poly(α-methylstyrene) [64–66] has been studied. Many of these blends also demonstrate full miscibility, probably due to the interaction of the electron-deficient methyl groups of PPE with the electron-rich phenyl rings of PS [63], [67]. The reduction of impact performance of the ductile PPE by blending with brittle PS (Fig. 13) can be attributed to a reduction of the entanglement density and some loss of free volume [63], [67]. Impact performance is greatly improved by incorporation of impact modifiers. Thermograms of commercial blends have two glass transition temperatures, with one of them being attributable to the rubber impact modifier [68].

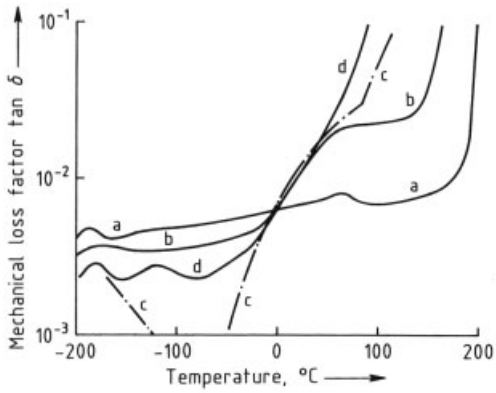

Figure 13. Impact strength of PPE – PS blends as a function of temperature
a) PPE; b) PPE – PS (70 : 30); c) PPE – PS (30 : 70); d) PS

Impact strength increases with PPE molecular mass. Commercial PPEs have weight-average molecular masses in the range 30 000 – 40 000. The rubber particle size is also important. Since their commercial introduction in 1964 by General Electric, PPE – HIPS blends have become the engineering polymer blends with the largest annual sales. They are used for applications ranging from automotive interior parts including dashboards and instrument panels, to business machine housings and chassis; from electrical conduits and lighting covers to gears and sprockets.

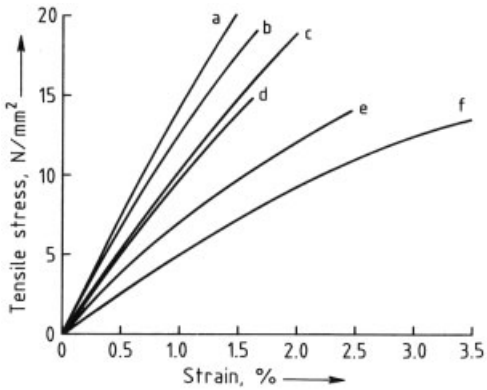

Figure 14. Creep behavior of PPE – PA 66 (Noryl GTX 900) at 23 °C and 50 % RH
a) PPE – PA 66, 1 h; b) 100 h; c) 5000 h; d) PA 66 1 h; e) 100 h; f) 5000 h

PPE – PA Blends. PPE – PA 6 and PPE – PA 66 blends have been extensively discussed elsewhere, see → Poly(Phenylene Oxides), Chap. 3.2. In these blends, the finely dispersed PPE phase raises the continuous use temperature. The tensile modulus is maintained up to almost 205 °C with only a minor drop over the temperature range from 0 – 50 °C due to the T_g of the amorphous PA phase. The moisture sensitivity of the system is also reduced. The PA 6 and PA 66 phases provide easy processibility and chemical resistance. These PAs can be easily replaced by PAs with higher glass transition and melting temperatures [69–71]. Rubbery impact modifiers have been added to improve low-temperature impact performance [69–71]. These blends exhibit mechanical properties which are less moisture-sensitive than PA 66 as shown by the stress – strain curves (Fig. 14). At a stress level of 7 MPa, PA 66 exhibits three times as much creep as the PPE – PA 66 blend.

Blends of PPE and PA have been specifically developed for exterior automotive parts with a Class A surface quality (e.g., fenders of the Renault Clio 16 S). Other applications are rear quarter panels, and hood materials for lawn and garden tractors.

PPE – Polyolefin Blends. The usefulness of unmodified PPE is limited by its notch sensitivity and difficult melt processibility. Addition of polyolefins improves flow by modifying the melt rheology of the system but sometimes lowers the heat distortion temperature. New blends have been developed in which PPE is combined with special rubbers (e.g., high molecular mass polyoctanomer [72], [73]) that have a high impact strength, a high softening temperature up to 205 °C, and facile processibility. These PPE – polyolefin blends have a high gloss and are used in metallizable head lamp reflectors. Other application areas are dimensionally stable exterior automotive body panels, and major appliances.

The intrinsic flame retardancy of PPE has been enhanced with additives resulting in a new, halogen-free, flame-retardant material that meets the most stringent low smoke [73], low toxicity, and low corrosivity standards. The key factor is increased char formation under fire conditions which reduces the amount of volatiles [73]. Potential applications of this material (Noryl Xtra LS) are in office buildings, trains, and tunnels.

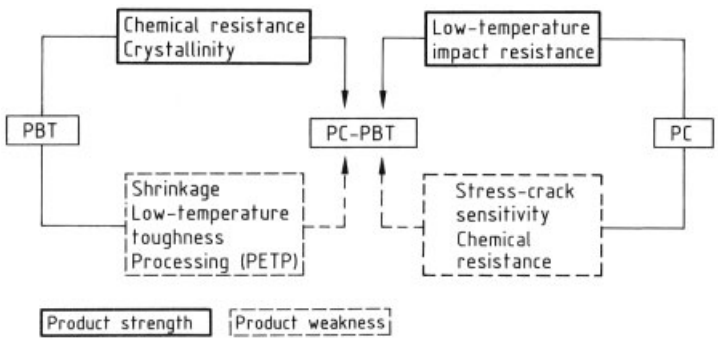

Figure 15. Blend motives for PC and PBT

4.6. Polycarbonates

The polycarbonate of bisphenol A [*25037-45-0*] (PC) is a linear polymer with a low rate of crystallization. See also → Polycarbonates. Thus, it is regarded as an amorphous polymer (T_g ca. 145 °C). Its maximum use temperature is approximately 130 °C and, because of its excellent thermal oxidative stability, this almost coincides with its continuous use temperature (125 °C). PC is noted for its high impact resistance, especially at low temperature. This is due to the high segmental mobility of the polymer chain (gamma transition at − 100 °C [74]). Polycarbonate suffers from stress – crack sensitivity coupled with poor chemical resistance. The latter prevents PC from being used in automotive applications where frequent contact with gasoline and oils would ruin its material properties. Chemical resistance can be improved by blending with semicrystalline polyester resins. The stress cracking sensitivity of PC can be overcome by blending with ABS or ASA.

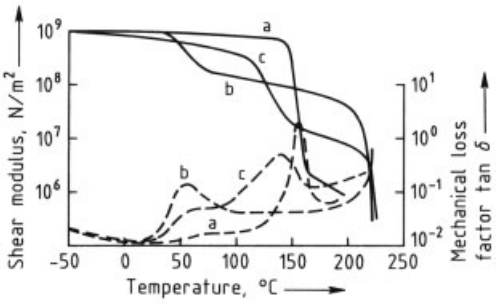

Figure 16. Dynamic mechanical properties of PC, PBT, and PC – PBT blends : shear modulus (— — —) and mechanical loss factor (— — —) as a function of temperature
a) PC; b) PBT; c) PC – PBT (55 : 45)

PC – Polyester Blends. The advantages resulting from blends between PC and PBT are summarized in Figure 15. Blends of PC with polyesters are based on a partial miscibility of both components in the amorphous phase [75–77]. Thermal and dynamic mechanical analysis shows a sharp T_g transition for PC at 150 °C (Fig. 16), and a sharp melting point at 225 °C with a broad glass transition for PBT due to the spectrum of relaxation times of the crystal – amorphous polymer interphase. Partial miscibility of PC and PBT is indicated by a shift of the PC T_g to 141 °C in a PC – PBT blend.

Blending both resins results in a complex co- continuous morphology. Resistance towards chemicals (e.g., oils and lubricants) is strongly enhanced as compared to PC. However, the impact properties of these blends are still insufficient for automotive applications. Impact modification with a rubber compound (frequently core – shell impact modifiers) solves this problem [78], [79].

The key to morphology control and thus mechanical properties is control of transesterification. Care should be taken to prevent these reactions (Figs. 17 and 18) catalyzed by residual polycondensation catalysts present in the PBT [79–81]. Addition of proprietary additive packages is commonly practised. In addition, processing temperatures of 270 – 290 °C should not be exceeded.

One of the most successful applications of the blend of PC, PBT, and impact modifier (Xenoy, GE Plastics) is a car bumper, developed for the Ford Sierra in Europe, and transferred to the Ford Escort in the United States. This bumper withstands 5 km/h impact at high and low temperature.

PC – PETP blends have similar properties to their PBT analogs [7], [82], [83]. The lower crystallization rate of PETP allows the development of transparent materials for use in medical applications by virtue of their good dimen-

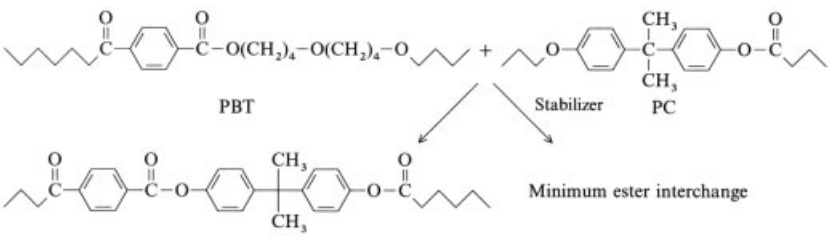

Figure 17. Ester interchange between PC and PBT

sional stability, clarity, and low water absorption. The products have a better chemical resistance than PC and are impact modified to give acceptable low-temperature impact performance. PC–PETP blends flow better than PC, and are injection-moldable. Unreinforced PETP cannot be injection molded due to its low rate of crystallization.

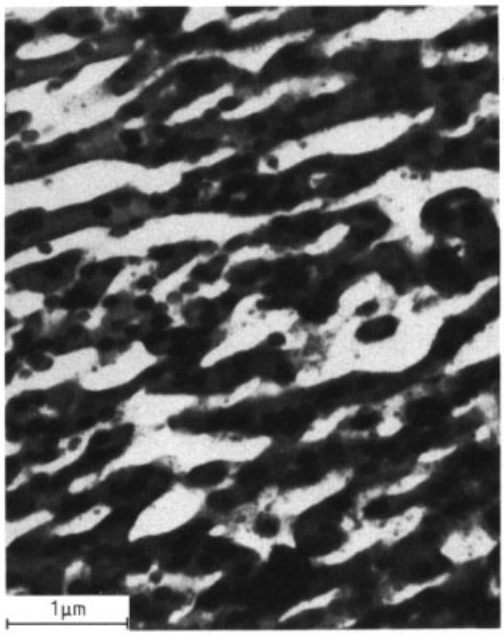

Figure 18. Morphology of impact-modified PC – PBT blend (Xenoy, GE Plastics)
The darker areas are PC, the black dots are impacted modifiers, the remainder is PBT.

PC – Polyolefin Blends. Despite its good toughness, PC has a distinct ductile to brittle transition which is dependent on temperature, thickness, test speed, stress configuration, and molecular mass. The dependency is reduced if PC is blended with low levels (4 – 8 %) of LDPE. This shifts the ductile – brittle transition of PC from about 0 °C to below − 20 °C. Moreover, the effect of part thickness on notch Izod impact strength is reduced [84].

PC – ABS Blends. Polycarbonate blends with ABS resins have been successfully used in automotive interior and exterior applications [43], [85]. The acrylonitrile content of the ABS plays a key role in determining the miscibility of the SAN phase with the PC phase. ABS has no low-temperature impact performance, whereas PC does through its gamma transition at − 100 °C and ample free volume [74]. The impact values of PC – ABS blends are directly correlated with their heat distortion temperatures. The presence of the rubbery impact modifier slightly reduces the tensile modulus. The HDT values are hardly affected at modulus values above 3 GPa. These are important values because they are required for high-speed impact tests for automotive applications. In addition, PC – ABS blends provide paintability and better UV stability if PC is the continuous phase [86–89]. The addition of UV absorbers to the (transparent) PC phase improves the protection of ABS against the damaging effects of UV light (see Section 4.3). PC–ASA blends have superior weather resistance, however.

PC – ABS swells in aromatics, ketones, esters, and hydrocarbons. It has good electrical insulating capability. Major applications are instrument panels [7] and business equipment housings [90].

4.7. Polyesters

Thermoplastic polyesters have a low moisture absorption and are chemically resistant apart from being susceptible to hydrolysis at high temperature in alkaline environments. They have a low dimensional stability due to their low T_g; glass-filled materials are required to achieve acceptable use temperatures. Impact performance for these highly filled materials is poor, but can be improved by blending with impact modifiers. The most important polyesters are poly(ethylene terephthalate) [*25038-59-9*] (PETP) and poly(butylene terephthalate) [*24968-12-5*] (PBT). See also, → Polyesters PBT crystallizes within 1 – 3 s whereas PETP crystallizes within 15 – 20 s. PBT's higher melt flow and faster crystallization lead to a shorter cycle time. Special processing conditions are required if (partially) crystalline PETP is to be obtained: high mold temperature (120 – 140 °C) and long cycle times (2 min). The advantages of PETP over PBT are its higher melt temperature (256 vs. 226 °C) and strength.

PBT – PETP blends have enhanced surface quality and reduced warpage when reinforced with glass fibers. Both qualities are attributable to the lower rate of PBT crystallization caused by the addition of PETP. The weight ratios used in PBT – PETP blends are usually 85 : 15 to 75 : 25. HDT values are not affected, but costs are reduced due to the lower price of PETP. Typical ap-plications include hot-appliance housings (e.g., toasters, irons, and frying pans).

PBT – PC Blends. Addition of PC to glass-filled PBT grades increases impact strength values and improves thermal expansion, flow, and flatness (reduced warpage). The chemical resistance of these blends is slightly lower than that of a PBT compound. Core – shell impact modifiers are used for toughening [78], [79], [91]. The importance of interfacial adhesion with the matrix has been proven [91], [92]. Applications include grill doors, air cleaners, and appliance components where improved dimensional stability over PBT is required.

Miscelleneous Polyester Blends. Noteworthy are blends of PBT with copolyester elastomers containing short crystallizable PBT blocks [93]. The tensile strength and modulus of these compounds are increased upon blending with PBT while retaining its flexibility. Excellent surface quality and improved paint adhesion qualify these resins for bumper fascias. Similar uses hold for PETP – acrylic and PBT – acrylic materials. Toughened PBT and PETP blends are used to extend existing product lines by varying the amount and type of impact modifier. Ignition compounds, housings, and windshield wipers are typical automotive applications; flame-retardant grades are used for connectors and switches.

4.8. Polyetherimides

Polyimides are condensation polymers derived from bifunctional carboxylic anhydrides and primary diamines. Polyetherimides [*61128-24-3*] (PEI) contain an ether group in either or both the aromatic dianhydride or aromatic diamine building block. Ultem polyetherimide (GE Plastics) produced from bisphenol A bisether dianhydride is the only commercially available thermoplastic polyetherimide [94] (see also → Polymers, High-Temperature, Chap. 2.2.). It has a glass transition temperature of 217 °C, and can be subjected to conventional melt-processing techniques. Its thermal stability is maintained up to ca. 500 °C under nitrogen [95].

In commercial PEI grades, additives (e.g., reinforcing fibers) and modified base resins are used to increase the glass transition temperature and chemical resistance. Primary reasons for blending are therefore to improve processibility, impact strength, and chemical resistance. PEI is used mainly in under-the-hood automotive applications, head-lamp reflectors, injection-molded printed circuit boards, trays for microwave ovens, and high-performance composite sheets with PEI as the matrix.

PEI – PC Blends. PEI – PC blends combine the high temperature resistance and intrinsic flame retardance of PEI with the good impact performance and ease of processing of PC. Interactions are strong enough to form an excellent blend without using compatibilizers. The high continuous use temperature of PEI is not decreased because PC is stable to thermal oxidation up to ca. 175 °C. Blending with PC reduces the price of the material and its notch sensitivity.

Two grades of PEI – PC thermoformable sheets exist, one is targeted to household applications (e.g., Tupperware microwave cookware), the other to aircraft interior components where the material satisfies all regulations pertaining to flammability, smoke, and toxicity. PEI – PC is used for aircraft window housings and seat backs.

Thermoformable PEI – PC products are also used as high-temperature paint masks in the automotive industry to prevent overspray. The high heat resistance and excellent dimensional stability enable the material to withstand the temperature conditions in the painting operation.

PEI – PEEK Blends. PEI is an amorphous resin. Its chemical resistance can therefore be improved by blending with a semicrystalline resin. PEI is miscible with the amorphous phases of polyetherketone [*27380-27-4*] (PEK) and polyetheretherketone [*31694-16-3*] (PEEK) over the entire composition range. Crystallinity

and crystallization are not affected by incorporation of up to 40 wt % PEI [96]. The product range of these materials can be further enlarged by copolymerization of the PEI structure leading to enhanced inherent properties.

Markets for PEI – PEK and PEI – PEEK blends have not yet been fully exploited [97]. A 70 : 30 PEI – PEEK grade developed by Amoco Performance Polymers was not commercialized due to competition with the liquid crystalline polymer Xydar [7].

4.9. Polysulfones

Polysulfones [*25135-51-7*] (PSU) and polyethersulfones [*25667-42-9*] (PES) are amorphous high-T_g polymers with excellent thermal and oxidative stability (see also, → Polymers, High-Temperature). Their environmental stress-cracking resistance against most liquids encountered in the food and automotive industries is good because of the polar sulfone group.

Polysulfones are, however, expensive, demonstrate a severe notch sensitivity, and have a limited chemical resistance resulting in stress cracking sensitivity under certain conditions. Both problems can be overcome by blending with ABS (Fig. 19), PAs, PETP, and PBT. Chemical resistance is strongly improved if the semicrystalline resin constitutes the continuous phase. The principles and objectives followed for developing these blends are the same as those for PC – ABS, PC – PETP, and PPE – PA blends. Polar interactions are possible between the carboxyl groups of PSU and the nitrile groups of ABS.

Applications include household water faucets, electric, and electronic applications. PSU – PBT replaces PBT in selected connectors and switches due to its higher strength [7], [48].

4.10. Polyoxymethylenes

Polyoxymethylene [*9002-81-7*] homo- and copolymers are highly crystalline materials well known for their hardness, low friction coefficient, chemical stability, and relative brittleness; see also → Polyoxymethylenes. They are highly flammable (LOI 16) and their thermal stability is limited to 210 °C. In view of their melting temperature of 182 °C (homopolymer), careful processing is required.

Blends of polyoxymethylene homo- and copolymers are known. Silicone oils and polytetrafluoroethylene have been added to POM to enhance the high slip characteristics. Impact-resistant grades of POM use polybutadiene, acrylic rubber, grafted synthetic rubber, polyolefins, and polyurethane [98].

Impact-resistant POMs, comparable to supertough nylon, have been produced by blending with compatible linear thermoplastic polyurethanes. They demonstrate a special morphology since the polyurethane phase forms an interpenetrating network structure in the POM matrix.

Applications are hinges and connectors that demand improved low-temperature flexibility and impact strength.

4.11. Polyarylates and Copolyarylate – Carbonate Blends

In principle the polyarylates (PAr) comprise all fully aromatic polyesters. In practice this term denotes polyesters derived from bisphenol A and mixtures of isophthalic and terephthalic acids in ratios between 70 : 30 and 30 : 70, but most frequently 50 : 50 [99], [100]. Due to their fully aromatic structure, the PArs are amorphous, transparent, practically colorless, and heat resistant. The T_g of the commonly used PAr containing bisphenol A and isophthalic/terephthalic acid in a weight ratio of 50 : 50 is 184 °C. This polymer is soluble in chlorinated hydrocarbon solvents, has a reasonable hydrolytic stability, good mechanical properties, strength, modulus, impact resistance, and excellent weatherability (despite yellowing). The polyarylates are intrinsically self extinguishing in flammability tests [101], [102]. Processibility is, however, marginal in view of the high processing temperatures (320 – 360 °C) required. Copolyester carbonates (PPCs) have therefore been developed with T_g's between those of PC (145 °C) and PAr (184 °C). PPCs are aromatic copolyesters in which part of the ester groups are substituted with carbonates. This group of polymers fills the T_g gap which existed between PC and the polysulfones [103].

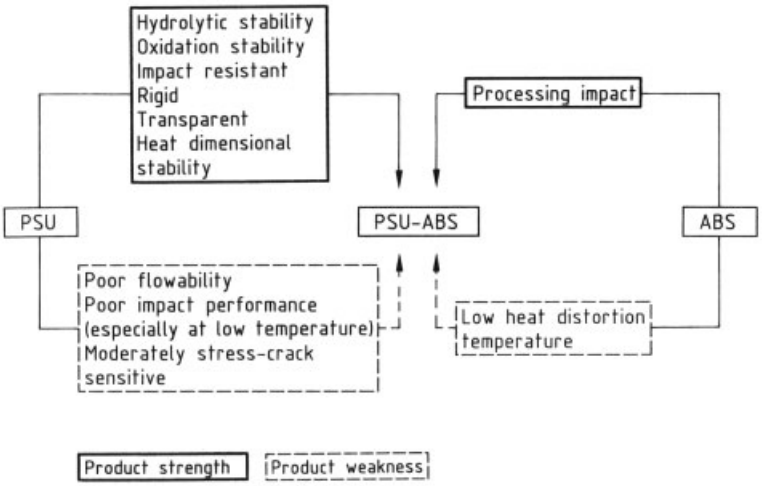

Figure 19. Blend motives for PSU and ABS

Polyarylates are mainly used in applications demanding high heat stability and transparency such as lenses and medical equipment (high-pressure steam and radiation resistance needed for sterilization). Since the ultimate thermal dimensional stability of polymers under loading conditions is determined by the T_g of the matrix polymer, polyarylates and copolyarylates contribute significantly to the engineering property profile of blends.

Many PAr blends, especially those with semicrystalline aliphatic – aromatic polyesters such as PETP, PBT, and poly(cyclohexane terephthalate) (PCHT) have been made but very few have been successful. These polyester – PAr blends show the same features as the polyester – PC blends. PETP and PBT blends are partly miscible with PAr, whereas PCHT is completely miscible over the whole composition range. Furthermore, under melt-processing conditions, transesterification reactions catalyzed by polymerization catalysts rapidly transform the physical blend into a homogeneous copolymer with an averaged T_g value. Only by adding phosphites can the catalytic system be deactivated, yielding a usable blend with different characteristics from the existing PPCs. This holds particularly for the PETP and PBT blends. The ester interchange reaction does not need to be stopped in PAr – PCHT blends due to its complete miscibility [104–107].

PETP – PAr blends are used in the automotive industry for their improved chemical resistance and heat distortion temperature over PC and PC – PETP blends. They are also used in the lighting industry for reflectors. PCHT – PAr blends are used in medical applications because they have a very good resistance to yellowing when exposed to high-energy radiation for sterilization [103].

4.12. Poly(Phenylene Sulfide)

Poly(phenylene sulfide) [*9016-75-5*] (PPS) is a versatile thermoplastic engineering resin (see → Polymers, High-Temperature, Chap. 5.). It possesses high temperature resistance, inherent flame resistance, good mechanical properties, and high chemical resistance. The mechanical properties and dimensional stability of parts are determined by the crystallinity and orientation of the polymer during processing. Unfortunately, PPS is very brittle resulting from stiff spherulites and weak interspherulitic boundaries. Research is therefore directed at improving ductility while maintaining overall crystallinity levels.

Two types of PPS are commercially available: (glass) filled grades and blends with polytetrafluoroethylene (to improve moldability and crack resistance). A large number of other blends have been reported (e.g., PPS – PEI [108] and PPS – PETP [109]) but have not been introduced commercially.

5. General Application Technology

5.1. Processing of Blends

The utilization of thermoplastics, including blends, involves processing and adjustment of the bulk rheological properties of each commercial plastic grade to the requirements of the processing technique (e.g., injection molding, extrusion, blow molding). Adjustment is only possible if an acceptable processing window exists within the molecular mass constraints of the polymers which are needed to obtain the desired mechanical properties. The processing window is determined by the thermal stability and the processibility (viscosity) of the melt. In blends, the maximum melt temperature is determined by the stability of the least stable component.

The rheology of a blend system is determined by its components. Blend processibility can only be predicted to a limited extent. The rule of logarithmic additivity only applies to completely miscible systems of homologous polymers such as mixtures of the same polymer with different molecular masses (Eqs. 10 and 11), all functions having a molecular mass > entanglement molecular mass.

$$log\eta_b = \Phi_1 \log\eta_1 + \Phi_2 \log\eta_2 \tag{10}$$

$$\eta_b^{1/3.4} = \Phi_1 \eta_1^{1/3.4} + \Phi_2 \eta_2^{1/3.4} \tag{11}$$

where η is the viscosity, Φ is the partial volume, and the subscripts b, 1, and 2 denote the blend, component 1, and component 2, respectively.

Even for a completely miscible blend system (e.g., PPE – PS), corrections have to be made for the "positive deviation" which is associated with the mixing enthalpy (Eq. 12):

$$\log\eta_b = \Phi_1 \log\eta_1 + \Phi_2 \log\eta_2 - \frac{\Phi_1 \Phi_2 \Delta H_{\text{mixing}}}{2.45RT} \tag{12}$$

This enthalpy factor depends on the temperature (LCST behavior, Section 3.1). In miscible or partly miscible blends with processing temperatures close to the LCST temperature this can have extremely important effects, such as shear-induced phase separation (temperature rises due to the introduction of mechanical energy via the processing equipment) leading to complete loss of morphology and mechanical properties. Furthermore, the elastic and viscous parameters that depend on temperature and shear rate also contribute to the rheology characteristics of a polymer, and are different for each polymer.

In impact-modified multiphase blend systems, the prediction of the rheology characteristics by means of (empirical) rules becomes highly unsatisfactory. Simple approaches based on the Einstein relation for dispersions of solid particles in a liquid are followed by assuming that two-phase systems behave as a dispersion of fixed particles of different deformability (from solid to rubbery). In more complex two-phase blends where the morphology of the dispersed phase is not fixed and delamination is possible, a combination of parallel and series model of rheological effects is used with an interfacial layer correction.

The impact modifiers interact strongly with the matrix to create a thixotropic melt behavior in which the viscosity value at zero shear rate is much higher than in the unmodified systems. These characteristics can be very useful in creating the higher melt strength needed for extrusion and blow molding. Their presence may, however, also lead to poor surface appearance and knitline problems under excessive shearing forces.

In conclusion, the rheology of polymer blends is extremely complicated to predict, but commercial products are tuned to the requirements of most standard processing methods. The tolerance level for deviation from the recommended processing conditions is not unlimited.

5.2. Design Technology

The design considerations of a thermoplastic part involves functional factors related to production and assembly, and material factors related to the performance of a material in service.

Design procedure can be subdivided into the selection of the production process (e.g., injection molding, extrusion), resin characteristics (mechanical and electrical properties), flow analysis, stress analysis, long-term considerations (creep, fatigue, stress relaxation), appearance of a part, its assembly (including automation), and recycling. Computer programs aid the designer in developing a part capable of meet-

ing all the mechanical requirements, while maintaining its production feasibility [28], [110].

The current emphasis on recycling demands that produced parts can be easily disassembled from final applications. In the case of cars, routes for facile dismantling and separation should be conceived in the designing step (see Section 6.1). Product designers and engineers must therefore consider all aspects of the primary and secondary life cycles.

Blends require special consideration of flow (viscosity), weld lines, processing temperature versus phase behavior (see Section 3.1), flow versus continuous phase, and high shear conditions. Thermodynamically controlled, shear-induced phase separation is possible leading to delamination and morphology changes (Section 5.1). Care should be taken to avoid high shear conditions.

The dependence of viscosity and shear rate on temperature is also very important for blends because the ratios between the blend components influence blend morphology and determination of the continuous phase (see Section 5.1). Prediction of rheology and flow behavior is more complicated than for single thermoplastic resins. Weld lines are an issue at positions where two flows encounter each other [111]. Design should result in placing weld lines in nonstressed regions.

6. Future Developments

6.1. Recycling

Recycling of plastics and polymer blends has drawn public attention, see also → Plastics, Recycling. Concern regarding plastic waste has risen because of the large volume of plastics (20 – 30 %) in municipal waste. It corresponds, however, to a weight fraction of only 8 % for which packaging materials are mainly responsible. The volume of waste must be reduced, particularly in densely populated areas. Thermoplastic blend producers and end users are faced with finding commercially effective recycling strategies. This is particularly acute for the automotive industry in Germany, where recycling laws demand the recycleability of 25 % of plastic car parts by the end of 1994.

The use of smaller quantities of better, more durable blend products may or may not have a positive overall environmental effect. One of the solutions to the waste problem practised in Japan is energy recycling by burning (70 %) and deposition of solid waste (30 %). This simple solution will soon disappear in the western world for political and economic reasons – waste disposal will become very expensive due to increasing tariffs. Other options (i.e., reuse of plastics in other applications) must therefore be considered.

The most important factor in plastic recycling is the purity and compatibility of the waste. Minimum waste is generated in plastic manufacturing and compounding facilities due to economic reasons; reworking of plastic waste is a standard procedure. Separation of the different plastics used facilitates backfeed and rework programs. Multilayer coathanger extrusion and two-component molding facilitate the usage of off- color and slightly off-grade materials.

The real problem starts with postconsumer plastic waste where purity and compatibility requirements are no longer fulfilled. Separation, sorting, and cleaning of the plastic waste from the bulk waste are necessary. This is very expensive and only economically feasible for a small number of cases. Any recycled material derived from plastics with different colors and grades therefore represents a downgrade product. In addition, this approach will only succeed if the recycle concept is economically viable, e.g., in the case of high-value engineering plastics.

Reusing the chemicals from which the polymers were made is another route. Thermal, pyrolytic, hydrolytic, or glycolytic processes are used to degrade the polymers [112]. This approach is only applicable in a few cases and has been proven technically feasible with PETP bottles [112]. However, the process economics are very limited.

Finally, the fuel value of the plastics should be considered. Polymers have a considerable caloric value being (mainly) organic materials. The energy in the material can be reclaimed by burning and then converted to electrical energy or steam.

Already 20 % of all plastics used can be considered blends. Blend thermodynamics, morphology, and careful processing are critical for obtaining the required/desired mechanical

properties (Chap. 3). Lower-grade applications should be found for recycled blends. Typical examples are blends of postconsumer PE with PA 6 [113], scrap rubber [114], and recycled PETP with PE [115], [116] and PC [117].

Two important principles govern the role of engineering blends in recycling strategies:

1) Designing and developing better, more durable products via blend technology.
2) Combining compatible resin systems (i.e., minimization of the number of different resins in a module for various applications). The target is to obtain materials from one product family which are mutually compatible (Section 3.2). Only these compatible materials can be combined to yield economically viable, useful materials (Fig. 20).

An example of a module is a dashboard structure made of a glass-filled, injection-molded PPE – HIPS blend, which could include, for example, a large proportion of recycle material from dashboards. The air ducting is made of compatible, blow-molded PPE – HIPS and is mounted on the carrier unit by means of PPE – HIPS clip fasteners. The complete unit is overmolded in PPE – PS foam to form the outer surfaces and to provide sound insulation. A soft-feel foil is used to give a "luxury" look.

This principle influences the material selection step of the design process. Functionality (i.e., good performance) and cost are no longer the most important selection criteria because they may lead to a combination of incompatible materials. Instead compatibility considerations dominate. The flexibility of blend technology allows the creation of materials that satisfy compatibility and use criteria. It is conceivable that concessions may be made regarding functionality, i.e., although not all requirements are met, certain materials are chosen to give the cheapest module costs.

Recycling materials after their life cycle involves several steps [118]: dismantling, transportation of parts to general assembly, cleaning, processing of reground material into an application. The most important issues are identification of the numerous blends in similar applications, definition of separation procedures in which contaminants are removed, and facile disassembly steps during designing. Costs involved with collection, separation, cleaning, and reprocessing materials mean that this is only a viable concept for blends of engineering plastics with their high property level and high profit margin.

The cascade model (GE Plastics) envisages a cycle of applications for a material in keeping with the deterioration of mechanical and other properties.

Developments are under way to compatibilize the coatings, paints, and other materials used for aesthetic purposes with the main component.

Examples of cascade models for thermoplastics employed by Opel combine PP from battery cases and bumpers to make new fender liners, bumper carriers, and air-filter housings, and convert PC – PBT from old painted bumpers into new spoiler supports [119].

6.2. New Developments

Several promising new developments can be mentioned. PPE – PS blends are used to make expandable beads, which give strong low-density (down to 20 g/dm^3) foams with a heat distortion temperature $> 100\,^{\circ}$C, i.e., a sterilizable, re-usable material [120].

The use of PPE as an organic, heat-resistant, flame-retardant filter that does not affect mechanical properties at high filling ratios has been demonstrated for PA – PPE blends, and can be applied to other blend systems. These new blends include a PPE – PBT blend under development (trade name Gemax, GE Plastics), which has an even lower moisture absorption and thus higher dimensional stability than PA – PPE [121]. The material is targeted for automotive and appliance markets. Blends of chemically modified PPEs will also find increasing use.

Hybrid systems are related blends and consist of thermoset networks filled with heat-resistant engineering thermoplastics [122]. High-performance thermoset materials are used in structural applications where dimensional stability is of the utmost importance. Their temperature stability is determined by the glass transition temperature, whereas their mechanical stability above this temperature is controlled by the network density. Stiffer, more heat-resistant materials can be obtained by using stiffer building blocks or increasing the network density. However, this also increases the brittleness.

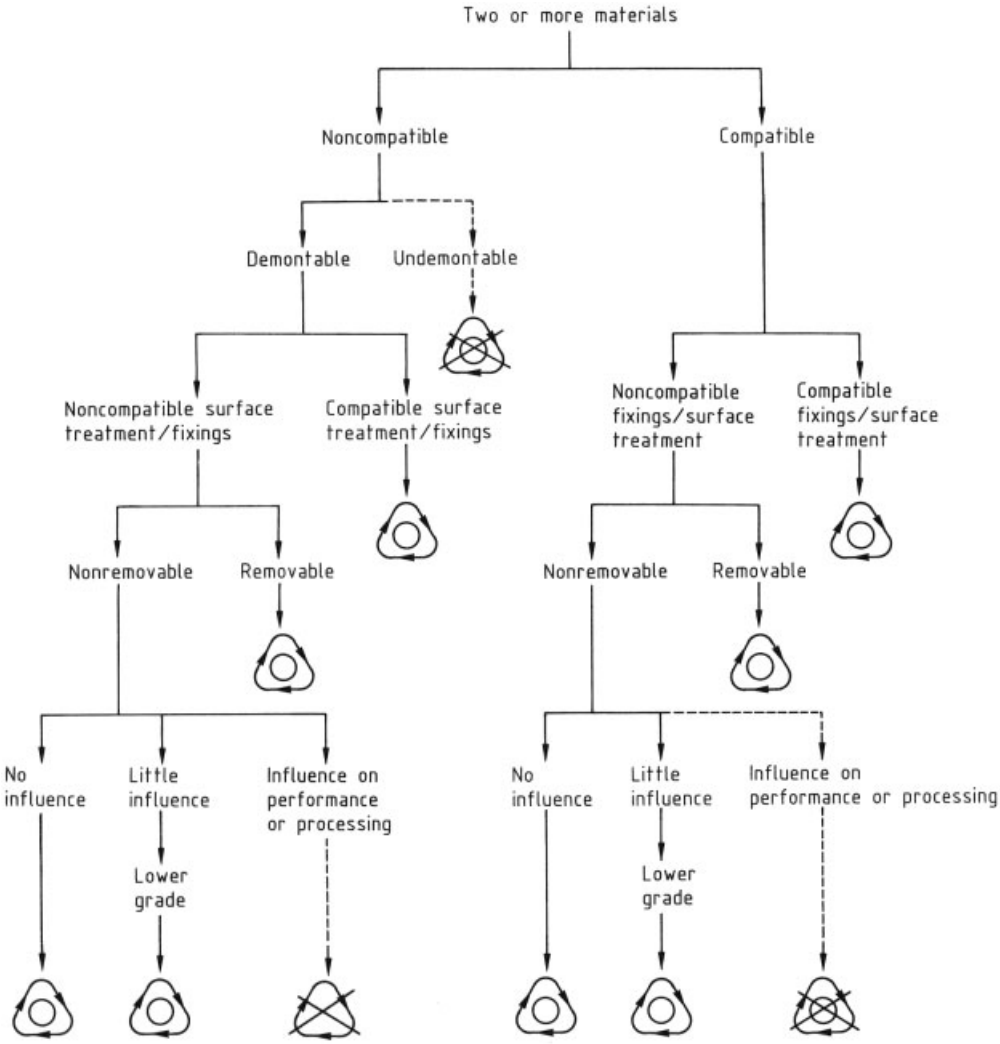

Figure 20. Generalized scheme illustrating the importance of resin compatibility in recycling

Incorporation of heat-resistant engineering thermoplastics improves impact behavior while maintaining dimensional stability at high temperatures. Dissolution of these resins in epoxy systems allows easy processing. The thermodynamically controlled phase separation that takes place during the curing of the epoxy resin leads to a high flexibility in controlling the ultimate network density and thus mechanical properties. Commercially available materials include PEI – epoxy systems (Brochier, France) [123–125] intended for metal-laminating structures in the aircraft industry, and PPE – epoxy systems (General Electric, USA) [126], [127]. The latter systems provides another solution to the poor melt processibility of PPE (Section 4.5); it combines the low viscosity and easy processibility of thermosets with the toughness of thermoplastics. A commercial product (Getek, GE Electromaterials) is used as a circuit board material, the presence of PPE allows compliance with highly demanding techniques used for surface mounting and impedance control. A commercially available PPE has been functionalized with head and tail groups to provide optimal adhesion between the PPE phase and final cured epoxy resin. The PPE volume fraction determines whether the product is an epoxy-filled PPE matrix or a con-

tinuous epoxy matrix with a highly dispersed PPE phase.

A new reactive polymer system is based on oxazolines. Copolymers of styrene with a functional oxazoline monomer can react with carboxyl-, amino-, or anhydride-functional polymers to produce graft copolymers [128]. This reactive PS has been used to compatibilize the combination of PPE and an ethylene – acrylic acid copolymer, producing a high-impact engineering blend with excellent strength and processing properties.

Also new are blends of thermoplastics with 5 – 25 % liquid crystalline polymers. The resulting "molecular composite" can be regarded as a material in which the rigid crystalline polymers are reinforced at a molecular level. Liquid crystalline polymer blends have been reported for PC [129], [130]; PEI [129]; PP [131]; PBT [132]; PPE – PS [133]; ABS [134]; PSU [135]; PETP [136], [137]; and PA 6 [138]. These blends can be processed by conventional techniques such as extrusion, injection molding, and fiber spinning. Nearly all combinations of liquid crystalline polymers lead to an immiscible polymer blend. Consequently the mechanical properties of the blends are determined by the morphology of the liquid crystalline polymer domains in the thermoplastics. This morphology depends on composition, processing conditions, viscosity ratio of the component polymers, and the rheological characteristics.

Efforts are being made to use more biodegradable plastics for packaging. Commercial blends of PE with starch have been introduced by Feruzzi. The biodegradability of blends of polycaprolactone and LDPE or PP with cornstarch has been reported [139].

The development of thermoplastic blends for high-temperature composites (at $> 200\,^{\circ}\mathrm{C}$) to be used in structural applications is described in [140]. Polymers with high temperature stability tend to be intractable or brittle, whereas easily processible polymers usually fall short of the projected temperature requirement. Luckily, many high-temperature, aromatic and heterocyclic polymers form miscible blends with each other which permits balancing of thermal response and processibility. Poly(2,2′-metaphenylene-5,5′-bibenzimidazole) (PBI), for example, is miscible with PEI and with PAr; PEI is miscible with PEEK and PEKK.

Blending is also applied to improve the processibility of conductive polymers. Examples are Versicon (Zipperling) intended for transparent electrostatic coatings and corrosion prevention; electrically conductive blend or polyaniline and PVC with EMI properties; and conductive blends of poly(*p*-phenylene vinylene) and polyacrylamide [141]. Blends of (conductive) polyaniline in a matrix of elongated/stretched ultrahigh molecular mass PE are being commercialized by Neste.

6.3. Outlook

Blend technology will accelerate penetration of thermoplastics in markets where its presence is still limited [142]. Designers will have to be made aware of the advantages of working with thermoplastics instead of conventional materials such as wood and metal: their superior performance in a large number of applications, their versatile property profiles, the design freedom, and recycle options. The construction industry is an area with enormous potential for thermoplastic blends and composites. Short-term growth will be driven by applications such as windows, electrical systems, and skylights, followed by the replacement of wood, steel, and concrete. Advantages are lighter constructions, labor savings, and low maintenance costs. A number of possibilities are demonstrated in GE Plastics "Living Environment" concept house and a similar principle developed by Neste. They show the possibilities of thermoplastics and their blends, as well as the use of recycle materials in novel design concepts.

A breakthrough in the automotive market would be large exterior parts and structural components [143]. Two factors limit the use of thermoplastics in body panels. Firstly, cars will continue to feature steel – plastic body construction throughout the 1990s, requiring plastics to withstand vehicle assembly processes designed for steel. Secondly, car makers must compare the overall cost of plastics to steel including tooling, assembly, and customer value rather than on a raw material basis.

Miniaturization is the main driving force for the electronics expanded use of synthetic poly-

mers in the electronics industry. With the development of new molding techniques for high-temperature resins such as polyarylsulfones and poly(phenyletheretherketone), new applications can be realized.

7. References

1. H. Rudolph, *Makromol. Chem. Macromol. Symp.* **16** (1988) 57.
2. A. Moro, A. Chiolle, L. Credall, G. Foschini, *Makromol. Chem. Macromol. Symp.* **16** (1988) 137.
3. J. Levy, *Perform. Chem.* **6** (1991) no. 2, 36.
4. L. M. Rossi, R. S. Stevenson, *Compalloy '90* (1990) 357.
5. D. T. Wark, *ANTEC '90* (1990) 1156.
6. S. Izawa, *Int. Polym. Sci. Tech.* **17** (1990) 35.
7. *Specialty Polymeric Blends and Alloys,* Skeist Incorporated, Whippany, N.J. 1991.
8. *Macplas,* November 1991, 65.
9. J. B. Crossfield, *Des. Eng. (N.Y.),* Nov 1990, 45.
10. J. Bussink in P. J. Lemstra, L. A. Kleintjes (eds.): *Integration of Fundamental Polymer Science and Technology,* vol. 3, Elsevier Applied Science, New York 1989.
11. *European Plastic News* **18** (1991) no. 10, 42.
12. O. Olabisi, L. M. Robeson, M. T. Shaw: *Polymer* – Polymer Miscibility, Academic Press, New York 1979.
13. W. J. Macknight, F. E. Karasz, in D. R. Pual, S. Newman (eds.): *Polymer Blends,* **vol. 1,** Academic Press, New York 1978. Chapter 4, 111.
14. W. J. Macknight, F. E. Karasz, J. R. Fried, in D. R. Pual, S. Newman (eds.): *Polymer Blends,* vol. 1, Academic Press, New York 1978.
15. L. A. Utracki: *Polymer Alloys and Blends,* Carl Hanser Verlag, München 1989.
16. M. Bosma, G. ten Brinke, T. S. Ellis, *Macromolecules* **21** (1988) 1465.
17. S. Wu: *Polymer Interface and Adhesion,* Marcel Dekker, New York 1982.
18. N. G. Gaylord, *J. Macromol. Sci. Chem.* **A 26** (1989) 1211.
19. L. Leibler, *Makromol. Chem. Macromol. Symp.* **16** (1988) 1.
20. C. H. Han: *Multiphase Flow in Polymer Processing,* Academic Press, New York 1981.
21. R. M. Ottenbrite, L. A. Utracki, S. Inoue: *Current Topics in Polymer Science,* vol. 2, Hanser Verlag, München 1987.
22. J. G. M. van Gisbergen: "Electron Beam Irradiation of Polymer Blends", Ph. D. Thesis, Eindhoven University 1991.
23. J. J. Elmendorp: "A Study on Polymer Blending Microrheology," Ph. D. Thesis, Delft University 1986.
24. *Aufbereiten von Polymerblends,* VDI Verlag, Düsseldorf 1989.
25. H. P. Grace, *Chem. Eng. Commun.* **14** (1982) 225.
26. S. Y. Hobs, M. E. J. Dekkers, V. H. Watkins, *Polymer* **29** (1988) 1598.
27. M. Xanthos, S. S. Dagli, *Compalloy '90* (1990) 277.
28. GE Plastics, Engineering Materials Design Guide, Bergen op Zoom 1991.
29. C. B. Bucknall: *Toughened Plastics,* Applied Science Publishers, London 1977.
30. H. Kim, H. Keskkula, D. R. Paul, *Polymer* **32** (1991) 2372.
31. H. T. van de Grampel: *RAPRA Review Report Series 49,* Pergamon Press, Oxford 1992.
32. G. Allen, *Polymer J.* **19** (1987) 1. L. A. Utracki in L. A. Utracki, R. A. Weiss (eds.): *Multiphase Polymeric Materials,* ACS Books, Washington, DC. 1989.
33. P. M. Subramanian in W. J. Koros (ed.): "Barrier Polymers and Structures," *ACS Symp. Series* **423** (1990).
34. A. D. Abshire: "Blends Containing Polyolefins", *SRI report* **175 A** (1990).
35. G. Serpe, J. Jarrin, F. Dawans, *Polym. Eng. Sci.* **30** (1990) 553; R. Bell, *ANTEC '90* (1990) 671.
36. J. Willis, B. D. Favis, *Compalloy '89* (1989) 173.
37. A. K. Bhowmik, H. L. Stevens(eds.): *Handbook of Elastomers,* Marcel Dekker, New York 1988.
38. A. J. Kinioch, R. J. Young: *Fracture Behavior of Polymers,* Applied Science Publishers, London 1983.
39. PVC Blends, Alloys and Graft Polymers, Conference Proceedings, Atlanta, Oct. 1990, SPE Vinyl Div.
40. D. Braun, B. Boehringer, J. Herth, *Makromol. Chem. Macromol. Symp.* **29** (1989) 227.
41. A. D. Roberts (ed.): *Natural Rubber Science and Technology,* Oxford University Press, Oxford 1988.
42. F. Mark et al. (ed.): *Encycl. Polym. Sci. Technol.* J. Wiley, New York 1991, vol. 9, p. 769.
43. G. Lindenschmidt, R. Theysohn, *Engineering Plastics* **3** (1990) 1.

44. W. K. Chin, L. J. Hwang, *ANTEC '87* (1987) 1379.
45. D. J. Stein et al., *Angew. Makromol. Chem.* **36** (1974) 89.
46. J. Huang, M. Wang, *Adv. Polym. Technol.* **9** (1989) 293.
47. J. M. Wefer, US 4 895 921, 1985.
48. H. F. Giles, *Modern Plastics Encycl.,* 1987 – 1988.
49. J. Maruta, T. Ougizawa, T. Inoue, *Polymer* **29** (1988) 2056.
50. J. G. Dolden, *Polymer* **17** (1976) 875.
51. P. Maskus, *ANTEC '91* (1991) 1385.
52. T. S. Ellis, *Polym. Commun.* **32** (1991) 489.
53. T. S. Ellis, *Macromolecules* **22** (1989) 742.
54. T. S. Ellis, *Macromolecules* **24** (1991) 3845.
55. V. J. Triacca et al., *Polymer* **32** (1991) 1401.
56. R. E. Lavengood et al., *SPE Tech. Pap.* **33** (1987) 1369.
57. E. Gattiglia et al., *J. Appl. Polym. Sci.* **38** (1989) 1807; **41** (1990) 1411.
58. S. S. Dagli et al., *ANTEC '90* (1990) 1924.
59. M. J. Modic, L. A. Pottick, *Plast. Eng.* **49,** July 1991, 37.
60. M. Wajs, M. Glotin, *Br. Plast. Rubber* 1989 March, 15.
61. S. Y. Hobbs, R. C. Bopp, V. H. Watkins, *Polym. Eng. Sci.* **23** (1983) 380.
62. B. N. Epstein, J. A. Rakeshaw, E. A. Flexman, D. D. Huang, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **32** (1991) no. 2, 42.
63. J. Bussink, W. Minderhout, W. Sederel in print (ed. Bottenbruch), Carl Hanser Verlag, München 1992.
64. M. B. Djordjevic, R. S. Porter, *Polym. Eng. Sci.* **23** (1983) 650.
65. General Electric, US 3 383 435, 1968 (E. P. Cizek).
66. V. T. Bui, Polymer Blends, Symposium Proceedings, Cambridge 1990, C 15/1-4.6125.
67. A. F. Yee, *Polym. Eng. Sci.* **17** (1977) 213.
68. T. Szabados, E. Galambos, *Int. Polym. Sci. Techn.* **13** (1986) 42.
69. S. Y. Hobbs, M. Dekkers, V. H. Watkins, *J. Mater. Sci.* **24** (1989) 2025.
70. H. J. Sue, A. F. Yee, *J. Mater. Sci.* **26** (1991) 3449.
71. J. R. Campbell, S. Y. Hobbs, T. J. Shea, V. H. Watkins, *Polym. Eng. Sci.* **30** (1990) 1056.
72. Chemische Werke Hüls, US 4 656 220, 1987 (H. Jadamus et al.).
73. C. Bailly, *Compalloy '90* (1990) 375.
74. D. Freitag, G. Fengler, L. Morbitzer, *Angew. Chem. Int. Ed. Engl.* **30** (1991) 1598.
75. G. J. Pratt, M. J. A. Smith, *Polymer* **30** (1989) 1113.
76. W. N. Kim, C. M. Burns, *J. Appl. Polym. Sci.* **41** (1990) 1575.
77. D. C. Wahrmund, D. R. Paul, J. W. Barlow, *J. Appl. Polym. Sci.* **22** (1978) 2155.
78. M. E. J. Dekkers, S. Y. Hobbs, V. H. Watkins, *J. Mater. Sci.* **23** (1988) 1225.
79. S. Y. Hobbs, M. E. J. Dekkers, V. H. Watkins, *J. Mater Sci.* **23** (1988) 1219.
80. A. W. Birley, X. Y. Chen, *Br. Polym. J.* **16** (1984) 77.
81. D. Delimoy et al., *Polym. Eng. Sci.* **28** (1988) 104.
82. W. N. Kim, C. M. Burns, *J. Appl. Polym. Sci.* **41** (1990) 1409.
83. R. W. Avakian, R. B. Allen, *Polym. Eng. Sci.* **25** (1985) 462.
84. General Electric, US 4 122 131, 1978.
85. K. W. McLaughlin, *Polym. Eng. Sci.* **22** (1989) 1550.
86. K. W. McLaughlin, *SAE Proceedings,* 1988.
87. D. Quintens, G. Groeninckx, M. Quest, L. Aerts, *Polym. Eng. Sci.* **30** (1990) 1474; **22** (1990) 1484.
88. H. Takahashi et al., *J. Appl. Polym. Sci.* **36** (1988) 1821.
89. W. Chiang, D. Hwung, *ANTEC '86,* 492.
90. H. Kress, D. Folajtan, N. Lazear, *ANTEC '88,* 1834.
91. M. E. J. Dekkers, S. Y. Hobbs, V. H. Watkins, *Polymer* **32** (1991) 2150.
92. D. J. Hourston et al., *Polymer* **32** (1991) 1140.
93. M. Gaztelumendi et al., *Makromol. Chem. Macromol. Symp.* **20** (1988) 269.
94. General Electric, US 3 730 946, 1973 (G. Wirth, D. R. Heath).
95. H. Farong, W. Xueqiu, L. Shijin, *Polym. Degrad Stab.* **18** (1987) 247.
96. G. Crevecoeur, G. Groeninckx, *Macromolecules* **24** (1991) 1190.
97. J. E. Harris, L. M. Robeson, *J. Appl. Polym. Sci.* **35** (1988) 1877.
98. Delrin General Guide to Products and Properties, E. I. du Pont de Nemours & Co., Wilmington, Del. 1991. Celcon Acetal Copolymer Properties, Celanese Corp., New York 1991.
99. G. Bier, *Polymer* **15** (1974) 527.
100. D. Freitag, K. Reinking, *Kunststoffe* **71** (1981) no. 1, 46.
101. M. Kimura, G. Salee, R. S. Porter, *J. Appl. Polym. Sci.* **29** (1984) 1629.
102. A. Golovoy, M.-F. Cheung, *Polym. Eng. Sci.* **29** (1989) no. 2, 85.

103. Bayer; Unikka; GE Plastics; Amoco; product brochures.
104. L. M. Robeson, *J. Appl. Polym. Sci.* **30** (1985) 4081.
105. A. Golovoy, M.-F. Cheung, K. R. Carduner, M. J. Rokosz, *Polym. Eng. Sci.* **29** (1989) 1226.
106. A. Ausin et al., *Polym. Eng. Sci.* **27** (1987) 529.
107. C. E. Desper, M. Kimura, R. S. Porter, *J. Polym. Sci. Polym. Phys. Ed.* **22** (1984) 1193.
108. S. Akhtar, J. L. White, *Polym. Eng. Sci.* **31** (1991) 84.
109. V. L. Shingankuli, P. Jog, V. M. Nadkarni, *J. Appl. Polym. Sci.* **36** (1988) 335.
110. D. V. Rosato, D. P. Di Mattia, D. V. Rosato: *Designing with Plastics and Composites: A Handbook,* Van Nostrand Reinhold, New York 1991.
111. G. Menges, *Makromol. Chem. Macromol. Chem.* **23** (1989) 13.
112. *Modern Plastics,* July 1991.
113. F. P. La Mantia, D. Curto, *Polym. Degrad. Stab.* **36** (1992) 131.
114. J. R. M. Duhaime, W. E. Baker, *Plast. Rubber Process Appl.* **15** (1991) 87.
115. E. Wissler, *ANTEC '90* (1990) 1434.
116. I. M. Chen, C. M. Shiah, *ANTEC '89* (1989) 1802.
117. A. G. Staniulis, *ANTEC '90* (1990) 794.
118. GE Plastics, Design for Recycling, Bergen op Zoom 1992.
119. N. H. Naitove, R. Monks. *Plast. Technol.,* **38** (1992) 101.
120. L. M. Martynowicz, *J. Cell. Plast.* **26** (1990) 423.
121. A. McHale, P. Y. Liu, SPE Conference Proceedings, Rosemont II, 1987, 147.
122. J. L. Hedrick et al., *Polymer* **32** (1991) 2020.
123. C. B. Bucknall, A. H. Gilbert, *Polymer* **30** (1989) 213.
124. A. H. Gilbert, C. B. Bucknall, *Makromol. Chem. Macromol. Symp.* **45** (1991) 289.
125. D. J. Hourston, J. M. Lane, N. A. Macbeath, *Polymer Int.* **26** (1991) 17.
126. S. J. Kubisen, P. C. Long, Proceedings Electronic Materials and Processing Conference, A.S.M. International, Chicago III., Sept. 1988.
127. S. J. Kubisen et al.: *A New Performance Material,* GE Laminated Products, Coshocton, Ohio 1991.
128. G. Sinai-Zingde et al., *Makromol. Chem. Macromol. Symp.* **42/43** (1991) 329.
129. K. G. Blizard, C. Federici, O. Federico, L. L. Chapoy, *Polym. Eng. Sci.* **30** (1990) 1442.
130. T. M. Malik et al., *Polym. Eng. Sci.* **29** (1989) 60.
131. H. J. O'Donell et al., *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **33** (1992) no. 1, 376.
132. M. Paci et al., *J. Polym. Sci. Polym. Phys.* **25** (1987) 1595.
133. G. Crevecoeur, G. Groeninckx, *Polym. Eng. Sci.* **30** (1990) 532.
134. M. Takayanagi, K. Goto, *J. Appl. Polym. Sci.* **29** (1984) 2547.
135. V. G. Kulichikhin et al., *J. Appl. Polym. Sci.* **42** (1991) 363.
136. A. M. Shukhadia, D. Done, D. G. Baird, *Polym. Eng. Sci.* **30** (1990) 519.
137. M. Kyotani et al., *Polym. Prepr. Jpn. (Engl. Ed.)* **39** (1990) E 1423.
138. F. P. La Mantia, A. Valenza, M. Paci, P. L. Magagnini, *Polym. Eng. Sci.* **30** (1990) 7.
139. Y. Tokiwa, A. Iwamoto, M. Koyama, *Polym. Mater. Sci. Eng.* **63** (1990) 742.
140. M. Jaffe et al., *Makromol. Chem. Macromol. Chem.* **53** (1992) 163.
141. J. Machado, F. E. Karasz, R. W. Lenz, *Polymer* **29** (1988) 1412.
142. J. Alper, G. L. Neldon: *Polymeric Materials: Chemistry for the Future,* ACS, Washington, DC, 1989.
143. *Plast. World,* September 1991, 55.

Polymerization Processes

Archie E. Hamielec, Institute for Polymer Production Technology, Department of Chemical Engineering, McMaster University, Hamilton, Ontario, L8S 4L7, Canada

Hidetaka Tobita, Department of Materials Science and Engineering, Fukui University, Fukui, 910, Japan

1. Introduction–Trends in Polymer Reaction Engineering . . . 7
2. Polymerization Mechanisms and Kinetics 8
2.1. Step-Growth Polymerization . . 9
2.1.1. Linear Polymerization 9
2.1.2. Interfacial Polymerization 12
2.1.3. Nonlinear Polymerization 12
2.2. Chain-Growth Polymerization . 14
2.2.1. Free-Radical Polymerization . . . 15
2.2.1.1. Initiation 16
2.2.1.2. Propagation 18
2.2.1.3. Termination 19
2.2.1.4. Chain Transfer to Small Molecules 21
2.2.1.5. Kinetics of Linear Polymerization 22
2.2.1.6. Effect of Temperature 25
2.2.1.7. Branching Reactions 26
2.2.2. Ionic Polymerization 28
2.2.2.1. Cationic Polymerization 29
2.2.2.2. Anionic Polymerization 30
2.2.2.3. Ziegler – Natta Polymerization . . 32
2.3. Copolymerization 34
2.3.1. Copolymer Composition 35
2.3.2. Kinetics of Copolymerization . . 37
2.3.3. Copolymerization of Vinyl and Divinyl Monomers 38
3. Polymerization Processes and Reactor Modeling 40
3.1. Introduction 40
3.2. Processes and Reactor Modeling for Step-Growth Polymerization 41
3.2.1. Types of Reactors and Reactor Modeling 41
3.2.2. Specific Processes 44
3.3. Processes and Reactor Modeling for Chain-Growth Polymerization 47
3.3.1. Material Balance Equations for Batch, Semi-Batch, and Continuous Reactors 47
3.3.1.1. Rates of Reaction and Copolymer Composition 48
3.3.1.2. Molecular Masses, Long-Chain Branching, and Cross-Linking . . 50
3.3.2. Examples of Free-Radical Polymerization 50
3.3.2.1. Homopolymerization – Linear Chains 50
3.3.2.2. Copolymerization – Linear Chains 54
3.3.2.3. Copolymerization – Long-Chain Branching 55
3.3.3. Polymerization Processes 55
3.3.3.1. Solution Polymerization 55
3.3.3.1.1. Polymer Soluble in Monomer . . 55
3.3.3.1.2. Addition of a Solvent in which both Monomer and Polymer are Miscible 55
3.3.3.1.3. Polymer – Polymer Demixing during Polymerization 56
3.3.3.2. Precipitation Polymerization . . . 57
3.3.3.2.1. Polymer Insoluble in its Monomer 57
3.3.3.2.2. Monomer Functioning as Solvent for the Polymer 60
3.3.3.3. Suspension Polymerization 63
3.3.3.3.1. Qualitative Description 64
3.3.3.3.2. Dispersants 66
3.3.3.3.3. Mechanism of Particle Formation 67
3.3.3.3.4. Industrial Applications 70
3.3.3.4. Emulsion Polymerization 73
3.3.3.4.1. Theories of Emulsion Polymerization 74
3.3.3.4.2. Physicochemical Parameters of Dispersions 85
3.3.3.4.3. Inverse Emulsion Polymerization 88
3.3.3.4.4. Semi-Batch Emulsion Polymerization 89
3.3.3.4.5. Continuous Emulsion Polymerization 90
3.3.4. Miscellaneous Processes 93
3.3.5. Ionic Polymerization Modeling . 96
3.3.5.1. Introduction 96
3.3.5.2. Heterogeneous Coordination Polymerization 96

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_305

3.3.6. Process Variables, Reactor Dynamics/ Stability, On-Line Monitoring and Control 97
3.3.6.1. Influence of Reactor Type and Configuration on Molecular Mass and Copolymer Composition Distributions, and on Long-Chain Branching and Cross-Linking . . 97
3.3.6.1.1. Monomer Coupling with Bimolecular Termination Plug Flow and Batch Reactors (CPFR/BR) 99
3.3.6.1.2. Monomer Coupling Without Termination Plug Flow and Batch Reactors (CPFR/BR) 103
3.3.6.1.3. Polymer Coupling 104
3.3.6.1.4. Copolymerization 107
3.3.6.1.5. Long-Chain Branching and Cross-Linking 109
3.3.6.2. Reactor Dynamics and Stability . 111
3.3.6.3. On-Line Monitoring and Control 112
4. References 114

Sections 3.3.3.1 – 3.3.3.5 and 3.3.6.1 were based on the article Polymerisationstechnik in Ullmann's, 4th ed. written by HEINZ GERRENS.

List of symbols

A chemical species; vacant adsorption site
[A] concentration of species A
$[A]_0$ initial concentration of species A
A_1, A_2, A_3 adjustable parameters
ABS acrylonitrile – butadiene – styrene rubber-modified copolymer
ACA aminocaproic acid
$A(h)$ energy required to separate to a distance $h=\infty$, two drops of diameter $d=1$ initially separated by a distance h_0
A_m surface area of micelles
A_p surface area of polymer particles
B chemical species
BHET bis-hydroxyethyl terephthalate
BR batch reactor
C_{pi} dimensionless moments of polymer distribution for chain transfer to polymer $[= K_{fp} Q_i/(K_p[M])]$
C_S surfactant concentration
CCD chemical composition distribution
CMC critical micelle concentration
CPFR continuous plug flow reactor
CSTR continuous stirred-tank reactor with an ideal residence-time distribution
CTA chain-transfer agent
d particle diameter
$\bar{d}$ average particle diameter
d_{32} Sauter mean diameter of a spherical-particle suspension
d_{50} diameter at which 50 wt % of particles pass through a sieve
d_{min} minimum particle diameter
d_{max} maximum particle diameter
D stirrer diameter
D_{op} mean diffusion coefficient for oligomeric radicals and latex particles
DMT dimethyl terephthalate
E_d activation energy for initiator decomposition
E/E_0 mass fraction of material passing out of reactor with a residence time t to $t + dt$
E_f activation energy for chain-transfer reaction
E_L activation energy for average chain lengths
E_N activation energy for polymer particle nucleation
E_p activation energy for propagation
E_R activation energy for polymerization
$E(t)$ residence-time distribution for a flow reactor at steady state
E_t activation energy for bimolecular termination
Eu modified power number
EG ethylene glycol
EGDMA ethylene glycol dimethacrylate
EPS expandable polystyrene
ESR electron spin resonance spectroscopy
f initiator efficiency; functionality of monomer
f_j mole fraction of monomer of type j
$F_{i,in}$ molar flow rate of monomer of type i into the reactor
F_{in} total molar flow rate (of all monomer types) into the reactor
$F_{Ii,in}$ molar flow rate of initiator of type i into the reactor
F_j mole fraction of monomer of type j, chemically bound in polymer produced instantaneously

$\bar{F}_j$ mole fraction of monomer of type j chemically bound in accumulated polymer
$\bar{F}_1$ mole fraction of monomer 1 (containing an abstractable atom) in accumulated polymer
$\bar{F}_2$ mole fraction of monomer 2 (containing a reactive carbon – carbon bond
$F_{\mathrm{p}i,\mathrm{in}}$ molar flow rate of monomer of type i chemically bound in polymer into the reactor
Fr Froude number
$F_{\mathrm{T,in}}$ molar flow rate of chain-transfer agent T into the reactor
G^+ counterion
GPC gel permeation chromatography
HCSTR homogeneous CSTR
HDPE high-density polyethylene
H – H Hui – Hamielec styrene polymerization model
HIPS high-impact polystyrene
I initiator or catalyst
[I] concentration of initiator or catalyst
K chemical rate constant; equilibrium constant
K_{a} absorption constant for oligomeric radicals entering polymer particles
K_{A} adsorption rate constant
K_{d} initiator decomposition constant
K_{dp} depropagation constant
K_{D} desorption rate constant
$K_{\mathrm{f}\,{}^{j}_{i}}$ rate constant for polymeric radical of type i abstracting an atom from monomer of type j chemically bound in polymer
K_{fm} transfer to monomer rate constant
K_{fp} rate constant for chain transfer to polymer
K_{fT} rate constant for chain transfer to CTA
K_{fT^i} rate constant for chain transfer from polymeric radical of type i to CTA
K_{fX} transfer to small molecule X rate constant
K_{i} rate constant for monomer adding to a primary radical
K_{p} propagation rate constant
K'_{p} propagation rate constant for transfer radical
K^-_{p} propagation rate constant for free ion
$K^{\pm}_{\mathrm{p}}$ propagation rate constant for ion pair
$K_{\mathrm{p}*}$ rate constant for polymeric radicals adding to pendant double bonds on polymer chains
$K_{\mathrm{p}\,{}^{j}_{i}}, K_{ij}$ propagation rate constant for monomer of type j adding to polymeric active center of type i
$K_{\mathrm{p}_{ij}}$ rate constant for polymeric radical of type i adding to a double bond on a monomer unit of type j chemically bound in the polymer
$K_{\mathrm{p}_{ijk}}$ propagation rate constant for monomer of type k adding to a polymeric active center of type ij
K_{t} total bimolecular termination constant $(K_{\mathrm{tc}} + K_{\mathrm{td}})$
K_{t0} total bimolecular termination constant at zero conversion of monomer
K_{tc} rate constant for bimolecular termination by combination
$K_{\mathrm{tc}_{ij}}$ termination by combination rate constant for polymeric radicals of types i and j (chemical control)
$\bar{K}_{\mathrm{tcN}}$ number-average bimolecular termination constant by combination
K_{td} rate constant for bimolecular termination by disproportionation
$K_{\mathrm{td}_{ij}}$ termination by disproportionation rate constant for polymeric radicals of types i and j (chemical control)
$\bar{K}_{\mathrm{tdN}}$ number-average bimolecular termination constant by disproportionation
$\bar{K}_{\mathrm{tN}}$ total number-average bimolecular termination constant
K_{tp} termination rate constant in polymer particles
$K_{\mathrm{t}}(r, s)$ total bivariate distribution for diffusion-controlled bimolecular termination of polymeric species of chain lengths r and s
K_{tw} termination rate constant in aqueous phase
$\bar{K}_{\mathrm{tW}}, \bar{K}_{\mathrm{t}Z}$ total weight- and z-average bimolecular termination constants
L characteristic length of energy-containing large eddies
L length of path traversed by a growing radical from its point of origin to the point where it precipitates
L reactor length
LALLS low-angle laser light scattering
LCB long-chain branching

LDPE low-density polyethylene
LLDPE linear low-density polyethylene
m_i number of moles of monomer i in terpolymer (Eq. 3.101)
M_c average molecular mass between cross-links
M_i monomer of type i
M_m aggregation number for emulsifier molecules in micelles
M_{mi} molecular mass of monomer of type i
$\bar{M}_N, \bar{M}_W, \bar{M}_Z, \bar{M}_{Z+1}$ number-, weight-, Z and $Z+1$-average molar mass (molecular mass, respectively)
[M] total monomer concentration
$[M]_0$ initial monomer concentration; monomer concentration in feed
$[M]_c$ equilibrium concentration of monomer at the ceiling temperature
$[M_i]$ concentration of monomer of type i
$[M]_p$ concentration of monomer in the polymer particles
M – H Marten – Hamielec polymerization model
MMA methyl methacrylate
MWD molecular mass distribution (molar mass distribution)
n number of monomer types
n order of reaction
$\bar{n}$ average number of radicals per particle
N_0, N number of functional groups at time zero and t
N total number of moles in the reactor; stirrer speed
N_A number of moles of A-functional groups; Avogadro number
N_{A_0} initial number of moles of A-functional groups
N_B number of moles of B-functional groups
N_{B_0} initial number of moles of B-functional groups
N_i moles of monomer of type i in the reactor
N_I number of moles of initiator in the reactor; number of growing chains
N_{I0} initial number of moles of initiator in the reactor
N_{Ii} moles of initiator of type i in the reactor
$N(r)$ number chain length distribution (number-fraction of polymer molecules of chain length r)
N_M number of monomer units; number of micelles; number of monomer molecules consumed
N_n number of polymer particles containing n radicals
N_p number of polymer particles per unit volume
N_T moles of CTA in reactor
NBR nitrile – butadiene rubber
NIRS near infrared spectroscopy
p conversion of functional groups
p_c critical threshold
P growing polymer particle; polymer
P_c conversion of functional groups at gelation point
P_{cr} critical chain length for precipitation
P_i moles of monomer of type i chemically bound in polymer in the reactor
P_{ij} polymer containing i units of monomer of type 1 and j units of monomer of type 2
$[P_m]$ concentration of polymer with chain length m
$P_{m,n}$ dead polymer chain containing m units of monomer 1 and n units of monomer 2
P_N number-average chain length of polymer produced instantaneously
$\bar{P}_N$ number-average chain length of accumulated polymer
$\bar{P}_N^{sol}$ number-average chain length of sol molecules
P_r polymer molecule of chain length r
P_W weight-average chain length of polymer produced instantaneously
$\bar{P}_W$ weight-average chain length of accumulated polymer
$\bar{P}_W^{sol}$ weight-average chain length of sol molecules
PDI polydispersity index of polymer produced instantaneously
$\overline{PDI}$ polydispersity index of accumulated polymer
PE polyethylene
PEK polyetherketone
PES polyethersulfone
PETP poly(ethylene terephthalate)
PFR plug-flow reactor
PMMA poly(methyl methacrylate)
PP polypropylene
PPS poly(phenylene sulfide)
PS polystyrene

PSD particle size distribution
PVAL poly(vinyl alcohol), partially hydrolyzed
PVC poly(vinyl chloride)
$P*$ polymeric active center
$[P*]$ concentration of polymeric active centers (ionic or radical type)
P_i^* polymeric active center with active center located on monomer of type i chemically bound in the polymer chain
P_{ij}^* polymeric active center with active center located on monomer of type j which is adjacent to monomer of type i chemically bound in the polymer chain
$P_{m,n,i}^*$ polymer chain containing m units of monomer 1, n units of monomer 2, with active center on monomer i
Q_i i-th moment of the dead polymer distribution
r polymer chain length
r polymer particle radius
r_{M} micelle radius
r_{p} polymer particle radius
r_1, r_2 reactivity ratios
R gas constant
$R^\bullet$ polymeric radical
$[R^\bullet]$ concentration of polymeric radicals
Re Reynolds number
R_{f} rate of chain transfer
$R_{\mathrm{FP}}(r)$ production rate of polymer molecules with chain length r
R_{i} rate of radical entry into polymer particles
R_{in}^* initiator/catalyst fragment with an active center
$\hat{R}_{\mathrm{in}}^\bullet$ initiator radical with a peroxide end group
$R_{\mathrm{in}}^\bullet$ initiator or primary radical
$[R_{\mathrm{in}}^\bullet]$ concentration of primary radicals
$R_{\mathrm{i,w}}$ radical generation rate in the aqueous phase via initiator decomposition
R_{I} rate of initiation (rate of generation of polymeric radicals with chain length unity via initiation)
RIM reaction injection molding
R_{p} rate of polymerization (monomer consumption rate via propagation reactions)
R_{p_i} consumption rate of monomer of type i via propagation reactions
$R_{\mathrm{p}_{ij}}$ consumption rate of monomer of type j by propagation with polymeric radicals of type i
$R_{\mathrm{p,o}}$ initial polymerization rate
$R_r^\bullet$ polymeric radical of chain length r
$[R_r^\bullet]$ concentration of polymeric radicals of chain length r
R_{t} rate of bimolecular termination
R_{tc} rate of bimolecular termination by combination
R_{td} rate of bimolecular termination by disproportionation
RTD residence time distribution
$[R^\bullet]_w$ concentration of radicals in the aqueous phase
ΔS^0 entropy change of polymerization at the standard state
S surface area of polymer particle
S surfactant
SAN styrene – acrylonitrile copolymer
SBR styrene – butadiene rubber
SCSTR segregated CSTR
SSH stationary-state hypothesis
t time
$t_{1/2}$ half-life of initiator
t_1 polymer particle nucleation time
t_{s} time when polymer particle nucleation ceases
T temperature
T chain-transfer agent
$T^\bullet$ CTA transfer radical
T_{c} ceiling temperature
T_{g} glass transition temperature
TPA terephthalic acid
TREF temperature rise elution fractionation
U MWD nonuniformity index
v volume of polymer particle
v volumetric flow rate into and out of reactor
v_{c} capture rate of radicals by polymer particles
v_{f} flocculation rate of precipitated (primary) polymer particles
V volume of reacting mixture in the reactor
V_0 intial volume of reacting mixture in the reactor
V_{in} total volumetric flow rate into the reactor
V_{out} total volumetric flow rate out of the reactor

V_m specific volume of monomer
V_p specific volume of polymer
V_s volume of solvent in the reactor
$V_{s,in}$ volumetric flow rate of solvent into the reactor
VCM vinyl chloride monomer
We Weber number
$\bar{W}_g$ weight fraction of gel
$W(r)$ weight chain length distribution (weight fraction of polymer of chain length r)
$W(r), W(r,t)$ "instantaneous" weight chain length distribution
$\bar{W}(r), \bar{W}(r,t)$ weight chain length distribution of accumulated polymer
W_1, W_2 weight fractions of homopolymers 1 and 2
x monomer conversion
X small molecule with a labile atom
$X^\bullet$ transfer radical
z exponent indicating dependence of N_p on emulsifier and initiator concentrations
α stoichiometric imbalance
$\Delta\alpha_1, \Delta\alpha_2$ differences between thermal expansion coefficients above and below T_g for homopolymers 1 and 2
β kinetic parameter (dimensionless)
γ kinetic parameter (dimensionless)
$*\gamma_{prec}$ volume fraction of precipitant
δ kinetic parameter (dimensionless)
ε characterizes the radical capture efficiency of latex particles relative to micelles; energy-dissipation rate
$\bar{\varepsilon}$ mean rate of energy dissipation per unit mass
η moles of monomer consumed per active site ($\equiv \bar{P}_N$)
η_c viscosity of continuous phase
η_d viscosity of disperse phase
ϱ_c density of continuous phase
ϱ_d density of disperse phase
ϱ_{el} elastic cross-link density
ϱ_{m_i} density of monomer i
ϱ_p density of polymer
σ standard deviation
σ^2 statistical variance
σ interfacial tension
σ_{SG} interfacial tension between solid and gas
σ_{SM} interfacial tension between solid and monomer
σ_{SW} interfacial tension between solid and water
σ_{WG} interfacial tension between water and gas
σ_{WM} interfacial tension between water and monomer
τ kinetic parameter (dimensionless)
τ mean residence time
φ phase volume ratio
φ, φ' probability of propagation
φ_m volume fraction of monomer
φ_p volume fraction of polymer
$\dot{\varphi}_i$ number fraction of polymeric radicals of type i (terminal model)
φ_i^* number fraction of active or live polymer molecules of type i (radical or ionic, terminal model)
$\dot{\varphi}_{ij}$ number fraction of polymer radicals of type ij (penultimate model)
Φ kinetic parameter (dimensionless)
χ Flory – Huggins polymer – solvent interaction parameter
$\psi(r)$ number fraction of polymeric radicals of chain length r

1. Introduction–Trends in Polymer Reaction Engineering

The worldwide production of synthetic polymers, estimated at ca. 100×10^6 t/a in 1990 [1] and at ca. 170×10^6 t/a in 2000 [957], continues to grow in spite of criticism from environmentalists. Polymer waste has become an urgent topic for industry, providing new and challenging areas of research and development on recycling, reuse, and degradation. The technical principles of polymer reaction engineering will no doubt play a significant role in the solution of some of these problems. With the increase in production volumes of commodity polymers (LLDPE, HDPE, PP, PVC, and PS copolymers), large-reactor technology (suspension PVC) and continuous processes (production of LLDPE in continuous fluidized-bed reactors, e.g., UNIPOL process) are being developed [1], [2]. In the early days of the polymer industry, polymers were specialty materials, produced in batch reactors by using faithfully followed recipes scaled up from the chemist's beaker. The process engineer, although versed in the principles of chem-

ical reaction engineering, had little background in polymer chemistry, polymerization kinetics, and polymer characterization techniques. This has changed dramatically in the last two decades, as evidenced by the rapid growth of the field of polymer reaction engineering within the chemical reaction engineering discipline. Process parameters, such as residence-time distribution, micromixing, and segregated flow, whose influence on productivity and selectivity of small molecule reactions has been studied for many years, appear to be far more important for polymerization reactors in that they influence polymer properties dramatically [1–7].

The development of engineering and specialty polymers with a better balance of properties or with a particular unique property has been growing rapidly. In this regard, it has been found to be often more economic to produce a new polymer from existing commodity polymers rather than to start with a new monomer and produce polymer in the usual manner. Techniques such as polymer alloying and blending are particularly attractive. These and other techniques use chemical modifications of existing polymers by chain scission, long-chain branching, cross-linking and grafting. These chemical modifications are usually carried out with polymer melts in an extruder reactor [1]. This process is often called reactive polymer processing. This is a new and commercially promising area where the principles of polymer reaction engineering could be profitably exploited.

Since 1980, modeling of polymerization reactors has become more comprehensive. Interest has focussed on the prediction of polymer properties (chemical composition and molecular mass distribution, long-chain branching, cross-link density, polymer particle size distribution, and particle morphology). To develop a predictive model, account must be taken of the chemistry and physics of all of the relevant microscopic processes which occur in the polymerization process. Detailed physical property and thermodynamic data on the partitioning of species among phases is required to quantitatively calculate the concentrations of reactants at the loci of polymerization. Valid kinetic rate constants (frequency factors and activation energies) are also required. In this regard, one should note that the values for individual elementary rate constants are often not required. In the models, groups of rate constants often appear when calculating rates and polymer properties. A knowledge of the Arrhenius equation (overall frequency factor and activation energy) is usually sufficient. Another factor which should be noted is that process models, no matter how detailed, cannot track polymerization rates and polymer properties in real time without feedback from online sensors. The variability in trace impurity levels cannot be accounted for without periodically adjusting kinetic parameters in a process model. The great effort made by chemical kineticists to measure individual elementary rate constants are not in vain. Elementary rate constants can be related to the structure of the reactants, but more importantly for process modelers, elementary rate constants can be used to discriminate kinetic models (for example, the terminal and penultimate models in copolymerization). At this point it is appropriate to emphasize the need for on-line sensors to monitor polymer properties so that process models can be used more effectively in state estimation and control.

2. Polymerization Mechanisms and Kinetics

Polymerization reactions can be classified as either step-growth or chain-growth reactions. It has been proposed that these mechanisms should be termed random and sequential polymerizations [18], [19] since these terms have more significance statistically and are devoid of inference concerning the chemistry of the reactions involved in the polymerizations. In this article, however, the conventional terms step-growth and chain-growth polymerization are used. It is important to note that this is a classification of reaction mechanisms, not of the structure of the repeating unit, since many polymers can be synthesized either by step-growth or chain-growth polymerization. Generally, however, polymer physical properties can differ significantly depending on the polymerization mechanism, and this is often due to the difference in molecular masses, i.e., polymers synthesized by chain-growth polymerization often have higher molecular masses.

These two types of growth reaction differ basically in terms of the time-scale of vari-

ous reaction events, namely, the size of polymer molecules increases at a relatively slow rate over a much longer period of time in step-growth polymerization. With step-growth polymerization, the reactions that link monomers, oligomers, and polymers involve the same reaction mechanism, and any two molecular species (monomer, oligomer, or polymer) can be coupled. The growth of a polymer chain proceeds slowly from monomer to dimer, trimer, tetramer, and so on, until full-sized polymer molecules are formed at high monomer conversions. Polymer chains continue to grow from both ends throughout the polymerization and, therefore, both chain lifetimes and polymerization times are usually of the order of hours.

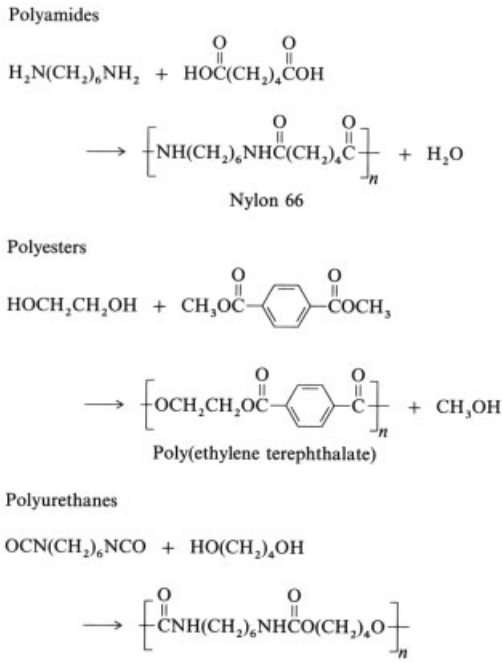

Figure 1. Linear polymers produced via step-growth polymerization

On the other hand, in chain-growth polymerization, polymer molecules generally grow to full size in a time-scale which is much smaller than the time required for high conversion of monomer to polymer. The lifetime of a growing polymer molecule may be less than a few seconds for a free-radical polymerization, which is a typical example of chain-growth polymerization, while a typical polymerization time to obtain high monomer conversion may be several hours. Chain-growth polymerizations require an active center, which may be a free radical, cation, or anion. Once an active center is created, a polymer chain grows extremely rapidly, and when the growing chain is deactivated by a termination reaction, the polymer chain is dead and no longer takes part as a reactant. With free-radical polymerization, however, the so-called dead polymer chain is not always truly dead because under certain circumstances it may itself react with radicals. The active center may initiate the growth of many polymer chains.

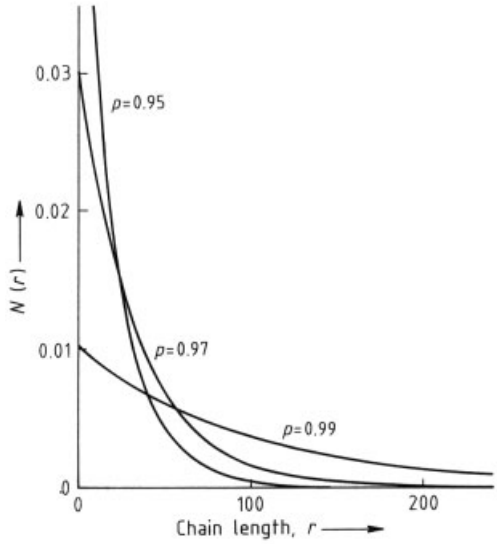

Figure 2. Number chain length distribution in linear step-growth polymerization

2.1. Step-Growth Polymerization

2.1.1. Linear Polymerization

Figure 1 shows some representative linear polymer chains produced commercially by step-growth polymerization. In step-growth polymerization, there is generally only one type of chemical reaction which links molecules of all sizes. Some of the typical chemical reactions are esterification, amidation, the formation of urethanes, and aromatic substitution. The growth reaction in step-growth polymerization can be represented by the general reaction

$$m\,\text{mer} + n\,\text{mer} \longrightarrow (m+n)\,\text{mer} \tag{2.1}$$

The kinetic study of such reactions would be extremely difficult if the rate constant for the coupling reaction depended on the size of both species. Fortunately, various kinetic studies have shown that the rate constant is effectively independent of chain length except perhaps for oligomers. This is often referred to as the concept of equal reactivity of functional groups.

Consider the example of step-growth polymerization shown below.

$$n\,\mathrm{A-A} + n\,\mathrm{B-B} \rightarrow \left[\mathrm{A-A-B-B}\right]_n \tag{2.2}$$

In the case of polyesterification of a diol and a diacid, A may be a hydroxyl group and B may be a carboxyl group, although the low molecular mass condensation byproduct is not shown. As will be shown later, an almost exact equivalence in the number of functional groups is necessary to obtain polymers with high molecular mass, although a nonstoichiometric condition may be used to control molecular mass. In the case of exact stoichiometric ratio of the two types of functional groups, i.e., [A] = [B], the polymerization rate or the rate of disappearance of functional groups is given by

$$-\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{A}])}{\mathrm{d}t} = K\,[\mathrm{A}]^2 \tag{2.3}$$

except for self-catalyzed polymerization, in which case the rate is third order in monomer (the self-catalyzed polymerization may not be a useful reaction from the practical point of view of productivity). Neglecting the volume change during polymerization, integration of Equation (2.3) gives

$$1/\,(1-p) = 1 + K\,[\mathrm{A}]_0\,t \tag{2.4}$$

where $[\mathrm{A}]_0$ is the initial (at $t = 0$) concentration of A groups, and p is the conversion of functional groups, which is defined as

$$p = \left(N_{\mathrm{A}_0} - N_{\mathrm{A}}\right)/N_{\mathrm{A}_0} \tag{2.5}$$

where N_{A_0} and N_{A} are the total number of moles of A groups at $t = 0$ and at any later time t, respectively. Equation (2.4) has been verified by several kinetic studies. As shown here, the rate expression for a step-growth polymerization is the same for monomer molecules, oligomers, and polymers.

The relationship between the average number of structural units (A–A and B–B in the above example), namely, the number-average chain length $\bar{P}_{\mathrm{N}}$ and the conversion of functional groups p for linear step-growth polymerization was first derived by Carothers [20]. $\bar{P}_{\mathrm{N}}$ is simply given as the total number of monomer molecules initially present divided by the total number of molecules present at time t.

$$\bar{P}_{\mathrm{N}} = N_{\mathrm{A}_0}/N_{\mathrm{A}} = 1/\,(1-p) \tag{2.6}$$

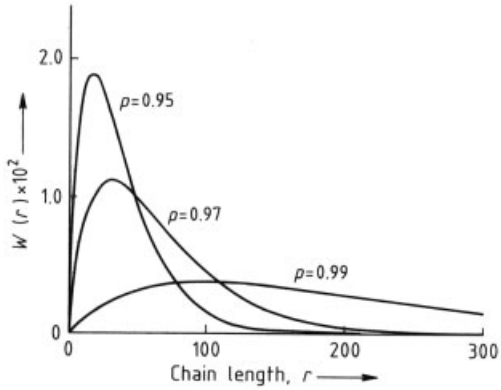

Figure 3. Weight chain distribution in linear step-growth polymerization

Equation (2.6) shows that very high conversions are necessary to obtain large chain lengths. For example, $\bar{P}_{\mathrm{N}} = 100$ requires a conversion of 99 %. Equation (2.6) assumes a stoichiometric ratio of unity. If a slight excess of one bifunctional monomer is used, all chain ends will eventually consist of the group present in excess. When $N_{\mathrm{A}_0} < N_{\mathrm{B}_0}$, the stoichiometric imbalance α is given by $\alpha = N_{\mathrm{A}_0}/N_{\mathrm{B}_0}$. The total number of monomer molecules initially present is given by $(N_{\mathrm{A}_0} + N_{\mathrm{B}_0})/2 = N_{\mathrm{A}_0}\,(1 + 1/\alpha)/2$. Now consider the situation at conversion p (p is usually defined with respect to the deficient group, so that p is defined for A groups in this example). Since each chain end is an unreacted functional group, the total number of end groups is $[N_{\mathrm{A}_0}\,(1-p) + N_{\mathrm{B}_0}\,(1-\alpha p)]$. Each molecule possesses two end groups, so that the total number of polymer (plus monomer) molecules is $[N_{\mathrm{A}_0}\,(1-p) + N_{\mathrm{B}_0}\,(1-\alpha p)]/2$. Therefore, Equation (2.6) can be modified as follows

$$\bar{P}_{\mathrm{N}} = \frac{N_{\mathrm{A}_0}\,(1 + 1/\alpha)\,/2}{\left\{N_{\mathrm{A}_0}\,(1-p) + N_{\mathrm{B}_0}\,(1-\alpha p)\right\}/2} = \frac{1+\alpha}{1+\alpha-2\alpha p} \tag{2.7}$$

As conversion p approaches unity, $\bar{P}_N$ approaches $(1+\alpha)/(1-\alpha)$. Thus if $\alpha = 0.99$, the maximum number-average chain length is only 199. This example illustrates the importance of precise control of the stoichiometric ratio to obtain a desired chain length.

In general, in order to produce high molecular mass polymer by step-growth polymerization, the system must satisfy the following requirements:

1) Very accurate control of the stoichiometric ratio of functional groups
2) Absence of side reactions
3) Availability of high-purity monomers
4) Reasonably high polymerization rate
5) Little tendency towards cyclization reactions

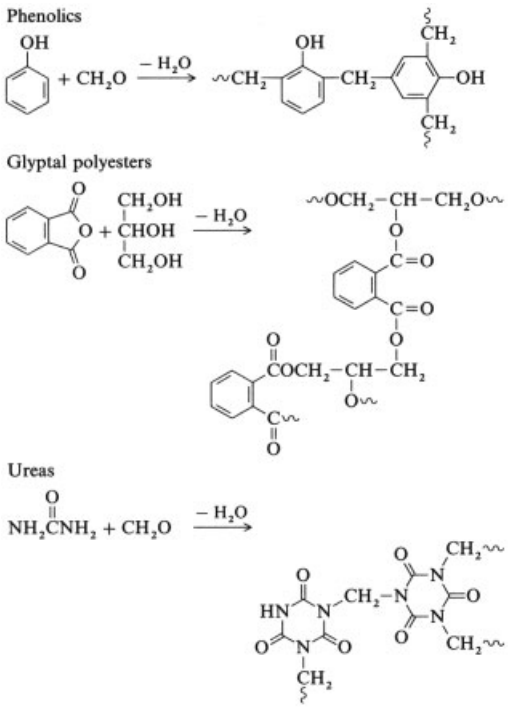

Figure 4. Nonlinear (network) polymers produced via step-growth polymerization

Since high molecular mass polymer is not produced until nearly complete conversion of monomer has occurred, the viscosity is relatively low throughout most of the conversion range. Thermal control and mixing is not overly difficult, which is opposite to the case for chain-growth polymerization. These are some of the reasons why bulk polymerization is quite often used commercially for the production of polyesters and polyamides.

The molecular mass distribution can most easily be derived by using statistical methods for a stoichiometric ratio of unity [21]. The conversion p can be interpreted as the probability that a functional group selected at random has reacted. Consider the probability that a randomly selected molecule consists of r monomer units (this quantity is equal to the number chain length distribution). This polymer molecule possesses $r-1$ reacted functional groups, and one unreacted functional group. Therefore, the number chain length distribution, $N(r)$ is given by

$$N(r) = p^{(r-1)}(1-p) \tag{2.8}$$

The weight chain length distribution $W(r)$ is given by

$$W(r) = r\,N(r) \Big/ \sum_{r=1}^{\infty} r\,N(r) = r\,p^{(r-1)}(1-p)^2 \tag{2.9}$$

The number and weight chain length distributions are shown schematically in Figures 2 and 3, respectively. The weight-average chain length is given by

$$\bar{P}_W = \sum_{r=1}^{\infty} r\,W(r) = (1+p)/(1-p) \tag{2.10}$$

Since the number-average chain length is given by Equation (2.6), the polydispersity index, $\overline{\mathrm{PDI}} = \bar{P}_W/\bar{P}_N$, is given by $(1+p)$, and therefore the $\overline{\mathrm{PDI}}$ approaches two as complete conversion is approached.

Various statistical treatments other than that shown above have been developed to calculate the molecular mass distribution for linear step-growth polymerization [22–26]. Although these statistical methods appear to work well, kinetic approaches based on the use of material balances may have greater generality [27–33]. For an A–B type monomer in a batch reactor, Equations (2.8) and (2.9) can also be derived from the following infinite set of differential equations.

$$\mathrm{d}\,[\mathrm{P}_1]/\mathrm{d}t = -2K\,[\mathrm{P}_1]\,[\mathrm{P}] \tag{2.11}$$

$$\mathrm{d}\,[P_m]/\mathrm{d}t = K \sum_{r=1}^{m-1} [\mathrm{P}_r]\,[\mathrm{P}_{m-r}] - 2K\,[\mathrm{P}_m]\,[\mathrm{P}] \quad (m \geq 2) \tag{2.12}$$

where $[P_m]$ is the concentration of polymer molecules with chain length m, and [P] is the total concentration of polymer and monomer.

For example, it is straightforward to derive the molecular mass distribution for the cases in which a slight amount of monofunctional reagent is used. Kinetic approaches would be easier to apply to reactors other than batch reactors, such as semi-batch and continuous flow reactors, although a statistical derivation for a stirred-tank reactor has been reported [34].

2.1.2. Interfacial Polymerization

Interfacial polymerization may provide a method to produce very high molecular mass polymers by step-growth polymerization [35], [36]. In interfacial polymerization, polymers are formed at or in the vicinity of the phase boundary of two immiscible monomer solutions. This technique requires an extremely fast polymerization. The best reaction type for step-growth polymerization would be Schotten – Baumann reactions involving acid chlorides. For example, polyamidation is performed at room temperature by placing an aqueous solution of diamine over an organic phase containing the diacid chloride. The polymer formed at the interface can be pulled off as a continuous film or filament. The amine – acid chloride reaction rate is so fast that the polymerization becomes diffusion controlled. Once the polymer molecules begin to grow and monomer molecules start to add to polymer chain ends, incoming monomer molecules tend to react with polymer chain ends before they can penetrate through the polymer film to start the growth of new chains. Thus, polymers with much higher molecular masses are formed. Since the reaction is diffusion controlled, there is no need to start with an exact balance of the two monomers. The lower temperatures used reduce the relative rates of side reactions, and, therefore, the purity of monomers is not as important as with most other step-growth polymerizations. In spite of the advantages that interfacial polymerization offers, this process has not attracted wide industrial use, mainly because of the high cost of the required reactive monomers and the large amount of solvent which must be removed and recovered.

2.1.3. Nonlinear Polymerization

Another important class of polymers produced by step-growth polymerization are nonlinear polymers formed by polymerization of monomers with more than two functional groups per molecule. Some of the nonlinear polymers produced commercially by step-growth polymerization are shown in Figure 4.

In the course of network formation, a polymer molecule of effectively infinite molecular mass may be formed. At this point, termed the gel point, the visible formation of a gel or insoluble polymer fraction is observed. The gel molecule is insoluble in a good solvent even at elevated temperatures under conditions at which degradation does not occur. Various physical properties of the system change abruptly at the gel point. Gelation should be understood as a critical phenomenon having similarities with other critical phenomena such as vapor – liquid condensation, nuclear chain reactions, and ferromagnetism.

It was Carothers [20], who first derived an equation for the extent of reaction at the gel point. He defined a gel molecule as one with infinite molecular mass. His criterion that gelation occurs when the number-average chain length $\bar{P}_N$ goes to infinity is not acceptable, since polymer molecules larger than $\bar{P}_N$ are always present and will become gel molecules earlier than this hypothetical gel point. However, the concept of the "infinitely large molecule" was fully established by Flory [37–39] using a statistical approach. His criterion for the onset of gelation is that it occurs when the weight-average chain length $\bar{P}_W$ goes to infinity. Since a gel molecule is the largest molecule in the reaction system, higher-order moments of the molecular mass distribution could also be used to determine the gel point. Fortunately, the second- and higher-order moments approach infinity simultaneously, at least for batch polymerizations [40], [41], and the criterion of infinite $\bar{P}_W$ is acceptable.

Flory devised a simple tree-like model, as shown in Figure 5, and used the following simplifying assumptions:

1) All functional groups of the same type are equally reactive
2) All functional groups react independently of one another

3) No intramolecular reactions occur in finite species

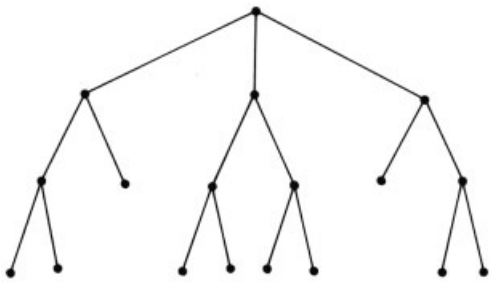

Figure 5. A schematic drawing of Flory's tree-like model (functionality $f = 3$)
The tree-like model is called the Bethe Lattice or Cayley tree by physicists

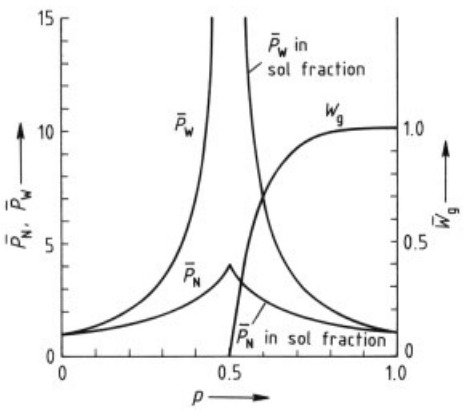

Figure 6. Molecular mass change and gel growth during network formation (functionality $f = 3$)

His basic proposal was that the gel point is reached when the expectancy of finding the next generation in a particular existing molecule is unity. For the tri-functional monomer units shown in Figure 5, the conversion at the gel point is given by $p_c = 1/(f - 1) = 0.5$, where f is the functionality of a monomer unit. His model was a brilliant development and it provides the starting point for most theories of polymer network formation. A few years later Stockmayer [42–44] further developed Flory's idea based on the most-probable size distribution, and their theory is usually called the Flory – Stockmayer theory. Examples of the calculated development of the number- and weight-average chain lengths of the sol fraction and of the weight fraction of gel are shown in Figure 6. Since the gel is a molecule with many functional groups, the gel once formed acts like a giant sponge, rapidly consuming sol polymer molecules.

This tree-like concept was generalized by Gordon et al. [45], [46] based on the theory of stochastic branching processes, which is considered to be a part of Graph Theory [47], [48]. This technique involves abstract mathematics and requires the derivation of the probability generating functions. The method is general but rather difficult to use for real problems. To avoid the use of probability generating functions, other probabilistic methods have been proposed [49–52]. Among them the Macosko – Miller model [50–52] using conditional probabilities is becoming popular due to its simplicity. All the models mentioned above are fully equivalent, that is, only the mathematical language is slightly different. These statistical models, which are sometimes called the classical theories, have a long history and have proven their power of refinability to accommodate highly system-specific effects such as unequal reactivity [53], [54], substitution effects [55], [56], and intramolecular reactions [18], [19], [57–59], which are important in real systems.

One drawback of the statistical theories mentioned above is that they assume an equilibrium system (i.e., the size distribution is calculated anew at each time) and they do not consider the kinetic buildup of the system. Therefore, the classical theories may not be applicable for kinetically-controlled systems. It has been shown that although there is no difference between equilibrium and kinetically-controlled systems under Flory's simplifying assumptions, the difference becomes significant as the conditions deviate further from Flory's assumptions [60], [61], so that in real systems the kinetic features may be dominant. It has been argued that the kinetic buildup can also be accounted for by using a statistical approach [62].

Another disadvantage of statistical approaches may be the excessive modifications required to generalize them for different reactor types (e.g., continuous reactors). The kinetics approach was originally shown in the appendix of a paper by Stockmayer [42]. Based on the chemical kinetics, the reaction rate would be proportional to the product of the number of unreacted functional groups in the respective reaction partner, so that an infinite set of differential

equations similar to Equations (2.11) and (2.12) can be set up. This idea has been applied to polymeric systems [60], [63–70].

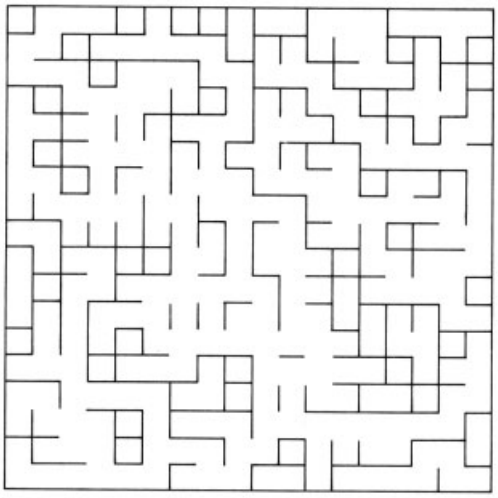

Figure 7. Example of percolation at the gel point in a square lattice ($p_c = 0.5$) [78]

All the theories mentioned above belong to a mean-field theory. On the other hand, the percolation theory [71–77], which is considered to be equivalent to a non-mean-field theory, has been applied to polymeric gelation [78], [79]. The percolation theory is usually associated with a lattice model to describe network structure. One of the simplest examples is the two-dimensional lattice shown in Figure 7. In this figure, each bond which has been formed is shown as a short line connecting two monomers, though the monomers are not shown. In the random (standard) percolation theory each site of a very large lattice is occupied randomly with probability p, independent of its neighbors. Some nearly "infinite" molecules can be seen in Figure 7, where "infinite" means that they span the whole sample. Mathematical methods to calculate this threshold exactly are restricted so far to two dimensions [77], and therefore, for practical calculations the Monte Carlo simulation is usually used. It is easy to understand why gelation is a critical phenomenon from the lattice model, because in the vicinity of the gel point only a few additional bonds are necessary to form a molecule which spans the whole sample. The percolation theory emphasizes the universality of critical phenomena and space dimensionality. De Gennes wrote in his book [80] that "it took more than thirty years to convince experimentalists that mean-field theory was wrong". However, at present the percolation models are far from simulating actual network formation quantitatively, because the bonds are too rigid, the movement of molecules is too suppressed, and necessary chemical rules of bond formation are ignored. The percolation theory is essentially devoted to describing the behavior near the critical threshold p_{cr}, where the system-specific features are not important.

Network polymers are increasingly used as engineering materials because of their excellent stability toward elevated temperature and physical stress. Since the three-dimensional polymers are neither soluble nor fusible once made, the final stage of polymerization is usually carried out in a mold of the desired shape.

2.2. Chain-Growth Polymerization

Chain-growth polymerization is initiated by a reactive species, R^*_{in}, produced from an initiator or catalyst I.

$$I \longrightarrow n\,R^*_{in} \tag{2.13}$$

Depending on the type of active center, chain-growth polymerization can be divided into free-radical, anionic, and cationic polymerization. The reactive species R^*_{in} adds to a monomer to form a new active center, and monomer molecules are added to the active center successively. This process is called the propagation reaction:

$$R^*_{in} \xrightarrow{+M} P^*_1 \xrightarrow{+M} P^*_2 \xrightarrow{+M} \cdots \tag{2.14}$$

where M represents a monomer molecule, and $P*_r$ is an active polymer molecule with chain length r. In general, the propagation reaction is represented by

$$P^*_r + M \xrightarrow{K_p} P^*_{r+1} \tag{2.15}$$

In chain-growth polymerization, only molecules with an active center can propagate, so that polymer molecules once formed may be considered dead polymer for linear chain-growth polymerization. Dead polymer

molecules do not take part as reactants thereafter. The active center is always on the chain end when linear chains are being produced exclusively. Polymer chain growth is terminated at some point by unimolecular and/or bimolecular termination. Bimolecular termination of active centers occurs only in free-radical polymerization.

Carbon – carbon double bonds and the carbon – oxygen double bond in aldehydes and ketones are the two main types of functional groups which undergo chain-growth polymerization. The polymerization of the carbon – carbon double bond is much more important, as most commercial monomers with carbon – carbon double bonds readily undergo free-radical polymerization (an important exception is propene). The carbonyl bond is not generally susceptible to polymerization by radical initiators due to its highly polarized structure. Another reason is that most of the carbonyl monomers (except formaldehyde) possess very low ceiling temperatures [81], [82] (the temperature above which active polymer chains depolymerize rather than grow).

Most of the commercial vinyl monomers (CH_2=CHX and CH_2=CXY, and monomers in which fluorine is substituted for hydrogen) can be polymerized with free radicals. Whether a vinyl monomer can be polymerized by anionic or cationic mechanisms strongly depends on the type of monomer. Monomers with electron-donating groups attached to the doubly bonded carbon atoms form stable carbenium ions and polymerize best with cationic initiators. Conversely, monomers with electron-withdrawing substituents form stable carbanions and require anionic initiators. It should be noted that ions of low stability would be expected to react with carbon – carbon double bonds; however, in many cases they cannot be formed or else are easily consumed by side reactions.

2.2.1. Free-Radical Polymerization

Generally, free-radical polymerization consists of four types of elementary reaction.

1) Initiation reactions, which continuously generate radicals during polymerization.

$$\mathrm{I} \xrightarrow{K_d} n\,\mathrm{R}^{\bullet}_{in} \tag{2.16}$$

$$\mathrm{R}^{\bullet}_{in} + \mathrm{M} \xrightarrow{K_i} \mathrm{R}^{\bullet}_{1} \tag{2.17}$$

The stoichiometric coefficient n is two for thermal decomposition of initiators. A free-radical $\mathrm{R}^{\bullet}_{in}$ derived from the initiator is called a primary or initiator radical.

2) Propagation reactions, which are responsible for the growth of polymer chains by addition of monomer to a radical center.

$$\mathrm{R}^{\bullet}_{r} + \mathrm{M} \xrightarrow{K_p} \mathrm{R}^{\bullet}_{r+1} \tag{2.18}$$

3) Bimolecular termination reactions between two radical centers, which give a net consumption of radicals. These consist of disproportionation (Eq. 2.19) and combination (Eq. 2.20).

$$\mathrm{R}^{\bullet}_{r} + \mathrm{R}^{\bullet}_{s} \xrightarrow{K_{td}} \mathrm{P}_{r} + \mathrm{P}_{s} \tag{2.19}$$

$$\mathrm{R}^{\bullet}_{r} + \mathrm{R}^{\bullet}_{s} \xrightarrow{K_{tc}} \mathrm{P}_{r+s} \tag{2.20}$$

where P_r is a polymer molecule of chain length r and does not have a radical center, while a polymer radical (or macroradical) of chain length r has the symbol $\mathrm{R}^{\bullet}_{r}$.

4) Chain transfer to small molecules which causes the cessation of growth of polymer radicals while generating small transfer radicals simultaneously. Chain-transfer reactions do not give a net consumption of radicals, and if the transfer radicals are as reactive as polymer radical (or more reactive) these reactions should not affect the polymerization rate or monomer consumption rate when the bimolecular termination reactions are chemically controlled. Chain-transfer reactions to small molecules reduce the size of polymer radicals and therefore would increase bimolecular termination rates when these reactions are diffusion controlled (bimolecular termination rates may be chain-length dependent under these conditions).

$$R_r^{\bullet} + X \xrightarrow{K_{rx}} P_r + X^{\bullet} \quad (2.21)$$

$$X^{\bullet} + M \xrightarrow{K_p'} R_1^{\bullet} \quad (2.22)$$

X may be monomer, a solvent molecule, or a chain-transfer agent. When X is a polymer molecule, polymer molecules with long-chain branches are formed. Long-branch formation is discussed in Section 2.2.1.7.

The sequence of elementary reactions, in Equations (2.16) – (2.22) results in total radical concentrations of the order $10^{-9} - 10^{-5}$ mol/L for most commercial polymerizations. Since polymer molecules with high molecular masses are produced from the very start of polymerization, the reacting solution can be quite viscous over most of the monomer conversion range. The high viscosities not only cause problems in mixing and heat removal, but also can affect reaction rates (reactions such as bimolecular termination of polymer radicals). This topic is discussed in Section 2.2.1.3.

Free-radical polymerization is the most commonly used method for the synthesis of polymers from vinyl and divinyl monomers. Some typical monomers which readily undergo free-radical polymerization are ethylene, styrene, vinyl chloride, vinylidene chloride, acrylonitrile, vinyl acetate, methyl methacrylate, methyl acrylate, acrylamide, etc. Of all chain-growth polymerization processes, it is the most widely studied and best understood.

2.2.1.1. Initiation

Free radicals may be generated in a monomer in a number of ways. The most often used method is to add chemical initiators, such as azo and peroxide compounds, to the monomer in low concentrations (usually < 1 wt % based on monomer). When heated, the initiator decomposes, generating radicals which act as active centers for monomer addition. For example, organic peroxides (ROOR′) decompose thermally by O – O bond cleavage to produce two initiator radicals as follows (other side reactions may of course occur).

$$ROOR' \xrightarrow{K_d} RO^{\bullet} + R'O^{\bullet} \quad (2.23)$$

where K_d is a thermal decomposition rate constant with units of inverse time, most often s^{-1}. For a batch reactor, the change in the number of moles of initiator N_I is given by

$$dN_I/dt = -K_d N_I \quad (2.24)$$

For isothermal decomposition (isothermal polymerization) Equation (2.24) can be integrated analytically to obtain

$$N_I = N_{I0} \exp(-K_d t) \quad (2.25)$$

where N_{I0} is the number of moles of initiator at time $t = 0$.

Thus the half-life of an initiator is given by

$$t_{1/2} = -\ln(0.5)/K_d = 0.693/K_d \quad (2.26)$$

Knowledge of K_d for an initiator therefore permits calculation of the initiator half-life $t_{1/2}$. Since K_d has an Arrhenius temperature dependence, K_d and $t_{1/2}$ both depend on temperature.

Activation energies for peroxides and azo initiators are ca. 120 kJ/mol, so the decomposition rate is highly temperature dependent, and the useful temperature range is quite small (decomposition rate is either too fast or too slow outside the useful temperature range, which normally spans about 30 °C).

To complete the initiation step, the initiator radicals ($R_{in}^{\bullet}$) must add to the double bond of a monomer molecule to generate a polymer radical of unit chain length $R_1^{\bullet}$. In most polymerizations, this step (Eq. 2.17) is much faster than the rate of initiator decomposition (Eq. 2.16). The homolysis of the initiator is the rate-determining step in the initiation sequence, and the initiation rate R_I is given by

$$R_I = 2K_d f [I] \quad (2.27)$$

where f is the initiator efficiency. The initiator efficiency is defined as the fraction of radicals produced by initiator decomposition that initiate polymer radicals. Note that not every initiator molecule, which in principle can produce two polymer radicals, does so. Some primary radicals may react with themselves or with other molecules to form stable species which do not form either polymer radicals or molecules. It should be noted that chain transfer to the initiator does not decrease f. The initiator efficiency

usually has values in the range 0.2 – 1.0 at low monomer conversions, where polymer concentrations are low. The major cause of low initiator efficiency is recombination of the radical pairs before they diffuse apart, which is called the cage effect [83], [84]. When an initiator decomposes, the primary radicals $R^{\bullet}_{in}$ are nearest neighbors for about $10^{-10} - 10^{-9}$ s. During this interval they are surrounded by a "cage" of solvent and monomer molecules through which they must diffuse to escape from the cage. Since reactions between radicals are extremely fast, there is reasonable probability that reaction between primary radicals occurs. Direct recombination may simply regenerate the original initiator molecules, but other reactions can also occur that consume initiator radicals without forming polymer chains. In particular, since azo initiators decompose with the elimination of a nitrogen molecule, recombination of the primary radicals results in the formation of a stable molecule that cannot generate radicals, and thus there may be a significant decrease in initiator efficiency. Efficiency decreases with increasing viscosity of the reaction medium [85], [86]. Thus f decreases during the course of polymerization and may approach zero at very high polymer concentrations where the diffusion coefficient of primary radicals in the "cage" is very small [87], [88].

When selecting an initiator type, in general one needs to consider the decomposition rate constant, water and oil solubility, stability of initiator fragments on chain ends, and other factors. Another important point is the activity of the initiator radical center towards the abstraction of atoms (e.g., hydrogen atoms) from the polymer backbone. This can lead to chain scission, long-chain branching, and possibly cross-linking.

Up to this point, only monofunctional initiators (initiators with one peroxide or one azo group per molecule) have been considered. There are commercially available bifunctional initiators (with two peroxide groups per molecule) with some potentially useful applications [89–92]. Their function can be illustrated as follows:

$$\underset{(\mathrm{I})}{R^1OOR^2OOR^1} \longrightarrow \underset{(R^{\bullet}_{in})}{R^1O^{\bullet}} + \underset{(\hat{R}^{\bullet}_{in})}{{}^{\bullet}OR^2OOR^1} \tag{2.28}$$

Even when the K_d's for both peroxide groups are the same, both groups on the same peroxide molecule do not decompose at the same time; thus a significant fraction of the polymer chains will have a peroxide end group. These terminal peroxide groups will later decompose, generating polymer radicals with an initiator fragment in the backbone. The practical benefits include higher molecular masses at the same temperature or comparable molecular masses at higher temperatures. Polymerization at higher temperatures results in higher productivity. These benefits will only accrue when most of the polymer chains are formed by bimolecular termination of polymer radicals. When chain transfer to small molecules produces most of the polymer chains, these benefits will no longer exist, and since bifunctional initiators are more expensive than monofunctional initiators it is recommended that the latter be used.

The decomposition rates of peroxy and azo compounds can be increased by irradiation with ultraviolet and visible light. Unlike thermal decomposition, the activation energy for photochemical initiation is approximately zero, so polymerization can be initiated at much lower temperatures. Compounds such as benzoin and disulfides, whose bonds are too strong to undergo thermal homolysis, are effective radical initiators under ultraviolet irradiation. Photochemical polymerization has been applied in coatings and inks for metal, paper, wood and plastics, in photo-imaging, printing circuits, and adhesives, although its use is limited by low penetration into the polymerizing mass.

Another method of lowering the activation energy of the peroxide decomposition reaction is to use redox initiation systems. The addition of a reducing agent results in radical formation from an oxidation – reduction reaction between the two components. Generally, the reaction is illustrated as follows.

$$A^{n+} + ROOR' \longrightarrow A^{(n+1)+} + R'O^{\bullet} + RO^{-} \tag{2.29}$$

where A is a reducing agent and ROOR′ is a peroxide. The integer n is two for Fe^{2+} and zero for N,N-dimethylaniline. The fate of the free radical depends on the relative concentration of reducing agent and monomer. Since the redox initiaton reactions do not produce a pair of radicals, the cage effect is not operative. At high monomer

concentrations most of the free radicals initiate polymerization, and ordinary second-order kinetics are obeyed. Redox initiation is usually used in the temperature range 0 – 50 °C.

All methods mentioned above employ initiators. There are, however, other means to initiate polymerization. Styrene (and some substituted styrenes such as *p*-methylstyrene) and methyl methacrylate polymerize at elevated temperatures in the absence of a free-radical initiating system. The accepted mechanism for thermal initiation of styrene is the Mayo mechanism [93], [94], which involves the formation of a Diels – Alder dimer intermediate which reacts with styrene to produce radicals. The Mayo mechanism is consistent with an observed initiation rate which can vary from second to third order in monomer during the course of polymerization [95–97], and some confirming evidence has been reported [98–100]. A serious disadvantage of the use of thermal initiation for styrene is the formation of undesirable byproducts (cyclic dimers and trimers) which are difficult to remove to give a high-quality polystyrene.

Irradiation with UV, high-energy electrons, and γ-rays can initiate polymerization with or without the presence of initiators. Radiation initiation has been used almost exclusively for polymer modification (chain scission, long-chain branching, cross-linking, and grafting). These radiation processes are characterized by a zero activation energy for radical generation and as a consequence a low activation energy for polymerization. Therefore, they are effective at both low and high temperatures. With radiation initiation, polymer molecular masses increase with increasing temperature, which is opposite to that for chemically initiated free-radical polymerizations (the high activation energy for initiator decomposition is responsible for this). Both UV and electrons have small penetration depths and are therefore used for polymerizations in thin layers. Gamma rays have high penetration depths but require expensive safety installations. Radiation polymerization may be initiated by radicals, cations, or anions. The effectiveness of a radical center depends on the chemistry of the monomer and the polymerization conditions. Most radiation polymerizations are free-radical except at very low temperatures where ionic species are sufficiently stable. With radiation initiation various active intermediates may be formed, leading to a very complex reaction mixture with the formation of many byproducts as well as the formation of long branches and possibly cross-linkages. Since photon energies for UV are lower, UV radiation generally gives cleaner polymerizations with the formation of linear chains, although monomers which undergo photolysis by UV radiation are limited.

2.2.1.2. Propagation

The propagation reaction (Eq. 2.18) controls both the rate of growth and the structure of the polymer chain. Monomers which undergo free-radical polymerization are commonly monosubstituted or 1,1-disubstituted ethylenes, CH_2=CHX or CH_2=CXY. With 1,1-disubstituted ethylenes both substituents should not be large, since propagation would be sterically hindered. 1,2-Disubstituted ethylenes are normally considered very difficult to polymerize since the approach of the propagating radicals to a monomer is sterically hindered. 1,2-Disubstituted ethylenes can, however, often be incorporated into copolymers.

Due to steric and resonance effects, vinyl monomers predominantly undergo head-to-tail addition:

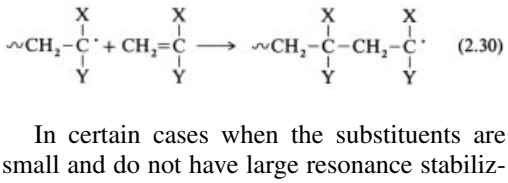

(2.30)

In certain cases when the substituents are small and do not have large resonance stabilizing effects, head-to-head propagation may occur. For example, approximately 16 % head-to-head placement has been reported for poly(vinyl fluoride) [101].

In free-radical polymerization, chain microstructure is largely independent of initiation mechanism and initiator type. Polymers produced by free-radical polymerization are largely atactic, since the terminal carbon – carbon bond can rotate freely during chain growth. The configuration of a monomer unit in the chain is not determined during its addition to the radical center but only when the next monomer molecule adds to it. The slight preference for syndiotactic over isotactic placement is caused by steric and/or electrical repulsion between

substituents in the chain, although at high temperatures their effects are progressively diminished. For example, the fraction of syndiotactic diads of poly(vinyl chloride) changes from 0.67 to 0.51 as the synthesis temperature increases from − 78 °C to 120 °C [102]. For methyl methacrylate, it is 0.86 at − 40 °C and 0.64 at 250 °C [103], [104].

For most chain-growth polymerizations (free-radical or ionic) the propagation reactions are reversible at elevated temperature and the rate of depropagation is significant [105], [106].

$$R_r^{\bullet} + M \underset{K_{dp}}{\overset{K_p}{\rightleftharpoons}} R^{\bullet}_{r+1} \tag{2.31}$$

where K_{dp} is the rate constant for depolymerization (depropagation). The ceiling temperature T_c, which is the temperature above which active polymer chains depolymerize rather than grow, is reached when the propagation and depropagation rates are equal. Based on thermodynamic arguments, the ceiling temperature can be related to the equilibrium monomer concentration $[M]_c$ as follows:

$$T_c = \Delta H / \left(\Delta S^0 + R \ln [M]_c\right) \tag{2.32}$$

where ΔH is the heat of polymerization, ΔS^0 the entropy change of polymerization at unit monomer concentration, and R the gas constant.

The ceiling temperature T_c is not a singular value but is a function of monomer concentration. At any temperature, a concentration of monomer exists at which the reaction in Equation (2.31) is at equilibrium. The existence of this equilibrium concentration prevents monomer conversion reaching 100 %.

Normally, this equilibrium monomer concentration is too low to detect. A notable exception is α-methylstyrene whose T_c for 100 % monomer concentration is 61 °C, and the equilibrium monomer concentration at 25 °C is 2.2 mol/L [107].

Conventionally, it has been assumed that the propagation rate constant K_p is independent of chain length. Experimental results have shown that K_p is independent of chain length at least for chain lengths > 16 for styrene and > 62 for methyl methacrylate [108]. The propagation rate constant K_p is relatively insensitive to the viscosity of the system except at very high polymer concentrations [88], [109].

It has long been recognized that some bulk polymerizations stop at well below 100 % conversion [110]. This phenomenon has successfully been explained as due to a glassy-state transition of the polymerizing mass [111]. Although the polymerization can proceed very slowly in the glassy state [109], [112], for practical purposes it can be assumed that polymerization stops when the system changes from a viscous liquid to a solid glass. It has been proposed that the initiator efficiency f approaches zero when the system reaches a glassy state [87], [113] and that this is mainly responsible for the cessation of polymerization.

2.2.1.3. Termination

An active center on a growing polymer radical may be destroyed by a variety of processes, including termination by added substances. The latter reactions are called inhibition and retardation processes, and are not considered here. This section discusses bimolecular termination reactions between polymer radicals. Although one of the radicals involved in bimolecular termination may be an initiator radical, under normal polymerization conditions such reactions may be negligible since the concentration of initiator radicals is much smaller than that of polymer radicals.

Bimolecular termination of two polymer radicals can occur by combination or coupling:

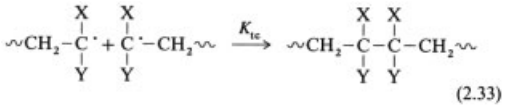

(2.33)

or by disproportionation, in which case a hydrogen radical is transferred from one polymer chain to the other. The result is the formation of two polymer molecules, one of which has a terminal double bond.

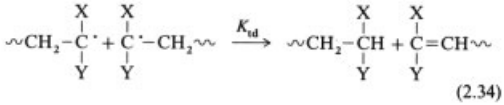

(2.34)

Termination by combination and disproportionation can occur simultaneously, and the relative importance of these two modes of termi-

nation depends on monomer type and polymerization temperature. Experimental data are not available for many monomers; however, radicals which undergo termination by combination appear exclusively to have the structure (**1**) [114]. A well-known example is styrene, which experiences termination by combination almost exclusively over a wide range of temperatures [114], [115]. On the other hand, radicals which undergo disproportionation and combination may have the structure (**2**).

1 **2**

For methyl methacrylate, combination and disproportionation are both important at low temperature, with disproportionation becoming the dominant mode at high temperatures [116–118].

Since bimolecular termination reactions are intrinsically very fast, these reactions are likely to be diffusion controlled when they involve radical centers on polymeric reactants. The autoacceleration of polymerization rate that is a consequence of diffusion-controlled termination is usually called the gel effect or Trommsdorff – Norrish effect. Figure 8 illustrates the autoacceleration in rate for the polymerization of methyl methacrylate [119]. The interpretation proposed was that the increase in rate is a consequence of a decrease in the rate of termination, due to the large increase in viscosity of the reacting medium, thus giving an increase in radical concentration.

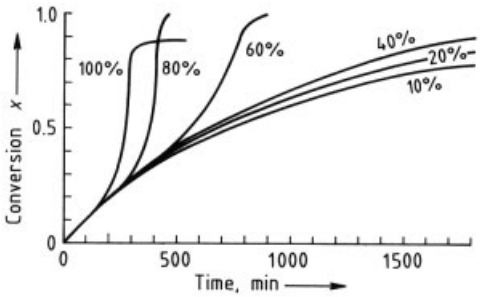

Figure 8. Conversion – time histories for the polymerization of methyl methacrylate in benzene initiated by benzoyl peroxide at 50 °C
The different curves are for various concentrations of monomer in solvent [119]

In general, the termination rate constant should depend on the size of the polymeric radical reactants, the concentration and molecular mass distribution of the accumulated polymer, solvent type, and temperature. As the concentration of entanglement points increases during the course of polymerization, the functional dependence of the termination rate constant on the chain lengths of the polymeric radical reactants should change dramatically. At some high monomer conversion (high polymer concentration) when the polymeric radicals are trapped (diffusion of the center of mass of polymer chains is essentially zero), radical centers may continue to move (due to monomer addition by propagation) and undergo bimolecular termination [118]. With this form of termination, the termination rate constant should be independent of the chain lengths of the polymeric radical reactants. Originally, upon examination of the autoacceleration of the conversion – time curve, it was believed that the termination reaction only became diffusion controlled at some monomer conversion greater than zero and that this occurred when the polymer chains were sufficiently entangled (with a sufficient number of physical entanglement points). It has been recognized that the bimolecular termination reactions may be diffusion controlled even at zero monomer conversion (zero polymer concentration) [114], [120–122]. At low monomer conversions, where polymer chains in a good solvent are isolated coils, translational diffusion of the center of mass of the chains is sufficiently rapid, and the limiting step is the so-called segmental diffusion rate of the radical center in the coil. With limited interpenetration of the coils, a finite and significant time is required for the radical centers on the partially penetrating coils to meet in a suitable reaction volume where bimolecular termination may occur. The probability of finding the radical centers in a suitable reaction volume near the coil surfaces decreases as the coil sizes increase, and, therefore, the termination constant should increase with decreasing polymer molecular mass, with decreasing goodness of the solvent (which depends on the nature and concentration of monomer, inert solvent, and polymer), and with increasing polymer concentration (increasing monomer conversion). It is well known that in dilute solution the poly-

mer coil size decreases with increasing polymer concentration.

At somewhat higher conversions, when there are a sufficient number of chain entanglement points, the translational diffusion rate of the center of mass of polymer coils decreases dramatically and bimolecular termination rates become translationally diffusion controlled. Of course, shorter polymer chains will experience translational diffusion-controlled termination at higher monomer conversions (higher polymer concentrations) than longer chains, and clearly the bimolecular termination rate constant will be chain-length dependent and the bivariate distribution $K_t(r, s)$ will change its shape dramatically with increasing polymer concentration. Finally, when the polymer coils are trapped, K_t should become independent of chain length [118].

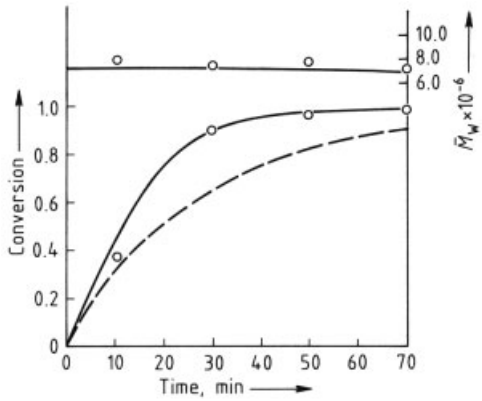

Figure 9. Polymerization of acrylamide – rate and molecular mass development ($T = 60\,°C$, $[M]_0 = 3.4$ mol/L, initiated by potassium peroxosulfate $[I]_0 = 5.2 \times 10^{-4}$ mol/L) – – – = Constant K_t; —·— = Use of Equation (2.35) [126]

The mechanisms of diffusion-controlled reactions in polymer systems are being clarified both theoretically and experimentally (see for example [122]), but at present developing a general formulation for K_t for the whole course of polymerization is a formidable task. Considering the complexity of the mechanisms of diffusion-controlled termination reactions, in some cases it may be necessary to use an empirical approach for reactor calculations [96], [97], [123–126]. For example, K_t may be approximated by

$$K_t = K_{t0} \exp\left[-\left(A_1 x + A_2 x^2 + A_3 x^3\right)\right] \tag{2.35}$$

where x is the monomer conversion, K_{t0} is the termination rate constant at zero monomer conversion ($x = 0$), and A_1, A_2, A_3 are adjustable parameters. An application is shown in Figure 9. The adjustable parameters A_1, A_2, A_3 are usually estimated by fitting isothermal conversion – time curves. The adjustable parameters should be functions of temperature and possibly initiator concentration (radical initiation rate). Note that K_t estimated by fitting polymerization rate and number-average molecular masses is a number-average termination rate constant. Although, this termination rate constant may predict rates of polymerization and number-average molecular masses adequately, its use to calculate higher averages will underestimate weight-average, and Z-average molecular mass [88].

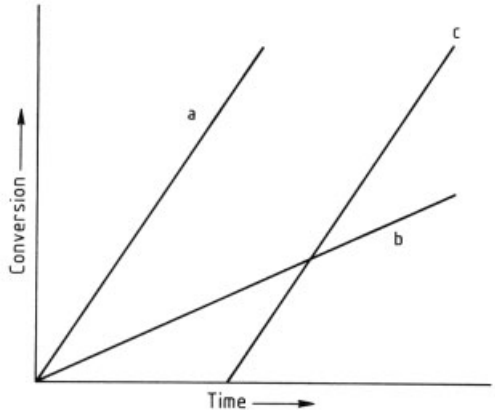

Figure 10. Effect of inhibitors and retarders
a) No retarder or inhibitor; b) With retarder; c) With inhibitor

2.2.1.4. Chain Transfer to Small Molecules

During free-radical polymerization, chain transfer to small molecules X may occur. The small molecule may be initiator, monomer, chain-transfer agent, solvent, inhibitor, or impurity. In general, these chain-transfer reactions can be represented by Equations (2.21) and (2.22).

$$R_r^{\bullet} + X \xrightarrow{K_{rX}} P_r + X^{\bullet} \tag{2.21}$$

$$X^{\bullet} + M \xrightarrow{K_p'} R_1^{\bullet} \tag{2.22}$$

When K'_p is approximately zero (i.e., $X^\bullet$ is a stable radical) X is called an inhibitor. If K'_p is smaller than the propagation rate constant K_p, X is called a retarder. Idealized behavior of inhibitors and retarders is shown schematically in Figure 10. The kinetics of inhibition and retardation can be found elsewhere [127].

For an added agent to be a chain-transfer agent, K'_p must be approximately equal to K_p (or $K_p < K'_p$, if chain length is large enough). Therefore, the chain-transfer agent reduces the molecular masses but does not affect rates of polymerization.

The chain-transfer rate constants for most monomers are about 10^4–10^5 times smaller than the propagation rate constant ($K_{fm}/K_p = 10^{-5}$–10^{-4}).

The presence of monomer molecules is inevitable, so that the value of K_{fm}/K_p places an upper limit on the polymer molecular mass that can be obtained with a given monomer. Larger K_{fm} values are observed when the propagating radicals have very high energies (high reactivities), such as in the case of ethylene, vinyl acetate, and vinyl chloride.

2.2.1.5. Kinetics of Linear Polymerization

The elementary reactions involved in linear free-radical polymerization (chains produced are linear with no branches or cross-links) are as follows:

Initiation

$$I \xrightarrow{K_d} 2\,R^\bullet_{in}$$

$$R^\bullet_{in} + M \xrightarrow{K_i} R^\bullet_1$$

Propagation

$$R^\bullet_r + M \xrightarrow{K_p} R^\bullet_{r+1}$$

Chain transfer to monomer

$$R^\bullet_r + M \xrightarrow{K_{fm}} P_r + R^\bullet_1$$

Chain transfer to small molecule (T)

$$R^\bullet_r + T \xrightarrow{K_{fT}} P_r + T^\bullet$$

$$T^\bullet + M \xrightarrow{K'_p} R^\bullet_1$$

Termination by disproportionation

$$R^\bullet_r + R^\bullet_s \xrightarrow{K_{td}} P_r + P_s$$

Termination by combination

$$R^\bullet_r + R^\bullet_s \xrightarrow{K_{tc}} P_{r+s}$$

To derive the kinetic rate equations, the following assumptions are usually made:

1) All rate constants are independent of chain length.
2) Chain lengths are sufficiently large that the total rate of monomer consumption may be equated to the rate of monomer consumption by the propagation reactions alone [this is often called the long-chain approximation (LCA)].
3) Radicals generated in chain-transfer reactions propagate with monomer rapidly and thus do not affect the polymerization rate.
4) The stationary-state hypothesis (SSH) is valid for radical reactions. One can therefore assume that both the rates of radical generation and consumption are much greater than the rate of change of radical concentration with respect to time [128], [129].

Let us first derive an expression for the polymerization rate R_p, applying the above assumptions. The balanced equation for polymer radicals with chain length r is given by

$$\frac{1}{V}\frac{d\,(V\,[R^\bullet_1])}{dt} = R_I + K_{fm}\,[M]\sum_{r=2}^{\infty}[R^\bullet_r] + K'_p\,[T^\bullet]\,[M] - K_{fT}\,[R^\bullet_1]\,[T] - K_p\,[R^\bullet_1]\,[M] - (K_{tc}+K_{td})\,[R^\bullet_1]\,[R^\bullet] \quad (2.36)$$

$$\frac{1}{V}\frac{d\,(V\,[R^\bullet_r])}{dt} = K_p\,[R^\bullet_{r-1}]\,[M] - K_p\,[R^\bullet_r]\,[M] - K_{fm}\,[R^\bullet_r]\,[M] - K_{fT}\,[R^\bullet_r]\,[T] - (K_{tc}+K_{td})\,[R^\bullet_r]\,[R^\bullet] \quad (r \geq 2) \quad (2.37)$$

where R_I is the initiation rate ($R_I = 2K_d f\,[I]$) and $[R^\bullet] = \sum_{r=1}^{\infty}[R^\bullet_r]$, which is the total polymer radical concentration. The transfer radical concentration $[T^\bullet]$ is given by

$$\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{T}^{\cdot}])}{\mathrm{d}t}=K_{\mathrm{fT}}\,[\mathrm{R}^{\cdot}]\,[\mathrm{T}]-K'_{\mathrm{p}}\,[\mathrm{T}^{\cdot}]\,[\mathrm{M}] \tag{2.38}$$

Applying the stationary-state hypothesis gives

$$K_{\mathrm{fT}}\,[\mathrm{R}^{\cdot}]\,[\mathrm{T}]=K'_{\mathrm{p}}\,[\mathrm{T}^{\cdot}]\,[\mathrm{M}] \tag{2.39}$$

Summation of Equations (2.36) and (2.37) over all chain lengths (1 to infinity) and substituting Equation (2.39) into the sum gives

$$\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{R}^{\cdot}])}{\mathrm{d}t}=\sum_{r=1}^{\infty}\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{R}^{\cdot}_{r}])}{\mathrm{d}t}$$
$$=R_{\mathrm{I}}+K'_{\mathrm{p}}\,[\mathrm{T}^{\cdot}]\,[\mathrm{M}]-K_{\mathrm{fT}}\,[\mathrm{R}^{\cdot}]\,[\mathrm{T}]-(K_{\mathrm{tc}}+K_{\mathrm{td}})\,[\mathrm{R}^{\cdot}]^2$$
$$=R_{\mathrm{I}}-(K_{\mathrm{tc}}+K_{\mathrm{td}})\,[\mathrm{R}^{\cdot}]^2 \tag{2.40}$$

Application of the stationary-state hypothesis for the total polymer radical concentration $[\mathrm{R}^{\bullet}]$, gives

$$R_{\mathrm{I}}=K_{\mathrm{t}}\,[\mathrm{R}^{\cdot}]^2 \tag{2.41}$$

where $K_{\mathrm{t}}=K_{\mathrm{tc}}+K_{\mathrm{td}}$. In the above formalism, the termination rate, R_{t} is given by

$$R_{\mathrm{t}}=K_{\mathrm{t}}\,[\mathrm{R}^{\cdot}]^2 \tag{2.42}$$

It is worth noting here that $R_{\mathrm{t}}=2\,K_{\mathrm{t}}\,[\mathrm{R}^{\bullet}]^2$ is often used in the literature, although $R_{\mathrm{t}}=K_{\mathrm{t}}\,[\mathrm{R}^{\bullet}]^2$ is more widely used in free-radical polymerization (e.g., the compilation of kinetic rate constants in [130]). One must distinguish carefully which type of termination rate constant is being used when consulting the literature on polymerization kinetics.

From Equation (2.41), the total polymer radical concentration is given by

$$[\mathrm{R}^{\cdot}]=(R_{\mathrm{I}}/K_{\mathrm{t}})^{0.5} \tag{2.43}$$

Based on the long-chain approximation, the polymerization rate R_{p} is given by

$$R_{\mathrm{p}}=-\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{M}])}{\mathrm{d}t}=K_{\mathrm{p}}\,[\mathrm{R}^{\cdot}]\,[\mathrm{M}]$$
$$=\left(\frac{K_{\mathrm{p}}}{K_{\mathrm{t}}^{0.5}}\right)R_{\mathrm{I}}^{0.5}\,[\mathrm{M}] \tag{2.44}$$

Since $R_{\mathrm{I}}=2K_{\mathrm{d}}\,f\,[\mathrm{I}]$,

$$R_{\mathrm{p}}=\left(\frac{K_{\mathrm{p}}}{K_{\mathrm{t}}^{0.5}}\right)(2K_{\mathrm{d}}f\,[\mathrm{I}])^{0.5}\,[\mathrm{M}] \tag{2.45}$$

Equation (2.45) predicts a first-order dependenceon monomer concentration and a square-root dependence on initiator concentration, the latter being a direct consequence of the bimolecular nature of the termination reaction.

Now consider the weight chain length distribution $W(r)$. Application of the stationary-state hypothesis for polymer radicals with chain length r (Eq. 2.36 and 2.37) gives.

$$[\mathrm{R}^{\cdot}_{1}] \tag{2.46}$$
$$=\frac{R_{\mathrm{I}}+K_{\mathrm{fm}}\,[\mathrm{M}]\,[\mathrm{R}^{\cdot}]+K_{\mathrm{fT}}\,[\mathrm{T}]\,[\mathrm{R}^{\cdot}]}{K_{\mathrm{p}}\,[\mathrm{M}]+K_{\mathrm{fm}}\,[\mathrm{M}]+K_{\mathrm{fT}}\,[\mathrm{T}]+(K_{\mathrm{tc}}+K_{\mathrm{td}})\,[\mathrm{R}^{\cdot}]}$$

$$[R^{\cdot}_{r}] \tag{2.47}$$
$$=\frac{K_{\mathrm{p}}\,[\mathrm{M}]\left[\mathrm{R}^{\cdot}_{r-1}\right]}{K_{\mathrm{p}}\,[\mathrm{M}]+K_{\mathrm{fm}}\,[\mathrm{M}]+K_{\mathrm{fT}}\,[\mathrm{T}]+(K_{\mathrm{tc}}+K_{\mathrm{td}})\,[\mathrm{R}^{\cdot}]}$$

Let us introduce the following dimensionless groups:

$$\tau=\frac{R_{\mathrm{td}}+R_{\mathrm{f}}}{R_{\mathrm{p}}}=\frac{K_{\mathrm{td}}\,[\mathrm{R}^{\cdot}]+K_{\mathrm{fm}}\,[\mathrm{M}]+K_{\mathrm{fT}}\,[\mathrm{T}]}{K_{\mathrm{p}}\,[\mathrm{M}]} \tag{2.48}$$

$$\beta=\frac{R_{\mathrm{tc}}}{R_{\mathrm{p}}}=\frac{K_{\mathrm{tc}}\,[\mathrm{R}^{\cdot}]}{K_{\mathrm{p}}\,[\mathrm{M}]} \tag{2.49}$$

where

$R_{\mathrm{p}}=K_{\mathrm{p}}\,[\mathrm{R}^{\bullet}]\,[\mathrm{M}]$; propagation rate
$R_{\mathrm{td}}=K_{\mathrm{td}}\,[\mathrm{R}^{\bullet}]^2$; rate of termination by disproportionation
$R_{\mathrm{tc}}=K_{\mathrm{tc}}\,[\mathrm{R}^{\bullet}]^2$; rate of termination by combination
$R_{\mathrm{f}}=K_{\mathrm{fm}}\,[\mathrm{R}^{\bullet}]\,[\mathrm{M}]+K_{\mathrm{fT}}\,[\mathrm{R}^{\bullet}]\,[\mathrm{T}]$; rate of chain transfer.

Since $R_{\mathrm{I}}=R_{\mathrm{td}}+R_{\mathrm{tc}}$, Equations (2.46) and (2.47) can be simplified as follows:

$$[\mathrm{R}^{\cdot}_{1}]=\frac{\tau+\beta}{1+\tau+\beta}\,[\mathrm{R}^{\cdot}] \tag{2.50}$$

$$[R^{\cdot}_{r}]=\frac{1}{1+\tau+\beta}\,[\mathrm{R}^{\cdot}_{r-1}] \tag{2.51}$$

Therefore,

$$[R^{\cdot}_{r}]=[\mathrm{R}^{\cdot}]\,(\tau+\beta)\,\Phi^{r} \tag{2.52}$$

where

$\Phi=1/(1+\tau+\beta)$

Now consider the production rate of polymer molecules with chain length r, $R_{\mathrm{FP}}(r)$, which is given by

$$R_{\mathrm{FP}}(r) = \frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{P}_r])}{\mathrm{d}t} = (K_{\mathrm{fm}}\,[\mathrm{M}] + K_{\mathrm{fT}}\,[\mathrm{T}] + K_{\mathrm{td}}\,[\mathrm{R}^{\cdot}])\,[\mathrm{R}^{\cdot}_r] + \frac{1}{2}K_{\mathrm{tc}}\sum_{s=1}^{r-1}[\mathrm{R}^{\cdot}_s]\,[\mathrm{R}^{\cdot}_{r-s}] \tag{2.53}$$

Substituting for $[\mathrm{R}^{\bullet}_r]$ using Equation (2.52) gives

$$R_{\mathrm{FP}}(r) = K_{\mathrm{p}}\,[\mathrm{R}^{\cdot}]\,[\mathrm{M}]\,(\tau+\beta)\left\{\tau+\frac{\beta}{2}(\tau+\beta)(r-1)\right\}\Phi^r \tag{2.54}$$

The instantaneous weight chain length distribution $W(r)$ is therefore given by

$$W(r) = \frac{rR_{\mathrm{FP}}(r)}{\sum_{r=1}^{\infty} rR_{\mathrm{FP}}(r)} = \frac{(\tau+\beta)\left\{\tau+\frac{\beta}{2}(\tau+\beta)(r-1)\right\}r\Phi^r}{1+\tau+\beta} = (\tau+\beta)\left\{\tau+\frac{\beta}{2}(\tau+\beta)(r-1)\right\}r\Phi^{r+1} \tag{2.55}$$

If $\beta \ll \tau$, that is most polymer chains are formed by chain transfer and/or termination by disproportionation, Equation (2.55) reduces to

$$W(r) = \tau^2 r\Phi^{r+1} = r\left(\frac{1}{1+\tau}\right)^{r-1}\left(\frac{\tau}{1+\tau}\right)^2 \tag{2.56}$$

Here, $1/(1+\tau)$ is the probability of growth for a given polymer radical, and $\tau/(1+\tau)$ is the probability that a polymer radical stops growing. Therefore, Equation (2.56) is essentially the same as the distribution derived for linear step-growth polymerization (Eq. 2.9), based on a statistical argument.

It is sometimes more convenient to describe the chain length distribution as a continuous function rather than a discrete function, and this can often be done with small error as r is usually very large. Therefore, the following approximation may be useful:

$$W(r) \approx (\tau+\beta)\left\{\tau+\frac{\beta}{2}(\tau+\beta)(r-1)\right\}r \cdot \exp\{-(\tau+\beta)\,r\} \tag{2.57}$$

Note that $(\tau+\beta)$ has the value of ca. 10^{-6}–10^{-2} for usual free-radical polymerization $[(\tau+\beta) \ll 1]$.

The weight-average chain length P_{W} for polymer produced instantaneously is given by

$$P_{\mathrm{W}} = \sum_{r=1}^{\infty} rW(r) = \frac{\tau(2+\tau+\beta)+\beta(3+\tau+\beta)}{(\tau+\beta)^2} \approx \frac{2\tau+3\beta}{(\tau+\beta)^2} \tag{2.58}$$

The instantaneous number-average chain length P_{N} is given by

$$P_{\mathrm{N}} = \frac{1}{\sum_{r=1}^{\infty} W(r)/r} = \frac{(1+\tau+\beta)}{(\tau+\beta/2)} \approx \frac{1}{(\tau+\beta/2)} \tag{2.59}$$

The polydispersity index PDI for polymer produced instantaneously is given by

$$\mathrm{PDI} = \frac{P_{\mathrm{W}}}{P_{\mathrm{N}}} \approx \frac{(2\tau+3\beta)(\tau+\beta/2)}{(\tau+\beta)^2} \tag{2.60}$$

If $\beta = 0$, i.e., termination by combination does not occur, the polydispersity index takes on the maximum value, PDI = 2. On the other hand, if $\tau = 0$, i.e., chain termination is solely by bimolecular termination through combination, PDI takes on the minimum value, 1.5. Figure 11 shows the PDI as a function of the fraction of chain termination by bimolecular termination by combination, $\beta/(\tau+\beta)$.

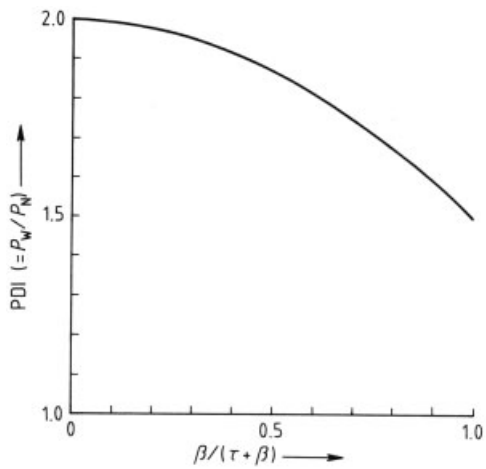

Figure 11. Effect of the type of chain termination on the polydispersity index

The $W(r)$, P_{W}, P_{N}, and PDI derived here give the instantaneous properties. In linear free-radical polymerization the polymer molecules once formed are inert and do not react further.

In general, since the concentrations of monomer, initiator, and chain-transfer agent change with time, the chain length distribution of the accumulated polymer is always broader than the instantaneous distribution. Particularly, when bimolecular termination is strongly diffusion controlled and most of the polymer chains are produced by bimolecular termination, the accumulated distribution broadens significantly with increasing conversion. The polydispersity index PDI for commercial polymers is usually larger than two and this is the result of a drift in molecular mass averages of the instantaneous distribution. The accumulated distribution and its averages can be calculated as follows:

$$\bar{W}(r) = \frac{1}{x}\int_0^x W(r)\,\mathrm{d}x \tag{2.61}$$

$$\bar{P}_\mathrm{W} = \frac{1}{x}\int_0^x P_\mathrm{W}\,\mathrm{d}x \tag{2.62}$$

$$\bar{P}_\mathrm{N} = \frac{x}{\int_0^x \frac{1}{P_\mathrm{N}}\,\mathrm{d}x} \tag{2.63}$$

$$\overline{\mathrm{PDI}} = \frac{\bar{P}_\mathrm{W}}{\bar{P}_\mathrm{N}} \tag{2.64}$$

Superscript bars denote accumulated properties. These integrals may be replaced with equivalent ordinary differential equations which then can be solved by using readily available differential equation solvers. The variation of the kinetic parameters of $W(r)$ with respect to conversion must be known to calculate the accumulated properties. In a similar manner, accumulated molecular mass properties may be calculated for semi-batch and continuous reactors. In a well mixed flow reactor with an ideal residence time distribution (CSTR), the chain length distribution produced is the instantaneous distribution. In batch reactors the distribution is therefore broader [131], [132]. Other mathematical techniques to derive the distribution functions can be found elsewhere [29], [32], [33], [133–135].

2.2.1.6. Effect of Temperature

In free-radical polymerization initiated by the thermal decomposition of an initiator, the polymerization rate is given by Equation (2.45), and the effect of temperature can be estimated by the change in the ratio of three rate constants, $K_\mathrm{p}\,(K_\mathrm{d}/K_\mathrm{t})^{0.5}$. Since each kinetic rate constant is considered to follow the Arrhenius equation, the activation energy of polymerization E_R is given by

$$E_\mathrm{R} = E_\mathrm{p} + \frac{E_\mathrm{d}}{2} - \frac{E_\mathrm{t}}{2} \tag{2.65}$$

where E_p, E_d, E_t are activation energies for propagation, initiator decomposition, and bimolecular termination. Typical values for E_p, E_d, E_t are 30, 120, and 15 kJ/mol, respectively. Therefore, E_R is ca. 80 kJ/mol and this is largely due to the very high activation energy for initiator decomposition. The high activation energy for polymerization means that the rate of polymerization increases strongly with increasing temperature. With redox initiation, E_d is ca. 40 kJ/mol and therefore E_R is considerably smaller at ca. 40 kJ/mol. With radiation initiation, E_d is close to zero and E_R is ca. 20 kJ/mol.

Now consider the effect of temperature on the molecular mass of polymer obtained. For a simple example, consider the case where $\tau \ll \beta$, that is, termination by combination produces most of the polymer chains. Equation (2.58) gives

$$P_\mathrm{W} \propto \frac{1}{\beta} = \frac{K_\mathrm{p}\,[\mathrm{M}]}{K_\mathrm{tc}\,(2K_\mathrm{d} f\,[\mathrm{I}]/K_\mathrm{tc})^{0.5}} \propto \frac{K_\mathrm{p}}{(K_\mathrm{tc}K_\mathrm{d})^{0.5}} \tag{2.66}$$

The activation energy for average chain lengths E_L is given by

$$E_\mathrm{L} = E_\mathrm{p} - \frac{E_\mathrm{d}}{2} - \frac{E_\mathrm{t}}{2} \tag{2.67}$$

With initiators E_L is about -40 kJ/mol. The average chain lengths decrease significantly with increasing temperature when initiators are used and bimolecular termination controls molecular mass development (most of the polymer chains are produced by bimolecular termination). With radiation initiation, E_L is about 20 kJ/mol and molecular masses increase moderately with temperature.

When unimolecular termination (chain transfer to small molecules) dominates in the production of polymer chains, E_L is given by

$$E_\mathrm{L} = E_\mathrm{p} - E_\mathrm{f} \tag{2.68}$$

where E_f is the activation energy for the chain-transfer reaction. E_f depends on the type of chain-transfer agent, but usually $E_p - E_f < 0$, and therefore molecular masses usually decrease with increasing temperature when unimolecular termination (chain transfer to small molecules) controls molecular mass development.

2.2.1.7. Branching Reactions

For the elementary reactions shown in Section 2.2.1.4, radical centers are always located on chain ends, and the dead polymer chains are chemically inert and are all linear. Branched or cross-linked polymer can be obtained by chemical treatment of linear chains using processes such as vulcanization, radiation, and melt processing with peroxides. This section, however, is mainly concerned with branching reactions which may occur during polymerization. Branched and cross-linked polymers are of significant commercial interest, but quantitative interpretation has been limited due to a lack of analytical techniques for the comprehensive characterization of branched polymers. Some of the newer analytical techniques are summarized in [136–138].

Chain Transfer to Polymer. Chain transfer to polymer involves the abstraction of an atom from the backbone of a polymer chain and results in the formation of a backbone radical center. Monomer addition to this radical center produces a tri-branching point and a long-chain branch whose average length is equal to that of the primary chains produced at the same instant.

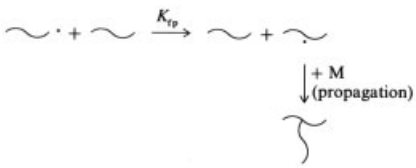

(2.69)

Chain transfer to polymer may be neglected at low monomer conversions as polymer concentrations are low. However, under more practical situations where polymerizations are carried to very high conversions, it may be important. In emulsion polymerization the polymer concentration in polymer particles is relatively high even near zero conversion and therefore long-chain branching reactions are more significant in emulsion than in solution or bulk polymerization. Examples of monomers for which long-chain branching via chain transfer to polymer is important include ethylene and vinyl acetate.

When long-chain branching is important the calculation of the full molecular mass distribution requires excessive computation. However, the leading moments and molecular mass averages can be readily calculated by using the method of moments for an infinite set of differential equations which describe the chemical kinetics [61], [134], [139–147]. The i-th order moment of the polymer distribution can be defined by

$$Q_i = \sum_{t=1}^{\infty} r^i \, [\mathrm{P}_r] \tag{2.70}$$

Therefore, the number- and weight-average chain lengths of the accumulated polymer are given by $\bar{P}_N = Q_1/Q_0$ and $\bar{P}_W = Q_2/Q_1$, respectively. For the elementary reactions shown in Section 2.2.1.4 plus Equation (2.69), some of the lower order moments for batch polymerization are given by [61], [145–147]:

$$\frac{1}{V_0}\frac{\mathrm{d}\,(VQ_0)}{\mathrm{d}x} = \tau + \beta/2 \tag{2.71}$$

$$\frac{1}{V_0}\frac{\mathrm{d}\,(VQ_1)}{\mathrm{d}x} = 1 \tag{2.72}$$

$$\frac{1}{V_0}\frac{\mathrm{d}\,(VQ_2)}{\mathrm{d}x} = \frac{2\,(1+C_{p2})}{\tau+\beta+C_{p1}} + \frac{\beta\,(1+C_{p2})^2}{(\tau+\beta+C_{p1})^2} \tag{2.73}$$

with initial conditions: $Q_0 = Q_1 = \ldots\ldots = 0$ at $x = 0$, where

$$C_{pi} = \frac{K_{fp} Q_i}{K_p\,[\mathrm{M}]}$$

$$\tau = \frac{K_{td}\,[\mathrm{R}^{\cdot}] + K_{fm}\,[\mathrm{M}] + K_{fT}\,[\mathrm{T}]}{K_p\,[\mathrm{M}]}$$

$$\beta = \frac{K_{tc}\,[\mathrm{R}^{\cdot}]}{K_p\,[\mathrm{M}]}$$

and V_0 is the initial volume of the reacting mixture. Figure 12 shows calculated weight-average chain lengths for a range of the kinetic parameters, keeping τ and β constant with conversion. Note that reactions involving the addition of a

polymer radical center to a double bond on the end of a polymer chain are not accounted for in the moment equations above (Eq. 2.71 and 2.73).

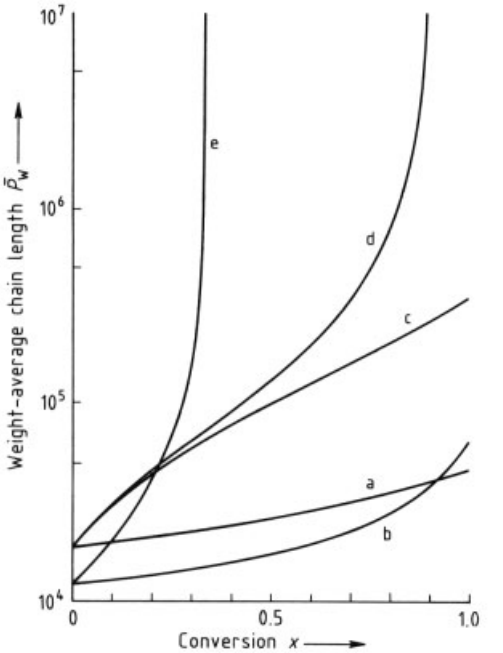

	a	b	c	d	e
τ	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
β	1×10^{-5}	1×10^{-4}	5×10^{-6}	1×10^{-5}	1×10^{-4}
K_{fp}/K_p	1×10^{-4}	1×10^{-4}	1×10^{-3}	1×10^{-3}	1×10^{-3}

Figure 12. Weight-average chain length development during polymerization with chain transfer to polymer [147]

It is of interest to know whether chain transfer to polymer can lead to gelation. FLORY [148], [149] predicted that this reaction type would not cause gelation without the assistance of other interlinking processes. By application of Equation (2.73) it has been clearly shown that chain transfer to polymer plus termination by disproportionation (i.e., there is no termination by combination) can never cause gelation. With termination by combination, however, gelation can occur under certain circumstances [61], [145], [147].

In general, when long-chain branching occurs, the molecular mass distribution broadens. This is especially true for polymerization in a continuous stirred-tank reactor as the reacting mixture always has a relatively high polymer concentration. Comparing long-chain branching frequencies in batch and CSTR reactors at the same monomer conversion levels shows that the residence time distribution of the CSTR favors branching [150], [151].

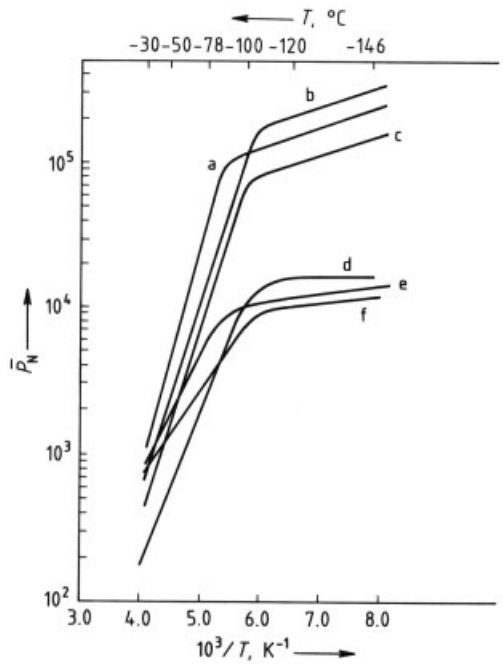

Figure 13. Effect of temperature on the number-average chain length of polyisobutylene and isobutene – isopropene copolymer [161]
a) Polyisobutylene with $AlEtCl_2$ as initiator; b) Polyisobutylene with BF_3 as initiator; c) Polyisobutylene with $AlCl_3$ as initiator; d) Isobutene – isopropene with BF_3 as initiator; e) Isobutene – isopropene with $AlEtCl_2$ as initiator; f) Isobutene – isopropene with $AlCl_3$ as initiator

When the reactivity of polymer radicals is high (as with polyethylene), intramolecular chain transfer (transfer to a location on its own chain) or backbiting may occur. In this case polymer chains with short branches are formed. In the synthesis of low-density polyethylene (LDPE) by high-pressure free-radical polymerization, short-chain branching frequencies are high (typically 20 – 40 branches per 1000 backbone carbon atoms). For polyethylene, the amount of short-chain branching is about ten times greater than long-chain branch-

ing. For polyethylene the backbiting reaction can be represented as follows

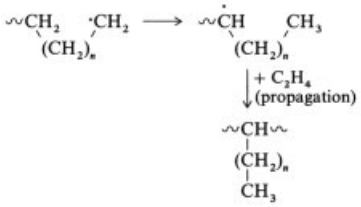

(2.74)

with $n = 3$ and 4 most probable [152].

Reactions with Double Bonds in Polymer Chains. When a radical center on the end of a polymer chain adds to a double bond on the end of a polymer chain, a tri-branching point is formed.

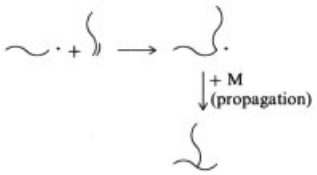

(2.75)

Terminal double bonds on the ends of polymer chains may be formed by chain transfer to monomer, termination by disproportionation, and by β-scission at backbone radical centers.

Reaction with double bonds located on the polymer backbone (internal or pendant double bonds) leads to formation of tetra-branching centers.

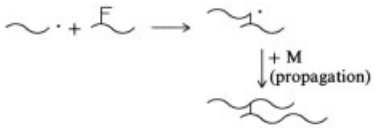

(2.76)

Pendant double bonds are obtained when monomers containing two or more reactive double bonds are polymerized. For example, the homopolymerization of a diene and the copolymerization of vinyl and divinyl monomers produces polymer chains with pendant double bonds. The photopolymerization of multifunctional monomers is used in surface coatings and in the replication of optical discs [153–155]. For densely cross-linked systems the buildup of radical concentration is so rapid that the stationary-state hypothesis may not be valid. Superabsorbent polymers used in diapers and porous organic packagings used in size exclusion chromatography are lightly cross-linked polymeric networks. The kinetics of copolymerization of vinyl and divinyl monomers are discussed in more detail in Section 2.3.3.

2.2.2. Ionic Polymerization

The energy required to form a pair of ions from a neutral molecule is large, and therefore these very unstable ions must be stabilized by solvation at low temperature before polymerization will occur. Polar solvents cannot be used to solvate ions because they are overly reactive and destroy the ionic initiators. Ionic polymerizations are usually carried out at low temperature in solvents of low polarity. These solvents may give ion pairs as well as free ions. Thus, a propagating ionic chain may have a counterion close to the active center during its growth. The proximity of the ion on the growing chain to its counterion depends on the type of counterion, which is determined by the initiator type, and the solvation power of the solvent. Therefore, unlike free-radical polymerization the type of initiator and the nature of the solvent have a large effect on monomer addition during chain growth. The propagation rate constant, therefore, depends not only on temperature but also on the type of initiator and the type and amount of solvent.

For ionic chain-growth polymerizations in solvents with high solvating power where the distance between the propagating active center and the counterion is large, the factors governing the stereochemistry are similar to those for free-radical polymerization. However, in solvents with poor solvating power, there may be extensive coordination between initiator, propagating chain end, and monomer, which results in isotactic (or syndiotactic) placements almost exclusively, i.e., stereospecific polymerization occurs.

Unlike radical polymerization, bimolecular termination between active centers does not occur in ionic polymerization. Termination of an active center on a polymer chain occurs by reaction with the counterion, solvent, monomer, or other species. Often, the initiation reactions are very fast, and the initiator is consumed in the

early stages of polymerization before the polymer chains have grown much beyond oligomeric size. In the absence of unimolecular termination a population of polymer chains having the same molecular mass can grow. The concentration of ionic reactive centers is usually much larger than that of radical centers. Ionic polymerization kinetics are not as well understood as those radical-based polymerizations because of the requirements of extreme purity for the components of the reacting mixture.

Elastomers such as butyl rubber and polyisoprene; high-density polyethylene; polypropylene and its copolymers are widely produced by ionic polymerization. Polypropylene of high molecular mass cannot be produced by radical polymerization. In addition to carbon – carbon double bonds, carbonyl double bonds [81], [82], alkynes [156], and carbon – nitrogen double bonds [157] can be polymerized via ionic mechanisms. However, the discussion here is limited to carbon – carbon double bonds.

2.2.2.1. Cationic Polymerization

Cationic polymerization proceeds through attack on the monomer by an electrophilic species, resulting in heterolytic splitting of the double bond to produce a carbenium ion [158], [159].

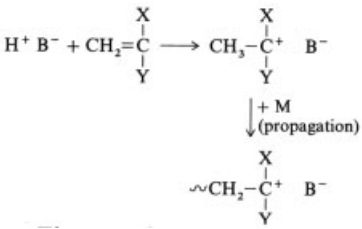

(2.77)

The most important commercial high polymers produced by cationic polymerization are polyisobutylenes and butyl rubber (a copolymer of isobutylene and a 1,3-diene, usually isoprene). A typical polymerization is carried out at about $-100\,°C$ in chlorinated solvents such as chloromethane and is initiated by $AlCl_3$. The polymerization is very fast.

In the following section, interesting features of cationic polymerization are discussed, but no attempt is made to develop quantitative rate expressions. Carbenium ions are very sensitive to polar impurities and this often precludes the establishment of a stationary state. It is also difficult to establish what proportion of the initiator produces growing polymer chains.

Cationic polymerization of vinyl monomers is essentially limited to those with electron-donating substituents such as 1,1-dialkyl, alkoxy, phenyl or vinyl groups. Cationic polymerization involves initiation, propagation, termination, and chain transfer to small molecules much as in free-radical polymerization.

To initiate cationic polymerization, protonic acids such as H_2SO_4, $HClO_4$, and H_3PO_4 or Lewis acids such as BF_3, $AlCl_3$, $TiCl_4$, and $SnCl_4$ are used. Lewis acids are by far the most important initiators for industrial cationic polymerizations. Initiation with Lewis acids requires the presence of a trace of proton donor such as water, alcohol, and organic acid or a cation donor such as alkyl halide.

$$AlCl_3 + (CH_3)_3CCl \rightleftharpoons (CH_3)_3C^+(AlCl_4)^- \tag{2.78}$$

$$(CH_3)_3C^+(AlCl_4)^- + CH_2{=}CH(X) \longrightarrow (CH_3)_3CCH_2{-}C^+(X)H(AlCl_4)^- \tag{2.79}$$

In general,

$$I + ZY \overset{K}{\rightleftharpoons} Y^+(IZ)^- \tag{2.80}$$

$$Y^+(IZ)^- + M \xrightarrow{K_i} YM^+(IZ)^- \tag{2.81}$$

Most polymerizations exhibit a maximum rate at some ratio of initiator and coinitiator concentrations. This optimum ratio varies widely from one initiator system to another and in some instances the solvent has an effect on the ratio. (Note that it has also been proposed that the protogen or catinogen be referred to as the initiator, and the Lewis acid as coinitiator [159], although conventional terminology is exactly the reverse.)

The propagation reaction involves the successive insertion of monomers into the partial bond between the propagating species and its counterion.

$$YM_n^+(IZ)^- + M \xrightarrow{K_p} YM_{n+1}^+(IZ)^- \tag{2.82}$$

The apparent propagation rate constant K_p depends on the type of initiator and solvent used. In general, K_p becomes larger as the acidity of the initiator and/or the dielectric constant of the solvent increases. This is due to the fact that the reactivity of free ions in monomer addition is much greater than that of ion pairs.

A wide variety of reactions may lead to termination of chain growth in cationic polymerization, but it is usually ambiguous and difficult to distinguish termination reactions from chain-transfer reactions. Chain transfer to monomer is most often responsible for the formation of dead polymer:

$$YM_n^+(IZ)^- + M \xrightarrow{K_{tm}} M_n + YM^+(IZ)^- \qquad (2.83)$$

If the counterion is sufficiently nucleophilic, termination by combination may occur:

$$YM_n^+(IZ)^- \xrightarrow{K_t} YM_nIZ \qquad (2.84)$$

However, this reaction is not very common. In fact, true termination in which the activity of a chain carrier is lost without regeneration of an active center is very rare if no impurities capable of destroying an active center are present. However, since the carbenium ion is highly reactive, it would be a formidable task to remove all these impurities from the reaction components. Under commercial conditions it is likely that active centers are consumed to a significant extent by impurities. Termination reactions in cationic polymerizations are unimolecular, as they are in anionic and in anionic-coordination polymerization. Generally, for cationic systems the polymerization rate is given by

$$R_p = K\,[M]^{1-3}\,[I] \qquad (2.85)$$

The first-order rate with respect to initiator concentration is a consequence of the fact that termination reactions are unimolecular. The overall activation energy for polymerization E_R may be ca. -40 to 60 kJ/mol [160]. For many polymerization systems E_R is negative and the rather unusual phenomenon of increasing polymerization rate with decreasing temperature is observed. The activation energy for the degree of polymerization is always negative, and therefore the average chain length decreases with increasing temperature. Usually, cationic polymerizations can produce polymer with sufficient molecular mass only at very low temperatures. Figure 13 shows the Arrhenius plot for the number-average chain length [161]. There is a change in the slope of this plot around $-100\,°C$. This has been attributed to a change in the chain termination step from chain transfer to monomer below $-100\,°C$ to chain transfer to solvent above $-100\,°C$.

Table 1. Effect of counterion on the propagation rate constants (K_p^- and $K_p^{\pm}$) in the anionic polymerization of styrene in THF at 25 °C

Counterion	$K_p^{\pm}$	K_p^-	$K \times 10^7$
Li^+	160	6.5×10^4	2.2
Na^+	80	6.5×10^4	1.5
K^+	60–80	6.5×10^4	0.8
Rb^+	50–80	6.5×10^4	0.1
Cs^+	22	6.5×10^4	0.02

2.2.2.2. Anionic Polymerization

Anionic polymerizations show many of the same characteristics as cationic polymerizations. However, since the nature of carbanions is different from carbenium ions, there are distinct differences. In contrast to cationic polymerization, neither termination nor chain transfer occur in many anionic polymerizations (living polymerization) especially when polar substances are absent. Anionic active centers are usually much more stable than cationic active centers. Although anionic polymerizations proceed rapidly at low temperatures, they are not usually as temperature sensitive as cationic polymerizations, and polymerizations usually proceed well at ambient temperature and higher.

A variety of basic initiators have been used to initiate anionic polymerizations [162], [163]. The initiation involves the addition of an anion (base) to the double bond of the monomer.

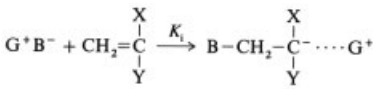

(2.86)

where the $C^- \cdots G^+$ bond can have character ranging from partially covalent to completely ionic. Alkyllithium initiators have been most widely used in the polymerization of butadiene and isoprene, since they are easy to prepare and are soluble in hydrocarbon solvents. A complication which arises when alkyllithiums are used

in nonpolar solvents such as benzene, toluene, cyclohexane, and *n*-hexane is association of various organolithium species. This phenomenon is important, since the associated species are essentially unreactive in propagation [162], [164].

Propagation occurs by the successive insertion of monomers into the partial bond between the propagating anion and its cationic counterion.

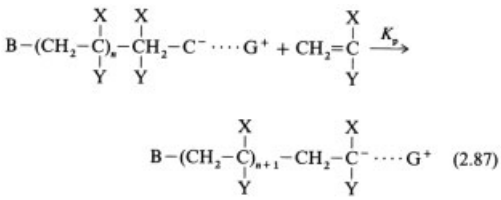

(2.87)

The polymerization rate is fundamentally expressed by $R_p = K_p\,[M]\,[M^-]$, where $[M^-]$ is the concentration of active species. When nonpolar solvents such as dioxane are used, all active species may be ion pairs. However, in polar solvents such as THF, the effect of dissociation to free ions on polymerization rate cannot be neglected. Therefore, the propagation rate is given by the sum of the rates for free propagating anions (R^-) and for ion pairs, $R^-(G^+)$.

$$R_p = K_p^- \,[R^-]\,[M] + K_p^{\pm}\,[R^-\,(G^+)]\,[M] \tag{2.88}$$

The two propagating species are in equilibrium,

$$R^-(G^+) \overset{K}{\rightleftharpoons} R^- + G^+ \tag{2.89}$$

where K is the dissociation constant and is given by

$$K = [R^-]\,[G^+]\,/\,[R^-\,(G^+)] \tag{2.90}$$

Some experimental values for $K_p^{\pm}$ and K_p^- are listed in Table 1 [165], [166]. Consider the simple example of the anionic polymerization of styrene in THF at 25 °C, initiated by sodium naphthalide with an initial concentration $[I] = 1 \times 10^{-3}$ mol/L. Assuming the initiation reaction is instantaneous, $[R^-] = 1.22 \times 10^{-5}$ and $[R^-\,(G^+)] = 9.88 \times 10^{-4}$ mol/L. Therefore,

$$\frac{K_p^-\,[R^-]\,[M]}{K_p^{\pm}\,[R^-\,(G^+)]\,[M]} = \frac{(6.5 \times 10^4)\,(1.22 \times 10^{-5})}{(80)\,(9.88 \times 10^{-4})} = 10.0 \tag{2.91}$$

As shown here, although the concentration of free ions is only 1.2 % of the total propagating species, approximately 90 % of the monomer is consumed by the free ion.

Since carbanions are relatively stable, anionic polymerization with carefully purified reagents may lead to systems in which chain termination is absent. Such polymers are referred to as living polymers [167], [168].

Let us now develop an expression for the chain length distribution for living polymer. Assuming that initiation is instantaneous, the total number of growing chains N_I is constant throughout the polymerization. Now a monomer molecule adds to a polymer chain as shown in Figure 14. The probability that it adds to polymer chain (1) is $1/N_I$, and that it does not add to polymer chain (1) is $1 - 1/N_I$. Therefore, when N_M monomer units have been consumed and bound into polymer chains the probability that a randomly selected polymer chain possesses r monomer units is given by a binomial distribution:

$$N\,(r) = \left(\frac{N_M}{r}\right)\left\{\frac{1}{N_I}\right\}^r \left\{1 - \frac{1}{N_I}\right\}^{N_M - r} \tag{2.92}$$

$N\,(r)$ is also the number chain length distribution. Since $1/N_I \ll 1$ and r is large, the binomial distribution reduces to the Poisson distribution:

$$N\,(r) = \frac{e^{-\eta}\eta^r}{r!} \tag{2.93}$$

where $\eta = N_M/N_I$, which is equal to the number-average chain length, $\bar{P}_N$.

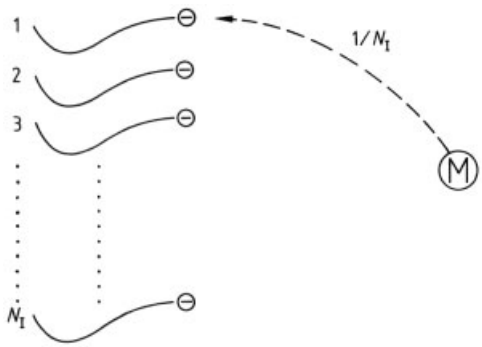

Figure 14. Schematic drawing for the derivation of chain length distribution in living polymerization

The weight chain length distribution $W\,(r)$ and the weight-average chain length $\bar{P}_W$ are given by

$$W(r) = \frac{rN(r)}{\sum_{r=1}^{\infty} rN(r)} = \frac{e^{-\eta}\eta^{r-1}}{(r-1)!} \tag{2.94}$$

$$\bar{P}_W = \sum_{r=1}^{\infty} rW(r) = \eta + 1 \tag{2.95}$$

Therefore, the polydispersity index $\overline{PDI}$ is given by

$$\overline{PDI} = \bar{P}_W / \bar{P}_N = 1 + 1/\bar{P}_N \tag{2.96}$$

With sufficiently large average chain length, the polydispersity should approach unity. To obtain a narrow molecular mass distribution (MWD), the rate of initiation should be greater than the rate of propagation. Polystyrenes having polydispersity as low as 1.01 have been synthesized by using sodium naphthalide as initiator, and these are widely used in the calibration of gel permeation chromatography. It is of interest to note that narrow MWD can be obtained even when there are two reactive site types (free ion and ion pair). This is a result of the equilibrium between the site types, and growing chains will have both site types attached to their ends for the same fraction of time during their growth.

Sequential addition of monomers to a living anionic polymerization system is at present the most useful method for synthesizing well-defined block copolymers.

2.2.2.3. Ziegler – Natta Polymerization

Due to the interactions between the propagating chain end, counterion, and incoming monomer molecule, ionic polymerizations tend to give stereospecific polymers. Polar monomers such as methacrylates and vinyl ethers undergo stereospecific polymerization initiated by conventional ionic initiators under certain conditions. However, the coordinating power of the Ziegler – Natta initiators is much stronger than the usual ionic initiators, so it is appropriate to treat Ziegler – Natta (ZN) systems separately. The use of ZN initiators for diene polymerization has yielded remarkable results that far surpass the stereospecificity exhibited by organolithium initiators. With butadiene, four different stereospecific polymer structures, namely, *cis*-1,4, *trans*-1,4, syndiotactic-1,2, and isotactic-1,2 can each separately be obtained to almost total exclusion of the others by an appropriate choice of initiator system, as shown in Table 2 [169].

The ZN initiators are the only ones which can be used to polymerize α-olefins such as propene and 1-butene. Although the phenomenon of stereoisomerism is not applicable to the symmetrical ethylene, the use of ZN initiators produces a polyethylene with much less long-chain branching than that obtained by free-radical polymerization.

The ZN initiators consist of a combination of alkyls or hydrides of group 1 – 3 metals with salts of group 4 – 8 metals. Some of the components of ZN initiators are as follows:

Group 1 – 3 metal	Transition metal
$(C_2H_5)_3Al$, $(C_2H_5)_2AlCl$	$TiCl_4$, $TiCl_3$, VCl_3
$(C_2H_5)AlCl_2$, $(C_2H_5)_2Be$	$VOCl_3$, $Ti(OC_4H_9)_4$
$C_5H_{11}Na$	$CrCl_3$, $MnCl_3$

The titanium – aluminum initiator systems, especially $(C_2H_5)_3Al - TiCl_3$, have been most thoroughly studied. Many important ZN initiators are solids and during polymerization are suspended in liquid or gaseous media. Heterogeneous initiator systems appear to be necessary for the production of isotactic polyolefins, although soluble initiators are used for the synthesis of syndiotactic polypropylene.

Discussions on the mechanism of ZN polymerization can be found elsewhere [169–173]. At present, however, none of the proposed mechanisms have been comprehensively verified and the observed kinetics are usually quite complex. Some examples of the initial stages of heterogeneous ZN polymerization are shown in Figure 15 [174]. The particles of the transition metal component usually consist of aggregates of smaller crystals, so that when the particle size of the transition metal component is relatively larger (curves c and d), the mechanical pressure of the growing polymer chains cleaves the particles. This increases both the surface area of the initiator and the number of active sites. After this initial period, a steady-state rate may be observed. The time required to reach this steady-state can be reduced by initially using smaller particles. When the transition metal particles are ground or crushed to a very small size just before the polymerization, the rate of polymerization increases rapidly and exceeds the steady-state values (curves a and b). This phenomenon may in-

Table 2. Stereospecific polymerization of butadiene [169]

Ziegler-type catalyst		Conditions	Structure of polymer*, %		
Metal alkyl	Transition metal salt		1,2	*cis*-1,4	*trans*-1,4
$AlEt_2Cl$ and/or $AlEtCl_2$	cobalt salts or solubilized cobalt compounds	Al/Co ≥ 100		95–98	
$AlEt_3$	VCl_4 or $VOCl_3$	Al/V = 2			95–100
$AlEt_3$	TiI_4	Al/Ti = 5	5	92	4
$AlEt_3$	chromium acetylacetonate	Al/Cr = 10	isotactic		
$AlEt_3$	chromium acetylacetonate	Al/Cr = 3	syndiotactic		
$AlEtCl_2$	β-$TiCl_3$	Al/Ti = 1–5			100
$AlEt_3$	$Ti(OBu)_4$	Al/Ti = 3	90	10	
$AlEt_3$	$TiCl_4$	Al/Ti = 1	2	49	49

* 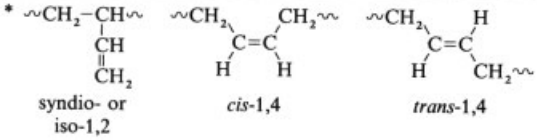

dicate the existence of highly active but short-lived sites. This kind of behavior can be avoided by aging the initiator system prior to addition of monomer.

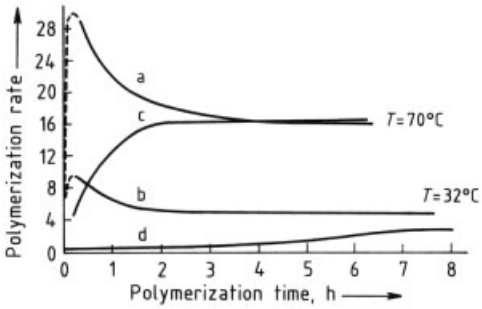

Figure 15. Effect of previous physical treatment on a sample of α–$TiCl_3$ on the propene polymerization rate (in grams C_3H_6 per gram $TiCl_3$ per hour) at constant pressure and temperature
a), b) Ground α–$TiCl_3$ (sizes $\leq 2\,\mu m$); c), d) Unground α–$TiCl_3$ (sizes $1-10\,\mu m$)

Many ZN polymerizations exhibit a continuous decrease in rate rather than reaching a steady-state in which the rate curve reaches a maximum and then continues to decline. The rate of decline varies with the polymerization system used. The decline in polymerization rate has been attributed to (1) a decrease in the number of active centers, (2) a lowering of activity of individual active centers due to structural changes, and (3) a lowering of activity of individual active centers due to encapsulation by polymer.

In this section a very simple example is considered, too approximate to accurately describe real systems, but useful to illustrate a model building process. Consider the following polymerization scheme [175].

1) Adsorption of monomer (M) on the active site

$$\mathrm{M} + \mathrm{A} \xrightarrow{K_A} \mathrm{P}_1^* \tag{2.97}$$

where P_1^* is a growing chain with unit chain length and A is a vacant active site.

2) Propagation

$$\mathrm{P}_r^* + \mathrm{M} \xrightarrow{K_P} \mathrm{P}_{r+1}^* \tag{2.98}$$

where P_r^* is the growing polymer chain with length r

3) Desorption of polymer chain from the active site

$$\mathrm{P}_r^* \xrightarrow{K_D} \mathrm{P}_r + \mathrm{A} \tag{2.99}$$

where P_r is a dead polymer molecule with chain length r.

Assuming that the polymer chain is sufficiently long (long-chain assumption), the polymerization rate R_p is given by

$$R_p = K_p\,[\mathrm{P}^*]\,[\mathrm{M}] \tag{2.100}$$

where [P∗] is the total concentration of active sites, namely, $[P^*] = \sum_{r=1}^{\infty} [\mathrm{P}_r^*]$.

If the stationary-state hypothesis is applied to [P∗],

$$K_A\,[M]\,[A] = K_D\,[P^*] \tag{2.101}$$

Each active site is either occupied by growing polymer chains or is vacant. Denoting the total concentration of adsorption sites by $[A]_0$,

$$[A]_0 = [A] + [P^*] \tag{2.102}$$

Combining Equations (2.101) and (2.102) gives

$$[P^*] = K\,[A]_0\,[M] / (1 + K\,[M]) \tag{2.103}$$

where $K = K_A/K_D$. Therefore, the polymerization rate is given by

$$R_p = K_p K\,[A]_0\,[M] / (1 + K\,[M]) \tag{2.104}$$

Next consider the instantaneous molecular mass distribution for this process. The population balance equations for growing polymer chains are

$$\frac{1}{V}\frac{d\left(V\,[P_1^*]\right)}{dt} = K_A\,[M]\,[A] - K_p\,[M]\,[P_1^*] - K_D\,[P_1^*] \tag{2.105}$$

$$\frac{1}{V}\frac{d\,(V\,[P_r^*])}{dt} = K_p\,[M]\left[P_{r-1}^*\right] - K_p\,[M]\,[P_r^*] - K_D\,[P_r^*] \quad (r \geq 2) \tag{2.106}$$

By application of the SSH for $[P_r^*]$,

$$[P_1^*] = \frac{1}{1+\tau'}\,[A] \tag{2.107}$$

$$[P_r^*] = \frac{1}{1+\tau'}\left[P_{r-1}^*\right] \tag{2.108}$$

where $\tau' = K_D/(K_p\,[M])$. Therefore,

$$[P_r] = K_D\,[P_r^*] = K_D\,[A]\left(\Phi'\right)^r \tag{2.109}$$

where $\Phi' = 1/(1 + \tau')$.

The number and weight chain length distributions are given by

$$N(r) = \frac{[P_r]}{\sum_{r=1}^{\infty}[P_r]} = \left(1-\Phi'\right)\left(\Phi'\right)^{r-1} \tag{2.110}$$

$$W(r) = \frac{r\,[P_r]}{\sum_{r=1}^{\infty} r\,[P_r]} = r\left(1-\Phi'\right)^2\left(\Phi'\right)^{r-1} \tag{2.111}$$

where Φ' is the probability that a growing polymer chain adds another monomer molecule, and $(1 - \Phi')$ is the probability that a growing chain stops growing.

Therefore, Equations (2.110) and (2.111) are essentially the same as Equations (2.8) and (2.9) or Equation (2.56). The polydispersity index (PDI) for instantaneously formed polymer is two.

An interesting feature of heterogeneous ZN initiators is the very broad MWD of the polymers produced, although the above simple model predicts PDI = 2 for instantaneously formed polymer. The PDI of accumulated polymer may be 10 or higher. In the case of copolymerization, the compositional heterogeneity is also large even when monomer ratios are kept constant during polymerization. Presently, there are two main theories which try to explain these large dispersities, namely, the presence of a distribution of activities for the active sites or diffusional effects which limit the transport of reactants to the active sites [176].

2.3. Copolymerization

Copolymerization permits the synthesis of an almost unlimited number of polymer types and is therefore often used to obtain a better balance of properties for commercial applications. Copolymers may be synthesized by chain-growth and step-growth polymerization. In step-growth polymerization, different monomers with the same type of functional group generally show only minor differences in reactivity. As a result, most copolymers prepared by step-growth polymerization contain essentially random placements of repeat units, with the composition of the copolymers essentially the same as those of the original monomer mixture.

In contrast, strong selective effects often occur in chain-growth copolymerizations, and the composition of the copolymer formed may differ greatly from the composition of the original monomer mixture. This section deals exclusively with chain-growth copolymerization. Chain-growth copolymerization can be carried out with various types of active centers including free-radical, cationic, and anionic species. Free-radical copolymerization is most commonly used due to its higher alternating tendencies.

2.3.1. Copolymer Composition

The composition of copolymers cannot be determined from a knowledge of the homopolymerization rates of each monomer. In 1944, the instantaneous copolymer composition equation was proposed independently by several researchers by assuming that the chemical activity of a propagating chain depends solely on the terminal monomer unit on which the active center is located [177–180]. This model is called the terminal model, and the copolymer chain can be considered as a first-order Markov chain. For binary systems, the following four propagation reactions are possible.

$$\mathrm{P}^*_{m,n,1} + \mathrm{M}_1 \xrightarrow{K_{11}} \mathrm{P}^*_{m+1,n,1} \tag{2.112}$$

$$\mathrm{P}^*_{m,n,1} + \mathrm{M}_2 \xrightarrow{K_{12}} \mathrm{P}^*_{m,n+1,2} \tag{2.113}$$

$$\mathrm{P}^*_{m,n,2} + \mathrm{M}_1 \xrightarrow{K_{21}} \mathrm{P}^*_{m+1,n,1} \tag{2.114}$$

$$\mathrm{P}^*_{m,n,2} + \mathrm{M}_2 \xrightarrow{K_{22}} \mathrm{P}^*_{m,n+1,2} \tag{2.115}$$

where $\mathrm{P}^*_{m,n,1}$ is a live copolymer chain with m units of monomer 1 (M_1) and n units of monomer 2 (M_2) bound in the polymer chain and with the active center located on terminal monomer unit 1.

The reactivity of the propagating species may be affected by the penultimate monomer unit. In such cases, the model is referred to as the penultimate model or a second-order Markov chain [181], [182] and propagation consists of eight reactions. Further expansion is possible by considering the effects of remote units preceding the penultimate unit, such as the pen-penultimate model [183–185].

It is customarily assumed that the propagation rate constants are independent of chain length and that the chains are sufficiently large (long-chain assumption, LCA). The LCA includes the approximation that monomer consumed in reactions other than propagation is negligible and that the SSH is valid for each type of active center; i.e., the rate of formation of any type of active center is equal to its rate of consumption. Statistically, these conditions are equivalent to the statistical stationary condition [185–187].

Now consider a binary copolymerization whose propagation reactions follow the terminal model. The specific polymerization rates of monomers 1 and 2 are given by

$$-\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{M}_1])}{\mathrm{d}t} = K_{11}\,[\mathrm{P}^*_1]\,[\mathrm{M}_1] + K_{21}\,[\mathrm{P}^*_2]\,[\mathrm{M}_1] \tag{2.116}$$

$$-\frac{1}{V}\frac{\mathrm{d}\,(V\,[\mathrm{M}_2])}{\mathrm{d}t} = K_{12}\,[\mathrm{P}^*_1]\,[\mathrm{M}_2] + K_{22}\,[\mathrm{P}^*_2]\,[\mathrm{M}_2] \tag{2.117}$$

where $[\mathrm{P}^*_1]$ and $[\mathrm{P}^*_2]$ are the total concentrations of active centers of types 1 and 2 with

$$[\mathrm{P}^*_1] = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}[\mathrm{P}^*_{m,n,1}] \tag{2.118}$$

$$[\mathrm{P}^*_2] = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty}[\mathrm{P}^*_{m,n,2}] \tag{2.119}$$

Application of the SSH gives

$$K_{21}\,[\mathrm{P}^*_2]\,[\mathrm{M}_1] = K_{12}\,[\mathrm{P}^*_1]\,[\mathrm{M}_2] \tag{2.120}$$

Dividing Equation (2.116) by Equation (2.117) and using Equation (2.120) gives the instantaneous copolymer composition equation:

$$\frac{F_1}{F_2} = \frac{-\mathrm{d}\,(V\,[\mathrm{M}_1])}{-\mathrm{d}\,(V\,[\mathrm{M}_2])} = \frac{[\mathrm{M}_1]\,(r_1\,[\mathrm{M}_1] + [\mathrm{M}_2])}{[\mathrm{M}_2]\,([\mathrm{M}_1] + r_2\,[\mathrm{M}_2])} \tag{2.121a}$$

or

$$F_1 = \frac{r_1 f_1^2 + f_1 f_2}{r_1 f_1^2 + 2 f_1 f_2 + r_2 f_2^2} \tag{2.121b}$$

where F_1 and F_2 are the mole fractions of monomers 1 and 2 in the copolymer produced instantaneously, and f_1 and f_2 are the mole fractions of unreacted monomer. The reactivity ratios are defined by

$$r_1 = \frac{K_{11}}{K_{12}} \tag{2.122}$$

$$r_2 = \frac{K_{22}}{K_{21}} \tag{2.123}$$

It is straightforward to extend this method to multicomponent polymerization [188–190]. Note that Equation (2.121 a or b) is strictly valid only for infinitely long polymer chains and its application to short chains may introduce significant error [191–193].

When applying Equation (2.121 a or b), it is usually assumed that all of the copolymer chains

have the same composition. However, since the chain length of a copolymer is finite, the composition and chain lengths of the individual polymer molecules cannot all be identical. Therefore, for copolymer chains produced instantaneously there is a bivariate distribution of composition and chain length [194–200]. The variance of the copolymer composition distribution is approximately inversely proportional to the chain length and, therefore, for sufficiently long chains it may be reasonable to neglect the composition distribution. In order to treat oligomeric molecules, application of discrete mathematics such as the finite Markov chain theory [184], [201] is necessary.

In batch copolymerization, there is a compositional drift due to the change in the composition of the unreacted monomer mixture with time. The total monomer conversion x is related to the mole fraction f_1 of unreacted monomer 1 by the following equation [202]:

$$\frac{\mathrm{d}f_1}{\mathrm{d}x} = \frac{f_1 - F_1}{1 - x} \tag{2.124}$$

$f_1 = f_{10}, F_1 = F_{10}$ at $x = 0$.

Therefore, if the relationship between f_1 and F_1 is known, Equation (2.124) can be solved numerically. If the terminal model is applicable to a binary system (Eq. 2.121 a or b), Equation (2.124) can be integrated analytically to obtain the following equation [203–208]:

$$x = 1 - \left(\frac{f_1}{f_{10}}\right)^{\alpha} \left(\frac{f_2}{f_{20}}\right)^{\beta} \left(\frac{f_{10} - \delta}{f_1 - \delta}\right)^{\gamma} \tag{2.125}$$

where
$\alpha = r_2/(1 - r_2)$,
$\beta = r_1/(1 - r_1)$,
$\gamma = (1 - r_1 r_2)/\{(1 - r_1)(1 - r_2)\}$
$\delta = (1 - r_2)/(2 - r_1 - r_2)$.

Except for copolymerizations at the azeotropic point, there will be some compositional drift. (At the azeotropic point, the composition of the instantaneously produced copolymer and of the unreacted monomers is the same; hence in a batch reactor there is no compositional drift at this point.) An example of compositional drift during batch polymerization is shown in Figure 16 [208]. One method to avoid compositional drift is to use a continuous stirred-tank reactor with an ideal residence time distribution. In this instance, the STOCKMAYER instantaneous bivariate distribution of chain length and composition will be obtained. Another method of composition control is to use semi-batch operation in which monomers are fed to maintain a constant ratio of monomer concentrations in the reactor [209], [210].

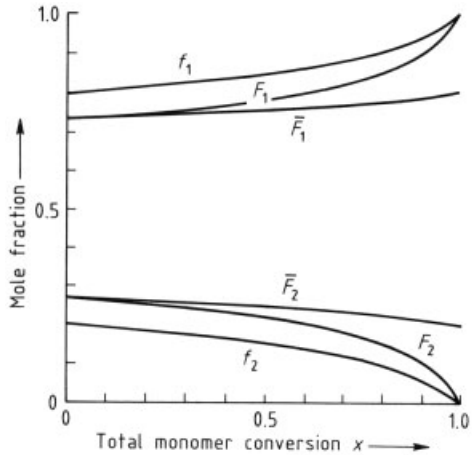

Figure 16. Batch copolymerization of styrene (M_1) and methyl methacrylate (M_2) at 60 °C for $f_{10} = 0.8$, $f_{20} = 0.2$, $r_1 = 0.53$, $r_2 = 0.56$
$\bar{F}_1$ and $\bar{F}_2$ are the accumulated mole fraction of monomer 1 and 2 bound in the copolymers.

In the terminal model, monomer sequence-length distribution depends only on the product of reactivity ratios $r_1 r_2$ [211], [212]. In free-radical copolymerizations $r_1 r_2$ is generally less than unity, indicating a higher alternating tendency. The styrene – maleic anhydride system is an example of a very highly alternating one with both r_1 and r_2 very close to zero. For ionic copolymerizations, there is a general lack of any tendency towards alternation, and $r_1 r_2$ is usually close to or greater than unity. Generally, in free-radical copolymerization, the reactivity ratios are relatively insensitive to the reaction medium and temperature. However, in ionic copolymerization, temperature and reaction medium can significantly affect reactivity ratios.

Traditional methods for estimating reactivity ratios [177], [213–215] are based on the transformation of the instantaneous copolymer composition equation into a form that is linear in the parameters r_1 and r_2. While these linearizations provide simple techniques for parameter estimation, they are generally statistically invalid be-

cause the independent variable has experimental error and the dependent variable does not have a constant variance [216–223]. Both of the latter assumptions are necessary for the linear least squares method to be a statistically valid estimation method. Although it has been shown that proper experimental design can allow the use of linear least squares analysis [222], [224], the estimation of the reactivity ratios is a typical example of a problem in nonlinear estimation. Reactivity ratios are now generally estimated by application of procedures based on the statistically valid error-in-variables-model (EVM) [220], [221], [225–230]. These methods allow all the sources of experimental error to be accounted for.

In spite of the fact that reactivity ratios obtained using invalid estimation procedures have often been used, the copolymer composition equation based on the terminal model has given useful predictions. It is generally ineffective to use the copolymer composition equation for comparing models. Information on the copolymer microstructure (e.g., monomer sequence length distribution) is needed to compare the validity of terminal and penultimate models [179], [231], [232]. Advancements in NMR techniques have provided suitable information on sequence length distribution. However, cases have been reported in which sequence length distribution information does not permit model discrimination [233].

As will be discussed in Section 2.3.2 the composition equation includes the rate constants of elementary reactions only in ratios, while the rate equation is dependent on their absolute values, and therefore success of the former equation does not insure success of the latter [234].

2.3.2. Kinetics of Copolymerization

Consider the kinetics of free-radical copolymerization of monomers M_1 and M_2, assuming the terminal model is applicable. Important elementary reactions are:

Initiation

$$\mathrm{I} \xrightarrow{K_d} 2\,\mathrm{R}^{\bullet}_{in}$$
$$\mathrm{R}^{\bullet}_{in} + \mathrm{M}_1 \xrightarrow{K_1} \mathrm{R}^{\bullet}_{1,0,1}$$
$$\mathrm{R}^{\bullet}_{in} + \mathrm{M}_2 \xrightarrow{K_2} \mathrm{R}^{\bullet}_{0,1,2}$$

Propagation

$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{M}_1 \xrightarrow{K_{11}} \mathrm{R}^{\bullet}_{m+1,n,1}$$
$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{M}_2 \xrightarrow{K_{12}} \mathrm{R}^{\bullet}_{m,n+1,2}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{M}_1 \xrightarrow{K_{21}} \mathrm{R}^{\bullet}_{m+1,n,1}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{M}_2 \xrightarrow{K_{22}} \mathrm{R}^{\bullet}_{m,n+1,2}$$

Transfer to monomer

$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{M}_1 \xrightarrow{K_{f11}} \mathrm{P}_{m,n} + \mathrm{R}^{\bullet}_{1,0,1}$$
$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{M}_2 \xrightarrow{K_{f12}} \mathrm{P}_{m,n} + \mathrm{R}^{\bullet}_{0,1,2}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{M}_1 \xrightarrow{K_{f21}} \mathrm{P}_{m,n} + \mathrm{R}^{\bullet}_{1,0,1}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{M}_2 \xrightarrow{K_{f22}} \mathrm{P}_{m,n} + \mathrm{R}^{\bullet}_{0,1,2}$$

Transfer to small molecule

$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{T} \xrightarrow{K_{fT1}} \mathrm{P}_{m,n} + \mathrm{T}^{\bullet}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{T} \xrightarrow{K_{fT2}} \mathrm{P}_{m,n} + \mathrm{T}^{\bullet}$$

Termination by disproportionation

$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{R}^{\bullet}_{r,s,1} \xrightarrow{K_{td11}} \mathrm{P}_{m,n} + \mathrm{P}_{r,s}$$
$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{R}^{\bullet}_{r,s,2} \xrightarrow{K_{td12}} \mathrm{P}_{m,n} + \mathrm{P}_{r,s}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{R}^{\bullet}_{r,s,2} \xrightarrow{K_{td22}} \mathrm{P}_{m,n} + \mathrm{P}_{r,s}$$

Termination by combination

$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{R}^{\bullet}_{r,s,1} \xrightarrow{K_{tc11}} \mathrm{P}_{m+r,n+s}$$
$$\mathrm{R}^{\bullet}_{m,n,1} + \mathrm{R}^{\bullet}_{r,s,2} \xrightarrow{K_{tc12}} \mathrm{P}_{m+r,n+s}$$
$$\mathrm{R}^{\bullet}_{m,n,2} + \mathrm{R}^{\bullet}_{r,s,2} \xrightarrow{K_{tc22}} \mathrm{P}_{m+r,n+s}$$

where $\mathrm{R}^{\bullet}_{m,n,1}$ is a polymer radical with m units of monomer 1 (M_1) and n units of monomer 2 (M_2) bound in the polymer chain with active center located on monomer unit 1. $P_{m,n}$ is a polymer molecule with m units of monomer 1 and n units of monomer 2. The polymerization rate, R_p is given by

$$\begin{aligned} R_p &= K_{11}[\mathrm{R}^{\bullet}_1][\mathrm{M}_1] + K_{12}[\mathrm{R}^{\bullet}_1][\mathrm{M}_2] \\ &\quad + K_{21}[\mathrm{R}^{\bullet}_2][\mathrm{M}_1] + K_{22}[\mathrm{R}^{\bullet}_2][\mathrm{M}_2] \\ &= K_p[\mathrm{R}^{\bullet}][\mathrm{M}] \end{aligned} \tag{2.126}$$

where $[\mathrm{R}^{\bullet}_1]$ and $[\mathrm{R}^{\bullet}_2]$ are given by Equations (2.118) and (2.119) and $[\mathrm{R}^{\bullet}] = [\mathrm{R}^{\bullet}_1] + [\mathrm{R}^{\bullet}_2]$, $[\mathrm{M}] = [\mathrm{M}_1] + [\mathrm{M}_2]$, and K_p is defined by

$$K_p = (K_{11}f_1 + K_{12}f_2)\varphi^{\bullet}_1 + (K_{21}f_1 + K_{22}f_2)\varphi^{\bullet}_2 \tag{2.127}$$

where $f_1 = [M_1]/[M]$, and $\varphi_1^{\bullet} = [R_1^{\bullet}]/[R^{\bullet}]$.

Equation (2.127) is an example of a pseudo-kinetic rate constant [145–147], [209], [210], [235–238]. For an N-component system, the pseudo-kinetic rate constants can be defined as follows

Propagation

$$K_p = \sum_{\substack{i=1\\j=1}}^{N} K_{ij}\,\varphi_i^{\bullet} f_j \tag{2.128}$$

Chain transfer to monomer

$$K_{fm} = \sum_{\substack{i=1\\j=1}}^{N} K_{fij}\,\varphi_i^{\bullet} f_j \tag{2.129}$$

Chain transfer to small molecule

$$K_{fT} = \sum_{i=1}^{N} K_{fT_i}\,\varphi_i^{\bullet} \tag{2.130}$$

Termination by disproportionation

$$K_{td} = \sum_{\substack{i=1\\j=1}}^{N} K_{td_{ij}}\,\varphi_i^{\bullet}\,\varphi_j^{\bullet} \tag{2.131}$$

Termination by combination

$$K_{tc} = \sum_{\substack{i=1\\j=1}}^{N} K_{tc_{ij}}\,\varphi_i^{\bullet}\,\varphi_j^{\bullet} \tag{2.132}$$

Even though these pseudo-kinetic rate constants depend on the monomer mole fraction and change with time in a batch reactor, the simplification achieved when dealing with multicomponent polymerizations is very great. Furthermore, since the mole fraction of each polymer radical type $\varphi_i^{\bullet}$ is independent of chain length for sufficiently long chains, these same pseudo-kinetic rate constants can be used to calculate molecular mass distribution as well as polymerization rate [61],[235] . Therefore, by using pseudo-kinetic rate constants, a multicomponent polymerization reduces to a homopolymerization, and therefore Equations (2.55) – (2.64) are all applicable for copolymerization. With appropriate definitions for pseudo-kinetic rate constants, this method may also be usefully applied when copolymerization kinetics follow the penultimate model [235].

In the early development of the kinetics of copolymerization, chemically controlled bimolecular termination was assumed to be operable [239–241]. However, since these reactions have been shown to be diffusion controlled [120–122], [242], Equations (2.131) and (2.132) must be modified by using appropriate models which account for this. In this context, the use of the cross-termination factor, $\varphi = K_{t12}/(K_{11}K_{12})^{0.5}$, is not acceptable. Several models for diffusion-controlled termination of binary copolymerization have been proposed. Atherton and North [243], [244] proposed the following for the termination constant.

$$K_t = K_{t11}F_1 + K_{t22}F_2 \tag{2.133}$$

The mole fractions of monomers bound in polymer chains F_1 and F_2 can be calculated by using the copolymer composition equation. On the other hand, Russo and Munari [245] proposed the use of penultimate effect for the termination reactions since segmental diffusion is highly dependent on the last two portions of the chain.

Until the mid 1980s, a prevailing view of copolymerization kinetics was that the propagation process is, in most cases, correctly described by the terminal model, whereas the termination process involves complexity still to be elucidated. However, it has also been speculated that it is the propagation step that needs further study, while the termination step is well described by simple models such as Equation (2.133), and the penultimate effect on propagation reactions is being examined from the point of view of copolymerization rates [233], [246–252].

2.3.3. Copolymerization of Vinyl and Divinyl Monomers

The use of the terms vinyl and divinyl here is not according to their strict definitions. Here, a vinyl monomer is defined as a monomer with a single reactive double bond (a double bond which will readily add to a radical center) and a divinyl monomer is a monomer which has two such double bonds.

The free-radical copolymerization of vinyl and divinyl monomers is important in the manufacture of ion-exchange resins, chromatographic packings, superabsorbent polymers (with a rapidly growing market in baby diapers and

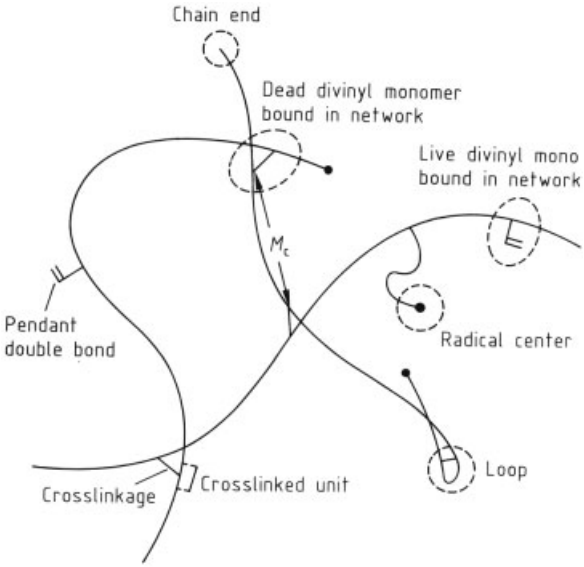

Figure 17. A schematic drawing of a cross-linked polymer network synthesized by free-radical copolymerization of vinyl and divinyl monomer

promise of the same in several other market areas), cross-linked latex polymers, and other products. Figure 17 shows a schematic of a polymer network synthesized by copolymerization of vinyl and divinyl monomers. There may be many radical centers on the polymer network during polymerization since the mobility of radical centers chemically bound to the network can be highly restricted. Strong autoacceleration in polymerization rate has been reported during network formation by free-radical copolymerization [253–260]. The live double bonds located on polymer chains are called pendant double bonds. Kinetic behavior of radical centers on polymer chains and of pendant double bonds are the most important factors influencing the kinetics of network formation during free-radical polymerization. From the point of view of the mobility of chains, network polymers may be considered as a heterogeneous reaction system, and the formation of microgels before the gel point is reached may be a general feature of network formation in free-radical polymerization [257], [261–265]. The number-average molecular mass between cross-links M_c is important from the point of view of the elastic properties of a gel molecule. As the mole fraction of divinyl monomer increases, the heterogeneity of the polymer network (i.e., regions having very different M_c) becomes significant [265–268].

Reactivity ratios for the copolymerization of vinyl and divinyl monomers are rather difficult to determine. If the reactivities of both double bonds on the divinyl monomer are the same and independent, and cyclization does not occur, the conventional copolymer composition equation (Eq. 2.121 a or b) is still valid when monomer concentration is replaced by double bond concentration and the reactivity ratios defined with respect to each type of double bond are used [269]. However, difficulties generally arise due to the complicated behavior of pendant double bonds, which may react inter- and intramolecularly, and whose reactivity may differ from that of monomer double bonds. At the moment when a divinyl monomer is chemically bound in a polymer chain, another double bond on the just-reacted divinyl monomer may be the nearest neighbour of the active center, and therefore cyclopolymerization may occur under certain conditions. For example, the free-radical polymerization of diallyl quaternary ammonium salts gives soluble, not cross-linked polymers with little or no residual unsaturation [270], [271].

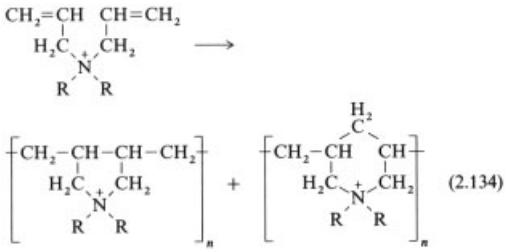

(2.134)

In general, cyclization may also involve two or more monomer units. Since cyclization reactions are controlled by the conformational statistics of the sequence of bonds and do not follow the conventional rate law, larger cycles, formed when an active center adds to a pendant double bond on its own chain or to two double bonds on another chain, may also affect the copolymer composition even when the terminal model is applicable. Various copolymer composition equations which account for cyclization reactions have been proposed [272–276]; however, it may be difficult to apply these equations to higher mole fractions of divinyl monomer or to high monomer conversions. Strictly speaking, in order to know the copolymer composition, it is necessary to know the kinetic behavior of pendant double bonds completely; that is, a knowledge of the reactivity ratios r_1 and r_2 is insufficient to estimate the change in composition during copolymerization. However, r_1 and r_2 are usually obtained without taking into account the monomer consumed by active centers located on just-reacted pendant double bonds. Such parameters may be better regarded as empirical parameters which do not reflect the chemical reactivity of radical centers with double bonds, and they may change with monomer conversion.

As for the kinetics of network formation, fundamental models which assume an equilibrium system (see Section 2.1.3) have been extensively applied to chain-growth polymerization [39], [43], [51], [52], [148], [277–279]. However, chain-growth polymerizations are kinetically controlled, so that the application of the conventional approaches may be in error. It has been shown that under FLORY's simplifying assumptions, namely that (1) the reactivities of all types of double bonds are equal, (2) all double bonds react independently of one another, and (3) there are no intramolecular reactions in finite molecules, the predictions of kinetic and equilibrium models are the same [61], [147], [280]. However, none of the above idealized assumptions are strictly applicable to a real system [265], [281–287]. Kinetic models for network formation by free-radical polymerization are being developed [61], [62], [145–147], [288–295]. Figure 18 shows the change in the average chain length and weight fraction of gel during network formation in the copolymerization of methyl methacrylate and ethylene glycol dimethacrylate [147].

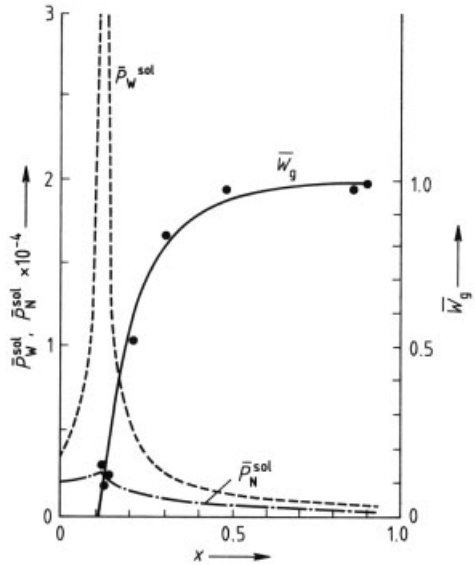

Figure 18. Development of average chain length within sol fraction and weight fraction of gel during batch copolymerization of methyl methacrylate and ethylene glycol dimethacrylate (0.25 mol %) at 70 °C initiated by AIBN [147]
● = Experimental data

3. Polymerization Processes and Reactor Modeling

3.1. Introduction

The polymer reactor model is now becoming accepted as a valuable tool whose use contributes significantly to all aspects of process technology for polymer manufacture. This includes process design, optimization, state estimation, and control. Through process design, polymers with a unique and desirable combination of properties can often be obtained. Process parameters such

as residence time distribution (RTD) are usually not considered by polymer synthesis chemists, although RTD can influence chemical composition distribution (CCD), molecular mass distribution (MWD), long-chain branching (LCD) and gel/sol ratios. In the early days of the polymer industry, the chemist played the major role in product and process development and scale-up. This has changed, with the process engineer now playing a significant role in all phases of commercialization of new and improved polymer products. His broad experience with process fundamentals and computer modeling are essential to obtain high-quality products, safely and economically.

Dynamic reactor models can be used in a variety of ways. Stability and control of polymer reactors should be considered at the design stage and control problems minimized then, rather than take corrective action after the plant is built. Complex interactions which are involved in polymerization (highly nonlinear temperature and concentration effects) preclude optimal design based on experimentation alone because the cost would be prohibitive. Models can be used to identify potential sources of product variability and strategies to minimize their effects. Models can be used to store information on process technology in a concise and readily retrievable and modifiable form.

Process models can be used to train chemists, chemical engineers, and plant operators and give them a feel for the dynamics of the polymerization process.

The most expensive aspect of model development is experimental estimation of model parameters; highly instrumented bench-, pilot-scale, and plant-scale reactors are required. Statistically designed experiments should be performed to permit efficient parameter estimation and model development. Modeling is an iterative process and the very act of developing a deterministic model permits a greater understanding of the relevant microscopic processes which occur during polymerization or polymer modification. As additional data (plant, pilot-plant, and bench-scale) become available, model structure and parameters can be updated.

This section considers recent developments in polymerization/polymer modification processes and discusses advances in polymer reactor modeling, state estimation, and control.

3.2. Processes and Reactor Modeling for Step-Growth Polymerization

3.2.1. Types of Reactors and Reactor Modeling

In step-growth polymerization, high molecular mass polymers are usually not produced until the final stage of reactions, so that thermal control and mixing of the reaction mixture do not present serious problems in the earlier stages. However, since the final stage of polymerization is very important for the production of polymers with high molecular mass, handling of very high viscosities and temperatures and high interfacial area to remove small molecules are required. Various polymerization processes and reactor types, both for batch and continuous production, have been proposed. Examples of reactors for high viscosities are shown in Figures 19 and 20. Careful selection of the polymerization reactor is very important to produce high-quality polymers [296], [297].

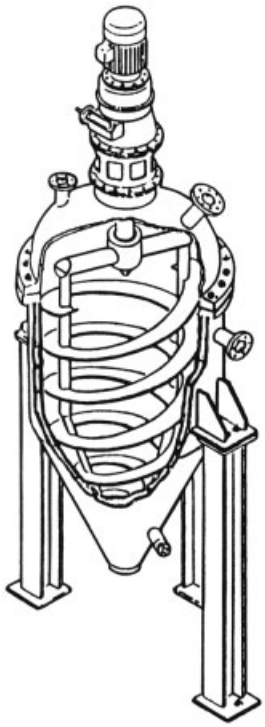

Figure 19. Vertical cone ribbon blade reactor (Mitsubishi Heavy Industries)

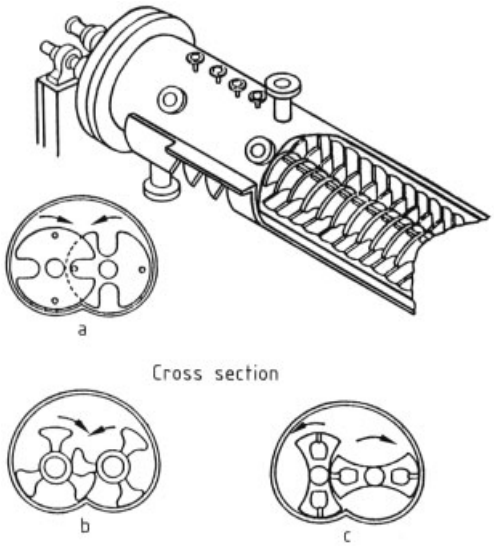

Figure 20. Horizontal high-viscosity reactor

The batch reactor is the most versatile reactor type and is used extensively for specialty polymers at low production volumes. Some examples of step-growth polymerizations carried out in such reactors are nylon 6, phenol – formaldehyde, urea – formaldehyde, and melamine – formaldehyde. In polycondensation reactions, it may be necessary to remove condensation products to attain sufficient conversion. When the volume of the reaction mass decreases continuously with time, such reactors are called semi-batch reactors. (A tank-type reactor which does not operate at steady state is defined as a semi-batch reactor). For example, in the production of poly(ethylene terephthalate), PETP, since methanol or ethylene glycol evaporates during polymerization, the batchwise production of PETP is considered a semi-batch operation.

On the other hand, newer high-capacity plants often use continuous processes. The first approximation for a continuous process is a model that consists of plug flow reactors (PFR) and continuous stirred-tank reactors (CSTR) in various combinations (Fig. 21), although various nonideal effects such as flow pattern in the reactor, mass- and heat-transfer limitations, and residence time distribution must be considered for a detailed analysis and design of real reactors [298–302].

Molecular mass distribution of linear polymers produced by step-growth polymerization in a batch or a PFR basically follows the most probable distribution [303]. (Note that batch reactor and PFR are, in principle, equivalent). The molecular mass distribution may be controlled by varying its reaction path if the reaction system is in a nonequilibrium state. Assuming irreversible step-growth polymerization without interchange reactions, the effect of reactor types, such as homogeneous CSTR, segregated CSTR and PFR with a recycle loop, on molecular mass distribution have been considered [301], [302], [304], [305]. An important feature of step-growth polymerization is that the variance of the molecular mass distribution is smallest in a batch reactor or PFR and is largest in a homogeneous CSTR, which is quite contrary to that for chain-growth polymerization. This result may be disappointing, since it is, in principle, impossible to produce polymers whose polydispersity index M_W/M_N is smaller than two in step-growth polymerization at sufficiently high conversions. The polydispersity index of polymers produced in a batch reactor is given by $1+p$, as shown in Section 2.1.1., where p is the conversion of the functional group. However, in commercial polymeric materials, polymers with narrower distributions are not always superior to those with broader distributions, since various levels of properties are required at the same time. The use of a cascade of CSTRs and/or PFRs with recycle loops may be one method to obtain a molecular mass distribution with a polydispersity index larger than two. However, in practice, these methods may have shortcomings because they need a long start-up period and, therefore, problems may occur with the stability of the reaction system. A method in which additional monomers are fed intermediately to a batch reactor or a PFR has been proposed [306]. Theoretical analysis of this intermediate monomer feed method has also been carried out [307], and it has been shown that the polydispersity index can be easily controlled over a wide range with values greater than 2.

In a batch reactor, the reverse reactions and the interchange reactions (redistribution reactions) do not change the MWD from the most probable distribution [303], [308–310]. However, these reactions do change the MWD of polymers produced in CSTR, PFR with a recycle

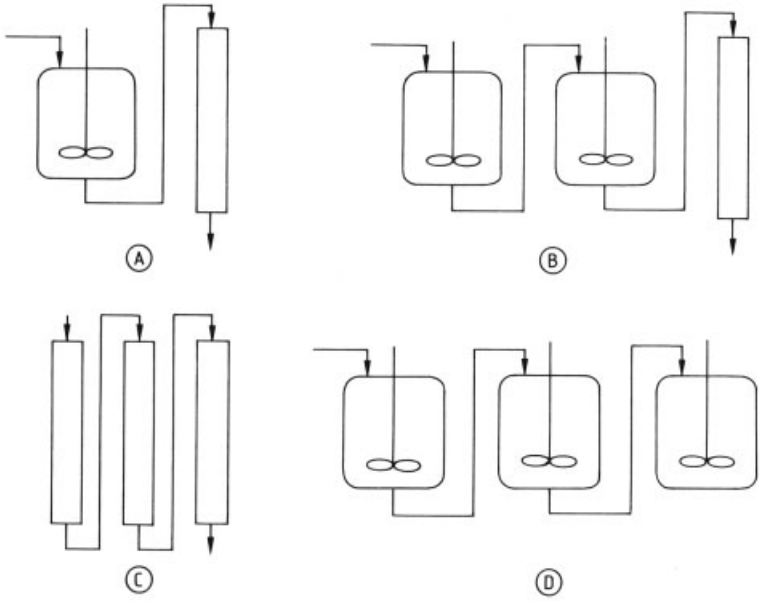

Figure 21. Representative models for polymerization reactors
A) CSTR + PFR; B) CSTR + CSTR + PFR; C) PFRs; D) CSTRs

loop, and intermediate monomer feed method. Some consideration of these reactions in a CSTR is given in [311]. Qualitatively, these effects lower the polydispersity and make the MWD approach the most probable distribution. This result seems reasonable, since any MWD approaches the most probable distribution with a polydispersity index of two when polymer chains are severed randomly [312], [313].

Other than the common reactor types discussed above, other special types of reactor system may be applied. For example, in the polymerization of urethanes, the reaction rates are so high that reaction takes place even when the monomers are being mixed and pumped into molds. In situ polymerization to form the desired articles directly from monomeric liquids is known as reaction injection molding (RIM) [314], [315]. A schematic of the RIM process is shown in Figure 22. The RIM processing of polyesters, epoxy resins, polyamides, and dicyclopentadienes has also been introduced, although more than 95 % of the total produced by RIM is polyurethane [315]. In the RIM process, the reaction is almost complete by the time the material fills the mold, and therefore the mixing and flow equations must be solved simultaneously with those for chemical reactions in a rational model for these complex situations.

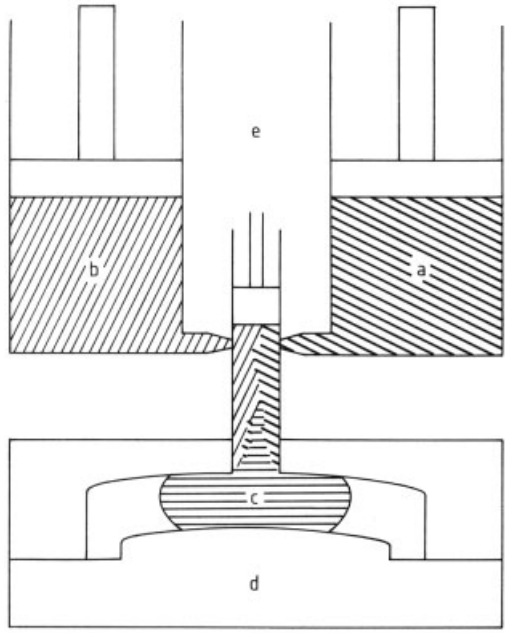

Figure 22. Schematic drawing of the RIM process
a) Monomer A; b) Monomer B; c) Polymerizing mixture; d) Mold; e) Mixer

In the finishing stage of nylon 6, nylon 66, and PETP polymerizations, higher molecular mass polymers may be obtained by solid-state polymerization in which polymerization occurs by heating chips or flakes of a material below its melting point in a stream of hot gases in a fluidized bed or in a drier operated under vacuum [316–320]. The monomer, condensation products, and various byproducts diffuse out, and further reaction takes place inside the solid. The progress of these types of reaction is affected significantly by the diffusion of the condensation products and the morphology of the solid.

Although step-growth polymerization has a very long history, a systematic kinetic treatment like that available for free-radical polymerization does not exist because of limitations due to system-specific side reactions and the scarcity of reliable kinetic data. However, this synthetic route is becoming more important due to the development of materials synthesized by step-growth polymerization such as aramides, PPS, PEK, and PES. The production technology of step-growth polymers seems to have revived as an attractive research area and is enjoying a Renaissance period.

3.2.2. Specific Processes

Polyamides (see also → Polyamides) are manufactured by two basic routes. One of these is synthesis from cyclic monomers such as lactams. Polymerization of these substances requires ring opening and subsequent chain growth. Another class of synthetic polyamides is formed from diamines and diacids.

The most common types of polyamides are nylon 6 and nylon 66. The term nylon is often used for synthetic aliphatic polyamides. One number indicates that the product was prepared from a single monomer and represents the number of carbon atoms in the repeating unit. Two numbers refer to the number of carbon atoms in the diamine and that in the diacid, respectively.

Nylon 6 is typically produced by the hydrolytic polymerization of ε-caprolactam [321–325], although polymers with higher molecular mass can be produced by ionic polymerization [321], [323], [326]. The major reactions in the hydrolytic polymerization are

1) Ring opening of ε-caprolactam by water (Eq. 3.1), which produces aminocaproic acid (ACA)
2) Polycondensation of ACA (Eq. 3.2)
3) Acid-catalyzed polyaddition (ring-opening polymerization) by nucleophilic attack of the amine nitrogen on the lactam (Eq. 3.3)

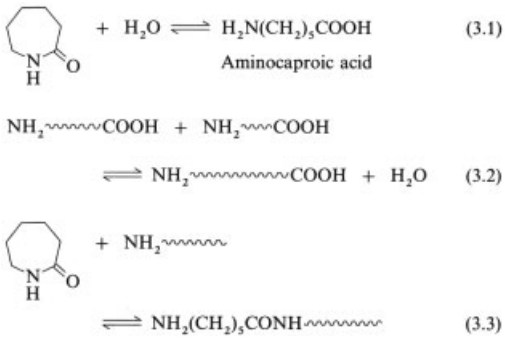

Step-growth polymerization of the amino acid (Eq. 3.2) accounts for only a few percent of total polymerization of ε-caprolactam. However, step-growth polymerization is important since it usually determines the final degree of polymerization at equilibrium. The molecular mass distribution is essentially the most probable distribution, except for the presence of monomer and cyclic oligomers. Since low molecular mass substances lower the polymer quality, they are usually removed by leaching or vacuum treatment of the polymer melt. The formation of cyclic oligomers is an important side reaction [327–329]. For the simulation and optimization of polymerization reactors, only small ring formation or overall ring formation is considered to make the analysis easier [330–333].

Various types of polymerization reactors have been proposed both for batch and continuous processes. A commonly used industrial reactor for a continuous process is a tubular reactor such as the conventional VK column (Vereinfacht Kontinuierliches Rohr) [321], [322], [334], which consists of a vertical tube operating at atmospheric pressure. The feed enters the top of the column and is heated to ca. 220 – 270 °C. The simplest model for this type of reactor is a PFR. However, according to impulse response experiments, the flow is approximately laminar rather than plug flow [322], [335], and the reactor should be modeled as a CSTR followed by

a tubular reactor when a large quantity of water is used since a significant convection current and mixing is provided by the evaporating water [322].

Nylon 66 is manufactured by polycondensation of hexamethylenediamine and adipic acid [336], usually in a multistage process. First, nylon salt (hexamethylenediammonium adipate) is prepared from stoichiometric quantities of hexamethylenediamine and adipic acid in water. The salt can easily be separated by precipitation with methanol.

$$H_2N(CH_2)_6NH_2 + HOOC(CH_2)_4COOH \longrightarrow \begin{bmatrix} {}^{-}O_2C(CH_2)_4CO_2^{-} \\ H_3\overset{+}{N}(CH_2)_6\overset{+}{N}H_3 \end{bmatrix} \quad (3.4)$$

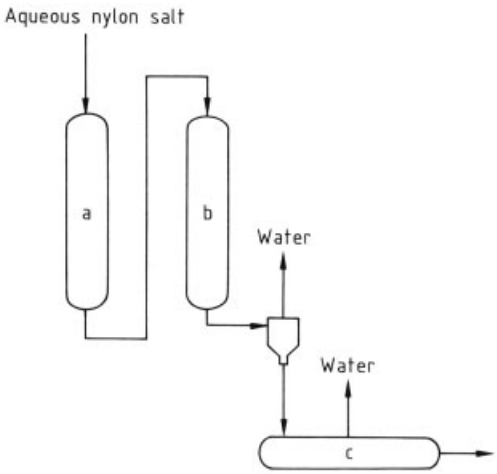

	a	b	c
Hold-up time, min	15–30	15–30	60
Temperature, °C	200–235	270–290	270–290
Pressure, 10^5 Pa	17–24	17–24	atmospheric

Figure 23. Continuous melt polymerization of nylon 66

The use of nylon salt guarantees the presence of equimolar amounts of $-NH_2$ and $-COOH$ groups. Close control of diamine–diacid balance is important to control the final polymer molecular mass and reactive end groups.

Nylon 66 is fairly unstable at high temperatures in the presence of oxygen. Not only degradation but also cross-linking may occur. Because of this instability, polymerization used to be carried out solely in batch processes. However, complete elimination of oxygen has made it possible to carry out continuous polymerization. An example of a continuous melt polymerization process is shown in Figure 23. The aqueous nylon salt solution is heated to above 200 °C at > 17 bar in an oxygen-free atmosphere. Thereafter, the pressure is reduced to atmospheric and vapor is separated from polymer to promote polymerization to the desired high molecular mass. It has also become possible to polymerize molten hexamethylenediamine and adipic acid directly [330], [336]. Polymerization can also be completed in the solid state.

Several kinetic studies on the synthesis of nylon 66 have been published [337–341]. However, more information is necessary for detailed simulation and optimization of nylon 66 reactors.

Polyesters (see also → Polyesters). The production of high molecular mass polyesters differs somewhat from that of polyamides. In the case of nylons, the chemical equilibrium favors the polyamide under polymerization conditions. With polyester formation, however, the equilibrium is much less favorable. In order to drive the reaction in the forward direction, the condensation product must be removed continuously, usually by application of high vacuum. For polyester reactors, a high vacuum, a high temperature, and a high interfacial area with sufficient surface renewal are required, especially at high conversions.

Both saturated and unsaturated polyesters are produced. Among the saturated polyesters, poly(ethylene terephthalate), PETP, is produced in the largest quantity, and is used for production of fibers, films, molding plastics, and beverage containers. In this section, the engineering aspects of PETP formation are illustrated as an example of a polyester production process.

There are two major routes to synthesize PETP industrially, although the objective in each case is to obtain an intermediate product – i.e., bis(hydroxyethyl)terephthalate (BHET).

Two major routes to synthesize BHET are ester interchange of dimethyl terephthalate (DMT) and direct esterification of terephthalic acid.

Figure 24 shows an example of the PETP production process via the ester interchange route. The ester interchange reaction

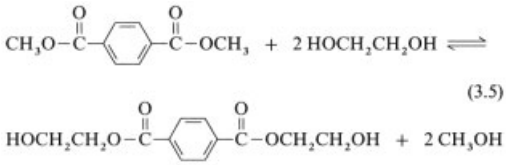

(3.5)

is operated in the temperature range 150 – 210 °C at atmospheric pressure. The use of a catalyst is common [342], [343]. The methanol and ethylene glycol (EG) emerging from the reactor are passed through a rectifying column and EG is fed back to the reactor. It is very difficult to force the ester interchange reaction to completion, and therefore after a particular conversion (usually 90 – 95 %), the reaction mixture is passed on to the polycondensation stage. The reaction mixture consists of oligomers of various types. Oligomers with degrees of polymerization as high as three may be formed [342], [344], [345]. Several reactor models for both batch and continuous processes have been proposed [342–350]. An optimization study showed that the ester interchange reactor should be operated initially at high temperature to obtain high conversion; the temperature should be lowered to reduce side reactions [343], [348].

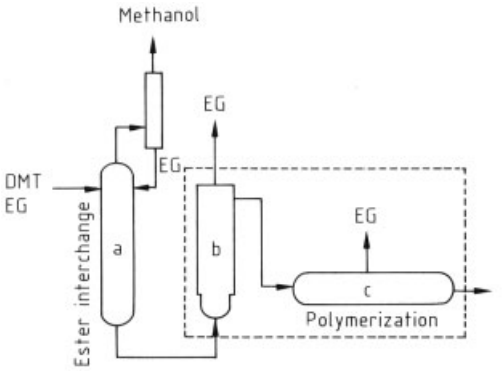

	a	b	c
Hold-up time, h	4 – 6	2 – 3	2 – 3
Temperature, °C	150 – 210	265 – 285	265 – 285
Pressure, 10^2 Pa	atmospheric	13 – 133	≪ 6

Figure 24. Continuous polymerization process of PETP via ester interchange route

In the polycondensation stage, the reaction temperature is raised to 265 – 285 °C to keep the reaction mixture molten and polymerization fast.

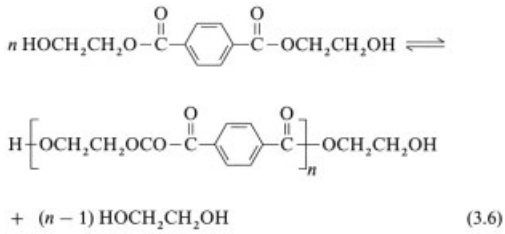

(3.6)

For PETP production, a dual catalyst system in which one component is specially active for ester interchange and the other for polymerization is quite often used [351]. The production of high molecular mass polymer requires the complete removal of ethylene glycol due to the unfavorable equilibrium, and therefore a vacuum is applied. Especially in the final stage of the polycondensation reaction, a very high vacuum is required since the reaction system becomes highly viscous. Consideration of the limitations of mass and heat transfer is very important. Various types of reactors such as rotating disc contactors, wiped film reactors, partially filled screw extruders have been developed as finishers for the polycondensation reaction [342], [343], [350]. Details of fluid mechanics, mixing, and mass and heat transfer characteristics are required for a rational analysis and design of such high-viscosity reactors. In addition to polycondensation reactions, various side reactions must also be considered since a very high temperature is used.

Melt polycondensation of PETP is not generally carried out beyond a particular extent of polymerization since the degradation reactions dominate the process and the product quality may suffer from various undesirable byproducts. To attain higher molecular masses, the products may be subjected to solid-state polymerization [316–320]. Newer processes, especially for beverage and food containers, prefer to stop melt polymerization at lower conversion, and solid-state polymerization is extensively applied.

Direct esterification of terephthalic acid (TPA) and ethylene glycol was generally not preferred earlier because of the difficulties in the purification of TPA due to its low solubility and high melting point. However, with improvements in technology, the direct esterifica-

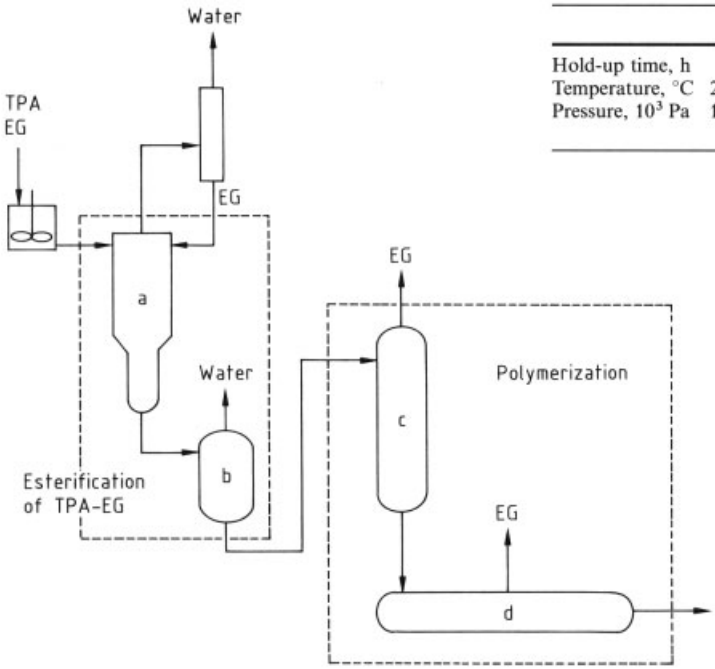

	a	b	c	d
Hold-up time, h	3–5	1–3	2–3	2–3
Temperature, °C	240–260	240–260	265–285	265–285
Pressure, 10^3 Pa	100–300	atmos-pheric	1–10	≪ 0.6

Figure 25. Continuous polymerization process of PETP via direct esterification route

tion method has been gaining in importance. The process is claimed to give polyesters with superior quality due to their low content of carboxyl end groups and diglycol linkages [351]. In the modeling of this process, aside from the difficulties caused by the various reactions and mass balances involved, it is necessary to take account of the heterogenity of the reactions due to the low solubility of TPA in EG. Simulation and control of the direct esterification reactors is reported in [343], [350], [352–354]. Figure 25 shows a flow diagram of a continuous process for PETP production by direct esterification.

3.3. Processes and Reactor Modeling for Chain-Growth Polymerization

3.3.1. Material Balance Equations for Batch, Semi-Batch, and Continuous Reactors

The following material balance equations apply for multicomponent polymerization, accommodate operation of a well-stirred reactor (no spatial variations in temperature and concentrations), and may be used to simulate different comonomer systems under a variety of operating conditions. Bulk (suspension) and solution polymerizations are considered first; extensions required for multiphase systems (emulsion, inverse emulsion, suspension, dispersion, and gas-phase processes) may be found elsewhere [355]. Special cases where spatial variations in temperature and concentrations are important (e.g., in tubular reactors or packed beds) are also considered.

Bulk, solution, and suspension polymerization systems are characterized by the fact that all of the reactions proceed in a single phase with no spatial variations in temperature and concentration. A model for a reactor carrying out such polymerizations would consist of a set of material balances giving the rates of accumulation, inflow, outflow and a reaction source (sink) term for the various monomers, initiators, chain-transfer agents, and polymer in the reactor. These balance equations are now given in general form

Monomer balances:

$$\mathrm{d}N_i/\mathrm{d}t = F_{i,\mathrm{in}} - (N_i/V)\,V_{\mathrm{out}} - R_{\mathrm{p}_i}V \tag{3.7}$$

where

N_i is the number of moles of monomer i in the reactor
$F_{i,\mathrm{in}}$ are molar flow rate of monomer i into the reactor
V is the reaction volume in the reactor
V_{out} is the total volumetric flow rate of all species out of the reactor
R_{p_i} is the net rate of disappearance of monomer i by reaction

Reaction Volume. Since the density of a polymer is usually significantly greater than that of its monomer, the reaction volume V decreases with conversion for isothermal polymerization in a batch reactor. This shrinkage must also be taken into account in semi-batch and continuous operations. Neglecting volume change on mixing polymer and monomers (thermodynamic data are most often not available and deviation from ideality is often not great) the change in reaction volume (V) may be calculated by using (assuming V is the equilibrium volume):

$$\begin{aligned} \mathrm{d}V/\mathrm{d}t &= V_{\mathrm{in}} - V_{\mathrm{out}} - \text{shrinkage rate} \\ &= V_{\mathrm{s,in}} + \sum_{i=1}^{n} F_{i,\mathrm{in}} M_{\mathrm{m}i}/\varrho_{\mathrm{m}_i} \\ &\quad - \sum_{i=1}^{n} R_{\mathrm{p_i}} M_{\mathrm{m}i} \left(\frac{1}{\varrho_{\mathrm{m}_i}} - \frac{1}{\varrho_{\mathrm{p}}} \right) V - V_{\mathrm{out}} \end{aligned} \quad (3.8)$$

where n is the number of monomer types, $V_{\mathrm{s,in}}$ is the volumetric flow rate of inert solvent into the reactor, $M_{\mathrm{m}i}$ is the molecular mass of monomer i, ϱ_{m_i} is the density of monomer i, ϱ_{p} is the density of polymer produced instantaneously.

Polymer Balances. With a batch reactor, where there is no inflow and outflow of polymer from the reactor, the total amount of polymer formed and its composition can be obtained directly from the monomer balances. However, with semi-batch and continuous operation, additional balances are required and these follow

$$\mathrm{d}P_i/\mathrm{d}t = F_{\mathrm{p}i,\mathrm{in}} - (P_{\mathrm{i}}/V)\, V_{\mathrm{out}} + R_{\mathrm{p_i}} V \quad (3.9)$$

where P_i is the number of moles of monomer i chemically bound in the polymer "in the reactor", $F_{\mathrm{p}i,\mathrm{in}}$ is the molar flow rate of monomer i bound in the polymer flowing into the reactor.

Additional Ingredient Balances. In order to calculate R_{p_i} for free-radical systems, the total polymer radical concentration and, therefore, initiator concentration are required. In addition, balances for the chain-transfer agent (for molecular mass calculations) and for the solvent in solution polymerizations are required. These balances follow:

$$\mathrm{d}N_{\mathrm{I}i}/\mathrm{d}t = F_{\mathrm{I}i,\mathrm{in}} - (N_{\mathrm{I}i}/V)\, V_{\mathrm{out}} - R_{\mathrm{I}i} V \quad (3.10)$$

$$dV_{\mathrm{s}}/\mathrm{d}t = V_{\mathrm{s,in}} - (V_{\mathrm{s}}/V)\, V_{\mathrm{out}} \quad (3.11)$$

$$dN_{\mathrm{T}}/\mathrm{d}t = F_{\mathrm{T,in}} - (N_{\mathrm{T}}/V)\, V_{\mathrm{out}} - R_{\mathrm{T}} V \quad (3.12)$$

where $N_{\mathrm{I}i}$ is the number of moles of initiator of type i in the reactor, $F_{\mathrm{I}i,\mathrm{in}}$ is the molar flow rate of initiator i into the reactor, $R_{\mathrm{I}i}$ is the consumption rate of initiator i by reaction, V_{s} is the volume of inert solvent in the reactor, N_{T} is the number of moles of chain-transfer agent (CTA) in the reactor, $F_{\mathrm{T,in}}$ is the molar flow rate of CTA into the reactor, and R_{T} is the consumption rate of CTA by reaction.

It is convenient to sum the monomer balance equation (Eq. 3.7) over n, the number of monomer types to give

$$\mathrm{d}N_{\mathrm{M}}/\mathrm{d}t = F_{\mathrm{in}} - (N_{\mathrm{M}}/V)\, V_{\mathrm{out}} - R_{\mathrm{p}} V \quad (3.7a)$$

where N_{M} is the total number of moles of monomer in the reactor, F_{in} is the total molar flow rate of monomer to the reactor, R_{p} is the total molar consumption rate of monomer by reaction.

3.3.1.1. Rates of Reaction and Copolymer Composition

With application of the pseudo-kinetic rate constant method, R_{p} can be expressed as

$$R_{\mathrm{p}} = K_{\mathrm{p}}\, [\mathrm{M}]\, [\mathrm{P}^*] \quad (3.13)$$

where

$$[\mathrm{M}] = N_{\mathrm{M}}/V = \sum_{i=1}^{n} [\mathrm{M}_i] \quad (3.13a)$$

and $[\mathrm{M}_i]$ is the concentration of monomer of type i in the reactor.

When the terminal model [356] for copolymerization is valid

$$[\mathrm{P}^*] = \sum_{i=1}^{n} [\mathrm{P}_i^*] \tag{3.13b}$$

where $[\mathrm{P}_i^*]$ is the concentration of active centers of type i in the reactor and

$$K_\mathrm{p} = \sum_{i=1}^{n} \sum_{j=1}^{n} K_{\mathrm{p_{ij}}} \varphi_i^* f_j \tag{3.13c}$$

where φ_i^* is the number fraction of active centers of type i in the reactor, and f_j is the mole fraction of monomer of type j in the reactor.

When the penultimate model [356] for copolymerization is valid

$$[\mathrm{P}^*] = \sum_{ij}^{n} [\mathrm{P}_{ij}^*] \tag{3.13d}$$

and

$$K_\mathrm{p} = \sum_{ijk}^{n} K_{\mathrm{p}_{ijk}} \varphi_{ij}^* f_k \tag{3.13e}$$

The fraction of active centers of type i (φ_i^*) can be found by using the stationary state hypothesis (SSH). Model development will continue assuming that the terminal model for copolymerization is valid. The φ_i^* may be found by using

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} R_{\mathrm{p}_{ji}} = \sum_{\substack{j=1 \\ j\neq i}}^{n} R_{\mathrm{p}_{ij}} \tag{3.14a}$$

with $i = 1, \ldots n$ and where $R_{\mathrm{p}_{ij}}$ is the consumption rate of monomer j adding to the active center i.

For anionic, cationic, and anionic coordination polymerization, the estimation of the total number of active centers is not always straight forward, and this quantity is sometimes used as an adjustable parameter. For free-radical polymerization, a balance between radical generation rate and bimolecular termination rates provides the following with use of the SSH

$$[\mathrm{R}^\cdot] = \left(R_\mathrm{I}/\bar{K}_\mathrm{tN}\right)^{1/2} \tag{3.14b}$$

where $[\mathrm{R}^\bullet]$ is the total concentration of polymeric radicals in the reactor, R_I is the generation rate of polymeric radicals of chain length unity, and $\bar{K}_\mathrm{tN}$ is the number-average bimolecular termination constant. When termination by combination and disproportionation are both significant

$$\bar{K}_\mathrm{tN} = \bar{K}_\mathrm{tcN} + \bar{K}_\mathrm{tdN} \tag{3.14c}$$

For chemically controlled bimolecular termination, pseudo-kinetic rate constants for $\bar{K}_\mathrm{tcN}$ and $\bar{K}_\mathrm{tdN}$ may be found in [355–357]. Details concerning the definitions and use of $\bar{K}_\mathrm{tN}$, $\bar{K}_\mathrm{tW}$, $\bar{K}_\mathrm{tZ}$, may be found elsewhere [358], [359]. The definition of $\bar{K}_\mathrm{tN}$ follows:

$$\bar{K}_\mathrm{tN} = \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} K_\mathrm{t}\,(r,s)\,\psi\,(r)\,\psi\,(s) \tag{3.14d}$$

where $K_\mathrm{t}\,(r, s)$ is an arbitrary bimolecular termination constant distribution for reaction of polymeric radicals of chain length r and s and $\psi\,(r)$ and $\psi\,(s)$ are the number fractions of polymeric radicals of chain length r and s. Note that $\bar{K}_\mathrm{tN}$ can be used to calculate R_p and $\bar{M}_\mathrm{N}$ but if used to calculate higher molecular mass averages they would be underestimated if a significant number of polymer chains are formed by bimolecular termination reactions [358].

Realistic calculations of [R] using Equation (3.14 b) requires that the SSH be valid and that diffusion-controlled bimolecular termination be accounted for. The first direct experimental test by ESR of the validity of the SSH for bulk polymerization of methyl methacrylate (MMA) (where linear polymer chains are produced) and of MMA – ethylene glycol dimethacrylate (EGDMA) (where chains with long branches and polymeric networks are formed) has shown that SSH is valid for bulk MMA polymerization, but is not valid when substantial crosslinking occurs [360]. Coyle et al [361] have confirmed the validity of SSH for bulk polymerization of MMA using numerical solutions of the full set of kinetic equations. Attempts to model diffusion-controlled bimolecular termination is considered later when discussing the polymerization of MMA (see page 53) which is considered the model system because of its extreme Trommsdorff – Norrish effect. At very high monomer conversions, when the polymerization temperature is below the glass transition temperature of the polymer being synthesized, the initiator efficiency and propagation constant

both begin to fall dramatically, and these effects should be properly accounted for. These effects will also be considered in detail later when MMA polymerization is discussed. The modeling of diffusion-controlled bimolecular termination, propagation, and the cage effect on initiator efficiency is not entirely clear, and much research must be done before these topics may be considered standard engineering practice.

The polymerization rates for individual monomer type j can be calculated from

$$R_{\mathrm{p}j} = \left(\sum_{i=1}^{n} K_{\mathrm{p}_{ij}} \varphi_i^* f_j \right) [\mathrm{M}]\,[\mathrm{P}^*] \tag{3.14e}$$

with $j = 1, \ldots n$.

The mole fraction of monomer of type j in copolymer produced instantaneously is given by

$$F_j = R_{\mathrm{p}_j} / R_{\mathrm{p}} \tag{3.14f}$$

Monomer sequence length distributions may be calculated by using equations given in [356].

3.3.1.2. Molecular Masses, Long-Chain Branching, and Cross-Linking

Linear Copolymer Chains. For chain lengths greater than about 50 and when the terminal model for copolymerization is valid, STOCKMAYER's bivariate weight chain length distribution may be used to calculate the composition and chain length of binary copolymer produced instantaneously [356], [362]. The bivariate distribution of the accumulated polymer is readily calculated by integrating the instantaneous distribution using weighting factors based on instantaneous rates of polymer production. Unfortunately, an analytical function for a multidimensional distribution (involving three or more monomer types) has yet to be derived. However, part of this problem has already been solved. In the limit of large chain lengths, the copolymer composition of chains produced instantaneously may be assumed to be independent of chain length, and therefore, the instantaneous bivariate weight chain length distribution is that given by the same expression as for homopolymerization with all of the chains having the same composition F_i. For more details, see Section 2.3.1 and [356].

Long-Chain Branching and Cross-Linking. The production of long branches and cross-links requires that so-called dead polymer chains take part in branching reactions to produce tri- and tetrafunctional branch points via transfer to polymer and by addition of polymer radical centers to polymer double bonds. The mechanisms of anionic, cationic, and anionic coordination polymerization invariably produce linear polymer chains. There may be a few exceptions, but these will not be considered.

Pseudo-kinetic rate constants for transfer to polymer and polymer radical addition to pendant double bonds may be defined as [355], [357], [363–365]

$$K_{\mathrm{fp}} = \sum_{ij}^{n} K_{\mathrm{fp}_{ij}} \varphi_i \bar{F}_j \tag{3.15}$$

$$K_{\mathrm{p}}^* = \sum_{ij}^{n} K_{\mathrm{p}_{ij}}^* \varphi_i \bar{F}_j \tag{3.16}$$

when the terminal model is valid; where $\bar{F}_j$ is the mole fraction of monomer j chemically bound in the accumulated polymer. When a significant number of labile atoms have been abstracted and when a significant number of double bonds have been consumed, Equations (3.15) and (3.16) should be modified to account for this [363], [364].

When polymer chains that are normally inert to further reaction undergo long-chain branching reactions, the instantaneous MWD is no longer a permanent quantity. Therefore, the method of moments should be used to calculate the molecular mass averages. For details on the use of the method of moments for the calculation of sol molecular mass averages before and after the gelation point and other methods for calculating sol – gel fraction and cross-linking density, see [363–365].

3.3.2. Examples of Free-Radical Polymerization

3.3.2.1. Homopolymerization – Linear Chains

The modeling techniques described in Section 3.3.1 are illustrated with actual monomer systems and experimental kinetic data. The simplest modeling examples (homopolymerization

producing linear chains) are considered first, beginning with the thermal bulk polymerization of styrene, which is relatively easy to model and the modeling has been most successful in a variety of applications.

Bulk Thermal Polymerization of Styrene ($T > 100\,°C$). This system is comparatively easy to model for the following reasons:

1) The Mayo mechanism for thermal initiation of radicals is valid [366–368]
2) The polystyrene chains are linear
3) The Trommsdorff – Norrish effect (see Section 2.2.1.2), although significantly affecting the polymerization rate, has at most a minor effect on molecular mass development because most of the polymer chains are produced by chain transfer to the Diels – Alder intermediate [366].

The size of the polymeric radicals (radical centers are exclusively on chain ends) depends on temperature and monomer concentration (or conversion), and the molecular mass distribution of the accumulated polymer differs little from that of the polymeric radicals. In general, the self-diffusion coefficients of polymer radicals should depend on the size of the macroradical, the concentration and MWD of the accumulated polymer, and temperature. The size of the macroradical is of greater importance than the MWD of the accumulated polymer [369], [370].

An empirical correlation of the bimolecular termination constant with polymerization temperature and monomer conversion should provide a model which is applicable over a range of process conditions (batch and continuous reactor operation, temperature programming, etc). Another important factor, revealed in the work of Kirchner and Riederle [367], is that reactions involved in thermal initiation do not become diffusion controlled at monomer conversions as high as 97 %. The first effective model for the bulk thermal polymerization of styrene was developed by Hui and Hamielec [371] and later extended to higher temperatures ($T < 230\,°C$) [372] and shown to be valid for a continuous stirred-tank reactor up to 280 °C [373]. This model has been evaluated by many workers over a wide range of reactor type and operational conditions. These include:

1) Wu et al. [374] used the model for a theoretical/experimental optimization study employing temperature programming in a batch reactor with highly successful results.
2) Kirchner et al. [375] applied the model to a CSTR and obtained accurate predictions.
3) Tien et al. [375] applied the model to a tubular reactor with internal mixers.
4) DIERS [376], [377] (Design Institute for Emergency Relief Systems) a consortium of chemical and insurance companies financed the design and construction of a unique adiabatic reactor system (see Fig. 26) for the measurement of temperature and pressure/time variations during adiabatic runaway exothermic reactions. They measured responses for the adiabatic thermal polymerization of styrene (Fig. 27 and 28) and compared them with those predicted by the Hui – Hamielec polystyrene model and found excellent agreement.

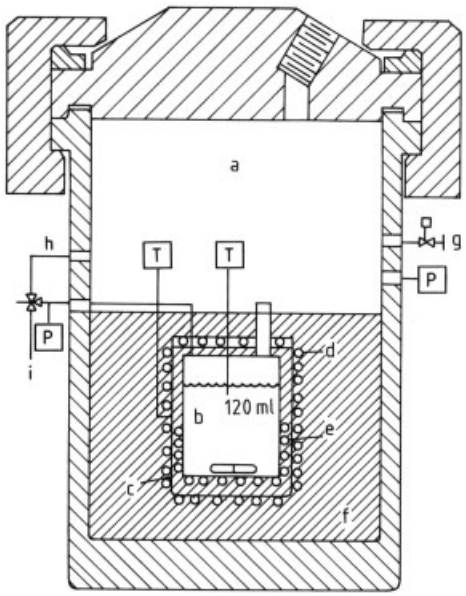

Figure 26. DIERS VSP (adiabatic batch reactor)–monitors temperature and pressure changes during runaway of highly exothermic reactions [377]
a) Containment vessel (ca. 4 L); b) Test cell; c) Outer can; d) Guard heater; e) Inner heater; f) Insulation; g) Exhaust and supply; h) Bypass; i) Fill

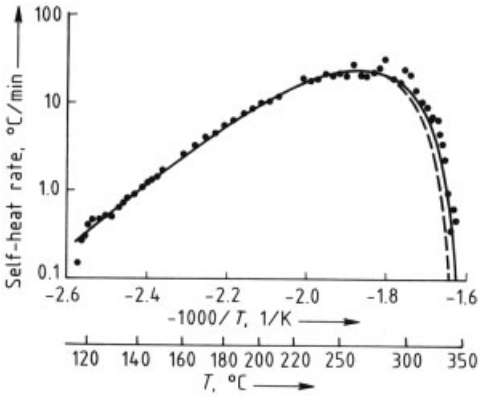

Figure 27. VSP self-heat rate response for the adiabatic thermal polymerization of styrene (80 wt % styrene and 20 wt % ethylbenzene) compared with the predictions of the H–H polystyrene model [377]
– – – Constant heat of reaction (– 700 kJ/kg); —- Temperature-dependent heat of reaction; • Experimental data

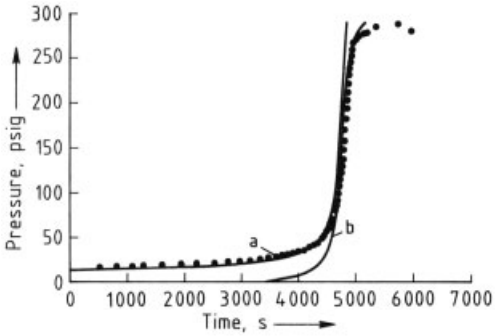

Figure 28. VSP Pressure – time response for the adiabatic thermal polymerization of styrene (80 wt % styrene and 20 wt % ethylbenzene) compared with the predictions of the H–H polystyrene model with temperature-dependent heat of reaction [377]
(kPa = psi × 6.89)
a) Prediction: total pressure (maximum 360 psig);
b) Prediction: partial vapor pressure (maximum 312 psig)
• Experimental data

The pressure calculation required the use of the Flory – Huggins equation (Eq. 3.17). These evaluations under extreme conditions confirm that the H–H model for the thermal polymerization of styrene is indeed useful in a variety of applications. These kinds of evaluations, which, however, have not been carried out for any other polymerization model, should be considered standard tests for polymerization models before they are accepted for general engineering use.

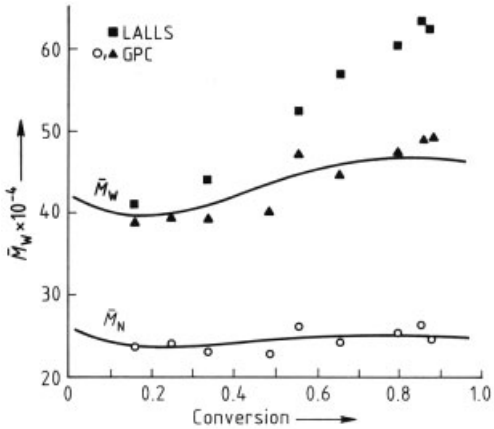

Figure 29. Thermal polymerization of *p*-methylstyrene: molecular mass development followed by GPC and light scattering (LALLS). Deviation for $\overline{M}_W$ between GPC and LALLS suggests that LCB frequency is increasing with monomer conversion. Solid curves are model predictions neglecting chain transfer to polymer [378].

A monomer which polymerizes thermally and almost identically to styrene is 4-methylstyrene [378]. With the latter monomer some long-chain branching occurs due to chain transfer to methyl hydrogens in the polymer backbone. An effective experimental technique to confirm the presence of long branches [378] is to measure $\bar{M}_W$ by light scattering and by GPC at different monomer conversions in isothermal batch polymerization. The molecular mass calibration curve for GPC should be constructed using linear poly(4-methylstyrene). At low conversions, where chains are almost all linear, $\bar{M}_W$ by light scattering and GPC should agree. However, at high conversions where long-chain branching should be significant (if it is present at all), $\bar{M}_W$ by light scattering should be greater than that by GPC and the difference should increase with increasing monomer conversion. Figure 29 shows an example of this behavior. Long-chain branching in poly(4-methylstyrene) has been satisfactorily modeled by accounting for chain transfer to polymer and using light scattering $\bar{M}_W$'s which should be valid for branched homopolymer [378]. Another effective procedure to verify the presence or absence of long-chain branching is to add a previously synthesized narrow-MWD sample of the polymer that is to be synthesized to the monomer solution and then polymerize. The newly synthesized polymer is then com-

pletely separated from the old polymer by GPC. If the GPC response for the added polymer shifts towards larger radii of gyration (hydrodynamic volume) this is evidence for long-chain branching [379].

Bulk polymerization of methyl methacrylate (MMA) with free-radical initiator is the model monomer system for the investigation of diffusion-controlled bimolecular termination and propagation and the decrease in initiator efficiency at high monomer conversions (cage effect). It has the largest Trommsdorff – Norrish effect because most of the polymer chains are produced by bimolecular termination and the higher average molecular masses ($\bar{M}_W$, $\bar{M}_Z$, $\bar{M}_{Z+1}$) increase dramatically in the absence of a chain-transfer agent. The MWD shifts to higher molecular masses and sometimes becomes bimodal [380]. GPC detector responses, multiplied by monomer conversion so that the total area under the detector response is proportional to the amount of polymer in the batch reactor, are shown in Figure 30. At the monomer conversion where the number of physical chain entanglement points becomes significant a spike of high molecular mass polymer is produced. Thereafter, polymer with lower molecular masses is no longer produced. The instantaneous MWD has clearly shifted towards higher molecular mass (higher molecular mass polymer has a lower GPC retention volume or retention time). MMA is generally polymerized below the glass transition temperature of PMMA (ca. 110 °C) and as a consequence the initiator efficiency and propagation constant decrease dramatically at high monomer conversions ($\sim$ 80 %) due to restrictions on diffusion rates during the glassy-state transition because of appreciable loss in free volume. There is also evidence that during the rapid autoacceleration in polymerization rate (due to large increase in polymer radical concentration [360]) radicals become frozen in the glassy-state [358]. Radical pairs which form in the cage may become frozen after a few monomer additions because of monomer starvation. This homopolymerization is difficult to model. Nevertheless, there have been many serious attempts to model this system [381–392]. More recent research has focused on predicting the fall in initiator efficiency and propagation constant at high conversion and on the use of a simple chain-length-dependent model for bimolecular termination. This work has been summarized in a recent article by Adams et al. [393]. The first important models [381–383] for MMA polymerization were carefully evaluated [384] using comprehensive rate and molecular mass data involving three different initiator types measured by Röhm. The Marten – Hamielec model was found to better satisfy the specifications for a polymer reactor model that can be used to optimize commercial production systems. It was pointed out however, that further work should be done to investigate the applicability of the model to systems that have been prepolymerized and also polymerized nonisothermally (perhaps with temperature programming or adiabatically) and for systems with mixed initiators. Panke [394] has more recently shown that when using prepolymer, simulations are better when the parameter n is changed from 1.75 to 0.5 in the M–H model. From a fundamental point of view the M–H model has several weaknesses which should be pointed out. Chain-length dependence of bimolecular termination is accounted for in an overly simplified manner. The weight-average molecular mass of the accumulated polymer does reflect the change in the size of the macroradicals but in a dampened fashion. For example, a sudden change in the size of the macroradicals would not be felt soon enough, particularly when a substantial amount of dead polymer has accumulated. This model effectively uses one termination constant (the number-average termination constant [358]). It can therefore predict polymerization rate and number-average molecular mass, but invariably underestimates $\bar{M}_W$ and higher molecular mass averages. The M–H model as well as most others have assumed that K_p becomes diffusion controlled while the initiator efficiency f remains constant. It has long been known that the initiator efficiency falls at high conversions [395–398]. To separate the product $f^{\frac{1}{2}}K_p$ which appears in the rate expression, accurate molecular mass data at high conversions are required. This has not been possible to date, because of the difficulty of measuring $\bar{M}_N$ and $\bar{M}_W$ for the very high molecular mass PMMA produced during the autoacceleration of reaction rate. For all of the models for which f was taken to be independent of monomer conversion, the observed decrease in K_p is actually

the decrease in $f^{\frac{1}{2}}K_p$. The change in K_p at very high monomer conversions can be measured by ESR [360], [399] and of course this would permit the estimate f from the product $f^{\frac{1}{2}}K_p$. Another factor which has been completely overlooked is that during the autoacceleration when approaching the glassy-state transition, the actual volume of the polymerizing mixture may be significantly greater than the local equilibrium volume. This is largely responsible for the fact that the limiting conversion depends on the initiator concentration at high initiator and radical concentration levels [400–402].

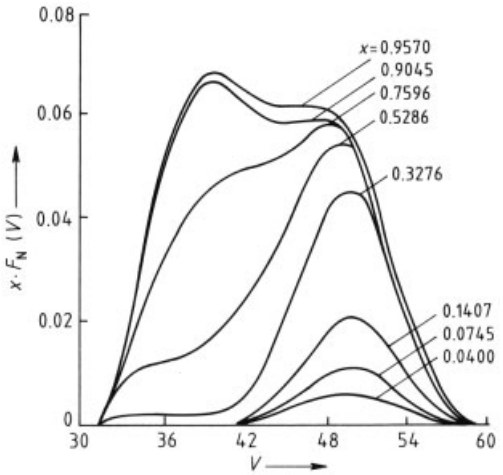

Figure 30. GPC chromatograms showing molecular mass development due to the Trommsdorff–Norrish effect for the isothermal bulk free-radical polymerization of methyl methacrylate in a batch reactor
Parameter x denotes monomer conversion ($T = 70\,°C$, [AIBN] = 0.3 wt %) [380]

During the autoacceleration in polymerization rate which accompanies diffusion-controlled bimolecular termination, it is difficult to maintain isothermal conditions. The use of large ampoule reactors with inadequate surface-to-volume ratios for heat transfer has been recently criticized [403], [404]. The use of ampoule reactors (3 – 5 mm diameter) is recommended for kinetic investigations of the Trommsdorf – Norrish effect [404].

3.3.2.2. Copolymerization – Linear Chains

The terminal model has been very useful for predicting the average compositions of binary copolymers synthesized with vinyl monomers having a wide variety of structures and range of reactivities. The terminal model has not been as successful in the prediction of comonomer sequence length distributions [405], [406] and the propagation rate [407–413]. Model discrimination has been based on the use of measurement techniques which provide estimates of the elementary overall (or pseudo-kinetic) rate constants K_p and $\bar{K}_{tN}$. These include rotating sector [407–409], spatially intermittent polymerization (SIP) reactor [414], and pulsed-laser techniques [411–413]. ESR measurements [410] have also been used to show the existence of penultimate effects. Most of these studies considered the binary copolymerization of styrene and methyl methacrylate; however, other monomer pairs have been investigated including: styrene – acrylonitrile [408] and styrene with a series of alkyl methacrylates [408], [415]. The use of the pulsed-laser technique has a serious limitation when applied to copolymerization. One must measure the MWD by GPC and the potential errors are several and potentially serious. For example, there is inevitable statistical broadening of composition distribution for short chains and the uncertainty that the concept of universal molecular mass calibration is valid. Nonuniform composition of chains in the detector cell makes the conversion of detector response to polymer solute concentration uncertain and in addition the errors due to deviation from the universal calibration curve may be significant.

Most of the serious attempts to model binary copolymerizations from an industrial perspective (accounting for diffusion-controlled bimolecular termination and propagation and for the fall in initiator efficiency at high conversions) have employed the terminal model and the pseudo-kinetic rate constant method [355], [357], [416–423]. There have been some recent attempts to employ the penultimate model in this regard [424], [425]. It should be stressed that whether terminal or penultimate models are used appropriate pseudo-kinetic rate constants can be defined for calculating rates of polymerization and molecular mass distribution [356].

3.3.2.3. Copolymerization – Long-Chain Branching

The few serious attempts to model polymerization rate and molecular mass development have employed the pseudo-kinetic rate constant method with the method of moments [355], [357], [363–365], [418], [419], [422]. The method of instantaneous MWD, which is such a powerful method for calculating the full MWD for linear multicomponent polymers is not useful when dead polymer chains can be reactivated during formation of long-chain branches. The instantaneous MWD is no longer a permanent quantity but loses some of its chains, which become branched during the course of polymerization. It should be noted here that the pseudo-kinetic rate constant method is equally valid for the modeling of linear and branched copolymer chain synthesis. The most common analytical technique for the measurement of long-chain branching frequency (average number of branches per polymer molecule or per 1000 backbone carbon atoms) is GPC with a dual detector system (mass concentration detector plus either a viscometer or light scattering photometer as detector in series) or GPC with off-line viscometry [426].

3.3.3. Polymerization Processes

3.3.3.1. Solution Polymerization

3.3.3.1.1. Polymer Soluble in Monomer

Kinetics. A polymerization in which the polymer being synthesized is soluble in its monomer may be called a bulk or mass as well as a solution polymerization. A full description of the polymerization kinetics and modeling techniques can be found in Section 2.2.1 and Section 3.3.2.

Polymerization Processes. The main technical problems associated with solution polymerization are heat removal, recovery of residual monomer and solvent, and the manipulation of highly viscous solutions and melts. It is well known that the ratio of cooling surface area to volume of reacting mixture decreases as the reactor volume increases. Length-to-diameter ratios for reactors are usually < 2 to achieve acceptable mixing. When jacket cooling is no longer sufficient to maintain isothermal polymerization, additional modes of heat transfer must be used, as shown in Figure 31. Internal cooling coils are often not practical because they tend to interfere with stirring. External tubular coolers can, in principle, provide a very large heat-transfer area, but they may have very large pumping requirements in the case of highly viscous solutions. Reflux cooling removes the heat of polymerization by evaporation of solvent and/or monomer; the condensed vapor is recycled to the reacting mass. Condensers may be as large as necessary, with the limiting factor usually the amount of vapor that can be treated without causing intense foaming or spattering of the polymer solution. Remixing of the condensed liquid with the more viscous reacting mass may also be difficult.

3.3.3.1.2. Addition of a Solvent in which both Monomer and Polymer are Miscible

The addition of a solvent in which both monomer and polymer are miscible lowers the viscosity of the reacting mass, thereby improving its flow and heat-transfer characteristics. As a result, and depending upon the nature and con-centration of the solvent, the Trommsdorff – Norrish effect can be either completely suppressed or significantly reduced. Use of solvent can be especially beneficial when evaporative cooling is used. In choosing a solvent, it is important to take into account the possibility of chain transfer to solvent with a concomitant reduction in polymer molecular mass. Removal of solvent and residual (unreacted) monomer from the highly viscous polymer solution requires very high surface areas to permit rapid devolatilization at the moderate temperatures required to minimize polymerization and degradation of chains during devolatilization to reduce off-spec polymer. Figure 32 illustrates possible designs of devolatilizers. One of these, the vacuum degasser, incorporating a type of spray device [427], operates adiabatically, which means that the heat for evaporation is supplied by the solution itself. In contrast, a degasser in series with a tubular heat exchanger causes some of the monomer/solvent to

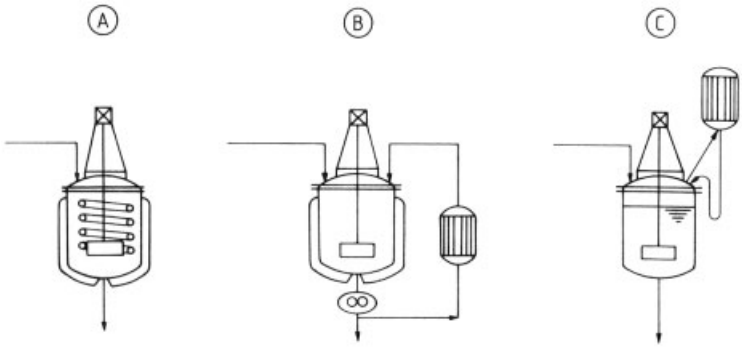

Figure 31. Alternative methods of heat removal for polymerization conducted in a stirred reactor
A) Internal cooling coils; B) External cooler; C) Reflux cooling

evaporate during passage through the heat exchanger. This evaporation accelerates the flow of product, thereby increasing the heat-transfer rate [428]. Twin-screw extruders with one or more vapor outlets that can be connected to a vacuum source are also utilized for the removal of monomer and solvent [429], [430]. Intermeshing and self-cleaning screws provide continuous renewal of the evaporating surface. The use of thin-film evaporators with rotating wiper blades has also been suggested [431].

Given stable operation of the devolatilizer, solution polymerization can be carried out in a continuous stirred-tank reactor [432], [433]. Conversion is normally lower than with tower processes, however, requiring large amounts of monomer/solvent to be recovered and recycled to the reactor. Special attention must therefore be paid to the buildup of impurities (e.g., inhibitors, chain-transfer agents) over time by using appropriate purge streams. Problems with processes involving step-growth polymerization are discussed in Section 3.2.

3.3.3.1.3. Polymer – Polymer Demixing during Polymerization

Polymer – polymer demixing is especially important in the production of thermoplastics whose application characteristics are enhanced by the presence of dispersed domains containing an elastomer. Generally, the thermodynamics are unfavorable for complete miscibility between different types of polymers. For a system polymer A – polymer B – solvent, the following generalizations are usually applicable [434], [435]: dilution with solvent and, in most cases, an increase in temperature increases the compatibility; increasing the molecular masses of the polymers has the opposite effect. Except in the case of high dilution, incompatibility in solution is the rule for pairs of polymers, even when the solvent is a good solvent for both polymer types.

Systems such as polymer A – polymer B – block or graft copolymer AB can be regarded as a polymeric oil-in-oil emulsion in which the copolymer functions as an emulsifying agent [436–439]. Such systems arise, for example, in the manufacture of high-impact polystyrene (HIPS), or in the preparation of ABS (acrylonitrile – butadiene – styrene) by solution polymerization. In the simplest case, ca. 5 – 10 % polybutadiene is dissolved in monomeric styrene to give a homogeneous solution suitable for polymerization. The polystyrene synthesized is incompatible with the polybutadiene present, causing phase separation even at very low monomer conversion and producing a polybutadiene – styrene continuous phase and a polystyrene – styrene disperse phase. Simultaneously, graft polymerization produces polystyrene branches on polybutadiene backbone. The graft copolymer serves as an emulsifier, accumulating at the interface and stabilizing the oil-in-oil emulsion.

With increasing conversion, the volume fraction of the polystyrene phase (which is initially small) increases considerably due to formation of additional polystyrene which absorbs styrene monomer. Finally, often at a phase-volume ratio of about unity, a phase reversal occurs, with the rubber phase now the disperse phase, and the

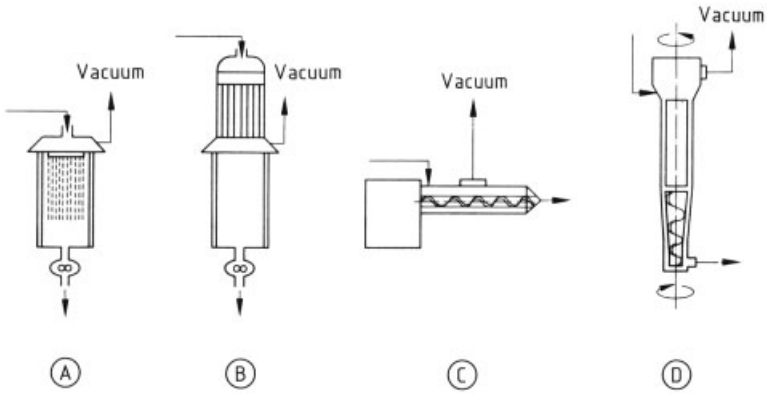

Figure 32. Options for removing residual monomer and solvent
A) Strand degasser; B) Tubular evaporator; C) Degassing extruder; D) Thin-film evaporator

polystyrene phase the continuous phase. Agitation is essential for the completion of phase reversal, since rapid approach to equilibrium with respect to transfer of monomer/polymer must be attained in a highly viscous medium. A common result of inadequate agitation is the interpenetration of two continuous phases [438], [440–442].

Phase reversal can also lead to various types of emulsions; e.g., polystyrene – styrene droplets can be occluded within the rubber particles of the disperse phase. This is actually desirable as it increases the volume fraction of the disperse rubber phase. A reinforcing effect is observed, with the resulting HIPS having higher impact strength for a given mass of rubber in comparison with materials prepared by emulsion grafting, in which occluded polystyrene is not formed [443], [444]. Rather than rely on in situ formation of grafted rubber, it is also possible from the outset to utilize block copolymers of styrene and butadiene as an additive in the recipe. The desired particle size distribution and morphology of the rubber particles in HIPS is then achieved by varying the intensity of agitation, the viscosities of the disperse and continuous phases, the graft activity of the primary radicals of the initiator, the molecular mass of the continuous phase, and the mass fraction and molecular mass of the blocks in the diblock copolymer [445–454].

Figure 33 provides examples of the morphology of these rubber-modified polystyrene systems. Morphologies of commercial significance are essentially limited to cell and capsule structures, since the other morphologies shown generally provide lower impact strength.

3.3.3.2. Precipitation Polymerization

The term precipitation polymerization refers to processes in which the initial ingredients of a recipe are soluble, giving a homogeneous solution, but the synthesized polymer precipitates during the course of polymerization. The precipitated polymer is generally swollen with monomer and with nonsolvent if present. To ensure convenient handling the disperse phase must be finely divided. This is achieved by effective agitation, and a protective colloid (steric stabilizer) is often used. The associated problems are similar to those with HIPS manufacture.

3.3.3.2.1. Polymer Insoluble in its Monomer

Polymers that are insoluble in their monomers include poly(vinyl chloride) and other polymers derived from halo or pseudohalo-substituted ethylenes such as vinyl bromide, vinylidene chloride, trifluoroethylene, and acrylonitrile. Polyethylene also falls into this category, at least when produced under moderately high pressure. For the system vinyl chloride – poly(vinyl chloride), for example, virtually no polymer dissolves in the monomer in the temperature range 30 – 60 °C. Solubility of the monomer in the polymer is described well by the simplified

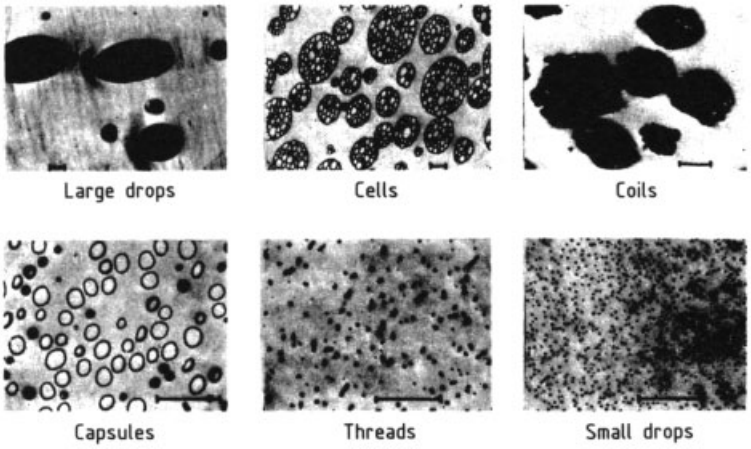

Figure 33. Possible morphologies for rubber particles in impact resistant polystyrene [454]
Length of the scale bar: 1 μm

Flory – Huggins equation (3.17) [435, p. 511], [455–457]:

$$\ln\left(\frac{p}{p_0}\right) = \ln\,(1-\varphi_{\mathrm{P}}) + \varphi_{\mathrm{P}} + \chi\varphi_{\mathrm{p}}^2 \qquad (3.17)$$

where

p_0 is the vapor pressure of pure vinyl chloride
p is the partial pressure of vinyl chloride
φ_{p} is the volume fraction of poly(vinyl chloride)
χ is the polymer – solvent interaction parameter.

According to [457], deviations from Equation (3.17) are expected at low vapor pressures. The conversion – time curves for the bulk polymerization of such monomer – polymer systems all display rate increases with increasing conversion. This phenomenon has been especially thoroughly investigated for the bulk polymerization of vinyl chloride. Its cause may be interpreted as follows: even at very low conversion, poly(vinyl chloride) precipitates, forming a monomer-swollen polymer-rich phase. Initiator is partitioned between the phases, and thus radical generation and polymerization occur in both phases. The bimolecular termination rate is diffusion controlled in the polymer-rich phase and hence the concentration of radicals is higher in this phase. Even though the monomer concentration is lower in the polymer-rich phase, the specific polymerization rate R_{p} is higher. In addition, the volume of the polymer-rich phase grows at the expense of the monomer-rich phase with increasing monomer conversion. This explains the autoacceleration in rate as monomer conversion increases.

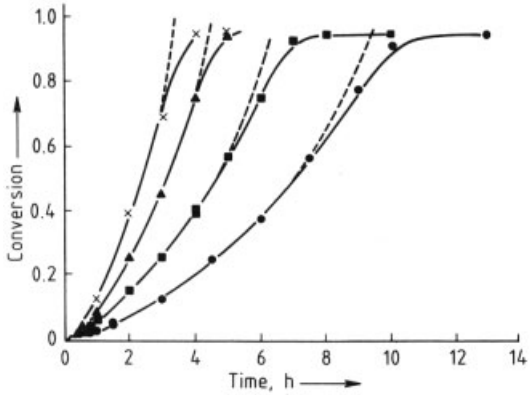

Figure 34. Time – conversion curve for the bulk polymerization of vinyl chloride [459]
Polymerization temperature 50 °C, initiator lauryl peroxide, theoretical curves dashed
[I]: • 0.78×10^{-3} ; ■ 1.57×10^{-3} ; △ 3.38×10^{-3} ; × 5.50×10^{-3} mol/mol vinyl chloride

Talamini has suggested that such a reaction may be described in terms of two homogeneous polymerizations occurring in parallel, with rapid monomer transfer from the monomer-rich to the polymer-rich phase [458], [459]. Assuming that the monomer/polymer mass ratio in the polymer-rich phase is independent of conversion or time, and that the ratio of the specific polymerization rates in the two phases re-

mains constant (Talamini suggests a ratio of ca. 19 at 50 °C), then the observed conversion – time curve can be described well up to ca. 70 % conversion, as shown in Figure 34. However, Talamini's model fails at higher conversions because the monomer-rich phase disappears, and polymerization now occurs only in the remaining polymer-rich phase, with a monomer concentration which decreases with increasing reaction time. Several authors have offered extensions and modifications of the original Talamini model [460–468].

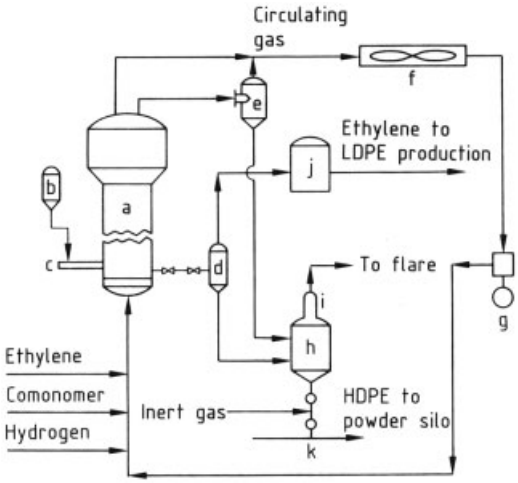

Figure 35. Schematic diagram of the Union Carbide gas-phase process for manufacturing HDPE [474]
a) Fluidized-bed reactor; b) Catalyst transfer tanks; c) Catalyst feeders; d) Product discharge tanks; e) Multiclone dust separator; f) Air coolers; g) Compressor; h) Product degassing tank; i) Filter; j) Ethylene tank; k) Pneumatic transport system

An industrial example of precipitation polymerization is the Pechiney – Saint Gobain two-stage bulk polymerization process [469–472], or the more popular suspension polymerization where poly(vinyl chloride) precipitates in the monomer droplets (→ Poly(Vinyl Chloride)). Precipitation polymerization also encompasses what has been called "gas-phase" polymerization, a process in which polymer particles form within a monomer vapor. Such polymerization does not actually occur in the gas phase, however, because the catalyst resides either within or on the surface of existing polymer particles, and a significant amount of monomer is dissolved in the polymer. The actual site of polymerization is thus within the polymer particle, to which a continuous supply of monomer flows from the gaseous phase. Commercial examples of gas-phase processes include the Union Carbide Unipol process [473–476] and the BASF process [477–481] for low-pressure polymerization of ethylene and propylene (→ Polyolefins, Chap. 1.5.3., → Polyolefins, Chap. 2.5.6.). Figure 35 shows a schematic of the Unipol process in which polyethylene powder is produced using ethylene as a fluidizing gas in a fluidized-bed reactor that contains a modified chromium catalyst. A high-density polyethylene (HDPE) which has a very broad molecular mass distribution and linear chains is produced. Copolymerization of ethylene with propene, 1-butene, and 1-hexene gives products with a controlled amount of short-chain branching and lower polymer density (LLDPE) [482]. The densities of LLDPE are between those of HDPE and LDPE made in the high-pressure free-radical processes.

Polymer – monomer – precipitant systems employ a nonsolvent for the polymer, while the precipitant is miscible with the monomer. The most important example is probably the so-called slurry process for the manufacture of high-density polyethylene (HDPE), isotactic polypropylene, and their copolymers (→ Polyolefins, Chap. 1.5.2., → Polyolefins, Chap. 2.5.1.) Ziegler – Natta (transition-metal catalysts) are used in these processes. Polymerization is often conducted continuously in a train of well-mixed reactors in the presence of a nonsolvent hydrocarbon, usually a C_{6-7} hydrocarbon, at ca. 50 – 100 °C and 5 – 30 bar. Polymer chains grow on suspended, very fine catalyst particles. Soluble, homogeneous catalysts of the Ziegler – Natta type are also known and have gained commerical interest. The productivity of modern catalysts is so high (greater than 100 kg polyethylene per gram transition metal) that subsequent removal of the catalyst from the polymer product is not required. An overview of earlier developments with respect to these catalysts is provided in [483].

Stirred-tank reactors are generally used for this process, often linked in series to give a cascade or train [484], [485]. A simplified schematic of the Hoechst process is shown in Figure 36. Similar slurry processes also exist for polypropylene [486] and copolymers

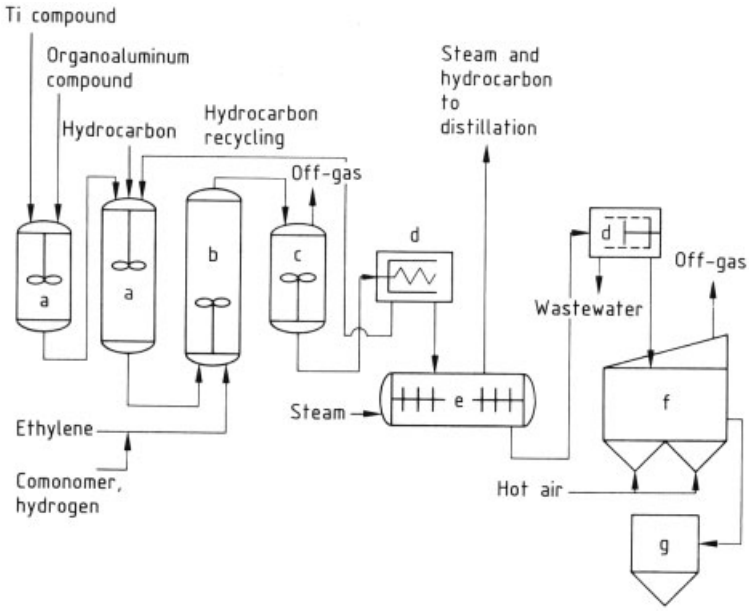

Figure 36. Schematic of the Hoechst slurry-process for HPDE [484]
a) Mixing vessels; b) Reactor; c) Postreactor; d) Centrifuge; e) Stripper; f) Dryer; g) Silo for HDPE powder

based on ethylene – propene (EPM) and ethylene – propene – diene (EPDM) mixtures. Acrylonitrile, sometimes together with a comonomer, can be polymerized in aqueous phase using a redox system such as $K_2S_2O_8$–$Na_2S_2O_5$ or H_2O_2–Fe^{2+} [487]. The polymer precipitates as a fine powder, which is then filtered, washed, and dried. Dissolution of the polymer powder in dimethylformamide produces a spinning solution that can be used to produce polyacrylonitrile fibers. This process has been conducted in a continuous manner [488]. One of the first commercial precipitation polymerizations was the belt conveyor process for polyisobutylene developed by BASF (Fig. 37; see also → Polyolefins, Chap. 4.2.4.). Equal amounts of isobutene and ethylene are first mixed and then deposited in the liquid state on a circulating steel belt. A second inlet is used to add a solution of BF_3 in ethylene. Cationic polymerization occurs very rapidly, and the heat of polymerization is dissipated by the evaporation of ethylene. Thus, ethylene functions here not only as a precipitant but also as a medium for evaporative cooling. The resulting gaseous ethylene is purified, liquefied, and recycled.

A similar process is used by Exxon for the synthesis of butyl rubber [489]. In this case, isoprene serves as the comonomer and methyl chloride takes the place of ethylene. Polymerization is carried out in a continuous stirred-tank reactor. Residual monomer and diluent are removed from the resulting polymer suspension by stripping with hot water.

3.3.3.2.2. Monomer Functioning as Solvent for the Polymer

The Polymer – Monomer – Precipitant System. If the polymer is soluble in its monomer, precipitation polymerization requires the addition of a precipitant that is miscible with the monomer. Polymerization then begins in a solvent – nonsolvent mixture and it ends in a phase consisting only of pure precipitant once conversion of the monomer is complete. Thus, the solubility relationships change during the course of the polymerization, and polymerization may initially occur in a homogeneous manner prior to the onset of precipitation, induced by the enrichment in precipitant that accompanies the consumption of monomer.

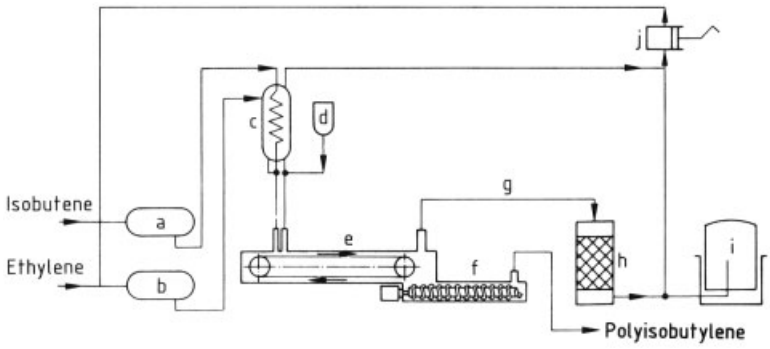

Figure 37. Schematic diagram of the BASF process for manufacturing polyisobutylene [433]
a) Storage vessel for liquid isobutene; b) Storage vessel for liquid ethylene; c) Refrigerating condenser; d) Storage vessel for ethylene containing 0.03 % boron trifluoride; e) Conveyor belt reactor; f) Heated degassing screw; g) Gaseous ethylene; h) Purification with calcium oxide ; i) Gasometer; j) Compressor

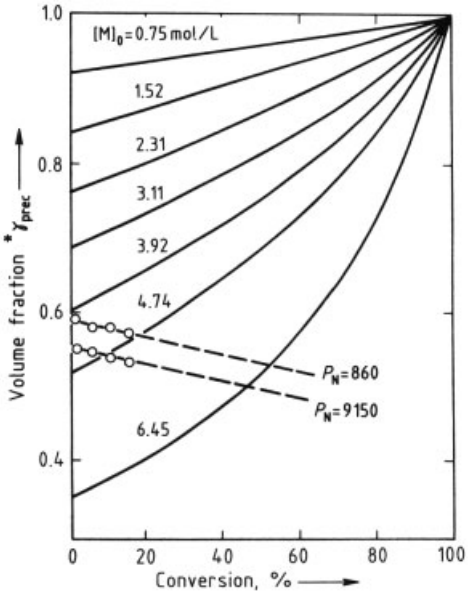

Figure 38. Volume-fraction $*\gamma_{\mathrm{prec}}$ (—-) of methanol as a function of conversion for various initial concentrations of monomer $[\mathrm{M}]_0$ [491]
Azeotropic precipitation copolymerization of styrene – acrylonitrile in methanol (– – –), precipitation point determined using cloud-point titration for two different degrees of polymerization

Figure 38 illustrates these relationships, using as an example the copolymerization (in methanol as precipitant) of 40 mol % acrylonitrile and 60 mol % styrene (at the azeotropic composition). The volume fraction $*\gamma_{\mathrm{prec}}$ of methanol in the reaction mixture, defined as

$$*\gamma_{\mathrm{prec}} = \frac{\text{volume of precipitant}}{\text{volume of precipitant} + \text{volume of solvent}} \tag{3.18}$$

is plotted in Figure 38 as a function of conversion for various initial concentrations of monomer $[\mathrm{M}]_0$. The dashed lines represent extrapolations from experimental cloud-point titration data, which give the $*\gamma_{\mathrm{prec}}$ values at which the first polymer fractions precipitate at 65 °C. For an initial concentration $[\mathrm{M}]_0 < 4.1$ mol/L, extremely small conversions suffice to cause precipitation. At high monomer concentrations the polymer initially remains in solution because the volume fraction of methanol required for polymer precipitation has not yet been reached. In such cases, the critical value of $*\gamma_{\mathrm{prec}}$ is exceeded only at higher conversion: the greater $[\mathrm{M}]_0$, the higher the conversion required.

In precipitation polymerization of this type it is especially important to ensure that the polymer precipitates in finely divided form. Among other things this ensures that the particles do not overheat during polymerization. The process has been employed on large scale by BASF [490–492], with batch polymerization occurring in stirred reactors under reflux at 65 °C. Azobisisobutyronitrile was employed as initator, and poly(vinyl pyrrolidone) or poly(vinyl ether) as a protective colloid. The resulting styrene – acrylonitrile copolymer was separated by centrifugation and the recovered methanol was distilled and recycled to the next batch.

Polymer dispersions in nonaqueous media with particle size in the range 0.01 – 10 µm (nonaqueous dispersions, colloids, organosols) have been thoroughly investigated by BARRET and coworkers [493]. BARRET characterizes "dispersion polymerization" as a process in which an insoluble – and thus dispersed – polymer is prepared from a monomer dissolved in an organic diluent to which has been added an amphiphatic block or graft copolymer to serve as a dispersant. However, here this is regarded as a special case of precipitation polymerization in which complete coagulation of polymer particles is prevented and the particle size is controlled.

The key to controlling particle size is the selection of dispersant type. Among the most effective dispersing agents are the so-called amphipathic molecules: block and graft copolymers made up of two polymeric components, only one of which is insoluble in the continuous, diluent-containing phase. Graft copolymers of this type often form during the polymerization as a result of grafting on the dissolved polymer, but it is not absolutely essential that the insoluble portion of the dispersant be identical to or soluble in the disperse phase. In many cases its insolubility in the diluent is sufficient to ensure adequate adsorption on the particle surface. These amphipathic dispersing agents act as steric stabilizers [494].

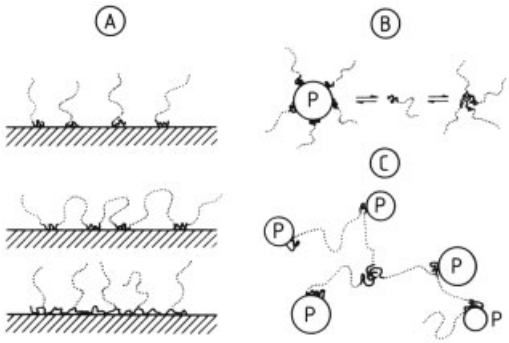

Figure 39. Steric stabilization of precipitating polymer/monomer particles with the aid of amphipathic block and graft copolymers [495]
A) Schematic representation of adsorption of amphipathic molecules on the polymer particles; —- insoluble groups, ····· soluble groups; B) Equilibrium established in the course of a precipitation polymerization P = (growing) polymer particle; C) Schematic representation of flocculation due to multifunctional amphipathic molecules

Figure 39 A is a schematic representation of the adsorption of di-block, multiple-block, and graft copolymers on the surface of growing polymer particles. Soluble and insoluble portions of the dispersant molecules must be kept carefully in balance. If the insoluble part is too small, or if it interacts too weakly with polymer particles, then adequate adsorption will occur only when the dispersant concentration in the continuous phase is very high. If the soluble portion is too large, the dispersant will be present largely as aggregates or micelles with little tendency to dissociate and be adsorbed on the interface. Figure 39 B depicts the equilibrium situation. Finally, it is important to note that multifunctional amphipathic molecules like those shown in Figure 39 C can also function as weak flocculating agents.

In a well-stabilized dispersion, each particle is covered by a layer of freely moving polymer chains, which are in turn dissolved in the continuous phase. These layers prevent the frequently colliding particles from approaching so closely that Van der Waal's attractive forces become dominant. A simplified model (Fig. 40) suggests that the mechanism of steric stabilization involves an increase in the local concentration of polymer chains or segments as a result of overlapping and mutual chain interpenetration [496] as two polymer particles approach. This induces an osmotic pressure and increase in the free energy ΔG_R. To compensate for this effect, solvent flows into the regions of higher polymer concentration and drives the particles apart. A positive value of $\Delta G_R = \Delta H_R - T\,\Delta S_R$ can be due either to enthalpic effects (ΔH_R) or to entropic effects ($T\,\Delta S_R$). It is therefore possible to divide the contributions to steric stabilization into three categories [498]:

1) Enthalpic stabilization
   ΔH_R and ΔS_R are both positive
   $\Delta H_R > T\,\Delta S_R$
   The dispersion flocculates on warming
2) Entropic stabilization
   ΔH_R and ΔS_R are both negative
   $T\,\Delta S_R > \Delta H_R$
   The dispersion flocculates on cooling
3) Combined enthalpic – entropic stabilization
   ΔH_R is positive
   ΔS_R is negative

The dispersion is stable over a wide range of temperature.

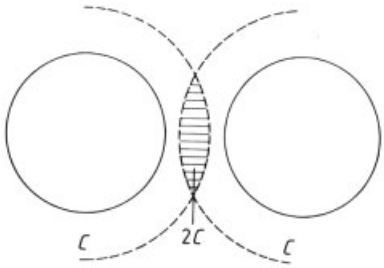

Figure 40. Model for steric repulsion caused by the overlap of two spheres containing dissolved molecular chains [496], [497]
C = Concentration of the polymer chains in the adsorption layers

More complete discussions of steric stabilization may be found elsewhere [495], [498–503]. A survey of precipitation polymerization is available [504].

3.3.3.3. Suspension Polymerization

Definition. The term "suspension polymerization" includes a series of processes, all of which involve emulsifying monomers to droplets by stirring them in a suspending medium in which they are insoluble in the presence of a free-radical initiator, usually one that is soluble in the monomer. When the polymer formed is soluble in the monomer, nonporous spherical "beads" are formed, hence the term "suspension bead polymerization". If, however, the polymer precipitates during polymerization, the resulting polymer particles are composed of many smaller primary particles. They are opaque, usually possess an irregular surface, and may have substantial internal porosity. This type of polymerization has been called "suspension powder polymerization". The dispersants (protective colloids) are either macromolecules that are insoluble in the suspending medium or insoluble, usually inorganic, powders, the so-called Pickering emulsifiers [505]. Their function is first to assist in the formation of the initial monomer emulsion and then to stabilize the resulting polymer particle suspension. The following recipes are examples for suspension polymerization involving protective colloid and inorganic powder dispersants.

With protective colloid:	
Temperature	80 °C
Reaction time	8 h
Styrene	100 parts (by mass)
Water	200 parts (by mass)
Benzoyl peroxide	0.4 parts (by mass)
Poly(vinyl alcohol) [partially hydrolyzed poly(vinyl acetate)]	0.5 parts (by mass)

With pickering emulsifier:	
Temperature	80 °C
Reaction time	10 h
Styrene	100 parts (by mass)
Water	200 parts (by mass)
Benzoyl peroxide	0.2 parts (by mass)
Barium sulfate	1.0 part (by mass)
$C_{18}SO_3Na$	0.002 parts (by mass)

The term "suspension polymerization" is perhaps inappropriate, because precipitation and emulsion polymerizations also produce suspensions of polymer particles in a continuous phase. The distinction from precipitation polymerization is that it is initiated in a homogeneous mixture, while suspension polymerization takes place in an emulsion. The beads or powder particles produced in a suspension polymerization are roughly of the same size as the original monomer droplets with diameters on the order of 10^{-3} to 0.5 cm. Emulsion polymerization also starts with a monomer emulsion, but the initiator is usually one that is soluble in the continuous suspending phase rather than in the monomer. Moreover, the resulting latex particles are very much smaller (diameter range 5×10^{-6} to 3×10^{-5} cm or 0.05 to 0.3 μm) than the original monomer droplets. Borderline cases which might be called emulsion or suspension polymerization are discussed later in Section 3.3.3.4.

In the vast majority of cases, the suspending medium for suspension polymerization is water, although inverse-suspension polymerizations are also known and used commercially to produce very high molecular weight polymers and copolymers based on the comonomer acrylamide. Here a water-soluble monomer is dispersed in a hydrophobic organic suspending medium, usually in the presence of water in the disperse phase.

3.3.3.3.1. Qualitative Description

Generally, dispersants are employed at a concentration (relative to the aqueous phase) of 0.1 – 5 wt % in the case of protective colloids and 0.1 – 2 wt % for Pickering emulsifiers. A typical initiator concentration is 0.1 – 1 wt % relative to the monomer. The volume ratio monomer/aqueous phase is usually between 25 : 75 and 50 : 50, and the stereometric limit, which cannot be exceeded with spheres of uniform size, is 74 : 26.

The reactor vessel is usually a stirred tank. The monomer is subjected either to turbulent pressure fluctuations or viscous shear forces, which break it into small droplets that assume a spherical shape under the influence of interfacial tension. These droplets undergo constant collision (collision rate $\geq 1\ s^{-1}$), with some of the collisions resulting in coalescence. In the absence of stabilizers, a dynamic equilibrium is eventually established, leading to a stationary mean particle size. Individual drops do not retain their unique identity, but instead undergo continuous breakup and coalescence. This phenomenon can easily be demonstrated by the addition of a small amount of a monomer that has been labeled with a water-insoluble dye. The dye is rapidly distributed uniformly in the disperse monomer phase.

In some cases, an appropriate dispersant can be used to induce the formation of a protective film on the droplet surface. As a result, pairs or clusters of drops that tend to coalesce are broken up by action of the stirrer before the critical coalescence period elapses. A stable state is ultimately reached in which individual drops maintain their identities over prolonged periods of time. In this case, addition of dye-bearing monomer does not result in migration of the dye into other droplets. Such a system is described as a turbulence-stabilized emulsion.

In the simple case of a polymer that is miscible in all proportions with its monomer (e.g., styrene and methyl methacrylate), various viscosity states of the disperse phase are traversed during the course of polymerization. The initially nonviscous, liquid monomer is transformed gradually into an increasingly viscous solution of polymer in monomer, and as conversion proceeds the disperse phase acquires the characteristics of a solid polymer. Particularly in the tacky intermediate stage, individual polymer particles tend to form incompletely fused clumps. Coagulation at this critical stage of conversion is somewhat inhibited by the action of the dispersant, but other effective measures to reduce coagulation may also be taken, including adjusting the densities of the two phases to make them more similar, or by increasing the viscosity of the aqueous continuous phase. Rapid polymerization during the sticky stage minimizes the number of collisions among polymer particles and thus should reduce coagulation.

An experiment with dye-labeled monomer is also applicable when polymerization occurs in the disperse phase [506–508]. In this case dye-containing polymer beads generated in a parallel polymerization are added to the suspension during polymerization. It is observed that beyond a certain conversion (which depends on the reaction conditions) coagulation and particle breakup cease entirely. This is known as the "particle identity point" or the limit of dynamic equilibrium. Increasing the dispersant concentration (i.e., reducing the size of particles) displaces the identity point towards lower conversion [509], [510].

The schematic diagram in Figure 41 shows these fundamental relationships. The upper part of the diagram corresponds to dynamic equilibrium in a monomer emulsion that contains no dispersant, or to an emulsion undergoing polymerization that has not yet reached the identity point. The monomer phase is broken up into long strands by the stirrer, and these in turn fragment into spherical droplets due to interfacial tension forces. The droplets may then agglomerate into larger aggregates, finally coalescing into larger drops, which either break apart again due to stirring or collect as an extended monomer phase. The lower portion of Figure 41 shows the various stages that lie between the identity point and the end of the polymerization process. In a turbulence-stabilized emulsion, one in which monomer drops maintain their identity, the probability is high that individual drops will polymerize directly to primary beads. Nevertheless, even in a suspension bead polymerization, irregular particles sometimes appear, composed of many individual smaller polymer particles, a consequence of some clusters surviving past the identity point.

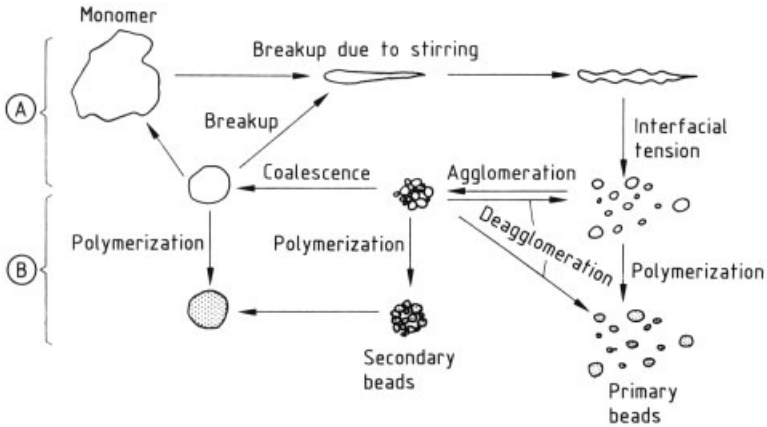

Figure 41. Schematic diagram of dispersion and polymerization in a suspension bead polymerization
A) Monomer emulsion in the absence of dispersant, or a polymerization mixture that has not yet reached the identity point; B) Polymerizing mixture after passing the identity point

A polymerizing bead with a diameter $d = 10^{-2}$ cm contains ca. 10^8 growing polymeric radicals. The effects of subdivision due to particle breakup in suspension polymerization are not anticipated to occur at the levels found in emulsion polymerization (polymer particles are submicron in size) where radical contents of 0.5 or lower per particle occur [511], [512]. Each bead may be regarded as a small, isolated reactor. For this reason, the observed polymerization kinetics correspond directly to those for bulk polymerization [506], [513], [514]. Here again the Trommsdorff – Norrish and glass effects must be taken into account, as must the effects associated with demixing when the polymer is insoluble in its monomer [459], [460], [462], [515]. Indeed, the process is sometimes referred to as a water-cooled bulk polymerization. To ensure that conversion is as complete as possible, it is common to employ mixtures of initiator types with different half-lifes, and to allow the polymerization temperature to increase in the final stages of conversion [516], [517]. With many industrially important suspension polymerizations, simply preparing a polymeric material is not sufficient. Often a particular particle-size distribution and morphology must be achieved, as in the manufacture of expandable polystyrene, or poly(vinyl chloride), whose particles must later absorb substantial quantities of plasticizer for some applications.

For what has already been said about the course of suspension polymerization, together with various results from the literature [507], [510], [513], [518], [519], it is possible to establish a number of special factors – apart from those common to all free-radical polymerizations – that exert an important influence on particle size and size distribution:

1) Geometric factors of the reactor: profile, type of stirrer, stirrer diameter D relative to the reactor dimensions, bottom clearance of the stirrer, and internal fittings
2) Operating parameters: stirrer velocity N, stirring and polymerization time, phase volume ratio φ, fill level of reactor, and temperature T
3) Substance parameters: dynamic viscosities η_c and η_d and densities ϱ_c and ϱ_d of the continuous and disperse phases, and interfacial the tension σ.

During monomer fragmentation, in collisions leading to agglomeration, and in deagglomeration of clusters, the most important consideration is the energy introduced into the reaction mixture per unit time. This can be expressed in terms of the mean rate of energy dissipation per unit mass $\bar{\varepsilon}$. In the case of a stirred reactor containing baffles and operated with a high Reynolds number, the following equation is applicable [520]:

$$\bar{\varepsilon} = KN^3D^2 \left[\mathrm{cm^2\,s^{-3}\ or\ W\,kg^{-1}}\right] \tag{3.19}$$

where K is a dimensionless constant that depends on the type of stirrer. Increasing the stirrer speed N causes a decrease in particle size.

The concentration of dispersant [S] determines the maximum particle surface area that can be stabilized and influences the interfacial tension. Increasing [S] leads to a decrease in particle diameter, as does lowering the interfacial tension. Increased disperse phase viscosity reduces particle breakup and thus leads to larger particles.

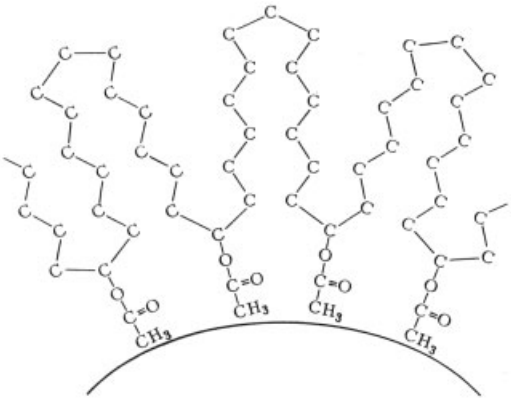

Figure 42. Schematic representation of the adsorption of partially hydrolyzed poly(vinyl alcohol) on the surface of a dispersed particle [510]
All − OH groups (omitted for clarity) are directed toward the aqueous phase, and residual acetate groups toward the disperse phase

3.3.3.3.2. Dispersants

Protective Colloid Dispersants. Organic protective colloids include natural products such as alginates, tragacanth, agar, and starch as well as modified natural polymers such as carboxymethylcellulose (sodium salt), hydroxyethylcellulose, and methylcellulose. Among the effective synthetic polymers are styrene – maleic anhydride copolymer, poly(methacrylic acid), poly(vinyl pyrrolidone), poly(vinyl alcohol), and partially hydrolyzed poly(vinyl acetate). The important feature of all these materials is their amphipathic character, which explains their ability to lower the interfacial tension and to concentrate at the monomer – water interface.

As shown in Figure 42, protective colloids are probably adsorbed in such a way that they form loops near the particle surface. In the case of partially hydrolyzed poly(vinyl alcohol), the residual acetate groups act as hydrophobic points of attachment, while the OH groups are directed into the continuous aqueous phase. The most important factor in dispersant effectiveness is an appropriate balance between hydrophilic and hydrophobic groups [521]; molecular mass is of considerably less significance. The wettability of a polymer can be varied by the addition of trace amounts of a low molecular mass surfactant.

Higher concentrations of protective colloid may cause a bead polymer to be contaminated by a small amount of a much more finely divided emulsion polymer. This should be avoided because it results in polymer loss during workup. The chance of encountering such a problem is enhanced by increased solubility of the monomer and initiator in the continuous phase. However, emulsion polymerization via homogeneous nucleation in the continuous phase can be suppressed by addition of a water-soluble inhibitor such as NH_4SCN or a copper salt [513], [522].

Powdered Dispersants (Pickering Emulsifiers). Finely divided, usually inorganic, insoluble solids may also be employed as dispersants in suspension polymerization [505], [513], [523]. Common choices include barium sulfate, talc, aluminum hydroxide, hydroxyapatite, tricalcium phosphate, calcium oxalate, magnesium carbonate, and calcium carbonate. It is advantageous to dissolve out the dispersant after the polymerization (e.g., in dilute acid), thereby minimizing polymer contamination. Monomer emulsions based on systems of this type are referred to as three-phase emulsions, a term introduced by Wenning [507]. In the case of a polymer that is insoluble in its monomer, a fourth phase occurs during the course of polymerization.

Solid dispersants must be wet by two immiscible liquids, and they must also exhibit a certain degree of self-adhesion. The wettability can be modified by adsorption of low molecular mass surfactants. This technique is referred to as modulation and it has much in common with techniques used in mineral flotation processes.

The wettability of a solid (S) by a liquid such as water (W) in the presence of a gas (G) depends upon the wetting angle α, which results

from equilibrium between three interfacial tensions:

$$\sigma_{SW} + \sigma_{WG} \cdot \cos\alpha = \sigma_{SG} \tag{3.20}$$

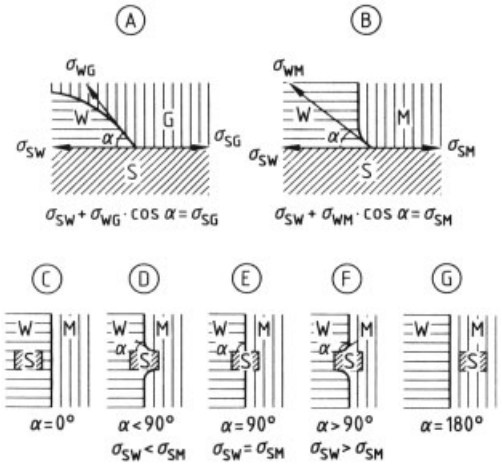

Figure 43. Wetting of a Pickering dispersant by monomer and water, modulated by addition of a surfactant
S solid; G gas; M monomer; W water
For further explanation, see text

Figure 43 A illustrates the significance of this relationship. An angle $\alpha = 0°$ corresponds to total wetting. For $\alpha = 90°$, $\sigma_{SW} = \sigma_{SG}$, and at $\alpha = 180°$ no wetting occurs. In a three-phase emulsion, two immiscible liquids compete for wetting the solid S: monomer M and water W. As indicated in Figure 43 B, there is again an equilibrium condition:

$$\sigma_{SW} + \sigma_{WM} \cdot \cos\alpha = \sigma_{SM} \tag{3.21}$$

Taking the example of barium sulfate as the dispersant in a styrene – water bead polymerization, the barium sulfate is at first wetted more effectively by water than by styrene, so $\sigma_{SW} < \sigma_{SM}$. Water wets the barium sulfate with a small contact angle α and $\cos\alpha$ is positive (Fig. 43 D). Addition of a surfactant whose polar groups are adsorbed on the barium sulfate makes the surface more hydrophobic (or lypophilic). As a result, σ_{SW} increases and σ_{SM} decreases, leading to the states shown in Figure 43 E and F. In the limiting case (Fig. 43 G), the particle is fully immersed in the monomer phase and loses its effectiveness as a dispersant. In the course of passing from D to F it is not uncommon for a phase reversal to occur: the oil-in-water emulsion is transformed into a water-in-oil emulsion and the polymer forms a block that contains water-filled voids [507]. It is clear that the limiting cases C and G are inappropriate for suspension polymerization. A good Pickering dispersant must possess amphiphatic characteristics, thus leading to one of the cases D – F. Many pure inorganic substances are completely wetted by water, the chief reason why modulation is so important in practice as a means of inducing the transition from C to D.

3.3.3.3.3. Mechanism of Particle Formation

Suspension polymerization always begins with dispersion of a monomer in water. It thus seems reasonable to first consider the formation of a monomer emulsion – or more generally, the formation of an emulsion of an organic liquid in water – and then examine the effects of polymerization in the disperse phase.

Droplet Size in an Emulsion Subject to Turbulent Mixing. Especially simple relationships should govern the suspension polymerization of a turbulence-stabilized emulsion in which single drops maintain their separate identities. However, such an emulsion must fulfill a series of conditions [524–526]:

1) Stirring must be sufficiently intensive to separate any droplet pairs or clusters that begin to form. The adhesive forces between two drops increase roughly linearly with the drop diameter d, while the forces exerted on the drops by stirring show a higher order dependence on drop diameter. The probability of separation therefore increases with diameter; given a particular stirring intensity, the diameter must remain above a specific minimum d_{min}:

$$d_{min} = C_1' \varrho_c^{-3/8} A(h)^{-3/8} \bar{\varepsilon}^{-1/4} \tag{3.22}$$

where C_1' is an empirical constant, ϱ_c is the density of the continuous phase and $A(h)$ is the energy required to separate to a distance $h = \infty$ two drops of diameter $d = 1$ initially separated by a distance h_0. The value of $A(h)$ is strongly dependent upon the thickness and characteristics of the adsorbed protective film. Equation (3.19) may be used

to convert Equation (3.22) to the following form:

$$d_{\min} = C_1' \varrho_c^{-3/8} A(h)^{-3/8} N^{-3/4} D^{-1/2} \quad (3.23)$$

If $d < d_{\min}$, drops will coalesce until a diameter $d_{\min}$ has been reached. If energy dissipation in the stirred reactor is not uniform (e.g., if only near the stirrer blades is it sufficient to break up drop clusters), then material circulation within the vessel must be sufficiently rapid to ensure that each cluster approaches the stirrer blades at least once during the critical coalescence time.

2) The dissipated stirring energy must not be so large as to cause significant breakup of individual drops, that is, the critical Weber number We (crit) or a maximum drop diameter $d_{\max}$ must not be exceeded:

$$We\,(\text{crit}) = C_2' d_{\max}^{5/3} \varrho_c \sigma^{-1} \bar{\varepsilon}^{2/3} \quad (3.24)$$

which is transformed by using Equation (3.19) into:

$$d_{\max} = C_2 \varrho_c^{-3/5} \sigma^{3/5} N^{-6/5} D^{-4/5} \quad (3.25)$$

where C_2' and C_2 are empirical constants and σ is the interfacial tension. Equation (3.25) is only applicable if the viscosity ratio η_d/η_c is not too large. Otherwise, the stable drop diameter increases with increasing values of η_d/η_c, because turbulent oscillations in the disperse phase are damped (viscosity stabilization of drop size).

3) A density difference between the phases results in a tendency towards settling, and this increases with increasing drop size. Stirring energy must be sufficient to counteract this tendency. The maximum drop diameter $d_{\max}$ that can be sustained in an emulsion at a particular rate of energy dissipation $\bar{\varepsilon}$ is given by:

$$d_{\max} = C_3' \left(\frac{\varrho_c}{\varrho_d - \varrho_c} \right)^3 \frac{1}{g^3} f(\varphi) \bar{\varepsilon}^2 \quad (3.26)$$

and using Equation (3.19) one obtains

$$d_{\max} = C_3 \left(\frac{\varrho_c}{\varrho_d - \varrho_c} \right)^3 \frac{1}{g^3} f(\varphi) N^6 D^4 \quad (3.27)$$

where C_3' and C_3 are empirical constants, g is the acceleration due to gravity, φ is the volume ratio (disperse phase volume/continuous phase volume), and $f(\varphi)$ is an empirical function.

According to [524–526], turbulence stabilization is only possible if the drop diameter is larger than $d_{\min}$ as given by Equation (3.23), and smaller than both $d_{\max}$ from Equation (3.25) and $d_{\max}$ from Equation (3.27). It is reasonable to assume that the Sauter mean diameter given by

$$d_{32} = \frac{\Sigma \eta_i d_i^3}{\Sigma \eta_i d_i^2} \quad (3.28)$$

is proportional to the maximum diameter $d_{\max}$. Figure 44 illustrates these relationships. Stability is achieved only in the shaded region between the lines defined by Equations (3.23), (3.25), and (3.27). If $\bar{\varepsilon} < \bar{\varepsilon}_{\min}$, then $d_{\min}$ (coalescence) $> d_{\max}$ (demixing) and the emulsion will separate into two phases. On the other hand, if $\bar{\varepsilon} > \bar{\varepsilon}_{\max}$, then $d_{\min}$ (coalescence) $> d_{\max}$ (breakup) and the emulsion will be unstable because the energy necessary for separating a droplet pair is sufficient to breakup the droplets themselves.

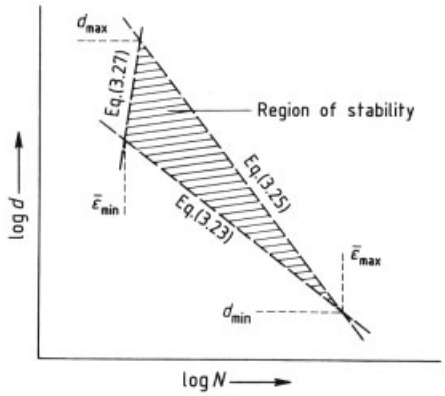

Figure 44. Theoretical relationship governing the fragmentation (Eqs. 3.22 and 3.23), coalescence (Eqs. 3.24 and 3.25), and demixing (Eqs. 3.26 and 3.27) of droplets in a stirred reactor. Only in the shaded region can a turbulence-stabilized emulsion form [525]

A prerequisite for the validity of Equations (3.22) – (3.27) is the applicability of Kolmogoroff's theory of local isotropy [527–529]. The characteristic length L of the energy-containing large eddies ($L \approx 0.08\,D$, according to

[244]) must be larger, and the characteristic length l of the energy-dissipating small eddies [$l = (\gamma_c^3/\varepsilon)^{1/4}$], where γ_c is the kinematic viscosity of the continuous phase] must be much smaller than the drop diameter d (i.e., $L \gg d \gg l$). Equations (3.24) and (3.25) [531–534] as well as (3.22) and (3.23) [525], [526], [535] have been verified experimentally. Nevertheless, they apply only for very small phase ratios $\varphi < 0.015$, which are unrealistic for industrial suspension polymerizations. The polymerizing system is also subject to nonnegligible effects of the viscosity ratio η_d/η_c. In [536], for example, this has been taken into account by introduction of the viscosity group $\eta_d(\varrho_d\,\sigma\,d)^{-1/2}$, which depends only on the physical properties of the disperse phase. At higher phase ratios, both d_{32} and d_{max} increase with increasing distance from the stirrer [534], [537], [538]. Various corrections [531], [534], [539], [540] have been suggested for Equation (3.25), all of which take the form:

$$\frac{d_{32}}{D} = A\,(1+B\varphi)\,We^{-3/5} \tag{3.29}$$

where A and B are numerical constants. For phase relationships such as those applicable on an industrial scale it is therefore not possible to assume a homogeneous energy-dissipation rate ε; instead, it must be anticipated that ε is large in the vicinity of the stirrer, becoming smaller in more remote circulation zones. A simulation program has been described [541–543] that approaches this problem by dividing the reactor into various stirring and circulation zones and then applying Monte Carlo methods.

The Size of the Polymer Particles. Conversion of a monomer into polymer increases the viscosity η_d and density ϱ_d of the disperse phase, and lowers the volume phase ratio φ. The interfacial tension σ also changes, and grafting reactions between polymeric radicals on the organic protective colloid and the monomer may have to be accounted for. As has been demonstrated by Hopff et al. [544–547] in the case of a suspension bead polymerization of methyl methacrylate with a relatively high concentration of a protective colloid [poly(vinyl alcohol), Mowiol N 70 (88)], systems exist in which the final particle size is already established at very low conversion. Polymerizing droplets maintain their identity, and subsequent reduction in stirring speed has no influence on particle size. Apparently, a turbulence-stabilized state is reached at the very onset of polymerization. Evaluation with respect to particle size is particularly straightforward in this case, since changes in the above-mentioned parameters during polymerization are irrelevant.

A series of experiments in geometrically similar reactors (see Fig. 45) coupled with dimensional analysis, permitted derivation of the following dimensionless equation in the case of this especially simple system:

$$\frac{d_{50}}{D} = kRe^{0.5}We^{-0.9}Fr^{-0.1}\left(\frac{\eta_d}{\eta_c}\right)^{0.1}\varphi^0 \tag{3.30}$$

where d_{50} is the bead diameter below which 50 wt % of the particles pass through the sieve (in other words 50 wt % of the particles have a diameter less than d_{50}), D is the stirrer diameter, which throughout the experiments was kept in a constant ratio to the reactor diameter, and k is a numerical constant.

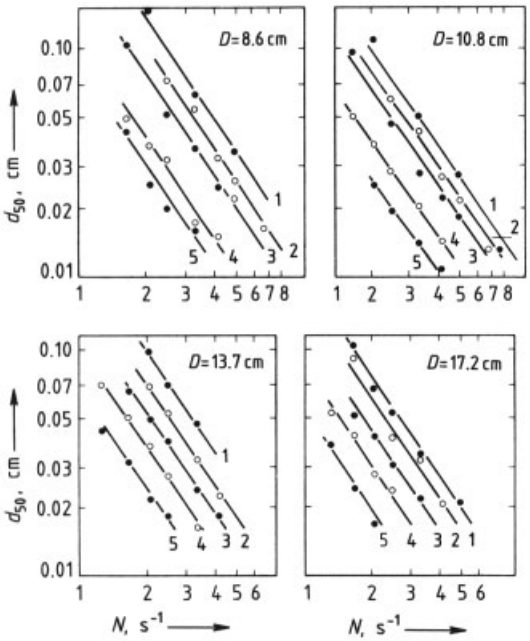

Figure 45. Dependence of the bead diameter d_{50} on the stirrer velocity N [545]
Experiments conducted in four geometrically similar stirred reactors with stirrer diameter D
Curve parameter: protective colloid concentration in g/100 cm^3

The dimensionless groups are:

$$Re = \left(\frac{D^2 N \varrho_c}{\eta_c}\right) = \text{ratio of inertial/viscous forces}$$

$$Fr = \left(\frac{DN^2}{g}\right) = \text{ratio of inertial/gravitational forces} \quad (3.30a)$$

$$We = \left(\frac{D^3 N^2 \varrho_c}{\sigma}\right) = \text{ratio of inertial/interfacial forces}$$

It follows that $d_{50} = kN^{-1.5}$

A relationship similar to Equation (3.30) has been found for the suspension polymerization of vinyl chloride [548], [549]. In this case, however, a large particle diameter region was identified with $d = kN^{-1.9}$ at low stirrer speeds (with correspondingly large particle diameters) as well as a fine-particle size region with $d = kN^{-0.6}$. An analogous effect has also been reported for styrene [550].

In contrast to the very early establishment of particle diameter in the experiments of HOPFF et al., dye-labeling experiments have shown that in other systems [506–508], [550], [551] the particles reach the so-called identity point only at high conversion. With these systems it is apparently necessary to take into account the time and conversion dependence of the parameters enumerated above. Another important observation [550], [552] is the fact that in certain systems the stirring time in which the unpolymerized emulsion reaches its final droplet diameter is longer than that taken by the polymerizing system to reach 50 % conversion. A faster polymerization would therefore give larger polymer particles due to viscosity stabilization.

The problems associated with scaling up from small to industrial scale, particularly for particle size, are an important concern in suspension polymerization. In principle, Equation (3.30) should be applicable, suggesting that all that is required for maintaining constant d/D when scaling up is to keep Re, We, Fr, and η_d/η_c constant. However, Equations (3.30 a) show that this is not possible without changing the composition of the system. Furthermore, it is d, not d/D, that must be kept constant in an industrial scaleup. To what extent it is possible at high Reynolds numbers to avoid an exact similarity transformation and scale the process up on the basis, for example, of [553], [554]

$$\frac{d}{D} = kW e^{-0.6} Eu^{-1.0} \quad (3.31)$$

or

$$d \sim D^{-0.8} N^{-1.2} \quad (3.32)$$

has not yet been established experimentally (Eu is here a modified power number in power per unit volume). A not further defined suspension polymerization of methyl methacrylate in 1 % polyacrylamide solution was transformed with constant particle size on the basis of

$$N_1 = N_2 \left(\frac{D_2}{D_1}\right)^{2/3} \quad (3.33)$$

from a laboratory reactor (1) with $D = 10.1$ cm to a half-tonne scale (2) with $D = 81.2$ cm [555].

3.3.3.3.4. Industrial Applications

Suspension polymerization is used for producing a wide variety of polymer types, the most important of which are mentioned briefly. Standard polystyrene for use in injection molding is manufactured by suspension bead polymerization, as is poly(methyl methacrylate) and its copolymers containing small amounts of acrylate esters. Clear transparent polymers are often required, so formulations involving Pickering dispersant (e.g., $MgCO_3$) that can be dissolved out of the polymer with dilute acid after polymerization are particularly advantageous.

In the case of styrene – acrylonitrile copolymers, the method of choice for batch suspension polymerization is normally that involving the azeotropic composition to minimize compositional drift. Nevertheless, complications often arise, because considerably more acrylonitrile than styrene dissolves in the aqueous continuous phase. As conversion proceeds, acrylonitrile diffuses into the polymer particles and the monomer ratio in the beads changes, causing the composition of the copolymer to change as well [556]. The same phenomenon can accompany copolymerization in heterogeneous systems generally [557].

High-impact polystyrene and ABS are often prepared in a combined process. This begins with a solution of polybutadiene in styrene or

styrene/acrylonitrile and permits bulk polymerization to occur under stirring until phase reversal or inversion has occurred. Water and dispersant are then added and the polymerization is completed in suspension [558], [559].

Expandable polystyrene is prepared by suspension polymerization in the presence of a blowing agent, such as pentane [560], [561]. It is also possible to introduce the blowing agent in a second step after polymerization, allowing it to diffuse into the beads [562], [563]. Warming to 80 – 110 °C, generally with steam, causes the beads to expand by foaming, their volume increasing by a factor of ca. 30 – 50. Particle size and size distributions both play important roles. For example, if thin-walled objects are to be made from expandable polystyrene, especially small polymer particles are required. The unfoamed product is fractionated by sieving prior to drying. Figure 46 shows a schematic of a typical EPS production facility.

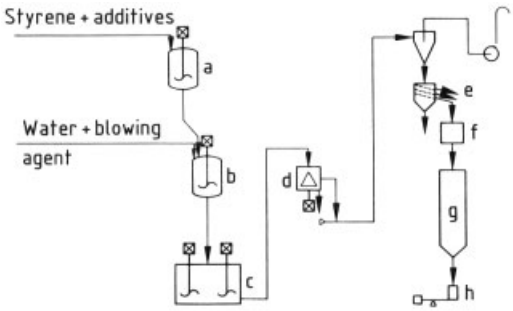

Figure 46. Schematic representation of the manufacture of Styropor by batch suspension polymerization
a) Mixing tank; b) Stirred reactor; c) Puffer tank; d) Centrifuge; e) Sieving; f) Drying; g) Silo; h) Packaging

The required amount of deionized water is loaded into the reactor at ambient temperature and agitation is started. Styrene and the initiator pair, AIBN or benzoyl peroxide plus a finishing initiator such as di-*tert*-butyl peroxide or *tert*-butyl peroxybenzoate are pumped into the reactor at a constant rate, while the stabilizer, either an inorganic, insoluble, finely divided powder or a polymeric steric stabilizer is added. The reactor is closed and the heating cycle starts. Droplet size development occurs during the heating cycle. When the polymerization temperature is reached, between 75 °C and 95 °C depending on the initiator type, the polymerization proceeds until a conversion in the range 32 – 35 % is obtained. This point marks the beginning of the so-called "sticky-stage" at which point the coalescence rate begins to increase, causing the beads to grow from a mean particle diameter of about 0.2 mm to the desired size, depending on the original stabilizer formulation, for the final application of the resin product. During this stage of the process, particle size is monitored by periodic sampling of particle size growth rate. Additional stabilizer may be added if the growth rate is too large. At 65 – 68 % conversion of monomer, the identity point is reached. At this point the particle viscosity is sufficiently large that collisions between particles are elastic and particle size growth ceases. In addition, the density of dispersed and continuous phases are almost identical, and the suspension is very stable. Autoacceleration of the polymerization rate becomes appreciable at the identity point and increases until ca. 95 % conversion where a glassy-state transition occurs and the beads become hard. Once the beads are hard the reaction mixture is heated to a temperature above the glass transition temperature of the polystyrene ($T_g \approx 100$ °C). During heating, the reactor is pressurized with a blowing agent (5 – 8 % with respect to polymer), a low-boiling hydrocarbon ($C_4 - C_7$). The reactor is then pressurized with nitrogen at 700 – 950 kPa and the so-called impregnation stage starts and proceeds for 3 – 8 h. During impregnation the blowing agent diffuses into the beads. At the same time, the free volume increases, and the finishing initiator rapidly generates radicals, causing a relatively rapid increase to a monomer conversion of ca. 99.9 %. The impregnation time should be sufficient to allow the blowing agent to reach the core of the particle and to cause the breakup of hydrocarbon domains within the polymer matrix, giving a uniform distribution. At the end of the impregnation stage the suspension is cooled to 20 – 30 °C, depending on the blowing agent type, to freeze in the ingredients and prevent bead expansion during handling. Some typical properties of an EPS grade are:

Monomer conversion	> 99.9 %
$\bar{M}_W \times 10^{-3}$	200 – 300
Polydispersity	2.2 – 2.4
Mean particle diameter $\bar{d}$	0.4 – 1.5 mm
PSD breadth $\sigma/\bar{d}$	0.20 – 0.25
Blowing agent concentration	5.0 – 8.0 %

Table 3 lists required properties of the bead for different applications.

Table 3. EPS bead properties and applications

$\bar{d}$, mm	$\sigma/\bar{d}$	Density, kg/m^3	Use
1.5	0.20	8 – 16	construction (insulation)
1.0	0.15	12 – 20	packaging
0.4	0.10	50 – 70	coffee cups

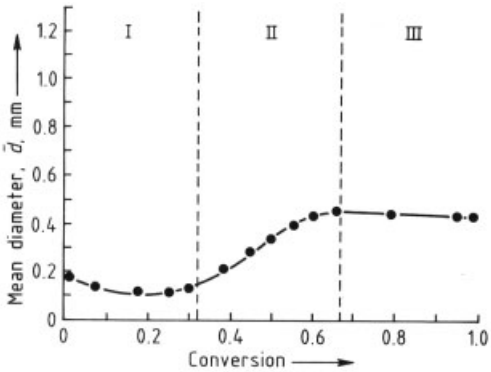

Figure 47. Growth path for EPS beads (mean diameter $\bar{d}$ versus monomer conversion) during suspension polymerization of styrene [564]
Stage I: breakup and coalescence rates are about equal
Stage II: coalescence rate exeeds breakup rate
Stage III: conversions exceed identity point (no further breakup and coalescence)

Suspension polymerization processes give a rather broad PSD. In some EPS applications, such as for coffee cups, a very narrow PSD is required and for this grade the suspension polymerization process is commonly carried out in two stages to minimize off-spec resin. In the first stage, styrene is polymerized to give the narrowest PSD possible and then the suspension is removed from the reactor and the polymer particles are classified by sieving to give the required narrow PSD. The particles are then resuspended in the reactor to permit impregnation with a blowing agent in the second stage. Suspension polymerization for EPS production is a three stage process. In the first stage, the particle viscosities are low and an equilibrium particle size is obtained (rates of droplet breakup and coalescence are equal). In the second stage, where the viscosity of the particles is sufficiently high, the particles grow (coalescence rate exceeds breakup rate). For EPS, VILLALOBOS [564] observed that no further breakup occurs beyond 50 % monomer conversion when the particle viscosity is about 10^4 cP. Addition of stabilizer is used to obtain the desired $\bar{d}$ at the identity point. At the identity point both breakup and coalescence rates are zero. At the end of stage II, the PSD has been established except for some minor shrinkage in the third stage due to density differences between monomer and polymer. Figure 47 shows a typical growth path for EPS beads with $\bar{d} = 0.425$ mm. Since $\bar{d}$ continually increases during stage II, the variance of the PSD also increases significantly. Figure 48 shows the broadening of the PSD with conversion observed by KONNO et al. [565].

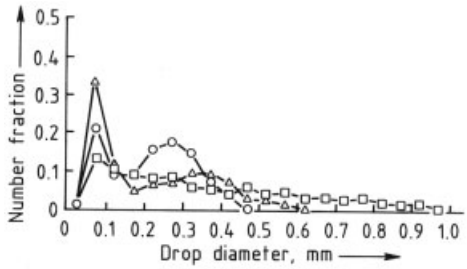

Figure 48. Transient drop size distributions for styrene suspension polymerization [565]
○ 3 % monomer conversion, 10 min
□ 20 % monomer conversion, 120 min
△ 45 % monomer conversion, 240 min

A potentially practical method of narrowing the PSD is to speed up the polymerization rate, by, for example, employing bifunctional initiators [566] to reduce the time for coalescence in stage II. This approach is based on the interaction of polymerization variables with the fluid dynamical variables to achieve a desired PSD.

Suspension polymerization is the most important commercial process for the manufacture of poly(vinyl chloride) [→ Poly(Vinyl Chloride), Chap. 4.1.]. Porous beads with a rough surface are advantageous for rapid incorporation of plasticizers. This effect can be achieved by suitable choice of dispersant [515], [567–569] or by the addition of a few percent *n*-butane [570], [571]. A review provides a comprehensive survey of polymer particle morphology development during suspension polymerization of vinyl chloride [572]. A serious attempt to predict the size distribution of primary polymer particles

and bead porosity has been made [573]. Ion-exchange resins are almost always made by suspension polymerization giving spherical beads in the size range ca. 0.3 – 1.2 mm. Most products use polystyrene cross-linked with divinylbenzene to which functional groups are subsequently attached (e.g., $-SO_3$ groups by treatment with sulfuric acid, chlorosulfonic acid, or SO_3). Weakly acidic cation exchangers are prepared by copolymerization of divinylbenzene with methacrylic acid, acrylate esters of lower alcohols, or acrylonitrile. In this case the functional groups are incorporated during polymerization, or they result from subsequent hydrolysis. Macroporosity is achieved by addition of an inert liquid which is a solvent for the monomer but a nonsolvent for the polymer and which can readily be removed after polymerization is complete [574].

With water-soluble monomers, especially acrylamide and comonomers (e.g., acrylic acid, dimethylaminoethyl acrylate) it is common to use "inverse microsuspension polymerization" in which a concentrated aqueous solution of the monomer is emulsified in an oily continuous phase, usually a hydrocarbon, and then polymerized with either an oil- or water-soluble initiator. When an oil-soluble initiator is used with an aromatic continuous phase, the kinetics have been shown to resemble those of emulsion polymerization [575–577]. The resulting polymer particles are much smaller than the original monomer droplets, and the number of radicals per polymer particle is small (< 1). However, when paraffinic oil continuous phases are used, as is most common commercially, the locus of polymerization is in the monomer droplets. This has been verified by dynamic light scattering measurements which failed to detect inverse micelles and indicated a constant particle morphology with increasing conversion [578–580]. The polymerization therefore physically and kinetically resembles a suspension polymerization and the prefix "micro" is added because average polymer particle diameter is nominally 1 μm, well below the usual size range for suspension polymerization. Suitable dispersants include Pickering emulsifiers, polymers bearing hydrophilic groups, or block or graft copolymers whose components differ in solubility [581–583].

Stirred batch reactors are by far the most common for suspension polymerization, reaching sizes up to 200 m^3. Their construction presents numerous engineering problems [554], [584–588]. The most effective stirring in large reactors is achieved with impellers driven from below. The patent literature describes a wide variety of approaches for continuous suspension polymerization. Apparently to date there are no continuous suspension polymerizations carried out commercially. The main problems relate to the fact that, on the one hand, it is important to assure the most uniform shear gradients possible (i.e., total back-mixing) to establish the desired particle size distribution and to prevent coagulation and clumping in dead spaces. On the other hand, maximum conversion is also desirable, and this is only possible in a continuous process with a narrow residence time distribution. Diffusion of monomer through the continuous phase may not be sufficiently rapid to ensure that conversion is the same in all particles, independent of their residence time in the reactor. A particle that has a long residence may have a very low monomer concentration with concomitant high level of branching, cross-linking, and gel.

3.3.3.4. Emulsion Polymerization

Emulsion polymerization is probably the most versatile of the polymerization techniques, applicable with many monomer types in batch, semi-batch (semicontinuous), and continuous processes. The product is a finely divided (particle diameters ca. 0.05 – 0.3 μm) aqueous polymer dispersion (latex) containing up to 60 wt % solids, which can either be used in latex form or first coagulated and dried. Particle size distribution may play an important role in the final application, as with emulsion PVC paste products. However, with some polymers (e.g., styrene – butadiene rubber), the latex particles are coagulated during work up to rubber bales and so their size is irrelevant, except perhaps in the coagulation stage of the process.

Suspension and emulsion polymerization are both begun in an aqueous emulsion of a monomer with low water solubility. In emulsion polymerization, however, the locus of polymerization is in latex particles and not monomer droplets, the latex particles being much smaller and having a much larger total interfacial area [589]. Various theories exist for the mechanism

of latex particle formation, differing primarily with respect to the effect of the degree of monomer solubility in water. The polymerization process in the latex particles is also quite different from that in monomer – polymer droplets in suspension polymerization, because the primary radicals generally form in the aqueous continuous phase and migrate from there into the latex particles. The most obvious difference in the kinetics compared to solution or suspension polymerization is the fact that polymerization rate and molecular mass of the polymer produced may be increased simultaneously in emulsion polymerization. Various explanations have been offered for the observed kinetic behavior. The differences relate primarily to the question of whether or not radicals, once they have entered a latex particle, are then capable of leaving again and entering other particles. The decisive factors are the prevalence of chain transfer to small molecules (e.g., monomer) and the water solubility of the monomer. Finally, the Trommsdorff – Norrish effect also plays an important role but with some modified features. It should also be noted that polymer particles in emulsion polymerization have a rather high polymer concentration even at their birthtime.

Several reviews discuss kinetics and mechanisms [590–602] as well as mathematical models [357], [603] applicable to emulsion polymerization. Proceedings from various symposia on emulsion polymerization and related topics have also been published [604–611].

3.3.3.4.1. Theories of Emulsion Polymerization

Qualitative Theory. The qualitative theory of batch emulsion polymerization is due primarily to the groups of Fikentscher [589], [612–614] and Harkins [615–618]. It is based on a system consisting of water, a monomer with low water solubility, an emulsifier, and a water-soluble initiator that decomposes to produce radicals in the aqueous phase (see Fig. 49). The emulsifier concentration is above the critical micelle concentration (CMC) and thus micelles form. The hydrophobic interior of the micelles contains solubilized monomer, which is apportioned by diffusion out of the emulsified monomer drops and through the aqueous phase. Initiator decomposes in the water phase to generate primary radicals, which propagate with monomer dissolved in water to form oligomeric radicals. When an oligomeric radical enters a micelle it propagates rapidly with solubilized monomer to form a polymer particle. In a typical emulsion polymerization there are about 10^{13} monomer droplets per liter of emulsion, with an average droplet size of about 3 µm. This compares with ca. 10^{18} micelles, each consisting of ca. 100 emulsifier molecules with a diameter of about 5 – 10 nm. The total interfacial area of the micelles is about three orders-of-magnitude larger than that of the monomer droplets. Consequently, oligomeric radicals in the aqueous phase are much more likely to diffuse into a micelle swollen with monomer than into a monomer droplet. Polymerization thus occurs almost exclusively in the micelles and polymer particles which are later formed, consuming monomer that arrives by diffusion through the aqueous phase from the monomer droplets. Micelles are thus gradually transformed into polymer (latex) particles with a diameter of ca. 0.1 µm and a concentration of ca. 10^{17} particles per liter. As polymerization proceeds a form of subdivision occurs, with monomer being transferred from large monomer droplets with a concomitant increase in total interfacial area. As a consequence, micelles are consumed by being "stung" with an oligomeric radical from the water phase and by being adsorbed on new interfacial area which is continuously being formed. When all the micelles are consumed and the concentration of emulsifier in the aqueous phase is just about to fall below the CMC, polymer particle nucleation (via micellar nucleation) ceases. The interval from the start of the generation of oligomeric radicals in the aqueous phase to the point where micelles have been consumed is called Stage I in the emulsion polymerization process. At the end of Stage I, as illustrated in Figure 50, there is a rapid drop in the free emulsifier concentration (which throughout Stage I is equal to the CMC due to equilibrium with micelles) and the surface tension (head-space gas/latex) rises rapidly from its previously stationary value. Because the fractional coverage of the surface of polymer particles falls after the end of Stage I problems with particle stability and coagulation may occur.

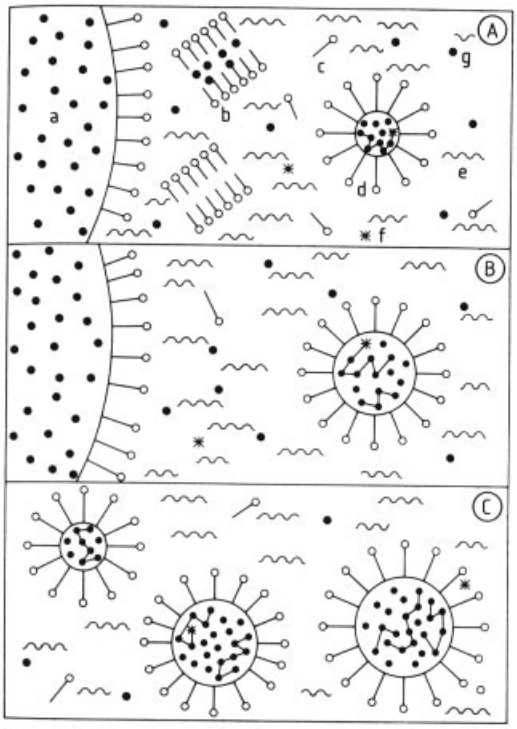

Figure 49. Schematic representation of an emulsion polymerization
A) Particle nucleation stage; B) Particle growth stage; C) Monomer finishing stage
a) Monomer droplet; b) Micelle; c) Emulsifier molecule; d) Latex particle; e) Water; f) Radical; g) Monomer molecule

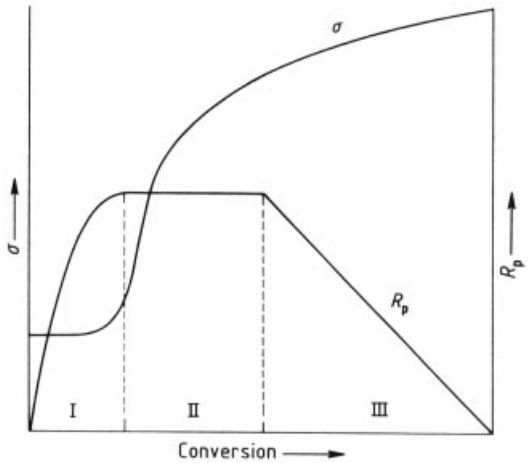

Figure 50. Overall rate of reaction R_p and surface tension σ as a function of conversion during the three phases of emulsion polymerization (schematic)

Stage II is known as the polymer particle growth stage, during which the number of particles remains constant (in the absence of coagulation), as does the monomer concentration $[M]_p$ in the latex particles as a result of monomer diffusion from the reservoir of monomer droplets. Because of the extremely high interfacial areas (polymer particle/water and monomer droplet/water) and associated very rapid mass transfer of monomer, there is an equilibrium with respect to monomer transfer from monomer droplets to polymer particles (the chemical potential of monomer is the same in all three phases, monomer droplet/water/polymer particle). The interfacial energy per unit volume for the small polymer particles contributes significantly to the free energy and thus must be accounted for as shown in Equation (3.34):

$$\frac{2V_1\sigma}{rRT} = -\left\{\ln(1-\varphi_p) + \varphi_p + \chi\varphi_p^2\right\} \tag{3.34}$$

where V_1 is the molar volume of monomer, σ is the interfacial tension (latex particle/aqueous phase), r is the radius of the latex particle; φ_p is the volume fraction of polymer in the latex particle ($\varphi_m = 1 - \varphi_p$, with φ_m the volume fraction of monomer); χ is the Flory – Huggins polymer – solvent interaction parameter [435, p. 522], [455].

Swelling and monomer concentration $[M]_p$ both increase with decreasing interfacial tension σ, increasing particle radius r, and with a decrease in the parameter χ (which is equivalent to the monomer being a better solvent for the polymer; only when $\chi < 0.5$ are monomer and polymer completely miscible).

Since σ and r increase simultaneously during Stage II, both φ_m and $[M]_p$ remain relatively constant [619], [620] provided monomer droplets are present. The constancy of N_p (the total number of polymer particles per liter) and $[M]_p$ usually results in a constant polymerization rate R_p during Stage II.

Stage III, known as the depletion or monomer finishing stage, begins with the disappearance of all monomer droplets. The only reservoir of monomer for the polymerization in the latex particles is the aqueous phase. This is hardly sufficient and in Stage III the monomer concentration $[M]_p$ falls with time and conversion. The viscosity of the latex particles increases dramatically due to increase in the number of physical chain

entanglement points as the polymer concentration increases. The self-diffusion coefficients of polymeric radicals fall and the Trommsdorff–Norrish effect, which was active in both Stages I and II, increases in intensity with monomer conversion [621], [622]. This can lead to a heat-kick when temperature control of the polymerization is inadequate. Another phenomenon which may occur is a glassy-state transition in the latex particles. This occurs when the polymerization temperature is lower than the glass transition temperature of the polymer being synthesized. The monomer acts as a plasticizer for the polymer, and propagation essentially ceases when the polymer – monomer solution undergoes glass transition at the polymerization temperature.

Overall Rate of Polymerization with a Given Number of Particles. To calculate the rate of polymerization it is necessary to know the number of polymer particles containing n radicals (or the average number of radicals per latex particle). Assuming a stationary state, the following particle population balance may be written [623]:

$$\underbrace{\frac{R_i}{N_p} \cdot N_{n-1}}_{\text{entry}} + \underbrace{\frac{K_0 S}{v}(n+1)N_{n+1}}_{\text{exit}} + \underbrace{\frac{K_{tp}}{v}(n+2)(n+1)N_{n+2}}_{\text{termination}} =$$

$$\underbrace{\frac{R_i}{N_p} \cdot N_n}_{\text{entry}} + \underbrace{\frac{K_0 S}{v} \cdot nN_n}_{\text{exit}} + \underbrace{\frac{K_{tp}}{v}(n)(n-1)N_n}_{\text{termination}} \quad (3.35)$$

The formation rate of particles with n radicals equals the disappearance rate of particles with n radicals.

where R_i is the rate of radical entry (radical $cm^{-3}\,s^{-1}$); K_0 is the rate constant for radical exit (cm/s); S is the particle surface area (cm^2); K_{tp} is the rate constant for bimolecular termination in the polymer particles (cm^3 $radical^{-1}\,s^{-1}$); and v is the particle volume.

Smith and Ewart have derived expressions for several limiting cases [623]:

Case 1) Rate of radical exit $\gg$ rate of radical entry

$$\frac{K_0 S}{v} \gg \frac{R_i}{N_p} \quad (3.36)$$

In this case the average number of radicals per particle which is given by

$$\bar{n} = \frac{\sum nN_n}{\sum N_n} \ll 1 \quad (3.37)$$

is much less than unity.

Case 2) Rate of termination > rate of entry $\gg$ rate of exit

$$\frac{K_{tp}}{v} > \frac{R_i}{N_p} \gg \frac{K_0 S}{v} \quad (3.38)$$

This case leads to $\bar{n} = 1/2$ on the basis of the following simple considerations [511]: when an oligomeric radical enters a polymer particle containing no radicals polymerization is rapid and a particle containing one radical is formed. When a second radical enters this particle, termination is instantaneous and two radicals are annihilated (two radicals cannot coexist for any significant time). Instantaneous termination occurs for small latex particles when the termination constant K_{tp} is sufficiently large. Since the exit rate is negligibly small, a given latex particle will alternately contain either one or no radical. Thus averaged over time $\bar{n} = 1/2$. One can also obtain this result with a simple kinetic analysis: application of the stationary-state hypothesis for radicals:

$$R_i = 2R_i\left(\frac{N_1}{N_p}\right) \text{ and } \bar{n} = \frac{N_1}{N_p} = 1/2 \quad (3.39)$$

where $N_p = N_0 + N_1$.

Since water-phase termination affects R_i by the same factor (on both sides of the equation) its magnitude is irrelevant. Radical scavengers in the water phase will for the same reasons have no effect on $\bar{n}$. Radical scavengers in the polymer particles would, however, lower N_1 and $\bar{n}$ to values below 1/2.

Case 3) Rate of exit $\ll$ rate of entry $\geq$ rate of termination

$$\frac{K_0 S}{v} \ll \frac{R_i}{N_p} \geq \frac{K_{tp}}{v} \quad (3.40)$$

Here the latex particles are effectively flooded with radicals, so that $\bar{n} \gg 1$, and it can be shown that the kinetics in the polymer particles correspond to those in bulk polymerization. During the buildup of radicals in the particles with time,

radical entry rate will exceed the bimolecular termination rate; however, a stationary-state with respect to radical concentration will be reached as with bulk polymerization.

A general solution to Equation (3.35) has been provided by STOCKMAYER [624] with minor corrections by O'TOOLE [625]. After multiplication by v/K_{tp} and the substitutions

$$\alpha = \frac{vR_i}{N_p K_{tp}}, \quad m = \frac{K_0 S}{K_{tp}}, \quad a = (8\alpha)^{1/2}$$

Equation (3.35) is transformed into:

$$N_{n+2}(n+2)(n+1) + N_{n+1}(n+1)m + N_{n-1}\alpha = N_n[n(n-1) + mn + \alpha]$$

Solutions to this equation follow:

$$\bar{n} = \frac{a}{4}\frac{I_0(a)}{I_1(a)} \quad \text{for } m = 0 \tag{3.41a}$$

$$\bar{n} = \frac{a}{4}\frac{I_m(a)}{I_{m-1}(a)} \quad \text{for } 0 < m < 1 \tag{3.41b}$$

$$\bar{n} = \frac{(m-1)}{2} + \frac{a}{4}\frac{I_{m-2}(a)}{I_{m-1}(a)} \quad \text{for } m \geq 1 \tag{3.41c}$$

Here the expressions $I_m(a) = i^{-m} J_m(ia)$ are first-order Bessel functions, tabulated in [626], [627]. An approximate solution to the Bessel functions is available in [597, pp. 556, 602 ff]. Figure 51 shows the dependence of $\bar{n}$ on a and m. Knowledge of $\bar{n}$ allows calculation of R_p using the relationship:

$$R_p = K_p [M]_p \bar{n} N_p / N_A \tag{3.42}$$

where N_A is Avogadro's number.

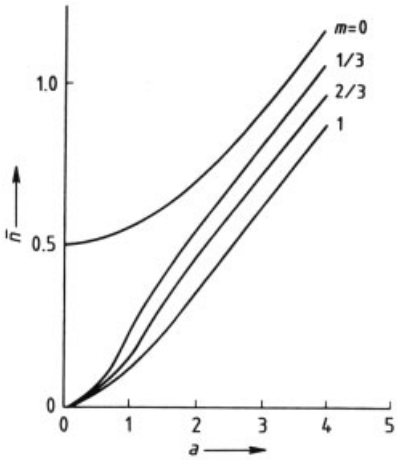

Figure 51. Average number of radials $\bar{n}$ in latex particles as a function of the parameters a and m Equation (3.41) [625]

GARDON has described a series of investigations in which the assumption $dN_n/dt = 0$ in Equation (3.35) was not made [620], [628]. Realistic parameters when used in the nonstationary form of Equation (3.35) led to the conclusion that negligible errors were introduced in Equation (3.35) by neglecting the transient accumulation term [597, pp. 566, 602 ff], [629]. One possible exception is a system subject to a very large Trommsdorff–Norrish effect and where $K_{tp} \gg K_p$ is no longer valid. In principle, Equations (3.41) and (3.42) may be used to estimate the magnitude of the Trommsdorff–Norrish effect in emulsion polymerization.

Using an appropriate relationship for the termination constant K_{tp} and its dependence on polymer concentration or monomer conversion in Stage III, one can calculate $\bar{n}$ versus conversion as shown in Figure 52 for styrene emulsion polymerization. Figure 53 shows calculated conversion – time curves for several monomer systems using appropriate expressions for K_{tp} as a function of conversion. It is clear that the Trommsdorff–Norrish effect is of greater importance in emulsion polymerization when Case 2 kinetics are obeyed and when the polymer particles are larger (vinyl acetate is clearly a Case 1 system).

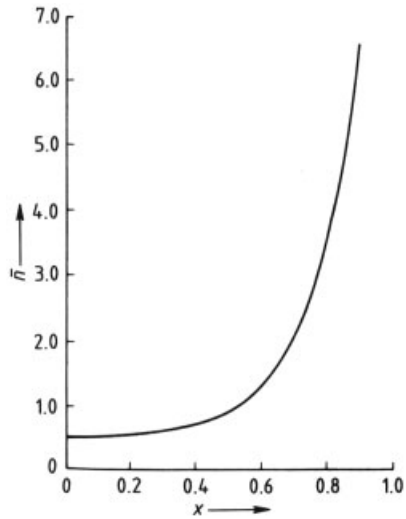

Figure 52. Theoretical course of the average number of radicals $\bar{n}$ per latex particle with increasing conversion [630] Emulsion polymerization of styrene, calculated on the basis of Equation (3.41)

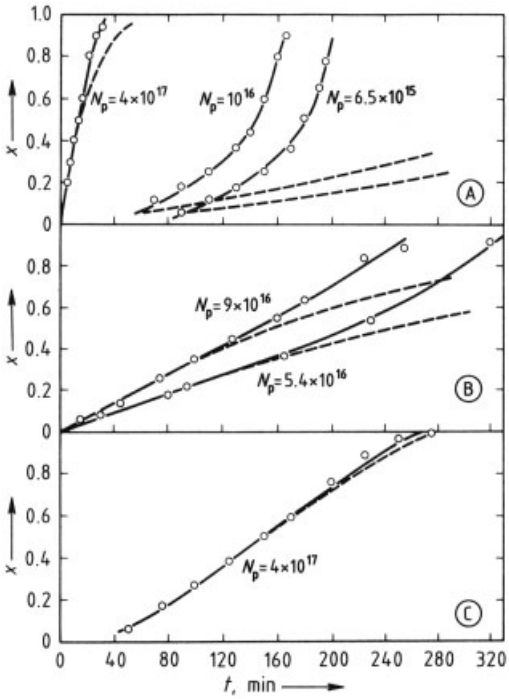

Figure 53. Time – conversion curves for emulsion polymerization
A) Methyl methacrylate (60 °C); B) Styrene (60 °C); C) Vinyl acetate (50 °C)
—- calculated with the Trommsdorff–Norrish effect
· · · · · calculated without the Trommsdorff–Norrish effect
○ experimental data

It seems remarkable that the empirical relationships for K_{tp} as a function of conversion found for bulk polymerization can be applied so effectively to emulsion polymerization, as shown in Figure 53. Part of the answer as to why this agreement exists may be that when two radicals cannot coexist in a polymer particle (instantaneous termination), $\bar{n}$ is independent of K_{tp}. At higher conversions when $\bar{n}$ is clearly > 0.5 for a Case 2 system, the size of polymeric radicals in the latex particles should depend on the rates of termination and radical entry when chain transfer to monomer is negligible. However, if chain transfer to monomer is significant, the size of polymeric radicals depends only on polymerization temperature and not on the type of polymerization process.

To explain kinetic results obtained with styrene, for which R_p fails to decline on entering Stage III, it has been suggested that one should consider uneven swelling of latex particles composed of a monomer-rich shell and a polymer-rich core [631–633]. The monomer concentration in the shell, which is the site of polymerization, should then remain constant even in Stage III. These kinetic arguments have been criticized by several other authors [630], [634], [635] and the observed facts can also be largely accounted for by considering the Trommsdorff – Norrish effect in which a simultaneous decrease in $[M]_p$ and K_{tp} effectively compensate for each other. However, experiments involving the polymerization of styrene on a seed latex where either the monomer or the seed latex is labeled [632] clearly suggest that in Stage II the monomer does indeed polymerize preferentially in the outer shell. The contradiction with respect to Equation (3.34), which has also been verified experimentally [620], [636], [637], has not been adequately explained. Neutron scattering experiments indicate that the polystyrene latex particles are uniformly swollen by styrene.

If the exit rate of radicals is negligible ($m = 0$), then a sudden reduction in the rate of initiation in the water phase to $R_i = 0$ should result in a constant R_p. In fact, however, experiments involving intermittent γ-Co [638] or UV radiation [639] as initiator reveal that R_p decreases rapidly as soon as irradiation ceases. This suggests that radicals can be desorbed from one latex particle and reabsorbed in another. There thus exists an alternative termination process: interparticle termination that reduces the mean concentration of radicals $\bar{n}$. Equation (3.35) makes no provision for possible reentry of a desorbed radical. According to UGELSTAD [640–643], the rate of radical entry R_i is a function of both the rate of radical formation in the aquous phease by initiator decomposition $R_{i,w}$ and the rate of desorption, diminished by the rate of termination in the aqueous phase:

$$R_i = R_{i,w} + K_d \sum_{n=1}^{\infty} nN_n - 2K_{tw}\,[R^{\cdot}]_w^2 \qquad (3.43)$$

In contrast to Eq. (3.35), this expression contains the desorption constant $K_d = K_0 S/v$ [s^{-1}], removing the precondition that the rate of radical exit be proportional to the latex particle surface area. Assuming that the radical absorption rate into polymer particles is proportional to $[R^{\bullet}]_W$ and substituting for $R_i = K_a\,[R^{\bullet}]_W$ and multipli-

cation by $\left(\frac{v}{K_{\mathrm{tp}} N_{\mathrm{p}}}\right)$ and making the following substitutions:

$$\alpha = \frac{R_{\mathrm{i}}\, v}{K_{\mathrm{tp}}\, N_{\mathrm{p}}}\,; \quad \alpha' = \frac{R_{\mathrm{i,\,w}}\, v}{K_{\mathrm{tp}}\, N_{\mathrm{p}}}\,;$$
$$m = \frac{K_{\mathrm{d}}\, v}{K_{\mathrm{tp}}}\,; \quad Y = \frac{2 N_{\mathrm{p}}\, K_{\mathrm{tp}}\, K_{\mathrm{tW}}}{K_a^2\, v}$$

gives

$$\alpha = \alpha' + m\bar{n} - Y\alpha^2 \tag{3.44}$$

where Y is a dimensionless group.

Equations (3.41) and (3.44) permit calculation of $\bar{n}$ as a function of a', m, Y. The results for $Y = 0$ (i.e., negligible termination in the water phase) are shown in Figure 54. It is apparent that there is a rather large range of parameter values (α' and m) for which $\bar{n} < 0.5$; this is discussed further in [640] and [597, p. 559 ff].

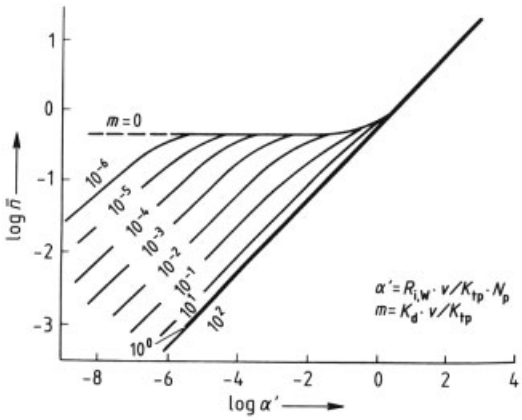

Figure 54. Average number of radicals $\bar{n}$ per latex particle, assuming interparticle termination and a negligible amount of termination in the aqueous phase ($Y = 0$), calculated by using Equations (3.41 and 3.44) [640]

The number of radicals in a polymer particle clearly depends on rates of radical entry, desorption, bimolecular termination, and chain transfer. Although radical capture efficiencies by polymer particles can be high [644], efficiencies considerably below 100 % have been observed [645]. Brooks [644] has presented general expressions for radical absorption rates which account for reabsorption and give capture efficiencies less than 100 %. Various experimental techniques for the measurement of desorption rates have been employed [645–647]. Lichti et al. [648] obtained significantly enhanced desorption rates from polystyrene particles containing soluble hydrocarbon diluents. It has been suggested that in some systems desorbed radicals cross-terminate in the water phase and are not reabsorbed. Nomura et al. [649] used the pseudokinetic rate constant method to develop expressions for $\bar{n}$ for binary copolymerizations that follow the terminal model. Nomura et al. [650] also observed that the extent of Stage I (particle nucleation stage) could be lengthened by radical desorption.

Mechanism of Particle Formation. As emulsion polymerization always involves a certain amount of monomer in the aqueous phase, and for this reason all the routes depicted in Figure 55 represent possible reactions of the radicals or radical ions that form in the water phase by initiator decomposition:

1) Entry into micelles, which are then transformed into latex particles
2) Propagation within the aqueous phase until a specific critical degree of polymerization P_{cr} is reached, at which point the growing macroradical precipitates to form a latex particle
3) Entry into an existing monomer droplet, which is converted into a latex particle
4) Entry into an existing latex particle, which is then subject once again to propagation (seed polymerization)

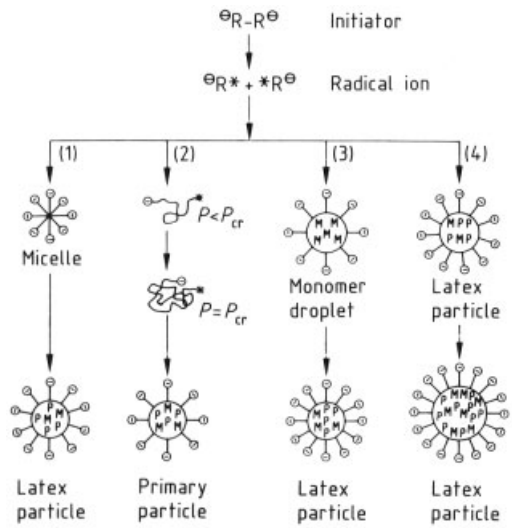

Figure 55. Possible radical reactions leading to particle formation during emulsion polymerization (schematic)

With all of these reactions, newly formed particle surface area must somehow be stabilized, either by emulsifier or through the charge originally present on the radical ion, which usually remains at the surface of the particle as the end group of the resulting polymer. If the available degree of stabilization is insufficient, the interfacial area will decrease by coagulation (or flocculation) until a point is reached at which the charge carriers present once again permit stabilization.

Depending on the nature of the rate-determining step, the rate of absorption of radicals into particles or micelles may be proportional to the particle radius, surface area, or volume. Latex particles and micelles can also differ in their capture efficiencies, and differences may even be observed between various latex particles as a result of differences in size, surface charge, and concentration of monomer and radicals [643].

In principle, the various reaction paths in Figure 55 may compete with each other, in which case the system is extremely complex. It is often possible in practice to select reaction conditions that cause one particular pathway to dominate. Further elaboration of the theory is frequently based on assumptions which led to one mechanism being dominant in polymer particle formation. Smith and Ewart [623] followed the Fikentscher – Harkins theory in assuming that latex particles form only as a result of the entry into an emulsifier micelle of a radical formed in the aqueous phase. Particle formation would then continue until the total surface area of the latex particles A_p corresponds exactly to the interface-covering capacity of the emulsifier $a_s \cdot$ [S] minus the area of monomer droplets. A correction for emulsifier dissolved in the water should also be made. In this expression [S] is the emulsifier concentration and a_s is the emulsifier covering capacity per unit of emulsifier (per mole or gram). Based on this assumption, two limiting cases can be described, in each of which it is further assumed that $\bar{n}$ = constant and $[M]_p$ = constant, and that the polymer particles grow with a constant volumetric growth rate, $dv/dt = \mu$ = constant.

Model 1) All of the radicals formed in the water phase by initiator decomposition enter micelles, so that the rate of formation of polymer particles equals the rate of generation of radicals.

$$\frac{dN_p}{dt} = R_{i,w} \tag{3.45}$$

at time $t = t_1$, where $A_p = a_s \cdot$ [S], it follows that

$$t_1 = \frac{0.53}{\mu^{2/5}} \left(\frac{a_s\,[S]}{R_{i,w}} \right)^{3/5} \tag{3.46}$$

Consequently, $N_p = R_{i,w} \cdot t_1$ and

$$N_p = 0.53 \left(\frac{R_{i,w}}{\mu} \right)^{2/5} (a_s\,[S])^{3/5} \tag{3.47}$$

This approximation leads to an upper limit for N_p, because even during Stage I some radicals enter existing polymer particles.

Model 2) Latex particles and micelles absorb radicals at a rate proportional to their current interfacial areas A_p and A_m. The rate of particle formation is therefore

$$\frac{dN_p}{dt} = R_{i,w} \left(1 - \frac{A_p}{a_s\,[S]} \right) \tag{3.48}$$

and it follows that

$$N_p = 0.37 \left(\frac{R_{i,w}}{\mu} \right)^{2/5} (a_s\,[S])^{3/5} \tag{3.49}$$

where μ is the volume growth rate of the latex particles.

Table 4. Influence of various parameters on the number of particles N_p, the overall rate of reaction R_p, and degree of polymerization P_N [612]

Parameter	N_p	R_p	P_N^*
Surfactant concentration [S]	$\sim [S]^{3/5}$	$\sim [S]^{3/5}$	$\sim [S]^{3/5}$
Initiator concentration [I]	$\sim [I]^{2/5}$	$\sim [I]^{2/5}$	$\sim [I]^{-3/5}$
Temperature	$E_N = 2/5\,(E_d - E_P)$	$E_R = E_P + E_N$	$3/5\,(E_P - E_d)$

$*$ Ignoring chain transfer

Table 4 illustrates the interdependency of the number of particles N_p, the rate of polymerization R_p, and the degree of polymerization P_N based on Equations (3.42) and (3.47) or (3.49) [612]. The relationships have been confirmed experimentally over wide ranges of the relevant variables for such relatively water-insoluble monomers as styrene, butadiene, isoprene, and chloroprene [651–658]. By contrast, other relationships are found to apply in certain situations,

especially in the case of more soluble monomers with large transfer-to-monomer rate constants, including vinyl acetate and vinyl chloride [640], [642], [659–664]. Even in the case of styrene it is not possible to explain all the experimental results on the basis of the Smith – Ewart theory. If particle formation were to conform to Model 1, then at the end of Stage I, each latex particle would contain precisely one radical, i.e., $\bar{n} = 1$ [665]. In addition the molecular mass of the polymer chain in a particle nucleated near $t = 0$ would be extremely large at the end of Stage I. For Model 2, $\bar{n} = 0.67$ at the end of Stage I [628]. Since $\bar{n} = 0.5$ during Stage II, R_p must peak at the end of Stage I. However, no such maximum has been observed with styrene. Moreover, N_p calculated by using Equations (3.47) or (3.49) often exceeds observed values by a factor of 2 – 3. These discrepancies can be rationalized by assuming that latex particles are more efficient at capturing radicals than micelles [665], [666]. Some particle coagulation would also reduce the discrepancy. If radicals formed in the aqueous phase can enter either micelles or latex particles, the condition for radical balance assures that

$$\frac{d\,[R^{\cdot}]}{dt} = R_{i,w} - K_1\,[M_i]\,[R^{\cdot}] - K_2 N_p\,[R^{\cdot}] \tag{3.50}$$

where $[M_i]$ is the micelle concentration. Assuming a stationary state is valid, one obtains

$$[R^{\cdot}] = \frac{R_{i,w}}{K_1\,[M_i] + K_2 N_p} \tag{3.51}$$

Particle formation should only occur after the entry of a radical into a micelle, so

$$\frac{dN_p}{dt} = K_1\,[M_i]\,[R^{\cdot}] \tag{3.52}$$

Substitution of Equation (3.51) into (3.52) gives

$$\frac{dN_p}{dt} = \frac{K_1\,[M_i]\,R_{i,w}}{K_1\,[M_i] + K_2 N_p} = \frac{R_{i,w}}{1 + \frac{K_2 N_p}{K_1\,[M_i]}} = \frac{R_{i,w}}{1 + \frac{\varepsilon N_p}{[S]}} \tag{3.53}$$

where $[S]/[M_i] = M_m$ is the aggregation number for emulsifier molecules in micelles (in many cases, $M_m \approx 100$); $K_2 N_p / K_1\,[M_i] = \varepsilon N_p/[S]$ is the ratio of the rate of entry into latex particles versus micelles, and $\varepsilon = (K_2/K_1) M_m$ characterizes the radical capture efficiency of latex particles relative to micelles. For $\varepsilon N_p/[S] \ll 1$, Equation (3.53) is equivalent to Equation (3.45) and Smith – Ewart kinetics are applicable in the form of Equation (3.47).

On the other hand, in the case where $\varepsilon N_p/[S] \gg 1$, there is a somewhat different dependency, with $N_p \sim (R_{i,w})^{2/7}$ and $N_p \sim [S]^{5/7}$. To obtain agreement between calculated and experimental N_p, it is necessary to introduce the values $\varepsilon = 1.3 \times 10^5$ for styrene [666] and $\varepsilon = 1.2 \times 10^7$ for vinyl acetate [667]. This is explained by assuming that radicals exit very rapidly from very small latex particles as well as micelles – so rapidly that they do not have sufficient time to propagate to the extent required for insolubility in the water phase. HANSEN and UGELSTAD [597, p. 556, 602 ff], [668] have applied Danckwerts theory of diffusion with homogeneous reaction [669] to this problem. The rate of absorption by latex particles that already contain a radical is significantly larger than that of latex particles and micelles containing no radicals, because the rate of termination between two radicals is much higher than the rate of propagation. Termination by combination with a second radical normally leads to molecules that are no longer able to desorb due to having exceeded the critical degree of polymerization P_{cr}.

In the derivation of Equations (3.47) and (3.49) the volumetric growth rate μ of a latex particle, given by

$$\mu = \frac{dv}{dt} = \frac{K_p \varphi_m}{N_A\,(1 + \varphi_m)} \cdot \frac{\varrho_m}{\varrho_p} \cdot \bar{n} \tag{3.54}$$

was considered constant since $\bar{n}$ was taken to be a constant, independent of the latex particle age. This assumption is no longer valid when Case 1 kinetics apply, and $\bar{n}$ increases with particle growth. Reference [643] has therefore adopted the following general relationship for the rate of particle formation:

$$\frac{dN_p}{dt} = R_i \cdot \frac{\delta N_M\, r_M^{\chi}}{\delta N_M\, r_M^{\chi} + N_M\, r_p^{\chi}} \tag{3.55}$$

where N_M is the number of micelles, r_M is the micelle radius, and r_p is the particle radius, χ (= 1, 2, 3) is the exponent that characterizes the dependence of the capture rate on the radii of the micelles and polymer particles, and δ is an efficiency factor for the capture of radicals by micelles relative to polymer particles. This in turn is a function of the factor ε in Equation (3.56):

$$\delta = \frac{M_m}{\varepsilon}\left(\frac{r_p}{r_M}\right)^{\chi} \tag{3.56}$$

Neglecting termination in the aqueous phase, Equation (3.43) is transformed into

$$R_i = R_{i,w} + K_d\,\bar{n}\,N_p \tag{3.57}$$

and Equations (3.55) and (3.57) then lead to

$$\frac{dN_p}{dt} = \frac{R_{i,w} + K_d\bar{n}N_p}{\left[1 + \frac{N_p}{\delta N_M}\left(\frac{r_p}{r_M}\right)^{x}\right]} \tag{3.58}$$

Assuming that only monomer transfer radicals are capable of desorbing from latex particles, HANSEN and UGELSTAD [643] constructed a balance equation for the formation and disappearance of latex particles containing a single monomer transfer radical, and then proceeded to solve it under the assumption of steady state. One can then assume either that desorbed radicals do not add monomer in the aqueous phase, and instead simply wander from one particle to another, or that such radicals do in fact add at least one monomer molecule in the aqueous phase. In the latter case, no monomer transfer radicals absorb into the latex particles, and the chain-transfer reaction becomes the sole source of monomer radicals in the particles. For further details refer to the original article [643]. What finally results from these calculations is a set of curves representing the number of particles N_p as a function of the emulsifier concentration [S] and the rate of initiation $R_{i,w}$ as shown in Figure 56 for various monomer types, assuming there is no chain transfer to monomer in the aqueous phase. Other graphs of interest corresponding to various values of δ and χ may be found in [643]. These curves serve as a basis for the rule that the sum of the orders with respect to [S] or [I] (or $R_{i,w}$) for styrene, methyl methacrylate, vinyl acetate, or vinyl chloride is always in the vicinity of 1.0. Stated more generally

$$N_p \sim [S]^z\,, N_p \sim [I]^{1-z} \tag{3.59}$$

where $0.6 < z < 1.0$. If desorption and reabsorption dominate (total radical entry rate $\gg R_{i,w}$) $z = 1.0$; if there is no desorption of radicals, $z = 0.6$ (usual value for styrene). The actual value of z may vary with [S] and [I]. If all other variables are held constant, then an increase in [S] or [I] leads to an increase in the order z relative to [S] or [I], while the order with respect to the variable that is held constant decreases [643].

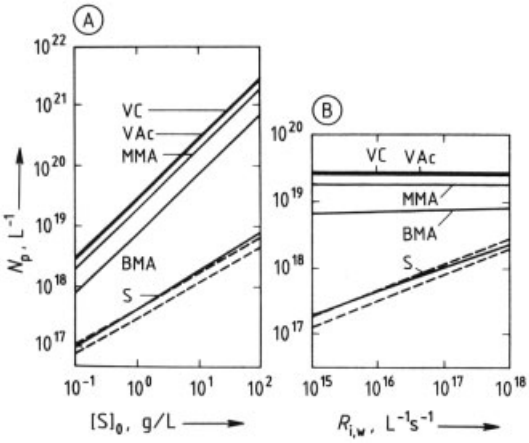

Figure 56. Number of particles N_p as a function of A) the initial emulsifier concentration $[S]_0$ and B) the rate of initiation $R_{i,w}$ for five different monomers [643] $\delta = 1.0$; x = 1; $R_{i,w}$ in A = 10^{16} L^{-1} s^{-1} ; $[S]_0$ in B = 1.0 g/L
dotted lines: Smith – Ewart theory (Eqs. 3.47 and 3.49)
VAc: vinyl acetate; VC: vinyl chloride; MMA: methyl methacrylate; BMA: *n*-butyl methacrylate; S: styrene

In the case of monomers that are significantly more soluble in water than styrene, it is also necessary to take into account the fact that latex particles can be formed by homogeneous nucleation in the aqueous phase [655]. However, Equations (3.47) and (3.49) based on the Smith – Ewart theory (which, incidentally, do not explicitly refer to particle formation from micelles) remain valid even when homogeneous nucleation is significant [670]. The rate of particle formation then becomes equal to either $R_{i,w}$ (for Model 1 and Eq. 3.47) or $R_{i,w}$ minus the rate of radical capture by previously formed latex particles (Model 2 and Eq. 3.49). The volumetric growth rate μ is taken here to be constant, and particle nucleation ceases when the total latex particle surface area corresponds to a_s [S]. The concentration of free emulsifier should at this point be somewhat lower than the CMC.

FITCH and coworkers have developed the theory of homogeneous particle nucleation further [606, pp. 73 – 102, 103 – 116], [671–675]. With [S] < CMC, a radical produced in the aqueous phase by initiator decomposition continues to grow in this phase until it is either absorbed by an existing latex particle or a specific critical degree of polymerization P_{cr} has been exceeded,

whereupon the macroradical precipitates and forms a primary polymer particle. In this case the rate of particle formation is described by

$$\frac{dN_p}{dt} = R_{i,w} - v_c \tag{3.60}$$

A steady-state (where $R_{i,w} = v_c$) is reached at $t = t_s$, which means that the number of polymer particles nucleated is given by

$$N_p = \int_0^{t_s} (R_{i,w} - v_c)\, dt \tag{3.61}$$

originally, the rate of capture v_c was taken to be [628], [673].

$$v_c = R_{i,w} (N_p \pi r^2)\, L \tag{3.62}$$

where L is the length of the path traversed by the growing radical from its point of origin to the point where it precipitates as a primary particle (collision theory). This proposal has been subjected to criticism [493, pp. 163 ff], [629] and the expression

$$v_c = 4\pi D_{op} C_S N_p r \tag{3.63}$$

where D_{op} is the mean diffusion coefficient for oligomeric radicals and latex particles in the aqueous phase and C_S is the concentration of oligomeric radicals in the same phase in fact provides a better fit than Equation (3.62) to experimental data on particle nucleation in the presence of seed particle [675]. The effectiveness of seed particles in capturing oligomeric radicals before they grow and precipitate to form new particles corresponds not to $N_p r^2$ but rather to $N_p r$. Hansen and Ugelstad [597, pp. 556, 602 ff], [643] have proposed a model for radical capture that takes into account not only the possibility of radical desorption, but also the electrostatic repulsion between radicals and particles.

One practical application of these considerations relates to stepwise seed polymerization as a means of preparing uniformly large (monodisperse) latex particles [676], [677], in which the key consideration is keeping v_c as large as possible; i.e., preventing the formation of new polymer particles.

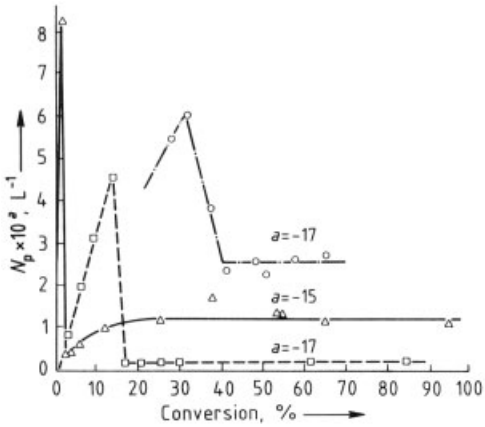

Symbol	[M]	[S]	[Fe^{2+}]
——	0.0361	0	0
·····	0.0364	0	2.4×10^{-6}
–·–·	0.0361	7.95×10^{-4}	

Note the various factors that apply to the ordinate

Figure 57. Number of particles N_p as a function of conversion in the emulsion polymerization of methyl methacrylate with sodium dodecylsulfate as emulsifier [606, pp. 73 – 102]
Concentrations in mol/L: $[S_2O_8^{2-}] = 7.35 \times 10^{-4}$, $[HSO_3^-] = 1.14 \times 10^{-3}$; $T = 30\,°C$

In certain formulations (e.g., in the absence of emulsifier, or where the emulsifier concentration is very low) charge density on the polymer particle surface is insufficient to stabilize the resulting particles. Moreover, with relatively soluble monomers such as methyl acrylate the interfacial tension is especially low, and emulsifiers are apparently less strongly adsorbed. The relationships here are further complicated by the fact that particles tend to coagulate until the surface area has been reduced (surface charge density has increased) sufficiently to restore stability, a process for which experimental evidence also exists [590, pp. 292 ff], [629], [678]. Figure 57 shows experimental data for the production of particles by homogeneous nucleation, in which methyl methacrylate was polymerized with various rates of initiation $R_{i,w}$ and sodium dodecyl sulfate concentrations ranging from zero to the CMC. With increasing conversion the number of particles passes through a maximum before falling to a constant value. The conversion at this maximum as well as the final constant number of polymer particles is lower for lower $R_{i,w}$ values and lower emulsifier concentrations. Very sim-

ilar results have been obtained with ethyl acrylate [679]. Electron micrographs of the resulting polyacrylate latex particles after etching with O_2 clearly show that they were formed by coagulation of a large number of smaller primary particles.

If limited coagulation (or flocculation) occurs, it is necessary to also take into account the rate of flocculation v_f, transforming Equation (3.60) into

$$\frac{dN_p}{dt} = R_{i,w} - v_c - v_f \tag{3.64}$$

The Smoluchowski theory [680] can be used to calculate the flocculation rate v_f [674].

If to an anionic or cationic emulsifier there is added a small amount of a straight-chain alcohol whose chain length corresponds at least to that of the emulsifier, then emulsification leads to an especially stable emulsion consisting of many very small monomer droplets. Polymerization of such an emulsion with an oil- or water-soluble initiator can, in an appropriate formulation, lead to almost total adsorption of the emulsifier on the surface of the monomer droplets. Since the number of micelles is no longer large compared to the number of monomer droplets, polymerization occurs partially or even principally within the monomer droplets. The consequence is a bimodal distribution of polymer particle sizes, where small particles formed in the aqueous phase are accompanied by large particles whose origin is in the monomer droplets. In the limiting case, the size distribution of the latex particles corresponds almost exactly to that of the original monomer droplet size distribution [681], [682]. These relatively large latex particles may contain considerably more than one radical and the observed polymerization kinetics are similar to those for homogeneous bulk polymerization [683], [684]. This therefore represents a transition toward suspension bead polymerization.

An emulsion polymerization in which polymerization occurs in both latex particles (formed from micellar or homogeneous nucleation) and monomer droplets is called miniemulsion polymerization (a term coined by the Lehigh University Group). For such systems bimodal PSD as well as polymerization rates have been both measured and predicted [685–687].

Song and Poehlein [688–690] have developed new approaches to modeling homogeneous nucleation for monomers having a range of water solubility. They account for the variation of the critical chain length for precipitation with concentration, and for sparingly soluble monomers suggest that oligomers generated in the water phase form micelles which are then stung by a radical to generate polymer particles. They also showed that the number of polymer particles N_p increases slowly with emulsifier concentration [S] below the CMC, where homogeneous nucleation is dominant, but rapidly increases above the CMC due to micellar nucleation. Others have concluded that micellar nucleation is dominant even for monomer with appreciable water solubility [691].

Radical scavengers (e.g., O_2 and monomer inhibitors such as *tert*-butyl catechol) can have interesting effects on the number of polymer particles formed and $\bar{n}$. Radical scavengers have been classified as water soluble and monomer (organic) soluble [692], [693]. Oxygen, which dissolves in water, monomer droplets, micelles, and polymer particles, mainly influences the initiation of particle nucleation. Oxygen dissolved in the aqueous phase reacts rapidly with primary radicals, preventing particle nucleation. This results in an induction period in which no polymerization occurs. The induction period ends when the oxygen content of the aqueous phase has been reduced to an almost indetectable level. After the induction period, apart from usually a minor loss of initiator, the polymerization takes its normal course. However, in the presence of, say, *tert*-butyl catechol (an oil-soluble inhibitor) in the micelles and polymer particles, polymer particles are nucleated, but $\bar{n}$ can be very low due to radicals reacting with inhibitor. This reduces the volumetric growth rate μ and hence a larger number of polymer particles are nucleated (refer to Equations (3.47), (3.49), and (3.54). Once the organic-soluble inhibitor has been consumed, the polymerization rate recovers but to a rate higher than originally expected in the absence of the inhibitor. This phenomenon can cause low-frequency oscillations in reactor trains when inhibitor contents in monomer recycle streams vary and potential safety problems for all reactor types (heat generation rate may be excessive).

3.3.3.4.2. Physicochemical Parameters of Dispersions

This section treats some properties of dispersions that can be unambiguously described in terms of physiochemical quantities. In addition to these, applications technology makes use of a number of other characteristics, primarily empirical quantities and tests, [694], [695].

Molecular Mass Distribution of Linear Homopolymer Chains. In emulsion polymerization, chain transfer to monomer is more important than in homogeneous bulk polymerization because bimolecular termination rates for a given polymerization rate are generally lower for emulsion polymerization. This is why it is possible to obtain high molecular mass at high polymerization rates in these systems. When transfer to monomer is dominant in controlling molecular mass development, the process type is irrelevant and molecular masses of linear chains depend solely on polymerization temperature. Therefore, equations developed in Section 2.2 for bulk and solution polymerization in homogeneous media may be used to calculate molecular mass distributions. A useful experimental test to determine whether chain-transfer reactions to small molecules (monomer or chain-transfer agent) are dominant in producing polymer is to change the concentration of initiator by several fold. If molecular masses decrease with increasing initiator concentration then a significant fraction of polymer chains are formed by bimolecular termination. However, if polymer molecular mass is independent of initiator concentration, then chain-transfer reactions are dominant in producing polymer chains [696].

When termination reactions are significant in molecular mass development, the calculation of MWD and average molecular masses is generally not straightforward. There are several complicating factors. Firstly, the MWD depends on the polymer particle size distribution, which is difficult to predict in commercial latex systems. Even if the PSD were known one would still need to know the mechanism by which radicals enter particles. Attempts to predict the PSD have been made by using population balance methods [697–700].

Experimental verification of these PSD predictions is lacking in most cases. The molecular masses depend on radical entry rate per particle, which in turn is some unknown function of particle size and number. For the special case of a monodisperse latex the radical entry rate is known, however, and this case can be easily treated when termination is instantaneous (rate of radical entry is much smaller than the rate of termination in the polymer particle; or in other words two radicals cannot coexist in a polymer particle). This is true for Case 2 kinetics ($\bar{n} = 1/2$). An expression is now derived which describes molecular mass development for this special case. Transfer to monomer and to chain-transfer agent are included in the analysis.

Consider an oligomeric radical in a polymer particle. The probability that this radical adds r monomer molecules and then terminates either by chain transfer to monomer or chain-transfer agent T or by instantaneous termination with a new oligomeric radical which enters the particle from the aqueous phase is given by $\varphi^{r}\ (1-\varphi)$, where

$$\varphi = \frac{K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}}}{K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}} + K_{\mathrm{fm}}\,[\mathrm{M}]_{\mathrm{p}} + K_{\mathrm{fT}}\,[\mathrm{T}]_{\mathrm{p}} + \left(\frac{R_{\mathrm{i,w}}}{N_{\mathrm{p}}}\right) N_{\mathrm{A}}} \tag{3.65}$$

The instantaneous weight chain length distribution $W(r,t)$ is therefore given by

$$W\ (r,t) = \frac{r\ (1-\varphi)\ \varphi^{r}}{\int_{0}^{\infty} r\ (1-\varphi)\ \varphi^{r}\mathrm{d}r}$$

$$= r\ (1-\varphi)^{2} \exp\left(-\ (1-\varphi)\ r\right) \tag{3.66}$$

where K_{fm} and K_{fT} are rate constants for chain transfer to monomer and chain-transfer agent T.

Equation (3.66) is the most probable distribution, with number- and weight-average degrees of polymerization given by

$$P_{\mathrm{N}} = \frac{1}{(1-\varphi)} \tag{3.67}$$

$$P_{\mathrm{W}}/P_{\mathrm{N}} = 2 \tag{3.68}$$

These expressions should be valid for systems such as styrene and methyl methacrylate when Case 2 kinetics apply and the PSD is quite narrow. It is worth mentioning that in bulk polymerization of styrene at low conversions the polydispersity ($P_{\mathrm{W}}/P_{\mathrm{N}}$) is 1.5 because bimolecular termination by combination produces most of

the polymer chains. A polydispersity of 2 in emulsion polymerization is due to the fact that when termination is instantaneous, termination by combination acts like termination by disproportionation. In fact the polydispersity (for polymer produced instantaneously) obtained in emulsion polymerization of any monomer type will always be equal to or greater than that in homogeneous solution or bulk polymerization. Equations (3.66) – (3.68) apply to an essentially monodisperse latex. For polydisperse PSD it is clear that instantaneous molecular mass properties will be broader as will the MWD of polymer accumulated in the polymer particles over polymerization times of several hours. For most commercial recipes, chain-transfer agents are used to control molecular mass development. In these cases, the MWD is independent of the PSD, and molecular mass calculations for both polymer produced instantaneously and accumulated polymer are straightforward. φ is then equal to $K_{\mathrm{fT}}\,[\mathrm{T}]_{\mathrm{p}}/K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}}$, the instantaneous weight chain length distribution is again given by Equation (3.66), and the MWD of the accumulated polymer for a batch reactor is given by Equation (3.69)

$$\bar{W}\,(r,t) = \frac{\int_0^t W\,(r,t)\,R_{\mathrm{p}}\mathrm{d}t}{\int_0^t R_{\mathrm{p}}\mathrm{d}t} \tag{3.69}$$

Several theoretical papers dealing with MWD development in emulsion homopolymerizations are available [701–705]. The problem has now been fully solved [706], [707] for the case when two or more radicals can coexist in a polymer particle (zero – one – two, systems). When the PSD is broad, the population balance equations for particle size and molecular mass are coupled, requiring numerical solution of a large set of partial differential equations [708].

Polymer Chains with Long-Chain Branching. Long-chain branching and cross-linking can occur as a result of chain transfer to polymer and by addition of polymeric radicals to terminal and pendant double bonds. The dimensionless groups which are relevant in polymer chain modification by these reactions are

$$\frac{\bar{F}_1 K_{\mathrm{fp}} Q_1}{K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}}},\ \frac{K_{\mathrm{p}}^{*} Q_0}{K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}}}\ \text{and}\ \frac{\bar{F}_2 K_{\mathrm{p}}^{*} Q_1}{K_{\mathrm{p}}\,[\mathrm{M}]_{\mathrm{p}}}$$

where K_{fp}, K_{p}^{*} are chain transfer to polymer rate constant and rate constant for the addition of polymeric radicals to terminal and pendant double bonds, respectively; Q_1 is the number of moles per liter of monomer molecules chemically bound to polymer chains (proportional to grams per liter of polymer); Q_0 is moles of polymer per litre which have a terminal double bond; $\bar{F}_1$ is the fraction of monomer molecules in the polymer chains which have an extractable atom; $\bar{F}_2$ is the fraction of monomer molecules in the polymer chains which have a pendant double bond. As these dimensionless groups increase in value, long-chain branching and cross-linking (cross-linking density and gel/sol ratio) increase. These groups increase with monomer conversion and usually increase with polymerization temperature. Long-chain branching and cross-linking reactions are more important in emulsion polymerization than in homogeneous solution, bulk, and suspension polymerization because even at the birth of a latex polymer particle, the concentration of polymer in the particle is large [709]. Chain branching, cross-linking, and gel formation are undesirable for the manufacture of cold SBR and NBR by continuous emulsion polymerization. To minimize long-chain branching, three important features of plant design are:

1) A large number of back-mixed reactors in series (this gives a narrow residence-time distribution which is desirable for low levels of branching)
2) The final conversion is kept rather low (ca. 63 % monomer conversion for cold SBR)
3) The polymerization temperature (ca. 5 °C for cold SBR) is kept low to reduce $K_{\mathrm{fp}}/K_{\mathrm{p}}$ and $K_{\mathrm{p}}^{*}/K_{\mathrm{p}}$

When carrying out semi-batch emulsion polymerization, extending the addition time of monomers lowers $[\mathrm{M}]_{\mathrm{p}}$, and this can lead to higher levels of long-chain branching and cross-linking.

Linear and Branched Copolymer Chains. For such calculations, it is recommended that the pseudo-kinetic rate constant method be employed. Emulsion copolymerizations can be successfully modeled (calculations of polymerization rate, chain microstructure, molecular mass,

and long-chain branching and cross-linking) by application of the pseudo-kinetic rate constant method [417], [649], [710] and properly accounting for changes in chemical and physical properties during the course of polymerization [711–713].

Copolymer composition is usually controlling semi-batch operation [357], [405], [714], [715]. The composition of polymer particles depends on the thermodynamic compatibility of polymer chains and is not always spatially uniform. Core/shell morphology, which is often the planned particle morphology for specialty products, can be effected by grafting and cross-linking reactions as well as compositional drift [715–718]. Special end use properties can be obtained by controlling morphology [719], [720]. The morphology of latex particles is also a function of the mechanism of particle formation. Limited coagulation (flocculation) leads to latex particles composed of a number of not entirely fused smaller primary particles. Such particles are often not smooth and spherical, displaying a rather rough surface and an irregular lumpy or "potato-like" shape [721].

Polymer Particle Size Distribution. In general, the PSD is narrower the lower the conversion at the end of Stage 1 is relative to the final conversion at the end of the polymerization. In the limiting case of stepwise seed polymerization, no new polymer particles are produced, even though a very high monomer conversion is reached, resulting in larger latex particles that are nearly monodisperse in PSD [676], [677]. A semiquantitative calculation [590, p. 275 ff] based on Smith – Ewart Case 2 kinetics shows that PSD narrows with increasing monomer/water ratio, as well as with increasing temperature, increasing initiator concentration in the aqueous phase, and decreasing emulsifier concentration [S]. Experiments involving styrene have confirmed these predictions. Quantitative calculations are also in good agreement with experiment [606, pp. 153 – 161] for styrene emulsion polymerization. A more comprehensive population balance model which accounts for coagulation in the seeded emulsion polymerization of vinyl chloride has been developed [722]. The predicted bimodal distributions of particle size are in good agreement with experimental data.

GIANETTI [699] in a theoretical study of the evolution of PSD in Stage 2 has reviewed the more important theoretical studies and discussed their limitations. The results obtained represent a generalization of previous theories in that

1) A comprehensive description for monomer systems which conform to zero – one kinetics has been made
2) The contribution of latex particles having more than one growing radical has been properly accounted for
3) Bimolecular termination has been explicitly included in the model (at least where mutual annihilation of two growing chains is not the main termination reaction)
4) General expressions for the time evolution of the cumulants of the PSDs (both in the transient and nontransient case) have been determined

GIANETTI's solutions not only revise theoretical misinterpretations which appeared in the literature, but also allow rate coefficients to be extracted with more precision from reliable experimental data. In this connection it has been shown that rate coefficients obtained from experimental PSD data are in good agreement with values found by using dilatometry to measure rates of polymerization. Recent interest in analyzing the experimental time evolution of PSD is due mainly to the possibility of determining kinetic rate constants with a small confidence interval [602], [723]. According to GIANETTI's solutions PSD and kinetic data are not independent of each other at steady state (in other words, if a certain volume dependence for the entry rate constant is assumed, from the analysis of an experimental PSD one must find the same volume dependence for the exit rate constant). Apparently, PSDs for systems with $\bar{n} \leq 1/2$ are insensitive to assumptions about the mechanisms of radical entry and exit. In this connection, a new approach based on the effect of PSD on MDW has been proposed [724]. In fact, a similar approach was proposed much earlier by DEGRAFF and POEHLEIN [725]. A CSTR was used to obtain a broad PSD for styrene emulsion polymerization, and assuming that radical capture was according to the collision theory (capture rate $\sim r^2$), a polydispersity of 4.84 was calculated for the MWD, DEGRAFF and POEHLEIN's MWD measurements showed that the molecular mass averages $\bar{M}_N$ and $\bar{M}_W$

were independent of mean residence time and that the polydispersity was ca. 3. The discrepancy between theory and experiment may be due to any number of possibilities, including:

1) Perhaps the collision model is not valid (try the diffusion model)
2) Larger latex particles may not obey Case 2 kinetics
3) Transfer to monomer may not be insignificant

Similar experimental data were found for continuous emulsion polymerization of styrene in a CSTR [726]. Simultaneous growth of latex particles by polymerization and coagulation may occur at various stages in emulsion polymerization [688–690], [723], [727–730]. For the calculation of PSD in the emulsion polymerization of vinyl chloride, it may be necessary to account for coagulation [731], [732].

The extent of coverage of the particle surfaces by emulsifier is an important factor in the stability of a dispersion. For example, this coverage should not be too large if the polymer is to be later recovered by controlled coagulation. The surface coverage is evaluated in a detergent titration [733], [734], in which the amount of emulsifier that must be added to give total monomolecular coverage of the surface is measured. Total coverage is indicated, for example, by the point where additional emulsifier no longer lowers the surface tension (surface tension remains constant after micelles are formed). If the surface coverage required for particle stability is known, emulsifier can be added during Stage II to maintain stability but not form new micelles with the danger of nucleating a second family of particles. More is involved here than simply the adsorption of an emulsifier on the particle surfaces, however. The majority of the hydrophilic end groups on polymer molecules also tend to migrate to the interface (e.g., $-OH$, $-SO_4$, $-COOH$), thereby contributing to stabilization of the polymer particles. A comparable effect is achieved by the addition of small amounts of hydrophilic comonomers (e.g., acrylic or methacrylic acid), which polymerize preferentially at the particle surface [721].

The minimum film-forming temperature of a polymer is the lowest temperature at which a dispersion can be converted into an integral film on a surface by evaporation of the water. This normally corresponds approximately to the glass transition temperature T_g of the polymer and is one of the most important application parameters of a polymer dispersion. Copolymerization of monomers whose homopolymers have a range of glass transition temperatures makes it possible to achieve desired values for both the glass temperature and the minimum film-forming temperature. The glass transition temperature $[T_g]$ of the copolymer can be calculated by using the equation

$$T_g = \frac{W_1 T_{g1} + K W_2 T_{g2}}{W_1 + K W_2} \text{ with } K = \frac{\Delta\alpha_1}{\Delta\alpha_2} \tag{3.70}$$

where W_1 and W_2 are mass fractions of monomers 1 and 2 in the copolymer, T_{g1} and T_{g2} are the glass transition temperatures of the corresponding homopolymers, and $\Delta\alpha_1$ and $\Delta\alpha_2$ are the differences between the thermal expansion coefficients above and below T_g (melt and glass) for the two homopolymers.

3.3.3.4.3. Inverse Emulsion Polymerization

In inverse emulsion polymerization, a hydrophilic monomer is dissolved in water, emulsified in a continuous hydrophobic oil phase with a water-in-oil emulsifier, and polymerized by using an oil- or water-soluble initiator. The product is a latex consisting of very small polymer particles, swollen with water and suspended in an oil continuous phase. The emulsified monomer droplets are also very small, and they may serve as the source of some of the latex particles (a form of miniemulsion polymerization). The overall rate of polymerization is significantly higher here than in the corresponding solution polymerization. This observation strongly supports the idea that what is occurring is a true inverse emulsion polymerization rather than an inverse suspension polymerization. When oil-soluble initiators are used with aromatic continuous phases the kinetics have been shown to resemble those of emulsion polymerization [576], [577], [735] with the locus of nucleation of particles in inverse micelles. However, when paraffinic oil continuous phases are used, as is most common commercially, the locus of particle nucleation and polymerization is in the

monomer droplets and the process is properly called inverse suspension polymerization (the term microsuspension is often used because the monomer droplets are nominally 1 μm in diameter) [575]. This has been verified by dynamic light scattering measurements which failed to detect inverse micelles and indicated a constant particle size throughout the polymerization [578–580]. Reviews of polymerization in inverse emulsions and microemulsions are available [736–738]. Inverse microsuspension polymerization has been used to study the kinetics of polymerization of acrylamide with cationic comonomers (quaternary ammonium cationic monomers) [739].

Water-in-oil emulsions of high molecular mass polyacrylamide and acrylamide copolymers (anionic and cationic) are particularly useful as flocculating agents in sewage treatment and as retention aids in papermaking. At the present time inverse emulsion and microsuspension polymerization is the only technology for the large-scale production of very high molecular mass polymers in a liquid-like form that affords a convenient suspension viscosity, high solids content, and easy phase inversion during application [740–742].

3.3.3.4.4. Semi-Batch Emulsion Polymerization

In semi-batch polymerizations, water, monomers, initiators, and emulsifiers may be added to a reactor over a time period which is comparable to the total polymerization time (usually over several hours). Several feed policies have been employed. For example, a starting mixture of water, emulsifier, and initiator is first prepared, usually also containing a portion of the monomer; the remaining monomer (often together with initiator) is added during the course of polymerization (monomer feed) [743]. With another procedure, the polymerization is begun in a portion of the total batch, the rest being fed into the reactor gradually after the onset of polymerization in the form of a monomer emulsion (emulsion feed), although the composition of the feed may well differ somewhat from the material used to initiate the process [744].

Semi-batch processes are intermediate between batch and continuous polymerization processes. A monomer-feed process based on a very large starter mix resembles more batch processes, whereas an emulsion semi-batch process involving a small initial charge with most of the monomers being fed in during the semi-batch period is more similar to continuous processes. These two semi-batch methods are often mentioned in the patent literature [745]. Their advantages include better control of polymerization and heat-generation rates by appropriate manipulation of feed rates. Moreover, the concentration of unreacted monomer can be minimized (offering a safer process which is less prone to runaway) and cold feed can give higher productivity (less demand for jacket or condenser cooling, for example). In the case of copolymerizations, it is possible to manipulate monomer feeds to control copolymer composition [746–751]. In particular, such methods are often exploited for the purpose of reaching a specific degree of polymerization, particle size, or particle size distribution without changing the overall composition of the copolymer in the latex particles. Adjustment of the process parameters can also have an influence, for example, on the stability of an acrylate dispersion [752–755].

Kinetic studies [756], [757] have shown that, provided the feed rate is not too large, both types of semi-batch polymerization experience a quasi-steady state in which the polymerization rate and feed rate of monomer remain comparable. During the semi-batch feeding period, the concentration of monomer $[M]_p$ is relatively constant and it stays below the saturation concentration predicted by Equation (3.34). The resulting conversion – time curves are linear over a wide range of monomer feed times. Increasing the monomer feed rate causes an increase in $[M]_p$ and according to Equation (3.42) this also leads to increases in both R_p and P_N. Slower feed rates prolong the steady state period, and this should result in a narrower MWD for the polymeric product [757], [758]. Prolonging the monomer feed period lowers $[M]_p$ and raises polymer concentration. For certain types of monomer this can lead to greater long-chain branching and cross-linking [355], [357].

When every ingredient except a reserve portion of the monomer is initially charged to the reactor, the phase ratio of monomer/water differs from the corresponding batch process, but this has no effect on the number of particles nucle-

ated. On the other hand, when polymerization is initiated in a portion of the complete recipe, the number of polymer particles nucleated per unit volume of emulsion may not be the same as for a batch process. Clearly the number of particles nucleated in the initial charge will be much less. If the onset of addition is delayed until some time after polymerization has begun, then Stage I (the nucleation stage) for the initial charge may have been completed. If no further particles are nucleated during the semi-batch feed period, the feed will be entirely directed towards the growth and stabilization of the existing polymer particles. The number of polymer particles is then fewer and their size at the end of the polymerization is larger. Based on the same total recipe, the two semi-batch processes just described can be compared and contrasted as follows:

When the initial charge and the semi-batch feed have the same recipe (composition), fewer (hence larger) latex particles are produced than for the batch process or the semi-batch process with just monomer fed to the reactor over time. If particle formation is avoided during the semi-batch period, then the initial rate of polymerization $R_{p,o}$ and the number of particles nucleated N_p will be proportional to the size of the initial charge.

Assuming particle formation does not occur during the semi-batch feed period, emulsion feed provides a narrower PSD than monomer feed or batch polymerization. If particle formation does occur during emulsion feed, then the PSD becomes broader, and in extreme cases it can be even broader than with the other processes. Particle formation during semi-batch feed can be suppressed by using a larger amount of initial charge and prolonging the time period before semi-batch feed is initiated (i.e., more polymer particles are nucleated and grow larger before semi-batch feed is initiated). Taking the increasing surface area into account also makes it possible to adjust the emulsifier feed so that new polymer particles are not nucleated and no coagulation occurs [759].

As long as the rate of addition of monomer is sufficiently large to keep the polymer particles saturated, then the rate and degree of polymerization for both semi-batch processes are independent of feed rate.

If the monomer concentration falls below its saturation value in the polymer particles (due to slow monomer feed, for example) then the rate of polymerization will equal the monomer feed rate. The rate of polymerization and heat-generation rate will increase with increasing monomer feed rate. The polymerization temperature can therefore be controlled by manipulating the monomer feed rate.

For the control of copolymer composition distribution, various monomer feed policies may be used and equations have been developed to determine the appropriate time-dependent feed rates required [356], [357], [405]. Consideration of molecular mass, long-chain branching, and cross-linking is also possible [357].

3.3.3.4.5. Continuous Emulsion Polymerization

Continuous emulsion polymerization is used particularly where the latex is to be coagulated and the solid polymer recovered. Important examples include the manufacture of styrene – butadiene rubber (cold SBR) [760–762], acrylonitrile – butadiene rubber (NBR) [763], polychloroprene [764], [765], SAN and ABS polymers [766–769], and poly(vinyl chloride) [733], [770]. For products used in latex form, most manufacturers prefer the flexibility associated with batch and semi-batch processes, especially with product lines that are subject to frequent change. Apparently, however, there is also a certain amount of large-scale polymer dispersion manufacture by the continuous process.

Systematic investigations have been carried out on the kinetics and mechanisms of particle formation in continuous stirred-tank or stirred-cascade reactors for styrene [771–781], styrene – butadiene [782], [418], [419], ethylene [783], vinyl chloride [733], [784], [785], vinyl acetate [786–791], methyl methacrylate [786], and methyl acrylate [776]. A tubular reactor has also been analyzed [780] and review articles are available [792], [793] on the general subject of continuous emulsion polymerization.

The most important elements of the theory of emulsion polymerization in a stirred reactor or stirred reactor cascade have been presented by Gerhsberg and Longfield [771]. They began with the Smith – Ewart assumption that latex particles originate with the entry of an oligomeric radical into a micelle, and that

radicals enter both latex particles and micelles according to their surface areas (collision theory, with no distinction between micelles and polymer particles). The following equation then describes the steady state with particle formation rate equal particle outflow rate (Case 2 kinetics apply)

$$\frac{\mathrm{d}N_\mathrm{p}}{\mathrm{d}t} = \frac{R_{\mathrm{i,\,w}}\,A_\mathrm{m}}{A_\mathrm{m}+A_\mathrm{p}} - \frac{N_\mathrm{p}}{\tau} = 0 \tag{3.71}$$

where A_m and A_p are the surface areas per unit volume of latex for micelles and polymer particles and τ is the mean residence time of a single CSTR. The residence-time distribution for latex particles in a CSTR is given by

$$E\,(t)\;\mathrm{d}t = \frac{\mathrm{d}N_\mathrm{p}}{N_\mathrm{p}} = \frac{1}{\tau}\exp\,(-t/\tau)\;\mathrm{d}t \tag{3.72}$$

where $E\,(t)\,\mathrm{d}t$ is the fraction of latex particles in the exit stream with age t. Because latex particles are very small and have a density close to that of water, the continuous phase, it is reasonable to assume that the residence-time distribution of latex particles and the aqueous phase are the same (latex particles follow streamlines).

It is further assumed that the role of molecularly dissolved emulsifier in the aqueous phase can be neglected relative to the amount of emulsifier in micelles and adsorbed on latex particles. Additional assumptions are that [S] $\gg$ CMC and that

$$a_\mathrm{s}\,[\mathrm{S}]\;N_\mathrm{A} = A_\mathrm{m} + A_\mathrm{p} \tag{3.73}$$

The number of particles is then given by

$$N_\mathrm{p} = \frac{R_{\mathrm{i,\,w}}\,\tau}{(1+A_\mathrm{p}/A_\mathrm{m})} \tag{3.74}$$

The particle surface area A_p for spherical latexparticles with various ages and sizes thus becomes

$$A_\mathrm{p} = 4.85\int_0^N v^{2/3}\,\mathrm{d}N_\mathrm{p} \tag{3.75}$$

Substitution of $v = f\,(K_\mathrm{p}, t, [\mathrm{M}]_\mathrm{p}, \bar{n})$ into Equation (3.75), where $[\mathrm{M}]_\mathrm{p}$ is constant (due to the very rapid diffusion of monomer from monomer droplets to polymer particles and, if monomer droplets are absent, due to rapid diffusion of monomer from polymer particle to polymer particle; the direction of monomer diffusion should be towards older and larger polymer particles) and $\bar{n} = 0.5$ and combination with Equations (3.72) and (3.74) then gives

$$N_\mathrm{p} = \frac{R_{\mathrm{i,w}}\,\tau}{\left[1+\frac{K_1 R_{i,\mathrm{w}}\,\tau}{a_\mathrm{s}\,[s]}\left(\frac{K_\mathrm{p}\,\tau\,[\mathrm{M}]_\mathrm{p}}{l-K_2\,[\mathrm{M}]_\mathrm{p}}\right)\right]} \tag{3.76}$$

As can be seen the comparison of Equations (3.74) and (3.76), the second term in the denominator of Equation (3.76) is analogous to the quantity $A_\mathrm{p}/A_\mathrm{m}$. At low concentrations of emulsifier [S] and high values of $R_{\mathrm{i,\,w}}$ and τ, it can be assumed that $A_\mathrm{p}/A_\mathrm{m} \gg 1$, so Equation (3.76) takes on the simplified form

$$N_\mathrm{p} = \frac{a_\mathrm{s}}{K_1}\,\frac{[\mathrm{S}]}{(K_\mathrm{p}\tau)^{2/3}}\left(\frac{1-K_2\,[\mathrm{M}]_\mathrm{p}}{[\mathrm{M}]_\mathrm{p}}\right)^{2/3} \tag{3.77}$$

Substitution into Equation (3.42) then gives

$$\mathrm{R_p} = K_3\,[\mathrm{M}]_\mathrm{p}\,\frac{a_\mathrm{s}\,[\mathrm{S}]\;K_\mathrm{p}^{1/3}}{\tau^{2/3}}\left(\frac{1-K_2\,[\mathrm{M}]_\mathrm{p}}{[\mathrm{M}]_\mathrm{p}}\right)^{2/3} \tag{3.78}$$

Here the constants K_1–K_3 are functions of the specific volumes V_m and V_p of monomer and polymer as well as of the molecular masses of the monomers; K_1 and K_3 contain in addition the assumed constant $\bar{n} = 0.5$ within the latex particles [771].

In order to compare exprimental data obtained over a wide range of conditions, it is convenient to rewrite Equation (3.76) in dimensionless form [772]:

$$\left(\frac{R_{\mathrm{i,\,w}}\,\tau}{N_\mathrm{p}}\right) = 1 + K_1\,\frac{R_{\mathrm{i,\,w}}\,\tau}{a_\mathrm{s}\,[\mathrm{S}]}\left(\frac{\tau K_\mathrm{p}\,[\mathrm{M}]_\mathrm{p}}{1-K_2\,[\mathrm{M}]_\mathrm{p}}\right)^{2/3} \tag{3.79}$$

or

$$\pi_1 = 1 + \pi_2$$

If N_p is expressed in terms of Equation (3.42) and introduced into Equation (3.79), what results is a second dimensionless equation for R_p:

$$\frac{R_{\mathrm{i,w}}\,\tau K_\mathrm{p}\,[\mathrm{M}]_\mathrm{p}\,\bar{n}}{R_\mathrm{p} N_\mathrm{A}} = 1 + K_1\,\frac{R_{\mathrm{i,w}}\,\tau}{a_\mathrm{s}\,[\mathrm{S}]}\left(\frac{\tau K_\mathrm{p}\,[\mathrm{M}]_\mathrm{p}}{1-K_2\,[\mathrm{M}]_\mathrm{p}}\right)^{2/3} \tag{3.80}$$

or

$$\pi_3 = 1 + \pi_2$$

Comparisons of data supplied by three different authors [771], [772], [776] are presented in

Figures 58 A and 58 B. In Figure 58 A, essentially all of the points lie above the 45° line; that is, fewer latex particles are formed than predicted by theory. By contrast, the experimental values for R_p in Figure 58 B, which are more precise, correspond well to the theoretical straight line. Overall, one may conclude that in the case of styrene the theory is well substantiated. The Gershberg – Longfield theory has been extended and modified a number of times. Thus DeGraff and Poehlein dropped the assumption that $\bar{n} = 0.5$ and instead calculated $\bar{n}$ using the Stockmayer – O'Toole solution for different particle sizes (Eq. 3.41) [772]. This accounts for the possibility of more than one radical per particle for the larger particles. Nomura et al. [773] also applied Equation (3.53), according to which radicals are absorbed more rapidly into latex particles than into micelles, to continuous emulsion polymerization and obtained a prediction that is in better agreement with experiment (as compared to that in Fig. 58 A). Brooks [794] discusses various equations for the entry of radicals into latex particles and micelles, assuming among other things that micelles disintegrate only slowly under subsaturation conditions, and that even in the presence of micelles the surface of the latex particle may not be saturated with emulsifier.

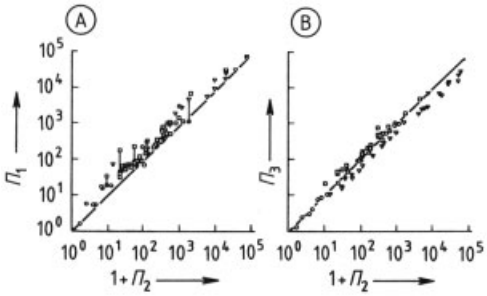

Figure 58. A) Number of particles according to Equation (3.79), and B) Overall rate of polymerization according to Equation (3.80) for the continuous emulsion polymerization of styrene (dimensionless representation) [772]
□ from [772]
▼ from [771]
○ from [776]
● calculated from values for R_p with the assumption that $\bar{n} = 1/2$

The existence of multiple steady states for the isothermal operation of a CSTR was first demonstrated in the case of continuous emulsion polymerization by Gerrens et al. [776–778], although it had earlier been predicted on theoretical grounds for autocatalytic processes in general [795–806]. Autocatalytic acceleration here is a consequence of the Trommsdorff –Norrish effect. The mass balance for monomer at steady state is given by

$$\dot{v}\,[\mathrm{M}]_0 - \dot{v}\,[\mathrm{M}] - R_p V = 0 \tag{3.81}$$

where $\dot{v}$ is the volumetric flow rate to and from the reactor, $[\mathrm{M}]_o$ is the monomer concentration in the feed, [M] is the monomer concentration in the reactor and exit stream, V is the volume of the reacting mass in the reactor, and R_p is the rate of polymerization in the reactor. Furthermore,

$$\frac{1}{\tau}\left([\mathrm{M}]_0 - [\mathrm{M}]\right) = R_p \tag{3.82}$$

where $\tau = V/\dot{v}$ is the mean residence time in the reactor.

Figure 59 shows a graphical solution of Equation (3.82) and R_p versus monomer concentration, in which data have been plotted as a function of both monomer concentration [M] and conversion X. At low values of mean residence time τ, there is one intersection (one solution) between experimental $R_p = f([\mathrm{M}])$ and the straight line (Eq. 3.82) which implies a single stable operating point at low conversion. By contrast, longer residence times lead to three points of intersection, with stable operating points at both low and high conversions and an unstable operating point at an intermediate conversion. For a train or cascade of n stirred reactors there may be as many as $n + n(n+1)/2$ stable operating points and n unstable operating points for the individual reactors as well as $n+1$ stable states for the entire cascade. Oscillations in the number of particles, the polymerization rate, and the interfacial tension have attracted considerable theoretical attention. These are attributed to bursts of particle formation that accompany swings between states of supersaturation of the latex particle surfaces, indicating large amounts of free emulsifier (and micelles), and states characterized by undersaturation, in which particle formation almost ceases. This phenomenon has long been recognized in industrial-scale continuous emulsion polymerization [733] and has now been thoroughly investigated in the case of styrene [771], [778], methyl methacrylate [786], [807], [808], and vinyl acetate [786], [798], [799], [809].

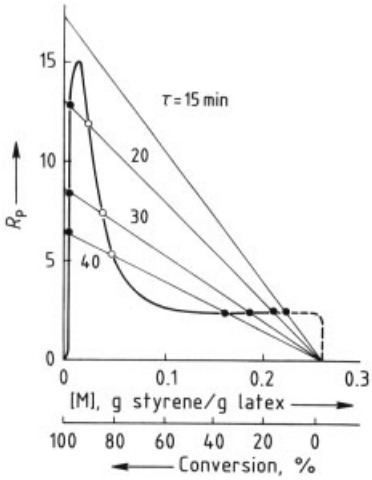

Figure 59. Emulsion polymerization of styrene. Stationary operating points for a continuous stirred reactor at various residence times τ [778]
• Stable, ○ unstable operating points; R_p in milligrams styrene per gram latex per minute

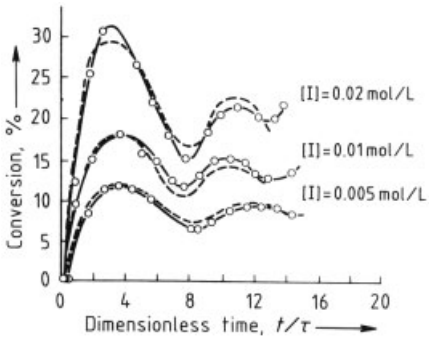

Figure 60. Oscillations in the continuous emulsion polymerization of vinyl acetate at three different initiator concentrations [I]. Comparison between experimental (○ · ○ · ○) [786] and calculated (– – –) values [787]

Model calculations are available for styrene [800], vinyl acetate [787], [790], [791], and vinyl chloride [785]. Figure 60 illustrates the close agreement between experiment and theory. Radical desorption has a strong influence (with accompanying more rapid polymer particle nucleation rates) on the oscillatory behavior of continuous emulsion polymerization in a CSTR [810–812] and radical desorption rate constants have been measured using a CSTR [810], [813]. In the absence of particle nucleation, oscillations (at least those not due to thermal effects) do not occur. A CSTR (with split flow) or tubular reactor operating as a polymer particle seed generator in series with one or more larger finishing reactors has been used to eliminate oscillations for Case 1 systems (vinyl acetate and vinyl chloride) [785], [646], [814], [815].

To complete the discussion of emulsion polymerization, a few brief comments will be made about an area which has almost been completely neglected in the literature. This is the effect of cross-linking in the polymer particle on emulsion polymerization kinetics and modeling (effect on particle nucleation and growth). The first and very important effect of cross-linking is on the equilibrium swelling of polymer particles by monomer. Taking into account the elastic free energy change due to the cross-linking network structure as well as the free energy contributions of mixing and interfacial tension, it is possible to derive an equation which gives the equilibrium monomer concentration in polymer particles

$$\frac{2V_1\sigma}{rRT} = -\left[\ln(1-\varphi_p) + \varphi_p + \chi\varphi_p^2\right] - \left(\varphi_p^{1/3} - \frac{\varphi_p}{2}\right)\left(\frac{\varrho_p}{\varrho_M}\right)\bar{\varrho}_{el} \tag{3.83}$$

where ϱ_{el} is the elastic cross-link density of the accumulated polymer [816]. The first thing to notice is that cross-linking reduces the monomer concentration and hence the volumetric growth rate of polymer particles. According to Equation (3.47) or (3.49) the number of polymer particles nucleated in Stage I will be larger. In an experimental study using styrene and divinyl monomers, NOMURA et al. have confirmed [817] that much higher numbers of particles are generated when initiated particles are cross-linked. An obvious effect of cross-linking is the strengthening of the Trommsdorff – Norrish effect with concomitant earlier and larger decrease with conversion of the termination constant and the consequences thereof. When the molecular mass of primary chains produced in smaller polymer particles is larger, one might expect gelation to occur earlier in smaller particles. There are clearly many interesting phenomena related to cross-linking in latex particles which await investigation.

3.3.4. Miscellaneous Processes

The modeling of high-pressure tubular and autoclave reactors for the manufacture of LDPE

has received a great deal of attention [818–836]. These models are notable for the detailed kinetics, which include transfer to polymer, backbiting, and β-scission in an attempt to calculate both short- and long-chain branching frequencies. The polymerizations are almost adiabatic, with large temperature increases, and phase separation between ethylene and polyethylene may occur in autoclave reactors. The tubular reactor is considered to be in plug flow, while mixing in the autoclave is rather complex and various mixing models have been considered. A comprehensive copolymerization model for ethylene and comonomers in a high-pressure continuous tubular reactor has been developed [836].

A comprehensive tubular reactor model for the solution polymerization of vinyl acetate has been developed by Hamer and Ray [837]. The kinetics employed account for transfer to polymer and reaction with terminal double bonds (trifunctional branching frequencies are calculated). The model solves the equations of motion for both axial and radial velocity profiles. Experimental validation of the model has been carried out.

The bulk and solution polymerization of styrene in tubular reactors has also been extensively studied both theoretically and experimentally [375], [838]. Spatial distributions of velocity, temperature, and concentration were modeled [838] and it was shown that for tube diameters > 2 cm, thermal runaway may occur. Internal mixers were used in larger diameter tubes to give better radial mixing and this technique appears to give polystyrenes with high molecular mass at a productivity of commercial interest [375]. References to other important studies on styrene production in tubular reactors may be found elsewhere [375], [838].

Chemical Modification of Polymers in Extruders The commercial incentive for the chemical modification of polymers by chain scission, long-chain branching, cross-linking, and grafting is the enhancement of the physical and chemical properties of polymers and polymer mixtures (alloys, blends, additives) and/or improvement of their processability. In many cases, the extruder has been found to be an effective chemical reactor in which chemical modification of high molecular mass polymers in the melt can be done economically for low-volume products. It is particularly effective for chemical modification of high molecular mass, solvent-free, commodity polymers to tailor-made products to meet the growing diversity in polymer applications. The extruder reactor provides short and controlled residence times; efficient mixing, particularly for reactants at low concentrations; and good heat transfer. Multistage reaction, with sequential feeding of reactants and removal of byproducts, is possible. There are certain advantages of twin-screw over single-screw extruders [839]. These include: intense shear mixing (with better surface renewal rates) giving molecular-level mixing (micromixing), improved heat transfer, multistage feeding, venting of volatile byproducts, and screw speed can be changed without changing the throughput. These attributes of the twin-screw extruders are ideal for chemical modification where a small amount of the low molecular mass compound to be grafted on the polymer chains must be homogeneously mixed with a highly viscous base polymer.

Chemical modifications are often carried out using free-radical reactions in the temperature range 170 °C – 350 °C. When a polymer containing a peroxide and a compound to be grafted is melt processed, chain scission, long-chain branching, cross-linking, and grafting reactions may occur simultaneously. To optimize product properties, the rates of these reactions must be controlled.

Chain Scission. Polymer chain scission occurs when chemical bonds (usually C – C) along the backbone break. Since a higher molecular mass chain has a greater number of such bonds, it experiences chain scission preferentially (for polypropylene, scission seems to be random with respect to the position of the bond along the chain backbone [840], [841]). Therefore, the broad MWDs, usually found for polyolefins narrow as the average molecular mass falls. With random scission occurring exclusively, the limiting polydispersity for all initial MWDs is two and the limiting MWD is Flory's most-probable distribution. A reduction in molecular mass improves processability. The outstanding commercial example of the use of controlled chain scission of polypropylene (PP) is in the production of controlled-rheology PP used for fiber manufacture, film-extrusion, and for fast molding applications [842].

Long-Chain Branching and Cross-Linking. Long-chain branches (tetrafunctional) are formed when radical centers on the backbone of polymer chains experience bimolecular termination by combination. If this long-chain branching process continues, a three-dimensional network (cross-linked gel) is formed. The production of higher molecular mass chains by long-chain branching produces a polymer that exhibits increased die swell and melt strength, and improved strain-hardening properties [842].

Cross-linking of polymer chains results when bimolecular coupling (via bimolecular termination by combination of polymer backbone radicals or the reaction of grafted chemical groups such as those with silane functionality [843], [844]) of polymer chains to form tetrafunctional branches occurs repeatedly. The cross-linked polymer is a mixture of polymeric gel, of effectively infinite molecular mass, and sol which is made up of linear and branched chains. Property enhancement due to cross-linking includes: increased service temperature, solvent resistance, flexural modulus, low-temperature impact strength, environmental stress cracking resistance, and reduced creep [842].

Grafting onto polymer chains can occur when a mixture of the polymer, peroxide, and unsaturated or saturated compound to be grafted are melt processed in an extruder. For example, acrylic acid or maleic anhydride can be grafted onto PP or PE to enhance compatability with more polar blends or to improve adhesion to metals, glass fibers, and to other polymer types [845], [846]. Vinyl silanes can be grafted to polyolefins to subsequently produce cross-links in the presence of moisture [843]. Functionalized additives (antioxidants, flame retardants, pigments) can be grafted onto the host polymer to give the desired property enhancement. These chemically bound additives are effectively mixed on a molecular scale and do not diffuse out of the matrix during use [847].

Source of Radicals. During normal extrusion of polyolefins, alkyl radicals may be formed thermomechanically, but at too low a rate to initiate chemical modification to the extent normally desired. To provide a suitable source of radicals, organic peroxides are added to the polymer before melt processing. In these applications the peroxides used must generate radicals which are sufficiently energetic to abstract backbone atoms and generate backbone radical centers.

Other reaction types, including anionic, cationic, anionic coordination, and condensation can be carried out in an extruder reactor [839]. Kowalski [848] describes an extruder process for halogenation of butyl rubber. Saleem and Baker [849] describe a process in which polystyrene and polyethylene along with their reactive counterparts (polystyrene with oxazoline groups, and polyethylene with carboxylic acid groups) are reacted to form polymeric compatabilizers in situ.

Kinetics and Extruder Reactor Modeling. Reviews of the kinetics of chemical modification of polymer by free-radical mechanisms are available [839], [850], [851]. Motha et al. [852] have developed a model for the grafting of unsaturated acids and silanes on polyethylene. This model does not attempt to predict the gelation point nor the gel/sol ratio in the post-gel region. A very general model which permits one to follow the change in the MWD of polymer being modified by simultaneous random scission, long-chain branching, cross-linking, and grafting has been developed [850], [851]. The model, which is in the form of an integro-differential equation is based on a simpler version developed earlier by Saito [851], [852]. Saito found an analytical solution for the limiting case of random scission occurring exclusively. Kimura [853] obtained a limiting analytical solution for the case in which random cross-linking is occurring exclusively. The general case can be handled by using the method of moments to give the molecular mass averages. The widely used Charlesby – Pinner equation was obtained by using an idealization which is questionable [854]. Simultaneous scission and cross-linking are treated as two processes in series, with random scission occurring first and then cross-linking. In addition, the initial MWD must be Flory's most probable distribution. Numerical methods must be used to solve the integro-differential equation for the case of simultaneous scission and cross-linking. This is not a trivial problem and has not been successfully solved to date.

The design of extruders for reactive processing is much more challenging than for purely physical processing [839]. Extruders must handle polymer melts with continuously changing

properties. Residence-time distribution (RTD) and micromixing can significantly affect the properties of the chemically modified polymer. One of the important advantages of twin-screw over single-screw extruders is the greater control they allow over RTD and mixing. The modeling of conventional or physical extrusion processes has been well documented in the literature, particularly for single-screw extruders. Details of the modeling of extruders may be found in the review article by TZOGANAKIS [839]. TZOGANAKIS et al. [855] developed a comprehensive rheokinetic model for the random scission of isotactic polypropylene in a single-screw extruder. Finite-element or finite-difference methods will play an important role in the solution of the three-dimensional flow equations for both single and twin-screw extruders [856].

3.3.5. Ionic Polymerization Modeling

3.3.5.1. Introduction

The most powerful processes for the control of MWD of linear chains are living anionic and cationic polymerizations. These processes are interesting from the point of view of process engineering because the MWD of the product can be varied widely by changing the RTD of the process or by oscillating the process in a controlled manner. These living ionic processes will be considered from this perspective in Section 3.3.6.1.

3.3.5.2. Heterogeneous Coordination Polymerization

The commercially important anionic coordination processes (heterogeneous ZN polymerization) are now discussed. SEPPALA and AUER have published an excellent review of much of the relevant subject matter [857]. Third- and fourth-generation catalysts have already been commercialized for slurry polymerizations. These newer catalysts possess high and stable activity, high stereospecificity independent of polymerization time, and increase in activity with increasing temperature up to about 80 °C. Electron donors may be used to regulate isotacticity. Catalysts may also be designed to control the size and shape of polymer particles.

The gas-phase UNIPOL process has become one of the most successful commercial processes for the production of polyolefins [858]. In this process, homo- and copolymerizations of ethylene with α-olefin comonomers are carried out in a fluidized-bed reactor using heterogeneous Ziegler – Natta or Philips supported metal oxide catalysts [859].

Modeling of these processes can be classified into three categories [857], [860]:

1) Microscale kinetic phenomena: coordination kinetic mechanism, multiple active site types, stereoregularity, copolymer composition and molecular mass distributions, and catalyst deactivation kinetics
2) Mesoscale particle phenomena: fracture of catalyst particles, heat generation in particles, mass transfer, and growth and influence of changing properties of reaction environment
3) Macroscale reactor phenomena: overall mass and energy balances, gas – liquid mass transfer effects, influence of reaction medium, effects of particle motion, and heat transfer limitations and influence of reactor type

Similar modeling categories could be considered for the manufacture of PVC and EPS by free-radical suspension polymerization, where particle morphology and foamed cell structure are important considerations.

One of the major challenges in modeling heterogeneous coordination polymerizations has been to rationalize the very broad MWDs and copolymer composition distributions (CCD) that are often observed. There have been two schools of thought. RAY and coworkers have until recently attempted to explain the broad MWDs in terms of diffusional resistances for reactants, which increase as the catalyst particles are enveloped by a growing shell of polymer. An additional resistance may result from the crystallization of polymer chains. Others have attempted to explain the broad MWDs and CCDs as due to multiple sites of different activity. These sites have different compositions. To model a binary copolymerization it can be assumed that each site type produces a polymer with STOCKMAYER's bivariate distribution [861], [862]. The more difficult problem is, of

course, to determine the number of active site types and the concentration of each. Two active-site-type models have been used [862], [863] as well as site-type distribution functions [864]. Among the diffusion-limited models, the multigrain model is the most comprehensive. It is assumed that fragmentation of the original catalyst particle is complete before appreciable polymer has been produced [865], [866]. Thus, the large macroparticle is composed of many catalyst fragments encapsulated by polymer [865], [867–869]. Both macroparticle and microparticle mass transfer can be important for large catalyst particles of high activity [857]. Ray et al. [868] found that the maximum possible temperature rise in both micro and macroparticles is negligible in slurry polymerization of propylene. However, in gas-phase processes the temperature inside the polymer particle could reach its melting point under some conditions [866]. Near the beginning of polymerization, the polymer particles may heat up because of the high catalyst activity and polymerization rate. In summary, it can be said that the very broad MWDs and CCDs can be explained by invoking multiple site types. However, under certain conditions, mass- and heat-transfer resistances may also play a significant role. The development of more realistic models for copolymerization on heterogeneous coordination catalysts will require the measurement of the bivariate distribution of molecular mass and composition by a combination of analytical techniques such as TREF/GPC and TREF/NMR to observe the detailed microstructure of the copolymer chains [870].

3.3.6. Process Variables, Reactor Dynamics/ Stability, On-Line Monitoring and Control

3.3.6.1. Influence of Reactor Type and Configuration on Molecular Mass and Copolymer Composition Distributions, and on Long-Chain Branching and Cross-Linking

Survey of Idealized Reactor Types and Simple Polymerization Reactions. Molecular mass and copolymer composition distributions, long-chain branching, and cross-linking can exert a significant influence on the properties of all polymers. It is thus important to know the distribution functions that correspond to various polymerization mechanisms and how these distributions change with reactor type and configuration. Another way of expressing this is to note that these molecular properties are strong functions of the residence-time distribution of the polymerizing mass and that once an optimal residence-time distribution has been established, one can choose reactor types and manipulate reactor configuration to obtain this residence-time distribution in a practical manner. Most deviations of predicted and measured molecular properties are likely to be due to assumptions about levels of micromixing and segregated flow and the comparison of predicted instantaneous molecular properties with measurements of these properties on accumulated polymer (e.g., instantaneous properties can change with time in a batch reactor as, say, [M] decreases and hence the distributions for molecular properties for the accumulated polymer will have larger variances). From an industrial standpoint, however, any process of interest involves high – if not complete – conversion, and a wide variety of batch and continuous reactors is utilized. The relationships between reactor types and configurations and the various polymerization mechanisms are often quite complex, and it will only be possible within the confines of this article to treat certain simple ideal reactors and reaction types.

Figure 61 shows four ideal reactors, characterizing them on the basis of residence-time distribution and the temporal and spatial course of chemical reaction. Batch reactors (BR), continuous plug flow reactors (CPFR), and homogeneous continuous stirred-tank reactors (HCSTR) are treated in the article Stirred –Tank and Loop Reactors. With a homogeneous continuous stirred-tank reactor, perfect mixing down to the molecular level is assumed, and for a CSTR with an ideal residence-time distribution, there are by definition no spatial variations in temperature and concentrations. Mixing at the molecular level is often called "micromixing". This condition is often not achieved in practice, however, and for this reason a fourth ideal reactor type is introduced, the segregated continuous stirred-tank reactor (SCSTR) [801–803]. Here the fluid phase is regarded as subdivided into many small

isolated compartments. Each compartment contains a large number of molecules, which are permanently confined within the limits of that compartment; therefore, the individual compartments function as miniature batch reactors with different residence times in the flow reactor. The compartments themselves are taken to be ideally mixed, leading to what is called "macromixing" despite total segregation of molecules in different compartments. Thus, the sum of all the compartments in a segregated continuous stirred-tank reactor have the same residence-time distibution as the contents of the homogeneous stirred-tank reactor. A macroscopic mean taken over all the compartments in the effluent stream and in the reactor itself would show concentrations and temperature that are constant both spatially and with respect to time. On the other hand, a probe capable of microscopic sampling of individual compartments would reveal concentrations that varied in a statistical manner from one compartment to another. Given the high viscosity of a typical polymerizing solution it is quite likely that many solution polymerizations actually occur in segregated systems [804]. Ideal reactors are discussed in greater detail in, for example, [795], [805].

Polymerization reactions can be divided according to their kinetics into three classes:

1) Monomer coupling with bimolecular termination (e.g., free-radical polymerization)

$$\left.\begin{array}{l} \mathrm{I} \xrightarrow{K_d} 2\,\mathrm{R}^{\bullet}_{in} \\ \mathrm{R}^{\bullet}_{in} + \mathrm{M} \xrightarrow{K_i} \mathrm{R}^{\bullet}_{1} \end{array}\right\} \text{initiation}$$

$$\mathrm{R}^{\bullet}_{r} + \mathrm{M} \xrightarrow{K_p} \mathrm{R}^{\bullet}_{r+1} \qquad \text{propagation}$$

$$\mathrm{R}^{\bullet}_{r} + \mathrm{R}^{\bullet}_{s} \left.\begin{array}{l} \xrightarrow{K_{td}} \mathrm{P}_r + \mathrm{P}_s \\ \quad\text{by disproportionation} \\ \xrightarrow{K_{tc}} \mathrm{P}_{r+s} \\ \quad\text{by combination} \end{array}\right\} \begin{array}{l}\text{bimolecular}\\ \text{termination}\end{array}$$

2) Monomer coupling without termination (e.g., living anionic polymerization)

$$\mathrm{I} + \mathrm{M} \xrightarrow{K_i} \mathrm{P}^{*}_{1} \qquad \text{initiation}$$

$$\mathrm{P}^{*}_{r} + \mathrm{M} \xrightarrow{K_p} \mathrm{P}^{*}_{r+1} \qquad \text{propagation}$$

3) Polymer coupling (e.g., polycondensation)

$$P_r + P_s \xrightarrow{K} P_{r+s} \qquad \text{propagation}$$

In both cases of monomer coupling, many individual monomer molecules add successively to the growing chains. Monomer coupling with bimolecular termination is characteristic of free-radical polymerizations. Termination may occur through the combination of two growing radicals (degree of coupling $K = 2$). Termination by disproportionation is an example of a process in which $K = 1$ [806]. Monomer coupling without termination is encountered, for example, in living anionic or cationic polymerizations. Polymer coupling between macromolecules is characteristic of polycondensation and polyaddition reactions. The discussion that follows has been restricted deliberately to the simplest reactions possible, with precisely simultaneous initiation of all active centers in the case of anionic polymerization and exact stoichiometric balances in polycondensation. Chain-transfer reactions will be neglected, as will exchange reactions (e.g., transesterification, transamidation). A few consequences of these complex reactions are outlined in [871].

The three types of polymerization differ considerably in the way the degree of polymerization of the accumulated polymer $\bar{P}_N$ (averaged over the course of the conversion process) changes as a function of conversion. As shown in Figure 62, a high degree of polymerization $\bar{P}_N$ is reached at very small conversions in free-radical polymerization. In a batch reactor this value subsequently decreases with increasing conversion as a consequence of decrease in the monomer concentration [M] (assuming that the Trommsdorff effect is negligible and that the radical initiation rate is essentially constant). With living polymerization in the absence of termination reactions, on the other hand, $\bar{P}_N$ increases linearly with conversion, and in polycondensation a high degree of polymerization is achieved only at very high conversion (e.g., 95 % conversion is required to reach $\bar{P}_N = 20$).

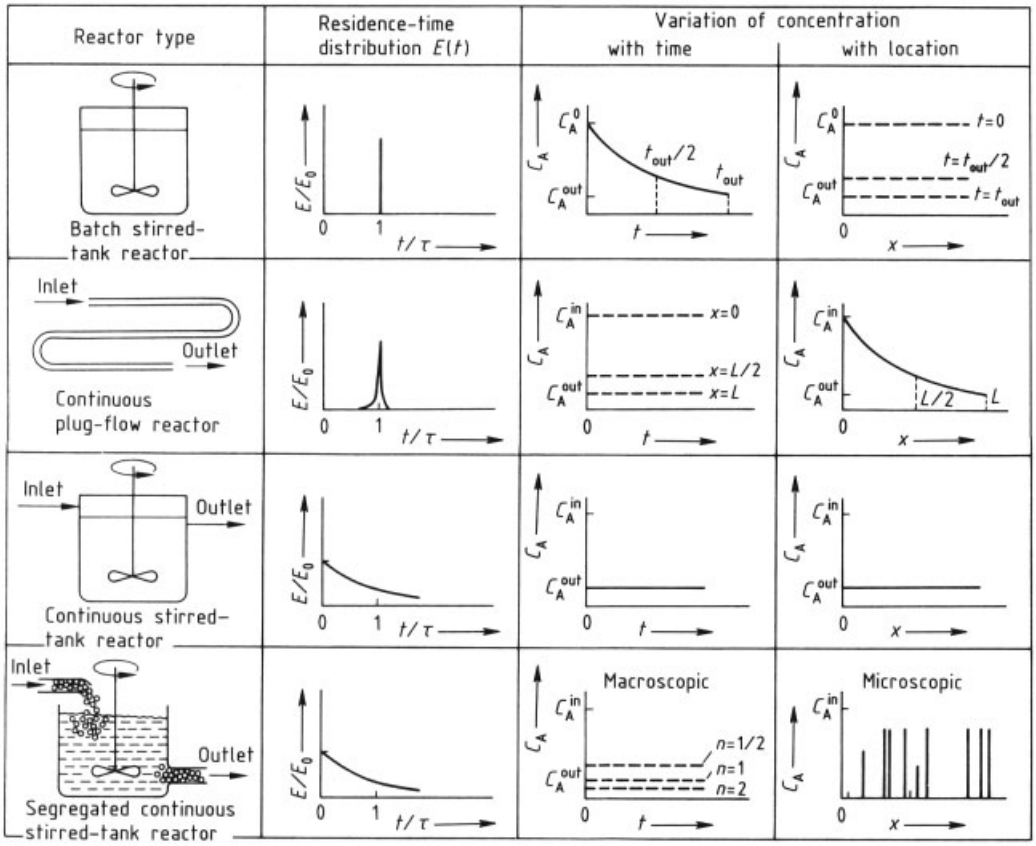

Figure 61. Schematic illustration of the course of reaction in various types of reactors
E/E_0 = mass fraction of the material throughput traversing the reactor with a residence time between t and $t + dt$; C_A = concentration of component A; t = time; τ = mean residence time; x = spatial coordinate; L = reactor length; n = order of the reaction

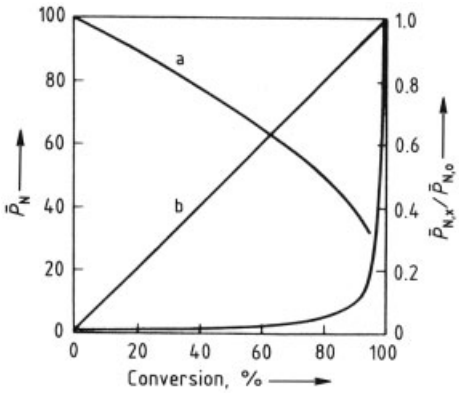

Figure 62. Cumulative degree of polymerization $\bar{P}_N$ as a function of conversion for various polymerization reactions a) Monomer coupling with termination; b) Monomer coupling without termination; c) Polymer coupling

Batch and plug flow reactors provide equivalent results, so they can conveniently be considered together. There are thus nine different combinations possible for the three types of polymerization and the three ideal reactor types, as summarized in Table 5, which also includes a qualitative description of the resulting molecular mass distributions.

3.3.6.1.1. Monomer Coupling with Bimolecular Termination Plug Flow and Batch Reactors (CPFR/BR)

Under the assumption that R_1 = constant (constant radical generation rate; half-life $t_{1/2}$ of the initiator $\gg$ residence time τ of reactor contents) and [M] = constant (small conversion increment), the instantaneous degree of polymer-

Table 5. Molecular mass distribution for polymerization reactions in various types of reactors

Reaction	Reactor		
	Continuous plug flow (CPFR) or batch (BR) reactor	Homogeneous continuous stirred-tank reactor (HC-STR)	Segregated continuous stirred-tank reactor (SCSTR)
1. Monomer coupling with termination (e.g., radical polymerization)	1.1. distribution increasingly broader than Schulz–Flory distribution with increasing conversion	1.2. Schulz–Flory distribution PDI = 2 for $k = 1$ PDI = 1.5 for $k = 2$	1.3. increasingly broader distribution than 1.1 with increasing conversion
2. Monomer coupling without termination (e.g., anionic living polymerization)	2.1. Poisson distribution for $K_I \geq K_p$ Gold distribution for $K_I < K_p$	2.2. Schulz–Flory distribution	2.3. Schulz–Flory distributon for $x = 0$ Poisson distribution for $x = 1.0$
3. Polymer coupling (e.g., polycondensation)	3.1. Schulz–Flory distribution	3.2. much broader than the Schulz–Flory distribution	3.3. between 3.1 and 3.2

ization P_N, when termination is exclusively by disproportionation, is given by

$$P_N = \frac{K_p\,[M]}{(K_{td} R_I)^{1/2}} = \frac{1}{(1-\varphi)} \tag{3.84}$$

and the instantaneous mass chain length distribution is given by the Schulz–Flory (or most probable) distribution [872], [873]:

$$W(r) = (1-\varphi)^2\, r\varphi^{r-1}$$
$$\approx (1-\varphi)^2\, r \exp = [-(1-\varphi)\, r] \tag{3.85}$$

where

$$\varphi = \frac{k_p\,[M]}{k_p\,[M] + (K_{td}\, R_I)^{1/2}}$$

The breadth of the distribution is described either by the nonuniformity U or by the polydispersity index PDI.

$$U = \frac{P_W}{P_N} - 1 \tag{3.86}$$

$$\overline{\mathrm{PDI}} = \frac{P_W}{P_N} = \frac{Q_2 Q_0}{Q_1^2} \tag{3.87}$$

with PDI = 2 for $k = 1$ (bimolecular termination by disproportionation); PDI = 1.5 for $k = 2$ (bimolecular termination by combination). The PDI (or U) can be expressed in terms of the moments of the frequency distribution of chain lengths Q as shown in Equation (3.87). Thus, if one examines the entire range of conversions (i.e., dropping the assumption that [M] = constant) what results with increasing conversion x (decreasing monomer concentration [M]) is a set of Schulz–Flory distributions with different P_N for decreasing degrees of polymerization, as shown in Figure 63.

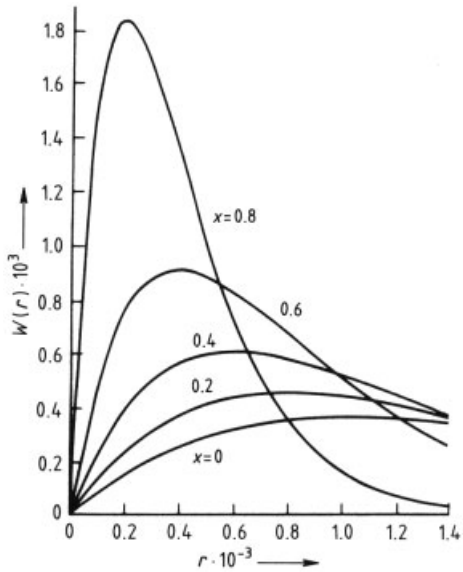

Figure 63. Mass distribution $W(r)$ of instantaneously formed polymer for monomer coupling with termination. Batch reactor at small conversion increments, or homogeneous continuous stirred-tank reactor.
$P_{N,0} = 1000$, $P_{N,x} = P_{N,0}(1-x)$; degree of coupling $k = 1$

Note that PDI = 2 for all the molecular mass distributions (actually mass chain length distributions to be more precise) in Figure 63. By contrast, the statistical variance σ^2 of the associated

frequency distribution decreases with increasing conversion. However, since $\sigma^2 = P_N^2$ (PDI − 1), σ^2 continuously decreases as P_N decreases with conversion, except in the case of a monodisperse MWD with PDI = 1. For this reason, one of the quantities PDI or U is preferred for expressing distribution breadth.

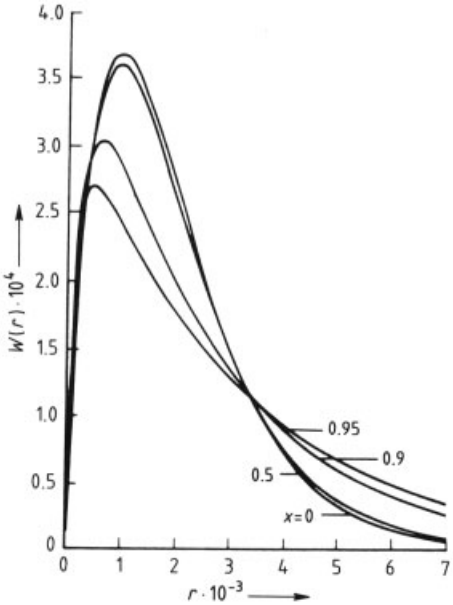

Figure 64. Cumulative mass distribution $W(r)$ for polymer formed throughout the conversion interval $x = 0$ to $x = x$ for monomer coupling with termination in a BR/CPFR $\bar{P}_{N,x} = 1000$, $k = 1$

Integration of the distributions in Figure 63 with appropriate weighting factors gives the molecular mass distribution of the final polymer product at the end of the batch [874–877]. To simplify comparison, the distributions in Figure 64 have been calculated in such a way that for each conversion x, the degree of polymerization of the accumulated polymer $\bar{P}_N = 1000$. It is clear that distributions broaden with increasing conversion. Representations of PDI and $\bar{P}_N$ as functions of conversion follow in Figures 68 and 69. Theoretical treatments of this case can be found in [878–888] and in Section 2.2. Before leaving CPFR/BR reactors, the following points should be noted. Firstly, the Trommsdorff effect can have a significant effect on molecular mass development with increasing conversion. The significant lowering of either K_{tc} or K_{td} when chain transfer to small molecules (monomer or chain-transfer agents) is negligible can cause dramatic increases of molecular mass with conversion. See, for example, data for the change of molecular mass distribution with conversion for the bulk homogeneous polymerization of methyl methacrylate [380]. On the other hand, when chain transfer to monomer (as in vinyl chloride polymerization) or to chain-transfer agent produces most of the polymer chains, the Trommsdorff effect has no influence on molecular mass development. In particular, when chain transfer to monomer produces most of the polymer chains, molecular mass development will depend on temperature and be independent of process type and reactor type [889].

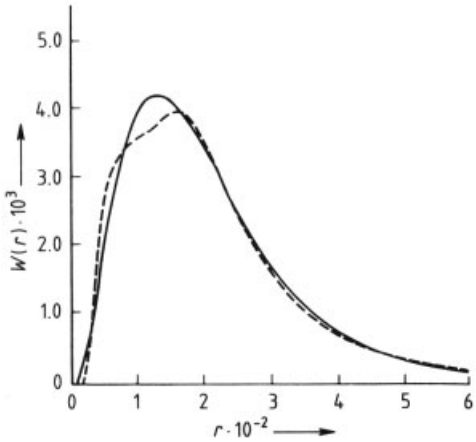

Figure 65. Radical polymerization of styrene in an HCSTR at low viscosity. Comparison between observed (– – –) and calculated (—-) molecular mass distributions [891]
Reaction conditions:
Feed: 2.39 mol/L monomer, 7.44 mol/L solvent (benzene), 0.0152 mol/L initiator (AIBN); temperature 74.1 °C; viscosity $\eta = 1.5$ mPa · s; mean residence time $\tau = 160$ min

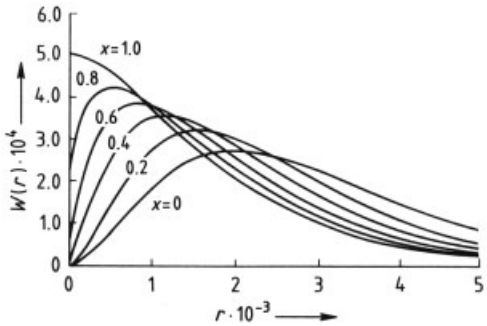

Figure 66. Mass distribution $W(r)$ of polymer formed by monomer coupling with termination in an SCSTR [883] $P_{N,0} = 2000$; $k = 2$

In a CPFR/BR, the width of the copolymer composition distribution can be substantial if one monomer type reacts much faster than the other in propagation. For the same final conversion, the residence-time distribution for CPFR/BR is ideal compared to the other reactor types when minimization of long-chain branching, cross-linking and gel/sol ratios is required.

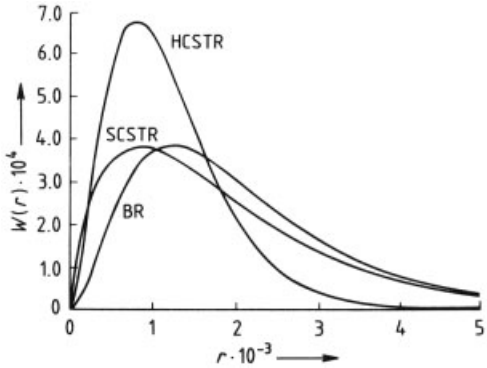

Figure 67. Mass distribution for radical polymerization in three different reactors [885]
Conversion $x = 0.6$; $P_{N,0} = 1000$; $k = 2$

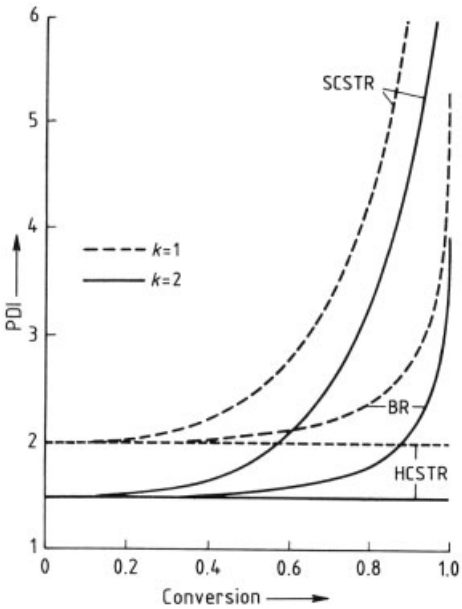

Figure 68. Polydispersity index PDI = P_W/P_N as a function of conversion for radical polymerization in three different reactors

Homogeneous Continuous Stirred-Tank Reactor (HCSTR). DENBIGH has shown that in a homogeneous CSTR two opposing factors influence the molecular mass distribution [878–881]. The constancy of [M] at steady state leads to a narrowing of the distribution relative to that obtained in a batch reactor, while the range of residence times leads in principle to broadening of the distribution. However, in radical polymerization, the lifetime of a growing radical is ca. 1 s, very small compared to the mean residence time of any industrial reactor. This first factor (constant [M]) thus predominates and the resulting distribution is narrower in an HCSTR than in a CPFR/BR. With polymerization reactions that do not involve termination, the lifetime of the growing chain is comparable to the mean residence time of the reactor, so that broadening of the MWD by the residence-time distribution is dominant.

In the steady state with [M] = constant, the same conditions are present in the stationary state as in a batch reactor at low conversion or at some higher conversion over a small conversion increment, resulting in the Schulz – Flory distribution described by Equations (3.85) – (3.87). A comparison of observed and calculated molecular mass distributions for an HCSTR, as shown in Figure 65, indicates very good agreement in regions of low viscosity where the Trommsdorff – Norrish effect plays no significant role. Even when the Trommsdorff – Norrish effect is appreciable, the Flory – Schulz distribution should still apply when diffusion-controlled bimolecular termination is not chain-length dependent [358]. Additional experimental studies on continuous free-radical polymerization are described in [367], [373], [890–895].

Experimental	Theoretical
$\bar{M}_W = 20\,990$	$M_W = 20\,440$
$\bar{M}_N = 14\,020$	$M_N = 13\,650$
PDI = 1.50	PDI = 1.50

Segregated Continuous Stirred-Tank Reactor (SCSTR). Individual compartments, all of which can be regarded as miniature batch reactors, display residence times or reaction times according to the exit-age distribution function of a CSTR which is

$$E(t) = \frac{1}{\tau}\exp(-t/\tau) \tag{3.88}$$

The effect of segregation is a broadening of the MWD to a breadth greater than that for a batch reactor, and the breadth increases with conversion. Figure 66 provides several examples, starting with the Schulz – Flory distribution at a conversion $x = 0$. Figure 67 compares molecular mass distributions for free-radical polymerization in the three reactor types. The corresponding changes in the polydispersity index PDI are illustrated in Figure 68, and Figure 69 shows the degree of polymerization for the accumulated polymer as a function of conversion. Thus, the narrowest possible molecular mass distribution for free-radical polymerization is the Schulz – Flory distribution and it is obtained in an HC-STR, and the broadest distribution is obtained in a SCSTR. For additional theoretical studies, see [883], [885], [896], [897].

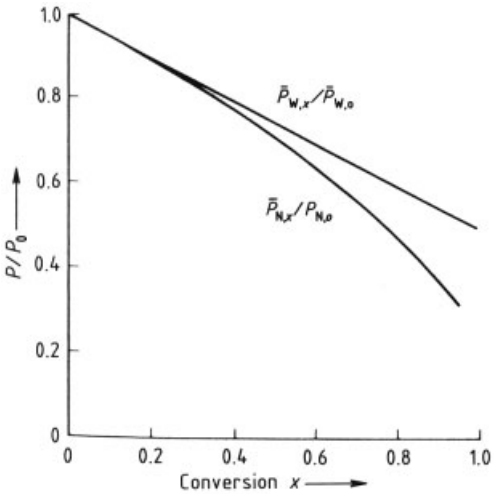

Figure 69. The course of the degrees of polymerization of the accumulated polymer $\bar{P}_N$ and $\bar{P}_W$ as a function of conversion during radical polymerization

3.3.6.1.2. Monomer Coupling Without Termination Plug Flow and Batch Reactors (CPFR/BR)

In the absence of termination reactions the kinetic chain length η corresponds to the ratio of the initial concentrations of monomer and initiator

$$\eta = \frac{[M]_0}{[I]_0} \tag{3.89}$$

or at incomplete conversion

$$\eta = \frac{[M]_0 - [M]}{[I]_0} = \frac{[M]_0\, x}{[I]_0} \tag{3.90}$$

If initiator molecules are counted as chain units

$$P_N = 1 + \eta \tag{3.91}$$

The assumption is made here that initiation is much faster than propagation, i.e., $K_i \gg K_p$. At $t = 0$, all active molecules will have a degree of polymerization $P_N = 1$, and the number of propagation steps per active site (i.e., the kinetic chain length) is $\eta = P_N - 1$. The case $K_i < K_p$ leads to a broader distribution [898], [899], but one that is still narrower than the Schulz – Flory distribution. The molecular mass distribution corresponds to a Poisson distribution [435, pp. 346 – 339], [900] (see also Section 2.2.2.2):

$$W(r) = \frac{e^{-\eta}\,\eta^{r-1}}{(1+\eta)\,(r-1)\,!} \tag{3.92}$$

with a polydispersity

$$\mathrm{PDI} = P_W/P_N = 1 + \frac{1}{\eta} \tag{3.93}$$

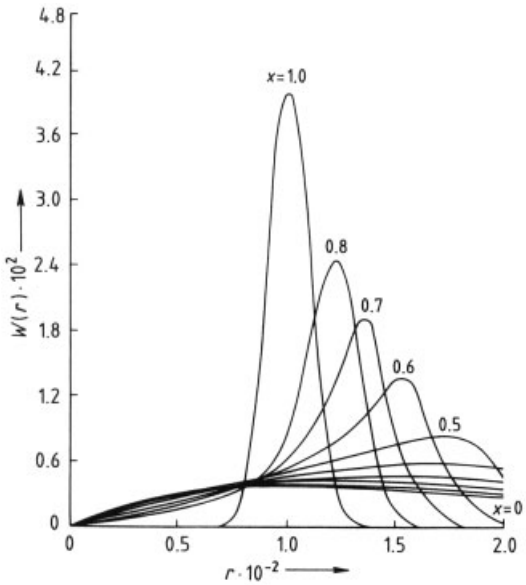

Figure 70. Mass distribution $W(r)$ of polymer formed by monomer coupling without termination in a SCSTR [885] $P_N = 100$; Schulz – Flory distribution for $x = 0$; Poisson distribution for $x = 1$

Thus the distribution becomes narrower as the contribution from statistical broadening decreases with chain growth (increase in η) (cf. the bottom curve in Fig. 71). A comparison of Equations (3.87) and (3.93), or inspection of Figure 70 indicates that Poisson distribution is significantly narrower than the Schulz – Flory distribution.

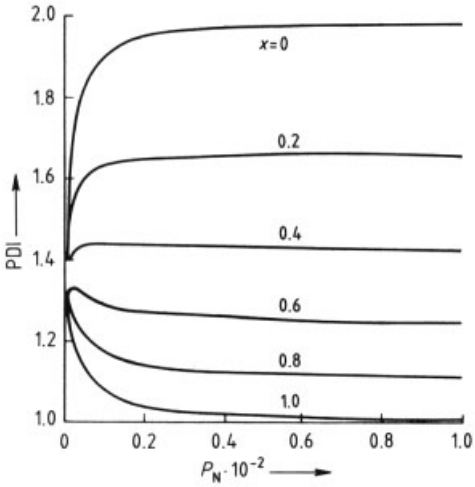

Figure 71. Polydispersity index PDI = P_W/P_N as a function of P_N for polymerization by monomer coupling without termination in a SCSTR [885]

Homogeneous Continuous Stirred-Tank Reactor (HCSTR). If it is assumed that chain growth terminates abruptly as growing chains leave the reactor, then the lifetime of an active growing chain is equal to its residence time in the reactor. According to [879–881], the molecular mass distribution in an HCSTR should be broader than for a CPFR/BR. Several authors have demonstrated that the Schulz – Flory distribution is obtained for living polymerization (with instantaneous initiation) in an HCSTR [882], [883], [885], [901–904].

Segregated Continuous Stirred-Tank Reactor (SCSTR). The rate of living polymerization is first-order with respect to [M], so R_p and x are the same for both HCSTR and SCSTR [795, pp. 309 – 331], [805]. Segregation inhibits total molecular mixing, whereas in an HCSTR all molecules are distributed according to Equation (3.88), this is true for a SCSTR only with respect to all compartments, each of which contains a large number of molecules with equal residence times. The molecular mass distribution is therefore subject to less broadening relative to a batch reactor than in the case of the HCSTR. At very low conversion, it is possible to assume constant monomer concentration, so the distributions in the three reactor types are almost identical. At high conversion the effect of the residence-time distribution diminishes. At complete conversion the lifetime of the active species must be significantly shorter than the mean residence time, and the molecular mass distribution becomes equivalent to that in a batch reactor. Figure 70 shows the transition from a Schulz – Flory to a Poisson distribution with increasing conversion. Figure 71 is a plot of polydispersity PDI as a function of P_N and x. The bottom curve (for $x = 1.0$) corresponds to Equation (3.93) for a batch reactor.

3.3.6.1.3. Polymer Coupling

Only the formation of linear chains, either from monomer of type A – B or from exact stoichiometric mixtures of monomer types A – A and B – B is considered here. Ring formation is also excluded.

Plug Flow and Batch Reactors (CPFR/BR). The functional group conversion p is defined as the fraction of functional groups that have reacted at a given time

$$p = \frac{N_0 - N}{N_0} \tag{3.94}$$

For the degree of polymerization it follows that

$$P_N = \frac{N_0}{N} = \frac{1}{(1-p)} \tag{3.95}$$

$$P_W = \frac{1+p}{1-p} \tag{3.96}$$

The molecular mass distribution is given by

$$W(r) = (1-p)^2\, r p^{r-1} \tag{3.97}$$

with

$$\mathrm{PDI} = P_W/P_N = 1+p \tag{3.98}$$

and is once again the Schulz – Flory distribution with PDI = 2 at high conversions [435,

pp. 336 – 339], [873], [882], [883], [902], [905–908].

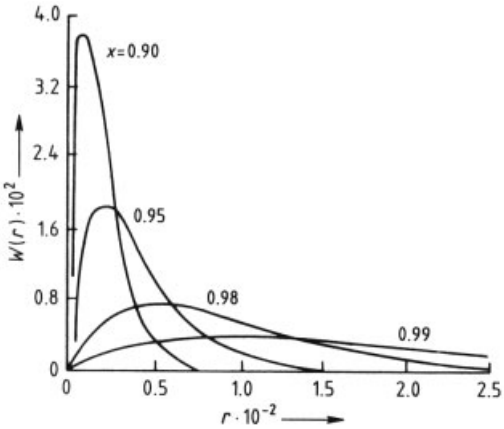

Figure 72. Molecular mass distribution $W(r)$ of polymer formed by polymer coupling in a CPFR/BR [873]

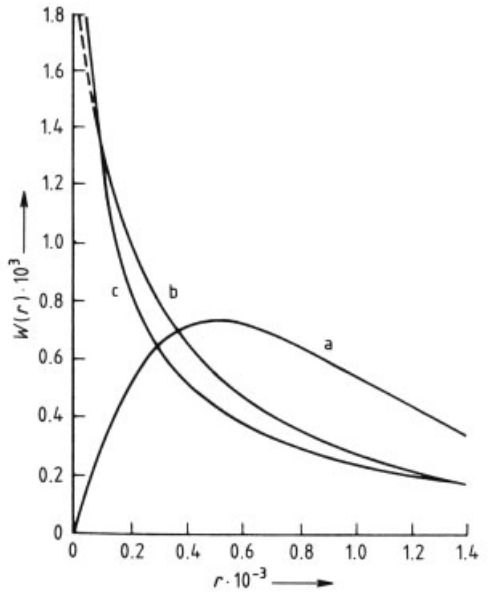

	p	P_N	P_W	PDI
a) BR	0.998	500	999	1.998
b) SCSTR	0.9887	88.41	1000	11.31
c) HCSTR	0.9561	22.8	995	43.6

Figure 73. Molecular mass distribution $W(r)$ for polymerization by polymer coupling in three different reactors [883]

Figure 72 shows the molecular mass distribution at various conversions. For free-radical polymerization the Schulz – Flory distribution always has a PDI = 2 because φ, the probability of propagation, is almost always very close to unity. This is not so for polycondensations, where p can vary from 0 to 1, and thus PDI varies from 1 to 2.

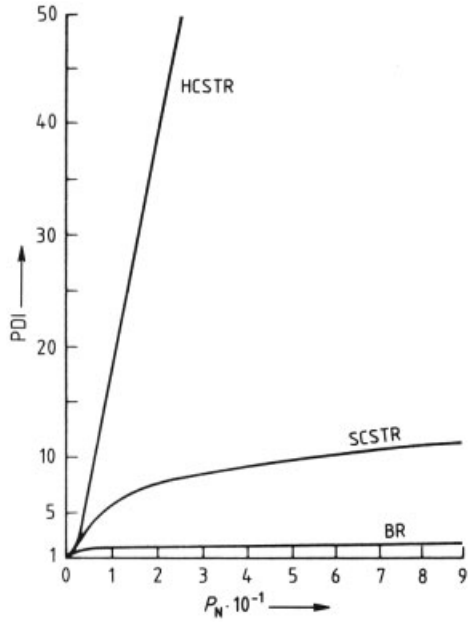

Figure 74. Polydispersity index PDI = P_W/P_N as a function of P_N for polymer coupling [885]

Homogeneous Continuous Stirred-Tank Reactor (HCSTR). Monomer and polymer molecules are capable of reacting with each other throughout their entire residence time in the reactor. Thus, according to the principle of Denbigh [879–881], a broader molecular mass distribution is to be expected in an HCSTR than in a batch reactor. The broad distribution of residence times means that material leaving the reactor always contains a relatively large amount of monomer and polymer with a low degree of polymerization. On the other hand, with increasing molecular mass the probability of the formation of very large macromolecules by the coupling of two smaller ones increases greatly. Calculations give a very broad distribution (Fig. 73) in which the number-average degree of polymerization P_N is the same as in a batch reactor, but the weight-average P_W is much higher [883], [885], [902], [907], [909], [910]:

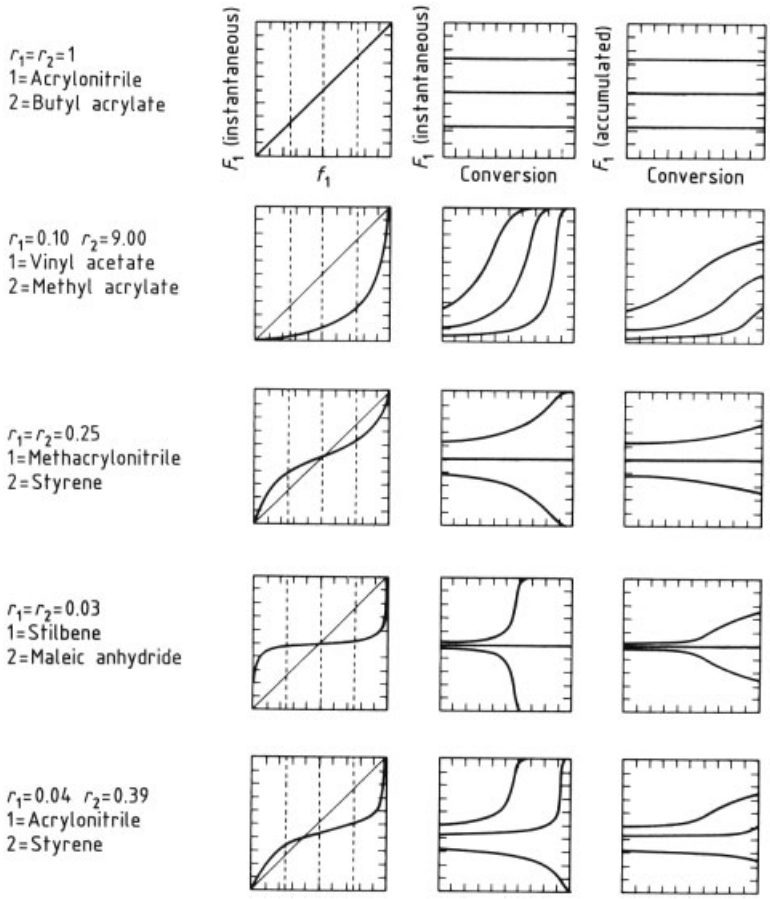

Figure 75. Composition of the copolymer produced in a batch reactor (BR) as a function of monomer composition of the charge as well as conversion
First column: instantaneous composition; second column: instantaneous compositions, starting with monomer ratios $[M_1]:[M_2] = 1:3$, $1:1$, and $3:1$; third column: cumulative compositions based on the same starting ratios

$$P_N = \frac{1}{1-p}$$

$$P_W = \frac{1+p^2}{(1-p)^2} \tag{3.99}$$

This corresponds also to a considerably higher PDI, which increases rapidly with increasing conversion (Fig. 74):

$$\mathrm{PDI} = \frac{1+p^2}{(1-p)} \tag{3.100}$$

Segregated Continuous Stirred-Tank Reactor (SCSTR). Segregation reduces the amount of polymer with very high or very low molecular mass [883], [885]. Figure 73 compares the molecular mass distributions expected for the three reactor types and Figure 74 shows the polydispersity index PDI as a function of P_N. In terms of both P_N and PDI, the SCSTR lies between the CPFR/BR and the HCSTR.

Note that the conclusions drawn in the preceding two sections are a result of purely theoretical considerations, and they actually have little to do with practice. On the one hand, Equation (3.95) (see also Fig. 62) indicates that attaining a reasonably high degree of polymerization requires a very high conversion; HC-

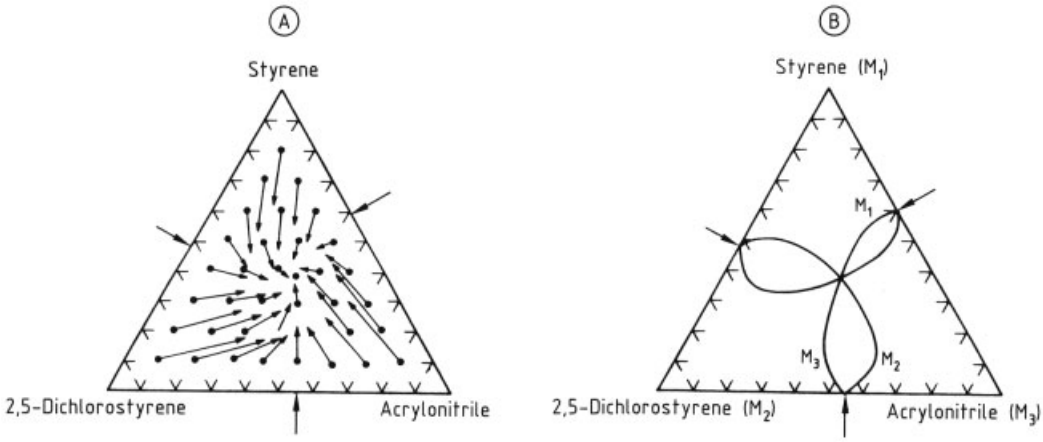

Figure 76. Ternary copolymerization in the system styrene (M_1), 2,5-dichlorostyrene (M_2), and acrylonitrile (M_3) [912]
A) Instantaneous copolymer composition (arrow point) as a function of composition of the monomer mixture (origin of the arrow); B) Partial azeotropes, where arrows outside the diagram indicate the positions of binary azeotropes

STR and SCSTR systems would therefore appear inappropriate (extreme viscosities would be inappropriate for a CSTR). On the other hand, many polycondensations and polyadditions are accompanied by rearrangements at the heteroatoms, including transesterifications or transamidations. This means that polymer with a Schulz – Flory distribution would be obtained even in an HCSTR [435, pp. 336 – 339], [911].

3.3.6.1.4. Copolymerization

The theory of copolymerization kinetics is covered in Section 2.3. However, the performance of the three reactor types (CPFR/BR, HCSTR, and SCSTR) were not compared and therefore this topic is considered here.

Copolymerization in CPFR/BR Reactors. Figure 75 gives several illustrations of compositional drift for binary copolymerization for various comonomer pairs in continuous plug flow and batch reactors. The first two columns are for the mole fraction of monomer 1 chemically bound in copolymer chains produced instantaneously (over a small conversion or time interval) as a function of mole fraction of free monomer 1 and of conversion. The third column shows the mole fraction of monomer 1 in the accumulated copolymer chains versus conversion. Azeotropes also occur in systems with more than two comonomers. Figure 76 shows a three-fold combination represented in terms of triangular coordinates. The compositions of various comonomer mixtures are indicated by the origins of arrows, with the corresponding arrow points showing the instantaneous compositions of the resulting polymers. The length of an arrow is thus a measure of the deviation between compositions of free monomer and copolymer. Such an arrow degenerates to a point in the case of an azeotropic composition. A phenomenon known as "partial azeotropy" is of interest in the context of multicomponent systems, with only one of the monomers represented by an identical mole fraction in both copolymer and monomer mixture. Thus, for ternary copolymerization

$$\frac{m_1}{m_1+m_2+m_3} = \frac{[\mathrm{M}_1]}{[\mathrm{M}_1]+[\mathrm{M}_2]+[\mathrm{M}_3]} \tag{3.101}$$

A systematic computer search of the literature has produced copolymerization parameters for 37 ternary azeotropes, 4 quaternary azeotropes, and one 5-component azeotrope [912]. Figure 77 shows copolymer composition distributions (CCDs) for the monomer pairs and compositions treated in Figure 75. Copolymer of uniform composition can be expected during a batch polymerization only at the azeotropic composition, or at extremely low conversions. Strictly speaking, a copolymer is not uniform at all molecular masses even under azeotropic conditions. What is in fact produced is more nearly a

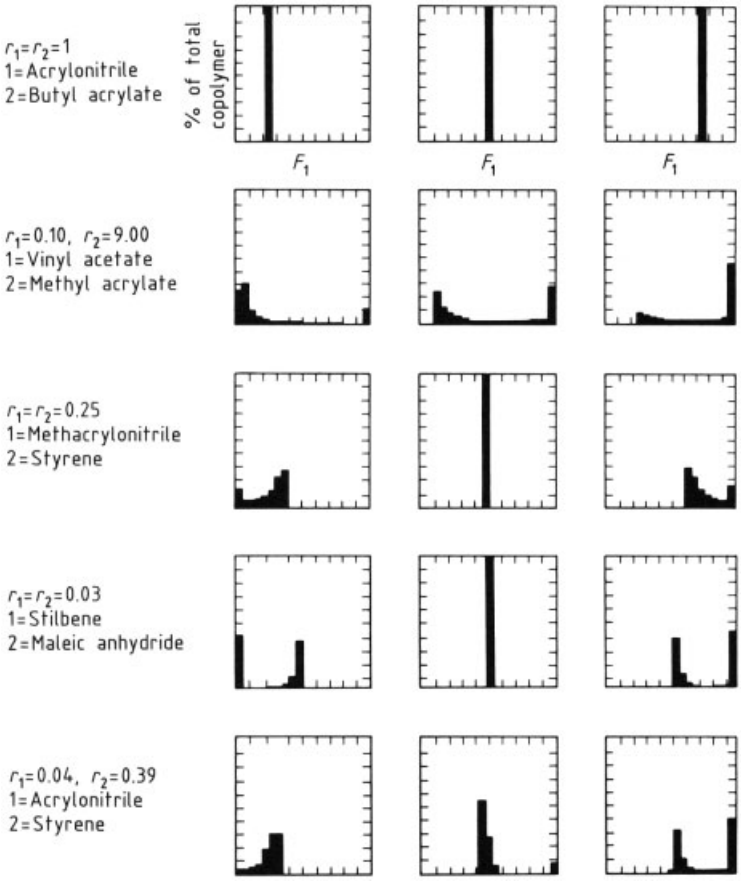

Figure 77. Copolymer composition distributions (CCDs) for the monomer pairs of Figure 75. CCDs of completely polymerized batches with molar ratios $[M_1]:[M_2] = 1:3$ (first column), 1 : 1 (second column), and 3 : 1 (third column)
The mean composition of the copolymer corresponds to that in the batch. The height of each block represents that fraction of the overall copolymer that has a composition within the block width of 0.05
Experimental data and calculated values as a function of f_1; 60 °C, [I] = 1 g/L azobisisobutyronitrile

statistical distribution about a mean [913], [914], although at high degrees of polymerization this variation is minor in comparison to the heterogeneity that results from progressive changes in the composition of the monomer mixture with conversion.

Copolymerization in an HCSTR. In an HCSTR operating at steady state the temperature and concentration are constant in both space and time. The result is therefore a chemically uniform copolymer. In the limiting case of very low conversion, the copolymer composition corresponds to the predictions of Equation (2.121) and Figure 75 when concentrations of monomer in the inlet stream are used to calculate monomer composition. The only other situation in which the composition of the copolymer corresponds precisely to the composition of the monomer mixture in the feed is the technically unrealistic limit of complete conversion, which would require impractically long residence times. The instantaneous copolymer equation, of course, can always be used when monomer composition in the exit stream is used (as opposed to inlet stream) [915], [916]. Figure 78 shows the change in steady-state copolymer composition for the system styrene/acrylonitrile (40/60) as a function of conversion. Thus, a copolymer with a particular composition can be made in practice (at least approximately) either by establishing in advance a preferred composition for

the monomer feed and then varying the extent of conversion, or by varying the feed composition at a fixed conversion.

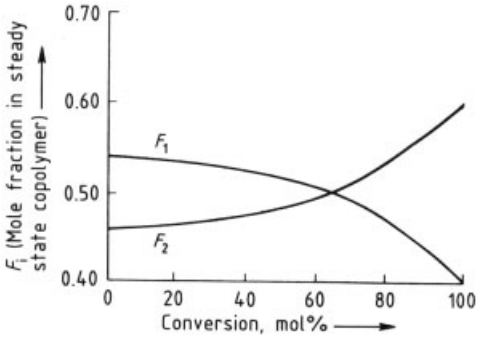

Figure 78. Continuous copolymerization in the system styrene ($f_1 = 0.4$) / acrylonitrile ($f_2 = 0.6$) [915]
Composition of accumulated copolymer F_i as a function of conversion

Copolymerization in a SCSTR. It was noted earlier that true mixing is often less complete than envisioned by the HCSTR concept in a polymerization reactor operating under conditions of high viscosity. Most industrial reactors probably provide an efficiency of mixing that falls between the values assumed for the HCSTR and SCSTR. By introducing certain simplifying assumptions, O'Driscoll and Knorr [917] were able to calculate CCDs corresponding to the three reactor models CPFR/BR, HCSTR, and SCSTR. Figure 79 A shows a representation of a typical instantaneous copolymer composition diagram. Figure 61 B shows CCD curves for the three ideal reactor types at two conversions. The feed composition is fixed at $f_1 = 0.4$ for all cases shown. Open circles represent mean copolymer compositions F_1 which are not very different in the three cases. The differences in the CCDs are greater, however. At 35 % conversion the CCD is relatively narrow in a CPFR/BR (curve 1), absolutely uniform in an HCSTR (curve 2), and already quite nonuniform in a SCSTR (curve 3). In the latter, more than 10 % of the polymer is homopolymer of monomer 2. A SCSTR always provides several batch-reactor-like compartments with very long residence times, high conversions, and extreme changes in monomer and copolymer compositions. At 73 % conversion the CCD shows nearly as much nonuniformity with a CPFR/Batch (curve 4) as with a SCSTR (curve 6), only the HCSTR (curve 5) again producing a uniform CCD. The amount of homopolymer produced is, of course, greater at 73 % than at 35 % conversion.

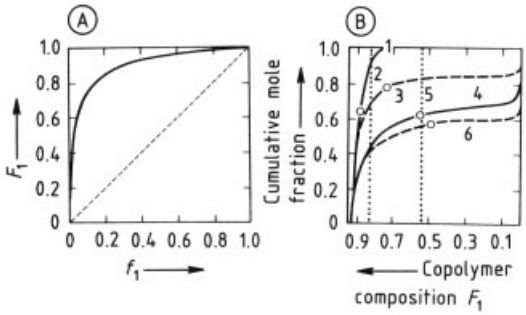

Figure 79. Copolymerization of methyl methacrylate (M_1) – vinyl acetate (M_2) in three types of reactor: BR, HCSTR, and SCSTR [917]
A) Instantaneous composition, calculated with $r_1 = 20$, $r_2 = 0.015$; B) Copolymer composition distribution (CCD) at $f_1 = 0.4$, 35 and 73 % conversion in a BR (curves 1 and 4), a HCSTR (2 and 5), and a SCSTR (3 and 6)

Curve:	1	2	3	4	5	6
Conversion x:	0.35	0.35	0.35	0.73	0.73	0.73

3.3.6.1.5. Long-Chain Branching and Cross-Linking

The amounts of long-chain branching and cross-linking (with gel formation) obtained in free-radical polymerizations depend strongly on the residence-time distribution of the reactor and the conversion (or polymer concentration). Compared to the HCSTR, the CPFR/BR gives less long-chain branching, cross-linking (cross-linking density), and gel for the same conversion. Apparenty, calculations on long-chain branching and cross-linking in a SCSTR have not been performed; however, one could speculate that the extent of long-chain branching in a SCSTR would lie between that for a CPFR/BR and HCSTR. Graessley et al. [918] have carried out theoretical modeling and experiments for CPFR/BR and HCSTR using vinyl acetate homopolymerization. Their experimental polydispersity data are shown in Figure 80 for CPFR/BR and HCSTR. Note that long-chain branching is due to chain transfer to polymer and addition of polymeric radicals to terminal double bonds on polymer chains and that such branching leads to increases in the polydispersity index starting at 2.0 for linear chains produced at

low conversion (in other words, the larger the PDI, the higher the long-chain branching frequency). A survey by REICHERT and MORITZ [918] covers the effect of reactor type as well as the effect of changes in the properties of the reacting mixture on the polymerization processes. For example, the reaction mixture viscosity may increase by several orders of magnitude with conversion and this may affect kinetic rates (e.g., Trommsdorff – Norrish effect, reduction in initiator efficiency, and propagation rate constant), as well as processes of heat, mass, and momentum transfer, the quality of mixing and the residence-time distribution of the continuous process. They also point out that the reactor type (RTD and degree of segregation) can significantly affect long-branching and cross-linking levels in free-radical polymerization.

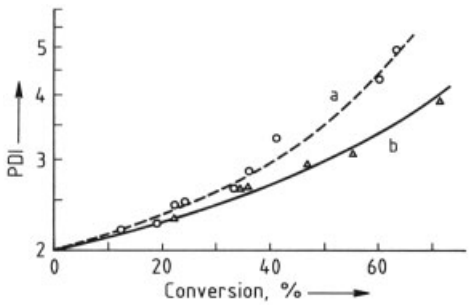

Figure 80. Polydispersity index PDI = P_W/P_N as a function of conversion for batch and CSTR bulk polymerization of vinyl acetate
a) Continuous; b) Bulk

A semi-batch process has been investigated using a comprehensive kinetic model for long-chain branching and cross-linking for the purposes of producing homogeneous polymeric networks [356], [363–365], [919], [920]. In the model, "the cross-linking/branching density distribution" was proposed. The use of the distribution enables one to calculate MWD irrespective of the reactor type [921]. This model has been successfully applied to various batch copolymerizations involving methyl methacrylate – ethylene glycol dimethacrylate and acrylamide –*N,N′*-methylene bisacrylamide. It has been shown that the variance of the cross-link density distribution is reduced with the use of a chain transfer agent. BOOTS [922] confirmed the effect of shorter primary chains qualitatively using a Monte Carlo simulation.

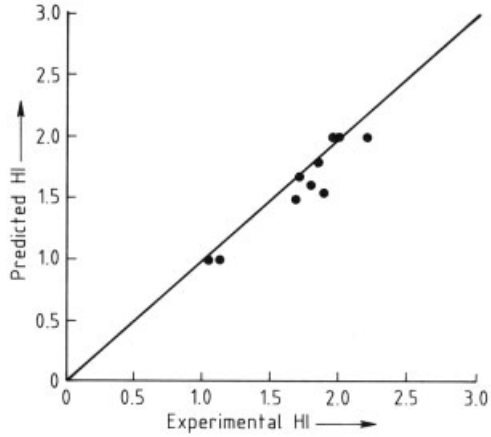

Figure 81. Anionic (living) polymerization of butadiene showing measured versus model-predicted heterogeneity index HI (polydispersity index PDI) [923]

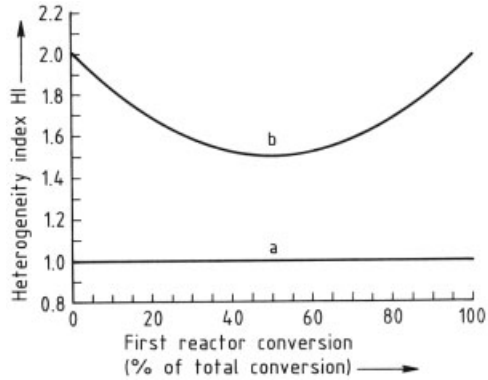

Figure 82. Anionic (living) polymerization of butadiene showing the effect of RTD on heterogeneity index or breadth of MWD [923]
a) Batch/plug flow reactor; b) 2 CSTRs

The MWDs produced by living anionic polymerization, as shown earlier, are strong functions of RTD. CHANG et al. [923] used changes in RTD to modify MWD in butadiene polymerization in experiments involving a reactor train, while MEIRA et al. [924–930] used periodic operation (in residence time) to broaden the MWD as well as controlled addition of killing agent in semi-batch operation. CHANG et al. [923] developed a comprehensive model for the anionic solution polymerization of butadiene with *n*-butyllithium initiator, tetramethylethylenediamine modifier, and hexane solvent. The model successfully predicted monomer conversion, vinyl structure, and polydispersity for both batch reactors and continuous reactor trains.

Figure 81 shows predicted versus measured heterogeneity index (polydispersity index or M_W/M_N), covering the range from that obtained in a PFR or batch reactor to that obtained in a single CSTR. Figure 82 shows the polydispersity for 2 CSTRs of equal volume in series with equal conversion of monomer in each, giving the lowest value of polydispersity, 1.5.

3.3.6.2. Reactor Dynamics and Stability

Reichert and Moritz [918] discuss reactor stability and safety under the subtopics thermal and concentration stability. The strong nonlinear dependence of polymerization rate (or heat production rate) on temperature is due to the Arrhenius dependence of rate on temperature and to the large activation energy and heat of polymerization. The dynamic behavior of polymer reactors is therefore greatly affected by temperature. The influence of the nonlinear dependence of polymerization rate on monomer concentration, on reactor dynamics is much smaller than that for thermal effects. Both static and dynamic instabilities may occur when stability limits are crossed. The reactor is statically unstable if temperature and conversion rapidly move from one state to another as if alternately experiencing "ignition" and "extinction". Dynamic instabilities give rise to periodic changes in temperature and monomer conversion with certain phase shifts observed [931], [932]. Thermal instabilities may occur in various reactor types (CSTR, PFR, nonideal reactors) [933], [934]. Henderson [935] points out that for all prior publications on multiple steady states and instability for free radical polymerization in a CSTR, three basic assumptions were made:

1) The heat-transfer coefficient of the vessel wall does not change with monomer conversion
2) The viscous dissipation term for the agitation is negligible
3) Perfect mixing (ideal RTD and micromixing)

Mechanical work of the agitator can be a significant factor in the energy balance for a pilot-scale reactor (perhaps for solution and bulk polymerization, but not for suspension and emulsion polymerization), while it is almost negligible for commercial scale. Assumptions (2) and (3) may be valid for low-viscosity solution processes; however, in bulk polymerization, the viscosity can increase by several orders of magnitude and, therefore, all 3 assumptions are questionable. Henderson again points out that a thermal runaway can occur when:

1) The process side heat-transfer coefficient decreases significantly
2) The agitator provides insufficient mixing of the reaction mass, allowing hot spots to develop
3) The agitator is putting more mechanical work into the system than designed for

Ray et al. [788–791], [936–945] have made the most significant contributions to the understanding of polymer reactor dynamics and stability. Comprehensive reactor models were developed for a range of polymerization processes and reactor types: nonisothermal solution homo- and copolymerization in a CSTR [931], [939], [940]; batch and continuous emulsion polymerization reactors [788–791], [941], [942]; heterogeneous coordination polymerization for both liquid and gas dispersion reactors [943–945]. The contributions by Ray et al. are notable not only for the comprehensive mathematical models but also for attempts at experimental verification. In this regard, the experimental investigation of the effect of chain-transfer agents on the stability of continuous latex reactors is very revealing. Whereas under most operation conditions the continuous emulsion polymerization of styrene is stable, the addition of a chain-transfer agent to the recipe can lead to sustained oscillations. Some of Ray's general conclusions include:

1) Bifurcation analysis reveals a parameter space rich in dynamic phenomena. It confirms limit cycle behavior (in which monomer conversion does not smoothly approach a constant steady-state value, but rather oscillates with constant frequency and amplitude indefinitely) and analyzes the stability of periodic branches in detail providing useful guidelines for experimental design.
2) Full-scale reactors exhibit dynamic behavior of high complexity.
3) Semi-batch reactors can be operated under conditions that insure a fast approach to a

steady-state operating value, hence permitting almost perfect control. The possibility of oscillatory behavior was, however, demonstrated.

4) Multiplicity (multiple steady states) occurs under isothermal conditions due to the Trommsdorff – Norrish effect and there appears to be no relationship between phenomena causing multiplicity and reactor oscillations.
5) The PSD plays a major role in latex reactor stability. Models which calculate only a mean diameter d (and not the full PSD) cannot predict oscillations under any operating conditions.
6) Polymer particle multiplicity, which leads to polymer melting and sticking (in gas-phase heterogeneous coordination polymerization, such as UNIPOL) can cause a thermal runaway of the reactor (UNIPOL is open-loop unstable). Multiplicity and Hopf bifurcation of the fluidized-bed reactor can lead to runaway and long periods of slowly damped oscillations. The reactor is prone to runaway even under feedback control because of cooling limitations and nonlinear sensitivity to process disturbances.

Lu and Brooks [946] have confirmed that latex reactor startup policy can have a significant effect on reactor stability.

It is clear, based on these studies of reactor stability, that a comprehensive dynamic model of the polymerization reactor is required to properly design a continuous manufacturing process (process stability, control, and startup procedures should be considered at the design stage). The interactions are too complex to permit an effective design based on costly experimentation alone.

3.3.6.3. On-Line Monitoring and Control

The goals in commercial polymerizations are consistent polymer properties, high productivity, and safe operation achieved in the most economic manner possible. Polymerizations (particularly ionic and free-radical) are very sensitive to temperature and impurities (radical scavengers and poisons at ppm levels) and, therefore, adjustments may have to be made during polymerization. Process models when applied in real time require measurements and parameter updating to track and control the process. The need for monitoring the polymerization is clear and to be most effective in controlling the process, on-line monitoring is highly desirable. A survey of on-line sensors for polymerization reactors was made by Chien and Penlidis [947]. Although considerable advances have been made in polymer reaction engineering over the past two decades, sensor technology for on-line monitoring is still in its infancy. The major problem has been the on-line measurement of polymer and particle properties (copolymer composition, molecular mass averages, long-chain branching, as well as cross-linked gel content and polymer particle size distribution in suspension and emulsion polymerization). Sensors in the reaction mixture are readily fouled by sticky polymer and polymer particles. Another factor to explain the lack of progress is the fact that sensor development is a multidisciplinary task with expertise required in statistics, mathematical modeling, process understanding, reactor design, advanced control concepts, electronics, and instrumentation engineering [947]. Even the measurement of polymer properties off-line is difficult (e.g., measurement of molecular mass averages and long-chain-branching frequencies for heterogeneous copolymers [870]). Chien and Penlidis [947] classify sensors as those which monitor reactor operation [by measuring temperature, pressure, flows and level (of reaction mixture in the reactor)], and those that monitor polymer properties (e.g., by measuring density, viscosity, concentrations, copolymer composition, particle size, using densitometers, viscometers, gas and liquid chromatographs, IR and UV spectrophotometers, light scattering, ultrasonics, fiber-optic techniques, robotics, reactor energy balances, and state estimation/filtering techniques). Most of the commercial applications of on-line monitoring involve those which monitor reactor operation, although reactor energy balances are commonly used to monitor monomer conversion. The main problem with the successful use of reactor energy balances lies in the evaluation of the transient accumulation terms (time derivatives of temperature) and in accounting for the propagation of measurement errors (temperature and flow rate) into the calculated heat release term. Acceptable results are

often obtained through rapid sampling and then averaging or filtering all the measurements at regular short intervals [949].

Two promising on-line monitoring techniques are based on the use of ultrasound and NIRS (near infrared spectroscopy) [948]. Ultrasonic velocity and attenuation measurements have been used for monitoring high-temperature composition in extruders and melt-transfer lines; high-temperature imaging studies in extruders; and for ultrasonic spectroscopy of dispersions. NIRS is an older technique which has been improved by newly available computing power. It is the most promising process spectroscopy available at present because of its process compatability and ability to analyze multicomponent mixtures. NIRS has been applied as an on-line monitoring technique for both extruders used for chemical modification and polymer reactors.

The on-line monitoring and control of polymer and polymer particle properties should provide a major improvement in product consistency and productivity.

There have been a number of excellent reviews on control of polymerization reactors [947], [949], [950]. Most industrial control practice to date has centered around batch reactor sequencing and the control of process variables such as temperature, pressure, and viscosity. Reactor energy balances, as mentioned earlier, have also been used to infer polymer production rates and monomer conversion.

Much of the academic literature has focused on the computation of optimal operating policies for batch reactors by using nonlinear programming or the maximum principle [374]. These policies are open-loop in that they provide optimal feed and temperature policies versus time. Kozub and MacGregor [951] presented a much simpler solution to the optimal trajectory problems based on an "instantaneous property" approach. This latter approach is applicable only to chain-growth polymerizations (see Section 2.2 for details of instantaneous property calculations), but it allows for the feedback implementation of these optimal policies via nonlinear control.

Due to the lack of sufficient on-line sensors to monitor polymer properties of interest or because some property measurements are only available infrequently from quality control laboratories, various methods for inferring or estimating these properties on-line by using fundamental mathematical models have been proposed. State estimators such as the Kalman Filter [952], [953] and recursive parameter estimators [954] have been advocated to update model parameters and predict polymer properties.

Feedback control over measured or inferred properties in continuous reactors has been accomplished by using empirical models [955], [956].

However, in spite of the many advances that have been made in the methodology for the advanced control of polymerization reactors, there are still few industrial applications.

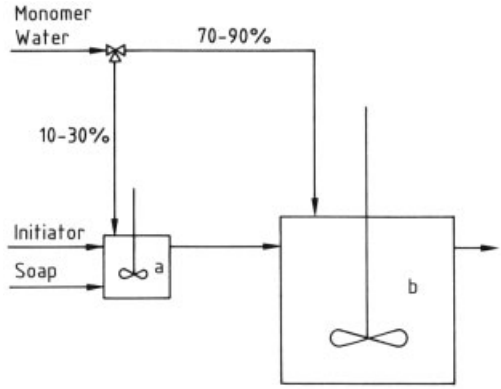

Figure 83. Continuous emulsion polymerization in a reactor train with a split feed (small polymer particle nucleating reactor followed by one or more larger finishing reactors) [956]
a) Small seeding reactor; b) First large reactor in train

To again emphasize the point that control problems should be considered at the design stage of a polymer manufacturing process, reference is made to the well-known stability problems that are experienced with continuous emulsion polymerization (in a single CSTR) involving monomer(s) that experience Case I kinetics (e.g., vinyl acetate and vinyl chloride). Modern control theory did not provide suitable control methodology for the single CSTR. It was decided, therefore to go back to the design stage, redesign the process by using a dynamic reactor model and minimize control problems at the design stage. The new design which is shown in Figure 83 has almost trivial control problems [949], [956]. Experimental verification of the increase in stability of this new design may be seen in Figure 84. The initial response (conversion

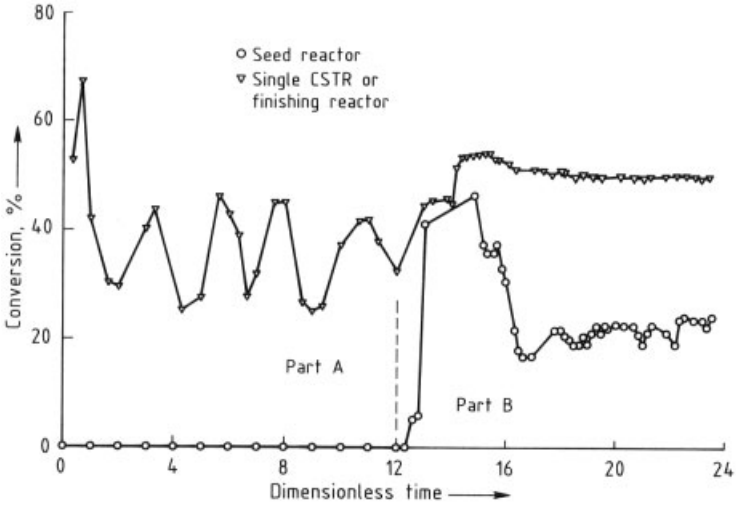

Figure 84. Conversion responses to CSTR reactor system shown in Figure 83 for continuous emulsion polymerization of vinyl acetate
Part A: operation with a single CSTR (shows sustained oscillations)
Part B: switching over to seed reactor followed by one finishing reactor (responses for both reactors are shown, with very stable response for downstream reactor and minor disturbance for seed reactor) [611]

versus time) is for a single CSTR. The following two responses are for the small seed reactor and for the larger finishing reactor.

4. References

1. J. V. Seppala, K. H. Reichert: "Trends in Polymer Reaction Engineering," *Kem.-Kemi* **17** (1990) 346.
2. W. H. Ray: "Current Problems in Polymerization Reaction Engineering," *ACS Symp. Ser.* **226** (1983) 101.
3. *Ullmann,* 4th ed., **19,** 109.
4. H. Gerrens, *Chem. Ing. Tech.* **52** (1980) 477; *Chemical Engineering* **4** (1981) 1.
5. H. Gerrens, *Chem. Tech.* (1982) June, 380; (1982) July, 434.
6. D. H. Sebastian, J. A. Biesenberger: *Polymerization Engineering,* Wiley, New York 1983.
7. K. H. Reichert, H. U. Moritz: "Polymer Reaction Engineering," *Comprehensive Polymer Science,* vol. 3, Pergamon Press, New York 1989, p. 327.

General References

8. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, New York 1953.
9. R. W. Lenz: *Organic Chemistry of Synthetic High Polymers,* Wiley-Interscience, New York 1968.
10. F. W. Billmeyer: *Te* xbook of Polymer Science, Wiley-Interscience, New York 1971.
11. D. H. Solomon (ed.): *Step-Growth Polymerization,* Marcel Dekker, New York 1972.
12. F. A. Bovey, F. W. Winslow (eds.): *Macromolecules,* Academic Press, New York 1979.
13. G. Odian: *Principles of Polymerization,* Wiley-Interscience, New York 1981.
14. A. Rudin: *The Elements of Polymer Science and Engineering,* Academic Press, Orlando, Fla. 1982.
15. P. C. Hiemenz: *Polymer Chemistry,* Marcel Dekker, New York 1984.
16. S. K. Gupta, A. Kumar: *Reaction Engineering of Step Growth Polymerization,* Plenum Press, New York 1987.
17. R. B. Seymour, C. E. Carraher: *Polymer Chemistry,* Marcel Dekker, New York 1988.

Specific References

18. J. L. Stanford, R. F. T. Stepto, *J. Chem. Soc., Faraday Trans.* **71** (1975) 1292.
19. R. F. T. Stepto: "Intra-Molecular Reaction and Gelation in Condensation or Random Polymerization," in R. N. Haward (ed.):

Developments in Polymerization-3, Applied Science Publ., Barking 1982, p. 81.
20. W. H. Carothers, *Trans. Faraday Soc.* **32** (1936) 39.
21. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y. 1953, chap. 8.
22. W. Burchard, *Polymer* **20** (1979) 589.
23. D. Durand, C. M. Bruneau, *Makromol. Chem.* **180** (1979) 2947.
24. T. M. Orlova, S.-S. A. Pavlova, L. V. Dubrovina, *J. Polym. Sci., Polym. Chem. Ed.* **17** (1979) 2209.
25. F. Lopez-Serrano, J. M. Castro, C. W. Macosko, M. Tirrell, *Polymer* **21** (1980) 263.
26. E. Ozizmir, G. Odian, *J. Polym. Sci., Polym. Chem. Ed.* **18** (1980) 2281.
27. W. H. Abraham, *Ind. Eng. Chem. Fundam.* **2** (1963) 221.
28. H. Kilkson, *Ind. Eng. Chem. Fundam.* **3** (1964) 281.
29. J. A. Biesenberger, *AIChE J.* **11** (1965) 369.
30. H. Kilkson, *Ind. Eng. Chem. Fundam.* **7** (1968) 354.
31. T. T. Szabo, J. F. Leathrum, *J. Appl. Polym. Sci.* **13** (1969) 477.
32. W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 8** (1972) 1.
33. W. H. Ray: "Polymerization Reaction Engineering," in L. Lapidus, N. R. Amundson (eds.): *Chemical Reactor Theory,* Prentice-Hall, Englewood Cliffs, N. J. 1977, p. 532.
34. N. A. Dotson, *Macromolecules* **22** (1989) 3690.
35. P. W. Morgan: *Condensation Polymers by Interfacial and Solution Methods,* Wiley-Interscience, New York 1965.
36. F. Millich, C. E. Carraher, Jr., J. Preston (eds.): *Interfacial Synthesis,* vol. III, Marcel Dekker, New York 1982.
37. P. J. Flory, *J. Am. Chem. Soc.* **63** (1941) 3083.
38. P. J. Flory, *J. Am. Chem. Soc.* **63** (1941) 3091.
39. P. J. Flory, *J. Am. Chem. Soc.* **63** (1941) 3096.
40. C. Cozewith, W. W. Graessley, G. Ver Strate, *Chem. Eng. Sci.* **34** (1979) 245.
41. S. K: Gupta, A. Kumar: *Reaction Engineering of Step Growth Polymerization,* Plenum Press, New York 1987, chap. 4.
42. W. H. Stockmayer, *J. Chem. Phys.* **11** (1943) 45.
43. W. H. Stockmayer, *J. Chem. Phys.* **12** (1944) 125.
44. W. H. Stockmayer, *J. Polym. Sci.* **9** (1952) 69; *J. Polym. Sci.* **11** (1953) 11.
45. M. Gordon, *Proc. R. Soc. London A* **268** (1962) 240.
46. M. Gordon, S. B. Ross-Murphy, *Pure Appl. Chem.* **43** (1975) 1.
47. M. Gordon, W. B. Temple: "The Graph-Like State of Matter and Polymer Science," in A. T. Balaban (ed.): *Chemical Application of Graph Theory,* Academic Press, New York 1976, p. 299.
48. S. I. Kuchanov, S. V. Koralev, S. V. Panyukov, *Adv. Chem. Phys.* **72** (1988) 115.
49. D. S. Pearson, W. W. Graessley, *Macromolecules* **11** (1978) 528. O. Durand, C. M. Bruneau, *Makromol. Chem.* **83** (1982) 1007, 1021.
50. D. R. Miller, C. W. Macosko, *J. Polym. Sci. Polym. Phys. Ed.* **26** (1988) 1.
51. C. W. Macosko, D. R. Miller, *Macromolecules* **9** (1976) 199.
52. D. R. Miller, C. W. Macosko, *Macromolecules* **9** (1976) 206.
53. D. R. Miller, C. W. Macosko, *Macromolecules* **11** (1978) 656.
54. D. Durand, C. M. Bruneau, *Polymer* **24** (1983) 587.
55. M. Gordon, G. R. Scantlebury, *Proc. R. Soc., London, A* **292** (1966) 380. M. Gordon, G. N. Malcom, *Proc. Roy. Soc, London, Ser. A* **292** (1966) 29. D. Durand, C. M. Bruneau, *Polymer* **24** (1983) 592.
56. D. R. Miller, C. W. Macosko, *Macromolecules* **13** (1980) 1063.
57. R. W. Kilb, *J. Phys. Chem.* **62** (1958) 969.
58. M. Gordon, G. R. Scantlebury, *J. Polym. Sci., Part C* **16** (1968) 3933.
59. W. B. Temple, *Makromol. Chem.* **160** (1972) 277.
60. H. Galina, A. Szustalewicz, *Macromolecules* **22** (1989) 3124.
61. H. Tobita, Ph. D. Thesis, McMaster University, Hamilton, Canada 1990.
62. A. N. Dotson, R. Galvan, C. W. Macosko, *Macromolecules* **21** (1988) 2560.
63. O. Saito, *J. Phys. Soc. Jpn.* **13** (1958) 198.
64. O. Saito: "Statistical Theory of Crosslinking", in M. Dole (ed.): *The Radiation of Macromolecules,* Academic Press, New York 1972, p. 223.
65. S. I. Kuchanov, L. M. Pis'men, *Polym. Sci. USSR (Engl. Transl.)* **14** (1972) 985. M. Gordon, G. N. Malcom, *Proc. R. Soc. London A,* **292** (1966) 29. D. Durand, C. M. Bruneau, *Polymer* **24** (1983) 592.
66. K. Dusek, *Makromol. Chem. Suppl.* **2** (1979) 35.

67. E. Donoghue, J. H. Gibbs, *J. Chem. Phys.* **70** (1979) 2346.
68. S. I. Kuchanov, Ye. S. Povolotskaya, *Polym. Sci. USSR (Engl. Transl.)* **24** (1982) 2499.
69. S. I. Kuchanov, Ye. S. Povolotskaya, *Polym. Sci. USSR (Engl. Transl.)* **24** (1982) 2512.
70. J. Mikes, K. Dusek, *Macromolecules* **15** (1982) 93.
71. S. R. Broadbent, J. M. Hammersley, *Proc. Cambridge Philos. Soc.* **53** (1957) 629.
72. J. M. Hammersley, *Proc. Cambridge Philos. Soc.* **53** (1957) 642.
73. H. L. Frisch, J. M. Hammersley, *J. Soc. Ind. Appl. Math.* **11** (1963) 894.
74. J. W. Essam: "Percolation and Cluster Size," in C. Domb, M. Green (eds.): *Phase Transitions and Critical Phenomena,* Academic Press, New York 1972, p. 197.
75. S. Kirkpatrick, *Rev. Mod. Phys.* **45** (1973) 574.
76. C. Domb, E. Stoll, T. Schneider, *Contemp. Phys.* **21** (1980) 577.
77. D. Stauffer: *Introduction to Percolation Theory,* Taylor and Francis, London 1985.
78. D. Stauffer, A. Conoglio, M. Adam, *Adv. Polym. Sci.* **44** (1982) 103.
79. D. Durand: "Network Formation – From Basic Theories Towards More Realistic Models," in R. A. Pethrick (ed.): *Polymer Yearbook 3,* Harwood Academic Publishers, New York 1986, p. 229.
80. P. G. de Gennes: *Scaling Concepts in Polymer Physics,* Cornell University Press, Ithaca, N.Y. 1979, p. 139.
81. O. Vogl: "Kinetics of Aldehyde Polymerization," C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* vol. 15, Elsevier, Amsterdam 1976, chap. 5.
82. O. Vogl: "Aldehyde Polymers," in *Encyclopedia of Polymer Science and Engineering,* vol. 1, Wiley-Interscience, New York 1985, p. 623.
83. R. N. Noyes: "Cage Effect," in *Encyclopedia of Polymer Science and Technology,* vol. 2, Wiley-Interscience, New York 1965, p. 796.
84. G. Odian: *Principles of Polymerization,* Wiley-Interscience, New York 1981, p. 216.
85. H. Kiefer, T. G. Traylor, *J. Am. Chem. Soc.* **89** (1967) 6667.
86. E. Niki, Y. Kamiya, *J. Am. Chem. Soc.* **96** (1974) 2129.
87. G. T. Russell, D. H. Napper, R. G. Gilbert, *Macromolecules* **21** (1988) 2141.
88. S. Zhu, A. E. Hamielec, *Macromolecules* **22** (1989) 3098.
89. S. S. Ivanchev, *Polym. Sci. USSR, (Engl. Transl.)* **20** (1978) 2157.
90. V. R. Kamath, *Mod. Plast.* **58** (1981) Sept., 106.
91. K. Y. Choi, W. R. Liang, G. D. Lei, *J. Appl. Polym. Sci.* **35** (1988) 1547.
92. M. A. Villalobos, M. Eng. Thesis, McMaster University, Hamilton, Canada 1989.
93. F. R. Mayo, *J. Am. Chem. Soc.* **90** (1968) 1289.
94. W. D. Graham, J. G. Green, W. A. Pryor, *J. Org. Chem.* **44** (1979) 907.
95. F. R. Mayo, *J. Am. Chem. Soc.* **75** (1953) 6133.
96. A. W. Hui, A. E. Hamielec, *J. Appl. Polym. Sci.* **16** (1972) 749.
97. A. Husain, A. E. Hamielec, *J. Appl. Polym. Sci.* **22** (1978) 1207.
98. N. J. Barr, W. I. Bengough, G. Beveridge, G. B. Park, *Eur. Polym. J.* **14** (1978) 245.
99. H. F. Kaufmann, *Makromol. Chem.* **180** (1979) 2649, 2665, 2681.
100. Y. K. Chong, E. Rizzardo, D. H. Solomon, *J. Am. Chem. Soc.* **105** (1983) 7761.
101. C. W. Wilson III, E. R. Santee, Jr., *J. Polym. Sci., Part C* **8** (1965) 97.
102. G. Talamini, G. Vidotto, *Makromol. Chem.* **100** (1967) 48.
103. T. G. Fox, H. W. Schnecko, *Polymer* **3** (1962) 575.
104. T. Otsu, B. Yamada, M. Imoto, *J. Macoromol. Sci. Chem.* **1** (1966) 61.
105. F. S. Dainton, K. J. Ivin, *Nature* **162** (1948) 705.
106. H. Sawada: *Thermodynamics of Polymerization,* Marcel Dekker, New York 1976.
107. H. W. McCormick, *J. Polym. Sci.* **25** (1957) 488.
108. T. P. Davis, K. F. O'Driscoll, M. C. Piton, M. A. Winnik, *Macromolecules* **22** (1989) 2785.
109. M. J. Ballard et al., *Macromolecules* **19** (1986) 1303.
110. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y. 1953, p. 127.
111. K. Horie, I. Mita, H. Kambe, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 2663.
112. M. Stickler, *Makromol. Chem.* **184** (1983) 2563.
113. K. Ito, *Polym. J. (Tokyo)* **16** (1984) 761.
114. G. C. Eastmond: "The Kinetics of Free-Radical Polymerization of Vinyl Monomers in Homogeneous Solution," in C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* vol. 14 A, Elsevier, Amsterdam 1976, chap. 1.

115. J. C. Bevington, H. W. Mellville, R. P. Taylor, *J. Polym. Sci.* **12** (1954) 449.
116. G. V. Schulz, G. Henrici-Olive, S. Olive, *Makromol. Chem.* **31** (1959) 88.
117. C. H. Bamford, G. C. Eastmond, D. Whittle, *Polymer* **10** (1969) 771.
118. M. Stickler, D. Panke, A. E. Hamielec, *J. Polym. Sci., Polym. Chem. Ed.* **22** (1984) 2243.
119. G. V. Schulz, G. Haborth, *Makromol. Chem.* **1** (1948) 106.
120. A. M. North: "The Influence of Chain Structure on the Free-Radical Termination Reaction," in A. D. Jenkins, A. Ledwith (eds.): *Reactivity, Mechanism and Structure in Polymer Chemistry,* Wiley-Interscience, New York 1974, chap. 5.
121. K. F. O'Driscoll: "Kinetics of Bimolecular Termination," in *Comprehensive Polymer Science,* vol. 3 Pergamon Press, London 1989, p. 161.
122. I. Mita, K. Horie, *J. Macromol. Sci., Rev. Macromol. Chem. Phys.* **C 27** (1987) 91.
123. A. W. Hui, A. E. Hamielec, *J. Polym. Sci., Part C* **25** (1968) 167.
124. J. H. Duerksen, A. E. Hamielec, *J. Polym. Sci.* **25** (1968) 155.
125. G. Z. A. Wu, L. A. Denton, R. L. Laurence, *Polym. Eng. Sci.* **22** (1982) 1.
126. C. J. Kim, A. E. Hamielec, *Polymer* **25** (1984) 845.
127. G. C. Eastmond: "Chain Transfer, Inhibition and Retardation," in C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* vol. 14 A, Elsevier, Amsterdam 1976, chap. 2.
128. C. H. Bamford, W. G. Barb, A. D. Jenkins, P. F. Onyon: *The Kinetics of Vinyl Polymerization by Free Radical Mechanism,* Butterworth Publications, London 1958.
129. C. H. Bamford: "Radical Polymerization," in *Encyclopedia of Polymer Science and Technology,* vol. 13, Wiley-Interscience, New York 1988, p. 708.
130. J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* Wiley-Interscience, New York 1975.
131. Z. Tadmor, J. A. Biesenberger, *Ind. Eng. Chem. Fundam.* **5** (1966) 336. A. E. Hamielec, J. W. Hodgins, K. Tebbens, *AIChE J.* **13** (1967) 1087.
132. R. Cintron-Cordero, R. A. Mostello, J. A. Biesenberger, *Canad. J. Chem. Eng.* **46** (1968) 434.
133. L. H. Peebles, Jr.: "Rates of Polymerization, Average Molecular Weights, and Molecular Weight Distribution of Polymers," in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* vol. II, Wiley-Interscience, New York 1975, p. 405.
134. H. Tompa: "The Calculation of Mole-Weight Distributions from Kinetic Schemes," in C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* Elsevier, Amsterdam 1976, chap. 7.
135. M. Tirrell, R. Galvan, R. L. Laurence: "Polymerization Reactors," in J. J. Carberry, A. Varma (eds.): *Chemical Reaction and Reactor Engineering,* Marcel Dekker, New York 1987, chap. 11.
136. D. J. P. Harrison, W. R. Yates, J. F. Johnson, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 25** (1985) 481.
137. M. Andreis, J. L. Koenig, *Adv. Polym. Sci.* **89** (1989) 69.
138. S. Shiga, *Polym. Plast. Technol. Eng.* **282** (1989) 17.
139. C. H. Bamford, H. Tompa, *J. Polym. Sci.* **10** (1953) 345.
140. C. H. Bamford, H. Tompa, *Trans. Faraday Soc.* **50** (1954) 1097.
141. C. H. Bamford, W. G. Barb, A. D. Jenkins, P. F. Onyon: *The Kinetics of Vinyl Polymerization by Free Radical Mechanism,* Butterworth Publications, London 1958, chap. 7.
142. W. W. Graessley, H. Mittelhauser, R. Maramba, *Makromol. Chem.* **86** (1965) 129.
143. W. W. Graessley, H. Mittelhauser, *J. Polym. Sci., Polym. Phys. Ed.* **5** (1967) 431.
144. O. Saito, K. Nagasubramanian, W. W. Graessely, *J. Polym. Sci., Polym. Phys. Ed.* **7** (1969) 1937.
145. H. Tobita, A. E. Hamielec, *Makromol. Chem., Macromol. Symp.* **20/21** (1988) 501.
146. H. Tobita, A. E. Hamielec, *Macromolecules* **22** (1989) 3098.
147. H. Tobita, A. E. Hamielec: "Cross-linking Kinetics in Free-Radical Copolymerization," in K.-H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Verlagsgesellschaft, Weinheim 1989, p. 43.
148. P. J. Flory, *J. Am. Chem. Soc.* **69** (1947) 2893.
149. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y. 1953, p. 384.
150. K. Nagasubramanian, W. W. Graessley, *Chem. Eng. Sci.* **25** (1970) 1549.
151. K. Nagasubramanian, W. W. Graessley, *Chem. Eng. Sci.* **25** (1970) 1559.
152. F. A. Bovey, F. C. Schilling, F. L. McCrackin, H. J. Wagner, *Macromolecules* **9** (1976) 76.

153. R. Dowbenko et al., *Prog. Org. Coat.* **11** (1983) 71.
154. G. A. Senich, R. E. Florin, *J. Macromol. Sci., Rev. Macromol. Chem. Phys.* **C 24** (1984) 239.
155. J. G. Kloosterboer, *Adv. Polym. Sci.* **84** (1988) 1.
156. T. Masuda, T. Higashimura, *Macromolecules* **7** (1974) 728.
157. G. C. East, H. Furukawa, *Polymer* **20** (1979) 659.
158. G. A. Olah, *J. Am. Chem. Soc.* **94** (1972) 808.
159. G. Odian: *Principles of Polymerization,* Wiley-Interscience, New York 1981, p. 342 – 344.
160. H. Sawada: *Thermodynamics of Polymerization,* Marcel Dekker, New York 1976, chap. 4.
161. J. P. Kennedy, R. G. Squires, *Polymer* **6** (1965) 579.
162. D. H. Richards: "Anionic Polymerization," in R. N. Haward (ed.): *Development in Polymerisation-1,* Applied Science Publishers, Essex 1979, chap. 1.
163. S. Bywater: "Anionic Polymerization," in *Encyclopedia of Polymer Science and Engineering,* vol. 2, Wiley-Interscience, New York 1985, p. 1.
164. M. Fontanille: "Carbanoic Polymerization: General Aspects and Initiation," in *Comprehensive Polymer Science,* vol. 3, Pergamon Press, London 1989, p. 365.
165. D. N. Bhattacharyya, C. L. Lee, J. Smid, M. Szwarc, *J. Phys. Chem.* **69** (1965) 612.
166. D. N. Bhattacharyya, J. Smid, M. Szwarc, *J. Phys. Chem.* **69** (1965) 624.
167. M. Szwarc, M. Levy, R. Milkovich, *J. Am. Chem. Soc.* **78** (1956) 2656.
168. M. Szwarc: "Living Polymers," in *Encyclopedia of Polymer Science and Technology,* vol. 8, Wiley-Interscience, New York 1968, p. 303.
169. J. Boor, Jr., *Macromol. Rev.* **2** (1967) 115.
170. A. Ledwith, D. C. Sherrington: "Reactivity and Mechanism in Polymerization by Complex Organometallic Derivatives," in A. D. Jenkins, A. Ledwith (eds.): *Reactivity, Mechanism and Structure in Polymer Chemistry,* Wiley-Interscience, New York 1974, p. 383.
171. W. Cooper: "Kinetics of Polymerization Initiated by Ziegler – Natta Catalysts," in C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* vol. 15, Elsevier, Amsterdam 1976, chap. 3.
172. J. Boor, Jr.: *Ziegler – Natta Catalysts and Polymerizations,* Academic Press, New York 1979.
173. G. Odian: *Principles of Polymerization,* Wiley-Interscience, New York 1981, chap. 8.
174. G. Natta, I. Pasquon, *Adv. Catal.* **11** (1959) 1.
175. J. A. Biesenberger, D. H. Sebastian: *Principles of Polymerization Engineering,* Wiley-Interscience, New York 1983, p. 79, 130.
176. A. B. de Carvalho, P. E. Gloor, A. E. Hamielec, *Polymer* **30** (1989) 280.
177. F. R. Mayo, F. M. Lewis, *J. Am. Chem. Soc.* **66** (1944) 1594.
178. T. Alfrey, Jr., G. Goldfinger, *J. Chem. Phys.* **12** (1944) 205.
179. F. T. Wall, *J. Am. Chem. Soc.* **66** (1944) 2050.
180. I. Sakurada: *Kyojugo Hanno,* Society of Polymer Chemistry, Tokyo 1944 p. 35.
181. E. Merz, T. Alfrey, Jr., G. Goldfinger, *J. Polym. Sci.* **I.** (1946) 75.
182. G. E. Ham: "Theory of Copolymerization," in G. E. Ham (ed.): *Copolymerization,* Wiley-Interscience, New York 1964, chap. 1.
183. G. E. Ham, *J. Polym. Sci.* **45** (1960) 169.
184. F. P. Price, *J. Chem. Phys.* **36** (1962) 209.
185. K. Ito, Y. Yamashita, *J. Polym. Sci., Part A* **3** (1965) 2165.
186. B. D. Coleman, T. G. Fox, *J. Polym. Sci., Part A* **1** (1963) 3183.
187. C. W. Pyun, *J. Polym. Sci.* **8** (1970) 1111.
188. T. Alfrey, Jr., G. Goldfinger, *J. Chem. Phys.* **12** (1944) 322.
189. C. Walling, E. R. Briggs, *J. Am. Chem. Soc.* **67** (1945) 1774.
190. T. Alfrey, Jr., G. Goldfinger, *J. Chem. Phys.* **14** (1946) 115.
191. T. Fueno, J. Furukawa, *J. Polym. Sci., Part A* **2** (1964) 3681.
192. M. N. Galbraith, G. Moad, D. H. Solomon, T. H. Spurling, *Macromolecules* **20** (1987) 675.
193. K. F. O'Driscoll, T. P. Davis, *Polym. Commun.* **30** (1989) 317.
194. R. Simha, H. Branson, *J. Chem. Phys.* **12** (1944) 253.
195. W. H. Stockmayer, *J. Chem. Phys.* **13** (1945) 199.
196. J. Stejskal, P. Kratochvil, D. Strakava, *Macromolecules* **14** (1981) 150.
197. C. Tosi, G. Catinella, *Makromol. Chem.* **137** (1970) 211.
198. J. Stejskal, P. Kratochvil, *Macromolecules* **20** (1987) 2624.
199. J. C. J. F. Tacx, H. N. Lissen, A. L. German, *J. Polym. Sci., Polym. Chem. Ed.* **26** (1988) 61.

200. J. C. J. F. Tacx, A. L. German, *Polymer* **30** (1989) 918.
201. F. P. Price: "Copolymer Composition and Tacticity," in G. G. Lowry (ed.): *Markov Chains and Monte Carlo Calculations in Polymer Science,* Marcel Dekker, New York 1970, p. 187.
202. I. Skeist, *J. Am. Chem. Soc.* **68** (1946) 1781.
203. I. H. Spinner, B. C.-Y. Lu, W. F. Graydon, *J. Am. Chem. Soc.* **77** (1955) 2198.
204. V. E. Meyer, G. G. Lowry, *J. Polym. Sci., Part A* **3** (1965) 2843.
205. V. E. Meyer, G. G. Lowry, *J. Polym. Sci., Polym. Chem. Ed.* **5** (1967) 1289.
206. W. Ring, *Makromol. Chem.* **101** (1967) 145.
207. R. L. Kruce, *J. Polym. Sci.,* Part B, 5 (1967) 437.
208. R. K. S. Chan, V. E. Meyer, *J. Polym. Sci., Part C* **25** (1968) 11.
209. A. E. Hamielec, J. F. MacGregor: "Modelling Copolymerization–Control of Composition, Chain Microstructure, Molecular Weight Distribution, Long Chain Branching and Cross-linking," in K.-H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Publishers, New York 1983, p. 21.
210. A. E. Hamielec, J. F. MacGregor, A. Penlidis, *Makromol. Chem., Macromol. Symp.* **10/11** (1987) 521.
211. S. Igarashi, *J. Polym. Sci., Part B* **1** (1963) 359.
212. C. Tosi, *Makromol. Chem.* **108** (1967) 307.
213. M. Fineman, S. D. Ross, *J. Polym. Sci.* **5** (1950) 259.
214. F. R. Mayo, C. Walling, *Chem. Rev.* **46** (1950) 191.
215. T. Kelen, F. Tudos, *J. Macromol. Sci., Chem.* **A 9** (1975) 1.
216. D. W. Behnken, *J. Polym. Sci., Part A* **2** (1964) 645.
217. P. W. Tidwell, G. A. Mortimer, *J. Polym. Sci., Part A* **3** (1965) 369.
218. P. W. Tidwell, G. A. Mortimer, *J. Macromol. Sci., Rev. Macromol. Chem.* **C 4** (1970) 281.
219. R. M. Joshi, *J. Macromol. Sci., Chem.* **A 7** (1973) 1231.
220. R. Van der Meer, H. N. Lissen, A. L. German, *J. Polym. Sci., Polym. Chem. Ed.* **16** (1978) 2915.
221. B. Yamada, M. Itahashi, T. Otsu, *J. Polym. Sci., Polym. Chem. Ed.* **16** (1978) 1719.
222. R. C. McFarlane, P. M. Reilly, K. F. O'Driscoll, *J. Polym. Sci., Polym. Chem. Ed.* **18** (1980) 251.
223. K. F. O'Driscoll, P. M. Reilly, *Makromol. Chem., Macromol. Symp.* **10/11** (1987) 355.
224. G. C. Laurier, K. F. O'Driscoll, P. M. Reilly, *J. Polym. Sci., Polym. Symp.* **72** (1985) 17.
225. M. J. Box, *Technometrics* **12** (1970) 219.
226. H. I. Britt, R. H. Luecke, *Technometrics* **15** (1973) 233.
227. T. L. Sutton, J. F. MacGregor, *Canad. J. Chem. Eng.* **55** (1977) 602.
228. H. Patino-Leal, P. M. Reilly, K. F. O'Driscoll, *J. Polym. Sci., Polym. Lett. Ed.* **18** (1980) 219.
229. P. M. Reilly, H. Patino-Leal, *Technometrics* **23** (1981) 221.
230. L. H. Garcia-Rubio, M. G. Lord, J. F. MacGregor, A. E. Hamielec, *Polymer* **26** (1985) 2001.
231. M. Berger, I. Kuntz, *J. Polym. Sci., Part A* **2** (1964) 1687.
232. D. J. T. Hill, J. H. O'Donnell, P. W. O'Sullivan, *Macromolecules* **15** (1982) 960.
233. K. F. O'Driscoll, T. P. Davis, *J. Polym. Sci., Polym. Lett. Ed.* **27** (1989) 417.
234. T. Fukuda, Y.-D. Ma, H. Inagaki, *Makromol. Chem., Phys. Suppl.* **12** (1985) 125.
235. H. Tobita, A. E. Hamietec, *Polymer* **32** (1991) 2641.
236. M. Nomura, M. Kubo, K. Fujita, *Research Reports of the Faculty of Engineering, Fukui University* **29** (1981) 167.
237. T. O. Broadhead, A. E. Hamielec and J. F. MacGregor, *Makromol. Chem., Phys. Suppl.* **10/11** (1985) 105.
238. H. Tobita, A. E. Hamielec, *ACS Symp. Ser.* **404** (1989) 242. H. Tobita, A. E. Hamielec. *Polymer* **32** (1991) 2641.
239. H. W. Melville, B. Noble, W. F. Watson, *J. Polym. Sci.* **2** (1947) 229.
240. H. W. Melville, B. Noble, W. F. Watson, *J. Polym. Sci.* **4** (1949) 629.
241. C. Walling, *J. Am. Chem. Soc.* **71** (1949) 1930.
242. P. Wittmer, *Makromol. Chem. Suppl.* **3** (1979) 129.
243. J. N. Atherton, A. M. North, *Trans. Faraday Soc.* **58** (1962) 2049.
244. A. M. North: *The Kinetics of Free Radical Polymerization,* Pergamon Press, London 1966, p. 94.
245. S. Russo, S. Munari, *J. Macromol. Sci., Chem.* **A 2** (1968) 1321.
246. T. Fukuda, Y.-D. Ma, H. Inagaki, *Polym. J. (Tokyo)* **14** (1982) 705.
247. T. Fukuda, Y.-D. Ma, H. Inagaki, *Macromolecules* **18** (1985) 17.
248. Y.-D. Ma, T. Fukuda, H. Inagaki, *Macromolecules* **18** (1985) 26.
249. T. Fukuda, K. Kubo, Y.-D. Ma, H. Inagaki, *Polym. J. (Tokyo)* **19** (1987) 523.

250. T. Fukuda, Y.-D. Ma, H. Inagaki, *Makromol. Chem., Rapid Commun.* **8** (1987) 495.
251. T. Fukuda, Y.-D. Ma, K. Kubo, A. Takada, *Polym. J. (Tokyo)* **12** (1989) 1003.
252. T. P. Davis, K. F. O'Driscoll, *Makromol Chem., Rapid Commun.* **10** (1989) 509.
253. M. Gordon, R.-J. Roe, *J. Polym. Sci.* **21** (1956) 27 – 90.
254. B. T. Storey, *J. Polym. Sci., Part A* **3** (1965) 265.
255. K. Ulbrich, M. Ilavsky, K. Dusek, J. Kopecek, *Eur. Polym. J.* **13** (1977) 579.
256. K. Ulbrich, K. Dusek, M. Ilavsky, J. Kopecek, *Eur. Polym. J.* **14** (1978) 45.
257. A. C. Shah, I. W. Parsons, R. N. Haward, *Polymer* **21** (1980) 825.
258. J. N. Nieto et al., *Eur. Polym. J.* **23** (1987) 551.
259. W.-H. Li, A. E. Hamielec, C. M. Crowe, *Polymer* **30** (1989) 1513.
260. S. Zhu, Y. Tian, A. E. Hamielec, D. R. Eaton, *Polymer* **31** (1990) 154.
261. K. Horie, A. Otagawa, M. Muaoka, I. Mita, *J. Polym. Sci., Polym. Chem. Ed.* **13** (1975) 445.
262. H. Kast, W. Funke, *Makromol. Chem.* **180** (1975) 1335.
263. J. Spevacek, K. Dusek., *J. Polym. Sci., Polym. Phys. Ed.* **18** (1980) 2027.
264. H. Galina et al., *Eur. Polym. J.* **16** (1980) 1043.
265. K. Dusek: "Network Formation by Chain Cross-linking (Co)polymerization," in R. N. Haward (ed.): *Developments in Polymerisation,* Applied Science Pub, London 1982, p. 143.
266. M. K. Gupta, R. Bansil, *J. Polym. Sci., Polym. Lett.* **21** (1983) 969.
267. M. Nagy, *Colloid Polym. Sci.* **263** (1985) 245.
268. J. Baselga, I. Hernandez-Fuentes, I. F. Pierola, M. A. Llorente, *Macromolecules* **20** (1987) 3060.
269. W. E. Gibbs, *J. Polym. Sci., Part A* **2** (1964) 4809.
270. G. B. Butler, R. J. Angelo, *J. Am. Chem. Soc.* **79** (1957) 3128.
271. C. L. McCormick, G. B. Butler, *J. Macromol. Sci., Rev. Macromol. Chem.* **C 8** (1972) 201.
272. J. Roovers, G. Smets, *Makromol. Chem.* **60** (1963) 89.
273. W. E. Gibbs, R. J. McHenry, *J. Polym. Sci., Part A* **2** (1964) 5277.
274. D. Braun, W. Brendlein, *Makromol. Chem.* **167** (1973) 203.
275. A. Matsumoto, T. Aso, S. Tanaka, M. Oiwa, *J. Polym. Sci., Polym. Chem. Ed.* **11** (1973) 2357.
276. K. Dusek, J. Spevacek, *Polymer* **21** (1980) 750.
277. P. J. Flory, *J. Am. Chem. Soc.* **69** (1947) 30.
278. D. Durand, C. M. Bruneau, *Eur. Polym. J.* **21** (1985) 527.
279. D. Durand, C. M. Bruneau, *Eur. Polym. J.* **21** (1985) 611.
280. H. Tobita, A. E. Hamielec: "Network Formation in Free Radical Polymerization," in P. J. Lemstra, L. A. Kleintjens (eds.): *Integration of Fundamental Polymer Science and Technology,* vol. IV, Elsevier Applied Science Publishers, London 1990, p. 33.
281. R. Okasha, G. Hild, P. Rempp, *Eur. Polym. J.* **15** (1979) 975.
282. R. S. Whitney, W. Burchard, *Makromol. Chem.* **181** (1980) 869.
283. C. D. Frick, A. Rudin, R. H. Wiley, *J. Macromol. Sci., Chem.* **A 16** (1981) 1275.
284. G. Hild, R. Okasha, *Makromol. Chem.* **186** (1985) 93.
285. G. Hild, R. Okasha, *Makromol. Chem.* **186** (1985) 389.
286. D. T. Landin, C. W. Macosko, *Macromolecules* **21** (1988) 846.
287. H. Tobita, A. E. Hamielec, *Polymer* **31** (1990) 1546.
288. S. I. Kuchanov, L. M. Pis'man, *Polym. Sci., USSR (Engl. Transl.)* **13** (1971) 2288.
289. H. Tobita, A. E. Hamielec, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 193.
290. H. J. Herrmann, D. Stauffer, D. P. Landau, *J. Phys., A: Math. Gen.* **16** (1983) 1221.
291. R. Bansil, H. J. Herrmann, D. Stauffer, *Macromolecules* **17** (1984) 998.
292. R. Bansil, H. J. Herrmann, D. Stauffer, *J. Polym. Sci., Polym. Symp.* **73** (1985) 175.
293. H. E. Stanley, F. Family, H. Gould, *J. Polym. Sci., Polym. Symp.* **73** (1985) 19.
294. H. M. J. Boots: "Simulation Model for Densely Cross-Linked Networks Formed by Chain-Reactions," in L. A. Kleintjens, P. J. Lemstra (eds.): *Integration of Fundamental Polymer Science and Technology,* Elsevier Applied Science, London 1986, p. 204.
295. G. P. Simon et al., *Macromolecules* **22** (1989) 3555.
296. H. Gerrens, *Chem. Tech.* **vol. 12,** (1982) June, 380.
297. H. Gerrens, *Chem. Tech.* (1982) July, 434.
298. O. Levenspiel: *Chemical Reaction Engineering,* John Wiley & Sons, New York 1972.
299. G. F. Froment, K. B. Bischoff: *Chemical Reactor Analysis and Design,* John Wiley & Sons, New York 1979.

300. J. M. Smith: *Chemical Engineering Kinetics,* McGraw-Hill, New York 1981.
301. J. A. Biesenberger, D. H. Sebastian: *Principles of Polymerization Engineering,* Wiley-Interscience, New York 1983.
302. S. K. Gupta, A. Kumar: *Reaction Engineering of Step Growth Polymerization,* Plenum Press, New York 1987.
303. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y. 1953, chap. 8.
304. H. Kilkson, *Ind. Eng. Chem., Fundam.* **2** (1964) 281.
305. Z. Tadmor, J. A. Biesenberger, *Ind. Eng. Chem., Fundam.* **5** (1966) 336.
306. Kao Corporation, JP-Kokai 27 0704, 1990.
307. H. Tobita, Y. Ohtani, *Polymer* **33** (1992) 801.
308. J. G. Watterson, J. W. Stafford, *J. Macromol. Sci., Chem.* **A 5** (1971) 679.
309. J. W. Stafford, *J. Polym. Sci., Polym. Chem. Ed.* **19** (1981) 3219.
310. H.-G. Elias, *J. Macromol. Sci., Chem.* **A 12** (1978) 183.
311. A. Kumar, R. K. Agarwal, S. Gupta, *J. Appl. Polym. Sci.* **27** (1982) 1759.
312. O. Saito, *J. Phys. Soc. Jpn.* **13** (1958) 198.
313. O. Saito in M. Dole (ed.): *The Radiation Chemistry of Macromolecules* **vol. 1,** Academic Press, New York 1972, p. 223.
314. L. J. Lee, *Rubber Chem. Technol.* **53** (1980) 542.
315. L. T. Manzione: "Reaction Injection Molding," in *Encyclopedia of Polymer Science and Engineering,* vol. 14, Wiley-Interscience, New York 1988, p. 72.
316. J. Zimmerman, *J. Polym. Sci.* **B 2** (1964) 955.
317. F. C. Chen, R. G. Griskey, G. H. Beyer, *AIChE J.* **15** (1969) 680.
318. T. M. Chang, *Polym. Eng. Sci.* **10** (1970) 364.
319. L. H. Buxbaum, *J. Appl. Polym. Sci., Appl. Polym. Symp.* **35** (1979) 59.
320. M. Thakur: "Solid-State Polymerization," *Encyclopedia of Polymer Science and Technology,* vol. 15, Wiley-Interscience, New York 1989, p.362.
321. H. K. Reimschessel, *Macromol. Rev.* **12** (1977) 65.
322. K. Tai, T. Tagawa, *Ind. Eng. Chem., Prod. Res. Dev.* **22** (1983) 192.
323. J. Sebenda, *Prog. Polym. Sci.* **6** (1978) 123.
324. S. K. Gupta, A. Kumar, *J. Macromol. Sci., Rev. Macromol. Chem. Phys.* **C 26** (1986) 183.
325. R. J. Welgos: "Polyamide, Plastics," *Encyclopedia of Polymer Science and Engineering,* vol. 11, Wiley-Interscience, New York 1988, p. 445.
326. G. Odian: *Principles of Polymerization,* Wiley-Interscience, New York 1981, p. 536.
327. P. H. Hermans, D. Heikens, P. F. van Velden, *J. Polym. Sci.* **30** (1958) 81.
328. J. M. Andrews, F. R. Jones, J. A. Semlyen, *Polymer* **15** (1974) 420.
329. J. A. Semlyen, *Adv. Polym. Sci.* **21** (1976) 41.
330. S. K. Gupta, A. Kumar; *Reaction Engineering of Step-Growth Polymerization,* Plenum Press, New York 1987, chap. 7.
331. S. Mochizuki, N. Ito, *Chem. Eng. Sci.* **28** (1973) 1139.
332. S. Mochizuki, N. Ito, *Chem. Eng. Sci.* **33** (1978) 1401.
333. S. K. Gupta, C. D. Naik, P. Tandon, A. Kumar, *J. Appl. Polym. Sci.* **26** (1981) 2153.
334. *Ullmann,* 4th ed. **19** 118.
335. K. Tai, Y. Arai, T. Tagawa, *J. Appl. Polym. Sci.* **27** (1982) 731.
336. D. B. Jacobs, J. Zimmerman in C. E. Schildknecht, I. Skeist (eds.): *Polymerization Processes,* Wiley-Interscience, New York 1977, p. 424.
337. N. Ogata, *Makromol. Chem.* **42** (1960) 52.
338. N. Ogata, *Makromol. Chem.* **43** (1961) 117.
339. Y. Murakami: *Fundamental Principles and Analysis of Polymerization Reactors,* Baifukan, Tokyo 1976, p. 127, in Japanese.
340. A. Kumar, S. Kuruville, A. R. Raman, S. K. Gupta, *Polymer* **22** (1981) 387.
341. A. Kumar, R. K. Agarwal, S. K. Gupta, *J. Appl. Polym. Sci.* **27** (1982) 1759.
342. M. Katz in C. E. Schildknecht, I. Skeist (eds.): *Polymerization Processes,* Wiley-Interscience, New York 1977, p. 468.
343. S. K. Gupta, A. Kumar: *Reaction Engineering of Step-Growth Polymerization,* Plenum Press, New York 1987, chap. 8.
344. M. J. Barandiaran, J. M. Asua, *Polymer* **31** (1990) 1347.
345. M. J. Barandiaran, J. M. Asua, *Polymer* **31** (1990) 1352.
346. J. W. Ault, D. A. Mellichamp, *Chem. Eng. Sci.* **27** (1972) 2219.
347. K. Ravindranath, R. A. Mashelker, *J. Appl. Polym. Sci.* **26** (1981) 3179.
348. A. Kumar, V. K. Sukthankar, C. P. Vaz, S. K. Gupta, *Polym. Eng. Sci.* **24** (1984) 185.
349. K. Ravindranath, R. A. Mashelker, *J. Appl. Polym. Sci.* **27** (1982) 471.
350. K. Ravindranath, R. A. Mashelker, *Chem. Eng. Sci.* **41** (1986) 2969.

351. I. Goodman: "Polyesters," *Encyclopedia of Polymer Science and Technology,* vol. 12, Wiley-Interscience, New York 1988, p. 1.
352. K. Ravindranath, R. A. Mashelkar, *Polym. Eng. Sci.* **22** (1982) 610.
353. T. Yamada, Y. Imamura, O. Makimura, *Polym. Eng. Sci.* **25** (1985) 788.
354. T. Yamada, Y. Imamura, O. Makimura, *Polym. Eng. Sci.* **26** (1986) 708.
355. A. E. Hamielec, J. F. MacGregor, A. Penlidis: "Multicomponent Free-Radical Polymerization in Batch, Semi-Batch and Continuous Reactors," *Makromol. Chem., Macromol. Symp.* **10/11** (1987) 521.
356. H. Tobita, A. E. Hamielec: "The Kinetics of Free-Radical Copolymerization–The Pseudo-Kinetic Rate Constant Method," *Polymer* **32** (1991) 2641.
357. A. E. Hamielec, J. F. MacGregor: "Modelling Copolymerizations – Control of Chain Microstructure, Long Chain Branching, Cross-linking and Molecular Weight Distributions," in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Publishers, New York 1983, p. 21.
358. S. Zhu, A. E. Hamielec, *Macromolecules* **22** (1989) 3093.
359. H. M. J. Boots, *J. Polym. Sci., Polym. Phys. Ed.* **20** (1982) 1695.
360. S. Zhu, Y. Tian, A. E. Hamielec, D. R. Eaton, *Polymer* **31** (1990) 154.
361. D. J. Coyle, T. J. Tulig, M. Tirrell, *Ind. Eng. Chem. Fundam.* **24** (1985) 343.
362. W. H. Stockmayer, *J. Chem. Phys.* **13** (1945) 199.
363. H. Tobita, A. E. Hamielec, *Makromol. Chem., Macromol. Symp.* **20/21** (1988) 501.
364. H. Tobita, A. E. Hamielec, *Macromolecules* **22** (1989) 3105.
365. H. Tobita, A. E. Hamielec, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 193.
366. W. A. Pryor: *Free Radicals,* McGraw-Hill, New York 1966.
367. K. Kirchner, K. Riederle, *Angew. Makromol. Chem.* **111** (1983) 1.
368. Y. K. Chong, E. Rizzardo, D. H. Solomon, *J. Am. Chem. Soc.* **105** (1983) 7761.
369. M. Tirrell, T. J. Tulig, in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Publishers, New York 1983, p. 247.
370. C. T. Liao, D. C. Sundberg, *Polymer* **27** (1986) 265.
371. A. W. Hui, A. E. Hamielec, *J. Applied Polym. Sci.* **16** (1972) 749.
372. A. Husain, A. E. Hamielec, *J. Applied Polym. Sci.* **22** (1978) 1207.
373. A. E. Hamielec, J. F. MacGregor, S. Webb, T. Spychaj in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig & Wepf Verlag, New York 1986, p. 185.
374. G. Z. Wu, L. A. Denton, R. L. Laurence, *Polym. Eng. Sci.* **22** (1982) 1.
375. T. Rintelen, K. Riederle, K. Kirchner in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Publishers, New York 1983, p. 269. N. Khac Tien, E. Flaschel, A. Renken in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Publishers, New York 1983, p. 175.
376. H. K. Fauske, J. C. Leung, *Chem. Eng. Progr.* **81** (1985) 39.
377. J. C. Leung, H. K. Fauske, *Thermochim. Acta* **104** (1986) 13.
378. O. Chiantore, A. E. Hamielec, *Polymer* **26** (1985) 608.
379. J. Braks, Ph. D. Thesis, University of Waterloo, Canada 1973.
380. S. T. Balke, A. E. Hamielec, *J. Applied Polym. Sci.* **17** (1973) 905.
381. J. Cardenas, K. F. O.'Driscoll, *J. Polym. Sci., Chem. Ed.* **14** (1976) 883.
382. J. Cardenas, K. F. O'Driscoll, *J. Polym. Sci., Chem. Ed.* **15** (1977) 1883, 2097.
383. F. L. Marten, A. E. Hamielec, *ACS Symp. Ser.* **104** (1979) 43.
384. M. Stickler, D. Panke, A. E. Hamielec, *J. Polym. Sci., Chem. Ed.* **22** (1984) 2243.
385. W. Y. Chiu, G. M. Carratt, D. S. Soong, *Macromolecules* **16** (1983) 348.
386. S. K. Soh, D. C. Sundberg, *J. Polym. Sci., Chem. Ed.* **20** (1982) 1299, 1315, 1331, 1345.
387. K. Ito, *Polym. J. (Tokyo)* **12** (1980) 499.
388. H. M. J. Boots, *J. Polym. Sci., Polym. Phys. Ed.* **20** (1982) 1695.
389. T. J. Tulig, M. V. Tirrell, *Macromolecules* **14** (1981) 1501.
390. T. J. Tulig, M. V. Tirrell, *Macromolecules* **15** (1982) 459.
391. B. M. Louie, G. M. Carratt, D. S. Soong, *J. Appl. Polym. Sci.* **30** (1985) 3985.
392. D. Achillias, C. Kiparissides, *J. Applied Polym. Sci.* **35** (1990) 1303.
393. M. E. Adams et al., *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 1.
394. D. Panke, *Makromol. Chem., Rapid Commun.* **7** (1986) 171.
395. J. C. Bevington, *Trans. Faraday Soc.* **51** (1955) 1392.
396. F. DeSchrijver, G. Smets, *J. Polym. Sci., Part A-1* **4** (1966) 2201.

397. J. H. Duerksen, A. E. Hamielec, *J. Polym. Sci.* **C 25** (1968) 155.
398. A. W. Hui, A. E. Hamielec, *J. Polym. Sci.* **C 25** (1968) 167.
399. M. J. Ballard et al., *Macromolecules* **19** (1986) 1303.
400. M. Stickler, *Makromol. Chem.* **184** (1983) 2563.
401. J. G. Kloosterboer, *Adv. Polym. Sci.* **84** (1988) 1.
402. J. G. Kloosterboer, G. F. C. Lijten, H. M. J. Boots, *Makromol. Chem., Macromol. Symp.* **24** (1989) 223.
403. P. D. Armitage et al., *Polymer* **29** (1988) 2221.
404. S. Zhu, A. E. Hamielec: "Heat Effects for Free-Radical Polymerization in Glass Ampoule Reactors," *Polymer* **32** (1991) 3021.
405. A. E. Hamielec, J. F. MacGregor, A. Penlidis: "Copolymerization Kinetics," *Comprehensive Polymer Science,* **vol. 3,** Pergamon Press, Oxford 1989, p. 17 – 32.
406. D. A. Tirrell: "Copolymer Composition," *Comprehensive Polymer Science,* vol. 3, Pergamon Press, Oxford 1989, p. 195 – 206.
407. T. Fukuda, Y. D. Ma, H. Inagaki, *Macromolecules* **18** (1985) 17.
408. Y. D. Ma, T. Fukuda, H. Inagaki, *Macromolecules* **18** (1985) 267.
409. T. Fukuda, Y. D. Ma, H. Inagaki, *Makromol. Chem., Rapid Commun.* **8** (1987) 495.
410. H. Tanaka, K. Sasai, T. Sato, T. Ota, *Macromolecules* **21** (1988) 3534.
411. O. F. Olaj, I. Bitai, G. Gleixner, *Makromol. Chem.* **186** (1985) 2569.
412. O. F. Olaj, I. Bitai, F. Hinkelmann, *Makromol. Chem.* **188** (1987) 1689.
413. O. F. Olaj, P. Kremminger, I. Schnoll-Bitai, *Makromol. Chem., Rapid Comm.* **9** (1988) 771.
414. H. K. Mahabadi, K. F. O'Driscoll, *J. Macromol. Sci. Chem.* **A 11** (1977) 967.
415. T. P. Davis, K. F. O'Driscoll, M. C. Piton, M. A. Winnik, *Macromolecules* **23** (1990) 2119.
416. J.-F. Kuo, C.-Y. Chen, C.-W. Chen, T.-C. Pan, *Polym. Eng. Sci.* **24** (1984) 22.
417. G. Storti et al., *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 213.
418. T. O. Broadhead, A. E. Hamielec, J. F. MacGregor, *Makromol. Chem., Suppl.* **10/11** (1985) 105.
419. J. Kanetakis, F. Y. Wong, A. E. Hamielec, J. F. MacGregor, *Chem. Eng. Commun.* **35** (1985) 123.
420. L. H. Garcia, M. G. Lord, J. F. MacGregor, A. E. Hamielec, *Polymer* **27** (1986) 602.
421. K. M. Jones, D. Bhattacharya, J. L. Brash, A. E. Hamielec, *Polymer* **27** (1986) 602.
422. D. Bhattacharya, A. E. Hamielec, *Polymer* **27** (1986) 611.
423. I. M. Yaraskavitch, J. L. Brash, A. E. Hamielec, *Polymer* **28** (1987) 489.
424. M. A. Dubé, A. Penlidis, K. F. O'Driscoll, *Chem. Eng. Sci.* **45** (1990) 2785.
425. K. F. O'Driscoll, J. Huang, *Eur. Polym. J.* **26** (1990) 643.
426. A. E. Hamielec, A. C. Ouano, L. L. Nebenzahl, *J. Liq. Chromatogr.* **1** (1978) 527. A. E. Hamielec, H. Meyer: "Online Molecular Weight and Long Chain Branching Measurements using SEC and Low-Angle Laser Light Scattering," *Developments in Polymer Characterization–5,* Elsevier Applied Science Publishers, London 1986, p. 95 – 130. M. Styring, J. Armonas, A. E. Hamielec, *Polym. Mat. Sci. Eng.* **54** (1986) 88. M. Styring, J. Armonas, A. E. Hamielec, *J. Liq. Chromatogr.* **10** (1987) 783. *Ullmann,* 4th. ed. 19, 107 – 165. H. Bieringer, K. Flatan, D. Reese, *Angew. Makromol Chem.* **123/124** (1984) 307.
427. Dow, US 2 530 409, 1948, US 2 727 884, 1953, US 2 849 430, 1955. Distrene, *Br. Plast.* **30** (1957) no. 1, 26 – 27.
428. H. Ritter, *Chem.-Ing.-Tech.* **41** (1969) 419 – 426.
429. R. Friedrich, M. Eisenstein, *Chem.-Ing.-Tech.* **50** (1978) 466 – 467.
430. K.-M. Hess, *Chem.-Ing.-Tech.* **51** (1979) 245.
431. F. Widmer, *Adv. Chem. Ser.* **128** (1973) 51 – 67.
432. G. Beckmann, E. F. Engel, *Chem.-Ing.-Tech.* **38** (1966) 1025 – 1031.
433. G. Daumiller, *Chem.-Ing.-Tech.* **40** (1968) 673 – 682.
434. R. L. Scott, *J. Chem. Phys.* **17** (1949) 279 – 284.
435. P. J. Flory: *Principles of Polymer Chemistry,* Cornell University Press, Ithaca, N.Y. 1953, p. 554 ff.
436. G. E. Molau, *J. Polym. Sci. Part A* **3** (1965) 1267 – 1278, 4235 – 4242.
437. G. E. Molau, *J. Polym. Sci. Polym. Lett. Ed.* **3** (1965) 1007 – 1015.
438. G. E. Molau, H. Keskkula, *J. Polym. Sci. Polym. Chem. Ed.* **4** (1966) 1595 – 1607.
439. G. E. Molau, W. M. Wittbrodt, V. E. Meyer, *J. Appl. Polym. Sci.* **13** (1969) 2735 – 2736. G. E. Molau, *Kolloid Z. Z. Polym.* **238** (1970) 493 – 498.

440. Dow, US 2 694 692, 1954.
441. J. L. Amos, *Polym. Eng. Sci.* **14** (1974) 1 – 11.
442. B. W. Bender, *J. Appl. Polym. Sci.* **9** (1965) 2887 – 2894.
443. E. R. Wagner, L. M. Robeson, *Rubber Chem. Technol.* **43** (1970) 1129 – 1137.
444. T. O. Craig, *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2105 – 2109.
445. H. Willersinn, *Makromol. Chem.* **101** (1967) 296 – 316.
446. G. E. Molau, W. M. Wittbrodt, *Macromolecules* **1** (1968) 260 – 264.
447. G. F. Freeguard, M. Karmarkar, *J. Appl. Polym. Sci.* **15** (1971) 1649 – 1655.
448. G. F. Freeguard, M. Karmarkar, *J. Appl. Polym. Sci.* **15** (1971) 1657 – 1663.
449. G. F. Freeguard, M. Karmarkar, *J. Appl. Polym. Sci.* **16** (1972) 69 – 82.
450. G. F. Freeguard, *Polymer* **13** (1972) 366 – 370.
451. G. F. Freeguard, *Br. Polym. J.* **6** (1974) 205 – 228.
452. J. P. Fischer, *Angew. Makromol. Chem.* **33** (1973) 35 – 74.
453. H. Cherdron, *Chimia* **28** (1974) 553 – 560.
454. A. Echte, *Angew. Makromol. Chem.* **58/59** (1977) 175 – 198.
455. M. L. Huggins: *Chemistry of High Polymers,* Wiley, New York 1958, p. 38 – 51.
456. H. Gerrens, W. Fink, E. Köhnlein, *J. Polym. Sci. Polym. Symp.* **16** (1967) 2781 – 2793.
457. A. R. Berens, *Angew. Makromol. Chem.* **47** (1975) 97 – 110.
458. G. Talamini, *J. Polym. Sci. Polym. Phys. Ed.* **4** (1966) 535 – 537.
459. A. Crosato-Arnaldi, P. Gasparini, G. Talamini, *Makromol. Chem.* **117** (1968) 140 – 152. J. Swoboda, J.-C. Heilig, *Kunststoff-Symposium Moskau,* 11 – 14 Oct, 1977.
460. A. F. Hauss, *J. Polym. Sci. Polym. Symp.* **33** (1971) 1 – 12.
461. A. H. Abdel-Alim, A. E. Hamielec, *J. Appl. Polym. Sci.* **16** (1972) 1093 – 1101.
462. C. Hoheisel, *Angew. Makromol. Chem.* **34** (1973) 19 – 33.
463. O. F. Olaj, *Angew. Makromol. Chem.* **47** (1975) 1 – 14.
464. O. F. Olaj, *J. Makromol. Sci. C A* **11** (1977) 1307 – 1317.
465. O. F. Olaj, J. W. Breitenbach, K. J. Parth, N. Philippovich, *J. Macromol. Sci. Chem. A* **11** (1977) 1319 – 1331.
466. J. Ugelstad, H. Lervik, B. Gardinovacki, E. Sund, *Pure Appl. Chem.* **25** (1971) 121 – 152.
467. J. Ugelstad, H. Fløgstad, T. Hertzberg, E. Sund, *Makromol. Chem.* **164** (1973) 171 – 181.
468. J. Ugelstad, *J. Macromol. Sci. Chem. A* **11** (1977) 1281 – 1305.
469. J. Jaeger, *Chem. Ztg.* **90** (1966) 70 – 73.
470. J.-C. Thomas, *SPE J.* **23** (1967) no. 10, 61 – 65.
471. J.-C. Thomas, *Hydrocarbon Process.* **47** (1968) no. 11, 192 – 196.
472. J. Chatelain, *Br. Polym. J.* **5** (1973) 457 – 465.
473. Union Carbide, BE 723 775, 1968.
474. D. M. Rasmussen, *Chem. Eng.* **79** (1972) no. 21, 104 – 105.
475. H. L. Batleman, *Plast. Eng.* **31** (1975) no. 4, 73 – 75.
476. Union Carbide, *Hydrocarbon Process.* **56** (1977) no. 11, 212.
477. BASF, US 3 634 282, 1968; BASF, US 3 639 377, 1967; BASF, US 3 652 527, 1968.
478. BASF, *Oil Gas J.* **68** (1970) no. 47, 64.
479. K. Wisseroth, *Angew. Makromol. Chem.* **8** (1969) 41 – 60.
480. K. Wisseroth, *Kolloid Z. Z. Polym.* **241** (1970) 943 – 954.
481. K. Wisseroth, *Chem. Ztg.* **101** (1977) 271 – 284.
482. J. C. Davis, *Chem. Eng.* **85** (1978) no. 1, 25 – 27.
483. B. Dietrich, *Appl. Polym. Symp.* **26** (1975) 1 – 11.
484. H. Kreuter, B. Dietrich, *Chem. Eng.* **81** (1974) no. 16, 62 – 63.
485. Hoechst, *Hydrocarbon Process.* **56** (1977) no. 11, 208. A. Kageyama, A. Hirotani, *Hydrocarbon Process.* **51** (1972) no. 11, 97 – 98. E. Susa, *Hydrocarbon Process.* **51** (1972) no. 7, 115 – 116. Montecatini Edison, *Hydrocarbon Process.* **50** (1971) no. 11, 195. Snam Progetti, *Hydrocarbon Process.* **50** (1971) no. 11, 197.
486. Hoechst, *Hydrocarbon Process.* **56** (1977) no. 11, 213. Mitsui – Montedison, *Hydrocarbon Process.* **56** (1977) no. 11, 214.
487. F. S. Dainton, P. H. Seaman, D. G. L. James, R. S. Eaton, *J. Polym. Sci.* **34** (1959) 209 – 228.
488. Hoechst, *Hydrocarbon Process.* **54** (1975) no. 11, 175. *Ullmann,* 3rd ed., **14**, pp. 276 – 280.
489. K. Thomas, *India Rubber World* **130** (1954) 203.
490. BASF, DT 961 309, 1954. BASF DT 1 003 446, 1955.
491. H. Gerrens, H. Ohlinger, R. Fricker, *Makromol. Chem.* **87** (1965) 209 – 227.

492. H. Gerrens, D. J. Stein, *Makromol. Chem.* **87** (1965) 228 – 247.
493. K. E. J. Barrett (ed.): *Dispersion Polymerization in Organic Media,* Wiley, London 1975.
494. W. Heller, T. L. Pugh, *J. Chem. Phys.* **22** (1954) 1778. W. Heller, *Pure Appl. Chem.* **12** (1966) 249 – 274.
495. D. J. Walbridge: *The Design and Synthesis of Dispersants for Dispersion Polymerization in Organic Media,* chap. 3 in [493], p. 45 – 114.
496. E. W. Fischer, *Kolloid Z.* **160** (1958) 120 – 141.
497. R. H. Ottewill, T. Walker, *Kolloid Z. Z. Polym.* **227** (1968) 108 – 116.
498. D. H. Napper, *Ind. Eng. Chem. Prod. Res. Dev.* **9** (1970) 467 – 477.
499. K. Jäckel, *Kolloid-Z.* **197** (1964) 143 – 151.
500. R. H. Ottewill: "Effect of Nonionic Surfactants on the Stability of Dispersion," in M. J. Schick (ed.): *Nonionic Surfactants,* vol. 1, chap. 19, Dekker, New York 1967, p. 627 – 682.
501. D. H. Napper, *Trans. Faraday Soc.* **64** (1968) 1701 – 1711. D. H. Napper, *J. Colloid Interface Sci.* **29** (1969) 168 – 170. D. H. Napper, *J. Colloid Interface Sci.* **32** (1970) 106 – 114.
502. D. H. Napper, A. Netschey, *J. Colloid Interface Sci.* **37** (1971) 528 – 535.
503. D. H. Napper, R. J. Hunter in M. Kerker (ed.): *Surface Chemistry and Colloids,* Butterworth, London 1972, p. 279 – 292. D. H. Napper: *Polymeric Stabilization of Calloidal Dispersions,* Academic Press, London 1983.
504. A. Guyot: "Precipitation Polymerization," in *Comprehensive Polymer Science,* Pergamon Press, Oxford 1989, p. 261.
505. S. U. Pickering, *J. Chem. Soc.* **91** (1907) 2001 –2021. S. U. Pickering, *Kolloid Z.* **7** (1910) 11 – 16.
506. W. P. Hohenstein, H. Mark, *J. Polym. Sci.* **1** (1946) 127 – 145.
507. H. Wenning, *Makromol. Chem.* **20** (1956) 196 – 213.
508. E. H. Merz, *J. Appl. Polym. Sci.* **3** (1960) 374.
509. H. Wenning, *Kolloid Z. Z. Polym.* **182** (1962) 60 –75.
510. H. Wenning, *Kunstst. Plast.* **5** (1958) 328 – 340.
511. R. N. Haward, *J. Polym. Sci.* **4** (1949) 273 – 287.
512. S. Kaichi, *Bull. Chem. Soc. Jpn.* **29** (1956) 241 –245.
513. E. Trommsdorff, *Makromol. Chem.* **13** (1954) 76 –89.
514. H. J. Reinhardt, R. Thiele, *Plaste Kautsch.* **19** (1972) 648 – 654.
515. E. Farber, M. Koral, *Polym. Eng. Sci.* **8** (1968) 11 – 18.
516. Dow, BE 677 224, BE 677 173, 1965.
517. Br. Petroleum, US 3 726 848, 1970.
518. E. Trommsdorff, C. E. Schildknecht: "Polymerizations in Suspension," in C. E. Schildknecht (ed.): *Polymer Processes,* chap. 3, Interscience, New York 1956, p. 69 – 109. M. Munzer, E. Trommsdorff: "Polymerizations in Suspension," in C. E. Schildknecht, I. Skeist (eds.): *Polymerization Processes,* chap. 5, Wiley, New York 1977, p. 106 – 142.
519. E. Farber: "Suspension Polymerization," in H. F. Mark, N. G. Gaylord, N. M. Bikales (eds.): *Encyclopedia of Polymer Science and Technology,* vol. 13, Interscience, New York 1970, p. 552 –571.
520. J. H. Rushton, E. W. Costich, H. J. Everett, *Chem. Eng. Prog.* **46** (1950) 395 – 404, 467 – 476. F. Wolf, B. Hoffbauer, S. Eckert, *Plaste Kautsch.* **19** (1972) 26 – 28.
521. F. H. Winslow, W. Matreyek, *Ind. Eng. Chem.* **43** (1951) 1108 – 1112.
522. BASF, DT 810 812, 1949.
523. Röhm & Haas, DT 735 284, 1935; DT 747 596, 1935; US 2 171 765, 1934.
524. R. Shinnar, J. M. Church, *Ind. Eng. Chem.* **52** (1960) 253 – 256.
525. J. M. Church, R. Shinnar, *Ind. Eng. Chem.* **53** (1961) 479 – 484.
526. R. Shinnar, *J. Fluid Mech.* **10** (1961) 259 – 275.
527. A. N. Kolmogoroff, *C. R. Acad. Sci USSR* **30** (1940) 301, 538.
528. A. N. Kolmogoroff, *J. Fluid Mech.* **13** (1962) 82 –85.
529. R. H. Kraichnan, *J. Fluid Mech.* **62** (1974) 305 –330.
530. L. A. Cutter, Ph. D. Thesis, Columbia University, New York 1960. H. G. Schwartzberg, R. E. Treybal, *Ind. Eng. Chem. Fundam.* **7** (1968) 1 – 6.
531. T. Vermeulen, G. M. Williams, G. E. Langlois, *Chem. Eng. Prog.* **51** (1955) no. 2, 85 F – 94 F.
532. F. B. Sprow, *Chem. Eng. Sci.* **22** (1967) 435 – 442.
533. H. T. Chen, S. Middleman, *AIChE J.* **13** (1967) 989 – 995.
534. Y. Mlynek, W. Resnick, *AIChE J.* **18** (1972) 122 – 127.
535. W. A. Rodger, V. G. Trice, J. H. Rushton, *Chem. Eng. Prog.* **52** (1956) 515 – 520.

536. J. O. Hinze, *AIChE J.* **1** (1955) 289 – 295.
537. F. B. Sprow, *AIChE J.* **13** (1967) 995 – 998.
538. H. D. Schindler, R. E. Treybal, *AIChE J.* **14** (1968) 790 – 798.
539. C. A. Coulaloglou, L. L. Tavlarides, *AIChE J.* **22** (1976) 289 – 297.
540. J. H. Vanderveen, M. S. Thesis, Univ. of California, Berkeley UCRL 8 733, Berkeley 1960.
541. M. A. Zeitlin, L. L. Tavlarides, *AIChE J.* **18** (1972) 1268 – 1271.
542. M. A. Zeitlin, L. L. Tavlarides, *Can. J. Chem. Eng.* **50** (1972) 207 – 215.
543. M. A. Zeitlin, L. L. Tavlarides, *Proc. 5th European/2nd Intern. Symp. on Chem. Reaction Eng.,* Elsevier, Amsterdam 1972, p. B 1, 1 – 12.
544. H. Hopff, E. Lutz, *Kunstst. Plast.* **5** (1958) 341 –344.
545. H. Hopff, H. Lüssi, P. Gerspacher, *Makromol. Chem.* **78** (1964) 24 – 36, 37 – 46.
546. H. Hopff, H. Lüssi, E. Hammer, *Makromol. Chem.* **82** (1965) 175 – 183, 184 – 189.
547. H. Hopff, H. Lüssi, E. Hammer, *Makromol. Chem.* **84** (1965) 274 – 281, 282 – 285, 286 – 289.
548. H. Hedden, Dissertation Univ. Stuttgart 1969.
549. H. Hedden, *Chem.-Ing.-Tech.* **42** (1970) 457 – 462.
550. F. Langner, H. U. Moritz, K. H. Reichert, *Chem. Ing. Tech.* **51** (1979) 746.
551. F. Wolf, S. Eckert, *Plaste Kautsch.* **18** (1971) 580 –583.
552. H. Wenning, *Kunstst. Plast.* **5** (1958) 328 – 340.
553. B. A. Hills, *Br. Chem. Eng.* **6** (1961) no. 2, 104 –106.
554. G. Beckmann, *Chem. Ing. Tech.* **36** (1964) 169 – 174.
555. D. B. Scully, *J. Appl. Polym. Sci.* **20** (1976) 2299 – 2303.
556. G. Mino, *J. Polym. Sci.* **22** (1956) 369 – 383.
557. J. Brandrup, *Faserforsch. Textiltechn.* **12** (1961) 133 – 140, 208 – 213.
558. Monsanto, US 2 862 906, 2 862 907, 1957.
559. Dow, US 3 267 178, 1962.
560. BASF, DT 845 264, 1950.
561. BASF, DT 941 389, 1951.
562. BASF, DT 951 299, 1953; 963 955, 1953.
563. Koppers, GB 756 654, 1953.
564. M. Villalobos: *Foamed Polymer Course Notes,* McMaster University, Hamilton, Canada 1990.
565. M. Konno, K. Arai, S. Saito, *J. Chem. Eng. Jpn.* **15** (1982) 131.
566. M. A. Villalobos, A. E. Hamielec, P. E. Wood, *J. Appl. Polym. Sci.* **42** (1991) 629.
567. Edison, US 3 228 919, 1960.
568. Chem. Werke Hüls, DE-AS 2 528 950, 1975.
569. Firestone, US 4 000 355, 1975.
570. Hoechst, DE-AS 1 021 165, 1956.
571. M. Macoveanu, K. H. Reichert, *Angew. Makromol. Chem.* **44** (1975) 141 – 149.
572. T. Y. Xie, A. E. Hamielec, P. E. Wood, D. R. Woods, *J. Vinyl Technol.* **13** (1991) 2.
573. C. Kiparissides, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 171.
574. Bayer, DT 1 045 102, 1957; DT 1 113 570, 1957.
575. D. Hunkeler, A. E. Hamielec, W. Baade, *Polymer* **30** (1989) 127.
576. P. Trijasson, P. Pith, M. Lambla, *Makromol. Chem. Macromol. Symp.* **35/36** (1990) 141.
577. C. Graillot, C. Pichot, A. Guyot, M. S. El-Aasser, *J. Polym. Sci., Polym. Chem. Ed.* **24** (1986) 427.
578. W. Baade, K. H. Reichert, *Makromol. Chem., Rapid Commun.* **7** (1986) 235.
579. F. Candau, Y. S. Leong, R. M. Fitch, *J. Polym. Sci., Polym. Chem. Ed.* **23** (1985) 193.
580. J. W. Vanderhof et al., *J. Dispersion Sci. Technol.* **5** (1984) 323.
581. Dow, DT 1 081 228, 1957.
582. Nalco, US 3 767 629, 1971.
583. Röhm GmbH, DE-OS 2 009 218, 1970.
584. G. Beckmann, *Adv. Chem. Ser.* **128** (1973) 37 – 50.
585. G. Beckmann, *Chem. Technol.* **3** (1973) 304 – 310.
586. G. Beckmann, *Chemie-Techn.* **5** (1976) no. 4, 135 – 139.
587. L. F. Albright, C. G. Bild, *Chem. Eng.* **82** (1975) no. 19, 121 – 128.
588. B. Terwiesch, *Hydrocarbon Process.* **55** (1976) no. 11, 117 – 121.
589. H. Fikentscher, *Angew. Chem.* **51** (1938) 433.
590. H. Gerrens, *Fortschr. Hochpolym. Forsch.* **1** (1959) 234 – 328.
591. E. W. Duck: "Emulsion Polymerization," in H. F. Mark, N. G. Gaylord, N. M. Bikales (eds.): *Encyclopedia of Polymer Science and Technology,* vol. 5, Interscience, New York 1966, p. 801 – 859.
592. B. M. E. van der Hoff: "Emulsion Polymerization," in K. Shinoda (ed.): *Solvent Properties of Surfactant Solutions,* Marcel Dekker, New York 1967, p. 285 – 340.
593. J. W. Vanderhoff: "The Mechanism of Emulsion Polymerization," in G. E. Ham (ed.): *Kinetics and Mechanisms of Polymerization,* vol. 1, part II, Marcel Dekker, New York 1969, p. 1 – 138.

594. A. E. Alexander, D. H. Napper: "Emulsion Polymerization," in A. D. Jenkins (ed.): *Progress in Polymer Sci.,* vol. 3, Pergamon Press, Oxford 1971, p. 145 – 197.
595. J. L. Gardon: "Mechanism of Emulsion Polymerization," in J. K. Graven, R. W. Tess (eds.): *Applied Polymer Science,* Org. Coatings and Plastic Div., ACS, Washington, D. C. 1975, p. 138 – 160.
596. J. L. Gardon: "Emulsion Polymerization, Theory," in H. F. Mark, N. M. Bikales (eds.): *Encyclopedia of Polymer Science and Technology,* Supplement vol. 1, Interscience, New York 1976, p. 238 – 259.
597. J. Ugelstad, F. K. Hansen, *Rubber Chem. Technol.* **49** (1976) 536 – 609.
598. G. W. Poehlein: in R. Buscall, T. Corner, J. F. Stageman (eds.): *Polymer Colloids,* Elsevier Applied Sci. Publishers, Barking 1985, p. 45 – 68.
599. A. Penlidis, J. F. MacGregor, A. E. Hamielec, *AIChE J.* **31** (1985) 881.
600. J. W. Vanderhoff, *J. Polym. Sci., Polym. Symp.* **72** (1985) 161.
601. B. W. Brooks: "Developments in Emulsion Polymerization," in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Publishers, New York 1989, p. 3.
602. D. H. Napper, R. G. Gilbert: "Polymerizations in Emulsions," *Comprehensive Polymer Science,* **vol. 4.**
603. K. W. Min, W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **11** (1974) no. 2, 177 – 255.
604. *J. Polym. Sci. Polym. Symp.* **27** (1969).
605. *Br. Polym. J.* **2** (1970) 1 – 66, 116 – 177.
606. R. M. Fitch (ed.): *Polymer Colloids,* Plenum Press, New York 1971.
607. *J. Makromol. Sci. Chem.* **7** (1973) 601 – 736.
608. *ACS Symp. Ser.* **24** (1976).
609. "Emulsion Copolymerization Mechanisms and Processes; Relations between Colloid Structure and Properties," *Die Makromolekulare Chemie, Macromolecular Chemistry and Physics,* **Suppl. 10/11,** March 1985.
610. "Copolymerization and Copolymers in Dispersed Media," *Die Makromolekulare Chemie, Macromolecular Chemistry and Physics,* **Suppl. 35/36,** May 1990.
611. G. W. Poehlein, R. H. Ottewill, J. W. Goodwin (eds.): *Science and Technology of Polymer Colloids,* vols. I and II, Nato ASI Series, Series E: Applied Sciences, no. 68, Martinus, Nijhoff Publishers, 1983.
612. H. Fikentscher, H. Gerrens, H. Schuller, *Angew. Chem.* **72** (1960) 856 – 864.
613. H. Finkentscher, *Kunststoffe* **53** (1963) 734 – 740.
614. C. Heuck, *Chem. Ztg.* **94** (1970) 147 – 157.
615. W. D. Harkins, *J. Chem. Phys.* **13** (1945) 381 – 382.
616. W. D. Harkins, *J. Chem. Phys.* **14** (1946) 47 – 48.
617. W. D. Harkins, *J. Am. Chem. Soc.* **69** (1947) 1428 – 1444.
618. W. D. Harkins, *J. Polym. Sci.* **5** (1950) 217 – 251.
619. B. M. E. van der Hoff, *Adv. Chem. Ser.* **34** (1962) 6 – 31.
620. J. Gardon, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 2859 – 2879.
621. H. Gerrens, *Z. Elektrochem.* **60** (1956) 400 – 404.
622. H. Gerrens, *Ber. Bunsenges. Phys. Chem.* **67** (1963) 741 – 753.
623. W. V. Smith, R. H. Ewart, *J. Chem. Phys.* **16** (1948) 592 – 599.
624. W. H. Stockmayer, *J. Polym. Sci.* **24** (1957) 314 –317.
625. J. T. O'Toole, *Appl. Polym. Sci.* **9** (1965) 1291 – 1297.
626. E. Jahnke, F. Emde: *Funktionentafeln mit Formeln und Kurven,* 2nd ed., B. G. Teubner, Leipzig 1933.
627. M. Abramowitz, I. A. Stegun: *Handbook of Mathematical Functions,* 10th ed., Dover, New York 1972.
628. J. L. Gardon, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) a 623 – 641, b 643 – 664, c 665 – 685, d 687 – 710, e 2853 – 2857.
629. J. Ugelstad, F. K. Hansen, K. Herder-Kaggerud, *Faserforsch. Textiltech.* **28** (1977) 309 – 320.
630. N. Friis, A. E. Hamielec, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 3321 – 3325.
631. M. R. Grancio, D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 2617 – 2629, 2733 – 2745.
632. P. Keusch, J. Prince, D. J. Williams, *J. Macromol. Sci. Chem.* **7** (1973) 623 – 646. P. Keusch, D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 143 – 162.
633. P. Keusch, R. A. Graff, D. J. Williams, *Macromolecules* **7** (1974) 304 – 310. D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 301 – 303. D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2123 – 2132.
634. J. L. Gardon, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 241 – 251.

635. J. L. Gardon, *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2133 – 2135.
636. M. Morton, S. Kaizerman, M. W. Altier, *J. Colloid Sci.* **9** (1954) 300 – 312.
637. E. Vanzo, R. H. Marchessault, V. Stannet, *J. Colloid Sci.* **20** (1965) 62 – 71.
638. D. Hummel, G. Ley, C. Schneider, *Adv. Chem. Ser.* **34** (1962) 60 – 86.
639. G. V. Schulz, J. Romatowski, *Makromol. Chem.* **85** (1965) 195 – 226. J. Romatowski, G. V. Schulz, *Makromol. Chem.* **85** (1965) 227 – 248.
640. J. Ugelstad, P. C. Mörk, J. O. Aasen, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 2281 – 2288.
641. J. Ugelstad, P. C. Mörk, P. Dahl, P. Rangnes, *J. Polym. Sci. Polym. Symp.* **27** (1969) 49 – 68.
642. J. Ugelstad, P. C. Mörk, *Br. Polym. J.* **2** (1970) 31 – 39.
643. F. K. Hansen, J. Ugelstad, *Makromol. Chem.* **180** (1979) 2423 – 2434. *Polym. Sci. C, Polym. Lett.* **26** (1988) 385.
644. B. W. Brooks, *Colloid Polym. Sci.* **265** (1987) 58.
645. B. S. Hawkett, D. H. Napper, R. G. Gilbert, *J. Chem. Soc., Faraday Trans.* **76** (1980) 1323.
646. H. C. Lee and G. W. Poehlein, *J. Dispersion Sci. Technol.* **5** (1984) 247.
647. B. W. Brooks, B. O. Makanjuola, *J. Chem. Soc., Faraday Trans.* **77** (1981) 2659.
648. G. Lichti et al., *J. Chem. Soc., Faraday Trans.* **80** (1984) 2911.
649. M. Nomura, M. Kubo, K. Fugita, *J. Appl. Polym. Sci.* **28** (1983) 2767.
650. M. Nomura, I. Horie, M. Kubo, K. Fujita, *J. Appl. Polym. Sci.* **37** (1989) 1029.
651. W. V. Smith, *J. Am. Chem. Soc.* **70** (1948) 2177 –2179; **71** (1949) 4077 – 4082.
652. M. Morton, P. P. Salatiello, H. Landfield, *J. Polym. Sci.* **8** (1952) 111 – 121.
653. M. Morton, P. P. Salatiello, *J. Polym. Sci.* **8** (1952) 279 – 287.
654. I. M. Kolthoff et al.: *Emulsion Polymerization,* Interscience, New York 1955, p. 191 – 201.
655. E. Bartholomé, H. Gerrens, R. Herbeck, H. M. Weitz, *Z. Elektrochem.* **60** (1956) 334 – 348.
656. H. Gerrens, E. Köhnlein, *Z. Elektrochem.* **64** (1960) 1199 – 1210.
657. Z. Maňyásek, A. Řežábek, *J. Polym. Sci.* **56** (1962) 47 – 56.
658. K. P. Paoletti, F. W. Billmeyer, *J. Polym. Sci., Part A* **2** (1964) 2049 – 2062.
659. D. M. French, *J. Polym. Sci.* **32** (1958) 395 – 411.
660. M. K. Lindemann: "The Mechanism of Vinyl Acetate Polymerization," in G. E. Ham (ed.): : *Kinetics and Mechanisms of Polymerization,* vol. 1 part I, Marcel Dekker, New York 1967, p. 207 – 329, 259 – 290.
661. R. Patsiga, M. Litt, V. Stannett, *J. Phys. Chem.* **64** (1960) 801 – 804.
662. M. Litt, R. Patsiga, V. Stannett, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 3607 – 3649.
663. N. Friis, L. Nyhagen, *J. Appl. Polym. Sci.* **17** (1973) 2311-2327.
664. M. Nomura et al., *J. Chem. Eng. Jpn.* **4** (1971) 160 – 166.
665. A. G. Parts, D. E. Moore, J. G. Watterson, *Makromol. Chem.* **89** (1965) 156 – 164.
666. M. Harada et al., *J. Appl. Polym. Sci.* **16** (1972) 811 – 833.
667. M. Nomura, M. Harada, W. Eguchi, S. Nagata, *ACS Div. of Polymer Chemistry, Polymer Preprints* **16** (1975) no. 1, 217 – 222.
668. F. K. Hansen, J. Ugelstad, *J. Polym. Sci. Polym. Chem. Ed.* **16** (1978) 1953 – 1979.
669. P. V. Danckwerts, *Chem. Eng. Sci.* **2** (1953) 1 – 13.
670. C. P. Roe, *Ind. Eng. Chem.* **60** (1968) 20 – 33.
671. R. M. Fitch, *Off. Dig J. Paint Technol. Eng.* **37** (1965), no. 10, 32 – 48.
672. R. M. Fitch, M. B. Prenosil, K. J. Sprick, *J. Polym. Sci. Polym. Symp.* **27** (1969) 95 – 118.
673. R. M. Fitch, C. H. Tsai, *J. Polym. Sci. Polym. Lett. Ed.* **8** (1970) 703 – 710.
674. R. M. Fitch, *Br. Polym. J.* **5** (1973) 467 – 483.
675. R. M. Fitch, Lih-Bin Shih, *Progr. Colloid Polym. Sci.* **56** (1975) 1 – 11.
676. J. W. Vanderhoff, J. F. Vitkuske, E. B. Bradford, T. Alfrey, *J. Polym. Sci.* **20** (1956) 225 – 234.
677. N. Dezelić, J. J. Petres, G. J. Dezelić, *Kolloid Z. Z. Polym.* **242** (1970) 1142 – 1150. N. Dezelić, J. J. Petres, G. J. Dezelić in J. E. Mulvaney (ed.): *Macromolecular Syntheses,* **vol. 6,** Wiley, New York 1977, p. 85 – 89.
678. A. R. Goodall, M. C. Wilkinson, J. Heran, *J. Colloid Interface Sci.* **53** (1975) 327 – 331.
679. V. I. Yeliseyeva, A. V. Zuikov, *ACS Symp. Ser.* **24** (1976) 62 – 81.
680. M. V. Smoluchowski, *Z. Phys. Chem.* **92** (1918) 129 – 168.
681. J. Ugelstad, M. S. El-Aasser, J. W. Vanderhoff, *J. Polym. Sci. Polym. Lett. Ed.* **11** (1973) 503 – 513.
682. J. Ugelstad, H. Fløgstad, F. K. Hansen, T. Ellingsen, *J. Polym. Sci. Polym. Symp.* **42** (1973) 473 – 485.
683. J. Ugelstad, F. K. Hansen, S. Lange, *Makromol. Chem.* **175** (1974) 507 – 521.

684. A. R. M. Azad, J. Ugelstad, R. M. Fitch, F. K. Hansen, *ACS Symp. Ser.* **24** (1976) 1 – 23. D. P. Durbin, M. S. El-Aasser, G. W. Poehlein, J. W. Vanderhoff, *J. Appl. Polym. Sci.* **24** (1979) 703 – 707.
685. H. Tang, P. L. Johnson, E. Gulari, *Polymer* **25** (1984) 1357.
686. G. I. Litvinenko, V. A. Kaminski, M. G. Slinko, *Khim. Promst. (Moscow)* **8** (1984) 463.
687. J. M. Asua, V. S. Rodriguez, C. A. Silebi, M. S. El-Aasser, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 59. Pergamon Press, Oxford 1989, p. 171.
688. Z. Song, G. W. Poehlein, *J. Macromol. Sci. Chem.* **A 25** (1988) 403, 1587.
689. Z. Song, G. W. Poehlein, *J. Colloid Interface Sci.* **128** (1989) 486, 501.
690. Z. Song, G. W. Poehlein, *Polym. Plast. Technol. Eng.* **29** (1990) 377.
691. M. Nomura, U. S. Satpathy, Y. Kouno, K. Fujita, *J. Polym. Sci.* **37** (1989) 1029.
692. B. P. Huo et al., *J. Appl. Polym. Sci.* **35** (1988) 2009.
693. A. Penlidis, J. F. MacGregor, A. E. Hamielec, *J. Appl. Polym. Sci.* **35** (1988) 2023.
694. *Dispersionen synthetischer Hochpolymerer,* Springer, Berlin 1969, F. Hölscher: "Eigenschaften, Herstellung und Prüfung," part 1, H. Reinhard: "Anwendung," part 2.
695. H. Warson: *The Applications of Synthetic Resin Emulsions,* Ernest Benn, London 1972.
696. N. Friis, A. E. Hamielec, *J. Appl. Polym. Sci.* **19** (1975) 97.
697. K. W. Min, W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 11** (1974) 177.
698. C. Kiparissides, S. R. Ponnuswamy, *Chem. Eng. Commun.* **10** (1981) 283.
699. E. Giannetti in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Publishers, Basel 1989, p. 343.
700. D. C. Sundberg, *J. Appl. Polym. Sci.* **23** (1979) 2197.
701. J. L. Gardon, *J. Polym. Sci.* **A 6** (1968) 665.
702. S. Katz, R. Shinnar, G. M. Saidel, *Adv. Chem. Ser.* **91** (1969) 145.
703. K. W. Min, W. H. Ray, *J. Appl. Polym. Sci.* **22** (1968) 89. K. W. Min, W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 11** (1974) 177.
704. D. C. Sundberg, J. D. Eliassen: *Polymer Colloids,* Plenum, New York 1971, p. 121 – 137.
705. C. C. Lin, W. Y. Chin, *J. Appl. Polym. Sci.* **23** (1979) 2049.
706. G. Lichti, R. G. Gilbert, D. H. Napper, *J. Polym. Sci. Polym. Ed.* **18** (1980) 1297.
707. E. Giannetti, G. Storti, M. Morbidelli, *J. Polym. Sci.,* Part A: Chem. Ed. 26 (1988) 1835, 2307.
708. K. W. Min, W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 11** (1974) 177.
709. N. Friis, A. E. Hamielec, *J. Appl. Polym. Sci.* **19** (1975) 97. H. Tobita, *Macromolecules* **25** (1992) 2671.
710. T. O. Broadhead, A. E. Hamielec, J. F. MacGregor, *Makromol. Chem., Suppl.* **10/11** (1985) 105.
711. X. Z. Kong, C. Pichot, J. Buillot, *Eur. Polym. J.* **24** (1988) 485.
712. J. Guillot in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig & Wepf, Basel 1986, p. 147.
713. F. V. Loncar, M. S. El-Aasser, J. W. Vanderhoff, *Polym. Mater. Sci. Eng.* **54** (1986) 453.
714. M. H. Andrus, *Polym. Mater. Sci. Eng.* **54** (1986) 603.
715. J. W. Vanderhoff, *NATO ASI Ser. E* : 138 (1987) 23.
716. V. Dimonie, M. S. El-Aasser, A. Klein, J. W. Vanderhoff, *J. Polym. Sci. Polym. Chem. Ed.* **22** (1984) 2197.
717. H. Schuller in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig & Wepf, Basel 1986, p. 137.
718. M. Okubo, A. Yamada, R. Matsumato, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 3219.
719. M. Okubo, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 307.
720. A. Zosel, W. Heckelmann, G. Ley, W. Mächtle, *Makromol. Chem., Macromol. Symp.* **35/36** (1990) 423.
721. D. Distler, G. Kanig, *Colloid Polym. Sci.* **256** (1978) 1052 – 1060.
722. K. W. Min, H. I. Gostin, *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1979) 272.
723. G. Lichti, R. G. Gilbert, D. H. Napper, *J. Polym. Sci. Polym. Chem. Ed.* **21** (1983) 269.
724. M. Cociani et al. in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Publishers, Basel 1989, p. 199.
725. A. W. DeGraff, G. W. Poehlein, *J. Polym. Sci. A 2* **9** (1971) 1955.
726. G. C. Lin, W. Y. Chin, L. C. Huang, *J. Appl. Polym. Sci.* **25** (1980) 567.
727. P. J. Feeney, D. H. Napper, R. G. Gilbert, *Macromolecules* **17** (1984) 2520.
728. J. L. Guillaume, C. Pichot, A. Revillon, *Macromol. Chem. Suppl.* **10/11** (1985) 69.
729. A. S. Dunn, *Macromol. Chem., Suppl.* **10/11** (1985) 1.

730. J. Snuparek, Z. Kleckova, *J. Appl. Polym. Sci.* **29** (1984) 1.
731. T. Hjertberg, E. M. Sorvik, *J. Polym. Sci., Part A, Polym. Chem. Ed.* **24** (1986) 1313.
732. K. Tauer, I. Mueller, *Chem. Abstracts* **101** (1984) 7729 g.
733. B. Jacobi, *Angew. Chem.* **64** (1952) 539 – 543.
734. S. H. Maron et al., *J. Colloid Sci.* **9** (1954) 89 – 103, 104 – 112, 263 – 268, 347 – 352, 353 – 358, 382 – 384.
735. W. Hübinger, K. H. Reichert in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Publishers, Basel 1989, p. 207.
736. A. S. Dunn: "Polymerization in Micelles and Microemulsions," in *Comprehensive Polymer Science,* Pergamon Press, Oxford 1989, p. 219.
737. F. Candau: "Polymerization in Inverse Microemulsions," in *Comprehensive Polymer Science,* Pergamon Press, Oxford 1989, p. 225.
738. F. Candau in F. Candau, R. H. Ottewill (eds.): *Scientific Methods for the Study of Polymer Colloids and their Applications,* **vol. XIII,** Kluwer Publishers, Dordrecht, Boston 1991, p. 73.
739. D. Hunkeler, A. E. Hamielec, *Polymer* **32** (1991) 2626.
740. Dow, US 3 284 393, 1959.
741. Nalco, US 3 624 019, 1970.
742. BASF, DAS 2 432 699, 1974.
743. IG-Farbenind., DT 663 469, 1930.
744. IG-Farbenind., DT 900 019, 1937.
745. Houben-Weyl, 4. ed., "Makromolekulare Stoffe," vol. XIV–1, part 1, p. 159, 186, 333, 337, 874, 878, 912, 1051.
746. IG-Farbenind., DT 629 220, 1933.
747. K. Chujo, Y. Harada, S. Tokuhara, *J. Polym. Sci. Polym. Symp.* **27** (1969) 321 – 332.
748. J. Sňupárek, *Angew. Makromol. Chem.* **25** (1972) 105 – 112, 113 – 119.
749. J. Šňupárek, *Angew. Makromol. Chem.* **37** (1974) 1 – 9.
750. J. Šňpárek, F. Krška, *J. Appl. Polym. Sci.* **20** (1976) 1753 – 1764.
751. J. Šňupárek, F. Krška, *J. Appl. Polym. Sci.* **21** (1977) 2253 – 2260.
752. V. I. Yeliseyeva, N. G. Zharkova, A. V. Chubarova, P. I. Zubov, *Polym. Sci. USSR (Engl. Transl.),* **7** (1965) 171 – 178.
753. V. I. Yeliseyeva, P. I. Zubov, U. F. Malofeyevskaya, *Polym. Sci. USSR (Engl. Transl.)* **7** (1965) 1496 – 1504.
754. V. I. Yeliseyeva et al., *Polym. Sci. USSR (Engl. Transl.)* **9** (1967) 813 – 822.
755. V. I. Yeliseyeva et al., *Polym. Sci. USSR (Engl. Transl.)* **11** (1969) 1136 – 1144.
756. H. Gerrens, *Kolloid Z. Z. Polym.* **227** (1968) 92 – 107.
757. J. J. Krackeler, H. Naidus, *J. Polym. Sci. Polym. Symp.* **27** (1969) 207 – 235.
758. S. Jovanović, J. Romatowski, G. V. Schulz, *Makromol. Chem.* **85** (1965) 187 – 194.
759. J. E. Vandegaer, *J. Appl. Polym. Sci.* **9** (1965) 2929 – 2938.
760. J. J. Owen, C. T. Steele, P. T. Parker, E. W. C. Carrier, *Ind. Eng. Chem.* **39** (1947) 110 – 113.
761. R. W. Laundrie, R. F. McCann, *Ind. Eng. Chem.* **41** (1949) 1568 – 1570.
762. P. Baumann, *Chimia* **12** (1958) 233 – 245.
763. W. Hofmann: *Nitrilkautschuk,* Berliner Union, Stuttgart 1965.
764. Du Pont, US 2 384 277, 1939.
765. Du Pont, US 2 831 842, 1953.
766. Union Carbide, US 2 984 703, 1956.
767. Uniroyal, GB 1 168 760, 1967.
768. Monsanto, GB 1 154 820, 1966.
769. Societá Italiana Resine, FR 2 074 118, 1970.
770. R. D. Dunlop, F. E. Reese, *Ind. Eng. Chem.* **40** (1948) 654 – 660.
771. D. B. Gershberg, J. E. Longfield, *Symp. Polymerization Kinetics and Catalyst Systems,* part 1, preprint no. 10, 54th Meeting A. I. Ch. E., New York, Dec. 2 – 7, 1961.
772. A. W. DeGraff, G. W. Poehlein, *J. Polym. Sci. Polym. Phys. Ed.* **9** (1971) 1955 – 1976.
773. M. Nomura et al., *J. Appl. Polym. Sci.* **15** (1971) 675 – 691.
774. T. Ueda, S. Omi, H. Kubota, *J. Chem. Eng. Jpn.* **4** (1971) 50 – 54.
775. S. Omi, T. Ueda, H. Kubota, *J. Chem. Eng. Jpn.* **2** (1969) 193 – 198.
776. H. Gerrens, K. Kuchner, *Br. Polym. J.* **2** (1970) 18 – 24.
777. H. Gerrens, K. Kuchner, G. Ley, *Chem.-Ing.-Tech.* **43** (1971) 693 – 698.
778. G. Ley, H. Gerrens, *Makromol. Chem.* **175** (1974) 563 – 581.
779. B. W. Brooks, H. W. Kropholler, S. N. Purt, *Polymer* **19** (1978) 193 – 196.
780. M. Ghosh, T. H. Forsyth, *ACS Symp. Ser.* **24** (1976) 367 – 378.
781. J. D. Stevens, J. O. Funderburk, *Ind. Eng. Chem. Process. Des. Dev.* **11** (1972) 360 – 369.
782. G. Beckmann, H. Matis, *Chem.-Ing.-Tech.* **38** (1966) 209 – 214.
783. S. Senrui, A. Kodama, M. Takehisa, *J. Polym. Sci. Chem. Ed.* **12** (1974) 2403 – 2417.
784. A. R. Berens, *J. Appl. Polym. Sci.* **18** (1974) 2379 – 2390.

785. A. Penlidis, A. E. Hamielec, J. F. MacGregor, *J. Vinyl Technol.* **6** (1984) 134.
786. R. K. Greene, R. A. Gonzalez, G. W. Poehlein, *ACS Symp. Ser.* **24** (1976) 341 – 358.
787. C. Kiparissides, J. F. MacGregor, A. E. Hamielec, *J. Appl. Polym. Sci.* **23** (1979) 401 – 418.
788. J. B. Rawlings, W. H. Ray, *AIChE J.* **33** (1987) 1663.
789. J. B. Rawlings, J. C. Prindle, W. H. Ray, in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig & Wepf Verlag, New York 1986, p. 1.
790. J. B. Rawlings, W. H. Ray, *Polym. Eng. Sci.* **42** (1987) 2767.
791. J. B. Rawlings, W. H. Ray, *Polym. Eng. Sci.* **28** (1988) 237, 257.
792. G. W. Poehlein, D. J. Dougherty, *Rubber Chem. Technol.* **50** (1977) 601 – 638.
793. A. Penlidis, J. F. MacGregor, A. E. Hamielec: "Continuous Emulsion Polymerization: Design and Control of CSTR Trains," *Chem. Eng. Sci.* **44** (1989) 273.
794. B. W. Brooks, *Br. Polym. J.* **5** (1973) 199 – 211.
795. O. Levenspiel: *Chemical Reaction Engineering,* Wiley, New York 1962, p. 145.
796. T. Matsuura, M. Kato, *Chem. Eng. Sci.* **22** (1967) 171 – 183.
797. R. S. Knorr, K. F. O'Driscoll, *J. Appl. Polym. Sci.* **14** (1970) 2683 – 2696.
798. M. Nomura, personal communication.
799. R. K. Greene, Ph. D. Thesis, Dept. of Chem. Eng., Lehigh Univ., Bethlehem, 1976.
800. R. F. Dickinson, C. E. Gall, *59th CIC Conference,* London, Ontario, June 1976.
801. P. V. Danckwerts, *Chem. Eng. Sci.* **8** (1958) 93 – 102.
802. Th. N. Zwietering, *Chem. Eng. Sci.* **8** (1958) 101.
803. Th. N. Zwietering, *Chem. Eng. Sci.* **11** (1959) 1 – 15.
804. E. B. Naumann, *J. Macromol. Sci. Rev. Macromol. Chem.* **10** (1974) 75 – 112.
805. E. Fitzer, W. Fritz: *Technische Chemie. Eine Einführung in die Chemische Reaktionstechnik,* Springer, Berlin 1975, p. 352 ff.
806. G. V. Schulz, *Z. Phys. Chem. Abt. B* **43** (1939) 25 – 46.
807. F. J. Schork, W. H. Ray in D. R. Bassett, A. E. Hamielec (eds.): "Emulsion Polymers and Emulsion Polymerization," *ACS Symp. Ser.* **165** (1981).
808. F. J. Schork, W. H. Ray, *J. Appl. Polym. Sci.* **28** (1983) 407.
809. M. S. Nomura et al., *ACS Meeting Preprints,* San Francisco 1980.
810. H. C. Lee, G. W. Poehlein, *Chem. Eng. Sci.* **41** (1986) 1023.
811. C. H. Lee, R. G. Mallinson, *AIChE J.* **34** (1988) 840.
812. J. B. Rawlings, W. H. Ray, *Chem. Eng. Sci.* **42** (1987) 2767.
813. R. N. Mead, G. W. Poehlein, *Ind. Eng. Chem.* **28** (1989) 51.
814. M. Nomura in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hüthig & Wepf, Basel 1986, p. 42.
815. M. Pollock, J. F. MacGregor, A. E. Hamielec, *ACS Symp. Ser.* **197** (1982) 209.
816. H. Tobita: "Crosslinking Kinetics in Emulsion Polymerization," *Polymer* **25** (1992) 2671.
817. M. Nomura, K. Fujita: "Particle Nucleation in Emulsion Copolymerization Containing Multifunctional Monomers," *Polym. Int.* (1992).
818. M. Buback, *Makromol. Chem.* **181** (1980) 373.
819. Y. Tatsukami, T. Takahashi, H. Yoshioka, *Makromol. Chem.* **181** (1980) 1108.
820. S. Seidl, G. Luft, *J. Macromol. Sci. Chem.* **A 15** (1981) 1.
821. S. Goto, K. Yamamoto, S. Furui, M. Sugimoto, *J. Appl. Polym. Sci., Appl. Polym. Symp.* **36** (1981) 21.
822. G. Donati et al., *ACS Symp. Ser.* **196** (1982) 579.
823. M. Buback, H. Lendle, *Makromol. Chem.* **184** (1983) 193.
824. W. Hollar, P. Ehrlich, *Chem. Eng. Commun.* **24** (1983) 57.
825. Y. Ogo: "Polymerizations at High Pressures," *J. Macromol. Sci. Revs. Macromol. Chem. Phys.* **C 24** (1984) 1.
826. P. Feucht, B. Tilger, G. Luft, *Chem. Eng. Sci.* **40** (1985) 1935.
827. H. Mavridis, C. Kiparissides, *Polym. Process Eng.* **3** (1985) 263.
828. S. K. Gupta, A. Kumar, M. V. G. Krishnamurthy, *Polym. Sci. Eng.* **25** (1985) 1.
829. B. J. Yoon, H. K. Rhee, *Chem. Eng. Commun.* **24** (1985) 253.
830. P. P. Shirodkar, G. O. Tsien, *Chem. Eng. Sci.* **41** (1986) 1031.
831. P. Postelnicescu, A. M. Dumitrescu, *Materiale Plastice* **24** (1987) 2.
832. S. K. Gupta, *Current Science* **56** (1987) 19.
833. M. Tjahjadi, S. K. Gupta, M. Morbidelli, A. Varma, *Chem. Eng. Sci.* **42** (1987) 2385.

834. A. Brandolin, N. J. Capiati, J. N. Farber, E. M. Valles, *Ind. Eng. Chem. Res.* **27** (1988) 784.
835. J. K. Beasley: "Polymerization at High Pressure," in *Comprehensive Polymer Science,* vol. 3, Pergamon Press, Oxford 1989, p. 273.
836. R. C. M. Zabisky, W.-M. Chan, P. E. Gloor, A. E. Hamielec: "A Kinetic Model for Olefin Polymerization in High Pressure Tubular Reactors–A Review and Update," *Polymer* **33** (1992) 2243.
837. J. W. Hamer, W. H. Ray, *Chem. Eng. Sci.* **41** (1986) 3083.
838. S. S. Agarwall, C. Kleinstreuer, *Chem. Eng. Sci.* **41** (1986) 3101.
839. C. Tzoganakis: "Reactive Extrusion of Polymers, A Review," *Adv. Polym. Techn.* **9** (1989) 321.
840. D. Suwanda, R. Lew, S. T. Balke, *J. Appl. Polym. Sci.* **35** (1988) 1019, 1033.
841. C. Tzoganakis, J. Vlachopoulos, A. E. Hamielec, *Polym. Eng. Sci.* **28** (1988) 170.
842. G. Foster: "Processing and Product Performance Control" in *Polymer Reaction Engineering Course Notes,* McMaster University, Hamilton, Canada 1990.
843. G. M. Gale: "Crosslinked Polyethylene Extrusions Using Direct Silane Injection," *ANTEC* /84, Society of Plastics Engineers, 1984, p. 2.
844. L. Panzer, "Advances in Silane Crosslinking of Polyethylene," *ANTEC* /88, Society of Plastics Engineers, 1988, p. 1421.
845. R. Greco et al., *Polym. Proc. Eng.* **4** (1986) 253.
846. A. Hogt, *ANTEC* /88, Society of Plastics Engineers, 1988, p. 1478.
847. S. Al-Malaika in J. L. Benham, J. F. Kinstle (eds.): *Chemical Reactions on Polymers,* ACS, Washington 1988, chap. 29.
848. R. C. Kowalski, *Chem. Eng. Prog.* **85** (1989) no. 5, 67.
849. M. Saleem, W. E. Baker, *J. Appl. Polym. Sci.* **39** (1990) 655.
850. A. E. Hamielec, P. E. Gloor, S. Zhu, "Chemical Modification of High Molecular Weight Polymers in Extruders – Experimentation and Computer Modelling of the Kinetics of Chain Scission, Long Chain Branching, Crosslinking and Grating" in *4th International Congress on Compatibilizers and Reactive Polymer Alloying,* New Orleans 1991.
851. A. E. Hamielec, P. E. Gloor, S. Zhu, *Canad. J. Chem. Eng.* **69** (1991) 611.
852. K. Motha, J. Seppala, C. Bergstrom, *Polym. Eng. Sci.* **29** (1989) 1579.
853. T. Kimura, *J. Phys. Soc. Jpn.* **17** (1962) 1884.
854. A. Charlesby and S. H. Pinner, *Proc. R. Soc. London A* **249** (1959) 367.
855. C. Tzoganakis, J. Vlachopoulos, A. E. Hamielec, *Intl. Polym. Proc.* **3** (1988) 141.
856. A. Hrymak, McMaster University, personal comunication 1990.
857. J. V. Seppala, M. Auer: "Factors Affecting Kinetics in Slurry Type Coordination Polymerization," *Prog. Polym. Sci.* **15** (1990) 147.
858. I. D. Burdett, "The Union Carbide UNIPOL Process: Polymerization of Olefins in a Gas-Phase Fluidized Bed," *AIChE Annual Meeting,* Washington, D.C., Nov./Dec., 1988.
859. K. Y. Choi, W. H. Ray, *J. Macromol. Sci., Rev. Macromol. Chem.* **25** (1985) 57.
860. W. H. Ray, *Ber. Bundesanst. Ges. Phys. Chem.* **90** (1986) 947.
861. B. de Carvalho, P. Gloor, A. E. Hamielec, *Polymer* **30** (1989) 280.
862. K. B. McAuley, J. F. MacGregor, A. E. Hamielec, *AIChE J.* **36** (1990), 837.
863. R. Galvan, M. Tirrell, *Chem. Eng. Sci.* **41** (1986) 2385.
864. B. de Carvalho, P. Gloor, A. E. Hamielec, *Polymer* **31** (1990) 1294.
865. S. Floyd et al., *J. Appl. Polym. Sci.* **33** (1987) 1021.
866. R. L. Laurence, G. N. Chiovetta in K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* Hanser Verlag, München 1983, p. 74.
867. E. D. Nagel, A. V. Kirillov, W. H. Ray, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 372.
868. S. Floyd, K. Y. Choi, T. W. Taylor, W. H. Ray, *J. Appl. Polym. Sci.* **32** (1986) 2231, 2935.
869. T. W. Taylor, K. Y. Choi, H. Yuan, W. H. Ray, "Physicochemical Kinetics of Liquid Phase Propylene Polymerization," in *Proceedings International Symposium on Transition Metal Catalyzed Polymerization,* Gordon and Breach, New York 1983.
870. J. M. Vela Estrada, A. E. Hamielec: "Measurement of the Bivariate Distribution of composition and Molecular weight for Binary Copolymers – A Review," *Polym. Reaction Engng.* (1992).
871. H. Gerrens, *Proc. 4th Intern./6th European Symp. Chem. Reaction Engng. Heidelberg,* **vol. 2,** April 6 –8, 1976, DECHEMA, Frankfurt/Main 1976, p. 585 – 614.
872. G. V. Schulz, *Z. Phys. Chem. Abt. B* **30** (1935) 379 –398.

873. P. J. Flory, *J. Am. Chem. Soc.* **58** (1936) 1877 – 1885.
874. E. F. G. Herington, A. Robertson, *Trans. Faraday Soc.* **38** (1942) 490 – 501.
875. G. Gee, H. M. Melville, *Trans. Faraday Soc.* **40** (1944) 240 – 251.
876. L. Küchler: *Polymerisationskinetik,* Springer, Berlin 1951, p. 74 ff.
877. C. H. Bamford, W. G. Barb, A. D. Jenkins, P. F. Onyon: *The Kinetics of Vinyl Polymerization by Radical Mechanisms,* Butterworth, London 1958, p. 279 ff.
878. K. G. Denbigh, *Trans. Faraday Soc.* **40** (1944) 352 –373.
879. K. G. Denbigh, *Trans. Faraday Soc.* **43** (1947) 648 –660.
880. K. G. Denbigh, *J. Appl. Chem.* **1** (1951) 227 – 236.
881. K. G. Denbigh: *Chemical Reactor Theory,* Cambridge University Press, Cambridge 1965, p. 104 – 107.
882. J. A. Biesenberger, Z. Tadmor, *J. Appl. Polym. Sci.* **9** (1965) 3409 – 3416.
883. J. A. Biesenberger, Z. Tadmor, *Polym. Eng. Sci.* **6** (1966) 299 – 305.
884. Z. Tadmor, J. A. Biesenberger, *J. Polym. Sci. Polym. Lett. Ed.* **3** (1965) 753 – 759.
885. Z. Tadmor, J. A. Biesenberger, *Ind. Eng. Chem. Fundam.* **5** (1966) 336 – 343.
886. S. Katz, G. M. Saidel, *AIChE J.* **13** (1967) 319 – 326.
887. R. Thiele, *Chem. Tech. (Leipzig)* **19** (1967) 221 –225.
888. R. Thiele, K.-D. Herbrich, J. Fischmann, *Chem. Tech. (Leipzig)* **22** (1970) 445 – 460.
889. N. Friis, A. E. Hamielec, *J. Appl. Polym. Sci.* **19** (1975) 97.
890. S. T. Balke, A. E. Hamielec, *J. Appl. Polym. Sci.* **17** (1973) 905 – 949.
891. J. H. Duerksen, A. E. Hamielec, J. W. Hodgins, *AIChE J.* **13** (1967) 1081 – 1086.
892. J. H. Duerksen, A. E. Hamielec, *J. Polym. Sci. Polym. Symp.* **25** (1968) 155 – 166.
893. A. E. Hamielec, J. W. Hodgins, K. Tebbens, *AIChE J.* **13** (1967) 1087 – 1091.
894. A. W. Hui, A. E. Hamielec, *J. Polym. Sci. Polym. Symp.* **25** (1968) 167 – 189.
895. N. Khac Tien, E. Flaschel, A. Renken in K. H. Reichert, W. Geiseler (eds.). *Polymer Reaction Engineering,* Hanser, New York 1983, p. 175. C. Cozewith, *J. Appl. Polym. Sci.* **15** (1971) 2855 – 2863.
896. R. W. Dunn, C. C. Hsu, *Adv. Chem. Ser.* **109** (1972) 85 – 86.
897. M. Rätzsch, R. Nitzsche, *Plaste Kautsch.* **17** (1970) 6 – 8.
898. L. Gold, *J. Chem. Phys.* **28** (1958) 91 – 99.
899. L. H. Peebles: *Molecular Weight Distributions in Polymers,* Interscience, New York 1971, p. 137 ff.
900. P. J. Flory, *J. Am. Soc.* **62** (1940) 1561 – 1565.
901. W. H. Abraham, *Ind. Eng. Chem. Fundam.* **2** (1963) 221 – 224.
902. H. Kilkson, *Ind. Eng. Chem. Fundam.* **3** (1964) 281 – 293.
903. T. E. Corrigan, M. J. Dean, *J. Makromol. Sci. Chem.* **2** (1968) 645 – 662.
904. L. L. Böhm et al., *Fortschr. Hochpolym. Forsch.* **9** (1972) 1 – 45.
905. H. Dosal, R. Raff, *Z. Phys. Chem. Abt. B* **32** (1936) 117 – 129.
906. P. J. Flory, *Chem. Rev.* **39** (1946) 137 – 197.
907. J. A. Biesenberger, *AIChE J.* **11** (1965) 369, 371 – 373.
908. W. H. Ray, *J. Macromol. Sci. Rev. Macromol. Chem.* **8** (1972) 1 – 56.
909. N. H. Smith, G. A. Sather, *Chem. Eng. Sci.* **20** (1965) 15 – 23.
910. W. H. Abraham, *Chem. Eng. Sci.* **25** (1970) 331 – 335.
911. W. Kuhn, *Ber. Dtsch. Chem. Ges. B* **63** (1930) 1503 – 1509.
912. P. Wittmer, F. Hafner, H. Gerrens, *Makromol. Chem.* **104** (1967) 101 – 119.
913. R. Simha, H. Branson, *J. Chem. Phys.* **12** (1944) 253 – 267.
914. W. H. Stockmayer, *J. Chem. Phys.* **13** (1945) 199 – 207.
915. W. Ring, *Dechema-Monogr.* **49** (1964) 75 – 97.
916. W. Ring, *Makromol. Chem.* **75** (1964) 203 – 207.
917. K. F. O'Driscoll, R. Knorr, *Macromolecules* **2** (1969) 507 – 515.
918. K. H. Reichert, H. U. Moritz: "Polymer Reaction Engineering," *Comprehensive Polymer Sience,* vol. 3, Pergamon Press, Oxford 1989, p. 327.
919. H. Tobita, A. E. Hamielec: "Control of Network Structure in Free-Radical Crosslinking Copolymerization," *Polymer* (1992)
920. H. Tobita, A. E. Hamielec, in "Computer Applications in Applied Science II," *ACS Symp. Ser.* **404** (1989) 242.
921. H. Tobita, A. E. Hamielec in P. J. Lemstra, L. A. Kleintjens (eds.): *Integration of Fundamental Polymer Science and Technology–4,* Elsevier Applied Science, London 1990, p. 33.

922. H. M. J. Boots: "Simulation Model for Densely Cross-linked Networks Formed by Chain-Reactions," in L. A. Kleintjens, P. J. Lemstra (eds.): *Integration of Fundamental Polymer Science and Technology,* Elsevier Applied Science, London 1986, p. 204.
923. C. C. Chang, J. W. Miller, Jr., G. R. Schorr, *J. Appl. Polym. Sci.* **39** (1990) 2395.
924. G. R. Meira, A. F. Johnson, *Polym. Eng. Sci.* **21** (1981) 415.
925. D. A. Couso, G. R. Meira, *Polym. Eng. Sci.* **24** (1984) 391.
926. D. A. Couso, L. M. Alassia, G. R. Meira, *J. Appl. Polym. Sci.* **30** (1985) 3249.
927. G. L. Frontini, G. E. Elicabe, D. A. Couso, G. R. Meira, *J. Appl. Polym. Sci.* **31** (1986) 1019.
928. L. M. Alassia, D. A. Couso, G. R. Meira, *J. Appl. Polym. Sci.* **36** (1988) 481.
929. G. L. Frontini, G. E. Elicabe, G. R. Meira, *J. Appl. Polym. Sci.* **33** (1987) 2165.
930. L. M. Alassia, G. L. Frontini, J. R. Vega, G. R. Meira, *J. Appl. Polym. Sci., Part C: Polym. Lett.* **26** (1988) 201.
931. J. W. Hamer, T. A. Akramov, W. H. Ray, *Chem. Eng. Sci.* **36** (1981) 1897.
932. A. D. Schmidt, W. H. Ray, *Chem. Eng. Sci.* **36** (1981) 1401.
933. J. A. Biesenberger, D. H. Sebastian, *Polym. Eng. Sci.* **16** (1976) 117.
934. D. H. Sebastian, J. A. Biesenberger, *Polym. Eng. Sci.* **19** (1979) 190.
935. L. S. Henderson, *Chem. Eng. Prog.* **83** (1987) 42.
936. A. Uppal, W. H. Ray, A. B. Poore, *Chem. Eng. Sci.* **29** (1974) 967.
937. A. Uppal, W. H. Ray, A. B. Poore, *Chem. Eng. Sci.* **31** (1976) 205.
938. J. B. Planeaux, K. F. Jensen, *Chem. Eng. Sci.* **41** (1986) 1497.
939. A. D. Schmidt, A. B. Clinch, W. H. Ray, *Chem. Eng. Sci.* **39** (1984) 419.
940. F. Teymour, W. H. Ray, *Chem. Eng. Sci.* **44** (1989) 1967.
941. M. J. Barandiaran, J. C. Prindle, W. H. Ray: "The Effects of Chain Transfer Agents on the Stability of Continuous Polymerization Reactors," *AIChE Meeting,* Washington, Nov. 27 – Dec. 2, 1988.
942. J. C. Prindle, W. H. Ray: "Emulsion Polymerization Model Development for Operation Below the CMC," *AIChE Meeting,* New York, Nov. 15 – 20, 1987.
943. R. A. Hutchinson, W. H. Ray, *J. Appl. Polym. Sci.* **34** (1987) 657.
944. K. Y. Choi, W. H. Ray, *Chem. Eng. Sci.* **40** (1985) 2261.
945. K. Y. Choi, W. H. Ray, *Chem. Eng. Sci.* **43** (1988) 2587.
946. Y. J. Lu, B. W. Brooks, *Chem. Eng. Sci.* **44** (1989) 857.
947. D. C. H. Chien, A. Penlidis,: "On-Line Sensors for Polymerization Reactors," *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 30** (1990) 1.
948. C. H. Sohl: "Process Sensors," in *Polymer Reaction Engineering,* Engineering Foundation (AIChE), Santa Barbara 1991.
949. J. F. MacGregor, A. Penlidis, A. E. Hamielec: "Control of Polymerization Reactors: A Review," *Polym. Process Eng.* **2** (1984) 179.
950. G. E. Elicabe, G. R. Meira: "Estimation and Control in Polymerization Reactors: A Review," *Polym. Eng. Sci.* **28** (1988) 121.
951. D. Kozub, J. F. MacGregor: "Feedback Control of Polymer Quality in Semi-Batch Copolymerization Reactors," *Chem. Eng. Sci.* **47** (1992).
952. H. Schuler: "Automation," *IFAC PRP-4,* Ghent, Belgium, Pergamon Press, London 1980, p. 369.
953. L. Gagnon, J. F. MacGregor: "State Estimation for Continuous Emulsion Polymerization," *Can J. Chem. Eng.* **69** (1991) 648.
954. K. McAuley, J. F. MacGregor: "Online Inference of Polymer Properties in an Industrial Polyethylene Reactor," *AIChE J.* **37** (1991) 825.
955. J. Roffel, T. W. Hoffman, J. F. MacGregor, *DYCORD 89,* Maastricht, Netherlands 1989.
956. J. F. MacGregor, A. E. Hamielec, A. Penlidis, M. Pollock, *IFAC PRP-5,* Antwerp, Belgium, Pergamon Press, London 1983, p. 291.
957. http://www.kvs.ch/bulletin2000/bulletin2_200-0/Kap2.htm

Polymerization Technology → Polymerization Processes

Polymers, Electrically Conducting

HERBERT NAARMANN, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **Synthetic Routes** 2
3. **Principles of Electrical Conduction** . . 3
4. **Orientation Processes** 4
5. **Types of Electrically Conducting Organic Materials** 5
5.1. **Polyacetylene** 5
5.2. **Polydiacetylenes** 7
5.3. **Polypyrrole** 8
5.4. **Polythiophene** 10
5.5. **Polyphenylene** 10
5.6. **Poly(Phenylene Sulfide)** 12
5.7. **Poly(Phenylene Vinylene)** 13
5.8. **Polyaniline** 13
5.9. **Miscellaneous Polymers** 15
6. **Uses** 18
7. **References** 18

1. Introduction

Electrically conducting polymers (ECPs) are materials with an extended system of C=C conjugated bonds. They are obtained by reduction or oxidation reactions (called doping), giving materials with electrical conductivities up to 10^5 S/cm. These materials differ from polymers filled with carbon black or metals because the latter are only conductive if the individual conductive particles are mutually in contact and form a coherent phase.

This review concerns the synthesis routes, polymerization techniques, doping, orientation, and development of well-defined, highly conducting polymeric materials. Their wide range of potential uses from electrodes in rechargeable batteries to organic transistors is limited by their vulnerability to air and moisture due to their highly conjugated structures and the doping agents. Electrically conducting materials are compiled, their specific properties and potential applications are described.

Numerous attempts have been made to synthesize "conductive organic materials". The first was the synthesis of polyaniline by F. GOPPELSROEDER in 1891 [1]. After decades interest grew in organic polymers as insulators, but not as electrical conductors.

In the late 1950s organic semiconductors became the focus of investigations. Preliminary studies in this field up until the mid 1960s are reviewed in [2]. The semiconducting polymers were termed "covalent organic polymers", "charge-transfer complexes", "organometallic polymers", "hydrogen-bonded polymers", and "mixed polymers". Highest conductivity values reached about 10^{-3} S/cm. In 1964 LITTLE theoretically evaluated the possibility of superconductivity in polymers and suggested a model, consisting of a polyene chain with cyanine, dye-like substituents [3]. In the same year systematic studies were presented based on aromatic and heterocyclic compounds exhibiting electrical conductivities of 0.5 S/cm [4], followed by studies correlating doping, pressure, irradiation, and chain length to conductivity, with values up to 100 S/cm [5].

Interest heightened and became acute from 1975 when IBM scientists showed that poly(sulfur nitride), $(SN)_n$, was superconductive [6] and MACDIARMID's group reported [7] the doping of polyacetylene films prepared by SHIRAKAVA [8] reaching conductivity values of 38 S/cm. Since then many expectations have been raised, but scientific progress and practical applications have been limited; they depend on the reproducible production of well-defined specimens, the determination of synthesis conditions, and the laws relating these conditions to product properties. Synthetic methods are improving; more easily processible, soluble, flexible materials are now being produced.

2. Synthetic Routes

The synthesis of electrically conducting polymers with conjugated −HC=CH− bonds requires the controlled coupling of a large number

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_429

of monomers (unsaturated compounds or compounds with difunctional groups). Alternatively, pendant groups attached to existing polymers can be cyclized to give conjugated systems.

Polymerization of unsaturated monomers gives polymers with long chains that are often branched and cross-linked and therefore insoluble. Alkynes polymerize as follows:

1) $n\,HC{\equiv}CH \longrightarrow \text{-}(CH{=}CH)\text{-}_n$
   insoluble polyacetylenes [9]
2) $n\,HC{\equiv}C{-}CH{=}CH_2 \longrightarrow \text{-}(CH{=}CH{-}CH{=}CH)\text{-}_n$
   insoluble polyacetylenes [9]
3) $n\,RC{\equiv}CH \longrightarrow \text{-}(CR{=}CH)\text{-}_n$
   soluble polyacetylenes where R denotes alkyl or aryl groups [10]
4) Alkyne or allene $\longrightarrow$ 460 °C thermal polymerization, insoluble polyenes and aromatics [11]

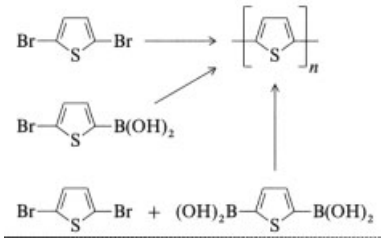

Stepwise coupling of monomers with difunctional groups runs in a more controlled fashion than polymerization of unsaturated monomers but problems are caused by low conversion rates and purification of the starting monomers. This method includes Wittig, MacMurry, retro Diels – Alder, and elimination reactions. It is accessible to a wide variety of soluble precursors and produces defined polymers. Examples follow:

1) Oxidative coupling [12] can be applied to pyrrole, thiophene, aromatic, and heteroaromatic systems (0 – 350 °C)

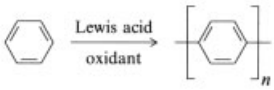

2) Grignard coupling [13]:

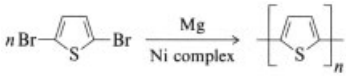

3) Boronic acid coupling [14]:

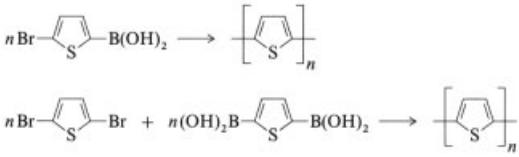

4) Stille reaction [15]:

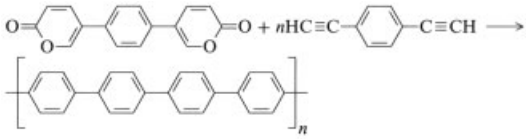

5) Elimination reactions [16]:

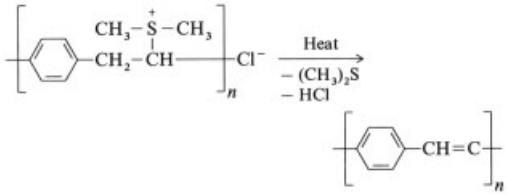

6) Feast reaction (retro- cycloaddition) [17]:

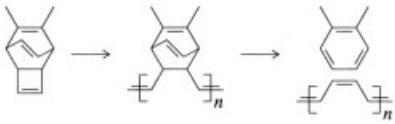

7) Grubb's method: polybenzvalene is isomerized to polyacetylene in the presence of $HgCl_2$ [18]:

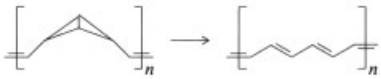

Cyclooctatetraene is also polymerized to give soluble polyenes [19]:

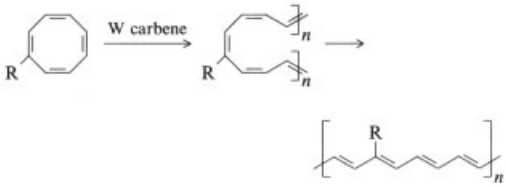

Polycyclization starts with polymers whose pendant side groups are cyclized to give condensed systems, often under pyrolytic conditions. Defects cannot be excluded.

These methods lead to the basic structures of electrically conducting organic materials (ECOMs) and are reviewed in [9]. Examples follow [12]:

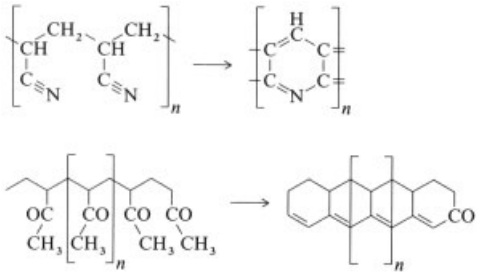

Analogous reactions occur with $-CH{=}CH_2$ pendant groups.

Monomers may also be cyclized [10, 20]:

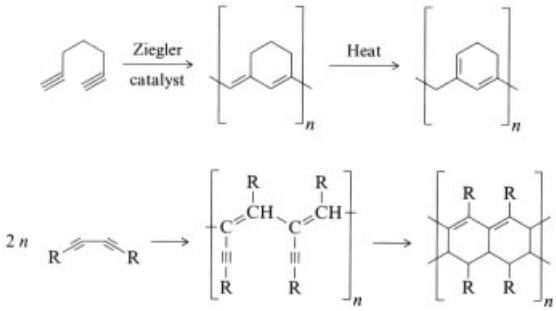

3. Principles of Electrical Conduction

The electrical properties of materials are determined by their electronic structure (Fig. 1). The band theory accounts for the different behaviors of metals, semiconductors, and insulators.

The band gap is the energy spacing between the highest occupied energy level (valence band) and the lowest unoccupied energy level (conduction band). Metals have a zero band gap which means that they have a high electron mobility, i.e., conductivity. Semiconductors have a narrow band gap (ca. 2.5 – 1.5 eV), conductivity only occurs on excitation of electrons from the valence band to the conduction band (e.g., by heating). If the band gap is larger (> 3 eV), electron excitation is difficult; electrons are unable to cross the gap and the material is an insulator [21]. Electrically conducting organic materials such as polyphenylene, polyacetylene, or polypyrrole are, however, peculiar in that the band theory cannot explain why the charge- carrying species (electrons or holes) are spinless. Conduction by polarons and bipolarons is now thought to be the dominant mechanism of charge transport in organic materials [22]. This concept also explains the drastic deepening of color changes produced by doping. A polaron (a term used in solid-state physics) is a radical cation that is partially delocalized over several monomer units (e.g., in a polymer segment). The bipolaron is a diradical dication. Low doping levels give rise to polarons, whereas higher doping levels produce bipolarons. Both polarons and bipolarons are mobile and can move along the polymer chain. See also, → Plastics, General Survey, Chap. 9.2.

Doping. The process that transforms insulating polymers (e.g., polyacetylene, conductivity 0.1 S/cm) to excellent conductors (Fig. 2) is the formation of change-transfer complexes by electron donors such as sodium or potassium (*n* doping, reduction) or by electron acceptors such as I_2, AsF_5, or $FeCl_3$ (*p* doping, oxidation). The doped polymer backbone becomes positively or negatively charged [23] with the dopant forming oppositely charged ions (Na^+, K^+, I_3^-, I_5^-, AsF_6^-, $FeCl_4^-$). The polymer can be switched between the doped, conductive state and the undoped, insulating state by applying an electric potential that makes the counterions move in and out. This switching corresponds to charging and discharging when these materials are used as electrodes in rechargeable batteries.

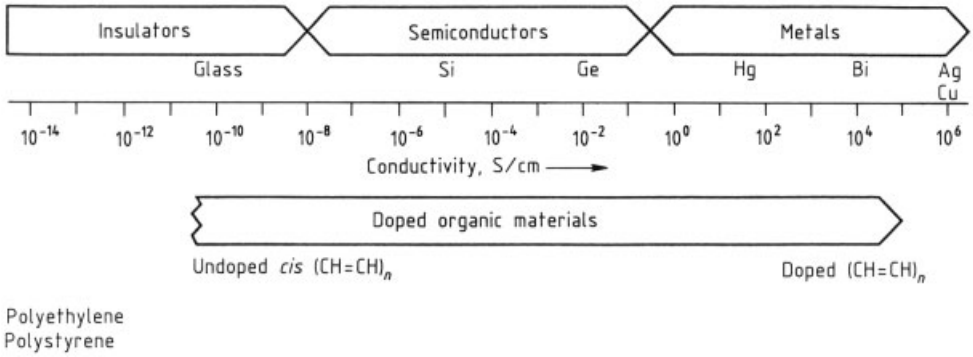

Figure 2. Comparison of the electrical conductivity (300 K) of organic and inorganic materials and the effect of doping

Measurement. Electrical conductance is a measure of the flow of current through a material for a given applied voltage [24].

The electrical conductivity σ in reciprocal ohms or siemens per centimeter (Ω^{-1} cm^{-1} or S/cm) denotes the current flow through a body of length 1 cm and with a cross-sectional area of 1 cm^2. The conductance σ is usually measured via its inverse, the electrical resistance R ($\sigma = 1/R$).

The resistivity is the inverse of the conductivity

$$\sigma = \frac{d}{F}\frac{1}{R}$$

(R in Ω, σ in Ω^{-1} cm^{-1}, d is the thickness of the specimen in cm, and F its surface area in cm^2).

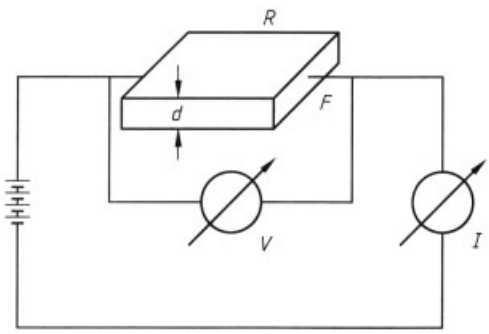

Figure 3. Conductivity measurement equipment

Ohm's law ($R = V/I$) is valid for electrical conductors (metals) and semiconductors and allows R to be determined by measuring the applied voltage V in volts and the current I in amperes with a simple measuring system (Fig. 3) or with a Wheatstone bridge.

With high resistances (i.e., $\sigma = 10^{-4}$ to 10^{-14} S/cm), two-point measurement is enough. Higher conductivities, as in electrically conducting polymers, require four-point measurements because the contact resistances in the circuit are frequently as great or greater than that of the material under investigation.

Contactless measurements of the conductivity are possible in a microwave field [24].

4. Orientation Processes

Orientation processes are powerful methods that are used to improve conductivity and other material properties (e.g., transparency, anisotropy). Orientation can be achieved in several ways, including stretching.

Mechanical stretching can be performed after polymerization, e.g., in noncross-linked polymers [25]. In the case of polyacetylenes prepared with aged Ziegler–Natta catalysts [26, 27], stretching increases conductivity from 2500 S/cm to values as high as 10^5 S/cm.

Continuous electrochemical polymerization (e.g., of polypyrrole) on the surface of a rotating drum permits simultaneous peeling off, mechanical stretching, and orientation

$\sigma \leq 200$ S/cm [25]. Greater stretching rates and therefore greater conductivities are reported in poly(pyrrole perchlorate) films (σ up to 10^3 S/cm) [28]. Biaxially stretched films yielded conductivities of 800 S/cm parallel to the stretching direction and 290 S/cm in the cross direction. Stretched poly(phenyl vinylenes) and poly(thienyl vinylenes) yielded conductivities of ca. 10^3 S/cm [29].

Orientation can also occur during polymerization or by performing polymerization in an oriented matrix consisting of liquid crystals and using magnetic fields [30]. Variants are the use of liquid crystal matrices during the electro chemical synthesis of polyheterocycles [31] and the synthesis of polymers (e.g., substituted thiophenes) with liquid crystal side chains that contain sulfonate groups [32]. The sulfonate groups act as " self dopants" and the liquid crystal side chains are responsible for orientation.

Polymerization of extremely thin polyacetylene films (< 1 µm) on crystal surfaces by epitaxial growth (e.g., on frozen benzene) [33] also induces orientation in the deposited polymer layer. Substituted polypyrrole films with a high anisotropy can be produced by the Langmuir–Blodgett technique [34–36].

5. Types of Electrically Conducting Organic Materials

Electrically conducting organic materials can be divided into three main groups: polyenes, polyheterocycles, and polyaminoaromatic compounds. Polyacetylene, polypyrrole, and polyaniline are the most intensively studied polymers.

5.1. Polyacetylene

Polyacetylene (PAC) exists in various isomeric forms:

cis-transoid (*cis*-polyacetylene)
[*25786-70-1*], [*74373-36-7*]

trans-cisoid
[*73589-68-1*]

trans-transoid (*trans*-polyacetylene)
[*25768-71-2*]

cis-Polyacetylene is relatively unstable and reverts to the thermodynamically stable *trans*-polyacetylene via the metastable *trans*- cisoid form. *cis*- cisoid PAC has not yet been prepared in pure form, it has a cyclic, helical structure [37].

The term "polyacetylene" was first mentioned in 1866 [38]. In 1898, E. Erdmann and P. Köthner [39] described a "high-molecular-weight copper acetylene" that was formed when acetylene gas was passed over copper powder below 250 °C. In 1948 Reppe [40] prepared Cuprene film with a metallic luster. In 1961 Hatano reported the polymerization of acetylene with a $AlEt_3/Ti(OBu)_4$ catalyst to give polymers with conductivities up to 0.001 S/cm [41]. Since then intensive work has been carried out on the various polymer types, reviews are given in [42]. Some important syntheses are listed in Table 1. Table 2 shows the conductivities of various types

The polyacetylenes grow primarily in the form of fibrils with an internal surface area of up to 200 m^2/g [45]. The fibrils are insoluble and can be rendered electrically conductive by incorporation of *n*-type or *p*-type dopants.

The polyacetylenes are chemically unstable and react, for example, with atmospheric oxygen to form epoxy, carbonyl, or carboxyl groups. $N-(CH{=}CH)_n$ compounds are more stable, probably due to the absence of defect sites at high crystallinity. Conductivity increases as a function of crystallinity and decreases as a function of defects (sp^3-bonded carbon atoms, $-CH_2CH_2-$) [46].

Figure 4 illustrates two types of polyacetylene. The standard Shirakawa type (Fig. 4A) is cross-linked and contains ca. 2 % sp^3 fractions. The new BASF technique involves polymerization at room temperature (instead of -76 °C) and the use of a tempered catalyst. The stretched polyacetylene product has parallel fibrils (Fig. 4B). It is linear (no sp^3 frac-

Table 1. Preparation and properties of polyacetylenes

Name of method	Catalyst	Polymerization-temperature, °C	*cis/trans* ratio	Defects (sp^3 fraction), %	Conductivity, S/cm (iodine doped)	References
Reppe	Ni acetylacetonate	100	5/95	15	10^{-3}	[43] (Table 1, no. 2)
Luttinger – Green	$NaBH_4$/CO (II)/Ni (II)	− 76	50/50	8	10^{-3}	[43] (Table 1, no. 6)
Hatano	$AlEt_3/Ti(OBu)_4$	100	20/80	10	10^{-5}	[43] (Table 2, no. 8)
Shirakawa	$AlEt_3/Ti(OBu)_4$ in toluene	− 76	90/10	3	520	[43] (Table 1, no. 16)
Naarmann (BASF)	1) $AlEt_3/Ti(OBu)_4$ in silicone oil	20	85/15	0	500	[44]
	2) as 1), but with an aged (> 100 °C) catalyst	20	85/15	0	2 000	[44]
	3) as 2) but oriented	20	85/15	0	$> 10^5$	[44,45]
Feast	thermal cleavage of tetracyclobis(trifluoromethyl)decatriene	80	75/25		100	[17]
Grubbs	isomerization of polybenzvalene,	20	40/60		1	[18]
	cyclooctatetraene ring opening by carbenes	20	40/60		350	[19]

Table 2. Conductivities of some metals and different types of polyacetylene

Metal or polyacetylene	Conductivity (volume), S/cm	Density[d], g/cm^3	Conductivity (weight), S cm^{-2} g^{-1}
Hg	10 365	13.54	767
Pt	101 522	21.45	4 733
Cr	77 519	7.19	10 781
Fe	102 986	7.87	13 085
Au	470 588	19.3	24 382
Ag	671 140	10.49	63 972
Cu	645 000	8.94	72 147
Shirakawa polyacetylene [8]	897 [a]		813
	30 [b]	1.23 (0.40)	22
Shirakawa polyacetylene, oriented in liquid-crystal matrix [30]	1 600 [b]	1.10 (0.50)	1 750
BASF polyacetylene, oriented [44]	23 000 [c]	1.12 (0.85)	20 000
BASF polyacetylene with additional reducing agent, oriented [44]	120 000 [c]	1.15 (0.90)	104 347

[a] Doped with $FeCl_3$. [b] Doped with I_2 vapor. [c] Doped with a saturated solution of I_2 in CCl_4. [d] Densities for polymers denote values after doping; values before doping are given in parentheses.of polyacetylenes and compares them with those of metals.

tions), is highly orientable (can be stretched by up to 660 %), and has a conductivity exceeding 10^5 S/cm. A convincing demonstration of the high anisotropy (1: 100) in the stretched samples is the fact that when strips of such a stretched polymer are laid across each other, polarized light (sunlight) is extinguished in the region of overlap [47].

Substituted Polyacetylenes. The synthesis of polymers from substituted acetylene monomers is directed toward the preparation of substituted, conjugated chains which ameliorate the negative properties of polyacetylenes (e.g., sensitivity to air, insolubility, and infusibility) while maintaining the desired electrical properties of acetylene's conjugated backbone [10, 48]. Alkyl- and aryl-substituted polymers result. They are soluble (e.g., in toluene and cyclohexane) and processible, but have low conductivities (< 0.1 S/cm) compared with the unsubstituted polyacetylene.

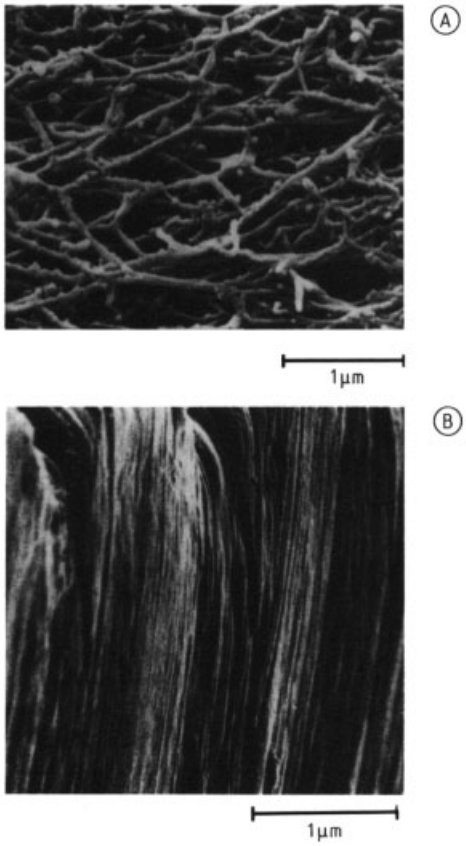

Figure 4. Scanning electron micrographs of iodine-doped polyacetylene
A) Standard, cross-linked polyacetylene (conductivity ca. 500 S/cm); B) Stretched polyacetylene (conductivity > 100 000 S/cm)

Modification of Polyacetylenes. The controlled modification of the polyene chain by radical reactions and Diels – Alder addition leads to functionalized systems with slightly improved atmospheric stability but diminished conjugation length and electrical conductivity. Examples include reaction with tosylmethyl isocyanide ($TosCH_2-N{=}C$) to give polyenes with pyrrole units [49], with nitrogen-substituted spin-labeled compounds to give spin labels (e.g., maleimido peroxyl groups) [50], with chlorosulfonyl isocyanate to give $-SO_3$ and $-COO^-$ groups [47], with tetrachloromethane to give $-Cl$ and $-CCl_3$ groups [27], and with 3-chloroperbenzoic acid to give epoxy groups [27]. Cycloaddition with 3,4-dichloromaleic anhydride leads to a fusible polyacetylene (165 – 180 °C) [47].

5.2. Polydiacetylenes

In contrast to polyacetylene, polydiacetylenes [*27987-87-7*] have limited electrical conductivity (≤ 0.1 S/cm), but can be obtained as large, single crystals.

The polymerization of diacetylene

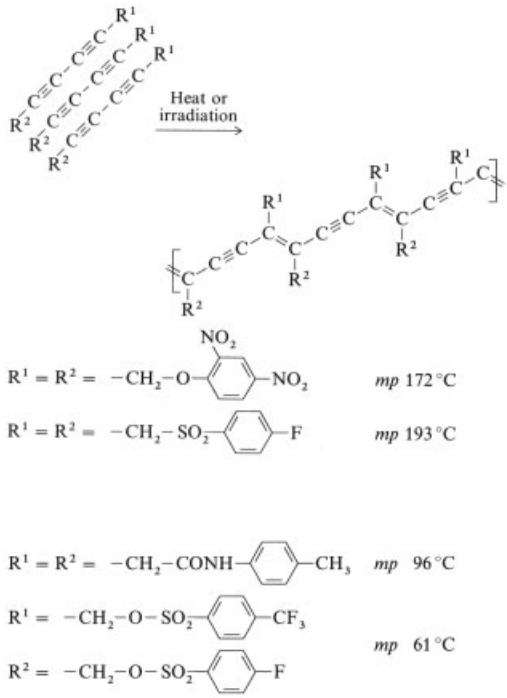

is an example of a topochemical polymerization in which 1,4-addition of 1,3-diyne units takes place in the crystalline state. The reaction does not require a catalyst and is performed by irradiation of the diacetylene crystals with visible or UV light, X-rays, or γ-rays, or by annealing the crystals below their melting point. The unreacted monomer is then extracted with a suitable solvent (e.g., hexane, toluene), leaving a single, dark red crystal of polydiacetylene. This unique polymerization process was already observed in 1882 [51] and has now been intensively studied [52]. A correct interpretation of the phenomena was first provided in the early 1970s [53].

5.3. Polypyrrole

Polypyrrole [*30604-81-0*] is a polysalt that can be produced in the form of powders, coatings, or films; it is intrinsically conductive.

Synthesis. The preparation of polypyrrole (pyrrole red and pyrrole black) by oxidation of pyrrole dates back to 1888 [54] and by electrochemical polymerization to 1957 [55]. A fairly long period elapsed before this organic π-system attracted general interest and was found to be electrically conductive [4].

Conductive polypyrrole films are obtained directly by anodic polymerization of pyrrole in aqueous or organic electrolytes (acetonitrile) [56]. They are black and, under suitable reaction conditions, can be detached from the anode in the form of self-supporting films (minimum thickness ca. 30 µm). Some of the conducting salt used in the electrolyte solution is incorporated in the film as a counterion.

In contrast to polyacetylene, polypyrrole has a high mechanical and chemical stability and can be produced continuously as flexible film (thickness 80 µm; trade name: Lutamer, BASF) by electrochemical techniques (Fig. 5) [57].

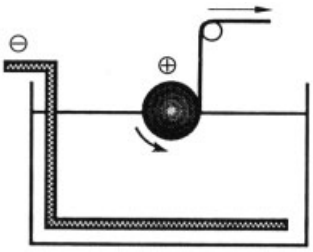

Figure 5. Continuous electrochemical process (BASF) for producing polypyrrole on a rotating drum electrode

Other electrochemically polymerizable heterocycles are thiophene, furan, and their substituted and oligomeric derivatives.

The polymerization reaction is very complex. The main reaction steps are shown in Figure 6 [58]. A radical cation is formed in the initial oxidation step. This is followed by a coupling reaction, deprotonation, and a one-electron oxidation which regenerates the radical cation system. Also important is the reaction introducing oxygen with formation of a labile −CONH group in the ring system.

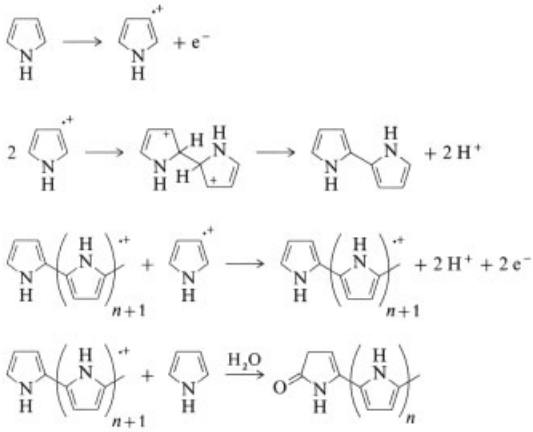

Figure 6. The oxidative coupling of pyrrole [58]

The quality of the polymers is greatly influenced by many factors, e.g., impurities, electrode material, pressure, concentrations, temperature, and comonomers. The most decisive, however, are the current density and the electrolyte, particularly the conducting anion X^- [59] because it is incorporated into the polymer as a counterion.

The properties of the counterion (e.g., its size, geometry, charge) influence the properties of the polymer. The amount of counterion (anion) incorporated depends on the reaction conditions. In general, one anion is incorporated for every three pyrrole units. Exceptions are pyrrole- or thiophenesulfonic acids where the counterion is coupled directly to the monomer (self doping) [60]. Some typical conducting anions are fluoroborate, perchlorate, aromatic sulfonic acids, penicillin, *n*-dodecyl sulfate [59,61], phthalocyanine sulfonic acid, poly(styrene sulfonic acid), camphor sulfonic acid, styrene sulfonic acid, and heparin [52].

By changing reaction conditions, polymers with different surface morphologies, (e.g., an open porous structure) can be obtained.

The anion X^- can also be released, e.g., by applying a negative potential. Release can be specifically controlled, offering interesting possibilities for active counterions of medical interest (e.g., heparin and monobactam) that are incorporated into polypyrrole.

Variation of the monomers and their substituents yields polymers with conductivities between 10^2 and 10^{-4} S/cm [59]. Alkyl substituents also increase the solubility of the poly-

mers with the result that electrically conducting polymers can be applied as coatings from solutions. This also applies to polymers derived from bridged pyrroles. Examples follow [62]:

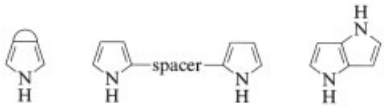

An interesting variant is the chemical oxidation of heterocycles (e.g., thiophene or pyrrole [63]) dissolved in an organic solvent (e.g., ethanol) on the surface of various materials. Conductive coatings (thickness 0.01 μm) can be produced on films of poly(phenylene sulfone), block copolymers of butadiene and styrene (Ultrason, Styrolux), poly(vinyl chloride), or other polymer films to give transparent, antistatic films with conductivities of about 0.001 S/cm. Ceramics and glass can also be coated in this way. Porous material (e.g., wood, fabrics, and open-celled foams) or fibers (e.g., polyamide, glass, or carbon fibers) can also be modified and rendered antistatic by this method.

Conductive powders (e.g., polypyrrole) with particle sizes of about 0.1 μm and conductivities of up to 10 S/cm can also be produced by chemical oxidation and can be incorporated as fillers in thermoplastics. These materials can be used for chip carriers.

Potential Uses. The conductivity of polypyrrole film suggests applications such as flexible conductive paths in printed circuits, heating films, and film keyboards.

Polypyrrole films show good electromagnetic shielding effects of about 40 dB over a wide range of frequencies (0 – 1500 MHz).

Although the new systems appear to be promising, stability problems may be encountered. High mechanical stability and/or stable conductivity are required under a variety of conditions (e.g., presence of air and moisture, temperature, and duration of exposure). Polypyrroles in which the counterion is 4-hydroxyphenyl sulfonic acid do not undergo any change in conductivity if they are exposed to nitrogen for two months at 140 °C. No change in conductivity was observed when these polymers were stored in the laboratory for three years at room temperature and 55 % R. H. Polypyrrole is sensitive to moisture because this leads to leaching of the counterion and thus to a decrease in conductivity. This can be avoided by use of appropriate hydrophobic or polymeric counterions [e.g., camphor sulfonic acid or poly(styrene sulfonic acid)] or by incorporating hydrophilic compounds. Polypyrroles obtained by synthesis in aqueous electrolytes maintain their conductivity at a level of about 20 S/cm. Polypyrroles with perchlorate as a counterion are unstable under atmospheric conditions but can be used as electrodes in rechargeable batteries [63].

The redox character of polypyrrole was first studied in 1975 [64]. It is a suitable electrode material for rechargeable electrochemical cells. The advantage of polymer electrodes is that they can be easily shaped, allowing novel battery designs (e.g., for the electronics sector) and less expensive production methods. Polymer cells with polypyrrole and lithium electrodes have been developed (Fig. 7) [63]. In the flat cell, the polypyrrole and lithium films are sandwiched together; in the cylindrical cell, the two films are wound concentrically. Their energy per unit mass and their discharge characteristics are similar to those of the nickel – cadmium cells now on the market. More than 500 charging and recharging cycles have been achieved with laboratory cells. Applications include dictaphones and pocket radios [63].

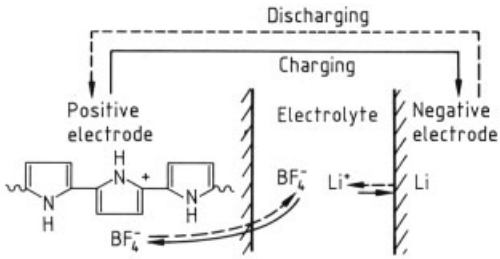

Figure 7. Electron reactions of a rechargeable polypyrrole – lithium battery
During discharge electrons flow from the negative electrode to the positive electrode and reduce the polypyrrole from the doped to the neutral state. During recharging an opposite potential is applied to the electrodes. The polymer takes up BF_4^- ions from the electrolyte and is oxidized. Lithium ions are also deposited at the other electrode.

5.4. Polythiophene

Since the first report of thiophene polymerization in 1883 [65] decades passed until thiophene was considered to be an attractive monomer for conducting polymers [66]. Interest is still growing because thiophene is easier to handle than pyrrole (less sensitive to oxygen) and allows the simple preparation of substituted monomers leading to soluble polymers. Three methods have been used to produce polythiophene [*25233-34-5*]: chemical oxidation (Lewis acid), coupling with organometallic agents or by the Grignard reaction, and electrochemical oxidation (anodic). The preparation techniques for films, powders, and coatings are similar to those of pyrroles but thiophene needs stronger oxidants for its polymerization than pyrrole or other heterocycles. Standard oxidation potentials follow:

Pyrrole	+0.8 V
Indole	+0.9 V
Azulene	+0.9 V
Thiophene	+1.6 V
Furan	+1.85 V

The oxidation potential of the monomers correlates to the ease of polymerization.

Figure 8 illustrates the correlation between UV – VIS absorption and the oxidation potential of thiophene oligomers [67]. The oxidation potential of the series $(\text{thiophene})_n$: $n = 1 - 6$ decreases with the increasing number of thiophene units.

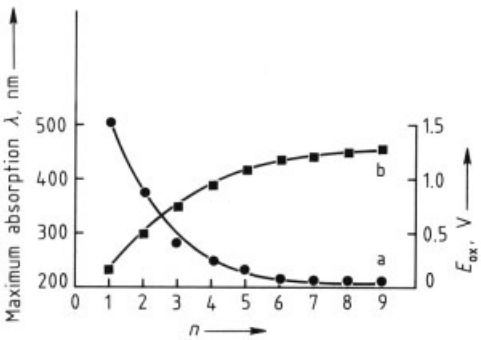

Figure 8. Correlation between maximum absorption wavelength (a) and oxidation potential (b) of thiophene oligomers Oxidation potential measured in 0.5 mol/L Et_4NBF_4 in acetonitrile versus saturated calomel electrode.

Substituted Polythiophenes. Alkyl-substituted thiophenes with alkyl chains up to C_{18} yield polymers that are soluble in common organic solvents with electrical conductivities up to 100 S/cm. Due to their good processibility [68] and high chemical and electrochemical stabilities in air and moisture, polythiophene and its derivatives are excellent candidates for practical applications (e.g., display devices, electrochromic devices [69], energy storage, field transistors [70], solar energy cells, and overchargeable batteries [66].

5.5. Polyphenylene

Synthesis. First attempts to prepare polyphenylene [*9033-83-4*], [*24991-24-0*], [*25190-62-9*], [*26499-97-8*] date back to 1842 [71]. Riese describes a process in which polyphenylene (n=13) is synthesized from 1,4-dibromobenzene and sodium. Further methods are the Ullmann reaction, thermal decomposition of diazonium salts, coupling of phenylene dihalogenide – Grignard compounds [72], and the oxidative coupling of benzene with $AlCl_3$ and $CuCl_2$ (Kovacic method [73]). In 1963 the Kovacic method was systematically extended by varying reaction conditions (temperature, catalysts, Lewis acid, and oxidants) and starting materials [4]. Many side reactions occur leading to cross-linked graphite-like lattices.

Other aromatic compounds, substitution products, condensed systems (e.g., diphenylene, terphenylene, pyrene, anthracene, chrysene), unsaturated

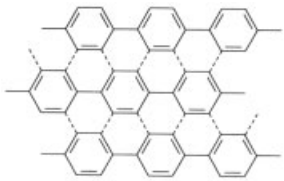

cyclic compounds (e.g., cyclopentadiene, cyclooctatetraene), heterocyclic compounds (e.g., thiophene, pyrrole, quinoline), and metal complexes (e.g., phthalocyanines, ferrocene) were investigated.

All Friedel – Crafts catalysts and the usual dehydrogenation agents can be used for oxidative coupling. Oxidizing Friedel – Crafts

agents, such as $FeCl_3$, assume both functions. The degree of suitability has been tested on benzene/$CuCl_2$ (Eq. 1) and on benzene/$AlCl_3$ (Eq. 2).

$$AlBr_3 > AlCl_3 > FeCl_3 > MoCl_6 > NbCl_5 > TiCl_4 > SnCl_4 > BF_3 > ZnCl_2 \quad (1)$$

$$Pd\,(II) > Co\,(III) > Mn\,(III) > Cu\,(II) > Fe\,(III) > \text{chloranil} > V\,(III) > I_2 \quad (2)$$

The oligomeric and polymeric reaction products were also characterized in terms of thermoelectric power, dark conductivity, and photoconductivity. They showed electrical conductivities up to 0.1 S/cm, the highest value obtained at that time [4]. The polymers were obtained as powders and were processed by press sintering. Later synthesis by Diels – Alder [74] or boronic acid coupling [75] lead to soluble or meltable materials.

A novel route to high molecular mass poly(*p*-phenylene) is based on the polymerization of derivatives of 5,6-dihydroxycyclo-1,3-hexadiene, a compound that is easily prepared by the bacterial oxidation of benzene [76].

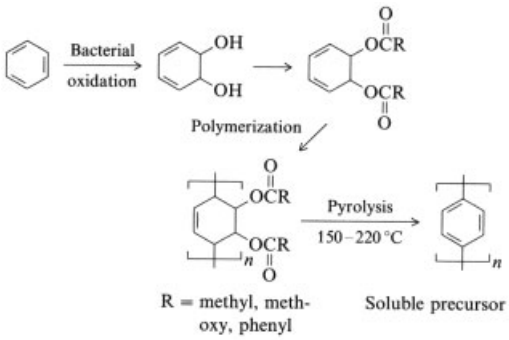

The advantage of this process is that it involves a soluble precursor polymer, which means that films can first be cast, objects coated, or fibers spun from solution; poly(*p*-phenylene) is then formed on heating.

The Marvel method is similar to the Kovacic method [77]:

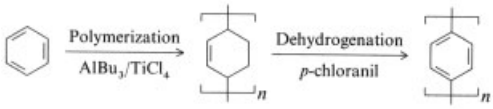

Properties. The properties of polyphenylenes depend mainly on their linkage (*m-, p-*), but also on the method used to prepare them. *o*-Polyphenylenes are unknown: *m*- and *p*-polyphenylenes are crystalline (> 50 %), intractable, and insoluble. *p*-Polyphenylenes are stable up to ca. 500 °C in air and up to 650 °C under nitrogen (10 % weight loss); the stability of *m*-polyphenylenes is approximately 200 °C lower [78]. Figure 9 demonstrates the drastic differences between *m*- and *p*-substituted polyphenylenes. The *para* derivatives are soluble in toluene up to $n = 6$, the *meta* derivatives are soluble up to $n = 16$. Polymer powders can be processed by press sintering (400 °C, > 100 bar).

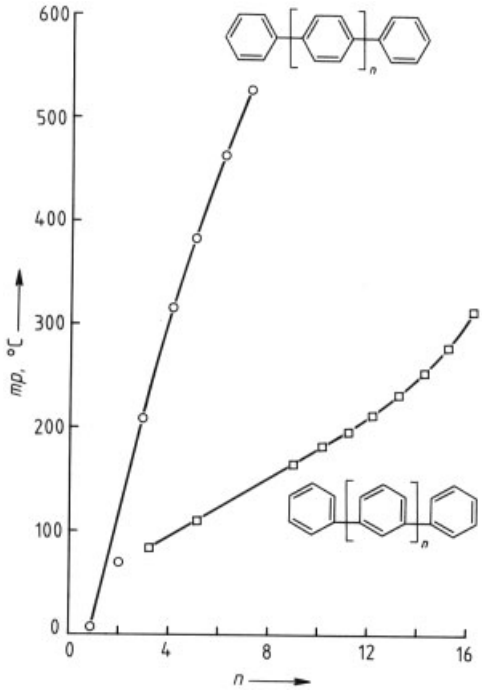

Figure 9. Correlation of the melting points and the degree of condensation (n) of *m*- and *p*-polyphenylenes

Poly(*p*-phenylene) can be *n*-doped with electron donors or *p*-doped with electron acceptors, electrical conductivities reach 10^2 S/cm. Donor doping is best accomplished with the aid of an electron-transfer agent (e.g., by adding naphthalene or α-methylstyrene in tetrahydrofuran) [79, 80]. Acceptor doping is usually done by exposing the powders to AsF_5 gas or heating with

nitromethane solutions of SbF_5. Electrochemical doping (anodic oxidation or cathodic reduction) is performed in suitable electrolytes (e.g., acetonitrile), the electrodes being separated by diaphragms.
Anodic oxidation:

$$p\text{-PP} + AgBF_4 \longrightarrow p\text{-PP}^+BF_4^- + Ag^+$$

Cathodic reduction:

$$p\text{-PP} + LiClO_4 \xrightarrow{+e^-} p\text{-PP}^-Li^+ + ClO_4^-$$

PP = polyphenylene
Oxidation potentials, conductivities, and spectra of polyphenylene are cited in [79]. Soluble, substituted polyphenylenes and polyaromatic systems are discussed in [81].

Potential Uses. Poly(*p*-phenylene) can be used as electrodes for rechargeable batteries.

5.6. Poly(Phenylene Sulfide)

Poly(*p*-phenylene sulfide) [*9016-75-5*], PPS, (white powder, T_g 92 °C, *mp* 270 – 290 °C, 65 % crystallinity) was the first melt-processible polymer to be doped with strong electron acceptors (e.g., AsF_5) to yield highly conductive products [82]. The first laboratory synthesis of PPS was reported by MACALLUM [83] and involved the melt reaction of 1,4-dichlorobenzene, sulfur, and sodium carbonate. A commercially product has been available as powder, film, or fiber since 1973 from Phillips Petroleum under the trade name Ryton; it is produced from 1,4-dichlorobenzene and sodium sulfide (high-pressure process) in a polar solvent (*N*-methylpyrrolidone) [84].

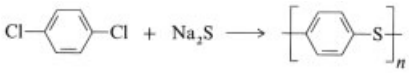

Another process involves self condensation of a halogenated thiophenol resulting in "head-to-tail" polymerization [85]:

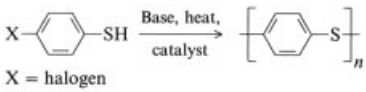

Poly(*m*-phenylene sulfide) and substituted (methyl-, fluoro-) poly(*p*-phenylene sulfide) are prepared by analogous reactions from appropriate monomers.

Doping of PPS with AsF_5 produces conductivities up to 2×10^{-2} S/cm [86]. Attempts to make *n*-doped PPS have been unsuccessful. In addition to AsF_5, doping with $FeCl_3$, H_2SO_4, $HClO_4$, FSO_3H, CF_3SO_3H, $AlCl_3$, TaF_5 has been investigated [87]. Heavily doped PPS undergoes intramolecular cross-linking to form benzothiophene rings [79]:

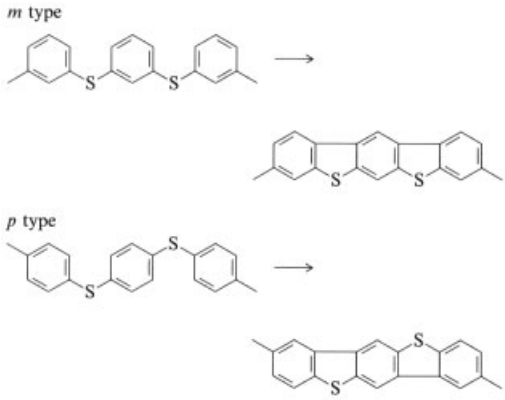

Poly(phenylene sulfides) are used where resistance to heat and chemicals is required, for example in the chemical process industry (pump housings, impellers, valves, metering devices, tanks, coil bobbins). See also, → Polymers, High-Temperature, Chap. 5.

Poly(*p*-phenylene selenide) [*52410-66-9*] and poly(*p*-phenylene telluride) [*84174-18-5*] have been prepared by techniques similar to those used to prepare PPS [79]. Conductivities of AsF_5-doped samples reach 0.01 S/cm.

5.7. Poly(Phenylene Vinylene)

Poly(*p*-phenylene vinylene) [88] [*96638-49-2*], [*87092-55-5*], [*26009-24-5*] was first synthesized by the Wittig condensation of terephthaldehyde and *p*-xylene bis(triphenylphosphonium) chloride. The method which is now mainly employed starts from a soluble polysulfonium salt [89], for example:

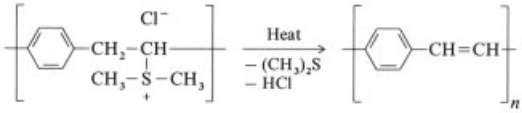

The polysulfonium salt is obtained by the reaction of α,α'-dichloro-*p*-xylene with excess dimethyl sulfide (50 °C, 20 h), and polymerization with sodium hydroxide (0 °C, 1 h). The first doped poly(phenylene vinylene) was reported by KARASZ using AsF_5 as dopant [90]; its conductivity was 3 S/cm. Stretching the precursor film at high temperature and doping with AsF_5 and SO_3 yielded conductivities up to 2800 S/cm [36,91].

Substituted derivatives [92], copolymers [93], and blends [94] show better thermal stability. Highly conductive graphite films have been prepared by pyrolysis of poly(phenylene vinylene) (> 3000 °C), stretched samples doped with SO_3 had conductivities of 10^5 S/cm [95]. Poly(phenylene vinylene) can be used as tunable polymer diodes or luminescent electrodes [96].

5.8. Polyaniline

The synthesis of polyaniline [*25233-30-1*] is remarkably simple and is summarized in Figure 10 [97,98]. Aniline can be chemically (1863, LIGHTFOOT) or electrochemically (1865, LETHEBY) oxidized and dehydrogenated to form a quinone diimine.

Oxidation of the primary aromatic amine usually does not stop at the monomeric quinone imines but gives rise to more complex quinone imine products. The type of product depends largely on the oxidants and the reaction conditions. 1,4-Benzoquinone-4-imino-1-phenylimine is an unstable intermediate which reacts with further aniline molecules and oligomerizes to give the final product, polyaniline.

Studies on polyaniline suggest that the polymer can exist in a wide range of structures [99], which can be regarded as copolymers of reduced (amine) and oxidized (imine) units of the form:

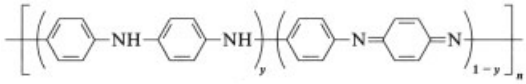

When $0 < y < 1$ these structures are poly(*p*-phenylene amine imines), in which the oxidation state of the polymer increases with increasing content of the imine form. The fully reduced form ($y = 1$) is leucoemeraldine, the fully oxidized form ($y = 0$) is pernigraniline, and the 50 %-oxidized structure ($y = 0.5$) is emeraldine. Each structure can exist as the base or as its salt, formed by protonation. Four repeating units can therefore be envisaged in the polymer chain in amounts which depend on the degree of oxidation and protonation of the structure. The " metallic state" is

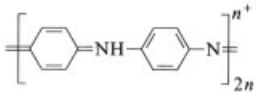

The reaction of aniline to "aniline black" proceeds through a number of intermediates [98]:

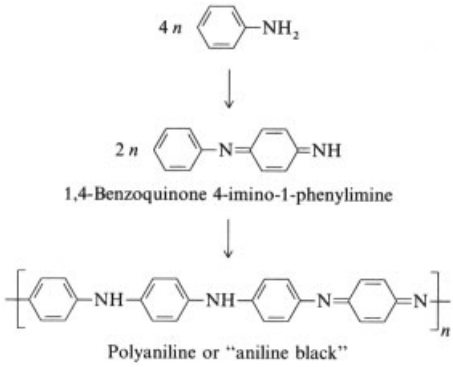

Emeraldine is very stable and can be directly prepared from aniline by oxidation with sodium perchlorate or V(III) chloride [100]. Nigraniline and pernigraniline are further oxidation products of emeraldine, both are very unstable. The structure of emeraldine and related molecules is described in [101].

Nitrogen- or ring-substituted homologues of aniline can apparently be oxidized similarly, yielding corresponding "aniline blacks".

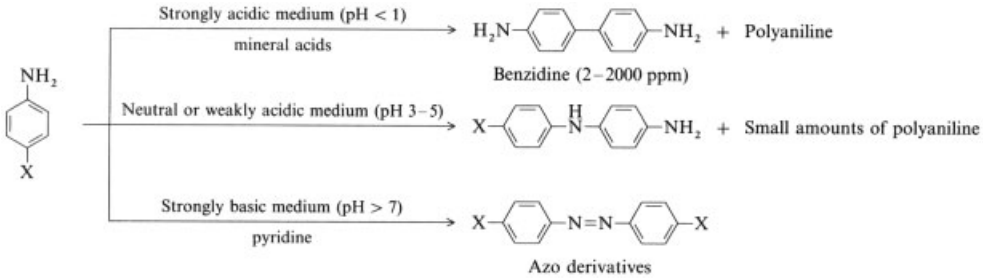

Figure 10. Products of aniline oxidation formed at different pH values

Oxidants used for preparing oligomers and polymers of aniline include $Cr_2O_7^{2-}$, $S_2O_8^{2-}$, Ce^{4+}, and Pb^{4+} ions. Aniline can also be polymerized by anodic oxidation.

The reaction conditions (temperature, duration, pH, oxidants, and catalysts) determine the yield and type of oxidation product [97].

A decisive handicap in the production of polyaniline is the formation of small quantities of carcinogenic benzidene (Fig. 10) [102].

Honzl synthesized benzidene-free polyanilines by avoiding the use of aniline as the starting monomer [103]:

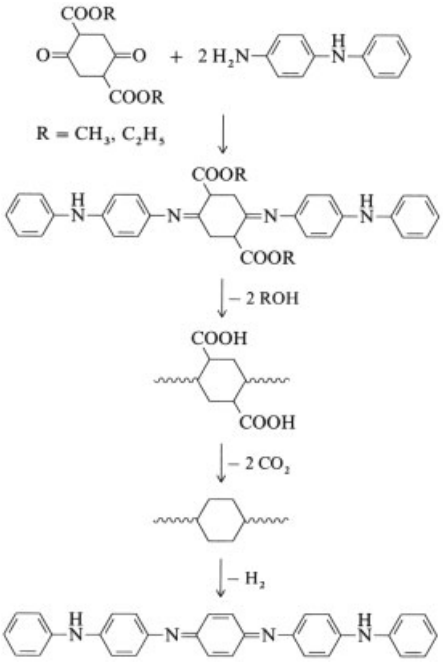

Similar structures can be obtained by polycondensation of quinones and aromatic amines [104].

Properties. The conductivity of polyaniline is 10 S/cm up to pH 4, but decreases to 10^{-10} S/cm above pH 4. The polymers are stable up to 250 °C (undoped) and up to ca. 150 °C (doped). Conduction in polyaniline can be accounted for in terms of electrons hopping under the assistance of proton transfer, for which the presence of water plays an essential role [105].

In contrast to other electrically conducting polymers, polyanilines may be doped with protons. Polyaniline is available as powder, film, or fibrils (Versicon, Allied Signal) [106, 107].

The electrochemistry of polyanilines was first investigated at the end of the nineteenth century. In the mid 1960s, interest in this topic was renewed. The use of polyaniline as a battery electrode on account of its redox and proton transfer behavior was described in 1968 [108]. In 1986, polyaniline was presented as a "novel conducting polymer"; its preparation and redox behavior were described [109]. The phenomena of anodic oxidation and discharge by means of reversed polarity have been known since 1891 [1].

Uses. Polyaniline is primarily of interest as an electrode material. Its discharge capacity is greater than that of polypyrrole(+)/perchlorate(−), its self-discharge is higher than that of polythiophene or Ni/Cd battery systems [66]. Genies has announced a rechargeable battery based on polyaniline (anode), propylene carbonate – $LiClO_4$ (electrolyte), and Li – Al (cathode) [110, 111]. This polyaniline is made by oxidizing aniline with ammonium persulfate in ammonia/2.3 mol/L hydrogen fluoride as solvent. The

discharge capacity of the polymer is 100 A · h/kg at 25 °C and 140 A · h/kg at 40 °C for current densities of 0.5 mA/cm^2 and for an amount of material giving a capacity of 10 mA · h. The voltage in open circuit for the fully charged battery developed by GENIES is 3.6 V. The average utilizable potential is 2.8 – 3 V. The energy density for the polymer lies between 280 and 420 W · h/kg. Its behavior with regard to self-discharge and to constant applied voltage (floating life) is excellent [111]. Like polypyrroles, polyanilines show excellent antistatic behavior and have a high shielding efficiency for electromagnetic interference.

5.9. Miscellaneous Polymers

There has been a flood of literature concerning new electrically conducting polymers, examples are listed below.

Bridged macrocyclic complexes [81] are mainly derivatives of tetraazaporphyrin or phthalocyanine:

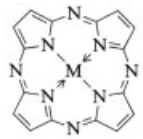

Tetraazaporphyrin (M = H_2)

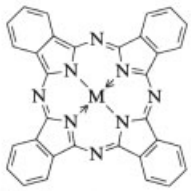

Phthalocyanine (M = H_2)

The synthesis and doping of bridged phthalocyanine polymers is illustrated in Figure 11. The bridging ligand can be linked by two σ bonds, by two coordinate bonds, or by one σ and one coordinate bond (Fig. 12).

After doping with iodine, conductivities up to 10.1 S/cm are obtained. Substituted phthalocyanines are soluble and can be cast as films or handled by the Langmuir – Blodgett technique to give ultrathin, well-defined molecular layers. The insoluble powders can only be processed by press sintering.

Polyazulene [*82451-56-7*] [112, 113] is synthesized by electrochemical polymerization with ClO_4^- as counterion (similar to that of polypyrrole) yielding amorphous polymer films which can be peeled from the anode. Conductivity is 0.01 S/cm. The films may be electrochemically and reversibly discharged to the nonconducting form.

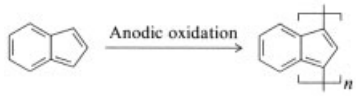

Polycarbazole [*51555-21-6*] [113]. Solutions of carbazole in acetonitrile may be electrochemically oxidized (counterion ClO_4^- or BF_4^-) at a platinum anode to give electrically conductive films with poor mechanical stability and conductivities up to 0.001 S/cm. Polycarbazole has also been obtained by vacuum evaporation of carbazole and by chemical condensation [114].

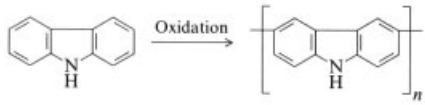

Polyindole [*82451-55-6*] [48]. Electrochemical oxidation of indole (counterion ClO_4^- or BF_4^-) yields brittle films with conductivities of ca. 0.01 S/cm. Polyindole has been employed as a polymer coating in a glucose sensor [113].

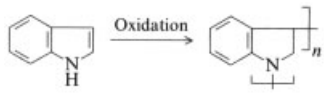

Polypyrene [*41496-25-7*] [113]. Electrochemical oxidation of pyrene (counterion BF_4^-, ClO_4^- or AsF_6^-) yields insoluble, brittle films with conductivities up to 1 S/cm.

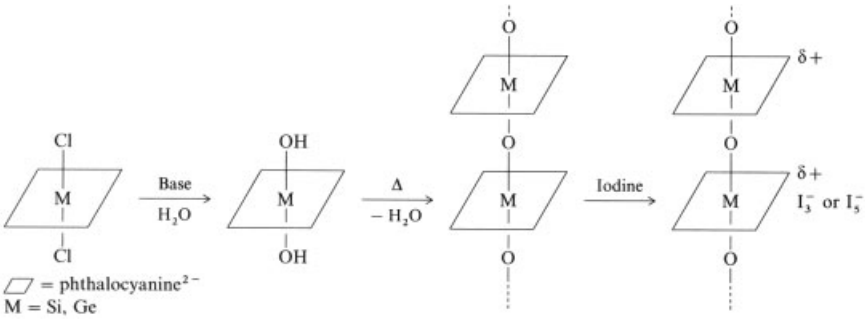

Figure 11. Synthesis and doping of bridged phthalocyanine polymers

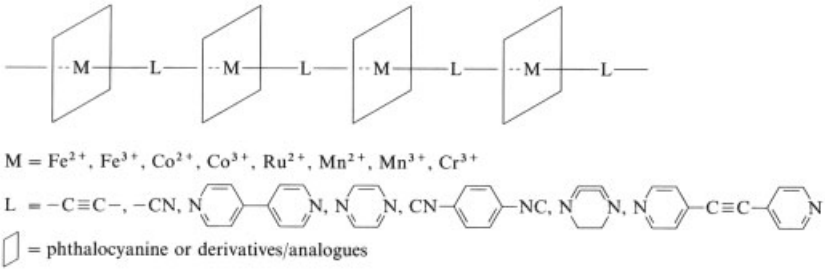

Figure 12. Construction of bridged macrocyclic transition-metal polymers

Polyazepine [*115483-32-4*] [115]. Phenyl azide can be photopolymerized in the gas phase to yield films of poly-1,2-azepine with a conductivity of 10^{-2} S/cm (after I_2 doping).

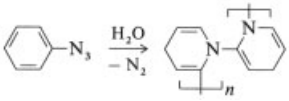

Polyfulvenes [*107889-59-8*] [116]. Polyfulvenes are formed by cationic polymerization of 6,6-dimethylfulvene followed by chemical or electrochemical dehydrogenation, conductivity 10^2 S/cm (after I_2 doping).

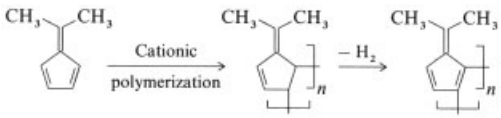

Low-band-gap aromatic polymers [117]. An important prerequisite for obtaining low-band-gap polymers is that the ground state of the polymer has a quinonoid contribution. These polymers have the general formula:

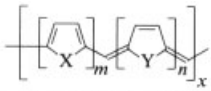

X, Y = SO_2, SO, S, N−R, CH=CH

Benzenoid and anthracenoid precursors are:

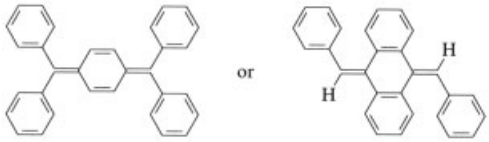

Low- or narrow-band-gap systems (gap < 0.5 eV) already contain a high proportion of thermally excited electrons at room temperature. These materials can be oxidized electrochemically at a low oxidation potential to give conducting polymers. Cross coupling with organometallic intermediates affords the undoped polymers. Conductivities up to 10 S/cm are reached (after iodine or electrochemical doping).

Polyindophenines [*138309-48-5*] [118] are prepared by reaction of isatin with thiophene and have conductivities up to 10^{-2} S/cm without additional dopant. The polymers are soluble in dimethyl sulfoxide and can be cast as films; the material is thermostable up to 230 °C.

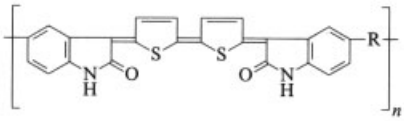

Charge-transfer complexes [119] are combinations of electron donor compounds with electron acceptor compounds. They are synthesized, for example, by electrochemical charging of aromatic hydrocarbons with ClO_4^-. The charge-transfer complexes are assembled in defined stacks and have conductivities up to 10^3 S/cm, in some cases with superconductivity. Examples are given below [120, 121]:

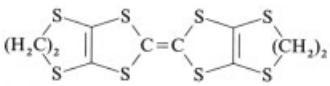

Bisethylene dithiotetrathiofulvalene – I_3 complex
T_c 8.1 K

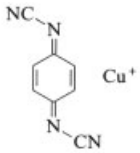

Dicyanodiimine quinone–copper complex
$\sigma = 10^3$ S/cm

Finally it should be remembered that all electrically conducting organic materials are due to the various dopant charge-transfer complexes [123].

Buckminsterfullerene [*99685-96-8*] [122] is an allotrope of carbon (Fig. 13). When doped with potassium, it reaches conductivities up to 500 S/cm; cooling to 18 K makes it superconducting.

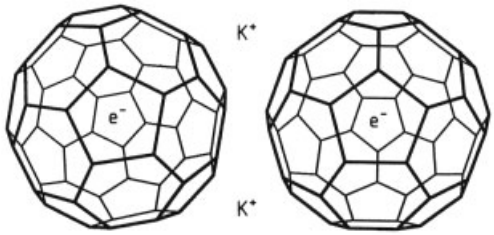

Figure 13. Potassium-doped C_{60} molecules

6. Uses

Few of the organic materials discussed above have been demonstrated to be sufficiently stable for critical applications.

Stability is a relative property and is the "Achilles heel" of electrically conducting polymers. For example, polypyrroles with BF_4^- as counterion are unstable under atmospheric conditions but are as stable as Ni/Cd electrodes in rechargeable batteries. It was hoped that electrically conducting organic materials could combine conductivities comparable to those of metals with the processibility of plastics. Work over the last decades has gone a long way in this direction but there are still many obstacles to be mastered before these materials become widely used.

Polyene chains with extended double bond systems are unstable against oxygen, forming epoxy, carbonyl, and carboxyl groups. They also undergo cross-linking and substitution reactions (Diels – Alder and radical reactions). Aromatic and heterocyclic structures are much more stable against environmental conditions (e.g., temperature, irradiation).

Doped materials (*n* or *p* doped) are saltlike compounds and are therefore unstable against humidity, oxygen, and temperature. Conductivity is lost as a result of interruption of the conju-

gation system, this destroys the charge-transfer complexes within a few days.

Defect-free oriented polyacetylenes are more stable, they do not loose any conductivity for 3 months [24]. Polypyrroles and polythiophenes show unchanged conductivities over several years (Sections 5.3, 5.4). Polyindophenines are "intramolecularly doped", containing covalently bonded electron donor and acceptor groups within the same molecule (Section 5.9).

How many uses do conducting polymers have? The short answer is "scarcely any" [123], but they do have some applicability [80, 124, 125] for:

1) Rechargeable batteries [120, 126] (Sections 5.3 and 5.4)
2) Coatings for electromagnetic radiation shielding or electrochemical metallization [63, 127]
3) Semiconducting devices and photovoltaics [128]
4) Catalysis [129]
5) Sensors and electrochromic displays [130], e.g., for gas detection (NH_3, CO_2)
6) Solar energy conversion, e.g., electroluminescence [131]

It seems that this class of electrically conducting materials will form the basis for a new generation of "intelligent" material systems [132].

7. References

1. F. Goppelsroeder, *Die Internationale Electrotechnische Ausstellung* **18** (1891) 978; **19** (1891) 1047.
2. H. A. Pohl: "Electronic Behavior of Organic Macromolecular Solids," B. A. Bolto: "Semiconducting Organic Polymers Containing Metal Groups," D. D. Eley: "Semiconducting Biological Polymers," in J. E. Katon (ed.): *Organic Semiconducting Polymers,* Marcel Dekker, New York 1968.
3. W. A. Little, *Phys. Rev.* **134** (1964) A 1416.
4. H. Naarmann, F. Beck: "Neuartige Polymerisate aus aromatischen und heterocyclischen Verbindungen und ihre elektrophysikalischen Eigenschaften," GDCh Meeting, Munich, Oct. 12, 1964. BASF, DE 1 178 529, 1964 (H. Naarmann, F. Beck, E. G. Kastning). BASF, DE 1 195 497, 1963 (H. Naarmann, F. Beck, E. G. Kastning). BASF, DE 1 092 137, 1964 (H. Naarmann, F. Beck, E. G. Kastning).
5. W. Slough et al. in [2] pp. 55, 118 – 163. H. Naarmann, *Angew. Chem. Int. Ed. Engl.* **8** (1969) 915. BASF, DE-OS 1 953 898, 1969 (H. Willersinn, H. Naarmann, K. Schneider).
6. R. Greene, G. B. Street, L. J. Süter, *Phys. Rev. Lett.* **34** (1975) 577.
7. A. J. Heeger et al., *J. Chem. Soc. Chem. Commun.* 1977, 578 – 580.
8. H. Shirakawa, S. Ikeda, *Polym. J. (Tokyo)* **2** (1971). T. Ito, H. Shirakawa, S. Ikeda, *J. Polym. Sci. Polym. Chem. Ed.* **12,** (1974) 11.
9. H. Naarmann: "Synthese elektrisch leitfähiger Polymere," *Angew. Makromol. Chem.* **109/110** (1982) 295 – 338. W. J. Feast: "Synthesis of Conducting Polymers," in T. A. Skotheim (ed.) : Handbook of Conducting Polymers. Marcel Dekker, New York 1986, pp. 1 – 44. N. C. Billingham, P. D. Calvert: "Electrically Conducting Polymers," *Adv. Polym. Sci.* **80** (1989) 1 – 90.
10. H. W. Gibson: "Substituted Polyacetylenes," in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* Marcel Dekker, New York 1986, pp. 405 –435.
11. H. Hopf, O. Kretschmer, H. Naarmann, *Adv. Mater.* **1** (1989) no. 12, 445.
12. H. Naarmann, *Adv. Mater.* **2** (1990) no. 8, 345 – 348. W. R. Sorenson, T. W. Campbell: *Präparative Methoden der Polymeren Chemie,* Verlag Chemie, Weinheim 1962, p. 167.
13. T. Yamanato, Y. Hayashi, A. Yamamoto, *Bull. Chem. Soc. Jpn.* **51** (1978) 2091.
14. N. Miyaura, T. Yagani, A. Suzuki, *Synth. Commun.* **11** (1981) 513 – 519.
15. K. J. Stille, *Makromol. Chem.* **154** (1972) 49.
16. M. Kanabe, M. Okawara, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 1058.
17. H. H. Edwards, W. J. Feast, *Polymer* **21** (1980) 595.
18. T. M. Swager, D. A. Dougherty, R. H. Grubbs, *J. Am. Chem. Soc.* **111** (1989) 4413.
19. E. J. Ginsburg et al.: "Conjugated Polymeric Materials," in J. L. Bredes, R. R. Chance (eds.): *NATO ASI Ser.* **182** (1990) 65.
20. A. W. Snow, *Nature (London)* **292** (1981) no. 40.
21. K. Menke, S. Roth, *Chem. Unserer Zeit* **20** (1986) no. 2, 33, 53.
22. R. R. Chance, D. S. Boudreaux, J. L. Bredas: "Solitons, Polarons, Bipolarons" in T. A. S. Skotheim (ed.): *Handbook of Conducting Polymers,* vol. 2, Marcel Dekker, New York

1986, pp. 825 – 857. P. Bätz, Dissertation, Universität Tübingen, Tübingen 1991.
23. H. Schäfer, Dissertation, Universität Göttingen, Göttingen 1983.
24. S. Pekker, A. Janossy in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* **vol. 1,** M. Dekker, New York 1986, p. 45.
25. G. Lugli, U. Pedretti, G. Perego, *J. Polym. Sci. Polym. Lett. Ed.* **23** (1985) 129.
26. BASF, BMFT-Forschungsbericht: Entwicklung von elektrisch leitfähigen Alternativ-Polymeren, 03 C 134-0, chap. II, Ludwigshafen, Aug. 1, 1982 – July 31, 1985, pp. 1 – 75.
27. H. Naarmann, *Synth. Met.* **17** (1987) 223. H. Naarmann, N. Theophilou, *Synth. Met.* **22** (1987) 1. T. Schimmel, M. Schwoerer, *Solid State Commun.* **65** (1988) no. 11, 1311. J. Tsukamoto, A. Takahashi, K. Kawasaki, *Jpn. J. Appl. Phys. Part* **29/1** (1990) 125.
28. M. Yamaura et al., *Synth. Met.* **28** (1989) C 157 –164.
29. D. Gagnon, J. Capistran, F. Karasz, R. Lenz, *Polym. Prepr. (Am. Chem. Soc. Dir. Polym. Chem.)* **25** (1984) 284 – 285. D. Gangon, F. Karasz, E. Thomas, R. Lenz, *Synth. Met.* **20** (1987) 85 – 95.
30. H. Shirakawa, Y.-C. Chen, K. Akagi, T. Norahara, *Synth. Met.* **14** (1986) 173, 199. M. Aldissi, *J. Polym. Sci. Polym. Lett. Ed.* **23** (1985) 167. H. Shirakawa, K. Akagi, S. Katayama, *J. Macromol. Sci. Chem.* **A25** (1988) no. 5 – 7, 643. H. Shirakawa, K. Akagi, M. Suezaki, *Synth. Met.* **28** (1989) D 1. A. Moutaner et al., *Synth. Met.* **28** (1989) D 19.
31. BASF, DE-OS 3 533 252 Al, 1985 (H. Naarmann, M. Portugall).
32. H. Naarmann, *Synth. Met.* **41 – 43** (1991) 1 – 6.
33. A. G. MacDiarmid, T. Woerner, A. G. Heeger, A. Feldblum, *J. Polym. Sci., Polym. Lett. Ed.* **20** (1982) 305; **22** (1984) 119.
34. T. Shimidzu, *Langmuir* **3** (1987) 1169.
35. H. Nakahara, *Thin Solid Films* **60** (1988) 87, 153. J. Watanabe, K. Hong, M. F. Rubner, *Synth. Met.* **28** (1989) C 473.
36. M. Matsumoto et al., *Synth. Met.* **27** (1988) B 601.
37. I. Bozovic, *Mod. Phys. Lett. B.* **1/3** (1987) 81.
38. H. Wiu, *Jahresber. Fortschr. Chem.* (1866) 516.
39. H. Erdmann, P. Köthner, *Z. Anorg. Allgem. Chem.* **18** (1898) 48.
40. W. Reppe, *Justus Liebigs Ann. Chem.* **560** (1948) no. 1, 140.
41. M. Hatano, S. Kambara, S. Okamoto, *J. Polym. Sci.* **61** (1961) 26.
42. *Houben-Weyl,* 4th ed., **20/2** 1212.
43. H. Naarmann, *Angew. Makromol. Chem.* **109/110** (1982) 295.
44. H. Naarmann, N. Theophilou, *Synth. Met.* **22** (1988) 1.
45. T. Schimmel et al., *Bayreuther Polym. Symp.,* April 12, 1989; *Synth. Met.* **37** (1990) 1 – 6; *Makromol. Chem.* **190** (1989) 3217.
46. H. Haberkorn et al., *Synth. Met.* **5** (1982) 51.
47. H. Naarmann in J. L. Bredas, R. R. Chance (eds.): *Conjugated Polymeric Materials,* Klüver Acad. Publ., Dordrecht 1990, pp. 11 – 51.
48. H. Naarmann, P. Strohriegel: "Conducting and Photoconducting Polymers," in H. Kricheldorf (ed.): *Handbook of Polymer Synthesis,* part 13, Marcel Dekker, New York 1992, pp. 1353 – 1425.
49. BASF, DE-OS 3 118 630, 1981 (H. Naarmann, G. Köhler).
50. H. Winter, G. Sachs, E. Dormann, H. Naarmann, *Synth. Met.* **36** (1990) 353.
51. A. Bayer, L. Landsberg, *Ber. Dtsch. Chem. Ges.* **15** (1882) 52. F. Bohlmann, *Angew. Chem.* **69** (1957) 82.
52. G. Wegner in E. W. Halfield (ed.): *Molecular Metals,* chap. 4.0, Plenum Publ., New York 1979, p. 209.
53. E. Hädicke et al., *Angew. Chem.* **83** (1971) no. 7, 253.
54. M. Dennstedt, J. Zimmermann, *Ber. Dtsch. Chem. Ges.* **21** (1888) 1478. A. Grossauer: *Die Chemie der Pyrrole,* Springer-Verlag, Berlin 1974, p. 149.
55. H. Lund, *Acta Chem. Scand.* **11** (1957) 1323. A. Stanienda, *Z. Naturforsch.* **228** (1967) 1107.
56. K. K. Kanazawa et al., *J. Chem. Soc. Chem. Commun.* 1979, 854. A. F. Diaz, J. Bargon: "Electrochemical Synthesis of Conducting Polymers," in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* vol. 1, Marcel Dekker, New York 1986, p. 82. G. B. Street: "From Powder to Plastics," in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* vol. 1, Marcel Dekker, New York 1986, p. 266.
57. BASF, US 4 468 291, 1982 (H. Naarmann, G. Köhler, J. Schlag).
58. E. M. Genies, G. Bidan, A. F. Diaz, *J. Electrochem. Soc.* **149** (1983) 101; **129** (1982) 1685.
59. H. Naarmann, *Macromol. Chem. Macromol. Symp.* **8** (1987) 1.

60. BASF, DE-OS 3 425 511, 1968 (H. Naarmann, G. Köhler). A. O. Patil, Y. Ikenoue, F. Wudl, A. Heeger, *J. Am. Chem. Soc.* **109** (1987) 1858.
61. BASF, EP 129 070, 1984 (G. Wegner, W. Wernet).
62. M.-A. Sato, S. Tanaka, K. Kaeriyama, *Synth. Met.* **18** (1987) 229. B. F. Goodrich, US 4 487 667, 1984 (L. Traynor). G. Schiavon, G. Zotti, *Synth. Met.* **28** (1989) C 199. N. Oyama, T. Ohsaka, H. Miyamoto, *Synth. Met.* **28** (1989) C 193. T. Inagaki et al., *Synth. Met.* **18** (1989) C 245. R. Bittihn, *Kunststoffe* **79** (1989) no. 6, 530
63. BASF, EP 0 206 133 Al, 1986 (H. Naarmann, W. Heckmann). BASF Plastics, Research and Development KVX 8611, Ludwigshafen, Oct. 1986. R. Bittihn, *Makromol. Chem. Symp.* **8** (1987) 51.
64. Agence National de Valorisation, FR 7 518 383, 1975 (R. Buvet, R. Vallot, I. Gal, L. T. Yu).
65. V. Meyer, *Ber. Dtsch. Chem. Ges.* **16** (1883) 1465.
66. G. Tourillon: "Polythiophene and its Derivatives," in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* vol. I, Marcel Dekker, New York 1986, p. 293.
67. F. Martinez, R. Voelkel, D. Naegele, H. Naarmann, *Mol. Cryst. Liq. Cryst.* **167** (1989) 227.
68. G. Gustafsson et al. in J. L. Bredas, R. Silbey (eds): *Conjugated Polymers,* Klüver Acad. Publ., Dordrecht 1991, p. 315.
69. M. Gazard: "Electrochromic Display Device," in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* Marcel Dekker, New York 1986, p. 673.
70. F. Garnier, G. Horowitz, X. Peng, D. Fichou, *Adv. Mater.* **2** (1990) no. 12, 592.
71. F. Riese, *Justus Liebigs Ann. Chem.* **164** (1872) 161.
72. H. Naarmann, *Angew. Makromol. Chem.* **109/110** (1982) 295.
73. P. Kovacic, A. Kyriakis, *Tetrahedron Lett.* **11** (1962) 467.
74. J. K. Stille, *Makromol. Chem.* **154** (1972) 49.
75. M. Miyaura, T. Yanagi, A. Suzuki, *Synth. Commun.* **11** (1981) 513.
76. D. G. H. Ballard, A. Courtis, I. M. Shirley, S. C. Taylor, *J. Chem. Soc. Chem. Commun.* 1983, 954.
77. C. S. Marvel, *J. Am. Chem. Soc.* **81** (1959) 448.
78. P. E. Cassidy: *Thermally Stable Polymers,* M. Dekker, New York 1980, p. 34. M. Busch, W. Weber, *J. Prakt. Chem.* **146** (1936) 1 –55.
79. R. L. Elsenbaumer, L. W. Schacklette in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* Marcel Dekker, New York 1986, p. 214.
80. R. H. Baughman, L. W. Shacklette in W. R. Salaneck, D. T. Clark, E. J. Samuelsen (eds): *Science and Applications of Conducting Polymers,* A. Hilger, New York 1990, p. 47.
81. H. Naarmann et al., Entwicklung neuer Polymerer mit konjugierten Bindungssystemen und definierten zweidimensionalen Strukturen, BMFT Report 03M4019, part 1, Feb. 1991.
82. L. W. Shacklette, R. L. Elsenbaumer, R. Chance, H. Eckhardt, *J. Chem. Phys.* **75** (1981) 1919.
83. A. Macallum, *J. Org. Chem.* **13** (1948) 154.
84. Phillips Petroleum, US 3 607 843, 1971 (S. Edmonds, H. Hill, jr.).
85. R. W. Lenz, C. E. Handlorits, H. A. Smith, *J. Polym. Sci.* **58** (1962) 351.
86. J. Rabolt et al., *J. Chem. Soc. Chem. Commun.* 1980, 347.
87. J. Tsukamoto, K. Matsumura, *Jpn. J. Appl. Phys.* **23** (1984) 584.
88. R. MacDonald, T. Campbell, *J. Am. Chem. Soc.* **82** (1960) 4669.
89. M. Kanabe, M. Okawava, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 1058; *Chem. Abstr.* **68** (1968) 105721t. R. Wessling, H. Zimmermann, US 3 706 677, 1972; *Chem. Abstr.* **78** (1973) 85306 n.
90. G. Wnek, J. Chien, F. Karasz, C. Lillya, *Polymer* **20** (1979) 1441.
91. J. Murase et al., *J. Mol. Cryst. Liq. Cryst.* **118** (1985) 335.
92. H. H. Hörhold et al., *Materials Sciences Forum,* **vols. 62 – 64,** (1990) 411 – 417. J. Murase, T. Onishi, T. Nogushi, M. Hirooku, *Synth. Met.* **17** (1987) 639.
93. M. Lux et al., *Proc. Macromol. Colloq.,* Freiburg, FRG 1989, p. 44.
94. J. Schlenhoff, J. Machado, P. Gletkowki, F. Karasz, *J. Polym. Sci. Polym. Phys. Ed.* **26** (1988) 2247.
95. F. Ohnishi, J. Murase, N. Takanobu, M. Hirooka, *Synth. Met.* **14** (1986) 207.
96. R. H. Friend et al., *Nature (London)* **356** (1992) 47.
97. *Houben Weyl,* 4th ed., **VII/3b,** 242.
98. R. Willstätter, C. W. Moore, *Ber. Dtsch. Chem. Ges.* **40** (1907) 2665.
99. A. G. Diarmid et al., *Synth. Met.* **18** (1987) 285.

100. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. II,** Academic Press, New York 1952, p. 776.
101. C. B. Duke et al., *Synth. Met.* **21** (1987) 143.
102. E. Genies, *New J. Chem.* **12** (1988) no. 4, 184.
103. J. Honzl, M. Metalova, *Tetrahedron* **25** (1969) 3641.
104. A. Everaets, S. Roberts, H. K. Hall, *J. Polym. Sci. Polym. Chem. Ed.* **24** (1968) 1703.
105. J. P. Travers, M. Nechtschein, *Synth. Met.* **21** (1987) 135.
106. M. Kryszewski: *Semiconducting Polymers PWN,* Polish Scientific Publ., Warzawa 1980, p. 76.
107. A. J. Heeger (pp. 1 – 12, 105 – 115), E. Genies (pp. 93 – 104), A. G. MacDiarmid, A. J. Epstein (pp. 117 –127) in W. R. Salaneck, D. T. Clark, E. J. Samuelsen (eds.): *Science and Application of Conducting Polymers,* Adam Hilger, New York 1990.
108. R. Buret et al., *Electrochim. Acta* **13** (1968) no. 2, 1441, 1451.
109. W. S. Huang, B. D. Humphrey, A. G. MacDiarmid, *J. Chem. Soc. Faraday Trans. 1* **82** (1986) 2385.
110. N. C. Billingham, P. O. Calvert: "Electrically Conducting Polymers," *Adv. Polym. Sci.* **80** (1989) 85.
111. E. Genies, P. Hany, C. Santier, *J. Appl. Electrochem.* **18** (1988) 751; *Synth. Met.* **28** (1989) C 647.
112. J. Bargon, S. Mohmand, R. J. Waltman, *Mol. Cryst. Liq. Cryst.* **93** (1983) 279.
113. P. C. Pandey, *J. Chem. Soc. Faraday Trans.* **1** (1988) no. 84, 2259.
114. S. A. Jenekhe, S. Reed, *J. Mol. Cryst. Liq. Cryst.* **105** (1984) 175; **106** (1984) 289.
115. A. W. Meijer, S. Nijhius, F. Van Vroonhoven, E. Havinga in L. J. Bredes, R. R. Chance (eds.): *Conjugated Polymeric Materials,* **vol. 182,** Klüver Acad. Publ., Dordrecht 1989, p. 115.
116. H. Naarmann, *Synth. Met.* **17** (1987) 223.
117. M. Hanack, G. Hieber, G. Dewald, H. Ritter in W. R. Salaneck, D. T. Clark, E. J. Samuelsen (eds.): *Science and Application of Conducting Polymers,* Adam Hilger, New York 1990, p. 153.
118. H. Naarmann in : Entwicklung neuer Polymerer mit konjugierten Doppelbindungssystemen und definierten zweidimensionalen Strukturen, BMFT Report 03M 40197, Feb. 1990, p. 8. G. Koßmehl in T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* M. Oelsler, New York 1986, vol. 1, p. 394.
119. E. P. Goddings, *Endeavour* **34** (1976) 125; *Synth. Met.* **27** (1988) 1 – 4.
120. S. Hünig et al., *Z. Naturforsch. A Phys. Phys. Chem. Kosmophys.* **44 A** (1989) 825; *Synth. Met.* **27** (1988) B 181.
121. K. Bechard, D. Jerome, *Spektrum Wissensch.* **9** (1982) 38. D. Jerome, *Synth. Met.* **21** (1988) A183.
122. A. W. Sleight: "Sooty Superconductors," *Nature (London)* **350** (1991) 557.
123. N. C. Billingham, P. D. Calvert, *Adv. Polym. Sci.* **90** (1989) 4.
124. T. A. Skotheim (ed.): *Handbook of Conducting Polymers,* vol. 1, Marcel Dekker, New York 1986.
125. J. R. Ellis: "Commercial Applications of Intrinsically Conducting Polymers," in [124]p. 489.
126. A. G. MacDiarmid, R. B. Kaner: "Electrochemistry of Polyacetylene, Application to Rechargeable Batteries," in [124] p. 689.
127. H. J. Mair, S. Roth (eds.): *Elektrisch leitende Kunststoffe,* Hanser Verlag, München 1989, pp. 59, 89, 105, 137, 149, 237.
128. J. Kanicki: "Polymeric Semiconductor Contacts and Photovoltaic Applications," in [124] p. 543.
129. K. Soga, S. Ikeda: "Catalytic Properties of Conducting Polymers," in [124] p. 661.
130. G. Weddingen, H. Suhr in [127] p. 545.
131. *Nobel Symp.* NS 81, June 12 – 18, 1991, Lulea, Sweden.
132. D. T. Clark in [107] p. 161.

Polymers, High-Temperature

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey ; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing . Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

David Parker, ICI Advanced Materials, Wilton, Middlesborough, United Kingdom (Chaps. 1, 3, and 4)

Jan Bussink, GE Plastics Europe, Bergen op Zoom, The Netherlands (Chap. 2)

Hendrik T. van de Grampel, GE Plastics Europe, Bergen op Zoom, The Netherlands (Chap. 2)

Gary W. Wheatley, Humberside Polytechnic, Hull, United Kingdom (Chaps. 3 and 4)

Ernst-Ulrich Dorf, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

Edgar Ostlinning, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

Klaus Reinking, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 5)

1. Introduction . . . 2
2. Polyetherimides . . . 3
2.1. Production . . . 3
2.2. Properties . . . 5
2.3. Processing . . . 10
2.4. Uses . . . 10
2.5. Economic Aspects . . . 11
3. Polyaryletherketones . . . 11
3.1. Production . . . 12
3.2. Properties . . . 13
3.3. Processing . . . 14
3.4. Uses . . . 14
4. Polysulfones . . . 14
4.1. Production . . . 15
4.2. Properties . . . 16
4.3. Processing . . . 17
4.4. Uses . . . 17
5. Poly(Phenylene Sulfide) . . . 17
5.1. Production . . . 18
5.2. Properties . . . 19
5.2.1. Material Properties . . . 19
5.2.2. Properties of Nonreinforced Moldings 19
5.2.3. Properties of Reinforced Moldings . . 20
5.3. Processing . . . 22
5.3.1. Injection Molding . . . 22
5.3.2. Other Molding Methods . . . 23
5.3.3. Finishing of Shaped Articles . . . 23
5.3.4. Extrusion . . . 23
5.3.5. Semifinished Products . . . 24
5.4. Uses . . . 24
5.5. Economic Aspects . . . 25
6. References . . . 25

1. Introduction

Definition. Any definition of a *high-temperature polymer* is necessarily arbitrary. Today's conventional engineering plastics [e.g., nylon, poly(ethylene terephthalate)] were yesterday's high-temperature polymers, and the quest for improved temperature performance is ongoing. In the context of this review, thermoplastic polymers exhibiting useful mechanical properties at temperatures in excess of 150 °C would seem an appropriate description.

Various other synonyms such as *high-performance polymers* or *advanced engineering polymers* have been employed. These materials are characterized by relatively low-volume sales compared with conventional engineering polymers and a relatively high selling price. Thus the ratio of the sales price for nylon to the sales price for the high-temperature polymers considered here varies from about 1 : 3 to 1 : 20. However, these polymers enjoy applications in varying market sectors such as the electrical–electronic, automotive, aerospace, and chemical process industries because they offer good performance to cost ratio.

The fact that the polymers can be used at high temperatures should not disguise that they possess many other useful and interesting proper-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_449

ties. The crystalline polymers, polyetheretherketone and poly(phenylene sulfide), for example,find many room-temperature applications as a result of their good environmental resistance, particularly to solvents.

Classification. As with all polymers, high-temperature polymers fall into two classes, amorphous and (semi)crystalline. Polysulfone (PSU), poly(ethersulfone) (PES), and polyetherimide (PEI) are amorphous whereas poly(phenylene sulfide) (PPS), polyetheretherketone (PEEK), and polyetherketone (PEK) are semicrystalline.

Despite the high linearity of the sulfone polymers (100 % 4,4′-specificity), the differences in the conformations of the phenyls bordering the sulfone and ether links preclude close packing of chains whereas the regular conformations of the phenyls in chains of PPS, PEEK, and PEK lend themselves to crystal formation. Interestingly, the ether and carbonyl moieties in PEEK are crystallographically equivalent.

Crystalline polymers (particularly filled grades – glass, carbon, mineral) retain useful mechanical properties above their glass transition temperature (T_g). Thus, although PEEK has a T_g of 143 °C its continuous service temperature (CST) is 250 °C. A further advantage of the crystalline range is inertness to aggressive media: PEEK shows outstanding resistance to aqueous acids, alkali, and organic solvents.

History [1]. Two synthetic routes lend themselves to the production of high-temperature polyaromatics: polycondensation (nucleophilic) or electrophilic processes. All of the polymers considered here may be made by polycondensation, and are currently manufactured by this process. Initially, electrophilic routes to PES (ICI) and various polyketones were developed but were quickly superseded by the polycondensation process, largely due to the lack of selectivity for linear polymers and the corrosive reactants used in the electrophilic process.

Raychem in 1972 launched their version of PEK (Stilan) produced via electrophilic chemistry, but this material was short-lived in the marketplace. Development of electrophilic processes continued, and polyketones were announced by Du Pont (PEKK), and BASF (Ultra PEK).

Routes to PPS (Philips), PES (ICI), and PSU (Union Carbide, now produced by Amoco) were developed during the 1960s with commercial launch in the early 1970s. Similarly, polycondensation routes to PEEK (ICI), PEK (ICI), and PEI (General Electric, GE) were developed in the 1970s with launches in the 1980s.

To put the timescale of one of these developments into perspective, ICI research for routes to polyketones started in 1972. The first 20 g sample of PEEK was made in ICI laboratories at Welwyn Garden City in 1977, followed quickly by scaleup to semitechnical quantities. PEEK manufacture moved to full production scale at Hillhouse in late 1978 and became commercially available in 1981. A dedicated plant for PEEK production was brought on-line at Hillhouse in 1987.

One of the keys for success for polymers made by polycondensation is the use of high-purity monomers to guarantee the high degree of reactivity required (> 99 %) for useful molecular masses, and the high degree of stereospecificity for properties such as toughness. Indeed, the history of the development of high-temperature polymers is mirrored by the development of their component monomers.

Stability. High-temperature thermoplastic polymers which have found commercial application have done so because they can be processed using conventional techniques such as extrusion or injection molding. However, for these materials to stand the processing temperatures employed they must exhibit stability against degradation. Typically an amorphous polymer is processed at 130 – 150 °C above its T_g and a crystalline polymer at 30 – 50 °C above its crystalline melting temperature (T_m). Thus, PES (T_g 225 °C) and PEI (T_g 217 °C) are processed at 350 – 370 °C while PEEK (T_m 343 °C) is processed at 370 –390 °C. Filled and high-viscosity grades may require even higher temperatures.

High temperature stability is attributable to high melting or softening points, high bond dissociation energies, structures not susceptible to chain (unzipping) degradation reactions, lack of inter- or intrabond formation, and low chemical (particularly to oxygen) reactivity. Resistance to chain degradation is perhaps the most important and is most easily provided by replac-

ing aliphatic with aromatic units. Thus although aliphatic polymers have high bond dissociation energies and the first step in degradation is a high-energy process, it yields highly reactive, free-radical species which "unzip" the chain. Unzipping is precluded in aromatic polymers because any free radicals generated are stabilized by delocalization via the π system. Furthermore, volatile fragments cannot be readily formed.

Clearly, the totally aromatic polyparaphenylene should therefore have optimum stability. It is indeed stable to above 500 °C. However, extreme chain stiffness results in an intractable material. In order to take advantage of aromatic stability but provide processibility, flexibilizing link units are incorporated into the chain [O, S, $C(CH_3)_2$]. Other, relatively inert moieties may be used to replace the aromatic ring (e.g., SO_2, CO). The balancing of these groupings gives the polymers their characteristic properties.

2. Polyetherimides

In 1982 General Electric Plastics introduced the first and still only thermally processible polyetherimide (PEI) resin under the trade name Ultem [*61128-24-3*]. It was the result of a ten-year development effort headed by J. Wirth of General Electric's Center of Research and Development in Schenectady (USA). The idea behind this development was to maintain the already known thermal and chemical stability associated with the aromatic ether and imide structures while building a polymer chain flexible enough to provide melt processability within the limits of the melt stability.

Polyimides are condensation polymers derived from bifunctional carboxylic anhydrides and primary diamines. Polyetherimides form a special class within these polyimides as they contain an ether group in either or both the aromatic dianhydride or aromatic diamine building block. A number of possible monomeric units are shown in Table 1. To date, Ultem produced from bisphenol A bisether dianhydride is the only commercially available thermoplastic polyetherimide. The synthesis thereof forms the basis for the development of these PEI resins [2, 3].

2.1. Production

Polyetherimides are synthesized by melt polycondensation of bisphenol A dianhydride with a diamine, usually *m*-phenylenediamine [4]. The starting dianhydride for PEI is produced as follows. First, an *N*-alkyl phthalimide is nitrated:

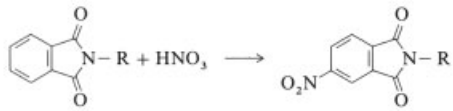

The NO_2 groups are then nucleophilically displaced by bisphenolate A (BPA) ions:

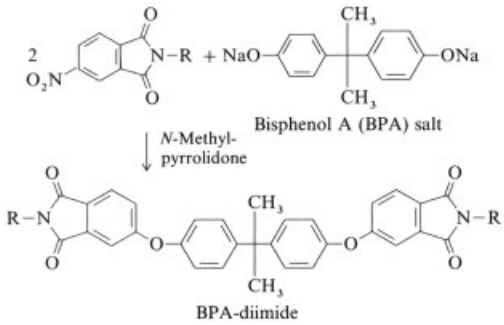

The resulting BPA diimide can be converted into the desired BPA dianhydride by an exchange reaction with phthalic anhydride:

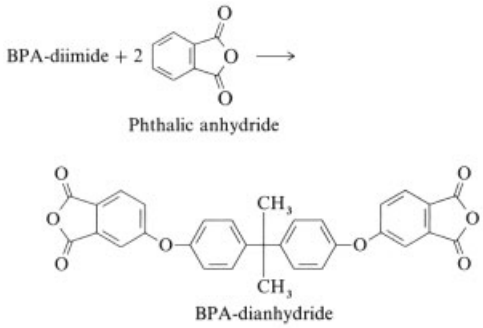

The final polycondensation of the BPA dianhydride with an aromatic diamine (e.g., *m*-phenylenediamine) can be performed in a number of ways, ultimately resulting in the formation of PEI:

Table 1. Possible monomers in polyetherimides

Name	Structure
Dianhydrides	
Hydroquinone bis(3,4-dicarboxyphenyl)ether dianhydride	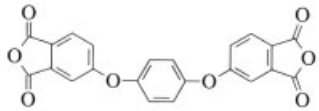
Bis(3,4-dicarboxyphenyl)ether dianhydride	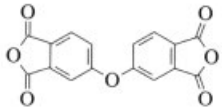
Bisphenol A bis(3,4-dicarboxyphenyl)ether dianhydride	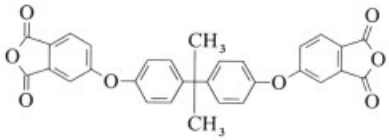
Diamines	
4,4′-Diaminodiphenyl ether	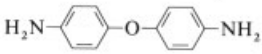
1,4-Bis(4-aminophenoxy)benzene	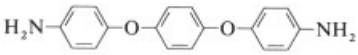
1,3-Bis(4-aminophenoxy)benzene	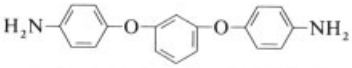

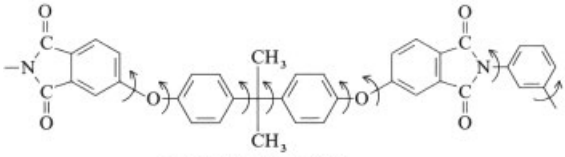

Ultem Polyetherimide
(GE Plastics)

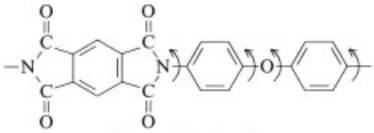

Kapton (Du Pont)

Figure 1. Polyetherimide (PEI) structures

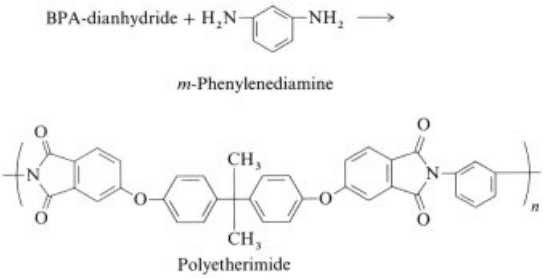

PEI resins with a number-average molecular mass of 15 000 – 17 000 are usually produced. The number-average molecular mass can be adjusted with monofunctional phthalic anhydride [5, 6]. Commercial PEI grades are tuned to the specific needs of an application and are obtained in three ways:

1) Use of additives, such as reinforcing fibers, mineral fillers, conductive fibers, pigments, polytetrafluoroethylene (to reduce frictional resistance), release agents, and flow improvers.

2) Modification of the base resin by copolycondensation with other monomers such as *p*-phenylenediamine and pyromellitic anhydride. This allows for higher glass transition temperatures and/or changes in chemical resistance.
3) Blending with other polymers (e.g., polycarbonate) improves processibility and impact resistance.

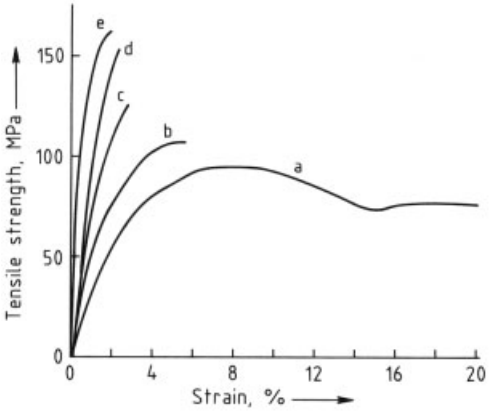

Figure 2. Stress – strain curves of unmodified PEI and glass reinforced PEI
a) Unmodified PEI; b) 10 % glass-reinforced PEI; c) 20 % glass-reinforced PEI; d) 30 % glass-reinforced PEI; e) 40 % glass-reinforced PEI

2.2. Properties

The structures of Ultem and Kapton, a polyetherimide produced by Du Pont (available as "Kapton polyimide film"), are shown in Figure 1 [7–9]. The segmental mainchain rotational freedom in Ultem (shown by arrows) provides this PEI molecule with many more configurational possibilities than the more symmetrical Kapton. Ultem therefore has a lower glass transition temperature (217 °C) than Kapton (360 °C) and can be more readily melt processed. The thermal stability of both polymers is almost the same (± 500 °C under nitrogen) [10]. Properties of some Ultem grades are listed in Table 2.

Mechanical Properties. Ultem has a very high tensile strength of about 110 MPa compared to, for example, polycarbonate (60 MPa), as well as a long-term, high-load, high-temperature dimensional stability. This is the combined effect of a high glass transition temperature, a high tensile modulus, a low thermal expansion coefficient, low creep, and a low water absorption. Stress – strain behavior is shown in Figure 2.

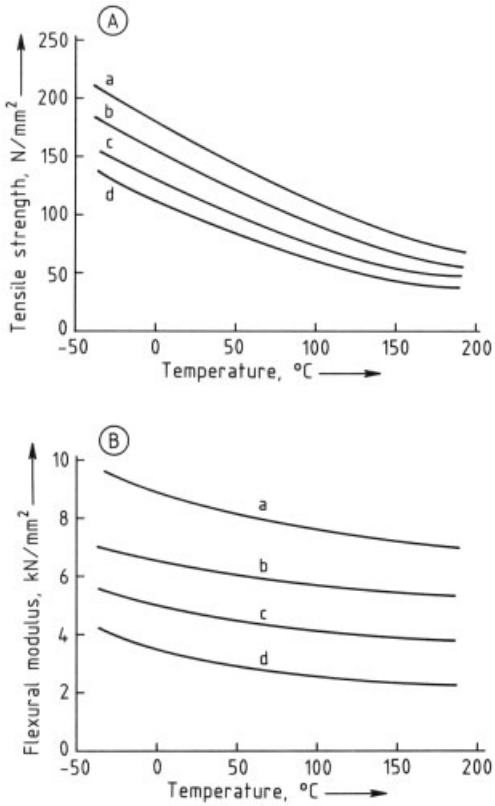

Figure 3. Tensile strength (A) and flexural modulus (B) of nonreinforced and glass-reinforced PEI
a) Ultem 2300 (30 % glass-reinforced); b) Ultem 2200 (20 % glass-reinforced); c) Ultem 2100 (10 % glass-reinforced); d) Ultem 1000 (nonreinforced)

The relatively constant mechanical property profile of the PEI grades as a function of glass fiber content over a wide temperature range (up to 175 °C) is a direct result of the high glass transition temperature (217 °C) of this amorphous resin (Fig. 3). The excellent adhesion of PEI resins to glass fibers gives a linear relationship between the flexural modulus or tensile strength and glass fiber content up to 30 wt % [11].

Fatigue properties are excellent. The endurance (stress) limit at $10^6 - 10^8$ cycles for PEI reinforced with 30 % glass fibers is 30 MPa. Isochronous stress – strain diagrams from which

Table 2. Properties of Ultem PEI

Property	Standard	Ultem grade (x_1, x_2, x_3, x_4) *					
		1000	1010	2100, 2110	2200, 2210	2300, 2310	2400, 2410
Mechanical							
Tensile strength, yield, MPa	ISO 527	105	105	120			186
Tensile strength, break, MPa	ISO 527	90	90	115	140	160	
Tensile elongation, yield, %	ISO 527	8	8	5			
Tensile elongation, break, %	ISO 527	60	60	6	3	3	2
Tensile modulus, GPa	ISO 527	3.0	3.0	4.5	6.9	9.0	11.7
Flexural strength, yield, MPa	ISO 178	145	145				
Flexural strength, break, MPa	ISO 178			200	210	230	250
Flexural modulus, GPa	ISO 178	3.3	3.3	4.5	6.2	9.0	11.7
Hardness, Rockwell		M 109	M 109	M 109	M 114	M 114	M 125
Izod impact strength notched, J/m	ISO 180/1A	50	30	60	90	100	105
Thermal							
Vicat softening temperature, rate B, °C	ISO 306	219	219	223	226	228	234
Vicat VST/B/120, °C	ISO 306	215		222	224	226	
DTUL** (0.45 MPa), °C	ISO 75	210	207	210	210	212	216
DTUL** (1.82 MPa), °C	ISO 75	200	197	207	209	210	213
Thermal conductivity, $W\,m^{-1}\,{}^{\circ}C^{-1}$	DIN 52 612	0.22	0.22				
Coefficient of linear expansion, flow, $mm^{-1}\,{}^{\circ}C^{-1}$	DIN 53 752	5.6×10^{-5}	5.6×10^{-5}	3.2×10^{-5}	2.5×10^{-5}	2.0×10^{-5}	1.4×10^{-5}
Physical							
Density, g/cm^3	ISO 1183	1.27	1.27	1.34	1.42	1.51	1.61
Water absorption, 24 h, %	DIN 53 495	0.25	0.25	0.28	0.26	0.18	0.12
Water absorption, equilibrium, %	DIN 53 495	1.25	1.25	1.00	1.00	0.90	0.90
Mold shrinkage, flow, %	ASTM D955	0.7	0.7	0.5 – 0.6	0.3 – 0.6	0.2 – 0.4	0.1 – 0.3
Electrical							
Volume resistivity, $\Omega\cdot m$	IEC 93	6.7×10^{15}	6.7×10^{15}	1.0×10^{15}	7.0×10^{14}	3.0×10^{14}	
Dissipation factor (2450 MHz)	IEC 250	0.0025	0.0025	0.0046	0.0049	0.0053	

Table 2. Continued

Property	Standard	Ultem grade (x_1, x_2, x_3, x_4) ∗					
		1000	1010	2100, 2110	2200, 2210	2300, 2310	2400, 2410
Comparative tracking index, V	IEC 112	100 – 175	100 – 175	100 – 175	100 – 175	100 – 175	100 – 175
Flame characteristics							
Oxygen index, %	ISO 4589	47	44	47	50	50	54
Flame retardancy (mm thickness)	UL 94	V-0 (1.60) 5 V (1.90)	V-0 (0.71)	V-0 (0.41)	V-0 (0.41)	V-0 (0.25)	V-0 (0.25)

∗ The first three digits (x_1, x_2, x_3) have the following meanings: x_1=1 unfilled resin; x_1=2 glass-fiber-reinforced resin; x_2=1, x_2=2, x_2=3, glass fiber content=10, 20, and 30 %, respectively; x_3=0 standard resin; x_3=1 easy flow. Ultem is a registered trade mark of GE Plastics. ∗∗ DTUL=deformation temperature ultimate limit.

the creep performance up to 100 000 h can be extrapolated, indicates a good resistance of PEI against creep (Fig. 4).

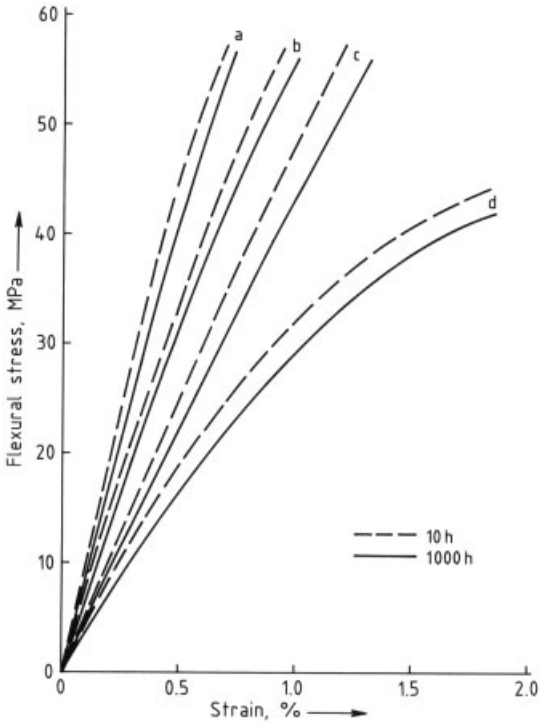

Figure 4. Isochronous stress – strain diagram of PEI (Ultem) at 23 °C
a) 30 % glass-reinforced; b) 20 % glass-reinforced; c) 10 % glass-reinforced; d) Unfilled

Aging characteristics (UL 94 temperature index) indicate a continuous service life up to 100 000 h at 175 °C.

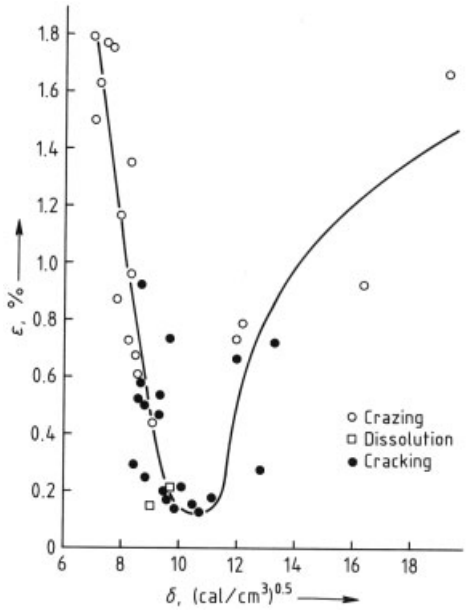

Figure 5. Dependence of critical strain ε_c for crazing, cracking, and dissolution of PEI versus liquid solubility parameter δ

Impact Behavior. The unfilled resin has excellent ductility due to the large number of possible segmental rotations and the high entanglement density that results from the kinked structure of the macromolecular chain. Its rather high notch sensitivity can be overcome by making sure that parts made of PEI have wellrounded edges.

Chemical Resistance. Environmental stress crazing and cracking are typical problems of amorphous engineering plastics, including to a

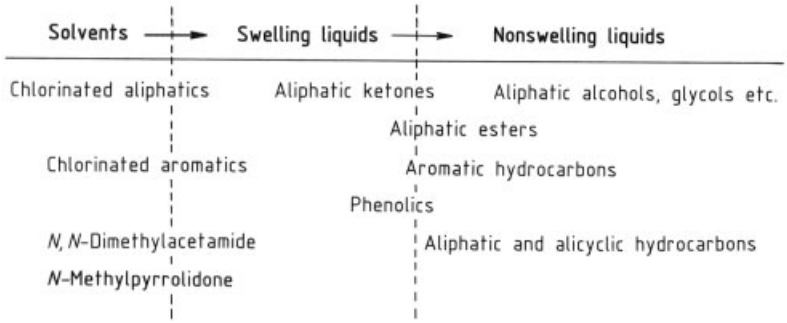

Figure 6. Chemical resistance of polyetherimides
Neutral basic, acidic and/or oxidative aqueous systems have no effect with the exception of aqueous amine and ammonia which cause hydrolytic breakdown of PEI resins.

lesser extent also PEI resins. Contact and subsequent interaction between the polymer and a liquid may result in dissolution or swelling, depending on the strength of the molecular interaction forces between the polymer and the liquid molecules. These forces are characteristic for each combination of liquid and amorphous polymer and are determined by the differences in solubility parameter δ, defined as

$$\delta = \sqrt{(\text{cohesive energy}/\text{molar volume})} = \sqrt{\text{cohesive energy density}}$$

For PEI δ is 9.3 – 11.7 $(\text{cal/cm}^3)^{0.5}$ [12]. Its behavior toward contact with liquids of different solubility parameters is shown in Figure 5. If the difference between the solubility parameters of PEI and the liquid is larger than 5, i.e. $-\delta_{\text{PEI}} - \delta_{\text{liquid}} - > 5\,(\text{cal/cm}^3)^{0.5}$, then no stress cracking problems are to be expected. However, if $-\delta_{\text{PEI}} - \delta_{\text{liquid}} - < 5\,(\text{cal/cm}^3)^{0.5}$, swelling of PEI is possible and stress cracking may result from load, leading to actual strain which in time may result in crack problems. Moreover, if $\delta_{\text{PEI}} \approx \delta_{\text{liquid}}$, the polymer dissolves and the part disintegrates. The chemical resistance of PEI towards several solvents is summarized in Figure 6.

Flammability. The flammability behavior of Ultem as characterized by a high limiting oxygen index (LOI) of 47 results from the highly aromatic structure with a hydrogen/carbon molar ratio of 0.64. Selfsustained burning requires an oxygen rich atmosphere of over 45 vol %, making Ultem a selfextinguishing engineering plastic [13]. Therefore PEI resins fulfil the most severe flammability requirements stipulated by various countries for different industrial segments and part thicknesses:

1) Electrical and electronics industry: Rating V-0 at 0.64 mm wall thickness according to UL 94 Rating 5 V at 1.5 mm wall thickness according to UL 94
2) Building and construction industry: Class B-1 at 1.5 mm according to DIN 4102 (Germany) M 1 rating at 1.5 mm according to NFP 92-501 (France) Class 0 at 1.5 mm according to BS 476, parts 6 and 7 (UK)
3) Aircraft interior construction Passed FAR 25.853 at 1.5 mm (Federal Aviation Regulations, USA)

Requirements with respect to heat, smoke, and toxic gas generation during burning can also be met. Both the heat release rate criterion of ≤ 65 W at a 65 W radiation energy input and the NBS smoke density criteria (measured according to ASTM D 662) can easily be complied with.

Electrical properties of unfilled PEI and PEI reinforced with 30 % glass fiber are presented in Table 3. The dielectric constant ε has a nearly constant value over the entire service temperature and frequency range. The temperature and frequency dependence of the dissipation factor (Fig. 7) are much lower for PEI than for other engineering resins. The values at 915 and 2450 MHz are of particular interest since these are the most well-developed microwave bands for power applications. The low value (in an absolute sense) ensures that dielectric losses are minimal, especially at high temperatures.

Table 3. Electrical properties of PEI

Property	Test	Unfilled	30 % reinforced with glass fiber
Dielectric strength, V/mm	D 149	2840	2520
Dielectric constant ε (1 Hz)	D 150	3.15	3.7
Dissipation factor (1 kHz, 50 % R.H.)	D 150	0.0013	0.0015
Volume resistivity, $\Omega \cdot$ cm	D 257	6.7×10^{17}	3×10^{16}

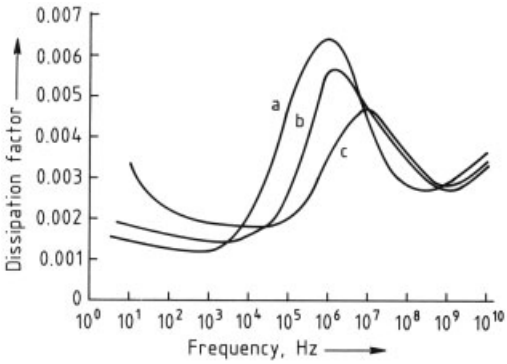

Figure 7. Dissipation factor of PEI unmodified resin versus frequency at 50 % relative humidity (ASTM D 150)
a) 23 °C; b) 49 °C; c) 82 °C

Weather and radiation resistance of PEI parts is excellent. A 1000-h weatherometer test had no effect on mechanical properties.

2.3. Processing

The rheological characteristics of PEI resins allow them to be subjected to all thermoplastic processing techniques, i.e., injection (blow) molding, extrusion, structural foam molding, and compression molding [14]. Processing should be based on guidelines applicable to amorphous engineering plastics, but with appropriate temperature adjustments. The high glass transition temperature (217 °C) requires a melt temperature typically in the range 360 – 425 °C. Clean, well-processed regrind can be recycled in the feed in amounts of 20 – 50 %. PEI resins absorb water and have to be extensively predried before they are melt processed to prevent severe surface splay upon molding or extrusion. Heating for at least 4 h at 150 °C is required to lower the water content to the desired level of 0.03 %.

Although processing fumes are negligible, proper ventilation is advised. Purging of the processing equipment can be achieved with high molecular mass polyethylene in combination with a gradual lowering of the melt temperature from 400 to 200 °C.

2.4. Uses

Commercial PEI grades are used in several industrial segments, of which the most important are reviewed below [15].

The *automotive industry* uses PEI mainly in under-the-hood applications, where long-term mechanical strength and dimensional stability are required at high service temperatures in combination with resistance to oil, grease, brake- and cooling fluids. Examples are sensor housings, oil pump parts, ball bearing cages, and bearings.

PEI can easily be metallized and is therefore used in head-lamp reflectors, especially in combination with energy-intensive halogen lamps.

The *lighting industry* has similar requirements to those mentioned above for head-lamp reflectors. A special feature of PEI is its IR transparency. Thus unfilled, unpigmented PEI reflectors made by dichroic layer technology can produce "cool" light. Heat is not concentrated by the reflector but diffuses through it. This principle is of importance in lighting systems employed in operating theaters, art exhibitions, etc. The dichroic coating technique uses the different refraction of various wavelengths at interfaces and its dependence on layer thickness to separate the IR and the visible part of the emitted light. The visible part of the spectrum is reflected, whereas the IR part diffuses through the PEI.

The *electrical and electronic industries* make use of PEI's intrinsic flame resistance and its excellent dielectric properties. Notable is the development of two- and three-dimensional injection-molded, printed circuit board structures where the design allows integration of mechanical and circuitry functionality. A special, newly developed, metal-plating technique for PEI makes this even more interesting

[16]. Other applications include circuit boards, connectors, switches, and controls. PEI's low tracking resistance (comparative tracking index=145 V, according to IEC standards) makes it unsuitable for high-voltage parts that carry direct current.

Polyetherimides are also used in the *food processing and beverage industries* on account of their high heat resistance, favorable mechanical properties, good chemical resistance to and cleanability (sterilization) of contact liquids, as well as lack of toxicity and disadvantageous effects in taste and odor. Specially developed PEI grades fulfil these requirements and have been formally approved by most European health agencies (e.g., Bundesgesundheitsamt in Germany) and the FDA in the United States. Polyetherimides are used in liquid-handling equipment, pumps, high-pressure dispensors, coffee makers, kitchen utensils, microwave trays, dishes and cups, cheese molds, and lightweight unbreakable baking forms. The food processing industry is becoming one of the most important volume markets for PEI.

PEI parts are also utilized in *household appliances*, e.g., water reservoirs for steam irons, hair dryers, doors of microwave ovens, and coffeemaker reservoirs.

Medical equipment often has to be sterilized and must be inert to body fluids. Here, parts made of PEI have a good performance and can be sterilized by all conventional methods (e.g., use of steam, irradiation, and chemicals). The intrinsic transparency of the material is also advantageous. Examples of applications in the medical equipment sector include surgical staplers and sterilization tanks.

The *aircraft industry* is also interested in materials that offer cost and weight savings without sacrificing safety (especially in relation to flammability issues). Components made of PEI with a high volume fraction of glass fibers or other high modulus fiber provide an excellent substitute for metal (especially aluminum) parts. The injection moldability of PEI resins allows the economic production of complex parts and multifunctional integration. High-performance composite sheets with PEI as the binding resin are used in monolithic and laminated structural parts, such as interior partition walls and floor elements. Ducts and casings are also made from PEI composites. The composites have a thermoplastic matrix and can easily be thermoformed into complex shapes.

2.5. Economic Aspects

The commercial development of PEI has been very successful. Since the introduction of the first commercial Ultem grade in 1982, annual worldwide sales volume exceeded 5000 t in 1990. Future growth is expected to increase because applications are expanding and further market development activities are underway. Potential large-volume markets in composites and in household and industrial appliances are emerging. Extra market growth may be expected from developments in high-speed trains, where the same composite materials can be used as in the aircraft industry due to the emphasis on fire safety issues. Possible markets for the compatible PEI – polycarbonate blends and the completely miscible PEI – polyetherketone or PEI – polyetheretherketone blends have not yet been fully exploited [17]. Moreover, the product range of these materials will be further enlarged by simple modification of the PEI structure through copolymerization, leading to enhanced inherent properties. An example of the flexibility offered by chemical modification of the polymer main chain is siloxane – etherimide block copolymers (PSEI) [18]. Siloxane segments provide not only flexibility, toughness, and reduced moisture absorption, but also specific surface properties and lower dielectric constant. The length of the siloxane blocks is the key to these properties [18]. PSEI copolymers have been commercialized by GE Plastics under the trade name Siltem as high-performance coating materials for electrical wires. They are also used with Ultem in aircraft interior applications where they improved the impact performance of PEI, while maintaining its flame retardancy and heat resistance. Applications of PEI do not pose any environmental issues. It is possible to recycle PEI, provided the material has been well processed and does not contain impurities.

3. Polyaryletherketones

The first reports of syntheses of polyaryletherketones appeared in the early 1960s [19, 20]

and were based on electrophilic substitution reactions, essentially Friedel–Crafts acylations. The problem of limited solubility of the crystalline polymers was solved by conducting polymerization in strongly acidic media, initially polyphosphoric acid [21] and later HF/BF_3 [22], which dissolved the polymers by protonation of their carbonyl groups. Despite the fact that high molecular mass polymers with useful mechanical properties could be produced, production problems associated with the highly aggressive reaction medium and problems in obtaining pure, stable polymers retarded the full commercialization of polyaryletherketones.

Poly(oxy-1,4-phenylenecarbonyl-1,4-phenylene) [*27380-27-4*] (polyetherketone, PEK), synthesized by polyacylation in HF/BF_3, was available on a limited scale from Raychem for a short period in the mid 1970s as "Stilan".

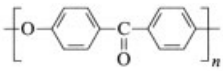

Polyetherketone PEK

The first synthesis of polyaryletherketones by nucleophilic displacement was reported by Johnson et al. [23] (Union Carbide) in the late 1960s, but the molecular masses of the polymers were low due to crystallization resulting in premature precipitation; mechanical properties were consequently poor.

The key breakthrough in the synthesis of polyaryletherketones by nucleophilic displacement was the identification by Rose et al. [24,25] (ICI) of diphenyl sulfone as a suitable solvent for the potentially highly crystalline, high molecular mass polymers. Nucleophilic displacement polymerization in this solvent formed the basis of the large scale industrial route to poly(oxy-1,4-phenyleneoxy-1,4-phenylenecarbonyl-1,4-phenylene) [*31694-16-3*]

(polyetheretherketone, PEEK), the first polyaryletherketone to be fully commercialized. PEEK has been commercially available from ICI since the early 1980s and PEK, synthesized by a similar nucleophilic route, since 1987.

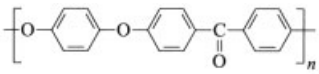

Polyetheretherketone PEEK

The major producer of polyaryletherketone resins is ICI (Victrex PEEK and PEK). Several other companies have introduced polyaryletherketone resins (Du Pont with PEKK, Amoco with Kadel, Hoechst with Hostatec, and BASF with Ultrapek).

3.1. Production

PEEK. The most commercially significant polyaryletherketone, PEEK, is prepared by reacting 1,4-benzenediol with 4,4′-difluorobenzophenone in diphenyl sulfone in the presence of alkali-metal carbonates under an inert atmosphere at temperatures approaching the melting point of the polymer (> 300 °C) [26]. The polymer is isolated from the diphenyl sulfone and inorganic impurities, which can destabilize the polymer melt, by solvent extraction (leaching). A slight excess of 4,4′-difluorobenzophenone is normally employed to prevent the formation of phenoxide chain ends which can destabilize the polymer. The water formed as a result of polycondensation could potentially hydrolyze the fluoro groups and prevent formation of high molecular mass polymer but is lost due to the high reaction temperature and the hydrophobicity of solvent. The expensive difluoro monomer is required rather than the less expensive dichloro analogue because the weakly electron-withdrawing carbonyl group does not strongly activate the halogen to nucleophilic displacement and consequently a good leaving group is necessary for efficient reaction.

PEK is produced by ICI by two nucleophilic routes [25] which are essentially similar to the route to PEEK. The two-monomer route employs 4,4′-dihydroxybenzophenone in place of 1,4-benzenediol. The single-monomer route employs the potassium salt of 4-fluoro-4′-hydroxybenzophenone. The latter route has the greatest tendency to exhibit side reactions leading to branching (via carbanion formation due to abstraction of a proton ortho to the fluoro group of the monomer) which results in polymers with inferior mechanical properties. The two-monomer route is therefore usually preferred.

PEK has been produced by electrophilic acylation [27]. The most convenient synthe-

sis involves the Friedel – Crafts polycondensation of 4-phenoxybenzoyl chloride. The preferred reaction medium of HF/BF_3 necessitates the use of special polytetrafluoroethylene reaction vessels. This electrophilic route employs relatively cheap nonfluoro-substituted aromatic monomers. If the problems associated with the aggressive reaction medium can be solved, this may prove to be the cheapest source of PEK.

3.2. Properties

The similar structures of polyaryletherketones result in polymers with similar properties and are mainly exemplified with data for PEEK (Table 4).

Table 4. Properties of PEEK

Property *	PEEK 450 P	30 % glass reinforced
T_g, °C	143	143
T_m, °C	334	334
Heat distortion temperature, °C (D 648)	160	315
Continuous service temperature, °C	250	250
Water absorption, wt % (D 370)	0.5	
Tensile strength, MPa (D 638)	92	157
Flexural modulus, GPa (D 790)	3.66	10.3
Izod impact strength, J/m (D 256)		
Unnotched	no break	725
Notched	83	96

* Appropriate ASTM test methods are given in parentheses.

Physicochemical Properties. Most wholly aromatic polyaryletherketones are semicrystalline due to the strong interchain forces between the polar groups in the backbone and the crystallographic near equivalence of the ether and carbonyl units [28]. Many of the desirable properties of commercial polyaryletherketones are related to the tendency to develop a high degree of crystallinity under normal processing conditions.

The semicrystalline morphology of polyaryletherketones is largely responsible for their excellent resistance to solvent attack. PEEK is insoluble in all common solvents. It is soluble at room temperature only in strongly acidic media (e.g., concentrated sulfuric acid, hydrofluoric acid, and trifluoromethanesulfonic acid) that can protonate the carbonyl groups. At temperatures well above the glass transition temperature the polymer is soluble in a few unusual media (e.g., diphenyl sulfone, some high boiling point esters, benzophenone, and 1-chloronaphthalene). PEEK absorbs small amounts of water [29] but can absorb significant amounts of dichloromethane and related solvents that contain a hydrogen atom activated by neighboring electronegative groups [29, 30].

The insolubility of polyaryletherketones causes problems in the molecular mass characterization of the polymers. Viscometry and light scattering methods using concentrated sulfuric acid and a high-temperature gel permeation chromatography method have been developed [31]. The number-average molecular mass of commercial PEEK grades is typically of the order of tens of thousands.

The essentially linear aromatic backbone of PEEK results in excellent resistance to chemical attack and ionizing radiation. The only common reagents that degrade PEEK are chlorine, bromine, and concentrated nitric acid [32]. Dissolution of polyaryletherketones in concentrated sulfuric acid results in monosulfonation of the aromatic rings substituted by two ether linkages [33].

PEEK displays limited resistance to degradation by UV light, as is the case for all linear polyaromatics, but the problem is only serious in high insolation environments (i.e., prolonged exposure to intense sunlight) and is much reduced by incorporation of a suitable pigment (e.g., carbon black).

Thermal Properties. Differential scanning calorimetry studies of quenched amorphous PEEK indicate a T_g at 143 °C, a crystallization exotherm centered at 180 °C, and a melting endotherm with a maximum at 334 °C [34]. Increasing the ratio of carbonyl to ether linkages in polyarylether backbones increases the T_g and the melting point, PEK therefore has a T_g at 160 °C and T_m at 365 °C.

Isothermal crystallization studies indicate that PEEK crystallizes very rapidly [34] and normal processing conditions generally result in degrees of crystallinity of 25 – 35 %.

The thermal stability of polyaryletherketones is excellent. PEEK melts are stable for > 1 h at 400 °C. Thermoxidative stability is lower with prolonged heating above 300 °C resulting in cross-linking. The presence of copper impurities

significantly accelerates the thermal degradation of PEEK melts [35].

Mechanical Properties. The high crystallinity rapidly developed by PEEK and PEK results in very high flexural modulus and tensile strength. These properties decrease substantially above the T_g but useful residual values are maintained to over 250 °C, close to the melting of the crystalline regions and the maximum service temperature of the polymers reflects this. The higher T_g and T_m of PEK result in improved high-temperature retention of properties.

PEEK compositions containing glass or carbon fibers have improved mechanical properties and high-temperature performance.

Commercial polyaryletherketones are hard and display excellent wear and abrasion resistance. They are notch-sensitive like many thermoplastics, i.e., the presence of a stress-raising feature lowers the otherwise acceptable impact strength.

Electrical Properties. PEEK has a high volume resistivity (i.e., is a good insulator) and a low dissipation factor for electromagnetic radiation. Furthermore these are stable over a wide temperature range up to the T_g.

3.3. Processing

The thermal stability of commercial polyaryletherketones allows melt processing in conventional equipment [36]. This, coupled with the ability to recycle scrap material, represents a major advantage over thermosets with otherwise comparable performance.

Articles are readily prepared by injection or compression molding. Many other common fabrication techniques, notably thermoforming, centrifugal casting, and machining, are also possible. Extrusion techniques can be used to make amorphous or crystalline film and monofilament or multifilament fibers.

3.4. Uses

The uses of polyaryletherketones are generally connected with their outstanding high-temperature performance and chemical resistance.

PEEK and PEK have been widely used for cable insulation in environments where a combination of chemical resistance, high-temperature performance, abrasion resistance, and radiation resistance is required (e.g., in oil wells, nuclear installations, or aerospace applications).

In general load-bearing applications the polyaryletherketones replace metal where weight saving is critical (aerospace applications) or where the cost of the finished article is lower than the metal equivalent (recycling of scrap material combined with lower energy costs during processing).

Excellent hydrolysis and chemical resistances have resulted in PEEK replacing metal in many fluid-handling applications. Low wear rate and high abrasion resistance favor PEEK for bearing applications in aggressive environments.

PEEK film (ICI Stabar film) has been used as a barrier material and as a substrate for flexible printed circuit boards to be used at high temperature. PEEK monofilament has been woven to produce temperature- and chemical-resistant filters and transmission belting. PEEK fibers are described elsewhere, see → High-Performance Fibers, Chap. 2.3.2.

An important end use of polyaryletherketone resins is in the production of a carbon fiber composite (ICI Advanced Polymer Composite, APC) [37] which is used widely in aerospace applications due to a high strength to weight ratio coupled with resistance to aggressive service environments.

4. Polysulfones

Polysulfone is the generic term for all sulfone-containing polymers. Polysulfones are an important class of engineering thermoplastics. The polymer backbone consists of *para*-linked aromatic groups connected by ether, sulfone, and in some cases alkyl groups.

Development work on polysulfones was active by the early 1960s. Early syntheses were electrophilic, being in effect polysulfonylations [38]. This type of reaction did not generally produce a linear *para*-linked chain, and the introduction of *ortho* linkages and branchpoints resulted in polymers with poor mechanical properties [39].

An alternative strategy for the synthesis of polysulfones based on nucleophilic aromatic substitution reactions [40, 41] proved more successful for the production of *para*-linked, essentially linear polymers with good mechanical and thermal properties. This route was subsequently employed in all large-scale commercial routes to polysulfones. By the late 1960s the first commercial polysulfone, poly(oxy-1,4-phenylenesulfonyl-1,4-phenyleneoxy-1,4-phenylene(1-methylethylidene)-1,4-phenylene) or bisphenol A polysulfone [*25135-51-7*]

(Udel, Union Carbide) was available.

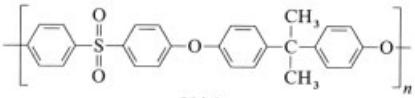

Udel

This was followed by the introduction in the early 1970s of the polyethersulfone poly(oxy-1,4-phenylenesulfonyl-1,4-phenylene) [*25667-42-9*]

(Victrex, PES) by ICI and Astrel 360 by 3 M, a biphenyl containing polysulfone with a high glass transition temperature but poor processability.

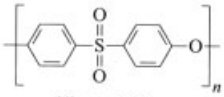

Victrex PES

More recent developments include the introduction of polysulfones containing biphenyl groups in the repeat unit (Radel R polyphenylsulfone [*68518-59-2*], Union Carbide; Victrex HTA [*121763-41-5*], ICI) which have higher T_g values but retain melt processibility.

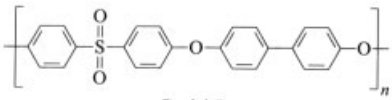

Radel R

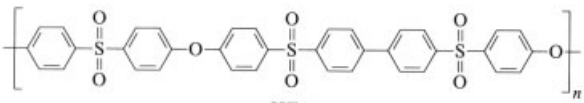

HTA

The major manufacturers of polysulfone resins are Amoco, which acquired the Union Carbide polysulfone business

(Udel, Radel A and R), and ICI (Victrex PES and Victrex HTA). BASF marketed comparable polysulfones to the ICI and Amoco polymers but on a smaller scale (Ultrason S Polysulfone, Ultrason E Polyethersulfone).

Since drafts were first submitted for publication, ICI has announced that it is to withdraw from the manufacture and sale of Victrex PES and HTA. BASF have announced a capacity increase for their Ultrason products.

4.1. Production

Bisphenol A polysulfone (Udel) is synthesized by nucleophilic substitution of 4,4′-dichlorodiphenyl sulfone (DCDPS) by 4,4′-(1-methylethylidene)bisphenol (bisphenol A) in a dipolar aprotic solvent such as dimethyl sulfoxide (DMSO) [42]. The reaction is carried out in two stages. Firstly addition of two equivalents of aqueous NaOH or KOH to a solution of the bisphenol in a mixture of DMSO and azeotroping solvent (e.g., chlorobenzene) generates the more nucleophilic bisphenate salt. Excess water is then removed by azeotropic distillation at 120 – 140 °C before the dichloro monomer, which could be hydrolyzed by water and lead to a low molecular mass product, is added and polymerization proceeds. Polymerization is carried out at 130 – 160 °C under an inert atmosphere to prevent oxidation of the bisphenate salt. The dipolar aprotic solvent promotes very rapid rates of nucleophilic substitution and the molecular mass of the polymer can become excessively high unless it is controlled by the addition of monofunctional halides or phenols.

Radel R and Radel A (structure not disclosed) are probably also synthesized by this type of

route but employing the appropriate bisphenol monomers.

The synthesis of Victrex PES is similar to the synthetic route to Udel with 4,4′-sulfonylbisphenol (bisphenol S) being condensed with DCDPS [43]. The lower solubility and reactivity of the bisphenol (nucleophilicity reduced by electron-withdrawing sulfone group para to the hydroxy group) necessitates a significantly higher reaction temperature (up to 285 °C). Consequently, the dipolar aprotic solvent used is diphenyl sulfone, phenate salts being notionally generated "in situ" by the action of K_2CO_3 or Na_2CO_3. The high reaction temperature and the hydrophobicity of the solvent result in the efficient removal of water that might otherwise hydrolyze the dichloro monomer. Azeotropic distillation is not necessary.

Victrex HTA is produced by an analogous route but with the DCDPS being replaced by bis(4-chlorophenylsulfonyl)biphenyl.

4.2. Properties

Physicochemical Properties. Key properties of Udel polysulfone and Victrex PES are listed in Table 5.

Table 5. Properties of Udel polysulfone and PES

Properties ∗	Udel polysulfone P 1700	Victrex PES 4100 G
T_g, °C	185	223
Heat distortion temperature, °C (D 648)	174	203
Water absorption, % (D 570)	0.22	0.66
Tensile strength, MPa (D 638)	70	84
Flexural modulus, GPa (D 790)	2.7	2.6
Izod impact strength, J/m (D 256)		
Unnotched	no break	no break
Notched	69	83

∗ Appropriate ASTM test method given in parentheses.

All commercial polysulfones are essentially amorphous as utilized. The unfilled polymers display good clarity.

The polysulfones are relatively polar and their solubility characteristics reflect this. Polysulfones are soluble to some extent in a variety of dipolar aprotic solvents (e.g., dimethyl sulfoxide, dimethylformamide, *N*-methylpyrrolidone) and certain chlorinated hydrocarbons (e.g., dichloromethane, chlorobenzene). Udel polysulfone is significantly soluble in low molecular mass ketones and some aromatic hydrocarbons [44]. Victrex PES and biphenyl-containing polysulfones are more solvent resistant than Udel polysulfone [45]. The polysulfones are crazed but not dissolved by solvents with polarities similar to that of the polymer and this tendency increases with stressed samples.

Polysulfones can absorb small amounts of water but this has a minimal adverse effect on dimensions and mechanical properties. Polysulfones are very resistant to hydrolysis, even by hot water or steam. They are stable to dilute acid and caustic but are degraded by concentrated nitric and sulfuric acids. Chlorine and bromine also degrade polysulfones.

Polysulfones show limited stability to UV light and the materials must be filled for outdoor applications. Resistance of the wholly aromatic polysulfones to ionizing radiation is excellent, Udel polysulfone has a lower resistance due to the presence of aliphatic groups.

The ready solubility of polysulfones allows the determination of average molecular masses and distributions by a variety of common techniques including gel permeation chromatography. Number-average molecular masses of commercial polysulfones are typically of the order of tens of thousands.

Thermal Properties. Commercial polysulfones cannot be crystallized from the melt or by annealing and are amorphous as utilized. The T_g tends therefore to be responsible for short-term temperature ceilings (e.g., heat deflection temperature) for applications of the polymers. Aromatic rings and sulfone groups introduce chain rigidity whereas ether links increase flexibility. Hence increasing the ratio of sulfone to ether links in the repeat unit increases the T_g. The presence of biphenyl groups (e.g., HTA and Radel R) also increases chain rigidity and hence T_g.

The thermal stability of polysulfones is excellent due to the strong bonds in the backbone and results in stable melts during processing below 400 °C. The aliphatic isopropylidene groups in Udel polysulfone significantly reduce the thermooxidative stability relative to the wholly aromatic polysulfones and prevent prolonged high-temperature use close to T_g. Melt processing

on conventional equipment is still possible however.

Commercial polysulfones possess selfextinguishing behavior. Despite the presence of sulfur-containing groups, they exhibit relatively low levels of smoke and toxic gas emission.

Mechanical Properties. The polar groups in the polysulfone chain result in a very high flexural modulus. Modulus decreases rapidly as T_g is approached. The impact strength of the unnotched polymers is also high, but the presence of molded or accidental notches dramatically reduces the impact strength. The notched impact strength of the biphenyl-containing polysulfones is significantly higher than that of other polysulfones.

Filling of the polymers with glass or carbon fiber substantially increases the modulus of the polysulfones but generally only increases the heat deflection temperature slightly.

Electrical Properties. The polysulfones have high resistivities and low dielectric constants and dissipation factors for electromagnetic radiation which are substantially retained up to the T_g of the polymer.

4.3. Processing [44,46]

Commercial polysulfones are all melt processible on conventional equipment, the required processing temperatures increasing with the T_g of the polymer. The polysulfones are dried before processing to prevent the evolution of adsorbed moisture from lowering the quality of moldings. Annealing at a temperature close to the T_g of the polymer is recommended to prevent stress cracking of moldings.

4.4. Uses

The mechanical properties of Udel polysulfone and Victrex PES are similar. If enhanced high-temperature performance or chemical resistance is required Victrex PES is preferred. The biphenyl-containing polysulfones are significantly more expensive than Udel and Victrex PES, their use is limited to situations where extreme heat or impact resistance is required. Their use as films and foamed plastics are described elsewhere, see → Films, Chap. 6.10., → Films, Chap. 6.11.,→ Films, Chap. 6.12.,→ Films, Chap. 6.13. and → Foamed Plastics.

Udel and Victrex PES have been widely used to produce injection-molded articles, replacing metals in low- to medium-temperature applications where weight saving is important – particularly aerospace applications. Notch sensitivity limits the application of the polymers in critical load-bearing components.

The high-temperature hydrolytic stability of polysulfones has permitted widespread use in fluid-handling components and steam-sterilizable biomedical moldings. Chemical resistance enables a range of chemical process and automotive applications.

Good insulating properties up to high temperatures favor polysulfones as insulation for electrical cables and as a circuit board material (→ Insulation, Electric, → Insulation, Electric).

5. Poly(Phenylene Sulfide)

Poly(phenylene sulfide) [*9016-75-5*] (PPS), poly(thio-1,4-phenylene), is the most important thermoplastically processible representative of the poly(arylene sulfides). The latter consist of repeating aromatic monomer units joined by sulfur bridges:

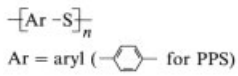

On account of its unusual properties, PPS is expected to experience an above-average increase in production and sales [47].

PPS was first obtained in 1888 as a polymeric byproduct by Friedel and Crafts [48,49]. Attempts by Genvresse (1897) to prepare PPS from benzene and sulfur with aluminum chloride resulted in cross-linked resins and a large number of low molecular mass byproducts [50, 51]. In the late 1940s Macallum developed a process for preparing poly(arylene sulfides) (including PPS). By reacting dichlorinated aromatics with sulfur, he obtained partially cross-linked products with higher sulfur contents than those predicted by their theoretical structure[52,53].

A series of publications describe the synthesis of alternating sequences of aromatic units by reacting dithiolates with dihalogenated compounds [54–57] and the first use of polar aprotic solvents [54].

In the early 1960s a process was developed by Dow which involved copper-induced self-condensation of aromatic halothiophenols and resulted in an experimental product [51,58,59].

In 1963 Edmonds and Hill developed a new process in which poly(arylene sulfides) were prepared from dihalogenated aromatics with sodium sulfide in a suitable polar aprotic solvent [60]. Since 1972 PPS has been produced on an industrial scale according to this method by Phillips Petroleum and marketed under the name Ryton. Other industrial methods are all based on this principle:

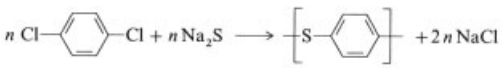

Some new synthetic pathways for producing PPS have been published, but have not yet resulted in industrial processes [61].

5.1. Production

The formation reaction of poly(arylene sulfides) is generally described as nucleophilic substitution of an aromatic compound, with elimination of alkali chloride.

Although the reaction appears to be extremely simple, it is a complex system of partial reactions. Synthesis is difficult to control due to extreme differences in reactivity in the course of the polycondensation [62]. Secondary reactions (in some cases radical-induced) lead to functionality losses, as well as production of undesirable byproducts and cyclic oligomers [63]. Industrial production is possible only if all the reaction parameters are strictly defined and controlled.

In the first process developed by Phillips Petroleum (see Section 5) sodium sulfide hydrate is dissolved in *N*-methylpyrrolidone. In a development of this process the sulfur component is incorporated in a two-stage reaction. In the first stage sodium hydroxide is reacted with the solvent *N*-methylpyrrolidone and in the second stage hydrogen sulfide gas is introduced [64]. Some of the water can be removed by distillation before adding 1,4-dichlorobenzene. The reaction is brought to completion under increasing pressure at temperatures up to 290 °C [60]. Modifications of this process include metering [65], specific temperature controls [66], and continuous mode of operation [67].

The reaction product is a suspension that contains PPS and sodium chloride as the solid phase. The suspension can be freed from volatile substances such as water, *N*-methylpyrrolidone, and dichlorobenzene by, for example, flash evaporation [68]. The resulting powder consists of polymer and salt. It is washed with water at various pressures and temperatures to remove the salt and then dried [69].

The procedure described above only produce low molecular mass PPS which is unsuitable for further processing (e.g., by extrusion or injection molding). In order to achieve the necessary molecular mass, the PPS powder is cured. The melt flow index (MFI) can be reduced from about 6000 g/10 min to 50–700 g/10 min (ASTM D 1238-70, 316 °C, 5 kg) by chain extension and branching after thermooxidation with an oxidizing agent (in the simplest form with air) [70].

Higher molecular mass PPS can be obtained by using catalysts (e.g., sodium or lithium acetate) and possibly branching agents (trifunctional halogenated aromatics) [71], and by establishing a defined water content [72].

The high molecular mass PPS forms a separate molten phase, up to 15 % of the total PPS remains as a low molecular mass fraction in the solvent phase. This latter fraction has to be separated by screening [73].

In a further development of this process, instead of using catalysts water is selectively added to the low molecular mass PPS solution which then undergoes complete phase separation and further condensation in the separated liquid phase. High molecular mass, linear polymers can thus be obtained [74].

In another variant, high molecular mass PPS is produced directly in a solvent (e.g., *N*-methylcaprolactam), in the absence of water. Any water contained in the reactants is subsequently removed by azeotropic distillation with the dihalogenated aromatic reactant. In this way the reaction can be brought to completion in a very controlled manner under simple reflux dis-

tillation. This procedure can be carried out as a continuous process [75].

The use of special catalysts (e.g., amino acids) and new workup methods result in a further improvement in product properties [76].

5.2. Properties

5.2.1. Material Properties

The properties of PPS may vary to a limited extent, depending on the production process and workup.

Various methods have been described for characterizing the melt behavior and estimating the molecular mass of PPS. A commonly employed method is the measurement of a melt index:

(melt flow index (MFI) in g/10 min, or melt volume index (MVI) in mL/10 min). In many publications the ASTM D 1238-70 standard with a modified application force (5 kg) and melt temperature (316 °C) is recommended [70–72]. The method determines only one parameter, however, as a single-point measurement on a shear-dependent melt curve. Furthermore, the standard does not permit measurement of higher values (i.e., high flowability). Measurement of the melt viscosity as a function of the shear force in a cone – plate or capillary viscosimeter is more useful [77].

Solution viscosity and light scattering measurements in combination with size-exclusion chromatography (gel permeation chromatography) have also been used to determine molecular masses [78]. Industrially produced poly(arylene sulfides) generally have broad distribution curves, which become wider after curing (see above). For further details of measurement of molecular mass, see → Plastics, Analysis.

PPS can be characterized by IR or nuclear magnetic resonance spectroscopy [79]. The thermal degradation of PPS has been investigated by direct pyrolysis in the ionization source with the aid of mass spectroscopy. In the presence of oxygen, aromatic ether bonds are formed by oxidation, and condensed aromatic compounds are produced by elimination of hydrogen sulfide. Sulfur dioxide is another inorganic gaseous product [80].

Surface oxidation produces sulfoxide and sulfone groups that can be detected by electron spectroscopy for chemical analysis (ESCA) methods [81]. Thermogravimetric measurements have shown that the high thermal stability of PPS compared to other plastics is due to cross-linking reactions [82].

Chloroaryl, thiol, or thiolate groups may be expected as end groups for poly(arylene sulfides) with a regular structure. The chlorine end group content can be determined by elemental trace analysis (e.g., Wickbold combustion) [83]. Determination of thiol groups is more difficult, a wet chemical method is described in [84].

The content of ionic impurities may be critical in certain applications, for example in the electronics sector. The polymer must be brought into solution to release and determine its salt content [85]. A physical method for determining chloride ions has also been published [86].

5.2.2. Properties of Nonreinforced Moldings

PPS is a partially crystalline material. Moldings thus consist of an amorphous phase and a crystalline phase. The amorphous phase has a glass transition temperature of 90 °C. The thermodynamic melting point of the crystalline phase for "ideal" crystals is between 300 and 310 °C [87]. The platelet-shaped, orthorhombic crystals (length and width $\leq$ 1000 nm, thickness $\leq$ 10 nm) in moldings have a melting point between 280 and 285 °C. The PPS crystallites form spherical spherulites from still melts, and stacks with a hard-fiber morphology from flowing melts [88].

During injection molding or extrusion, PPS crystallizes when the melt is cooled. Differential scanning calorimetry (DSC) shows that crystallization begins at about 250 °C. The final degree of crystallinity depends on whether the melt is cooled above or below the crystallization temperature of ca. 127 °C, and how quickly this takes place.

It is often impossible to obtain the maximum degree of crystallization during injection molding. Such moldings may become distorted due to post-shrinkage during subsequent use under thermal stress. In order to ensure that PPS moldings are dimensionally stable during subsequent use, it may therefore be necessary to

Table 6. Mechanical properties of nonreinforced PPS

Property	Standard	Melt viscosity ∗, Pa · s		
		14	66	168
Impact strength, kJ/m^2	ISO 180/1 C	14	26	34
Tensile strength, MPa	DIN 53 455	40	67	59
Elongation at break, %	DIN 53 455	1.1	2.3	10.9
Modulus of elasticity (tension), MPa	DIN 53 457	3940	3560	3460

∗ Measured in a rotation viscosimeter at 306 °C and a shear rate of 100 s^{-1} (measurements made at Bayer).

temper them beforehand (i.e., to store them at elevated temperature, e.g., 220 °C) to complete their crystallization under controlled conditions.

Injection moldings of nonreinforced PPS have high tensile and flexural strengths, as well as high moduli of elasticity (Table 6). They have a high rigidity and a low flexibility, with the result that they may fracture even under very small loads. The impact strength and notched impact strength are also extremely low. Whereas moldings of low-viscosity PPS exhibit brittle fracture even under slight deformation, moldings of high-viscosity PPS can be deformed before they fracture in a ductile manner [89].

Table 7. Thermal and electrical properties of nonreinforced PPS (Ryton R 6)

	Standard	Value
Thermal properties		
Vicat softening temperature (VST/B$_{120}$), °C	DIN 53 660	≥ 240
Heat deflection temperature (HDTA), °C	DIN 53 641	117
Linear coefficient of expansion, K^{-1}	DIN 53 752	
at 20 – 90 °C		60×10^{-6}
at 100 – 180 °C		110×10^{-6}
Electrical properties		
Dielectric strength, kV/mm	ASTM D 149-64	15
Volume resistivity, Ω · cm	ASTM D 257-66	4.5×10^{16}
Dielectric constant	ASTM D 150-70	
at 1 kHz		3.1
at 1 MHz		3.1
Dissipation factor	ASTM D 150-70	
at 1 kHz		0.0005
at 1 MHz		0.0009

Thermal and electrical properties of nonreinforced PPS are listed in Table 7. Moldings of nonreinforced PPS that are not mechanically stressed retain their shape up to the melting temperature. Under mechanical loads, however, they become deformed above ca. 90 °C, i.e., the glass transition temperature. For a partially crystalline thermoplastic the coefficient of linear thermal expansion is extremely low. The thermal conductivity of PPS is also very low [90].

On the basis of its dielectric values and resistance values PPS is a slightly polar insulator.

PPS is an inherently flame-resistant thermoplastics (oxygen index 0.44). It is classified as a V-0 material (UL 94) at a thickness of 0.8 mm (→ Flame Retardants, Chap. 3.) [90].

At low temperature PPS is resistant to all organic solvents and chemicals. PPS moldings swell, however, under the action of some organic compounds on heating, particularly after prolonged exposure (Table 8). PPS is extremely resistant to acids and alkalis but oxidizing agents, in particular nitric acid, attack PPS on heating and destroy the moldings. PPS is moderately permeable to diffusion of gases and water vapor [91].

Table 8. Chemical resistance of PPS on storage at 93 °C

	Weight increase after 24 h, %	Tensile strength after 24 h, MPa	Tensile strength after 3 months, MPa
	0	74	74
Gasoline benzene	0.07	71	64
Trichloroethylene	6.5	51	
Butanol	0.05	72	74
Butanone	1.02	77	52
Dibutyl ether	0	79	74
Amyl acetate	0.14	92	74
Butylamine	1.52	49	0
Hydrochloric acid (37 %)	0.57	75	41
Sulfuric acid (30 %)	0.14	81	74
Nitric acid (10 %)	0.32	83	23
Sodium hydroxide (30 %)	0.07	69	74

5.2.3. Properties of Reinforced Moldings

Reinforcement of PPS with fibers and/or mineral fillers increases its mechanical strength and thereby compensates for its low impact strength,

and also improves its dimensional stability at high temperature. Fibrous reinforcing materials are usually made of glass but carbon fibers and aramid fibers are also employed on a small scale. Mineral fillers are generally used in combination with glass fibers and include calcium carbonate, kaolin, calcium sulfate, mica, talc, and quartz. These fillers and reinforcing materials can be incorporated into PPS in high concentrations (up to about 70 wt %) by using conventional compounding equipment [91–94].

Injection moldings made of glass-fiber-reinforced PPS have a high hardness, ultimate tensile strength, flexural strength, and modulus of elasticity, but a very low formability and ductility and only moderate impact strength and notched impact strength. Such moldings can permanently withstand high vibrational stress. The slip properties and abrasion resistance of glass-fiber-reinforced PPS molding compounds are inferior to those of other partially crystalline thermoplastics reinforced with glass fibers [93].

Glass fiber – mineral-reinforced PPS injection molding compounds have relatively high hardness and rigidity, and relatively low impact strength and percentage elongation at break (Table 9). The properties may be altered within wide limits by varying the weight ratio of glass fibers to mineral fillers [91, 92].

Reinforcing PPS with mineral fillers also considerably improves its heat resistance, hardness, and rigidity (Table 10). It also results in moldings with very smooth surfaces and high dimensional stability. Omission of fibrous reinforcement means, however, that the flexural and tensile strengths of the moldings are very low: they can be deformed only very slightly before they fracture [91, 92].

Nonreinforced PPS exhibits partial softening at 90 °C and moldings cannot be mechanically stressed above this temperature. Reinforcement with glass fibers or minerals improves the strength to such an extent that moldings can withstand mechanical stress at elevated temperature. This is reflected in the heat resistance values according to ISO 75, i.e., the heat deflection temperature (HDT), which consistently lies between 260 and 270 °C [91, 92].

Maximum service temperature is characterized by relative temperature indices which are the temperatures at which 50 % loss of initial mechanical properties occurs after an exposure of 25 000 h. The relative temperature indices of reinforced and filled PPS molding materials are 200 – 240 °C. The thermal conductivity of reinforced PPS injection molding compounds is primarily determined by their glass fiber content and the fiber orientation in the molding. In moldings fabricated from materials reinforced with glass fibers and mineral fillers, the anisotropy of the thermal expansion is less. Reinforced PPS injection molding compounds are recommended as excellent electrical insulators on account of their high volume resistivity and surface resistivity (Table 11) [91, 92].

Table 11. Electrical properties of reinforced PPS molding compounds (measurements made at Bayer)

Property	KU 1-9511	KU 1-9521
Specific volume resistance, Ω · cm (VDE 0303, part 1)		
4 d, 23 °C, 50 % R.H.	7.2×10^{15}	2.5×10^{15}
4 d, 23 °C, 92 % R.H.	6.7×10^{14}	9.2×10^{14}
Surface resistivity, Ω · cm (VDE 0303, part 3)		
4 d, 23 °C, 50 % R.H.	6.4×10^{15}	2.7×10^{15}
4 d, 23 °C, 92 % R.H.	9.6×10^{13}	3.4×10^{13}

Reinforced PPS molding materials are extremely flame-resistant, with a very low fume density. In addition they pass the Ohio State University (OSU) test [91, 92].

Moldings are extremely resistant to most organic compounds, even at elevated temperatures. However, on prolonged storage at elevated temperature in water or in water vapor they lose up to 60 % of their initial strength [94]. Water absorption is extremely low, even in boiling water, and accordingly dimensional changes due to water absorption do not occur. Under prolonged storage in hot air, above ca. 120 °C PPS injection moldings turn brown. However, this is not associated with reduction of mechanical strength (maximum use temperature: 200 – 240 °C) [91, 92].

PPS undergoes photooxidation in the light, and prolonged weathering results in yellowing and moldings with a dull, rough surface. Their mechanical strength decreases only very slowly however. PPS is resistant to high-energy radiation [91, 92].

Repeated processing of PPS injection molding compounds does not result in degradation of the PPS macromolecules, as is confirmed by the constant melt viscosity. Recycling of regen-

Table 9. Mechanical properties of PPS molding compounds reinforced with 40 wt % glass fiber

Property	Standard	Craston XMB 3100 (Ciba-Geigy)	Fortron 1140 A 4 (Hoechst Celanese)	Ryton R 4 (Phillips Petroleum)	Supec G 401 (GE Plastics)	Tedur KU 1-9510-1 (Bayer)
Tensile strength, MPa	DIN 53 455	175	159	116	170	190
Elongation at break, %	DIN 53 455	1.6	1.4	0.9	1.0	1.6
Modulus of elasticity (tension), GPa	DIN 53 457	17.5	14.2		14.5	15.0
Flexural strength, MPa	DIN 53 452	250	240	190	240	300
Modulus of elasticity (bending), GPa	DIN 53 457	15.5	12.4	13.0	14.0	13.0
Impact strength (Charpy), kJ/m^2	DIN 53 453	20	29	11	25	35
Notched tensile impact strength (Charpy), kJ/m^2	DIN 53 453	8	9	6	8	9
Ball indentation hardness, MPa	DIN 53 456	300	305	300	300	300

Table 10. Mechanical properties of PPS molding compounds reinforced with glass fibers and mineral fillers

Property	Standard	Craston X MB 3110	Fortron 6165 B 4	Ryton R 10	Tedur KU 1-9521
Density, g/cm^3	ASTM D 1505-68	1.93	2.1	2.0	1.9
Glass fiber/mineral content, wt %		60	65	70	60
Tensile strength, MPa	DIN 53 455	110	111	86	110
Elongation, %	DIN 53 455	1	1	0.7	1
Modulus of elasticity (tension), GPa	DIN 53 457	21.5	18.2	18.0	20
Flexural strength, MPa	DIN 53 452	165	190	140	180
Modulus of elasticity (bending), GPa	DIN 53 453	20.0	16.7	15.0	19
Impact strength (Charpy), kJ/m^2	DIN 53 453	11	19	6	10
Notched tensile impact strength (Charpy), kJ/m^2	DIN 53 453	6	7	4	6
Ball indentation hardness, MPa	DIN 53 456	370	430	350	300

erated material from sprues and defective parts is therefore acceptable [91, 92].

5.3. Processing

5.3.1. Injection Molding [95]

PPS injection molding materials containing glass fibers and/or minerals have an extremely good processibility on account of their low melt viscosity.

Injection Conditions. It has proved advantageous to store the molding powder for 3 – 6 h at ca. 150 °C prior to injection molding to remove any volatile constituents. Materials that contain hydrophilic fillers should always be dried beforehand under these conditions.

PPS injection molding compounds are processed at melt temperatures of 350 – 370 °C, the most common range is 320 – 340 °C. PPS only crystallizes extremely slowly and incompletely below 120 – 130 °C, the temperature of the molds must therefore be above 130 °C and, if possible, between 150 and 170 °C to obtain moldings with smooth, uniform surfaces.

Solidification during filling results in surface damage and weak spots on the moldings. The injection rate should therefore be as high as possible and is usually performed at 75 – 150 MPa. High injection pressures are advantageous since they ensure uniform moldings, good surfaces, and favorable mechanical properties. Overinjection can easily occur however, with the formation of mold marks. After-pressures of 50 – 70 % of the injection pressure are normally employed.

PPS molding compounds can be processed in all conventional screw-injection molding devices, triple-zone screw arrangements with compression ratios of 1 : 2 to 1 : 3 are often used. PPS materials generally contain glass fibers or mineral fillers and thus, wear down screws and cylinders, particularly in the feed section. Reinforced (armor-plated) screws and cylinders are therefore recommended, with shutoff nozzles instead of open nozzles. Heated channel nozzles may also be used.

Molds. Since the PPS melt has a low viscosity, it may penetrate between the mold parting

surfaces, and thus cause formation of burrs and marks. High clamping pressures are used to prevent this.

PPS moldings reinforced with glass fibers shrink anisotropically. Nonwarping moldings can, however, be obtained by modifying the mold shape, injection rate, mold temperature, and the temperature of the melt. Shrinkage is also lower in PPS materials that contain both mineral fillers and glass fibers. Such materials are consequently better suited to the production of nonwarping, dimensionally stable moldings.

The high content of glass fibers and fillers in most PPS injection molding compounds can lead to considerable abrasion of the molds, and wear-resistant mold steels are therefore used. Steels with Rockwell hardnesses of 59 HRC to 62 HRC are preferred.

The mold cavities should be vented sufficiently rapidly, normally through venting slits in the mold parting surfaces. The slits should not be wider than 0.01 mm to avoid the formation of burrs and marks on the moldings.

5.3.2. Other Molding Methods

PPS moldings can be produced not only by injection molding (Section 5.3.1), but also by compression or by sintering PPS powders or mixtures of PPS powders and glass fibers, fillers, and/or additives (e.g., molybdenum sulfide or polytetrafluoroethylene).

The property spectrum of PPS pressed moldings is substantially inferior to that of injection moldings. The mechanical strength of sintered parts roughly corresponds to that of pressed parts of the same composition. An advantage over compression molding is that hot (360 °C) molds do not need to be handled [96].

Low-viscosity PPS powders are used to coat metallic parts (coating from aqueous dispersions, electrostatic powder coating, fluidized-bed coating). Coatings of oxidatively cross-linked PPS are rigid, flexible, and wear-resistant, and can withstand high thermal stress over prolonged periods. The chemical resistance of PPS coatings is decisive for their use in protecting metal parts against corrosion [97].

5.3.3. Finishing of Shaped Articles

Most PPS materials contain a high concentration of glass fibers and mineral fillers. The use of hard-metal drills, milling machines, and saws to finish PPS moldings is therefore recommended.

PPS moldings can be strongly bonded to one another by ultrasonic welding, hot-plate welding, and with two-component adhesives, in particular epoxy and polyurethane adhesives.

Coatings on PPS moldings can be obtained only if a polyurethane primer is applied.

5.3.4. Extrusion

Films [98]. See also → Films, Chap. 6.12.4..

Films, particularly biaxially stretched films, constitute one of the main uses of unfilled PPS, and are mainly used in capacitors. Biaxially stretched PPS films have good mechanical properties, their outstanding property being their high thermal resistance. If ignited, they rapidly cease burning after the flame has been removed. PPS films are resistant to bases, acids, solvents, and other chemicals. However they are rapidly destroyed by concentrated sulfuric or nitric acid at room temperature.

PPS films can be coated and printed after corona or plasma pretreatment. The very low polarity of the PPS molecules accounts for the low dielectric constant and the low dielectric loss factor.

To produce biaxially stretched PPS films the filtered PPS melt is extruded at ca. 320 °C onto cooled rollers. It has also proved advantageous to remove volatile constituents by degassing the melt before the extrusion, this is achieved by releasing the pressure and applying a vacuum. The amorphous film is then stretched at 90 – 120 °C successively or simultaneously in the longitudinal and transverse directions.

Fibers [99]. See also → High-Performance Fibers, Chap. 2.3.2.

Fibers constitute the second largest area of use of unfilled PPS, and are employed particularly in the form of fabrics and nonwovens for filtering hot and/or aggressive media.

PPS fibers have a high mechanical strength, and can be used for short periods up to about 230 °C, whereby they retain 40 % of their tensile

strength at 23 °C. The longterm service temperature is 190 °C. If fabrics made from PPS are ignited, they rapidly cease burning once the source of ignition is removed.

PPS fibers have a high resistance to acids, alkalis, and most organic solvents, even at elevated temperature. Oxidizing agents attack PPS fibers however. In boiling water PPS fibers may shrink by up to 10 %.

Extrusion of the raw fibers is followed by two-stage stretching and heat setting, in which the fibers are heated under tension at 265–270 °C for 1 min to ensure satisfactory crystallization of the PPS.

5.3.5. Semifinished Products

Reinforced PPS materials are processed into continuous profiled sections, rods, pipes, etc. On account of its high resistance to heat and solvents, poly(phenylene sulfide) is also an excellent matrix for high-performance composite materials [100].

5.4. Uses

The high price of PPS injection molding materials limits their uses and applications. They are therefore mainly used where technical and economic requirements cannot be satisfied by other industrial thermoplastics.

The applications of PPS molding materials are determined by their unique combination of technical advantages, the most important being:

1) High dimensional stability even at high temperature
2) Very high rigidity and hardness
3) Very good long-term static and dynamic loadability
4) Low static and dynamic coefficients of friction
5) Extremely high resistance to almost all chemicals
6) Low flammability
7) Good processability by injection molding
8) Machinability and weldability
9) Inscribability of the moldings by printing or laser inscription

The main areas of use for PPS molding materials include the electrical/electronics, automotive, and lighting sectors. PPS moldings are also used in many other market sectors, for example in household appliances, chemical plant construction, and mechanical engineering.

In the *electronics sector*, special, highly filled PPS molding compounds with particularly low melt viscosities have to some extent replaced epoxy compositions in the housing of electronics components such as transistors, coils, and capacitors. Glass fiber-reinforced PPS is also used for integrated circuit housings (chips). Connectors (e.g., high-density connectors) made of PPS reinforced with glass fibers (and minerals) are produced on a large scale in the United States. The high mechanical strength of moldings made from glass fiber-reinforced PPS accounts for its use in components of electric motors (e.g., brush holders and wound cores).

PPS molding compounds are also being used increasingly in the *automotive sector*. For example pumps in petrol tanks are fabricated from PPS reinforced with glass fiber and minerals. In pump impellers of vacuum pumps used in steering and braking systems, the PPS moldings have to satisfy extremely stringent requirements as regards dimensional accuracy, dimensional stability, mechanical strength, dynamic longterm stress, and abrasion resistance.

In the *lighting sector* the trend towards small automobile headlamps and high light intensities means that a great deal of heat is generated in a confined space. Housings are therefore made of special, highly-filled PPS molding compounds. Caps and holders of halogen lamps are also produced from glass fiber-reinforced PPS.

PPS molding compounds are employed in *chemical plant construction* (e.g., for valves, column packings, and distillation columns). In feed pumps the danger of spark formation as a result of buildup of electrostatic charge can be excluded by using PPS molding materials whose electrical conductivity has been increased by incorporating carbon black.

On account of its resistance to chemicals, PPS is suitable for *applications involving contact with food*. Certain PPS compounds have been approved by the BGA and FDA for use as a material for kitchen utensils [101].

5.5. Economic Aspects

Despite the fact that PPS has been commercially available since the mid 1970s, it is a very "young" product. On account of its outstanding property profile it has a broad, constantly increasing range of applications. Newly developed copolymers and poly(arylene sulfides) based on dihalogenated aromatics with tailor-made property profiles are reinforcing this development [102].

PPS compounds are used in strongly expanding sectors. In parallel with growing awareness among customers, further processors are beginning to use PPS which provides a good precondition for high growth rates for PPS compounds. Average annual growth rates exceeding 11 % are expected in North America, the Far East (including Japan), and Western Europe (see Table 12) [47, 103].

Table 12. PPS turnover forecast [47]

Region	Turnover, t		Average annual
	1989	1995	growth, %
North America	5 500	10 500	11.4
Far East (including Japan)	4 000	7 700	11.5
Western Europe	2 500	4 800	11.5
Total	12 000	23 000	11.5

Trade Names. In the mid 1970s only Phillips Petroleum was represented on the market, with its Ryton types. Following the expiry of the basic patents other companies are becoming increasingly active both as regards resin production and compounding [103].

Trade names of companies with their own primary resin production include Ryton (Phillips); Tedur, Tregalon (Bayer); Fortron (Kureha); Tonen PPS (Tohpren); Susteel (Tosoh Susteel); Torelina (Toray); (Sung Yong)

Trade names of companies without their own primary resin production include Supec (General Electric, United States); Tedur (Miles, United States); OFL and OCL (LNP, United States); Fortron (Hoechst Celanese, United States, Western Europe); Craston (Ciba-Geigy, Western Europe); Primef (Solvay, Western Europe), Larton (Lati; Western Europe); DIC-PPS (Dainippon Ink, Japan); Sumicon (Sumitomo Bakelite, Japan); Nova PPS (Mitsubishi Kasei, Japan); Fortron (Polyplastics, Japan); Denka PPS (Denka, Japan); Toyobo PPS (Toyobo, Japan); PPS (Ube, Japan); Petchem Idemitsu PPS (Idemitsu, Japan).

6. References

1. R. B. Seymour, G. S. Kirshenbawm: *High Performance Polymers: Their Origin and Development,* Elsevier, Amsterdam 1986.
2. J. G. Wirth, D. R. Heath, US 3 730 946, 1973.
3. T. Takekoshi, J. E. Kochanowsky, J. S. Manello, M. J. Webber, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **20** (1979) no. 1, 179.
4. T. Takekoshi, J. E. Kochanowsky, J. S. Manello, M. J. Webber, *J. Polym. Sci. Polym. Chem. Ed.* **23** (1985) 1759 – 1769.
5. T. Takekoshi, J. E. Kochanowsky, J. S. Manello, M. J. Webber, *J. Polym. Sci. Polym. Symp.* **74** (1986) 93 – 108.
6. L. R. Schmidt, E. M. Lovgren, P. G. Meissner, *Int. Polym. Proc. IX* **4** (1989) 270 – 276.
7. J. W. Verbicky in H. F. Mark, N. M. Bikales, C. G. Overberger, G. Menges (eds.): *Encyclopedia of Polymer Science and Engineering,* **vol. 12,** John Wiley, New York, 1988, 364 – 382.
8. *Kirk-Othmer,* 3rd ed., **18,** 611 – 615.
9. I. W. Serfati, *Proceedings of the First Technical Conference on Polyetherimides,* New York 1982, 149 –161.
10. H. Farong, W. Xueqiu, L. Shijin, *Polym. Degrad. Stab.* **18** (1987) 247 – 259.
11. R. O. Johson, E. O. Teutsch, *Polym. Compos.* **4** (1983) 162.
12. S. A. White, S. R. Weissman, R. P. Kambour, *J. Appl. Polym. Chem.* **27** (1982) 2675 – 2682.
13. D. E. Floryan, G. L. Nelson, *J. Fire Flammability* **11** (1980) 284.
14. H. D. Burks, T. L. St. Clair, *J. Appl. Polym. Sci.* **30** (1985) 2401 – 2411.
15. R. O. John, H. S. Burklis, *J. Polym. Sci. Polym. Symp.* **70** (1983) 129.
16. General Electric, US 4 959 121, 1990 (W. V. Dumas, D. F. Foust).
17. J. E. Harris, L. M. Robeson, *J. Appl. Polym. Sci.* **35** (1988) 1877.
18. C. A. Arnold et al., *Polymer. Prepr.* **28** (1987) 217; **29** (1988) 349.
19. Du Pont, US 3 065 205, 1962 (W. H. Bonner).
20. ICI, GB 971 227, 1964 (I. E. Goodman et al.).
21. Y. Iwakura et al., *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 3345.

22. Du Pont, US 3 441 538, 1969 (B. M. Marks).
23. R. N. Johnson et al., *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 2375.
24. ICI, GB 1 414 421 and 1 414 422, 1975 (J. B. Rose, P. A. Staniland).
25. T. E. Attwood et al., *Polymer* **22** (1981) 1096.
26. ICI, EP 0 001 879, 1977 (J. B. Rose and P. A. Staniland).
27. Raychem, US 3 953 400, 1976 (K. J. Dahl).
28. P. C. Dawson, D. J. Blundell, *Polymer* **21** (1980) 577.
29. E. J. Stober et al., *Polymer* **25** (1984) 1845.
30. J. N. Hay, D. J. Kemmish, *Polymer* **29** (1988) 613.
31. J. Devaux et al., *Polymer* **26** (1985) 1994.
32. ICI Advanced Materials, Victrex PEEK, Lit. Ref. VK 2/0988, Wilton.
33. X. Jin et al., *Br. Polym. J.* **17** (1985) 4.
34. D. J. Blundell, B. N. Osborn, *Polymer* **24** (1983) 953.
35. R. B. Prime, J. C. Seferis, *J. Polym. Sci. Polym. Symp.* **24** (1986) 641.
36. ICI Advanced Materials, Victrex PEEK, Lit. Ref. VK 3/0988, Wilton 1988.
37. G. R. Belben, *I Mech. E Proc.* **198** (1984) 47.
38. 3 M, GB 1 060 546, 1963 (H. A. Vogel).
39. J. B. Rose, *Polymer* **15** (1974) 456.
40. ICI, GB 1 153 035, 1965 (D. A. Barr, J. B. Rose).
41. R. N. Johnson et al., *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 2375.
42. Union Carbide, GB 1 078 234, 1963 (A. G. Farnham, R. N. Johnson).
43. T. E. Attwood et al., *Polymer* **18** (1977) 259.
44. Union Carbide, Udel Polysulphone: An Outstanding Engineering Polymer for Moulding and Extrusion, 1978.
45. ICI Advanced Materials, Victrex PES: Chemical Resistance Data, Lit. Ref. VS 7/0689, Wilton 1989.
46. ICI Advanced Materials, Victrex PES: Properties and Processing, Lit. Ref. VS 10/1089, Wilton 1989.
47. Nippon Chemtec Consulting Inc.: "PPS & Competitive Resins," *Evaluation of High Performance Engineering Resins,* part 2, R & D Evaluation Rep. Ser. No. 50, Osaka 1990.
48. C. Friedel, J. M. Crafts, *Ann. Chim. Phys.* **14** (1888) 433.
49. J. W. Cleary, *Polym. Sci. Technol. (Plenum)* **31** (1985) 159.
50. P. Genvresse, *Bull. Soc. Chim.* **17** (1897) 599.
51. J. W. Cleary, *Polym. Sci. Technol. (Plenum)* **31** (1985) 173.
52. A. D. Macallum, *J. Org. Chem.* **13** (1948) 154;US 2 538 941, 1951.
53. R. W. Lenz, C. E. Handlovits, *J. Polym. Sci.* **43** (1960) 167.
54. A. Kreuchunas, US 2 822 351, 1955.
55. B. Hortling, M. Söder, J. J. Lindberg, *Angew. Makromol. Chem.* **107** (1982) 163.
56. Inventa, DE-AS 2 009 323, 1969.
57. I. Haddad, S. Harley, C. S. Marvel, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 2793; **12** (1974) 1301.
58. H. A. Smith. "Poly(Arylene Sulfides)," *Encycl. Polym. Sci. Technol.* **10** (1969) 653.
59. R. W. Lenz, C. E. Handlovits, H. A. Smith, *J. Polym. Sci.* **58** (1962) 351. Dow Chemical, US 3 274 165, 1965.
60. Phillips Petroleum, US 3 354 129, 1963.
61. E. Tsuchida, K. Yamamoto, H. Nishida, S. Yoshida, *Macromolecules* **20** (1987) 2030. Kodak, EP-A 395 720, 1990. Bayer, DE-OS 3 738 276, 1989;DE-OS 3 738 277, 1989.
62. W. Koch: "Modelluntersuchungen zur Kinetik und zum Mechanismus der Bildung von Poly(thio-1,4-phenylen)," Dissertation, Universität Marburg 1983, p. 76 ff. J. Koschinski, K. H. Reichert, Makromolekulares Kolloquium, Freiburg 1985.
63. W. Koch: "Modelluntersuchungen zur Kinetik und zum Mechanismus der Bildung von Poly(thio-1,4-Phenylen)," Dissertation, Universität Marburg 1983, p. 48 ff. E., Heitz, W. Koch, *Makromol. Chem.* **184** (1983) 779. V. A. Sergeev, V. K. Shitikov, V. J. Nedelkin, *Usp. Khim.* **47** (1978) 2065. V. A. Sergeev, V. J. Nedelkin, *Vysokomol. Soedin. Ser. B* **20** (1978) 395; *Acta Polym.* **33** (1982) 647. R. A. Rossi, R. A. Alonso, *J. Org. Chem.* **45** (1980) 1239. V. Z. Annenkova, et al. *Zh. Obsk. Khim.* **53** (1983) 2409. M. L. Kaplan, W. D. Reents, Jr., *Tetrahedron Lett.* **23** (1982) 373.
64. Phillips Petroleum, EP 39 508, 1980.
65. Phillips Petroleum, US 3 790 536, 1972.
66. Phillips Petroleum, US 3 756 993, 1970.
67. Phillips Petroleum, US 4 060 520, 1971.
68. Phillips Petroleum, US 3 707 528, 1971.
69. Phillips Petroleum, US 4 526 684, 1985; EP-A 103 261, 1984.
70. Phillips Petroleum, US 4 383 080, 1982; R. T. Hawkins, *Macromolecules* **9** (1976) 189.
71. Phillips Petroleum, US 3 919 177, 1975; US 4 038 260, 1977; US 4 038 261, 1977.
72. Phillips Petroleum, DE 3 030 488, 1979.
73. Phillips Petroleum, US 4 415 729, 1982; Dainippon, DE-OS 30 527 492, 1984.

74. Phillips Petroleum, US 4 524 200, 1985; Kureha, EP-A 166 368, 1984; EP-A 256 757, 1986, EP-A 259 984, 1986.
75. Bayer, DE-OS 3 318 401, 1983; DE-OS 3 338 501, 1983; DE-OS 3 243 189, 1983.
76. Bayer, EP-A 142 024, 1985; DE-OS 3 339 233, 1985; EP-A 162 210, 1985; EP-A 171 021, 1986; EP-A 175 968,; DE-OS 3 713 669, 1988; DE-OS 3 723 071, 1989.
77. J. Schurz: *Viskositätsmessung an Hochpolymeren,* Berliner Union Verlag, Stuttgart 1972; Struktur-Rheologie, Berliner Union Verlag, Stuttgart 1974.
78. C. J. Stacy, *J. Appl. Polym. Sci.* **32** (1986) 3959. T. Housaki, K. Satoh, *Polym. J. (Tokyo)* **20** (1988) 1163 Bayer DE-OS 3 909 599, 1990.
79. D. O. Hummel: *Atlas der Polymer- und Kunststoffanalyse* 2nd ed., Hanser Verlag, München 1978. C. E. Brown, I. Khoury, M. D. Bezoari, J. Kovacic. *J. Polym. Sci. Polym. Chem. Ed.* **20** (1982) 1697.
80. G. Montaudo, M. Przybilski, H. Ringsdorf, *Makromol. Chem.* **176** (1975) 1763. G. Montaudo, C. Puglisi, E. Scamporrino, D. Vitalini, *Macromolecules* **19** (1986) 2157. F. Quella, *Kunststoffe* **71** (1981) 386.
81. A. Kaul, *Polym. Prep. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) 229.
82. P. A. Lowell, R. H. Still, *Polym. Degrad. Stab.* **18** (1987) 33. A. Kinugawa, *Kobunshi Ronbunshu* **44** (1987) 139.
83. F. Ehrenberger, S. Gorbach: *Methoden der organischen Elementar- und Spurenanalyse,* Verlag Chemie, Weinheim 1973. T. S. Ma, R. C. Rittner: *Modern Organic Elemental Analysis,* Marcel Dekker, New York 1979.
84. Kureha, EP-A 238 193, 1986.
85. Bayer, DE-OS 3 828 056, 1990.
86. Toyo Soda, Hodogaya, EP-A 225 471, 1987.
87. L. C. Lopez, G. L. Wilkes, *Polymer* **30** (1989) 147.
88. B. J. Tabor, E. P. Magré, J. Boon, *Eur. Polym. J.* **7** (1971) 7. J. Garbarczyk, *Polym. Commun.* **27** 81986) 335.
89. Phillips Petroleum, Ryton PPS-Information no. 2, Bartlesville 1986. Celanese Eng. Resins, Fortron Polyphenylene Sulfide Introduction, Chatham 1987. R. S. Shue, *Plastics Eng.* **4** (1983) 37.
90. Phillips Petroleum; Ryton PPS, Polyphenylene Sulfide Resins, Bartlesville 1986.
91. Phillips Petroleum, Ryton PPS Tesins, Technical Service Memorandum, TSM-275, Bartlesville 1976; TSM-266, Bartlesville 1983.
92. Ciba Geigy, Craston Polyphenylene Sulfide Moulding Compounds, Cambridge 1988. Celanese Eng. Resins, Fortron Polyphenylene Sulfide (PPS), Chatham 1987. Susteel PPS, *Plas. Ind. News* **33** (1987) 4. Bayer, Polyphenylensulfid Tedur, Leverkusen 1990.
93. LNP Eng. Plastics, Selbstschmierende verstärkte Thermoplaste, LNP-Bulletin 254-983, Ramsdonkveer 1983.
94. *Mod. Plast. Int.* **5** (1974) 86. M. W. Woods, *1st Int. Ryder Conf. Special Performance Plast. Markets* **2** (1987).
95. M. W. Hill, *Polym. Eng. Sci.* **16** (1976) 831. C. W. Osborn, *Plas. Eng.* **33** (1977) 25. Phillips Petroleum, Ryton PPS, Spritzgießverarbeitung, Bartlesville 1979.
96. Phillips Petroleum, Plastics Techn. Center Report SSL-243, Bartlesville 1974. Phillips Petroleum, Ryton PPS Resins, Technical Service Memorandum, TSM-260, Bartlesville 1974, 1976.
97. E. Heitz, *Gummi, Asbest, Kunstst.* **28** (1975) 578. W. D. Powell, *31st Ann. Conf. Exhibition Soc. Plast. Ind. Canada,* Vancouver 1973. Phillips Petroleum, Ryton PPS Resins, Technical Service Memorandum, TSM-275, Bartlesville 1976, TSM-278, Bartlesville 1977.
98. Toray, PPS-Film Torelina, Tokyo 1988. M. Ito, *J. Polym. Sci. Polym. Chem. Ed.* **23** (1985) 245.
99. Phillips Petroleum, Preliminary Bulletin, Bartlesville 1983.
100. F. N. Cogswell, *Int. Polym. Process.* **1** (1987) 157. H. Gupta, *Gummi, Fasern, Kunststoffe* **39** (1986) 288. J. B. Cattanach, G. Guff, F. N. Cogswell, *J. Polym. Eng.* **6** (1986) 348. W. Soll, T. G. Gutowski, *SAMPE J.* **24** (1988) 15. H. Heißler, D. Muser, H. Wurtinger: *Kunststoffeinsatz im Automobilbau,* VDI-Verlag, Düsseldorf 1984, pp. 167 – 177.
101. FDA Regulation 177.2490, Title 21, Code of Federal Regulations.
102. D. G. Brady, J. F. Geibel, *ACS-Meeting, Division Polym. Chem.,* Reno 1987, p. 7. M. Wood, J. Geibel, V. Vives: *ANTEC* 1988, p. 1673.
103. A. S. Wood, *Mod. Plast. Int.* 1988 April, 34.

Polymethacrylamide → Polyacrylamides and Poly(Acrylic Acids)

Polymethacrylates

For poly(methacrylic acid), see Polyacrylamides and Poly(Acrylic Acids). General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Manfred Stickler, Röhm GmbH, Darmstadt, Federal Republic of Germany

Thoma Rhein, Röhm GmbH, Darmstadt, Federal Republic of Germany

1. Introduction 1
2. Properties 2
3. Raw Materials 4
4. Production Processes 5
4.1. Bulk Polymerization 6
4.1.1. Casting of Acrylic Glass 6
4.1.2. Continuous Production of Cast Acrylic Glass 7
4.1.3. Tube Polymerization 7
4.2. Suspension Polymerization 7
4.3. Continuous Polymerization of Molding Compounds 8
4.4. Batchwise Solution Polymerization 8
4.5. Emulsion Polymerization 8
5. Uses 9
5.1. Molding Compounds 9
5.1.1. Standard Molding Compounds 9
5.1.2. Heat-Resistant Molding Compounds . 9
5.1.3. Impact-Resistant Molding Compounds 10
5.1.4. Molding Compounds for Special Optical Uses 11
5.2. Semifinished Products 11
5.2.1. Extruded Acrylic Glass 11
5.2.2. Cast Acrylic Glass 11
5.3. Binders and Paints 12
5.4. Additives for the Petroleum Industry 12
5.5. Miscellaneous Uses 12
6. Toxicology 13
7. Environmental Aspects 13
8. References 13

1. Introduction

Poly(methyl methacrylate) [*9011-14-7*] (PMMA) has been produced industrially since the early 1930s. The history and development of this polymer are very closely associated with the chemist Otto Röhm [1,2]. In 1932 Röhm and his collaborator W. Bauer were the first to polymerize methyl methacrylate (MMA) into transparent sheets [3]. Almost parallel to these developments at Röhm & Haas in Germany, corresponding activities were in progress at Rohm & Haas and Du Pont in the United States and at ICI in England. A few years later the first thermoplastically processable molding compounds based on PMMA were available.

PMMA occupies an intermediate position between "commodities" and "high-tech polymers" or "engineering plastics"; its market has in the meantime developed greatly. In 1988 the MMA capacities worldwide were ca. 1.5×10^6 t/a (Western Europe 4.7×10^5 t/a, North America 5.1×10^5 t/a, Japan 3.5×10^5 t/a [4], in 1992, 1.9×10^6 t/a [5]). About 50 % of this amount is used for further processing into acrylic glass and PMMA molding compounds; the remainder being distributed among dispersions, paints, binders, adhesives, additives, etc.

PMMA is the main representative of this polymer class; polymers and copolymers of higher alkyl methacrylates are of minor economic importance.

2. Properties

The properties that make PMMA a highly-prized plastic are its good mechanical strength, its outstanding optical properties (clarity, brilliance, transparency), and its extremely good weather resistance [6]. This favorable profile can be supplemented by the surface modification of semifinished products (e.g., scratch-resistant or

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_473

antistatic coating), coloring or pigmenting, and treatment with flame retardants to open up a wide variety of potential uses for this polymer (see Chap. 5).

Table 1. Physical properties of PMMA (standard values at 23 °C and 50 % R.H. for Plexiglas GS 233 and Plexiglas XT 7H)

Property	Cast (extruded)∗	Test procedure
Density, g/cm^3	1.18	DIN 53 479
Charpy impact strength, kJ/m^2	15	ISO 179/1 D
Izod-test value, kJ/m^2	1.6	ISO 180/1 A
Tensile strength (1/1 test bar 3, v=5 mm/min), MPa	80 (72)	DIN 53 455
Elongation at break (1/1 test bar 3, v=5 mm/min), %	5.5 (4.5)	DIN 53 455
Flexural strength (standard bar 80×10×4 mm), MPa	115 (105)	DIN 53 452
Long-term tensile strength (1/1 test bar 3, t=10 000 h), MPa	38 (30)	DIN 53 444
Modulus of elasticity (short-term value), MPa	3300	DIN 53 457
Shear modulus (ca. 10 Hz), MPa	1700	DIN 53 445
Ball indentation hardness $H_{961/30}$, MPa	200 (190)	DIN-ISO 2039-1
Degree of transmission (3 mm thickness, illumination D 65, visible region, λ=380 – 780 nm), %	92	DIN 5036
Refractive index n_D^{20}	1.491	DIN 53 491
Linear thermal expansion coefficient (0 – 50 °C), K^{-1}	70×10^{-6}	DIN 53 752-A
Shrinkage onset temperature, °C	> 80	
Vicat softening point (method B), °C	115 (102)	DIN-ISO 306
Heat distortion temperature (at flexural stress of 1.8 MPa), °C	105 (90)	ISO 75
Water absorption (compared with dry state after 24 h, specimen 50×50×4 mm), mg	30	DIN 53 495

∗ Extruded from molding compound.

Pure PMMA is an amorphous plastic with a high surface gloss, high brilliance, and crystal-clear transparency. It is classified as a hard, rigid, but brittle material. The tensile, compressive, and flexural strengths are considered satisfactory; scratch resistance is also good and can be improved further by the use of special coatings. The most important physical properties of PMMA are summarized in Table 1; a distinction is made between cast PMMA ($M_r>10^6$) and PMMA extruded from molding compounds ($M_r<5\times10^5$). The properties listed in Table 1 were determined according to the so- called basic value table [7]; further data can be found in the CAMPUS data bank [8]. This data bank also includes information about copolymers of MMA with small proportions of acrylate esters (< 10 % methyl acrylate). Incorporation of acrylate allows the rheological properties of the thermoplastics to be varied within a wide range to satisfy the processing requirements during injection molding and extrusion.

PMMA is resistant to aliphatic hydrocarbons, nonpolar solvents, aqueous alkalis, and aqueous acids. Aromatic hydrocarbons, chlorinated hydrocarbons, esters, ketones, and other polar solvents attack PMMA, causing it to dissolve or swell. Whereas alcohols and water alone are regarded as nonsolvents, their mixtures may have a solubilizing effect (cosolvency). Detailed information on solvents and precipitating agents for PMMA can be found in [9].

Most physical properties E of polymers depend on the molecular mass. For example, the viscosity of a polymer melt increases continuously with the molecular mass. Some other properties (e.g., glass transition temperature, density, refractive index) show an asymptotic behavior, which can be described by the following equation:

$$E=E_\infty+\frac{k}{\bar{M}_n} \tag{1}$$

For sufficiently high number-average molecular masses $\bar{M}_n$, a polymer-specific limiting value E_∞ is accordingly reached (typically at ca. $\bar{M}_n>50\,000$); the constant k is influenced by the terminal groups of the polymer.

Table 2 summarizes some physical properties of the most important polymethacrylates. The values were obtained at sufficiently high molecular mass (i.e., E_∞), so that the influence of the alcohol residue of the ester group can be recognized.

The glass transition temperatures T_g decrease sharply with the length of the linear, aliphatic side chain. Whereas PMMA, the first member of this homologous series, is a hard, rather rigid material, the polymers with residues larger than n-hexyl are rubber-like. The glass transition temperature decreases up to about poly(dodecyl methacrylate). With longer linear ester residues, side-chain crystallization occurs, which masks the glass transition. Polymers with linear alkyl ester residues always have a lower glass transition temperature than the branched and alicyclic isomers. Large bulky residues considerably restrict the mobility of the polymer main chain, particularly if they are joined directly to the ester group.

Table 2. Physical properties of the most important polymethacrylates

Polymer	CAS registry no.	ρ (20 °C), g/cm^3	n_D^{20}	T_g, °C	δ, (J/cm^3)$^{1/2}$	References
Poly(methyl methacrylate)	[*9011-14-7*]	1.190	1.490	125	18.6	[10, 13, 14]
Poly(ethyl methacrylate)	[*9003-42-3*]	1.119	1.485	66	18.3	[10]
Poly(*n*-propyl methacrylate)	[*25609-74-9*]	1.085	1.484	35	18.0	[10]
Poly(isopropyl methacrylate)	[*26655-94-7*]	1.033	1.4728	81		[10, 13]
Poly(*n*-butyl methacrylate)	[*9003-63-8*]	1.058	1.483	27	18.0	[10]
Poly(isobutyl methacrylate)	[*9011-15-8*]	1.045	1.477	53	17.7	[10]
Poly(*sec*-butyl methacrylate)	[*29356-88-5*]	1.052	1.480	62		[10, 11, 13]
Poly(*tert*-butyl methacrylate)	[*25189-00-8*]	1.022	1.4638	122	17.0	[10, 13]
Poly(*n*-hexyl methacrylate)	[*25087-17-6*]	1.007 *	1.481	−5	17.6	[10]
Poly(cyclohexyl methacrylate)	[*25768-50-7*]	1.100	1.5065	92		[10, 13]
Poly(*n*-octyl methacrylate)	[*25087-18-7*]	0.971 *		−20	17.2	[10]
Poly(2-ethylhexyl methacrylate)	[*25719-51-1*]			−10		[10]
Poly(*n*-decyl methacrylate)	[*29320-53-4*]			−70		[11]
Poly(dodecyl methacrylate)	[*25719-52-2*]	0.929 *	1.4740 **	−55	16.8	[10]
Poly(hexadecyl methacrylate)	[*25986-80-5*]		1.4750 **			[10]
Poly(octadecyl methacrylate)	[*25639-21-8*]			−100	16.0	[10, 14]
Poly(bornyl methacrylate)			1.5059			[10]
Poly(isobornyl methacrylate)	[*28854-39-9*]	1.060	1.500	110	16.6	[10, 14]
Poly(menthyl methacrylate)	[*30847-50-8*]	1.010	1.4890	90		[10]
Poly(phenyl methacrylate)	[*25189-01-9*]	1.210	1.5715	120		[10]
Poly(1-naphthyl methacrylate)	[*31547-85-0*]		1.6410	75		[10]
Poly(benzyl methacrylate)	[*25085-83-0*]	1.179	1.5680	54	20.3	[10]
Poly(2-hydroxyethyl methacrylate)	[*25249-16-5*]		1.5119	55		[10]
Poly(ethylthioethyl methacrylate)	[*27273-87-6*]		1.5300	−20		[14]
Poly(furfuryl methacrylate)	[*29320-19-2*]	1.2603	1.5381	60		[10, 13]
Poly(tetrahydrofurfuryl methacrylate)	[*25035-85-2*]		1.5096	60		[10]
Poly(3,3,5-trimethylcyclohexyl methacrylate)	[*75673-26-6*]		1.485	79		[10]
Poly(tricyclodecyl methacrylate)				150		[12]
Poly(2,2,3,3-tetrafluoropropyl methacrylate)	[*29991-77-3*]	1.4934	1.4215	64		[13]

* At 25 °C.
** n_D^{30}.

When discussing the data in Table 2 it has to be borne in mind that properties of a polymer, in particular the glass transition temperature, also depend on its tacticity. If a tactic polymer is regarded as a terpolymer of isotactic, syndiotactic, and heterotactic triads, then its glass transition temperature can be described by the following linear combination.

$$T_g = x_{ii}T_{g,i} + x_{is}T_{g,h} + x_{ss}T_{g,s} \quad (2)$$

where $T_{g,i}$, $T_{g,h}$, and $T_{g,s}$ denote the glass transition temperatures of the (imaginary) polymers constructed purely from isotactic, heterotactic, or syndiotactic triads, respectively, and x_{ii}, x_{is}, and x_{ss} denote the associated mole fractions of these triads in the relevant real polymer. In the case of PMMA, the values $T_{g,i}$=42 °C, $T_{g,h}$ = 139 °C, and $T_{g,s}$ = 124 °C were determined by fitting experimental T_g values to x_{ii}, x_{is}, and x_{ss} mole fractions as obtained from NMR analysis [15]; data for other polymethacrylates can be found in the literature [14]. The mole fractions x_{ii}, x_{is}, and x_{ss} depend on the polymerization temperature and the polymerization mechanism (and are possibly influenced also by any solvent used in the polymerization). These relationships have been investigated in detail in the case of PMMA [16].

All the influences on the glass transition temperature described above have not been taken into account in Table 2. Most of the listed T_g values were measured on atactic polymers synthesized by radical polymerization. This explains to some extent the considerable differences that individual authors find in one and the same polymethacrylate. In addition to this, variations also arise due to the use of different methods of measurement of T_g (dilatometry, differential scanning calorimetry) and evaluation procedures (e.g., "midpoint" and "on-set" evaluation in differential scanning calorimetry).

The residual monomer content of the polymers represents a large source of errors in the determination of T_g; even relatively small propor-

tions can lead to considerable reduction of the glass transition temperature. Methacrylates with bulky side groups (e.g., isopropyl, *tert*-butyl, or cyclohexyl methacrylate) are especially critical because even in the case of a very sharp end polymerization they yield a product that still contains a few percent of residual monomers [17]. The values listed in Table 2 were obtained on reprecipitated, carefully dried material and are considerably higher than previous literature values.

Polymers used as materials for optical application require good transmission behavior; the refractive index and Abbe number are other important parameters. Polymethacrylates exhibit a considerable range of refractive indices; plastics with a particularly low refractive index have been in demand for polymer optic fibers (POF). Fluorinated polymethacrylates are particularly suitable in this connection. A summary of relevant data (T_g, n_D^{20}) is given in [18].

The influence of tacticity on the refractive index and density should also be taken into account, but up to now little information has been available.

3. Raw Materials

The starting material for PMMA is the monomer methyl methacrylate (MMA). The most important route to this monomer (as well as to methacrylic acid) is still methanolysis (or hydrolysis) of methacrylamide sulfate. The latter is produced from acetone and hydrogen cyanide via the intermediate acetone cyanohydrin. Alternative synthesis pathways have also been adopted: the C-4 route involves two-stage oxidation of isobutene (or *tert*-butanol) to methacrylic acid and MMA. In the C-3 route propene is reacted with carbon monoxide to give isobutyric acid, from which methacrylic acid is then obtained by oxidative dehydrogenation. The C-2 pathway involves condensation of propanal (or propionic acid) with formaldehyde to methacrolein (or methacrylic acid). Propanal can be obtained from ethylene by hydroformylation. For further details of the synthesis of methacrylic acid and MMA, see → Methacrylic Acid and Derivatives, Chap. 3.

MMA for polymerization is usually available with a purity of >99.9 %; impurities are mainly water, methanol, and traces of acid.

MMA is stabilized for storage or transportation with phenol derivatives (e.g., hydroquinone or hydroquinone monomethyl ether) in amounts of 10 – 100 ppm; the stabilizing action of these compounds requires the presence of oxygen. Nitrogen oxides or aromatic nitro compounds can be used in the absence of oxygen. MMA should be stored under exclusion of light and at as low a temperature as possible (<10 °C).

To obtain higher methacrylate esters, methacrylic acid is esterified with the relevant alcohols. Longer-chain methacrylate esters are also synthesized by transesterification of MMA.

4. Production Processes

Methyl methacrylate can be polymerized by a free radical mechanism [19, 20] or an anionic mechanism [21–23]. Anionic polymerization is, however, of no industrial importance because the monomer has to be subjected to extensive purification and polymerization has to be performed at low temperatures. The group transfer polymerization technique (GTP) is mechanistically very similar to anionic polymerization, but is less complicated [24, 25]. It remains to be seen whether this polymerization technique will become industrially important.

Radical polymerization of MMA may be carried out homogeneously (bulk or solution polymerization) or heterogeneously (suspension or emulsion polymerization). It can be initiated with radiation (light, γ-radiation), heat, or radical-forming chemical agents. Spontaneous thermal initiation of MMA polymerization is extremely slow [26] and is therefore of no industrial relevance. Photoinitiation of MMA polymerization can be performed without sensitizers, but photochemically labile compounds are generally added. On exposure to light these compounds form radicals either directly (e.g., azo compounds, disulfides, derivatives of benzoin or acetophenone) or are converted into an excited state and then form radicals after abstracting a hydrogen atom from the monomer or solvent (e.g., derivatives of benzophenone or anthraquinone) [27–31]. Initiation of polymerization by γ-rays (e.g., ^{60}Co source) may be of interest if the addition of a thermal initiator is undesirable and a conventional light source cannot be used because the polymerization batch absorbs too

strongly (e.g., colored solutions, porous materials such as wood or stone impregnated with monomer [32]).

The polymerization of MMA is most commonly initiated by using thermal radical-forming agents (i.e., initiators). Azo compounds [33,34] and organic peroxy compounds [34] (diacyl peroxides, dialkyl peroxides, alkyl peresters, peroxy dicarbonates, perketals, alkylhydroperoxides) are often employed. A wide range of initiators is available so that the polymer manufacturer can select an initiator with an appropriate half-life for the relevant polymerization temperature. Redox systems also play an important role in the formation of initiating radicals, particularly in polymerizations in aqueous solution [35].

The polymerization of MMA up to high conversion (particularly in bulk polymerization) is governed by the gel (Trommsdorff) effect [36]. The extremely large increase in the viscosity of the reaction medium during polymerization hinders diffusion of the polymer radicals considerably. Bimolecular chain termination thus becomes increasingly difficult and the concentration of the growing radicals increases by several orders of magnitude. The result is a sharp rise in the polymerization rate and mean molecular mass [37], which leads to a further increase in the viscosity of the reaction batch. Under these circumstances removal of the heat of polymerization (57.7 kJ/mol for MMA [38]) may become a serious problem or even a potential hazard.

When polymerization is combined with a shaping or forming operation, the considerable decrease in volume associated with polymerization of the monomer may also lead to problems. For the bulk polymerization of MMA shrinkage can be described by the following equation:

$$v = v\,(1 + e \cdot x)\,\mathrm{M} \tag{3}$$

where v_{M} and v denote the specific volumes of the pure monomer and reaction mixture respectively, x is the degree of conversion, and ε is the contraction constant. The dependence of ε on the reaction temperature θ (in °C) is given by [39]:

$$\varepsilon = -0.2256 - 4.81 \times 10^{-4}\theta - 4.1 \times 10^{-7}\theta^2 \tag{4}$$

At 60 °C, for example, the volume shrinkage of a fully polymerized sample amounts to 25.6 %. Shrinkage decreases with an increase in the size of the alcohol residue of the methacrylate; ε calculated on a molar basis remains roughly constant, however.

The method selected for the industrial polymerization of MMA (e.g., bulk, bead, emulsion, or solution polymerization) is generally governed by the form in which the polymer is finally used. Semifinished products (sheets, blocks, rods, tubes) or pellets, beads, dispersions, and solutions may be produced.

4.1. Bulk Polymerization

Sheets, rods, and tubes of PMMA can be produced by extrusion of molding powders or by bulk polymerization of monomers or "syrup." The molding powder is obtained by bulk, suspension, emulsion, or solution polymerization.

4.1.1. Casting of Acrylic Glass

The conventional casting process is still the most important method for producing high-quality acrylic glass with a mirror-smooth surface (e.g., Plexiglas GS, Perspex, or Vedril). Casting molds are used that consist of two plane-parallel silicate glass plates (e.g., float glass with a typical length, width, and thickness of 3000, 2000, and 8 mm, respectively) which are held apart by a spacer (sealing) and clamped together with metal clamps. The sealing thread runs around the edge of the plates and is usually made of poly(vinyl chloride) or ethylene – vinyl acetate copolymer. It must have sufficient compressibility to allow for the shrinkage occurring during polymerization so that the acrylic sheet does not become detached from the glass plates. The thickness of the resulting sheets is between a few millimeters and up to ca. 30 cm.

The reaction solution consisting of monomer, initiator, and other additives (e.g., comonomers, cross-linking agents, stabilizers, flame retardants, dyes, or pigments) is introduced into the chamber through a small hole in the sealing. Care must be taken to avoid formation of air bubbles, if necessary the solution should first be evacuated. A monomer – polymer solution (syrup) may be used instead of the monomer

solution. This syrup contains 5 – 20 % PMMA and is obtained by prepolymerization of MMA in a stirred-tank reactor. It is advantageous in that the shrinkage and heat of polymerization in the chamber are lower, and the polymerization time is also reduced. Furthermore, the viscous syrup ensures better dispersion of pigments and prevents leakage from the chamber.

The contents of the sealed cells are polymerized in horizontal wracks mounted in hot-air ovens with a high air circulation rate, in autoclaves using spray water (90 – 100 °C), or in water tanks (20 – 60 °C). The residence time may range from a few hours to several days depending on the sheet thickness. Conversion only reaches 80 –90 %, however, because the glasslike solidification of the reaction mixture that occurs during polymerization below the glass transition temperature of the pure polymer prevents further reaction (glass effect). Subsequent polymerization at 110 – 120 °C is therefore necessary to obtain maximum conversion [17,40]. After cooling to room temperature, the polymer sheet is separated from the glass plates and can be removed. The glass plates are cleaned before being used again.

The conventional casting process described above is labor-intensive and time-consuming; attempts have therefore been made to develop more efficient methods for producing cast PMMA glass. One approach is to pack several plane-parallel glass plates and spacers together. The simplest case is a double cell (three plates with two spacers). More complex constructions have been known for a long time and include a process patented in 1933 in which several glass plates with intermediate seals are stacked in vertical rows. The plates are enclosed on both sides by heatable and coolable metal chambers through which a heat-transfer agent flows, and are held together by a screw press [41,42]. The heat-transfer agent flowing through the metal chambers allows accurate temperature control. Shrinkage is compensated for by compression of the packet, resulting in strict tolerances of the cast sheets.

In a special variant of this process, the entire apparatus is tightly encapsulated to facilitate degassing of the reaction batch by evacuation. The system is then pressurized to suppress the formation of boiling bubbles caused by local superheating during polymerization. The advantages of this method are the smaller space requirement and reduced workforce. A disadvantage is the fact that the use of metal chambers means that there is less flexibility as regards variation of the sheet sizes. The process is still used on a small scale in Europe (Rostero) and in the United States (Polycast Corp).

In addition to the production of high molecular mass, high-quality, optical acrylic glass, casting polymerization is still occasionally used to produce regular molding compounds and paint resins. Since these products must not have the perfect plane-parallel surfaces required in semifinished products, cells made of steel or aluminum sheet metal are generally used that are lined with a film (e.g., polyester film). The cells thus enclose polymerization sacks that allow simple mechanical handling. After polymerization the film is removed and the polymethacrylate sheet is ground and granulated in an extruder; residual monomer is removed by degassing. For economic reasons this procedure is only used on a small scale.

4.1.2. Continuous Production of Cast Acrylic Glass

The improvements in the conventional chamber process described in Section 4.1.1 cannot disguise the fact that batch production of acrylic glass is relatively complex and labor-intensive. There have therefore been many attempts to develop continuous processes; the best-known of these is the double-belt process which allows efficient production of large series of sheets up to ca. 6 mm thick [43–46]. The MMA – PMMA syrup is continuously fed between two endless, highly polished steel belts (width up to 3 m, length ca. 250 m). The belts are sealed at the edge with flexible plastic hoses (e.g., of PVC) and guided at a rate of about 1 – 3 m/min over roller systems through several accurately controlled temperature zones. The continuous product sheet can either be wound on rollers or cut into rectangles. This process was first used in the United States by Swedlow and subsequently by Rohm & Haas, Mitsubishi, and Du Pont. The surface quality of the acrylic glass product depends solely on the polishing of the steel belt; the brilliance of acrylic glass produced by the cell polymerization process is seldom ob-

tained. Polymerization in the double-belt process is faster than in the conventional casting process, resulting in somewhat lower molecular masses and therefore some minor deterioration of mechanical properties.

4.1.3. Tube Polymerization

Tubes of PMMA can be produced by extrusion from a molding compound, but also by a special type of bulk polymerization, i.e., centrifugal polymerization [47]. Monomer or syrup is fed into tubular molds (e.g., of stainless steel) that are rotated during polymerization. A film is produced on the internal wall of the pipe by centrifugal forces and polymerizes to form pipes of specific diameter (13 – 650 mm) and wall thickness (1 – 10 mm). This technique is extremely labor-intensive and requires high-precision apparatus. Centrifugal polymerization is therefore seldom used on an industrial scale.

4.2. Suspension Polymerization

PMMA can be produced in the form of small beads by suspension polymerization [48,49]. The monomer (with or without comonomer) is stirred with about twice its volume of water and dispersants (distributors). A dropletlike distribution of the monomer phase in water is then formed. The droplets become increasingly viscous during the course of polymerization. The distributors stabilize the suspension and prevent the droplets from adhering to one another during collisions. Gelatin, cellulose derivatives, synthetic, water-soluble polymers [e.g., poly(vinyl alcohol) or alkali salts of (co)polymers of methacrylic acid] [50], and finely suspended inorganic substances (kaolin, magnesium carbonate, pre-cipitated aluminum hydroxide) have proved suitable as distributors [51,52]. The particle size can be controlled within wide limits (50 – 1000 μm) by choosing appropriate stirring conditions and distributor systems [53].

Suspension polymerization is initiated by adding radical-forming agents, whose distribution equilibrium between the aqueous and the monomer phases should be in favor of the monomer. This also applies to other polymerization auxiliaries, such as chain-transfer agents. The mechanism of suspension polymerization corresponds to that of bulk polymerization. One of the main advantages of suspension polymerization is the extremely high surface to volume ratio of the polymerizing phase. This facilitates removal of the heat of polymerization, which permits higher polymerization rates and thus higher space – time yields than in bulk polymerization. Another advantage is the substantially lower viscosity of the suspension.

After the completion of polymerization, the polymer beads are separated from the aqueous phase (e.g., by filtration), washed with water to remove adhering distributor and salts, and dried.

4.3. Continuous Polymerization of Molding Compounds

The continuous production of cast acrylic glass is of minor importance compared with the batch casting process (see Section 4.1.2). Continuous polymerization processes for the production of PMMA used as injection molding and extrusion material are, however, of major interest. These molding compounds often contain a small proportion of acrylate ester as comonomer (up to ca. 20 %) and have a relatively low weight-average molecular mass (100 000 – 200 000) compared with cast acrylic glass.

The oldest continuous process for producing molding compounds is similar to the double-belt process described in Section 4.1.2 [54,55]: syrup is metered into an endless hose of sheet material and polymerized by passage through a water bath (ca. 75 °C) and an annealing oven (ca. 110 °C). In a plant with a total length of ca. 50 m the residence time is about 1 h.

Technically more elegant processes are generally performed in two stages. In the first stage the monomer is continuously prepolymerized at high temperature (130 – 170 °C) in a stirred-tank reactor. With residence times of 4 – 5 h conversions of 40 – 80 % are obtained. The highly viscous polymer is fed into an extruder. In older processes, the aim was to achieve maximum polymerization in the extruder [56,57]. These techniques are, however, relatively complex and the high polymerization temperatures cause problems with the ceiling equilibrium [38] and thus result in a lower efficiency. More economic processes [58–65] dispense with high conversion in

the extruder. Instead, residual monomer is removed in the degassing zone and recycled to the metering feed device of the stirred tank reactor (if necessary after processing).

The extruded polymer can be pelletized. In order to produce colored pellets, dyes or pigments are mixed with the polymer melt before it is fed into the degassing extruder. Further additives such as chain-transfer agents for adjusting the molecular mass, lubricants, and stabilizers are added either to the melt or to the stirred-tank reactor.

4.4. Batchwise Solution Polymerization

The continuous two-stage polymerization of MMA described in Section 4.3 can be carried out both in bulk and in solution. Batchwise solution polymerization of MMA and in particular of longer-chain methacrylates is also of interest, especially if the resultant polymer solutions constitute the form in which the polymer is used. Typical examples include adhesives and paint resins (see also → Adhesives; → Paints and Coatings, Chap. 2.5.; → Polyacrylates) or additives for the mineral oil sector. Criteria for the choice of solvents include price, solvent quality, toxicity, flammability, boiling point, and the chain-transfer action during polymerization. Solution polymerization can be performed batchwise or with a continuous feed. In order to remove the heat of polymerization it may be advantageous to operate under evaporative cooling conditions.

4.5. Emulsion Polymerization

Methacrylates play an important role as comonomers for acrylates (see → Polyacrylates) and other vinyl compounds (e.g., vinyl acetate) in the production of aqueous plastics dispersions, which are used as paint resins, paper coating agents, textile binders, paper processing agents, and auxiliaries for leather treatment (→ Leather, Chap. 8.4.). Colloidal dispersion and film-forming properties can be widely varied by adding small amounts of methacrylic acid, methacrylamide, or its methylol or methylol ether derivatives. Dispersions of pure polymethacrylates do not, however, form films.

Emulsion polymers with a core – shell structure that contain methacrylates as monomers have become important. Such emulsion polymers are used as impact-resistant modifiers [66] for PMMA [67–69] or other thermoplastics [70, 71]. The solids of a dispersion can be produced by spray drying, freeze coagulation, or squeezing out in an extruder.

5. Uses

Bulk polymers produced from MMA are generally more expensive than many other plastics. Their use is therefore only justified if their special advantages can be fully exploited. These advantages include low density, resistance to fracture, and, above all, excellent optical properties (transparency, gloss) and weather resistance. PMMA also has favorable processing properties (especially good thermoforming) and can be modified in many ways (e.g., coloring, pigmenting, addition of flame retardants, or surface treatment with scratch-resistant coatings).

PMMA has a wide range of applications, particularly in glazing, roofing, vehicles, machinery, and instruments. Pipelines and vessels made of acrylic glass are also of interest for the storage and transportation of food due to their transparency and lack of smell and taste. Optical components such as lenses, reflectors, prisms, and spectacle lenses are also made from polymethacrylates.

5.1. Molding Compounds

PMMA molding compounds are extremely important and versatile as regards applications and uses. The thermoplastically deformable polymers can be used for injection molding and extrusion. The molecular masses (M_w) of PMMA molding compounds are in the range $(1-2)\times10^5$. The processing temperature is generally between 200 and 280 °C.

5.1.1. Standard Molding Compounds

Standard molding compounds consist of MMA homopolymers and copolymers containing up

to ca. 20 % acrylate esters. The comonomer permits variation of the melt flow rate of the polymers.

Standard molding compounds are processed by injection molding. PMMA is especially suitable for this on account of its high dimensional stability. Typical uses include precision parts such as dials, rulers, and other drawing instruments; optical components such as lenses or spectacles; plastic optical fibers (→ Fiber Optics, Chap. 3.3.); and domestic and household articles. Colored injection-molded parts (e.g., covers for automobile rear lights) represent a very important market.

Standard PMMA molding compounds are also extruded on a large scale into flat sheets (width up to 3 m, thickness up to 20 mm), sinusoidally profiled sheets and rods, as well as hollow shapes, double-skin sheets, triple-skin sheets, and tubes (up to 30 cm diameter). If necessary the extrudates are also subjected to surface treatment by coating or coextrusion (e.g., scratch-resistant, antistatic, or nondrip water spreading coating). After cutting to size and covering with a protective packaging film (e.g., polyethylene) they are then marketed as semifinished products.

Trade names of standard PMMA molding materials include Acrylic/Acrycal (Continental Polymers); Acrylite (CYRO); Acrypet (Mitsubishi); Degalan (Degussa); Delpet (Asahi); Diakon (ICI); Lucite (Du Pont); Lucryl/Resarit (BASF/Resart); Oroglas (Rohm & Haas, in the USA: Plexiglas); Plexiglas (Röhm); Polyacrylate (American Cyanamid); Vedril (Atochem).

The worldwide market for PMMA standard molding compounds is ca. 160 000 – 200 000 t/a, of which Western Europe and the United States each account for ca. 70 000 t/a.

5.1.2. Heat-Resistant Molding Compounds

The long-term service temperatures of articles made of standard molding compounds are considerably below 100 °C. Heat-resistant molding compounds have service temperatures of up to ca. 150 °C. They consist of copolymers of MMA with monomers that stiffen the polymer chains. Examples of such comonomers include maleic anhydride [72] (trade names of copolymers: Acrypet ST, Mitsubishi; Degalan HT 120, Degussa; Delpet 980 N, Asahi; Plexiglas hw 55, Röhm) or also methacrylates with bulky ester groups, such as tricyclodecyl methacrylate [12] (Optorez OZ 1000, Hitachi Chemical).

Chain stiffening can, however, also be achieved by subsequent polymer-analogous reaction of the functional groups in the (co)polymer. For example, adjacent acid groups of a copolymer of methyl methacrylate and methacrylic acid are converted into polymethacrylic anhydride structures under the influence of a strong base [73]. The resulting polymers have Vicat softening temperatures of up to 150 °C, depending on the degree of cyclization. These products are, however, still under development.

Poly(*N*-methyl methacrylimide) [*119499-71-7*] (PMMI), also known as poly(*N*-methylglutarimide), a thermoplastic molding compound, is obtained by reacting PMMA with methylamine, → Polyimides, Chap. 5., [74–76].

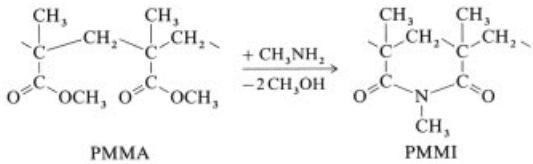

PMMI, which is marketed under the trade names Kamax (Rohm & Haas), Pleximid (Röhm), or PMI resin (Mitsubishi Rayon), has favorable optical properties and good weather resistance. It also has a high modulus of elasticity, (ca. 4×10^3 N/mm^2). The Vicat softening temperature of PMMI can be selectively varied via the degree of imidation, and exceeds 170 °C in highly imidated products.

The imidation reaction can be performed in the melt or in solution. Other primary amines besides methylamine can be used. In addition to the glutarimide units resulting from the cyclization and unreacted methyl methacrylate groups, PMMI may also contain varying proportions (depending on the reaction conditions) of structural elements derived from methacrylic acid and methacrylic anhydride.

The market volume for heat-resistant methacrylate molding compounds is estimated at ca. 3000 t/a worldwide and is expected to grow. The most important customers for heat-resistant

molding compounds are the automobile and lighting industries.

5.1.3. Impact-Resistant Molding Compounds

Impact-resistant PMMA molding compounds are multiphase polymer blends based on a soft elastomer phase dispersed in the rigid PMMA matrix. The elastomer phase may consist of EPDM rubbers, poly(butyl acrylate), polybutadiene, and, in rare cases, polysiloxanes. The choice of rubber is especially important for the notched impact resistance at low temperature; acrylic rubbers have a good weather resistance.

The bonding of the disperse elastomer phase to the rigid matrix and its morphology play an important role in determining the mechanical properties of impact-resistant thermoplastics [66]. If the impact-resistant thermoplastics are to be transparent, the refractive indices of the rigid and soft phases must be identical (as long as the domain size of the elastomer phase does not remain below a critical value, e.g., below the wavelength of visible light).

Various methods can be used to produce impact-resistant PMMA: in older processes a suitable type of rubber was grafted with MMA (and comonomers to adjust the refractive index) by bulk or solution polymerization. Increasing conversion leads to phase inversion, whereby the elastomer phase, which frequently includes PMMA (or, in the presence of comonomers, the corresponding copolymer), becomes the disperse phase. This desired morphology and the size of the elastomer particles can be controlled by the viscosity of the reaction medium and the shear forces resulting from stirring. They are then fixed by cross-linking when conversion is nearly complete. Graft polymerization by two-stage suspension polymerization is carried out on an industrial scale, too [48].

Emulsion polymerization is particularly effective if particle size and particle morphology have to be selectively modified (see Section 4.5). Impact-resistance modifiers with a core – shell structure have therefore been developed for PMMA [67–69].

Worldwide demand for impact-resistant PMMA is estimated to be ca. 35 000 – 40 000 t/a. Products in which butadiene constitutes the tough phase include methacry-late – butadiene – styrene (MBS) [*25053-09-2*] (e.g., Plexalloy-F-PAB, Röhm; Cyrolite G 20, CYRO) and methacrylate – acrylonitrile – butadiene – styrene (MABS) [*9010-94-0*] (e.g., Terluran, BASF; Cyrolite XT Polymer, CYRO). PMMA modified with butyl acrylate is commercially available under the trade names Delpet SR (Asahi), Diakon TD (ICI), Lucryl KR (BASF), Oroglas DR (Rohm & Haas), Plexiglas zk (Röhm), and Vedril V (Atochem).

5.1.4. Molding Compounds for Special Optical Uses

PMMA is a highly prized material mainly due to its very good weather resistance, outstanding optical quality, and good processability (e.g., by injection molding). Optical components (e.g., lenses, prisms, mirrors) represent an important area of use for PMMA [77, 78]. The relatively high water absorption of PMMA (ca. 2 %) causes dimensional changes and change in refractive index. This disadvantage can be overcome by using cycloaliphatic methacrylates such as cyclohexylmethyl methacrylate [79] or tricyclodecyl methacrylate [12] as comonomers. Other methacrylates are used in soft contact lenses, in particular the hydrophilic poly(2-hydroxyethyl methacrylate) [80, 81].

The excellent optical properties of PMMA (especially its low orientation birefringence) favor its use as a substrate for optical data storage media. However, in asymmetric, single-sided audio compact discs the relatively high water absorption of the polymer and the associated warpage may detract from these advantages [82, 83]. This problem does not arise with the two-sided video disc; PMMA molding materials are therefore used for these products [84]. See also → Information Storage Materials, Chap. 2.3.2.

A relatively promising area of use for extremely pure PMMA is plastic optic fibers (POF) [85, 86], → Fiber Optics, Chap. 3.3. The theoretical limiting values of transmission loss of PMMA in the 650 and 567 nm minima of the damping spectrum are 100 and 27 dB/km, respectively [87]. Limits of 5 – 6 dB/km at 680 and 650 nm can be achieved with special deuterated or fluorinated methacrylates [88]. PMMI (see

Section 5.1.2) could become important as heat-resistant core material for POFs.

PMMA is also employed in lithographic materials in semiconductor technology. Use as a high-resolution positive resist is based on the depolymerization of this polymer by radiation, see → Imaging Technology. The irradiated regions are accurately dissolved out with suitable solvents. Sensitivity to electron beams and UV light can be increased further by copolymerization with suitable monomers [89, 90].

5.2. Semifinished Products

5.2.1. Extruded Acrylic Glass

A substantial proportion of standard PMMA molding compounds (and in the United States particularly impact-resistant PMMA) is extruded into semifinished products (sheets, double- and triple-skin sheets, corrugated profile sheets, tubes) [91]. Reference has already been made to possible surface modifications of these extrudates (see Section 5.1.1). Cast acrylic glass with a matt surface is difficult to produce, extruded semifinished products with matt surface structures are therfore an interesting alternative. It would be desirable if such products could also be thermoplastically formed without losing their matt structure. One possible solution might be the addition of a highly cross-linked bead polymer to the extrusion compound which would certainly open up new applications extending beyond the traditional areas of use of illuminated advertising signs, projection systems, and picture glass.

The most important market sectors for extruded semifinished products are illuminated advertising signs and glazing (including lighting domes). In Europe the market volume is ca. 70 000 t/a. Important trade names include Acrylite FF (CYRO); Altulex and Vedril E (Atochem); Deglas (Degussa); Perspex TX (ICI); Plexiglas XT and Plexiglas SDP (Röhm).

5.2.2. Cast Acrylic Glass

The substantially more economic production of extruded acrylic glass has led to a considerable decline in the production of cast semifinished products. Nevertheless, cast material has retained its position in applications with stringent requirements as regards surface quality and optical properties. For example, high growth rates for cast acrylic glass are found in the sanitary ware sector (e.g., baths, shower units, and washbasins). On account of certain deficiencies revealed by the hot water alternating test (DIN EN 198), (co)extruded products have still not been widely accepted in this sector.

A more recent use of cast acrylic glass (optionally with flame retardants) are noise protection walls on motorways. This market sector is expected to show a high growth rate. Large sheets ($\geq 2\times4$ m) are needed to prevent drivers from being disturbed by the frames used for mounting.

Aviation glazing, one of the oldest applications of cast acrylic glass, has become increasingly important on account of the enormous increase in air traffic. Acrylic glass has the advantages of low density and outstanding resistance to extreme weather conditions (high UV radiation intensities at high altitudes). In order to improve strength, biaxially stretched polymer is often used. The extremely stringent optical requirements [e.g., freedom from optical distortion, low deflection, lack of internal defects (e.g., bubbles, inclusions), lack of surface defects (e.g., pimples)] predestine cast acrylic glass for aviation glazing. Aviation standards prohibit the use of extruded material.

The total market volume for cast acrylic glass in Europe is ca. 90 000 t/a. Important trade names and manufacturers include Acrylite (CYRO); Altuglas, Casocryl, and Vedril C (Atochem); Paraglas (Degussa); Paralinx (Kyowa Gas); Perspex (ICI); Plexiglas GS (Röhm); Plexiglas (Rohm & Haas, United States); Resartglas (BASF); Shinkolite (Mitsubishi); Sumipex (Sumitomo Chemical).

5.3. Binders and Paints

Methacrylates are used in large amounts as comonomers in the production of acrylate–methacrylate copolymers for binders and paint resins (→ Paints and Coatings, Chap. 2.5.). Methacrylate is often used in this connection to obtain the desired glass transition temperature.

The use of such copolymers in the form of solution polymers, aqueous dispersions, or solid products extends over a wide field (e.g., weather-resistant paints, textile binders, paper coating materials, and adhesives); further details are given elsewhere, see → Polyacrylates.

5.4. Additives for the Petroleum Industry

Long-chain poly(alkyl methacrylates), PAMA, are used as additives in the petroleum sector for many purposes [92–95]. The best-known application is probably their use as viscosity index improvers in engine and gear oils (→ Lubricants and Related Products, Chap. 7.1.). At low temperatures viscosity index improvers increase the viscosity of the mixture only slightly. At high temperatures, however, they compensate for the sharp decrease in viscosity of the base oil by virtue of their intrinsically high contribution to the solution viscosity of the whole system.

This is because the mineral oil becomes an increasingly better solvent for PAMA at higher temperatures and the polymer coils expand greatly. Viscosity index improvers thus prevent destruction of the lubricating oil film and resultant damage to the engine. Viscosity index improvers are generally constituents of an extremely complex additive formulation and interact with other components not only in improving flow behavior, but also in cleaning the engine, dispersing dirt and contaminants, and preventing oxidation of the oil. This dispersing effect can be achieved by incorporating polar comonomers (e.g., *N*-vinylpyrrolidone) into the PAMA chain.

PAMA-based additives also act as pour-point depressants (→ Lubricants and Related Products). The pour point denotes the temperature at which an oil solidifies as a result of precipitation of paraffins. At concentrations of less than 1 %, PAMAs hinder the growth of paraffin crystals and thus prevent the formation of aggregates of crystal needles. The oil remains free-flowing and the oil pump is able to function during cold starting.

5.5. Miscellaneous Uses

Polymethacrylates are used in many applications besides those described in Sections 5.1, 5.2, 5.3, 5.4.

Most poly(vinyl chloride) processing auxiliaries consist of high molecular mass polymers based on MMA. They are generally marketed in powder form as spray-dried emulsion polymers [96]. Important examples of such products include Diakon APA (ICI), Kane ACE (Kanegafuchi Chemical), Metablen (Mitsubishi Rayon), Paraloid K (Rohm & Haas), and Vestiform (Hüls).

Methacrylate reaction resins (monomer–polymer systems) are used in the building industry in mortars and concretes, and in synthetic floor screeds [97, 98].

Alkali-soluble suspension polymers, e.g., Rohagit (Röhm), are used as thickeners for aqueous dispersions and cosmetics, as sizings and finishes in the textile industry, and as dispersion and stabilizing auxiliaries in suspension polymerization. Cation-exchange resins can be obtained by hydrolysis of PMMA, thus producing weakly acidic carboxylic functions (→ Ion Exchangers, Chap. 2.2.1.).

Polymethacrylates are used as dental materials for fillings, false teeth, and prostheses [99], and in surgery as a bone cement for implants. Examples of such products include Kallodoc (ICI) and Plexidon (Röhm).

Coatings of polymethacrylate copolymers such as Eudragit E-30 D [*9010-88-2*] or Eudragit RS [*39316-06-8*] (Röhm) are used as galenic auxiliaries for controlled-release tablets [100, 101].

6. Toxicology

No toxic effects of PMMA have been described, and the polymer is regarded as inert and nontoxic. In individual cases, however, contact of the mucous membranes with polymer powders or dusts may cause allergic and irritating reactions. Skin reactions (presumably due to residual monomer) and asthma have also been noted [102, 103].

Since PMMA is used as a bone cement, its potential carcinogenic effect has been monitored in animal implant tests. Malignant tumors were

found in some cases at the implant site, but also occurred with inert materials of the same shape. Such tumors are therefore most likely to be foreign-body carcinoma resulting from mechanical irritation. Tumors have not yet been observed in humans after many years' use [102].

7. Environmental Aspects

In contrast to most other polymers, PMMA can be almost quantitatively depolymerized to the monomer at high temperature. Conversion of production waste and waste from semifinished product blanks to MMA by cracking had already been investigated at the start of large-scale industrial production in the 1930s [104, 105]. In these earlier processes the polymer was mixed with sand and depolymerized batchwise in retorts or with superheated steam (to avoid local overheating and thus uncontrolled pyrolysis). Continuous processes now operate with an indirectly heated fluidized bed and are substantially more economical [106–108].

8. References

1. H. von Pechmann, O. Röhm, *Ber. Dtsch. Chem. Ges.* **34** (1901) 427.
2. E. Trommsdorff: *Dr. Otto Röhm, Chemiker und Unternehmer,* Econ, Düsseldorf 1976.
3. Röhm & Haas, DE 724 229, 1932 (O. Röhm, W. Bauer).
4. *Kunststoffe* **79** (1989) 923.
5. *Kirk-Othmer,* **vol. 16,** Wiley, New York, 1995, p. 497.
6. G. Schreyer, *Angew. Makromol. Chem.* **11** (1970) 159.
7. K. Oberbach, *Kunststoffe* **79,** (1989) 713.
8. J. Schmitz, E. Bornschlegl, G. Dupp, G. Erhard, *Plastverarbeiter* **39/4** (1988) 50.
9. O. Fuchs: *Gnamm-Fuchs: Lösungs- und Weichmachungsmittel,* **3 vols.,** Wissenschaftl. Verlagsgesellschaft, Stuttgart 1980.
10. O. G. Lewis: *Physical Constants of Linear Homopolymers,* Springer Verlag, Berlin 1968.
11. Perspex Acrylic Sheet, ICI Bulletin.
12. H. Kawai, F. Kanega, H. Kohkame, *Proc. SPIE,* **896,** (1988) 68.
13. Röhm GmbH, unpublished data.
14. *Kirk-Othmer,* 3rd ed., **15,** 377.
15. M. Stickler, unpublished results.
16. M. A. Müller, M. Stickler, *Makromol. Chem., Rapid Commun.* **7** (1986) 575.
17. W. Wunderlich, M. Stickler in B. M. Culbertson, J. E. McGrath (eds.): *Advances in Polymer Synthesis,* Plenum Publishing, New York 1985, p. 505.
18. B. Boutevin, Y. Pietrasanta: *Les Acrylates et Polyacrylates Fluorés Dérivés et Applications,* EREC, Puteaux 1988, p. 235.
19. G. Odian: *Principles of Polymerization,* McGraw-Hill, New York 1970.
20. G. C. Eastmond in C. H. Bamford, C. F. H. Tipper (eds.): *Comprehensive Chemical Kinetics,* **vol. 14 A,** Elsevier, Amsterdam 1976, p. 196.
21. H. Yuki, K. Hatada, *Adv. Polym. Sci.* **31** (1979) 1.
22. A. H. E. Müller in J. E. McGrath (ed.): "Anionic Polymerization, Kinetics, Mechanisms, and Synthesis," *ACS Symp. Ser.* **166** (1981) 441.
23. M. Morton: *Anionic Polymerization: Principles and Practice,* Academic Press, New York 1983.
24. O. W. Webster et al., *J. Am. Chem. Soc.* **105** (1983) 5706.
25. W. R. Hertler, D. Y. Sogah, O. W. Webster, G. M. Cohen, *Macromolecules* **20** (1987) 1473.
26. M. Stickler, G. Meyerhoff, *Makromol. Chem.* **179** (1978) 2729.
27. G. Oster, N.-L. Yang, *Chem. Rev.* **68** (1968) 125.
28. S. S. Labana, *J. Macromol., Sci.-Revs. Macromol. Chem.* **C 11** (1974) 299.
29. A. Ledwith, *J. Oil Colour Chem. Assoc.* **59** (1976) 157.
30. A. Ledwith, *Pure Appl. Chem.* **49** (1977) 431.
31. J. P. Fouassier, P. Jacques, D. J. Lougnot, T. Pilot, *Polym. Photochem.* **5** (1984) 57.
32. F. Gütlbauer, E. Proksch, H. Bildstein, *Österr. Chem. Z.* **67** (1966) 349.
33. P. S. Engel, *Chem. Rev.* **80** (1980) 99.
34. G. Moad, D. H. Solomon in G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science* **vol. 3,** part I, Pergamon Press, Oxford 1989, p. 97.
35. C. H. Bamford in G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science,* **vol. 3,** part I, Pergamon Press, Oxford 1989, p. 123.
36. G. P. Gladyshev, K. M. Gibov: *Polymerization at Advanced Degrees of Conversion,* Israel Program for Scientific Translations, Jerusalem 1970.

37. K. F. O'Driscoll in G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science,* **vol. 3,** part I, Pergamon Press, Oxford 1989, p. 161.
38. H. Sawada: *Thermodynamics of Polymerization,* Marcel Dekker, New York 1976.
39. D. Panke, M. Stickler, W. Wunderlich, *Makromol. Chem.* **184** (1983) 175.
40. M. Stickler, *Makromol. Chem.* **184** (1983) 2563.
41. Röhm & Haas, DE 659 469, 1933 (O. Röhm).
42. Röhm & Haas, DE 639 095, 1934 (W. Bauer).
43. ICI, US 2 500 728, 1945 (E. G. Williams).
44. Swedlow Incorp., BE 687 405, 1966 (C. J. Opel, P. H. Bottoms).
45. Swedlow Incorp., BE 687 406, 1966 (O. H. Heilsund).
46. Swedlow Incorp., US 3 376 371, 1968 (C. J. Opel, P. H. Bottoms).
47. Röhm & Haas, DE 673 394, 1936 (W. Bauer).
48. M. Munzer, E. Trommsdorff in C. E. Schildknecht (ed.): "High Polymers," *Polymerization Processes,* vol. 29, J. Wiley & Sons, New York 1977.
49. *Winnacker-Küchler,* 4th ed., **6,** 414.
50. ICI, GB 427 494, 1933 (J. W. C. Crawford, J. McGrath).
51. Röhm & Haas, DE 735 284, 1935 (O. Röhm, E. Trommsdorff).
52. Röhm & Haas, DE 747 596, 1935 (O. Röhm, E. Trommsdorff).
53. H. Hopf, H. Lüssi, P. Gerspacher, *Makromol. Chem.* **78** (1964) 37.
54. Degussa, DE 1 212 301, 1963 (E. Bäder, W. Unseld, G. Morlock).
55. Degussa, DE 1 923 393, 1969 (E. Bäder, H. Amann, W. Unseld).
56. ICI, DE-OS 1 253 459, 1967 (F. Bild, A. Jukes).
57. W. A. Mack, *Chem. Eng. (N.Y.)* **79** (1972) 99.
58. Dow Chemical, GB 715 666, 1952 (A. Handsen).
59. Monsanto, CA 493 550, 1953 (H. F. Park, R. M. Dickey).
60. ICI, GB 875 853, 1960 (F. Bild, A. William).
61. Amer. Cyanamid, US 3 252 950, 1966 (J. Schmitt, J. Terenzil).
62. Mitsubishi Rayon, DE-OS 2 341 318, 1973 (K. Shimada et al.).
63. Mitsubishi Rayon, DE-OS 2 504 417, 1975 (K. Shimada et al.).
64. Bayer AG, DE-OS 2 724 360, 1977 (F. Wingler et al.).
65. Mitsubishi Rayon, DE-OS 2 438 927, 1974 (T. Maeda, T. Nishizawa, M. Toda, Y. Iwaoka).
66. C. B. Bucknall: *Toughened Plastics,* Applied Science Publ., London 1977.
67. Rohm & Haas, US 3 793 402, 1971 (F. Owens).
68. Röhm, EP 0 113 924, 1983 (N. Sütterlin et al.).
69. P. A. Lovell et al., *Proc. ACS Div. Polym. Mat. Sci. Eng.* **63** (1990) 583.
70. Röhm, EP 0 326 938, 1989 (J. Fischer, W. Siol, M. Munzer, T. Rhein).
71. Röhm, DE 3 743 199, 1987 (J. Fischer, W. Siol, M. Munzer, T. Rhein).
72. B. C. Trivedi, B. M. Culbertson: *Maleic Anhydride,* Plenum Press, New York 1982.
73. Sumitomo, EP 0 264 508, 1986 (Y. Kato et al.).
74. Röhm & Haas, DE 1 113 308, 1960 (G. Schröder, K. Tressmar, U. Baumann).
75. Röhm & Haas, DE 2 652 118, 1976 (R. M. Kopchik).
76. Mitsubishi Rayon, EP 0 200 530, 1986 (H. Anzai et al.).
77. R. H. Waxler, D. Horowitz, A. Feldmann, *Appl. Opt.* **18** (1979) 101.
78. C. Hofmann, *Plaste Kautsch.* **36** (1989) 338.
79. G. Klepek, F & M *Feinwerktech. Meßtechn.* **96** (1988) 43.
80. H. K. Singer, E. C. Bellantoni, A. R. LeBoef in J. I. Kroschwitz (ed.): *Encyclopedia of Polymer Science and Engineering,* **vol. 4,** Wiley, New York 1985, p. 164.
81. T. B. Harvey, W. E. Meyers, L. M. Bowmann, *Polym. Mater. Sci. Eng.* **59** (1988) 73.
82. J. Hennig, *Jpn. J. Appl. Phys.* **26** (1987) 26 (Suppl.).
83. H. Ryn, K. Akagane, H. Kori, *CEER Chem. Econ. Eng. Rev.* **18** (1986) 26.
84. G. Kämpf, *Polym. Eng. Sci.* **27** (1987) 1421.
85. R. M. Glen, *Chemtronics* **1** (1986) 98.
86. C. Emslie, *J. Mater. Sci.* **23** (1988) 2281.
87. T. Kaino, M. Fujiki, K. Jinguji, *Rev. Electr. Commun. Lab.* **32** (1984) 478.
88. T. Kaino, Y. Katayama, *Polym. Eng. Sci.* **29** (1989) 1209.
89. R. Pethrick, *Progr. Rubber Plast. Technol.* **2** (1986) 1.
90. E. Reichmanis, L. F. Thompson, *Chem. Revs.* **89** (1989) 1273.
91. M. Buck, *Kunststoffe* **80** (1990) 1132.
92. C. W. Georgi: *Motor Oils and Engine Lubrication,* Reinhold Publ. Co., New York 1950.
93. C. V. Smalheer, K. Smith: *Lubricant Additives,* The Lezius-Hiles, Cleveland 1967.
94. A. Schilling: *Motor Oils and Engine Lubrication,* 2nd ed., Scientific Publication, London 1968.

95. M. W. Ranney: *Lubricant Additives,* Noyes Data Corp., Park Ridge, N.Y. 1973.
96. D. L. Dunkelberger, *J. Vinyl Technol.* **9** (1987) 173.
97. H. J. Peschke, *Zentralbl. Industriebau* **33** (1987) 10.
98. H. Schuhmann, *BmK (Bauen mit Kunststoffen)* **5** (1988) 5.
99. I. E. Ruyter, H. Oeysaed, *CRC Crit. Rev. Biocompat.* **4** (1988) 247.
100. R. K. Chang, C. H. Hsiao, J. R. Robinson, *Pharm. Technol.* **11** (1987) 56.
101. K. O. R. Lehmann, *Drugs Pharm. Sci.* **36** (1989) 153.
102. D. Henschler, *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründungen von MAK-Werten,* VCH Verlagsgesellschaft, Weinheim 1988.
103. B. Kennes, P. Garcia-Herreros, P. Dierckx, *Clin.Allerg.* **11** (1981) 49.
104. Röhm & Haas, DE 642 289, 1935 (P. Weisert).
105. Röhm & Haas, DE 729 730, 1940 (K.-Th. Kautter, W. Leitenberger).
106. H. Sinn, W. Kaminsky, J. Janning, *Angew. Chem.* **88** (1976) 737.
107. W. Kaminsky, J. Menzel, H. Sinn, *Conserv. Recycling* **1** (1977) 91.
108. W. Kaminsky, H. Sinn, J. Janning, *Chem. Ing. Tech.* **51** (1979) 419.

Polymethine Dyes → Methine Dyes and Pigments
Polymethylbenzenes → Hydrocarbons
Polymethylbenzenes → Xylenes
Polynorbornene → Rubber, 3.Synthetic
Polyoctenamers → Rubber, 3.Synthetic

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Polyolefins

Standard Article

Kenneth S. Whiteley[1], T. Geoffrey Heggs[2], Hartmut Koch[3], Ralph L. Mawer[4], Wolfgang Immel[5]

[1]formerly ICI Plastics Division, Welwyn Garden City, United Kingdom
[2]current address: The Mill, Great Ayrton, TS9 6PX, United Kingdom
[3]Shell Chemical International, Shell Center, London, United Kingdom
[4]Shell Chemical International, Shell Center, London, United Kingdom
[5]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a21_487
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Polyethylene
1.1. Introduction
1.2. Properties of Polyethylenes
1.2.1. Molecular Structure and Morphology
1.2.2. General Properties
1.2.3. High Molecular Mass (Bimodal) Polyethylene (HMWPE)
1.2.4. Ultra High Molecular Mass Polyethylene (UHMWPE)
1.2.5. Properties of Ethylene Copolymers
1.3. Polymerization Chemistry
1.3.1. Free-Radical Catalysis
1.3.1.1. Introduction
1.3.1.2. Copolymerization
1.3.2. Coordination Catalysis
1.3.2.1. Phillips Catalysts
1.3.2.2. Ziegler Catalysts
1.3.2.3. Single-Site Catalysts (Metallocenes)
1.3.2.4. Copolymerization
1.4. Raw Materials
1.4.1. Ethylene
1.4.2. Comonomers
1.4.3. Other Materials
1.5. Production Processes
1.5.1. High-Pressure Process
1.5.1.1. Autoclave Reactor
1.5.1.2. Tubular Reactor
1.5.1.3. High-Pressure Copolymers
1.5.1.4. Linear Low-Density Polyethylene (LLDPE)
1.5.2. Suspension (Slurry) Process
1.5.2.1. Autoclave Process
1.5.2.2. Loop Reactor Process
1.5.3. Gas-Phase Process
1.5.4. Solution Process
1.6. Uses
1.6.1. Film
1.6.2. Extrusion Coating
1.6.3. Blow Molding
1.6.4. Injection Molding
1.6.5. Pipe
1.6.6. Wire and Cable Insulation
1.6.7. Ethylene Copolymers
1.6.8. Ultra High Modulus Polyethylene Fibers
1.6.9. Joining Polyethylene
1.7. Chemically Modified Polyethylenes
1.7.1. Cross-Linked Polyethylene
1.7.2. Chlorinated Polyethylene
1.7.3. Fluorinated Polyethylene
1.8. Environmental Aspects
1.8.1. Manufacture
1.8.2. Polymer Disposal and Recycling
2. Polypropylene
2.1. Historical Survey
2.2. Polymer Structure
2.2.1. Molecular and Chain Structure
2.2.2. Crystallization and Morphology
2.3. Raw Materials
2.3.1. Propene
2.3.2. Polymerization Diluents
2.3.3. Catalyst Preparation

2.3.3.1. $TiCl_3$-Based Catalysts
2.3.3.2. Supported Catalysts
2.3.3.3. Homogeneous Catalysts
2.3.3.4. Aluminum Alkyl Cocatalysts
2.3.4. Hydrogen
2.4. Polymerization Mechanism
2.5. Industrial Processes
2.5.1. Suspension Homopolymerization (Early Diluent Processes)
2.5.2. Bulk Polymerization in Liquid Propene
2.5.3. Solution Polymerization
2.5.4. Spheripol Process
2.5.5. Hypol Process
2.5.6. Gas-Phase Processes
2.5.7. Copolymerization
2.5.7.1. Random Copolymerization
2.5.7.2. Impact (Block) Copolymerization
2.5.8. Product Finishing
2.5.9. Additives
2.6. Compounding
2.7. Properties
2.7.1. Homopolymer
2.7.2. Copolymers
2.7.3. Elastomer Blends with Polypropylene
2.8. Uses
2.8.1. Injection Molding
2.8.2. Blow Molding
2.8.3. Fibers and Flat Yarns
2.8.4. Film
2.8.5. Foil and Sheet
2.8.6. Extruded Pipe
2.9. Environmental Aspects
2.9.1. Recycling
2.9.2. Pyrolysis
2.9.3. Incineration
2.9.4. Environmental Interactions
3. Poly(1-Butene)
3.1. Production
3.2. Properties
3.2.1. Structure and Physical Properties
3.2.2. Application Properties
3.3. Processing
3.4. Applications
4. Polyisobutylene
4.1. Introduction
4.2. Production
4.2.1. Isobutene as Raw Material
4.2.2. Low Molecular Mass Polyisobutylenes
4.2.3. Medium Molecular Mass Polyisobutylenes
4.2.4. High Molecular Mass Polyisbutylenes
4.3. Structure
4.4. Properties
4.5. Uses and Processing
4.6. Commercial Products
4.7. Copolymers
5. Poly(4-Methyl-1-Pentene)
5.1. Raw Materials
5.1.1. Monomers
5.1.2. High-Melting Nucleating Polymers
5.1.3. Catalysts
5.2. Industrial Manufacture
5.3. Polymer Structure
5.3.1. Homopolymer
5.3.2. Copolymers
5.4. Properties of PMP Products
5.4.1. Optical Properties
5.4.2. Stability and Chemical Resistance
5.4.3. Mechanical Properties
5.4.4. Electrical Properties
5.4.5. Surface Properties and Permeability
5.4.6. Rheology
5.5. Economic Aspects
5.6. Fabrication and Applications
5.6.1. Fabrication
5.6.2. Applications
5.7. Health and Environmental Aspects

[Top of Page]

1. Polyethylene

Kenneth S. Whiteley

1.1. Introduction

Despite ethylene's simple structure, the field of polyethylene is a complex one with a very wide range of types and many different manufacturing processes. From a comparatively late start, polyethylene production has increased rapidly to make polyethylene the major tonnage plastics material worldwide (45×10^6 t capacity in 1995). In the 1920s research into the polymerization of unsaturated compounds such as vinyl chloride, vinyl acetate, and styrene led to industrial processes being introduced in the 1930s, but the use of the same techniques with ethylene did not lead to high polymers. The chance observation in 1933 by an ICI research team that traces of a waxy polymer were formed when ethylene and benzaldehyde were subjected to a temperature of 170 °C and a pressure of 190 MPa, led to the first patent in 1936 and small-scale production in 1939. The polymers made in this way, by using free radical initiators, were partially crystalline, and measurement of the density of the product was quickly established as a means of determining the crystallinity. Due to the side reactions occurring at the high temperatures employed, the polymer chains were branched, and densities of 915 – 925 kg/m^3 were typically obtained. The densities of completely amorphous and completely crystalline polyethylene would be 880 and 1000 kg/m^3, respectively.

During the 1950s three research groups working independently discovered three different catalysts which allowed the production of essentially linear polyethylene at low pressure and temperature. These polymers had densities in the region of 960 kg/m^3, and became known as high-density polyethylenes (HDPE), in contrast to the polymers produced by the extensively commercialized high-pressure process, which were named low-density polyethylenes (LDPE). These discoveries laid the basis for the coordination catalysis of ethylene polymerization, which has continued to diversify. Of the three discoveries at Standard Oil (Indiana), Phillips Petroleum, and by KARL ZIEGLER at the Max-Planck-Institut für Kohlenforschung, the latter two have been extensively commercialized. More recently the observation that traces of water can dramatically increase the polymerization rate of certain Ziegler catalysts has led to major developments in soluble coordination catalysts and later their supported variants.

The coordination catalysts allowed for the first time the copolymerizaton of ethylene with other olefins such as butene, which by introducing side branches reduces the crystallinity and allows a low-density polyethylene to be produced at comparatively low pressures. Although Du Pont of Canada introduced such a process in 1960, worldwide the products remained a small-volume specialty until 1978 when Union Carbide announced their Unipol process and coined the name linear low-density polyethylene (LLDPE). In addition to developing a cheaper production process, Union Carbide introduced the concept of exploiting the different molecular structure of the linear product to make tougher film. Following this lead, LLDPE processes have been introduced by many other manufacturers.

The history of these discoveries is covered in [1-4], [20].

The three types of polyethylene outlined above account for the major part of polyethylene production (Table 1), but the picture is slightly confused since many plants have the capability of producing more than one type of product (so- called swing plants). Additionally, copolymers are made by both types of process. The free-radical process is used to produce copolymers of vinyl acetate, acrylates, methacrylates, and the corresponding acids, but chain transfer prevents the use of higher olefins because of the drastic reduction in molecular mass of the polymer. The coordination catalysts are able to copolymerize olefins, but are deactivated by more polar materials. Because of the complex interplay of the capabilities of modern plants, it is convenient to treat separately the products, the catalysts, and the processes.

Table 1. Polyethylene production capacities in 10^3 t/a * (1995)

	North America	Western Europe	Eastern Europe	Japan	Rest of World	Total
LDPE	3891	5783	1918	1444	4210	17 246
LLDPE	4422	1848	100	1059	3728	11 157
HDPE	6198	4008	873	1024	4715	16 891
Total PE	14 511	11 639	2891	3527	12 653	45 221

* Data from Chem Systems, London.

1.2. Properties of Polyethylenes

1.2.1. *Molecular Structure and Morphology*

Figure 1 shows schematic structures for the three polyethylenes, with the main features exaggerated for emphasis. LDPE has a random long-branching structure, with branches on branches. The short branches are not uniform in length but are mainly four or two carbon atoms long. The ethyl branches probably occur in pairs [21], and there may be some clustering of other branches [22]. The molecular mass distribution (MMD) is moderately broad.

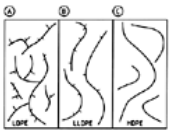

Figure 1. Schematic molecular structure

A) Low-density polyethylene; B) Linear low-density polyethylene; C) High-density polyethylene

LLDPE has branching of uniform length which is randomly distributed along a given chain, but there is a spread of average concentrations between chains, the highest concentrations of branches being generally in the shorter chains [23]. The catalysts used to minimize this effect generally also produce fairly narrow MMDs.

HDPE is essentially free of both long and short branching, although very small amounts may be deliberately incorporated to achieve specific product targets. The MMD depends on the catalyst type but is typically of medium width.

Polyethylene crystallizes in the form of platelets (lamellae) with a unit cell similar to that of low molecular mass paraffin waxes [24]. Due to chain folding, the molecular axes are oriented perpendicular to the longest dimension of the lamella and not parallel to it as might be expected (Fig. 2). The thickness of the lamellae is determined by the crystallization conditions and the concentration of branches and is typically in the range of 8 – 20 nm. Thicker lamellae are associated with higher melting points and higher overall crystallinities. Slow cooling from the melt or annealing just below the melting point produces thicker lamellae. Where long molecules emerge from the lamella they may either loop back elsewhere into the same lamella or crystallize in one or more adjacent lamellae, thereby forming “tie molecules”.

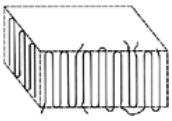

Figure 2. Folded-chain lamellar crystal of polyethylene

Thermodynamically the side branches are excluded from the crystalline region because their geometry is too different from that of the main chains to enter the crystalline lamellae. Therefore, the branches initiate chain folding, which results in thinner lamellae with the branches mainly situated on the chain folds on the surface of the lamellae. However, on rapid cooling these energetically preferred placements may not always occur, and some branches may become incorporated as crystal defects in the crystalline regions. Detailed measurements by solid-state NMR and Raman spectroscopy show that the categorization into crystalline and amorphous phases is too simplistic and a significant fraction of the polymer is present in the form of an “interfacial” fraction which has neither the freedom of motion of a liquid, nor the well-defined order of a crystal [25], [26]. A further result of a side branch is that having been prevented from folding directly into the same lamella, the polymer chain may form a tie molecule that links to one or more further lamellae.

Under moderately slow cooling conditions, crystallization may be nucleated at a comparatively small number of sites. Crystallization then propagates outwards from these centers until the surfaces of the growing spheres meet. The resulting spherulites show a characteristic banded structure under a polarizing optical microscope. The typical milkiness of polyethylene is due to light scattered by spherulites or other, less well defined aggregates of crystallites, rather than by the crystallites themselves, which are much smaller than the wavelength of light [27]. Ethylene copolymers may be transparent, although partially crystalline.

1.2.2. *General Properties*

LDPE and LLDPE are translucent whitish solids and are fairly flexible. In the form of films they have a limp feel and are transparent with only a slight milkiness. HDPE on the other hand is a white opaque solid that is more rigid and forms films which have a more turbid appearance and a crisp feel.

Polyethylene does not dissolve in any solvent at room temperature, but dissolves readily in aromatic and chlorinated hydrocarbons above its melting point. On cooling, the

solutions tend to form gels which are difficult to filter. Although LDPE and LLDPE do not dissolve at room temperature, they may swell in certain solvents with a deterioration in mechanical strength. Manufacturers issue data sheets detailing the suitability of their products for use in contact with a wide range of materials. In addition to solvents, polyethylene is also susceptible to surface active agents which encourage the formation of cracks in stressed areas over prolonged periods of exposure. This phenomenon, known as environmental stress cracking (ESC), is believed to be due to lowering of the crack propagation energy [28]. In general, HDPE is the preferred polyethylene for liquid containers.

Some properties of typical LDPE, HDPE, and LLDPE are listed in Table 2. Polyethylenes are routinely characterized by their density and melt flow index (MFI). The MFI test was originally chosen for LDPE to give a measure of the melt characteristics under conditions related to its processing. It is carried out by applying a standard force to a piston and measuring the rate of extrusion (in g/10 min) of the polyethylene through a standard die (Fig. 3). Other standard conditions are sometimes used on the same equipment to extend the range of information and because higher loads are sometimes considered more appropriate for HDPE. The short parallel section of the standard die introduces errors which mean that the MFI cannot be accurately related to viscosity (it would be an inverse relationship). For LDPE and HDPE, the MFI increases disproportionately with the applied load. The ratio of the two MFIs gives a measure of the ease of flow at high shear and is sometimes known as the flow ratio. The die swell ratio can also be measured in the MFI test and gives a measure of the elastic memory of the melt. This parameter correlates usefully with extrusion processes, where a low value is desirable for tubular film, and a high value is necessary for extrusion coating.

Table 2. Properties of some typical polyethylenes (data from Repsol Quimica)

Property	LDPE	HDPE	LLDPE	Method	Standard
Polymer grade	Repsol PE077/A	Hoechst GD-4755	BP LL 0209		
Melt flow index (MFI), g/600 s	1.1	1.1	0.85	190 °C/2.16 kg	ASTM D1238
High load MFI, g/600 s	57.9	50.3	24.8	190 °C/21.6 kg	ASTM D1238
Die swell ratio (SR)	1.43	1.46	1.11		
Density, kg/m^3	924.3	961.0	922.0	slow annealed	ASTM D1505
Crystallinity, %	40	67	40	DSC	
Temperature of fusion (max.), °C	110	131	122	DSC	
Vicat softening point, °C	93	127	101	5 °C/h	ASTM D1525
Short branches **	23	1.2	26	IR	ASTM D2238
Comonomer		butene	butene	NMR	
Molecular mass*					
M_w	200 000	136 300	158 100	SEC	
M_n	44 200	18 400	35 800	SEC	
Tensile yield strength, MPa	12.4	26.5	10.3	50 mm/min	ASTM D638
Tensile rupture strength, MPa	12.0	21.1	25.3		
Elongation at rupture, %	653	906	811		
Modulus of elasticity, MPa	240	885	199	flexure	ASTM D790
Impact energy,					
unnotched, kJ/m^2	74	187	72		ASTM D256
notched, kJ/m^2	61	5	63		ASTM D256
Permittivity at 1 MHz	2.28				ASTM D1531
Loss tangent at 1 MHz	100×10^{-6}				ASTM D1531
Volume resistivity, Ω · m	10^{16}				
Dielectric strength, kV/mm	20				

* Corrected for effects of long branching by on-line viscometry.
** Number of methyl groups per 1000 carbon atoms.

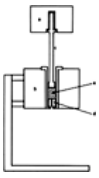

Figure 3. Melt flow index equipment

a) Interchangable piston loading weight; b) Electrically heated barrel; c) Piston; d) Die; e) Polyethylene melt

The behavior of polyethylene under shear is shown in Figure 4, which compares a LDPE and a LLDPE similar to those in Table 2. At sufficiently low shear rates the viscosity of all polyethylenes becomes Newtonian, i.e., independent of shear rate. Due to its narrow MMD the viscosity of the LLDPE is less shear dependent than that of the LDPE and has a higher viscosity under the higher shear conditions used in processing equipment.

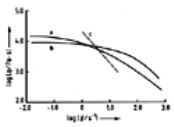

Figure 4. Dependence of viscosity η on shear rate $\dot{\gamma}$ for two polyethylenes

a) Low-density polyethylene; b) Linear low-density polyethylene; c) Apparent shear rate in melt flow index test

As indicated above, the crystalline properties are affected by the rate of cooling from the melt and the subsequent thermal history. For the purposes of reproducibility it is usual to apply a standard annealing treatment to test samples, such as annealing at 100 °C for 5 min followed by slow cooling to room temperature. The crystallinity correlates with the density, reflecting the average properties of the polymer, but the melting point and the softening points for LLDPE are higher than for LDPE due to the presence of some relatively sparsely branched species in the former type of polymer. In addition to the differential scanning calorimetry (DSC) technique used to measure the fusion point, the relative heterogeneity of the LLDPEs is demonstrated even more clearly by temperature rise elution fractionation (TREF) [29], [30]. After depositing the sample by slow cooling from solution onto an inert support, elution is carried out over a programmed temperature range to measure the concentration of eluted polymer as a function of elution temperature. Examples are shown in Figure 5. LLDPE produced with a single-site catalyst (see Section Single-Site Catalysts (Metallocenes)) shows a single sharp peak by this test.

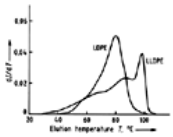

Figure 5. Temperature rise elution fractionation curves (*I* is the cumulative weight fraction)

The weight-average and number-average molecular masses determined by size exclusion chromatography (SEC, also known as gel permeation chromatography, GPC) are listed in Table 2, and Figure 6 shows their molecular mass distribution curves.

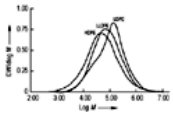

Figure 6. Molecular mass distribution curves for the polyethylenes of Table 2 (data from Repsol Quimica)

I is the cumulative weight fraction.

In the tensile test the yield strengths and the elastic moduli are as expected from the respective crystallinities. The higher rupture strength of the LLDPE relative to the LDPE is typical when the two polymers are chosen to have equal MFIs, but if the criterion were constant high-shear extrudability (e.g., the high load MFI) the tensile strengths would be very similar. The example chosen for the LLDPE is a butene copolymer. For higher performance applications longer chain olefins are used as comonomers [31]; e.g., octene (Du Pont, Dow), hexene (UCC, Exxon), 4-methylpentene (BP). These produce higher tensile strengths and impact energies, particularly in the form of film.

In common with other polymers, polyethylene is viscoelastic in the solid state. This means that the strain produced by applying a stress is time dependent, and in defining an elastic modulus it is important to specify the timescale of the measurement. Figure 7 shows the results of a measurement of strain versus time for a HDPE specimen at a constant tensile stress of 6.5 MPa [32]. The strain continues to increase approximately linearly with log *t* and after 4 d would have reached almost four times the value measured after 10 s. Thus the effective Young's modulus after 4 d is only 25 % of the short-term value and after several years would be lower still. This creep behavior is particularly marked in the case of polyethylenes because the amorphous regions are relatively mobile at room temperature [33]. For design purposes the modulus must be estimated for the timescale and temperature expected for the application. A related problem in designing for long-term use is that a prolonged high stress may lead to crack formation and failure at a stress significantly below the conventionally measured yield stress [34].

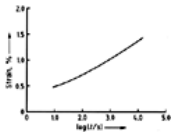

Figure 7. Tensile creep of HDPE (Rigidex 006–60) [32]

The extremely low loss tangent (power factor) makes LDPE ideally suitable as an insulator for high-frequency cables. Since pure polyethylene contains no polar groups, a very low power factor is expected, and most of the measured value is due to traces of oxidation, catalyst residues, antioxidant, etc. Because they contain Ziegler or Phillips catalyst residues, such low power factors are not generally achievable with HDPE or LLDPE.

Figure 8 shows the permeability of LDPE and ethylene – vinyl acetate copolymer (EVA) films to oxygen and water vapor. Although LDPE has a very useful role as a packaging material, the gas and water vapor permeabilities are not particularly low compared to other film forming materials (see ⟶). Since permeation takes place in the mobile amorphous phase, the permeability of HDPE is appreciably lower.

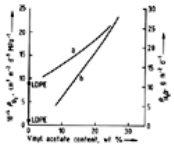

Figure 8. Permeability *P* of 0.025 mm films of LDPE and EVA at 23°C (data are based on Exxon Escorene Ultra EVA copolymers [35])

a) Oxygen; b) Water vapor (50 % R.H.)

Polyethylene is a high molecular mass hydrocarbon which can be considered as toxicologically inert, and indeed high-purity forms are used in medical prostheses. The suitability of a polyethylene for use in contact with food or in medical prostheses depends on its content of catalyst residues and principally on additives such as antioxidants. Acceptable limits on process residues and additives are normally controlled by national regulatory bodies, but in many cases they are based on the standards defined by the American Food and Drug Administration (FDA) or the German Bundesgesundheitsamt.

1.2.3. *High Molecular Mass (Bimodal) Polyethylene (HMWPE)*

This description is usually applied to HDPEs with MFIs in the range 0.01 – 0.1 by the normal test, although a higher load is usually used to characterize them. They are processable with some difficulty on conventional PE processing machinery, although some advantages can be gained by using equipment optimized for these grades. In general, they are not simply higher molecular mass versions of normal HDPE grades, but are specially designed broad-MMD polymers, often produced by using multiple reactors or combinations of catalysts [36]. The broad MMD, which may be bimodal, produces a highly shear dependent melt viscosity combined with good mechanical strength.

A particularly important type of bimodal polymer is produced by combining a high molecular mass component having a relatively high concentration of short branches with a low molecular mass component containing few or no branches. In this way the low molecular mass component crystallizes as folded chain lamellae and the high molecular mass component forms tie molecules which crystallize in several of the lamellae. This combination of branching and molecular mass is the reverse of what is usually observed in LLDPEs (see Sections Molecular Structure and Morphology and Single-Site Catalysts (Metallocenes)), where the more higly branched components have low molecular masses. The desired bimodal distribution is achieved by careful process modifications using multiple reactors. The optimization of the tie molecules confers excellent resistance to long term crazing and crack growth [37], making the polymers particularly suitable for pipe and liquid containers.

1.2.4. *Ultra High Molecular Mass Polyethylene (UHMWPE)*

These polyethylenes were commercialized shortly after HDPE and have extremely high molecular mass (3 – 6×10^6 according to ASTM D4020), but do not usually have a broad molecular mass distribution. The viscosity is too high to be measured by the MFI test, and they are usually characterized either by the relative solution viscosity in decalin at 135 °C, or by specially developed melt flow tests. The material is processed by techniques similar to those developed for PTFE, which involve fusing by sintering, rather than plastification to a true melt. UHMWPE has a remarkable combination of abrasion resistance, chemical inertness, low friction, toughness, and acceptability in contact with foodstuffs. Some properties are listed in Table 3. The density is only 940 kg/m^3, even though the polymer is unbranched, because the extremely high viscosity hinders the crystallization process.

Table 3. Properties of an ultra high molecular mass polyethylene (Hoechst GUR 412) [38]

Property	UHMWPE	Method	Standard
Molecular mass	4.5×10^6	viscometric	DIN 53 728
Reduced specific viscosity, cm^3/g	2300	0.05 % in decalin	
Density, kg/m^3	940		DIN 53 479
Crystalline melting range, °C	135 – 138	polarizing microscope	
Tensile yield strength, MPa	22		DIN 53 455
Tensile rupture strength, MPa	44		
Elongation at rupture, %	> 350		
Impact energy, notched, kJ/m^2	210	double V-notch	DIN 53 453
Volume resistivity, Ω · m	> 10^{13}		DIN 53 482

1.2.5. *Properties of Ethylene Copolymers*

Table 4 lists the principal types of ethylene copolymers which are in commercial production. Of these the vinyl acetate copolymers are produced in the largest quantities. All the ester and acid copolymers are produced by the high-pressure free-radical process, but most of the plants in the rapidly expanding LLDPE sector have the capability to produce VLDPE copolymers with higher content of α-olefin and this is an area with a large potential for expansion.

Table 4. Principal types of ethylene copolymer

Comonomer	Abbreviation	Feature	Catalysis [*]
Vinyl acetate	EVA	flexibility	FR
Methyl acrylate	EMA	flexibility, thermal stability	FR
Ethyl acrylate	EEA	flexibility at low temperature, thermal stability	FR
Butyl acrylate	EBA	as for EEA	FR
Methyl methacrylate	EMMA	flexibility, thermal stability	FR
Butene	VLDPE	flexibility at low temperature, thermal stability	Z, SS
Hexene	VLDPE	as for butene copolymer	Z, SS
Octene	VLDPE	as for butene copolymer	Z, SS
Acrylic acid	EAA	adhesion	FR
Methacrylic acid	EMAA	adhesion	FR
Methacrylic acid + Na^+ or Zn^{2+} (ionomer)		adhesion, toughness, stiffness	FR
Acrylic acid + Zn^{2+} (ionomer)		adhesion, toughness, stiffness	FR
Carbon monoxide	PK	polyketone, stiffness, high melting	[**]
Norbornene	COC	cycloolefin copolymer, transparency	SS

[*] FR = free radical, Z = Ziegler catalysis, SS = single-site catalyst.
[**] Novel group VIII metal catalyst.

The principal effect of copolymerization is to reduce the crystallinity. The effect is approximately the same for all comonomers on a molar basis, with the exception of propene which can be incorporated into the crystal lattice. At room temperature the amorphous regions of the copolymer (where the comonomer groups are concentrated) are mobile and the effect of introducing a comonomer is to progressively reduce the stiffness (Fig. 9). In addition to the branches due to the comonomer, the branches which occur under the high pressure synthesis conditions also contribute to the reduction in crystallinity. Below room temperature there are differences between ester copolymers due to the differences in the glass transition temperature T_g of the amorphous material. This follows the T_g of the ester homopolymer and results in EEA and EBA copolymers being flexible down to lower temperatures than EVA, with EMMA having the highest T_g. The VLDPEs produced by the LLDPE processes should have the best low-temperature properties, particularly in the case of narrow composition distribution polymers. Properties of three typical EVAs and a VLDPE [42] are listed in Table 5.

Table 5. Properties of ethylene copolymers

Property	EVA (Repsol PA-501)	EVA (Repsol PA-538)	EVA (Repsol PA-440)	VLDPE (DSM TMX 1000) [a]	Condition	Standard
Vinyl acetate content, wt %	7.5	18	28			
Melt flow index, g/600 s	2	2	6	3	190 °C/2.16 kg	ASTM D1238
Density, kg/m^3	926	937	950	902	rapid annealed	ASTM D1505
Vicat softening point, °C	83	64		66		
Tensile strength, MPa	16	16	11	11.5	50 mm/min	ASTM D638
Elongation at rupture, %	700	700	800		TMX 1000: 0.4 m/s	
Modulus of elasticity, MPa	156[b]	47[b]	24[b]	95[c]		
Permittivity at 1 MHz	2.46	2.70				ASTM D1531
Loss tangent at 1 MHz	0.014	0.035				ASTM D1531
Volume resistivity, Ω · m	2.0×10^{15}	2.5×10^{14}				
Dielectric strength, kV/mm	19	20				

[a] Octene copolymer.
[b] 0.2 % strain, 100 s.
[c] ASTM D790.

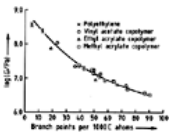

Figure 9. Dynamic modulus *G* (ca. 2 Hz) at 23°C as a function of total branch points

The methacrylic and acrylic acid copolymers are produced to provide enhanced adhesion, particularly in coextruded films or laminates. Not included in Table 4 are terpolymers in which acid monomers are used together with ester comonomers to improve adhesive properties.

The ionomers are produced by the partial neutralization of acidic copolymers containing 10 wt % of (meth)acrylic acid by sodium or zinc ions. The ionic salts and the unneutralized acid groups form strong interchain interactions, producing a form of thermally labile cross-linking in both the solid and the molten states [43], [44]. The T_g of the amorphous material in ionomers is slightly above room temperature, resulting in a stiffness similar to that of LDPE at room temperature. However, the stiffness of most ionomers decreases rapidly with increasing temperature above 40 °C. The polyketone (PK) and cycloolefin (COC) copolymers are both recent developments in which ethylene is present as a 50 mol % alternating copolymer [45]. The PK materials are believed to be made with a novel group VIII catalyst [46]. They are targeted at medium-stiffness applications and also have good barrier properties for hydrocarbons. The carbonyl group, however, makes the polymers susceptible to degradation by sunlight. The COC materials from Hoechst make use of the remarkable catalyst activity and copolymerization ability of the new single site catalysts [47] (Section Single-Site Catalysts (Metallocenes)) and the high transparency and low birefringence will allow its use in compact discs and transparent packaging.

1.3. Polymerization Chemistry

Heat of Reaction. The heat of polymerization of ethylene is 93.6 kJ/mol (3.34 kJ/g). Since the specific heat of ethylene is 2.08 J °C^{-1} g^{-1}, the temperature rise in the gas phase is ca. 16 °C for each 1 % conversion to polymer. Heat removal is thus a key factor in a commercial polymerization process. Some processes (e.g., ICI autoclave, Unipol fluidized bed) employ only a limited conversion per pass, and the heat of reaction is absorbed by the cool reactants. The unreacted monomer is then cooled in the recycle stage. In other cases (e.g., UCC or BASF tubular and various slurry processes) more surface area or more residence time is provided, and heat is removed through the reactor walls.

1.3.1. *Free-Radical Catalysis*

1.3.1.1. *Introduction*

Free-radical catalysis is used exclusively in the high-pressure process, that is at pressures above 100 MPa. The reason for the use of such high pressures is a combination of historic, economic, and technical factors [39]. Because ethylene is gaseous above its critical temperature of 9 °C, a pressure of ca. 20 MPa would be necessary in any case to achieve a reasonable concentration of monomer. Employing pressures of ca. 200 MPa and temperatures above 160 °C enables the polyethylene produced to dissolve in the unreacted ethylene, and the high reaction rate makes the best use of the very expensive high-pressure equipment. 20 % conversion of the monomer is typically achieved in 40 s. A schematic phase diagram for an ethylene – polyethylene system is shown in Figure 10 [48]. For a more detailed account of the effect of molecular mass and MMD on the phase equilibria, see [49].

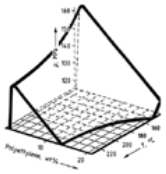

Figure 10. Schematic cloud point surface for ethylene – polyethylene [48]

The single-phase ethylene – polyethylene mixture allows the reaction to take place as a classical free-radical-initiated solution polymerization (⟶ Polymerization Processes). Some aspects which are particularly important for ethylene systems are as follows:

1. In addition to the effect of concentrating the gaseous monomer, the pressure also influences the reaction rate constants, as is also the case for liquid systems subjected to high pressures [50]. This is generally considered in terms of a volume of activation, analogous to the energy of activation. High pressure affects the configuration change necessary for the reactants to reach the transition state. The overall contribution of the effect of pressure on the rate constants over the pressure range 0 to 200 MPa is to increase the polymerization rate by a factor of ca. 12 [51].
2. The reaction temperatures employed are also high, an average of ca. 220 °C being typical. In parts of the reactor the temperature may be as low as 140 °C but in other parts may reach over 300 °C. These high temperatures also contribute to the high reaction rates, the activation energy being 32 kJ/mol.
3. The growing polymer chains are linear alkyl radicals and as such are very reactive not merely in the addition to double bonds, but also in abstracting hydrogen atoms from other molecules, thereby forming saturated alkyl chains and new radicals. The process is called chain transfer. Since these hydrogen atom abstraction reactions have higher activation energies than polymerization, they become increasingly important as the polymerization temperature rises. On the one hand they put stringent demands on the monomer purity to avoid traces of compounds which could give rise to chain transfer and thereby reduce the molecular mass, and on the other hand it is possible to use low concentrations of suitable materials (so-called chain-transfer agents or modifiers) to control the polymer molecular mass. Chain-transfer agents which have been used commercially include hydrogen, propane, propene, acetone, and methyl ethyl ketone. Chain transfer to some compounds with very active hydrogen atoms such as propene, and particularly the higher alkenes, can lead to radicals which are insufficiently reactive towards ethylene to reinitiate new chains rapidly, and reduced reaction rates result.
4. Chain transfer can also occur with the polyethylene chains themselves, either to the same growing chain (intramolecular) or to other polyethylene chains (intermolecular transfer). These reactions create the characteristic structural features of LDPE which distinguish it from HDPE. As can be demonstrated with models, the most probable intramolecular chain-transfer reaction is to the carbon atom four carbons back down the chain, which produces butyl groups (Fig. 11 A). This so-called back-biting mechanism was first published by M. J. Roedel of Du Pont [52]. If, after the addition of one ethylene molecule to the newly formed secondary radical, a further back-bite occurs (Fig. 11 B), a pair of ethyl branches or a 2-ethylhexyl group is formed. A further possibility, shown in Figure 11 C, is that a back-bite occurs to a branch point, and the tertiary radical then decomposes into a new short radical, leaving a vinylidene group at the end of the polyethylene chain. This process is the principal chain-termination mechanism in LDPE, and concentrations of vinylidene groups approach one per number average molecule for LDPEs produced at high temperature. These three reactions account for the principal features observed in the infrared spectrum of LDPE, but to a lesser extent other intramolecular transfers also occur [9]. Since the activation energy for transfer is higher than for the polymerization reaction and the activation volume is smaller, branching and unsaturation increase with increasing polymerization temperature and decrease with increasing reaction pressure [53], [54].
5. Intermolecular transfer leads to long branching and broadening of the molecular mass distribution. Since they have more hydrogen atoms available for chain transfer, the long chains tend to be the most highly branched, and there may be branches on the branches. Statistically each new radical produced by initiation or chain-transfer reactions has a range of probabilities of growing to various lengths before being terminated by one of these same chain-transfer or radical combination reactions. The probable length is the same whether the chain grows from a new initiating radical or a branch point. Thus branches are statistically the same length as the backbone itself and, taking into account statistical variations, the branches may in some cases be longer than the initial backbone. Since the long-branching reaction is a chain-transfer mechanism, the average chain length of an unbranched molecule or a branch is shorter than it would be in the absence of the long-branching reaction. The effects of temperature and pressure are similar to those for short branching, but additionally the amount of long branching is proportional to the concentration of dissolved polymer. In principle, this leads to a clear difference between plug flow (tubular) and continuous stirred-tank reactors (CSTRs, autoclaves). Theoretical analyses have been presented for autoclave [55], [56], [57] and tubular [58] reactors.
6. Initiation is very similar to that in many other free-radical polymerizations, but there are some limitations. Initiators are commonly referred to as “catalysts”. In the sense that one mole of initiator will achieve the polymerization of several thousand moles of ethylene this is so, but the initiator is destroyed in the process and so the term is, strictly speaking, incorrect. Oxygen was used as initiator in the early commercial processes because of the ease of introducing it into the process. With the development of high-pressure pumps and new initiators, modern plants are able to maintain more precise control of temperature profiles by the injection of solutions of liquid catalysts. The mechanism by which oxygen forms free radicals is rather complicated, and at lower temperatures oxygen can act as an inhibitor [59]. In the autoclave process the use of oxygen has been largely superseded, but in the tubular reactor process it is still widely used, sometimes in combination with liquid initiators. The two overall limitations on initiators are that they should be readily soluble in alkanes and they should produce reactive radicals, ideally alkyl or alkoxy radicals. On the former count, all the aqueous systems and most azo compounds are excluded, and on the second count dibenzoyl peroxide is unsuitable. The initiators are selected for use on the basis of their half-lifes at the reaction temperature. Since the residence time in the reactor zone may be of the order of 20 s or less, to obtain good control of the reaction rate, an initiator half-life of about 1 s is required. For a tubular reactor the same initiator is active over a wider range of temperatures, but its selection is equally critical. Typical initiators are listed in Table 6 [60].
7. The mechanism of kinetic chain termination is by combination of radicals. This further widens the MMD in long-branched systems when the rate of inititation-combination is high [57].
8. Although the conversion of ethylene to polyethylene is thermodynamically favorable, the decomposition into carbon and a mixture of methane and hydrogen is also highly exothermic:

$$C_2H_4 \rightarrow C + CH_4 \; \Delta H = 127\,\mathrm{kJ/mol}$$

$$C_2H_4 \rightarrow 2\,C + 2\,H_2 \; \Delta H = 53\,\mathrm{kJ/mol.}$$

Table 6. Peroxide initiators for the high-pressure process [60]

Peroxide	Formula	Temperature for half-life of 1 s, °C	Storage temperature, °C
Di-(2-ethylhexyl) peroxydicarbonate	$(C_4H_9)(C_2H_5)CH-CH_2-O-C(=O)-O-O-C(=O)-O-CH_2-CH(C_2H_5)(C_4H_9)$	130	< − 15
tert-Butyl 2,2-dimethylperoxypropanoate	$(CH_3)_3C-O-O-C(=O)-C(CH_3)_2-CH_3$	154	< 0
Di-(3,5,5-trimethylhexanoyl) peroxide	$(CH_3)_3C-CH_2-CH(CH_3)-CH_2-C(=O)-O-O-C(=O)-CH_2-CH(CH_3)-CH_2-C(CH_3)_3$	173	− 10 to 0
tert-Butyl 3,5,5-trimethylperoxyhexanoate	$(CH_3)_3C-O-O-C(=O)-CH_2-CH(CH_3)-CH_2-C(CH_3)_3$	211	< + 30
Di-*tert*-butyl peroxide	$(CH_3)_3C-O-O-C(CH_3)_3$	244	< + 30

Figure 11. Principal intramolecular chain-transfer reactions

For kinetic reasons these reactions are only important at the high temperature and pressures of the high-pressure process. In a confined system the large amount of heat released can raise the temperature, and hence the pressure, of the methane and hydrogen to potentially dangerous levels. The theoretical final temperature and pressure for a contained decomposition starting at 250 °C and 200 MPa are 1400 °C and 620 MPa. In practice much of the heat would be absorbed by the walls of the vessel or pipework. High-pressure plants are designed with relief valves or bursting disks to protect the equipment from overpressurization due to decomposition. Decompositions usually start as a runaway polymerization reaction, but then they can propagate as a slow flame front even into cold gas. Experimental decompositions usually show low propagation velocities of ca. 0.2 m/s [61], but under the more turbulent conditions of commercial plant operation, propagation can be more rapid.

1.3.1.2. *Copolymerization*

High-pressure ethylene copolymerization follows the classical free-radical copolymerization mechanism (⟶ Polymerization Processes). One of the most important characteristics are the reactivity ratios r_1 and r_2. The rates v of the four growth reactions are

$$-R_1^\bullet + M_1 \rightarrow \; -R_1^\bullet \quad v_{11} = k_{11}[R_1][M_1]$$

$$-R_1^\bullet + M_2 \rightarrow \; -R_2^\bullet \quad v_{12} = k_{12}[R_1][M_2]$$

$$-R_2^\bullet + M_1 \rightarrow \; -R_1^\bullet \quad v_{21} = k_{21}[R_2][M_1]$$

$$-R_2^\bullet + M_2 \rightarrow \; -R_2^\bullet \quad v_{22} = k_{22}[R_2][M_2]$$

$$r_1 = k_{11}/k_{12} \; r_2 = k_{22}/k_{21}$$

$-R_1^\bullet$ represents a growing radical where the last unit added is monomer 1 (ethylene for the purpose of this discussion).

For low concentrations of monomer 2 it follows from the well-known copolymerization equation that:

$$1/r_1 = \frac{\text{concentration of monomer 2 in copolymer}}{\text{concentration of monomer 2 in reactor}}$$

The value of r_1 can thus give guidance on the effectiveness of incorporation of comonomer relative to the feed concentration. Some measured values are shown in Table 7. A more comprehensive compilation is given in [9]. Vinyl acetate has a reactivity ratio of almost exactly 1.0 which means that the copolymer has the same composition as the reactor feed. Acrylates on the other hand have r_1 much lower than 1 which means that the copolymer is much richer in acrylate than the reactant mixture. This has implications for the type of reactor. A continuous stirred-tank reactor (CSTR) operates with the comonomer concentrations in a dynamic equilibrium. To make a copolymer containing 20 wt % ethyl acrylate, the feed composition would contain typically 4 % of ethyl acrylate, but the reactor and the stream leaving the reactor would contain only 1.6 %. In the case of a continuous plug flow reactor (CPFR), corresponding to a tubular reactor or a laboratory batch reactor, the corresponding figures would be 3.1 and 0.7 %. The copolymer would in this case be a continuous blend of compositions ranging from 38 % to 9 % produced as the comonomer was used up progressively along the reactor. A useful feature of r_1 for vinyl acetate being 1.0 (and r_2 is also 1.0) is that values of reactivity ratios reported in the literature for other monomers copolymerizing with vinyl acetate can be used to estimate the reactivity ratios for ethylene at high pressure [39]. This is illustrated in Table 7.

Table 7. Copolymerization reactivity ratios and chain transfer [a]

Comonomer M_2	r_1 (rel. to ethylene)	r_1 (rel. to VA) [b]	Chain transfer [c]
Vinyl acetate (VA)	1.05		0.7
Methyl acrylate	0.13	0.1	0.4
Ethyl acrylate	0.08		3.3
Methyl methacrylate	0.06	0.015	2.2
Methacrylic acid	0.015	0.01	
Styrene	0.01	0.01	
Vinyl chloride	0.18	0.23	> 10
Propene	1.3		1.1
Butene	2.0	2.0	20
Isobutene	1.0		2.1

Hexene	1.4	> 20

[a] Reactivity ratios and chain transfer measured in a continuous reactor at 190 °C and 157 MPa, except for the olefins where the data are from a batch reactor at 140 °C and 196 MPa.
[b] Literature values from [40], [41].
[c] d[log(MFI)]/d[M_2] in decades of MFI change per mol % in reaction mixture.

Comonomers also act as chain-transfer agents. Table 7 shows some practical measurements of these effects on the MFI under standard conditions. Reference [9] tabulates the chain-transfer effects in terms of molecular mass reduction. Clearly a moderate chain-transfer activity can be tolerated more readily for a comonomer with a low r_1 since the concentration in the reactor will be lower. The chain-transfer activity of the higher α-olefins such as butene is very high, precluding their use in the free-radical process. The radicals produced by chain transfer are relatively stable and retard the reaction. Propene is less extreme in these effects and is sometimes used in the high-pressure process.

Copolymerization theory shows that if r_2 is greater than 1 (it usually approximates to $1/r_1$), then the overall reaction rate is reduced. In a continuous process this translates to an increased initiator demand which in extreme cases makes the process inoperable. As indicated above for homopolymers, a high initiation rate broadens the MMD. Thus acrylate and methacrylate copolymers require higher initiator injection rates and produce hazier films due to the wider MMD.

1.3.2. *Coordination Catalysis*

The three independent discoveries of low-pressure routes to linear polyethylene had one thing in common: they used catalysts containing transition metals. Despite their very different methods of preparation, there is general agreement that the basic mechanism of polymerization by these catalysts is the same. At some stage a σ-bonded alkyl group is formed. Ethylene is coordinated to the transition metal by a π-bond. This then facilitates the insertion of the ethylene molecule into the metal – alkyl bond producing a longer chain alkyl and a vacant coordination site.

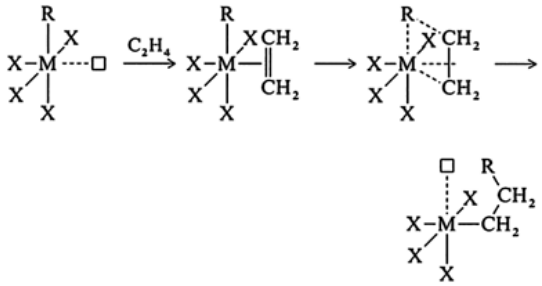

The direct insertion of ethylene into σ-bonded aluminum alkyls was discovered by ZIEGLER in 1950, but the reaction was slow and did not lead to high polymers. The prior coordination of ethylene is clearly crucial. The active species in commercial catalysts is complex, but there is a very large worldwide activity in investigating coordination catalysts, usually by way of model compounds [16], [17]. Commercial catalysts are virtually all heterogeneous solids (at least on a microscopic scale) and require careful attention to the particle shape and size in the development of the manufacturing process. A feature of modern coordination catalysts (except those used in solution processes) is that the catalyst particles grow by a process of replication [13]. This means that the overall shape of each particle is maintained as it grows by polymerization, and thus the distribution of polymer particle sizes is related to the distribution of catalyst particle sizes. For this type of growth to occur, the polymerization process must break down the catalyst particles into much smaller entities which remain held together by the polymer formed, sometimes in the form of fibrils (Fig. 12). Electron microscopy has shown, for one type of supported catalyst at least, that the growth occurs in the form of cylinders, in the growing end of which a catalyst fragment is embedded[62]. This type of growth explains why polymerization rate does not decrease as the overall particle size increases, since the monomer diffusion path remains short. The active catalyst fragments range in size from about 4 nm for $TiCl_3$ to 100 nm for oxide-supported catalysts.

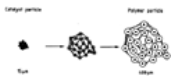

Figure 12. Catalyst particle growth by replication [13]

By suitable technology, catalysts can be prepared so that replication leads to polymer particles with a spherical shape and a diameter of ca. 1 mm, suitable for direct use without a pelletization step. Examples of this are the Phillips Particle Form and UCC's Unipol processes based on silica-supported catalysts, and the Montedison Group's Spheripol catalyst based on $MgCl_2$. Outside the United States there is considerable resistance by fabricators to the direct use of polymer powder, and much of the product is pelletized.

Not being simple isolated molecules, the active catalyst sites are not all identical and their activities and the average chain lengths which they produce vary. Instead of a simple statistical distribution of molecular mass, which would lead to a ratio of weight-average to number-average molecular mass of 2.0, appreciably broader MMDs are obtained (Fig. 13).

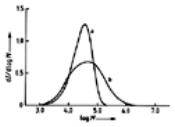

Figure 13. MMD curve of a Ziegler HDPE compared with simple statistical theory

a) Most probable distribution; b) Hostalen GD-4755

1.3.2.1. *Phillips Catalysts*

A typical Phillips catalyst is produced by impregnating silica particles with a solution of CrO_3 to give a chromium content of ca. 1 %. The powder is then calcined in a current of air with increasing temperature to a final value of ca. 800 °C. At this temperature all of the physically absorbed water and most of the surface hydroxyl groups are driven off [63] and the chromium is present as surface chromate. The formation of a surface silyl chromate

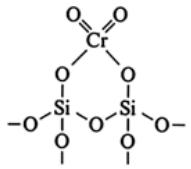

is significant, because at the calcining temperature CrO_3 would decompose to lower valent oxides. This catalyst, which is now moisture sensitive, will polymerize ethylene, but with an induction period. During the initial reaction with ethylene the catalyst's color changes from orange to blue. This is believed to be due to the reduction to Cr^{2+}. The reduction can also be brought about by treatment with carbon monoxide, and in this case the catalyst reacts immediately with ethylene without an induction period.

The selection and treatment of the support is fundamental to the process, and a plant may use catalysts made from a variety of supports to produce the whole range of products. The optimum silica is claimed to have a high percentage of pores with a pore diameter of 20 – 50 nm. A frequently used grade, Grace Davison 952, combines the required pore characteristics with a microspheroidal form. Where the spherical form is not necessary, other silicas are also used, made from crushed and graded silica gel with the required pore structure. The supports are normally supplied with the chromium already impregnated, leaving only the activation stage to be carried out at the polyethylene plant.

As a means of increasing the catalyst productivity and facilitating the production of lower molecular mass (higher MFI) polymers, treatment with titanium compounds is

frequently used [64]. The resulting polymers have MFIs more suitable for many applications and also have somewhat broader MMDs. Molecular mass control is a problem with the Phillips process, since it is not possible to use a chain-transfer agent such as hydrogen, which is oxidized to water by the chromate groups and acts as a catalyst poison. Molecular mass control is therefore effected mainly by the choice and treatment of the support. The use of titanium compounds is believed to result in the chromium atoms being bound to the support via titanate bonds [65]. A further parameter available for modifying the surface characteristics is treatment with fluorine compounds which convert surface hydroxyl groups to fluoride and reduce the surface area. To some extent the effect is similar to high temperature calcination [65], [66].

Catalyst productivities are of the order of 5 kg PE per gram of catalyst [63] or higher, with a corresponding chromium content of 2 ppm or less. The percentage of chromium atoms which form active polymerization centers has been estimated as 12 % [67]. With a number-average chain length of the order of 1000 monomer units, each chromium atom thus produces about 1000 molecules. The chains are terminated by a β-hydrogen shift reaction:

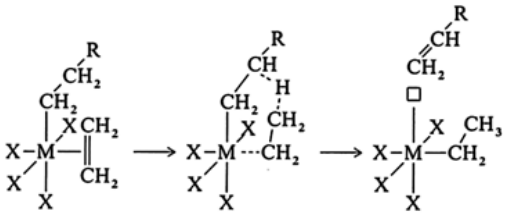

forming an unsaturated molecule and an ethyl ligand attached to the chromium atom [66].

The primary form of molecular mass (and hence MFI) control is by selection of a catalyst which favors the hydrogen shift reaction, but fine adjustments can be made by varying the reaction temperature, since a higher temperature favors the shift reaction.

There have been two developments of chromium catalysts by UCC which fall conceptually in the field of Phillips catalysts. CARRICK et al. reported that bis(triphenylsilyl) chromate, which is closely related to the proposed active site of a Phillips catalyst, polymerizes ethylene at high pressure [69]. When supported on silica it forms a very active catalyst for low-pressure polymerization [70]. The second type of catalyst is formed by the reaction of chromium compounds having π-bonded ligands with the hydroxyl groups on silica [16], [17], [71], [72]. Paricularly favored is dicyclopentadienyl chromium (chromacene). Unlike the Phillips catalyst it is believed that the chromium is attached to the support by only one bond, with one bond remaining to a cyclopentadienyl group. A very useful feature of these catalysts is their sensitivity to hydrogen, which allows a wide range of molecular masses to be produced. The MMD produced is fairly narrow, but not as narrow as that from some of UCC's Ziegler catalysts.

Somewhat similar to the Phillips catalyst is the Standard Oil (Indiana) catalyst which was the first of the coordination catalysts to be discovered [73]. It typically consists of MoO_3 supported on alumina or silica and calcined in air at high temperature. Unlike the Phillips catalyst it is necessary to reduce the precursor with hydrogen at elevated temperature before using in the polymerization reactor. Despite extensive development it has not been widely commercialized.

1.3.2.2. *Ziegler Catalysts*

The range of catalysts which function by the Ziegler mechanism is extremely broad and it would be impossible to cover all the variants here. However, those catalysts which meet the requirements of modern polyethylene processes are more restricted in number and tend to follow a common pattern. Some of the developments in Ziegler catalysts for polyethylene have arisen out of developments for polypropylene (see Section Historical Survey), but since ethylene has a higher reactivity and the extra parameter of stereoregularity does not have to be dealt with, the requirements of the polyethylene catalysts have generally been met more easily.

Early Ziegler catalysts for ethylene were based on β-$TiCl_3$, produced by reducing $TiCl_4$ with $AlEt_2Cl$ or $Al_2Et_3Cl_3$ at low temperature. As cocatalyst, $AlEt_3$ was used. The $TiCl_3$ produced in this way contains $\frac{1}{3}$ mole of cocrystallized $AlCl_3$. Although the catalyst productivities were much higher than for propene, it was necessary to extract the catalyst residues in order to reduce the quantities of Ti and Cl in the product to acceptable levels. Later, by the use of longer chain aluminum alkyls as cocatalyst and alkoxytitanium chlorides as transition metal compound [74], [75], the catalyst productivity was improved sufficiently to allow the costly residue removal stages to be eliminated.

Kinetic studies have shown [17], [67] that only a small proportion of the titanium atoms in $TiCl_3$ catalysts form active centers (typically < 1 %). This is believed to be due to the fact that, even with the small polymerizing particles, only a small fraction of the titanium atoms are at the crystallite surface. In order to make a higher proportion of the titanium atoms available to form active centers, various developments were made to support the transition metal compound on a carrier. Early attempts to support $TiCl_4$ directly onto silica, alumina, or magnesia did not lead to a sufficient increase in productivity [74]. The first useful high-yield catalyst used Mg(OH)Cl as support [76]. Since then a variety of magnesium compounds have been used successfully as supports [17], [74], but preeminent amongst these is $MgCl_2$ or reaction mixtures which can produce this compound, at least on the support surface. Most modern Ziegler processes seem to use catalysts which fall into this category. The massive increase in activity of $MgCl_2$ supported catalysts has been claimed to be due to an increase in the percentage of titanium atoms forming active centers (approaching 100 %) and not to a significant increase in reaction rate at the active center [16], [67]. The $MgCl_2$ supported catalysts produce polyethylenes with narrower MMDs than unsupported catalysts [16] and a narrower distribution of composition in the case of LLDPEs. As with polypropylene, electron donors such as esters or THF may be employed to modify the characteristics of the $MgCl_2$ based catalysts [16], but are not invariably used in the case of polyethylene.

For some processes, particularly the fluidized-bed and loop reactors, the catalyst shape and size is very important, and procedures such as ball-milling and chemical reaction used to produce the basic catalysts do not lead to particle sizes and shapes suitable for direct use. Further processing is required to produce a defined particle size range. Spray-drying can be used to produce spherical paricles. UCC have filed patents on the use of the microspheroidal silicas used in the Phillips process for supporting $MgCl_2$ – $TiCl_4$ systems for their gas-phase process [77]. For solution processes, as small a particle size as possible is desirable, and this may be achieved either by the method of catalyst preparation [78] or by prepolymerization treatment with a higher olefin [79].

The molecular mass of polyethylene is normally controlled by the use of hydrogen. With Ziegler catalysts at 70 – 100 °C this can require 20 mol % hydrogen in the gaseous phase of a slurry process or a gas-phase process. When other olefins are copolymerized with ethylene the concentration of hydrogen required is lower, due to chain transfer to the comonomer. At the high temperatures employed in the high-pressure process, the hydrogen concentrations used are much lower due to a relative acceleration of the rate of chain transfer to hydrogen and also a greater proportion of chain transfer by the β-shift reaction. Chain transfer to monomer or comonomer by the β-shift reaction results in unsaturation, primarily in the form of vinyl groups, but hydrogen chain transfer forms methyl-terminated chains.

1.3.2.3. *Single-Site Catalysts (Metallocenes)*

Typical Phillips or Ziegler catalysts do not follow classical polymerization kinetics and the resulting MMDs are substantially broader than the simple statistical case (see Section Coordination Catalysis). For homopolymers this is probably an advantage and gives rise to the good processing characteristics of HDPE. In the case of copolymers this type of kinetic behavior leads to a broad distribution of composition, in which some chains have low comonomer concentrations and others high concentrations. For LLDPE there are advantages such as higher stiffness than LDPE of the same density, but disadvantages of higher extractable fractions and stickiness in the lower density versions. A catalyst which could produce a narrow composition distribution would overcome these disadvantages, and at the same time less of the expensive comonomer would be required for a given density reduction.

Catalysts which behave in a uniform manner to produce simple statistical distributions of molecular mass and composition have been known for many years, but they are low-yield systems and have not been commercialized for polyethylene [80]. Such catalysts, made from vanadium oxide trichloride [80] and bis(cyclopentadienyl)titanium dichloride [81], remain soluble in the presence of the cocatalyst.

In 1983 KAMINSKY described a catalyst which combined ideal MMD and composition distribution with high yield [82], [83]. A basic form of the catalyst was bis (cyclopentadienyl)dimethylzirconium with a massive excess of methyl aluminoxane $(CH_3 – Al – O)_n$ as cocatalyst. Al/Zr ratios as high as 10 000 or more are typically employed. Since then there has been intensive development in academia and industry with the objectives of improving the thermal stability of such catalysts, reducing the concentration of cocatalyst, and, above all, making supported versions.

Because the original catalysts were bis(cyclopentadienyl) transition metal compounds, known as metallocenes, there is a tendency to refer to these recent developments as “metallocene catalysts”, but a more general term is “single-site catalysts”. Given that metallocenes had previously been investigated as catalysts with conventional alkylaluminum compounds as cocatalysts [84], it could be said that the key discovery was the use of methyl aluminoxane (MAO) as cocatalyst. The latter is made by controlled hydrolysis of trimethylaluminum and has a rather ill-defined structure. To avoid too rapid a reaction leading to aluminum hydroxide and unreacted trimethyl, SINN et al. initially used the water of crystallization of materials such as copper(II) sulfate pentahydrate as the source of water [85]. Later developments have led to physical methods for the controlled hydrolysis and the material is commercially available.

Since the transition metal components of the catalysts are stable entities with well-defined structures, investigators have been able to make systematic changes to the structure and achieve the sort of control over the polymerization process which had been hoped for in the early days of Ziegler catalysis. Principal achievements include stereospecific polymerization (see Chapter Polypropylene), higher reactivity towards other olefins and also towards other monomers not industrially polymerizable by Ziegler catalysis. It is generally accepted that the active catalytic species is a transition metal cation associated with an aluminoxane counteranion [19], [20]. Some investigators believe that the function of the MAO is to act as a source of free trimethylaluminum and also to act as a receptor of the anion produced by the reaction between the catalyst and cocatalyst [86], [87]. Polymerization proceeds more rapidly in aromatic solvents and even more rapidly in chlorinated aromatic solvents, and this supports the view that a higher dielectric constant aids the formation of an optimal ion pair [87]:

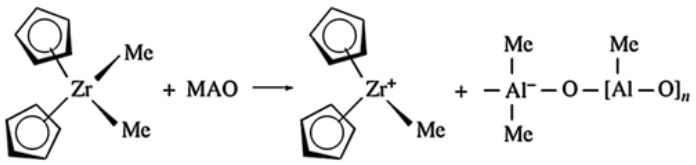

Both metallocene dichlorides and dimethyl forms of the catalysts are frequently described, but MAO transforms the former into methyl-substituted species. Chemical changes have been generally directed to modifying the stereochemical environment of the active center and include changes such as replacing the cyclopentadienyl groups with indenyl or fluorenyl, alkyl substitution of the cyclopentadienyl rings, and linking two indenyl units to form a more rigid cage. Alkyl substituents can have a major effect on the catalyst activity [88]. Single-site catalysts have also been made which have only one cyclopentadienyl ligand [89]. Some general forms of single-site catalysts include:

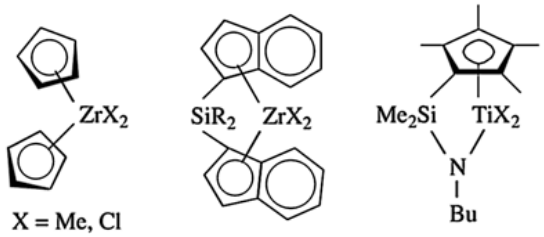

The active sites are tetrahedrally coordinated, in contrast to heterogeneous Ziegler catalysts which are octahedrally coordinated in a $TiCl_3$ or $MgCl_2$ crystal lattice. In principle the greater openness of the tetrahedral configuration allows more ready access to bulkier monomers than the octahedral structure, but the ligands themselves are generally rather large. The improved incorporation of monomers larger than ethylene is generally only achieved by bridged systems, in which the opening angle can be greater than the 109° of a regular tetrahedron. Dow refer to their catalysts as “constrained geometry” catalysts and specify that they reduce the bond angles of the coordinated ligands to less than 109°, so that a greater solid angle is available for the entry of monomers to the free coordination site [90].

Developments in cocatalysts have been directed to reducing the ratio of cocatalyst to transition metal and developing better defined chemical entities as cocatalysts. Montell have shown that highly active aluminoxanes can be formed from trialkylaluminum compounds other than the methyl derivative, and that in some cases they are even more active than MAO [91]. When made from higher molar mass aluminum alkyls, well-defined cocatalyst compounds such as tetraalkyldialuminoxanes can be isolated.

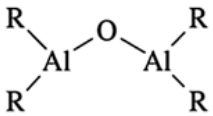

The large excess of MAO found to be necessary to obtain optimum activity in many studies on metallocene catalysts may be due in part to moderately high molar concentration of the cocatalyst needed to generate the active centers. When catalyst and cocatalyst are precontacted at high concentrations and then diluted for polymerization, Al/Zr ratios as low as 20 can be effective [92]. Apart from aluminum-based cocatalysts, boron compounds have been used and allow much lower cocatalyst/catalyst molar ratios to be achieved [93]:

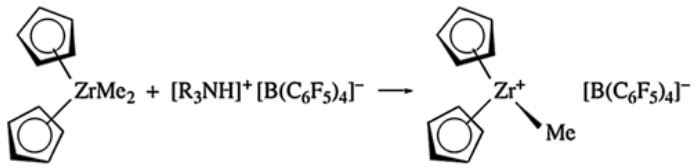

The general aspects of ethylene polymerization with metallocene catalysts are similar to those of Ziegler catalysis: molecular mass control can be effected by chain transfer to hydrogen; at higher temperatures chain transfer by the β-shift reaction produces vinyl groups; metallocene catalysts tend to be thermally labile and decompose to inactive species [87].

Industrially, the fact that the original metallocene catalysts were soluble meant that the immediate applications were in solution processes. Exxon took the view that the catalysts were short-lived and expensive and developed a high-pressure process to give a high catalyst yield [94], [95]. Product targets were VLDPEs with a lower content of extractables and less tackiness. Dow, which operates a solution process for HDPE and LLDPE, has developed high-temperature catalysts for producing a range of homopolymers and copolymers. They claim improved processability for these products due to incorporation of long branches by copolymerization of vinyl end groups of polymer chains formed by β-scission [90]. The solution process favors end-group polymerization because the high temperature leads to formation of a high proportion of chain ends by β-scission; the polymer to monomer concentration ratio is high; and the catalyst is chosen to readily incorporate higher molecular mass olefin monomers. Other routes to improve the processability of the product include the use of mixed catalysts with different molar mass dependence on hydrogen concentration or temperature [96], [97].

For use in fluidized bed or slurry phase reactors, a supported catalyst is required. This has been achieved by many companies by treating a silica having the required particle size with MAO and then with the metallocene catalyst. The characteristics of narrow MMD etc. are maintained. In 1996 most major manufacturers are developing single-site catalyzed PEs, but the amount actually being sold is a very small fraction of PE sales. However, the catalysts can be used in all the major PE processes now using Phillips and Ziegler catalysts, and a changeover to single-site catalysts could occur comparatively quickly, depending on the advantages and disadvantages the new products present to plastics fabricators and end-users.

1.3.2.4. *Copolymerization*

Copolymerization of ethylene and α-olefins is the basis of the large and expanding LLDPE market. As with the free radical case the important factors in addition to those of homopolymerization are the relative reactivities and chain transfer. Both Phillips and Ziegler catalysts copolymerize olefins, but the latter are more widely used because they can be more readily tailored to produce narrow distributions of composition and molecular mass. As has been noted in Section Coordination Catalysis, polymerizations with typical coordination catalysts do not follow classical polymerization kinetics and hence the reactivity ratio concept is not strictly applicable. For completely soluble polymerization systems r_1 is very high and increases with the comonomer chain length (e.g., 29 for butene and 73 for octene, calculated from data in [41]). However, in the case of heterogeneous catalysts and heterogeneous polymerization systems, in particular, the apparent reactivity ratios are lower and may even decrease with increasing olefin chain length. This is presumably due to mass transfer and solubility effects in the growing particles. Chain transfer is enhanced in the presence of olefin comonomer, but this can be compensated by lowering the hydrogen concentration.

Literature articles have appeared on the copolymerization of ethylene with monomers such as styrene, butadiene, vinyl chloride, esters, and others but these have involved Ziegler catalyst variants with poor yields. For practical purposes the range of comonomers polymerizable by Ziegler catalysis is at present limited to olefins and nonconjugated diolefins. The advent of single-site catalysts with high intrinsic activity is removing some of these limitations.

1.4. Raw Materials

1.4.1. *Ethylene*

The first polyethylene plant built by ICI used ethylene produced by dehydration of ethanol. Modern ethylene production plants are based on the thermal cracking of

hydrocarbon feedstocks at ca. 850 °C. Due to the availability of raw materials, the feedstock used predominantly in North America was originally ethane, with naphtha being used more commonly in Europe. The tendency now is for crackers to be designed to accept a wider range of feedstocks, so as to match the range of coproducts such as propene to the market requirements.

Modern plants produce ethylene with a quality which in many cases is suitable for polymerization with little or no further purification. In North America and Europe, producers supply ethylene to an agreed specification via a common ethylene grid, and polymerization plants take their supplies from these pipelines. In the case of sensitive catalyst systems such as the Phillips catalyst some further purification may be necessary to ensure that maximum impurity levels are not exceeded. Table 8 lists specifications for a polymerization-grade ethylene suitable for most processes. There are overall limits on inert materials such as ethane or nitrogen which, because of the efficient recycle system, could build up and dilute the process stream. The main impurities of importance to the free-radical process are oxygen and water. The former could cause inhibition of low-temperature initiators, or uncontrolled initiation of reaction at higher temperatures. At moderate concentrations (after concentration by the recycle) water can form ethylene hydrate [98] in the cooler parts of the high-pressure process, completely blocking the pipework. Some of the other specified compounds can produce problems of molecular mass control or inhibition in the free-radical process, but only at very much higher concentrations.

Table 8. Specifications for polymerization-grade ethylene *

C_2H_4	> 99.9 vol %
CH_4, C_2H_6, N_2	< 1000 vol ppm
Olefins + diolefins	< 10 vol ppm
Acetylene	< 2 vol ppm
H_2	< 5 vol ppm
CO	< 1 vol ppm
CO_2	< 1 vol ppm
O_2	< 5 vol ppm
Alcohols (as MeOH)	< 1 vol ppm
H_2O	< 2.5 vol ppm
Sulfur	< 1 vol ppm
Carbonyl sulfide	< 1 vol ppm

* Data obtained from Repsol.

In the case of the Phillips and Ziegler processes, the materials included in the specification act as catalyst poisons. Hydrogen is a poison only for the Phillips catalyst.

1.4.2. *Comonomers*

The vinyl acetate and acrylate esters used as comonomers in the free-radical process are normal commercial quality materials containing sufficient stabilizer to prevent homopolymerization in storage or during pumping, but not so much as to affect the copolymerization reaction. They must be freed of dissolved oxygen by nitrogen sparging.

The butene used in LLDPE is normally a purified refinery product, although in the United States material is available from the oligomerization of ethylene. As an alternative to buying butene with its associated transport problems, IFP have developed the Alphabutol process [99], [100] for dimerizing ethylene as a compact adjunct to a polymerization plant. There is also the possibility of using a polymerization catalyst capable of simultaneously dimerizing ethylene [101], [102].

The higher α-olefins such as hexene and octene used in LLDPE processes are ethylene oligomerization products. They must be freed of oxygen and water before use in the polymerization process. The 4-methylpentene used by BP is a propylene dimer made by an alkali metal catalyzed process.

1.4.3. *Other Materials*

The initiators and catalysts are described in Section Free-Radical Catalysis. All the free radical initiators and many of the Ziegler and Phillips catalysts are manufactured by specialist suppliers. Some of the simpler Ziegler catalysts may be made on-site from basic chemicals such as $TiCl_4$, $MgCl_2$, etc. The aluminum alkyls are made by a very few international suppliers.

Initiator solvents and compressor lubricants for the free-radical process are carefully selected to be free of aromatic compounds, since much of the product may be used for food packaging.

1.5. Production Processes

Modern polyethylene production processes offer the possibility of a versatile range of products. High-pressure processes can produce LLDPE in addition to the normal range of LDPEs and ester copolymers. As well as HDPE some low-pressure plants can also produce LLDPE and VLDPE, and in many cases these compete for the same market as LDPE and the ester copolymers. A guide to the applicability of the various types of processes is given in Table 9.

Table 9. The technical applicability of processes for polyethylene manufacture *

	High-pressure autoclave	High-pressure tubular	Gas-phase fluidized bed	Slurry phase autoclave/loop	Solution autoclave
Installed capacity ** worldwide, 10^6 t/a	8.8	8.3	12.8	12.5	3.2
LDPE	+	+	–	–	–
EVA copolymers	+	+	–	–	–
Acrylate copolymers	+	+	–	–	–
HDPE	0	–	+	+	+
HMW HDPE	–	–	+	+	–
UHMPE	–	–	0	+	–
LLDPE	+	0	+	0	+
VLDPE	+	0	+	–	+

* + = suitable; 0 = technically feasible with some limitations; – = unsuitable or not possible.
** Capacity data (1995) obtained from Chem Systems, London.

Table 10 compares the capital and operating costs for various types of polyethylene plant. The comparison is on the basis of construction on a United States Gulf Coast site using the largest stream size currently (1996) available for licensing. The costings assume the production of pellets and this results in higher costs for the gas-phase and slurry plants than would be the case if 100 % sales of ex-reactor granules could be assumed. Some differences due to the economy of scale occur as a result of the available stream size. The production costs are dominated by the cost of ethylene, although some differences such as the extra cost of electrical energy can be noted. In the case of LLDPE these costings assume a unit cost for the butene equal to 1.05 times that of ethylene. In many other parts of the world the price of butene, including transport, is considerably higher. The unit price of higher olefins is generally appreciably higher than that of ethylene, from which they are derived by oligomerization.

Table 10. Production costs for polyethylene processes in $/t (1996 U.S. Gulf Coast prices) *

Product	LDPE	LLDPE	HDPE
Process	Autoclave Tubular	Fluid bed Solution	Fluid bed Ziegler Phillips

Capacity, 10^3 t/a	117	200	225	200	200	200	200
Capital cost, 10^6 $	85	116	98	138	90	135	105
Monomer costs	447	443	450 **	452 **	449	456	445
Catalysts, chemicals	20	18	29	31	26	22	20
Electricity	31	33	15	9	15	16	16
Other utilities	5	2	5	17	5	10	11
Manpower	10	6	6	6	6	6	6
Maintenance	15	13	9	15	9	15	13
Overheads	35	29	22	31	22	26	29
Production costs	565	544	534	561	553	553	539
Depreciation	71	59	44	69	45	68	53
Total costs	636	603	578	630	577	620	592

* Data from Chem Systems, London.
** Includes cost of butene monomer at a unit price equal to1.05 times that of ethylene; other locations or the use of other olefin comonomers could lead to a higher monomer cost.

1.5.1. *High-Pressure Process*

A flowsheet of the high-pressure polyethylene process is shown in Figure 14. The reactor may take one of two forms: a high-pressure autoclave or a jacketted tube, but otherwise the processes are similar. The reaction pressure is typically in the range 150 – 200 MPa for the autoclave process and 200 – 350 MPa for a tubular reactor. Such high pressures call for very specialized technology and many key features have remained proprietary information. The design of thick-walled cylinders requires a different type of analysis [103], [104] from that for lower pressure vessels, and fatigue is a major design consideration for pumps and compressors. A drawing of the second stage cylinder of a high-pressure ethylene compressor is shown in Figure 15, which illustrates the massive construction necessary and the avoidance of cross-bores to improve fatigue resistance at the very highest pressures. Specialized forms of sealing joints in vessels and pipework have been developed which make use of the pressure itself to increase the sealing forces [105].

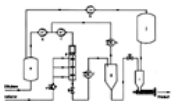

Figure 14. High-pressure autoclave process

a) Ethylene stock tank (5 MPa); b) Primary compressor; c) Secondary compressor (200 MPa); d) Autoclave reactor; e) Initiator pumps; f) Product cooler; g) Separator (25 MPa); h) Recycle cooler; i) Low-pressure separator and melt extruder; j) Low-pressure stock tank (0.2 MPa); k) Booster compressor

Figure 15. Second stage cylinder of a Nuovo Pignone compressor with 350 MPa maximum output pressure

a) Piston; b) Packings; c) Lubricant injection to packings; d) Valves

Referring to Figure 14, the fresh ethylene enters from the refinery at ca. 5 MPa, mixes with the low-pressure recycle and is compressed to 25 MPa. After mixing with the intermediate-pressure recycle, the pressure is raised in the secondary (or hyper) compressor to 150 – 350 MPa for feeding to the reactor. The pressure in the reactor is controlled automatically by a flow control valve at the reactor outlet. The reaction mixture then passes through a cooler to reduce the polymer temperature to a value suitable for feeding the pelletizing extruder. The polyethylene is separated from the majority of the unreacted monomer in the intermediate separator at ca. 27 MPa. This pressure is chosen to give a compromise between separation efficiency and compression energy savings. The remaining monomer is removed in the low-pressure separator that feeds the pelletizing extruder. The extrudate is pelletized underwater by a die-face cutter, and the pellets are then dried and conveyed to temporary storage hoppers to await quality control clearance. Finally the pellets are transferred to silos for blending and storage, before off-loading to tankers or sacks.

1.5.1.1. *Autoclave Reactor*

A typical autoclave design is shown in Figure 16. The autoclave volume is chosen to give an overall residence time of ca. 30 – 60 s, with corresponding volumes in larger plants of ≥ 1 m^3. A novel feature is the internal stirrer motor. The Du Pont process uses an external motor. The elongated cylindrical form arises partly from the fabrication constraints of making a thick-walled forging, and partly from the requirements of the process for multiple zones. Cross-bores are provided along the length of the reactor for thermocouples, and monomer and initiator entries. Bursting disks or other relief devices are mounted directly into the reactor walls to provide unrestricted passage for the reactor contents in the event of a pressure rise due to a decomposition.

Figure 16. High-pressure autoclave reactor

a) Stirrer motor; b) Stirrer shaft; c) Bursting disk ports

The autoclave functions as an adiabatic continuous stirred-tank reactor (CSTR), with the heat of reaction being removed by the fresh ethylene entering the reactor. The conversion of monomer to polymer is thus related to the difference in temperature between the feed gas and the final reaction temperature. For practical purposes percentage conversion = 0.075× ΔT. Most modern reactors have two or more zones with increasing temperatures. The reaction temperatures are maintained constant by controlling the speeds of the pumps feeding initiators into the respective zones. The first zone is typically 180 °C and the final zone 290 °C. For adequate control the initiators must have decomposition half-lifes of ca. 1 s under the reaction conditions in the zone. Table 6 lists a range of commercial initiators used in both the autoclave and tubular processes.

1.5.1.2. *Tubular Reactor*

A tubular reactor typically consists of several hundred meters of jacketted high-pressure tubing arranged as a series of straight sections connected by 180° bends. Inner diameters of 25 – 75 mm have been quoted, but 60 mm or somewhat larger is probably typical of modern tubular reactors. A ratio of outer to inner diameters of about 2.5 is used to provide the necessary strength for the high pressures involved. At many of the pipe junctions thermocouples are introduced to follow the course of the reaction, and initiator and gas inlets or pressure relief devices may also be incorporated. Unlike the autoclave process, no after-cooler is required for the secondary compressor, but the first section of the tubular reactor must function as a preheater to raise the ethylene to a sufficiently high temperature for the reaction to start. This temperature depends on the initiator employed, ranging from 190 °C for oxygen to 140 °C for a peroxydicarbonate. The latter part of the reactor functions as a product cooler similar to that of the autoclave process.

A tubular reactor works in the plug flow regime with heat transfer to the jacket. Plug flow is achieved by the correct choice of pipe diameter relative to the flow rate [106] so as to give sufficient turbulence and good axial mixing. Although heat is transferred through the reactor wall, it is not generally possible to maintain isothermal conditions, and temperature peaks occur. Because of the temperature peaks, which may not occur at exactly constant position in the tube, automatic temperature control must be more sophisticated than in the autoclave process; i.e., it must be possible to calculate average temperatures for appropriate regions of the reactor. When oxygen is used as initiator, the temperature control acts on the rate of addition of oxygen in the lower pressure part of the system. When peroxide initiators are used, the speeds of the high-pressure pumps are controlled. Oxygen is still widely used in the tubular reactor process, either alone or in conjunction with peroxides. Because of its complex initiation mechanism, oxygen tends to give more gentle temperature peaks with less tendency for decomposition. In the case of reactors with multiple initiator injection, liquid initiators offer more flexibility, since they can be injected at points where there is no fresh ethylene injection (which would be required to carry in oxygen as initiator). Injection of initiator at various positions along the tube produces new temperature peaks, increasing the overall conversion. By using these techniques higher conversions than in the autoclave reactor can be achieved, but at a higher cost in compression energy. Although conversions of up to 35 % (compared with 20 % for the autoclave) have been claimed, the maximum useful conversion depends on the product quality required, since quality deteriorates markedly with increasing conversion.

1.5.1.3. *High-Pressure Copolymers*

The high-pressure processes described are also suitable for the copolymerization of monomers such as vinyl acetate or acrylic esters. The autoclave process is generally

preferred for its well-defined operating conditions and its ability to produce a useful conversion at a low maximum temperature. In the case of vinyl acetate the reactivities of the two monomers are virtually identical and so they are consumed at the same rate. This means that the reacting monomer mixture and the copolymer produced maintain a constant composition even in multizoned reactors, but the recycle system must handle high concentrations of vinyl acetate. The converse is the case for the acrylate esters. If a multizoned or a tubular reactor is used the composition varies in the different zones, but the recycle is nearly pure ethylene. The principal modifications to a high-pressure polyethylene plant to enable it to manufacture copolymers are:

1. Installation of liquid pumps, which usually pump the comonomer into the suction of the secondary compressor
2. In the case of vinyl acetate, it is necessary to collect the monomer which condenses in the low-pressure recycle system and purify it before returning it to the liquid pump
3. The system for removing final traces of monomer must be improved, because the comonomers are more soluble than ethylene and they have more offensive odors

Technically a copolymer plant could also copolymerize acrylic or methacrylic acid, but there are long-term corrosion problems. Major producers of these copolymers have constructed plants especially for their manufacture, using corrosion-resistant steels.

1.5.1.4. *Linear Low-Density Polyethylene (LLDPE)*

CdF Chimie converted high-pressure process equipment to use Ziegler catalysts to make HDPE. These plants were later used to copolymerize ethylene with butene and other comonomers to make LLDPE. The catalysts used are generally of the Ziegler type but have been specially developed for the high temperatures of the high-pressure process [107]. Several other manufacturers have modified existing high-pressure plants to enable them to make a rapid, but limited entry into the LLDPE market, but this route is not seen as suitable for a large new investment. The required modifications (generally similar to those for free-radical co-polymers together with some additional ones associated with the different type of catalyst) are as follows:

1. Purification columns to remove polar impurities from the ethylene and olefin comonomer.
2. Hydrogen injection for MFI control.
3. Compressor modifications to take account of the lower compressibility and poorer lubricating properties of the monomer mixture [108]. The concentrations of butene or other olefin of ca. 50 % are much higher than in the free-radical copolymer case.
4. Catalyst handling equipment to produce pumpable dispersions, and to maintain the catalyst under a nitrogen atmosphere.
5. A system for injecting a catalyst deactivator such as a suspension of calcium stearate [109], after the reactor and before the separator.
6. Usually some modifications to the pelletizing extruder to take account of the higher torque generated by LLDPE.

1.5.2. *Suspension (Slurry) Process*

The formation of polyethylene suspended in a hydrocarbon diluent was envisaged as a convenient form of production process from the earliest patenting by ZIEGLER [110]. This was reinforced by the fact that, for a given pressure, most Ziegler catalysts give their highest yields at temperatures at which polyethylene is insoluble. The Phillips process on the other hand originated in the laboratory as a solution process that uses a fixed bed of catalyst [111] and was commercialized in 1956 as a solution process using a powdered catalyst which had to be removed by filtration. Laboratory developments led to the discovery of high-activity catalysts which fragment at temperatures below 105 °C to enable the low concentration of finely divided catalyst residues to be left in the product. Since 1961 all Phillips plants have been of this type using novel process technology to implement their Particle Form suspension process.

Many companies have built plants to make polyethylene by using Ziegler catalysts, but because the license was only for the use of the catalyst, there has been a diversity of process designs, even amongst the suspension processes. The Phillips process was licensed as a package and the plants themselves tended to be very similar. More recently the picture has become blurred as chromium-based catalysts have been used in fluidized-bed reactors and Ziegler catalysts are being employed in loop reactors. Early Ziegler plants included a catalyst residue removal stage which added considerably to the complexity and cost. Since the late 1960s it has been possible to eliminate this step. Other variations arise from the selection of the diluent. A high-boiling diluent generally requires more energy to remove the final traces from the polymer, and stripping with steam is frequently employed. Because of the low flash point, the use of a low-boiling diluent such as hexane requires more care in the design and operation of the plant, but this seems to be the preferred route for modern plants.

The suspension process has been used extensively for the production of HDPE, and in many cases these polymers incorporate a small amount of comonomer to increase the toughness or resistance to stress cracking. The use of higher concentrations of comonomer to produce LLDPE presents problems however, as a significant fraction of the product dissolves in the diluent. Because of the poorer solvent properties of isobutane, the Phillips process is best suited to making lower density materials. With the original Phillips catalysts the lower density limit was considered to be about 930 kg/m^3, but with more recent catalysts, particularly supported single-site catalysts, the limit is reduced to around 920 kg/m^3.

1.5.2.1. *Autoclave Process*

Figure 17 shows a flowsheet for a suspension (slurry) process based on various descriptions of the Hoechst process [8], [74], [112], [113]. The pressure employed is between 0.5 and 1.0 MPa, allowing the use of large reactors (ca. 100 m^3). The reaction temperature is 80 – 90 °C. The diluent is a low-boiling hydrocarbon such as hexane. The catalyst compounds and aluminum alkyl are slurried with diluent in the catalyst mixing vessel before being fed to the reactor at a rate sufficient to maintain the required polymerization rate. Only one main reactor is shown but, at least in the case of bimodal MMD high molecular mass polymers [114], two or more reactors in cascade may be used. The reaction mixture is then passed to a run-down reactor where the dissolved ethylene is consumed almost completely, avoiding the need for an ethylene recycle. The slurry concentration is an important parameter in the process. A high concentration allows higher outputs from a given reactor volume, but the heat transfer to the cooling jacket is worse and stirring becomes more difficult. The maximum usable slurry concentration depends on a number of factors, including solvent type, particle size and shape, but principally on the bulk density of the polymer particles. Slurry concentrations vary from 15 to 45 wt % [115], with many patents reporting values in the range 30 –35 % in a heavy diluent.

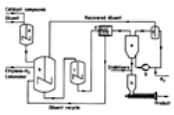

Figure 17. Hoechst suspension polymerization process

a) Catalyst preparation vessel; b) Polymerization reactor; c) Run-down reactor; d) Centrifuge; e) Fluidized-bed drier; f) Diluent condenser; g) Nitrogen circulator; h) Powder-fed extruder

The slurry from the run-down reactor then passes to a centrifuge to remove the bulk of the diluent, which is recycled directly to the reactor. This diluent contains aluminum alkyl and comonomer, if used, and the injection rates of these raw materials are adjusted to take into account the amounts reintroduced by the diluent recycle. The polymer is dried in a continuous fluidized-bed drier in a stream of hot nitrogen to remove residual diluent. Further quantities of diluent are condensed from the circulating nitrogen stream and recycled. Before extrusion into pellets, stabilizers are added to neutralize the catalyst residues, and other additives such as antioxidants may be added at this point.

The powder-fed extruder has a longer barrel than is the case for the high-pressure process because in addition to generating sufficient pressure for the pelletizing head, it must also melt the polymer powder. Although a single long barreled extruder is shown, new installations increasingly make use of a combination of a short extruder to melt the polymer followed by a gear pump to generate the pelletizing pressure. This combination has a lower energy requirement.

1.5.2.2. *Loop Reactor Process*

A simplified flowsheet for the Phillips Particle Form process is shown in Figure 18. The novel double loop reactor constructed from wide-bore jacketted pipe was developed by Phillips engineers to avoid deposits, which had been troublesome in a stirred autoclave [111], [116]. It also has a high surface-to-volume ratio, facilitating heat removal and allowing short residence times. The impeller forces the reaction mixture through the pipework in a turbulent regime with a velocity of 5 – 10 m/s. The reaction conditions of 100 °C and 3 – 4 MPa correspond to the needs of the chromium-based Phillips catalyst and the required productivity. The diluent used is isobutane which facilitates the subsequent flash separation and, being a poor solvent for polyethylene, permits higher operating temperatures than the higher alkanes. The catalyst is flushed into the reactor with diluent from the metering device at the base of the catalyst slurry tank. The polymer is taken off from a sedimentation leg which enables the slurry to be passed to the flash tank at a concentration of 55 – 65 % instead of the 30 – 35 % circulating in the loop reactor [111]. The isobutane diluent evaporates in the flash tank and is then condensed and recycled. Residual isobutane is removed in a nitrogen-flushed conveyor. Pelletization is carried out in a powder-fed extruder in a similar way to the Hoechst process.

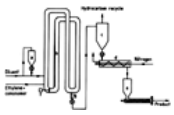

Figure 18. Flowsheet of the Phillips Particle Form process

a) Catalyst hopper and feed valve; b) Double loop reactor; c) Flash tank; d) Purge drier; e) Powder-fed extruder; f) Impeller; g) Sedimentation leg

1.5.3. *Gas-Phase Process*

Fluidized-bed processes for the production of HDPE were developed in the late 1960s by Union Carbide and somewhat later by Naftachemie (now BP). Although an innovative technology, the process did not offer clear economic advantages over the established slurry processes (see Table 10). A rapid increase in the building of fluidized-bed plants followed Union Carbide's announcement of their Unipol process for LLDPE in 1977. In this case the fluidized-bed process is competing with solution and high-pressure processes compared to which it has lower capital and operating costs, particularly if it can be assumed that customers are able to use ex-reactor granules directly (not assumed in Table 10). Many fluidized-bed units have been built as dual-purpose plants ("swing plants") with the ability to produce either LLDPE or HDPE according to demand. The fluidized-bed process can produce a very wide range of MFIs and densities since it is free of the viscosity constraints of the solution process and the solubility constraints of the slurry process. Ranges of < 0.01 to > 100 in MFI and densities from 890 to 970 kg/m^3 have been claimed [117], [118]. Originally the process used butene as the comonomer for LLDPEs, but later hexene was introduced for high-performance copolymers. BP uses 4-methylpentene for its high-performance LLDPEs.

Figure 19 shows a flowsheet for the Union Carbide fluidized-bed process [119]. The reactor has a characteristic shape with a cylindrical reaction section, and an expanded section in which the gas velocity is reduced to allow entrained particles to fall back into the bed. The bed diameter is ca. 4 m with a working height of 10 m and an overall height of the reaction unit of ca. 30 m. The gas enters the reactor through a distributor plate which provides an even distribution of gas and must also prevent powder falling through when the gas flow is stopped. The fluidized bed functions more or less as a continuous stirred-tank reactor in which mass transfer provides back mixing of material and heat throughout the reactor. There is a predominant upflow at the center and downflow at the walls. Reaction temperature is 80 – 100 °C, depending on the density of the product being made, and the pressure is in the range of 0.7 – 2.0 MPa. Originally, lower pressures were used for butene and hexene to avoid condensation in the recycle cooler, but it is now claimed that condensation can be used to increase the output since the heat of vaporization of the liquid olefin absorbs more heat from the polymerizing particles [120]. The conversion per pass is ca. 2 % for HDPE, but higher when using an olefin comonomer in the condensing mode. The positioning of the catalyst feed point and the polymer offtake is important to minimize the loss of catalyst particles, which results in low conversion. A cyclone and/or filter prevent fine particles reaching the recycle cooler and compressor. The polymer is removed via a sequenced valve to a powder cyclone, from which residual monomers are recovered and recompressed. The main recycle compressor circulates gas at a high flow rate, but with a small pressure rise. Since the process operates close to the melting point of the polymer, accurate temperature control is necessary by regulating the rate of catalyst addition. If a runaway reaction is detected, a gas such as carbon dioxide can be injected to poison the catalyst.

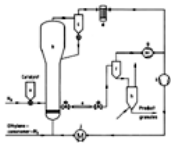

Figure 19. Fluidized-bed process

a) Catalyst hopper and feed valve; b) Fluidized-bed reactor; c) Cyclone; d) Filter; e) Polymer take-off system; f) Product recovery cyclone; g) Monomer recovery compressor; h) Purge hopper; i) Recycle compressor; j) Recycle gas cooler

The selection of the catalyst is critical for the success of the process. The catalyst particle grows by a process of replication to about 15 – 20 times its initial size. Not only does this affect the fluidization characteristics, but the polymerization rate and heat transfer must be controlled to prevent fusion of the particles or a thermal runaway reaction. Catalysts based on microspheroidal silica or $MgCl_2$ with a mean particle size of about 50 µm are claimed to be particularly suitable [77].

1.5.4. *Solution Process*

Solution processes have been developed by various companies including Du Pont, Dow, DSM, and Mitsui for the manufacture of LLDPE or HDPE/LLDPE on a swing basis. The advantages are that they readily handle a wide range of comonomer types and product densities and, depending on catalyst type, tend to produce narrow MMD products more readily. Like the high-pressure process (which is also a solution process) they are unable to handle high-viscosity products. The Du Pont and DSM processes are similar to the high-pressure process in that they function adiabatically with short residence times, whereas the Dow and Mitsui processes remove heat from the reaction mixture. Since the Du Pont process is available under license and more plants of this type have been constructed, this is the process included in the table of cost estimates (Table 10) and described below. The Du Pont process has a catalyst residue removal stage and for this reason is probably not the most economic solution process.

Figure 20 shows a flowsheet of the Sclairtech process of Du Pont of Canada [121]. The ethylene is dissolved in a diluent such as cyclohexane and pumped to the reactor at ca. 10 MPa. The reaction step is adiabatic and the reaction temperature is in the range 200 – 300 °C. The feed contains ca. 25 wt % ethylene, of which 95 % is converted to poylethylene in the reactor. The residence time is ca. 2 min. The catalyst most frequently quoted in patents is a mixture of $VOCl_3$ and $TiCl_4$ activated by an aluminum alkyl. Although the catalyst components may be initially soluble, the active catalyst species appears to be heterogeneous [80]. The polyethylene solution leaving the reactor is treated with a deactivating agent and the mixture then passes through a bed of alumina where the deactivated catalyst residues are adsorbed. Two depressurization stages follow, similar to the high-pressure process, in which solvent and unreacted monomers are volatilized. After extrusion into pellets, further removal of solvent residues is achieved by passing a heated stream of gas through the bed of pellets.

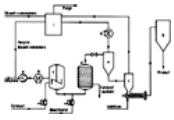

Figure 20. Solution polymerization process

a) Reactor feed pump; b) Temperature control; c) Reactor; d) Catalyst adsorber; e) First separator; f) Low-pressure separator and melt-fed extruder; g) Purged product hopper; h) Recycle cooler; i) Diluent and monomer purification unit

1.6. Uses

Table 11 shows an analysis by end-use of approximately half of the world market for polyethylenes. The pattern is very different for the two classes. LDPE (together with LLDPE which is sold into the same market) is used predominantly for films, not all of which is for packaging. Because of its greater rigidity and better creep properties, HDPE is used in more structural applications, and also has important applications in the packaging of aggressive liquids such as bleach, detergent, and hydrocarbons. The basic techniques of plastics fabrication are treated in ⟶ Plastics, Processing and so only the aspects particular to polyethylene are discussed here.

Table 11. Consumption by application*, %

Market	HDPE	LDPE and LLDPE
Film	24	71
Blow molding	36	1
Injection molding	21	6
Extrusion coating		9
Pipe	10	2
Wire and cable insulation	1	4
Others	8	6

* Percentages based on data for Western Europe, Japan, and North America 1996 [122].

1.6.1. *Film*

LDPE retains its position as a preferred packaging material because of its limp feel, transparency, toughness, and the ability to rapidly take up the shape of the contents of the bag. Other materials such as thinner HDPE film or paper may in some cases be more economical, but are less acceptable to the customer. Most LDPE film is produced by the film blowing process (⟶) but flat film extruded onto chilled rolls (⟶) is also made, particularly in the United States. An LDPE for high-clarity film typically has a MFI of 2 and a density of 920 kg/m^3 and is extruded at 160 – 180 °C. The cast film process is usually used with higher density polyethylenes (930 – 935 kg/m^3), where the quenching on the chilled rolls enables good optical properties to be achieved with a higher film stiffness. The bubble diameter in the tubular film process may be up to 2 m for general purpose packaging, and larger for heavy gauge industrial film. Originally, bags were made directly from tubular film by welding one end, but the tendency now is to make wide film on large machines and fabricate the bags by welding and cutting. Apart from packaging film and heavy duty sacks, increasing quantities of polyethylene are used for impermeable or stabilizing membranes in civil engineering construction.

Additives play a particularly important role in the production of LDPE film by the film blowing process. Without additives, the pressure of the windup rollers on the warm film forces the surfaces into such close contact that subsequently it may be virtually impossible to separate them. High-gloss films are the worst affected. This is overcome by adding an antiblocking agent such as very fine silica, which roughens the surface on a submicroscopic scale without significantly affecting the optical properties. To reduce the friction between the surfaces a slip agent such as oleamide or erucamide is added. Other additives may be added to achieve effects in the final product such as

oxidation resistance, UV resistance, or antistatic properties.

A film extruder designed for LDPE requires extensive modifications to allow it to extrude LLDPE at comparable rates [123-125]. To avoid this investment and the lack of flexibility, many film manufacturers (particularly in Europe) use blends of LDPE and LLDPE. The equipment modifications required are:

1. The screw must have greater clearance between the flights and the barrel to avoid temperature buildup due to the higher shear viscosity. Ideally the screw should be shorter.
2. The die gap must be widened to reduce the shear rate and so avoid a type of surface defect, known as shark skin, to which narrow MMD polymers are prone. To some extent this problem can be avoided by incorporating additives based on fluoroelastomers (Du Pont, 3M) or silicones (UCC) [126-128].
3. The cooling ring must give more rapid quenching of the melt to avoid the development of excessive crystalline haze, and also to give more support to the bubble because of the lower tensile viscosity under the inflation conditions. Apart from this lack of inherent stability of the bubble, the low tensile viscosity is responsible for the major product advantage of LLDPE, since high stresses do not occur when the melt is drawn down to thin film. LLDPE can be drawn down to much thinner film without the bubble tearing than can a LDPE with equivalent mechanical properties.

HDPE film is generally made on units specifically optimized for the high molecular mass grades normally used. To produce tough films a balanced orientation of the film is necessary, and a high ratio of die diameter to bubble diameter ("blow ratio") is used so as to balance the machine direction draw with a high transverse draw. A characteristic stalk-shaped bubble is used in which substantial machine-direction draw takes place before a rapid expansion occurs some distance above the die. Mechanical guides are needed to stabilize the bubble. HDPE film is opaque and is usually used in much thinner gauges than LDPE. It competes with LDPE in areas such as carrier bags and supermarket convenience bags.

1.6.2. *Extrusion Coating*

LDPE is used extensively for coating cardboard, paper, and aluminum for milk cartons etc., for which a very low level of impurities is required. Extrusion is carried out at ca. 300 °C through a wide-slit die. Polymers with a high die swell due to long branching perform best, since the tendency to expand on leaving the die opposes the "neck-in" tendency when the melt is drawn onto the substrate. A typical LDPE used for this application has a MFI of 4 and a density of 920 kg/m^3.

1.6.3. *Blow Molding*

HDPE is now the preferred material for blow-molded containers for liquids, combining adequate environmental stress crack resistance with higher rigidity than LDPE, and hence permitting lower bottle weights for a given duty. The major uses are in the domestic market for bleach, detergents, and milk. For this application a HDPE with a MFI of 0.2 and a density of 950 g/m^3 is suitable. Other uses include a variety of industrial containers and gasoline tanks. There is a trend to higher molecular mass and to broader MMD polymers, with bimodal MMDs achieved by using catalyst or reactor developments [129].

1.6.4. *Injection Molding*

Injection molding is used for a variety of products. Some, such as caps and lids are used in the packaging field. Other applications such as housewares, toys, and industrial containers are more durable. According to whether high flexibility is required or not, LDPE or HDPE may be used. Some LLDPEs with a narrow MMD are particularly well suited for this method of fabrication. MFIs range from 2 to > 40. A narrow MMD gives the best compromise between toughness, flow into the mold, and freedom from warpage in the finished product.

1.6.5. *Pipe*

The use of polyethylene for pipes is one of the few engineering applications, where the applied stresses are carefully assessed and a lifetime of at least 50 years is required. Polyethylene is used for water and natural gas local distribution systems and, unlike many other uses, there is a substantial stress applied continuously. The potential problem is thus long-term stress rupture, and this must be carefully assessed by extrapolation and accelerated testing. Large quantities of pipe made of a medium-density polyethylene designed to withstand a stress in the PE of 8 MPa have been laid, but recent advances in bimodal HDPE polymers now enable a 50-year lifetime at a design stress of 10 MPa to be met. The first polymers which met this specification (PE 100) were introduced by Solvay in their TUB 120 range, but now all major producers have similar polymers in their ranges [130], [36].

1.6.6. *Wire and Cable Insulation*

Because of its outstanding dielectric properties, the first application of polyethylene was the insulation of very high frequency and submarine telephone cables. These have continued to be made from LDPE, but uses have expanded into telephone and power cables. In the latter case cross-linking is increasingly used to enhance the properties and allow a higher current rating for a given size of cable.

1.6.7. *Ethylene Copolymers*

The processing techniques used for copolymers are generally the same as used for polyethylene homopolymers, but the range of applications is generally more diverse [12]. The property most generally exploited is the decreased rigidity to give more flexible film, moldings, and tubing than is possible with LDPE. Copolymers with high comonomer content are generally competitive with plasticized PVC and rubbers, with a low extractable additive content as an advantage. For flexibility down to the lowest temperatures acrylate copolymers are preferred over EVA copolymers. VLDPE copolymers, particularly those made by single-site catalysts, are increasingly being used in applications requiring flexibility combined with thermal stability and good low-temperature flexibility.

Ethylene copolymers are also used in blends, in the same way as rubbers, to confer increased toughness. Particular applications include blends with wax and with bitumen for paper coating and road surfacing, respectively. Copolymers are also used as constituents of hot melt adhesives.

1.6.8. *Ultra High Modulus Polyethylene Fibers*

Because of their low softening point polyethylene fibers have not been of interest for general purpose textile use, although substantial amounts are used for producing monofilament. In 1973 Capaccio and Ward showed that solid HDPE could be drawn at elevated temperatures to very high draw ratios to produce fibers with a tensile modulus of 70 GPa and a tensile strength of 1 GPa (Kevlar 149: 172 GPa and 3.4 GPa, respectively). Theoretical values for a perfect extended-chain crystal are 300 GPa and ca. 25 GPa, and thus these early results achieved a significant fraction of the maximum modulus possible. Further developments have been on parallel fronts: melt spinning followed by solid-state drawing, and gel spinning.

Melt spinning has been adopted commercially by Celanese, SNIA, and Mitsui. Gel spinning consists of forming a dilute solution in a heavy solvent and then extruding it so that gelled PE fibers are formed on cooling and can then be drawn on conventional fiber-making equipment. The solvent has to be removed by evaporation or solvent extraction. The gel-spinning process is operated by Allied Signal and DSM. The melt-spinning process is limited to moderate molecular masses because of the polymer viscosity, but the gel-spinning process can use higher molecular mass PEs and achieves higher modulus and tensile strength (172 GPa and 3.0 GPa). Further advances may be expected on using PEs made with single-site catalysts. Since the production costs of the gel-spinning process are higher there seems to be a market for both types of product.

Limitations on the uses include the low softening temperature and creep. Creep can be reduced considerably by cross-linking [131]. For use in composites the fibers must exhibit good adhesion to epoxy resins, and this can be achieved by various methods, principally by plasma treatment [132]. PE fibers have poor compressive strength and so for composites they are usually used in combination with other reinforcing fibers such as glass or carbon. The principal market is probably in energy-absorption applications such as personal armour where the characteristics which are responsible for creep give rise to energy absorption superior to existing materials. Other applications which make use if its high modulus are in leading-edge sporting goods where it competes with carbon fiber, aramid etc. [133-132].

1.6.9. *Joining Polyethylene*

The most satisfactory method of joining PE is by welding but considerable experience is required to achieve satisfactory results. A widely used method uses a stream of hot gas (usually nitrogen to avoid oxidation) in a similar manner to gas welding of metals. The edges to be joined are chamfered to allow a bead of PE to be melted into the join. The heating causes changes in the annealed state of the PE adjacent to the join and it may become weaker than the original material, but this can be overcome by welding a reinforcing strip over the joint line [135].

Welding is used by the gas and water industries to join pipe in their distribution networks. The most satisfactory method is butt welding, in which the ends are machined true and then heated by a PTFE-insulated electrical heater. The automated machine then removes the heater and presses the ends together at a predetermined force. The procedure produces a bead of PE on the inside and outside surfaces, and the external bead is machined flush with the pipe surface.

Because of its low surface energy PE cannot be joined in its natural state by any of the strong adhesives such as epoxy resins. Rubbery adhesives such as pressure-sensitive and hot-melt adhesives function satisfactorily, but the mechanical strength of the bonding material is not very high. By treatment of the surface it is possible to form a bond with epoxies. A recommended method is oxidation with chromic acid, after which the surfaces can be glued with epoxy resin [135]. A recent development for bonding polyolefins is the use of an aliphatic amine primer that penetrates into the surface of the PE. When a cyanoacrylate adhesive is then applied, polymerization to form a bond occurs in the usual way [136].

1.7. Chemically Modified Polyethylenes

1.7.1. *Cross-Linked Polyethylene*

As noted in Section Free-Radical Catalysis, polyethylene free radicals combine to form larger molecules, unlike materials such as polypropylene, which tends to degrade. This cross-linking reaction has been made use of in a number of ways to effectively increase the molecular mass with corresponding improvements in some properties. The principal commercial methods of cross-linking [137] are:

1. Free radicals can be generated by mixing in a peroxide such as dicumyl peroxide or di-*tert*-butyl peroxide in an extruder at as low a temperature as possible and then curing the extruded cable in an external heated zone [138]. This process uses ca. 2 % peroxide and the curing process must be carried out under pressure to avoid excessive formation of voids. Peroxide together with a blowing agent is used for making foamed sheeting or molded articles, particularly of EVA.
2. Polyethylene can be cross-linked by radiation either above or below the melting point, but commercially, products are cross-linked in the solid state after the basic fabrication step. There are some differences in properties between peroxide and radiation cross-linked polyethylene due to the fact that in the solid state the cross-links occur in the existing amorphous regions, and the existing crystallinity is largely retained. Electron accelerators are used to provide the radiation in a readily controlled process [139]. Products include shrink tubing and sheeting, wire insulation, and water pipes.
3. Dow has developed a process (Sioplas-E) in which a peroxide is used to graft a vinylsilane to the polyethylene. The silane groups can then react together in the presence of moisture to form cross-links. A catalyst such as dibutyltin dilaurate is added to accelerate the cross-linking reaction. The process has the advantage that only a tenth of the amount of peroxide is required and the curing stage is simpler. However, since water must diffuse in from the surface, curing times become excessive for thick sections. The original process produced the graft polymer as a separate stage from the cable extrusion, but the Monosil process achieves the grafting in the cable extruder [140].
4. A further variant of silane cross-linking is the production of a silane copolymer by direct copolymerization of a vinylsilane in a high-pressure polyethylene reactor. This process developed by Mitsubishi for their Linklon-X polymers has been licensed to A. T. Plastics for their Aqualink process [141].

1.7.2. *Chlorinated Polyethylene*

Both LDPE and HDPE can be chlorinated to form what are in effect ethylene/vinyl chloride copolymers which are not accessible by direct polymerization. Chlorine contents of 34 – 44 wt % are reported [142]. By selection of the polyethylene and the chlorination conditions, products can be made with a range of processing viscosities and with the chlorination either randomly distributed or concentrated in segments, leaving some polyethylene crystallinity. The principal application is as a toughening agent for PVC, but there are other uses in the same general area as flexible materials such as EVA, but with the emphasis on outdoor weather resistance.

Du Pont manufacture a chlorosulfonated polyethylene (Hypalon) in which the sulfonyl groups are used in a propietary vulcanizing process. The materials are used as rubbers with good weathering properties.

1.7.3. *Fluorinated Polyethylene*

Fluorinated polyethylene is not produced as a bulk material, but surface treatment of articles with fluorine gas is used to produce a highly fluorinated surface layer. In this way the permeability of containers to a wide range of solvents can be reduced drastically. The treatment can be applied to finished articles, or fluorination can be carried out in-line by employing a dilute mixture of fluorine gas in nitrogen as the blow gas in the blow molding process [143]. Gasoline tanks are a major application, but the process is also used for containers for materials such as solvents and pesticides.

1.8. Environmental Aspects

1.8.1. *Manufacture*

Under normal operating conditions the impact on the environment of polyethylene production plants is minimal. There are some hydrocarbon emissions, particularly from those plants using solvents, but these are regulated by local legislation. In some cases hazardous materials such as the Phillips catalyst or aluminum alkyls are handled safely by attention to detail in the plant design and operation. The preparation of both peroxide initiators and Ziegler catalysts involves some extensive chemical operations with the attendant problems of waste disposal. All peroxide initiators are now made by specialist producers and there is also a tendency for this to be the case for Ziegler and Phillips catalysts.

The high-pressure autoclave process is protected against over-pressurization by bursting disks. When a disk bursts, the vessel contents, including carbonized polyethylene, are discharged quite safely to the atmosphere, but with the impression of an explosion. Thus, it is not feasible to site this type of plant as close to residential areas as would otherwise be possible.

1.8.2. *Polymer Disposal and Recycling*

At the time of writing (1997), much of the polyethylene produced is dumped into landfill within a few months of manufacture. This represents the fraction used in packaging. However, the use of disposable packaging is being increasingly called into question. In the 1970s and 1980s the concern was with the unsightliness of a material which was littering the environment, and did not degrade. Various schemes were developed to promote more rapid degradation based either on photosensitized oxidation [144] or biodegradation of blends with starch [145]. Some of these developments have reached commercialization, but now public opinion is moving towards the need to conserve plastics. Table 12 shows some of the options available to deal with the problem of disposable packaging.

Table 12. Polyethylene packaging disposal strategy

Disposal method	Advantages	Disadvantages
Incineration	avoids landfill	generates CO_2 without benefits
Incineration with energy recovery	displaces existing fossil fuel use	loss of fabrication costs
Biodegradation	removes unsightly refuse	doubts about completeness of degradation; generates CO_2 without benefits
Promoted degradation	removes unsightly refuse; lower labor costs for agricultural uses	degradation slow away from sunlight; generates CO_2 without benefits
Recycling via melt	makes use of scarce resource	some deterioration in quality
Recycling via monomer	top-quality product; processes mixed plastics	fairly large scale processing plant
Reuse	saves on fabrication costs	PE may not be the best material

Much emphasis centers on recycling, that is converting the polyethylene back into granules and then using it to produce more film etc. There is some oxidation during use and recovery, leading to some cross-linking and a slight deterioration in extrusion properties. It will be difficult therefore to match the original manufacturer's tightly controlled specification, particularly if the recovered material is of a variety of grades from different producers. There will also be a problem of suitability for use in contact with food, since the recovered material may contain a range of unknown organic and inorganic impurities. At present the intentions are to use recycled material in uncritical nonfood uses. However, a successful recycling program could outstrip such uses. Ideally recycling should be at the monomer level, but polyethylene does not readily decompose to form ethylene.

Processes have been developed which crack the plastics waste to a liquid hydrocarbon suitable for use as the feedstock to a conventional cracker, which effectively allows recycling via the monomer [146]. In the BP process plastics waste is fed to a fluidized bed reactor maintained at about 500 °C, and the liquid feedstock is recovered from the recirculating gas stream. Gaseous byproducts are used as fuel to maintain the temperature of the fluidized bed. This process has advantages over the hypothetical process of cracking PE to ethylene in that it tolerates the presence of other plastics materials, even PVC. The ever present problem of collection and transport is also eased in that the cracked feedstock could be produced in relatively small local units and transported in tankers to the refineries.

Another approach is to burn the polyethylene to recover electrical or thermal energy, thereby saving on fossil fuel which in most economies is still being used for this purpose.

[Top of Page]

2. Polypropylene

T. Geoffrey Heggs

2.1. Historical Survey

In the 1950s KARL ZIEGLER discovered new catalysts which were to revolutionize the plastics industry. Before then, propene could not be polymerized to high molecular mass products with the catalysts then available, namely, free-radical, anionic and cationic systems. Even the most favorable of these yielded only liquids consisting of many isomers and greaselike materials unsuitable for making hard plastic products.

In 1953 KARL ZIEGLER, Professor at the Max-Planck-Institut für Kohlenforschung, demonstrated that ethylene could be polymerized to a high molecular mass crystalline polymer at moderate pressure and temperature. His group had for some years been using aluminum alkyls to convert ethylene into a range of oligomers having an even number of carbon atoms. Unexpectedly, almost total conversion to butenes occurred in one experiment. The cause was eventually traced to nickel-ion contamination in the autoclave. Tests with many other metallic compounds revealed that a combination of zirconium ion with aluminum alkyl converted ethylene to high molecular mass polymer. Further searching revealed that titanium ion was even more active in producing polyethylene (November 1953). Such combinations of transition metal compounds with an aluminum alkyl later became known as Ziegler – Natta Catalysts, thereby acknowledging the immense contribution of GIULIO NATTA to discovering and characterizing polypropylene (PP) and other α-olefin polymers in 1954.

NATTA, Professor and Director of the Milan Institute of Industrial Chemistry, was a consultant to Montecatini, to whom NATTA assigned his patents. The close collaboration between ZIEGLER and NATTA was acknowledged in 1963 by the joint award of a Nobel prize in Chemistry for their outstanding contributions to polymer science. A fascinating account of ZIEGLER and NATTA's personal approach to these discoveries is provided by MCMILLAN [147].

In the the United States, patents on crystalline polypropylene were awarded originally to NATTA et al. (on behalf of Montecatini), but were overturned later in favor of Phillips Petroleum Company. The immense legal procedings to establish inventor rights to crystalline polypropylene in the United States are outlined by HOGAN and BANKS [148], with more recent comments by PINO and MORETTI on this costly litigation [149]. Phillip's polymerization catalyst, consisting of chromium-ion promoted silica – alumina, has not been used in any commercial plant to manufacture crystalline polypropylene.

Notwithstanding the complexities and hazardous nature of these new coordinated catalysts, industrialists and academe alike were stimulated by ZIEGLER and NATTA's prompt and prolific disclosures concerning this new, high-melting polymer. Industrial-scale production of PP started at Ferrara, Italy, in 1957 with a 6×10^3 t/a plant. By the end of 1994 global capacity had risen to 20.5×10^3 t/a (Table 13), spread over 35 countries [150]; today, new plants are expected to be capable of producing 200×10^3 t/a to achieve reasonable economies of scale.

Table 13. Global PP capacity (10^3 t/a) and its distribution

Country/region	End 1990	End 1994	Growth * 1990 – 1994, %
Western Europe	4530	5705	26
North America	4400	5440	24
Japan	1870	2656	42
Other Far East/Australia	2734	4090	50
Middle East/Africa	280	655	34
South America	450	1035	30
Eastern Europe	790	930	18
Total/average	15 054	20 511	36

* 7 % per annum growth for four years is 31 % total growth.

There are several reasons for this phenomenal growth, which exceeds that of other bulk plastics [151] (Table 14). At first propene was readily available, almost at byproduct prices, from petrochemical cracker plants making ethylene. The polymer itself was suitable for a wide range of existing and new applications such as films, various fibers, large and small moldings from boat hulls to instrument parts. In addition, new manufacturing technologies based on $MgCl_2$ supported catalyst systems yielded both cost savings and improved products by eliminating atactic PP removal steps and then the deashing stage. Some 40 years on there are expectations that the new metallocene catalyts will extend these uses, even to the extent of challenging products such as polyamide, ABS, and flexible PVC for part of their market [152].

Table 14. Major polymer usage in Western Europe

Polymer	Consumption, 10^3 t/a			Growth rate, %	
	1974	1984	1994	'74/'84	'84/'94
LLDPE +LDPE	3080	4089	6015	33	47
HDPE	1035	1704	3570	65	110
PS	1118	1290	1908	15	48
PVC	3360	3825	5480	14	43
PP	680	1895	5178	179	173

2.2. Polymer Structure

2.2.1. *Molecular and Chain Structure*

Polypropylene is a synthetic, high molecular mass linear addition polymer of propene. In 1954/5, NATTA et al. prepared sufficiently large amounts of PP to isolate and characterize various stereoisomers. This led to a new descriptive nomenclature still used today. Commercial interest lies primarily in highly crystalline PP, together with its further modifications through copolymerization.

In perfectly isotactic PP, each monomer unit in the chain is arranged in a regular head-to-tail assembly without any branching or 1,3 additions (Fig. 21 A). This is the result of template-type constraints by the heterogeneous stereospecific catalyst. In practice such perfection is hard to achieve. An occasional error (Fig. 21 B), averaging about 0.3 – 1.5 per hundred chain links, occurs in some, but not equally in all chains. The multisite nature of solid catalysts is thought to be responsible for this behavior.

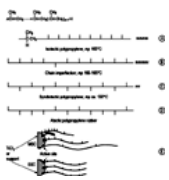

Figure 21. PP chain structures

Solvay and $MgCl_2$ supported systems reduce the extent of these errors, and with them the amount of noncrystalline (atactic) polymer. Still at the development stage, metallocene-based single-site catalysts (SSCs) also produce stereoregularity errors, but these seem to be distributed more uniformly along the chains, regardless of their length (Fig. 21 E). This contrasts with multisite catalysts (MSCs), for which misplacements vary in frequency and in number. The tendency here is for the high molecular mass fraction to contain fewer faults, while the short chains suffer from excessive dispruptions, which lead to reduced crystallinity.

Until recently, syndiotactic PP (Fig. 21 C) was not commercially important because of unacceptable polymerization difficulties at low temperatures. Some metallocene catalysts overcome this problem to produce a regular sequence of racemic (r) propene placements yielding a crystalline polymer melting at 130 °C. Mitsui Toatsu and Fina have entered into a joint venture to assess costs and customer reaction to syndiotactic PP.

Complete loss of steric regulation generates atactic polymer chains (Fig. 21 D), incapable of crystallizing. Hitherto this low molecular mass byproduct had only a few commercial outlets for adhesives. The position could change with the emergence of high molecular mass versions deliberately made with a metallocene-based catalyst, $[(Me_2SiFlu_2)ZrCl_2]$ (Flu = fluorenyl)[153].

All Ziegler – Natta commercial polymers have broad molecular mass distributions within the range M_w/M_n = 5 – 10 for ex-reactor material. Any reduction is achieved by

peroxide-induced chain clipping in an extruder. Metallocene polymers have narrower distributions (M_w/M_n = 2 – 4) as a consequence of their single-site mechanism. This avoids the need for peroxide additives, which produce objectionable odor and volatile byproducts.

2.2.2. *Crystallization and Morphology*

Polypropylene's strength and versatility stem from a matrix of interlocking crystallites that allow formation of rigid and tough articles. Highly isotactic polymer, with its regular structure, forms a helical coil having three monomer units per turn. These coils stack together into thin lamellar crystallites which force the chains to fold several times as they emerge and reenter lamellae. Polyethylene behaves similarly (see Fig. 2)

While the equilibrium (i.e., thermodynamic) melting point is 188 °C, at normal analysis heating rates the final melting point is ca. 160 – 170 °C. Three crystalline forms are known (Table 15), of which the α-form is the most stable. Rapid quenching yields the β-form with a lower density and melting point of 150 °C. Polymers of lower stereoregularity and random copolymers usually contain γ- crystallites in addition to the α-form.

Table 15. Crystal habits of PP [154-156]

Crystal form	System	Crystal density (20 °C), g/cm^3	Chains per unit cell	*mp*, °C *
Iso α	monoclinic	0.932 – 0.943	4	171
Iso β	pseudo hexagonal	0.922	9	150
Iso γ	triclinic	0.939	1	131
Smectic		0.916		
Amorphous		0.85		
Syndio	orthorhombic	0.93	2 (4/1 helix)	138

* The enthalpy of melting for 100 % crystalline isotactic PP is variously reported, due to different techniques. WUNDERLICH [154] gives 165 ± 18 J/g, BRANDRUP and IMMERGUT report 209 J/g [155], the ICI DSC standard is 185 J/g, and the Shell DSC standard is 188emsp14;J/g [157, p. 590].

Crystallization is complicated further by assembly of these lamellar crystals into larger structures, called spherulites. These have radial symmetry and can adopt five different forms (Table 16). While the size of the crystallites is in the region of 10 – 50 nm, too small to scatter visible light (400 – 600 nm), spherulites are much larger (10^3 – 10^5 nm). The latter are responsible for the translucency of slowly cooled PP moldings. Rapid crystallization from the melt by fast cooling, adding nucleating agents, and reducing spherulite sizes by stretching a partially solidified melt all improve transparency.

Table 16. Spherulitic structures

Spherulite type	Crystal type	Features
I	α	positive birefringence, lost at 160 – 170°C; epitaxial branching; always associated with type II
II	α	negative birefringence, lost at 160 – 170 °C; fibrillar crystals radiating from nucleus; most stable of all spherulites
III	β	strongly negative birefringence; associated with β-crystals and high shear in cooling; melts and reverts to type II at 145 –150 °C
IV	β	special case of Type III, characterized by banded concentric rings
V	γ	rare variant formed under high pressure with γ-crystals

Crystallization from a viscous melt is hindered by chain entanglements and by the need for helices to fold as they close-pack into lamellae. This, and the formation of nuclei might contribute to the low degree of crystallization even in highly stereoregular polymers. Typically, the onset of crystallization occurs at ca. 110 – 120 °C in DSC instruments. This point coincides with the cooling peak for polyethylene, whose presence would be masked in blends and in ethylene copolymers. Using the melting curve overcomes this analytical difficulty. Another aspect of tardy crystallization is the low level of crystallinity in fabricated items, even when highly stereoregular polymers are used. Commercial articles vary from 30 % crystallinity in rapidly quenched films to 50 – 60 % in moldings. Even purified and annealed samples of highly stereoregular PP rarely achieve more than 70 % crystallinity. Polypropylene is properly regarded as a semicrystalline polymer for these reasons.

2.3. Raw Materials

2.3.1. *Propene*

Sensitive Ziegler – Natta catalysts require pure propene to ensure high polymerization rates and good yields of isotactic polymer, without wasteful byproducts. Harmful impurities can be grouped into two main categories which have different effects on the process. The most destructive contaminants are polar and highly unsaturated compounds which react directly with the catalyst systems, retarding polymerization and impairing stereospecificity. Examples include acetylenes, dienes, carbon monoxide and dioxide, carbonyl sulfide, water, alcohols, and ammonia. The second group contains materials which adversely influence the process, but not the chemistry, by raising the pressure during reaction due to an accumulation of inert substances. These further complicate the monomer recovery section and recycle streams. This group contains materials such as methane, ethane, propane, butane, and nitrogen.

Most manufacturers use polymer-grade propene as feedstock, which can be further refined by passing it through local guard columns containing alkali, molecular sieve, alumina, supported copper, etc. to guarantee consistently high purity [158], [159]. Typically, monomer conforming to the following specification, is adequate for most plants, but certain sensitive supported catalysts require the lower levels listed for carbonyl sulfide, carbon monoxide, and oxygen (this quality is said to be suitable for metallocene catalysts) [160]:

Propene	99.5	wt % min.
Propadiene	5	vol ppm max.
Propyne	5	vol ppm max.
Butadienes	10	vol ppm max.
Oxygen	5 – 2	vol ppm max.
Carbon monoxide	3 – 0.3	vol ppm max.
Carbon dioxide	5	vol ppm max.
Carbonyl sulfide	0.5 – 0.03	vol ppm max.
Total sulfur	1	wt ppm max.
Water	5	wt ppm max.
Methanol	5	vol ppm max.
Ammonia	1	vol ppm max.
Ethane	500	vol ppm
Propane	0.5	wt %
Butane	500	vol ppm
$N_2 + CH_4$	300	vol ppm
Isobutene	50	vol ppm (estimated)
Hydrogen	10	vol ppm max.
Ethylene	50	vol ppm max.

1-Butene	50 vol ppm max.

Ethylene and 1-butene copolymerize with propene to marginally reduce the polymer softening point. Hydrogen is a powerful chain-transfer agent whose concentration must be accurately controlled in the reactor.

Propene is normally manufactured outside the PP plant. An important source of both ethylene and propene, in the approximate weight ratio of 2: 1, is from steam cracking naphtha or gas oil at 700 – 950 °C, followed by fractionation. Refinery catalytic cracking of petroleum products yields 2 – 5 % of propene, which is isolated as a chemical grade containing 90 – 95 % of the monomer with much propane. This can be purified further to polymer-grade specification, but some PP plants which are able to handle the extra propane may bypass further refining or add a simplified hydrogenation unit to reduce dienes [161]. These possibilities arise at the milder end of the cracking range of 400 – 500 °C, which then generates lower amounts of harmful highly unsaturated products. Currently there is some interest in building propane dehydrogenation plants as a third source of pure propene [161], [162].

Some properties of propene relevant to polypropylene production are as follows:

Heat of polymerization	2514 kJ/kg
bp	– 47.7 °C
Critical temperature	92 °C
Critical pressure	4.6 MPa
Vapor pressure at 20 °C	0.98 MPa

2.3.2. *Polymerization Diluents*

Most early processes and some current plants use an inert hydrocarbon diluent in the reactor to help transfer propene to the solid catalyst and to convey reaction heat to the water- cooled jacket. Typically there is twice as much diluent in the reactor as polymer; thus, the purity demands on the diluent are as important as for the monomer. Refined petroleum fractions and synthetic hydrocarbons in the C_6 – C_8 range perform well, while solvents ranging from butane to dodecane are also used commercially.

Polar impurities, such as alcohols, carbonyl compounds, water, and sulfur- containing compounds must be kept below 1 – 5 ppm. The content of aromatics should be below 0.1 – 0.5 vol %, depending on the catalyst employed. Here UV spectroscopy, supplemented with IR spectroscopy and refractive index measurement, is used to monitor quality. Regular checks must be made on recycled-diluent quality to prevent oxidized species and catalyst fragments accumulating unnoticed.

Polypropylene articles, such as films and containers, are widely used for food packaging and handling. In these cases the diluents and other additives must conform with the appropriate health regulations.

Since polymerizations carried out in the gas phase or in liquid propene only use small amounts of diluent as catalyst carriers, their purity requirements are less stringent.

2.3.3. *Catalyst Preparation [163]*

All current commercial manufacturing processes for highly isotactic PP use heterogeneous Ziegler – Natta catalyst systems. These two-component initiators consist of a solid transition-metal halide, usually $TiCl_3$, and an organoaluminum alkylating agent such as diethylaluminum chloride (DEAC). This system was the basis of PP plants until about 1980 when supported catalysts were commercialized. An overview of the four development phases is given in Table 17.

Table 17. Catalyst development

System	Catalyst performance		Plant
	Activity, kg PP/g catalyst	I.I. *, wt %	process
1st Generation			
1957 – 1970	0.8 – 1.2	88 – 93	deash and remove
$TiCl_3 \cdot AlEt_2Cl$			atactic
2nd Generation			
1970 – 1980	3 – 5	92 – 97	deash/deactivate,
$TiCl_3 \cdot AlEt_2Cl$ +			atactic usually remains
Lewis base (Solvay)			
3rd Generation			
1980 – 1990	5 – 20	≥ 98	no deash and
$MgCl_2$-supported $TiCl_4 \cdot AlEt_3$			no atactic removal
4th Generation			
1995	ca. 20	?	no deash and
Metallocene/MAO systems			no atactic removal
			MMD ca. 2 – 4

* Isotactic index (% insoluble in boiling heptane).

2.3.3.1. *$TiCl_3$-Based Catalysts*

Of the four known crystalline forms of $TiCl_3$, three are purple (α, γ, δ) while the fourth is brown (β). The latter is not used in PP manufacture because of its poor stereospecificity. All the purple forms have a layer lattice consisting of closely packed chlorine atoms (Fig. 22) either hexagonal (α), cubic (γ), or intermediate between the two (δ). Preparative methods are critical to achieve good stereoregulation and high activity with these catalyst components.

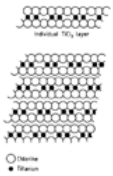

Figure 22. Layer lattice of α-$TiCl_3$ [165] (printed with permission of Harwood Academic Publishers)

Early work by the NATTA school used pure $TiCl_3$ made by hydrogen reduction of $TiCl_4$, followed by ball milling to give a high surface area.

$$TiCl_4 + 1/2\,H_2 \rightarrow TiCl_3 + HCl$$

A more active form is made by reduction with aluminum powder, followed by ball milling.

$$3\,TiCl_4 + Al \rightarrow 3\,TiCl_3 \cdot AlCl_3$$

Much of the aluminum chloride is present as a true solid solution in crystalline δ-$TiCl_3$. This catalyst has a larger surface area of 10 – 40 m^2/g and is commercially available

from several companies. Nowadays, a Lewis base, such as an ester or ketone, may be introduced at the milling stage to increase stereospecificity.

Another important route dispenses with ball milling by reducing a hydrocarbon solution of $TiCl_4$ with an organoaluminum compound below room temperature. Di- and trialkylaluminum compounds are suitable. In this process the titanium trichloride precipitates in the brown, β form. Heating to 80 – 120 °C completes the reduction and transforms this solid to the required purple variety. Several hydrocarbon washes remove some of the aluminum compounds to leave an active, stereospecific catalyst. A useful aspect of this technique is that conditions can be selected to precipitate spherical catalyst particles. Subsequent propene polymerization then yields spherical polymer particles that exhibit good flow characteristics. This ability to replicate the shape of catalyst particles is quite general in Ziegler – Natta polymerization. Polymer forms throughout the agglomerate of small catalyst clusters, expanding the particle uniformly in all dimensions as polymerization continues.

Major refinements in alkyl-reduced $TiCl_3$ systems were made by Solvay [164-166] in their three-stage process, which gives a 4 – 5-fold increase in activity:

1. Titanium tetrachloride solution is reduced with diethylaluminum chloride at 0 °C
   $TiCl_4 + AlEt_2Cl \longrightarrow TiCl_3 \cdot x\,AlCl_3 \cdot z\,AlEtCl_2$
   where $x = 0.15$, $z = 0.20$
2. The brown β-$TiCl_3$ solid is treated with diisoamyl ether to dissolve out most of the aluminum compounds
3. The extracted solid is reacted with neat titanium tetrachloride at 65 °C to transform the β crystallites to the violet δ-form

The novelty lies in removing much of the aluminum residues at the β stage, without collapsing the particle, and then transforming to the stereoregulating δ-form while retaining a spongelike structure with a surface area of 150 m^2/g [164]. Analogous commercial products are available in Japan and the United States.

2.3.3.2. *Supported Catalysts*

The third-generation, supported titanium catalysts, are based on research by Montedison (now Montell) in Italy and Mitsui Petrochemical Industries (now Mitsui Sekka) in Japan. Collaboration and cross-licensing between these two companies avoided incipient interference claims. Starting around 1968 with finely milled magnesium chloride, which has a similar layer structure to violet $TiCl_3$, ways were discovered of depositing stereoregulating titanium compounds on the surface.

A particle-form supported catalyst may be prepared as follows [167]:

1. A spheroidal support is prepared by making a fine dispersion of molten (125 °C) magnesium chloride ethanolate complex in hot kerosene containing sorbitan distearate as surfactant. The dispersion is quenched by pumping into kerosene at – 15 °C to give particles with a diameter of 5 – 30 µm.
2. A slurry of the particles is added to neat $TiCl_4$ at 20 °C.
3. Diisobutyl phthalate (17 % v/w) is added and the mixture heated to 120 °C.
4. Decant, and heat with further $TiCl_4$ to 130 °C. Isolate by hot filtration and washing. The product contains 2.3 wt % Ti, 63 wt % Cl, 20 wt % Mg, and 9.9 wt % diisobutyl phthalate.

For the polymerization of propene, triethylaluminum is used as a cocatalyst, and phenyltriethoxysilane as an additional external Lewis base stereoregulator. Potential corrosion problems have been substantially reduced by using the chlorine-free triethylaluminum.

By ca.1980, commercial plants were achieving 20 – 30 kg polymer per gram catalyst, and the isotactic index progressively improved from 88 to 99 %.

These catalysts form the basis of advanced technology plants in which deashing (removal of catalyst residues) and atactic polymer separation can be eliminated entirely [168]. It is estimated that by 1998 70 % of global plants will be operated with such “high mileage” technology. Coupled with this is the possibility of exploiting the good handling characteristics of the Spheripol coarse particle-form powders, thus dispensing with the need for expensive pelletizing plant. The segregation of added powdered stabilizers during storage and the problems of handling such mixtures are addressed, to some degree, by Montell's Valtec products, in which the additives are treated to provide a hard coating on the polymer. Some customers still prefer to handle traditional granules from an extruder.

2.3.3.3. *Homogeneous Catalysts [163],[169], [170],*

Metallocenes have been used for 30 years as model compounds for Ziegler – Natta reactions. They were quite unsuitable for commercial reactors because of their extremely low activity and poor stereocontrol. In the early 1980s, Sinn and Kaminsky serendipitously discovered that replacing triethylaluminum with methyaluminoxane (MAO) enormously increased the polymerization rate of ethylene when combined with a substituted zirconocene dichloride. There were several other major hurdles to overcome before a suitable candidate was available for use in PP plant. Low stereospecificity succumbed to a more stereorigid metallocene, and the poor reactivity and polymerization temperature restrictions were solved by further molecular tailoring. Trials were conducted in 1996 on commercial-scale streams.

Metallocences suitable for isotactic PP manufacture generally seem to be based on zirconocenes supported on inert solids to preserve particle size and shape. This makes them more compatible with the advanced process technologies of the major operators who refer to them as “drop-in catalysts”. Sophisticated chemistry (Figure 23) is needed to prepare the base metallocenes, many of which have been made only in laboratory-scale equipment. An additional constraint is the heavy patenting, already amounting to some 900 applications since 1984. Hoechst, Exxon, Fina, Mitsui, and BASF together hold half of these.

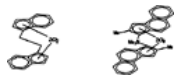

Figure 23. Hoechst (left) and BASF metallocenes

The great versatility of single-site metallocenes has been demonstrated in trials that yielded improved and novel polymers. It remains to be seen how many of these desirable features can be incorporated in a single catalyst and how easily appropriate changes can be introduced at plant level. Some of these polymer properties are mentioned in Section Homopolymer [169].

2.3.3.4. *Aluminum Alkyl Cocatalysts.*

Nowadays, aluminum alkyls are purchased from outside manufacturers. In the case of DEAC systems it is necessary to tailor the chlorine content to the type of $TiCl_3$ used. As the various alkyls interact rapidly, it is simple to adjust the composition by adding $AlEt_3$ or $AlEtCl_2$:

$$2\,AlEt_2Cl \rightleftharpoons AlEtCl_2 + AlEt_3$$

Triethylaluminum is made commercially from aluminum, ethylene, and hydrogen. Care should be taken to avoid the high levels of the hydride that are sometimes present. Trimethylaluminum is available from outside maufacturers, as is methyl aluminoxane (MAO). The latter is made by the controlled addition of water to the trialkyl with various techniques to moderate the potentially violent reaction. One method uses a salt hydrate, such as such as hydrated potassium aluminum sulfate (potassium alum). Various grades of MAO are available, possibly due to the different manufacturing methods and the polymeric nature of the product, which contains linear and cyclic short chains.

2.3.4. ***Hydrogen***

Pure hydrogen is used at concentrations of 0.05 – 1 vol % based on propene to control molecular mass. Polar and highly unsaturated contaminants should be excluded, as usual.

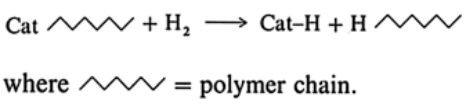

where ⌇⌇⌇ = polymer chain.

2.4. Polymerization Mechanism

Many kinetic and theoretical models have been applied to Ziegler catalysts. Only a very small proportion (< 1 %) of the titanium atoms in violet $TiCl_3$ are active for stereospecific polymerization. Cossee proposed that an active site is generated at those surface titanium atoms which have one chloride vacancy and one exposed and weakly bound chloride anion, the remaining four chlorine atoms being firmly held in the lattice. These sites are located on the lateral faces of violet $TiCl_3$ crystals where the protruding chlorine is alkylated by aluminum alkyl (Fig. 24 A). An incoming propene molecule approaches this active site (Fig. 24. B), whose immediate crystal geometry controls both the initial coordination and the configuration as the monomer inserts into the titanium – alkyl σ-bond (Fig. 24 C). Accordingly, this is a template type [171] polymerization controlled by the surface shape, and not by the previously inserted monomer unit. Atactic polymer is formed at more open surface sites having two vacancies

or two weakly bonded chlorine atoms. While there are a number of alternative proposals [163], [172], [173] the essentials of the Cossee – Arlman model are widely accepted.

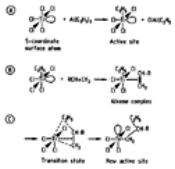

Figure 24. Cossee – Arlman model for polymerization site

A) Alkylation of lateral surface chlorine; B) Propene complex; C) Insertion into Ti – C_2H_5 σ-bond

Since each active site on the crystal is unlikely to be unique, a distribution of imperfections is to be expected. Sometimes Ziegler – Natta catalysts are referred to now as multisite catalysts to emphasize this wide distribution of polymer chain lengths and stereoregularities.

In $MgCl_2$ supported titanium chloride systems, 10 – 20 % of the titanium is thought to be in the form of active sites, substantially more than in Solvay's improved catalyst, which contains ca. 0.8 – 2.7 % of the titanium in active form. Surface geometry still controls stereoregularity, but Lewis bases are much involved in this regulation [174].

The new (1996) metallocene single-site catalysts have unique structures which secure precise control and definition of the active site. Consequently, a narrow molecular mass distribution and better control of chain irregularities, now including new types of chain defects, may be obtained with selected metallocenes/aluminoxanes. These catalysts can be used in solution, but manufacturers usually prefer them supported to secure better control of particle size and shape, together with freedom from reactor build-up [170]. Apparently certain supports can be used without spoiling catalyst or plant performance. Hence the name drop-in catalysts to emphasize the ease with which they can be introduced into existing plant. Their considerable potential and performance are being tested at the pilot-plant stage [175].

2.5. Industrial Processes

In the laboratory, it is a fairly simple matter to make a sample of PP from a given catalyst. All that is necessary is to suspend the solid transition metal compound in dry, inert-gas-blanketed, pure heptane containing aluminum alkyl cocatalyst and then bubble in propene gas at about 60 °C and atmospheric pressure. Polymer forms as a white solid permeating the catalyst particles which give an overall red-pink hue (with $TiCl_3$). Elevated pressures increase reaction rates according to a first order law.

Industrial manufacturing plants must incorporate many other stages to secure salable product: some of these complications are eliminated in modern high-technology plants by using advanced catalysts. Here the residual catalyst and atactic levels can be sufficiently low not to discolor polymer nor to generate odor and smoking. Although older plants still using the traditional process make high quality products, their higher cost base means that they are becoming increasingly uncompetitive, except in some highly specialized and low-tonnage applications. Some 70 % of global PP plants have installed new "High Mileage" processes.

The following sections describe one older slurry type of plant and four state-of-the-art advanced processes. From these it will be possible to visualize a number of hybrid alternatives to suit particular catalyst characteristics [178]. Copolymerization is dealt with separately for each system in Section Copolymerization, though for convenience the process flow sheets illustrating homopolymerization (Figs. 26, 27 and 29) include the bolt-on copolymer plant.

2.5.1. *Suspension Homopolymerization (Early Diluent Processes)*

Details of the individual steps shown in Figure 25 vary between manufacturers as regards process conditions and type of equipment.

Figure 25. Schematic of early PP manufacturing process

Polymerization. For continuous systems, which have largely superseded batch polymerization, the mean residence time of catalyst in the reactor(s) is usually 1.3 – 3 h. Individual autoclaves started out as 10 – 30 m^3 vessels, but as stream capacities increased sizes up to 100 m^3 have been installed. Another common practice involves linking reactors together in series or in parallel configurations. The maximum output rate of a reactor is determined by the heat removal system, which no longer relies simply on water- cooled polymerizer jackets. Internal and external cooling circuits provide the necessary extra cooling at these high rates (> 0.3 t h^{-1} m^{-3}).

It is only necessary to meter catalyst into the first reactor of a cascade system, both because the activity decay is fairly small, and because it is not desirable on quality grounds to generate a second family of particles downstream. Propylene and hydrogen feeds to each reactor maintain the required pressure and gas compositions to secure the correct polymerization rate and molecular mass. Most of the reaction diluent is metered into the first vessel in the train, but some can be added further downstream to minimize fouling as the polymer concentration increases. Slurry viscosity increases rapidly during the latter stages of polymerization as the swollen polymer particles approach close packing. This point is strongly dependent on the size and shape of the catalyst particles. Seldom will the slurry concentration be allowed to exceed 42 wt % unless a very low density diluent is used.

As a rule, polymerization temperatures lie between 50 and 75 °C; the desirability of using a higher temperature being countered by decreased catalyst stereospecificity causing increased solubles and poorer properties. While reactor pressures of 0.5 – 1.0 MPa typified earlier plants, intensified processes expand this to 2 MPa.

Propene Flashing and Recovery. Two ways of removing and recovering unreacted propene, stages b and k in Figure 25, are practiced commercially. Most commonly, slurry from the last reactor discharges into a heated vessel at lower pressure to flash off propene vapor and other volatiles. After cooling and condensation, fractional distillation recovers pure propene as an overhead stream with a bottom discharge of propane and some diluent. In certain cases it is more economical to return the flashed propene to an adjacent plant for use there. The alternative scheme simply allows polymerization to continue without feeding any fresh propene to the latter part of the reaction train. Propene in solution is gradually converted into polymer as the polymerization rate falls exponentionally in the first-order reaction. This simple scheme entirely eliminates the recovery stage, but it does call for additional reactors working at low rates.

Catalyst Removal (Deashing). Catalyst removal, steps c and d in Figure 25, involves the extraction of highly polar particulate $TiCl_3$ from a hydrophobic polymer particle. Procedures differ appreciably, but the first step always converts titanium and aluminum residues into complexes or alkoxides that are soluble in the diluent, usually by heating to 60 °C with 2 – 20 % of an alcohol, though acetylacetone has also been used. The choice of ethanol, propanol, or butanol is influenced by the diluent boiling point and subsequent recovery schemes.

Step d, Figure 25, shows a widely practised arrangement whereby water washing is used to transfer all these catalyst complexes into an aqueous phase, leaving behind a purified polymer slurry.

$$\underset{\text{(in hydrocarbon)}}{TiCl_3 \cdot (ROH)_4} + H_2O \longrightarrow \underset{\text{(in aqueous soln.)}}{Ti^{3+} + 3\,Cl^-}$$

It is vital to work with either a strongly acidic or strongly alkaline complexing medium to repress hydrolysis and polycondensation, which impair the extraction efficiency. This technique normally yields polymer containing 10 – 30 ppm of Ti, 10 – 40 ppm of Al, and 20 – 40 ppm of Cl residues. Very pure polymer can be made in an alternative form of step d. Here, the polymer is isolated by filtration, and is repeatedly washed by reslurrying in diluent/alcohol mixtures to remove all solubilized catalyst without introducing water. This process adds to the load on diluent and alcohol recovery, but PP quality improves.

Centrifuging. Centrifuging the deashed slurry (step e, Figure 25) removes most of the free diluent, but the wet cake can still hold up to 50 wt % (dry basis) of diluent containing dissolved atactic polymer. Depending on the grade of polymer being produced, it is sometimes beneficial to introduce a washing stage with diluent while centrifuging to reduce the amount of soluble polymer in the final product. Centrifuges can be of the filtering (basket) or decanting type according to the powder characteristics.

Drying. Drying the wet cake sometimes requires more than one stage, particularly in processes using high-boiling solvents. Steam distillation removes most of the diluent to leave an aqueous slurry from which the polymer is readily isolated by filtration. A conventional drier removes the small amount of water still adhering to the hydrophobic

particles. More elaborate driers, operating under a nitrogen atmosphere, remove diluent directly in flash drying, fluidized beds, or hot surface contact type systems. Before the advent of high-technology systems (Section Spheripol Process), some manufacturers had already improved catalyst stereospecifity such that the low level of atactic polymer in the diluent could be left in the product. In this case, the total slurry can be dried without previous centrifugation. This avoids all atactic polymer handling problems. All these drying stages must use nitrogen blankets throughout to minimize hazards with flammable hydrocarbons, as well as to protect the polymer.

Extrusion and Pelletizing. Powder leaving the driers at about 100 – 120 °C can be fed directly to a pelletizing extruder, in which it is stabilized and converted into dense granules. As intermediate powder storage is often necessary, common practice introduces a cooling stage to 70 – 80 °C to prevent adventitious oxidation when antioxidants are absent.

Diluent and Alcohol Recovery. An essential part of the schematic plant shown in Figure 25 is solvent and alcohol recovery, and atactic polymer separation (stages h – j). Diluent from the centrifuge contains soluble polymer and traces of alcohol. This mixture, together with liquid from the driers, is separated by multistage distillation. Pure diluent is then recycled to the polymerizers, while viscous atactic polymer separates out as a heavy end. Wiped film evaporators can be used to strip the last traces of diluent from highly viscous solutions.

Most of the alcohol is recovered from the first few water washes by fractional distillation (step h) followed by azeotrope splitting.

2.5.2. *Bulk Polymerization in Liquid Propene*

A special case of the slurry is to use liquid propene, with its poorer solvent power, instead of an inert diluent. Polymerization rates increase considerably with pressures around 3 – 4 MPa. Stirred autoclaves incorporating evaporative cooling, and loop reactors provide good heat-transfer rates. The chemistry parallels that described in Section Suspension Homopolymerization (Early Diluent Processes), but the engineering demands are greater [178]. After polymerization, the catalyst is solubilized with polar complexing agents to permit extraction from the polymer by countercurrent washing with liquid propene. Hybrid plants are known which use liquid propene for polymerization, followed by flashing and resuspension in heptane for polymer deashing.

2.5.3. *Solution Polymerization*

In contradistinction to polyethylene manufacture, solution polymerization at high temperature is rarely practiced for isotactic PP. Special catalysts are necessary to minimize the customary reduction in stereospecificity above 100 °C, but the activity and stereoregularity still remains too low to dispense with catalyst extraction and removal of atactic PP. The original Eastman – Kodak process makes grades for some special markets, but there are no plans for further new plants along these lines. While polymerization at ca. 150 – 200 °C should permit some useful recovery of polymerization heat, overall costs are high for isotactic PP. It is reported [179] that the solution process is being used to make atactic PP, for which it seems more suited. Commercial plants dedicated to make only atactic PP have been announced by Himont (Canada), El Paso (Texas), and Huls (Germany).

2.5.4. *Spheripol Process [180]*

Figure 26 depicts a modern state-of-the-art plant based on Himont technology [180] for making homopolymer and impact copolymers (Section Impact (Block) Copolymerization) with supported catalysts. The latter are sufficiently active not to require any catalyst extraction, nor removal of atactic polymer because of the high stereospecificity. Homopolymerization takes place at ca. 70 °C and 4 MPa in liquid propene circulating round one or more loop reactors. A single axial flow agitator in each loop maintains high flow rates to ensure good heat transfer to the water-cooled jackets, whilst also preventing any polymer particles settling from the slurry [181]. Typically the PP concentration is ca. 40 wt %.

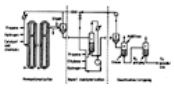

Figure 26. Spheripol process

a) Loop reactors; b) Primary cyclone; c) Copolymer fluidized bed; d) Secondary and copolymer cyclone; e) Deactivation; f) Purging

Continuously metered catalyst, triethylaluminum, and a Lewis base stereoregulator such as a dialkyldimethoxysilane are fed into the reactor to maintain polymerization and stereocontrol. The initial few seconds of polymerization with a new high-activity catalyst particle are quite critical to secure good performance. For this reason, some processes have a prepolymerization stage in which the catalyst components react at lower temperature and monomer concentration. This can be either a batch or continuous pretreatment which produces only small amounts of polymer (< 100 g/g) in the catalyst. This prepolymerized catalyst is then fed into the loop reactor as usual. Mean residence time in a single polymerizer is 1 – 2 h. Two loop reactors can be operated in series to narrow residence time distributions, modify the polymer, and increase output.

A continuous stream of polymer slurry discharges through a heated zone for the first stage of pressure let down in cyclone (b). For homopolymers, this connects directly to the secondary cyclone (d), bypassing the copolymerization unit (Section Impact (Block) Copolymerization). Unreacted propene flashes off from the first cyclone and is condensed with cooling water and recycled to the reactor. A compressor is required for gas from cyclone (d). Polymer powder from the cyclone is fed into vessel (e) for deactivation with small amounts of steam and undisclosed additives. Residual moisture and volatiles are removed by a hot nitrogen purge in vessel (f) before conveying polymer to storage silos for conventional finishing as stabilized powder or extruded pellets.

Provided the expensive extrusion stage can be dispensed with by selling powder, then it is claimed that the capital cost of Himonts Spheripol process is only 50 % of the suspension process described in Section Suspension Homopolymerization (Early Diluent Processes) [180].

2.5.5. *Hypol Process*

Mitsui Petrochemical, co-inventor with Himont of the supported catalyst system, has developed an analogous cheap process using bulk polymerization with their own supported catalysts. It differs from the Spheripol system in that batch prepolymerization is used with washing. Two conventional stirred polymerizers are used in series, with heat removal by evaporative cooling of liquid propene in the reactors. Slurry is then discharged into a stirred, heated flash vessel, where propene is recovered from polymer, as in the Himont system.

2.5.6. *Gas-Phase Processes*

In gas-phase processes, gaseous propene is contacted with solid catalyst intimately dispersed in dry polymer powder. Industry uses two different methods of carrying out this reaction depending on the chosen method of heat removal. The Union Carbide/Shell process uses an adaptation of the Unipol polyethylene fluidized-bed system. BASF and Amoco use mechanically agitated dry powder beds with evaporative cooling in vertical and horizontal autoclaves, respectively.

BASF Novolen Process [182]. Figure 27 shows the BASF continuous process for making homopolymers, impact copolymers, and random propylene – ethylene copolymers using high-activity, highly stereospecific catalysts. The reactor vessels, of 25, 50, or 75 m^3 capacity, are equipped with proprietary helical agitators, which give excellent agitation. Homopolymerization needs only the primary reactor, into which the catalyst components are fed. These must be very well dispersed in the powder bed to avoid build-up. The reaction conditions of 70 – 80 °C and 3 – 4 MPa ensure that the monomer phase is gaseous in the reactor. Low concentrations of hydrogen are used to control molecular mass over wide ranges. The temperature is controlled by removing gaseous propene from the reactor head space, condensing it with cooling water, and then recirculating it back into the reactor, where its evaporation provides the required cooling, as well as further aeration of the stirred powder bed. Each tonne of polymer made requires ca. 6 t of liquid propene to be evaporated as coolant.

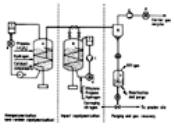

Figure 27. BASF gas-phase Novolen process

a) Primary reactor; b) Copolymerizer; c) Compressors; d) Condensers; e) Liquid pump; f) Filters; g) Primary cyclone; h) Deactivation/purge

Powder and associated gas discharge continuously from the primary reactor dip tube directly into a low-pressure cyclone (g). Propene carrier gas from this cyclone is recycled to the reactor after compression, liquifaction, and, sometimes, distillation. The powder then passes to a purge vessel where a deactivator quenches all residual catalyst activity, and nitrogen strips out traces of propene from the hot powder. From here powder is conveyed into silos for stabilization and extrusion into granules. BASF also offers a post-granulation steam-stripping package to remove any oligomers and oxidized residues from the granules for demanding applications.

BASF pioneered their gas-phase process with commercial production in 1969. The products made, Novolen 1300 series, were based on high molecular mass total work up polymers (i.e., containing atactic PP and catalyst residues) having reduced stereoregularity. Today, such grades still find niche markets, although they are vulnerable to competition from random copolymers. Production is to be phased out shortly.

BASF also uses their process with a cheaper second-generation catalyst $TiCl_3/AlEt_2Cl$, which then requires an additional dry-powder dechlorination stage.

Amoco – Chisso Stirred-Bed Process. Collaboration between Amoco and Chisso resulted in joint licensing arrangements from 1985. This process uses a horizontally stirred reactor, Figure 28, instead of the vertical helical agitator of the BASF process. Condensed recycled monomer sprayed into the top of the reactor provides cooling, while uncondensed monomer and hydrogen injected into the base maintain the gas composition. Figure 28 also includes a fluidized-bed deactivation system for use with second-generation, chlorine-rich catalysts. Amoco – Chisso claim that their reactor achieves some degree of plug flow, roughly equivalent to that of 2 – 3 fully back mixed reactors in series [183].

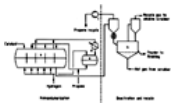

Figure 28. Amoco – Chisso gas-phase process

a) Horizontal reactor; b) Fluidized-bed deactivation; c) Compressor; d) Condenser; e) Hold/separator tank

Unipol – Shell Fluidized-Bed Process [184]. The Unipol – Shell plants, commissioned in 1986, combine technologies from Union Carbide and Shell. Most conspicuous in this process, Figure 29, is the tall fluidized-bed reactor with its expanded upper section to reduce gas velocity and powder entrainment. Continuous feeds of catalyst components, comonomer, if any, hydrogen, and propene are thoroughly mixed in the dense-phase fluidized bed of powder. A large cooler in the gas recirculation loop removes all the reaction heat from the considerable gas flow. In this system the fluidized bed is said to behave as a fully back mixed reactor, without undue separation of coarse particles. No mechanical agitation is needed. Reaction conditions are reported as < 88 °C and < 4 MPa.

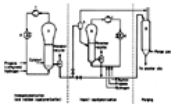

Figure 29. UCC/Shell – Unipol fluidized-bed process

a) Primary fluidized bed; b) Copolymer fluidized bed; c) Compressors; d) Coolers; e), f) Discharge cyclones; g) Purge

Product powder and associated gas are discharged from just above the distributor plate by timed valves into a cyclone separator (e) and then directly to a purge vessel (g) to remove residual monomer. Neither catalyst removal nor atactic polymer extraction is necessary with the modern Shell catalysts used in the Unipol process.

2.5.7. *Copolymerization*

Copolymers of propene with other α-olefins account for ca. 30 % of all PP sales. There are two distinct classes of copolymer having different application areas. Random copolymers, obtained by copolymerizing mixtures of propene and other α-olefins have lower melting points and improved clarity. Impact (block) copolymers, made in a two-stage polymerization process, are high-impact-strength grades that contain dispersed propene – ethylene elastomers.

2.5.7.1. *Random Copolymerization*

Random copolymers are made in the same way as homopolymers, but a mixture of propene and the comonomer is used in place of pure propene. Usually, polymers containing 2 – 6 wt % of combined ethylene cover most applications. Sometimes the resulting copolymers are referred to as “statistical” because heterogeneous Ziegler catalysts rarely yield truly random copolymers [157] with mixed feeds of monomers. 1-Butene can also be used, especially as a termonomer along with ethylene. The higher α-olefins are more expensive and less reactive than propene and can complicate the monomer recycle stage. However, attractive product properties are stimulating manufacture [185], [186].

In the case of suspension polymerization, either in hydrocarbon diluent or in bulk, significant amounts of soluble byproducts add to handling difficulties due to viscosity increases at higher comonomer contents. Drying temperatures must be reduced because of the lower softening point and more cohesive nature of the particles.

Gas-phase systems, especially those having some degree of mechanical agitation, are less adversely affected, though maximum temperatures still need careful control.

Random copolymers often find use in food packaging (see Section Copolymers), where the polymer must not contain high proportions of extractable material. Clarity improves and melting points fall as comonomer and solubles increase. A compromise is reached at about 3 wt % of ethylene. Changes to catalyst and polymerization techniques can raise this by securing a more uniform disruption of the isotactic sequences, which in turn encourages formation of the lower melting γ- crystallites [187].

2.5.7.2. *Impact (Block) Copolymerization [188], [189]*

Impact copolymerization is regarded as the lowest cost process for making toughened PP. It can be visualized as a method for producing an intimate blend of PP homopolymer with a toughening ethylene – propene elastomer. Approximately 10 – 20 wt % of this elastomer suffices for most applications, though the recent availability of products containing 30 – 40 wt % will expand their use.

The minimum requirement for block copolymerization is a two-stage process. In one stage homopolymer is prepared by any of the methods described in Sections Suspension Homopolymerization (Early Diluent Processes), Bulk Polymerization in Liquid Propene, Solution Polymerization, Spheripol Process , Hypol Process, Gas-Phase Processes. The other stage uses a mixed monomer feed of ethylene and propene to produce a largely amorphous, elastomeric phase within the polymer particles. Many other more complex configurations have claimed to be advantageous for certain products, but two stages are generally adequate [188], [189].

Homopolymer, or a very low ethylene copolymer, is always made in the first reactor because the elastomer fragments catalyst particles in diluent systems and produces sticky products in gas-phase reactors. This first stage of reaction, using propene and hydrogen as molecular mass controller, is indistinguishable from conventional homopolymerization. In batch slurry systems, propene is metered into a reactor until the amount of polymer formed is 80 – 90 % of the intended final make. The remaining 10 – 20 % is made in the second stage as follows. Normally the propene concentration is allowed to fall, by continued polymerization, to a predetermined value before introducing ethylene, propene, and possibly hydrogen. No further catalyst is needed for the second stage, and the reaction temperature might be lowered to reduce the amount of soluble polymer diffusing out of the particles. The high reactivity of ethylene promotes rapid copolymerization in a more exothermic reaction. Venting and polymer isolation follow the procedure already described in Section Suspension Homopolymerization (Early Diluent Processes), but filtration and atactic polymer separation are more difficult because of additional amounts of viscous soluble copolymer.

Gas-phase systems using modern catalysts avoid much of this difficulty. Figure 27 shows a typical BASF continuous process using two stirred reactors for making impact copolymers [182]. Active polymer and gas from the first reactor discharge directly into the copolymerizer, operated at 1 – 2 MPa lower pressure and lower temperature. Normally, the propene/ethylene/hydrogen ratio is adjusted such that the rubbery copolymer made in this reactor contains 40 – 60 wt % ethylene, depending on the particular grade. This gas composition is not readily condensed with cooling water alone, so a compression stage is added to liquify some of the recycled gas returned to the reactor base. Copolymer powder, with its associated carrier gas, discharges into the low-pressure cyclone separator (g), which passes this carrier gas to recompression for return to the copolymerizer. Residual monomer in the copolymer is removed in a combined purging and deactivation vessel (h) before conveying powder into silos for extrusion.

Analogous gas-phase techniques are used by other manufacturers. Amoco – Chisso (see Fig. 28) add a second horizontal stirred reactor when making impact copolymers, and El Paso also add a similar polymerizer after their bulk first stage. Unipol – Shell introduce a second, but smaller, fluidized bed (Fig. 29), operating at ca. 1.5 MPa, for their impact copolymers, thus taking advantage of enhanced copolymerization activity. Sumitomo's new process also uses two fluidized beds. Himont (Fig. 26) and Mitsui Petrochemical retain their normal bulk polymerization for the first, homopolymerization step, but use gas-phase copolymerization in a fluidized bed for producing impact copolymers. Such dry powder reactions overcome the problems arising from extraction of soluble copolymer species in liquid-based processes. At the same time, copolymer diffusion to the surface should be reduced by the lower degree of swelling in dry powder systems.

While copolymers containing 40 %, or more, of elastomer can be handled satisfactorily in plant processes, their cohesiveness calls for more care than with homopolymers. Himont patents suggest that their fluidized-bed technology recognizes this aspect. Overall, there is complete unanimity that a gas-phase copolymerization stage is preferred for this class of copolymer in modern plant.

2.5.8. *Product Finishing*

Considerable capital is involved in the finishing areas of all plants after the dry powder stage. The great majority of this powder is extruded with a variety of additives, pelletized, tested, and then stored in various packages for sale. Direct sale of suitably stabilized reactor powder circumvents some, but not all, of these costs.

Extrusion and Pelletization. Proprietary equipment from specialist manufacturers will always be used for high-throughput plants operating at 5 – 20 t/h and 250 °C. In selecting extruders the following factors are relevant (in addition to costs): powder morphology, product MFI range, melt filtration needs, devolatalization, turn down ability (i.e., efficient operation at low rates), temperature control, mixing needs, and additive feeds. The types of extruder available include long, single-screw machines and two-stage systems in which powder is melted in a continuous high-speed mixer which discharges the melt into a short, pressure-generating extruder. Twin-screw machines with corotating and segmented screws are increasingly used for additional flexibility. Versatility and output can also be increased by installing a gear pump between the die plate and the discharge end of the screw in any of these systems.

Pelletization takes two forms. Extruded laces from the die plate are quenched in water to give rigid strands. High-speed cutters then chop the continuous strands into 2 –

5 mm lengths to produce so-called lace-cut pellets. In the other system, high-speed knives rotate against the extruder die plate to cut off short lengths of the molten extrudate. Either the insulated die plate is immersed in water, or the molten pellets are instantly solidified by being flung into a water ring quench zone to prevent cohesion. These die face cut granules are increasingly common in large plants.

Salable Powder. Densification is avoided entirely for those outlets willing to use suitably stabilized powder. Spherical particles, up to ca. 0.3 mm diameter, have been available for many years, but in most applications pellets are preferred because of their better housekeeping (spillages, dust, etc.), easier handling in standard systems, and absence of concern about possible dust explosions due to powder attrition. Emergence of coarse Spheripol powders from Himont, 1 – 5 mm diameter, should dispel some of these concerns provided adequate and nonseparating stabilization systems are established with mechanically robust coarse particles. This will entail some additional manufacturing costs, and perhaps more product constraints, but this expenditure should be less than that for capital and energy intensive extrusion and pelletization operations. Himont is constructing a PP plant in Brazil, due for completion in 1991, which is claimed to have no extrusion facilities. However, these have subsequently been installed.

Additive Incorporation. Extrusion provides an efficient and convenient way of incorporating antioxidants and other modifiers into polymers. Metered amounts are mixed with the main powder feed to the extruder and homogenized in the molten polymer. Good mixing is essential to prevent extrusion of unstabilized product at high temperature. Nitrogen blanketing is common practice. Highly specialized and low-tonnage formulations are either made on a separate plant having appropriately sized equipment, or the customer himself adds concentrates in the form of stabilizer master batches.

Liquid additives, such as some antistats, can be metered directly into the melt, given good subsequent homogenization in the extruder.

2.5.9. *Additives*

Stabilizers [190]. Like most hydrocarbons, PP must be protected against oxidation, particularly above 100 °C. The dominant reaction is chain scission by free-radical attack at the tertiary carbon atoms of the backbone. This generates hydroxyl and carbonyl groups, accompanied by yellow discoloration and brittleness. Incorporating radical scavengers considerably delays the onset of embrittlement and increase of the MFI. Antioxidants are always incorporated into PP before or during extrusion at manufacturing plants. Usually a peroxide decomposer, such as a phosphite, is introduced to improve melt stability, together with hindered phenols to confer long-term aging resistance. Typically, the amount of each of these additives is 0.05 – 0.2 wt %, but in more aggressive environments the levels might rise to 0.5 %, with thiodipropionate esters as peroxide decomposers. The choice of stabilizer is influenced by the application, for example, interaction with spin finishes [191], resistance to extraction by hot water, and regulatory constraints.

Traces of copper or cobalt powerfully accelerate thermal oxidation, especially at sustained temperatures above 60 °C. In these cases a chelating agent such as a bishydrazone is added [192].

Continuous outdoor exposure also requires protection from the damaging effects of UV radiation. The severity of embrittlement increases at high temperature and with rainfall in some polypropylenes. Atmospheric pollution can deposit protective layers on the surface. Ultraviolet-absorbing compounds, such as alkoxybenzophenones and hydroxybenzotriazoles for thick sections, and hindered amine light stabilizers (HALS) for films, provide protection at contents of 0.1 – 1.0 wt % [193]. Screening with some fine pigments or carbon black is effective, but restrictive on appearance.

Gamma radiation, used for packaged food and sterilization of medical equipment, degrades PP, causing embrittlement. The effects can be mitigated by the careful choice of stabilizer.

Antistats. The familiar static charges and resulting dust deposits on many plastics can be overcome in PP by incorporating 0.2 – 1.0 wt % of polar additives. Suitable materials, such as polyether fatty amide and fatty amine condensates and glyceryl monostearate reduce the surface resistivity from > 10^{13} to ca. 10^{7} Ω. The mechanism involves slow diffusion of the additive to the surface where it picks up atmospheric moisture. Hence, time and humidity are important. At 50 % RH most of the recommended additives are satisfactory, but at 30 % RH the fatty amine derivatives are advantageous. A few days storage at room temperature is sufficient to develop the conducting surface layer.

Compositions containing moderate amounts of conductive carbon black are also effective, as are small amounts of polyacetylenes. The latter, and the incorporation of small amounts of fine metallic wire, are too expensive for PP applications.

Antacids. Most of the earlier PP formulations contained 0.05 – 0.2 wt % of calcium stearate to prevent traces of acidic catalyst residues from corroding customers' equipment. It behaves as a mild slip agent and suppresses attack of HCl on certain antioxidants. Zinc, magnesium, and sodium stearates have also been used, and more recently synthetic hydrotalcite (a hydrated magnesium aluminum hydroxycarbonate). Lower levels of such additives are still widely used.

Slip and antiblock agents are mainly used in film production where high-speed handling requires a controlled reduction in frictional forces. 0.1 – 0.5 wt % of oleamide, erucamide, glyceryl monostearate, etc., usually suffices. Fine, spherical particles of silica incorporated at the pelletization stage prevent layers of film adhering to each other during storage.

Crystallization Nucleants. Although pure PP has a final melting point of 176 °C, the melt supercools by 50 – 60 °C before crystallization sets in at ca. 115 °C. Small amounts of certain additives [194] nucleate the melt, raising the crystallization temperature by 10 – 15 °C. Smaller spherulites result, sometimes having different structures. Advantages of nucleated grades include shorter injection molding cycles, higher moduli, greater transparency, and higher heat distortion temperatures. Suitable additives include dibenzylidene sorbitol and substituted variants, di-*tert*-butylbenzoate salts, sodium benzoate, finely divided talc, sodium diaryl phosphates, and some pthalocyanine pigments. These materials are somewhat selective in their effects, i.e., some are particularly good for transparency improvement, but less effective for modulus enhancement. Careful balancing of the overall additives recipe is also essential because both adverse and beneficial interactions occur between some of the components [195].

2.6. Compounding

Compounding includes incorporating mineral fillers, glass fibers, elastomers, flame retardants, pigments, and carbon black. It is a specialized operation dealing with relatively short run lengths in specified compounding extruders or mixers. For this reason, and for concern about cross contaminating unmodified PP grades, compounding plants usually operate separately from main PP production plants.

Substantial property enhancements enable polypropylene-based products to compete in quite demanding areas where unmodified PP would be inadequate (see Table 18). Examples include tough front and rear bumpers for cars, rigid coupled glass compositions for washing machine tubs, dense grades for outdoor garden furniture, and temperature-resistant compositions for car under hood applications. Long-fiber-reinforced PP secures further increases in toughness and stiffness over short-glass-fiber compounds [196]. Fibers with a minimum length of between 0.8 and 3 mm, depending on the coupling between matrix and fiber, are likely to be satisfactory. These form an interlocking skeletal structure which dissipates the impact energy over a large area of the molding.

Table 18. PP compounds — property enhancement

Property	Homopolymer	Impact copolymer	Compound	Composition
Toughness, kJ/m^2 (N Izod, 0 °C)	3.0	10.0	> 50	rubber + filler (car spoilers)
Modulus, GPa	1.5	1.0	7.6	coupled glass
	1.5	1.0	2.7	chalk loaded
Heat distortion temperature	105	90	160	coupled glass
(0.45 MPa), °C	105	90	130	talc filled (heater casings)

The combination of lower production rates, specialized equipment, and dense additives mean that PP compounds are always more expensive than natural grades.

2.7. Properties

2.7.1. *Homopolymer*

Basic mechanical properties are largely influenced by molecular mass and molecular mass distribution (MMD), chain stereoregularity, and processing conditions, which introduce orientation, strain, etc. Various additive packages must be tailored to the application to achieve good and durable performance.

Molecular Mass. At low shear rates melt viscosities are proportional to the 3.4 power of the weight-average molecular mass. At the higher shear rates found with most processing equipment, the dependence is still greater than a power of 2 because much chain entanglement still exists. As a result of this sensitivity, and the direct relevance to processing, the melt flow index (MFI) is always used to describe the viscosity characteristics of each polymer grade. The dependence of melt viscosity on shear rate is also related to MMD as a result of the strong influence of very long chains. Hence MFI is often measured at two different shear rates to give some idea of MMD. Increasing

importance is being attached to a more comprehensive rheological characterization [197], [198].

A typical, medium-flow polymer has an MFI (230 °C/2.16 kg) value of 3 dg/min, corresponding to a M_w of 460 000 and M_n of 54 000. The practical grade range of 0.3 – 50 MFI covers applications from sheet and pipe extrusion to high-speed fiber spinning.

Heterogeneous Ziegler – Natta catalysts yield broad MMD polymers having M_w/M_n of 5 – 10, the lower values being associated with the more recent Solvay and supported catalysts. Very narrow MMDs are still made by preferentially cleaving long chains in an extruder, usually with the assistance of organic peroxides. Such grades, commonly referred to as controlled rheology (CR) polymers, reduce extrusion and molding orientation caused by high molecular mass species with their long relaxation times [199]. Note that reworked polymer is likely to have a higher MFI and a correspondingly narrower MMD than the starting material.

Stereoregularity. Very high stereoregularity is now achievable with modern catalysts, even without removing any soluble byproduct. In some applications processing is helped by a small lowering (detuning) of the isotactic index. This is readily achieved by minor changes to the polymerization recipe, and by introducing small amounts of copolymerized ethylene.

Metallocene-Based PP. An expanded range of products is imminent following the encouraging metallocene developments. Further pilot trials and customer assessment of processing, product, and economics is needed before precise definition is possible. The advantages already seen include higher extrusion rates for films and fibers, enhanced clarity, absence of oligomers, superior property balances, and the ability to control further melting point reductions by copolymerization without significantly increasing the low level of xylene-soluble material.

Other Properties. Table 19 lists the properties of several polymers measured by standard procedures, with some copolymer data included for comparison. Care must be taken in translating these exact values to fabricated products because properties are dependent on built in stress and orientation. Increasing the molecular mass, i.e., lowering MFI, raises toughness somewhat, but the modulus and hardness decrease with the lower crystallinity.

Table 19. Mechanical and thermal properties of PP Origin "Propathene" (with permission of ICI Chemicals & Polymers)

Property	Method	Temperature	Homopolymer				Impact copolymer		Random copolymer
			GSE 16	GWM 22	LYM42	LZM 55/52	GWM 101	GWM 213	LXF 301
Melt flow index (230 °C/2.16 kg), g/10 min (dg/min)	ISO 1133		1.00	4.00	12.00	33.00	6.00	2.00	7.50
Tensile yield stress, MPa	ISO 527		33.50	34.50	34.50	34.00	27.00	23.50	25.00
Flex. modulus, GPa	ISO 178 (10 mm/min)		1.45	1.50	1.50	1.55	1.15	1.00	0.85
Izod impact strength, kJ/m^2	ISO 180 (0.25 mm notch radius)	23 °C	4.50	4.50	4.00	3.00	9.50	no break	5.00
		0 °C	3.00	3.00	2.50	2.00	5.50	10.00	2.50
		–20 °C	2.00	2.00	2.00		3.50	7.50	
		–40 °C						5.00	
Instrumented drop weight impact strength, J	ICI method	23 °C					7.00	10.00	10.00
		0 °C					6.00	11.00	2.00
		–20 °C					6.00	11.00	
		–40 °C						9.50	
Embrittlement temperature, °C	ICI method		> 23	> 23	> 23	> 23	–20.00	–45.00	
Rockwell hardness	ISO 2039/2, ASTM D785 R scale		93.00	95.00	95.00	95.00	90.00	75.00	
Vicat softening temperature (10N force), °C	ISO 306A, BS 2782: 120A		154.00	154.00	154.00	154.00	147.00	143.00	135.00
Heat distortion temperature, °C	ISO 75/A & /B								
A – 1.8 MPa			65.00	65.00	65.00	68.00	60.00	50.00	46.00
B – 0.45 MPa			100.00	100.00	100.00	102.00	95.00	90.00	71.00

Some general and electrical properties of polypropylene are as follows:

Density, kg/m^3	
20 °C	905
200 °C	770
220 °C	760
240 °C	740
260 °C	710

Coefficient of linear expansion, K^{-1}	
20 °C	1.1×10^{-4}
80 °C	1.7×10^{-4}

Specific heat, $kJ\ kg^{-1}\ C^{-1}$	
20 °C	1.93
50 °C	2.04
100 °C	2.37
	4.0

150 °C

Glass transition temperature T_g, °C	
isotactic	–13 to 0
atactic	–18 to –5

Flammability	
UL 94, 3 mm thick	UL 94 HB
ASTM D635, 3 mm	23 mm/min
FMVSS 302	
1 mm	43 mm/min
2 mm	38 mm/min
3 mm	20 mm/min

Thermal conductivity, W m^{-1} $°C^{-1}$	0.209
Combustion heat, kJ/g	44.0
Autoignition temperature, °C	360
O_2 index (ASTM D28) at 600 °C	3.5
Relative permittivity at 1 kHz (IEC 250)	2.25
Volume resistivity, 1 min, Ω · m (IEC 93)	> 10^{15}
Electrical strength (20 – 80 °C, 10 – 20 s, 3.2 mm, IEC 243)	> 12 kV/mm

Dissipation factor tan δ (IEC 250)	
50 Hz	0.00015
1 kHz	0.0002
1 MHz	0.00015

The final melting point of commercial PP lies in the range 160 – 170 °C, with purified polymer reaching 176 °C. However the safe upper working temperature limit should be 100 – 120 °C, depending on the stress, with short excursions up to 140 °C being permissible. This comfortably extends the upper operating range available with high-density polyethylene by ca. 20 °C, which has a similar room temperature modulus.

Favorable long-term creep characteristics under load permit useful increases in the stacking height of PP crates or other pelletized containers. The repose angle of PP moldings on wood or canvas, unpredictably 30 % higher than for HD polyethylene, is allegedly a useful stabilizing feature in mechanical conveyors.

The crystalline and nonpolar nature of PP confers good resistance to a wide range of aqueous and polar media, including emulsifier solutions with their strong stress cracking abilities. Powerful oxidizing agents and highly aromatic or chlorinated solvents are too aggressive for safe storage in PP containers.

The ease with which homopolymers can be drawn into high strength structures has been heavily exploited in the fiber and film industries (see Section Fibers and Flat Yarns).

2.7.2. *Copolymers*

Random Copolymers. Incorporating 2 – 7 wt % of ethylene into the main chain lowers the overall crystallinity, gives a broader softening range with reduced melting points (Figs. 30 A, 30 B), increases the fraction of soluble polymer, and improves transparency and surface gloss. These changes are dependent on the amount of comonomer and its distribution along the chain. Table 19 shows how the modulus decreases for a random copolymer containing ca. 3 wt % of combined ethylene.

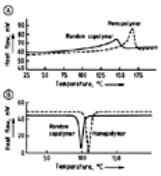

Figure 30. A) DSC heating endotherm; B) DSC cooling exotherm

Courtesy of the "Propathene" Business, ICI Chemicals & Polymers

Several applications take advantage of the lower melting point to make thermoweldable film. Ideally, it would be useful to have a low melting range without undue increases in the proportion of xylene-soluble residues. This is claimed to be possible with butene copolymers, perhaps because the solubles are produced at semi-exposed catalyst sites more easily accesible to ethylene [186]. Some advantages are also claimed for propene – ethylene – butene terpolymers [180], [185]. Butene helps the development of low-melting γ-triclinic crystallites. Random copolymers are somewhat tougher than homopolymers, and do not exhibit the familiar stress-whitening behavior of the tougher copolymer impact grades.

Impact Copolymers. Section Impact (Block) Copolymerization describes block copolymerization systems for making high impact strength grades of PP. These are often referred to as block copolymers, toughened polypropylene, or more pedantically heterophasic copolymers [180].

Essentially, these products are best viewed as an intimate dispersion of elastomer in a PP matrix. The particulate rubber phase confers good toughness, while the homopolymer matrix is responsible for retaining good high-temperature performance and adequate stiffness. In this respect they are far superior to random copolymers, and become the first choice when impact strength and rigidity are called for. In extruded pellets or molded articles the rubbery copolymer phase is well distributed in the form of droplets 1 – 5 µm in diameter. Such a structure (Fig. 31) is produced during melting and recrystallization. Powder direct from the reactor has elastomer permeating throughout the particle in a more continuous form. Many of the complex interactions which control the stability of this phase structure and the resulting toughness, await further study, but current views are very well discussed by VAN DER VEN [189].

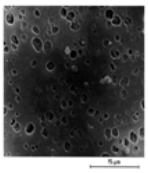

Figure 31. Electron micrograph of high impact copolymer molding after hexane surface wash

Toughness, stiffness, and softening points are given in Table 19 for two impact copolymers containing ca. 10 and 20 % of elastomer. Gas-phase polymerization systems are able to increase the rubber content to 40 %, but the products are not available in large commercial quantities at present. These impact copolymers naturally contain more soluble polymer than homopolymer, and this second phase reduces the translucency. Severe deformation causes some whitening or blushing [200] in the stressed regions, particularly in very high impact grades. In other respects the rules relating processing and properties to MFI, nucleation, controlled rheology, stabilization, etc., are the same as for homopolymer.

2.7.3. *Elastomer Blends with Polypropylene*

Originally, PP was toughened by blending with ca. 10 wt % of polyisobutylene (PIB) elastomer. In some situations the same technique is used today, with ethylene – propene rubber (EPR) or ethylene – propene – diene rubber (EPDM) replacing PIB.

As rubbers are incompatible with PP, considerable mechanical work is necessary to secure the vital fine dispersion of elastomer. Good blending also requires a reasonable viscosity match between the two components. Common practice is to use a two-stage process in which the elastomer is first compounded with PP to give a 50 – 70 wt % masterbatch of the rubber as free flowing granules. Such masterbatches can be purchased at extra cost. Good mixing devices are essential for the final compounding step with PP. Twin-screw extruders or high-speed mixing devices perform satisfactorily.

Well-prepared blends have physical property combinations only just short of those of directly polymerized impact copolymers. However, the latter are generally preferred on account of consistency and grade versatility. Elastomer blending is also part of the broad spectrum of compounds referred to in Section Compounding. Here they are used to uprate the toughness of homopolymer and impact copolymers in these complex formulations. Some manufacturers also introduce HD polyethylene, in addition to rubber, to secure further toughness improvements. In these cases, electron microscopy reveals that in well-dispersed systems the polyethylene is always located inside and towards the center of these rubber droplets. This encapsulation of PE makes it behave more like rubber towards the matrix, thereby increasing toughness at the expense of modulus in such ternary blends.

2.8. Uses

Polypropylene is readily processed in conventional equipment used for other thermoplastics. Injection molding, commonly using screw-type reciprocating machines, accounts for 40 – 50 % of all applications. Extrusion processes account for the remainder with domination by fiber and film.

A more detailed picture is obtained by analysing the North American 1996 sales data (in 10^3 t/a) listed in the following [202] :

Blow molding	
Containers	74
Others	5
Total blow molding	79
Extrusion	
Coatings	5
Fibers and filaments	1463
Film (up to 0.25 mm)	
Oriented	503
Unoriented	111
Pipe and conduit	
Sheet (over 0.25 mm)	110
Wire and cable	5
Others	43
Total extrusion	2240
Injection molding	
Appliances	139
Consumer products	565
Rigid packaging	523
Transportation	186
Others	131
Total injection molding	1544
Export	503
All others	1177
Total	5543

2.8.1. *Injection Molding*

An exceptionally wide range of injection-molded products stems from the rigidity, toughness, and chemical and temperature resistance of PP. Examples include automotive trim and ventilation components, bottle crates, industrial containers, washing machine tops and tubs, kitchenware, tool handles, domestic waste systems, and small boat hulls. Such a variety of articles requires careful grade selection according to the required impact strength of the product and the melt flow constraints of the processing equipment. An MFI (230 °C/2.16 kg) range of 2 – 20 dg/min meets most needs. Increasing the MFI not only assists mold filling and reduces cycle times, but it helps manufacture of complex moldings having high flow ratio geometries (flow ratio is the ratio of longest path to the section thickness). The inevitable fall in impact strength at high MFI can be recouped, in some cases, by moving from homopolymer to impact copolymer grades.

While molding conditions vary with the size and complexity of the article, the following machine settings will cover most products [199]:

Melt temperature:	230 – 275 °C
Injection pressure:	55 – 100 MPa
Mold temperature:	40 – 80 °C, lower still for thin walled items

The low moisture absorption of PP largely dispenses with any need to dry granular feedstock. Neither do gaseous or carbonaceous decomposition products form in extruder dead spots. Instead, polymer remaining there becomes progressively more fluid through chain scission. Machine shot weight capacities normally refer to performance with polystyrene, i.e., the polystyrene yardstick figure. The lower melt density of PP (0.75 g/cm^3) lowers the machine's weight capacity by ca. 25 % when molding PP. Plasticizing capacity is slightly greater than for polyethylene, but is lower than for amorphous polystyrene. Machine outputs with PP benefit from the ability to eject moldings

at higher temperatures than with their lower softening counterparts.

Linear post-molding shrinkage of PP is 1 – 2 %, of which 85 % occurs within the first 24 h. Distortion in molded products reflects internal stress caused by nonuniform cooling and polymer orientation. The latter is especially pronounced in surface skins where the viscous melt stretches as it is pulled along by the advancing melt front adjacent to cool walls. Controlled rheology (CR) polymer, containing fewer of the orientation-prone long chains, reduces this type of warping [199].

An important practical aspect of PP concerns its ability to form strong integral hinges, such as in a lidded box, where the lid is permanently attached to the base by a thin web of polymer along the whole of one edge (Fig. 32). This web, about 0.25 – 0.6 mm thick, can be produced directly in the molding process or by post forming operations. An initial flexing of the hinge while slightly warm induces the correct molecular orientation to permit repeated opening, even at subzero temperatures. In the laboratory, such hinges have withstood 23×10^6 flexes without failure [203].

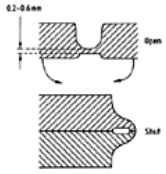

Figure 32. Integral hinge

Courtesy of the "Propathene" Business, ICI Chemicals & Polymers

Various types of structural foamed moldings can be produced having a cellular core sandwiched between solid integral skins giving an overall density in the range 600 – 800 kg/m^3 [204]. These lightweight, but rigid parts can vary in size from a paint brush handle to a sack pallet or car body panel. Foaming results from using a tumble-blended feedstock of granules and a chemical blowing agent. Alternatively, high-pressure nitrogen can be injected into the melt. Molding techniques are altered to achieve extremely rapid polymer injection rates into the mold. This prevents premature foaming before the solid skin forms. Foaming in the core arises by allowing the melt pressure to fall rather than maintain the customary pressure hold to keep the mold full.

Four alternative foam processes exist:

1. In the ICI sandwich molding process, different polymers are used for the skin and core by using two separate injection barrels. As the skin is unfoamed, high-quality surfaces result.
2. In the Varitherm Process (BASF, Krauss Maffei), improved surface finish is achieved by using a hot mold (100 – 120 °C) followed by rapid cooling after a short shot of the melt (i.e., an injection of polymer insufficient to fill the mold).
3. In the counterpressure method the mold is pressurized prior to filling to prevent degassing at the melt front.
4. In hollow section molding (e.g., CINPRES) nitrogen is injected in a continuous stream into the mold at the same time as the polymer melt. Under appropriate conditions, hollow or ribbed moldings are produced when gas flows preferentially down thick, hot sections of the molding. This technique is of increasing interest.

2.8.2. *Blow Molding*

The extrusion blow molding process for PP copies that already well established for low- and high-density polyethylene. However, application to PP has been slow while suitably viscous and tough polymers were developed at attractive prices. At present, the important areas are smaller packaging containers which take advantage of good transparency, good form stability up to 140 °C for hot filling or sterilization, and greater freedom from environmental stress cracking with aggressive products. Brake fluid reservoirs exemplify this. Very large containers are dominated by high-density polyethylene.

Extrusion blow molding employs two linked processes [205] in which an extruded molten parison is transferred to a mold for inflation with low-pressure air at 0.4 – 1.0 MPa. High molecular mass polymer, MFI(230 °C/2.16 kg) 0.4 – 2.0 dg/min, combined with low melt temperatures at the die of 200 – 210 °C, provide a suitably stiff melt. If the melt stress in a heavy PP parison exceeds ca. 20 kPa, then the resulting tension thinning will be troublesome as regards parison stability and uniform wall thickness.

Injection blow molding replaces the extruded parison with an injection-molded preform. Blowing takes place in a second mold which maintains the high-definition neck of the preform. This process eliminates scrap and makes it easier to produce wide mouth containers having improved dimensional accuracy and appearance.

Injection stretch blow molding is a further refinement which introduces a plunger stretching stage to elongate a conditioned parison or preform just before the final blow (see ⟶ Plastics, Processing). Better mechanical properties and enhanced transparency stem from partial orientation. This technique is sometimes referred to as the melt phase process.

In stretch blow molding, the extruded or molded preform is carefully conditioned to just below the polymer melting point before stretching. Axial stretching of the preform with simultaneous or almost immediate blowing in the mold produces a very high quality container. Correctly carried out, this process secures increased rigidity and toughness, high transparancies, and reduced permeabilities as a direct result of significant amounts of orientation. The more complex technique increases capital and operating costs, but it is of growing interest [206] and is already being applied extensively to biaxially oriented PETP bottles.

Coextruded parisons can be used as precursers for blow molding when special barrier properties are required [207]. In particular, a central layer of ethylene vinyl alcohol copolymer (EVOH) to reduce oxygen diffusion, combined with moisture protective outer layers of propene – ethylene random copolymer, give containers suitable for storing oxygen-sensitive food products.

2.8.3. *Fibers and Flat Yarns*

Very large amounts of polypropylene homopolymers are used for fiber manufacture in Europe and the United States. These applications exploit the wide range of physical forms, including increasing amounts of versatile nonwoven fabrics [208-210]. Monoaxial orientation can be applied to conventional spinneret yarns, as with polyamides and polyesters, and to flat tapes made from extruded film. These differ in the following respects (Table 20).

Table 20. Polypropylene fiber processes

Process	Filament count, tex *	Product
Long spin	0.2 – 3.0	high-tenacity monofilament; drawing integral or separate; high output
BCF yarn	0.2 – 2.0	special case of long spin making only bulked continuous fiber
Spunbonded	0.2 – 2.0	venturi haul off; no 2nd stage draw. bonded mat output
Shortspin	0.2 – 40	compact unit; tow stretched and cut in line for staple
Melt blown	0.002 – 0.02	low orientation, very fine fiber; only bonded mat output
Fibrillated yarn	110 – 500	oriented slit film; fibrillated for baler twine, rope, etc.
Weaving tape	ca. 110	nonfibrillated slit film for carpet backing, sacks, etc.
Strapping tape	500 – 1000	thick, oriented tape as a steel alternative

* tex = weight in grams of 1000 m of yarn; equivalent cylindrical fiber diameter in $\mu m = 37\sqrt{tex}$.

Spinneret Type Yarns. Here melt is extruded through a die plate perforated by many small holes to generate individual thread lines. In the long spin process [211] (see ⟶ Fibers, 4. Synthetic Organic – Staple Fiber Production Process), which has integral spinning and finishing stages as well as out of line drawing options, there may be 50 – 250 holes per spinneret. An air cooling gap of 2 – 5 m is needed between the die plate and the wind-up roll. Line speeds up to 1000 m/min at the spinning stage, increasing to 3000 m/min during solid drawing over hot rolls, call for complex and expensive haul off, drawing, and wind-up sections. Product may be packed off as continuous yarn, tow, or staple. Suitable polymers have MFI(230 °C/2.16 kg) in the range 12 – 25 dg/min, with narrower molecular mass distribution grades offering some advantage with such high rates of melt draw. These are high-throughput plants, best suited to long runs of a single grade of product.

In a variation of this technique [212], the integrated spunbonded process (see ⟶ Fibers, 3. General Production Technology – Process Technologies for Spunbonded Nonwovens), yields a continuous bonded mat of partially oriented yarn as a consequence of some draw in the venturi type haul off operating at up to 5000 m/min. There is no further drawing as the fiber is collected as a mat on the take-off belt for conversion into nonwoven fabrics. These are used, for example, in geotextiles, furnishings, carpet backing, etc. The balance of drawdown rate and orientation in this process is influenced by the polymer molecular mass distribution.

Figure 33 depicts a compact shortspin process for staple production using a die containing ca. 40 000 holes. Cooling air jets freeze the melt within about 20 mm so that Godet haul off rolls can be placed only 1 m away from the die. Orientation is achieved by the in-line stretching of this tow in an air oven, followed by chopping in line to produce staple [207]. Broad molecular mass distribution polymers give the necessary melt strength, while MFI(230 °C/2.16 kg) requirements increase from 7 to 20 dg/min as the filaments become finer. These find use in geotextiles, carpet face fiber, and diaper cover stock applications. A useful feature of PP fabrics made from fine fibers is the ability of underwear garments to reduce claminess by transferring moisture away from the skin to an outer absorbent layer. Polypropylene based fabrics are also used for sports clothing and socks, but the low melting point of PP calls for special attention to the heat setting of the average domestic iron.

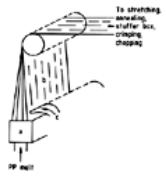

Figure 33. Shortspin process

a) Spinneret; b) Quenched fibers; c) Air-jet cooling Courtesy of the “Propathene” Business, ICI Chemicals & Polymers

In the third, and most recent production process (Fig. 34), air is blown through a very fluid polymer melt maintained at high temperature to assist chain scission in the spinning process itself [212], [213]. Polymer MFI(230 °C/2.16 kg) at the nozzle can range from 35 to 300 dg/min or higher. The specially designed die, having a row of nozzles along its width, sprays a stream of short, low-orientation fibers onto a moving belt where they are bonded together and wound onto a roll. This process produces nonwoven fabrics which are of increasing importance for industrial wipes and surgical wrap, and for a variety of filtration applications, such as face masks made from bonded mats of very fine fibers.

Figure 34. Melt blowing process

a) Extruder; b) Gear pump; c) Heated die; d) Hot air manifold; e) Collector; f) Take-off roll

Flat Yarns [214] (see also ⟶ Fibers, 4. Synthetic Organic – Polyolefin Film Yarns and Split Yarns). In this process (Fig. 35), extruded film is the basis of a cheap and different type of product to that achievable with filament-based yarns. Melt from a conventional slot die is rapidly quenched with cold water or water- cooled steel rolls to give low- crystallinity film, 50 – 250 µm thick. Razor blade type knives then slit this broad film into strips 5 – 20 mm wide, which are oriented in a hot air oven at 120 – 180 °C by drawing between Godet rolls at ratios in the range 5 : 1 to 10 : 1. Both film thickness and width diminish in this operation by approximately the square root of the draw ratio.

Figure 35. Flat tape/yarn production unit

a) Wind up frame; b) 3rd Godet stand; c) Hot air annealing oven; d) 2nd Godet stand; e) Hot air orientation oven; f) 1st Godet stand; g) Simple film costing unit; h) Extruder; i) Film die; j) Water- cooled chromium-plated rolls or water bath; k) Slitting unit; l) Starting drum; m) Individual precision flangeless bobbin wind up

Courtesy of the “Propathene” Business, ICI Chemicals & Polymers

Fibrillated Tape (see also ⟶ Fibers, 4. Synthetic Organic – Polyolefin Film Yarns and Split Yarns). While flat yarns are quite strong in the direction of draw, they are very weak in the transverse direction. With normal homopolymer, simply twisting high-draw-ratio tape induces spontaneous fibrillation. A web of irregular fine fibers is formed as the film splits down its length. The resulting product resembles a coarse sisal twine, for which it is a good substitute. Such fibrillated yarns can be tailored to specific applications by controlling draw ratios, thickness, and polymer. Nowadays, pin fibrillation, in which a corotating spiked roll contacts the film just after slitting, offers more precise control over this step. The products are familiar as baler twine, string, and some ropes.

Weaving Tapes. High strength tapes, without fibrillation, are made by controlling extrusion, orientation, and polymer type. Introduction of finely divided chalk, together with some polyethylene, usually as the chalk masterbatch, is particularly effective in reducing splitting. Impact copolymers are a more expensive, and rarely used, alternative to homopolymer for this application.

Both tapes and fibers contain residual stress from the extrusion and orientation processes. This results in linear shrinkages of up to 8 % immediately after manufacture, but after storage for a few days at room temperature the value falls to about 2.5 %. For those applications, such as weaving tapes for carpet backing, where zero shrinkage is often needed, a heat setting stage is essential. Out-of-line storage of the tapes on bobbins at 130 – 135 °C for several hours often suffices.

Thick, 0.3 – 0.9 mm, oriented tapes are established alternatives to steel bands in many strapping applications. They are generally made by the water quench process operating at draw ratios of 9 – 10. Thicker tapes are made from foil or individually extruded film tapes. High molecular mass, MFI(230 °C/2.16 kg) 0.4 – 2.0 dg/min, homopolymers and block copolymers with added chalk and polyethylene largely suppress fibrillation. A further precaution here is to emboss the finished tape with a diamond pattern. A useful aspect of PP strapping tapes is their higher elongation and elastic recovery which helps them to remain tight on packages prone to shrink or settle.

2.8.4. *Film*

Packaging film is an important outlet for PP, accounting for about one sixth of polymer sales in Europe, and rather more in Japan (see Section Compounding). By far the major proportion of this is biaxially oriented PP (BOPP), an established replacement for regenerated cellulose. Although the equipment needed for orientation is complex and expensive, it secures superior physical properties.

Good clarity in unoriented film is dependent on very rapid quenching of the melt to reduce spherulite size. The chill roll extrusion process (see ⟶ Films – Extrusion) for making 30 – 250 µm film extrudes a mobile, hot melt from a slot die onto water-cooled, highly polished steel rolls. Extrusion temperatures of 240 – 270 °C combined with an MFI(230 °C/2.16 kg) of about 7 dg/min minimize surface irregularities in the melt. Random copolymers containing 2 – 3 wt % of combined ethylene enhance toughness, clarity, and gloss, with some loss of modulus. These soft-feel transparent films are used as textile packages, bakery wraps, and document sleaves at thicker gauges.

Tubular quenched (TQ) film made in a downwardly extruded bubble (Fig. 36) uses some external water cooling which achieves reasonable clarity in film for packaging. In contrast, the less common tubular film from the air- cooled upwardly extruded bubble is more hazy, but finds use in the production of weaving tapes (Section Fibers and Flat Yarns). For most film applications, slip and antiblock additives are needed.

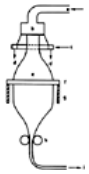

Figure 36. Tubular quenching process

a) Melt from extruder; b) Die; c) Air ring; d) Air flow; e) Film; f) Water quench weir; g) Cooling skirt; h) Nip rolls; i) Dewatering, slitting, wind up etc.

Biaxially oriented polypropylene (BOPP) film [215], [216], typically 12 – 40 µm thick, has excellent transparency, toughness, and stiffness, and is widely used in cigarette and biscuit packages, potato chip bags, and carton overwrap. Film modulus is three times that of the unoriented cast product. These complex film manufacturing plants are usually proprietary systems based either on a bubble process, or on a flat die (stenter) process. In the bubble process (Fig. 37), thick cast tube is reheated and then blown into a bubble with simultaneous stretching in the transverse and machine directions to produce a balanced film. The stenter process (see ⟶ Films – Biaxial Stretching) uses thick cast sheet which is reheated and conditioned for drawing. Longitudinal and lateral stretching are performed separately and can be adjusted to secure the required orientation balance. The typical area draw in these processes is 30 : 1 to 50 : 1.

Figure 37. BOPP bubble process

a) Cast tube; b), c) Stretching heaters; d) Bubble guides; f) Air ring; g) Wind up, slitting, etc.

It is impossible to heat seal such highly oriented material because unacceptable shrinkage then occurs. Low-melting surface coatings overcome the problem by reducing the sealing temperature to a level which the films can safely withstand. Suitable materials are vinylidene chloride copolymers, polyethylene, poly(1-butene), as well as butene, propene, and ethylene co- and terpolymers. Some of these may be applied as aqueous emulsions, but common practice is to coextrude the low-melting polymer as an outer layer of the cast tube or sheet. After orientation, this layer can be less than 1 µm in thickness. These more compatible polyolefin coextruded coatings also simplify scrap recovery and recycling without fear of the coating instability inherent in halogenated polymers.

Polypropylene film, however made, is not receptive to printing inks and adhesives because of the low polarity. Mild surface oxidation by corona discharge treatment overcomes these problems.

2.8.5. *Foil and Sheet*

Slow heat transfer rates in sheet and thick foil call for extended cooling times at the extrusion stage. Foil, up to about 0.7 mm thick, can be made by the cast tube process (Section Film), but the more versatile three roll stack system (Fig. 38) is able to produce sheet with thicknesses of 0.3 – 15 mm. Recent techniques involving additional and larger diameter rolls extend this range to 40 mm. Thicker sheets are made by compression molding of granules or large molded slabs, and by injection – compression molding. This can make 2 m square slabs with thicknesses up to 240 mm. High molecular mass homo- and copolymer, MFI(230 °C/2.16 kg) ca. 0.3 – 1.0 dg/min, provide higher dimensional stability in slow- cooling, thick-section melts whose temperature at the die exit is ca. 210 °C. Higher flow grades with MFI(230 °C/2.16 kg) of up to 5 dg/min may be used for thin foil, depending on the available equipment.

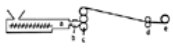

Figure 38. Three roll stack extrusion

a) Extruder; b) Slot die; c) Polishing stack; d) Nip rolls; e) Wind up

Commercial sheets, normally up to 2 m wide provide stock for fabricating into welded tanks for storing corrosive liquids. They are also used as linings for steel and concrete containers, for ducting construction, and as an inert construction stock for chemical plant. It is possible to bond woven glass or polyester fabric to the heated surface for subsequent further strengthening using traditional glass fiber lamination techniques. Punched sheets can be biaxially oriented to generate heavy duty meshlike structures for geotextiles [217].

Thick sheet, as made, has a nonuniform structure across its thickness as a result of considerable differences in cooling rate.

2.8.6. *Extruded Pipe*

Compared with polyethylene, tubing applications for PP are quite small. However there are several areas where PP's increased resistance to aggressive media and its better performance at elevated temperatures are advantageous.

Extrusion at ca. 225 °C is straightforward; a central torpedo is used to make the hollow pipe [218]. Homogeneous melts require an extruder having a length-to-diameter ratio of 25 : 1 and a compression ratio of 4 : 1 to reduce the risk of trapping air in the extrudate. Pipe up to 300 mm outside diameter is generally vacuum sized in a water bath: semi-molten pipe passes through a finely perforated steel tube having vacuum applied to its external surface. Larger pipes are made by pressure sizing where nitrogen or air is injected via a small capillary into the torpedo to expand the pipe to the sizing die form. Both homopolymers and impact copolymers are used in the MFI(230 °C/2.16 kg) range 0.3 to 2.0 dg/min, increasing to 4.0 for drinking straws and ballpoint pen tubing.

Copolymers, providing superior impact strength with some flexibility, find use in domestic waste systems. Homopolymers have higher resistance to stress. It is essential to have a heavy duty stabilization system which will withstand extraction for applications involving aggressive and hot environments, such as chemical plants and effluent pipes. Manufacturers offer guidance [218] on polymer selection and pipe design conforming to the various pressure classes in BS 4991, and to BS 52 254 and BS 5255 for domestic waste pipes and fittings. Creep data are available [218] for pressure pipes predicting performance for up to 50 years.

Pipe jointing uses all the common techniques, with the exception of solvent welding. Ring seals with rubber inserts allow some movement in domestic waste pipes. Also used are flanged joints, hot plate butt welding, and socket fusion welding, the latter being optionally supplemented with hot-gas bead welding. Compression joints involving copper alloys should not be used at temperatures exceeding 60 °C because of reduced life expectancy.

2.9. Environmental Aspects

2.9.1. *Recycling*

The structure and properties of PP make it well suited to recycling operations. There is no fear about cross-linking, nor of complications with plasticizers or chlorine-rich species. Re-extruding PP usually lowers its molecular mass and narrows the molecular mass distribution, particularly in aged feedstock. This can be helpful in some applications that require enhanced flow. Adverse factors which must be assessed include some loss of strength from reprocessing/ageing, oxidation stability, discoloration, and contamination.

PP Manufacturers. Manufacturers usually produce small proportions of PP which do not meet their full quality requirements. Examples include deviations in MFI/MMD, antioxidant levels, copolymer characteristics, and gels. Provided these faults are within predetermined limits in the quality protocol, the material may be recycled in small proportions, sometimes with further corrective adjustments to the composition. This operation calls for additional work, and can reduce plant output by having to process the polymer twice. For this reason, and in those cases where recycle is not permitted or the deviations are too large, manufacturers either reclassify the product to a lower grade, or sell to an outside compounder for further reprocessing. This minimizes disturbances to continuous manufacturing processes.

A second source of much lower grade material, arises when manufacturing conditions are severely disturbed. This can be due to instrumental and human error, blockages, equipment failure, or extrusion disturbances. In these cases scrap can vary from substandard reactor powder to blocks of polymer from extruder dumps. External reprocessors convert such products into surprisingly useful materials by blending and recompounding. To them, the availability of essentially pure PP is the prime requirement.

Plastics Converters. Turning PP into moldings, pipe, film, etc., invariably involves some scrap production. Apart from substandard articles, there are sprues from moldings, edge trim from film and sheet, parison waste, and stamped sheet from thermoforming. These are recovered for recycling by the converter himself if the application allows, or more often by an outside compounder. In favorable situations involving long runs of a standard product, the converter segregates waste, grinds it, and then recycles it by blending small amounts with virgin feedstock when specifications permit. Occasionally there is an undemanding application which can accept large proportions of such regrind without difficulty. Storage and segregation problems are reduced when the converter simply collects his scrap together in the form of regrind for sale to a compounder.

The British Plastics Federation (BPF) estimates [219] that there are about 60 reprocessing companies in the United Kingdom handling at least 150 000 t/a of polymer. Forty three of these process an estimated 25 000 t/a of PP [220]. Only a few of these also process recovered scrap into molded products such as horticultural, highway, safety, and street furniture, which do not make excessive demands on polymer behavior. Included in the estimated figure are 15 000 t of recovered used battery cases and crates which can be converted into quite respectable compounds. Most reprocessors only handle single polymers, not mixtures. Their main supply sources are plastics converters and industrial or commercial establishments as suppliers of fairly well defined PP scrap [221].

Recycling PP After a Full Service Life. Nearly half the thermoplastics used in Europe are converted into packaging [222]. In the United Kingdom 13 % of this is based on PP, involving an estimated 142 000 t of polymer in 1988.

Polymers used in United Kingdom for packaging in 1988 (in 10^3 t) broke down as follows [222]:

Low-density polyethylene (LLDPE and LDPE)	510
High-density polyethylene	255

Polypropylene	142
Polystyrene	78
Poly(vinyl chloride)	68
Poly(ethylene terepthalate)	53
Expanded Polystyrene	21
Total	1127

Inevitably, much of this has only a short life before it ends up as waste in some form. Domestic waste contains 5 – 7 wt % (up to 20 % by volume) of such products [220], very little of which is currently recovered. Public concern about the environment naturally includes highly visible plastic litter, which is made unduly prominent by its low density and durability. This is an additional factor accelerating legislation for recovery of all types of packaging, including glass, metal, paper, as well as plastic [223-225].

In the case of PP, notable success rests with recycling old plastic bottle crates and battery boxes in Western Europe and the United States. This is mainly attributable to the ease with which large amounts of a clean single polymer can be collected without cross contamination from other plastics. Cooksons Plastics [226] and others [227], [228] have expanded on lead recovery to retrieve PP battery cases themselves. A mechanised sorting system has to separate stones, metal, wood, string, ebonite paper, and residual acid. The Cookson process involves crushing, several stages of washing, flotation, and screening followed by drying and extrusion. The final products include good-quality PP capable of being used in automotive products and for certain types of battery cases. Success here rests in quality upgrading by adding elastomer to enhance toughness, stabilizers, and antacid incorporation to recoup oxidative stability, and molecular mass tailoring with peroxides for specific applications. Hoechst [229] plan to collect plastic waste from scrap cars, household appliance industries, and presorted packaging waste. The 5000 t/a plant near Cologne should be in operation by mid-1992. ICI studies on scrap recycling from used PP crates stress the need to incorporate additional stabilizer in weathered products, and to allow for some reduction in impact strength for seven-year-old crates.

Recovery and reuse of mixed plastics is more difficult. The real need is to develop applications and fabrication processes for such comingled polymers, which are essentially incompatible and contaminated. Generally, they will be channeled into some massive products, such as posts, construction boards, and other bulky structures where resistance to water and decay is more important than high inherent strength. Some 20 different European schemes aim to make recycled plastic components as wood and concrete replacements for agricultural applications [223], [230]. Comingled scrap can be upgraded by blending with other recycled polymers to assist processing.

2.9.2. *Pyrolysis*

PP can be cracked to liquid hydrocarbons by thermal pyrolysis at 400 – 550 °C. In a recent process developed in Japan [231], zeolite catalysts enable this temperature to be reduced to 200 – 420 °C by using a two-stage heating process. Other polyolefins, such as polyethylene and polystyrene, and some chlorinated plastics can also be pyrolized to liquid products. Process cost estimates are $ 235 per ton of plastic, which yields 0.5 m^3 of gasoline, 0.5 m^3 of kerosene, and diesel products. Plants are expected to be built in Taiwan and Korea in 1991.

2.9.3. *Incineration*

Municipal waste is normally disposed of in landfill sites, with incineration being the next most important method [232]. As suitable landfill sites become rarer and more expensive, pressure is exerted to reduce waste volume either by incineration or by scrap recovery.

Modern incineration technology, with energy recovery in the form of electricity generation, can take advantage of the high calorific value of thermoplastics. Municipal waste calorific values average 10 kJ/g, compared with 30 – 35 kJ/g for coal. At 44 kJ/g, PP and other polyolefins have the same calorific value as fuel oil. Moreover, PP has an exceptionally low sulfur content, making it a very clean fuel with no apprehensions about toxic flue-gas emissions. These incineration plants are costly to build. Even after allowing for the sale of electricity, the net disposal cost in the United States is estimated at 40 – 70 $/t of processed waste [224]. Special burners have been developed to handle atactic and isotactic waste PP feedstock, but more profitable outlets are now available.

2.9.4. *Environmental Interactions*

Overall PP manufacture, like any other product, involves some interaction with the environment. Various essential inputs, such as raw materials and services need to be considered along with all the plant output streams.

Inputs. The principal input for PP is pure propene, with lesser amounts of ethylene for copolymer grades. Energy is required to manufacture these monomers from petroleum feedstocks, and still more is needed for conversion into polymer. At present, most of this power comes from burning fossil fuel which then discharges carbon dioxide and water vapor to the atmosphere. Table 21 compares the overall energy input needed to make various raw materials. This can be expressed either as the weight of oil to produce this energy, or in terms of joules per unit volume. Both values include energy for monomer production, but exclude the hydrocarbon content of the monomer, which is transformed into polymer without combustion. On a weight basis, polyolefin energy requirements are broadly similar to those of most alternative raw materials. On a more realistic volume basis, the strength and lightness of plastics gives energy savings. Usually this is enhanced further in the final article where, for example, thin and strong plastic bottles show savings over their thicker glass counterparts.

Table 21. U.K. energy requirements for the manufacture of natural and synthetic materials [232]

Material	Energy requirements	
	TOE *, overall conversion	kJ/cm^3
Aluminum	5.6	661
Tinplate	1.3	427
Copper billet	1.2	469
Glass bottles	0.5	46
Paper and board	1.4	50
Cellulose film	4.4	293
Low-density polyethylene	1.2	51
PP, old process, with extrusion	1.5	61
PP, new process, with extrusion	1.3	52

* TOE = tonnes of oil equivalent to the energy requirement for overall conversion.

Propene and ethylene monomers themselves are nontoxic asphyxiant gases whose flammable nature calls for the usual extensive precautions. Ignition by static electricity is possible with high-pressure leaks and accidental discharges.

Autoxidation of hot, unstabilised PP powder in storage silos containing air is another known cause of fire.

Effluents. Although catalyst components constitute only a very small proportion of plant materials, both their manufacture and disposal involve noxious substances. Titanium tetrachloride is used for all catalysts, sometimes in considerable excess for high-activity systems. Whereas expensive recycle and recovery stages may be installed as capacities increase, some remaining complexed residues will require disposal. Often, this involves hydrolysis with water to generate corrosive HCl and titanium dioxide suspensions. After neutralization with alkali, these slurries are either discharged as aqueous effluent or settled out and used in landfill. Some incineration is planned. Disposal facilities also need to be capable of receiving occasional faulty catalyst batches. Older plants still using catalyst extraction processes simply use aqueous slurry discharges, or dump concentrated residues from the base of stills.

Lower aluminum alkyls are pyrophoric, react violently with water, and are intensely aggressive towards exposed skin. Safe handling usually calls for full protective clothing. Disposal can be by burning to generate alumina, or quite often, by controlled hydrolysis with water after deactivation with alcohols. Some alkyl manufacturing plants will accept returned waste alkyls for disposal in their own dedicated facilities.

Atactic PP. Older plants needed to dispose up to 5 % of their output as low molecular mass atactic residues from solvent recovery streams. Landfill was a common outlet, with some incineration. Useful low- cost applications appeared in adhesives and roofing felt compositions, which now consume nearly all of the increasingly rare atactic PP byproduct.

Gaseous Effluent. Traditional flare stacks receive most of the voluntary and involuntary process gaseous discharges. Normally, these comprise small flows from distillation column purges which can be dealt with by screened low level flares. These reduce the visual impact of flares which some people find objectionable, especially at night. Some nitrogen purges containing traces of hydrocarbons are discharged to the atmosphere. Scrubbing systems and carbon adsorption beds reduce hydrocarbon emissions, but only a few manufacturers use them because of the cost.

[Top of Page]

3. Poly(1-Butene)

Hartmut Koch and Ralph L. Mawer

3.1. Production

Introduction. Poly(1-butene) is a thermoplastic polymer produced by solution polymerization of 1-butene with a Ziegler – Natta catalyst system. Hüls developed a process during the 1960s and produced the polymer in a 12 000 t/a plant in the Federal Republic of Germany. Mobil started to develop a process in 1968 and, together with Witco, built a plant in Taft, Louisiana with a capacity of 30 000 t/a. Shell (USA) bought the plant in 1977 and is at present the only commercial producer.

The feedstock of pure 1-butene is supplied from the Shell Higher Olefins Plant in Geismar, Louisiana.

The poly(1-butene) plant produces grades specifically developed for pipe and speciality film applications.

Process. The current commercial poly(1-butene) production process consists of seven primary unit operations: (1) polymerization, (2) deashing, (3) polymer/monomer separation, (4) monomer purification, (5) extrusion, (6) pellet blending, and (7) packaging (Fig. 39).

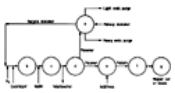

Figure 39. Flow sheet for the production of poly(1-butene)

a) Monomer purification; b) Polymerization; c) Deashing; d) Monomer/polymer separation; e) Extrusion and pelletizing; f) Blending and storage; g) Packaging

In the polymerization step the catalyst system is fed into the reactor with the purified 1-butene recycle stream. Hydrogen is added as a chain terminator to control the melt flow index of the polymer. Ethylene is fed to the reactor when producing random 1-butene – ethylene copolymers.

Polymerization catalyst residues must be removed from the product to preserve its long-term properties. In the deashing step, the poly (1-butene) is mixed with dilute caustic soda to dissolve and extract the catalyst components. The cleaned hydrocarbon – polymer phase and the catalyst- containing aqueous phase are then separated.

During the polymer/monomer separation step, unreacted 1-butene is separated from the poly(1-butene) for recycle to the reactor. Separation is carried out by heating the polymer solution under pressure then flashing off the monomer by rapidly reducing the pressure. The remaining molten polymer feeds the extrusion step, while the vaporized monomer flows to the monomer purification section of the process.

The monomer purification step accomplishes separation of impurities both heavier and lighter than 1-butene. Water is removed by azeotropic distillation. Fresh 1-butene is added to the process at this stage.

In the extrusion step further volatile hydrocarbons are removed, additives for color, product performance, processing, and stabilization are added, and the polymer is pelletized.

The pellets are pneumatically conveyed to the batch blending stage and then transferred to the packaging area. Pellets are typically sold in 450-kg boxes.

3.2. Properties

3.2.1. *Structure and Physical Properties*

The properties of a high-molecular material are essentially determined by the dimensions and structure of the molecular chain (fine structure) and by the thermomechanical history [233]. Since the fine structure of poly(1-butene) can be varied across a broad range by means of the polymerization process, a knowledge of this is essential for the spectrum of properties.

Molecular Structure. Polymerization with Ziegler – Natta catalysts produces both atactic and isotactic polymer. Atactic poly(1-butene) is virtually impossible to crystallize, while isotactic poly(1-butene) has a crystallinity of 40 – 55 % at normal cooling speeds.

The molecular chains in the crystal form helices. Depending on the crystallization conditions, poly(1-butene) forms 11_3, 3_1 or 10_3 helices, which differ in their helical dimensions [234-236] and give rise to different crystalline modifications (see below).

The isotactic content of poly(1-butene) is generally determined by the solubility in boiling diethyl ether ("ether extract"), whereby it is assumed that only the atactic components are soluble. However, infrared spectroscopy and X-ray crystallography have shown that the atactic extract contains crystallizable components, mostly of low molecular mass. A quicker and more reliable method of determining the atactic content is by evaluating the infrared band in the region of 920 cm^{-1} [237]. This leads to the following definition of the atactic content Y_a:

$$Y_a = \frac{\bar{\nu} - 918\,\mathrm{cm}^{-1}}{30\,\mathrm{cm}^{-1}}$$

where $\bar{\nu}$ is the band position of the unknown sample in cm^{-1}.

The advantage of this method is that it is independent of the molecular mass and of the morphology of the sample.

Molecular Mass. The average molecular mass may be determined either by direct methods such as light scattering or osmometry, or by indirect methods such as determination of the intrinsic viscosity [238], [239]. The molecular mass is calculated from the viscosity number using the Mark – Houwink equation [239]. Mark – Houwink equations for isotactic poly(1-butene), which give virtually identical molecular masses are as follows:

$$[\eta]_{n-\mathrm{nonane}} = 5.9 \times 10^{-5} \times M^{0.80}$$
$$[\eta]_{\mathrm{dekalin},135^\circ\mathrm{C}} = 6.2 \times 10^{-5} \times M^{0.79}$$

These equations also apply to atactic poly(1-butene).

The molecular mass distribution can be determined easily by gel permeation chromatography in 1- chloronaphthalene as solvent. Industrial products have molecular nonuniformities $U = (M_w/M_n) - 1$ between 5 and 12 (M_w = weight-average value, M_n = number-average value).

Crystal Structure and Superstructure. Like most polymers isotactic poly(1-butene) can crystallize in various modifications. However, the polymorphism is more marked and of greater industrial significance in poly(1-butene) than in any other polymer. Four crystal forms of poly(1-butene) are known at present and are designated modification I, I′, II, and III [234], [240]. Characteristic physical data of these modifications are listed in Table 22.

Table 22. Characteristic data for crystal modifications of poly(1-butene)

	Modification II (tetragonal)	Modification I (hexagonal)	Modification III (orthorhombic)	Modification I′ (hexagonal)	Amorphous phase	Reference
Helix type	$11_3/40_{11}$	3_1	10_3	3_1		[235], [240], [242]
Helix length per monomer unit, nm	0.187	0.217		0.217		

Helix cross section, nm^2	0.055	0.044				
Specific volume at 25 °C, cm^3/g	1.1364	1.0453			1.1579	[243]
					1.16 for mod. II	
					1.14 for mod. I	[243]
Specific coefficient of volume expansion, $cm^3\ g^{-1}\ K^{-1}$	3.05×10^{-4}	3.13×10^{-4}			8.12×10^{-4}	[243]
Melting point, °C	ca. 125	ca. 140	93 – 98	100		[241], [244] [245][246]
Heat of fusion, kJ/mol	ca. 3.1	ca. 6.7				[247][248] [249]

Modifications I′ and III are formed on crystallization from solution. They are of no industrial importance. However, modifications II and I, which are obtained by crystallization from the melt, are industrially significant. All the crystal forms can be obtained from modification III in a single step [246], as shown in Figure 40. The transitions from one crystal form to another are monotropic. Modification I′ crystallizes preferentially from solution at lower temperatures [240], while at high temperature the orthorhombic modification III is produced [250], [251]. Modification III is also formed by precipitation, e.g., by addition of methanol to a benzene solution of poly(1-butene) [252]. Modifications I′ and III are also formed in solution or suspension polymerization [253].

Figure 40. Monotropic interconversion of the four modifications of poly(1-butene)

The tetragonal modification II and the hexagonal form I are industrially important. As a melt cools, i.e., after most processing operations, form II crystallizes out first. This is thermodynamically unstable and in the absence of external mechanical stresses it changes monotropically into the stable hexagonal modification I after a period ranging from a few hours to several days, depending on the temperature and thermal history [234]. The transformation from II to I is associated with a change in various physical properties, e.g., an increase in melting point, density, [243] and modulus (Fig. 41).

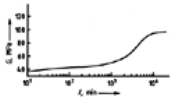

Figure 41. Time dependence of the shear modulus *G* of modification II of poly(1-butene)

Thermal history: 1 mm thick test bar cooled from the melt (160 °C) in air to the test temperature of 20 °C

These changes in properties are primarily due to the change in crystal modification. The degree of crystallization and the mean spherulite size remain virtually constant during the transition from II to I [243]. The crystallite size also remains unchanged [244]. This is a purely crystal – crystal transformation without experimentally detectable melting of modification II crystals.

The rate of conversion, as indicated for example by the half-value time $t_{1/2}$ depends on the thermal history (Fig. 42) and on the temperature (Fig. 42).

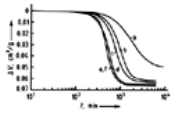

Figure 42. Time dependence of specific volume during transformation (II $\longrightarrow$ I) of poly(1-butene) at 20 °C

a) Quenched; b) Crystallized for 200 min at 92 °C; c) Crystallized for 5700 min at 102 °C; d) Crystallized for 5600 min at 102 °C; e) Crystallized for 7500 min at 106 °C; f) Crystallized for 7200 min at 108 °C

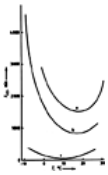

Figure 43. Temperature dependence of half-value time $t_{1/2}$ for the transition II $\longrightarrow$ I

a) Pure poly(1-butene); b) Melt mixture of poly(1-butene) and polypropylene in the ratio of 1: 1; c) Poly(1-butene) with 30 % butyl stearate

Mechanical stresses may significantly increase the conversion rate [254], [255]. Figure 44 shows the influence of constant tensile stress on the conversion rate, monitored by IR spectroscopy. Whereas these measurements show a stress-free sample to have a half-value time of 5500 min, the half value for a sample under a tensile load of 4.4 MPa is only ca. 10 min. Internal stresses and orientations have a similar effect to external stresses. The conversion of II to I is also accelerated by hydrostatic pressure [256]. The conversion rate can also be increased by certain additives, such as polypropylene or plasticizers (see Fig. 43) [240], [257], [258].

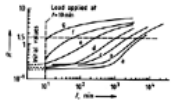

Figure 44. Poly(1-butene) conversion curves for modification II $\longrightarrow$ I for various constant tensile stresses

R is the ratio of the IR extinction of modification I at 920 cm^{-1} to the extinction of modification II at 920 cm^{-1}. The half-value time $t_{1/2}$ (II $\longrightarrow$ I) is reached when *R* has a value of ca. 1.5

a) 0 MPa; b) 0.12 MPa; c) 1.48 MPa; d) 1.91 MPa; e) 2.55 MPa; f) 2.95 MPa; g) 4.23 MPa

The molecular mass and the molecular mass distribution have virtually no effect on the conversion rate. However, the molecular mass does have significant influence on the rate of crystallization of modification II (Fig. 45). The crystallite and spherulite sizes of all crystal modifications can be varied over a broad range by modifying the cooling and crystallization conditions of the melt [248], [259], [260]. Surprisingly, the mean crystallite size and size distribution are not increased by subsequent tempering [259], [260]. This may be a reason for the low time dependence of certain mechanical properties (see Mechanical Properties.).

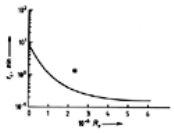

Figure 45. Crystallization time t_c of poly(1-butene) at 78 °C as a function of the viscosity-average molecular mass $\bar{M}_v$

Curve: measurements on fractions, ▪ unfractionated product

The four modifications are distinguished analytically and characterized quantitatively by a combination of X-ray, infrared, and differential thermal analysis [245-247], [261].

3.2.2. *Application Properties*

In normal processing operations, isotactic poly(1-butene) is heated to above its melting point. Therefore it is the unstable modification II which forms immediately after cooling, and is converted to the stable form I during storage or on application of mechanical stress. Since only the properties of form I are of interest from the point of view of applications, the statements below refer only to this modification.

In terms of density, melting point, and glass transition temperature, poly(1-butene) resembles polyethylene rather than polypropylene [262]. This similarity extends to cross-linked polyethylenes, although some of their properties naturally vary, depending on their cross-linking density [263]. Commercial material is around 40 % to 55 % crystalline. Since the crystallization rate is slow by comparison with polyethylene and polypropylene, superstructures which are of interest for applications can be obtained by the choice of cooling conditions [264]. Nucleators increase the rate of crystallization, leading in some conditions to a material that is somewhat more brittle at low temperature [264].

Mechanical Properties. Table 23 lists some mechanical properties of poly(1-butene), polyethylene, and polypropylene. According to the VDE hardness, poly(1-butene) is softer than polyethylene and polypropylene. However, the Shore hardness is virtually the same as for polyethylene. The surface gloss of components made from poly(1-butene) is similar to those made from polypropylene [265].

Table 23. Hardness, rigidity and strength of low-pressure polyolefins [265]

Property	Test method	Polyethylene	Polypropylene	Poly(1-butene)
VDE hardness (60 s), MPa	VDE 0302	44 – 64	61 – 80	25 – 45
Shore D hardness	DIN 53 505	62 – 69	70 – 75	60 – 68
Flexural stress at conventional deflection, MPa	DIN 53 542	26 – 43	34 – 50	15 – 25
Modulus of elasticity, MPa	[a]	900 – 1200	1000 – 1500	500 – 900
Impact strength	DIN 53 453	no break	no break	no break
Notched impact strength, kJ/m^2	DIN 53 453	5 to no break	3 – 15	20 to no break
Brittle point, °C	ASTM D 746–57T	below – 70	+ 20 to – 5	– 10 to – 30

[a] Measured in Hooke range by bending tests.

The rigidity of poly(1-butene) is relatively low. The flexural stress at conventional deflection and modulus of elasticity are lower than for polyethylene (Table 23).

The impact strength and notched impact strength are significantly higher than those of polypropylene. Above 0 °C the values are similar to those for high molecular mass polyethylene, but at lower temperature they are inferior [265]. Nevertheless the low-temperature strength, as characterized by the brittle point, is relatively good [254], [265].

The stress – strain behavior of modification I differs noticeably from that of other polyolefins. The differences are not only in the shape of the stress – strain curve, which for industrial material has virtually no maximum [265], [266], but also in the dependence on time and temperature [267]. Figure 46 shows the yield strength σ_{ys}, measured in an uniaxial tensile test as a function of the stretching time t_s. Whereas the yield strength of polyethylene and polypropylene declines sharply with increasing stretching time, the values for poly(1-butene) remain virtually constant above 0.1 s. The good creep resistance and resistance to stress cracking of poly (1-butene) pipes subject to internal pressure [262] is due to the low time dependence of the mechanical properties.

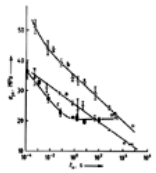

Figure 46. Yield strength σ_{ys}, measured in a tensile test at 22 °C, as a function of the stretching time t_s [266]

a) Polyethylene; b) Polypropylene; c) Poly(1-butene)

Since this low time dependence of the mechanical characteristics is maintained at higher temperatures, poly(1-butene) is particularly suitable for hot-water pressure pipes [263], [268]. Figure 47 shows the burst hoop stress performance of poly(1-butene) pipe under internal pressure according to ISO-SEM (DIS 9080) [269].

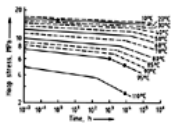

Figure 47. Burst hoop stress performance of poly(1-butene) pipes

As well as the outstanding resistance to stress cracking, the unusually good creep behavior is significant. The tendency to creep, i.e., to undergo plastic deformation below the yield point, is significantly lower than for other polyolefins, even at temperatures up to 100 °C [263], [265], [270]. This is also an essential requirement for the use of poly (1-butene) as a film material, particularly for the packaging sector [270].

Electrical Properties. The electrical behavior of poly(1-butene) does not differ significantly from that of polyethylene and polypropylene [265]. Apart from the usual quantities of stabilizers, it only contains residues of polymerization catalysts. Because of its good resistance to chemicals, outstanding resistance to stress cracking, and flexibility, poly(1-butene) is an important material for special applications in the cable and electrical industry.

Resistance to Chemicals and Gas Permeability. Because of its paraffinic structure, poly(1-butene) is resistant to most common chemicals such as acids, alkalies, salts, alcohols, and ketones. However, it swells or dissolves in hot aromatic and chlorinated hydrocarbons. It is less inert to aliphatic hydrocarbons than polyethylene and polypropylene and more prone to permeation [265], [271], [272]. A detailed table on the resistance of poly(1-butene) to chemicals is given in [270].

Permeability to gases and vapors, in particular water and oxygen, is comparable to that of polyethylene of similar crystallinity. Permeability to carbon dioxide, however, is higher [264].

3.3. Processing

Processing poly(1-butene) on conventional machines presents no problems. Because of its low crystallite melting point and particularly because of the rapid decrease in melt viscosity with increasing temperature, poly(1-butene) can be processed by extrusion (pipes, profiles, sheets), extrusion blow molding (hollow bodies), and injection molding (engineering products, sheathing for metals) [265]. Sheets made by extrusion or compression molding can be shaped hot or cold. A special feature is cold shaping of the relatively soft modification II. Articles shaped from this modification then become increasingly rigid as the conversion to modification I progresses.

3.4. Applications

Pipe. The major outlet for poly(1-butene) is currently in high-performance pipes, which exploit its creep resistance and long lifetime at elevated service temperatures, flexibility, light weight, corrosion resistance, and its ability to be heat welded [263]. It competes in plumbing and radiator heating applications with copper, where progress is mainly limited by resistance to replacing traditional working practises. In the underfloor heating pipe market it competes with other plastics such as cross-linked polyethylene and polypropylene.

The advantage of poly(1-butene) over other polymers in both these applications is its better ability to retain its mechanical and performance properties at elevated temperature. For instance a comparison of tangent modulus, yield stress, ultimate stress, and elongation at break within and transverse to the extrusion direction of the pipes shows a more balanced performance for pipes made of poly(1-butene) [273]. In the last decade the long-term properties of poly(1-butene) pipes and their lifetime at temperatures up to 120 °C have been studied [274] and the methodology for such long-term tests has been developed for various plastics [275-277].

Extrapolation methods give estimated lifetimes of 50 years at 70 °C [273]. A recent evalua-tion of pipes based on polymers as well as traditional materials, such as copper, favored poly(1-butene) [278].

The problem of oxygen diffusion in plastic piping systems, which was given as one explanation of corrosion in metal radiators, tanks, and boilers in the past, has been resolved by the use of multilayer pipes containing an oxygen-barrier layer and the addition of inhibitors to the circulating hot water [278].

Packaging. In the packaging sector, poly (1-butene) has become established in speciality film applications which require resistance to cooking temperatures, and in heat-seal layers for oriented polypropylene, which exploit the high compatibility of poly(1-butene) with polypropylene. The incompatibility of poly(1-butene) in blends with various polyethylenes and ethylene copolymers, such as low-density polyethylene, linear low-density polyethylene, high-density polyethylene and ethylene – vinyl acetate copolymers, forms the basis of peelable seals (see Fig. 48).

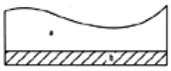

Figure 48. Cross section of a coextruded peel-seal layer on an LDPE substrate

a) LDPE substrate; b) Peel-seal blend LDPE in which incompatible poly(1-butene) particles are dispersed

Peel-seal technology provides a robust pack, yet allows the consumer to open it easily without ripping or tearing. The key attributes of systems based on poly(1-butene) are the broad heat sealing range and the cohesive failure mechanism, which results in an aesthetically attractive package after opening. This is because cohesive failure propagates through the peel-seal layer itself due to the weaker intermolecular forces between the dissimilar materials, and results in clean uniform separation. Additional advantages are good clarity and a clean peel with no stringing [279], [280]. The more conventional system relies on delamination between layers as a result of adhesive failure and is also more difficult to control.

A correlation of peel-seal strength with blend composition and temperature is shown in Figure 49.

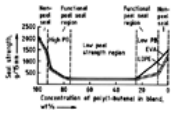

Figure 49. Dependence of peal-seal strength on blend composition

Manufacture of packaging film or sheet incorporating a poly(1-butene) peel-seal blend is possible by all the conventional processes such as film blowing or casting, sheet extrusion or coextrusion, extrusion or coextrusion coating, and lamination.

Poly(1-butene) peel seals are used in flexible packaging such as box liners for cereals, cake mixtures, and biscuits. In combination with more sophisticated barrier films, they are used for processed meat and vegetable packs. They are also used for packaging medical products.

An important growth area for peelable seals is in lids for rigid or semirigid containers. These range from the relatively simple yogurt, milk, and cream pots to individual-portion containers for microwave meals.

Two features of poly(1-butene) peel seals are considered to be of particular importance to the development of the next generation of high-performance packages. The first is their resistance to high-temperature cooking or sterilization. Often referred to as “retortability”, this is an increasingly important feature both in the food and nonfood sectors. Poly(1-butene) peel seals can be made heat resistant by formulation with higher-temperature-resistant materials such as polypropylene or high-density polyethylene.

The second important feature is a distinct whitening effect when the seals are opened because of the characteristic cohesive failure mechanism. This gives a clear indication that the package has been opened or tampered with.

[Top of Page]

4. Polyisobutylene

Wolfgang Immel

4.1. Introduction

Industrially important polyisobutylenes can be divided into five groups:

1. Di-, tri-, and tetraisobutenes
2. Low molecular mass polyisobutylenes with $\bar{M}_n$ = 330 – 1600
3. Medium molecular mass polyisobutylenes with $\bar{M}_n$ = 20 000 – 45 000
4. High molecular mass polyisobutylenes with $\bar{M}_n$ = 75 000 – 600 000
5. Ultra high molecular mass polyisobutylenes with $\bar{M}_n$ > 760 000

History. In 1925 LEBEDEW was the first to prepare polyisobutylene [281], [282]. He obtained products of varying chain length: liquids of oily to honey-like consistency; sticky, doughy materials; or solid, rubber-like materials.

Relatively uniform low molecular mass polyisobutylenes were first obtained by the company de Bataafse [283] by polymerizing isobutene at + 10 to – 10 °C in the presence of metal halides.

High molecular mass polyisobutylenes were first prepared by OTTO and MÜLLER-CUNRADI by polymerizing isobutene below – 10 °C in presence of volatile inorganic halides [284].

Polymerization Mechanism. Even in the early studies, two important aspects of the polymerization of isobutene were recognized: (1) catalysis with initiators of the Friedel – Crafts type, and (2) the inverse dependence of the molecular mass on the temperature. Subsequent studies led to the first formulations of a reaction mechanism [285].

According to these ideas, which essentially remain valid today, the reaction consists of the following steps:

Initiation:

$$BF_3 + HX \longrightarrow H[BF_3X]$$

$$H[BF_3X] + \begin{array}{c} CH_3 \\ | \\ C{=}CH_2 \\ | \\ CH_3 \end{array} \longrightarrow \left[CH_3{-}\begin{array}{c} CH_3 \\ | \\ C \\ | \\ CH_3 \end{array} \right]^+ [BF_3X]^-$$

Chain growth:

$$\left[CH_3{-}\begin{array}{c} CH_3 \\ | \\ C \\ | \\ CH_3 \end{array} \right]^+ + n\, CH_2{=}\begin{array}{c} CH_3 \\ | \\ C \\ | \\ CH_3 \end{array}$$

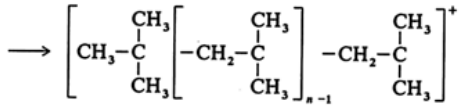

Chain termination:

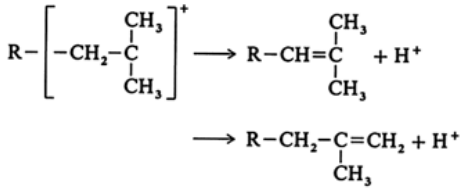

The initiator is a Lewis acid (BF_3, $AlCl_3$) in combination with a proton source, which is usually present as a trace impurity, e.g., water, hydrogen, fluoride, or hydrogen chloride.

The competition between the growth reaction and chain termination determines the chain length (i.e., the molecular mass). Since the activation energy of chain termination is higher, increasing the polymerization temperature lowers the polymer chain length (ln M ~1/T).

Meanwhile [286], a series of studies have appeared which have revealed a considerably refined view of the course of the reaction; yet a comprehensive mechanism which is consistent with all experimental findings is still lacking. Other studies deal with new initiators (aluminum alkyls; inorganic initiators such as acid clays and molecular sieves; Ziegler catalysts; high-energy radiation as an initiator). However, none of these new catalyst systems are used industrially. Industrial processes for polyisobutylene production still use the conventional Friedel – Crafts catalysts, boron trifluoride and aluminum chloride.

4.2. Production

4.2.1. *Isobutene as Raw Material*

The production of isobutene is described in ⟶ Butenes – Resources and Raw Materials, ⟶ Butenes – Upgrading of Butenes.

The purity requirements for the isobutene differ according to the type of polyisobutylene produced.

For the production of low molecular mass polyisobutylenes, butadiene-free C_4 fractions can be used, which apart from isobutene contain *n*-butane, isobutane, and *n*-butenes; stricter requirements are set for the isobutene for high molecular mass and solid polyisobutylenes.

During isolation of isobutene from C_4 fractions, an isobutene must be obtained which is as free from *n*-butene as possible. Primarily the content of *cis*- and *trans*-2-butene must be minimized. Both are powerful polymerization poisons and chain terminators. On the other hand, 1-butene can be used as a regulator in the production of high molecular mass and solid polyisobutylenes.

4.2.2. *Low Molecular Mass Polyisobutylenes*

Figure 50 shows the layout of the Cosden Petroleum continuous process, which can be regarded as a kind of fluidized-bed process.

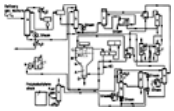

Figure 50. Cosden Petroleum process for producing low molecular mass polyisobutylene

a) Preliminary distillation; b) Absorption column; c) Caustic soda solution wash tower; d) Heat exchanger; e) Water separator; f) Silica gel towers; g) Heat exchanger; h) Reaction vessel; i) Mixing tank; j) Refrigeration unit; k) Gear pump; l) Pump; m) Offtake pipe; n) Cooler; o) Stripping column; p) Clay-filled tower; q) Distillation column; r) Water cooler; s) Pressure column; (For A to D, see text)

A liquefied refinery-gas mixture with C_1 – C_5 hydrocarbons and an isobutene content of 3 – 50 % is freed from pentenes in a preliminary distillation, enriched in isobutene in an absorption column, washed in a wash tower with ca. 20 % caustic soda solution to remove sulfur, cooled in a heat exchanger, dewatered in a water separator, dried in silica gel towers and, after further low-temperature cooling in a second heat exchanger, introduced into the lower part of the reaction vessel. The polymerization takes place here at temperatures between – 43 and +16 °C and pressures of 0.1 – 0.35 MPa, so that the liquid phase is maintained. The catalyst used is a mushy suspension of finely divided (50 – 100 mesh) aluminum chloride in dry, liquid polyisobutylene, which is prepared in a mixing tank and supplied to the reaction vessel by a gear pump.

The quantity of aluminum chloride is about 10 – 20 % of the total hydrocarbons in the reaction mixture. The activity of the aluminum chloride is increased by adding 0.08 – 0.12 % hydrogen chloride (relative to aluminum chloride) as activator and accelerator. Water or chloroform have the same effect. The diameter of the reaction zone (A) and the feed rate of the catalyst slurry are so matched to each other that the aluminum chloride particles do not settle. Zone (B) of the reaction vessel, with a diameter two to six times that of zone (A), reduces the flow rate by a factor of about three to four, and functions as a calming zone. It merges into the settling zone (C), in which the aluminum chloride particles settle and are pumped off as a suspension through the offtake pipe which extends into zone (B). This catalyst slurry returns via a cooler into the reaction zone (A). Spent, aggregated catalyst is discharged at the bottom of the reaction zone.

A variant of the process uses separate vessels for polymerization and for catalyst settling.

From the settling zone flows a clear, light solution of liquid polyisobutylene in unreacted starting material. To improve the yield, the reaction mixture is pumped at least four times, but usually more than eight times, through the reaction zone (A). The clear polymer solution flows through the heat exchangers (g) and (d) to cool the isobutene and, after a water wash to remove acidic constituents, into a stripping column in which about one to two thirds of the inert, unreacted starting material is distilled off. After post-treatment of the concentrated polymer solution in a clay-filled tower, which may precede the stripping column, all volatile constituents of the starting material, especially unpolymerized isobutene, are vaporized in a distillation column and returned via a water cooler into the absorption column. In a further column, the last traces of volatile constituents are distilled from the liquid polyisobutylene at atmospheric pressure.

Aside from this process of the Cosden Petroleum Corp., others have been developed. For a description of the BASF and Exxon processes, see [287, pp. 94 – 97], [288].

4.2.3. *Medium Molecular Mass Polyisobutylenes*

The basic principle of the Exxon process for producing medium molecular mass polyisobutylenes (Fig. 51) [289] resembles the process for low molecular mass polyisobutylenes. However, a purer isobutene is used as starting material and pentane or hexane is added as solvent. The polymerization temperature must be lower to obtain a higher molecular mass. A mixture prepared from fresh isobutene, recycled isobutene, and hexane with 30 % isobutene and 70 % hexane is brought via a heat exchanger and a cooler to the polymerization temperature of – 40 °C and fed into the reaction vessel. Simultaneously, a slurry of 5 % finely divided aluminum chloride in hexane at a temperature of below – 23 °C flows into the reaction vessel.

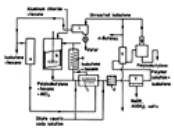

Figure 51. Exxon process for producing medium molecular mass polyisobutylenes

a) Mixing vessel; b) Reaction vessel; c) Heat exchanger; d) Nozzle mixer; e) Cooler; f) Settling tank; g) Degasser; h) Degassing vessel; i) Water separator

The polymerization takes place with powerful stirring so that the aluminum chloride does not settle, and external cooling. The throughput is 378.5 L/h isobutene – hexane mixture and 0.454 kg/h $AlCl_3$. The resulting solution of polyisobutylene in hexane with suspended aluminum chloride is drawn off, and after it has passed through a heat exchanger an excess of dilute caustic soda solution is added to destroy the catalyst. After thorough mixing in a nozzle mixer, the mixture of polymer solution and dilute caustic soda solution runs into a settling tank, where it forms two layers. The lower layer, consisting of dilute caustic soda solution, aluminum hydroxide, and salts, is discarded. The upper layer, consisting of a solution of polymer in hexane and unreacted starting material, is fed into a degassing vessel, in which the lower boiling

components, principally unreacted isobutene, are stripped off at 99 °C and 0.35 bar. *n*-Butenes if present are removed from the system to avoid poisoning the reaction. In a second degasser, the last traces of volatile constituents are removed. The unreacted isobutene is condensed, dewatered in a water separator, and added as recycle isobutene to the starting mixture.

In its batch process for producing medium molecular mass polyisobutylenes, BASF makes use of the chain-terminating effect of diisobutene [287, pp. 88 – 91], [290].

In a conical reaction vessel with good thermal insulation, pure, dry liquid isobutene and pure, dry liquid ethylene as internal coolant are mixed in the ratio 1 : 2. The quantity of diisobutene added to the isobutene depends on the desired molecular mass of the polyisobutylene. A sufficient quantity of gaseous boron trifluoride for the polymerization is then carefully introduced into the mixture. The mixture starts to boil, and the evaporating ethylene removes the heat of polymerization. The polyisobutylene remains behind in the reaction vessel in the form of a white foam. It is removed from the vessel and freed from all volatile constituents in a degassing machine.

4.2.4. *High Molecular Mass Polyisbutylenes*

BASF and Exxon have developed continuous processes for producing high molecular mass solid polyisobutylenes [287, pp. 97 – 104].

BASF Belt Process. In the belt process, the conical polymerization vessel of the BASF batch process described above is replaced by a continuous polymerization belt (Fig. 52) [291].

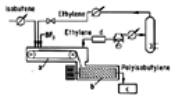

Figure 52. BASF belt process for high molecular mass polyisobutylenes

a) Steel belt; b) Twin-screw machine; c) Cooling and compounding; d) Purification

A gently inclined trough-shaped continuous steel belt, 16 – 18 m long and 50 cm wide, runs in a gastight housing over a drive roller and a tension roller. A mixture of pure, dry liquid ethylene and isobutene in the ratio 1 : 1 flows through a pipe onto one end of the steel belt. From a second pipe a solution of boron trifluoride in the same quantity of ethylene as is mixed with the isobutene is added. The quantities of cocatalyst and chain-length regulator [287, pp. 82 – 91]depend on the desired molecular mass. The polymerization starts instantly. It continues at the boiling point of the liquid ethylene (–103.7 °C) and in a short time is ended. The heat of polymerization is absorbed by the evaporation of ethylene. The vaporized ethylene is recycled via a purification and liquefaction system. The polyisobutylene is stripped from the steel belt with a scraper and degassed and homogenized in a twin-screw machine.

Exxon Process. Polyisobutylene is also produced by the process used for butyl rubber (see ⟶ Rubber, 3. Synthetic) [292].

4.3. Structure

The polymer obtained by cationic polymerization has the following structure:

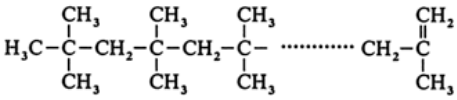

In the stretched state polyisobutylene gives a well-developed X-ray diffraction pattern which shows that 8 monomers with their 16 methyl groups are arranged in an almost regular helix around the principal axis, forming one translation period of the chain after 3 rotations [293].

It has been shown by IR spectroscopy that the principal end group is the methylene group [294].

4.4. Properties

The rubber-like polyisobutylenes, like all other polyisobutylenes, owe their physical and chemical properties to their paraffinic character. This results in a low glass transition temperature, low specific thermal conductivity, very low water vapor permeability, high dielectric strength, high resistivity, low dielectric constant, low dielectric loss factor, high ageing resistance, and rot resistance as well as an unusually wide-ranging chemical resistance. The good solubility in hydrocarbons and chlorohydrocarbons and the insolubility in alcohols, esters, and ketones also result from the paraffinic nature of the polyisobutylenes.

At room temperature, polyisobutylenes are resistant to dilute and concentrated hydrochloric acid, sulfuric acid, phosphoric acid, chlorosulfonic acid, phenolsulfonic acid, naphthalenesulfonic acid, formic acid, acetic acid, ammonia, potassium hydroxide solution, caustic soda solution, as well as aqueous calcium hydroxide, aqueous hydrogensulfite, copper sulfate solution, potassium permanganate solution, hydrogen peroxide, chromic acid, and dichromate solution. Above 80 °C, concentrated sulfuric acid causes carbonization and concentrated nitric acid causes decomposition.

Some properties of polyisobutylene that are independent of the degree of polymerization are as follows:

Density, g/cm^3	0.92
Volume coefficient of thermal expansion at 20 °C, K^{-1}	6.3×10^{-4}
Specific heat, kJ kg^{-1} K^{-1}	2.0
Glass transition temperature (DSC), °C	– 60
Thermal conductivity, W K^{-1} m^{-1}	0.19
Refractive index n_{20}^{D}	1.51
Dielectric constant ϵ_r (50 Hz, 20 °C) DIN 53 483	2.2
Dielectric loss factor tan δ (50 Hz, 20 °C) DIN 53 483	$\leq 5\times10^{-4}$
Water vapor permeability, g m^{-1} h^{-1} mbar^{-1}	2.5×10^{-7}

Further properties are listed in Tables 24 and 25.

Table 24. Properties of polyisobutylenes dependent on degree of polymerization, exemplified by Oppanol B (BASF)

Oppanol	Viscosity, Pa · s		Intrinsic viscosity	Number-average molecular
	at 20 °C	at 100 °C	[η], cm^3/g	mass $\bar{M}_n$*
Low molecular mass				
B 3	25	0.2 ± 0.02		820
Medium molecular mass				
B 10	5.0×10^4	2.2×10^2	27.5 – 31.2	24 000
B 12			34.0 – 39.1	30 000
B 15	5.0×10^5	3.0×10^3	45.9 – 51.6	40 000

High molecular mass				
B 50	1.5×10^{8}	8.0×10^{5}	113 – 143	120 000
B 80			178 – 236	180 000
B 100	3.6×10^{8}	6.7×10^{7}	241 – 294	250 000
B 120			295 – 361	300 000
B 150			416 – 479	425 000
B 200	1.5×10^{11}	1.0×10^{9}	551 – 661	600 000
Ultra high molecular mass				
B 246			> 770	> 760 000

* $\overline{M_n} = 0.94\sqrt{\frac{[\eta]10^3}{2.27}}$

Table 25. Behavior of Oppanol B100 and B200 towards solvents (weight increase by swelling in wt % at room temperature)

Storage time	**3½ days**		**30 days**	
Oppanol type	**100**	**200**	**100**	**200**
Acetone	1.7	1.5	3.9	3.6
Alcohol	0.1	< 0.1	0.35	0.2
Benzene	400	400	500	500
Butyl acetate	34	24	34	37
Carbon tetrachloride			soluble	
Chlorobenzene			soluble	
Chloroform			soluble	
Cyclohexane			soluble	
Cyclohexanone	9	7.4	24	18
Dimethylformamide	0.6	0.6	2.1	1.7
Ether	200	100		
Ethyl acetate	18	12	21	20
Ethylene glycol	0.6	1.4	0.9	2.0
Gasoline			soluble	
Glycerol	0.6	0.6	1.0	1.6
Hydrogen sulfide			soluble	
Isobutanol	1	0.3	2.4	0.7
Isooctane			soluble	
Methanol	< 0.1	< 0.1	0.1	0.1
Methyl acetate	8	5	7.8	10
Methylene chloride	200	200	200	200
Mineral oil			soluble	
Nitrobenzene	3	2	5.4	4.5
Olive oil	4	1.5	4.2	3.1
Paraffin			soluble	
Paraffin oil			soluble	
Tetrahydrofuran			soluble	
Toluene			soluble	
Water	< 0.1	< 0.1	0.1	0.1
Xylene			soluble	

4.5. Uses and Processing

The low molecular mass polyisobutylenes are used as plasticizing components for modifying sealants. In addition, they are used for producing insulating oils.

After further functionalization with polar groups, the low molecular mass polyisobutylenes are used as dispersants in fuel and lubricant additives.

The medium molecular mass polyisobutylenes are chiefly used as raw materials for adhesives and sealants. Due to their very low water-vapor permeability, permanently flexible sealants based on polyisobutylene – carbon black are important in the production of sealed double glazing units. Combinations of polyisobutylene with paraffin or wax are used for impregnating paper and board.

The breaking point of bitumen is lowered and its flexibility improved by addition of medium molecular mass polyisobutylene.

High molecular mass polyisobutylenes are used as blends with medium molecular mass grades for producing self-adhesive compounds, e.g., for adhesive plasters.

Because of their good tolerance by the skin, filled polyisobutylenes are used to produce sealing rings for colostomy bags.

Unstabilized polyisobutylenes are used in chewing gum bases.

Blending polyisobutylenes with paraffin oil and inorganic fillers gives permanently elastic sealants, which are usually marketed in extruded form as strip or round-profile string.

High molecular mass polyisobutylenes with high filler contents are used to produce sheets that are used for sealing buildings against groundwater and seeping water as well as for protection against corrosion and radiation. Conductive and magnetic polyisobutylene sheetings are also available.

The cold impact resistance of polypropylene is improved by blending with polyisobutylene. The sensitivity of polyethylene to environmental stress cracking is reduced by adding polyisobutylene.

Polyisobutylenes are compatible with natural and synthetic rubber and with reclaimed rubber.

Solutions of ultra high molecular mass polyisobutylenes ($\overline{M}_n$ > 760 000) in organic solvents even in low concentrations (<< 1 %), have pronounced viscoelastic properties and distinctly increased extensional viscosity.

Polyisobutylene can be compounded on a roll mill or in an internal mixer. Compounding is carried out at 140 – 170 °C to minimize thermal degradation of the polyisobutylene.

4.6. Commercial Products

BASF, Germany:	Oppanol B 3 (liquid); Oppanol B 10, B 12, B 15 (flexible resin-like, tacky); Oppanol B 50, B 80, B 100, B 120, B 150, B 200, B 246 (solid, rubber-like)
Exxon, United States:	Vistanex LM-MS, LM-MH (flexible resin-like, tacky); Vistanex MM L-80, MM L-100, MM L-120, MM L-140 (solid, rubber-like)

In addition, a range of low molecular mass polyisobutylenes are marketed.

4.7. Copolymers [295]

Isobutene – isoprene – divinylbenzene terpolymers are produced under conditions similar to those used for butyl rubber, with the addition of divinylbenzene. Unlike butyl rubber, they exhibit no cold flow.

Isobutene – styrene copolymers behave similarly to polyisobutylenes or butyl rubber.

The copolymers have no economic significance.

[Top of Page]

5. Poly(4-Methyl-1-Pentene)

T. Geoffrey Heggs

As described for polypropylene in Chapter Polypropylene, Ziegler – Natta stereospecific catalysts discovered in 1954/1955 polymerize α-olefins to crystalline products. Foremost of these are polypropylene and, to a lesser extent, poly(1-butene), but higher α-olefins also yield crystalline products having helical chains in the crystallites [296-298]. Much laboratory work on C_5 – C_{18} olefins centers on polymerization mechanisms and the interesting crystallization behavior of homopolymers and copolymers, but only poly(4-methyl-1-pentene), PMP, based resins have reached commercial production status. Their excellent optical clarity combined with a high melting point, reasonable toughness, and general inertness has led to use in demanding applications where the much higher raw material cost is a secondary consideration. Operating temperatures for PMP in service can be as high as 200 °C for short times, after which thermal oxidation is likely to start. For this reason and because of impact strength limitations, the potential advan-tages of even higher melting poly(α-olefins) cannot be realized in practice. Engineering resins such as polyimides, polysulfones, and polysulfides must be used for sustained high-temperature operations and extra toughness.

4-Methyl-1-pentene monomer can be polymerized with cationic and Ziegler – Natta catalysts. The former cause some isomerization and yield irregular chains, giving amorphous viscous liquids or rubber-like solid products of little commercial interest [299], [300].

In 1965 ICI made several grades of PMP, under the trade name TPX, in its semitechnical plant at Welwyn Garden City for distribution and in-depth evaluation [301]. Manufacture on its Wilton production site started in June 1968 on a 2000 t/a plant producing transparent and opaque grades. In 1973 the process was sold to Mitsui Petrochemical Industries, which became the sole manufacturer on its plant at Iwakuni-Ohtake works in Japan. Capacity was increased from 2000 t/a to 3000 t/a in 1982, and then to 6000 t/a in 1983. In 1989, Phillips 66 in Bartlesville, United States, started supplying development quantities from a 250 t/a pilot plant in the United States under the trade name Crystalor.

5.1. Raw Materials

5.1.1. *Monomers*

Several monomers are used in TPX production to overcome the deficiencies of haze and brittleness in pure homopolymers of 4-methyl-1-pentene. The restricted availability of these monomers in commercial quantities in the early 1960s proved troublesome for ICI's development program.

Polar and highly unsaturated contaminants retard polymerizations with Ziegler – Natta catalysts. Handling precautions must recognize flammability hazards and facile peroxidation of these liquid monomers on exposure to air. At high concentrations they exert a narcotic effect.

4-Methyl-1-pentene (4MP) is prepared by selectively dimerizing propene at ca. 150 °C and 10 MPa with supported alkali metals, particularly potassium. British Petroleum first established a commercial process in 1967, followed by further patents to Chevron and Idemitsu (Table 26). All these processes yield a liquid fraction containing ca. 80 % of 4MP. Separation of pure 4MP from the C_6 fraction requires superfractionation with a 100-plate column.

Table 26. Dimerization of propene to 4-methyl-1-pentene [298]

Company	Process	Catalyst	Temperature, °C	Pressure, MPa	Yield *, % 4MP	Patent priority
British Petroleum [302]	rocking bed	K on graphite	160	11	85	12.12.62
British Petroleum [303]	fixed bed	Na on K_2CO_3	160	12	75	8.8.63
Chevron [304]	slurry	K on $MgSiO_3$	150	5	80	17.5.66
Idemitsu [305]	slurry	K/Cu/*tert*-amine	180	7	90	13.4.76

* In liquid fraction.

Apart from being the main feedstock for PMP manufacture, 4MP is now used as comonomer in certain grades of linear low-density polyethylene. However, 4MP is not traded commercially, and there is no quoted price. The current name plate capacity of Mitsui Petrochemical Industries plant for 4MP is 13 000 t/a.

A typical analysis (wt %) of 4MP monomer used by ICI to make TPX is as follows:

4-Methyl-1-pentene	97.4
Lighter than C_6	0.1
cis-4-Methyl-2-pentene	0.9
trans-4-Methyl-2-pentene	1.5
1-Hexene	0.1

Some physical properties of 4MP [306], [307], relevant to its polymerization are:

bp, °C	53.86
mp, °C	– 153.64
Density at 20 °C, g/cm^3	0.6637
Vapor pressure at 20 °C, kPa	29.47
Vapor pressure at 50 °C, kPa	89.34
Refractive index n_{20}^{D}	1.38267
Heat capacity of liquid at 25 °C, J g^{-1} K^{-1}	2.191
Heat of vaporization at *bp*, J/g	0.3221
Flash point, °C	– 7.0
Autoignition temperature, °C	300

Linear C_6 – C_{12} α-Olefins. Most commercial grades of PMP contain small amounts of higher α-olefins as comonomer to modify the main chain. Usually, the even carbon

number olefins are made by oligomerizing ethylene with triethylaluminum or soluble nickel complexes. Commercial quantities of these pure olefins up to C_{18} are priced at about £1/kg. Table 27 lists some of their physical properties, together with those of useful branched comonomers.

Table 27. Physical properties of some comonomers * [306]

Comonomer	*bp*, °C	*mp*, °C	Density at 20 °C, g/cm^3	*mp* of isotactic homopolymer, °C
1-Pentene	29.96	– 165.22	0.6405	75 – 80
3-Methyl-1-butene	20.05	– 168.49	0.6272	310
1-Hexene	63.48	– 139.83	0.6732	amorphous at 20 °C
3-Methyl-1-pentene	54.17	– 153.0	0.6674	> 350
4-Methyl-1-pentene	53.86	– 153.64	0.6637	235 – 250
4,4-Dimethyl-1-pentene	72		0.6827	> 380
1-Octene	121.29	– 101.72	0.7149	34
1-Decene	170.60	– 66.28	0.7408	45
1-Dodecene	213.30	– 33.60	0.7583	60

* Data for 4-methyl-1-pentene included for comparison.

5.1.2. *High-Melting Nucleating Polymers*

It is reported [308], [309] that < 1 % of well dispersed high-melting polymers from branched α-olefins nucleate PMP crystallization to enhance transparency. Isotactic polymers of 3-methyl-1-butene, 3-methyl-1-pentene, and 4,4-dimethyl-1-pentene are preferred. These branched monomers are much more expensive than linear α-olefins because they arise only as minor byproducts from refining and cracking operations, or by special synthesis. For example, 3-methyl-1-butene can be prepared by dehydrating isoamyl alcohol. The same isospecific catalysts used to polymerize 4MP are also suitable for making these branched olefin polymers (see Section Industrial Manufacture). Preferably these high-melting polymers are prepared as a stage of the PMP polymerization process, rather than as separate additives.

5.1.3. *Catalysts*

Most Ziegler – Natta catalysts suitable for polypropylene polymerize and copolymerize 4MP to give linear head-to-tail polymers. The reactivity of the higher α-olefins is appreciably lower than that of propene under comparable conditions. Copolymerization data compiled by KISSIN [310], [311] illustrate how branching and chain length impede reaction with titanium-based catalysts (Table 28). Steric hindrance by bulky groups in the 3-position, adjacent to the double bond, is particularly pronounced (3-methyl-1-butene and 3-methyl-1-pentene). 4-Methyl-1-pentene is several times less active than propene, as are all the higher linear α-olefins. Low overall reaction rates arise when some of the active sites on certain catalysts are inaccessible to bulky monomers [311].

Table 28. Relative copolymerization activities r_2 of higher α-olefins relative to propene [310]

Monomer	r_2 *
Ethylene	8 – 20
Propene	1
1-Pentene	0.2 – 0.5
1-Hexene	0.2 – 0.4
1-Decene	0.1 – 0.3
1-Octadecene	0.1 – 0.15
3-Methyl-1-butene	0.02 – 0.06
3-Methyl-1-pentene	0.05
4-Methyl-1-pentene	0.15

* Ratio of propagation rates, $r_2 = k_2^2 / k_2^2$.

It is necessary to use highly isospecific catalysts to achieve good yields of crystalline polymer without much atactic byproduct. Preferred systems use combinations of $TiCl_3$ with $Al(C_2H_5)_2Cl$ as activator [312], or possibly the newer supported catalyst systems such as $TiCl_4$ – $MgCl_2$ – $Al(C_2H_5)_3$ with an appropriate Lewis base [313]. Catalysts made directly from $TiCl_4$ and organoaluminum compounds only yield 60 % of crystalline polymer, and even violet $TiCl_3$ activated with aluminum trialkyls produces some 20 % of solubles. These trialkylaluminum – titanium halide systems also cause some isomerization of 4MP, but this is completely avoided with $TiCl_3$ – $AlEt_2Cl$. The original ICI process used a coarse, preformed spherical $TiCl_3$ made by controlled reduction of $TiCl_4$ with alkyl aluminum chlorides. Experiments showed that it was important to secure a very mobile slurry to permit thorough catalyst extraction for the highest polymer transparency.

5.2. Industrial Manufacture

Published information on the production of TPX is very limited apart from the disclosures of ICI. Details which follow relate to their early slurry process, but some improvements have been introduced by Mitsui Petrochemical Industries as their manufacturing capacity expanded from 2000 to 6000 t/a.

Semibatch polymerizations used a conventional water-jacketed stirred reactor containing inert aliphatic hydrocarbon diluent, $AlEt_2Cl$, and particle-form purple $TiCl_3$ catalyst. A nucleating polymer of a high-melting polyolefin can be formed on this catalyst either prior to its introduction into the reactor, or at the end of the main polymerization stage [301]. Overall reaction cycles are shortened when the catalyst is first subjected to prepolymerization as a concentrated slurry in a separate reactor to generate the high-melting polymer. This procedure yields a supply of treated catalyst ready for use in the main reactor without delays caused by the low-activity monomer. Sometimes this simple technique yields prepolymer which is poorly dispersed. The problem is reportedly overcome [309] by first polymerizing a small amount of 4MP on the catalyst before adding monomer to generate nucleating polymer. An even simpler procedure apparently works: If a mixture of 4MP and the preferred monomer for nucleation is used for prepolymerization, 4MP is consumed first because of its higher reactivity, leaving the virtually pure second monomer free to polymerize on the expanded catalyst particles. Less than 1 % of this additive is needed.

Following the addition of (prepolymerized) catalyst to the main reactor, an initial charge of 4MP and comonomer establishes both the required composition and the polymerization rate, which is first order with respect to catalyst and monomer. As polymerization proceeds at ca. 60 °C, continuous injection of monomers maintains the reaction conditions until the prescribed amount of polymer has formed. This reaction lasts a few hours, during which small injections of hydrogen control molecular mass, as in polypropylene manufacture. Termination of monomer feeds, followed by a short run-down period ensures that most of the monomer is consumed before deashing starts.

All catalyst components are solubilized by adding a C_3 – C_6 alcohol to the slurry in a second heated vessel. Centrifuging, followed by several stages of washing with alcohol – diluent mixtures extracts catalyst residues and soluble polymer. Importantly, addition of water, used in some polypropylene deashing processes, results in lower transparency of molded articles. Ash levels below 0.01 %, and < 10 ppm of Al and Ti are readily achieved. All filtrates are recycled after separation by fractional distillation.

Conventional hot-nitrogen driers remove diluent from the centrifuged cake. After addition of stabilizers, the dry powder is densified at 270 – 300 °C in an extruder with a length/diameter ratio of about 30/1. Various grades differ in comonomer content, molecular mass, and additive formulation (e.g., reinforcement with glass fibers, mica, and mineral fillers).

A Mitsui patent [314] describes techniques for making broad MMD (co)polymers from 4MP using $MgCl_2$ supported catalysts in a two-stage polymerization system. This can be either conventional diluent-based slurry for each stage, or slurry for the first stage followed by a fluidized-bed reactor operating at 100 °C and 0.2 MPa absolute pressure.

In 1989, Phillips Petroleum commissioned a 250 t/a integrated pilot plant at Bartlesville, United States, making 4MP monomer to convert into PMP for evaluation and sale. This plant operates a slurry process using a conventional semibatch stirred reactor.

Patents to BASF [315], [316] describe bulk polymerization of 4MP at 90 – 150 °C under inert gas pressures up to 2 MPa to increase the yield. Catalysts used were $TiCl_3 \cdot AlCl_3$/tributyl phosphine/$AlEt_2Cl$. Hot-heptane solubles of ca. 30 % were, however, high, and no process has yet been announced.

Institut Français du Petrole's intriguing isomerization – polymerization of potentially cheap 4-methyl-2-pentene to PMP, using $TiCl_3$, nickel acetylacetonate, and $AlEt_2Cl$, has not been commercialized [317], [318].

5.3. Polymer Structure

5.3.1. *Homopolymer*

PMP polymer, made with highly isospecific catalysts, has the following regular head-to-tail structure in which all isobutyl groups lie on one side of the chain in a two dimensional Fischer projection.

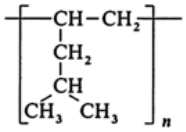

NMR spectroscopy, so useful in polypropylene and ethylene – propene copolymer characterization, is of limited use for higher poly (α-olefins) where overlapping resonances of the main- chain atoms and pendant groups severely impair resolution and interpretation [319]. Deductions concerning polymer stereoregularity largely rely on crystallinity related phenomena, such as X-ray and IR spectra, melting behavior, solvent fractionation, and various physical properties.

Stereoregular poly(α-olefin) chains pack together to form high-melting crystallites, which then assemble into a superstructure of radially symmetric spherulites during crystallization from a melt phase. In unbranched polyethylene, the chains adopt a zig-zag arrangement in the crystal. The methyl side group in polypropylene forces the chain to adopt a helical configuration involving three monomer units per turn (3/1 helix). With still larger branches the helix unwinds to 4/1. In the case of PMP, the bulky isobutyl group is accommodated in a 7/2 helix with 3.5 monomer units in each turn, and four chains per unit cell of 28 monomer units. Unit cell parameters are $a = b = 1.86$ nm and c (chain axis) = 1.38 nm [320]. This gives an exceptionally low calculated crystal density of 0.822 ± 0.006 g/cm^3, largely because of the loose space packing of helices. At room temperature, the amorphous phase in this semicrystalline polymer has the higher density of 0.838 g/cm^3 [320], [321]. In a typical molding, the average polymer density of PMP is ca. 0.83 g/cm^3. Such a unique reversal of crystalline and amorphous polymer densities has some interesting implications for molding operations. At 60 °C, a common mold temperature, the densities of the two phases are judged to be equal, i.e., there would be no volume change for crystallization at this temperature.

Crystalline PMP finally melts at about 250 ± 5 °C (slow heating) and has a glass transition temperature T_g of 50 – 60 °C at 100 Hz [312] (18 – 30 °C by dilatometry [321]). Other transitions have been reported [322]: 130 °C (start of crystal disordering) and at – 120 °C to – 150 °C (relaxation of short chain segments). Crystallization from the melt does not require massive supercooling. Isothermal crystallization half time at 220 °C is 0.5 min, increasing to 480 min at 236 °C. As with polypropylene, PMP is properly viewed as a semicrystalline polymer, achieving a crystallinity of ca. 75 % in fully annealed samples, 40 % in quenched moldings, and 55 – 65 % in normally fabricated articles [323]. The only stable crystallite formed from polymer melts is tetragonal, but up to five polymorphs can be obtained by crystallization from solution [324].

5.3.2. *Copolymers*

Random copolymers of 4MP, with ca. 2 – 8 wt % of C_6 – C_{12} linear α-olefins, are important commercially because of their good transparency and robust processing behavior.

All linear α-olefins copolymerize with 4MP to some degree. Products are complex, with a single copolymer preparation yielding chains with a wide range of compositions and degree of block structure. The more active comonomers, such as ethylene and propene, produce a dispersed phase of slightly modified PE or PP, even when only 5 mol % of the comonomer is fed [325]. The amount of this phase increases as more comonomer is used, thereby further reducing transparency through light scattering. 1-Hexene and a few other comonomers behave differently in that they are able to substitute for 4MP units in the crystal without significantly distorting or altering the crystal structure of PMP; nor do they generate a second dispersed phase [326]. In these cases, the physical properties still resemble those of PMP at low comonomer contents. Small decreases in crystallinity and melting point are measurable but they are less severe than with non- cocrystallizing monomers. 1-Pentene copolymers also readily cocrystallize, causing some slight lattice contraction. Only small proportions of 1-butene can be accommodated within the lattice.

Increasing the monomer chain length to octene, decene, and above makes cocrystallization increasingly difficult. However, small amounts of 1-octene and 1-decene are accommodated within the lattice by expansion, but octadecene is far too large. The latter monomer severely reduces crystallinity and forces the copolymer into amorphous regions. The magnitude of these crystal changes is detailed in [326]. α-Axis spacing in cocrystallized copolymers can expand from 1.86 to 1.91 nm with octene, and shrink to 1.81 nm with 1-pentene copolymers.

Glass transition temperature is favorably reduced by 15 – 20 °C in 5 mol % 1-decene copolymers. This in itself is only sufficient to give a small toughness improvement because of the high T_g of PMP homopolymer. Copolymerization usually lowers the modulus.

5.4. Properties of PMP Products

5.4.1. *Optical Properties*

Crystalline PMP polymers exhibit unusually high optical transparency, both in terms of clarity (narrow-angle scatter) and haze (high-angle scatter). Optical transmittancies of commercial resins are in the region of 90 – 95 %, compared with 88 – 92 % for polystyrene and 90 – 92 % for acrylics. Other, more familiar crystalline polyolefins, such as polyethylene and polypropylene, are at best translucent in thin sections. In these cases light is scattered from spherulitic assemblies of crystalline and amorphous regions that have different optical refractivities. Additionally, the crystallites themselves are anisotropic.

With spherulitic PMP, high transparency is favored by the similar densities of the crystalline and amorphous regions, and by the low optical anisotropy of helical chains in the spherulites. Providing polymers have been thoroughly deashed, homopolymer optical transmittances of 80 – 85 % can be realized in quenched, 3 mm thick moldings, and > 90 % in 1.5 mm thick samples. Plaques cooled more slowly develop considerable haze and lower transparency (20 %) as a result of increased interspherulitic voiding. These voids, about 1 µm in diameter, cluster round the boundaries of spherulites in the form of “shells” which are responsible for haze [327]. Quenched moldings have fewer microvoids because crystallization takes place at lower temperatures. This lowers the volume reduction on crystallization as the densities of amorphous and crystalline phases become more similar. At this lower temperature, the stronger melt is more able to resist shrinkage stresses without forming voids.

Copolymers of the type described in Section Copolymers have improved transparencies because of such reduced voiding. With these products the crystallization temperature is reduced by 10 – 30 °C, depending on composition, giving improved transparency in both quenched and slowly cooled moldings. With 11 wt % 1-octene copolymer, the transmittance increases to 91 % and 85 % in 3 mm thick quenched and slowly cooled moldings, respectively. This improvement with copolymers is useful, even with 5 wt % comonomer, because high transparency then becomes less dependent on maintaining very precise molding conditions. It is interesting to note that spherulite size is largely unaffected by copolymerization, or by quenching a molding [328], except in the thin surface skin.

The final, but major improvement to optical transparency comes from introducing properly dispersed nucleating agents into homo- or copolymers to reduce the size of spherulites from 50 µm to < 5 µm. In quenched, nucleated homopolymers optical transmission reaches 96 %, which only falls to about 88 % with slower cooling (Tables 29 and 30). Copolymers, already quite transparent when quenched, now achieve transmittancies around 96 % in slow- cooled moldings after nucleation (Table 30). The combination of nucleation, rapid cooling, and copolymerization virtually eliminates microvoiding in most moldings.

Table 29. Optical transmission of PMP homopolymer with poly(4,4-dimethyl-1-pentene) nucleant [309]

Spherulite size, µm			Transmission of
Max.	Mean	Min.	quenched moldings, %
100	50	< 1	69
15	10	< 1	86
< 1	< 1	< 1	96

Table 30. Copolymer nucleation [a] with poly(3-methyl-1-pentene) [309]

Comonomer	Optical transmission, %			
	Quenched moldings		Slow-cooled moldings	
	No nucleant	Nucleated	No nucleant	Nucleated
1-Pentene [b], 2 wt %	84	97	40	94
1-Hexene [b], 2 wt %	84	97	40	94
1-Octene [b], 2 wt %	84	97	40	94
1-Decene [b], 2 wt %	84	97	40	94
None	84	95	20	88
1-Hexene, 2 wt %	84	97	40	94
1-Hexene, 5 wt %	87	98	72	96

[a] Nucleated samples all had mean spherulite diameters < 2 µm, while unnucleated samples had mean spherulite diameters > 25 µm.
[b] Optical transmission similar for all four comonomers.

Excellent light transmission in the near UV region makes PMP superior to glass and other transparent plastics in this respect, and inferior only to quartz (Figure 53).

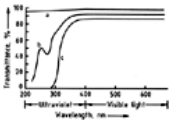

Figure 53. Light transmission [329]

a) Quartz; b) Poly(4-methyl-1-pentene); c) Glass

5.4.2. *Stability and Chemical Resistance*

The semicrystalline and saturated hydrocarbon nature of PMP confer excellent resistance to polar solvents and inorganic substances, and to strong aqueous solutions of most mineral acids, alkalies, and salts. Vulnerable tertiary C – H bonds make PMP unsuitable for use in highly oxidizing environments, such as chromic acid, concentrated nitric acid, and chlorine. It is permeable to some light hydrocarbons, aromatic liquids, and halogenated solvents, which cause swelling and some loss of strength. Repeated sterilization with superheated steam at 135 °C over 60 cycles causes only minor reductions in mechanical properties [330]. Accordingly, PMP is especially suited to medical uses. None of the alternative transparent polymers (polystyrene, acrylics, polycarbonate, and polyester) can offer PMP's broad range of resistance at room and elevated (60 °C) temperature. Resistance to detergent stress cracking, while better than other transparent polymers, is not quite up to the standard of the best bottle-blowing grades of polyethylene.

Above 100 °C, several aromatic and aliphatic solvents will effect complete dissolution. Stable gels containing 0.5 – 2 mm diameter spherulites form on rapidly quenching cyclohexane solutions containing 150 g/L of PMP [331]. Decalin at 135 °C is suitable for conventional molecular mass measurements by viscometry according to the Mark – Houwink equation [332] $[\eta] = 1.94\times10^{-4} M_v^{0.81}$.

Thermal Degradation. Processing temperatures of ca. 300 °C are necessary for the majority of PMP grades. Even under high vacuum these severe conditions promote degradation as evidenced by a marked decrease in molecular mass and gas evolution [333], principally isobutene (55 %) and propane (35 %). Carbon- carbon bond cleavage occurs in chain reactions in which hydrogen is abstracted from tertiary C–H bonds. The initial high rate seems to stem from a few weak links, possibly resulting from adventitious oxidation. Degradation subsequently continues at a lower rate. Suitable melt stabilizers incorporated by manufacturers severely retard much, but not all degradation.

Thermo-oxidation. Oxygen greatly accelerates the rate of chain scission over heat alone [334-336]. Precautions must be taken not only at high melt processing temperatures, but also in the solid state. Oxidation is accompanied by molecular mass reduction along with hydroxyl and carbonyl formation in the polymer, leading to yellow discoloration, brittleness, and odor at long exposures. Low molecular mass volatiles such as aldehydes, ketones, acids, water, and carbon dioxide form in chain reactions involving peroxidation of tertiary C–H bonds and radical-induced chain cleavage. The well-established faster oxidation of PMP compared to PP is attributed to the more rapid diffusion of oxygen in the open PMP structure. Studies indicate [335] that oxygen also attacks PMP crystallites, and not just the amorphous phase, as in PP.

Stabilizers derived from bulky hindered phenols and polyphenols, phosphites, and thioesters provide reasonable protection against deterioration during processing and in service at elevated temperature. The useful life with ca. 0.2 – 0.5 % of these additives increases from around one day at 200 °C to seven years at 100 °C [337]. Manufacturers advise on the most suitable polymer grade and further appropriate antioxidants for demanding applications.

Photo-oxidation. Photodegradation of PMP is basically similar to thermal oxidation, but UV light of wavelength < 400 nm seriously accelerates decomposition because chromophores, perhaps from impurities or oxygen charge-transfer complexes, assist oxidation by peroxide intermediates. Chain decomposition then yields an array of oxidized products and lowers the molecular mass [338-340]. These reactions can be minimized by incorporating into the polymer specific stabilizers to deactivate excited chromophores, scavenge radicals, and reduce peroxides. Hindered amine light stabilizers (HALS), which behave as UV-stable antioxidants, and metal chelates are quite effective [340]. However, the latter can cause mild blue-green discoloration.

Enhanced oxygen diffusion in PMP is cited as a principal reason for its greater susceptability to photodecomposition than PP. Applications involving extended outdoor weathering need careful appraisal, even with UV-stabilized grades, unless there is heavy protection by pigment screening with carbon black, ferric oxide, rutile, etc. Light reaching earth is screened of wavelengths < 290 nm by the atmosphere.

Suitably formulated PMP is also able to withstand up to 50 kGy of high energy β- and γ-radiation, used for sterilization, without discoloration or embrittlement [341]. Irradiation at higher doses (3 MGy) causes cross-linking and generation of volatiles [342], whereas chain scission is the dominant reaction with PP except at high doses.

5.4.3. *Mechanical Properties*

PMP has mechanical properties reasonably similar to those of PP at 0 – 100 °C. This applies to flexural modulus, yield strength, and hardness, though the precise comparison is dependent on the chosen grade (Table 31). Copolymers are more flexible with lower yield strengths and softening points. Melting and crystallization curves of two commercial copolymer grades, provided by Mitsui Petrochemical Industries (Figs. 54 and 54) illustrate some differences. Figure 54 shows DSC (differential scanning calorimetry) exotherms and endotherms for a lightly copolymerized PMP. While the final melting is at 242 °C, with a peak melting rate at 233 °C, the onset of melting is clearly visible at 200 °C. The second copolymer (Figure 55) is much less crystalline (melting peak areas 28.2 and 37.2 J/g respectively), with a slightly lower final melting point of 231 °C and a peak at 226 °C. This reduction in crystallinity is responsible for lowering modulus, heat deflection temperature, and surface hardness. A presumed lowering of the glass transition temperature accounts for the increased unnotched impact strength of MX004, although this is still inade-quate to raise notched impact strengths at room temperature.

Table 31. Mechanical properties of PMP [a]

Property	Method	Poly(4-methyl-1-pentene)					Polypropylene	
		Crystalor		TPX		Crystalor, TPX	Propathene	
		HBC-020 (homopolymer)	CBN-020 (copolymer)	RT18 (copolymer)	MX004 (copolymer)	30 % glass fiber filled	GWM 22 (homopolymer)	GWM 101 (copolymer)

Density (20 °C), g/cm^3	ASTM D 1505	0.835	0.835	0.833	0.834	1.04	0.905	0.905
Melt flow index (MFI) at 260 °C, 5 kg, g/10 min	ASTM D 1238	20	20	26	26		4.0 [b]	6.0 [b]
Melting point, °C	Manufacturers	240	235	240	240	230 – 240	170	165 – 170
Vicat softening point, °C	ASTM D 1525	188	175	173	160		154	147
Heat deflection temperature, °C	ASTM D 648							
0.46 MPa load		90	85	90	85	227	100	95
1.80 MPa load		54	49		40	189	65	60
Tensile yield strength, MPa	ASTM D 638 5 mm/min	32.4	27.6	23.0	19.6	59 – 67	34.5	27
Flexural modulus, GPa	ASTM D 790	1.92	1.24	1.27	0.74	5.5 – 5.9	1.5	1.15
Izod unnotched impact strength, kJ/m^2	ASTM D 256	12	19	15	49	22	60	tough
Charpy notched impact strength, kJ/m^2	ASTM D 256	5	5	5	4	8	4.5	9.5
Hardness (Rockwell)	ASTM D 785			85	60	107	95	90

[a] Taken from manufacturer's literature of Phillips 66, Mitsui Petrochemical Industries, and ICI.
[b] Measured at 230 °C, 2.16 kg.

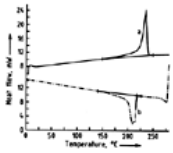

Figure 54. DSC traces for TPX, RT 8XB [343]

a) Melting endotherm; b) Crystallization exotherm

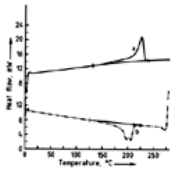

Figure 55. DSC traces for TPX, MX004 (food grade) [343]

a) Melting endotherm; b) Crystallization exotherm

Carefully molded PMP articles are less brittle than transparent polystyrene and acrylics, and they are nearly as tough as polypropylene homopolymer. The latter usually fails with some yielding in unnotched Izod tests at room temperature, thereby giving high failure energies. PMP's higher T_g causes failure in a brittle mode unless it is sufficiently modified by copolymerization. Above 150 °C, PMP has greater strength and form stability than PP and polycarbonate as a result of its high melting point. At 20 °C PMP creeps less in long-term tests than either polypropylene or high-density polyethylene.

Polymers filled with glass fiber and mineral additives are available from Mitsui, Phillips, and Schulman in Belgium. Composites with 30 % glass fiber are particularly useful due to their enhanced modulus and greater form stability at 200 – 250 °C. With 30 wt % glass fiber, room temperature modulus is triple that of PMP homopolymer, while the heat deflection temperature increases from 90 °C to 227 °C (Table 30).

5.4.4. *Electrical Properties*

The excellent electrical insulation properties of PMP (Table 30) are similar to those of PTFE and the best cable-grade polyethylene, whilst in the frequency range of 10 kHz to 20 MHz PMP is superior. Its permittivity (dielectric constant) is the smallest of all synthetic plastics, and it remains essentially constant over wide temperature and frequency ranges. A combination of low dielectric loss factor and good form stability at high temperature makes it very suitable for microwave oven applications. Due to its negligible moisture absorption, PMP preserves these electrical characteristics under adverse conditions.

Table 32. General properties of PMP products

	PMP	PP
Coefficient of linear expansion, K^{-1}	1.17×10^{-4}	1.1×10^{-4}
Thermal conductivity, $Wm^{-1}K^{-1}$	0.168	0.209
Specific heat (20 °C), $Jg^{-1}K^{-1}$	1.97	1.93
Heat of melting[a], J/g	143[b]	185
Burning rate, cm/min	2.54	2.3
Mold shrinkage, %	1.5–3.0	1–2
Refractive index, n_D^{20}	1.463–1.470	1.503
Volume resistivity, $\Omega \cdot cm$	$> 10^{16}$	$> 10^{16}$
Dielectric breakdown voltage, kV/mm	42–65	75
Permittivity at 100 Hz–1 MHz	2.1	2.25
Dielectric loss factor tan δ		
(20 °C) 50 Hz 10^{-6}	60	150
1 kHz 10^{-6}	35–140	200
1 MHz 10^{-6}	25–50	150

[a] For 100 % crystalline polymer.
[b] Mitsui value; for other measurements see [320] [323] [344].

5.4.5. *Surface Properties and Permeability*

PMP's surface tension of only 24 mN/m is the second lowest to PTFE (20 mN/m) [345]. This confers good stain resistance and excellent release properties towards other materials.

Confirmation of PMP's loose space packing is to be found in its high permeability towards gases (Table 30). Oxygen, nitrogen, and carbon dioxide diffuse some 25 times more rapidly through PMP film than through cast PP or HDPE films. Moisture vapor transmission is substantial, despite the low absorption of only 0.01 %, even after total immersion in water.

Table 33. Gas and water vapor permeability of 50 µm PMP film [346]

Permeability to	PMP	HDPE	PP
Oxygen, L m^{-2} d^{-1} atm^{-1}	40	1.1	1.6
Nitrogen, L m^{-2} d^{-1} atm^{-1}	9.5	0.3	0.36
Carbon dioxide, L m^{-2} d^{-1} atm^{-1}	120	4.6	6.5
Water vapor, g m^{-2} d^{-1} atm^{-1}	65	4.0	7.0

5.4.6. *Rheology*

PMP melts are strongly non-Newtonian. Their viscosity decreases rapidly at high shear rates and with increasing temperature (Fig. 56). These two effects produce particularly fluid melts under optimum processing conditions, and they call for some alteration to processing techniques compared to PP. In most cases friction heating alone in the extruder barrel is insufficient for uniform melting, and pressure generation also needs improving. Allowance should be made for low melt strength. There is a suggestion [323] of some melt order in view of the low entropy of fusion, 10.3 J K^{-1} mol^{-1}. The particularly rapid decline in viscosity with increasing temperature, around 280 °C, also has been linked with helical association in the melt [344], [347], [348].

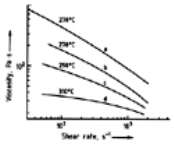

Figure 56. Viscosity curves [347]

a) Polypropylene (Propathene) MFI 4.0 (230 °C/2.16 kg); b) – d) Poly(4-methyl-1-pentene), TPX, grade DX810, MFI 70 (260 °C/5.0 kg)

5.5. Economic Aspects

The following list prices (in £/t) for polymerization grade materials and polymers prevailed in the fourth quarter of 1991:

Propene (Western Europe)	210
4-Methyl-1-pentene	Not traded
1-Hexene	1010
1-Octene	920
1-Decene	900
Polypropylene (Western Europe and United States)	360 – 450
PMP (grade and volume dependent)	2500 – 4300

The higher cost of PMP over PP reflects higher raw material prices without the economies of scale now realized for PP, as well as 10 years difference in experience.

5.6. Fabrication and Applications

5.6.1. *Fabrication*

In principle, most of the normal fabrication techniques used for polyolefins can be used to make PMP articles, provided due recognition is given to its high melting point and low melt viscosity. Further changes to processing techniques have been introduced over the past five years to improve output rates and quality.

Injection Molding. While both piston and screw injection molding machines can be used, the ideal type employs a slow- compression screw with an L/D ratio of 20 – 24. The cylinder must be capable of being heated to 350 °C, with precise temperature control over the preferred operating range of 280 – 320 °C. Friction heating is inadequate, and can cause local decomposition. An alternative is to feed the machine with granules preheated to 130 °C. A valved nozzle prevents die drool between shots with very fluid melts. Injection pressures only need to be 50 and 25 MPa for the primary and secondary stages, with a relatively slow injection speed (< 50 % of the machine capacity) into a mold at 20 – 60 °C. Gas venting of molds is imperative to avoid burning from any volatile decomposition products formed at high melt temperatures.

Extrusion Blow Molding. The attraction of making high-quality transparent containers by blow molding PMP is tempered by poor melt strength and additional heating requirements. Some simple adjustments to secure accurate melt temperature control in the range 220 – 250 °C, the use of a parison controller and accumulator, and pressure generation with a barrier type torpedo are satisfactory for certain applications. Still higher quality and better production rates call for a number of additional alterations. These include continuous parison extrusion, modification of the pinch-off blades, a long L/D extruder, and mold vents. Manufacturers advise on the best conditions for their resins.

Injection stretch blow molding is the ideal method for making high- clarity containers having excellent physical properties due to biaxial orientation (see Section Blow Molding). A standard low- compression screw normally suffices, but improved heating to 200 °C is called for in the conditioning station before transfering to the final mold. Here, the blowing air must have a supply pressure of 3 MPa. Standard machines are easily modified to process PMP.

Extrusion. PMP can be used in most plastic extrusion processes for making film, fiber, sheet, pipe, and coatings for wire and paper. In all cases the extrusion stage should have ample barrel heating arrangements. Twin and single conventional full-flighted or three-stage screws used for PE or PP are suitable, provided that those generating excessive shear are avoided. However, the throughput is reduced with PMP, and surging tends to occur when the screw speed is increased. A more suitable single screw has a double flight in the compression zone to separate melt and solid polymer before it is plasticized (Fig. 57). This secures stable operation without any throughput penalties.

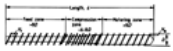

Figure 57. Screw for PMP [349]

$L/D \geq 28$; $CR = H_f/H_m = 2.5 – 3.0$; H_f = Depth of feed zone; H_m = Depth of metering zone

The downstream equipment used for making sheet, film, tubes, and coatings is similar to that used for PP, but with appropriate temperature adjustments. Fine fibers can be made by the blown melt method or by the long spin process (see Section Fibers and Flat Yarns).

Difficult and Impractical Processes. Thermoforming, either by vacuum, pressure, or matched-mold processes is difficult, but possible, with PMP thin sheet. The need to use high temperatures, followed by a long cooling period gives a cycle time of approximately 1.5 min. However, attractive food trays for microwave ovens are being made from thin sheet by using low draw down ratios. Rotational molding and mechanical forming are unsuited to PMP. Small amounts of foamed moldings are marketed.

5.6.2. *Applications*

The sales pattern of PMP worldwide in 1991 is reported by Mitsui Petrochemical Industries (MPC) as follows [350]:

Release film and paper	22 %
Industrial	20 %
Medical and laboratory	13 %
Cookware	11 %
Electric	7 %
Packaging	6 %
Others	21 %

PMP film, manufactured by MPC under the trade name Opulent, is available in different thicknesses (25 – 500 µm), with different flexibilities (modulus 0.7 – 1.6 MPa), and with alternative surface finishes. Three-layer laminates with Kraft paper or low-softening polyolefins extend the range. These products find extensive use in applications requiring good release characteristics combined with high temperature resistance and chemical inertness. Important examples include release films in manufacturing electronic circuit boards, in curing advanced composites for aircraft parts, and in continuous curing of rubber with heat or by UHF vulcanization. Inert mandrels made from flexible grades of PMP are used as supports during the manufacture and curing of reinforced high-pressure rubber hoses.

Good electrical properties and heat resistance meet specifications for use as covering for heat resistant wire, power, oil well, and communication cables. Glass-fiber-filled grades have sufficient shape retention after a few minutes at 260 °C to be used in dip soldering applications. Other industrial uses include replacement of stainless steel chemical treatment trays, sight glasses and floats, and a multitude of electrical insulators.

Labware and medical products were the first to take advantage of PMP. Transparency, convenience of sterilization with either γ-rays, ethylene oxide, dry heat, or steam autoclaves, and resistance to soiling are all relevant. Examples include blood test cells, medical instruments, tubing connectors, centrifuge cups, syringes and other measuring devices, urine collection tubes, and blood transfusion equipment. In the laboratory, PMP cages for small animals are easily cleaned and sterilized. Shatterproof transparent replacements for glass embrace most of the familiar items such as beakers, measuring cylinders and jugs, funnels, bottles, and Petri dishes.

Important outlets exist in food preparation where resistance to heat, oil, and steam is needed. Baking trays for bread, and thermoformed PMP- coated paper cartons for cooking foods, either in microwave or sometimes in gas-heated ovens, have superior releasability to aluminum foil containers. Application extend to coffee maker parts, popcorn popper covers, packages suitable for food retorting, microwave oven cookware, and service trays.

New areas under investigation include outlets for PMP fibers, such as nonwoven fabric, gas separation systems, and optical fibers.

5.7. Health and Environmental Aspects

PMP itself is regarded as inert and nontoxic, like polyethylene and polypropylene. Formulation systems for various grades are available to conform with food regulations in Japan, Germany, the United States, and the United Kingdom. Being a saturated hydrocarbon, PMP presents no problems in disposal by complete combustion. Flammability is the same as for PP. In principle, PMP can be recycled for reuse in similar or lower-quality applications, but attention must be paid to molecular mass reduction as a consequence of high-temperature processing.

[Top of Page]

References

General References

1. *Golden Jubilee Conference, Polyethylenes 1933 – 1983*, Preprints, The Plastics and Rubber Institute, London 1983.
2. *Golden Jubilee Conference, Polyethylenes 1933 – 1983*, Plenary Papers, The Plastics and Rubber Institute, London 1983.
3. R. B. Seymour, T. Cheng (eds.): *History of Polyolefins*, Reidel Publ. Co., Dordrecht 1986.
4. F. M. McMillan: *The Chain Straighteners*, Macmillan, London 1981.
5. R. A. V. Raff, K. W. Doak: *Crystalline Olefin Polymers*, parts I and II, Interscience Publishing, London 1965.
6. A. Renfrew, P. Morgan (eds.): *Polythene*, 2nd ed., Iliffe, London 1960.
7. S. van der Ven: *Polypropylene and other Polyolefins*, Elsevier, Amsterdam 1990.
8. *Aufbereiten von Polyolefinen*, VDI-Verlag, Düsseldorf 1984.
9. P. Ehrlich, G. A. Mortimer: “Fundamentals of the Free Radical Polymerization of Ethylene,” *Adv. Polym. Sci.* **7** (1970) 386 – 448.
10. Papers and Proceedings, *Polycon '84* LLDPE, Plastics and Rubber Institute, London 1984.
11. Conference Papers, *Polyethylene — The 1990s and Beyond*, Plastics and Rubber Institute, London 1992.
12. G. W. Gilby: “Ethylene-Vinyl Acetate Copolymers,” in A. Whelan, K. S. Lee (eds.): *Developments in Rubber Technology-3*, Applied Science Publishers, Barking 1983, chap. 4.
13. J. Boor Jr. *Ziegler-Natta Catalysis and Polymerization*, Academic Press, New York 1979.
14. W. Kaminsky, H. Sinn (eds.): *Transition Metals and Organometallics as Catalysts for Olefin Polymerization*, Springer Verlag, Berlin 1988.
15. R. P. Quirk (ed.): *Transition Metal Catalyzed Polymerizations*, Cambridge University Press, Cambridge 1988.
16. F. J. Karol, *Catal. Rev.-Sci. Eng.* **26** (1984) no. 3, 4, 557 – 595.
17. P. J. T. Tait: “Ziegler-Natta and Related Catalysts,” in R. N. Haward (ed.): *Developments in Polymerization-2*, Applied Science Publishers, London 1979, chap. 3.
18. Papers presented at MetCon 96, Worldwide Metallocene Conference, Houston, TX, USA 1996.
19. A. D. Horton: “Metallocene Catalysis: Polymers by Design?”, *Trends Polym. Sci.* **2** (1994) no. 5, 158 – 166. Links
20. M. Bochmann: Cationic Group 4 Metallocene Complexes and their Role in Polymerisation Catalysis: the Chemistry of Well-Defined Ziegler Catalysts. *Chem. Soc.*

Dalton Trans., **1994** ,255 – 270. Links

Specific References

21. W. L. Mattice, F. C. Stehling, *Macromolecules* **14** (1981) 1479. Links
22. see General Ref. [7].
23. F. M. Mirabella, E. A. Ford, *J. Polym. Sci. Polym. Phys. Ed.* **25** (1987) 777.
24. A. Keller, *Philos. Mag.* **2** (1957) 1171. *Prog. Phys.* **31** (1968) 623 See General Ref. [1], paper D1.
25. K. Bergmann, K. Nawotki, *Kolloid Z.* **219** (1967) 132.
26. S.-D. Clas, D. C. McFaddin, K. E. Russell, *J. Polym. Sci. Part B* **25** (1987) 1057. Links
27. R. S. Stein, R. Prud'homme, *J. Polym. Sci. Part B* **9** (1971) 595.
28. H. Schonhorn, H. L. Frisch, R. V. Albarino, *J. Polym. Sci. Polym. Phys. Ed.* **11** (1973) 1013.
29. L. Wild, T. Ryle, D. Knobeloch, *J. Polym. Sci. Polym. Phys. Ed.* **20** (1982) 441. G. Glöckner, *J. Appl. Polym. Sci.: Appl. Polym. Symp.* **45** (1990) 1 (review).
30. M. Aroca Hervás, *Revista de Plásticos Modernos* (1996) no. 481, 61 – 68. Links
31. Du Pont, US 4 076 698, 1957 (A. W. Anderson, G. S. Stamatoff).
32. M. Philip, I. M. Ward, B. Parsons, *J. Mat. Sci.* **21** (1986) 879.
33. L. C. E. Struik, *Polymer* **28** (1987) 1521. Links
34. P. C. Powell: *Engineering with Polymers*, Chapman and Hall, London 1983, p. 130.
35. Exxon Chemicals, Escorene Ultra EVA Copolymers, company brochure, Kraainem, Belgium.
36. J. Scheirs, L. L. Böhm, J. C. Boot, P. S. Leevers, *Trends Polym. Sci.* **4** (1996) no. 12, 408 – 415. Links
37. Y-L. Huang, N. Brown, *J. Polym. Sci. Part B* **29** (1991) 129 – 137; Links
38. Hoechst, Hostalen GUR Ultra High Molecular Weight Polyethylene, company brochure, Frankfurt 1989.
39. K. S. Whiteley in General Ref. [1], paper B2.
40. L. J. Young in Brandrup and Immergut (eds.): *Polymer Handbook*, 2nd ed., Wiley, New York 1975, Section II, p. 105.
41. L. J. Young, *J. Polym. Sci.* **54** (1961) 411.
42. DSM, Teamex Octene based VLDPE, company brochure, Geleen, The Netherlands, 1990.
43. R. J. Statz in General Refs. [1], paper B5 and [3], chap. 11.
44. W. J. MacKnight, T. R. Earnest, *Macromol. Rev.* **16** (1981) 41.
45. P. Mapleston, *Modern Plastics International*, June 1996, 56 – 61.
46. Shell, US 4 808 699, 1987 (J. A. M. Van Broekhoven, M. J. Doyle).
47. Hoechst, US 5 610 253, 1995 (W. Hatke, F. Osan).
48. A. Gilchrist (ICI): unpublished work, 1965.
49. R. Spahl, G. Luft, *Ber. Bunsenges. Phys. Chem.* **86** (1982) 621. G. Jungermann, G. Luft, *Ber. Bunsenges. Phys. Chem.* **91** (1987) 541.
50. K. E. Weale: *Chemical Reactions at High Pressures*, Spon, London 1967.
51. P.-C. Lim, G. Luft, *Makromol. Chem.* **184** (1983) 849.
52. M. J. Roedel, *J. Am. Chem. Soc.* **75** (1953) 6110. Links
53. M. E. A. Cudby in General Ref. [1], paper D6.
54. G. Luft, R. Kämpf, H. Seidl, *Angew. Makromol. Chem.* **108** (1982) 203. Links
55. L. Nicolas, *J. Chim. Phys.* **55** (1958) 185.
56. H. Tobita, *J. Polym. Sci. Polym. Phys.* **33** (1995) 841 – 853. Links
57. R. A. Jackson, P. A. Small, K. S. Whiteley, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 1781.
58. P. Ehrlich *et al.*, *A. I. Ch. E. Journal* **22** (1976) no. 3, 463.
59. Y. Tatsukami, T. Takahashi, H. Yoshioka, *Makromol. Chem.* **181** (1980) 1107.
60. Peroxid Chemie, Organic Peroxides P 3.2.1 (Half-Lifetimes), A 3.7.4 (Ethylene Polymerization I, II), company brochure, München.
61. W. W. Lawrence, S. E. Cook: Loss Prevention S.10, Chemical Engineering Progress Technical Manual, A.I.Ch.E. 1967.
62. D. G. H. Ballard *et al.*, *Polymer* **15** (1974) 169. Links
63. M. P. McDaniel, M. B. Welch, *J. Catal.* **82** (1983) 98. Links
64. C. E. Marsden: Paper 4A3 in General Ref. [11].
65. M. P. McDaniel, M. B. Welch, M. J. Dreiling, *J. Catal.* **82** (1983) 118. Links
66. M. P. McDaniel, M. M. Johnson, *Macromolecules* **20** (1987) 773. Links
67. F. J. Karol in General Ref. [16], p. 573.
68. V. A. Zakharov, Yu. I. Yermakov, *Cat Rev.-Sci. Eng.* **19** (1979) no. 1, 67.
69. W. L. Carrick *et al.*, *J. Polym. Sci.* A1, 10 (1972) 2609.
70. L. M. Baker, W. L. Carrick, *J. Org. Chem.* **35** (1970) 714.
71. F. J. Karol *et al.*, *J. Polym. Sci.* A1, 10 (1972) 2621.
72. Union Carbide, NL 68 16149, 1968.
73. Standard Oil (Indiana), US 2 692 257, 1951 (A. Zletz).
74. B. Diedrich, *Appl. Polym. Symp.* **26** (1975) 1.
75. Hoechst, DE 1 595 666, 1966 (B. Diedrich, K. D. Keil).
76. Solvay, BE 650 679, 1963.
77. Union Carbide, US 4 293 673, 1979 (A. D. Hamer, F. J. Karol).
78. Stamicarbon, BE 751 315, 1969.
79. ICI, GB 1 253 046, 1968 (J. C. Greaves, W. G. Oakes).
80. Du Pont, GB 1 209 825, 1967.
81. G. Henrici-Olivé, S. Olivé, *Adv. Polym. Sci.* **15** (1974) 1 – 30. Links
82. J. Herwig, W. Kaminsky, *Polym. Bull.* **9** (1983) 464. Links
83. W. Kaminsky: Paper 5.1 in General Ref. [1].
84. For a review of earlier work, see General Ref. [20].
85. BASF, US 4 404 344, 1981 (H. W. Sinn, W. O. Kaminsky, H. C. Vollmer, R. O. H. H. Woldt)
86. L. Resconi, S. Bossi, L. Abis, *Macromolecules* **23** (1990) 4489 – 4491. Links
87. P. J. Tait, M. G. K. Monteiro, M. Yang, J. L. Richardson: “Some Recent Advances in Supported Metallocene Catalysts,” in General Ref. [20].
88. M. Arndt, I Beulich, W. Kaminsky: “C1 Symmetric Metallocenes for Olefin Copolymerization: Catalytic Performance and Polymer Properties,” in General Ref. [18].
89. Dow, US 5 064 802, 1990 (J. C. Stevens, D. R. Neithamer). Exxon, US 5 026 798, 1990 (J. A. M. Canich).
90. Dow, EP 416 815, 1990 (J. C. Stevens, F. J. Timmers, D. R. Wilson, G. F. Schmidt, P. N. Nickias, R. K. Rosen, G. W. Knight, S. Lai).
91. M. Galimberti, T. Dall'Occo, F. Piemontesi, I. Camurati, G. Collina: “Metallocenes and Alumoxanes: Do they determine the Copolymer's Fingerprint,” in General Ref. [18].
92. Exxon, US 4 752 597, 1987 (H. W. Turner).
93. Dow, US 5 064 802, 1990 (J. C. Stevens, D. R. Neithamer).
94. P. Mapleston, *Modern Plastics International* **21** (1991) no. 3, 63; **19** (1989) no. 11, 36.
95. N. Baron: “A New Polyolefin Product Family Based on EXPOL Catalyst Technology”, paper presented at 2nd International Symposium on High Pressure Chemical Engineering, Erlangen 1990.
96. Exxon, US 4 530 914, 1983 (J. A. Ewen, H. C. Welborn).
97. Phillips, US 5 594 078, 1994 (M. B. Welch, R. L. Geerts, S. J. Palackal, T. M. Pettjohn).

98. G. A. M. Diepen, M. A. de Zeeuw in S. A. Miller (ed): *Ethylene and its Industrial Derivatives*, Ernest Benn, London 1969.
99. D. Commereuc *et al.*, *Hydrocarbon Process.* **63** (1984) no. 11, 118.
100. A. Chauvel, Y. Chauvin, J. Gaillard, J. Leonhard: *LLDPE in Europe*, Preprints to Conference, Madrid 1986, Plastics and Rubber Institute.
101. ICI, GB 1 456 841, 1973 (R. R. Cooper, K. S. Whiteley).
102. Y. V. Kissin, D. L. Beach, *J. Polym. Sci., Polym. Chem. Ed.* **24** (1986) 1069.
103. W. R. D. Manning, S. Labrow: *High Pressure Engineering, Chemical and Process Engineering Series*, Leonard Hill, London 1971.
104. B. C. Crossland: "Philosophical Thoughts on Standards for High Pressure Systems," *ASME Conference on Pressure Vessels and Piping*, Chicago 1986.
105. K. E. Bett, B. C. Crossland, H. Ford, A. K. Gardner in General Ref. [1], paper B1.
106. Esso, Research and Engineering, DT 2 151 649, 1970 (K. L. O'Hara).
107. J. P. Machon in Ref. [15] p. 344.
108. Charbonnages de France, GB 2 141 491, 1983 (D. Grasset).
109. Charbonnages de France, US 4 105 609, 1978 (J. P. Machon, B. Levresse, P. Gloriod).
110. K. Ziegler, GB 799 392, 1953.
111. J. P. Hogan in General Ref. [1], paper B3.
112. *Hydrocarbon Process* **62** (1983) no. 11; **64** (1985) no. 11.
113. L. Böhm, *Chem.-Ing.-Tech.* **56** (1984) 674.
114. Hoechst, DE 1 720 611, 1967 (B. Diedrich, K. D. Keil).
115. *Ullmann*, 4th ed., **19** p. 185.
116. Phillips Petroleum, US 3 405 109, 1960 (R. G. Rohlfing).
117. R. B. Staub in General Ref. [1], paper 5.4.
118. R. B. Staub: *LLDPE in Europe*, Preprints to Conference, Madrid 1986, Plastics and Rubber Institute, paper 1.
119. Union Carbide, US 4 003 712, 1975 (A. R. Miller).
120. D. E. James in Ref. [116], paper 2. Union Carbide, US 4 543 399, 1984 (J. M. Jenkins *et al.*).
121. T. K. Moynihan in General Ref. [1], paper 5.3.1.
122. *Mod. Plast. Int.* **27** (1997) no. 1 . Links
123. C. Krumm: *Polycon '84*, Papers and Proceedings, Plastics and Rubber Institute, London 1984, paper D.
124. Miller *et al.*, *Plast. Eng.* **42** (1986) 37. Links
125. S. Bork, *Kunststoffe* **74** (1984) 474.
126. G. R. Chapman, D. E. Priester, R. D. Souffie: *RETEC Polyolefins V*, Houston 1987, Society of Plastics Engineers, Brookfield, Conn., p. 271.
127. D. F. Klein in Ref. [126], p. 285.
128. W. A. Fraser, W. B. Herdle in Ref. [126] p.301.
129. R. D. Leaversuch, *Modern Plastics International* **20** (1990) no. 9, 55.
130. J-M. Yernaux: Paper S3B in General Ref. [11].
131. D. W. Woods, I. M. Ward, *Plastics, Rubber and Composites Processing and Applications* **18** (1992) no. 4, 255 – 261.
132. A. B. Brennan: "Surface Modification of Polyethylene Fibres for Enhanced Performance in Composites", *Trends Polym. Sci.* **3** (1995) no. 1, 4 – 11. Links
133. I. M. Ward: "A Review of Recent Developments in the Processing and Properties of Oriented Polyethylene", *Plastics, Rubber and Composites Processing and Applications* **19** (1993) no. 1, . Links
134. D. C. Prevorsek: "Preparation, Structure, Properties and Applications of Gel-spun Ultrastrong Polyethylene Fibers", *Trends Polym. Sci.* **3** (1995) no. 1, 12 – 21. Links
135. Rigidex HDPE Technical Manual No. 22 (BP).
136. Loctite Data Sheets: Primer 770 and Adhesive 406.
137. B.-A. Sultan, M. Palmlöf in General Ref. [11], paper 4B/3
138. B. E. Roberts, S. Verne in General Ref. [1], paper C3.1.
139. B. J. Lyons in General Ref. [1], paper C3.2.
140. L. Panzer: *ANTEC '88*, Conference Proceedings Society of Plastics Engineers, Brookfield, Conn., p. 1421.
141. D. C. Eagles, *Wire Technology International* **18** (1990) no. 6, 15. Mitsubishi, US 4 413 066, 1981 (T. Isaka, M. Ishioka, T. Shimada, T. Inoue).
142. Hoechst, Hostapren Chlorinated Polyethylene, company brochure, Frankfurt 1986.
143. D. A. Buck, F. A. Milcetich, P. D. Marsh, K. J. Kallish: *ANTEC '86*, Conference Proceedings, Society of Plastics Engineers, p. 803. Air Products, US 3 862 242, 1972 (R. C. Terrell).
144. G. Scott in S. A. Barenberg, J. L. Brash, R. Narayan, A. E. Redpath (eds.): *Degradable Materials*, CRC Press, Boca Raton 1990, p. 143.
145. W. J. Maddever, P. D. Campbell in Ref. [144], p. 237.
146. I. Dent, S. Hardman: "Plastics Waste Recycling — an Alternative Approach", I.Chem.E. Environmental Protection Bulletin Issue 044, 1996.
147. F. M. McMillan: *The Chain Straighteners*, McMillan Press, London 1979.
148. J. P. Hogan, R. L. Banks in R. B. Seymour, Tai Cheng (eds.): *History of Polyolefins*, Reidel Publ. Co., Dordrecht 1986, p. 103.
149. P. Pino, G. Moretti: "The Impact of the Discovery of the Polymerization of α-Olefins on the Development of Stereospecific Polymerization of Vinyl-monomers," *Polymer* **28** (1987) 683. Links
150. D. J. B. Potter, The Manor, Goose Lane, Ilminster, Somerset TA19 9RU: Consultant for Polypropylene Producers and Their Technologies, page preceding p. 55, consultant's Fig. 4.16.
151. D. J. B. Potter, The Manor, Goose Lane, Ilminster, Somerset TA19 9RU: Consultant for Polypropylene Producers and Their Technologies, page following p. 30, consultant's Fig. 4.1.
152. E. Seiler, *Kunststoffe Plast. Eur.* 1995, no. 8, 31.
153. R. Silvestri, L. Resconi, A. Pelliconi, Metallocenes '95, 1st Int. Congr. Mettalocene Polym., Brussels, April 26 – 27, 1995, p. 209.
154. B. Wunderlich: *Macromolecular Physics*, vol. **3**, Academic Press, New York 1980, p. 63.
155. J. Brandrupp, E. H. Immergut: *Polymer Handbook*, 3rd ed., J. Wiley, New York 1989, V27.
156. A. Turner-Jones, *Makromol. Chem.* **75** (1964) 134 –158.
157. S. van der Ven: *Polypropylene and other Polyolefins*, Elsevier, Amsterdam 1990, chap. 4.
158. A. W. Sloley, S. W. Golden, *Hydrocarbon Technol. Int. HTI Q.* **Spring 1995**, 111.
159. D. L. Smith, Am. Inst. Chem Eng., Summer Meeting, San Diego, Aug. 1990, Paper 34a.
160. H. Mark *et al.* (eds.): *Encyclopaedia of Polymer Science and Engineering*, vol. **13**, J. Wiley, New York 1987, p. 472 – 473.
161. B. V. Vora, P. R. Pujado, N. D. Adams, C. A. LeMerle, *International Conference, Polypropylene – the Way Ahead*, Madrid, Nov. 9 – 10, 1989, Paper 1/1 – 1/5.
162. *Eur. Chem. News* **53** (1989) 18th Dec., 14.
163. B. L. Goodall in van der Ven (ed.): *Polypropylene and other Polyolefins*, Elsevier, Amsterdam 1990, chap. 1.
164. Solvay et Cie, US 4 210 738, 1980 (J. P. Hermans, P. Henriqulie); US 3 769 233, 1973 (J. P. Hermans, P. Henriqulie).
165. R. P. Nielson in R. P. Quirke (ed.): *Transition Metal Catalysed Polymerizations*, vol. **4**, "part A", Harwood Academic Publ., New York 1983, p. 61.
166. P. Fiasse, *International Conference, Polypropylene – the Way Ahead*, Madrid, Nov. 9 – 10, 1989, paper 4/1.
167. Mitsui Petrochem. Ind., AU 89462 82, 1981, Example 1 (N. Kashiwa, A. Toyota).
168. *Chem. Eng.* (1985) July 8, 22.
169. W. Biddel et al., Metallocenes '96, 2nd Int. Congr. Mettalocene Polym., Düsseldorf, March 6 – 7, 1996, Session 3.
170. F. Langhauser, D. Fischer, S. Seelert, Metallocenes '95, 1st Int. Congr. Mettalocene Polym., Brussels, April 26 – 27, 1995, p. 245.
171. H. V. Vendilla, G. Guerra, P. Corradini, *Eur. Polym. J.* **27** (1991) no. 1, 45 – 54.
172. G. Natta, I. Pasquon, *Adv. Catal.* **XI**, Academic Press, London 1959, pp. 1 – 66.
173. T. Keii, *Kinetics of Ziegler-Natta Polymerization*, Kodansha, Tokyo 1972.
174. U. Giannini, G. Giunchi, E. Albizzati, P. C. Barbe in M. Fontanille, G. Guyot (eds.): *Recent Advances in Mechanistic & Synthetic Aspects of Polymerization*, Series C,

vol. 215, Reidel Publ. Co., Dordrecht 1987, pp. 473 – 484.
175. M.-J. Brekner, Metallocenes '96, 2nd Int. Congr. Mettalocene Polym., Düsseldorf, March 6 – 7, 1996, p. 155.
176. D. B. J. Potter, *Chem. Week* (1990) April, 22 –25.
177. E. Johnson, *Chem. Eng.* **97** (1990) no. 5, 30 – 37.
178. N. F. Brockmeier in R. P. Quirk (ed.): *Transition Metal Catalysed Polymerization, Alkenes & Dienes*, vol. **4**, "part B", Harwood Academic Press, New York 1983, pp. 671 – 696.
179. T. C. Cheng in R. B. Seymour, T. Cheng (eds.): *History of Polyolefins*, Reidel Publ. Co., Dordrecht 1986, p. 119.
180. D. Del Duca: "Le Polypropylène," *Euretec'88'*, Société Française des Ingénieurs Plasticiens, Paris, April 20 – 21, 1988, paper 5.
181. M. A. Ferrero, M. G. Chiovetta, *Polym. Plast. Technol. Eng.* **29** (1990) no. 3, 263 – 287.
182. K. D. Rumpler, J. F. R. Jaggard, R. A. Werner: "Le Polypropylène," *Euretec '88*, Société Française des Ingénieurs Plasticiens, Paris, April 20 –21, 1988, paper 4.
183. N. Hattori, *Chem. Econ. Eng. Rev.* **18** (1986) no. 3, 21.
184. H. K. Ficker, G. L. Goeke, G. W. Powers, *Plast. Eng.* **43** (1987) no. 2, 29 – 32. Links
185. R. D. Leaversuch, *Mod. Plast.* **20** (1990) no. 10, 46 – 48.
186. H. K. Ficker, D. A. Walker, *International Conference, Polypropylene – the Way Ahead*, Madrid, Nov. 9 – 10, 1989, paper 5/4.
187. P. Galli, T. Simonazzi, P. C. Barbe, *Conv. Ital. Sci. Macromol. (Atti) 6th*, **1** (1983) 113 – 126.
188. T. G. Heggs in D. C. Allport, W. Janes (eds.): *Block Copolymers*, Applied Science, London 1973, chapt. 4, 8D.
189. S. van der Ven: *Polypropylene and other Polyolefins*, Elsevier, Amsterdam 1990, chap. 5, 6.
190. R. Gachter, H. Muller: *Plastics Additives Handbook*, Hanser Verlag, München 1985.
191. H. Hinksen, F. K. Meyer, *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 25.
192. T. J. Henman: *World Index of Polyolefin Stabilizers*, Kogan Page, London 1982.
193. W. Meyer, H. Zweifel: "Le Polypropylène," *Euretec '88*, Société Française des Ingénieurs Plasticiens (SFIP), Paris 1988, paper 20.
194. B. Wunderlich: *Macromolecular Physics*, vol. **2**, Academic Press, New York 1976, p. 1.
195. K. T. Collington, *International Conference, Polypropylene – the Way Ahead*, Madrid, Nov. 9 – 10, 1989, paper 8/1.
196. A. Lucke, *Kunststoffe Plast. Eur.* **87** (1997) no. 3, 7 – 8. Links
197. R. C. Kowalski in R. B. Seymour, T. Cheng (eds.): *History of Polyolefins*, Reidel Publ. Co., Dordrecht 1986, p. 307.
198. G. R. Zeichner, C. W. Macosko, *Soc. Plast. Eng. 40th Ann. Tech. Conf., Eng. Props Struct. Div.*, San Francisco, May 1982, pp. 79 – 81.
199. ICI, Propathene Technical Data, P11/1 Injection Moulding, Middlesbrough, UK, 1989.
200. M. Kojima, *J. Macromol. Sci., Phys.* **B19** (1981) no. 3, 523 – 541.
201. *Mod. Plast.* **20** (1990) no. 1, 31, 39.
202. *Mod. Plast. Int.* **27** (1997) 65 – 68. Links
203. ICI, Propathene Technical Data, P12/1 Integral Hinges, Middlesbrough, UK.
204. ICI, Propathene Technical Data, P13/1 Structural Foam, Middlesbrough, UK.
205. ICI, Propathene Technical Data P14/1, Blow Moulding, Middlesbrough, UK.
206. R. D. Leaversuch, *Mod. Plast.* **20** (1990) no. 10, 46 – 48.
207. R. Mikielski: "Le Polypropylène," *Euretec '88*, Société Française des Ingénieurs Plasticiens (SFIP), Paris, April 4, 1988, paper 19.
208. J. Bonus: "Market Facts and Figures," *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham, Sept. 23 – 25, 1987, paper 2.
209. R. G. Mansfield: "Structure of Unsaturated Polypropylene Fiber Business," *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 3.
210. J. E. Ford: "Comparative Properties of Polypropylene and Other Man Made Fibers," *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 5.
211. M. Navone: "New Concepts and Development in Machinery and Process for Production of Polypropylene Fibers, *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 15.
212. N. D. Scott: "Polypropylene into Non Wovens," *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 38.
213. L. C. Wadsworth, A. O. Muschelewicz: "Effects of Resin and Process Variables on the Production of Ultrafine Melt Blown Polypropylene Fibers", *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 45.
214. ICI, Propathene Technical Data, P7/1, Tapes, Film Yarns, Fibers, Middlesbrough, UK, 1989.
215. P. Mapleston, *Mod. Plast.* **20** (1990) no. 6, 38.
216. N. Petiniot, A. Gemmi, M. Gillet, *Euretec 88*, Paris, April 21, 1988, paper 15.
217. K. F. Martin, J. Paul: "Manufacture and Civil Engineering Applications of High Strength Polypropylene Grids," *Fourth International Conference on Polypropylene Fibers and Textiles*, Nottingham 1987, paper 50.
218. ICI, Propathene Technical Data, P10/1, Pipework, Middlesbrough, UK, 1989.
219. Directory of UK Companies Involved in Recycling Plastics, 2nd ed., Brit. Plast. Federation, London 1989.
220. *Plastics & the Environment. Fact Sheet – Recycling*, Brit. Plast. Federation, London.
221. R. Armson, *Resources, Conservation and Recycling* **2** (1988) no. 1, 13 – 16.
222. T. W. Moffit: *Plastics Packaging – The Environmental Dimension*, Brit. Plast. Federation, London 1990.
223. *Eur. Plast. News* **15** (1988) no. 12, 25 – 31.
224. C. H. Kline, *Chem. Ind. (London)* 1989, July, 440.
225. C. H. Kline, *Chem. Ind. (London)* 1989, Aug., 483.
226. G. G. David, *Chem. Ind. (London)* 1989, Aug., 487.
227. Hüls: "Recycling Batteries," *Chem. Rundsch.* Nov. 10, 1989, p. 6.
228. British Plastics & Rubber: *Case Study by DTI (London) of Waste Minimisation at Chloride Battery Plant*, Feb. 1990, p. 15.
229. *Plast. Rubber Wkly*, Dec. 8, 1990, no. 1365, 2.
230. J. Maczko, *Plast. Eng.* **44** (1988) no. 6, 39. Links
231. *Chem. Eng. (N.Y.)* **97** (1990) no. 6, 15.
232. ICI, The Competiveness of LDPE, PP, PVC after the 1973 Oil Crisis, the ICI View, Welwyn Garden City, 1974.
233. B. Wunderlich: *Macromolecular Physics*, Academic Press, New York "Crystal Structure, Morphology, Defects,", vol. 1, 1973; "Crystal Nucleation, Growth, Annealing," vol. 2, 1976.
234. G. Natta, P. Corradini, I. W. Bassi, *Makromol. Chem.* **21** (1956) 240 – 244.
235. A. Turner Jones, *J. Polym. Sci. Part B* **1** (1963) 455 – 456; *Polymer* **7** (1966) 23 – 59. Links
236. S. W. Cornell, J. L. Koenig, *J. Polym. Sci. Part A-2* **7** (1969) 1965 – 1982.
237. I. T. Luongo, R. Salovey, *J. Polym. Sci. Part A-2* **4** (1966) 997 – 1008.
238. I. D. Rubin in H. Marowetz (ed.): *Polymer Monographs*, vol. 1, Gordon & Break Sci. Publ., New York 1968.
239. H. Tompa: *Polymer Solutions*, Butterworths, London 1956.
240. V. F. Holland, R. L. Miller, *J. Appl. Phys.* **35** (1964) 3241 – 3248. Links
241. G. Natta, *Angew. Chem.* **68** (1956) 393 – 403.
242. S. W. Cornell, I. L. Koenig, *J. Polym. Sci. Part A-2* **7** (1969) 1965 – 1982.
243. G. Videtto, A. I. Kovacs, *Kolloid Z. Z. Polym.* **220** (1967) 1 – 18.
244. W. Glenz, G. Goldbach, unpublished data from measurements made by Chemische Werke Hüls AG, Marl.
245. C. Geacintow, R. B. Milles, H. I. L. Schuurmans, *J. Polym. Sci. Part A-1* **4** (1966) 431 – 433.
246. G. Goldbach, G. Peistscher, *J. Polym. Sci. Part B* **6** (1968) 783 – 788.
247. F. Danusso, G. Gianotto, *Makromol. Chem.* **61** (1963) 139 – 156.
248. I. Powers, I. D. Hoffman, I. I. Weeks, F. A. Quinn, Jr., *J. Res. Natl. Bur. Stand. Sect. A* **69** (1965) 335 – 345.

249. H. Wilski, T. Grewer, *J. Polym. Sci. Part C* **6** (1964) 33 – 41.
250. Allied Chem. Corp., BE 652 381, 1965.
251. C. Geacintow, R. S. Schotland, R. B. Miles, *J. Polym. Sci. Part B* **1** (1963) 587 – 591; *J. Polym. Sci. Part C* **6** (1964) 197 – 207.
252. I. Boor, Jr., E. A. Youngman, *J. Polym. Sci.* **2** (1964) 903 – 907.
253. J. P. Rakus, C. D. Mason, *J. Polym. Sci. Part B* **4** (1966) 467 – 468.
254. G. Goldbach, *Angew. Makromol. Chem.* **29/30** (1973) 213 – 227. Links
255. G. Goldbach, *Angew. Makromol. Chem.* **39** (1974) 175 – 188. Links
256. C. D. Armeniades, E. Baer, *J. Makromol. Sci. Phys. B* **1** (1967) no. 2, 309 – 334.
257. I. D. Rubin, *J. Polym. Sci. Part A* **3** (1965) 3803 –3813.
258. I. Boor, I. C. Mitchel, *J. Polym. Sci.* **62** (1962) 70 –73.
259. J. Haase, R. Hosemann, S. Köhler, *Polymer* **19** (1978) 1358 – 1361. Links
260. J. Haase, S. Köhler, R. Hosemann, *Z. Naturforsch. A* **33** (1978) 1472 – 1483.
261. M. Goldstein, M. E. Seeley, H. A. Willis, V. J. I. Zichy, *Polymer* **19** (1978) 1118 – 1122. Links
262. J. Plenikowski, O. Umminger, *Kunststoffe* **55** (1965) 822 – 827.
263. M. L. Kasakevich, *Kunststoffe* **80** (1990) 54 – 57.
264. A. I. Foglia, *Appl. Polym. Symp.* **11** (1969) 1 – 18.
265. J. Plenikowski, *Kunststoffe* **55** (1965) 431 – 437.
266. I. P. Rakus, C. D. Mason, R. J. Schaffhauser, *J. Polym. Sci. Part B* **7** (1969) 591 – 595.
267. G. Goldbach, *Kunststoffe* **64** (1974) 475 – 481; *Angew. Makromol. Chem.* **74** (1978) 123 – 146. Links
268. W. Bollmann, *NDZ, Neue Deliwa-Z.* **9** (1972) 385 –391.
269. Unpublished data from measurements made on behalf of Shell Research.
270. C. R. Lindgren, *Polym. Eng. Sci.* **10** (1970) 163 –169.
271. J. Park *et al.*, *J. Am. Water Works Assoc.*, Oct. 1991, 71 – 78.
272. J. Park *et al.*, *J. Am. Water Works Assoc.*, Nov. 1991, 85 – 91.
273. M. Kasakevich *et al.*, *Kunststoffe* **80** (1990) 1361 –1362.
274. M. Ifwarson, T. Traenker, *Kunststoffe* **79** (1989) 827 – 830.
275. M. Ifwarson, P. Eriksson, *Kunststoffe* **76** (1986) 245 – 248.
276. M. Ifwarson, *Kunststoffe* **79** (1989) 525 – 529.
277. M. Ifwarson, P. Eriksson, "40A: Experience from 12 Years Evaluation of Cross-Linked Polyethylene," report at *6th International Conference Plastic Pipes VI*, York 1985, p. 40A.1.
278. J. Schultze, M. Homann: *C_4 Hydrocarbons and Derivatives*, Springer-Verlag, Berlin 1989.
279. A. Wouters, D. Hopper: "Polybutylene Film for Non-retortable Applications," *Pira 5th International Packaging Conference*, London Oct. 3 – 4, 1990.
280. A. Wouters: "Polybutylene Film for Non-retortable Applications," *Packforst Seminar Scanplast '91*, Gothenburg, March 6, 1991. R. Lee: "High Performance Lidding Films from Polybutylene Blends," *Pira International Conference*, Birmingham, Nov. 22, 1990.
281. S. W. Lebedew, E. P. Filonenko, *Ber. Dtsch. Chem. Ges.* **58** (1925) 163 – 168.
282. S. W. Lebedew, G. G. Kobliansky, *Ber. Dtsch. Chem. Ges.* **63** (1930) 103 – 112.
283. De Bataafse, GB 358 068, 1930 (H. D. Elkington).
284. I. G. Farbenindustrie, DT 641–284, 1931 (M. Otto, M. Müller-Cunradi).
285. P. H. Plesch: *The Chemistry of Cationic Polymerization*, McMillan Publ., New York 1963.
286. J. P. Kennedy: *Cationic Polymerization of Olefins*, A Critical Inventory, Wiley-Interscience, New York 1975, pp. 86 – 137.
287. H. Güterbock: *Polyisobutylen and Isobutylen-Mischpolymerisate*, Springer Verlag, Berlin 1959.
288. Cosden Petroleum, US 2 957 930, 1956 (W. K. Jackson); BE 584 683, 1959 (W. K. Jackson); *Chem. Eng. N. Y.* **67** (1960) no. 16, 98 – 101.
289. Standard Oil Development Co., US 2 698 320, 1952 (A. R. Garabrant, H. G. Goering, H. G. Schneider). Esso Research and Engineering, US 2 849 433, 1953 (H. G. Schneider, H. G. Goering, V. F. Mistretta); US 8 42 533, 1955 (R. P. Cahn).
290. Standard Oil Dev., US 2 176 194, 1936 (L. A. Bannon). Jasco, US 2 296 399, 1939 (M. Otto, H. G. Schneider).
291. I. G. Farbenindustrie, DT 496 482, 1937 (M. Otto, H. Güterbock, A. Hellemanns).
292. Exxon Research and Eng. Comp., US 2 893 981, 1954 (J. L. Ernst, A. B. Small, L. W. McLean); US 2 856 394, 1954 (A. B. Small, J. L. Ernst).
293. H.-U. Lenne, *Kolloid Z.* **137** (1954) 65 – 70. A. M. Liquori, *Acta Crystallogr.* **8** (1955) 345 – 347. Links
294. J. F. Gross, K. W. Nelson, J. M. Slobodin, *Zh. Fiz. Khim.* **25** (1951) 504 – 512. M. S. C. Flett, O. H. Plesch, *J. Chem. Soc.* 1952, 3355 – 3358.
295. J. P. Kennedy, E. Marechal: *Carbocationic Polymerization*, Wiley Interscience, New York 1982.
296. G. Natta *et al.*, *Atti Accad. Naz. Lincei, Cl. Sci. Fis., Mat. Nat., Rend.* **19** (1955) 397, 404. G. Natta, F. Danusso (eds.): *Stereoregular Polymers and Stereospecific Polymerization*, vol. **1**, Pergamon Press, New York 1967, p. 86.
297. Van der Ven: *Polypropylene and other Polyolefins*, Elsevier, Amsterdam 1990, p. 525.
298. *Kirk-Othmer*, 3rd. ed., **16**, John Wiley pp. 470 ff.
299. P. D. Trifedi, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Sci.)* **17** (1976) 803 – 807.
300. J. P. Kennedy, J. E. Johnston, *Adv. Polym. Sci.* **19** (1975) 37.
301. *Kirk-Othmer*, 2nd ed., **suppl. vol.**, pp. 800 – 807.
302. Brit. Petroleum, GB 912 824, 1962 (J. K. Hamblin).
303. Brit. Petroleum, GB 933 253, 1963 (G. W. Alderson, J. K. Hamblin).
304. Chevron Res. Co., US 3 251 895, 1966 (J. B. Wilkes).
305. Idemitsu Petr. Co., US 3 950 450, 1976 (H. Hashimoto, S. Omiya).
306. ASTM DS4A, Physical Constants of Hydrocarbons C_1 – C_{10}, American Soc. for Testing and Materials, Philadelphia 1971, p. 44.
307. Y. V. Kissin, in J. I. Kroschwitz (ed.): Encyclopedia of Polymer Science Engineering, vol. 9, Wiley, New York 1987, pp. 707 – 718.
308. In [301] p. 799.
309. ICI, GB 1 085 914, 1967 (K. J. Clark).
310. Y. V. Kissin, *Isospecific Polymerization of Olefins with Heterogeneous Ziegler-Natta Catalysts*, Springer-Verlag, New York 1985, pp. 227, 237.
311. V. S. Steinbeck *et al.*, *Eur. Polym. J.* **11** (1975) 457 –465.
312. K. J. Clark, R. P. Palmer, *The Chemistry of Polymerization Processes*, Monograph no. 20, Soc. of Chemical Industry, London 1966, pp. 82 – 103.
313. P. J. G. Tait, M. Ab Fid, A. E. Enenmo in R. B. Seymour, T. Cheng (eds.): *Advances in Polyolefins*, Plenum Press, New York 1987, p. 309.
314. Mitsui Petrochemical Industries, EP 0 253 625, 1986 (J. Yoshitake *et al.*)
315. BASF, GB 1 493 166, 1974 (H. Schick, H. Mueller-Tamm, N. Johann).
316. BASF, GB 1 489 312, 1974 (H. Schick) *et al.*.
317. Instit. Français du Petrole, FR 1 540 270, 1968 (Y. Chauvin, G. Lefebvre, M. Uchino).
318. Instit. Français du Petrole, FR 1 540 271, (Y. Chauvin, G. Lefebvre). See also [301] p. 794.
319. In [297] p. 529.
320. J. Brandrup, E. H. Immergut: *Polymer Handbook*, 3rd ed., John Wiley and Sons, New York 1989, V1/15.
321. J. H. Griffith, B. G. Ranby, *J. Polym. Sci.* **44** (1960) 369.
322. Y. Tanada, N. Kawasaki, K. Imada, M. Takayanagi, *Rep. Progr. Polym. Phys. Jpn* **9** (1966) 165 –168.
323. P. Zoller, H. W. Starkweather, G. A. Jones, *J. Polym. Sci. Polym. Phys. Ed.* **24** (1986) 1451.
324. G. Charlet, G. Delmas, J. H. Revol, R. St. J. Manley, *Polymer* **25** (1984) 1613 – 1619. Links
325. In [301] p. 795.

326. A. Turner-Jones, *Polymer* **6** (1965) 249 – 268. Links
327. In [312] p. 89.
328. In [312] p. 91.
329. Mitsui Petrochemical Industries, Technical Brochure, TPX, p. 6.
330. Mitsui Petrochemical Industries, TPX, no. 2312A, p. 8.
331. T. Tanigami, H. Suzuki, K. Yamamura, S. Matsuzawa, *Macromolecules* **18** (1985) 2595. Links
332. A. S. Hoffman, P. A. Frier, P. C. Condit, *J. Polym. Sci., Polym. Symp.* **4** (1963) 109.
333. L. Reginato, *Makromol. Chem.* **132** (1970) 113, 125.
334. N. C. Billingham, T. J. Walker, *J. Polym. Sci., Polym. Chem. Ed.* **13** (1975) 1209.
335. N. C. Billingham, P. Prentice, T. J. Walker, *J. Polym. Sci., Polym. Symp.* **57** (1976) 287 – 297.
336. F. Tudos, M. Iring, T. Kelen, in A. V. Patsis (ed.): *Advances in Stabilization and Controlled Degradation of Polymers*, vol. **1**, Technomic Publ. Comp., Pennsylvania 1989, p. 86.
337. Mitsui Petrochemical Industries, TPX, General Brochure, p. 7.
338. R. Arnaud, J. Lacoste, J. Lemaire, *Makromol. Chem. Rapid Commun.* **6** (1985) 803.
339. R. Chandra *et al.*, *Acta Polym.* **39** (1988) 265 – 267. Links
340. N. S. Allen, J. F. McKellar, *Br. Polym. J.* **9** (1977) 302 – 311.
341. ICI, early data on TPX, Welwyn Garden City.
342. N. S. Soboleva, S. S. Leshchenko, V. L. Karpov, *Vysokomol. Soedin, Ser. A* **25** (1983) 381. *Chem. Abstr.* (1983) 98 127069h.
343. Mitsui and ICI data, 1992.
344. G. Charlet, G. Delmas, *J. Polym. Sci. Part B, Polymer Phys.* **26** (1988) 1123.
345. Mitsui Petrochemicals Industries, TPX Document 2312A, p. 13.
346. Mitsui Petrochemicals Industries, TPX Film Brochure, p. 5.
347. Mitsui Petrochemicals Industries, TPX Document, TM-550, p. 6.
348. M. N. Matrosovich *et al.*, *Vysokomol. Soedin, Ser.* **B 22** (1980) 357. H. Ohi, K. Furuya in I. I. Rubin (ed.): *Handbook of Plastics Materials and Technology*, Wiley Interscience, New York 1990, pp. 375 – 383. G. C. Garland, L. M. Fodor in *Modern Plastics*, October 1991, p. 81. Mitsui Petrochemicals Industries and Phillips 66, Data sheets.
349. Mitsui Petrochemicals Industries, TPX Document TM-506, p. 3.
350. Mitsui Petrochemical Industries, personal communication.

[Top of Page]

Related Articles

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey Plastics, Properties and Testing Plastics, Processing Plastics, Additives and Plastics, Analysis, respectively. States of order are treated in ⟶ Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in ⟶ Polymerization Processes

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Polyoxyalkylenes

Polyoxymethylenes is a separate keyword. General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Frederick E. Bailey, Union Carbide Technical Center, Charleston, WV 25303, United States

Joseph V. Koleske, Consultant, 1513 Brentwood Road, Charleston, WV 25314, United States

1. Introduction 1
2. Properties 1
3. Production 5
4. Toxicity 7
5. Handling 8
6. Uses 8
7. Economic Aspects 9
8. References 10

1. Introduction

This article deals with polyoxyethylene, polyoxypropylene, and polytetrahydrofuran or poly(1,4-epoxybutane). Other names include poly(ethylene oxide), PEO; poly(propylene oxide), PPO; and poly(tetramethylene oxide), PTMO. Strictly speaking, the latter polymer is not derived from an alkene and thus is not a poly(alkylene oxide); however, it is included here, and the term poly(alkylene oxides) is used as a collective term for these three polymers. The poly(alkylene oxides) have an ether linkage in their main chain backbone and thus are polyethers. They are described by the repeat unit

$$-[(CH_2)_x-CHR-O]-$$

Poly(ethylene oxide):	R=H, x=1
Poly(propylene oxide):	R=CH_3, x=1
Poly(tetramethylene oxide):	R=H, x=3

Many details about the alkylene oxides and their polymers are given in [1].

2. Properties

Physical Properties. Typical physical properties of PEO, PPO, and PTMO are presented in Table 1. Except for PTMO, these polyethers can be linear or branched and are commercially available in a broad range of molecular masses from less than a few hundred to several million.

Low molecular mass polymers of ethylene oxide are clear, essentially colorless liquids and become partially crystalline solids at a molecular mass of ca. 800. As the molecular mass increases, they change from soft waxy solids to hard waxes. At high molecular mass, they are tough thermoplastics. Polyoxyethylenes of less than ca. 20 000 number-average molecular mass are usually referred to as poly(ethylene glycols). The crystalline polymer is spherulitic, has a unit cell (see Table 2) with four molecular chains, and has a fiber identity period of seven repeat units in two helical turns. The molecules are in an array of dihedral symmetry with twofold axes. One axis bisects the carbon–carbon bond and the other passes through the ether oxygens.

The glass transition temperature of PEO is listed as -52 °C in Table 1; however, this is the value for a high molecular mass, partially crystalline polymer. This parameter is complicated by crystallinity in the molecular mass range 10^3–10^5 [22], and abnormal values as high as -17 °C are found for the polymer with M = 6000, which is known to have the highest crystallinity (ca. 95 %). Macrospherulites or spherulites that can be seen without magnification have been prepared [23].

PEO is soluble in water, chlorinated hydrocarbons, aromatic hydrocarbons, methyl ethyl ketone, 2-ethoxyethyl acetate, butyl acetate, cyclohexanone, esters, dimethylformamide, and

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_579

Table 1. Selected physical properties of the poly(alkylene oxides)

	Poly(ethylene oxide)	Poly(propylene oxide)	Poly(tetramethylene oxide)
Density, g/cm^3			
Amorphous	1.127 – 1.128 [a] [2]	ca. 1.01 – 1.025 [a] [3]	0.975 (25 °C) [4]
Crystalline	1.21 (20 °C) [5]	1.096 (25 °C) [6]	1.07 – 1.08 [7]
Melting point, °C			
High molecular mass	65 [5]	75 [8]	58 – 60 [9]
Oligomeric [b]	30 – 50 [10]	[c]	10 – 45
Glass transition temperature, °C	− 52 [11]	− 73 to − 78 [12]	− 86 [4], [12]

[a] d_{20}^{20}.
[b] M_n ca. 400 – 3000.
[c] Liquid under ambient conditions.

Table 2. Crystal data for poly(alkylene oxides)

	Poly(ethylene oxide) [13–15]	Poly(propylene oxide) [16], [17]	Poly(tetramethylene oxide) [18–21]
Unit cell	Monoclinic	Orthorhombic	Monoclinic
a, nm	0.796	1.04	0.548 – 0.561
b, nm	1.311	0.467	0.873 – 0.892
c, nm	1.939	0.716	1.207 – 1.225
β	124.8°		134.2 – 134.5°

Table 3. Mark – Houwink constants for poly(alkylene oxides)

Solvent	Temperature, °C	$K \times 10^5$	*a*	Approximate molecular mass range	Average molecular mass	Reference
Poly(ethylene oxide)						
Water	30	12.5	0.78	$10^4 - 10^7$	Weight	[25]
Water	35	6.4	0.82	$10^4 - 10^7$	Weight	[26]
Aq. 0.45 M K_2SO_4	35	130	0.50	$10^4 - 10^7$	Weight	[26]
Aq. 0.39 M $MgSO_4$	45	100	0.50	$10^4 - 10^7$	Weight	[26]
Benzene	25	39.7	0.69	$8 \times 10^4 - 5 \times 10^6$	Weight	[27]
Carbon tetrachloride	20	69	0.61	200 – 10 000	Number	[28]
Toluene	35	145	0.70	400 – 4000	Number	[29]
Poly(propylene oxide)						
Acetone	25	75.5	0.56	1000 – 4000	Number	[30]
Benzene	20	11.1	0.77	700 – 3300	Number	[31]
Benzene	25	11.2	0.77	$3 \times 10^4 - 7 \times 10^5$	Weight	[32]
Toluene	20	20.8	0.72	700 – 3300	Number	[31]
Toluene	25	12.9	0.75	$3 \times 10^4 - 7 \times 10^5$	Weight	[32]
Poly(tetramethylene oxide)						
Benzene	30	131	0.60	$3 \times 10^4 - 10^6$	Weight	[33]
Cyclohexane	30	176	0.54	$3 \times 10^4 - 10^6$	Weight	[33]
Ethyl acetate	30	42.2	0.65	$3 \times 10^4 - 10^6$	Weight	[33]
Toluene	28	25.1	0.78	$3 \times 10^4 - 10^5$	Number	[34]

other solvents [1]. It has a solubility parameter of 10.3 [24]. It is insoluble in ethers and aliphatic hydrocarbons. Mark – Houwink constants for the intrinsic viscosity/molecular mass relationship $[\eta]=KM^a$ are given in Table 3. In water, PEO exhibits an inverse solubility – temperature behavior, becoming less soluble as the temperature is increased.

Intermolecular or association complexes between PEO and other compounds are readily formed due to the strong hydrogen bonding character of the ether oxygens. These complexes, which involve PEO and polymeric acids such as poly(acrylic acid) [35], urea, and thiourea [36], as well as various other inorganic and organic compounds [1], [37], have potential applications

in mechanochemical energy conversion systems [38], chemical valves for altering the porosity of membranes [39], and biosensors and carriers for controlled drug delivery [40]. Alkali-metal salt complexes have electrical properties that make them of interest as battery electrodes [41].

Polymers of propylene oxide remain clear, almost colorless liquids or amorphous solids over broad ranges of molecular mass. They can be prepared as amorphous or stereoregular polymers, with stereoregularity resulting in crystallinity. Polyoxypropylenes with $M <$ ca. 20 000 are usually referred to as poly(propylene oxide) polyols, and these polymers are by far the most important products from a commercial standpoint. They are usually marketed as di- and trihydroxyl polyols that are available in a broad range of molecular mass. (The third hydroxyl group comes from the initiator; for example if glycerine, which is the most common initiator for PPO polyols, is used, the PPO grows off all three hydroxyl groups and a trifunctional polyol results.) High molecular mass PPO polymers are used as specialty rubber and represent only a small worldwide demand of about 1000 t/a. The crystalline polymer has a unit cell (see Table 2) that contains two chains in a bowed zigzag conformation, with the methyl group of one molecule located between the ether oxygen and methylene group of the adjacent molecule.

PPO is soluble in most organic solvents. Very low molecular mass polyols (< 1000) are insoluble in aliphatic hydrocarbons such as hexane due to the high hydroxyl content. However, polyols with such high hydroxyl contents do have some solubility in water. Mark – Houwink constants for the intrinsic viscosity/molecular mass relationship $[\eta]=K M^{a}$ are given in Table 3.

Although not so thoroughly studied as those of poly(ethylene oxide), association complexes of poly(propylene oxide) are also known. These complexes are formed with a variety of salts such as lithium perchlorate [42] and metal halides, particularly chlorides [43]. Complexes of propylene oxide – ethylene oxide block copolymers with sodium and lithium salts exhibit marked increases in electrical conductivity [44] as compared to the uncomplexed copolymers.

Tetrahydrofuran can be polymerized to high molecular mass, but the commercially available poly(tetramethylene oxide) diols have molecular masses of ca. 650 – 3000 and a narrow molecular mass distribution. Although these polyols are usually crystalline in nature, their narrow molecular mass distribution [45] makes them easy to use, since even partial melting yields a product with effectively the same molecular mass distribution as that of a completely melted product. The crystalline polymer has a monoclinic unit cell whose characteristic dimensions are given in Table 2.

Solvents for PTMO, which has a solubility parameter of 17.3 – 17.6 [46], include aromatic hydrocarbons such as benzene and toluene, chlorobenzene, dichloromethane, chloroform, carbon tetrachloride, esters such as ethyl acetate, and nitrobenzene. Low molecular mass products are soluble in acetone, alcohols, and similar polar solvents. The polymers are insoluble in aliphatic hydrocarbons such as heptane, hexane, pentane, and petroleum ether. Only cursory studies of association complexes of PTMO have been conducted [43].

Chemical Properties. Poly(alkylene oxides) are chemically characterized as usually having primary (PEO and PTMO) and secondary (PPO) hydroxyl end groups [47]. PEO and PPO have been prepared with a variety of end groups including allyl or propenyl ethers, which arise during polymerization of ethylene oxide and propylene oxide through the termination reaction [48]; carboxyl [49], [50]; amine [51]; cyano [52]; and vinyl ether [53]. Alkyl or aryl ester terminated PEOs are used as nonionic surfactants [54], [55], and alkoxy-terminated PEOs prepared from alcohols or other monohydroxyl compounds are used for similar purposes as well as in cosmetic and pharmaceutical preparations. In addition to primary hydroxyl termination, PTMOs have been prepared with a majority of primary amino terminal groups [56]. Isocyanate-terminated poly(propylene oxides) and poly(tetramethylene oxides) are known as isocyanate-terminated prepolymers for preparation of polyurethanes.

Copolymers of various types can be prepared from the alkylene oxides. The best known of these are the polyurethanes, though they are not normally prepared from the poly(ethylene oxides). Polyurethanes are formed in a rearrangement reaction from poly(alkylene oxide) polyols, short chain diols such as 1,4-butanediol and ethylene glycol, and multifunctional isocy-

anates. The reaction takes place at relatively low temperatures, including room temperature, as depicted in the general, idealized manner (Fig. 1) where Y represents the residue of a polyfunctional isocyanate without its isocyanato groups. (In this generalization, it should be noted that the starter molecule for the poly(alkylene oxide) is omitted. If the alkylene oxide involved were propylene oxide, on the average both end groups would have been secondary hydroxyl groups.)

This unit would be the average repeating unit in the final polyurethane. A wide variety of isocyanates are used, including toluene-2,4-diisocyanate, 4,4′-diphenylmethane diisocyanate and its polymeric forms, 4,4′-dicyclohexylmethane diisocyanate, and isophorone diisocyanate. The latter two isocyanates are used when light-stable polyurethanes are required.

Urethane acrylates are formed by preparing an isocyanate-terminated prepolymer and then capping it with a hydroxyalkyl acrylate. These acrylate-terminated oligomers are widely used in radiation curable coatings.

Isocyanate-terminated PTMO polyols have been marketed as Adiprene (Du Pont) and Vibrathane (Uniroyal) precursors. Block copolymers of poly(tetramethylene oxide) and poly (butylene terephthalate) are marketed as Hytrel (Du Pont) for a variety of end uses.

Other copolymers include the very important ethylene oxide – propylene oxide copolymers [57], which occur in random, block, and graft form. Random copolymers are usually prepared from mixtures of monomers, and block copolymers by alternately feeding the two monomers. Block copolymers are often used as surfactants, metal lubricants, and functional fluids. Most of the block copolymers have poly(ethylene oxide) as the hydrophilic block and poly(propylene oxide) as the hydrophobic block. Other hydrophobic blocks include poly(styrene oxide), poly(tetramethylene oxide), and polystyrene [58]. ABA block copolymers of poly(1-proline) and amine-terminated PEO [59], as well as those of poly(ε-caprolactone) [60] and polyesters [61] grown onto hydroxyl-terminated PEO, have been prepared. A well-known graft copolymer, hydroxyethyl cellulose (→ Cellulose Ethers, Chap. 5.1.1.) [62], [63] is formed by swelling cellulose in a caustic solution and then adding ethylene oxide to the hydroxyl groups of the anhydroglucose units. Silica [64], gelatin and collagen [65], and chitin [66] are other polymeric substrates to which ethylene oxide has been grafted. Low molecular mass PEOs terminated with 1,2,4-triazoline-3,5-dione have been grafted onto polybutadiene [66].

A particularly useful class of grafted poly(propylene oxide) polyols was discovered in 1966 [67], [68]. These so-called polymer polyols are composed of vinyl polymers (e.g., acrylonitrile, styrene) and copolymers which are grafted to PPO or to PEO- capped PPO. They are produced by adding the vinyl monomers to the PPO polyol and dispersing them by agitation. A catalyst is then added to polymerize the vinyl monomers, which are believed to graft or partially graft to the polyol. The polymer polyol retains the end-group hydroxyl functionality, which is thus available for reaction with, for example, isocyanates. The vinyl polymer, which is incompatible with the polyol, is present in the form of a stable dispersion that acts as a reinforcing filler. These polymer polyols are used to manufacture polyurethane foams that have enhanced moduli while maintaining other desirable characteristics. Commercial products contain either grafted polyacrylonitrile or acrylonitrile – styrene copolymers, and they are used in high-resiliency automotive seating [69], microcellular elastomers [70], and other high-performance applications.

A deficiency of polyethers is their susceptibility to oxidation. PEO degrades by an autoxidation mechanism that is associated with formation and cleavage of intrachain hydroperoxides [71], [72]. Mechanical actions such as high shear can also cause degradation [73]. Hydroperoxide formation is facilitated by heavy metal ions, strong acids, and ultraviolet light. PPO and PTMO degrade by both chain scission and crosslinking [74], [75].

A number of investigations indicate that aromatic compounds such as substituted phenols, cresols, and quinones [76–78], mixtures of aliphatic amines and dihydric phenols (e.g., hydroquinone, *tert*-butylhydroquinone) [79], carbamates [80], and similar compounds function as stabilizers. A prime stabilizer for the poly(alkylene oxides) is 2,6-di-*tert*-butyl-4-methylphenol. Stabilization during storage is usually effected with 200 – 300 ppm of the stabilizer. However, in end uses involving solutions or slab stock polyurethane foam produced in

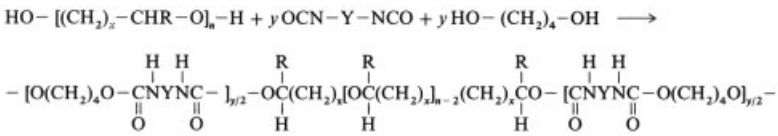

Figure 1. Formation of polyurethanes

bun form [1, p. 237] from poly(propylene oxide) polyols, 1000 – 5000 ppm are used to compensate for the long times at high temperature that are encountered due to retention of the heat of polymerization caused by the excellent insulation properties of the foam.

3. Production

Ethylene oxide was first polymerized in 1863 by Wurtz [81], who heated the oxide with water in a sealed tube. Earlier, in 1859, Lourenço [82] had obtained the polymer from ethylene glycol and ethylene bromide. Seven decades later Staudinger and Lohmann [83] prepared a number of polymers with a variety of catalysts, including alkali and alkaline earth metals.

There are three main ring-opening reactions for the polymerization of ethylene oxide and propylene oxide: acid, base, and coordinate initiation. Acid or cationic initiation [84–86] with an acid, HZ, involves addition of an active-hydrogen compound such as an alcohol or glycol, $(HO)_mX$ (m=1 – 4), to the 1,2-epoxide ring. In the initiation step, an oxonium complex forms and ring opening takes place by cleavage of the oxygen – carbon bond to form an initiated species that is available for propagation.

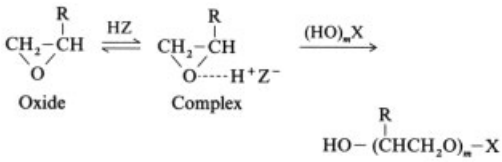

In the propagation step, n molecules of oxide are added to give a polymer having the formula $HO-[(CHRCH_2O)_n]_m-X$. Termination can occur by proton transfer, which yields an unsaturated product, or by anion transfer, which generates the starting acid.

Brønsted or Lewis acids such as $AlCl_3$, $SnCl_4$, PF_6, and BF_3 are used as initiators. The reaction is very fast, molecular mass is relatively low, and molecular-mass distribution is broad.

Base or anionic initiation [87–89] takes place in the following manner, where R′ is usually an alkyl group.

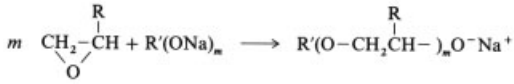

Propagation takes place in a manner similar to that described above. Termination can be by hydrogen abstraction from the backbone, which results in terminal unsaturation and formation of a new initiation site, or by rearrangement of the active species.

Polymerization rates for the two mechanisms when perchloric acid and sodium methoxide are used are given in Table 4 for a few selected alkylene oxides.

Table 4. Polymerization rate for selected alkylene oxides relative to ethylene oxide [90], [91]

Oxide	Initiator	
	Perchloric acid	Sodium ethoxide
Epichlorohydrin	0.065	4.8
Ethylene oxide	1.0	1.0
Propylene oxide	54	0.5
Isobutylene oxide	5100	0.06

Anionic initiators are usually metal alkoxides such as sodium or potassium methoxide or ethoxide. The reaction is usually slower than that obtained with acid initiation, except for epichlorohydrin and glycidol [90], [91]; relatively high molecular masses can be obtained and the molecular mass distribution is relatively narrow.

Ionic coordinative polymerization [92–95] involves a complex of a metal, M, such as Al, Sn, Fe, Mg, Li, etc., with ligands, L (e.g., OH, OR, Cl) that coordinates in an ionic manner with the oxygen atom of the epoxide. Complexation is followed by nucleophilic attack by the ligand, which results in opening of the oxirane ring to

give an initiated species that can undergo propagation to form the polymer chain.

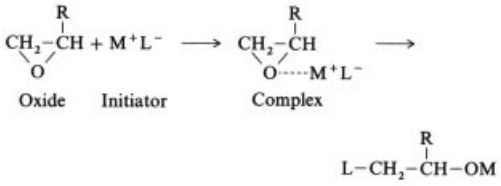

There are two categories of ionic coordination initiators [94], [96]. Initiators in the first category involve the formation of metal – oxygen coordination bonds, and those in the second category are alkaline earth compounds that have an anionic surface capable of coordinating with the oxide monomer.

In the case of propylene oxide, the monomer has an asymmetric carbon atom (C∗)

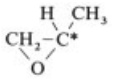

and when the ring is opened this asymmetric center is incorporated in the polymer chain. PPO can contain both head-to-tail and head-to-head sequences.

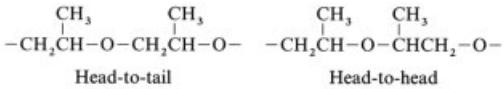

The head-to-tail sequences are formed by the expected carbon – oxygen cleavage between the methylene group and the oxygen atom of the epoxide, whereas the head-to-head sequences occur when unexpected cleavage takes place between the methyl-substituted carbon and the oxirane oxygen. Up to 30 % head-to-head structure has been found in certain polyoxypropylenes [97], [98].

Tetrahydrofuran was first polymerized by Meerwein and Kroning in 1937 [99], [100]. An I.G. Farbenindustrie patent [101] indicates that polymers of this monomer were the subject of numerous investigations from 1939 to 1943. Ring strain, a necessity for favorable polymerization, is small in this monomer compared to other cyclic ethers [102], and thus THF has less tendency to polymerize than ethylene oxide and propylene oxide. Although commercial polymers are dihydroxyl functional and low in molecular mass, a high molecular mass polymer with a melting point above 100 °C has been reported [103].

THF is a nucleophilic monomer that forms a tertiary oxonium ion during initiation and only polymerizes by a cationic ring-opening mechanism [104–108]. A variety of mechanisms for formation of the tertiary oxonium ion are known. These include direct alkylation or acylation, in situ formation from reactive organic halides and metal salts or Lewis acids, and direct addition of strong protic acids to the oxygen atom. For example [109], direct alkylation of the monomer by exchange with trialkyloxonium salts illustrates the initiation reaction.

Propagation then takes place by collisions between the initiated species and THF, whereby

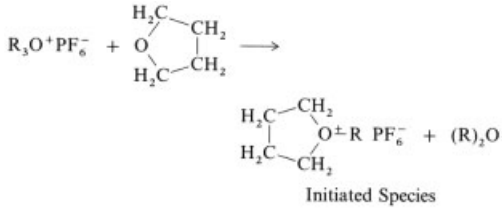

only collisions that occur at the α- carbon atom of the THF molecule result in propagation.

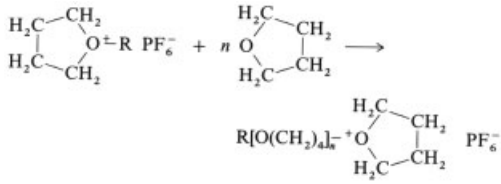

Water, oxygen, and other nucleophilic reagents can be used to terminate the polymerization.

Industrial poly(tetramethylene oxides) are hydroxyl terminated, and control of the termination reaction is important to the process. Hydroxyl groups, which have a profound effect on molecular mass, can be formed if strong protic acids are used to form the initiated species [110], [111]. The proton gives rise to an hydroxyl group on the end of the propagating species, and when water is used to terminate the polymerization, a second hydroxyl group is formed and the initiator acid is regenerated.

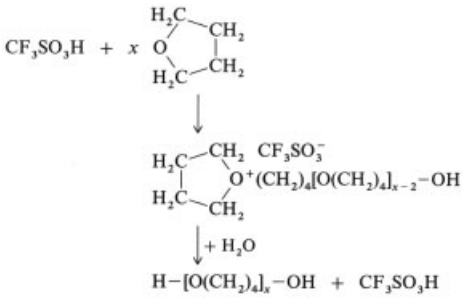

4. Toxicity [112]

Poly(ethylene oxides) have a very low single-dose oral toxicity that decreases with increasing molecular mass. Single-dose oral LD_{50} values in g/kg, for polymers of 200, 1000, 2000, and 6000 are, respectively, 34, 45, 45, and > 50 for male rats; 28, 32, > 50, and > 50 for female rats; 14, (not available), > 50, and > 50 for male rabbits; and 34, > 50, > 50, and > 50 for mice. PEOs of 400, 1540, and 4000 molecular mass were fed to dogs at a 2 % level in the diet with no adverse effect. When fed to rats at 2 %, PEO of 400 molecular mass had no effect, and at 4 % PEOs of 1540 and 4000 molecular mass had no effect. Doubling the dietary amounts for rats resulted in weight loss, an increase in liver weight, and slight histologic changes in the liver with the 400 compound, an increase in liver weight with the 1540 compound, and weight loss with the 4000 compound. Feeding of a 4 000 000 molecular mass compound at 5 % (rat) and 2 % (dog) diet had no detectable effect after 24 months. Testing with a radioactively labeled sample of the high molecular mass PEO indicated no significant absorption of the polymer from the gastrointestinal tract of the rat or dog.

PEOs do not appear to cause appreciable irritation to the eyes of rabbits. Skin irritation is apparently absent, which is attested to by the wide use of these compounds in cosmetic preparations. Overall, there are no reported adverse effects or human injuries from handling of PEOs [111].

Poly(propylene oxides) of molecular mass ca. 200 – 2000 have a single-dose oral toxicity that is noticeable, and they have mild eye irritability. They are not irritating to the skin, and even though there is some absorption through the skin, they do not appear to present a serious industrial health hazard. Inhalation of mists, particularly of the low molecular mass compounds (ca. 400 to 1200), could be hazardous. The low molecular mass compounds are rapidly adsorbed from the gastrointestinal tract and are potent CNS stimulants that readily cause cardiac arrhythmia. Excitement and convulsions are apparent within minutes after oral administration. Higher molecular mass compounds (> 2000) are very low in toxicity by all administration routes and do not stimulate the CNS. Oral LD_{50} values (rat) in g/kg for polymers of 400, 1200, and 4000 molecular mass are 1.6, 1.4, and > 15 for males and 2.1, 1.1, and > 15 for females, respectively.

Poly(tetramethylene oxides) are commercially available in four number-average molecular masses (650, 1000, 2000, and 2900) as dihydroxyl functional compounds [113], [114]. Acute oral toxicity (LD_{50} in g/kg) for albino rats was 11.3, 18.8, and > 34 for the 650, 1000, and 2000 molecular mass polyols, respectively. Acute dermal toxicity in albino rabbits was 8.37, 10.25, and > 10.25 g/kg for the same polyols. The poly-ols are rated as being minimally irritating to the eyes and skin of albino rabbits [112]; however, a Material Safety Data Sheet indicates that the products may irritate the skin and eyes [115].

5. Handling

Most manufacturers and suppliers of poly (alkylene oxides) have detailed, handling methods available for bulk, drum, and laboratory handling and storage, which should be consulted when working with the chemicals.

The poly(ethylene oxides) range from liquids at low molecular mass to hard waxes at moderate molecular mass, and to tough thermoplastics at high molecular mass (> ca. 100 000). They are not corrosive to most common construction materials. Although not strongly corrosive to copper and its alloys, some discoloration may occur if the compounds are kept in contact with copper- containing materials for prolonged times. Aqueous solutions of PEOs are corrosive to regular carbon steel; hence, stainless steel should be used for their storage. Tanks for bulk storage are usually made of plain carbon steel.

For smaller quantities, lined steel drums are used to minimize the pickup of traces of iron [116]. When short storage times of a month or less are involved, storage under air is acceptable. For prolonged storage, particularly in hot climates or under heated conditions, the products should be kept under an inert (nitrogen) atmosphere. Since PEOs are hygroscopic, dry conditions are important for most uses. PEOs have low vapor pressures and are not fire hazards unless high temperatures are involved. For example, polymers of molecular mass 200, 1000, 4000, and 9000 have Cleveland Open Cup flash points of 171 – 182 °C, 254 – 265 °C, 271 °C, and 271 °C, and Cleveland Closed Cup flash points of > 148 °C, > 177 °C, > 177 °C, and > 177 °C, respectively.

The poly(propylene oxides) are usually liquid, stable, noncorrosive compounds that have a very low vapor pressure. They are preferably stored at temperatures below 50 °C, usually in carbon steel tanks. However, if there is concern about the pickup of iron, tanks protected with a baked phenolic coating can be used [117]. PPOs should be blanketed with dry air or nitrogen to prevent absorption of moisture during storage. They are usually dried before use in polyurethane manufacture. Nitrogen is preferred for prolonged storage or storage at elevated temperature. They represent fire hazards only at high temperatures. For example, 400, 2000, and 4000 molecular mass polymers have Cleveland Open Cup flash points of 166 °C, 229 °C, and 229 °C, respectively.

The poly(tetramethylene oxides) are liquids or low-melting solids under ambient conditions. Mild steel and phenolic-lined steel tanks are suitable for storage. These products are not considered fire hazards unless high temperatures are encountered, and they have a Tag Open Cup flash point of > 163 °C. The compounds are unstable and may decompose at temperatures above 100 to 200 °C or at lower temperatures if in contact with high surface area substrates such as fibrous insulation [115]. When they decompose, very flammable tetrahydrofuran and carbon monoxide can be released.

6. Uses

Low molecular mass poly(ethylene oxides) (M <800) are usually used as intermediates for the manufacture of fatty acid ester surfactants that are used in a variety of formulated products. Illustrative of these products are soap or detergent suds control agents, foam stabilizers, fabric softeners, and optical brighteners. Typical detergents contain up to 20 % of a nonionic or anionic surfactant. Polymers with molecular mass of ca. 1000 – 2000 are low-melting waxes used in pharmaceutical ointments, cosmetic creams, and lotions. Their low toxicity, nonirritating nature, and smooth, creamy texture along with liquefaction at body temperatures make them particularly useful in cosmetic and pharmaceutical applications such as formulations for salves, suppositories, and ointments. Esters of these compounds are used as thickeners in cosmetic formulations.

The medium molecular mass PEO compounds, up to about 20 000, are hard, brittle waxes. They are used as adhesives, binders, plasticizers, molding compounds, wood impregnants, stiffening agents, and metal-working lubricants. Various PEOs are used in artifact preservation, particularly for items recovered after prolonged water immersion [118], [119]. In particular, copolymers of ethylene oxide and propylene oxide are often used in metal-working fluids and as quenchants, lubricants, and hydraulic fluids.

High molecular mass PEOs are used as textile sizes, as hydrodynamic friction reduction agents, in cosmetics and toiletries, as water-soluble films for seeds, in packaging, in adhesives, and in dentistry. The ability of the high molecular mass polymer to reduce hydrodynamic friction has led to applications in fire fighting and municipal sewage discharge control. The polymer is so effective in reducing the friction of water that it has been banned from use in competitive skiing and boating, where it markedly improved speed when incorporated into waxes and polishes. The association complexes of poly(ethylene oxide) have found use in controlled release of drugs, as adhesives, in artificial kidneys, and microencapsulation.

Poly(propylene oxides) are used as plasticizers, lubricants, solvents, softening agents, antifoaming agents, surface active agents, mold release agents, hydraulic fluid compositions, quenchants, and as intermediates in the production of resins (polyurethanes and alkyds). The largest use of the poly(propylene ox-

ides) and ethylene-oxide- capped PPO polyols is in the production of polyurethane foams and elastomers. The main polyurethanes are those based on toluene-2,4-diisocyanate and 4,4′-diphenylmethane diisocyanate, although light-stable versions based on cycloaliphatic polyisocyanates are made for specialty end uses. The polyurethanes are used as automobile seating, automotive crash pads, furniture cushioning, insulation, bedding, packaging, carpet underlay, molded parts, and many other applications.

The low molecular mass poly(tetramethylene oxide) polyols are used in coatings, to make block copolymers that are used for fabricated parts, and polyurethanes used as elastomeric fibers (spandex fibers). PTMOs have FDA approval for use in polyurethanes and adhesives that come into direct contact with dry foods [120]. The fibers have low hysteresis loss, very good tensile strength, and good tear and abrasion resistance. They are particularly useful for clothing, support garments, knitwear, swimwear, and other similar end uses. No commercial uses are known for high molecular mass PTMO.

7. Economic Aspects [121]

Estimates of the worldwide production of poly(alkylene oxide) polyols for polyurethanes is given in Table 5. Poly(propylene oxide) polyols account for about 95 percent of the polyols produced for polyurethanes, and poly(tetramethylene oxide) polyols account for about 2 percent of the total.

Table 5. Major worldwide production (1990) or consumption of poly(alkylene oxide) polyols [121]

Country	Production, 10^3 t
Production for polyurethane end uses	
Western Europe	931
United States	835
Japan	260
Korea and Taiwan	54
Total	2080
Production for nonpolyurethane end uses	
United States	159
Western Europe	92
Japan	57
Total	308

It is difficult to estimate amounts produced for nonpolyurethane end uses since major companies use the same facilities to produce for both markets. The production for nonpolyurethane end uses is given in Table 5. The majority of the polyol for nonpolyurethane use is for surface active agents (40 %) and for functional fluids and lubricants (31 %). The remainder is used in a wide variety of end uses (see Chap. 6). In the United States, production is about 18.6 % poly(propylene oxides), 31.4 % poly(ethylene oxides), and 50 % ethylene oxide – propylene oxide copolymers. In Western Europe production is about 38 % PPO and PEO, and the remainder copolymers. In Japan production is about evenly divided between PPO, PEO, and copolymers.

The largest manufacturers are:

United States	ARCO, BASF, Dow, E. R. Carpenter, Mobay, Olin, and Union Carbide
Western Europe	ARCO, BASF, Bayer, Dow, Montedipe, and Shell
Japan	Asahi Denka, Asahi Olin, Dow, and Mitsui Toatsu
Republic of Korea	Korea Polyol and Miwon
Taiwan	Chiunglong Petrochemical

Although production of poly(tetramethylene oxide) polyols is included in the production values for polyols as given in Table 5, production of this specialty polyol is delineated by company in Table 6. Polyurethane fibers and elastomers produced from these polyols have tensile properties as good as or better than those of polyesters, together with especially good hydrolytic stability and excellent low hysteresis loss [122].

Table 6. Worldwide production (1990) of poly(tetramethylene oxides) [121]

Company	Production, 10^3 t	Trade name
United States		
Du Pont	30	Terethane
QO Chemicals	11	Polymeg
Wesern Europe		
BASF	4	PolyTHF
Du Pont (Netherlands)	30	Terethane
Japan		
Hodogaya Chemical	3	
Mitsubishi Kasei	4	
Sanyo Chemical	1	

8. References

1. F. E. Bailey, Jr., J. V. Koleske: *Alkylene Oxides and Their Polymers,* Marcel Dekker, New York 1991.
2. G. M. Powell in R. L. Davidson (ed.): *Handbook of Water-Soluble Gums and Resins,* McGraw-Hill, New York 1980.
3. Union Carbide: Specification Polyols, United States 1987.
4. G. S. Trick, J. M. Ryan, *J. Polym. Sci C* **18** (1967) 93.
5. D. R. Braun in R. L. Davidson (ed.): *Handbook of Water-Soluble Gums and Resins,* McGraw-Hill, New York 1980.
6. G. Natta, L. Porri, P. Corradini, D. Morero, Atti Accad. Naz. Lincei, *Cl. Sci. Fis. Mat. Nat. Rend.* **20** (1956) 408.
7. P. Dreyfuss, M. P. Dreyfuss, *Adv. Polym. Sci.* **4** (1967) 528.
8. C. C. Price, M. Osgau, R. E. Hughes, C. Shambelan, *J. Am. Chem. Soc.* **78** (1956) 690.
9. A. C. Farthing in N. G. Gaylor (ed.): *High Polymers,* vol. 13, part I, Interscience, New York 1963.
10. B. Boener, H. Heitz, W. Kern, *J. Chromatogr.* **53** (1970) 51.
11. J. A. Faucher et al., *J. Appl. Phys.* **37** (1966) 3962.
12. J. Lehmann, P. Dreyfuss, *Adv. Chem. Ser.* **176** (1979) 587.
13. A. H. Kovacs, B. Lotz, A. Keller, *J. Macromol. Sci.-Phys.* **B 3** (1969) no. 3, 385.
14. J. L. Koenig, A. C. Angood, *J. Polym. Sci. Polym. Phys. Ed.* **8** (1970) 1787.
15. J. F. Rabot, K. W. Johnson, R. N. Zitter, *J. Chem. Phys.* **61** (1974) 504.
16. E. Stanley, M. Litt, *J. Polym. Sci.* **43** (1960) 453.
17. M. Cesari, G. Perego, A. Marconi, *Makromol. Chem.* **99** (1966) 194.
18. M. Cesari, G. Perego, A. Mazzei, *Makromol. Chem.* **83** (1965) 106.
19. K. Imada et al., *Makromol. Chem.* **83** (1965) 113.
20. H. Tadokoro, *J. Polym. Sci. Part C* **15** (1966) 1.
21. S. Kobayashi, K. Murakami, Y. Chatani, H. Tadakoro, *J. Polym. Sci. Polym. Lett. Ed.* **14** (1976) 591.
22. F. E. Bailey, Jr., J. V. Koleske: *Poly(ethylene oxide),* Academic Press, New York 1976.
23. F. E. Bailey, Jr., J. V. Koleske, *J. Chem. Ed.* **55** (1973) 761.
24. P. M. Galin, *Polymer* **24** (1983) 323.
25. F. E. Bailey, Jr., J. L. Kucera, L. G. Imhof, *J. Polym. Sci.* **32** (1958) 517.
26. F. E. Bailey, Jr., R. W. Callard, *J. Appl. Polym. Sci.* **1** (1959) 56, 73.
27. G. Allen et al., *Polymer* **8** (1967) 391.
28. C. Sadron, R. Rempp, *J. Polym. Sci.* **29** (1958) 127.
29. D. K. Thomas, A. Charlesby, *J. Polym. Sci.* **42** (1960) 195.
30. G. Meyerhoff, U. Maritz, *Makromol. Chem.* **107** (1967) 143.
31. W. Scholtan, S. Y. Lie, *Makromol. Chem.* **108** (1967) 104.
32. G. Allen, C. Booth, M. N. Jones, *Polymer* **5** (1964) 195.
33. M. Kurata, U. Utiyama, K. Kamada, *Makromol. Chem.* **88** (1965) 281.
34. S. U. Ali, M. B. Huglin, *Makromol. Chem.* **84** (1965) 117.
35. K. L. Smith, A. E. Winslow, D. E. Peterson, *Ind. Eng. Chem.* **51** (1959) 1361.
36. F. E. Bailey, Jr., H. G. France, *J. Polym. Sci.* **49** (1961) 397.
37. E. Tsuchida, K. Abe, *Adv. Polym. Sci. Ser.* **45** (1982) 1.
38. Y. Osada, *J. Polym. Sci. Polym. Chem. Ed.* **15** (1977) 255; **17** (1979) 3485.
39. S. Nishi, T. Kotaka, *Macromolecules* **19** (1986) 978; *Polym. J. (Tokyo)* **21** (1989) no. 5, 393.
40. R. K. Drummond, J. Klier, J. A. Alameda, N. A. Peppas, *Macromolecules* **22** (1989) 3818.
41. R. L. Peck, US 4 797 190, 1989.
42. J. Moacanin, E. F. Cuddihy, *J. Polym. Sci. Part C* **14** (1966) 313.
43. D. B. James, R. E. Wetton, D. S. Brown, *Polymer* **20** (1979) 187.
44. R. Xue, C. A. Angell: Report TR-10, (1987) Order no. AD-A 183502, *Chem. Abstr.* **108** (1988) 205 458.
45. T. G. Croucher, R. E. Wetton, *Polymer* **17** (1976) 205.
46. M. B. Huglin, D. J. Pass, *J. Appl. Polym. Sci.* **12** (1968) 473.
47. J. D. Ingram et al., *J. Macromol. Chem.* **1** (1961) 75.
48. D. M. Simmons, J. J. Verbanc, *J. Polym. Sci.* **44** (1960) 303.
49. K. B. Wagener, C. Thompson, S. Wanigatunga, *Macromolecules* **21** (1988) 2668.
50. S. Matsumura, N. Yoda, S. Yoshikawa, *Makromol. Chem. Rapid Commun.* **10** (1989) 63.

51. Dow Chemical, US 3 236 895, 1966 (J. M. Lee, J. C. Winfrey).
52. Japan Kokai Tokkyo Koho, JP 63 210 128, 1988 (A. Suekane, T. Masuda, H. Kawakami); *Chem.Abstr.* **110** (1989) 76 274.
53. K. S. Kazanskii, N. V. Ptitsyna, *Makromol. Chem.* **190** (1989) 255.
54. W. B. Satkowski, S. K. Huang, R. L. Liss in M. J. Schick (ed.): *Nonionic Surfactants,* Marcel Dekker, New York 1967.
55. A. N. Wrigley, F. D. Smith, A. J. Stirton, *J. Am. Oil Chem. Soc.* **34** (1957) 39.
56. S. Smith, A. Hubin, *J. Macromol. Sci. Chem.* **A 7** (1973) 1399.
57. F. T. Wall, *J. Am. Chem. Soc.* **63** (1941) 803.
58. R. A. Quirk, J.-J. Ma, *J. Polym. Sci. Polym. Chem. Ed.* **26** (1988) 2031.
59. S. H. Jeon, S. M. Park, T. Ree, *J. Polym. Sci. Polym. Chem. Ed.* **27** (1989) 1721.
60. R. Perret, A. Skoulois, *C. R. Heled. Seances Acad. Sci.* **268** (1969) 230.
61. Union Carbide, US 3 670 045, 1972 (J. V. Koleske, R. M.-J. Roberts, F. D. DelGuidice).
62. R. T. K. Cornwell in R. L. Whistler, J. N. BeMiller (eds.): *Industrial Gums,* Academic Press, New York 1959.
63. H. M. Spurlin in E. Ott, H. M. Spurlin, M. W. Graffin (eds): *Cellulose and Cellulose Derivatives,* part II, Interscience, New York 1954.
64. T. Tajouri et al., *Chim. Phys. Phys.-Chem. Biol.* **84** (1987) no. 2, 243. A. Gantar et al., *Polymer* **28** (1987) 1403.
65. T. Okujama, *Tanpakushitsu Kakusan Koso* **15** (1970) 47.
66. M. Weber, R. Stadler, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **30** (1989) no. 1, 109.
67. W. C. Kuryla, F. E. Critchfield, L. W. Platt, P. Stamberger, *J. Cell. Plast.* **2** (1966) 84.
68. Union Carbide, US 3 304 273, 1967 (P. Stamberger).
69. W. Patten, D. C. Priest, *J. Cell. Plast.* **8** (1972) 134.
70. R. A. Dunleavy, *J. Elastoplast.* **2** (1970) 1.
71. C. W. McGary, Jr., *J. Polym. Sci.* **46** (1960) 51.
72. L. Reich, S. S. Stivain, *J. Appl. Poly. Sci.* **13** (1969) 977.
73. A. Nakano, Y. Minoura, *J. Appl. Poly. Sci.* **15** (1971) 927.
74. L. E. St. Pierre, C. C. Price, *J. Am. Chem. Soc.* **78** (1956) 3432.
75. A. Davis, J. H. Golden, *Makromol. Chem.* **81** (1965) 38.
76. P. A. Okunev, *Nauchni Tr. Ivanov. Energ. Inst.* **14** (1972) 130.
77. Union Carbide, US 3 388 169, 1968 (J. E. Tyre, F. G. Willeboordse).
78. K. Konishi, Y. Sano, E. Ajisaka, JP 7 0 03 224, 1970.
79. California Research, US 2 687 378, 1954 (A. Goldschmidt, W. T. Stewart, E. Cerrito, O. L. Harle).
80. Celanese, US 3 144 431, 1964 (T. J. Dolce, F. M. Berardinelli, D. E. Hudgin).
81. A. Wurtz, *Ann. Chim. Phys.* **69** (1863) 330, 334.
82. A. Lourenço, *Compt. rend.* **49** (1859) 619.
83. H. Staudinger, H. Lehmann, *Ann. Chim.* **505** (1933) 41.
84. P. H. Plesch: *The Chemistry of Cationic Polymerization,* Pergamon Press, New York 1963.
85. J. P. Kennedy, E. Marachal: *Carbocationic Polymerization,* J. Wiley & Sons, New York 1968.
86. S. Sakai, T. Sugiyama, Y. Ishii, *Kogyo Kagaku Zasshi* **69** (1966) 699.
87. G. Gee, W. C. E. Higginson, K. Taylor, M. W. Trenholme, *J. Chem. Soc.* 1961, 1345.
88. B. Wojtech, *Makromol. Chem.* **66** (1966) 180.
89. S. Shigematsu, Y. Miura, Y. Ishii, *Kogyo Kagaku Zasshi* **64** (1961) 153; **65** (1962) 360.
90. R. E. Parker, N. S. Isaacs, *Chem. Rev.* **59** (1959) 737 (1959).
91. G. Gee, W. C. E. Higginson, P. Levesley, K. J. Taylor, *J. Chem. Soc.* 1959, 1338.
92. Union Carbide, US 2 706 181, 1955 (M. E. Pruitt, J. M. Baggett).
93. Union Carbide, US 2 914 419, 1959 (F. E. Bailey).
94. J. Furukawa et al., *Kogyo Kagaku Zasshi* **62** (1959) 1269.
95. Union Carbide, BE 557 883, 1957; US 2 987 988, 1961 (F. N. Hill, F. E. Bailey, Jr., J. T. Fitzpatrick).
96. J. Furukawa, T. Saegusa: *Polymerization of Aldehydes and Oxides,* Wiley-Interscience, New York 1963.
97. E. J. Vandenberg, *J. Polym. Sci. Part B* **2** (1964) 1085.
98. C. C. Price, R. Spector, *J. Am. Chem. Soc.* **87** (1965) 2069.
99. H. Meerwein, J. Kroning, *J. Prakt. Chem.* **2** (1937) no. 147, 257.
100. I.G. Farbenindustrie, DE 741 478, 1939 (H. Meerwein).
101. I.G. Farbenindustrie, FR 898 269, 1945.
102. H. de V. Robles, *Rec. Trav. Chim. Pays-Bas* **59** (1940) 184.

103. Du Pont, US 2 856 370, 1958 (E. L. Muetterties).
104. P. Dreyfuss, M. P. Dreyfuss, *Adv. Polym. Sci.* **4** (1967) 528.
105. G. Pruckmayr in K. C. Frisch (ed.): *Cyclic Monomers,* Wiley-Interscience, New York 1972.
106. B. A. Rosenberg, W. I. Ioshak, N. S. Enikolopian: *Interchain Reactions in Polymeric Systems,* Chemistry, Moscow 1975.
107. Y. Eckstein, P. Dreyfuss, *Anal. Chem.* **52** (1980) 537.
108. K. Yasuda, Y. Yokoyama, S. Matsumoto, K. Harada, in E. J. Goethals (ed.): *Cationic Polymerization and Related Processes,* Academic Press, New York 1984.
109. P. Dreyfuss, M. P. Dreyfuss, *Adv. Chem. Ser.* **91** (1969) 335.
110. G. Pruckmayr, T. K. Wu, *Macromolecules* **11** (1978) 265, 662.
111. K. Matsuda, Y. Tanaka, T. Sakai, *J. Appl. Polym. Sci.* **20** (1976) 2821.
112. *Patty* 3rd ed., **2 C.**
113. Quaker Chemical, POLYMEG Polyols, Applications and Properties Data, United States, 1976.
114. Du Pont, TERETHANE Polyether Glycol, Properties, Uses, Storage and Handling of Du Pont Glycols, United States, 1985.
115. Du Pont, Material Safety Data Sheet, TERATHANE Polyether Glycols, Wilmington 1985.
116. Union Carbide, CARBOWAX Polyethylene Glycols, United States, 1986.
117. Union Carbide, Specification Polyols, United States, 1987.
118. A. M. Rosenqvist, *Stud. Conserv.* **4** (1959) 62.
119. K. Morris, B. L. Siefert, *J. Am. Inst. Conserv.* **18** (1978) 33.
120. United States Food and Drug Administration Regulations 121.2550 and 121.2562.
121. SRI International: *Chemical Economic Handbook,* Menlo Park 1991.
122. W. J. Pentz, R. G. Krawiec, *Rubber Age* 1975, Dec., 31.

Polyoxymethylenes

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

GUENTER SEXTRO, Hoechst AG, Frankfurt, Federal Republic of Germany

1. Introduction 1
2. Physical and Chemical Properties . . 2
3. Production 6
3.1. Polymerization of Formaldehyde and Trioxane 6
3.2. Homopolymers 6
3.3. Copolymers 7
3.4. Additives 8
3.5. Impact-Resistant Types 9
4. Economic Aspects 9
5. Quality Specifications and Analysis . 9
6. Product Range and Processing 10
7. Uses 11
8. Recycling 11
9. Toxicology and Occupational Health . 11
10. References 12

1. Introduction

Polyoxymethylenes are thermoplastically processable plastics whose structure is predominantly composed of unbranched chains of oxymethylene units:

$$HO(CH_2O)_nH$$

Polyoxymethylenediol

The term polyacetals is also commonly used for such products. The first systematic investigations of the production and properties of polyoxymethylenes were carried out by STAUDINGER and KERN [1] in the 1920s. Basic properties such as the relatively high softening temperature of ca. 175 °C, high crystallinity, and the ability to form films and fibers were already recognized at that time. However, the insufficient thermal stability of polyoxymethylenes at processing temperatures prevented any practical use of the products.

Following intensive work begun at the end of the 1940s, Du Pont made a breakthrough in 1956 with the production of a high-molecular, thermally stable polyoxymethylene that could be processed thermoplastically [2]. The decisive step consisted of blocking the unstable hemiacetal terminal groups by esterification with acetic anhydride. A production plant with ca. 7000 t/a capacity came on stream in the United States in 1959. The product was introduced on the market under the name Delrin.

$$CH_3-\overset{\overset{O}{\|}}{C}-O(CH_2O)_n\overset{\overset{O}{\|}}{C}-CH_3$$

Polyoxymethylene diacetate (homopolymer)

A few years later Celanese in the United States and Hoechst in Europe succeeded in producing a thermally stable acetal copolymer [3], [4] by copolymerizing trioxane with small amounts of cyclic ethers or cyclic acetals. Compared with homopolymers with blocked terminal groups, the copolymers have the advantage that they are resistant to alkalies, though they have a somewhat lower crystallinity.

$$HO-CH_2CH_2O(CH_2O)_nCH_2CH_2O(CH_2O)_m \cdot \ldots -CH_2O-CH_2CH_2OH$$

Polyoxymethylene produced from trioxane and ethylene oxide (copolymer)

Production of the acetal copolymer Celcon came on stream in 1962 in a 7000 t/a plant at the Celanese plant in Bishop, Texas. The year before Celanese and Hoechst in Germany had already established the joint venture Ticona Polymerwerke, which commenced production of the acetal copolymer Hostaform in 1964. In 1962 Celanese and Daicel formed the joint com-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_591

pany Polyplastics in Japan and commenced production of the acetal copolymer Duracon. This was followed in the 1970s by Ultraform (established jointly by BASF and Degussa; product name: Ultraform), the Asahi Chemical Industry Co. (product name: Tenac), and the Mitsubishi Gas Chemical Co. (product name: Iupital). Further plants were built in Poland (product name: Tarnoform) and the former Soviet Union. In the 1980s and early 1990s new plants were constructed mostly in Asia (Korea Engineering Plastics: Kepital; Lucky-Goldstar: Lucel; Taiwan Engineering Plastics).

Polyoxymethylenes are now widely used in areas that were for a long time the preserve of metals. Their low weight and economic processing together with an advantageous combination of mechanical, thermal, electrical, and chemical properties have led to a wide range of applications.

About one third of the market for polyoxymethylenes is accounted for by homopolymers, and about two thirds by copolymers. Capacities of ca. 440×10^3 t/a can be compared with a world demand of ca. 390×10^3 t/a for 1990.

Further details of the production, properties, processing, and uses of polyoxymethylenes can be found in [5], [6].

2. Physical and Chemical Properties

Important physical properties of homopolymers are summarized in [7], and those of copolymers in [8], [9]. Some physical data for polyoxymethylene are given in Table 1 and [10], [11]. Computer disks with materials properties can be obtained from the Western European producers (Material Data Bank CAMPUS: Computer Aided Material Preselection by Uniform Standards [21]).

Morphology. Polyoxymethylene crystallizes in a hexagonal structure with the lattice constants a = 474 pm and c = 1739 pm. The polymer chains have a helical conformation with an identity period of 9/5 [22], [23].

The crystalline density is 1.49 g/cm^3 [22], and that of the amorphous state 1.324 g/cm^3 (homopolymer) and 1.320 g/cm^3 (copolymer) [14]. The density of the commercially available partially crystalline polyacetals is 1.42 g/cm^3 and 1.41 g/cm^3 for homopolymers and copolymers, respectively [7], [8], [24].

The crystallinity of polyoxymethylenes varies between 50 % and 80 %, depending on the comonomer content, molecular mass, and sample pretreatment. Crystallinities of 64 – 69 % for homopolymers and 56 – 59 % for copolymers were found for identical thermal histories [14]. The degree of crystallinity can be increased by thermal posttreatment. Considerably higher values for the crystallite melting point and degree of crystallinity have been reported for the as-polymerized virgin polymer [25].

During crystallization from the melt spherulitic superstructures form. The size of the spherulites depends, among others, on the crystallization temperature and number of nucleation centers. The overall crystallization rate and number of spherulites formed can be increased by adding nucleation agents (see Section 3.4).

Mechanical Properties. Polyoxymethylenes exhibit high hardness and rigidity, high toughness down to $-40\,^{\circ}$C, high thermal dimensional stability, low water absorption, good electrical and dielectric properties, good elasticity, favorable slip and wear behavior, and are easy to process. This combination of favorable properties means that polyoxymethylenes are widely used construction materials.

The structural difference between homopolymers and copolymers results in a lower degree of crystallinity of copolymers. As a consequence homopolymers have about 10 – 12 % higher hardness, rigidity, and strength [5]. However, copolymers possess advantages with regard to thermal and chemical stability.

Standardized test methods are used to determine the physical properties. Since polyoxymethylenes are generally processed by injection molding, the mechanical properties are usually measured on injection-molded test specimens. Slight differences in test specimen production and mechanical testing can result in changes in the measured values. For this reason it is only possible to a limited extent to make an absolute comparison of the property values given in the manufacturers' product brochures [7–9]. Further details are given in [11].

Table 1. Physical properties of polyoxymethylene

Property	Value	Reference
Activation energy of polymerization, kJ/mol		
of cationic polymerization of trioxane in dichloroethane	79.5	[12]
in the solid state	36	[12]
Ceiling temperature, °C		
polymerization of formaldehyde (gas) → polyoxymethylene (solid) at 1 bar	116 – 137	[10]
Coefficient of linear thermal expansion, k^{-1} (ASTM D 696)		
copolymer	0.85×10^{-4}	[13]
homopolymer	0.81×10^{-4}	[13]
Crystallinity, %		
Delrin 550, 20 °C, from density data	68.7	[14]
Hostaform C 9020, 20 °C, from density data	58.7	[14]
Density, g/cm^3		
homopolymer, Delrin 550, 20 °C	1.435	[14]
copolymer, Hostaform C 9020, 20 °C	1.416	[14]
Dielectric constant at 10^3 Hz (DIN 53 483, ASTM D 150-59 T)	3.6 – 4.0	[10]
Dielectric loss tan δ at 10^3 Hz	$(15-50) \times 10^{-4}$	[10]
Dielectric strength, kV/mm (DIN 53 481, 50 Hz, 0.5 kV/s, 0.2 mm film)	50 – 70	[10]
Enthalpy, kJ/kg		
homopolymer, 300 K	222.5	[15]
copolymer, 300 K	211.4	[15]
Entropy, kJ kg^{-1} K^{-1}		
homopolymer, 300 K	1.487	[15]
copolymer, 300 K	1.439	[15]
Heat capacity, J mol^{-1} K^{-1}		
homopolymer, 300 K	42.79	[15]
copolymer, 300 K	41.11	[15]
molten polyoxymethylene, 300 K	57.67	[16]
Heat distortion temperature under load, °C (ASTM D 648)		
homopolymer	124	[10]
copolymer	110	[10]
Heat of combustion of polyoxymethylene, kJ/mol	-511.6 ± 1.7	[17]
Heat of fusion, kJ/kg		
homopolymer	236	[18]
copolymer	181 – 192	[19]
Heat of polymerization, kJ/mol		
formaldehyde (gas, 1 bar, 298 K)	− 55 to −72	[10]
trioxane (liquid → solid)	− 22.6	
Melting point, °C		
Delrin 500	175	[7]
Hostaform C 9021	164 – 167	[8]
Specific volume, cm^3/g		
Delrin 550, at 20 °C	0.6968	[14]
Hostaform C 9020	0.7064	[14]
Surface resistivity, Ω		
homopolymer (ASTM D 257)	2×10^{13}	[13]
Thermal conductivity at 20 °C (VDE 0304), W m^{-1} K^{-1}	0.292	[20]
Volume resistivity (film 0.2 mm, DIN 53 482/ASTM D 257-61), Ω · cm	$> 10^{15}$	[8]
Water absorption (ASTM D 570), %		
after 24 h immersion		
homopolymer	0.25	[13]
copolymer	0.22	[13]
equilibrium, 50 % R.H.		
homopolymer	0.2	[13]
copolymer	0.16	[13]

The suitability of a structural component for a given purpose depends not only on the mechanical properties of the selected raw material, but to a large extent on its structural form and shape [26], [27]. In addition there are influences resulting from the production and environmental conditions, as well as time, temperature, and stress level.

Thermal Properties. The thermal properties of polyoxymethylenes essentially depend on the crystallinity and chemical structure, which are also responsible for differences in the thermal behavior of homopolymers and copolymers.

Homopolymers have higher crystallite melting points (175 °C) than copolymers (164 – 167 °C). Further information on structural changes with temperature are obtained from the torsion vibration test (DIN 53 445), from the temperature function of the mechanical shear modulus G', and the mechanical loss factor d. The transition region (relaxation region) at – 60 °C represents an important embrittlement limit under impact conditions. A second relaxation region exists at 0 °C. A third, subsidiary maximum above 100 °C is attributed to molecular motion in the crystalline regions [11] and is important for the thermal dimensional stability.

The continuous use temperatures are, depending on the type of stress, about 90 – 110 °C [11] and 80 – 100 °C [28]. According to Underwriters Laboratory, Hostaform is permitted for use up to 80 – 105 °C [8]. Experience shows that short-term thermal loading (several hours) up to 140 °C is permissible [28].

The continuous use temperature is largely determined by thermal aging processes, which generally involve thermal-oxidative damage at high temperature. Thermal-oxidative decomposition and decomposition caused by acidic degradation products can be suppressed by adding oxidation stabilizers and costabilizers (see Section 3.4 and [29]).

Thermal stability and continuous use temperatures are, however, not accurately definable material-specific properties and must always be considered in the context of the relevant application.

Temperatures of 190 – 210 °C are generally recommended for processing polyoxymethylenes. Higher temperatures can be used with short-er residence times (e.g., up to 230 °C for copolymers) [6], [8].

The specific heat of copolymers is 1.47 kJ kg^{-1} K^{-1}, and the thermal conductivity is 0.31 W m^{-1} K^{-1} (at 20 °C) [8]. Further thermal data are given in Table 1 and [10], [11].

Chemical Stability. Owing to their high degree of crystallinity, polyoxymethylenes exhibit good resistance to numerous chemicals. Copolymers are also resistant to strong alkalies. Polyoxymethylenes are not resistant to strongly acidic and oxidizing substances.

Only hexafluoroisopropanol and hexafluoroacetone sesquihydrate are suitable as solvents at room temperature. Most other solvents only dissolve polyoxymethylenes above the melting point; these include polar substances such as benzyl alcohol, dimethylformamide, γ-butyrolactone, and phenol [30].

Various investigations on the stability towards numerous chemicals have been published, for example [8], [31], [32].

Conventional organic solvents such as alcohols, esters, ethers, ketones, aliphatic and aromatic hydrocarbons, as well as perhalohydrocarbons do not attack polyoxymethylenes at room temperature even under continuous contact. The mechanical properties are only slightly affected, and the degree of swelling is small.

The good resistance to all common fuels (regular, premium, also fuels containing 15 – 20 % methanol, diesel), as well as to oils and greases, coolants, and brake fluids means that polyoxymethylenes are widely used as construction materials in the automotive sector [8], [33].

Polyoxymethylenes swell only very slightly in contact with water; the water absorption is 0.6 % at 20 °C, and 1.6 % at 100 °C [8]. Copolymers in particular are extremely stable at high temperatures. For example, the tensile strength decreases by only 25 % after two years in boiling water. The copolymers are therefore widely used in washing machines and dishwashers, as well as in water boilers. Alkaline detergents and rinsing agents do not affect the stability of the copolymers. Polyoxymethylenes are generally not subject to stress- cracking.

As regards use in the sanitary sector, varying chlorine contents in tap water, as well as temperature and pressure fluctuations mean that the

suitability must be individually assessed in each case.

UV and Weather Resistance. Polyoxymethylenes can be protected to a certain extent against UV radiation by the use of additives. Suitable UV absorbers include those based on benzophenone or benzotriazole derivatives. In practice, they are combined with HALS stabilizers, which gives a synergistic effect. A further improvement can be achieved by adding carbon black or pigments such as titanium dioxide (see Section 3.4).

Weathering results in a loss of surface gloss, followed by so-called chalking, caused by a layer of degraded, low molecular mass polyoxymethylene. This layer protects deeper-lying layers, so that the reduction in the mechanical properties is greater at the start of weathering, but then decreases. The damage to thin-walled parts is thus greater than to thick-walled parts. The behavior of non-light-stabilized copolymers under natural and artificial weathering is described in [34]. Investigations of the mechanism are given in [35–38].

Various tests are used to evaluate the UV and weathering resistance. Artificial irradiation in the Xenotest and in the Fade-O-Meter produces similar results to natural weathering [37]. Weathering in the Florida test can, depending on the intended use, be performed in the open air or behind window glass. It is not possible to interconvert the various results.

Polyoxymethylenes are only moderately stable to γ-radiation. Medical equipment can be sterilized by using conventional radiation doses of 2.5×10^4 J/kg [8]. The deterioration in the mechanical properties increases with increasing radiation dose [39].

Permeability. Polyoxymethylenes are only very slightly permeable to many gases and solvents. In particular, the very low permeability to propane; butane; aliphatic, aromatic, and halogenated hydrocarbons; alcohols; and esters should be emphasized. The permeability to hydrogen, CO_2, water, and methanol [20], [33], [40] is higher. A comparison with other plastics is given in [11].

Numerous uses have accordingly been found for polyoxymethylenes; for example, cigarette lighters, aerosol containers, and automotive parts in contact with fuels.

Fire Behavior. Polyoxymethylenes burn slowly with a pale blue flame, forming droplets. The external ignition temperature is 350 °C, and the self-ignition temperature is 360 °C.

According to Underwriters Laboratories polyoxymethylenes are classified in Class HB (UL Safety Standard 94). According to the U.S. Federal Motor Vehicle Safety Standard FMVSS 302 and DIN 75 200, the combustion rate does not reach the maximum permitted combustion rate of polyoxymethylenes at wall thicknesses of ≥ 1 mm, and accordingly use in automobiles is permitted. Further information on flammability can be found in [9].

A formulation with the classification V – O is not known.

Optical Properties. Polyoxymethylene moldings are translucent to opaque-white, depending on the thickness. The light transmission of 2 mm thick injection-molded plates is 50 %. The refractive index n_D^{20} is 1.48. The gloss depends mainly on the surface quality of the mold.

Electrical Properties. Polyoxymethylenes have good electrical insulation properties and a favorable dielectric behavior. This, together with the good mechanical properties, means that polyoxymethylenes are useful materials in electrical engineering.

The specific resistivity is $> 10^{15}\,\Omega\cdot$cm, and the surface resistance is $> 10^{13}\,\Omega$. The relative dielectric constant ε_r is 3.6 – 4.0. Since the loss factor tan δ is low, high-frequency heating and high-frequency welding are not possible [8].

The dielectric strength describes the behavior under short-acting stress produced by high voltages. Values of 600 – 700 kV/cm have been measured for polyoxymethylene sheets [8].

Polyoxymethylenes generally do not accumulate static charge. Antistatic and electrically conducting grades have been developed for specialty uses (e.g., Hostaform C 32 021 AST and Hostaform C 9021 ELS [8]).

3. Production

3.1. Polymerization of Formaldehyde and Trioxane

The polymerization of formaldehyde and trioxane proceeds according to an ionic mechanism. Whereas formaldehyde can be polymerized anionically as well as cationically, trioxane, being a cyclic acetal, can be polymerized only under cationic conditions. Polymerization can be performed in the gas phase, liquid phase, in dispersion, and in solution. Trioxane also polymerizes in the solid state.

In practice homopolymers are produced by anionic polymerization of formaldehyde, while copolymers are mainly produced by cationic copolymerization of trioxane with cyclic ethers or cyclic acetals. Processes for producing copolymers that use formaldehyde as starting material have also been described [41].

A review of the historical development of polyoxymethylenes is given in [42–44]. Comprehensive summaries of the homopolymerization and copolymerization of trioxane can be found in [45], [46].

For information on the most important process steps in the production of homopolymers and copolymers, together with numerous literature and patent references, see [47].

3.2. Homopolymers

Terminal-group-stabilized homopolymers of formaldehyde with adequate thermal stability were first developed by Du Pont [2], [47–49].

The production process can be divided into the following steps:

1) Oxidation of methanol to formaldehyde
2) Purification of formaldehyde
3) Polymerization
4) Terminal group capping
5) Homogenization of the polymer in the melt; addition of auxiliaries

The production of formaldehyde from methanol is described in → Formaldehyde, Chap. 4. The formox or silver catalyst process is usually used; see also [50].

High-purity formaldehyde is required for the polymerization step. Purification on an industrial scale can be performed, for example, by absorbing the reaction gases from the formox or silver catalyst process in cyclohexanol [51], [52]. Cyclohexanol hemiformal is formed from formaldehyde and cyclohexanol; water and other impurities are then removed by vacuum distillation. Vacuum distillation is followed by thermal cracking. The resultant formaldehyde gas can then be passed after further purification (e.g., removal of reactive impurities by partial polymerization [53], [54]) to the actual polymerization process [48]. Amines [54]; alkoxides [55]; ammonium and sulfonium salts [56]; metal carbonyls [57]; and phosphines, arsines, and stibines [58] are suitable initiators for anionic polymerization.

The high heat of polymerization (ca. 60 kJ/mol [10]) must be removed. Suspension polymerization has proved convenient in practice. Aliphatic or cycloaliphatic hydrocarbons are suitable suspension agents, for which the same purity requirements apply as for the monomer.

The purified formaldehyde is added to a reaction vessel containing the suspension agent (e.g., cyclohexane) and initiator (e.g., tri-*n*-butylamine). The polymerization can be carried out at room temperature and under normal or reduced pressure. The polymer is formed as a fine suspension.

In the next step the thermally unstable hemiacetal terminal groups of the polymer are converted into alkyl or acyl groups. In particular esterification with acetic anhydride has become industrially important [2]. The reaction can, for example, be carried out in the presence of small amounts of sodium acetate as catalyst in boiling acetic anhydride.

This conversion stage is followed by several purification steps. Finally, the polymer powder is melted in extruders with the addition of antioxidants, co-stabilizers, and other additives, and converted into granules.

3.3. Copolymers

The copolymer production process is divided into the following steps:

1) Oxidation of methanol to formaldehyde

2) Trimerization of formaldehyde to trioxane
3) Purification of trioxane
4) Copolymerization of trioxane with cyclic ethers or cyclic acetals
5) Removal of unstable terminal groups
6) Homogenization of the polymers in the melt; addition of auxiliaries

The production of formaldehyde and trioxane is described in → Formaldehyde, Chap. 4., → Formaldehyde, Chap. 6.2. The trioxane must satisfy strict purity requirements for the subsequent cationic polymerization. Compounds having a chain-transfer action such as water, methanol, methyl formate, formic acid, and methylal must be removed as far as possible. Chain-transfer reactions lower the attainable molecular mass [59]:

$$\sim\sim O-CH_2-O-CH_2^+ + H_2O \longrightarrow \sim\sim O-CH_2-O-CH_2-OH + H^+$$

Growing chain end Unstable chain end

The proton then initiates a new chain. The final purification to polymerization grade trioxane is generally carried out by crystallization [60],[61] or distillation.

A large number of compounds may be used as initiators for the cationic polymerization. Important catalysts include

1) Lewis acids, such as BF_3, and their complexes (e.g., $BF_3 \cdot Bu_2O$) [62]
2) Protic acids such as perchloric acid, perfluoroalkanesulfonic acids, as well as their alkyl esters or anhydrides [63]
3) Heteropoly acids such as tungstophosphoric acid [64]
4) Carbenium salts such as triphenylmethyl hexafluoroarsenate [65], oxocarbenium salts, onium compounds, and metal chelates

The desired molecular mass is regulated by chain-transfer agents such as methylal and butylal [66].

Suitable comonomers include cyclic ethers and cyclic acetals. In particular, ethylene oxide [67], 1,3-dioxolane, diethylene glycol formal [68], and 1,4-butanediol formal are used. The comonomers are generally used in amounts of 2 – 5 wt %.

The polymerization of trioxane can be performed in bulk, solution [69], or suspension [70]. For solution and suspension polymerization the same solvents may be used as for formaldehyde polymerization. Bulk polymerization is the preferred method in practice.

Bulk polymerization is generally carried out in self- cleaning kneaders or double-screw reactors [71], [72] at 70 – 115 °C. Polymerization is complete within a few minutes, and a very hard, crystalline polymer is formed. The heat of polymerization, ca. 23 kJ/mol [73], must be removed to prevent monomer losses due to evaporation of trioxane.

Other polymerization processes have been reported, including solid-state copolymerization of trioxane/1,3-dioxolane [74], homogeneous-phase polymerization [75], polymerization in an open vessel [76], and a drum process [77].

The copolymers contain the comonomer units in statistically distributed form. Only part of the terminal groups is formed from comonomer units. The unstable hemiacetal and formate terminal groups must be removed by heat treatment. In practice, both solution [78], [79] and melt [80], [81] processes have proved suitable. Solution processes are generally carried out at 150 – 200 °C, and melt processes at 170 – 250 °C. Amines or phosphines, for example, are added to deactivate the acid catalysts. The objective of all the processes is to completely remove unstable terminal groups and residual monomers from the polymer.

The solution processes are very efficient but have the disadvantage of high energy consumption. After thermal treatment, the product is precipitated in powder form by lowering the temperature. The adhering residual solvent is removed mechanically and thermally. The dry powder, together with stabilizers and other, optional additives, is melted and homogenized in extruders and processed into granules.

Although the melt processes do not achieve the efficiency of the solution processes, they require less energy. After addition of stabilizers the melt can be processed directly into granules.

A large number of patent applications have been filed by various companies for the polymerization as well as the work-up process, e.g., [82–90]. The Asahi process, in which formaldehyde is prepared from methylal instead of by direct oxidation of methanol [91] is described in [92].

The mechanical properties are measured by the conventional methods used for polymer materials. Test methods and values can be obtained from the brochures of the raw material manufacturers [7], [8], [9]. See also Table 1 as well as [11] and Chapter 5.

3.4. Additives

Stabilizers. Polyoxymethylenes can be decomposed by the action of high temperatures, oxygen, shear forces, and/or acidic substances. These influences are especially important in the processing of the products in the melt. Further influences may also occur when using the finished parts.

In the case of terminal-group-stabilized polyoxymethylenes, decomposition begins by chain scission. Above the melting point, chain scission is followed by depolymerization with elimination of formaldehyde. For copolymers, this involves elimination of formaldehyde units until the next stabilizing comonomer unit is reached. In the case of homopolymers complete decomposition of the polymer chain occurs. Formic acid, which can initiate an acidolytic chain scission, is formed as byproduct of the decomposition by oxidation of formaldehyde or by disproportionation. Acidic catalyst residues can also produce chain scission [93], [94]; for further details, see [95].

Only multicomponent systems are used to stabilize polyoxymethylenes. Phenolic compounds in particular have proved suitable against thermal-oxidative decomposition (antioxidants). Alkaline-earth metal salts and nitrogen-containing compounds, among others, are used to prevent acidolytic decomposition (acid trapping agents). Formaldehyde trapping agents can be used to further improve the stability.

The following compounds are used as antioxidants:

1) 4,4′-Butylidenebis(6-*tert*-butyl-3-methylphenol) [96]
2) Triethylene glycol bis[3-(3,5-di-*tert*-butyl-4-hydroxyphenyl)propionate] [97]
3) 1,6-Hexamethylene bis[3-(3,5-di-*tert*-butyl-4-hydroxyphenyl)propionate] [98]
4) 2,2′-Methylenebis(4-methyl-6-*tert*-butylphenol) [99]

The following compounds, among others, are used as acid trapping agents:

1) Dicyanodiamide, melamine [100]
2) Calcium carbonate, stearate, or montanate [101]
3) Polycondensates of formaldehyde, isophthalamide and 2-imidazolidinone [102]
4) Polyamides [103]
5) Triazine derivatives [104]

The nitrogen-containing compounds also act as formaldehyde trapping agents. Other formaldehyde trapping agents include polyhydroxy compounds such as poly(vinyl alcohol) [105].

The aforementioned classes of stabilizers can be combined in numerous ways. Optimum utilization of the possible synergistic effects can be achieved only by carefully matching all the components.

Damage caused to polyoxymethylenes by UV radiation and weathering is described in Chapter 2. The addition of organic light stabilizers and special pigments has proved useful in improving the resistance to UV radiation and weathering.

Compounds from the following classes are used as light stabilizers [106–108]: benzotriazoles, benzophenones, benzoates, oxanilides, and sterically hindered amines (HALS stabilizers). Synergistic effects can also be achieved here by suitable combinations of stabilizers of different classes of compound, in particular combination with HALS stabilizers. Titanium dioxide and, in particular, carbon black are suitable pigments.

All the above-mentioned stabilizers and pigments are incorporated into the polymer melt.

Nucleation Agents. The following are suitable as nucleation agents for polyoxymethylenes:

1) Branched or cross-linked polyoxymethylenes [109]
2) Trioxane block copolymers [110]
3) Special melamine – formaldehyde resins [104], [111]
4) Boron nitride [112]
5) Talcum [113]
6) Antimony trioxide [114]

Even in amounts of less than 1 wt % these compounds accelerate crystallization of the melt. Nucleated polyoxymethylene molding compounds often allow earlier removal from the mold, and thus a reduction of the cycle time, in injection molding processing. The resulting spherulite structure is finer and more uniform, which results in increased strength and toughness [11], [115].

Release Agents. The addition of release agents may be advantageous for the production of complicated molded parts. Suitable additives include, for example, polyethylene waxes, montanate esters and salts, *N,N′*(bis-stearoyl)ethylenediamine, fatty alcohols, and fatty acid salts.

3.5. Impact-Resistant Types

Polyoxymethylenes are materials that already have a high basic toughness, without the need for further modification. However, the toughness can be improved further by impact resistance modification with elastomers.

The first commercial impact-resistance-modified polyoxymethylenes were introduced in 1982, and in the meantime several manufacturers have developed products of this type. The various products differ according to their flow behavior and degree of impact resistance modification. Compared with the unmodified products the impact-resistant types have slightly reduced rigidity, hardness, and thermal dimensional stability, though this is compensated by a substantially higher energy absorption under impact and tensile stress.

The conventional types are based on blends of polyoxymethylene and thermoplastic polyurethane elastomers [116–120]. Products in which cross-linked rubbers having a particularly low glass transition temperature and a fixed particle size are used for the impact strength modification have been developed more recently [121].

Impact-resistant polyoxymethylenes are produced by mixing the components in the melt, using single-screw or double-screw extruders. Processing into moldings is performed by the same methods as are used for unmodified polyoxymethylenes. Whereas polyurethane-modified products have, depending on the extrusion and processing conditions, a variable morphology of the impact-resistant component (size, shape, and distribution of the particles), the rubber-modified products have a uniform particle morphology. In addition to the particularly high low-temperature impact resistance, the new products have further advantages, including the extremely good weld line strength and the absolute absence of delamination during processing [122].

4. Economic Aspects

Table 2 lists producers of polyoxymethylenes, together with their capacities and trade names [123].

A demand for polyoxymethylenes of about 390×10^3 t/a can be compared with a nominal capacity of ca. 440×10^3 t/a for 1990. Homopolymers accounted for about 30 % of the total market.

5. Quality Specifications and Analysis

Residual monomer content, molecular mass, thermal stability, and mechanical properties are important quality criteria for polyoxymethylenes. For the copolymers there is in addition the comonomer content.

The so-called lutidine or acetylacetone method [124], which is specified by the BGA to evaluate the formaldehyde content of plastic molded parts that come into contact with foodstuffs [125], [126], is generally used to measure the content of extractable formaldehyde in granules or in molded parts. Formaldehyde can also be determined with chromotropic acid or by titration with sodium sulfite solution. The tests are normally performed on aqueous extracts of the materials.

The molecular mass is monitored in practice via the melt viscosity (melt flow index, MFI). Various $\eta - M$ relationships are given in the literature for molecular mass determination by means of the solution viscosity. Usual solvents are dimethylformamide and hexafluoroacetone – water mixtures [127], [128].

The weight loss that occurs in the melt at a specific temperature over a certain time and in

Table 2. Producers, capacities, and trade names for polyoxymethylenes in 1990

Region	Producer	Capacity, 10^3 t/a	Product type	Trade name
Western Europe	Hoechst (Ticona)	50	copolymer	Hostaform
	BASF (Ultraform)	30	copolymer	Ultraform
	Du Pont	55	homopolymer	Delrin
United States	Hoechst Celanese	68	copolymer	Celcon
	BASF (Ultraform)	16	copolymer	Ultraform
	Du Pont	55	homopolymer	Delrin
Japan	Polyplastics	75	copolymer	Duracon
	Mitsubishi Gas	20	copolymer	Iupital
	Asahi Chemical	15	copolymer	Tenac-C
	Asahi Chemical	20	homopolymer	Tenac
South Korea	Korea Engineering Plastics	10	copolymer	Kepital
	Lucky-Goldstar International	10	copolymer	Lucel
Former Soviet Union		10	copolymer	
Poland		10	copolymer	
China		10	copolymer	

a defined atmosphere (air, nitrogen) can be used to evaluate the thermal stability [129], [130].

NMR spectroscopy is suitable for monitoring the comonomer content in the end product [131]. When the comonomer is ethylene oxide, the frequency of occurrence of mono-, di-, and triethylene glycol units can also be determined with this method.

IR spectroscopic methods have been described for the determination of terminal groups in acylated homopolymers [132], [133].

Conventional analytical techniques such as UV, IR, NMR, and ashing, optionally in combination with suitable separation methods, can be used for the qualitative and quantitative determination of additives, fillers, and reinforcement agents in the end product.

6. Product Range and Processing

Polyoxymethylenes are produced in various degrees of polymerization. The individual types differ as regards their flow behavior (melt index) and in the choice and concentration of additives and colored pigments. Reinforced types, types with improved slide and wear behavior and impact resistance, and products with anti-UV additives are also available [7–9]. The usual supply form is granules of about 3 mm grain size.

Polyoxymethylenes can be processed with all conventional machinery for thermoplastics, such as injection molding machines, extruders, injection and extrusion blow units, as well as by compression. Temperatures between 190 °C and 210 °C are generally recommended for processing. The temperature in the melt should not exceed ca. 230 °C [8], [9].

Posttreatment of injection moldings is generally not necessary. Machining, in which parts are fashioned from the bulk plastic (semifinished product), has, however, become important. Machining is economically practicable up to medium-size production runs, particularly if computer numerical control techniques are used.

In the "outsert" technique a wide range of polyoxymethylene functional components are bonded to a metal blank in a single step in an injection molding process [134]. Such components include guiding and sliding el-ements, bearings, axles, columns, snap connections, and spring components. They are used, for example, in compact disc drives and in cassette and video recorders.

Further information on processing is given in [5], [6], [11].

7. Uses

Polyoxymethylenes have a balanced combination of favorable properties for technical use, such as hardness, rigidity, toughness, spring elasticity, and resistance to fuels. The range of uses has been broadened further by a large number of modified special grades. Polyoxymethy-

lenes are used in almost all sectors of industry and technology. The main areas of polyoxymethylene consumption in Western Europe are the automotive sector (39 %), domestic, household and office equipment (14.7 %), and in electrical engineering and the electronics industry (10.8 %) [122]. Further uses include mechanical engineering, machine construction, sanitary uses, medical technology, toys, and cosmetics.

Utility articles and parts fabricated from polyacetals vary widely as regards their range of application and are accordingly so numerous that it is only possible to list the most important uses. Some examples of applications are as follows [5], [6]:

Automotive sector: parts for gasoline pumps, gasoline gauges, carburettor floats, brake boosters, chain wheels, clips, tank closures, water separators, lock housings, fan blades, and gas bubble separators.

Electrical industry: gearwheels, bearings, springs, radial cams, lifters, tappets, axles, pinions, gears, transmissions, worm gears, computer keyboards, water boilers, and parts for leakage current protection circuits.

Mechanical engineering: valves, pistons, sliding bearings, and ball-type nipple cages.

Sanitary sector: handsprays, fittings, control flaps for WC flush units, and functional parts in water and gas meters.

8. Recycling

On account of the high commercial value of polyoxymethylenes a functioning material recycling system for polyoxymethylenes already exists. Many plastics processors use material residues directly themselves and contribute substantially by means of this internal recycling process to reduce waste. Polyoxymethylene waste is also processed by compounders into regranulate. This regranulate is, however, of poorer quality than the original material.

Hoechst has developed a closed-loop recycling process, in which polyoxymethylene waste is decomposed into the monomer starting materials. The decomposition products are purified and repolymerized [135]. The cycle can be repeated as often as desired without affecting the quality.

9. Toxicology and Occupational Health

Polyoxymethylenes are odorless and tasteless. Feeding tests have not produced any positive clinical results. Skin compatibility investigations showed that polyoxymethylene does not irritate intact skin [40], [136].

There are no objections to the use of polyoxymethylene moldings that come into contact with foods, as long as the regulations of the licensing authorities (BGA [125], FDA [137], [138]) are satisfied and the parts are suitable for the intended use. The regulations cover the nature and amount of monomers, comonomers, catalysts, deactivators, regulators, stabilizers, lubricants, and fillers, and specified contents of fluorine, boron, zinc, and amines must not be exceeded. A maximum permissible upper limit is specified for formaldehyde traces that can be extracted from finished parts.

The maximum permitted workplace concentration for formaldehyde should also be observed when processing polyoxymethylenes. In Germany the MAK is 0.5 ppm (0.6 mg/m^3), and in the United States the TLV-TWA is 1.0 ppm (1.2 mg/m^3) [as of 1991]. These limits may be exceeded for a short period only if certain preconditions are satisfied. For further details see → Formaldehyde, Chap. 10..

10. References

1. W. Kern in H. Staudinger (ed.): *Die hochmolekularen organischen Verbindungen Kautschuk und Cellulose,* Springer Verlag, Berlin 1932.
2. C. E. Schweitzer, R. N. MacDonald, J. O. Punderson, *J. Appl. Polym. Sci.* **1** (1959) no. 2, 158 – 163.
3. Celanese, US 3 027 352, 1960 (C. T. Walling, F. Brown, K. W. Bartz).
4. Hoechst, DE 1 124 703, 1959 (K. Küllmar, E. Fischer, K. Weissermel).
5. H.-D. Sabel, H. Schlaf, U. Struth, P. Unger: "Polyacetale," in: *Kunststoff-Handbuch,* Chap. 4: "Technische Thermoplaste," Hanser Verlag, München, in preparation.
6. H. Schmidt, U. van Spankeren: "Polyacetale," in: *Kunststoff-Handbuch,* vol. XI, Hanser Verlag, München 1971.

7. Du Pont, company in formation, *Delrin, Acetal Resins,* no. E-96 324 1288, 1988.
8. Hoechst, company in formation Hoechst Plastics, *Hostaform,* July 1987.
9. BASF, company in formation, no. B 563 d, *Ultraform-Polyoxymethylen,* 1990.
10. G. Sextro: "Physical Constants of Poly(oxymethylene)," in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* 3rd ed., vol. 1, J. Wiley & Sons, New York 1989, V/p. 87 – V/p. 99.
11. H. Domininghaus: *Die Kunststoffe und ihre Eigenschaften,* 3rd. ed., VDI-Verlag, Düsseldorf 1988, pp. 307 – 336.
12. S. Okamura et al., *J. Polym. Sci. Part C* **4** (1963) 827 – 838.
13. J. C. Bevington, H. May: Aldehyde Polymers, in *Encycl. Polym. Sci. Technol.* **1** (1964) 609 – 628.
14. H. Wilski, *Makromol. Chem.* **150** (1971) 209.
15. F. S. Dainton, D. M. Evans, F. E. Hoare, T. P. Melia, *Polymer* **3** (1962) 263.
16. H. Suzuki, B. Wunderlich, *J. Polym. Sci. Polym. Phys. Ed.* **23** (1985) 1671.
17. F. S. Dainton, K. J. Ivin, D. A. G. Walmsley, *Trans. Faraday Soc.* **55** (1959) 61.
18. F. Salaris, A. Turturro, U. Bianchi, E. Martuscelli, *Polymer* **19** (1978) 1163.
19. H. Wilski, *Kolloid-Z. Z. Polym.* **248** (1971) 867.
20. VDI/VDE: *Werkstoffe der Feinwerktechnik-Polyacetal-Formstoffe,* Richtlinie 2477, Beuth Vlg., Berlin 1974.
21. H. Breuer, G. Dupp, J. Schmitz, R. Tüllmann, *Kunststoffe* **80** (1990) 1289 – 1294.
22. T. Uchida, H. Todokoro, *J. Polym. Sci. Polym. Phys. Ed.* **5** (1967) 63.
23. U. Gaur, B. Wunderlich, *J. Phys. Chem. Ref. Data* **4** (1981) no. 10, 1001.
24. B. E. Read, G. Williams, *Polymer* **2** (1961) 239.
25. M. Dröscher, K. Hertwig, H. Reimann, G. Wegner, *Makromol. Chem.* **177** (1976) 1696 – 1706.
26. Hoechst, company information, no. HBKB 0395 C.3.3, *Gestalten von Formteilen aus technischen Kunststoffen.*
27. Hoechst, company information, no. HBKP 0395/C.3.4 D, *Richtlinien für das Gestalten von Formteilen aus technischen Kunststoffen.* (English version HBKP 0395/C.3.4 E).
28. E. Wolters, H. Racké, *Kunststoffe* **63** (1973) 608 –612.
29. F. Gügümus in R. Gächter, R. Müller (eds): *Taschenbuch der Kunststoffe – Additive,* Hanser Verlag, Munich 1989.
30. R. G. Alsup, J. O. Punderson, G. F. Leverett, *J. Appl. Polym. Sci.* **1** (1959) 185.
31. BASF, company in formation, no. TI-PA-28 d/82 243, *BASF Kunststoffe, Verhalten von Ultramid, Ultraform und Ultradur gegen Chemikalien,* Oct. 1980.
32. H. Schmidt, *Br. Plast.* **38** (1965) no. 9, 546 – 551.
33. K. Kohlhepp, R. Orth, *Kunststoffe* **77** (1987) 686.
34. R. Vesely, M. Kalenda, *Kunststoffe* **59** (1969) 107 – 110.
35. J.-L. Gardette, H.-D. Sabel, J. Lemaire, *Angew. Makromol. Chem.,* **188** (1991) 113 – 128.
36. N. Grassie, R. S. Roche, *Makromol. Chem.* **112** (1968) 34 – 39.
37. H. Schmidt, E. Wolters, *Kunststoffe* **61** (1971) no. 4, 261 – 265.
38. W. Kern, *Chem. Ztg.* **91** (1967) 255 – 262.
39. E. Wolters, S. Rösinger, *Kunststoffe* **63** (1973) 605 –608.
40. Du Pont, company information, *Delrin-Polyacetal,* no. A 84 029-2.
41. Du Pont, US 3 183 211, 1961 (K. C. Brinker, C. E. Schweitzer).
42. W. Kern, H. Cherdron, V. Jaacks, *Angew. Chem.* **73** (1961) no. 6, 177 – 224.
43. W. Kern, V. Jaacks, *Kolloid-Z. Z. Polym.* **216 – 217** (1967) 286 – 298.
44. W. Kern, H. Deibig, A. Giefer, V. Jaacks, *Pure Appl. Chem.* **12** (1966) 371 – 386.
45. K. Weissermel et al., *Angew. Chem.* **79** (1967) no. 11, 512 – 520; *Angew. Chem. Int. Ed. Engl.* **6** (1967) no. 6, 526 – 533.
46. K. Weissermel, E. Fischer, K. Gutweiler, H. D. Hermann, *Kunststoffe* **54** (1964) no. 7, 410 – 415.
47. G. E. Haddeland in: *Process Economics Program,* no. 69: "Acetal Resins," Stanford Res. Inst. Menlo Park, California, June 1971.
48. Du Pont, US 2 768 994, 1954 (R. N. MacDonald).
49. W. H. Linton, H. H. Goodman, *J. Appl. Polym. Sci.* **1** (1959) no. 2, 179 – 184.
50. J. F. Walker: *Formaldehyde,* 3rd. ed., reprint of orig. ed. 1964, Krieger Publ. Comp., Malabar, Florida.
51. Du Pont, US 4 539 387, 1984 (L. M. Blair).
52. Du Pont, US 2 943 701, 1957 (D. L. Funck).
53. Du Pont, DE 1 000 152, 1953 (R. N. MacDonald).
54. Du Pont, US 2 841 570, 1953 (R. N. MacDonald).

55. Du Pont, US 2 848 437, 1954 (W. P. Langsdorf, G. S. Stamatoff).
56. Du Pont, GB 793 673, 1956 (H. H. Goodman, L. T. Shelwood).
57. Du Pont, US 2 734 889, 1954 (F. C. Starr).
58. Du Pont, FR 1 082 519, 1954 (R. N. MacDonald).
59. G. L. Collins, R. K. Greene, F. M. Berardinelli, W. H. Ray, *J. Polym. Sci.* **19** (1981) 1597 – 1607.
60. Mitsubishi Gas Chem. Ind., DE-OS 2 855 710, 1978 (A. Sugio et al.).
61. Hoechst, EP 162 252, 1985 (S. Rittner, K.-H. Burg, H. Schlaf).
62. Celanese, US 2 989 507, 1957 (D. E. Hudgin, F. M. Berardinelli).
63. Degussa, DE 1 152 818, 1960 (W. Kern, A. Wildenau, V. Jaacks).
64. Polyplastics, EP 325 052, 1988 (K. Yamamoto, H. Sano, S. Chino).
65. British Ind. Plastics, GB 1 054 205, 1964 (H. May, B. J. Kendall-Smith, J. A. Dodd).
66. Degussa, DE 1 194 145, 1959 (W. Kern, V. Jaacks).
67. Celanese, US 3 027 352, 1959 (C. T. Walling, F. Brown, K. W. Bartz).
68. Hoechst, US 3 194 788, 1959 (K. Küllmar, E. Fischer, K. Weissermel).
69. Celanese, US 2 989 511, 1958 (A. W. Schnizer).
70. St. Penczek, J. Fejgin, W. Sadowska, M. Tomaszewicz, *Makromol. Chem.* **116** (1968) 203.
71. Hoechst, DE 1 161 421, 1961 (K. Küllmar, H. Schmidt, G. Roos, E. Fischer).
72. Celanese Corp., US 4 105 637, 1974 (M. Semanchik, D. M. Braunstein).
73. H. Sawada: *Thermodynamics of Polymerization,* Marcel Dekker, New York 1976.
74. Hoechst, DE 2 263 300, 1972 (G. Sextro, K.-H. Burg, K. Gutweiler).
75. BASF, EP 80 656, 1982 (J. Sadlowski, M. Walter, K. Hinselmann).
76. BASF, DE 2 003 270, 1970 (E. Bäder, H. Amann).
77. BASF, DE 1 720 300, 1967 (H. Heel, H. Klingenspohr, W. Klee, G. Göschel).
78. Celanese, DE 1 445 294, 1961 (F. M. Berardinelli, J. E. Wall, E. T. Smith, G. J. Fisher).
79. Hoechst, DE 2 452 736, 1974 (G. Sextro, K.-H. Burg, H. J. Leugering, H. Schlaf, A. Heller).
80. Celanese, DE 2 319 972, 1973 (M. D. Golder).
81. BASF, DE 1 963 402, 1969 (B. Sander, H. Sperber).
82. Asahi, JP 59 159 812, 1983. *Chem. Abstr.* **102** (1985) 204 475 t.
83. Asahi, JP 57 139 113, 1981. *Chem. Abstr.* **98** (1983) 35 186 r.
84. Asahi, JP 63 162 686, 1986 (H. Hata, J. Masamoto). *Chem. Abstr.* **109** (1988) 231 712 u.
85. Mitsubishi Gas, DE 3 617 754, 1986 (T. Unemura, I. Masumoto, Y. Iha).
86. Mitsubishi Gas, DE 3 738 632, 1987 (T. Suzumori, I. Masumoto, M. Nakagawa, Y. Iha).
87. UBE, GB 2 132 213, 1983 (K. Tsunemie, T. Suzuki, Y. Yamamoto).
88. UBE, JP 01 108 211, 1987 (K. Suzuki, K. Hisatomi, K. Kamimura). *Chem. Abstr.* **111** (1989) 215 164 c.
89. UBE, JP 01 121 313, 1987 (K. Suzuki, K. Tokunaga, N. Seki). *Chem. Abstr.* **111** (1989) 215 147 z.
90. UBE, JP 61 151 218, 1984 (K. Hisatomi, H. Ninomiya, Y. Yamamoto). *Chem. Abstr.* **105** (1986) 227 877 b.
91. J. Masamoto et al., *Macromol. Chem. Macromol. Symp.* **42/43** (1991) 409 – 423.
92. W. S. Fong in: *Process Economics Program* no. 89-2-3: "Acetal Copolymers by an Asahi Chemical Process," SRI Int., Menlo Park, California, June 1990.
93. W. Kern, H. Cherdron, *Makromol. Chem.* **40** (1961) 101.
94. W. Kern, H. Cherdron, *Makromol. Chem.* **40** (1961) 177.
95. G. Opitz, *Plaste Kautsch.* **28** (1981) no. 1, 15 – 19.
96. Du Pont, DE 1 082 404, 1957 (R. D. Kralovec, P. N. Richardson).
97. Ciba Geigy, DE 2 133 374, 1971 (E. K. Kleiner, J. D. Spivack).
98. Ciba Geigy, DE 2 203 836, 1971 (H. Müller).
99. Du Pont, GB 748 856, 1953 (M. A. Kubico, R. N. MacDonald, R. L. Stearns, F. A. Wolff).
100. Celanese, US 3 313 767, 1967 (F. M. Berardinelli, R. J. Kray, T. J. Dolce).
101. Hoechst, US 3 743 614, 1973 (E. Wolters, G. Roos).
102. BASF, DE 1 669 692, 1967 (E. Schwartz, F. Schmidt, B. Sander, H. Sperber, H. Wilhelm, E. Ricker).
103. Du Pont, DE 1 066 739, 1957 (R. G. Alsup, P. E. Lindvig).
104. Degussa, EP 0 202 530, 1986 (H. Amann, G. Morlock, E. Schöla).

105. Du Pont, EP 242 176, 1987 (N. E. West).
106. Du Pont, US 3 043 709, 1958 (L. E. Amborski).
107. Du Pont, GB 921 808, 1961 (D. B. Carlson).
108. F. R. Stohler, K. Berger, *Angew. Makromol. Chem.* **176/177** (1990) 323 – 332.
109. Hoechst, DE 2 101 817, 1971 (R. Kern, H. Schmidt, K.-H. Burg, E. Wolters).
110. Hoechst, DE 2 142 091, 1973 (G. Sextro, K.-H. Burg).
111. G. Morlock: *6. Internationales Makromolekulares Symposium,* Interlaken (Switzerland), 8th/9th June 1978.
112. Hoechst, US 3 767 610, 1971 (K.-H. Burg, L. Brinkmann, E. Wolters).
113. Bayer, DE 1 247 645, 1964 (W. Hechelhammer).
114. Hoechst, US 3 783 154, 1971 (H.-H. Grossmann, K.-H. Burg, E. Wolters).
115. H. J. Leugering, K.-H. Burg, H. Cherdron, *Makromol. Chem. Suppl.* **1** (1975) 621 – 636.
116. F. Kloos, E. Wolters, *Kunststoffe* **75** (1985) 735 – 739.
117. Hoechst, EP 115 847, 1984 (R. Reske, E. Wolters).
118. Celanese, DE 2 051 028, 1970 (F. McAndrew).
119. Bayer, DE 1 193 240, 1962 (K. Wagner, H. Röhr).
120. Du Pont, EP 0 116 456, 1984 (E. A. Flexman, Jr.).
121. Hoechst, EP 0 156 285, 1985 (K.-H. Burg, H. Cherdron, F. Kloos, H. Schlaf).
122. H.-D. Sabel, U. Struth, *Kunststoffe* **80** (1990) no. 10, 1118 – 1122.
123. A. S. Wood, *Mod. Plast.* **67** (1990) no. 6, 56.
124. T. Nash, *Biochem. J.* **55** (1953) 416 – 421.
125. R. Franck, H. Wieczorek: *Kunststoffe im Lebensmittelverkehr,* Teil A, XXXIII: "Acetalharze", Stand: 1. Sept. 1987, 38. Lieferung, 1989, Carl Heymanns Verlag, Köln.
126. R. Franck: *Kunststoffe im Lebensmittelverkehr,* Teil B II, XVIII: "Bestimmung von Formaldehyd in Kunststoffgefäßen aus Melaminharz", Stand 1. Sept. 1961, 33. Lieferung, August 1984, Carl Heymanns Verlag, Köln.
127. L. Hoehr et al., *Makromol. Chem.* **103** (1967) 279.
128. I. M. Bel'gorskii, N. S. Enikolopyan, L. S. Sakhonenko, *Vysokomol. Soedin.* **4** (1962) 1179; *Polym. Sci. (USSR) (Engl. Transl.)* **4** (1963) 367.
129. Du Pont, US 2 993 025, 1958 (R. G. Alsup, P. E. Lindvig).
130. Celanese, FR 1 221 148, 1959 (T. Walling, F. Brown, K. W. Bartz, G. W. Polly).
131. D. Fleischer, R. C. Schulz, *Makromol. Chem.* **176** (1975) 677 – 689.
132. H. Frank, V. Jaacks, W. Kern, *Makromol. Chem.* **114** (1968) 92 – 112.
133. T. A. Koch, P. E. Linvig, *J. Appl. Polym. Sci.* **1** (1959) 164 – 168.
134. Hoechst, company information, no. C.3.5, *Outsert-Technik mit Hostaform.*
135. Hoechst, DE 4 035 495, 1990 (K.-F. Mück, G. Reuschel, D. Fleischer).
136. R. B. Akin: *Acetal Resins,* Reinhold, New York 1962.
137. Federal Register, Code of Federal Regulations, Title 21, Part 177, Section 2470, U.S. Government Printing Office, Washington 1989.
138. Federal Register, Code of Federal Regulations, Title 21, Part 177, Section 2480, U.S. Government Printing Office, Washington 1989.

Poly(Phenylene Ethers) → Poly(Phenylene Oxides)

Poly(Phenylene Oxides)

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated inPlastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes

Jan Bussink, GE Plastics BV, Bergen op Zoom, The Netherlands

Hendrik T. van de Grampel, GE Plastics BV, Bergen op Zoom, The Netherlands

1. Introduction 1
2. Production 2
3. PPE Blends 3
3.1. PPE – PS Blends 3
3.2. PA – PPE Blends 6
3.3. PPE – Polyolefin Blends 9
4. Processing 9
5. Uses 9
6. References 10

Acknomledgement. The authors thank Veronique Hopmans for her contributions.

1. Introduction

Poly(phenylene oxides) [*9041-80-9*], also known as poly(phenylene ethers) or poly-oxyphenylenes, refer to a class of polymers made by the oxidative polymerization of substituted phenols. A large variety of polymers have been investigated based on differently substituted phenols (Table 1). However, only poly(2,6-dimethyl-1,4-phenylene ether) [*25134-01-4*] (PPE) has significant utility, albeit in blends with other polymers such as polystyrene or polyamide:

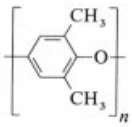

PPE

History. The discovery of poly(2,6-dimethyl-1,4-phenylene ether) in 1956 by Hay (General Electric) [1] marked the introduction of the catalyzed oxidative coupling of phenols as a means of obtaining engineering resins. This poly(phenylene oxide) made from 2,6-dimethylphenol can be regarded as amorphous for all practical purposes [2], and possesses a high glass transition temperature (T_g 211 °C) in addition to an extremely high melt viscosity, which requires high processing temperatures that lead to side reactions [3], [4]. Only through the exceptional miscibility of PPE and polystyrene (PS) [5] could acceptable processing temperatures and melt viscosities be achieved. This development led to the marketing of the first commercial poly(phenylene oxide) in 1964 – Noryl, a blend of PPE and impact modified PS (HIPS). These PPE – HIPS blends are unique because their ultimate use temperature can be easily adapted to application requirements by adjustment of the PPE – HIPS ratio. GE Plastics was the sole supplier of modified PPE until 1983 when its original patents expired; other companies (e.g., BASF with Ultranyl and Hüls with Vestoran) have now introduced their own PPE grades. A blend of HIPS and a PPE copolymer with a similar property profile to poly(2,6-dimethyl-1,4-phenylene ether) was introduced in the early 1970s under the trade name Prevex (Borg – Warner and Mitsubishi Chemicals):

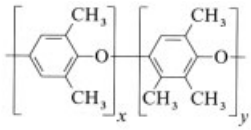

Over 200 000 t/a of modified PPE are sold worldwide, mainly for applications in the automotive, electronics, and electrical industries.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_605

Table 1. 2,6-Substituted aromatic polyethers formed by oxidative coupling of di-*ortho*-substituted phenols

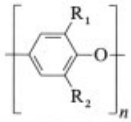

R_1	R_2	Polymer	Diphenoquinone compounds	CAS registry no.	T_g, °C	T_m, °C
H	H	branched	branched	[*25667-40-7*]	82	298
H	CH_3	branched	branched			
CH_3	CH_3	X		[*25134-01-4*]	211	268
CH_3	C_2H_5	X		[*26635-16-5*]		
CH_3	$CH(CH_3)_2$	X		[*31985-12-3*]	144	
CH_3	$C(CH_3)_3$		X	[*35916-68-8*]		
CH_3	$CH_2C_6H_5$	X		[*26545-37-9*]	99	
CH_3	C_6H_5			[*25805-39-4*]	169	
CH_3	OCH_3	X				
CH_3	Cl	X		[*26498-58-8*]		
C_2H_5	C_2H_5	X			120	
$CH(CH_3)_2$	$CH(CH_3)_2$		X			
$C(CH_3)_3$	$C(CH_3)_3$	X				
Cl	Cl	X		[*26023-26-7*]	228	269
OCH_3	OCH_3		X	[*25667-13-4*]	167	
C_6H_5	C_6H_5	X		[*24938-68-9*]	225	460 (decomp.)
C_6H_5	3-$CH_3C_6H_4$	X		[*79569-12-3*]	219	
C_6H_5	4-$CH_3C_6H_4$	X		[*79569-09-8*]	218	
C_6H_5	4-$(CH_3)_3C_6H_4$	X		[*79569-10-1*]	240	
C_6H_5	2-naphthyl	X		[*79569-11-2*]	234	

2. Production

The synthesis of PPE from 2,6-dimethylphenol, using an amine complex of a basic cupric salt as a homogeneous oxidation catalyst, follows a complicated polymerization mechanism which has not been completely resolved [6], [7].

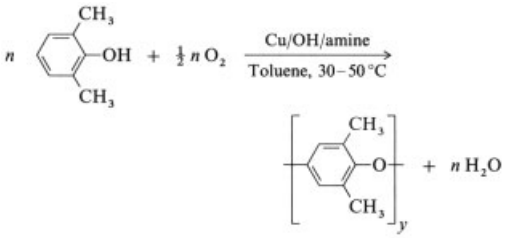

The main byproduct of the polymerization is 3,3′,5,5′-tetramethyldiphenoquinone (TMDPQ):

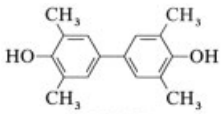

TMDPQ

The oxidative coupling chemistry is very versatile because different di-*ortho*-substituted phenols can be used for the formation of 2,6-substituted aromatic (co)polyethers (Table 1). However, only PPE is of commercial importance.

Commercial PPE production can be performed batchwise or continuously. Different reactor designs are possible depending upon the choice of the solvent and introduction of oxygen (pure oxygen or air – oxygen mixtures). A gas – liquid reactor is used to avoid low reaction rates associated with diffusion of air into the liquid. Reaction temperatures are typically 20 – 65 °C. At higher temperatures, formation of TMDPQ increases and should be avoided. The solubility and viscosity of the final polymer solution depend on the solvent, temperature, polymer concentration, and its molecular mass. Commercial PPE is produced with a weight-average molecular mass of 32 000 – 40 000.

The following stages can be discerned (Fig. 1):

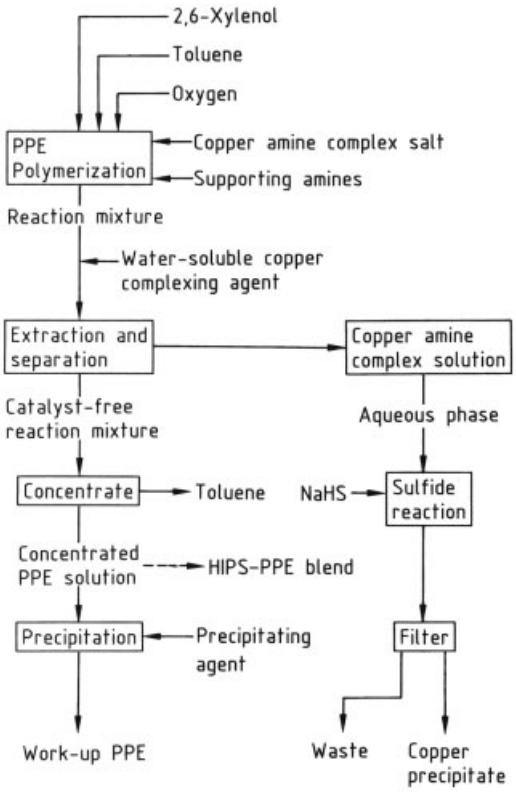

Figure 1. Commercial production of PPE

1) Preparation of the catalyst solution
2) Preparation of reaction mixture
3) Polymerization using oxygen or oxygen – air mixtures
4) Termination of polymerization at a certain degree of polymerization
5) Removal of copper catalysts from the reaction mixture using a copper complex as complexing agent and liquid – liquid extraction
6) Concentration of PPE solution
7) Precipitation of PPE in a nonsolvent
8) Repeated washing and drying (purification of resin)

3. PPE Blends

3.1. PPE – PS Blends

A processible, commercially attractive material is obtained by blending PPE with PS (Fig. 2). The unique, complete miscibility of this blend presumably results from interaction between the π-electron donor PS and the electron-deficient methyl groups of PPE [8]. The T_g of the blend can be varied almost linearly between 100 and 215 °C over the complete composition range of PS and PPE (Fig. 3). Table 2 shows some important properties of PPE – HIPS blends.

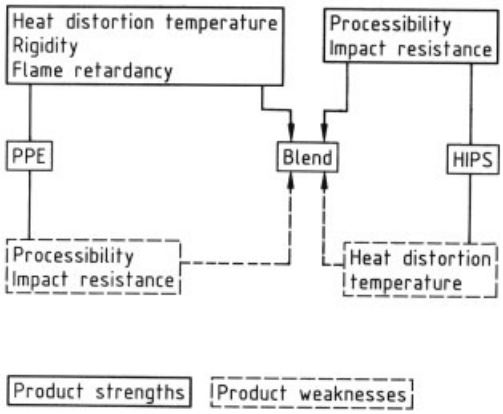

Figure 2. Blending criteria for PPE – PS

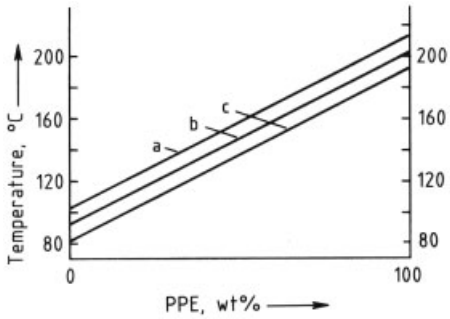

Figure 3. Thermal properties of PPE – HIPS blends
a) Glass transition temperature, T_g; b) Vicat B softening temperature; c) Heat distortion temperature

Commercial blends such as Noryl (GE Plastics) or Luranyl (BASF) make use of PS impact-modified with polybutadiene (high-impact polystyrene, HIPS). Impact modification of these blends comprising a ductile resin (PPE) and a brittle one (PS) poses an interesting problem because the two components require different rubber particle sizes for optimal impact modification (Fig. 4). Diluting HIPS with PPE yields a more ductile matrix in which smaller rubber particles (diameter 0.5 – 2 µm) are more effective in impact modification than the larger particles (2 – 12 µm) [9], [10]. The larger particles

Table 2. Selected properties of PPE – HIPS blends (Noryl, GE Plastics)

Property	Standards		Unreinforced	Unreinforced	Glass-reinforced	Glass-reinforced	Structural
	ASTM[f]	DIN[f]	blend[a]	flame retardant[b]	blend[c]	flame retardant[d]	foam[e]
Mechanical							
Tensile strength (break), MPa	D 638	53 455	50	45	90	77	22
Tensile elongation (break), %	D 638	53 455	50	50	3	2	8
Tensile modulus, MPa	D 638	53 457	2500	2400	6500	7300	1600
Flexural strength (yield), MPa	D 790	53 452	98	90			
Flexural strength (break), MPa	D 790	53 457			140	120	46
Flexural modulus, MPa	D 790	53 457	2500	2500	6000	5000	1800
Hardness (H 358/30), MPa		53 456	100	87	122	147	
Hardness, Rockwell	D 785		R 115	R 115	L 106		
Flexural creep (300 h, 6.9 MPa), %	D 674						0.60
Impact strength							
Charpy impact (notched), kJ/m^2		53 453	15	15	9	9	5
Izod impact (notched), J/m	D 256		200	250	80	60	
Izod impact (notched, −40 °C), J/m	D 256		140	120	70	50	
Thermal							
Vicat softening point (rate B), °C	D 1525		148	105	145	148	
Vicat VST/B/120, °C		53 460	135	100	140	135	
Deflection temperature under load (0.45 MPa), °C	D 648	53 461				133	96
Deflection temperature under load (1.82 MPa), °C	D 648	53 461	130	90	132	126	83
Thermal conductivity, $W\,m^{-1}K^{-1}$	C 177		0.22	0.16	0.24		0.12
Linear thermal expansion coefficient (flow), K^{-1}	D 696	VDE 0304/1	6×10^{-5}	7×10^{-5}	4×10^{-5}	2.4×10^{-5}	$(7-8) \times 10^{-5}$
Physical							
Relative density	D 792	53 479	1.06	1.10	1.21	1.45	0.88
Water absorption (24 h), %	D 570		0.07	0.08	0.06	0.05	
Water absorption (equilibrium), %	D 570		0.14	0.37	0.14		
Water absorption (equilibrium, 100 °C), %	D 570		0.30	0.55	0.32		
Mold shrinkage (flow), %	D 955		0.5 – 0.7	0.5 – 0.7	0.2 – 0.4	0.1 – 0.3	0.6 – 0.9
Electrical							
Volume resistivity, $\Omega \cdot m$	D 257		$>10^{15}$	$>10^{15}$	$>10^{15}$	$>10^{15}$	
Dielectric strength (3.2 mm), kV/mm	D 149		22.0	16.0	17.0		7.5 (6 mm)
Dielectric constant (50 Hz)	D 150		2.70	2.70	2.90		2.30
Dielectric constant (1 MHz)	D 150		2.60	2.60	2.90		2.20
Dissipation factor (50 Hz)	D 150		0.0004	0.0004	0.00008		0.0047
Dissipation factor (1 MHz)	D 150		0.0009	0.0009	0.0014		0.0039
Arc resistance (tungsten), s	D 495						
Flame characteristics							
Oxygen index, %	D 2863		24	33	26		28
Flame retardancy (mm thickness)	UL-94		HB (1.65)	V-0 (1.52) 5 V (3.12)	HB (1.47)	V-0 (1.60) 5 V (3.20)	V-0 (6.29) 5 V (6.29)

[a] Noryl 731.
[b] Noryl N 190, V – 0 rating according to UL 94; V-1 and V-2 grades also available.
[c] 20 wt % glass-filled grade Noryl GFN 2; 10, 20, and 30 wt % glass-filled grades also available.
[d] V-0 rating according to UL 94; 35 wt % glass-filled grade Noryl VO 3525.
[e] Noryl FN 150 D.
[f] Unless otherwise stated.

are more effective at higher HIPS levels [11]. Processing temperatures of PPE-rich compositions above 80 °C cause additional cross-linking of the polybutadiene rubber phase, rendering it less effective.

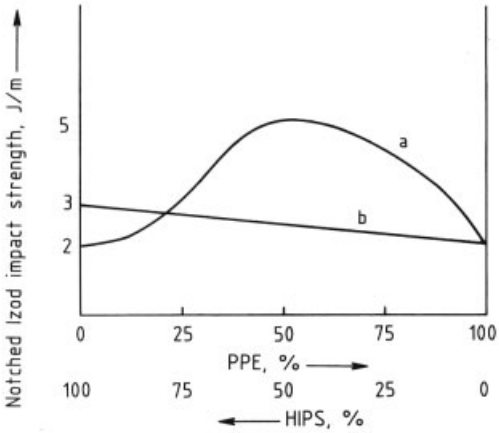

Figure 4. Notched impact strength of PPE – HIPS blends impact-modified with polybutadiene rubber (10 wt %, content in HIPS)
a) 0.5 – 2 µm polybutadiene rubber; b) 2 – 12 µm polybutadiene rubber

PPE – HIPS grades offer relatively constant mechanical properties over a wide temperature range. Ultimate use temperatures of PPE – HIPS blends, as estimated from UL temperature ratings, are between 90 and 100 °C (see Table 2).

The very low water absorption of PPE –HIPS blends and their favorable mechanical properties contribute to a high dimensional stability under load, the most important property of engineering thermoplastics.

Polymer miscibility is not restricted to the PPE – HIPS system, but also holds for combinations of other substituted poly(1,4-phenylene oxides) (see Table 1) with styrene homopolymers such as poly(α-methylstyrene), poly(4-methylstyrene), and styrene copolymers in which the comonomer content is limited to maintain full miscibility with PPE.

Adjustment of the maximum use temperature offered by blend technology also applies to the UL-94 flammability ratings of the PPE – HIPS grades (Fig. 5). For a description of UL-94 ratings, see → Flame Retardants, Chap. 3. PPE is intrinsically flame extinguishing as exemplified by its high limiting oxygen index (LOI) value of 29. HIPS, on the other hand, has a LOI of 17 and burns easily and completely [12]. Blending these two miscible polymers provides an intermediate flammability behavior which can be further improved by adding triaryl phosphates to produce flame-retardant PPE – HIPS grades that meet required industrial specifications without utilizing halogen-containing compounds [13]. Since the organic phosphate also functions as a plasticizer, a higher percentage of PPE is required to maintain a given heat distortion temperature (HDT) (Fig. 3). This leads to a well-balanced relation between flame retardancy behavior, composition, and HDT. The flame-retardant PPE – HIPS grades also have excellent melt processibility.

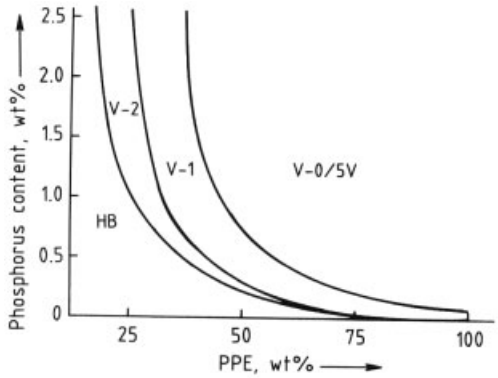

Figure 5. UL-94 flammability ratings of PPE – HIPS blends with different triaryl phosphate contents

Enhanced flame retardancy, however, decreases the comparative tracking index (CTI), a criterion used for applications in the electrical and electronics industry (→ Insulation, Electric, Chap. 2.1.). This problem can be solved by the addition of inorganic fillers (e.g., glass fibers) and has led to the development of special grades.

PPE – HIPS blends are, like most polymers, excellent electrical insulators. The dielectric loss factor, a measure for the loss of electrical energy through heat buildup, is < 0.002 in the range 1 – 16 MHz over a wide temperature span [14]. For other electrical properties of poly(phenylene oxides), see → Insulation, Electric.

Chemical resistance of PPE – HIPS blends is summarized in Table 3. Temperature, strain level, contact time, and the chemical nature of the contact liquid strongly affect the related environmental stress cracking phenomena. PPE – HIPS blends have an intrinsically high resistance

against hydrolysis under acidic and basic conditions.

Table 3. Chemical resistance of PPE – HIPS (Noryl) and PA – PPE (Noryl GTX) blends

Chemical	PPE – HIPS blend	PA – PEE blend
Water and aqueous solutions		
Cold water	excellent	very good
Hot water	excellent	good
Cleaning agent	very good	satisfactory/good
Salts	very good	good
Acidic solutions	good	good
Basic solutions	good	satisfactory
Concentrated bases	good	satisfactory
Concentrated acids	satisfactory	bad
Organic solvents		
Alcohols	very good	very good
Ketones	very good	satisfactory
Aliphatic hydrocarbons	swelling	very good
Aromatic hydrocarbons	bad	good
Chlorinated hydrocarbons	bad	bad
Esters	bad	satisfactory
Miscellaneous		
Gasoline	bad	(very) bad
Fats	type-dependent	(very) good
Oils	type-dependent	(very) good
Freon	type-depentent	(very) good

3.2. PA – PPE Blends

PPE can also be used as an organic filler in semicrystalline matrices such as polyamides (PA 66 or PA 6). PA – PPE blends have been specifically developed for on-line painted exterior automotive parts with excellent ("class A") surface quality.

The favorable processibility properties, chemical resistance, and impact resistance of PA are combined with the high-temperature performance of PPE. PA – PPE blends were the first commercial blends of two completely incompatible polymers and are different from the fully miscible PPE – HIPS system. Blends of immiscible polymers are thermodynamically unstableand lack morphological stability, which leads to phase separation and delamination. This can be overcome by the use of compatibilizers (concen-tration usually ≥ 10 wt %), which are frequently A–B block copolymers. The compatibilizers reduce the interfacial tension between the constituent polymers, resulting in greatly improved dispersion of PPE in PA. They also provide strong bonding between the dispersed and the continuous phases (important for mechanical properties), and increase morphological stability by preventing agglomeration of dispersed particles. The latter is important as it allows processing of the blends under practical shear conditions such as exist in injection-molding equipment.

Compatibilization is achieved either by adding PA – PPE block copolymers or by in situ formation of these block copolymers. The latter is achieved by addition of PPE end-capped with functionalized maleic anhydride (Fig. 6 A) or by addition of PS grafted with maleic anhydride (Fig. 6 B). The anhydride groups react with the amine end groups of the PA polymers and form a compatibilizing linear PA – PPE block copolymer (Fig. 6 A) or a grafted PS – PA block copolymer (Fig. 6 B). Figure 7 shows the effect of compatibilization as determined by scanning electron microscopy [15]. Higher concentrations of PA – PPE copolymers reduce the dimensions of the dispersed PPE phase.

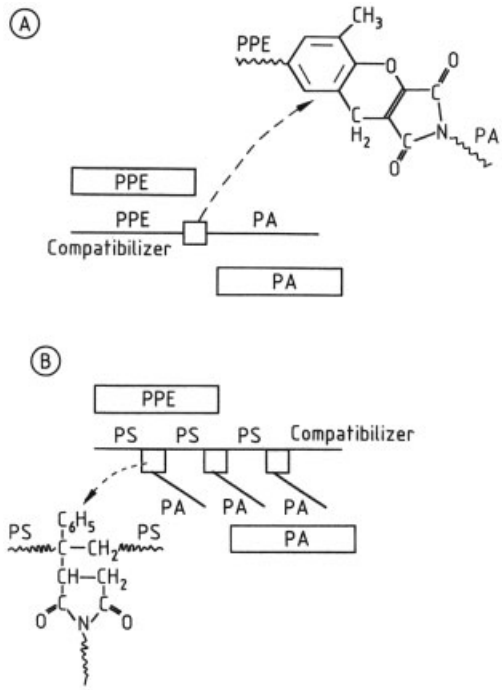

Figure 6. Schematic representation of compatibilization of PA – PPE using PPE end-capped with functionalized maleic anhydride (A) or PS grafted with maleic anhydride (B)

PPE is a ductile material which can be impact-modified to form extremely tough blends. The high glass transition temperature of

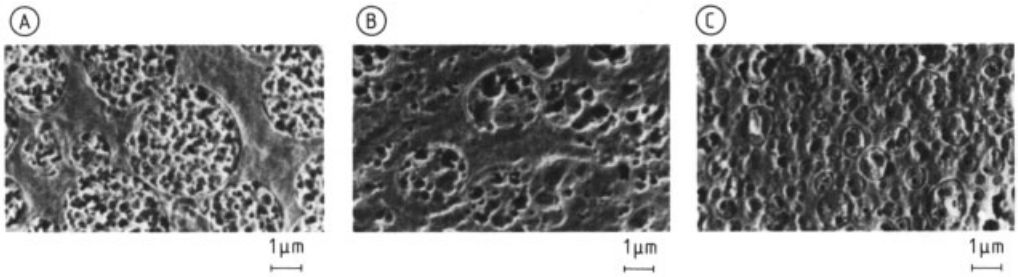

Figure 7. Effect of compatibilization on PA – PPE morphology as determined by scanning electron microscopy with PA as the continuous phase (courtesy of S. Y. Hobbs, General Electric)
A) No compatibilizer; B) 8 wt % compatibilizer; C) 15 wt % compatibilizer

PPE also increases the strength, rigidity, and dimensional stability of the PA – PPE blends up to the temperature required for automotive lacquering. The blends are usually impact-modified phase selectively with various rubbers located in the PPE or PPE – PS phase dispersed in the PA phase. Two-phase impact modification of the PPE (or PPE – HIPS) and the PA phases is also possible by appropriate selection of the rubbers. The PA – PPE blends can thus easily be adapted to give the ultimate use temperatures and impact levels required for specific lacquering and painting applications of exterior automotive parts (see Fig. 8). Ranges of temperature requirements for various coating techniques are indicated.

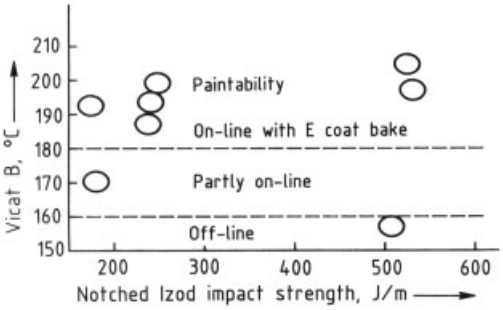

Figure 8. Ultimate use temperature plotted against impact strength of commercially available PA – PPE blends (Noryl GTX, GE Plastics), ranges of possible coating techniques are indicated

Table 4 shows important properties of PA – PPE blends. Polyamides are known for their high moisture absorption which reduces their dimensional stability. Addition of PPE to PA lowers the water absorption but, remarkably, PA – PPE blends absorb less moisture than that calculated from their PA volume fraction (Fig. 9). The resulting blend thus has a greater dimensional stability than PA, providing a more constant mechanical property profile with temperature and relative humidity. This dimensional stability can be further enhanced by reinforcing the material with glass fibers. Special grades exist for bumper beams that combine high-temperature paintability with low-temperature impact performance.

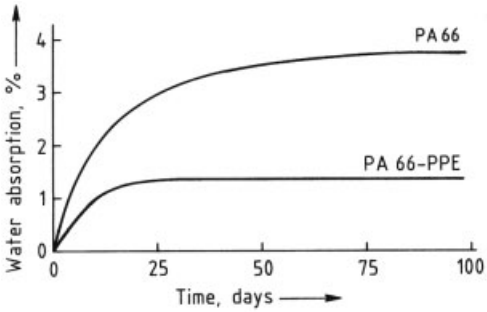

Figure 9. Water absorption of PA and PA – PPE at 23 °C and 50 % R.H.

Advantages of PA – PPE blends over PPE – HIPS blends are their better chemical resistance (see Table 3), improved environmental stress cracking resistance, and a higher ultimate use temperature range.

3.3. PPE – Polyolefin Blends

In 1989 GE Plastics introduced a blend series (Noryl Xtra) containing ≥ 90 % PPE. The high melt viscosities of PPE (see Chap. 1) were reduced using plasticizers and olefinic components of proprietary composition as flow promoters. The resins combine good processibility

Table 4. Properties of PA – PPE blends (Noryl GTX, GE Plastics)

Property	Standards		Unreinforced [b]	30 wt % Glassfiber-reinforced[c]
	ASTM[a]	DIN[a]		
Mechanical				
Tensile strength (yield), MPa	D 638	53 455	55	
Tensile strength (break), MPa	D 638	53 455	50	160
Tensile elongation (yield), %	D 638	53 455	5	
Tensile elongation (break), %	D 638	53 455	100	3
Tensile modulus, MPa	D 638	53 457	2000	9000
Flexural stength (yield), MPa	D 790	53 452	85	
Flexural strength (break), MPa	D 790	53 452		220
Flexural modulus, MPa	D 790	53 457	2100	7000
Hardness (H 358/30), MPa		53 456	115	141
Impact strength				
Charpy impact (notched), kJ/m^2		53 453	15	7
Izod impact (notched), J/m	D 256		240	85
Izod impact (notched, − 20 °C), J/m	D 256		190	75
Thermal				
Vicat softening point (rate B), °C	D 1525		220	240
Vicat VST/B/120, °C		53 460	190	240
Deflection temperature under load (1.82 MPa), °C	D 648	53 461		225
Thermal conductivity, $Wm^{-1}K^{-1}$	C 177		0.23	0.26
Linear thermal expansion coefficient (flow), K^{-1}	D 696	VDE 0304/1	9×10^{-5}	$(2-3) \times 10^{-5}$
Physical				
Relative density	D 792	53 479	1.10	1.31
Water absorption (24 h), %	D 570		0.40	0.50
Water absorption (equilibrium), %	D 570		3.50	3.60
Mold shrinkage (flow), %	D 955		1.2 – 1.6	0.3 – 0.5
Flame characteristics				
Oxygen index, %	D 2863		20	29
Flame retardancy (mm thickness)	UL-94		HB (1.60)	HB (1.60)

[a] Except where otherwise stated.
[b] Noryl GTX 900.
[c] Noryl GTX 830

with high dimensional and temperature stability up to 195 °C as well as an UL-94 V-0 flammability rating. The materials have a high surface gloss and can be used as light reflectors.

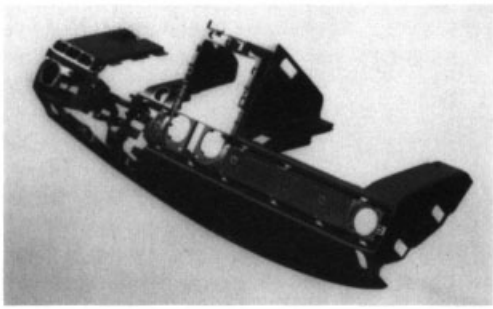

Figure 10. PPE – HIPS dashboard (Noryl, GE Plastics)

In addition to excellent intrinsic flame retardancy properties, the special-grade Noryl Xtra LS generates an extremely low amount of smoke during burning because it contains PPE-specific char promoters. This feature is extremely important for its use in electrical insulation applications in enclosed or inaccessible areas (airplanes, ships, public buildings).

4. Processing

Melt viscosities of all PPE – HIPS and PA – PPE grades conform to the processing capabilities of conventional melt-processing equipment (injection molding, extruding etc.). However, PA – PPE blends require a more stringent drying procedure, and care should be taken with PPE – HIPS blends not to exceed the recommended drying time of 4 h at 110 °C, since this leads to loss of impact resistance. Processing temperatures of standard grades are typi-

cally 220 – 270 °C for PPE – HIPS, 280 – 300 °C for PA – PPE, and 320 – 340 °C for PPE – olefin blends. Injection molding and glass-filled grades require slightly higher temperatures.

5. Uses

The most important application areas for PPE blends are in the automotive industry (35 %), in electric and electronic appliances (25 %), office equipment (15 %), and (hot) water distribution and environmental applications (8 %) [13].

Interior Automotive Applications. The development of PPE – HIPS blends marked the break-through of plastics for interior automotive applications. The most significant applications are dashboards of PPE – HIPS (Fig. 10), a development that started in Europe and was followed worldwide. Other applications include radio housings, loudspeakers, and heating and ventilation systems. So called "antisqueak" materials used for handgrips and ventilation grills.

Figure 11. Fenders of 1991 Renault Clio 165 (Noryl GTX, GE Plastics)

Exterior Automotive Applications. The on-line and off-line paintability of PA – PPE blends provided by their high-temperature performance, their "class A" surface with excellent dimensional stability, and good mechanical properties in combination with low-temperature impact resistance makes them suitable for vertical body panels, bumpers, petrol filling caps, wheel trims, grilles, spoilers, electromechanical systems (EMS), and fenders for the 1991 Renault Clio 16 S (Fig. 11). These applications allow for easy styling changes, weight savings, and, most importantly, cost savings. The latter is particularly important for "small" production series of 50 000 vehicles per year.

Electrical and Electronics Industry. Flame-retardant PPE – HIPS grades are used mostly on account of their good processability, good mechanical properties, and flame retardancy behavior without using halogens. Typical applications are housings for electrical equipment (Fig. 12), connectors, switches, and interior parts in laundromats. PA – PPE blends can be used in paintable extrusion and blow molding parts.

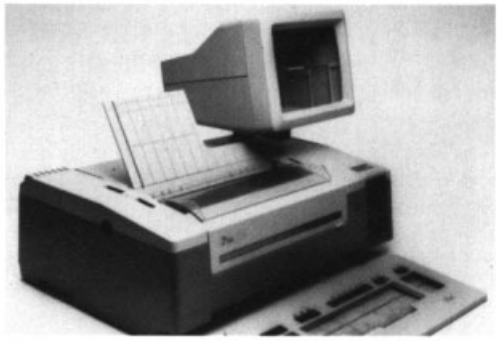

Figure 12. PPE – HIPS housing (Noryl, GE Plastics)

Data Processing and Office Equipment Industry. Flame-retardant PPE – HIPS blends or, alternatively, painted structural foams are used for in-house appliances such as computer housings. These materials are provided with a UV-stabilizing system to comply with the most stringent demands for color stability.

The hydrolytic and dimensional stability of PPE – HIPS blends and the reduced stress – crack sensitivity provided by glass fiber reinforcement makes them suitable for applications in water and air purification systems. All blends in applications involving food contact or drinking water are approved by the FDA.

Building and Construction. This industry segment constitutes a new market for PPE blends. The low sales percentages result from the fact that traditional materials are still used in potentially large-volume applications such as insulation systems or solar collectors.

New Developments. A few promising new developments can be mentioned. In situ PPE – PS blends used to make expandable beads give strong foams with a low density (to 0.02 g/cm^3) and a heat distortion temperature above 100 °C, a sterilizable and reusable material (Fig. 13).

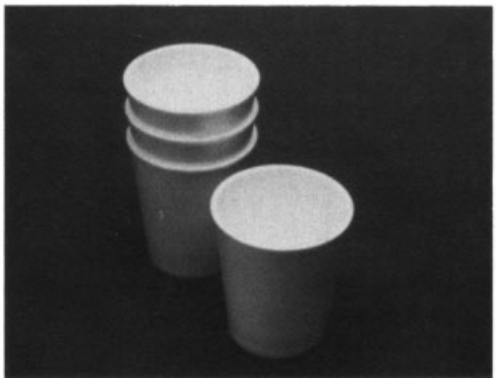

Figure 13. Beakers made from PPE – PS expandable beads (Caril, Shell)
Bead size 300 – 500 µm, density 60 g/L, T_g 110 °C

The use of PPE as an organic, highly heat-resistant, flame-retardant filler that does not affect mechanical properties at high filling ratios, as demonstrated for PA – PPE blends, will be applied to other blend systems (Section 3.2). These new blends include a blend of PPE and poly(butylene terephthalate) which has an even lower moisture absorption than PA – PPE blends and thus higher dimensional stability. Blends of chemically modified PPEs will also find increased use, aimed at applications with specific requirements.

6. References

1. General Electric, US 3 306 874, 1967 (A. S. Hay).
2. D. Aycock, W. Abolius, D. M. White: "Poly(Phenylene ether)," *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vol. 13,** John Wiley & Sons, New York 1988, p. 1.
3. Z. S. Slama, *Acta Polym.* **31** (1980) 746.
4. M. Kryszewski, J. Jachowicz in N. Grassie (ed.): *Developments in Polymer Degradation,* **vol. 4,** Applied Science Publishers, London 1982 p. 1.
5. General Electric, US 3 063 872, 1962 (E. M. Boldenbruch).
6. H. L. Finkbeiner, A. S. Hay, D. M. White: "Polymerization by Oxidative Coupling," in C. E. Schildknecht, I. S. Skeist (eds.): *Polymerization Processes, High Polymers,* vol. 29, John Wiley & Sons, New York 1977, p. 537.
7. A. S. Hay, *Polym. Eng. Sci.* **16** (1976) 1. A. J. Schouten, G. Challa, J. Reedijk, *J. Mol. Catal.* **9** (1980) 41. C. E. Koning, R. Brinkhuis, R. Wevers, G. Challa, *Polymer* **28** (1987) 2310. F. J. Viersen, G. Challa, J. Reedijk, *Polymer* **31** (1990) 1369.
8. J. R. A. Pearson, *Polym. Eng. Sci.* **18** (1978) 222.
9. C. B. Bucknall, D. Clayton, W. E. Keast, *J. Mater. Sci.* **7** (1972) 1443.
10. General Electric, US 4 128 602, 1978 (G. Lee, A. Katchman).
11. C. B. Bucknall: *Toughened Plastics,* Applied Science Publishers, London 1977.
12. D. W. van Krevelen: *Properties of Polymers,* 3rd ed., Elsevier, Amsterdam 1990.
13. H. Frank, *Kunststoffe* **76** (1986) 859.
14. R. I. Warren, *Polym. Eng. Sci.* **25** (1985) 477.
15. J. R. Campbell, S. Y. Hobbs, T. J. Shea, V. H. Watkins, *Polym. Eng. Sci.* **30** (1990) 1056.

Polystyrene and Styrene Copolymers

Jürgen Maul, Hüls AG, Marl, Germany (Chap. 1)

Bruce G. Frushour, Monsanto Chemical Company, Springfield, MA 01150, United States (Chap. 2, 3 and 4)

Jeffrey R. Kontoff, Monsanto Chemical Company, Springfield, MA 01150, United States (Chap. 2, 3 and 4)

Herbert Eichenauer, Bayer AG, Dormagen, Germany (Chap. 5)

Karl-Heinz Ott, Bayer AG, Dormagen, Germany (Chap. 5)

Christian Schade, BASF AG, Ludwigshafen, Germany

1. Polystyrene . . . 1
1.1. Introduction . . . 1
1.2. Production . . . 3
1.2.1. Bulk Polymerization . . . 4
1.2.2. Suspension Polymerization . . . 6
1.3. Properties . . . 7
1.3.1. Chemical Properties . . . 7
1.3.2. Physical and Processing Properties . 7
1.4. Processing and Uses . . . 9
1.5. Quality Specifications . . . 10
1.6. Storage and Transportation . . . 11
1.7. Recycling . . . 11
1.8. Environmental Aspects . . . 11
2. Styrene – Acrylonitrile (SAN) Copolymers . . . 12
2.1. Production . . . 12
2.2. Properties . . . 14
2.3. Processing . . . 15
2.4. Uses . . . 16
2.5. Blending of SAN . . . 16
3. Barrier Resins . . . 16
4. Other Copolymers . . . 17
4.1. α-Methylstyrene – Acrylonitrile Copolymers . . . 17
4.2. Styrene – Methyl Methacrylate Copolymers . . . 18
4.3. Styrene – Maleic Anhydride Copolymers . . . 18
4.4. Styrene – Maleimide Copolymers 19
4.5. Styrene – Acrylate Copolymers . . 19
5. Acrylonitrile – Butadiene – Styrene (ABS) Polymers . . . 19
5.1. Definition and Structure . . . 19
5.1.1. Historical Aspects . . . 19
5.1.2. Structural Principles . . . 20
5.1.3. Synthesis of the Two-Phase Structure 21
5.1.4. Properties of the Resin Matrix . . . 22
5.2. Structure – Property Relationships . . . 23
5.3. Production of ABS Polymers . . . 24
5.3.1. ABS Production by Emulsion Polymerization . . . 25
5.3.2. Production of the Matrix Copolymer 29
5.3.3. ABS Production by Bulk Polymerization . . . 31
5.3.4. ABS Production by Bulk Suspension Polymerization . . . 33
5.3.5. Process Combinations . . . 34
5.3.6. Additives . . . 35
5.4. Quality Assurance and Standardization . . . 35
5.5. Properties . . . 36
5.6. Special Product Modifications . . . 38
5.7. Legal Aspects . . . 38
5.8. Storage and Transportation . . . 39
5.9. Uses . . . 39
5.10. Economic Aspects of Acrylonitrile-Based Styrene Copolymers . . . 39
5.11. Recycling . . . 40
5.12. ABS-Analogous Systems . . . 40
5.12.1. ASA, AES, and ACS Polymers . . . 40
5.12.2. MBS and MABS Polymers . . . 41
5.13. ABS Blends . . . 43
6. References . . . 44

1. Polystyrene

1.1. Introduction

Polystyrene (PS) belongs to the group of standard thermoplastics that also includes polyethylene, polypropylene, and poly(vinyl chloride). Because of its special properties, polystyrene can be used in an extremely wide range of applications. The annual consumption worldwide in 2004 was ca. 11.5×10^6 t; thus polystyrene is one of the most quantitatively important chemicals.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_615.pub2

Historical Aspects. Styrene [*100-42-4*] (known as styrax) was first isolated in 1831 by Bonastre from the resin of the amber tree. In 1839 E. Simon, who also first described the polymer, gave the monomer its name. He observed that styrene was slowly converted into a viscous solution on standing. Around 1925 the development of an industrial production process for polystyrene began; this work achieved success in the plants of the IG Farbenindustrie in Germany in 1930 [1]. In the United States polystyrene was first produced on a commercial scale in 1938 by the Dow Chemical Company.

Polystyrene and Impact-Resistant Polystyrene. Chapter 1 deals with the clear polystyrene molding materials [*9003-53-6*] and rubber modified polystyrene [*9003-55-8*], also known as styrene – butadiene molding material.

Polystyrene molding materials are hard, transparent materials with a high gloss. They are most commonly described as general-purpose polystyrene (GPPS) but terms such as standard polystyrene, normal polystyrene, clear polystyrene, crystal polystyrene, or styrene homopolymer are also in use. In this article the definition polystyrene molding material (PS) is used according to ISO 1622–2.

Below 100 °C PS molding materials solidify to give a glasslike material with adequate mechanical strength, good dielectric properties, and resistance toward a large number of chemicals for many areas of application.

Above its softening point clear polystyrene occurs as a melt which can be readily processed by techniques such as injection molding or extrusion. Small quantities of lubricants can be used internally or externally as processing aids. The addition of antistatic agents, UV stabilizers, glass fibers, or colorants is also common.

The mechanical properties of the relatively brittle PS molding materials can be considerably improved by adding rubbers, generally polybutadiene [*106-99-0*]. Styrene – butadiene molding materials are generally referred to as high-impact polystyrene (HIPS), this term is used here. They are also known as toughened PS or rubber-modified PS; ISO 2897–2 defines them as impact-resistant polystyrene (IPS).

Early production processes for HIPS were based on mixing PS molding materials with a rubber component. Polymerization of a styrene – polybutadiene solution is, however, much more effective. A two-phase system is formed due to the immiscibility of polystyrene and polybutadiene. Polystyrene forms the continuous phase (matrix) and polybutadiene the disperse phase. The globular rubber particles contain small inclusions of polystyrene (Fig. 1).

The additives commonly used with PS molding materials can also be employed in HIPS. Paraffinic white oil is commonly used as lubricant, zinc stearate as mold release agent. Antioxidants are used for rubber stabilization. Flame retardants, antistatic agents, and other additives are added for special applications.

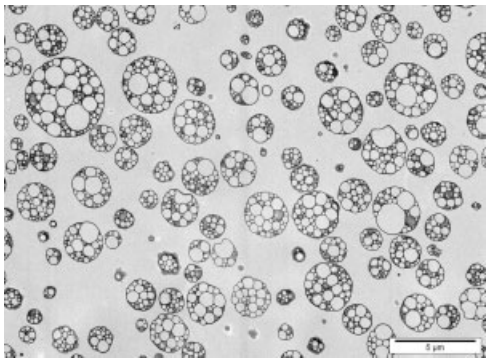

Figure 1. Electron micrograph of HIPS molding material The dispersed rubber particles (dark as a result of OsO_4 treatment) are embedded in the polystyrene matrix (lighter regions). The rubber particles contain lighter polystyrene inclusions
Courtesy of W. Heckmann.

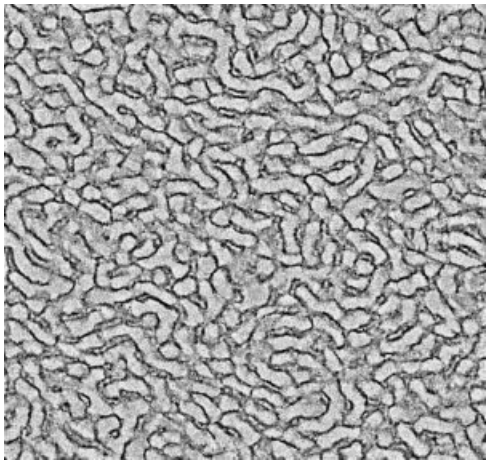

Figure 2. Electron micrograph of a transparent IPS molding material (SBC, contrasting as in Fig. 1). The size of the square is approx. 1.5 μm × 1.5 μm.
Courtesy of W. Heckmann.

The rubber particles in HIPS generally have a mean diameter of 0.5 – 9 μm. They thus scatter visible light and the transparency of the PS molding materials is lost. Sequential anionic block copolymerization of styrene and butadiene allows the production of transparent [3, 4]. Phase separation of the different blocks occurs. Typically lamellar or similar impact resistant molding materials, also called styrene block copolymers (SBC) structures are formed instead of discrete particles (Fig. 2). The thickness of the layers is far less than the particle size in HIPS, so that light scattering does not occur at the phase boundaries.

Styrene Copolymers. Styrene can be copolymerized with many other monomers. Styrene – acrylonitrile molding materials, in particular, have achieved great economic importance in transparent and rubber-modified forms (SAN [*9003-54-7*], Chap. 2; ABS [*9003-56-9*], Chap. 5). Compared with the pure styrene polymers they have advantages with regard to hardness, strength, and resistance to heat distortion and environmental stress cracking. However, these advantages are offset by a higher price and more difficult processing.

Copolymers of styrene and maleic anhydride [*9011-13-6*] (Section 4.3) have a softening point that is up to 30 °C higher than that of PS molding materials. These products are used in the form of foams in the automotive industry (e.g, as interior headliners).

Polystyrene Foams. Expandable polystyrene (EPS) is the starting material for PS hard foam materials. It is produced from polystyrene by the addition of ca. 6 % of a low-boiling hydrocarbon (e.g., pentane) as the foaming agent. Extruded polystyrene (XPS) foams are produced from polystyrene and a blowing agent, e. g., carbon dioxide. Foamed polystyrene is discussed in detail elsewhere, → Foamed Plastics.

1.2. Production

Styrene. The properties and production of the raw material styrene are described elsewhere (→ Styrene). The styrene used for polymerization should have a purity greater than 99.6 % because the contaminants arising from the production process, mainly ethylbenzene, cumene, and xylenes, affect the molecular mass of the polystyrene. Commonly, purities of 99.9 % or more are used.

For storage and transport styrene is stabilized with inhibitors such as 4-*tert*-butylcatechol to prevent polymerization at low temperatures. These inhibitory radical scavengers must be removed by distillation or adsorption on alumina.

Polymerization. Styrene can function as an electron donor or an electron acceptor. It can therefore be polymerized by radical, cationic, or anionic mechanisms, or by coordination propagation steps.

The industrial polymerization of styrene to PS and HIPS molding materials is carried out exclusively by a free radical mechanism. The chain reaction does not necessarily have to be started by the addition of radical-forming, readily decomposing initiators because styrene itself can form polymerization-initiating radicals [5]. The propagation mechanism for the chain growth proceeds by addition of further monomer to the radical chain end. Growth of the radical chain is mainly terminated by recombination. With increasing temperature, disproportion and back-biting (see → Polyolefins, Section 1.3.1.1) followed by β-scission [6] play an increasing role. The growth of a chain proceeds within seconds; the gross rate is largely determined by the radical initiation reaction. Average molecular masses between 100 000 and 400 000 are obtained within a short time. Molecular masses are mainly controlled by reaction temperature, chain-transfer agents — including solvents — and by initiators. The kinetics of the process are well understood [7] and may be modeled by commercially available software.

Cationic polymerization of styrene (e.g., by strong Lewis acids such as $AlCl_3$, $HClO_3$, or BF_3) is unimportant industrially because control of the molecular mass is difficult. Very small quantities of very low molecular mass polystyrene are synthesized via such routes.

Transparent HIPS (SBC) is, however, produced industrially by anionic block copolymerization of butadiene and styrene. Organolithium compounds (e.g., *n*-butyllithium) are used as initiators. The monomers and solvent used in anionic polymerization have extremely high pu-

rity requirements. The polymerization gross rate is higher than that of the radical chain reaction due to the simultaneous growth of all initiated chains. Polymerization is carried out at low temperature in solution to enable the reaction to be controlled.

The coordination polymerization of styrene with conventional Ziegler catalysts produces a sterically ordered, isotactic polymer. These highly ordered chains can form polymer crystals. The melting temperature of the crystalline zones of isotactic polystyrene is ca. 230 °C. The rate of crystallization is, however, so low that this property cannot be exploited industrially [8].

Syndiotactic polystyrene crystallizes very rapidly and is produced by using special titanium – aluminoxan catalysts [9, 10]. The melting temperature of the syndiotactic sequences is ca. 270 °C. The material has found limited use in heat-resistant applications, often glass-fiber reinforced or blended with polyamide.

HIPS Synthesis. Rubbers used for HIPS synthesis are medium-*cis*-polybutadienes with a glass transition temperature around – 85 °C. They are obtained by butyllithium-initiated anionic polymerization of butadiene. Usually the polybutadienes are long-chain branched to prevent cold flow during storage. Other rubbers (e. g., high-*cis* grades, styrene – butadiene block copolymers, EPDM rubbers) may be also employed, but have less significance.

The synthesis of HIPS requires careful control of the rubber particle size and morphology. Initially, the solution consists of polybutadiene rubber dissolved in styrene and further additives such as diluents, chain-transfer agents, or initiators. With the onset of polymerization, the polystyrene forms a separate phase. Styrene and optional diluents are the common solvents for the elastomer and the polystyrene phase (oil-in-oil emulsion). The polystyrene phase grows at the expense of the rubber phase with further conversion. The rubber phase initially is the continuous phase and remains so if unstirred. In a stirred reactor, phase inversion occurs, largely depending on the phase volume ratios and viscosities of the two phases. The reaction conditions during phase inversion determine the rubber morphology and particle size to a large extent.

Simple blending of polystyrene and polybutadiene leads to a mechanically inferior product. The mechanical strength of HIPS is supported by polystyrene molecules, which are grafted onto rubber molecules. Grafting occurs especially at low styrene conversions, when the volume of the polystyrene phase is still small. Quantity and nature of these copolymers have big impact on the particle size and morphology of the HIPS rubber particles [11, 15].

1.2.1. Bulk Polymerization

Currently, PS and HIPS molding materials are mainly produced by continuous polymerization processes. The polystyrene product has consistent quality, high purity, and a low residual monomer content (generally less than 500 ppm monomeric styrene). The most difficult task from an industrial standpoint is the control of the highly viscous melt and the removal of the heat of polymerization (71 kJ/mol).

Solvents such as toluene or ethylbenzene are frequently present in small proportions (typically 5 – 15 %) to provide better control of the polymerization rate, the viscosity of the polymer melt, and the cross-linking of the rubber phase. This technique is known as *solution polymerization*. Due to the limited quantities of the diluent, the processes are often also referred to as *continuous bulk polymerization*. The solvent and the unreacted monomer must be removed by degassing. They are generally recovered, optionally purified and transferred back to the feed. In a running continuous operation, solvents are in essence introduced via styrene impurities (mostly ethylbenzene).

The plant setup generally comprises a polymerization section, a degassing and solvent recovery section, and a pelletizing section.

The most common reactor for *PS molding materials* is an ebullient single stirred-tank reactor. A second stirred-tank reactor or a tubular plug-flow reactor may be used to increase conversion. PS molding materials may also be produced in more complex reactor configurations like HIPS units.

HIPS plants generally consist of a series of three to five reactors, split into a prepolymerization and postpolymerization section. In the prepolymerization section phase inversion takes place, and rubber morphology and particle size are largely adjusted. Various reactor designs

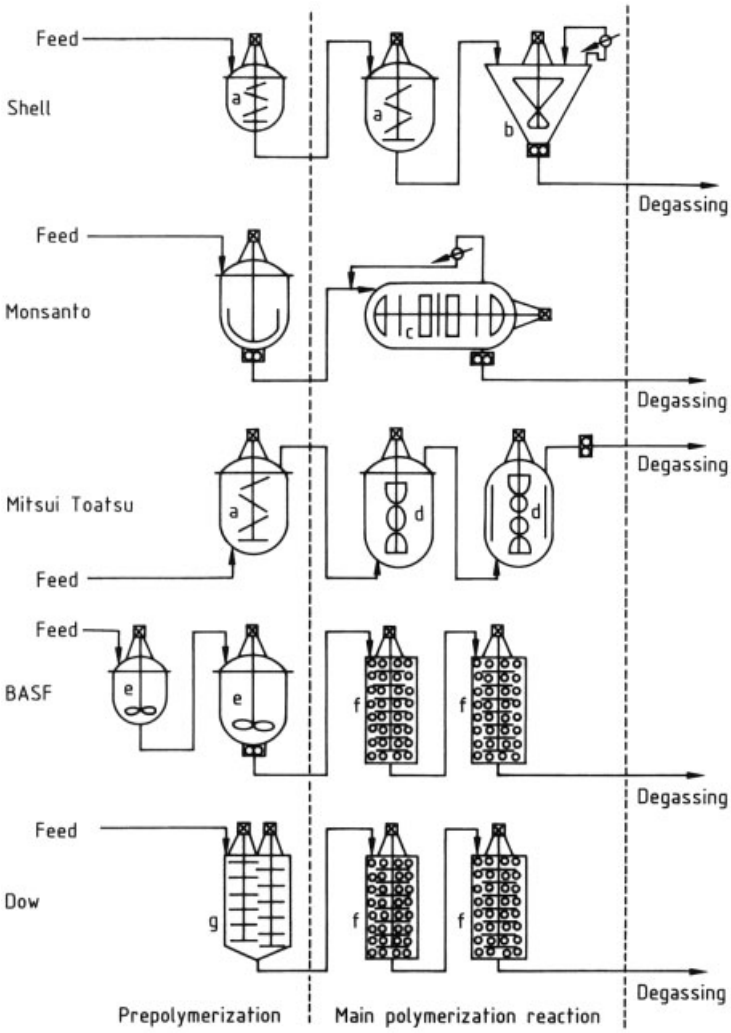

Figure 3. Various reactor configurations for HIPS production
a), b), d), e), g) Stirred-tank reactors; c) Horizontal plug-flow reactor; f) Tower reactors
Reproduced from [14] with permission of Hanser Publishers.

have been proposed for this part of the plant. Common are one or two stirred-tank reactors, tower reactors, loop reactors with static mixers, or combinations of these. The reaction is typically carried out at 100 – 150 °C to 15 – 30 % conversion. The postpolymerization part commonly consists of several plug-flow type reactors to achieve high styrene conversions, but continuous stirred-tank reactors may be used as well. Common temperatures are 140 – 190 °C in this part of the plant. Figure 3 shows several common reactor configurations.

In the *continuous stirred-tank reactor* part of the heat is removed by means of reflux cooling. The heat of polymerization causes part of the styrene to evaporate. The gaseous monomer is condensed in condensers with large surface areas and recycled back to the reactor. The prerequisite for this type of heat removal is effective mixing of the reactor contents. Anchor stirrers, impellers, or helical ribbon agitators are commonly used. The reaction temperature is the most important parameter affecting the molecular mass of the polystyrene and thus the properties of the product. It is controlled by varying the

pressure and thus the boiling point of styrene in the stirred tank. The temperatures in the reactors are between 100 and 170 °C, corresponding to pressures of ca. 0.5 – 2 bar.

The reactor feed can also be used to remove heat of polymerization. Reactors converting more than 40 % of the styrene feed usually need additional cooling. The lower solid content of the HIPS prepolymerization reactors, on the other hand, necessitates preheating of the reactor feed.

Tower reactors are equipped with special baffles in which a cooling agent circulates to ensure the removal of the heat of polymerization. Different temperature zones can be set up in a reactor by adjusting the temperature of the cooling agent. The tower reactors are also equipped with agitator blades.

Tubular reactors with static mixing elements and cooling are also in use. Adiabatic or isothermal tubular reactors are sometimes used for high conversions. Also known are axially segmented and agitated tubular reactors operating under reflux.

The polystyrene melt leaving the reactor section has a concentration of 60 – 90 %. Unreacted styrene monomer and solvent are removed by flash devolatilization. The polymer is fed into an evacuated, heated vessel where the volatile components evaporate, the melt is then pumped out. The styrene – solvent mixture formed is recycled to the first reactor. A small fraction must be discarded, however, to avoid increasing the concentration of styrene contaminants.

A second flash tank or another degassing unit (e.g., falling strand degassers, steam strippers, thin-film evaporators, degassing extruders) may be used to further decrease content of volatile compounds.

The polymer melts can be mixed with additives (e.g., dyes, stabilizers, or lubricants) by using static or mechanical mixing equipment.

Degassing of the polystyrene melt is followed by pelletization. In extrusion pelletization the polymer is forced through a perforated disk in continuous strands. The necessary pressure is applied by means of the discharge pump in the degassing unit or by an extruder. The strands are cooled in a water bath and cut into cylindrical pellets.

Underwater and water-ring pelletizing systems are also used and yield lentil-shaped pellets.

The pellets can be coated with an externally applied lubricant (e.g., zinc stearate or *N,N'*-ethylene-bis(stearamide)) to improve the sliding properties.

1.2.2. Suspension Polymerization

Suspension processes are chiefly used for the production of expandable polystyrene. PS molding materials may be produced by this process when the blowing agent is omitted. The combined bulk – suspension process is sometimes still used for HIPS production.

In suspension polymerization styrene is dispersed in water in the form of small droplets by stirring. The addition of suspension agents such as poly(vinyl pyrrolidone), poly(vinyl alcohol), alkaline-earth phosphates, or celluloses is necessary to stabilize the organic phase. Polymerization must be started with radical-forming initiators because the process temperatures are considerably lower than those of continuous bulk polymerization (80 – 110 °C).

The main advantage of suspension polymerization is that the heat of polymerization can easily be removed via the aqueous phase. A high degree of conversion can also be achieved.

Generally, however, the polystyrene beads must be repelletized after separation of the water and drying to satisfy market requirements with regard to processing of the molding materials (e.g., injection molding equipment). Additives and dyes can be incorporated during regranulation.

In the case of HIPS, suspension polymerization must be preceded by bulk polymerization to ensure that rubber particles are formed. The reaction can only be carried out in suspension after the formation of the disperse polybutadiene phase. The product has comparable material properties to HIPS produced by the bulk process.

The batch suspension process for styrene polymerization has been gradually replaced by the more economically favorable continuous bulk processes. Very few of the older plants based on batch suspension technology are still run commercially.

Table 1. Physical properties of PS and HIPS resins

Property	Test method	Typical values	
		PS resins	HIPS resins
Density, g/cm^3	DIN 53 479 ASTM D 792	1.05	1.03 – 1.05
Melt flow index (200 °C/5 kg), g/10 min	DIN 53 735 ASTM D 1238/G	1 – 30	2 – 20
Tensile modulus, MPa	DIN 53 457 ASTM D 638	3000 – 3400	1600 – 2800
Yield tensile strength, MPa	DIN 53 455 ASTM D 638	35 – 60	20 – 45
Tensile elongation at break, %	DIN 53 455 ASTM D 638	1.5 – 3	25 – 70
Ball indentation hardness, MPa	DIN 53 456 ISO 2039	150 – 160	75 – 140
Flexural strength, MPa	DIN 53 452 ASTM D 790/1	60 – 120	35 – 80
Charpy impact strength, kJ/m^2	ISO 179	10 – 25	35 – no break
Notched Charpy impact strength, kJ/m^2	DIN 53 453	1 – 2	5 – 17
Izod impact strength, kJ/m^2	ISO 180/1 A	10 – 25	35 – no break
Notched Izod impact strength, J/m	ASTM D 256/A	15 – 30	50 – 150
Vicat softening temperature, °C			
VST/B (50°/h, 5 kg)	DIN 53 460	78 – 102	75 – 97
VST/A (120°/h, 1 kg)	ASTM D 1525/B	85 – 110	80 – 100
Thermal conductivity, W K^{-1} m^{-1}	DIN 52 612	0.16	0.16
Dielectric constant (1 MHz)	DIN 53 483 ASTM D 150	2.5	2.5 – 2.7
Dissipation factor (1 MHz)	DIN 53 483 ASTM D 150	$(1 – 4) \times 10^{-4}$	$(2 – 8) \times 10^{-4}$
Thermal coefficient of linear expansion, K^{-1}	DIN 53 752	$(6 – 8) \times 10^{-5}$	$(8 – 10) \times 10^{-5}$
Refractive index	DIN 53 491	1.59	

1.3. Properties

1.3.1. Chemical Properties

The chemical properties of atactic PS and HIPS are similar.

Under certain conditions oxidizing acids lead to oxidative degradation. However, the molding materials are resistant to dilute acid, alkali, and salt solutions. Exposure to UV radiation in the presence of oxygen causes yellowing and embrittlement, particularly in HIPS modified with polybutadiene. Reactions at the remaining double bonds bring about a gradual loss in elastic properties of the soft components. This damage can be suppressed but not completely inhibited by the addition of light stabilizers. Polystyrene is therefore not generally used in areas of application where weather resistance is required.

The replacement of polybutadiene rubber by EPDM rubber and the addition of special antioxidants lengthen the service lifetime of materials exposed to weathering by a factor of 15 – 20.

Oils and fats (particularly important in the food packaging sector) produce stress cracking. This can, however, be effectively counteracted by a range of measures [13], for example, by moderately increasing the average particle size to more than 5 μm.

The self-ignition temperature of PS and HIPS is 490 °C. Polystyrene articles are generally classified as being moderately flammable (DIN 4102) and as slow-burning (Underwriters Laboratories UL 94 HB). The flammability of special HIPS molding materials can be reduced considerably with additives (e.g., organobromine compounds) to give products that are classified as having a low flammability. The main area of application for these products is in the electrical sector. They are also increasingly being used for housings for technical appliances (e.g., photocopiers, televisions).

1.3.2. Physical and Processing Properties

Polystyrene, being a linear hydrocarbon chain with phenyl side groups linked, on average, to every second carbon atom, dissolves in aromatic hydrocarbons (toluene, ethylbenzene), halogenated hydrocarbons, aliphatic ethers (tetrahydrofuran), esters (ethyl acetate), and ketones

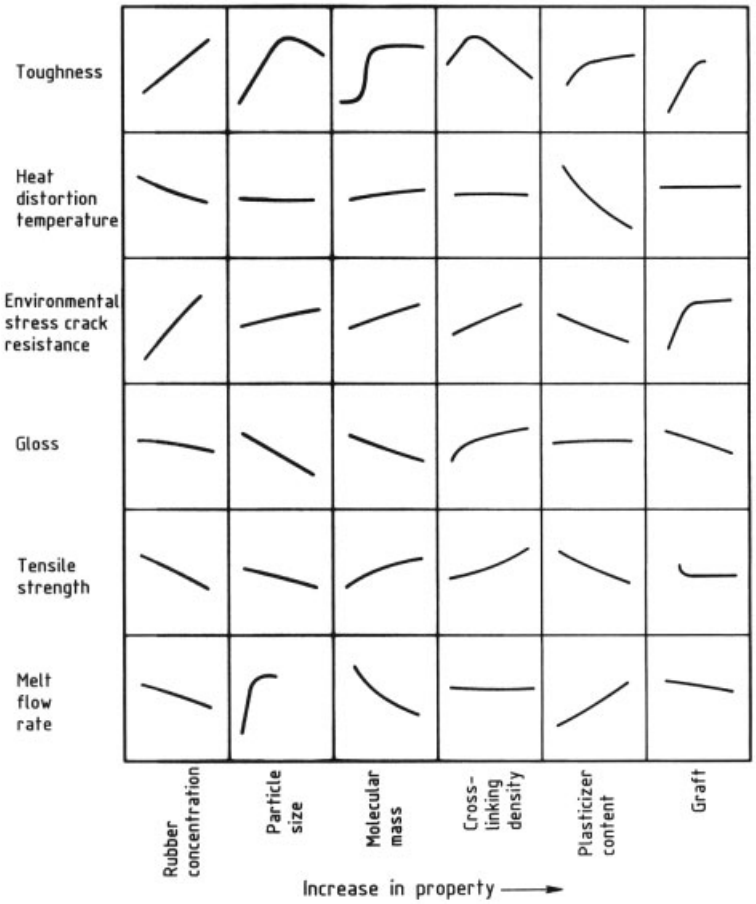

Figure 4. Summary of property relationships of commercial mass-produced HIPS
Reproduced from [12] with permission of The Dow Chemical Company.

(butanone), but not in aliphatic hydrocarbons (hexane) and alcohols (methanol). A summary of the solubility behavior of polystyrene can be found in [12].

The most important physical data for PS and HIPS are given in Table 1.

Special PS molding materials are available with specific property profiles for different processing technologies such as injection molding, extrusion, or thermoforming.

The mechanical and rheological behavior of polystyrene is predominantly determined by its average molecular mass (normally 150 000 – 400 000). The strength improves with increasing chain length but the melt viscosity increases making processing more difficult. Lubricants such as butyl stearate or paraffinic oil can be added to lower the melt viscosity of PS and HIPS. However these agents also lower the softening point.

PS grades with very good or good flow properties are mainly used for processing in injection molding machines. High molecular mass grades with higher strength and resistance to heat distortion are preferred for extrusion of films and sheets.

In extrusion and injection molding, the polystyrene melt is forced through narrow nozzles or slits. The polymer molecules orientate themselves parallel to the flow direction due to the high shear rate. The melt often solidifies in the mold so quickly that this initial orientation cannot be reversed by molecular diffusion. Polystyrene end products may therefore display anisotropy of their mechanical properties parallel and perpendicular to the processing direction.

Tensile properties are much increased in parallel to the processing direction, while they are generally weakened perpendicularly to it. Biaxially oriented polystyrene has increased mechanical strength over the entire film. It may dissipate mechanical energy by a so-called "shear yielding" rather than the more common and less efficient "crazing" mechanism.

The impact resistance of HIPS is higher than that of PS due to the rubber component, but transparency is lost [exception: transparent styrene – butadiene multiblock copolymers (SBC) produced by anionic polymerization (Section 1.1)]. Processing properties are not altered significantly by the polybutadiene phase. However, the softening point, the modulus of elasticity, and the hardness are lower than those of PS molding materials.

In HIPS, the rubber phase (generally polybutadiene) is distributed in the form of discrete particles in the polystyrene matrix. The most important parameters affecting the properties of the end products are the rubber content, morphology, average particle size, particle-size distribution, and the degree of cross-linking of the polybutadiene phase (Fig. 4).

Rubber Content. The HIPS grades from different producers may be classified according to their polybutadiene (PB) content: semi-impact-resistant (ca. 3 – 6 % PB) and impact-resistant (6 – 10 % PB). The real criterion is not the rubber content, but the gel content that is obtained with a given rubber content (i.e., the mass fraction of the disperse phase including the polystyrene enclosed within the particles and grafted onto the polybutadiene chains) [14]. The gel content is roughly three to four times the quantity of rubber added.

Rubber Morphology. A wide range of different particle morphologies can be obtained depending on the chemical composition of the rubber and the polymerization conditions. Standard HIPS grades containing polybutadiene usually consist of cellular particles (Fig. 1). If part of the polybutadiene is replaced by styrene – butadiene block copolymers, coils, mazes, shell, or capsule particles are formed [15].

Particle Size and Particle-Size Distribution. The particle-size spectrum of the rubber phase is the factor determining the impact resistance of the molding materials. Commercial grades have average particle sizes in the range 0.5 – 9 μm. The highest mechanical efficiency is in the range 1.5 – 3 µm. If the particles are significantly larger, less crazes and less cavitation can be supported by the same rubber quantity [16 17]. Grades with particles sizes > 5 μm have increased environmental stress crack resistance. Rubber particles smaller than 1 µm are less efficient in craze initiation. Such HIPS grades have reduced impact strength and are used for high gloss applications.

The average particle size depends on the polymerization conditions during phase inversion. The most important parameters listed below can be modified as required:

1) The ratio between the solution viscosities of the disperse polybutadiene and homogeneous polystyrene phases
2) The shear force applied to the polymerizing material via the agitator
3) The surface tension between the two phases [18]—largely affected by the degree of grafting

Degree of Cross-Linking. Another important parameter is the cross-linking density of the polybutadiene phase. The rubber chains must be cross-linked with one another to a certain degree so that they can resist the shear forces used in subsequent processing, thus guaranteeing a stable particle form. However, the cross-linking density must not be too high because this would lead to the loss of the desired rubber-elastic properties. Cross-linking is a thermally initiated radical reaction and occurs mainly at high degrees of conversion and in the degassing section of the polymerization plant. The parameters in this process step (residence time, temperature) must therefore be adjusted to give the desired degree of cross-linking in the end product.

1.4. Processing and Uses

PS and HIPS molding materials are processed using molding processes generally used for thermoplastics (e.g., injection molding, extrusion, or thermoforming). About 50 % of the total quantity is processed with screw injection molding machines at 200 – 280 °C.

Films and sheets are produced by extrusion at ca. 200 °C with broad-slit dies. Thermoforming is used to produce articles such as refrigerator

linings from sheets and yogurt cups and disposable tableware from films.

PS and HIPS molding materials are mutually compatible and can be mixed with one another in all proportions. The properties of the mixtures depend on the proportions of the components.

The main areas of application of PS and HIPS molding materials are in packaging, technical items, household and consumer goods, and refrigeration equipment. Roughly two thirds of PS and HIPS are consumed for packaging and for technical items. Packaging accounts for more than 50 % of the consumption in Western Europe and the United States, where technical items represent 20 – 25 % of the market. In Asia, 20 – 25 % of the material goes to the packaging sector, while more than 50 % are used for technical items. Asia accounts for more than 40 % of the total polystyrene consumption, the NAFTA region and the European Union for about 25 % each (2005). Recent annual growth rates have been ca. 2 %.

Technical items include housings for television, computers, printers, air conditioners and the like, electrical appliances, computer accessories, and sanitary ware.

Household and consumer goods include toys, containers, furniture, disposable tableware, and cups for drink vending machines. Polystyrene is used for the packaging of compact discs, audiocassettes, foods and pharmaceuticals. Polystyrene is especially suitable for the packaging of dairy products. Biaxially oriented polystyrene and PS blown film are used for display packaging, wrap film, and envelope windows.

PS and HIPS molding materials do not affect odor or taste and are licensed worldwide for contact with foods. Legal requirements in individual countries specify permitted concentrations of materials used in production and processing. All standard commercial PS and HIPS molding materials comply with these guidelines.

Trade Names and Producers. PS and HIPS molding materials are produced by more than 100 raw materials manufacturers throughout the world who offer an extensive range of products classified according to strength, processing behavior, and dimensional stability when subjected to heat. Total installed capacity is about 15×10^6 t. Examples of trade names are given in alphabetical order by producer below:

Polystyrol	BASF Aktiengesellschaft, Germany
Starex	Cheil Industries, Korea
Polyrex	Chi Mei, Taiwan
-	Chevron – Phillips Chemical Comp.
Styron	Dow Chemical, USA
Edistir	Enichem Polimeri, Italy
-	Estizulia, Venezuela
-	Innova, Brazil
-	Ineos, Great Britain
-	Kumho, Korea
LG	LG Chemicals, Korea
-	Japan Polystyrene
-	Nizhnekamskneftekhim, Russia
-	Nova Chemicals, Canada
-	Resirene, Mexico
-	PS Japan
SC	Supreme Petrochem. Ltd., India
PS	Sabic, Saudi Arabia
-	Secco, China
-	Shantou, China
Lacqrene	Total, France
-	Videolar, Brazil
-	Wuxi Wei Da, China

The producers (and trade names) of transparent HIPS molding materials are Asahi Chemicals, Japan (Asaflex), BASF, Germany (Styrolux); Chevron – Phillips Chemicals, USA (K-resin); Chi Mei, Taiwan (Kibiton), Denka, Japan (Clearen), and Kraton Polymers, USA (Kraton). Total installed capacity is about 0.5×10^6 t (2006).

Syndiotactic polystyrene is produced by Idemitsu Petrochemical, Japan, under the trade name Xarec. Installed capacity is 5000 t/a (2001).

1.5. Quality Specifications

Most PS and HIPS molding materials are processed automatically which means that product quality must be as uniform as possible. Quality assurance systems (ISO 9000 – 9004) are common in the production of molding materials to guarantee maximum uniformity. The ISO standard specifies the quality of all raw materials, the production conditions, and the properties of the end products.

The rheological, mechanical, thermal, and electrical properties of the molding materials are defined in ISO 1622–2 (PS) and 2897–2 (HIPS). The Vicat softening temperature (VST B/50, ISO 306), the melt flow index (MFI 200/5,

ISO 1133), and the impact resistance (Charpy, ISO 179) are among the most important properties. National standards such as ASTM or DIN (Table 1) largely correspond to the ISO standards.

1.6. Storage and Transportation

The storage and transportation of PS and HIPS molding materials do not pose any special problems. They are subject to the general fire protection requirements applicable to solid combustible materials such as wood. Roofed sheds or silos are suitable for storage. Polystyrene is not moisture-sensitive, but must be dry for processing. PS and HIPS molding materials are transported in paper or polyethylene bags, in large cardboard containers (capacity ca. 1 t), in silo vehicles or containers. In the NAFTA region, rail cars (hopper cars, capacity 80 t) are also used.

1.7. Recycling

Waste PS and HIPS can be reutilized in various ways. In *material recycling* the molding materials are broken into smaller pieces, melted, and reprocessed to give finished products. *Chemical recycling* involves the breaking up of the polymer chain into lower molecular mass components which can be used as raw materials for chemical processes. In *thermal recycling* the plastic is incinerated and the liberated energy is used, for example, as process heat or for general heating purposes. See also → Plastics, Recycling.

Material Recycling. Recycling of industrial waste, such as trims and scrap, is not problematic. Polystyrene end products can also be recycled after use provided that the material can be cleaned and sorted into grades. In no case, however, should food packaging be produced from this recycled material. Some countries stipulate legal quotas for material recycling of plastics.

Chemical Recycling. Two processes are in the development stage: pyrolysis and cleavage by hydrogenation cracking. Both processes can be applied to pure plastics fractions (e.g., polystyrene) and mixtures of plastics.

Pyrolysis involves the cleavage of long-chain molecules at high temperatures (500 – 900 °C) with the exclusion of oxygen [22]. In hydrogenating cleavage the polymer chains are broken up under higher pressure at ca. 400 °C, and in the presence of hydrogen, into volatile saturated compounds. Both processes yield a nonspecific spectrum of products (60 – 80 % of styrene and ethylbenzene, respectively), which can be used as feedstock for other petrochemical processes.

Thermal Recycling. In the presence of sufficient oxygen, polystyrene burns almost completely to give carbon dioxide, water, and a small amount of nontoxic ash residue. The enthalpy of combustion (41 MJ/kg) is close to the energy content of fuel oil (43 MJ/kg). In modern waste incineration plants, polystyrene comprises part of the plastics fraction of household waste, adherence to the necessary combustion conditions is guaranteed.

1.8. Environmental Aspects

PS and HIPS molding materials are not water soluble. No gaseous or liquid components are liberated during transportation and storage. Liquid and gaseous styrene must, however, be handled safely during production processes. Polymerization is therefore carried out in closed systems. The maximum permitted workplace concentration of styrene is between 10 and 50 ppm (42 – 210 mg/m^3) in all important producer countries.

In bulk polymerization processes, the product only comes into contact with water during the cooling of the extruded strands. Contamination of the water does not occur. In the suspension process, polymerization takes place in an aqueous medium and the wastewater must be treated by using standard sewage technology.

In the processing of PS and HIPS, small quantities of styrene are liberated above ca. 280 °C as a result of the residual monomer content and of thermal degradation. The styrene can be removed by special ventilation devices; however, normal room ventilation is usually adequate to ensure compliance with workplace concentration limits.

2. Styrene – Acrylonitrile (SAN) Copolymers

Many applications requiring stiffness and transparency are served by polystyrene, a moderately priced commodity thermoplastic (Chap. 1). A mature market has also continued to thrive for resins offering a better balance of optical clarity, rigidity, good chemical and heat resistance, high strength, and flexible processing. One class of thermoplastics having these properties is the styrenic copolymers. The following four copolymers continue to have growing commercial importance:

1) Styrene – acrylonitrile (Chap. 2)
2) α-Methylstyrene – acrylonitrile (Section 4.1)
3) Styrene – methyl methacrylate (Section 4.2)
4) Styrene – maleic anhydride (Section 4.3)

These copolymers are produced commercially by free-radical polymerization [23]. Their specific properties are determined by their molecular mass and composition. The initial concentration of the two monomers (M_1 and M_2) prior to initiation of polymerization must be carefully chosen to give the required copolymer composition for the finished product. Techniques are available for calculating the correct monomer concentration that utilize parameters known as reactivity ratios (r) that reflect the relative reactivity of each monomer with the free-radical end of the growing chain [24]. The dependence of the copolymer composition on the monomer concentration for each of the above styrene copolymers is given in Figure 5.

The styrene copolymers having the greatest commercial importance are the styrene – acrylonitrile (SAN) copolymers [*9003-54-7*]. Incorporation of acrylonitrile improves heat resistance, toughness, chemical resistance, and barrier properties when compared against polystyrene. The SAN polymers can be used by themselves in applications requiring transparency, or they can be toughened by incorporation of rubber particles. Use of butadiene rubber gives acrylonitrile – butadiene – styrene (ABS) polymers (see Chap. 5).

The acrylonitrile (AN) levels in commercial SAN polymers fall into two ranges. The lower range (20 – 35 wt %) covers most injection-molded articles, and the higher range (60 – 85 wt %) represents a small but growing market for barrier plastics where low gas permeability is needed for packaging of food and beverages (see Chap. 3).

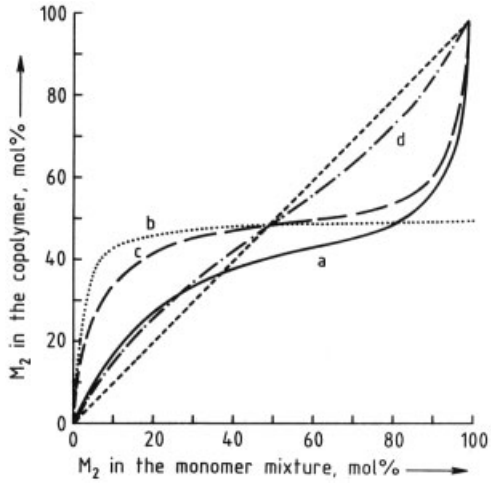

Figure 5. Composition diagram for common styrene copolymers
a) Styrene (M_1) – acrylonitrile (M_2) (r_1= 0.40, r_2= 0.04); b) Styrene (M_1) – maleic anhydride (M_2) (r_1= 0.04, r_2= 0); c) α-Methylstyrene (M_1) – acrylonitrile (M_2) (r_1= 0.10, r_2= 0.06); d) Styrene (M_1) – methyl methacrylate (M_2) (r_1= 0.59, r_2= 0.54)

2.1. Production

The relative reactivities of styrene and acrylonitrile differ. Therefore, the polymer composition also differs from the monomer composition, except at the azeotrope (ca. 24 wt % acrylonitrile). This, along with the fact that acrylonitrile is soluble in water, requires special considerations for the various production methods.

Using reactivity ratios [25] and Equation (1) [26] for two monomers M_1 and M_2:

$$C = \frac{P-1+\sqrt{(1-P)^2+4Pr_1r_2}}{2\,r_1} \qquad (1)$$

C = molar ratio M_1/M_2 in the monomer mixture
P = molar ratio M_1/M_2 in the copolymer
r_1 = reactivity ratio for monomer 1
r_2 = reactivity ratio for monomer 2

the monomer – polymer composition curve can be determined for SAN and other common

copolymer pairs (Fig. 5). Reactivity ratios (styrene: r_1 = 0.4 and acrylonitrile, r_2 = 0.04; Fig. 5.) imply that a mixture of about 75 wt % styrene and 25 wt % acrylonitrile shows no compositional drift ("azeotropic composition").

Reactivity ratios are a function of temperature [27] and the composition curves should be adjusted if the temperature range during manufacture is wide. This is not as critical when the azeotrope composition is approached.

Maintaining tight composition control of SAN is important to obtain acceptable optical and mechanical properties. Differences of a few weight percent in the SAN composition can result in incompatibility [28]. For example, addition of 2 wt % SAN with 32 % acrylonitrile to SAN with 24 % acrylonitrile leads to hazy material. As conversion increases, severe composition drift occurs (Fig. 6) unless corrective actions are taken to replenish the more reactive monomer [29].

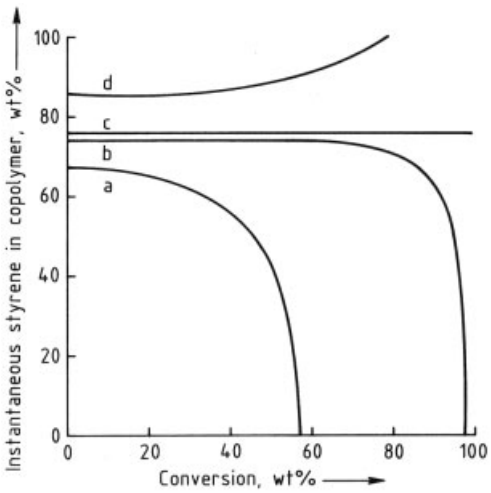

Figure 6. Approximate compositions of SAN copolymers formed at different conversions, starting with various monomer mixtures [29]
Styrene : acrylonitrile weight ratio
a) 35/65; b) 70/30; c) 76/24; d) 90/10

Commercially, SAN is produced worldwide by emulsion, suspension, and continuous bulk technologies [30]. These methods are described in detail elsewhere, → Polymerization Processes. The majority (85 %) of SAN is used captively as the matrix component for ABS. Most SANs used for the manufacture of ABS contain 24 – 32 wt % acrylonitrile. The properties of ABS can be influenced by the composition and molecular mass distribution of the SAN, which are manipulated by the polymerization method employed to make the SAN. This is discussed in detail in Section 5.3.2.

SAN materials sold as finished products are almost exclusively made by continuous bulk methods which results in superior optical properties. Also, continuous bulk technology is the most cost efficient and generates the least waste of the three methods.

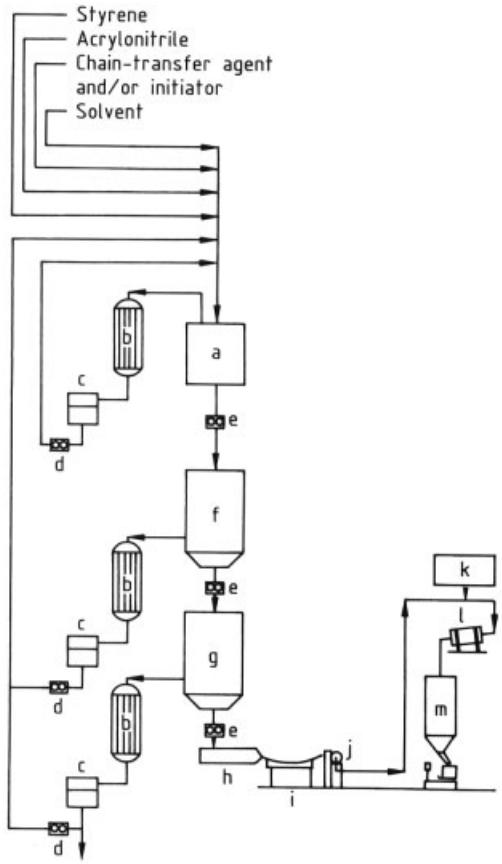

Figure 7. Process for the continuous bulk and solution polymerization of SAN [31]
a) Reactor; b) Condenser; c) Condensate receiver; d) Recycle pump; e) Melt pump; f) First-stage devolatilizer; g) Second-stage devolatilizer; h) Discharge screw pump; i) Water bath; j) Pelletizer; k) External lube; l) Screen; m) Silo

Figure 7 shows an example of the manufacture of SAN via continuous bulk polymeriza-

tion. The process consists of continuous feeds to one or more reactors (a), followed by devolatilization units (f, g) and pelletizer (j). Unreacted monomers and solvent are recycled to maintain conversion and composition at desired levels. The technology is very similar to that employed for PS molding materials (see Section 1.2.1).

The reactor(s) can be plug flow or partially filled types, cooled by jacket and/or reflux [31, 32]. Adequate agitation is critical for proper temperature and composition control. Polymerization temperatures range from 80 – 170 °C, depending on the process configuration and whether chemical or thermal initiation is used. Conversion ranges from 30 to 70 %.

Devolatilization is usually accomplished via wiped-film or falling-strand type units, but can also be done with a devolatilizing extruder. To achieve the low levels of residual monomer required to meet food contact applications, devolatilization is usually done in a two-stage process with different vacuum levels.

2.2. Properties

SAN copolymers with acrylonitrile contents between 20 and 35 wt % are transparent plastics having better mechanical and chemical resistance properties than polystyrene, thus enabling SAN to command a premium price. Properties for commercial grades of polystyrene and SAN are compared in Table 2 [33]. Next to a higher rigidity and impact resistance, the SANs are stronger and more scratch resistant and have a better heat resistance.

SAN can be reinforced with glass fiber to give higher stiffness, breaking strength, and notched impact strength [34, 35]. The effect of adding 20 % glass fiber to SAN is shown in Table 3 [36].

Table 3. Comparison of properties of filled and unfilled SAN

Property	Unfilled	Filled with 20 % glass fiber
Tensile strength, MPa	69 – 76	74 – 83
Elongation at break, %	2 – 3	1.4 – 1.9
Tensile modulus, GPa	3.3 – 3.9	8.3 – 11.7
Notched Izod impact strength, J/m	21.4 – 32	133 – 160
Heat deflection temperature, °C at 1.82 MPa load	101 – 104	99 – 110
Density, g/cm^3	1.07 – 1.08	1.22 – 1.40

The conventional method for making glass-fiber-reinforced SAN is to mix the polymer pellets and glass fiber in an extruder, followed by injection molding of the glass-filled pellets. This approach suffers from the inadvertent reduction in glass-fiber length (12 mm) to ca. 0.4 – 0.5 mm.

An alternative approach is to make a glass concentrate (collimate) by polymerizing the SAN in the presence of the glass fibers, and then blending the collimate with SAN pellets. Commercial collimates are available containing 80 % glass [37]. The glass fiber length in the final product is about 1 mm, and superior reinforcing properties are claimed with the collimate approach. The SAN collimate can be added to other thermoplastic resins.

The heat resistance (softening temperature) of typical commercial grades of SAN is ca. 5 – 6 °C higher than that of polystyrene (Table 2), thus allowing SAN to resist softening when exposed to boiling water. The softening of styrene polymers (and copolymers) is a consequence of the glass to rubber transition that occurs when glassy polymers are heated [38].

Addition of acrylonitrile to the styrene chain continuously increases the glass transition temperature (T_g) until a level of approximately 35 wt % acrylonitrile is reached. The T_g then falls with further addition of acrylonitrile [39]. The location and width of the glass transition for any composition is affected by the polydispersity of the molecular mass.

Commercial SANs contain oligomers left from polymerization which lower the T_g via plasticization. Removal of the oligomers in a commercial SAN (24 % acrylonitrile) increases the glass transition from 108.5 to 114.5 °C [40]. Therefore, any commercial SAN that is carefully treated to remove oligomers will benefit by having a higher heat resistance. The oligomers can be reduced by polymerizing at low temperatures ($<$ 120 °C) or by stripping them out during devolatization and subsequent purging from the recycle stream. If a large increase in heat resistance over polystyrene is needed, it is preferable to use the α-methylstyrene – acrylonitrile copolymers (see Section 4.1).

SAN has a slight yellow color in comparison to clear plastics such as polystyrene, acrylics, and polycarbonate. This color arises from cyclic chromophores associated with the acrylonitrile

Table 2. Comparison of the physical properties of standard polystyrene and SAN copolymers

Property	DIN test method	Standard polystyrene	SAN copolymers (20 – 35 wt % acrylonitrile)
Density, g/cm^3	53 479	1.05	1.08
Tensile strength, MPa	54 355	50 – 65	75 – 85
Tensile elongation to fail, %	53 455	2 – 4	5
Flexural strength, MPa	53 452	100 – 110	110 – 140
Flexural modulus, MPa	53 457	3200 – 3400	3600 – 3800
Impact resistance, kJ/m^2	53 453	15 – 20	20 – 25
Notched impact test, kJ/m^2	53 453	2 – 3	3 – 4
Heat resistance, °C	53 461, version B	80 – 100	90 – 105
Thermal conductivity, kJ m^{-1} h^{-1} K^{-1}		0.71	1.17
Linear thermal expansion coefficient, K^{-1}		$(6 – 8) \times 10^{-5}$	$(6 – 8) \times 10^{-5}$
Dielectric constant	53 483	2.5 – 2.7	2.9 – 3.1
Dielectric loss factor	53 483	$(1.8 – 2.1) \times 10^{-4}$	$(6 – 9) \times 10^{-3}$
Specific resistance, $\Omega \cdot$ cm	53 482	$> 10^{16}$	$> 10^{16}$
Dielectric strength, kV/mm	53 481	> 100	> 50
Water absorption, 24 h, %	53 472	0.1	0.2 – 0.6
Processing temperature, °C		150 – 260	175 – 300

segments, and it becomes stronger as the acrylontrile content increases. Some new commercial, "water-white" grades have been introduced for applications requiring minimum yellowness [27, 28]. The reduction in color is brought about by lowering the level of acrylonitrile and fine-tuning the polymerization process to reduce haze and eliminate contaminants.

One negative consequence of the addition of the acrylonitrile to the styrene chain is an increase in melt viscosity. The nitrile groups in acrylonitrile are strong dipoles, and, as the acrylonitrile level increases, the dipolar attraction between the acrylonitrile groups increases the melt viscosity. This can be compensated for by increasing the processing temperature (Table 2). However, raising the temperature also increases the yellow color, so a compromise must be reached between processibility, properties, and color.

The polar nature of the acrylonitrile also results in slight deterioration of electrical properties in comparison to polystyrene (Table 2). SAN still has adequate properties as an insulation material.

Excellent chemical resistance is a very important feature of SAN resins. They are resistant to aliphatic hydrocarbons, nonoxidizing acids, alkali, vegetable oils, foods, some alcohols, and detergents. Many of these materials readily attack polystyrene, but the polarity of the acrylonitrile groups in SAN is responsible for better resistance. SAN resins are attacked by some aromatic hydrocarbons, ketones, esters, and chlorinated hydrocarbons [41]. The moisture absorption of SAN is also slightly higher (Table 2), again because of the polarity of acrylonitrile.

The UV resistance of SAN is not exceptionally good, due to the styrene unit. However, the addition of UV stabilizers improves the UV resistance to the point that some SAN grades are used for glazing and other outdoor applications [41].

2.3. Processing

SAN resins can be processed by all methods common to thermoplastics, with injection molding being the most prevalent [42 – 44]. The resins are slightly hygroscopic and should therefore be dried prior to processing. Since SAN resins are immiscible with most other polymers it is important that the equipment be thoroughly purged prior to processing.

Injection molding barrel operating temperatures are typically ca. 220 – 280 °C, which is about 10 – 20 °C higher than is used for polystyrene. Yellowing and some mechanical property deterioration occurs if the temperature is too high or if too much regrind is used [45].

The mold should be filled as rapidly as possible. This can be accomplished by using large nozzles and gates, high pressure, fast ram, and adequate mold vents. Warm molds are necessary to achieve maximum surface gloss and improve impact strength—the warm mold lowers

the cooling rate and allows some of the flow orientation to relax before the part reaches the glassy state. Oven annealing of the part also reduces this orientation and may be required for special applications involving high mechanical stresses and pressures.

Finishing operations applicable to SAN include hot stamping, painting, screen printing, sonic welding, machining, vacuum metallizing, solvent welding, and most other operations used for ABS, acrylics, and polystyrene [46]. The good chemical resistance of SAN facilitates finishing operations requiring chemical exposure. Coloring and provision of antistatic agents, stabilizers, and flame retardants, as well as the reuse of scrap, are handled as with polystyrene (see Chap. 1).

2.4. Uses

The largest end use for SAN is as the rigid component for ABS (Chap. 5). All major ABS producers manufacture SAN for this captive use. SAN is used in the merchant market for a range of applications such as cookware, transparent parts in electronics and electrical appliances, instrument panels, sanitary and medical goods or cosmetic packaging. Chemical resistance and transparency are primary reasons for use [41, 42, 46].

The largest producers of SAN and their trade names are summarized in Table 4.

Table 4. Major producers and trade names of SAN

Producer	Trade name
BASF (Germany)	Luran
Daicel (Japan)	Cevian
Dow Chemical (USA)	Tyril
Techno Polymer Co. (Japan)	Sanrex
Polimeri (Italy)	Kostil

2.5. Blending of SAN

A new resin with unique properties may be obtained by blending (alloying) two or more polymers. Many studies have focused on SAN-based polymers as a major component of the alloy.

Commercial efforts have largely involved blending of ABS with major commercial polymers such as poly(ethylene terephthalate), polycarbonate, poly(vinyl chloride), and polyamide nylon to make resins with improved impact strength, heat resistance, and chemical resistance when compared to the individual polymers (see Section 5.12) [47]. These blends are not transparent because the rubber incorporated into the ABS for impact strength scatters light. The SAN does not become miscible with the other polymer. Rather, distinct domains of ABS and the other polymer phases remain, and the key to developing good mechanical properties is to reduce the domain size and promote good interphase adhesion [48].

Blending of SAN with other polymers to yield a transparent resin would require that the component polymers be thermodynamically miscible, alternatively the domains would have to stay below the critical size for light scattering.

Although polymers are generally not mutually miscible [47], SAN polymers show limited miscibility with poly(methyl methacrylate) [49], styrene – maleimide copolymers [50], styrene – maleic anhydride copolymers [51, 52], polycaprolactone [53], and poly(vinyl chloride) [54].

3. Barrier Resins

Acrylonitrile polymers and copolymers afford efficient barriers to gases, such as oxygen and carbon dioxide. Polyacrylonitrile [*25014-41-9*] (PAN) has the lowest permeability of any polymer, but its inherent rigidity and poor processibility (it degrades upon heating prior to melting) has spurred development of high acrylonitrile copolymers containing 60 – 85 wt % acrylonitrile.

Rubber-modified acrylonitrile – methyl acrylate copolymers are offered under the trade name Barex by Innovene. The acrylonitrile level in the Barex resin is 74 %. The copolymer is toughened by addition of styrene – butadiene rubber at a level of 10 %. The resins may be processed by all major thermoplastic processing methods.

Permeation of oxygen into food and beverages adversely affects their taste [55], and storage of carbonated beverages requires low permeation of carbon dioxide out of the bottle. The high acrylonitrile resins have much lower permeation rates than other thermoplastic packaging resins (Table 5). These high-acrylonitrile

Table 5. Permeation rates to oxygen and carbon dioxide at room temperature for polymers used in food packaging

Polymer	Permeation rate		
	Oxygen [a]	Carbon dioxide [a]	Ratio CO_2/O_2
Poly(vinylidene chloride)	0.0393	0.118	3.0
Barex 210	0.43	1.22	2.8
AN – styrene copolymer, 70–30	0.39	1.18	3.0
Poly(ethylene terephthalate)	2 – 4 [b]	4.7 – 7.8 [b]	
Poly(vinyl chloride) unplasticized	3 – 6 [c]	7.8 – 16 [c]	
Cellophane (dry, uncoated)	0.05	0.118	2.3
High-density polyethylene	43	118	2.7
Polypropylene	59	177	3.0
Low-density polyethylene	189	591	3.1

[a] $\frac{\mathrm{cm}^3,\mathrm{cm}}{\mathrm{m}^2,\mathrm{d},\mathrm{atm}} \times 10^5$ at $23^\circ C$, 100 % relative humidity.
[b] Depends on crystallinity and orientation.
[c] Depends on compound formulation.

resins are also quite good barriers to moisture, which is another requirement for food packaging.

The effect of acrylonitrile level on permeation rates for SAN copolymers is shown in Figure 8. The highest practical limit for acrylonitrile content is approximately 80 %, at which point the copolymer begins to take on the pseudocrystalline structure of polyacrylonitrile and becomes difficult to process [56, 57].

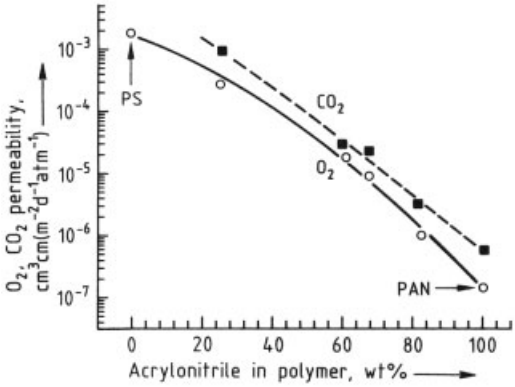

Figure 8. Permeability of oxygen and carbon dioxide through styrene – acrylonitrile copolymers
Values for the two homopolymers, polystyrene (PS) and polyacrylonitrile (PAN), are indicated.

Barrier resins are produced by suspension or emulsion polymerization [58, 59]. Maintaining constant composition at high acrylonitrile levels requires special techniques. A certain amount of the more reactive comonomer (e.g., styrene) is held outside the reactor and fed in a controlled manner during the polymerization cycle.

As animal tests have shown acrylonitrile to be a potent carcinogen, levels of acrylonitrile in the containers must be kept exceedingly low [60].

4. Other Copolymers

Producers and trade names of other important styrene copolymers are listed in Table 6.

Table 6. Examples of producers and trade names of common thermoplastic styrene copolymers

Producer (country)	Comonomers	Trade name
BASF (Germany)	α-methylstyrene – acrylonitrile	Luran KR 2556
Nova Chemicals (Canada)	styrene – methyl methacrylate	NAS, Zylar
Arco Chemical (USA)	styrene – maleic anhydride	Dylark
Denki Kagaku Kogyo(Japan)	styrene – maleimide	Malecca
Techno Polymer Co. (Japan)	styrene – maleimide	Superex

4.1. α-Methylstyrene – Acrylonitrile Copolymers

The copolymer of α-methylstyrene (AMS) and acrylonitrile [*25747-74-4*] has a higher heat resistance and softer flow than SAN of the same acrylonitrile content. The heat deflection temperatures of the AMS – AN and SAN copolymers are 104 and 96 °C, respectively [61]. Physical properties and chemical resistance are similar to those of SAN. As with SAN, most

AMS – AN copolymers are used as diluents to make high-temperature ABS (Sections 5.1.4 and 5.3.2) and acrylonitrile – styrene – acrylate (ASA) materials. AMS – AN copolymers are extensively compatible with SAN copolymers if they both have an acrylonitrile content of ca. 30 wt % [159]. The AMS copolymer is miscible with poly(vinyl chloride) [62], and could therefore be used to make both toughened and clear blends.

The manufacturing methods used for SAN can also be used for the AMS – AN copolymer. However, AMS reacts more slowly than styrene, greatly increasing the process residence times for a given acrylonitrile content. It is also more difficult to obtain high molecular mass AMS copolymers due to the higher termination rate constant compared with styrene and the strong chain-transfer effect of AMS oligomers and its monomer impurities. AMS-based materials are also yellower than their styrene counterparts.

By utilizing the difference, versus styrene, in reactivity ratio with acrylonitrile, a method for greatly increasing the production rate of AMS – AN copolymer has been developed [63]. Introducing low levels of styrene into the process to make an α-methylstyrene – acrylonitrile – styrene terpolymer is another way to increase reaction rates, with minor sacrifices in heat resistance.

4.2. Styrene – Methyl Methacrylate Copolymers

Methyl methacrylate readily polymerizes with styrene to form a styrene – methyl methacrylate copolymer [*25034-86-0*], SMMA, with similar physical properties to SAN, but with superior optical properties (low haze and glass-like appearance). Off-azeotrope copolymers are much easier to make, due to the similar reactivities of the monomers (Fig. 5). Commercial materials contain 20 – 60 % methyl methacrylate and are produced by methods similar to SAN [64, 65]. Methyl methacrylate is more commonly used to make clear ABS-type polymers (see Section 5.12.2).

4.3. Styrene – Maleic Anhydride Copolymers

Styrene and maleic anhydride react readily (Fig. 5) to give a copolymer [*9011-13-6*], SMA, with alternating monomer units. Reaction rates are significantly higher than for styrene alone or SAN, so polymerization conditions must be carefully chosen to control heat removal. Polystyrene technology is commonly used to produce thermoplastic SMA resins. The continuous stirred tank reactor technology permits synthesis of SMA copolymers with well defined S : MA ratio. The copolymerization parameters otherwise lead to the preferential formation of alternating copolymers.

The primary benefit of maleic anhydride is the greatly increased heat resistance. Also, the cost of this system is low, as both monomers are commodity materials. The products offer high rigidity, are commonly rubber modified, and are easily reinforced with fillers such as glass.

SMA resins generally fall into two categories. Low maleic anhydride ($<$ 25 wt %), high molecular mass thermoplastic materials are produced for molding or extrusion applications, low molecular mass materials with a high maleic anhydride content (25 – 50 wt %) are produced for alkali-soluble coatings and adhesives.

Thermoplastic SMA grades span a heat resistance range of 110 – 135 °C (Vicat softening point). Grades developed specifically for expandable foam applications and blends of SMA and polycarbonate for power tool housings are also available [66, 67].

For structurally demanding parts, rubber-modified SMA is reinforced with glass fiber (typically up to 20 %) by melt compounding. These materials are often used for interior automobile applications [67].

SMA resins are unstable at high temperature, liberating carbon dioxide. They must be adequately stabilized and processed at temperatures below 260 °C, otherwise splaying of parts results. The resin is stabilized by incorporating up to 1 % of a hindered phenol antioxidant with a thioester synergist.

4.4. Styrene – Maleimide Copolymers

For applications where heat resistance above the levels available from SMA resins is desirable, copolymers of styrene and maleimides can be used. These polymers are produced by copolymerizing styrene with a maleimide comonomer (e.g., *N*-phenylmaleimide) or by imidizing SMA with ammonia or an amine.

Polymerization of styrene and maleimides can be accomplished by any of the methods used for SAN. Imidization of SMA is achieved by injecting ammonia or an amine into a vented extruder, or by direct addition to the polymerization [66, 68].

SMI resins are used in a similar manner as SMA. They are rubber modified, glass reinforced, alloyed, or polymerized with termonomers such as acrylonitrile or acrylates. Unlike SMA, they are thermally stable at high temperatures and do not splay.

4.5. Styrene – Acrylate Copolymers

A recyclable, water-soluble copolymer of styrene and acrylate is available from Belland AG [69]. This copolymer contains carboxyl side groups which give it solubility in basic aqueous media. It is recoverable by precipitation via acidification. End uses include typical coatings applications for packaging laminates, book binding glues, and hygiene products.

5. Acrylonitrile – Butadiene – Styrene (ABS) Polymers

5.1. Definition and Structure

ABS [*9003-56-9*] and HIPS (Chap. 1) represent the industrially most important thermoplastic two-phase systems with an amorphous structure. The ABS polymers are based on three monomers: acrylonitrile (A), butadiene (B), and styrene (S). The polymer components have different chemical compositions and coexist as two separate phases whose compatibility is controlled by their structure and chemical microstructure.

In all classical ABS molding compounds the continuous phase (matrix) consists of copolymers of styrene (or an alkylstyrene) and acrylonitrile. An elastomer based on butadiene forms the disperse phase that is distributed in the continuous phase and has a characteristic morphology (Fig. 9).

ABS is mainly produced via two major industrial processes: Emulsion-ABS by compounding of emulsion-made rubbers into the SAN matrix; and solution-ABS via dissolving a rubber (commonly: butadiene homopolymer) in styrene – acrylonitrile and subsequent bulk polymerization, similar to the HIPS process (see Section 2.1). However, process combinations also exist.

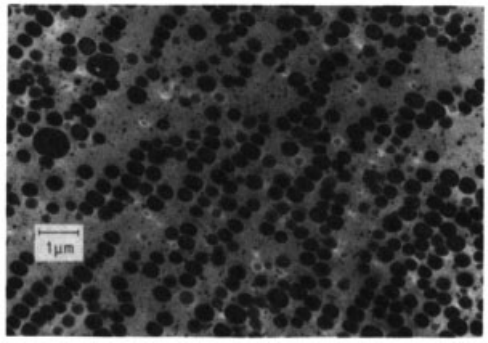

Figure 9. ABS graft polymer showing rubber particles embedded in a styrene – acrylonitrile (SAN) copolymer matrix

Attempts to synthesize styrene-co-acrylonitrile-block-butadiene copolymers via living free radical polymerization have not resulted in commercial products so far [70].

5.1.1. Historical Aspects

The history of ABS polymers began in the mid-1940s. In attempts to produce bulletproof plastic sheets during the last years of World War II, polymer systems were developed from special butadiene – acrylonitrile copolymers and styrene – acrylonitrile copolymers with high molecular masses. These materials had a high impact resistance on account of their low thermoplastic flow, but could only be processed with extruders. Semifinished products (i.e., predominantly sheets, profiles, and pipes) were the first

molded parts to be made from ABS polymers and had dull or matt surfaces.

A drastic increase in the flow of the material allowed processing by injection molding and thus opened up the way for the production of engineering plastics. Improved processibility was obtained by the use of graft polymerization. This technique allowed the production of molded parts with glossy or high-gloss surfaces.

Systematic research into ABS polymers has led to the development of other multiphase plastics with a chemically different make-up. These contain other monomer units in addition to or as a replacement for acrylonitrile, butadiene, or styrene.

Recent developments include modifications of the A, B, and S monomer units as well as complex blend systems, in which the ABS polymers are components of polymer mixtures with an extremely wide range of properties. See also, → Polymer Blends.

5.1.2. Structural Principles

A knowledge of the structural principles of ABS polymers and the parameters on which they depend is required in order to understand production processes, their relationship to product properties, and the basis for the development of other ABS polymers.

The most important physical properties for the user are good thermoplastic flow, high rigidity, satisfactory heat resistance, high surface gloss (emulsion-ABS), and good (notched) impact strength down to temperatures as low as ca. − 40 °C.

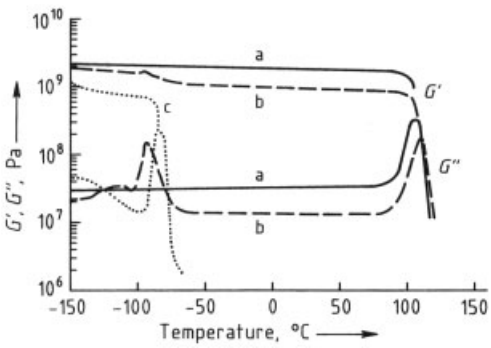

Figure 10. Temperature dependence of the complex shear modulus of ABS graft polymer and its polymer components
a) SAN copolymer (matrix); b) ABS graft polymer; c) Polybutadiene (elastomeric graft base)

The simultaneous occurrence of high toughness and high rigidity results from proper bonding between the two polymer phases [71 – 78].

The elastic properties disappear on cooling to just above the glass transition temperature of the elastomer phase. The two-phase system then behaves as a brittle quasi-single-phase system.

The effect of the elastomer phase dispersed in the resin matrix is seen in plots of the modulus of elasticity as a function of temperature (Fig. 10) and of the stress – strain behavior at constant temperature (Fig. 11) [75, 76].

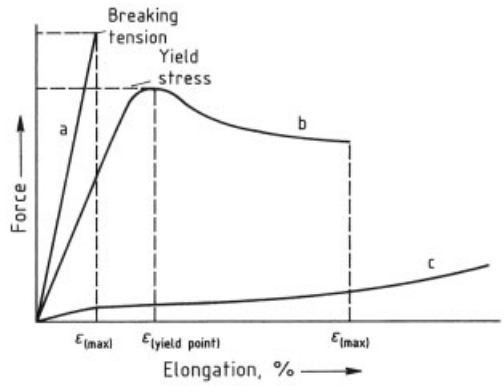

Figure 11. Stress – strain diagram for ABS graft polymer and its polymer components
a) SAN copolymer (matrix); b) ABS graft polymer; c) Polybutadiene (elastomeric graft base)

In Figure 11 the ABS polymer (b) absorbs significantly more energy during deformation up to the breaking point than the SAN resin (a). This is responsible for the high toughness of ABS polymers. The increased deformability in comparison with the pure SAN copolymer and the increased strain in comparison with the pure elastomer component (c) are due to yield processes,in which the applied energy causes the formation of wedge-shaped microcavities (crazes, Fig. 12 A) or initiates shear yielding of the matrix copolymer chains (Fig. 12 B). Both deformation processes are initiated by mechanical loading of the elastomer particles dispersed in the matrix.

The existence of more than one phase is therefore a prerequisite for high toughness. The deformation mechanisms and the effect of different elastomer particles and matrix parameters can be

investigated by electron microscopy techniques [79 – 84].

The glass transition temperature (T_g) depends on the stress frequency (it increases by ca. 30 – 40 °C after impact). The rubber used as the elastomer phase should therefore have as low a T_g as possible. Apart from special rubbers (e.g., silicone rubbers) polybutadiene is the most suitable. Polybutadiene (T_g – 85 °C) guarantees good low-temperature properties. Copolymers of butadiene with styrene (SBR) or with acrylonitrile (NBR) have much higher T_g values and are mainly used in special types of ABS.

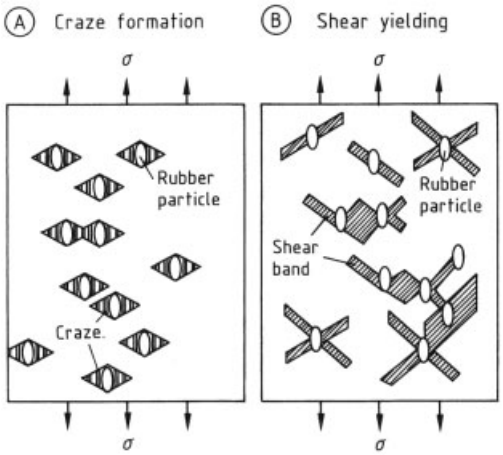

Figure 12. Deformation mechanisms (schematic) in ABS polymers

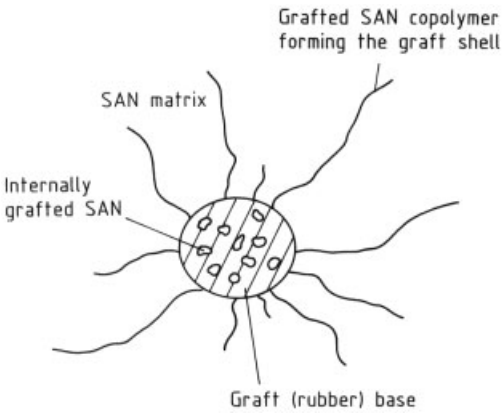

Figure 13. Schematic structure of ABS graft polymers

Tough ABS resins require sufficient coupling between the elastomer particles and the resin matrix to ensure that forces can be transferred across the phase interfaces (phase coupling). Optimal phase coupling between the elastomer particles and the matrix is achieved by graft polymerization (Fig. 13). The graft shell surrounding the elastomer particles is synthesized (1) from the same monomers that are used to synthesize the matrix resin or (2) from monomers that produce a copolymer which is miscible (compatible) with the matrix.

Various polymerization processes (e.g., emulsion, solution, bulk suspension, and bulk processes) can be used to manufacture ABS molding compounds. These processes are described in detail elsewhere, → Polymerization Processes. The producer can also decide whether to use a single-step process involving reaction of polybutadiene and the styrene and acrylonitrile monomers (solution-ABS), or whether to produce the grafted polybutadiene and SAN separately (emulsion-ABS). The second route allows wider variation of individual parameters such as morphology of the elastomer phase (average particle size [85], and degree of cross-linking [75]). The styrene and acrylonitrile can also be copolymerized in different weight ratios [86], and the average molecular mass of the SAN copolymer modified as desired.

Styrene can be replaced partially or completely by α-methylstyrene (AMS). For special requirements (e.g., transparency or high heat resistance), terpolymers are also produced. Monomers used for this purpose, in addition to styrene and acrylonitrile, include methyl methacrylate, maleic anhydride, and *N*-phenylmaleimide.

5.1.3. Synthesis of the Two-Phase Structure

ABS polymers can be divided into ABS blend types and graft types.

In *ABS blend types*, a rubber copolymer is blended with a SAN copolymer. The rubber copolymer contains a matrix monomer (e.g., acrylonitrile in NBR rubber) or a monomer whose solubility parameters resemble those of the matrix [87 – 89]. These ABS systems have very little importance today.

Graft ABS polymers contain a polybutadiene graft copolymer as the elastomer component. The graft shell is constructed from the same monomer units as the resin matrix to give optimal phase interaction between the cross-linked rubber particles and the thermoplastic matrix [90, 91].

Grafting Reaction. The radical-initiated grafting of vinyl monomers (e.g., styrene, acrylonitrile, or methyl methacrylate) only produces satisfactory graft yields with a few rubbers. Industrially, butadiene polymers are used almost exclusively as the rubber (graft) base, polybutadiene being the most common because of its low glass transition temperature.

Grafting of vinyl monomers (e.g., styrene) on polybutadiene can take place by two mechanisms [74]:

1) Addition to the double bond

$$-CH_2-CH=CH-CH_2- + R^{\bullet} \longrightarrow -CH_2-\overset{\overset{R}{|}}{C}H-\overset{\bullet}{C}H-CH_2- \quad (1)$$

2) Abstraction of hydrogen.

$$-CH_2-CH=CH-CH_2 + R^{\bullet} \longrightarrow \quad (2)$$

$$RH + -\overset{\bullet}{C}H-CH=CH-CH_2- \rightleftharpoons -CH=CH-\overset{\bullet}{C}H-CH_2-$$

where R· is a polymer or initiator radical.

The polybutadiene radicals (R′ ·) formed according to Equation (1) or (2) react further with monomeric styrene to give the graft copolymer (3):

$$R'^{\bullet} + H_2C=\underset{\underset{C_6H_5}{|}}{CH} \longrightarrow R'-CH_2-\underset{\underset{C_6H_5}{|}}{\overset{\bullet}{C}H} \quad (3)$$

The allyl radical (Eq. 2) is not very reactive because of its high resonance stabilization. Its resonance energy is comparable with that of a styrene radical [92, 93]. Common polybutadiene rubbers are composed of 1,4-*cis*, 1,4-*trans* and 1,2-vinyl units. Grafting occurs mostly via abstraction of the allylic hydrogen atom of the 1,2-vinyl group.

Recently a copolymerization mechanism has also been proposed for the grafting reaction [94, 95].

The kinetics of the radical grafting reaction largely correspond to the normal kinetics of radical polymerization. At low rubber concentrations deviations occur because of cross-linking reactions. Grafting cannot be carried out with 2,2′-azobisisobutyronitrile as an initiator. The resonance stabilization of the $(CH_3)_2(CN)C\cdot$ radical prevents H-abstraction from the rubber backbone [96]. In the thermally initiated grafting of styrene onto polybutadiene the graft yield increases with increasing temperature and with the 1,2-vinyl group content of the graft base [97].

The degree of grafting can be varied within wide limits by factors such as the choice of polymerization process, the graft base : monomer ratio, the degree of conversion, the activating and regulating systems, and the use of redox initiators based on organic peroxides [98].

Cross-Linking of the Rubber Phase. The high hardness and toughness of ABS polymers requires controlled cross-linking of the rubber phase. If cross-linking is insufficient, the rubber particles dispersed in the SAN matrix (particle size ca. 100 – 1000 nm) are destroyed by the action of shear forces during processing. Rubber which is not cross-linked has no elasticity. The rubber may be cross-linked during the production of the graft base (emulsion polymerization), during the grafting reaction, or in the subsequent workup (bulk polymerization). The swelling index in a good solvent is usually used as a measure of the degree of cross-linking [99]. Quantitative relationships between the degree of swelling and the cross-linking density have also been established [100].

5.1.4. Properties of the Resin Matrix

The resin matrix of ABS polymers is mainly responsible for properties such as processibility and heat resistance of the molding compounds produced from it. Processibility depends predominantly on molecular mass and molecular mass distribution, whereas heat resistance is mainly determined by the chemical composition (i.e., the monomer units and their arrangement in the polymer chain) and the resulting glass transition temperature.

Table 7 summarizes the important properties of ABS matrix components.

Table 7. Properties of ABS matrix resins

Matrix component	Chemical construction	T_g, °C	Concentrations in the ABS blend, wt %	Heat resistance (Vicat B 120), °C	Other properties
SAN	random copolymer S : AN 80 – 65 : 35 – 20	115	95 – 50	104	decrease in thermostability at high acrylonitrile contents
AMS – AN – S copolymer	random copolymer AMS : S : AN 45 : 35 : 20		95 – 50	108 110	
AMS – AN copolymer	random copolymer AMS : AN ca. 70 : 30	128	95 – 50	117	depolymerization begins at 280 °C
AMS – AN sequence polymer	AMS : AN ca. 70 : 30 high proportion of AMS pairs	140	95 – 50	∼ 130	toughness lower than with random copolymers
Styrene – AN – NPMI terpolymer	random terpolymer S : AN : NPMI ca. 67 : 28 : 5	140	95 – 80	∼ 130	
Styrene – AN – MA terpolymer					high reactivity of the anhydride ring, heat resistance limited

AN = acrylonitrile; AMS = α-methylstyrene; MA = maleic anhydride; NPMI = *N*-phenylmaleimide; S = styrene

Styrene – Acrylonitrile (SAN) Copolymers. See also Chapter 2. Styrene – acrylonitrile copolymer resins with an acrylonitrile content of ca. 20 – 35 wt % are usually used for ABS polymers. The weight-average molecular masses are in the range ca. 50 000 to 180 000; the molecular nonuniformities $U = (\overline{M}_w/\overline{M}_n) - 1$ are ≥ 2 (for emulsion polymers) and between 1 and 2 (for bulk and suspension polymers); the glass transition temperature is ca. 115 °C. SAN resins can be synthesized by separate copolymerization or during graft polymerization. All ABS polymers also contain ungrafted SAN resin formed during graft polymerization.

α-Methylstyrene – Acrylonitrile (AMS – AN) Copolymers. See also Section 4.1. Copolymers of α-methylstyrene and acrylonitrile are closely related to SAN polymers. Their high glass transition temperatures and their chemical behavior, which is similar to that of SAN resins, yield ABS systems with improved heat resistance. Their thermostability decreases with increasing α-methylstyrene content (back cleavage of triads of α-methylstyrene units).

Resin production is carried out by emulsion [101], solution, or bulk polymerization [102]. Products usually have acrylonitrile contents of ca. 20 – 35 wt % with weight-average molecular masses of ca. 50 000 – 150 000.

Special copolymers with a high proportion of α-methylstyrene-containing sequences can be produced by emulsion polymerization and yield ABS polymers with high heat resistance (Vicat B temperature) of up to ca. 130 °C [103].

Other Vinyl Polymers. Copolymers of styrene and methyl methacrylate are used in transparent ABS. 4-Methylstyrene – acrylonitrile copolymers [104], styrene – maleic anhydride copolymers (Section 4.3), polyglutarimides [105], and *N*-phenylmaleimide – styrene – acrylonitrile terpolymers [106] are used for the production of systems that are analogous to ABS but with high heat resistances.

5.2. Structure – Property Relationships

The individual components described in Sections 5.1.3 and 5.1.4 can be used to draw up structure – property relationships for ABS polymers which, in turn, can be employed to predict properties from a small amount of basic data. The physical behavior of ABS polymers is partially based on the properties of the matrix and the rubber component.

The room-temperature toughness of ABS polymers generally passes through a maximum as the proportion of the rubber phase increases (Fig. 14). In parallel to this the modulus of elasticity, heat resistance, and flow properties decrease [107]. The low-temperature toughness behaves analogously to the room-temperature toughness; its absolute value, however, is determined by the glass transition temperature of the rubber component. Polymer properties are related to the acrylonitrile content of the resin matrix in the range relevant to industry (i.e., ca. 20 – 35 wt % acrylonitrile). Increasing the acrylonitrile content slightly improves chemical resistance and toughness (maximum at 34 %) but flow properties deteriorate [108].

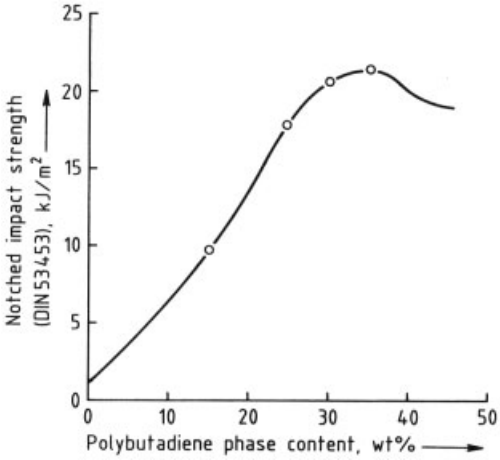

Figure 14. Variation of the toughness of ABS with content of the elastomer (polybutadiene) phase

The most important structural parameters for a given matrix resin and rubber and a constant proportion of the rubber phase are the molecular mass and molecular mass distribution of the matrix resin and the structure of the grafted rubber, especially the particle size, particle size distribution, and degree of grafting and cross-linking.

An increase in the molecular mass of the matrix resin results in an increase in toughness and chemical resistance (environmental stress cracking) but simultaneous deterioration in the thermoplastic flow properties [109]. Broadening of the molecular mass distribution results in easier flow.

The use of cross-linked matrix resins in ABS polymers leads to embrittlement. Even a low degree of cross-linking is detectable because it lowers the surface gloss of molded products.

The properties of ABS polymers are most strongly influenced by the characteristics of the grafted rubber components. Toughness increases with particle diameter.

Grafted rubbers with average particle diameters below ca. 150 nm and a narrow particle size distribution lead to ABS systems which are brittle when subjected to high-velocity impact. In the particle size range of ca. 200 – 300 nm, the toughening effect of the grafted rubber is very strongly dependent on the degree of cross-linking and the cross-linking structure of the rubber base. A good toughening effect can also be achieved with relatively small particles if the rubber is weakly cross-linked. At particle sizes exceeding 300 nm, grafted rubbers with high toughness are generally obtained. However, with very large, compact particles (emulsion polymers with average particle sizes $\geq$ ca. 450 nm) toughness may decrease due to insufficient coupling to the matrix.

ABS molding compounds produced by bulk polymerization (particle size ca. 1 – 10 µm) generally have enhanced toughness because of their high proportion of grafting and matrix inclusion within the rubber phase (rubber phase volume).

Thermoplastic flow properties and the surface gloss of ABS polymers are also strongly dependent on the grafted rubber. Increasing particle size for a given rubber content (i.e., smaller number of particles) improves processibility [110], but leads to loss of surface gloss as a result of light scattering [111].

The degree of grafting, the graft density, and the length of the grafted chains of the grafted rubber also affect the properties of the ABS polymers [112]. An increasing degree of grafting improves coupling to the resin matrix resulting in a more uniform distribution within the matrix and a higher toughness. Grafted rubbers with a low degree of grafting tend to agglomerate so that the properties of the molded articles depend increasingly on processing conditions.

At a fixed degree of grafting, the length of the grafted chain and thus the graft density can be varied. Low graft density corresponds to a small number of very long grafted chains and a high graft density to very many short chains. A fine balance during the grafting reaction allows adjustment of the special requirements of ABS molding compounds (e.g., high toughness, improved flow properties).

5.3. Production of ABS Polymers

ABS can be produced by emulsion polymerization, bulk polymerization, or combined processes (Fig. 15).

There are two main types of *emulsion polymerization* which yield ABS graft rubbers with a rubber content of 5 – 85 wt %.

1) Separate production of the ABS graft rubber and SAN matrix with a subsequent mixing and compounding step

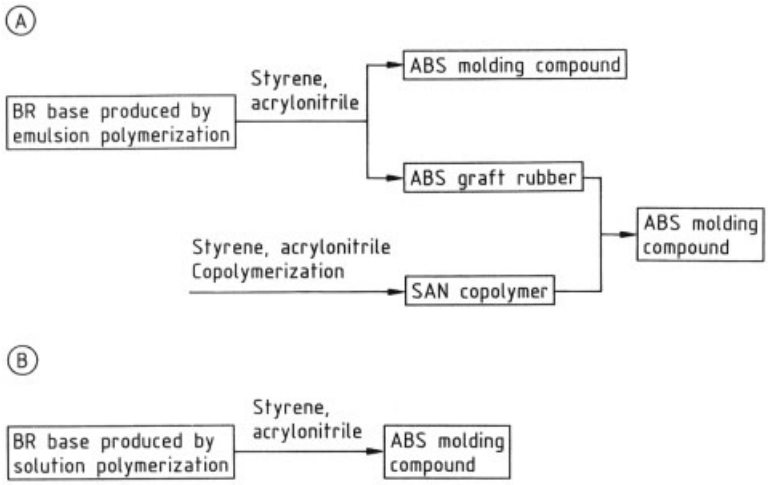

Figure 15. Overview of ABS production processes
A) Emulsion polymerization; B) Bulk polymerization

2) Total emulsion process, i.e., production of the molding compound containing both the grafted rubber and the SAN matrix by emulsion polymerization (the synthesis of a polybutadiene base precedes the ABS grafted polymer production)

Bulk polymerization starts from a separately produced rubber base (BR or SB solution rubbers). The compound is produced in a bulk process run continuously over several stages. Rubber content is limited to ≤ 20 wt % for viscosity reasons.

The *combined processes* include, e.g., the addition of emulsion graft rubber to bulk ABS molding compounds

The separate production of the grafted rubber and the SAN matrix has the advantage that the properties of each component can be controlled independently and different production processes can be used for the grafted rubber phase and the SAN matrix. SAN matrix copolymers are produced almost exclusively by thermally or radically initiated bulk polymerization [113, 114]. The two-stage bulk suspension process is also still used.

Other styrene copolymers besides SAN can also be combined with the ABS graft rubber (see Section 5.1.4). α-Methylstyrene – acrylonitrile (AMS – AN) copolymers or styrene – acrylonitrile – *N*-phenylmaleimide terpolymers, for example, give products with a high heat resistance because of their high glass transition temperatures [115, 116]. Some copolymers are only accessible via the separate bulk polymerization route [117]. Bulk or emulsion polymerization is used for AMS – AN copolymers [118, 119]. The bulk process has the advantage of an inherent degassing stage, so that products with very low residual monomer contents and thus high heat resistance can be obtained. However, yields and molecular masses are limited. The advantages of the bulk process lie in the fact that completely continuous processing is possible and that the grafted rubber particles are formed in situ, i.e., at the same time as SAN polymerization. The proportion of internal grafting is thus higher [120, 121] and products have a more favorable notched impact strength – hardness relationship than those produced by emulsion polymerization. The bulk ABS process includes a degassing step which means that unreacted monomers can be directly recycled.

5.3.1. ABS Production by Emulsion Polymerization

The emulsion process is the standard ABS production process but the principle of this method has also been applied to other rubber-modified, two-phase styrene copolymer systems [122 – 124].

The production of grafted rubbers with high rubber contents and their mixing with a separately produced SAN matrix is used worldwide.

The total emulsion process is of minor importance and is not described here.

The production of ABS grafted rubber consists of three stages:

1) Production of the graft base, i.e., a polybutadiene or butadiene copolymer latex with defined particle size, particle size distribution, and gel content
2) Grafting the styrene – acrylonitrile monomer mixture onto the graft base
3) Workup of the grafted rubber latex

The particle size, particle size distribution, and cross-linking density of the graft base; the degree of grafting; the proportion and morphology of the rubber-phase volume; and the molecular mass of the grafted SAN chains largely determine the properties of the resulting grafted rubber system. Bimodal ABS systems have optimal surface properties and toughness synergies. These systems are produced from a graft base with a bimodal particle size distribution, or from two grafted rubbers of differing particle size that are mixed at the latex stage and then worked up together. The particle size of ABS emulsion grafted rubbers lies in the range 50 – 600 nm, values of 100 – 400 nm being preferred. Figure 16 shows the production of uni- and bimodal ABS graft rubber systems.

Production of the Graft Base. The production of graft bases mostly takes place batchwise. The butadiene is polymerized at 5 – 75 °C and $\leq$12 bar. In the case of butadiene copolymers, the reactivities of the individual monomers must be taken into account. If compositions of the copolymers are chosen which lie outside the "azeotrope region" of the monomer mixture, then high chemical nonuniformity is avoided by using a continuous process and carrying out the polymerization until a given monomer conversion is reached. Unreacted butadiene is fed back into the first reactor of the reactor cascade.

Monomer – water weight ratios of 1 : 0.6 to 1 : 2 are used. Anionic emulsifiers are added in quantities of 1 – 5 parts per 100 parts of monomer to disperse the monomers and stabilize the polybutadiene latex particles. The alkali salts of disproportionated abietic acid or fatty acids, C_{12} – C_{18} *n*-alkyl- and *n*-alkylarylsulfonates are commonly used as emulsifiers. Emulsifiers are dosed such that the resulting latices have a surface tension of < 0.065 N/m and the particles are extensively coated with emulsifier. Emulsifier systems are preferred which are also used for the subsequent grafting reaction and which readily allow coagulation of the dispersed system on addition of acid or electrolytes to facilitate product isolation.

Polymerization is initiated with alkali peroxydisulfates or combinations of reducing agents with peroxydisulfates or organic (hydro)peroxides. The use of (hydro)peroxide systems allows a sufficiently high reaction rate at low temperature and modification of the polybutadiene microstructure [125].

The particle size can be controlled by varying the monomer – water ratio, by stepwise addition of emulsifier, and by adjusting the reaction temperature. A much more elegant method involves the use of seed latex technology, mostly in a semibatch process [126].

Polybutadiene bases with a bimodal particle-size distribution can also be produced by chemical or physical agglomeration of latex particles (Table 8). The advantage of this process lies in the fact that latices with a small particle size can be produced much faster than latices with a large particle size.

The molecular mass and cross-linking density of the polybutadiene bases are adjusted by means of modifiers (e.g., *n*-alkylthiols). Increasing quantities of modifier decrease the average polymer chain length.

Polymerization is usually terminated above 80 % conversion to ensure sufficient cross-linking of the rubber (gel contents > 60 %, preferably > 70 %). No shortstopping agents are used (apart from pure reducing agents) to prevent disruption of the subsequent grafting reaction.

The latex product is devolatilized to remove residual monomers by using vacuum and/or steam in a batch reaction, or continuously in a bubble column.

Grafted Rubber Production. A styrene – acrylonitrile monomer mixture is polymerized in the presence of the graft base. An increase in the ratio of monomers to polybutadiene is accompanied by increases in the net reaction rate, in the molecular masses of the grafted chains and of the "free" SAN copolymer, and in the

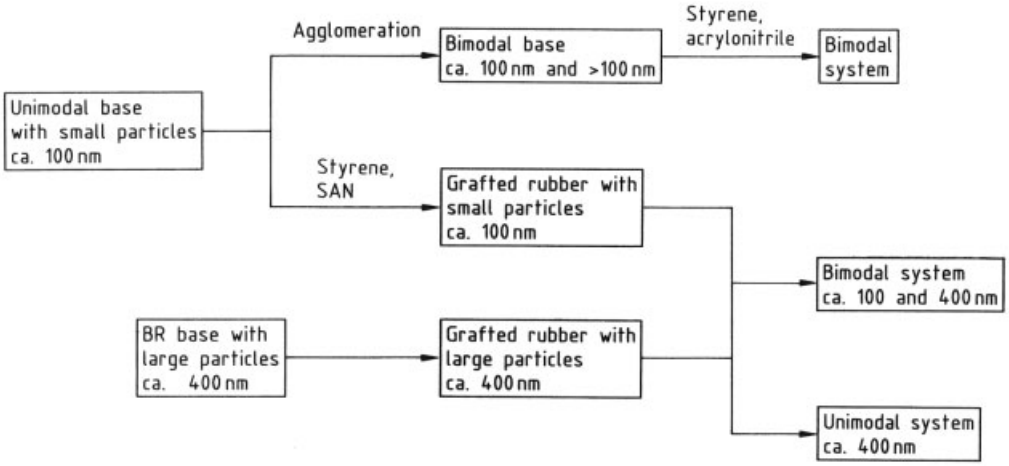

Figure 16. Composition of unimodal and bimodal ABS graft rubbers

Table 8. Production of bimodal ABS graft bases by agglomeration processes

Starting latex (particle size)	Emulsifier	Agglomeration process	Mean final particle size, nm	Reference
BR latex (100 nm)	alkylsulfonate	chemical agglomeration by addition of latices with hydrophilic groups	200 – 500	[127]
Butadiene – acrylonitrile copolymer latices (100 nm)	Na stearate	chemical agglomeration by addition of carboxylic acid anhydrides	200 – 400	[128]
BR latex (100 nm)	Na oleate, cannot be substituted	pressurized agglomeration using Manton – Gaulin homogenizer	200 – 800	[129]
Butadiene – styrene copolymer latex (50 nm)	alkali salts of abietic acid and fatty acids	chemical agglomeration	200	[130]
Butadiene – styrene copolymer latex, polybutadiene	alkali salts of fatty acids and abietic acid	agglomeration by freezing	200 – 400	[131]

degree of grafting. Since the reaction is exothermic ($\Delta H = -840$ kJ/kg), a monomer-feed process or a completely continuous process is used [132, 133]. If the copolymer compositions lie outside the azeotrope region of styrene and acrylonitrile, an increase in chemical nonuniformity is avoided by adding the more rapidly reacting monomer later (staggered dosing). The kinetics of the reaction are described by the Smith – Ewart theory, [134, 135]. Details of the grafting mechanism are given in [136 – 139].

The particle size, particle size distribution, and cross-linking density of the graft bases are directly related to the degree of grafting, graft density, rubber phase volume, and morphology of the grafted rubber [140]. The proportion of internal grafting generally increases with increasing particle size and decreasing cross-linking density of the rubber base. At constant particle size and rubber cross-linking density, the degree of grafting and the graft density become the factors determining the properties of the ABS product [77, 141]. The choice of initiator can influence the ratio of SAN copolymerization to graft reaction and the ratio of external to internal grafting. In comparison with monomer-soluble initiators, water-soluble peroxydisulfates lower the proportion of internal grafting and increase the proportion of free SAN copolymer [77].

The degree of grafting can also be affected by stepwise grafting whereby the formation of the graft shell determines the diffusion rate of the monomers [142].

If the weight ratio of monomers to graft base exceeds unity, molecular mass modifiers are usually added to the monomer mixture. They affect the graft reaction and the styrene – acrylonitrile copolymerization, the molecular masses of the

free and bound SAN chains, and the degree of grafting decrease.

The emulsifier systems used in the grafting reaction are usually the same as those used in the production of the graft base. High latex stability during the polymerization and ready coagulation during the product workup are also desired.

Stabilization and Workup of the Grafted Rubber. After completion of polymerization, excess (hydro)peroxide initiator is reduced with sodium hydroxymethanesulfinite. Stabilizers are added, generally in the form of a synergistically acting system, to protect the grafted rubber from oxidation during subsequent powder drying and compounding. The stabilizers usually comprise one or more phenolic antioxidants and a heat stabilizer. Long-chain alkyl thiodipropionates, long-chain alkyl phosphites, and low molecular SAN polymers with *S*-alkyl terminal groups are particularly effective [143 – 145]. The stabilizer systems should have a low volatility and must fulfil all ecological and toxicological legal requirements. To obtain a homogeneous distribution, the stabilizer system is added to the grafted rubber latex as an aqueous dispersion (0.2 – 1.5 parts by weight/100 parts of polymer) and coprecipitated with the grafted rubber after mixing.

To isolate the ABS graft rubber from the polymer latex, the emulsifier must be deactivated. Emulsifiers based on alkali carboxylates are coagulated and precipitated by adding dilute acid (acetic or mineral acid) to adjust the pH to below 3. Emulsifiers based on alkyl(aryl) sulfonates are coagulated by adding dilute aqueous solutions of inorganic electrolytes (e.g., $MgSO_4 \cdot 7\ H_2O$ or $CaCl_2$).

Strict control of the precipitation conditions is necessary to obtain a powder with a uniform particle structure and size distribution, from which all water-soluble polymerization and coagulation additives can be easily washed out. Coarse powder particles contain inclusions of precipitants which adversely effect drying and processing. Finely divided powders lead to filtration and wastewater problems. Powder structure, particle size, and particle size distribution can be affected by:

1) Consecutive or parallel addition of two or more grafted rubber latices and the precipitating agent [146]
2) The ratio of latex to precipitant solution
3) The temperature at which the precipitating agent and the grafted rubber latex are charged to the coagulation reactor or coagulation cascade
4) The type of stirring used to mix the process streams, the temperature control in the coagulation reactor cascade, and the average residence time

After coagulation the material is filtered with band, plane, or rotary filters and washed. The filter cake is then fed into a reslurry tank and mixed with fresh water. The slurry is processed in filter presses, extruders, or sieve centrifuges and the wastewater recycled to the process. The powder products have water contents of ca. 15 – 40 wt %. They are dried to moisture contents of < 1.0 wt % with two-stage paddle dryers, flow dryers, or vacuum tray dryers.

Alternatively, further processing of the moist grafted rubber can be performed in screw extrusion machines. In this case, dewatering leads to grafted rubber granulate [147] or, after mixing with a thermoplastic resin melt, to solid ABS products [148].

During the workup, measures are taken to lower the residual monomer content to comply with legal limits. Examples include latex degassing with a perforated plate column and countercurrent steam, the use of a precipitation – stripping process, and powder workup with a helical sieve followed by compacting and degassing in a twin-screw extruder [149, 150].

Production Process (Fig. 17). The first stage involves production of a polybutadiene latex with a defined particle size. An aqueous soap solution is formed in the polymerization reactor (a) at ca. 65 °C and the molecular mass modifier is added. After flushing with nitrogen, butadiene is pumped in and the polymerization is initiated with potassium peroxydisulfate solution. The reaction rate is determined by the heat removal capacity of the reactor. The maximum conversion rate is ca. 10 %/h. At conversions of ca. 80 – 90 % the reaction rate decreases; the monomer is then extensively consumed and cross-linking reactions begin. Polymerization is terminated by lowering the temperature to below 50 °C. Unconverted butadiene is vented and stirred out under vacuum.

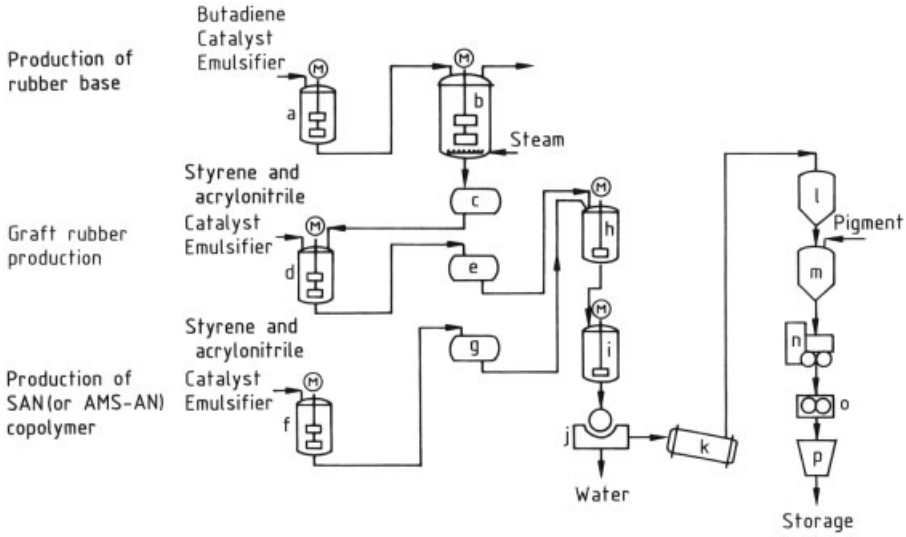

Figure 17. Production of ABS graft rubber and ABS by emulsion polymerization
a) Polybutadiene reactor; b) Degassing reactor; c) Polybutadiene latex tank; d) Grafting reactor; e) Graft latex tank; f) SAN reactor; g) SAN latex tank; h) Latex mixing and stabilization tank; i) Coagulation reactor; j) Filter; k) Dryer; l) Powder silo; m) Premixer; n) Internal kneader; o) Rollers; p) Granulator

In the second stage (the ABS graft rubber production process), the polybutadiene latex, an aqueous potassium persulfate solution, and part of the emulsifier solution are charged to the grafting reactor (d). The temperature is increased to ca. 60 °C and a styrene – acrylonitrile monomer mixture and the remainder of the emulsifier solution are charged to the reactor at a constant rate over 5 h. The reaction temperature is then increased to ca. 65 °C to complete polymerization (conversion $\geq$ 96 %, cycle time ca. 8 h, average reactor size is ca. 40 m^3). After running through the stabilization tank (h), the latex is coagulated. Washing and drying yield a powder which is subjected to a compounding step. After the addition of lubricants, dyes, and other additives, the polymer is pelletized (p).

5.3.2. Production of the Matrix Copolymer

The properties of the matrix copolymers are summarized in Section 5.1.4.

SAN Copolymers. The main ABS matrix components are styrene – acrylonitrile (SAN) copolymers with an acrylonitrile content of 20 – 35 wt % and a molecular mass in the range 50 000 – 180 000.

The production of SAN copolymers can take place by bulk, bulk suspension, or emulsion polymerization. In bulk polymerization the "azeotrope composition" is about 75 wt % styrene and 25 wt % acrylonitrile. In emulsion polymerization the quasi-azeotrope composition is 71.5 wt % styrene and 28.5 wt % acrylonitrile due to monomer partitioning [151].

Chemically and physically uniform SAN copolymers are obtained by *continuous bulk polymerization* (see also Section 2.1) [114, 152]. This also applies to compositions that lie outside the azeotrope region. Compared with suspension and emulsion polymerization, bulk polymerization has environmental advantages: production of wastewater is avoided and virtually all volatile components are removed from the product during workup. A disadvantage is, however, the fact that it is not possible to produce very high molecular masses due to high viscosities. Most of the matrix material is produced by this process.

Uniform SAN copolymers can also be obtained by *bulk suspension polymerization*, within the styrene – acrylonitrile azeotrope range [151]. A SAN prepolymer is first produced which is dispersed in water after addition of water-soluble dispersants and polymerized to give small beads [153]. Unreacted monomers are removed from the beads by means of steam or a

degassing extruder before or after the separation of the polymer – water mixture.

In *emulsion polymerization* (Fig. 17, f) the monomers are dispersed in water using an emulsifier. After addition of an activator dispersion, polymerization occurs to give latex particles (diameter ca. 50 – 200 nm). SAN production by emulsion polymerization is analogous to the production of ABS graft rubbers (see page 26) except that polybutadiene latex is not present.

Semibatch or continuous emulsion polymerization can be used to obtain product compositions with extensive chemical uniformity outside the azeotrope region. The semibatch process leads to molecular nonuniformities of ca. 1.5 – 2. In the continuous process the molecular nonuniformities are in the range 1.5 – 5 depending on the number of reactors and the average residence time. The emulsion-polymerized SAN copolymers show better flow properties and better resistance to stress cracking in the ABS blend compared with bulk-polymerized SAN. The toughness properties of bulk-polymerized SAN copolymers are not, however, achieved. The emulsion process always offers advantages if high molecular masses are desired [154]. The SAN emulsion copolymers are not usually isolated as such, but are mixed with the ABS grafted rubber latex in the latex step (Fig. 17, h). The mixture is then subjected to workup and residual monomer degassing to give an ABS powder blend.

α-Methylstyrene – Acrylonitrile (AMS – AN) Copolymers. ABS grades with a high heat resistance (HT-ABS) are produced by partial or complete exchange of the SAN matrix with compatible (AMS – AN) copolymers or α-methylstyrene – styrene – acrylonitrile terpolymers [155]. AMS – AN copolymers have a higher glass transition temperature and a higher heat resistance in the ABS blend. This system has an azeotropic mixture (69 wt % α-methylstyrene, 31 wt % acrylonitrile) [156].

Like SAN copolymers, AMS – AN copolymers can be produced by bulk and emulsion polymerization. The AMS – AN system has a much lower reaction rate, however, because of the small AMS growth-rate constant. The reaction rate cannot be increased by raising the temperature because depolymerization occurs [157, 158]. Two alternative approaches are used industrially:

1) A drastic increase in the acrylonitrile concentration of the monomer mixture in the starting phase of the reaction
2) Addition of styrene as a third monomer to give terpolymers with glass transition temperatures between those of SAN and AMS – AN copolymers [159]

The equipment requirements for the production of AMS – AN copolymers are similar to those for SAN bulk polymerization. However, higher average residence times are needed and thermal initiation is not possible. Multistep degassing is used to remove residual monomers and to achieve maximum heat resistance.

The most common route for the production of AMS – AN copolymers is semibatch emulsion polymerization. A high monomer purity ($\geq$ 99.5 %) is necessary. Long incubation times can be overcome by using seed latex technology [101]. Process conditions are similar to those of ABS graft rubber production (page 26). To achieve monomer conversions of $>$ 95 %, a pH of $\leq$ 6 must be maintained and an appropriate emulsifier chosen. More details on chemical composition are given in [160].

The emulsion copolymer is usually subjected to workup after latex mixing with ABS graft rubbers. A degassing step must be included between polymer workup and product compounding.

The glass transition temperature of the binary AMS – AN copolymer (ca. 128 °C) can be raised to 140 – 150 °C if the α-methylstyrene concentration at the beginning of emulsion copolymerization is raised above the azeotrope ratio and the reaction is terminated at conversions of $\leq$ 90 %. This increase is due to formation of sequence polymers containing AMS – AMS – AN blocks [161, 162]. The reaction rate and latex stability can be improved by using low concentrations of other monomers [163]. In the ABS blend these AMS – AN sequence polymers produce heat resistance (Vicat B 120) of 130 °C. Toughness diminishes slightly, however, because of incompatibility between the grafted rubber and the matrix.

ABS molding compounds with an AMS – AN matrix have less favorable flow properties than those with a SAN matrix. Although styrene can be replaced by α-methylstyrene in the ABS

graft rubber system, this has not been used industrially for processing reasons [164].

Styrene – Acrylonitrile – NPMI Terpolymers. Incorporation of imide comonomers into the SAN matrix yields heat-resistant ABS polymers [165, 166]. *N*-Phenylmaleimide (NPMI) has been used to produce styrene – NPMI – AN terpolymers [167]. Compare Section 5.3.3.

Other Styrene Copolymers. Attempts have been made to replace the SAN matrix by co- or terpolymers other than those described above. Although styrene can be copolymerized with almost all important vinyl monomers [168], none of the efforts made so far has been used industrially. Attempts to shift the heat distortion temperature of ABS polymers into the Vicat B region 120 – 130 °C by adding styrene – maleic anhydride co- and terpolymers failed because the anhydride ring opened under processing conditions [169, 170].

5.3.3. ABS Production by Bulk Polymerization

ABS production by bulk polymerization (Fig. 15 B) is based on the polymerization of styrene – acrylonitrile mixtures in the presence of a rubber substrate dissolved in this monomer phase. The process can be divided into three steps:

1) The rubber substrate is dissolved in the monomer mixture
2) The rubber – monomer mixture is prepolymerized with continuous mixing (conversion 15 – 30 wt %)
3) The polymer – monomer mixture is polymerized further by bulk suspension polymerization or by continuous bulk polymerization in a high-viscosity reactor cascade

Unlike the emulsion process (Section 5.3.1), no preformed rubber particles are present at the beginning of the grafting reaction. In emulsion polymerization the particle size distribution is determined by the particle size of the grafting base, in bulk polymerization it is mainly determined by the choice of the rubber base and the reaction conditions during prepolymerization and phase inversion. BR rubbers with a glass transition temperature of ≤ -80 °C are usually used as the grafting base. BR rubbers containing ca. 40 % of units with a 1,4-*cis* configuration are preferred because they do not crystallize and the resulting ABS blend has a good low-temperature toughness. The molecular mass and degree of branching are very important for adjusting the particle size and the rubber phase volume in the prepolymerization stage. Long-chain branched BR rubbers with an average molecular mass of 180 000 – 260 000 are commonly used [171]. High glass grades are made from star-branched BR rubbers with low solution viscosity. The rubber content is limited to ca. 20 wt % in ABS bulk processes for viscosity reasons.

The rubber is chopped to pieces and charged to a dissolving vessel. Dissolution in the monomer mixture and optionally a solvent (ethylbenzene, toluene and the like; usually 5 – 15 %) may take up to 10 h. Acrylonitrile acts as a precipitation agent for the butadiene rubber; with increasing acrylonitrile content of the solution less rubber can be dissolved and additional solvents are necessary. The rubber solution is transferred to a feed tank and continuously fed to the reactor cascade.

The prepolymerization stage is performed with continuous stirring. Phase separation and phase inversion begin during prepolymerization and largely determine the size, structure, and particle size distribution of the rubber phase in the resin matrix and thus the properties of the products. In the radical-initiated prepolymerization step, an oil-in-oil emulsion is formed at conversions of ca. 2 % [172 – 174]. Monomer-charged SAN droplets are emulsified in the rubber solution by graft SAN polymer chains which are simultaneously formed on the rubber. With increasing SAN formation the phase-volume ratio changes [175]. When the phase-volume ratio reaches unity, phase inversion begins and the rubber phase is then dispersed in the SAN phase [176], i.e., the SAN – styrene – acrylonitrile monomer phase becomes the continuous phase [177]. SAN is enclosed in the newly formed polybutadiene particles in the form of subinclusions. The phase inversion process, which is initially reversible, is terminated at conversions of ca. 12 – 15 %, depending on the rubber content of the starting solution.

The rubber particle size is determined by three main parameters:

1) The shear forces applied during stirring [178]
2) The viscosity ratio of the two phases [179]
3) The interfacial surface tension between the two phases [180]

Phase inversion can be detected by viscosity measurements. After phase inversion polymerization is continued to a conversion of ca. 25 – 30 %. The viscosity of the polymerizing mass is then already so high that the phase inversion becomes irreversible.

Another important factor affecting ABS production by bulk polymerization is the degree of grafting [181]. Grafting occurs as a result of hydrogen abstraction in the allyl position, grafting by copolymerization has also been discussed (Section 5.1.3) [93]. A first attempt to calculate the maximum possible degree of grafting as a function of conversion is described in [182]. The degree of grafting depends on the nature of the initiators and modifiers, as well as the different solubilities of styrene and acrylonitrile in the rubber and matrix phases [183]. This difference in solubility lowers the degree of grafting and leads to differing acrylonitrile contents in the grafted rubber and the matrix resin.

The molecular mass and molecular mass distribution are principally determined by process parameters (temperature, initiator concentration, solvent, and chain-transfer agents). Thiols are especially effective as molecular mass modifiers in the initial phase of polymerization, chain-transfer agents (e.g., terpinols) are only effective at higher conversions because of their smaller transfer constants.

Another important parameter is the establishment of defined rubber cross-linking densities. The rubber must be highly cross-linked to avoid crack formation within the particles on continuing "hardening out" of the system. If the rubber phase volume is known, the cross-linking density achieved in ABS polymers can be described quantitatively by a modified Flory – Rehner equation [100]. Since the cross-linking density is thermally controlled, this parameter is mainly determined in the degassing stage of the continuous bulk ABS process by variation of the temperature and the residence time [99, 184].

In the continuous bulk solution process, polymerization is carried out in a cascade of two or more reactors [185 186]. Solvents or mineral oils are often added to lower the viscosity and improve heat transfer.

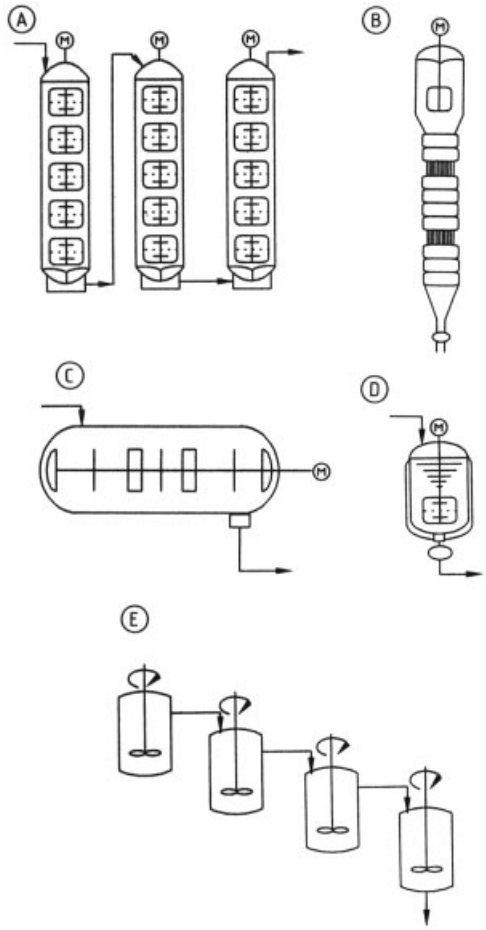

Figure 18. Continuous stirred-tank reactor types [189]
A) Tower cascade; B) Stirred tank – tower combination; C) Ring disk; D) Continuous stirred-tank reactor; E) Stirred-tank cascade

Prepolymerization is performed in tower or stirred-tank reactors. Prepolymerization is followed by polymerization in a high-viscosity reactor or a tower reactor cascade [187]. Producers use specially constructed reactors for bulk solution polymerization (Fig. 18) mostly with laminar flow conditions [190]. The average residence times are ca. 10 – 15 h, and the reaction temperatures 170 – 270 °C [191]. Final conversions of 80 – 85 % are established. Various processes are used to separate the polymer and residual monomer at the end of the bulk poly-

merization stage: strand degassers, degassing extruders, thin-film evaporators, and spray evaporators. Common are two-stage processes involving a flash tank and a second facility.

The SAN matrix of ABS polymers usually has molecular masses in the range 50 000 – 170 000, the molecular nonuniformity $\overline{M}_w/\overline{M}_n$ is ca. 1.5 – 3.0. To achieve maximum toughening, an average rubber particle size of ca. 2.5 μm is aimed at. The particles generally have a cellular structure but other structures (e.g., capsules or lamellae) can be obtained by using polybutadiene or butadiene – styrene block copolymer bases or special phase and reaction conditions in the prepolymerization stage [192 – 194].

Bimodal or multimodal bulk ABS systems are being increasingly described [141, 195, 196]. There are two ways of producing them:

1) Two separately produced prepolymers with different particle sizes and graft structures are mixed with one another before charging to the bulk polymerization stage
2) Two polymers produced in separate prepolymerization and bulk polymerization stages are charged together to a workup zone

In 1990s, intensive research has been carried out on the production of bulk ABS products with a particle size of the rubber phase comparable to that of the emulsion rubber particles ($\leq$ 0.6 μm) [197] with the aim of producing lustrous moldings.

Common to all bulk ABS processes is the fact that the polymers are formed directly as pellets and are also sold in this form for certain uses. More demanding areas of use with defined color requirements need a subsequent compounding stage.

The bulk polymerization process is considered to be more economical than the emulsion and bulk suspension processes because of lower investment costs.

5.3.4. ABS Production by Bulk Suspension Polymerization

After prepolymerization, subsequent polymerization in the suspension process is carried out batchwise in cycles in two or more reactors.

Prepolymerization is started as a batch process. After the end of the phase inversion (conversion 25 – 30 %), the prepolymer is discharged from this reactor and is subjected to further polymerization in a second reactor or reactor cascade. Depending on the initiator system, the polymerization temperature is 80 – 120 °C [188] and the average residence time 2 – 4 h.

In the suspension stage the prepolymer is dispersed in the form of droplets (diameter ca. 0.1 – 1 mm) in an aqueous medium by means of intensive stirring. The water – prepolymer ratio is in the range 3 : 1 to 1 : 1. The suspension is stabilized by adding water-soluble organic polymers, e.g., poly(vinyl alcohol), or colloidal, water-insoluble compounds, e.g., calcium phosphate. Each droplet may be considered as a well-cooled bulk reactor. The bead size depends on the viscosity of the prepolymer syrup, the stirring conditions, the density difference between the organic and the aqueous phase, and the surface tension between the prepolymer syrup and the aqueous medium. The temperature is raised stepwise to 80 – 170 °C. "Oil-soluble" initiators are used. Since the necessary rubber cross-linking is only achieved at conversions of $\geq$80 %, it is important to adopt a definite temperature – time program. The average residence time of the suspension polymerization stage is ca. 10 h. The polymer dispersion is freed from unreacted monomer by the introduction of steam and transferred to a buffer tank. The polymer is separated from the aqueous phase in centrifuges; the polymer beads have a moisture content of $<$ 10 % and are dried with air in a rotating drier.

Production Process (Fig. 19). The BR or SBR rubber bales are cut into pellets (b) and charged to the dissolving vessel (c). The monomer – modifier mixture (a) is added.

The rubber – monomer mixture is transferred to the prepolymerization reactor (e), which is equipped for mixing highly viscous systems. After addition of the initiator, bulk solution polymerization is started by raising the temperature to ca. 90 °C. After ca. 8 h a conversion of ca. 20 – 30 % is reached. (The start of phase inversion is recognized by a sharp increase in the solution viscosity). After lowering the temperature to ca. 60 °C the peroxide initiators are added for the subsequent suspension stage. A ca. 20 % poly(vinyl alcohol) solution is often used as suspension agent (f).

The prepolymer syrup is heated to 120 °C and then pumped into the suspension polymerizer

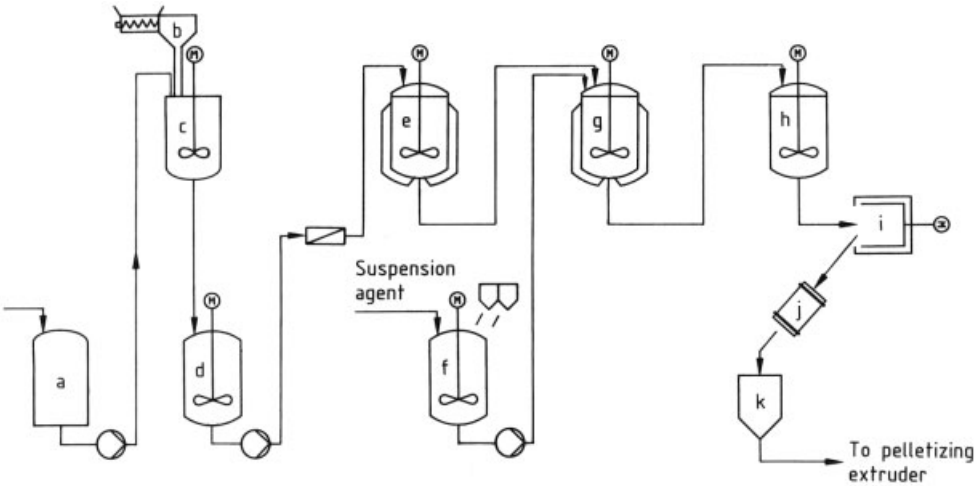

Figure 19. Production of ABS by bulk suspension polymerization [186]
a) Monomer feed; b) Rubber chopper; c) Rubber dissolver; d) Solution storage; e) Prepolymerizer; f) Suspension makeup; g) Suspension polymerizer; h) Suspension buffer tank; i) Centrifuge; j) Dryer; k) Bead storage

(g). The ratios of aqueous phase to prepolymer syrup are chosen to give a solids concentration of ca. 40 %. After 3 – 4 h the temperature is raised to ca. 140 °C and maintained at that level for 3 – 4 h. At strongly decreasing monomer concentrations the required rubber cross-linking is thus obtained.

The resulting polymer suspension is led via a suspension buffer tank (h) into a sieve centrifuge (i) and separated. The aqueous phase is filtered and transferred to a water treatment plant. The polymer beads are dried in the presence of atmospheric oxygen (k) and then compounded.

5.3.5. Process Combinations

Producers have gone over to combining individual processes or process steps to exploit their advantages. Two possibilities are available:

1) The individual components are produced separately and worked up or compounded together
2) Different polymerization processes follow one another and are used to produce one product

Examples of the first alternative are:

1) An ABS polymer produced by emulsion polymerization is mixed (e.g., in a twin-screw extruder, with a SAN copolymer produced by bulk polymerization [198]).
2) ABS graft rubber is added to the workup stage of a continuous bulk SAN process [199].
3) An ABS polymer produced by bulk suspension polymerization is mixed in the compounding stage with ABS graft rubbers obtained by emulsion polymerization [200, 201]. This process can be used to obtain bimodal bigraft ABS systems [77, 202].

Examples of the second alternative include a latex suspension process (which is also used for high-impact polystyrene [203]) and a bulk dispersion process. Latex suspension polymerization can be performed batchwise or continuously, whereas the bulk dispersion process must be run continuously. Neither variant has achieved industrial importance so far. The bulk dispersion process combines the advantages of emulsion polymerization of a rubber base that has a defined particle size and particle size distribution with a bulk grafting step and a bulk SAN polymerization running parallel to it [204].

Like the bulk dispersion process, the latex suspension process starts from a rubber latex produced by emulsion polymerization. The latex is added to a styrene – acrylonitrile monomer mixture and the swollen mixture then converted into a monomer – polymer – water suspension by addition of dispersing and precipitating agents. Conventional bulk suspension polymerization then follows [205].

The main advantages of these two methods over the emulsion process are: there is no precipitation stage, wastewater contamination is reduced, larger particle sizes can be achieved more easily, and less polymerization auxiliaries are required. Advantages over the conventional bulk process are that the rubber chopping step and the production of a rubber solution are omitted. Because of the defined structure of the rubber particles, higher rubber contents can be used without incurring viscosity and heat removal problems. The critical phase inversion stage is circumvented and the process is reduced to a bulk process.

5.3.6. Additives [137]

See also, → Plastics, Additives.

Antioxidants. The rubber phase of ABS polymers is degraded by thermooxidation and light in the presence of atmospheric oxygen. Degradation is thought to proceed via a free-radical mechanism [206], resulting in yellowing, deterioration of mechanical properties, and embrittlement. Stabilizers or stabilizer systems are added at concentrations of 0.1 – 1 wt % to prevent degradation [207, 208] by scavenging radical intermediates or destroying hydroperoxides. They can be divided into sterically hindered phenols, organic compounds with divalent sulfur (thioethers) [143, 209], and organic compounds of trivalent phosphorus (phosphites) [207 – 210].

Light Stabilizers. ABS polymers undergo degradation reactions when exposed to UV light in the wavelength region 250 – 320 nm. They can be protected by UV absorbers, mainly derivatives of 2-hydroxyphenylbenzotriazole, 2-hydroxybenzophenone, and cinnamate esters. Hindered amine light stabilizers (HALS) [211] act by scavenging radicals produced during photooxidation [212].

Optimal protection can be attained by combining UV stabilizers and HALS compounds, generally at levels of ca. 0.2 – 1.5 %.

Lubricants. Long-chain alkyl compounds are often added to ABS polymers as lubricants to improve thermoplastic flow properties and facilitate removal of parts made by injection molding. Fatty acid esters, fatty acid amides, metal stearates (e.g., calcium stearate), or combinations of these substances are used at levels of ca. 0.5 – 2 %.

Antistatic Agents. Like all plastics of low polarity, ABS polymers tend to build up electrostatic charge on their surfaces which causes adverse effects such as static discharge or dust attraction. Polar antistatic agents are added at levels of 0.5 – 3 % to prevent this. They can be divided into cationic, anionic, or nonionic compounds [213, 214].

5.4. Quality Assurance and Standardization

Assurance of uniform quality is important both for the raw materials and the final products. A quality assurance scheme is shown in Figure 20. General guidelines on quality assurance, quality assurance programs, and quality audits are laid down in the EC standard EN 29 001 [215]. The test procedures include:

1) Chemical and physical analysis and methods of determination [216, 217]
2) Product-specific approval tests

Analytical procedures are based on appropriate DIN, ISO, and internal company or client-specific standards. They relate to monitoring of the production process and the final products.

Approval tests are based on the properties of the end product. They include mechanical properties (e.g., notched impact strength, hardness), thermal properties (e.g., heat distortion temperature), processing behavior (e.g., flow properties, dispersing), and fire behavior. They are carried out according to relevant national or international standards.

ABS plastics with an elastomer phase based on butadiene and that do not contain fibrous fillers are standardized according to ASTM D 1788–81, DIN 16 772 part 1, and ISO/DIS 2580–1. The large number of ABS types are classified in a grid system (cells), based on three or four properties, each with several ranges of values.

ASTM D 1788–81	
Izod impact strength at 23 °C (D 256)	6
Deflection temperature under load 264 psi (D 648)	6
Tensile stress at yield point (D 638)	6
Density (D 792)	1.0 – 1.2
Acrylonitrile content, %	min. 13 %
DIN 16 772 Parts 1 and 2 (Sept. 87, Dec. 88)	
Acrylonitrile content of the continuous phase	2
Vicat softening temperature VST B/50	4
Melt flow index, 220/10	4
Notched impact strength at 23 °C	5
ISO 2580–1 (1990)	
Acrylonitrile content of the continuous phase	2
Vicat softening temperature	4
Melt flow index, MFI 220/10	4
Notched impact strength at 23 °C	5
Modulus of elasticity from the bending test (ISO 178)	4

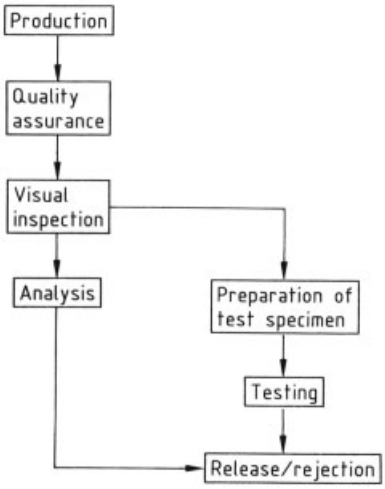

Figure 20. Quality assurance scheme
Visual inspection: Testing the pellets for impurities, pellet shape, and foreign bodies
Analysis: Chemical and physical analysis
Preparation of test specimen: Preparation under standardized conditions and processing parameters
Testing: Product-specific properties
Release/rejection: Authorized by quality assurance department according to prescribed specifications and guidelines

5.5. Properties

ABS polymers have a high toughness (even in the cold), satisfactory rigidity, and good resistance to heat, chemicals, and environmental stress cracking. Molded articles with high dimensional stability and good surface quality can be produced by simple processing techniques. Although each of these properties can be bettered by other thermoplastics, no other system displays such a good combination of technically important properties.

A deeper knowledge of structure – property relationships [77, 218, 219] allows producers to supply a range of commercial products with user-oriented property profiles. Table 9 gives an overview of some commercial grades [220].

The properties of ABS polymers are determined by molecular and morphological parameters. The matrix composition and molecular mass, the type of rubber, the volume ratio of the rubber to the continuous phase, the rubber particle size, the grafted rubber structure, and the additive content are also important.

Standard ABS systems are opaque because their two phases have different refractive indices. Opacity depends on the particle size of the grafted rubber phase and the difference between the refractive indices of the two phases (Mie's theory) [221]. If the particle diameter is small enough compared with the wavelength of visible light, the two phases can be considered as optically homogeneous, and they appear translucent. Completely transparent systems are obtained if the refractive indices are equal. This is achieved, for example, by using a MABS resin matrix [222] or by completely substituting methyl methacrylate for acrylonitrile [223] (see also Section 5.12.2). In bimodal graft rubber systems, depending on ratio of the graft base to the grafted polymer, opacity may be observed despite equal refractive indices. The same applies if grafted rubber particles accumulate to form clusters during processing or tempering because of an incompletely closed graft shell.

In all ABS polymers the continuous phase (SAN and/or AMS – AN copolymer) is responsible for most of the chemical properties. Since only C–C bonds are present in the polymer chains, hydrolysis reactions are unimportant. ABS polymers are therefore generally resistant to aqueous solutions of salts, acids, or bases. They can absorb up to 1.5 wt % water on storage in aqueous media due to residual emulsifier and the polarity of the nitrile side groups. Under normal conditions, dimensional or property changes are negligible.

Table 9. Physical properties of various ABS types

Modulus of elasticity, N/mm^2	Yield stress, N/mm^2	Izod notched impact strength, kJ/m^2		Ball indentation hardness, N/mm^2	Vicat softening temperature, °C	Melt flow rate cm^3/10 min
(ISO 178)	(ISO R 527)	(ISO 180, 80 × 10 × 4 mm)		(ISO 2039)	(ISO 306)	(ISO 1133)
		23 °C	− 30 °C			
Injection molding and standard grades						
2500	45	10	4	120	95	45
2400	45	15	7	115	96	31
2200	40	22	8	100	98	17
2100	40	29	12	90	97	5
1700	34	38	25	70	96	3
2400	44	16	5	115	93	37
Heat-resistant grades						
2500	47	18	8	105	103	8
2500	53	15	6	110	108	3
2700	53	15	5	115	113	2
Galvanizable grades						
2100	40	24	10	90	95	25
2800	55	12	5	115	111	4
Flameproofed grades						
2400	45	8	5	110	87	45
2300	43	17	6	105	88	19
Extrusion grades						
2400	44	19	10	105	98	7
1900	38	31	19	80	96	4
Special grades						
2300	49	30	10	105	105	5
2900	50	5	4	130	98	50
1900	39	35	22	78	98	12
2000	38	12	6	90	94	10

Paraffinic hydrocarbons do not dissolve ABS polymers. A certain weight gain can be detected on storage in such media, and depends on the nature and quantity of the rubber phase. ABS polymers show equally good resistance to animal and vegetable fats and to a range of cosmetic creams.

Other organic liquids, such as halogenated hydrocarbons, aromatics, esters, and ketones dissolve the SAN phase. Oxidizing agents, especially inorganic acids, break up the chains and thus degrade the polymer. The resistance tables published by ABS producers only have limited meaning for practical applications. In order to answer specific questions, the running of a creep test under constant deformation, or, better, under constant tension is recommended. States of stress exist in each molded article as a result of its production. Additional stress is often produced by binding elements (e.g., rivets, screws, molded-in or enclosed metal parts) or on use of the article. Environmental stress cracking is low compared with polystyrene, but cannot always be excluded. Sensitivity towards stress cracking can be further lowered by increasing the acrylonitrile content and the molecular mass of the SAN phase, as well as by relaxation processes that are favored by the disperse rubber phase.

The double bonds in the elastic rubber phase (BR rubber) are responsible for the relatively high sensitivity of ABS polymers towards the long-term effects of heat, light, and weathering which result in yellowing and graying of the surface and a decrease in toughness [224]. Although satisfactory protection can be provided by using stabilizer combinations (phenolic antioxidant, a thiodipropionate ester, and/or a phosphite ester) during production and processing at 220 – 260 °C, long-term stabilization, particularly during simultaneous exposure to UV light and weathering, is difficult. Use of established stabilizer systems allows, however, repeated processing (e.g., recycling) without a serious decrease in toughness. For all indoor applications, ABS is aging-resistant. For outdoor use, adequate resistance to weathering is necessary. Three routes are currently used for this:

1) "Pigmenting" with soot (dark grey shades and black) to absorb UV radiation.

2) Coating semifinished or finished articles with barrier layers having a low oxygen permeability, with less sensitive materials, or with UV-protective paints.
3) Using grafted rubber systems with few or no double bonds. Although these systems do not contain the butadiene monomer unit they are nevertheless members of the ABS family because they have the same basic structure and morphology. They are SAN copolymers that are rendered impact-resistant by a dispersed elastomer phase.

At processing temperatures above 280 °C, ABS polymers lose their toughness because of damage of the rubber phase. Depolymerization begins above 300 °C. In the presence of atmospheric oxygen, decomposition can begin at ca. 280 °C.

5.6. Special Product Modifications

Heat-Resistant ABS Polymers. To increase heat resistance the SAN matrix is generally replaced totally or partially by an AMS – AN copolymer matrix. This is associated with small reductions in toughness, flow properties, and surface quality [225, 226]. To counteract these disadvantages, several methods are available to the producer:

1) Use of an AMS – AN sequence polymer matrix with an increased proportion of AMS – AMS – AN structural units (Vicat B values of ca. 130 °C can be reached) [161, 162]
2) Use of a matrix consisting of a styrene – acrylonitrile – *N*-phenylmaleimide terpolymer (Vicat B values of 130 °C)
3) Use of blends with polymers having higher heat resistance, e.g., polycarbonates or poly (butylene terephthalate) [227 – 229]

Commercial high-temperature ABS (HT-ABS) grades may either have good toughness with only a limited increase in the Vicat values or a limited toughness and Vicat values in the range 115 – 130 °C.

Glass-Fiber-Reinforced ABS Polymers. Reinforcement of ABS with 15 – 20 wt % glass fiber increases the modulus of elasticity by a factor of 2 – 2.5 [230]. The reinforced material has a lower thermal expansion coefficient (and thus high dimensional stability) as well as high rigidity, hardness, and heat resistance. Finished articles with smooth surfaces can then be obtained by appropriate coordination of the flow properties, glass fiber length, and glass fiber sizing material.

Flameproofed ABS Polymers. Standard ABS plastics are classified as flammable according to ASTM D 635 (rate of combustion 5 – 7 cm/min) and as class HB according to the UL 94 test. They also pass the Motor Vehicle Safety Standards (MVSS) test 302 and the so-called Kleinbrenner-Test (limited burning test) DIN 4102 B 2 (see also → Flame Retardants, Chap. 2). For certain applications addition of flame retardants is necessary (mostly halogen compounds in combination with antimony trioxide) to obtain UL 94 V-0 grades.

The brominated diphenyl ethers used in the past as flame retardants have been replaced by other compounds (e.g., brominated oligostyrenes and polycarbonates).

5.7. Legal Aspects

ABS polymers are subject to a large number of legal regulations during production, processing, use, and disposal. A few examples are given here and they relate to the situation in Germany. Similar regulations apply in other countries.

The monomer building blocks A, B, S, and AMS are classified as hazardous substances [231]. Contact with these monomers requires observance of maximum workplace concentrations. See also, → Acrylonitrile, Chap. 9; → Butadiene, Section 8.3; → Nitriles, Section 2.5; → Styrene.

Emissions of ABS production plants are controlled by clean air regulations (TA-Luft) and the law governing wastewater discharge (Abwasserabgabengesetz). Waste materials are subject to the waste disposal law (Abfallbeseitigungsgesetz).

In the processing of ABS pellets, traces of volatile components are formed, which may include residual monomers. As in ABS production, the maximum workplace concentrations must be observed.

Table 10. Product development of acrylonitrile based styrene copolymers (ABS, SAN, ASA, etc.) (10^3 t)

Region	1995	2000	2005
EU	650	800	800
NAFTA	700	800	800
Asia	2300	3400	4700
Others	200	250	300
Total	3850	5250	6600

Finished articles based on ABS polymers can be used in industry without limitation. EC guidelines govern toys and contact with foods [232]. Similar guidelines apply in other countries. These stipulate the monomers permitted for plastics production, polymerization auxiliaries, and additives. They also prescribe limiting values for the use of individual components and for their maximum transfer from the plastic into foods and food substitutes. The maximum proportion of heavy metals in pigments is an important criterion for toys [233].

In the area of disposal, restrictive legal measures are being prepared that prescribe recycling of used components (e.g., in the automotive and electrical sectors).

5.8. Storage and Transportation

ABS pellets can be stored and transported in polyethylene sacks stacked on pellets, in octatainers with polyethylene liners, in big bags, and also in silo containers. Labeling according to GefStoffV and EC guidelines is not necessary. Safety data sheets must be enclosed with the goods for road, rail, or marine transport. The surfaces of the pellets absorb water, which must be removed by predrying (e.g., 2 h at 70 °C) before processing.

ABS powders which are used as additives (e.g., for PVC) must be stored and transported in explosion-proof systems because of their ready oxidizability and large specific surface area. Measures should also be taken against buildup of electrostatic charge. The powders are stored and transported in polyethylene sacks or silo containers. If there is a danger of dust formation, dust masks must be worn.

5.9. Uses

ABS polymers are widely used as engineering materials [234 – 236]. The main consumers are the automotive industry, the domestic appliances industry, the data technology and telecommunications area, and producers of refrigeration equipment, toys, sports articles, and semifinished articles. ABS polymers have a very favorable price – performance ratio. In the automotive industry ABS often competes with modified polypropylenes, but in other applications it also replaces industrial thermoplastics and thermoplastic blends.

Trade names and producers of ABS polymers follow:

BASF	Terluran
Chi Mei	Polylac
Cheil	Starex
Dow Chemical	Magnum
Formosa Plastics	-
GE Plastics	Cycolac
Lanxess	Novodur
LG Chemicals	-
Techno Polymer	-
Toray Industries	Toyolac

5.10. Economic Aspects of Acrylonitrile-Based Styrene Copolymers

The different polymer production and sales options only permit the joint evaluation of data for all acrylonitrile-based styrene copolymers (ABS, SAN, ASA, AES, and ACS). The major part ($> 75\,\%$) is assignable to ABS.

Consumption of acrylonitrile-based styrene copolymers has grown considerably during the last years (Table 10) reflecting their wide use as engineering materials and blend components for other industrial thermoplastics. Nominal production capacities and producers are listed in Table 11. Consumption is estimated to grow by 5 – 6 % over the next years.

Table 11. Worldwide production capacities for acrylonitrile-based styrene copolymers in 2006 (10^6 t)

Company	Region	Capacity
Chi Mei	Asia	1.2
BASF	Asia, EU, NAFTA	0.9
LG Chem	Asia	0.75
Lanxess	Asia, EU, NAFTA	0.7
GE Plastics	Asia, EU, NAFTA	0.6
Formosa Chemicals	Asia	0.45
Cheil Industries	Asia	0.4
Dow Chemical	Asia, EU, NAFTA	0.4
Techno Polymer	Asia	0.3
Grand Pacific	Asia	0.25
Kumho	Asia	0.2
Jilin Chemical	Asia	0.2
others		1.2

5.11. Recycling

ABS articles can be recycled in two ways:

1) The used articles are returned and used directly in the production process (direct recycling)
2) They are degraded chemically into their individual components (i.e., acrylonitrile, butadiene, styrene) or other organic compounds which are then recycled (chemical recycling) [237]

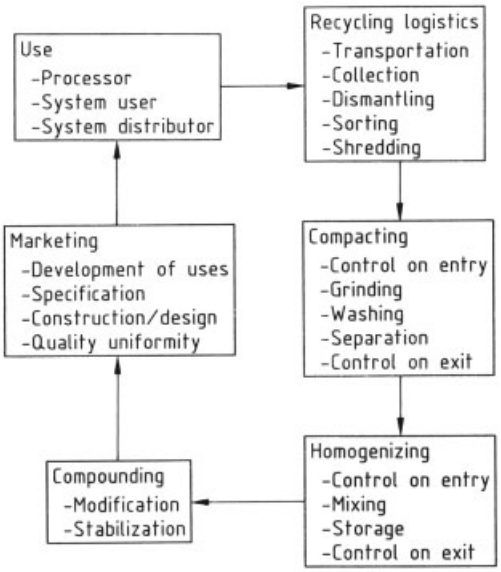

Figure 21. ABS recycling operations
Impact strength of ASA, AES, and ABS

The interest of ABS producers and consumers is concentrated on the first method. Investigations carried out by raw materials producers and the automotive industry show that sorted ABS polymers are well suited to direct recycling (i.e., defined quantities of recycled material are added to the primary material) [238, 239]. Tests have shown that incorporation of 20 % and 50 % recycled ABS derived from articles sorted into pure grades has very little effect on (notched) impact strength, heat resistance (Vicat B 120), and the melt flow index [240]. The basic operations involved in ABS recycling are summarized in Figure 21.

5.12. ABS-Analogous Systems

ABS-analogous systems are elastomer-modified, two-phase thermoplastic systems with the same basic structure as ABS (i.e., a SAN-grafted rubber phase embedded in a SAN matrix). Production methods and principles are the same as those used as for ABS.

5.12.1. ASA, AES, and ACS Polymers

ASA, AES, and ACS polymers consist of acrylonitrile and styrene grafted onto acrylate, EPDM, or chlorinated polyethylene rubbers, respectively. In these polymers, aging resistance is increased by substituting the polybutadiene rubber of ABS by a rubber base which is only slightly or not at all susceptible to oxidation. Other features of ABS (e.g., toughness – hardness and toughness – heat resistance behavior or the coupling behavior of the grafted rubber) should be retained. These conditions can, however, only be partly satisfied because appropriate "saturated" elastomer candidates behave less favorably than butadiene polymers; they have a higher glass transition temperature and are less prone to grafting reactions. The SAN matrix itself does not have unlimited aging and weather resistance; SAN copolymers undergo oxidative degradation in the presence of air and this is independent of the degradation behavior of the grafted rubber system [241]. Addition of light stabilizers can delay, but not prevent degradation. To obtain rubber-modified, two-phase plastics that are completely aging- and

weather-resistant, another matrix system such as poly(methyl methacrylate) must be used.

Acrylate rubbers (ACM), ethylene – propene rubbers (EPM or EPDM) (→ Rubber, 3. Synthetic), chlorinated polyethylene, and other oxidation-resistant rubbers (e.g., silicone rubbers) have been suggested for the production of elastomer-modified, two-phase plastic systems [242, 243]. However, only the first three of these elastomers have been used in aging-resistant ABS polymers. Only ASA [244], AES [245], and ACS polymers are of industrial importance.

ASA production starts from a poly(butyl acrylate) rubber produced by emulsion polymerization and with a uni- or bimodal particle size distribution. To achieve a sufficiently high degree of grafting in the subsequent grafting stage, one of the following grafting bases is used:

1) A terpolymer from *n*-butyl acrylate, vinyl methyl ether, and butadiene [246]
2) An *n*-butyl acrylate – butadiene copolymer [247]
3) A copolymer consisting of butyl acrylate and a monomer acting as a cross-linking or grafting center [248]

Production and workup of ASA graft rubber obtained by emulsion polymerization require the same process parameters and reaction conditions as those described for ABS in Section 5.3.1.

Seed latex technology offers a further possibility, starting from a finely divided polybutadiene seed latex [249].

ASA products are obtained by compounding the grafted polymers and separately produced SAN copolymers with additives and colorants.

AES production usually starts with an EPDM rubber system with 7 – 12 double bonds for every 1000 carbon atoms [250]. At a rubber content of 10 – 50 % the AES and grafted rubber are produced by continuous bulk solution polymerization as described for ABS in Section 5.3.3 [251]. The rubber is dissolved in a monomer – solvent mixture and then charged to the first reactor of a bulk solution polymerization plant. The final product can be isolated analogously to ABS but a precipitation stripping process can also be used for grafted rubber systems with high EPDM contents.

ACS production is analogous to AES production, but is declining in importance.

Table 12 lists producers and trade names of ASA, AES, and ACS polymers.

ASA and AES polymers exhibit better aging behavior than ABS polymers (Fig. 22). The principal disadvantages of AES and ASA polymers compared with ABS polymers are their lower low-temperature toughness and lower heat resistance. Deeply colored modifications are difficult to produce (increased opacity due to the different refractive indices of the rubber base and the matrix).

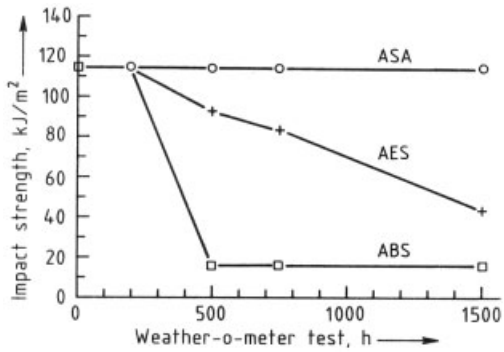

Figure 22. Graft rubber systems after exposure to light (weatherometer test)

5.12.2. MBS and MABS Polymers

MBS polymers contain a methyl methacrylate (M) – styrene (S) copolymer matrix. The grafted rubber system embedded in the matrix consists of a methyl methacrylate – styrene copolymer shell grafted onto a polybutadiene or butadiene – styrene copolymer core. MABS polymers denote systems in which acrylonitrile is incorporated into the polymer matrix or the grafted rubber shell as an additional monomer. Table 13 lists producers and trade names of MBS and MABS polymers.

The composition of MBS and MABS polymers is usually chosen to ensure good transparency: the refractive indices of the matrix, the grafted rubber shell, and the rubber core should be the same; the grafted rubber particle diameter should be less than half the wavelength of visible light (i.e., < 200 nm).

Table 12. Producers and trade names of two-phase ASA, AES, and ACS plastics

Abbreviation	Matrix	Elastomer base	T_g of the elastomer base, °C	Producer	Trade name
ASA	SAN	acrylate rubber	ca. − 48	BASF	Luran S
				Lanxess	Centrex
				GE Plastics	Geloy
				Hitachi Chemicals	Vitax
				UMG ABS Ltd.	Dialac
AES	SAN	EPDM rubber	− 50 to − 60	Dow Chemical	Rovel
				Polimeri	Koblend
				Techno Polymer	AES
				Nippon A& L	Unibrite
ACS	SAN	chlorinated polyethylene	− 20 to − 30	Asahi Kasei	Stylac ACS

Table 13. Examples of producers and trade names of ABS, MBS, MABS, and MACR modifiers

Producer	Country	Product	Modifier for *	Trade name
BASF	Germany	ABS, MBS, MACR	PVC	Vinuran
Lanxess	Germany	ABS, MBS, MABS, MACR	PVC	Baymod
GE Plastics	United States	ABS, MABS	PVC, PC/PBT	Blendex
Rohm & Haas	United States	MBS, MABS	PVC, PC/PBT	Acryloid/Paraloid
Kaneka	Japan	MBS, MABS	PVC	Kane ACE
Mitsubishi Rayon	Japan	MBS, MABS	PVC, PBT	Metablen
Nippon A & L	Japan	ABS, MBS, MABS	PVC	Kralastic

* PBT = poly(butylene terephthalate); PC = polycarbonate; PETP = poly(ethylene terephthalate); PVC = poly(vinyl chloride).

MBS and MABS polymers are seldom used in the pure form; they are generally employed to modify other thermoplastics. They improve processing behavior, notched impact strength, and heat resistance of rigid poly(vinyl chloride) (PVC). The modifiers must be in powder form because rigid PVC is mostly processed as a powder. MBS and MABS polymers are therefore produced by emulsion polymerization or latex suspension polymerization.

MABS polymers used to modify PVC have the following requirements [252]:

1) The particle diameter of the grafted rubber must be < 200 nm to ensure high transparency [253]
2) To obtain transparent, colorless polymer blends, the composition of the modifier must be chosen such that its refractive index is the same as that of PVC ($n_D^{20} = 1.54$) [254, 255]
3) The modifier should have high activity at low dosage to minimize loss in modulus and heat resistance

A butadiene – styrene copolymer (weight ratio 75 : 25) is usually used as the rubber base to achieve a refractive index n_D^{20} of 1.54. Since methyl methacrylate is hydrolyzed relatively rapidly at alkaline pH [256], the grafting reaction is carried out at neutral or weakly acidic pH. Long-chain alkyl or arylsulfonates are used as emulsifiers and yield stable dispersions even below pH 7. Inorganic peroxides (e.g., potassium peroxydisulfate) and redox systems (e.g., cumene hydroperoxide and sodium hydroxymethane-sulfinite) are generally used as initiators [257]. The degree of grafting is controlled by adjusting the ratio of the grafted monomers to the rubber base and by the addition of molecular mass modifiers. The particle size of the grafted rubber is determined by that of the rubber base [120]. This, in turn, is controlled by the water – monomer ratio used in the polymerization, by the emulsifier concentration, or, in seed latex polymerization, by the particle size of the seed latex.

It is advantageous to carry out the grafting reaction in several stages [257, 258], or to use blends of grafted polymers which have a synergistic effect on the toughness of the PVC blend.

The grafted rubber and resin matrix can be synthesized separately and mixed in a workup [259] or compounding stage. Further workup and powder isolation are similar to those of ABS emulsion polymers. Like ABS polymers, MBS and MABS polymers must be protected from attack by atmospheric oxygen by the addition of stabilizers. MBS polymers are being increasingly used for modifying aromatic polyesters and polyester – polycarbonate blends [260, 261]. Since the processing temperature

Table 14. Overview of ABS blends

Grafted rubber	Matrix component I	Matrix component II	Phase ratio grafted rubber to matrix I/matrix II	Advantages	Uses
ABS	SAN [a] AMS – AN [c]	BPA – PC [b]	0.05 – 0.8/0.95 – 0.2	higher heat resistance, higher toughness than ABS, better processibility and environmental stress cracking behavior than PC	interior car fittings, housings for electronic instruments
ABS	SAN	PBT or PBT – PC [d]	0.25 – 0.45/0.75 – 0.55	higher toughness, higher low-temperature toughness than PBT or PC – PBT	car bumpers, housings
ABS		PVC [e]	0.05 – 0.5/0.95 – 0.55	higher toughness, higher heat resistance, better processibility than PVC, better flame resistance than ABS	containers, housings
ABS	AMS – AN	PVC			
ABS	SAN	PVC			
ABS	SAN	PUR [f]	0.05 – 0.25	higher low-temperature toughness, higher rigidity than PUR	ski boots, other injection-molded articles
ABS (modif.) [g]	SAN	PA 6, PA 66	0.05 – 0.4 0.95 – 0.6	higher toughness, better dimensional stability than PA, or PA 66	car parts

[a] SAN: Styrene – acrylonitrile copolymer, weight ratio S : AN 0.9 – 0.6/0.1 – 0.4;
[b] BPA – PC: Polycarbonate based on bisphenol A (BPA);
[c] AMS – AN: α-Methylstyrene – acrylonitrile copolymer in weight ratio AMS : AN 0.90 – 0.65/0.10 – 0.35;
[d] PBT – PC: Blend of poly(butylene terephthalate) and BPA polycarbonate in weight ratio 0.9 – 0.3/0.1 – 0.7;
[e] PVC: Rigid PVC, i.e., without addition of plasticizer;
[f] TPU: Thermoplastic polyurethane elastomer, preferably based on methylene diphenylene diisocyanate;
[g] ABS modif.: Incorporation of functional groups into the graft shell to achieve coupling to the amide group.

exceeds 280 °C, MACR polymer modifiers are used in which the SBR base is replaced by an acrylate rubber base (see also Table 13). Production requirements for MACR polymers are the same as for ASA and MBS graft rubber systems [262].

5.13. ABS Blends

Polymeric materials with novel properties can be obtained by the combination of known polymers in blends (alloys). Table 14 summarizes ABS-containing polymer blends. See also → Polymer Blends, Section 4.3.

Mixtures of ABS with polycarbonate (PC), poly(butylene terephthalate) (PBT), PBT – polycarbonate (PC), PVC, polyamide and thermoplastic polyurethane (TPU) are industrially important, both with regard to quantity and properties [263 – 265]. Other blends not mentioned in Table 14, (e.g., ABS – PETP and ABS – polysulfone blends) have not yet become important [266]. Blending of polycarbonate with ABS improves heat resistance, processing behavior, low-temperature toughness, and leads to synergistic effects [267, 268].

The combination of ABS – polycarbonate blends with triphenyl phosphate as a flame retardant and polytetrafluoroethylene as an antidripping agent yields molding compounds with a low flammability (UL 94 V-0) that are highly suited as housing materials.

ABS – polycarbonate and ASA – polycarbonate blends are predicted to have growth rates of 7 – 8 % in the next years.

The use of plastics for external automotive parts makes certain demands concerning heat resistance, low-temperature toughness (no spall fracturing), resistance to gasoline, and rigidity. Rubber-modified polymer blends have been developed as a solution to these problems. ABS, PBT, and ABS – PBT – PC blends are widely used. The addition of ABS graft rubbers to PBT improves impact properties [269].

The addition of 5 – 20 wt % of ABS graft rubbers to PVC increases its toughness and thus its possible uses. Opaque and transparent blends can be made depending on the size of the grafted rubber particles. The ABS modifier also improves processing behavior, particularly in the production of calendered films [270].

Addition of heat-resistant ABS polymers to rigid PVC (exchange of the SAN matrix for an AMS – AN or styrene – acrylonitrile – NPMI copolymer) increases the Vicat softening temperature and improves toughness. The addition of PVC to ABS polymers improves their

flammability behavior and, in combination with antimony trioxide, gives UL 94 V-0 materials. These blends are used on a large scale in the United States, Japan, and the Pacific Rim.

Crashpad sheets are used for covering car dashboards. They contain an ABS – PVC blend as the main component and their flexibility can be altered by extending the blend system with NBR (butadiene – acrylonitrile copolymer) or other polymer plasticizers.

The addition of ABS graft rubbers to thermoplastic polyurethanes in an in-situ reaction [271] can improve their rigidity and low-temperature toughness. Blends of this type (trade name Cycoloy, GE Plastics or Desmopan, Bayer) are widely used in ski boots. Addition of soft TPU types to ABS polymers allows wide variation of low-temperature toughness and processing behavior [272].

Polyamide 6 and polyamide 66 do not exhibit sufficient compatibility with standard ABS grades. This can be overcome by incorporating functional groups (preferably –COOH groups) into the shell of the ABS graft rubber system or using compatibilizing agents [e.g., imidized poly(methyl methacrylate) [273], or maleic anhydride copolymers [274].

ABS polymers and all other thermoplastics mentioned in Table 14 are produced as powders or pellets. Compounding can be carried out in two ways, usually in internal kneaders or twin-screw extruders [275]:

1) Powder mixing with subsequent melt compounding
2) Direct melt compounding in which the individual components are charged simultaneously or consecutively into the compounding equipment

6. References

General References

1. H. Gausepohl, R. Gellert: "Polystyrol", *Kunststoff Handbuch*, vol. 4, Hanser Verlag, München 1996.
2. J. Scheirs, D. Priddy: "Modern Styrenic Polymers: Polystyrenes and Styrenic Copolymers", John Wiley & Sons, Chichester 2003.

Specific References

3. Phillips Petroleum, US 3 639 517, 1969 (A. G. Kitchen, F. J. Szalla).
4. BASF, US 4 167 545, 1975 (G. Fahrbach, K. Gerberding, E. Seiler, D. Stein).
5. A. W. Hui, A. E. Hamielec, *J. Appl. Polym. Sci.* **16** (1972) 749 – 769.
   K. Kirchner, K. Riederle, *Angew. Makromol. Chem.* **111** (1983) 1 – 16.
   K. S. Khuong, W. H. Jones, W. A. Pryor, K. N. Houk, *J. Am. Chem. Soc.* **127** (2005) 1265 – 1277.
6. J. D. Campbell, F. Teymour, M. Morbidelli, *Macromolecules* **36** (2003) 5491 – 5501.
7. J. Gao, K. D. Hungenberg, A. Penlidis, *Macromol. Symp.* **206** (2004) 509 – 522.
8. N. Overbergh et al., *J. Polym. Sci. Polym. Phys. Ed.* **14** (1976) 1177 – 1186.
9. N. Ishihara et al., *Macromolecules* **19** (1986) 2465 – 2466.
10. A. Zambelli et al., *Makromol. Chem. Rapid. Commun.* **8** (1987) 277 – 279.
11. M. Fischer, G. P. Hellmann, *Macromolecules* **29** (1996) 2498 – 2509.
12. E. R. Moore et al., *Encyl. Polym. Sci. Eng.* **16** (1989) 68.
13. R. A. Bubeck et al., *Polym. Eng. Sci.* **21** (1981) 624 – 633.
14. A. Echte in C. K. Riew (ed.): *Rubber Toughened Plastics*, ACS, Washington, D.C. 1989, p. 23.
15. A. Echte, *Angew. Makromol. Chem.* **58/59** (1977) 175 – 198.
16. G. H. Michler, *Kunststoffe* **81** (1991) 449 – 454.
17. G. H. Michler, *Plaste und Kautschuk* **26** (1979) 680 – 684.
18. A. Echte in [14] pp. 24 – 29.
19. U. Reichert, *Kunststoffe* **80** (1990) 1092 – 1096.
20. J. Le Blanc et al., *Digest of Polym. Developments, Styrenics and Acrylics Series III* **58/59** (1991) May/June, 37.
21. National Polystyrene Recycling Company, company brochure, Washington, D.C. 1989.
22. W. Kaminsky, *Makromol. Chem., Macromol. Symp.* **48/49** (1991) 381 – 393.
23. J. M. G. Cowie: "General Concepts in Addition Polymerization," in *Comprehensive Polymer Science*, vol. 3, part 1, chap. 2, Pergamon Press, Oxford 1989, pp. 1 – 16.
24. D. M. Tirrell: "Copolymer Composition," in *Comprehensive Polymer Science*, vol. 3, part 1, chap. 15, Pergamon Press, Oxford 1989, pp. 195 – 206.

25. R. Greenly in J. Brandrup, E. Immergut (eds.): *Polymer Handbook*, 3rd ed., John Wiley and Sons, New York 1989, pp. II-216, 221, 222, 228.
26. American Cyanamid: *The Chemistry of Acrylonitrile*, The Beacon Press, New York 1957, p. 33.
27. K. F. O'Driscoll: "Temperature Dependence of Copolymerization Reactivity Ratios," *J. Macromol. Sci. Chem.* **A 3** (1969) no. 2, 307 – 309.
28. G. E. Moleau, *J. Polym. Sci., Part B* **3** (1965) 1007 – 1015.
29. C. H. Basdekis: *ABS Plastics*, Reinhold Publishing Corp., New York 1964, p. 47.
30. J. LeBlanc, O. Deex, F. Duston, Y. Phillips in M. Hartung (ed.): *Digest of Polymer Developments—Styrenics and Acrylics, Series 3*, **53** (1989) no. 12, 135 – 145.
31. Monsanto, US 3 747 899, 1971 (L. F. Carter, G. A. Latinen); DE-OS 2 240 294, 1972 (G. A. Latinen).
32. Dow Chemical, US 2 714 101, 1950 (J. L. Amos, J. C. Frank, K. E. Stober); US 2 727 884, 1953 (K. E. Coulter, J. L. McCurdy, D. L. McDonald).
33. *Ullmann* 4th ed., **19**, 274.
34. Plastics, ed. 9, Thermoplastics and Thermosets Desk-Top Data Bank, International Plastics Selector, San Diego 1990, pp. 870 – 877.
35. H. Jenne, *Kunststoffe* **62** (1972) 616; **66** (1976) 581.
36. P. L. Ku: "Polystyrene and Styrene Copolymers; Their Manufacture and Application," *Adv. Polym. Sci.*, **8** (1988) no. 2, 177 – 196.
37. Monsanto Kasei, Collimate M-800, Technical Data Sheet, Yokkaichi, Japan, 1991.
38. I. M. Ward: *Mechanical Properties of Solid Polymers*, 2nd ed., John Wiley & Sons, New York 1983, p. 168.
39. N. W. Johnston, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **14** (1973) no. 1, 46.
40. M. J. Guest, J. H. Daly, *European Polymer J.* **25** (1989) 985.
41. T. E. Evans: "Styrene – Acrylonitrile," in *Modern Plastics Encyclopedia*, McGraw-Hill, New York 1982 – 1983, pp. 116 – 117.
42. F. M. Peng: "Acrylontrile Polymers," in *Encyclopedia of Polymer Science and Technology*, vol. 1, 2nd ed. John Wiley & Sons, New York 1985, pp. 455 – 464.
43. *Plastverarbeiter* **27** (1976) 355 – 367.
44. F. Forster: "Polystyrol," in R. Vieweg, G. Daumiller (eds.): *Kunststoff-Handbuch*, vol. 5, Hanser Verlag, München 1969, p. 230.
45. G. Obieglo, *Kunstst. Ger. Plast.* **69** (1979) no. 7, 8.
46. D. Caldwell, *Modern Plastics Encyclopedia*, vol. **67**, no. 11, McGraw-Hill, New York 1990, pp. 101 – 102.
47. D. R. Paul, J. W. Barlow, H. Keskkula: "Polymer Blends," in *Encyclopedia of Polymer Science and Technology*, 2nd ed., vol. 12, John Wiley and Sons, New York 1978, pp. 399 – 461.
48. D. R. Paul: "Interfacial Agents for Polymer Blends," in D. R. Paul, S. Newman (eds.): *Polymer Blends*, vol. 2, Academic Press, New York 1978, chap. 12.
49. R. Bernstein et al., *Macromolecules* **10** (1977) 681.
50. K. Ogura et al., *Annu. Tech. Conf. Soc. Plast. Eng.* (1987), 1365.
51. W. J. Hall et al., *ACS Symp. Ser.* **229** (1983) 49.
52. J. H. Kim et al., *J. Polym. Sci. Polym. Phys. Ed.* **27** (1989) 223 – 244.
53. P. B. Rim, J. Runt, *J. Appl. Polym. Sci.* **30** (1985) 1545.
54. J. H. Kim et al., *J. Polym. Sci. Polym. Phys. Ed.* **27** (1989) 2211 – 2227.
55. I. Louis Gomez: *High Nitrile Polymers for Beverage Container Applications*, Technomic, Lancaster 1990, pp. 17 – 18.
56. F. M. Peng: "Acrylonitrile Polymers," in *Encyclopedia of Polymer Science and Technology*, 2nd ed., vol. 1, John Wiley & Sons, New York 1985, pp. 455 – 464.
57. B. G. Frushour, R. S. Knorr: "Acrylic Fibers," in M. Lewin, E. M. Pearce (eds.): *Handbook of Fiber Science and Technology*, vol. IV, Marcel Dekker, New York 1985, pp. 228 – 235.
58. Monsanto US 3 540 577, 1967 (Q. A. Trementozzi); DE-OS 2 050 535, 1971 (C. Lee, S. P. Nemphos).
59. Standard Oil, DE-OS 1 929 860, 1970 (J. T. Duke, D. C. Prem); DE-OS 2 007 519, 1970 (J. T. Duke).
60. W. A. Combellick: "Barrier Polymers," in *Encyclopedia of Polymer Science and Technology*, 2nd ed., vol. 2, John Wiley & Sons, New York 1985, pp. 176 – 192.
61. "Plastics Technology," *Manufacturing Handbook and Buyers' Guide*, 1989/1990, Bill Communications, New York 1990, p. 675.
62. J. H. Kim et al., *Polym. Eng. Sci.* **29** (1981) 581.

63. Monsanto, US 4 874 829, 1989 (C. E. Schwier, W. C. Wu).
64. Dow Chemical, US 2 769 804, 1951 (A. W. Hanson);
US 3 080 348, 1963 (A. F. Roche).
65. Richardson Company, US 4 001 484, 1974 (I. H. Song).
66. J. LeBlanc, O. Deex, F. Duston, in S. Englehart (ed.): *Digest of Polymer Developments—Styrenics and Acrylics, Series 3*, **54/55** (1990), May/June 128 – 148.
67. A. Wambach: "Styrene Maleic Anhydride", *Modern Plastics Encyclopedia*, McGraw-Hill, New York 1991, pp. 104 – 106.
68. N. Platzer et al. in M. Hartung (ed.): *Digest of Polymer Developments—Styrenics and Acrylics, Series 3*, **45** (1987/1988) 119 – 141.
69. *Europa-Chemie* **47** (1991) 2.
70. Dow, WO 30944, 36944, 1997.
71. C. B. Bucknall, R. R. Smith, *Polymer* **6** (1965) 437.
72. K. Dinges, *Kunststoffe* **56** (1966) 548.
73. J. Frazer, *Chem. Ind. (London)* **33** (1966) 1397.
74. K. Dinges, H. Schuster, *Makromol. Chem.* **101** (1967) 202.
75. L. Bohn, *Angew. Makromol. Chem.* **20** (1971) 129.
76. L. Morbitzer, K. H. Ott, H. Schuster, D. Kranz, *Angew. Makromol. Chem.* **27** (1972) 57.
77. L. Morbitzer, D. Kranz, G. Humme, K. H. Ott, *J. Appl. Polym. Sci.* **20** (1976) 2691.
78. M. Yokouchi, S. Seto, Y. Kobayashi, *J. Appl. Polym. Sci.* **28** (1983) 2209.
79. J. A. Sauer, C. C. Chen, *Polym. Eng. Sci.* **24** (1984) 786.
80. F. Lednicky, Z. Pelzbauer, *Polym. Test.* **7** (1987) 91.
81. M.-J. Zhang, F.-X. Zhi, X.-R. Su, *Polym. Eng. Sci.* **29** (1989) 1142.
82. C. C. Chen, J. A. Sauer, *J. Appl. Polym. Sci.* **40** (1990) 503.
83. G. Michler, *Plaste Kautsch.* **35** (1988) 347.
84. G. H. Michler, *Kunststoffe* **81** (1991) 449.
85. C. F. Pearsons, E. L. Suck, *Adv. Chem. Sci.* **99** (1971) 340.
86. H. Kim, H. Keskkula, D. R. Paul, *Polymer* **32** (1991) 1447.
87. P. J. Flory: *Principles of Polymer Chemistry*, Cornell University Press, New York 1953.
88. S. Krause, *J. Macromol. Sci. Rev. Macromol. Chem.* **7** (1972) 251.
89. M. T. Shaw, *J. Appl. Polym. Sci.* **18** (1974) 449.
90. B. J. Schmitt, *Angew. Chem.* **91** (1979) 286.
91. K. Dinges, *Kunststoffe* **32** (1979) 748.
92. A. Brydon, G. M. Burnett, G. G. Cameron, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 3255; **12** (1974) 1011.
93. V. K. Gupta, G. S. Bhargava, K. K. Bhattacharyya, *J. Macromol. Sci. Chem.* **A 16** (1981) 1107.
94. R. A. Hayes, S. Futamura, *J. Polym. Sci. Polym. Chem. Ed.* **19** (1981) 985.
95. R. A. Hayes, *J. Polym. Sci. Polym. Chem. Ed.* **19** (1981) 993.
96. H. A. J. Battaerd, G. W. Tregear: *Graft Copolymers*, Wiley Interscience, New York 1970.
97. J. P. Fischer, *Angew. Makromol. Chem.* **33** (1973) 35.
98. E. S. Daniels, V. L. Dimonie, M. S. El-Aasser, J. W. Vanderhoff, *J. Appl. Polym. Sci.* **41** (1990) 2463.
99. K. Mc Creedy, H. Keskkula, *Polymer* **20** (1979) 1155.
100. H. J. Karam, L. Tien, *J. Appl. Polym. Sci.* **30** (1985) 1969.
101. Bayer, EP 331 999, 1989.
102. Bayer, EP 255 889, 1988.
103. N. W. Johnston, *Macromolecules* **6** (1973) 453.
104. Mobil Oil, EP 003 639, 1979.
105. Rohm & Haas, US 4 217 424, 1979.
106. Denki Kagaku Kogyo, US 4 404 322, 1983.
107. M. Bresson, P. Hygounenc, *Rev. Gen. Caoutch. Plast.* **642** (1984) 109.
108. H. Kim, H. Keskkula, D. R. Paul, *Polymer* **31** (1990) 869.
109. J. W. F. Bley, S. A. H. Mohammed, *Polym. Plast. Technol. Eng.* **20** (1983) 161.
110. L. Li, T. Masuda, M. Takahashi, *J. Rheol. (N.Y.)* **34** (1990) 103.
111. F. Lednicky, Z. Pelzbauer, *Angew. Makromol. Chem.* **141** (1986) 151.
112. Y. Aoki, *Macromolecules* **20** (1987) 2208.
113. Bayer, EP 255 889, 1988.
114. Dow Chemical, DE 26 19 969, 1976.
115. Uniroyal, US 3 111 501, 1958.
116. ICI, US 3 652 726, 1969.
117. Monsanto, US 4 374 951, 1983.
118. Japan Synthetic Rubber EP 164 874, 1985.
119. BASF, DE-AS 2 809 180, 1980.
120. G. Kämpf, H. Schuster, *Angew. Makromol. Chem.* **27** (1972) 81.
121. E. R. Wagner, L. M. Robeson, *Rubber Chem. Technol.* **43** (1970) 1129.
122. Bayer, EP 34 748, 1981.
123. H. Ebneth, *Kunststoffe* **53** (1969) 70.
124. T. O. Purcell, *Encycl. Polym. Sci. Technol.*, **suppl. 1**, 1976, 319.

125. H. F. Mark et al., *Encyclopedia of Polymer Science and Technology*, vol. **2**, Wiley Interscience, New York 1965, p. 696.
126. Bayer, EP 394 779, 1990.
127. H. G. Keppler, H. Wesslau, H. Stabenow, *Angew. Makromol. Chem.* **2** (1968) 1.
128. Monsanto, US 3 558 541, 1971.
129. ISR, US 35 73 242, 1971; US 3 573 246, 1971.
130. Chemische Werke Hüls, DE 1 208 879, 1966.
131. B. F. Goodrich, DE-OS 1 494 093, 1969.
132. US Rubber, US 2 820 777, 1955. Röhm, DE 33 19 340, 1984.
133. The International Synthetic Rubber, DE-OS 22 33 287, 1973.
134. W. V. Smith, R. H. Ewart, *J. Chem. Phys.* **16** (1948) 592.
135. J. L. Gardon, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 623, 665, 687, 2853, 2859.
136. M. R. Granico, D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 2617.
137. D. J. Williams, *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2133.
138. P. Keusch, R. A. Graff, D. J. Williams, *Macromolecules* **7, 304** (1974).
139. N. Friis, A. E. Hamielec, *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 3321.
140. Bayer, EP 717 077, 1996.
141. F. Haaf et al., *J. Sci. Ind. Res.* **40** (1981) 659.
142. Japan Synthetic Rubber, JA 3113/67, 1967.
143. Bayer, EP 192 151, 1986; EP 225 510, 1987.
144. R. Gächter, H. Müller: *Taschenbuch der Kunststoff-Additive*, Hanser Verlag, München 1983.
145. J. Stepek, H. Daoust: *Additives for Plastics*, Springer Verlag, New York 1983, p. 167.
146. Bayer, US 4 522 959, 1985; US 4 522 964, 1985.
147. BASF, EP 665 095, 1995.
148. BASF, EP 734 826, 1996.
149. A. K. Ghosh, J. T. Lindt: *International Polymer Processing*, vol. **3**, Hanser Verlag, München 1990, p. 195.
150. A. K. Ghosh, J. T. Lindt, *J. Appl. Polym. Sci.* **39** (1990) 1553.
151. C. C. Lin et al., *J. Appl. Polym. Sci.* **26** (1981) 1327.
152. Monsanto Chemical, US 3 813 369, 1974. Dow Chemical, EP 897 165, 1983.
153. Monsanto, US 3 708 658, 1973.
154. Ugine Kuhlmann, DOS 19 64 915, 1970. Bayer, EP 167 772, 1986.
155. F. E. Karasz, W. J. Macknight, in D. R. Paul, L. H. Sperling (eds.): *Multicomponent Polymer Materials 211 Advances Chemistry Series*, American Chemical Society, Washington, D.C. 1986.
156. A. Rudin, M. C. Samanta, *J. Appl. Polym. Sci.* **24** (1979) 1665.
157. P. Wittmer, *Makromol. Chem.* **103** (1967) 188.
158. P. Wittmer, *Makromol. Chem.* **177** (1976) 991.
159. Borg Warner, US 3 010 936, 1961.
160. H. Lange, H. Baumann, *Angew. Makromol. Chem.* **43** (1975) 167.
161. N. W. Johnston, *Appl. Polym. Symp.* **25** (1974) 19.
162. N. W. Johnston, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **14** (1973) 46.
163. Bayer, EP 330 039, 1989.
164. Borg-Warner, US 2 908 661, 1959.
165. Nitto Chemical, EP 177 031, 1985. Nippon Shokubai K.K., EP 213 933, 1986.
166. Mitsui Toatsu Chemicals, EP 216 524, 1986.
167. Monsanto, US 4 567 233, 1986.
168. C. Walling, *J. Am. Chem. Soc.* **71** (1949) 1930.
169. Monsanto, US 4 305 869, 1981; US 4 197 376, 1980; US 4 223 096, 1980.
170. W. J. Hall, R. L. Kruse, R. A. Mendelson, Q. A. Trementozzi, *Polym. Coat. Appl. Polym. Sci. Prep. Div. Am. Chem. Soc.* **47** (1982) 298.
171. Dow Chemical, EP 103 657, 1984.
172. J. L. White, R. D. Patel, *J. Appl. Polym. Sci.* **19** (1975) 1775.
173. G. E. Molau, *J. Polym. Sci. Part A* **3** (1965) 4235.
174. S. Zhiqiang, Y. Huigen, Z. Yuan, *J. Polym. Sci.* **32** (1986) 3349.
175. G. F. Freeguard, M. Karmarkar, *J. Appl. Polym. Sci.* **15** (1971) 1657.
176. G. E. Molau, H. J. Keskkula, *J. Polym. Sci. Part A* **4** (1966) 1595.
177. B. W. Bender, *J. Appl. Polym. Sci.* **9** (1965) 2887.
178. J. L. Amos, *Polym. Eng. Sci.* **14** (1974) 1.
179. F. D. Rumscheidt, S. G. Mason, *J. Colloid Sci.* **16** (1961) 238.
180. R. W. Flumerfelt, *Ind. Eng. Chem. Fundam.* **11** (1972) 312.
181. M. Baer, *J. Appl. Polym. Sci.* **16** (1972) 1109.
182. S. L. Rosen, *J. Appl. Polym. Sci.* **17** (1973) 1805.
183. J. L. Locatelli, G. Riess, *Makromol. Chem.* **175** (1974) 3523; *J. Polym. Sci. Polym. Chem. Ed.* **11** (1973) 3309.
184. A. E. Platt in: *Encyclopedia of Polymer Science and Technology*, vol. 13, Wiley Interscience, New York 1970, p. 156.

185. N. Platzer: "Design of Continuous and Batch Polymerization Processes," *Ind. Eng. Chem.* **62** (1970) 6.
186. H. G. Pohlemann, A. Echte: "Polymer Science Overview," in G. A. Stahl (ed.): *ACS Symp. Ser.* **175** (1981) 265.
187. Dow Chemical, US 2 727 884, 1955; US 2 694 692, 1954.
188. H. J. Karam, J. C. Bellinger, *Ind. Eng. Chem. Fundam.* **7** (1968) 576.
189. W. H. Ray, *ACS Symp. Ser.* **226** (1983) 101.
190. Toray Ind. K.R., US 3 981 944, 1976.
191. Mitsui Toatsu Chemicals, Toyo Engineering Corporation: *Continuous Buck Polymerization Process for ABS Products*, company brochure, Tokyo 1985.
192. A. M. Donald, E. J. Kramer, *J. Appl. Polym. Sci.* **27** (1982) 3729.
193. C. B. Bucknall, P. Davies, I. K. Partidge, *J. Mater. Sci.* **22** (1987) 1341.
194. D. G. Gilbert, A. M. Donald, *J. Mater. Sci.* **21** (1986) 1819.
195. BASF, US 44 93 922, 1985.
196. BASF, EP 507 117, 1992; Dow Chemical, EP 412 801, 1992.
197. BASF, EP 477 671, 1992; EP 477 764, 1992; EP 505 798, 1992. General Electric, EP 657 479, 1995. Mitsui Toatsu Chemicals, US 5 506 304, 1996.
198. BASF, DE 20 37 984, 1972.
199. Monsanto, US 3 928 495, 1975.
200. Societa Edison, FR 14 56 141, 1966.
201. Monsanto, US 3 509 237, 1971.
202. Bayer, BE 828 318, 1975; BE 828 319, 1975.
203. S. H. Roth, *J. Appl. Polym. Sci.* **18** (1974) 3305.
204. Toray Ind. K.K., DE-AS 2 044 427, 1970.
205. D. E. Bell Richardson, US 3 370 105, 1968.
206. Y. A. Shlyapnikov, *Russ. Chem. Rev. (Engl. Transl.)* **50** (1981) 581.
207. G. Scott, *J. Appl. Polym. Sci. Appl. Polym. Symp.* **35** (1979) 123.
208. C. R. H. I. de Jonge, *Pure Appl. Chem.* **55** (1983) 1637.
209. Pennwalt, US 3 652 680, 1972.
210. K. Schwarzenbach in [144] p. 1.
211. F. Gugumus in [144] p. 101.
212. D. M. Wiles, J. P. P. Jensen, D. J. Carlsson, *Pure Appl. Chem.* **55** (1983) 1651.
213. H. W. Finck in [144] p. 581.
214. BASF, DE 31 12 428, 1982; Bayer, EP 278 347, 1988; Bayer, EP 278 349, 1988.
215. European Standardization Committee, EN 29 001, 1987.
216. G. Lawson, J. A. Sidwell, *Plast. Rubber Int.* **8** (1983) 15.
217. J. R. Fried, *Plast. Eng.* **38** (1982) no. 8, 27.
218. C. B. Bucknall: *Toughened Plastics*, Applied Science Publishers, London 1977.
219. D. R. Paul, S. Newman: "Polymer Blends," in: *Encyclopedia of Polymer Science and Engineering*, 2nd ed., vol. 12, Wiley, New York 1988, p. 399.
220. Bayer AG: *Novodur*, company brochure no. KU 41 662, ed. 12/95.
221. B. F. Conaghan, S. L. Rosen, *Polym. Eng. Sci.* **12** (1972) 134.
222. H. Ebneth, *Plastverarbeiter* **4** (1968) 1.
223. H. Jenne, *Kunststoffe* **66** (1976) 581.
224. E. Zahn, *Appl. Polym. Symp.* **11** (1969) 209. P. Pagan, *Polym. Paint Colour J.* **117** (1987) 704.
225. K. Stange, *Chem. Ind. (Düsseldorf)* **20** (1968) 804.
226. H. Jenne, *Kunststoffe* **62** (1972) 616; **66** (1976) 581.
227. W. Nouvertne, H. Peters, H. Beicher, *Plastverarbeiter* **33** (1982) 1070.
228. W. Witt, *Kunststoffe* **74** (1984) 592.
229. P. R. Müller, *Kunststoffe* **74** (1984) 569.
230. A. S. Wood, *Mod. Plast. Int.* **20** (1990) June, 58.
231. Verordnung über gefährliche Stoffe, Aug. 26, 1986, BGBL I, p. 1470, and from April 23, 1990, BGBL V, p. 790.
232. EG Kunststoffrichtlinie 90/128, EWG no. 349/26 from Dec. 13, 1990.
233. Positiv-Liste für Additive und Polymerisationshilfsstoffe, Synoptic Document REV 5 (III/3141/89-EN), Aug. 01, 1991.
234. K. Schneider, E. Frohberg, *Kunststoffe* **80** (1990) 1099.
235. R. Theyson, *Kunststoffe* **79** (1989) 913.
236. E. Frohberg, *Kunststoffe* **86** (1996) 1498.
237. S. Menges, R. Fischer, *Kunststoffe* **81** (1991) 6.
238. R. D. Leaversuch, *Mod. Plast. Int.* **21** (1991) Feb., 12.
239. M. de Braaf, Dow Chemical, company brochure 1991.
240. Bayer AG: *Kunststoff-Recycling*, company brochure no. KU 40 237, 11/91.
241. G. M. Ruhnke, L. F. Biritz, *Kunststoffe* **62** (1972) 250.
242. Dow Corning, DE 2 321 904, 1973; US 3 898 300, 1975.

243. Bayer, EP 246 537, 1987;
EP 258 746, 1988.
244. E. Zahn, H. W. Otto, *Kunststoffe* **57** (1967) 921.
245. Japan Synthetic Rubber, DE 30 01 766, 1980;
EP 96 527, 1983;
US 4 814 388, 1989.
Montedipe, EP 286 071, 1988.
246. BASF, DE 1 174 069, 1964;
DE 1 180 126, 1964;
DE 1 182 811, 1964;
1 238 207, 1976.
247. Stauffer Chem., US 3 994 631, 1976.
248. Rohm and Haas, US 4 096 202, 1978.
249. Bayer, EP 34 748, 1981.
250. C. L. Meredith, *Rubber Chem. Technol.* **44** (1971) 1130.
251. Japan Synthetic Rubber, DE 30 36 921, 1981.
252. Kanegafuchi Chem., FR 1 356 523, 1964.
253. Kanegafuchi Chem., JA 10 735 67, 1967;
JA 190 535 64, 1964.
254. American Cyanamid, GB 994 924, 1963.
255. Foster Grant, GB 1 009 360, 1962.
256. W. C. Mast, C. H. Fischer, *Ind. Eng. Chem.* **41** (1949) 790.
257. Kanegafuchi Chem., DE-OS 1 570 855, 1965.
258. Japan Geon K.K., GB 1 469 868, 1977.
259. Bayer, DE 38 13 363, 1989;
DE 33 12 541, 1984.
260. Teijin, US 38 64 428, 1975.
261. Montedison, DE 23 48 377, 1975.
262. J. T. Lutz, *Polym. Plast. Technol. Eng.* **21** (1983) 99.
263. W. Witt, *Kunststoffe* **77** (1987) 1009.
264. D. R. Paul, S. Newman: *Polymer Blends*, vols. I and II, Academic Press, New York 1978.
265. K. Kheit, D. V. Hove, *Polymer Blends and Alloys Symposium*, Luzern, June 1988.
266. Uniroyal, US 3 555 119, 1971;
US 3 636 140, 1972.
267. L. Morbitzer, H. J. Kress, C. Lindner, K. H. Ott, *Angew. Makromol. Chem.* **132** (1985) 19.
268. H. Peters, L. Morbitzer: *Plasticon 81 Polymer Blends*, Plastic and Rubber Institute, University of Warwick, Preprint 29.
269. Bayer, US 4 292 233, 1981.
270. Bayer, EP 101 900, 1984;
EP 101 904, 1984;
EP 101 899, 1984.
271. Bayer, DE 28 54 407, 1980.
272. W. J. Farrissey et al. in: *Advances in Polymer Blends and Alloys Technology*, vol. 1, Technomic Publishing Company, Lancaster 1988.
273. Rohm and Haas, US 4 436 871, 1984.
274. M. Weber, W. Heckmann, A. Goeldel, *Macromol. Symp.* **233** (2006) 1 – 10.
275. A. P. Plochocki, S. S. Dagli, H. H. Mack, *Kunststoffe* **78** (1988) 254.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Polyurethanes

Standard Article

Norbert Adam[1], Geza Avar[2], Herbert Blankenheim[3], Wolfgang Friederichs[4], Manfred Giersig[5], Eckehard Weigand[5], Michael Halfmann[6], Friedrich-Wilhelm Wittbecker[6], Donald-Richard Larimer[7], Udo Maier[8], Sven Meyer-Ahrens[9], Karl-Ludwig Noble[10], Hans-Georg Wussow[11]

[1]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[2]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[3]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[4]Bayer MaterialScience AG, Dormagen, Federal Republic of Germany
[5]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[6]Bergische Universität Wuppertal, Wuppertal, Federal Republic of Germany
[7]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[8]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[9]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[10]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany
[11]Bayer MaterialScience AG, Leverkusen, Federal Republic of Germany

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a21_665.pub2
Article Online Posting Date: January 15, 2005

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Basic Reactions
3. Starting Materials
3.1. Polyisocyanates
3.1.1. Aromatic Polyisocyanates
3.1.2. Aliphatic Polyisocyanates
3.1.3. Blocked Isocyanates
3.2. Polyols
3.2.1. Polyether Polyols
3.2.2. Polyester Polyols
3.2.3. Polycarbonate Polyols
3.2.4. Other Polyols
3.3. Diamines and Amino-Terminated Polyethers
3.4. Special Building Blocks
3.5. Catalysts and Additives
4. Structure and Morphology
4.1. Polyurethanes Without Segmented Structure
4.2. Polyurethanes with Segmented Structure
4.2.1. Hard and Soft Segments
4.2.2. Segregation and Morphology
4.3. Cross-linking of Polyurethane

4.4. Polyisocyanurates
5. Production of Polyurethanes
5.1. Stoichiometry
5.2. Reaction without Solvents
5.2.1. One-Shot Process
5.2.2. Prepolymer Processes
5.3. Reaction in Solvents
5.3.1. One-Component Systems
5.3.2. Two-Component Systems
5.4. Reactive One-Pack Systems
5.5. Other Processes
6. Processing of Polyurethanes
6.1. Supply, Storage, and Preparation of Raw Materials
6.2. Metering and Mixing Technology
6.3. Processing Plants
7. Foams
7.1. Flexible Foams
7.1.1. Flexible Slabstock Foam
7.1.1.1. Raw Materials
7.1.1.2. Production
7.1.1.3. Properties
7.1.1.4. Trimming and Processing
7.1.1.5. Applications
7.1.2. Molded Flexible Foam
7.1.2.1. Production
7.1.2.2. Molding Process
7.1.2.3. Properties
7.1.2.4. Applications
7.2. Semirigid Foams
7.2.1. Applications
7.2.2. Production
7.2.3. Properties
7.3. Rigid Foams
7.3.1. Raw Materials
7.3.2. Processing
7.3.3. Properties
7.3.4. Special Types
7.4. Integral Skin Foams and RIM Materials
7.4.1. Applications
7.4.2. Production
7.4.3. Properties
8. Noncellular Polyurethanes
8.1. Cast Elastomers
8.1.1. Applications
8.1.2. Production
8.1.3. Properties
8.2. Thermoplastic Polyurethane Elastomers (TPU)
9. Polyurethane Coatings
10. Polyurethane Adhesives
11. Polyurethane Fibers
12. Polyurethanes and Isocyanates as Binders
13. Special Products
14. Safety and Ecology
14.1. Safety Precautions when Handling the Raw Materials
14.2. Emissions, Accidental Release, and Waste Disposal
14.3. Recycling/Recovery of Polyurethanes
14.4. Fire Performance of Polyurethanes
15. Economic Aspects

[Top of Page]

1. Introduction

Wolfgang Friederichs

Polyurethane (PUR) is the collective name for an extensive group of polymers with very different compositions and correspondingly varied property profiles. All PURs are built on a common principle: they are produced by the polyaddition process [1, 2] of (poly)isocyanates typically with (poly)alcohols. The characteristic chain link is the urethane group. In most cases two or more closely positioned urethane groups link poly(alkylene ether) and/or polyester sequences with molecular masses between ca. 200 and 6000 g/mol. Thus, the urethane group usually is only present to a minor extent.

PURs also include polymers produced from polyisocyanates that have urea, isocyanurate, or carbodiimide groups as their characteristic structural unit. As a result of the increasing importance of di- and polyamines and water as reactants, most of the industrially produced PURs contain urea groups (including the isocyanurates as cyclic ureas) as their main property-determining structural units.

A PUR is usually understood to be the fully reacted finished product (e.g., foams and coatings). These products are free from monomeric isocyanates, and unconsumed isocyanate groups, which may be present in small amounts, are bound to the PUR matrix. In adhesives and paints, the isocyanate-containing precursors are often called “PUR adhesive,” “PUR hardener,” or “PUR paint” to indicate the field of application of the end product produced from them.

The diisocyanate polyaddition process for synthesizing PURs is characterized by the following statements:

1. A large number of different reactants (monomers and oligomers) are available.
2. It is an exothermic reaction that can even take place at room temperature. No byproducts need be separated.
3. The reaction rate can be controlled over a wide range by catalysts, and in some cases a specific reaction can be accelerated by careful choice of the catalyst (e.g., tin compounds catalyze the urethane reaction, tertiary amines catalyze the isocyanate-water reaction; see Section Flexible Slabstock Foam).
4. Intermediates (prepolymers) with reactive end groups can be produced.
5. The stoichiometry influences the molecular mass (see Section Prepolymer Processes).
6. Production and composition can be adapted to various process techniques and extremely diverse product requirements.
7. A large number of end products can be made.

Besides their unique properties, one of the main advantages of PURs is that all types of articles (foams, elastomers, thermosets) can be produced starting from liquid reactive components. This has led to their great industrial importance (see Section One-Shot Process and Chap. Economic Aspects).

PUR products include highly elastic foams (mattresses, cushions, car seats), rigid foams (insulation materials); rigid and flexible moldings with compact skins (window frames, housings, skis, damping units such as car shock absorbers, front- and rear-end body parts, steering wheels, and shoe soles); engineering moldings with a high hardness and elasticity, ski boots, films, hoses, blow-molded parts, noise shields for trucks, seals for stoneware pipes, roller coatings, sealants, grouting compounds, surfacings for sport and play areas, off-road tires, windsurfing equipment, hydrocyclones, fenders, printing rollers, cable sheathing, catheters, high-quality paints, corrosion protection for steel-reinforced concrete, adhesives, textile coatings, high-gloss paper coatings, leather finishes, poromerics, glass fiber sizes, and wool finishing agents. The number and range of applications are constantly growing. The production, properties, technology, and use of PURs are reviewed in [3-13].

[Top of Page]

2. Basic Reactions [3, 4, 8]

Wolfgang Friederichs

The addition reaction of an alcohol to an isocyanate to give an urethane (carbamate) has been known since 1849 (WURTZ):

The extension of this reaction to di- and polyfunctional isocyanates and hydroxy compounds by OTTO BAYER in 1937 led to the principle of the polyaddition reaction resulting in the formation of linear, branched, or cross-linked polymers:

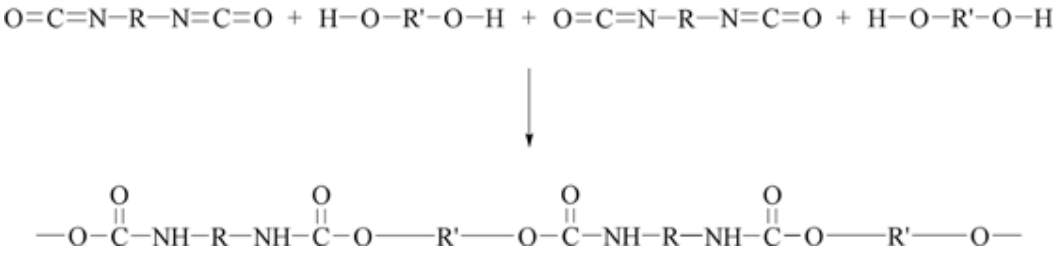

In addition to alcohols, the most important group of NCO-reactive compounds are the amines, which lead to the formation

of ureas:

Disubstituted urea

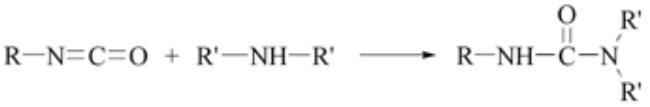

Trisubstituted urea

Ureas are also formed by the reaction of water with isocyanates, in which the carbamic acid formed in the first step of the reaction spontaneously decomposes to an amine with elimination of carbon dioxide. This amine then reacts with excess isocyanate to yield symmetrically substituted ureas:

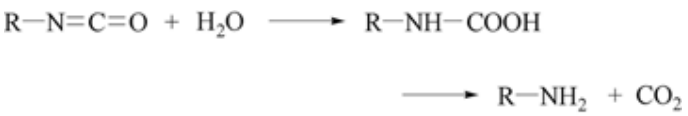

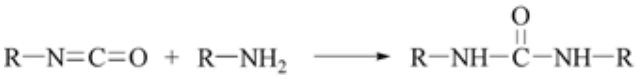

Disubstituted urea

This reaction is the basic reaction leading to PUR foams, especially flexible foams.

Another process which leads to the formation of carbon dioxide is the reaction of isocyanates with organic acids. The reaction product formed in this way is an amide:

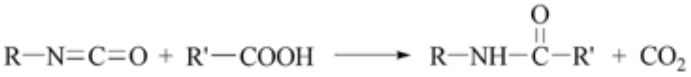

Amide

As the urethanes and ureas formed in the reaction mentioned above still contain reactive hydrogen atoms they can react with excess isocyanate to form allophanates and biurets, respectively.

Allophanate

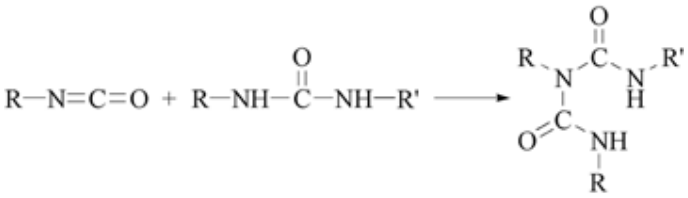

Biuret

These cross-linking reactions are important in the production of PURs and play a key role in the development of polymeric materials with tailor-made properties.

Apart from these polyaddition reactions, the following reactions involving two or more isocyanate groups are also important in PUR chemistry (catalyst in parentheses):

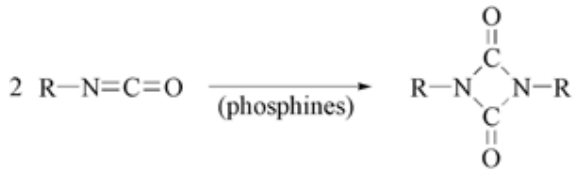

Uretdione

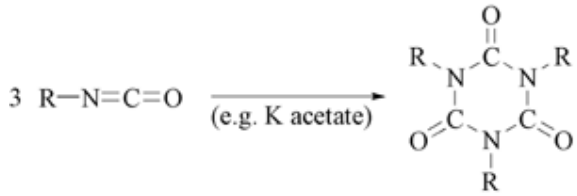

Isocyanurate

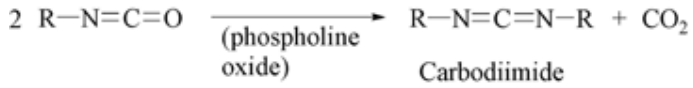

Carbodiimide

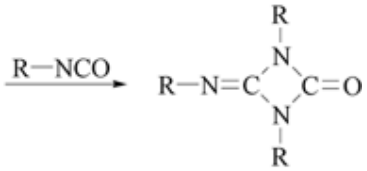

Uretonimine

Due to the high reactivity of the isocyanate group some further reactions are well known but only of minor importance in the production of PURs. Examples are the reactions of amides and epoxides with isocyanates:

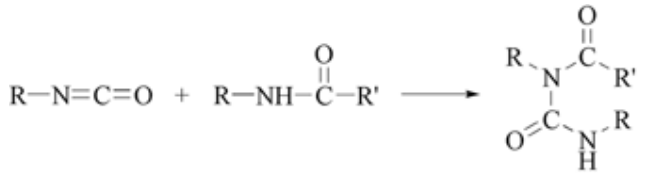

Acylurea

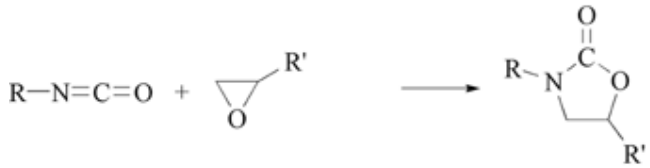

Oxazolidone

The tremendous number of different possible reactions not only play a role in the production of the final PUR polymers but also in the production of modified polyisocyanates (see Section Polyisocyanates) used as raw materials for PURs.

Usually the reactions described above proceed in parallel. For example, the production of foams is based on the combination of the NCO/ROH reaction leading to urethanes and the NCO/H_2O reaction leading to ureas and causing foaming of the reaction mixture by the evolution of carbon dioxide.

Other methods for preparing PURs which are usually not used in industrial applications are the polycondensation of bis-chloroformate esters and diamines, and the reaction of polyamines with carbonylbiscaprolactam (CBC) or polycyclocarbonates.

Thermally stable *N*-alkyl PURs are formed by the reaction of isocyanates with secondary diamines (e.g., piperazine). Since these products are unable to form hydrogen bonds, their properties are different from those of conventional PURs.

[Top of Page]

3. Starting Materials [3]

Wolfgang Friederichs

The range of PUR starting materials developed since the early 1950s is based on light petroleum fractions (ethylene, propene, naphtha). To a minor part renewable raw materials (natural products such as sugar, castor oil, etc.) are used [14-16]. The suitability of natural fatty acid esters [17, 18], lignin [19, 20], polysaccharides [19, 21, 22], and chitin as raw materials has been examined for special applications.

3.1. Polyisocyanates

See also ⟶ Isocyanates, Organic

More than 90 % of PURs are produced from aromatic polyisocyanates. Compared to aliphatic and cycloaliphatic polyisocyanates, the aromatic polyisocyanates show higher reactivities towards hydroxyl compounds and give PURs with better mechanical properties.

The aliphatic and cycloaliphatic polyisocyanates are used to obtain PURs that do not become discolored on exposure to light or heat. This is particularly important for coating materials (paints, varnishes, textile coatings, and glass-fiber sizes). The stability of aliphatic or cycloaliphatic polyisocyanates PURs can be improved by incorporating urea, biuret, or hydrazide groups into the polymer [23-25].

3.1.1. *Aromatic Polyisocyanates*

Aromatic polyisocyanates are primarily used for a wide variety of PUR foamed plastics, elastomers, thermosets, and adhesives.

2,4-Toluenediisocyanate [584-84-9] (TDI), 2,4-diisocyanatotoluene (**1a**), *bp* 121 °C (1.33 kPa), *mp* 21.8 °C.

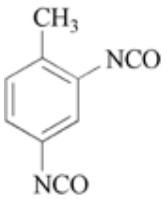

1a

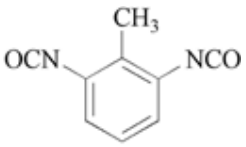

1b

Usually the pure 2,4 isomer or mixtures of 2,4-TDI (**1a**) with 2,6-TDI (**1b**) are used for industrial applications:

TDI 80 (80 % 2,4-TDI, 20 % 2,6-TDI), *bp* 121 °C (1.33 kPa) *mp* 13.5 °C, vapor pressure (25 °C) 3 Pa, equilibrium vapor concentration 30 ppm, MAK value 0.005 ppm (according to TRGS 900).

TDI 65 (65 % 2,4-, 35 % 2,6-), *bp* 121 °C (1.33 kPa), *mp* 5 °C.

1,3-Bis(3-isocyanato-4-methylphenyl)-2,4-di-oxo-1,3-diazetidine [26747-90-0] (**2**) (uretdione of 2,4-TDI, TDI-dimer; Desmodur TT, Bayer MaterialScience), *mp* 153 °C.

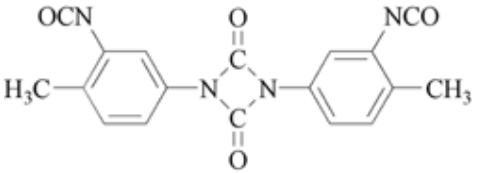

2

Uretdiones react as blocked isocyanates: the four-membered ring opens at elevated temperature to release the NCO groups.

"TDI-urea" [5206-52-0] (**3**) (TDIH, Bayer MaterialScience) *mp* 180–184 °C [26, 27]

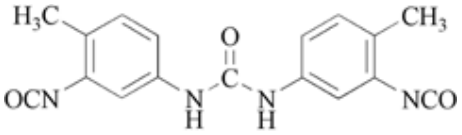

3

1,3,5-Tris(3-isocyanato-4-methylphenyl)-2,4,6-trioxohexahydro-1,3,5-triazine [20649-21-6] (**4**) (TDI-trimer), ca. 51 % in butyl acetate (Desmodur IL, Bayer MaterialScience): NCO content ca. 8 %, viscosity (20 °C) 2000 ± 500 mPa · s.

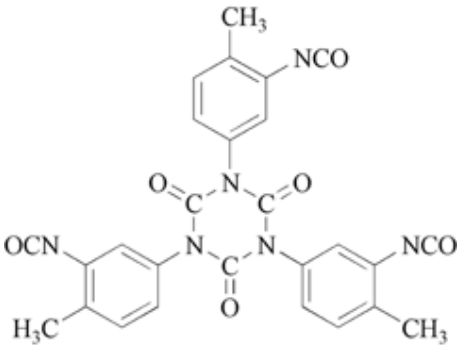

4 (idealized)

Adduct of TDI and trimethylolpropane [21092-47-7] (**5**), 75 % solution in ethyl acetate (Desmodur L, Bayer MaterialScience): NCO content 13 ± 0.5 %, viscosity (20 °C) 2000 ± 500 mPa · s.

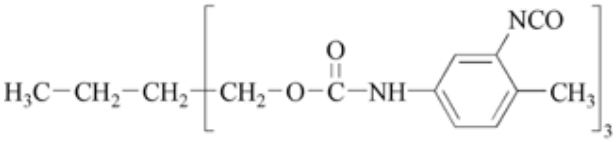

5

Monomeric MDI, 4,4′-Methylene diphenyl diisocyanate [101-68-8], 4,4′-methylene-bis(phenyl isocyanate), bis(4-isocyanatophenyl)methane, 4,4′-diisocyanatodiphenylmethane (**6a**), Monomer-MDI, *bp* 208 °C (1 kPa), *mp* 39.5 °C, vapor pressure (25 °C) <1 mPa, equilibrium vapor concentration 0.009 ppm, MAK value 0.005 ppm (according to TRGS 900)

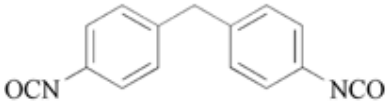

6a

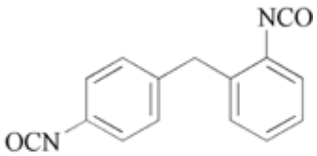

6b

60 % 2,4′-MDI (**6a**), 40 % 4,4′-MDI (**6b**) (Desmodur 1806, Bayer MaterialScience), *mp* 16–17 °C.

Polymeric MDIs (PMDIs) [9016-87-9] (**7**) typically contain 30–70 % dinuclear components (n = 0), 14–40 % trinuclear components (n = 1), and 15–30 % higher nuclear components (n > 1), viscosity 50–20000 mPa · s, NCO content 25–33 %.

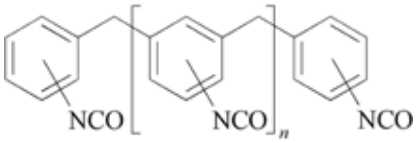

7

Urethanized 4,4′-MDI [60440-22-4] (Desmodur PF, Bayer MaterialScience) *mp* 10–15 °C.

Carbodiimidized 4,4′-MDI [37353-55-2], *mp* 10–15 °C (Desmodur CD, Bayer MaterialScience; for carbodiimidization, see Chap. Basic Reactions).

1,5-Naphthalene diisocyanate (NDI) [3173-72-6], 1,5-diisocyanatonaphthalene (**8**), *mp* 127 °C (Desmodur 15, Bayer MaterialScience).

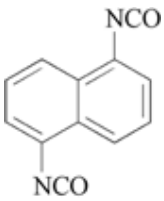

8

Tris(4-isocyanatophenyl)methane [2422-91-5] (**9**), 20 % solution in ethyl acetate, NCO content 7 ± 0.1 %, viscosity (20 °C) ca. 1 mPa · s.

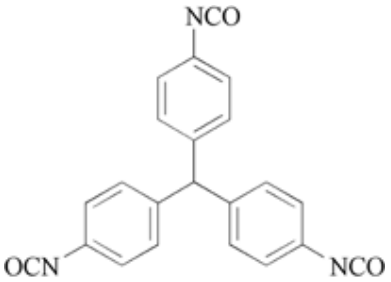

9

Aromatic isocyanates with the NCO group in the benzylic position behave similarly to the (cyclo-)aliphatic isocyanates with regard to reactivity and light stability:

1,3-Bis(1-isocyanato-1-methylethyl)benzene [2778-42-9] (**10**), m-tetramethylxylylene diisocyanate (TMXDI, American Cyanamid) [28, 29], *mp*–10 °C, *bp* 150 °C (49.5 kPa).

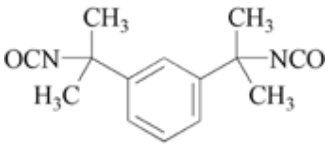

10

3.1.2. *Aliphatic Polyisocyanates*

Aliphatic and cycloaliphatic polyisocyanates are used to obtain light-stable PURs. This is particularly important for coating materials (paints, varnishes, textile coatings, and glass-fiber sizes).

Hexamethylene diisocyanate (HDI) [822-06-0], 1,6-diisocyanatohexane (**11**) (Desmodur H, Bayer MaterialScience), *bp* 127 °C (1.33 kPa).

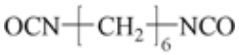

11

Biuret based on HDI, *N*-isocyanatohexylaminocarbonyl-*N*,*N′*-bis-(isocyanatohexyl) urea [28182-81-2] (**12**), biuret triisocyanate (technical mixture), ca. 75 % solution in methoxypropyl acetate/xylene (1/1) (Desmodur N75, Bayer MaterialScience), NCO content 16–17 %, viscosity (20 °C) 250 ± 100 mPa · s.

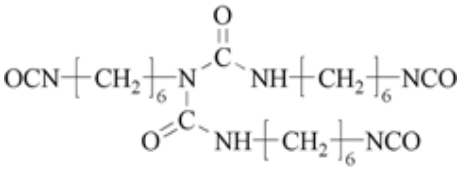

12 (idealized)

Trimer based on HDI, 2,4,6-trioxo-1,3,5-tris(6-isocyanatohexyl)hexahydro-1,3,5-triazine [28182-81-2] (**13**), trimeric hexane diisocyanate, ca. 90 % solution in butyl acetate (Desmodur N3390, Bayer MaterialScience), viscosity (25 °C) 2500 ± 500 mPa · s, NCO content 21.5 %, monomer content < 0.5 %.

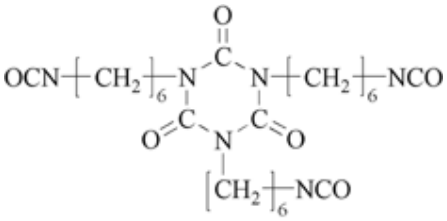

13 (idealized)

Unsymmetrical trimer based on HDI, 2,4-dioxo-3,5-bis(6-isocyanatohexyl)-6-(6-isocyanatohexyl)hexahydro-1-oxa-3,5-diazine [28182-81-2] (**14**), unsymmetrical trimeric hexane diisocyanate [30] (Desmodur XP2410, Bayer MaterialScience), viscosity (23 °C) ca. 700 mPa · s, NCO content ca. 24 %, monomer content <0.3 %.

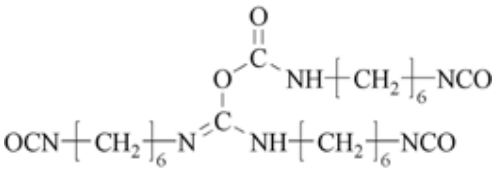

14 (idealized)

Isophorone diisocyanate (IPDI) [4098-71-9], 3,5,5,-trimethyl-1-isocyanato-3-isocyanatomethylcyclohexane (**15**) (IPDI, VEBA), *bp* 158 °C (1.33 kPa), 70 % cis, 30 % trans [31, 32]. For reactivity, see [33].

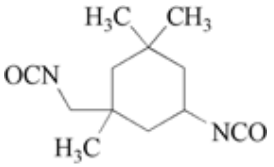

15

Uretdione of IPDI (Crelan TP.LS 2147) [34], *mp* 105–115 °C, ca. 15.4 % NCO.

Trimer based on IPDI, 2,4,6-trioxo-1,3,5-tris(5-isocyanato-1,3,3-trimethylcyclohexylmethyl)hexahydro-1,3,5-triazine [53880-05-0], ca. 70 % solution in xylene/methoxypropyl acetate (1/1) (Desmodur Z4470, Bayer MaterialScience), softening range 90–110 °C, NCO content 17 %, monomer content < 0.5 %.

1,1-Methylenebis(4-isocyanatocyclohexane), H12-MDI [5124-30-1], 4,4′-diisocyanatodicyclohexylmethane (**16**), trans,trans-4,4′-diisocyanatodicyclohexylmethane, *mp* 83 °C [35]. Technical isomer mixture (24 % *cis*,*cis*, 43 % *cis*,*trans*, 20 % *trans*,*trans*, 13 % 2,4 and 2,2′ isomers): *mp* 15 °C (Desmodur W, Bayer MaterialScience)

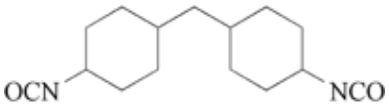

16

3.1.3. *Blocked Isocyanates [36-39]*

Polyisocyanates or prepolymers containing NCO groups are not stable on storage in the presence of atmospheric humidity or compounds that contain reactive hydrogen atoms.

Reaction of the NCO groups with certain organic compounds that have acidic hydrogen atoms such as phenol, cresol, nonylphenol, caprolactam, and methyl ethyl ketoxime leads to “blocked isocyanates,” which are stable (storable) at room temperature and labile at elevated temperatures. These adducts are inert to polyols at room temperature and can be mixed with them in one-pack systems. At higher temperatures, usually at 140–180 °C the blocking agent is released, and the free isocyanate reacts as usual with the NCO-reactive components. Due to their thermal lability, uretdiones and uretonimines also behave like blocked isocyanates.

For the physical blocking of solid polyisocyanates by a diffusion barrier layer, see Section Reactive One-Pack Systems.

3.2. Polyols

Polyols are the largest group of PUR starting materials. The wide range of properties of the resulting PURs is largely determined by the chemical composition and molecular mass of the polyols. The largest group of polyols used for PURs is the group of polyethers. The volume of polyesters is much smaller, but they nevertheless have become more important, especially for flame-retarded rigid foams (see Section Special Types).

3.2.1. *Polyether Polyols*

See also ⟶ Polyoxyalkylenes

Polyethers (polyether polyols) with terminal hydroxyl groups are produced by addition of cyclic ethers, especially propylene oxide (PO, R′ ═CH_3) and ethylene oxide (EO, R′ ═H), to polyfunctional “starter” molecules.

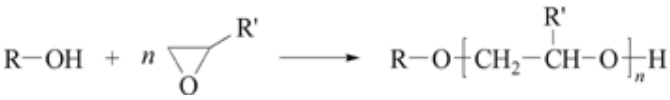

Usually, the EO and PO units can be structured as homo blocks, produced by batchwise addition of EO and PO to the starter molecule, or as mixed blocks by feeding an EO/PO mixture to the starter. The combination PO block-mixed block-EO block is common.

The structure determines the reactivity, the hydrophobic-hydrophilic properties, the surface activity, and consequently the foaming behavior of the polyethers. Polyols with terminal EO units contain primary hydroxyl groups and consequently have

a higher reactivity than polyols with the secondary hydroxyl groups of terminating PO units.

The following starters are typically used industrially: ethylene glycol (EG), 1,2-propanediol (PG), bisphenol A (BPA), trimethylolpropane (TMP), glycerol (GLY), pentaerythritol, sorbitol, sucrose, water, ethylenediamine (EDA), and diaminotoluene (TDA). Monofunctional polyethers are usually started with 1-butanol.

Depending on their structure, two groups of polyethers can be distinguished: long-chain polyols, mainly used for flexible foams, and short-chain polyols, mainly used for rigid foams.

The long-chain polyols typically have hydroxyl numbers below 100 mg KOH/g and functionalities of 2 to 3, which corresponds to molecular weights of 2000 g/mol or more. The properties of some long-chain polyols are given in Table 1.

Table 1. Properties of long-chain polyether polyols

Starter	Alkylene oxide	Functionality	M_r, g/mol	OH no., mg KOH/g	Viscosity (25°C), mPas
PG	PO/EO	2	4000	28	870
PG	PO	2	2000	56	310
TMP/PG	PO	2.8	3750	42	630
GLY	PO/EO	3	4800	35	860

The short chain polyols usually have OH numbers of 200 mg KOH/g or more, functionalities of up to 6, and molecular weights of 1000 g/mol or less (Table 2).

Table 2. Properties of short-chain polyether polyols

Starter	Alkylene oxide	Functionality	M_r	OH no., mg KOH/g	Viscosity (25°C), mPa s
TMP	PO	3	435	385	600
TMP	PO	3	305	550	1800
EDA	PO	4	350	620	19200
TDA	PO/EO	4	475	470	8000
Sucrose/PG	PO	3,1	385	450	15000

.

Polyethers produced from trifunctional starters and bifunctional cyclic ethers (e.g., epoxy resins) are highly branched and, to some extent, intramolecularly cross-linked [40]. Special polyethers contain, for example, quaternary ammonium or sulfonate groups [41, 42].

An important special polyether is polytetrahydrofuran [43], which is described in detail elsewhere (⟶ Polyoxyalkylenes). It is used mainly for the production of high-grade elastomers for the engineering field and is liquid above 40 °C [44, 45].

A newer process, known as the IMPACT process, uses double metal cyanides (DMC) as catalysts for the production of especially long chain polyether polyols [46-48] that show improved properties compared to polyols produced by the classical KOH process. During the addition of propylene oxide to the starter molecule, base-catalyzed isomerization of propylene oxide to allylic alcohol, which itself can act as starter molecule, leads to a mono-ol containing a double bond. These side products lower the average functionality of the polyols and reduce, for example, their stability against oxidation. The IMPACT process leads to polyols with extremely low mono-ol contents and therefore a low level of unsaturation and results in polyols with functionalities very close to the theoretical number. The IMPACT polyols have become more and more important, especially in the production of flexible foams.

Products based on polyethers have good low-temperature behavior and usually high hydrolytic stability. The viscosity of the liquid polyethers is relatively low (40–15000 mPa · s). A disadvantage of polyether polyols is that they undergo thermooxidative degradation on exposure to heat and atmospheric oxygen; UV radiation is mainly responsible for light-induced degradation changes. Both antioxidants and UV stabilizers counteract this degradation. In some special applications polyether polyols containing hydrazide or urea groups, both of which have a stabilizing effect, are used to avoid thermooxidative degradation.

Polyether dispersions (two-phase systems with a solid polymer as the disperse phase) are a further group of polyethers. Mainly styrene-acrylonitrile polymers (SAN polyols), polyurea, or poly(hydrazodicarbonamide) polymers (PHD polyols) are used as fillers. The dispersions are milky white, of higher viscosity than the corresponding base polyols, and completely stable to sedimentation. The average particle diameter is ca. 1 µm.

3.2.2. *Polyester Polyols*

Compared with polyethers, polyesters with terminal hydroxyl groups show outstanding resistance to light and thermal aging. PURs produced from them are therefore primarily used for paints, coating materials, and, due to their thermal

stability, for flame-retarded rigid foams. Special types are suitable for high-grade PUR elastomers [49, 50].

Polyesters are produced by polycondensation of di- and trifunctional polyols with dicarboxylic acids or their anhydrides. Polyols commonly used are ethylene glycol, 1,2-propanediol, 1,4-butanediol, 1,6-hexanediol, neopentyl glycol, diethylene glycol, glycerol, and trimethylolpropane. Common dicarboxylic acids or anhydrides are succinic acid, glutaric acid, adipic acid, phthalic anhydride, isophthalic acid, and terephthalic acid.

The hydrolytic stability of a polyester urethane is determined by the constitution of the polyester. Polyesters synthesized from glycols and dicarboxylic acids that each contain at least five carbon atoms have good hydrolytic stability. This is because the increase in hydrophobicity of the polyester segments repels moisture [49, 50]. The best results have been obtained with polyesters derived from 1,6-hexanediol and adipic acid.

Common polyesters based on adipic acid and ethylene glycol or ethylene glycol/1,4-butanediol provide inexpensive PUR elastomers of excellent quality. Their hydrolytic stability is lower than that of the polyesters mentioned above and can be considerably improved by appropriate stabilizers (see Section Catalysts and Additives).

Polyesters have a wide molecular mass distribution and, compared with polyethers, a high viscosity; many are glassy solids (polyphthalates) or crystalline (polyadipates). Polyesters containing diethylene glycol are ether esters; their viscosities are relatively low and their stability to hydrolysis and UV radiation is moderate. The average molecular mass of polyesters is 500–4000 g/mol. For properties, see Table 3.

Table 3. Properties of polyester polyols

Acid	Glycol	Functionality	M_r	OH no., mg KOH/g	Viscosity, mPas
Adipic acid	ethylene glycol	2	2000	56	500 – 600 (75° C)
Adipic acid	diethylene glycol/TMP	2.5	2300	60	900 – 1100 (75° C)
Phthalic anhydrid/adipic acid	ethylene glycol	2	1750	64	2200 – 3200 (75°C)
Phthalic anhydrid/adipic acid	diethylene glycol	2	560	200	6300 – 7600 (25°C)
Terephthalic acid	diethylene glycol	2	450	250	2700 – 7700 (25°C)

The polyesters always contain significant proportions of the monomeric glycol used in their synthesis, as well as short-chain oligomers and often nonfunctional cyclic esters. These substances can be removed for a limited time by thin-film distillation, but they are gradually reformed by transesterification.

3.2.3. *Polycarbonate Polyols*

Aliphatic carbonate esters (⟶ Polycarbonates) are produced by transesterification of dimethyl carbonate (DMC) or diphenyl carbonate (DPC) with glycols (e.g., 1,6-hexanediol).

Many investigations have been carried out on the production of polycarbonate by copolymerization of alkylene oxides with carbon dioxide [51-55], but this reaction is not used on an industrial scale as yet.

Polyaddition of caprolactone or pivalolactone to low molecular mass diols yields hydroxy-functional polylactones.

As the susceptibility of polyesters to hydrolysis increases with increasing dissociation constant of the acid on which they are based, the polycarbonates are highly stable against hydrolysis. In particular, polyester urethanes based on 1,6-hexanediol polycarbonate show an unmatched resistance to hydrolysis.

3.2.4. *Other Polyols*

Other oligomers and polymers with functional groups (OH, SH, COOH, NHR) can also be used for PUR synthesis (e.g., polythioethers, polyacetals, polyester amides, alkyd resins, and polysiloxanes).

Hydroxy-functional poly(1,3-butadienes) contain the hydroxyl function mainly in the terminal position as a result of polymerization with hydrogen peroxide. These compounds yield PURs with moderate mechanical properties but extremely high hydrolytic stability and good low-temperature behavior. Bifunctional poly(1,3-butadienes) can also be obtained by anionic polymerization.

Hydroxyl-bearing polyacrylates based on 2-hydroxyethyl acrylate or methacrylate monomers are used in the production of high-quality, two-component paints [56] (see also ⟶ Paints and Coatings – Polyurethane Coatings).

Low molecular mass polyols, in contrast to the higher molecular mass polyols mentioned above (Section Polyether Polyols), are used as chain extenders. In the production of PUR elastomers they are generally used in the synthesis of the “hard” segment (see Section Polyurethanes with Segmented Structure). Important representatives are ethylene glycol, 1,4-butanediol, 1,6-hexanediol, and 1,4-bis(hydroxyethoxy)benzene.

Trimethylolpropane or glycerol is added to introduce a defined small degree of branching.

3.3. Diamines and Amino-Terminated Polyethers

Diamines are used as building blocks for PURs that contain urea groups. Due to their high reactivity the polyaddition is very fast even without a catalyst, leading to polymers with considerably improved properties compared with “standard” PURs.

For polyaddition in bulk, sterically hindered diamines (e.g., isomeric diaminodiethylmethylbenzenes **17** [57]) or other aromatic diamines with mono- or dialkyl ring substituents [58-61] are preferred.

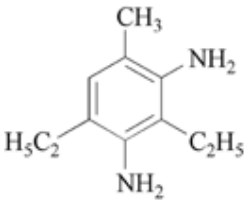

17a

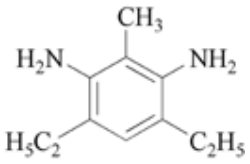

17b

These amines are liquid at room temperature and therefore suitable for processing in two-component devices (see Section Metering and Mixing Technology).

Less reactive diamines with electron-withdrawing substituents are used for producing elastomers by the prepolymer process (see Section Prepolymer Processes); examples include bis(4-amino-3-chlorophenyl)methane [101-14-4] (MBOCA) and 2-methylpropyl-4-chloro-3,5-diaminobenzoate [32961-44-7] [62-64]. For the toxicological and adverse health effects of MBOCA, see [65].

Amino-terminated polyethers (ATPE) are obtained primarily from the appropriate polyether polyols (see Section Polyether Polyols) by exchanging the hydroxyl groups with ammonia [66-68] or primary amines [69]. Aminopropoxy end groups are obtained by cyanoethylation of polyols followed by hydrogenation. The high reactivity of the amino groups can be reduced by using their adducts with carbon dioxide (carbonates, carbamates). During the polyaddition reaction the carbon dioxide is released and acts as a blowing agent.

Various processes have been proposed for producing ATPEs with aromatic amine end groups [70]. For example, NCO prepolymers can be hydrolyzed to give the corresponding amino-terminated prepolymers [71]. Other methods include the preparation of N-substituted crotonic acid esters by acetoacetylation of polyols followed by amination [72, 73] and nitrophenylation followed by hydrogenation [74-76].

The adducts of (cyclo-)aliphatic di- and polyamines with ketones are bis- and polyketimines, respectively. They function as capped amines, with which they are in equilibrium in the presence of water (see Section Blocked Isocyanates). In the tautomeric enamine form, they also react with NCO groups to form urea and amides. Combinations of ketimines and isocyanates are thus both moisture- and heat-curing [77-79].

Polyether ketimines based on aliphatic polyether amines have a rather low reactivity and therefore a very good processability [80-83].

3.4. Special Building Blocks

Aqueous PUR-ionomer dispersions are synthesized from diols or diamines that contain ionic groups or groups that can be converted to ionic groups by alkylation or neutralization with compounds such as *N*-methyldiethanolamine, 2-(*N*,*N*-diethylaminomethyl)-2-ethyl-1,3-propanediol, and dimethylolpropionic acid, as well as salts of 2-sulfo-1,4-butanediol, *N*-sulfoethylethylenediamine, *N*-carboxyethylethylenediamine, lysine, and glycerol 2-phosphate [84, 85].

Hydroxy-functional polyisobutylene [86, 87], siliconediamines [88-90], and chlorosulfonated polyolefins [91] can be used as oligomer building blocks leading to segmented PUR polymers.

Other special polyols can be obtained by glycolytic degradation of polymer waste [92-94] (see Section Recycling/Recovery of Polyurethanes).

3.5. Catalysts and Additives

Catalysts. In order to balance the two main reactions, the isocyanate-water reaction (blowing reaction) and the isocyanate-polyol reaction (gel reaction), various catalysts are used. The reactions of the isocyanate group are extraordinarily sensitive to many different kinds of catalysts:

1. Lewis bases: 1,4-diazabicyclo[2.2.2]octane [280-57-9] (DABCO), triethylamine [121-44-8], dimethylbenzylamine [103-83-3], bis[2-(dimethylamino)ethyl] ether [3033-62-3], 1,1,3,3-tetramethylguanidine [80-70-6], 1,8-diazabicyclo[5.4.0]undec-7-ene [6674-22-2] (DBU), 2,4-bis(dimethylaminomethyl)phenol [5424-54-4]
2. Lewis acids: bis(ethylhexanoyloxy)tin [301-10-0] (tin dioctanoate), bis(dodecanoyloxy)dibutyltin [77-58-7] (dibutyltin dilaurate), dichlorodimethyltin [753-73-1]
3. Insertion catalysts: bis(dodecylthio)dibutyltin, organotin alkoxides, organotin oxides, organotin thiolates, organotin sulfides, 1,3-dicarbonyl compounds

Almost every catalyst has a specific “activity profile” [13, 95-99]. The situation is extraordinarily complex, especially since

the urethane group itself exerts a catalytic effect. As hydrogen bonds and the nature of solvents used in the reaction influence the reaction rates, most kinetic studies in solvents are not applicable to solvent-free PUR production [100]. Due to the desire to reduce emanations from PUR foams less volatile or reactive catalysts have been developed [101, 102].

Metal catalysts activate the isocyanate groups by making them more electrophilic, while bases make the hydroxyl groups more nucleophilic [103].

Alkali metal salts of organic acids and phenols, as well as phenol Mannich bases catalyze not only the urethane reaction but also the trimerization of the isocyanate group. NCO prepolymers trimerize more rapidly than the corresponding diisocyanates because of the presence of urethane groups which act as cocatalysts (see Section Polyisocyanurates).

Phosphines are special catalysts which dimerize the NCO group; phospholine oxides cause carbodiimidization.

Additives (see also ⟶ Plastics, Additives).

Foam Stabilizers. Besides the chemical structure and molecular weight of the polyols and isocyanates, cell size and fraction of open cells can be controlled to a certain extent by surfactants or foam stabilizers [104]. In some cases, foam formation is virtually impossible without these surface-active compounds. The majority of foam stabilizers are polysiloxane-polyether block copolymers [105-108]. Specific types, e.g., for liquid CO_2 as blowing agent [109], for flame-retarded foams [110], or for other specific foam types are available as well [111, 112].

Hydrolysis Stabilizers [113-116]. Monocarbodiimides, polycarbodiimides, epoxides, cyanates, phenyliminooxazolidines.

Oxidation Stabilizers [113-115], [117-119]. These additives are used in foamed materials, especially in flexible foams, to prevent core scorching. They include sterically hindered phenols, alkylated anilines, phosphites, hydrazides, sulfides, and thioethers.

UV Stabilizers [113-115, 117-119]. Piperidines with bulky substituents in the 2,6-positions, benzophenones, benzotriazoles, cyanoacetate ester derivatives, quenchers, Tinuvins (Ciba-Geigy).

Blowing Agents (see also ⟶ Foamed Plastics – Blowing Agents). According to the implementation of the Montreal Protocol (and the consecutive amendments) blowing agents with an ozone-depleting potential are no longer used as blowing agents for flexible foam by the member states since 1992. Several alternatives have been developed [120-127].

For flexible foams the predominant technology meanwhile is the addition of liquid or supercritical carbon dioxide (e.g., Novaflex by Hennecke, Cardio by Cannon). Another method is the application of varied ambient pressure, e.g., MegaFoam by Hennecke or VPF by Beamech (see Section Production) and depending on regulations of the local authorities, blowing agents like dichloromethane or acetone may be applied as well.

For rigid foams nowadays usually alkanes, namely, n-pentane, isopentane, and cyclopentane are used (see Section Raw Materials).

Flame Retardants (see also ⟶ Plastics, Additives – Flame Retardants). Various classes of chemicals are used to modify the combustion behavior of PUR flexible foams [128]. The selection of the appropriate flame retardant depends on the test that must be passed, as well as the chemistry of the foam (e.g., polyester vs polyether, HR vs. conventional foam). Most common are phosphorus- and/or nitrogen-containing compounds and halogenated phosphates such as tris(2-chloropropyl) phosphate (TCPP). As in the case of catalysts, less volatile or reactive flame-retardants have been developed in order to reduce the emissions of volatile organic compounds (VOCs) from the finished foam. The liquid flame retardants may be combined with solid ones such as melamine and ammonium polyphosphate. Brominated diphenyl ethers, which have been used for some special foam grades or in some regional markets, will be phased out due to concerns regarding their potential for bioaccumulation.

For rigid foams halogenated compounds such as TCPP, polyesters based on tetrabromophthalic acid, and special bromopolyethers are usually used. The use of polyisocyanurate rigid foams (see Section Special Types), which exhibit a superior flame behavior compared to PUR rigid foams, has increased. This has led to a reduction in flame retardant consumption.

Aluminum oxide trihydrate, magnesium hydroxide, and expandable graphite are also used as flame-retardants. The first two act by splitting off water vapor and the latter acts by forming protective layers.

[Top of Page]

4. Structure and Morphology

Wolfgang Friederichs

Depending on their structure, PURs cover a broad range of properties. Besides the chemical composition, the chain length, and cross-linking, the properties are significantly affected by the interactions between the polymer chains (see Fig. 1), that is to say by the supramolecular structure and morphology.

Figure 1. Hydrogen bond interactions between polymer chains of a PUR based on HDI and 1,4-butanediol

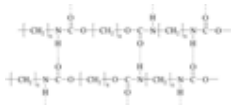

4.1. Polyurethanes Without Segmented Structure

The properties of linear PURs, which are produced by polyaddition of monomeric diisocyanates and short- or long-chain diols, are similar to those of polyamides.

Short-Chain Polyols. The polyaddition of hexamethylene diisocyanate (HDI) and 1,4-butanediol leads to a very tough, crystalline material (mp 180–185 °C), which is used for manufacturing of fibers and bristles. The polyaddition of toluene diisocyanate (TDI) and butanediol or diethylene glycol yields very hard, transparent products with low heat resistance.

Long-Chain Polyols. Linear PURs based on diisocyanates (preferably MDI) and long-chain polyester or polyether diols are soft, rubbery, high molecular mass polymers with only a small proportion of urethane groups. Therefore, these materials are homogeneous (single phase), and the intermolecular physical interactions are mainly determined by the polyester and polyether segments (van der Waals forces [129]).

With increasing stiffness of the polyol chain (e.g., highly crystalline polyesters), and decreasing chain length, the resulting PUR becomes harder and less elastic.

4.2. Polyurethanes with Segmented Structure

To obtain PUR elastomers with superior properties, the polyaddition reaction mixture usually contains three components, namely, a diisocyanate, a long-chain polyether or polyester polyol, and a short-chain diol (chain extender). The resulting block copolymer type PURs are characterized by a segmented structure.

The secondary and tertiary structures, and as a consequence, the morphology of these PURs, depend on the chemical structure and the block length of the different segments. Due to the two- or multi-phase morphology, segmented PURs can exhibit extraordinary high-grade properties.

4.2.1. *Hard and Soft Segments [12, 130-135]*

The reaction of one equivalent of long-chain diol, one equivalent of short-chain diol, and two equivalents of diisocyanate leads to the idealized structure shown in Figure 2.

The long-chain polyols, which are flexible and are usually tangled (soft segments), alternate along the polymer chain with oligomeric rigid urethane units (hard segments).

Figure 2. Schematic primary structure of a segmented PUR (idealized)

Usually the structure of the PUR as well as the structure of the soft segment itself is determined by a Flory distribution [136, 137]. This results in a smaller number of larger hard segments than expected in theory. If polyester polyols are used, the low molecular glycols, which are usually present in considerable amounts, even enhance this effect. Further effects leading to deviations from theory include kinetic effects (if end groups with different reactivity are used) and phase transitions due to temperature changes during the polyaddition reaction.

4.2.2. *Segregation and Morphology*

On cooling, the initially homogeneous PUR melt converts to a two-phase system. This is due to the incompatibility of the polar low-molecular glycols with the low-polarity polyether or polyester polyols and polyisocyanates. As the reaction progresses, the formation of polar urethane groups at the ends of the long-chain polyols leads to a certain degree of compatibilization. Nevertheless, the low miscibility of the hard and soft segments leads to segregation, which results in covalently bonded microphases in which paracrystalline oligourethane domains appear (tertiary structure, Fig. 3). On further cooling, these can aggregate and grow to form spherulitic microcrystallites (quaternary structure).

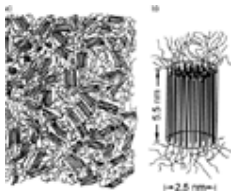

Figure 3. A) Tertiary structure of a linear segmented PUR with polydisperse soft segments and uniform (monodisperse) hard segments; B) Cylindrical model of a hard domain

The hard segment domains form three-dimensional cross-linking centers of a noncovalent type (physical cross-linking). The formation of a large number of weak hydrogen bonds results in a high bonding strength, but allows shifts and rearrangements under load, where a covalent bond would rupture. As the hard segments are covalently bonded to the soft segments, flowability of the chains is restricted. This gives the segmented PURs their high elasticity. The hard segments

are responsible for tensile strength, tear resistance, hardness, permanent elongation, and compression set, while the soft segments determine the elastic expansion and the glass transition temperature.

The interaction between different hard segment domains is mainly determined by the symmetry of the isocyanates and the nature of the chain extenders, which influence the possibility to form highly organized structures.

Most commercially produced PURs contain urea groups, which often determine the property profile more strongly than the comparatively small content of urethane groups. Since the urea group is more polar than the urethane group, the segregation of hard and soft segments is strengthened, the melting range of the hard-segment domains is higher, and the intermolecular (interchain) interaction stronger [138]. Due to these strong interactions, a single urea group shows hard-segment character. These interactions also explain the superior mechanical properties of polyureas and their low thermoplasticity.

The balance between these hard segment/hard segment interactions and the interactions between hard segments and soft segments determine the degree of segregation. Due to decreasing polarity, segregation increases from polyesters to polyethers to polybutadiene polyethers.

In the soft segments, strain-induced crystallization can be initiated by stretching or disentanglement. This strain-induced crystallization explains the high hysteresis of typical PUR elastomers [139, 140]. The mechanical properties of segmented PUR depend on the melting temperature of the crystalline domains.

Products based on 1,5-naphthalene diisocyanate (NDI) are no longer thermoplastic because the melting temperature of the hard segments is too high. They are usually produced with a small excess of isocyanate as casting elastomers (see Section Cast Elastomers). These products have extremely good mechanical properties: high tear strength, high abrasion resistance, high elasticity at high hardness, low damping, high load-bearing capacity, and excellent long-term dynamic endurance (load-bearing property) [135].

High-quality PUR thermoplastic elastomers are produced primarily from MDI. The melting temperature and the solubility of these segmented thermoplastic PURs can be influenced by modifying the chemical composition, especially of the hard segments. Use of a second short-chain diol lowers and widens the melting temperature range, and solubility is also improved.

As the properties of segmented PUR are highly determined by the phase segregation and the crystallinity of the domains, numerous production parameters [130, 131, 141-143] have tremendous influence on the behavior of thermoplastic PUR materials. Even thermoplastic PUR with monodisperse (uniform-length) N-alkylurethane hard segments, which are incapable of hydrogen bond formation show differing morphology and properties depending on their thermal pretreatment [144].

4.3. Cross-linking of Polyurethane

PUR flexible and rigid foams, cast elastomers, and most reactive systems are usually cross-linked. Even if these cross-linked PURs are not thermoplastic, the melting of the hard-segment domains is still observable, e.g., by softening (decrease in shear modulus) at elevated temperatures. With increasing cross-linking, the glass transition temperature increases, whereas the tendency to crystallize decreases.

If the cross-linking is not caused by subsequent reactions (formation of allophanate or biuret), but introduced by the use of highly functional raw materials, then the formation of segregated domains can be disturbed due to the rapid formation of a highly cross-linked PUR network, fixing the polymer chains or at least reducing their flexibility. In this case, even heating to elevated temperatures does not improve the material properties, as domain formation is not possible (e.g., in flexible foams).

4.4. Polyisocyanurates [145]

The general term PURs also covers products that contain cyclic urea structures (isocyanurates), which are formed by cyclotrimerization of isocyanate groups. The mechanism of isocyanurate formation is not completely understood. The most probable mechanism of base catalyzed trimerization [146] is shown in Figure 4.

Addition of the catalyst II to the carbon-nitrogen double bond leads to an urethane anion III, which by subsequent addition of two more isocyanate molecules I is transformed via IV to the adduct V. This adduct collapses into the stable isocyanurate VI and regenerates the catalyst II.

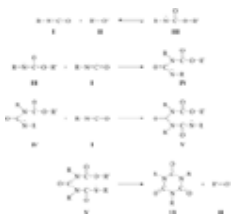

Figure 4. Mechanism of base-catalyzed isocyanate trimerization

The isocyanurates exhibit improved mechanical properties such as compression strength and dimensional stability and generally show superior fire retardancy [147]. These advantages have led to the increasing importance of isocyanurates, especially since the blowing agent for rigid foams was switched from CFC to pentane (see Section Special Types).

[Top of Page]

5. Production of Polyurethanes

Wolfgang Friederichs

5.1. Stoichiometry

Isocyanates (RNCO) are specified by their percentage NCO content and in some cases also by their equivalent mass (molecular mass of NCO moiety/functionality). Functionality means the number of isocyanate groups per molecule.

$$\text{Equivalent mass} = \frac{\text{molecular mass of RNCO}}{\text{functionality}} = \frac{42}{\%\text{NCO}} \cdot 100 \tag{1}$$

Polyols are specified by their hydroxyl number (OH no.) and sometimes also by their percentage OH content or their equivalent mass. The OH no. is given in mg KOH/g, owing to the method used for analyzing the OH groups: they are usually treated with acetic anhydride and, after hydrolysis with water, the excess acetic anhydride is titrated with potassium hydroxide. This means that a polyol containing 1 mmol OH groups per gram has an OH no. of 56 mg KOH/g.

$$\text{OH no.} = \%\,\text{OH} \cdot \frac{56 \cdot 1000}{17 \cdot 100} = \%\,\text{OH} \cdot 33 \tag{2}$$

$$\text{Equivalent mass} = \frac{56 \cdot 1000}{\text{OH no.}} \tag{3}$$

Water, *acids* (acid number AN), and secondary and tertiary amines are also most simply characterized by their equivalent mass. For mixed amines, a mean functionality must be used where necessary. For calculating quantities, 1 gram equivalent isocyanate corresponds to 1 gram equivalent Zerewitinoff-active hydrogen compound (e.g., polyol). In practice, the quantity of isocyanate is calculated almost exclusively for 100 parts by weight of the active-hydrogen compound(s). If the raw materials also contain water, then an additional quantity of isocyanate is required, since the latter reacts with water to form carbon dioxide and substituted urea.

The index (or isocyanate index) is equal to 100 (isocyanate index = 1) if the quantity of isocyanate added equals the amount necessary for reaction with all Zerewitinoff-active hydrogen atoms.

$$\text{Index} = \frac{\text{Used isocyanate quantity}}{\text{Necessary isocyanate quantity}} \cdot 100 \tag{4}$$

In practice, the index can be used to influence the properties of the PUR decisively, e.g., by “under-” or “over-cross-linking.” Under-cross-linking means that free hydroxyl groups remain in the PUR; over-cross-linking means that the urethane moieties undergo consecutive reactions, i.e., the formation of allophanates. High indices of 300 or more are used to produce polyisocyanurates (PIR), in which the isocyanate groups undergo trimerization by reaction with each other (see Section Polyisocyanurates).

5.2. Reaction without Solvents

The possibility of manufacturing PURs from liquid raw materials without the use of any solvent is one of the biggest advantages of PUR chemistry. As a result, most PURs are manufactured by a mass polyaddition process. This process can be a single-step reaction (one-shot process) or a two- or multiple-step reaction (prepolymer process). The reactivity is controlled by temperature or by the addition of catalysts. Due to the high exothermicity, reaction temperatures of about 100 to 180 °C are not unusual. These high temperatures guarantee good end curing of the molded PUR parts.

5.2.1. *One-Shot Process*

Bulk reactions without solvents are usually very fast, especially when catalysts are added. Therefore, the production of PUR foams is usually carried out by the one-shot process, in which all components are mixed directly, generally with simultaneous addition of auxiliaries such as catalysts, foam stabilizers, reinforcing agents, fillers, and fire retardants. The reaction is highly exothermic and largely completed in 0.5–30 min, depending upon the catalyst. The final properties, however, are frequently reached only after 24–48 h (post-curing).

The one-shot process requires that the reactivity of all isocyanate species is quite similar. Amines usually lead to excellent PUR properties but react much more rapidly than polyols. If diamines or polyamines are to be used it must be ensured that the urea formed at the beginning of the reaction does not precipitate and is no longer available for incorporation into the hard segments. This means that the use of polyamines is limited to amines which cannot undergo cross-linking (secondary amines), are sterically hindered [57-61, 148], or are deactivated by electron-withdrawing substituents [62-64, 149].

For easy processability all components should be liquid, of low viscosity, and miscible or at least be able to form a homogeneous reaction mixture for the very first moments of the reaction. Usually all the nucleophilic components, that is, polyols and/or polyamines, are premixed and homogenized. If these mixtures are incompatible with the short-chain diols (chain extenders), the miscibility must be enhanced by raising the temperature or adding emulsifiers.

In general, the polyaddition reaction of a multicomponent system by the one-shot process leads to a predominantly statistical structure of the polymer chains.

The term reaction injection molding (RIM) technology usually is used when highly reactive components are mixed and injected into a mold on a very short timescale (see Section Integral Skin Foams and RIM Materials). This technique allows for the handling of reactive systems with pot lives (i.e., the period in which no significant increase in viscosity is observed) of less than 1 s and reaction times of about 10–60 s.

RIM technology has initialized a development, which started with catalyzed highly reactive PUR systems [150], and, by the use of polyurethane-polyurea systems, chain extended by diamines [151-153], led to pure polyurea systems, in which the

only isocyanate-reactive components are diamines [154-163].

5.2.2. *Prepolymer Processes*

In PUR chemistry, prepolymers are defined as intermediate reaction products of the polyaddition reaction which still contain free isocyanate or isocyanate-reactive groups. In contrast to the one-shot process, the prepolymer process allows for a highly controlled polyaddition reaction, e.g., to construct tailor-made segment structures.

In the first step of the process, polyether or polyester polyols are usually treated with excess diisocyanate to yield homogeneous reaction mixtures, which may contain considerable amounts of unconverted isocyanate groups. If diisocyanates with NCO groups of different reactivity are used (2,4-diisocyanatotoluene, isophorone diisocyanate), the resulting NCO prepolymers exhibit narrow molecular weight distribution with only minor amounts of unconverted monomeric diisocyanate.

If prepolymers free of monomeric diisocyanate are required, the latter must be removed, e.g., by thin-film distillation or extraction with a hydrocarbon solvent which does not dissolve the prepolymer. The monomer content is then lowered to below 0.1 % [164, 165].

The production of storage-stable NCO prepolymers from aromatic diisocyanates is difficult and requires special precautions (e.g., the exclusion of all traces of water and catalysts) to avoid side and consecutive reactions [166].

Low-viscosity, monomer-free NCO prepolymers can be obtained from silylated hydroxy compounds and isocyanates with a carboxylic acid chloride group [167, 168].

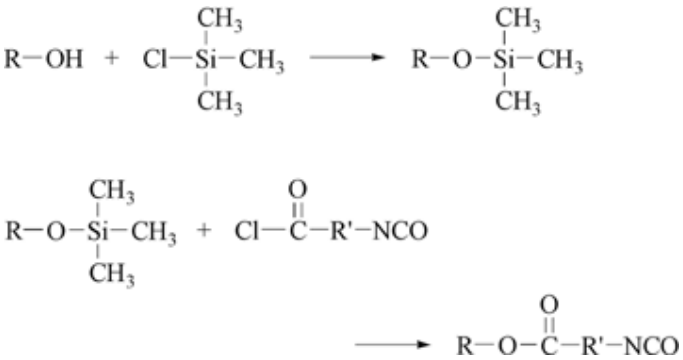

5.3. Reaction in Solvents

5.3.1. *One-Component Systems*

Linear or slightly branched PURs or polyurethane polyureas with molecular weights of 50 000 to 150 000 g/mol can be obtained by polyaddition reaction in solution. The PURs are usually isolated by evaporation of the solvents or by precipitation, e.g., by pouring the solution into a nonsolvent liquid.

The choice of solvent for the polyaddition is crucial, as the solvent must be able to dissolve both the soft and hard segments of the segmented PUR structure. Polar organic solvents (dimethylformamide, dimethylacetamide, 1-methyl-2-pyrrolidone, sulfolane, etc.) usually fulfill this requirement. In some cases, ketones like acetone or methyl ethyl ketone are also suitable. The ketones are especially suitable for chain extension of NCO prepolymers with aliphatic diamines as they reduce their activity by reversibly forming an adduct.

In industrial processes, linear PURs containing only urethane groups are usually produced without using solvents, but can be dissolved in a solvent if necessary for the application.

PURs containing urethane and urea groups are mainly produced by a two-step process. In the first step a prepolymer is made without the use of any solvents. In the second step this prepolymer is then treated with chain extenders such as diols or diamines. In the second reaction step, the solvent is usually added stepwise to avoid a too rapid increase in the viscosity of the solution. The obtained PUR solutions have PUR contents of ca. 20 to 30 %.

With increasing viscosity it is very difficult to achieve complete conversion of the functional groups. This problem is overcome by using NCO/OH ratios (indexes) slightly above 100 and running the reaction till the viscosity reaches a specified value, which must be determined empirically for each product. Then the reaction is stopped by adding a monofunctional stopper, e.g., dibutylamine.

If the reaction is carried out in highly polar solvents, especially in mixtures of water and acetone, methyl ethyl ketone, or DMF, even highly reactive aliphatic diamines such as ethylenediamine can be used. With increasing molecular mass the products may tend to gelation or precipitation. The solubility of the PUR can be improved by adding small portions of a second chain extender such as 1,2-diaminopropane, piperazine, or water to disturb the formation of the hard-segment domains.

5.3.2. *Two-Component Systems [169]*

Two-component systems are of tremendous importance for coatings and for leather and paper finishing. They usually consist of a polyhydroxy compound and an isocyanate cross-linker.

The polyhydroxy compounds are usually obtained by chain extension of common polyether or polyester polyols with diisocyanates; the isocyanate cross-linkers, which for ecological reasons must be free of isocyanate monomers, may be

NCO prepolymers made by the reaction of diisocyanates with diols, diamines, or water; or they may be modified isocyanates, e.g., trimerization products of HDI (see Section Aliphatic Polyisocyanates).

To increase the amount of reactive components in the solvent-containing two-component systems, numerous improvements have been introduced to industrial processes: part of the solvent can be substituted by reactive thinners [170], which are incorporated into the cross-linked PUR. By using new catalysts an unsymmetrical trimerization product of HDI [30], with particularly low viscosity, has been developed, allowing for higher concentrations (high-solids systems) without deterioration in the processability.

5.4. Reactive One-Pack Systems

The term “one-pack system” denotes the storage-stable, ready-to-use formulation of a polymer precursor which cures without addition of further components, e.g., curing agents. Curing generally is caused by humidity, atmospheric oxygen (mainly used in paints), or by the thermal elimination of a blocking agent (see Section Blocked Isocyanates), which either vaporizes or remains in the cured polymer.

Other curing systems involve encapsulated materials with a diffusion barrier layer, which is broken, e.g., on heating. They cure without uptake or release of a substance.

Moisture Curing. Relatively low molecular mass NCO prepolymers can be dissolved in small amounts of low-polarity solvents (e.g., 75 % solution in ethyl acetate) and cured by the action of atmospheric moisture. The curing rate depends on the atmospheric humidity and temperature. Usually these solutions are used as paints or coatings, and the carbon dioxide formed as a byproduct escapes from the thin film without bubble formation.

Eliminating Blocking Agents. Mixtures of blocked NCO prepolymers with polyols are stable at storage temperatures (20–50 °C) and are cured by heating (> 120 °C) to eliminate the blocking agent (see Section Blocked Isocyanates). The reaction can be conducted in such a way that an isocyanate intermediate does not appear [39]. Appropriate hydrophilic modification can be used to produce such combinations as low-viscosity aqueous emulsions or dispersions, in which one of the components usually has a high molecular mass [38].

Removal of a Diffusion Barrier Layer. Some solid diisocyanates such as NDI, TDI-urea, and TDI-uretdione (TDI-dimer) can be stabilized by encapsulation with a diffusion barrier layer. The barrier layer is produced by dispersing the solid isocyanate in a solvent and reacting the NCO groups at the particle surface with (cyclo-)aliphatic diamines. After removal of the solvent the encapsulated isocyanates are dispersed in polyols or polyamines.

Liquid, pasty, or granular one-pack systems are produced in this way. They can be stored for several months and are made to react by heat shock, e.g., 1 min at 120 °C. Heat-curable reactive powders can also be produced on this basis [171].

5.5. Other Processes

In addition to the industrial isocyanate polyaddition processes described previously and in subsequent sections, several other synthetic methods for PURs or polyureas are of scientific interest [172].

PURs can also be obtained by the polycondensation reaction of bis-chloroformate esters and diamines. This method has the advantage over the polyaddition method that it can be carried out under very mild conditions (at ≤20 °C).

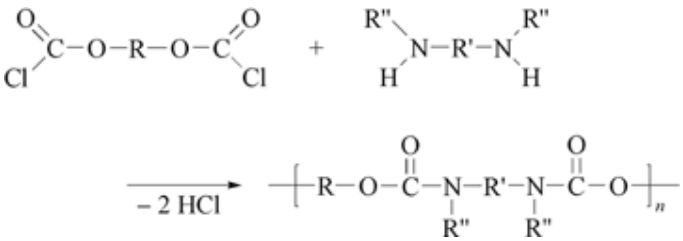

If secondary diamines such as piperazine are used in the polycondensation reaction, thermostable poly(N-alkylurethanes) are obtained. This principle has been used to produce, e.g., high-melting PURs [173] for fibers or PURs with liquid-crystalline character [174]. Liquid-crystalline PURs have also been produced from *N*-alkylcarbamic acid chlorides and diols.

Another method for preparing PURs is the ring-opening polyaddition of oligomers with cyclic ethylene carbonate groups and di- or polyamines, which leads to segmented PURs with free hydroxyl groups [175, 176]:

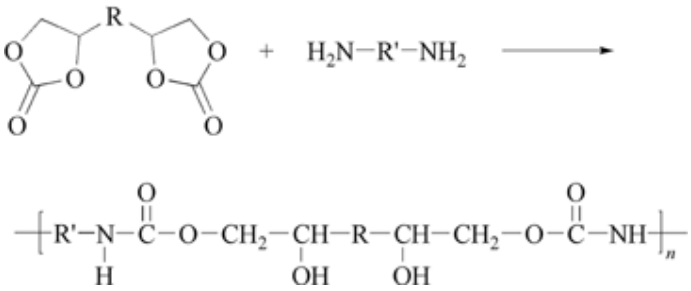

Functionalized PURs can be obtained if carbonylbiscaprolactone (CBC) is used as carbonyl source [177]. Due to its high

selectivity, reactions with diamines can be carried out even in the presence of additional hydroxyl groups.

High-molecular mass α, ω-bis(amino) prepolymers for use in RIM technology (see Section Integral Skin Foams and RIM Materials) have been produced by the reaction of α, ω-bis(amino)polyethers with urea [178].

[Top of Page]

6. Processing of Polyurethanes

Udo Maier

6.1. Supply, Storage, and Preparation of Raw Materials

The raw materials are transported in small (< 100 kg) and normal-sized drums (200–250 kg), and (road) tank trucks and trailers (ca. 23 t). The regulations applicable to the transportation of chemicals (including labeling) must be observed, as well as the transportation conditions prescribed by the manufacturer for preserving the quality of the raw materials. If raw materials are delivered by tank truck, storage tanks for polyol and isocyanate should be installed in pairs at the processing site (Fig. 5). The size of the storage tanks should be adapted to the delivery volume (frequently 15 or 30 m^3). Since the raw materials in the tank truck are usually delivered at elevated temperature, they must be cooled to the processing temperature (22–25 °C) in the storage tank. Ordinary thick plate can be used as the tank material.

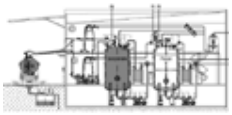

Figure 5. Tank truck unloading and storage tanks

6.2. Metering and Mixing Technology

In contrast to the machines used for processing thermoplastics, which consist of devices for melting, molding, and solidifying, the machines used for producing PUR moldings or semifinished parts contain all the elements of a chemical reactor.

The raw materials are conveyed from the storage tank, optionally via premixing stations for adding auxiliary materials, to the processing machine. Often the metering and mixing machine is called the wet part of the processing plant. Figure 6 schematically depicts a two-component metering and mixing machine with an impingement mixer as the mixing unit. In practice, machines using the impingement mixing principle are available for mixing up to six components. In the working tanks (a), the raw materials are prepared for processing, i.e., primarily their temperatures are adjusted precisely. Any temperature change, especially of the polyols, results in a viscosity change, which can impair later processing. The working tanks (volume 20–500 L according to the machine size) are therefore usually provided with heating and cooling jackets. Separate temperature-control circuits via heat exchangers are also common. The working tank is also equipped with an agitator for homogenization as well as an automatic refilling system and indicators for level, pressure and temperature. The conditioned raw materials are conveyed via metering pumps (b) to the most important part of the machine, the mixing head (c), which delivers the liquid reaction mixture.

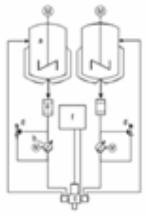

Figure 6. Principle of a high-pressure machine (circulation system)

a) Working tank; b) Metering pump; c) Mixing head; d) Safety valve; e) Filter; f) Hydraulic unit

Depending on the PUR system to be processed, precision metering pumps of various outputs (up to 150 L/min) and pressure ranges (up to 250 bar) are used. The pumps must always satisfy extremely high requirements for metering accuracy in order to maintain a certain reaction mixture. In addition to conventional gear pumps and in-line piston pumps, which have been established in PUR technology for many years, axial piston pumps (Fig. 7) and, for systems containing fillers and reinforcements, slow-running single-cylinder positive-displacement pumps are used (Fig. 8).

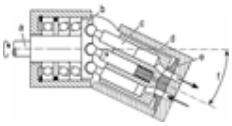

Figure 7. Axial piston pump

a) Shaft; b) Rod plate; c) Piston; d) Cylinder unit; e) Cam plate; f) Angle of inclination to adjust volume flow rate

Figure 8. Plunger pump

a) Stepping motor; b) Hydraulic cylinder; c) Hydraulic piston; d) Metering piston; e) Metering cylinder; f) Stop valve:

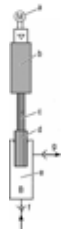

feeding side; g) Stop valve: delivery side

Two types of mixing principles exist: low-pressure mixing heads with an agitator (3–40 bar) and high-pressure mixing heads (100–250 bar). In mixing heads with an agitator, the raw materials entering the upper part of the mixing chamber are intensively mixed on their way to the outlet by rotating, variable-speed agitators. Agitators include spiked, paddle, or screw agitators, propellers, or turbine impellers. After the end of the shot the agitators are cleaned with a rinsing medium. High-pressure mixing heads operate according to the countercurrent injection principle. The component streams are injected into the mixing chamber via nozzles (Fig. 9) and are mixed by virtue of the kinetic energy used and the chaotic flow behavior created [179].

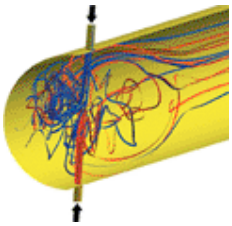

Figure 9. View into an impingement mixer during shot; visualization of streamlines

In controlled-circulation mixing heads, the raw material streams are changed from the “circulation position” (Fig. 10A) (tank ⟶ pump ⟶ mixing head ⟶ tank) to the “injection position” (Fig. 10 B). After injection has occurred, a piston cleans the remaining reaction mixture from the mixing chamber and outlet pipe. These are therefore known as self-cleaning mixing heads.

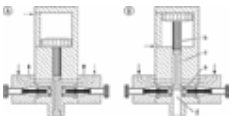

Figure 10. Circulating-groove mixing head

A) Circulation/cleaning; B) Mixing

a) Nozzle; b) Piston; c) Groove; d) Outlet tube

So-called L-shaped mixing heads are able to improve the homogeneity of the reaction mixture by a perpendicular deflection of the flow direction and simultaneous acceleration of the flow velocity (Fig. 11). Furthermore, the larger diameter of outlet tube compared with the dimension of the mixing tube allows the required laminar flow out of the mixing head.

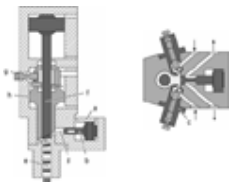

Figure 11. L-shaped mixing head with ring throttle (type Hennecke)

a) Impingement mixer; b) Actuating piston; c) Injection nozzle; d) Mixing chamber; e) Outlet tube; f) Cleaning piston; g) Adjustment of ring throttle; h) Ring throttle

6.3. Processing Plants

A two- or multicomponent metering and mixing machine is part of each processing plant. Further equipment and machines define the kind of processing plant.

In continuous plants, semifinished products are produced in the form of blocks, slabs, boards, profiles, etc. The scheme in Figure 12 depicts a plant for the continuous production of foam slabs. Four components are metered, mixed with an agitator mixing head, and poured onto the conveyor belt. The bottom and side paper are used to keep the conveyor belt clean and to avoid sticking of the foam to the conveyor belt during rising and curing. Figure 13 is a three-dimensional view of a plant for the manufacture of rectangular slab stocks.

Figure 12. Principle of a plant for continuous production of foam slabs

a) Working tank; b) Metering pump; c) Agitator mixing head; d) Conveyor; e) Bottom paper; f) Side paper; g) Rising foam; h) Cured foam

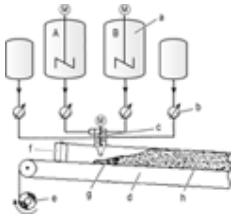

Figure 13. Plant for the manufacture of rectangular slab stocks (Quadro-Foamat-Machine: QFM, source Hennecke)

a) Agitator mixing head; b) Fall plate path adapted to foam expansion; c) Conveyor; d) Device to keep block shape rectangular; e) Paper winch

A three-dimensional drawing of a plant for manufacturing metal sandwich panels (Fig. 14) gives an impression of how large and complex PUR plants can be. The length over all stations of processing steps in this case is ca. 350 m.

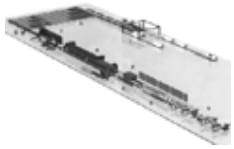

Figure 14. Plant for the manufacture of metal sandwich panels (source Hennecke)

a) Uncoiling station; b) Profiling machine; c) Heater; d) Twin-head foaming portal; e) Laminator; f) Cutting machine/bandsaw; g) Turn over system for roof elements; h) Cooling section; i) Stacker

Plants for producing moldings operate batchwise. The example in Figure 15 shows a top view of a processing plant: On the left side is the wet part of the plant with a total of six components; a robot handles the high-pressure mixing head, and two clamping units are available for pouring into open moulds.

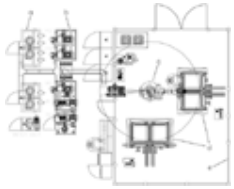

Figure 15. Layout of a plant for producing moldings by open-mold filling technology

a) Tank unit; b) Pump unit; c) Robot with mixing head; d) Clamping units; e) Protective guard

A distinction can be made between plants with movable (Fig. 16) and those with stationary molds (Fig. 17). Movable molds are supplied by a single mixing head, whereas in RIM plants, several mixing heads are used, each of them being permanently assigned to one (stationary) mold. The plants are preferably automatically controlled with electronic equipment.

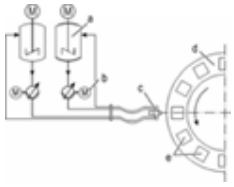

Figure 16. Principle of a plant with movable molds and stationary mixing head

a) Working tank; b) Metering pump; c) Mixing head; d) Turntable; e) Molds

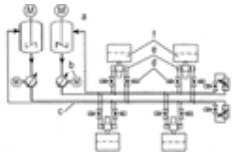

Figure 17. Principle of a RIM plant

a) Working tank; b) Metering pump; c) Circulation system; d) Selection valves; e) Mixing heads; f) Clamping units with tools

[Top of Page]

7. Foams

Norbert Adam, Geza Avar, Sven Meyer-Ahrens

Whilst the reaction of isocyanates with water in most PUR applications is an unwanted side reaction, for foams this reaction is of utmost importance, due to the two reaction products formed. These are CO_2, which acts as the blowing agent for foam formation, and the more or less crystalline polyurea domains, which control cell opening and act as reinforcing filler in the PUR backbone. This chemical blowing process was supplemented in the early 1960s by a physical

blowing process using low-boiling liquids as blowing agents. These liquids are vaporized by the heat of reaction of the polyaddition reaction between isocyanates and polyols and/or water. The blowing agent vapor is trapped in the viscous reaction mixture and thus leads to the cellular structure of the foam. Since foam formation is a very sensitive and complex process of interdependent chemical, physical, and rheological processes, specific additives and catalysts are required (see Section Catalysts and Additives).

PUR foams are generally divided into flexible, rigid, and semiflexible products. Integral skin foams, which have a densified surface layer forming an elastomerlike skin and a foamed core, are described in Section Integral Skin Foams and RIM Materials .

Polyols with a relatively high molecular mass (2000–8000 g/mol) and a nominal functionality of 2–6 yield flexible foams. Combinations of these polyols with low molecular mass glycols or amines as chain extenders and preferably polyisocyanates with a functionality of more than two yield semiflexible foams. Highly branched polyols, e.g., polyethers based on sorbitol or sucrose, with a relatively low molecular mass (<1000 g/mol) are the basis of rigid foams.

For all these classes of foams different types of manufacturing processes may be applied. An alternative way of classifying foams is therefore based upon the manufacturing process.

Depending on the processing, a distinction is made between molded foam and slabstock foam. Besides these most prominent groups of flexible foams, other foam types include cavity-filling foam (e.g., filling of cavities in car bodies for acoustical or reinforcing purposes) and spray foam (e.g., for thermal insulation of roofs). The term “filling foam” essentially refers to semiflexible foams behind appropriate overlays, e.g., flexible plastic sheets, films, etc.

7.1. Flexible Foams

Most flexible foams are produced continuously as slabs or batchwise by molding. They are characterized by reversible deformability and an open-cell structure, associated with air permeability.

7.1.1. *Flexible Slabstock Foam*

7.1.1.1. *Raw Materials*

Based on different chemistries on the polyol side, three fundamentally different types of slabstock have to be considered: Polyester foams, made from polyester polyols, conventional polyether foam, made from polyether polyols with mainly secondary OH groups and high-resilient (HR) foamf, made from polyether foams with mainly primary OH groups.

The isocyanate used for polyester foams is usually TDI 65 (see Section Aromatic Polyisocyanates) or blends of TDI 65 and TDI 80; the typical isocyanate for polyether foams is TDI 80. High-resilient slab stock foam is produced from polyethers and TDI 80, both optionally modified.

Main components (parts by weight) of a formulation for flexible polyester foam:

Polyester polyol (OH no. 65), e.g., Desmophen 2200	100
TDI 65 and/or TDI 80, e.g., Desmodur T65, Desmodur T80	variable
Index	95
Water	2.5–4.8
Amine catalyst	0.3–0.6
Foam stabilizer	0.7–1.3

Main components (parts by weight) of a formulation for flexible polyether foam:

Polyether polyol (OH no. 48), e.g., Arcol 1108	100
TDI 80, e.g., Desmodur T80	variable
Index	98–115
Water	2.0–4.5
Amine catalyst	0.15
Tin dioctanoate	0.12–0.28
Foam stabilizer	0.6–1.5

Main components (parts by weight) of a formulation for high-resilient (HR) flexible foam:

Grafted polyol (OH no. 28), e.g., Desmophen 7653	100
TDI 80, e.g., Desmodur T80	variable
Index factor	95–113
Water	1.5–4.0
Amine catalyst	0.10–0.20
Diethanolamine	0.8–1.5

Tin dioctanoate	0.1–0.35
Foam stabilizer	0.3–1.0

The polyesters (M_r 2000–3000 g/mol) are usually based on adipic acid and linear glycols (e.g., diethylene glycol), or triols (glycerol or trimethylolpropane) if slight branching is desired.

The polyethers used (M_r 1000–6000 g/mol) are di- and/or triols with a varying propylene oxide/ethylene oxide ratio, which is tailored to the foam's application. An important factor for polyether foaming is a well-balanced ratio of the metal and amine catalysts, which controls the simultaneous reactions of the NCO group with hydroxyl groups (polyurethane reaction) and with water (blowing reaction). If the polyurethane reaction rate predominates, the foam may shrink due to the formation of a high proportion of closed cells. If the water reaction proceeds too rapidly, the carbon dioxide formed cannot be contained in the cells; splits result and the foam can even collapse.

High-resilient foams are produced from highly reactive grafted polyethers which contain a high proportion of primary OH groups (from ethylene oxide). Grafting is achieved by in situ polymerization of styrene and acrylonitrile to form a stable SAN suspension in the polyol. Another class of grafted polyols is based on the polyaddition of TDI and hydrazine to yield PHD polyols which contain suspended polyhydrazodicarbonamide/polyurea particles. If highly functional (pre-cross-linked) modified TDI types are used, unmodified polyols can be applied, but due to advantages in foam processing, this approach has been generally replaced by grafted-polyol technology.

7.1.1.2. *Production*

The raw materials are processed in fully continuous plants to give flexible foam slabs up to 220 cm wide, 120 cm high, and of any length. Depending on the slab dimensions, internal temperatures of up to 165 °C are reached as a result of the exothermic reaction. This temperature must not be exceeded, otherwise severe scorching and even self-ignition of open-celled foam may occur. The foaming process is completed after ca. 3 min, and final curing takes from about 10–12 h up to 72 h, depending on the foam type. The slabs are therefore stored in curing storage facilities until they have cooled and attained their final mechanical properties, so that they can be transferred to further processing, fabrication, etc. The length of the slabs varies from 10 m (short slabs) to 120 m (long slabs). To minimize trimming losses of the foam buns, which would be dome-shaped under free-rise conditions, various technologies are applied to produce rectangular slabs.

Figure 18 shows the principle of a rectangular slabstock plant (Hennecke Planibloc process) [180]. A cover is applied to the rising foam. The pressure of the skids avoids the development of the usual slab dome without the foam becoming compacted. Besides a rectangular slab cross section, this process gives an increased yield of foam due to partial foaming of the mixture at the slab surface, which would otherwise form the surface skin.

Figure 18. Principle of a plant for production of rectangular flexible foam slab stocks

a) Covering paper; b) Skids; c) Mixing head; d) Conveyor belt; e) Foam

In other processes a polyethylene film is drawn up to the slab sides in synchronization with the rising movement of the foam (Draka-Petzetakis) [181, 182]; or the reaction mixture is discharged into a trough with a sloping bottom, in which it rises and then flows off over a downward incline (fall plates) onto the conveyor (Maxfoam) [183]. A more recent development (Multiflex, Quadrofoamat, see Fig. 13) combines the fall plates with liquid laydown principles.

The application of reduced air pressure during the foaming process (variable-pressure foaming, Beamech [184]) or the introduction of carbon dioxide into the mixing chamber (e.g., Multistream (Fig. 19) and Novaflex (Fig. 20), Hennecke [185]) has allowed for the complete elimination of other physical blowing agents such as CFCs.

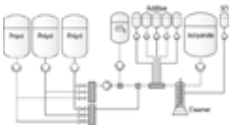

Figure 19. Multistream (Hennecke)

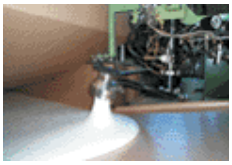

Figure 20. Novaflex creamer (Hennecke)

Rectangular blocks can also simply be produced batchwise by introducing the reaction mixture into metal or wooden boxes and at a given time putting a floating cover on the rising foam. In addition to primitive methods (“box foam”), mechanized box processes are also offered for this purpose with relatively low investment costs and a processing capacity of up to 1000 t/a (e.g., Hennecke Bloc-Foamat BFM 100). Special versions of these batch plants are available so that different ambient pressures can be applied during the foaming stage (e.g., Hennecke Megafoam).

In addition to the production of blocks with a rectangular cross-section, continuously produced blocks with a round cross section (diameter ca. 120 cm) have become industrially important (Reeves Brothers, General Foam [186, 187]). These buns can be peeled seamlessly.

7.1.1.3. *Properties*

The mechanical properties of the foam are largely determined by the raw materials and the formulation used and by its apparent density. For flexible foam slabstocks, the usual densities are 20–40 kg/m^3, but extremely light and extremely heavy types of foam (from 16 kg/m^3 and up to 130 kg/m^3) can be produced industrially for special fields. The choice of density is determined by the specific application of the foam: the higher the density, the better the properties or the lower the fatigue of the foam when in continuous use.

The hardness (measured as compressive strength at 40 % compression) is an important characteristic of PUR foams (Table 4). The hardness can be varied within wide limits by altering the formulation. Considerably softer conventional foams are obtained, e.g., by including polyethers of high ethylene oxide content and special additives. Hardness control is also possible through the index or the addition of grafted polyols.

Table 4. Relation between apparent density and compressive strength of various grades of flexible foam slab stock

Flexible foam system	Apparent density (ISO 845), kg/m^3	Compressive strength at 40 % compression (ISO 3386), kPa
Polyester foam, based on Desmophen 2200B/Desmodur T65	35	6.5
Polyether foam, based on Polyether 10WF15/Desmodur T80/T65	36	6.0
Polyether foam, based on Arcol 1108/Desmodur T80	36	4.5
HR slab stock foam, based on Desmophen 7619W /Desmodur T80	34	2.7

Apart from the hardness, other characteristic values can be derived from the hysteresis loop (Fig. 21). The slope of the loading curve and the area enclosed by the loading and unloading curves (hysteresis) are measures of the elasticity and energy absorption of the foam: the closer the curves are together, the more elastic the foam; and the larger the area, the better the energy absorption or damping properties of the foam. HR slabstock foam has the best elasticity, followed by polyether foam and polyester foam.

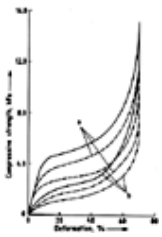

Figure 21. Hysteresis loops of different grades of flexible foam slab stock (apparent density 35 kg/m^3)

a) Loading lines; b) Unloading lines

— Polyester-based foam; – – – Polyether-based foam; – · – · – HR foam

Some important mechanical properties are summarized in Table 5 to demonstrate their range of values. Table 6 shows the values for fatigue in continuous use under the conditions of short-duration tests.

Table 5. Mechanical properties of various grades of flexible foam slabstock

Flexible foam system	Tensile strength (ISO 1798), kPa	Elongation at break (ISO 1798), %	Compressive set (after 90% deformation; ISO 1856), %
Polyester foams	160 – 220	100 – 400	5 – 15
Polyether foams	100 – 180	100 – 400	2 – 10
HR slabstock foams	50 – 120	80 – 200	2 – 10

Table 6. Change of hardness on ageing of various grades of flexible foam slab stock

Flexible	Apparent density	Decrease of indentation	Decrease of compressive

foam system	(ISO 845), kg/m^3	load deflection (ISO 3385), %	strength (5 d , 90°C, 95% rel. hum.), %
Polyester foams	25	43	35
	35	29	40
	45	26	40
Polyether foams	20	34	10
	25	29	14
	35	19	18
	50	16	16
HR slabstock foams	25	24	8
	35	18	13
	45	16	18

The following list contains some additional properties of flexible PUR slabstock foams:

Thermal conductivity: 0.033–0.040 W m^{-1} K^{-1} (at 0–70 °C and densities of 20–70 kg/m^3)

Temperature stability: from –40 to +80 °C without embrittlement.

Open-cell character (see also reticulation capability): 98–100 % with respect to air permeability, not with respect to resistance to air flow.

Chemical resistance to acids and alkalis: polyether foam has a better resistance to acids and alkalis than polyester foam.

Chemical resistance to organic solvents: polyester foam shows a better resistance to organic solvents than polyether foam.

Chemical resistance to atmospheric oxygen: oxygen attacks foam slab stocks only superficially without harming the serviceability. Polyether foam is more easily oxidized than polyester foam. However, if aromatic polyisocyanates and polyether polyols are used for foam production, the foam will decompose over time if exposed to oxygen and light.

Combustibility (see Section Fire Performance of Polyurethanes): Flexible slabstock PURs are combustible, like all organic (carbon-containing) materials. By addition of suitable flame retardants the combustion behavior of foams can be adjusted to meet several flammability standards such as MVSS 302, BS 5852 part 2 Crib V, or even Fire Class B1 (DIN 4102). Also a post-treatment with latices containing flame retardants is possible, in particular if demanding tests have to be met. Besides flame retardants, the type of foam (polyester, conventional, or high resilient) is important if certain flammability standards have to be met.

Bactericidal and fungicidal treatment is possible by including appropriate additives in the formulation or by post-treatment.

Reticulation capability (see also open-cell character): the flow resistance can be greatly reduced by using chemical or physical processes (Chemotronics) to remove residual cell membranes (filter foams).

Physiological aspects: According to extensive studies (both dermatological and oral toxicity in animals), flexible PUR foam of defined composition is physiologically harmless.

7.1.1.4. *Trimming and Processing*

Various possibilities and machines are available for processing the cured slabstock foam to finished articles. The simplest processes for foam blanks are horizontal-cutting (e.g., long-block cutting/peeling), vertical cutting, and angular cutting. These are carried out with continuously rotating knives.

Profile cutting in the form of copy cutting and punching, compression cutting, and milling is used for further forming. With profile cutting tailor-made cuts can be made by following prefabricated templates or drawings (Fig. 22).

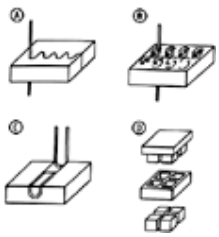

Figure 22. Cutting methods and punching

A) Hot-wire cutting; B) Copy cutting; C) Profile cutting; D) Punching

Compression-cutting is a special variant of profile-cutting. All these processes exploit the easy deformability and excellent recovery of PUR flexible foam (Fig. 23).

Figure 23. Compression cutting

A) Convoluting process; B) Skiving process

Advances in foam trimming equipment such as fully automated CNC processing lines allow for economical conversion of foam buns into very sophisticated contoured parts (Fig. 24). At the same time, trim losses are minimized by specific software [188].

Figure 24. Contoured parts, produced by different trimming technologies

Form milling is used for rounding edges on seat backs and armrests as well as on seat pads. Due to the large variety of milling tools, any shape is possible (Fig. 25).

Figure 25. Shaping of flexible foam by milling

Other post-treatment methods include reticulation, impregnation, compression, welding, cementing, laminating with textiles or films by flame lamination or adhesive bonding, quilting, and covering.

The trimmed-off foam is flaked in suitable mills. The flakes can be used as filling material or are bonded to form flake-composite blocks, referred to as rebonded foam. These composite blocks are then fabricated similarly to normal foam slab stocks (see Section Recycling/Recovery of Polyurethanes).

7.1.1.5. *Applications*

The variety of properties and the continuous development of new foam grades and further processing techniques have made slabstock indispensable for many applications.

For furniture, the use of polyether foams extends from complete foam upholstery to the simple seat cushion. Very comfortable cushioning is achieved by combining types with different properties (e.g., density, hardness, or elasticity).

Full PUR foam mattresses at present account for approximately one-third of all mattresses sold in Europe. The relative low weight (e.g., versus latex foam) and the freedom of design by combining various foam types (high resilient, conventional, viscoelastic) and grades (density, hardness) in conjunction with advanced cutting systems have turned these mattresses into high-tech applications. All required grades of hardness are available, from the child's to the intervertebral disk mattress. Even in steel-spring mattresses, a layer of slabstock foam is used as cushioning and lining material.

An important field of application for slabstock foam is the lamination of fabrics. Polyester foams are preferred to polyether foams due to their melting properties and their specific cell structure. The foams can be bonded on one or two sides with textile widths by adhesive or flame lamination and are widely used for automotive interior trim applications such as seat covers or headliners.

Other applications in automotive industry are sun visor fillings, which can be manufactured from semirigid polyester foams, and sound-absorption materials for passenger, engine, and trunk compartments. PUR foams are also used as sealing and filter materials in the ventilation system. All foams for automobiles must satisfy the fire test MVSS 302 (see Section Fire Performance of Polyurethanes). Depending on the density and the chemical nature of the PUR foam (HR or conventional foam) flame retardants must be added to the foam formulation to meet this requirement.

In the clothing industry the above-mentioned laminated fabrics are incorporated into jackets and coats. Other applications are paddings in ski and sports boots and underlinings in the purse-making industry.

Flexible PUR foam is particularly suitable as a packaging material for fragile goods. Because of the variety of cutting processes, extremely complicated shapes can be produced for the transportation of goods.

The main application in the household sector is sponges. They are produced from both polyether and polyester foam in very diverse colors and shapes. They can be fine- or coarse-pored and range from extremely flexible to almost rigid. The foam's hydrophilicity can be increased by using special additives. The range of uses in the household includes adhesive, sealing, and jointing tapes as well as undercarpet padding, doormats, and cleaning brushes. Special grades can be obtained by bonding with thermoplastic films by high-frequency welding.

The ability of flexible foams to absorb airborne noise efficiently is exploited for noise abatement. The low thermal conductivity of flexible block foams is used for the thermal insulation of heat accumulators and pipes. For use as filter material, e.g., in ventilation and air-conditioning systems or as dust filter in breathing masks, the flow resistance of the foam must be reduced by post-treatment. Depending on technical requirements, the foam can also be used as a filter for liquids. Since the foams do not absorb X rays, they are an ideal material for the comfortable support of parts of the body during X ray exposure.

7.1.2. *Molded Flexible Foam*

The foam molding process is used when it is impossible or uneconomic to produce a complex geometry by trimming slabstock foam, or when the incorporation of metal inserts (for fastening cushioning units or coverings) or even steel springs (for furniture applications) is desired. A distinction is made between hot-cure molded foam and cold-cure molded foam.

7.1.2.1. *Production*

To produce hot-cure foam, trifunctional polyethers with a low content of highly reactive primary OH groups in the molecular mass range of 3000–5000 g/mol are reacted with toluene diisocyanate (TDI) with ca. 80 % of the 2,4′ isomer, e.g., Desmodur T80. The blowing agent is carbon dioxide, which is formed by the water-isocyanate reaction. Polyether polysiloxanes are employed for stabilizing the foam and due to the lower reactivity of the polyols, tin dioctanoate or other tin compounds are used to catalyze the polyurethane reaction, amines are used to catalyze the blowing reaction, and rather high mold temperatures must be applied. The hardness of the foams can be adjusted by adding grafted polyols for increased load-bearing properties or by using special polyols or additives for softer grades. The resulting foam is open-celled and does not require crushing for cell opening.

Cold-cure foams are produced from trifunctional polyethers with enhanced reactivity (high proportion of primary OH groups) in the molecular mass range of 5000–6000 g/mol and grafted polyols. A variety of different isocyanates are used, usually with a functionality exceeding two (modified TDI types, blends of TDI and higher functional MDI; special MDI types). If pure TDI 80 is applied, polyethers with a functionality slightly higher than three are usually used (e.g., Hyperlite). No further blowing agents are used; the carbon dioxide formed from the reaction of water with isocyanate is the sole blowing agent. Due to the higher reactivity of the polyols no metal catalysts are required for curing, and the mold temperatures are relatively low (“cold” cure) compared to hot-cure foam. On demolding the cold-cure foam is closed-celled and must be mechanically crushed to avoid shrinkage and to achieve consistent physical properties.

7.1.2.2. *Molding Process*

The molding methods for hot- and cold-cure foam systems differ sharply from each other. This is mainly because external heat must be supplied to a hot-cure molded foam to cure the foam's surface fast enough.

Figure 26 shows the individual process steps of the hot-cure foam molding process. This production sequence has a cycle time of ca. 20–25 min, depending on the nature and size of the foam parts.

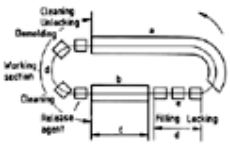

Figure 26. Plant for hot-cure moldings

a) Oven (160–250°C); b) Recirculating air/water bath; c) Cooling section; d) Working section; e) Inserts

The molds consist of black iron plate (wall thickness 1.5–2.0 mm), aluminum sheet (wall thickness 4–6 mm), or cast aluminum (wall thickness 6–10 mm). Since the foaming process for this foam type must proceed unpressurized, good mold venting is necessary. The arrangement of the vent holes is determined experimentally. In manufacturing the molds, a volume loss during the curing process of ca. 2 % must be taken into consideration. The precise value is specific for the formulation and must be determined experimentally. Release agents are mainly aqueous wax or soap dispersions. High- or low-pressure machines, operating batchwise, are used for metering the raw materials. The molds are usually heated by hot-air ovens. The heating capacity of the ovens for the full-cure process must be adapted to the mold material and the required cycle time. After removing the molded foam parts, the molds are cooled by water or cold air to the loading temperature.

The curing process can be shortened by rapid adjustment of the internal wall temperature of the mold to ca. 120 °C after the end of the rising time. In this case, the molds are heated directly with liquid heating media to conserve energy.

Cold-cure foam systems require much lower mold temperatures. To achieve the optimum quality of the foam skin, the molds are adjusted to the optimum temperature for the mold material and the release agent (between 40 and 65 °C). Cooling equipment is not required (Fig. 27). Depending on the combination of polyols, the isocyanate, the catalyst package, and the shape of the cavity, the foam can be demolded after 2–6 min.

Figure 27. Plant for cold-cure moldings

T = thermostatting units

In contrast to open-celled, hot-cure molded foam, the closed cells of cold-cure foam parts must be mechanically opened. This can be achieved by roller crushers with counter-rotating rollers, compressing the parts to a fraction of their original

height, or by fast alteration of the ambient pressure in a pressure chamber.

One hour after demolding, cold-cure molded foams have reached only 50–80 % of their final hardness. Therefore, cold-cure foam parts need to be handled with care in the first few hours after production. Depending on the chemistry and the climatic conditions, final properties are achieved after 5–12 h (for comparison: a hot-cure foam will reach its final hardness within the first 2 h after demolding).

The foam molds must be designed for an internal mold overpressure of 0.3–0.6 bar. When fabricating the master patterns for the mold replica, a loss in volume of the molded part of up to 2 % must be taken into account.

Larger plants (greater space requirement, more molds) are required for the production of hot-cure foam than for cold-cure foam, owing to both the heating/cooling cycle and the longer mold holdup times of ca. 12 min compared to 4–8 min for cold-cure foam. Under ideal conditions, hot-cure molded foam has the advantage of a lower molding mass compared to cold foams at the same compression-load deflection.

7.1.2.3. *Properties*

Hot-cure and cold-cure foams differ fundamentally in their mechanical properties. The lowest densities can be achieved with hot-cure foam. At the same density, a hot-cure foam will usually provide a higher hardness. However, due to higher venting losses, savings in material for a hot-molded piece of foam are significantly lower than expected based on the simple comparison of apparent densities. The compression load deflection curves of molded foams can be compared with that of free-rise flexible foams (Section Properties): the characteristics of hot-cure molded foam are similar to those of conventional slabstock foam; cold-cure molded foam can be compared with HR slabstock. The hysteresis curves show that cold-cure foam has a higher elasticity (narrower hysteresis loop) than molded parts from hot-cure foam. Table 7 gives a general overview of the properties of hot- and cold-cure foams.

Table 7. Average properties for seat grades of flexible molded foams

Property	Hot-cure foam	Cold-cure foam	
		TDI based	MDI based
	HC	HR-T	HR-M
Apparent density (EN-ISO 845), kg/m^3	28 – 40	35 – 50	45 – 65
CLD-Hardness at 40 % (EN-ISO 3386-1), kPa	3 – 7	2 – 8	4 – 12
Compression set 50 %, 70 °C, 22 h (EN-ISO 1856), %	2 – 4	4 – 8	4 – 8
Elongation (EN-ISO 1798), %	120 – 160	100 – 150	90 – 120
Tensile strength (EN-ISO 1798), kPa	80 – 140	130 – 200	100 – 160
Tear strength (ASTM D 3574-F), N/m	200 – 400	200 – 450	160 – 250
Humid ageing (OEM specific)	very good	good	good – very good
Hysteresis (Energy Dissipation)	high	low	medium

With a cold-cure foam cushion, greater comfort is experienced due to its higher elasticity. However, for an automobile seat this higher elasticity means a reduced damping of the vehicle's vibrations. Since the construction of the seat (full foam or combination of foam and springs) and the damping behavior of the vehicle structure must be matched, the question whether to use hot-cure or cold-cure molded foam must be decided for each individual case. A second and more basic difference between hot- and cold-cure foam is their fire behavior. Cold-cure foams down to a density of ca. 35 kg/m^3 usually meet the fire standards of the automobile industry, even without added flame retardants. For hot-molded foam, flame retardants must be used.

7.1.2.4. *Applications*

Flexible molded foams are principally used in all kinds of seat pads, seat cushions, backrests, bench seats, headrests, etc. in passenger cars (see Fig. 35), motorcycles, and commercial vehicles, aircraft, and railways. Another important area of application in the automotive industry is sound abatement. Flexible molded foam is used either on its own for absorption of airborne noise or as a composite in conjunction with a heavy layer (EPDM, EVA, or PUR elastomers filled with, e.g., barium sulfate) or as backfilling foam on the vehicle’s floor cover. These composites may act either as sound barriers or can be tailored to dampen structureborne noise. In addition, flexible molded foams are used in the furniture industry, e.g., for office chairs, armrests, and in specific combinations with steel springs, which are incorporated into the molded foam. Sealing lips for passenger car tail lamps and for filters can also be produced from molded foams.

7.2. Semirigid Foams

7.2.1. *Applications*

Semirigid foams are used primarily for improving inner safety and comfort in automobiles. Combined with an external layer of ABS/PVS or PUR skin and a metal or plastic carrier, they are used to produce instrument panels (dash boards), door

and pillar covers, armrests, and consoles. The foam has a density range of 80–180 kg/m^3 in the finished part. Key properties are long-term adhesion to skin and carrier, a uniform, fine cell structure, low emission, and high resistance to aging and heat.

Special grades (EA foams) are optimized for energy absorption. On impact, they dissipate the kinetic energy by deformation and thus protect the passengers in a crash (Fig. 28).

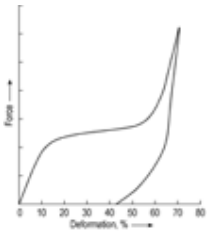

Figure 28. Dynamic force–deformation characteristics of a semi-rigid PUR foam

7.2.2. *Production*

Parts made of semirigid foam are produced mostly in a two-step process. First, the preformed skin (ABS/PVC by thermoforming or PUR by spraying) is positioned in the cavity of the mold. A plastic or metal carrier is fixed on the lid of the mold. In the second step, the mold is closed and the semirigid foam is injected. Alternately, the semirigid foam is spread in the open mold and the lid is closed before the foam expands.

The two PUR components are mixed by high-pressure machines. The molds are made from aluminum, often coated with epoxy resin. If the foam has contact to the surface of a mold, a wax-based release agent must be used. The demold times are between 1 and 6 min, depending on the size, thickness, and geometry of the parts and on the reactivity of the foam system.

7.2.3. *Properties*

Semirigid foams are mostly open-celled and have a higher compression load deflection and lower elasticity than flexible foams. Typical properties are:

Apparent density (ISO 845)	80–180 kg/m^3
Tensile strength (ISO 1798)	300–400 kPa
Elongation at break (ISO 1798)	30–60 %
Compression load deflection at 40 % compression (ISO 3386)	80–200 kPa
Compression set at 50 % compression (ISO 1856) 22h at 70 °C	< 10 %

The dynamic impact characteristics (Fig. 28) are important for the dampening and energy absorbing behavior.

7.3. Rigid Foams

Rigid foam is in most cases a closed-cell material that is used for thermal insulation. Besides providing thermal insulation the foam may also have to bear mechanical loads in some applications. Composite structures in which the self-adhesive effect is used to bond the foam to a substrate during the foaming process are often found in the field of rigid foam. This reinforcing sandwich-effect improves the mechanical properties significantly.

Rigid foam is used in many ways. The main applications for rigid foam are appliances (refrigerators, freezers), insulation boards (with and without facings), composite panels (mostly with metal facings), pipe insulation (as insulating shells or as composites), insulated containers, and in situ foam (spray foam).

7.3.1. *Raw Materials*

Rigid PURs require a highly cross-linked molecular structure. To achieve this, raw materials with high functionalities and low molecular weight are preferred on the polyol and on the isocyanate side. Therefore, short-chain polyether polyols and polymeric MDI are the materials of choice. Low molecular weight polyester polyols are also used.

Typical polyether polyols (see Section Polyether Polyols) are produced from a highly functional starter molecule (sucrose, sorbitol, glycerol, diethylene glycol, ethylenediamine, and others) or mixtures thereof by adding alkylene oxides, mainly propylene oxide but sometimes ethylene oxide. The resulting products are viscous liquids.

The *polyester polyols* (see Section Polyester Polyols) are typically based on aromatic acids like phthalic acid. (Di)ethylene glycol is mostly used as alcoholic component. Glycerol or other highly functional alcohol components may be used to increase the functionality. However, the use of starting materials with higher functionalities is limited because of the increasing viscosities of the resulting products.

The *polymeric* MDI types (see Section Aromatic Polyisocyanates) that are used have functionalities in the range of roughly 2.5 to 3.5, corresponding to viscosities of about 100 to 1000 mPa · s at 25 °C.

Blowing agents (see Section Catalysts and Additives) are not only used to create the foam. Since the blowing agent remains inside the closed cells it has a strong influence on the thermal conductivity of the resulting foam. The choice of the

blowing agent depends on the most important criteria in the respective application, such as the lowest possible thermal conductivity, processability, or relationship between cost and performance. In most cases, a combination of a physical blowing agent (a low-boiling liquid or a gas) and a chemical blowing agent (mostly carbon dioxide formed by the chemical reaction of water with isocyanate) is used. While, according to the Montreal Protocol, the use of chlorine-containing halogenated hydrocarbons is still allowed in certain areas of the world, the use of these products in Western Europe is banned. Here only fluorinated hydrocarbons such as tetrafluoroethane (R 134a), pentafluoropropane (R 245fa), and pentafluorobutane (R 365mfc) may be used. Hydrocarbons such as n-, iso-, and cyclopentane or mixtures with other hydrocarbons like propane or butane are widely used, despite their flammability. Methods for a safe handling of these products are industrially available and installed. All these blowing agents have very small coefficients of diffusion; they remain within the foam during its lifetime. Where no (metallic) facings are provided, the thermal conductivity increases very slowly because of air diffusing into the foam. Therefore, in some applications, e.g., the construction industry, increments are added to the measured values of the thermal conductivity to take this ageing process into account. In this way it is ensured that a sufficient level of insulation is installed.

Catalysts (see Section Catalysts and Additives), sometimes called activators, are used to control the foaming process. Products used as catalysts are tertiary amines, organometallic compounds, and alkali metal and quarternary ammonium salts of carboxylic acids. The products may be classified according to their performance-without really being able to draw sharp lines-as blowing, gel, and trimerization catalysts. A few examples taken from a wide range of products are given here: bis[2-(N,N-dimethylamino)ethyl]ether and *N*,*N*-bis(*N′*,*N′*-dimethylaminoethyl)methylamine as blowing catalysts; triethylenediamine (dissolved) and dibutyltin dilaurate as gel catalysts and potassium acetate (dissolved) and 1,3,5-tris(3-(dimethylamino)propyl)hexahydrotriazine as trimerization catalysts.

Foam stabilizers (see Section Catalysts and Additives), also called surfactants, are required to prevent the small bubbles in the liquid reaction mixture from coagulating. Fine-celled foams can be prepared by using these products. To produce foams with a fine cell structure, some air (or other inert gases) may be dispersed into the polyol component or into the reaction mixture for nucleation. A fine cell structure is important for a low thermal conductivity. Chemically, most of the products used in the industry are polyether-modified polysiloxanes. The structural variations possible by using different polyethers and different polysiloxanes lead to the great number of products available on the market, optimized for various purposes. Some of these purposes are: flowability (describes the expansion of the foam into narrow cavities), adhesion to facings (the force required to remove the facing), fine cell structure (small cells of similar size), and flammability (resistivity to flames or high temperatures).

Flame retardants (see Section Catalysts and Additives) are used to make the foams less flammable. The methods for testing flammability are manifold. However, official requirements regarding flammability are in place for some applications, e.g., in the construction industry. Since PUR foams are based on organic molecules, it is only possible to achieve certain flammability classifications within the group of flammable materials. Flame retardants used in the field are usually substances based on halogens and/or phosphorus. Some products have OH groups, so they can react with isocyanate and are incorporated into the macromolecule; others are nonreactive. Examples of reactive flame retardants are polyethers made from dibromobutenediol and epichlorohydrin or brominated aromatic polyester polyols. Examples of nonreactive flame-retardants are tris(2-chloroisopropyl)phosphate, diphenyl cresyl phosphate, and dimethyl propyl phosphonate. Apart from red phosphorus, which is used in the production of rigid slabstock foam by low-pressure mixing, solid flame retardants have hardly been used in the field of rigid foams. A technique has been developed for processing solid flame retardants continuously on high-pressure machines [189].

7.3.2. *Processing*

Mixtures of the raw materials mentioned above are processed to produce rigid foams. The foaming process is described and monitored by periods of time taken from the start of mixing:

The *cream time* is the time until the reaction mixture starts to expand. At this point in time, the exothermic reaction has heated the reaction mixture to a temperature at which the blowing agent is starting to evaporate. The gas formed remains in the reaction mixture and forms small bubbles that grow with increasing temperature and further formation of gas.

When the cross-linking of the polymer has progressed to the point where the liquid reaction mixture starts turning into a solid the *gel time* can be observed. This is mostly done by prodding the foam with a small stick. When fibers of polymer are pulled from the rising foam the gel time has been reached.

After that the foam continues to rise to its final height (*rise time*).

The *tack-free time* is the time required for the surface of the foam to lose its tackiness. The tack-free time may occur before or after the rise time.

The time required to cool down the foam depends on its thickness. For slabs 1 m in height it takes more than 24 h until ambient temperature is reached in the core. In all processes the foam applies pressure on the compartment in which it is contained. To keep the parts produced within their specified dimensions, the molds must withstand this pressure until the foam has reached a minimum strength and the part can be demolded (demold time). In continuous processes, the line speed needs to be adjusted to keep the foam in the machine for the time required to cure sufficiently. For thicker parts the demold times are longer than for thinner parts, and therefore the line speeds in a continuous production are slower for thicker parts.

Preblending some of the components-usually into the polyol side-for easier operation can be done at different levels. It can be performed by the raw material supplier or by the processor. This depends mostly on the variability required for the respective process (e.g., adjusting reactivity in continuous processes).

A discontinuous process (sometimes called a one-shot process) is used in cases where particular cavities are to be filled,

such as in the production of refrigerators, freezers, insulated containers, and certain sandwich panels. Figure 29 shows the principle of filling refrigerator cabinets.

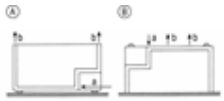

Figure 29. Principle of filling refrigerator cabinets

A) Door opening on top; B) Door opening on bottom

a) Pour-in opening; b) Ventilation openings

Figure 30 shows the discontinuous production of sandwich panels in a multistage press using the mixing head pull technique.

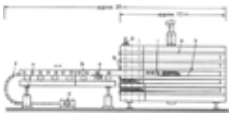

Figure 30. Production line for the discontinuous production of PUR composite elements

a) Plates of the press; b) Carrier on rolls for the hoses leading to the mixing head; c) Lifting device; d) Foam machine; e) Driving elements for horizontal movement; f) Cable holder; g) Device for closing the press; h) Closing cap for each unit; i) Mixing head with hoses; j) Opening device; k) Metal facers

In most of these applications only one type of reaction mixture is used for the whole production range. Here all the necessary components are preblended into the polyol so that only two components, i.e, this polyol formulation and the isocyanate component, must be provided. The components are tempered and metered to the mixing head by high-pressure pumps. The mixing itself is performed by using the high-pressure impingement mixing technique. The amount of reaction mixture necessary to fill the cavity is determined by the volume of the cavity and the density of the foam. The duration of the dosage is then given by the output of the pumps. This shot time must be shorter than the cream time to avoid pumping liquid reaction mixture into an already expanding foam. In so doing, the foam would be partially destroyed with negative effects on all of its properties.

In the continuous production of composite panels and insulation boards on double conveyor belts, the high-pressure impingement mixing technique is also used in most cases. The principle of this process is shown in Figure 31. Here the low weight of the mixing head and the hoses leading to it enable the oscillating of the mixing head across the moving bottom facing for an even distribution of the reaction mixture.

Figure 31. Principle of the continuous production of insulation board

A) With flexible facings; B) With profiled metal facings

a) Unwinding device; b) Lay-down with mixing head; c) Means for folding paper; d) Foaming reaction mixture; e) Compression; f) Cooling area; g) Longitudinal cutting; h) Transverse cutting; i) Coil unwinding; j) Corrugation rolls; k) Heating elements; l) Gate with moving high-pressure impingement mixing head

In the special case of the production of insulation boards on high-speed laminators at line speeds of up to 60 m/min, the high-pressure impingement mixing technique is used in most cases, but without oscillating the mixing head. The oscillation frequencies to ensure an even distribution of the reaction mixture on the bottom facing would be too high.

Here even distribution is achieved by passing the reaction mixture through a narrow gap between the top and the bottom facing (calibration technique). By this technique the small voids near the top facing that occur with the oscillating mixing head technique are eliminated. In all these applications it is necessary to adjust the reactivity of the reaction mixture according to the thickness of the board or panel. Also the amount of blowing agent must be adjusted. This means that the degree of preblending is lower than in a discontinuous process. Catalyst, blowing agent, and possibly other additives are metered separately into the polyol stream. Thus, only two components are finally led to the mixing head, keeping the moving weight low for the oscillation.

In the continuous production of rigid slabstock foam low-pressure mixing is used due to the high output. Here the components are metered into a mixing chamber and mixed by a mechanical stirrer. Separate metering of the components into the mixing head can be used. Preparing a polyol component containing all necessary components is also possible and allows for a certain quality control prior to foaming.

7.3.3. *Properties*

The most important property of rigid PUR foams is its low thermal conductivity, which is the lowest among the different insulation materials available by a significant amount. This allows for a smaller thickness of the insulating layer to achieve the same insulation value. The thermal conductivity of closed-cell PUR foams in the density range 25–60 kg/m^3 depends not so much on the apparent density (usually the thermal conductivity increases slightly with increasing density) but is dominated by the thermal conductivity of the blowing gas. The thermal conductivity can initially be near 0.020 W m^{-1} K^{-1}, and after ageing (cell gas diffusion) typically is in the range of 0.025–0.030 W m^{-1} K^{-1}.

The adhesion to substrates is important for composite constructions. The force needed to remove the foam from the substrate is the criterion used to judge adhesion.

Mechanical properties like compressive strength, tear strength, and shear strength are also important in load-bearing applications. Their values strongly depend on the density and the cell structure (Fig. 32).

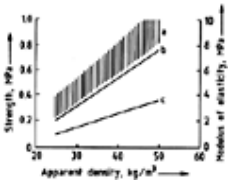

Figure 32. Relationship between typical properties and apparent density of rigid PUR foam (values at room temperature)

a) Modulus of elasticity; b) Tensile strength; c) Compression strength

Dimensional stability must be given under the conditions of use of the respective application. Dimensional stability is tested at low temperatures (e.g., for freezers) and at high temperatures (e.g., for hot water tanks).

Flammability is a very important property of the foam or the composite or both, depending on the test that needs to be passed. Various test methods are in place, most of them in the construction and transportation industries. For example, a harmonized test for the construction industry has been introduced in the EU. The new test procedure comprises a test on the foam (small burner) and the single burning item (SBI) test on the final product under end-use conditions. Other test procedures, e.g., those set up by insurance companies, are also in place.

For the different applications the importance of the various properties is prioritized differently.

For appliances (commercial and domestic refrigerators, freezers, hot water tanks), a low thermal conductivity is of top priority. In these applications a maximum useable space is desired at minimum or given outer dimensions. For this reason, the thickness of the insulation is limited, and only with a foam that offers low thermal conductivity can the energy consumption be minimized. Dimensional stability at low and high temperatures is the other important property here. In the low temperature applications the foam also fulfils a structural task, it forms a type of sandwich structure with the inliner and the outer facing.

Preinsulated pipes are mostly used for district heating. Here the thermal conductivity and the lifetime expectancy are the important properties. Accelerated ageing tests at elevated temperatures are performed to ensure a lifetime of, for example, at least 30 years according to EN 253.

Insulation boards, predominantly those with facings (aluminum foil, mineral fleece, papers) produced on double conveyor belts, also require low thermal conductivity and dimensional stability. On larger insulated areas, e.g., flat roofs, shrinkage would lead to areas without insulation due to gaps between the boards, while swelling might make the boards lift up and break the protective cover. For decorative boards also the appearance is important. Wrinkled facings, mostly aluminum foil, or sink marks due to voids under the facing are unacceptable here. Since most of the products are used in construction, flammability is also an important property.

Composite panels, mostly produced continuously on double conveyor belts with profiled steel sheets as facings, are mainly used in the construction of walls and roofs of large buildings like warehouses or factory buildings. The prefabricated self-supporting panels are attached to a bearing frame construction, so that the complete wall or roof can be put together very fast. The most important properties here are adhesion of the facings, dimensional stability, flammability, and thermal conductivity. The appearance is also important, especially for panels in facades. The thickness of the panels depends on the use of the building. For cold storage houses, where goods are stored at -40 °C, panels up to 24 cm thick are used. The thickness of panels for factories (from about 20 to 100 mm) depends mostly on the local climate. Discontinuously produced panels may have locks for connecting panels with one another that are incorporated into the foam during the foaming process. These panels can be used without bearing constructions, e.g., for walk-in-coolers. Another application is the use of panels in roll-up doors; large ones for truck entrances in industrial buildings, and small ones for private garages.

Slabstock foam is made for applications in very different fields. A few examples are given here: low-temperature insulation (liquid gas tanks, also in ships); refrigerated containments for lorries, where the foam is covered with glass reinforced plastic; insulation for parts that are difficult to mold such as valves and flanges (here the foam is mechanically shaped); wedge-shaped boards to give flat roofs a slope; and others. Different properties are required in these different applications, so rigid slabstock foam is a specialty field. Foams with a range of densities from about 30 to over 200 kg/m^3, different flammability classifications, foams with glass fiber reinforcements, and other variations are found here.

7.3.4. *Special Types*

Foam from pressurized cans, mainly one-component-foam, is used for mounting and fixing applications such as windows and doorframes. Here all the raw materials, including blowing agent and isocyanate, are blended in the can to form a prepolymer with residual NCO groups. The propellant, that is, liquid under pressure, acts first as thinner for the prepolymer. When expelled the propellant acts as blowing agent to produce the foam. The foam itself is usually an open-cell product.

A few special types of open-cell foam exist, such as thermoformable rigid foam that is used in passenger car headliners, foam for flower-arranging purposes, and packaging foam. Recently some attempts have been taken to evacuate open cell foam to reduce the thermal conductivity even further. One problem here is to find suitable facings that ensure that the vacuum is maintained during the lifetime of the product.

In certain cases the foam must be applied in situ. Two methods are predominant: spraying and layer-by-layer application. Spray foam is mostly used for the industrial insulation of storage tanks or in the insulation of buildings. While spray foam on walls is only for insulation, foam on roofs is also applied for sealing purposes. Two-component systems are used here. The components are fed through heated hoses to the mixing head, which is designed like a spray gun. The reactivity of these systems is so high that the foam cures within a few seconds. In layer-by-layer application the reaction mixture is

poured between two substrates, e.g., the outside of a tank and the casing. This method is used where the application of the foam in a factory is not possible, e.g., because of the size of the tank. Here also two-component systems are used. In both cases special attention must be paid to temperature and humidity conditions. Foams should not be applied on cold or moist surfaces. Otherwise, problems with bonding the foam to the substrate may result.

Polyisocyanurate (PIR) foams [190-195], more precisely to be understood as polyisocyanurate-modified PUR foams, are products in which isocyanurate structures are incorporated into the macromolecule. These structures are formed due to the reaction of excess isocyanate with itself. Special (trimerization) catalysts are required. Three NCO groups form the cyclic isocyanurate structure, which shows a higher decomposition temperature than PUR groups. Therefore, these foams are used where flammability or heat resistance is an issue. However, this highly cross-linked structure increases the brittleness of the foams. The overall cross-linking density is controlled by the use of low-functionality polyols. In the continuous production of insulation boards on double conveyor belts, which is the major application for this type of foam, the PIR structure is usually combined with (aromatic) polyester polyols. The higher decomposition temperatures of esters compared to ethers also leads to an additional positive effect on the flammability of the foam. In the manufacturing of these foams, special attention must be paid to the temperature of the laminator because the trimerization only takes place at temperatures above 60 °C. Apart from the reduction of brittleness the polyols have the task of heating the reaction mixture by their exothermic reaction with isocyanate. When the required temperature is reached, the trimerization starts, which can often be observed as a second phase in the expansion of the foam.

7.4. Integral Skin Foams and RIM Materials [196-198]

Integral skin or self-skinning foams are products with a cellular core and a cell-free surface that are formed in a mold in a single operation (Fig. 33). Parts made by the reaction injection molding (RIM) process are microcellular materials with a high surface quality that are mostly painted.

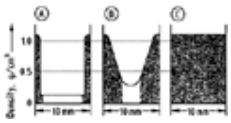

Figure 33. Principle of density distribution for integral foam boards of 10 mm thickness and various apparent densities:

A) 0.2 g/cm^3; B) 0.6 g/cm^3; C) 1.1 g/cm^3

7.4.1. *Applications*

Flexible self-skinning foams are used mainly in the automotive, furniture, and footwear industries. Typical automotive applications are steering wheels and gearshift knobs, but also armrests and headrests (Fig. 34). In the footwear industry, integral skin foams are used for soles of sport, leisure, and safety shoes.

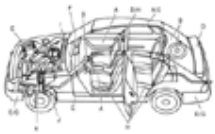

Figure 34. Use of PURs in the passenger car

A) Rigid integral foam; B) Rigid integral foam, glass-mat reinforced; C) Hot-moldable rigid foam; D) Semi-flexible integral foam; E) Semi-flexible integral foam, glass-fiber reinforced; F) Semi-flexible filling foam; G) Semi-flexible filling foam, impact-energy-absorbing ; H) Flexible molded foam; I) rigid foam; J) Hot-curing PUR elastomer; K) PUR rubber

RIM materials are used for exterior auto parts like fenders (Fig. 35A), panels, side protection and window encapsulation, but also for other exterior parts in agricultural machines or recreational vehicles like snowmobiles (Fig. 35B).

Figure 35. Use of integral skin foams and RIM materials

A) Fender of the VW Touareg; B) Snowmobile

Rigid integral skin foams or RIM materials can be used in many different applications, mainly for structural parts. They can favorably compete with wood, thermoplastics, thermosets, and metals. They are used for furniture, skis, pump housings, window frames, decorative applications, and housings, just to name a few characteristic applications.

7.4.2. *Production*

Integral skin foams are mostly manufactured in metallic (aluminum) or epoxy molds. Skin formation is controlled by addition of a low-boiling solvent in amounts of about 2–10 wt %, relative to the polyol mixture. After the banning of CFCs and HCFCs, currently HFCs (hydrofluorocarbons) or alkanes are used. When using alkanes, special precautions must be taken because of their flammability.

Thermodynamic equilibrium between the liquid and gaseous phases of the blowing agent in the reaction mixture is crucial for the formation of the closed surface. The mold temperature is kept between 40 and 60 °C, and the reaction mixture in the vicinity of the mold surface has a lower temperature than the core (up to 140 °C). In the core the blowing agent vaporizes, while near the edge it condenses, with formation of the solid surface.

The quality of the surface depends very much on the quality of the mold, but precise control of the processing pressures and temperatures is also required.

In special cases formation of an integral skin is possible without the addition of physical blowing agents. For this, temperature-controlled carbon dioxide formation from the NCO/H_2O reaction must be achieved.

High-pressure metering and mixing units are necessary for manufacturing of RIM parts. The processing of these materials, which are mostly of high reactivity, also requires closed, metallic molds and a clamping unit. To obtain void-free parts without porosity the reaction mixture must enter the mold cavity through a gate designed to ensure laminar flow. Also, correct venting of the mold at the highest points and nucleation of the reaction mixture by air or nitrogen are important for good quality.

After demolding, the parts must be trimmed, and gate and flash must be removed. The surface is prepared for painting by removal of the release agent by washing or sanding.

7.4.3. *Properties*

Depending on the field of use, soft-elastic moldings have densities of 100–300 kg/m^3, flexible molded articles 400–600 kg/m^3, and the relatively hard, microcellular elastomers 900–1100 kg/m^3.

Technical data for a typical soft integral skin foam of low density:

Density (ISO 845)	400 kg/m^3
Shore A hardness (DIN 53505)	35
Tensile strength (DIN 53504)	0.8 MPa
Elongation at break (DIN 53504)	85 %
Tear resistance (ISO 34)	1.4 kN/m
Compression set (25 %/70 h, RT; ISO 1856)	2.5 %

Mechanical properties of a flexible molding of medium density used in the footwear industry (e.g., polyester system):

Density (ISO 845)	500–600 kg/m^3
Shore A hardness (DIN 53505)	ca. 50
Tensile strength (ISO 1798)	6–8 MPa
Elongation at break (ISO 1798)	450–550 %
Tear propagation strength (ISO 34)	10–15 kN/m
Abrasion loss (DIN 53516)	50–100 mg

Mechanical properties of a microcellular elastomer molding of high density (e.g., automotive body parts):

Density (ISO 845)	1100 ± 50 kg/m^3
Shore D hardness (DIN 53505)	64 ± 3
Tensile strength (DIN 53504)	35 ± 3 MPa
Elongation at break (DIN 53504)	300 ± 30 %
Tear propagation strength with cut (ISO 34)	80 ± 10 kN/m
Modulus of elasticity according to Roelig	
at–30 °C	1400 ± 100 MPa
at + 20 °C	600 ± 50 MPa
at + 65 °C	360 ± 30 MPa

The technical data of rigid integral skin foams depend very much on the density, which can be anywhere between 150 and 1100 kg/m^3. Typical properties of a low-density rigid integral skin foam, as used for imitation wood follow:

Density (ISO 845)	400 kg/m^3
Shore D hardness (DIN 53505)	58
Flexural strength (ISO 178)	20 MPa
Flexural modulus of elasticity (ISO 178)	520 MPa
Impact strength (ISO 6603)	10 KJ/m^2
Heat deflection temperature (ISO 75)	90 °C

High-density rigid RIM materials, which are typically used at wall thickness between 3 and 5 mm, have the following characteristics:

Density (ISO 1183)	1100 kg/m^3

Shore D hardness (DIN 53505)	76
Flexural strength (ISO 178)	25–45 MPa
Flexural modulus of elasticity (ISO 178)	2000 MPa
Tensile strength (ISO 527)	53 MPa
Elongation at break (ISO 527)	14 %
Impact strength (ISO 179)	45 KJ/m^2
Heat deflection temperature (ISO 75)	115 °C

When flame retardancy tests are also considered, RIM materials are capable of passing burning tests such as UL 94 V-0 (Underwriters' Laboratories) or B2 (DIN 4102) by the addition of flame retardants.

Both flexible and rigid RIM materials can be reinforced with short or long glass or mineral fibers and also glass mats or mats made of natural fibers. These composite materials show considerably higher stiffness and especially impact strength.

[Top of Page]

8. Noncellular Polyurethanes

Geza Avar and Hans-Georg Wussow

Although the term "noncellular" is unambiguous, the terminology of the products treated in this chapter requires explanation. "Cellular Vulkollan," well known from the early days of PURs, already breached the borderline of precise definition. In practice, this material was not considered a "foam," although it can be undoubtedly classified as a foam from a strictly physical viewpoint because of its cellular structure, however small the few cells may be. The RIM technology has made the definitions even less distinct due to the terminology of "microcellular elastomers" on the one hand and "solid integral parts" on the other. In this chapter the problem of definition is solved as follows: All products which are based on a formulation that usually leads to a foam are classified as "foams," even if cell-free solid materials can be obtained by raising the apparent density (less blowing agent). Similarly, products which are based on formulations that usually result in noncellular materials are classified as noncellular PUR, even though they may be cellular, and these materials are treated in this chapter.

There are two different processing technologies for manufacturing these noncellular PURs. Liquid, low molecular mass raw material systems are converted to high molecular mass products by the reaction of the polyol and polyisocyanate components by cold or hot curing. Linear, high-molecular-mass raw material systems are either thermoplastically processable pellets or urethane rubber are processed by vulcanization technology.

8.1. Cast Elastomers

Cast elastomers are mostly solid, noncellular materials, produced in a low-pressure casting process from ester or ether prepolymers of MDI, TDI, and NDI (see Section Aromatic Polyisocyanates) and a cross-linker such as butanediol. If water-containing cross-linkers are used, cellular elastomers can also be obtained.

8.1.1. *Applications*

The highest quality elastomers are those derived from NDI [Tradenames: Vulkollan (Bayer MaterialScience), Cellasto (Elastogran)]. They are used where materials must meet extremely high requirements regarding durability and mechanical properties. Examples are high-load rolls and wheels for forklift trucks and machines, hydraulic seals, membranes, valve seats and balls, scrapers, and sieves. Cellular NDI-based elastomers are widely used in the automotive industry for bouncers and spring supports (Figs. 34 and 36).

Figure 36. Use of cellular Vulkollan as supplementary spring in the passenger car

Hot-curing elastomers based on MDI or TDI are used, e.g., for bottle stars in automatic filling machines, cutting wheels for fiber production, roll covers for the paper and printing industry, and wheels for in-line skates and skateboards.

Cold-curing PUR elastomers are used as binders for the surface of playgrounds and sport tracks. These surfaces are very durable, require little maintenance, can be used under all weather conditions, and reduce the risk of injuries. Other applications are seals for stoneware pipes, polishing disks, protective edges, and permanent fenders for boats. Cold-cure elastomers can also be sprayed on concrete or metal for corrosion protection and are used for the surfacing of bridges, parking decks, and terraces and for lining of wagons and trucks.

8.1.2. *Production.*

The raw materials used are low molecular mass glycols or amines, polyester or polyether polyols, and (modified) MDI-, TDI-, or NDI-type isocyanates.

The elastomers can be manufactured in a two-step or a one-step procedure. Figure 37 shows a typical line for the production of high-quality elastomers, as is used for NDI-based high-performance materials. The system passes through several stages: dehydration of the polyester (a), production of prepolymer from polyester and diisocyanate (b), degassing of the feed component (c), admixing of cross-linking agent (d), atmospheric-pressure casting in open molds (e, f), and fully curing the molding at 100–300 °C for several hours (g). This processing technique even allows the production of voluminous and thick-walled parts in relatively simple, inexpensive molds with low energy consumption. Depending on the system to be cast, steel, aluminum, and in special cases plastics or wood can be used as the mold material.

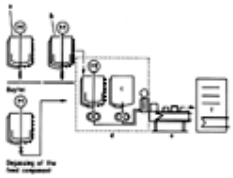

Figure 37. Flow chart for production of high-quality elastomer (e.g., Vulkollan/Baytec)

a) Dehydrogenation of the polyester; b) Prepolymer production; c) Cross-linker; d) Casting machine; e) Mold frame; f) Heating oven (thermal posttreatment)

For MDI and TDI stable prepolymers are commercially available, so that a one-step process can be used. The reactivity of the raw materials varies widely. Casting times between 2 and 60 min can be realized.

Cold-casting systems are usually processed with a casting machine. For some formulations that have a sufficiently long pot life, the starting components can also be hand mixed and cast. In this case the charge volumes are limited to a few liters. The reaction mixtures cure, without an external heat supply, within widely differing cure times, from less than 10 min to a maximum of 24 h. The castings reach their final strength after ca. 7 d storage at room temperature.

8.1.3. *Properties*

NDI-based elastomers have a temperature range for continuous use between –30 and +80 °C. At temperatures below 0 °C the material becomes increasingly hard and inelastic, though there is no danger of fracture. Embrittlement occurs only at very low temperatures (below –30 °C).

NDI elastomer (Vulkollan) is resistant to mineral oils, grease, gasoline, and most organic solvents. Gasoline and some solvents cause some swelling, but this does not impair its mechanical properties if the solvents are able to re-evaporate. Concentrated acids, alkalis, and other hydrolyzing media attack NDI elastomers. By incorporating additives or using special polyesters the hydrolytic stability can be increased so that the service life of the elastomers becomes adequate for a variety of uses even under extreme conditions. Mechanical properties of NDI elastomers are listed in Table 8.

Table 8. Mechanical properties of a hot-casting system (e.g., Vulkollan)*

Property	Vulkollan type			
	18/40	18	30	50
Shore hardness A/D (DIN 53505)	68/20	83/31	92/36	96/58
Density (ISO 1183), kg/m^3	1.25	1.25	1.26	1.26
Abrasion loss** (DIN 53516), mm^3	45 – 50	45 – 50	40 – 50	43 – 48
Tear propagation strength according to Graves (ISO 34), kN/m	20 – 30	40 – 55	67 – 85	85 – 96
Impact resilience (DIN 53512), %	48 – 55	45 – 50	45 – 50	45 – 50
Tensile strength (DIN 53504), MPa	42 – 49	50 – 60	37 – 45	40 – 43
Elongation at break (DIN 53504), %	600 – 700	630 – 750	550 – 650	330 – 370

*The properties listed here can only be regarded as approximate. For Vulkollan articles on the market, the quality characteristics specified by the manufacturer are applicable.
**Relative to abrasion-resistant grade of natural rubber (= 100).

Cellular NDI elastomers can be produced with densities of 300–700 kg/m^3 and are characterized by their high capability of absorbing energy on impact. The material can be compressed by up to 80 % of the original volume without damage. Damping units with various compression deformation characteristics can be manufactured from cellular Vulkollan by varying its apparent density, preferably in the range 300–600 kg/m^3.

Hot-casting systems with a base other than NDI are produced with Shore hardness grades from ca. 55A to 65D. The other mechanical properties and the chemical stability are basically comparable with those of NDI elastomers without reaching the same levels, especially in dynamic stability. They do not, however, reach the high values of Vulkollan in all points, as is also clear from the lower dynamic stability under load. MDI and TDI elastomers cannot replace Vulkollan but are used as supplementary products, primarily in fields where the high quality of Vulkollan is not absolutely necessary, where a system with long casting time is advantageous for producing large-volume and thick-walled moldings, and where short demolding times are desirable for producing large series.

Cold-curing products have a good elasticity, high flexibility, and a variable Shore hardness in the range of ca. 20A to 65D. They are highly resistant to aging and microorganisms and are not attacked by dilute acids and alkalis, mineral oils, soaps, or detergents. Organic solvents cause various degrees of swelling.

8.2. Thermoplastic Polyurethane Elastomers (TPU)

See also ⟶ Rubber, 3. Synthetic – Thermoplastic Polyurethanes, ⟶ Thermoplastic Elastomers – Thermoplastic Polyurethane Elastomers

Production. Diisocyanate (usually MDI), polyols (polyester and polyether polyols), chain extenders (butanediol), and the necessary auxiliaries are reacted in continuous plants via mixing heads or in reaction extruders. After solidifying, the product is processed further to pellets. The TPU can be processed further by, e.g., injection molding, extrusion, calendering, or blow molding to give finished or semi-finished products.

Properties. The principle of "tailor-made plastics" applies to the properties of thermoplastic PURs, as to all PURs. A very wide range of properties is available by careful choice of the starting materials. This is shown in Figures 38 and 39 as well as Table 9 for three selected examples.

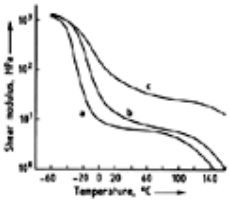

Figure 38. Shear modulus for various TPUs (e.g., Desmopan)

a) Type 385; b) Type 790; c) Type 460

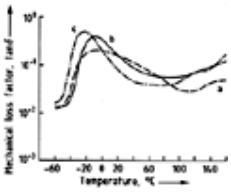

Figure 39. Mechanical loss factor for various TPUs (e.g., Desmopan)

a) Type 385; b) Type 790; c) Type 460

Table 9. Mechanical properties of different types of TPU (e.g., Desmopan)

Property	Desmopan system		
	385	460	790
Shore hardness A/D (ISO 868)	86/33	97/59	92/42
Density (ISO 1183), kg/m^3	1200	1210	1210
Tensile strength (ISO 527-1/-3), MPa	40	35	55
Elongation at break (ISO 527-1/-3), %	450	350	450
Tear propagation resistance (ISO 34-1), kN/m	80	120	85
Impact resilience (ISO 4662), %	42	35	32
Abrasion loss (ISO 4669), mm^3	30	40	30

Uses. The TPUs are valued as high-grade polymeric materials in many industrial branches for a great variety of constructional parts. Table 10 shows examples of their use in relation to their Shore hardness. Glass-fiber reinforced types are used for bodywork units.

Table 10. Typical applications for different grades of TPU

Hardness range		Uses
Shore A	Shore D	
75 – 90	< 35	membranes, seals, damping units, sieve units, hoses, cable sheatings
90 – 95	35 – 45	animal identification markings, roll coverings, toothed belts, sport shoe soles
> 95	45 – 55	seals, flanges, rolls, pneumatic hoses
	55 – 70	bearings, bushes, connectors, ball sockets, seals, coupling units, ski bootlegs, heel pieces

[Top of Page]

9. Polyurethane Coatings [199-201]

Karl-Ludwig Noble

See also ⟶ Paints and Coatings – Polyurethane Coatings

PUR coatings offer a large variety of outstanding properties, which make them suitable for many high-performance applications. PUR coatings have polymeric structures with urethane, urea, biuret, or allophanate coupling groups. They are generated either from reactive systems by polyaddition of oligomeric polyisocyanates with polyol components, or from high molecular mass adducts which are further cross-linked by physical drying or other mechanisms. The nature of the isocyanate building blocks (aliphatic or aromatic) and the structure of the backbone are important. Film mechanical properties are dependent on the structure of the polymer (segmented, interpenetrated networks) and cross-link density. PUR coatings are available solvent-borne and, as a result of the trend to reduced emission systems (low VOC), as high-solids, waterborne, solvent-free, and powder systems. Important PUR systems are summarized below.

Two-component coatings are the most important (solvent-borne) PUR paint systems. The primary component is always polyisocyanate (mainly HDI, IPDI, or TDI derivatives). The other component consists of polyols (polyester, polyacrylic) or amines or a mixture thereof and any additional ingredients such as pigments or solvents. The reactive components must be kept separate. Mixing, preferably of an equimolar equivalent, is done immediately before application. The coatings system can be applied by all conventional coating methods, but preferably by spraying or, industrially, with automatic two-component equipment. The solids contents of the coatings are usually higher than with standard coatings and can reach 80 % (pigmented high-solids), depending on the formulation. Curing takes place at ambient temperature, but can be accelerated by heat. Outstanding film properties such as high mechanical resistance, high chemical resistance, and, with aliphatic polyisocyanates, excellent lightfastness and weather resistance, open up a broad range of uses. Major areas of application are transportation (large motor and rail vehicles, aircraft, automobile finishes and refinishes), the building sector (wood and mineral substrates), industrial paints, and steel construction.

One-component coatings include air-drying and stoving systems. The most important are solvent-borne, isocyanate-terminated prepolymers of higher molecular weight made from aromatic or aliphatic isocyanates (MDI, TDI, HDI, IPDI) with polyols. They react with atmospheric moisture to form urea linking groups. Although the final properties are comparable to those of two-component PURs, curing of these systems is slower, which limits their use mainly to wood-flooring coatings, corrosion protection, and applications on mineral surfaces.

The reaction of TDI or IPDI with polyol-modified drying or semi-drying oils generates urethane alkyds, which can be cured oxidatively (by air drying) like alkyds.

The isocyanate groups of prepolymers can be reacted with agents containing, e.g., acidic CH or NH groups to form blocked isocyanates (see Section Blocked Isocyanates). Their mixtures with polyols are stable at room temperature, but react at elevated temperatures (120–220 °C, depending on the nature of the blocking agent). The resulting films have a very high mechanical resistance. They are particularly important for automotive finishes, industrial goods and coil coating.

Powder coatings consist of a combination of solid polyols (mostly polyester or polyacrylate polyols) and solid blocked polyisocyanates as the cross-linker. ε-Caprolactam was the most commonly used blocking agent but is being replaced by other blocking agents. An alternative method, avoiding the use of an added blocking agent is the dimerization of the isocyanate groups to uretdiones. During stoving (10–25 min at 150–210 °C) isocyanate groups are liberated and undergo cross-linking. PUR powder systems are used in many industrial applications.

Waterborne coatings are growing continuously in importance, due to the requirement that emissions should be reduced. A variety of systems are available, comparable to solvent-borne coatings. One-component systems are based on self-emulsifying polyurethane or polyurethane-polyurea dispersions. They are manufactured by processes which allow the production, in contrast to solvent-borne PURs, of any molecular weight without having a significant impact on viscosity, and with only small contents of solvents or no solvent at all. These products are used in physically dried films for more flexible substrates like textile, paper, and plastics. One-component polyurethane/polyacrylate dispersions based on aliphatic or aromatic isocyanates are used for coating parquet flooring. In addition, UV-curing PUR acrylate dispersions are used in industrial wood and furniture finishes.

One-component stoving systems contain blocked polyisocyanates emulsified in OH- or NH-containing dispersions (epoxy polyester, acrylic). Important applications are automotive coatings (electrodeposition coating, primer surfacer) and other industrial coatings.

Reactive waterborne two-component systems are based on polyisocyanates (HDI, IPDI) and aqueous OH-functional dispersions (polyacrylate, polyester, polyester polyurethane). Self-emulsifying polyisocyanates reduce interfacial tension between the components and ease homogenization. The film properties are influenced by the extent of side reactions, which can occur in the presence of water. The main applications of these reactive systems are in all types of industrial, wood, and plastic coatings.

[Top of Page]

10. Polyurethane Adhesives [202-204]

Herbert Blankenheim

See also ⟶ Adhesives

PUR adhesives exhibit unique properties due to their wide variety of compositions, their excellent adhesion to most polar substrates (ability to form hydrogen bonds), and the possibility of chemical cross-linking.

To achieve good cohesive strength of the adhesive bond, high molecular weight polymers are necessary. On the other hand, for good adhesion to the substrate, efficient wetting of the surface is essential. This can be done either by "liquefying" the polymer by the use of solvents, aqueous dispersions, or heat or by in situ polyaddition of low-viscosity polyols and polyisocyanates to give high molecular weight products. Depending on the application, the adhesives can be differentiated as follows:

Solvent-borne adhesives are made by dissolving granules of a linear high molecular weight OH-terminated PUR in a solvent like methyl ethyl ketone, acetone, ethyl acetate, or mixtures thereof to a solids content of up to 15–20 %. The PUR building blocks are crystalline polyester diols or polycaprolactones, aromatic diisocyanates (mainly MDI or TDI), and possibly a low molecular weight chain extender like butanediol or hexanediol. After evaporation of the solvent a nonblocking adhesive film results on the substrate. Due to the crystalline nature of the polyester soft segment, this adhesive film can be thermally reactivated at temperatures between 50 and 80 °C, and in this tacky state bond formation to a second coated and activated substrate can be achieved. On cooling down again, recrystallization leads to high initial and final bond strength ("physical setting"). To improve properties such as heat resistance, resistance to solvents, water, oils, or plastisizers, two-component processing is possible. This is achieved by adding a solvent-based triisocyanate (e.g., Desmodur R-Series, Bayer MaterialScience, cf. Section Aromatic Polyisocyanates) to the adhesive solution shortly before application, which results in chemical cross-linking by reaction with the OH groups of the base polymer and other isocyanate-reactive groups on the substrate (including water). The main application for this kind of adhesive is in the footwear industry (bonding shoe soles to uppers).

Waterborne adhesives usually consist of a 40–50 wt. % aqueous dispersion of a high molecular weight PUR made by the acetone or the prepolymer mixing process. Besides crystalline polyester diols, aliphatic isocyanates such as HDI or IPDI and ionic or hydrophilic groups containing chain extenders are used as starting materials. Properties and processing are comparable to those of solvent-borne adhesives. Cross-linking can be achieved by using solvent-free emulsifiable polyisocyanates (e.g.- Desmodur D series, Bayer MaterialScience). The main applications are in sport shoe fabrication, the furniture industry (MDF kitchen cabinet door lamination with PVC films), and the automotive industry (lamination of interior trim parts with films).

Hot-melt adhesives are available in nonreactive and reactive form. The starting materials for nonreactive systems are similar to those used for solvent-borne adhesives, but instead of dissolving the granules, these are either directly extruded onto the substrate or (co)extruded to form hot-melt films or fleeces. Another possibility for bringing the adhesive polymer into intimate contact with the substrate (without using solvents or water as a carrier) is cold grinding of the granules. The resulting powders (particle size < 600 µm) can be easily brought onto the surface by scatter coating. Common to all these nonreactive hot melt application forms is the (in this case only one-component) processing by the thermo-activation procedure described above. Application fields are textile and film lamination.

Reactive hot melts are prepolymers of a crystalline or high-T_g polyester with an excess of a diisocyanate (mostly MDI-based). They are applied at temperatures of 60–80 °C higher than the melting point of the soft segment and combine the aforementioned advantages of physical setting and chemical cross-linking (NCO/water reaction). Typical application areas are bookbinding, woodworking, packaging, and the automotive and construction industries.

Reactive adhesives can be divided into one- and two-component systems. One-component reactive adhesives liquid at room temperature are NCO-terminated prepolymers of (usually) polyether polyols and an excess of an aromatic polyisocyanate (mainly MDI derivatives). Setting occurs by reaction with humidity and therefore requires at least one porous substrate. Applications are foam rebonding, rubber crumb bonding, and in the construction and transportation fields.

Two-component reactive adhesives consist of a liquid polyether polyol (or a low-viscosity polyester polyol) and a polyisocyanate (mainly polymeric MDI). The two components are either mixed and applied immediately before use by an automatic machine or by using a handheld double cartridge/static mixer combination. A broad spectrum of polyisocyanates and polyols is available, and the mixing ratio of the two components can be varied over a wide range. Thus, it is possible to design the adhesive for the individual substrate/bond requirements, e.g., from flexible to rigid. Application fields are again transport and construction, and flexible packaging/film lamination (no porous substrate necessary).

[Top of Page]

11. Polyurethane Fibers [205]

Wolfgang Friederichs

See also ⟶ Fibers, 4. Synthetic Organic – Polyurethane Fibers

Synthetic fibers containing at least 85 wt % of high molecular weight segmented PUR [206] are known as Elastane or in the USA as Spandex. The segmented structure (see Section Polyurethanes with Segmented Structure) gives these fibers high elasticity and makes them easy to stretch. The properties are determined not only by the raw materials used, but also by the synthesis conditions, the spinning process, and the post-treatment.

Compared to rubber fibers, the PUR fibers exhibit superior properties such as the ability to form fine filaments, good colorability and chemical resistance, and high resistance to physical impact.

They are produced by a two step process. In the first step, long-chain diols are reacted with aromatic diisocyanates; in the second step, the high-melting urethane and/or urea groups which are necessary for the high-level properties are introduced by chain extension with short-chain diols or diamines (see Section Polyurethanes with Segmented Structure) [207-211].

In industrial production, usually poly(tetraethylene glycol) [212, 213], manufactured by ring-opening polymerization of tetrahydrofuran, is used as the long-chain polyol. In some cases, copolyethers of tetrahydrofuran and 3-methyltetrahydrofuran are used [214]. If polyesters are used as diols, they are usually based on adipic acid and mixtures of ethylene glycol, propylene gylcol, butylene glycol, and/or 1,6-hexandiol.

Aromatic diisocyanates, mainly 4,4′-MDI and in some cases 2,4-TDI, are used exclusively. Due to its unsymmetrical structure, 2,4-TDI leads to a weaker segmented structure and therefore to lower thermostability. Aliphatic diisocyanates, e.g., HDI, are only used in small amounts as modifiers.

Chain extension with diamines usually gives high-performance Elastane [215-218]. Mostly primary cycloaliphatic or aliphatic diamines such as ethylenediamine and 1,2-propylenediamine are used. Modification of the hard segments is achieved by adding small amounts of, e.g., 1,3-diaminocyclohexane.

[Top of Page]

12. Polyurethanes and Isocyanates as Binders [219]

Donald-Richard Larimer

Panels and molded parts made of wood chips, fibers, and veneers are used extensively in the furniture, building, packaging, and automotive industries. These products are produced under heat and pressure by using a variety of organic binders. PURs are excellent starting materials for composite binders and are used extensively in their production. They are classified according to binder type as polyurethane and polyurea bonding agents.

Polyurethane bonding can be accomplished with one-component moisture-curing systems or two-component ambient-cure systems. The latter can be adjusted in reactivity to meet processing conditions.

PUR dispersions are also utilized in special cases. In addition to adhesion, PUR dispersions exhibit excellent film-forming properties. This is an especially important characteristic for the sizing of glass fibers. PUR dispersions meet very specific requirements regarding film formation and glass adhesion, critical for “matrix” bonding. The properties of glass-reinforced plastic depend to a high degree on the adhesion of the polymer to the surface of the glass fibers.

In polyurea bonding, polyisocyanates are transformed into polyureas at high temperatures in the presence of moisture and, if necessary, catalysts. In addition to the physical adhesion properties of polyureas, there is also the possibility of adhesion by way of chemical reactions between polyisocyanate and reactive groups of the substrate material. Compared to the PUR binders, polyureas are more rigid and hydrophobic.

Binders based on polymeric diphenylmethane diisocyanate (PMDI) have been industrially used in the wood composite industry since 1973. PMDI binders are used worldwide in the manufacture of particleboard, oriented strand board (OSB), laminated strand lumber (LSL), medium-density fiberboard (MDF), and other specially engineered wood composites, as well as panel products based on annual plants, e.g., straw and sugar cane bagasse. The manufacture of straw-based panels was indeed first made possible with PMDI binders.

The success of PMDI binders in the wood and straw composites industry has its origin in the unique properties of the polyurea bond.

Moisture resistance: polyurea is highly cross-linked, chemically stable, and hydrolytically resistant. PMDI binders are one of two binders listed in the European standard for particleboard EN 312-3/5, option 2 that require no special testing to prove hydrolytic resistance.

Dimensional stability: polyurea is also hydrophobic and lends wooden panels improved dimensional stability. This was demonstrated on particleboard during three years of outdoor exposure, where they showed the lowest moisture absorbance and weight changes. This favorable behavior was also confirmed by long-term stress bending (creep testing) under severe conditions of 40 °C and 95 % R.H.

Emission-free: PMDI wood binders contain no formaldehyde. The resultant polyurea bond in the finished panel is free from emissions.

High-strength properties: The excellent adhesion properties of polyurea and the deep anchoring of the bond in the wood structure leads to the high strengths of the different wood composite panels.

During the manufacture and processing of rigid PUR foams, cutting and milling about 5–25 % of scrap is generated. Utilization of this scrap is possible by bonding it into panels or molded parts with varying properties. This can technically be realized with the aid of the hot-press technique employing PMDI binders, analogous to the manufacture of wood composites. The physical and mechanical properties of articles made from rigid PUR foam scrap can be varied over a wide range. The properties depend upon the type and pretreatment of the raw material, the amount of bonding agent, and the density.

Cork granulates and remilled rubber waste, especially peeled tire tread from the recapping process, as well as recovered scrap from milling of frozen old tires, can be converted with PUR bonding agents to composite materials. The bonding

agents are isocyanate-terminated, one-component, moisture-curing prepolymers or reactive two-component systems. They form elastic films during curing. The prepolymers are synthesized from long-chain linear or slightly branched polyethers and MDI types or toluene diisocyanate and/or from mixtures of both. The one-component prepolymers generally have an NCO content of ca. 10 % and a viscosity of 1500 to 3000 mPa · s at 23 °C.

The most important properties are elasticity, flexibility, damping capacity and void content. They can be varied over a broad range through the shape and size of the rubber particles, type and amount of binder, and degree of compaction.

[Top of Page]

13. Special Products

Wolfgang Friederichs

PURs are particularly suitable as wall materials for microcapsules (see also ⟶ Microencapsulation). Capsules can be impermeable or permeable to the core material. The capsule wall is usually produced by interfacial reaction of multifunctional isocyanates or NCO prepolymers and multifunctional amines at an oil/water interface. The microcapsules have a diameter of 1 to 5000 µm, depending on the droplet size of the emulsion used for polymer formation. The capsules are used for a wide variety of purposes, e.g., for the production of carbonless copying paper or for the delayed release of plant-protection products.

Polyoxazolidinone polyisocyanurates are produced from polyisocyanates and epoxy resins in the presence of special catalysts (see Chap. Basic Reactions). They have high fire resistance, high long-term heat resistance, and excellent electrical insulation properties [220, 221].

Water-soluble PURs are of interest as retaining agents and dyeing auxiliaries, thickeners, paper-sizing agents, and other sizers.

Oligourethanes with unsaturated (e.g., acrylate) groups can be cured by UV radiation in the presence of photoinitiators [222]. Systems based on acrylate groups containing blocked NCO prepolymers have been developed. They can be cured either by heat or by radiation (dual curing [223-227]) and are used especially for automotive refinishing paint.

[Top of Page]

14. Safety and Ecology

Manfred Giersig, Eckehard Weigand, Michael Halfmann, Friedrich-Wilhelm Wittbecker

PUR-containing products have become an essential part of modern living. For production and processing of all of these varieties of PUR, it is estimated that worldwide well over 500 000 people are handling PUR raw materials every day. For the toxicology of the raw materials, see also ⟶ Isocyanates, Organic – Toxicology and Occupational Health and ⟶ Polyoxyalkylenes – Toxicity .

Health, safety, and environmental aspects, especially for MDI and TDI (see Section Aliphatic Polyisocyanates), are described in [228].

14.1. Safety Precautions when Handling the Raw Materials

When building and operating plants for producing PUR, all laws, decrees, and regulations in force for chemical plants must be observed. Many years' experience in handling the raw and auxiliary materials for PUR has shown that these products can be handled safely. It is a prerequisite for this, however, that the persons working with them are supplied with all the necessary information and above all, that the required technical and personal protective equipment is available and used.

All isocyanates, activators, and most alkanes and haloalkanes are regarded as products having acute or chronic harmful properties.

Di- and polyisocyanates have only a moderately acute oral and cutaneous toxicity. Di- and polyisocyanates are known to have allergenic potential, and the human health risk due to the acute or chronic effect of isocyanate vapors and aerosols (of liquid and dust forms) is considerably higher. Therefore, it is of extreme importance that appropriate safety measures, e.g., protective clothing and ventilation, are provided.

Occupational exposure limits have been set in many countries. As examples, the currently (2003) MAK values for Germany are listed in Table 11.

Table 11. German air concentration limits (according to TRGS 900) for typical polyisocyanates (see Section Polyisocyanates)

Isocyanate	Air concentration limit	
	ppm	mg/m^3
TDI (2,4-/2,6-)	0.005	0.035
MDI	0.005	0.050
NDI	0.010	0.087

HDI	0.005	0.035
IPDI	0.010	0.092
H12-MDI	0.005	0.054

The vapor pressures of the industrially most important isocyanates at their usual processing temperatures are listed in Table 12.

Table 12. Vapor pressures of the industrially most important isocyanates at processing temperature

Isocyanate	Processing temp., °C	Vapor pressure, Pa	Conc. of saturated vapor, mg/m^3
TDI (2,4-/2,6-)	25	3	255.5
PMDI	25	$< 10^{-2}$	< 0.1
MDI	42	6×10^{-2}	0.65
NDI	130	ca. 120	7850

Many years' experience in handling the various isocyanates has shown that very effective air-extraction systems must be present at all workplaces where TDI is processed at room temperature or MDI and NDI at temperatures above 40 °C. If, for any reason, the aforementioned maximum value at the workplace is exceeded (e.g., on construction sites), the persons present in the working area must be protected against inhaling isocyanates, wherever possible by wearing suitable masks. TRGS 430 is a special German guideline on how to analyze and monitor isocyanates at workplaces where PURs are produced or processed.

More information can be obtained from literature available from the isocyanate producers, or their associations e.g.,

ISOPA: http://www.isopa.org
API: http://www.polyurethane.org/

Polyols have been well studied with respect to their toxicity and, if used according to relevant regulations, do not present an industrial hygiene problem. Aminopolyethers are moderately irritating to the eyes and skin, and thus contact with them should be avoided.

Auxiliary Materials. The ready-to-use polyol preparations, which are used to a large extent nowadays, contain the auxiliary materials necessary for PUR production (especially activators, emulsifiers, and stabilizers) in quantities of up to ca. 1.5 wt %. Toxicological studies have shown that these small quantities of the auxiliary materials scarcely affect the biological behavior of the polyols.

Polyols and polyol preparations have an exceptionally low vapor pressure; a health hazard via inhalation is therefore not to be expected. If, however, they contain readily volatile aliphatic amines as activators, an unpleasant odor can be expected.

For the *tertiary amines* used as catalysts, occupational threshold values in air (MAK) defined by the German TRGS 900 (2003) are available only in some cases, e.g., triethylamine: 1 mL/m^3 (ppm) corresponding to an eight-hour average value of 4 mg/m^3. However, the intense odor of most tertiary amines gives a clear warning of excessive concentrations.

Low-boiling chlorofluoroalkanes were used as blowing agents for producing rigid and flexible foams in the past. Since these substances are suspected of damaging the ozone layer, their use was phased out in the late 1980s. These substances have therefore been replaced by alkanes and to some extent by partly halogenated alkanes, because the blowing process cannot be carried out with carbon dioxide (from the NCO/H_2O reaction) in all cases. Since direct contact of partly halogenated alkanes and alkanes with the skin leads to intense defatting and roughness, it is important to wear impermeable gloves and to take good care of the skin. The inhalation of vapors from these blowing agents in high concentrations can lead to intoxication.

14.2. Emissions, Accidental Release, and Waste Disposal [229]

The emission of PUR raw materials into the atmosphere during production and processing is usually small. In special cases (e.g., filling of storage tanks, flexible foam slabstock production), measures to reduce emissions are advisable or required. Since a certain concentration limit of a harmful substance must not be exceeded in the air at the workplace, the air is extracted at the processing plant, and this is one source of emissions. The volatility of the raw materials determines the extent of emission. Hence blowing agents (e.g., carbon dioxide, halogenated hydrocarbons, hydrocarbons) and tertiary amines are the main constituents, whereas isocyanates normally do not play a significant role. Since the carbon dioxide formed from the reaction of water with isocyanate is a natural constituent of air, it presents no problem. The other blowing agents are discharged into the atmosphere in varying quantities depending on the foam type. The emissions of volatile organic chemicals are being increasingly regulated in industrialized countries. National regulations must be observed.

During storage, tank-filling, and transportation of PUR raw materials, discharge due to accidents or leaks cannot be entirely ruled out. Should this happen, as a standard precaution, the accidentally released raw material should be prevented from entering soil or water by suitable measures (e.g., covering sewers). Spilled product should be covered with

a liquid-binding material such as a chemical binder based on hydrated calcium silicate (e.g., Hybilat), sand, or sawdust. This can then be transferred to a waste container. In the case of spilled isocyanates, the liquid-binding material should be moist and the material in the waste container stored under moist conditions in a secure place in the open air. The lid of the container must not be tightly sealed because of danger of bursting (CO_2 evolution). The polluted areas must then be decontaminated (see below). In all cases, the waste collected in the containers is best disposed of by incineration in a suitable facility (see below). In the unlikely event that a large pool of escaped isocyanate has formed, all unauthorized persons should be kept away from the scene of the accident. Persons downwind should be evacuated. Further evaporation of isocyanate can be prevented very effectively by covering the pool with protein foam. Large quantities of liquids can be pumped into waste containers, before the procedure given above is applied to the remaining residues. Nevertheless, in the case that isocyanate raw materials have leaked into the soil or water, inert polyurea is the major end product of their reaction with water. According to a three-month soil study with ^{14}C-labeled TDI and MDI, degradation products (e.g., amines) were at no time detected in the aqueous extract.

An experimental pond study indicated low environmental risk arising from spillage of polymeric MDI into a natural aquatic ecosystem. No direct ecotoxic effects and no bioaccumulation of MDI or MDA were observed. Furthermore no MDI or MDA could be detected in water.

Waste Disposal. Liquid or solid PUR production wastes which are not completely reacted are categorized as hazardous waste and should not be dumped but burned in industrial waste incineration facilities, equipped with state-of-the-art flue-gas scrubbing.

Drum decontamination. Any isocyanates residue in “empty” containers that are to be sent for reconditioning or disposal can be rendered harmless with a special neutralizing solution (for formulations, see the specialist literature). The same solution can also be used to decontaminate clothing, equipment and floors. Fully cured PUR end-product wastes, such as trimming waste from flexible foam slabs, can be recycled or recovered (see below). Wastes for which no utilization is yet possible can have their energy content recovered in appropriate incineration plants or can be disposed of in landfill sites, as a last resort (and as long as local regulations do not exclude organic materials from landfilling).

14.3. Recycling/Recovery of Polyurethanes

Because of their excellent durability, PURs in general contribute significantly to a long service life of the products containing them. This is an important contribution to waste minimization. To handle PUR wastes that cannot be prevented, technologies for recycling or recovery have been developed by many industrial associations, joint ventures, and cooperations between producers and users of raw materials throughout the world, e.g., the Polyurethane Recycle and Recovery Council (PURRC, USA), European Isocyanate and Polyol Producers Association (ISOPA, Europe), Tecpol (Germany), and the Japan Urethane Industrial Institute (JUII, Japan). The most appropriate method may vary from case to case, and depends on the properties of the PUR, the intended application, and the related capacity of the market for the recycled material, and very much on logistical, economic, and ecological factors. After all, recycling should not only serve to minimize waste; it ought to contribute also to the saving of resources and to the reduction of environmental burdens. Only then can recycling be counted as a contribution to sustainable development. Currently, the following possibilities for recycling/recovery of PUR exist.

Mechanical Recycling. Mechanical (or physical) recycling of PURs means one of the many forms of “particle recycling.” The rebonding of flexible foam or adhesive pressing of rigid foam, for example, involves binding roughly 90 wt % of particles with about 10 wt % PUR binders under heat and pressure. Particle bonding uses up to 70 wt % PUR as the matrix for any kind of particles (e.g., rubber chips for sports ground surfaces). PUR powder can be incorporated into new PUR articles at a loading of round about 20 wt %. PUR powders/particles are also being reused as oil binders.

Feedstock Recycling/Chemical Recycling. Large-scale feedstock recycling processes recover oil and gas products from mixed plastic waste streams of a hundred thousand tons per year or more, of which PUR materials can be one constituent. These processes include pyrolysis, hydrogenation, synthesis gas generation, and reduction of iron ore in blast furnaces. When smaller, but pure streams of particular polymers are available, “chemical recycling” processes can be applied. It is thus possible to obtain liquid degradation products from PURs that are suitable, together with new material, for the manufacture of new PURs. For example glycolysis has been applied more often in favorable case where a suitable application for the glycolysate has been identified. Much research has resulted in several process variations [230]. Some of them include purification and chemical processing of the regenerate before use in PUR applications. Observation of appropriate industrial hygiene and safety is essential, because hazardous substances could be formed.

Energy Recovery. Combustion, or incineration with the recovery of energy, is currently the most effective way to reduce the volume of organic material which otherwise would have to be sent to landfill. Combustion is suitable for all materials for which material recycling is ruled out on ecological or economic grounds or simply because of logistical difficulties. Rotary kilns, fluidized beds, and mass burning equipment, for example, have been shown to be suitable for combustion of plastic scrap. These processes are applied, e.g., in cement production, in industrial power stations, and in municipal solid waste combustors. The heat content of PURs (lower heating value 24–30 MJ/kg) is comparable to that of coal. Numerous test runs have shown that this energy content is recovered in an environmentally sound manner in modern plants equipped with state-of-the-art flue-gas treatment facilities.

14.4. Fire Performance of Polyurethanes [128], [231]

The fire performance of a product is characterized at the various stages of a fire by the parameters ignitability, flame spread, and heat release, and by phenomena such as smoke density and the toxic potency of the combustion products. The dripping behavior, the amount of debris in general, and the corrosive effect of effluents are also determined. Fire test methods differ greatly from country to country, and also depend on the field of application. Different requirements exist, e.g., for the electrical, transportation, furniture, and building sectors. Fire testing, classification and regulatory requirements are very complex and reference must be made to summarizing literature.

In general, the fire hazards posed by a PUR foam and its fire effluents are comparable to those of natural materials like wool and wood. The chemical composition of the material is not the main determining factor. The behavior of the material in fires depends on, e.g., the relative surface of the material (e.g., massive wood compared to sawdust), the combination of materials, thermal conductivity, nature, duration, and intensity of ignition sources, and ventilation.

The qualitative and quantitative composition of fire effluents depends on the decomposition conditions (temperature and ventilation) as well as the amount of material involved. Comparative tests taking into account different stages of a fire have confirmed that the toxicity of fire effluents from PUR products is similar to that of other natural products like wood, cork, and wool. Fire effluents of natural and artificial organic products always pose a toxic hazard irrespective of the type of burning material.

As effluents from PUR do not differ greatly from natural materials, a similar effect on the environment is assumed. Due to their reactivity, isocyanates are rapidly decomposed photochemically or converted chemically.

A special fire risk from PUR is not evident: risk of ignition and flame spread as well as emissions mainly depend on the scenario and the amount of burning material, and are not determined by the specific material involved.

[Top of Page]

15. Economic Aspects

Geza Avar

Even though PUR makes up only for ca. 6 % of the total consumption of plastics, the annual production reached the 10 × 10^6 t level in 2004. The growth rate since 1970 (Fig. 40) on average exceeded 5 % and thus grew faster than GDP. The fastest growing market over the last few years was Asia, in particular China with growth rates exceeding 10 %.

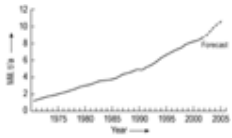

Figure 40. World PUR consumption

Figures from 2001 show that there is still a big difference in per capita consumption between the regions (Fig. 41), which allows an optimistic outlook for further development.

Figure 41. PUR per capita consumption 2001 by regions

Figure 42 gives an overview of the main application areas of PURs and Figure 43 the split for the basic raw materials MDI, TDI, and polyethers.

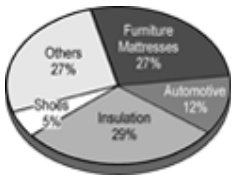

Figure 42. World PUR consumption 2004 by applications (total: 10 × 10^6 t/a)

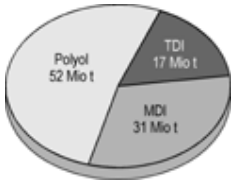

Figure 43. World PUR consumption 2004 by raw materials (total: 10 × 10^6 Mio t/a)

Table 13 lists trade names of raw materials.

Table 13. Trade names of raw materials (exemplary)

Type	Tradename	Company
Isocyanates	Caradate	Shell

	Cosmonate	Mitsui Takeda
	Desmodur	Bayer MaterialScience
	Isonate	DOW
	Lupranate	BASF
	Mondur	Bayer MaterialScience
	Rubinate	Huntsman
	Suprasec	Huntsman
	Takenate	Mitsui Takeda
	Tedimon	DOW
	Voranate	DOW
Polyols	Acclaim	Bayer MaterialScience
	Actcol	Mitsui Takeda
	Arcol	Bayer MaterialScience
	Caradol	Shell
	Daltolac	Huntsman
	Desmophen	Bayer MaterialScience
	Isonol	DOW
	Jeffol	Huntsman
	Lupranol	BASF
	Luprapren	BASF
	Multranol	Bayer MaterialScience
	Pluracol	BASF
	Rubinol	Huntsman
	Tercarol	DOW
	Voranol	DOW
Catalysts	DABCO	Air Products
	Desmorapid	Bayer MaterialScience
	Jeffcat	Huntsman
	Niax	Witco
	Polycat	Air Products
	Toyocat	Tosoh

[Top of Page]

References

General References

1. O. Bayer, *Angew. Chem.* **59** (1947) 257 – 272.
2. I. G. Farbenindustrie AG, DE 728981, 1937 (O. Bayer).
3. G. Oertel: “Polyurethane,” in G. W. Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., **vol. 7**, Carl Hanser-Verlag, München 1992.
4. G. Oertel: *Polyurethane Handbook*, 2nd ed., Carl Hanser Verlag, München, 1993.
5. J. H. Saunders, K. C. Frisch: *Polyurethanes*, Interscience, New York 1962, 1964. “Chemistry,” part I; “Technology,” part II.
6. K. Uhlig: *Discovering Polyurethanes*, 1st ed., Carl Hanser Verlag, München, 1999.
7. R. Leppkes: “Polyurethanes” in *Die Bibliothek der Technik*, 1st ed., **vol. 91**, moderne industrie Verlag, Landsberg, 2002.
8. *Houben-Weyl*, 4th ed., **E20**, 1561 – 1757.
9. K. N Edwards: “Urethane Chemistry and Applications,” *ACS Symp. Ser.* **172** (1981).
10. C. Hepburn: *Polyurethane Elastomers*, 2nd ed., Elsevier Applied Science Publishers, London, 1991.
11. G. Woods: *The ICI Polyurethane Book*, 2nd ed., ICI Polyurethanes and John Wiley and Sons, Chichester, New York, 1990.
12. W. Meckel, W. Goyert, W. Wieder: “Thermoplastic PUR Elastomers” in G. Holden, N. R Legge, R. Quirk, H. Schroeder (eds.): *Thermoplastic Elastomers*, 2nd ed., Hanser-Verlag, München 1996, pp. 15 – 44.
13. K. C Frisch, S. L Reegen: *Advances in Urethane Science and Technology*, **vols. 1–6**, Technomic Publ., Westport, CT, 1971 – 1978. K. C Frisch, D. Klempner: *Advances in Urethane Science and Technology*, **vols. 8–10**, Technomic Publ., Lancaster, PA, 1981–1987.
14. J. M Methven: “Polymeric Materials from Renewable Resources,” *Rapra Rev. Rep.* **4** (1991) no. 6, Report 43, 1 – 121.
15. H. Zoebelein, *Chem unserer Zeit* **26** (1992) 27 – 34. Links
16. P. L Nayak, S. Lenka, S. K Panda, T. Pattnaik, *J. Appl. Polym. Sci.* **47** (1993) 1089 – 96.

17. I. Javni, Z. Petrovic, *Annu. Tech. Conf. Soc. Plast. Eng.* 1997, 791 – 795.
18. Bayer AG, EP 672694, 1995 (W. von Bonin, H.-P. Müller, M. Kapps).
19. A. Gandini, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **39** (1998) 69 – 70.
20. P. Ni, R. W Thring, *Int. J. Polym. Mat.* **52** (2003) 685 – 707.
21. S. Desai, I. M Thakore, B. D Sarawade, S. Devi, *Polym. Eng. Sci.* **40** (2000) 1200 – 1210.
22. Agency of Ind. Science & Technol., JP 5186556, 1993 (H. Hatakeyama *et al.*).
23. Hüls AG, EP 787754, 1996 (R. Gras, E. Wolf).
24. BASF AG, EP 716080, 1994 (W. Heider, W. Langer, H. Renz, S. Wolff).
25. Bayer AG, EP 802210, 1997 (B. Baumbach, L. Kahl, E. König, N. Yuva).
26. Bayer AG, EP 270804, 1987 (H. Heß, K. König, R. Kopp, G. Grögler).
27. R. Kopp, H. Heß, G. Grögler, Proc. SPI/FSK Polyurethane World Congress 1987, *Aachen, Technomic Publ., Lancaster*, pp. 316 – 322.
28. American Cyanamid Company, EP 234353, 1982 (B. Singh, L. W Chang, P. S Forgione).
29. V. D Arendt, R. E Logan, R. Saxon, *J. Cell. Plast.* **18** (1982) 376. Links
30. Bayer AG, EP 798299, 1999 (H. Mertes, F. Richter, J. Pedain, C.-G. Dieris).
31. VEBA Chemie AG, DE 2323299, 1973 (K. Schmitt, J. Disteldorf, J. Reiffer).
32. L. Born, D. Wendisch, H. Reiff, D. Dieterich, *Angew. Makromol. Chem.* **171** (1989) 213.
33. N. Bialas, H. Höcker, M. Marschner, W. Ritter, *Makromol. Chem.* **191** (1990) 1843.
34. Hüls AG, US 4912210, 1987 (J. Disteldorf, W. Hübel, K. Schmitz).
35. Mobay Corp., EP 453914, 1990 (S. D. Seneker, T. A Potter, K. L Dunlap, M. K Lowery).
36. Z. W Wicks, *Prog. Org. Coat.* **3** (1975) 73. Links
37. Z. W Wicks, *Prog. Org. Coat.* **9** (1981) 3.
38. J. W Rosthauser, J. L Williams, *Polym. Mater. Sci. Eng.* **50** (1984) 344.
39. V. Mirgel, K. Nachtkamp, *Farbe Lack* **89** (1983) 928.
40. B. W Peterson, T. H Austin, *Proc. SPI Annu. Polyurethane Tech./Mark. Conf.*, 1986, Toronto, Technomic Publ., Lancaster, pp. 155 – 158.
41. Th. Goldschmidt AG, US 5001189, 1986 (J. Fock, D. Schedlitzki).
42. Th. Goldschmidt AG, EP 0326890, 1989 (J. Fock, E. Esselborn, F. Nickel).
43. BASF AG, DE 4316137, 1994 (W. Franzischka, R. Becker, C. Palm, C. Sigwart).
44. W. J Pentz, R. G Krawiec, *Rubber Age (New York)* **12** (1975) 39.
45. J. R Harrison, *Plast. Ind. News* **30** (1984) 120.
46. J. L Schuchardt, S. D Harper, *Proc. 32nd Annual Polyurethane Technical Marketing Conf.*, 1989, San Francisko, pp. 360 – 364.
47. G. Wegener, M. Brandt, L. Duda, J. Hofmann, B. Klesczewski, D. Koch, R.-J. Kumpf, H. Orzesek, H.-G. Pirkl, C. Six, C. Steinlein, M. Weisbeck, *Appl. Catal. A: Gen.* **221** (2001), 303 – 335.
48. ARCO Chemicals, EP 700 949, 1994 (B. Le-Khac).
49. E. Müller, *Angew. Makromol. Chem.* **14** (1970) 75 – 86.
50. E. Müller, *Angew. Makromol. Chem.* **16/17** (1971) 117 – 128..
51. S. Inoue, H. Koinuma und T. Tsuruta, *J. Polym. Sci. Polym. Lett.* **7** (1969) 287.
52. S. Inoue, H. Koinuma und T. Tsuruta, *Makromol. Chem.* **130** (1969) 210.
53. M. Ree, J. Y. Bae, J. H. Jung, T. J. Shin, J. Korea, *Polym. J.* **7** (1999) 333.
54. W. Kuran, S. Pasynkiewicz, J. Skupinska, A. Rokicki, *Makromol. Chem.* **177** (1976) 11.
55. T. Aida und S. Inoue, *J. Am. Chem. Soc.* **107** (1985) 1358.
56. Bayer AG, EP 68383, 1981 (J. Probst, G. Kolb, B. Riberi, P. Höhlein).
57. Bayer AG, DE 2622951, 1976 (Ch. Weber, H. Schäfer).
58. Dow Chemical, US 4477644, 1983 (T. R. Sutton, D. W Hughes).
59. BASF Wyandotte Corp., EP 107108, 1983 (J. W Lightsey, R. A Markovs, G. G Ramlow, P. T Kan).
60. Air Products, US 4816543, 1985 (W. F Burgoyne, D. D Dixon, B. Milligan, J. P Casey).
61. C. J Nalepa, A. A Eisenbraun, *Proc. S. P. I. Annu. Polyurethane Tech./Mark. Conf.*, 1986, Toronto, Technomic Publ., Lancaster, pp. 228 – 233.
62. Bayer, DE 1803635, 1968 (W. Meckel, E. Müller).
63. H. Kleimann, *Proc. SPI/FSK Polyurethane World Congress* 1987, Technomic Publ., Aachen, Lancaster, pp. 815 – 819.
64. Th. Voelker, P. Balling, *Proc. SPI Annu. Polyurethane Tech./Mark. Conf.*, 1986, Toronto, Technomic Publ., Lancaster, pp. 133 – 136.
65. U. S. Department of Health and Human Services. Hazardous Substances Data Bank, National Toxicology Information Program, National Library of Medicine, Bethesda, MD. 1993.
66. Jefferson Chem., US 3847992, 1973 (P. H Moss).
67. BP Chem. Int. Ltd, EP 356046, 1988 (I. D Dobson).
68. J.-M. Jehng, C.-M. Chen, *Catal. Lett.* **77** (2001) 147 – 154. Links
69. Union Carbide, EP 297536, 1988 (R. M Gerkin, D. J Schreck, D. E Smith).

70. *Houben-Weyl*, 4th ed., **E20**, 1602, 1610.
71. Bayer AG, EP 219035, 1985 (W. Raßhofer, K. König, H.-J. Meiners, G. Grögler).
72. Bayer AG, DE 1935484, 1969 (G. Oertel, G. Grögler).
73. Mobay Corp., EP 457129, 1990 (R. Mafoti, J. Sanders).
74. Dow Chemical, EP 268849, 1986 (V. R Durvasula, F. A Stuber).
75. Dow Chemical, EP 335274, 1988 (V. R Durvasula).
76. Bayer, DE 3834749, 1988 (U. Thiery, H. Reiff, D. Dieterich, G. Grögler, J. Sanders).
77. M. Bock, R. Halpaap, *Farbe Lack* **93** (1987) 264.
78. M. Bock, R. Halpaap, *Polym. Mat. Sci. Eng.* **55** (1986) 448.
79. M. Bock, R. Halpaap, *J. Coat. Technol.* **59** (1987) 131.
80. E. F. Cassidy, *Urethanes Technology* (1989), pp. 26 – 34.
81. ICI, EP 361704, 1988 (E. F Cassidy, H. R Gillis, M. Hannaby).
82. D. G Schlotterbeck, G. Matzke, P. Horn, H.-U. Schmidt, *J. Elastomers Plast.* **21** (1989) 49.
83. D. G Schlotterbeck, G. Matzke, P. Horn, H.-U. Schmidt, *Elastomerics* **121** (1989) 18.
84. *Houben-Weyl*, 4th ed., **E20**, 1603.
85. P. K. H. Lam, M. H George, J. A Barrie, *Polym. Commun.* **32** (1991) 80.
86. The University of Akron, PCT WO 9010657, 1989 (J. P Kennedy).
87. J. P Kennedy, B. Ivan, V. S. C Chang, *Adv. Urethane Sci. Technol.* **8** (1981) 245.
88. Minnesota Mining & Manufacturing, EP 380236, 1989 (C. Leir).
89. Dow Corning, EP 378079, 1989 (G. T. Decker, C.-L. Lee, G. A Gornowicz).
90. Union Carbide, EP 405494, 1989 (R. S Neale, C. L Schilling).
91. Du Pont, EP 420223, 1989 (E. G Brugel).
92. DE 3702495, 1987 (G. Bauer).
93. B.-U. Kettemann, M. Melchiorre, T. Muenzmay, W. Rasshofer, *Kunststoffe* **85** (1995) 1947 – 1950.
94. M. Kunh, M. Kugler, *Kunststoffe* **87** (1997) 729 – 730, 732.
95. M. L. Listemann, A. L Wressell, K. R Lassila, H. C Klotz, G. L Johnson, A. C Savoca, *Proc. Polyurethanes World Congress*, 1993, Vancouver, Technomic Publ., Lancaster, pp. 595 – 608.
96. A. L Silva, J. C Bordado, *Catal. Rev.* **46** (2004) 31 – 51.
97. K. C Frisch, L. P Rumao, *J. Macromol. Sci., Rev. Macromol. Chem. Phys.* **C5** (1970) 103 – 150.
98. G. Oertel: *Polyurethane Handbook*, 2nd ed., Carl Hanser Verlag, München, 1993, pp. 98 – 104.
99. D. C Fondots, *J. Cell. Plast.* **11** (1975) 250. Links
100. M. L Listemann, A. C Savoca, A. L Wressell, *J. Cell. Plast.* **28** (1992) 360 – 398.
101. J. G Kniss, R. Fard-Aghaie, A. R Arnold, S. Wendel, *Proc. API Polyurethanes Expo*, 2003, Orlando, pp. 361 – 366.
102. R. Hoffmann, H. Schloens, *Proc. API Polyurethane Conf.*, 2002, Salt Lake City, pp. 380 – 386.
103. L. Thiele, *Acta Polym.* **30** (1979) 323. Links
104. G. D Andrew, J. G Kniss, M. L Listemann, L. A Mercando, J. D Tobias, S. Wendel: “Dimensional Stabilizing Additives for Flexible Polyurethane Foams” in D. Klempner, K. C Frisch (eds.): *Advances in Urethane Science and Technology*, Rapra Technology Limited, 2001, pp. 3 – 83.
105. R. M Hill: *Silicone Surfactants*, Marcel Dekker, New York, 1999.
106. G. Rossmy, *J. Cell Plast.* **17** (1981) 319. Links
107. Th. Goldschmidt AG, EP 193815, 1985 (H.-J. Kollmeier, R.-D. Langenhagen).
108. B. Kanner, B. Prokai, C. S Eschbach, G. I Murphy, *J. Cell. Plast.* **15** (1979) 315. Links
109. O. Eyrisch, G. Burkhart, R. Borgogelli, *Proc. API Polyurethanes Expo*, 2001, Columbus, pp. 417 – 422.
110. T. Boinowitz, R. Borgogelli, W. Gower, *Proc. API Polyurethanes Expo*, 2003, Orlando, pp. 324 – 330.
111. O. Eyrisch, G. Burkhart, *Proc. API Polyurethanes Conf.*, 2002, Salt Lake City, pp 409 – 414.
112. T. Boinowitz, G. Burkhart, J. Klietsch, W. Bunting, *Proc. API Polyurethanes Expo*, 2001, Columbus, pp. 41 – 46.
113. T. Timm, *Kautsch. Gummi Kunstst.* **35** (1982) 568.
114. T. Timm, *Kautsch. Gummi Kunstst.* **36** (1983) 257.
115. T. Timm, *Kautsch. Gummi Kunstst.* **37** (1984) 933, 1021.
116. C. S Schollenberger, F. D Stewart, *Adv. Urethane Sci. Technol.* **1** (1971) 65.
117. G. Capocci, *Proc. SPI. Annu. Polyurethane Tech./Mark. Conf.*, 1986, Toronto, Technomic Publ., Lancaster, pp. 220 – 227.
118. B. P Thaspliyal *et al.*, *Progr. Polym. Sci.* **15** (1990) 735 – 750.
119. G. Mathur, J. E Kresta, K. C Frisch, *Adv. Urethane Sci. Technol.* **6** (1978) 103.
120. R. Wiedermann, G. Heilig, *Kunststoffe* **80** (1990) 909.
121. G. Heilig, *Kunststoffe* **81** (1991) 622.
122. U. Leyrer, D. Polke, *Kunststoffe* **88** (1998) 235.
123. H. Fleurent, *Cell. Polym.* **13** (1994) 419. Links
124. R. D Leaversuch, *Mod. Plast. Int.* **24** (1994) 19.
125. H. Krähling, L. Zipfel, *Proc. API Polyurethanes Expo*, 2001, Columbus, pp. 333 – 338.

126. L. Zipfel, K. Boerner, P. Dournel, *Proc. API Polyurethanes Expo*, 2001, Columbus, pp. 637 – 642.
127. D. Williams, *Proc. API Polyurethanes Conf.*, 2002, Salt Lake City, pp 135 – 143.
128. J. Troitzsch, *Plastics Flammability Handbook*, 3rd ed., Carl Hanser Verlag, München 2004, pp. 78 – 83, 165 – 167.
129. G. Neumann, H. U Schimpfle, R. Becker, *Plaste Kautsch.* **24** (1977) 27.
130. G. Oertel: "Polyurethane" in G. W Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., **vol. 7**, Carl Hanser Verlag, München 1993, pp. 40 – 47.
131. W. Goyert, H. Hespe, *Kunststoffe* **68** (1978) 819.
132. C. S Schollenberger, K. Dinbergs, *Adv. Urethane Sci. Technol.* **6** (1978) 60.
133. K. Hoffmann, R. Bonart, *Makromol. Chem.* **184** (1983) 1529.
134. I. Blackwell, K. H Gardner, *Polymer* **20** (1979) 13. Links
135. E. C Prolingheuer, J. J Lindsey, H. Kleimann, *J. Elastomers Plast.* **21** (1989) 100.
136. M. Tanaka, *Makromol. Chem.* **187** (1986) 2345.
137. M. Tanaka, T. Nakaya, *Macromol. Sci. Chem.* **A24** (1987) 777.
138. L. Born, H. Hespe, *Colloid Polym. Sci.* **263** (1985) 335. Links
139. L. Morbitzer, H. Hespe, *J. Appl. Polym. Sci.* **16** (1972) 2697.
140. H. F Hespe *et al.*, *J. Appl. Polym. Sci.* **44** (1992) 2029.
141. G. Zeitler, *Kunststoffberater* **34** (1989) 65.
142. G. Zeitler, *K-Plast. Kautsch.-Ztg.* **403** (1989) 85.
143. U. Rotermund, *J. Elastomers Plast.* **21** (1989) 122.
144. D. Boese *et al.*, *Makromol. Chemie, Macromol. Symp.* **50** (1991) 191.
145. *Houben-Weyl*, 4th ed., **E20**, pp. 1739 – 1751.
146. S. M Clift, J. Grimminger, K. Muha, *Proc. SPI Polyurethanes Conf.*, 1994, Boston, pp. 546 – 560.
147. S. B Burns, E. L Schmidt, *Proc. Polyurethane World Congress* 1993, Vancouver, pp. 234 – 240.
148. Bayer AG, DE 3914718, 1989 (A. Ruckes, H.-J. Meiners, H. Boden, M. Schmidt).
149. J. Blahak, W. Meckel, E. Müller, *Angew. Makromol. Chem.* **26** (1972) 29.
150. Bayer AG, DE 2513817, 1975 (H. Schäfer, C. Weber).
151. Bayer AG, DE 2622951, 1976 (H. Schäfer, C. Weber).
152. D. Nissen, R. A Markovs, *J. Elastomers Plast.* **15** (1983) 96.
153. J. Blackwell, J. R Quay, R. B Turner, *Polym. Eng. Sci.* **23** (1983) 816. Links
154. Bayer AG, DE 3147736, 1981 (K. Seel, C. Weber, H. Wirtz).
155. Texaco Inc., EP 92672, 1982 (R. J. G Dominguez).
156. Texaco Inc., EP 93862, 1982 (R. J. G Dominguez, D. M Rice, R. F Lloyd).
157. Texaco Inc., EP 128691, 1983 (R. J. G Dominguez).
158. Texaco Inc., US 4474901, 1983 (R. J. G Dominguez).
159. Texaco Inc., US 4420570, 1982 (R. J. G Dominguez).
160. R. J. G Dominguez, *J. Cell. Plast.* **20** (1985) 433.
161. Dow, US 4495081, 1983 (J. A Vanderhider, G. M Lancaster).
162. BASF, DE 3215907, 1982 (D. Nissen, W. Schoenleben, M. Marx, K. H Illers, P. Simak).
163. BASF, DE 3215909, 1982 (D. Nissen, W. Schoenleben, M. Marx, K. H Illers, P. Simak).
164. Air Products & Chemicals, US 4892920, 1988 (J. R Quay, J. P Casey).
165. Air Products & Chemicals, EP 420026, 1989 (R. M Machado, A. J Siuta,'W. E Starner, B. A Toseland).
166. *Houben-Weyl*, 4th ed., **E20**, 1613 – 1617.
167. L. Schmalstieg, L. Kahl, *Congr. FATIPEC*, 20th Nice, 1990, pp. 425 – 428.
168. Bayer AG, EP 394759, 1989 (L. Schmalstieg, K. Nachtkamp, J. Pedain).
169. C. Zwiener, L. Schmalstieg, M. Sonntag, K. Nachtkamp, J. Pedain, *farbe+lack* (1991) 1052.
170. H. Kittel, *Adhäsion* **21** (1977) 162.
171. Bayer AG, DE 3940271, 1989 (G. Grögler, H. Heß, R. Kopp, W. Raßhofer).
172. *Houben-Weyl*, 4th ed., **E20**, pp. 1703 – 1716, 1729 – 1735.
173. P. W Morgan, *J. Appl. Polym. Sci.* **40** (1990) 1771.
174. Bayer AG, DE 3808274, 1988 (H.-P. Müller, B. Jansen, W. Calaminus, R. Dhein).
175. L. Ubaghs, N. Fricke, H. Keul, H. Höcker, *Macromol. Rapid Commun.* **25** (2004) 517.
176. Eurotech ltd, US 6120905, 2000 (O. L Figovsky).
177. S. Maier, T. Loontjens, B. Scholtens, R. Mülhaupt, *Macromolecules* **36** (2003) 4727. Links
178. R. F Harris, C. D DePorter, R. B Potter, *Macromolecules* **24** (1991) 2973. Links
179. M. Bierdel: *Numerische und experimentelle Untersuchungen zum Gegenstrom-Injektions-Mischer*, Logos Verlag, Berlin 2001.
180. Maschinenfabrik Hennecke GmbH, DE 2123216, 1971 (W. Schmitz, R. Raffel, F. Althausen, F. Proksa).
181. Hollandsche Draad en Kabelfabrik (Draka), DE 1207072, 1958 (B. Hackert).
182. Hellenic Plastics & Rubber Ind., DE 1629458, 1965 (N. G. Petzetakis).
183. Unifoam AG, DE 2142450, 1970 (L. Berg).

184. Beamech Variable Pressure Foaming, Tenax Road, Trafford Park, Manchaster, M17 1JT, England.
185. NovaFlex Hennecke, Hennecke GmbH, Birlinghofener Str. 30, D-53754 Sankt-Augustin, Germany.
186. Reeves Brothers Inc., US 3325573, 1964 (J. D Boon, K. F Hager).
187. F. Buff, M. French, W. Pollock, US 3476845, 1965 (F. Buff, M. French, W. Pollock).
188. Fecken Kirfel GmbH & Co, Prager Ring 1–15, D-52070 Aachen, Germany.
189. U. Leyrer, Conf. Papers UTECH 2003, The Hague.
190. H.-U. Schmidt, E. Calgua, *Kunststoffe* **89** (1999) 104.
191. Air Products & Chemicals, US 5958990, 1980, (J. Grimminger).
192. W. W Reichmann, B. A Phillips, *J. Cell. Plast.* **24** (1988) 601. Links
193. J. Wu, D. Dillon, R. Crooker, *Proc. API Polyurethane Conf.*, 2002, Salt Lake City, pp. 144 – 150.
194. C. Chittolini, *Proc. Polyurethanes World Congress*, 1997, Amsterdam, pp. 655 – 659.
195. R. Franco, M. Checchin, C. Cecchini, *Proceedings of UTECH* 2000, The Hague, pp. 10 – 12.
196. V. Knipp: *Herstellung von Großteilen aus Polyurethanschaumstoffen, Zechner & Höthig*, Speyer 1974.
197. W. Schuhmacher, *Gummi Asbest Kunstst.* 1974, 698.
198. H. Piechota, H. Röhr: *Integralschaumstoffe*, Hanser Verlag, München 1975.
199. G. Oertel: “Polyurethane,” in G. W. Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., **vol. 7**, Carl Hanser Verlag, München 1993, pp. 599 – 642.
200. M. Bock: *Polyurethanes for Coatings*, Vincents Verlag, Hannover 2001.
201. D. Stoye, W. Freitag: *Resins For Coatings*, Hanser Gardner, Chicago 1996.
202. G. Oertel: “Polyurethane,” in G. W Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., **vol. 7**, Carl Hanser Verlag, München 1993, pp. 643 – 663.
203. G. Festel, A. Proß, H. Stepanski, H. Blankenheim, R. Witkowski, *Adhäsion Kleben Dichten* 1997, 16.
204. S. Lee: *The Polyurethanes Book*, Jon Wiley & Sons, New York 2002, pp. 379 – 394.
205. G. Oertel: “Polyurethane,” in G. W Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., vol. **7**, Carl Hanser Verlag, München 1993, pp. 679 – 694.
206. DIN 60001 (1970).
207. D. C Allport, A. A Mohajer in D. C Allport, W. H Janes: *Block Copolymers*, Section 5.2.6, Applied Publishers Ltd., London, 1973.
208. H. Oertel in B. v. Falkai: *Synthesefasern-Grundlagen, Technologie, Verarbeitung und Anwendung*, Section 6.1.3, Verlag Chemie, Weinheim, Germany, 1981.
209. M. Couper, in M. Levin, J. Preston: *Handbook of Fibre Science and Technology Fibers*, **Vol. 3**, part A, Marcel Dekker, New York 1985, pp 51 – 85.
210. K. H Wolf, *Text.-Prax. Int.* **36** (1981) 839.
211. J. H Saunders, K. C Frisch: *Polyurethanes*, Interscience, New York 1962, 1964. “Chemistry,” part I; “Technology,” part II, p 129.
212. DuPont, US 2726219, 1951 (F. B Hill).
213. E. M Hicks Jr., *Am. Dyest. Reptr.* **52** (1963) 33.
214. DuPont, EP 343985, 1988 (N. E Houser, R. L Dreibelbis).
215. R. Bonart, L. Morbitzer, H. Rinke, *Kolloid. Z. Z. Polym.* **240** (1970) 807.
216. A. Blejenberg, D. Heikens, H. Meijers, H. Lampe, P. v. Reth, *Brit. Polym. J.* **4** (1972) 125.
217. H. Oertel, *Chemiefasern Textilind.* **27/79** (1977) 1090.
218. H. Oertel, *Chemiefasern Textilind.* **28/80** (1978) 43.
219. G. Oertel: “Polyurethane,” in G. W Becker, D. Braun (eds.): *Kunststoff-Handbuch*, 3rd ed., vol. **7**, Carl Hanser Verlag, München 1993, pp. 665 – 678.
220. J. Franke, *Kunststoffe* **79** (1989) 999.
221. J. Franke, H.-P. Müller, *Kunststoffe* **80** (1990) 1200.
222. W. Y Chiang, S. C Chan, *Angew. Makromol. Chem.* **182** (1990) 9.
223. W. R. Grace & Co., EP 549116, 1991 (D. R Kyle).
224. W. Fischer, J. Weikard, E. Luehmann, T. Faecke, *FAPU* **14** (2002) 24.
225. H.-U. Meier-Westhues, European Coatings Conf., 2001, Zürich, Vincentz Verlag, Hannover, pp.149 – 62.
226. K. Studer, C. Decker, E. Beck, R. Schwalm, European Coatings Conf. 2004, Berlin, Vincentz Verlag, Hannover, pp. 49 – 63.
227. L. W Arndt, L. J Junker, S. P Patel, D. B Pourreau, W. Wang, *Paint Coatings Ind.* **20** (2004) 42.
228. D. C Allport, D. S Gilbert, S. M. Outterside: *MDI & TDI—Safety, Health and the Environment*, John Wiley & Sons Ltd, Chichester 2003.
229. F. K Brochhagen: “Isocyanates” in O. Hutzinger (ed.): *The Handbook of Environmental Chemistry*, **vol. 3**, part G, Springer-Verlag, Berlin 1991, p. 73.
230. W. Raßhofer: *Recycling von Polyurethan-Kunststoffen*, Hüthig Verlag, Heidelberg, 1998.
231. F. H Prager: *Polyurethane and Fire*, Wiley VCH, Weinheim, Germany, 2004.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Poly(Vinyl Chloride)

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Michael W. Allsopp, Independent PVC Technology Consultant, Heswall, Wirral, England (Chaps. 1 – 3, 5 – 10, Section 4.1, 4.2, 4.4)

Giovanni Vianello, European Vinyls Corportaion (IT), Porto Maguero, Venice, Italy (Section 4.3)

1. Introduction 2
1.1. Morphology 2
1.2. Versatility 2
1.3. Molecular Structure 2
2. Physical and Chemical Properties 3
2.1. Vinyl Chloride Monomer (VCM) 3
2.2. Poly(Vinyl Chloride) (PVC) 3
3. Resources and Raw Materials . . 3
4. Production 4
4.1. Suspension Polymerization of Vinyl Chloride 4
4.1.1. Introduction 4
4.1.2. Additives 6
4.1.3. Kinetics 8
4.1.4. Morphology 9
4.1.4.1. Control of Grain Size 9
4.1.4.2. Control of Porosity and Bulk Density 10
4.1.4.3. K-Value (Molecular Mass) 11
4.1.4.4. Problem Areas 13
4.2. Bulk or Mass Polymerization of Vinyl Chloride 13
4.3. Emulsion Polymerization of Vinyl Chloride 15
4.3.1. Introduction 15
4.3.2. Emulsion Polymerization Processes 15
4.3.3. Description of the Batch Process . . 16
4.3.4. Plastisols 18
4.3.5. Polymer – Plasticizer Interaction . . 19
4.3.6. Gelation and Fusion 21
4.3.7. Applications 21
4.4. Chlorinated Poly(Vinyl Chloride), CPVC 22
5. Environmental Protection 22
6. Quality Specifications and Analysis 23
7. Storage and Transportation 23
8. Processing and End Uses 23
9. Economic Aspects 25
10. Toxicology and Occupational Health 25
11. References 28

1. Introduction

Poly(vinyl chloride), PVC, a polymer prepared from vinyl chloride monomer (VCM),

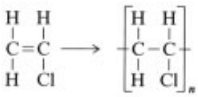

where $n = 700 - 1500$, holds a unique position amongst all the polymers produced today. It is relatively inexpensive and is used in such a wide range of applications that its versatility is almost unlimited. However, if it were discovered today it would probably be shelved as a somewhat intractable and thermally unstable material. How can this apparent contradiction be explained?

The uniqueness of PVC can be considered under three headings: morphology, versatility, and molecular structure.

1.1. Morphology

As made, PVC is particulate in nature and comes in two main sizes depending on the process used. Suspension and mass polymerizations give grains (particles) of 100 – 180 µm in diameter, whereas the emulsion process affords a latex of particle size 0.1 – 3.0 µm. The latter is dried to yield friable grain-like structures of 5 –50 µm.

Because of this unique particulate structure the most frequently used word in the vocabulary

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_717

of the PVC technologist is morphology. In no other polymer is it as important as it is in PVC.

In its polymerisation, a growing PVC chain becomes insoluble in VCM above a chain length of about 10 units [6] so PVC is essentially insoluble in its monomer and the process is thus classified as a precipitation suspension polymerization. However, PVC is heavily swollen and partly solvated by the monomer to the extent of 27 wt % [7] and this has a major influence on the polymerization itself (see Section 4.1) as well as the final properties and end uses of PVC. Hence, the way the PVC separates from the monomer, its future growth mechanism, and the swelling of the polymer by the monomer are critically important in its formation, handling, and subsequent processing.

1.2. Versatility

Poly(vinyl chloride) is a generic name. Each producer makes a range of PVC polymers which vary in morphology and in molecular mass, depending on the intended end use. In industry, the K-value and viscosity number are used to represent molecular mass, and producers often reflect these parameters in the grade codes used to define different products (e.g., S 68/173 refers to a suspension type material with a K-value of 68, and VY 110/57 to a resin with a viscosity number of 110). The calculation of the K-value is given in [8] and the relationship between these parameters and molecular mass in [9].

PVC with K = 66 – 68 can be processed in rigid formulations to give pipes, conduit, sheet, and window profiles; K = 65 – 71 in flexible formulations for flexible sheet, flooring, wallpapers, cable coverings, hoses, tubing, and medical products, and PVC with low K-values (55 – 60) in formulations for injection molding of pipe and conduit fittings, integral electrical plugs, and blow molding of bottles and other containers.

In contrast, poly(methyl methacrylate) of even very high molecular mass is still soluble in its monomer and produces only glasslike solid beads at the end of polymerization which have no internal morphology and as such are limited to a narrow range of end uses (e.g., moldings for optical applications such as covers for car lights and illuminated signs.

1.3. Molecular Structure

Amongst the range of polymeric materials produced today PVC is unique because the bulky chlorine atom imparts a strongly polar nature to the PVC polymer chain, and the essentially syndiotactic conformation of the repeat unit in the chain leads to a limited level of crystallinity [10]. This results in good mechanical properties, particularly stiffness at low wall thickness, high melt viscosity at relatively low molecular mass, and the ability to maintain good mechanical properties even when highly plasticized. This enables a wide range of softness and flexibility to be achieved and hence leads to an even wider variety of end uses.

2. Physical and Chemical Properties

2.1. Vinyl Chloride Monomer (VCM)

Vinyl chloride monomer (VCM), *bp* − 13.4 °C, is a gas at room temperature and pressure. Therefore, it is handled as a compressed volatile liquid in all polymerization operations. Its vapor pressure over the typical polymerization temperature range of 50° to 70 °C is 800 – 1250 kPa. As a result PVC polymerization reactors are thick-walled jacketed steel vessels with a pressure rating of 1725 kPa. VCM is slightly soluble in water (0.11 wt % at 20 °C). Whilst this has some influence on the suspension polymerization process it is critically important to the success of the emulsion polymerization process described in Section 4.3. The polymerization of VCM is strongly exothermic, and its specific heat and heat of evaporation of 1.352 kJ kg^{-1} K^{-1} and 20.6 kJ/mol, respectively, allow the use of a condenser to remove the heat of reaction as well as by the more conventional jacketed vessel systems. Its explosive limits in air are 4 – 22 vol % and plant design, particularly when handling unreacted VCM in the recovery system, must be designed and operated accordingly.

Further physical and chemical properties of VCM are given in → Chlorinated Hydrocarbons, Chap. 3.1.1.

2.2. Poly(Vinyl Chloride) (PVC)

PVC is never used alone. It is always mixed with heat stabilizers, lubricants, plasticizers, fillers, and other additives to make processing possible, all of which can influence its physical and mechanical properties. Table 1 lists properties of rigid (unplasticized) PVC with a total additives content of <10 %. Table 2 list properties of flexible (plasticized) PVC where the range of physical properties varies widely, depending on the plasticizer content. This can vary between 20 and 100 phr so as a guide properties typical of a plasticizer content of 50 phr are given.

Table 1. Typical properties of rigid PVC (UPVC)

Property	Test	Value
Tensile strength at 23 °C, MPa	BS 2782 : 301G ISO R527	55
Tensile modulus (1 % strain, 100 s), GPa	BS 4618 ISO R899	2.7 – 3.0
Tensile modulus (1 % strain, 3 years), GPa	ISO R899	1.7
Izod impact, ft lb/in	BS 2782 : 306A ISO R180	2 (unmodified) 10 (modified)
Specific gravity	ISO R1183	1.38 – 1.45
Coefficient of linear thermal expansion, K^{-1}	BS 4618 : 3.1	6×10^{-5}
Coefficient of thermal conductivity, W/mk		0.14
Flammability (oxygen index)	ASTM D2863 (Fenimore Martin)	45
Weathering resistance		very good (especially white)
Resistance to concentrated mineral acids (at 20 °C)		excellent
Maximum continuous operating temperature	field experience	60 °C

Table 2. Typical properties of flexible PVC

Property	Test	Value
Specific gravity	ISO R1183	1.19 – 1.68
Tensile strength, MPa	BS 2782.320A	7.5 – 30
Elongation at break, %	BS 2782.320A	140 – 400
BS softness	BS 2782 365A	5 – 100
Cold flex temperature, °C	BS 2782 159B	− 20 to − 60
Volume resistivity at 23 °C, Ω · cm	BS 2782 202A	$10^{10} - 10^{15}$
Ageing resistance	field experience	excellent
Ozone resistance	field experience	very good

In addition the K-value (molecular mass) of the PVC can also influence properties significantly (see Section 4.1.4.3).

PVC has extremely good chemical resistance to all but low molecular mass chlorinated solvents. Therefore, it is widely used in the construction and lining of chemical plants.

3. Resources and Raw Materials

VCM is produced industrially by two main reactions:

1) Hydrochlorination of acetylene
2) Thermal cracking of 1,2-dichloroethane produced by direct chlorination or oxychlorination of ethylene in the balanced process. Presently, more than 90 % of the VCM produced is based on this route, full details of which are given in →Chlorinated Hydrocarbons , Chap. 3.1.

In an ideal situation a PVC plant is fully integrated, beginning with ethylene and chlorine (salt), but various levels of integration are employed worldwide. Stand-alone PVC plants, supplied with VCM by sea, road, or rail, are well known as are ones making the bulk of their VCM requirements from scratch on site, with supplementary supplies of 1,2-dichloroethane for cracking to satisfy peak demand. Finally, there are fully integrated plants which have the benefit of uninterrupted supplies of base raw materials (ethylene and salt) and where the monomer is supplied by pipeline at a significant cost benefit. It can cost up to 65 Euro/t to transport VCM within continental Western Europe depending on distance and type of freight used, i.e., road, rail or ship.

4. Production

There are three main processes used for the commercial production of PVC: suspension (providing 80 % of world production), emulsion (12 %) and mass, also called bulk (8 %).

4.1. Suspension Polymerization of Vinyl Chloride

4.1.1. Introduction

The layout of a typical suspension PVC plant is shown in Figure 1.

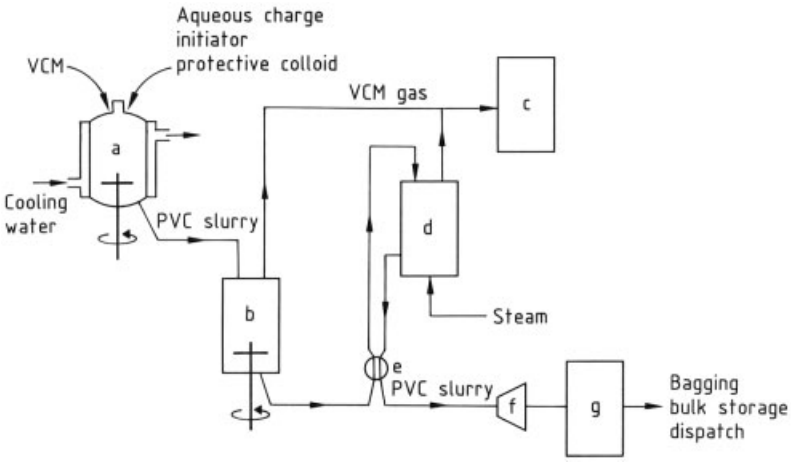

Figure 1. Suspension PVC plant
a) Reactor; b) Blowdown vessel; c) VCM recovery plant; d) Stripping column; e) Heat exchanger; f) Centrifuge; g) Driers

The suspension polymerization process is essentially a bulk polymerization process carried out in millions of small "reactors" (droplets). Liquid vinyl chloride under its autogeneous vapor pressure is dispersed in water by vigorous stirring in a reactor (autoclave) of 25 – 150 m^3 capacity, fitted with a jacket and/or condenser for heat removal and baffles for optimum agitation. This results in the formation of droplets of average diameter 30 – 40 µm which are stabilized against coalescence by one or more protective colloids (granulating agent). The other essential ingredient is a monomer-soluble free radical initiator.

The formulations charged to the reactors (otherwise known as autoclaves, kettles, or polymerizers) are normally referred to as recipes. A basic recipe for suspension processes can be simply water, VCM, initiator, and suspension agent, for example:

VCM	100 parts
Water	90 – 130 parts
Protective colloid	0.05 – 0.15 parts
Initiator	0.03 – 0.08 parts

These quantities vary depending on the PVC grade, reactor size, plant type, etc.

Whilst this will produce a particulate PVC it is unlikely to have the optimum morphology. This will only be achieved if other additives are employed, such as oxygen, buffers, secondary or tertiary granulating agents, chain-transfer or chain-extending agents, comonomers, antioxidants, together with the right level of agitation and homogenization and the correct charging procedure and timing for each additive addition. A combination of just the right number and degree of the above parameters to achieve the optimum morphology usually determines the quality of a PVC grade.

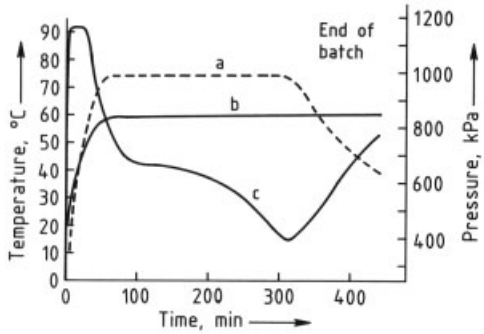

Figure 2. Reaction profile for a typical suspension polymerization
a) Pressure; b) Batch temperature; c) Jacket temperature

After charging, the reactor contents are heated to the reaction temperature of 45 – 75 °C. Reactors are 80 – 95 % full at this stage. Heat causes some of the initiator to decompose into free radicals, and the monomer in the droplets begins to polymerize. The strongly exothermic reaction (1540 kJ/kg) is controlled by removing heat via the jacket and/or by boil-off into a con-

denser from which the condensed monomer is returned to the reactor.

Progress of the reaction in a jacketed reactor can be followed by continuously monitoring the water temperature in the jacket since the flow rate is constant and a constant batch temperature is maintained by progressively reducing the cooling water temperature (see Fig. 2).

Although PVC is insoluble in its monomer (see Section 1.1) the polymer is swollen by ca. 27 wt % VCM to form a coherent gel [7]. Therefore, as the conversion increases towards 70 % and beyond, the pressure in the reactor starts to fall; slowly at first, but then much more rapidly as the last of the free liquid monomer is consumed. Polymerization continues in the gel phase, very rapidly at first since chain termination is hindered by lack of mobility of the growing chain, but, as conversion increases beyond 80 – 85 %, monomer starvation rapidly reduces the rate.

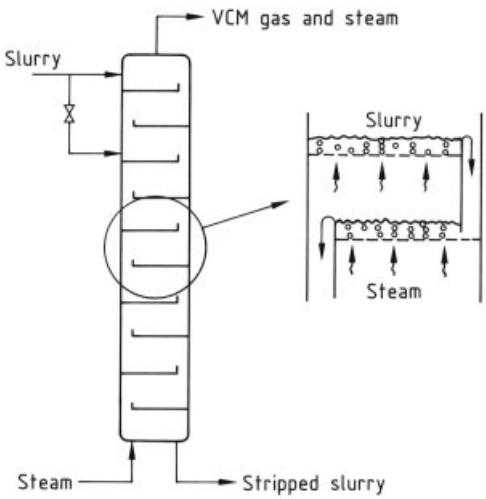

Figure 3. Stripping column

The reaction is terminated at a predetermined pressure by either adding a chain terminator and/or venting off the unreacted monomer to the recovery plant. After venting, the resin in the aqueous slurry can still contain 2 – 3 % unreacted monomer, which is removed by stripping, either batchwise (as in older plants) or continuously in a column. The batch is discharged from the reactor to a feed vessel and then fed continuously through a stripper column (Fig. 3). The unreacted monomer is recovered, liquified, and reused in later polymerizations. After passing through a heat exchanger the slurry is fed to a continuous centrifuge to give a wet-cake with 20 – 30 % moisture content. The remaining water is then removed by conventional flash and/or fluid-bed drying to give a dry, free-flowing powder with a residual VCM content below 1 ppm.

4.1.2. Additives

Water. The water used in the process should be demineralized (i.e., of low conductivity) since ionic species, especially sodium ions, can affect the performance of other additives such as protective colloids and influence final polymer properties such as volume resistivity.

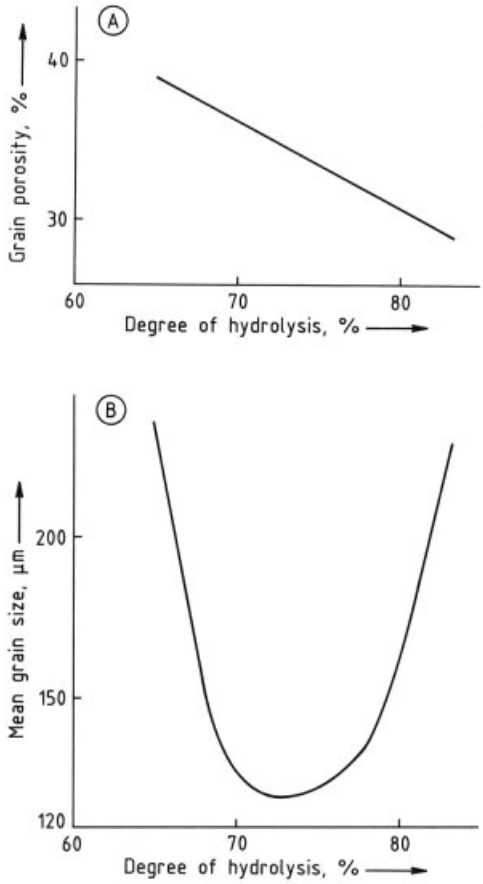

Figure 4. The effect of the degree of hydrolysis of poly(vinyl acetate) on PVC grain properties
A) Grain porosity; B) Mean grain size

Protective Colloids (Granulating Agents). The protective colloids used in the suspension polymerization process are of two types: primary and secondary. The main function of the primary granulating agents is to control grain size, but they also affect porosity and other morphological properties. The demand for polymers with reduced residual VCM levels and better plasticizer adsorption led to the need for higher and more uniform porosity within each grain and this is achieved by using a secondary stabilizer in addition to the primary.

Although a wide variety of surface active materials have been or are still used, the substances most commonly used today are cellulose ether derivatives and partially hydrolyzed poly(vinyl acetates) or poly(vinyl alcohols) (PVAs). The latter are usually preferred and are used alone or in combination with the cellulose type, which they are gradually replacing. Typical PVA protective colloids are listed in Table 3.

As PVA primary granulating agents are prepared by alkaline hydrolysis of poly(vinyl acetates) they are complex mixtures containing fractions with different degrees of hydrolysis and molecular mass and having a block distribution of the OH groups [11]. Because of this complexity suspension processes need fine-tuning around the chosen agent since decreasing the average degree of hydrolysis of the PVA increases the porosity of the PVC grains and affects the grain size, as shown in Figure 4. This behavior is used in conjunction with agitation to achieve the optimum size and porosity. Below 70 % hydrolysis the PVAs are insoluble in water and lose their ability to stabilize the VCM droplets adequately so the normal range of PVA granulating agents used are in the range 70 to 85 %.

Many surfactants of low molecular mass have been claimed as useful secondary stabilizers but in practice few can be used without adverse side effects. One exception is the nonionic sorbitan monolaurate. Again, PVAs are much more widely used. However, this family is prepared by acid hydrolysis of low molecular mass poly(vinyl acetates) and has a random distribution of acetate groups [11]. Initially these materials, which are insoluble in water, were prepared as methanolic solutions (e.g., Polivic S 202) but more recently aqueous-based, low-hydrolysis PVAs (e.g., Alcotex A 55, and Polivic S 404 W), which are safer to handle and are environmentally more acceptable, have been developed and are now very widely used in the industry. Typical secondary protective colloids are listed in Table 3.

The use of secondary granulating agents affects the mean grain size and grain size distribution so the concentration of the primary must be adjusted to compensate. Since each producer uses autoclaves of different size and shape, equipped with different agitation and baffle systems, with or without condensers, and makes many different types of resin, no absolute formula is available which determines the magnitude of this change in primary concentration; this must be developed in-house by the suspension polymerization technologist.

Initiators. A range of initiators have been used, but currently a relative few are widely employed. At one time benzoyl peroxide and azo initiators were typically used, but diacetyl peroxides, peroxydicarbonates (PDC), and alkyl peroxyesters are now preferred. Typical initiators are listed in Table 4.

The half life $t_{1/2}$ at a particular temperature is the main factor which determines the choice of an initiator for a given PVC grade. The initiator should have a half-life of ca. 2 h at reaction temperature to optimize reaction time. Since the best plant output is obtained when the reaction profile is as square as possible (see Section 4.1.3) two initiators of different activity are often chosen to try to achieve this ideal. The more reactive initiator is used to start the polymerization but then as this exhausts itself the slower one takes over to completion. When preparing K 70 – 80 resins at temperatures at or below 52 °C, a peroxydicarbonate (e.g., **6**, **7**, **8** or **9**) is used as the primary initiator, with an active initiator such as **1** helping to bolster the slow initial rate which is obtained at these low temperatures. In the intermediate range of K 63 – 68 a single peroxydicarbonate (PDC) is usually sufficient, but at or below K 60 the PDC needs the extra radical flux of a slow initiator to prevent exhaustion of the reaction, so a less active diacyl peroxide such as **4** is used as a secondary.

Three other factors are also important: water solubility, storage and handling, and reactor fouling.

The initiators' water solubility can affect their efficiency by making them more susceptible to

Table 3. Typical primary and secondary protective colloids used in the suspension polymerization of vinyl chloride

Trade name	Manufacturer	Degree of hydrolysis	Viscosity of 4 % aqueous solution, mPa · s
Primaries			
Alcotex 72.5/B72	Synthomer	71.5 – 73.5	5.5 – 7.5
Alcotex 78	Synthomer	76 – 79	5.5 – 7.5
Poval L9	Kuraray	70 – 72	6.0 – 6.5
Poval L10	Kuraray	71 – 73	5.5 – 7.5
Gohsenol KP08	Nippon Gohsei	71 – 75	6 – 9
Gohsenol KH17	Nippon Gohsei	78.5 – 81.5	32 – 38
Gohsenol KH20	Nippon Gohsei	78.5 – 81.5	44 – 52
Gohsenol GH20	Nippon Gohsei	86.5 – 89	40 – 46
Secondaries			*Solvent Base*
Alcotex 55-002H/552P	Synthomer	54 – 57	aqueous
Polivic S202	3V-Sigma	47	methanolic
Polivic S404W	3V-Sigma	55	aqueous
Poval LM10HD	Kuraray	38 – 42	solid
Gohsenol LL02	Nippon Gohsei	45 – 51	solid
Atlas G72143	ICI	50	methanolic
Span 20 (sorbitan monolaurate)	ICI		red-brown liquid, no solvent

Table 4. Initiators for suspension or mass polymerization of VCM

No.	Name	Storage temperature[a], °C	M_r	Supplied form	$T_{1/2}$, °C [b]			$t_{1/2}$, min [c]		
					10 h	1 h	6 min	50 °C	60 °C	70 °C
	Diacyl peroxides									
1	acetylcyclohexane sulfonyl peroxide		– 15 222	solution	37	51	66	70	15	
2	dioctanoyl peroxide		10 286	solid	61	79	100		700	190
3	didecanoyl peroxide	25	343	solid	60	79	99		700	190
4	dilauroyl peroxide	30	398	solid	62	80	99		850	220
	Peroxydicarbonates									
5	bis(2-ethylhexyl) peroxydicarbonate	– 15	346	solution	44	61	80	250	70	21
6	dicyclohexyl peroxydicarbonate	10	286	solid	44	59	76	250	58	16
7	dicetyl peroxydicarbonate	25	571	solid or dispersion	45	61	79	320	70	20
8	bis(4-*tert*-butyl cyclohexyl) peroxydicarbonate	30	398	solid or dispersion	43	60	78	250	60	18
9	dimyristyl peroxydicarbonate	20	515	solid	46	62	79	320	70	20
	Alkyl peroxyesters									
10	*tert*-butyl peroxyneodecanoate	– 5	246	solution	42	62	84	250	70	25
11	*tert*-amyl peroxyneodecanoate	– 5	260	solution	30	59	83	560	180	55
12	cumyl peroxyneodecanoate	– 20	306	solution	36	54	74	100	30	9
13	*tert*-butyl peroxypivalate	– 5	174	solution	51	71	94	700	200	70
	Azo initiator									
14	Azobisisobutyronitrile (AIBN)	20	164	solution	59	79	101		500	170

[a] Maximum.
[b] Temperature at which $t_{1/2}$ = 10 h, 1 h, or 6 min, as indicated.
[c] Approximate.

hydrolysis and can affect their distribution between VCM droplets; the higher the solubility the more homogeneous the distribution.

The ease of storage and handling has a major influence on the acceptability of an initiator to a PVC producer. In general the shorter the half-life of the initiator, the more unstable it is, and the lower its self-decomposition temperature. The stability also depends on the radical energy and physical form. Solid crystalline initiators such as **8** (BCHPC) are more stable than their liquid analogues, e.g., **5** (EHPC). Many of the more active initiators must to be stored under refrigeration which tends to restrict their acceptability. Liquid initiators have the advantage that they can be automatically metered into reactors, which is

essential in larger computer-controlled plants. However, the trend is to use initiators as finely divided aqueous dispersions. This has the advantage of improving their homogeneity of distribution throughout the VCM droplets, allows them to be added at any time or temperature during the charging procedure, and renders them very safe to handle since their heat of decomposition is less than the heat of evaporation of the carrier. Dispersion of 25 – 40 % are typical, and systems must be carefully designed to ensure complete homogenization of the dispersion is maintained and accurate metering achieved. As the water soulbility of the initiator increases so does its tendency to cause reactor fouling. This is of increasing importance as more and more processes switch from the open to the closed mode of operation (see Section 4.1.4.4).

4.1.3. Kinetics

The kinetic profile seen in Figure 2 is typical of many VCM suspension polymerizations conducted at the lower end of the temperature range and exhibits the classical acceleration with conversion as shown by the gradually decreasing jacket temperature. At higher temperatures the reaction rate is more constant and a square-wave profile is obtained. Three different phases in the reaction can be identified:

Low Conversion (< 5 %). Polymerization occurs largely in the monomer phase because the quantity of polymer produced is so small. It is only this early stage that follows classical kinetics where the rate of polymerization is dependent on the square root of the initiator concentration:

$$\frac{-\mathrm{d}\,[\mathrm{M}]}{\mathrm{d}t} = \frac{k_\mathrm{p}\,[\mathrm{M}]\cdot k_\mathrm{d}\,[\mathrm{I}]^{1/2}}{k_\mathrm{t}}$$

where

[M] = monomer concentration
[I] = initiator concentration
k_p = propagation rate constant
k_d = initiator decomposition rate constant
k_t = termination rate constant

As the amount of polymer phase increases with conversion the rate deviates more and more from this equation due to a decrease in k_t. A full description of the models used to explain the deviations from ideality is given in [12], [13].

Medium Conversion (5 – 65 %). Polymerization continues in both the free monomer and polymer – monomer gel phases. This region shows the classical acceleration with increasing conversion [14]. Since termination occurs primarily by the diffusion together and collision of two bulky macroradicals, it is considerably slower in the viscous polymer phase. As a result the termination rate constant is dramatically reduced and the rate of polymerization in this phase is much faster. As the amount of polymer phase increases the rate of polymerization accelerates.

High Conversion (> 65 %). No free monomer is left to maintain the autogeneous pressure, which falls, and polymerization occurs in the polymer gel phase of rapidly increasing viscosity as the remaining VCM is consumed. Hence, the rate of polymerization increases still further to reach a maximum just after pressure drop but then decreases progressively due to increasing monomer starvation. When the instantaneous polymerization rate falls below the overall rate of reaction the batch is terminated, either by venting off unreacted VCM to the recovery plant or by adding a chain terminator and then venting. Depending on the properties required, a conversion of 70 – 95 % is chosen. It is uneconomic to continue to very high conversion, and as properties such as plasticizer absorption, dry up rate (i.e., the time taken to achieve a free-flowing PVC – plasticizer dry blend), and initial color – heat stability deteriorate rapidly above 85 – 90 % conversion, most flexible resins are stopped at around 85 % and rigid resins at around 90 % conversion.

4.1.4. Morphology

Control of the morphology of each PVC grain is the main parameter controlling the quality of a PVC resin. This subject covers mean grain size, grain size distribution, and, most important of all, porosity (cold plasticizer absorption) and bulk density, which are closely interrelated.

4.1.4.1. Control of Grain Size

At the beginning of the batch the bulk vinyl chloride phase is broken down into droplets of 30 – 40 μm mean size under the influence of agitation and by the presence of the water-soluble protective colloid, which is adsorbed at the monomer – water interface. As polymerization proceeds, graft copolymerization of PVC onto the interfacial layer of protective colloid reduces its mobility and protective ability, and droplet coalescence begins to occur at around 4 – 5 % conversion. As the conversion increases, changes in the surface layer of the droplet due to further polymerization suppress the droplet coalescence step, which ceases at 20 % conversion, and the grain size remains largely constant for the remainder of the polymerization. Altering the stirrer speed after this point has no further effect on the mean grain size (MGS) [15] of the final product, which is controlled to between 100 and 180 μm, depending on resin type, by a combination of initial agitator speed and protective colloid(s) concentration and type.

Polymerizations with high concentrations of colloid and less agitation give spherical grains of low porosity ($\leq$ 10 %) and very high bulk density ($\geq$ 700 g/L) and are of little interest, generally due to problems of VCM removal and poor gelation. Flexible resins are made with a low MGS (100 – 130 μm), rigid polymers higher (150 – 180 μm), and low K resins with intermediate size (130 –160 μm).

In extreme cases it is possible to stabilize the droplets to such a degree that no coalescence occurs. Specialized VCM – vinyl acetate copolymers for gramophone record production fall into this category as do blending resins (30 μm MGS) for use as viscosity depressants in plastisols (see Section 4.3).

As well as achieving the desired MGS it is equally important to control the size distribution. A Gaussean distribution is normally produced and it is important to make this as narrow as possible since a coarse fraction ($>$ 250 μm) can give rise to fisheyes (dispersion faults), and make VCM removal difficult as well as resulting in loss of yield due to screen rejects, while fines ($<$ 60 μm) can be lost in aqueous and gaseous effluent or cause powder flow problems.

4.1.4.2. Control of Porosity and Bulk Density

Porosity in PVC grains is created by a complex series of interrelated steps in the formation and growth of the submicroscopic structure within each droplet and also depends on the unique nature of the VCM – PVC system.

PVC is insoluble in its monomer, and the growing oligomers start to precipitate when the chain length exceeds 10 monomer limits [6]. This initial precipitation results in the formation of microdomains about 15 – 20 nm in diameter [16]. This occurs at as low as 0.001 % conversion [17]. The microdomains rapidly become unstable and aggregate to form domains (primary particle nuclei) of about 0.1 μm diameter. After this stage no further domains are formed and the number of primary particles is fixed. All new PVC radicals that are produced in the monomer phase collide with and are precipitated onto the vast existing surface area of primary particles before a new domain can be formed. Since the existing PVC phase is swollen with monomer and contains initiator, initiator radicals, and growing chains undergoing chain transfer, primary particles grow rapidly by accretion of new micro radicals and microdomains and by gel-phase polymerization. They in turn become unstable and a final agglomeration step to form 1 – 2 μm primary particle aggregates takes place. By the end of the polymerization the primary particles have grown to 1 μm within the aggregates, which themselves have grown to 2 – 5 μm. The whole mechanism is shown in diagrammatic form in Figure 5 and discussed in detail in [18].

The arrangement of the aggregates of primary particles within each monomer droplet has a major influence on the morphology of the final PVC grain. The conversion of VCM of density 0.85 g/cm^3 at 52 °C to PVC of 1.4 g/cm^3 produces a volume contraction of 39 %. If the polymerizing droplets contract totally the final grain has no porosity. If, however, the contraction is prevented totally the final grain porosity can be as high as 39 %. Therefore, the better the three-dimensional spatial arrangement of the aggregates relative to one another the higher the porosity that is achieved.

The main aim of the polymer technologist is to create the best possible spatial network of aggregates relative to one another so that a network is formed that is strong enough to resist droplet

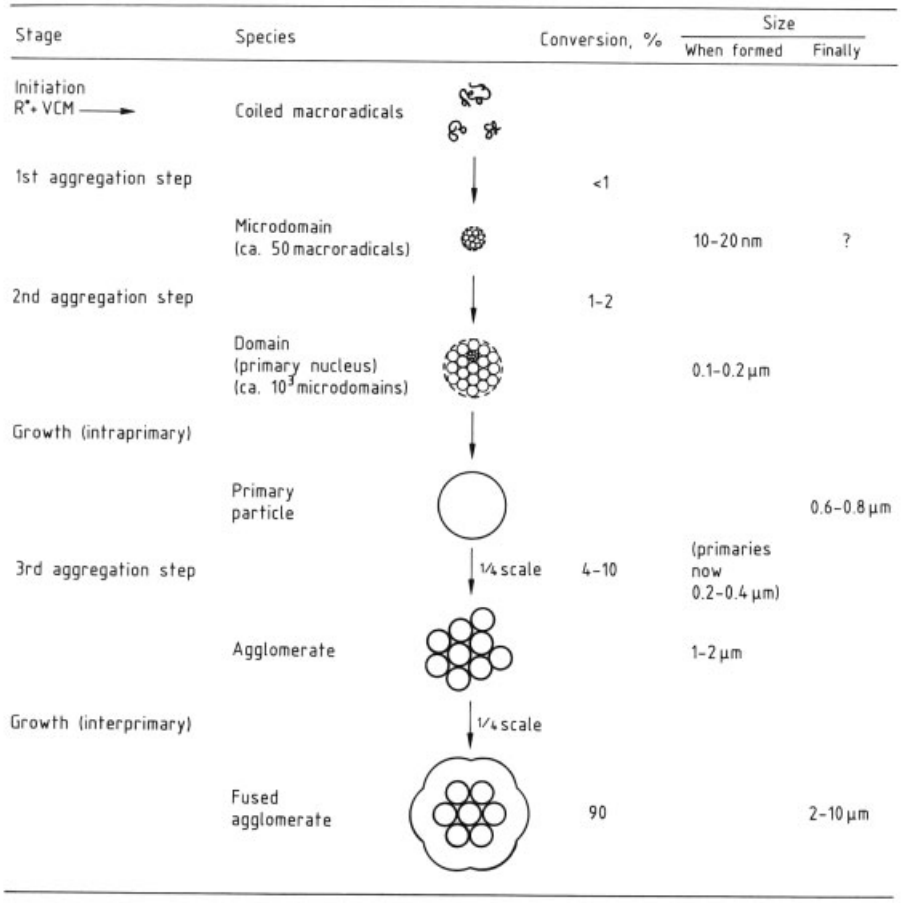

Figure 5. Schematic representation of the mechanism of VCM polymerization (from [18], courtesy of Elsevier Applied Science Publishers)

contraction and hence maximize porosity. The sooner this is achieved in the polymerization the better the porosity attained. In practice 34 – 35 % is easily achievable. At ca. 65 – 70 % conversion all of the free monomer is consumed, and the only VCM remaining is that swelling the polymer gel (27 %) and in the vapor space and dissolved in the water. At this stage the voids in the grain (the porosity) are filled with batch water, so that the quantity of water in interstitial spaces between the grains is reduced and a significant increase in slurry viscosity is observed since less mobilizing water is available to separate the grains.

The control of bulk density or apparent density (AD) is more important than even porosity to the processor. The latter is concerned with charge size and cycle times in high speed mixers, output rates, motor torque and degree of gelation in extruders, and mechanical properties of the article being produced. These parameters are very dependent on AD. The reason for introducing the control of porosity before control of AD is that the latter is very heavily influenced by the former. Apparent density depends on porosity, grain shape, and grain size distribution in order of priority and is best described by examining the packing fraction (PF) concept, i.e., the ability of the somewhat irregular PVC grains to fill a space to the maximum efficiency.

$$\mathrm{PF} = \frac{(1 + 0.014 \cdot \mathrm{CPA})\,(\mathrm{AD} \cdot 0.1)}{1.4}$$

CPA = cold plasticizer absorption (porosity)
AD = apparent density

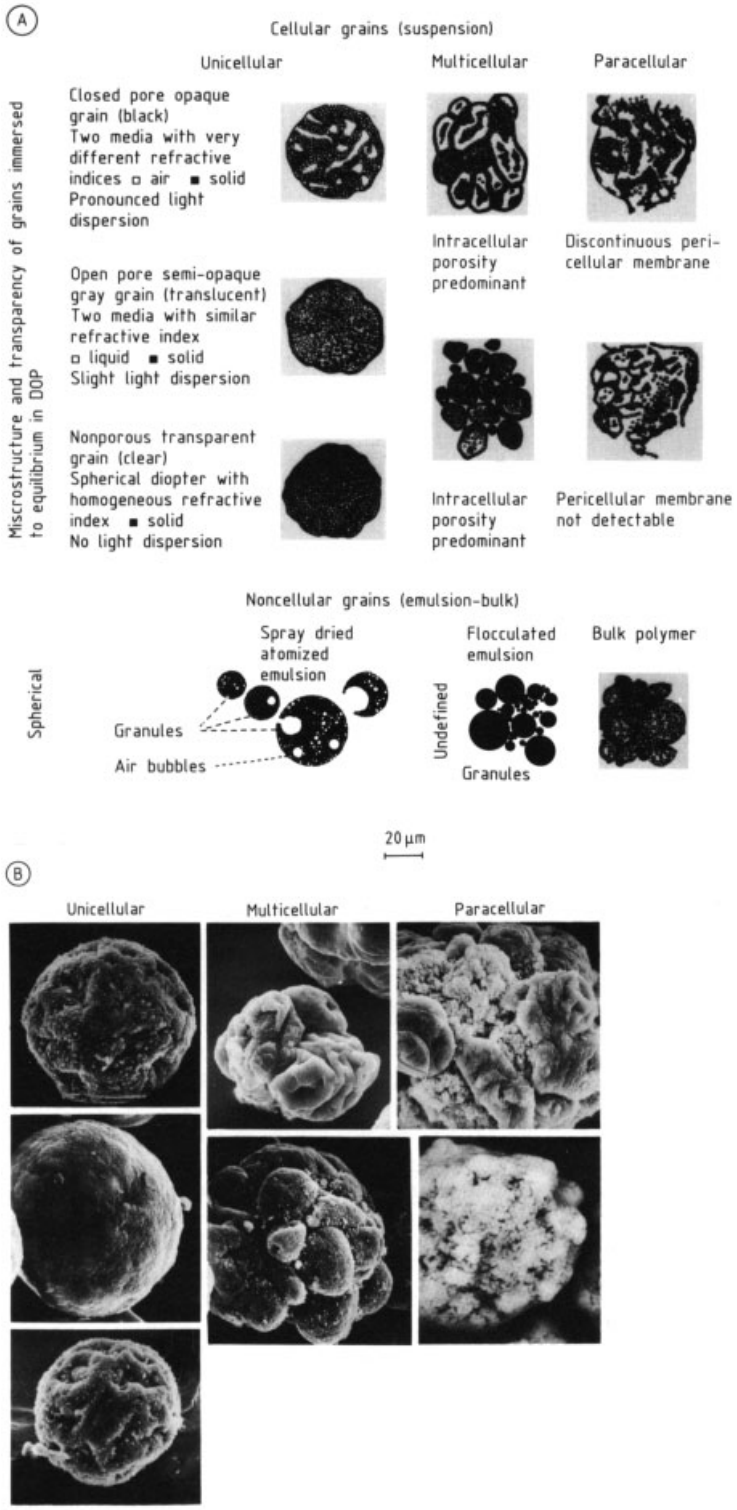

Figure 6. A) Classification of PVC grains; B) Typical cellular PVC grains – layout as in Figure 6 A
From [18], courtesy of Elsevier Applied Science Publishers

Most extruders used in processing PVC are of the twin screw type and these operate like a simple pump: increasing the feed rate by 10 % (10 % higher AD) increases the discharge or output rate by 8 – 9 % (see Chap. 8 in [14]). In practise, controlling porosity is relatively easy since a number of effective secondary stabilizers are available. Maximizing the AD is much more difficult since control of grain shape comes from a complex interaction of both agitation and choice of primary and secondary granulating agents together with in-house expertise of charging techniques and use of other additives. A range of grain shapes which can be produced is shown in Figure 6.

In practice typical property targets for porosity and AD are:

1) For flexible resins (K 70): porosity of ≥ 30 % and AD of ≥ 500 g/L
2) For rigid extrusion resins (K 66 – 68): porosity 20 % and AD 580 g/L
3) For bottle resins (K 57 – 60): porosity 18 – 20 % and AD 560 g/L

4.1.4.3. K-Value (Molecular Mass)

PVC is again rather unique in its relationship between reaction temperature and K-value. With PVC the rate of polymer formation by chain – chain termination is small due to the influence of gel phase polymerization described above. Instead, since chain transfer to monomer is a predominant step in the polymerization of VCM, the length of the polymer chain (degree of polymerization, DP) is determined by the ratio of the rate of chain propagation to chain transfer:

$$\mathrm{DP} = \frac{\text{Rate of propagation}}{\text{Rate of chain transfer}} = \frac{k_\mathrm{p}\,[\mathrm{R}^{\cdot}_n]\,[\mathrm{M}]}{k_\mathrm{tr}\,[\mathrm{R}^{\cdot}_n]\,[\mathrm{M}]} = \frac{k_\mathrm{p}}{k_\mathrm{tr}}$$

[M] = monomer concentration
$[\mathrm{R}\cdot_n]$ = radical concentration
k_p = propagation rate constant
k_tr = chain-transfer rate constant

As k_p and k_tr depend only on temperature the molecular mass of PVC is controlled by the reaction temperature (see Figure 7).

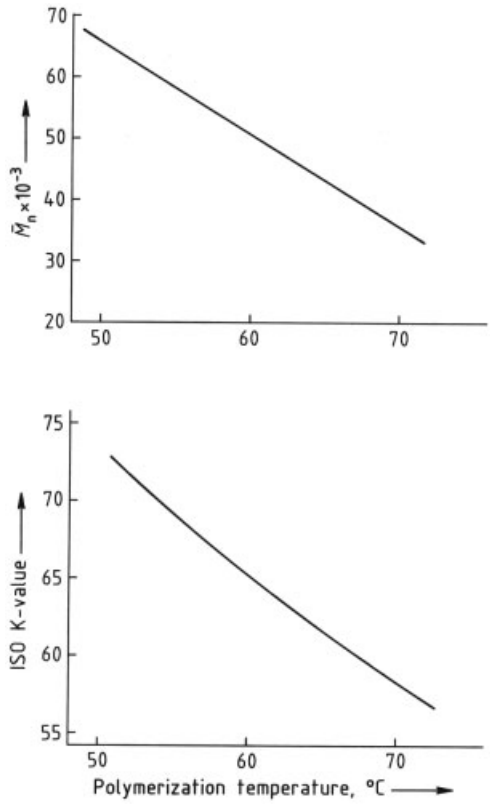

Figure 7. The effect of temperature on molecular mass and K-value

If a particularly low molecular mass polymer is required or if a reactor's pressure rating is too low to operate at very high temperatures an additional chain-transfer agent can be employed. This reduces the molecular mass by increasing the chain-transfer rate constant above that for monomer alone in the above equation. As a result DP decreases, depending on the quantity of chain transfer agent used.

The general K-value range of commercial resins is K 55 to K 75, with a few producers providing speciality resins outside this range to K 50 and K 80 and with just one or two K 90 – 100 polymers also being available. Despite the relatively narrow K-value/molecular mass range of most commercial resins (see Table 5) the polar nature of the PVC chain gives rise to a considerable melt viscosity span. Mechanical properties increase slightly with increasing K-value so wherever possible the highest K-value polymer is used. However, the power and mechanical ro-

Table 5. Correlation of various PVC molecular mass (expressions) ∗

European methods						Japanese method	American methods				
Viscosity number ISO/R 174 (1974)	Specific viscosity 0.5 g/100 mL cyclohexanone at 25 °C	K-value 0.5 g/100 mL^{-1} cyclohexanone at 25 °C	Relative viscosity 0.5 g/100 mL ethylene dichloride at 25 °C	Specific viscosity 1.0 g/100 mL cyclohexanone at 25 °C	K-value 1.0 g/100 mL cyclohexanone at 25 °C	Polymerization degree JIS K 6721	Inherent viscosity ASTM D 1243-58 T (method A)	Specific viscosity ASTM D 1243-58 T (method B)	Specific viscosity 0.2 g/100 mL cyclohexanone at 30 °C	Specific viscosity 0.4 g/100 mL nitrobenzene at 25 °C	Approximate number–average molecular mass
70	0.35	53.9	1.304	0.70	54	560	0.62	0.239	0.13	0.24	36 000
73	0.37	55	1.316	0.74	55	600	0.65	0.25	0.14	0.25	
77	0.38	56.1	1.329	0.77	57	640	0.67	0.264	0.145	0.26	40 000
80	0.40	57.2	1.342	0.80	58	680	0.70	0.275	0.15	0.27	
83	0.42	58.3	1.355	0.83	59	720	0.73	0.285	0.155	0.29	
87	0.44	59.5	1.369	0.86	60	760	0.75	0.3	0.16	0.30	45 500
90	0.45	60.6	1.383	0.87	61	800	0.78	0.31	0.17	0.31	
94	0.47	61.9	1.397	0.92	62	840	0.80	0.32	0.175	0.32	50 000
98	0.49	62.9	1.412	0.95	63	885	0.83	0.33	0.18	0.33	
102	0.51	64	1.427	0.98	64	930	0.85	0.34	0.19	0.34	
105	0.53	65.2	1.443	1.01	65	975	0.83	0.36	0.195	0.35	55 000
109	0.55	66.3	1.458	1.04	66	1025	0.91	0.37	0.20	0.36	
113	0.57	67.4	1.474	1.07	67	1070	0.92	0.38	0.205	0.37	
117	0.59	68.5	1.491	1.10	68	1120	0.95	0.39	0.21	0.39	60 000
121	0.61	69.7	1.508	1.13	69	1175	0.99	0.40	0.22	0.39	
125	0.63	70.5	1.525	1.16	70	1230	1.01	0.41	0.225	0.41	64 000
130	0.65	71.2	1.543	1.19	70.5	1300	1.03	0.43	0.23	0.43	
134	0.67	72.2	1.562	1.22	71	1350	1.06	0.44	0.235	0.44	
138	0.69	73.0	1.581	1.25	72	1420	1.08	0.45	0.24	0.46	70 000
142	0.71	74.0	1.60	1.28	73	1490	1.11	0.46	0.25	0.48	
145	0.73	74.5	1.62	1.31	74	1570	1.13	0.47	0.255	0.49	73 000
149	0.75	75.5	1.64			1650	1.16	0.49	0.26	0.50	
153	0.77	76.3	1.661			1720	1.18	0.50	0.27	0.51	
157	0.79	77.0	1.682			1810	1.21	0.51	0.275	0.52	
161	0.81	77.5	1.704			1900	1.23	0.53	0.28	0.53	80 000
165	0.83	78.5	1.726			1980	1.26	0.54	0.29	0.54	82 000

∗ From [65], courtesy of Elsevier Applied Science Publishers.

bustness of equipment and the complicated melt flow pattern requirements in molds often restrict the K-value which can be used. For example, most bottles are made from K 57 polymers, some from K 60, but this is the limit for the application because bottles are blow molded so restricting the melt viscosity which can be handled.

4.1.4.4. Problem Areas

There are relatively few problems in the operation of the suspension polymerization process. However, to obtain the best possible quality and consistency some points need special attention.

In the past reactor fouling was bad and mechanical cleaning between each batch essential. With the advent of clean-wall technology, high-pressure rinse systems, and improved recipes it is now possible to operate in a closed manner, thereby reducing problems of VCM exposure (see Chap. 5 in [14]). Reactor opening for cleaning is reduced to once in several hundred batches. It is essential to ensure that the reactor is rendered completely free of polymer between batches since this is the most common cause of fisheyes. Equally important is to ensure there is no chance of cross-grade contamination in the downstream operations of slurry handling, stripping, drying, and storage. Contamination of the product by foreign substances is eliminated by good housekeeping and plant management together with the right choice of metals of construction and design. Stainless steel is widely used in many sections of PVC plants.

4.2. Bulk or Mass Polymerization of Vinyl Chloride

The mass polymerization process is virtually identical to that taking place within each monomer droplet in the suspension process. As such the mechanism and kinetics are very similar to those described in Section 4.1.3. The major difference is in the mechanical operation of the process.

About 10 % of world PVC production is provided by the mass process. The process was developed by Pechiney St. Gobain (PSG) who are now part of the French ATO Company. All mass producers are licensees of the PSG technology. The process is carried out in two stages:

1) In the first stage monomer and initiator are charged and vigorously agitated in a vertical, stainless-steel autoclave (prepolymerizer) of 8 – 25 m^3 capacity, fitted with a water- cooled jacket and condenser. Rapid polymerization takes place at 62 – 75 °C to give ca. 100 µm aggregated spherical flocs composed of 0.1 µm primary particles, which form the basis or seeds of the final PVC grain. Conversion is taken to 7 – 12 % in ca. 30 min, by which time the initiator used is exhausted.
2) The slurry from the prepolymerizer is discharged into the second-stage reactor, which has a volume of 12 – 50 m^3, with fresh initiator and more VCM. Up to five second-stages can be fed from one prepolymerizer. In the second stage the 0.1 µm primary particles making up each seed grow in size and fuse together to give the final grain of PVC of 130 – 160 µm diameter. As the conversion increases the physical nature changes from a wet powder at ca. 20 % to a dry one at 40 % conversion. The heat of polymerization is removed from the growing grains by evaporation of the VCM and condensation on the cooled wall of the reactor or in a water-cooled reflex condenser(s). The second stage takes 3 – 9 h, depending on K-value.

Originally, the second stage reactor was a horizontal jacketed vessel with one or more condensers, agitated by slowly rotating (6 – 10 rpm) ribbon blenders with some blades set very close to the reactor wall to keep it clean. However, it proved difficult to discharge and clean the reactor to a sufficiently high standard to prevent the formation of fisheyes (dispersion faults) in PVC grades intended for plasticized applications [19]. The product can contain up to 10 % of coarse material which must be sieved out and recovered by grinding, milling, and reclassifying.

This led to the development in 1978 of a vertical second-stage reactor of up to 50 m^3 capacity fitted with two independent agitators designed to overcome the above discharge problems. However, there has been little interest in this development and all recent investment in new PVC plants has been in the suspension process. In theory the mass process should be cheaper since it does not need a drying stage. However, it is a two-stage process, it is difficult to remove traces of unreacted monomer, and it gives a significant amount of oversized material, all of which offset its advantages.

Several of the smaller and older mass-polymerization plants have closed in recent years because they find it more and more difficult to meet increasingly stringent environmental standards and are becoming less and less competetive compared to suspension-polymerization plants on the basis of production rate.

The mass process is discussed in more detail in [19], [20].

4.3. Emulsion Polymerization of Vinyl Chloride

4.3.1. Introduction

The emulsion polymerization process involves the polymerization of monomer in an aqueous medium containing surfactant and a water-soluble initiator, producing PVC latices. PVC latices are colloidal dispersions of spherical particles, ranging in size between 0.1 and 3.0 µm. Most PVC latices are spray dried and then milled to obtain fine powders, made up of agglomerates of latex particles. When mixed with plasticizers they disperse readily to form stable suspensions. During mixing most of the agglomerates are broken down into the original latex particles. Such dispersion of fine particles in plasticizers are known as plastisols or pastes, and the powder is called dispersion or paste polymer.

The surfactant layer around the particle surface prevents their adsorbing the plasticizer at room temperature so they can be used as liquids and may then be spread on to fabric or other substrates, poured on molds, or deposited on formers to produce flooring, wall covering, artificial leather, balls, toys, or protective gloves. There are other grades of PVC polymers, produced by emulsion polymerization, that do not form plastisols and that are used as blends with suspension PVC grades for extrusion application or in the manufacture of battery separator plates. These so-called emulsion polymers are of only minor economic interest. Sales in latex form are very limited; latices are used in water-based paints, printing inks, and impregnated fabrics. The total production capacity of emulsion PVC is given in Table 6.

Table 6. World emulsion PVC production in 1989

Region	Production, 10^3 t	Percentage
Western Europe	662	41
United States	278	17
Japan	155	10
Rest of the world	513	32
Total capacity	1608	100

4.3.2. Emulsion Polymerization Processes

The ways to produce PVC latices are:

1) The classic emulsion polymerization, in which the particle size and particle size distribution are controlled by a water-soluble initiator and a surfactant [21–25]. They are monodisperse with a particle size typically less than 0.1 µm.
2) Seeded emulsion polymerization, in which the particle size distribution of the final latex depends on the amount and the size of the latex used as seed and on the amount of surfactant and the way in which it is added [26], [27]. A typical material has 0.2 and 1.2 µm particles (see Fig. 8).
3) Microsuspension polymerization, in which a monomer-soluble initiator is used. Polymerization takes place within the fine droplets in which the monomer is dispersed by passing a coarse emulsion of monomer, initiator, and surfactant through a mechanical homogenizer [28–30]. The resulting particle size distribution is Gaussian between 0.2 and 1.2 µm.
4) Polymerization of fine emulsions achieved by using a combination of a typical surfactant such as sodium dodecyl sulfate and a long-chain fatty alcohol. This gives rise to spontaneous emulsification of the monomer into very fine droplets. In this case the polymerization can be carried out by using a water-soluble or a monomer-soluble initiator [31]. The latex particle size distribution is the same as for microsuspension polymerization.

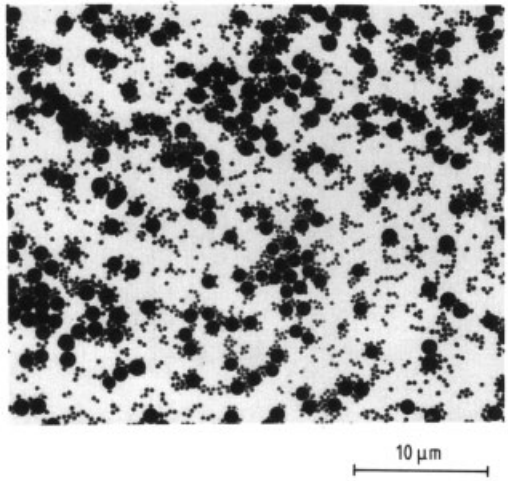

Figure 8. Electron micrograph of latex particles obtained by seeded emulsion polymerization

The industrial polymerization techniques used for carrying out these processes are the following:

1) Batch polymerization where all the monomer is charged at the beginning
2) Semibatch or semicontinuous polymerization, with the monomer added in stages or continuously as reaction proceeds
3) Continuous polymerization where monomer and other components are fed continuously into the reactor and the latex is withdrawn from the bottom of the reactor

These processes are used to achieve a wide range of latex characteristics, which are related to the rheological properties of the PVC plastisols [32].

4.3.3. Description of the Batch Process

The emulsion polymerization of PVC consists of the following stages:

1) Polymerization,
2) VCM removal,
3) latex storage,
4) drying,
5) milling,
6) packing and storage.

Figure 9 shows a schematic representation of the stages required for a plant which produces all the emulsion, seeded, and microsuspension grades.

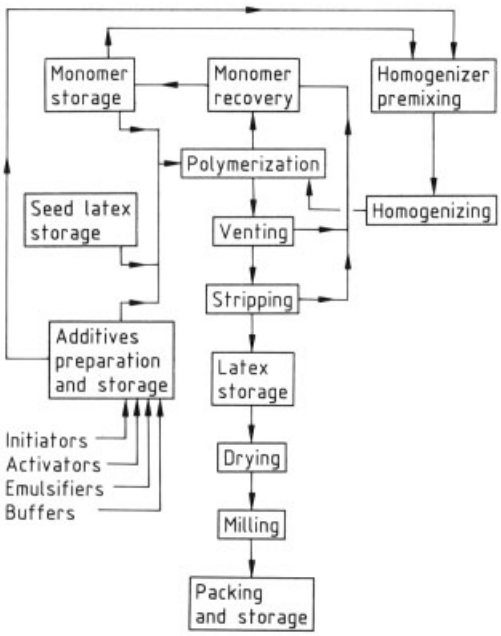

Figure 9. Schematic representation of the stages required for a plant which produces all emulsion, seeded, and microsuspension grades

A recipe for a simple batch emulsion polymerization (in parts by weight) is as follows:

Demineralized water	110 – 140
Vinyl chloride	100
Emulsifier	0.1 – 1
Initiator	0.1 – 0.2

The polymerization takes place in an autoclave resistant to the VCM vapor pressure. Pressures of 6.4 – 10.0 bar correspond to polymerization temperatures of 40 – 60 °C, depending on the desired molecular mass. The polymerization reaction is strongly exothermic with a heat of polymerization of 1534 kJ/kg. Heat removal can be a problem as the size of the autoclave increases because the ratio between surface area and volume becomes less favorable. For this reason industrial plants commonly use water cooling, and external condensers. Modern plants use autoclaves with capacities ranging between 30 and 80 m^3. Large autoclaves can present agitation problems because low shear rates are required to maintain mechanical stability of the latex and to avoid coagulation.

Initiators. Although organic peroxides slightly soluble in water such as methyl ethyl ketone peroxide and 1-hydroperoxy-1′-hydroxydicyclohexyl peroxide can be used, the most widely used initiators are ammonium or potassium peroxosulfate. Their decomposition in aqueous solution is first order with respect to the peroxosulfate ion concentration [33]. The rate may be increased significantly by using a reducing agent. Redox catalysis involves the use of transition metal ions such as copper or iron [34]. The systems widely used in emulsion polymerization of vinyl chloride are potassium peroxosulfate – sodium bisulfate – copper or iron. Redox systems can also be used in microsuspension, an example being dilauroyl peroxide – copper – ascorbic acid [35].

Emulsifiers are very important in the emulsion polymerization of vinly chloride because they determine not only the latex characteristics but also the properties of the final product [36–38]. Moreover, the nature and the concentration of the emulsifier, together with the agitation, are major parameters controlling the mechanical stability of the latex during polymerization and handling, and the major cause of reactor wall encrustation and lump formation. Since the latex is dried by spray drying, the emulsifier remains in the resin and can influence properties of the product such as the flowability of the powder, plastisol formation, the viscosity of the plastisol, and also the properties of the final article such as heat stability, water adsorption, and coat adhesion. For this reason a wide variety of emulsifiers are used. During polymerization anionic types are usually used, while nonionics are generally added to the latex after polymeriza-

tion. Typical emulsifiers are the sodium salts of alkyl sulfates, alkyl sulfonates, alkylbenzenesulfonates, dialkyl sulfosuccinates, alkyl ethoxysulfates, fatty acid soaps, alkyl phenol ethoxylates, and fatty acid ethoxylates.

Particle Formation. Emulsifiers are characterized by the critical concentration at which their molecules aggregate to form micelles [39], [40]. In Figure 10 a plot of surface tension versus emulsifier concentration is shown; the point at which the curve exhibits a marked change is known as the critical micelle concentration (CMC). Conventional emulsion polymerization is carried out at a concentration above the CMC. In this way, particle formation is controlled by the micelles, which solubilize the monomer by swelling and become the loci of polymerization. The nucleation phase stops when the micelles disappear, and the emulsifier is absorbed on to the surface of the particles. Depending on the balance between the hydrophilic and hydrophobic part of the molecule, each emulsifier plays a specific role in the emulsion polymerization, as they have different CMC values, surface covering powers, and stabilizing effects.

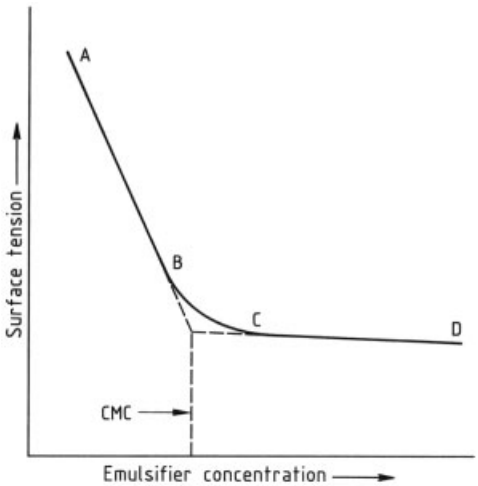

Figure 10. Dependence of surface tension on emulsifier concentration

VCM Removal. The polymerization is terminated at 90 % conversion and is followed by the venting of unreacted monomer to a gas holder. After degassing, the latex contains 3 % vinyl chloride and further reductions must be made to fulfil legal requirements for environmental protection. The residual VCM is removed by stripping under vacuum in a stripping vessel. Although VCM loss from particles by diffusion is fast, VCM removal from an emulsion is much more difficult than from suspension slurries as the surfactants lead to large amounts of foam formation. Continuous processes using a thin-film evaporator or which spray the latex into an evacuated chamber against a countercurrent stream of gas reduce considerably the problems of foaming and coagulation [41–44]. The residual vinyl chloride monomer content of the latex after stripping ranges between 200 and 2000 ppm, depending on the type of latex and the operating conditions.

Drying and Milling. Although there are several different techniques to separate polymer from water, such as freeze drying, drum drying, or coagulation followed by centrifugation, the most widely used is spray drying because this confers a special structure to the dried particle that is important for the production of plastisols. A fine spray of latex is introduced into a hot air stream where evaporation takes place.

The spray is obtained by various types of atomizers, such as:

1) Spinning disks or wheels rotating at 15 000 – 20 000 rpm
2) Single pressure nozzles operating at 100 – 500 bar
3) Two-fluid pressure nozzles using compressed air up to 4 bar as atomizing fluid [45]

The drying air temperature at the inlet is 150 – 240 °C and at the outlet, 55 – 80 °C, depending on the quality of the required product and the type of latex. By varying the drying conditions it is possible to change the resin characteristics, and hence its use, even from a typical paste grade to a general-purpose emulsion resin, e.g., for rigid extrusion. The dry powder consists of spherical aggregates, with the size usually ranging from 1 to 100 µm diameter with an average of 30 – 40 µm. The aggregates are called secondary particles, to distinguish them from those of the latex, which are known as primary particles. The way the aggregates form is complex and depends on many parameters such as primary particle size, type and amount of emulsifier, mean residence time in the drying cham-

ber, droplet size of the spray, and temperature [46]. Figure 11 shows a scanning electron micrograph of spray-dried PVC particles formed from bimodal seeded latex.

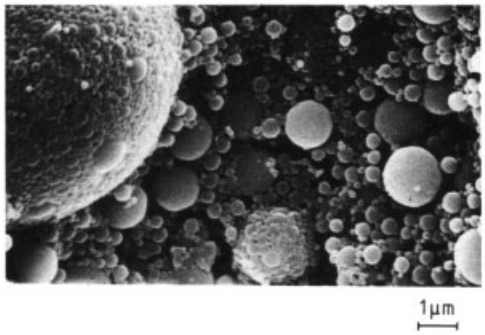

Figure 11. Scanning electron micrograph of spray-dried PVC particles produced from bimodal seeded latex

Plastisol-grade resins, usually made from multimodal latex (i.e., with a Gaussian particle size distribution), are milled to reduce the size of the coarse fraction, which can cause problems in coating applications of the plastisol. Generally the viscosity of the plastisol increases when the polymer has been milled, so it is desirable to reduce milling as far as possible and to use air classification to remove the coarse fraction ($> 65\,\mu m$) under control. In industrial production, air classifiers or mills or a combination of both are used. After milling, the average particle size of the resin is in the range of $5-15\,\mu m$.

Polymers for rigid extrusion and for manufacturing plate separators in batteries are sold in unmilled form. Resins are characterized by analysis of molecular mass, emulsifier and salt content, pH of aqueous extract, granulometry, and viscosity and rheological behavior of the pastes.

4.3.4. Plastisols

Plastisols, commonly called pastes, are dispersions of PVC powders in plasticizers. A wide range of plasticizers are used in plastisol formulations, but the most widely used are C_4–C_{12} esters such as phthalates, adipates, azelates, sebacates, trimellitates, and phosphates. Chlorinated paraffins are used as secondary plasticizers. A good plasticizer should have low volatility, low color value, neutral reaction, resistance to hydrolysis, insolubility in water, flame resistance, and nontoxicity. As no plasticizer can satisfy all these properties several plasticizers are mixed to achieve the desired properties of the final product. Table 7 shows the influence of plasticizers on plastisol properties; Table 8, the influence of plasticizers on final product properties; and Table 9, the influence of resin characteristics on plastisol and final product properties.

Table 7. Influence of plasticizers on plastisol properties

Plastisol properties	Plasticizers *
Low fusion temperature	DBP, BBP, TCP
Low viscosity	DOA, DOS, DIPA
High viscosity	BBP, TCP, polymerics
High fusion temperature	DIDA, DOZ, DTDP, DOS, polymerics

* DBP = dibutyl phthalate, BBP = butyl benzyl phthalate, TCP = Tricresyl phthalate, DOA = Dioctyl adipate, DOS = Dioctyl sebacate, DIDA = Diisodecyl adipate, DOZ = Dioctyl azelate, DTDP = Ditridecyl phthalate.

Table 8. Influence of plasticizers on final product properties

Product properties	Plasticizers
Flame retardancy	phosphates, chlorinated paraffins
Light and heat resistance	epoxy plasticizers
Low-temperature flexibility	adipates, azelates, sebacates, straight-chain phthalates
Low volatility	polymerics, trimellitates, linear phthalates, epoxy plasticizers
Stain resistance	butyl benzyl phthalate
Low migration	polymerics, trimellitates

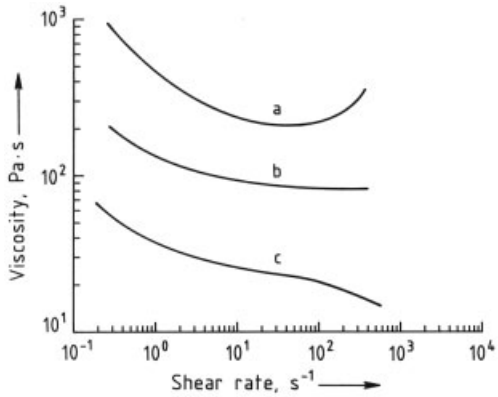

Figure 12. Influence of plasticizer content on plastisol viscosity
a) PVC/DOP = 100/40; b) PVC/DOP = 100/50; c) PVC/DOP = 100/60
DOP = dioctyl phthalate

Table 9. Influence of resin characteristics on plastisol and final product properties ∗

Plastisol and product properties	Effect of resin characteristics			
	Molecular mass	Particle size	Particle size distribution	Emulsifier type and quantity
Paste making		S	S	S
Plastisol viscosity		S	S	S
Air release				S
Gelation	M	M	M	S
Fusion	S			
Mechanical properties	S			
Gloss		S	M	
Transparency		S		S
Foaming	S	M		S
Fogging				S
Streaking flow		S		
Water resistance				S

∗ M = moderate effect; S = strong effect.

The amount of plasticizer used in plastisol formulations plays a major role in determining the hardness and flexibility of the end product and can vary from 40 to 130 phr. In Figure 12 the influence of plasticizer content on plastisol viscosity is shown.

The other essential ingredients in plastisol formulations are:

1) Heat stabilizers, which generally are the same compounds used for suspension polymers.
2) Dyes and pigments suitable for coloring plastisols. The most widely used is titanium dioxide.
3) Fillers, generally inorganic materials such as calcium carbonate, barite, silicate, kaolin or china clay, are used primarily to lower costs and also to impart special properties such as hardness, abrasion resistance, and no sticking.
4) Blowing agents for foamed PVC production such as azodicarbonamide.
5) UV adsorbers to prevent the decomposition action of sunlight.
6) Esters of poly(ethylene glycol), phosphate esters, and fatty acid amines and amides are usually employed as antistatic agents.
7) Viscosity depressants, such as white spirit, poly(ethylene glycol) monolaurate, alkylphenols, low molecular mass paraffins.
8) Extender resin or filler polymer – a PVC grade of spherical form, made by the suspension process and therefore cheaper than dispersion resin, which is used to reduce cost and viscosity of the plastisol [47], [48].

In Table 10 the influence of plastisol ingredients on the final product characteristics is shown.

4.3.5. Polymer – Plasticizer Interaction

Figure 13 illustrates the types of dependence of viscosity on shear rate;

1) Newtonian (viscosity is independent of shear rate)
2) Pseudoplastic (viscosity decreases as shear rate increases)
3) Dilatant (viscosity increases with increasing shear rate)

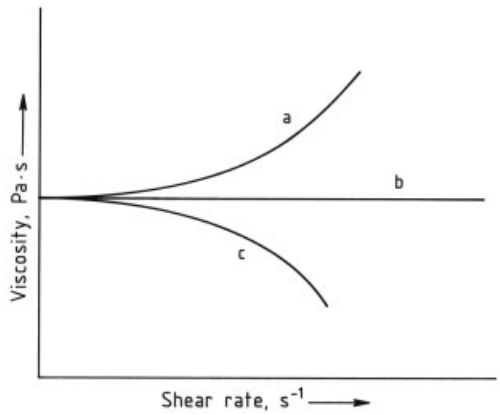

Figure 13. Effect of shear rate on dilatant (a), Newtonian (b), and pseudoplastic (c) fluids

Table 10. Influence of plastisol ingredients on the final product characteristics ∗

Characteristics	Filler	Pigment	Stabilizers	Quantity of plasticizer	Plasticizer type	Resin
Tensile properties	M			S	M	S
Flexibility	M			S	M	
Hardness	M			S	M	
Low-temperature bending resistance	M			S	S	M
Gloss	S	S				S
Volatility				S	S	
Extraction				S	S	
Color	S	S	M			
Light resistance		S	S		M	
Cost	S		M	M	M	

∗ M = moderate effect; S = strong effect.

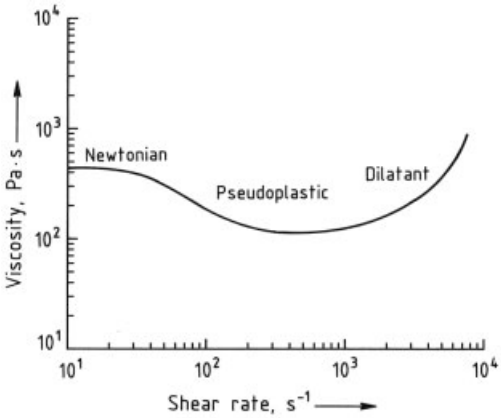

Figure 14. Effect of shear rate on the PVC plastisol viscosity (DOP/PVC = 100/50 at $T = 25\,°C$)

Over a wide range of shear rate a plastisol may exhibit all three types of behavior since the viscosity of plastisols rarely displays true Newtonian behavior. At low shear rates they are generally pseudoplastic, at intermediate values they show dilatancy, and may sometimes become pseudoplastic again at higher shear rates, as shown in Figure 14. There is no certainty on the exact mechanism as the rheology of PVC plastisols is complex and affected by many factors such as:

1) Polymer/plasticizer mixing ratio
2) Type of plasticizer and its interaction with polymer
3) Particle size and their distribution of primary particles
4) Structure, shape, porosity, and surface of the secondary particles
5) Surfactants, added during polymerization or in the paste
6) Molecular mass and molecular mass distribution
7) Paste-making conditions such as mixing temperature, time, and intensity [49].

4.3.6. Gelation and Fusion

Gelation is defined as the change from liquid to semisolid state, and fusion, when an homogeneous phase at molecular level between polymer and plasticizer is reached. These processes have been examined by determining viscosity changes of the liquid system, the tensile strength, and changes in complex viscosity and viscoelastic properties over the temperature range 25 – 200 °C.

At the beginning of heating, plasticizer viscosity decreases, as does the plastisol viscosity. At the same time, polymer particles adsorb plasticizer and swell, reducing the fraction of liquid plasticizer, and the viscosity then increases rapidly. When the liquid phase is completely absorbed by the particles, the system becomes dry. As the process continues the interparticle boundaries disappear and the polymer chains become increasingly entangled and development of the physical properties begins. At this point the gelation stage is complete and the fusion process starts, as illustrated in Figure 15. Further heating of plastisol makes the polymer flow into a continuous mass. The melt viscosity, after having reached a maximum, starts to decrease and the process passes through the fusion stage.

Morphology of the particles, particle size, particle size distribution and molecular mass of the resin affect the viscosity behavior. Low molecular mass resin starts the gelation phase earlier and completes the gelation process at lower temperature. The same result is obtained by using plasticizers with higher solvating action, while larger particle size increases gelation temperature or time because it slows the penetration of the plasticizer into the resin [50].

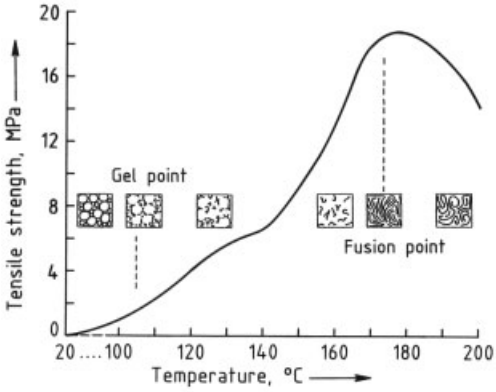

Figure 15. Changes in morphology and mechanical properties of PVC plastisol during gelation and fusion

4.3.7. Applications

Vinyl plastisols can be applied by spread coating, knife coating, roll coating, molding, dipping and spraying; the most widely used application is coating.

Coating. Vinyl plastisol can be used to coat paper, fabrics, metals, felt, and glass fibers to produce wallpapers, floor coverings, vinyl leather, conveyor belts, and tarpaulins. Each application needs its own formulation, in which the appropriate resin grade and plasticizer must be used to give the required rheological characteristics. In the case of fabric coating where direct spreading is employed, it is important to avoid penetration of plastisol into the fabric; therefore, the viscosity behavior of the paste must be strongly pseudoplastic at low shear rate and have moderate viscosity at high shear rate. In wallpapers, where thin coatings are applied, a low-viscosity resin with Newtonian behavior is required. In production of floor coverings, three or even four coatings are applied: an impregnation layer to saturate the fibers, one or two foam coats, and a wear layer with high mechanical strength. Each coat, which needs different grade resin to achieve the required properties, is gelled at 150 °C before application of the next layer. When the last coat is applied the entire coating is fused in an oven at 200 °C. The foam layer can be chemically or mechanically foamed [51]. Usually mechanical embossing is used to create surface textures and apply colored patterns.

Dipping. The product, mainly work, household, and surgical gloves, is obtained by dipping metal or ceramic molds or an article to be coated into the plastisol and then withdrawing it, draining it, and then curing it in an oven.

Molding is used for producing hollow articles such as balls, dolls, and toys. The process requires a plastisol with low viscosity at low shear rate that has a short gelation time and is easy to deaerate. In rotational molding the plastisol is poured into a cold mold that rotates around two perpendicular axes whilst entering an oven where it is heated with air to 200 – 250 °C. After gelation and fusion, the mold is cooled in a waterbath.

Further information is given in [32], [39], [40], [45], [48], [52–54].

4.4. Chlorinated Poly(Vinyl Chloride), CPVC

The softening point (heat distortion temperature or T_g) of PVC can be increased by chlorination of the polymer, either in a solvent or as a suspension. The normal chlorine content of 56.7 % is raised to 63 – 68 %, depending on type, and this increases the softening point from ca. 78 – 83 °C (depending on formulation) to at least 120 °C.

The older method involves chlorination of the PVC dissolved in a chlorinated solvent, but the process is expensive because of the need for an extensive solvent recovery system. However, it has the benefit of yielding a homogeneous material. The product is difficult to handle physically because of its very low bulk density.

The most common commercial process disperses the PVC grains in water and swells them with a chlorinated hydrocarbon. The reactor is

degassed to remove oxygen and heated to reaction temperature (50 – 60 °C); chlorine is added continuously and the system irradiated with UV light to produce chlorine radicals. Hydrogen chloride is produced as a byproduct and is removed at the end by washing with an inorganic base, after which the product is dried [55]. The degree of chlorination is variable, with the surface layers of the grains being preferentially treated.

CPVC is less thermally stable than conventional PVC and requires higher temperatures to develop good mechanical properties. Hence, it is difficult to process and quite expensive, but blends with suspension PVC can be used to achieve a specific softening point. CPVC products have not been very successful commercially but have found a specific niche in the market. They are used for hot-water systems in mobile homes in the United States and for the manufacture of fibers, for use in thermal underwear, in France.

5. Environmental Protection

Potential problem areas associated with the production, handling, and use of PVC are small if a few simple measures to avoid environmental contact and pollution are followed.

Losses of VCM from polymerization plants have been reduced to extremely low levels following conversion to closed-lid operation as a result of the toxicological problems discussed in Chapter 10 in [14]. Although modern column strippers reduce VCM levels in PVC to less than 10 ppm, a small quantity of VCM is lost from the drier section. The older batch strippers present more of a problem since they cannot achieve such low residual VCM contents.

PVC powder can be lost in centrifuge and scrubber effluent and by overfilling stock tanks and from pipeline leaks. It is important that all plant drains which are likely to contain PVC are directed to settling pits where flocculation treatment can be used so that the final discharge is free of PVC. All wet waste is collected and reclaimed as second- or third-grade material for use in noncritical applications such as one-trip horticultural products and fencing.

Powder losses from drier stacks can be minimized by ensuring that drier cyclones are adequately sized to the air conveying rates used. Some modern plants use bag filter systems to meet regulations that restrict the amount of PVC dust loss to 50 mg/m^3 air [56].

Compared to other polymers and typical metals PVC requires less energy per unit weight for its production [57]. Whilst plastics in general are blamed for long-lasting litter problems, it is significant that only 12 % of PVC is used for consumables, while 64 % is incorporated into products with an expected useful life of 15 years or more, probably much more. Disposal is therefore less of an immediate problem than is often thought.

As landfill sites are used up more municipal waste is being incinerated. Though inherently nonflammable, PVC will burn in an existing fire to release hydrogen chloride gas, which can be scrubbed out satisfactorily. Fears about the increased release of dioxins during incineration due to the presence of PVC in waste have proved to be unfounded [58].

6. Quality Specifications and Analysis

All PVC producers test a number of parameters on every batch produced as part of routine quality control operation. These include Kvalue, mean grain size and grain size distribution, apparent density, porosity, contamination, dispersion, color and heat stability, and residual VCM content. Most plants use statistical process control techniques to ensure that product consistency from batch to batch meets the stringent requirements of today's customers. It is difficult to give a typical testing regime since every plant makes batches of different size, has slurry blending facilities of differing degrees of sophistication, and has hoppers of different capacity on their drying plants. For a modern plant operating large autoclaves there is considerable testing of each polymerization batch (wet stage) together with additional testing of the dried and blended material as a cross check.

In-house tests have been collected and correlated by national and international standards organizations to make them universally applicable. ISO has defined tests and standards set up by other organisations such as the ASTM and DIN

so that ISO methods are now widely accepted worldwide. ASTM D 1755 defines a system for classifying general purpose (mass and suspension) and dispersion (paste) resins [59].

7. Storage and Transportation

Most plants handle more than 80 % of their production in bulk, and storage and supply in 25 kg bags is diminishing all the time as customers convert. After drying, the material is conveyed by blow-eggs (vessels in which the powder is pressurized and fluidized) of 2 – 15 t capacity via pipeline to the silo farm where it is stored in silos of capacity 100 – 500 t. PVC is discharged either directly or via a smaller loading hopper into 20 t road tankers, the most popular form of PVC transportation. Sometimes big-bags made of woven polypropylene are used for 1 t lots of PVC, and 20 t ISO metal containers are used for long distance deliveries. PVC is not a difficult material to handle and the logistics side of the operation is straightforward. It is essential to prevent cross-contamination of the PVC by other products and thorough inter-trip cleaning isessential, since many operators use their tankers for back carriage of a wide range of products, many nonpolymeric.

8. Processing and End Uses

Due to its unique combination of properties PVC is never handled on its own. Instead a complex formulation incorporating several additives is used. A typical base formulation contains: PVC resin, heat stabilizer(s), internal lubricant(s), external lubricant(s), processing aid, and additionally, impact modifier, filler(s), pigment, UV stabilizer, as well as primary and secondary plasticizers (for flexible applications; see → Plasticizers).

PVC is intrinsically unstable because of molecular defects in some of the polymer chains [60], [61] and when subjected to heat they initiate a self-accelerating dehydrochlorination reaction. Stabilizers neutralize the HCl produced and introduce nucleophilic substitution reactions that prevent further degradation [60], [61] (see also → Plastics, Additives, Chap. 4.).

The Ca – Zn stabilizer family is becoming increasingly popular as objections to the use of heavy-metal systems increases. In Europe withdrawal of cadmium stabilizers was planned for the end of 2001, and although it has been shown that use of the highly effective lead stabilizers is safe, adverse publicity will slowly erode their market position.

Table 11 gives a summary of the *heat stabilizers* in regular use.

Table 11. Heat stabilizers for PVC

Stabilizer	Properties
Lead stabilizers	
Lead soaps	Good heat stability
Lead salts	Sulfur staining
Liquid lead complex	Lead phosphites give good light stability
Lead one-pack systems	Kickers for chemical blown foam
Tin stabilizers	
Organo sulfur tin	Excellent heat stability, good clarity; poor light stability
Organo tin carboxylates	Medium heat stability
Excellent light stability	
Barium – zinc stabilizers	
Ba – Zn liquid	Good heat and light stability on high-quality grades; general-purpose stabilizers.
Calcium – zinc stabilizers	
Ca – Zn soaps	Low toxicity
Ca – Zn liquid	Medium heat and light stability
Organic co-stabilizers	
Thiourea derivatives	
Indol derivatives	
Amino crotonic acid esters	Synergistic effects with epoxy plasticizer and/or Ca – Zn, Ba – Zn
Diketones	
Organic phosphites	
Antioxidants	
Epoxidized soybean oil (ESBO)	Synergistic effect with Ca – Zn and Ba – Zn
Epoxy esters	

The polar, highly viscous PVC melt sticks easily to metal walls of extruder barrels, calenders, mills, etc. so an *external lubricant* is employed to assist the smooth passage of the melt. *Internal lubricants* help to reduce melt viscosity and prevent overheating and so help to ensure good color of the final product. *Processing aids* improve the surface appearance of extruded sections and reduce melt defects such as screw memory, where the helical nature of the screw

can be seen as regular ripples in the pipe surface. Many other additives can be used, e.g., *impact modifiers* for bottles and *UV stabilizers* and *fillers* (e.g., TiO_2) for house sidings and window frames to ensure the best possible in-service performance or longevity.

In all extrusion and some other conversion processes the PVC grain is not broken down to its constituent primary particles [62], unlike emulsion PVC processing (see Section 4.3.6). Instead the suspension grains (150 µm) gradually lose their original form by fusion and elongation under the influence of heat, pressure, and shear so that in badly processed PVC grain memory can be detected by optical microscopy of cross sections of extrudates [62]. In practise converters take great care to ensure that the degree of gelation (i.e., lack of grain memory) in the final PVC article is very high. Degree of gelation can be measured by determining the degree of attack by a poor solvent such as acetone or methylene chloride or by measurement of flow pressure to assess the strength and elasticity of partially fused material directly [63].

The enormous subject of processing, including different aspects such as formulations, types of processing equipment, gelation, rheology and mechanical properties is covered more widely in [62], [64–68].

Table 12. Typical end uses for PVC

Application	Rigid PVC	Flexible PVC
Construction	window frames, gutters, pipes, cars, housesiding, ports, roofing	waterproof membranes, cable insulation, roof lining, greenhouses
Domestic	curtain rails, drawer sides, laminates, audio and video tape cases, records	flooring, wallcoverings, shower curtains, leathercloth, hosepipes
Packaging	bottles, blister packs, transparent packs and punnets	cling film
Transport	car seat backs	underseal, roof linings, leathercloth upholstery, wiring insulation, window seals, decorative trim
Medical		oxygen tents, bags and tubing for blood transfusions, drips and dialysis liquids
Clothing	safety equipment	waterproofs for fishermen and emergency services, life-jackets, shoes, wellington boots, aprons and baby pants
Others	floppy-disk covers, credit cards	conveyor belts, inflatables, sports goods, toys

The versatility of PVC can be gauged by the very broad summary of typical end uses given in Table 12.

9. Economic Aspects

Fear of litigation, the need to handle VCM (a known carcinogen), and reduced funding have all but eliminated fundamental research into VCM polymerization. Work on processing of PVC continues, but at a much reduced level. Many companies have scaled down R & D and process development, and raw materials suppliers are now increasingly responsible for technological developments.

Since 1992 the main trend in the PVC industry has been survival by amalgamation. Some well-known companies, including pioneers from the 1940s (e.g., B. F Goodrich/GEON, ICI, and Chisso) have disappeared through aquisition, and many others have formed larger conglomerates to survive. The current feeling in the industry is that a minimum capacity of 500 000 t/a is required for a viable operation.

Irresponsible investment in new plant near the peak of the economic cycle has resulted in over-capacity, weakened prices, lower company profits, and ecomomic caualties, just as it did ca. 1990. Even the staid Japanese PVC industry has seen amalgamations and plant closures since ca. 1995, and the former 15 companies have been reduced to just five major production groups. Changes in the Western European PVC scene are summarized in Figure 16.

Table 13 lists the regional capacities (in 10^3 t/a) of the top ten producers, which represent 48 % of world capacity of 29.8×10^6 t/a. World production in 2000 was 23.3×10^6 t (78 % of capacity).

Since ca. 1990 China has emerged as a major force on the global PVC market, and its tremendous import capacity has had a large impact on the world price/consumption relationship. The eight main Chinese producers use ethylene-based VCM and produce 42 % of total output, while about 70 % of producers manufacture the remaining 42 % using VCM from a calcium carbide based process. Between 1996 and 2000 output of PVC increased on average by 12 % per annum, consumption by 15 %, and imports by 20% per annum.

Table 13. Regional capacities (in 10^3 t/a) of the top ten PVC producers in 2000

Company	S. America	Asia	NAFTA	W. Europe	Total
Shin-Etsu	140	550	1500	520	2710
Formosa		1240	1170		2410
Oxy-Vinyls			2020		2020
Solvay	450	190		1250	1890
EVC				1320	1320
Georgia-Gulf			1270		1270
Atofina				890	890
Borden			690		690
L.G. Chemical		600			600
Taiyo Vinyl		600			600

The top 10 producers represent about 48 % of world capacity of 29.8 million tonnes. World production in 2000 was 23.3 million tonnes (78 % occupacity).

In China

8 main producers use ethylene based VCM production process and produce 58 % of total output.
70 % of production enterprises use VCM from a calcium carbide based process and produce 42 %.
PVC Production Plants:

Consumption of PVC Resin	Number of Plants
> 100 ktpa	8
> 50 ktpa	10
< 50 ktpa	51

Between 1996 and 2000
Domestic output of PVC increased 12 % in yearly average.
Domestic consumption increased 15 % in yearly average.
Imported quantities of PVC increased to 20 % in yearly average.
Consumption per capita

West Germany	27.1
North America	15.1
Western Europe	14.6
South America	3.6
East Europe	3.5
China	3.1
Asia/Pacific	2.6
Africa/Middle East	1.9
India	1.0

Table 14. World supply/demand balance for PVC in 1999 (in 10^3 t/a)

	Capacity	Operating rate, %	Production	Imports	Exports	Consumption
N. America	7480	91	6800	110	370	6540
S. America	1930	81	1560	470	460	1570
W. Europe	6100	89	5410	610	390	5630
Japan	2810	87	2450	10	780	1680
Korea	1150	100	1150	30	380	820
Taiwan	1760	79	1390	160	360	1190
Other Asia	4620	74	3430	1950	450	4930
E. Europe	2290	54	1240	250	760	730
Africa	360	85	310	210	10	510
Middle East	910	88	810	510	340	980
Others	3820	67	2550	1050	1120	2480

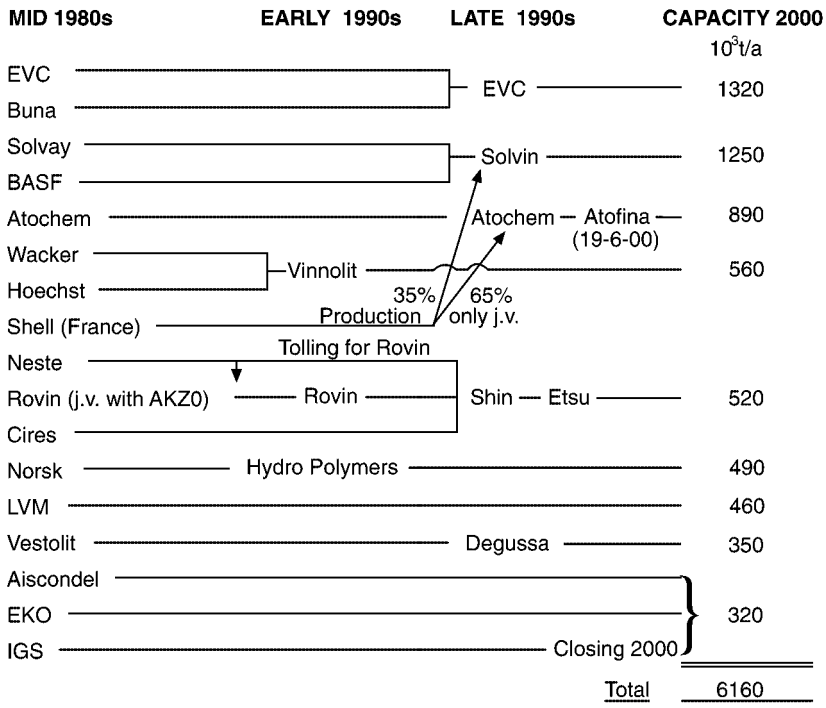

Figure 16. Evolution of the PVC industry in Western Europe
j.v. = joint venture

World PVC production was 23.3×10^6 t/a in 2000, and growth ranged from −3 (Japan) to +8 % (China). Overall long-term growth worldwide is estimated at ca. 4 %. Table 14 summarizes the world supply/demand balance.

10. Toxicology and Occupational Health

Vinyl Chloride Monomer (VCM). Since vinyl chloride boils at − 13 °C it is handled as a compressed liquid. It has a vapor density greater than air so precautions are taken in design not to contain the vapor in restricted areas. Protection of reactors by relief valves or bursting disks and double-valve isolation of all lines containing VCM are basic precautions taken in the design and construction of any PVC plant vessel which contains monomer. The gas is explosive between 3.6 and 25 vol % with air and 12 vol % oxygen is required for ignition [69]. Therefore, a special series of working conditions must be satisfied in the working environment and great care is always taken to purge lines, pumps, valves, etc. before any maintenance work is started. Thorough documentation (e.g., clearance to work certificates) are always used to ensure proper isolation of plant equipment is carried out before any entry. The whole plant is contained within a "red-fence" area where smoking is not permitted and where all electrical equipment is of a gas-tight and "flame-free" construction.

VCM has a narcotic effect at 8 – 12 vol % concentration and can cause death at higher concentrations [69]. For this reason early VCM and PVC plants were designed to avoid build-up of gas above 1000 ppm in the atomosphere. VCM has a pleasant ethereal smell which is first noticed at concentrations of 500 – 2000 ppm.

Exposure to VCM has been shown to lead to two distinct problems. In the mid-1960s it became clear that a number of workers involved in reactor cleaning suffered from a bone con-

dition called acroosteolysis (AOL), which affects mainly the hands and feet. Removal of the worker from exposure to VCM leads to an almost complete recovery. Since improvements to ventilating reactors and the introduction of high-pressure water cleaning the incidence of AOL has ceased [70]. A study of long-term exposure of rats to 30 000 ppm VCM did not reproduce any sign of AOL but they did develop cancers in various sites [71].

Subsequently, an extensive study with a wide range of dose conditions relating to occupational exposure was begun in 1971 at the Institute of Oncology in Bologna [72]. This showed that a very rare liver cancer, angiosarcoma could be formed in rats at levels of 250 – 500 ppm VCM. The first correlation between exposure to VCM and cancer in humans was made in 1973 when three workers at a plant in Louisville, Kentucky were shown to have died from angiosarcoma [73]. Since then other cases have been identified, all of which involved workers exposed to high concentrations of VCM as reactor cleaners or charging operators. As the period between first exposure to VCM and the appearance of angiosarcoma is 20 to 25 years it is likely that the incidence of the disease will soon decrease following improvements made to operation of PVC plants since 1973.

The knowledge that VCM is a human carcinogen has led to the introduction of very stringent controls by governmental and regulatory authorities to limit the exposure of workers and the general public to the monomer. In the United States, OSHA has set an 8-h TWA of 1 ppm VCM. During a shift an employee's exposure must not exceed 5 ppm over a period of 15 min or less [74]. The European regulations, EC directive 78/610, require an annual average of < 3 ppm in the plant's airspace with alarm values of 15 ppm over 1 h, 20 ppm over 20 min, and 30 ppm over 2 min.

Poly(Vinyl Chloride). Strict limits have been set for the quantity of residual VCM in processed PVC articles intended for use in food contact applications.

In 1979 the EC set an upper limit of 1.0 ppm in the article and 10 ppb in the food. In 1986 the FDA suggested levels of ≤ 5 ppb in rigid and ≤ 10 ppb in plasticized PVC articles in food contact applications. To ensure these limits are met, PVC producers aim for these concentrations in the polymer grains themselves so that any further loss during processing ensures even greater safety.

There is little evidence that PVC powder itself causes any significant medical problems but steps are always taken to reduce powder emissions because of its nuisance value.

11. References

General References

1. R. H. Burgess (ed.): *Manufacturing and Processing of PVC,* Applied Science Publishers, London 1982.
2. G. Butters (ed.): *Particulate Nature of PVC,* Applied Science Publishers, London 1982.
3. L. I. Nass, C. S. Heiberger (eds.): *Encyclopaedia of PVC,* 2nd ed., **vols. 1, 2, 3,** Marcel Dekker, New York 1986.
4. G. Matthews: "Vinyl Chloride and Vinyl Acetate Polymers", *Vinyl and Allied Polymers,* **vol. 2,** Iliffe, London 1972.
5. A. Whelan, J. L. Craft (eds.): *Development in PVC Production and Processing,* Applied Science Publishers, London 1977.

Specific References

6. J. D. Cotman, M. F. Gonzales, G. C. Claver, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 1137 – 1164.
7. M. R. Meeks, *Polym. Eng. Sci.* **9** (1969) 141.
8. H. Fikentscher, *Kolloid Z.* **53** (1932) 34.
9. G. A. R. Matthews, R. B. Pearson, *Plastics (London)* **28** (1963) 98.
10. W. F. Maddams in E. D. Owen (ed.): *Degradation and Stabilisation of PVC,* Elsevier Applied Science Publishers, Barking 1984.
11. C. A. Finch (ed.): *Poly(vinyl alcohol), Properties and Applications,* John Wiley & Sons, New York 1973.
12. D. G. Kelsall, G. C. Maitland in K. H. Reichert, W. Geisseler (eds.): *Polymer Reaction Engineering, Influence of Reaction Engineering on Polymer Properties,* Hanser Publishers, München 1983.
13. G. Wieckert, G. Henschel, K. D. Weissenborn, *Angew. Makromol. Chem.* **147** (1987) 1.
14. R. H. Burgess (ed.): *Manufacturing and Processing of PVC,* Applied Science Publishers, London 1982.

15. B. Mariasi, *J. Vinyl Technol.* **8** (1986) no. 1, 20.
16. M. W. Allsopp in R. A. Burgess (ed.): *Manufacturing and Processing of PVC,* Applied Science Publishers, London 1982, p. 152.
17. J. Boissel, N. Fischer, *J. Macromol. Sci. Chem.* **A 11** (1977) no. 7, 1249.
18. M. W. Allsopp in [14] p. 181.
19. N. Fischer, *J. Vinyl Technol.* **6** (1984) no. 1, 35.
20. M. W. Allsopp in [14]chap. 2.
21. I. G. Farben, US 2 068 424, 1931 (H. Fikentscher).
22. W. D. Harkins, *J. Am. Chem. Soc.* **69** (1947) 1428.
23. W. V. Smith, R. H. Ewart, *J. Chem. Phys.* **16** (1948) 592.
24. J. Ulgestad, P. C. Mork, J. O. Aasen, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 2281.
25. E. Peggion, F. Testa, G. Talamini, *Makromol. Chem.* **71** (1964) 173.
26. G. Gatta, G. Benetta, G. P. Talamini, G. Vianello: "Addition and Condensation Polymerization," in R. F. Gould (ed.): *Adv. Chem. Ser.* 1969, no. 91, 158.
27. J. Ugelstad, H. Flogstad, F. K. Hansen, T. Ellingsen, *J. Polym. Sci. Polym. Symp.* **473** (1973) no. 42, 473.
28. ICI, GB 978 875, 1981 (A. Frangou).
29. Rhône-Progil, GB 1 435 425, 1976 (T. Kemp).
30. Rhône-Poulenc, GB 1 503 247, 1977.
31. J. Ugelstad, F. K. Hansen, S. Lange, *Makromol Chem.* **175** (1974) 507.
32. D. E. M. Evans in R. H. Burgess (ed.): *Manufacture and Processing of PVC,* Applied Science Publishers, London 1982, pp. 67 – 75.
33. I. M. Koltoff, I. K. Miller, *J. Am. Chem. Soc.* **73** (1951) 3055.
34. R. G. R. Bacon, *Q. Rev.* **9** (1955) 283.
35. ICI, EP 38 634, 1981.
36. P. Rangnes, O. Palmgren, *J. Polym. Sci., Part C* **33** (1970) 181 – 192.
37. Hüls, DE 1 964 029, 1970.
38. Hüls, BE 857 235, 1970.
39. P. Becher: *Emulsion: Theory and Practice,* Reinhold, New York 1955, p. 173.
40. F. A. Bovey, I. M. Koltoff, A. I. Medalia, A. J. Meehan: *Emulsion Polymerisation,* Interscience, New York 1954.
41. Hoechst, DE 2 429 776, 1980.
42. Hüls, US 4 233 437, 1980.
43. Montedison, DE-OS 3 020 237, 1980.
44. ICI, US 4 283 526, 1981.
45. K. Masters: *Spray Drying Handbook,* Georg Godwin, London 1979.
46. Niro Atomizer, US 4 229 249, 1980 (K. S. Felsvang, O. E. Hansen).
47. K. Kern Sears, J. R. Darby: *The Technology of Plasticizers,* Wiley-Interscience, New York 1982.
48. H. A. Sarvetnick: *Plastisols and Organosols,* Van Nostrand Reinhold, New York 1972.
49. H. Zecha, J. Schneider, K. Wulf, *Kunststoffe* **76** (1986) no. 1, 51.
50. N. Nakajima, E. R. Harrel, *Adv. Polym. Technol.* **6** (1986) no. 4, 409.
51. Congoleum-Nairn, US 3 293 094, 1966.
52. F. Candau, R. H. Ottewill: *Scientific Methods for the Study of Polymer Colloids and their Application,* Kluwer, London 1990.
53. G. Matthews: *Vinyl and Applied Polymers,* vol. 2, Iliffe, London 1972.
54. L. I. Nass: *Encyclopaedia of PVC,* M. Dekker, New York 1976.
55. R. G. Parker, G. Martello in [61]2nd ed., 1986, vol. 1, p. 619.
56. TA Luft 1986.
57. British Plastics Federation, *The Energy Content of Plastics Articles,* London 1986.
58. Company brochure, Solvay et Cie, *PVC and The Environment.*
59. D. J. Brandt, R. S. Guise, B. A. McCoy in L. I. Naas, C. Heiberger (eds.): *Encyclopaedia of PVC,* 2nd ed., vol. 1, Marcel Dekker, New York 1986.
60. T. Hjerberg, E. M. Sörvik in E. D. Owen (ed.): *Degradation and Stabilisation of PVC,* Elsevier Applied Science Publishers, Barking 1984.
61. D. Braun, E. Bazdadea in L. I. Nass, C. A. Heiberger (eds.): *Encyclopaedia of PVC,* 2nd ed., vol. 1, Marcel Dekker, New York 1986.
62. M. W. Allsopp in [14] chap. 8.
63. A. Gonze, *Plastica* **24** (1971) no. 2, 49.
64. D. A. Tester in [14]chap. 9 and 10.
65. G. C. Portingell in G. Butters (ed.): *Particulate Nature of PVC,* Applied Science Publishers, London 1982, pp. 135 – 234.
66. N. L. Perry in [61] vol. 2, p. 601, 1st ed., 1976; vol. 2, p. 1, 2nd ed., 1988.
67. W. Henschel, P. Franz in W. V. Titow (ed.): *PVC Technology,* 4th ed., Elsevier Applied Science Publishers, Barking 1984.
68. E. D. Owen: *Degradation and Stabilisation of PVC,* Elsevier Applied Science Publishers, Barking 1984.
69. J. T. Barr in [61] vol. 1, chap. 5, 2nd ed., 1986.
70. R. H. Burgess in [14] p. 101.
71. P. L. Viola, A. Bigotti, A. Caputo, *Cancer Res.* **31** (1971) 516.

72. C. Maltoni in *Vinyl Chloride Carcinogenicity; Available Scientific Evidence and Control Measures,* Association of Plastics Manufacturers in Europe, Brussels 1986.
73. J. L. Creech, M. N. Johnson, *JOM, J. Occup. Med.* **16** (1974) 150.
74. *Fed. Regist.* **39** (1974) 35890 – 35898.

Polyvinyl Compounds, Others

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

Manfred L. Hallensleben, Universität Hannover, Institut für Makromolekulare Chemie, Hannover, Federal Republic of Germany

1. Poly(Vinyl Alcohol) 1
1.1. Structure and Crystallinity 2
1.2. Production 2
1.3. Properties 4
1.4. Uses, Toxicology, and Economic Aspects 5
2. Poly(Vinyl Acetals) 6
2.1. General Aspects 6
2.2. Poly(Vinyl Formal) 7
2.3. Poly(Vinyl Butyral) 8
2.4. Other Poly(Vinyl Acetals) 11
3. Poly(N-Vinyllactams) and Poly(N-Vinylamines) 11
3.1. Polyvinylpyrrolidone 11
3.2. Polymers of other *N*-Vinyllactams . . 13
3.3. Polyvinylamines 14
4. References 15

1. Poly(Vinyl Alcohol)

In 1924, W. O. Herrmann and W. Haehnel [1] were the first to prepare poly(vinyl alcohol) by saponifying poly(vinyl esters) with stoichiometric amounts of caustic soda solution. At about the same time, work was resumed in the Hoechst Works of the IG Farbenindustrie on the saponification of poly(vinyl ester) derivatives such as poly(vinyl chloroacetate), which F. Klatte and E. Zacharias had begun in 1912 in the Chemische Fabrik Griesheim Elektron.

In 1932, W. O. Herrmann, W. Haehnel, and H. Berg [2] discovered that poly(vinyl alcohol) can be prepared from poly(vinyl esters) by transesterification with absolute alcohols in the presence of catalytic amounts of alkali. The transesterification principle is still used by all poly(vinyl alcohol) producers. Work on the conversion of poly(vinyl acetate) to poly(vinyl alcohol) contributed substantially to the proof of Staudinger's theory of the structure of macromolecules [3].

In the transesterification of poly(vinyl acetate) to poly(vinyl alcohol), products with different acetyl group contents can be obtained depending on production conditions. The range of commercial poly(vinyl alcohols) extends from fully saponified types to products with degrees of hydrolysis of ca. 70 mol %.

In addition to these poly(vinyl alcohols), which dominate the market, saponification products of vinyl acetate – ethylene copolymers are also commercially important [4]. Numerous saponification products of other vinyl acetate copolymers (e.g., with other vinyl esters, longer-chain α-olefins, vinyl ethers, vinyl halides, acrylate and methacrylate esters, and unsaturated acids, nitriles, and amides) are described in [5].

1.1. Structure and Crystallinity

Poly(vinyl alcohol) contains mainly 1,3-diol units. The content of 1,2-diol units in poly(vinyl alcohols) obtained by hydrolysis of poly(vinyl acetates) is $<1-2\,\%$. The content of 1,2-diol sequences can be reduced by lowering the polymerization temperature of vinyl acetate [6] or using other vinyl esters such as vinyl formate [7] or vinyl benzoate [8]. The content of 1,2-diol units influences some properties of poly(vinyl alcohols) (e.g., the degree of swelling of films in water [9]).

Carboxyl and carbonyl end groups are more abundant than initiator end groups in poly(vinyl alcohols) [10]. Poly(vinyl alcohol) is slightly branched as a result of chain transfer reactions occurring during the polymerization of vinyl acetate [11]; → Poly(Vinyl Esters).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a21_743

In partially saponified poly(vinyl alcohols) the acetyl groups can be distributed statistically or blockwise, depending on production conditions. The nature of the acetyl group distribution influences important properties of poly(vinyl alcohols), such as melting point [11], surface tension of aqueous solutions [12], and emulsifying and protective-colloid properties [13].

The stereochemistry of poly(vinyl alcohols) is fixed during the polymerization of vinyl monomers. It is determined from the position and intensity in the ^{1}H NMR spectrum of the hydroxyl proton signal, which consists of three doublets assigned to syndiotactic (S), atactic (A), and isotactic (I) triads of poly(vinyl alcohol) [14, 15]. According to this, poly(vinyl alcohol) obtained by radical polymerization of vinyl acetate followed by hydrolysis is predominantly atactic in structure (24 % I, 44 % A, 32 % S). Stereochemically more uniform poly(vinyl alcohols) are obtained when the initial polymer is poly(vinyl trifluoroacetate) (23 % I, 36 % A, 40 % S) or poly(vinyl *tert*-butyl ether) (55 % I, 32 % A, 13 % S). Poly(vinyl alcohols) of different tacticity have different properties. For example, on acetalization (see Chap. 2), the isotactic sequences react first, and the highest stability toward water is shown by syndiotactic poly(vinyl alcohols) [16].

Poly(vinyl alcohol) is a crystalline polymer (for X-ray structure, see [17]). The degree of crystallinity depends strongly on the structure and previous history of the poly(vinyl alcohol). Acetyl groups and other incorporated groups of any kind lower the crystallizability [18]. The degree of crystallinity of fully saponified poly(vinyl alcohols) is increased by heat treatment, which also lowers their solubility in water [19]. This effect is less marked for poly(vinyl alcohols) that contain acetyl groups.

1.2. Production

Since the monomer of poly(vinyl alcohol), vinyl alcohol, is nonexistent – the tautomeric equilibrium lies on the acetaldehyde side – only indirect methods are available for the production of poly(vinyl alcohol).

Attempts to prepare poly(vinyl alcohol) by polyaldol condensation of acetaldehyde have had little success up to now [20]. Polymerization of the metal vinyl compounds (which are relatively complicated to prepare), followed by saponification, leads in general to poly(vinyl alcohols) of low molecular mass [21, 22].

The most important manufacturing process for poly(vinyl alcohol), which is used worldwide, is the polymerization of vinyl esters or ethers, with subsequent saponification or transesterification. The preferred starting material is vinyl acetate, but derivatives of vinyl acetate (such as vinyl mono- and dichloroacetate, vinyl bromoacetate [23], and vinyl trifluoroacetate [24]), vinyl esters of other carboxylic acids (such as vinyl formate [25], vinyl butyrate [26], and vinyl benzoate [27]) and vinyl ethers (such as benzyl vinyl ether [28]) can also be used. The hydrolysis of poly(vinyl acetate) is easily controllable, and no side reactions occur.

Production Conditions. Since the properties of poly(vinyl alcohols) depend primarily on their molecular mass and residual content of acetyl groups, industrial production processes are optimized for precise control of these two characteristics.

The preferred process for the production of poly(vinyl acetate) for further processing to poly(vinyl alcohol) is polymerization in methanol.

Factors determining the molecular mass of poly(vinyl acetate) and, accordingly, of poly(vinyl alcohol) are the polymerization temperature, vinyl acetate – methanol ratio, and polymerization conversion [29]. Lowering the polymerization temperature, the proportion of methanol, and the conversion increases molecular mass.

In general, poly(vinyl acetate) should be as free as possible of monomeric vinyl acetate because the latter is cosaponified. Any acetaldehyde formed colors the poly(vinyl alcohol) yellow in alkaline saponification as a result of aldehyde resin formation and acetalizes it in acid saponification.

Conversion of poly(vinyl acetate) to poly(vinyl alcohol) can be carried out in solution, suspension, or emulsion with alkaline or acidic catalysts. The preferred process is transesterification in methanol in the presence of catalytic amounts of sodium methoxide, with formation of poly(vinyl alcohol) and methyl acetate.

Hydrolysis conditions influence the structure of the poly(vinyl alcohol) formed. By varying catalyst concentration, reaction temperature, and the reaction time, the content of residual acetyl groups can be adjusted practically at will. The nature of the distribution of the remaining acetyl groups in partially saponified poly(vinyl alcohol) is determined by the choice of catalyst and by the solvent. Thus, in alkaline saponification the residual acetyl groups are distributed largely blockwise [30], and in acidic saponification largely statistically [31].

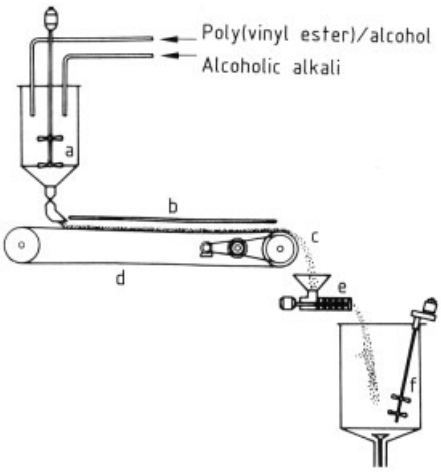

Figure 1. Belt saponification process for poly(vinyl alcohol) [38,39]
a) Mixing vessel; b) Cover plate; c) Discharge; d) Conveyor belt; e) Mill; f) Washing vessel

If methyl acetate [31,32] or benzene [18] is added to methanol in transesterification, poly(vinyl alcohols) with increased block formation by the remaining acetyl groups are obtained. In aqueous solution with acid catalysis, equilibrium conditions occur. The degree of saponification can be adjusted by the addition of different amounts of acetic acid [33]. Partly acetylated poly(vinyl alcohols) prepared by reacetylation of fully saponified products in homogeneous aqueous systems have a completely statistical distribution of the acetyl groups [31].

Poly(vinyl alcohol) is manufactured in both batch and continuous processes. For the continuous polymerization of vinyl acetate in methanol, high molecular masses can be achieved only with relatively low methanol content and low vinyl acetate conversion [34]. Part of the vinyl acetate must therefore be recycled.

In the alkaline hydrolysis of poly(vinyl acetate) in methanol, a highly swollen gel is formed initially. Then, toward the end of saponification this swollen gel contracts and releases solvent (syneresis), becoming a tough gel of poly(vinyl alcohol).

An example of a continuous process is the belt saponification of Shawinigan Chemicals (Fig. 1) [35,36].

Suspension- or solution-polymerized poly(vinyl acetate) is mixed with methanolic alkali in a mixing vessel to give a 30 % methanolic solution, which is fed to a conveyor belt, where it remains until syneresis begins. The gel is ground in a mill, neutralized with acid, washed with methanol, separated from the mother liquor, and dried. The mother liquor, consisting of excess acid, methanol, and methyl acetate, is worked up in further process stages.

Saponification in a kneader is an example of a discontinuous process (Fig. 2) [37,38]. Poly(vinyl acetate) is dissolved in methanol in a kneader, and the amount of sodium methylate in methanol required for the desired degree of saponification is added. The transesterification process at 20 – 30 °C usually lasts about 8 h. Then the mixture of methanol and methyl acetate is driven from the kneader at 60 – 70 °C with a nitrogen stream, and poly(vinyl alcohol) is fed to subsequent processing stages.

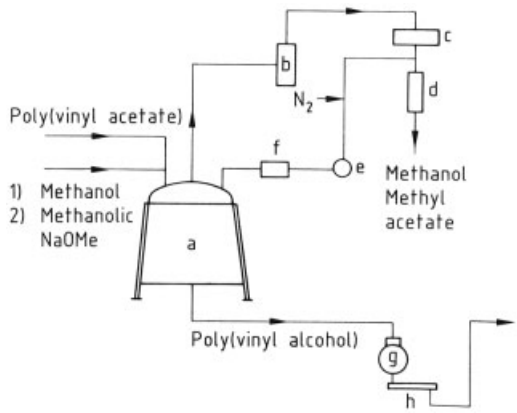

Figure 2. Manufacture of poly(vinyl alcohol) in a kneader [37,40]
a) Kneader; b) Dust separator; c) Trap; d) Cooler; e) Blower; f) Heater; g) Mill; h) Sieve.

Subsequent process stages [39,40] include separation of vinyl acetate from partly reacted solution polymers in stripping columns, separation of the saponification mother liquor into methyl acetate and methanol, hydrolytic cleavage of methyl acetate with subsequent distillation of methanol and acetic acid, and drying of poly(vinyl alcohol), which can be carried out batchwise or continuously.

Trade Names. Poly(vinyl alcohol) is supplied by manufacturers under different trade names: Mowiol (Hoechst, Germany), Polyviol (Wacker Chemie, Germany), Rhodoviol (Rhône-Poulenc, France), Alcotex (Revertex, England), Polivinol (Rhodiatoce, Italy), Denka Poval (Denki Kagaku Kogyo, Japan), Gohsenol (Nippon Gohsei, Japan), Kurashiki Poval (Kuraray, Japan), Shinetsu Poval (Shinetsu Chem. Ind., Japan), Unitika Poval (Unitika, Japan), Elvanol (Du Pont, US), Gelvatol (Shawinigan Resins, US), Lemol (Borden, US).

Storage and transport of poly(vinyl alcohol) require no special precautions: the polymers must merely be protected from the effects of moisture and heat.

Wastewater Problems. Poly(vinyl alcohol) is degradable in sewage treatment plants with an adapted biological clarification stage. However, when discharged into receiving waters it is degraded only with difficulty, as shown by the 5-d BOD values. Tests showed that fish (guppies) are not harmed, even at a poly(vinyl alcohol) concentration of 500 mg/L of water [41].

1.3. Properties

Physical Properties. Normal commercial poly(vinyl alcohols) are white to yellowish and are supplied as powder and granules. The main factors influencing physical properties are the degrees of polymerization and hydrolysis. Most commercial products fall into two main groups: those with a degree of hydrolysis of about 98 mol % of the acetyl groups, and those with a degree of hydrolysis of 87 – 89 mol %. Both groups include products with degrees of polymerization from ca. 500 to 2500. Data such as the viscosity of the aqueous poly(vinyl alcohol) solutions, the molecular mass of the precursor poly(vinyl acetates), and the degree of hydrolysis or ester number are frequently used by manufacturers of poly(vinyl alcohol) to characterize their products.

The melting point and glass transition temperature depend not only on the content and distribution of acetyl groups but also on the tacticity and water content. For fully saponified poly(vinyl alcohols) a melting point of 228 °C and a glass transition temperature of 85 °C are reported [42]. Prolonged heating of dry poly(vinyl alcohol) above 110 °C can lead to evolution of gaseous cleavage products.

Poly(vinyl alcohol) Solutions. Water is the most important solvent for poly(vinyl alcohol). Certain polar solvents such as diethylenetriamine [43], dimethyl sulfoxide [44], formamide, dimethylformamide, and hexamethylphosphoric triamide [45] are relatively good solvents. The rate of dissolution in water increases with decreasing degree of polymerization and degree of hydrolysis. The degree of polymerization has a greater effect with fully saponified than with partially saponified poly(vinyl alcohols). Poly(vinyl alcohols) with degree of hydrolysis below 88 % dissolve better in water at low than at elevated temperature [46].

Poly(vinyl alcohol) solutions are obtained by stirring poly(vinyl alcohol) into water at ca. 90 °C. Industrially they are produced in indirectly heated dissolving tanks or by introducing steam into poly(vinyl alcohol) suspensions. For prolonged storage, solutions must be stabilized against microbial attack by adding preservatives.

The viscosity of aqueous solutions depends on the degrees of polymerization and hydrolysis of poly(vinyl alcohols) as well as on the concentration and temperature. Fully saponified poly(vinyl alcohols) yield more viscous solutions than partially saponified poly(vinyl alcohols) at the same degree of polymerization [46]. Concentrated solutions of highly saponified poly(vinyl alcohols) undergo a viscosity increase on prolonged standing, which can be reversed by heating.

Many inorganic salts, especially sulfates and phosphates, precipitate poly(vinyl alcohols) from aqueous solution [42]. Boric acid and borax act as thickeners [47]. The conductivity of aqueous poly(vinyl alcohol) solutions depends on their electrolyte content, which is carried over

from the precipitation process: for 10 % solutions it is ca. 0.1 – 0.7 mS/cm [48].

Chemical Properties. Poly(vinyl alcohol) undergoes the chemical reactions expected of a polyglycol with secondary hydroxyl groups. The most important are the formation of esters, ethers, and acetals [42,49,50]. Only a few typical reactions are mentioned here. Examples of inorganic esters are poly(vinyl nitrates), phosphates, and sulfates. Organic esters can be obtained by reaction with acid anhydrides or acid chlorides. The cinnamate esters are important for the production of light-sensitive polymers [51]. Many poly(vinyl alcohol) ethers have been described. The products of the reaction of poly(vinyl alcohol) with ethylene oxide are suitable for the production of films with good solubility in cold water. Of the reaction products of poly(vinyl alcohol) and aldehydes, poly(vinyl butyrals) have attained importance as raw materials in the manufacture of safety glass.

1.4. Uses, Toxicology, and Economic Aspects

Protective Colloid in Emulsion and Suspension Polymerization. Poly(vinyl alcohol) has proved outstandingly effective as a protective colloid for the production of plastic dispersions, especially of poly(vinyl acetate) homo- and copolymers, and as a suspension stabilizer. By selection of suitable types and combination with emulsifiers, dispersions with extremely varied properties can be produced, most of which are used in adhesives [31,52–55].

Adhesives. Poly(vinyl alcohol) is used in aqueous solution, alone or in combination with plastic dispersions, for packaging adhesives, cigarette adhesives, and moistenable gumming [56].

Sizes. The most important field of application in the textile industry is the sizing sector. Poly(vinyl alcohol) is an excellent raw material for the sizing of stable fiber yarns and filaments [56–59].

Paper Industry. Applications in the paper industry are surface sizing and the production of coated papers and special papers with specific barrier properties. Poly(vinyl alcohol) has excellent pigment binding power and is a good carrier for optical brighteners [60–63].

Other Uses. Poly(vinyl alcohol) is used for bonding nonwoven fabrics of all kinds, especially glass fiber; in temporary bonding agents for special ceramics; in secondary brighteners in electroplating and electroforming; for the production of protective lacquers and decorator sizing and solvent-resistant dipped products, such as gloves and aprons; for the modification of surface coating formulations; as a release agent for cast resin moldings; in the production of photoresists for the printing industry and of highly absorbent sponges; and in binders and thickeners for cosmetics.

Poly(vinyl alcohol) is also very suitable for the thermoplastic production of packaging films with graded water solubility. Good plasticizers for poly(vinyl alcohol) are ethylene glycol, glycerol, trimethylolpropane, neopentyl glycol, poly(ethylene glycols) with molecular masses up to ca. 400, and ethoxylated phosphate esters. Films of plasticized poly(vinyl alcohols) are practically impermeable to nitrogen, oxygen, and carbon dioxide [64]. Saponification products of vinyl acetate copolymers (e.g., with ethylene or long-chain α-olefins) are also used in the film sector [60,64].

Toxicology. In testing poly(vinyl alcohol) for toxicity and for compatibility with skin and mucous membranes, no negative effects were found in animals (LD_{50}; 90-d feeding test on rats; FDA patch test; dermal application) [48].

Economic Aspects. The worldwide production capacities (in 10^3 t/a) are as follows:

World	650
Japan	239
United States	127
Western Europe	60

2. Poly(Vinyl Acetals)

The first poly(vinyl acetal) was obtained by W. Haehnel and W. O. Herrmann in 1924 by

the reaction of poly(vinyl alcohol) with benzaldehyde in concentrated hydrochloric acid [65]. In the following years, more methods were developed, in which the reactions were carried out in concentrated acids and anhydrous solvents. In 1929, H. Hopf and E. Kühn succeeded in acetalizing aqueous poly(vinyl alcohol) solutions in the presence of dilute acid [66]. Shortly afterward G. Kränzlein et al. reacted poly(vinyl acetate) with aldehydes with simultaneous saponification and acetalization to give poly(vinyl acetals) [67].

2.1. General Aspects

Since most aldehydes and many ketones react with vinyl alcohol homo- and copolymers, a number of poly(vinyl acetals) can be obtained. However, at present, only poly(vinyl butyral) and poly(vinyl formal) are commercially important. All other poly(vinyl acetals) have until now been used only in isolated cases in specialty applications.

The main reason for the limited use of poly(vinyl acetals) is that their manufacture, which requires several stages, is expensive in comparison with other plastics. Their use is restricted also by their sensitivity to acids and oxidation. However, they have excellent mechanical and adhesive properties, and good processibility.

Production and Properties. The intermediate for all poly(vinyl acetal) syntheses is poly(vinyl alcohol). In principle, the way in which poly(vinyl alcohol) is produced, is immaterial. The only industrially important method is radical polymerization of vinyl acetate. Poly(vinyl acetate) is then saponified with catalysis by alkali to poly(vinyl alcohol), which is acetalized in the presence of mineral acids (two-stage process). Poly(vinyl acetate) can also be saponified in acid, and the saponification combined with acetalization, in a one-process operation (single-stage process).

Important properties of poly(vinyl acetals), above all the molecular mass and molecular mass distribution, are fixed during polymerization of vinyl acetate. A good starting material for acetalization is poly(vinyl acetate) with a narrow molecular mass distribution, such as is obtained by solution polymerization of vinyl acetate with incomplete conversion.

The poly(vinyl alcohols) used to produce poly(vinyl acetals) always contain unsaponified vinyl acetate units (generally 1 – 8 wt %). This figure is usually unchanged by acetalization in the two-stage process. Poly(vinyl acetals) produced in the single-stage process often have considerably higher acetate content since the acid saponification of the acetate groups is relatively slow.

Poly(vinyl acetals) also differ in the degree of acetalization, which can vary within wide limits and strongly influences the polymer properties, especially strength, solubility, and adhesiveness. According to Flory [68], for statistical reasons, no more than 81.6 % of the alcohol groups in poly(vinyl alcohol) can react with aldehydes. The acetalization reaction is assumed to be exclusively intramolecular (i.e., it proceeds without cross-linking). In fact, only in the reaction of poly(vinyl alcohol) with formaldehyde can cross-linking be produced relatively easily, although even in this case, intramolecular acetalization is preferred. In the reaction with higher aldehydes, intermolecular acetalization is much more difficult or impossible [69].

Thus poly(vinyl acetals) are terpolymers of the formula

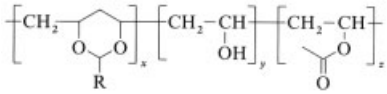

where x, y, and z can vary over wide ranges, and the alcohol, acetal, and acetate groups are generally distributed statistically. Although indications are that poly(vinyl acetals) with block structures can be prepared, little is known about such structures.

In the reaction of formaldehyde with poly(vinyl alcohol), isotactic vinyl alcohol units, which give *cis* rings on acetalization, react more readily than syndiotactic vinyl alcohol units, which give more highly strained *trans* rings [70]. Therefore, in the acetalization of poly(vinyl alcohol), the isotactic vinyl alcohol units react first while the syndiotactic units remain. Individual acetalization processes appear to differ, however, so poly(vinyl butyrals) of the same

empirical composition can vary not only in the blocklike or statistical arrangement of their repeating units but also in the tacticity of the residual vinyl alcohol sequences. The properties of poly(vinyl acetals) are affected as a result. For example, syndiotactic vinyl alcohol polymers exhibit pronounced intermolecular hydrogen bonding, which is the reason for their high strength and poor solubility.

Stabilization. Like all acetals, poly(vinyl acetals) are sensitive to acid hydrolysis. If oxygen is also present, peroxidic oxidation products of the aldehyde that splits off can cause cleavage of the molecular chain and increase the acidity of the polymer so that the decomposition accelerates autocatalytically.

Poly(vinyl acetals) can be stabilized effectively by addition of alkali [71]. However, poly(vinyl acetals) are also sensitive to oxidation in neutral or alkaline medium. Therefore, stabilization with antioxidants, in addition to alkali, is advisable. Of the numerous antioxidants proposed [72], phenols have proved effective [73], whereas phosphites [74] or triazole derivatives [75] are effective as costabilizers.

Unprocessed poly(vinyl butyral) powders cannot be stored indefinitely, even in the stabilized state. Manufacturers therefore recommend a maximum storage period of 6 – 12 months.

Raw Materials. The only significant raw materials are poly(vinyl acetate) or poly(vinyl alcohol) and formaldehyde or *n*-butyraldehyde. The polymer starting materials should be as pure and colorless as possible. Commercial products can generally be used without further purification.

The poly(vinyl alcohol) must not contain insoluble components. The aqueous poly(vinyl alcohol) solution must therefore be filtered before acetalization. Although technical-grade formaldehyde can be used directly for acetalization, butyraldehyde should be purified by distillation and stored under nitrogen. It must not contain more than 2 % impurities, especially other aldehydes and acids. The inorganic acids used for acetalization should be free of heavy metals.

2.2. Poly(Vinyl Formal)

The importance of poly(vinyl formal) is limited. Its production probably makes up only 10 % of total poly(vinyl acetal) production, if the acetalized fibers produced from poly(vinyl alcohol) are not considered.

Production. Since the commercially important poly(vinyl formals) contain a relatively large number of vinyl acetate units, a single-stage process for manufacturing directly from poly(vinyl acetate) is particularly suitable.

In the processes developed in the United States [76, 77], 1000 parts by weight of poly(vinyl acetate) are dissolved in 2000 parts of pure acetic acid; a mixture of 875 parts of 30 % formalin and 70 parts of concentrated sulfuric acid is added; and the mixture is stirred for 6 h at 85 °C. After the mixture is cooled to 50 °C and the sulfuric acid is neutralized, poly(vinyl formal) is precipitated with water and washed. A pale yellow powder is obtained. Instead of acetic acid, Hoechst [78] recommends the use of ethyl acetate, optionally mixed with methanol.

By altering the proportions of starting materials and varying reaction conditions, numerous types of poly(vinyl formal) can easily be produced. For the reaction mechanism, see [79, 80].

Commercial Products. Poly(vinyl formals) are marketed by Wacker-Chemie under the name Pioloform F and by Monsanto under the name Formvar.

Properties [81, 82]. Commercial poly(vinyl formal) powders in contrast to formalized fibers from poly(vinyl alcohol), are not cross-linked. Individual types differ in molecular mass, which is usually between 15 000 and 45 000, and in the proportion of vinyl acetate units, which is usually 10 – 30 wt % of the polymer. The proportion of vinyl alcohol units is between 5 and 9 wt %.

The viscosity of polymer solutions increases not only with increasing molecular mass but also with increasing poly(vinyl alcohol) content. At the same time the glass transition temperature increases from ca. 75 to 95 °C. The softening temperature is between 140 and 170 °C, depending on molecular mass. The tensile strength of the polymers is up to 70 MPa with a fracture strain

of 30 – 70 %. Depending on the type of stabilization, polymers can withstand temperatures of 120 – 150 °C for up to 1 h. The solubility in polar solvents increases with increasing proportion of poly(vinyl acetate). Poly(vinyl formals) are insoluble in hydrocarbons and alcohols, but soluble in mixtures of alcohols and aromatic hydrocarbons.

Uses. The most important use of poly(vinyl formal) is in the production of coatings for electrical wires [81], in which the poly(vinyl formal) is combined with phenolic resins. The coatings, which are cured at elevated temperature, are insoluble, tough, and resistant to abrasion and heat, and also have very good insulating properties [83, 84]. Its use in adhesive joints [85, 86] and metal coating [81] is less important.

The use of poly(vinyl formal) for foamed plastics has frequently been proposed. To produce such foamed plastics, an aqueous solution of poly(vinyl alcohol) is foamed and formalized in the presence of a water-soluble substance such as starch or sodium chloride. After washing out, an insoluble soft and spongy material is obtained [87, 88].

Toxicity. Poly(vinyl formals) are nontoxic. Poly(vinyl formals) are therefore permitted for coatings that come in contact with food [81, 82].

2.3. Poly(Vinyl Butyral)

Poly(vinyl butyrals) are commercially the most important poly(vinyl acetals). They are produced in numerous types. In 1990, world consumption of poly(vinyl butyral) was ca. 80 000 t.

Production. For the commercial production of poly(vinyl butyrals), only those processes that start from poly(vinyl alcohol) isolated before acetalization are significant. Two important processes exist. One begins with aqueous poly(vinyl alcohol) solutions, which are reacted with *n*-butyraldehyde in the presence of mineral acid. After passage through a highly viscous intermediate stage, which the stirrer must be able to cope with, the polymer precipitates as fine grains. The reaction is then completed heterogeneously. The difficulty with this process is to achieve a sufficiently homogeneous polymer.

In the second process, an alcoholic poly(vinyl alcohol) suspension is acetalized. A homogeneous poly(vinyl butyral) solution is obtained that can be freed from insolubles by filtration. The problems of this process are the costly precipitation of the polymer and the need to recover the solvent.

The effect of production conditions on polymer properties is described in [89].

Acetalization in Water.

In the IG Farbenindustrie process [90], 40 kg of poly(vinyl acetate) is saponified in a kneader with 60 kg of methanol and 3.2 L of 10 % methanolic sodium hydroxide solution. The poly(vinyl alcohol) is centrifuged from the excess methanol and methyl acetate and, still wet with solvent, is dissolved in 230 kg of water. The solution is cooled to 0 °C, and 12 kg of butyraldehyde also cooled to 0 °C, is added; the mixture is stirred for 10 min, and 50 kg of 18 % hydrochloric acid is added in one batch. After 2 min a viscous mass forms, which after another 2 min thins again. The mixture is stirred for 30 min at 0 °C, then slowly heated to 40 °C or higher, depending on whether a finer or coarser powder is desired. After cooling, the product is filtered by suction, washed with water until neutral, and dried at 40 °C.

In the Du Pont process [91], poly(vinyl alcohol) is reacted with butyraldehyde in water at 90 °C in the presence of sulfuric acid and *p*-toluenesulfonic acid. Agglomeration of the polymer that precipitates is prevented by a small amount of sodium lauryl sulfate.

Acetalization in Alcohols.

In the Hoechst process [92], 250 parts by weight of poly(vinyl acetate) is dissolved in 750 parts of methanol and 90 % saponified with one part of sodium hydroxide in 10 parts of methanol under reflux. Then, 84 parts of butyraldehyde and 35 parts of 30 % hydrochloric acid are added, and acetalization is carried out for 6 h at 60 °C. The solution is filtered, and the polymer is precipitated in fine-grained form by adding ice-cold distilled water.

Shawinigan Resins describes the saponification of poly(vinyl acetate) in ethanol in the presence of sulfuric acid for 6 – 7 h at 75 °C [93]. The separated poly(vinyl alcohol) is suspended moist in ethanol, and after sulfuric acid has been added, it is acetalized with butyraldehyde for 1 – 2 h at the boiling point.

Commercial Products. Poly(vinyl butyral) is sold as a powder. The most important manufacturers are Hoechst (Mowital B), Wacker-Chemie (Pioloform B), Monsanto (Butvar), Rhône-Poulenc (Rhovinal B), Sekisui Chem. Ind. (S-lec), and Union Carbide (Vinylite).

Properties. The interest in poly(vinyl butyrals) is based on their structure, which consists of reactive, markedly hydrophobic and hydrophilic polymer units, the ratio between the two groups being variable within wide limits.

The hydrophobic vinyl butyral units give the polymer good thermoplastic processibility, solubility in numerous solvents, elasticity, and toughness, as well as compatibility with many resins and plasticizers. The hydrophilic vinyl alcohol units are responsible for the high adhesion to inorganic materials such as glass and metals, high strength, cross-linking ability, and anticorrosive action.

Poly(vinyl butyrals) are amorphous and transparent. At low degrees of acetalization they are water soluble. Like vinyl alcohol – vinyl acetate copolymers, they are good dispersants [94]. Also like these copolymers, their solubility in water decreases with increasing temperature [95].

More highly acetalized poly(vinyl butyrals) with ca. 1 – 3 wt % of vinyl acetate units are commercially more important. They differ considerably, however, in their molecular mass and degree of acetalization.

Table 1 lists the composition and some average properties of various poly(vinyl butyrals), derived from product information pamphlets [81,96,97]. Their accuracy is, however, sufficient to show the effect of molecular mass on rheological, thermal, and mechanical properties. In addition, the effect of the proportion of vinyl alcohol units in the polymer is recognizable: with increasing vinyl alcohol content the intermolecular forces increase, and the strength and modulus of elasticity rise as do the softening point and the glass transition temperature.

In comparison with poly(vinyl formal), poly(vinyl butyrals) have a lower strength and lower modulus of elasticity. They are softer and more extensible. Their density is ca. 1.1 g/cm^3. All poly(vinyl butyrals) are soluble in $C_2 - C_4$ alcohols, glycol ethers, and alcohol – aromatic mixtures, but insoluble in aliphatic hydrocarbons.

Their compatibility with plasticizers such as esters of phthalic, sebacic, ricinoleic, and citric acids is good [98]. Other important plasticizers are ethers and esters of ethylene glycol, and esters of its oligomers with butyric or hexanoic acid. Dihexyl adipate is also used as a plasticizer for poly(vinyl butyral) [99].

Uses. The most important uses of poly(vinyl butyral) are in the manufacture of laminated glass sheets and paints.

Laminated Glass Sheets. This field of application, in which poly(vinyl butyral) is almost without competition, has become very important.

The only industrial process for producing poly(vinyl butyral) sheets is extrusion. Poly(vinyl butyral) powder is mixed with ca. 30 wt % of a plasticizer and extruded above 150 °C [100,101]. The most frequently used plasticizer is the di-2-ethylbutyric acid ester of triethylene glycol (Flexol 3 GH, Union Carbide).

A high-strength sheet is required to achieve high puncture resistance of the glass composite. Only high molecular mass poly(vinyl butyral) with ca. 17 – 22 wt % of vinyl alcohol units can meet this requirement.

In the manufacture of automobile windshields a sheet thickness of ca. 0.8 mm is generally used, whereas for structural or bullet-proof glass, thinner sheets suffice. The surface of the sheet is roughened by various techniques [100]. This reduces the tendency to stick together and facilitates deaeration during production of the glass composite. To completely prevent sticking, the sheet is either dusted with sodium hydrogencarbonate, which must be washed off before producing the composite, or cooled below 10 °C.

In manufacturing the composite [102], sheets are laid between two or more sheets of glass and subjected in an autoclave to a special pressure and temperature program [103].

Accurately defined adhesion of the plastic sheet to the glass is particularly important [104]. Upon impact of a foreign body on the composite, the glass breaks first. If the sheet adheres very firmly to the glass, it is at once overstretched and tears when the glass breaks, and the composite

Table 1. Properties of poly(vinyl butyrals) with various percentages of vinyl alcohol units and various molecular masses

Vinyl alcohol units, wt %	Molecular mass	Viscosity [a], mPa · s	Water absorption [b], %	Solubility [c], in			Tensile strength [d], N/mm^2	Elongation at break, %	Softening point [e], °C	Modulus of elasticity [f], N/mm^2	Glass temperature [g], °C
				Acetone	Methylene chloride	Methanol					
25	30 000	25	13	i	i	s					
20	30 000	25	8	p	p	s	44	110	107	2200	50
25	45 000	75	13	i	i	s					
20	45 000	60	6	p	p	s	45	100	125	2200	53
13	45 000	55	4	p	p	p	44	110	112	2000	48
20	100 000	200	5	i	p	s	50	75	135	2300	55

[a] 5 % solution in butanol, 20 °C. Determination in the Höppler viscometer according to DIN 35 015. [b] 0.1 mm film, 24 h in H_2O at 20 °C. [c] i = insoluble, s = soluble, p = partially soluble. [d] According to ASTM D 63858 T. [e] According to ASTM B 569-59. [f] According to ASTM D 63858 T. [g] According to ASTM D 1043-51.

is punctured. The adhesion must therefore be reduced so that when the glass breaks the sheet detaches slightly and a larger area can then absorb the impact elastically. However, the adhesion must not be so low that splinters of glass become completely detached from the sheet.

The adhesion of poly(vinyl butyral) sheets to glass is normally too high, and numerous substances are known that cause a defined reduction. These include particularly salts, of which various potassium salts such as potassium acetate are especially important [105]. Lecithin is also highly effective [106].

Plasticized poly(vinyl butyral) sheets for the production of automobile windshields and structural or bullet-proof glass are sold, for example, by the following companies: Du Pont

(Butacite), Dynamit Nobel (Trosifol), Fatra Napajedla CSFR (Butaflex), Monsanto (Saflex), and Sekisui Chem. Ind. (S-lec Sheet). Monsanto and Du Pont have the largest market shares.

Paints. Because of their elasticity and good adhesion, poly(vinyl butyrals) are important raw materials for paints, particularly for metal coatings. They are frequently combined with phenolic, alkyd, epoxy, urea, or melamine resins. The low-acetalized and lower molecular mass types, in particular, show good compatibility with these resins. When the coatings are baked, cross-linking occurs [107].

Poly(vinyl butyrals) are outstandingly effective as primers with high corrosion protection. In wash primers they are used in dissolved form (e.g., together with pigments, chromium salts, and phosphoric acid). Shop primers, which are used especially in shipbuilding for priming metal plates, must have extended durability. In addition to poly(vinyl butyral) and solvents, they contain, for example, a phenolic resin, zinc chromate, iron oxide, and phosphoric acid. The coatings are abrasion resistant and adhere very well. The coated plates can be welded easily [96, 108].

In addition, poly(vinyl butyrals) are used as binders in special paints (e.g., for leather or plastics), as well as in printing inks [109].

Poly(Vinyl Butyral) Dispersions. When poly(vinyl butyral) is used dispersed in the aqueous phase, pollution of the environment by solvents is avoided. Nevertheless, such dispersions have minor importance because of their high price and their relatively high emulsifier content, which impairs the adhesive properties and water resistance of coatings. Poly(vinyl butyral) dispersions can be produced by acetalization of aqueous poly(vinyl alcohol) solutions in the presence of an emulsifier or by emulsification of poly(vinyl butyral). A plasticizer-containing dispersion is sold by Monsanto under the name Butavar Dispersion BR for use as a textile coating [81].

Other uses for which poly(vinyl butyrals) have been proposed are adhesives [107]; temporary, strippable coatings [110]; powder coatings [111]; and blends with other polymers.

Toxicity. Poly(vinyl butyrals), like poly(vinyl formals), are nontoxic and permitted by law in packaging for foodstuffs [81, 96].

2.4. Other Poly(Vinyl Acetals)

The numerous aldehydes that have been reacted with poly(vinyl alcohol) include furfural, benzaldehyde, crotonaldehyde, halogenated aldehydes, cyanopropionaldehyde, nitrobenzaldehyde, and glyoxylic acid. Nothing is known about the economic importance of the poly(vinyl acetals) prepared from them. Poly(vinyl acetaldehyde acetal), on the other hand, temporarily played a role in Germany and Canada, but was displaced by poly(vinyl butyral) [112].

Studies of poly(vinyl acetals) from aliphatic C_2-C_8 aldehydes with different acetal contents in the presence and absence of plasticizers [113] have shown that the glass transition temperature decreases with increasing chain length of the aldehyde component. At the same time the poly(vinyl acetals) become softer and more hydrophobic.

Of the poly(vinyl acetals) from substituted aldehydes, the sodium salt of poly(vinyl butyral sulfonic acid) is produced commercially (Afrilan, Hoechst). The substance is water soluble and is important in sizing. It is produced by passing sulfur dioxide into an aqueous solution of poly(vinyl alcohol) in the presence of crotonaldehyde at 0 °C and then heating the solution. Sulfurous acid adds to crotonaldehyde simultaneously with acetalization, and the sulfonic acid formed acts as catalyst for the acetalization, so that further addition of acid is unnecessary [114].

3. Poly(N-Vinyllactams) and Poly(N-Vinylamines)

3.1. Polyvinylpyrrolidone

When *N*-vinylpyrrolidone and other *N*-vinyl compounds became commercially available by Reppe vinylation of lactams in about 1940, the polymers were soon produced on an industrial scale. Polyvinylpyrrolidone is used as an auxiliary in cosmetics and pharmaceuticals.

N-vinylpyrrolidone

Production. Only radical polymerization is important. *N*-vinylpyrrolidone (*mp* 13.9 °C, *bp* at 1.3 kPa 90–93 °C) is usually polymerized batchwise in aqueous solution with hydrogen peroxide as initiator and regulator. The average molecular mass can be adjusted by varying the quantity of hydrogen peroxide: with 0.05 % hydrogen peroxide, molecular masses of ca. 750 000 are obtained; with 3 %, ca. 25 000. With increasing quantity of hydrogen peroxide the polymerization rate increases. The rate also increases with increasing concentration of *N*-vinylpyrrolidone in the aqueous solution, reaches a maximum at 40–60 % monomer content, and then decreases. Due to the viscosity of the final polymer solution the concentration is adjusted within a range between 10 % at high molecular mass and 60 % at low. To control polymerization, the monomer solution must be added slowly to the polymerization mixture with cooling. Polymerization is complete after 6–12 h. The ready saponifiability of *N*-vinylpyrrolidone in aqueous acidic solution is a problem. Aqueous solutions containing monomeric *N*-vinylpyrrolidone should therefore always be neutral to basic. The pH value is best controlled with ammonia, which also activates the polymerization reaction [115].

Polyvinylpyrrolidone can be obtained as a white powder from polymer solutions by drum or spray drying. Because it is highly hygroscopic it must be protected from moisture during extended storage.

Polymerization can also be carried out in organic solvents. The initiators preferably used then are organic peroxides or azocompounds, preferably alkyl hydroperoxides, dialkyl peroxides, and peroxy esters [116]. The polymerization temperature is chosen according to the decomposition temperature of the initiator, but the polymerization rate also depends on the nature of the solvent [117, 118].

Bulk polymerization [119] and polymerization in suspension [120] have no commercial importance.

N-Vinylpyrrolidone can be copolymerized with many other vinyl compounds such as acrylic acid [119, 121, 122], methacrylates esters [115, 123], and methacrylamide [124]; for the *Q* and *e* values, see [125]. The most important are the polymers with vinyl esters [124, 126, 127],

whose properties lie between those of the pure polymers.

When *N*-vinylpyrrolidone is heated for several hours with hydroxides and alkoxides of the alkali and alkaline earth metals [128], an insoluble polymer slightly swellable by water is formed in a spontaneous reaction. This polymer is also formed on heating *N*-vinylpyrrolidone with divinyl compounds in the absence of atmospheric oxygen [129] and is known as popcorn polymer [130].

Commercial Products. Polyvinylpyrrolidones with an average molecular mass of 2500 to 750 000 are on the market.

Trade names are Kollidon, Luviskol, Albigen A, and Divergan (BASF); PVP and Plasdone (General Aniline and Film Corp.); Collacral and Luviskol VA (copolymers with vinyl esters, BASF); PVP/VA (copolymers with vinyl acetate, General Aniline and Film Corp.); Polyclar and Polyplasdone XL (insoluble polyvinylpyrrolidone, General Aniline and Film Corp.).

Properties. Polyvinylpyrrolidone is a white powder. The glass transition temperature depends on the water content; extrapolation to 0 % water content gives a glass transition temperature of 175 °C. The mean molecular mass can be adjusted in the range of 1000 to 10^6.

Polyvinylpyrrolidone is highly hygroscopic, absorbing ca. 30 % water at 60 % humidity. It dissolves readily in water, and aqueous solutions are stable to electrolyte addition. The viscosity of the solution decreases with increasing temperature and with strong shearing [131].

Polyvinylpyrrolidone dissolves readily in polar solvents such as alcohols, amines, acids, and chlorinated hydrocarbons. It is insoluble in esters, ethers, ketones, and hydrocarbons. Completely anhydrous polyvinylpyrrolidone is also soluble in toluene.

Polyvinylpyrrolidone can be precipitated from aqueous solution by dropwise addition to acetone and from solution in organic solvents by addition of diethyl ether.

Polyvinylpyrrolidone has good compatibility with numerous film formers, water-soluble binders, and plasticizers. On solution casting, a clear, high-gloss, hard film is formed.

The polymer is chemically inert. The lactam group is saponified only by the action of concentrated acids, with the formation of poly[vinyl(γ-amino)butyric acid] [132].

With some substances such as iodine, polyphenols, tannin, dyes, and toxins, polyvinylpyrrolidone forms complexes [133]. Frequently, only a reversible association occurs. The protective colloid action is utilized in dispersions and in suspension polymerization.

For qualitative analysis [134], aqueous ammonium cobalt thiocyanate solutions are added to an acidic aqueous polyvinylpyrrolidone solution. The flocculation of a bright blue complex indicates polyvinylpyrrolidone. The IR spectrum is also suitable for identification. To determine concentration, the brownish red polyvinylpyrrolidone – iodine complex can be analyzed colorimetrically.

Uses. In cosmetics, polyvinylpyrrolidone and some copolymers, especially with vinyl acetate, are used as film formers in setting lotions and hairsprays. Fairly high molecular mass polyvinylpyrrolidone is used as a thickening agent and protective colloid in cosmetic emulsions. Its use as a blood plasma substitute has been abandoned.

In pharmacy, polyvinylpyrrolidone has a wide spectrum of applications: as a solubilizer, as a crystallization retarder, for detoxification, for reducing the irritant action and toxicity of certain substances, as a tablet binding and coating agent, as a suspension stabilizer, and as a dispersant for pigments in tablet-coating suspensions. The polyvinylpyrrolidone (PVP) complex with iodine, PVP-iodine, is used as a disinfectant and does not show the side effects of elementary iodine [135]. The insoluble and slightly water-swellable popcorn polymer is used as a tablet-disintegrating agent. Polyvinylpyrrolidone and PVP-iodine are described in a series of pharmacopoeias (e.g., in the U.S.).

Since polyvinylpyrrolidone forms complexes with tannins it is used for the clarification of beer and other beverages. The cross-linked, insoluble popcorn polymer is particularly suitable [136].

Polyvinylpyrrolidone forms hard, transparent, strongly adherent films on glass, metal, plastics, and cellulose, and is used in the adhesives industry as a binder and water-soluble hot-melt adhesive.

Polyvinylpyrrolidone is also used as an auxiliary in textile finishes, as a dye acceptor for synthetic fibers, as a leveling and stripping agent for dyes, as a thickener for printing inks and latex paints, as a dispersant in laundry detergents, as a protective colloid in the emulsion and suspension polymerization of many polymers, and as a water-binding agent for the concentration of protein solutions [137].

Toxicology. Polyvinylpyrrolidone is chemically and biologically largely indifferent [138, 139] with regard to both acute toxicity and skin irritation.

In feeding tests on rats with radioactively labeled polyvinylpyrrolidone of average molecular mass 40 000, more than 99 % was excreted through the intestinal tract [138]. Even when large amounts of polyvinylpyrrolidone were fed for years, no harm to experimental animals was established. Swelling of the lymph nodes, which can be interpreted as a transition stage in the excretion of polyvinylpyrrolidone, was observed only with dogs; no indications of accumulation or extensive degenerative changes to organs were detected.

Carcinogenicity of polyvinylpyrrolidone was not observed in any of the experimental procedures [140]. On the basis of extensive experience on humans as well as on the results of further experimental studies on animals with parenteral or oral administration, no evidence of carcinogenic risk to humans was found [141].

More than one million people have shown good tolerance for polyvinylpyrrolidone infusions as a blood plasma substitute without side effects. However, the high molecular mass fractions of polyvinylpyrrolidone, with molecular masses above 40 000, are excreted slowly or not at all. Polyvinylpyrrolidone is therefore no longer used for infusions because the low molecular masses types do not have a plasma-expander action. Animal experiments have shown that larger molecules are stored in the cells of the reticuloendothelial system of the spleen, the liver, and the lymph nodes, as well as in bone marrow. However, during such polyvinylpyrrolidone storage, neither morphological nor functional damage has been observed [142].

On the basis of inhalation tests on female volunteers with six different kinds of hairspray that caused no pathological changes of any kind [143], the FDA reported that the normal use of aerosol hairsprays containing PVP and PVP–vinyl alcohol as film formers is not accompanied by a health risk.

3.2. Polymers of other *N*-Vinyllactams

In the addition to polyvinylpyrrolidone, the polymers of other *N*-vinyllactams such as *N*-vinylpiperidone [137] and *N*-vinylcaprolactam [144] are known.

N-vinylpiperidone

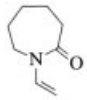

N-vinylcaprolactam

These polymers have better solubility than polyvinylpyrrolidone in organic solvents such as aromatic hydrocarbons, ketones, and aliphatic hydrocarbons. Polyvinylcaprolactam is readily soluble in cold water: the higher its average molecular mass, the better is its solubility. In warm water (30 – 40 °C), polyvinylcaprolactam precipitates from solution.

Vinylcaprolactam can be polymerized in bulk [145] or in solution. The polymers are used for gluing woven synthetic textiles and as components of lubricating oils [146]. The copolymers with ethylene are used as coatings for textiles and paper, and as adhesives [147].

3.3. Polyvinylamines

Poly(*N*-vinylcarbazole) is a transparent thermoplastic material. The glass transition temperature is 211 °C and the glass point 173 °C. It starts to decompose only above 300 °C [148].

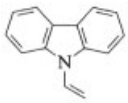

Vinylcarbazole

Poly(*N*-vinylcarbazole) dissolves in aromatic and chlorinated hydrocarbons and in tetrahydrofuran. It swells in benzene-containing fuels and is insoluble in water, dilute acid and alkali, aliphatic hydrocarbons, and alcohols.

The electrical properties are unusual. Photoconductivity depends on the molecular mass [149]; doping effects are observed upon addition of various substances such as 1,5-diaminonaphthalene, tetracyanoethylene, or anthracene [150]. The existence of traps for electronic charge carriers (electrons and defect electrons) has been demonstrated [151]. For the mechanism of conductivity, see [152].

Before polymerization, *N*-vinylcarbazole (*mp* 63 – 64 °C, *bp* at 0.3 kPa 155 °C) should be purified by distillation or recrystallization. When handling *N*-vinylcarbazole, safety instructions should be observed [153], because it has very strong sensitizing properties. No harmful side ef-fects are known for monomer-free poly(*N*-vinylcarbazole).

For radical polymerization in bulk, azo initiators or peroxides are used [154, 155]. Suspension polymerization in water [156] or methanol [157] is also known. Cationic polymerization of *N*-vinylcarbazole proceeds easily, but whether polymerization in the presence of benzoyl peroxide occurs via a radical or a cationic route is unknown [158]. Cationic polymerization has a rate constant five times higher than radical polymerization and is used for the production of low molecular mass polymers [159].

Radical copolymerization with other vinyl compounds such as styrene, acrylonitrile, vinyl esters, and acrylate esters has been described, and a Q value of 0.28 and an e of -1.49 are given [160]. The stereospecific polymerization of *N*-vinylcarbazole is also known [161].

Because its softening point is ca. 200 °C, polyvinylcarbazole was formerly used for moldings that are exposed to elevated temperature. Fibers and films have also been manufactured from polyvinylcarbazole [162]. Today, it is used as a heat-resistant insulator in radio and radar technology and in capacitor manufacture [163]. Because of its photoconductivity it is important for the manufacture of electrostatic photocopiers and television recording tubes [164].

Commercial Product: Luvican (BASF).

Polyvinylimidazole. Because of the second nitrogen atom, polymers of *N*-vinylimidazole have a basic character. The solubility therefore depends on the degree of protonation. Unquaternized polyvinylimidazole is soluble in organic solvents such as alcohol, but in water it is only swellable. It dissolves in acids, with salt formation. Quaternized polyvinylimidazole is water soluble. Depending on the length of the alkyl group used in quaternization, polymers with different hydrophobic – hydrophilic character are obtained [165].

N-vinylimidazole (*bp* 190 – 193 °C) and *N*-vinylimidazoles quaternized with alkyl halides or dialkyl sulfates can be polymerized by the radical mechanism using azo initiators and peroxides. Copolymerization with acrylamide, acrylates, butadiene, styrene, *N*-vinylpyrrolidone, and vinyl acetate has been described: values of 0.1 and 0.24 have been given for Q, and -0.91 and -1.73 for e [166].

N-vinylimidazole

Polyvinylimidazole and copolymers are used as hairdressing preparations, antistatic agents for hydrophobic plastics, liquid detergents, and in the production of hydrogels.

Other Vinylamines. Because of their instability, aliphatic vinylamines can be prepared only with difficulty, and the corresponding polyvinylamines are obtainable only by indirect routes [167]. Cyclic amines such as carbazole and imidazole, indole, and pyrrole, on the other hand, can easily be converted to *N*-vinylamines by Reppe vinylation.

Polymers of 2-vinylpyridine and 4-vinylpyridine have also been produced. 4-Vinylpyridine can be copolymerized with other vinyl compounds such as styrene and vinyl chloride; values of 1.91 for Q, and -0.51 for e are given [168]. Polyvinylpyridines have been studied as polymeric reagents in organic synthesis, catalysts, and ion exchangers.

4. References

1. Chem. Forschungsgemeinschaft, DE-OS 450 286 (W. O. Herrmann, W. Haehnel).
2. Chem. Forschungsgemeinschaft, DE 642 531, 1937 (W. O. Herrmann, W. Haehnel, H. Berg); *Chem. Abstr.* **31,** 59 059 (1937).

3. H. Staudinger, K. Frey, W. Starck, *Ber. Dtsch. Chem. Ges.* **60** (1927) 1782.
4. Du Pont, Elvon Hydroxyvinyl Resins, Technical Information, Wilmington, Del., 1969.
5. K. Noro: "Preparation of Modified Polyvinyl Alcohols from Copolymers," in: C. Finch (ed.): *Polyvinylalcohol,* Wiley Interscience, New York 1973, pp. 147 ff.
6. K. Noro, H. Takida, *Kobunshi Kagaku* **19** (1962) 261; *Chem. Abstr.* **58** (1963) 4650 c.
7. K. Fujii, S. Imoto, J. Ukida, M. Matsumoto, *J. Polym. Sci. Part B* **1** (1963) 497.
8. T. Ito, K. Nomo, *Kobunshi Kagaku* **15** (1968) 310; *Chem. Abstr.* **54** (1960) 8140 d.
9. J. Ukida, R. Naito, *Kogyo Kagaku Zasshi* **58** (1955) 717; *Chem. Abstr.* **50** (1956) 8245 h.
10. M. Shiraishi, *Kobunshi Kagaku* **19** (1962) 676.
11. K. R. Tubbs, *J. Polym. Sci. Polym. Chem. Ed.* **4** (1966) 623.
12. S. Hayashi, C. Nakano, T. Motoyama, *Kobunshi Kagaku* **21** (1964) 300; *Chem. Abstr.* **62** (1964) 9244 g.
13. S. Hayashi, C. Nakano, T. Motoyama, *Kobunshi Kagaku* **22** (1965) 354; *Chem. Abstr.* **63** (1965) 16476 h.
14. J. R. Member, H. C. Haas, R. L. MacDonald, *J. Polym. Sci. Part B* **10** (1972) 385.
15. T. Moritani, J. Kuruma, K. Shibatani, Y. Fujiwara, *Macromolecules* **5** (1972) 577.
16. J. F. Kenney, G. W. Willcocksen, *J. Polym. Sci. Polym. Chem. Ed.* **4** (1966) 679; H. N. Friedlander, H. E. Harris, J. G. Pritchard, **4** (1966) 649; H. E. Harris et al., **4** (1966) 665.
17. C. W. Bunn, *Nature (London)* **161** (1966) 929.
18. S. Hayashi, C. Nakano, T. Motoyama, *Kobunshi Kagaku* **20** (1963) 303; *Chem. Abstr.* **61** (1964) 5802 b.
19. I. Sakurada, Y. Nukushima, Y. Sone, *Kobunshi Kagaku* **12** (1955) 506.
20. T. Yamamoto, S. Konagaya, A. Yamamoto, *J. Polym. Sci. Polym. Lett. Ed.* **16** (1978) 7.
21. S. Nozakura, S. Kida, *J. Polym. Sci. Polym. Chem. Ed.* **12** (1974) 2337.
22. Consort. f. Ind. Elektrochem., DE 1 299 879; FR 1 361 830, 1964; DE-AS 1962; *Chem. Abstr.* **62** (1965) 6591 b.
23. F. Kainer: *Polyvinylalkohole,* Enke Verlag, Stuttgart 1949, p. 11.
24. C. E. Schildknecht: *Vinyl and Related Polymers,* Wiley Interscience, New York 1952, pp. 323 ff.
25. A. A. Vanscheidt, L. F. Chelpanova, *J. Gen. Chem. USSR Engl. Transl.* **20** (1950) 2261; *Chem. Abstr.* **45** (1951) 4482.
26. J. P. Flory, F. S. Leutner, *J. Polym. Sci.* **5** (1950) 267.
27. K. Imai, U. Maeda, *Kobunshi Kagaku* **16** (1959) 222.
28. Monsanto, BE 636 139, 1964 (R. J. Kern); *Chem. Abstr.* **61** (1964) 16187 d.
29. J. Sakurada, Y. Sakaguchi, K. Kashimoto, *Kobunshi Kagaku* **18** (1961) 694; *Chem. Abstr.* **56** (1962) 14461.
30. Y. Sakaguchi, Z. Sawada, M. Koizumi, K. Tamaki, *Kobunshi Kagaku* **23** (1966) 890; *Chem. Abstr.* **66** (1967) 65995 k.
31. K. Noro, *Br. Polym. J.* **2** (1970) 128.
32. F. Gregor, E. Engel, *Chem. Prum.* **10** (1960) 53.
33. Du Pont, US 2 657 201, 1953 (R. W. Nebel); *Chem. Abstr.* **48** (1954) 1066 a.
34. H. Shohota: "Continuous Polymerization of Vinylacetate for Polyvinylalcohol Production," in C. A. Finch (ed.): *Properties and Applications of Polyvinyl Alcohol, SCI Monogr.* **30** (1968) 18.
35. Shawinigan Chemicals, US 2 643 994, 1953 (L. M. Germain); *Chem. Abstr.* **47** (1953) 9058 e.
36. Eastman Kodak, US 2 642 419, 1953 (G. P. Waugh, W. O. Kenyon); *Chem. Abstr.* **47** (1953) 7925 h.
37. M. K. Lindemann in G. E. Ham (ed.): *Vinyl Polymerization,* **vol. 1,** Marcel Dekker, New York 1967, p. 252.
38. A. Hill, D. K. Hale, *Biostat Tech. Rep. (Amherst),* no. 1418 (1946).
39. K. Noro: "Manufacturing and Engineering Aspects of the Commercial Production of Polyvinylalcohol," in[61] p. 121.
40. G. Heck, A. Schmidt, *Chem. Ing. Tech.* **47** (1975) 541.
41. T. Suziki, Y. Ichahara, M. Yamada, K. Tonumura, *Agric. Biol. Chem.* **37** (1973) 747.
42. M. K. Lindemann, *Encycl. Polym. Sci. Technol.* **14** (1971) 149.
43. H. C. Haas, A. S. Makas, *J. Polym. Sci.* **46** (1960) 524.
44. R. Naito, *Kobunshi Kagaku* **15** (1958) 597; *Chem. Abstr.* **54** (1960) 16113 i.
45. Farbwerke Hoechst, DE 1 111 819, 1959 (G. Lohaus); *Chem. Abstr.* **56** (1962) 4973 g.
46. K. Toyoshima: "General Properties of Polyvinyl Alcohol in Relation to its Application," in[61], p. 17.
47. T. Motoyama, S. Okamura, *Kobunshi Kagaku* **11** (1954) 23; *Chem. Abstr.* **50** (1956) 2200 f.
48. Hoechst, Druckschrift über Mowiol, Frankfurt, Sept. 1976, D 1.

49. C. A. Finch: "Chemical Properties of Polyvinylalcohol," in [61]p. 183.
50. J. G. Pritchard: *Poly(vinylalcohol),* Gordon & Breach Sci. Publ., London 1970, p. 81.
51. M. Tsuda, *J. Polym. Sci. Part B* **1** (1963) 215.
52. M. Shiraishi, *Br. Polym. J.* **2** (1970) 135.
53. H. Lamont, *Adhes. Age* **16** (1973) 24.
54. A. S. Dunn, C. J. Tonge, S. A. B. Anabtawi, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **16** (1975) 223.
55. M. Shiraishi, K. Toyoshima, *Br. Polym. J.* **5** (1973) 419.
56. K. Toyoshima: "Applications of Polyvinylalcohol in Adhesives," in [61]p. 413.
57. E. P. Czerwin, *Mod. Text. Mag.* 1966, 22.
58. C. R. Blumenstein, *Text. Ind. (Atlanta)* **130** (1966) 63.
59. V. Heap, *Text. Inst. Ind.* **11** (1973) 172.
60. H. L. Jaffe, *Pap. Trade J.* **147** (1963) no. 9, 44.
61. H. Schaefer, H. G. Oesterlin, *Wochenbl. Papierfabr.* **102** (1974) 335.
62. R. Vesanto, H. Schaefer, D. Wolf, *Pap. Puu* **57** (1975) 399; *Chem. Abstr.* **83** (1975) 117430 v.
63. H. G. Oesterlin, H. Schaefer, *Papier (Darmstadt)* **32** (1978) V 13.
64. K. Toyoshima: "Properties of Polyvinyl Alcohol Films," in [61]p. 339.
65. Consortium f. elektrochem. Industrie, DE 480 866, 1929 (W. Haehnel, W. O. Herrmann).
66. I. G. Farbenindustrie, DE 683 165, 1939 (H. Hopff, E. Kühn); *Chem. Abstr.* **36** (1942) 38799.
67. I. G. Farbenindustrie, DE 692 988, 1940 (A. Voss, V. Starek); *Chem. Abstr.* **45** (1941) 45203; DE 737 630, 1943; *Chem. Abstr.* **38** (1944) 36647.
68. P. J. Flory, *J. Am. Chem. Soc.* **61** (1938) 1518.
69. H. Kawase, *Kogyo Kagaku Zasshi* **74** (1971) 1228; *Chem. Abstr.* **75** (1971) 119111 a.
70. K. Shibatani et al., *J. Polym. Sci. Part C* **23** (1968) 647.
71. S. Okamura, T. Motoyama, *Bull. Inst. Chem. Res. Kyoto Univ.* **30** (1952) 45; *Chem. Abstr.* **47** (1953) 6175 e. The Fiberloid, GB 503 634, 1939; *Chem. Abstr.* **33** (1939) 69927;Monsanto Chem., US 2 258 410, 1942 (J. Dahle); *Chem. Abstr.* **36** (1942) 5922.
72. J. Schreiber: *Chemie und Technologie der künstlichen Harze,* 2nd ed., **vol. 1,** Wissenschaftliche Verlagsgesell., Stuttgart 1961, p. 468.
73. General Electric, US 2 195 122, 1940 (B. W. Nordlander); *Chem. Abstr.* **34** (1940) 52063.
74. Farbwerke Hoechst, DE-OS 2 208 167, 1973 (E. Schmidt); *Chem. Abstr.* **80** (1974) 15739 y.
75. Monsanto, US 3 823 113, 1974 (A. J. Reisman); *Chem. Abstr.* **83** (1975) 60333 j.
76. General Electric, US 2 085 995, 1935 (W. I. Patnode, E. J. Flynn); *Chem. Abstr.* **31** (1937) 59029.
77. Shawinigan Chemicals, US 2 168 827, 1934 (G. O. Morrison, A. F. Price); *Chem. Abstr.* **33** (1939) 94901.
78. Farbwerke Hoechst, DE 878 861, 1951 (H. J. Hahn); *Chem. Abstr.* **51** (1957) 2324 g; DE 889 367, 1951 (W. Fitzky).
79. G. N. Kormanovskaja, I. N. Vlodavec, *Izv. Akad. Nauk. SSSR, Ser. Khim.* **10** (1964) 1748; *Chem. Abstr.* **62** (1965) 4130 f.
80. M. Chanda, *Angew. Makromol. Chem.* **62** (1977) 229.
81. Monsanto: Butvar, Polyvinylbutyral and Formvar, Polyvinylformal, Techn. bull. 6070 and 6130, 1969.
82. M. K. Lindemann, *Encycl. Polym. Sci. Technol.* **14** (1971) 208.
83. General Electric, US 2 307 588, 1943 (E. H. Jackson, R. W. Hall); *Chem. Abstr.* **37** (1943) 35343.
84. A. F. Fitzhugh, E. Lavin, G. O. Morrison, *J. Electrochem. Soc.* **100** (1953) 351.
85. L. Dimter, H. Schulz, K. Thinius, *Plaste Kautsch.* **9** (1962) 318.
86. T. Goldschmidt, DE-OS 2 503 288, 1976 (R. Mitgan, D. Schedlitzki, H. Wacker); *Chem. Abstr.* **85** (1976) 161476 e.
87. Kanebo, Tokio, DE 2 323 968, 1972 (T. Koide et al.).
88. H. Schindler, *Kunstharz Nachr.* **9** (1975) 22.
89. O. V. Piastro, L. L. Ezhenkova, N. J. Tyazhlo, M. E. Rozenberg, *Plast. Massy* 1970 no. 2, 13; *Chem. Abstr.* **72** (1970) 133504 a. N. I. Tyazhlo et al., *Zh. Prikl. Khim. (Leningrad)* **47** (1974) 2285; *Chem. Abstr.* **82** (1975) 58806 b.
90. I. G. Farbenindustrie, DT 891 745, 1939 (A. Weihe, F. Herrlein); *Chem. Abstr.* **50** (1956) 13506 c; DT 904 592, 1939 (A. Weihe, F. Herrlein); *Chem. Abstr.* **50** (1956) 13506 f.
91. Du Pont, DE 1 247 663, 1967 (L. H. Rombach); *Chem. Abstr.* **67** (1967) 91324 f.
92. Farbwerke Hoechst, DE 899 864, 1953 (H. J. Hahn); *Chem. Abstr.* **50** (1953) 14269 h.
93. Shawinigan Resins, US 2 496 480, 1950 (E. Lavin, A. T. Marinaro, W. R. Richard); *Chem. Abstr.* **44** (1950) 5639 f.
94. Farbwerke Hoechst, DE-OS 1 770 580, 1968 (H. Fritz, F. Sorg); *Chem. Abstr.* **72** (1970) 133690 h.
95. L. D. Taylor, B. Biasotti, *J. Appl. Polym. Sci.* **20** (1976) 1721.

96. Hoechst AG, Mowital B, Polyvinylbutyral, Product Information, 1976.
97. Sekisui Chem. Ind.: S-lec B Polyvinylbutyral Resin, Technical data, 1977.
98. *Houben-Weyl,* 4th ed., **14/2,** 717.
99. Monsanto, DE-OS 2 429 032, 1973 (R. H. Fariss, J. A. Snellgrove); *Chem. Abstr.* **82** (1975) 157268 j.
100. H. Kessler in *Kunststoff-Handbuch,* **vol. 11,** Carl Hanser Verlag, München 1971, p. 518.
101. G. Schulz, G. Müller, *Kunststoffe* **42** (1952) 298.
102. Sekurit Glas Union, Verbundsicherheitsglas, brochure no. 4, Aachen 1977.
103. *Glass* **54** (1977) 190.
104. J. Svoboda, V. Matys, *Plasty Kauc.* **12** (1975) 131; *Chem. Abstr.* **83** (1975) 60851 b.
105. Monsanto, DT 1 596 895, 1965;US 3 271 233, 1966 (E. Lavin, G. E. Mont); *Chem. Abstr.* **65** (1966) 20323 g.
106. Farbwerke Hoechst, DE 206 123, 1961 (F. Winkler, H. Staerk, H. Kessler, W. Grundmann); *Chem. Abstr.* **64** (1966) 6277 c.
107. E. Lavin, J. A. Snelgrove in *Adhesives,* 2nd ed., Van Nostrand-Reinhold, New York 1977, p. 507.
108. A. Steen, *Ind. Lackierbetr.* **44** (1976) 367.
109. BASF, DE-OS 2 538 097, 1977 (G. Reuss); *Chem. Abstr.* **86** (1977) 156973 f.
110. M. P. Portyanko, V. F. Yavoroskaya, A. E. Martinovich, G. M. Lishanova, *Tekhnol. Organ. Proizvod* **14** (1937), no. 45; *Chem. Abstr.* **80** (1974) 97372 k.
111. A. D. Yakovlev, I. S. Okhrimenko, *Mashinostroitel* 1973 no. 9, 20; *Chem. Abstr.* **84** (1976) 6535 k.
112. *Kirk-Othmer,* 2nd ed., **21,** 304.
113. A. F. Fitzhugh, R. N. Crozier, *J. Polym. Sci.* **8** (1952) 225.
114. Farbwerke Hoechst, DE 947 144, 1954 (W. Stark, W. Langbein); *Chem. Abstr.* **67** (1958) 6794 e.
115. BASF, DE 922 378, 1943 (W. Reppe, K. Herrle, H. Fikentscher); *Chem. Abstr.* **52** (1958) 19252 e.
116. BASF, DE-OS 2 439 196, 1974 (K. Herrle, W. Denzinger, K. Seelert); *Chem. Abstr.* **84** (1976) 165420 j.
117. J. W. Breitenbach, A. Schmidt, *Monatsh. Chem.* **83** (1952) 1288.
118. E. Senogles, R. Thomas, *J. Polym. Sci. Polym. Symp.* **49** (1975) 203.
119. I. G. Farbenindustrie, DE 757 355, 1939 (C. Schuster, R. Sauerbier, H. Fikentscher).
120. General Aniline & Film, DE-OS 2 602 917, 1976 (D. H. Lorenz, E. P. Williams, H. S. Schultz); *Chem. Abstr.* **86** (1977) 56214 r.
121. H. Uelzmann, *J. Polym. Sci.* **33** (1958) 377.
122. S. Ponratnam, S. L. Kapur, *J. Polym. Sci. Polym. Chem. Ed.* **14** (1976) 1987.
123. BASF, DE 954 197, 1956 (H. Fikentscher, H. Wilhelm); *Chem. Abstr.* **53** (1959) 13674 c.
124. J. F. Bork, E. Coleman, *J. Polym. Sci.* **43** (1960) 413.
125. B. Vollmert: *Polymer Chemistry,* Springer Verlag, New York 1973, p. 139.
126. K. Hayashi, G. Smets, *J. Polym. Sci.* **27** (1958) 275.
127. D. J. Kahn, H. H. Horowitz, *J. Polym. Sci.* **54** (1961) 363.
128. General Aniline & Film, US 2 938 017, 1956 (F. Grosser); *Chem. Abstr.* **54** (1960) 20335 f.
129. BASF, DE-OS 2 437 629, 1976 (W. Denzinger, E. Hoffmann, K. Herrle); *Chem. Abstr.* **84** (1976) 165586 t.
130. W. Breitenbach, H. F. Kauffmann, *Angew. Makromol. Chem.* **45** (1975) 167.
131. A. Nakano, Y. Minoura, *J. Appl. Polym. Sci.* **21** (1977) 2877.
132. Chemical-Technological Institute, Moskau, SU 471 370, 1975 (A. S. Tevlina, V. G. Chelnokov, M. M. Mardanyan, G. V. Makarov); *Chem. Abstr.* **83** (1975) 9873 u.
133. B. Wurzschmitt, *Fresenius Z. Anal. Chem.* **130** (1949/50) 128.
134. U. Müller, *Pharm. Acta Helv.* **43** (1968) 108.
135. H. A. Shelansky, M. V. Shelansky, *J. Int. Coll. Surg.* XXC (1956) no. 6, 727.
136. BASF, DE-OS 2 437 640, 1976 (W. Denzinger, E. Hoffmann, K. Herrle); *Chem. Abstr.* **84** (1976) 165587 u.
137. P. C. Brown, R. Cousden, *Chem. Ind. (London)* 1955, 1452.
138. BASF, Brochure B 359 d, Ludwigshafen, 1976, p. 29.
139. W. Wessel, M. Schoog, E. Winkler, *Arzneim. Forsch.* **21** (1971) 1468.
140. L. W. Burnette, *Proc. Sci. Sect. Toilet Goods Assoc.* **38** (1962) 1.
141. H. Zeller, BASF Bericht WNT, Ludwigshafen, 19.01.1978.
142. O. Fresen, W. Weese, *Beitr. Pathol. Anat. Allg. Pathol.* **112** (1952) 47.
143. J. H. Draize, *Proc. Sci. Sect. Toilet Goods Assoc.* **31** (1959) 28.
144. N. Cobianu, S. D. Vasilescu, S. Matache, *Mater. Plast. Chem. Abstr.* **10** (1973) no. 2, 75; *Chem. Abstr.* **79** (1973) 66822 g.
145. O. F. Solomon, M. Corciovei, C. Boghina, *J. Appl. Polym. Sci.* **12** (1968) 1843.

146. General Aniline & Film, US 3 287 272, 1966 (W. Katzenstein); *Chem. Abstr.* **66** (1967) 48127 m.
147. Bayer, DT 1 392 354, 1964.
148. J. Pielichowski, *J. Therm. Anal.* **4** (1972) 339.
149. K. Tanikawa, *Makromol. Chem.* **176** (1975) 3025.
150. K. Okamoto, *Bull. Chem. Soc. Jpn.* **46** (1973) 2613.
151. H. Bauser, *Kunststoffe* **62** (1972) 192.
152. P. J. Rencroft, *J. Polym. Sci. Polym. Phys. Ed.* **10** (1972) 2305.
153. I. R. Tabershaw, J. B. Skinner, *J. Ind. Hyg. Toxicol.* **26** (1944) 313.
154. BASF, DE 931 731, 1953 (H. Fikentscher, R. Frikker); *Chem. Abstr.* **52** (1958) 12458 g; DE 936 421, 1955 (H. Fikentscher, R. Fricker); *Chem. Abstr.* **52** (1958) 1685 a.
155. British Oxygen, GB 1 034 309, 1966 (L. P. Ellinger); *Chem. Abstr.* **65** (1966) 9104 f.
156. Ricoh, NL 6 905 628, 1968.
157. General Aniline & Film, DE-OS 2 111 293, 1971 (E. V. Hort); *Chem. Abstr.* **76** (1972) 46676 x.
158. J. C. Bevington, *Makromol. Chem.* **178** (1977) 2741.
159. General Aniline, DE-OS 2 111 294, 1971 (E. V. Hort); *Chem. Abstr.* **76** (1972) 46678 z.
160. J. Negulesen, D. Feldmann, *Polymer* **13** (1972) 149.
161. K. Okamoto et al., *Macromolecules* **9** (1976) 645.
162. Montecatini Societa Generale per l'Industria Mineralia e Chimica, GB 914 418, 1963; *Chem. Abstr.* **58** (1963) 9252 f.
163. Standard Telephones & Cables, GB 1 007 040, 1965 (E. H. Hornisch); *Chem. Abstr.* **64** (1966) 836 a.
164. Kalle, DE 1 127 218, 1962 (H. Hoegl); *Chem. Abstr.* **58** (1963) 138 c.
165. J. C. Salamone, *J. Polym. Sci. Polym. Symp.* **45** (1974) 65.
166. J. C. Salamone, P. Taylor, B. Smider, *J. Polym. Sci. Polym. Chem. Ed.* **13** (1975) 161.
167. G. Franzmann, H. Ringdorf, *Makromol. Chem.* **177** (1976) 2547.
168. K. Matsuoka, M. Otsuka, K. Takemoto, M. Imoto, *Kogyo Kagaku Zasshi* **69** (1966) 137; *Chem. Abstr.* **65** (1966) 15515 f.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Polysaccharides

Standard Article

Alphons C. J. Voragen[1], Claus Rolin[2], Beinta U. Marr[3], Ian Challen[4], Abdelwahab Riad[5], Rachid Lebbar[6], Svein Halvor Knutsen[7]

[1]Wageningen, Agricultural University, Department of Food Science, Wageningen, The Netherlands
[2]The Copenhagen Pectin Factory Ltd., Lille Skensved, Denmark
[3]The Copenhagen Pectin Factory Ltd., Lille Skensved, Denmark
[4]NutraSweet Kelco, Waterfield, Tadworth, Surrey, United Kingdom
[5]Setexam, Kenitra, Morocco
[6]Setexam, Kenitra, Morocco
[7]Norwegian Food Research Institute, MATFORSK, Norway

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a21_a25.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Analysis and Characterization
3. Pectin
3.1. Occurrence and Structure
3.2. Pectolytic Enzymes
3.3. Production
3.4. Properties
3.4.1. Physical Properties
3.4.2. Gel Properties
3.4.3. Stability and Chemical Reactions
3.5. Analysis
3.5.1. Measurement and Standardization of Gel-Forming Capacity
3.5.2. Chemical Analysis
3.6. Pharmaceutical and Nutritional Characteristics
3.7. Application in the Food Industry
3.8. Market
4. Alginates
4.1. Occurrence
4.2. Production
4.3. Structure
4.4. Properties
4.5. Propylene Glycol (Propane-1,2-diol) Alginate
4.6. Bacterial Alginates
4.7. Analysis
4.8. Applications
4.9. Market
5. Carrageenan
5.1. Structure
5.2. Sources and Raw Materials
5.3. Production
5.4. Analysis
5.5. Properties
5.6. Applications
5.7. Physiological Properties
6. Agar
6.1. Production
6.2. Structure and Gelling Mechanism
6.3. Quick Soluble Agar
7. Gum Arabic
8. Gum Tragacanth
9. Gum Karaya
10. Gum Ghatti
11. Xanthan Gum
11.1. Production

11.2. Structure and Properties
11.3. Analysis
11.4. Applications, Market
12. Gellan Gum
13. Galactomannans
13.1. Structure
13.2. Production
13.3. Properties
13.4. Analysis and Composition of Commercial Preparations
13.5. Derivatives
13.6. Applications
13.7. Market
14. Acknowledgement

[Top of Page]

1. Introduction [1-12]

Alphons C. J. Voragen

For a general definition of the term polysaccharides, see ⟶ Carbohydrates.

Polysaccharides made up of only one type of neutral monosaccharide structural unit and with only one type of glycosidic linkage — as in cellulose or amylose — are denoted as *perfectly linear* polysaccharides. In branched polysaccharides the frequency of branching sites and the length of the side chains can vary greatly. Molecules with a long “backbone” chain and many short side chains are called *linearly branched* polysaccharides.

Polysaccharides are water soluble or swell in water, giving colloidal, highly viscous solutions or dispersions with plastic or pseudoplastic flow properties. Functional properties such as thickening, water holding and binding, stabilization of suspensions and emulsions, and gelling, are based on this behavior. Therefore, polysaccharides are often referred to as gelling or thickening agents, stabilizers, water binders, or fillers. A more generic name is gums. The terms hydrocolloids or, in the food industry, food colloids are also in use, but they include proteins with similar functional properties.

In this article the economically important polysaccharides, with the exception of starch (see ⟶ Starch) and cellulose (⟶ Cellulose) are described. Table 1 lists the polysaccharides, and gives an overview of their sources and composition. Their most important and various applications are summarized in Table 2. In only a few cases is their use specific. Examples are the use of pectin in commercial jam and marmalade production, where its heat stability coincides with its ability to gel in the typical pH range of these products, or the stabilization of drinking chocolate with carrageenan, which has a specific interaction with milk proteins and chocolate particles. In many cases, the manufacturer has various options and price plays an important role. For papermaking choice may be between, for example, alginates, galactomannans, and starch derivatives; for oil drilling, xanthan, starch derivatives, cellulose derivatives, galactomannans, or alginates can be used. Milk gels can be prepared with pectin, starch, alginates, or carrageenans. Often mixtures of polysaccharides are preferred to single polysaccharides because they combine various desirable properties (e.g., in ice cream or salad dressing). In a number of cases a synergistic interaction on the molecular level between polysaccharides can be exploited (e.g., improvement of the structure of carrageenan gels by the addition of guar gum or the preparation of gels by mixing xanthan and locust bean gum).

Table 1. Origin and main sugar moieties of polysaccharides

Name	Origin	Main sugar moieties
Pectin	cell walls and middle lamella of higher land plants	D-galacturonic acid D-galacturonic acid methylester
Alginate	cells walls of brown seaweeds exopolysacchrides of *Azetobacter vinelandii*	D-mannuronic acid L-guluronic acid and their acetyl derivatives
Carrageenan	cell walls of red seaweeds	D-galactose 3,6-anhydro-D-galactose both sugars sulfated to higher or lower degree
Agar	cell walls of red seaweeds	D-galactose 3,6-anhydro-L-galactose few sulfate groups
Gum arabic	exudate of acacia species	L-arabinose, D-galactose L-rhamnose, (4-*O*-methyl) D-glucuronic acid
Gum tragacanth	exudate of astragalus species	L-arabinose, D-galactose D-galacturonic acid methylester D-xylose, L-rhamnose, L-fucose
Gum karaya	exudate of sterculia species	D-galacturonic acid, L-rhamnose, D-galactose D-glucuronic acid
Gum ghatti	exudate of *Anageissus latifolia*	L-arabinose, D-galactose D-mannose, D-xylose, D-glucuronic acid, L-rhamnose
Guar gum	endosperm of seeds of the guar plant (*Xyampsis tetragonolobus* L. Taub)	D-mannose D-galactose
Locust bean gum	endosperm of the seeds of the carob tree (*Ceratnia siliqua* L.)	D-mannose D-galactose
Tara gum	endosperm of the seeds of the tara tree (*Caesalpinia spinoza*)	D-mannose D-galactose
Xanthan gum	exopolysaccharide of *Xanthomonas campestris*	D-glucose, D-glucoronic acid, D-mannose substituted with acetate or pyruvate groups
Gellan gum	exopolysacchride *Pseudomonas elodea*	D-glucose, D-glucuronic acid, L-rhamnose

Table 2. Areas of application of polysaccharides *

Foods (⟶ Foods, 3. Food Additives)

1. Thickening and gelatinizing of fruit with a higher or lower content of sugar: preserves, marmalades, jellies, fruit for yogurt and ice cream, marmalade for cooking.
2. Gelatinizing of heat-reversible jelly and sugar glaze. Jellied sugar products, jellied sweets, dessert jellies (powder form).
3. Prevention of starch breakdown in bread and baked products, and in fruit-filled products based on starch (⟶ Bread and Other Baked Products).
4. Cloud stabilization in fruit juice. Providing body in refreshing powdered drinks (⟶).
5. Stabilization of powdered flavor emulsions. Microencapsulation of flavorings (⟶ Microencapsulation).
6. Gelatinization and thickening of milk: puddings and cream for hot and cold preparations; also as powdered instant products (⟶).
7. Stabilization of chocolate milk and evaporated milk: prevention of sedimentation or separation of the cream fraction (⟶).
8. Stabilization of ice cream: prevention of formation of ice crystals and dripping. Stabilization of the fat emulsion, imparting favorable slip properties for mold release (⟶).
9. Water binding in cream cheese-type spreads, soft cheese, and cheese preparations.
10. Stabilization of fat emulsion in powdered coffee whiteners and low-fat margarine.
11. Stabilization of egg-white froth, whipped cream, and imitation whipped cream, meringues, and beer foam.
12. Stabilization and thickening of soups, sauces, dips, salad cream, mayonnaise, catsup, etc. Prevention of syneresis. Imparting freeze – thaw stability, and body in fat and starch-reduced products.
13. Binding agents in ground meat products (e.g., corned beef, sausages) and in canned dog and cat foods.
14. Gelatinized (jellied meat) or thickened stock from meat, fish, and vegetables.
15. Binding agents for (potato flour) snacks.

Cosmetics

16. Thickening agents in creams, ointments, lotions, and hair gels (⟶ Skin Cosmetics).

Household Goods and Industry

17. Solid air fresheners.
18. Stabilization of emulsified and suspended active agents in insecticides and herbicide sprays.
19. Stabilization of suspended abrasion materials in polishes and cleaning agents.
20. Glue.

Pharmaceuticals

21. Thickening agents in creams, lotions, ointments, and syrups.
22. Toothpaste (water-binding, consistency, ⟶ Dental Materials).
23. Denture adhesives.
24. Tablet-binding agents.
25. Dental impression materials (⟶ Dental Materials).

Industry

26. Binding agents in explosives.
27. Binding agents for layered coating of welding electrodes.
28. Different functions of the rinsing liquid used in deep drilling technology.
29. Paper binding agents. Paper coating agents (fat-resistant properties, improved printability, uniform distribution of pigment despite extremely high speeds of modern papermaking machines (⟶ Paper and Pulp, Production); additives to improve absorbency.
30. Sizing and finishing agents used in the textile industry (⟶ Textile Auxiliaries).
31. Thickening of textile printing pastes and coloring agents to ease application and prevent color bleeding.
32. Flocculation agents for water processing (⟶ ; ⟶ Wastewater).
33. Pigment stabilization and paint consistency (thixotropy in emulsion paints).

---

* The numbering has been introduced merely to allow reference in the text and does not rank the applications in order of importance.

[Top of Page]

2. Analysis and Characterization

Alphons C. J. Voragen

The analysis and characterization of polysaccharides can be aimed at the following:

1. Estimation of their content in raw materials
2. Qualitative and quantitative identification of one or more polysaccharides in a preparation
3. Qualitative and quantitative identification of one or more polysaccharides in a (food) product
4. Estimation of specific properties such as gelling strength or viscosity
5. Estimation of specifications for purity (e.g., color, content of heavy metals or cations, moisture, ash, or microbial count)

The manufacturer is interested primarily in item 1); items 2) and 3) are of concern to the user and 3) for state inspection services; item 4) is of interest to the manufacturer in terms of standardization, and the user in terms of applications; item 5) is of particular interest to the consumer and to state inspection services in pharmaceutical and food applications.

For determination of items 2) and 3), several possiblities exist [13], [14]. In most cases, solutions or extracts must be prepared in which the polysaccharides are estimated qualitatively by precipitation reactions [15], or a special analytical procedure is used to separate anionic polysaccharides from neutral polysaccharides by precipitation with cetylpyridinium chloride [16], [17]. When the necessary equipment is available, quantitative analysis, based on the gas chromatographic determination of the constituent sugar moieties after hydrolysis and derivatization of the purified polysaccharide is often quicker and more reliable [29]. Recently, HPLC has been used more often in the analysis of constituent sugar moieties. Another new development is the use of immunological methods based on the development of antibodies against the polysaccharides. The requirements indicated under item 5) and the corresponding analytical methods are defined in pharmacopeias, food legislation, and specifications of national and international organizations [41, 62, 63]. The latter also describe identification reactions. Additional information on the analysis of individual polysaccharides is given in the corresponding sections.

[Top of Page]

3. Pectin

Alphons C. J. Voragen, Claus Rolin, Beinta U. Marr

Pectin or pectic substances, also called galacturonans or rhamnogalacturonans in scientific literature, is a collective name for heteropolysaccharides that consist predominantly of partially methylated galacturonic acid residues [18, 74, 85, 96, 104-111]. The name pectin (Greek: *Pectos* = gelled) was coined by BRACONNOT [112] who first described this compound in detail in 1825 and indicated its primary use as a gelling agent. Native pectin plays an important role in the consistency of fruits and vegetables, and in textural changes during ripening and storage [85], [113]. The often desired cloud stability of fruit juices and fruit drinks is lost when enzymes endogenous to the fruit degrade pectin (tomatoes) or cause it to precipitate as calcium pectate after demethylation. On the other hand, enzymes may be added as processing aids to degrade native pectin, for example, to apple juice to facilitate clarification or to the pulp of berries to improve press and color yield [114-116]. Pectin is extracted on an industrial scale from the press residues in apple and citrus juice manufacture and used mainly as gelling and stabilizing agents in the food industry [117].

3.1. Occurrence and Structure

Occurrence. Pectins occur in virtually all higher plants, Zosteraceae seaweed, and certain freshwater algae. Pectins are major structural components of the primary cell wall and the middle lamella of young growing plant tissues (meristimatic and parenchymatic) but do not occur in more mature tissue. The composition of the cell wall therefore is of major importance in the texture of fruit and vegetables [19], [20]. The primary cell wall consists of 90 % polysaccharide and 10 % glycoprotein on a dry matter basis [21]. The polysaccharide composition is 30 – 60 % cellulose, 15 – 45 % pectic substances, and 15 – 25 % hemicellulose [22]. A schematic structure of the plant cell wall is illustrated in Figure 1.

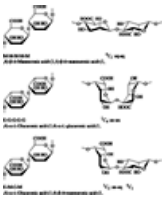

Figure 1. Schematic structure of the plant cell wall [23]

The biosynthesis of pectin takes place in the cell plate during cell division. Pectin is formed as polygalacturonic acid with UDP-D-galacturonic acid — arising from UDP-D-glucose by an epimerase- catalyzed reaction — as the most active glycosyl donor. Immediately after the galacturonan chain has been formed, methoxyl groups are formed with *S*-adenosylmethionine as the methyl group donor [120].

Pectic substances can be partially solubilized from plant tissues without degradation by using weakly acidic, aqueous solvents with or without calcium chelating agents. The pectin fraction that is not extractable with these extractants because of its attachment to other cell wall components by chemical, physical, or mechanical (enmeshment) bonds, is often designated as *protopectin*. Commercial pectin extraction must break down protopectin to a soluble, high molar mass pectin. This is achieved by acid hydrolysis in which, on the one hand, the molar mass of pectin molecules is lowered and, on the other hand, connections to the hemicellulose fractions are split. This transition from protopectin to soluble pectin also occurs during ripening of fruit or cooking of vegetables, resulting in textural changes [85], [113]. Pectin technology is therefore interested in the nature of protopectin and its fixation in the cell walls in an attempt to develop more efficient methods of extraction. FRY [119] has discussed cross-links in cell walls and agents used to cleave the individual bonds. Pure enzymes may also be used for extraction of firmly bound pectin [24], [25].

Structure. Pectin is composed of 1,4-linked α-D-galactopyranosyluronic acid units in the $^4C_1$ conformation, with the glycosidic linkages arranged diaxially (Fig. 2). A proportion of the carboxyl groups is esterified with methanol. Commercial pectins are divided into low-ester and high-ester pectin. In *low-ester pectins (LM-pectins)* less than 50 % of the carboxyl groups are methylated (typical range is from 20 to 40 %) whereas in *high-ester pectins (HM-pectins)* more than 50 % are methylated (typical range is from 55 to 75 %). If less than 10 % of the carboxyl groups are methylated the polysaccharide is called *pectate* or *pectic acid*. Pectins prepared from pears, potatoes, sugar beets, and sunflower heads are acetylated to varying degrees at the secondary hydroxyl groups of the galacturonic acid residues [118].

Figure 2. α-D-Galactopyranosyluronic acid in $^4C_1$ conformation (above); fragment of galacturonan chain, 40 % methylated (below)

The heteropolysaccharide nature of pectin derives from the fact that other sugars are incorporated in the pectin molecule. The most common ones being L-rhamnose (Fig. 3), occasionally inserted by α-1,2-linkages in the galacturonan backbone, providing “kinks” in the molecular chain. Other sugars are β-D-xylose, attached as single-unit side chains mainly to O-3 of the galactopyranosyluronic acid residues in the backbone; and D-galactose and L-arabinose, which occur in long side chains, only attached to rhamnopyranosyl residues (for projection formulae, see ⟶ Carbohydrates).

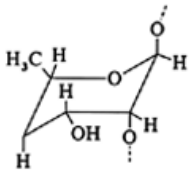

Figure 3. 1,2-Linked L-rhamnopyranosyl unit in $^1C_4$ conformation

The frequency of rhamnose occurrence remains to be established, however it has been suggested that α-rhamnosyl units are concentrated in rhamnose-rich areas. In other words, the soluble pectin is built up of homogalacturonan-dominated areas, so-called *smooth regions* linked to rhamnogalacturonan areas rich in neutral sugars, so-called *hairy regions* (Fig. 4). The neutral sugars account for 10 – 15 % of the weight of the pectin. In the hairy regions, the neutral sugar chain length may be in the range 8 to 20 residues [26], [27]. By degradation with chemical β-elimination and endo-polygalacturonase, it was found that 90 % of the rhamnose units are found in the hairy region [28], [30].

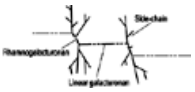

Figure 4. Schematic representation of pectin backbone, showing the “hairy” regions (rhamnogalacturonan and side-chains) and the “smooth” regions (linear galacturonan) [31]

By acid hydrolysis, i.e., splitting the acid labile glycosidic bonds between rhamnose and galacturonic acid, nearly pure homogalacturonic acids with molecular masses in the range 20 000 to 25 000 have been obtained [32], [33]. This corresponds to a chain length of 75 to 100 galacturonic acid residues. For comparison, a molecular mass of 90 000 has been quoted for intact pectin [34], but the molecular mass of pectin is somewhat

uncertain (see Section Chemical Analysis).

Commercial pectins may also contain neutral polysaccharides that are not covalently attached to the pectic backbone, such as galactans, arabinans, arabinogalactans, and starch, often referred to as “ballast” compounds. Purified pectins prepared from apple pomace or citrus peel may contain 75 – 90 % anhydrogalacturonic acid on an ash- and moisture-free basis.

3.2. Pectolytic Enzymes

Pectins can be attacked by various enzymes [114], [115], [121]. The significance of native pectic substances in food technology can be evaluated properly only when the activity of these enzymes is taken into account.

Pectinesterase (PE, pectin methylesterase, pectase, pectin demethoxylase, pectin pectylhydrolase, EC 3.1.1.11) splits off the methoxyl groups and converts high-methoxyl pectins to low-methoxyl pectins. The latter are extremely sensitive to complex formation and precipitation with Ca^{2+} ions, particularly when a pectinesterase of plant origin is used. This type of enzyme does not saponify methyl esters in a random fashion as microbial PEs do, but acts along the galacturonan chain, creating blocks of free carboxyl groups. Pectinesterase occurs in many higher plants, particularly tomatoes, citrus, and other fruits; it is also produced by many fungi and bacteria.

Polygalacturonase (PG, pectinase, pectate hydrolase, poly- α-1,4-D-galacturonide glycanohydrolase, EC 3.2.1.15 and 3.2.1.67) preferentially hydrolyzes low-methoxyl pectins or pectic acid because these enzymes can cleave glycosidic linkages only next to free carboxyl groups. PGs can be divided into enzymes that degrade their substrate by an endo attack (splitting randomly in the backbone, *endoPG*), and enzymes that act from the nonreducing end removing mono- or digalacturonic acid (*exoPG*). PGs are produced by fungi and certain bacteria, and also occur in higher plants (tomatoes). The endoPGs, with their strong depolymerizing action, are of particular technological importance.

Pectate lyase (PAL = pectic acid lyase, PATE = pectic acid transeliminase, LMPL = low-methoxyl pectin lyase, poly- α-1,4-D-galacturonide lyase, EC 4.2.2.2 and 4.2.2.9) also splits glycosidic linkages next to free carboxyl groups. In this group of enzymes, endo and exo enzymes also exist. The preferential substrates for *endoPAL* are LM-pectins rather than pectic acid. Pectate lyases have an absolute requirement for Ca^{2+} ions. The glycosidic linkages are split by a *trans*-elimination reaction. PALs are produced predominantly by bacteria and are not important in fruit and vegetable processing because of their high optimum pH (8.5 – 9.5).

Pectin lyase (PL, PTE = pectin transeliminase, pectinase, poly- α-1,4-D-methoxygalacturonide lyase, EC 4.2.2.10) splits glycosidic linkages between methoxylated galacturonide residues by a *trans*-elimination reaction. These enzymes therefore have a preference for HM-pectins. Pectin lyases are produced only by fungi.

A newer pectin-degrading enzyme acts in cooperation with a pectin acetylesterase only on highly branched regions of pectin to release oligosaccharides consisting of alternating sequences of α-1,2-linked L-rhamnosyl residues and α-1,4-linked D-galacturonosyl residues, with galactosyl residues β-1,4-linked to part of the rhamnosyl units. The nonreducing end is always a rhamnose unit [122], [123].

Commercially available pectinases, used on an industrial scale in fruit and vegetable processing are of fungal origin and generally contain in addition to PE, PG, and PL, proteases and various hemicellulases and cellulases.

3.3. Production [106], [109], [117], [122], [124]

The production of pectin is summarized in Figure 5.

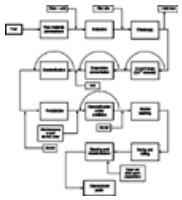

Figure 5. Flow chart of pectin production

Raw materials of importance to pectin manufacturing are currently various kinds of citrus peel, and apple pomace. Lemon and lime are the preferred citrus sources, and more pectin is produced from these than from apple or the less preferred citrus materials, orange and grapefruit. Some of the pectin producing companies which historically developed in connection to apple production now partly or wholly base their production on imported citrus peel. Sugar beet was once used as a pectin source to some extent [35], but the pectin is inferior as a gelling agent compared to citrus or apple pectin. It has been reintroduced, and is currently being marketed as a stabilizer. Numerous other sources like mango [36], pea hulls [37], sunflower heads [38], [39], and pumpkin [40] have been suggested.

Citrus peel and apple pomace are available as byproducts from juice manufacturing. They are usually washed in water and dried before being used for pectin manufacturing, but some citrus material is used in pectin plants neighboring the juice production without previous drying. In either case processing of the raw materials has to commence immediately after juice production in order to prevent microbial degradation. The washing leaches out organic acids, sugar, and pigments and is thus one of the separation processes in the purification of pectin. Most importantly, it prevents discoloration either from browning of pigments or from caramelization during the raw materials drying.

Extraction. The pretreated raw material is extracted in water which has been acidified with e.g., hydrochloric or nitric acid. Typical conditions are: pH 1 to 3, temperature 50 to 90 °C, duration 3 to 12 h. During the extraction, limited depolymerization of the pectin and possibly of other connecting biopolymers takes place, and the pectin dissolves. The low pH further dissociates ionic linkages which hold the pectin in the plant tissue. In addition to hydrolyzing glycosidic bonds, the extraction conditions also hydrolyze ester linkages, more specifically the methyl ester at C-6, and the acetate to which pectin may be esterified by its hydroxyl groups. The extraction process thus causes a reduction in degree of polymerization as well as in degree of methyl and acetate esterification. The pectin yield increases with the acidity, the temperature, and the duration, but the product will lose too much in degree of polymerization if all these parameters are at their maximum. The combination of low pH and low temperature favors hydrolysis of ester linkages over hydrolysis of glycosidic bonds, and it is thus preferred for production of pectin with a relatively low degree of esterification.

Filtration in one or more stages separates the extract containing the solubilized pectin from the insoluble, but at this stage very soft and fragile, plant tissue. The rather difficult filtration requires reasonably low viscosity, and as a consequence the pectin concentration must be less than 0.6 to 1 %, depending upon the pectin type. Further, the solids must not have been comminuted by excessive mechanical treatment such as vigorous agitation. Water-insoluble materials like wood cellulose or diatomaceous earth may be added in order to improve the porosity and mechanical strength of filter cakes. Amylase may be added to remove starch from apple pectin extracts.

The spent plant raw material is typically used for cattle feed.

Isolation. Following filtration, the extract may optionally be passed through a column with cation-exchange resin and concentrated by evaporation. The pectin is then precipitated by mixing the extract with an appropriate alcohol, e.g., 2-propanol. Finally, the precipitate is separated from the spent alcohol, washed in more alcohol, pressed to drain as much liquid as possible, and then dried and milled. The powder is now ready for

standardization, i.e., mixing with other pectin batches and/or sucrose in order to ensure uniformity. The alcohol is recovered by distillation. An alternative to alcohol precipitation is precipitation by adding appropriate metal salts to the extract. Pectin forms insoluble salts with, e.g., Cu^{2+} and Al^{3+}. The Al^{3+} precipitation [125] was previously used industrially. Removal of the metal ions from the precipitated pectin is done by washing in acidified aqueous alcohol.

Modification. Pectin derived from citrus or apple raw material as described above will normally have a degree of esterification between 55 and 75. A lower degree of esterification can be achieved by acidifying the extract and leaving it for some time before precipitation, or by treating precipitated (but not dried) pectin with acid or alkali during suspension in aqueous alcohol. Ammonia may convert methyl-esterified carboxylate groups of a pectin to primary amides [42], [43]. This is done industrially by suspending precipitated pectin in a mixture of alcohol and water with dissolved ammonia [127], [128]. Deesterification takes place concurrently. By choosing proper conditions with respect to ammonia concentration, water activity, and temperature, pectins with various proportions of amidated, methyl esterified and free carboxylate groups can be produced.

Standardization. The properties of botanical raw materials like those used for pectin fluctuate due to, e.g., weather conditions or sorts variation. Pectin as it appears directly from milling contains this variation. In order to maintain a constant quality, the pectin manufacturer may mix different batches. Further, pectin intended for food is typically diluted with sucrose in order to achieve a uniform grade (i.e., “strength,” for a definition, see Section Measurement and Standardization of Gel-Forming Capacity). Pectin without admixed sugar, e.g., for pharmaceutical purposes, is also available from the major manufacturers.

Due to the multitude of ways in which pectins may vary, it is not possible to ensure batch to batch consistency with respect to all possible attributes at the same time. Pectin is normally standardized with respect to a few properties which are measured in defined chemical systems which simulate the applications, e.g., breaking strength of a gel, gelation temperature, etc. The major pectin manufacturers have developed a great number of specialty types which are tailor-made for individual applications and which have each their set of standardization criteria (control methods). In fact, a pectin type is defined by its set of standardization criteria. When using pectin, it is obviously important to choose a type which has been standardized in a way that corresponds reasonably to the intended use.

3.4. Properties

A range of parameters — intrinsic and extrinsic — are important for the performance of pectin which is in most cases used to impart certain rheological properties, e.g., by forming a gel. The *intrinsic parameters* determine the nature of the gel and may include molecular mass, degree of esterification (DE), degree of acetylation (DA), neutral sugar content, and composition. The *extrinsic parameters* which determine the gelation process may include pectin concentration, pH, ionic strength, water activity, and temperature.

The most important properties of pectin preparations depend on their molecular mass and DE and DA. The proportion and nature of neutral sugars in the side chains as well as in the “ballast” are also of significance. Pectin with a high degree of polymerization is more viscous in solution compared to a pectin with a lower degree of polymerization. Further, the gel strength will typically increase with the degree of polymerization, i.e., less pectin is needed with a high molecular mass pectin [44-46]. The DE strongly influences the functional properties of pectin with the two main groups, HM- and LM-pectin being influenced differently. At the typical conditions in HM-pectin applications, high DE means high gelling temperature whereas at typical use conditions for LM-pectins, low DE means high gelling temperature. No experimental evidence is available to demonstrate the influence of the hairy regions (see Structure.) on the functionality of pectin. It could be speculated that by being bulky and providing kinks in the molecular chain, these parts will prevent molecules from aligning throughout their entire length. This may contribute to preventing precipitation and reducing potential syneresis (spontaneous exudation of solvent from the gel).

Chemical Reactions. By treating pectin with ammonia under alkaline conditions in alcohol suspensions, ca. 20 % of the methyl ester groups are converted to acid amide groups, and amidated pectins are obtained [129]. Amidated pectins have a higher calcium reactivity than LM-pectins, and gels can be obtained with very few Ca^{2+} ions [134]. Carboxyl groups in pectin can be esterified easily with methanol [126], glycol, and glycerol but poorly with ethanol [135]. By using polyols, cross-linked, insoluble systems are obtained. Insoluble pectates can also be prepared by cross-linking with epichlorohydrin; these pectates have ion-exchange properties with a certain selectivity for calcium and heavy-metal ions. They are successfully used for the isolation of pectolytic enzymes [136], [137]. Pectins are readily degraded by oxidants [138] except for chlorite and chlorine dioxide, which can selectively oxidize aldehyde groups at the reducing chain end [139].

3.4.1. *Physical Properties*

Pectin is water-soluble, exhibiting an increased solubility with increasing DE and decreasing degree of polymerization. In order to ensure complete dissolution of pectin, it is necessary that it is properly dispersed without lumping. Once formed, lumps are extremely difficult to dissolve. Pectin, like any other gum or gelling agent, will not dissolve in media where gelling conditions exist. In order to add pectin to complex formulations such as food systems three alternative procedures are recommended: (1) dissolve the pectin in pure water and add the solution; (2) dry blend the pectin with five parts of sugar and add the mixture; (3) disperse the pectin in a liquid in which it is not soluble and add the dispersion.

Aqueous dilute pectin solutions, i.e., with a pectin content below 0.5 % are almost Newtonian whereas more concentrated pectin solutions exhibit pseudoplasticity, i.e., shear-thinning behavior. From dilute viscosity data, the intrinsic viscosity, $[\eta]$ (dL/g) may be determined. $[\eta]$ indicates the hydrodynamic volume of a polymer molecule and depends primarily on the molecular mass, however viscosity is also influenced in a complex manner by many other factors such as the DE, pH (dissociation), and ionic strength. $[\eta]$-values for pectins typically lie in the range of 1.0 to 6.0 dL/g [47-51]. The molecular mass of pectin is often estimated by intrinsic viscosity methods using the Mark – Houwink relationship $[\eta] = KM^{\alpha}$ (see Section Analysis). Originally an α-value of 1.34 [52] was suggested corresponding to a rigid rod-like molecular structure, however more recent findings estimate α to be in the range 0.7 – 0.8 indicating a random coil structure [47], [49], [53], [54].

Pectin is insoluble in most organic solvents such as alcohols and acetone. Pectin can also be precipitated from aqueous solutions by quaternary detergents, water-soluble cationic polymers including proteins, and multivalent cations. LM-pectins can be precipitated by calcium ions; pectates by alkali cations and by acid.

Pectin is a polycarboxylic acid. Dissolved pectin is negatively charged at neutral pH and approaches zero charge at low pH. Since pectin is a polyprotic acid it is not possible to determine an exact value of the apparent dissociation constant, pK_a. Rather, pK_a is different for varying carbohydrate concentrations and for varying degree of dissociation, α. The negative charge density, in turn, is dependent on DE which implies that pK_a-values are increasing with increasing content of unesterified galacturonic acid units. Typically, pK_a-values at 50 % dissociation, i.e., $\alpha = 0.5$, lie in the range 3.5 – 4.5 [55-59]. The usual dependence of pK_a with polymer concentration and ionic strength is observed, i.e., pK_a is lowered with increasing concentration and ionic strength.

3.4.2. *Gel Properties*

Pectin is used mainly as a gelling agent in industry; therefore its gelling properties are most important [4], [6], [106], [117]. The gel formation mechanism of pectin is similar to that of other gelling polysaccharides: Some regions of the polymer molecules associate in junction zones to form a three-dimensional network, which traps the solvent with cosolutes; free stretches of the molecules provide elasticity to the gel obtained [111], [140], [141] (Fig. 6). Irregularities in the pectin molecule, such as the distribution pattern of methyl ester or *O*-acetyl groups, rhamnosyl residues in the backbone, or the presence of side chains, limit the length of the junction zones and give shape to the free stretches of the macromolecules [142].

Figure 6. Network of a gel

A comprehensive, coherent description of pectin gelation does not exist. Conventionally, a distinction is made between HM- and LM-pectin gelation, however, in reality this paradigm is too simple. It applies that different gelation mechanisms may act simultaneously. This is e.g., illustrated by the fact that gelation of LM-pectin, which is normally claimed to gel in the presence of certain metal ions, is further favored by a decrease in pH. If the gelation was solely determined by the formation of calcium-bridges between molecules, the opposite effect would be expected, an increase in pH leading to an increase in gel strength.

In order to adapt the conventional and still widely accepted theories for pectin gelation, the gelation phenomenon will, however, be treated as two distinct mechanisms, ie HM- and LM-pectin gelation. Gels used for jams and jellies are typically formed with HM-pectin at an acidic pH and require the presence of a high concentration of sugar. LM-pectin is typically used for yogurt fruit preparations; these gels can be formed without sugar over a wide pH range, however, the presence of a divalent cation is necessary. In most cases the cation, i.e., calcium, is inherently present in the fruit material. LM-pectin gels can be remelted whereas usually it is not possible to melt an HM-pectin gel, i.e., with HM-pectin gel preparations, the difference between the apparent temperatures of setting and melting is so large that the gels are said to be thermo-irreversible. Further, LM-pectin gels solidify almost immediately after gelling conditions have been introduced, while an HM-pectin gel will build up over time.

HM-pectins form so-called *low-water-activity* or *pectin – sugar – acid gels* and are used in jam, jelly, and marmalade production. The basic galacturonan chain (smooth regions) of the pectin molecule apparently contains blocks with conformational regularity to provide opportunities to build up junction zones. The homogalacturonan part of the molecule is configured as helices with three anhydrogalacturonic acid units per turn, with the methyl ester groups protruding from the helix. According to OAKENFULL and SCOTT [143], junction zones are stabilized by different forces between pectin chain molecules: hydrogen bonds between undissociated carboxyl and secondary alcohol groups, and hydrophobic interactions between methoxyl groups (Fig. 7). Both types of forces are fortified by sucrose; the low pH suppresses the dissociation of carboxyl groups. Sugar reduces the water activity of the system and thereby influences hydrogen bonding by decreasing polymer – water interactions and increasing polymer – polymer interactions. To a certain extent, sugar and acid are interchangeable: at lower sugar concentration, lower pH is required, but at higher sugar concentration, higher pH values are possible and necessary to avoid setting during the boil. The lower limit for the sugar concentration is 55 %. At this concentration, the pH should not be higher than 2.8. At a sugar concentration of 80 % (jellies), the mass will also gel at pH 3.5. This means that within the gelling range, at the same sugar concentration, more acid will give a stronger gel and, at the same pH value, this is achieved by adding more sugar. Addition of urea to a gel cancels out hydrogen bonds and results in a weaker gel with a lower setting temperature [144].

Figure 7. A) Junction zone in high ester pectin gel by hydrogen bonds (dotted lines) and hydrophobic interactions (filled circles); B) Junction zone in low ester pectin gel by dimerization of polygalacturonate (polyguluronate blocks) induced by their strong binding power for Ca^{2+} ions which fill the oxygen-lined cavities between the polysaccharide chains [110]

In general, pectin – sugar – acid gels are prepared by a boiling process followed by cooling. At a certain temperature the system sets to a gel. The food technologist is interested in the rate of setting. At the same rate of cooling the rate of setting determines the setting temperature or the setting time. The parameters that contribute to stronger gels also accelerate the setting rate. Based on the pH limits of gelling as well as on the setting rate, HM-pectins can be subdivided into rapid-set (DM > 70 %) and slow-set (DM 60 – 65 %) pectins. At the same sugar concentration, rapid-set pectins have a higher pH limit for gelling because these very highly methyl esterifiedpectins have few carboxyl groups that must be protonated. A fully methyl esterified pectin gels with sugar alone and does not need acid. The concentration of pectin in the gel influences the rheological properties of the gel, not the gelling rate.

LM-pectins form so-called *calcium pectate gels.* Theories about the chemical structure of calcium pectate gel junction zones were first developed in comparison with alginate gels (see Chapter Alginates). In calcium alginate gels, the junction zones are formed by α-1,4-L-polyguluronate blocks in which the diaxial configuration of the glycosidic linkages leads to a buckled ribbon with limited flexibility and a strong binding power for Ca^{2+} ions, which induce dimerization of alginate chains by filling the oxygen-lined cavities between them. This has evoked the picture of an eggbox, and the expression eggbox-type junction zones has become universally accepted [111], [141], [145], [146] (Fig. 7). In comparing the primary structures of poly-L-guluronate and poly-D-galacturonate (pectin), they are seen to be stereochemically analogous mirror images of each other, except at C-3. The two-fold helix, however, has not been observed with X-ray diffraction techniques, but molecular modeling calculations indicate that it can exist [60]. Circular dichroism data [61] suggest that a conformational change takes place when dissolved or Ca^{2+}-gelled pectate is dried and it is suggested that a transition takes place from a 2_1 ribbon-like to a 3_1 helical symmetry. Similar to the low-water-activity gels, junction zones in calcium pectate gels are terminated by rhamnosyl residues in the backbone, side chains attached to the backbone, or acetyl groups. The presence of some methoxyl groups does not inhibit formation of eggbox-type junction zones. With LM-pectins, acids or sugars are not so important. Gels of acidic fruit juices with LM-pectins can be made by addition of a calcium salt (low-sugar jams) or of milk with its neutral pH and calcium ions (desserts). The amount of calcium necessary for gelation depends on the degree of esterification, the way the LM-pectin has been prepared, and the types and amounts of other ingredients. Coagulation as a result of the addition of calcium salts must be absolutely avoided. Slow availability of Ca^{2+} for the pectin molecules is a prerequisite for obtaining a gel network. This can be accomplished in various ways: (1) an insoluble calcium salt (phosphate, citrate, tartrate) may be used resulting in a slow exchange of Ca^{2+} ions with the LM-pectin and formation of a gel; (2) use of calcium chelating agents such as diphosphates help to retard the availability of Ca^{2+} ions. The fact that calcium pectate gels and precipitates are often thermoreversible (i.e., they are soluble under conditions of gel formation at high temperature) can also be used to advantage. Soluble calcium salts such as calcium lactate or calcium chloride can therefore be added at boiling temperature, and gelling occurs on cooling. LM-pectins can be solubilized in milk by heating; the calcium caseinate of the milk provides the calcium necessary for the system to gel on cooling. Addition of sugars to such gels gives stronger gels; however, at higher concentrations the risk of coagulation increases. A solution to this problem is offered by *amidated pectins*. Gel formation with amidated pectins is also explained by chain associations via eggbox junction zones. Eggbox junction zones can accommodate amide groups, which, however, provide less drive for Ca^{2+} ion binding [146], [147]. On the one hand, they need fewer Ca^{2+} ions for gelation, and on the other hand in the presence of excess Ca^{2+} ions they are not as sensitive to coagulation [148]. In the United States, amidated LM-pectin is used for all applications of LM-pectins.

3.4.3. *Stability and Chemical Reactions*

Stability of pectin molecules in aqueous solution depends upon the temperature and the pH. Pectin has, in contrast to most other hydrocolloids, optimal stability at pH 3.5 to 4. Figure 8 shows the stability of pectin and some other thickening agents at various pH values, expressed as residual viscosity of buffered solutions after heating for 10 min at 90 °C [131]. As is evident from the figure, highly esterified pectin is vulnerable to high pH. Even at pH 5, depolymerization is considerable, in particular at elevated temperatures [133]. Consequently, it is difficult to raise the pH of pectin solutions without causing a decline in the average degree of polymerization, because when trying to mix an alkaline solution into a pectin solution, too high pH cannot be avoided locally. It is recommended to ensure good agitation and low temperature, and to avoid the use of hydroxides.

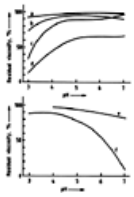

Figure 8. Stability of some polysaccharides at various pHs [131]: Residual viscosity after 10 min incubation at 90 °C

a) Carboxymethylcellulose – methyl cellulose; b) Locust bean gum; c) Agar; d) Carrageenan; e) Pectate; f) Pectin

Low pH hydrolyzes ester bonds causing a decline in DE as well as in the content of *O*-acetyl groups, and it causes a decline in degree of polymerization by hydrolysis of glycosidic bonds [64], in particular at rhamnose insertions in the molecular backbone [65]. Very high acid concentration may degrade galacturonic acid to CO_2, furfural, reductic acid (2,3-dihydroxy-2-cyclopenten-1-one; $C_5H_6O_3$) and alginetin. Carbon dioxide production which is quantitative by boiling in 12 mol/L HCl has in the past been used for the quantitative determination of pectin [85], [132].

High pH depolymerization is due to β-elimination [66] (Fig. 9), it requires the presence of a methyl ester group at the anhydrogalacturonic acid residue which has its 4-C attached to the bond being split. Since the presence of methyl esters is required for β-elimination, vulnerability to this degradation mechanism is related to the DE. High pH further reduces DE (whereby the β-elimination becomes incomplete) as well as the content of *O*-acetyl groups.

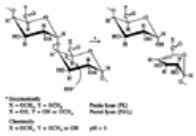

Figure 9. β-Eliminative depolymerization of a galacturonan chain by pectin lyase (PL) or pectate lyase (PAL) or chemically at pH ≥ 5

Derivatization. The low-pH hydrolysis of natural ester linkages may be reversed under conditions of low water activity, e.g., in mixtures of methanol and concentrated sulfuric acid [67], [68], [126], or, for introduction of *O*-acetyl groups, mixtures of concentrated sulfuric acid and acetic anhydride [69]. Further, the carboxyl group may readily be esterified with glycol or glycerol but poorly with ethanol [135]. By using polyols, cross-linked, insoluble systems are obtained.

At high pH, ammonia may convert methyl esterified carboxylate groups to amides [42], [129]. This is used industrially since amidated pectins are of commercial importance.

Insoluble pectates can be prepared by cross-linking with epichlorohydrin; these pectates have ion-exchange properties with a certain selectivity for calcium and heavy-metal ions. They are successfully used for the isolation of pectolytic enzymes [136], [137].

Pectins are readily degraded by oxidants [138] except for chlorite and chlorine dioxide which can selectively oxidize aldehyde groups at the reducing chain end [139].

3.5. Analysis

3.5.1. *Measurement and Standardization of Gel-Forming Capacity*

HM-pectins are generally standardized to uniform strength at specified constant conditions. Expressing the sugar binding capacity of the pectin, the USA – SAG method suggested by the IFT Committee for Pectin Standardization, has been universally accepted for grading HM-pectins [149]. A standard gel is prepared in conical test glasses with the following conditions: soluble solids 65 %, pH 2.20 – 2.40, gel strength 23.5 %, SAG measured with a *ridgelimeter* [106], [153]. After 24 h at 25 ± 3 °C the gel is deposited on a glass plate, and the sagging of the gel under its own weight is measured after 2 min. From the SAG value and the pectin quantity, the grade can be calculated. Most commercial HM-pectins are standardized to 150 grade USA – SAG. (A gel strength of 150 grade SAG implies that 1 part of pectin is able to transform 150 parts of sucrose into a jelly with above standard properties.)

LM-pectin may be standardized by closely analogous procedures, however, no universally accepted method exists. With LM-pectins it is difficult to set up a single universal test because the conditions under which LM-pectins are used may differ widely with respect to soluble solids, calcium content, and pH (see also Standardization.).

3.5.2. *Chemical Analysis [70]*

In addition to gel-forming ability, the analysis of pectin preparations is concerned particularly with the *degree of of esterification, DE*. This is determined by converting the pectin to its acid form by passing it over an ion-exchange resin or washing it in an alcohol suspension, first with hydrochloric acid – alcohol and then with neutral alcohol. The acid and saponification equivalent is determined by titration, and from these values the anhydrogalacturonic acid content and the DE (in percent) can be calculated [63]. This principle for determination of DE has been adopted by the major legislative bodies [71-73]. Presence of *O*-acetyl groups in the pectin will result in an overestimation of DE as well as the anhydrogalacturonic acid content, therefore, it should be evaluated in a separate aliquot of the preparation. The pectin can also be precipitated with copper ions before and after saponification to determine the copper in the well-washed precipitate. From the amount of copper ions bound to the original pectin and the saponified pectin the anhydrogalacturonic acid and the DM can then be calculated without the interference of acetyl groups [151], [152]. Various methods exist for the separate determination of methyl ester and acetyl groups. Methanol is released on saponification of methyl ester and can be determined by GLC either directly [153] or in the headspace of a closed vial after conversion of the methanol to volatile methyl nitrite [154], [155]. *Acetyl groups* can be conveniently determined by an enzymatic spectrophotometric method (supplier Boehringer), by GLC [156], or by distillation and titration after alkaline saponification [157]. A convenient new method is to saponify the pectin preparation in alkaline alcohol, which is then analyzed by HPLC for methanol and acetic acid [158]. Methods measuring the relative content of carboxylate groups to total material, either using size exclusion chromatography with combined detection by conductivity and refractive index [75], or using capillary electrophoresis [76], have also been published. The DE can be inferred from calibration curves if it can be assumed that the anhydrogalacturonic acid content is the same in the samples compared. Previously, the analytical methods used by legislative bodies did not comprise a correction for *O*-acetyl groups since the content of those is small in citrus pectin and modest in apple pectin, but it has now been included in the latest version of the FAO/WHO specification for pectin [71].

In commerical pectins, up to 25 % of the carboxyl groups may be amidated. The *degree of amidation* is calculated from the amount of ammonia released on alkaline distillation [63], [159]. Amidated pectic acids undergo β-elimination reactions, whereas pectic acid does not, which permits quantitative analysis of mixtures of amidated and nonamidated pectins [160]. The anhydrogalacturonide content of pectins can also be determined from aqueous solutions by colorimetric methods with carbazole [161], the more specific *m*-hydroxydiphenyl [162], or sulfamate – *m*-hydroxydiphenyl [163]. Sometimes corrections must be made for interfering compounds (neutral sugars, amide groups, azide). The colorimetric methods can be automated easily and used for routine analyses of large series of samples [130].

With the above-mentioned methods, only the galacturonide residues in the backbone can be analyzed. The *neutral sugars* can be analyzed conveniently by gas chromatography after acid hydrolysis and conversion to volatile derivatives. By a preceding precipitation of pectins with copper

ions the neutral sugars covalently attached to the galacturonan can be analyzed specifically. If starch is present it can be removed by enzymatic degradation [164].

To analyze the *pectin of plant material* the so-called alcohol-insoluble residue is usually prepared first by washing the plant material with refluxing alcohol. This inactivates endogenous enzymes and removes alcohol-soluble constituents. The pectin content is then determined in extracts of the alcohol-insoluble residue. The total pectin is determined in an alkaline extract or in the combined extracts of enzymatic and acid extraction. Another approach is the gradual extraction first with cold or hot water (HM-pectin), oxalate, ethylenediaminetetraacetic acid (EDTA), or cyclohexanediamine tetraacetate (CyDTA) (LM-pectin), and then acid or alkali (protopectin) [165], [166]. The total pectin content and the average degree of methoxylation can also be determined in the alcohol-insoluble residue when this is converted to the acid form before and after saponification. When treated with alcoholic calcium acetate solutions, the free carboxyl groups of pectin set free an equivalent amount of acetic acid that can be determined by titration. An alternative is determination of the bound copper ions from copper solutions [151].

Molecular mass of pectin may be determined with viscosimetry [44], [77], membrane osmometry [78], size exclusion chromatography [47], [48], [54], light scattering [54], [77], [79], [80], ultracentrifugation [34], and analysis of reducing end-groups [80]. Quoted results vary, partly because the intermolecular distribution of molecular mass is broad, partly because pectin molecules aggregate and may contain slight amounts of insoluble material, and, of course, partly because samples are different. As an example, a weight average molecular mass of 90 000 ± 10 000 was reported by HARDING et al. [34]. Quoted values for the Mark – Houwink exponent, relating intrinsic viscosity to molecular mass, are generally in the vicinity of 0.8 [47], [49], [53], [54], suggesting a random coil molecular shape. Integrated systems combining high performance size exclusion chromatography, viscosity detection and light scattering detection are commercially available, including software with which data for molecular mass and Mark – Houwink parameters can be extracted.

3.6. Pharmaceutical and Nutritional Characteristics [167]

Pectin is not significantly degraded in the upper digestive tract of humans, and it can be recovered almost intact from the small intestine [81]. In the cecum and colon, it is fermented by microorganisms mainly to short-chain fatty acids, as can be concluded from in vitro studies [82], rat studies [83] and studies comparing the degradation patterns of humans and rats [84]. Since pectin is a dietary fiber, much attention has been paid to the possible health benefits of pectin which are: (1) reduced glycemic response [86-88], (2) prolonged gastric residence time [89-91], (3) reduced serum cholesterol level [86], and (4) effect against diarrhoea [92]. Pectic polysaccharides from certain botanical sources like ginseng root (*Panax ginseng*) [93], eel grass (*Zosteraceae*), and *Bupleurum falcatum* [94], [95] have shown healing effect on gastric and duodenal ulcers. Most studies, e.g., [97-99], conclude that pectin, in spite of its metal-binding ability, apparently does not inhibit the uptake of minerals from the diet. All of the above effects must be thought of as general tendencies in a vast amount of published findings which are not all mutually consistent. Some discrepancies may be explicable because it is attempted to generalize results achieved with different systems (in vitro, animal, or human) and with different pectins or pectin-rich plant materials. Publications often fail to specify important details about the pectin being used for the study, such as botanical origin, DE, etc.

3.7. Application in the Food Industry [168]

Indigenous manufacture of jams and marmalades, before commercial stabilizers were available, involved the use of pectin-rich fruit and in situ extraction of the pectin during prolonged cooking. Partly owing to this tradition, but mostly due to its superior stability and gelling ability at relatively low pH, pectin is the dominating gelling agent in modern production of jams as well as other products which are gelled, acidulous, and sweet. Examples are jelly fruits, and fruit preparations for industrial production of fruit-containing yogurt (Table 2 groups 1 and 2). Commercial pectins for these applications are tailor-made to yield specific gelation temperatures or gelling rates under specified conditions, and to exhibit specific functionalities such as heat reversibility, heat resistance, firmly gelled textures, pumpable semi-gelled textures, etc. *HM pectin* is used as a stabilizer in yogurt beverages and beverages in which milk proteins are heat-treated at relatively low pH during the production [100-102]. LM pectin finds use for thickening spoonable yogurt (Table 2 group 6). Particles of calcium pectinate are used as a substitute for fat in low-calorie foods [103] (Table 2 group 12). Other applications from Table 2 include groups 4, 8, 9, 10, and 11.

3.8. Market

Annual pectin production is estimated at 25 000 t (80 % citrus pectin), sold mostly in standardized form. Pectin production takes place in Brazil, Denmark, France, Germany, Mexico, Switzerland, and United Kingdom. Smaller amounts are produced in Eastern Europe and the former Soviet Union. Average prices on the world market are $10 – 11 per kilogram HM-pectin and $ 12 – 13 per kilogram LM-pectin.

[Top of Page]

4. Alginates

Alphons C. J. Voragen and Ian Challen

Alginate is a collective term for a family of linear 1,4-linked α-L-gulurono- β-D-mannuronans of widely varying composition and sequential structures [169]. Commercial preparations are usually designated as alginates and include alginic acid, its salts, and derivatives [2], [3], [5], [170-175].

4.1. Occurrence

Alginates form a group of polysaccharides that occur as structural components of the cell walls of brown seaweed (Phaeophyceae; Fig. 10) in which they make up to 40 % of the total dry matter and play an essential role in maintaining the structure of the algal tissue. They were isolated for the first time in 1880 by STANFORD [176], who also introduced the names algin for the soluble substances in an aqueous sodium carbonate extract and alginic acid for the material that could be precipitated from this extract by addition of acid.

Figure 10. Origin of seaweed extracts — general classification

[a] Species of economic significance; [b] Contains only component mentioned; [c] Contains predominantly underlined component

The brown weeds occur in many parts of the world. For example, on the west coast of North America the giant kelp *Macrocystis pyrifera* is harvested mechanically by special ships, which cut the weeds at a depth of 1 m. The weeds grow so fast that they can be harvested several times a year. Other species are harvested semi-mechanically by fishing boats, collected by reaping hooks, harvested from shore at low tide, or collected directly from the beach. Many regions have a tradition of harvesting or collecting brown seaweed, which in the past was used for the extraction of iodine and sodium carbonate.

4.2. Production

Various brown seaweed species are used as raw materials to obtain a range of specific properties (Table 3).

Table 3. Species of brown seaweeds (Phaeophyceae); alginate content and ratio between mannuronic and guluronic acid residues (data from various sources [283])

Species	% Alginate DM *	Man : Gul
Macrocystis pyrifera	13 – 24	1.4 – 1.8
Ascophyllum nodosum	20 – 30	1.4 – 1.9
Laminaria digitata	15 – 27	1.3 – 1.6
Laminaria hyperborea (leaf)	15 – 26	1.3 – 6.0
(stem)	27 – 33	0.4 – 1.0
Ecklonia maxima	30 – 38	1.4 – 1.8

	Mannuronan	Guluronan	Alternating
Macrocystis pyrifera	40.6 %	17.7 %	41.7 %
Ascophyllum nodosum	38.4 %	20.7 %	41.0 %
Laminaria hyperborea	12.7 %	60.5 %	26.8 %
Azetobacter vinelandii (acetylated)	17.8 %	0.5 %	81.7 %

* DM = Dry matter.

The production of alginates is shown schematically in Figure 11. Current processes are still based on patents that were filed in the 1930s by U.S. companies [185], [186]. Alginate production is based on a series of ion-exchange processes: The water-insoluble calcium alginate in the raw material is, for the purpose of extraction and purification, converted first to the soluble sodium form and then, possibly via the calcium form, to insoluble alginic acid, which is neutralized to obtain the finished product.

Figure 11. Flow sheet for the production of sodium alginate [171]

While the *Macrocystis* is taken to the processing factory directly by boat other species are air dried before being transported for processing. The weeds are dried mechanically, ground, and then transported to factories. Salts, which might influence the solubility, and other impurities such as the high-polymeric laminarin (a β-1,3; β-1,6-D-glucan) are removed by washing. The alginate remains in an insoluble form as the calcium salt or as alginic acid. Extraction of the material with cold or hot soda follows, combined with a mechanical disintegration. A homogeneous mass with a pH of ca. 10 – 11 is obtained. After dilution with water, the slurry separates into a liquid phase (sodium alginate) and a solid phase consisting predominantly of cellulose. For removal of these solid impurities, a flotation step may precede the classical separation processes such as sieving and filtration. Alginic acid is obtained from purified sodium alginate by direct precipitation with acid or by precipitation with Ca^{2+} as calcium alginate and subsequent conversion to alginic acid by washing with acid. For the destruction of chloroplasts, either the sodium alginate solution or the calcium alginate precipitate is bleached with hypochlorite. Washed and mechanically dewatered alginic acid can be dried as such; in general, however, it is treated with sodium carbonate or other bases to produce alginate salts. The finished products are highly refined, odorless white powders, which are permitted as food additives. For industrial purposes, less purified preparations are often sufficient.

4.3. Structure

Until 1954, alginate was considered to be composed of polymannuronic acid. In 1955, FISCHER and DÖRFEL [177] discovered L-guluronic acid in addition to D-mannuronic acid, and since ca. 1964, alginic acid has been known to be a copolymer of these two residues. In principle, alginates are composed of three structural elements: the homopolysaccharides α-1,4-L-guluronan and β-1,4-D-mannuronan, and a heteropolysaccharide consisting of alternating 1,4-linked α-L-guluronic (G) and β-D-mannuronic acid (M) residues [178] (Fig. 12). The structure of alginate can therefore be represented schematically as

$$– M – G – M – (M – M)_n – M – G – (M – G)_q – M – G – (G – G)_p – G – M – G –$$

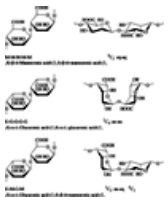

Figure 12. Structural units in alginates

Alginates, like pectins, are linear polymers. The D-mannuronic acid residues are in the $^4C_1$ configuration, and therefore the glycosidic linkage between them is equatorial – equatorial. Since the L-guluronic acid residues are in the $^1C_4$ configuration, they are connected by axial – axial glycosidic linkages similar to the linkages between galacturonosyl residues in pectin. However, unlike pectin, alginates do not contain neutral sugar residues. The properties of alginates are determined largely by their molecular mass and by the ratio in which the three structural elements occur in the polymer. This ratio is determined by the variety of seaweed from which the alginate is extracted [172], [175], [179-182].

The *biosynthesis in plants* starts from D-mannose which, via guanosine diphosphate mannose, is oxidized to guanosine diphosphate mannuronic acid by a NAD dehydrogenase system and then polymerized to polymannuronic acid by a transferase enzyme. In homogenates of brown seaweed, guanosine diphosphate guluronic acid has also been identified, so at the end of the pathway copolymerization may also occur [173].

4.4. Properties [188]

Solubility. *Alginic acid* itself is insoluble in water and precipitates at pH < 3.5. The ammonium, alkali and magnesium salts, and alginate salts with

organic bases are water soluble; calcium ions and other multivalent ions precipitate with alginates. *Alginates* are insoluble in organic solvents and can be precipitated with alcohols. They are, however, slightly miscible with simple alcohols, glycerol, and glycols, and these compounds can therefore be included to a certain content in alginate solutions. The dissociation constant p*K* of monomeric mannuronic acid is 3.38, and of guluronic acid 3.65. Alginic acid with a high proportion of guluronan has a p*K* of 3.74, when the proportion of mannuronan is high, this value is 3.42.

Degree of Polymerization. The properties of alginates depend largely on the degree of polymerization and the ratio of guluronan : mannuronan blocks in the molecules. The mannuronan regions in the molecules with their diequatorial glycosidic linkages are stretched and flat, while the diaxial glycosidic linkages in the guluronan regions give this part of the alginate molecule a buckled ribbon shape with limited flexibility.

Viscosity. The high viscosity of aqueous alginate solutions is a function of the degree of polymerization and is one of their most important properties (Table 4).

Table 4. Dynamic viscosities (in mPa · s) of polysaccharide solutions at different concentrations (data from various sources [283], [41], [62], [63]) *

Concentration	Pectin	Alginate	Carrageenan	Agar	Gum arabic	Gum tragacanth	Gum karaya	Gum ghatti	Xanthan	Locust bean gum	Guar gum
1 %	50	214	57	4		54	3 000	6	2 000	59	3 025
2 %	200	3 760	397	25		906	8 500	10	7 000	1 114	25 060
3 %	550	29 400	4 411			10 605	20 000	40	11 500	8 260	111 150
4 %		39 660	25 356	400		440 265	30 000	60		39 660	302 500
5 %			51 425		7	111 000	45 000			121 000	510 000
10 %					17						
30 %					200						
50 %					4 163						

* Reported value give only general trends. Depending on preparation and measuring conditions differences may occur.

By mixing alginates of different sources, and by oxidative degradation, manufacturers can offer products with viscosities varying from 4 – 1000 mPa · s for a 1 % solution. Viscosity numbers alone are meaningless without the shear stress or shear rate because alginate solutions are Newtonian fluids only at very low shear rates and low concentrations. Under other conditions, they are pseudoplastic. Table 4 also shows the strong concentration dependence of viscosity. Cations have a significant effect on viscosity. Low concentrations of Ca^{2+} ions considerably increase the viscosity as a result of complex formation. In measuring the viscosity of alginate solutions a distinction is made between the “direct” viscosity and the “sequestered” viscosity. The latter is always lower because a sequestrant removes the Ca^{2+} ions.

Effect of Shear. In common with most other food hydrocolloids, sodium alginate solutions are pseudoplastic. In solution the alginate molecule is long, thin, and rigid. When the solution is at rest, the alginate molecules interfere with each other forming loose entanglements. When shear is applied to the solution, the molecules are oriented in the direction of the shearing force. The association between the molecules is greatly reduced and there are no longer any entanglements. The result is an immediate lowering of the viscosity. When the shear is removed, the molecules entangle again resulting in the recovery of the viscosity.

Effect of pH. Between about pH 4.5 and 10, there is little change in solution viscosity. At about pH 3.8 the solution viscosity increases dramatically and, as the pH is reduced further the alginate is precipitated out of solution as alginic acid. At higher pHs viscosity is lost by a β-elimination process. The glycosidic linkages between mannuronic and guluronic acid residues are known to be less stable to acid hydrolysis than the linkages between two mannuronic or two guluronic acid residues [190]. The homopolymers are slightly more stable than heteropolymers.

Effect of Sequestrants. Sequestrants are used with alginates to soften the water so that the alginate can hydrate completely and thus modify the rate of the calcium – alginate reaction. They may also be used to sequester the calcium ions naturally present in the alginate itself. Different sequestrants bind different amounts of calcium ions at different pHs. Therefore, it is important to understand which sequestrant is being used and at what pH it will be used at.

Effect of Salts. The presence of *monovalent salts* reduce the viscosity of sodium alginate solutions. The viscosity drop occurs because alginates, in common with other polyelectrolytes, contract as the ionic strength increases.

As *polyvalent cations*, usually calcium ions, are gradually introduced into a sodium alginate solution, they cause the alginate polymer chains to start to align — this is seen as an increase in the solution viscosity. As more calcium is added the solution viscosity increases further to a point where there is evidence of some gel structure. At this stage, the solution shows thixotropic properties — the viscosity decreases when sheared, pumped, or shaken but the structure is regained when the shear is removed. Higher levels of calcium ions result in coherent gel formation. Typically, this occurs when more than about 70 % of the alginate has been converted into the calcium alginate form. The introduction of even more calcium ions increases the strength of the gel until, finally, all of the alginate is precipitated out of solution. At this stage all solution viscosity is lost.

Gel Properties. In food applications, sodium alginates are used as gelling agents, which in many aspects resemble LM-pectins and also form gels with Ca^{2+} ions. As pure polyuronic acids, alginates are, however, more sensitive to acid than pectins and are insoluble at pH < 3.5. Alginates also have a higher sensitivity to calcium than pectins, which is probably due to their higher degree of polymerization. For the preparation of acid gels, calcium salts are used that are soluble only under acidic conditions, together with compounds that release H^+ ions slowly, for instance, slowly soluble fumaric acid or slowly hydrolyzing glucono-1,5-lactone. The use of sequestrants has also been mentioned. The gel-forming ability of alginates is related mainly to the proportion of L-guluronan blocks. These blocks bind with Ca^{2+} ions according to a cooperative mechanism, giving rise to the formation of junction zones, which are referred to as the eggbox model (see Section Physical Properties). The junction zones in alginate gels are stronger than those in LM-pectin gels, and they are not thermally reversible even in the presence of sugar. Alginates rich in guluronan blocks form strong, brittle gels with a tendency to syneresis, whereas alginates rich in mannuronic acid form weaker more elastic gels that are less prone to syneresis [193]. Single mannuronic acid residues or mannuronan blocks interrupt the guluronan junction zones. *Propylene glycol alginates* form weaker gels. They can be prepared more easily because propylene glycol alginate is less sensitive to acid and Ca^{2+} ions and, therefore, less liable to coagulation. Alginates also gel in the presence of equimolar quantities of an HM-pectin. The strongest gels are obtained with HM-pectin and alginates with a high proportion of guluronic acid residues at low pH and in the absence of calcium ions. Sugar is not essential for gelation; but it does affect the gel properties [144], [194], [195].

Oxidation and Degradation. By using high acid concentrations and high temperatures, alginates and alginate esters can be quantitatively decarboxylated. Oxidants such as halogens or periodate, and redox systems such as polyphenols, ascorbic acid, or thiols, are able to depolymerize alginates. For production of low viscosity alginates, hydrogen peroxide is used. Bacterial alginate depolymerases have also been described [173],

[191]. These enzymes have been used to study the distribution of L-guluronic and D-mannuronic residues along the backbone. They have no commercial significance; however, they condition the chemical conservation of alginate solutions. The degree of polymerization of commercial alginates is in the range of ca. 80 to ca. 800.

4.5. Propylene Glycol (Propane-1,2-diol) Alginate

Propylene glycol alginate (pga) is prepared by treating partially neutralized alginic acid with propylene oxide [187] (Fig. 13).

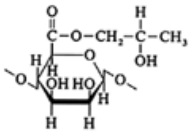

Figure 13. Propylene glycol mannuronate unit

Pga's are completely soluble in both hot and cold water at neutral pH. Depending on the degree of esterification, pga can be water soluble at a pH as low as 2.5. As with the soluble alginate salts, the viscosity of a pga solution is a function of the degree of polymerization. Typically, pgas are available in the viscosity range 25 to 500 mPa · s for a 1 % solution and degrees of esterification from about 50 % to about 90 %. Highly esterified products cannot be made with highly polymerized alginates due to steric hindrance.

The flow properties of pga solutions depend upon the gum concentration and the shear rate: at high concentrations (e.g., > 1 %) they are shear thinning over a wide range of shear rates. At low concentrations they exhibit long, smooth flowing characteristics.

With increasing temperature the viscosity – shear rate curves are approximately linear over a wide range of shear rates and show progressively less pseudoplasticity with increasing temperature. This effect is reversed on cooling.

The bulky propylene glycol groups are responsible for the other different properties of the molecule. They physically hinder the aggregation of the polymer chains and also afford protection to the β,1–4 linkages. As a consequence of esterification, fewer carboxyl groups are available for interaction with divalent metal cations such as calcium. Propylene glycol alginates with less than 60 % esterification will react with calcium ions to give enhanced viscosity — such solutions are thixotropic but there is no evidence of gel formation or precipitation. Highly esterified pgas are not affected by the presence of calcium ions.

Propylene glycol alginates are tolerant to ionic materials such as salt and acid. The property of salt tolerance is made use of in barbecue sauces, salad dressings, and other liquid products. The esterification process reduces the ionic character of the pga molecule, thereby increasing its tolerance to low pH. The higher the degree of esterification the more tolerant the pga is to lower pH, but in strong acid the alginate chains depolymerize, with a resultant loss of viscosity [171].

In acid solutions proteins behave as bases and can interact with pga. This leads to some thickening and a degree of solution thixotropy, i.e., the solution exhibits high viscosity at rest but low viscosity when shear is applied. This property is used in meringues, cakes, and in the stabilization of beer foam.

The lipophilic ester side chains interact with fatty substances. Propylene glycol alginates build a molecular film around oil droplets thereby enhancing emulsification by inhibiting coalescence and droplet growth. Thus pgas provide dual functionality — they act as emulsifiers and show other rheological benefits. This film-forming property is also reflected in pga's abiltiy to form stable foams in aqueous solution. Thus, they are useful in protein-containing systems such as marshmallows and the stabilization of beer foam.

Legal Aspects. The *safety* of pga was evaluated by the Joint FAD/WHO Expert Committee on Food Additives (JECFA) in 1971, when it was assigned an Acceptable Daily Intake (ADI) of 0.25 mg per kilogram body weight. Pga has been an EC-approved food additive for over 25 years. It is listed as E405, in the Emulsifiers, Stabilisers, Thickeners and Gelling Agents Directive, 74/329/EEC.

4.6. Bacterial Alginates

Alginate production has also been reported in bacteria (i.e., *Pseudomonas aeruginosa* (pathogenic), *P. mendocina, P. putida, P. fluorescens*, and *Azetobacter vinelandii*). These alginates are also formed of the same three structural elements and are further characterized by the presence of *O*-acetyl groups. For *A. vinelandii* it has been shown that polymannuronic acid is synthesized first; mannuronic acid residues can then be transformed to guluronic acid residues by a mannuronan C-5 epimerase. With this enzyme both bacterial and algal alginates can be epimerized. The activity of the enzyme depends strongly on the Ca^{2+} concentration [169], [183], [184].

4.7. Analysis

The alginate content in alginate preparations can be determined by titration as described for pectins. From this analysis the degree of esterification can be deduced if propylene glycol alginate is present. In mixtures or foods, the uronic acid content can be determined by the colorimetric methods used for pectins [162]. If necessary this analysis can be preceded by a purification step. Titration as well as colorimetric methods do not, however, distinguish between pectins and alginates. Pectins can be degraded in solution with specific enzymes whereupon they become alcohol soluble. Alginates can then be determined in the alcohol-insoluble residue. Polyuronides can also be precipitated from preparations, mixtures, or foods with Ca^{2+} ions and thus separated from other polysaccharides. After hydrolysis to monomeric uronic acid residues [181], the latter can be analyzed by gas chromatography [29] or HPLC [196]. This method also allows analysis of the uronic acid composition of alginates and, after partial hydrolysis, the proportions of mannuronan and guluronan blocks. The proportions of M and G residues in an alginate preparation and their distribution in M, G, and MG blocks can also be determined by ^{1}H-NMR [182], [197] and circular dichroism [198]. As a result of their wide range of applications, no internationally recognized standard methods exist to assess the quality of alginates. Manufacturers therefore usually specify the viscosity or gel strength of their products by methods agreed upon with clients.

4.8. Applications

Food Industry. The applications of *alginates* in the food industry are based on

1. Their potential to form highly viscous solutions with outstanding suspension-stabilizing properties
2. Their stability at high temperature and high pH
3. Their reactivity with Ca^{2+} ions, which enables them to gel
4. The thermal stability of these gels [168], [171]

As a result alginates are used in almost all groups listed in Table 2. For acid products, *propylene glycol alginates* are used. In all of these applications, lump formation must be considered when adding the alginate to an aqueous system, similar to pectins. When sugar is part of the ingredients, some may be admixed with the alginate. When the application allows the addition of alginate as a solution, the alginate can best be solubilized by dosing it gradually in the vortex of a powerful mixer. For laboratory use the alginate can be moistened with alcohol. For all listed applications, only a few grams of alginate are added per kilogram of product; for the stabilization of foams a few milligrams per kilogram suffice.

Other Applications. Twice as much alginate is used in applications outside the food industry. Except for groups 17, 20, 23, and 26, all applications

in Table 2 have been described. Of greatest impact is group 31, which claims to cover ca. one-third of alginate sales. Groups 27 and 29 also represent a substantial market. One of the first applications of alginates was in the treatment of boiler water (group 32). The calcium is flocculated as calcium alginate and not deposited as scale. For this application, alginate paste is used, which does not have to be refined to the extent necessary for food applications. The strong swelling power of alginic acid makes it a good binding agent for tablets (group 24), and its calcium reactivity enables denture replicas to be made (group 25). The triethanolamine salt is often used as a thickening agent in cosmetic and pharmaceutical ointments. Another interesting application mentioned in the commercial literature is as a gel for embedding the roots of plants to protect them from mechanical damage and from dehydration after replanting. In the last decade alginates have been used in biotechnology research to immobilize bioactive cells and to purify enzymes [137].

4.9. Market

The alginate industry is limited to a small number of companies. Production plants are found in locations where suitable weed species are present (i.e., North America, Scotland, Brittany, and Japan). Alginate is also produced in the former Soviet Union. About 25 000 t of alginate is produced annually; ca. one-third of this is used in the food industry. The price of sodium alginate for food applications is ca. $ 9.0 – $ 15 per kilogram.

[Top of Page]

5. Carrageenan

Alphons C. J. Voragen and Svein Halvor Knutsen

Carrageenans belong to a family of polydisperse long chain galactans, which can be extracted from red seaweeds (Rhodophyta) [304], [305]. The major uses are as high quality ingredients in food and cosmetics with estimated worldwide sales approximately $ 263 × 10^6 in 1997 [306].

5.1. Structure

Carrageenans are build up by alternating 1,3-linked β-galactose (**G**-units) and 1,4-linked α-galactose (**D**-units) that can be partly esterified with sulfuric acid (**S**). In carrageenans with the ability to form a gel the major part of the 4-linked units are in the 3,6-anhydro form (**DA**). A short hand nomenclature system based on letters has been introduced [307] to simplify the old system based on Greek letters [308], [309].

Historically the three major commercial carrageenans types were named κ, ι, and λ with their corresponding structures and short names given in Table 5. Originally the two former (gelling family) were isolated based on their insolubility in KCl [310], whereas the latter formed the soluble fraction (nongelling family). These major types are representation in Figure 14. In the text the Greek nomenclature normally will be used for the three common commercial types only, whereas the letter system will be used to specify structural characteristics when appropriate.

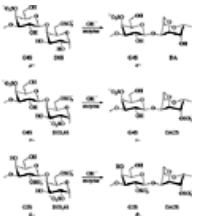

Figure 14. The repeating structures of the most common commercial carrageenans and the relation to their 6-sulfated analogues

Table 5. Most common carrageenan structures in commercial samples

4-linked unit	3-linked unit	Greek letter name
DA	**G4S**	kappa (κ)
DA2S-G4S	**G4S**	iota (ι)
D2S,6S-G2S	**G2S**	lambda (λ)
DA	**G**	beta (β)

There are several biological factors that complicate the structural picture of carrageenans.

In general the **DA** (in κ-carrageenan) or the **DA2S** (in ι-carrageenan) might be replaced with their 6-sulfated precursor residues, **D6S** (μ-carrageenan) and **D2S,6S** (ν-carrageenan), respectively [311]. The conversion of precursors to 3,6-anhydrogalactose (**DA** or **DA2S**) occurs during biosynthesis in the algae [312], but can also be catalyzed by hot alkali during the commercial phycocolloid production. During the alkali treatment or the enzyme catalyzed elimination reaction the $^4C_1$ conformation of the 4-linked 6-sulfated precursor units (**D2S,6S** or **D6S**) is transformed into $^1C_4$ in 3,6-anhydrogalactose with the liberation of sulfate from position 6. The desulfatase enzyme has also been called “dekinkase” because it effects a change in the configuration of the transformed galactosyl unit and removes “kinks” present as equatorial-axial glycosidic linkages [204]. Some algae have a gametophyte generation producing carrageenans of the gelling family, and a sporophyte generation (both often identical in appearance), producing **D2S,6S-G2S** (λ) or structurally related carrageenans [313]. Such carrageenans have traditionally been ill defined due to difficulties in their characterization and are mostly obtained from *Chondrus* and *Gigartina* genera [306]. λ-Carrageenan can most probably only be transformed into a component containing a 3,6-anhydro bridge (**DA2S-G2S**, termed θ-carrageenan) during a hot alkaline treatment. Enzymes capable of catalyzing this reaction in the algae have not been detected and naturally occurring **DA2S-G2S** is rare [314].

In general the repeating sequence that dominates the structural character of a carrageenan is the basis for any nomenclature system. However, it should be stressed that the idealized structures most often are masked by precursors or other carrageenan elements. In fact by selecting the algal source carrageenans with wide **DA/DA2S** ratio can be obtained [315-317]. In addition to the carrageenan mentioned above in Table 5, which are the main constituents of commercial preparations, different carrageenan structures are constantly isolated from new sources [309], [314], [318] and/or found in small quantities in regular samples. Finally, algae containing molecules with a mixed carrageenan and agar (4-linked D-galactose replaced with 4-linked L-galactose) structure have been detected [319], [320], [321]. So far these extracts are not economically important.

In addition to sulfate, substitution by methyl ethers and pyruvate or other sugars, such as xylose, glucose, and uronic acid, may be present in a carrageenan sample. Nitrogen containing components such as proteins and pigments are common impurities (0.1–1 wt %). Since the viscosity of several commercial carrageenan solution can be reduced by the addition of protein degrading enzymes, some possible linkage to the carrageenan chain is suspected [322]. Carrageenans may therefore be considered as proteoglycans [304].

To complete the description of the carrageenan structure, the mean number of constituent sugars, their weight average molecular mass M_w, and the distribution of the molecules (polydispersity factor = $\langle M_w \rangle / \langle M_n \rangle$ must be given. Representative values are in the range of 300 000–600 000 and 2.3–5.1, respectively [305].

5.2. Sources and Raw Materials

Carrageenan has been used for several hundred years in Europe and the Far East, and the name is linked to the Irish village Carraghen, where Irish Moss (*Chondrus crispus*) was harvested and utilized in milk products [323]. During the last years the harvesting of natural Irish Moss populations has been reduced and is being discussed as an environmental issue in several countries. Natural resources are mostly harvested in temperate regions (Canada, Chile, and France) and at present the largest consumption is based on cultivated tropical seaweeds such as *Kappapycus alvarezii* [306]. For the major commercial products and sources see Table 6.

Table 6. Examples of carrageenans that are present in neutral* or alkali-treated extracts. n is restricted to the gametophytic generation and 2n is restricted to the sporophytic generation. The species considered as commercially most important are in bold phase.

Dominating structure (s)	Generic examples
DA-G4S	***Kappaphycus alvarezii***[a],*Hypnea musciformis*, ***Chondrus crispus***(*n*), *Chondrus ocellatus (n)*, *Gigartina tenella(n)*
DA2S-G4S	***Eucheuma denticulatum***[b], *Agardhiella tenera*, *Gymnogongrus concinna*,
DA2S-G4S and **DA-G4S**[c]	***Sarcothalia crispate***[d](*n*), ***Gigartina skottsbergii***[e](*n*)
DA-G4S ≈ **DA-G**	***Furcellaria lumbricalis***[f]
DA-G > **DA-G4S**	*Eucheuma gelatinae*
D2S,6S-G2S*[g]	***Chondrus crispus***[h](2*n*), ***Gigartina pistillata***[i](2*n*), *Sarcothalia crispate(2n)*, *Gigartina skottsbergii(2n)*
DA2S-G2S[j]	*Sarcothalia crispate(2n), Gigartina skottsbergii(2n)*
DA-G6S	*Phyllophora sp, Risoella verruculosa*

[a]Trade name (Eucheuma) cottonii or cottoni.
[b]Trade name (Euchuma) spinosum.
[c]Copolymers referred to as kappa-2 with weaker and less milk reactivity than **DA-G4S**. For a full discussion of several species see [317].
[d]Sandpaper [324].
[e]Pigskin [324].
[f]Danish agar, Furcellaran or κ-furcellaran.
[g]May contain other structures such as pyruvated residues (**GP,2S**).
[h]Cultivated in Canada.
[i]Obtained from Morocco.
[j]Occur only in alkali treated sporophytic extracts. For a more detailed explanation of nomenclature see [305], [307]. Information mostly extracted from [306] and [317].

5.3. Production

After harvesting the seaweeds material is washed to remove sand and stones and dried to a dry matter content of ca. 25 wt %. The carrageenan content can vary from 15–70 %, depending on the seaweed. The production of refined carrageenan is shown schematically in Figure 15. The dried algae are treated with alkali and ground to a paste. Alkaline conditions facilitate extraction of the macerated algae, retard acid-catalyzed depolymerization of the galactan and catalyze the conversion of C-6 sulfated precursor residues to 3,6-anhydrogalactopyranosyl residues, thus increasing the gelling potential of the product. The crude extracts are further purified by sieving and filtration by using a filter aid such as celite. If drum drying is anticipated, pigments are first removed with activated carbon. To prevent gelling, all of these operations must be carried out at higher temperatures.

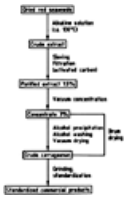

Figure 15. Flow sheet for the production of carrageenans

Because as with pectin and alginates, dilute extracts must be processed to a dry product, the energy cost for dewatering and recovery of alcohol is an important factor. Isopropanol is used as the alcohol for precipitation of the carrageenan. An alternative process is available for κ-carrageenan by extruding the extract into a KCl solution, pressing the precipitate and removing water by freeze-thaw cycle. After drying and grinding several batches might be blended and/or standardized by the addition of salts and sugars. Furthermore, by selecting a milder alkaline extraction process than the conventional alkali modification, carrageenans with a remaining small fraction of precursor units and hence altered function might be obtained [324]. The most innovative processing procedures include enzymic tools based on molecular biology, mostly developed at the Goemar Laboratories, (Roscoff, France). Both stable enzyme preparations of glycosylhydrolases—to produce carrageenan oligosaccharides—as well as sulfohydrolases — to convert precursor residues into 3,6-anhydrogalactose on the polymer level have been cloned [325]. In the late 1990s [326] a semirefined carrageenan has been offered on the market. The product, Processed Euchuma Seaweed (PES), is regulated as E407b. Such carrageenan products are obtained by first treating the algae with hot KOH solutions, washing with water, and finally drying, bleaching, and grinding to a suitable particle size [327]. Under the conditions used the carrageenan is not extracted and the precursors, if present, are converted to 3,6-anhydrogalactose inside the algal material. PES will contain more acid insoluble matters and fiber components such as cellulose [328].

5.4. Analysis

Carrageenan preparations can be separated based on their solubility in aqueous KCl and further analyzed by several tools. The content of sulfate can be determined by gravimetry (FAO, 1992) and the content of 3,6-anhydrogalactose by the resorcinol reagent [329]. The content of both constituents can be determined by HPLC [330]. By a simple recording of the increase in the 3,6-anhydrogalactose content by the recorcinol reagent upon a hot alkaline treatment the level of precursors can be evaluated in a carrageenan sample.

The constituent sugars can be analyzed by *gas chromatography* after the introduction of a specialized hydrolysis and derivatization procedure to conserve the 3,6-anhydro units. This reductive hydrolysis in the presence of methylmorpholine-borane [363] can even be applied directly to seaweed samples for evaluation of carrageenan structure and taxonomy [331]. The use of partial reductive hydrolysis after methylation can discriminate between agaran and carrageenan backbone and its substitution [332]. *IR-spectroscopy* can be applied to study the position of sulfation but has some limitation for quantitative analysis. This technique can be applied directly to carrageenan [333] or to dried and milled material by the use of FTIR diffuse reflectance spectroscopy [334].

By far the most powerful technique for carrageenan analysis is *NMR spectroscopy* [335]. ^{13}*C-NMR* spectroscopy is preferred for identifying the different residues in a carrageenan sample since most of the signals (resonances) from the residues and even some signal splitting due to sequence effects are well separated [336]. However, a low signal to noise ratio is encountered and components below 5 % may not be detected. For ^{1}*H-NMR* the sensitivity is better but a distinction between the different components is more difficult since the different signals may overlap in a quite narrow region of the spectra. For routine applications ^{1}H-NMR can be applied to 1 mg samples and analyzed within 15 min [316], [337], [338]. Spectroscopy can also be used for the estimation of carrageenan in foods [339]. Combined with the selective degradation with carrageenase [311] and chromatography the major as well as the minor constituents may be identified [340]. Apart from data for sugar composition and sulfate substitution an important parameter for carrageenans is their molecular mass. At present the common method is HPLC based *size exclusion chromatography* (HPSEC), coupled to a Multi-Angle Laser-Light Scattering (MALLS) detector, absolute mass and the molecular mass distribution can be obtained [341], [342]. Typical tests to demonstrate the presence of carrageenan are the formation of a blue coagulate with methylene blue, positive sulfate reaction after acid hydrolysis, and the formation of much stronger gels by the addition of potassium chloride. Recent progress in the analysis of carrageenans in food is reviewed elsewhere [343].

5.5. Properties

Gelation. All forms of λ-carrageenan as well as the sodium salts of κ- and ι-carrageenan are soluble in cold water. The potassium and calcium salts of κ- and ι-carrageenans, however, dissolve only at 70 °C and form gels or—depending on the ionic strength—viscous systems upon cooling (Table 4). κ-Carrageenan is more sensitive to potassium ions, whereas ι-Carrageenan is more sensitive to calcium ions. Once a highly potassium sensitive carrageenan resembling κ-carrageenan was produced from *Furcellaria*, i.e., furcellaran or Danish agar [344]. This carrageenan has a mixed **DA-G4S** and **DA-G** structure. However, since this free floating algae became extinct from Danish waters the holdfast variety is at present collected in Canada and processed together with *Chondrus crispus*. From a solution of the three major types of carrageenans, κ- and ι-carrageenan can be precipitated by potassium chloride. To obtain a gel the carrageenan molecules must undergo a transition from a random coil structure into helices that aggregate upon cooling. In general the ion induced network formation [345] is an intermolecular association that requires a minimum degree of polymerization of about 100 [346]. It is still under debate whether the fundamental ordered state is a single or a double-stranded helix. Results obtained by high-performance size exclusion chromatography-low angle light scattering (HPSEC-LALLS) (see Section Analysis) originally suggested a doubling of the molecular mass for the ordered conformation [341], [347] and hence double helixes (Fig. 16). At present the common method is HPLC based Size Exclusion Chromatography, coupled to a Low- or a Multi-Angle Laser-Light Scattering (LALLS or MALLS, respectively) detector.

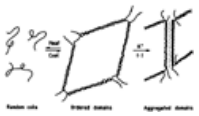

Figure 16. Gelation of carrageenans by formation of double helices and aggregation of helices [348]

However, the findings based on light scattering data have been questioned due to the preparation method of the solution and the concentrations used, and result have been presented in favor of single helixes of ι-carrageenan [349]. The carrageenan gels are thermoreversible and upon heating, the helices unfold, the molecules go in solution again as random coils and the gel melts. In the gel state the aggregation of helices may continue, the network contracts, and the gel becomes brittle (short) and shows syneresis. The inability of λ-*carrageenan* to form gels can be explained by the occurrence of **D2S,6S** units in its backbone. A solubilized preparation rich in κ- *carrageenan* gives a firm gel with a setting point of 40 °C and a melting point of 55 °C at a concentration of 0.5 %, in the presence of 0.2 % potassium salt. The gel strength increases strongly with increasing concentration of potassium ions; however, potassium ion concentrations > 0.2 % are unacceptable because of taste. Also, addition of sugar increases the gel strength. Both additions also increase the setting temperature, as well as the melting temperature, of the gels. The hysteresis, i.e. the difference between melting and setting temperature, remains small and is ca. 15 °C. The gels are brittle and have a tendency to become opaque and show syneresis. This can be prevented by adding ι-carrageenan. Gels of ι-*carrageenan* alone are transparent, they show no syneresis and little hysteresis. Due to the presence of **DA2S** residues, which may act as a wedging group and prevent tight aggregation of double helices, the gels are rather weak. Furthermore, the observation that gelation of a commercial ι-carrageenan (**DA2S**-units) showed a small specificity towards monovalent cations was interpreted as being due to the inclusion of a small proportion of κ-carrageenan (**DA**-units) [350].

Viscosity. In a solution with low ionic strength the carrageenan chains are extended due to the electrostatic repulsion of the negatively charged sulfate groups and the resulting solutions are highly viscous. The viscosity increases with concentration and decreases by addition of salts due to charge shielding, and will decrease exponential upon heating. Viscosity will be reasonable reversible upon cooling as long as the condition for maximum stability of the molecular mass is kept (pH 9). The viscosity of soluble carrageenan forms can be measured in 1 % solutions at room temperature, but because of gelation at intermediate temperature, viscosity is normally measured in 1.5 % solution at 75 °C. At this temperature values up to 800 mPa·s have been measured for a 1.5 % solution. In industrial applications measurements are performed with rotational viscometers. Since carrageenan solutions have pseudoplastic behavior the shear rate must also be reported. The molecular mass (M_w) dependence of the intrinsic viscosity [η] is given by the Mark-Houwink equation:

$$[\eta] = K(M_w)^a$$

The constants *K* and *a* are dependent of the carrageenan type and the solvent. A value for *a* close to unity indicates rigid rod-like structure [351].

Synergistic Effect with Other Gums. Of particular interest is the interaction of κ-carrageenan with locust bean gum galactomannan. By partially replacing the κ-carrageenan with locust bean gum, which does not gel on its own, a stronger gel with improved properties is obtained. It becomes more elastic, and has a lower tendency to syneresis and to become opaque. Similar observations were made with konjac glucomannans. The smooth regions of the mannan chain (i.e., regions with no galactose or glucose side groups) are thought to bind to the double helices of the κ-carrageenan, forming mixed junction zones [207], (Fig. 17). These mixed junction zones have a lower tendency to form tightly packed aggregates. The industry has placed preparations on the market that exhibit constant gel strength and gel characteristics by blending different lots and adding potassium salts, sugar, and possibly locust bean gum. Because carrageenans may form lumps when added to water, premixing with sugar or adding the carrageenan in a powerful mixer is recommended. Carrageenans are also useful in altering the textural properties of a starch system. Adding *i*-carrageenan (0.5 %) to a starch system increases the viscosity as much as 10 times, whereas no effect is obtained from *k*-carrageenan [352].

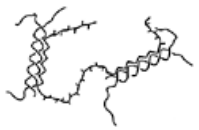

Figure 17. Synergistic gelation between helix-forming carrageenans and agar with galactomannans by direct association of smooth regions of the mannan chain with double helices in the seaweed polysaccharides [209]

Stability. Carrageenans are not stable under acidic conditions, because the 3,6-anhydro ring and the 1 ⟶3 linkages are easily hydrolyzed. The substitution with 2-sulfate (**DA2S**) introduces some stability. Carrageenan solutions can be sterilized at pH > 4.5 but some reduction in molecular mass will occur [353]. In general decrease in molecular mass might be encountered during dialysis and due to autohydrolysis during freeze-drying or long storage. Gelled carrageenans are more stable and the common microorganisms found outside the marine environment do in general not degrade carrageenans.

Reactivity with Proteins. Carrageenans possess a strong anionic character because of their sulfate groups. These charges as well as associated

ions (Na^+ versus K^+ and Ca^{2+}) and the conformation of the sugars in the chain determine the properties of carrageenans. Reactivity with proteins can be observed both with carrageenan of the gelling and nongelling family, but some chain regularity is important in different types of interactions [351]:

1. Below the pH of the isoelectric point (I-pH) of the protein the positively charged protein and the negatively carrageenan form a complex, which might result in a precipitate depending on the net charge ratio.
2. Above I-pH, when both polymers are negatively charged, the interactions are mediated by polyvalent cations such as Ca^{2+}.
3. Finally, interaction with a positively charged part of a molecule with a net negative charge may occur.

In milk systems a highly specific interaction between κ-casein and carrageenans has been established that causes κ-casein particles or casein micelles to attach to the carrageenan chain. When the molecular mass of the carrageenans is sufficient, helical regions can form and aggregate and a gel network is obtained. This can occur only with the gel forming κ- and ι-types. Addition of 0.2 % κ-carrageenan to milk at the pasteurization temperature and cooling to 43 °C, gives a firm gel that melts again at 60 °C.

Weak networks form at carrageenan concentrations as low as 0.02 %, which can fix casein particles. In chocolate milk, for instance, this network holds the cocoa suspension and, in creams the lipid globules. It is suggested that both type (2) and (3) are important in the gelation of milk and to stabilize chocolate milk. The ability to stabilize casein micelles is in the order **DA-G4S**>**DA2S-G2S**>**D2S,6S-G2S** [351]. The reaction between milk proteins and carrageenan may synergistically increase the gel strength about 10 times, and carrageenans forms milk gels such as flans at a concentration of 0.2 %. ι-Carrageenan forms elastic, κ forms brittle, and λ-carrageenan forms weak milk gels.

5.6. Applications

Medical Applications. Carrageenans are widely used in clinical tests for anti-inflammatory agents by their ability to induce so-called rat-paw edema. Furthermore, it has been shown that carrageenan may act as anticoagulant (heparin analogue) and by activating the Hagemann factor [354]. Carrageenans may also act as immunosuppressor agents [355] and show toxicity against macrophages [356]. It should be noted that some of the earliest studies may have been performed with quite ill defined carrageenans. Carrageenans have more recently been considered as potential antiviral agents [357], [358]. It has been shown that the activity is related to the sulfate substitution pattern and that **D2S,6S**-residues as found in precursor- and λ-carrageenan have the highest reactivity [359].

Other Applications. Carrageenan oligosaccharides have been investigated as fertilizers, stimulants of growth and reproduction and to stimulate defense mechanisms in plants [360].

Discussions on carrageenans for use as paint stabilizer, beverage clarifier, controlled release agent of drugs, in viscosifier in food and tooth paste and in a waste different other applications in the food, nonfood, and medicine sector are found elsewhere [328], [351], [361].

5.7. Physiological Properties

Carrageenan has been given GRAS status by the FDA, based on the assumptions that high molecular mass carrageenans are not absorbed in the intestine, and food grade carrageenans are not ulcerogenic in humans. However, it has been shown that commercial carrageenan samples have a quite wide molecular mass distribution and that carrageenan may be degraded under stomach juice simulated conditions [362]. Therefore the matter is constantly under debate.

[Top of Page]

6. Agar

Alphons C. J. Voragen, Abdelwahab Riad, Rachid Lebbar

Agar, also called agar-agar, kanten, or gelose, is the oldest known gel-forming polysaccharide.

Discovered in the 17th century in Japan and consumed for 200 years, agar is extracted from certain marine red algae of the class Rhodophyceae mainly from *Gelidium* and *Gracilaria* species, growing essentially along the coasts of Morocco, Spain, Portugal, Chile, Japan, and Korea [210].

The extreme inertness of agar led KOCH and PETRI in 1882 to use it as a medium in which to grow bacteria [211]. In his now classical preliminary note on the tubercle bacillus, published in 1882, KOCH formally announced agar as a new solid culture medium. Up to date, no better solidifying agent in microbiological media has been found. At present, agar has become an indispensable substance in microbiological, biotechnological, and public health laboratories, and an important colloid in other industries. Agar is also a permitted gelling, stabilizing, and thickening agent for food applications, authorized in all countries without limitations of daily intake [210]. It finds use in many food industries such as confectionery, bakery, pastry, beverage, sauces, wines, spreads, spices and condiments, meats and fishes, dairy, jams, etc.

Apart from its ability to gelify aqueous solutions and produce gel without the support of other agents, agar can also be used as a safe source of dietary fiber since it is not digestible by the human body.

Agar is insoluble in cold water, hardly soluble in hot water, and readily soluble in boiling water. In food or technical preparations of a high total solids content, agar should be solubilized separately before adding the other ingredients of the formulation. Usual agar is therefore not compatible with the continuous industrial processes.

Quick soluble agar (QSA) is a new agar version developed in 1993 by Setexam. It is obtained from a selected red seaweed and according to a patented manufacturing process without any chemical or genetic modifications.

Unlike traditional agar QSA has a great ability to dissolve in water or in milk at temperatures > 65 °C or during the pasteurization stage, without need to be cooked or heated to boiling, and is therefore compatible with most of the continuous industrial processes.

6.1. Production

Agar is insoluble in cold water but is colloidally dispersible in boiling water. Hence, most commercial agars are produced by hot water extraction. Other methods are possible, such as an extraction with glycerol, anhydrous ammonia, or other solvents [212], but the traditional process employing hot water is the most recommended for the agar destined to food applications.

The agar extraction procedures vary according to the treated seaweed variety. They generally follow three stages (see Fig. 18):

Extraction
Purification
Dehydration

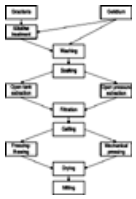

Figure 18. Flow sheet of traditional agar extraction

First, the seaweeds are carefully washed with water in order to eliminate marine salt and foreign matter such as sand. The extraction is made by hot water under pressure or in open tank if the seaweeds have undergone an alkaline treatment. This latter is mainly applied to *Gracilaria* seaweed in order to lower the sulfate content in the resultant agar.

The extraction juice, composed of 99 % water and 1 % agar, is filtered at hot temperature, then cooled at room temperature. The resulting gel is dehydrated under mechanical pressure or by freezing-thawing.

In *pressure dehydration* a pressure of 49 – 98 N/cm^2 is applied to press the gel through intercalated filters. Water passes the filters and the agar molecules are retained as filter cake.

Gel freezing is carried out at – 18 °C. At that temperature the agar molecules are excluded selectively from the ice network and form strips. The thawing step melts the ice and agar strips are recovered selectively.

These two operations of dehydration taken individually remove about 70 % of the initial water content. A part of the residual moisture is then eliminated by drying with hot air. The dry agar flakes are ground to prepare a final product in powder form with different mesh sizes.

6.2. Structure and Gelling Mechanism

Structure. The elucidation of the complex chemical structure of agar by ARAKI had revealed the presence of two dominating polysaccharides: a virtually neutral polymer termed agarose, and a charged polymer termed agaropectin [213]. ARAKI deduced the structure of *agarose* to be 1,4-linked 3,6-anhydro- α-L-galactose alternating with 1,3-linked β-D-galactose [214-216].

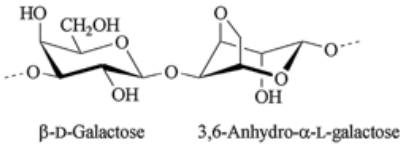

Agaropectin is thought to have the same repeating unit as agarose, although some of the 3,6-anhydro-L-galactose residues can be replaced with L-galactose sulfate residues [217], [218] and the D-galactose residues are partially replaced with the pyruvic acid acetal 4,6-*O*-(1-carboxyethylidene)-D-galactose [217], [219]. Fractionation studies on DEAE Sephadex A-50 carried out by DUCKWORTH and YAPHE [217], [220] have indicated that agar is not made of one neutral and one charged polysaccharide but is composed of a complex series of related polysaccharides which range from a virtually neutral molecule to a highly charged galactan. These studies indicated that in agars containing 4,6-*O*-(1-carboxyethylidene)-D-galactose, this charged residue is always found in regions of the molecule with low sulfate content. Agars in which the D-galactose residues are partially replaced with 6-*O*-methyl-D-galactose contain this 6-*O*-methylated monosaccharide evenly distributed throughout the neutral and charged regions of the complex [217], [218].

Studies by GUISELEY in 1970 [221] indicated that in naturally occurring agars, an increase in gelling temperature occurred with increasing methoxyl group content. In this way, the gelling temperature of an agar from a *Gracilaria* species with high methoxyl content is higher than that of an agar from a *Gelidium* species with little or no methoxyl content.

Gelling Mechanism. The gelling properties of an agar are essentially effected by the agarose part [222], [223]. The rheological properties of the gel are strongly influenced by agaropectin.

Agar forms high strength gels which are completely reversible with a hysteresis cycle over a range of 40 °C (see Fig. 19): Upon dissolution in boiling water agar acquires a random coil configuration. The subsequent gel formation proceeds in two steps. The decrease in temperature of the hot solution, slightly above the gelification temperature, provokes at first the formation of a double-helical conformation of the agarose chains [224-226]. The second stage is the self-associations by nucleation of these double helices to form aggregates [227], [228]. The obtained agar gel is completely heat-reversible since it remelts at above 85 °C (see Fig. 20).

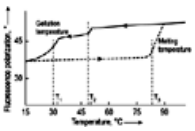

Figure 19. Hysteresis of 1.5 % agar gels

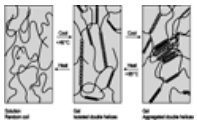

Figure 20. Gel formation mechanism in aqueous agar solutions

The gelation process of agar is not linked with the cations present such as in the case of alginates, some pectins, or the carrageenans, but is only due to formation of hydrogen bonds which explains the thermoreversibility of agar gels. Three equatorial hydrogen atoms of the 3,6-anhydro- α-L-galactose residue are believed responsible for constraining the molecule so as to form a helix with a threefold screw axis [229]. Interaction of these helices causes gel formation.

Replacement of 3,6-anhydro- α-L-galactose residues with L-galactose sulfate causes kinks in the helix, and hence a polysaccharide of lower gel strength is formed [230].

6.3. Quick Soluble Agar

Production. If the gelling mechanism is superposed to the extraction procedure, it can be concluded that agar powder is a dehydrated gel and the

molecules adapt the aggregated double helix conformation. That is why a certain amount of heat energy is required to dissociate the hydrogen bonds and to dissolve the agar molecules in random coil form.

The amount of heat energy required can be considerably lowered if the molecules in agar powder are not obtained in aggregated double helix but in random coil configuration. Setexam managed to finalize a new procedure including neither a chemical transformation nor a genetic modification for the production of Quick Soluble Agar (see Fig. 21).

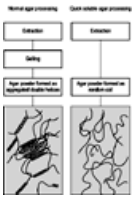

Figure 21. Comparison of production processes of traditional agar and QSA

Properties. QSA has the same physical and chemical characteristics as traditional agar, i.e., gel strength, pH, color, total ash, etc., except for solubility. Contrary to traditional agar that dissolves only in boiling water, QSA dissolves instantly in water or milk at 65 °C or during pasteurization. This property of QSA simplifies the manufacturing of food or technical products, in batch and continuous processes. It eliminates all the disadvantages met with the traditional agar utilization such as the necessity of a separated dissolution stage, a large initial volume of water, blackening on cooking inner sides, degradation of the gelling power, etc.

Application of QSA is very easy, as it is introduced during the preliminary stages of any food or technical preparation and dissolves instantaneously during the heating, the pasteurization or the sterilization stage, which makes it compatible with most of the continuous industrial processes.

The thermoreversibility of QSA gels facilitates its handling and recycling; and its unmatched natural hysteresis offers a definite advantage particularly with regards to the shelf life of food or technical preparations.

The QSA reacts selectively with the water portion of any recipe in a wide range of pH values without interaction with neither the present cations nor with the present organic matters.

[Top of Page]

7. Gum Arabic

Alphons C. J. Voragen

Gum arabic is a dried exudation obtained from the stem and branches of acacia trees, *Acacia senegal* L. or *Acaica seyal* (fam. Leguminosae). For commercial production the trees are tapped or drilled. After drying in air, the exudate tears are harvested by hand and transported to central collecting stations where they are sorted according to size, color, and contamination with impurities. Harvesting takes place during October to June. A young tree yields ca. 0.9 kg and an older tree up to 12 kg of gum annually. The trees grow in the tropical and subtropical areas of Africa. Exproting countries are Sudan, Tschad, Nigeria, Mali, Mauretania, Senegal, and Ethiopia. Acacia species can be found worldwide, but only the exudates of *A. senegal* and *A. seyal* may be traded as gum arabic. World production is around 40 000 t/a.

Raw gum is still sold in substantial quantities; however, a precleaning by air classification to remove sand, fines, and bark is required. The product can be further processed to a kibbled, granular, or powdered form. Importers in Western Europe and the United States produce speciality products for particular applications. For this purpose the gum is dissolved, and the gum solution is exposed to sieving or centrifugation and then drum or spray dried [1-4], [231].

Structure. Gum arabic consists of three principal fractions, an arabinogalactan fraction (AG) with very little protein, representing 88 % of the gum und having a molecular mass of 250 000; a high molecular mass (1.5×10^6) arabinogalactan – protein complex (AGP) representing about 10 % of the gum; and a glycoprotein (GI) with a molecular mass of 200 000, making up 1.2 % of the gum. By treatment with proteolytic enzymes the AGP fraction is degraded to fragments with a molar mass similar to AG. The wattle blossom model has been proposed to describe the structure of AGP: Carbohydrate blocks with molecular mass of ca. 250 000 are linked together by a main polypeptide chain. The carbohydrate blocks consist of a β-1,3-linked galactopyranose backbone chain, with numerous side chains linked through β-1,6-galactopyranose residues and containing arabinofuranose, arabinopyranose (Ara), rhamnopyranose (Rha), glucuronic acid (GlcA), and 4-*O*-methylglucuronic acid. The sugar moieties occur in the ratio Gal 36, Ara 31, Rha 13, and GlcA 18. The structure of a representative segment is shown in Figure 22 [232-235]; depending on growth conditions and age, some structural variations exist.

Figure 22. Structural features of gum arabic

Properties. The molecular mass of gum arabic averages around 500 000; the glucuronic acid content (including 4-*O*-methylglucuronic acid) of ca. 18 % makes gum arabic an acidic polysaccharide. With a cation exchanger or by washing with acidified alcohol, gum arabic can be converted to arabinic acid, which exhibits the titration curve of a strong acid. A 5 % solution has an apparent dissociation constant of 10^{-3}. Solutions in water have a pH of 2.2 – 2.7.

The most striking feature of gum arabic is its *extreme solubility in water*. Due to its highly branched, compact structure, solutions containing up to 40 % gum arabic can be obtained, which still exhibit typical Newtonian flow behavior. Compared to other polysaccharides of similar molar mass, the viscosity of these solutions is very low (Table 4). Viscous solutions are obtained only at concentrations > 30 %, where effective molecular overlap begins to occur and solutions gradually assume pseudoplastic behavior. The *viscosity* of gum arabic solutions is highest in the pH range 4.5 – 5.5; in this range the carboxyl groups are ionized to a large degree. Addition of salts strongly decreases the viscosity due to suppression of the electrostatic charge. Gum arabic is compatible with most other hydrocolloids except for alginates and gelatin. It is also surface active and has been widely used to *stabilize oil-in-water emulsions*. The AGP fraction is responsible for its emulsifying ability. The polypeptide chain is believed to be at the periphery of the molecule and to adsorb onto hydrophobic surfaces. Gum arabic has been shown to be heat sensitive; proteinaceous material (AGP and GI) precipitates on prolonged heating, and as a consequence the solution viscosity drops considerably and the emulsifying ability is lost [231], [232], [235], [236].

Gum arabic can be *precipitated* with trivalent cations and by salts such as silicates, borates, and mercury nitrate. It is insoluble in oils and most organic solvents but soluble in aqueous ethanol up to ca. 60 % ethanol. In gum arabic preparations enzyme activities have often been

demonstrated, for example, oxidases, peroxidases, amylases, and pectinases. For applications in foods, these activities are undesirable, and the enzymes must be inactivated by heat [2].

Applications, Market. Gum arabic was a commodity 4000 years ago in Egypt, and was used for the manufacture of cosmetics, dyes, and tinctures. Half of the amount imported by the United States is used in the food industry; this makes gum arabic one of the most useful plant hydrocolloids. The annual production has risen to > 50 000 t/a [231], [234]. Gum arabic is used in groups 2, 4, 8, and 11 (Table 2). Dried, powdered flavors (group 5) constitute a major application. High gum concentrations allow spray drying without loss of flavor. Its emulsifying and stabilizing ability, in combination with its use in high concentrations, makes gum arabic attractive for pharmaceutical and cosmetic applications (groups 16 and 21). In industrial applications such as an adhesive on gummed envelopes and stamps (group 20), and as a dispersive agent in dyes and tinctures, gum arabic is being replaced by starch and cellulose derivatives. Prices vary between $ 3 and $ 8.5 per kilogram.

[Top of Page]

8. Gum Tragacanth

Alphons C. J. Voragen

Gum tragacanth is the exudate from breaks or wounds inflicted in the bark of shrubs of the *Astragalus* species, particularly *Astragalus gummifer* Labillardière (Leguminosae family), which are found in the mountainous regions of Iran, Syria, and Turkey. The gum was described by THEOPHRASTUS in the third century B.C. The name tragacanth is of Greek origin and means goat horn. It probably refers to the curved shape of the 2 – 3-mm-thick ribbons that represent the better quality gum. Gum tragacanth also exists as flakes, which are of poorer quality. The bushes are carved to stimulate gum production. Ribbons are obtained from April to September; flakes are formed in the following months until November. The translucent, whitish, elastic ribbons, which originate predominantly in Iran, are considered the best quality. The harvest is collected at central stations and is exported mainly to the United States and the United Kingdom. Importers process the raw materials to standardized powders [1-4], [232], [233], [237]. Prices range from $ 16 – $ 21 per kilogram; the annual harvest is ca. 3000 t/a.

Structure. Gum tragacanth is composed of tragacanthic acid, which is the major component, and a nearly neutral *arabinogalactan* that is water insoluble but swells to a gel. The constituent sugar residues of this arabinogalactan are: 75 % L-arabinose, 12 % D-galactose, 3 % D-galacturonic acid methyl ester, and L-rhamnose. *Tragacanthic acid* is water soluble and consists of a backbone of 1,4-linked α-D-galactopyranosyluronic acid residues (polygalacturonic acid chain as in pectins) with short side chains including single-unit β-1,3-linked D-xylopyranose, dimeric β-1,3-linked D-xylopyropyranosyl-1,2- α-L-fucopyranose, and dimeric β-1,3-linked D-xylopyranose-1,2-D-galactopyranose. The structure is shown schematically in Figure 23. The sugar moieties are present in the relative proportions D-galacturonic acid 43 %, D-xylose 40 %, L-fucose 10 %, and D-galactose 4 %. Tragacanthic acid has a high molecular mass (840 000) and gives accordingly highly viscous solutions (Table 4) that exhibit pseudoplastic behavior at concentrations > 0.5 %. A 1 wt % solution can have a viscosity of 3600 mPa · s. Accompanying Ca^{2+} and Mg^{2+} ions may contribute to the high viscosity and gelling properties of gum tragacanth solutions. The viscosity changes little in the pH range 2 – 10 and is stable during heating in a 1 % acetic acid solution [1-4], [232], [233], [237].

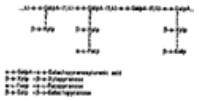

Figure 23. Structural features of gum tragacanth

Applications, Market. The application of gum tragacanth in foods and acid products is based on its heat stability, and its stabilizing ability in emulsions and suspensions. Therefore, gum tragacanth is of particular interest for group 12 (Table 2). Because of the slimy consistency of gum tragacanth dispersions, this gum can also function as a fat replacer. Other applications are mentioned in groups 8 and 9. Outside the food industry gum tragacanth is also used in groups 16, 18, and 21. A special application is its use in spermicidal jelly.

[Top of Page]

9. Gum Karaya

Alphons C. J. Voragen

Gum karaya is also named Sterculia gum, Kadaya gum, or Indian gum; it is the dried exudate from the bark of *Sterculia urens* or other *Sterculia* species (Sterculaceae family), trees occurring throughout northern and central India. The formation of exudate is artificially initiated by carving the bark. After a few days the exudate can be harvested as large irregularly shaped tears, which may weigh 1 kg or more. From April to June, 4.5 kg can be obtained per tree. The gum is sold to merchants at central trade stations and precleaned by removing pieces of bark; the tears are broken and graded according to purity and color. In consumer countries, the gum is then milled to a homogeneous powder. In the past, gum karaya was sold in a blend with gum tragacanth; in the meantime, however, it has found many applications because of its specific characteristics. The annual production is estimated at 5000 t, the larger part of which is used in the United States. Prices are ca. $ 4.8 –$ 6.4 per kilogram. Expansion of the production requires afforestation and takes at least 10 years [1-4], [232]. Gum karaya is a highly acetylated, acidic polysaccharide with a backbone chain of alternating 1,4-linked α-D-galacturonopyranosyluronic acid residues and 1,2-linked α-L-rhamnopyranosyl residues. Singleunit β-D-galactopyranosyl- and β-D-glucuronopyranosyluronic acid side chains are attached to O-2 and O-3 of the galacturonic acid residues, whereas half of the rhamnose residues have 1,4-linked β-D-galactopyranosyl residues at O-4 (Fig. 24). The molecule also has 13 wt % acetyl groups. Gum karaya is poorly water soluble; however, it swells readily in cold water to many times its original volume, particularly when finely ground (<200 mesh), giving a dispersion that appears homogeneous. By using an autoclave, solutions of up to 20 % can be obtained; their viscosity is, however, decreased as a result of degradation. Gum karaya solutions deviate from Newtonian behavior and exhibit viscoelasticity. Under alkaline conditions, deacetylation occurs and converts a gum karaya suspension into a ropy mucilage [232], [233]. The ability to swell and its resistance to enzymic and microbial degradation make powders of gum karaya good adhesives for dentures. Particles of ca. 0.6 mm swell to 60 – 100 times their volume and form a mucilage that allows drainage. Food applications are in groups 9, 11, 12, and 13 (Table 2). Gum karaya has also found use in the paper industry (group 29) and textile painting (group 30).

Figure 24. Structural features of gum karaya

[Top of Page]

10. Gum Ghatti

Alphons C. J. Voragen

Gum ghatti, also called India gum, is an exudate of *Anogeissus latifolia* (Combretaceae family), trees that grow in the deciduous forests of India and Sri Lanka. The Indian word gath means mountain pass and refers to the trade routes for this gum. The exudate usually has the shape of 1- cm-thick

beads or tears, which may be colored by absorption of tannins from the bark. Also, dust and other particles may stick to the gum. Gum ghatti is harvested mainly in April and graded by traders according to color and impurities. The highest quality is pure and only lightly colored. Importers, particularly in the United States, purify and standardize the gum by grinding, sieving, and blending or by solubilization in water, filtration, and spray drying. It is of minor importance, with an annual production of ca. 1000 t [4].

Gum ghatti is composed of D-galactose, D-mannose, D-glucuronic acid, and L-rhamnose. A complex array of single-unit and oligomeric side is attached to a chain of alternating 1,4- β-D-glucuronic acid and 1,2- α-D-mannopyranose residues. The side chains comprise several 1,6-linked D-galactopyranose units, some terminated by glucuronic acid, joined to the mannopyranose residues at O-3 through 1,3-linked L-arabinopyranose residues. Single-unit L-arabinofuranose residues can also be found, some attached to the O-6 of mannopyranose residues, and other short sequences of L-arabinofuranose [233], [238]. *O*-acetyl groups have also been found. A schematic presentation of the structure is shown in Figure 25. In its acid form, gum ghatti has an equivalent mass of 1067; the gum occurs as the calcium or magnesium salt. The gum is not entirely soluble in cold water, but 90 % can be solubilized by maceration and heating at 90 °C. Removal of the divalent cations enhances solubility. The remaining 10 %, which is chemically identical to the soluble portion, is strongly swollen and also contributes to the viscosity [192]. Spray-dried preparations are completely soluble in water; they give, however, a lower viscosity of the aqueous solutions. The viscosity of gum ghatti solutions is higher than that of gum arabic solutions; viscosity increases quadratically with concentration. Like gum arabic, gum ghatti is an excellent stabilizer of emulsions and suspensions, and has found application in wax suspensions and pastes, oil emulsions, polishes, syrups, and as a dispersive agent in dried, vitamin-rich oil suspensions.

Figure 25. Structural features of gum ghatti

[Top of Page]

11. Xanthan Gum

Alphons C. J. Voragen and Ian Challen

Xanthan gum is the extracellular, high molecular mass heteropolysaccharide produced by the aerobic fermentation of *Xanthomonas campestris* NRRL-B1459 [1-5], [239]. The production of xanthan gum goes back to the investigations of JEANES and coworkers in the early 1960s at the Northern Regional Research Laboratory of the United States Department of Agriculture, Peoria, Illinois, in the framework of a large screening program for slime-producing microorganisms of industrial interest [240].

11.1. Production

Commercial production is carried out by a batchwise, submersed fermentation under strong aeration. During fermentation the bacterial cells are kept under a constant stress to direct their metabolism to metabolite production instead of growth. This is achieved by proper selection of medium composition. The medium cost forms a significant part (ca. 25 %) of the total production cost; the sugar concentration is adjusted so that it falls as low as possible at the end of the fermentation cycle, but depletion must be avoided to prevent the microorganisms from using the polysaccharide as a substrate [241]. For xanthan gum production, a typical medium contains a nitrogen source, phosphate and magnesium ions, trace elements, and glucose, which is kept at < 5 %. The pH must not be below 7. After 96 h at 30 °C, less than 0.1 % glucose is left in the medium, and a conversion to polysaccharide of more than 50 % can be obtained.

In the patent literature [242], much attention is devoted to propagation of the starter culture. In this stage, nitrogen is supplied by organic sources (e.g., soy peptone or corn steep liquor), which also contain some growth factors. During the main fermentation, nitrogen is supplied by mineral salts (e.g., ammonium nitrate). A typical composition of the medium at this stage is 0.06 % ammonium nitrate, 0.5 % potassium dihydrogenphosphate, 0.01 % magnesium sulfate heptahydrate, 2.25 % glucose, and 97.18 % water (Fig. 26). After inocculation of 5300 L of medium with 227 L of starter culture and vigorous aeration for 72 h at 28 – 31 °C, 77 kg of hydrocolloid can be recovered. The remaining medium contains < 0.1 % glucose. This means a conversion of 64.7 % and a final concentration of the polysaccharide of 1.45 %, which gives the medium a viscosity of 3000 mPa · s at 25 °C (Brookfield Viscometer, 60 rpm).

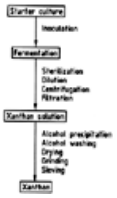

Figure 26. Flow sheet for the manufacture of xanthan gum

The usual precautions must be taken to preclude contamination of the batch fermentation through the culture medium, equipment and accessories, air, and neutralizer; thus it is doubtful if the batch operation can be replaced by a continuous process. Also the stability of the bacterial strain is a point of consideration. Since *Xanthomonas* bacteria do not form spores, the fermentation can be stopped by heat treatment in a heat exchanger. Further downstream processing involves centrifugation, filtration, and precipitation with ethanol or isopropanol as used for other hydrocolloids. The high viscosity of the fermentation liquor must be reduced by dilution with water or alcohol. The precipitate, dried by vacuum or with hot air, is processed to a marketable article by grinding and sieving. The powder is off-white to yellow in color.

11.2. Structure and Properties

Chemical Structure. Xanthan is a heteropolysaccharide composed of D-glucose, D-mannose, and D-glucuronic acid. It has a β-1,4-D-glucan (cellulosic) backbone substituted through C-3 on alternate glucose residues with a trisaccharide side chain consisting of β-D-mannose-(1,4)- β-D-glucuronic acid-(1,2)- α-D-mannose. The terminal mannose moiety may have pyruvate residues linked to the 4- and 6-positions, the internal mannose unit is acetylated at C-6 (Fig. 27) [243]. X-ray diffraction studies on oriented xanthan gum fibers [244] identified the molecular conformation as a right-handed five-fold helix with a pitch of 4.7 nm. In this conformation the trisaccharide side chain is aligned with the backbone and stabilizes the overall conformation by hydrogen bonding [245].

By modifying the biosynthetic pathway for xanthan production, the carbohydrate structure and the substitution pattern of the polymer can be genetically controlled. In this way, polysaccharides with quite different properties can be obtained [246].

Xanthan gum has a *molecular mass* of ca. 2×10^6. Since the structure and the molecular mass are genetically controlled, hydrocolloids with similar molecular mass and similar properties are obtained.

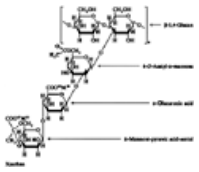

Figure 27. Structural features of xanthan gum

Solution Properties. Xanthan gum is completely soluble in hot or cold water. The gum can be hydrated in sugar solutions up to 60 %. The viscosity of the gum solution is a function of its concentration in the dispersion [247]. *Aqueous solutions* of xanthan gum are extremely viscous (Table 4) and pseudoplastic (i.e., they exhibit a reversible, highly shear-thinning behavior) [248] resulting from the conformation (rod-like) structure and the high molecular mass of the polymer [249]. A 0.25 % solution in water containing 1000 µg/g NaCl and 147 µg/g $CaCl_2 \cdot H_2O$ has a viscosity of ca. 1000 mPa · s at a shear rate of 1 s^{-1} which decreases to ca. 50 mPa · s at a shear rate of 100 s^{-1}. The property of shear thinning is made use of in food processing and in improving the sensory qualities of food [250]. Equations relating apparent viscosity, concentration, shear rate, and temperature have been reported [251].

Xanthan gum solutions exhibit higher viscosities than other gums at the same low concentration. Shear thinning is more pronounced due to the semi-rigid conformation (compared to the random coil conformation of the other gums). The formation of a weak gel network results in high yield point values. Theses interactions are shear-reversible. 1 % Xanthan gum in a 1 % KCl solution has a yield value of ca. 11 300 mPa compared to 4000 mPa for guar gum, 830 mPa for hydroxypropyl methyl cellulose (HPMC) and 360 mPa for locust bean gum (LBG). At a concentration of 0.3 % xanthan gum exhibits a yield value of 500 mPa while the other gums mentioned have no significant yield value. This quality is responsible for xanthan gum's ability to stabilize emulsions and dispersions [252].

Effect of Salts. In the presence of salts the solution viscosity is stabilized and maintained; polyvalent cations, however, can precipitate xanthan gum under alkaline conditions. Xanthan gum solutions are compatible with high concentrations of various salts [253].

Effect of Temperature. At low ionic strength xanthan gum solutions undergo a thermal transition (i.e., melting), as can be detected by a viscosity change [254]. Optical rotatory and circular dichroic transitions are coincident with the viscosity change, indicating a conformation transition. These data are consistent with the unwinding of an ordered structure into a random coil [255]. The transition temperature can be increased to > 90 °C when small amounts of salt are added [257].

The viscosity of a xanthan gum solution containing salt is almost unaffected over the temperature range 10 – 90 °C. Sterilized food products containing xanthan gum loose only ca. 10 % of their viscosity — much less than products stabilized with other hydrocolloids such as guar gum, alginate, and carboxymethyl cellulose (CMC).

In foods intended for use in microwave ovens xanthan gum eliminates moisture separation. At a gum level of 0.4 % the viscosity is not affected by electrolytes while at a 1 % gum level the viscosity increases significantly.

Effect of pH. The secondary structure (the side chains wrapped around the backbone) imparts xanthan gum stability against degradation by acids, bases, heat treatment, freeze – thaw cycles, enzymes and extended mixing processes [239].

Xanthan gum solutions are stable over a wide pH range (ca. < 2.5 to > 11) [247]. At pH $\simeq$ 9 and above the gum gradually deacetylates. Its stability depends on the gum concentration: the higher the concentration the more stable the solution.

Enzymes commonly present in food products or added to such products (amylase, pectinase, cellulase) do not degrade xanthan gum although it can be partly degraded by endo- β-1,4-glucannases after removal of calcium ions [256].

Interactions and precipitation can occur in acidic and/or heat-treated dairy-protein systems [239].

Oxidizing Agents. Xanthan gum solutions are degraded by high levels of strong oxidants such as hypochlorite, persulfate, and hydrogen peroxide [247], particularly at high temperature. The gum will dissolve in 8 % solutions of sulfuric, nitric and acetic acids, 10 % hydrochloric acid, and 25 % phosphoric acid [257]. Such solutions remain stable for several months.

Solvents. Up to 50 % ethanol and propylene glycol can be added to aqueous xanthan gum solutions without causing precipitation.

Other Hydrocolloids. Xanthan gum is highly compatible with alginates and starch exhibiting long term stability. It is completely compatible with most natural and synthetic thickeners.

Gel Formation. Two models have been proposed for weak gel formation in xanthan gum solutions; both models are based on the occurrence of xanthan molecules in helical structures (Fig. 28). In the model of NORTON et al., single helical regions associate to form a network [258]. The model of MORRIS [259] is based on a double-helical model for xanthan. The primary mechanism is considered to be end-to-end association into fibrous aggregates via double-helix formation.

Xanthan gum interacts with the galactomannans (locust bean, guar, and tara gum) and glucomannans synergistically, leading to enhanced viscosity or gel formation. The formation of thermoreversible gels by synergistic mixtures of xanthan gum with certain galacto- and glucomannans has been ascribed to intermolecular binding through cocrystallization of denatured xanthan chains within the mannan crystallite [260]. The cellulosic backbone of xanthan gum and the stereochemically similar mannan backbone both form ribbonlike structures. The mixed crystallites probably act as strong junction zones to consolidate the weak xanthan gum network.

Gel Formation with Galactonmannans. At all concentrations and blend ratios xanthan gum and *guar gum* interact to produce a synergistic viscosity increase. This interaction will occur in the cold although higher viscosities are generated if the system is heated and cooled.

It is presumed that the interaction occurs between the xanthan gum and the “smooth” regions of the galactomannans, thereby explaining the stronger interaction with locust bean gum than with guar gum [261-263]. Xanthan gum exhibits a similar reactivity with tara gum [35].

The xanthan gum – guar gum interaction leads to an increase in solution viscosity and elastic modulus (*G*'). Determinations of elastic modulus and apparent viscosity are important for the stabilization of emulsions and suspensions (e.g., salad dressing requires a high elastic modulus while sauces and soups need a high apparent viscosity). Maximum viscosity synergism is produced at total gum concentrations of up to 1 %. The synergistic ratio varies with the total gum level; for example, at a total gum level of 0.5 % the maximum viscosity is produced with a ratio of xanthan gum:guar gum of 30 : 70 while at a total gum concentration of 0.2 % the ratio is nearer 50 : 50. The synergism decreases in the presence of ionic materials such as salt and acid. These synergistic blends have a lower elastic modulus which is important in some food applications. The presence of xanthan gum improves the heat stability of guar gum. These blends have a wide variety of applications due to the good sensory properties and lower cost than xanthan gum alone.

In *locust bean gum* the ratio of mannose (backbone) to galactose (side chains) is ca. 4 : 1 but, unlike guar gum, the distribution of the side chains is

more irregular — there are regions that are highly substituted ("hairy") and unsubstituted ("smooth") regions. In xanthan gum – LBG blends at total gum levels of < 0.1 % synergistic viscosity increases are produced; at gum levels > 0.1 % strong cohesive thermoreversible gels exhibiting very little syneresis are obtained. The optimal gum ratios are xanthan gum : LBG of 40 : 60 to 60 : 40. The interaction, as demonstrated by gel strength, is stronger at neutral pH and decreases with decreasing pH. In food products this gel texture by itself is undesirable but it can be modified by the inclusion of starches and/or proteins [239].

Rheological and ultracentrifugation studies were conducted on heated and unheated mixtures of xanthan gum with whole LBG, and temperature fractions of the latter possessing different mannose : galactose ratios [264]. Results suggest that xanthan gum and LBG interact via two distinct mechanisms. One mechanism takes place at room temperature, gives weak elastic gels, and has little dependence upon the galactose content of the galactomannan, whereas the second mechanism requires significant heating of the polysaccharide mixture, gives stronger gels, and is highly dependent upon galactomannan composition [264].

Mixed gels of xanthan gum and galactomannans (guar gum and fractionated LBG) were compared by means of oscillatory shear measurements [265]. Whatever the proportion of LBG, maximal synergism was observed when the ratio of xanthan gum to galactomannan was 1 : 1. However, the magnitude of this maximum varied with LBG content in the mixture.

If guar gum is modified through partial removal of galactose substituents with an α-galactosidase, it also gels with xanthan gum [260]. The sedimentation rate in aqueous mixtures of xanthan gum and enzyme-modified galactomannans (the galactose content being decreased by the enzyme) was studied [266]. Results were related to rheological properties of the systems at large deformations. The shear modulus increased while the strain at yielding and sedimentation velocity decreased with decreasing galactose content of galactomannan. Properties of the xanthan gum enzyme-modified galactomannan system were compared with those of xanthan gum – LBG systems. LBG was suggested to be a better stabilizer than the enzyme-modified galactomannans studied [266].

Gel Formation with Glucomannans. Xanthan gum – konjac mannan (glucomannan) interactions are strong at total gum concentrations above 0.2 % [267]. The optimum ratio is between 40 : 60 and 30 : 70. The presence of salt decreases the interaction. At temperatures > 50 °C, the system's *G*' is higher than its *G*'', indicating well defined elastic properties. At ca. 55 °C the gel structure collapses and there is a sharp decrease in the value of *G*'. At values > 55 °C the behavior is typical of a liquid which has *G*'' significantly higher than *G*'. At ca. 1.0 % total gum concentration, maximum gel strength is achieved, with no further increase in this property with increasing gum concentration. The mechanical properties of xanthan gum – konjac mannan are also similar to those of xanthan gum – LBG.

Others. Small amounts of xanthan gum are sufficient to stabilize *starch solutions* during storage. Xanthan gum at 0.1 – 0.2 % prevents starch solution retrogradation and increases its stability, especially at low pH (around 3). These small amounts of xanthan gum also improve starch stability during freeze – thaw cycles.

Reactions of xanthan gum with specific dextrins and incompatibility between xanthan gum and gum arabic at pH values below 5 have been reported. Xanthan gum also reacts with casein below its isoelectric point [268].

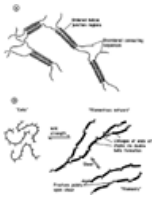

Figure 28. A) Weak gel formation in xanthan gum solutions by self-association of single helices and B) End-to-end association via double-helix formation into fibrous aggregates [260]

11.3. Analysis

The presence of xanthan gum can be indicated by the formation of a gel during cooling of a 1: 1 mixture with carob gum in a 1 % solution. Its presence can also be indicated from the reaction of pyruvate with 2,4-dinitrophenylhydrazone in a hydrolysate.

11.4. Applications, Market

The applications of xanthan gum are based especially on the viscous behavior and stabilizing properties of this hydrocolloid. In industry, xanthan gum is used in groups 18, 19, 26, 27, 28, 29, and 33 (Table 2). These applications enabled Kelco, the first producer of xanthan gum, to carry out a comprehensive toxicological evaluation that resulted, in 1969, in the acceptance of this hydrocolloid in foods, where it has found applications especially in groups 4, 5, 8, and 12.

Xanthan gum is produced by several companies. Present production is estimated at ca. 20 000 t/a. For food applications prices are ca. $ 11 – $ 13 per kilogram. Because of its specific properties the potential for this polysaccharide is great.

[Top of Page]

12. Gellan Gum

Alphons C. J. Voragen and Ian Challen

Gellan gum is produced by the bacterium *Pseudomonas elodea* (ATCC 31461). The bacterium was originally found growing in a lily pond in Pennsylvania, USA. Gellan gum has been approved for food uses [260], [269], [270]. It is a broad-spectrum, multipurpose gelling agent, which is fully compatible with many food systems and processes. Gellan gum is both cold setting and thermosetting; its melting point (thermoreversible and thermoirreversible) depends on ionic strength. Gelled products can be made to be stable or to melt when subjected to retorting and microwave cooking.

Production. The gum is produced commercially by inoculating a fermentation medium with the microorganisms. The medium contains a carbon source, phosphate, organic and inorganic nitrogen salts, and trace elements. The fermentation is carried out under sterile conditions with careful control of aeration, agitation, pH, and temperature [271]. When fermentation is complete, the viscous broth is pasteurized to kill viable bacteria before the gum is recovered. Recovery can be achieved in several ways. Direct recovery yields the gum in its native or *high acyl form*. Treatment of the broth with alkali causes deacylation resulting in the *low acyl version*. The presence of the acyl substituents largely influences the resulting solution properties and gel characteristics [272].

Structure. Gellan gum is a linear anionic polysaccharide with a tetrasaccharide repeating unit:

-3)- β-D-Glc-(1,4)- β-D-GlcA-(1,4)- β-D-Glc-(1,4)- α-L-Rha-(1-

Light scattering and intrinsic viscosity measurements give a molecular mass of approximately 5×10^5 for the deacylated gum [273].

In the native polysaccharide two acyl substituents, L-glycerate and acetate, are present. Both substituents are located on the same 1,3-linked glucosyl residue. The glycerate is esterified on the hydroxyl group on C-2, the acetate on the hydroxyl group on C-6. On average, there is one

glycerate per repeating unit and one acetate per every two repeating units [274].

Chemical analysis established that glycerate substitution predominates over that with acetate [275]. In addition, glycerate substitution dramatically influences gellan gum properties since this bulky substituent hinders chain association and accounts for the change in gel texture brought about by de-esterification [270].

X-ray diffraction studies of oriented fibers [275-277] shows that gellan gum exists as a three-fold, left handed, parallel double helix. The pair of molecules that comprise the helix is stabilized by hydrogen bonds at each carboxylate group. In the potassium salt of the deacylated version the potassium ion is coordinated to the carboxylate group, which, in turn, is involved in interchain hydrogen bonds to stabilize the duplex. In this structure the location of the potassium ions on both the inside and the outside of the helix provides a mechanism for both stabilization of the helix and the association of the helices that is necessary to produce a macromolecular gel structure to aggregate [278]. The calcium salt form fits the same model with the divalent calcium ion replacing the potassium – water – potassium bridge [279]. In these salt forms of the gel, helix aggregation is responsible for the brittle nature of the gel. In the native form the acyl substituents do not prevent helix formation. However, the acetate groups interfere with aggregation, resulting in elastic rather than brittle gels.

Hydration. The pure sodium ion form of gellan gum hydrates at room temperature in deionized water. However, the commercial product is a mixed salt — predominantly in the potassium ion form but also containing other ions such as sodium, calcium, and magnesium. This form is not soluble in cold water since the low level of divalent ions is sufficient to hold the molecular chains together and, thereby, inhibit hydration. Hydration can be achieved by dispersing the gum in deionized water and heating to 75 °C. In tap water, the hydration temperature depends upon the concentration and types of ions in solution. For example, at around 180 ppm calcium carbonate hardness (0.0072 % calcium) the gum will not dissolve even in boiling water but it will hydrate at 90 °C in a 1 % sodium chloride solution [272].

In many systems, hydration of gellan gum by simply dispersing the gum in water and heating is not practical. In such systems hydration is effected by the addition of a sequestrant such as sodium citrate. The sequestrant removes the inhibitory divalent ions from both the gum and the water, but (if used at the appropriate amounts) does not introduce sufficient sodium ions into the system to prevent gum hydration. Depending upon the ions initially present and their concentrations, gellan gum can be hydrated at room temperature when a sequestrant is used [272].

Solution Properties. Highly viscous solutions of *low acyl gellan gum* can be prepared using a sequestrant. At a concentration of 1 wt % gellan gum solutions are less pseudoplastic than xanthan gum solutions but more pseudoplastic than solutions of high molecular mass sodium alginate [276]. A large drop in the solution viscosity of gellan gum is observed at 25 – 40 °C. This reflects a conformational change from some form of relatively ordered, nonaggregated double helix to a random coil. This change is reversible.

Native, or *high acyl gellan gum* dispersions in cold water have very high solution viscosities. Such solutions are very sensitive to salt concentration. In equally concentrated solutions of xanthan gum and gellan gum containing increasing levels of sodium chloride the viscosity of the native gellan gum is strongly dependent upon the salt concentration whereas that of the xanthan gum is not. High acyl gellan gum solutions are highly thixotropic; such behavior is the outcome of gel-like network formation. Similar thixotropic behavior is observed with solutions containing a blend of xanthan gum with locust bean gum.

Gel Formation. Unlike other gums gellan gum forms gels with many ions including hydrogen ion. However, its affinity for each ion is different. Divalent ions have a much stronger affinity than monovalent ones. Divalent ions like calcium and magnesium produce optimal gel strength at 3 % of the level required with sodium and potassium ions [280]. Since ions promote chain association in forming gels, the ion levels needed to cause gelation tend to prevent the gum from hydrating. Furthermore, in keeping with their respective affinities, divalent ions inhibit hydration much more strongly than monovalent ions.

Ions are needed to set up the gel network. The simplest way to make a gellan gum gel is to hydrate the gum in deionized water by heating the gum dispersion to above 75 °C, add ions such as calcium or sodium ions to the hot solution, and cool. When sequestrants are used for hydration, ions are already present in the system. Since the level of these ions is such that they do not prevent hydration, they are also not sufficiently abundant to form strong gels. In several applications such weak gels are quite acceptable but, to optimize gel strength, additional ions must be added. Again, these ions can be added to the hot gellan gum solution which will later gel when allowed to cool.

Gel Texture. The texture of gellan gum and other polysaccharide gels has been investigated using the technique of Texture Profile Analysis on an Instron (a testing machine produced by Instron). Free-standing gel disks were prepared by allowing hot solutions to cool in ring molds. The set, tempered, gels were compressed twice in succession to a given percentage of their original height at a given rate. The textural parameters (modulus, hardness, brittleness, elasticity and cohesiveness) were calculated from the resulting force/deformation curve [272]. The results show that gels made with low acyl gellan gum have a high modulus and are very brittle; in contrast high acyl gellan gum gels are very elastic, nonbrittle and cohesive [281], [282].

Low acyl gellan gum gels are clear, firm to the touch and do not melt, but they release liquid when eaten giving a melt-in-the-mouth sensation. This property is important to flavor release and can be useful in, for example, dessert products [283]. Other investigations have suggested that overall flavor increases as gel hardness decreases and gel brittleness increases [284].

Melting and setting points, which are important gel properties, depend upon ion concentration and type and, to a lesser extent, on gum concentration [269]. Gels differ in their setting temperature. For example, those containing calcium ions set between 25 and 40 °C, and those with sodium ions set at 40 – 50 °C. At lower ion levels, gels remelt upon heating, at higher ion levels they do not melt below 100 °C.

Applications. Gellan gum is used in many applications including microbiological media, tissue culture media, food and pet feeds, photographic emulsions, and microcapsules. It is also used as a model for studying gel swelling and gel shrinkage [269], [285], [286]. Gellan gum can be used as a fining agent for alcoholic beverages including beer, wines, and fortified wines [287].

Gellan gum is mainly used in confectioneries, jams and jellies, fabricated foods, water-based gels, pie fillings and puddings, icings and frostings, and dairy products. Gellan gum and polydextrose were used to produce low-calorie or sugar-free jelly sweets where the final solids content exceeded 80 %. In comparison to other gelling agents gellan gum does not appreciably add to the viscosity of the depositing mix and is, therefore, an ideal gelling agent for such systems [288].

Gellan gum is useful in cocoa and chocolate products [289] and in alcoholic and nonalcoholic beverages [290].

Gellan gum is used in a variety of Japanese food including Mitsumame jelly cubes, soft and hard red bean jelly and Tokoroten noodles [269]. It has been reported to stabilize oil-in-water dispersions [291] and can be used to prepare gels that are dried and subsequently rehydrated to their original shape [292].

Legal Aspects. Gellan gum has been shown to be completely safe for use in food. It has been given a “not specified” ADI by the Joint FAO/WHO Expert Committee on Food Additives [293] and by the EC's Scientific Committee for Food [294].

Gellan gum is approved in over 30 countries worldwide, including the United States, Canada, Australia, South Africa, Japan and most of South East Asia, e.g., Hong Kong, Philippines, Taiwan and Thailand, most of South America, and all 15 EU member states.

In the European Union gellan gum is one of the generally permitted additives listed (as E418) in Appendix I of the European Parliament and Council Directive 95/1/EU on food additives other than colors and sweeteners [295].

Specification and Purity. Kelcogel gellan gum complies with the specifications for gellan gum prepared by the Joint FAO/WHO Expert Committee on Food Additives (JECFA) [296] and the Food Chemicals Codex (FCC) [297].

Market. The commercially available products are K9A50, a nonclarified form of the gum for industrial applications; Kelcogel LT100, a nonclarified form and Kelcogel, a clarified form for food and industrial uses; and Gelrite, which is used for microbiological plating media. All gellan gum is produced by NutraSweet Kelco (a Monsanto company). Gellan gum is priced at about $ 53 per kg.

[Top of Page]

13. Galactomannans [2-6], [11], [298]

Alphons C. J. Voragen

Galactomannans are plant reserve carbohydrates, like starch, that occur in the endosperm of the seeds of many Leguminoseae. During sprouting, the galactomannans are enzymatically degraded and used as nutrition. Of the many known galactomannans, up to now only three have been processed and used on an industrial scale:

1. **Locust bean gum (carob gum)** has been known for a long time. It is derived from seeds of the carob tree (*Ceratonia siliqua* L.) which grows around the Mediterranean Sea, in Spain, Portugal, and Morocco, where more attention is paid to its systematic cultivation. The annual production is estimated at 10 000 – 12 000 t.
2. **Guar gum** is at present the most important galactomannan. It is derived from the seeds of the guar plant (*Cyamopsis tetragonoloba* L. Taub.), which grows in India and Pakistan. Since 1944, guar plants have also been cultivated on a large scale in the southern United States. They have been introduced into Brazil, Africa, and Australia, where two crops a year can be obtained. The industrial production of guar gum began around 1940. The annual production is estimated at ca. 125 000 t.
3. **Tara gum** has been produced only since the 1970s, to a much lesser degree (1000 t/a), from the seeds of the tara tree (*Cesalpina spinosa* L.), which occurs mainly in Peru.

The galactomannans on the market are flours prepared from the endosperm of the corresponding seeds. The carob and tara kernels are particularly difficult to process because of their tough, hard hulls. Remnants of the hulls and seeds are, therefore, always present in small amounts. The pure polysaccharide extracted from the flour of locust bean gum is often designated carobin gum and the polysaccharide from guar flour, guaran.

13.1. Structure

Galactomannans have a backbone of β-(1,4)-glycosidically linked mannopyranosyl residues. This backbone is substituted with single-unit α-D-galactopyranosyl residues linked to the O-6 of certain mannose moieties. This substitution renders the mannans soluble. Within the various mannans a wide spectrum of chemical structures occurs. This diversity of structure includes not only a wide variation in their mannose: galactose ratio, but also significant differences in the distribution of galactose units along the mannan backbone (Fig. 29) [299], [300]. For *locust bean gum* a mannose: galactose ratio of 3.5 has been established, compared to a ratio of 3 for *tara gum* and 1.5 for *guar gum*. A measure of the differences in the distribution of galactose side groups has been obtained from an examination of the degree of hydrolysis and the characteristic array of oligosaccharides obtained by digestion with an endomannanase. In this way, DEA et al. [301] established that a *tara gum* has a more statistically random distribution of side chains while *locust bean gum* has a nonregular, nonstatistically random distribution with a certain proportion of unsubstituted blocks of intermediate length. *Guar gum* was found to have few, if any, regions that were unsubstituted (smooth regions) with galactose. Galactomannans differ further in their molar mass distribution; the highest values were established for guar gum. These differences in fine structure between the galactomannans, involving the extent and pattern of galactosyl substitution and molar mass distribution, account for their very different functional properties.

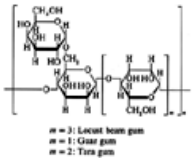

Figure 29. Structural features of galactomannans

13.2. Production [4], [11], [298]

Guar Gum. For the separation of the germ and the endosperm halves, guar seeds are first screened to remove foreign matter and then fed to an attrition mill to split the seeds in two endosperm halves; finally, the germ material is sifted off. Despite their relatively low galactomannan content the remaining endosperm halves, covered with hull (splits), are suitable for many technical applications. For higher-quality demands, and particularly for food applications, the endosperm must be further purified and liberated from the adhering hull. Various processes are used for this; in general, the splits are treated with moist or dry, hot air, which results in the loosening of the hull as a result of different swelling properties. After removal of the hull by sifting, pure endosperm is obtained with a galactomannan content of 85 – 95 % based on dry matter. By suitable milling and screening techniques the endosperm can be worked up to the desired commercial products.

Locust Bean and Tara Gum. Since the hull of carob and tara gum is very hard and tough, these gums are difficult to process. For removal of the hulls, two different processes exist. In the *acid process* the hulls are carbonized by treating the kernels with moderately dilute sulfuric acid at elevated temperature. Similar to the guar gum process, the remaining hull fragments are then removed by washing and brushing operations. After a drying step the dehulled kernels are cracked and the germs sifted off from the endosperm. Commercial products are obtained by milling and screening.

In the vanishing *roasting process* the kernels are roasted in a rotating furnace, where the hulls pop. The endosperm is then obtained as described above. This process yields products of somewhat darker color. However, the use of sulfuric acid is avoided.

13.3. Properties

The most prominent properties of galactomannans are their high water-binding capacity, their potential to form highly viscous solutions even at low concentrations, and the interaction of locust bean gum with carrageenans, agar, and xanthan gum. In double logarithmic plots of zero shear viscosity versus concentration multiplied by intrinsic viscosity, galactomannans exhibit a different type of viscosity behavior from most polysaccharides for which nonspecific physical entanglement between fluctuating random coil molecules is the underlying mechanism. Galactomannans show an earlier onset of concentrated solution behavior and a substantially greater concentration dependence thereafter [299], [300]. This behavior is attributed to the occurrence of interchain associations inherent to β-1,4-D-mannan chain segments. Such self-association has also been observed for β-1,4-D-glucan and β-1,4-D-xylan structures.

Locust bean gum and guar gum are both *water soluble*; they differ, however, in their dissolution behavior. Whereas guar gum is readily soluble in cold water, locust bean gum can be solubilized only by boiling. Solutions of both galactomannans are, depending on their purity, somewhat cloudy because of the presence of some proteins and crude fiber.

Even 1 wt % solutions are highly *viscous* (2000 – 6000 mPa · s, measured with an RTV Brookfield viscometer at 20 rpm); 2 – 5 wt % swollen dispersions show gellike behavior. The viscosity of solutions is hardly influenced by ionic strength. Locust bean gum solutions can gel at high

sodium hydroxide concentrations. Also boric acid and alkaline copper salt solutions cause flocculation or gelling of the galactomannans. Mixing of solutions of locust bean gum and xanthan gum results in an increase of the viscosity or even in gel formation (after boiling [209]). Such synergistic interactions have also been observed for combinations of locust bean gum with carrageenan or agar (Section Production). Dilute galactomannan solutions form elastic, transparent films upon dehydration.

Galactomannans are rapidly *degraded* in acidic, aqueous solutions at elevated temperature, resulting in a rapid loss of viscosity. They can also be degraded by oxidants and by microbial enzymes (mannanases); under alkaline conditions, they are rather stable [302]. When stored for longer periods of time, solutions of galactomannans must therefore be protected against microbial degradation by addition of preservatives.

13.4. Analysis and Composition of Commercial Preparations

Commercial galactomannan preparations still may contain remnants of germs and hull that are rich in protein and fiber, and reduce the galactomannan content. The galactomannan content is therefore an important factor in quality. This *gum content* is determined by analyzing the impurities and subtracting their total from 100 %. Depending on the application, products with different galactomannan content are used.

Another quality factor is the *viscosity* of the aqueous solution, usually measured in 1 wt % systems. Locust bean gum must first be solubilized by heating.

The *presence of galactomannans* can be demonstrated by various precipitation reactions (e.g., gelation with borax). For determination of the *mannose : galactose ratio*, the preparation is hydrolyzed and the sugar moieties are analzyed by using gas chromatography [15], [29].

13.5. Derivatives

As with cellulose, galactomannans contain primary and secondary hydroxyl groups, which, in principle, can be derivatized and substituted like cellulose. Only hydroxyalkyl, carboxymethyl, and cationic derivatives, as well as partially depolymerized products, are of industrial significance, however.

Hydroxyalkyl derivatives are obtained by treating galactomannans with ethylene or propylene oxide in an alkaline medium. These derivatives differ from the parent material by having a faster rate of hydration and being almost completely soluble.

Carboxymethylated products are obtained by “etherification” of a certain portion of the hydroxyl groups with monochloric acid or sodium monochloroacetate. Unlike carboxymethyl celluloses, which are soluble only at degrees of substitution in the range of 0.7 – 1.0, galactomannans — if not already soluble prior to derivatization — become completely soluble at a degree of substitution of <0.05.

Cationic derivatives are obtained by treatment of galactomannans with suitable reactive organic amines such as 2-hydroxy-3-chloropropyltrimethylammonium chloride or its reaction product 2,3-epoxypropyltrimethylammonium chloride, which forms in alkaline medium.

Partially depolymerized galactomannans with reduced viscosities are obtained by oxidative, hydrolytic (acid), or enzymatic degradation. Conversely, by reaction with bifunctional molecules such as epichlorohydrin, *cross-linked galactomannans* can be obtained, which — depending on their degree of cross-linking — are only partly water soluble or fully insoluble. Plant α-galactosidase that can remove galactose from galactomannans has been cloned in yeast and can be produced conveniently. This enzyme is active in systems containing up to 35 % galactomannan [303].

13.6. Applications

Guar gum has found application in many industries; its use is still increasing. Most uses are based on its good solubility in water and the high viscosity of aqueous solutions. In many applications, its inertness against high concentrations of salts of monovalent cations, its ability to form gels with certain multivalent cations, and its flocculating action are also decisive. *Derivatives of guar gum* are used in the oil industry in secondary or tertiary oil recovery of nearly exhausted oil wells. For applications in the textile industry (Table 2, groups 30 and 31), a variety of native and derivatized types are produced. They are also used in the manufacture of paper (group 29), explosives (group 26), and cables. In the latter applications, guar gum is added in advance to immobilize the inevitably penetrating water. They are also used in the pharmaceutical industry and as flocculating agents in the processing of minerals and effluent purification (group 32). In the food industry, guar gums are used as efficient thickeners of aqueous systems, thereby controlling the mobilization of water. This influences the consistency, body, and shelf life of aqueous food systems, as well as the stability of oil – water and water – oil emulsions. They are used in groups 3, 5, 6, 8, 9, and 12, and have also found application in pet food (group 13) and cattle feed.

Due to its higher price, *locust bean gum* is used mainly in the food industry, in the same groups as the guar gums. Of special interest for locust bean gum is its synergistic action with other hydrocolloids such as carrageenans, agar, and xanthan gum.

13.7. Market

Of the 125 000 t of *guar gum* produced annually, ca. one-third is used in the food industry; the remainder has found technical application. *Locust bean gum* is used almost completely in the food and pet food industry. The prices on the world market average ca. $ 1.1 – $ 1.3 per kilogram for guar gum and $ 6.4 – $ 7.5 per kilogram for locust bean gum.

[Top of Page]

14. Acknowledgement

Updated from the corresponding chapter in Ullmann's 4th ed. written by WALTER PILNIK.

[Top of Page]

References

1. M. Glicksman: *Gum Technology in the Food Industry*, Academic Press, New York 1969.
2. R. L. Whistler, J. N. BeMiller (eds.): *Industrial Gums*, 2nd ed., Academic Press, New York 1973.
3. H. D. Graham (ed.): *Food Colloids*, AVI, Westport, Connecticut 1977.
4. H. Neukom, W. Pilnik: *Gelier- und Verdickungsmittel in Lebensmitteln*, Forster Verlag, Zürich 1980.
5. G. O. Aspinall: *The Polysaccharides*, vol. 2, Academic Press, New York 1983.
6. P. Harris: *Food Gels*, Elsevier Applied Science Publishers, London 1990.
7. J. R. Mitchell, D. A. Ledwàrd: *Functional Properties of Food Macromolecules*, Elsevier Applied Science Publishers, London 1986.
8. M. L. Fishman, J. J. Jen: “Chemistry and Functions of Pectin,” *ACS Symp. Ser.* **310** (1986)
9. D. Oakenfull: “Gelling Agents,” *CRC Crit. Rev. Food Sci. Nutr.* **26** (1987), 1 – 25.
10. M. Yalpani: “Industrial Polysaccharides; Genetic Engineering, Structures/Property Relations and Applications,” *Prog. Biotechnol.* **3** (1987)
11. M. Glicksman: *Food Hydrocolloids*, CRC Press, Boca Raton Fla. 1986.
12. R. H. Walter: *The Chemistry and Technology of Pectin*, Academic Press, New York 1991.
13. W. Schmolck, E. Mergenthaler: “Beiträge zur Analytik von Polysacchariden, die als Lebensmittelzusatzstoffe verwendet werden,” *Z. Lebensm. Unters. Forsch.* **152** (1973) 87 – 99.

14. M. E. Endean: The Separation and Identification of Polysaccharide Stabilizers, Thickeners and Gums from Foods, Technical circular no. 575, The British Food Manufacturing Industrie Research Association, Leatherhead, Surrey 1974.
15. *Schweizerisches Lebensmittelbuch*, vol. 3, chap. 40, "Gelier- und Verdickungsmittel," Eidgen. Drucksachen- und Materialzentrale, Bern 1989.
16. R. G. Morley, G. O. Phillips, D. M. Power, *Analyst (London)* **97** (1972) 315 – 319.
17. J. E. Scott, D. Glick (eds.): *Methods of Biochemical Analysis*, Interscience, New York 1960, pp. 145 –197.
18. C. Rolin in R. L. Whistler, J. N. BeMiller (eds.): *Industrial Gums*, 3rd ed., Academic Press, San Diego 1993, pp. 257 – 293.
19. R. Ilker, A. S. Szcezesniak: "Structural and chemical bases for texture of plant foodstuffs," *J. Text. Studies.* **21** (1990) 1.
20. J. P. Van Buren: "The Chemistry of Texture in Fruits and Vegetables," *J. Text. Studies.* **10** (1979) 1.
21. M. McNeil, A. G. Darvill, S. C. Fry, P. Albersheim, "Structure and Function of the primary Cell Walls of Plants," *Ann. Rev. Biochem.* **53** (1984) 625.
22. R. R. Selvendran: "The Chemistry of Plant Cell Walls," in G. G. Birch, J. Parker (eds.): *Dietary Fibre*, Applied Sci. Publ., London 1995.
23. M. C. McCann, K. Roberts: *The Cytoskeletal Basis of Plant Growth and Form*, Academic Press, Oxford 1991, pp. 109 – 129.
24. C. M. G. C. Renard, H. A. Schols, A. G. J. Voragen, J. F. Thibault, W. Pilnik: "Studies on Apple Protopectin. III. Characterization of the Material Extracted by Pure Polysaccharidases from Apple Cell Walls," *Carbohydr. Polym.* **15** (1991) 13.
25. C. M. G. C. Renard, A. G. J. Voragen, J. F. Thibault, W. Pilnik: "Comparison between enzymatically and chemically extracted pectins from apple cell walls," *Animal Feed Sci. Technol.* **32** (1991) 69.
26. J. A. De Vries, F. M. Rombouts, A. G. J. Voragen, W. Pilnik: "Enzymic Degradation of Apple Pectins," *Carbohydr. Polym.* **2** (1982) 25.
27. R. R. Selvendran: "Development in the Chemistry and Biochemistry of Pectic and Hemicellulosic Polymers," *J. Cell Sci. Suppl.* **2** (1985) 51. Links
28. T. P. Kravtchenko, A. G. J. Voragen, W. Pilnik: "Analytical comparison ot three industrial pectin preparations," *Carbohydr. Polym.* **18** (1992) 17.
29. H. Thier: "Identification and Quantification of Natural Polysaccharides in Food Stuffs," in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry 2, Applications of Hydrocolloids*, Pergamon Press, Oxford 1984.
30. T. P. Kravtchenko, M. Penci, A. G. J. Voragen, W. Pilnik: "Enzymic and chemical degradation of some industrial pectins," *Carbohydr. Polym.* **20** (1993) 195.
31. R. H. Walter (ed.): *The Chemistry and Technology of Pectins*, Academic Press, New York 1991, p. 113.
32. J.- F. Thibault, C. M. G. C. Renard, M. A. V. Axelos, P. Roger, M.- J. Crepeau: "Studies of the length of homogalacturonic regions in pectins by acid hydrolysis," *Carbohydr. Res.* **238** (1993) 271. Links
33. J. A. de Vries: "Repeating units in the structure of pectin," in G. O. Phillips, P. A. Williams, D. J. Wedlock (eds.): *Gums and Stabilisers for the Food Industry*, vol. **4**, IRL Press, Oxford 1988, p. 25.
34. S. E. Harding, G. Berth, A. Ball, J. R. Mitchell, J. Garcia de la Torre: "The Molecular Weight Distribution and Conformation of Citrus Pectins in Solution Studies by Hydrodynamics," *Carbohydr. Polym.* **16** (1991) 1.
35. Z. I. Kertesz: *The Pectic Substances*, Interscience Publishers, New York 1951.
36. O. P. Beerh, B. Raghuramaiah, G. V. Krisnamurthy: "Utilization of mango waste: peel as a source of pectin," *J. Food Sci. Technol.* **13** (1976) 96.
37. R. M. Weightman, C. M. G. C. Renard, J.- F. Thibault: "Structure and properties of the polysaccharides from pea hulls. Part 1: chemical extraction and fractionation of the polysaccharides," *Carbohydr. Polym.* **24** (1994) 139.
38. M. J. Y. Lin, E. S. Humbert: "Extraction of pectins from sunflowers heads," *Can. Inst. Food Sci. Technol. J.* **11** (1978) 75.
39. W. J. Kim, F. Sosulski, S. C. K. Lee: "Chemical and gelation characteristics of ammonia-demethylated sunflower pectins," *J. Food Sci.* **43** (1978) 1436. Links
40. N. M. Ptitchkina, I. A. Danilova, G. Doxastakis, S. Kasapis, E. R. Morris: "Pumpkin pectin: gel formation at unusually low concentration," *Carbohydr. Polym.* **23** (1994) 265.
41. EC Proposal for Counsil Directive on Food Additives Other than Colours and Sweeteners (III/3624/91/EM-REV3) 1992.
42. G. H. Joseph, A. H. Keiser, E. F. Bryant: "High-polymer ammonia-demethylated pectinates and their gelation," *Food Technol.* **3** (1949) 85.
43. J. C. E. Reitsma, J.-F. Thibault, W. Pilnik: "Properties of amidated pectins. I. Preparation and characterization of amidated pectins and amidated pectic acids," *Food Hydrocoll.* **1** (1986) 121.
44. P. E. Christensen: "Methods of Grading Pectin in Relation to the Molecular Weight (Intrinsic Viscosity) of Pectin," *Food Res.* **19** (1954) 163.
45. A. Kawabata: "Studies on chemical and physical properties of pectic substances from fruits," *Mem. Tokyo Univ. Agr.* **19** (1977) 115.
46. W. J. Kim, V. N. M. Rao, C. J. B. Smit: "Effect of chemical composition on compressive mechanical properties of low ester pectin gels," *J. Food Sci.* **43** (1978) 572. Links
47. H. A. Deckers, C. Olieman, F. M. Rombouts, W. Pilnik: "Calibration and Application of High-Performance Size Exclusion Columns for Molecular Weight Distribution of Pectins," *Carbohydr. Polym.* **6** (1986) 361.
48. M. L. Fishman, D. T. Gillespie, S. M. Sondey, Y. S. El-Atawy: "Intrinsic Viscosity and Molecular Weight of Pectin Components," *Carbohydr. Res.* **215** (1991) 91. Links
49. M. A. V. Axelos, J.- F. Thibault, J. Lefebvre: "Structure of Citrus Pectins and Viscometric Study of their Solution Properties," *Int. J. Biol. Macromol.* **11** (1989) 186. Links
50. J.-F. Thibault, M. Rinaudo: "Interactions of mono- and divalent counterions with alkali- and enzyme-deesterified pectins in salt-free solutions," *Biopolymers* **24** (1985) 2131. Links
51. C. Garnier, M. A. V. Axelos, J.- F. Thibault: "Phase Diagrams of Pectin-Calcium Systems: Influence of pH, Ionic Strenght, and Temperature on the Gelation of Pectins with Different Degrees of Methylation," *Carbohydr. Res.* **240** (1993) 219. Links
52. H. S. Owens, H. Lotzkar, T. H. Schultz, W. D. Maclay: "Shape and Size of Pectinic Acid Molecules Deduced from Viscometric Measurements," *J. Am. Chem. Soc.* **68** (1946) 1628. Links
53. G. Berth, H. Anger, F. Linow: "Streulichtphotometrische und Viskosimetrische Untersuchungen an Pektinen in wäßrigen Lösungen zur Molmassenbestimmung," *Nahrung* **21** (1977) 939. Links
54. H. Anger, G. Berth: "Gel Permeation Chromatography and the Mark Houwink Relation for Pectins with Different Degrees of Esterification," *Carbohydr. Polym.* **6** (1986) 193.
55. G. Ravanat, M. Rinaudo: "Investigation on Oligo- and Polygalacturonic Acids by Potentiometry and Circular Dichroism," *Biopolymers* **19** (1980) 2209. Links
56. F. Michel, J.- F. Thibault, J.- L. Doublier: "Viscometric and Potentiometric Study of High-Methoxyl Pectins in the Presence of Sucrose," *Carbohydr. Polym.* **4** (1984) 283.
57. E. Racapé, J.- F. Thibault, J. C. E. Reitsma, W. Pilnik: "Properties of Amidated Pectins. II. Polyelectrolyte Behavior and Calcium Binding of Amidated Pectins and Amidated Pectic Acids," *Biopolymers* **28** (1989) 1435.
58. I. G. Plaschina, E. E. Braudo, V. B. Tolstoguzov: "Circular-Dichroism Studies of Pectin Solutions," *Carbohydr. Res.* **60** (1978) 1. Links
59. A. Cesàro, F. Delben, S. Paoletti: "Thermodynamics of the Proton Dissociation of Natural Polyuronic Acids," *Int. J. Biol. Macromol.* **12** (1990) 170. Links
60. M. D. Walkinshaw, S. Arnott: "Conformations and interactions of pectins, I. X-ray diffraction analysis of sodium pectate in neutral and acidified forms," *J. Mol. Biol.* **153** (1981) 1055. Links

61. M. J. Gidley, E. R. Morris, E. J. Murray, D. A. Powell, D. A. Rees: “Spectroscopic and stoichiometric characterization of the calcium-mediated association of pectate chains in gels and in the solid state,” *J. Chem. Soc. Chem. Commun.* **22** (1979) 990.
62. FAO Food and Nutrition Papers 34 and 37, Specifications for Identity and Purity of Certain Food Additives. Distribution and Sales Section, FAO, Rome 1986.
63. Food Chemicals Codex, 3rd ed., National Academy of Sciences, Washington, D.C. 1992.
64. J. N. BeMiller: “Acid-catalyzed hydrolysis of glycosides,” *Adv. Carbohydr. Chem.* **22** (1967) 25.
65. D. A. Powell, E. R. Morris, M. J. Gidley, D. A. Rees: “Conformations and Interactions of Pectins, II. Influence of Residue Sequence on Chain Association in Calcium Pectate Gels,” *J. Mol. Biol.* **155** (1982) 517. Links
66. P. Albersheim, H. Neukom, H. Deuel: “Splitting of pectin chain molecules in neutral solutions,” *Arch. Biochem. Biophys.* **90** (1960) 46. Links
67. W. Heri, H. Neukom, H. Deuel, *Helv. Chim. Acta* **44** (1961) 1939. Links
68. R. Kohn, I. Furda: “Distribution of free carboxyl groups in the molecule of pectin after esterification of pectic and pectinic acid by methanol,” *Collect. Czech. Chem. Commun.* **34** (1969) 641. Links
69. E. L. Pippen, R. M. McCready, H. S. Owens: “Gelation properties of partially acetylated pectins,” *J. Am. Chem. Soc.* **72** (1950) 813. Links
70. L. W. Doner: “Analytical methods for determining pectin composition,” *A.C.S. Symp. Ser.* **310** (1986) 13.
71. FAO, Food and Nutrition Paper 52, Addendum 1, Rome 1992, p. 87.
72. *Food Chemicals Codex*, FCC III Monographs, 3rd ed. (including supplements), National Academy Press, Washington D.C. 1981, p. 215.
73. EEC, Council Directive 78/663, *Off. J. EEC* 14 08 78 (plus Updates), 1978.
74. Z. I. Kertesz: *The Pectic Substances*, Interscience, New York 1951.
75. A. Plöger: “Conductivity detection of pectin: a rapid HPLC method to analyze degree of esterification,” *J. Food Sci.* **57** (1992) 1185.
76. H.- J. Zhong, M. A. K. Williams, R. D. Keenan, D. M. Goodall, C. Rolin: *Carbohydr. Polym.* **32** (1997) 27.
77. G. Berth: “Studies on the Heterogeneity of Citrus Pectin by Gel Permeation Chromatography on Sepharose 2B/Sepharose 4B,” *Carbohydr. Polym.* **8** (1988) 105.
78. R. C. Jordan, D. A. Brant: “An Investigation of Pectin and Pectic Acid in Dilute Aqueous Solution,” *Biopolymers* **17** (1978) 2885. Links
79. S. Sawayama, A. Kawabata, H. Nakahara, T. Kamata: “A Light Scattering Study on the Effects of pH on Pectin Aggregation in Aqueous Solution,” *Food Hydrocoll.* **2** (1988) 3.
80. R. H. Walter, H. L. Matias: “Pectin Aggregation Number by Light Scattering and Reducing End-Group Analysis,” *Carbohydr. Polym.* **15** (1991) 33.
81. W. D. Holloway, C. Tasman-Jones, K. Maher: “Pectin digestion in humans,” *Am. J. Clin. Nutr.* **37** (1983) 253. Links
82. E. C. Titgemeyer, L. D. Burquin, G. C. Fahey, Jr., K. A. Garleb: “Fermentability of various fiber sources by human fecal bacteria in vitro,” *Am. J. Clin. Nutr.* **53** (1991) 1418. Links
83. M. Nyman, N.- G. Asp: “Fermentation of dietary fibre components in the rat intestinal tract,” *Br. J. Nutr.* **47** (1982) 357. Links
84. M. Nyman, N.- G. Asp, J. Cummings, H. Wiggins: “Fermentation of dietary fiber in the intestinal tract: comparison between man and rat,” *Brit. J. Nutr.* **55** (1986) 487. Links
85. J. J. Doesburg: “Pectic Substances in Fresh and Preserved Fruits and Vegetables,” I.B.V.T. Communication no. 25, Sprenger Institute, Wageningen 1965.
86. R. A. Baker: “Potential Dietary Benefits of Citrus Pectin and Fiber,” *Food Technol.* (1994) Nov., 133.
87. A. Siddhu, S. Sud, R. L. Bijlani, M. G. Karmarkar, U. Nayar: “Nutrient interaction in relation to glycaemic and insulaemic response,” *Indian. J. Physiol. Phamacol.* **36** (1992) 21 – 28.
88. D. J. A. Jenkins, *et al.*: “Pectin and complications after gastric surgery: normalisation of postprandial glucose and endocrine responses,” *Gut* **21** (1980) 574. Links
89. S. E. Schwartz, R. A. Levine, A. Singh, J. R. Scheidecker, N. S. Track: “Sustained Pectin Ingestion Delays Gastric Emptying,” *Gastroenterology* **83** (1982) 812.
90. S. Satchithanandam, M. Vargofcak-Apker, R. J. Calvert, A. R. Leeds, M. M. Cassidy: “Alteration of gastrointestinal mucin by fiber feeding in rats,” *J. Nutr.* **120** (1990) 1179. Links
91. C. D. Lorenzo, C. M. Williams, F. Hajnal, J. E. Valenzuela: “Pectin Delays Gastric Emptying and Increases Satiety in Obese Subjects,” *Gastroenterology* **95** (1988) 1211.
92. D. M. Zimmaro, *et al.*: “Isotonic Tube Feeding Formula Induces Liquid Stool in Normal Subjects: Reversal by Pectin,” *J. Parenteral and Enteral Nutr.* **13** (1989) 117.
93. X.-B. Sun, T. Matsumoto, H. Yamada: “Anti-Ulcer Activity and Mode of Action of the Polysaccharide Fraction from the Leaves of Panax ginseng,” *Planta Med.* **58** (1993) 432. Links
94. H. Yamada, M. Hirano, H. Kiyohara: “Partial structure of an anti-ulcer pectic polysaccharide from the roots of Bupleurum falcatum L.,” *Carbohydr. Res.* **219** (1991) 173. Links
95. X.-B. Sun, T. Matsumoto, H. Yamada: “Effects of a Polysaccharide Fraction from the Roots of Bupleurum falcatum L. on Experimental Gastric Ulcer Models in Rats and Mice,” *J. Pharm. Pharmacol.* **43** (1991) 699. Links
96. A. Kawabata: *Studies on Chemical and Physical Properties of Pectic Substances from Fruits*, Memoirs of the Tokyo University of Agriculture XIX, 1977, pp. 115 – 200.
97. L. M. Drews, C. Kies, H. M. Fox: “Effect of dietary fiber on copper, zinc, and magnesium utilization by adolescent boys,” *Am. J. Clin. Nutr.* **32** (1979) 1893. Links
98. K. Y. Lei, M. W. Davis, M. M. Fang, L. C. Young: “Effect of pectin on zinc, copper and iron balances in humans,” *Nutr. Rep. Int.* **22** (1980) 459.
99. J. H. Cummings, *et al.*: “The digestion of pectin in the human gut and its effect on calcium absorption and large bowel function,” *Br. J. Nutr.* **41** (1979) 477. Links
100. P.- E. Glahn: “Hydrocolloid stabilization of protein suspensions at low pH,” *Prog. Fd. Nutr. Sci.* **6** (1982) 171.
101. A. Parker, P. Boulenguer, T. P Kravtchenko: “Effect of the addition of high methoxyl pectin on the rheology and colloidal stability of acid milk drinks,” in K. Nishinari, E. Doi (eds.): *Food Hydrocolloids, Structures, Properties and Functions*, Plenum Press, New York 1994, p. 307.
102. P.- E. Glahn, C. Rolin: “Casein-pectin interaction in sour milk beverages,” *Food Ingred. Eur. Conf. Proc.* 1994, 252.
103. B. U. Nielsen: “Fiber-Based Fat Mimetics: Pectin,” in S. Roller, S. Jones (eds.): *Handbook of Fat Replacers*, CRC Press, Boca Raton FL, 1996, p. 161.
104. W. Pilnik, P. Zwiker: “Pektine,” *Gordian* **70** (1970) 202 – 204, 252 – 257, 302 – 305, 343 – 346.
105. D. B. Nelson, C. J. B. Smit, R. R. Wiles in H. D. Graham (ed.): *Food Colloids*, Avi, Westport, Connecticut 1977, p. 418.
106. C. Rolin, J. De Vries in P. Harris (ed.): *Food Gels*, Elsevier Applied Science, London 1990, pp. 401 – 434.
107. G. O. Aspinall in J. Preiss (ed.): *The Biochemistry of Plants*, vol. **3**, Academic Press, New York 1980, p. 473.
108. J. N. BeMiller, *ACS Symp. Ser.* **310** (1986) 12.
109. J. K. Pederson in R. L. Davidson (ed.): *Handbook of Water Soluble Gums and Resins*, Chap. 15 – 1, McGraw-Hill, New York 1980.
110. S. Hojgaard Christensen: “Pectins” in M. Glicksman (ed.): *Food Hydrocolloids*, vol. 3, CRC Press, Boca Raton, Fla. 1986, p. 205.
111. W. Pilnik: “Pectin – a Many Splendoured Thing” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry 5*, IRL Press, Oxford 1990, pp. 209 – 221.

112. H. Braconnot, *Ann. Chim. Phys. Ser. 2* **28** (1825) 173 – 178.
113. W. Pilnik, A. G. J. Voragen: "Pectic Substances and Other Uronides," in A. C. Hulme (ed.): *The Biochemistry of Fruits and their Products*, vol. 1, Academic Press, London 1970, pp. 53 – 87.
114. F. M. Rombouts, W. Pilnik: "Pectic Enzymes" in A. H. Rose, (ed.): *Economic Microbiology*, vol. 5, "Enzymes and Enzymic Conversions," Academic Press, London 1980, pp. 225 – 280.
115. A. G. J. Voragen, W. Pilnik: "Pectin-Degrading Enzymes in Fruit and Vegetable Processing," *ACS Symp. Ser.* **389** (1989) 93 – 115.
116. W. Pilnik, A. G. J. Voragen: "The Significance of Endogenous and Exogenous Pectic Enzymes in Fruit and Vegetable Processing" in P. F. Fox (ed.): *Food Enzymology*, vol. 1, Elsevier Applied Science, London 1991, pp. 303 – 337.
117. C. D. May: "Industrial Pectins: Sources, Production and Application," *Carbohydr. Polymers* **12** (1990) 79 –90.
118. E. M. McComb, R. M. McCready, *Anal. Chem.* **29** (1957) 819 – 821. Links
119. S. C. Frey: "Cross-Linking of Matrix Polymers in the Growing Cell Wall of Angiosperms," *Ann. Rev. Plant. Physiol.* **37** (1986) 165 – 186.
120. D. H. Northcote, *ACS Symp. Ser.* **310** (1986) 134.
121. L. Rexova-Benkova, O. Markovic, *Adv. Carbohyd. Chem. Biochem.* **33** (1976) 101 – 130.
122. H. A. Schols *et al.*, *Carbohydr. Res.* **206** (1990) 105 –115. Links
123. M. J. F. Searle-van Leeuwen *et al.*, *Eur. J. Appl. Microbiol. Biotechnol.* **38** (1992) 347 – 349.
124. W. Pilnik: "Pektine" in R. Heiss (ed.): *Lebensmitteltechnologie*, Springer Verlag, Berlin 1988, pp. 228 – 234.
125. M. A. Joslyn, G. de Luca, *J. Colloid Sci.* **12** (1957) 108 – 130.
126. M. Manabe, *Nippon Nogei Kagaku Kaishi* **45** (1971) no. 4, 195 – 199.
127. Sunkist Growers, US 1 332 985, 1971.
128. California Fruit Growers Exchange, US 2 480 710, 1949.
129. W. J. Kim, C. J. B. Smit, V. N. M. Rao, *J. Food Sci.* **43** (1978) 74 – 78. Links
130. W. Pilnik, A. G. J. Voragen: "Gelling Agents (Pectins) from Plants for the Food Industry," *Adv. Plant Cell Biochem. Biotechnol.* **1** (1992) 219 – 270.
131. W. Pilnik, R. A. MacDonald, *Gordian* **68** (1968) 531 – 535.
132. M. Byland, A. Donutzhuber, *Sven. Papperstidn.* **15** (1968) 505 – 508.
133. T. P. Kravtchenko, I. Arnould, A. G. J. Voragen, W. Pilnik, *Carbohyd. Polym.* (in press).
134. G. H. Joseph, A. H. Kieser, E. F. Bryant, *Food. Technol (Chigago)* **3** (1949) 85 – 90.
135. M. Manabe, *Nippon Nogie Kogaku Kaishi* **45** (1971) no. 9, 417 – 422.
136. F. M. Rombouts, A. K. Wissenburg, W. Pilnik, *J. Chromatogr.* **168** (1979) 151 – 161. Links
137. W. Somers, J. Visser, F. M. Rombouts, K. van t'Riet: "Developments in Downstream Processing of (Poly)saccharide Converting Enzymes," *J. Biotechnol.* **11** (1989) 199 – 222. Links
138. H. Neukom, *Schweiz. Landwirtsch. Forsch.* **2** (1963) 112 – 122.
139. H. F. Launer, Y. Tomimatsu, *Anal. Chem.* **31** (1959) no. 9, 1569 – 1574. Links
140. D. A. Rees, *Adv. Carbohydr. Chem. Biochem.* **24** (1969) 267 – 332. Links
141. D. A. Rees, E. R. Morris, D. Thom, J. K. Madden in G. O. Aspinall (ed.): *Polysaccharides*, vol. **1**, Academic Press, New York 1982, pp. 195 – 290.
142. D. A. Rees, A. W. Wight, *J. Chem. Soc. B* 1971, 1366.
143. D. Oakenfull, A. Scott: "New Approaches to the Study of Food Gels," in G. O. Philips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, vol. 3, Elsevier Applied Science Publishers **1985,** London, pp. 465 – 475.
144. E. R. Morris *et al.*, *Int. J. Biol. Macromol.* **2** (1980) 237.
145. D. Thom, I. C. M. Dea, E. R. Morris, D. A. Powell, *Progr. Fd. Nutr. Sci.* **6** (1982) 87 – 108.
146. V. J. Morris in J. R. Mitchell, D. A. Ledward (eds.): *Functional Properties of Food Macromolecules*, Elsevier, Amsterdam 1986, pp. 121 – 169.
147. Racapé *et al.*, *Biopolymers* **28** (1989) 1435 – 1448.
148. B. Lockwood, *Food Process. Ind.* **41** (1972) 493, 47 –51.
149. I. F. T. Pectin Standardization, *Food Technol. (Chicago)* **13** (1959) 496 – 500.
150. P. C. Crandall, L. Wicker: "Pectin Internal Gel Strength: Theory, Measurement and Methodology," *ACS Symp. Ser.* **310** (1986) 79 – 102.
151. M. J. H. Keijbets, W. Pilnik, *Potato Res.* **17** (1974) 169 – 177. Links
152. M. F. Katan, P. v.d. Bovenkamp, *Basic Clin. Nutr.* **3** (1981) 217 – 239.
153. R. F. McFeeters, S. A. Armstrong, *Anal. Biochem.* **139** (1984) 212 – 217. Links
154. L. G. Bartolome, J. E. Hoff, *J. Agric. Food Chem.* **20** (1972) 266 – 270. Links
155. M. A. Litchman, R. P. Upton, *Anal. Chem.* **44** (1972) 1495 – 1497. Links
156. G. C. Cochrane, *J. Chromatogr. Sci.* **13** (1975) 440 –447. Links
157. E. L. Pippen, R. M. McCready, H. S. Owen, *Anal. Chem.* **22** (1950) 1457 – 1458. Links
158. A. G. J. Voragen, H. A. Schols, W. Pilnik, *Food Hydrocolloids* **1** (1986) 65 – 70. Links
159. FAO Food and Nutrition Paper 31/2 (Rome) 1984, p. 75.
160. J. C. E. Reitsma, W. Pilnik: "Analysis of Mixtures of Pectins and Amidated Pectins," *Carbohydr. Polym.* **10** (1989) 315 – 319.
161. T. Bitter, H. M. Muir, *Anal. Biochem.* **4** (1962) 330. Links
162. N. Blumenkranz, G. Asboe-Hansen, *Anal. Biochem.* **54** (1973) 484 – 489. Links
163. T. M. C. C. Filisetti-Cozzi, N. C. Carpita, *Anal. Biochem.* **197** (1991) 157 – 162. Links
164. T. P. Kravtchenko, A. G. J. Voragen, W. Pilnik, *Carbohydr. Polym.* **18** (1992) 17 – 25.
165. R. R. Selvendran, P. Ryden: "Isolation and Analysis of Plant Cell Walls," in P. M. Dey, J. B. Harborne (eds.): *Methods in Plant Biochemistry* **2** (1990) 549 –579.
166. C. M. G. C. Renard, A. G. J. Voragen, J. F. Thibault, W. Pilnik, *Carbohydr. Polym.* **12** (1990) 9 –25.
167. H. U. Endress: in [12], pp. 251 – 267.
168. P. A. Sanderson, J. Baird: "Industrial Utilization of Polysaccharides" in G. O. Aspinall (eds.): *The Polysaccharides*, vol. 2, Academic Press, London 1983, pp. 411 – 490.
169. T. J. Painter in G. O. Aspinall (ed.): *The Polysaccharides*, vol. **2**, Academic Press, London 1983, pp. 257 – 275.
170. P. A. Sandford in R. R. Colwell, E. R. Pariser, R. J. Sinskey (eds.): *Biotechnology of Marine Polysaccharides* Hemisphere, Washington 1985, pp. 454 – 516.
171. W. J. Sime. Alignates in P. Harris (ed.): *Food Gels*, Elsevier Applied Science Press, New York 1990, pp. 53 – 78.
172. W. Pilnik, A. G. J. Voragen: "Pektine und Alginate" in H. Neukom and W. Pilnik (eds.): *Gelier- und Verdickungsmittel in Lebensmitteln*, Forster Verlag, Zürich 1980, pp. 67 – 94.
173. E. Percival, R. H. McDowell: "Algal Polysaccharides," *Methods in Plant Biochemistry* **2** (1990) 523 – 548.

174. V. J. Chapman: *Seaweeds and Their Uses*, 2nd ed., Methuen & Co. Ltd., London 1970.
175. A. Haug: “Composition and Properties of Alginates,” Report no. 30, Norwegian Institute for Seaweed Research, Trondheim 1964.
176. E. C. C. Stanford, GB 142, 1881.
177. F. G. Fischer, H. Dörfel, *Hoppe Seyler's Z. Physiol. Chem.* **302** (1955) 186 – 203.
178. A. Haug, B. Larsen, O. Smidsrod, *Acta Chem. Scand.* **21** (1967) 691 – 704.
179. W. A. P. Black, *J. Mar. Biol. Assoc. U.K.* **29** (1950) 45 – 72. Links
180. H. A. Hoppe, O. J. Schmid, *Bot. Mar. Suppl.* **3** (1962) 16 – 66.
181. A. Haug, B. Larsen, *Acta Chem. Scand.* **16** (1962) 1908 – 1918.
182. A. Penman, G. R. Sanderson, *Carbohydr. Res.* **25** (1972) 273 – 285. Links
183. D. F. Pindar, C. Bucke, *Biochem. J.* **152** (1975) 617 – 622. Links
184. G. Skjäk-Break, B. Larsen, *Carbohydr. Res.* **139** (1985) 273 – 283.
185. Marine Colloids, US 2 128 551, 1938.
186. Kelco, US 2 036 934, 1936.
187. A. B. Steiner, W. H. McNeely, US 2 494 911, 1950.
188. A. P. Imeson in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry 2*, Pergamon Press, Oxford 1984, p. 189.
189. A. Haug, B. Larsen, O. Smidsrod, *Acta Chem. Scand.* **20** (1966) 183 – 190.
190. A. Haug, B. Larsen, O. Smidsrod, *Acta Chem. Scand.* **21** (1967) 2859 – 2870.
191. J. Preiss, G. Ashwell, *J. Biol. Chem.* **237** (1962) 309 – 316. Links
192. M. Jefferies, G. Pass, G. O. Phillips, *J. Sci. Fd. Agric.* **28** (1977) 173 – 179.
193. O. Smidsrod, *Faraday Discuss. Chem. Soc.* **57** (1974) 263.
194. D. Thom, I. C. M. Dea, E. R. Morris, D. A. Powell, *Progr. Food Nutr. Sci.* **6** (1982) 97.
195. K. Taft, *Progr. Food Nutr. Sci.* **6** (1982) 89.
196. H. A. Schols, *Analyt. Biochem.* (in press).
197. H. Grasdalen, *Carbohydr. Res.* **118** (1983) 225 – 260.
198. J. S. Craigie, E. R. Morris, D. A. Rees, D. Thom, *Carbohydr. Polym.* **4** (1984) 237 – 252.
199. G. A. Towle in [281], pp. 237 – 252.
200. J. K. Pedersen in [283], pp. 113 – 133.
201. N. F. Stanley in [285], pp. 79 – 119.
202. Marine Colloids, US 3 094 517, 1963.
203. O. Smidsrod, B. Larsen, A. Pernas, A. Haug, *Acta Chem. Scand.* **21** (1967) 2585 – 2598.
204. C. J. Lawson, D. A. Rees, *Nature (London)* **227** (1970) 390 – 395.
205. E. R. Morris, D. A. Rees, G. R. Robinson, *J. Mol. Biol.* **138** (1980) 349. Links
206. O. Smidsrod, H. Grasdalen, *Carbohydr. Polym.* **2** (1982) 270.
207. I. C. M. Dea *et al.*, *Carbohydr. Res.* **57** (1977) 249–272. Links
208. B. V. McCleary, *Carbohydr. Res.* **71** (1977) 249.
209. E. R. Morris, (1990) In *Food Gels*, (Ed P. Harris), Applied Science Publishers, London , pp. 291–359.
210. Marinalg International, world association of seaweed processors, Paris, www.marinalg.org.
211. E. Percival, R. H. Mc Dowel, “Algal Polysaccharides,” in P. M. Dey, J. B. Harborne (eds.): *Plant Biochemistry*, Vol. **2**, Carbohydrates, Academic Press, New York 1990, pp. 536 – 537.
212. H. H. Selby, W. H. Wynne, “Agar in Industrial Gum., ” in R. L. Whistler (ed.): *Polysaccharides and Their Derivatives*, 2nd ed., Academic Press, New York 1973, pp. 29 – 48.
213. C. Araki, *J. Chem. Japan* **58** (1973) 1338.
214. C. Araki, *Bull. Chem. Soc. Japan* **29** (1956) 543.
215. C. Araki, S. Hirase, *Bull. Chem. Soc. Japan* **33** (1960) 291.
216. C. Araki, K. Arai, *Bull. Chem. Soc. Japan* **30** (1957) 287.
217. M. Duckwork, K. C. Hong, W. Yaphe, *Carbohydr. Res.* **18** (1971) 1 – 9.
218. C. Araki, *Proc. 5th Intern. Seaweed Symp.*, Pergamon Press, Oxford 1966, p. 3.
219. S. Hirase, *Bull. Chem. Soc. Japan* **30** (1957) 68.
220. M. Duckwork, W. Yaphe, *Carbohydr. Res.* **16** (1971) 189.
221. K. B. Guisely, *Carbohydr. Res.* **13** (1970) 247 – 256.
222. D. A. Rees, *Chem. Ind.* (1972) 630 – 636. Links
223. M. Glicksman, Polysaccharides in Food, ed. J. M. V and Mitchell, J. R., In Blanshard (1979) 185 – 24, Butterworths, London.
224. A. Hayashi, *Polym. J.* **9** (1977) 219 – 225. Links
225. A. Hayashi, *Polym. J.* **10** (1978) 485 – 494. Links
226. A. Hayashi, *Polym. J.* **12** (1980) 447 – 453. Links
227. J. P. Flory, *Proc. R. Soc. London. Ser. A* **234** (1956) 73.
228. E. Pines, W. Prins *Macromolecules* **6** (1973) no. 6 888 – 895. Links
229. D. A. Rees, *Adv. Carbohyd. Chem. Biochem.* **24** (1969) 267.
230. D. A. Rees, *J. Chem. Soc.* (1961) 5168.
231. C. R. Williams: “The Processing of Gum Arabic to give Improved Functional Properties,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 37 – 40.
232. A. P. Imeson: “Exudate Gums,” in A. Imseon (ed.): *Thickening and Gelling Agents for Food*, Blackie Academic and Professional, London 1992, pp. 66 – 97.
233. A. M. Stephen: “Structure and Properties of Exudate Gums,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 3 – 16.
234. S. Conolly, J. C. Fenyo, M. C. Vandevelde, *Food Hydrocolloids* **1** (1987) 477 – 480.
235. J. C. Fenyo, M. C. Vandevelde: “Physico-Chemical Properties of Gum Arabic in Relation to Structure,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabiliserer for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 17 – 23.
236. P. A. Williams, G. O. Phillips, R. C. Randall: “Structure-Relationships of Gum Arabic,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 25 – 36.
237. D. M. W. Anderson, *Food Hydrocolloids* **2** (1988) 417 – 423. Links

238. G. O. Aspinall, V. P. Bhavanadan, T. B. Christensen, *J. Chem. Soc.* (1965) 2677 – 2684.
239. B. Urlacher, B. Dalbe: “Xanthan Gum,” in A. Imeson (ed.): *Thickening and Gelling Agents for Food*, Blackie Academic and Professional, London 1992, pp. 202 – 226.
240. A. R. Jeanes, J. E. Pittsley, F. R. Santi, *J. Appl. Polymer Sci.* **5** (1961) 519 – 526.
241. P. Delest: “Fermentation Technology of Microbial Polysaccharides,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 301 – 313.
242. A. A. Lawrence: *Edible Gums and Related Substances*, Noyes Data Corp., Park Ridge, N.J. 1973.
243. P. E. Janson, L. Keene, B. Lindberg, *Carbohydr. Res.* **45** (1975) 275 – 282. Links
244. R. Moorhouse, M. D. Walkinshaw, S. Anott: “Xanthan gum – molecular conformation and interactions,” in P. A. Sandford (ed.): *Extracellular Microbial Polysaccharides. ACS Symposium Series*, No. 45.: American Chemical Society, Washington, D.C. 1977, pp. 90 – 102.
245. R. Moorhouse, in A. P. Imeson (ed.): *Thickening and Gelling Agents for Food*, "Chap. 9", Blackie A & O, Glasgow 1992, pp. 202 – 226.
246. M. R. Betlach et al.: “Genetic Engineering, Sturcture/Property Relations and Applications,” in M. Yalpani (ed.): *Industrial Polysaccharides*, Elsevier, Amsterdam 1987, pp. 35 – 50.
247. D. J. Petit: “Xanthan gum,” in M. Glicksman (ed.): *Food Hydrocolloids*, vol. **1**, CRC Press, Boca Raton, FL 1982, pp. 127 – 149.
248. P. J. Whitcomb, C. W. Makosko, *J. Rheol.* **22** (1978) no. 5, 493 – 505. Links
249. Whitcomb, C. W. Makosko, “Rheology of xanthan gum solutions,” *J P.J. Rheol.* **22** (1978) no. 5, 493 – 505.
250. Kelco Co.: “Xanthan gum Keltrol/Kelzan, in Natural Biopolysaccharides for Scientific Water Control,” 2nd ed., Kelco Co., San Diego, CA 1975.
251. R. A. Speers, M. A. Tung: “Concentration and temperature dependence of flow behaviour of xanthan gum dispersions,” *J. Food Sci.* **51** (1986) no. 1, 96 – 98, 103. Links
252. B. Launay, J. L. Doublier, G. Cuvelier, in J. R. Mitchell and D. A. Ledward (eds.): *Functional Properties of Macromolecules*, Elsevier Applied Science, London 1986, pp. 1 – 78.
253. W. H. McNeely, K. S. Kang in R. J. Whistler, N. J. BeMiller (eds): *Industrial Gums*, Academic Press, New York 1973, pp. 473 – 521.
254. A. R. Jeanes, J. E. Pittsley, F. R. Samtis: “Polysaccharide B-1459; a new hydrocolloid polyelectrolyte produced from glucose by bacterial fermentation,” *J. Appl. Polym. Sci.* **5** (1961) 519 – 526. Links
255. E. R. Morris *et al.*: “Order – disorder transition for a bacterial polysaccharide in solution: a role for polysaccharide conformation in recognition between Xanthomonas pathogen and its plant host,” *J. Mol. Biol.* **110** (1977) no. 1, 1 – 16. Links
256. M. Rinaudo, M. Milos, *Int. J. Biol. Macromol.* **2** (1980) 45 – 48. Links
257. G. A. R. Carnie: “Evaluation of a new polysaccharide gum,” *Aust. J. Pharm.* **45** (1964) no. 19, 580 – 583.
258. I. T. Norton *et al.*, *J. Mol. Biol.* **175** (1984) 371. Links
259. V. J. Morris, *Food Biotechnol.* **4** (1990) 45 – 57. Links
260. V. J. Morris: “Science, Structure and Applications of Microbial Polysaccharides,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 315 – 328.
261. A. W. Thomas, H. A. Murray: “Gum arabic,” *J. Phys. Chem.* **32** (1928) no. 6, 676 – 697.
262. US 3 413 125, 1971 (H. R. Schuppner).
263. US 3 721 571, 1973 (M. Glicksman, H. E. Farks).
264. R. O. Mannion *et al.*: “Xanthan/locust bean gum interactions at room temperature,” *Carbohydr. Polym.* **19** (1992) no. 2, 91 – 97.
265. P. B. Fernandez: “Influence of galactomannan on the structure and thermal behaviour of xanthan/galactomannan mixtures,” *J. Food. Eng.* **2** (1995) 269 – 283.
266. H. Luyten, W. Kloek, T. Van Vliet: “Yielding behaviour of mixtures of xanthan and enzyme-modified galactomannans,” *Food Hydrocolloids* **8** (1994) no. 5, 431 – 440. Links
267. B. Dalbe: “Interactions between xanthan gum and konjac mannan,” in G. O. Phillips, D. J. Williams (eds.): *Gums and Stabilisers for the Food Industry*, vol. **6**, Oxford University Press, Oxford 1992 pp. 201 – 208.
268. G. R. Sanderson: “The interaction of xanthan gum in food systems,” *Prog. Food. Nut. Sci.* **6** (1982) 77 – 87.
269. G. R Sanderson: “Gellan Gum,” in [285] , pp. 201 –232.
270. W. Gibson: “Gellan Gum,” in A. Imeson (ed.): *Thickening and Gelling Agents for Food*, Blackie Academic and Professional, London 1992, pp. 227 – 249.
271. K. S. Lang *et al.*: “Agar-like polysaccharide produced by Pseudomonas species: production and basic properties,” *Appl. Environ. Microbiol.* **43** (1982) 1086.
272. “Gellan Gum: Multifunctional Polysaccharide for Gelling and Texturising,” NutraSweet Kelco publication, 1994.
273. H. Grasdalen, O. Smidsrod: “Gelation of gellan gum,” *Carbohydr. Polym.* **7** (1987) 371.
274. M. S. Kuo, A. Dell, A. J. Mort: “Identification and location of L-glycerate, an unusual substituent in gellan gum,” *Carbohydr. Res.* **156** (1968) 173.
275. C. Upstill, E. D. T. Atkins, P. T. Atwool: “Helical conformations of gellan gum,” *Int. J. Biol. Macromol.* **8** (1986) 275. Links
276. R. Chandrasekaran, A. Radha, V. G. Thailambal: “Roles of potassium ions, acetyl and L-glyceryl groups in native gellan double helix: an x-ray study,” *Carbohydr. Res.* **224** (1992) 1. Links
277. R. Chandrasekaran, R. P. Millane, S. Arnott: “Crystal structure of gellan,” *Carbohydr. Res.* **175** (1988) 1. Links
278. R. Chandrasekaran, L. C. Puigjaner, K. L. Joyce, S. Arnott: “Cation interactions in gellan: an x-ray study of the potassium salt,” *Carbohydr. Res.* **181** (1988) 23. Links
279. R. Chandrasekaran, V. G. Thilambal, *Carbohydr. Polym.* **12** (1990) no. 4, 431 – 442.
280. G. R. Sanderson, R. C. Clark: “Gellan Gum”, *Food Technol.* **37** (1983) 63.
281. G. R. Sanderson, V. L. Bell, R. C. Clark, D. Ortega: “The texture of gellan gum gels,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 4th ed., IRL Press, Oxford 1988, p. 219.
282. G. R. Sanderson, V. L. Bell, D. Ortega: “A comparison of gellan gum, agar, κ-carageenan and algin,” *Cereal Foods World* **34** (1989) 991. Links
283. G. Owen: “Gellan-gum quick setting gelling systems,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, vol. **5**, IRL Press, Oxford 1989, pp. 345 – 349.
284. R. C. Clark: “Flavour and texture factors in model gel systems,” in A. Turner (ed.): *Food Technology International Europe*, Sterling Publications International, London 1990, pp. 272 – 277.
285. US 6027 930115, 1994 (W. F. Chalupa, G. R. Sanderson).
286. A. Nissinovitch, N. Peleg, E. Mey-Tal: “Continuous monitoring of changes in shrinking gels,” *Lebensm. Wissen. Tech.* **28** (1995) no. 3, 347 – 349.
287. C. K Darty: “Applications of gellan gum as a fining agent in alcoholic beverages,” Res. Disclosure, No. 348, 256 (1993).
288. W. Gibson et al.: “Production of low calorie (low joule) or sugar-free jelly sweets using polydextrose and gellan gum, where final total solids exceeds 80 %,” Res. Disclosure, No. 361, 276–7 (1994).
289. “Gellan Gum: a stabiliser of many means,” *Prepared Foods* **160** (1995) no. 6, 125.

290. J. Giese: “Developments in beverage additives,” *Food Technol.* **49** (1995) no. 5, 63 – 65, 68 – 70, 72.
291. EP-A 0 473 854 A1, 1992 (I. T. Norton).
292. V. L. Bell et al.: “Use of gellan gum to prepare gels that can be dried and subsequently rehydrated to their original shape,” Res. Disclosure, No 345, 42 (1993).
293. Toxicological evaluation of certain food additives, WHO Food Additives Series 28, World Health Organisation, Geneva 1991.
294. Reports of the Scientific Committee for Food 26th Series, Commission of the European Committees, Report EUR 13913, Luxemburg 1992.
295. Official Journal of the European Committee, L61, April 18, 1995.
296. Compendium of Food Additive Specification, FAO Food and Nutrition Paper 52. Add 5., FAO, Rome 1997.
297. *Food Chemicals Codex*, 4th ed., National Academy Press, Washington D.C. 1996.
298. J. E. Fox: “Seed Gums,” in A. Imeson (ed.) *Thickening and Gelling Agents for Food*, Blackie Academic and Professional, London 1992, pp. 153 – 170.
299. I. C. M. Dea: “Structure/Function Relationships of Galactomannans and Food Grade Cellulosics,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 5th ed., IRL Press, Oxford 1990, pp. 373 – 382.
300. A. H. Clark, I. C. M. Dea, B. V. McCleary: “The Effect of the Galactomannan Fine Structure on Their Interaction Properties,” in G. O. Phillips, D. J. Wedlock, P. A. Williams (eds.): *Gums and Stabilisers for the Food Industry*, 3rd ed., Elsevier Applied Science Publishers, London 1985, pp. 429 – 440.
301. I. C. M. Dea, A. H. Clark, B. V. McCleary, *Carbohyd. Res.* **142** (1986) 275 – 294.
302. P. M. Dey, *Adv. Carbohydr. Chem. Biochem.* **35** (1978) 341 – 376. Links
303. B. V. McCleary, *Int. J. Biol. Macromol.* **8** (1986) 349 – 354. Links
304. T. J. Painter, in G Aspinall (ed.): *The Polysaccharides*, Vol. 2, Academic Press, London, 1983, pp. 195–285.
305. M. Lahaye, *J. Appl. Phycol.* **13** (2001) 173–184. Links
306. G. A. DeRuiter, B. Rudolph, *Trends Food Sci. Technol.* **8** (1997) 389–395.
307. S. H. Knutsen, D. E. Myslabodski, B. Larsen, A. I. Usov: “A modified system of nomenclature for red algal galactans”, *Bot. Mar.* **37** (1994) 163–169. Links
308. D. A. Rees, *Adv. Carb. Biochem.* **24** (1969) 267–332.
309. E. Zablackis, G. A. Santos, *Botanica Marina* **29** (1986) 319–322. Links
310. D. N. Smidt, W. H. Cook, *Arch. Biochem. Biophys.* **45** (1953) 232–233. Links
311. C. Bellion *et al.*, *Carbohydr. Res.* **119** (1983) 31–48. Links
312. K. F. Wong, J. S. Craigie, *Plant. Physiol.* **61** (1978) 663–666. Links
313. E. L. McCandless, J. S. Craigie, J. A. Walter, *Planta (Berl.)* **112** (1973) 201–212.
314. R. Falshaw, R. H. Furneaux, *Carbohydr. Res.* **276** (1995) 155–165. Links
315. R. Falshaw, H. J. Bixler, K. Johndro, *Food Hydrocoll.* **15** (2001) 441–452.
316. F. Vande Velde, H. A. Peppelman, H. S. Rollema, R. H. Tromp, *Carbohydr. Res.* **331** (2001) 271–283. Links
317. H. J. Bixler, K. Johndro, R. Falshaw, *Food Hydrocoll.* **15** (2001) 619–630.
318. A. Amimi *et al.*, *Cabohydr. Res.* **333** (2001) 271–279.
319. A. I. Usov, V. V. Barbakadze, *Bioorg. Khim.* **4** (1978) 1107–1115.
320. M. Ciancia, M. C. Matulewicz, A. S. Cerezo, *Phytochemistry* **34** (1993) 1541–1543. Links
321. R. Takano *et al.*, *Phytochemistry* **37** (1994) 1615–1619. Links
322. G. M. King, G. E. Lauterbach, *Botanica Marina* **30** (1987) 33–39.
323. C. K. Tseng, *Science* **101** (1945) 597–602. Links
324. H. J. Bixler, K. Johndro, R. Falshaw: “Kappa-2 carrageenan: structure and performance of commercial extracts II. Performance in two simulated dairy applications”, *Food Hydrocolloids* **15** (2001) 619. Links
325. T. Barbeyron, D. Flament, G. Michel, P. Potin, B. Kloareg, *Cahiers De Biologie Marine* **42** (2001) 169.
326. H. J. Bixler, *Hydrobiologia* **327** (1996) 35.
327. R. A. Hoffmann, M. J. Gidley, D. Cooke, W. J. Frith, *Food Hydrocoll.* **9** (1995) 281–289.
328. A. P. Imeson, in G. O. Phillips, P. A. Williams (eds.): *Handbook of Hydrocolloids*, Woodhead Publishing Ltd., Cambridge, 2000, pp. 87–102.
329. W. Yaphe, G. P. Arsenault, *Anal. Biochem.* **13** (1965) 143–148. Links
330. C. N. Jol *et al.*, *Anal. Biochem.* **268** (1999) 213–222. Links
331. A. I. Usov, N. G. Klochkova, *Botanica Marina* **35** (1992) 371–378. Links
332. R. Falshaw, R. H. Furneaux, *Carbohydr. Res.* **269** (1995) 183–189. Links
333. D. J. Stancioff, N. F. Stanley, *Proc. Int. Seaweed Symp.* **6** (1969) 595–609.
334. T. Chopin, E. Whalen, *Carbohydr. Res.* **246** (1993) 51–59. Links
335. A. I. Usov, *Botanica Marina* **27** (1984) 189–202. Links
336. M. Ciancia, M. C. Matulewicz, P. Finc, A. S. Cerezo, *Carbohydr. Res.* **238** (1993) 241–248. Links
337. D. Welti, *J. Chem. Res. (M)* (1977) 3566–3587.
338. S. H. Knutsen, D. E. Myslabodski, H. Grasdalen, *Carbohydr. Res.* **206** (1990) 367–372. Links
339. T. Turquois, S. Acquistapace, F. A. Vera, D. H. Welti, *Carbohydr. Polym.* **31** (1996) 269–278.
340. S. H. Knutsen, H. Grasdalen, *Carbohydr. Polym.* **19** (1992) 199–210.
341. T. Hjerde, O. Smidsrod, B. E. Christensen, *Biopolymers* **49** (1999) 71–80. Links
342. C. Viebke, P. A. Williams, *Food Hydrocoll.* **14** (2000) 265–270.
343. M. A. Roberts, B. Quemener, *Trends Food Sci. Technol.* **10** (1999) 169–181.
344. E. Bjerre-Petersen, J. Christensen, P. Hemmingsen, in R. L. Whistler and B. J. N. (eds.): *Industrial gums. Polysaccharides and their derivatives*, Academic Press, New York, 1973, pp. 123–136.
345. L. Piculell, in A. M. Stephen (ed.): *Food polysaccharides and their applications*, Marcel Dekker, Inc., New York, 1995, pp. 205–244.
346. T. Hjerde, O. Smidsrod, B. T. Stokke, B. E. Christensen, *Macromolecules* **31** (1998) 1842–1851. Links
347. C. Viebke, J. Borgstrom, L. Piculell, *Carbohydr. Polym.* **27** (1995) 145–154.
348. N. F. Stanley, in P. Harris (ed.): *Food gels*, Elsevier Applied Science, London, 1990, pp. 79–119.
349. K. Bongaerts *et al. Macromolecules* **33** (2000) 8709–8719. Links
350. L. Piculell, C. Hakansson, S. Nilsson, *Int. J. Biol. Macromol.* **9** (1987) 297–301. Links
351. K. B. Guiseley, N. F. Stanley, P. A. Whitehouse in R. L. Davidson (ed.): *Handbook of water soluble gums and resins*, McGraw-Hill, New York 1980.
352. FMC Biopolymer: “Not Just Products. Partners. (2000), p. 10 (www.fmcbiopolymer.com).

353. G. H. Therkelsen, in R. L. Whistler, J. N. BeMiller (eds.): *Industrial gums: Polysaccharides and their derivatives*, Academic Press, San Diego, 1993, pp. 145–180.
354. M. DiRosa, *J. Pharm. Pharmacol.* **24** (1972) 89–102. Links
355. L. Aschheim, S. Raffel, *Res. J. Reticuloendothel. Soc.* (1972) 253–262.
356. P. Cantanzaro, H. Schwartz, R. Graham, *Am. J. Pathol.* **64** (1971) 387–404. Links
357. D. J. Schaeffer, V. S. Krylov, *Ecotoxicol. Environ. Saf.* **45** (2000) 208–227. Links
358. C. Haslin, M. Lahaye, M. Pellegrini, J. C. Chermann, *Planta Medica* **67** (2001) 301–305. Links
359. M. J. Carlucci *et al. Antiviral Res.* **43** (1999) 93–102. Links
360. P. Potin, K. Bouarab, F. Kupper, B. Kloareg, *Curr. Opin. Microbiol.* **2** (1999) 276–283. Links
361. F. Vande Velde, G. A. DeRuiter, in S. De Baets, E. J. VanDamme, S. Steinbüchel (eds.): *Biopolymers*, vol. **6** Polysaccharides II, Wiley-VCH, Weinheim, Germany, 2002.
362. L. G. Ekstrøm, *Carbohydr. Res.* **135** (1985) 283–289.
363. T. T. Stevenson, R. H. Furneaux, *Carbohydr. Res.* **210** (1991) 277–298. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Poly(Vinyl Esters)

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of Order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

HELMUT RINNO, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Raw Materials 1
3. Polymer Structure 3
4. Production 4
4.1. Bulk Polymerization 4
4.2. Solution Polymerization 4
4.3. Suspension (Bead) Polymerization . . 5
4.4. Emulsion Polymerization 5
5. Industrially Important Poly(Vinyl Esters) 7
6. Properties 8
7. Quality Specifications, Trade Names, Storage and Transport, Wastewater . 9
8. Uses, Toxicology, and Economic Aspects 10
9. References 10

1. Introduction

Poly(vinyl esters) are polymers with structural units of the general formula:

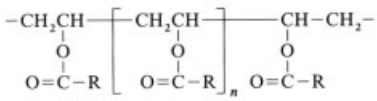

They are synthesized by polymerization of vinyl ester monomers $CH_2{=}CH{-}O{-}CO{-}R$.

In addition to vinyl ester homopolymers, co- and terpolymers of mixed vinyl esters or of vinyl esters and other comonomers are also important industrially.

Poly(vinyl acetate) and vinyl acetate copolymers are the most important poly(vinyl esters).

In 1912 and 1913 KLATTE synthesized the first poly(vinyl esters): poly(vinyl chloroacetate) ($R = CH_2Cl$) and poly(vinyl acetate) ($R = CH_3$) [1].

2. Raw Materials

Vinyl Esters. Vinyl acetate is predominantly produced by acetoxylation of ethylene with acetic acid on palladium catalysts. Vinyl esters of $\geq C_3$ acids are produced by addition of acetylene to the acid, or by palladium- or mercury-catalyzed transvinylation of vinyl acetate (see → Vinyl Esters).

On an industrial scale vinyl acetate ($R = CH_3$), vinyl propionate ($R = CH_2CH_3$), vinyl pivalate [$R = C(CH_3)_3$], vinyl 2-ethylhexanoate [$R = CH(C_2H_5)-(CH_2)_3CH_3$], vinyl laurate [$R = (CH_2)_{10}CH_3$], VeoVa 10 [$R = C(CH_3)R'_2$] and VeoVa 9 [$R = C(CH_3)R'_2$] are used for the production of polymers. VeoVa 9 and VeoVa 10 are the vinyl esters of versatic acids – synthetic tertiary carboxylic acids with nine or ten carbon atoms (see → Carboxylic Acids, Aliphatic, Chap. 10.). Previously, vinyl chloroacetate, vinyl isobutyrate, vinyl caproate, vinyl isononanate, vinyl stearate, VeoVa 911, and vinyl benzoate were also used in industry for a time or were envisaged for this purpose.

Data for the industrially most important monomers are summarized in Table 1. High-purity monomers are required for the reproducible production of polymers. The content of other components can be determined by gas chromatography. The purity determination of vinyl acetate is standardized in ASTM standards [2]. Impurities which are particularly disruptive are those which severely inhibit polymerization, such as crotonaldehyde and vinylacetylene, or compounds that act as chain-transfer agents, such as acetic acid, acetaldehyde, acetone, benzene, or toluene [3], [4]. Impurities with two readily copolymerizable double bonds,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_001

Table 1. Physical data for industrial vinyl esters

	Vinyl acetate	Vinyl propionate	Vinyl pivalate	Vinyl 2-ethyl-hexanoate	VeoVa 9	VeoVa 10	Vinyl laurate
CAS no.	[*108-05-4*]	[*105-38-4*]	[*3377-92-2*]	[*94-04-2*]	[*54423-67-5*]	[*51000-52-3*]	[*2146-71-6*]
Vinyl ester content, %	≥ 99.9 [a]	≥ 99.9 [b]	99.4 [c]	99.5 [c]	99.4	> 98.0	99.4 ± 0.1
Water (DIN 51 777), %	≤ 0.03 [a]	≤ 0.1 [b]	< 0.10	< 0.10	< 0.1 [d]	≤ 0.1 [d]	< 0.005
Acid content, %	≤ 0.0005	≤ 0.03 [b]	< 0.10	< 0.10	< 0.3	≤ 0.2	< 0.5 [e]
Color number (ISO 6271)	≤ 5	≤ 10 [b]	20	< 20	< 20	< 15	≤ 10
Inhibitor content, ppm	3 – 20	10 [b]	9 – 13	6 – 10	5 ± 2 [d]	5 ± 2 [d]	without [e]
bp, °C	72 – 73 [a]	ca. 95	112.1	185.6	185 – 200	136 (13.3 kPa)	254
Vapor pressure, kPa							
at 20 °C	12	4.58	2.412	0.057	< 0.1	< 0.1	
at 50 °C	42.6	19.63					0.01
at 70 °C					1.2	0.7	0.38
Flash point (DIN 51 755), °C	− 8	4.5 [i]	14 [i]	65	45	75 [h]	136
Upper/lower explosion limit (vapor in air at 20 °C), vol %	2.6/13.4	1.8 [j] – 13.5 [k]	n.d.	n.d.	0.7/n.d.	n.d.	0.5/n.d.
Solubility in water at 25 °C, %	0.9 [f]	0.6	0.08	< 0.01	< 0.1	< 0.1	< 0.1
Solubility of water in vinyl ester at 25 °C, %	2.3 [f]	0.8	n.d.	n.d.	< 0.06	< 0.05	< 0.1
Heat of polymerisation, kJ/kg	1035.8	882			ca. 522	ca. 485	n.d.
T_g of homopolymer [g], °C	33	− 7 (33)	+ 70 (33)	− 36 (33)	60 (28)	− 3 (28)	− 53 [l]

[a] Hoechst specification. [b] BASF specification. [c] Union Carbide Specification. [d] Shell specification. [e] Values from Wacker Chemie. [f] At 20 °C. [g] Values depend on measurement method; values in parentheses give T_g of poly(vinyl acetate) when same method is used. [h] DIN 51 758 (PM). [i] Tag closed cup (ASTM D56 – 61). [j] At 4 °C. [k] At 41.5 °C. [l] Extrapolated from copolymer data.
n.d. = not determined.

such as vinyl crotonate, also cause problems because they can lead to cross-linking reactions and thus to the formation of insoluble polymers. Monomers used for emulsion polymerization must be completely free from previously formed polymer.

To prevent polymerization during storage and transport, 3 – 20 ppm of a phenolic compound such as hydroquinone or hydroquinone monomethyl ether is used as an inhibitor. The inhibitors generally need not to be removed before polymerization, since they can be rendered inactive during polymerization by using a sufficiently large amount of initiator.

To assess the purity and the suitability for polymer production, particularly in the case of vinyl acetate, a simple polymerization test [5] is used. A given quantity of monomer is warmed with 0.5 % benzoyl peroxide (relative to monomer) to 70 °C under defined conditions, and the time elapsing until the start of polymerization is recorded. The delay in comparison with pure, inhibitor-free monomer is a measure of the impurity content or of the amount of inhibitor added for stabilization.

Initiators. In industrial poly(vinyl ester) processes polymerization is initiated by radical

initiators. Other processes, such as photopolymerization, radiation-induced polymerization, and the use of organometallic compounds as initiators have not achieved importance.

Azo compounds (e.g., azobisisobutyronitrile), diacyl peroxides (e.g., dibenzoyl peroxide and dilauryl peroxide), peroxyesters (e.g., *tert*-butyl peroxybenzoate and *tert*-butyl peroxyoctanoate), hydroperoxides (e.g., *tert*-butyl hydroperoxide and cumene hydroperoxide), and water-soluble inorganic peroxides (e.g., hydrogen peroxide, potassium, ammonium, or sodium peroxodisulfate) are used as radical initiators. Redox systems are also widely used, in which a peroxy compound is combined with a reducing agent (e.g., H_2O_2–ascorbic acid; H_2O_2–Fe^{2+}, ammonium peroxodisulfate – sodium sulfite, or *tert*-butyl hydroperoxide – sodium formaldehyde sulfoxalate) [6].

Initiators are chosen according to their solubility in the solvent or monomer, and the decomposition half-life of the substance or the redox combination under polymerization conditions.

Sufficient initiator must be used to oxidize any inhibitors present in the monomer, and to ensure a constant reaction rate in spite of continuous decomposition of the peroxide under the reaction conditions. High radical concentrations result in polymers of low molecular mass. Generally 0.1 – 1 % of an initiator based on the monomer is used. The initiator can be added at the start of the polymerization in one portion or added gradually during the course of the reaction.

Molecular Mass Regulators. To regulate the molecular mass of the polymer, C_1–C_4 alcohols, C_2–C_4 aldehydes, thiols ($>C_{10}$), or chlorinated hydrocarbons (e.g., carbon tetrachloride) are used; for chain-transfer constants, see [7].

3. Polymer Structure

The kinetics of homo- and copolymerization of vinyl esters has been widely studied; for a summary, see [3]. Three features are particularly characteristic of the polymerization of vinyl acetate:

1) Very high reactivity of the vinyl acetate radical
2) Relatively low reactivity of vinyl acetate monomer towards attack by radicals
3) High chain-transfer constants of vinyl acetate, poly(vinyl acetate), and other polymers relative to the vinyl acetate radical or the growing chain

The differing reactivity of the radical and the monomer, which is also the case for other vinyl esters, makes the production of uniform copolymers from vinyl esters with comonomers that behave differently (e.g., where the monomer reacts more rapidly with radicals) difficult. For copolymerization parameters *r*, *q*, and *e*, see [7]. Chain-transfer reactions involving the monomer or the polymer results in products with a high degree of branching being formed in the polymerization of vinyl acetate. Three types of branching are observed in the final polymer:

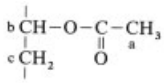

The side chains can be bound to the main chain via an ester group (a) and also directly by a C–C bond (b and c). In the case of chain-transfer in (a), the side chains can be cleaved hydrolytically. The extent of side chain formation at positions (a), (b), and (c) depends strongly on the reaction conditions. The data vary between 97 % and 75 % for (a) [8–10]. The lowering of the average degree of polymerization occurring on saponification of polyvinylacetate, the reduction in the number of side chains in the polymer, and the broadening of the molecular mass distribution, which depends on the proportion of (a) branches, must be taken into consideration in the production of downstream products of poly(vinyl acetate) (see → Polyvinyl Compounds, Others).

In poly(vinyl acetate) the monomer units are linked predominantly head-to-tail. The proportion of head-to-head segments decreases with decreasing polymerization temperature [3].

4. Production

Homo- and copolymers of vinyl esters are produced both in the homogeneous phase by bulk or solution polymerization, and in the heterogeneous phase by suspension and emulsion polymerization.

4.1. Bulk Polymerization

Bulk polymerization can be carried out discontinuously by monomer addition in stirred reactors or continuously in tubular reactors. The monomers are heated to the reaction temperature in the presence of monomer-soluble initiators. The heat of reaction can be removed for low-boiling monomers by a reflux condenser, and in all other cases through the reactor wall. The process is limited to polymers with low to medium molecular mass. Molecular mass and melt viscosity must be sufficiently low at the end of the reaction that the product can be removed from the reactor without difficulty.

Discontinuous processes for poly(vinyl acetate) with a molecular mass of $M_w = 35\,000$ have been described. The molecular mass can be controlled by addition of regulators [11]. A similar process is used for the production of graft copolymers from poly(ethylene glycol) and vinyl acetate [12].

Poly(vinyl acetate) with $M_w = 300\,000$ is obtained in tubular reactors, which give almost quantitative polymerization [11].

Terpolymers can be produced from VeoVa 10, styrene, and maleic or acrylic esters by substance polymerization at 155 – 170 °C [13]. Vinyl acetate – ethylene copolymers with an ethylene content of $> 60\,\%$ are produced by using the high-pressure polyethylene process (see → Polyolefins, Chap. 1.5.1.3.).

4.2. Solution Polymerization

Solution polymerization allows the production of polymers with a more uniform structure and lower degree of branching in a simpler process. The removal of the heat of reaction is less problematic than in bulk polymerization, since the reaction can be carried out under reflux at the boiling point of the solvent or one of the monomers. The degree of polymerization depends on the nature and amount of solvent [14].

A process which allows the production of relatively uniform, high molecular mass poly(vinyl acetate) is solution polymerization with relatively low conversion. The unreacted vinyl acetate acts as an additional solvent. The reaction is stopped at a conversion of, for example, 50 % by adding an inhibitor, and unreacted monomer is removed.

Poly(vinyl acetate) which is to be further processed to poly(vinyl alcohol) [7], is generally produced in methanolic solution. Use of *tert*-butanol permits the production of high molecular mass solution polymers because of its low chain-transfer constants relative to the vinyl acetate radical. This is exploited industrially in the production of vinyl acetate – ethylene copolymers with an ethylene content of 50 – 70 % (see → Rubber, Synthetic). Solution polymerization can be carried out as a discontinuous process in simple stirred reactors, which operate under pressure (for vinyl acetate – ethylene) or pressureless, depending on the monomer used and the reaction conditions; or continuously in a tubular reactor or a cascade of stirred reactors, which are particularly suitable for this process.

4.3. Suspension (Bead) Polymerization

In suspension polymerization the monomer, usually vinyl acetate, is suspended in water by stirring together with a monomer-soluble initiator, and heated to the boiling point of the vinyl acetate – water azeotrope for polymerization. The polymer is produced as beads (diameter ca. 0.2 – 3 mm) and can easily be separated after cooling, and dried. Suspension polymerization combines the advantages of bulk and solution polymerizations: heat removal during polymerization is straightforward, and the product can be easily isolated and handled, and is free of solvent. Suspending agents are used to stabilize the two-phase system of water and the monomer or the solution of polymer formed in the monomer during the reaction. In the older literature a large number of active substances are described, including neutral electrolytes such as phosphates, water-soluble polymers such as modified starch, poly(vinyl alcohol), polyacrylamide, polyvinylpyrrolidone, copolymers with

carboxyl groups such as styrene – maleic acid copolymers [11], or crotonic acid – vinyl acetate copolymers.

Apart from the nature of the suspending agent, the shape of the reactor, the stirrer speed, and the stirrer geometry significantly affect the stability of the system and the particle size distribution of the beads formed.

4.4. Emulsion Polymerization

Emulsion polymerization is the most important process for the production of vinyl ester homo- and copolymers. The emulsions, which contain water as the continuous and the polymers as the dispersed phase, are mainly used in the liquid form. Smaller amounts are converted to the polymer by spray-drying to give a powder which can be redispersed in water.

The properties of the product are determined not only by polymer composition and structure, but also by the stabilizing additives used in production, and by the particle size and particle size distribution of the dispersed polymer phase.

In the production of poly(vinyl ester) emulsions three types of stabilizer are used:

1) Water-soluble, polymeric protective colloids, such as poly(vinyl alcohol) [15], [16], [17], hydroxyethyl cellulose, polyvinylpyrrolidone [18], poly(*N*-vinylamides) [19], or starch.
2) Anionic emulsifiers that contain hydrophilic groups such as $-SO_3^-$, $-OSO_3^-$, $-(OCH_2-CH_2)_n-SO_3^-$, or $-OPO_3^{2-}$, in addition to a hydrophobic group. The hydrophobic part of the molecule can consist of $C_{12}-C_{24}$ alkyl or alkylaryl groups,

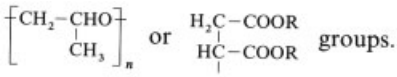

Typical compounds of this type are dodecylbenzenesulfonate, sodium laurylsulfate, and dioctyl sulfosuccinate. Ethylene sulfonate is an anionic stabilizer [20] that leads to formation of high molecular mass polymers with $-SO_3Na$ groups by incorporation into the polymer.

3) Nonionic emulsifiers with hydrophobic groups and polyether chains as hydrophilic groups. Oxyethylation products of nonylphenol, isooctylphenol, $C_{12}-C_{18}$ fatty alcohols, or low molecular mass poly(propylene oxide) are most widely used. The degree of oxyethylation is 4 – 50.

Protective colloids usually give emulsions with an average particle size of ca. 1 – 10 µm. Ionic emulsifiers give an average particle diameter of 0.1 – 0.3 µm. With a combination of the two types the average particle diameter is ca. 0.2 –2 µm.

The choice of emulsifiers or protective colloids depends on the desired properties of the emulsions. There is a range of known relationships between the emulsifier/protective colloid system, the amount of emulsifier or protective colloid, the polymerization reaction conditions, and the polymer properties [7], [21], [22]. However, they have not yet led to a general theory for the choice of emulsifiers and protective colloids. Therefore, for the production of new poly(vinyl ester) emulsions the most favorable emulsifiers and protective colloids and the optimal concentration must be determined experimentally.

The protective colloid most thoroughly investigated and most often used is poly(vinyl alcohol). Its emulsifying action is strongly dependent on the content of residual acetyl groups. It decreases with decreasing acetyl group content, as does the reaction rate of the emulsion polymerization [23]. Poly(vinyl alcohols) with a residual acetyl content of 0 – 12 % are generally used for the emulsion polymerization of vinyl esters. During polymerization, part of the poly(vinyl acetate) is grafted on to the poly(vinyl alcohol). In peroxodisulfate-initiated polymerization, the content of graft polymer increases with increasing molecular mass and increasing acetyl group content of the poly(vinyl alcohol) [24]. It is also strongly dependent on the initiator. With H_2O_2 and azobisisobutyronitrile as initiators significantly less graft polymer is formed.

Kinetic investigations of the emulsion polymerization of vinyl esters are mainly concerned with vinyl acetate homo- and copolymerization in the presence of ionic emulsifiers [3], [25]. Because of the considerable water solubility of the monomer, in the case of vinyl acetate a mechanism of particle formation differing from that in

the emulsion polymerization of other monomers has been discussed. It is assumed that polymerization initially takes place in aqueous solution since the rate is almost independent of the emulsifier concentration [7], [25–27]. In contrast, experiments on emulsion polymerization of water-insoluble vinyl esters of long- chain carboxylic acids have shown, that here the behavior is similar to that of other hydrophobic monomers such as styrene.

For emulsion polymerization with protective colloids ideas about the stabilizing effect are mainly qualitative. It is assumed that high molecular mass hydrophilic protective colloids are adsorbed as loops on the surface of the particles and stabilize them by forming a hydrate layer.

Discontinuous Polymerization. On an industrial scale, discontinuous emulsion processes predominate. In batch processes the total quantity of monomer is heated to the polymerization temperature with the aqueous solution of the protective colloid or emulsifier and the initiator required to form the final emulsion. In the addition process the monomer or, in the production of copolymers, a monomer mixture is added to the total aqueous phase during the course of polymerization. A special case of the addition process is the preemulsion technique, in which part of the aqueous phase is charged to the reactor and a preemulsion is formed from the other part and the monomer. This preemulsion is then added during the course of polymerization.

After polymerization is complete, the emulsion formed is cooled, treated if necessary with stabilizing additives, plasticizers, or by pH adjustment, and then removed from the reactor. The treatment can also be carried out in additional stirred vessels connected to the reactor.

In the discontinuous processes stirred vessels are used. The stirrer geometry and speed have a major influence on the quality of the emulsion. Anchor stirrers, propellers, impellers, and MIG stirrers are used. In the case of vinyl acetate homo- and copolymers produced at atmospheric pressure, the heat of reaction can be removed by a reflux condenser. For other vinyl esters, and for polymerization of vinyl acetate below its boiling point, jacket cooling or the use of internal or external heat exchangers is necessary.

If low-boiling or gaseous monomers such as ethylene or vinyl chloride are copolymerized with vinyl esters, stirred autoclaves must be used. The amount of ethylene which is copolymerized mainly depends on the ethylene pressure above the reaction medium but also on other process parameters such as reaction temperature and monomer concentration during polymerization [28], [29]. Addition processes [30] and batch processes [31] at ethylene pressures up to 100 bar have been described.

Continuous Processes. For the continuous production of poly(vinyl ester) emulsions, both atmospheric-pressure, and pressurized-vessel processes (for copolymerization with ethylene [32] or vinyl chloride) have been described. The monomers, initiators, emulsifiers and protective colloids, and water are fed, either together as a preemulsion or separately as aqueous and organic phases, to a tubular reactor [33], [34], a stirred-tank cascade, a continuous stirred-tank reactor, or a loop reactor with a circulating pump [35].

Construction Materials. Vinyl esters, poly (vinyl esters), and the additives and initiators used in their production are not very corrosive. To produce poly(vinyl esters) reactors made of stainless steel or glass-lined steel are used. For parts of the plant which only come into contact with the monomers, such as pipes and heat exchangers, aluminum can also be used. Copper and copper alloys should be avoided, since copper ions can inhibit polymerization even at very low concentrations [3].

5. Industrially Important Poly(Vinyl Esters)

The only industrially important homopolymers are poly(vinyl acetate) and, to a much lesser extent, poly(vinyl propionate). All vinyl esters which are accessible on an industrial scale are also used to prepare a large number of co- and terpolymers by combination with one another and with other monomers. In this group vinyl acetate is also by far the most important vinyl ester.

In the following the most important processes for the production of the individual polymers are denoted by letters as follows:

(B) Bulk polymerization
(S) Solution polymerization
(BS) Bead suspension polymerization
(G) Graft polymerization
(E) Emulsion polymerization

Polymers are produced from the following monomers by the methods shown:

Vinyl acetate (B, S, BS, E)
Vinyl acetate – dibutyl maleate (B, E) [36]
Vinyl acetate – *n*-butyl acrylate (E)
Vinyl acetate – 2-ethylhexyl acrylate (E)
Vinyl acetate – *n*-butyl acrylate – *N*-hydroxymethylacrylamide (E) [37]
Vinyl acetate – crotonic acid (S, E) [38]
Vinyl acetate – VeoVa 10 (B, E) [13]
Vinyl acetate – VeoVa 10 – acrylic acid (E) [13]
Vinyl acetate – VeoVa 10 – *n*-butyl acrylate (E)
Vinyl acetate – *N*-hydroxymethylacrylamide (E)
Vinyl acetate – vinyl laurate (B, E) [39]
Vinyl acetate – vinyl laurate – vinyl chloride (E) [40]
Vinyl acetate – ethylene (B, S, E)
Vinyl acetate – ethylene – vinyl chloride (E) [41]
Vinyl acetate – ethylene – acrylic ester (E)
Vinyl acetate – ethylene – acrylamide (E) [42]
Vinyl acetate – ethylene – *N*-hydroxymethylacrylamide (E) [43]
Vinyl acetate – crotonic acid grafted on polyglycol (G) [12]
Vinyl acetate – *N*-vinylpyrrolidone (S)
Vinyl propionate (E)
Vinyl propionate – vinyl chloride (E)
Vinyl propionate – *tert*-butyl acrylate (E)
VeoVa 10 – vinyl chloride (E) [22]
VeoVa 10 – styrene – acrylic ester and/or maleate (B) [13]
VeoVa 10 – VeoVa 9 – methyl methacrylate – butyl acrylate (E)

VeoVa 10 can be completely or partly replaced by VeoVa 9 to achieve special properties.

6. Properties

Poly(Vinyl Acetate) Homopolymers. Poly (vinyl acetate) is amorphous, odorless, tasteless, and has high lightfastness, and weather resistance. The glass transition temperature (T_g = 28 °C) depends to a certain extent on the molecular mass [7]. For electrical, mechanical, and thermal properties, see [7]. The polymers are soluble in many esters, ketones, cyclic ethers, phenols, halogenated aliphatic hydrocarbons, methanol, 95 % ethanol, and 90 % isopropanol.

Commercial poly(vinyl acetate) emulsions with poly(vinyl alcohol) as the protective colloid usually have a particle size of ca. 1 – 10 µm, a solids content of 40 – 65 %, and a viscosity range of 1 – 50 Pa · s. The latex particles are frequently aggregated to produce grape-like formations [44].

Emulsifier-containing homopolymer emulsions, some of which contain protective colloids such as hydroxyethyl cellulose, have lower particle size (0.1 – 2 µm).

The minimum film formation temperature of homopolymer emulsions, which is important for their applications, is 15 – 18 °C. Therefore, for many uses plasticizers, such as dibutyl phthalate, tricresyl phosphate, or auxiliary film formers, are added.

Coarse-particle emulsions stabilized with poly(vinyl alcohol) are mainly important as wood glues and adhesives. Fine- and medium-particle emulsions stabilized by emulsifiers are mainly used as binders for emulsion paints and textiles.

Vinyl Acetate Copolymers. The facile alkaline saponification of the ester groups in vinyl acetate homopolymers is a disadvantage in many applications, since the saponification products are hydrophilic and water sensitive. This low resistance to hydrolysis is one reason for combination with comonomers. The second is the glass transition temperature of 28 °C, which is too high for many uses (e.g., paints and adhesives) and renders the poly(vinyl acetate) too hard and inflexible at the usual temperatures of use. The glass transition temperature can be adjusted by the addition of plasticizers. However, for use where plasticizers are undesirable and increased resistance to saponi- fication is required, soften-

ing comonomers are used, e.g., dibutyl maleate; *n*-butyl acrylate; 2-ethylhexyl acrylate; ethylene; or ethylene in combination with vinyl chloride, VeoVa 10, or vinyl laurate. The glass transition temperature is most effectively lowered by ethylene, vinyl laurate, and acrylic esters, while the resistance to alkali is best improved by ethylene, or ethylene combined with vinyl chloride or by VeoVa 10. Copolymerization with > 20 wt % VeoVa 9 increases the glass transition temperature and the resistance to alkali.

Since for a particular application of a copolymer, a narrow range for the glass transition temperature or for the minimum film formation temperature is required, the ratio of vinyl acetate to the comonomers can only be varied within certain limits. The additional use of VeoVa 9 in the pressureless production of emulsions or the additional use of vinyl chloride for high-pressure polymers allows the production of polymers with a desired glass transition temperature but different vinyl acetate contents. For the evaluation of a copolymer it is critical that, at a monomer ratio required by the glass transition temperature, the other properties are favorable. For example, emulsions used as binders in paints must be stable towards alkalies. LINDEMANN [28] has determined the compositions necessary to achieve glass transition temperatures of + 10 °C, 0 °C, and− 10 °C for the most important comonomers used in the production of vinyl acetate copolymers. Saponification experiments on vinyl acetate copolymers have shown that vinyl acetate – ethylene, vinyl acetate – ethylene – vinyl chloride, vinyl acetate – VeoVa 10, and vinyl acetate – vinyl laurate – vinyl chloride, co- and terpolymers with 50 – 80 % vinyl acetate have particularly favorable properties for paints [41], [45–47]. Use of *N*-hydroxymethylacrylamide allows the production of cross-linkable vinyl acetate copolymers, and the use of crotonic acid gives water-soluble vinyl acetate copolymers with good film-forming properties.

Vinyl Proprionate Homopolymers. Poly(vinyl propionate) differs from poly(vinyl acetate) in its lower glass transition temperature (T_g − 7 °C) and its higher stability towards alkalies [47]. Under the same conditions vinyl propionate polymerizes more slowly than vinyl acetate and gives products with lower molecular mass [48]. Poly(vinyl propionate) is produced industrially by emulsion polymerization.

Vinyl Propionate Copolymers. For many applications the glass transition temperature of poly(vinyl propionate) is too low. To obtain industrially more desirable polymer properties (e.g., T_g, resistance to saponification, weather resistance) it is therefore more commonly used together with comonomers that raise the glass transition temperature. Vinyl propionate – vinyl chloride and vinyl propionate – *tert*-butyl acrylate copolymers are produced by emulsion polymerization. Coatings made from these polymers are relatively resistant to saponification and have good weather resistance.

VeoVa 10 and VeoVa 9 Polymers. Homopolymers of VeoVa 10 or VeoVa 9 have not yet achieved any importance. Copolymers with vinyl acetate (see above) and vinyl acetate and acrylic esters are the most important. A range of other co- and terpolymers have been described [13], [22]. VeoVa 10 – vinyl chloride copolymers in the ratio 2 : 1 to 1 : 2 can be produced by emulsion polymerization in stirred autoclaves. The polymers are very weather resistant and have a high pigment-binding capacity.

Styrene – VeoVa 10 – acrylic ester or maleic ester terpolymers are clear, colorless, saponification resistant, and, on copolymerization with monomers with functional groups, cross-linkable [49].

7. Quality Specifications, Trade Names, Storage and Transport, Wastewater

Quality Specifications. The applications of vinyl ester homo- and copolymers are so varied that uniform, generally valid quality specifications cannot be defined.

The content of unpolymerized monomers, which is generally < 0.5 % can be determined by GC [50], [51]. Molecular mass can be determined by light scattering or gel permeation chromatography.

For emulsions the determination of dry residue (ISO 1625), sieve residue (ISO 4576), pH value (ISO 1148), viscosity (ISO 3219 and ISO 2555), freeze/thaw resistance (ISO 1147),

density (ISO 8962), minimum film formation temperature, and the white point (ISO 2115) is standardized. The size of the latex particles is determined by optical or electron microscopy, by light scattering [52], by the Marshall method [53], or by aerosol spectroscopy [54]. In Germany, for polymers subject to foodstuffs regulations, the chewing gum regulation [55], the cheese regulation [56], and Recommendation XIV [57] apply.

Trade Names and Producers. Airflex (Air Products); Appretan, Mowilith, Mowiton, Imperon-Binder (Hoechst); Emultex (Harlow Chemical Comp.); Everflex (Grace); Propiofan (BASF); Rhodopas (Rhone Poulenc Polymers); Uramul (DSM); Vinamul (Vinyl Products); Ucar (UCC); Vinavil (Montedison); Vinnapas, Vipolit (Wacker).

Storage and Transport. Poly(vinyl esters) are combustible, but not readily flammable. They are generally not corrosive.

For poly(vinyl ester) emulsions, transport and storage containers must be resistant towards the aqueous phase, including low molecular mass substances dissolved therein. Irreversible damage to emulsions can arise through external effects which change the composition or the state of colloidal distribution. For example water loss can result in skin formation on the surface.

Many emulsions are frost resistant. However some products coagulate when the aqueous phase freezes. Producers generally give information on the frost resistance of emulsions.

Undesired aggregation or coagulation of the latex particles can occur on strong shearing in unsuitable pumps or in spray nozzles.

Poly(vinyl ester) emulsions must frequently be protected against attack by microorganisms [58].

Wastewater. With the exception of the less widely used crotonic acid copolymers, the polymers are insoluble in water. They can, however, enter the wastewater system as emulsions. Poly(vinyl esters) are nontoxic but are degraded extremely slowly in water. Emulsions therefore must not be discharged directly into receiving waters. Wastewater that contains emulsion can generally be treated in biological sewage treatment plants. The polymer enters the sewage sludge.

8. Uses, Toxicology, and Economic Aspects

Uses. The most important uses of poly(vinyl esters) are in emulsion paints (→ Paints and Coatings, Chap. 2.4.4.→ Paints and Coatings, Chap. 2.4.5.) and adhesives (→ Adhesives.). Other uses include antinoise compounds, chewing gum bases, concrete additives, soil stabilization, glass fiber sizing material, binders for nonwoven fabrics, fibrous leather substitute, pigment printing pastes, coatings for undersides of carpets, cheese rinds, and coatings for paper.

For some uses (building, adhesives, paints) dispersion powders, which can be redispersed in water before use, are produced from the emul-sions by spray drying [59]. Here vinyl acetate homopolymers, copolymers of vinyl acetate with ethylene, VeoVa 10, or with VeoVa 10 and acrylic esters stabilized with poly(vinyl alcohol) are predominantly used. The main advantage of these products lies in their ability to be stored without any problem, and that they can be combined with hydraulic binding agents as one- component systems [60].

Toxicology. A differentiation must be made between the polymers themselves and the other substances always present in the industrial products. All investigations published to date on poly(vinyl esters) refer to industrial products; they are therefore only valid for the composition used in each case.

There are no indications that the poly(vinyl esters) investigated are toxic by oral administration; poly(vinyl acetate) see [61]. Poly(vinylesters) from C_2–C_{18} unbranched fatty acids are therefore permitted together with the comonomers listed individually in the corresponding regulations [55–57] and the accompanying substances for coating cheese and as additives for the production of chewing gum.

Economic Aspects. For the production of polymers based on vinyl esters there are no reliable statistical figures. The production capacity for vinyl acetate in 1990 (in 10^3 t/a) gives an indication of the order of magnitude [62], [63]:

	1991	1994
Western Europe	590	655
Germany	290	–
Eastern Europe	300	116
Former Soviet Union	115	–
North and South America	1480	1775
United States	1280	1500
Asia	930	953
Japan	570	–

The consumption of vinyl acetate is assumed to be ca. 800 000 t in the United States, ca. 660 000 t in Western Europe, and 550 000 t in Japan in 1990. The distribution of the consumption of vinyl acetate among the different products varies in the individual regions. In Western Europe the following quantities of vinyl acetate are consumed in the most important areas of use:

Poly(vinyl acetate) emulsions	ca. 430 000 t
Poly(vinyl alcohol)	ca. 120 000 t
Poly(vinyl acetal) and other uses	ca. 110 000 t

9. References

1. Dr. Klattes: Pionierpatente, Dokumente aus Hoechster Archiven, vol. 10, Hoechst, Frankfurt 1965.
2. ASTM D 2086, D 2190, D 2191, D 2193.
3. M. K. Lindemann in G. E. Ham (ed.): *Vinyl Polymerization,* **vol. 1,** Dekker, New York 1967, pp. 252 – 255.
4. K. K. Georgieff et al., *J. Appl. Polym. Sci.* **8** (1964) 889 – 896.
5. K. K. Georgieff et al., *J. Appl. Polym. Sci.* **5** (1961) 212 – 217.
6. *Houben-Weyl,* **XIV/1,** 263.
7. M. K. Lindemann in H. F. Mark, N. G. Gaylord (eds.): *Encyclopedia of Polymer Science and Technology,* **vol. 15,** Interscience, New York 1971, pp. 577 –703.
8. W. W. Graessley, R. D. Hartung, W. C. Uy, *J. Polym. Sci. Polym. Phys. Ed.* **7** (1969) 1919 – 1935.
9. D. J. Stein, *Makromol. Chem.* **76** (1964) 170 – 182.
10. P. Mehnert, *Kolloid Z. Z. Polym.* **251** (1973) 587 – 593.
11. J. M. de Bell, W. C. Groggin, W. E. Gloor: *German Plastics Practice,* De Bell & Richardson, Springfield, USA 1946.
12. Hoechst, DE 1 077 430, 1958 (K.-H. Kahrs et al.).
13. Shell, DE-OS 2 422 043, 1974 (A. M. C. Steenis, W. J. Westrenen).
14. Hoechst, DE 1 177 825, 1961 (W. Ehmann, K.-H. Kahrs).
15. I.G. Farbenind., DE 727 955, 1934 (W. Starck, H. Freudenberger).
16. Wacker-Chemie, DE 887 411, 1938 (H. Berg, H. Mader).
17. K. Noro, *Br. Polym. J.* **2** (1970) 128.
18. BASF, DE-AS 2 256 154, 1972 (H. Grubert, W. Druschke, Sliwka).
19. Hoechst, DE 1 206 592, 1962 (W. Bartmann, C. Beermann, W. Ehmann, D. Ulmschneider).
20. I.G. Farbenind., DE 744 318, 1940 (W. Starck).
21. Shawinigan, GB 574 863, 1942 (M. Kiar).
22. Shell Chemie: Die Herstellung von VeoVa-Mischpolymer-Dispersionen. Technische Broschüre RES/VVx/4(G), 3rd ed., Oct. 1976.
23. T. Motoyama, S. Yamamoto, S. Okamura, *Kobunshi Kagaku* **10** (1953) 108 – 116.
24. S. Okamura, T. Yamashita, *Kobunshi Kagaku* **15** (1962) 271.
25. M. Nomura, M. Harada, W. Eguchi, S. Nagata, *ACS Symp Ser.* **24** (1976) 102 – 121.
26. F. Candau, R. H. Ottewill: *An Introduction to Polymer Colloids,* Kluwer Academic Publishes, Dordrecht 1990, p. 1, p. 159 ff.
27. M. S. El Aasser, R. M. Fitch: *Future Directions in Polymer Colloids,* Martinus Nijhoff, Dordrecht 1987.
28. M. K. Lindemann, *Paint Manuf.* **38** (1968) no. 9, 30.
29. G. Löhr, *Plast, Rubber Mater. Appl.* **4** (1979) 4, 141.
30. Hoechst, DE 1 127 085, 1960 (K.-H. Kahrs, A. Staller).
31. Air Reduction, DE-OS 1 595 402, 1966 (M. K. Lindemann, R. P. Volpe).
32. BASF, FR 1 226 382, 1959 (H. Fikentscher, E. G. Kastning, N. Rudolphi).
33. Bayer, DE-OS 2 007 793, 1970 (D. Glabisch, W.-D. Schellenberg, K. Nöthen, H. Bärtl).
34. Hoechst, AU 250 009, 1963 (D. Backmann, A. Kühlkamp, E. Paszthory, H. Schreiber).
35. Shawinigan, DE-OS 1 900 112, 1969 (R. Lanthier).
36. Deutsche Amphibolin-Werke, DE-AS 1 092 656, 1958 (G. Penell).
37. Union Oil Comp. of California, US 3 714 096, 1970 (G. Biale).
38. Henkel, DE 1 469 400, 1963 (M. Dohr, G. Tauber, J. Galinke).
39. Wacker-Chemie, DE-OS 2 229 569, 1972 (E. Bergmeister, A. Stoll).

40. Wacker-Chemie, DE 1 745 555, 1967 (E. Bergmeister, J. Heckmaier, G. P. Kirst, H. West).
41. Vinyl Products, *Pigm. Resin Technol.* **4** (1975) no. 9, 11.
42. Hoechst, US 3 870 673, 1973 (K. J. Rauterkus, J. Blazek).
43. Sumitomo Chem., DE-OS 2 727 205, 1977 (T. Oyamada et al.).
44. Hoechst: *Mowilith,* 4th. ed., 1969, p. 63.
45. G. P. Kirst, *Fette Seifen Anstrichm.* **69** (1967) 919–923.
46. G. Florus, *Congr. FATIPEC* **7** (1964) 149–153.
47. W. Sliwka, *Angew. Makromol. Chem.* **4/5** (1968) 310–350.
48. A. J. Buselli, M. K. Lindemann, C. E. Blades, *J. Polym. Sci.* **28** (1958) 485–498.
49. Shell Chemicals: VeoVa 10 Mass Polymerisation. Prel. Techn. Information RES 76 : 1, 1976.
50. H. Hachenberg: *Perkin Elmer Tips* **41, GC,** April 1970.
51. G. Schmötzer, *Fresenius Z. Anal. Chem.* **260** (1972) 10–24.
52. M. Kerker: *The Scattering of Light,* Academic Press, New York 1969, pp. 311–413.
53. A. Kuhn: *Kolloidchemisches Taschenbuch.* Akad. Verlagsges., Leipzig 1953.
54. G. Löhr, R. Reinecke, *Angew. Makromol. Chem.* **85** (1980) 181.
55. Lebensmittelrecht, Bundesgesetze und -verordnungen über Lebensmittel und Bedarfsgegenstände, C. H. Becksche Verlagsbuchhandlung, München, 111. Kaugummiverordnung lt. ÄndV. Dec. 20, 1977 (BGBl. I, S. 2802, 2806).
56. Lebensmittelrecht, Bundesgesetze und -verordnungen über Lebensmittel und Bedarfsgegenstände, C. H. Becksche Verlagsbuchhandlung, München, vol. 1, Nov. 1, 1990. 64. Käseverordnung, Neufassung April 14, 1986.
57. R. Frank, H. Mühlschlegel: Kunststoffe im Lebensmittelverkehr; Empfehlungen des Bundesgesundheitsamtes. Empfehlung XIV: Kunststoff-Dispersionen, 178. Mitteilung, Heymanns Vlg., 1990.
58. K. H. Wallhäusser, W. Fink, *Farbe + Lack* **76** (1970) 471; **82** (1976) 108.
59. Hoechst, DE 2 214 410, 1972 (K. Matschke et al.).
60. J. Schulze, *Beton* **41** (1991) 232–237.
61. W. M. Carpenter, M. F. Grower, G. Nash, *Oral Surg. Oral Med. Oral Pathol.* **42** (1976) 461–469.
62. Hoechst, Zentrale Marktforschung 1991.
63. *Kirk-Othmer* , **24**, Wiley, 1997, p. 947 .

Poly(Vinyl Ethers)

General aspects of polymers, properties and testing, processing, additives, and analysis are discussed in Plastics, General Survey; Plastics, Properties and Testing; Plastics, Processing; Plastics, Additives; and Plastics, Analysis, respectively. States of order are treated in Plastics, Properties and Testing. Fundamental aspects of polymerization reactions are treated in Polymerization Processes.

GERD SCHRÖDER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
2. Production 1
3. Industrial Processes 3
3.1. Batch Bulk Polymerization of Vinyl Methyl Ether 3
3.2. Continuous Bulk Polymerization of Vinyl Ethyl Ether 3
3.3. Continuous Solution Polymerization of Vinyl Isobutyl Ether 3
3.4. Environmental Protection 3
4. Properties and Uses 4
5. Copolymers 4
6. Toxicology 4
7. References 4

1. Introduction

Industrially important poly(vinyl ethers) are compounds with the general formula: -$(CH_2CHOR\text{-})_n$ where R is an alkyl group such as methyl, ethyl, isobutyl, or octadecyl.

Poly(vinyl ethers) were first prepared by WISLICENUS, who in 1878 produced a balsam-like product from 700 mols of vinyl ethyl ether [*109-92-2*] and 1 mol of iodine [1]. This reaction was recognized as a chain reaction involving a cationic polymerization initiated by I^+ and chain growth via carbonium ions.

In the 1920s the monomeric vinyl ethers [2] became readily available by Reppe vinylation of alcohols with acetylene and grew in industrial importance. Their polymerization was researched systematically.

After 1930 vinyl ethers and poly(vinyl ethers) were produced on an industrial scale, initially at the Ludwigshafen works of the I. G. Farbenindustrie (now BASF), and since 1940 at the General Aniline and Film Corporation (GAF) in the United States, and later also by Union Carbide (UCC) in the United States.

Today (1991) poly(vinyl ether) homopolymers are only produced by BASF; UCC and GAF have ceased production. Poly(vinyl ether) homopolymers are used as raw materials for adhesives.

Copolymers of vinyl ethers with other monomers, such as maleic anhydride, vinyl chloride, or acrylic esters, are produced by GAF and BASF. They are used as auxiliary materials for adhesives and paints, and in many other industries. General information on poly(vinyl ethers) can be found in [3–5].

2. Production

Monomers. Poly(vinyl ethers) are produced exclusively by polymerization of vinyl ether monomers.

In addition to the vinylation of alcohols with acetylene, the vinylation of alcohols with ethylene also leads to alkyl vinyl ethers, according to patents of Union Oil of California [6]. However the process has not become established industrially.

The very reactive α-methylvinyl ethers [7]

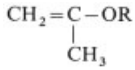

are only of academic interest.

With both strong and weak initiators and independent of the polarity of the solvent used, they give polymers with syndiotactic structures [8].

The stereochemistry of the polymers from β-methylvinyl ethers

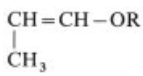

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_011

has been intensively investigated [9–15]. Polymers of vinyl ethers which are substituted at the alkyl group by, for example, chlorine [16], oxyethylene groups [17], a phthalimide group [18], an acetoxy group [19], a 9-carbazolyl group [20], a trifluoroalkyl group [21], or an acrylate or sorbate group [22], also have no industrial importance.

Of the industrially important vinyl ethers (methyl [*107-25-5*], ethyl [*109-92-2*], isobutyl [*109-53-5*], and octadecyl [*930-02-9*]), vinyl methyl ether has the greatest economic growth rate.

Initiators. *Anionic initiators* are unsuitable for the polymerization of vinyl ethers.

With *radical initiators* such as azoisobutyronitrile [23], peroxides [24], or redox systems [25], vinyl ethers polymerize to give homopolymers with a low reaction rate and a low degree of polymerization. However, copolymers from vinyl alkyl ethers and maleic anhydride in a molar ratio of 1 : 1 can be produced at a high rate with radical initiation, the polymer chains being constructed from both monomer units in a strictly alternating sequence.

Copolymers of vinyl ethers and vinyl chloride or vinyl ethers and acrylic esters contain a maximum of 50 mol % vinyl ether.

Cationic initiators are very suitable for the polymerization of vinyl ethers. They require high-purity monomers and solvents and clean apparatus; the absence of water is particularly important.

Friedel–Crafts catalysts , such as BF_3, $AlCl_3$, $SnCl_4$, or their complexes (e.g., BF_3 dietherate or BF_3 dihydrate), and also acidic aluminum sulfate, are widely used [26].

Other initiators such as halogens [27], protic acids [28], halogenated metal alkyls [29], cation-forming salts (e.g., $Ph_3C^+BF_4^-$) [30], Ziegler–Natta catalysts [31], and high-energy radiation [32], [33] are of theoretical interest.

Stereoregularity. Poly(vinyl ethers) were the first group of compounds in which the phenomenon of stereoregularity was recognized [34], [35]. Depending on the choice of polymerization conditions, either an amorphous or a crystalline polymer, which were later recognized as atactic or isotactic, can be produced from vinyl isobutyl ether [36]. For further details see [4].

Living Polymers. In 1984 it was discovered that in nonpolar solvents the initiator system $HI-I_2$ can convert vinyl isobutyl ether to living polymers [37]. The number-average molecular mass of the poly(vinyl isobutyl ether) increases with increasing monomer conversion, but is inversely proportional to the initial concentration of initiator. The molecular mass distribution is very narrow: the quotient of the weight-average and number-average molecular masses is close to 1, whereas for conventional poly(vinyl ethers) it is ca. 3. This interesting process is, however, not used commercially.

3. Industrial Processes

3.1. Batch Bulk Polymerization of Vinyl Methyl Ether

In one process 200 mL of an initiator solution (3 % boron trifluoride dihydrate in dioxane) is charged to a stirred-tank reactor containing 300 L vinyl methyl ether (*bp* 6 °C) at 12 °C. The reactor is cooled by jacket cooling and is equipped with a reflux condenser.

After the strongly exothermic polymerization is initiated, 5300 L vinyl methyl ether (in 10 h) and 6 L initiator solution (in 16 h) are continuously pumped into the stirred reactor against a regulated internal pressure of ca. 0.3 bar. The reactor temperature is ca. 15 °C. The heat of reaction (1400 kJ/kg monomer) is removed via the reactor jacket and the reflux condenser. Three hours after addition of the monomer is complete the temperature is increased to 60 °C, and 2 h later to 90 °C. After a further 2 h the reflux condenser is removed and the automatic cooling switched off. Two hours later the stirrer is switched off.

After a further 2 hours the reactor is degassed for 4 hours. The contents of the reactor are then either emptied out directly as a hot polymer melt or transferred to a mixer where they are converted into 50–70 % solutions with solvents such as water, ethanol, or toluene.

About 97 % of the monomeric vinyl methyl ether is converted to a light yellow soft resin, with $\bar{M}_w \approx 60\,000$ and $\bar{M}_n \approx 20\,000$ (commercial name Lutonal M 40). Polymers from vinyl ethyl ether (*bp* 36 °C) and vinyl isobutyl ether (*bp* 83 °C) can be produced analogously.

3.2. Continuous Bulk Polymerization of Vinyl Ethyl Ether

4600 mol/h (440 L/h) of vinyl ethyl ether and 0.4 mol/h of an initiator in solution are continuously pumped into a 2.5 m^3 stirred reactor. The reaction temperature is 98 °C, and the heat of reaction (ca. 1100 kJ/kg) is removed by jacket cooling and a reflux condenser.

The liquid level in the stirred reactor is kept constant by a siphon overflow. The polymer running out is collected in a larger receiver in which unreacted monomer is polymerized. The product is homogenized and then removed in batches as a viscous liquid. The molecular mass $\bar{M}_w$ is 3300 and $\bar{M}_n$, 1400 (commercial name Lutonal A 25). A viscous polymer with $\bar{M}_w = 17\,000$ can be produced analogously from vinyl isobutyl ether (commercial name Lutonal I 30).

3.3. Continuous Solution Polymerization of Vinyl Isobutyl Ether

The low-temperature polymerization of vinyl isobutyl ether in liquid propane on a conveyor belt [26] is no longer used due to poor economics.

3.4. Environmental Protection

The off-gases formed in the production of poly(vinyl ethers) are contaminated by solvent and monomer vapors; they are collected and incinerated. The off-gases from incineration are discharged to the atmosphere. In the market, solvent- containing poly(vinyl ether) products are declining in importance, while 100 % polymers and aqueous systems are exhibiting growth – a favorable tendency for environmental protection.

4. Properties and Uses

Consistency and Compatibility. Commercial poly(vinyl ethers) are highly viscous oils, soft adhesive resins, or nonadhesive elastomers, depending on the molecular mass. Poly(vinyl octadecyl ether) has a waxlike consistency.

Poly(vinyl ethers) are resistant to saponification by dilute acids and bases. Their solubility in water and organic solvents and their compatibility with other polymers depends on the nature of the alkyl group. For example, poly(vinyl methyl ether) (Lutonal M 40) is soluble in aromatic hydrocarbons, insoluble in aliphatic hydrocarbons, and soluble in alcohols and cold water. Formation of hydrogen bonds between the ether oxygen and water is responsible for the solubility in cold water. These bonds are destroyed on heating, and above 28 °C the polymer precipitates from aqueous solution.

Poly(vinyl methyl ether) forms clear films when mixed with many other polymers such as methyl cellulose, ethyl cellulose, nitrocellulose, epoxy resins, natural rubber, poly(vinyl acetate), vinyl acetate – vinyl chloride copolymers, and polystyrene. It can also be used as a plasticizer.

Poly(vinyl ethyl ether), a highly viscous oil, soft resin,or nonadhesive material,depending on the molecular mass, is soluble in ethers, esters, ketones, and hydrocarbons, but insoluble in water.

Poly(vinyl isobutyl ether) is produced as a highly viscous oil and as a soft resin, but also as a secondary emulsion in water for mixing with other aqueous polymer emulsions.

Poly(vinyl octadecyl ether) is a brittle solid which melts at 50 °C and is compatible with many waxes.

Uses. Uses of some commercial poly(vinyl ethers) are listed in Table 1.

The economically most important use of poly (vinyl ethers) is in the adhesives industry, mostly as modifying additives for other (cheaper) raw materials. Depending on the nature of the poly-(vinyl ether), adhesives for labels, for flexible tapes, or adhesives which stick well to polyethylene, on a moist background, or at low temperatures can be produced. For further information on the uses of poly(vinyl ethers) see [5], [38].

Table 1. Uses of poly(vinyl ethers)

Poly(vinyl ether)	Trade name	Consistency	Use
Methyl	Lutonal M 40	soft resin	raw material for adhesives
	ca. 50 % in water	highly viscous solution	plasticizer thickener heat sensitizer
	70 % in ethanol	highly viscous solution	raw material for adhesives
	70 % in toluene	highly viscous solution	raw material for paints
Ethyl	Lutonal A 25	highly viscous oil	raw material for paints auxiliary for printing inks
	Lutonal A 50	soft resin	raw material for adhesives
	70 % in naphtha	highly viscous solution	plasticizer
	70 % in ethanol	highly viscous solution	plasticizer
	70 % in toluene	highly viscous solution	plasticizer
	Lutonal A 100, ca. 25 % in hexane	highly viscous solution	raw material for adhesives
Isobutyl	Lutonal I 30	highly viscous oil	raw material for adhesives
	Lutonal I 60	soft resin	raw material for adhesives
	50 % in heavy naphtha	highly viscous solution	raw material for paints
	80 % in light naphtha	highly viscous solution	raw material for adhesives
	Lutonal I 60 D	medium-viscosity aqueous dispersion	raw material for adhesives
	Lutonal I 65 D	medium-viscosity aqueous dispersion	raw material for adhesives
Octadecyl	Luwax V	wax	polish for floors and other surfaces

5. Copolymers

Vinyl ethers readily undergo cationic copolymerization with one another, but the resulting copolymers are unimportant industrially. Radically produced copolymers from vinyl ethers and other monomers contain a maximum of 50 mol % vinyl ether. Solutions (e.g., in ethyl acetate) or aqueous emulsions of radical copolymers containing up to 30 % vinyl isobutyl ether are used as raw materials for adhesives with special adhesive and compatibility properties. Copolymers with 75 % vinyl chloride and 25 % vinyl isobutyl ether are used as binding agents for ageing-resistant, flexible anticorrosion paints.

Copolymers of vinyl methyl ether and maleic anhydride are very versatile in their applications, being used as protective colloids, thickeners, pigment dispersants, film formers, and binding agents. Products of this type are produced by GAF (trade name Gantrez) and BASF (trade names Sokalan, Luviform, Lupasol) for use as detergent additives, in the paper and textile industries, in cosmetics, and in the production of microcapsules. For further information see [5]. GAF sells the copolymer as a white powder, which, on dissolving in water, forms a maleic acid – vinyl methyl ether copolymer and, on dissolving in an alcohol, a maleic acid half ester – vinyl methyl ether copolymer.

6. Toxicology

For the common poly(vinyl alkyl ethers) known for nearly 50 years (methyl, ethyl, isobutyl, and octadecyl), no effects which are hazardous to health have so far been found. In animal experiments they are classified as nonirritant to the skin and nonmutagenic. The oral toxicity LD_{50} (rat) is > 500 mg/kg.

The toxicity of organic poly(vinyl ether) solutions is determined essentially by the solvent. Toluene and *n*-hexane present in naphtha have a narcotic effect, irritate the eyes and skin, and can produce permanent damage to the nervous system. Good workplace ventilation is necessary.

7. References

1. I. Wislicenus, *Justus Liebigs Ann. Chem.* **192** (1878) 106.
2. *Ullmann,* 4th ed., **23,** 608 – 611.
3. J. P. Kennedy, E. Marechal: *Carbocationic Polymerization,* Wiley-Interscience, New York 1982, pp. 499 ff.
4. T. Higashimura, M. Sawamoto: "Carbocationic Polymerization: Vinyl Ethers," in G. C. Eastmond et al. (eds.): *Comprehensive Polymer Science,* vol. 3, Pergamon Press, Oxford 1989, pp. 673 – 696.

5. M. Biswas, A. Mazumdar, P. Mitra: "Vinyl Ether Polymers," *Encycl. Polym. Sci. Eng.* **17** (1989) 446–468.
6. Union Oil of California, US 4 057 575, 1977 (D. L. Klass); US 4 161 610, 1979 (D. L. Klass).
7. M. Goodman, Y. L. Fan, *J. Am. Chem. Soc.* **86** (1964) 4922, 5712.
8. K. Matsuzaki, S. Okuzono, T. Kanai, *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 3447.
9. G. Natta, *J. Polym. Sci.* **48** (1960) 219.
10. G. Natta, M. Peraldo, M. Farina, G. Bressan, *Makromol. Chem.* **55** (1962) 139.
11. T. Higashimura, Y. Hirokawa, K. Matsuzaki, T. Uryu, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 1489.
12. K. Matsuzaki et al., *Makromol. Chem.* **182** (1981) 2421.
13. T. Higashimura et al., *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 2511.
14. T. Higashimura et al., *J. Polym. Sci. Polym. Chem. Ed.* **15** (1977) 2691; *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 1473.
15. Y. Hirokawa et al., *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 3923; *Polym. Bull. (Berlin)* **1** (1979) 365.
16. T. Higashimura, Y. M. Law, M. Sawamoto, *Polym. J.* **16** (1984) no. 5, 401–406.
17. T. Nakamura, S. Aoshima, T. Higashimura, *Polym. Bull. (Berlin)* **14** (1985) no. 6, 515–521.
18. T. Hashimoto, H. Ibuki, M. Sawamoto, T. Higashimura, *J. Polym. Sci. Polym. Chem. Ed.* **26** (1988) no. 12, 3361–3374.
19. S. Aoshima, T. Nakamura, N. Vesugi, M. Sawamoto, *Macromolecules* **18** (1985) no. 11, 2097–2101.
20. S. A. Haque, T. Uryu, H. Ohkawa, *Makromol. Chem.* **188** (1987) no. 11, 2523–2533.
21. W. O. Choi, M. Sawamoto, T. Higashimura, *Polym. J.* **20** (1988) no. 3, 201–206.
22. S. Aoshima, O. Hasegawa, T. Higashimura, *Polym. Bull. (Berlin)* **14** (1985) 417–423.
23. N. M. Bortnik, S. Melamed, US 2 734 890, 1956.
24. J. F. Nelson, F. W. Banes, W. P. Fitzgerald, US 2 967 203, 1961.
25. R. R. Dreisbach, J. L. Lang, US 2 859 209, 1958.
26. *Ullmann,* 4th. ed., **19,** 382–385.
27. D. D. Fley, J. Saunders, *J. Chem. Soc.* 1954, 1668.
28. F. Bolza, F. E. Treloar, *Makromol. Chem.* **181** (1980) 839.
29. Y. Kishimoto, S. Aoshima, T. Higashimura, *Macromolecules* **22** (1989) no. 10, 3877–3882.
30. T. Kunitake, K. Takarabe, *Makromol. Chem.* **182** (1981) 817.
31. C. E. Schildknecht, I. Skeist: "Vinyl Ether Polymerizations," in: *High Polymers,* vol. 29, Wiley-Interscience, New York 1977, pp. 321–327.
32. Ka. Hayashi, Ko. Hayashi, S. Okamura, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 2305.
33. A. Deffieux et al., *Polymer* **24** (1983) 573.
34. C. E. Schildknecht, A. O. Zoss, C. McKinley, *Ind. Eng. Chem.* **39** (1947) 180.
35. C. E. Schildknecht et al., *Ind. Eng. Chem.* **40** (1948) 2104.
36. G. Natta, I. Bassi, P. Corradini, *Makromol Chem.* **18/19** (1956) 455.
37. M. Miyamoto, M. Sawamoto, T. Higashimura, *Macromolecules* **17** (1984) 265.
38. W. J. Müller, *Adhäsion* **25** (1981) no. 5, 208–213.

Poly(Vinyl Fluoride) → Fluoropolymers, Organic

Poly(Vinylidene Chloride)

YANNICK VANDENDAEL, Solvay & Cie., S.A., Brussels, Belgium

1. Introduction 1
2. Production 2
2.1. Emulsion Polymerization 2
2.2. Suspension Polymerization 3
3. Structure and Properties 5
3.1. Homopolymers 5
3.2. Copolymers 5
4. Processing and Uses 8
4.1. Extrusion 8
4.2. Films and Sheets 9
4.3. Lacquering and Coating 10
4.4. Uses 12
5. Economic Aspects 12
6. Toxicology and Occupational Health . 13
7. References 13

Based in part on the corresponding article in Ullmann's, 4th ed., written by Karsten Flatau.

1. Introduction

Vinylidene chloride (1,1-dichloroethylene), VDC, was first obtained by V. REGNAULT in 1838 by treating 1,1,2-trichlorethane with alcoholic caustic potash solution to produce a colorless, chemically impure liquid having a boiling point of 35–40 °C which was subsequently recognized as being impure VDC [2]. The processes used today are still based on this reaction (see → Chlorinated Hydrocarbons, Chap. 3.2.).

When vinylidene chloride is stored and particularly when exposed to light (E. BAUMANN, 1872), a white precipitate is formed.

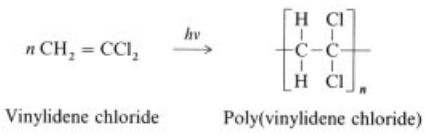

This precipitate was identified by FEISST and STAUDINGER (1930), who carried out the first systematic polymerization experiments [3], [4], as poly(vinylidene chloride) homopolymer. It was found that poly(vinylidene chloride) (PVDC) could not be processed thermoplastically because of its thermal instability. R. M. WILEY [5] succeeded in obtaining processable materials by copolymerization and plasticization. Because of the special properties of PVDC, this polymer was intensively researched and further developed in the 1930s (in the United States by Dow in particular, and in Europe by BASF). For further information on the historical development, see [6–12].

The first poly(vinylidene chlorides) were marketed more than 50 years ago. They were originally used mainly for resistant fibers and bristles. Coatings of PVDC were first applied in significant quantities around 1940 in the United States and in Germany. The first extruded shrink PVDC films for meat packaging also appeared at that time.

However, large-scale development of this class of products did not start until the 1950s when improved PVDC resins for coatings came onto the market; early examples of this class of high-grade polymers were the soluble Saran resins (F and B) and Saran dispersions of Dow and the Diofan dispersions of BASF.

All conventional commercial PVDC products are now copolymers of vinylidene chloride with small amounts of other monomers. The VDC content of these polymers is usually between 75 and 95 %, and in special cases is as low as 50 % (e.g., VDC – VCM copolymers).

2. Production

Due to its asymmetric structure, VDC polymerizes readily at its double bonds; comonomers are used for modification. The main comonomers are methyl acrylate, ethyl acrylate, acrylonitrile, methyl methacrylate, butyl acrylate, and vinyl chloride.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_017

Small amounts of unsaturated carboxylic acids or other products with adhesion-promoting groups are also used to obtain improved anchorage or other special coating properties.

Two processes are used to produce PVDC copolymers on an industrial scale, which are of comparable commercial importance: emulsion polymerization and suspension polymerization, also known as bead polymerization.

Vinylidene chloride polymerizes in the presence of ionic and radical initiators and Ziegler catalysts and under the influence of radiation. Ionic polymerization has the disadvantage that initiators such as butylithium react with the polymer and accelerate its degradation. Up to now, only radical polymerization in aqueous systems has been of industrial interest.

Comparison of the Processes. On the commercial scale, the heterogeneous polymerization of VDC is carried out by the suspension and emulsion processes in presence of radical sources.

Radiochemical and anionic initiation suffer from the impossibility of excluding polymer degradation, and cationic polymerization gives low molecular mass polymers. Ziegler polymerization of VDC has never been developed.

Homogeneous solution polymerization has not been described in patents. Bulk polymerization and precipitation polymerization in inert solvents such as cyclohexane or benzene cannot be controlled in practice owing to the associated engineering problems such as stirring and heat removal. In bulk polymerization, the reaction mixture thickens to a paste at 10 – 20 % conversion. In the presence of inert solvents the reaction is too slow, only low molecular masses are attained, and the polymer must be freed of solvent.

Suspension polymerization in water can be well controlled but, especially in the case of VDC – VC copolymers, requires quite long reaction times to reach the high molecular masses attained in emulsion polymerization. Owing to their low concentration of polymerization additives, suspension polymers are generally more stable and less sensitive to water.

The composition of the copolymer can be better controlled in emulsion polymerization by the metering in of monomer.

Suspension and emulsion polymerization are carried out in stirred, glass-lined or stainless steel autoclaves. Heat is supplied or removed by jacket heating or cooling. The reactors are flushed free of oxygen before polymerization. The water added must be free of metallic impurities, or the thermal stability of the polymers is impaired.

Purity and Storage of the Monomers. The VDC monomer used for polymerization contains 200 ppm *p*-methoxyphenol or 0.6 – 0.8 % phenol as stabilizer. Polymerization is practically unaffected by *p*-methoxyphenol, while phenol must be removed before polymerization by distillation or alkaline extraction.

Vinylidene chloride is stored under a nitrogen protective atmosphere in nickel-plated or phenol- coated vessels. The oxygen content in the stock tank must not exceed 100 ppm.

On storage VDC can form peroxides, especially with nonstabilized VDC and oxygen. These peroxides can lead to spontaneous polymerization, but are also explosive in the presence of air or oxygen, even at low temperatures (– 40 °C). The peroxides can be detected by their pungent, acrid smell or by the formation of flocculent deposits of polymer. They are deactivated on storage under water, and can be destroyed by washing with a 5 % solution of methanol in tetrachloroethylene. PVDC peroxide decomposes into phosgene and formaldehyde [13].

2.1. Emulsion Polymerization

VDC is polymerized by batch [14] or monomer-addition [15–19] processes. The mechanism of the polymerization cannot be described by the Smith – Ewart theory [20]. Typical polymerization recipes in parts by weight are as follows:

Batch process

Vinylidene chloride	78
Vinyl chloride	22
Water	180
Potassium peroxosulfate	0.22
Sodium metabisulfite	0.3
Sodium dioctylsulfosuccinate (80 %)	3.58

Monomer-addition process

Vinylidene chloride	468
Comonomer	52
Dihexylsulfosuccinate	15
Sodium lauryl sulfate (5 %)	5
Ammonium peroxosulfate (5 %)	10
Sodium metabisulfite	10
Water	440

In batch polymerization at 30 °C, 95 – 98 % conversion is reached after 7 – 8 h [15]. Polymerization to completion of batch charges gives copolymers with broad comonomer distributions [21], except when the copolymerization parameters r_1 and r_2 equal 1.

Figure 1 shows a flow sheet of the monomer-addition process.

The monomers are charged to a mixing tank, and aqueous solutions of the emulsifiers and the initiator and of the redox activator are prepared. After the reactor has been closed and heated to 40 °C, one-tenth of the monomer mixture and one-tenth of the aqueous initiator and redox activator solutions are run in via computer-controlled metering pumps. After the reaction has started, the three streams are added at preprogrammed rates. The reactor internal temperature is kept constant by means of the jacket temperature, and the reactor pressure approximately constant by addition of the activator. After completion of the main reaction, further initiator and activator are added to allow the system to react to completion. The residual monomers are withdrawn at elevated temperature under vacuum, or unreacted VDC is stripped out by treating the dispersion with steam or hot inert gases. The resulting latex has a solids content of 52 % and can be transported as such or coagulated to give a resin. The wall deposits formed during polymerization are removed between batches, e.g., by spraying with high-pressure water.

In VDC emulsion (co)polymerization, anionic emulsifiers such as sodium alkylarylsulfonates, alkyl esters of sodium sulfosuccinic acid, and fatty alcohol sulfates are used alone or in combination with nonionic surfactants such as ethoxy-lated alkylphenols. The usual initiators are peroxosulfates, hydrogen peroxide, organic hydroperoxides, peroxoborates, and peroxocarbonates in the presences of redox activators such as bisulfites or formaldehyde sodium bisulfite. Redox systems are used to obtain high molecular mass products at low temperatures and acceptable space – time yields.

Polymerization is preferably carried out at pH 2 – 4 to avoid attack by bases, and the operating temperature must be below 80 °C to suppress polymer degradation. Feed tanks, polymerization reactors, and monomer removal units are often made of glass-lined steel to avoid metallic impurities.

Because the reaction is carried out under controlled conditions, copolymers manufactured by the continuous addition process have uniform comonomer distribution [15–19]. The copolymers are present in the latex state in the amorphous form and can be used directly as coating materials with good film-forming properties.

The principal VDC copolymers produced by emulsion polymerization are those with acrylonitrile and/or various (meth)acrylates [22–27], while VDC – vinyl chloride copolymers are produced in smaller volume [14], [28], [29]. The copolymers are used as dispersions of 50 – 60 % solids content, or first coagulated, washed, and dried, and then processed as solution. For coagulation, the dispersion is diluted to 30 % solids content to lower its viscosity, and aluminum or calcium salts are added. The polymer is then filtered, washed, and dried.

2.2. Suspension Polymerization [11], [12]

A typical charge for the suspension polymerization of VDC (in parts by weight) is as follows

Vinylidene chloride	80
Vinyl chloride (VC)	20
Water	200
Methyl hydroxypropyl cellulose	0.2
Dilauroyl peroxide	0.6

Figure 2 shows a flow sheet of the process. In 30 – 60 h at a polymerization temperature of 60 °C a conversion of 85 – 90 % is achieved.

Residual monomers can be removed by steam treatment in downstream stripping columns or in the polymerization reactor itself by vacuum degassing or steam treatment at reduced pressure. Residual monomers are recycled. The reactors are cleaned with high-pressure water after each batch.

The monomer-free suspension enters intermediate tanks of rustproof material in which sev-

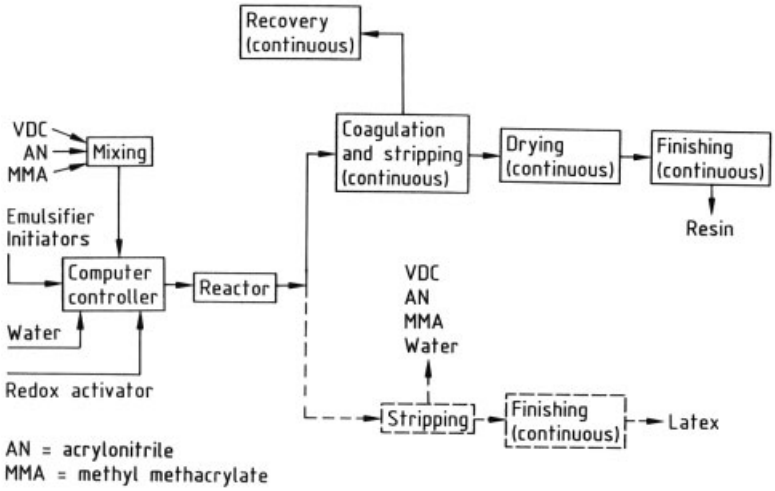

Figure 1. Flow sheet for the production of VDC copolymers by monomer-addition emulsion polymerization

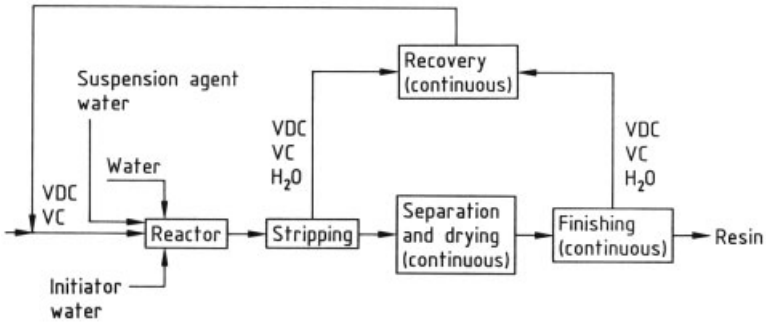

Figure 2. Flow sheet for the production of VDC – VC copolymers by suspension polymerization

eral batches can be homogenized. The tanks also provide a buffer for the continuous processing of the product.

After centrifuging off and washing with water, the product is dried at ca. 50 – 70 °C in cocurrent flow in a drum dryer, sieved, and then conveyed pneumatically to storage bins.

Whereas polymerization was formerly carried out only in glass-lined reactors, since ca. 1970 stainless steel reactors have also been used.

In suspension polymerization organic peroxides, peroxocarbonates, and azo compounds are used as initiators, and poly(vinyl alcohols) and methyl celluloses as suspension agents. Before polymerization, the initiator must be distributed uniformly in the monomer mixture to avoid different rates of polymerization in the dispersed monomer droplets. Droplets that polymerize rapidly are converted to hard, glassy suspended particles which are difficult to process. The polymerization rate increases with increasing initiator concentration and temperature. The suspension agents influence the particle size distribution, porosity of the particles, and clarity and transparency of the processed products. Secondary suspension agents such as sorbitan monostearate improve the porosity and plasticizer absorption of the polymer powder.

Reproducible setting of adefined comonomer distribution in the polymer is more difficult in suspension polymerization than in emulsion polymerization. Unmodified batch operation [30], as well as temperature-programmed polymerization and the use of initiators with different decomposition rates [31], lead to a polymer mixture with heterogeneous polymer composition. In VDC – VC copolymerization, uniform monomer incorporation is achieved by contin-

uously removing the less reactive monomer to maintain a constant volume ratio in the unreacted monomer mixture [32]. The polymerization rate can be increased by stepwise polymerization with more active initiators such as peroxodicarbonates [33].

Suspension polymerization is mainly used to produce VDC – VC copolymers [31], [34], which are suitable for injection molding and extrusion molding. Modified VDC – alkyl acrylate [35] and VDC – methyl acrylate copolymers [36] can be processed by blow molding.

3. Structure and Properties [10–12]

3.1. Homopolymers

PVDC does not dissolve in monomeric vinylidene chloride, and polymer and monomer are practically insoluble in water. Vinylidene chloride polymerizes in the presence of radical sources to give a linear, symmetrical macromolecule with head-to-tail structure and the repeat unit

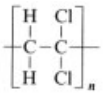

The theoretical chlorine content of the polymer is 73.14 %, although in practice chlorine contents of 72 – 73 % are obtained.

The polymer chain is tightly packed in the solid state, and in the thermodynamically stable state it is highly crystalline (ca. 75 %), indicating the absence of defects or branches. The crystallinity and rate of crystallization are influenced by the polymerization conditions, the mechanical and thermal treatment of the polymer, and, in the case of copolymers, the comonomer content.

The combination of high crystallinity and high density results in PVDC having very good barrier properties against gases and vapors and resistance to solvents.

The absolute molecular mass of PVDC cannot be determined by conventional methods due to the lack of a suitable solvent. Like other high-melting polymers, PVDC is sparingly soluble. Only at temperatures of 130 °C can 1 % solutions be prepared in, e.g., 1,3-dibromopropane or bromobenzene. At room temperature PVDC dissolves only in polar solvents such as hexamethylphosphoric triamide or tetramethylene sulfoxide; amorphous PVDC dissolves in tetrahydrofuran.

The structure and solubility of PVDC are independent of the polymerization temperature.

PVDC decomposes at temperatures above 125 °C, chiefly with elimination of hydrogen chloride and formation of polyene sequences. The degraded polymer is infusible, and the crystalline structure is destroyed [37]. The polymer is also degraded by the action of high-energy radiation, bases, and heavy metal ions.

Some properties of PVDC are as follows:

Melting point	198 – 205 °C
Glass transition temperature	− 19 to − 11 °C
Amorphous density (25 °C)	1.67 – 1.775 g/cm^3
Crystalline density (25 °C)	1.80 – 1.97 g/cm^3
Degree of polymerization	100 – 10 000
M_w/M_n	1.5 – 2.0
Softening range	100 – 150 °C
Ceiling temperature	350 – 400 °C

3.2. Copolymers

Industrially important copolymers are those manufactured by radical polymerization of VDC with vinyl chloride, acrylonitrile, and (meth)-acrylate esters. Table 1 lists the copolymerization parameters of VDC with the most important comonomers [10]. The Q and e values of VDC are $Q = 0.20$ and $e = 0.36$ [38].

Table 1. Copolymerization parameters of vinylidene chloride with various comonomers

Comonomer	r_1	r_2	Temperature, °C
Styrene	0.14	2.0	60
Vinyl chloride	3.2	0.3	60
Acrylonitrile	0.37	0.91	60
Methyl acrylate	1.0	1.0	60 – 70
Methyl methacrylate	0.24	2.53	60
Vinyl acetate	6	0.1	68

Copolymers with high VDC content have lower melting points and, at least at low comonomer content, higher glass transition temperatures than the PVDC homopolymer. Figures 3 and 4 show the effect of the copolymer composition on the melting point and glass transition temperature.

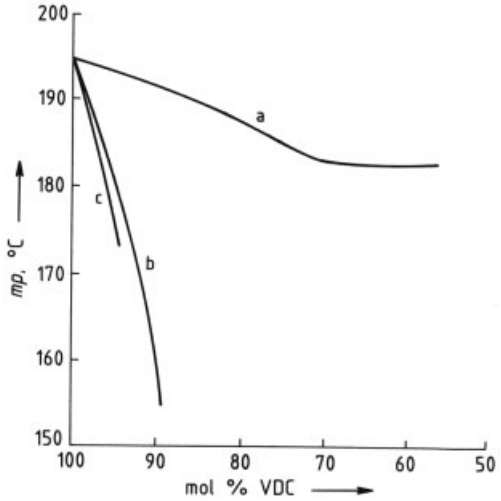

Figure 3. Dependence of melting point on copolymer composition
a) VDC – vinyl chloride; b) VDC – methyl acrylate; c) VDC – octyl acrylate

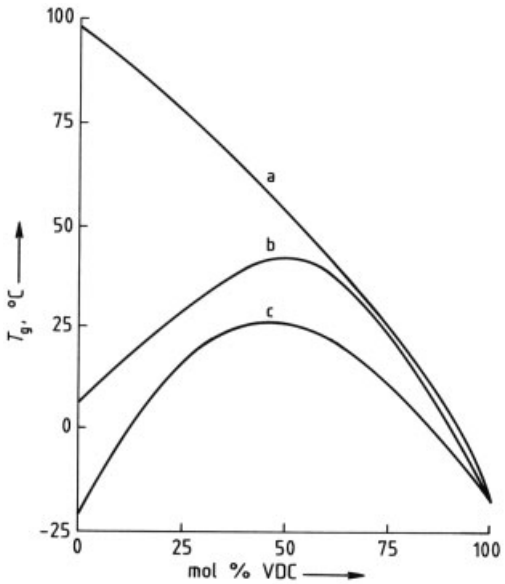

Figure 4. Dependence of glass transition temperature on copolymer composition
a) VDC – acrylonitrile; b) VDC – methyl acrylate;c) VDC – ethyl acrylate

VDC copolymers are more stable to degradation than pure PVDC. Comonomers interrupt the unstable poly(vinylidene chloride) chain. The copolymerization must therefore be conducted so that as few long PVDC sequences as possible are present in the copolymer. Furthermore, VDC – VC copolymers are more stable than PVDC since the chlorine – carbon bond is stronger in the vinyl chloride unit than in the VDC unit. Polar functional groups, however, weaken the chlorine – carbon bond, so that copolymers with polar functional groups such as the nitrile group are less stable due to neighboring group effects.

The mechanical properties of VDC copolymers depend on temperature, molecular mass, crystallinity, degree of orientation, and plasticizer content. With increasing crystallinity the strength increases, whereas toughness and elongation decrease. The above properties are improved by orientation.

Incorporation of vinyl chloride, for example, in the VDC polymer chain raises the glass transition temperature, which leads to less favorable low-temperature behavior (increased brittleness). The glass transition temperature can be lowered again without loss of strength properties by plasticization (2 – 10 % plasticizer).

Crystallinity. The performance of PVDC films and coatings is greatly influenced by the high chlorine content of the polymers and especially their crystallinity.

The polymer molecule is very symmetrical if the proportion of VDC in the polymer is high, which leads to crystallization. It has been shown that the crystallinity is directly proportional to the mole fraction of VDC in the polymer. Thus, homopolymers of VDC are highly crystalline, i.e., they contain between 50 and 80 % of crystalline material, the actual figure depending on the polymerization conditions. For this reason, they are normally unsuitable for further processing.

The symmetry of VDC polymer chains is disturbed by the incorporation of comonomers such as methyl acrylate. The crystallinity decreases steadily with decreasing mole fraction of VDC until it is completely suppressed at ca. 80 mol % VDC.

If the mole fraction of VDC in the copolymer is very high (93 % VDC), the polymer particles of a PVDC dispersion may already crystallize in the aqueous phase. It is very difficult for such highly crystalline latex particles to fuse together, with the result that film formation on drying is hardly possible. For this reason, the VDC mole fraction in PVDC dispersions for general applications usually lies between 87 and 92 mol %.

The partially crystalline morphology of PVDC copolymers can best be described by a network of continuously interpenetrating crystalline and amorphous zones.

Studies on crystallization of coatings have revealed that if the mole fraction of VDC in the copolymer is ca. 88 %, the amorphous state is retained for a long time (200 – 300 d under normal conditions) and that the crystalline phase is formed very slowly.

If the coatings contain a high proportion of VDC (91 – 92 %), crystallization already commences after 1 d and attains a maximum within 5 – 20 d, depending on the product. This stage is referred to as primary crystallization. Further, secondary crystallization is slight and proceeds slowly.

Table 2. Demands imposed on the properties of PVDC in relation to the mole fraction of VDC in the copolymers and the crystallinity

Demands on properties of PVDC	Effect of *	
	Increasing VDC mole fraction	Increasing crystallinity
Specific properties of PVDC		
impermeability to :		
water and water vapor	+	+
gases especially oxygen	+	+
aroma	+	+
resistance to solvents and chemicals	+	+
resistance to blocking	+	+
slip	+	+
flexibility	—	—
heat sealability	—	—
clarity	+	+
PVDC processability		
solubility (soluble grades)	—	—
film forming (aqueous dispersions)	—	—
schemal stability (extrudable)	—	—
film blowing aptitude (extrudable)	—	—

* + = improvement, — = deterioration.

Table 2 reviews the effect of increasing the VDC content in copolymers and the crystallinity on the properties of coatings and the processability of PVDC.

Permeability [39]. It has been found that the most water vapor-resistant polymers are the most oxygen permeable and vice versa, with the exception of PVDC, which acts as a very good barrier to both (see → Films, Chap. 2. for comparison of permeability data). Table 3 lists the permeability of three PVDC grades.

Aromaproofness is closely related to the impermeability to water vapor or gases. It is important in practice that the contents of a pack should not lose their aroma as a result of diffusion from the inside to the outside. Likewise, their aroma must not be impaired by the diffusion of odoriferous substances from the outside to the inside.

The very good aromaproofness derives from its high resistance to diffusion and low solubility, particularly that of the crystalline sections of the polymers.

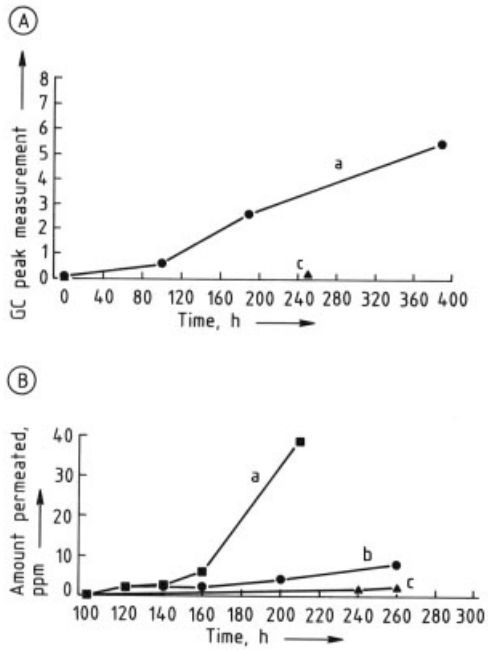

Figure 5. Aroma barrier properties of 25 µm plastic films
A) Menthol permeation at 0.2 ppm concentration (GC measurements); B) Limonene permeation at 2.5 ppm
a) Coextruded OPP film without barrier; b) OPP film coated with acrylic; c) OPP film coated with PVDC

Table 3. Permeability of PVDC

PVDC grade	Permeability		
	to oxygen (20 – 25 °C) *	to carbon dioxide (20 – 25 °C) *	to water vapor (38 °C, 90 % RH) **
Coating from PVDC lacquer	15 – 25	45 – 80	15 – 25
	10 – 15	30 – 45	10 – 15
Coating from PVDC dispersion	40 – 80	90 – 115	15 – 25
	15 – 40	35 – 90	8 – 15
PVDC film: plasticised slightly or unplasticised	300 – 1000	650 – 2200	50 – 100
	40 – 80	135 – 265	15 – 25

* Permeability in cm^3 µm m^{-2} d^{-1} bar^{-1} (dry gas).
** Permeability in g µm m^{-2} d^{-1}.

Figure 5 illustrates the good aroma barrier of PVDC- coated BOPP (biaxially oriented polypropylene) films (26 µm) in comparison with coextruded OPP and OPP coated on both sides with acrylic for limonene and menthol.

Resistance to Oils, Fats, and Chemicals. Pinhole-free films or coatings of PVDC offer high resistance to animal and vegetable fats and oils. It is not swollen by oil or grease, nor does it permit their diffusion. Exceptions are strongly unsaturated vegetable oils, which may cause amorphous grades of PVDC to swell somewhat.

Poly(vinylidene chloride) exhibits excellent chemical resistance. The crystallinity, the large volume proportion of chlorine, as well as the high density are responsible for the high resistance to numerous solvents and chemicals.

4. Processing and Uses

4.1. Extrusion

The commercial extrusion grades of PVDC are copolymers of vinylidene chloride with comonomers such as vinyl chloride, acrylic or methacrylic acid esters, and others. They are polymerized in emulsion and coagulated, or directly polymerized in suspension; they are generally available in powder form.

Normal processing formulations of those copolymers require a plasticizer and/or a stabilizer; some include, if necessary, a lubricant for the extrusion process, a slip agent if requested by the final properties of the extrudate, pigments, light stabilizers, and impact modifiers.

A typical PVDC film formulation, for instance, contains 0 – 10 % plasticizer (e.g., dibutyl sebacate, acetyltributylcitrate), 0.5 – 5 % stabilizer (e.g., epoxidized soybean oil, sodium pyrophosphate), 0 – 1 % lubricant (e.g., fatty acids and fatty acid amines, montan wax). The dry blend (formulated PVDC) is extruded in single-screw extruders at a compression ratio from 1/3.5 to 1/4.5. A ratio of screw length to screw diameter of 10 – 24 has proven to be satisfactory.

Since PVDC, even stabilized, does not tolerate long periods at high temperature, which causes the release of hydrogen chloride and leads to carbonisation, all metal parts of the extruder and dies in contact with hot molten polymer must be made of corrosion-resistant, noncatalytic metals. Recommended materials are chromium (particularly for the screw) and nickel alloys.

4.2. Films and Sheets

One of the major outputs for PVDC resins is the manufacture of films.

It is possible to process PVDC as monolayer film or as part of a coextruded sandwich, both techniques allowing the preparation of shrink or nonshrink materials.

Shrinkage can be obtained by applying tension before cooling (orientation) or by crosslinking.

Monofilm Biorientation. Usual film process for monolayer PVDC film requires the extrusion of a molten tube; the temperature of the

PVDC melt is normally 150 – 160 °C, depending on the selected additives formulation.

The shrink film bubble process can be applied to both mono and coextruded film. Immediately after the die, the molten tube passes into a cooling bath (2 – 10 °C) which keeps the quenched cooled material in an amorphous state. The quenched, nontacky material is then slightly rewarmed in an oven and blown into a bubble. As the tube of amorphous polymer is oriented, crystallization takes place rapidly.

If a heat-stabilized film is desired, a further annealing step is applied by re-inflating the film and passing it through an infrared heating tunnel.

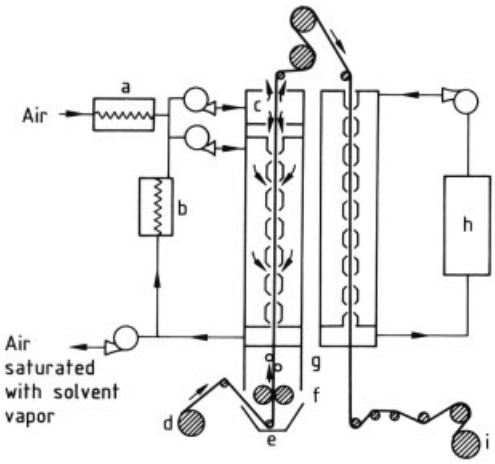

Figure 6. Cellulose film coating machine
a) Air preheater; b) Air heater; c) High-pressure chamber; d) Supply roll; e) Coating bath; f) Calibrating rolls; g) Smoothing roll; h) Rehumidifier; i) Receiving roll

Coextruded Film. Since PVDC monofilm does not meet all requirements of film packaging, although formulations already allow great versatility, PVDC films are being increasingly coextruded with other materials. This technique is used to optimize the film properties by combining materials in such a way, so as to add their own advantageous properties, while minimizing their drawbacks. Furthermore, it provides better PVDC barrier properties. Since the PVDC is no longer a self-supporting material, it is possible to reduce the amount of liquid additives that adversely affect barrier properties but are necessary in monolayer extrusion to insure mechanical properties. Also, it allows lower thickness of the PVDC film to be used if required.

Two coextrusion techniques have reached a high level of development: multimanifold die coextrusion and feedblock coextrusion.

The common feature of both techniques is the ability to process the materials that will be used in the final complex, on different extruders, the difference being how the different melts are joined.

In *multimanifold coextrusion*, each molten material is brought into the die through its own flow and distribution channel.

In the *feedblock system* the materials are combined before they enter the die.

Limitations of the feedblock system are that relative viscosities and extrusion temperatures of the different melts must be similar.

Permeability data for extruded and coextruded PVDC films are summarized in Table 4.

Coextruded sheets are produced with feedblock systems in the same way as for coextruded films. The more common structures are PP – TIE – PVDC – TIE – PP or PE; PS – TIE – PVDC – TIE – PS or PE; and PVC – PVDC – PVC, where TIE denotes an adhesive (tie) layer.

4.3. Lacquering and Coating [40–43]

PVDC lacquer resins (powder form) are vinylidene chloride copolymers modified to give solvent solubility; vinylidene chloride homopolymer is almost insoluble in usual solvents. The usual comonomers are acrylonitrile, vinyl chloride, acrylates, and methacrylate esters.

Copolymers with very high vinylidene chloride contents (> 90 %) are generally only soluble in THF mixtures.

Resins containing between 80 and 90 % VDC are less crystalline and therefore are also soluble in methyl ethyl ketone (MEK) mixtures.

Materials with lower VDC content are soluble in common organic solvents at room temperature, such as esters, ketones, and cyclohexanone.

Coating of Regenerated Cellulose (Cellophane). A major field of application for PVDC lacquers with a high VDC content in the copolymer is the coating of cellulose film.

These tailor-made PVDC resins develop high crystallinity and are therefore soluble only in aggressive solvents such as mixtures of

Table 4. Permeability of various PVDC structures

PVDC grade	Structure	Permeability to water vapor, g m^{-2} d^{-1}	Permeability to oxygen, cm^3 m^{-2} d^{-1} bar^{-1}
Aqueous dispersions	CBK ∗ + 15 g/m^2 PVDC	2.8 – 3.6	11 – 15
	CBK ∗ + LDPE + 20 g/m^2 PVDC	2 – 2.2	3.5 – 6.5
	PETP (12 µm) + primer + 3 – 5 g/m^2 PVDC	6 – 10	10 – 15
	PA6 (30 µm) + primer + 5 – 6 g/m^2 PVDC	6 – 11	7 – 10
	PVC (200 µm) + primer + 40 g/m^2 PVDC	0.6 – 1.1	2.6 – 3.3
	Cellulose film (37 – 48 g/m^2) + primer + 3 g/m^2 PVDC per side	5 – 8	3 – 6
Lacquer resins	Cellulose film (28 – 50 g/m^2) + primer + 1.5 g/m^2 PVDC per side	6 – 15	4 – 10
	PETP (15 µm) + 1.5 g/m^2 PVDC per side	9 – 10	12 – 14
	BOPP (20 µm) + 3 – 5 g/m^2 PVDC	3 – 5	5 – 11
Extrusion grades	EVA (20 µm)/PVDC (10 µm)/EVA (20 µm) nonoriented films		
	unplasticized PDVC	1 – 8	5 – 22
	2 % plasticizer	3 – 10	10 – 40
	4 % plasticizer	10 – 28	40 – 140
	PVDC monolayer (30 – 70 µm), 6 – 9 % plasticizer, heat shrinkable	8 – 15	30 – 200
	PVDC monolayer (15 – 25 µm), 4 – 6 % plasticizer, slightly heat shrinkable	5 – 15	15 – 60

∗ CBK = calendered bleached kraft.

THF/toluene or MEK/toluene. These solvent mixtures (for instance, THF/toluene 70/30 to 55/45) generally allow 10 – 20 % PVDC lacquers to be prepared. The required solution temperature is ca. 50 °C; the MEK/toluene mixtures must be heated to 70 – 75 °C to ensure complete dissolution.

Coating of cellulose film with PVDC is carried out by dipping on machines equipped with a vertical drying oven (Fig. 6).

To avoid blocking or sticking to the jaws of sealing machines, and to obtain the required slip properties, the following additives must be incorporated in the lacquer:

1) An additive to reduce sticking to the jaws of the sealing machine, such as fatty acids (behenic or stearic acid, for instance)
2) A natural wax (candellila, carnauba, etc.) or a synthetic wax (ketone or fatty amide) to improve blocking resistance
3) A mineral charge that roughens the surface to improve slip properties (silica, bentonite, talc, etc.)

Stabilizers are also incorporated in the solution in some cases to lower the corrosivity of the PVDC solution to metal parts of the machine.

As long as its moisture content is low, cellulose film constitutes a good barrier against gases; PVDC is thus applied to lower the water vapor permeability.

Barrier properties of cellulose film PVDC coated materials depend on the following factors:

1) Cellulose film composition
2) PVDC composition
3) PVDC coating thickness
4) Additives in the coating formula
5) Degree of crystallinity of the coating

Permeability data are listed in Table 4.

Lacquering of Paper, Foil, and Plastic Films. Some copolymers of vinylidene chloride are tailor polymerized to obtain good solubility at room temperature in a wide range of common organic solvents such as ketones and esters.

The interest of this family of product is:

1) Easier solubility
2) Heat sealability
3) Transparency and brilliance
4) Adhesion to a wide range of substrates
5) Good barrier properties and chemical resistance

These products, although less impermeable than the PVDC for cellulose film, are used for coating paper, plastic films, foils, and cardboards to lower their permeability (see Table 4) and make them heat sealable. They are also used as protective coatings. Coatings can be made by:

1) Coating on roll machines (reverse roll coating, squeeze coating, kiss coating, gravure roll coating, etc.)

2) Dip coating
3) Spray coating
4) Use of manual rollers or brushes

Depending on the final use, plasticizers, waxes, fillers, pigments, and dyes may be added to the solution.

Vinylidene chloride copolymers for coating are aqueous dispersions of copolymers with a high content of vinylidene chloride, in the form of a milky white liquid of low viscosity.

The usual comonomers are acrylic esters, unsaturated carboxylic acids, and vinyl chloride.

The PVDC aqueous dispersions are produced by emulsion polymerization. The dry solid content is 40 – 60 %.

They can be applied to papers and boards; plastics, films, and sheets; and aluminum. The nature and quantity of the comonomer is adjusted according to the substrate leading to the following coating grades of PVDC:

1) Grades for the production of nonblocking, sealable top coatings for paper, board, or paper-based laminates (paper – PE, paper – aluminum, etc.); these substrates may be pretreated, if necessary, with a grade of PVDC giving more flexible coatings and/or with a primer.
2) Grades for the production of sealable or nonsealable, nonblocking top coatings for plastic films (OPETP, BOPP, PE, rigid PVC, PA6, OPA, etc.). The films are coated with a primer or a first layer of PVDC (flexible grade) before coating with the nonblocking top coat.
3) Grades giving coatings that are more flexible, but also slightly more blocking and more permeable, than those made from the above grades. They are used mainly to coat paper, board, and paper-based laminates (with primer if necessary), and occasionally heavy-gauge plastic films with a primer.
4) Grades giving special coatings that adhere directly to certain plastic films without a primer. Their principal application is as an undercoat for subsequent coatings of another PVDC or PE applied by extrusion coating.
5) Grades for special coating or impregnation applications.

Although PVDC dispersions are in most cases supplied ready for use, under certain circumstances the user may need to introduce one or more additives to the aqueous dispersions prior to processing to modify the properties of the coating (surface characteristics, opacity, pigmentation, etc.) or the behavior of the product when applied by a coating or impregnation machine (corrosiveness, foaming, adjustment of coating weight, etc.).

PVDC dispersions are applied in roll-to-roll operations (web coating) in widths usually not exceeding 2500 mm. Processing speed has increased considerably in the last few years to 100 – 400 m/min.

The barrier properties of PVDC coatings depend on

1) Nature of PVDC (comonomers)
2) Quality of the substrate (porosity)
3) Coating weight and number of layers
4) Flexibility of PVDC and the substrate
5) Drying conditions
6) Crystallinity

Permeability data are listed in Table 4.

4.4. Uses

Extrudable grades are used for:

1) Monofilaments for the production of military combat boots and insoles, draperies and curtains, and filter applications
2) Medical applications, e.g., colostomy bags are made of monolayer or multilayer structures (EVA – PVDC – EVA)
3) Food packaging:
   Cling films, which are very thin films, slightly adhesive, used for wrapping foodstuffs, are monolayer films
   Shrink multilayer films (EVA – PVDC – EVA) are used to transport and store fresh cuts of meat between the slaughterhouse and the point of sale
   Some synthetic sausage skins are made of monolayer PVDC
   Rigid sheets (PP – PVDC – PP) are stable towards sterilization process and give shelf-stable storage

Aqueous dispersions are used for:

1) Pharmaceutical packaging blister packs for drugs which are very sensitive to water are made of PVC coated with PVDC.

2) Plastic food packaging:
   Biscuits are often packed in PVDC-coated OPP. Delicate chips or peanuts are packed in PVDC-coated PET laminated with LDPE. Cellulose casing for sausages is often coated with PVDC.
3) PVDC coated paper is often used for biscuits, cereals, and dried food.
4) Technical applications:
   PVDC is generally used for its water-resistance, fireproof or binder properties in applications such as coating carpet backings, wood chip panels, and cork coating.

Soluble grades are generally used for coating cellophane, and in some cases for plastic films (e.g., PETP).

5. Economic Aspects

PVDC copolymers as they arise from the polymerization are marketed by the raw material manufacturer as dispersions or powders under various names.

The best known trade names and producers are:

Amsco-res, Unocal (aqueous dispersions)
Daran, W. R. Grace (aqueous dispersions)
Diofan, BASF (aqueous dispersions)
Ixan, Solvay (aqueous dispersions, extrusion grades, lacquer resins)
Saran, Dow Chemical (extrusion grades, lacquer resins, PVDC films)
Serfene, Morton (aqueous dispersions)
Viclan, ICI (aqueous dispersions, lacquer resins, PVDC films)
Kureha (aqueous dispersions, PVDC films)
Asahi (aqueous dispersions, lacquer resins, PVDC films)

The PVDC world market (1990) is estimated at 125 000 t:

By countries:		By grades:	
Europe	26 %	Aqueous dispersions	32 %
Japan	38 %	Lacquer resins	6 %
United States	34 %	Extrusion grades	62 %
Others	2 %		

6. Toxicology and Occupational Health [44], [45]

Monomeric VDC is highly volatile. The vapor is denser than air and detectable by smell only at concentrations of 500 – 1000 ppm. When handling VDC the working area must be well ventilated. VDC vapors are highly inflammable and burn fiercely in air in the concentration range 7 – 16 vol %.

Animal tests and epidemiological investigations have not so far proved that VDC has a specific carcinogenic effect [46]. The mean VDC concentration in the workplace has been limited in the United States to 5 ppm (TLV) and in Germany to 2 ppm (MAK).

PVDC is mainly used for packaging foodstuffs or pharmaceutical products, involving both direct and indirect contact with the contents. Such packaging is governed by highly stringent regulations at an international level.

These regulations are generally based on positive lists with specific migration limits. PVDC satisfies all of these regulations.

Residual vinylidene chloride monomer in commercial polymers generally ranges from indetectable up to a few ppm since all modern producers have now adapted processes to outgas PVDC. These quantities are even lower when measured on coated or extruded materials.

In the United Kingdom, information on the content of VDC in food sold in retail outlets was collected by the MAFF [47]. Of the foods analysed, only snack foods and foods contained in VDC – VC copolymer-coated packs were found to contain measurable amounts of VDC. In most other cases of food packed in barrier-coated materials (polymer films and paper), the VDC content was below the detection limit of 0.03 – 0.05 ppm. From all the information available, the maximum possible intake of VDC per person, when averaged over the entire U.K. population, would be no more than 1 µg/d, in the worst-case assumption that all the residual VDC in food packaging was in fact ingested [47].

7. References

General References

1. R. A. Wessling, D. S. Gibbs, P. T. Delassus: "Vinylidene Chloride Polymers," *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **Vol. 17,** Wiley, New York 1987.

Specific References

2. V. Regnault, *Ann. Chim. Phys.* **69** (1838) 151.
3. H. Staudinger, W. Feist, *Helv. Chim. Acta* **13** (1930) 805.
4. H. Staudinger, W. Feist, *Chem. Ind.* **45** (1947) 685.
5. Dow, US 2 160 931, 1939.
6. R. C. Reinhardt, *Ind. Eng. Chem.* **35** (1943) 422.
7. J. J. P. Staudinger, *Br. Plast.* **20** (1947) 381.
8. C. E. Schildknecht: *Vinyl and Related Polymers,* J. Wiley and Sons, New York 1952, Chap. VIII.
9. J. F. Gabett, W. M. Smith in G. E. Ham: *Copolymerisation,* Interscience, New York 1964, Chap. X.
10. G. Talamini, E. Peggion in G. E. Ham: *Vinyl Polymerization,* Decker, New York 1967, Chap. V.
11. R. A. Wessling, F. G. Edwards, *Encycl. Polym. Sci. Technol.* **14** (1971) 540.
12. R. A. Wessling: *Polyvinylidene Chloride,* Gordon and Breach, New York 1977.
13. K. Mayumi et al., CA 53: 5225 a.
14. Grace, US 3 033 812, 1962.
15. Scott Bader, US 3 291 769, 1966. Scott Bader, GB 1 100 398, 1968.
16. D. M. Woodford, *Chem. Ind.* 1966, 316.
17. Dow, US 3 617 368, 1971.
18. Grace, US 3 317 449, 1967.
19. Dow, US 2 956 047, 1960.
20. R. A. Wessling, D. S. Gibbs, *J. Macromol. Sci. Chem. A* **7** (1973) no. 3, 647.
21. L. Marker et al., *J. Polym. Sci.* **57** (1962) 855.
22. Dow, US 3 879 359, 1975.
23. ICI, GB 1 338 561, 1973.
24. Chemische Werke Hüls, DE 1 919 705, 1970.
25. Eastman Kodak, US 3 379 665, 1968.
26. BASF, GB 1 039 383, 1966.
27. Du Pont, US 3 002 956, 1961.
28. Dow, US 3 662 028, 1972.
29. Grace, US 3 088 037, 1963.
30. Dow, US 2 968 651, 1961.
31. Dow, US 3 642 743, 1972.
32. Dow, US 2 482 771, 1944.
33. Kureha Chem. Ind., GB 1 040 826, 1966; GB 1 111 087, 1968.
34. Kureha, DE 2 244 957, 1973.
35. ICI, GB 1 171 245, 1971.
36. ICI, GB 1 162 576, 1969.
37. M. P. Farr, I. R. Harrison: "A study of Thermal Degradation Kinetics of H – H Poly(Vinylidene Chloride)," *J. Polym. Sci. Part C* **24** (1986) 257 – 261.
38. J. Brandrup, E. H. Immergut: *Polymer Handbook,* Interscience, New York 1966.
39. I. M. Mohr, D. R. Paul: "Comparison of Gas Permeation in Vinyl and Vinylidene Polymers," *J. Appl. Polym. Sci.* **42** (1991) 1711 – 1720.
40. Y. Vandendael: "Barrier Polymers: A Complex, Rapidly Developing Field with a Wealth of Possibili- ties and a Promising Future," Biotechnology International Century Press, London 1990, pp. 311 – 315.
41. J. P. Golstein: "New Trends in Barrier PVDC Copolymers," Belgian Technical Seminar organized by Belgian Office of Foreign Trade, Tokyo, November 19, 1991.
42. Y. Vandendael: "What is Barrier Packaging?", lecture given at Mobil's Pack-It Basel Seminar, March 20th, 1992.
43. C. Michel: "Developments in PVDC Barrier Resin Technology," Pira International Conference, London, April 2nd – 3rd, 1987.
44. European Chemical Industry Ecology and Toxicology Centre, Ecetoc Report no. 5, Vinylidene Chloride, 14 August 1985.
45. International Program on Chemical Safety (I.P.C.S.), Environmental Health Criteria 100 Vinylidene Chloride, World Health Organisation, Geneva 1990.
46. A. Collard, *Kunstst. Plast. (Solothurn, Switz.)* **25** (1978) 40.
47. Ministry for Agriculture, Fisheries and Food (MAFF), Food Suveillance Paper 3, HMSO 1980.

Poly(Vinylidene Fluoride) → Fluoropolymers, Organic
Poly(Vinylpyrrolidone) → Polyvinyl Compounds, Others
Polyxylenes → Specialty Plastics
Porcelain → Whitewares
Porcelain Enamel → Ceramics, Ceramic-Metal Systems

Poromerics → Leather Imitates
Positive Inotropic Agents → Cardiac Glycosides and Synthetic Cardiotonic Drugs
Potash → Potassium Compounds

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Potassium and Potassium Alloys

Standard Article

Elizabeth R. Burkhardt[1]

[1]BASF Corporation, Evans City, PA 16033, USA

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a22_031.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Potassium
1.1. Properties
1.2. Production
1.3. Handling and Occupational Safety
2. Potassium Alloys
2.1. Properties of NaK Alloy
2.2. Production of NaK Alloy
2.3 Handling and Occupational Safety of NaK Alloy
3. Uses of Potassium and Potassium Compounds
3.1. Potassium and NaK Alloy
3.2. Potassium Hydride
3.3. Potassium Bases
3.3.1 Potassium Alkoxides
3.3.2 Potassium Amides
3.4. Potassium Superoxide
4. Acknowledgement

Potassium is the third element in the alkali metal group. This highly reactive metal is used to make potassium superoxide for use in self-contained breathing apparatus. Potassium in the form of alloys with sodium is used in industrial synthesis of alkylated aromatic hydrocarbons such as isobutylbenzene. Compounds of potassium derived from alcohols and amines are routinely used as bases in pharmaceutical and agricultural synthesis.

[Top of Page]

1. Potassium

HUMPHREY DAVY discovered potassium metal in 1807 upon electrolysis of molten potash (KOH) [1]. The English (and French) name for the element potassium is derived from the word “Potash”. The elemental symbol K (and the German name *Kalium*) comes from the Arabic *al kalzum*, meaning ashes.

Potassium [7440-09-7], atomic no. 19, is included in the group 1 elements comprising the alkali-metal family [2] and has the electron configuration $1s^2 2s^2 p^6 3s^2 p^6 4s^1$. All alkali metals carry a single, unpaired electron in the outermost electron shell. The alkali metals are the most electropositive elements known because of their great tendency to attain an inert gas structure by donation of the lone electron. This unique property renders all alkali metals strong reducing agents and good electron donors.

Potassium has three natural isotopes: ^{39}K (93.26 %), ^{40}K (0.012 %), and ^{41}K (6.73 %) [3]. The isotope ^{40}K, with a half-life of 1.25×10^9 years, is useful in determining geological age, disintegrating by electron capture to 89.33 ± 0.11 % ^{40}Ca and 10.67 ± 0.11 % ^{40}Ar. Several artificially produced isotopes of potassium are also known.

Potassium, the seventh most abundant element, does not occur in nature as a free metal. The abundance of potassium in the earth's crust is 2.59 %, mainly in silicates (feldspars, micas). The average content of potassium in seawater is 0.039 wt

% due to the weathering of minerals and varies depending on the distance from a river mouth [4].

Potassium minerals [5] are mined from salt deposits formed by crystallization and/or evaporation of marine regions during various stages of the earth's evolution. The most important potassium minerals are sylvite, sylvinite, carnallite, and langbeinite. Sylvinite is primarily potassium chloride (KCl, sylvite [6]) and sodium chloride (NaCl, halite) containing 20 – 30 % K calculated as potassium oxide (K_2O). The largest potassium chloride deposit is located in Saskatchewan (Canada). Langbeinite [7] is mainly potassium sulfate and magnesium sulfate with about 23 % K as K_2O. The largest deposit of carnallite, the double salt $KCl \cdot MgCl_2 \cdot 6\ H_2O$, is in Solikamsk (Russia). KCl can also be obtained by evaporation from the brines of Searles Lake in California, the Great Salt Lake in Utah, and the Dead Sea in the Middle East. Potassium nitrate is mined in Chile. Caustic potash (KOH), another important source of potassium, is primarily mined in Germany, New Mexico, California and Utah.

Potassium is indispensable for plants since it plays an important role in carbohydrate metabolism and the swelling of cells. Many plants accumulate 1 – 2 % potassium to grow at a normal rate. Therefore, KCl is widely used as a fertilizer to increase crop yields. In animal organisms, K^+ ions stimulate glycolysis, lipolysis, respiration of tissue, and the synthesis of proteins and acetylcholine. Potassium and sodium levels in higher animals are regulated by the kidneys.

1.1. Properties

Physical Properties. Potassium is a very soft, silvery metal with low density and melting point and very high thermal and electrical conductivities. The physical properties are summarized below [8].

Z	19
A_r	39.098
Crystal structure	body-centered cubic
Lattice parameter (17.3 °C)	0.531 nm
mp	63.25 °C
bp	759.9 °C
Atomic radius	277 pm
Ionic radius	151 pm
Ionization potential [9]	4.3407 V
Work function	2.24 eV
Standard reduction potential [10]	– 2.924 V
Electron emission threshold	0.65 μm
Electrical resistance	
0 °C	6.1 μΩ · cm
liquid, 63.2 °C	13.31 μΩ · cm
Density, 20.0 °C	0.856 g/cm^3
liquid, 63.2 °C	0.82 g/cm^3
300.0 °C	0.771 g/cm^3
Thermal conductivity	
21.0 °C	9.63×10^{-3} W m^{-1} K^{-1}
400.0 °C	4.00×10^{-3} W m^{-1} K^{-1}
Specific heat	
0.0 °C	0.737 J g^{-1} K^{-1}
50 – 100.0 °C	0.837 J g^{-1} K^{-1}
Heat capacity at 25 °C [11]	29.522 J g^{-1} K^{-1}
Latent heat of fusion	59.71 J g^{-1}
Latent heat of vaporization	2043 J g^{-1}
Hardness (Mohs scale)	0.5
Spectral lines	769.9, 766.5, 404.5 nm

Vapors of potassium metal remain diatomic (K_2) until the required dissociation energy of 49.6 ± 0.4 kJ (at 0 °C and 1.013 bar) is attained, causing the metal – metal bond to rupture: K_2 2 K. Potassium's vapor pressure is 164 mbar at 570 – 600 °C, and 521 mbar at 685 –700 °C. For the temperature dependence of the vapor pressure (vapor pressure equation), see [12]; for the vapor pressure curve, see [13].

Inelastic X-ray scattering methods were used to measure the high-frequency collective dynamics of molten potassium [14].

Chemical Properties. Potassium shares many common features with other members of the alkali metal group (Li, Na, Rb,

and Cs). Electropositivity increases from lithium to cesium. However, rubidium and cesium are too reactive for safe handling and they are not readily available on an industrial scale. Thus, potassium is the most electropositive and powerful industrial reducing agent from a practical and economical standpoint.

Although K, Na, and Li share certain attributes, each possesses unique chemical characteristics as well. Significant differences in ionic size and ionization energy lead to the distinctive chemical traits of each individual alkali metal. Potassium is the largest cation of the three, so organopotassium compounds are generally the most ionic. In polar solvents, potassium cations form loose ion pairs (or solvent-separated ion pairs) with their counteranions, whereas organolithium compounds tend to exist as aggregates (tetramers and hexamers), since lithium cations form tight ion pairs.

The smaller lithium cation is better suited to stabilizing localized (hard) carbanions, while the potassium cation stabilizes delocalized (softer) carbanions. Accordingly, the chemistry of sodium lies between those of potassium and lithium. Potassium's high reducing potential is demonstrated by the facile displacement of hydrogen atoms in weakly acidic hydrocarbons, amines, and alcohols (Table 1).

Table 1. Acidic compounds and their potassium derivatives

Compound	pK_a*	Potassium derivatives
1,3-Diaminopropane	35	$KNH(CH_2)_3NH_2$
Ammonia	35	KNH_2
Aniline	27	$KNHC_6H_5$
Acetylene	25	KC≡CH
tert-Butyl alcohol	18	$KOC(CH_3)_3$
Methanol	16	$KOCH_3$
Phenol	10	KOC_6H_5

*The expression pK_a refers to the pK_a of the conjugate acid from which the potassium base is formed.

The strong reducing potential of potassium metal is not limited to compounds containing active hydrogen. Potassium also reacts readily with various gaseous elements and inorganic substances (Table 2).

Table 2. Miscellaneous reactions of potassium

Element, compound	Conditions	Products
H_2	200 – 411 °C	KH
O_2	room temp.	KO_2
	200 – 300 °C	K_2O_2
X_2 (halogen)	room temp.	KX
S	70 °C	K_2S
H_2S	room temp.	K_2S
MX_n (metal salt)	various	M (metal)
C (graphite)	150 – 400 °C	C_8K
		$C_{24}K$, $C_{36}K$, etc.

Potassium forms three oxides: potassium monoxide, K_2O; the peroxide, K_2O_2; and the superoxide, KO_2. In dry air or dry oxygen, metallic potassium burns completely, forming potassium superoxide (KO_2). The propensity to form the superoxide resembles that of rubidium and cesium more than sodium. Under comparable oxidation conditions, sodium yields Na_2O_2, while lithium gives Li_2O.

Potassium does not react with saturated hydrocarbons, but potassium superoxide may explode in the presence of hydrocarbons [15]. Consequently, potassium is usually stored under inert gas because potassium can react with entrapped air in oils to form potassium superoxide. The metal has been used as a dispersion [16] in hydrocarbons, and is used in this form for chemical reactions. For many organic reactions, potassium is most easily handled as a liquid alloy with sodium.

Potassium metal is less dense than water and will float while reacting extremely rapidly with the water to form a colorless solution of potassium hydroxide (KOH) and hydrogen gas (H_2). The evolving hydrogen gas can explode violently. The

potassium metal melts due to the exothermic reaction then catches fire and burns with a characteristic pale lilac color. The resulting aqueous solution is caustic because of the dissolved potassium hydroxide.

Like other alkali metals, considerable amounts of potassium dissolve in liquid ammonia (ca. 480 g of potassium in 1000 g of ammonia at 0 °C). Depending on the concentration, potassium – ammonia solutions are blue to copper-colored, with electrical conductivities approaching those of liquid metals [17]. These special solvation characteristics have been the subject of comprehensive investigations [18]. Potassium dissolved in liquid ammonia reacts very slowly to form the amide (KNH_2). Transition-metal salts greatly accelerate this process. The transformation of gaseous ammonia into potassium amide proceeds rapidly at 300 °C.

Potassium metal along with a crown ether (18-crown-6 or 15-crown-5) in THF forms an alkalide composed of $K^-K^+(crown)_n$ that can be used in deprotonation reactions to generate ϒ-lactone enolates [19]. The alkalide with 15-crown-5 was shown to be more stable [20]. Alkalides of NaK have been studied recently and are presumed to exist as a K^+Na^-(15-crown-5) ionic pair [21].

Potassium reduces oxides, hydroxides, and salts of many heavy metals to the metallic state, and at high temperature it decomposes silicates (e.g., glass, porcelain, and mica).

Potassium and cesium react violently with Teflon equipment even under inert atmosphere; therefore, Teflon-coated stirbars or gaskets in equipment or valves should not be allowed to contact the metal.

1.2. Production

Degussa's electrolytic process for producing relatively pure potassium metal in the Castner cell was perfected in the late 1920s. This method employed an electrolyte containing 66 % KOH, 19 % K_2CO_3, and 15 % KCl. Later, a salt mixture (*mp* ca. 600 °C) of potassium chloride and potassium fluoride containing a small percentage of potassium carbonate was utilized [22]. Electrolysis of molten salt by the Downs process (1921) was developed into an industrial process for the manufacture of sodium, but corresponding efforts in the case of potassium have been unsuccessful.

In the industrial production of potassium metal, a thermal process is preferred over electrolysis because potassium reacts with the carbon electrodes in electrolysis.

$$KCl + Na \rightarrow K + NaCl$$

Reduction of potassium chloride with sodium metal was introduced in the 1950s by Mine Safety Appliance Company [23]. This process can produce from the same installation either high-purity potassium metal or potassium – sodium alloys of specified compositions. Sodium metal and molten potassium chloride are combined and a metal – vapor mixture of sodium and potassium is fractionated with a column. Sodium chloride produced from the reaction is drained off. Because potassium vapor is continuously withdrawn from the reaction zone, the equilibrium is shifted to the thermodynamically unfavorable potassium side.

The Griesheimer process, which was commonly employed in the past, is thermodynamically favorable starting from potassium fluoride and calcium carbide [24]. This thermal process is used in Baroda, Gujarat for consumption of potassium in the Heavy Water Plant operated by the Department of Atomic Energy in India. The endothermic reaction is presumably more complex than suggested by the simple chemical equation:

$$2\,KF + CaC_2 \rightarrow 2\,K + CaF_2 + 2\,C$$

Other thermal processes, in which potassium carbonate, caustic alkali, potassium fluoride, and potassium sulfate are treated with carbon, iron, silicon, or other reducing agents, have no commercial importance [25], [26], [27], [28], [29].

Potassium is commercially available with a purity of greater than 99.0 % [30] but very little is shipped as solid potassium metal. The preferred industrial compound is the alloy of sodium and potassium.

1.3. Handling and Occupational Safety

Potassium is packaged in steel or stainless steel drums and cylinders under a nitrogen atmosphere. It is classified as a substance which in contact with water, emits flammable gases. Transport regulations for potassium are as follows: [31]

UN No.	2257
Classification	4.3, PG I

Potassium is hazardous due to its rapid reaction with moisture in mucous membranes and skin. The potassium hydroxide formed on water contact is severely corrosive and can cause severe burns. When handling potassium, gloves and a nonflammable apron (leather is best) should be worn along with protective goggles and face shield. Where the potential exists for airborne exposure to potassium or potassium combustion products, an approved respirator equipped with particulate filters should be worn.

Potassium is highly flammable and reacts with water or moist air to form potassium hydroxide and hydrogen. The high heat of reaction causes the potassium to melt and spray, and it also ignites the released hydrogen. Because burning potassium is difficult to extinguish, conventional fire extinguishers are unsuitable. Such fires should be extinguished with a suitable dry powder mixture, dry sodium chloride, potassium chloride, water-free soda, dry sand, cement or Ansul's Met-L-X [32]. The use of foam extinguishers, carbon dioxide extinguishers, and halogenated hydrocarbons (e.g., carbon

tetrachloride) should be avoided.

Potassium should be stored in an inert atmosphere (argon or nitrogen) or under a high-boiling hydrocarbon (for limited storage periods). Small quantities can be kept in glass bottles under an inert atmosphere. Potassium stored under petroleum or other hydrocarbons can form the peroxide and superoxide at room temperature, which may provoke a massive explosion when the corresponding incrustation is removed (as by cutting). Potassium incrusted with a yellow oxide or peroxide layer should be destroyed in small quantities by careful, controlled disposal [33]. Small amounts of potassium residues (in containers, for example) may be dissolved in *tert*-butanol or other high-boiling alcohols [34]. The resulting alkoxide is destroyed by careful addition of methanol followed by water. The aqueous caustic potassium waste is then discarded in accordance with local regulations.

[Top of Page]

2. Potassium Alloys

Potassium forms alloys and, in some cases, intermetallic compounds with several metals (e.g., Na, Zn, Cd, Hg, In, Tl, Sn, Pb, and Bi). These alloys and intermetallic compounds are reactive substances, but are generally less reactive than free potassium metal. Except for alloys of sodium and potassium, intermetallic compounds are also less ductile than pure potassium metal.

When combined with either rubidium or cesium, potassium forms solid solutions over the entire range of binary compositions. The formation of solid solutions between potassium and metals from other groups in the Periodic Table is limited.

Liquid potassium amalgam with a low potassium concentration (ca. 0.3 % K) forms during the electrolysis of aqueous potassium chloride solutions in mercury cells.

Various alloys in the potassium – sodium system (Fig. 1) have achieved commercial importance, especially the liquid alloy containing ca. 78 % potassium (called “NaK alloy” or “nack” in the USA). Potassium – sodium alloys are liquid over a wide range of composition (ca. 40 – 90 wt % of potassium).

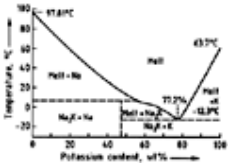

Figure 1. Sodium – potassium phase diagram [35]

2.1. Properties of NaK Alloy

The eutectic composition, with a eutectic temperature of –12.6 °C, contains 77.8 wt % K (corresponding to 66 atom % K). Na_2K is the only intermetallic compound existing with certainty in the sodium – potassium system (peritectic temperature 6.90 °C). The solubility of potassium in sodium at temperatures above 7 °C is given by the expression (57.1 – 0.586 *t*) wt %, where *t* is the temperature in °C [36]. Other physical properties follow: [37]

Composition	77.8 wt % K, 22.2 wt % Na
mp	–12.6 °C
bp (estimate)	785 °C
Electrical resistance	
at *mp*	33.5 μΩ · cm
at 20 °C	38.0 μΩ · cm
Density	
at 20 °C	0.867 g/cm^3
at 100 °C	0.855 g/cm^3
at 550 °C	0.749 g/cm^3
Thermal conductivity	
at 20 °C	2.18×10^{-3} W m^{-1} K^{-1}
at 550 °C	2.62×10^{-3} W m^{-1} K^{-1}
Specific heat	
at *mp*	0.975 J g^{-1} K^{-1}
at 100 °C	0.937 J g^{-1} K^{-1}

Trace impurities found in NaK alloys are H, C, N, O, OH, Cl, Ca, and Fe (together with its alloying partners in steels), as well as lead and other heavy metals. Metallic impurities are mainly introduced during the blending of potassium with sodium containing impurities. Purification of NaK alloys is performed (as with potassium) in a closed purification installation involving prepurification and subsequent distillation [38], [39]

2.2. Production of NaK Alloy

Alloys of potassium and sodium can be prepared by several methods:

1. Pure potassium and sodium metals can be melted together in an inert atmosphere or under an inert hydrocarbon
2. Various compositions of NaK can be prepared continuously from potassium chloride and sodium metal
3. An electrolytic process in a Castner crucible is a viable alternative for producing the alloy, in contrast to pure potassium metal

2.3. Handling and Occupational Safety of NaK Alloy

Commercial potassium – sodium alloys (see Potassium Alloys) are shipped in airtight steel cylinders equipped with internal piping and valves. In the United States, it is shipped in stainless steel or carbon steel containers containing 0.5, 14, 75, 84, 90, or 340 kg. Transport regulations for potassium – sodium alloys are as follows [40]:

Physical state	liquid at room temperature
UN No.	1422
Physical state	solid
UN No.	3404
Classification	4.3 PG I

Potassium – sodium alloys are highly flammable and react with water or moist air to form potassium hydroxide and hydrogen, which may ignite. The dangers are increased because spilled metal may flow toward reactive materials since these alloys are liquid at ambient temperature. The technology for handling NaK on a large scale from ambient temperature to ca. 825 °C is well developed, and equipment is available commercially for pumping and metering NaK safely and accurately on any scale. Handling information and a safety and handling video are available from BASF Corporation.

Potassium – sodium alloys rapidly react with moisture in mucous membranes and skin. The potassium hydroxide formed upon water contact is highly irritating and leads to burns and skin cauterization. When handling cylinders containing potassium – sodium alloy, gloves and a nonflammable apron (leather is best) should be worn along with protective goggles and face shield. Where the potential exists for airborne exposure to the alloy or combustion products, an approved respirator equipped with appropriate filters should be worn.

NaK fires should be extinguished with a suitable dry powder mixture, dry sodium chloride, potassium chloride, water-free soda, dry sand, cement, or Ansul's Met-L-X [41]. The use of foam extinguishers, carbon dioxide extinguishers, and halogenated extinguishing agents (e.g., carbon tetrachloride) should be avoided.

Potassium – sodium alloys should be stored in an unbreakable container under an inert atmosphere (argon or nitrogen). It is not recommended to store potassium – sodium alloy under petroleum or other hydrocarbons, since potassium superoxide can form from dissolved oxygen. Potassium superoxide is insoluble in NaK and may provoke an explosion when physically manipulated in contact with organic material.

[Top of Page]

3. Uses of Potassium and Potassium Compounds

The understanding of potassium chemistry has been greatly expanded over the last 40 – 50 years [42], [43], [44], [45], [46], [47], [48] and this has led to numerous applications of potassium compounds. Potassium, sodium – potassium alloys, and potassium compounds (alkoxides, amides, and hydride) are extensively used in academia and industrially to synthesize both organic and inorganic compounds.

3.1. Potassium and NaK Alloy

In general, organopotassium compounds, i.e., those with a potassium – carbon bond, are not thermally stable and so must be formed at subambient temperature and used immediately [49]. Potassium vapor was used for doping of single-walled carbon nanotubes [50] and superfulleride thin films [51]. These materials have been studied for their electronic properties as quantum dots. Potassium-doped carbon nanotubes were also found to absorb hydrogen at ambient temperature and release it at higher temperature [52]. These compounds may find use for hydrogen storage, since the sorption – desorption cycle can be repeated.

Potassium – Sodium Alloys (NaK). On an industrial scale, the chemistry of potassium is most easily conducted by using the liquid alloy of sodium and potassium. Sodium – potassium alloys are used for performing reactions of potassium metal in organic synthesis. For chemical reactions involving unstable intermediates such as carbanions and free radicals, NaK is preferred to the individual metals due to its liquid state down to –12 °C. Additionally, NaK alloy is used in applications that exploit its unique physical properties.

The advantages of NaK compared to Na or K are:

- Higher reactivity than solid alkali metals at ambient temperature
- Higher surface area of reactive metal than either sodium or potassium
- Elimination of the need to produce and handle reactive metal dispersions

– Easily applied to solid supports

Physical Applications. In addition to easier handling, the unique physical properties of NaK led to its use as a heat-exchange fluid in nuclear reactors [53], [54], [55], [56]. Although the market for NaK in the nuclear industry is insignificant, the favorable thermal properties are still being exploited in other heat-exchange applications [57], such as a cooling liquid in hollow valve stems [58] and in solar power generators [59]. The electrical conductivity and the wide range of compositions in the liquid state render the NaK alloys ideal for use in nonaqueous batteries, [60] high-temperature thermostats, and as a hydraulic fluid [61].

An interesting application of NaK may involve both its reduction and alloying capacities; the treatment of brass plumbing fixtures with NaK removes Pb from the surface and decreases the leaching of Pb into water during use [62].

Chemical Applications. Sodium – potassium alloy is used in a number of industrial applications, such as alkylation, isomerization, condensation, and transesterification reactions. One of the largest commercial uses of NaK as a catalyst is for side-chain alkylation of toluene and xylenes [63]. An example is the side-chain alkylation of toluene to form isobutylbenzene, a starting material for the synthesis of ibuprofen. The synergistic effect of the two alkali metals has been studied [64].

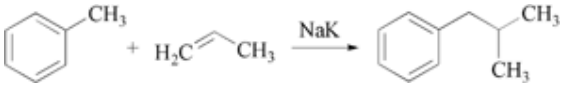

Side-chain alkylation of xylene followed by isomerization and oxidation yields naphthalenedicarboxylate [65], a specialty monomer. In these applications, the reactivity of the liquid metal is increased by dispersion on a solid support.

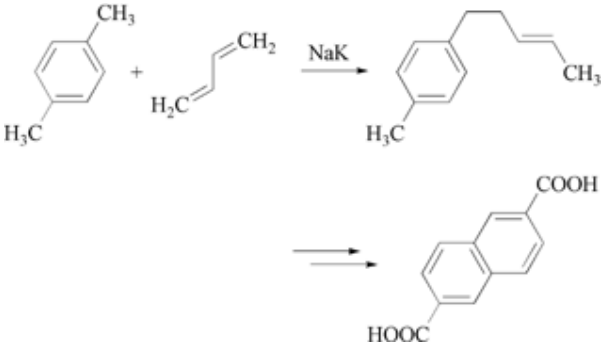

NaK has also been used extensively for isomerization of α-olefins to internal olefins [66].

Industrially, this process is used to isomerize norbornene for use in the synthesis of EPDM rubbers.

NaK alloys have found applications in dehydrogenation [67], free-radical [68], condensation [69], [70], [71], [72], and polymerization [73] reactions. Polymerization of alkylsilicon trichlorides with NaK generates Si – Si bonded materials [74]. Additionally, NaK has been used for the synthesis of polycarbynes [75], in the derivatization of poly(isobutylene-*tert*-butyl methacrylate) block copolymers [76] and as initiator for anionic polymerization [77].

As a reducing agent, NaK can be used to reduce metal halides to highly reactive metal powders [78]. Potassium complexes with crown ethers have also been used in metal-salt reductions [79]. For example, lanthanide nanoparticles have been prepared by reduction of the salts with NaK(crown ether) complexes [80]. Synthetic utility of NaK includes reduction of organic functional groups such as arenes, ketones, aldehydes, and alkyl halides [81], [82] and reductive cleavage of functional groups containing C–X, C–O, and O–S bonds [83], [84], [85], [86], [87].

The complete dechlorination of polyhalogenated hydrocarbons is a difficult process but extremely important for the environmental remediation of contaminated waste and soil. Hexachlorobenzene was efficiently dechlorinated at nearly 100 % conversion in 30 min with NaK alloy due to its high reactivity [88].

Since NaK is a liquid at room temperature it has been used as an efficient impurity scavenger in both organic liquids and gases. Examples are the removal of acetylenic and allenic contaminants [89] and purifying gas streams for semiconductor processes [90]. This application for gas purification is no longer recommended because of the explosion [91] hazards due to KO_2 formation.

An interesting use of NaK for a consumer product is the interesterification of vegetable oil to margarine [92] and interesterification of tallow with cotton seed oil to improve the pour and clarification temperatures [93].

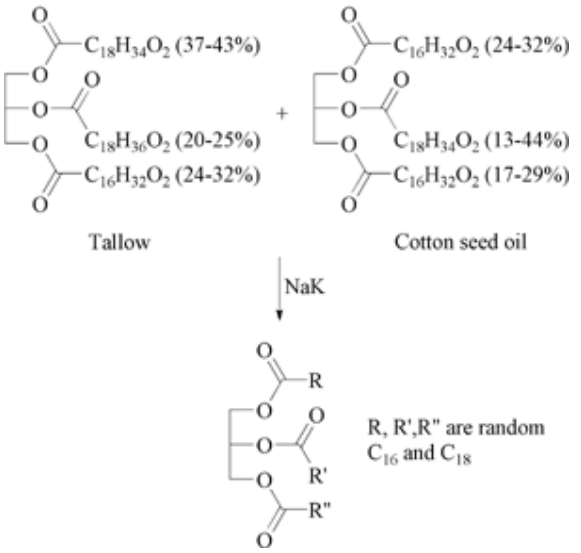

Potassium Graphite (C_8K). Potassium, rubidium, and cesium react with graphite (⟶ Carbon) and activated carbon to form compounds with the following formulas: C_8M, $C_{24}M$, $C_{36}M$, $C_{48}M$, and $C_{60}M$ [94]. C_8K corresponds to all carbon layers separated by a layer of potassium. In $C_{24}K$, $C_{36}K$, etc., intercalation occurs in each second, third, fourth, etc., interlamellar graphite spacing. At these higher stages of intercalation the structure is less ordered than that of C_8K (⟶ Carbon – Graphite Intercalation Compounds). Potassium graphite [12081-88-8] (C_8K) in the form of bronze-colored flakes is prepared by melting potassium over graphite at 120 – 150 °C under an inert atmosphere.

Potassium graphite is a powerful solid reducing agent because the free electrons from the ionized potassium are trapped within the graphite lattice. In packed columns or beds, it can effectively eliminate undesirable organic contaminants such as halogenated hydrocarbon impurities from gas streams and liquid solutions. The scope of its reducing reactions include [94], [95] Birch-type reactions of α, β-unsaturated ketones, carboxylic acids, and Schiff's bases; reductive cleavage of vinylic and allylic sulfones; selective alkylation of aliphatic esters, imines, and nitriles; reductive decyanation of nitriles; and reduction of glycosyl halides to glycols [96].

Potassium graphite has also been used to prepare active metals supported on graphite (Gr) by reduction of the metal salt in THF: [94], [97] Zn – Gr [69704-06-9] for Reformatsky reactions, Sn – Gr [89248-61-3] for allylic organometallics, Ti – Gr [6970-05-8] for coupling carbonyl compounds, Fe – Gr [55957-21-6] for debromination, Pd – Gr [59873-73-3] for vinylic and allylic substitution, and Pd – Gr, Ni – Gr [59873-69-7] for hydrogenation. Potassium graphite has been used to prepare chromium carbenes [98] and has also been applied in polymerization reactions [99]. Potassium graphite or similar intercalation compounds may have applications in pyrotechnic compositions [100].

Potassium on Alumina and Silica Support. Finely dispersed potassium on aluminum oxide has been used in double-bond isomerizations and as a metalating agent [101]. Methods to apply and use alkali metals on silica support have been studied [102].

3.2. Potassium Hydride

Potassium hydride [7693-26-7], KH, is made from potassium metal and hydrogen gas at ca. 200 °C in an oil dispersion or at lower temperatures with a catalyst [103] (⟶ Hydrides – Potassium Hydride). The principal supplier is Pressure Chemical Company, Pittsburgh, PA (USA).

The synthesis [104] and use of KH as a base and as a nucleophile is reviewed elsewhere [105]. Compared to NaH or LiH, KH [106], [107] is an effective reagent for ketone enolization reactions, although soluble potassium bases (such as potassium hexamethyldisilazane) are the best choice for kinetic deprotonation. Potassium hydride has been used as an initiator for anionic polymerization of oxiranes [108]. A practical cycling hydrogen storage medium was prepared by addition of KH to a mixture of NaH and $NaAlH_4$ doped with Ti [109].

A noteworthy development is conversion of the basic character of potassium hydride to a hydridic character by complexing with alkoxyboranes and alkylboranes. The potassium borohydride complexes formed vary from mild to strong reducing agents and are used primarily in the pharmaceutical industry. Potassium triethylborohydride [22560-21-0], $KB[C_2H_5]_3H$, potassium tri-*sec*-butylborohydride [54575-50-7], $KB[CH(CH_3)C_2H_5]_3H$, and potassium trisamylborohydride [67966-25-0], $KB[CH(CH_3)CH(CH_3)_2]_3H$, are made by the reaction of KH with triethylborane [97-94-4], tri-*sec*-butylborane [1113-78-6], and trisamylborane [32327-52-9], respectively, in THF. Potassium triisopropoxyborohydride [42278-67-1] can be made by the reaction of KH with isopropyl borate [5419-55-6] [110] and is much less reactive.

The potassium trialkylborohydrides find a variety of uses as selective reducing agents [111], [112]. Potassium triethylborohydride is a strong reducing agent that rapidly reduces a number of functional groups [113]. For example, disulfides are reduced to thiols, epoxides are reduced to give the more substituted alcohol, and primary bromides and iodides are reduced to the corresponding hydrocarbon.

Because of its bulky nature, potassium tri-sec-butylborohydride (KTSBBH) exhibits high stereoselectivity in the reduction of

hindered cyclohexenones [114]. An example of high diastereofacial stereoselectivity achieved using potassium tri-sec-butylborohydride is the reduction of optically pure acyl lactones to protected *syn*-diols [115]. In comparison, reduction with sodium borohydride gave a *syn*:*anti* ratio of 35:65.

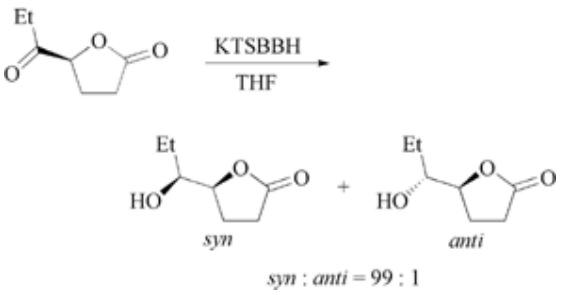

Regioselective reduction of unsymmetrical cyclic anhydrides by trialkylborohydrides affords lactones [116]. Potassium tri-*sec*-butylborohydride cleanly reduces α, β-unsaturated enones in a 1,4-addition to give ketones [117], [118] In addition, potassium trialkylborohydrides and potassium triisopropoxyborohydride have been used in organometallic synthesis [119].

3.3. Potassium Bases

Another major application of potassium metal involves preparing potassium bases by treating the metal with alcohols and amines (⟶ Alcohols, Aliphatic – Alkoxides).

3.3.1. *Potassium Alkoxides*

Potassium alkoxides (also called alcoholates) are moderately strong bases that are used in countless drug- and chemical-synthesis applications such as alkylation, arylation, acylation, solvolysis of esters, condensation, elimination, isomerization, rearrangements, transfer hydrogenation, Wittig reactions, and redox reactions. Many potassium bases are commercially produced on the multitonne scale [120]. The most commonly used bases are potassium methylate [865-33-8], potassium *tert*-butylate (KTB) [865-47-4], and potassium *tert*-amylate (KTA) [41233-93-6]. Annual world consumption of potassium alkoxides (as $KOCH_3$) exceeds 2000 t. The bases derived from tertiary alcohols offer such advantages as strength (pK_a = 18), solubility (see Table 3), selectivity (due to the bulky *tert*-alkyl groups), and stability (because of the lack of α-protons). In addition, KTB and KTA powder have long shelf lives of several years when stored under an inert atmosphere.

Table 3. Solubility of KTA and KTB at 25 °C (wt %)

Solvent	KTA	KTB
Toluene	36	2.3
Cyclohexane	36	0.4
Hexane	30	0.3
2-Methoxyethyl ether	47	42
Tetrahydrofuran	54	44

The base strength can be varied with choice of solvent and the structure of the alcohol. In DMSO, the ion aggregates are solvated, thus increasing the basicity. As alkyl groups are added to the alcohol, the alcoholate becomes a stronger base compared to primary alcoholates and hydroxide. Furthermore, the higher base strength and steric bulk of the tertiary alcoholates influence the speed, selectivity, and specificity.

KTB and KTA are usually preferred over NaH for deprotonation reactions because they have good solubility, leave no oil residue, and produce no hydrogen. Selective deprotonation can be achieved with KTB due to the steric hindrance provided by the tertiary butyl group. The following examples of potassium alkoxides illustrate the diverse applications of these bases.

Aryl- or heteroaryluracils, useful as herbicides, were prepared with substantially increased yields over previous methods by cycloaddition of isocyanates in the presence of 0.1 – 0.3 equivalents of KTB in DMF or DMSO [121].

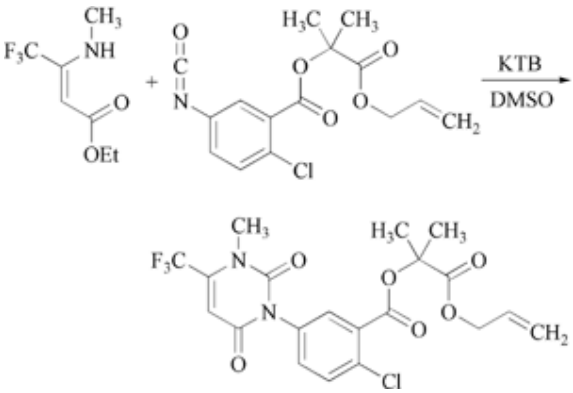

By a Wittig reaction, cephalosporin derivatives were prepared in 75 % yield via conversion of the phosphonium bromide to the corresponding ylide with KTB, then treating the ylide with an aldehyde [122].

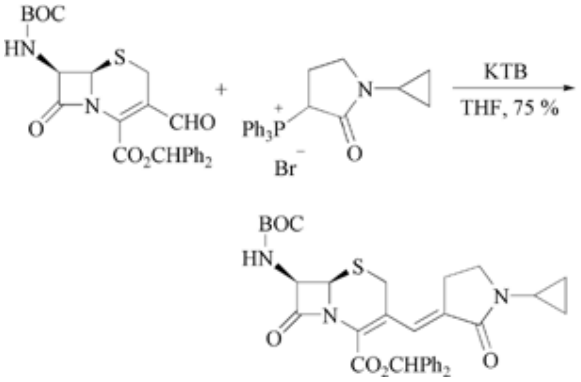

The hydrocarbon solubility of potassium tert-amylate (KTA) gives it an advantage over sodium and potassium alcoholates derived from *tert*-butanol or primary alcohols. In the synthesis of α-chloroketones from amino acid phenyl esters, a sulfur ylide was formed by deprotonation with KTA. The amount of desired product was dramatically increased with KTA versus NaH [123].

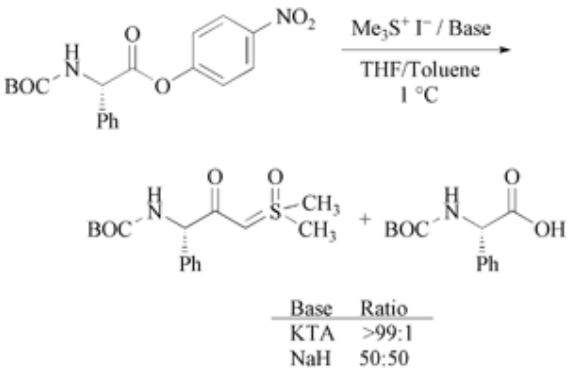

Base	Ratio
KTA	>99:1
NaH	50:50

For higher base strength, mixed metal reagents derived from organolithium compounds and potassium alkoxides (Superbase or LICKOR reagent [124]) can provide unique reactivity and selectivity. Superbases are capable of deprotonating compounds of very low acidity, to about pK_a = 45. *ortho,ortho′*-Disubstituted diphenylacetylenes were prepared by use of a superbase made from n-butyllithium and KTB. The products were prepared in moderate yields but with high regioselectivity [125].

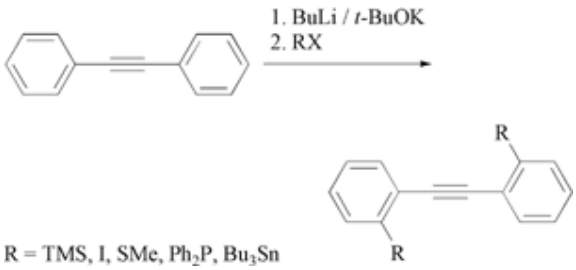

A growing use of potassium methoxide is in the interesterification of vegetable oils and waste cooking oil to make biodiesel [126]. Typically, potassium methoxide is a more efficient catalyst that NaOH or KOH for this process.

3.3.2. *Potassium Amides*

The most useful of the potassium amides is potassium hexamethyldisilazane [40949-94-8] (KHMDS), $KN[Si(CH_3)_3]_2$. This strong (pK_a = 28), extremely soluble, stable and nonnucleophilic potassium amide base is available in commercial quantities [127] in THF, toluene, and 1,3,5-trimethylbenzene solutions. KHMDS is ideal for regio- and stereospecific deprotonation and has demonstrated benefits for reactions involving kinetic enolates [128], alkylation and acylation [129], Wittig reaction [130], epoxidation [131], Ireland – Claisen rearrangement [132], [133], isomerization [134], [135], Darzen reaction, [136] Dieckmann condensation [137], cyclization [138], chain [139]/ring expansion [140], and elimination [141].

KHMDS efficiently forms the kinetic enolate of a carbonyl compound with α-hydrogen atoms [142]. Potassium enolates are highly reactive [143], for example, a potassium oxindol enolate reacted about ten times faster than the corresponding lithium enolate [144].

The cation can have an effect on the subsequent selectivity of the enolate alkylation. Sodium and lithium enolates yield primarily C-alkylation products while potassium enolates O-alkylation. For example, in the multikilogram synthesis of intermediates for production of LFA-1 inhibitors, a bis-lactam was deprotonated with KHMDS and O-alkylated with $(EtO)_2POCl$ [145]. Equilibration to thermodynamic enolates by warming to ambient temperature allowed O alkylation of an enolate to prepare a Claisen precursor [146].

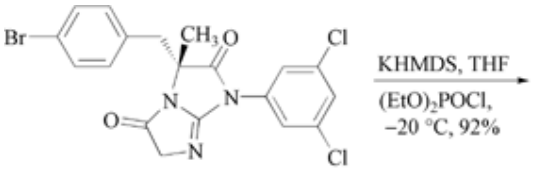

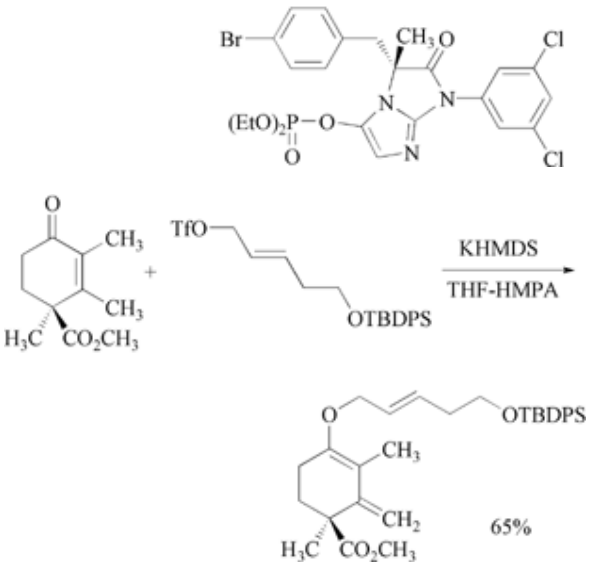

A Wittig reaction is an effective way to replace a carbonyl group with an olefinic residue [147]. Both NaHMDS and KHMDS have been used effectively to generate phosphorous ylides under lithium-free conditions to attain high Z selectivity [148]. KHMDS was the reagent of choice in the Wittig synthesis of Z trisubstituted allylic alcohols [149]. A synthesis of a glaucoma drug used KHMDS in the Wittig reaction to deliver 97.8 % *cis* product along with only 2.2 % *trans* product, whereas KTB gave 3.5 % *trans* [150].

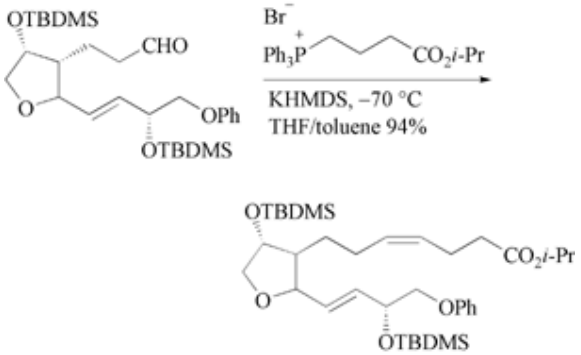

Vinylaziridines are required for a precursor to the synthesis of frog alkaloids [151]. Use of KHMDS as the base gave the desired product with an *E*:*Z* ratio of 1:17.

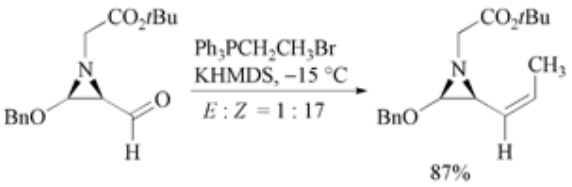

The anionic version of the oxy-Cope rearrangement proceeds more rapidly and is called the oxyanionic Cope rearrangement. In the synthesis of taxol derivatives, KHMDS was used in the anionic oxy-Cope rearrangement to give optically pure tricyclic compounds [152].

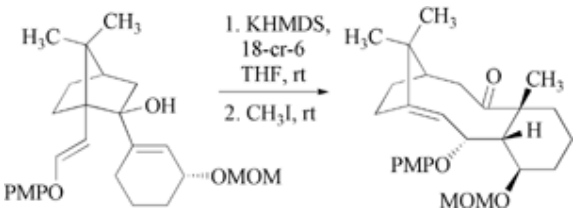

Other potassium amides such as potassium diisopropylamide [67459-71-6], potassium 2,2,6,6-tetramethylpiperidide [110426-49-8], and potassium 3-aminopropylamide [56038-00-7] ($KNHCH_2CH_2CH_2NH_2$, KAPA) have not yet found widespread application. Due to their tendency to decompose or react with solvent, these reagents are usually prepared and used in situ. KAPA, pK_a 35, can be prepared by the reaction of 1,3-diaminopropane with potassium metal or

potassium hydride [153], [154], [155]. However, due to the extremely unstable nature of this amide, in situ use is strongly advised. KAPA is extremely effective in converting internal acetylenes to the terminal isomers, which are stabilized as potassium acetylides. For example, KAPA isomerizes 3-methyl-4-nonyne to 7-methyl-1-nonyne in about 1 min at 0 °C [156].

3.4. Potassium Superoxide

Potassium behaves similarly to rubidium and cesium on oxidation by oxygen, forming potassium superoxide [12030-88-5], KO_2. The Mine Safety Appliances Company first developed potassium superoxide for use as an oxygen source in self-contained breathing equipment, since it liberates oxygen in contact with exhaled moisture and carbon dioxide [157].

Yellow potassium superoxide forms on the surface of potassium metal on exposure to air or oxygen dissolved in solvents. Extreme care must be taken when this yellow solid is found on potassium stored in hydrocarbon oils, because explosions have occurred when the yellow KO_2 is disturbed, such as by cutting or scraping.

Since potassium superoxide is a strong oxidizing agent, its utility in organic synthesis is limited but has been briefly studied. The unique chemical behavior of KO_2 is a result of its dual character as a radical anion and a strong oxidizing agent [158], [159]. The reactivity and solubility of KO_2 is greatly enhanced by crown ethers [160]. Its usefulness in furnishing oxygen anions is demonstrated by its applications in S_N2-type reactions to displace methanesulfonate and bromide groups [161], the oxidation of benzylic methylene compounds to ketones [162], and the synthesis of a-hydroxyketones from ketones [163]. Appropriate caution must be taken due to the hazard associated with potassium superoxide in the presence of organic materials. For example, acetone vapors have exploded in the presence of KO_2.

[Top of Page]

4. Acknowledgement

The author thanks colleagues at BASF Corporation for helpful comments on the manuscript.

[Top of Page]

References

General References

1. D. R. Lide (ed.): *Handbook of Chemistry and Physics*, 78th Ed., CRC Press, Baton Rouge 1997–1998, pp. 4–23.
2. Gmelin, system no. 22 (1939) 64–69.
3. Commission on Atomic Weights and Isotopic Abundances Report for the International Union of Pure and Applied Chemistry in Isotopic Compositions of the Elements 1989, *Pure Appl. Chem.* **70** (1998) 217.
4. See website http://www.palomar.edu/oceanography/salty_ocean.htm.
5. J. P. Searls, *U.S. Geological Survey Minerals* Yearbook 2003.
6. Locations where this mineral is found: http://www.mindat.org/show.php?id=3850&ld=1&pho=.
7. Locations where this mineral is found: http://www.mindat.org/min-2320.html.
8. *Landolt-Börnstein* (1965) **4**, 2 C, 351. D. R. Lide (ed.); *Handbook of Chemistry and Physics*, 78th ed., CRC Press, Baton Rouge 1997–1998.
9. Properties of the Elements in Knovel Critical Tables 2003 at www.Knovel.com.
10. J. A. Dean (ed.): *Lange's Handbook of Chemistry*, McGraw-Hill, New York 1999.
11. *Working equations available also* in C. L. Yaws (ed.): *Chemical Properties Handbook*, McGraw-Hill, New York 1999.
12. D. W. Green, J. O. Maloney (eds.): *Perry's Chemical Engineers' Handbook*, 7th ed., McGraw-Hill, New York 1997, Tables 2–7. J. C. Bailar *et al.* (eds.): *Comprehensive Inorganic Chemistry*, vol. **1**, Pergamon Press, Oxford 1973, pp. 381, 387.
13. R. E. Honig, D. A. Kramer, *Vapor Pressure Data for the Solid and Liquid Elements*, RCA Rev. 1969, 285.
14. A. Monaco, *J. Chem. Phys*. **120** (2004) 8089. Links
15. Chem. Eng. News 80 (2002) no. 37, 2. J. Evers, T. M. Klapoetke, G. Oehlinger, pubs.acs.org/cen/safety/potassium.html.
16. I. Fatt, M. Tashima: *Alkali Metal Dispersions*, Van Nostrand, New Jersey 1961.
17. W. L. Jolly: *The Inorganic Chemistry of Nitrogen*, W. A. Benjamin Inc., New York 1964, p. 32.
18. J. C. Bailar *et al.* (eds.): *Comprehensive Inorganic Chemistry*, vol. **1**, Pergamon Press, Oxford 1973, pp. 381 – 387.
19. Z. Jedlinski, M. Kowalczuk, P. Kurcok, M. Grzegorzek, J. Ermel, *J. Org. Chem.* **52** (1987) 4601. Links Z. Grobelny, A. Stolarzewicz, *J. Org. Chem.* **70** (2005) 8201. Links
20. Z. Grobelny, A. Stolarzewicz, M. Sokol, J. Grobelny, H. Janaczek, *J. Phys. Chem.* **92** (1992) 5193.
21. A. Winiarski, Z. Grobelny, A. Stolarzewicz, *Appl. Crystallogr.* **19** (2004) 397. Z. Grobelny, A. Stolarzewicz, B. Morejko-Buz, A. Winiarski, *Tetrahedron* **59** (2003) 2147. Links
22. *Winnacker-Küchler*, 4th ed., **6**, 102. R. Thilenius, Z. Elektrochem. *Angew. Phys. Chem.* **37** (1931) 740.
23. Mine Safety Appliances Co., US 2 480 655, 1949 (C. B. Jackson, R. C. Werner).
24. Chem. Fabrik Griesheim-Elektron, DT 239 106, 1909. H. Speketer, Nernst-Festschrift, Knapp, Halle/Saale 1912, p. 427.
25. A. S. Mikulinskii *et al.*, *Akad. Nauk SSSR Inst. Met.* 1966, 58–62; *Tr. Inst. Metall. Sverdlovsk* **9** (1966) no. 11, 100–

118; *Chem. Abstr.* **67** (1976) 24 743, **67** (1967) 110 738, **68** (1968) 52 188, **67** (1967) 74 780. V. A. Kosarev *et al.*, *Tr. Inst. Met. Sverdlovsk* **9** (1966) no. 11, 197–209; *Chem. Abstr.* **67** (1967) 55 622, 55 623.

26. G. N. Zviadadze, D. Sh. Oziashvili *et al.*, *Tr. Gruz. Inst. Metall.* **15** (1966) 27–35; *Chem. Abstr.* **68** (1968) 108 326.
27. G. N. Zviadadze, D. Sh. Oziashvili *et al.*, *Tr. Gruz. Inst. Metall.* **15** (1966) 56–63; *Chem. Abstr.* **68** (1968) 52 189.
28. Houillères du Bassin du Nord et du Pas-de-Calais, FR 1 469 634, 1966.
29. Houillères du Bassin du Nord et du Pas-de-Calais, FR 1 469 635, 1966.
30. Potassium metal is available from BASF Corporation, Metaux Speciaux, and Suparna Chemicals Ltd.
31. Recommendations on the Transport of Dangerous Goods Model Regulations, 14th revised edition, copyright United Nations, 2005 (ISBN 92-1-139106-7) United Nations Publications. http://www.unece.org/trans/danger/publi/unrec/rev14/14files_e.html.
32. MSDS from BASF Corp. 1424 Mars-Evans City Rd. Evans City, PA 16033.
33. M. A. Armour, L. M. Browne, G. L. Weir: *Hazardous Chemical Information and Disposal Guide*, 2nd ed., University of Alberta, Edmonton, Alberta, Canada 1984, p. 202.
34. L. Bretherick: *Hazards in the Chemical Laboratory*, 3rd ed., Royal Society of Chemistry, London 1981.
35. W. Mialki, *Metall (Berlin)* **3** (1959) 174.
36. M. Hansen, K. Anderko(eds.): *Constitution of Binary Alloys*, McGraw-Hill, New York 1958, pp. 876 – 877. R. P. Elliott (ed.): *Constitution of Binary Alloys*, 1st Suppl., McGraw-Hill, New York 1965, p. 567. F. A. Shunk (ed.): *Constitution of Binary Alloys*, 2nd Suppl., McGraw-Hill, New York 1969, p. 469.
37. O. J. Foust (ed.): *Sodium-NaK Engineering Handbook*, vol. **1**, Gordon and Breach, Science Publishers, New York 1972.
38. K. H. Spiller, D. Droxte, *Chem. Labor Betr.* **22** (1971) 481 – 490, 538 – 542.
39. R. L. McKisson in M. Zief (ed.): *Preparation of Pure Sodium Potassium*, Ultrapurity, New York 1972.
40. Recommendations on the Transport of Dangerous Goods Model Regulations, 14th revised ed., copyright United Nations, 2005 (ISBN 92-1-139106-7) United Nations Publications. http://www.unece.org/trans/danger/publi/unrec/rev14/14files_e.html.
41. MSDS from BASF Corp.
42. H. Pines, W. M. Stalick, H. B. Injovanovich (eds.): *Base-Catalyzed Reactions of Hydrocarbons and Related Compounds*, Academic Press, New York 1977.
43. M. Szwarc: *Carbanions, Living Polymers, and Electron Transfer Processes*, Interscience, New York 1968.
44. E. Grovenstein, Jr., K. W. Chiu, B. B. Patil, *J. Am. Chem. Soc.* **102** (1980) 5848. Links E. Grovenstein, *Adv. Organomet. Chem.* **16** (1977) 167. Links
45. L. Lochmann, J. Pospisil, D. Lim, *Tetrahedron Lett.* (1966), 257 – 262. Links
46. L. Lochmann, J. Petranek, *Tetrahedron Lett.* **32** (1991) 1483. Links
47. L. Lochmann, J. Trekoval, *J. Organomet. Chem.* **326** (1987) 1–7. Links
48. M. Schlosser, S. Strank, *Tetrahedron Lett.* **25** (1984) 741–744. Links
49. M. Schlosser, *Angew. Chem. Int. Ed. Engl.* **4** (1965) 362.
50. J. Kong *et al.*, *Appl. Phys. Lett.* **77** (2000) 3977. Links
51. N. Swami, M. E. Thompson, *J. Appl. Phys.* **85** (1999) 3696. Links
52. P. Chen, X. Wu, J. Lin, K. L. Tan, *Science* **285** (1999) 91. Links
53. *Chem. Eng. News* **33** (1955) 648.
54. W. G. Anderson *et al.*, *Proc. 25th Intersoc. Energy Convers. Eng. Conf.* **25** (1990) no. 5, 268 – 273. JP 2003262690, 2003 (M. Toda).
55. JP 03 263 881, 1991 T. Yagakinai *et al.*). *Chem. Eng. News* **25** (1955) 648.
56. P. L. Kirillov, P. A. Ushakov, *Teploenergitika (Moscow)* 2002, 49.
57. H. S. Rhee *et al.*, *Proc. Intersoc. Energy Convers. Eng. Conf.* **25** (1990) no. 4, 8 – 13. L. Barleon, U. Burr, K. J. Mack, R. Steiglitz, *Fusion Eng. Des.* 51 – 52 (2000) 723. J. Zhang, Z. Hu, G. Mei, Z. Shuang, S. Zhang, B. Zhang, CN 1182869 A 5-27, 1998. P. L. Kirillov, P. A. Ushakov, *Teploenergetika* **1** (2001) 49.
58. US 5 126 530, 1992 (H. Bergmer). US 4 406 046, 1983 P. Adam).
59. F. Hensel, G. F. Hohl, D. Schaumloffel, W. C. Pilgrim, *Z. Phys. Chem.* **214** (2000) 823.
60. Dow Chemical Co., US 5 085 955, 1992 R. A. Cipriano).
61. D. A. Wallace, *et al.*, *Proc. Intersoc. Energy Convers. Eng. Conf.* **26** (1991) no. 5, 349-354.
62. US 5 904 783, 1999 (J. P. Hager, J. P. Downey).
63. Ethyl Corp., US 4 977 124, 1990 (R. S. Smith). Ethyl Corp., US 4 950 831, 1990 (J. S. Staton). Ethyl Corp., US 4 929 783, 1990 (R. S. Smith). US 4 179 580, 1979 (R. L. Cobb). J. S. Lee, S. C. Lee, M. H. Kil, S. Choi, J. S. Lee, J. C. Kim, *Hwahak Konghak* **40** (2002) 669.
64. S. Rummel, S. M. Yunusov, H. Langguth, V. B. Shur, *Russ. Chem. Bull.* **48** (1999) 2083.
65. Teijin, US 5 365 001, 1994 (K. Sumitani *et al.*); US 4 018 840, 1977; US 3 953 535, 1976; US 3 865 889, 1975; US 3 766 288, 1973. J. S. Lee *et al.*, Hwahak Konghak 40 (2002) 669.
66. Sumitomo Chemical Co. US 4 720 601, 1988 (G. S. Osaka, M. F. Shiga). G. Suzukamo, M. Fukao, M. Minobe, A. Sakamoto, EP-A 211448, 1987.
67. Union Carbide, US 5 321 180, 1994 (C. S. Davis). G. P. Pez *et al.*, *J. Am. Chem. Soc.* **102** (1985) 4098–4100. D. J. Mathre, W. C. Guida, *Tetrahedron Lett.* **21** (1980) 4773–4776. Links
68. GB 820 263, 1959 (Z. Karl).

69. C. W. Carlson, R. West, *Organometallics* **2** (1983) no. 12, 1972 – 1977.
70. P. A. Bianconi, T. W. Weidman, *J. Am. Chem. Soc.* **110** (1988) no. 7, 2342 – 2344. Links
71. H. Y. Qui, Z. D. Du, *Goadeng Xuexiao Huaxue Xuebao* **10** (1989) no. 4, 423.5.
72. US 4 800 221, 1989 (O. W. Marko).
73. Z. Jedlinski *et al.*, *Makromol. Chem. Rapid. Commun.* **4** (1983) no. 6, 153 – 168.
74. V. Harabagiu, M. Pinteala, C. Cotzur, C. Iojoiu, B. C. Simionescu, *Bul. Inst. Politehnic Iasi Sect. 2: Chim. Ing. Chim.* **43** (1997) 83. EP-A 632 086, 1995 K. Ikai *et al.*). G. T. Visscher, P. A. Bianconi, *J. Am. Chem. Soc.* **116** (1994) 1804–1811. P. A. Bianconi, F. C. Schilling, T. W. Weidman, *Macromolecules* **22** (1989) 1697–1704. Links
75. PTC Int. Appl. WO 9 524 368, 1995 (P. A. Bianconi, B. L. Aldrich). Z. Sun, Y. Sun, S. R. Wilson, *Thin Solid Films* 377 – 378 (2000) 203. Links
76. J. Feldthusen, B. Ivan, A. H. E. Muller: “Functional Polymers”, *ACS Symp. Ser.* **704** (1998) 121.
77. JP 11286507 A2 10-19- 1999 (K. Takami).
78. R. D. Rieke, *Acc. Chem. Res.* **10** (1977) 301. Links R. D. Rieke, *Science* **246** (1989) 1260. Links R. D. Rieke, L. D. Rhyne, *J. Org. Chem.* **44** (1979) 3445. Links
79. K. L. Tsai, J. L. Dye, *J. Am. Chem. Soc.* **113** (1991) 1650. Links
80. J. A. Nelson, L. H. Bennett, M. J. Wagner, *J. Am. Chem. Soc.* **124** (2002) 2979. Links J. A. Nelson, M. J. Wagner, *Chem. Mater.* **14** (2002) 915. Links
81. R. M. Schramm, C. E. Langlois, *Prepr. Div. Petr. Chem. Am. Chem. Soc.* **4** (1959) no. 4, B-53.
82. PCT Int. Appl., WO 8 607 097, 1986 (W. R. B. Martin).
83. T. Ohsawa, T. Takagaki, A. Haneda, T. Oishi, *Tetrahedron Lett*. 22 (1981) 2583–2586. Links
84. T. Ohsawa *et al.*, *Chem. Phar. Bull.* **30** (1982) no. 9, 3178 – 3186.
85. C. A. Ogle, T. E. Wilson, J. A. Stowe, Synthesis 1989, 495 – 496.
86. T. A. Thornton *et al.*, *J. Am. Chem. Soc.* **111** (1989) 2434–2440. Links
87. Jpn. Kokai Tokyo Koho, JP 63 218 705, 1988 M. Ohata *et al.*).
88. K. Miyoshi, T. Nishio, A. Yasuhara, M. Morita, T. Shibamoto, *Chemosphere* **55** (2004) 1439. Links
89. DD 235 184, 1986 (G. Houbelein, D. Stadermann, H. Hartung, W. Muller).
90. PCT Int. Appl. WO 9 501 409, 1995 (G. M. Tom, V. McManus). US 4 604 270, 1986 (G M. Tom).
91. J. Desreumaux, M. Calais, R. Andriano, S. Trambaud, C. Kappenstein, M. Hguefack, *Eur. J. Inorg. Chem.* **9** (2000) 2031.
92. BE 890 414, 1982 (T. A. Pellose, L. Kogan). Safinco, EP-A 9207, 1980 R. de Lathauwer, *et al.*). Proctor & Gamble Co., US 2 733 251, 1956 (K. H. Harold, R. D. Dobson).
93. Thomas Hedley & Co. Ltd., GB 785 147, 1957. D. Kazimierz, *Tluszcze i Srodki Piorace* **5** (1961) 143–149.
94. W. Rudorff in H. J. Emeleus, A. G. Sharpe (eds.): *Graphite Intercalation Compounds, Advances in Inorganic Chemistry and Radiochemistry*, vol. **1**, Academic Press, New York 1959, pp. 224 – 264. G. R. Henning in F. A. Cotton (ed.): *Interstitial Compounds of Graphite, Progress in Inorganic Chemistry*, vol. **1** Interscience, New York 1959, pp. 125 – 205.
95. R. Csuk, B. I. Glanzer, A. Fuerstner, *Adv. Organomet. Chem.* **28** (1988) 85. Links A. Fuerstner, *Angew. Chem. Int. Ed. Engl.* **32** (1993) 164.
96. A. Fuerstner, H. Weidmann, *J. Carbohydr. Chem.* **7** (1989) 773.
97. R. Csuk, B. I. Glanzer, A. Fuerstner, *Adv. Organomet. Chem.* **28** (1988) 85. Links A. Fuerstner, *Angew. Chem. Int. Ed. Engl.* **32** (1993) 164.
98. L. S. Hegedus, *Pure Appl. Chem.* **62** (1990) 691. Links
99. W. Uhlig, *Z. Naturforsch. B* **50** (1995) 1674 – 1678. B. Lacave-Goffin, L. Hevesi, J. Devaux, *J. Chem. Soc. Chem. Commun.* 1995, 769, 770.
100. PCT Int. Appl. WO 9634825 A1, 1996 (G. Leighter, J.-C. Cachon, D. Guerard).
101. A. Fuerstner in L. A. Paquette (ed.): *Encyclopedia of Reagents for Organic Synthesis* vol. **6**, Wiley, New York 1995, p. 4180. R. Rienaecker, J. Graefe, *Angew. Chem. Int. Ed. Eng.* **24** (1985) 320.
102. PCT Int. WO 2005051839, 2005 (M. Lefenfeld, J. L. Dye).
103. Mine Safety Appliances Co., US 4 327 071, 1982 (K. W. Chiu, J. R. Strickler). Y. Zhang, S. Laio, Y. Xu, *J. Mol. Catal.* **84** (1993) 211.
104. J. B. Condon, *Diffus. Defect Data Part B* **49050** (1996) 25.
105. H. W. Pinnick, *Org. Prep. Proced. Int.* **15** (1983) 199. Links
106. C. A. Brown, *J. Org. Chem.* **39** (1974) 3913. Links
107. C. A. Brown, *Synthesis* 1975, 326. Links
108. A. Stolarzewicz, D. Neugebaure, J. Grobelny, *Polym. Macromol. Rapid Commun.* **17** (1996) 787.
109. P. Wang, X.-D. Kang, H.-M. Cheng *J. Appl. Phys.* **98** (2005) 74905.
110. H. C. Brown, *et al.*, *J. Org. Chem.* **49** (1984) 885 – 892. Links H. C. Brown *et al.*, *Inorg. Chem.* **23** (1984) 2929-2931.
111. C. A. Brown, *J. Am. Chem. Soc.* **59** (1973) 4100.
112. S. Krishnamurthy, *Aldrichimica Acta* **7** (1974) 55.
113. N. M. Yoon, H. S. Yang, Y. S. Hwang, *Bull. Korean Chem. Soc.* **8** (1987) 285.
114. Ciba-Geigy Corp., US 5 428 159, 1995 (W. C. Shieh, J. A. Carlson).
115. L. Larcheveque, *J. Chem. Soc. Chem. Commun.* 1985, 83.

116. S. Krishnamurthy, W. B. Vreeland, *Heterocycles* **18** (1982) 265–270. Links
117. B. Ganem, *J. Org. Chem.* **40** (1975) 146. Links
118. J. M. Fortunato, B. Ganem, *J. Org. Chem.* **41** (1976) 2194–2200. Links
119. J. A. Gladysz, *Aldrichimica Acta* **12** (1979) 13–17. Links
120. Potassium alcoholates are available from BASF Corporation, 1424 Evans-City Rd, Evans City, PA 16033.
121. Syngenta Crop Protection, Inc., US 6 207 830, 2001 (A. R. Sting).
122. F. Hoffmann-La Roche AG, EP-A 1067131, 2001 (P. Spurr, G. Trickes).
123. Bristol-Meyers Squibb, PCT Int. Appl. WO 0214256, 2002 (D. Krothenthal, M. D. Schwinden).
124. L. Lochmann, *Eur. J. Inorg. Chem.* **6** (2000) 1115.
125. J. Kowalik, L. M. Tolbert, *J. Org. Chem.* **66** (2001) 3229 Links
126. http://www.basf.com/businesses/chemicals/alcoholates/biodiesel/catalysts.html.
127. Available from BASF Corporation, 1424 Evans-City Rd, Evans City, PA 16033.
128. J. Tsuji, I. Minami, I. Shimizu, *Tetrahedron Lett.* **24** (1983) 1973.
129. G. M. Coppola, *Synth. Commun.* **15** (1985) 135.
130. E. J. Corey, M. M. Mehrotra, *Tetrahedron Lett.* **27** (1986) 5173. Links
131. W. C. Still, V. J. Novack, *J. Am. Chem. Soc.* **103** (1982) 1282.
132. A. B. Smith *et al.*, *J. Am. Chem. Soc.* **104** (1982) 4105.
133. J. Kallmerter, T. J. Gould, *J. Org. Chem.* **51** (1986) 1155.
134. Y. Ikeda, H. Yamamoto, *Tetrahedron Lett.* **25** (1984) 5581.
135. Y. Ikeda, J. Ukai, N. Ikeda, H. Yamamoto, *Tetrahedron Lett.* **25** (1984) 5177. Links
136. Pfizer, EP 0 076 643, 1982 (P. Weeks).
137. Pfizer, EP 0 076 643, 1982 (P. Weeks).
138. K. Takeda *et al.*, *J. Org. Chem.* **51** (1986) 4735. Links
139. D. Horne, J. Gaudino, W. J. Thompson, *Tetrahedron Lett.* **25** (1984) 3529. Links
140. T. Tekahashi, S. Hashiguchi, K. Kasuga, J. Tsuji, *J. Am. Chem. Soc.* **100** (1978) 7424.
141. S. Hanessiam, A. Ugolini, P. J. Hodges, D. Dube, *Tetrahedron Lett.* **27** (1986) 2699.
142. D. A. Evans in J. D. Morrison (ed.): *Asymmetric Synthesis, Academic, New York, 1984*, vol.**3**, p. 1. C. A. Brown, *J. Org. Chem.* **39** (1974) 3913. Links
143. D. A. Evans, T. C. Britton, *J. Am. Chem. Soc.* **109** (1987) 6881. Links D. A. Evans, J. A. Ellman, *J. Am. Chem. Soc.* **111** (1989) 1063. Links
144. L. E. Overman *et al.*, *J. Am. Chem. Soc.* **126** (2004) 14043. Links
145. R. P. Frutos *et al.*, *Org. Process Res. Dev.* **9** (2005) 137. Links
146. R. K. Boeckman, Jr. *et al.*, *J. Am. Chem. Soc.* **124** (2002) 190. Links
147. B. C. Hamann, J. F. Hartwig, *J. Am. Chem. Soc.* **119** (1997) 12382. Links K. H. Shaughnessy, B. C. Hamann, J. F. Hartwig, *J. Org. Chem.* **63** (1998) 6546. Links
148. H. J. Bestmann, W. Stransky, O. Vostrowsky, *Chem. Br.* **109** (1976) 1694. G. Storck, K. Zhao, *Tetrahedron Lett.* **30** (1989) 2173. K. Tago, M. Arai, H. Kogen, *Perkin Trans. 1* **13** (2000) 2073.
149. C. Sreekumar, K. P. Darst, W. C. Still, *J. Org. Chem.* **45** (1980) 4260. Links
150. M. E. Fox *et al.*, *J. Org. Chem.* **70** (2005) 1227. Links
151. J. Ahman, T. Jarevang, P. J. Simfai, *Org. Chem.* **61** (1996) 8148.
152. L. A Paquette, S. K. Huber, R. C. Thompson, *J. Org. Chem.* **58** (1993) 6874. Links
153. C. H. Heathcock, J. A. Stafford, *J. Org. Chem.* **57** (1992) 2566. Links
154. C. Almonsa, A. Moyano, M. A. Pericas, F. Serratosa, *Synthesis* 1988, 707.
155. J. H. Wotiz, P. M. Barelski, D. F. Koster, *J. Org. Chem.* **38** (1973) 489. Links
156. C. A. Brown, A. Yamashita, *J. Am. Chem. Soc.* **97** (1975) 891. Links
157. J. W. Mausteller: “Characterization of Carbon Dioxide Absorbing Agents Life Support Equipment,” *Winter Annu. Meet. Am. Soc. Mech. Eng.* **97** (1982) 23-31.
158. D. T. Sawyer, J. S. Valentine, *Acc. Chem. Res.* **14** (1981) no. 12, 393. Links E. Lee-Ruff, *Chem. Soc. Rev.* **6** (1977) 195. Links
159. E. Lee-Ruff, *Chem. Soc. Rev.* **6** (1977) 195. Links L. Sotirion, W. Lee, R. W. Giese, *J. Org. Chem.* **55** (1990) 2159-2164. Links
160. E. Lee-Ruff, *Chem. Soc. Rev.* **6** (1977) 195. Links L. Sotirion, W. Lee, R. W. Giese, *J. Org. Chem.* **55** (1990) 2159 – 2164.
161. R. A. Johnson, E. G. Nidy, L. Baczynski, R. R. Gorman, *J. Am. Chem. Soc.* **99** (1977) 7738. Links N. A. Porter, J. D. Byers, R. C. Mebane, D. W. Gilmore, J. R. Nixon, *J. Org. Chem.* **43** (1978) 2088. Links
162. Y. H. Kim, K. S. Kim, H. K. Lee, *Tetrahedron Lett.* **30** (1989) 6357–6360. Links
163. C. Betancor, C. G. Francisco, R. Freire, E. Suarez, *J. Chem. Soc. Chem. Commun.* 1988. Links

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Potassium Compounds

Potassium Acetate → Acetic Acid ; Potassium Alkoxides → Potassium Compounds ; Potassium Aluminum Sulfate → Aluminum Compounds, Inorganic ; Potassium Amides → Potassium Compounds ; Potassium Ammonium Nitrate → Ammonium Compounds ; Potassium Arsenate → Arsenic and Arsenic Compounds ; Potassium Benzoate → Benzoic Acid and Derivatives ; Potassium Bromide, Potassium Bromate → Bromine Compounds ; Potassium Chlorate, Potassium Perchlorate → Chlorine Oxides and Chlorine Oxygen Acids ; Potassium Chromate, Potassium Dichromate → Chromium Compounds ; Potassium Cyanide, Potassium Cyanate, Potassium Hexacyanoferrate → Cyano Compounds, Inorganic ; Potassium Fluoride → Fluorine Compounds, Inorganic ; Potassium Hydride → Hydrides ; Potassium Iodide → Iodine and Iodine Compounds ; Potassium Nitrate → Nitrates and Nitrites ; Potassium Permanganate → Manganese Compounds ; Potassium Phosphate → Phosphoric Acid and Phosphates ; Potassium Superoxide → Potassium and Alloys

Heinz Schultz, Kali und Salz AG, Kassel, Federal Republic of Germany (Chaps. 1, 4, 6 – 10; Sections 3.3, 5.1, 5.2, 5.4 – 5.6, 5.8 – 5.12)

Günter Bauer, Kali und Salz AG, Kassel, Federal Republic of Germany (Chap. 2; Sections 3.1 and 3.2)

Erich Schachl, Kali und Salz AG, Kassel, Federal Republic of Germany (Chap. 2; Sections Sections 3.1 and 3.2)

Fritz Hagedorn, Kali und Salz AG, Kassel, Federal Republic of Germany (Sections 5.3 and 5.7)

Peter Schmittinger, Hüls Aktiengesellschaft, Werk Lülsdorf, Niederkassel, Federal Republic of Germany (Chaps. 11– 13)

1. Introduction . . . 3
1.1. Occurrence . . . 3
1.2. History . . . 3
2. Potash Salt Deposits . . . 4
2.1. Minerals . . . 4
2.2. Geology of Potash Deposits . . . 6
3. Mining of Potash Salts . . . 10
3.1. Shaft Mining . . . 10
3.2. Extraction, Conveying, and Haulage 10
3.3. Solution Mining . . . 11
4. Treatment of Potash Ores . . . 12
4.1. Intergrowth and Degree of Liberation . . . 12
4.2. Grinding . . . 13
5. Potassium Chloride . . . 15
5.1. Properties . . . 15
5.2. Production by Crystallization from Solution . . . 16
5.2.1. Phase Theory . . . 16
5.2.2. Hot Leaching Process . . . 18
5.2.3. Processing of Carnallite . . . 22
5.2.4. Equipment . . . 24
5.3. Flotation . . . 29
5.3.1. Potash Ores Suitable for Flotation . . 29
5.3.2. Carrier Solutions . . . 29
5.3.3. Flotation Agents . . . 29
5.3.4. Theory . . . 30
5.3.5. Flotation Equipment . . . 32
5.3.6. Processes . . . 33
5.4. Electrostatic Separation . . . 36
5.4.1. Theoretical Basis . . . 36
5.4.2. Equipment and Processes . . . 36
5.5. Heavy-Media Separation . . . 38
5.6. Debrining and Drying . . . 38
5.7. Process Measurement and Control 41
5.8. Waste Disposal and Environmental Aspects . . . 42
5.9. Granulation . . . 43
5.10. Quality Specifications . . . 45
5.11. Toxicology and Occupational Health 46
5.12. Economic Aspects and Uses . . . 47
6. Potassium Sulfate . . . 49
6.1. Properties . . . 49
6.2. Raw Materials . . . 49
6.3. Production . . . 49
6.3.1. From Potassium Chloride and Sulfuric Acid (Mannheim Process) . . . 50
6.3.2. From Potassium Chloride and Magnesium Sulfate . . . 50
6.3.3. From Potassium Chloride and Langbeinite . . . 52
6.3.4. From Potassium Chloride and Kainite 53
6.3.5. From Potassium Chloride and Sodium Sulfate . . . 53
6.3.6. From Potassium Chloride and Calcium Sulfate . . . 53

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_039

6.3.7. From Alunite 54
6.3.8. From Natural Brines and Bitterns . . . 54
6.4. Granulation 54
6.5. Quality Specifications 55
6.6. Toxicology and Occupational Health 55
6.7. Economic Aspects and Uses 55
7. Potash – Magnesia 56
8. Production of Potassium Salts from Other Raw Materials 56
8.1. The Dead Sea 56
8.2. The Great Salt Lake 57
8.3. Searles Lake 57
8.4. Other Sources 58
9. Storage and Transportation 58
10. Analysis of Potassium Compounds . 58
11. Potassium Hydroxide 59
11.1. Properties 59
11.2. Production 59
11.3. Quality Specifications 60
11.4. Economic Aspects and Uses 60
12. Potassium Carbonate 60
12.1. Properties 60
12.2. Production 61
12.3. Quality Specifications and Analysis 63
12.4. Storage and Transportation 63
12.5. Economic Aspects and Uses 64
13. Potassium Hydrogencarbonate . . . 64
13.1. Properties and Production 64
13.2. Uses 64
14. References 65

1. Introduction

1.1. Occurrence

Potassium occurs in nature only in the form of its compounds. It is one of the ten most common elements in the earth's crust. Several widely distributed silicate minerals contain potassium, in particular, the feldspars (→ Silicates) and micas (→ Mica). The weathering of these minerals produces soluble potassium compounds, which are present in seawater and occur in extensive salt deposits. Potassium is important in the metabolism of plants and animals, and is therefore found in ash from plant materials and in the bodies of animals. Potassium compounds are obtained almost entirely by the mining of salt deposits.

1.2. History

Before the discovery and exploitation of potassium salt deposits, the production of potassium compounds consisted almost entirely of potash (K_2CO_3) obtained from natural sources such as wood ash, residues from distilleries, Bengal saltpeter, wool grease, and mother liquors from sea salt production. Quantities were small and were used only for the production of soap, glass, and explosives.

In 1840, J. von Liebig laid the foundations of the theory of mineral fertilizers in his paper "The Application of Organic Chemistry to Agriculture and Physiology" (→ Fertilizers, Chap. 2.). This spread the knowledge that potassium was one of the most important plant nutrition elements. In 1851, some deep mine workings in Stassfurt struck minerals containing potassium and magnesium, although these could not be used as fertilizers. In 1861, Adolph Frank started the first plant using the process that he developed for producing from carnallite a potassium salt that could be employed as a fertilizer. This soon led to the development of other processes and to the establishment of many potash mines and works. Many attempts were made up to the end of World War I to hinder the building of new factories in Germany, which had a world monopoly in the production of potash fertilizers, and attempts were made to ration production and supply, and to regulate prices. In spite of this, 69 factories were in existence in 1910, and 198 in 1918. When World War I ended, Alsace was returned to France, and the potash works that were built there shortly after 1900 became French property, so that Germany lost her monopoly in potash.

The German national assembly of 1919 enacted the potash regulations and the so-called closure order, which reduced the number of operating factories to 29 by the year 1938. This also caused production to be concentrated in a few large potash companies [17, 18].

After the end of World War II, ca. 60 % of German production capacity went to the area that was later to become the German Democratic Republic, and became the VEB Kombinat Kali. After German unification in 1990, the newly formed Mitteldeutsche Kali AG took over these

operations. In the Federal Republic of Germany, several groups of works were formed, leading in 1970 to the formation of the company Kali und Salz AG following a series of amalgamations. This company owns seven potash works, all located in the former Federal Republic of Germany [19, 20].

Most of the French potash works in Alsace were nationalized after World War I. During World War II, production was greatly increased. Several of the deposits have now become exhausted, and many of the works have been closed. Production has been concentrated in two large works, although these are likely to be closed down soon after the end of this century [21].

Potash production in Spain began in 1926 with the start up of a factory in Catalonia. Other factories were established later both there and in Navarra.

The kainite deposits in Sicily have been worked since 1959 – 1960; the salts are used for potassium sulfate production [22].

The Soviet Union began potash production in 1931 at a plant in the Northern Urals. In 1939, plants that had been operating in Eastern Poland since 1920 were taken over by the Soviet Union. In 1963, the first plant was started to exploit a very extensive deposit in White Russia. The CIS today has several large operations in the Urals and White Russia, giving it the largest capacity of all potash-producing countries.

In the United States, potash production began during World War I because the United States economy could no longer buy German potash fertilizers. Potassium salts were obtained from Searles Lake in California and in northern Nebraska. Most of these operations ceased production in the 1920s. After the discovery of potash deposits in the area of Carlsbad, New Mexico, a large number of potash works were founded since 1931 [23]. Since the foundation of the Canadian potash industry, which has the United States as its main outlet, the number of works in the Carlsbad region and their production rate have continually declined. New potash works were started in southern Utah at Moab in 1964 and at the Great Salt Lake in 1970 [24].

The most important potash deposit in North America was discovered during World War II in Saskatchewan, Canada. After initial problems due to its great depth and the presence of water-bearing overlying rock, which was difficult to deal with, several potash works were founded in the early 1960s, and today Canada is second only to the CIS as a potash producer. Two more potash plants were established in the 1980s in New Brunswick on the east coast of Canada.

In 1986, a potash plant began production at Sergipe in Brazil.

In 1974, a potash mine was opened in Yorkshire, United Kingdom.

Potash production in Palestine began on the north bank of the Dead Sea in 1931. Potash plants have now been producing since 1952 at the southern end on the Israeli side and since 1982 on the Jordanian side.

Potash ores are treated today by three basic processes: leaching – crystallization, flotation, and electrostatic treatment. Gravity separation is of minor importance because of the small density differences between salt minerals.

The oldest process is leaching – crystallization. In this process, salt solutions were originally cooled in open vessels. Vacuum cooling of the solutions in crystallizers was introduced in 1918 in the United States, so that much less energy was required and cooling times were reduced from days or weeks to minutes. Flotation was introduced in 1935 in the United States. This proved so efficient, especially for the treatment of sylvinite ores, that it is now the main potash treatment process worldwide. The electrostatic process was first used on a large scale in the German potash industry in 1974. It is now widely used in Germany for treating complex hard salts.

2. Potash Salt Deposits

2.1. Minerals

Potash salt deposits were formed by the evaporation of seawater [25]. Their composition is often affected by secondary changes in the primary mineral deposits. More than 40 salt minerals are now known, which contain some or all of the small number of cations Na^+, K^+, Mg^{2+}, and Ca^{2+}; the anions Cl^- and SO_4^{2-}; and occasionally Fe^{2+} and BO_3^{3-}, as well. Most of these are listed in Table 1 [26].

The more important salt minerals are halite, anhydrite, sylvite, carnallite, kieserite, poly-

Table 1. Principal salt minerals

Mineral	CAS registry no.	Formula	Crystal system/ crystal class	Refractive indices $n_\alpha : n_\beta : n_\gamma$ or $n_{\mathrm{opt.}} : n_{\mathrm{C}}$; optical activity	Density, g/cm^3	Hardness (Mohs)
Anhydrite	[*14798-04-0*]	$CaSO_4$	rhombic D_{2h}-mmm	1.570 : 1 .575 : 1.614 opt. +	2.96	3.8
Ascharite	[*13768-64-4*]	$Mg_2[B_2O_5] \cdot H_2O$	monoclinic C_{2h}-2/m	1.575 : 1.646 : 1.650 opt. −	2.70	3
Astrakhanite	[*15083-77-9*]	$Na_2Mg[SO_4]_2 \cdot 4\,H_2O$	monoclinic C_{2h}-2/m	1.483 : 1.486 : 1.487 opt. −	2.23	3
Bischofite	[*13778-96-6*]	$MgCl_2 \cdot 6\,H_2O$	monoclinic C_{2h}-2/m	1.495 : 1.507 : 1.528 opt. +	1.60	1.5
Boracite (stassfurtite: lumpy form of boracite)	[*1303-91-9*]	$Mg_3[Cl/B_7O_{13}]$	rhombic C_{2v}-mm 2	1.662 : 1.667 : 1.673 opt. +	2.95	7
Carnallite	[*1318-27-0*]	$KMgCl_3 \cdot 6\,H_2O$	rhombic D_{2h}-mmm	1.467 : 1.475 : 1.495 opt. +	1.60	2.7
D'Ansite	[*12381-13-4*]	$Na_{21}Mg[Cl_3(SO_4)_{10}]$	cubic Td-43 m	1.489	2.65	
Epsomite	[*14457-55-7*]	$MgSO_4 \cdot 7\,H_2O$	rhombic D_2-222	1.432 : 1.455 : 1.461 opt. −	1.68	2.5
Glaserite	[*13932-19-9*]	$K_3Na[SO_4]_2$	trigonal D_{3d}-3 m	1.491 : 1.498 opt. +	2.70	2.7
Glauberite	[*13767-89-0*]	$CaNa_2[SO_4]_2$	monoclinic C_{2h}-2/m	1.515 : 1.532 : 1.536 opt. −	2.85	3
Gypsum	[*13397-24-5*]	$CaSO_4 \cdot 2\,H_2O$	monoclinic C_{2h}-2/m	1.521 : 1.523 : 1.530 opt. +	2.32	2
Halite (rock salt)	[*14762-51-7*]	NaCl	cubic O_h-m3m	1.5443	2.168	2.5
Kainite	[*1318-75-2*]	$(KMg[ClSO_4])_4 \cdot 11\,H_2O$	monoclinic C_{2h}-2/m	1.494 : 1.506 : 1.516 opt. −	2.13	3
Kieserite	[*14567-64-7*]	$MgSO_4 \cdot H_2O$	monoclinic C_{2h}-2/m	1.518 : 1.531 : 1.583 opt. +	2.57	3.7
Koenenite	[*12252-18-5*]	$[Mg_7Al_4(OH)_{22}]$ $[Na_4(CaMg)_2Cl_{12}]$	trigonal D_{3d}-3 m	1.55 : 1 .58 opt. +	2.15	1
Langbeinite	[*14977-37-8*]	$K_2Mg_2[SO_4]_3$	cubic T-23	1.534	2.83	4.2
Leonite	[*15226-80-9*]	$K_2Mg[SO_4]_2 \cdot 4\,H_2O$	monoclinic C_{2h}-2/m	1.479 : 1.483 : 1.488 opt. +	2.20	2.7
Löweite	[*16633-52-6*]	$Na_{12}Mg_7[SO_4]_{13} \cdot 15\,H_2O$	trigonal C_{31}-3	1.495 : 1.478 opt. −	2.34	2.5 – 3
Mirabilite (Glauber's salt)	[*14567-58-9*]	$Na_2SO_4 \cdot 10\,H_2O$	monoclinic C_{2h}-2/m	1.394 : 1.396 : 1.398 opt. −	1.49	1.7
Polyhalite	[*15278-29-2*]	$K_2MgCa_2[SO_4]_4 \cdot 2\,H_2O$	triclinic C_1-1	1.548 : 1.562 : 1.567 opt. −	2.78	3 – 3.6
Rinneite	[*15976-45-1*]	$K_3Na[FeCl_6]$	trigonal D_{3d}-3m	1.588 : 1.589 opt. +	2.35	3
Schoenite	[*15491-86-8*]	$K_2Mg[SO_4]_2 \cdot 6\,H_2O$	monoclinic C_{2h}-2/m	1.461 : 1.463 : 1.476 opt. +	2.03	2.6
Sylvite	[*14336-88-0*]	KCl	cubic O_h-m3m	1.4903	1.99	2
Syngenite	[*13780-13-7*]	$K_2Ca[SO_4]_2 \cdot H_2O$	monclinic C_{2h}-2/m	1.501 : 1.517 : 1.518 opt. −	2.58	2.5
Tachhydrite	[*12194-70-6*]	$CaMg_2Cl_6 \cdot 12\,H_2O$	trigonal D_{3d}-3m	1.520 : 1.512 opt. −	1.67	2
Thenardite	[*13759-07-4*]	Na_2SO_4	rhombic D_{2h}-mmm	1.471 : 1.477 : 1.484 opt. +	2.67	2.7
Vanthoffite	[*15557-33-2*]	$Na_6Mg[SO_4]_4$	monoclinic C_{2h}-2/m	1.485 : 1.4876 : 1.489 opt. −	2.69	3.6

halite, langbeinite, and kainite. Gypsum occurs at the edges of salt deposits and in the overlying strata. Bischofite, tachhydrite, glauberite, thenardite, glaserite, and leonite occur additionally in some deposits [4, 27, 28].

Other minerals, not described in detail here, are useful in elucidating difficult geological questions with regard to the origin of salt deposits. In special geochemical investigations, small amounts (ppm) of Rb^+ and Cs^+ in place of K^+; Sr^{2+} replacing Ca^{2+}; Mn^{2+} replacing Fe^{2+}; Br^- replacing Cl^-; etc., are important [29, 30]. The individual minerals can be identified microscopically (grains or thin sections) and by X-ray analysis.

Potassium salt deposits always consist of a combination of several minerals (Table 2). The German term *Hartsalz* (hard salt) refers to the greater hardness of sulfate-containing potash minerals in potash deposits.

Table 2. Marine salt rocks: mineral constituents ∗

Salt rock	Main components	Secondary components
Rock salt	Na	A, Po, Ki, La, clay minerals, etc.
Anhydrite	A	Na, Dol, Mag, gypsum, Sy, C, clay minerals, borates
Carnallitite	C, Na	Ki, Sy, A
Sylvinite	Sy, Na	C, Ki, A
Hard salt		
kieseritic	Na, Ki, Sy	C, A, La, Po, borates
langbeinitic	Na, La	Ki, A, Sy, C
anhydritic	Na, Sy, A	Ki, C
Kainitite	Na, Kai	Ki, Sy, C
Bischofitite	Bi, C and/or Ta	Na, Ki, Po, A, borates
Tachhydritite	Ta, Bi	C, Na
Claystone	Clay minerals, quartz, mica, A, Dol, Mag	Na, Sy, C, Bi, Ta, La, coenenite, rinneite

∗ Abbreviations: A = anhydrite; Bi = bischofite; C = carnallite; Dol = dolomite; Kai = kainite; Ki = kieserite; La = langbeinite; Mag = magnesite; Na = halite; Po = polyhalite; Sy = sylvite; Ta = tachhydrite.

2.2. Geology of Potash Deposits

Potash deposits occur worldwide in almost all geological systems. The most important deposits were formed in the Devonian, Carboniferous, Permian, Cretaceous, and Tertiary periods [4, 25, 27, 31–37]. All major potash deposits are of marine origin. Bodies of seawater became isolated from the open ocean when bars formed under the water surface, and under arid climatic conditions, the seawater became concentrated, finally depositing the dissolved salts. The important feature is that exchange between normal seawater and concentrated salt solution generally does not occur. During sedimentation, the less soluble salts were deposited first, and the most soluble salts last. In most salt-forming sea basins, this process was repeated many times, resulting in cyclical salt formation [36, 38].

Table 3. Important potash deposits, mineral composition [a], geologic age [b]

Country	Geological age									
	Rect	Plei	Tert	Cret	Jura	Perm	Carb	Dev	Sil	Camb
Australia	B									
Brazil				SC						
Canada							S	SC		
CIS			KS		LKS	C		S		S
Congo				SC						
Germany						HS				
Ethiopia		SCK								
France			S							
United Kingdom						S				
Israel	B									
Italy			K							
Jordan	B									
Netherlands						C				
Spain			S							
Thailand				CS						
United States	B					SL	S	SC	S	

[a] S = sylvinite; C = carnallite; H = hard salt; L = langbeinite; K = kainite; B = brine.
[b] Rect = recent; Plei = Pleistocene; Tert = Tertiary; Cret = Cretaceous; Jura = Jurassic; Perm = Permian; Carb = Carboniferous; Dev = Devonian; Sil = Silurian; Camb = Cambrian.

A complete salt deposition cycle begins with basic carbonates (limestone, dolomite, and sometimes magnesite), followed by sulfates (gypsum and anhydrite), rock salt, and finally potassium and magnesium salts (Table 1). Intermediate layers of clay sediments are the result of repeated influxes of fresh water, which can lead to partial redissolution of the salt deposits. Also, eolian (airborne) transportation into the basin can occur. A recent example of a salt deposition basin is provided by the Kara-Bugas, a lagoon on the eastern side of the Caspian Sea.

In the course of the earth's history, the deposited salts underwent many changes, often leading to the formation of a modified mineral constitution. The original formations were affected by relatively low-temperature thermal

and hydraulic metamorphoses, and sometimes by volcanic action, producing the mineral compositions that exist today (Table 2) [39,40].

Rock movements have changed the original horizontal stratification in many salt deposits. Salt migration, folding, and upward movement to form diapirs have led to tilting of the potash layer, sometimes to an acute angle; to thinning; or to local accumulation. As a result of movements, salt deposits often came into contact with groundwater and were partially or completely redissolved.

Table 3 lists the most important potash salt deposits, showing the main potash minerals present, the geological systems, and the geographical locations. Table 4 shows the distribution of potash deposits, with estimates of the reserves [4,37,41].

Table 4. Minable potash reserves in units of 10^6 t K_2O

Canada, Saskatchewan (conventional mining only)	4500 – 6000
New Brunswick	60 – 80
United States	100 – 150
Brazil	10 – 40
Chile and Peru	30 – 50
Congo	ca. 20
Germany	400 – 800
United Kingdom	30 – 50
Italy	10 – 20
France	ca. 20
Spain	20 – 30
CIS	2000 – 3000
Dead Sea (Israel and Jordan)	100 – 200
China	10 – 100
Thailand	up to 160
Laos	up to 20
World *	7500 – 10 000

* If reserves only extractable by solution mining are included (particularly those in Saskatchewan, but also including those in the CIS), the figure for the total minable reserves increases by a factor of 4 or 5. Other deposits of potash salts are either unimportant or of minor local importance compared with the above figures. These exist in Australia, Ethiopia, Iran, Libya, Morocco, Poland, and Tunisia.

The concept of a reserve presupposes the possibility of economical extraction, which depends on the presence of a useful potash content together with usable quantities of other materials. Other important factors are the type of deposit, a uniform and usable seam thickness at a depth that is neither too great nor too low (water problems), economic workability, the possibility of solution extraction, and above all, proximity to consumers and profitability. Losses incurred during the extraction process (10 – 20 %) must be deducted from the total size of the reserve.

The two largest known potash deposits in the world, which are in Saskatchewan (Canada) and White Russia, are of Devonian origin. The Permian deposits (Germany, United States, and CIS) were for a long period the classical salt deposits and were the most important potash reserves, but these lost their economic importance after World War II.

However, the known potential of extractable potash deposits is so large that the world supply is guaranteed for many hundreds of years.

Cambrian. In the central part of the Angara-Lena Trough, ca. 800 km north-northeast of Irkutsk (CIS) in the vast eastern Siberian basin, an evaporite series more than 2 km thick contains a potash deposit. A sylvinite series up to 28 m thick is situated at a depth of 600 – 900 m in the interior of this Nepa potash basin. At the edges it merges into carnallitite, with over- and underlying carnallite – halite. The potash content of the best sylvinite seam, up to 13 m thick, is 19 – 30 % K_2O, with an extremely low $MgCl_2$ content [4, 27,37].

Silurian. In the center of the Michigan basin is a sylvinite zone having a total thickness of 28 m within a 900-m-thick evaporite series of the Salina group covering an area of 34 000 km^2. A solution-mining pilot plant (see Section 3.3) is in operation [4,37].

Devonian. *Canada.* The western Canadian basin between the Rocky Mountains and Hudson Bay accomodates eight Devonian salt-bearing units, of which the Prairie Evaporite is the most important both in extent and in economic importance, having four potash seams, partly sylvinitic and partly carnallitic. The upper limit of the Prairie deposit is inclined to the south, beginning at a depth of 600 m in the northeast and ending at >3000 m in the southwest. Of the total area (ca. 200 000 km^2) of the potash seams, ca. 50 000 km^2 is extractable by mining. Other areas can only be exploited by solution mining. The clay-bearing sylvinites, each 3 – 5 m thick, contain 25 – 30 % K_2O, but no potassium magnesium sulfate. At present, ten plants are in operation [4,27,37].

United States. The potash deposit described above, which is of great importance for Canada, stretches across the border to Montana and North Dakota. The potash-bearing area covers more than 30 000 km^2, but because it is at a depth of >1000 m and up to 3500 m, extraction by conventional mining is impossible [4, 37].

CIS. The potash deposits of Soligorsk and Starobinsk lie within the Pripyat marshes, 25 000 km^2 in area, ca. 120 km south of Minsk and 800 km from the Baltic harbor of Ventspils. Potash extraction began in 1963, and output reached 4.2×10^6 t of K_2O by 1980. The Upper Salt is 3200 m thick and contains ca. 60 potash zones in the upper half. These can be grouped into four workable potash-bearing formations. Each of these consists of interbedded layers of salt clays, rock salt, sylvinite seams, and some carnallitite. Only stage II (thickness: 1.8 – 4.4 m, K_2O: 17.7 %, insolubles: 5 %) and the lower part of stage III (thickness: up to 2.8 m, K_2O: 13.4 – 16.4 %, insolubles: 9 %) are extracted. Potash is mined at depths of 350 – 950 m, where level seams with only slight deformation exist [4, 28, 37].

Carboniferous. *United States.* The potash deposit of the Paradox formation at Moab, Utah, is extracted by Cane Creek Mine. The valuable sylvinite deposit is at present extracted by solution mining. The extent of the salt deposit is ca. 25 000 km^2, of which 15 000 km^2 is potash bearing. It contains about ten potash seams, of which four are workable, containing sylvinite (25 % K_2O) and some carnallitite. The salt has been formed by tectonic action into anticlines, sometimes diapiric, stretching from northwest to southeast [4, 27, 37]. Because the deposit is undulating, conventional mining is complicated and made hazardous by the presence of oil and gas.

Canada. In the eastern provinces of Canada on the Atlantic coast (i.e., New Brunswick, Nova Scotia, and Newfoundland) are a number of small sedimentary basins with evaporites of the Lower Carboniferous Windsor group. Gypsum, anhydrite, and slightly to strongly folded rock salt are present, along with several mainly tectonically bounded small structures containing potash salts. In New Brunswick a high-quality sylvinite seam occurs that contains 25 – 30 % K_2O up to 40 m thick. At present, potash is extracted by PCA (Potash Corporation of America) 7 km north of Sussex at a depth of 460 – 760 m, and by PMC (Potacan Mining Corporation) 25 km south-southwest of Sussex at a depth of 600 – 1000 m [4, 27, 37].

Permian. The Permian period was one of immense salt accumulation, which took place in three vast evaporite basins: (1) the Central European Basin, (2) the East European Basin, and (3) the American Midcontinental Basin, as well as numerous smaller sedimentation areas in Europe, Asia, North and South America, and the Arctic [27].

1) *The Central European Basin* extends from Yorkshire to Central Poland and Lithuania, and from the River Main in northern Germany to the northern part of the North Sea [42]. In the Zechstein, seven sedimentation periods are distinguished, of which the three lowest–the Werra, Stassfurt, and Leine series–are of economic importance for potash extraction.
Germany. In the Werra and Fulda areas, the Hessen and Thuringia potash seams of the Werra series are mined (hard salt and carnallitite in level deposits at a depth of 400 – 1000 m with a thickness of 2 – 5 m, K_2O: 9 – 12 %). The Stassfurt potash seam of the Stassfurt series is mined in the Harz – Unstrut – Saale area (hard salt and carnallitite at a depth of 500 – 1000 m and a thickness of 5 m, K_2O: 20 %). The potash seams Ronnenberg and Riedel of the Leine series are mined in the Hanover area in salt diapirs (sylvinite in inclined deposits, depth: 350 – 1500 m, thickness: 2 –40 m, K_2O: 12 – 30 %). Hard salts of the Stassfurt series are at present extracted to only a minor extent. Finally, potash is mined on the Massif of Calvörde near Zielitz (depth: 600 – 900 m, Ronnenberg sylvinite inclined at <18 – 25°, thickness: up to 10 m, K_2O: 14 – 20 %) [4, 37].
United Kingdom. In Yorkshire, a level deposit of carnallitic sylvinite is extracted, which correlates with the German Riedel seam both petrographically and stratigraphically (average thickness: 5 m, K_2O: ca. 27 %, depth: 1100 m) [4] *The Netherlands.* In northeast Holland, the carnallitic potash

salts of the Leine series have been drilled at various places. The plans are to extract carnallitite by solution mining [4].
Poland. The Zechstein series in Poland correlates with that in Germany. The carnallitic formations of the Stassfurt seam have been drilled, and the Leine deposits also contain some carnallitite and sylvinite.

2) *The Eastern European Basin* extends from the Barents Sea to the Caspian Sea and covers an area of 1.5×10^6 km^2.
CIS. Potash deposits are known in several parts of the western foothills of the Urals from the Pechora Depression to the Orenburg District, and in the region of the Caspian Depression. The most important reserve is in the area of Solikamsk and Berezniki on the upper Kama, 200 km north of the town of Perm. It has been mined since the late 1920s. Potash seams occur in the upper part of a salt series >400 m thick that is part of the Kungurian. These seams are sylvinitic below and sylvinitic – carnallitic above, interlayered with rock salt. The lower section, consisting of four sylvinite seams of total average thickness 20 m, with a K_2O content of 15 – 20 %, is mined at a depth of 300 – 500 m in predominantly level deposits [4, 27, 37].

3) *The Midcontinental Basin of North America* stretches from North Dakota to Texas but is broken up into numerous smaller basins. Horizons of rock salt are present in almost the whole of the Permian system. The deposition of potash in large quantities apparently took place only during a small part of the Upper Permian period in a subregion of the Delaware basin of New Mexico and Texas [27].
United States. In Carlsbad, New Mexico, sylvinite and langbeinite are mined in the 50 –140-m-thick McNutt potash zone of the Salado formation in five out of a total of eleven potash seams up to 3 m thick in a level deposit at a depth of 250 – 575 m. The sylvinite contains 12 – 30 % K_2O, and the langbeinite 7 – 9 % K_2O. Massive and disseminated polyhalite, carnallite, kieserite, kainite, leonite, glauberite, and thenardite are also present [4, 27, 37].

4) *Brazil.* In the Amazon basin at Nova Olinda, 150 km south of Manaus, drilling of a level deposit of Lower Permian sylvinite at a depth of 1000 – 1150 m has been carried out. The thickness is up to 4.5 m, and the K_2O content 17 %. Extractable reserves amount to ca. 35×10^6 t of K_2O [4, 37].

Jurassic. *Turkmenistan/Uzbekistan.* Up to six seams of sylvinite can be found at a depth of 200 – 1200 m in the Gaurdak formation in the areas of Gaurdak – Tyubegatan, Kugitang, and Okuzbulak; and carnallite may be found at Karlyuk. The seams of sylvinite, up to 6 m thick, contain 12.5 – 25 % K_2O, sometimes with as much as 12 % insoluble material. A deposit at Karshi in Uzbekistan is to be exploited, and solution mining on a pilot scale has been started at Karlyuk [4, 37].

Cretaceous. Along the east coast of Brazil and the west coast of Africa from Gabon to Angola, Lower Cretaceous potash-bearing evaporites occur. Their similar formations indicate the existence of a common sedimentation area before the two southern continents began to drift apart during the tectonic movements of the Cretaceous period [4, 37].

Brazil, Sergipe. The Taquari-Vassouras deposit of the Muribeca formation has been mined since 1985. It has a basal layer of tachhydrite, above which is a thin layer of carnallitite and two seams of sylvinite, 10 m and 8 m thick, containing 15 % K_2O, with a rock salt bed 4 – 10 m thick between the two seams. The reserves of sylvinite amount to 525×10^6 t, with $16\,000 \times 10^6$ t of carnallitite [4, 37].

Congo. Salt deposits have been formed in ten sedimentation cycles, of which the second highest contains four sylvinite seams. The other cycles contain large masses of carnallitite, and, as in Sergipe, tachhydrite. Mining of the sylvinite in the third layer (depth of 370 m, 3 – 16 % K_2O) began in 1969 but ceased in 1976 when the mine became flooded in the space of two days [4, 37].

Thailand. The lower part of the Maha-Sarakham formation below the Khorat Plateau northeast of Bangkok consists of evaporites with a potash sequence up to 90 m overall thickness at depths of 90 – 530 m. The K_2O content of the carnallitite, which can be 15 m thick, is usually ca. 10 %, but can occasionally reach 14 %. Sylvinite, with a maximum thickness of 3.7 m and a K_2O content of 18 %, has been found in only a few of the 48 exploration boreholes, with tach-

hydrite in some locations. Test mining has begun in the concession areas, which cover 300 km^2 [4, 37].

Tertiary. *France.* The Lower Oligocene deposit in Alsace contains two sylvinite seams in a marl – rock salt series. The upper layer has a thickness of up to 2 m and contains 19 – 25 % K_2O; the lower, up to 5.5-m-thick layer, with 15 – 23 % K_2O, also contains 15 % insolubles (clay, anhydrite, and dolomite). Mining is carried out at comparatively high rock temperatures at a depth of 500 – 1000 m in flat or slightly inclined seams that have been disturbed by faults [4, 37].

Spain. Deposits are located in two areas of the Ebro basin. In Catalonia and Navarra, potash salts lie above rock salt. These deposits are up to 10 m thick in Catalonia and up to 15 m in Navarra. Above this occurs an interbedded deposit of rock salt, carnallitite, marl, and anhydrite. Only the sylvinite seams A and B are mined. These are up to 4 m in total thickness at a depth of 1020 m, some deposits being level and some inclined. The crude salt contains 12.5 – 14 % K_2O [4, 37].

Italy. The Miocene evaporites starts with an anhydrite – polyhalite – halite series followed by rock salt with up to six kainite interbeds and a roof of kieserite – sylvinite. Since 1976, only the kainite seams, which are up to 32 m thick, and contain 12 – 14 % K_2O, have been mined at ca. 800-m depth in seams inclined at 25 – 65° [4, 37].

CIS. In the eastern Galician foredeep of Carpathia, potash salts have been known since 1804 (mainly langbeinite and kainite, with a little sylvinite) and have been mined since 1864 at Stebnik and Kalusch. The individual lenticular deposits occur in five horizons of 1 – 5 m thickness, the kainite containing 10 – 12 % K_2O and the sylvinite 8 – 19 %. The deposit is inclined at 45° [4].

Pleistocene. In the Danakil depression, Ethiopia, is a salt basin of ca. 40 × 100 km, 100 m below the surface. Boreholes have revealed a potash zone containing sylvinite, carnallitite, and kainite at a depth of ca. 740 m [4].

Salt Lakes and Subterranean Brines. *Israel and Jordan.* Potassium chloride, magnesium chloride, rock salt, soda, chlorine, and bromine are obtained from the Dead Sea [4, 37].

United States. Potassium and sodium sulfate are extracted from the Great Salt Lake [4, 37]. Potassium and magnesium chloride are obtained from subterranean brines from the Great Salt Lake Desert [4]. Potassium chloride and sulfate, sodium borate, and boric acid are obtained from subterranean brines in Searles Lake [4, 37].

Australia. Langbeinite is obtained from subterranean brines at Lake McLeod.

China. Tsaerhan Lake, a dry lake, ca. 1100 km northeast of Lhasa, the largest salt lake in the Chaidamu basin, appears to be the most important potential reserve of potassium chloride in China [4, 37].

3. Mining of Potash Salts

3.1. Shaft Mining

Mineral salts are very soluble, and therefore any flow of water into the mine from the overlying strata, which are normally water bearing, must be prevented. This causes difficulties when sinking a shaft, and severe accidents have been caused by influxes of water. The freezing technique, especially deep freezing, is comparatively safe and is used to sink most mineshafts. The shaft itself is protected from water-bearing rocks by the use of tubbings, which are segments bolted together to form rings and are usually made from cast iron or steel-reinforced concrete. The shafts generally have a diameter of 5 – 7 m, and the depth can exceed 1000 m.

3.2. Extraction, Conveying, and Haulage

In most potash mines, the salt is mined from subhorizontal deposits. Generally, rooms are created by removing the salt, and pillars are left between these to prevent the cover rocks from collapsing. This enables an extraction rate of 25 – 60 % to be achieved. For cost reasons the mined-out rooms are not backfilled. In some mines, the total ore is extracted, which causes substantial subsidence of the overlying strata (Alsace).

In steeply dipping deposits (e.g., in the salt domes of northern Germany), roof mining was

originally carried out. This was later replaced by floor mining and then by funnel mining, which is now being used increasingly in numerous variations [43]. Entry drifts are driven one above the other at intervals of 15 – 20 m, and the remaining potash salt is mined by sublevel stoping. Material loosened by explosives falls via the lowest funnel-shaped region into the main haulage level underneath. The mined-out room, 100 – 250 m in height, is usually backfilled with salt waste after mining. Funnel mining is much safer and cheaper than roof mining, because both ore stoping and backfilling take place under gravity, and the mined room need not be reentered. During drilling of the blast holes, the miners are protected by horizontal pillars between the sublevels.

Drilling and blasting operations are carried out with the help of trackless vehicles. Large holes with a diameter of ca. 40 cm and a length of 7 m are drilled by mobile drilling equipment, and another mobile drilling rig is used to drill blast holes around the larger hole in a predetermined pattern. The explosive (generally ammonium nitrate with addition of oil) is brought to the workplace in tanks carried by diesel vehicles and is blown into the shot holes by compressed air.

In predominantly horizontal sylvinite deposits, the most frequent method of extraction is by cutting with heavy machinery (180 t per unit) at high cutting rates. These machines produce material suitable for transport by conveyor belt, enabling continuous extraction in 30 – 60-m sections.

Extraction by borers with two to four rotors is mainly used to produce long pillars in a room and pillar system (Canada), the length of the chambers being as much as 1000 m. In longwall mining, however, which is usually carried out as a caving operation, two or three drum shearers are used (e.g., in Alsace and the CIS). These machines enable daily outputs between 1500 and 4000 t to be achieved.

The recovered material is transported by trackless diesel or cable-fed electric front-end loaders with a capacity of 15 t. Transport along underground roads is increasingly by conveyor belts but also by electric or diesel trains with 30-t-capacity wagons or by dumper trucks with a hopper capacity of 40 t. For transportation by conveyor belt, the material must first be broken into suitably sized pieces.

The network of roadways for conveying, traveling, and ventilation usually extends >100 km in large potash works. A radio communication system is generally used.

Since the introduction of very heavy machinery and diesel-powered vehicles, extra attention must be given to ventilation. Powerful fans supply fresh air at up to 30 000 m^3/min.

In hoisting shafts, the skips have a capacity of up to 25 t. These operate automatically and supply large intermediate storage bins in the filling station, so that the continuously operating manufacturing plants can be supplied with ore at a steady rate. Average daily throughput can be ca. 30 000 t of ore.

Improvements in mining methods and the introduction of new techniques have enabled the output per worker underground to be increased considerably. In Germany during 1965 – 1974 the output of potash ore increased by 20 %, despite a >50 % reduction in the number of employees.

3.3. Solution Mining

Solution mining is an alternative to conventional mining for the extraction of potash ore. The advantages of this method are that the high expense of sinking a shaft is eliminated and reserves can be exploited where conventional mining is impossible (e.g., at great depth). Also, this method can be used where existing mine workings are available but conventional mining methods are no longer feasible, even though extensive reserves may still exist.

Since 1964, Kalium Chemicals in Saskatchewan has operated a plant in which brine extracted from a potash deposit at a depth of ca. 1500 m is used to produce very pure potassium chloride. The process is based on a series of patents [44], but details have not yet been published. Water or an unsaturated solution of potassium chloride is passed through a system of boreholes into the potash seam, which is 20 –25 m thick; potassium chloride and sodium chloride are dissolved. The almost saturated solution is pumped to the surface and fed to the production plant. Rock salt above the potash seam is protected from dissolution by an oil

or air cushion (see Fig. 1). The brine produced passes through multiple-effect evaporators in which sodium chloride crystallizes. Potassium chloride is then crystallized in a series of vacuum coolers [45].

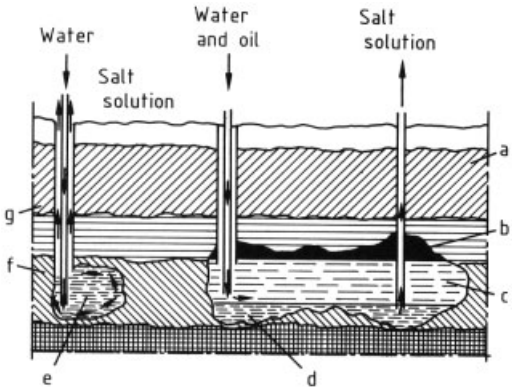

Figure 1. Solution mining
a) Overlying rock; b) Oil cushion; c) Partly unsaturated salt solution; d) Saturated salt solution; e) Cavity produced by dissolution; f) Potash deposit; g) Sodium chloride layer

In Utah (United States), where it was necessary to terminate a conventional mining operation due to severe geological and technological problems, operation was resumed in 1972 by using solution mining. Shafts and underground cavities were flooded, providing a route for the brine formed when water was fed in to dissolve the salt. The brine was passed into surface ponds where solar evaporation caused a mixture of potassium chloride and sodium chloride to crystallize, which was treated in a flotation plant to produce 60 % K_2O potassium chloride [46].

In Canada a conventional mining operation was also converted to the solution mining method after penetration of water led to complete flooding of the mine and forced operations to cease. Water is passed via boreholes into the flooded mine and is converted into a concentrated brine, which is withdrawn and cooled in a pond during the very cold Canadian winter. The potassium chloride that crystallizes is recovered and processed to give a salable product [47].

In the former German Democratic Republic, extensive research into the solution mining of carnallite or potassium chloride from carnallitic deposits has been carried out. An experimental plant with a KCl capacity of ca. 50 000 t/a was operated for a long period (see Section 5.2.3) [48].

4. Treatment of Potash Ores

4.1. Intergrowth and Degree of Liberation [49, 50]

The salt minerals in potash ores are intergrown to varying extents. Before the minerals can be separated and the useful components recovered, the ore must be sufficiently reduced in size so that individual components are accessible to the processing method to be used. In the hot leaching process, sylvite is extracted, and therefore it must first be liberated (i.e., it must not be occluded inside other minerals). To achieve this, it is sufficient to break down the ore to a particle size of 4 – 5 mm or less.

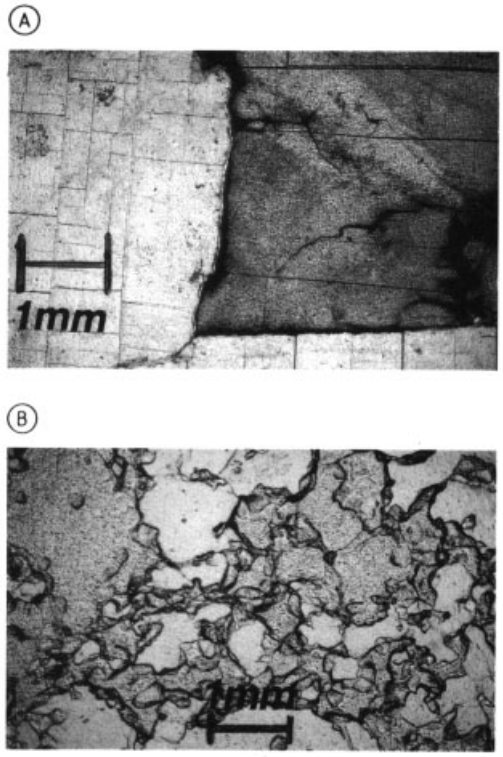

Figure 2. Thin sections of sylvinite ores
A) Coarsely intergrown (potash works in Lanigan, Canada); B) Finely intergrown (Kaliwerk Sigmundshall, Germany)

For the mechanical treatment processes (i.e., flotation, electrostatic treatment, and gravity separation), liberation of the minerals must be complete (i.e., individual grains must consist as much as possible of pure minerals). The extent to which the minerals in the potash ore are intergrown can vary greatly from deposit to deposit (see Fig. 2), which means that the crude salt must be size-reduced to varying degrees before further processing.

The degree of intergrowth of individual minerals can be determined by examination of a thin section. A photomicrograph shows the sizes and shapes of individual minerals in relation to each other [51]. The disadvantage of this method is that a thin section gives only a two-dimensional view of a relatively small region of the salt mineral, and a very small sample of the substance is examined. The three-dimensional arrangement of minerals present and their distribution are not observed. For these reasons, and also because of the high cost of preparing thin sections, this method is now of only minor importance for the industrial processing of potash ores.

More usually, the degree of liberation in size-reduced samples is determined. The degree of liberation of a mineral means the percentage ratio of fully liberated mineral particles to the total content of the mineral in the sample.

Degree of liberation:

$$\frac{\%\,\text{Free mineral}}{\%\,\text{Free} + \%\,\text{Intergrown mineral}} \cdot 100$$

The degree of liberation L depends not only on the grain size achieved by the grinding operation but also on the type of grinding. It can be given for particular ranges of grain size so that the way in which it depends on grain size can be determined, or it can be expressed as an average for the total sample (the integral degree of liberation $\bar{L}$). The arithmetic mean of the degree of liberation of each range of grain sizes is obtained from the following formula:

$$\bar{L} = \frac{\sum_{i=1}^{n} L_i p_i a_i}{100a}$$

where p_i is the mass fraction of the ith size range in percent; a_i the percentage of useful mineral in the ith size range; and a the percentage of useful mineral in the total sample. Apart from the fraction of useful mineral a, which is determined by chemical analysis, the liberated, nonintergrown fraction of useful mineral in each grain size range must be determined. Two methods for doing this are possible:

1) Visual estimation (by counting under the microscope) of the proportion of intergrown particles
2) Heavy-medium separation of the free or almost free grains

The first method is easy to apply to salt minerals because the intergrowth effects are readily recognized owing to the transparency of the grains. For coarse-grained materials such as those usually found in ground products from coarsely intergrown potash ores, this process cannot be used.

The separation of free or nearly free mineral grains from a size fraction is carried out by using heavy liquids of appropriate density, such as tetrabromoethane – toluene mixtures. A float–sink separation is carried out to determine the fraction of free mineral grains. This method cannot be used for salt particles with a grain size <0.5 mm because of the agglomeration of fine grains. In this case, the method of counting under the microscope must be used, or the degrees of liberation of the coarser size ranges must be extrapolated.

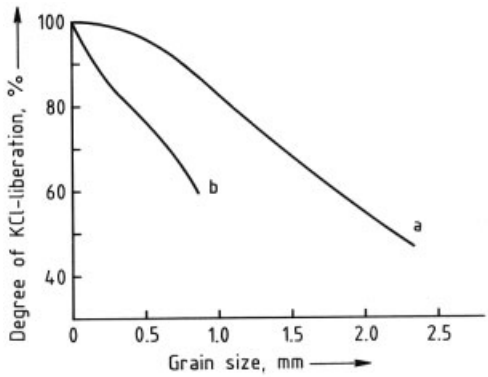

Figure 3. Liberation curves of sylvinite ores
a) Coarsely intergrown sylvinite ore (from New Brunswick, Canada); b) Finely intergrown sylvinite ore (from a north German salt deposit)

The results are expressed in liberation curves, which give the degree of liberation as a function of grain size (Fig. 3).

4.2. Grinding [52,53]

Potash salts are easily size-reduced. Therefore, fines may be formed, which can cause problems in later stages of processing. Great care must be taken in selection of the equipment for various stages of grinding.

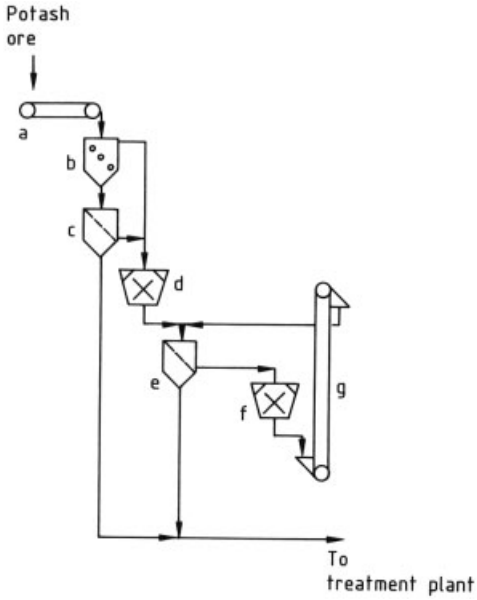

Figure 4. Two-stage grinding system
a) Conveyor belt; b) Grid; c) First screening stage; d) First grinding stage; e) Second screening stage; f) Second grinding stage; g) Bucket elevator

The maximum grain size to which the potash ore is ground depends on the processing method used and the degree of intergrowth of the ore. For the hot leaching process, an upper grain size limit of 4 – 5 mm is adequate. For mechanical processing, the ore must be ground to a degree of liberation > 75 %. For German sylvinite ores and hard salts, this is achieved by grinding to a maximum grain size of 0.8 – 1.0 mm. For the much coarser sylvinite ores of New Mexico, a maximum grain size limit of 2.4 mm is sufficient. The sylvinite ores of Saskatchewan are even more coarsely intergrown, so that size reduction to < 9 mm would give adequate liberation. However, such large crystals cannot be treated by conventional flotation, and the material is therefore normally ground to < 2.3 mm. One large Canadian manufacturer produces a coarse crystalline product by grinding the potash ore to < 9 mm, removing grains < 1.7 mm, and treating this fine material by conventional flotation. The remaining fraction (1.7 – 9 mm) is treated in a heavy-medium separation plant (see Section 5.5), giving a product with 60 % K_2O (95 % KCl) and a size distribution of the granular commercial grade.

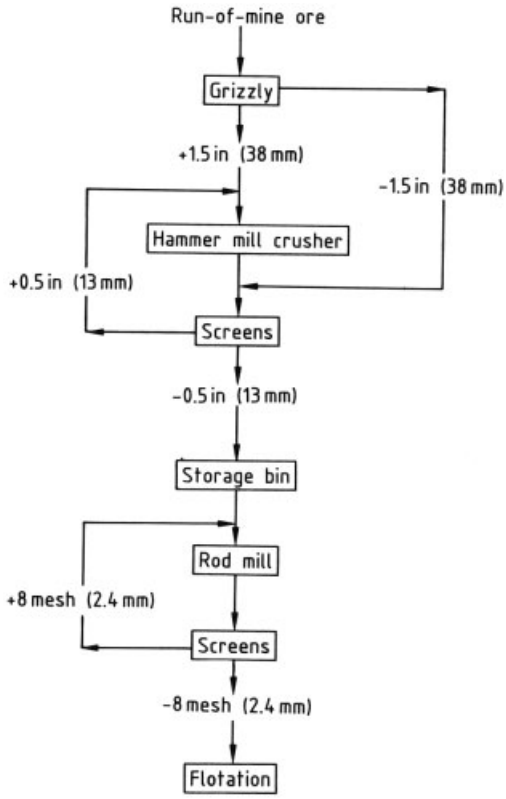

Figure 5. Wet grinding and screening of coarsely intergrown potash ore for flotation [5]
Reprinted with permission of the Society for Mining, Metallurgy, and Exploration, Inc.

The preliminary size reduction of potash ore is carried out by underground mobile crushers, usually in the vicinity of a mining operation. When the ore is mined by continuous mining this gives sufficient size reduction for it to go directly to the haulage line. Further size reduction to the grain size required for processing is carried out in two stages on the surface. An initial size reduction with impact or hammer mills (→ Size Reduction, Chap. 3.4.1.) is always carried out to produce 4 – 12-mm particles, depending on the raw material and processing method

to be used. A coarse grinding plant usually includes two grinding and screening stages (see Fig. 4).

The final fine grinding stage is carried out either by wet grinding in rod mills or by dry grinding with rollers or impact crushers (Figs. 5 and 6). Wet grinding with rod mills in a recirculating system with classification is standard for most flotation plants. Classification is by spiral classifiers, vibrating screens, curved screens, or cyclones. Wet screening produces only a small amount of fines and has the further advantage of providing a scrubbing effect that facilitates the removal of clay from clay-bearing ores; a necessary step before the flotation process.

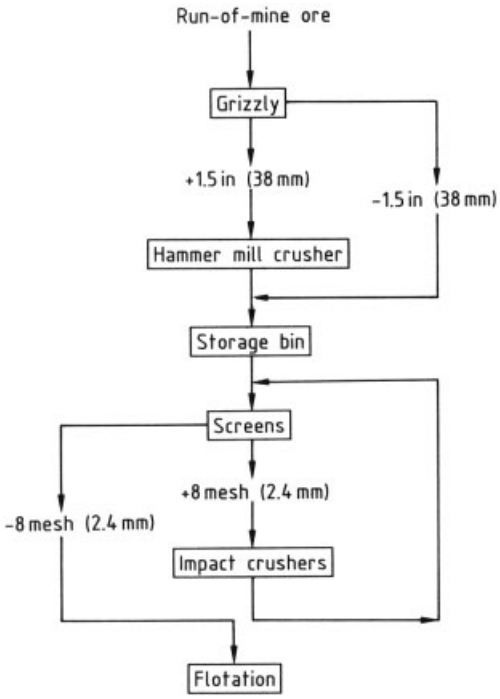

Figure 6. Dry grinding and screening of coarsely intergrown potash ore for flotation [5]
Reprinted with permission of the Society for Mining, Metallurgy, and Exploration, Inc.

The production of a fine product by dry grinding is rarely used in flotation plants. It is indispensable as a preparation for electrostatic processing, which is not compatible either with the changes to the mineral surfaces caused by aqueous solutions or with excess moisture. The grinding operation must be carried out carefully to give a product low in fines. Roller mills and impact mills can be used. Roller mills have the disadvantage that throughputs are relatively low and maintenance costs are high. Although this was the preferred method in the early days of electrostatic processing owing to its gentle grinding action, it has now been largely replaced by impact grinding, which is also used for size reduction of the middle product from electrostatic separation [54].

5. Potassium Chloride

Potassium chloride [*7447-40-7*], KCl, mineral name sylvite [*14336-88-0*], forms colorless nonhygroscopic crystals. It occurs in many salt deposits (see Section 2.2) mixed with halite and other salt minerals. Natural sylvite is usually opalescent or milky white, as are crystals obtained from an aqueous solution. Sylvite is often colored red by hematite. With magnesium chloride it forms the double salt carnallite [*1318-27-0*], $KCl \cdot MgCl_2 \cdot 6\,H_2O$, which is also commonly found in salt deposits. Potassium chloride is produced in large quantities from mined potash ores and from salt-containing surface waters. More than 90 % of the potassium chloride produced is used in single- or multi-nutrient fertilizers, either directly or after conversion to potassium sulfate (→ Fertilizers, Chap. 3.). The remainder has various industrial uses and is the raw material for the manufacture of potassium and its compounds.

5.1. Properties

Potassium chloride crystallizes in the cubic system, usually as actual cubes. Some physical properties described are as follows:

Relative molecular mass	74.55
Melting point	771 °C
Crystal system and type	cubic 0_h^5
Refractive index n_D^{20}	1.4903
Density	1.987 g/cm^3
Specific heat c_p	693.7 J kg^{-1} K^{-1}
Heat of fusion	337.7 kJ/kg
Enthalpy of formation ΔH^0	− 436.7 kJ/mol
Entropy S^0	82.55 J mol^{-1} K^{-1}
Dielectric constant (at 10^6 Hz)	4.68
Thermal coefficient of expansion (15 – 25 °C)	33.7×10^{-6} K^{-1}

Solubilities in water at various temperatures appear in Table 5, and the phase diagram of the system $KCl-H_2O$ is shown in Figure 7.

Table 5. Solubility of potassium chloride in water (g/100 g) [16]

Temperature, °C	0	10	20	30	40	50
Solubility	28.1	31.2	34.2	37.2	40.2	43.1
Temperature, °C	60	70	80	90	100	
Solubility	45.9	48.6	51.3	53.8	56.2	

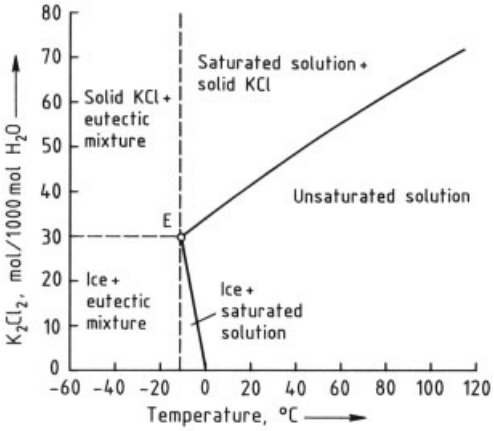

Figure 7. Solubility curves for potassium chloride in water
E = cryohydric point : Ice – potassium chloride solution

In the system $KCl-H_2O$, the only solid phases formed are KCl and ice. The cryohydric point (ice + KCl) is $-10.7\,°C$ (29.7 mol K_2Cl_2/1000 mol H_2O). The boiling point of the saturated solution at 1.013 bar is 108.6 °C (71.6 mol K_2Cl_2/1000 mol H_2O).

5.2. Production by Crystallization from Solution

5.2.1. Phase Theory

The salt deposits were formed by the evaporation of seawater, which contains the principal ions Na^+, K^+, Mg^{2+}, Ca^{2+}, Cl^-, and SO_4^{2-}. With water, these ions constitute a six-component system. The concentration of Ca^{2+} in the most interesting region of the system is negligibly small, so that the system reduces to quinary. It is nevertheless very complicated with, 23 different salts being formed between 0 and 100 °C, depending on the temperature and the ratio of concentrations of the components.

Understanding how the salt deposits were formed and how they behave in dissolution processes requires knowledge of the solution equilibria.

The theoretical foundations for this were laid by VAN'T HOFF et al., who between 1896 and 1906 investigated the formation of oceanic salt deposits [55]. Many investigators have continued this work up to the present.

J. D'ANS critically evaluated all published data up to 1933, expressing his results in graphical form [56]. In the same book, he described experimental methods for determining solution equilibria and gave recommendations for the graphical representation of experimental results [57].

Much of the equilibrium data published up to 1967 are given in [58, 59].

In the years following World War II, in the Kaliforschungs-Institut (Potash Research Institute) of Hanover, AUTENRIETH carried out comprehensive and detailed research into the stable and metastable equilibria of most relevance to potash production (especially from hard salt), giving the results in a form suitable for practical application [60–66].

The intensive investigations carried out into the quinary system make it the most thoroughly investigated system with more than four components. However, only parts of the system that are of most relevance to potash production have been thoroughly investigated. An obstacle in the application of equilibrium data to practical problems is that such a complex system is very difficult to represent in a two-dimensional diagram. However, by fixing parameters, working with projections on a plane, and using diagrams showing lines of constant parameters, even nonexperts can work with them.

The best-known region is that in which the solutions are saturated with sodium chloride. This is also the most important region for potassium chloride manufacture, both by the hot leaching process and by flotation, because in both cases a solution saturated with sodium chloride is used. The most important part of the so-called NaCl saturation space at 25 °C is shown in Figure 8 as a three-dimensional view.

Each point in the interior of this space corresponds to a solution saturated with NaCl, in which the concentrations of $MgCl_2$, K_2Cl_2, and

$MgSO_4$ are given by the distance of the point from the axes. For practical reasons, the concentration figures are given in moles per 1000 mol of H_2O, and the concentrations of KCl and NaCl are given in double moles. If one or more of the salt concentrations increase, and if the saturation concentration of another salt is exceeded, this salt separates out. Its identity depends on the concentration ratio in the solution. Such two-salt solutions (saturated with two salts) lie on two-salt surfaces, which form the boundaries of the saturation space. Of the twelve other salts whose saturation spaces form the boundary of the NaCl saturation space, seven can be seen on the figure because their two-salt surfaces lie in the concentration range of Figure 8.

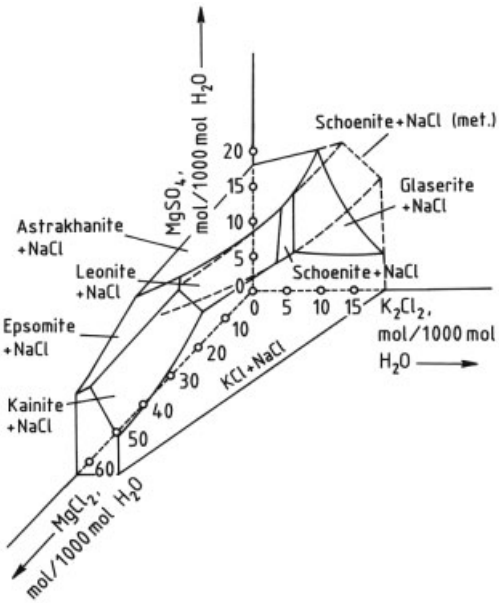

Figure 8. Three-dimensional view of the quinary system (saturated with NaCl) at 25 °C with 0 – 65 mol $MgCl_2$/1000 mol H_2O, showing stable and metastable regions

The KCl – NaCl two-salt surface on the front side of Figure 8 is of special significance for the potash industry, because it allows solutions to be made up that are saturated with both salts at 25 °C. Such solutions are important in the treatment of sylvinitic potash ores.

In many instances, a salt does not crystallize spontaneously when its concentration exceeds that indicated on a two-salt surface. Instead, highly supersaturated solutions are formed, which can remain stable for hours or days, depending on temperature and composition. These supersaturated solutions are termed metastable. They become stable saturated solutions by crystallization of a salt. In Figure 8, continuations of the stable schoenite – NaCl two-salt surface to the right and left into the unstable region are shown as broken lines. In potash manufacture, stable solution equilibria are seldom attained. This is particularly true for hard salt processing in which NaCl-saturated solutions with high $MgSO_4$ content often lead to the undesired crystallization of double sulfates such as schoenite, leonite, langbeinite, and glaserite. Here, the rates of dissolution, nucleation, and crystallization of these salts as a function of temperature and composition of the solutions are especially important [64–66].

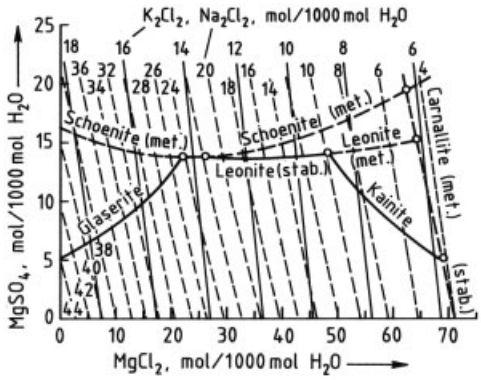

Figure 9. 25 °C isotherms of the quinary system saturated with NaCl
Stable (—) and metastable (– – –) surfaces saturated with KCl and NaCl. Concentrations of K_2Cl_2 (—) and Na_2Cl_2 (– – –) are indicated by lines of equal concentration

For practical application of equilibrium data, the boundary surface of the NaCl saturation space is projected, for example, in the direction of the K_2Cl_2 axis in the $MgSO_4 - MgCl_2$ plane. In Figure 9, the two-salt surface NaCl – KCl for the stable and metastable 25 °C isotherm is shown. The indicated K_2Cl_2 and Na_2Cl_2 lines of constant concentration enable the complete composition to be read off for each of the solutions shown here. By using this diagram, the dissolution and crystallization processes possible in this part of the quinary system can be described quantitatively. To control the crystallization of

potassium chloride from 90 °C solutions in an industrial plant, for example, the most important boundary surfaces of the 90 °C isotherm of the system (Fig. 10) are additionally required.

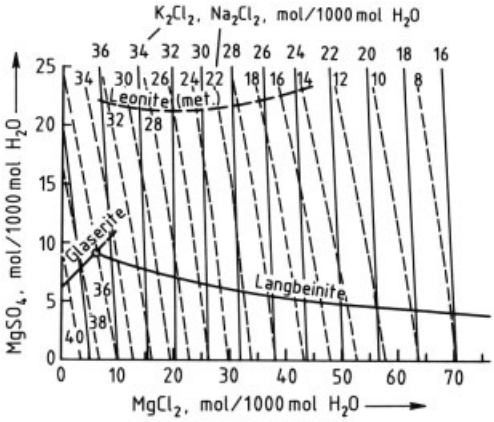

Figure 10. 90 °C isotherms of the quinary system saturated with NaCl
Stable (—-) and metastable (– – –) surfaces saturated with KCl and NaCl. Concentrations of K_2Cl_2 (—-) and Na_2Cl_2 (– – –) are indicated by lines of equal concentration

If any of the components of the quinary system are present in such small quantities that they have a negligible effect on the process, the mathematical treatment can be simplified by dealing only with the remaining subsystem. The following subsystems are of importance:

1) Na^+, K^+, Mg^{2+}, Cl^-, and H_2O with NaCl saturation (see Fig. 17) for the conversion of carnallite into potassium chloride and bischofite
2) K^+, Mg^{2+}, Cl^-, and H_2O (see Fig. 16) for the decomposition of carnallite
3) Na^+, K^+, Cl^-, and H_2O [57] for the selective dissolution of NaCl (e.g., from crystalline product obtained from hard salt in the hot leaching process)
4) K^+, Mg^{2+}, Cl^-, SO_4^{2-}, and H_2O [61] for the production of potassium sulfate and potash–magnesia
5) Na^+, Mg^+, Cl^-, SO_4^{2-}, and H_2O [57] for the production of thenardite or Glauber's salt

5.2.2. Hot Leaching Process

The hot leaching process is the oldest industrial process used to produce potassium chloride from potash ore. It was first used in 1860 in Stassfurt and since then has been developed further in Germany, where it is still the dominant process. It is especially suitable for treating very finely intergrown ores or ores that contain other salt minerals or insoluble minerals in addition to the sylvite and halite. It enables a high-purity product with a uniform grain size to be produced. In many plants, especially in Canada, where flotation is the main production process, small hot leaching plants are also operated, in which the product fines (<0.2 mm) are recrystallized, or potassium chloride is extracted and crystallized from the flotation residues or thickened clay slurries. These procedures give a considerable improvement in total yield and result in a very pure, completely water-soluble product.

Two different processes are used, depending on the composition of the ore. In the *sylvinite* hot leaching process, the other salts present in addition to KCl and NaCl play only a minor role in the process solutions. In *hard salt* leaching, process solutions contain appreciable amounts of $MgCl_2$ and $MgSO_4$. In the case of carnallite-containing hard salts, preliminary carnallite decomposition must be carried out (see Section 5.2.3.) if the amount of carnallite present exceeds a critical level.

The different solubility properties of sodium chloride and potassium chloride are shown in Figure 11. The solubility of potassium chloride is lower in hard salt solutions than in sylvinite solutions. The difference between the potassium chloride contents of saturated solutions at low and high temperatures is less for solutions of hard salt than for solutions of sylvinite, so that the amount of potassium chloride that can be crystallized from a given amount of solution is smaller, which has a marked effect on the energy requirement. Furthermore, an important difference between the two solution types is that the solubility of sodium chloride in sylvinite solutions decreases with increasing temperature, whereas it increases in hard salt solutions. This is apparent in Figure 11B, which shows the behavior of process solutions in a carnallitic hard salt plant with a magnesium chloride content of ca. 240 g/L. This dependence of the solubility of sodium chloride on temperature and magnesium salt content explains why the sodium chloride

contents of crystallized products differ, depending on whether they came from the treatment of sylvinite or hard salt.

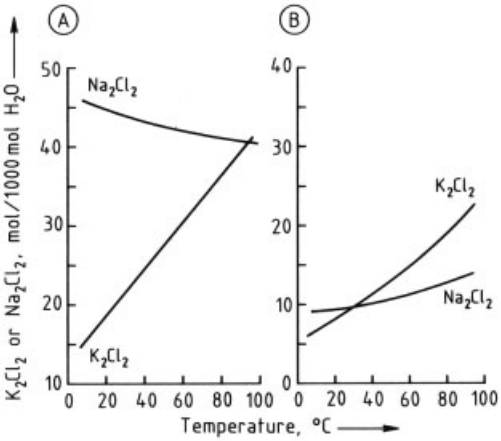

Figure 11. Solubility curves for KCl and NaCl (schematic) A) Sylvinite leaching (with nonevaporative cooling); B) Hard salt leaching

The Process (Fig. 12). The potash ore, ground to a fineness of <4 – 5 mm, is stirred in a continuous dissolver with leaching brine heated to just below its boiling point. The leaching brine is the mother liquor from the crystallization stage of a previous cycle of the process. The quantity of leaching brine required is determined by the amount of potassium chloride in the ore. The potassium chloride should be extracted from the ore as completely as possible, and the resulting product solution should be as nearly saturated as possible. The residue consists of two fractions of different particle size. The coarse fraction is removed from the dissolver and debrined. The fine fraction (fine residue or slime) is removed from the dissolver along with the crude solution, which is clarified with the aid of clarifying agents. The slime that separates is filtered off, and the filtrate from the coarse and fine residues is recycled to the recirculating brine. The residues are washed with water or plant brines low in potassium chloride to remove the adhering crude solution, which has a high potassium chloride content. The residues are then disposed of by dumping (see Section 5.8).

The hot, clarified, crude solution is cooled by evaporation in vacuum equipment. Potassium chloride and sodium chloride crystallize as the water is removed. The sodium chloride content of the crystals formed can be controlled by complete or partial replacement of the evaporated water during crystallization. The crystals formed are separated from the mother liquor and processed further. The mother liquor is heated and recycled to the dissolver as leaching brine.

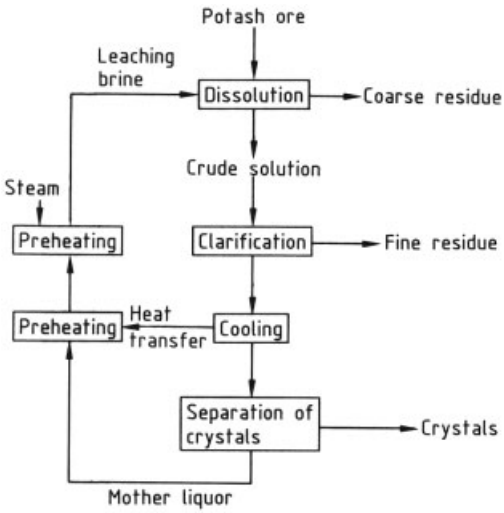

Figure 12. Overall schematic of a hot leaching process

The leaching process is usually carried out in two stages in a main dissolver and a secondary dissolver. The ground ore is first added to the main dissolver where it is mixed with the already partially saturated solution from the secondary dissolver. This causes the solution to be almost completely saturated, and it is then removed from the leaching equipment. The partly extracted ore is next fed to the secondary dissolver where it comes in contact with fresh leaching brine, and the potassium chloride that was not extracted in the main dissolver is taken up by the solution, which is then fed to the main dissolver. The leaching process in the main dissolver can be cocurrent or countercurrent (Fig. 13).

The residence time in the leaching equipment is insufficient to give complete KCl – NaCl equilibrium. The crude solution always contains less potassium chloride and more sodium chloride than corresponds to equilibrium. Another reason for the potassium chloride concentration to be maintained below saturation is that complete recovery of potassium chloride from the ore is

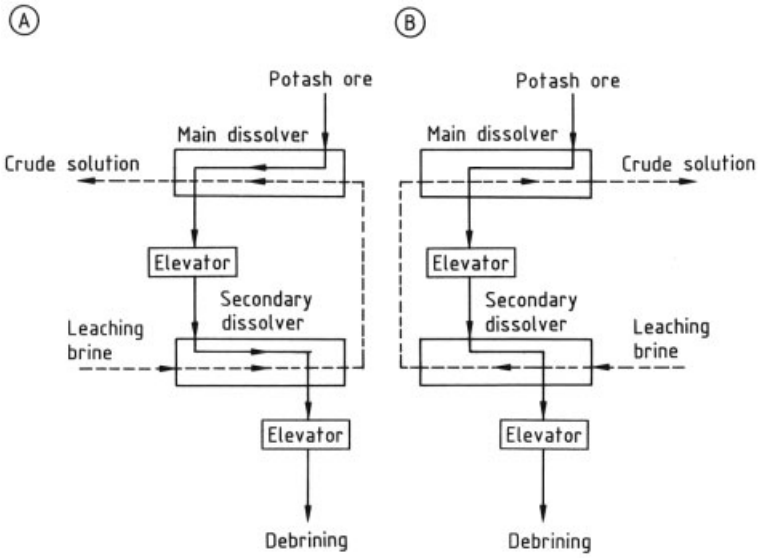

Figure 13. Schematic arrangements of a hot leaching apparatus
A) Cocurrent flow; B) Countercurrent flow

possible only if the hot solution at the end of the leaching process still has a small capacity for dissolving potassium chloride. Also, the crude solution cools by 1 – 2 °C as it travels from the main dissolver to the overflow from the following hot clarification stage. If the solution leaving the main dissolver were completely saturated in potassium chloride, the potassium chloride would begin to crystallize at this point, leading to losses in the slime.

For the crystallization process the potassium chloride content should be maintained as near to saturation as possible in the leaching equipment, so that the concentration difference between the mother liquor and the hot crude solution is as great as possible. The greater this difference, the smaller is the amount of solution used, and the lower is the energy consumption.

Also, a high concentration of potassium chloride in the solution leads to a high potassium chloride content in the crystalline product obtained on cooling. If the solution is unsaturated with respect to potassium chloride, a corresponding amount of sodium chloride in excess of its saturation concentration is taken up by the solution to compensate for the missing potassium chloride. This means that, on cooling, a quantity of sodium chloride crystallizes, causing a decrease in the potassium chloride content of the product even in sylvinitic solutions in which the solubility of sodium chloride increases with decreasing temperature.

Cooling the crude solution by evaporation increases the concentration of salts present below saturation, and the salts with which the solution is saturated crystallize. Unwanted crystallization of sodium chloride can be prevented by adding water to the vacuum cooling equipment, especially in the case of solutions from sylvinite ore leaching. If the required potassium chloride content in the crystalline product cannot be obtained by adding water to the solution, which is often the case when hard salt is being processed, the same result can be obtained by treating the product with cold water. A product containing 60 % K_2O (95 % KCl) is usually required. By this cold-water treatment technique, it is even possible to obtain a product with 62 % K_2O (98 % KCl). The spent solution is recycled to the process.

For KCl of analytical or pharmaceutical quality, potassium chloride produced by the hot leaching process must be purified by single or multiple recrystallization.

Processing of Hard Salt. Unlike the sylvinitic potash ores, whose principal constituents are sodium and potassium chloride, hard salts contain not only the alkali chlorides but also large amounts of kieserite, usually with

varying amounts of carnallite, langbeinite, and anhydrite. Therefore, process brines produced by the leaching of hard salt are characterized by high contents of magnesium chloride and magnesium sulfate, which make potassium chloride production more difficult.

Magnesium sulfate comes mainly from kieserite, which is very soluble (Fig. 14) but has a slow rate of dissolution that becomes even slower if large amounts of dissolved $MgCl_2$ are present. The amount of magnesium sulfate that dissolves depends on the grain size of the ore fed to the dissolving equipment, the kieserite content of the ore, the time for which the crude salt – solution mixture is stirred, the temperature, and the magnesium chloride content of the brine.

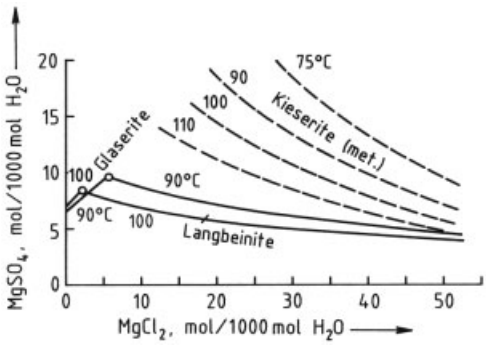

Figure 14. Metastable solubility of kieserite in the quinary system saturated with KCl and NaCl between 75 and 110 °C, and solubilities of langbeinite and glaserite at 90 and 100 °C

Magnesium chloride results from reaction of dissolved magnesium sulfate with potassium chloride to give sulfate-containing double salts, and also from any carnallite ($KCl \cdot MgCl_2 \cdot 6\,H_2O$) present in the hard salt. In general, the $MgCl_2$ and $MgSO_4$ contents are kept constant in the circulating brines, so that the rate at which magnesium is taken up from the ore is balanced by the rate at which it leaves the system in the residues and products. If the ore contains large amounts of carnallite, part of the circulating brine must be removed continuously from the system to prevent the $MgCl_2$ level becoming too high.

Because of the large amounts of magnesium salts in process brines, the hard salt leaching process has three important disadvantages compared with the sylvinite process:

1) In solutions of hard salt, the range of solubilities is considerably less than in solutions of sylvinite (Fig. 11). To extract a given amount of potassium chloride, much more liquor, and hence more energy, is required compared with sylvinite.
2) With solutions of hard salt, crystallization of double sulfates such as schoenite, leonite, and langbeinite often occurs. These salts can appear in the residue or the product, which leads either to potassium losses in the residue or to a lower K_2O content of the product. Also, double salts can crystallize in pipelines, vessels, and pumps, interfering with the process and in extreme cases bringing it to a standstill. The particular double salt that crystallizes depends on the $MgCl_2$ content of the solutions, the temperature, and other operating conditions (Figs. 9 and 10). The extent of supersaturation with double sulfates must therefore be controlled to prevent uncontrolled crystallization of double salts and consequent introduction of impurities into the product or disturbance of the process. To achieve this, the circulating brine, or part of it, is fed to a reactor in which it is agitated intensively in the presence of nuclei (20 – 40 wt %) of the double salt to be removed until the brine is no longer supersaturated with respect to it [67].
3) Another disadvantage of the hard salt leaching process is that, with higher magnesium salt contents, the temperature dependence of the solubility of NaCl becomes unfavorable (Fig. 11), that is, when the potassium chloride is crystallized by cooling, considerable amounts of sodium chloride can also crystallize (Fig. 15). In practice, equilibrium is not completely reached when the KCl – NaCl is dissolved, and cooling of the solution often occurs by water removal, so that the K_2O content of the crystals formed is usually only a little more than 40 %. Since the usual potassium chloride for fertilizers (excluding special products) has a minimum potash content of 60 % K_2O, this primary product was formerly crystallized in a second leaching plant. Alternatively, a product can be made with 60 % K_2O directly, if the crystallization of sodium chloride during vacuum cooling is prevented by addition of sufficient water before or during crystallization of the solution

to ensure that only potassium chloride crystallizes [67]. The K_2O content in the product of a primary crystallization can be increased to the required level by treating the product with water or a plant solution unsaturated with respect to sodium chloride. Both the excess water added during vacuum cooling and the water used for treating the product must be removed from the recirculating system of the plant. Both cause a loss of yield whose extent depends on operating conditions. To avoid this, excess water must be evaporated from process liquors.

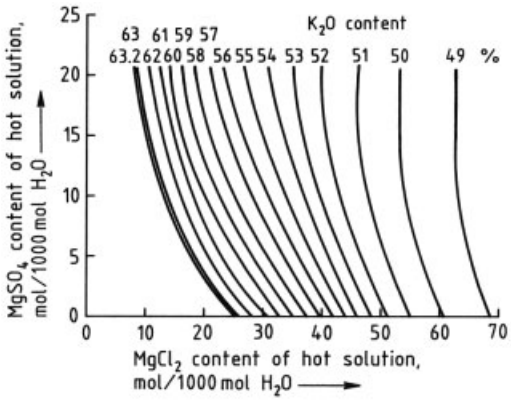

Figure 15. K_2O content of the crystalline product obtained by cooling an equilibrium solution saturated with NaCl and KCl from 90 to 25 °C without evaporation while maintaining saturation

5.2.3. Processing of Carnallite

Carnallite, $KCl \cdot MgCl_2 \cdot 6\,H_2O$, is the most abundant potassium mineral in salt deposits and occurs widely in mixtures with halite or with halite and kieserite in the form of carnallitite ore (see Section 2.2). In the early days of the potash industry in Germany, it was the preferred starting material for the production of potassium chloride. Today, sylvinitic ores and hard salts are used almost exclusively, because the extraction and processing of carnallite ore are considerably more difficult and expensive for the following reasons:

1) Carnallitite ore has unfavorable mechanical properties that make mining more difficult.
2) The K_2O content of pure carnallite is ca. 17 %, compared with ca. 63 % for pure sylvite.
3) Whereas the separation and purification of sylvite from sylvinite ore can usually be carried out by flotation, which does not involve a phase change, the extraction of potassium chloride from carnallitite ore necessitates dissolution or decomposition of the carnallite, and a high energy consumption for decomposition or purification of the decomposition product, depending on the process.
4) The treatment of carnallite generates large quantities of concentrated magnesium chloride solution, which must be disposed of.

For these reasons, mined carnallitite ore is today seldom used as a raw material. However, carnallite is often a major component in mixed salts of the hard salt type, and hence influences the choice of processing method.

Large quantities of a carnallite – halite mixture obtained by solar evaporation of water from the Dead Sea are used for the production of potassium chloride in Israel and Jordan.

Theoretical Basis. The theory of the production of potassium chloride from carnallite and carnallite-containing mixed salts is based on the $K_2Cl_2 - MgCl_2 - H_2O$ system shown in Figure 16, which is valid between − 3 and 117 °C. In VAN'T HOFF coordinates (moles of salt per 1000 mol H_2O) the points representing the composition of water ($P^{O}_{H^2}$), bischofite ($P_{bischofite}$), and carnallite ($P_{carnallite}$) are indicated. All possible mixtures of water and carnallite are shown on the straight line between $P^{O}_{H^2}$ and $P_{carnallite}$, the molar ratio $K_2Cl_2 : MgCl_2$ here being always 1 : 2, as in carnallite. The curve from L_4 via R and E to point L_0 represents an arbitrary isotherm, and indicates compositions in which solutions are in equilibrium with the corresponding solid phase. Point L_0 gives the solubility of KCl in water, and the solutions L_1 are in equilibrium with KCl as the solid phase. Solution E is in equilibrium with the solid phases KCl and carnallite, and the solutions L_2 are in equilibrium with carnallite only. Solution R is in equilibrium with bischofite and carnallite. The solutions L_3 are in equilibrium with bischofite, and point L_4 indicates the solubility of bischofite in water.

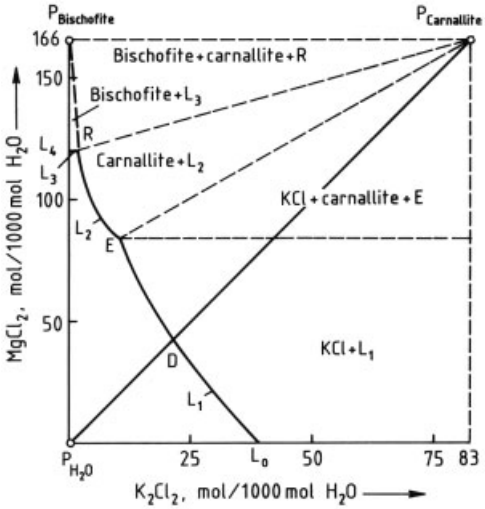

Figure 16. The system $K_2Cl_2 - MgCl_2 - H_2O$ (not to scale) represented by using van't Hoff coordinates

If carnallite is dissolved in water, the composition of the solution follows the straight line $P_{H_2O} - P_{carnallite}$, which intersects curve L_1 at point D. Here, the solution is saturated with KCl, and further addition of carnallite results in dissolution of $MgCl_2$ and crystallization of KCl until point E is reached. This incongruent solubility is the basis for the simple method of processing carnallite (i.e., cold decomposition by mother liquor).

Table 6. Composition of stable saturated solutions in Figure ∗

Point	Temperature, °C	Density, g/cm^3	Concentration, mol/1000 mol H_2O			
			K_2Cl_2	$MgCl_2$	$MgSO_4$	Na_2Cl_2
E_{25}	25	1.275	5.8	70.8		4.4
Q_{25}	25	1.291	5.8	68.0	5.2	4.2
E_{105}	105	1.328	13.2	93.7		4.4
Q_{105}	105	1.325	13.7	92.9	1.0	4.3

∗ Solutions Q_2^5 and Q_1^{05} are saturated with $MgSO_4$ and NaCl, and correspond to solutions E_2^5 and E_1^{05}, which are saturated with NaCl.

Carnallitic potash ores or crystallized products from solar evaporation always contain so much halite that solutions produced during processing are saturated with sodium chloride. Figure 17 is a section from the quaternary system $K_2Cl_2 - NaCl - MgCl_2 - H_2O$ saturated with NaCl, showing the 25 °C and 105 °C isotherms. The kieserite content in the ore in the region of the solutions E has in practice a negligible effect on the composition of the solutions. Some analyses of equilibrium solutions of the quaternary system are given in Table 6.

Processing Methods. Many processes for the treatment of carnallite have been described in the literature and used [10], but only a few are important today. By far the most important is cold decomposition by mother liquor. The complete dissolution process is still used occasionally.

Cold decomposition is carried out at ambient temperature (e.g., 25 °C). Carnallitite ore is mixed and agitated with water or with a solution of low $MgCl_2$ content such that the composition of the mixture corresponds to point B in Figure 17. This causes the crystallization of an amount of potassium chloride corresponding to the line $B - E_{25}$, with formation of a solution of composition E_{25}. The very fine potassium chloride produced still contains fine, salted-out sodium chloride, undissolved halite, and sometimes kieserite and clay, depending on the composition of the carnallitite ore used. Approximately 85 % of the potassium chloride contained in carnallite can be obtained as a crystalline product by this process. The yield can be increased by evaporative concentration of the decomposition mother liquor E_{25} or Q_{25} (analyses are given in Table 7). During evaporation, sodium chloride and sometimes magnesium sulfate crystallize and must be removed. The synthetic carnallite that crystallizes when the concentrated solution is cooled is fed to the cold decomposition process.

Potassium chloride from the decomposition, which consists of very finely divided particles and is rather impure, must be purified. Purification by flotation is difficult owing to the fineness of the decomposition product (see, however, Section 5.3.6). For this reason, purification is nearly always carried out by the hot leaching process, which yields a very pure, completely water-soluble, coarse-grained product.

The mother liquor E_{25} or Q_{25} from the decomposition process can take up a certain amount of sodium chloride (see Table 7). If the carnallitite used in the cold decomposition contains only small amounts of halite and if water is used for the decomposition, potassium chloride can be produced that requires no further purification, because sodium chloride has dissolved

in the decomposition liquor. In Israel, a large proportion of the carnallite – halite mixture recovered from the Dead Sea by solar evaporation is so completely freed from halite by fractional crystallization and hydraulic classification that a grade of potassium chloride with > 60 % K_2O can be produced directly by cold decomposition. This so-called cold crystallization process is carried out so that the crystals produced have the same grain size as those from the crystallization plant of the hot leaching process and no further increase in grain size is required. A schematic diagram showing various processes used in the treatment of carnallite from the Dead Sea is given in Figure 18.

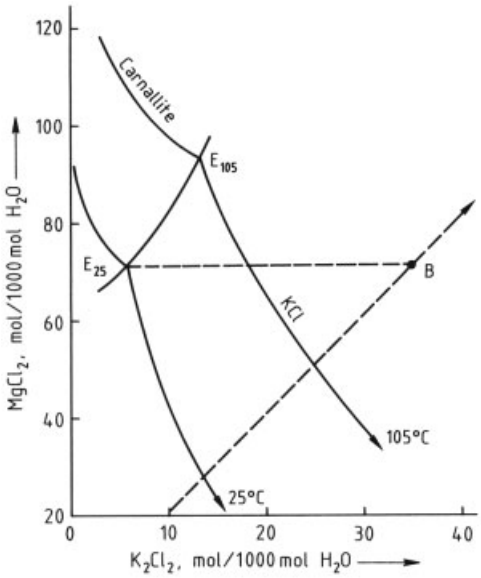

Figure 17. Quaternary system (K_2Cl_2 – $MgCl_2$ – Na_2Cl_2 – H_2O) saturated with NaCl
See text for explanation

In the electrostatic treatment of carnallitic hard salts (see Section 5.4.2), concentrates can be produced that contain various amounts of halite together with sylvite and carnallite. These concentrates are reacted with suitable process brines or with water to bring about cold decomposition of the carnallite. The K_2O content of the product, and hence its potential use, depend on the halite content of the concentrate.

The complete dissolution process is used in an experimental plant in central Germany in which carnallite is extracted from a salt deposit by solution mining (see Section 3.3) [48].

The crude salt is dissolved by a solution in which the magnesium chloride content is adjusted so that at the chosen temperature of ca. 80 °C, no decomposition of carnallite takes place, and hence no crystallization of potassium chloride, and the carnallite is completely dissolved. The solution, which is almost saturated with carnallite, is pumped out, evaporated, and cooled, causing crystallization of the carnallite, which is then treated by the cold decomposition process to produce potassium chloride.

For mixed salts containing less than ca. 15 % carnallite, carnallite decomposition is generally dispersed with, and the ore is treated directly by the hot leaching process. The magnesium chloride that enters the process in the carnallite must be removed from the brine circuit in appropriate process brines. If the carnallite content of the ore is high, a carnallite decomposition stage is carried out before the hot leaching process.

5.2.4. Equipment

Leaching. The choice of leaching equipment depends on the properties of the material to be leached and the throughput required. The techniques used in plants in which the potash is completely leached (e.g., in many German and French plants) differ fundamentally from those in plants (e.g., in Canada) where only the fine materials from a screening operation, cyclone fines, and slimes are recrystallized or extracted.

Potash ores are leached at rates of up to 1000 t/h salt, producing up to 2000 m^3/h solution. Screw dissolvers, up to 14 m long and 3 m in diameter, are widely used (Fig. 19). High material transfer is achieved by fitting partitions in the upper part of the dissolver at intervals of 2 – 3 m. These are immersed in the suspension of crystals and force the solution to flow perpendicular to the main flow direction. At the outlet of the dissolver, the solids are scooped from the solid – liquid mixture by an elevator system with perforated buckets from which liquid drains during conveying; the residual water content is ca. 15 – 20 %. Alternatively, bucket wheels can be used.

Elevators or bucket wheels remove only the coarse residue (ca. 75 % of the total residue)

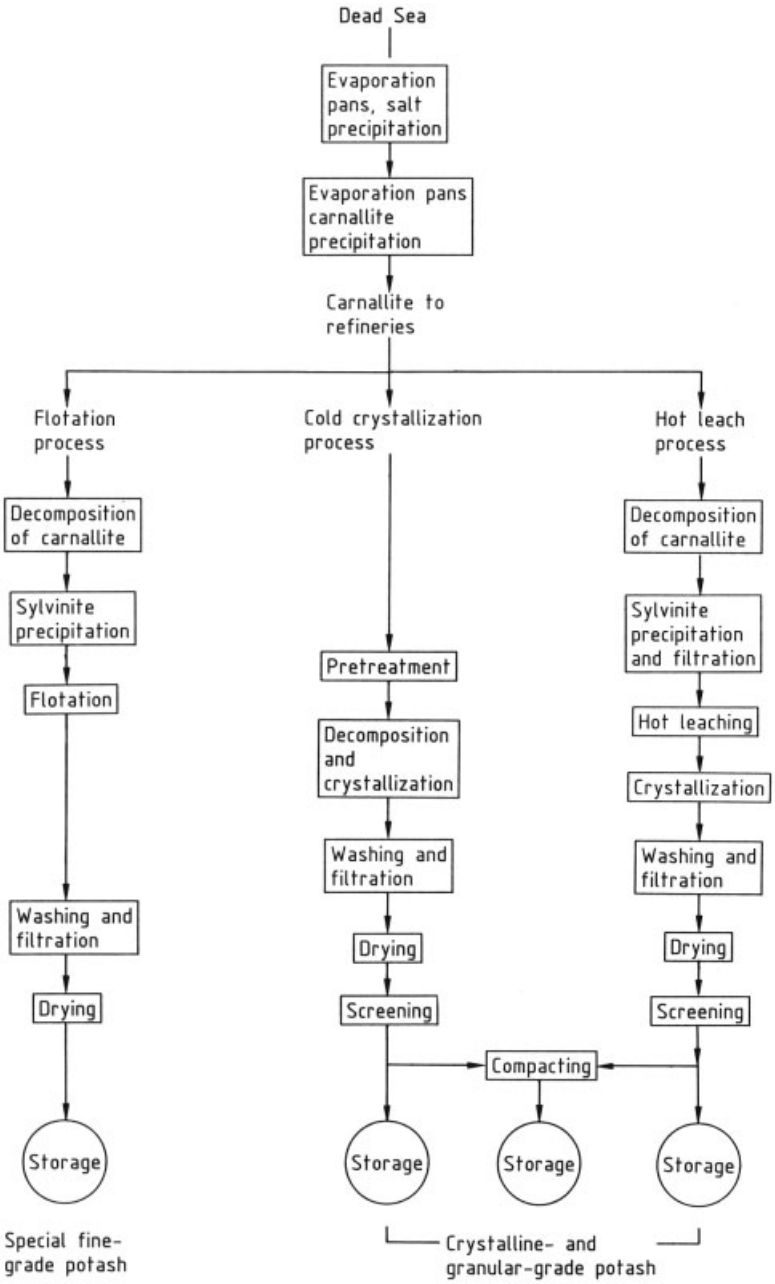

Figure 18. Flow diagram for potassium chloride production by the Dead Sea Works Ltd. (Israel) [4]
Reproduced from [4] with permission

from the screw dissolver. The amount of potassium chloride-containing solution adhering to the coarse residue is usually lowered to 2 – 4 % by centrifugation.

The fine residue (ca. 25 % of the total residue), which consists of very fine salt and insoluble components of the ore (mostly anhydrite and clay), flows with the raw solution from the screw dissolver into the hot clarifier, where flocculation agents are added and sedimentation occurs. Sealed, insulated circular clarifiers are used. The fine residue is removed from the hot

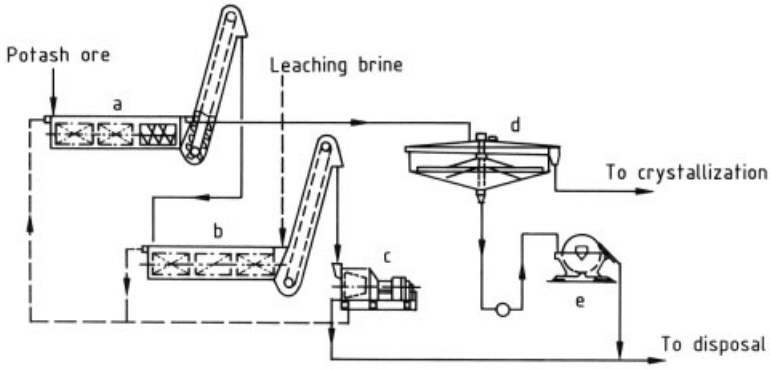

Figure 19. Leaching system with residue debrining
a), b) Screw leachers with draining elevators; c) Vibratory screen centrifuge; d) Hot clarifier; e) Rotary filter

clarifier as a suspension with a solids content of ca. 50 % and debrined on rotary filters to a residual water content of 10 – 16 %. To avoid losses of potassium chloride, the solution adhering to the residue must be removed. For small quantities of residue, this is done by washing on the filter with water or a process brine having a low potassium chloride content. For large amounts of residue, two- or three-stage water washing is used; the filter cake is slurried with the water, and the resulting suspension is filtered. This treatment is repeated once or twice.

In North America, where the flotation process predominates and only small quantities of ore are treated by the hot leaching process, a series of agitated vessels is used. The leaching process takes place in either cocurrent or countercurrent flow. Between the stages of countercurrent leaching, solid – liquid separation is carried out with hydrocyclones or hydroclassifiers [68]. The layout of a leaching plant including crystallization is shown in Figure 20.

Crystallization [69, 70]. The hot solution from the clarifier of the leaching plant is almost saturated with potassium chloride and is cooled by expansion evaporation in vacuum equipment to cause crystallization. The vapors are condensed in surface or barometric condensers, with process brines used for higher temperatures and cooling water for lower temperatures. The amount of heat that can be recovered depends on the number of stages in the cooling system. In the past, vacuum cooling plants were constructed with up to 24 stages, to give maximum possible heat recovery. However, the number of stages in a modern plant is usually between four and eight.

The high cooling rate of a vacuum plant results in a high degree of supersaturation that produces very fine crystals, unless the design of the plant and the crystallizing conditions are optimized to give a coarse product. Several suitable crystallizers are now available (→ Crystallization and Precipitation).

If it is not important to have coarse crystals (e.g., if they are to be reprocessed), stirred crystallizers or conical-based evaporators are used. The former consist of horizontal cylindrical vessels divided into chambers by vertical walls. The solution or suspension flows from chamber to chamber through openings in the walls as it is cooled. These crystallizers are provided with mechanical stirrers or air agitation (Fig. 21).

The conical-based evaporator consists of a vertical cylinder with a lower conical section. The solution is sprayed into the top, and much of the water evaporates. A certain liquid level is maintained in the evaporator, into which liquid from the nozzle falls as a supersaturated solution. Several evaporators are arranged in series one above the other so that the suspension flows through each stage in the direction of decreasing pressure without the need for intermediate pumps (Fig. 22). The advantage of this arrangement lies in its very simple and economical construction, with no moving parts. However, the product is very finely divided.

The aim of the crystallization process is usually a dust-free product with optimum grain size. Two main requirements must be met to achieve this: the solution must have only a slight degree

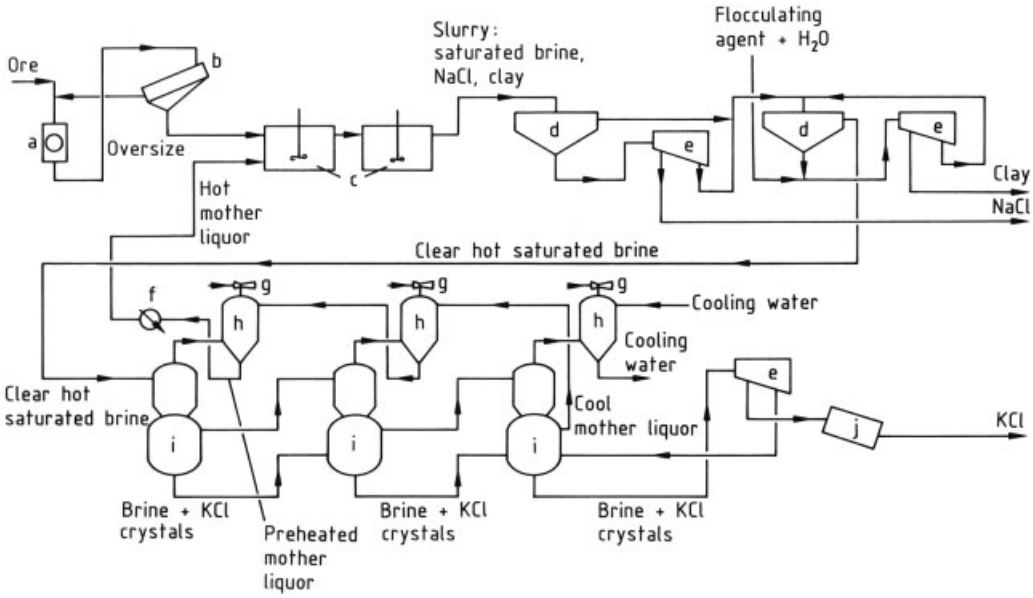

Figure 20. Leaching – crystallization process for production of potassium chloride from potash ore [8]
Reproduced from [8] with permission
a) Crusher; b) Screen; c) Leach tanks; d) Thickener; e) Centrifuge; f) Heater; g) Steam ejector; h) Barometric condenser; i) Vacuum cooler – crystallizer; j) Dryer

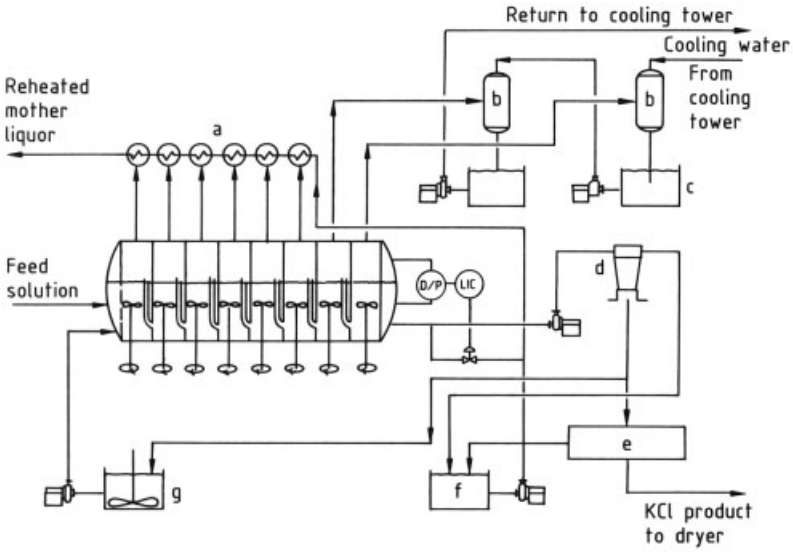

Figure 21. Eight-stage horizontal crystallizer
a) Surface condensers; b) Barometric condenser; c) Hot well; d) Cyclone; e) Centrifuge; f) Filtrate tank; g) Slurry holding tank.
Reprinted with permission from R. M. McKercher et al., *Proc. 1st Int. Potash Technol. Conf.,* Oct. 3 – 5, 1983, Pergamon Press.

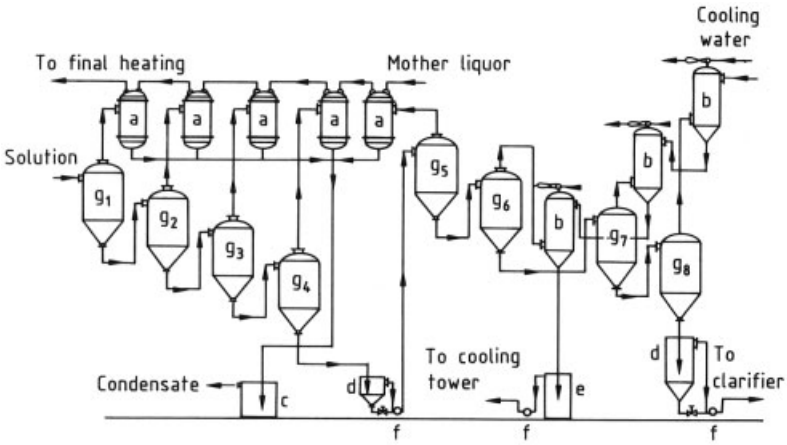

Figure 22. Conical-based evaporator – crystallizer plant with eight evaporation stages
a) Surface condensers; b) Barometric condensers; c) Barometric collection vessel for condensate; d) Barometric collection vessel for salts and brines; e) Barometric collection vessel for cooling water; f) Pumps; $g_1 - g_8$) Evaporators

of supersaturation, and this should be removed by crystallization on seed crystals that are already present. The slight degree of supersaturation is achieved by providing internal circulation within each crystallization stage, in which the cooled solution is mixed with the added hot solution in an exactly controlled ratio. The continuous presence of seed crystals is ensured by keeping the solids content in each crystallizer at ca. 30 wt %.

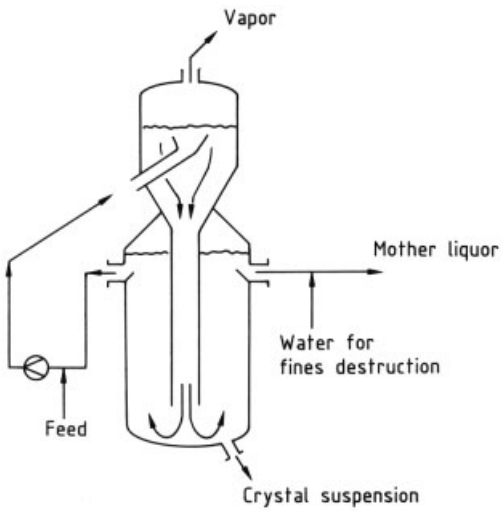

Figure 23. Oslo crystallizer

These conditions are fulfilled by two types of crystallizer: the fluidized-bed (Oslo) type and the draft-tube agitated type.

In the Oslo crystallizer (Fig. 23), the solution is supersaturated in a separate evaporator and flows into a crystal slurry vessel below, where the supersaturation is removed by crystallization on KCl crystals already present. To reduce the degree of supersaturation in the evaporator, a large amount of clarified solution is pumped from the crystal slurry vessel into the evaporator, together with the incoming hot feed solution. Clarified solution for the next stage is removed from the vessel at another takeoff point. A crystal suspension is also removed from the vessel at a rate corresponding to the rate of crystallization, and is fed either to the next stage as seed crystals or to the debrining stage. Crystallizers of this type are constructed by companies such as Lurgi, Swenson, and Struther-Wells.

In draft-tube agitated crystallizers, a state of supersaturation is created and removed in the same crystallizer. This type is manufactured by Swenson (DTB=draft tube baffle), Standard-Messo, and Kali und Salz. In the Swenson (→ Crystallization and Precipitation) and Standard-Messo (→ Crystallization and Precipitation) crystallizers, the entire suspension is brought into continuous contact with the evaporator surface by a propeller stirrer that provides internal circulation; immediately after supersat-

uration is achieved, a large quantity of salt is made available to remove the state of supersaturation. The coarser crystals collect preferentially in the lower part of the crystallizer, where they are removed as a suspension and fed to the next stage or to the debrining stage. The hot solution is passed through a clarifying zone and is fed to the next stage.

All the processes described so far are operated cocurrently (i.e., the crystallizing salt is transported with the cooling solution from one stage to the next). In contrast, the Kali und Salz process (Fig. 24) operates countercurrently [71]. The salt produced in each crystallizer is passed through a classifier at the bottom of the crystallizer as soon as a predetermined minimum crystal size has been reached; then, unlike the other systems, it is passed to the next hottest stage. The classifier is fitted with equipment that enables the upward flow rate, and hence crystal concentration and product crystal size, to be controlled independently of the throughput rate of the solution. Mixing of the cooler solution with the hotter solution in a crystallizer is prevented by transferring the salt with the help of the solution from the stage into which it is pumped.

5.3. Flotation

Since the early 1900s, ores of many different kinds have been processed by froth flotation. A general description of this process is given in → Flotation.

Investigations into the flotation of potash ores began in the early 1930s in the United States and the Soviet Union [72,73]. The first full-scale plant began production in 1935 in Carlsbad (New Mexico, United States) [74].

The first reports of investigations into the flotation of potassium salts in Germany appeared in 1939 [75]; the first full-scale plant began operating in 1953 [76,77].

At present, most of the worldwide production of potash fertilizers is by flotation. This is the most widely used process in Canada, the United States, the CIS, and France. Small hot leaching plants are often attached to the flotation plants for treating slimes and intermediate products. The flotation process is not widely used in Germany, where a combination of electrostatic treatment with the hot leaching process predominates.

5.3.1. Potash Ores Suitable for Flotation

Various types of crude potash salts can be treated by flotation:

Sylvinite ores are mixtures of sylvite (KCl) and halite (NaCl) in varying ratios. They represent the majority of potash ores treated. The sylvinites of Canada, the United States, and the CIS also contain up to 8 % clay components [78].

Hard salts contain kieserite ($MgSO_4 \cdot H_2O$), as well as sylvite and halite, and sometimes also anhydrite ($CaSO_4$).

Mixed salts consist of a mixture of sylvinite ore or hard salt with carnallite ($KCl \cdot MgCl_2 \cdot 6\,H_2O$).

Polymineral salts contain not only sylvite, halite, and kieserite, but also langbeinite ($K_2SO_4 \cdot 2\,MgSO_4$), kainite ($4\,KCl \cdot 4\,MgSO_4 \cdot 11\,H_2O$), polyhalite ($K_2SO_4 \cdot MgSO_4 \cdot 2\,CaSO_4 \cdot 2\,H_2O$), and clay.

5.3.2. Carrier Solutions

In the flotation of water-soluble salts the carrier liquids are salt solutions that are saturated with the components of the raw material.

Thus, sylvinite ore flotation is carried out in a KCl – NaCl solution. For the flotation of hard salt, the brine also contains various amounts of magnesium sulfate and magnesium chloride.

5.3.3. Flotation Agents

The *collectors* are the true flotation agents, which selectively coat the surface of the component to be floated. For the flotation of sylvite, straight-chain primary aliphatic amines are used in the form of their hydrochlorides or acetates [79]. Mixtures of amines of various chain lengths,which largely eliminate the effects of pulp temperature variation, are extremely useful. A typical flotation amine has, for example,

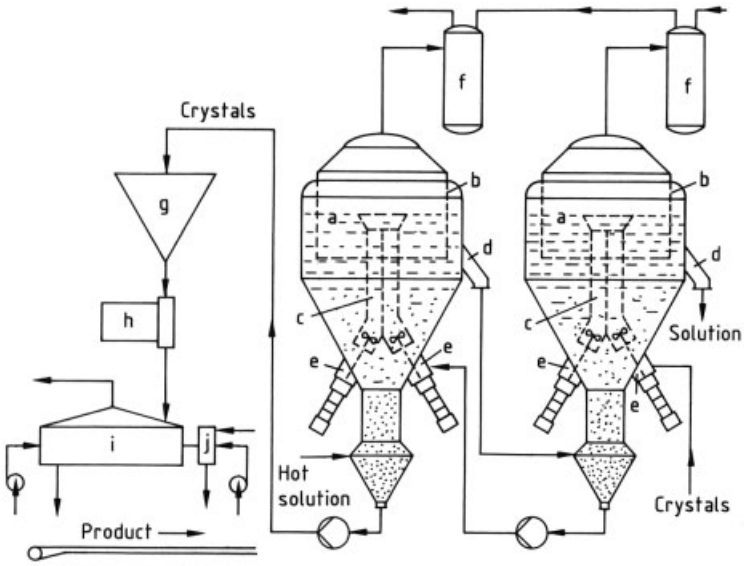

Figure 24. Kali und Salz countercurrent crystallization process
a) Crystallizers; b) Bells; c) Draft tubes; d) Liquor overflow; e) Stirrers; f) Condenser; g) Thickener; h) Centrifuge; i) Dryer; j) Combustion chamber

the following composition: 5 %$C_{14}-NH_3Cl$, 30 %$C_{16}-NH_3Cl$, 65 % $C_{18}-NH_3Cl$, iodine number: 4.

Foamers contribute to the dispersion of long-chain amines and to the stabilization and homogeneous distribution of amine micelles [80]. The following substances are preferred: aliphatic alcohols with chain lengths $>C_4$, terpene alcohols, alkylpolyglycol ethers, and methyl isobutyl carbinol, which is used mainly in Canada and the United States.

Extenders for sylvite flotation are nonpolar materials, especially oils of various types. They are probably incorporated in the micelles and increase their hydrophobic properties. Extenders are especially effective in the flotation of coarse particles [81].

Clay depressants are used in salt flotation to block clay components, which would otherwise bind large amounts of flotation agent. Clay contents of 1.5 – 2 % can be controlled by treatment with these depressants. If larger amounts are present, additional steps must be taken (clay flotation or classification). Clay depressants include guar and starch products, carboxymethyl cellulose, and polyacrylamide [78].

5.3.4. Theory

The combinations of reagents used for sylvite flotation have been found by empirical investigation.

Theoretical studies of the separation of potassium and sodium chloride by flotation have led to various interpretations.

The theories developed up to 1961 are thoroughly discussed in [82], which also describes experimental results that have contributed greatly to the understanding of salt flotation.

The most important theories are reviewed below:

According to the *exchange theory* [83–85], hydrophobic properties can be imparted to a mineral if the polar group of a collector can be incorporated into the crystal lattice in place of an ion. In the flotation of sylvinite ore, for example, exchange between the K^+ ion (radius: 0.135 nm) and the NH_3^+ group of the amine (radius: 0.143 nm) is assumed to occur, whereas the Na^+ ion (radius: 0.095 nm) is too small (Fig. 25). This does not account for the fact that kieserite ($MgSO_4 \cdot H_2O$), for example, is readily floated by amines in water although the Mg^{2+} ion has a radius of only 0.065 nm.

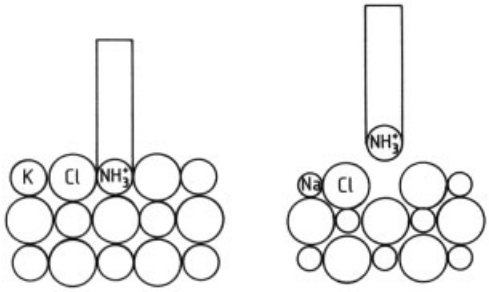

Figure 25. Incorporation of a polar group of the collector into the crystal lattice [83]
Reprinted with permission of the Society for Mining, Metallurgy, and Exploration, Inc.

The *structure theory* [86] postulates that the interanion distance in the crystal lattice of the salt matches the intermolecular distance in the lattice of the amine hydrochloride within ± 20 %. In sylvite, this condition for oriented growth of the collector on the crystal is fulfilled (Fig. 26), whereas the interionic distance of 0.398 nm in the halite lattice is too small. However, langbeinite has good flotation properties, although these steric requirements are not met [86].

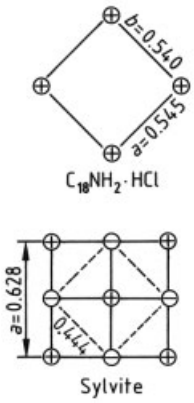

Figure 26. Lattice analogy between the amine and the salt [86] (distances in nm)

According to the *hydration theory* [87], flotation of a mineral is impossible if it has a positive heat of solution, whereas salts with negative heats of solution can be flotated. If heat is evolved when a salt dissolves, more energy is produced by hydration of the ions than is required for breakdown of the lattice. This suggests that such salts are strongly bonded to water molecules, preventing adhesion of the collector. A comment on this theory is given in [88].

Another hydration theory is based on new information on the hydration of ions in dilute solutions [89]. This differentiates between positive local hydration (in which the ion reduces the mobility of the neighboring water molecules) and negative local hydration (water molecules close to the ions are more mobile than in pure water).

According to [90], in the case of positive local hydration (e.g., with Na^+ ions), the Cl^- ions are shielded by water molecules, so that the bond between a cationic amine and sodium chloride can only be weak. In the case of negative local hydration (e.g., with K^+ ions), formation of a bond with the collector is unhindered. Application of these theories, developed for dilute electrolyte solutions, to the hydration state of salt surfaces is difficult, but should be qualitatively valid [90].

The G zone theory (G = Grenzflächen = interfaces) [91] applies both to the flotation of sylvinite with amines and to the flotation of kieserite with an anionic surfactant (the sodium salt of a highly sulfated fatty acid). In both cases, adsorption of the collector proceeds by the same mechanism.

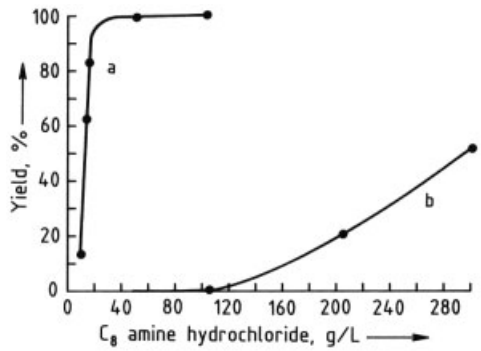

Figure 27. Flotation of potassium chloride with octylamine in saturated solutions of KCl and KCl + NaCl (Feed material: 20 % KCl, 80 % NaCl; 0.1 – 0.315 mm)
a) Carrier liquid saturated in KCl and NaCl (low solubility of octylamine); b) Carrier liquid saturated only in KCl (high solubility of octylamine)

Amines and the anionic surfactant, in a saturated solution of salts that can be floated with these reagents, are present either as micelles or in true solution. In the usual carrier brines,

flocculation (amines) or formation of very small droplets occurs (anionic surfactant).

Based on experimental results, zones are assumed to be formed on the surfaces of crystals suspended in a mixed-salt solution (carrier liquid) in which ions from the lattice are present in the form of a saturated solution.

When collectors are introduced into the suspension, they migrate to the zones where their solubility is greatest due to their thermodynamic tendency to dissolve (i.e., to the boundary zones of those particles that they cause to float). The bonding of the collectors in these zones is reinforced if their solubility in the surrounding carrier brine is low. The success of flotation depends on the strength of this bond (Fig. 27).

Froth cannot be produced in the usual carrier liquids either by collectors or by foaming agents. The collision of an air bubble with a crystal coated with the collector material causes the latter to spread over the surface of the bubble, stabilizing it and causing the crystal to float.

Large-scale operations of sylvite and kieserite flotation has thus been placed on a good unified theoretical foundation.

5.3.5. Flotation Equipment

The flotation equipment used in the potash industry resembles that used for flotation of other ores (→ Flotation, Chap. 5.3.).

Mechanical flotation cells operate with a rotor – stator system that causes both thorough mixing of the pulp and thorough distribution of the air, which can be drawn in by suction or injected from a compressed air network.

In recent decades, high-capacity flotation has been introduced almost everywhere. In this process, several stirrers operate in a single trough [92].

Agitair high-capacity flotation equipment is used in Germany, and the preconcentrate is purified in individual cells of the Denver type. Wemco high-capacity machines are used by the French potash industry, and Mechanobr high-capacity cells by the former Eastern-bloc countries.

Modified Mechanobr M 7 fluidized-bed units (Fig. 28), manufactured by Mescerjakov, are standard equipment in the CIS [93, 94]. Typical features include the presence of a grid and circulation of the suspension via a box on the front wall of the cell, connected to the agitator by a circulation pipe.

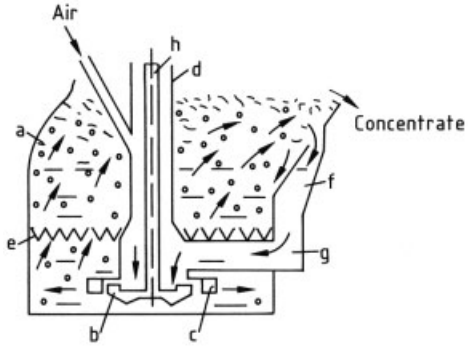

Figure 28. Mechanical flotation cell with fluidized bed
a) Flotation cell; b) Stirrer; c) Stator; d) Tube of the stirrer system; e) Grid; f) Circulation box; g) Circulation pipe; h) Shaft

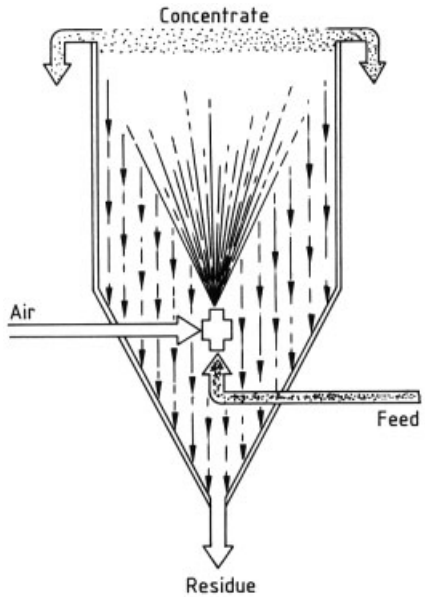

Figure 29. Principle of pneumatic free-jet flotation
Reproduced from [95] with permission

Pneumatic flotation equipment operates without stirrers. In the free-jet flotation process developed in Germany [95], the conditioned pulp is pumped through an aerator fitted with a porous gas distributor or annular nozzle, either outside or inside the separation vessel, and flows

as a free jet into the separation vessel. Air bubbles with the adhering minerals rise to the top and form a froth that flows over the edge of the vessel into a channel (Fig. 29). The pneumatic flotation process has been tested successfully for a special case of sylvite flotation. It is used industrially for the recovery of kieserite from leaching residues.

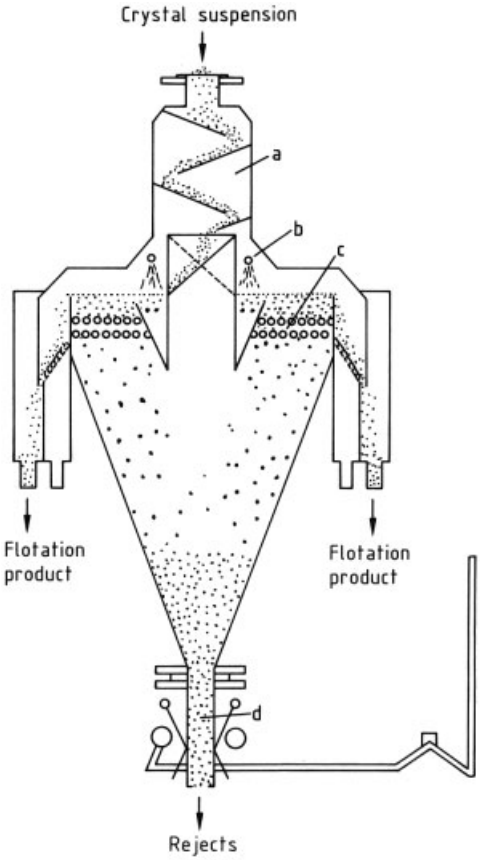

Figure 30. Pneumatic coarse-grain flotation cell (Gogorchimproect type) [96]
a) Feeding device; b) Spray nozzle; c) Aeration system (perforated rubber tubes); d) Device for removing reject material

In the CIS, a pneumatic flotation cell has been developed for the treatment of coarse sylvinite ores with a maximum grain size of 3–5 mm [96, 97] (Fig. 30). The conditioned pulp is fed from above onto a layer of froth immediately above the perforated-tube aeration system. The hydrophobic sylvine crystals are retained in the froth, so that mechanical damage of the bubble–crystal combination is kept to a minimum. The flotation rate is very high.

5.3.6. Processes

Sylvinite Ore Flotation. Pure sylvinite of German origin can be floated without serious problems. The components of the ore are relatively strongly intergrown [50] and must be broken down by grinding until each grain consists as far as possible of only one component.

The ore is first ground to a particle size of < 4 mm, screened to remove fines, and then slurried with carrier brine and ground to a grain size of $< 0.8 - 1.0$ mm in a circulating system that includes rod mills, spiral classifiers, or wet screens (Fig. 31).

The solids content of the flotation pulp is adjusted to 30–40 wt %, and it is then mixed with flotation agents (ca. 40 g of oil, 20 g of foaming agent, and 40–80 g of collector per tonne of crude salt). As much of the desired product as possible is then extracted by the three-stage high-capacity rougher flotation. The rougher tails are classified in hydrocyclones, debrined (the coarse fraction in centrifuges, and the fine fraction in rotary filters), and dumped. Since the filtrate still contains finely divided salt, it is clarified in a thickener and recycled for slurrying the ore.

The material floated in the rougher flotation is then concentrated in a cleaner flotation process, usually in three stages. This does not produce a marketable product (with $> 60\,\%$ K_2O). The concentrate is therefore separated from the brine and washed with water in salt washing equipment.

This washing process dissolves more halite than sylvite, producing a potash salt that meets quality requirements. The washing liquor is removed by centrifuges, and the product is dried and conveyed to silos.

The cleaner flotation process also produces some intermediate-quality product with a fairly high K_2O content, whose coarse particles contain most of the residual intergrown material from the ore. They are therefore recovered by hydrocyclones and fed to the screening–grinding stage for size reduction. The energy consumption for the entire process is ca. 10 kW · h/t ore.

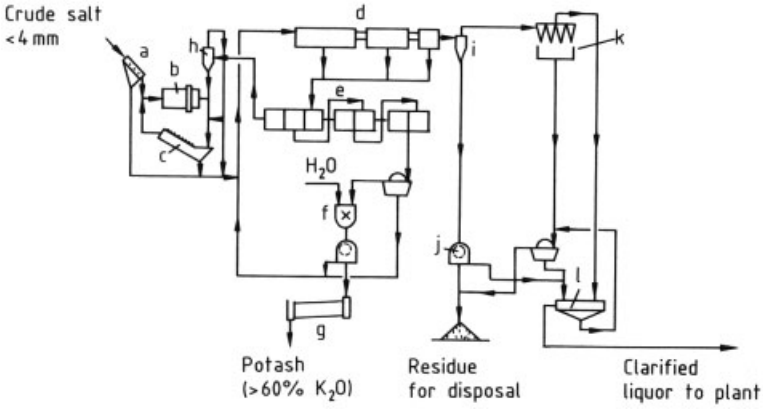

Figure 31. Flow diagram of the flotation of potash ores
a) Fine screen; b) Wet grinding; c) Classification; d) Rougher flotation (high volume); e) Cleaner flotation (three stage); f) Water washing for concentrate and debrining; g) Drying; h) Classification of intermediate-quality product; i) Residue (tails) collection; j) Residue debrining; k) Cyclone; l) Clarification of brine for recirculation

The treatment of sylvinite ores from Canada, the United States, and the CIS is more difficult owing to the clay content of up to 8 %. The clay components disintegrate in the brine, forming very fine slimes that absorb large quantities of flotation agent. The flotation process is therefore sometimes preceded by a desliming stage (Fig. 32). The clay is first detached from the surface of the salt by vigorous agitation in scrubbers, so that it can be floated in a separate operation. The clay is usually removed by multistage classification [98]. The fine material is then thickened and washed with water to reduce losses of K_2O. Up to 1.5 – 2 % of residual clay in the crude salt can be handled by adding clay depressants (see Section 5.3.3).

An advantage of clay-bearing sylvinite ores is that they are usually sufficiently liberated at grain sizes of < 3 – 5 mm. Sometimes, coarse crystals are floated separately from fine crystals (→ Flotation, Chap. 5.4.4.→ Flotation, Chap. 5.4.5.).

Hard-salt flotation is similar to the sylvinite flotation process (Fig. 31). However, the flotation product also contains kieserite, which cannot be removed by washing because it dissolves only slowly in water.

Sylvite, which is intergrown with kieserite, must therefore be liberated as fully as possible, which leads to increased grinding costs. Because of the increased production of fines, more flota-tion agent is needed. Also, in warm summer months, potassium-containing double salts (leonite, schoenite) can separate from the circulating brine. Since these salts end up in the residue, the yield is reduced.

Mixed-Salt Flotation. Carnallite ($KCl \cdot MgCl_2 \cdot 6\,H_2O$) in the mixed salt is usually decomposed by a brine with low $MgCl_2$ content. The potassium chloride produced has a grain size of < 0.04 mm and is accompanied by sylvinite ore or hard salt components.

In the 1980s, this complex salt mixture was successfully treated by flotation alone for the first time [99, 100].

Potassium chloride produced by the decomposition is first floated with a fairly small amount of amine in pneumatic or stirred flotation cells; then the sylvite component of the sylvinite ore or hard salt is floated with a further measured amount of amine, and a discardable residue is obtained (Fig. 33). The flotation froth is then treated by flotation in cleaner cells to give a concentrate with > 55 % K_2O. A final washing process gives a potash fertilizer salt containing > 60 % K_2O.

Polymineral salts are rarely treated by flotation. Potash ore from Stebnik (CIS) contains not only sylvite, halite, and clay materials, but also large amounts of kainite, langbeinite, and polyhalite [101]. Clay depressants are added first, followed by flotation of halite with a mixture of $C_7 - C_9$ fatty acids. Sodium hydroxide solution is used to adjust pH [101].

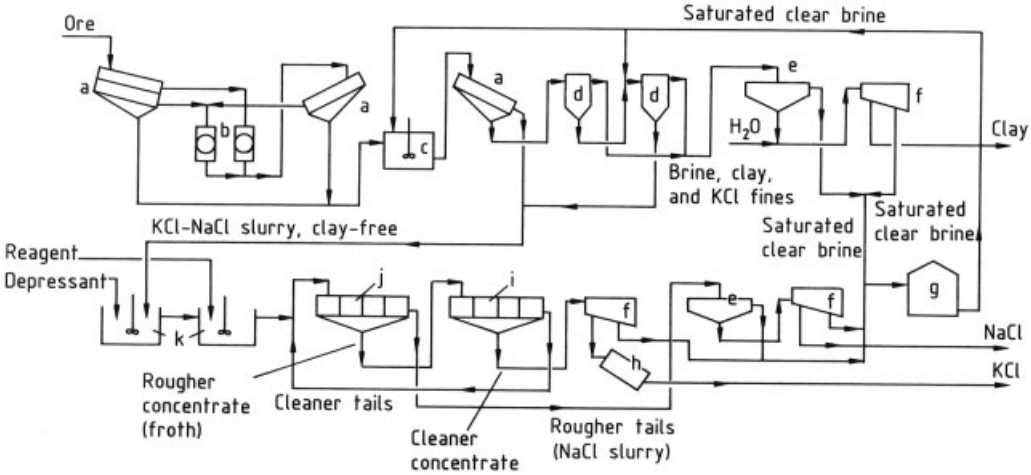

Figure 32. Schematic of flotation of potash ores
a) Screen; b) Crushers; c) Scrubber; d) Classifier; e) Thickener; f) Centrifuge; g) Brine tank; h) Dryer; i) Cleaner flotation cells; j) Rougher flotation cells; k) Conditioners
Reproduced from [98] with permission

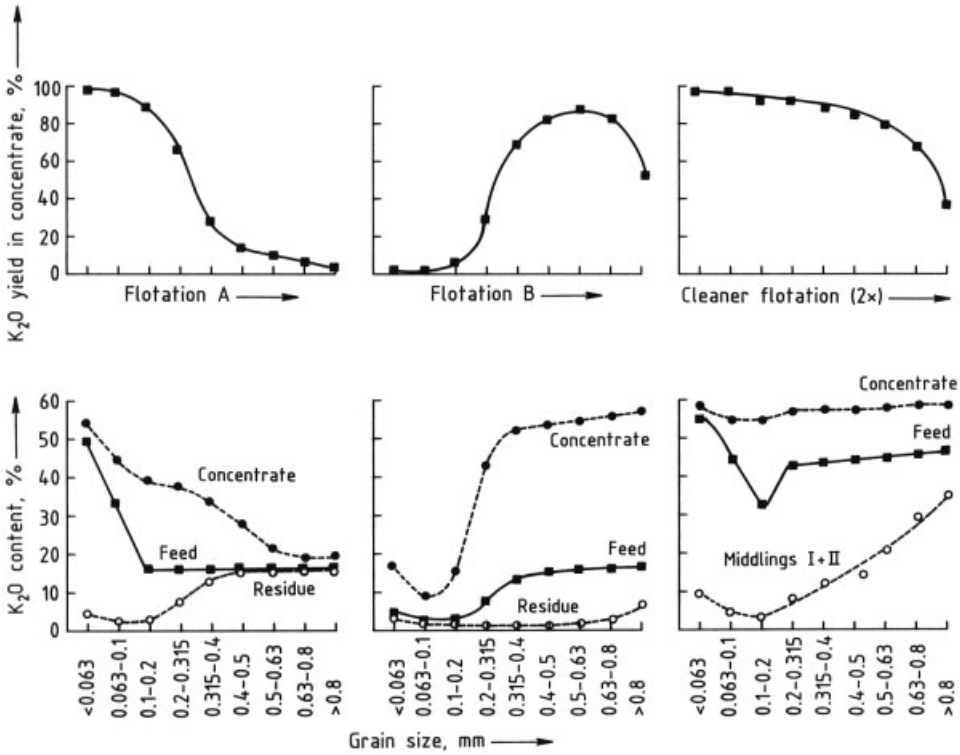

Figure 33. Flotation of fine salt (< 0.1 mm) from carnallite decomposition, showing K_2O content of each size fraction [100]
Yields are based on amounts of K_2O in each fraction

Schoenite can be floated from a salt mixture by using coconut oil acids as the collector [102].

The flotation of anhydrite to produce a salt mixture of kieserite, langbeinite, and polyhalite is recommended in [103].

A combination of cold leaching and flotation for the treatment of salt mixtures containing langbeinite and polyhalite is described in [104].

5.4. Electrostatic Separation

Electrostatic separation depends on the directional movement of electrically charged bodies or particles in an electric field. The processes used differ according to the various methods of producing or separating the charges and according to the separation equipment used (→ Electrostatic Separation). The separation of mixtures of salt minerals (e.g., in the treatment of crude potash ores) invariably involves the selective exchange of electric charges that takes place on contact between the various mineral particles, followed by separation in free-fall separators. Although roll separators are widely used for treating other minerals, they have not been used for salts, mainly because of their low throughput.

The first investigations into the industrial-scale separation of potassium chloride and sodium chloride were carried out after World War II by the International Minerals & Chemical Corp. in Carlsbad, New Mexico [105]. The ore, which contains alkali-metal chlorides and various amounts of clay and sulfate minerals, was ground, heated to ca. 500 °C, cooled to ca. 110 °C, and separated in a free-fall separator with a field strength of 2 – 6 kV/cm. Results were not encouraging and the method was abandoned.

Research in the Kaliforschungs-Institut in Hannover in 1956 led to an industrial breakthrough. The addition of organic and inorganic reagents much improved the electrical charge ex-change between the mineral components, and the separating temperature could be reduced to < 100 °C [106, 107]. The potash works of Neuhof-Ellers, a subsidiary of Kali und Salz, was the first industrial plant to use this process to produce kieserite in 1974. In the following years, plants for the electrostatic production of kieserite and potash concentrates, and for the dry removal of residues, were installed in three other factories, with capacities up to ca. 1000 t/h [108]. Investigations aimed at the introduction of this process for processing sylvinite ores from the potash deposits in Saskatchewan, Canada, have been carried out by PCS Mining, Saskatoon [109].

5.4.1. Theoretical Basis

The basis of the process is the mutual selective exchange of electrical charge between the salt components, which occurs on contact. The direction, selectivity, and intensity of the charge exchange can be influenced by a large number of reagents (conditioning agents) [110]. In addition to this chemical conditioning, treatment with air of specific relative humidity is necessary. This is usually controlled by means of the air temperature and should be between 5 and 25 % [110, 111]. By the appropriate choice of conditioning agent and relative humidity, the charging properties of the individual mineral components in a potash ore can be controlled so that the desired components are recovered. In this way, an ore of complex composition can be completely separated into its components [112].

The mechanism of charging by contact depends on the transfer of electrons between the touching mineral surfaces, which must have suitable surface properties (i.e., surface energies appropriate for the exchange of charge). These energies are influenced (or created) by chemical conditioning agents and partial water vapor pressure [113–115].

5.4.2. Equipment and Processes

Before the separation process, the potash ore must be size-reduced to give complete liberation of its components. Fine particles (< 0.1 mm) behave nonselectively in an electrostatic field, so the grinding process must be carried out as carefully as possible (e.g., by impact grinding). The conditioning agents (20 – 100 g/t ore) are added to the ore in a mixer or introduced in the vapor state into the fluidized-bed dryer that heats the salt to the separation temperature (25 – 80 °C), whereby the relative humidity is adjusted to a suitable value for selective charging of the salt particles [116–118]. Alternatively, a rotary dryer can be used. Depending on the conditioning

agent used and the relative humidity and temperature, sylvite generally becomes positively charged, whereas halite and kieserite can be positive or negative. Accompanying minerals such as langbeinite, carnallite, or kainite can be separated individually or together with other mineral components [119–122].

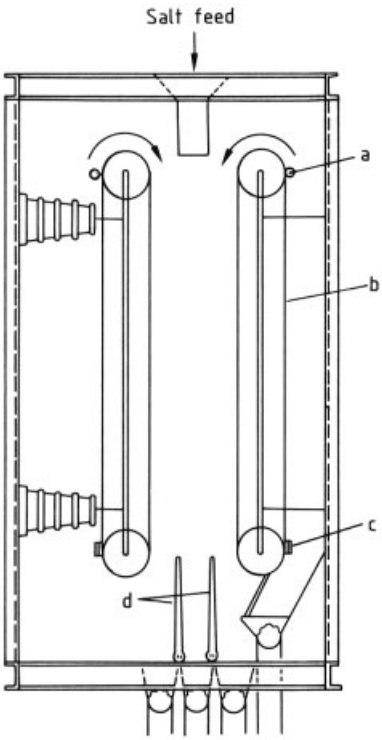

Figure 34. Vertical belt separator
a) Rotary brush; b) Rubber belt; c) Stationary brush; d) Flaps

Separation of the mixture of charged minerals is carried out in free-fall separators with an electrical field strength of 4 – 5 kV/cm. For a distance between the electrodes of 25 cm, the applied voltage is 100 – 125 kV. Formerly, vertical belt separators (Fig. 34) were common, but they are now used only for special applications. The electrodes consist of rotating rubber belts with conducting coatings. Brushes on the side opposite the electric field remove salt fines, which otherwise settle on the electrode surface, forming a coating that weakens the electric field and hinders separation. The separator used exclusively today is the tube free-fall separator (Fig. 35), which was developed by the potash industry [123]. It consists of two opposed rows of steel tubes, ca. 2 m long. The tubes rotate on their axes, and salt fines are removed by brushes on the side remote from the falling salt. The maximum working length of separators of this type is ca. 10 m. The charged salt mixture leaving the fluidized bed is fed to the top of the separator and falls through the electric field between the electrodes. This causes the particles to move sideways, the direction depending on the sign of the charge. At the bottom of the separator, adjustable flaps enable a cathodic fraction, an anodic fraction, and a middle fraction to be recovered. These fractions are then treated further or removed from the process as product and waste (or middlings).

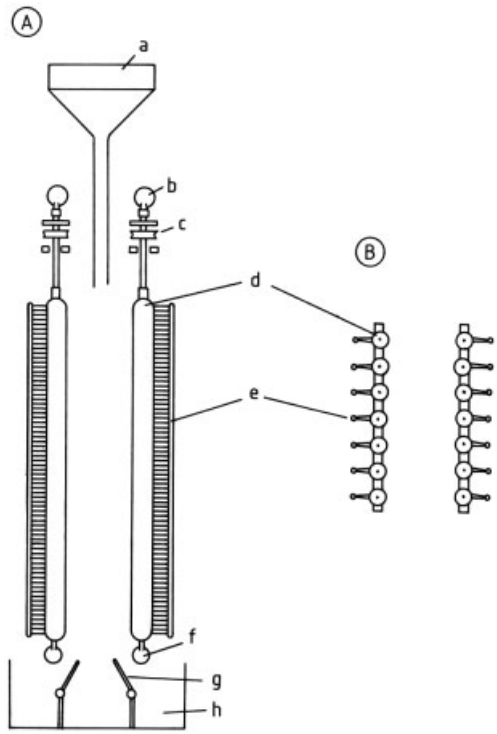

Figure 35. Tubular free-fall separator
A) Cross section; B) Plan viewa) Salt feed; b) Upper bearing; c) Motor; d) Tubes; e) Brushes; f) Lower bearing; g) Flaps; h) Receiver for products

The particle sizes for complete liberation should not exceed 1.5 –2 mm for sylvite and halite, or 1.2 mm for kieserite, because the movement of particles in the electric field of the separator is determined by both the horizontal

electrical force and the vertical effect of gravity. Since the electrical force depends on the charge on the particle, which in turn is a function of the surface area, the surface – volume ratio is a very important factor in determining the extent of sideways movement in the electric field. For fine materials, the effect of the electrical force is the greater; the converse is true for coarse particles.

Modern tube free-fall separators have a throughput of $20-30\,t\,h^{-1}\,m^{-1}$. Their energy consumption is very low because for this throughput the current is only ca. 2 mA.

Generally, even for a single-stage separation step two free-fall separators are arranged one above the other so that the concentrate and waste material from the upper separator bypass the lower one, while the middlings flow directly to the lower separator where further separation occurs. This is shown schematically in Figure 36. In most cases, multistage separation or treatment is necessary, in which the concentrate produced in the first stage is purified or concentrated in another single stage or in multistage separation (Fig. 37).

Electrostatic separation has thus far been used industrially on a large scale only for the treatment of hard salts with the principal components sylvite, halite, kieserite, and various percentages of carnallite. Three basic methods of separation into the individual components are

1) Separation of sylvite in the first stage [105, 111, 124], and separation of kieserite and halite in the second stage
2) Separation of kieserite in the first stage [125, 126], and separation of sylvite and halite in the second stage
3) Separation of halite in the first stage [127–129], and separation of kieserite and sylvite in the second stage

In all cases, carnallite appears in the sylvite fraction.

The first stage of method 2 is used to produce kieserite on a large scale, and method 3 is used in several variations for the production of a dry halite residue and of kieserite and potash concentrates [108]. So far, these plants have been operated almost entirely in association with leaching and flotation plants because electrostatic treatment alone gives unsatisfactory yields.

5.5. Heavy-Media Separation

Heavy-media separation is not used widely in the potash industry because of the generally unfavorable extent of intergrowth and the small density difference between the main components sylvite ($\varrho = 1.99$) and halite ($\varrho = 2.17$). However, the process can be used for very coarsely intergrown high-quality sylvinite ores as mined in Saskatchewan. International Minerals & Chemical Corporation (IMCC) operates a plant of this type in Esterhazy, Saskatchewan [130]. The heavy medium is a suspension of magnetite in a saturated salt solution whose density is adjusted to 2.10. Small salt crystals (< 1.7 mm) are very difficult to separate from magnetite; therefore, only materials with a particle size of 1.7 – 8 mm are treated by this process. Separation is carried out in hydrocyclones in two stages. A process scheme is given in Figure 38.

IMCC also operates a heavy-media separation plant in Carlsbad, New Mexico, where langbeinite ($\varrho = 2.83$) is recovered from a potash ore in which sylvite and halite are the other main components.

5.6. Debrining and Drying

The products and waste from all the treatment processes except for the dry electrostatic process are obtained as suspensions with various solids contents and must be debrined. The two main aims of debrining potassium chloride product are (1) to achieve as low a moisture content as possible to minimize drying costs and (2) to remove as much of the adhering brine as possible to maximize product purity. As the brine adhering to the residues contains potassium chloride, its recovery minimizes yield losses.

Suspensions must often be thickened in circular thickeners or hydrocyclones before debrining.

The choice of equipment is determined mainly by the particle size of the material to be treated. The extent to which adhering brine can be removed by washing with water is also important.

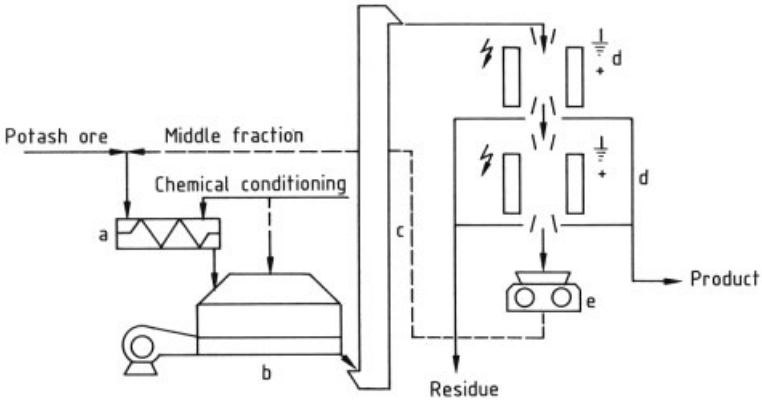

Figure 36. Schematic of single-stage electrostatic separation
a) Mixer; b) Fluidized-bed preheater; c) Elevator; d) Free-fall separator; e) Middle-fraction grinding

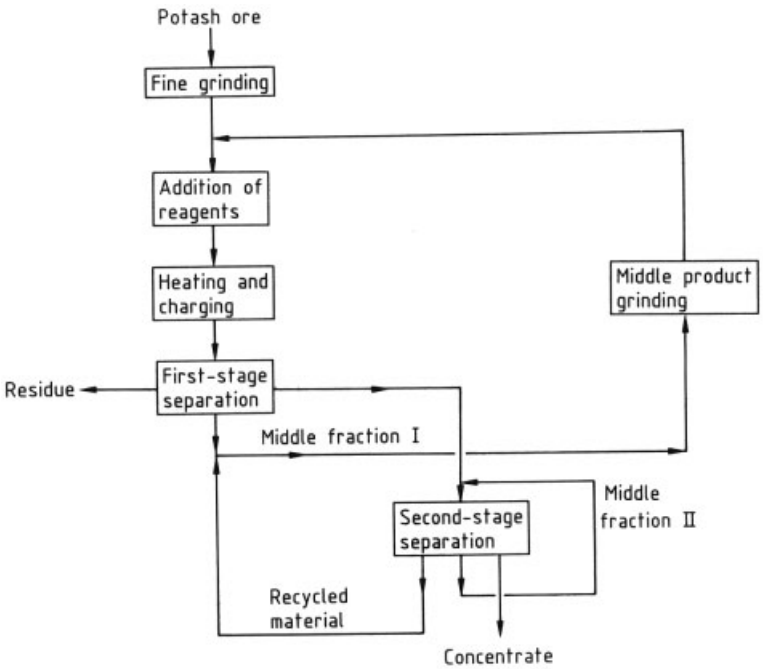

Figure 37. Flow diagram of electrostatic separation plant

Pan filters are used where debrining of a fine-grained product is combined with water washing to increase the K_2O content and when the product is to be stored intermediately. Residual moisture content in this case is 12 – 14 %.

Drum filters (→ Filtration, Chap. 8.6.) are generally used for debrining fine residues or when washing of the filter cake is necessary; they give a residual moisture content of 9 – 11 %. Alternatively, belt filters (→ Filtration, Chap. 8.2.) are used because they have a high capacity and allow the filter cake to be washed with recovery of the washing liquids.

The most commonly used debrining apparatus consists of centrifuges of various designs (→ Centrifuges, Filtering) [131]. In potash works in Canada, the United States, Jordan, and the CIS, screen-bowl and solid-bowl centrifuges of 1400-mm diameter are in general use. Screen-bowl centrifuges have throughputs

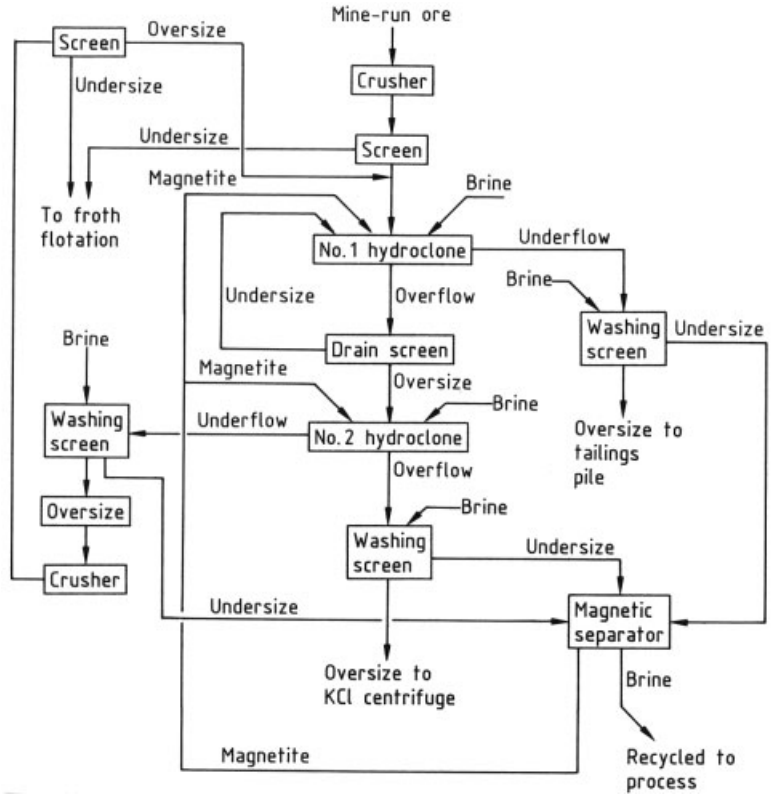

Figure 38. Heavy-media separation of potash ore [130]
Reprinted from [130] with permission of John Wiley & Sons, Inc.

of 60 – 110 t/h and achieve a residual moisture content of 3 – 8 %. They are used to treat both products and residues. Solid-bowl centrifuges mostly are used for fine residues, and achieve throughputs of 70 – 120 t/h and a residual moisture content of 6 – 8 %. In European and some Canadian factories, large-diameter (> 900 mm) pusher centrifuges are used, with throughputs of 40 – 50 t/h and a residual moisture content of 3 – 6 %. For coarse products and residues, vibratory screen and screw screen centrifuges are generally used. These have a diameter of 900 – 1200 mm and a throughput of 35 – 70 t/h, and give a residual moisture content of 2 – 4 %.

The products are usually dried in drum dryers (→ Drying of Solid Materials, Chap. 2.1.2.) with diameters up to 3 m and lengths of 20 m [132]. They are heated with oil or gas in cocurrent flow at throughputs up to 120 t/h. The dryers are fitted with internals to promote heat exchange and prevent caking of the salt. Exhaust gases are dedusted, first by cyclones and then by electrostatic filters, wet scrubbers, or fabric filters. The main reasons for the widespread use of drum dryers are their ruggedness, safe operation, lack of sensitivity to throughput variations, and adaptability to differing grain sizes and moisture content. Figure 39 gives a schematic diagram of a drying plant of this type.

Since the early 1960s, fluidized-bed dryers have been used to an increasing extent (→ Drying of Solid Materials, Chap. 2.1.3.). At first, they were used mainly for coarse products (> 0.5 mm) with low initial moisture content. Later, improvements in fluidized-bed technology enabled products with grain size down to 0.1 mm and initial moisture content up to 8 % to be dried. Their most important advantages compared with drum dryers are improved heat and mass transfer, more efficient use of energy, and much smaller floor space requirement. Cyclones and bag or electrostatic filters are used for dedusting the exhaust gases from a fluidized-bed dryer. A typical plant is shown in Figure 40.

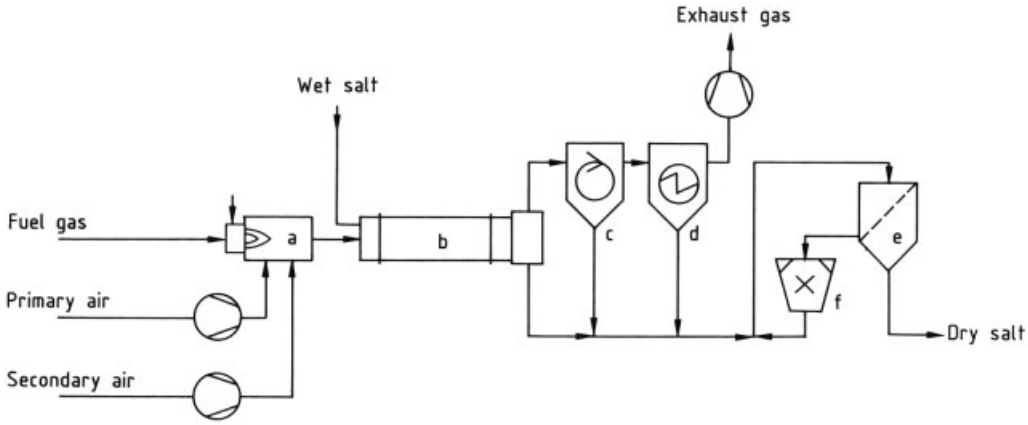

Figure 39. Drying plant for potassium chloride
a) Combustion chamber; b) Rotary dryer; c) Deduster cyclone; d) Electrostatic filter; e) Screen; f) Grinding equipment

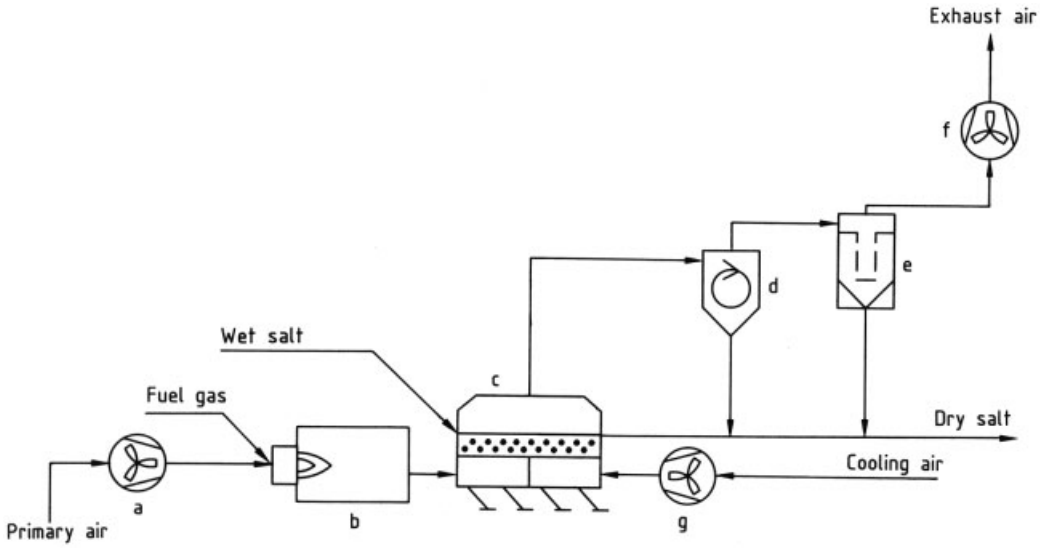

Figure 40. Fluidized-bed dryer with dedusting equipment
a) Fan; b) Combustion chamber; c) Fluidized-bed dryer; d) Deduster cyclone; e) Suction bag filter; f) Exhaust fan; g) Fan for cooling air

5.7. Process Measurement and Control

The methods normally used in the chemical industry for measuring and controlling process parameters are also used in potash plants. Special methods include the analytical determination of potassium by means of its natural radioactivity [133], the use of flame photometry for so-called ratio analysis, and mineral analysis by infrared spectrometry.

The natural mixture of potassium isotopes includes ca. 0.012 % of the radioactive isotope ^{40}K, which is both a β and a γ emitter. Both types of radiation are measured to determine potassium content. The β emissions are usually measured in the laboratory, whereas on-line measurement of γ emissions is widely used for process control. Owing to the great penetrating power of γ rays, the reliability of measurement is strongly dependent on the geometry of the measuring equipment. Equipment for the radiometric determination of potassium in bulk products is shown in Figure 41 [134].

The on-line method of ratio analysis [135, 136] of KCl – NaCl mixtures is based on measuring the potassium and sodium contents of

a sample with a double-beam flame photometer, and calculating the ratio from the measured data. Since this ratio is independent of the weight and dilution of the sample wet materials with varying water content can be analyzed without weighing the sample. Ratio analysis is used for determining residual NaCl content in flotation concentrates that have been water washed. The NaCl content is used to determine the amount of washing water required, and enables losses of K_2O due to excessive use of washing water to be avoided. The method has also been used for controlling the K_2O content in products from leaching – crystallization plants.

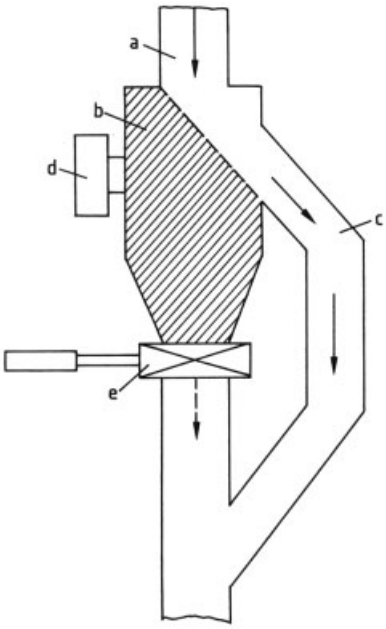

Figure 41. Equipment for radiometric determination of potassium in solid materials
a) Inlet for material to be analyzed; b) Container for material to be analyzed; c) Bypass; d) Detector; e) Time-controlled valve

Infrared spectroscopy is suitable for continuous determination of carnallite or kieserite in crude potash salts. Reflection photometry can be used to measure water of crystallization, and hence the content of these minerals. In some circumstances, the carnallite content can be determined even when kieserite is present [136].

5.8. Waste Disposal and Environmental Aspects [137, 138]

The main environmental problem of the potash industry is disposal of process waste. The total world production of potash ore is ca. 250×10^6 t/a, whose processing necessitates the disposal of ca. 200×10^6 t of waste without damage to the environment [137]. In the Canadian province of Saskatchewan alone, 300×10^6 t of solid waste has been generated during the last 30 years, covering an area of ca. 35 km^2 with solid and liquid materials [139].

The composition of the waste depends on the type of ore treated. Waste from the treatment of sylvinite consists mainly of halite. Waste materials from hard salt treatment are halite and kieserite, and from carnallitite ore processing, halite and magnesium chloride, which is always produced in the form of an aqueous solution. Salt solutions that must be disposed of are also generated during the production of potassium sulfate from potassium chloride and magnesium sulfate, and during the recovery of kieserite from residues from the treatment of hard salt by dissolving the halite (→ Magnesium Compounds, Chap. 5.3.1.).

Four methods for disposing of waste are dumping, backfilling, pumping into the ground, and discharge into natural water systems.

The disposal of waste by dumping is by far the most important method. Salt solutions that run off the dumped materials must be demonstrated not to harm the environment when they are absorbed into the ground. Salt solutions can originate from the brine adhering to wet residues or from the carrier liquid for transporting solid waste to the dump, or they can form when atmospheric precipitation dissolves the salt from waste material. If the ground underneath the dump is not impermeable, it must be sealed by layers of clay or plastic sheeting, or a combination of the two. The salt-containing runoff water is collected in ditches at the edges of the dump, and as much as possible is returned to the recirculating brine system in the plant. Excess brine is disposed of along with other liquid waste. In Germany, solid waste is formed into steep conical heaps after drying or debrining as fully as possible. This reduces the amount of salt-containing runoff water formed by atmospheric precipitation and also minimizes the

ground area required. In most Canadian installations, filtration residues (tails) are slurried with brines and pumped as a suspension into a large lagoon surrounded by dykes. Flat deposits are formed over a very large area. The brine that runs off is pumped back to the plant and reused for slurrying solid waste (Fig. 42) [140, 141]. If the salt solution enters groundwater-bearing layers despite sealing, boreholes are sunk, and salt-bearing groundwater is pumped back into the lagoon. Attempts to cover dumped waste material to prevent the formation of salt-containing water by the leaching effect of rainfall have in the long run proved unsuccessful [139].

Under certain geological conditions, and if mining methods are suitable, solid residues can be transported underground for backfilling. Since the bulk density of the residue is much lower than that of the potash ore, often only a part of the residue can be accommodated by the space left after extraction of the ore. Backfilling is the main method of waste disposal in North German salt works where the salt beds are steeply inclined, as well as the potash works of New Brunswick in Canada. In most potash works, where the potash is mined from level deposits, backfilling is not possible for technical and economic reasons.

Pumping salt solutions back into the ground is possible if certain geological requirements are met. The formation used for this purpose must possess sufficient porosity and permeability, and must have no contact with formations that could provide a water supply or contain salt deposits. Salt solutions are generally pumped under pressure through lined boreholes into the porous formation.

Since 1926, in the Werra potash region of Germany, large quantities of brine have been pumped into a porous dolomite layer 20 – 25 m thick. Injection wells have an absorption capacity up to 1000 m^3/h at head pressures up to 11 bar [142]. Since most of this waste brine comes from kieserite production, the amount produced has decreased drastically with the introduction of electrostatic ore treatment [143]. In Saskatchewan, excess brine is pumped into deep formations of dolomite or sandstone. The capacity of these wells is up to ca. 200 m^3/h at head pressures up to 60 bar. Kalium Chemicals, Saskatchewan, has been disposing of waste brine since 1979 in caverns produced by solution mining.

The possibility of disposing waste brine in rivers and lakes depends very much on the location. Sea disposal, as practiced in the United Kingdom and by one of the potash works in New Brunswick, presents few major problems if the outfall is sufficiently remote from the coast. The potash works on the Dead Sea can dispose of waste brine in the Dead Sea itself without harmful consequences. The Potash Company of America in Saskatchewan discharges its solid and dissolved wastes into the southern end of the salt-containing Lake Patience, which is isolated from the rest of the lake by a dam.

For most potash works, waste brine cannot be disposed in natural salt water. If underground disposal is impossible, disposal into flowing natural water is the only alternative. In all countries, this is subject to ever-stricter regulation. One particular problem, for which a solution in the foreseeable future is being sought, is the high salt content in the Werra River due to brine from the potash works in the eastern part of the Werra region [143].

The only significant atmospheric pollution caused by potash works is salt dust emitted by the drying plant and from the handling of the ore and products during production and supply. Dust removal from waste gases from dryers is discussed in Section 5.6. The dusts produced during production and conveying are usually removed by air extraction, and trapped in fabric filters or wet scrubbers. During the drying of products containing magnesium chloride brines, hydrolysis of the magnesium chloride can lead to the emission of hydrogen chloride, which can be removed from the gas by wet scrubbing or absorption in calcium hydroxide in combination with a woven filter [144]. This procedure can also greatly reduce the level of sulfur dioxide if it is present in the exhaust gas.

5.9. Granulation [145–149]

Potassium chloride can be produced in a wide range of crystal sizes, depending on the composition of the potash ore, its degree of intergrowth, and the process used. Different particle-size distributions are needed for various applications. To meet these requirements, the potash industry

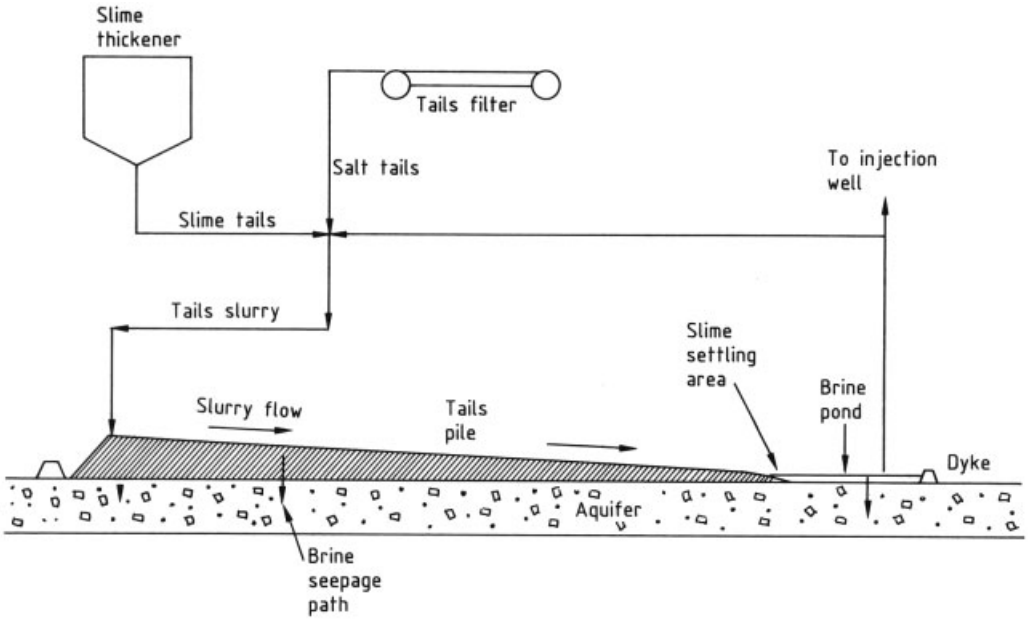

Figure 42. Conventional potash waste handling in Saskatchewan [139]
Reproduced with permission

offers products with standardized size distributions (see Section 5.10) obtained by screening to give the various fractions. The resultant distribution of the product among the various standardized grades does not always correspond to market requirements. Demand for products with a particle size of ca. 1 – 5 mm (coarse and granular grades) exceeds their normal production rate in a potash works. Most potash works must therefore increase the proportion of coarse product by granulating part of the primary product.

The main reason for the high demand for granulated potassium chloride is the technique of applying fertilizers that was developed mainly in North America and Western Europe in the postwar years. This technique, which is now the most widely used, requires coarse particles with a rather narrow size range. This is needed both for single nutrient fertilizers and for bulk-blended materials, which are widely used in North America. Also, granulated potassium chloride has a lower tendency to cake or form dust than the fine product.

Two methods of granulation are commonly used in the fertilizer industry: agglomeration of molten or wet material in rotating drums or dishes, or compaction of dry material in roll presses. The latter is the method most often used for potash fertilizers.

The starting material for compaction usually consists of a mixture of fine material from the production process with recycled material from the grinding and screening system of the compaction plant. Although dusts from the plant can also be compacted, a high proportion of fines (< 0.1 mm) must be avoided because they cause problems in the compaction equipment. Amines from the flotation process on the surfaces of the particles can also interfere with compaction and must be destroyed or rendered inactive by heat or chemical treatment.

Roll presses used in the potash industry usually have 60 – 125-cm-long rolls with diameters of 60 – 100 cm (→ Size Enlargement, Chap. 4.5.3.). The feed material, which is usually at 100 –120 °C, is generally predensified by force feeders that feed it into the nip between rolls, where it is deaerated and compressed, with plastic deformation of each particle, to produce a dense sheet of material. The compression force is 40 – 50 kN/cm of roller length. The sheet of material is fed to a prebreaker that size-reduces it for ease of transportation. It then goes to a grinding – sieving plant, which produces either a granular product or granular

and coarse products. Undersize material is recycled to the compaction press (→ Fertilizers, Chap. 5.3.5.). A typical flow diagram for a compaction – granulation plant is given in Figure 43.

The compaction product consists of irregularly shaped angular particles. Handling causes abrasion of the edges and corners to form unwanted dust. The granular material is therefore usually treated with water after screening and then dried in a fluidized-bed dryer at 180 – 200 °C. This smooths the corners and edges, and gives a dense surface to the granules. Another method of dust reduction is to treat the granules with liquids that bind the dust formed by abrasion (see Section 5.10).

5.10. Quality Specifications

The main application of potassium chloride is in potash fertilizers, either as a single-nutrient fertilizer or as the potash component in mixed or complex fertilizers. In English-speaking areas and on the international market it is often called muriate of potash (MOP).

Single-nutrient fertilizers, which were formerly the most common type, have been replaced largely by complex or mixed fertilizers. Also, the finely divided fertilizers formerly used have increasingly been replaced by granulated products. Single potash fertilizers often contain the additional nutrient magnesium sulfate.

Potassium chloride must have differing grain structure and nutrient content depending on its intended use, e.g., for the production of granulated NPK or PK fertilizers, for application as suspension or liquid fertilizers, for bulk blending with other components (→ Fertilizers, Chap. 3.), or as a single-nutrient fertilizer. The potash industry has developed internationally accepted quality standards.

The three main grades differ in grain size:

Standard	0.2 – 0.8 mm
Coarse	0.8 – 2.0 mm
Granular	1.2 – 3.5 mm

Other products, such as those designated soluble, special standard, or special fine, can contain a high proportion of grains <0.2 mm. Grain-size distributions of the principal grades are listed in Table 7. Products from various manufacturers can diverge considerably from these specifications.

Table 7. Typical grain-size distribution in potash fertilizers (cumulative percentage retained)

Mesh	Tyler	Cumulative subsieve fraction		
width, mm	mesh	Granular	Coarse	Standard
+3.36	+ 6	2 – 12		
+2.38	+ 8	30 – 45	2 – 20	
+1.68	+ 10	75 – 90	25 – 50	
+1.19	+ 14	95 – 98	70 – 90	5 – 15
+0.84	+ 20	99	90 – 98	20 – 45
+0.60	+ 28		99	50 – 75
+0.42	+ 35			70 – 90
+0.30	+ 48			85 – 95
+0.21	+ 65			96 – 98
+0.15	+100			97 – 99

The granular and coarse products are used for bulk blending with other granulated fertilizers or as single-nutrient fertilizers. The standard and other fine materials are used in the manufacture of granulated multinutrient fertilizers, and for suspension and liquid fertilizers.

The potassium chloride content, which for fertilizers is usually given as % K_2O (100 % KCl = 63 % K_2O), is generally at least 60 % K_2O (95 % KCl) for this application. The main impurities include sodium chloride, or magnesium sulfate, and sometimes anhydrite or clay minerals, depending on the raw materials and production method. Fertilizer-quality potassium chloride produced by flotation is often colored red to red-brown by hematite inclusions or salt clay.

For single-nutrient fertilization, especially in European agriculture, potash fertilizers with a low K_2O content are often used. These generally contain a guaranteed level of water-soluble magnesium salts. In Germany, for example, two granulated potash fertilizers with a guaranteed MgO content are widely used: 40er Kornkali with MgO (40 % K_2O, 6 % MgO) and Magnesia-Kainit coarse (11 % K_2O, 5 % MgO, 24 % Na_2O).

About 5 % of the potassium chloride produced worldwide is used as an industrial chemical, mainly for the production of potassium hydroxide (see Chap. 11) and chlorine by chloralkali electrolysis. Material of this quality is produced by recrystallizing low-purity potassium chloride. A distinction is made between industrial-grade and chemical-grade material, depending on the impurity content and area of use. Typical analyses of these products are

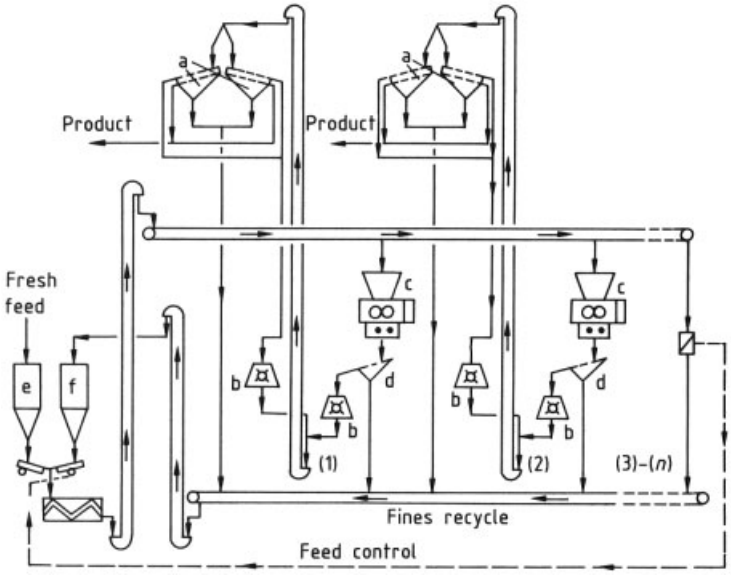

Figure 43. Flow sheet of a modern potash compaction – granulation plant [149]
a) Double screens; b) Hammer mills; c) Compactors; d) Screens; e) Fresh feed vessel; f) Recycled materials vessel
Reproduced from [149] with permission

Industrial grade:	KCl:	99.0 – 99.5 %
	NaCl:	0.8 – 0.3 %
Chemical grade:	KCl:	99.8 – 99.9 %
	NaCl:	0.05 – 0.03 %

The remaining impurities consist mainly of bromide and alkaline-earth sulfates, depending on the raw material and production process.

Dust content and free-flowing properties are important quality criteria, and depend on chemical composition and grain size. With granulated materials, dust formation by abrasion occurs due to handling during manufacture and transport even if the materials are specially treated during production (see Section 5.9). To bind the dust, the granulated material is treated with conditioning agents such as polyglycols, mineral oil products, vegetable oils and waxes, or mixtures of these. The addition rates are usually 0.3 – 5.0 kg/t [150, 151]. Various methods have been developed to assess or measure the abrasion resistance and dust content of the granulated product. Abrasion resistance is tested by subject-ing samples of the granules to a process of screening or rolling in a rotating drum with steel spheres or rods, and weighing the fine material formed after a given time [152]. Dust content is determined by air blowing a sample under precisely defined conditions and weighing the dust collected [150].

Fine potassium chloride (standard quality or finer) tends to cake on storage or transport over long distances. Anticaking agents are therefore generally added, usually mixtures of aliphatic amines, or sometimes higher fatty acids. Salt products that do not contain alkaline-earth sulfates are treated with potassium hexacyanoferrate(II). Addition rates for organic anticaking agents are 50 – 300 g/t, but potassium hexacyanoferrate(II) is effective even at 15 – 25 g/t [150, 151]. Anticaking agents can affect the wettability of potassium chloride, which in turn can affect its granulation properties when complex fertilizers are being produced. The anticaking treatment must therefore suit the user's requirements exactly [153].

5.11. Toxicology and Occupational Health

No toxic hazards are associated with the normal handling of potassium chloride. According to USP XVII of 1970, the usual therapeutic dose (e.g., for treating potassium deficiency) taken orally would be 1 – 10 g/d. The LD_{50} (oral, guinea pig) is 2500 mg/kg.

Protective measures for storage and handling and personal protection such as breathing apparatus or gloves are unnecessary.

5.12. Economic Aspects and Uses

A review of world potash production since 1900 is given in Table 8. The figures show the total output from the potash industry including potassium sulfate and potassium products for industrial use [4, 154, 155].

Germany was the sole producer up to World War I, and was later joined by France and the United States, and then Spain, the Soviet Union, and Poland. After World War II, the leading producers were the central and western European countries and the United States. In the 1960s, the Soviet potash industry grew strongly, and the Soviet Union became the leading producer. Also in the 1960s, the first potash works in Saskatchewan was started up. In a few years, several large potash works were in operation there, and Canada became the second largest producer after the Soviet Union. The capital investment in the Soviet Union and Canada and the rapidly increasing use of fertilizers in agriculture in the 1960s and 1970s led to a steep increase in world potash production. Since 1980, the average annual increase in world potash production has been only 0.7 %.

Almost two-thirds of world potash production is exported. All the potash-producing countries are exporters except for Brazil and China. Canada is by far the largest exporter. World exports of potash in 1989 (in 10^3 t K_2O) were as follows [154]:

Canada	6 808
France	517
former East Germany	2 251
West Germany	1 339
Israel	1 007
Italy	64
Jordan	755
Spain	354
United Kingdom	251
United States	482
former USSR	3 407
Total	17 235

The economic situation, particularly in developed countries, greatly influences the extent and regional distribution of exports. Both the quantity exported and its distribution among consumers are greatly affected by the state of their agriculture, especially in developed regions, and by the demand for or availability of convertible currency in the exporting or importing country. Transport costs for potash fertilizers have a considerable bearing on total cost to the consumer, and logistical considerations also influence the direction and size of exports or imports. Finally, fluctuations in the rate of exchange of currencies between the countries concerned are very important.

Because of the conditions described above, certain special regional relationships developed. The agricultural requirements of the former COMECON countries were satisfied by the Soviet and East German potash industries only. In western Europe, the market was supplied almost entirely by western European producers. In North America, Canadian and United States producers were in a dominating position. Canada had a good export market in Asia, as did Europe and Jordan. The Latin American market was supplied mainly by Canadian producers, together with former East Germany and the former Soviet Union. The political and economic changes in the former Eastern Bloc, the unification of Germany, and the collapse of the dollar have all led to changes in the supply–demand relationships described above, although the distance between the producer and the consumer is still of overriding importance. The future development of potash exports will be influenced greatly by imports into China and Brazil. Both countries have only minimal production and are compelled to import fertilizers on a large scale. Problems associated with their internal economies and with foreign exchange have thus far limited imports [156, 157].

The capacities of the potash producers in various countries for 1990–1991 (in 10^3 t K_2O) were [155]:

Middle and Western Europe	8 700
Germany	5 700
France	1 500
Spain	750
Italy	250
United Kingdom	500
North America	13 028
United States	1 838
Canada	11 190
former Soviet Union	12 880
Western Asia	2 220
Israel	1 380
Jordan	840

Table 8. World potash production (in 1000 t K_2O) by country

Country	Year													
	1900	1910	1920	1930	1940	1950	1960	1970	1980	1985	1986	1987	1988	1989
Germany	322	905	924	1381	1746									
West Germany						906	1978	2306	2737	2583	2161	2201	2290	2186
former East Germany						1200	1598	2419	3422	3465	3485	3510	3510	3200
France			194	506	517	896	1580	1765	1894	1745	1629	1539	1502	1195
Spain				25	88	159	265	521	658	659	702	740	766	742
Italy							25	152	102	143	109	117	126	154
United Kingdom									306	337	391	429	460	463
Middle and Western Europe	322	905	1118	1912	2351	3161	5446	7163	9119	8932	8477	8536	8654	7940
United States			44	56	344	1168	2394	2467	2240	1245	1171	1218	1461	1580
Canada								3173	7300	6637	6698	7267	8327	7360
North America			44	56	344	1168	2394	5640	9540	7882	7869	8485	9788	8940
Poland			1	70										
former Soviet Union					221	312	1084	4087	8064	10 367	10 228	10 888	11 300	10 231
Eastern Europe			1	70	221	312	1084	4087	8064	10 367	10 228	10 888	11 300	10 231
Israel					45		83	546	790	1163	1240	1253	1244	1273
Jordan										545	662	722	805	792
Dead Sea					45		83	546	790	1708	1902	1975	2049	2065
Brazil											11	37	48	109
Congo								123						
China								26	20	40	24	40	54	56
Others								149	20	40	35	77	102	165
Total	322	905	1163	2038	2961	4641	9007	17 585	27 533	28 929	28 511	29 961	31 893	29 341

Brazil	150
China	50
Others	200
Total	37 218

Estimated world demand for potash fertilizers in the business year 1990 – 1991 by regions (in 1000 t K_2O) was as follows:

Europe	7 520
Eastern Europe	2 120
Western Europe	5 400
former Soviet Union	5 600
America	7 320
North America	5 100
Central America	350
South America	1 870
Asia	4 705
Western Asia	155
Southern Asia	1 360
Eastern Asia	3 190
Africa	522
Oceania	260
World	25 927

In using these figures, it should be borne in mind that the capacities quoted by individual producers are generally too high and that only ca. 95 % of the total potash production is used in the form of fertilizers. Hence, the total consumption of products of the potash industry is ca. 1.5×10^6 t of K_2O higher than the figure given above. Nevertheless, considerable overcapacity exists worldwide, as can be seen by comparing the two tables.

For the use of potassium compounds in fertilizers, see → Fertilizers.

Industrial-grade and chemical-grade potassium chloride are used mainly for the electrolytic production of potassium hydroxide. Other important uses include the production of drilling fluids for the oil industry, aluminum smelting, metal plating, production of various potassium compounds, and applications in the food and pharmaceutical industries [158].

6. Potassium Sulfate [159]

Potassium sulfate [*7778-80-5*], K_2SO_4, mineral name arcanite, forms colorless, nonhygroscopic crystals. It occasionally occurs in nature in the pure state in salt deposits (e.g., in Germany, the United States, and the CIS) but is more widely found in the form of mineral double salts in combination with sulfates of calcium, magnesium, and sodium (see Table 1). Potassium sulfate is, after potassium chloride, the most important potassium-containing fertilizer, being used mainly for special crops. Potassium sulfate constitutes ca. 5 % of the world demand for potash fertilizers.

6.1. Properties

Potassium sulfate forms orthorhombic crystals, which transform to the trigonal modification at 583 °C. Some properties of potassium sulfate are listed below:

M_r	174.25
fp	1069 °C
Crystal system and type	orthorhombic D_{2h}^{16}
Phase change at	583 °C
Crystal system and type at > 583 °C	trigonal D_{3d}^{3}
Refractive indices n_D^{20}	1.4933; 1.4946; 1.4973
Density	2.662 g/cm^3
Specific heat capacity c_p	752.9 J kg^{-1} K^{-1}
Heat of fusion	197.4 kJ/kg^1
Heat of transformation (orthorhombic/trigonal)	48.5 kJ/kg^1
Enthalpy of formation ΔH^0	− 1438 kJ/mol
Entropy S^0	175.6 J mol^{-1} K^{-1}
Dielectric constant (at 4×10^8 Hz)	6.3
Thermal coefficient of expansion (cubic)	130×10^{-6} K^{-1}

Apart from the naturally occurring double-salt minerals mentioned above, potassium sulfate also forms double salts and mixed crystals with ammonium sulfate and the sulfates of beryllium, magnesium, calcium, strontium, barium, and lead. It is reduced to potassium sulfide or potassium polysulfides by reducing agents such as hydrogen and carbon monoxide at high temperature.

The solubility of potassium sulfate in water is listed in Table 9. The cryohydric point is − 1.51 °C (7.1 g K_2SO_4/100 g H_2O), and the boiling point of the saturated solution is 101.4 °C (24.3 g K_2SO_4/100 g H_2O). The solid phases formed in the system $K_2SO_4 - H_2O$ are K_2SO_4, $K_2SO_4 \cdot H_2O$, and ice. In aqueous ammoniacal solution, the solubility decreases rapidly with increasing ammonia concentration [160]. Potassium sulfate is virtually insoluble in industrial organic solvents.

Table 9. Solubility of potassium sulfate in water (g/100 g) [15]

Temperature, °C	0	10	20	30	40	50
Solubility	7.35	9.24	11.1	12.9	14.8	16.6
Temperature, °C	60	70	80	90	100	
Solubility	18.4	20.0	21.5	22.8	24.0	

6.2. Raw Materials

Potassium sulfate is produced from single or mixed minerals or brines, or by the reaction of potassium chloride with sulfuric acid or sulfates [161]. The economically important minerals include the deposits of hard salt in Germany, langbeinite in New Mexico (United States), and kainite in Sicily. In the United States, sulfate-containing crystalline products from the evaporation of water from the Great Salt Lake (Utah) and Searles Lake (California) are also used for the production of potassium sulfate.

Potassium chloride is usually converted to potassium sulfate by reaction with sulfuric acid, but SO_2 – air mixtures can also be used. Reactions with sodium sulfate and gypsum are also of interest. Small amounts of potassium sulfate are obtained during the production of alumina from alunite.

6.3. Production

The choice of production method and the location of a potassium sulfate plant depend on having a plentiful economic supply of the starting materials and being able to utilize or dispose of the byproducts or waste. Most production plants are located on salt deposits from which at least some of the raw materials can be obtained. Plants in which potassium chloride is reacted with sulfuric acid with liberation of hydrogen chloride are usually located in regions with a demand for hydrochloric acid, (e.g., for a acidification of petroleum boreholes), or they operate in conjunction with a chemical works having a process that uses hydrogen chloride. In such cases, the

exploitability of hydrogen chloride determines the capacity of the potassium sulfate plant.

6.3.1. From Potassium Chloride and Sulfuric Acid (Mannheim Process)

The reaction of sulfuric acid with potassium chloride takes place in two stages:

$$KCl + H_2SO_4 \longrightarrow KHSO_4 + HCl$$
$$KCl + KHSO_4 \longrightarrow K_2SO_4 + HCl$$

The first reaction step is exothermic and proceeds at relatively low temperature. The second is endothermic and must be carried out at higher temperature. The relationship between total reaction time and temperature is shown in Figure 44. In practice, the process is operated at 600–700 °C. To minimize the chloride content of the product, a small excess of sulfuric acid is used, which is later neutralized with calcium carbonate or potassium carbonate, depending on the purity requirements for the product.

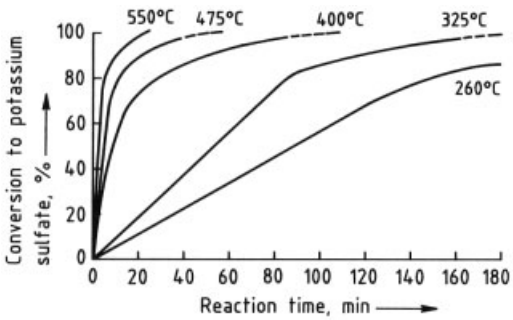

Figure 44. Temperature dependence of the reaction between potassium chloride and sulfuric acid
Reproduced from [162] with permission

The reaction is usually carried out in so-called Mannheim furnaces (Fig. 45) [162].

The furnace has a closed dish-shaped chamber, with diameter up to 6 m, heated externally by an oil or gas burner. Potassium chloride and sulfuric acid are fed into the chamber in the required ratio at an overhead central point. The mixture reacts with evolution of heat and is mixed by a slowly moving stirrer fitted with stirring arms with scrapers (rabbles), which propels the mixture from the center of the chamber to the periphery. Potassium sulfate leaves the reaction chamber at this point and is neutralized and cooled. It normally contains 50–52 % K_2O and 1.5–2 % chloride. Hydrogen chloride gas formed is absorbed in water to form hydrochloric acid or used in gaseous form.

The Mannheim process is the most widely used method of producing potassium sulfate due to its simplicity, high yield, and the many ways in which the byproduct can be utilized. Hydrogen chloride is used to produce dicalcium phosphate, vinyl chloride, or calcium chloride if it cannot be sold as hydrochloric acid.

Disadvantages of the process include high energy consumption, severe corrosion, and high capital cost. In the United States, reductions in corrosion and energy consumption are achieved by using the Cannon process, in which the reaction is carried out in a directly fired fluidized bed.

Another variation is the Hargreaves process, which is also used in the United States. Briquetted potassium chloride is heated in reaction chambers in a stream of sulfur dioxide from the combustion of sulfur, excess air, and water vapor. The yield and the degree of conversion are both ca. 95 %.

The Mannheim and Hargreaves processes are also used to produce sodium sulfate from sodium chloride and sulfuric acid. Mannheim furnaces can be used to produce potassium and sodium sulfates alternately. Research into the reaction of potassium chloride with sulfuric acid in a liquid–liquid extraction process did not result in the construction of a production plant [163, 164].

6.3.2. From Potassium Chloride and Magnesium Sulfate [165, 166]

In a process used mainly in Germany, the sulfate required is provided by kieserite, $MgSO_4 \cdot H_2O$, a component of German hard salt deposits. Methods of extraction are described in → Magnesium Compounds, Chap. 5. The reaction can be represented by the following overall equation:

$$2\,KCl + MgSO_4 \longrightarrow K_2SO_4 + MgCl_2$$

Kieserite reacts very slowly and must be ground finely before reaction. Alternatively, it can first be recrystallized to give epsomite, $MgSO_4 \cdot 7\,H_2O$.

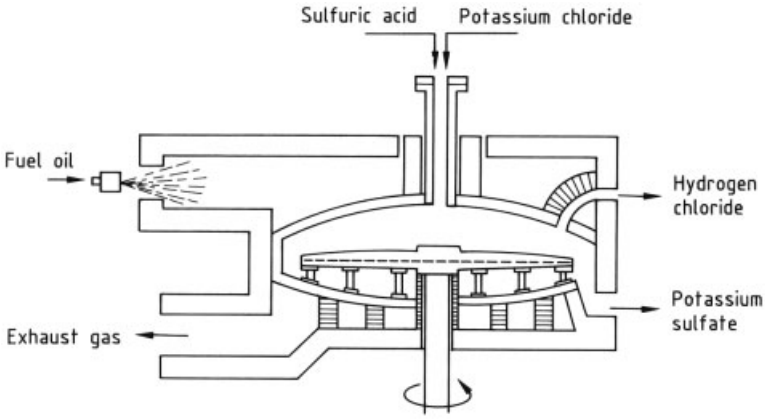

Figure 45. Schematic diagram of a Mannheim furnace
Reproduced from [162] with permission

The basis of the process is explained in Figure 46. The fundamental relationships for the single-stage process of KUBIERSCHKY and the two-stage process of KOELICHEN and PRZIBYLLA are shown as broken lines on the isotherm diagram.

For the single-stage process, the most favorable mixing ratio of the starting materials is given by point C. In the presence of sufficient water, this mixture reacts to form potassium sulfate and a sulfate mother liquor (point m). This solution has the highest magnesium chloride content attainable by direct reaction, which determines the yield. The magnesium content of solution m reaches a maximum at 25 °C, and the process is therefore carried out at this temperature. The single-stage process achieves a theoretical potassium yield of only 46.1 % and sulfate yield of 67.5 %.

For this reason, the two-stage process is now used exclusively. In this process, the starting materials are first mixed in the presence of a definite quantity of water corresponding to point S to form schoenite, $K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O$. So-called potash – magnesia liquor, which has a high magnesium chloride content (point P), is also formed. The schoenite is reacted with additional potassium chloride (point D) to form potassium sulfate and sulfate mother liquor:

$$2\,KCl + 2\,MgSO_4 + x\,H_2O \longrightarrow K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O + MgCl_2\ (aq.)$$

$$2\,KCl + K_2SO_4 \cdot MgSO_4 \cdot 6\,H_2O + x\,H_2O \longrightarrow 2\,K_2SO_4 + MgCl_2\ (aq.)$$

This process gives a theoretical potassium yield of 68 % and sulfate yield of 83.7 %.

The flow diagram of the process is shown schematically in Figure 47.

In the first stage, called the potash – magnesia stage, schoenite or leonite, $K_2SO_4 \cdot MgSO_4 \cdot 4\,H_2O$, is produced by stirring solid epsomite or finely ground kieserite with potassium chloride in sulfate mother liquor recycled from the second stage. The suspension produced is filtered on rotary filters; the potash – magnesia brine, which contains 180 – 200 g/L magnesium chloride, is removed; and the solid crystalline product, also known as potash – magnesia, is fed to the next stage, sometimes after being washed with sulfate mother liquor, where it is stirred with potassium chloride solution at ca. 70 °C. The temperature of the mixture is 35 – 40 °C, and solid potassium sulfate is formed. This is thicknened, debrined by centrifuges, and dried in drum or fluidized-bed dryers.

If the sulfate reaction is carried out in a classifying crystallizer at a high solids content [167], a very pure, coarsely crystalline product is obtained with K_2O content of 53 % and chloride content of < 0.5 %.

In the industrial process, sodium chloride is always present. If the molar ratio of $Na_2 : K_2$ in the sulfate mother liquor exceeds 2 : 5, glaserite, $3\,K_2SO_4 \cdot Na_2SO_4$, is formed instead of potassium sulfate. To prevent this, the potassium chloride used must be of adequate purity.

The yield is determined by losses to the waste brine. The higher the magnesium chloride con-

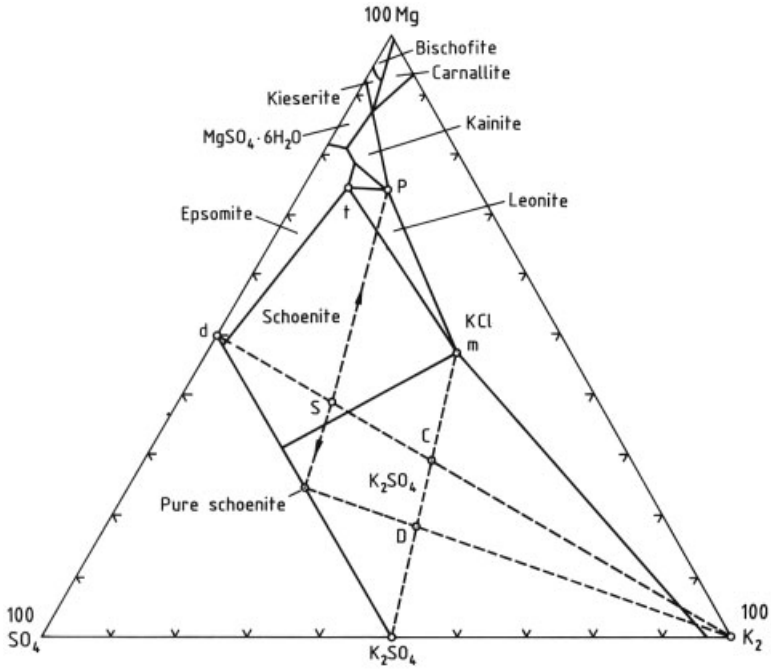

Figure 46. Isothermals of the system $K_2 - Mg - Cl_2 - SO_4^{2-} - H_2O$ at 25 °C according to JÄNECKE

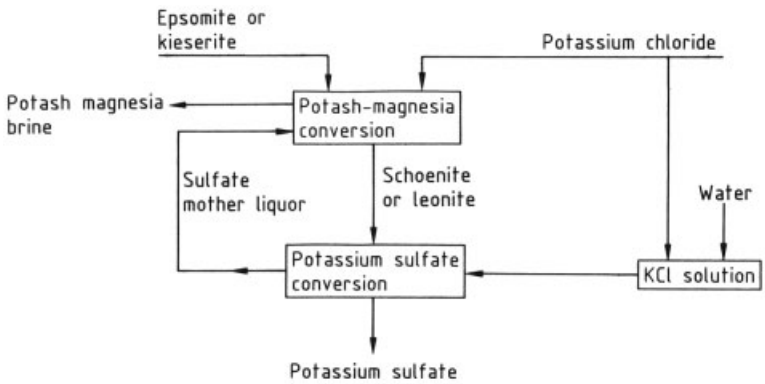

Figure 47. Flow diagram of the two-stage production of potassium sulfate from potassium chloride and magnesium sulfate

tent of the potash – magnesia brine is, the lower is the potassium content, and therefore the potassium loss. Excess potassium-rich sulfate mother liquor must be recovered.

6.3.3. From Potassium Chloride and Langbeinite [168]

Large deposits of langbeinite, $K_2SO_4 \cdot 2\,MgSO_4$, can be found in New Mexico (United States). Langbeinite can be converted to potassium sulfate according to the following overall equation:

$$K_2SO_4 \cdot 2\,MgSO_4 + 4\,KCl \longrightarrow 3\,K_2SO_4 + 2\,MgCl_2$$

The potash ore also contains halite and varying amounts of sylvite, from which langbeinite is separated by gravity separation, flotation, and dissolution of halite, giving various crystal sizes. The coarser langbeinite fraction is sold as potash – magnesium fertilizer, and the finer fraction is reacted with potassium chloride to pro-

duce potassium sulfate. A flow diagram for the process is given in Figure 48.

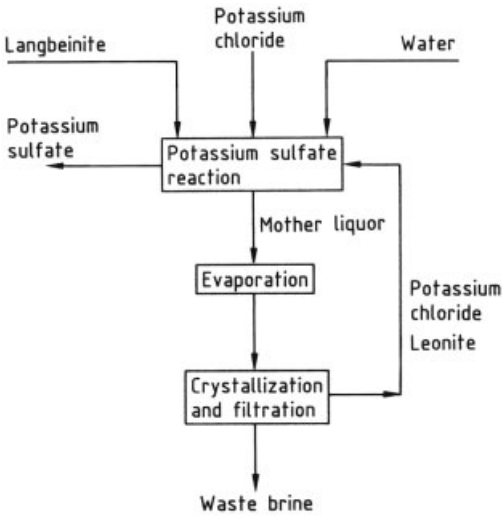

Figure 48. Flow diagram of the production of potassium sulfate from langbeinite

The potassium sulfate formed is granulated and marketed in three different grain sizes: granular (0.8 – 3.4 mm), standard (0.2 – 1.6 mm), and special standard. The latter has a high content of grains < 0.2 mm.

6.3.4. From Potassium Chloride and Kainite [161]

In Sicily, kainite, $KCl \cdot MgSO_4 \cdot 2.75\,H_2O$, is obtained from a potash ore by flotation. It is then converted into schoenite at ca. 25 °C by stirring with mother liquor containing the sulfates of potassium and magnesium from the later stages of the process. Schoenite is filtered off and decomposed with water at ca. 48 °C. This causes magnesium sulfate and part of the potassium sulfate to dissolve and most of the potassium sulfate to crystallize. The crystals are filtered and dried. The sulfate mother liquor is recycled to the kainite – schoenite conversion stage. The mother liquor produced there, which still contains ca. 30 % of the potassium used, is treated with gypsum, $CaSO_4 \cdot 2\,H_2O$, causing sparingly soluble syngenite [*13780-13-7*], $K_2SO_4 \cdot CaSO_4 \cdot H_2O$, to precipitate. Syngenite is decomposed with water at ca. 50 °C, which dissolves potassium sulfate and reprecipitates gypsum. The potassium sulfate solution is recycled to the schoenite decomposition stage, and gypsum is reused to precipitate syngenite. A simplified flow diagram of the process is given in Figure 49.

6.3.5. From Potassium Chloride and Sodium Sulfate [164]

The production of potassium sulfate from potassium chloride and sodium sulfate takes place in two stages, with glaserite, $Na_2SO_4 \cdot 3\,K_2SO_4$, as an intermediate, according to the following equations:

$$4\,Na_2SO_4 + 6\,KCl \longrightarrow Na_2SO_4 \cdot 3\,K_2SO_4 + 6\,NaCl$$
$$Na_2SO_4 \cdot 3\,K_2SO_4 + 2\,KCl \longrightarrow 4\,K_2SO_4 + 2\,NaCl$$

Potassium chloride and sodium sulfate are reacted at 20 – 50 °C in water and recycled process brines to form glaserite, which is filtered and then reacted with more potassium chloride and water to form potassium sulfate. Because the mother liquor from the glaserite stage has a high potassium and sulfate content, the maximum potassium yield is 73 %, and the maximum sulfate yield is 78 %. The yield can be increased considerably by cooling the mother liquor to produce more crystals and by including a final evaporation stage.

A production plant in the CIS uses the glaserite process [166], and an experimental plant is operating in Canada [169].

Alternatively, sodium sulfate solution can be used to charge an anion exchanger with sulfate, which reacts with potassium chloride solution to give a high yield of potassium sulfate [170, 171]. This process is used at Quill Lake, Saskatchewan, Canada, which has a high sodium sulfate content.

6.3.6. From Potassium Chloride and Calcium Sulfate

Processes based on gypsum, $CaSO_4 \cdot 2\,H_2O$, have often been proposed, because it is so readily available. Two processes have been tested in experimental plants:

1) Reaction with potassium chloride in strongly ammoniacal solution

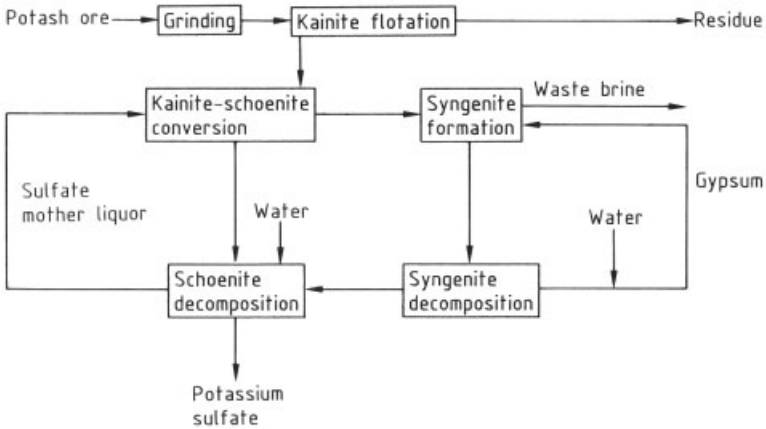

Figure 49. Flow diagram of the production of potassium sulfate from kainite

2) Reaction with anion exchangers

Potassium chloride reacts with gypsum in water to give syngenite. If the reaction is carried out in a concentrated solution of ammonia at low temperature, potassium sulfate with a very low syngenite content is obtained [163, 172, 173].

$$CaSO_4 \cdot 2\,H_2O + 2\,KCl \longrightarrow K_2SO_4 + CaCl_2\ (aq.)$$

By carrying out the reaction in two or more stages, high concentrations of calcium chloride in the waste brine can be produced, and hence very high yields. The ammonia required for the reaction medium must be recovered by distillation. This process has attracted some interest [174].

The anion-exchange process is carried out in two stages [175]. First, the ion-exchange resin (R) is treated with a suspension of gypsum, charging it with sulfate. The charged resin is then treated with a concentrated solution of potassium chloride, to replace the dissolved chloride by sulfate. Potassium sulfate crystallizes from the solution, sometimes after the addition of solid potassium chloride. The process takes place according to the following equations:

$$R_2Cl_2 + CaSO_4 \longrightarrow R_2SO_4 + CaCl_2$$

$$R_2SO_4 + 2\,KCl \longrightarrow K_2SO_4 + R_2Cl_2$$

6.3.7. From Alunite [176]

Alunite [*12588-67-9*], $K_2SO_4 \cdot Al_2(SO_4)_3 \cdot 4\,Al(OH)_3$, occurs in several extensive deposits. On being heated to 800 – 1000 °C, it decomposes with liberation of sulfur trioxide to form a mixture of alumina and potassium sulfate. The latter can be extracted from the mixture. A plant in the CIS uses this process.

6.3.8. From Natural Brines and Bitterns

In the United States, large quantities of potassium sulfate are produced from the brines of the Great Salt Lake and smaller amounts from the brines of Searles Lake (see Section 8.3). Extensive investigations and project studies have been carried out in various countries into the extraction of potassium sulfate from the mother liquors (bitterns) produced when salt is extracted from seawater, and from concentrated brines in Tunisia and the Atacama desert of Chile [164].

6.4. Granulation

The demand for granulated potassium sulfate has increased greatly for the same reasons that the demand for potassium chloride has increased. The method of production with compaction rolls is widely used (see Section 5.9), although potassium sulfate does not behave under pressure like potassium chloride, which undergoes plastic deformation with merging of the grain boundaries. However, the granulated product is much more dense and solid if the

potassium sulfate is wetted before compaction by adding up to 2 % water, or if steam is introduced through force feeders located above the compaction rolls. Also, the pressure used must be considerably higher than that for potassium chloride [177, 178]. A compaction plant for potassium sulfate is otherwise similar to that for potassium chloride, although the throughput is much lower.

6.5. Quality Specifications

The important quality specifications for agricultural-grade potassium sulfate, known in the English-speaking world and on the international market as sulfate of potash (abbreviated to SOP), are the K_2O and chloride content. All producers guarantee a minimum K_2O content of 50 % (92.5 % K_2SO_4), typical values being 50.5 –51.0 %. A completely water-soluble potassium sulfate with K_2O content of > 52 % is supplied for the production of liquid fertilizers.

The maximum permissible chloride content is 3 %, according to EC Guidelines, but the usual commercial upper limit is 2.5 %. The chloride content depends very much on the production method, and some producers offer potassium sulfate with chloride content < 0.5 %.

The most common impurities, apart from alkali chlorides, are the sulfates of calcium and magnesium.

The demand for granulated potassium sulfate as a single-nutrient fertilizer and for bulk blending has greatly increased. The crystal size and size distribution of the products granular, coarse, and standard are the same as those for potassium chloride (see Section 5.10).

Potassium sulfate does not cake as readily as potassium chloride and is therefore not treated with anticaking agents. Dust-reducing agents are added at the rate of 5 kg/t to the fine and granulated products; the same additives are used as for potassium chloride (see Section 5.10).

Industrial-grade potassium sulfate with a K_2SO_4 content of 99.6 – 99.9 %, purified by recrystallization, is supplied as a raw material for the production of other potassium compounds and as an auxiliary material or reagent for various branches of industry. Specifications with respect to chemical purity and crystal size distribution are suited to the individual user's requirements and can sometimes be very detailed.

6.6. Toxicology and Occupational Health

No health hazard is associated with potassium sulfate if it is handled in accordance with regulations. According to Swiss law relating to toxic substances, it is a Class 4 material (LD_{50}: 500 – 5000 mg/kg). Otherwise, the information given in Section 5.11 for potassium chloride applies.

6.7. Economic Aspects and Uses

Potassium sulfate accounts for ca. 5 % of world production by the potash industry, expressed in K_2O units. It is produced in 11 countries. Of the total, ca. two-thirds comes from Belgium and Germany. Other European producers are Italy, Spain, Finland, and Sweden. Potassium sulfate is also produced by the CIS. Other production plants are in the United States, Japan, South Korea, Taiwan, the Philippines, and China. Present world production and consumption both exceed 1.5×10^6 t K_2O [179, 180].

A list of the most important exporting countries (exports for 1985 – 1986 in 10^3 t K_2O) is given below:

Western Europe	586
Belgium	256
Finland	18
West Germany	227
Italy	54
Spain	31
former East Germany	61
United States	74
Asia/Middle East	16
Israel	3
South Korea	8
Taiwan	5
Total	737

The consumption (1985 – 1986 in 10^3 t K_2O) amounted to:

Western Europe	558
Eastern Europe	106
Africa	121
North America	115
Latin America	51
Asia/Middle East	271
Oceania	6
Total	1228

Potassium sulfate is up to twice as expensive as potassium chloride based on K_2O content due to the costs of the raw material (potassium chloride) and processing. It is therefore only used as a potash fertilizer for applications where it performs much better than potassium chloride.

The sulfur content of potassium sulfate is an advantage where there is a deficiency of sulfur in the soil. Also, it has only a slight oversalting effect on soil in arid or semiarid areas. It is useful for fertilizing crops that are sensitive to chloride or whose quality is improved by chloride-free fertilization. This applies particularly to tobacco, vegetables, potatoes, vines, and citrus or various other fruit [179, 181].

Potassium sulfate is used in industry for the production of other potassium compounds, accelerators for rapid-setting cements, synthetic rubbers, desensitizers for explosives, lubricants, powdered fire extinguishers, dyes, explosives, and pharmaceuticals.

7. Potash – Magnesia

Crops sensitive to chloride can be fertilized with potassium sulfate alone or by fertilizers containing the sulfates of potassium and magnesium. These consist either of a mixture of the two sulfates or of the double sulfates schoenite, leonite, or langbeinite in dehydrated form.

Formerly, large quantities of a mixture of fine-grained potassium sulfate and kieserite were marketed under the name potash – magnesia (K_2O: 27 – 30 %, MgO: 9 – 12 %). This has since been almost completely replaced by a granulated potash – magnesia fertilizer, which is produced by first mixing potassium sulfate and kieserite at 95 °C with hot synthetic langbeinite. The hot mixture is granulated in a drum granulator, and the product is quickly cooled to < 60 °C. It contains 29 – 30 % K_2O and 10 % MgO.

Langbeinite from the potash deposits in Carlsbad, New Mexico (United States), can be used in the pure state directly as a potash – magnesia fertilizer. Crude langbeinite also contains halite and varying amounts of sylvite, and the method of purification utilizes the low rate of dissolution of langbeinite in water. If the sylvite content is low, the potash ore is first ground to < 6 mm and then simply washed with water to dissolve the halite. The langbeinite that remains is dried and screened to obtain the commercial size gradings. For higher sylvite content, an additional stage is required to separate the components by a combination of gravity separation and flotation. The langbeinite produced has a K_2O content of 20 – 22 % and an MgO content of 18 – 19 %, and is marketed as Sulpomag or K-Mag.

8. Production of Potassium Salts from Other Raw Materials

Potassium salts occur not only in salt deposits, but also in solution in many types of inland lakes and in seawater. Where the concentration of the salt solution is high enough and the climatic and topographical conditions are suitable, potassium chloride or potassium sulfate can be produced by solar evaporation. Typical analyses of some of the more important sources are listed in Table 10.

Table 10. Composition of some natural brines compared with seawater (in wt %)

	Dead Sea	Wendover brine	Great Salt Lake	Seawater
K^+	0.6	0.6	0.7	0.04
Na^+	2.9	9.4	7.6	1.08
Mg^{2+}	3.4	0.4	1.1	0.13
Ca^{2+}	1.3		0.016	0.04
Cl^-	17.0	16.0	14.1	1.94
SO_4^{2-}	0.04	0.2	2.0	0.27
Br^-	0.5		0.01	0.006
H_2O	74.3	73.3	74.5	96.5

8.1. The Dead Sea

Potash production by solar evaporation began at the northern end of the Dead Sea in 1931. Later, production switched to the southern end but was interrupted by the war of 1947 – 1948. In 1952, potash production was resumed in Sodom by the Israeli State *Dead Sea Works*. The capacity was built up in several stages and now amounts to $> 2 \times 10^6$ t of potassium chloride in all the usual commercial grades [182, 183].

Brine from the lake is concentrated in evaporation ponds with a total area of ca. 90 km^2, from which the crystallized salts are recovered. Most of the dissolved sodium chloride separates out

first in the primary evaporation ponds. A mixture of carnallite and sodium chloride then crystallizes in the main production ponds. This is removed as a suspension by suction dredgers and pumped to the works where the crystals are filtered off and treated by the carnallite cold decomposition process (see Section 5.2.3). Most of the resulting NaCl – KCl mixture is then treated in a hot leaching plant (see Section 5.2.2), where potassium chloride with $> 60\,\%$ K_2O is produced in crystallizers. Part of the crop of crystals from the main production ponds is treated by a cold crystallization process developed by the Dead Sea Works. This decomposition process directly yields a product containing $> 60\,\%$ K_2O.

A potash works was started by the *Arab Potash Co.* in 1982 on the southeast bank of the Dead Sea near Safi, Jordan. Evaporation and carnallite production are carried out in a pond system with a total area of $100\,km^2$. The process resembles that of the Dead Sea Works. Magnesium chloride is removed from carnallite in a cold decomposition process, and the salt mixture formed is then treated in a hot leaching plant to give potassium chloride [182, 184].

8.2. The Great Salt Lake [185]

The Great Salt Lake is the result of the shrinkage by evaporation of the former Lake Bonneville and lies in the eastern part of the basin. It has a high salt content and is the reason for the existence of several plants that produce sodium chloride and, since 1968, potassium salts.

At the western end of the Great Salt Lake, in the Great Salt Lake Desert near Wendover, are the Bonneville Salt Flats. Here, under a salt crust, are porous sediments containing brines, which are regenerated by water from atmospheric precipitation. Potassium chloride has been produced from these brines since 1937. They are collected by a system of ditches and evaporated in ponds. A mixture of potassium and sodium chloride crystallizes, from which potassium chloride is obtained by flotation [186].

Unlike the Wendover brines, the Great Salt Lake brines contain considerable amounts of sulfate (see Table 10). The process used in Wendover cannot therefore be used here. The opportunity exists of obtaining substantial quantities of valuable potassium sulfate rather than potassium chloride, which has been carried out since 1970 in the Great Salt Lake Minerals & Chemicals Corporation (GSLM & CC) plant at Ogden on the east bank. Sodium chloride is first crystallized in the 56-km^2 pond system until the solution is saturated in potassium salts. Further solar evaporation then takes place in the main production ponds, producing a mixture containing varying proportions of kainite, carnallite, and schoenite, with small amounts of sodium chloride [187]. This mixture is converted into schoenite in the plant by treatment with recycled process brine. The sodium chloride that is not dissolved by this reaction must be removed before further treatment. This is carried out by flotation of a side stream [102]. Schoenite is then decomposed by water, which produces very pure potassium sulfate. The brine from this decomposition stage has a high potassium content and is recycled to the first stage of the process. The brine produced by the reaction at this stage is recycled to the evaporation pond. The sulfate content of the main crop of crystals is higher than the potassium content, so that further potassium sulfate can be produced by addition of potassium chloride from an external source [188].

The level of the Great Salt Lake increased so much during the 1970s and 1980s that the pond system of the Great Salt Lake Minerals & Chemicals Corporation (GSLM & CC) overflowed during 1984, and production had to stop [189]. The lake level then fell, the pond system was reinstated, and production started again in 1989.

8.3. Searles Lake [190]

In Trona, on the northwest bank of the almost dry Searles Lake, brines are obtained that contain not only sodium and potassium chlo-rides, but also considerable quantities of sulfate, carbonate, and borate ions. Recycled process brines are first added, and evaporation produces sodium chloride and the double salt burkeite [*12179-88-3*], $Na_2CO_3 \cdot 3\,Na_2SO_4$. Potassium chloride is obtained by vacuum cooling of the potassium- and borate-containing mother liquor. Part of the chloride is reacted with part of the burkeite to form glaserite, $Na_2SO_4 \cdot 3\,K_2SO_4$, an intermediate stage in potassium sulfate production. An-

other reaction used for potassium sulfate production is that of potassium borate in the end brines with sulfuric acid, to form potassium sulfate and boric acid:

$$K_2B_{10}O_{16} \cdot 8\,H_2O + 6\,H_2O + H_2SO_4 \longrightarrow K_2SO_4 + 10\,H_3BO_3$$

8.4. Other Sources

In addition to the sources mentioned under potassium sulfate, other salt lakes exist whose potassium content would appear to offer the possibility of extracting potassium salts. In 1969 at Lake McLeod in Western Australia, a works produced langbeinite, $K_2SO_4 \cdot 2\,MgSO_4$, for a short period of time but ceased operations for unknown reasons. In China, in the Qinghai Province, potassium chloride production has been carried out for several years at Tsarhan Lake by using solar evaporation. Large increases in production are planned [191].

Seawater has a low potassium content (Table 10) so that economical extraction of potassium salts is not possible. The production of sea salt by solar evaporation in salt gardens yields residual brines with increased potassium content. In some places (e.g., India), small quantities of low-percentage potassium salts are produced from these bitterns. However, economical production of potash fertilizers of marketable quality is not possible from the amounts of mother liquor available from even the largest sea salt producers.

9. Storage and Transportation

(→ Fertilizers, Chap. 8.1.)

The demand for potash fertilizers fluctuates greatly throughout the year, but because potash plants need to produce at as steady a rate as possible, large storage capacities are needed to accommodate periods of low demand. Therefore, potash plants usually have high-capacity product storage facilities. Also, in seaports, where fertilizers are loaded onto ships, the largest potash companies or their subsidiaries have large storage capacities. In both cases, long storage sheds are used, usually with walls sloping to match the angle of repose of the potash salt. The sheds are usually filled by means of conveyor belts located under the shed roof. They are emptied either by bucket loaders or scrapers that move the salt into a channel under the floor of the silo or onto a conveyor belt at the side, which carries it via sloping bands or elevators to the loading plant. More recently, especially where there is a shortage of land, round silos have been used, often arranged in rows. These too are filled from above by conveyor belts and are emptied through openings at ground level. The majority of potash fertilizer is transported in bulk in self-discharging wagons with a capacity up to 100 t, by rail, truck, or inland waterway. Transport from potash plants remote from a seaport or the main consuming area is usually by special trains that run on a fixed timetable between the potash plant and the seaport or intermediate storage facility. For example, the transport of potash fertilizers from Saskatchewan to a cargo-handling plant in Vancouver uses continuous-loop train tracks that enable 10 000 t to be delivered in 102 wagons with a capacity of 98 t each [192].

A small proportion of potash salts is supplied in sacks, usually containing 50 kg. The sacks are filled either by the supplier or at the loading plant at the seaport by automatic sack-filling machines. The sacks are usually paletized.

10. Analysis of Potassium Compounds

Potassium is usually determined gravimetrically. In the United States, it is precipitated as the hexachloroplatinate [193]. Precipitation as the tetraphenylborate is another widely used method, being the standard ISO method for fertilizers, and can be either a gravimetric or a volumetric procedure (ISO 5318 and 5310) [194]. Precipitation as the perchlorate or tartrate is seldom used. Flame photometry is used for both laboratory and process control analysis. X-ray fluorescence can be used for the analysis of solids or brines.

A review of methods recommended for potassium determination in fertilizers is given in → Fertilizers.

11. Potassium Hydroxide [195–198]

11.1. Properties

Pure, solid potassium hydroxide [1310-58-3], KOH, caustic potash, M_r 56.11, ϱ 2.044 g/cm^3, mp 410 °C, bp 1327 °C, heat of fusion 7.5 kJ/mol, is a hard, white substance. It is deliquescent and absorbs water vapor and carbon dioxide from the air. Potassium hydroxide dissolves readily in alcohols and water (heat of solution 53.51 kJ/mol). The solubility of KOH (g KOH/100 g H_2O) in water is shown below:

Temperature, °C	0	10	20	30	50	100
Solubility	97	103	112	126	140	178

The mono-, di-, and tetrahydrates are formed with water. Aqueous potassium hydroxide is a colorless, strongly basic, soapy, caustic liquid, whose density depends on the concentration:

Concentration, wt %	10	20	30	40	50
Density, g/cm^3	1.092	1.188	1.291	1.395	1.514

Technical caustic potash (90 – 92 % KOH) melts at ca. 250 °C; the heat of fusion is ca. 6.7 kJ/mol.

11.2. Production

Today, potassium hydroxide is manufactured almost exclusively by potassium chloride electrolysis. The diaphragm, mercury, and membrane processes (→ Chlorine) are all suitable for the production of potassium hydroxide, but the mercury process is preferred because it yields a chemically pure 50 % potassium hydroxide solution.

In the *diaphragm process*, a KCl-containing, 8 – 10 % potassium hydroxide solution is initially formed, whose salt content can be reduced to ca. 1.0 – 1.5 % KCl by evaporation to a 50 % liquor. Further purification is complicated, and the quality of liquor from mercury cells cannot be achieved.

In the *mercury process* a very pure KCl brine must be utilized, because even traces (ppb range) of heavy metals such as chromium, tungsten, molybdenum, and vanadium, as well as small amounts (ppm range) of calcium or magnesium, lead to strong evolution of hydrogen at the amalgam cathode. The very pure potassium hydroxide solution running off the decomposers is cooled, freed from small amounts of mercury in precoated filters, and in some cases sent immediately to the consumer as a 45 – 50 % liquor in drums, tank cars, or barges.

Since about 1985, new cell rooms for the manufacture of potassium hydroxide solution have used the *membrane process*. At present, the cell liquor has a low chloride content (10 – 50 ppm); the KOH concentration is 32 %. Before dispatch, it is concentrated to 45 – 50 % by evaporation.

Nonelectrochemical processes have been proposed for the manufacture of chlorine and potassium hydroxide from KCl by thermal decomposition of potassium nitrite in the presence of Fe_2O_3 [199].

This method involves reacting KCl with NO_2 to obtain Cl_2 and potassium nitrite, reacting the KNO_2 with iron(III) oxide and oxygen to give potassium ferrate ($K_2Fe_2O_4$), and reacting the ferrate with water to produce KOH. Another method consists of reacting an aqueous solution of KCl with NO_2 and O_2 to give Cl_2 and KNO_3, which is reacted with water in the presence of Fe_2O_3 to produce KOH.

Largely water-free, ca. 90 – 95 % potassium hydroxide (caustic potash) is obtained by evaporating potassium hydroxide solution. The residual content of 5 – 10 % H_2O is bound as a monohydrate.

Suitable evaporation processes are single- or multistage falling-film evaporators [200], Badger single-tube evaporators, or boilers connected in cascade. Heating is carried out with steam or by means of heat-transfer agents (salt melts, Dow-therm). Flash evaporators are used as the final stage in large-capacity plants [201].

To counter the strong corrosiveness of the potassium hydroxide solution and retain the purity of the caustic potash, the equipment is made largely from high-purity nickel (LC 99.2) or is silverplated. The equipment is often protected by polarization.

For dispatch, caustic potash comes on the market poured directly into drums or packed in polyethylene bags after cooling; in blocks, molded pieces, flakes, prills, and as a powder. Potassium hydroxide is classified as a corrosive material:

UN no.	1814	(for aqueous solution)
UN no.	1813	(for dry material)
GGVS/GGVE	Class 8	
RID/ADR	Class 8	

Handling is described in [202].

11.3. Quality Specifications

Potassium hydroxide solution is supplied in pure quality [total alkalinity 49.7 – 50.3 %, KOH 48.8 % (min.), NaOH 0.5 % (max.), CO_3^{2-} 0.1 % (max.)] or in technical quality [total alkalinity 49.7 – 50.3 %, NaOH 1.0 % (max.), CO_3^{2-} 0.3 % (max.)]. The contents of Cl^-, SO_4^{2-}, Fe^{2+}, and Ca^{2+} are < 30 ppm. Solid caustic potash produced from amalgam liquor has a total alkalinity (calculated as KOH) of 89 – 92 %, NaOH 1.5 % (max.), CO_3^{2-} 0.5 % (max.), Cl^- 0.01 % (max.). The values for SO_4^{2-}, Fe^{2+}, and Ni^{2+} are < 50 ppm. Caustic potash from diaphragm electrolysis has a Cl^- content of 2.5 – 3.0 % and higher content of heavy metals.

For analysis, see 12.3.

11.4. Economic Aspects and Uses

Pure-quality potassium hydroxide is used as a raw material for the chemical and pharmaceutical industry, in dye synthesis, for photography as a developer alkali, and as an electrolyte in batteries and in the electrolysis of water. Technical-quality KOH is used as a raw material in the detergent and soap industry; as a starting material for inorganic and organic potassium compounds and salts (e.g., potassium carbonate, phosphates, silicate, permanganate, cyanide); for the manufacture of cosmetics, glass, and textiles; for desulfurizing crude oil; as a drying agent; and as an absorbent for carbon dioxide and nitrogen oxides from gases.

World production is estimated at ca. 700 – 800 $\times 10^3$ t/a. Main producers are the United States [203], Germany, Japan, and France. Other important producer countries are Belgium, the United Kingdom, Italy, Spain, South Korea, India, Israel, former Yugoslavia, former Czechoslovakia, Sweden, and Romania.

12. Potassium Carbonate [195–198]

Potassium carbonate was produced in antiquity and used for many purposes. In the Old Testament, potash is mentioned in Jeremiah (written in the 7th century B.C.). ARISTOTLE describes the extraction of wood ash with water; the Romans manufactured soap from fat and potash. LAVOISIER identified potash as potassium carbonate.

The production of potash from wood ash for the manufacture of glass and soap was a flourishing industry in the Middle Ages in areas having a plentiful supply of wood such as Russia, and also in Scotland. Since 1860, potash salts have replaced wood as a raw material for the manufacture of potassium carbonate.

In Anglo-American usage, the term potash today includes potassium carbonate as well as all potassium salts, such as KCl, K_2SO_4, and $K_2SO_4 \cdot MgSO_4 \cdot x\,H_2O$, that are used as fertilizers; the potassium content is given as K_2O.

Potassium carbonate occurs in small amounts in a few African lakes (e.g., Lake Chad and in the vicinity of Lake Victoria), as well as in the Dead Sea.

12.1. Properties

Anhydrous potassium carbonate [*584-08-7*], K_2CO_3, M_r 138.21, ϱ 2.428 g/cm^3, *mp* 891 °C, is a white, hygroscopic, powdery material that deliquesces in moist air. It is readily soluble in water with the formation of an alkaline solution. The solubility of K_2CO_3 (g K_2CO_3/100 g H_2O) in water is given below:

Temperature, °C	0	10	20	30	40	50
Solubility	105.5	108.0	110.5	113.7	116.9	121.2
Temperature, °C	60	70	80	90	100	
Solubility	126.8	133.1	139.8	147.5	155.7	

On addition of acid, potassium carbonate reacts with the evolution of carbon dioxide:

$$K_2CO_3 + H_2SO_4 \longrightarrow K_2SO_4 + CO_2 + H_2O$$

K_2CO_3 forms several hydrates, of which $K_2CO_3 \cdot 1.5\,H_2O$ is the stable phase in contact with the saturated solution from 0 °C to ca. 110 °C. This hydrate (M_r 165.24, ϱ 2.155 g/cm^3) crystallizes in glassy, virtually

dust-free crystals. It is also hygroscopic and deliquesces in moist air. It is completely dehydrated at 130 – 160 °C.

12.2. Production

From Caustic Potash and Carbon Dioxide. The most important process for the manufacture of potassium carbonate begins with electrolytically produced potassium hydroxide solution. The almost chemically pure solution obtained by the mercury process (see 11.2) is reacted with carbon dioxide or CO_2-containing off-gases (flue gas, lime kiln gas).

$$2\,KOH + CO_2 \longrightarrow K_2CO_3 + H_2O$$

Solid potassium carbonate is then obtained by crystallization (under vacuum and with cooling) from liquors or in the fluidized-bed process.

In the continuous crystallization process (Fig. 50), the filtered, fresh carbonate solution is mixed with mother liquor and concentrated in several preliminary evaporators connected in series until the hydrate $K_2CO_3 \cdot 1.5\,H_2O$ finally precipitates in the crystallizer after cooling under vacuum [204]. The mother liquor is separated from the crystal suspension in hydrocyclones and centrifuges, filtered, and fed back to the process. The crystals are dried at ca. 110 – 120 °C in rotary kilns or fluidized-bed dryers and packed for sale as potash hydrate, or they are calcined at 200 – 350 °C to give 98 – 100 % K_2CO_3. Impurities such as soda, sulfate, silicic acid, and iron that concentrate in the mother liquors can be partially removed [205] by removing a partial stream of the mother liquor, which is either used for brine purification in the electrolysis process or sold as a low-grade potassium carbonate solution, or by crystallizing the double-salt $NaKCO_3$ at elevated temperature in a separate crystallization and drying process.

The resulting potassium carbonate is very pure and meets the requirements of USP, BP, DAB, and JP if the process is operated in appropriate manner.

Starting from potassium carbonate solution, prills can be produced in a combined reactor, in which spray drying and fluidized-bed granulation take place simultaneously [206].

In the fluidized-bed process, aqueous potassium hydroxide solution is sprayed into a fluidized-bed reactor from above and exposed to a countercurrent of CO_2-containing hot gas (Fig. 51) [207, 208]. Carbonization and calcination take place in the same reactor. Hard, spherical potassium carbonate prills are formed having a high packing density. The prills are discharged and sieved. The coarse grains are ground and returned to the reactor together with the very fine grains, where they act as crystallization seeds. The salable, dust-free, medium grains are cooled and packed. Because no mother liquor is formed, the quality of the potassium carbonate depends on that of the raw materials. Compared to the crystallization process the chlo-ride, soda, and sulfate contents are usually higher, but the investment and production costs are lower.

Amine Process. In the Mines de Potasse d'Alsace process, potassium chloride is reacted under pressure in autoclaves with carbon dioxide in precarbonated isopropylamine solution. Potassium hydrogencarbonate precipitates and is filtered off, carefully purified of amine by intensive washing, and dried. It can be converted to potassium carbonate by calcination. Free amine, containing carbon dioxide, is recovered from the mother liquor by distillation and recycled. The chloride, predominantly present in the mother liquor as amine chlorohydrate, is reacted with hydrated lime to give free amine and an aqueous solution of calcium chloride [209].

The use of triethylamine [210], hexamethylenimine [211], or piperidine [212] is also patented. All the processes have the disadvantage that calcium chloride liquor is obtained, which can be utilized today only to a small extent and therefore represents an environmental pollutant.

Nepheline Decomposition Process [213]. In the CIS, considerable amounts of potassium carbonate are formed as a byproduct in the nepheline decomposition process for aluminum hydroxide production. The mineral nepheline is decomposed with limestone by sintering at 1300 °C:

$$(Na,K)_2 \cdot Al_2O_3 \cdot 2\,SiO_2 + 4\,CaCO_3 \xrightarrow{1300\,^\circ C} (Na,K)_2O \cdot Al_2O_3 + 2(2\,CaO \cdot SiO_2) + 4\,CO_2$$

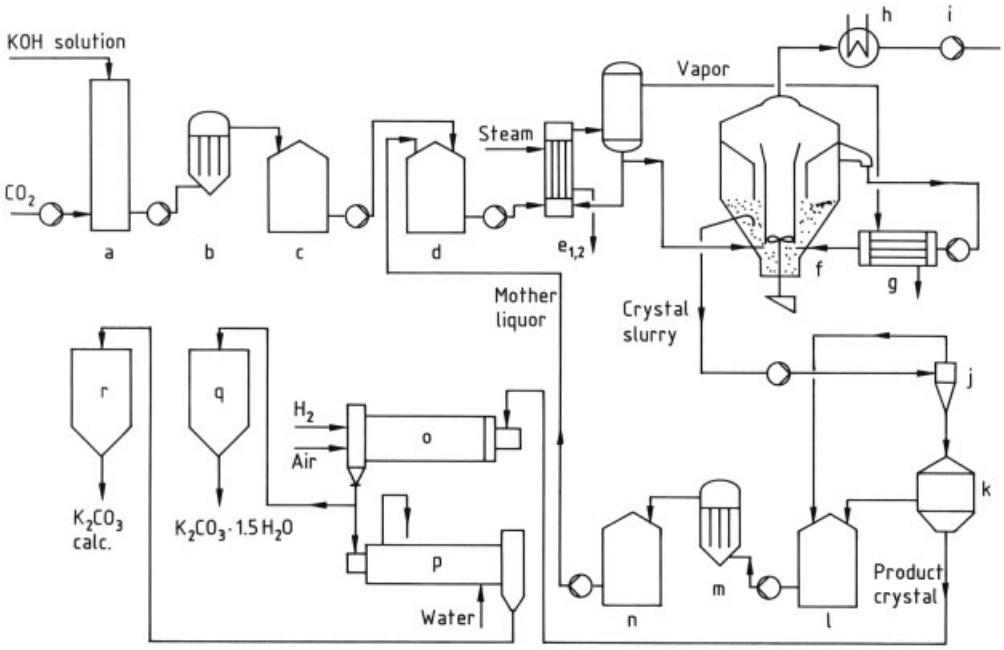

Figure 50. Preparation of potassium carbonate with continuous crystallization
a) Carbonization; b) Crude liquor filter; c) Fresh liquor tank; d) Mixed liquor tank; e_1), e_2) Preliminary evaporation; f) Vacuum/cooling crystallization (Chemietechnik Messo system); g) Preheater; h) Vapor condenser; i) Vacuum pump; j) Hydrocyclone; k) Centrifuge; l) Centrifuge liquor tank; m) Filter for mother liquor; n) Mother liquor tank; o) Drying or calcining rotary kiln; p) Cooling device for calcined K_2CO_3; q) Storage for hydrated potash; r) Storage for calcined potash

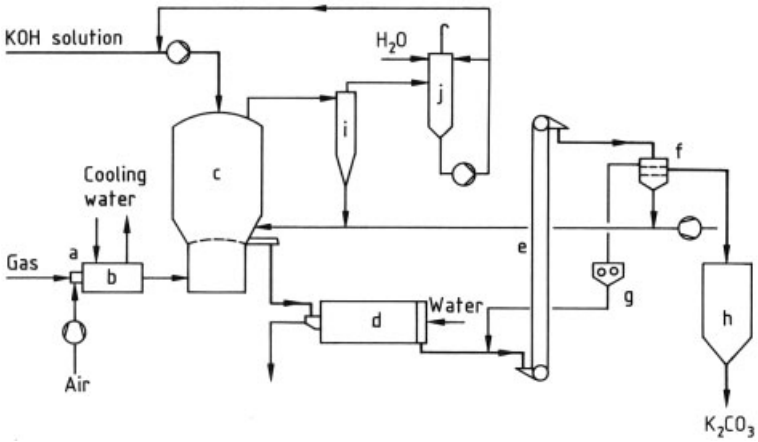

Figure 51. Production of potassium carbonate by the fluidized-bed process
a) Burner; b) Gas cooler; c) Fluidized-bed reactor; d) Cooler; e) Elevator; f) Screen; g) Mill; h) Silo; i) Cyclone; j) Exhaust gas scrubber

Alumina, portland cement, soda, and potash are obtained from the product in a complex process. The sinter product is leached with an Na_2CO_3 – NaOH solution. After filtration, a filter cake is obtained that is processed to give portland cement and an aluminate solution containing silicic acid. After precipitation of the silicic acid as alkaline aluminum silicate the purified aluminate solution is reacted with carbon dioxide:

$$2\,(Na,K)AlO_2 + CO_2 + 3\,H_2O \longrightarrow 2\,Al(OH)_3 + (Na,K)_2CO_3$$

The aluminum oxide hydrate is filtered off, and the carbonate solution is concentrated by fractional crystallization in a three-stage process and separated into sodium carbonate and $K_2CO_3 \cdot 1.5\,H_2O$.

Feldspar ($KAlSi_3O_8$) and leucite ($KAlSi_2O_6$) can also be decomposed analogously and used for alumina, cement, and potassium carbonate manufacture [214].

The magnesia process (Engel – Precht process) is of limited interest:

$$3\,(MgCO_3 \cdot 3\,H_2O) + 2\,KCl + CO_2 \longrightarrow$$
$$2\,(MgCO_3 \cdot KHCO_3 \cdot 4\,H_2O) + MgCl_2$$

In hot water the double salt ($MgCO_3 \cdot KHCO_3 \cdot 4\,H_2O$) decomposes under pressure into magnesium carbonate and dissolved potassium carbonate.

Other processes. Le Blanc process:

$$K_2SO_4 + CaCO_3 + 2\,C \longrightarrow CaS + K_2CO_3 + 2\,CO_2$$

Formate process :

$$K_2SO_4 + Ca(OH)_2 + 2\,CO \longrightarrow 2\,HCOOK + CaSO_4$$
$$K_2SO_4 + Ca(OH)_2 + 2\,CO \longrightarrow 2\,HCOOK + CaSO_4$$
$$HCOOK + KOH + 1/2\,O_2 \longrightarrow K_2CO_3 + H_2O$$

"Piesteritz" process :

$$K_2SO_4 + 2\,CaCN_2 + 2\,H_2O \longrightarrow$$
$$2\,KHCN_2 + CaSO_4 + Ca(OH)_2$$
$$2\,KHCN_2 + 5\,H_2O \longrightarrow K_2CO_3 + 4\,NH_3 + CO_2$$

These processes are uneconomical today because of high energy consumption and poor product quality, and are no longer used.

In the decomposition of chromium ores with potassium hydroxide solution, a chromate-containing potassium carbonate is obtained as byproduct. The production of potassium permanganate yields considerable amounts of potassium carbonate solution [215].

Organic raw materials, such as sunflower stalks, molasses, and suint, are used to a small extent for potash manufacture. They are ashed, leached with water, and processed to potash by fractional crystallization and calcination [216].

Ion-Exchange Process [217, 218]. An acidic ion exchanger loaded with ammonium ions is charged with KCl solution, K^+ being absorbed and an ammonium chloride solution running off. The ion exchanger is then eluted with an excess of ammonium carbonate solution (regeneration of the exchanger). The eluate, a K_2CO_3 – $(NH_4)_2CO_3$ solution, is separated by thermal cleavage to give ammonia and carbon dioxide. The ammonium chloride is reacted with magnesium hydroxide to give magnesium chloride and ammonia, which is recycled.

12.3. Quality Specifications and Analysis

Depending on the intended use, potassium carbonate is offered in varying commercial forms and degrees of purity: as granules, as powder, and as potassium carbonate hydrate ($K_2CO_3 \cdot 1.5\,H_2O$). The material manufactured from mercury-process potassium hydroxide solution is of high purity, particularly with respect to chloride content. In the amine process, the chloride content is higher and the sodium carbonate content lower, while nepheline decomposition gives high sodium carbonate contents and relatively high sulfate contents (Table 11).

Analysis. The *total alkalinity* includes $K_2CO_3 + KOH + Na_2CO_3$; it is determined with $0.5\,N\,H_2SO_4$ by potentiometric titration or with a methyl orange indicator (change to brown-red). Sodium is determined by flame photometry. The *chloride content* is determined by turbidity measurement after addition of $AgNO_3$. The *sulfate content* is determined by ion chromatography or gravimetrically after precipitation as barium sulfate. The *metal content* is determined by atomic absorption spectroscopy or photometrically by complex formation (Fe^{2+} as sulfosalicylate, Si^{4+} as the molybdato complex, Cu^{2+} as pyrrolidinodithiocarbamate, and Ni^{2+} as the diacetylglyoxime complex). Test methods for *photographic-grade* potassium carbonate, anhydrous are described in ISO 3623-1976 (E).

12.4. Storage and Transportation

Potassium carbonate is stored in bunkers; the ventilation air must be dry because of the hygroscopicity of the product.

Table 11. Analyses of calcined potassium carbonate of varying origin (data in %, remainder H_2O)

	From potassium hydroxide solution				
	Mercury process		Diaphragm	Amine	Nepheline
	Crystallization	Fluidized bed	process	process	decomposition
K_2CO_3	98.0 – 99.8	98.5 – 99.5	97 – 99	99	97.5 – 98.5
Na_2CO_3	0.1 – 0.5	0.1 – 0.5	0.5 – 1.0	0.01	0.1 – 1.0
Cl	0.001 – 0.002	0.004 – 0.013	0.2 – 1.0	0.19	0.01 – 0.03
SO_4	0.003 – 0.005	0.005 – 0.013	0.005 – 0.010		0.25 – 0.60
Si + Ca	0.005	0.006	0.010	0.009	
Fe	0.0003 – 0.0005	0.0002 – 0.0006	< 0.0005	< 0.0024	0.0007 – 0.0021

Silo vehicles and bulk containers are used for dispatch to bulk customers. Smaller amounts are packed in polyethylene valve sacks of 25 – 50 kg. The material is not hazardous; for pharmaceutical use it is classified as GRAS (generally recognized as safe) by FDA [219].

12.5. Economic Aspects and Uses

The glass industry is the most important consumer of K_2CO_3. Large amounts are also required for potassium silicate manufacture.

Potassium carbonate is used for many organic syntheses. Numerous inorganic and organic potassium salts are manufactured from potassium carbonate (potassium phosphate, bromide, iodide, dichromate, cyanide, and ferrocyanide); in addition it is a starting material for drying, neutralization, and condensation agents. As a regenerable absorbent for carbon dioxide, hydrogen sulfide, and sulfur oxides, it is attaining importance in environmental protection. Potassium carbonate is used as a fertilizer for acidic soil.

Other users are the electrical industry, the dye industry, the printing trade, the textile industry, the leather goods industry, and the ceramic industry. Soft soap manufacture has lost its earlier importance as a customer. Potash solutions are used as fire retardants and as cooling brines (freezing point − 36 °C at 576 g/L=40.5 wt % K_2CO_3).

The food industry uses potassium carbonate as a leavening agent in baked goods, as a debitterizing agent for cocoa beans, and as an additive for drying raisins. Potash in DAB quality is frequently used in the pharmaceutical industry as a raw material and auxiliary.

The most important producer countries for potassium carbonate are the CIS, France, Germany, the United States, and Japan. Other producers are Israel, Spain, India, South Korea, Belgium, Italy, former Yugoslavia, and China.

13. Potassium Hydrogencarbonate

[195–197]

13.1. Properties and Production

Potassium hydrogencarbonate [*298-14-6*], $KHCO_3$, M_r 100.12, ϱ 2.17 g/cm^3, is a white, crystalline powder that is sparingly soluble in water and insoluble in alcohol. When heated above 120 °C, it decomposes into potassium carbonate, water, and carbon dioxide.

It is manufactured industrially by passing carbon dioxide into concentrated potassium carbonate solutions or exposing these to a countercurrent of purified, cold flue gas in trickle towers (overcarbonization). Because of its low water solubility (22.4 g of $KHCO_3$ in 100 mL of H_2O at 20 °C) it precipitates in crystalline form, is separated by centrifugation, and dried at ca. 110 °C. In some potassium carbonate production processes, potassium hydrogencarbonate is obtained as a precursor (see 12.2).

The total alkalinity of the industrial material, calculated as $KHCO_3$, is at least 98 – 100 % [$KHCO_3$ 98 % (min.), Na^+ 0.1 % (max.), Cl^- 0.01 % (max.), SO_4^{2-} 0.02 % (max.); for analysis see potassium carbonate, p. 98].

13.2. Uses

Potassium hydrogencarbonate is used in the manufacture of fire extinguisher powders, in the

food industry as a leavening agent, and in the chemical and pharmaceutical industry for the manufacture of high-purity potassium carbonate and other pure potassium salts. Producing countries are the United States, Germany, and France.

14. References

General References

1. *Fertilizer Manual,* International Fertilizer Development Center, Muscle Shoals, Ala., 1979, pp. 225 – 247.
2. *Kirk-Othmer,* 3rd ed., **18,** 920 – 950.
3. *Winnacker-Küchler,* 4th ed., **2,** 268 – 333.
4. The British Sulphur Corporation, *World Survey of Potash Resources,* London 1985.
5. V. A. Zandon in N. L. Weiss (ed.): *SME Mineral Processing Handbook,* Society of Mining Engineers, New York 1985, section 22.
6. R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference,* Saskatchewan, Canada, Oct. 3 – 5, 1983, Pergamon Press, Toronto 1983.
7. W. H. Eatock: "Potash Refining in Saskatchewan," *Min. Eng. (Littleton, Colo.)* **34,** (1982) no. 9, 1350 –1353.
8. M. P. Kurtanjek: "Mining and Beneficiating Potash, Recent Developments and Trends outside North America," *Phosphorus Potassium* **128** (1983) Nov./Dec., 26 – 31.
9. *Gmelin,* **System no. 22.**
10. F. Serowy: *Verarbeitungsmethoden der Kalirohsalze,* W. Knapp, Halle/Saale 1952.
11. H. Schubert: *Aufbereitung fester mineralischer Rohstoffe,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig, vol. I, 1968; vol. II, 1986; vol. III, 1984.
12. D. Fulda et al.: *Kali, das bunte, bittere Salz,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1990.
13. A. Heinz, R. v. d. Osten (eds.): *ABC Kali und Steinsalz,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1982.
14. I. Barin: *Thermochemical Data of Pure Substances,* VCH, Weinheim 1989.
15. M. Broul, J. Nývlt, O. Söhnel: *Solubility in Inorganic Two-component Systems,* Elsevier, Amsterdam 1981.
16. *Ullmann,* 4th ed., **13,** 447 – 496.

Specific References

17. R. Slotta: "Die Kali- und Steinsalzindustrie," *Technische Denkmäler in der Bundesrepublik Deutschland,* vol. 3, Deutsches Bergbaumuseum, Bochum 1980.
18. D. Hoffmann: *Elf Jahrzehnte Deutscher Kalisalzbergbau,* Verlag Glückauf, Essen 1972.
19. Kaliverein e.V.: *Die Kaliindustrie in der Bundesrepublik Deutschland,* 6th ed.,1988.
20. *Phosphorus Potassium* **166** (1990) March/April, 17 –19.
21. R. Weissenberger: *Chronique des mines de potasse d'Alsace,* Ziegler, Bergholtz 1985.
22. *Phosphorus Potassium* **168** (1990) July/Aug., 12 – 13.
23. J. W. Turrentine: *Potash in North America,* Reinhold Publ., New York 1943.
24. *Phosphorus Potassium* **68** (1973) Nov./Dec., 39 – 43.
25. O. Braitsch: *Entstehung und Stoffbestand der Salzlagerstätten,* Springer Verlag, Berlin 1962.
26. *Ullmann,* 4th ed., **13,**450.
27. M. A. Zharkov: *Paleozoic Salt Bearing Formations of the World,* Springer Verlag, Berlin 1984.
28. E. A. Vysotsky, V. Z. Kislik: "Epochs of Bischofite Deposition in Geological History," *Int. Geol. Rev.* **29** (1987) no. 2, 134 – 139.
29. R. Kühn, *Kali Steinsalz* **1** (1955) no. 9, 3 – 16.
30. R. Kühn, *Geol. Jahrb.* **90** (1972) 127 – 220.
31. H. Borchert: *Ozeane Salzlagerstätten,* Bornträger, Berlin 1959.
32. F. Lotze: "Steinsalz und Kalisalze," *Die wichtigsten Lagerstätten der "Nicht-Erze",* **vol. III, part 1,** Bornträger, Berlin 1938.
33. F. Lotze: *Steinsalz und Kalisalze,* 2nd ed., **part 1,** Bornträger, Berlin 1957.
34. F. Lotze: *Die Salzlagerstätten in Zeit und Raum,* Arbeitsgem. Forsch. Landes Nordrhein-Westfalen, no. 195, Westdeutscher Verlag, Köln-Opladen 1969.
35. G. Richter-Bernburg: "Salzlagerstätten," in Bentz-Martini (ed.): *Lehrbuch der Angewandten Geologie,* vol. 2, part 1, Enke Verlag, Stuttgart 1968, pp. 916 –1061.
36. G. Richter-Bernburg, *Z. Dtsch. Geol. Ges.* **105** (1953) 593 – 645.
37. M. Brongersma, *Mar. Geol.* **11** (1972) 123 – 144.
38. C. Kippenberger et al.: *Untersuchungen über Angebot und Nachfrage mineralischer Rohstoffe, XX Kali,* BGR Hannover-DIW Berlin, Schweizerbart, Stuttgart 1986.

39. J. D'Ans, *Naturwissenschaften* **34** (1947) 295 – 301.
40. J. D'Ans, R. Kühn, *Kali Steinsalz* **3** (1960) 69 – 84.
41. H. Mayrhofer: "World Reserves of Mineable Potash Salts Based on Structural Analysis," *Proceedings of the 6th International Symposium on Salt,* **vol. 1,** The Salt Institute, Alexandria, USA, 1983.
42. P. A. Ziegler, *Geological Atlas of Western and Central Europe,* Elsevier, Amsterdam 1982.
43. E. Messer, *Kali Steinsalz* **5** (1970) 244 – 251.
44. Pittsburgh Plate Glass Co., US 3 058 729, 1962 (J. B. Dahms, B. P. Edmonds); CA 627 308, 1963 (J. B. Dahms); US 4 007 964, 1977 (E. L. Goldsmith); US 3 262 741, 1966 (B. P. Edmonds, J. B. Dahms); US 3 433 530, 1969 (J. B. Dahms, B. P. Edmonds); US 4 329 287, 1980 (E. L. Goldsmith).
45. *Phosphorus Potassium* **138** (1985) July/Aug., 32 – 33.
46. D. Jackson, *Eng. Min. J.* **174** (1973) no. 7, 59 – 68.
47. *Phosphorus Potassium* **168** (1990) July/Aug., 23 – 28.
48. G. Duchrow, I. Fitz, N. Grüschow, *Phosphorus Potassium* **167** (1990) May/June, 26 – 32.
49. H. Schubert: *Kali, das bunte, bittere Salz,* **1**VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 199072 – 84.
50. R. Kühn, *Kali Steinsalz* **5** (1970) 307 – 317.
51. R. Kühn, *Erzmetall* **8** (1955) Suppl. B 93 – B 107, B 115.
52. H. Autenrieth, O. Braun, W. Otto: *Winnacker-Küchler,* 4th ed., **2,** 281 – 283.
53. V. A. Zandon in N. L. Weiss (ed.): *SME Mineral Processing Handbook,* Society of Mining Engineers, New York 1985,22-4 – 22-5.
54. J. Götte, *Kali Steinsalz* **10** (1990) 261 – 264.
55. J. H. van't Hoff: *Untersuchungen über die Bildungsverhältnisse der ozeanischen Salzablagerungen,* Akad. Verlagsges., Leipzig 1912.
56. H. D'Ans: *Die Lösungsgleichgewichte der Systeme der Salze ozeanischer Salzablagerungen,* Verlagsges. für Ackerbau, Berlin 1933.
57. J. D'Ans, *Z. Elektrochem.* **56** (1952) 497 – 505.
58. *Gmelin,* System no. 22.
59. A. B. Sdanowsky, E. I. Lyakhowskaya, R. E. Schleymowitch: *Handbuch der Löslichkeit der Salzsysteme,* Gaskhimisdat, Leningrad, vol. I, 1953; vol. II, 1954; vol. III, 1961; vol. IV, 1963 (Russ.).
60. H. Autenrieth, *Kali Steinsalz* **1** (1953) no. 2, 3 – 17.
61. H. Autenrieth, *Kali Steinsalz* **1** (1954) no. 7, 3 – 22.
62. H. Autenrieth, *Kali Steinsalz* **1** (1955) no. 11, 18 – 32.
63. H. Autenrieth, *Kali Steinsalz* **2** (1958) no. 6, 181 –200.
64. H. Autenrieth, *Kali Steinsalz* **5** (1969) no. 5, 158 – 165.
65. H. Autenrieth, *Rev. chim. minér.* **7** (1970) 217 – 229.
66. H. Autenrieth, *Kali Steinsalz* **5** (1970) no. 9, 289 – 306.
67. G. Peuschel, *Kali Steinsalz* **9** (1986) no. 9, 296 – 303.
68. W. P. Wilson, A. G. McKee: *Proceedings of the 4th Symposium on Salt,* **vol. I,** The Northern Ohio Geological Society, Cleveland, Ohio, 1974, 517 – 525.
69. W. P. Wilson: *Proceedings of the 3rd Symposium on Salt,* **vol. II,** The Northern Ohio Geological Society, Cleveland, Ohio, 1969, 20 – 29.
70. J. H. Wolf: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto1983, 711 – 716.
71. H. Domning, *Kali Steinsalz* **7** (1977) no. 4, 155 – 160.
72. W. H. Coghill, J. R. De Vaney, J. B. Clemmer, S. R. B. Cooke, Report of Investigations of the US-Bureau of Mines, Report no. 3271, 1935.
73. A. S. Kusin, *Kalii* **6** (1937) 17 – 27.
74. R. A. Pierce, L. D. Anderson, *Eng. Min. J.* **143** (1942) 38 – 41.
75. Kreller, *Kali, Verw. Salze Erdöl* **33** (1939) 35 – 37, 46 – 47, 53 – 57.
76. E. Rüsberg, *Chem.-Ing.-Tech.* **27** (1955) 1 – 4.
77. O. Karsten, in W. Gründer: *Erzaufbereitungsanlagen in Westdeutschland,* Springer Verlag, Berlin 1955, 343 – 345.
78. V. A. Arsentiev, J. Leja, *CIM Bull.* **3** (1977) 154 – 158.
79. Du Pont, US 2 088 325, 1937 (J. E. Kirby).
80. C. M. Aleksandrovic, *Freiberg. Forschungsh.* **A 544** (1975) 73 – 81.
81. H. Schubert, *Aufbereit. Tech.* **7** (1967) 365 – 368.
82. A. Singewald, *Chem.-Ing.-Tech.* **33** (1961) 376 – 393, 558 – 572, 676 – 688.
83. D. W. u. M. C. Fuerstenau, *Min. Eng. (Littleton, Colo.)* **8** (1956) 302 – 307.

84. A. F. Taggart, *Elements of Ore Dressing,* John Wiley & Sons, New York 1951.
85. A. M. Goudin, Testimony in Transcript of Evidence, Civil Action no. 1829, District of New Mexico, p. 1255.
86. R. Bachmann, *Erzmetall* **8** (1955) Suppl. B 109 – B 116.
87. J. Rogers, J. H. Schulman: *Second International Congress of Surface Activity,* **vol. II,** Reprints, Butterworths, London 1957, 330 – 338.
88. A. Singewald, *Erzmetall* **12** (1959) 121 – 135.
89. O. J. Somojlov: *Struktur von wäßrigen Elektrolytlösungen,* B. G. Teubner, Leipzig 1961.
90. H. Schubert, *Aufbereit. Tech.* **6** (1966) 305 – 313.
91. F. Hagedorn, *Kali Steinsalz* **10** (1991) 315 – 328.
92. D. Uhlig, *Neue Bergbautech.* **5** (1975) 145 – 155.
93. H. Köhler et al., *Neue Bergbautech.* **16** (1986) 45 – 50.
94. N. F. Mescerjakov, Y. W. Rjabov, V. N. Kuznetzov, *Freiberg. Forschungsh.* **A 594** (1978) 33 – 54.
95. A. Bahr, K. Legner, H. Lüdke, F. W. Mehrhoff, *Aufbereit. Tech.* **1** (1957) 1 – 9.
96. N. F. Mescerjakov: *Flotacionnye maschiny,* Isdatel'stwo Nedra, Moskau 1972.
97. H. Schubert, *Neue Bergbautech.* **4** (1974) 223 – 228.
98. *Phosphorus Potassium* **145** (1986) 29 – 33.
99. Kali & Salz, DE 3 435 124, 1987 (F. Hagedorn, G. Peuschel, A. Singewald).
100. F. Hagedorn, *Kali Steinsalz* **9** (1986) 232 – 238.
101. H. Köhler, W. Kramer, *Neue Bergbautech.* **11** (1981) 362 – 366.
102. R. B. Tippin, *Chem. Eng. (N.Y.)* **184** (1977) no. 15, part 1, 73 – 75.
103. VEB Kali, DD 220 237, (L. Herrmann et al.).
104. S. Mildner, R. Ecke, DD 35 637, 1965.
105. I. M. Le Baron, W. C. Knopf, *Min. Eng. (Littleton, Colo.)* **10** (1958) 1081 – 1083.
106. H. Autenrieth, *Kali Steinsalz* **5** (1969) 171 – 177.
107. Kali-Forschungsanstalt, DE 1 056 551, 1957 (H. Autenrieth).
108. G. Fricke, *Kali Steinsalz* **9** (1986) 287 – 295.
109. D. Larmour in: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983597 – 602.
110. R. Bock, *Chem.-Ing.-Tech.* **53** (1981) 916 – 924.
111. Kali & Salz, DE 1 249 783, 1966 (A. Singewald, G. Fricke).
112. A. Singewald, U. Neitzel: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983* Pergamon Press, Toronto 1983,589 – 595.
113. L. Ernst, *Kali Steinsalz* **9** (1986) 275 – 286.
114. L. Ernst, *Ber. Bunsenges. Phys. Chem.* **93** (1989) 857 – 863.
115. L. Ernst, *Ber. Bunsenges. Phys. Chem.* **94** (1990) 1435 – 1439.
116. Kali & Salz, DE 3 603 166, 1986 (G. Fricke, I. Giesler, R. Diekmann).
117. Kali & Salz, DE 3 603 165, 1986 (H. Balzer, H. Burghardt, F. Maikranz).
118. Kali & Salz, DE 3 603 167, 1986 (U. Neitzel, G. Fricke).
119. Kali & Salz, DE 1 077 611, 1959 (H. Autenrieth, G. Peuschel).
120. Kali & Salz, DE 1 142 802, 1961 (H. Autenrieth, G. Peuschel, G. Weichart).
121. Kali & Salz, DE 2 007 677, 1970 (A. Singewald, G. Fricke, D. Jung).
122. Kali & Salz, DE 2 052 993, 1970 (A. Singewald, G. Fricke, D. Jung).
123. Kali & Salz, DE 1 154 052, 1960 (H. Autenrieth, H. Dust).
124. Kali & Salz, DE 1 076 593, 1957 (H. Autenrieth).
125. Kali & Salz, DE 1 261 453, 1967 (A. Singewald, G. Fricke).
126. Kali & Salz, DE 1 667 814, 1968 (G. Fricke, A. Singewald).
127. Kali & Salz, DE 1 283 772, 1967 (H. Autenrieth, H. Wirries).
128. Kali & Salz, DE 1 792 120, 1968 (A. Singewald, G. Fricke).
129. Kali & Salz, DE 1 953 534, 1969 (A. Singewald, G. Fricke).
130. W. B. Dancy: *Kirk-Othmer,* 3rd ed., **18,** 931 – 933.
131. T. E. Burus, B. J. Clarke, W. B. Coome, A. H. Newcombe: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983541 – 546.
132. R. Diekmann: Lecture held at *Int. Potash Technol. Conf., 2nd,* Hamburg, May 26 – 29, 1991.
133. E. Weps, *Kali Steinsalz* **8** (1981) 177 – 183.

134. Kali & Salz, DE 3 434 190, 1984 (O. Pfoh, C. Radick, H. Thenert).
135. F. Hagedorn, *Kali Steinsalz* **7** (1977) 161 – 164.
136. T. Fleischer, *Kali Steinsalz* **9** (1986) 304 – 313.
137. H. J. Scharf: "Environmental Aspects of K-Fertilizers in Production, Handling and Application," *Development of K-Fertilizer Recommendations,* International Potash Institute, Worblaufen-Bern 1990, 395 – 402.
138. *Phosphorus Potassium* **148** (1987) March/April, 30 –35.
139. M. D. Haug, K. W. Reid: Lecture held at *Int. Potash Technol. Conf., 2nd,* Hamburg, May 26 – 29, 1991.
140. J. E. Tallin, D. E. Pufahl: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983755 – 760.
141. K. W. Reid, G. A. Maki: Lecture held at *Int. Potash Technol. Conf., 2nd.,* Hamburg, May 26 – 29, 1991.
142. H. E. Schroth, *Phosphorus Potassium* **67** (1973) Sept./Oct., 38.
143. A. Singewald, *Die Weser* **57** (1983) no. 516, 1 – 8.
144. N. Knöpfel: Lecture held at *Int. Potash Technol. Conf., 2nd.,* Hamburg, May 26 – 29, 1991.
145. H. Stahl, *Aufbereit. Tech.* **20** (1980) 525 – 533.
146. W. B. Pietsch: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983, 661 – 669 .
147. L. Medemblik: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto1983, 653 – 659.
148. A. S. Middleton, D. A. Cormode, J. E. Scotten: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983, 647 – 651.
149. *Phosphorus Potassium* **173** (1991) May/June 28 – 36.
150. K. Kahle, G. Leib: Lecture held at *Int. Potash Technol. Conf., 2nd.,* Hamburg, May 26 – 29, 1991.
151. L. I. Skvirski, A. A. Chityakov, Z. L. Kozel: Lecture held at *Int. Potash Technol. Conf., 2nd.,* Hamburg, May 26 – 29, 1991.
152. H. Rieschel, K. Zech, *Phosphorus Potassium* **115** (1981) Sept./Oct., 33 – 39.
153. H. Rug, K. Kahle, *Phosphorus Potassium* **170** (1990) Nov./Dec., 23 – 27.
154. International Fertilizer Industry Association (IFA): *Potash Statistics 1989,* Paris.
155. Prognose-Arbeitsgruppe Weltbank/FAO/Unido, 1991, unpublished.
156. C. Childers, *Phosphorus Potassium* **169** (1990) Sept./Oct., 16 – 20.
157. O. Walterspiel, *Kali Steinsalz* **10** (1989) 168 – 174.
158. *Phosphorus Potassium* **165** (1990) Jan./Febr., 18 –19.
159. *Gmelin,* System no. 22, Suppl. vol., pp. 1280 – 1338.
160. J. Näther, H. H. Emons, *Bergakademie* **21** (1969) 310 – 313.
161. *Phosphorus Potassium* **156** (1988) July/Aug., 27 – 34.
162. *Phosphorus Potassium* **122** (1982) Nov./Dec., 36 – 39.
163. N. P. Finkelstein, S. H. Garnett, L. Kogan: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto 1983,, 571 –576.
164. U. Neitzel, *Kali Steinsalz* **9** (1986) 257 – 261.
165. H. Autenrieth, O. Braun, W. Otto: *Winnacker-Küchler,* 4th ed., **2,** 320 – 322.
166. H. Scherzberg, G. Döring: Lecture held at *Int. Potash Technol. Conf., 2nd.,* Hamburg, May 26 – 29, 1991.
167. Kali & Salz, DE 3 418 147, 1984 (E. Menche, H. G. Diehl, H. Eberle).
168. W. B. Dancy: *Kirk-Othmer,* 3rd ed., **18,** 945.
169. D. K. Storer: R. M. McKercher (ed.): *"Potash Technology," 1st International Potash Technology Conference, Saskatchewan, Canada, Oct. 3 – 5, 1983,* Pergamon Press, Toronto, 1983,577 – 582.
170. R. Phinney, EP 0 199 104, 1986.
171. Kali & Salz, DE 3 607 641, 1986 (S. Vajna, G. Peuschel).
172. Société d'Études Chimiques pour L'Industrie et l'Agriculture (SECPIA), DE 956 304, 1954 (J. Lafont).
173. J. A. Fernandez Lozano, A. Wint, *Chem. Eng. J. (Lausanne)* **23** (1982) 53 – 61.
174. *Phosphorus Potassium* **167** (1990) May/June, 11.
175. Superfos A/S, Vedbaek, DK, DE 3 331 416, 1983 (K. C. B. Knudsen); US 4 504 458, 1983 (K. C. B. Knudsen).

176. *Chem. Eng. (N.Y.)* **81** (1974) 98 – 99.
177. Kali & Salz, DE 2 810 640, 1978 (N. Knöpfel, F. Wartenpfuhl, A. Hollstein).
178. A. Hollstein, *Kali Steinsalz* **7** (1979) 498 – 500.
179. *Phosphorus Potassium* **141** (1986) Jan./Feb., 17 – 21.
180. *Phosphorus Potassium* **151** (1987) Sept./Oct. , 16 – 21.
181. G. Kemmler, *Kali Steinsalz* **9** (1985) 167 – 169.
182. The British Sulphur Corporation, *World Survey of Potash Resources,* London 1985 pp. 62 – 64.
183. *Phosphorus Potassium* **24** (1966) 40 – 44.
184. A. M. Amarin, K. Manasrah, *Proc. IFA-NFC Joint Middle East-South Asia Fertilizer Conference,* Lahore, Pakistan, Dec. 3 – 6, 1988.
185. *Kirk-Othmer,* 2nd ed., Suppl. vol., 438 – 467.
186. The British Sulphur Corporation, *World Survey of Potash Resources,* London 1985., pp. 38 – 39.
187. U. Neitzel, *Kali Steinsalz* **5** (1971) 327 – 334.
188. P. Behrens, Industrial Processing of Great Salt Lake Brines, Utah Geological and Mineral Survey Bulletin 116, 1980, 223 – 228.
189. *Phosphorus Potassium* **132** (1984) July/Aug., 6.
190. The British Sulphur Corporation, *World Survey of Potash Resources,* London 1985 , pp. 42 – 43.
191. The British Sulphur Corporation, *World Survey of Potash Resources,* London 1985 , p. 70.
192. *Phosphorus Potassium* **173** (1991) May/June, 26 – 27.
193. W. Horwitz (ed.): *Official Methods of Analysis of the AOAC,* 11th ed., Association of Official Analytical Chemists (AOAC), Washington D.C. 1970.
194. Verband Deutscher Landwirtschaftlicher Untersuchungs- und Forschungsanstalten (LUFA): "Die Untersuchung von Düngemitteln," *Methodenbuch,* **vol. II,** Verlag J. Neumann-Neudamm, Melsungen 1972, method 4.1.

General References
195. *Kirk-Othmer,* 3rd ed., **18,** 936 – 939.
196. J. Ford: "Caustic potash," *Encycl. Chem. Process. Des.* **7** (1978) 22 – 34.
197. *Ullmann,* 4th ed., **13,** 489 – 496.
198. Hüls, Handbook KOH-, K2CO3-, KHCO3-Products, Marl 1992.

Specific References
199. N. Takeuchi, *Soda to Enso* **39** (1988) no. 461, 277 –290.
200. Bertrams, Concentration Plants for NaOH-, KOH-, Na2S- and CaCl2-liquors, Muttenz, Switzerland, 1979.
201. Sulzer-Escher-Wyss, US 4 927 494, 1990 (R. Winkler et al.).
202. Oxy-Occidental Chem. Corp., Caustic Potash Handbook, Irving 1987.
203. Oil Paint Drugs, Chemical Marketing Reporter, 28th May, 1990.
204. Messo Chemietechnik, Brochure, Mass Crystallization, Duisburg, Germany, 1990.
205. Mannesmann, DE 3 816 061, 1989 (R. Schmitz).
206. VEB Kombinat Kali, DD 255 328 A, 1986 (K. Will, G. Elberling).
207. Bertrams, Fluid Bed Processes, Muttenz, Switzerland, 1979.
208. Diamond Shamrock Corp., Company brochure, Cleveland, Ohio, 1969.
209. Inf. Chim. **99** (1971) Aug./Sept., 125.
210. Kali-Chemie, DT 1 220 401, 1962 (P. Schmid).
211. J. N. Shokin et al., *Khim. Promst. (Moscow)* **9** (1978) 685.
212. FMC Corp., BE 616 193, 1962 (A. B. Gency, M. J. McCarthy).
213. D. M. Ginzburg, A. A. Tripolskii, *Tr. Gos. Nauchno-Issled. Proektn. Inst. Osnon. Khim.* **30** (1973) 26.
214. IMC Corp., US 3 073 443, 1960 (R. E. Snow).
215. A. Schmidt, *Angewandte Elektrochemie,* Verlag Chemie, Weinheim 1976, p. 183.
216. Lemar Developments, AU 563 487, 1987 (B. W. Levy).
217. *Chem. Age Int.,* 29th Sept., 1972.
218. Dynamit Nobel, DT 1 812 769, 1968 (D. Labriola et al.).
219. FDA, *Fed. Regist.* **48** (1983) no. 224, 52 440-3.

Potassium Nitrate → Nitrates and Nitrites
Pottery → Whitewares
Powder Coatings → Paints and Coatings

Powder Metallurgy and Sintered Materials

Sintered Steel and Iron is a separate keyword.

WOLFGANG A. KAYSSER, German Aerospace Research Establishment, Köln, Federal Republic of Germany (Chaps. 1 – 8, 9.1 – 9.6, 10)

WOLFGANG WEISE, Degussa AG, Hanau, Federal Republic of Germany (Section 9.7)

1. Introduction 2
1.1. Production Routes 2
1.2. Product Groups 2
1.3. History and Developments 6
1.4. Economic Aspects 6
2. Powder Production 6
2.1. Mechanical Methods 6
2.1.1. Mechanical Disintegration of Melts . 7
2.1.2. Mechanical Disintegration of Solids . 8
2.2. Physicochemical Methods 8
2.3. Chemical Methods 9
3. Powder Properties and Characterization 10
4. Powder Conditioning 11
5. Pressing and Forming at Ambient Temperatures 12
5.1. Uniaxial Pressing 13
5.2. Isostatic Pressing, Cold Rolling, and Extrusion 13
5.3. Pouring, Preforming, and Injection Molding 15
6. Sintering 15
6.1. The Sintering Process 15
6.1.1. Solid-State Sintering 15
6.1.2. Liquid-Phase Sintering 18
6.2. Sintering Furnaces and Atmospheres 19
7. Pressing and Shaping at Elevated Temperature 20
7.1. Hot Pressing and Hot Isostatic Pressing 21
7.2. Sinter Forging and Other Methods 21
8. Structural Parts 22
8.1. Structural Ferrous Parts 22
8.1.1. Low-Alloyed Steels 23
8.1.2. High-Alloyed Steels 23
8.1.3. Secondary Operations 24
8.2. Copper and Copper Alloys 25
8.3. Aluminum and Aluminum Alloys . 25
9. High-Performance Materials 25
9.1. High-Melting Metals and Alloys . . 25
9.2. Light Alloys 26
9.2.1. High-Performance Aluminum Alloys 26
9.2.2. Titanium and Titanium Alloys 28
9.2.3. Magnesium and Beryllium 28
9.3. High-Temperature Nickel and Cobalt Superalloys 28
9.4. Magnetic Materials 29
9.5. Hardmetals and Cermets 31
9.6. Metal – Ceramic Composites 31
9.7. Electric Contact Materials 32
9.7.1. Contact Materials for Low-Voltage Switchgear 33
9.7.2. Contact Materials for High-Voltage Switchgear 34
10. Parts with Inherent, Functional Porosity 35
11. References 35

1. Introduction

1.1. Production Routes

Powder metallurgy (PM) is a technique used to transform metal powders into consolidated material. Production methods include the pro-cessing steps [1–5]: powder preparation (Chap. 2), powder conditioning (Chap. 4), shaping, consolidation (Chaps. 567), and secondary operations (Fig. 1). The final products may or may not contain pores.

Powder production methods include chemical reactions from precursors or disintegration of bulk solids or liquids. The main *powder conditioning steps* are sieving, blending, and the addition of binders or pressing aids.

Shaping and hot consolidation can be realized in two consecutive steps or in a single shaping – hot consolidation step. In *two-step shaping – consolidation* the powders are shaped at ambient temperature by (almost) pressureless methods into preforms or by compacting in dies under an uniaxial load or in thin-walled capsules

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_105

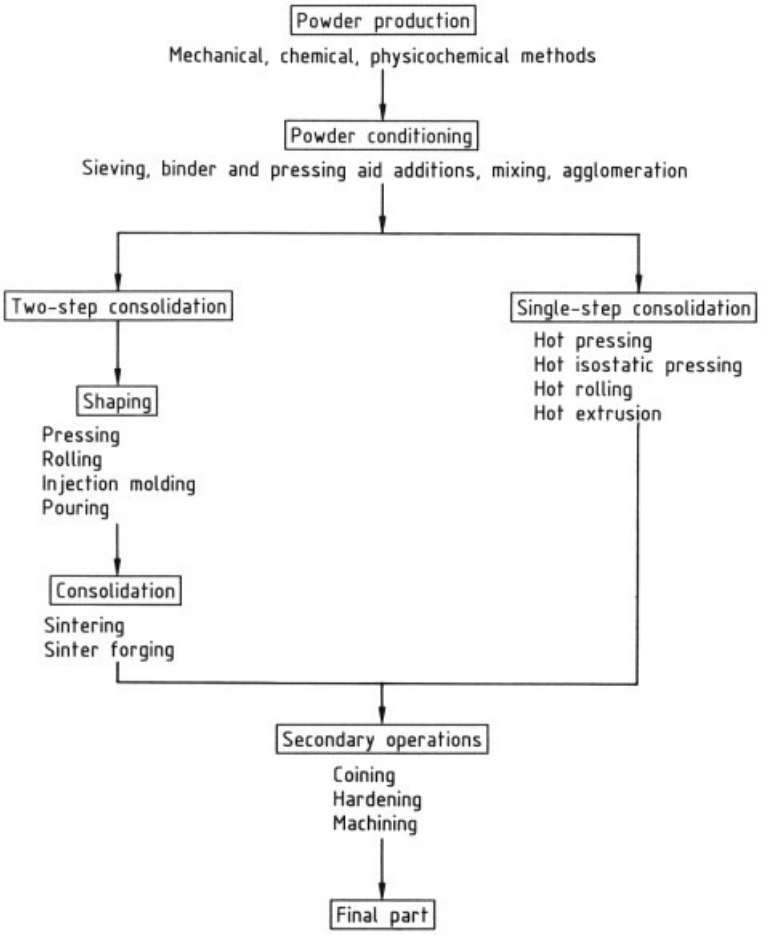

Figure 1. Basic steps of PM production

under isostatic pressure. These shaping methods are supplemented by indirect shaping methods (e.g., machining of loosely-packed and slightly sintered bulks).

Parts shaped at room temperature contain pores between the individual particles (usually porosity 10 – 60 vol %). The purpose of subsequent sintering (heat treatment below the melting temperature of the major constituents of the parts) is to initiate and enlarge the solid contacts between the particles. The driving force for contact growth is the reduction of the sum of the surface and interfacial energies. Depending on the geometric arrangement of the particles and the active transport mechanism, the same driving forces may lead to shrinkage or elimination of the pores. Low- or zero-porosity parts are also obtained by double sintering in which two pressing and sintering cycles are used.

Single-step shaping – hot consolidation methods simultaneously apply elevated temperature, external compacting, and/or forming pressures to obtain high or full density. Typical direct consolidation methods are hot pressing, hot isostatic pressing, extrusion, and rolling. Products of the two-step consolidation route (sintered parts) are sometimes used as preforms for direct consolidation methods such as hot isostatic pressing and in particular sinter forging.

Some of the *secondary operations* are common to non-PM parts (e.g., hardening by heat treatment, joining, and surface modification).

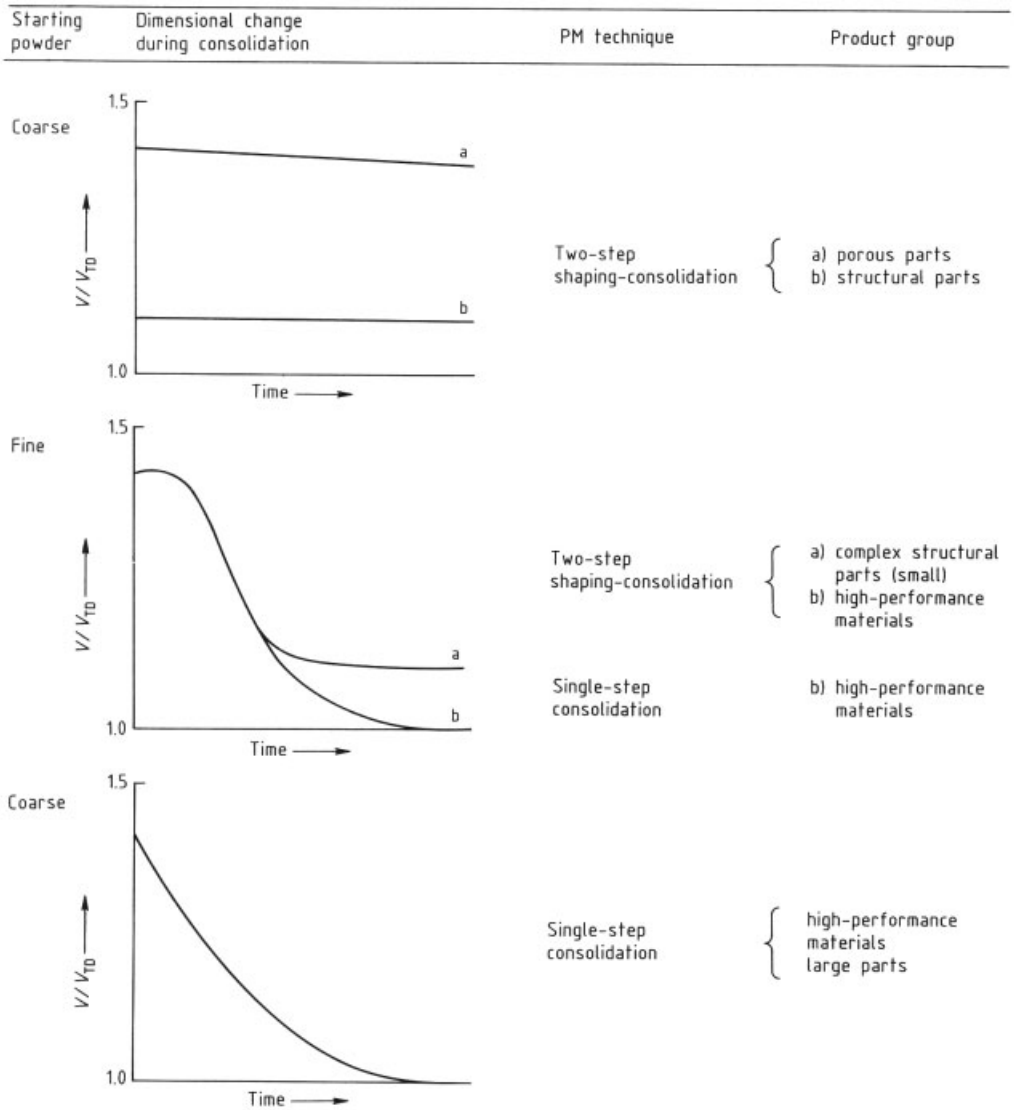

Figure 2. Initial powder size and dimensional change during sintering for different types of products
V/V_{TD} = relative volume; V_{TD} = theoretical volume

Coining operations and pore filling by steam reaction treatments or by infiltration with metals or polymers are unique to sintered porous parts.

1.2. Product Groups

PM products are classified as follows:

1) Structural parts (Chap. 8)
2) High-performance materials (Chap. 9)
3) Parts with inherent functional porosity (Chap. 10)

These products require different initial powders and different dimensional changes during consolidation (Fig. 2).

Structural Parts. Competitive structural PM parts with an accurate shape and size are cheaper than equivalent parts produced by machining cast and wrought bulks (Fig. 3) [6], [7].

Major arguments for the production of structural parts by PM are exact size and shape, exact weight, negligible machining, high material efficiency, and low energy consumption.

Figure 3. Structural PM parts for the automotive industry

The main emphasis is put on the control of dimensional changes occurring during sintering to avoid subsequent extensive machining of the sintered parts. The dimensions are well controlled if shrinkage or swelling during pressureless sintering are kept small; as a result the porosities after cold compaction and sintering should be approximately equal. The simultaneous requirements for minimal dimensional changes and a high final density (required for good mechanical properties) at low powder prices are satisfied by the use of coarse, unalloyed powders ($> 100\,\mu m$) which are cold compacted into compacts with density and shape similar to those of the final parts. Mixtures of powders are usually used to obtain alloys.

Formation of chemical bonds between the particles and homogenization of the alloying components during sintering yields reasonable mechanical properties as a result of solid-solution hardening, precipitation hardening, or the formation of martensite. In most cases the final material properties of PM parts are slightly inferior to the properties of parts machined from cast and wrought bulks due to a residual porosity of 5 – 15 %.

Metal-injection molding (MIM) is used for production of small structural PM parts with complex shapes. A blend of binder (e.g., waxes) and fine powder is injection molded into dies. After ejection and binder removal the porosity of the parts has to be reduced to a few percent by sintering. Extensive shrinkage occurs, requiring fine powders and limiting the dimensional accuracy of the final parts [8], [9]. The argument for the use of MIM is the lower price compared to machining or precision casting.

Figure 4. Hot isostatically-pressed stainless steel valve component for off-shore oil exploitation

High-Performance Materials. PM can be used to produce parts or semifinished parts made of high-performance materials. Major technical arguments for using PM instead of other methods are based on material properties such as high melting point, insufficient machinability, and microstructures that cannot be obtained via the casting route (phase combination and distribution).

Fine-grained, segregation-free material is obtained as well as materials and phase combinations that cannot be prepared by foundry procedures. Examples are refractory metals (W, Mo) [10], hard metals (WC – Co) [11], heavy-metal alloys (W – Ni – X) [12], hard magnets (Co – Sm, Fe – Nd – B), and many "high-tech" ceramics [13]. In the two-step shaping – consolidation route, these high-performance materials are sintered to complete density, usually at a small grain size. This production route generally requires fine starting powders and powder mixtures that accelerate

shrinkage (e.g., by forming a limited amount of liquid phase [5]).

High-performance materials are processed by the single-step shaping – hot consolidation method if they are produced from powders or powder mixtures that have an unsufficient sinterability (e.g., metal – matrix composites) or require low processing temperatures (e.g., rapidly solidified atomized powders) if the metastable microstructural features are to be maintained [14].

Very large parts with microstructures of an extremely long-range homogeneity (Fig. 4) can be obtained by hot isostatic pressing of coarse powders to near net shapes; material waste due to machining is minimized [15]. The argument that PM single-step shaping – consolidation is used to obtain superior high-performance material properties can be replaced by an economic argument which justifies the use of single-step shaping PM when material costs and the ratios of starting to final weight of the parts increase.

Parts with Inherent, Functional Porosity. Particles with shapes ranging from spherical powders to elongated fiber chunks or combinations with metal fiber fleeces are used for parts with inherent, functional porosity [16]. Sintering is stopped when contacts between adjacent particles provide sufficient mechanical strength, but interconnected (open) porosity still exists. This means that pore channels still connect the surface to the interior of the parts. Examples are electrodes of cardiac pacemakers, noise-reducing filter elements, and self-lubricating bearings. For the majority of these parts PM is the only cost-effective production method.

1.3. History and Developments

The Iron Pillar of Delhi illustrates the early use of a PM-related technique in the 2nd century B.C. Iron-rich nuggets were separated from nonmetallic material by warm forging and then sinter-forged into a pillar weighing more than 6 t. Similar techniques were used for martial weapons such as the Damascus swords. Platinum powder from platinum sponges was consolidated by PM techniques between 1800 and 1860 in Russia. Around 1900 COOLIDGE developed a production method for tungsten and molybdenum heating filaments used in incandescent lamps. In the 1920s hard PM materials (e.g., WC – Co hardmetal) were developed and pro-duced. Self-lubricating PM bearings based on bronze and brass penetrated the market in the 1930s. Sintered magnets and ceramic cutting tools followed shortly after and the technology to produce parts made of niobium, zirconium, titanium, and rhenium became available in 1940. Sintered structural parts, made mainly of ferrous materials, were developed and applied at the same time. Development of new PM materials and processing techniques was boosted after the World War II. Heating elements from $MoSi_2$ oxide-dispersion-strengthened alloys (ODS) on the basis of aluminum, nickel, and titanium alloys from rapidly quenched powders or aluminum composites are examples of a broad variety of materials developed for applications in the nuclear, electronics, aircraft and space industries. Continuous progress in press technology and starting powder quality has led to structural parts of high precision and complex shapes.

Developments in sensor and control techniques of presses and sintering furnace atmospheres [17], [18] will expand the production of hardenable, structural, ferrous parts for high-strength applications. The cheaper PM production route is increasingly being applied for production of multifunctional parts and parts with complex shapes not particularly suitable for the PM method by using secondary operations, especially machining and joining.

PM-generated materials are expected to play a key role worldwide as high-performance materials for critical applications in technologically advanced products such as high-speed trains or space planes. Major progress is also expected from hybrid technologies, where advanced PM processing is combined with other techniques such as chemical vapor deposition (CVD) . An example is the development of ultrafine-grained, hard-metal parts which have superior properties compared to standard qualities. New ultrafine powders are processed to cutting chips by hot isostatic pressing and subsequently CVD-coated with TiN-based surface layers [19].

1.4. Economic Aspects

Metal powder shipments in 1987 (without former Eastern Bloc) follow [20]:

Iron and steel	470 000 t
Copper and copper alloys	45 000 t
Aluminum	50 000 t
Stainless steel	15 000 t
Nickel	40 000 t
High-speed steels	10 000 t
Tungsten and tungsten carbide	25 000 t
Total shipment	655 000 t

By far the largest quantity consists of ferrous powders with copper and nickel as the main alloying powder additives for PM steel production. The major portion (ca. 65 %) of the ferrous parts is supplied to the automotive industry. The state of the ferrous PM industry is therefore intimately coupled with the economic situation of the automotive industry [21].

The second most important PM branch comprises hardmetals. Their annual gross product is approximately equal to that of the ferrous parts PM industry.

2. Powder Production

Powders for PM production are produced by mechanical, chemical, and physicochemical methods. The raw materials and the production method determine the size, size distribution, shape, extent of impurities, and the price of the powder [1], [22].

2.1. Mechanical Methods

Mechanical disintegration may be applied to metallic solids or melts.

2.1.1. Mechanical Disintegration of Melts

Atomization. During atomization, free-falling or confined liquid metal streams are disintegrated by high-pressure gas jets or liquid jets (predominantly water) into small melt ligaments and fine droplets (Fig. 5). The usual gas or liquid jet pressures (gas atomization < 3 MPa, water atomization < 15 MPa) result in average powder sizes of 30 – 100 µm. High-pressure atomization (8 – 60 MPa) and special methods (e.g., ultrasonic gas atomization) yield average particle sizes between 3 and 30 µm [22–26].

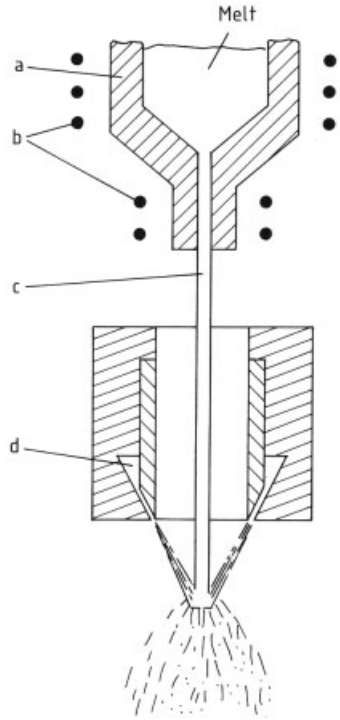

Figure 5. Principal arrangement of an atomizer
a) Melt crucible; b) Induction coils; c) Melt stream; d) Gas or water

Hydrodynamic vertical interaction between the gas jets and the metal stream is based on Taylor instabilities, parallel interaction results from shearing of the melt crests. Similar disintegration principles hold for the water – melt interaction. Water jets may also disintegrate prior to interaction and splash on the melt stream surface as individual water droplets [26], [27]. After penetrating the melt these droplets form collapsing vapor bubbles which disintegrate the melt stream by thermal breakup.

One of the most popular empirical formulas used to estimate the influence of processing parameters on the average particle size d_m is given for gas atomization by LUBANSKA [28]

$$d_m = K \frac{\eta_m}{\eta_g} \left(1 + \frac{V_m}{V_g}\right)^{1/2} \quad (1)$$

where η_m and η_g are the dynamic viscosities of the melt and the gas stream, respectively, and

V_m/V_g represents the mass flow ratio of the melt stream and the gas jet. Fine powders require the transmission of a high momentum from the gas jet to the melt stream and a high ratio of the momentum to the melt mass. An increased impulse is obtained by increasing the gas jet pressure and/or by using confined nozzle systems where the gas jets and the melt stream are guided closely together before interaction starts. The powder size distribution of atomized powder can usually be described by a logarithmic normal distribution [29].

The atomization conditions also determine the shape (morphology) and the internal microstructure of the powders. The atomizing media may form reaction products (e.g., oxides or hydroxides) at the powder surface which hinder the spheroidization of the initially irregular melt ligaments. Air and water atomization are especially susceptible to the formation of reaction products and irregularly shaped powder, in contrast to protective gas atomization with helium, argon, or nitrogen.

Irregular powder is also obtained when solidification starts before spheroidization is complete. Powders that are atomized by liquids cool more rapidly and therefore often show irregular shapes. Powders that are atomized with protective gases have lower cooling rates and minor surface reaction products, they are therefore usually spherical.

The cooling rate in a droplet dT_d/dt during gas atomization is determined by the droplet size d_d, the heat-transfer coefficient h, the thermal conductivity K_m, and the heat capacity of the melt c_{pm}:

$$dT_d/dt = \frac{6h\Delta T\Omega}{c_{pm}d_d} \tag{2}$$

where ΔT is the temperature difference between the droplet and the atomization medium and Ω is the molar volume of the melt.

$$h = (2K_d/d_d) + 0.6\,(\Delta v/d_d)^{1/2} \cdot \left(K_m^2 c_{pm}\right)^{1/3} (\varrho_g/\eta_g)^{1/6} \tag{3}$$

where Δv is the velocity difference between the droplet and the atomization medium, and ϱ_g and η_g are the density and dynamic viscosity of the atomization gas, respectively [30]. The cooling rate is independent of the thermal conductivity in the interior of the droplet, i.e., the temperature gradients in the small droplets are vanishingly small. Typical cooling rates in argon-atomized copper droplets with diameters of 100 µm and 10 µm are 5×10^3 K/s and 5×10^5 K/s, respectively. During water atomization cooling rates are one to two orders of magnitude higher.

Fragmentation by Centrifugal Processes. A thin liquid film on a rotating electrode or disk is fragmented by instability caused by centrifugal forces [31]. The surface tension due to the specific surface energy and the curvature of the liquid surface are restraining forces. Higher centrifugal forces and a slower rate of melt supply lead to finer powders.

Particles produced by centrifugal processes show a narrow size distribution. Their average size is much coarser, however, than that of gas- or water-atomized powder. Centrifugal methods are often applied for high-melting alloys (W, Pt – Ir alloys) and for alloys which react in the molten state with the crucible materials (Ti alloys).

2.1.2. Mechanical Disintegration of Solids

The mechanical disintegration of solids is decreasing in importance as a primary powder production process. Nevertheless, it is still of importance as a secondary process (e.g., to obtain defined particle-size distributions for brittle powders, to destroy aggregates in ceramic and refractory metal powder, or to recycle hardmetals). Interaction between the mill and the powder involves microforging, fracture, and agglomeration. Typical mills are tumbling ball mills, vibration ball mills, attrition ball mills, hammer and rod mills, and the coldstream impact process [1]. See also → Size Reduction.

High-energy milling in tumbling ball mills or attritors is a new method. Large rotating cylindrical drums contain the powder and milling balls which fall onto the powder leading to microforging (strain hardening and rejoining) and fracture of the powder particles. The average particle size approaches a constant value which increases with the ductility of the powder and the milling temperature [32], [33].

For ductile metal powders average particle size is usually > 100 µm. Brittle ceramic or

metal powder may be $< 1\ \mu m$. Liquid milling additives, e.g., 2-propanol reduce the average particle size. Frequent fracture and rewelding of the powders refine their microstructures and allows (1) production of "alloyed" powder by milling mixtures of different elemental powders or (2) the working of dispersions of fine hard particles (e.g., ZrO_2 or Al_2O_3) into a metal matrix [34], [35].

The coldstream method is based on the collision of fine powder particles which are transported in a gas stream to a target plate. This method is used particularly for recycling hardmetal under clean conditions. The jet-mill method is based on the counterflow of two particle-loaded gas streams. Particles fracture when they collide with each other.

Mechanical disintegration increases the dislocation density in the metal powders up to $10^{15}\ m^{-2}$ and reduces their compactability. In many cases subsequent heat treatment is applied to soften the milled powders before compaction.

2.2. Physicochemical Methods

The physicochemical methods depend on the controlled variation of chemical and physical parameters, such as pressure, temperature, and electrical current to obtain powders of metal elements, alloys, or compounds. The carbonyl and electrolytical methods are most frequently applied.

Carbonyl Method. A number of metal compounds (e.g., carbonyls and chlorides) decompose under certain conditions of pressure, temperature, and time into gas and fine-grained metal powders that are often of high purity and excellent sinterability. The carbonyl method is mainly used for the production of iron and nickel powders (→ Iron Compounds, Chap. 3.5.; → Nickel, Chap. 6.3.) [36]. Scraps of metal (metal junk) are reacted with carbon monoxide at high pressure (30 MPa) and temperature (200 – 250 °C) to give metal carbonyls. After distillation the metal carbonyls are decomposed at lower pressures in the same temperature range to form mostly spheroidal powders (size 1 – 10 μm). The impurity content of silicon, phosphorus, sulfur, manganese, and copper is in the low ppm range. However, the contamination due to carbon is 0.1 wt % and that of oxygen is 0.1 wt %.

Electrolytical Methods. Electrolytical powder production from aqueous solutions is mainly used for copper, iron, nickel, and cobalt powder [37]. Metal cations are obtained from dissolving anodes or appropriate electrolytes. The powders formed at the cathode are usually of high purity but brittle and require a soft annealing to improve their compressibility.

Average particle size, size distribution, and shape are determined by the current density at the cathode, the temperature, concentration, acidity, and the vigor of stirring of the bath. Dendritic particles are usually obtained. Copper powder, for example, is obtained by dissolution of copper anodes into acidic $CuSO_4$ electrolytes and subsequent reduction of the Cu^{2+} cations at the cathode (→ Copper, Chap. 6.4.) [38]. Cobalt powder is precipitated from $CoCl_2$ (→ Cobalt and Cobalt Compounds, Chap. 5.), and nickel from aqueous $NiSO_4$ solutions (→ Nickel, Chap. 6.2.1.). The possibility of precipitating metal alloys has been documented for copper alloys but is less common [39].

Highly reactive metal powders are electrolytically precipitated from molten salts. Appropriate choice of the parameters for the electrolysis of titanium, niobium, and tantalum powders provides flaky powders which are used for the production of capacitors (Nb, Ta). Beryllium powder is obtained by electrolysis of a molten $BeCl_2$/NaCl bath at 330 – 380 °C [40].

Other Physicochemical Methods. Reduction of an aqueous solution of nickel amine sulfate with hydrogen at 1.4 MPa and 200 °C is known as the Sheritt – Gordon process [41]. To start precipitation the solution is seeded with fine nickel particles and catalytic ferrous sulfate. Variation of processing parameters allows the removal of accompanying alloying elements (e.g., copper and cobalt) by selective reduction. See also → Nickel, Chap. 6.4.

2.3. Chemical Methods

Reduction of Metal Compounds. Approximately half of the amount of ferrous powders used in the PM industry is produced by the

reduction of magnetite (Fe_3O_4) with carbon. Magnetite, coke, and lime (lime binds the sulfur impurities from the magnetite and coke) are placed into silicon carbide saggers (reaction vessels) which are conveyed through a fuel-fired tunnel kiln. The carbon monoxide produced from the coke reduces the ore to iron at ca. 1200 °C. Subsequent milling, sieving, and final annealing in hydrogen result in irregular sponge-iron powder (particle size 40 – 150 µm) [42].

Other powder production processes involving the reduction of solid compounds by gases, metals, carbon, or compounds include the reduction of WO_3 and MoO_3 with hydrogen [43]. This method is also applied to a small extent for iron, cobalt, nickel, and copper powders.

Titanium powder is obtained by the Kroll process. Titanium chloride vapor is heated above 800 °C and reacted with liquid magnesium forming magnesium chloride and titanium metal. The magnesium chloride is removed by vacuum sublimation yielding a titanium sponge which is mechanically disintegrated into powder. See also → Titanium, Titanium Alloys, and Titanium Compounds.

Carbide Powder. Tungsten carbide powder is the major raw material for cemented carbides (Section 9.4). Powder (particle size 0.5 – 5 µm) is manufactured by heating an intimate mixture of tungsten and carbon powders in graphite tube furnaces in a hydrogen atmosphere at 1400 – 2000 °C (→ Carbides, Chap. 2.1.) [44], [45]. TiC and TaC are obtained from mixtures of their oxides with soot (i.e., of TiO_2 and Ta_2O_5 with C), heated in vacuum to 2300 and 1700 °C, respectively (→ Carbides, Chap. 2.2. → Carbides, Chap. 2.3.). The reaction products are broken up in jaw crushers and milled into powder in carbide-lined ball mills.

3. Powder Properties and Characterization

The economical use of the powders often requires their characterization according to the following criteria:

1) Bulk and surface composition
2) Shape, size, and microstructure
3) Technical properties

Bulk and Surface Composition. The chemical composition of the powders is determined by usual analytical methods which, however, yield average values; the composition of individual particles may still vary widely. The large specific surface area of fine powders requires distinction between internal composition and surface contamination. Typical surface contaminants are adsorbed gases or oxide/hydroxide scales which are analyzed by Auger electron spectroscopy [46]. Testing for hydrogen loss provides a guide to the oxygen content. The amount of acid-insoluble, nonmetallic inclusions can be estimated by dissolving the metal in mineral acid.

Table 1. Measurement methods for particle size and particle-size distribution

Method	Particle size, µm
Microscopy	
Light microscopy	> 1
Scanning electron microscopy	> 0.05
Transmission electron microscopy	0.001 – 1.0
Classification methods	
Sieve analysis	30 – 500
Air classification	0.01 – 50
Sedimentation analysis	
In a gravity field	1 – 60
In a centrifugal field	0.01 – 10
Field distortion analysis	
Electric sensing zone method	> 0.5
Light absorption	> 2.0
Light diffraction	2 – 300
Light scattering	0.1 – 0.7
Electrophoresis	< 0.1
X-ray techniques	
Line broadening	0.002 – 0.2
Low-angle scattering	< 0.05
Surface analysis	
BET method	< 1.0
Gas permeability analysis	< 40

Shape, Size, and Microstructure. *Powder morphologies* are qualitatively documented by scanning or transmission electron microscopy (for extremely fine powders). Quantitative morphology description by Fourrier analysis has not found commercial application [47].

A variety of methods are available for the determination of the average particle size and the particle size distribution (Table 1). See also → Particle Size Analysis. An important technical approach for the determination of the *average particle size* is the gas-flow technique em-

ployed in the Fisher subsieve sizer [48]. The pressure drop of gas passing through a slightly tapped powder bed depends on the specific surface area of the powder. The average particle size is calculated from the measured specific surface area. Other techniques are line broadening in X-ray diffraction measurements and the adsorption of noble gases by the BET method [49].

For many applications the *particle size distributions* are also required. The simplest method is sieving in a sieve column with a mesh that decreases from top to bottom. The method yields approximate results and is applicable only to powders $> 30\,\mu m$. Air classification methods are also used.

Sedimentation in liquids is based on the principle that particles move under constant acceleration (gravity or centrifugal) in viscous media after an initial period with a constant velocity which increases with their size. The time-dependent attenuation of white light or finely collimated X rays transmitted through settling suspensions is used to measure the particle size distribution [49], [50].

Field distortions caused by individual particles are measured by following changes in the electrical conductivity of electrolytes (electric sensing zone [51]) or by the light blockage principle (obscuration [52], [53]). The most frequently used method depends on the scattering of laser light. When coherent monochromatic laser light is scattered by small particles the angular intensity distribution of the diffraction pattern depends on the particle size. Scattering by particles with sizes between 2 and $300\,\mu m$ is analyzed by Fraunhofer diffraction. The Mie scattering theory is applied to sizes ranging from 0.3 to $2\,\mu m$. Very fine powders ($0.1-0.5\,\mu m$) are analyzed by the scattering of chromatic light perpendicular to the direction of the main beam [54].

Technological Properties. Technological properties of the powders are flow behavior, apparent density (filling behavior and tapping response), compressibility, and compactability. They are tested with methods that simulate practical conditions [55].

Flow behavior (flowability) improves with the integrity (compact shape) of the individual particles and their sphericity. The flow behavior of fine powders also depends on the surface properties and atmospheric humidity which may result in the formation of condensates in particle contact areas [55], [56]. These liquid bridges exert high capillary forces on the particles which reduce the flow capability compared to a dry, anhydrous powder. Very fine powders (e.g., ceramics, refractory metals, or carbides) usually possess a very limited flowability which can be measured in a funnel of defined geometry (Hall flowmeter) [57]. They are conditioned into flowable agglomerates (Chap. 4).

Good filling behavior and response to tapping provide high apparent or green densities (green density denotes the density after compaction). Powders with high apparent densities offer beneficial conditions for compaction into parts due to the smaller ratio of the initial to the final part size. The *apparent density* is measured as the mass of powder which fills a given container under free flow. The apparent density of a spherical powder is superior to that of an irregu-lar or dendritic powder. Large powders are superior to fine powders [58]. Broad particle-size distribution is superior to powders of uniform size [59]. Additional tapping or vibration increases the apparent density [57].

The *compressibility* is defined by the relative green density obtained by compacting a powder at a certain pressure, and is usually tested in dies which simulate the geometries of real applications. At the same pressure powders of ductile material yield higher densities than brittle powder. Impurities can reduce the compressibility dramatically (e.g., N_2 in Fe) [23]. Coarse powders ($60-150\,\mu m$) are more easily compressed than fine ($0.5-10\,\mu m$) or ultrafine ($< 0.5\,\mu m$) powders. The *compactability* of a powder is defined by the green strength, the edge strength, and the density in the neutral zone of the compacts [23].

4. Powder Conditioning

Powder conditioning of ferrous powders is frequently performed by the powder producer who delivers a ready-to-sinter powder. Some steps of powder conditioning for high-alloyed steels, hardmetals, and special low-volume production batches (e.g., hard magnets) are mostly done by the part manufacturers. Powder conditioning includes

1) Milling, annealing, and classification
2) Addition of binders and pressing aids
3) Blending
4) Agglomeration and granulation

Milling, Annealing, and Classification. Many production methods, such as the chemical reduction of Fe_3O_4 or WO_3 (Section 2.3), provide metal powder cakes which have to be disintegrated by soft milling to obtain powder suitable for PM production (Section 2.1.2).

The mechanical strain hardening of the powder during milling and uptake of humidity or oxidation during improper storage require a softening or reduction annealing at temperatures below the sintering temperature of the powder. Typical reduction temperatures for copper, nickel, and iron powder in a dry hydrogen atmosphere are 250, 650, and 700 °C, respectively [60]. The particle-size distributions of as-produced or granulated powders usually do not result in optimal flowability, filling density, and sintering be-havior. Different size classes are obtained by sieving, wet sedimentation, or air classification of the powder (Chap. 3), thus removing unfavorably small or large particles. Particle-size distributions with optimal properties are generated by mixing different powder size fractions.

Binders and Pressing Aids. Addition of binders and pressing aids reduces friction between the particles and between the powder and the pressing dies [61]. This results in higher, more homogeneous pressing densities; extended tool life; and an increased cohesiveness, i.e., reduced susceptibility to the formation of cracks or laminations during ejection from the pressing dies [62]. The pressing aids (usually zinc stearate) for ferrous structural parts are added at levels below 1 wt %. The common additive for hardmetals is paraffin which is added at levels to several weight percent [45]. These additives decompose during the sintering operation in the preheating zone of the furnace (see Fig. 13). Binders (e.g., wax) are added at levels of up to 45 vol % in metal injection molding [63], [64].

Blending. Different powders are blended by dry mixing or mixing of suspensions. Mixing quality depends on the loading, the rotational speed, and the type of the blender (usually double cones and tumblers). A statistical distribution is often not achieved if small and large, spherical and irregular, or low- and high-density powders are blended.

Demixing during further handling of ferrous powders is avoided by slight sintering of fine alloying powder particles onto the surface of the coarse iron particles (Distalloy) [65].

Agglomeration and Granulation. Micrometer-sized powders usually show inadequate flow and mixing behavior. For processing of hardmetal or ceramics, flowable powder agglomerates are produced by agglomeration, spray drying, or granulation. Bonding forces between the powder particles are provided by solid-state particle contacts, capillary forces at liquid menisci, adhesive binders, or van der Waals forces [66].

During *agglomeration* the powder particles are moved in tumbling cylinders or on inclined plates. Agglomerates form due to joining of fine particles to the surface of agglomerate nuclei. Friction among agglomerates yields spherical shapes. The cores of the agglomerates usually have a lower density than the peripheries [67].

During *spray drying* a liquid suspension (in water or organic solvents) of fine particles is atomized into fine droplets [68]. The water or solvent evaporates during free flight of the droplets which are transformed into dry agglomerates.

For *granulation* fine powder is slightly compacted into a cake-like block which is subsequently pressed through a sieve with a mesh size $>100\,\mu m$.

5. Pressing and Forming at Ambient Temperatures

Powders are consolidated to obtain parts or semifinished products with defined shape and density. Densification and forming for the two-step consolidation route (Fig. 1) aim to provide intermediate products with a density and strength that are sufficient for further handling and result in satisfactory properties after further processing. Densification and/or forming are performed at ambient temperature, most frequently by uniaxial pressing (Section 5.1). Cold isostatic pressing (Section 5.2) or injection molding (Section 5.3) is used for special

purposes. Less common are powder rolling or extrusion at ambient temperature (Section 5.2). The parts are subsequently further consolidated, most frequently by sintering, to give structural parts or semifinished products composed of high-performance materials. In some cases the preconsolidated parts are used for subsequent powder forging or hot isostatic pressing.

Shaping and consolidation are characterized by [69]

1) Initial rearrangement of particles which yields a denser packing without major particle deformation
2) Plastic deformation of particle contacts related to the approach of the particle centers, and increase both of the average contact area and the average number of contacts per particle [70]
3) A decrease in the further compressibility of the powder associated with increasing density is caused by the pressure decrease in the enlarging contacts and by deformation hardening [71]

Although, theoretical equations relating compact density and applied pressure exist [e.g., 70], empirical relations with fitting parameters are common in industrial practice. For structural parts the Konopcky relation [71]

$$P = C \ln U_0/U$$

is often used, where P is the compaction pressure, C is a constant to be determined for a given number of compactions, and U_0/U is the ratio of initial to final porosity after pressing with P. More detailed information on densification in different regions of the dies, resulting in optimized compaction parameters and die design, is given in [72–75].

5.1. Uniaxial Pressing

Powder is unidirectionally compacted in rigid dies with profiles that are perpendicular to the pressing axis and resemble the shape of the final part. The pressing tool set consists of the die, upper and lower punches, and possibly core rods (Fig. 6). The integral or split die is usually automatically filled by a powder-filled tube (Fig. 6 a) that sweeps over the die plate with the upper punch removed. Compaction is achieved by vertical movement of the upper (6 A) or the upper and lower (6 B) punches with a fixed (6 A, B, C, D) or floating (6 E, F) die arrangement. The green compact is ejected either by the lower punch moving upward (6 C, D, E) or by pulling the die downward (6 F). The ejected compact is removed either by the sweeping filling tube or by a separate fork.

During the movement of the punches, friction between the powder and the die walls leads to a decrease of the pressure transmitted into the powder. Maximal press densities are therefore measured close to the moving punches (6 C – F). Lowest densities are found in regions of minor displacements between the powder bed and the die walls, e.g., close to fixed punches in a single moving punch – fixed die arrangement (6 C) or in the central area (neutral zone) between two moving punches in a fixed (6 D) or floating (6 E, F) die [75]. The floating die is displaced by the friction forces between the moving powder bed and the die walls. Local variations in compact density often result in crack formation during ejection (elastic relaxation) and the loss of the initial compact shape as well as possible crack formation during sintering (differential sintering).

Uniaxial compaction of parts with different thicknesses in the pressing direction requires split punches to maintain a constant densification ratio, i.e., a constant ratio between the filling height and final local height of the compact. Movement of the split die is controlled by mechanical, pneumatic, or hydraulic driving systems which are precisely tuned by sensor systems [76]. Parts with severe undercuts are compacted with split-die and split-punch arrangements (Olivetti process).

Compaction pressures between 400 and 600 MPa are applied for ferrous structural parts. The range 50 – 800 MPa is used for other materials such as aluminum alloys and high-speed steels. The maximum force of modern presses is in the order of 15 MN, which limits the compaction area of the structural parts to $< 60\,cm^2$.

5.2. Isostatic Pressing, Cold Rolling, and Extrusion

For *cold isostatic pressing*, powder is filled into elastic molds that may have a rigid core or mantle

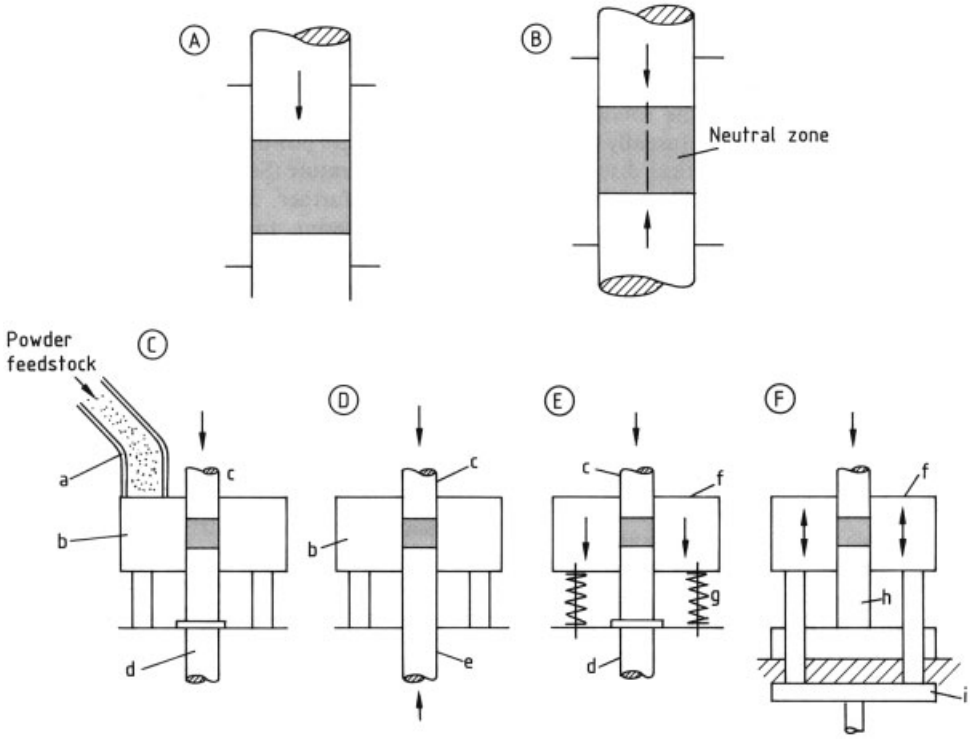

Figure 6. Uniaxial compaction of metal powder
A) Density distribution in the compacted powder bed when upper punch moves, die and lower punch are fixed (case C); B) Density distribution when upper and lower punch move and die is fixed or with floating die if upper punch moves and lower punch is fixed (cases D, E, F); C) Single action pressing; D) Ejection process; E) Floating die process; F) Withdrawal process
a) Filling tube; b) Fixed die; c) Upper punch; d) Ejector; e) Lower punch; f) Floating die; g) Spring system; h) Fixed lower punch; i) Withdrawal rods

and are sealed and subjected to hydrostatic pressure in an oil bath [77], [78]. Densification of the powders is usually isotropic. Shape retention of the elastic molds is moderate, but improves considerably if rigid cores or mantles are used. The economical pressure range is 30 –700 MPa.

In the wet-bag arrangement the powder is sealed in a flexible envelope made of poly(vinyl chloride) or polyurethane which is completely immersed in the pressure medium. The mold cavity is filled with powder, sealed with the cover plate, pressure is applied and released, the cover plate is raised, the green compact removed, and the cycle repeated. The wet-bag method has a low production rate compared to the dry-bag variant where the envelope (membrane) is permanently sealed into the pressure vessel (Fig. 7). The dry-bag method is mainly used for the production of spark plugs and thin-walled tubes.

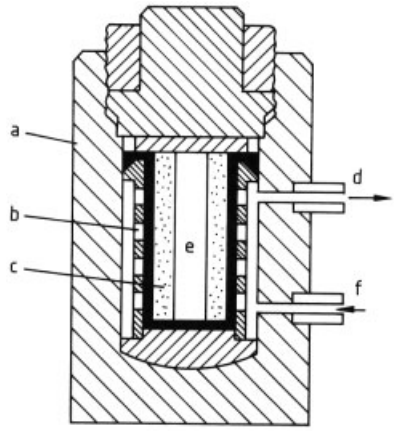

Figure 7. Simple dry-bag tool for cold isostatic pressing [77]
a) Pressure vessel; b) Membrane; c) Powder; d) Pressure outlet; e) Core rod; f) Pressure inlet

Advantages of isostatic pressing are

1) Lower pressure required to obtain a given density
2) More homogeneous axial density distribution in the compacts
3) Greater freedom in the shape choice, e.g., for parts with undercuts and thin walls or with large height to diameter ratios
4) Pressing aids can be avoided

Powder rolling and extrusion are methods used to obtain compacts for semifinished products from powders. In *powder rolling* powder is fed from a hopper into the gap of a rolling mill and emerges as a continuous compacted strip or sheet which is then sintered. The sintered green strip is still porous and can be used to produce large stainless-steel filter plates. The sintered strips have to be rolled again to obtain dense filter plates [79]. Powder rolling is economically advantageous for metallurgical refining methods that provide cheap powders which would otherwise have to be transformed into bulk material by melting. The classical example is the rolling of nickel and cobalt powder produced by the Sherrit – Gordon process [80]. An appropriate powder feed system allows rolling of bimetallic strips.

*Cold extrusion of powders*plastified with a high binder content is used for the production of hardmetal parts with high length to diameter ratios [81]. The plastic binder components must be carefully removed prior to sintering.

5.3. Pouring, Preforming, and Injection Molding

Shapes of filters or electrodes for cardiac pacemakers are manufactured by pouring powder into molds, density may be increased by vibration and subsequent sintering. For slip casting aqueous metal powder suspensions are filled into plaster molds [82]. The suspensions are dehydrated via the plaster capillaries to leave a shell of metal powder covering the mold walls. Slip casting is more popular for shaping ceramics.

Preforming requires presintering of poured or slightly compacted powders which are subsequently machined into parts of different geometry. It is frequently used in the hardmetal industry if shaping is otherwise impossible or if a limited number of parts are required (production of a compaction tool set uneconomical).

Plastified powder – binder mixtures are extensively used for powder injection molding, a process commonly employed for shaping thermoplastics. The process allows the forming of complex shapes such as turbine blades or integral turbochargers [8], [9].

High-density compacts are manufactured by high-speed, high-energy compaction methods which apply high local pressures ($< 10^5$ MPa) for short time intervals ($10^{-3} - 10^{-6}$ s) initiated for example by the ignition of explosives [83].

6. Sintering

6.1. The Sintering Process

Sintering occurs when a system of loose or compacted powders is subjected to heat treatment below the melting temperature of all its constituents (solid-state sintering) or of the major constituents (liquid-phase sintering). Solid-state sintering starts when solid contacts between the particles are formed and/or enlarged. This process is driven by a reduction of the net surface energies of the solid – vapor and solid – solid interfaces. Shrinkage (porosity decrease) may occur with a simultaneous decrease or increase of the average pore size and is driven by the same driving force. Porosity reduction means that an increasing fraction of solid material is found in each unit volume cell as a result of rearrangement and/or shape change of the particles (Fig. 8) [84], [85].

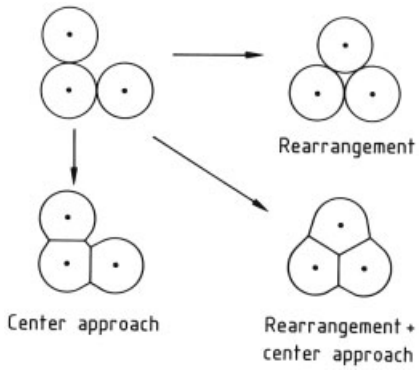

Figure 8. Shrinkage by rearrangement and center approach (shape change) during sintering

6.1.1. Solid-State Sintering

Figure 9 shows the densification of compacts from copper powder during isothermal sintering at 700 – 1000 °C. The shrinkage rate declines continuously during isothermal sintering, indicating a decreasing rate of material transport responsible for changes in the arrangement or shape of the particles. The reduced transport rates are due to lower driving forces, longer transport paths, and the fact that increasing volumes have to be transported to produce equal macroscopic density changes. The shrinkage rate for materials of equal density increases with decreasing particle size and increasing temperature, the transport process is thermally activated [84].

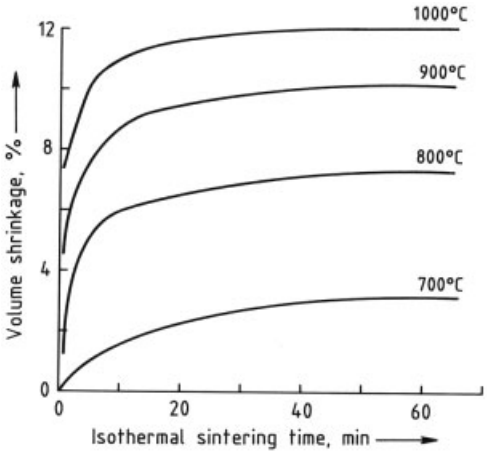

Figure 9. Densification of copper powder compacts (150 MPa, average particle size 45 µm) during sintering at different temperatures

A grain comprises a single crystal with a long-range periodic lattice order. If two such grains come together, the lattices point in different directions and form a grain boundary. A particle may consist of one or many grains.

Metallographic sections of the copper compacts at densities of 70, 86, and 96 % of the theoretical density (i.e., 30, 24, and 4 vol % porosity, respectively) characterize three subsequent stages of different pore arrangement during sintering (Fig. 10). In stage 1, solid material bridges form in the necks between individual particles. The reduction in porosity is modest until the size of necks is approximately one-third of the particle diameter [86]. Grain growth to sizes above that of the initial powder is usually suppressed. If a loose arrangement of spheres is considered (initial density 58 – 68 % [87]) the transition to stage 2 occurs at densities of ca. 75 % of the theoretical. Individual particles can no longer be distinguished and pore channels are situated along triple junctions of adjacent grains. Stage 2 is characterized by rapid shrinkage and slow grain growth [88], [89]. Stage 3 starts at densities between 91 and 95 %; the interconnected pore channels become unstable, forming isolated pores which lie at the grain boundaries or inside the grains [90], [91]. The open porosity is thus transformed into a closed porosity. During the final stage of sintering both slow densification (pore shrinkage) and fast coarsening (pore coarsening and grain coarsening) occur. During the production of parts with inherent porosity and structural parts, sintering is terminated during stage 1 or early in stage 2. Due to the high initial (green) densities of structural parts, the transi-tion from stage 1 to stage 2 occurs at densities of 86 – 92 %. Sintering of structural parts is thus usually restricted to stage 1, i.e., sintering is terminated after the formation of sintered contacts between adjacent particles without pronounced shrinkage.

Sintering theory provides a helpful general picture of the material transport processes that occur during sintering. The influence of parameters such as particle size, initial density (green density), temperature, and time has to be explained to allow control of dimensional and microstructural development during sintering. Sintering involves many particles, as a result theory cannot explain all details, especially in more complex systems, nor can it provide accurate quantitative predictions.

During solid-state sintering of amorphous material [92], center approach occurs by viscous flow in the neck region. During solid-state sintering of crystalline particles, center approach occurs when vacancies that emanate from the surface of the neck regions (source) are eliminated at the grain boundaries in the particle necks (sinks) [93]. The vacancy flux and the equivalent counter flux of atoms are driven by stress gradients in the contact areas that result from the surface curvatures in the neck regions. Figure 11 shows the five transport paths for the atoms (as counter flux to the flux of vacancies): grain boundary diffusion, bulk diffusion from

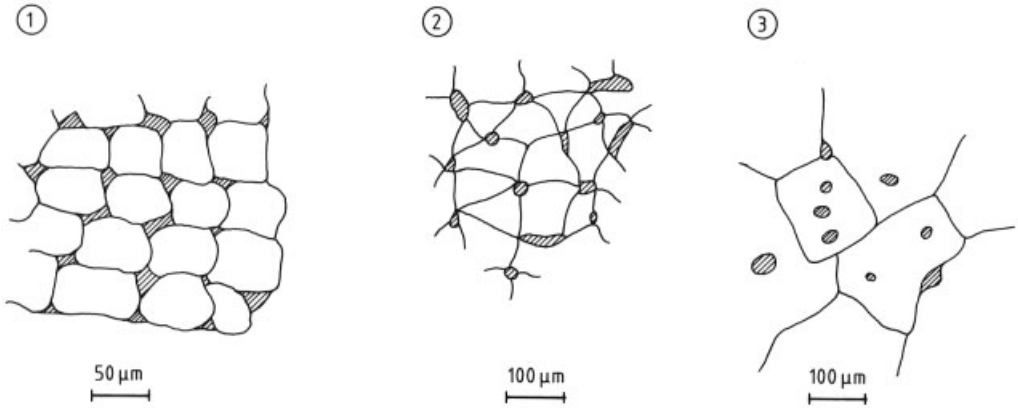

Figure 10. Stages of solid-state sintering

the neck to the grain boundary, bulk diffusion from distant surfaces to the neck surface, surface diffusion, and evaporation and recondensation. All fluxes lead to neck growth, but only grain boundary diffusion (Gb) and bulk diffusion from the neck to the grain boundary (V_1) are involved in the center approach of the particles and macroscopic shrinkage.

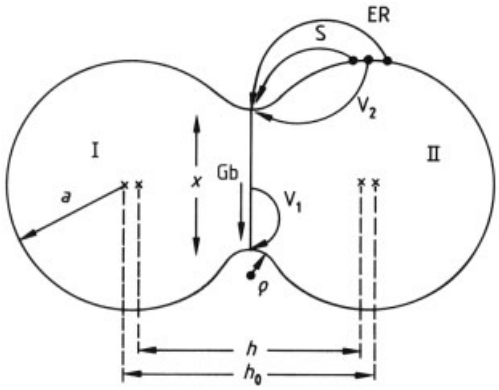

Figure 11. Two-particle model of solid-state sintering
ER = evaporation and recondensation; Gb = grain boundary diffusion; S = surface diffusion; V_1 = bulk diffusion from the grain boundary to the neck; V_2 = bulk diffusion from distant surfaces to the neck surface
a = radius of the initial sphere; h_0 and h = initial and actual distances between centres of the initial spheres, respectively; x = neck diameter; ϱ = main radius of curvature of the neck surface which is assumed to be part of a circle

Equations describing the neck size, x/a, and the relative shrinkage, $\Delta h/a$ (see Fig. 11) during initial sintering (stage 1) are of the general form [84]

$$(x/a)^n = Ct/a^m \tag{4}$$

and

$$(\Delta h/a)^{n/2} = (Ct) / (2^n a^m) \tag{5}$$

The parameters n, m, and C for the different transport processes are given in Table 2 [94].

The models for stage 2 assume pores either to be situated as continuous constant diameter channels along the triple junctions of an arrangement of grains of space-filling Kelvin bodies or as continuous channels with varying diameter along the triple junctions of irregular grains [98]. The equations are similar to those for stage 1 (Kelvin bodies are regular bodies with flat surfaces which can fill space perfectly in a regular arrangement, e.g., a cube but not an octahedron).

A similar approach was used for stage 3, when isolated pores of uniform size were assumed to be present at all corners of the space-filling polyhedrons (grains). This approach becomes invalid, however, when pores separate from grain boundaries.

During grain growth in stage 3, pores remain at the boundaries or become separated and "trapped" inside the grains (Fig. 12). Trapped pores usually shrink at much lower rates than the pores at the grain boundaries, since bulk diffusivity is much smaller than grain boundary diffusivity [90], [91], [99].

On the basis of the sintering equations described above, Ashby developed a graphical representation of the sintering behavior of powders [100], [101]. These sintering diagrams plot the relative neck radius (x/a) or the relative density as a function of temperature and time. Areas

Table 2. Parameters for initial sintering (stage 1) of spherical particles for Equations (4) and (5)

Transport	Measured		Plausible			Reference
mechanism	n	m	n	m	C	
Viscous flow	2	1	2	1	$3\gamma_{SV}/2\eta$	[95]
Evaporation and condensation (ER)	3 – 7	2 – 4	3	2	$2.4\varepsilon Z$	[96]
Grain boundary diffusion (Gb)	6	4	6	4	$48 D_g \delta Z$	[97]
Bulk diffusion from neck (V_1)	4 – 5	3	5	3	$16 D_V Z$	[97]
Bulk diffusion from surface (V_2)	4	3	4	3	$20 D_V Z$	[97]
Surface diffusion (S)	3 – 7	2 – 4	7	4	$23 D_s \cdot (\Omega k/R)^{1/3} Z$	[97]

δ = effective thickness of the grain boundary; D_g, D_s and D_v = grain boundary, surface, and bulk diffusion coefficients, respectively; ε = reaction rate constant; γ_{SV} = specific surface energy; η = viscosity; k = Boltzmann constant; R = gas constant; T = absolute temperature; Ω = molar volume; $Z = \gamma_{sv}\ \Omega/RT$.

are identified in which a single transport mechanism has a dominating effect on the neck growth rate. The sintering diagrams are valuable tools for initial assessment of the sintering conditions required for a particular material of known composition and grain size. The accuracy of the sintering diagrams is, however, limited by the accuracy of the models from which the equations are obtained and the diffusion and interface energy data [100].

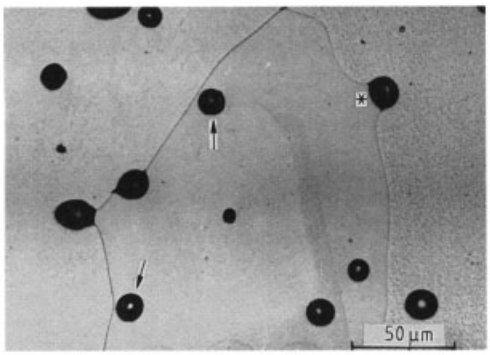

Figure 12. Microstructure showing trapped pores (e.g.∗) and pore/grain boundary separation (→) during final stage (3) sintering of tungsten containing 0.15 wt % nickel at 1400 °C for 3 h

Shrinkage and grain growth of materials as well as dimensional changes and homogenization of metallic structural parts are often manipulated by using sintering additives. Additives, unlike impurities, are intentionally added to the sintering systems. The usual impurity level is below 1000 ppm [1], whereas the amount of additive may vary from 0.1 to 10 wt %. Some additives are present as solid phases, as solid solutions with the host metals, or as segregates at the interfaces of the host metals. Other additives form liquid phases which often penetrate as thin films between the particles (which consist of one or more grains) and grains of the host powder [102]. Additives are frequently used to accelerate shrinkage (e.g., Ni for W or MgO for Al_2O_3), they enhance one of the diffusion mechanisms leading to shrinkage (e.g., grain boundary diffusivity) or suppress mechanisms which lower the driving force (e.g., surface diffusion) [103–107].

Densification, dimensional control, and homogenization can also be influenced by additives which are present as a second solid phase. These additives may provide a fast diffusion path for the host atoms. The additive may also form inclusions or obstacles which retard densification and grain growth. The broad variety of chemical interactions between the solid phases has a complex influence on densification and dimensional control.

6.1.2. Liquid-Phase Sintering

Densification during liquid-phase sintering is also based on rearrangement and shape change of the solid constituent particles (see Fig. 8) [84], [85]. The driving force for both liquid- and solid-state sintering is based on a decrease in energy connected with the changes in the areas of the liquid – vapor, liquid – solid, and solid – vapor interfaces. In liquid-phase sintering melt may originate from low-melting particles or areas which have low-melting point compositions resulting from diffusion during heating. With good wetting, the liquid phase is drawn into particle

necks and small pores by capillary forces [108–111]. As a reaction to the capillary forces, solid particles rearrange if they have sufficient mobility. This rearrangement is a prerequisite for effective densification throughout the sintering process even during later stages when, for example, the rates of particle shape change also limit the rearrangement rate.

Rearrangement may be primary or secondary. In primary rearrangement, the rate is controlled by mechanical movements produced by capillary forces. In secondary rearrangement, the rate is controlled by dissolution and reprecipitation [112]. Shrinkage resulting from the shape change of the solid constituents always involves dissolution – reprecipitation.

During liquid-phase sintering the initial irregular or spherical shape of the single crystal grains becomes polyhedral. Contact flattening is assumed to be a responsible mechanism for this shape change and the shrinkage that occurs during liquid-phase sintering [113]. Contact flattening is a dissolution – precipitation process which is driven by the high chemical potential of the solid matter in the contact area. A thin residual liquid film is assumed to exist between the particles. Under the assumption of local equilibrium the solid phase dissolves at the contact region of two adjacent particles. In the liquid film material is transported out of the contact area in a diffusion- or reaction-controlled manner and reprecipitated at sites of solid matter outside the contact region. The removal of material from the contact area leads to contact flattening. In contrast to the contact flattening theory the shape accomodation theory assumes shape accommodation and shrinkage to be closely connected to grain growth [114], [115]. Shrinkage is described as secondary rearrangement (controlled by a solution – precipitation process) which occurs when small particles dissolve, making way for the movement of larger particles.

Grain growth (coarsening) during liquid-phase sintering of PM materials determines their densification rate and their properties after processing. In many systems grain coarsening follows the relation

$$G_m^n = (G_{m0}^n + kt)^{1/n} \tag{6}$$

which indicates diffusion- or reaction-controlled Ostwald ripening ($n = 3$ or 2, respectively) as the rate-controlling step [116], [117] where G_m and G_{m0} are the actual and the initial grain size, respectively, k is a growth constant, and t is the growth time.

The penetration of melt into grain boundaries is a further important dissolution–reprecipitation mechanism for liquid-phase sintering. Grain boundary penetration leads to skeleton disintegration, but also to contact dilatation, particle dilatation, and particle disintegration. Other important consequences are boundary alloying and liquid film migration [112], [118–120]. Dimensional changes during liquid-phase sintering due to grain boundary penetration are of commercial interest for structural parts. For example, in high-green-density compacts from iron and copper powder blends, copper-rich melt penetrates into the contact regions between adjacent iron particles and along grain boundaries. Additional melt in prior contact areas and at prior grain boundaries inside the iron particles causes contact and particle dilatation leading to macroscopic swelling [121].

6.2. Sintering Furnaces and Atmospheres

Sintering temperatures depend on the alloy type and whether solid-state or liquid-phase sintering is anticipated. Common sintering temperatures are ca. 600 °C for aluminum alloys, 750 °C for copper bronzes, 900 °C for brass, 1120 – 1200 °C for copper, nickel, and Cu – Ni – Mo distalloy steels, 1200 – 1280 °C for high-alloyed steels, 1400 °C for hardmetals and heavy-metal alloys, 1600 – 2500 °C for refractory metals such as molybdenum and tungsten [122]. The most frequently used sintering furnaces are of the belt, walking beam, or batch type (Fig. 13).

Belt Furnaces. Mesh belt furnaces (Fig. 13 A) are used at temperatures up to 1150 °C. The compacts to be sintered are put directly on conveyor belts made of stainless steel nets or are packed into boxes which are transported by the belt through the sintering furnace at constant speed. The nickel – chromium heating elements allow annealing in protective gas which burns as a flame curtain at the entrance and the exit of the furnace. The initial preheating (dewaxing)

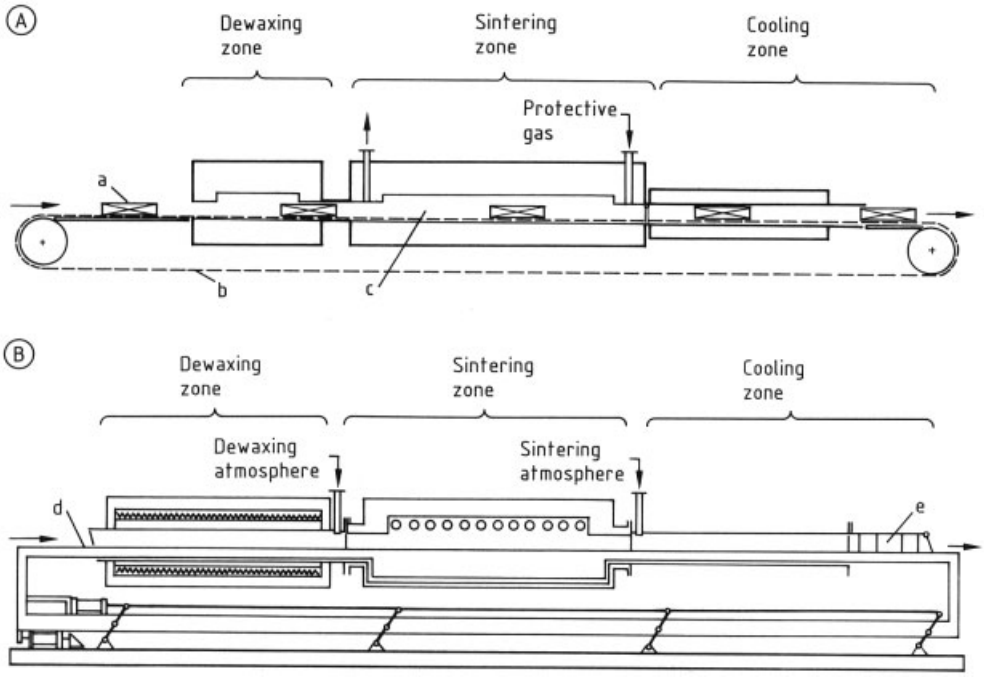

Figure 13. Section through a conveyor belt furnace (A) and a walking beam furnace (B)
a) Sintering tray; b) Conveyor belt; c) Muffle; d) Walking beam; e) Atmosphere lock

zone is operated at 600 – 800 °C; binders, pressing aids, and humidity are decomposed and removed from the parts. The central sintering zone is followed by a cooling zone.

Endo- and exogas are most frequently used as protective gas, they are both prereacted mixtures of methane and air containing carbon monoxide, carbon dioxide, hydrogen, and nitrogen. Exogas has a high partial pressure of carbon dioxide and reduces the oxides of iron, nickel, and copper. Endogas has a low partial pressure of carbon dioxide. The CO/CO_2 ratio determines the equilibrium carbon concentration in sintered PM steels. Exogas stabilizes a high, endogas a higher carbon concentration [123]. Hydrogen or catalytically decomposed ammonia is used less frequently.

Walking beam furnaces (Fig. 13 B) are used at temperatures up to 1350 °C and have molybdenum heating elements. Boats or trays are conveyed and periodically lifted, advanced a short distance, and then settled back on a ceramic ledge. Protective gas flow is controlled by sluices. In addition to the gases described for mesh belt furnaces, nitrogen is also used.

Batch sintering furnaces are used for hydrogen, nitrogen, or vacuum sintering. Hydrogen is used for sintering of tungsten, molybdenum, and iron of high purity, nitrogen is used for aluminum alloys. Vacuum sintering is applied for high-alloyed steel, hardmetals, magnets, and materials containing alloying elements sensitive to oxidation (e.g., chromium and aluminum). Extremely high vacuum is required for sintering niobium and tantalum. Sintering temperatures above 1500 °C are obtained with molybdenum, tungsten, or tantalum heating elements.

7. Pressing and Shaping at Elevated Temperature

Direct consolidation at elevated temperature by processes such as hot pressing, hot isostatic pressing, sinter forging, or hot extrusion results in simultaneous shaping and consolidation. Powder, precompacts, or porous presin-

tered parts are mostly consolidated to full-density products.

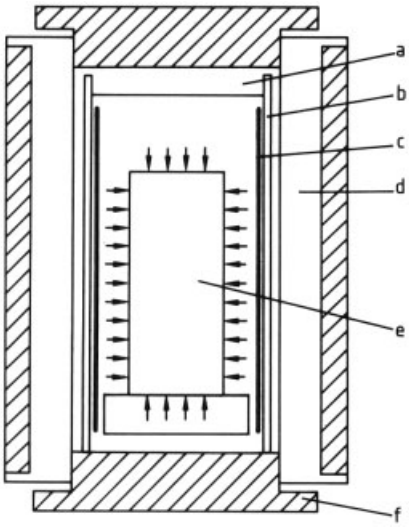

Figure 14. Pressure vessel for hot isostatic compaction a) Isolating lid; b) Isolating mantle; c) Furnace; d) Wire wound pressure vessel; e) Part to be pressed; f) Vessel lid

7.1. Hot Pressing and Hot Isostatic Pressing

Hot pressing is applied for simple geometries such as small cylinders or plates. The pressing tool sets are mostly simplified analogs of those used at ambient temperature and are generally built from graphite, in rare cases from stainless steel, refractory metals, or ceramics. Conventional compaction pressures are between 20 and 50 MPa at temperatures of 600 – 2100 °C. The main application of hot pressing is for cutting and grinding composite tools of diamonds embedded in cobalt or other metal matrices [124].

Hot isostatic pressing (*HIP*) has a broader range of applications than hot pressing [125], [126]. The parts are consolidated and densified in pressure vessels at elevated temperature (up to 2300 °C) and elevated pressure of helium or argon (50 – 200 MPa) (Fig. 14). Parts with open porosity require prior encapsulation in steel or glass containers. Open porosity may be defined as sintering stage 2 porosity which involves continuous channels connecting the interior of the part with the surrounding gas. Powders or precompacts are filled into the containers, which are usually evacuated at elevated temperature before sealing. The heating furnace is placed inside the pressure vessel and well isolated against heat transmission to the pressure vessel walls.

Homogeneous isotropic densification yields final part geometries of good accuracy (near net shape method) which only require minor final machining and show an excellent (low) ratio of initial to final weight. Large parts made of stainless steel (Fig. 4) or other high-value alloys (e.g., superalloys or Ti alloys) are important applications. A superior economy of HIP is realized if encapsulation is not required (i.e., in parts with closed porosity or closed porous defects).

Sintering to closed porosity and immediate complete densification to full density by HIP in one furnace and a single processing cycle (sinter – HIP) is frequently used in hardmetal manufacture. The additional HIP treatment allows the use of cheaper, less well defined initial powder or the production of extremely fine-grained hardmetal materials [19]. Sinter – HIP furnaces work at < 10 MPa and are therefore much larger and cheaper than full pressure (≤ 200 MPa) HIP furnaces [126]. HIP can also be used for defect (porous) healing in sintered materials [127] or, for example, in superalloy turbine parts after a prolonged lifetime [125].

Densification during hot pressing and hot isostatic pressing is mainly controlled by power-law creep and diffusion processes similar to those occurring during sintering. Densification kinetics are described satisfactorily by approaches which include consideration of rate changes caused by grain growth at high densities [128–130]. The micromechanical relations are the basis of the calculation routines for HIP diagrams. They yield predictions of the density of the powder compact depending on the pressing parameter and the initial particle size [128].

More precise predictions of densification and shape stability are obtained by finite element calculations [131], [132].

7.2. Sinter Forging and Other Methods

In some cases the required geometry and mechanical properties of parts make *sinter forging* (powder forging) an economic alternative to other production methods. Presintered, reheated

parts of exact mass and dimensions can be forged in closed dies without flash to a precision not obtained by conventional forging. Compared to conventionally forged parts, sinter-forged connecting rods show superior reliability and fatigue strength due to their improved surface quality and the absence of chemical inhomogeneities or large nonmetallic inclusions [133]. In addition, the properties of sinter forged products are essentially isotropic [134]. Prealloyed Mn – Cr and Mn – Cr – Mo steel powders are usually used for sinter forging of ferrous parts. Aluminum- and titanium-based alloy powders and superalloy powder are increasingly used for superplastic sinter forging.

Hot extrusion of powders provides profiled, dense rods at temperatures that are much lower than those required for densification by hot isostatic pressing. Most powders are encapsulated in a metal container and cold isostatically precompacted in a similar manner to the hot isostatic pressing route. Seamless stainless steel tubes with chemical homogeneities superior to cast and wrought materials are manufactured by this procedure (Asea – Nyby process) [135]. Hot extrusion became very popular in the late 1980s for full densification of metal – matrix composites which did not sinter to full or closed porosity without additional external pressure [136]. The large contribution of shear strains to densification and particle deformation results in the advantageous destruction of reaction products at the surface of the powders. Sometimes this makes hot extrusion preferable to hot isostatic pressing, despite the fact that hot-extruded bars show highly anisotropic material properties. Similar anisotropic properties are obtained after *hot rolling* of encapsulated powders.

Deformation during powder forging, hot extrusion, and hot rolling is rapid and depends mainly on the yield strength of the material which decreases with temperature and on dynamic recrystallization or recovery.

8. Structural Parts

8.1. Structural Ferrous Parts

PM structural parts are produced from metal powder (ferrous alloys, to a smaller extent also aluminum alloys and copper alloys) by compacting and sintering to closely controlled dimensions.

Powder metallurgy is often economically advantageous over other methods for the production of complex-shaped structural ferrous parts. The mechanical properties of PM structural ferrous parts are adequate for many applications but usually slightly inferior to the properties of machined parts, mainly due to the residual porosity [137]. Main customers are the automotive, electrical, and general machinery industries.

Structural parts with densities between 78 and 92 % of theoretical density are obtained by the single press – sinter process. Densities of 92 –95.5 % require a double-sinter process comprising uniaxial cold pressing, sintering, cold repressing, and resintering. Densities > 95.5 % are achieved by powder forging (see Section 7.2) [6].

Data on the influence of the porosity on the mechanical properties of pressed and sintered PM structural parts are given in [138], [139]. Figure 15 shows a compilation of data for iron [140].

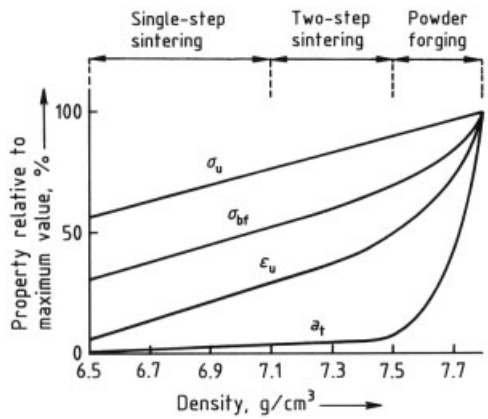

Figure 15. Mechanical properties of ferrous PM material as a function of the density [140]
σ_u = ultimate tensile strength; σ_{bf} = bending fatigue strength; ε_u = ultimate fracture elongation; a_t = fracture toughness

In general, the mechanical properties increase with increasing density. The pore shape determines the extent of their stress intensifica-

tion which particularly affects elongation and impact strength [141], [142]. Tensile strength and hardness show an almost linear increase. The strength of compacts from straight iron powder is limited. Higher strength PM structural parts are produced from powder mixtures of iron and alloying elements, the most important being graphite, copper, nickel, molybdenum, and phosphorus powder [143]. In special cases (e.g., for sinter forging), powders mainly prealloyed with nickel, manganese, molybdenum, and nickel are blended with graphite [144].

8.1.1. Low-Alloyed Steels

Table 3 lists the compositions and mechanical properties of typical ferrous steels densified to three density ranges [139]. Alloying elements of high oxygen affinity such as vanadium, manganese, and chromium are avoided in powder blends to be sintered but common for prealloyed powders to be powder forged. The use of alloying elements (e.g., Cu and P) which lead to grain boundary embrittlement and detrimental problems during warm forming in conventional steels is possible since no further forming step (except coining at room temperature) is required after sintering to obtain the final parts.

Copper additions of 1.5 – 5 wt % provided classical PM steels which required only minor control of the carbon potential of the sintering atmosphere. Iron – copper alloys are sintered at 1120 °C, i.e., above the melting point of copper, which leads to swelling of the compacts by melt penetration into particle contacts and grain boundaries (Section 6.1.2). The swelling may compensate the usual shrinkage during sintering of powder compacts [121], [145]. Copper acts as a solid solution and precipitation hardener [146], [147]. Nickel additions improve the toughness and weldability of ferrous steels.

Fe/Cu/C and Fe/Cu/Ni/C steels are stronger than low-carbon steels but require precise carbon control during sintering. Fe/P PM steels are characterized by a good combination of strength and toughness due to the formation of rounded pores during sintering. The major disadvantage of phosphorus-alloyed steels is their extensive shrinkage due to the formation of a eutectic liquid phase between the Fe_3C alloy powder and the iron particles. In addition, the presence of phosphorus in iron stabilizes the ferrite phase which shows a much higher bulk diffusion coefficient than the austenite.

FeCuNiMoC alloys are mainly manufactured from distalloy powder. Molybdenum suppresses the local enrichment of nickel which would lead to soft spots in the sintered material caused by retention of austenite after cooling.

8.1.2. High-Alloyed Steels

Parts made of austenitic and martensitic stainless steel are produced from prealloyed powders which are usually obtained by water atomization and less frequently by nitrogen gas atomization. The compositions are close to those of cast and wrought stainless steels (e.g., 316 L with the composition Fe13Ni17Cr2.1Mo0.7Si [148]). The high content of alloying elements re-sults in powders with low compressibilities which require lubricants to obtain sufficiently high green densities. These powders have to be dewaxed carefully before sintering to avoid the pickup of carbon. Sintering at 1120 °C in vacuum or hydrogen provides stainless steel parts of higher ductility but lower strength than sintering in a nitrogen atmosphere [149]. Small amounts of free carbon may be added to allow carbon monoxide formation during sintering thereby reducing the oxygen pickup. Austenitic stainless steels sintered in a vacuum to densities of 6.5 g/cm^3 show an ultimate tensile strength of 300 MPa and ultimate fracture elongations of 2 – 4 %. Martensitic stainless steels possess strength values > 450 MPa and negligible elongation (< 1 %). The open porosity of PM stainless steel structural parts has a disadvantageous effect on corrosion behavior, they are therefore often impregnated and sealed with resins or waxes.

The applications for stainless steel PM parts have been extended by the *Asea – Nyby process* for seamless tubes [135], [150] and the manufacturing of small parts with complex shapes by *powder injection molding* (e.g., for orthodontic purposes [151]).

High-speed tool steels are hard, tough, wear-resistant, high-alloyed ferrous materials which contain up to 20 wt % of tungsten, molybdenum, chromium, vanadium, cobalt, and sufficient carbon to form high-volume contents of

Table 3. Composition (in wt %) and mechanical properties of typical ferrous PM steels densified to three density ranges

Composition	Process	$\varrho = 6.4 - 6.8$ g/cm^3		$\varrho = 6.8 - 7.2$ g/cm^3		$\varrho = 7.2$ g/cm^3		
		σ_u, N/mm^2	ε_u, %	σ_u, N/mm^2	ε_u, %	δ, g/cm^3	σ_u, N/mm^2	ε_u, %
Fe	sintered	130	4	190	10	7.3	260	8
$Fe_1Cu_{0.6}$	sintered	460	2	570	2			
Fe_5Ni	sintered	250	5	330	9	7.4	410	18
$Fe_3Cu_{2.5}Ni$	sintered	300	3	440	5	7.4	550	14
$Fe_{0.45}P$	sintered	310	11	330	12			
$Fe_3Cu_{2.5}Ni_{0.8}Mo_{0.50}$	sintered	390	2	570	3	7.3	680	5
$Fe_{1.75}Ni_{1.5}Cu_{0.5}Mo_{0.5}C$	sintered heat treated	680	1.5	1000	1	7.4	1400	1
$Fe_{12}Cr_{0.2}C$	sintered	510	1					
$Fe_{1.9}Ni_{0.55}Mo_{0.25}Mn_{0.4}C$	sintered forged					7.8	770	30
	heat treated					7.8	1150	25

ϱ = density; σ_u = ultimate tensile strength; ε_u = ultimate fracture elongation

tungsten, molybdenum, chromium, and vanadium carbides [152]. Cast high-speed tool steels develop coarse, inhomogeneously distributed carbide networks which have to be homogenized by thermomechanical treatment. Gas or water atomization (with protective N_2 atmosphere) yields powders with homogeneous microstructures and very fine carbide precipitates. The fine microstructures are retained during consolidation. The *ASEA – Stora process* involves inert gas atomization of high-speed tool-steel powder which is filled into steel containers and cold isostatically pressed [150]. After hot degassing at 450 °C and sealing, hot isostatic pressing at 100 MPa and 1100 °C yields dense billets weighing up to 1000 kg. This technique avoids carbide segregation and yields billets with a uniform macro- and microstructure comprising evenly sized and distributed carbide particles with a fine, uniform grain size. Shaped parts are produced from the billets by machining.

Conventional uniaxial or isostatic pressing and vacuum sintering of water-atomized high-speed steel powder is used in the *Powdrex process* to manufacture high-speed tool-steel parts. The sintering temperature has to be kept to within a few degrees of 1240 °C, and the C/O_2 level of the compact must be carefully controlled to allow sintering to full density with a limited amount of liquid phase [153]. At higher temperatures coarse carbides form, at lower temperatures residual porosity is maintained. This net-shape sintering process gives superior material and economy compared to the cast and wrought route [154]. PM steels with high wear resistance are based on high contents of titanium carbide [155].

8.1.3. Secondary Operations

Some secondary operations are applied to both PM and non-PM parts, e.g., hardening by heat treatment, joining, and surface modification. Particular to sintered porous parts are coining operations and pore filling by reaction with steam or by infiltration with metals or polymers.

When ferrous PM structural parts are subjected to quench *hardening* the lower heat conductivity of the porous material must be considered. Materials with open porosity are sensitive to corrosion or oxidation and require protective atmospheres and oil instead of water as quenching medium. For case-hardening of parts with closed porosity, carbonizing or nitrocarburizing in high-carbon or carbon – nitrogen atmospheres and subsequent quenching in oil is used. Parts with open porosity require impregnation (e.g., with waterglass) or prior surface coating to avoid hardening embrittlement and severe dimensional changes in the interior of the parts. Plasma nitriding is increasingly preferred to gas and salt bath nitriding due to the defined high-quality surface layers formed [156].

Almost 90 % of the ferrous structural parts are coined after sintering. *Coining* improves the accuracy of dimensions and shape, the finish, and the mechanical strength of the surface ar-

eas. Threads and holes are machined perpendicular to the pressing direction. Machining is also applied to those surface areas which require exact geometry and superior surface finish [157], [158].

*Surface coating and modification*are used to improve corrosion or wear resistance. Galvanic precipitation, chemical or physical vapor deposition (CVD or PVD), and steam treatment are commonly employed. The first two methods require closed porosity or impregnation of the open porosity with waterglass, waxes, or resins prior to coating. Nickel, copper, chromium, and zinc are mainly deposited galvanically, whereas hard layers of nitrides or carbides are formed by CVD or PVD. During steam treatment at 450 – 570 °C, iron(III) oxide forms in the pores and adheres to the ferrous material, providing gastight closure of the initially open pore space. Self-lubricating bearings are manufactured by impregnating highly porous ferrous parts with lubricants (Chap. 10).

8.2. Copper and Copper Alloys

PM structural parts from pure copper are used for applications where good electrical and thermal conductivity are important [159]. High-purity electrolytic or gas-atomized copper powder is compacted to densities up to 90 %. Higher green densities are possible but outgassing becomes impossible due to the mainly closed porosity which leads to swelling of the copper compacts. (Outgassing is the decomposition of residual impurities or binders into volatile products which leave the compact through the open porosity, usually at slightly elevated temperature and under a partial vacuum.) Parts with a higher final density are obtained by outgassing and sintering compacts with a low green density that have previously been repressed and resintered after initial sintering. At 98 % of the theoretical density ca. 97 % of the International Annealed Copper Standard (IACS, standard electrical conductivity, → Copper Alloys, Chap. 1.) is maintained at an ultimate tensile strength of 200 MPa and at ultimate fracture elongations ≥ 30 %[160].

PM bronze structural parts are generally produced from mixtures of copper and tin powder by methods very similar to those used for self-lubricating bronze bearings (see Chap. 10). Structural parts from brass, leaded brass, and nickel silver (Cu18Ni17Zn) are produced from gas-atomized (usually air) powders [159]. Materials with higher lead and zinc contents require careful control of the lead and zinc evaporation rates during sintering at elevated temperature.

8.3. Aluminum and Aluminum Alloys

See also → Aluminum Alloys, Chap. 4.5.

PM aluminum alloys are successfully manufactured from mixtures of air-atomized aluminum powder with lubricants (e.g., zinc stearate) and up to 8 wt % of alloying elements (e.g., Cu, Mg, and Si). High green densities of 90 – 95 % are obtained with moderate compaction pressures of 200 – 400 MPa due to the low yield strength of the aluminum powders. Lubrication is required to avoid galling and seizing during pressing. The alloying additions form a liquid phase which penetrates below the thin oxide layers coating the surfaces of the aluminum powder particles. Similar to wrought alloys, the PM alloys may be strengthened by precipitation hardening and surface treated by chemical polishing, anodizing, and coloring. Depending on the thermal treatment an ultimate tensile strength between 120 and 300 MPa is achieved at low ultimate fracture elongations (between 4 and 1 %) [137].

9. High-Performance Materials

9.1. High-Melting Metals and Alloys

See also → High-Temperature Materials, Chap. 8.

The refractory metals (i.e., W, Mo, Ta, Nb, Re, Hf, Os, Ru, and Ir) possess melting temperatures above 2000 °C and have a low vapor pressure, low thermal expansion coefficient, high elastic modulus, high recrystallization temperature, and a remarkably high temperature strength [161]. Their mechanical properties sensitively depend on their microstructure (i.e., grain size, dislocation density), and on substitutional and interstitial impurities. Recrystallized tungsten and molybdenum are embrittled by oxygen, sulfur, and phosphorus impurities that segregate at the grain boundaries as atoms or as oxides, for

example. To avoid spalling and splitting as a result of brittle grain boundary fracture, plastic forming of tungsten and molybdenum requires elevated temperature. Molybdenum, tungsten, and rhenium are mainly processed in a hydrogen atmosphere. During application they require protective atmospheres or a vacuum.

In contrast to tungsten and molybdenum, tantalum and niobium dissolve considerable amounts of oxygen at elevated temperature and are less susceptible to grain boundary embrittlement. Tantalum and niobium are usually processed in a vacuum and also require a vacuum or a protective atmosphere during application at elevated temperature. Osmium, ruthenium, and iridium can be used in air [161].

The Coolidge process was invented in 1908 and provided fine-grained tungsten and molybdenum products. Uniaxially (in split dies) or isostatically compacted bars of molybdenum or tungsten powders were presintered at 1200 °C to improve their strength for easier handling. In a subsequent sintering treatment densities > 90 % were reached by passing a low-voltage, high-amperage current through the presintered bars, which reached temperatures > 3000 °C for up to 30 min [162].

In modern production lines, the compacted bars are usually subjected to indirect sintering in tungsten furnaces under a hydrogen atmosphere at > 2000 °C to give densities of 92 – 95 % [163]. Tantalum and niobium are usually sintered by the Coolidge method under vacuum.

The sintered refractory metals are brittle, with porosities up to 10 %. Complete densification and increased toughness are obtained by a multistep thermomechanical treatment. A typicals sequence for manufacturing tungsten wires for incandescent lamps is rolling of the sintered rods at 1600 °C, followed by swaging at 1500 °C. Temperatures are lowered to 900 °C with progressing deformation. The swaged bars are drawn into wires in cemented carbide and diamond dies. The drawing temperatures are gradually lowered from 800 to 400 °C [163]. Tungsten wires for lamps contain the so-called KAS dopant (a combination of K_2O, Al_2O_3, and SiO_2). The dopant is blended with WO_3 before reduction to tungsten powder under hydrogen. The KAS dopant forms strings of fine bubbles parallel to the wire-drawing direction during the final processing steps. Due to the presence of the bubble rows the wires develop a recrystallized structure of elongated grains. Undoped tungsten or molybdenum wires develop an equiaxial recrystallization structure with a considerably reduced lifetime compared to KAS-doped materials. The addition of ThO_2 and KAS simultaneously improves the thermal emissivity of electrons and is mainly used for electronic tungsten or molybdenum materials.

Alloys based on solid solutions of tungsten or molybdenum with rhenium have improved ductility and high-temperature strength compared to the unalloyed materials. They are only rarely used, however, due to the high cost of rhenium metal [80]. The technically most important molybdenum alloy is TZM which contains 0.5 wt % Ti, 0.07 wt % Zr and 0.05 wt % C. See also → Molybdenum, Chap. 4.5.

Liquid-phase sintered tungsten alloys with 2 – 6 wt % Ni, 2 – 4 wt % Cu, and/or 1 – 4 wt % Fe have found major application as high-density shells for penetrators in artillery bullets. At densities of 17.5 – 18.5 g/cm^3 their ultimate tensile strength exceeds 1200 MPa at fracture strains between 10 and 30 % [164].

Niobium and tantalum products are mainly used as sheets and wires for the electronic and chemical industry [80]. A major application of tantalum powder as a substrate for capacitors is described in Chapter 10.

9.2. Light Alloys

9.2.1. High-Performance Aluminum Alloys

For a complete description of aluminum alloys, see → Aluminum Alloys.

High-performance PM aluminum alloys provide improved high-temperature creep resistance, corrosion resistance, and a high specific stiffness. These properties are obtained by rapid solidification, oxide- or carbide-dispersion hardening, and particle reinforcement [165], [166].

Rapid solidification is typically performed by atomization, melt spinning, or planar-flow casting; it mainly provides prealloyed powders with fine microstructures and solid solutions. The rapidly solidified ribbons or sheets obtained after melt spinning or planar flow casting have to be milled to fine powders before processing. Processing usually involves degassing, cold

isostatic compaction, encapsulation, degassing, hot isostatic pressing, decapsulation, and extrusion. An alternative to cold isostatic compaction, encapsulation, degassing, hot isostatic pressing, and decapsulation is vacuum hot pressing without encapsulation.

Rapidly solidified aluminum alloys can be classified into high-strength, high-temperature, and lightweight alloys [167].

High-Strength Alloys. The compositions of the high-strength aluminum alloys are generally identical to those of conventional ingot cast and wrought alloys. See also → High-Temperature Materials, Chap. 4. The strength is slightly improved by the refinement of the microstructure due to rapid solidification and the addition of alloying elements that form finely dispersed precipitates. Typical compositions are cast (ingot metallurgy) IM 2618 and powder metallurgy PM 2618 with Al6Cu0.5Mg1Fe1Ni; ultimate tensile strength values reach 600 MPa in the aged conditions. Other efforts have concentrated on the 7000 series with main alloying elements zinc, copper, and magnesium and small additions of cobalt, nickel, zirconium, or molybdenum [168]. The development of low-density aluminum alloys, including "light" alloying elements (e.g., Li and Be), was dampened by successful IM Al – Li alloy developments [166].

High-Temperature Alloys. See also → High-Temperature Materials, Chap. 4. Excellent mechanical properties (especially thermal stability, creep strength, corrosion resistance, and fatigue strength) have been demonstrated for rapidly solidified aluminum – transition metal alloys. These high-temperature alloys have a high-volume content of precipitates with an improved aging behavior, this requires a low diffusion coefficient and a low solubility of the precipitate-forming alloying elements in the aluminum matrix. The alloying element iron satisfies the conditions of low solubility and diffusion coefficient and is used at levels between 8 and 12 wt % together with up to 4 wt % of a ternary additive (e.g., V, Mo, or Ce). To achieve a sufficient volume fraction of precipitates in the matrix, segregation-free material with extended solid solutions is solidified by argon atomization [169]. Compared to conventional high-temperature aluminum alloys creep resistance is increased by ca. 100 K in an Al – 8 Fe – 2 Mo PM alloy [165]. A second series of high-temperature aluminum alloys receiving attention comprises those based on the Al – Cr system [170].

Aluminum alloys strengthened with dispersoids (very fine inert particles, e.g., oxides, carbides) are manufactured by high-energy milling of aluminum powder blends with dispersoid precursors in attrition or planetary mills [171], [172]. More recently aluminum alloys containing Al_4C_3 and Al_2O_3 dispersoids have been commercialized. Mixtures of aluminum, soot, and an organic milling agent (e.g., 2-propanol) are milled at high energy to obtain homogeneous distribution of the very fine Al_2O_3 particles from fractured powder surfaces and of the graphite particles by continuous fracturing and microforging of the aluminum powder particles (see Section 2.1.2). Reaction of the finely distributed graphite with the aluminum during hot extrusion of the milled powders yields fine Al_4C_3 dispersoids. The thermal stability of the dispersoids makes these dispersoid-strengthened alloys suitable for high-temperature applications.

Lightweight Alloys. Recent years have seen the development of a wide range of lightweight, high-performance aluminum PM composites with a variety of discontinuous reinforcements (e.g., Al_2O_3, B_4C, and in particular SiC). Reinforcement shapes considered for such metal – matrix composites include whiskers, fiber junks, platelets, and spherical particles with mean sizes of 3 – 30 μm. Age-hardenable aluminum alloys of the 2000, 6000, and 7000 series (for applications $< 250\,°C$) as well as rapidly solidified high temperature aluminum alloys (for applications $> 250\,°C$) are used as matrix materials. The composites combine strength values up to 750 MPa with high stiffness [166].

To produce the composites, the reinforcement is subjected to dry or wet blending with the aluminum powder. Mechanical alloying can be used as a processing route for particulate reinforced metal-matrix composites (MMC). Final billet fabrication involves cold compaction, hot outgassing, hot isostatic pressing, vacuum hot pressing, or hot extrusion. At high temperatures (425 – 450 °C) and high strain rates ($0.7\,s^{-1}$) the mechanical alloying composite materials show

extended ductility up to 300 % total elongation due to superplastic behavior [172].

9.2.2. Titanium and Titanium Alloys

See also → Titanium, Titanium Alloys, and Titanium Compounds.

Complex shaped parts from titanium or titanium alloys are manufactured by PM in cases where considerable machining can be avoided and to obtain a low initial to final weight ratio of the finished part. The main user is the aircraft industry (engines and frames). Powders are atomized under vacuum or by the plasma rotating electrode process, outgassed, and filled into containers for hot isostatic pressing. Full density and near net shape processing are required to satisfy the economic criteria. Mechanical properties are comparable to those of thermomechanically manufactured titanium alloys.

The most conventional PM composition is Ti6Al4V. Considerable research and development efforts are being made to exploit rapid solidification techniques for titanium alloys. Preliminary results on a metalloid former Ti8V5Fe1Al powder product obtained by the plasma rotating electrode process indicate that strengths > 1380 MPa can be obtained with ductility levels close to 10 % ultimate fracture elongation [173].

9.2.3. Magnesium and Beryllium

Magnesium is the lightest (density 1.74 g/cm^3), nontoxic structural metal available for engineering applications. Most of the work to date on rapidly solidified magnesium alloys has focused on the improvement of strength levels and corrosion resistance [166], [174].

Practically all fine-grained beryllium is commercially produced by PM. Fine-grained beryllium powders produced by electrolysis of salt baths are vacuum hot pressed or hot isostatically pressed at ≤ 1150 °C. An ultimate tensile strength up to 600 MPa is associated with tensile fracture elongations below 6 %. Beryllium finds applications in the aircraft and space industry for stiff structural constructions, as heat sinks due to its high thermal conductivity and specific heat capacity, and as neutron moderators in nuclear power plants [175]. See also → Beryllium and Beryllium Compounds.

9.3. High-Temperature Nickel and Cobalt Superalloys

See also → Cobalt and Cobalt Compounds, Chap. 6.2. → High-Temperature Materials, Chap. 7.3.

The increase in strength exhibited by nickel- and cobalt-based superalloys is accompanied by a significant decrease in conventional hot workability which increases the overall component cost. Powder metallurgy techniques offer a means of producing these highly alloyed, high-temperature materials with a uniform, homogeneous microstructure that is often superplastic and, therefore, suitable for hot working [176]. Typical compositions of superalloys mainly used for components of military jet engines such as compressor disks, seals, high- and low-pressure turbine disks, and blade retainers are listed in Table 4.

The PM production sequence generally includes hot isostatic pressing of noble-gas-atomized powders into billets of 100 % density at 100 MPa and 1120 – 1200 °C. Consolidated billets are then hot forged slightly above 300 °C in several steps into jet engine parts [177]. Hot isostatically pressed and wrought PM materials show better mechanical properties than cast and wrought material and slightly improved economy due to higher material utilization and fewer forging steps. Still better economy is obtained by near net shape production via hot isostatic pressing alone; the reduction of the initial to final weight ratio is limited by the shape requirements for ultrasonic inspection of the near net shape part before final machining.

Efforts have been made to replace small, internally cooled, directionally-solidified cast blades in aircraft gas turbines by directionally recrystallized blades made of oxygen-dispersion-strengthened (ODS) superalloys (e.g., MA 6000 in Table 4). Mechanical testing indicates that the operating temperatures can be increased by 80 °C at equivalent stress rupture levels of 138 MPa as compared to a directionally solidified MAR – M 247 cast alloy [178].

Table 4. Composition of three PM superalloys (main elements)

Alloy	C	Cr	Mo	W	Ti	Nb	Co	Al	Y_2O_3	Ni
IN – 100	0.07	12.5	3.2		4.3		18.5	5.0		rem.
Rene 95	0.07	13.0	3.5	3.5	2.5	3.5	8.0	3.5		rem.
MA 6000	0.05	15.0	2.0	4.0	2.5			4.5	1.1	rem.

rem. = remaining.

Table 5. Soft and permanent magnetic sintered materials and composites [179]

Material type	Soft magnetic materials	Permanent magnets (energy density, kJ/m^3)	
		Classical	Modern
Metals	Fe, FeP, FeSi	steel (3) AlNiCo (40)	$SmCo_5$ (160)
	NiFe, CoFe	CrCoFe (40)	$Nd_2Fe_{14}B$ (240)
	NiFeMo	$Sm_2(Co, Cu, Fe, Zr)_{17}$ (200)	
Ceramics	MnZn ferrites		Ba ferrites (30)
	NiZn ferrites		Sr ferrites (30)
Composites	resin and inorganically bonded Fe, NiFe, SiFe powder (crystalline, amorphous) cores and parts	resin-bonded AlNiCo	resin-bonded ferrites, $SmCo_5$, $Nd_2Fe_{14}B$
		resin-bonded $Sm_2(Co, Cu, Fe, Zr)_{17}$	

9.4. Magnetic Materials

See also → Magnetic Materials.

The intrinsic magnetization of magnetic materials is determined by the magnetic moments of the atoms and their mode of alignment. A ferromagnetic or ferrimagnetic material is divided up into many small regions (Weiss domains) each of which is fully magnetized, i.e., all the moments within each domain are aligned in the same direction. The demagnetization and, therefore, the hysteresis of nearly all magnetic materials result from the nucleation and displacement of the Bloch walls (boundaries of finite thickness between Weiss domains with different directions of magnetization). Soft magnets and permanent magnets show a small and large hysteresis, respectively. Sintered magnetic materials are listed in Table 5.

Soft Magnets. In soft magnetic materials the motion of the Bloch walls with minimum friction is essential. Therefore the density of the effective centers pinning the Bloch walls should be as low as possible [179]. Pinning centers are most effective if their dimensions are in the order of the Bloch wall thickness, e.g., nonmagnetic inclusions, segregations, grain boundaries, or local lattice distortions (dislocation stress fields).

Soft magnetic PM materials must be single phase, of high purity, and large grain size. Any residual porosity should be present as large pores. In addition the magnetic crystal anisotropy and the product of magnetostriction (volume change associated with magnetization) and internal stresses should be kept low. The advantage of PM materials is based on their homogeneity, near net shaping, and the small amount of scrap. High-quality magnetic alloys for sintered parts may consist of Ni – Fe (36, 50, 80 wt % Ni), Fe – Si (3 wt % Si), Fe – P (0.3 – 0.8 wt % P), and Fe – Co – (V) (49 wt % Co, 2 wt % V) [179]. The porosities are below 5 %. The materials cover a range of coercitivities from 0.02 to 1.5 A/cm and saturation polarization up to 2.3 T.

Permanent Magnets. For permanent magnets high resistance against further change in the direction of magnetization is essential. In the "classical" permanent magnets (e.g., steel or AlNiCo) this was achieved by increasing the density of efficient pinning centers. In modern permanent magnets (e.g., $SmCo_5$) the extremely high magnetic crystal anisotropy constants are the basis of the superior permanent magnet properties.

The increased density of efficient pinning centers in classical permanent magnets results from high internal stresses (dislocation density) and often from a finely dispersed two-phase

microstructure. Unfortunately it is practically impossible to induce the magnetically desirable crystallographic texture by powder metallurgy; these structures are, however, readily obtained by directional solidification. The fraction of classical permanent magnets produced by PM is therefore decreasing. Oblong AlNiCo magnets have a permanent high-saturation polarization up to values of 1.2 T and an excellent temperature stability [179].

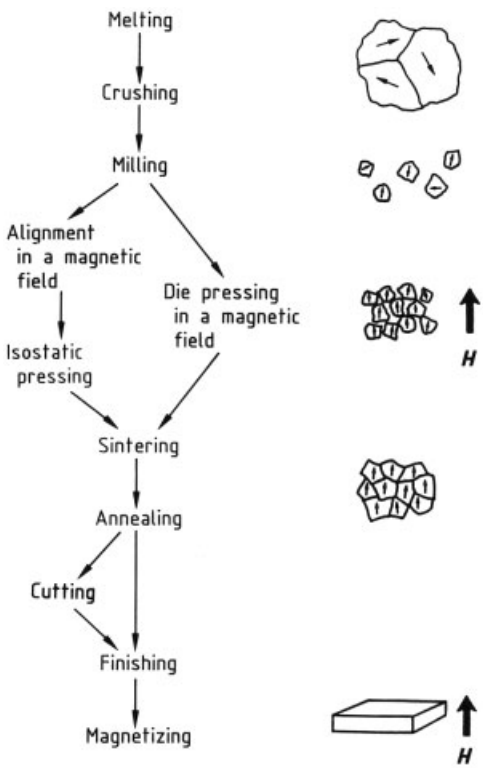

Figure 16. Processing flow chart of modern cut die-pressed permanent magnets [179]

Modern permanent magnets employ ceramic and metallic compounds with crystal anisotropy constants 100-fold higher than those of iron. The most important materials are barium-ferrites, $SmCo_5$ [180], and $Nd_2Fe_{14}B$ [181]. Other basic requirements (e.g., a high saturation polarization and a high Curie temperature) are satisfied mainly by the intermetallic rare earth – transition metal magnets. The basic microstructure of sintered magnets requires defect-free grains of the ferromagnetic phase to avoid local decrease of the anisotropy, the presence of impurities or nonferromagnetic phases at the grain boundaries to prevent magnetic domain walls and flux fields from crossing the boundaries, and well-aligned grains to optimize the total magnetic moment.

For processing cast polycrystalline rare earth – transition metal alloys are crushed and milled to a powder consisting of single crystals (particle size ca. 5 μm). The powders are aligned in a magnetic field, isostatically pressed, sintered, and mechanically cut or ground to customer specifications (Fig. 16). Near net shape magnets are manufactured by die pressing the powders under a strong directional magnetic field before sintering. The parts from both production routes are finally magnetized again either before or after mounting to their final device (e.g., small motors or switches).

Sintering to full density with maintenance of fine average grain size requires careful control of the sintering atmosphere and is often performed in the presence of small amounts of liquid phase (e.g., $Nd_2Fe_{14}B$).

The ferrites are low-cost permanent magnets with typical low saturation polarization values < 0.4 T and coercivities < 300 kA/m. The saturation polarizations of the expensive $SmCo_5$ and $Sm(Co, Cu, Fe, Zr)_{17}$ magnets are 0.9 and 1.15 T, respectively, with simultaneously high coercivities of 1500 – 2400 kA/m. Still higher saturation polarizations of 1.2 – 1.3 T but slightly lower coercivities and a lower Curie temperature (250 –310 °C) are typical for the Fe – Nd – B permanent magnets. The ability of the Sm – Co and Fe – Nd – B magnets to withstand even strong reversed magnetic fields without irreversible losses permits optimization of many devices and reduction of the costs of magnetic systems [179], [182].

Magnetic Composites. Soft magnetic powder composites consist of ferromagnetic particles, which are compacted, and molded. The particles are electrically insulated by an organic or inorganic binder (e.g., a polymer or ceramic). The composites have various advantages (e.g., high resistivity which reduces the losses by eddy currents and isotropic flux conducting properties). Limits are set by the small grain size and the temperature sensitivity of the binder materials. Permanent magnet composites are often

polymer bonded and offer the advantage of easy shaping. A severe drawback is the low magnetic energy density and permanent saturation polarization resulting from a low particle packing density and insufficient magnetic alignment of the particles [179].

9.5. Hardmetals and Cermets

Hardmetals. Hardmetals (cemented carbides, → Hard Materials) consist of a network of hard carbide particles (mainly WC, TiC, and TaC) embedded in a ductile metal matrix (usually Co) [44], [183]. Tungsten carbide or solid solutions of WC – TiC – TaC (for powder production see Section 2.2) are blended with cobalt powder by intensive ball milling for several days under an organic liquid (e.g., acetone, hexane).

A faster method for obtaining homogeneous distribution of the powder components is attrition milling. The addition of paraffin, granulation, or spray drying (Chap. 3) before pressing improves the filling behavior, the homogeneity of the green density, and the green strength of the powders.

Shaping and consolidation at pressures in the range 100 – 200 MPa in uniaxial (e.g., of inserts for metal cutting) or isostatic presses (large parts) or after plastification with an organic binder by extrusion (rods, tubes) yields green densities of ca. 50 % [45]. Dewaxing under humid hydrogen and sintering under dry hydrogen or vacuum at 1350 – 1550 °C are often performed in a single furnace without intermediate cooling. Powder injection molding (Section 5.3) is used for small parts with complex shapes [184]. Sinter – HIP (Section 7.1) yields products with special microstructures such as ultrafine grain sizes [185] or pore-free products from powders with low sinterability.

During sintering of WC and cobalt, a liquid phase forms at or slightly above the eutectic temperature of 1280 °C of the W – C – Co system. At the temperatures used for isothermal sintering (1350 – 1550 °C depending on the composition of the hardmetal) cobalt is completely molten and dissolves up to 25 wt % of the WC. The dissolution – reprecipitation processes described in Chapter 6 lead to densification and grain coarsening.

A major criterion for the quality of hardmetal tools is their tool life which is limited by wear and distortion due to mechanical and chemical interaction with the working pieces. The optimal cutting material has high warm strength, high toughness, and inertness against the working piece. The hardness, pressure strength, and wear resistance decrease with increasing cobalt content whereas the bending strength increases [45]. A reduced WC grain size in the sintered hardmetal improves the hardness and, with fine grain sizes ($<1.5\,\mu m$), the bending strength [185]. Coarse-grained material is used in mining operations. Extremely fine-grained materials are required for drilling in hybrid materials in the electronic industry. Hardmetals containing TiC and TaC are harder than WC – Co and less reactive with steel working pieces at higher cutting velocities. Most of the high-value cutting tool inserts are coated by chemical vapor deposition (CVD) or physical vapor deposition (PVD) with multilayers of TiC, TiN, and Al_2O_3 for improved tool life [185].

Cermets. Today cermets are understood to be hard materials based on multicomponent Ti(C, N) – Ni derivatives sintered in the presence of liquid phase [186]. The expression "Cermets" is an acronym of the words cer(amics) and met(al)s and was originally used to denote metal – ceramic composites. See also → Ceramics, Ceramic – Metal Systems, Chap. 6. Compared to WC – Co hard metals, cermets show enhanced chemical inertness against steel working pieces and reduced topochemical reactions. They allow faster cutting speeds and a reduced cutting wear rate due to higher hardness and cutting edge strength.

Advanced compositions contain molybdenum, tantalum, tungsten, vanadium, and nickel in addition to TiC, TiN, and nickel [187]. During liquid-phase sintering the hard material grains develop a core – mantle structure with a hard core and a tougher mantle [188]. Advanced cermets are beneficially PVD or CVD coated with protective layers.

9.6. Metal – Ceramic Composites

Metal/ceramic composites consist of (1) interpenetrating matrices of a metal and a ceramic

phase, (2) a high volume of ceramic inclusions in a metal matrix, or (3) metal inclusions in a ceramic matrix [189]. They are distinguished from metal – matrix or ceramic – matrix composites in that the main purpose of the ceramic or metal phase is not to improve the mechanical behavior of the base matrix material (as in SiC-reinforced aluminum alloys, hardmetals, or cermets) but to produce materials with specific properties such as thermal expansion or corrosion resistance. Often metal/ceramic and similar composites are manufactured from mixtures of metal and oxide, carbide or boride ceramics, diamond or glass powders by PM processing.

Oxide-based cermets composed primarily of U_3O_8 inclusions sintered in a matrix of aluminum, steel, or tungsten are used in the fuel elements of nuclear reactor cores. They provide better retention of the fission products than pure oxide material and reduce the danger of melting due to their improved thermal conductivity [190].

Excellent corrosion resistance against molten metals and high-temperature strength result in applications for Zr_2O – Mo and Zr_2O – Ti composites as protective tubes for thermocouples to measure melt bath temperatures and for extrusion dies for forming nonferrous metals. Al_2O_3 – Mo composites are used as seals, SiO_2 – (brass/bronze) composites are found as friction materials for heavy-duty brakes or clutches. Electrical contacts are produced on the basis of Ag – CdO and Ag – CrO_2 composites [191].

Boride-based metal – ceramic composites such as ZrB_2 – B are usually used because of their excellent high-temperature corrosion resistance against molten reactive metals.

A major group of metal-bonded composites are metal – diamond materials for grinding, lapping, sawing, cutting, dressing, and truing tools. The metallic matrix is frequently based on cobalt, more rarely on bronze or heavy-metal alloys [192].

Graphite-containing metallic friction materials for brake linings and clutch facings have a predominantly metallic matrix that exploits the high thermal conductivity for rapid energy adsorption. Graphite – copper composites are used for brushes in motors. For brushes with lower rubbing speed and low contact pressure graphite – silver composites are applied.

9.7. Electric Contact Materials

Transmission of electric power and information involves the passage of electric current through numerous devices and components that are joined to one another directly or via wires and leads. The path for power transmission from the generator to the consumer and of information from the sender to the receiver contains ca. 1000 contact points, of which ca. 100 are switching contacts. These switching contacts must switch the current on and off reliably, conduct the current without loss when switched on, and insulate the current paths from one another when switched off. The currents to be switched range from a few milliamps to ca. 100 000 A (e.g., in the case of short circuits at unfavorable points of the power supply network).

Contact components in power engineering generally consist of silver-based composite materials produced by PM. Even precious-metal alloys based on gold and platinum are used for contact components in communication and information technology. Composite materials produced by PM are used in numerous areas in electrical switching devices and switchgear [193–196]. Examples include high-voltage distribution switchgear, industrial and installation switchgear, switches for domestic equipments, switchgear for railways and electrically powered vehicles, and switches for motor vehicles.

Switching contacts have to satisfy many requirements and can be described on the basis of their operational conditions: make operation, current carrying, and break operation.

During *make operation* an arc may be formed after springback of the contacts, this is termed bouncing. As a result the contact surfaces heat up extremely and may weld together when they finally close. This welding tendency of the contact materials should be minimized.

In *current carrying* the contact surfaces only touch one another at a few points, resulting in an electrical contact resistance that causes heating of the contact components. The contact resistance and thus the heating effect should be minimized.

During *break operation* an arc is formed at the contact components, resulting in a high thermal load of the contact material surfaces. The

contact materials should have a low erosion rate and thus a long service life.

9.7.1. Contact Materials for Low-Voltage Switchgear

Most low-voltage switching devices operate in air. Silver is used as the base metal because it does not form nonconducting oxides on the contact surfaces and because it has a high electrical conductivity [195]. Silver is used together with metallic and nonmetallic components in the form of an alloy or composite material to obtain high mechanical and electrical resistance to wear, to suppress the welding tendency, and to assist arc migration and extinction. Powder metallurgy is very important because it is often the only way of producing such materials with their diverse combinations of properties. Worldwide consumption of silver for contact technology in the early 1990s was currently 1600 – 1900 t/a.

Contact materials for air-switching devices can be divided into four groups:

1) Silver and silver alloys
2) Silver – metal composites
3) Silver – metal oxide composites
4) Silver – graphite composites

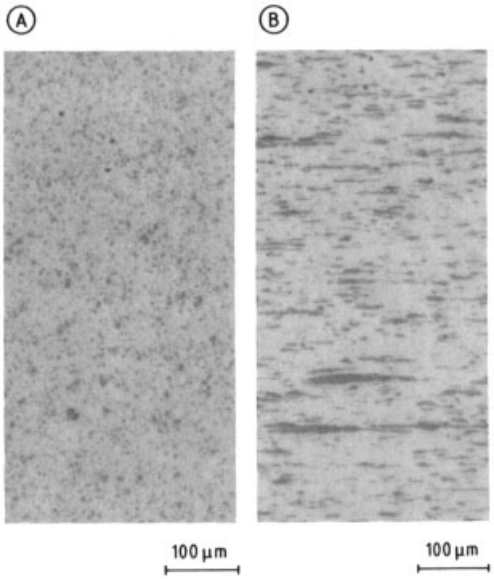

Figure 17. Transverse (A) and longitudinal (B) cross sections of a AgNi (90/10) composite material

Silver and silver alloys are only used to a small extent in power engineering. They have largely been replaced by particle-reinforced silver composites which have a lower tendency to weld and a longer service life.

Silver – metal composites are mainly produced by powder metallurgy. The most widely used material of this group is silver – nickel; its main advantages are its low contact resistance and simple processibility. It forms nonconducting erosion products, which means that switches made from this material for domestic appliances and process control have a small structural volume combined with a high dielectric strength. Figure 17 shows the fiber-like structure of an AgNi composite material formed by extruding sintered blocks.

Compared to silver – nickel, *silver – metal oxide composites* ($AgSnO_2$, AgCdO, and AgZnO) have a much higher welding resistance at high currents and a longer service life. They are therefore used in switches that have to withstand high current peaks during make operation (e.g., in engines, capacitors, and lamps).

Silver – tin oxide ($AgSnO_2$) is an important development in the field of contact materials [197] because it fulfils technical switching requirements and the need for more environmentally compatible materials with avoidance of toxic substances (cadmium oxide is carcinogenic).

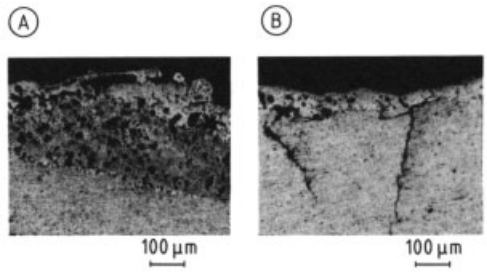

Figure 18. Transverse cross sections of $AgSnO_2$ (88/12) SPW VI (A) and $AgSnO_2$ (88/12) SPW VII (B) after switching on load

Silver – tin oxide is replacing silver – cadmium oxide (AgCdO) which has been successfully used since the early 1950s for applications in the current range from 30 A to several kiloamps. Silver – cadmium oxide exhibits a high resistance to welding, comparatively low heating in continuous operation, and a low tendency to

cut-out burning. These properties are also shared by $AgSnO_2$ materials manufactured by special processes involving special additives. In addition $AgSnO_2$ has a better wear resistance when high currents are switched. Figure 18 shows the structures of two types of $AgSnO_2$ after switching on load [198]. $AgSnO_2$VI (Fig. 18 A) is widely used at present, $AgSnO_2$ VII (Fig. 18 B) is a variant of $AgSnO_2$ VI; the materials differ as regards their additives. Figure 18 A shows $AgSnO_2$ SPW VI after switching loading. The material in the melt zone is permeated by many pores. The "spongy" structure weakens the surface of the material. In contrast, SPW VII (Fig. 18 B) exhibits a stable, "closed" structure in the melt zone, and only a small number of pores are formed.

Silver – graphite composites are also produced by PM. Of all silver-based materials silver – graphite is the most resistant to welding. Its most important field of application is therefore in automatic circuit breakers (e.g., in power switches and power circuit breakers) which must still be able to switch off reliably in the event of malfunction.

The contact materials behave very differently depending on the type of loading they are exposed to. Material tables roughly classify materials according to the type of load [199]. The optimum material combination for a specific application can, however, only be determined by comprehensive switching tests with the contactor in question [200].

The materials are classified in this section according to the criteria used in [199]: welding tendency during make operation and break erosion. The criteria make erosion, contact resistance, etc. are omitted from this classification.

If currents are to be reliably interrupted, for example with automatic circuit breakers, a material must be used that does not weld. Materials are listed below in order of increasing welding tendency.

AgC	3 – 5 % C
$AgSnO_2$	8 – 12 % SnO_2
AgCdO	10 – 15 % CdO
AgZnO	6 – 8 % ZnO
AgWNi	Ni content 10 – 40 %
AgCu	hard silver with 3 % Cu
AgNi	fine-grain silver with 0.15 % Ni
Cu	electrolytic copper
Ag	fine silver

The contact materials exhibit widely differing erosion behaviors in different current ranges. The materials listed below may be regarded as optimal for the indicated switching current ranges:

10 – 1000 A	AgNi
100 – 3000 A	AgCdO
200 – 5000 A	$AgSnO_2$
3000 – 5000 A	AgZnO
> 5000 A	AgW, AgWC

The limits between the current ranges are flexible, especially as the erosion behavior is influenced not only by the current but also by the constructional characteristics of the switching device.

The areas of use of the materials listed above apply to devices that perform frequent switching in air. Since graphite burns under these conditions, AgC has a limited service life. Therefore, AgC is only used in those devices which have a protective function and which must switch rarely occurring fault currents, for example in safety switches.

Production Technology. The PM production of contact materials can be effected in two ways: by extrusion and by individual sintering. Powder metallurgy has the advantage that a wide range of different additives can be combined.

Extrusion comprises the following stages: mixing of the powders, cold isostatic pressing (CIP) of the powders to form extrusion billets, sintering, and extrusion of the billets to form rectangular profiles or wires.

In *individual sintering*, mixing of the powders is also the first step. The individual contacts are then uniaxially pressed, sintered, and recompressed to obtain the required density.

Silver – metal oxide materials can also be produced by *internal oxidation.* Silver alloy billet is first produced by melting and casting. After extruding the billet to profiles or wires, these profiles are cut into contact pieces. These contacts are then annealed in an oxidizing atmosphere to selectively oxidize the alloying elements (e.g., Sn).

9.7.2. Contact Materials for High-Voltage Switchgear

A further group of contact materials are WCu or WAg composites. These materials are used in

high-voltage switches under oil, under SF_6, and in air with currents up to 100 kA. They are usually produced by impregnation methods. High-melting tungsten powder is sintered to form a porous, spongy structure. The pores are filled with molten copper or silver by infiltration techniques. The silver or copper is often alloyed with iron, nickel, or cobalt to improve the wetting of the tungsten structure.

The last group to be discussed comprises contact materials for vacuum switches. The contacts switch throughout their service life under a vacuum $< 10^{-4}$ mbar. One of the advantages of vacuum switches is that they can be used in atmospheres where there is a risk of explosion.

CuCr materials for use in vacuum power switches are most important, particularly in Europe. In materials with a chromium content > 45 wt %, a skeleton is first sintered from chromium powder and then infiltrated with the lower melting component. Chromium contents of 25 – 40 wt % guarantee a satisfactory combination of high welding resistance, high breaking capacity, and low contact heating. These materials are produced by pressing and sintering a mixture of copper and chromium powders. Other switching properties can be improved by using small amounts of additives.

10. Parts with Inherent, Functional Porosity

For a number of products the inherent porosity which results from PM processing route serves a function and is therefore intentionally created and preserved. The most important product groups are filters, bearings, electrodes for biomedical applications, and electrode supports for batteries and storage batteries.

Sintered metal filters are applied for filtering, flame stops, and sound absorption. Most frequently powders from bronze, stainless steel, nickel, titanium, tungsten, tantalum alloys are used. Defined porosities and permeabilities are obtained by the use of spherical initial powders with narrow size distribution. Sintered metal fiber nets also provide filters with effective channel diameters of 0.5 – 200 µm [201].

Sintered self-lubricating bearings are a classical application of powder metallurgy (Section 8.2). The lubricants are stored in the open pores (porosity 15 – 25 %) of the bearings. During rotation of the shaft the lubricant forms a wedge-shaped reservoir and a continuous lubricating film between shaft and bearing. When rotation stops most of the lubricant is withdrawn into the pore space. Due to this reversible process self-lubricating PM bearings require little, if any, maintenance. Materials for bearings are mainly bronze or brass on copper basis but also Fe – Cu – C and Fe – Ni – Cr steels for higher load-bearing applications in the electrical engineering and automotive industries [202].

Sintered electrode plates with porosities up to 80 % are used as electrically conductive frames for the electrode material (e.g., Ni frames for nickel hydroxide in Ni – Cd storage batteries [203]).

11. References

1. E. Klar (ed.): "Powder Metallurgy," *Metals Handbook,* 9th ed., **vol. 7,** American Society for Metals, Metals Park, Ohio 44073, 1984.
2. W. Schatt: *Pulvermetallurgie,* Hüthig Verlag, Heidelberg 1986.
3. R. M. German: *Powder Metallurgy Science,* "Metal Powder Industries Federation," Princeton, N.J., 1984.
4. R. Oberacker, F. Thümmler: "Produkte der Pulvermetallurgie" in *Chemische Technology,* 4th ed., vol. 4, Hanser Verlag, München 1986.
5. W. A. Kaysser: "Sintern mit Zusätzen," in G. Petzow, F. Jeglitsch (eds.): *Materialkundlich-Technische Reihe,* vol. 11, Gebrüder Bornträger, Berlin 1992.
6. G. Zapf: "Optimum Materials and Energy Utilization in the Powder Metallurgy Production of Automotive Structural Parts," *Powder Metall. Int.* **10** (1978) 143.
7. K. Schekulin: "Sinterbauteile im Wettbewerb mit sonstigen spanlos genau geformten Bauteilen," *Powder Metall. Int.* **14** (1982).
8. L. F. Pease III, R. J. Sansoucy (eds.): "Powder Injection Molding," *Advances in Powder Metallurgy,* vol. 2, Metal Powder Industries Federation, Princeton, N.J., 1991.
9. R. M. German: *Powder Injection Molding,* Metal Powder Industries Federation, Princeton, N.J., 1990.
10. J. Heyes, R. G. R. Sellors: "Industrial Applications of Refractory Metals" in I. Jenkins, J. V. Wood: *Powder Metallurgy – an*

Overview, The Institute of Metals, London 1991, pp. 218 – 331.
11. R. Kieffer, F. Benesovky: *Hartmetalle,* Springer Verlag, Wien 1965.
12. A. Bose, S. C. Yang, R. M. German, "Development of a New W – Ni – Mn Heavy Alloy" in L. F. Pease III, R. J. Sansoucy (eds.): *Advances in Powder Metallurgy,* vol. 6, Metal Powder Industries Federation, Princeton, N.J., 1991, pp. 425 – 437.
13. H. Nosbusch, I. V. Mitchell (eds.): *Technical Ceramics,* Elsevier Applied Science, London 1988.
14. R. J. Schaefer, M. Linzer (eds.): *Hot Isostatic Pressing* – Theory and Applications, ASM International Materials Park, Ohio 1991.
15. K. Torsell: "HIP Steel Components for the Manufacturing Industry," in M. Koizumi (ed.): *Hot Isostatic Pressing – Theory and Applications,* Elsevier Applied Science, London 1992, pp. 195 – 208.
16. P. Beiss: "Bauteile mit gesteuerter Porosität," *Powder Metall. Int.* **15** (1983) 2.
17. T. G. Mulcavage: "Electronic/Hydraulic Control for a Multi-platen Compacting System on a Mechanical Press," in L. F. Pease III, R. J. Sansoucy (eds.): *Advances in Powder Metallurgy,* vol. 1, Metal Powder Industries Federation, Princeton, N.J., 1991, p. 23.
18. P. R. Wilyman, M. Vandermeiren: "Selection of Atmospheres for Sintering," in I. Jenkins, J. V. Wood (eds.): *Powder Metallurgy – an Overview,* The Institute of Metals, London 1991, pp. 208 – 218.
19. H. Kolaska, K. Dreyer: "Hartmetalle, Cermets und Keramiken als verschleißbeständige Werkstoffe," *Metall (Berlin)* **45** (1991) 224 – 236.
20. D. G. White: "The Growing Internationalization of Powder Metallurgy" in P. U. Gummeson, D. A. Gustavson (eds.): *Modern Developments in Powder Metallurgy,* MPIF/APMI Princeton, N.J., 1988, pp. 1 – 8.
21. K. H. Roll, "P/M Global Perspective – North America" in E. N. Aqua, C. I. Whitman (eds.): *Modern Developments in Powder Metallurgy,* MPIF/AMPI Princeton, N.J., 1984, 1 – 11.
22. M. Wahlster, H. Stephan, R. Ruthardt: "Cost Estimates and Economic Considerations in the Production Processing of High Quality Powders," *Powder Metall. Int.* **12** (1980) 173.
23. S. Takajo, "Innovations in Ferrous Powders and their Production" in W. J. Huppmann, W. A. Kaysser and G. Petzow (eds.): *Powder Metallurgy – State of the Art,* Schmid-Verlag, Freiburg 1986, pp. 7 –28.
24. P. Hellmann, P. Billgren: "Commercial Gas Atomization" in A. Lawley, E. Klar (eds.): *Atomization and Processes: Current and Future,* Metal Powder Industries Federation, Princeton, N.J., 1984, pp. 33 –50.
25. E. Klar (ed.): "Atomization," in *Metals Handbook,* 9th ed., vol. 7, American Society for Metals, Metals Park 1984, pp. 25 – 51.
26. J. A. Tallmadge: "Powder Production by Gas and Water Atomization of Liquid Metals," in H. A. Kuhn and A. Lawley (eds.): *Powder Metallurgy Processing,* Academic Press, New York 1978, pp. 1 –32.
27. M. Bürger, E. von Berg, S. H. Cho, A. Schatz: "Analysis of Fragmentation Processes in Gas and Water Atomization Plants for Process Optimization Purposes," *Powder Metall. Int.* **21** (1989) 6; **24** (1992) 10.
28. H. Lubanska, "Correlation of Spray Ring Data for Gas Atomization of Liquid Metals," *J. Met.* **22** (1970) 45 – 49.
29. W. A. Kaysser, K. Rzesnitzek, "Principles of Atomization" in D. P. Uskokovic, H. Palmour III, R. M. Spriggs (eds.): *Science of Sintering,* Plenum Press, New York 1989, pp. 157 – 176.
30. T. W. Clyne, R. A. Ricks, P. J. Goodhew: "The Production of Rapidly – Solidified Aluminum Powder by Ultrasonic Gas Atomization," *Int. J. Rapid Solidif.* **1** (1984) 59 – 80.
31. P. R. Roberts: "Rotating Electrode Process," in E. Klar (ed.): *Metals Handbook,* 9th ed., vol. 7, American Society for Metals, Metals Park 1984, pp. 39 – 41.
32. J. J. Gillvary, B. M. Borgstrom: "Fracture and Comminution of Brittle Solidas," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **220** (1961) 380.
33. W. E. Lucky, H. Lucky: "Characterization of Ball Milled Type 316 L Stainless Steel Powder," in W. Kun, W. Ehretsman (eds.): *Fine Particles,* The Electrochemical Society Inc. (1974) pp. 95 – 121.
34. J. S. Benjamin, M. J. Bonaford, *Metall. Trans. A.* **8 A** (1977) 1301.
35. G. Jangg, F. Kutner, G. Korb, *Powder Metall Int.* **9** (1977) 24 – 26.
36. W. D. Jones: *Fundamental Principles of Powder Metallurgy,* E. Arnolds, London 1960, p. 128.
37. R. Walker, A. R. B. Sandford, *Chem. Ind. (London)* 1979, 624.
38. E. Peissker, *Z. Werkstofftech.* **6** (1975) 58.
39. R. Walker, *Chem. Ind. (London)* 1980 260.
40. In [2]58.

41. B. Meddings: "Production of Nickel Powder by Hydrometallurgical Processing," in [1]138.
42. M. Wiberg, *Jernkontorets Ann.* **142** (1958) 289.
43. P. Taskinen, M. H. Tikannen, *Scand. J. Metall* **6** (1977) 223.
44. R. Kieffer, F. Benesovsky (eds.): *Hartmetalle,* Springer Verlag, Wien 1965.
45. R. Glätzle, in F. Benesovsky (ed.): *Pulvermetallurgie und Sinterwerkstoffe,* Metallwerk Plansee, Reutte, 1980, pp. 121.
46. L. Nyborg, I. Olefjord: "Surface Analysis of REP-atomized Martensitic Steel Powder," *Powder Metall. Int.* **20** (1988) no. 2, 11.
47. J. K. Beddow: *Particulate Science and Technology,* Chemical Publishers, New York 1980. J. K. Beddow in [1] p. 241.
48. D. S. Parsons, in [1]pp. 230.
49. E. Robens: "The Measurement of Specific Surface Area and Pore Size Distribution of Powders," *Powder Metall. Int.* **18** (1986) 12.
50. B. Frisch, W. R. Thiele: "A Measuring Method for the Determination of Density, Specific Surface Area and Porosity of Powders and Compacted Bodies," *Powder Metall. Int.* **18** (1986) 17.
51. R. H. Berg, in [1] p. 220.
52. T. Allen: *Particle Size Measurement,* 3rd ed., Chapman and Hall, London.
53. H. A. Kupfer: "Automatic Particle Size Distribution Analysis of Powdered Material by the Light Blockage Prinziple," *Powder Metall.* **10** (1979) 96 – 97.
54. H. N. Frock, in [1]p. 216.
55. A. Adler: "Flow Properties of Metal Powders," *Int. J. Powder Metall.* **5** (1969) no. 1, 7 – 20.
56. G. Mattei, N. Claussen, H. H. Hausner: "Influence of Relative Humidity on Flow of Metal and Ceramic Powders," *Mod. Dev. Powder Metall.* **8** (1974) 5 – 11.
57. S. Tepper: *Pulvermetallurgie* – Metallpulver, Sintermetalle, Hartmetalle, DIN-Taschenbuch, Beuth Verlag, Köln 1991, pp. 182 – 184.
58. In [2]p. 108 .
59. R. M. German: *Particle Packing Characteristics,* Metal Powder Industries Federation, Princeton, N.J., 1989.
60. E. Klar: "Chemical Methods of Powder Production," in [1] p. 53.
61. E. Ernst et al.: "Friction Measurements During Powder Compaction," *Powder Metall. Int.* **23** (1991) 77.
62. S. Gasiorek, K. Z. Korczak: "Cohesiveness of a Compact" in L. F. Pease III, R. J. Sansoucy (eds.): *Advances in Powder Metallurgy,* vol. 1, Metal Powder Industries Federation, Princeton 1991 p. 287.
63. B. K. Lograsso, R. M. German: "Thermal Debinding of Injection Molded Powder Compacts," *Powder Metall. Int.* **22** (1990) 17 – 22.
64. E. Lange, N. Müller: "P/M Injection Molding Technique for Ceramic and Metal Parts," *Powder Metall. Int.* **18** (1986) 416.
65. R. J. Causton, W. B. James, J. J. Fulmer: "Performance Characteristics of New Sinter-Hardening Low-Alloy Steel" in L. F. Pease III, R. J. Sansoucy (eds.): *Advances in Powder Metallurgy,* vol. 5, Metal Powder Industries Federation, Princeton, N.J., 1991, pp. 91 – 104.
66. W. Pietsch, H. Rumpf: "Nährungsrechnung der Kapillarkräfte zwischen Kugeln," *Chem. Ing. Techn.* **39** (1967) 885 – 893.
67. N. Claussen, G. Petzow: "Preparation of Small Spheres by the Dry Agglomeration of Powders," in *Powder Metal Conference Supplement,* Part 1, 1971, pp. 225 – 248.
68. K. Masters: *Spray Drying Handbook,* 3rd ed., J. Wiley & Sons, New York 1979.
69. R. P. Seelig, J. Wulff: "The Pressing Operation in the Fabrication of Articles by Powder Metallurgy," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **166** (1946) 492 – 504.
70. E. Arzt: "The Influence of an Increasing Particle Coordination on the Densification of Spherical Powders," *Acta Metall.* **30** (1982) 1883 – 1890.
71. G. Zapf in G. Spur (ed.): *Handbuch der Fertigungstechnik,* **vol. 1,** Hanser, München 1980, p. 824.
72. L. F. Pease III, R. J. Sansoucy (eds.): "Compaction Quality Control and Training," *Advances in Powder Metallurgy,* vol. 4, Metal Powder Industries Federation, Princeton, N.J. 1991.
73. G. Petrus, T. M. Krauss, B. L. Ferguson: "Die Compaction Process Design," part 1, "Knowledgebased Design," in [72]pp. 101 – 110.
74. G. Petrus, T. M. Krauss, B. L. Ferguson: "Die Compaction Process Design," part 2, "FEM Simulation," in [72]pp. 111 – 122.
75. E. Ernst et al.: "Friction Measurements during Powder Compaction," *Powder Metall. Int.* **23** (1991) no. 2, 1 – 8.
76. T. G. Mulcavage: "Electronic/Hydraulic Control for Multi-platen Compacting System on a Mechanical Press," [72]pp. 23 – 46.

77. P. Lindskog: "Das kaltisostatische Pressen bei der Formgebung metallischer Pulver – wet and dry bag-Verfahren" in H. Kolaska (ed.): *Fortschritte bei der Formgebung in Pulvermetallurgie und Keramik,* vol. 7, VDI-Verlag, Düsseldorf 1991, pp. 141 –167.
78. V. T. Morgan, R. L. Sands: "Isostatic Compaction of Metal Powders, Metallurgical Reviews," *J. Inst. Met.* **134** (1969).
79. G. Naeser, F. Zirm, *Stahl Eisen* **70** (1950) 995 –1003.
80. M. H. D. Blore et al.: "Pure Nickel Strip by Powder Rolling," *Met. Eng. Q.* **6** (1966) 54 – 60. R. K. Dube, *Powder Metall. Int.* **15** (1983) 36.
81. B. L. Mordike, J. Kaczmar, M. Kielbinski, K. U. Kainer: "Cold Extruded Cu – W Composite Materials," *Powder Metall. Int.* **23** (1991) 91.
82. H. H. Hausner (ed.): *Perspectives in Powder Metallurgy,* **vol. 1,** Plenum Press, New York 1967, p. 221.
83. S. Clyens, W. Johnson, *Mater. Sci. Eng.* **30** (1977) 121.
84. W. A. Kaysser: "Solid State Sintering" in I. Jenkins, J. Wood (eds.): *The Institute of Metals Series on Powder Metallurgy, Powder Metallurgy – an Overview,* The Institute of Metals, London 1991, pp. 168 –182.
85. W. A. Kaysser: "Liquid Phase Sintering" in I. Jenkins, J. Wood (eds.): *The Institute of Metals Series on Powder Metallurgy, Powder Metallurgy – An Overview,* The Institute of Metals, London 1991, pp. 183 – 197.
86. G. C. Kuczynski, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **185** (1949) 169.
87. M. Mitkov, M. Aslan, W. A. Kaysser, *Powder Metall. Int.* **21** (1989) 7.
88. G. C. Kuczynski, *Z. Metallkd.* **67** (1976) 606.
89. R. L. Coble, in *Proc. Second Conference on Sintering and Related Phenomena,* Notre Dame 1965, Gordon & Breach, New York (1967) p. 423.
90. R. J. Brook, *J. Am. Ceram. Soc.* **52** (1969) 56.
91. C. H. Hsueh, A. G. Evans, R. L. Coble, *Acta Metall.* **30** (1982) 1269.
92. D. S. Wilkensen, M. F. Ashby, *Acta Metall.* **23** (1975) 1277.
93. B. H. Alexander, R. W. Balluffi, *Acta Metall.* **5** (1957) 666.
94. H. E. Exner: "Grundlagen von Sintervorgängen," in G. Petzow (ed.): *Materialkundliche Stuttgarter Technische Reihe,* Gebr. Borntraeger, Berlin, 1978, p. 4.
95. J. Frenkel, *J. Phys. (Moscow)* **9** (1945) 385.
96. W. D. Kingery, M. Berg, *J. Appl. Phys.* **26** (1955) 1205.
97. G. R. Rockland, *Z. Metallkd.* **58** (1969) 476.
98. R. L. Coble, *J. Appl. Phys.* **32** (1961) 787.
99. K. Uematsu et al. in H. Saito, S. Somiya (eds.): *Int. Symposium of Factors in Densification of Oxide and Non-Oxide Ceramics,* Gakujutsu Bunken Fukyu-Kai, Tokyo, 1978, p. 190.
100. M. F. Ashby, *Acta Metall.* **22** (1974) 275.
101. V. Smolej, *Z. Metallkd.* **74** (1983) 689.
102. W. A. Kaysser: "Ceramic Transactions," in G. L. Messing, E. R. Fuller, H. Hausner (eds.): *Ceramic Powder Science,* vol. 2, Am. Ceram. Society, Westerville, OH, 1988, p. 955.
103. W. A. Kaysser, M. Hofmann-Amtenbrink, G. Petzow, *Powder Metall.* **8** (1985) 199.
104. J. H. Brophy, L. A. Shepard, J. Wulff in W. Leszynski (ed.): *Powder Metallurgy,* AIME-MPI, Interscience, New York 1961, p. 113.
105. I. J. Toth, N. A. Lockington, *J. Less Common. Met.* **12** (1967) 353.
106. R. M. German, Z. A. Munir in *Reviews on Powder Metallurgy and Physical Ceramics,* Publishing House Freund, Saint-Saphorin, 1982, pp. 2, 9.
107. F. A. Kröger, V. J. Vink in F. Seitz, D. Turnbull (eds.): *Solid State Physics,* **vol. 3,** Academic Press, New York 1966, 307.
108. W. J. Huppmann, H. Riegger: "Modelling of Rearrangement Processes in Liquid Phase Sintering," *Acta Metall.* **23** (1975) 965 – 971.
109. K. S. Hwang, R. M. German, F. V. Lenel: "Capillary Forces between Spheres During Agglomeration and Liquid Phase Sintering," *Metall. Trans. A.* **18 A** (1984) 11 – 17.
110. H. H. Park, O. J. Kwon, D. N. Yoon: "The Critical Grain Size for Liquid Flow into Press during Liquid Phase Sintering," *Metall. Trans. A* **17 A** (1986) 325 –330.
111. W. A. Kaysser, G. Petzow: "Geometry Models for the Elimination of Pores during Liquid Phase Sintering in Systems with Incomplete Wetting," *Sci. Sintering* **16** (1984) 157 – 175.
112. W. A. Kaysser, G. Petzow: "Present State of Liquid Phase Sintering," *Powder Metall.* **28** (1985) 145 – 150.
113. W. D. Kingery: "Densification During Sintering in the Presence of a Liquid Phase I," *J. Appl. Phys.* **30** (1959) 301 – 309.
114. G. H. S. Price, C. J. Smithells, S. V. Williams: "Copper-Nickel-Tungsten Alloy Sintered with a Liquid Phase Present," *J. Inst. Met.* **62** (1938) 239 – 243.

115. D. N. Yoon, W. J. Huppmann: "Grain Growth and Densification During Liquid Phase Sintering of W – Ni," *Acta Metall.* **27** (1979) 693 – 698.
116. W. A. Kaysser, S. Takajo, G. Petzow: "Particle Growth by Coalescence during Liquid Phase Sintering of Large Particles," *Powder Metall. Int.* **14** (1981) 126 – 128.
117. W. A. Kaysser, G. Petzow: "Ostwald Ripening and Shrinkage during Liquid Phase Sintering," *Z. Metall.* **76** (1985) 687 – 692.
118. M. Hillert, G. Purdy: "Chemically Induced Grain Boundary Migration," *Acta Metall.* **26** (1978) 333 –340.
119. D. N. Yoon et al. : *Coherency Strain Induced Migration of Liquid Films through Slids, Interface Migration and Control of Microstructure,* NBS, Washington 1985.
120. C. A. Handwerker, J. W. Cahn, D. N. Yoon, J. E. Blendell in G. E. Murch, M. A. Dayananda (eds.): *The Effect of Coherency Strains on Alloy Formation, Migration of Liquid Films, Atomic Transport in Alloys: Recent Developments,* TMS/AIME Publications, Dayton 1985.
121. W. A. Kaysser, W. J. Huppmann, G. Petzow: "Analysis of Dimensional Changes during Sintering of Fe – Cu," *Powder Metall.* **23** (1980) 86 – 91.
122. V. Gemenetzis, R. Gräbner, R. Thümmler, *Z. Werkstofftech.* **10** (1979) 436.
123. D. Warga: "Controlled Furnace Atmospheres for Sintering," *Powder Metall. Int.* **23** (1991) 311.
124. C. G. Goetzel, in [1]pp. 501 – 521.
125. M. Koizumi (ed.): *Hot Isostatic Pressing* – Theory and Applications, Elsevier Applied Science, London 1992.
126. R. Widmer, in [125] pp. 3 – 10.
127. A. Frisch, W. A. Kaysser, G. Petzow: "Densification Maps and Healing Mechanisms During Sinter/HIP of Polyphase Materials," in T. G. Gasbarre, W. F. Jandeska Jr. (eds.): *1989 Advances in Powder Metallurgy,* MPIF/APMI, Princeton, N.J., 1989, pp. 431 –444.
128. M. F. Ashby: "The Role of Modelling in the Control of Hot Isostatic Pressing" in *Proceeding of International Conference on Hot Isostatic Pressing of Materials*: Applications and Developments, Antwerpen, 25. – 27. April 1988, Royal Flemish Society of Engineers, 1. 1. – 1. 9. 1988.
129. W. A. Kaysser: "Present State of Modelling of Hot Isostatic Pressing" in R. Schaefer, M. Linzer (eds.): *Hot Isostatic Pressing – Theory and Applications,* ASM International, Materials Park, Ohio, 1991, pp. 1 – 13.
130. T. Soh, A. Nohara, T. Nakagawa: "HIP Process Simulation by the Finite Element Method" in T. Garvare (ed): *HIP Proc. International Conference on Hot Isostatic Pressing-Theories and Applications,* CENTEK Publishers, Lulea, Sweden, 1988, pp. 81 –88.
131. J. Besson, M. Abouaf: "Numerical Simulation of Hot Isostatic Pressing of Ceramic Powders" in *Proceedings of the International Conference on Hot Isostatic Pressing of Materials: Applications and Developments,* Antwerpen, April 25–27 1988, Royal Flemish Society of Engineers 1988, 1.17 – 1.22.
132. D. Bouvard, M. Lafer: "Determination of the Densification Kinetics of Metal Powders by Interrupted Hot Isostatic Pressing Tests," *Powder Metall Int.* **21** (1989) no. 4, 11.
133. W. J. Huppmann, G. T. Brown: "The Steel Powder – Foraging Process – a General Review," *Powder Metall.* **10** (1978) 105 – 114.
134. W. J. Huppmann, M. Hirschvogel: "Powder Forging," *Int. Met. Rev.* **no. 233, vol. 23** (1978) 5, 209 –239.
135. H. Fischmeister: "Isostatic Hot Compation – a Review," *Powder Metall. Int.* **10** (1978) 119 – 123.
136. R. F. Singer, M. J. Couper: "Powder Metallurgy in Science and Practical Technology," in W. J. Huppmann, W. A. Kaysser, G. Petzow (eds.): *Powder Metallurgy 1986 – State of the Art,* vol. 2, Verlag Schmid, Freiburg 1986, pp. 177 – 204.
137. F. V. Lenel: *Powder Metallurgy* – Principles and Applications, Metal Powder Industries Federation, Princeton, New Jersey 1980, 401 – 444.
138. H. H. Hausner: *Handbook of Powder Metallurgy,* Chemical Publishing Corporation, New York 1973.
139. Werkstoff Leistungsblätter, in [57]pp. 41 – 48.
140. A. Squire, "Density Relationship of Iron Powder Compacts," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **171** (1947) 485 – 503.
141. H. E. Exner, D. Pohl: "Fracture Behavior of Sintered Iron," *Powder Metall. Int.* **10** (1978) 193.
142. M. Forskitt, P. A. Brook, J. R. Moon: "The Effect of Non-Circular Holes on the Elastic Properties of Materials," *Powder Metall. Int.* **19** (1987) 19.
143. U. Engström, C. Lindberg, J. Tengzelius: "Powders and Processes for High-Performance

PM Steels," *Powders Metall.* **35** (1992) 67 – 72.
144. A. Salak, V. Benke: "Powder-Forged Roller Bearing Races from Fe – Ni – Mo and Fe – Mn – Cr-Powders," *Powder Metall. Int.* **22** (1990) no. 5, 9.
145. N. Dautzenberg, H. J. Dörrweiler: "Maßverhalten von Kupfer-Kohlenstoff-Sinterstählen," *Tagungsbände der VII. Internationalen Pulvermetallurgischen Tagung in der DDR* **3** (1981) 177 – 189.
146. S. S. Kang, D. N. Yoon: "Effect of Cooling Rate on the Strength of Fe – Cu Compacts," *Powder Metall.* **20** (1977) 70 – 73.
147. V. B. Phadke: "Precipitation Hardening in Furnace-Cooled Sintered Fe – Cu Alloys," *Powder Metall.* **24** (1981) 25 – 31.
148. D. Peckner: *Handbook of Stainless Steels,* McGrawHill, New York 1977.
149. P. Beiss: "Processing of Stainless Steel Parts," *Powder Metall.* **34** (1991) 259 – 261.
150. P. Hellmann, H. Larker, J. B. Pfeiffer, I. Stromblad: "ASEA/Stora Process, a New Process for the Manufacture of Tool Steel and Other Alloy Steels from Powder" in H. H. Hausner (ed.): *Modern Developments in Powder Metallurgy,* vol. 4, MPIF-APMI, Princeton 1971, 573 – 582.
151. D. Whittaker: "Technical Advances, Applications and Business Outlook for Metal Injection Moulding," *Powder Metall.* **35** (1992) 8 – 10.
152. P. Beiss: "Wärmebehandlung und Eigenschaften von pulvermetallurgischen Werkzeugstählen" in H. Kolaska (ed.): *Konsolidierung und Wärmebehandlung von Sinterwerkstoffen, Pulvermetallurgie in Wissenschaft und Praxis,* vol. 6, Verlag Schmid, Freiburg 1990, 173 – 202.
153. R. H. Palma, V. Martinez, J. J. Urcola: "Sintering Behaviour of T 42 Water Atomized High Speed Steel Powder under Vacuum and Industrial Atmospheres with Free Carbon Additions," *Powder Metall.* **32** (1989) 291 – 299.
154. V. Arnhold, D. Duda, R. Wähling, "Schneidwerkzeuge aus pulvermetallurgischen Schnellarbeitsstählen" in H. Kolaska (ed.): *Schneidstoffe, Pulvermetallurgie in Wissenschaft und Praxis,* vol. 4, Verlag Schmid, Freiburg 1988, pp. 75 – 114.
155. F. Thümmler, C. Gutsfeld: "Sintered Steels with High Content of Hard Phases: a New Class of Wear Resistant Materials," *Powder Metall. Int.* **23** (1991) 285.
156. T. Bell: "Surface Treatment and Coating of PM Components," *Powder Metall.* **34** (1991) 253 – 258.
157. P. Beiss: "Finishing Processes in Powder Metallurgy," *Powder Metall.* **32** (1989) 277 – 284.
158. P. J. James: "The Machinability of Sintered Steels" *Powder Metall. Int.* **22** (1990) no. 6, 19.
159. P. W. Taubenblat: "Importance of Copper in PM," *Int. J. Powder Metall. Powder Technol.* **15** (1979) 169 – 184.
160. H. H. Hirsch: "Improving Conductivity of Copper PM Parts by Pretreatment of Green Compacts," *Powder Metall.* **22** (1979) 49 – 61.
161. K. Sedlatschek: "Eigenschaften und Anwendungen hochschmelzender Metalle," in F. Benesovsky (ed.): *Pulvermetallurgie und Sinterwerkstoffe,* Schriftenreihe für Mitarbeiter der Metallwerk Plansee Aktiengesellschaft, Reutte 1980, 104 – 120.
162. W. D. Coolidge, *Trans. Am. Inst. Elec. Eng.* **19 part II,** 1910, 961 – 965.
163. R. Eck: "Die Technologie der hochschmelzenden Metalle" in F. Benesovsky (ed.): *Pulvermetallurgie und Sinterwerkstoffe,* Schriftenreihe für Mitarbeiter der Metallwerk Plansee Aktiengesellschaft, Reutte 1980, pp. 76 – 103.
164. G. Jangg, R. Kieffer, B. Childeric, E. Ertl, *Planseeber. Pulvermetall.* **22** (1974) 15 – 28.
165. M. J. Couper, J. W. Luster, M. Thumann: "Development of Elevated Temperature Powder Metallurgy Aluminium Alloys," *Powder Metall. Int.* **23** (1991) 7 – 15.
166. F. H. Froes, Y. W. Kim, S. Krishnamurthy, R. Sundaresan: "Current Status of PM Technology of Light Metals" in I. Jenkins, J. V. Wood (eds.): *Powder Metallurgy – An Overview,* The Institute of Metals, London 1992, pp. 220 – 253.
167. R. F. Singer, M. J. Couper: "Powder Metallurgy in Science and Practical Technology," in W. J. Huppmann, W. A. Kaysser, G. Petzow (eds.): *Powder Metallurgy 1986 – State of the Art,* vol. 2, Verlag Schmid, Freiburg 1986, pp. 177 – 204.
168. G. J. Hildeman, L. C. Labarre, A. Hafeez, L. M. Angers in G. J. Hildeman, M. J. Koczak (eds.): *High Strength Powder Metallurgy Aluminum Alloys II,* TMS, Warrendale, PA, 1986, 25 – 43.

169. D. J. Skinner, P. A. Chipko, K. Okazaki, US 4 743 317, 1988.
170. I. G. Palmer, M. P. Thomas, G. J. Marshall: "Development of Thermally Stable Al – Cr – Zr Alloys Using Rapid Solidification Technology" in Y. W. Kim, W. M. Griffith (eds.): *Dispersion Strenghtened Aluminium Alloys,* TMS, Warrendale 1988, pp. 217 – 241.
171. J. S. Benjamin, M. J. Bonaford, *Metall. Trans. A* **8 A** (1977) 1301.
172. G. Jangg, F. Kutner, G. Korb, *Powder Metall. Int.* **9** (1977) 24 – 26.
173. T. G. Nieh, C. A. Henshal, J. Wadsorth, *Scripta Metall.* **18** (1984) 1405.
174. F. Hehmann, F. Sommer, H. Jones in F. H. Froes, S. J. Savage (eds.): *Processing of Structural Metals by Rapid Solidification,* ASM International, Metals Park, Ohio, 1987, pp. 379 – 398.
175. G. J. London, G. H. Keith, N. P. Pinto: "Grain Size and Oxide Content affect Beryllium Properties," *Met. Eng. Q.* **16** (1976) 45 – 57.
176. B. L. Ferguson: "Aerospace Applications," in E. Klar (ed.): *Metals Handbook,* 9th ed., vol. 7, American Society for Metals, Metals Park, Ohio, 1984, pp. 646 – 656.
177. V. A. Tracey, C. P. Cutter, *Powder Metall.* **24** (1981) 32 – 40.
178. G. S. Hoppins III, F. A. Schweizer: "MA 6000 E for Small Aircraft Gas Turbine Blades," *Frontiers of High Temperature Materials Conference,* INCO, May 1981.
179. E. Adler, W. Reppel, W. Rodewald, H. Warlimont: "How to Match Powder Metallurgy and Physics of Magnetic Materials," in W. J. Huppmann, W. A. Kaysser, G. Petzow (eds.): *Powder Metallurgy in Science and Practical Technology,* vol. 2, Verlag Schmid, Freiburg 1986, pp. 155 – 176.
180. K. J. Stranat, R. M. W. Stranat: "Rare Earth-Cobalt Permanent Magnets," *J. Magn. Magn. Mater.* **100** (1991) 38 – 56.
181. J. F. Herbst: "$R_2Fe_{14}B$ Materials: Intrinsic Properties and Technological Aspects," *Rev. Mod. Phys.* **83** (1991) no. 4, 819 – 898.
182. H. H. Stadelmaier, E.-T. Henig: "Permanent Magnet Materials – Development During the Last 18 Months," *J. O. M.* **43** (1991) no. 2, 32 – 35.
183. P. Schwarzkopf, R. Kieffer, W. Leszynski, F. Benesovsky: *Cemented Carbides,* Macmillan Co., New York 1960.
184. M. T. Martyn, P. J. James, B. Haworth: "Injection Moulding of Hardmetal Components," *Met. Powder Rep.* 1988, 812 – 815.
185. N. Reiter, H. Kolaska: "Schneidstoffe heute und morgen – Hartmetalle und beschichtete Hartmetalle," in H. Kolaska (ed.): *Schneidstoffe – Pulvermetallurgie in Wissenschaft und Praxis,* vol. 4, Verlag Schmid, Freiburg 1988, pp. 135 – 162.
186. P. Ettmayer, H. Kolaska: "Cermets (wolframcarbidfreie Hartmetalle)," in H. Kolaska (ed.): *Schneidstoffe – Pulvermetallurgie in Wissenschaft und Praxis,* vol. 4, Verlag Schmid, Freiburg 1988, pp. 163 –193.
187. F. Krupp GmbH DE-OS 3 806 602, 1988 (H. Kolaska, P. Ettmayer).
188. E. Rudy: "Boundary Phase Stability and Critical Phenomena in Higher Order Solid Solution Systems," *J. Less-Common Met.* **33** (1973) 43 – 70.
189. C. G. Goetzel: "Cermets," in E. Klar (ed,): *Metals Handbook,* 9th ed., vol. 7, American Society of Metals, Metals Park, Ohio, 1984, pp. 778 – 815.
190. A. N. Holden: *Dispersion Fuel Elements,* Gordon and Breach, New York 1967, pp. 80 – 167.
191. T. S. Shevlin: "Oxide Based Cermets," in: *Cermets,* Reinhold Publ. Co., New York 1960, pp. 97 –109.
192. C. G. Goetzel: *Treatise on Powder Metallurgy,* **vol. 2,** Interscience, New York 1950, pp. 171 – 174.
193. A. Keil, W. A. Merl, E. Vinaricky (eds.): *Elektrische Kontakte und ihre Werkstoffe,* Springer Verlag, Berlin 1984.
194. A. Erk, M. Schmelzle: *Grundlagen der Schaltgerätetechnik,* Springer Verlag, Berlin 1974.
195. K.-H. Schröder et al.: *Werkstoffe für elektrische Kontakte und ihre Anwendungen,* Expert Verlag, Ehningen 1992.
196. K.-H. Schröder: "The Future of Electrical Switching Contacts as Affected by Electronic Components," *14th Int. Conf. Electr. Contacts,* Paris 1988, pp. 307 –309.
197. W. Böhm, N. Behrens, M. Clasing: "Pulvermetallurgisch hergestellte Kontaktwerkstoffe für die Energietechnik auf der Basis $AgSnO_2$," *Metall.* **35** (1981) no. 7, 539 ff.
198. P. Braumann, H. Kolb, K.-H. Schröder, W. Weise: "Optimization of $AgSnO_2$ Contact Material for Use in Contactors," *16th Int.*

Conf. Electr. Contacts, Loughborough, UK, Sept. 7 – 11, 1992.
199. K.-H. Schröder: *Kontaktverhalten und Schalten in der elektrischen Energietechnik,* VDE-Fachbericht no. 42, VDE Verlag, Berlin 1991, p. 29 ff.
200. P. Braumann: "Investigation of Contact Materials for Switching Devices in Motor Vehicles," *14th Int. Conf. Electr. Contacts,* Paris 1988, pp. 423 – 428.
201. J. K. Lancaster, *Tribology* **6** (1973) 219 – 251.
202. V. T. Morgan: "Bearings from Metal Powders," *Powder Metall.* **21** (1978) 80 – 85.
203. V. A. Tracy: "Production of Porous Nickel for Alkaline-Battery and Fuel-Cell Electrodes," *Powder Metall.* **8** (1966) 241 – 255.

Praseodymium → Rare Earth Elements
Primary Cells → Batteries
Primary Explosives → Explosives

Printing Inks

Erich Frank, Flint Group Germany GmbH, Stuttgart, Germany
Rolf Rupp, Flint Group Germany GmbH, Stuttgart, Germany

Related Articles → Imaging Technology and → Textile Printing are separate keywords.

1. Introduction 1
2. Typographic Printing 2
2.1. Letterpress Printing 2
2.1.1. Cylinder and Flatbed Printing 2
2.1.2. Newspaper Letterpress Printing 2
2.1.3. Indirect Letterpress and Dry Offset Printing 3
2.2. Flexographic Printing 3
2.2.1. Solvent-Containing Flexographic Printing Inks 3
2.2.2. Water-Based Flexographic Systems . 3
2.2.3. UV-Curing Flexographic Inks 4
3. Planographic (Flatbed) Printing — Offset Printing 4
3.1. Sheet-Fed Offset Printing 5
3.1.1. Sheet-Fed Metal Decoration Printing 6
3.1.2. UV-Curing Offset Inks 6
3.2. Heatset Web Offset Printing 6
3.3. Newspaper Offset Printing 6
3.4. Fountain Solutions 7
4. Gravure Printing 7
4.1. Rotogravure 7
4.1.1. Publication Gravure 8
4.1.2. Packaging Gravure 8
4.1.3. Special Gravure Inks 10
4.2. Intaglio (Steel-Plate Gravure) 10
5. Screen Printing 10
6. Multicolor Printing 11
7. Properties and Testing of Printing Inks 11
7.1. Fastness Properties 11
7.2. Consistency 12
8. Pigments 12
8.1. Inorganic Pigments 12
8.2. Organic Pigments 13
8.3. Fillers and Extenders 14
9. Economic Aspects 14
10. References 15

Printing inks for the graphics industry are preparations that are transferred to a substrate by means of a printing form. Printing can be either informative, as in texts, or decorative, as with ornaments or pictorial representations.Typographic printing includes the letterpress printing involving raised, hard printing elements and a paste printing ink, and flexographic printing with liquid, quickly drying inks. The offset printing is based on the mutual repulsion between water and fatty printing inks and allows even relatively rough materials to be printed. Other printing processes including gravure printing, screen printing, multicolor printing are described in detail.Pigments in the printing ink industry must fulfill several requirements concerning dispersibility, rheological properties, and transparency. The article gives a list of inorganic and organic pigments, as well as additives that are used in the printing ink technology. A survey of toxicity and economic aspects is also given.

1. Introduction

Printing inks for the graphics industry are preparations that are transferred to a substrate, to which they subsequently adhere, by means of a printing form (definition of the European Printing Ink Assotiation, EuPIA). Printing can be either informative, as in texts, or decorative, as with ornaments or pictorial representations. Materials such as paper, cardboard, sheet metal, glass, plastics, foils, and wood may serve as substrates. In contrast to paints and coatings, print is usually incapable of protecting the substrate material. Typography always presupposes use of some kind of *printing form*. Indeed, printing processes can be classified as a function of the type of form employed (Table 1). Electronic and other special printing processes, such as textile printing, are treated elsewhere and will not be considered here (→ Imaging Technology, → Textile Printing).

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_143.pub2

Table 1. Characteristics of printing processes and the nature of the corresponding inks

Process	Inking system	Ink	Drying
Typographic			
Letterpress	long, many rollers	paste, remains fresh for long periods	slow; penetration, oxidation
Flexography	very short, 1 or 2 rollers	liquid, rapidly drying	very rapid; evaporation, UV
Lithographic			
Offset	long, many rollers	paste, remains fresh for long periods	slow; penetration, thermosetting, oxidation, UV, electron-beam
Gravure			
Gravure, rotogravure	direct inking	liquid, rapidly drying	evaporates very rapidly; two-component system, UV
Intaglio gravure	short, several rollers	paste, remains fresh for long periods	oxidation, penetration, evaporation
Screen	direct or squeegee	very soft	variable; thick ink layers

The mode of operation of the printing form determines the technology used for inking and thus the drying rate and consistency (viscosity and tack) of the corresponding printing ink (see Section 7.2). These conditions limit the possible formulations, with the result that not every printing process is equally suitable for all printing materials.

In the course of printing, ink must be transported from a reservoir on the printing press to the printing form. The resulting image is transferred in its fresh state either directly or via an intermediate carrier (e.g., a rubber blanket) to the substrate to be printed. The printing ink should then dry very quickly, or at least harden to such a degree that prints will not stick together in a stack or roll, even at high printing speeds. In this process, *resins* are needed to ensure abrasion-resistant binding of the ink solids, possibly to confer gloss to the print, and to produce an ink film with maximum chemical resistance.

2. Typographic Printing

2.1. Letterpress Printing

Forerunners of the letterpress printing process utilized bricks and cut wooden slabs to produce incunabula. Movable type made of clay, wood, or metal was also introduced at an early date in Eastern Asia. Casting instruments for movable type made of a lead alloy, the letterpress, and the associated printing ink were invented by Johannes Gutenberg (1397 – 1468) in Mainz, Germany.

Printing Process. Letterpress printing involves raised, hard printing elements that are evenly and lightly inked with rubber rollers (→ Imaging Technology, Chap. 3.4.2). The paste printing ink is uniformly rolled to a layer thickness of about 1 – 3 µm in long inking devices with the aid of approximately 10 – 20 rubber and metal rollers. The composition of the inks must be adapted so that they dry relatively slowly and do not begin to harden in the inking device.

2.1.1. Cylinder and Flatbed Printing

Printing speeds for sheet-fed flatbed machines reach ca. 3500 sheets/h. Letterpress printing is performed almost exclusively with offset printing inks. If necessary, these inks may be modified slightly with printing aids (see Section 3.1). The dominant position occupied until the 1960s by the letterpress process for high-quality printing on paper has now been assumed by offset printing. The main reasons for this development were the relatively expensive production costs of letterpress forms for text and graphics, time-consuming preparation of the printing press, and problems with ink trapping in multicolor wet-on-wet printing.

For these reasons letterpress machines of both the cylinder and flatbed types are now employed primarily for processes like embossing, stamping, perforation, numbering, and hot embossing, i.e., for work that demands devices with raised patterns and high mechanical pressure.

2.1.2. Newspaper Letterpress Printing

With the plastic plates now used in newsprint rotary letterpress production, it is possible to

achieve rates of ca. 25 000 cylinder revolutions per hour (ca. 8 m/s). The presses themselves are designed to operate at rates as high as about 12 m/s. However, the demand for high print quality on inexpensive newsprint, as in the case of color pictures, encourages the use of rival processes, such as offset and water-based flexographic newspaper rotary printing (Sections 2.2.2 and 3.3).

The print must be still wet when it is drawn over the former of the printing press; otherwise partially dried ink will be deposited on the metal parts. For this reason, dark, rather viscous mineral oils, which penetrate only slowly into paper, serve as "binders". It must also be possible to deink the prints in order to recycle wastepaper. Therefore, like offset newspaper inks (Section 3.3), such inks should be free of substances that undergo oxidative drying or cross-linking. In high-grade black and colored printing inks based on organic pigments, thinner, lighter mineral oils with hard resins function as binders (also bitumen in the case of black ink).

2.1.3. Indirect Letterpress and Dry Offset Printing

When printing uneven, embossed surfaces or molded articles (e.g., cups), an elastic rubber blanket is placed between the printing form and the material to be printed, just as in offset printing. Special printing inks are frequently applied in the printing of plastic beakers. These inks dry when subjected to an energy source (e.g., UV radiation) or partially penetrate into the plastic material (plasticizer as solvent).

The number of possible ink formulations is larger than in wet offset printing because interaction with water need not be taken into consideration.

2.2. Flexographic Printing

Flexography originated in about 1900 as aniline-dye rubber printing, used for simple packaging prints. With the introduction of photopolymer plates, flexography developed in the 1960s into an established graphic printing process using pigmented systems. Flexographic printing inks are liquid, unlike letterpress and offset inks. They dry very quickly.

As in letterpress printing, raised, inked elements of a rubber or plastic printing plate press directly against the material to be printed (→ Imaging Technology, Chap. 3.4.2).

2.2.1. Solvent-Containing Flexographic Printing Inks

A very compact inking system permits the use of low viscosity printing inks that dry quickly by evaporation of solvents (see Table 1). Modern flexo printing presses use laser-engraved ceramic anilox rollers which transfer a well-defined amount of ink to the flexible printing form. The open ink fountains of former times have been replaced by chambered doctor blades. The printing speed is about 400 – 600 m/min, depending on the type of impression. The flexographic process is used for the total range of flexible substrates from paper (shopping bags, gift-wrapping paper, napkins) and board (corrugated, substrates) to nonabsorbing films (e.g., PE, PP, PET, PA) and foils or metallized films.

The composition of flexographic printing inks is very similar to those of packaging gravure inks (see Section 4.1.2). The main difference is the solvent composition and therefore the drying properties, owing to the different distance from doctor blade to printing nip for the two processes. This means that flexographic printing inks contain a higher amount of solvents with lower evaporation rate (retarder). The drying equipment of a common flexo press consists of a hot air dryer.

Because of some advantages in costs (e.g., cheaper print form production), the flexographic process competes with gravure printing even for high-quality reproductions.

2.2.2. Water-Based Flexographic Systems

For environmental reasons the solvents in the inks are increasingly being replaced by water. Especially for absorbent substrates like paper or board, water-based inks are totally replacing solvent-based inks. Modern water-based inks contain only a small amount of solvents (e.g., less than 1 %), and in many applications these ink systems give excellent results without using

any solvent. Much progress has been made in printing water-based inks on nonabsorbent substrates such as PE or PP (shopping bags, heavy duty sacks). Even for these inks the solvent content is below 5 % (alcohol). Wallpaper (with or without plastic coating) can also be printed with specially adapted inks. The printing rate is 200 – 300 m/min, depending on the type of impression.

Drying of water-based inks is performed with the same equipment as for solvent-based inks, but in some cases additional drying with an IR dryer coupled with a cold-air blower can be useful, especially for increasing the printing speed for nonabsorbent films. The drying properties and the final fastness properties are adjusted as required by a special combination of mainly acrylic solutions and acrylic dispersions, solubilized by amines. On drying, the amines are evaporating and a water-resistant film is formed (see also Section 4.1.2).

Typical formulations for water-based flexographic inks are given in Table 2.

Table 2. Typical percentage formulations for water-based flexographic inks

Component	Paper printing	Film printing
Organic pigments	12 – 15	12 – 15
Resin system neutralized with amine (ammonia)	10 – 25	15 – 25
Alcohol (ethanol)		2 – 5
Additives: waxes, antifoamers, dispersing and wetting agents, microbiocides	5 – 7	6 – 10
Water	73 – 53	65 – 45

Water-based flexographic printing has also been introduced in some countries for newspapers, in particular in the United States. The resulting quality is usually superior to that of newspaper letterpress printing with paste inks (Section 2.1.2). The printing speed with waterbased flexography may reach ca. 13 m/s.

The aqueous, hydrophilic flexographic printing inks used today interfere with the deinking flotation process (i.e., the decolorization and reprocessing of old newspapers into fresh newsprint). However, new developments promise solutions to this problem in the near future.

2.2.3. UV-Curing Flexographic Inks

UV-Curing flexographic inks do not lead to air-quality problems because they do not evaporate. Their printability on paper and film is superior to that of water-based inks, and they produce higher gloss and better fastness (e.g., abrasion resistance, → Paints and Coatings). Their composition is similar to that of the UV-curing offset printing inks (Section 3.1.2) but their viscosity is lower.

3. Planographic (Flatbed) Printing — Offset Printing

Planographic printing was introduced as "lithography" (literally writing on stone) by Aloys Senefelder (1771 – 1834) in about 1798. Near the end of the nineteenth century it was developed further into *indirect offset printing*, in which the printing ink is transferred by means of a rubber blanket.

Printing Process. Offset printing is based on the mutual repulsion between water (or an aqueous fountain solution) and fatty printing inks. Whereas printing regions on the plate freely accept the fatty oleophilic inks, nonprinting areas on the same plane are subject to wetting by water (hydrophilic), causing them to repel the ink (see Section 3.4). (In waterless offset printing, a layer of silicone rubber replaces the ink-repelling aqueous wetting agent.) The resulting image is transferred to the printing material via a rubber blanket (for the platemaking process, see → Imaging Technology, Chap. 3.4.1).

A thin, uniform layer of ink is produced with a long inking system consisting of several rubber and metal rollers. For good printability, the inks must be viscous and tacky, and they should not harden in the inking system. For information on the ink – water balance, see Section 3.4. The elastic rubber blanket protects the sensitive printing plate, and it also clings to rough surfaces, permitting even embossed and relatively rough materials to be printed by the offset process. Precoated or raw sheet metal can be printed using special presses. High-quality prints with layer thicknesses of ca. 1 – 3 μm can be achieved.

Production. Pigments are usually supplied as powders that are dispersed in dissolvers and then ground in bead mills and/or three-roll mills with an appropriate oleophilic binder until a very fine dispersion is achieved. Larger particles remain behind and are eliminated.

In another method the finely divided pigments are precipitated from the aqueous phase. They are not dried but transferred directly to the oleophilic phase by a flushing process. The pigments do not agglomerate.

Occupational Health. Laws applicable in Germany stipulate that offset inks be nontoxic or only slightly toxic when used in accordance with regulations. Certain substances must, however, be specified in some countries. Many countries have special regulations for the use of printing inks in food packaging, toys, etc. Waste printing inks are often classified and treated as waste that has to be stored in special, authorized dumps or incinerated under defined conditions.

Substances classified as dangerous have in most cases been replaced by other, harmless materials. Rare exceptions are indicated by appropriate markings on each package.

3.1. Sheet-Fed Offset Printing

Sheet-fed offset printing is used for brochures, calendars, posters, business papers, and packaging (carton and soft packings). Printing speeds of ca. 15 000 sheets/h can now be attained, and speeds of ca. 12000 – 15 000 sheets/h are common.

In the case of absorptive printing materials, drying occurs primarily by penetration of the thin oils into the surface of the paper. This causes the viscosity of the ink to increase rapidly. The resulting firmness is usually sufficient to prevent printed sheets from sticking together in a stack. Depending on the printing material and the ink, the process is complete within a few minutes. Subsequently, cross-linking of oxidatively drying components in the binder occurs under the influence of oxygen. The print achieves a state of final hardness after several hours.

Papers, cartons, etc., often display differing absorptive capacities. Moreover, the oxidative drying process can be appreciably retarded by the presence of unsuitable components in the wetting agent or excessively high water absorption by the inks. Sticking of the drying prints in a stack can usually be prevented by dusting them with finely screened powdered chalk or plant starch.

Typical formulation for sheet-fed offset inks:

Organic pigments	12 – 20 wt %
Binders	
Hard resins	20 – 25 wt %
Soft resins + drying oils	20 – 30 wt %
Mineral oils	20 – 30 wt %
(*bp* 250 – 300 °C)	
Additives	5 – 10 wt %

Sheet-fed offset inks are also used in letterpress printing.

Binders. Interactions with the offset wetting agent must be considered when selecting the types and quantities of raw materials.

Hard resins not only give inks the tack required for good transfer and sharp print, but also confer excellent gloss on the finished print. The most important types of hard resins are rosin modified with phenols and aldehydes and displaying mineral-oil solubilities that lie within a definite range.

The mineral oils should, of course, readily dissolve the resins and keep the ink fresh in the inking system, but on absorptive printing materials, they should also quickly separate from the resin and penetrate into the substrate. In addition to their viscosity and boiling range, their solution properties are also important and are a function of their content of aromatics, naphthenes, and paraffins. Low viscosity accelerates penetration, and good solvent properties retard penetration.

In special sheet fed inks, the mineral oils are replaced by vegetable oils (renewable raw materials) or at least by derivatives of vegetable oils. These modifications are possible without impairing the printability.

Soft resins and drying oils are also good dissolution agents for hard resins. In the presence of driers and atmospheric oxygen these materials undergo oxidative cross-linkage and thereby confer mechanical stability on an ink film. The soft resins are usually fatty alkyds (fatty-acid-modified isophthalic acid polyesters). Drying oils include linseed, soybean, tung, ricinic, and safflower oils.

Additives comprise driers (oxidation catalysts), waxy lubricants for better abrasion resis-

tance (e.g., polyethylene), and anti-setoff powders (e.g., starch) that help prevent sticking of freshly printed sheets.

Pigments. Black paste inks contain special types of carbon black as pigment (see Section 8.1). Reflex blue pigments (e.g., Pigment Blue 56) or other, mostly blue, pigments may be added for toning purposes or to intensify the black effect on the human eye. Gold and silver bronzes and titanium white are somewhat important inorganic pigments (see Section 8.1). For colored organic pigments, see Section 8.2. Inorganic fillers (extenders, Section 8.3) are added as required.

3.1.1. Sheet-Fed Metal Decoration Printing

Metal plates and other nonabsorbing materials obviously are not compatible with the use of high-boiling mineral oils because penetration cannot occur. Oxidatively drying oils and alkyds assume the role of the resin-dissolving agents.

Drying usually takes place in a hot-air oven within several minutes. The baked sheets, perhaps after further lacquering, can be punched or deep-drawn into containers and crown corks. Ink films on cans for the food industry must be resistant to sterilization.

3.1.2. UV-Curing Offset Inks

UV-Curing offset inks are useful in the printing of packaging, labels, and nonabsorbing printing materials. UV printing inks typically contain unsaturated acrylates, polyesters, photoinitiators, and additives. At the end of the press the print is exposed to UV radiation and hardens within a fraction of a second. Printing speeds up to ca. 400 m/min are attained during continuous printing. UV-curing inks are of special interest in continuous form printing when the prints are intended for subsequent processing in a laser printer, where conventional inks cause problems such as evaporation of mineral oils and adhesion of ink on fuser rollers. Because of their high stability and excellent gloss, UV-curing printing lacquers (varnishes) are of interest as gloss coatings for conventional offset and letterpress products.

3.2. Heatset Web Offset Printing

Heatset web offset printing is used for double-sided, four- or five-color printing of medium to large print runs of brochures, magazines, books, etc., with a maximum speed of about 15 m/s. Immediately after the printing process, the paper web passes through a drying oven where it is dried with a high-speed hot air current. The residence time in the oven has to be at least 1 s, which means an oven length for fast-running presses up to 15 m. Mineral oils that volatilize are removed from the exhaust air by afterburning. Heatset inks are expected to release their volatile components as quickly as possible and produce high-gloss printed images. For this reason they contain relatively low-boiling mineral oils (*bp* ca. 200 – 300 °C) and gloss-producing hard resins. The mineral oils should not, however, evaporate on the printing press itself. Their boiling range must therefore be kept within very narrow limits.

Typical formulation for heatset web offset inks:

Organic pigments	15 – 25 wt %
Hard resins	25 – 35 wt %
Soft resins + drying oils	5 – 15 wt %
Mineral oils (*bp* 200 – 300 °C)	25 – 40 wt %
Additives	5 – 10 wt %

Because of the process conditions, compositions for these inks are similar to those of sheet-fed offset inks, although the proportion of oxidatively drying resins and oils is lower, the mineral oils have lower boiling ranges and driers are normally excluded. Process colors predominate in four-color printing, sometimes combined with a spot color.

3.3. Newspaper Offset Printing

As a result of quick, cost-effective platemaking processes and better image reproduction, newspaper offset printing (usually referred to as coldset printing in the UK) has largely replaced newspaper letterpress printing (see Section 2.1.2). Printing inks for this application dry by penetration, and it should be possible to remove them from waste paper by deinking (Section 2.1.2). For this reason they do not contain

binder components that undergo oxidative cross-linking, but rather hard resins dissolved in high-boiling mineral oils (*bp* ca. 260 – 350 °C). Black newspaper offset inks often contain dark, viscous mineral oils and asphalts or bitumen. The pigments are the same as those used in sheet-fed offset printing (see Chap. 8).

Use of an IR dryer to support the drying process allows thicker layers of ink to be printed and, as a result, more brilliant prints even on smooth newsprint paper.

The corresponding printing presses are designed for speeds up to ca. 13 m/s. Actual printing speed approaches 35 000 impressions per hour = ca. 11 m/s.

3.4. Fountain Solutions

The function of a fountain solution in offset printing is to moisten the hydrophilic areas of the printing plate and displace oleophilic inks (see 2.1.3). During the printing process, the fountain solution aids separation of the more viscous ink layer. The fountain solution is emulsified with the ink; emulsification speed, the water-absorbing capacity of the ink, and the particle size of the fountain solution droplets all exert decisive influences on the printing process. A balance must be established: too little fountain solution leads to printing of ink in nonprinting areas, but too much produces buttery, "short" inks in the inking system that fail to transfer properly.

The viscosity and tack of the ink, as well as the static and dynamic surface tension and pH of the fountain solution, represent important physical characteristics with respect to the offset process. Unsuitable ink components may upset the ink – water balance to the detriment of the product and disturb the printing process. Consumption of the fountain solution is in the same range as that of the printing ink and depends on the printed area and the material to be printed.

Typical formulation for wetting-agent concentrates (applied at a concentration of 2 – 5 %):

Buffer system	4 – 8 wt %
Alcohols	5 – 20 wt %
Surfactants	0 – 1 wt %
Hydrophilic polymers	1 – 7 wt %
Complexing agents	0 – 2 wt %
Preservatives	0 – 4 wt %
Anticorrosion additives	0 – 2 wt %
Water, added to make up to	100 wt %

Buffer systems (e.g., citrate buffer) produce a slightly acidic pH, which is effective in the regeneration of the printing plate and offers protection against alkaline substrates (→ Imaging Technology, Chap. 3.4.1). *Alcohols* and *surfactants* reduce the surface tension and improve the wetting power. *Hydrophilic polymers* have been used since the invention of offset printing. Gum arabic is being replaced to some extent by other polysaccharides or cellulose derivatives. The many alcoholic groups in these materials adhere very firmly to the oxide layer of the printing plate, thereby rendering it permanently hydrophilic. *Complexing agents* such as ethylenediamine tetraacetic acid (EDTA) deactivate interfering ions (e.g., Ca^{2+}). *Preservatives* inhibit fungal growth in the fountain solution. *Anticorrosion agents* prevent metallic parts of the printing press from damages.

4. Gravure Printing

Printing processes using printing forms into which the printing areas have been engraved are known as gravure printing (DIN 16 528).

4.1. Rotogravure

Copperplate engraving and etching were forerunners of rotogravure. The first industrial application was the printing of fabrics during the eighteenth century. The beginning of the twentieth century saw the introduction of photomechanical cylinder etching. For modern platemaking, see → Imaging Technology, Chap. 3.4.3.

Printing Process. The printing form in this case is usually a chromium-plated copper cylinder whose printing areas consist of tiny cells. The cylinder is dipped into a thin printing ink which fills the cells. Before the cylinder touches the substrate, excess ink is first wiped off in such a way that ink is transferred only from the cells (see → Imaging Technology, Chap. 3.4.3). Gravure printing is used mainly for printing large numbers of copies (e.g., > 500 000). The preparation of the cylinder is expensive, but very

high printing speeds can be attained because drying proceeds rapidly. The short path from the ink reservoir to the material to be printed via the deeply etched, mechanically and chemically resistant cylinder permits use of rapidly evaporating solvents and a wide range of binders.

Production of Gravure Inks. Powdered pigments and binders, including solvents, are mixed in dissolvers and subsequently dispersed in closed bead mills. Impurities and coarse pigment agglomerates are removed with filter devices such as bag filters.

Pigment preparations (chips) supplied by the pigment industry contain high concentrations of pigments finely dispersed in appropriate binders. Ready-for-use gravure inks are obtained simply by stirring and dissolving these preparations in dilute resin solutions. Such inks are often glossier, more brilliant, and usually more transparent than those made by grinding a powder. However, with this production method the binder cannot be changed at such short notice as in the case of dispersion of dry pigment.

4.1.1. Publication Gravure

This process is employed in the web printing of very large press runs of magazines, catalogs, etc., usually in multicolor applications. Printing speeds attained in magazine printing range up to ca. 15 m/s, with paper widths exceeding 4 m.

Printing Inks. Publication gravure utilizes hard resins dissolved in toluene, in which drying is achieved by solvent evaporation, most of which is recovered from the exhaust air with recovery systems (activated charcoal), thereby ensuring compliance with environmental legislation. Instead of toluene, low-boiling aliphatics and appropriately adapted colophony-based binders may also be used.

Water-based inks like those used in packaging gravure and flexography are not common because of the ease with which toluene can be economically recovered.

Typical formulations for publication gravure inks:

Pigments	ca. 8 – 15 wt %
Resins	ca. 15 – 20 wt %
Toluene (solvent)	ca. 60 – 70 wt %
Additives	ca. 0.5 – 5 wt %

Binders applicable to illustration publication gravure include maleinate resins (rosin modified with maleic acid), phenolic resins (colophony resins modified with phenols and aldehydes), calcium and zinc resinates (rosin "salts"), hydrocarbon resins, ethylcellulose, and bitumen (used only for black inks).

Additives. *Waxes*, often based on polyethylene, enhance surface smoothness (i.e., abrasion resistance) of the print. *Softeners* improve elasticity, and *defoaming agents* prevent foaming in the ink reservoir.

Pigments. Only a few types of pigments are applied in publication gravure because this usually involves four-color process printing with standard colors.

Black:	Pigment Black 7, C.I. 77 266, often shaded toward blue with Pigment Blue 27, Pigment Blue 15 : 3, or Pigment Blue 1, Pigment Blue 56 and 18
Yellow:	Pigment Yellow 12, Pigment Yellow 13
Magenta:	Pigment Red 57 : 1, Pigment Red 81 : 1
Cyan:	Pigment Blue 15 : 3, possibly containing Pigment Blue 27

Occupational Health. According to EC regulations, special labeling is obligatory for toluene. This substance is both highly flammable and damaging to health (→ Toluene, Chap. 8.2). Its classification as a health hazard is continually under review. Printing plants must protect their employees from the effects of toluene by installing equipment that ensures the removal of vapors, encasement of presses, etc., with verification by air monitoring. Pigments and binders are not toxic when used according to directions.

4.1.2. Packaging Gravure

Packaging gravure must be adapted to many different printing materials and the varying demands placed on the printed product. This requires much experience and great flexibility on the part of the printer. Rolls of paper, plastic films, aluminum foils, and laminated stock are all subject to such printing. Depending on the printing press and the object to be printed, printing speeds range between ca. 200 and 400 m/min. In contrast to publication gravure, toluene and aliphatic solvents are not employed. Ethyl acetate and alcohols (e.g., ethanol, 1- and

2-propanol, ethoxypropanol) are the most important solvents. The binders must be adapted to diverse requirements and their compositions therefore vary widely. In food-packaging materials, for instance, only a minimal residual solvent content is permissible.

Cellulose nitrates, polyvinylbutyral, polyurethanes, and other binders in combination with plasticizers are used with alcohols and/or aliphatic esters for many types of paper, lacquered (NC, acrylic) aluminum foils, metallized films, and the whole range of polymer films (PE, PP, polyester, polyamide). Polyethylene and polypropylene films reject printing inks. Print therefore only adheres to them after the film has been subjected to preliminary electrostatic treatment. The most important binder components in this application are cellulose nitrate containing special plasticizers, acrylate resins, polyurethane resins, and polyamide resins with ethanol and ethyl acetate as solvents. Poly(vinylidene chloride) surfaces can be difficult to print. Adherance can be achieved with, for example, mixed polymers of vinyl chloride – vinyl acetate – vinyl alcohol dissolved in ethanol and ethyl acetate. Blocking resistance can be improved with methacrylates or cellulose acetobutyrates. The pigments must be resistant to the plasticizers present.

Binders. Cellulose nitrates are classified according to their degree of nitration into ester-soluble types (nitrogen content ca. 12 %) and alcohol-soluble types, which are also soluble in esters (nitrogen content ca. 11 %). Both types are supplied in varying degrees of polymerization that confer varying levels of solution viscosity. The low-viscosity, alcohol-soluble grades are especially important for printing inks. Their solution viscosity is very high compared to other resins, so concentrations are limited to 12 – 18 %. Hard resins (e.g., maleic resin) are added to increase the solids content, thereby improving the film properties and increasing the gloss of a print.

Since cellulose nitrate is very brittle, the ink film must be made elastic (i.e., softened). Plasticizers are used, and include citrates, and adipates (35 – 50 % based on cellulose nitrate).

Pigments. Corresponding to the wide range of applications of packaging printing, a large variety of pigments is used. In addition to pigments employed in publication gravure, almost all types of organic pigments find application, just as in offset printing. Inorganic pigments include gold bronzes, silver bronzes (aluminum), and nacreous pigments (coated mica), as well as titanium dioxide for white backgrounds on transparent plastic films.

The pigments are responsible for fastness properties (e.g., lightfastness and resistance to migration). An especially good glossy, gold effect can be achieved on aluminum foil with soluble yellow zapon dyes and cellulose nitrate binders.

Additives. Apart from the usual substances applicable to publication gravure, matting agents or antilubricants and other specialties are employed.

Occupational Health. Pigments, binders, and solvents used to print packaging materials are not classified as toxic or even moderately toxic materials. However, almost all the solvents are flammable. If the printing ink contains hazardous substances, it must be labeled in accordance with appropriate regulations. It is the responsibility of the printer to make sure that limiting values (e.g., MAK, TLV values) for volatile substances are not exceeded, and that these defined limits in the outlet air from the dryer are not exceeded. Limiting values with respect to exhaust air can be satisfied if the air is treated before it is discharged. In contrast to publication gravure, this cannot be achieved by adsorption on activated charcoal because the solvent mixtures vary, and may include water-soluble components. The only recourse is to use a thermal afterburner.

Gravure inks must be excluded from wastewater systems.

Gravure with Water-Based Inks. For reasons of environmental protection and occupational safety, gravure inks that are free of, or low in solvent, are becoming increasingly important. Not only paper, but also polyolefin and aluminum foils and other nonabsorbing materials are printed with water-based inks. Inks of this type are also used in the gravure printing of wallpaper.

The advantage of water-based inks is that they create few problems with regard to the outlet air, and they are not flammable. In addition, occupa-

tional safety considerations (MAK, TLV values) do not apply.

As in water-based flexography, the disadvantages are slower drying and the limited choice of binders. Any attempt to accelerate drying by using high temperatures at high printing rates damage a thermoplastic film.

Use of an IR irradiator together with a cold-air blower or longer drying ovens at lower temperatures result in some improvement.

Printing Inks. As in water-based flexography, additives include acidic acrylates which become water soluble when combined with amines (e.g., ammonia, dimethylethanolamine, triethanolamine). A water-resistant ink film remains after evaporation of water and the amine because acidic acrylates are insoluble in water. Prints of this type are not, however, resistant to alkali and detergent. As with water-based flexographic newspaper printing (Section 2.2.2), the extent to which resulting waste paper can be deinked by the flotation process is extremely low, although newer methods have been developed for making water-based inks flotable.

Gravure Transfer Inks. Gravure is the method of choice for printing with transfer inks. Aqueous printing inks containing selected textile disperse dyes are usually printed and dried on special papers. The printed image can then be transferred to the textile materials (e.g., polyester fabric) by pressing at ca. 200 °C. The dyes sublime and are absorbed by the textile fibers, in which they are soluble.

An advantage of this printing process is that production and storage of the paper prints are relatively inexpensive, and high-quality textiles become available at short notice.

Typical formulation:

Disperse dyes	10 – 12 wt %
Acid acrylates + amines	15 wt %
Water	73 – 75 wt %

4.1.3. Special Gravure Inks

The printing of wallpaper, decorative wood, and poly(vinyl chloride) floor coverings imposes special demands. The printed product must be resistant to light and other influences for long periods of time and capable of enduring severe stress, such as high temperatures, during processing. Water-based inks are rapidly gaining prominence here. Special binders and pigments that are resistant to both light and solvents are required. Inorganic pigments retain their importance in this application.

4.2. Intaglio (Steel-Plate Gravure)

Line intaglio or steel-plate gravure is based on the same principle as traditional copperplate engraving. The printing lines are incised into a shiny, polished printing form. However, the cylindrical form is made of Ni – Cr alloy and is chromium-plated. It is prepared for use by stamping it over a matrix.

This process is applicable to the mass production of high-quality items (e.g., banknotes and postage stamps). The corresponding viscous printing inks are transferred from a cutout rubber cylinder to the printing areas of the cylindrical form via a roller inking device. Excess ink is wiped off with a reverse-rotating plastic roller.

Printing occurs directly on the paper under high contact pressure. The wiping cylinder is cleaned after each revolution in an aqueous, alkaline cleaning solution with the help of brushes and doctor blades. It is then dried with air, leaving it clean for the next revolution. High demands for fastness require that selected types of pigments be used. Fillers confer higher viscosity without increasing the tack of the printing ink. Excessive tackiness would impair the wiping process, causing adhesion of the relief-like fresh prints in the stack.

The printing inks dry as a result of oxidation and evaporation of solvent. For this reason, the prints are released for circulation only after a prolonged period of storage. In contrast to gravure, the inks remain raised on the paper even after drying, and they can be felt with the finger.

5. Screen Printing

Screen printing uses stencils made of a fine netlike fabric as the printing form (→ Imaging Technology). The printing inks are viscous, flow easily, and produce layer thicknesses greater than those achieved with other printing processes.

Difficult surfaces (e.g., metals, glass, ceramics, wood, and many different plastics) predominate as printing materials, resulting in many types of prints; some show especially high resistance to light and weather. Examples are posters made of paper or plastic sheet (up to several square meters in size), signs (e.g., reflecting), transparencies, and self-adhesive labels, bottle crates, glass bottles, ampoules, candles, and even pills and eggs. High-quality multicolor artistic posters can also be produced.

Electrical printed circuits, including electrical track conductors, discharge resists, and solder masks, represent a special application of screen printing. Highly efficient equipment has been developed for printing conical or cylindrical objects. In rotary screen-printing machines, the screen is stretched over a cylinder, resulting in what is known as web printing. Printing ink inside the cylinder is forced outward with a squeegee. Printing speeds vary from about 100 prints/h for manual printing to ca. 3000 prints/h (120 m/min) with fully automatic rotary screen devices and fast-drying inks.

Depending on the substrate, the process, and the stability required, a choice may be made from a very wide selection of screen printing inks (two-component inks, inks that are baked or undergo physical, oxidative, or UV drying).

Acrylates are frequently utilized as binder for printing plastics, in some cases together with other resins or polymers. As in gravure printing, vinyl chloride copolymers are suitable for difficult surfaces.

Appropriate organic solvents (e.g., aromatics, aliphatics, ketones, alcohols, and esters) are higher-boiling than those normally employed in gravure printing. Solvent-containing cellulose inks for printing posters may contain up to 75 % aromatic hydrocarbons.

Physically dried inks permit stacking of the prints after passage through a continuous hot-air dryer. Small numbers of slow-drying copies are laid out on a frame.

Screen printing inks that are dilutable with water are becoming increasingly important despite the fact that they are not very weather-resistant.

6. Multicolor Printing

In accordance with color theory, multicolored images are reproduced by superposed printing of the three primary colors cyan (greenish blue), magenta (bluish red), and yellow, as well as black (→ Imaging Technology, Chap. 3.1). Industrialized nations have taken steps toward standardizing all of the stages involved in the use of three-colored and black inks for offset and letterpress printing, including color separation, platemaking, optical properties of the inks, and the actual printing process. The pigments themselves are not specified, however, so technical progress is not impeded. Nevertheless, pigments described in Chapter 8 and recognized as belonging to a "set of pigments for multicolor printing" are normally employed.

The color shades of the inks for four-colour offset printing are standardized in ISO 2846. In the corresponding standards ISO 12647 working procedures for print shops are defined, to reach the standard shades.

Similar standards for publication gravure and flexograhic printing are in preparation.

7. Properties and Testing of Printing Inks

7.1. Fastness Properties

Prints are expected to have fastness properties that are dictated by their use and are obtained by appropriate choice of pigments and binders. National and international standards have been established in the industrialized nations to guarantee the availability of consistent grades of ink with certified characteristics. The fastness of a particular final product also depends on the material to which it will be applied and several other factors; it can therefore vary considerably.

Important properties are:

- Lightfastness (e.g., for posters ISO 2835)
- Fastness to water (ISO 2836) (e.g., for shopping bags)
- Fastness to solvents (ISO 2836) (e.g., fastness to lacquering)
- Fastness to alkali (ISO 2836)
- Fastness to soap (ISO 2836)
- Fastness to detergents (ISO 2836)

- Fastness to cheeses (ISO 2836)
- Fastness to edible fats and oils (ISO 2836)
- Fastness to wax (ISO 2836)
- Fastness to spices (ISO 2836)
- Resistance to sterilization and heat (e.g., for metal cans)
- Resistance to migration (e.g., in printed plastic films)

Print samples are made with inks and tested in accordance with relevant standards. If no testing standard exists, tests are conducted after consultation with the customer.

Lightfastness is assessed in a quick-exposure device (ISO 2835). Half of the sample is covered and the rest is exposed to filtered xenon light until a definite change in color is observed. An eight-level blue wool scale is exposed simultaneously as a reference. Change is characterized according to the following levels:

8	outstanding	4	quite good
7	excellent	3	moderate
6	very good	2	poor
5	good	1	very poor

The length of time over which prints remain resistant to sunlight depends on the strength of the light. For instance, a print with a lightfastness of 6 may require from October to the beginning of May (ca. 225 days) to show the first visible signs of a color change, while the same print exposed from June onward might show the same change after only ca. 45 days.

7.2. Consistency

A printing ink will only produce a satisfactory, standardizable printed result with a particular printing press and print material when its flow properties and tack are appropriate. Flow behavior refers to such properties as viscosity, yield value, and thixotropy. They are highly dependent on the pigment and therefore often differ from ink to ink.

High-speed printing devices require relatively thin inks, whereas slower machines need thicker inks. Both the type of inking unit and the path traversed by the ink from the ink fountain to the substrate to be printed determine the ultimate viscosity range (i.e., whether the ink is thin or viscous, see Chap. 1). Viscosity is measured with a flow cup in the case of thin gravure and flexographic inks, and with a falling bar or rotation viscosimeter for paste inks (offset inks).

The *tack* reflects whether a substance is "short" (like butter) or "long" (flowing but sticky, like honey). It denotes the resistance of an ink film to splitting when distributed on rollers. Ink that is too short cannot be transferred accurately to the substrate to be printed. Ink that is too long (i.e., too tacky) causes "picking" of the substrate when the print rate exceeds a certain value. The tack is not a physically defined measure, but refers instead to a specific tackmeter which measures the drag force. A driven roller is loaded with a certain amount of printing ink, which is distributed with a spreading roller. Measurements of the resulting drag force on a nondriven roller placed above the driven roller are then made. The speed of the driven roller can be varied.

8. Pigments

The production and general properties of pigments are described elsewhere, → Pigments, Inorganic; → Pigments, Organic. Only properties that are important for printing purposes are treated here. Apart from characteristics that depend on the chemical composition of the pigments (e.g., shade, coloring strength, and fastness), the most important considerations in the ink industry are dispersibility, rheological properties, and transparency. These properties are in turn mainly a function of particle shape, particle size, particle-size distribution, and the character of the particle surface, all of which can be heavily influenced by the pigment production method.

8.1. Inorganic Pigments

In general, inorganic pigments are opaque (hiding power), lightfast, and heat-resistant materials that generally display relatively pale colors in impure shades. They also have relatively high densities. The most important products for the ink industry follow.

Pigment White 6 [*1317-80-2*] (titanium dioxide, titanium white), C.I. 77 891, is the pigment with the highest hiding power. It is employed in the anatase and rutile forms, either alone or mixed with colored pigments, in many types of printing inks, especially for flexographic, gravure, and screen printing, as a way of obtaining covering prints on such substrates as transparent plastic films, aluminum foils, and dark paper.

Nacreous pigments are obtained by coating mica with titanium dioxide, leading to mother-of-pearl effects in gravure printing.

Pigment Red 101 and 102, Pigment Brown 6 and 7 (iron oxide and iron silicate pigments), C.I. 77 015, 77 538, and 77 491, are yellow, reddish-brown, and black pigments that are resistant to light, weather, and heat. The transparent reddish-brown pigments are useful, for example, in gravure printing of imitation wood grains. Black and brown iron oxides with magnetic properties are employed for magnetizable printing inks (e.g., for magnetic stripes on credit cards).

Pigment Blue 27 (iron cyan blue, Milori blue, or Prussian blue), C.I. 77 510 is relatively strongly colored. It is used for blue printing inks and as shading components in black letterpress, offset, and publication gravure inks. Apart from poor fastness toward alkalies, its general fastness is good.

Bronzes C.I. 77 000, 77 400: Aluminum powder (silver bronze) and powder made from copper – zinc alloys (gold bronze) are processed to various degrees of fineness and quality for silver and gold effects on packaging and advertising.

Pigment Black 7 [*1333-86-4*] (carbon black), C.I. 77 266: carbon-black pigments are the most important pigments in most black printing inks because of their high coloring strength and excellent fastness.

Important application properties such as color depth, shade, gloss, rheological characteristics, and dispersibility depend upon particle size, surface conditions, and structure. Thus, relatively coarse varieties are used for newspaper inks, and fine types are employed for high-quality black printing inks.

8.2. Organic Pigments

Organic pigments are very important constituents of most printing inks. Organic dyes are losing importance because they display moderate or poor fastness. Approximately 140 000 t of organic pigments were consumed worldwide in 2004 in the production of process printing inks. Pigments are characterized by high coloring strength, pure shades, transparency, and good dispersibility. Lightfastness, resistance to heat, fastness to lacquering, and chemical resistance vary considerably. About 100 organic pigments are suitable for use in printing inks; only the most prominent are mentioned here.

Pigment Yellow 12 [*6358-85-6*] and *13* [*5102-83-0*] (diarylide yellow), C.I. 21 090, 21 100, 21 105, are generally intensely colored, transparent, and resistant to solvents and lacquering. They are utilized in all types of printing inks, especially for the yellow process ink in multicolor printing. Toxic decomposition products may be formed at high temperatures.

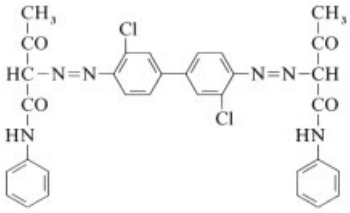

Pigment Yellow 12

Pigment Orange 34 [*15793-73-4*] (diarylide orange), C.I. 21 115, is very pure, transparent, and offers good lightfastness and offset printability.

Pigment Red 2 [*6041-94-7*] (naphthol AS red), C.I. 12 310, is a slightly bluish, pure red pigment with good lightfastness.

Pigment Red 57 : 1 [*5281-04-9*] (Rubin 4 B calcium salt), C.I. 15 850, provides moderate to good lightfastness, fastness to lacquering, and fairly good resistance to migration, but it is not stable to many chemicals. In all graphic multicolor printing processes this is the most important pigment for producing the red printing ink "magenta".

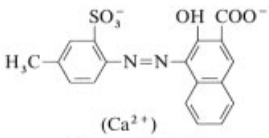

Pigment Red 57

Salts of basic dyes are very pure and brilliant. Their lightfastness is usually good, but they are very sensitive to solvents and chemicals. *Pigment Red 81*, rhodamine 6 G laked by precipitation with phosphorus – tungsten – molybdenum acid (PTMA), was previously used as a purple-red scale ink. Other examples are *Pigment Violet 1*, rhodamine B with PTMA, and *Pigment Blue 1*, Victoria blue with PTMA.

Pigment Blue 15 : 3 [*147-14-8*] (phthalocyanine blue), C.I. 74 160, is the most resistant organic pigment. The greenish β-modification is used in all graphic printing processes to provide the cyan for multicolor printing.

Pigment Blue 56 [*64742-56-9*] (reflex blue), C.I. 42 800, is a reddish, intensively colored pigment with average fastness but little resistance to lacquer solvents. It is frequently used to shade black letterpress and offset inks.

Pigment Green 7 [*1328-53-6*] (phthalocyanine green), C.I. 74 260, is a chlorinated phthalocyanine with high fastness, comparable to pigment blue 15. It is pure and transparent.

Polycyclic pigments include the highly resistant but expensive quinacridone, anthraquinone, and dioxazine pigments. They are employed for shading purposes and in specialty inks. The structure of a quinacridone pigment is shown below:

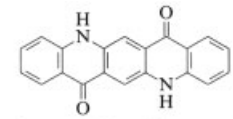

Pigment Violet 19 [*1047-16-1*]

Fluorescent pigments are solid solutions of various dyes in plastic or synthetic resins. They are capable of converting short-wavelength light into long-wavelength light, and they glow brightly when exposed to daylight or, better still, UV light. Resistances are moderate. Fluorescent pigments are used for inks applied to advertising and warning signs as well as packaging.

8.3. Fillers and Extenders

Extenders increase the viscosity of printing inks without simultaneously increasing their tack. They usually offer good transparency, but even small amounts interfere with the gloss of the print.

Aluminum hydrate C.I. 77 002, is used in letterpress and offset inks. Ground with binders, it is also employed as a "transparent white" to tint printing inks. The heavier *clay* (*blanc fixe*), which has somewhat better hiding power, is applied similarly.

Chalk ($CaCO_3$) may be finely ground or precipitated and sometimes hydrophobically coated (e.g., with waxes, fatty acids). It provides a slightly mat effect in paste printing inks.

Silica (SiO_2) is employed in an amorphous, highly disperse form, and is sometimes hydrophobically coated.

Bentonite (montmorillonite) is also coated. Even in small amounts it has a powerful thickening effect in many formulations.

9. Economic Aspects

Production data for printing inks in 2003 (estimation):

World	3 075 000 t/a
North America	830 000 t/a
Europe	1000 000 t/a
Southeast Asia	1 065 000 t/a
Latin America	180 000 t/a

Use in Europe is distributed as follows (2003, CEPE):

Germany	30 %
United Kingdom	17 %
France	12 %
Italy	12 %
Benelux countries	7 %

Market segments in Germany (2004, VdD):

Publication gravure	31 %
Specialty gravure	13 %
Flexography	6 %
Web offset (heatset)	23 %
Web offset (coldset)	12 %
Sheet-fed offset	6 %
Others	7 %

Suppliers (market share):

Worldwide:	7 large producers	(ca. 80 %)
	230 small producers	
Europe:	4 large producers	(90 %)
	80 small producers	
United States:	6 large producers	(80 %)
	110 small producers	

10. References

General References

1. R. H. Leach: *The Printing Ink Manual – I*, 4th ed., Society of British Printing Ink Manufacturers Ltd. 1988.
2. W. Baumann, B. Herberg-Liedtke: *Druckereichemikalien*, Springer Verlag, Berlin 1991.
3. J. D. Sanders: *Pigments for Inkmakers*, Select. Industrial Training Association Ltd., London 1989.
4. D. J. Owen: *Printing Inks for Lithography*, Sita Technology, London 1990.
5. W. Herbst, K. Hunger: *Industrial Organic Pigments*, VCH Verlagsgesellschaft, Weinheim 1993.
6. *Am. Ink Maker*, PTN Publications, Melville, New York.
7. *Ink + Print International*, Batiste Publications, Hornsey, London.

Propanal

ANTHONY J. PAPA, Union Carbide Chemicals and Plastics Company, Inc., South Charleston, WV 25303, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties and Uses** 1
4. **Production** 3
5. **Quality Specifications** 4
6. **Economic Aspects** 4
7. **Storage and Transportation** 4
8. **Toxicology and Occupational Health** . . 4
9. **References** 5

1. Introduction

Propanal [*123-38-6*], propionaldehyde, CH_3CH_2CHO, M_r 58.08, is a low-boiling, colorless, flammable liquid with a sharp, suffocating odor. It is found in vegetables such as onions, in cheese and other dairy products, and as an active ingredient in coffee beans, contributing to the aroma of coffee [1].

Propanal is highly reactive, and is employed primarily as a chemical intermediate to prepare C-3 and C-6 compounds. It is generally produced by hydroformylation of ethylene. Annual world production was estimated at 154×10^3 t in 1988 [2].

2. Physical Properties

Propanal is completely miscible with organic solvents such as alcohols, ether, and benzene. Solubility in water decreases at elevated temperature. For example, propanal is miscible with water in all proportions below 15 °C, but becomes less soluble and separates from solution in aqueous mixtures containing ca. 20 – 70 wt % propanal as the temperature increases. Propanal forms azeotropic mixtures with water (98 wt % propanal, *bp* 47.5 °C) and such other compounds as propyl nitrite (82 wt % propanal, *bp* 47.3 °C), 1-chloropropane (*bp* 46.4 °C), and cyclopropyl methyl ether (*bp* 43 °C). Some physical properties of propanal follow [3–7]:

Critical temperature	235.95 °C
Critical pressure	4.660 MPa
mp	−80.0 °C
bp	
101.3 kPa	48.0 °C
40.0 kPa	23.6 °C
1.33 kPa	−38.0 °C

Vapor pressure (20 °C)	34.391 kPa
Refractive index (25 °C)	1.3593
Liquid density (20 °C)	0.797 g/cm^3
Liquid viscosity (20 °C)	0.346 mPa · s
Surface tension (20 °C)	22.95 mN/m
Flash point	
Closed cup	−30.0 °C
Open cup	−7.2 – 9.4 °C
Autoignition temperature	206.85 °C
Lower flammability limit	2.9 vol %
Upper flammability limit	17.0 vol %
Solubility in water (20 °C)	35.6 wt %
Solubility of water in propanal (20 °C)	21.1 wt %

3. Chemical Properties and Uses

Propanal exhibits the characteristic high reactivity of low molecular mass aldehydes such as ease of oxidation and polymerization (trimerization). The reactions are caused by the polarity of the carbonyl group and acidity of the α-hydrogen atoms.

Uses. Propanal has no uses as such; it serves as a source of a propyl group for chemical synthesis. Propanal is primarily converted to 1-propanol, propionic acid, and trimethylolethane (trihydroxymethylethane). Minor applications

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_157

include the production of pharmaceuticals (e.g., meprobmat), agricultural chemicals, rubber additives, and corrosion inhibitors.

Reduction. 1-Propanol [*71-23-8*] is the major product of catalytic reduction of propanal (→Propanols). Reduction is carried out most economically by a continuous vapor-phase process over a heterogeneous catalyst of supported reduced nickel, copper, and/or zinc and manganese metals [8–10]. Reaction is normally conducted at 125 – 200 °C and a pressure of about 700 kPa. An advantage of the Cu – Zn catalyst is that it promotes hydrogenolysis of the propanal Tischenko ester (propyl propionate) byproduct to 1-propanol [11]. Liquid-phase reduction on a fixed bed of nickel and copper catalyst is also highly effective in reducing aldehyde to alcohol, but the technique suffers from low efficiency and high cost on a commercial scale.

Oxidation. Propionic acid [*79-09-4*] is produced commercially by liquid-phase noncatalytic air oxidation of propanal [12–14], see also → Propionic Acid and Derivatives, Chap. 4.2. Oxidation proceeds readily at relatively low temperatures (50 – 100 °C), and high aldehyde conversions and acid selectivities (>90 %) are achieved. Continuing efforts are being made to improve the process, particularly oxidation rates, using transition-metal catalysts [15–17]. Major byproducts result from reactions associated with the oxidation process itself, and only a few arise from propanal condensation reactions. Propionic acid is used as a preservative for animal feed and as a chemical intermediate for preparation of sodium and calcium salts added to baked goods, tobacco, and certain vegetables to inhibit the growth of mold and fungus. It is also used to prepare diethyl ketone by passing hot vapors over heterogeneous catalysts such as manganese(II) carbonate [18].

Peroxypropionic acid [*4212-43-5*] ($C_2H_5CO_3H$), useful for peroxidations of olefins, can be obtained by oxidation of propanal with oxygen over a complex-metal catalyst in acetone at 10 – 30 °C and atmospheric pressure in yields >85 % without decomposition of the peroxy acid [19].

Condensation and Addition. Propanal is a key intermediate for the manufacture of trimethylolethane [*77-85-0*] (2-hydroxymethyl-2-methyl-1,3-propanediol) by aldol condensation with excess formaldehyde in the presence of an alkaline catalyst [20–22]. In 1988, 5.4×10^3 t of trimethylolethane was produced and used largely to impart hardness, gloss retention, and durability in alkyd-resin formulations [2].

The two α-hydrogens of propanal can be readily replaced by formaldehyde in a modified Perkin-type reaction catalyzed with a mixture of secondary amine and a carboxylic acid to produce methacrolein [*78-85-3*] [23–26]. Methacrolein yields >90 % are obtained under relatively mild conditions (50 – 100 °C and <1 MPa). Atmospheric oxidation of methacrolein to methacrylic acid and subsequent esterification produces methyl methacrylate [*80-62-6*]. This route to methyl methacrylate appears to be more economical than similar processes starting with propionic acid. If successful, it could open a new major market for propanal [27], [28].

Propanal undergoes self-aldol condensation or cross-aldol condensation with aldehydes or ketones in the presence of aqueous strong alkali in a conventional two-phase liquid process [29]. Self-condensation gives initially the aldol 3-hydroxy-2-methylpentanal [30], which can be dehydrated by heating to 2-methyl-2-pentenal [*623-36-9*]. Typical impurities from the aldol condensation of propanal in alkali at ca. 80 – 100 °C have been identified, and include 2,4,6-triethyl-1,3,5-trioxane, 2,4-dimethyl-2,4-heptadienal, 5-hydroxy-2,4-dimethyl-2-heptenal, 7-hydroxy-2,4,6-trimethyl-2,4-nonadienal, and 3-ethyl-2,4-dimethylvalerolactone [30]. Most of these impurities result from chain lengthening by linear aldolization with 5-hydroxy-2,4-dimethyl-2-heptenal, which is present in significant amounts. The 2-methyl-2-pentenal obtained from dehydration of the propanal aldol intermediate can be readily reduced, oxidized, or made to undergo a variety of addition and condensation reactions. The most useful of these is reduction to 2-methyl-1-pentanol [*105-30-6*] (trade name UCAR Hexanol, Union Carbide Chemicals and Plastics Company, Inc.). Because of its branching, 2-methyl-1-pentanol exhibits lower evaporation rates and lower volatility than 1-pentanol or "primary amyl alcohol" (an industrial blend of 1-pentanol and 2-methylbutanol),

making it useful as a solvent extender and solution viscosity reducer. About 900 t of 2-methyl-1-pentanol was produced in 1990. The corresponding acetate ester was commercially available in 1985 – 1988, but has since been withdrawn from the market.

A continuous process for specific production of the aldol intermediate (97 – 99 % yield) has also been developed [30]. Because of its difunctionality, this β-hydroxyaldehyde is useful for the preparation of pharmaceuticals, plasticizers, and polymers.

Patents describing vapor-phase heterogeneous processes using catalysts such as Pd – rare earth oxides [31], anatase TiO_2 [32], group 11 and/or 12B metals supported on SiO_2 [33], B_2O_3 and/or P_2O_5 supported on TiO_2 or La_2O_3 [34], and La_2O_3, Nd_2O_3, and Sm_2O_3 [35] have been reported. This technique usually gives lower yields (ca. 90 %) and lower aldehyde conversions (<50 %), and it does not challenge the separation simplicity of the two-phase alkali process. Improved conversions (95 %) and yields (88 %) to 2-methylpentanal [*123-15-9*] were reported with a heterogeneous vapor-phase reductive condensation of propanal in the presence of Pd –PrO_3 catalyst on Al_2O_3 [31]. Other vapor-phase reductive condensation catalysts have been described for direct preparation of saturated ketones from cross-aldol condensations of propanal, including Ni, Co, and Zn supported on Al_2O_3 [36] and Mg, Al, Zn, or Te oxides or phosphates and a group 8 metal [37].

The Tischenko reaction can be used for one-step preparation of propyl propionate [*106-36-5*] (a specialty solvent) from propanal, but problems are associated with the aluminum alkoxide catalysts employed in this process. Propyl, butyl, and pentyl propionates are available as specialty solvents from conventional esterification of propionic acid with the corresponding alcohol [38].

A propanal trimer, 2-methyl-1,3-pentanediol monopropionate, used as a solvent modifier for paints, can be obtained from Tischenko rearrangement of the adduct formed between the aldol of propanal and a third molecule of propanal [39]. The nature of the Tischenko reaction of aldols of propanal is discussed in [40].

Propanal is converted to 2,4,6-triethyl-1,3,5-trioxane in the presence of acidic catalysts [41]. The ease of thermal reversion of the cyclic trimer has precluded an utility for materials of this class.

Propanal undergoes reductive catalytic amination with ammonia or primary, secondary, or tertiary amines to the corresponding propyl amine derivatives, which are used as corrosion inhibitors, fuel additives, pharmaceuticals, agricultural chemicals, and rubber additives [42–44]. Substituted pyridines, propionitrile (ethyl cyanide), or acrylonitrile can all be prepared by reaction of ammonia and propanal in the vapor phase depending on the temperature and catalyst employed [45–48].

Propanal can be converted to diethyl ketone, a herbicide intermediate, at high temperature (450 – 500 °C) in the presence of steam and a rare-earth metal-oxide catalyst [49], but propionic acid is the preferred starting material for this ketone [2].

4. Production

Low-pressure rhodium-catalyzed hydroformylation of ethylene with carbon monoxide and hydrogen is the most important process for the production of propanal (→ Oxo Synthesis). Some is still produced in the United States (and all in Europe) by the cobalt-catalyzed hydroformylation process at higher pressure, and a small fraction is obtained in Japan as a byproduct from Wacker oxidation of propene during acetone production [2].

The low-pressure hydroformylation process for production of propanal uses a soluble rhodium catalyst complex with excess triphenylphosphine. Reaction occurs in the liquid phase at 90 – 130 °C and a total pressure of <2.8 MPa with CO pressure <0.38 MPa and H_2 pressure <1.38 MPa [50]. Selectivities are >90 % because ethylene can lead to only one isomer and impurities can be minimized. Major impurities include 2-methyl-2-pentenal, which arises by aldol condensation; trimers from the Tischenko reaction afford 2-methyl-1,3-propanediol monopropionate, which can in turn disproportionate to a mixture of 2-methyl-1,3-propanediol dipropionate and 2-methyl-1,3-propanediol. Byproduct formation during propanal production can be inhibited by exclusion of oxygen, organic acids, and metal contaminants such as iron, which act

as Lewis-acid catalysts for its condensation to higher products.

5. Quality Specifications

Propanal purity is determined by gas chromatography using a capillary column. Care should be taken to ensure that the injection port is clean and low in temperature (<200 °C) to avoid condensation of propanal in the port.

Common impurities include butanal, 2-methylpropanal, aldol products, and various trimers. Typical specifications for high-purity-grade propanal follow [51], [52]:

Propanal	99.5 wt %, min. (dry basis)
Acidity	0.1 wt %, max. (calculated as propionic acid)
Distillation (103 kPa)	initial *bp* 46.0 °C, min.; after 97 mL, 50.0 °C max.
Water content	1.0 – 2.5 wt %
Color	20 Pt – Co, max. (ASTM)
Iron content	20 ppm by weight, max.

Water is generally added to stabilize the pure concentrated aldehyde against condensation to cyclic trimers or acyclic oligomers, catalyzed by Lewis acids such as iron, and to decrease peroxide formation on standing [53].

Table 1. Production of propanal derivatives in 1988 [2] ∗

Compound	Production, 10^3 t	
	United States	Western Europe
1-Propanol	88.5	11
Propionic acid	47.2	
Trimethylolethane	2.7	
Others	1.4	2

∗ Courtesy of SRI International.

6. Economic Aspects

Propanal is produced in the United States by Eastman Kodak, Hoechst Celanese, and Union Carbide Chemicals and Plastics Company, Inc. About 138 000 t was produced in 1988 in the United States [54]. In Europe, production is lower. The average annual growth rate for propanal during 1988 – 1993 is estimated to be ca. 0.8 % [2]. In 1988 an excess capacity for propanal manufacture of ca. 25 %.

Traditionally, propanal production has depended on the demand for 1-propanol and propionic acid, which account for more than 90 % of the aldehyde consumed (Table 1).

7. Storage and Transportation

Propanal should be stored in stainless steel, baked phenolic-lined steel, or aluminum containers to avoid contamination with metal salts present in the construction materials, which at >3 ppm can cause propanal condensation. Stainless steel is most commonly employed. Condensation can also be caused by light and oxygen, and it is particularly troublesome in high-purity samples. Storage at low temperatures is not protective.

Propanal is highly reactive and forms explosive peroxides on exposure to air. Storage under a nitrogen atmosphere is recommended; addition of water also reduces peroxide formation and provides stabilization against metal-catalyzed condensations [53]. Contact with other oxidizing materials should be avoided. Caution is advisable with emptied containers containing residual vapors, which may explode on ignition.

Propanal is a volatile and flammable liquid that can present a flammability hazard under normal storage conditions. It has a relatively high vapor pressure and low flash point. Important flammability data are given in Chapter 3.

8. Toxicology and Occupational Health

Acute exposure through inhalation of propanal vapor causes extreme irritation of the eyes and may irritate the respiratory tract. High concentrations may also cause drowsiness, headache, nausea, and difficulty in breathing. Contact of the liquid with the eyes or skin causes burns. Ingestion is irritating to the gastrointestinal tract, causing burning in the mouth and throat, nausea, abdominal discomfort, and diarrhoea [55].

The oral acute toxicity data reported for rats is LD_{50} 1410 mg/kg; specific toxic effects have not been described [56]. The oral LDLo for mice is 800 mg/kg [55]. The dermal LD_{50} for rabbits is 5040 mg/kg, showing slight absorption; no toxic

effects were noted [56]. Severe skin irritation resulted with guinea pigs.

Propanal is a severe irritant to rabbit eyes [56]. The inhalation LCLo for rats is 8000 ppm (4 h) and in humans the LC_{50} is 21.8 g/m^3 [55]. Propanal is an anesthetic when inhaled in relatively high amounts by rats [57]. Preliminary edema was the cause of death in mice and rabbits after inhalation of high levels of propanal [57]. Exposure of rats to vapor (0, 90, or 1300 ppm, 6 h/d, for 20 consecutive exposures) showed the lowest-observed-effect level (LOEL) at 1300 ppm (liver damage) and the no-observed-effect level (NOEL) at 90 ppm [53].

TLV and OSHA PEL values have not been established for propanal.

9. References

1. Comline News Service: Comline Industrial Machinery & Mechanical Engineering, Tokyo, Feb. 18, 1988, p. 3.
2. SRI International: *Chemical Economics Handbook.*
3. Design Institute for Physical Property Data (DIPPR), ONLINE data base, Project 801, Source File Tape, revision dates Aug. 1988, 1989, 1990.
4. Thermodynamics Research Center: *TRC Thermodynamic Tables – Non-Hydrocarbons (Loose-leaf data sheets),* The Texas A & M University System, College Station, Texas 1980.
5. J. M. Sorensen, W. Arlt: "Liquid-liquid Equilibrium Data Collection, Binary Systems," *DECHEMA Chemistry Data Series,* **vol. V, part 1,** Schon & Wetzel, Frankfurt/Main 1979.
6. Thermodynamics Research Center: *TRC Thermodynamic Table,* 23-2-1-(1.1100)-k, p. 1, Dec. 31, 1961, p. k-5300.
7. N. I. Sax, R. J. Lewis, Sr.: *Dangerous Properties of Industrial Materials,* 7th ed., Van Nostrand Reinhold, New York 1989, p. 2892.
8. Union Carbide, US 4 762 817, 1988 and US 4 876 402, 1989 (J. E. Logsdon, R. A. Loke, J. S. Merriam, R. W. Voight).
9. Davy McKee, EP 73 129, 1983 (M. W. Bradley, A. G. Hiles, J. W. Kippax).
10. Union Carbide, CA 1 137 519, 1982 (C. C. Pai).
11. Davy McKee, EP 74 193, 1983 (M. W. Bradley, N. Harris, K. Turner).
12. Centrul de Chimie Timisoara, RO 84 438, 1984 (R. Valceanu, V. D. Gazdac).
13. W. K. Langdon, E. J. Schwoegler, *Ind. Eng. Chem.* **43** (1951) 1011.
14. Imperial Chemical Industries, DE 1 237 092, 1967 (C. B. Cotterill).
15. Intreprinderea "Plafar", RO 87 470, 1985 (E. V. Ceausescu, O. Costisor, T. Cadariu, M. Brezeanu).
16. Babcock-Hitachi, JP 60 181 032, 1984 (T. Kamiguchi, K. Arikawa, R. Yamada, H. Tanimoto).
17. Institute of Chemical Physics, Academy of Sciences, SU 793 988, 1981 (S. A. Maslov, E. A. Blyumberg, G. N. Gvozdovskii, Yu. N. Koshelev).
18. F. C. Whitmore: *Organic Chemistry,* Van Nostrand Company, New York 1951, pp. 123, 220.
19. Intreprinderea "Plafar", RO 89 953, 1986 (E. V. Ceausescu, M. Brezeanu, O. Costisor, T. Cadariu).
20. BASF, DE 2 507 461, 1975 (F. Merger, S. Winderl, H. Toussaint).
21. VEB Leuna-Werke "Walter Ulbricht", DE 1 952 738, 1970 (H. D. Eilhauer, G. Reckling).
22. Commercial Solvents, US 3 478 115, 1969 (J. B. Bronstein, Jr.).
23. Bayer, DE 2 813 201, 1979 (O. Immel, H. H. Schwarz, H. Quast).
24. Ruhrchemie, DE 2 855 504, 1980 (W. Bernhagen et al.).
25. BASF, EP 58 927, 1982 (F. Merger, H. J. Foerster).
26. Hoechst, DE 3 740 293, 1989 (G. Diekhaus, H. Kappesser).
27. *Chem. Week* (1988) Jan. 20, 64.
28. A. S. Bakshi, *Oil Gas J.* **83** (1985) no. 47, 99.
29. BASF, DE 2 727 330, 1979 (W. Schoenlebe, H. Hoffmann, W. Lengsfeld).
30. P.-Y. Blanc, A. Perret, F. Teppa, *Helv. Chim. Acta* **47** (1964) no. 2, 567 – 575.
31. BASF, US 4 270 006, 1981 (G. Heilen, A. Nissen, O. Woertz).
32. Eastman Kodak, US 4 316 990, 1982 (D. L. Morris).
33. Amoco Corp., JP 61 15737, 1986.
34. Rohm & Haas, US 4 490 476, 1984 (R. J. Piccolini, M. J. Smith).
35. Kh. M. Minachev, O. K. Atal'yan, M. A. Markov, *Izv. Akad. Nauk SSSR, Ser. Kim.* **12** (1978) 2731 – 2737.
36. BASF, DE 2 625 541, 1977 (A. Nissen et al.).

37. BASF, DE 3 319 430, 1984 (W. Gramlich et al.).
38. Union Carbide Chemicals and Plastics, "UCAR Esters for Coating Applications," F-48589, Danbury, Conn., Sept. 1984.
39. Eastman Kodak, US 4 225 726, 1980 (D. L. Morris, A. W. McCollum).
40. G. Fouquet, F. Merger, R. Platz, *Liebigs Ann. Chem.* (1979) 1591 – 1601.
41. Marathon Oil, US 4 169 110, 1979 (J. M. Holovka, E. Hurley, Jr.).
42. Mitsubishi Chemical Ind., JP 62 252 746, 1987 (S. Imaki, Y. Takuma, M. Oishi).
43. Nippon Oils and Fats Co., EP 142 868, 1985 (M. Koyama et al.).
44. Nippon Oils and Fats Co., JP 62 164 653, 1987 (M. Koyama, F. Takahashi).
45. Nepera, EP 371 615, 1990 (D. Feitler, H. Wetstein).
46. E. I. Du Pont de Nemours & Co., US 2 452 187, 1948 (W. F. Gresham).
47. Phillips Petroleum Co., US 2 412 437, 1946 (C. R. Wagner).
48. VEB Leuna-Werke "Walter Ulbricht", DE 1 272 915, 1968 (K. Smeykal, H. J. Naumann).
49. Eastman Kodak, FR 1 524 596, 1968 (J. C. Fleischer, J. W. Reynolds, S. H. Young, C. R. Hargis).
50. Union Carbide, US 4 277 627, 1981 (D. R. Bryant, E. Billig).
51. Union Carbide Chemicals and Plastics Company, Product Specification, 1-4A3-1h, Danbury, Conn., Aug. 1, 1989.
52. Eastman Kodak, Eastman Chemical Products, Sales Specification No. 3827-7, Kingsport, Tenn., Sept. 1989.
53. Eastman Kodak, Eastman Chemical Products, Material Safety Data Sheet for Propanal, Kingsport, Tenn., April 24, 1990.
54. U.S. Production and Sales, U.S. Tariff Commission, "Synthetic Organic Chemicals," U.S. Government Printing Office, Washington, D.C., 1988.
55. Union Carbide Chemicals and Plastics Company,Solvents and Coatings Materials Division, Material Safety Data Sheet, Danbury, Conn., Aug. 29, 1990.
56. D. V. Sweet (ed.): "Registry of Toxic Effects of Chemical Substances," *DHHS (NIOSH) Publ. U.S.* **4** (1987) no. 87 – 114, 3862.
57. *Patty's* 3rd ed., **vol. II A,** pp. 2629 – 2669.

Propane → Hydrocarbons

Propanediols

Carl J. Sullivan, ARCO Chemical Company, Newtown Square, Pennsylvania 19073, United States

1. 1,2-Propanediol and Higher Propylene Glycols . . . 1
1.1. Physical Properties . . . 1
1.2. Chemical Properties . . . 3
1.3. Production . . . 4
1.4. Quality Specifications . . . 4
1.5. Transportation and Storage . . . 5
1.6. Economic Aspects . . . 5
1.7. Uses . . . 5
1.8. Derivatives . . . 6
2. 1,3-Propanediol . . . 6
2.1. Properties . . . 7
2.2. Production . . . 7
2.3. Uses . . . 8
3. Toxicology . . . 8
4. References . . . 9

1. 1,2-Propanediol and Higher Propylene Glycols

1,2-Propanediol, [*57-55-6*], propylene glycol, $HOCH_2CH(CH_3)OH$, is very similar to ethylene glycol in its physical and chemical properties (→Ethylene Glycol). The first reported description of propylene glycol was by Wurtz in 1859 [1].

Industrial-scale synthesis of propylene glycol from propylene oxide (→Propylene Oxide) and water began in the 1930s. Current production uses this same process, which leads simultaneously to di- and tripropylene glycols. The worldwide capacity for propylene glycol is estimated to be 9×10^5 t (1990) [2]. Production may be lower because some propylene glycol production units are designed to permit switching to ethylene glycol output depending upon relative demand for the glycols.

Propylene glycol is used in many diverse applications such as unsaturated polyesters (→Polyester Resins, Unsaturated) for thermoset composites, food chemistry, food processing equipment, cosmetics, and pharmaceuticals. A growing application is in the field of deicers and automotive antifreeze components (→Antifreezes).

1.1. Physical Properties

Propylene glycol is a clear, colorless, strongly hygroscopic liquid. The higher propylene glycols (di- and tripropylene glycol) have similar properties. Dipropylene glycol [*110-98-5*] exists as three isomers differing in the positions of the methyl substituents, and tripropylene glycol [*24800-44-0*] is typically a mixture of several isomers. All three glycols are moderately viscous liquids that do not readily crystallize, but solidify to vitreous masses upon cooling. Fundamental physical properties of these diols are listed in Table 1 [3–9].

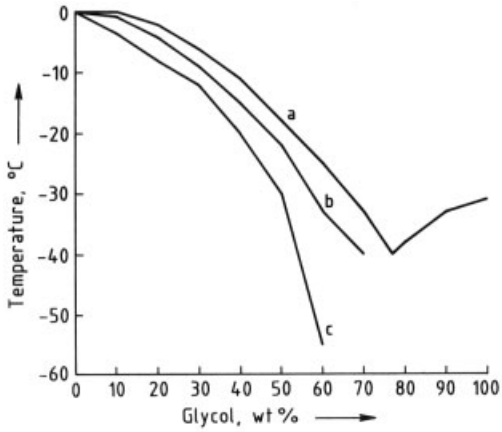

Figure 1. Solidification temperatures for aqueous propylene glycols
a) Tripropylene glycol; b) Dipropylene glycol; c) Propylene glycol

Propylene glycol is readily miscible with water and other polar solvents (e.g., alcohols and acetone). Propylene glycols are good solvents for many polar organic substances, such as phenols, alcohols, dyes, natural products, and some resins. Propylene glycol is insoluble to sparingly soluble in such nonpolar solvents as petroleum ethers, benzene, aromatic hydrocarbons, and carbon tetrachloride.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_163

Table 1. Physical properties of propylene glycols

Property	Propylene glycol	Dipropylene glycol *	Tripropylene glycol *
M_r	76.10	134.18	196.26
bp, °C	187.9	232.8	271
mp, °C	<-60	<-40	<-30
d_{20}^{20}	1.036	1.023	1.019
Refractive index n_D^{20}	1.4328	1.4415	1.4449
Vapor pressure (20 °C), kPa	0.011	<0.001	<0.001
Heat of vaporization at bp, kJ/kg	711	400	330
Specific heat capacity (20 °C), kJ kg^{-1} K^{-1}	2.49	2.36	2.26
Electrical conductivity (20 °C), µSm	4.4	0.19	
Cubic coefficient of expansion (20 °C), K^{-1}	7.2×10^{-4}	7.4×10^{-4}	7.7×10^{-4}
Heat conductivity (20 °C), Wm^{-1} K^{-1}	0.20	0.16	0.11
Dynamic viscosity (20 °C), mPa · s	56	107	84
Surface tension (20 °C), mNm	38	35	35
Dielectric constant (20 °C)	28	21	14
Flash point (DIN 51 758), °C	103	120	149
Ignition temperature (DIN 51 794), °C	410	350	260
Lower explosion limit in air, vol %	2.6	1.26	1.3

* Physical properties may be affected by isomer ratios.

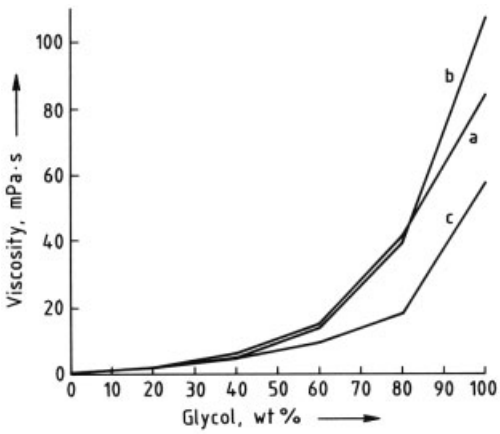

Figure 2. Viscosity of aqueous propylene glycols
a) Tripropylene glycol; b) Dipropylene glycol; c) Propylene glycol

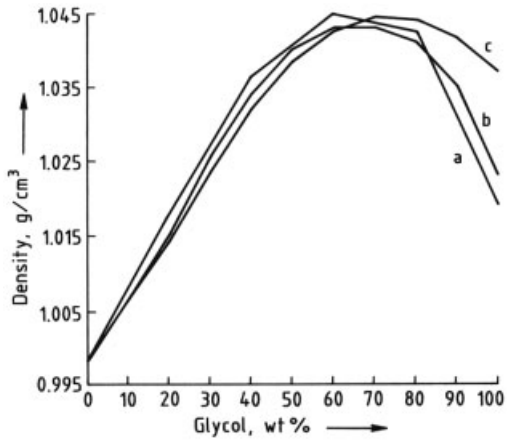

Figure 3. Density of aqueous propylene glycols at 20 °C
a) Tripropylene glycol; b) Dipropylene glycol; c) Propylene glycol

Propylene glycol greatly reduces the freezing point of water and is therefore being developed as an alternative to ethylene glycol in automotive antifreeze systems [10], [11] and aircraft deicers [12], [13]. The solidification temperatures of propylene glycol and its higher homologues at varying water ratios are plotted in Figure 1. The viscosities and densities of aqueous propylene glycol solutions are plotted in Figures 2 and 3 [4].

Vapor pressure versus temperature curves for the three propylene glycol homologues are plotted in Figure 4. Propylene glycol forms azeotropes with the solvents listed in Table 2 [14].

1.2. Chemical Properties

The chemical properties of propylene glycol are determined predominantly by its hydroxyl groups, and its reactions are typical of alcohols (→ Alcohols, Aliphatic). Propylene glycol condenses with carboxylic acids at elevated temperature to yield esters and water. It also reacts readily with isocyanates and acid chlorides to yield carbamates and esters, respectively.

Propylene glycol undergoes polycondensation reactions with diacids to yield polyesters. Predominant among such applications is the re-

Table 2. Azeotropes with propylene glycol [14]

Cosolvent	Cosolvent *bp*, °C	Azeotrope *bp*, °C	Propylene glycol, wt %
o-Nitrophenol	217	186	>62
Aniline	184	179.5	43
Toluene	110.6	110.5	1.5
o-Xylene	144.4	135.8	10
Dodecane	216	175	67
N,N-Dimethyl-*p*-toluidine	185.3	174	37
Dibutyl ether	142	136	
Tetradecane	252	179	70
Methyl hexyl ketone	173	<170	
N-Methylaniline	196	<181	>46
N,N-Dimethylaniline	194	<177	>45
Propylene glycol	187.9		

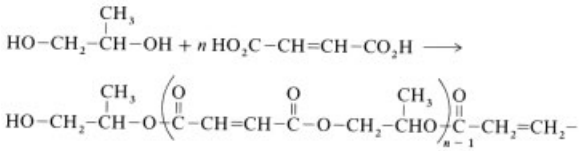

action with maleic anhydride and other diacid components to yield unsaturated polyesters [15], [16] (see above).

Propylene glycol reacts with propylene oxide to form dipropylene glycol, tripropylene glycol, and polyether polyols. The resulting polypropylene glycol ethers are important industrial building blocks for polyurethane foams and elastomers (→Polyurethanes) [17], [18].

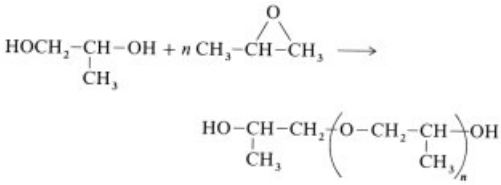

Because of its 1,2-diol structure, propylene glycol also undergoes a variety of interesting cyclization reactions. Cyclic acetals and ketals are formed with aldehydes and ketones, respectively [19].

At elevated temperature in the presence of an acidic catalyst, propylene glycol can undergo dehydration to yield cyclic ethers. Very low levels of these dehydration products are common in aqueous byproduct streams of polycondensation reactions involving propylene glycol [20].

Propylene glycol is unusual in that the commercial product is a racemic mixture of two chiral isomers. Only small quantities of chiral material are available from specialty chemical distributors.

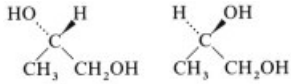

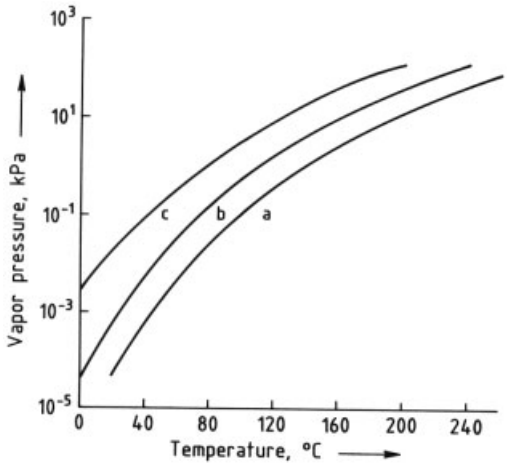

Figure 4. Vapor pressure versus temperature for propylene glycols
a) Tripropylene glycol; b) Dipropylene glycol; c) Propylene glycol

1.3. Production

The only industrial process for manufacturing propylene glycol is *direct hydrolysis of propylene oxide* with water. Dipropylene glycol and tripropylene glycol are formed by sequential addition of propylene oxide to propylene glycol. Consequently, all three products are produced simultaneously and separated by distillation.

Propylene oxide and water are combined in the initial stage of the process at a molar ratio of 1 : 15 at an initial temperature of 125 °C and a pressure of approximately 2 MPa (Fig. 5). Due to the exothermic reaction [21], the reactor effluent temperature will typically rise to 190 °C. At the above ratio of water to oxide, the resulting propylene glycol – dipropylene glycol – tripropylene glycol mixture is approximately in the ratio of 100 : 10 : 1. Higher propylene glycol ratios are attained by increasing the water-to-oxide ratio; however, such an increase also increases recycle rates, lowers throughput, and raises energy costs.

The water – glycol mixture exits the reactor zone (a) at about 200 °C and is stripped of water in dehydration columns (b). Three successive vacuum distillations (c – e) separate high-purity propylene glycol, dipropylene glycol, and tripropylene glycol. The residue represents a mixture of higher glycols, which have some limited commercial use. Waste from this process is extraordinarily low. Because of the very low toxicity of propylene glycol, dipropylene glycol, and tripropylene glycol, these materials represent a minimal hazard to the environment. All are readily biodegradable.

Alternative routes to propylene glycol have been investigated. Acetoxidation of propene to yield glycol acetates followed by hydrolysis to the glycol has been reviewed [22]. This technique involves lower production costs than the propylene oxide route, but fixed capital investment costs are much higher. Direct hydroxylation of propene to propylene glycol using oxygen in the presence of osmium catalysts has been reported [23–25]. However, because of the ready availability of propylene oxide, the simplicity of the current process, and the extremely low level of waste generated by the propylene oxide route, no alternative process has yet been commercialized.

1.4. Quality Specifications

At least two commercial grades of monopropylene glycol exist – a high-purity industrial grade, and a higher-purity grade called "super pure" or "U.S.P." grade. Typical assays for the two grades of propylene glycol are the same; however, the U.S.P. grade is used in applications such as pharmaceuticals, cosmetics, and food additives, and entails higher quality standards (see Table 3) [4].

Table 3. Typical properties of commercial propylene glycol [4], [26]

Property	Industrial grade	U.S.P. grade
Purity, %	99.5	99.5
d_{20}^{20}	1.0377	1.035 – 1.037
Distillation initial *bp*, °C	185	185
dry point, °C	190	189
Acidity, wt % as acetic acid	0.005	0.002
Chlorides, ppm as Cl (max.)	10.0	1.0
Iron, ppm (max.)	1.0	0.5
Heavy metals, ppm [a]		5.0
Arsenic, ppm as As (max.)		3.0
Sulfate, wt % (max.)		0.006
Water, wt % (max.)		0.2

[a] All metal residues are calculated as if they were lead.

1.5. Transportation and Storage

Propylene glycol, dipropylene glycol, and tripropylene glycol are stable materials that require no special handling procedures. They are noncorrosive and can be transported or stored in stainless steel, aluminum, and lined steel containers. Carbon steel is also acceptable, although slight iron contamination may occur upon extended periods of storage. U.S.P.-grade propylene glycol is typically stored in stainless steel. The glycols are hygroscopic, and should be protected from unnecessary exposure to the atmosphere.

1.6. Economic Aspects

Worldwide capacity for propylene glycol production is summarized in Table 4. These capacity figures are flexible, because the same facilities can also be used to produce ethylene glycol depending upon market pressures. Actual production in the United States in 1989 is estimated

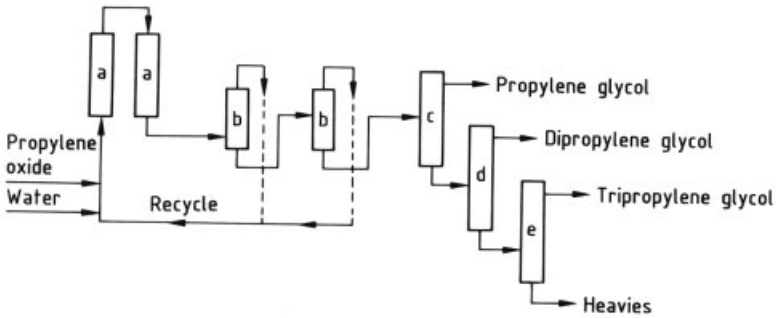

Figure 5. Schematic of propylene glycol production
a) Reactor; b) Evaporator (typically 2 – 4 columns; c) Propylene glycol distillation column; d) Dipropylene glycol distillation column; e) Tripropylene glycol distillation column

to have been 3.65×10^5 t, approximately 95 % of capacity [27].

Table 4. Worldwide capacity for propylene glycol production in 1990 [2]

Region	Number of manufacturers	Capacity, 10^3 t/a
North America	5	384
Western Europe	6	361
East Asia	7	118
Total	18	863

1.7. Uses

Propylene Glycol. The single largest use of propylene glycol is in the manufacture of unsaturated polyester resins. It is reacted with saturated and unsaturated carboxylic diacids such as isophthalic acid and maleic anhydride. The resulting resins are dissolved in styrene or another polymerizable monomer, then combined with filler, chopped glass, a peroxide polymerization initiator, and other additives, and cured to yield a hard, cross-linked, thermoset composite [15], [16], [21]. The end-use applications for such products include automotive plastics, fiberglass boats, and construction. Roughly 45 % of the propylene glycol produced goes to this one broad application [5].

Propylene glycol has many diverse applications in the food and pharmaceutical industries because of the "Generally Regarded As Safe" (GRAS) status conferred on it by the FDA in the United States. Propylene glycol is used as a humectant, solvent, and preservative in food and pet food products. In the manufacture of food products, propylene glycol is used as a lubricant for machinery, as a solvent in food processing, in food wraps, and as an antifreeze agent in machinery cooling water. Propylene glycol is also used as an emollient, softening agent, and humectant in skin- care products and cosmetics, and as a carrier for livestock medicinal products [4], [7].

Propylene glycol is utilized in aircraft deicing and anti-icing fluids because of its proven performance, low toxicity, ready biodegradability, and environmental acceptance [13]. It offers freeze protection similar to that of ethylene glycol, the principle component of automotive antifreeze solutions. Automotive antifreeze compositions incorporating propylene glycol have been investigated [10], [11], [28].

Propylene glycol is useful as a lubricant in combination with di- and tripropylene glycols. It is commonly employed as a humectant for tobacco. Propylene glycol is also introduced into many latex paints as a freeze – thaw protector, and to provide evaporation control in hot and dry environments. Propylene glycol is reacted with fatty acids or long- chain carboxylic acids to produce ester lubricants, emulsifiers, and plasticizers [29–31].

Propylene glycol is a precursor of many polyether polyols used in the urethane foam, elastomer, adhesives, and sealants industry, whereby the compound is reacted with propylene oxide and/or ethylene oxide in order to generate a low molecular mass polyether [17], [18].

Dipropylene and Tripropylene Glycols. Both glycols are used in hydraulic fluids and brake fluids because of their high stability and low corrosion properties. They are also used in inks and some paints because of their good solvency for pigments and binders, very low evaporation rates, and low toxicity. Stamp pad inks are a prime example [32]. Dipropylene glycol has some application in refining operations because it facilitates the separation of aromatic and aliphatic hydrocarbons. Di- and tripropylene glycols are precursors of low molecular mass esters and polyesters, which have utility as plasticizers and lubricants. Both glycols are also used to make polyester polyols for urethane applications.

1.8. Derivatives

The major end-use derivatives of propylene glycol, dipropylene glycol, and tripropylene glycol are unsaturated polyester resins as well as low molecular mass ester and polyester plasticizers. Propylene glycol-initiated polyethers are also modest-volume commodity products.

The class of materials known as glycol ethers (propylene glycol alkyl ethers) is commonly referred to as propylene glycol derivatives, but most of these are actually propylene oxide-derived products because of the ease with which they can be prepared from the oxide. Nevertheless, some glycol ether products actually are derived from the glycol – most notably propylene glycol mono-*tert*-butyl ether, which is produced by reacting propylene glycol with isobutene [33]. These derivatives find application predominantly in coatings and cleaners.

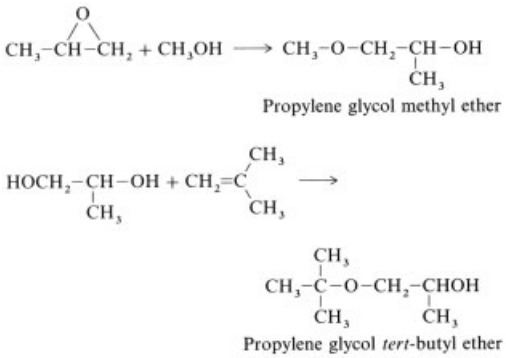

2. 1,3- Propanediol

1,3-Propanediol [*504-63-2*], trimethylene glycol, $HOCH_2CH_2CH_2OH$, is produced on a much smaller scale than its isomer, propylene glycol. In spite of some interesting areas of application, total production remains relatively small. The difficulties of manufacturing the product result in a poorly competitive price structure relative to other diols.

2.1. Properties

Key physical properties of 1,3-propanediol are as follows [3], [34]:

M_r	76.10
mp	−27 °C
bp (101.3 kPa)	214 °C
bp (0.7 kPa)	94 °C
Vapor pressure (20 °C)	0.008 kPa
Vapor pressure (100 °C)	0.98 kPa
d_{20}^{20}	1.0554
d_{4}^{20}	1.0529
Refractive index n_D^{20}	1.4389
Kinematic viscosity	46 m^2/s
Flash point	80 °C

1,3-Propanediol is a clear, colorless, odorless liquid that is miscible with water, alcohols, ethers, and formamide. It is sparingly soluble in benzene and chloroform.

The chemical properties of 1,3-propanediol are typical of alcohols. Like 1,2-propanediol, 1,3-propanediol condenses with carboxylic acids at elevated temperature to yield esters. It also reacts with isocyanates and acid chlorides to yield urethanes and esters, respectively. Unlike 1,2-propanediol, 1,3-propanediol has two primary hydroxyl groups with equivalent reactivity.

1,3-Propanediol readily forms ethers. 3,3′-Dihydroxydipropyl ether forms upon continued reflux of the diol. 1,3-Propanediol reacts with aldehydes and ketones, often in the presence of acidic catalysts, to form 1,3-dioxanes [35], [36]:

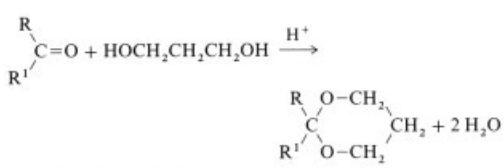

Over aluminum oxide at temperatures above 250 °C, 1,3-propanediol decomposes to allyl alcohol, propanol, and other products.

1,3-Propanediol reacts with diacids to form polyesters. The polyester with terephthalic acid is reported to have a crystalline melting point of 220 °C [37]. Polyurethanes can also be synthesized from 1,3-propanediol.

2.2. Production

1,3-Propanediol is prepared by a two-step process involving the hydrolysis of acrolein to 3-hydroxypropanal followed by hydrogenation [3], [38]:

$$CH_2{=}CH{-}CH{=}O + H_2O \longrightarrow HO{-}CH_2CH_2CH{=}O$$

$$\xrightarrow[\text{Cat.}]{H_2} HO{-}CH_2CH_2CH_2{-}OH$$

Hydrolysis is carried out under weakly acidic conditions in water containing initially ca. 20 % acrolein. Higher concentrations of acrolein tend to lead to greater amounts of undesired byproducts as a result of reaction between acrolein and hydroxypropanal.

3-Hydroxypropanal can be hydrogenated in the aqueous phase directly; however, the preferred technique is to extract the aldehyde into an organic solvent – particularly 2-methylpropanol – and then hydrogenate the aldehyde to yield the diol. Hydrogenation is conducted with Raney nickel under pressure in the aqueous phase and with nickel-supported catalysts at 2 – 4 MPa and 110 – 150 °C in the organic phase [39]. The diol is subsequently separated from solvent and water by distillation.

The yield of desired product by this route is relatively low – approximately 45 %. Alternative techniques for synthesizing the diol have appeared in the patent literature. Hydroformylation of ethylene oxide followed by hydrogenation yields 1,3-propanediol in good yield (92 %), but a high catalyst concentration and a very large excess of solvent render the process uneconomical [40]. More recently, hydroformylation of ethylene oxide directly to 1,3-propanediol with a rhodium – phosphine catalyst system has been disclosed [41]. The reaction of ethylene with formaldehyde and carboxylic acids has also not been commericialized because of low selectivity [42].

Handling of 1,3-propanediol is similar to that of propylene glycol. The material is noncorrosive and can be transported in stainless steel, aluminum, or lined steel vessels.

Quality analysis is readily accomplished by gas chromatography. A good technical grade generally displays a purity between 98.5 and 99.5 %.

The environmental impact of this product is expected to be low because of its low toxicity and ready biodegradability. The handling of acrolein in the production process must be carefully controlled.

2.3. Uses

As a diol, 1,3-propanediol is subject to many of the same polymeric applications as other low molecular mass diols (e.g., ethylene glycol, propylene glycol, and 1,4-butanediol). However, its relatively high price limits its use to applications requiring very specific performance characteristics. It is a raw-material source for 1,3-dioxanes. 1,3-Propanediol-bis(4-aminobenzoate) [*57609-64-0*] can be used as a chain extender in polyurethane elastomers. This bisbenzoate, which can also be synthesized from 1,3-dichloro-propane, finds other applications as a cross-linking agent in epoxy formulations and as a rubber additive.

3. Toxicology

1,2-Propanediol. The toxicological properties of 1,2-propanediol have been reviewed [3], [4], [43]. The substance has such a low order of acute toxicity that it is considered practically nontoxic. Since 1942 it has been classified as an acceptable ingredient in pharmaceutical applications, and the United States FDA permits its use in food products and cosmetics. Official requirements for 1,2-propanediol used as an adjuvant in pharmaceuticals can be found in international and national pharmacopeias.

Acute oral LD_{50} values for rats range from 21.0 to 33.7 g/kg. For mice the values range from 23.9 to 31.8 g/kg. Tests with other animals (guinea pig, rabbit, mouse, and dog) yield LD_{50}

values within the same range [43]. These values correspond to ingestion of 1.5 – 2.0 L by a typical adult human.

Administration of single, almost lethal doses to laboratory animals have caused balance disorders resulting from central nervous system depression. Significant impairment of motion was observed in rats after oral administration of 19.5 g/kg. A similar level of impairment was observed with an ethanol dosage of only 1.5 g/kg.

Propylene glycol is not injurious to the eye of rabbits, and it is not known to have caused any eye injuries in humans. Slight transitory stinging and tear formation may result from direct contact with the eye.

Propylene glycol does not produce any significant irritation effects in skin contact with rabbits or humans, although skin irritation has been reported in some individuals when high concentrations were held in contact with skin under closed conditions [44], [45].

Subchronic and chronic studies also indicate that propylene glycol has a low order of toxicity. A two-year study on rats demonstrated that incorporation of 5 % propylene glycol into the diet did not increase or accelerate mortality. No tissue damage was noted at 7.5 % of the diet for 20 weeks, but at 20 % of the diet the rats exhibited nonspecific minor changes in liver and kidney cell structure [46].

Studies also indicate that propylene glycol does not present a health risk after inhalation. Monkeys and rats exposed to propylene-glycol-saturated atmospheres for 12 – 18 months showed no adverse effects [47].

In vivo and in vitro mutagenicity tests show no mutagenic behavior [48]. Reproductive studies with mice, rats, hamsters, and rabbits show no effects upon fetal development, and no teratogenic effects. In a reproductive study with rats, a level of 7.5 % propylene glycol in food showed no effects over six generations [49].

In the body, propylene glycol is metabolized mainly to carbon dioxide via pyruvic acid, but it also enters into carbohydrate metabolism and promotes glycogen formation. At higher dose levels, some propylene glycol is excreted in the urine.

No permissible exposure levels have been established, because none appear to be necessary.

Dipropylene Glycol and Tripropylene Glycol. These two glycols have not been studied as extensively as propylene glycol, but both are considered to possess a low order of toxicity. The oral LD_{50} values in rats for dipropylene glycol and tripropylene glycol are 15 g/kg and 12 g/kg, respectively. In a subchronic study involving 5 wt % dipropylene glycol in drinking water, rats were not affected. However, at 10 % some rats died, and there was evidence of liver and kidney involvement.

Both di- and tripropylene glycol demonstrate negligible eye and skin irritation. Continuous or repetitive application of the two glycols provides no indication that toxic quantities are absorbed through the skin. Because of low vapor pressures and low systemic toxicity, inhalation hazards appear negligible.

1,3-Propanediol. 1,3-Propanediol is mildly toxic. The oral LD_{50} for rats is 14 – 15 g/kg, whereas intraperitoneal LD_{50} values for rats and rabbits are 6 – 7 g/kg and 3 g/kg, respectively [50]. Prolonged administration to chickens inhibits growth [51]. Unlike its isomer, propylene glycol, 1,3-propanediol is not permitted for use in food or livestock feed products.

4. References

1. A. Wurtz, *Justus Liebigs Ann. Chem.* **55** (1859) no. 3, 406.
2. *Directory of Chemical Producers,* SRI International, 1990.
3. *Ullmann,* 4th ed., **19,** 425 – 432.
4. ARCO Chemical Company, Product Brochure, Propylene Glycols, USA, 1983.
5. Dow Chemical, Product Brochure, Propylene Glycol U.S.P., USA, 1979.
6. *Kirk-Othmer,* 3rd ed., **11,** 933 – 956.
7. E. W. Flick (ed.): *Industrial Solvents Handbook,* Noyes Data Corp., Park Ridge, NJ, 1985.
8. G. O. Curme, F. Johnston: *Glycols,* Reinhold Publ. Co., New York 1952.
9. J. Mellon: *Polyhydric Alcohols, Spartan Books,* Washington 1962.
10. F. E. Mark, W. Jetten, *ASTM Spec. Tech. Publ.* **88** (1986) 61 – 77; *Chem. Abstr.* **105** (1986) 45 989.
11. C. Fiaud, P. Netter, M. Tadjamoli, M. Tainmann, *ASTM Spec. Tech. Publ.* **887** (1986) 162 – 175; *Chem. Abstr.* **105** (1986) 45 990 a.

12. S. H. Bloom, M. A. Bloom, US 4 585 571, 1986.
13. Kilfrost, Ltd., Product Brochure, Kilfrost ABC-3, 1989.
14. L. H. Horsley: "Azeotropic Data-III," *Adv. Chem. Ser.* **116** (1973) .
15. J. Selley: "Polyesters, Unsaturated," *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vol. 12,** Wiley Interscience, New York 1988, pp. 156 – 290.
16. Amoco Chemicals Company, Product Brochure, Processing Unsaturated Polyesters on Amoco Isophthalic Acid, Bulletin IP-430, USA, 1989.
17. S. D. Gagnon: "1,2-Epoxide Polymers," *Encyclopedia of Polymer Science and Engineering,* 2nd ed., **vol. 6,** Wiley Interscience, New York 1986, pp. 275 – 307.
18. G. Woods: *The ICI Polyurethane Book,* 1987.
19. F. A. Meskens, *Synthesis* **7** (1981) 501.
20. E. E. Parker, *Ind. Eng. Chem.* **58** (1966) no. 4, 54.
21. T. McMilland (ed.): *Process Economics Program Yearbook International,* SRI International, 1990.
22. H. W. Scheeline, H. Naka: "Propylene Glycol," *Process Economics Review,* Report No. PEP77-3, SRI International, 1978.
23. Exxon, US 4 533 772, 1985 (R. C. Michaelson, R. G. Austin).
24. Exxon, EP 77 201, 1981 (R. C. Michaelson, R. G. Austin).
25. Gulf Research & Developmemt, US 4 308 409 (C. Y. Wu, T. P. Kobylinski, J. E. Bozik).
26. The United States Pharmacopeia, 21st Revision, Official January 1, 1985, United States Pharmacopeial Convention, Inc., Rockville, MD, 1984.
27. Synthetic Organic Chemicals, United States Production and Sales, 1989, United States International Trade Commission Publication no. 2338, 1990.
28. R. D. Hercamp, R. D. Hudgens, G. E. Coughenour: "Aqueous Propylene Glycol Coolant for Heavy Duty Engines," SAE Technical Paper Ser. no. 900 434, SAE International, 1990.
29. VEB Petrochemisches Kombinat Schwedt, DD 208 478, 1984 (R. Steinhauer, et al.); *Chem. Abstr.* **101** (1984) P 194 972 q.
30. Meiji Seika Kaisha, Ltd., JP-Kokai 61 149 064, 1986 (Y. Takahashi, T. Yoshida); *Chem. Abstr.* **105** (1986) P 189 731 q.
31. Institut Francais du Petrole, FR 2 560 884, 1985 (A. Bre, M. Mollard, M. Osgan); *Chem. Abstr.* **104** (1986) P 150 114 v.
32. Dow Product Brochure, "The Glycol Ethers Handbook," 1981.
33. L. R. Nudy, W. A. Johnston. *Household Pers. Prod. Ind.* **27** (1990) no. 4, 90.
34. *Aldrich Catalog Handbook of Fine Chemicals,* Aldrich Chemical Company, Milwaukee, WI, 1988.
35. J. L. Mateo, O. RuizMurillo, R. Sastre, *An. Quim. Ser C* **80** (1984) no. 2, 178.
36. L. F. Lapuka et al., *Khim Geterotsikl. Soedin.* 1981, no. 9, 1182; *Chem. Abstr.* **95** (1981) 20 039 e.
37. R. W. Lenz: *Organic Chemistry of Synthetic High Polymers,* Interscience, New York 1967, p. 93.
38. R. Hall, E. S. Stern, *J. Chem. Soc.* 1950, 490.
39. Ruhrchemie AG, US 4 094 914, 1978 (W. Rottig et al.).
40. Shell Oil Company, US 3 463 910, 1970 (C. N. Smith, G. N. Schrauzer, K. F. Koetitz, R. J. Windgassen).
41. Hoechst Celanese, US 4 873 378, 1989 (M. Murphy et al.).
42. National Distillers and Chemical Corp., US 4 322 355, 1980 (D. Horvitz, W. D. Bargh).
43. J. A. Ruddick, *Toxicol. Appl. Pharmacol.* **21** (1972) 102.
44. K. Motoyoshi et al.: "The Safety of Propylene Glycol and Other Humectants," *Cosmet. Toiletries* **99** (1984) 83 – 91.
45. R. J. Trancik, H. I. Maibach: "Propylene Glycol: Irritation or Sensitization," *Contact Dermatitis* **8** (1982) 185 – 189.
46. F. Gaunt et al., *Food Cosmet. Toxicol.* **10** (1972) 151.
47. O. H. Robertson et al., *J. Pharmacol. Exp. Ther.* **91** (1947) 52.
48. FDA report PB-245 450 (1974).
49. *Fed. Regist.* **42** (1977) June 17, 30 865.
50. W. Van Winkle, *J. Pharmacol.* **72** (1941) 227.
51. R. D. Creek, *Poult. Sci.* **49** (1970) no. 6, 1686; *Chem. Abstr.* **74** (1971) 74 453.

Propanolamines → Ethanolamines and Propanolamines

Propanols

Anthony J. Papa, Union Carbide Chemicals and Plastics Company, Inc., South Charleston, WV 25303, United States

1.	**Introduction**	1
2.	**Physical Properties**	1
3.	**Chemical Properties**	2
4.	**Production**	4
4.1.	**Production of 1-Propanol**	4
4.2.	**Production of 2-Propanol**	5
5.	**Uses**	7
6.	**Specifications**	8
7.	**Economic Aspects**	8
8.	**Storage and Transportation**	9
9.	**Toxicology and Occupational Health**	10
10.	**References**	10

1. Introduction

The propanols C_3H_7OH, M_r 60.10, comprise two isomers, 1-propanol [*71-23-8*] and 2-propanol [*67-63-0*], also called isopropyl alcohol, of which the latter is industrially the more important. Both are clear, colorless, flammable liquids with a slight odor resembling that of ethanol. They occur in nature in crude fusel oils and as fermentation and decomposition products of various vegetables.

The propanols are used mainly as solvents for coatings; in antifreeze compositions and household and personal products; and as chemical intermediates for the production of esters, amines, and other organic derivatives. 2-Propanol is produced by hydration of propene, while 1-propanol is manufactured by the hydrogenation of propanal, in turn derived from hydroformylation of ethylene. Annual U.S. production of 1-propanol and 2-propanol is estimated in 1988 to have been 98×10^3 t and 630×10^3 t, respectively [1].

2. Physical Properties

The propanols are completely miscible with water and readily soluble in a variety of common organic solvents (e.g., ethers, esters, acids, ketones, and other alcohols). Physical properties of anhydrous 1- and 2-propanol as well as a 91 vol % azeotropic mixture of 2-propanol with water are provided in Table 1. Physical properties of the propanols reflect the position of the hydroxyl group. Associative properties in solution cause the propanols to form azeotropes with a variety of compounds, including aromatics, esters, amines, and ketones. Examples of binary azeotropes of 1- and 2-propanol are given in Table 2.

Freezing points of 1-propanol – water and 2-propanol – water mixtures are plotted in Figure 1. These plots show that the advantages of a substantially lower freezing point for pure 1-propanol relative to 2-propanol are lost in aqueous solutions of the two alcohols.

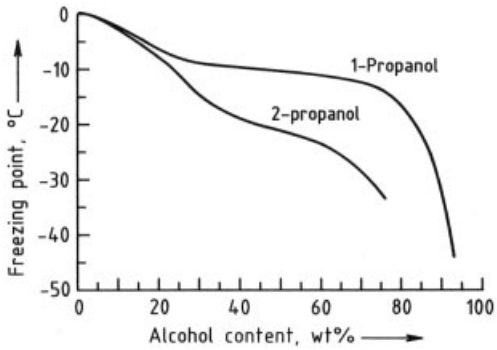

Figure 1. Freezing points of 1- and 2-propanol – water mixtures [2], [17]

3. Chemical Properties

The differences in reactivity between 1-propanol and 2-propanol reflect the influence of the nature of the hydrocarbon radicals in primary and secondary alcohols, respectively. Characteristic reactions of the hydroxyl group include dehydrogenation, oxidation, esterification, amination, and dehydration. The chemical properties of greatest commercial importance are discussed below.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_173

Table 1. Physical properties of propanols [2–15]

	1-Propanol	2-Propanol (anhydrous)	2-Propanol (91 vol % in water)
M_r	60.096	60.096	60.096
Critical temperature, °C	263.63	235.15	
Critical pressure, kPa	5175	4762	
fp, °C	−126.1	−88.5	−50
bp, °C			
101.325 kPa	97.15	82.26	80.40
39.997 kPa	74.479	60.66	
2.0 kPa	20.04		
1.3333 kPa		2.49	
Vapor pressure at 20 °C, kPa		4.4136	4.5
Refractive index at 25 °C	1.3837	1.3752	1.3769
Liquid density at 20 °C, kg/m^3	803.78	785.39	
Liquid viscosity at 20 °C, mPa · s (cP)	2.21	2.37	2.1 (25 °C)
Surface tension at 20 °C, mN/m (dyn/cm)	23.7	21.32	21.40
Flash point (closed cup), °C	15	11.85	18.3
Flash point (Tag open cup), °C	27	17.2	21.7
Autoignition temperature, °C	370.85	398.85	
Lower flammability limit in air, vol %	2.1	2.5	
Upper flammability limit in air, vol %	13.5	12	
Solubility in water at 20 °C, wt %	miscible	miscible	miscible

Table 2. Binary azeotropic mixtures with 1-propanol and 2-propanol [2], [16]

Component	1-Propanol		2-Propanol	
	bp at 101.3 kPa, °C	Alcohol, wt %	*bp* at 101.3 kPa, °C	Alcohol, wt %
Water	88.1	71.8	80.4	87.8
Ethyl acetate	nonazeotrope		75.9	25
Cyclohexane	74.3	20	69.4	32
2-Butanone	nonazeotrope		77.9	32
3-Pentanone	96.0	63	nonazeotrope	
Benzene	77.12	16.9	71.9	33.3
Propyl formate	80.6	9.8	75.5	∼ 36
Hexane	65.65	4	62.7	23
Dioxane	95.3	55	nonazeotrope	
Propyl ether	85.8	32.2	78.2	52

Dehydrogenation (Oxidation). 2-Propanol can be readily dehydrogenated to acetone (→ Acetone, Chap. 4.2.), although the process is being replaced by the cumene – phenol process and by direct oxidation of propene (as developed by Wacker – Hoechst).

$$(CH_3)_2CHOH \xrightarrow{\text{Catalyst}} (CH_3)_2CO + H_2 \quad \Delta H = 67\ \text{kJ/mol}$$

Pure dehydrogenation of 2-propanol can be accomplished in the liquid or gas phase with a zinc – copper catalyst at 300 – 500 °C and 300 kPa to give ca. 90 % selectivity to acetone and 98 % 2-propanol conversion. Selectivity can be increased to 96.8 % by employing a Cu – Zn – Cr catalyst [18]. Dehydrogenation is endothermic and requires high temperatures.

Dehydrogenation can also be conducted oxidatively in the presence of silver or copper catalysts at 400 – 600 °C [19].

Patents describe a 2-propanol dehydrogenation process for the production of a mixture of acetone, methyl isobutyl ketone, and higher ketones [20], [21]. For example, vapor-phase dehydrogenation of an azeotropic mixture of 2-propanol and water over a copper-based catalyst at 220 °C yields a product mixture containing acetone (52.4 %), 2-propanol (11.4 %), methyl isobutyl ketone (21.6 %), diisobutyl ketone (6.5 %), and 4-methyl-2-pentanol (2.2 %) [20].

Partial oxidation of 1-propanol can be achieved with air in the presence of catalysts to form propanal and propionic acid [*79-09-4*]

[22]. Strong oxidizing agents such as nitric acid react exothermically with both alcohols, giving a complex mixture of products.

Esterification. The propanols can be converted into propyl esters by reaction with organic and inorganic acids using the conventional method for esterification in the presence of mineral-acid catalysts. The commercially important *n*-propyl acetate [*109-60-4*] and isopropyl acetate [*108-21-4*] are produced by direct esterification of the corresponding alcohols with acetic acid in the presence of sulfuric acid, *p*-toluenesulfonic acid, methanesulfonic acid [23], [24], or a strong cationic resin [25] as catalyst. A higher-pressure (170 – 310 kPa) continuous process at 110 – 160 °C with continuous removal of product ester and water has been described, which should offer higher production rates [26]. 1-Propanol can also undergo ester interchange with methyl or ethyl acetate in the presence of a strong cationic exchange resin to give *n*-propyl acetate [27].

Reaction of 2-propanol with CS_2 in alkaline solution produces xanthate esters. The method simply involves refluxing the alcohol with alkali-metal hydroxide in trichlorofluoromethane solvent while adding CS_2 [28], [29]. Sodium isopropyl xanthate, $(CH_3)_2CHOCSSNa$ [*140-93-2*], is a herbicide used for weed control.

Reaction of 2-propanol with phosphorus halides proceeds readily to form phosphite esters. Generally, reaction is conducted in a solvent containing an acid scavenger, but a new solventless spray method could have merit [30].

Titanium(IV) isopropoxide [*546-68-9*] and titanium(IV) propoxide [*3087-37-4*] are made by reaction of the corresponding alcohol with $TiCl_4$ in refluxing heptane with removal of HCl coproduct [31]. These esters are especially useful as polymerization and transesterification catalysts.

The propanols also form esters upon reaction with active metals. For example, aluminum isopropoxide [*555-31-7*] and aluminum propoxide can be prepared by contacting aluminum with alcohol vapor [32], [33]. These materials find use as catalysts, in the preparation of aluminum soaps, as dispersants in paint formulations, and as starting materials for ceramics.

Reaction with Ammonia and Amines. Both 1- and 2-propanol can be reductively aminated by reaction with ammonia or lower amines to give propylated amine derivatives. The reactions are normally conducted at 1.0 – 10.0 MPa and 190 – 250 °C using catalysts consisting of nickel, aluminum, cobalt, molybdenum, and/or chromium. Alcohol conversions are usually ca. 90 % [34]. For example, $H_2 - NH_3$ – 2-propanol (12: 8: 1 mole ratio) was passed over a nickel-based catalyst at 110 °C in the vapor phase at a liquid hourly space velocity (LHSV) of $1\,h^{-1}$ to give a mixture of 93.5 % isopropyl amine [*75-31-0*] and 6.3 % diisopropyl amine [*108-18-9*] with a 2-propanol conversion of 93 % [35].

Etherification and Dehydration. 1-Propanol and 2-propanol can be reacted with one, two, or three moles of ethylene oxide, propylene oxide, or both to give a family of glycol ethers having broad utility as solvents. Reaction is generally catalyzed by alkali hydroxide.

Diisopropyl ether [*108-20-3*] can be prepared by liquid-phase dehydration of 2-propanol at 130 – 190 °C and 1.96 – 7.85 MPa over acidic catalysts containing aluminum [36], [37]. Dehydration of 1-propanol to di-*n*-propylether [*111-43-3*] can be accomplished with a cation-exchangeable layered-clay catalyst containing titanium or zirconium at 180 °C [38].

The dehydration of 1- or 2-propanol to give propene has no practical value. Nevertheless, dehydration is most facile with 2-propanol in the presence of mineral acid catalysts at room temperature or higher, conditions that should otherwise be avoided in most instances.

Others. The propanols give hemiacetal adducts readily on addition to aldehydes. The products can in turn be converted to acetals with a second mole of alcohol under dehydration conditions in the presence of an acidic catalyst [39]. Acetals find use as intermediates in the preparation of pharmaceuticals.

2-Propanol can be condensed with such aromatic compounds as toluene and phenol to produce isopropoxylated derivatives.

4. Production

1-Propanol and 2-propanol are produced by very different processes starting from ethylene and propene, respectively.

4.1. Production of 1-Propanol

Propanal, obtained by hydroformylation of ethylene (→ Propanal), is the primary commercial source of 1-propanol. In a second step the propanal is hydrogenated to 1-propanol.

$$CH_2 = CH_2 + CO + H_2 \xrightleftharpoons[\text{pressure}]{\text{Catalyst, heat,}} CH_3CH_2CHO$$

$$CH_3CH_2CHO + H_2 \xrightleftharpoons[\text{pressure}]{\text{Catalyst, heat,}} CH_3CH_2CH_2OH$$

Hydroformylation of Ethylene. Union Carbide Chemicals and Plastics Company, Inc. [40], and Hoechst Celanese practice the two-step oxo process based on rhodium-substituted phosphine-catalyzed low-pressure hydroformylation technology (→Oxo Synthesis, Chap. 3.1.), while Eastman Kodak Company uses a higher-pressure cobalt-based process for production of the intermediate propanal [1]. Typical oxo conditions employed in the low-pressure rhodium-catalyzed oxo process are 90 – 130 °C, <2.8 MPa total pressure, with CO at <380 kPa, H_2 at <1.4 MPa, and <500 ppm rhodium [41]. Conditions of the cobalt oxo process are 110 – 180 °C, 20 – 30 MPa (200 – 300 atm), 1: 1 to 1: 2 H_2: CO ratios, and 0.1 – 1.0 wt % cobalt based on ethylene [42]. The crude propanal is removed by vapor stripping with excess synthesis gas, and the rhodium – triphenylphosphine catalyst is recycled [40]. Traces of CO are removed in a stripping column from the crude, condensed propanal before hydrogenation to prevent poisoning of the hydrogenation catalyst in the second step. The higher selectivity and milder conditions of the rhodium-catalyzed process provide crude 1-propanol for hydrogenation containing fewer impurities, such as higher aldol-condensation products and aldehyde trimers (Tischenko esters).

Hydrogenation of propanal is conventionally carried out as a heterogeneous process in either the vapor or liquid phase over a variety of metal catalysts. Heterogeneous vapor-phase processes are effective at ca. 110 – 150 °C, and 0.14 – 1.0 MPa at a 20: 1 mole ratio of hydrogen to propanal [43–45]. Reductions are accomplished in excess hydrogen, and the heat of reaction is removed by circulating the vapor through external heat exchangers or by cooling the reactor internally [46].

Hydrogen efficiencies are >90 %, with aldehyde conversions as high as 99.9 % and alcohol yields >99 %. The commonly used commercial hydrogenation catalysts include combinations of copper, zinc, nickel, and chromium compounds. Major impurities at vapor-phase process temperatures are dipropyl ether, ethane, and propyl propionate. In addition to reduced zinc oxide and copper oxide catalysts, selectivity enhancers such as alkali and transition metals can be added to significantly reduce the formation of esters (by Tischenko reaction of the starting aldehyde) and ethers [44]. Addition of 1 – 10 % water to the reactor feed also suppresses ether formation [47].

The propyl propionate byproduct formed in the catalytic hydrogenation can be separated and hydrogenolyzed in the presence of reduced CuO – ZnO catalyst at 75 – 300 °C and 9.8 kPa – 9.8 MPa to give 1-propanol as the major product [48].

Liquid-phase heterogeneous processes are conducted at higher pressures. For example, hydrogenation at 95 – 120 °C and 3.5 MPa of hydrogen gives 1-propanol with >99.9 % purity. Nickel- and copper-based catalysts containing molybdenum, manganese, and sodium promoters are preferred [49]. The catalysts are usually supported on alumina.

Other processes have been described for the preparation of 1-propanol; for example, in two steps by isomerization and hydrogenation of propylene oxide [50]; from synthesis gas [51], [52]; from methanol [*67-56-1*] and synthesis gas [53]; and by the Guerbet reaction [54].

4.2. Production of 2-Propanol

There are two major commercial processes for the production of 2-propanol: indirect and direct hydration of propene. Smaller quantities are produced by hydrogenation of acetone.

Indirect hydration involves two steps. In the first, a mixture of mono- and diisopropyl sulfate esters is formed from reaction of propene with sulfuric acid. The reaction is exothermic, yielding about 50 kJ/mol.

$$3\,CH_3CH{=}CH_2 + 2\,H_2SO_4 \rightleftharpoons (CH_3)_2CHOSO_3H + [(CH_3)_2CHO]_2SO_2$$

The sulfate esters are then hydrolyzed to product 2-propanol in a second step.

$$(CH_3)_2CHOSO_3H + [(CH_3)_2CHO]_2SO_2 \xrightarrow{H_2O} 3\,(CH_3)_2CHOH + 2\,H_2SO_4$$

A major byproduct in the indirect process is diisopropyl ether, which is formed from reaction of the sulfate esters with product 2-propanol.

$$(CH_3)_2CHOSO_3H + [(CH_3)_2CHO]_2SO_2 \xrightarrow{(CH_3)_2CHOH} 2\,[(CH_3)_2CH]_2O + 2\,H_2SO_4$$

Other impurities include hydrocarbons, char, polymeric residues (oils), propanal, acetone, and odorous sulfur-containing compounds. Isopropyl sulfate esters are particularly unstable, decomposing at temperatures as low as 50 – 100 °C [55]. Mono- and disulfate esters can be thermally desulfonated to give SO_2, SO_3, and a mixture of propene and acetone.

$$(CH_3)_2CHOSO_3H \xrightarrow{Heat} CH_3CH{=}CH_2 + SO_3 + H_2O$$

$$(CH_3)_2CHOSO_3H \xrightarrow{Heat} CH_3COCH_3 + SO_2 + H_2O$$

The indirect hydration process can be conducted with either strong or weak acid. A schematic for an indirect *weak-acid hydration* is shown in Figure 2.

Propene-containing feed gas (propene content $\leq 60\,\%$) is reacted with ca. 60 % H_2SO_4 in a series of absorption columns (a) at 75 – 85 °C and 0.59 – 1.0 MPa. Production rates and efficiencies are linked to a balance of initial sulfuric acid concentration, temperature, total pressure, and propene partial pressure. High absorber temperatures favor ether formation. The propene is sparged into the lower portion of the absorbers while the H_2SO_4 is fed to the top to establish a countercurrent flow pattern. The "sulfates" product stream is taken from the bottom of the absorbers and fed to vacuum strippers (b) where it is hydrolyzed with steam and the resulting 2-propanol is flashed as overhead product. Product-free acid bottoms (recycle acid) from the strippers are concentrated to about 60 % and recycled to the absorbers. The required reconcentration of acid in this process is less costly than in the strong-acid process, where reconcentration to >90 % is required. Vaporized stripper product is neutralized by scrubbing with caustic (c), and uncondensed vapors (vent gas) are combined with the blow-off gas from the scrubbers and recycled to the reaction.

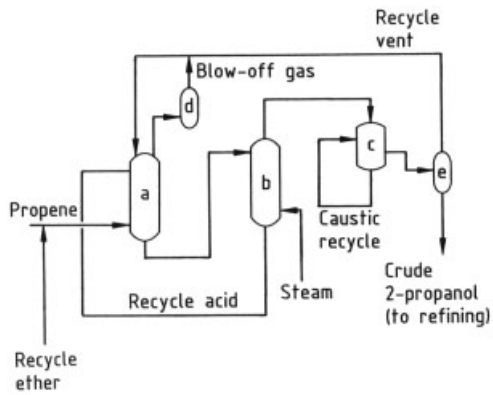

Figure 2. 2-Propanol production by indirect, weak-acid process
a) Absorbers; b) Strippers; c) Caustic scrubber; d) Scrubber; e) Liquifier

Crude 2-propanol is fed to a multicolumn refining system where a constant-boiling mixture (CBM) of 2-propanol is obtained. Byproduct diisopropyl ether is recovered as a lights stream from the first column in the refining train and recycled to the reaction absorbers, where it is hydrolyzed back to product, thus preventing ether accumulation in the process:

$$[(CH_3)_2CH]_2O + H_2O \xrightleftharpoons{H_2SO_4} 2\,(CH_3)_2CHOH$$

Heavy-end concentrated oils are also collected and discarded.

The main problem encountered in purification of 2-propanol is its separation from water. Enrichment of the CBM to >99 % or absolute 2-propanol is accomplished by use of an azeotroping agent, which forms a ternary constant-boiling mixture. Diisopropyl ether and cyclohexane are common entraining agents for 2-propanol enrichment [56]. Undesirable intense odors inevitably arise from sulfur-containing compounds, and this can cause difficulties in such commercial uses as cosmetic products, spray products, and medicinal formulations. Odor is generally removed by contacting the 2-propanol with ion-exchange resins, activated carbon, activated alumina, or metals (e.g., copper, nickel) [57–59].

The weak-acid indirect hydration process suffers from high corrosion rates and disposal problems with water, acid, caustic, and off-gas wastes. A major advantage is the ability to use low-purity propene feed. Advances in waste minimization include replacement of the lead-lined absorbers with corrosion-resistant ceramics and synthetics.

Indirect hydration is still the primary process conducted in the United States by Exxon, Shell Oil, and Union Carbide Chemicals and Plastics [60], while Lyondell Petrochemical produces a smaller amount by hydrogenation of acetone [60]. There are indications that a few companies in Europe and Japan may also be employing this older technology [19].

In the *strong-acid process* reaction is conducted with a high sulfuric acid concentration (> 90 wt %), while the pressure and temperature are low compared to the weak-acid process (1.0 – 1.2 MPa and 20 – 30 °C, respectively). Hydrolysis is accomplished in a separate second stage. Both processes offer 2-propanol selectivities of >90 wt %.

It is reported that for 1 t of 2-propanol produced, 0.8 t propene, 0.35 t H_2SO_4, 3.5 t steam, and 40 – 50 kW · h of electricity are required in an optimized process [61]. It is believed that the strong acid process is not employed as a major commercial source of 2-propanol in the world today because of the requirement for high-purity propene feedstock and waste-disposal problems.

Direct Hydration. The direct hydration of propene has been in commercial use since 1951.

$$CH_3CH{=}CH_2 + H_2O \underset{}{\overset{\text{Catalyst}}{\rightleftharpoons}} (CH_3)_2CHOH \quad \Delta H = -50\ \text{kJ/mol}$$

High pressures and low temperatures over an acidic fixed-bed catalyst characterize this process, causing the exothermic equilibrium reaction to be displaced to the right. Three versions of the direct hydration process are practiced commercially today.

1) Low-temperature (130 – 160 °C), high-pressure (8.0 – 10.0 MPa) vapor – liquid phase hydration over a sulfonated polystyrene ion-exchange resin catalyst was pioneered by Deutsche Texaco [62–69]. The feed, consisting of propene gas and liquid water, is fed in a supercritical state to the top of a fixed-bed reactor and allowed to trickle downward. Feed containing about 92 % propene can be used, resulting in propene conversions of about 75 % per pass. About 5 % diisopropyl ether and about 1.5 % oxygenates (alcohols) of propene oligomers (hexenes) are formed as byproducts. This technique requires high-pressure equipment, and could suffer from problems of short catalyst life. This technology and plants based on it are licensed outside Europe [70].
2) High-temperature, high-pressure (270 – 300 °C, 20 MPa) vapor – liquid hydration of propene over a reduced tungsten oxide catalyst was developed by Tokuyama Soda [71–73]. This process utilizes a molar ratio of water to propene of about 2.5: 1. Water is present in both the gas and liquid phases, which increases conversion (because equilibrium is shifted farther to the right due to the solubility of 2-propanol in water). This technology requires high-pressure equipment, but features high propene conversions of 60 – 70 % per pass and 2-propanol selectivities of 98 – 99 % based on converted propene. The catalyst must be durable (stable in the presence of water). The process has low gas-recycle requirements, so propene of only 95 % purity can be used.
3) ICI developed technology for vapor-phase hydration of propene involving medium to high pressures. This process uses a WO_3 – SiO_2 catalyst at 250 °C and 25 MPa [19], [74], and it gives yields of about 95 %. In Germany, VEBA developed a similar hydra-

tion process based on a phosphoric acid catalyst supported on SiO_2 and operated at 180–260 °C and 2.5–6.5 MPa [75–77]. Typically, these processes require high propene recycle (less than 10 % conversion per pass) and utility costs are probably high. High-purity propene (~99 %) is required. The phosphoric acid process is commercial in Germany, the Netherlands,the United Kingdom,and Japan.

Hydrogenation of Acetone. Hydrogenation can be conducted in the liquid phase over a fixed catalyst bed of a Raney-nickel catalyst to give 99.9 % selectivity and 99.9 % conversion of the acetone [78], [79]. Hydrogenation over copper oxide–chromium oxide at 120 °C and 196 kPa gives lower selectivities and conversions (98 % and 94 %, respectively) [80]. It is not essential that the acetone be pure. This process is particularly advantageous where excess acetone is available as a byproduct from another process (e.g., the cumene–phenol process).

Other Processes. Patents have been issued describing the manufacture of 2-propanol from cellulosic materials (e.g., cotton, corn, and wood) [81], [82].

5. Uses

Uses of the two propanols are dictated by their solvent properties, their high water miscibility, and by their potential for introducing the propyl group into chemical intermediates.

1-Propanol. In 1988 over 75 % of the 1-propanol in the United States was employed in solvent applications, either directly or in the form of acetate ester or glycol ether derivatives [1]:

Solvent uses	24 900 t
1-Propyl acetate	19 500 t
n-Propylamines	11 300 t
Glycol ethers	5 400 t
Others	2 300 t
Total	63 400 t

As a solvent, 1-propanol is used principally in printing inks, paint, cosmetics, pesticides, and insecticides [83].

The EC used about 100 000 t of 1-propanol in 1988 [84]. In Germany, BASF converts most of its propanal into 1-propanol for printing inks, cosmetics, solvents, and intermediates for propylamines used in pharmaceuticals and pesticides [1]. In Japan, 1000–2000 t of 1-propanol was consumed in 1988 for printing inks and paints, all of which was imported [1].

1-Propanol is used commercially to produce glycol ethers. These are characterized by dual functionality, which imparts high solvency, chemical stability, and water compatibility. Glycol ethers such as ethylene glycol monopropyl ether (from 1-propanol and 1 mol of ethylene oxide), diethylene glycol monopropyl ether (from 1-propanol and 2 mol of ethylene oxide), propylene glycol monopropyl ether (from 1-propanol and 1 mol of propylene oxide) and dipropylene glycol monopropyl ether (from 1-propanol and 2 mol of propylene oxide) are marketed by Union Carbide Chemicals and Plastics [85] and Eastman Chemical Products [86].

n-Propyl propionate [*106-36-5*] is a new solvent available from Union Carbide Chemicals and Plastics and developed for use in automotive refinishing, appliance coatings, and high-solids coating systems [85]. *n*-Propyl propionate is touted as a replacement for the isomeric *n*-butyl acetate in coatings, where improved odor characteristics are desirable.

2-Propanol. 2-Propanol is used primarily as a solvent in inks and surfactants [70]. Other applications include its role as an antiseptic alcohol, as a reaction solvent for cellulose carboxymethyl ether [*9000-11-7*] (CMC), in the production of cosmetic base materials and pesticide carriers, as a source of material for organic synthesis, for washing of a flux used in soldering electrical circuits, and for removal of water from gasoline tanks in cars [70]. Table 3 shows approximate consumption data for 2-propanol in both 1986 and 1990 in the United States.

The coatings use of 2-propanol has reached a plateau due to a shift toward waterborne and high-solids coatings. Its use as an electronics solvent is growing, but the market is small [89]. 2-Propanol is a particularly versatile solvent for producing high-performance (e.g., gelation-resistant and polyamide-containing) printing inks [90]. In the pharmaceutical industry, 2-

propanol enjoys favor as a processing solvent during manufacture of drugs.

Use of 2-propanol for the production of acetone has decreased dramatically. There remains only one acetone plant in the United States (compared to four in 1985), and the market is still oversupplied [91]. However, since the cumene route to phenol provides equal amounts of phenol and acetone (→Acetone), a strong demand for acetone relative to phenol can increase the need for acetone production from 2-propanol [88].

2-Propanol is widely used as a chemical intermediate, for example in reductive amination to produce monoisopropyl amine (for herbicide and pesticide production), and as a source of isopropyl acetate. Some diisopropyl amine is utilized as an intermediate in the synthesis of diisopropylammonium nitrate, a corrosion inhibitor [91].

Nippon Petrochemical has begun marketing pure 2-propanol (99.99 %) for use in high-purity large-scale integration (LSI) and silicon wafers for the electronics industry [92]. Another attractive new application is as an octane enhancer, carburetor anti-icing additive, and methanol cosolvent in motor gasoline blends [93].

2-Propanol is useful in extraction processes; thus, aqueous solutions are utilized in liquid–liquid extractions of fatty acids from vegetable oil at temperatures as low as −2 °C [94].

Table 3. 2-Propanol consumption in the United States (10^3 t) [87], [88]

	Year	
	1986	1990
Coating solvents	8.2	9.1
Processing solvents	6.4	6.8
Household/personal	6.4	5.4
Pharmaceuticals	6.4	5.4
Acetone	4.5	3.6
Miscellaneous solvent applications and chemical intermediates	4.5	3.2
Exports	9.1	11.8
Total	45.5	45.0

6. Specifications

ASTM standard specifications for propanol are given in Table 4.

Product purity is generally determined by capillary gas chromatography. Typical purities specified by manufactures for 1- and 2-propanol are 99.7 – 99.9 wt % and 99.8 wt %, respectively. Because of its use in packaging, 1-propanol may sometimes be required to pass a specific odor test conducted for consumers by an odor panel. Occasionally, alkalinity and propanal content are included in the requirements for 1-propanol.

7. Economic Aspects

Manufacture of the propanols in major industrial countries is summarized in Table 5.

Since 2-propanol is no longer the primary feedstock for acetone it will see little, if any, growth in the United States unless new applications are found. In 1989 production units there were currently operating at about 60 – 80 % of capacity [96]. United States output of 2-propanol was 649×10^3 t in 1989 [97], and forecasted to grow to only 726×10^3 t by the year 2000 [98]. Demand fell by 4.4 %/a in 1977 – 1986 to 578×10^3 t in 1986 due to the decrease in use for acetone production [99]. The capacity for 2-propanol in Western Europe is roughly comparable to that in the United States. It was estimated based on U.S. import/export data that Western Europe in 1991 operated at or near full 2-propanol capacity.

2-Propanol production in Japan has been rising by about 5 %/a. Demand in Japan reached 130×10^3 t in 1989 [100], with steady growth (3 – 4 %) in paints and solvents and rapid growth (15 – 20 %) in detergent applications [101]. In 1990 there was a shortage in Japan [100].

Nippon Petrochemicals (50 %), Tokuyama Soda (25 %), and Mitsui Toatsu Chemicals (25 %) are the three major producers of 2-propanol in Japan [70]. Tokuyama Soda employs its own direct hydration technology based on a tungsten catalyst, while Nippon Petrochemicals uses technology developed by VEBA [70].

8. Storage and Transportation

The propanols are flammable liquids, and should be considered dangerous fire risks when exposed to heat or flame. Explosive limits for propanols in air are 2 – 12 %. Some important data for the safe handling of propanols are given in Table 6.

Table 4. ASTM Specifications for propanols

	1-Propanol	2-Propanol	Specification method
ASTM standard	D 3622-90	D 770-90	
d_{20}^{20}	0.804 – 0.807	0.785 – 0.7870	
d_{25}^{25}	0.801 – 0.804	0.782 – 0.784	D 268 or D 4052
Color, Pt-Co (ASTM)	10	10	D 1209
Distillation range, 101.3 kPa	distill within a 2 °C range that includes 97.2 °C	distill within a 1.5 °C range that includes 82.3 °C	D 1078
Nonvolatile matter, mg/100 mL	5	5	D 1353
Water, max, wt %	0.1	0.2	D 1364 or D 1476
Acidity, acetic acid, max, wt %	0.003	0.002	D 1613

Table 5. 1-propanol and 2-propanol manufacture in major producing countries (10^3 t) [95]

Country	Production of 1-propanol in 1988	Capacities of 2-propanol		Production of 2-propanol in 1994
		1991	1996	
United States	98	791	740	658
France		115		
Germany	5 [a]	215	220	175
Japan	0	209	160	130
Netherlands		250		
United Kingdom		190		
Western Europe			750	570

[a] Production for merchant market.

Table 6. Safety data and transport regulations for propanols [102]

	1-Propanol	2-Propanol
Flash point (open cup), °C	15.0	11.7
Flammability limits in air, vol %	2.0 – 12.0	2.0 – 12.0
Autoignition temperature, °C	371.11	398.89
Vapor pressure at 20 °C, kPa	1.9854	4.4053
Percent in saturated air, 25 °C	2.7	5.8
Density in saturated air, (air = 1), 25 °C	1.028	1.06
Heat of combustion, MJ/kg	−33.60	−33.35
U.N. No.	1274	1219
RID/ADR	Class 3, number 3 b	Class 3, number 3 b
IMDG Code	Class 3.2	Class 3.2
CFR 49	172.101 flammable liquid	172.101 flammable liquid

Care should be taken during heating (e.g., distillation) of old samples of 2-propanol. Several violent explosions have been reported during attempts to concentrate old samples to small volume, suggesting that this compound should be classed as peroxidizable (via autoxidation of the tertiary hydrogen) [103].

Baked phenolic-lined steel or stainless steel tanks are recommended for storage and shipping of the propanols in order to maintain high quality. Mild steel tanks are also adequate providing a filtering system is installed to remove rust [104]. Storage and transport under dry nitrogen is also recommended to protect against moisture pickup and to minimize flammability hazards. Aluminum is not recommended. The flash points of the propanols are low enough so that the compounds can be stored underground under a nitrogen atmosphere. Pipes and pumps can be constructed from the same metals as the storage facility. Centrifugal pumps with explosion-proof electric motors are suitable [104].

Table 7. Toxicity of propanols in animals [107]

	1-Propanol	2-Propanol
LD_{50} rat (oral), mg/kg	1 870 [a]	5 045 [b]
LD_{50} rabbit (intravenous), mg/kg	483 [a]	1 184 [a]
LCLo rat (inhalation), ppm/4 h	4 000 [a]	12 000 [a], ppm/8 h
Eye injury, rabbit	20 mg/24 h, moderate 4 mg, open severe	10 mg, moderate
LD_{50} rabbit (cutaneous), mg/kg	5 040 [a]	12 800 [a]

[a] No description of toxic effects.
[b] Behavioral (altered sleep time).

Table 8. Odor detection limits for propanols [106]

	Absolute odor threshold [a], ppm	50 % Odor recognition [b], ppm	100 % Odor recognition [b], ppm
1-Propanol	<0.03	0.08	0.13
2-Propanol (anhyd.)	3.20	7.50	28.2

[a] This is the concentration at which 50 % of an odor panel observed an odor.

[b] The concentrations at which 50 % and 100 % of the panel defined the odor as due to propanol, respectively.

9. Toxicology and Occupational Health

The propanols are considered to have generally low toxicity in humans [102]. Ingestion of excessive dosages of 2-propanol by alcoholics and suicidals has been reported to cause poisoning due to depression of the CNS and respiratory irritation. Repeated or prolonged contact with propanols may cause drying of the skin. The propanols display an exposure warning because they can cause mild to severe irritation to the eyes, nose, and throat. However, adverse effects are rare in light of widespread use by the general public and considerable industrial exposure [102], [105].

Animal studies reveal irritation to the eyes, low acute oral toxicity, and low irritation to the skin, as shown by the data in Table 7.

The mean exposure limits for 1- and 2-propanol in air at the workplace by ACGIH (TLV) and OSHA (PEL) are 200 ppm and 400 ppm, respectively. The MAK value for 2-propanol is 400 ppm. Odor threshold data are provided in Table 8.

The propanols are registered in the Toxic Substance Control Act (TSCA) and the European Inventory of Existing Commercial Substances (EINECS).

10. References

1. SRI International: *Chemical Economics Handbook,* "Oxo Chemicals Report," Jan. 1991.
2. Unpublished physical property data, Union Carbide Chemicals and Plastics Company Inc., Danbury, Conn.
3. Data from DIPPR (Design Institute for Physical Property Data) (ONLINE data base) Project 801 Source File Tape, revision dates Aug. 1988, 1989, 1990.
4. Thermodynamics Research Center: "TRC Thermodynamic Tables – Non-Hydrocarbons," Loose-leaf Data Sheets, The Texas A & M University System, College Station, Texas.
5. N. I. Sax, R. J. Lewis, Sr.: *Dangerous Properties of Industrial Materials,* 7th ed., Van Nostrand Reinhold, New York 1989.
6. G. H. Tryon (ed.): *Fire Protection Handbook,* 12th ed., National Fire Protection Association, Boston, Mass. 1962.
7. G. S. Ross et al.: "Polymorphism in Isopropyl Alcohol," *Science (Washington D.C.)* **141** (1963) 1043.
8. Thermodynamics Research Center: *TRC Table* **23-2-1-(1.1000)-k, p. 1,** Dec. 31, 1960; June 30, 1965; Dec. 31, 1976; p. k-5000.
9. Thermodynamics Research Center: *TRC Table* **23-2-1-(1.1001)-k, p. 1,** June 30, 1965, p. k-5010.
10. Thermodynamics Research Center, "Selected Values of Properties of Chemical Compounds," Data Project, Loose-leaf Data Sheets, Texas A & M University, College Station, Texas 1980.
11. R. C. Wilhoit, B. J. Zwolinski, "Physical and Thermodynamic Properties of Aliphatic Alcohols," *J. Phys. Chem. Ref. Data Suppl.* **2** (1973) no. 1.
12. J. J. Jasper, "The Surface Tension of Pure Liquid Compounds," *J. Phys. Chem. Ref. Data* **1** (1972) no. 4, 841.
13. J. S. Riddick, W. B. Bunger: *Organic Solvents: Physical Properties and Methods of Purification,* 3rd ed., Wiley-Interscience, New York 1970.
14. Thermodynamics Research Center, "TRC Thermodynamic Tables – Nonhydrocarbons," The Texas A & M University System, College Station, Tex 1987.
15. D. Ambrose, J. Walton: "Vapor Pressures Up to Their Critical Temperatures of Normal Alkanes and Alkanols," *Pure Appl. Chem.* **61** (1989) no. 8, 1395.
16. L. H. Horsley, *Azeotropic Data–III,* American Chemical Society, Washington, D.C., 1973.

17. J. Timmermans: *The Physico-chemical Constants of Binary Systems in Concentrated Solutions,* vol. 4, Interscience, New York 1960.
18. Shell Oil Co., US 4 472 593, 1984 (L. H. Slaugh, G. W. Schoenthal, J. D. Richardson).
19. K. Weissermel, H. J. Arpe: *Industrial Organic Chemistry,* Verlag Chemie, Weinheim 1978.
20. Z. Hejda, R. Zidek, J. Kozuch, CS 241 425, 1988.
21. J. Pasek, V. Pexidr, R. Zidek, J. Hajek, CS 234 604, 1987.
22. Polska Akademia Nauk, PL 137 060, 1987 (A. Kowai, J. Haber).
23. BP Chemicals Ltd., EP 158 499, 1985 (J. Russell, J. A. Stevenson).
24. BP Chemicals Ltd., EP 9 886, 1980 (D. C. Buttle).
25. Chemische Werke Huels AG, EP 66 059, 1982 (H. Alfs, W. Boexkes, E. Vangermain).
26. BP Chemicals Ltd., JP 55 038 399, 1978.
27. Kuraray Co., Ltd., JP 59 137 444, 1984.
28. CIL Inc., ZA 8 004 227, 1981 (D. J. Gannon, D. T. F. Fung).
29. Institutul de Cercetari, RO 89 624, 1986 (G. Duda et al.).
30. Borsodi Vegyi Kombinat, HU 40 450, 1986 (J. Kalacska et al.).
31. Huels AG, DE 3 739 577, 1988 (U. Hdo, H. G. Srebny, H. J. Vahlensieck).
32. Mitsubishi Mining and Cement Co., Ltd., JP 60 218 344, 1985 (Y. Fukuda, M. Shimura, K. Miyahara).
33. International Business Machines Corp., EP 248 195, 1987 (J. J. Cuomo, P. A. Leary, J. M. Woodall).
34. Peti Nitrogenmuvek, HU 35 241, 1985 (G. Kincses et al.).
35. Daicel Chemical Industries, Ltd., JP 58 174 348, 1983.
36. Mitsui Toatsu Chemicals, Inc., JP 03 041 042, 1991 (A. Hiai, J. Ono, H. Kato, N. Kitano).
37. Mitsui Toatsu Chemicals, Inc., JP 03 074 343, 1991 (A. Hiai, J. Ono, H. Kato, N. Kitano).
38. Showa Denko K.K., JP 60 215 642, 1985 (T. Kametaka, T. Nozawa).
39. Nippon Synthetic Chemical Industry Co., Ltd., JP 62 019 556, 1987 (T. Kakimoto).
40. Union Carbide Corp., US 4 247 486, 1981 (E. A. V. Brewester, R. L. Pruett).
41. Union Carbide Corp., US 4 277 627, 1981 (D. R. Bryant, E. Billig).
42. J. Falbe in E. G. Hancock (ed.): *Propylene and its Derivatives,* J. Wiley & Sons, New York 1973, p. 333.
43. Union Carbide Corp., EP 8 767, 1980 (C. C. Pai).
44. Union Carbide Corp., US 4 762 817, 1988 (J. E. Logsdon, R. A. Loke, J. S. Merriam, R. W. Voight).
45. Union Carbide Chemicals and Plastics Company Inc., US 4 876 402, 1989 (J. E. Logsdon, R. A. Loke, J. S. Merriam, R. W. Voight).
46. J. B. Cropley, L. M. Burgess, R. A. Loke, *CHEMTECH* **14** (1984) 374.
47. W. L. Faith, D. B. Keyes, R. L. Clark: *Industrial Chemicals,* 3rd ed., Wiley & Sons, New York 1965, p. 183.
48. Davy McKee Ltd., EP 74 193, 1983 (M. W. Bradley, N. Harris, K. Turner).
49. BASF, DE 3 228 881, 1984 (K. Baer et al.).
50. VEB Leuna-Werke, DD 240 742, 1986 (P. Franke et al.).
51. VEB Leuna-Werke, DD 273 163, 1989 (G. Kohl et al.).
52. N. L. Holy, T. F. Carey, Jr., *Appl. Catal.* **19** (1985) no. 2, 219.
53. Ruhrchemie AG, EP 84 833, 1983 (B. Cornils, C. D. Frohning, H. Bahrmann, W. Lipps).
54. W. Ueda, T. Kuwabara, T. Ohshida, Y. Morikawa, *J. Chem. Soc. Chem. Commun.* 1990, no. 22, 1558.
55. M. H. Carr, H. P. Brown, *J. Am. Chem. Soc.* **69** (1947) 1170.
56. Tokuyama Soda, JP 77 012-166, 1977 (T. Sato, R. Ohuji, H. Yamanovchi).
57. Exxon Research and Engineering Co., GB 2 004 538, 1979.
58. Exxon Research and Engineering Co., US 4 219 685, 1980 (C. Savini).
59. Takeda Chemical Ind. KK, JP 55 091 359, 1980.
60. *Chemical Economics Handbook,* "Isopropyl Alcohol", United States Data Summary, SRI International, July 1990.
61. V. R. Gurevick et al., *Azerb. Neft. Khoz.* 1990, no. 6, 54.
62. *Hydrocarbon Process.* **68** (1989) no. 11, 111.
63. *Hydrocarbon Process.* **70** (1991) no. 3, 185.
64. W. Neier, J. Woellner, *CHEMTECH* 1973, Feb., 95.
65. *Hydrocarbon Process.* **60** (1981) no. 11, 173.
66. W. Neier, J. Woellner, *Erdoel Kohle Erdgas Petrochem.* **28** (1975) no. 1, 19.
67. W. Neier, J. Woellner, *Hydrocarbon Process.* **5** (1972) no. 11, 113.
68. Deutsche Texaco AG, DE 86-3628007, 1987 (R. R. Carls, G. Osterburg, M. Prezlj, W. Webers).

69. Deutsche Texaco AG, DE 3 512 518, 1986 (F. Henn, W. Neier, G. Strehlke, W. Webers).
70. *Jpn. Chem. Week* (1987) Dec., 10, 6, 8.
71. Y. Onoue, Y. Mizutani, S. Akiyama, Y. Izumi, *CHEMTECH* 1978, July, 432.
72. Y. Onoue, Y. Izumi, *CEER Chem. Econ. Eng. Rev.* **6** (1974) no. 7, 48.
73. Tokuyama Soda, US 3 758 616, 1973 (Y. Izumu, Y. Kawasaki, M. Tani).
74. J. C. Fielding in E. G. Hancock (ed.): *Propylene and its Industrial Derivatives,* J. Wiley & Sons, New York 1973.
75. Hibernia-Chemie GmbH, BE 683 923, 1966.
76. VEBA-Chemie AG, US 3 955 939, 1976 (A. Sommer, M. Urban).
77. *Eur. Chem. News* (1970) July 24, 32.
78. Mitsui Petrochemical Ind., Ltd., JP 03 133 941, 1991 (H. Fukuhara, Y. Shibuta, T. Tate, T. Isaka).
79. Mitsubishi Petrochemical Co., Ltd., JP 03 141 235, 1991 (H. Fukuhara, K. Taniguchi).
80. Mitsui Toatsu Chemicals, Inc., JP 03 041 038, 1991 (A. Hiai, J. Ono, H. Kato, N. Kitano).
81. New Fuel Oil Development Technology Research Assoc., JP 61 067 493, 1986 (S. Onuma, M. Naganuma, H. Oiwa).
82. Services de Consultation D et B Plus Ltee., CA 1 162 867, 1984 (H. C. Rothlisberger).
83. *Manuf. Chem.* **60** (1989) July, 7.
84. *Financ. Times (North Am. Ed.)* (1989) June, 7, 8.
85. UCAR n-Propyl Propionate, Brochure F-60741 A, Union Carbide Chemicals and Plastics Corporation, Danbury, Conn., Oct. 1989.
86. *The World of Eastman Chemicals,* Brochure ECP 3131, Eastman Chemical Products, Inc., Kingsport, Tenn, Jan. 1989.
87. *Chem. Mark. Rep.* **232** (Aug. 31, 1987) 46.
88. *Chem. Mark. Rep.* **238** (Sept. 3, 1990) 50.
89. *Chem. Mark. Rep.* **232** (Aug. 31, 1987) 19.
90. *Polym. Paint Colour J.* **172** (1989) June, 462.
91. G. T. Austin, Isopropyl Acetate Production, *Chem. Eng.* **141** (1974) May 101.
92. *Jpn. Chem. Week* (1987) Aug. 13, 3, 5.
93. M. Prezelj, *Hydrocarbon Process.* **66** (1987) no. 9, 68.
94. K. J. Shah, T. K. Venkatesan, *J. Am. Oil Chemists Soc.* (1989) June, 783.
95. Directory of Chemical Producers – United States, Western Europe, and East Asia, SRI International, 1991.
96. CPI Purchasing (Apr. 1989) 38 – 43.
97. *Chem. & Eng. News* **68** (1990) April 9, 12.
98. *Chem. Week* **143** (1989) Nov. 1, 15.
99. *Chem. Mark. Rep.* **232** (1987) Aug. 31, 46.
100. *Jpn. Chem. Week* **144** (1990) Aug. 16, 3.
101. *Jpn. Chem. Week* (Apr. 18, 1991), 2.
102. *Patty,* 3rd ed., **2 C.**
103. L. Bretherick: *Bretherick's Handbook of Reactive Chemical Hazards,* Butterworth, London 1990, p. 391.
104. "UCAR Alcohols for Coatings Applications," Brochure No. F-48588 A, Union Carbide Chemicals and Plastics Corp., Danbury, Conn., July 1989.
105. N. I. Sax, R. J. Lewis, Sr.: *Dangerous Properties of Industrial Materials,* 7th ed., vol. 3, Van Nostrand Reinhold, New York 1988.
106. D. V. Sweet (ed.): *Registry of Toxic Effects of Chemical Substances,* Publication No. 87-114, vol. 4, National Institute for Occupational Safety and Health, DHHS (NIOSH), Washington, D.C., Apr. 1987.
107. T. M. Hellman, F. H. Small, *J. Air Pollut. Control Assoc.* **24** (1974) no. 10, 979.
108. *Fed. Regist.* **54** (1989) no. 203, 43252 – 43264.

Propargyl Alcohol → Alcohols, Aliphatic

Propellants

→ Explosives and → Pyrotechnics are separate keywords

HILTMAR SCHUBERT, Fraunhofer-Institute for Chemical Technology, Pfinztal, Federal Republic of Germany

1. Introduction 1
2. History 2
3. Classification and Formulation of Propellants 3
3.1. Solid Propellants 4
3.1.1. Homogeneous Propellants 4
3.1.2. Composite Propellants 5
3.2. Liquid Propellants 6
3.2.1. Monopropellants 6
3.2.2. Bipropellants 8
4. Reactions and Properties 8
4.1. Combustion Process 8
4.2. Propellant Performance 10
4.3. Mechanical Properties 11
4.4. Chemical Stability 12
4.5. Service Life 12
4.6. Sensitivity and Vulnerability 12
5. Uses 14
5.1. Gun Propellants 14
5.1.1. Solid Propellants 14
5.1.2. Liquid Propellants 15
5.2. Rocket Propellants 16
5.2.1. Solid Propellants 16
5.2.2. Liquid Propellants 18
5.3. Propellants for Gas Generators . . . 18
6. Production 18
6.1. Homogeneous Propellants 19
6.1.1. Single-Base Propellants 19
6.1.2. Double- and Multibase Propellants . . 19
6.2. Composite Propellants 21
6.3. Special Propellant Devices 22
7. Test Methods 24
8. Safety Precautions and Environmental Protection 24
9. References 25

1. Introduction

Propellants are explosive agents that generate large amounts of hot gases by an exothermic reaction without the need for atmospheric oxygen. This process is generally initiated by supplying thermal energy. Burning occurs vertically to the burning surface at a rate of 10^{-3} to 1 m/s. Rates of several 100 m/s can also be achieved under certain conditions (deflagration). Detonation is chemically the same as deflagration but the chemical reaction is initiated by a shock wave which is propagated through the charge.

The burning rate depends on the prevailing pressure and differs from the pressure-independent detonation rate ($10^3 - 10^4$ m/s) by several orders of magnitude. Depending on the external conditions (confinement), chemical composition, and physical structure of the charge, a deflagration can become a detonation and vice versa.

Propellants are used whereever reproducible, time- controlled amounts of working fluid are required to propel rockets and missiles or to launch projectiles from guns (gun powder). They are also used in gas generators that are designed to jettison loads, drive turbines and pumps, actuate mechanical devices, convey liquids, or inflate balloons. The working gases of gas generators should generally have as low a temperature as possible. Gas generators are becoming increasingly important for civil applications (e.g., airbags in automobiles).

Since propellants burn without atmospheric oxygen to produce the working gas, the propellant system must contain the oxidizer required for the gas-generating reaction. The oxidizer may be available in the form of molecularly bound functional groups or in the form of discrete salts containing bound oxygen.

Propellants may exist in solid or liquid form and generally consist of several components.

The linear burning rate and its pressure dependence can be modified only very slightly by varying the chemical composition of the propellant. In order to satisfy practical requirements, the burning surface and/or the geometry of the propellant charge is therefore varied widely: this

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_185

means for rocket propellant as well as for gun propellant charges.

Solid propellants and some liquid propellants for rockets and guns may contain the fuel and oxidizer in a single phase (monopropellants). Alternatively, the liquid propellant fuel may undergo a combustion reaction with a separately added oxidizer; the combination of fuel and oxidizer (bipropellant) constitutes the propellant. Such systems involve combustion techniques similar to those used in a conventional combustion engine. The throughput can be varied by modifying the mode of injection into the combustion chamber.

The solid propellant charge and individual propellant components are subjected to mechanical stress and must have appropriate mechanical properties. With the exception of compressed powder propellants for gas generators, solid propellant charges consist of a polymeric binder matrix that may contain crystalline additives.

Single- and double-base propellants comprise plasticized nitrocellulose which may be filled with a crystalline explosive agent. Composite propellants consist of a polymeric binder that contains energetic materials and oxygen-containing salts.

Propellants also contain far more additives than explosives: stabilizers, burning moderators, energy suppliers, processing auxiliaries, etc. Previously, an explosive could be distinguished from a propellant on the basis of its chemical composition; nowadays, however, this is not always possible. For example, large proportions of high-performance explosives are found in propellant compositions, and the oxygen- containing salt ammonium perchlorate (NH_4ClO_4), commonly used in composite rocket propellants, may be found in explosive mixtures. Compositions for rocket propellants can only be distinguished from those for guns by the specific additives. There are two reasons for this. The first is the development of formulations that are particularly insensitive to external influences such as heat, impact, shock, shaped charges, or fragments (low vulnerability ammunition, LOVA). The second is the desire to obtain maximum performance.

Progressive automation of production and weapon technology within the last few decades has provided further incentives for development. Examples include the continuous production of propellants and the development of caseless ammunition, combustible casings, unit charges, and liquid propellants for guns. The use of environmentally friendly products and the possibility of recycling will also become important in the propellant sector.

2. History

The first incendiary agents were the precursors of solid propellants and were used in ancient times to attack defense fortifications and ships, and also as pyrotechnics for religious purposes.

Black powder-like compositions were discovered by Chinese alchemists in the eighth century. In the middle of the eleventh century such formulations were packed in bamboo tubes and used as rocket-propelled flame throwers. Black powder (a mixture of potassium nitrate, sulfur, and charcoal) probably reached Europe through India and Arabia by the thirteenth century: it was mentioned by Roger Bacon in 1257. Black-powder mixtures were also described by Albertus Magnus before 1280. The first drawing of a cannon operated with black powder was made by Walter de Milimete in 1326. Black powder remained the only propellant for guns and rockets for many centuries until the introduction of "smokeless powder" (plasticized nitrocellulose).

Nitrocellulose was first synthesized by Schönbein and Böttcher in 1846. In 1883 Nobel used nitroglycerine to plasticize nitrocellulose. In 1885 Vielle used an ether – alcohol mixture to produce a plasticized nitrocellulose. In 1899 Abel and Dewar used acetone as a plasticizing auxiliary. The replacement of black powder by nitrocellulose containing propellants lasted up to World War I.

In order to avoid the high barrel erosion produced by the high-performance powders, "cold powder" was developed in Germany; here, nitroglycerin was replaced by di- or triglycerine dinitrate (1935), and nitroguanidine was also added (triple-base propellant, Gudol powder 1937). "Ball powder," which can easily be metered and is suitable for use in small- caliber weapons, was developed in the United States in 1935.

Prior to the twentieth century the use of rockets propelled by black-powder charges was restricted almost exclusively to fireworks and dis-

play pyrotechnics. Few reports exist concerning the military application of solid propellant rockets. The reason for this was the short range and low accuracy of the rockets.

The development of compressed propellant charges consisting of smokeless powders began after World War I. The development of composite propellants consisting of a finely ground perchlorate- or nitrate salt that was incorporated in a polymeric binder, cast, and cured, began in 1940.

Combustible cases for larger caliber weapons were developed in the 1960s in the United States, and caseless ammunition for small caliber weapons was developed in the 1970s in the United States and Europe. Since the 1980s internationally unit charges have been produced for howitzers to automate the charging process. Since the 1970s attempts have been made to produce charges for all types of explosives (high explosives, gun propellants, rocket propellants) that display maximum insensitivity to external influences (shock, impact, temperature, etc.) while retaining their high performance (low vulnerability ammunition, LOVA). These trends, combined with efforts to produce environmentally compatible, possibly recyclable compositions, will continue.

The use of liquid propellants for rockets and missiles achieved great importance after experiments were carried out at the beginning of the twentieth century (Goddard, Ziolkowski) and before and during World War II (Oberth, Dornberger, von Braun, Nebel). This work was continued on a very large scale after 1945 in the United States and former USSR. In contrast to solid propellants, the choice of liquid propellants is still restricted to relatively few types for military application. After World War II the use of liquid propellants for guns was investigated. It remains to be seen whether these systems have any future [23–28].

3. Classification and Formulation of Propellants

Propellants are defined as substances or mixtures of substances of an explosive nature in a solid to liquid state that burn exothermically without atmospheric oxygen to form a large volume of gas. The time course of gas formation can be reproducibly predicted from the ambient pressure and the chemical composition, shape, particle size, and density of the propellant.

In order to obtain a high gas yield, the propellants must produce combustion gases that have high gas mole number, a low mean molecular mass, a high flame temperature, and a low ratio of specific heats (c_p/c_v).

Propellants are used as charges for guns, rockets, and gas generators. Solid or liquid propellant charges can be used for all three purposes. Pasty or gel-like systems have not achieved any importance.

Propellant charges serve simply to produce working gas to launch a missile from a tube or barrel or to propel rockets. In gas generators, however, they are used for a wide variety of purposes. The generated gases are used to drive pumps, empty tanks, actuate valves, bolts and separating devices, to inflate air bags, or jettison sub-systems. In contrast to propellant charges for guns and rockets, working gases for gas generators should have a medium or low temperature ($\leq 1000-1500\,^{\circ}\mathrm{C}$).

3.1. Solid Propellants

Solid propellants are used in guns, rockets, and gas generators. An extremely large variety of types is available with different forms and chemical compositions that depend on the intended use. *Single-base propellants* are based on *plasticized nitrocellulose.*

The plasticized nitrocellulose may, however, also contain an explosive plasticizer such as nitroglycerin (*double-base propellants*). Nitroguanidine may also be incorporated in the double-base powder matrix (*triple-base propellants*) in certain artillery systems that are susceptible to high barrel erosion.

In addition to nitrocellulose- containing propellant systems, a large range of *heterogeneous or composite propellants* is available. These propellants are formulated from a plastic fuel that is blended with oxygen- containing salts and/or energetic substances; they may also contain energy-rich plasticizers. The polymers themselves may have functional groups that increase the energy content of the propellant composition.

Finally, the principal constituents of double-base propellants can be combined with those of composite propellants. These combinations are termed *composite-modified double-base propellants (CMDB)* or *composite double-base propellants (CDB)*.

The most important constituents of solid propellants and their functions are listed below:

Binders (cross-linked)	polyurethane (PUR) hydroxy-terminated polybutadiene (HTPB) carboxy-terminated polybutadiene (DTPB)
Thermoplastic binder	cellulose acetate butyrate (CAB)
Energetic binders, thermoplastic	nitrocellulose (NC) poly(vinyl nitrate) (PVN)
Energetic binder, cross-linked	polyglycidylacid (GAP)
Plasticizers	triacetin dibutyl phthalate diethyl phthalate
Energetic plasticizers	nitroglycerine (NG) trimethylolethane trinitrate (TMETN) butanetriol trinitrate (BTTN) bis(2,2-dinitropropyl)acetal/formal (BTNA/F) dinitrotoluene (DNT)
Oxidizers	ammonium perchlorate (AP) ammonium nitrate (AN)
Energetic fuels	cyclotrimethylenetrinitramine (hexogen, cyclonite, RDX) cyclotetramethylenetetranitramine (octogen, homocyclonite, HMX) nitroguanidine (NQ) triaminoguanidine nitrate (TAGN)
Metallic fuel	aluminum
Ballistic modifiers for NC propellants	organic lead and/or copper salts
Ballistic modifiers for composite propellants	ferrocene derivatives, copper chromite
Stabilizers	diphenylamine nitrodiphenylamine diethyldiphenylurea (Centralite I) methyldiphenylurea (Akardite II)
Wetting agent	lecithin, polyoxyethylenesorbitantrioleate
Bonding agents	aziridine, triethanolamine
Flash reducers	potassium sulfate, potassium nitrate
Opacifier	carbon black

Details of the properties and production of the explosive components are given in [3–21], see also → Explosives.

3.1.1. Homogeneous Propellants

Single-base propellants consist mainly of plasticized nitrocellulose (nitrogen content e.g., 13.15 %). They contain plasticizers such as dibutyl phthalate and possibly additives to improve processing, suppress muzzle flash, etc. Older types of powders are processed also with dinitrotoluene. Stabilizers (e.g., diphenylamine) are added to the propellant for good storage stability. Single-base propellants are produced with maximum web thicknesses of 12 mm to facilitate solvent removal during processing with the result that they can be used only as gun propellants or in bulk powder fillings for gas generators.

Double-base propellants consist of nitrocellulose (nitrogen content 13.25 % to ca. 11.5 %) and an explosive plasticizer such as nitroglycerine. The remaining additives are similar to those used in single-base propellants. Both propellant types contain 0.5 ± 0.3 % water.

Types in which crystalline explosives such as HMX and RDX are incorporated in the double-base matrix represent a transition between double-base and composite propellants.

Triple-Base Propellants. Nitroguanidine has a relatively low flame temperature but produces a large volume of gas, and is used in "cold powders" (triple-base propellants) to reduce the flame temperature and thus protect the barrel. In these propellants nitroglycerine is replaced by explosive plasticizers with a low heat of explosion (e.g., diethyleneglycol dinitrate or butanetriol trinitrate). Examples of gun propellants are given in Table 1 and rocket propellants in Table 2.

Compositions for rockets also contain ballistic modifiers (e.g., lead salts and/or copper salts) to reduce the pressure exponent.

3.1.2. Composite Propellants

The heterogeneously formulated composite propellants consist of a polymeric binder in which crystalline oxidizers (e.g., ammonium perchlorate) and/or energy suppliers (e.g., RDX or HMX) are incorporated in amounts of up to 90 wt %. The binders are generally viscoelastic cross-linked elastomers. Compositions of composite propellants for rockets and gas generators are given in Table 3, and for LOVA guns in Table 4.

Many different types of plastics have been used but interest is now concentrated on castable

Table 1. Composition (wt %) and performance of gun propellant

Component or property	M 1	M 8	M 30	A 502	JA 2
Nitrocellulose	84.2	52.15	28.0	93.5	63.5
(nitrogen, %)	(13.15)	(13.25)	(12.6)	(13.1)	(13.0)
Nitroglycerine		43.0	22.5		14.0
Diglycol dinitrate					21.7
Nitroguanidine			47.7		
Ethylcentralite		0.6	1.5	1.4	
Diphenylamine	1.0			0.7	
Methyldiphenylurea					0.7
Dinitrotoluene	9.9				
Dibutyl phthalate	4.9			3.5	
Diethyl phthalate		3.0			
Additives		1.25	0.3	0.9	0.1
Flame temperature, K	2585	3757	3093	2807	3448
Force, J/g	962	1172	1102	998	1155
Mean molecular mass of products	22.3	26.7	23.3	23.4	24.8

M 1, A 502 = Single-base propellant, A 502 is a German 20 mm caliber propellant; M 8, JA 2 = Double-base propellants; M 30 = Triple-base propellant

polybutadienes. Their prepolymers are terminated with carboxyl or hydroxyl groups and cross-linked with difunctional or trifunctional isocyanates in the presence of catalysts. Bonding and wetting agents and plasticizers are added to improve processing and mechanical properties. Thermoplastics (e.g., cellulose acetate butyrate) are also suitable for certain applications, in particular for gun propellants, and are compressed after incorporation of a filler [11].

The mean particle size and particle distribution of the fillers influence processing, burning behavior, mechanical properties, and shock sensitivity. Rocket propellants contain aluminum powder to increase the flame temperature, although this has the disadvantage that a large amount of smoke is produced due to formation of aluminum oxide (primary smoke). Ballistic modifiers and additives for reducing pressure oscillation are added to improve burning behavior. Composite propellants based on ammonium perchlorate release hydrogen chloride on burning, which produces intense smoke in the presence of atmospheric moisture (secondary smoke). Ammonium perchlorate is replaced by chlorine-free oxidizers (e.g., ammonium nitrate) to reduce or avoid the formation of smoke and corrosive combustion gases. Energetic substances (e.g., RDX or HMX), explosive plasticizers (e.g., nitroglycerine, BTNA/F), and energy-rich binders (e.g., GAP) are used to improve performance and/or avoid or reduce the ammonium perchlorate content. If nitrocellulose or poly(vinyl nitrate) is also added to the nonexplosive binder, composite-modified double-base propellants are obtained (CMDB, C in Table 3).

Table 2. Compositions (wt %) and performance of double-base rocket propellants

Component or property	Extruded	Cast
Nitrocellulose	48.17	49.8
(nitrogen, %)	(12.75)	(12.6)
Nitrogylcerine	41.06	33.9
Dibutyl phthalate	3.91	
Dimethylformamide		1.2
Triacetin		9.1
Methyldiphenylurea	1.50	
Diethylphenylurea	1.0	3.6
Lead salts	2.85	2.2
Additives	1.51	0.2
Flame temperature, K	2549	2193
Specific impulse (70: 1) equilibrium, Ns/kg	2252	2119
Mean molecular mass of products	24.10	22.97
c_p/c_v	1.241	1.252

Composite propellants are particularly suitable for rockets up to very large diameters, but have also been used as insensitive LOVA gun propellants (Table 4). Corresponding developments can also be observed in the field of rocket propellants (insensitive rocket propellants, IRP). In such less- or low-sensitive propellants, nitrate esters are avoided wherever possible; instead nitramines (e.g., RDX or HMX of small grain size, $d < 10\,\mu m$) and/or nitroguanidine (spherical or high bulk density) or triaminoguanidine nitrate, are used as energy compounds [29–32].

Nitro compounds (e.g., BTNA/F) are used as explosive plasticizers. Thermoplastics (e.g., CAB) are used as binders for gun propellants. Filled viscoelastic cross-linked polymers that have been extruded and then cured may also be used as an alternative.

Performance may be improved by employing energetic binders such as GAP. Insensitive rocket propellants have similar compositions, both ammonium nitrate and ammonium perchlorate are used. The disadvantages of ammonium nitrate are its high hygroscopicity and the transformation of its crystal structure at +32.3 °C, which is accompanied by a change in density. This transformation can be displaced to higher temperatures by additives like diammine complexes of copper zinc, or nickel [33].

Propellants for larger gas generators are generally of the composite type. For small charges also single- or double-base propellants are used. Composite types have a low binder content and contain crystalline fillers (e.g., nitroguanidine, triaminoguanidine nitrate) that, on combustion, produce a large amount of gas with a relatively low flame temperature. Endothermically decomposing compounds (e.g., oxalates) are added to reduce the flame temperature. Passenger restraint systems for vehicles (e.g., air bags) generally contain sodium azide as a gas-generating charge which minimizes the concentration of carbon monoxide and carbon dioxide formed after the reaction.

3.2. Liquid Propellants [3], [15]

Liquid propellants contain the fuel and oxidizer as a single substance or mixture of substances (monopropellants), or in two separate phases (bipropellants). In contrast to solid propellants, where the solid for combustion is located in the combustion chamber, the liquid propellant is injected into the combustion chamber from a storage tank. The type of combustion and the propellant throughput is principally determined by metering. The first rockets and guns based on liquid propellants employed single substances or simple mixtures of substances. Only later did the compositions become more complex through the use of additives, which are used to modify the combustion process [34].

3.2.1. Monopropellants

Monopropellants consist of substances that already contain the oxidizer necessary for combustion (nitro compounds, nitrate esters), or of a mixture of fuel and oxidizer. These monopropellants must not react in the applied temperature range (−40 °C up to +50 °C) and must be stable on storage. Monopropellants have litte importance for rocket propulsion systems, apart from auxiliary systems (e.g., correction of satellite position).

Examples of monopropellants are 80 – 99 % *hydrogen peroxide* (which eventually decomposes, however, despite the addition of stabilizers) and *hydrazine* and its derivatives. Both compounds are injected into the combustion chamber over decomposition catalysts.

Monopropellants consisting of a mixture of components were initially used in addition to hydrogen peroxide to propel torpedoes. Examples of such mixtures are hydrazine, hydrazine nitrate, and water and compositions containing hydroxylammonium nitrate (HAN). These mixtures are known by the names Otto Fuel, Duplex, Solex, and NOS. Monopropellants were also used as propellants for guns; mixtures of nitromethane and isopropyl nitrate or methanol, as well as mixtures of hydroxylammonium nitrate and alkyl-substituted ammonium nitrate and water are of particular interest [35]. The following composition is typical:

LP 1845	63.2 % hydroxylammonium nitrate (HAN) 20.0 % triethanolammonium nitrate (TEAN) 16.8 % water

3.2.2. Bipropellants

Bipropellants consist of an oxidizer and a fuel that are injected into the combustion chamber in two separate phases. Both the oxidizer and fuel may consist of more than one component.

Conventional systems for rocket propulsion consist of hydrogen peroxide, concentrated nitric acid, nitrogen dioxide, or liquid oxygen as oxidizer, and hydrocarbons (e.g., rocket propellant, RP 1; jet propellant, JP-5), alcohols, amines, hydrazines, and alkyl derivatives as fuel. Liquid fluorine or a mixture of fluorine and oxygen (FLOX) and liquid hydrogen are also used in

Table 3. Composition (wt %) and performance of composite rocket propellants and gas generators

Component ∗ or property	A	B	C	D	E
AP	85	70	10	10	
RDX			52		
NQ				50	
TAGN					75
Aluminum		18		20	
PUR and additives			13.5		10
HTPB and additives	15	12			
GAP and additives				8	
NG and additives			24.5		
Triacetin					15
BTNA/F				12	
Flame temperature, K	2813	3253	2995	2791	1269
Specific impulse (70: 1), Ns/kg	2390	2681	2453	2486	1759
Mean molecular mass of products	24.5	27.2	24.7	24.2	18.3
c_p/c_v	1.22	1.20	1.23	1.21	1.29
Density, g/cm^3	1.66	1.79	1.70	1.76	1.40

∗ For explanation of abbreviations, see 3.1
A) Composite propellant with secondary smoke; B) Composite propellant with primary and secondary smoke; C) Composite modified double-base propellant with reduced smoke; D) Insensitive composite propellant; E) Gas-generator composition

Table 4. Composition (wt %) and performance of LOVA gun propellants

Component ∗ or property	A	B	C
RDX	76.0	65	61
NC	4.0		
NQ			26
TAN		20	
HTPB and additives		15	
GAP and additives			13
CAB and additives	12.0		
Acetyltriethyl citrate	7.6		
Diethyldiphenylurea	0.4		
Flame temperature, K	2814.0	2257	3202
Force, J/g	1120.5	1046	1243
Mean molecular mass of products	20.9	17.94	21.4

∗ For explanation of abbreviations, see 3.1
A) Propellant with thermoplastic binder; B) Propellant with cross-linked binder; C) Propellant with cross-linked energetic binder.

aerospace applications to improve performance. Some propellant combinations are hypergolic (i.e., the fuel and oxidizer ignite spontaneously on contact). Examples are concentrated nitric acid and hydrazine or alkyl derivatives. This hypergolic behavior is exploited in guns, where the propellant combination ignites immediately when injected into the combustion chamber; an example of such a propellant is concentrated nitric acid and monomethyl hydrazine (MMH) or triethanolamine.

4. Reactions and Properties

Propellants have to satisfy various criteria and must also have specific properties, depend-ing on their use. When choosing and developing a propellant its properties must therefore be matched against one another to achieve the best solution. This requires interdisciplinary cooperation between the system developer and propellant manufacturer. Factors that have to be observed when choosing a propellant are listed below:

Ignition conditions
Burning rate (depending on pressure and temperature)
Performance
Internal ballistic behavior (combustion instability and pressure irregularities, pressure – time curve)
Density and bulk density
Chemical stability and compatibility
Long storage behavior – service life
Mechanical properties (depending on temperature and stress rate)
Deflagration to detonation behavior
Sensitivity and vulnerability
Manufacturing characteristics (availability, cost, hazards etc.)
Rocket exhaust smoke
Gun smoke and flash

4.1. Combustion Process [12], [36–38]

The burning behavior of a solid propellant depends on its composition, pressure, propellant temperature, and on the characteristics of the flow field. The propellant burns in parallel layers vertically to the burning surface, provided that the combustion gases can escape unimpeded. The mass flow rate depends on the amount of heat produced during combustion and its feedback to the burning surface. Ignition is effected by heating the propellant surface above the decomposition temperature with hot gases, hot particles, or lasers.

Many combustion models exist for different compositions and pressure ranges. In practice the following equations are used:
For propellants in guns ($p > 200$ kPa)

$$r = a + bp$$

and for rockets ($p < 30$ kPa)

$$r = cp^n$$

where

r = linear burning rate
p = pressure
a, b, c = temperature-dependent constants
n = pressure exponent

The gaseous combustion zone above the burning surface constitutes the heat source for propellant decomposition. A rise in pressure compresses the combustion zone, thus increasing the heat feedback to the surface and producing a higher burning rate. A higher propellant temperature also increases the burning rate because less heat is then required to heat up the surface. At higher temperatures, higher pressures are therefore obtained in the gun or rocket combustion chamber.

Burning mechanisms differ according to the composition of the propellant. With *homogeneous* (single- and double-base) propellants the nitrate ester decomposes near the propellant surface (fizz zone) at ca. 600 K (Fig. 1). The primary products react in the nonluminous dark zone at ca. 1050 – 1750 K, which disappears at higher pressures. Further reaction up to thermodynamic equilibrium then occurs in the overlying flame zone (1650 – 2200 K).

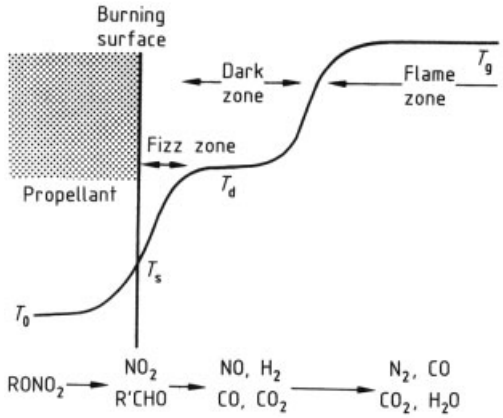

Figure 1. Diagram of the flame structure of a double-base propellant [36]

In the pressure range ca. 7 – 20 kPa, the use of burning modifiers consisting of lead and/or copper salts of aliphatic and aromatic acids in NC- containing propellants may result in a pressure-independent burning rate (plateau $n = 0$) or a reduction in the rate due to a pressure rise (mesa effect, n becomes negative) (Fig. 2). In general the pressure exponent in rockets should not exceed $n = 0.6$.

Nitramines such as RDX or HMX melt before or during decomposition, depending on the heating rate. The decomposition products then

react like the other combustion products in the zones described above [39].

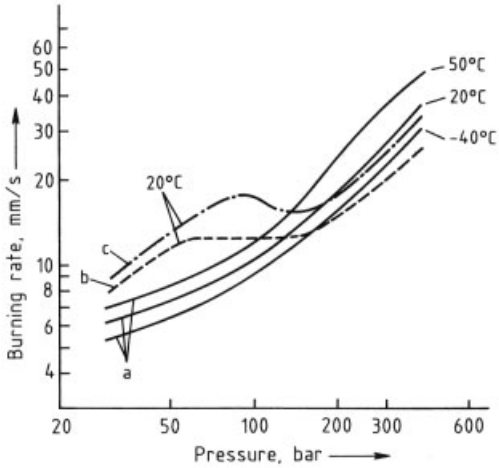

Figure 2. Burning behavior in solid, double-base propellants
a) Propellant without catalyst; b) Plateau propellant; c) Mesa propellantPropellant temperatures are indicated

Oxygen- containing salts, polymeric binders, and metal powders (aluminum) exhibit different combustion mechanisms, which can also change with pressure and temperature [40], [41]. The oxygen-rich gases of the decomposing oxygen-containing salt and the pyrolysis products of the organic binder produce a diffusion flame. The grain size of the ammonium perchlorate greatly influences the burning rate: the smaller the grain size the higher the burning rate. Burning modifiers (e.g., copper chromite) also increase the burning rate (Fig. 3). The grain size of ammonium nitrate has far less influence; these propellants have a relatively low burning rate, which can be increased by adding similar burning modifiers like ammonium dichromate. Oxidation of the metal powder occurs a long way from the burning surface and has only a slight effect on the burning rate. The small metal particles often agglomerate to form larger aggregates.

In *liquid monopropellants* vaporization and decomposition of the liquid occur on the droplet surface and depend on the composition and heating rate. Combustion follows in the gaseous phase. With *bipropellants* a diffusion flame is formed between the oxidizer and fuel.

4.2. Propellant Performance

The performance of a propellant is determined by its composition, its reaction products, and the difference between the heat of formation of the reactants before and after the reaction. The most important thermodynamic features of the combustion products are the flame temperature, gas volume, heat capacity, ratio of specific heats (c_p/c_v), and, where appropriate, the covolume of the gases at high pressures.

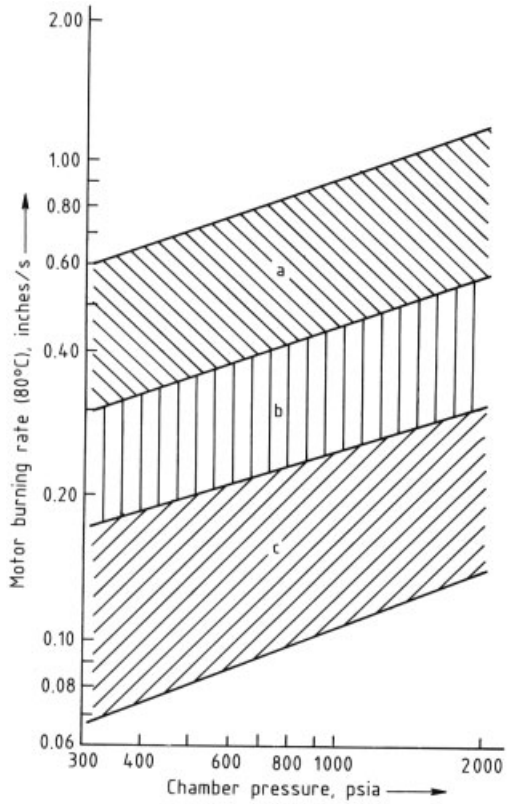

Figure 3. Burning rate range for polybutadiene propellants [16]
a) AP propellants with burning rate modifiers like metal oxides; b) AP propellant with different particle size; c) Propellant with mixed AP and AN oxidizers

Since the early 1970s computer programs have been developed to calculate the performance of gun and rocket propellants, they generally have an accuracy of 1 – 2 % [42], [43–45]. Differences between the results of the programs are generally attributed to different thermodynamic input data.

The specific energy or force f (in $\mathrm{m \cdot t\,kg^{-1}}$ or J/g) is derived from the equation of state for gases ($f = pV = nRT$) and is used as the performance parameter for gun propellants, where

p = pressure
V = volume
n = number of moles of gas
R = universal gas constant
T = explosion temperature

Since propellant powders in a gun at virtually constant volume and varying pressure (≤ 600 kPa) burn at temperatures up to 4000 K, the calculation requires an equation of state that takes into account the behavior of the gases at high densities (virial equation) [43–45].

The computer programs also provide information on the heat of explosion, explosion temperature, average molecular mass of the explosion products, total mole number, composition of the explosion gases, specific heat ratio (c_p/c_v), and covolumes. Examples of performance data of gun propellants are given in Tables 1 and 4.

The most important performance parameter of rocket propellants is the specific impulse I_s:

$$I_s = \frac{\bar{F}t}{w} \text{ in Ns/kg}$$

where

$\bar{F}$ = average thrust
t = burning time
w = weight of propellant

As the following equation shows, the specific impulse depends on the flame temperature in the chamber T_c, the average molecular mass of the flame gases $\bar{m}$, the specific heat ratio $k = c_p/c_v$, and the pressure expansion ratio p_c/p_e

$$I_s = \sqrt{\frac{2}{g} \cdot \frac{RT_c}{\bar{m}} \cdot \frac{k}{k-1} \left[1 - \frac{p_e^{(k-1)/k}}{p_c}\right]} \text{ Ns/kg}$$

where g is the acceleration due to gravity, p_c is the chamber pressure and p_e is the pressure at the nozzle end.

In contrast to gun propellants, rocket propellants generally burn at a uniform pressure. The specific impulse is often given at an expansion ratio of 70: 1, a distinction is made as to whether the combustion gases at the nozzle end are in equilibrium $I_{s,\,equil}$ or in a frozen state $I_{s,\,froz}$.

As well as the specific impulse the following information can be obtained from the computer programs:

Chamber temperature (adiabatic flame temperature)
Temperature of exit gas (equilibrium or frozen)
Composition of the burnt gases
Average molecular mass of the burnt gases
Total mole number of the burnt gases
Ratio of specific heats c_p/c_v

Since, in practice, the gas jet does not expand fully due to the restricted nozzle dimensions, a difference of up to several percent is found between the theoretical and experimentally determined specific impulse. Aluminum oxide occurs as condensed particles in the combustion jet and these particles often do not reach the gas flow rate. The difference between theory and practice is thus even greater in aluminum- containing propellants.

The combustion gases only reach equilibrium at the nozzle end in rockets with very large diameters. Examples of rocket-propellant performance data are shown in Tables 2 and 3.

Experiments in standard combustion chambers are carried out on a kilogram scale to determine the performance, burning behavior, and chamber pressure at different "Klemmung" (ratio of the burning surface area of the propellant charge to the nozzle throat area).

Since the volume of the loading chamber of a gun and the volume of a rocket combustion chamber are restricted, a charge with a propellant of a high density has advantages as regards performance. The velocity of a rocket V_b on completion of burning is also influenced by the ratio of the rocket mass at the start and completion of burning. For vertical liftoff the following equation applies:

$$V_b = \bar{V}_g \ln \frac{m_a}{m_e} - gt$$

where

V_g = velocity of the combustion gases
m_a = rocket mass at start
m_e = rocket mass at completion of burning
g = acceleration due to gravity
t = burning time

The densities of single- and double-base propellants are 1.54 – 1.64 g/cm^3. Composite propellants may have densities as high as 1.76 g/cm^3, particularly if they contain aluminum powder.

4.3. Mechanical Properties

Propellant charges are subjected to mechanical stresses that depend on their shape, size and the form in which they are used. Since propellants mainly consist of polymers, their mechanical properties also depend on the temperature and the stress rate. According to the superposition principle, the low-temperature behavior of the embrittlement, tensile strength, and elongation is nearly the same as that of propellants at room temperature but at high stress rates [46–48].

Gun Propellants. In guns the propellant powders are subjected to the force of the pressure wave in the powder bed. In extreme cases they are thrown against the base of the projectile during ignition. This leads to the danger of embrittlement, especially at low temperature with double- and triple-base propellants. Embrittlement results in an uncontrolled rise in pressure in the lower temperature range.

Rocket Propellants. The mechanical properties of rocket propellants are of greater importance. The thermoplastic double-base propellants generally have a high tensile strength and a relatively low elongation at break. They are therefore used as cartridge case propellant charges. The temperature range is limited due to embrittlement in the low temperature range (ca. −30 °C) and softening above +70 °C.

Viscoelastic composite propellants generally have a high elongation at break and a relatively low tensile strength. Composite propellants are therefore almost always used as case-bonded propellants, the large expansion capacity of the propellant absorbs the chamber movement as a result of temperature cycling. The tensile strength needs only be sufficiently large to ensure stability of the propellant under load. The larger the diameter and length of the propellant charge, the greater must be its elongation at break.

During the service life of a solid propellant the mechanical properties change in varying degrees. With composite propellants the elongation at break can deteriorate resulting in embrittlement caused by postcuring effects. Softening can also occur as a result of depolymerization, and leads to deterioration of the tensile strength [48].

The mechanical properties are also greatly affected by adhesion between the filler and binder. Moisture, frequent temperature variations, and vibration can cause dewetting of the filler from the binder, resulting in a decrease of the mechanical properties. In nitrocellulose the mean molecular weight decreases during aging, which likewise negatively affects the mechanical properties [49].

Surfaces of rocket propellants that are not meant to burn for internal ballistic reasons are provided with firmly adhering insulation layers made of polymers that are compatible with the propellant (e.g., ethylcellulose for double-base propellants and PVC or HTPB for composite propellants). The insulation layers may also contain inert fillers and plasticizers. The mechanical properties of the propellant and/or insulation can also change due to mutual migration of their constituents, and may result in malfunctions.

4.4. Chemical Stability

Nitrate esters decompose to form nitrogen oxides which provide the decomposition of ester groups (autocatalysis). This process is exothermic and can lead to self-ignition of the propellant. Stabilizers (e.g., substituted diphenylamines, ureas, or urethanes) are added to inhibit this reaction. Examples include diphenylamine, nitrodiphenylamine, and diphenylureas (Centralite, Akardite) [50–52].

The stabilizers trap the released nitrogen oxides to form polynitro compounds. The final stability of such a propellant is measured by the residual capacities of the stabilizers to trap nitrogen oxides. The process is temperature dependent and obeys the Arrhenius equation. Information of the service life of a propellant at normal temperature can thus be obtained from stability testing at elevated temperature.

Besides NO_x also N_2O, CO, CO_2 and H_2O may be formed during the aging process by re-

action of NO_x with the remaining reactants. In large propellant charges these gases lead to crack formation. Especially CO_2 may also accumulate underneath the insulation and detach it from the propellant surface.

The conventional composite propellants containing ammonium perchlorate are extremely stable. Chemical reactions involving the binder affect the mechanical properties. Nitramines and plasticizers containing nitro groups are also more stable than nitrate esters.

Hot-storage tests are used to check the chemical stability: samples of propellants are stored at 65.5 – 132 °C and the weight loss or release of gasis measured. Propellants that come into contact with other materials (insulation materials, liners, metal parts, etc.) are subjected to compatibility testing, i.e. the propellant is investigated to determine whether it reacts with the other material at elevated temperature (release of gas).

4.5. Service Life

The service life of a propellant ends when the reproducibility of performance and function can no longer be guaranteed within the prescribed limits.

The duration of the service life depends on the chemical composition, size of the propellant charge, type of use, conditions of storage and use (storage temperature, temperature fluctuations, moisture, mechanical stress, etc.), and the prescribed operational conditions.

The service life of gun propellants may be assumed to be about 30 years or even higher under good storage conditions and with very stable powders. Rocket propellants generally have a shorter service life, in some cases only 12 years [53].

4.6. Sensitivity and Vulnerability

Sensitivity. As in the case of high explosives, propellants are also screened to investigate their handling safety. Various methods and equipment are used to test the sensitivity with respect to impact, friction, electrostatic discharge, heat, and flame. Standard national test regulations exist for the certification of propellants and obtaining permission for transportation [54–57].

In 1987 the United Nations Committee of Experts on the Transport of Dangerous Goods issued recommendations for tests and criteria and has classified dangerous goods in nine classes [58]. Class 1 includes five explosive divisions

1.1) Substances and articles which have a mass explosion hazard
1.2) Substances and articles which have a projection hazard but not a mass explosion hazard
1.3) Substances and articles which have a minor blast hazard or a minor projection hazard or both, but not a mass explosion hazard
1.4) Substances and articles which present no significant hazard
1.5) Very insensitive substances which have a mass explosion hazard but are so insensitive that there is very little probability of initiation or of transition from burning to detonation under normal conditions of transport. As a minimum requirement they must not explode in the external fire test.

Handling and transport regulations are governed by the classification according to these five subdivisions. In some countries these subdivisions also have consequences for munition storage regulations.

The United Nations scheme provides for six test series which can also be applied to liquid propellants (monopropellants). The following parameters are tested:

1) Detonative shock under confinement
2) Intensive heat (fast cook-off)
3) Impact and friction
4) Thermal stability
5) Ignition
6) Free-fall impact of a test unit
7) Shock from a detonator cap
8) Transition from deflagration to detonation
9) Exposure of substance in packing to an external fire

With an additional Test Series 7, articles of Division 1.5 and extremely insensitive detonating substances (EIDS) are identified and specified precisely. For example, for EIDS the following are additionally specified:

1) Threshold for propagation of steady-state detonation conditions = Critical Diameter (confined and unconfined)

2) Detonative shock under confinement = EIDS GAP Test
3) Relative behavior of high explosive loads in projectiles and fired against a target = Susan Test
4) Mechanical behavior of a compact high explosive under a large impact (velocity of pieces = 150 m/s) = Friability Test
5) Response of an EIDS to kinetic energy associated with impact and penetration = Bullet Impact Test
6) Reaction to a gradually increasing thermal environment (3 K/h) = Slow Cook-Off Test

Vulnerability. The vulnerability of munition depends on the behavior of its explosive components with respect to external influences. These may vary in nature and intensity according to the use of the munition and the expected military threat. The inherent vulnerability is a combination of

1) The susceptibility of munition to initiate an unintentional reaction (i.e., its sensitivity)
2) The violence of such a reaction (i.e., its explosiveness) caused by accident (e.g., shock, impact), environmental stress factors (e.g., aerodynamic heating, electrostatic charging), and enemy action (e.g.,fragmentation shock).

The vulnerability of munition can be improved by modifying munition construction, implementing appropriate mechanical measures, and using insensitive energetic materials such as insensitive high explosive (IHE), insensitive gun propellants (LOVA), and insensitive rocket propellants (IRP). These insensitive energetic materials should react reliably on demand and fulfill performance requirements. Regardless of whether such materials are confined in munition or not, undesirable external influences caused by heat, shock, fragmentation projectiles, impact, friction, and electrostatic discharge should not lead to violent reactions that produce hazardous fragments or blast waves.

In order to check the vulnerability of gun and rocket propellant charges, the tests of the abovementioned Test Series 7 (Division 1.5) are used to the specific threat.

5. Uses

5.1. Gun Propellants

5.1.1. Solid Propellants

The Firing Process. In the firing process the propellant charge is ignited by a pyrotechnic device. Solid propellants for guns are also known as gun powder. Ignition is effected by hot gases, and is often assisted by the emission of hot particles. Lasers may be also used in future. The objective is to initiate ignition simultaneously over the whole surface of the propellant charge. Ignition lasts between < 1 ms and several 10 ms depending on the caliber size. The propellant charge then reacts to form hot gases, and the generated gas pressure ejects the projectile with increasing velocity from the barrel. Only 30 –40 % of the chemical energy is converted into kinetic energy to propel the projectile; 40 – 50 % is converted into kinetic energy of the gases or remains as heat in the gases; a further 20 – 25 % heats the barrel, case, and projectile. The pressure falls due to the energy losses of the gases and the increase in volume as a result of movement of the projectile. Thus, the gas pressure usually reaches a maximum value as a result of combustion of the propellant charge, and then falls. The propulsion process generally lasts < 10 ms, and accelerations of more than 10^5 g are produced. Muzzle velocities may reach 2200 m/s with high-performance artillery, at maximum gas pressures of more than 700 kPa [59], [60].

The firing process involves not only thermodynamic and gas dynamic processes, but also the reaction kinetics of the decomposition products of the propellant. The firing process can be simulated with internal ballistic computer models.

The gas pressure curve produced on burning the powder can be influenced by the following parameters:

1) Linear burning rate r of the propellant.
2) Treatment of the powder grains. The surface of the powder grain is modified so that it burns more quickly or more slowly than the underlying layers; treatment with plasticizers (e.g., camphor, dibutyl phthalate) produces an initially slower burning rate (progressive behavior). Incorporation of salts

(potassium nitrate) into the powder grain and subsequent dissolution yield a porous powder structure with a large, initial rise in pressure (manoeuver powder).

3) Loading density Δ of the propellant

$$\Delta = \frac{\text{Propellant weight}}{\text{Combustion volume}}\,\text{g/cm}^3$$

The loading density is in general between 0.1 (mortars) and 1.1 g/cm^3.

4) Total free propellant surface area.
5) Shape of the propellant particle.

The quotient of the powder area at time z (A_z) and the initial area A_a is termed the form function $f = A_z/A_a$. The dynamic vivacity is defined as

$$L = \frac{\dot{p}}{p_m \cdot p} \text{ where } p_m = \text{maximum pressure.}$$

Gun Barrel Erosion. Chemical and physical interactions between the high-velocity, hot combustion gases and the wall of the barrel result in barrel wear due to erosion. These phenomena occur particularly in high-performance weapons and powders with high combustion temperatures. Barrel erosion and wear can be reduced by using additives (e.g., titanium dioxide or talcum). To avoid erosion the combustion temperature can be lowered by adding nitroguanidine and replacing nitroglycerine by less energetic explosive plasticizers (cold powders) [59].

Muzzle Flash. Only substoichiometric amounts of oxygen are available to react with the propellant charge powders. Combustible gases (CO, CH_4, etc.) therefore leave the muzzle during firing, which can mix with the air and ignite (muzzle flash). This can be suppressed by adding appropriate compounds (e.g., K_2SO_4 or KNO_3) to the propellant [61].

Charge Technology. The propellant charge can be combined with the projectile to form a munition unit (cartridge munition) which is packed in the loading chamber. Alternatively, the propellant charge is introduced separately into the weapon in powder bags in some cases enclosed in a metal cartridge. Gun propellants may be classified according to the shapes into which they are formed: long tubes (perforated or unperforated), strips, sticks (slotted or unslotted), short-cut perforated tubes, short-cut multiperforated tubes, flakes, and ball powder [62], [63], [64].

Types of Ammunition. Since the early 1980s *caseless ammunition* has been developed for small-caliber weapons. The advantage of this type of ammunition is a short, compact cartridge in which an empty case no longer has to be ejected after firing. It consists of a porous powder body which partially or completely encloses the projectile (Fig. 4).

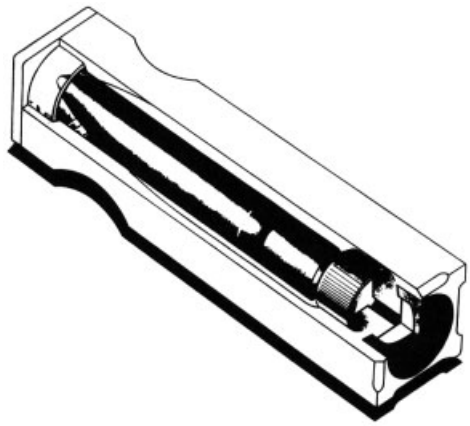

Figure 4. Caseless ammunition (4.73 mm X 33, Dynamit Nobel) for the G11 rifle

Ignition is initiated by a combustible cap. The powder charge often contains a small priming charge that comminutes the propellant charge into small elements during ignition. The propellant charge body is also provided with a combustible protective layer to prevent ignition outside the weapon. Unintentional ignition of caseless ammunition on contact with the hot walls of the loading chamber after previous firing (cook off) may be a problem. Solid propellant compositions having a high thermal stability are therefore used. An example of such a high ignition temperature propellant (HITP) is a finely crystalline, plastic-bonded nitramine (e.g., HMX) [65].

The desire to save material and avoid having to eject powder cartridges has led to the development of *combustible cases* for ammunition for large-caliber weapons. The case consists of a mechanically and thermally stable material like

condensed cellulose fibers or a polymeric material which burns together with the powder charge under the high pressure existing in the weapon without forming a residue. The outer wall of the case also has a protective layer to prevent ignition outside the weapon.

In the case of ammunition that is charged separately, the cartridge contains several partial charges for the different firing zones. *Modular charges* have been developed to reduce the number of partial charges. The ultimate aim is to use only one universal partial charge which can be snapped together to build up the complete charge and provide the required range.

The "unicharge" is packed in a rigid combustible cartridge case which contains its own priming charge. The tubular unicharges are closed at both ends to protect the igniter and propellant from the environment. To ensure that the flame spreads along the full length of the unicharge, barriers are ruptured when the charges are combined.

5.1.2. Liquid Propellants

Spurred on by the rapid development of rocket technology using liquid propellants, experiments began soon after the end of World War II to see whether liquid propellants could also be used for guns. This concept promised considerable advantages:

1) Improvement of specific performance
2) Optimization of the pressure – time curve
3) Avoidance of a cartridge
4) Less space required for the ammunition and thus better space utilization
5) Adjustment of the firing range by controlled metering of the liquids
6) Increase in the firing rate (cadence)
7) Lower propellant costs

This concept also opens up potential solutions for automating many aspects of artillery technology.

Disadvantages were also anticipated, for example a complicated conveyance and metering system, corrosion and handling problems, and an internal ballistics behavior that was largely unknown.

After 45 years of international development, it was found that such systems are feasible [66]. Two systems exist:

1) Bulk-loaded propellant guns (BLPG)
2) Regenerative liquid propellant guns (RLPG)

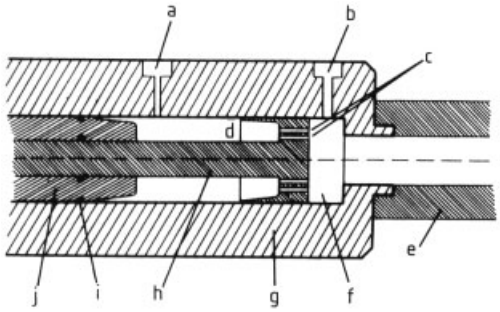

Figure 5. Schematic of a regenerative test fixture for a liquid-propellant gun chamber [59]
a) Propellant fill port; b) Ignition port; c) Injection orifices; d) Propellant reservoir; e) Barrel; f) Combustion chamber; g) Chamber; h) Regenerative piston; i) Seals; j) Breech plug

These systems can be operated with monopropellants (see Section 3.2.1), or bipropellants (see Section 3.2.2).

In BLPG using monopropellants, the loading chamber behind the projectile is filled with the liquid propellant (charging density 1 – 1.45 g/cm^3). Ignition is produced with a pyrotechnic device or submerged electric spark. Although the concept appears extremely simple, the internal ballistics are extremely complex and difficult to control. For example the energy and geometry of the igniter as well as the geometry of the chamber greatly influence the reaction of the liquid. The use of non-hypergolic bipropellants in BLPG has not been investigated in detail since it is even more difficult to control.

With RLPG the combustion process can be controlled more easily. The liquid is injected into the chamber as a result of gas pressure. A model of a monopropellant regenerative test fixture is shown in Figure 5 [59].

The regenerative system can also be used for bipropellants, both liquids are injected separately into the chamber. Such a system is more complex but offers better performance. The advantages of a hypergolic combination (e.g., nitric acid – triethylamine) can be exploited.

Gun systems with liquid propellants are unsuitable for small- caliber weapons. They are becoming increasingly important for machine guns (> ca. 30 mm caliber), but especially for cannons, howitzers and mortars. The regenerative

system is preferred to bulk loading. The suppression of pressure oscillations during combustion poses a special problem, monopropellants have a greater tendency than bipropellants to exhibit such oscillations.

5.2. Rocket Propellants

5.2.1. Solid Propellants

Solid propellants are used in rocket motors and ram jet engines. Rocket motors designed for solid propellants are relatively simple and generally contain no movable parts. Depending on the intended application, they operate at 2 – 50 kPa with burning times of fractions of a second up to several minutes. They are used for propelling projectiles or as a starting aid for aircraft [10], [13], [18]. Up to 70 % of the chemical energy of the propellant is converted into kinetic energy of the exhaust gas jet, of this no more than 70 % can be utilized for propulsion.

Hybrid propellants for rocket motors consist of a liquid and solid propellant, for example a liquid oxidizer that is injected into a combustion chamber containing a fuel grain.

Propellants for ram jet engines produce a combustion gas that is highly deficient in oxygen, and which can be almost completely burnt with the air of the ram jet engine.

The thrust program of a solid propellant motor is fixed and depends on the mission to be fulfilled by the rocket; it is mainly determined by the burning rate of the propellant as a function of pressure and temperature and by the propellant charge configuration. Examples of configurations are shown in Figure 6.

Depending on the physical and mechanical properties of the propellant, the charge may be free-standing (ends are supported against the chamber wall by mountings) or securely fixed to the chamber wall over its entire length (case-bonded grains). The first type include charges with a relatively high tensile strength and low elongation, and are mainly double-base propellants. Composite types (i.e., compositions with rubber-elastic properties) can follow the thermal expansion of the chamber wall and may therefore be securely joined to the latter via an insulation layer or liner. Such charges are "internal burners", which means that the flame reaches the chamber wall only at the end of the combustion.

Propellant charge surfaces that are not intended to burn are covered by an insulation that adheres firmly to the propellant charge and should have similar mechanical properties. These insulations consist of polymers, which may be filled with inert materials. Secure adhesion of the insulation to the propellant charge surface while the rocket motor is in use is important.

The propellant charge can be ignited from the nozzle side or the opposite side. Sealing of the nozzle ensures a rapid rise in pressure during ignition, excludes environmental climatic influences, and ensures safer handling.

Burning of the propellant in the combustion chamber is accompanied by phenomena that are influenced by the propellant, its burning surface, the combustion products, and the design of the propellant charge and combustion chamber. *Erosive burning* denotes the increase in burning rate caused by the flow of product gases across the burning surface of the propellant. The rate may more than double, especially at high gas flow rates. Increased erosion changes the pressure – time curve and may cause failure of the rocket motor [67].

Combustion instabilities may be either acoustic or irregular. In acoustic instability the pressure in the combustion chamber oscillates at high or low frequency. The burning zone acts as an amplifier, the propellant charge and thermal losses as a damping element, and the combustion products as the oscillating medium. The oscillation frequency is related to the dimensions of the combustion chamber.

Irregular instabilities may take the form of pressure variations or a sudden increase and decrease of the combustion chamber pressure caused by specific properties of the propellant [68].

5.2.2. Liquid Propellants

In rocket combustion chambers with a solid propellant charge, the nature and configuration of the charge predetermine the thrust program of the rocket. In rocket motors with a liquid propulsion system, however, the thrust can be varied within wide limits during the mission.

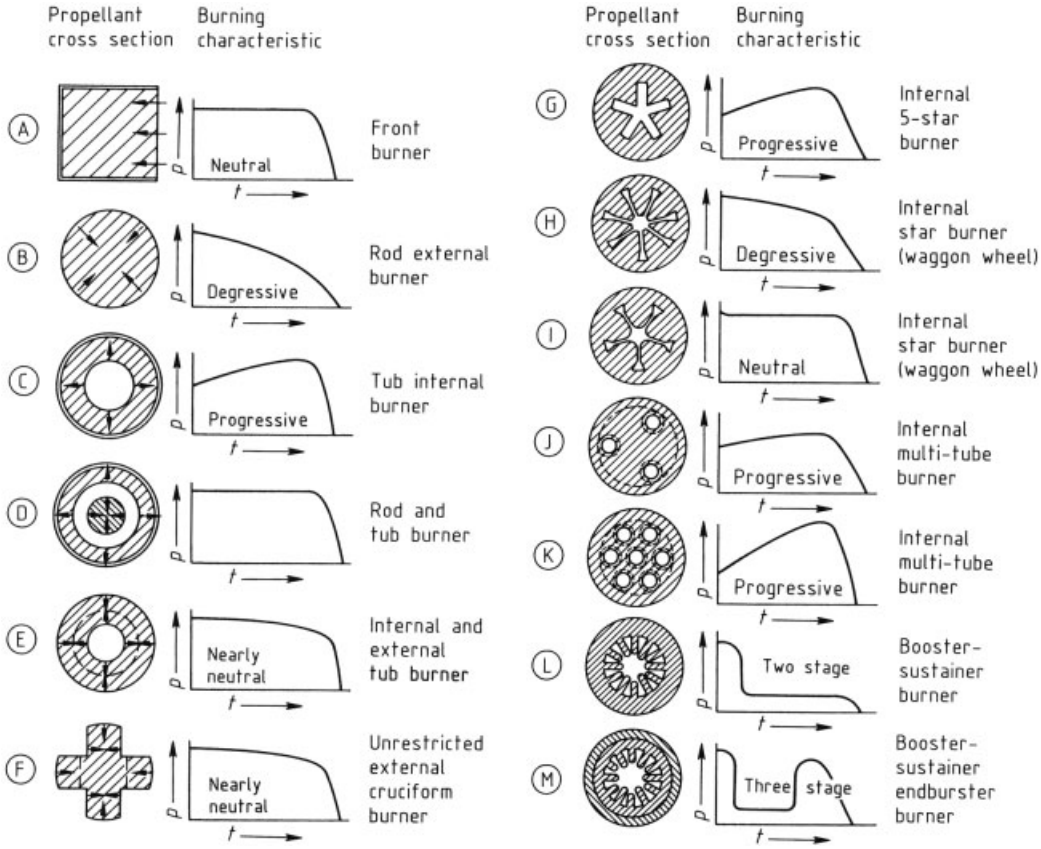

Figure 6. Different grain configurations and their burning characteristics

The propulsion unit consists of a combustion chamber and a nozzle (which are regeneratively cooled in larger rockets), complex injection devices, conveying devices with regulating units, pumps, turbines, and the propellant containers. In smaller rockets the propellants are conveyed to the combustion chamber by the action of pressure on the storage containers.

Monopropellants are injected into the combustion chamber either directly or over a catalyst bed. Rocket motors with monopropellants are normally used as pulsed propulsion units for controlling the position of satellites or rockets. Bipropellants are injected separately into the combustion chamber with a separate conveying system. The injectors are designed so that rapid compression mixing can take place. Ignition is effected either hypergolically or by a pyrotechnic or electric device. The ignition must be accurately coordinated with the injection process to avoid excess pressure in the ignition phase. The few liquid propellant rockets still used for military purposes are already with supplied propellants. Most liquid propellant rockets for civil purposes (space probes and rockets) are fuelled just before the start of flight. Fuelling is extremely slow, particularly with cryogenic propellants, and accordingly the countdown of a large rocket with a liquid propulsion unit lasts many hours [15].

5.3. Propellants for Gas Generators

The function of gas generators is to produce working gases. The charges are of different design that depend on the desired function time and required gas volume. They may be in the form of granules, tablets, or configurations similar to the propellant charges of projectiles or rockets. The gases that are produced activate mechanical devices and inflate or empty vessels and containers.

The combustion charges should therefore produce noncorrosive, low-temperature gases. For reasons of safety and a defined action time, the combustion charges are accommodated in vessels provided with rupture membranes. A typical example is the automobile airbag [69].

Gas generators are used to fill the low-pressure region formed at the rear of a flying projectile and thus by reducing the drag the range can be extended by ca. 20 % (*base bleed*). The mechanical properties of the charges must be able to withstand the high acceleration produced during launching [70], [71].

6. Production

Propellants are energetic materials and have to be handled, manufactured, and transported in such a way as to minimize the risk of accidents. In order to reduce the risk (defined as the probability of an accident times the extent of the damage), a series of independent safety measures are adopted during production and a maximum amount of explosives that can be handled by a defined workforce is specified. Where possible the individual process steps are carried out in special buildings or are separated from one another to minimize the risk of secondary explosions. A remote control is necessary with all processes in which explosive materials are involved and exposed to thermal or mechanical stresses. The manufacturer has to observe safety regulations, this is confirmed and monitored by appropriate safety authorities. A trend towards continuous processing (small amounts of explosive in the individual process steps) can be seen because this improves productivity and reproducibility, and also reduces the risk of accidents. Since the compositions and production processes for gun and rocket solid propellants exhibit many similarities, both types will be discussed jointly. They are classified according to Chapter 3 [3], [5], [7], [42].

Liquid propellants are chemical compounds or their mixtures that are produced by the chemical industry; they do not require special manufacturing processes.

The production of small- caliber ammunition for sport and hunting guns has remained relatively constant over the past few decades. A wide variety of ammunition is produced, largely by private companies. Research, development, and production for military purposes are carried out by government organizations and private companies, in some countries there is a trend toward privatization. The same applies to military rocket propellants. Propellants for aerospace rockets are mainly supplied by producers for the military sector.

Recent world political developments have led to a large reduction of military engineering activities, production is estimated to have declined by 25 – 40 %.

6.1. Homogeneous Propellants

6.1.1. Single-Base Propellants

Batch Solvent Process. This process involves the plasticization of fibrous nitrocellulose by means of solvents (ether – alcohol or acetone) [72–74] which subsequently have to be removed. The process is therefore suitable only for small powder grains with maximum web thickness of at most 13 mm.

The water-wet nitrocellulose (30 %) in the form of a mixed material (e.g., gun cotton containing 13.4 % N and collodium wool containing 12.5 % N in an 80: 20 ratio) or a uniform nitrocellulose (e.g., 12.8 % N) is subjected to displacement with 95 % ethanol in presses or centrifuges. The compressed nitrocellulose is broken up, sieved, and kneaded in a water-jacketed bladed mixer with the addition of appropriate stabilizers, plasticizers, and solvents. The plastic material is then shaped into strips, noodles, or tubes by extruders or pot presses with suitably formed dies. The strips etc. are cut to the desired lengths, graphitized in drums to prevent electrostatic charging, and dried to remove residual solvent. Drying is accompanied by shrinkage of the powder grains. Residual solvent can also be removed by intermediate soaking in hot water. Finally, surface treatment with plasticizers to slow the burning rate at the beginning of the grain combustion can be carried out in drums depending on the desired internal ballistic properties. The individual batches are mixed to obtain lots with reproducible ballistic properties.

Continuous Solvent Process. Single-base propellants can be manufactured continuously

[75]. The alcohol-moist nitrocellulose and additives are fed on a conveyor belt equipped with a continuous weigh-belt into a compounder consisting of a horizontal rotary plow. Solvent is added to form a homogeneous mass, which is then fed to a temperature- controlled screw mixer. The compacted paste is extruded through a die plate and cut into granules. These granules are then fed continuously into an extruder and the extrudate is cut into granules by rotating knives. The solvent is then removed as described above. The continuous mixing and kneading machines and the extruder are all temperature-controlled. On account of the complexity of the machinery and comprehensive control and monitoring system, such a plant is generally only used for large-scale production.

6.1.2. Double- and Multibase Propellants

Solvent Process. Double-base propellants for gun propellant charges can be manufactured by the processes used for single-base propellants [76]. The nitroglycerine is added in the form of a nitrocellulose – nitroglycerine powder preconcentrate or is mixed with the solvent. Propellants containing a small proportion of nitroglycerine are manufactured in this way because the gelatinization effect of the nitroglycerine in a low percentage is insufficient to form a homogeneous material. Triple-base propellants also contain nitroguanidine of generally very small grain size ($d_m = 4\ \mu m$) [77] and can likewise be manufactured by the solvent process. Nitroguanidine has a high gas yield at relatively low combustion temperatures. In order to reduce the combustion temperature of the propellant still further in order to reduce barrel erosion, the combustion temperature of the charge can be lowered by replacing nitroglycerine by less energetic plasticizers such as diglycol dinitrate, 1,2,4-butanetriol trinitrate, or metriol trinitrate.

Solventless Process. Roll and Extrusion Process. The use of solvents can be avoided by plas-ticizing nitrocelluloses with a lower nitrogen content (11.8 – 12.7 % N) with nitroglycerine or other liquid nitrate esters to form a homogeneous colloid. This widespread process is suitable for manufacturing gun- and rocket propellant charges. It has the advantage that the removal of solvent is avoided, larger propellant charges can therefore be produced.

Nitrocellulose and additives are suspended in water at 50 °C, and nitroglycerine and other plasticizers are added. The nitroglycerine is adsorbed by the nitrocellulose and the water is removed in a centrifuge to lower its concentration to 15 –30 %. The thoroughly mixed crude material is freed from residual water by means of heated rolls (95 °C) under friction, and is gelatinized. The homogeneous gelled material then passed between rolls running in the same direction at 60 °C to form carpet rolls of uniform thickness. The carpet rolls are cut into sections, laid together, and fed to an extrusion press where they are pressed at 70 – 100 °C and up to ca. 100 MPa. The feed rate is some centimeters per minute.

Generally powder strands with a diameter between one and up to 200 mm can be produced. The propellant strands are cut into the desired length. Rocket propellant charges are stored under temperature- controlled conditions to eliminate stresses and are checked by X-ray examination for cracks or voids.

This batch process has the usual disadvantages such as inadequate reproducibility and quality assurance. The roll process is labor-intensive and environmentally harmful. Extruding large quantities of propellant material in one batch involves a high safety risk. Individual work stages have therefore been improved (e.g., by remote- controlled operation during rolling). The use of shearing rolls permits continuous operation [78].

Screw Extrusion Process. In order to avoid the disadvantages associated with rolling and extrusion batch by batch in the mid 1960s Dynamit AG began to develop a continuous, fully automatic screw extrusion process for rocket propellants.

The crude material, which contains 25 % water, is fed automatically and continuously to twin-disk mixing devices. After sufficient dewatering and preplasticization, granules are produced in a screw extruder. Residual moisture is removed and the granules are shaped in a twin-screw extruder. Internal propellant with an external charge diameter of 300 mm were produced with this equipment [79], [80].

In the late 1970s twin-screw extruders were used for mixing, dewatering, plasticization, and shaping in the continuous manufacture of double- and triple-base gun propellants [81], [82].

Ball Powder Process. This process was developed in the United States at the end of World War II and yields spherical powder (ball powder) with a diameter of 0.4 – 1 mm. It can be used to recycle nitrocellulose from types of powder that are no longer required. Since the process is carried out in the presence of water, it is relatively safe. Ball powder is mainly used in small- caliber munition because it permits a high packing density and automatic cartridge filling [83–85].

Nitrocellulose is dispersed in warm water in a stirrer, after adding ethylacetate. The solvent evaporates on raising the temperature, and the nitrocellulose precipitates as small droplets. After adding a protective colloid, the solvent is completely evaporated. The solid nitrocellulose balls are then discharged into a washing tank. The balls are washed with water, screened, and passed to a coating tank where they may be enriched with nitroglycerine. Surface treatment with a plasticizer also takes place in the coating tank. Rolled ball powder is used in shotgun cartridges.

Fluid ball powder consists of small nitrocellulose balls with diameters of 0.05 – 0.15 mm. It is produced similarly but with a greater stirring energy, and is used as a starting material for the manufacture of double-base rocket solid propellants by the slurry process. The ball powder may also be treated with nitroglycerine [3], [86].

Casting process. The casting or Steinberger process (Hercules) allows the production of double- and multibase rocket propellants charges with large diameters [86], [87]. A single- or double-base granulate (casting powder, diameter ca. 1 mm) containing the solid additives (e.g., burning modifiers, stabilizers) is first prepared by a solvent process similar to that used for gun propellants. Oxygen- containing salts (e.g., ammonium perchlorate), aluminum powder, and explosives (e.g., RDX or HMX) can also be incorporated to improve the performance. Nitrocellulose with a moderate nitrogen content (e.g., 12.6 % N) ensures satisfactory mechanical properties.

The granular material is metered into a rocket combustion chamber or into molds that are provided with liners. The vessel is vibrated and then evacuated to obtain a high bulk density. A mixture of nitroglycerine and inert plasticizers is then slowly pressed into the powder bed. The liquid fills the cavities in the powder bed and gelatinizes with the casting powder within several days at ca. 60 °C to form a compact body. The process is relatively sophisticated, but has main advantages: (1) propellant charges of any size and configuration can be manufactured, (2) charges consisting of combinations of propellants can be manufactured by using different casting powders.

Composite modified double-base propellants with a high solids content also contain a polymer binder that is compatible with the double-base matrix.

Slurry Casting Process. The slurry casting process is an alternative to the two-stage casting process. A pourable slurry is prepared using small spherical gelatinized nitrocellulose particles (fluid ball) and plasticizers.

The thixotropic fluid ball is stirred with a mixture of nitroglycerine and inert plasticizers in planetary stirrers in vacuo, other additives follow. The slurry is then poured under vibration and reduced pressure into casting molds and hardened for several days at 50 – 70 °C to form compact bodies. The components used in this process for producing nitrocellulose plastisol propellants have been widely varied. For example, other energetic plasticizers have been used instead of nitroglycerin, such as pentaerythritol trinitrate (PETriN). As with the casting process, oxygen- containing salts, aluminum powder, and nitramines can be incorporated into the slurry. Cross-linked polymers can also be added. Instead of fluid balls, poly(vinyl nitrate) or poly(vinyl chloride) granules have also been used. By using cross-linked polymers and crystalline fillers, the slurry casting process and its products form the transition to composite propellants produced by analogous processes [16], [88].

6.2. Composite Propellants

Casting Process. The manufacture of most rocket propellant charges based on composite

propellants follows the same process, namely by mixing the components into a viscous liquid which is then poured into molds and cured at elevated temperatures. Composite propellants consist of a polymeric binder and a powdery filler (Section 3.1.2). The binder serves both as a fuel and as a matrix for accommodating the filler; the latter contains chemically bound oxygen for combustion (e.g., ammonium or potassium perchlorate, ammonium nitrate). Typical explosives such as HMX, RDX or NQ and TAGN may also be used. Metal powder (e.g., aluminum powder) is sometimes added to increase the combustion temperature and avoid instability during burning [11], [16], [89], [90].

Pretreatment. Although a large number of cross-linked polymers have been tested, hydroxyl-terminated polybutadienes (HTPB) are now mainly used for the casting process. The liquid prepolymer is carefully dried (moisture content $< 0.1\,\%$) because water can react with the cross-linking agents or adversely affect the wetting of the fillers. The prepolymers are subjected to strict quality control, particularly as regards their proportion of reactive groups, to obtain optimum mechanical properties.

The concentration and grain size distribution of the filler influence the viscosity and thus the processing of the casting materials, as well as the mechanical properties and burning behavior of the propellant. The fillers therefore have defined grain size distributions. Compositions containing two or three different grain sizes are used to obtain good casting properties in the sense of a high packing density. The oxygen- containing salts are carefully dried and then ground to the desired size in mechanical mills or fluid-energy pulverizers (grain size $< 5\,\mu m$). An anticaking agent is often added to prevent agglomeration of the finely ground material which should be processed within a few hours. Friction-sensitive fillers such as explosives are ground in the presence of water in an apparatus equipped with high-speed rotors, or are produced by crystallization.

Mixing. The binder prepolymer is first mixed with the inert additives (metal powder, burning modifiers, wetting agents, etc.) and, if necessary, with inert or explosive plasticizers. Modern mixers have vertically operating planetary kneader blades and are equipped with a vacuum device. The mixer has a double jacket for heating the materials to ca. 40 – 60 °C. After the binder and additives have been thoroughly mixed, the oxygen-containing salts and explosives are added. The mixture is kneaded under vacuum for several hours to ensure good wetting and degassing. After adding and mixing the curing agents are added.

Casting. The homogenized casting material is fed into the pouring device by an outlet at the bottom of the mixer or by tipping it. Highly viscous materials can be forced out by means of a plunger. The pretreated and evacuable casting molds often have a perforated plate at the filling end to ensure uniform filling. In many cases the materials are poured directly into the rocket chambers which contain a profiled core in the case of an internal burner. With highly viscous materials this core may be also pressed into the propellant after pouring. The casting device and the molds can be made to vibrate to improve the flow. The formulation of the propellant must be adjusted so as to ensure a sufficiently long pot-life. This is especially important with large propellant charges where casting can take several hours. The rocket chambers or casting molds are placed in heated chambers where they were cured for several days at ca. 60 °C. Rapid curing at relatively high temperatures should be avoided to prevent internal stresses, which are especially important with charges of large diameters. The casting devices are then allowed to cool, and any cores are removed.

Propellant charges of any size can, in principle, be manufactured by the casting process. The production of charges with small diameters is, however, only economical if extrusion is more expensive or more hazardous. The upper limit is governed by the capacity of the mixer and permitted safety distance. Several time-phased mixers are used for very large propellant charges.

Attempts have been made to employ continuously operating mixing and kneading machines, such as extruders, to avoid mixing of large amounts of propellant. Since the residence time in extruders is limited, the mixing and kneading time is often insufficient to produce an optimum propellant in the case of highly filled composite propellants. To overcome this problem equipment has been developed in which several individually controlled continuously fed, mixing chambers are connected in series. The

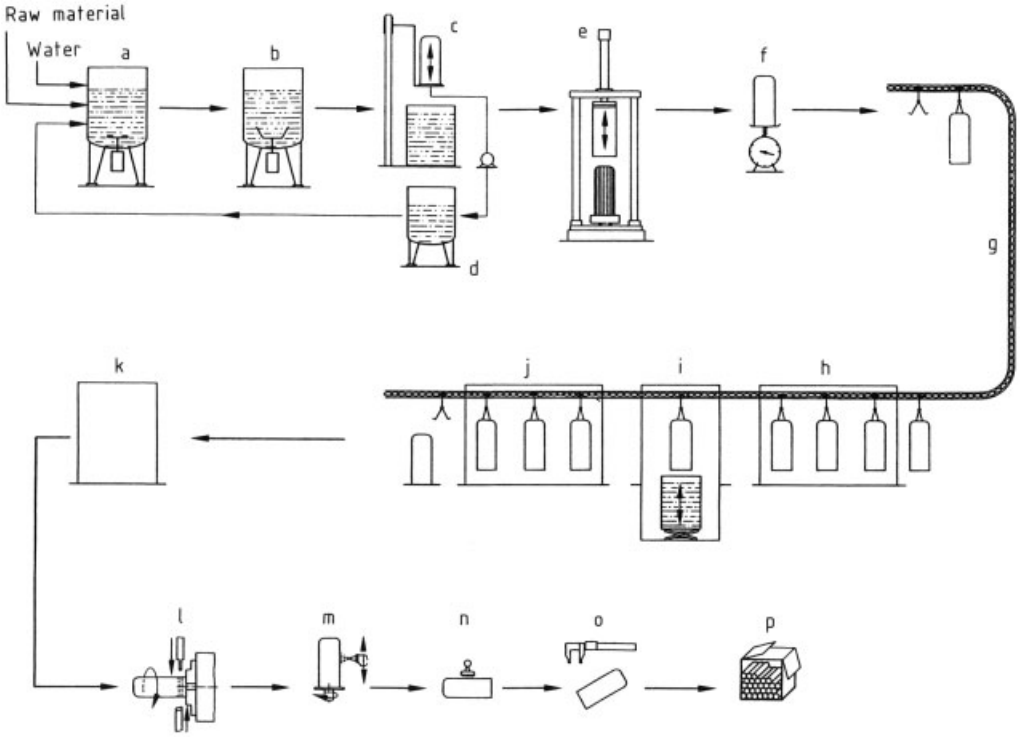

Figure 7. Flow sheet of production of combustible cases (WNC-Nitrochemie, Germany)
a) Mixer, b) Pulp tank; c) Matting; d) Water tank; e) Press; f) Weighing; g) Chain conveyor belt; h) Drying tunnel; i) Impregnation; j) Drying tunnel; k) Conditioning (21 °C, 65 % R.H.); l) Machining; m) External coating; n) Marking; o) Quality control; p) Packaging and storage

casting material can be removed continuously via an extruder outlet [91].

Extrusion Process. An extrusion process has to be employed for propellants with thermoplastic binders or propellant mixtures with a viscosity that is too high for casting. The pretreatment of the components and the mixing of the propellant materials are carried out as described before [92]. The highly viscous materials are kneaded in extruders and then forced through suitable dies into molds or combustion chambers. The processing temperatures (80 – 120 °C) are considerably higher than in the casting process. Horizontal mixers have also been used, which have an outlet screw in the bottom of the trough.

The development of insensitive gun propellants (LOVA) has led to products with chemical compositions comparable to those of composite propellants. They consist of a thermoplastic binder or a cross-linked polymer, explosive and nonexplosive plasticizers, and nitramines (e.g., RDX, HMX, NQ, or NTO). Also gun propellants of the composite type with thermoplastic binders are extruded, after conventional mixing and kneading, from extruders through dies. Propellants containing cross-linked polymers have to be after-cured at elevated temperatures (60 °C). For a continuous process double-screw extruders are used.

6.3. Special Propellant Devices

Caseless Ammunition. The first generation of caseless ammunition consisted of powder or granules that were compressed into mechanically stable grains with the aid of solvents or adhesives. The projectile and the consumable ignition device were embedded in the grain. Although the grains were provided with a protective coating they sometimes ignited unintentionally in previously fired guns (cook off). Com-

posite propellants containing energetic materials with a high thermal stability are used to overcome these problems. Finely crystalline HMX is homogenized with an inert binder and the mass formed into a porous grain. The porosity is chosen such that the specimen reacts within the required time. Polynitrophenylene is employed as an energetic material for caseless ammunition for the G11 rifle by Dynamit Nobel (Fig. 4) [64], [65], [93].

Combustible Cases. The production of combustible cases from matted cellulose fibers was developed in the United States. Figure 7 shows a modification of this process for the manufacture of 120-mm ammunition (Fig. 8) [94]. Cases for modular propellant charges are produced similarly [95].

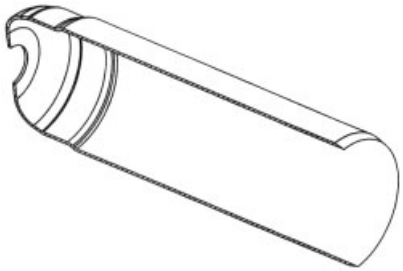

Figure 8. Combustible case for 120-mm ammunition of the Leopard 2 armed tank (Rheinmetall)
The projectile is subsequently inserted at the front of the case (right-hand side). The shaped part at the back (left-hand side) accomodates the steel base.

A pulp consisting of cellulose fibers and other raw materials in a water bath are deposited on a mold by sucking off the water. The fibers are distributed uniformly in all directions, resulting in matting. The case already has some mechanical stability, its shape is determined by the geometry of the mold. The final shape is obtained by pressing and machining. After pressing, the cases may be immersed in an impregnation bath to give the required mechanical stability. This step is not necessary if special impregnating additives are included in the pulp (Figs. 7, 8).

7. Test Methods

The properties of propellants must be studied in research and during their development, production, and use. Tests are used to determine safety of handling (permission for transport), chemical stability and compatibility, mechanical and physical properties, and internal ballistic properties. These tests can be supplemented by other methods to measure process parameters and sensitivity toward external influences.

Many test methods have been standardized at a national or international level. Test methods for the *military sector* are specified in

France	Manuel des modes operatives
Germany	Technische Lieferbedingungen (TL)
NATO	Standardisation Agreement (STANAG)
United Kingdom	Defence Standard (DEFSTAN)
United States	Military Standard (MIL-Std.)

The following organizations are responsible for the *civilian sector*:

France	Centre d'Etudes et Recherches de Charbonnages de France (CERCHFR)
Germany	Bundesanstalt für Materialforschung (BAM)
Japan	Japanese Industrial Standards
United States	National Bureau of Standards
UNO	United Nations Committee of Experts on the Transport of Dangerous Goods

Tests for safe handling and chemical and physical properties are the same as those used for explosives (→ Explosives, Chap. 8.). Other special internal ballistic tests must, however, also be carried out.

Burning Behavior (Ballistic Bomb). The burning behavior of a propellant powder is investigated with a pressure proof (≤ 1 MPa) screw-lid vessel made of steel and fitted with a piezoelectric pressure recorder. The pressure in the *ballistic bomb* is determined as a function of time. Alternatively, the rate of pressure increase is measured as a function of pressure. These parameters, especially the dynamic vivacity of a powder, are used for the preliminary design of a powder charge. If the form function is known, the linear burning rate of the powder can be calculated [96].

Gas pressure (crusher) gages similar to those used in weapons are used to measure the pressure curve of gun propellant charge. In the gage, however, the gun barrel is replaced by a rupture diaphragm.

Burning Rate. The linear burning rate of solid rocket propellants is measured in a pressure vessel known as a *Crawford Bomb* . Propellant rods (diameter 4 – 5 mm) are burned at a specified pressure (2 – 25 kPa). The completely insulated rods are ignited at one end and burn perpendicular to the burning surface. The burning rate is measured with wire detectors. The pressure is kept virtually constant by means of an equalization vessel. Similarly constructed bombs are used to determine the linear burning rate of gun propellant charges ($\leq$ ca. 600 kPa) [97].

Standardized combustion chambers of widely differing design are employed to determine the burning behavior of rocket propellant charges.

Heat of Explosion. The heat of explosion of a propellant powder (loading density generally 0.1 g/cm^3) can be calculated from the difference between the energies of formation of the powder constituents and the combustion products. Unlike the heat of combustion, the experimentally determined heat of explosion depends on the loading density. The calorific value is usually related to liquid water [96].

8. Safety Precautions and Environmental Protection

The production, handling, and transport of propellants are governed by national laws and regulations and international recommendations [58]. Safety measures are based on the classification of the energetic material as a dangerous substance. Regulations for the construction of storage and production facilities, quantitative limits, safety areas etc. are designed to minimize the effects on the environment in the case of an incident. The environmental effects caused by the production, use, and disposal of propellants depend on their chemical composition. In the case of liquid propellants, hydrazine, its derivatives, and fluorine- or chlorine- containing oxidizers are especially dangerous as potential pollutants. The following components of solid propellants must be classified as environmental pollutants: perchlorates, heavy-metal salts, and aromatic compounds (stabilizers).

Environmentally hazardous products, such as nitrogen oxides and hydrogen chloride, can be formed when propellants are burnt. Less nitrogen oxides are formed when combustion takes place at high pressure. The large quantities of hydrogen chloride that are produced during the start of a space rocket are especially hazardous for the environment if these are released at high altitude.

The environmentally safe disposal of propellants associated with the demilitarization of munition and abandoned polluted areas (e.g., ammunition factories, training areas, airports) presents a serious task for relevant experts.

As in the case of civilian products, material recycling is preferable to chemical, biological, or thermal treatment [98]. Biological methods will become increasingly important for treating contaminated soil.

The toxicological effects of propellant components and processing aids are described in national dangerous goods regulations.

9. References

1. Periodicals *Propellant Explos.* **1** (1976) – **6** (1981); *Prop. Expl. Pyrotech.* **7** (1982) to date. *Explosivstoffe* **1** (1952) – **22** (1974) to date. *J. Energetic Materials,* Dowden, Brodman and Devine, Stroudsburg, PA, USA, 1 (1983) to date. *ARS J.* **1** (1930) – **32** (1962); *AIAA J.* **1** (1963) to date. *J. Propulsion and Power,* Americ. Inst. of Aeron. a. Astron., Washington, USA, 1985 to date. *Combustion and Flame,* Elsevier, New York, N.Y., USA, 1 (1901) to date. *J. Propulsion and Power,* AIAA, Washington, USA, 1 (1985) to date. *J. Spacecr. Rockets* **1** (1963) to date. *Jet Propul.* **1** (1930) – **28** (1958).
2. Proceedings of Symposia *Int. Jahrestag Inst. Chem. Treib- Explosivst.* 1970 – 1987, continued as Institut für Chemische Technologie (ICT), Pfinztal/Karlsruhe, Germany, 1988 to date. *Symposium (Intern.) on Combustion,* Combustion Institute, Pittsburgh, USA, 1 (1928) to date. *Int. Symp. Ballist.* **1** (1974). Joint Intern. Symposium on Compatibility of Plastics and other Material with Explosives, Propellants and Pyrotechnics and Processing of Explosives, Propellants and Ingredients, followed 1992 by Joint Intern. Symposium on Energetic Materials Technology, ADPA, Arlington, Va, USA 1979 to date. *AIAA-Conference,* Am. Inst. of Aeron.

and Astron., Washington, USA, 1964 to date. *Int. Pyrotechnics Seminar,* Intern. Pyrotechnics Society, Chicago, Il., USA 1968 to date. J. Hansson (ed.): *Pyrotechnidagen* **1** (1971) to date. *International Gun Propellant Symposium* **1** (1973) to date.

General References

3. *Ullmann,* 4th ed., **20,** 91 – 112, **21,** 682 – 693.
4. J. Köhler, R. Meyer: *Explosives,* 4th ed., VCH Verlagsgesellschaft, Weinheim 1993.
5. T. Urbanski: *Chemistry and Technology of Explosives,* Pergamon Press, Oxford 1983 – 1985.
6. B. T. Federoff et al.: *Encyclopedia of Explosives and Related Items,* U.S. Army and National Technical Information Service, Springfield, Va, USA 1960 – 1983.
7. *Kirk-Othmer,* 3rd ed., **9,** 620 – 663.
8. J. Quinchon et al.: *Les Poudres, Propergols et Explosifs,* **vol. 1 – 4,** Technique et Documentation, Paris 1982 – 1991.
9. S. Fordham: *High Explosives and Propellants,* 2nd ed., Pergamon Press, Oxford 1980.
10. Y. M. Timnat: *Advanced Chemical Rocket Propulsion,* Academic Press, London 1987.
11. A. E. Oberth: *Principles of Solid Propellant Development,* A. E. Oberth, USA, 1986, US Government, DAR clause 8-104.9.
12. Summerfield: "Gun Propulsion Technology," in *Progress in Astronautics and Aeronautics,* vol. 109, Am. Inst. of Aeron. and Astron., Washington 1988.
13. G. P. Sutton: *Rocket Propulsion Elements,* J. Wiley & Sons, New York 1975.
14. F. A. Warren: "Solid Propellant Technology," *Selected Reprint Series,* **vol. 10,** Am. Inst. of Aeron. and Astron., Washington 1970.
15. A. Dadieu et al.: *Raketentreibstoffe,* Springer Verlag, Wien 1968.
16. C. Boyars, K. Klager: *Propellants Manufacture, Hazards and Testing,* Am. Chem. Soc., Washington 1969.
17. S. F. Sarner: *Propellant Chemistry,* Reinhold Publ. Corp., New York 1966.
18. M. Barrère et al.: *Raketenantriebe,* Elsevier Publ. Comp., London 1961.
19. S. S. Penner, J. DuCarme: *The Chemistry of Propellants,* Pergamon Press, Oxford 1960.
20. Chem. Propuls. Info Agency: *Hazards of Chemical Rockets and Propellants Handbook,* vol. 1 – 3, The Johns Hopkins University, Silver Spring, Maryland, USA 1972.
21. Military Explosives, Depts. of the Army and the Air Force, TM 9-1300-214, Washington 1967.
22. *A. Davenas: Solid Rocket Propulsion Technology,* Pergamon Press, Oxford 1993.

Specific References

23. B. T. Fedoroff, O. E. Sheffield: *Encyclopedia of Explosives and Related Items,* PATR 2700, vol. 7, Picatinny Arsenal, Dover, N.J., USA, 1975, pp. H114 –H159.
24. E. Gray, H. Marsh, M. McLaren, *J. Mater. Sci.* **17** (1982) 3385 – 3400.
25. *Fang-Toh Sun, 35th Congress of the Int. Astron. Federation,* IAA Paper 84-258 (1984).
26. Tin Ching et al., *Kogyo Kayaku* **47** (1986) no. 2, 111 –118.
27. A. Langhans, *Z. Gesamte Schieß Sprengstoffwesen* **33** (1938) no. 1, 1 – 65.
28. F. Trimborn, *Nobel Hefte* **55** (1989) 107 – 159.
29. *ICT Intern. Jahrestg.* **24** (1993) Energetic Materials – Insensitivity and Environmental Awareness.
30. W. Klöhn et al., *Propellants Explos. Pyrotech.* **12** (1987) 71.
31. S. Eisele et al., *Propellants Explos. Pyrotech.* **17** (1992) 155.
32. F. Schedlbauer, *Propellants Explos. Pyrotech.* **17** (1992) 164.
33. W. Engel, *Explosivstoffe* **21** (1973) 9.
34. G. Langer et al., *Propellants Explos. Pyrotech.* **17** (1992) 185.
35. W. F. Morrison et al.: *Liquid Propellant Guns,* US Army Ball. Res. Lab., AIAA, Washington, OMB Na 0704 – 0188, 1986.
36. K. K. Kuo, M. Summerfield: "Fundamentals of Solid Propellant Combustion," *Prog. Astronaut Aeronaut* **90** (1984).
37. B. Zeller et al.: "Combustion of Solid Propellants," *Lect. Ser.* **180** (1991).
38. N. S. Cohen, *AIAA J.* **18** (1980) 277.
39. N. S. Cohen, *AIAA J.* **23** (1985) 276.
40. M. Summerfield et al., *Prog. Astronaut Rocketry* **1** (1960) 141.
41. M. W. Beckstead et al., *13th Symp. (Int.) Combust. [Proc.]* (1971) 1047.
42. E. Freedman: Blake, A Thermodynamics Code Based on Tiger "Users" Guide and Manual, ARBTL-TR-02411, 1981.
43. M. Cowperthwaite, W. H. Zwisler, Tiger Computer Program Documentation, SRI Publication no. Z 106, 1973.
44. M. Hund et al.: "Determination of Interior Ballistics Parameters of Solid Propellants by Different Methods," *Int. Ballistics Symposium Proceedings,* Orlando 1981.

45. F. Volk, H. Bathelt: *User's Manual for the ICT-Thermodynamic Code,* **vol. 1 – 3,** ICT, Berghausen-Karlsruhe 1988/1991; *JANAF Interim. Thermochemical Tables,* **vols. 1, 2,** The Dow Chemical Company Midland, Michigan 1960.
46. F. N. Kelley: in [16], p. 188.
47. H. Schubert, D. Schmitt, *Explosivstoffe* **11** (1963) 1; **14** (1966) 8.
48. F. Volk et al., *Propellants Explos. Pyrotech.* **12** (1987) 81.
49. F. Volk, G. Wunsch, *ICT Intern. Jahrestg.* **17** (1986) 13.
50. J. Quinchon: in [8], vol. 3, p. 116.
51. H. Schubert et al., *AIAA/SAE Pap. Propul. Conference,* 13th, Orlando, Fla., USA 1977.
52. M. A. Bohn, F. Volk, *Propellants Explos. Pyrotech.* **17** (1992) 171.
53. "Lifetime of Rocket Propellants, Gun Propellants and Explosive Charges," *ICT Int. Jahrestg.,* ICT, Pfinztal (1971).
54. Propellants, Sampling, Examination and Testing, U.S. Military Standard 286 A, 1961.
55. "Testing Methods for Propellants and Explosives," *ICT Intern. Jahrestg.,* ICT, Pfinztal (1980).
56. Explosives Hazard Classification Procedures, TB 7002 NAVORD Inst. 8020 3 to 11 A-1-47 DSAR 8 220.1 Depts. of the Army, Air Force, and Defence Supply Agency, Washington, D.C. 1967.
57. Allied Ordonance Publication (AOP 7), Manual of Tests for the Qualification of Explosive Materials for Military Use, NATO 1988.
58. United Nations Organization: *Transport of Dangerous Goods,* 3rd ed., United Nations Publication, New York 1987.
59. L. Stiefel: "Gun Propulsion Technology," *Prog. Astronaut and Aeronaut* **109** (1988).
60. P. Baer: "Practical Interior Ballistic Analysis of Guns," *Prog. Astronaut Aeronaut* **66** (1979) 37.
61. G. Klingenberg, J. Heimerl: "Gun Muzzle Blast and Flash," *Prog. Astronaut Aeronaut Sec. Textbook* **vol. 13,** AIAA Publication, Washington, D.C. 1991.
62. J. Quinchon: in [8], vol. 3 (1986) 157
63. Ullmann, 4th ed. **21** (1982) 682.
64. J. H. Kelley, Gun Propulsion: Energing Army Technologies, Advances Prop. Tech. Study, Jet Prop. Laboratory, Pasadena, 1986, JPLD –3294.
65. K. H. Redecker et al., *Propellants Explos. Pyrotech.* **12** (1987) 196.
66. N. Klein: in [58], p. 473.
67. M. K. Razdan, K. K. Kuo: in [37], p. 515.
68. E. W. Price, J. S. T'ien: in [37], chap. 13, 14.
69. Intern. Symposium: *Airbag 2000, Proceedings,* ICT, Pfinztal/Karlsruhe 1992.
70. N. E. Gunners: in [37], chap. 16.
71. K. Andersson et al., *Propellants Explos.* **1** (1976) 69.
72. K. Fabel, Nitrocellulose, Enke Verlag, Stuttgart 1950.
73. in [5], vol. 2, 3
74. A. M. Ball: *Solid Propellants, Engineering Design Handbook,* **Part 1,** AMCP USA Material Command, Washington 1964, pp. 706 – 775.
75. C. J. v. Helle, *Propellants Explos. Pyrotech.* **13** (1988) 55.
76. G. Buschmann, H. U. Reich, *ICT Intern. Jahrestg.* **9** (1978) 199.
77. R. Goldstein et al.: *Process Engineering of Triple Base Propellants,* DT-TR-8-16, Picatinny Arsenal, Dover, N.Y. 1960.
78. J. Knobloch et al., *ICT Intern. Jahrestg.* **18** (1987) 5 – 1.
79. A. Homburg, H. Brachert, *Propellants Explos.* **2** (1977) 69.
80. D. Girke, H. P. Mackowiak, *ICT Intern. Jahrestg.* **18** (1987) 70 – 1.
81. D. Müller, *ICT Intern. Jahrestg.* **15** (1984) 97.
82. D. Müller, Joint Intern. Symp. on Compatibility of Plastics and other Materials; Processing of Explosives and Propellants, ADPA New Orleans 1988; The Pointe, Phoenix, Ar., USA (1982).
83. in [6]vol. 1, TR 2700 PAT (1962) B 11.
84. in [5], vol. 3, 1967, p. 632.
85. T. R. Olive, *Chem. Eng. (N.Y.)* **53** (1946) 136.
86. R. Steinberger in [19], p. 246.
87. R. Steinberger in [16], p. 1.
88. R. F. Preckel, *AIAA J.* **3** (1965) 2346.
89. J. W. Keating, F. Sage, *ICT Intern. Jahrestg.* **9** (1978) 87.
90. W. Klöhn, *ICT Intern. Jahrestg.* **9** (1978) 109.
91. W. Klöhn, *ICT Intern. Jahrestg.* **18** (1987) 4 – 1.
92. W. Klöhn, A. Rassinfosse, *ICT Intern. Jahrestg.* **12** (1981) 289.
93. D. Schmitt, *Propellants Explos.* **2** (1977) 49.
94. P. C. Bonnett, M. Collins, *ICT Intern. Jahrestg.* **18** (1987) 21 – 1.
95. J. Eich, *ICT Intern. Jahrestg.* **18** (1987) 22 – 1.
96. In [8], vol. 3, chap. 11.
97. B. L. Crawford et al., *Anal. Chem.* **19** (1947) 630.
98. Waste Management of Energetic Materials and Polymers, *ICT Intern. Jahrestg.* **23** (1992).

Propene

PETER EISELE, Linde AG, Werksgruppe Verfahrenstechnik und Anlagenbau, Höllriegelskreuth, Federal Republic of Germany (Chaps. 1, 2, 3, 5; Sections 4.1, 4.2)

RICHARD KILLPACK, Shell International Chemical Company Ltd., Shell Centre, London, United Kingdom (Chaps. 6, 7; Section 4.3)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 2
4.1. Production as Byproduct of Ethylene Production 2
4.1.1. Variants of Propene – Propane Separation 4
4.1.2. Apparatus and Construction Materials 6
4.2. Production as Byproduct of Refinery Processes 6
4.3. Propane Dehydrogenation 7
4.3.1. Oleflex Process 7
4.3.2. Catofin Process 8
4.3.3. Phillips STAR Process 8
4.3.4. Linde Process 8
5. Storage, Transportation, Quality Requirements 10
6. Uses 10
6.1. Thermal Uses 10
6.2. Motor Gasoline Uses 10
6.2.1. Polygasoline Production–Dimerization 11
6.2.2. Alkylation 11
6.3. Chemical Uses 12
7. Economic Aspects 13
8. References 13

1. Introduction

Propene, $CH_3-CH{=}CH_2$, M_r 42.081, [*115-07-1*] was the first petrochemical raw material to be employed on an industrial scale and was used more than 60 years ago in the production of isopropanol. For a long time, propene was to some extent overshadowed by its olefin homologue ethylene. However, part of propene's turbulent development and expansion since 1965 is also due to the success of ethylene: Being a byproduct of ethylene production, many important areas of application were opened up to propene by the chemical industry. Secondary products of propene, such as polypropylene, acrylonitrile, propylene oxide, and in Europe cumene, have now clearly overtaken the classic secondary product isopropanol in importance.

2. Physical Properties

Important physical properties are as follows:

mp	− 185.25 °C
bp	− 47.70 °C
t_{crit}	91.76 °C
p_{crit}	4.621 MPa
ϱ_{crit}	0.22 g/cm^3
Molar volume	21.976 L (STP)
Gas density	1.9149 g/L (STP)
Density relative to air	1.49
Heat of fusion	71.37 kJ/kg
Lower heating value	45 813 kJ/kg
Enthalpy of formation ΔH_{298}	20.43 kJ/mol
Entropy S_{298}	0.227 kJ mol^{-1} K^{-1}
Free energy of formation ΔG_{298}	62.65 kJ/mol
Explosion limits in air (at 1 bar and 20 °C)	
Lower	2.0 vol % (35 g/m^3)
Upper	11.1 vol % (200 g/m^3)
Ignition temperature	455 °C

The temperature dependence of typical properties is given in Table 1.

3. Chemical Properties

The chemical properties of propene are, like those of ethylene, characterized by the reactivity

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_211

Table 1. Temperature dependence of some physical properties of propene, according to [4]

Temperature, °C	Vapor pressure, kPa	Heat of vaporization, kJ/kg	Specific heat, kJ kg^{-1} K^{-1}		Density, g/cm^3	η_L, mPa · s	η_G, 10^{-4} mPa · s	Surface tension, mN/m	Thermal conductivity, mW m^{-1} K^{-1}	
			c_{pL}	c_{pg}					λ_L	λ_G
−185.3		568	2.17		0.759	6.76		37.7	237	
−175		560	2.16		0.749	3.52		36.0	231	
−150		540	2.12		0.725	1.14	33	32.1	216	
−125	0.33	518	2.07		0.699	0.53	40	28.2	201	
−100	4.0	494	2.04		0.672	0.33	48	24.4	186	5.8
− 75	24	469	2.04		0.644	0.235	55	20.7	171	8.1
− 50	91	441	2.09		0.614	0.180	62	17.1	156	10.6
− 47.7	101	438	2.09		0.611	0.176	63	16.8	154	10.8
− 25	256	409	2.20		0.582	0.140	70	13.6	140	13.1
0	584	374	2.40	1.41	0.546	0.108	77	10.2	125	15.7
25	1150	331	2.70	1.52	0.506	0.081	84	7.0	110	18.4
50	2060	277	3.12	1.62	0.457	0.058	91	4.1	95	21
75	3400	196		1.71	0.389	0.039	98	1.4	80	24
91.9	4620			1.78	0.225		102			26
100				1.81			105			27
125				1.90			111			30
150				1.99			118			33
175				2.08			125			36
200				2.16			131			40
400				2.75			181			67
600				3.20			225			96
800				3.54			264			125
1000				3.80			298			152

of its double bond. Propene undergoes a number of industrially important polymerization, addition, and oxidation reactions (see Chap. 6).

4. Production

Although propene is one of the most important feedstocks for the organic chemicals industry, it is produced almost entirely as a byproduct because it is obtained in sufficient amounts in ethylene production by steam cracking and in some refinery processes (primarily cat cracking).

Whereas in Europe, *refineries* satisfy on average only 20 % of the chemical industry's consumption of propene [5], in the United States they meet more than 40 % of the consumption demand [6], [7]. Despite this high figure, refineries in the United States consume ca. 75 % of their propene production in in-house, nonchemical applications [8]. This propene excess is utilized in gasoline production (alkylate and polymer gasoline; see Section 6.2), (→ Oil Refining, Chap. 3.7.), to produce liquefied petroleum gas (LPG, → Liquefied Petroleum Gas), and as heating gas.

A trend toward less severe cracking conditions and thus to increased propene production has been observed in *steam cracker plants* using liquid feedstock that have been designed since the mid-1980s. The increased consumption of propene has also boosted the demand for processes for propene production by *catalytic dehydrogenation* of propane. Various processes suitable for this purpose have been developed and in some cases tested.

4.1. Production as Byproduct of Ethylene Production

The influences of cracking feedstocks and cracking conditions on the composition of the cracked gas are described in detail in Ethylene (→ Ethylene, Chap. 5.1.).

A sour-gas-free, anhydrous C_3 fraction is obtained as overhead product from the depropanizer (a) in the processing of cracked gas. This fraction contains the C_3 hydrocarbons of the cracked gas (i.e., propane, propene, propadiene, and propyne) as well as traces of C_2 and C_4 hydrocarbons. Because of its propadiene and propyne content (depending on the cracking conditions, these can total up to 8 mol %), this C_3 fraction does not meet product specifications and therefore requires further treatment.

Figure 1 shows a typical flow diagram for the further processing of the C_3 fraction in a steam cracker plant. The C_3 fraction is fed to a selective hydrogenation unit (b) to remove propadiene and propyne. This hydrogenation can be performed in the gas and liquid phases (pressure ca. 18 bar), palladium catalysts being used in both. The amount of hydrogen added is calculated so that, on the one hand, complete conversion of C_3H_4 to C_3H_6 is achieved, and on the other, the smallest possible amount of propene is hydrogenated to propane. In practice a molar $H_2 : C_3H_4$ ratio of ca. 1.5 has proved suitable.

The reaction conditions of C_3 hydrogenation in the gas phase differ from those in the liquid phase. In gas-phase hydrogenation the reaction is controlled by means of the operating temperature, which may be between 50 and 120 °C, depending on the preparation and aging state of the catalyst. With liquid-phase hydrogenation the reaction is controlled by the hydrogen partial pressure. The operating temperature of 15 – 25 °C is considerably lower in this case. Adiabatic fixed-bed reactors with intermediate coolers, as well as isothermal tubular reactors, have proved suitable for the exothermic hydrogenation reaction in both phases.

Characteristic of a hydrogenation with metered addition of hydrogen is the formation of smaller amounts of oligomers (mainly dimers and trimers). If required by the propene product specification these can be scrubbed out following hydrogenation with a small amount of C_3 in scrubber (c). The mixture of oligomers and scrubbing agent is recycled to the depropanizer for C_3 recovery, and the oligomers are subsequently led into the pyrolysis gasoline fraction.

The methane introduced with the hydrogen (in liquid-phase hydrogenations, the unconsumed hydrogen as well) is then stripped off (d) and recycled to the cracked gas processing unit of the steam cracker to recover entrained propene. The bottom product of the C_3 stripper meets the product specification of *chemical-grade propene.*

Chemical-grade propene is purified to polymer-grade propene in a downstream

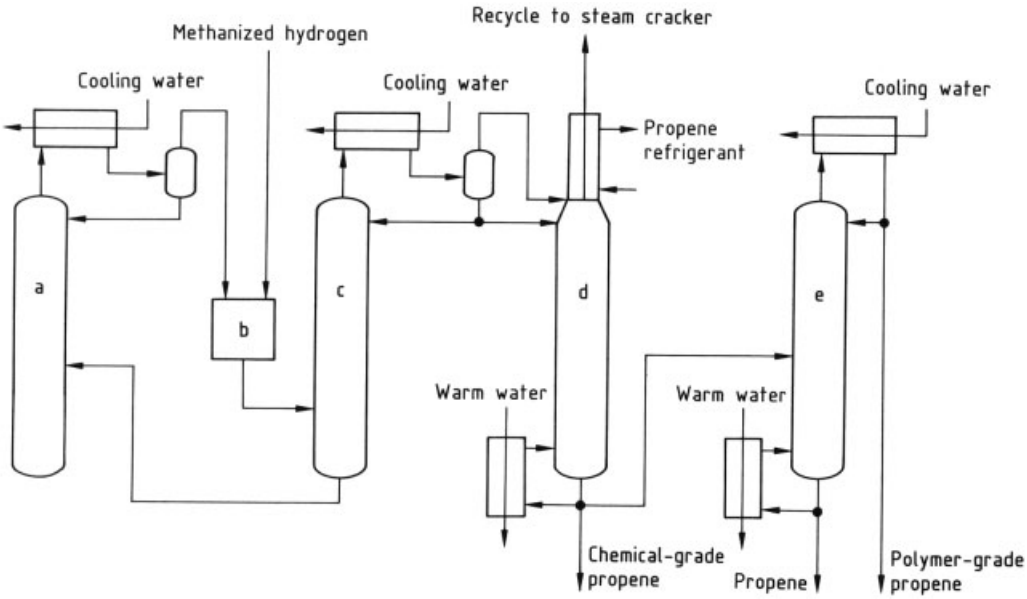

Figure 1. Typical flow diagram for work-up of the C_3 fraction in a steam cracker plant
a) Depropanizer; b) C_3 hydrogenation; c) Polymer scrubbing; d) C_3 stripper; e) C_3 splitter

propene – propane separation column (e). A single water- cooled column designed for this purpose is illustrated in Figure 1. Such a process is only one of several variants that are discussed in more detail in Section 4.1.1. At this point, reference may be made only to one advantage, which provides for the (not essential) integration of propene – propane separation in the steam cracker plant. In modern plants the cracked gases are usually cooled to ambient temperature in a water scrubbing tower after the oil fractionation. The circulating water is heated to 80 to 85 °C in this operation and must be cooled again to be reused in the water scrubbing tower. The heated circulating water can therefore be used to heat the propene – propane separation stage and thus save energy. Also, other heat consumers (e.g., the reboiler of the C_3 stripper) may be supplied economically from this energy source. The propane separated in the C_3 splitter is usually (like ethane) recycled as cracking feedstock.

4.1.1. Variants of Propene – Propane Separation

The closely similar boiling curves of propene and propane (*bp* of propane −42.1 °C) require highly complex separation units. The internal reflux ratio is generally between 0.90 and 0.97 (depending on the feedstock composition, product purities, column pressure, and number of trays). For this reason the propene – propane separation is a process unit in which the development of energy-saving solutions is worthwhile.

Single-Column Process. In principle, separation can be carried out in a single column (generally containing 150 – 200 trays), as illustrated in Figure 1 (e). The reflux can be condensed with cooling water (column pressure 16 – 19 bar) or in air coolers (column pressure 21 – 26 bar). However, for the throughputs that are common nowadays, huge column diameters are required, and transportation of these huge units (or parts of them) by rail or road is difficult. Thus the double-column process has been developed where the separation is carried out in two parallel columns that are smaller in diameter and can therefore be transported more easily. This double- column process requires only ca. 55 % of the original amount of cooling water.

Double-Column Process. The double-column process is illustrated in Figure 2. In this process, only the reflux from the second

separation column (b) is condensed with cooling water. The pressure of the first column (a) is so high (ca. 25 bar) that its overhead vapors (ca. 59 °C) can be liquefied in the reboiler of the second column (ca. 18 bar, 51 °C) and serve as heat medium. Each column produces ca. 50 % of the propene product. The bottom product from the first column is the feedstock for the second column. Heating the first column with warm water is still possible with this process.

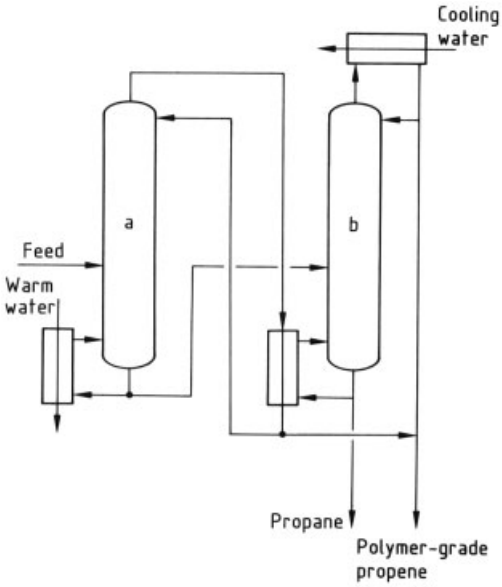

Figure 2. Propene – propane separation by the double-column process
a) High-pressure C_3 splitter (ca. 25 bar); b) Low-pressure C_3 splitter (ca. 18 bar)

Heat Pump Process. If no cost-free (or even cost-saving) heating medium such as the warm circulating water of the steam cracker is available, then separation according to the heat pump process is possible (see Fig. 3). In this process the higher propene – propane ratio in the gas phase at lower pressure is utilized. The overhead vapors of the C_3 splitter (a), which operates at ca. 10 bar, are heated slightly in the reflux subcooler (b) and then compressed to a pressure that enables them to be liquefied in the reboiler (e). Apart from the energy needed to drive the compressor (c), only a small amount of cooling water is required in the aftercooler (d).

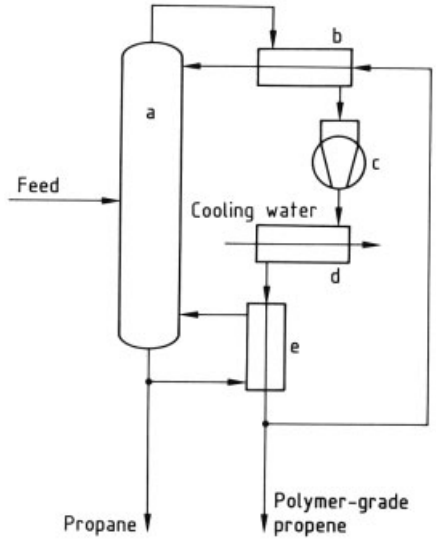

Figure 3. Propene – propane separation by the heat pump process
a) C_3 splitter; b) Reflux subcooler; c) Compressor; d) Postcooler; e) Condenser and reboiler

Lower Product Quality. Energy savings in the propene – propane separation can be achieved not only by appropriate process design, but also by restricting the purity of the polymer-grade propene to the absolute minimum necessary. Figure 4 shows the relative complexity of separation required for higher purities, relative to a minimum propene purity of 99.0 mol %. The complexity of separation and thus the energy consumption increases considerably with propene purity, and even more sharply as the number of column trays decreases.

4.1.2. Apparatus and Construction Materials

In steam cracker plants, work-up of the C_3 fraction occurs above 0 °C and below 100 °C (with the exception of gas-phase hydrogenation, where the temperature can rise to ca. 200 °C). The C_3-treatment section of the plant thus does not place any special requirements on materials. Normal carbon steels and fine-grained steels (such as WSt 36) are used; aluminum is employed for special heat exchangers (such as propene-heated reboilers of propene – propane splitters).

In general, turbocompressors are used as compressors for the *heat pump variant* of the propene – propane separation. Their power requirements for a propene capacity of 100 000 t/a and a product purity of 99.5 mol % are 1400 – 1500 kW. The compressors are driven by electric motors or steam turbines, which can be integrated into the steam system of the steam cracker.

Highly chargeable sieve trays spaced relatively close together are incorporated in the generally 70 – 90-m-high C_3 splitters. The apparatus used in steam cracker plants is discussed in more detail in → Ethylene, Chap. 5.3.2.

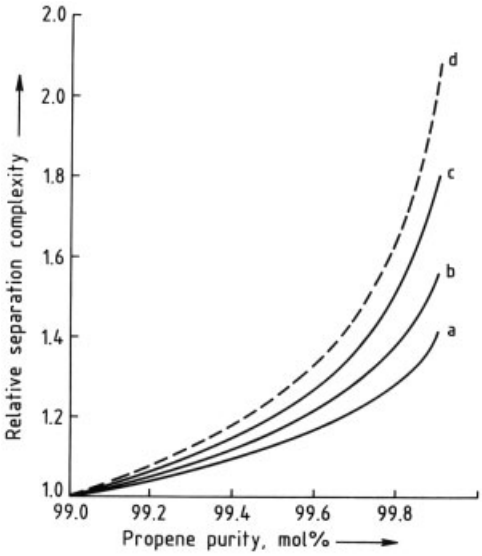

Figure 4. Relative complexity of the propene – propane separation as a function of enhanced propene purity (relative to a minimum propene purity of 99 mol %)
Base: Feedstock containing 93 mol % propene, 7 mol % propane; bottom product containing 30 mol % propene, 70 mol % propane
—- single column with 200 trays
– – – single column with 150 trays
a) Separation at 10 bar (heat pump process); b, d) Separation at 18 bar (reflux condensation with cooling water); c) Separation at 25 bar (reflux condensation with air cooler)

4.2. Production as Byproduct of Refinery Processes

The propene produced in refineries also originates from cracking processes. However, these processes can be compared to only a limited extent with the steam cracker for ethylene production because they use completely different feedstocks and have different production objectives.

Refinery cracking processes operate either purely thermally or thermally – catalytically. By far the most important process for propene production is the *fluid-catalytic cracking* (FCC) process in which the powdery catalyst flows as a fluidized bed through the reaction and regeneration phases (→ Oil Refining, Chap. 3.2.). This process converts heavy gas oil preferentially into gasoline and light gas oil.

Purely thermal cracking processes, which in refineries contribute to propene production, are employed in coking and visbreaker units (→ Oil Refining, Chap. 3.6.4.1., → Petroleum Coke, Chap. 3.). In *coking units* (*delayed coking* and *fluid coking*), residues from the atmospheric and vacuum distillation of the crude oil undergo relatively severe cracking and are thereby converted into gas oil, coke, gasoline, and smaller amounts of cracked gas (6 – 12 wt % of C_4 and lighter). The cracked gas from the coking unit normally contains 10 – 15 mol % C_3, mostly propane. In *visbreaker units*, vacuum residues are subjected to mild cracking, with the object of reducing the viscosity of the residue oil. Smaller amounts of gas oil, gasoline, and cracked gas (2 – 3 wt % of C_4 and lighter) are formed here.

The working-up of refinery gases to isolate the C_3 fraction generally occurs as follows: first, the light components are separated at ca. 15 bar in a deethanizer, whose upper part is operated as an absorber with gasoline and oil as scrubbing agents. The C_5 and heavier hydrocarbons are removed from the bottom product in a connected debutanizer. The C_3–C_4 fraction obtained as overhead product is normally desulfurized on molecular sieves and dried, before being split into individual fractions in a final depropanizer.

4.3. Propane Dehydrogenation

The principal sources of propene are the steam cracking of hydrocarbon feedstocks (Section 4.1) and refinery conversion processes [e.g., fluid catalytic cracking, visbreaking, and coking (Section 4.2)]. Despite the magnitude of the sources, in these cases propene is a byproduct of processes for the manufacture of other products, such as ethylene in the case of steam cracking and motor gasoline in the case of catalytic cracking. The availability of propene is

determined primarily by the demand for the main products, although factors such as feedstock and operating conditions have a significant influence on propene yield. The increasing demand for propene derivatives throughout the 1980s, especially for polypropylene, outstripped the availability from these established sources, and processes for the "on-purpose" production of propene by the dehydrogenation of propane from natural LPG fields have been developed commercially.

Propane dehydrogenation is an endothermic equilibrium reaction that is generally carried out in the presence of a noble- or heavy-metal catalyst such as platinum or chromium.

$$C_3H_8 \longrightarrow C_3H_6 + H_2$$

The process is highly selective; overall yields of propene from propane of ca. 90 % are claimed for commercially available processes. Higher temperature and lower pressure increase propene yield. However, increased process temperature also causes pyrolysis (cracking) of propane to coke in addition to its dehydrogenation to propene (i.e., reduced selectivity), whereas lower operating pressure increases selectivity. Coke formation wastes feedstock and deactivates the dehydrogenation catalyst. Consequently, propane dehydrogenation processes are operated near atmospheric pressure at around 500 – 700 °C.

A number of technologies are available commercially for the dehydrogenation of propane to propene: these include *Oleflex* developed by UOP, Des Plaines, Illinois, United States; *Catofin* developed by Air Products and Chemicals, Allentown, Pennsylvania, United States; and *STAR* developed by Phillips Petroleum, Bartlesville, Oklahoma, United States. These processes differ in their modes of operation, the dehydrogenation catalyst, and the methods of catalyst regeneration.

4.3.1. Oleflex Process

The Oleflex process was developed from the *Pacol* process (UOP), which is used to dehydrogenate $C_{10} - C_{14}$ paraffins to olefin feedstocks for the production of intermediates for synthetic detergents [9] (→ Hydrocarbons). The Oleflex process is an adiabatic process in which the heat of reaction is supplied by reheating the process stream between the different reaction stages. The process is operated at a slight positive pressure, and a proprietary platinum catalyst is used. Fresh propane feed is mixed with recycled hydrogen and unconverted propane and admitted to a train of three radial-flow moving-bed catalytic reaction vessels (Fig. 5). The process is continuous, and overall selectivities for propene of 89 – 91 % are claimed. The dehydrogenation catalyst circulates through the reactor section before passing to a separate regeneration vessel where coke is removed from the surface of the catalyst by combustion in air (UOP Continuous Catalyst Regeneration technology). The regenerated catalyst is returned to the first of the dehydrogenation reactors. Propene is recovered by conventional deethanizer – depropanizer splitting (see Section 4.1). Advantages claimed by the Oleflex process include continuous operation; a uniform, time-invariant catalyst activity profile; and isolation of the oxidative catalyst regeneration phase from the dehydrogenation reactor [9].

4.3.2. Catofin Process

The Houdry (Air Products) Catofin process (see Fig. 6) operates under a slight vacuum at 550 – 750 °C [10]. The process utilizes adiabatic, fixed-bed multiple reactors. The dehydrogenation catalyst consists of activated alumina pellets impregnated with 18 – 20 wt % chromium. The process is cyclic and includes a reaction period, discharge of the reactor, and regeneration of the catalyst in situ. Multiple reactors are used in parallel to achieve continuous plant throughput, with the reactor containing the coked catalyst being taken off stream for the regeneration step. The mixed fresh and recycle propane stream is preheated to 600 – 700 °C and fed to the reactor at ca. 30 kPa (one-third atmospheric pressure): the influence of reactor pressure and temperature on propene yield is shown in Figure 7. Combustion of the deposited coke on the catalyst during regeneration heats the catalyst bed, and this energy is released to the endothermic dehydrogenation reaction when that reactor is returned to operation. The overall selectivity of propane to propene is reported to be about 87 % [10]. Propene is recovered by a con-

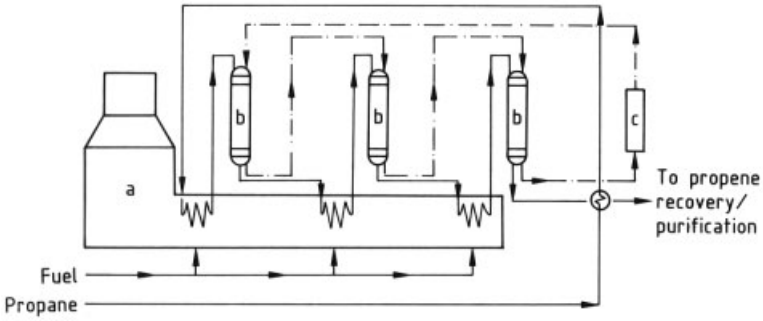

Figure 5. UOP Oleflex process
a) Reheat furnace; b) Moving-bed reactors; c) Continuous catalyst separation system

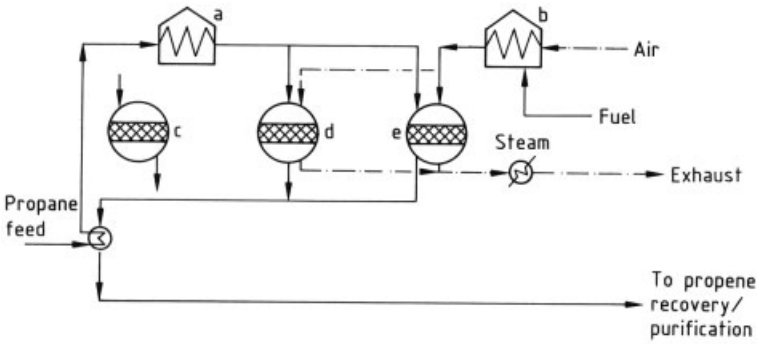

Figure 6. Houdry (Air Products) Catofin process
a) Charge heater; b) Air heater; c) Reactor on purge; d) Reactor on stream; e) Reactor on regeneration

ventional propane –propene splitter. Advantages claimed for the Catofin process include extended catalyst life and utilization of the heat generated in the exothermic regeneration step to assist dehydrogenation.

4.3.3. Phillips STAR Process

The Phillips Steam Active Reforming (STAR) process for the dehydrogenation of paraffins ($\leq C_5$) [11] differs significantly from other dehydrogenation processes. Here, steam is used as a diluent to maintain an overall positive pressure in the process reactor and simultaneously reduce the partial pressure of the hydrocarbons and hydrogen present. Thus, the equilibrium is shifted toward increased conversion. In addition, the process is performed isothermally. The preheated feedstream containing steam is admitted to a chain of multiple fixed-bed reactors. Each reactor consists of multiple catalyst-packed tubes in a furnace firebox, which supplies heat to the catalyst. Reactor operation is cyclic (i.e., one reactor is taken offline sequentially for catalyst regeneration), whereas dehydrogenation is maintained continuously. Catalyst deactivation occurs due to coke deposition, and after about 7 h on-line, an off-line catalyst regeneration by combustion for ca. 1 h is required. The catalyst is completely recovered, and overall catalyst lifetimes of one to two years have been reported [11]. Advantages claimed for the STAR process include 80 % yield of propene on propane; isothermal operation that ensures sufficient heat input to promote dehydrogenation; and the steam diluent that reduces the hydrocarbon pressures while maintaining a more practical process pressure. Carbon dioxide (formed by side reactions) must be removed from the reactor product stream prior to propene separation.

4.3.4. Linde Process

The Linde process [12] operates at low reaction temperature and nearly isothermal condi-

tions which are designed to minimize thermal cracking and coke formation. The process uses a chromium oxide catalyst in a fixed-bed tubular reactor. The catalyst has a relatively long cycle time (9 h) before regeneration is necessary. The propane feedstream is not diluted with hydrogen or steam as in other processes, resulting in high selectivity (e.g., 91 %). The process is claimed to have low power requirements and low capital costs. Product separation yields polymer-grade propene.

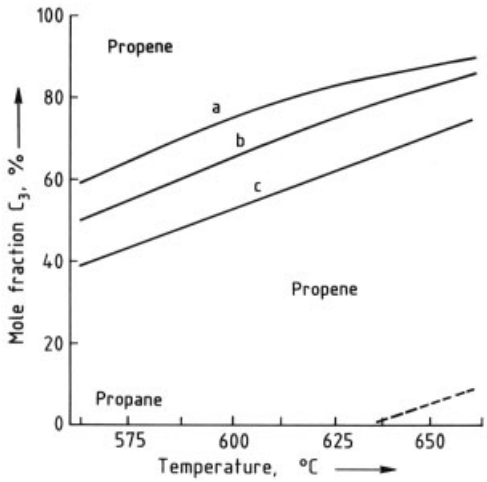

Figure 7. Influence of reactor pressure and temperature on propene yield by propane dehydrogenation
a) 30 kPa (one-third atmospheric pressure); b) 50 kPa (half atmospheric pressure); c) 101.3 kPa (atmospheric pressure)
– – – mole fraction of methyl acetylene

5. Storage, Transportation, Quality Requirements

There is an extensive propene pipeline network in Texas and Louisiana in the United States and a very limited system in Belgium in Western Europe. In other locations propene transportation is largely by road, rail, and ship. This discontinuous delivery and supplying of consumers requires large storage capacities on the part of both propene producers and propene consumers.

Liquid propene is normally *stored* at ambient temperature in spherical pressure tanks with diameters up to 20 m. Propene is however also stored virtually pressureless at −47 °C, particularly for very large amounts (→ Liquefied Petroleum Gas, Chap. 5.). The latter form of storage requires reliquefaction devices.

Propene is *transported* by road and rail in cylindrical pressurized tanks at ambient temperature. The standard railway tanker holds 42 t of propene. In the case of road transport the maximum permitted overall truck weight of ca. 40 t restricts the propene load to ca. 20 t (corresponding to 40 m^3). In transportation by ship, batteries of smaller pressurized tanks as well as large atmospheric tanks with reliquefaction devices are used.

Quality Requirements. The quality requirements placed on propene vary widely depending on its subsequent use. Corresponding to production in steam cracker plants, two qualities are basically employed namely, the less pure chemical-grade propene and the highly pure polymer-grade propene. Examples of specifications for these two product qualities are given in Table 2. Depending on different posttreatment processes (in the case of polymer-grade propene) or different intended main uses (in the case of chemical-grade propene) the permissible concentrations of impurities may vary within relatively wide limits.

6. Uses

Global propene production for chemical uses increased strongly through the 1980s [13], equivalent to a global average annual increase of 5 % (see Table 3).

The extent of steam cracker and refinery production of propene varies from region to region [13]. In the United States, refinery production has a much greater share of total availability than in other regions due to a large motor gasoline market that requires an adequate FCC capacity and a preponderance of ethane feedstock for steam crackers, resulting in lower yields of steam cracker propene. The proportion of propene produced by steam cracking between 1980 and 1990 is given in Table 4.

The three commercial grades of propene are used for different applications. *Refinery-grade* propene (i.e., 50 – 70 % pure propene in propane) is obtained from refinery processes. The main uses of refinery propene are in LPG for

Table 2. Typical product specification of propene

	Chemical grade	Polymer grade	Detection method ∗	Detection limit, ppm
Propene, mol %	92 – 95	99.5 – 99.8		
Acetylene, mol ppm	< 10	< 2	GC	0.5
Ethylene, mol ppm	< 20	< 20	GC	0.5
Ethane, mol ppm	< 2000	< 100	GC	0.5
Propyne, mol ppm	< 20	< 5	GC	1
Propadiene, mol ppm	< 20	< 5	GC	1
C_{4+}, mol ppm	< 1000	< 10	GC	2
Hydrogen, mol ppm	< 10	< 10	GC	2
Nitrogen, mol ppm	< 50	< 50	GC	2
Oxygen, mol ppm	< 5	< 5	colorimetry	1
Carbon monoxide, mol ppm	< 5	< 5	GC (via methanation)	0.5
Carbon dioxide, mol ppm	< 5	< 5	GC (via methanation)	1
Sulfur, mass ppm	< 5	< 1	according to Wickbold (DIN EN 41)	1
Water, mol ppm	< 25	< 10	Karl Fischer	1
Propane	remainder	remainder	GC	1

∗ GC = gas chromatography.

Table 3. Propene production for chemicals in 10^3 t

Region	1980	1982	1984	1986	1988	1990
North America	6 797	6 378	7 714	8 350	9 577	10 003
South America	660	732	982	1 094	1 182	1 466
Western Europe	5 823	5 984	7 191	7 333	8 702	9 142
Eastern Europe	1 686	2 068	2 272	2 557	2 571	2 706
Japan	2 637	2 565	2 981	3 147	3 682	4 214
Asia – Pacific	920	1 014	1 369	1 654	2 035	2 399
Others	76	72	55	86	271	390
Total	18 599	18 813	22 564	24 165	28 020	30 320

Table 4. Proportion of propene produced by steam cracking (in wt %)

Region	1980	1982	1984	1986	1988	1990
North America	59	46	54	58	58	56
South America	64	65	72	73	72	68
Western Europe	92	87	86	84	83	78
Eastern Europe	86	87	87	85	88	87
Japan	95	88	85	79	83	82
Asia – Pacific	88	86	89	88	89	88

thermal use or as an octane-enhancing component in motor gasoline. Refinery-grade propene can also be used in some chemical syntheses (e.g., of cumene or isopropanol).

Chemical-grade propene is used extensively for most chemical derivatives (e.g., oxo alcohols, acrylonitrile, or polypropylene).

*Polymer-grade*propene contains minimal levels of impurities such as carbonyl sulfide that can poison catalysts used in polypropylene and propylene oxide manufacture.

6.1. Thermal Uses

Propene has a calorific value of 45 813 kJ/kg, and refinery grade propene can be used as fuel if more valuable uses are unavailable locally (i.e., propane – propene splitting to chemical-grade purity). Refinery-grade propene can also be blended into LPG for commercial sale, provided no low limits exist on the olefin content of LPG (unsaturated hydrocarbons produce more smoke than saturated hydrocarbons and have a lower calorific value).

6.2. Motor Gasoline Uses

Propene is used as a motor gasoline component for octane enhancement, via dimerization – formation of polygasoline or alkylation.

6.2.1. Polygasoline Production-Dimerization

Polymerization of refinery-grade propene at high temperature and pressure with a catalyst such as phosphoric acid produces *polygasoline* containing dimers, trimers, and tetramers. Polygasoline is used as a blendstock in motor gasoline. In 1977, *dimerization* of propene to isohexenes at lower temperature and pressure by using nickel complex – alkylaluminum halide catalysts was developed by Institut Français de Petrole (Dimersol process) [14]. Propene conversions of $> 90\,\%$ are achieved. The product, *dimate*, has a low vapor pressure and a high blending octane rating (R + M)/2 of 89 (for a definition of blending octane rating, see → Octane Enhancers, Chap. 5.1.), compared to 86 for cat cracker gasoline and 87 for polygasoline.

6.2.2. Alkylation

Alkylation is an acid-catalyzed reaction between propene (or other light olefins) and isobutane to yield a highly branched higher alkane (2,3-dimethylpentane in the case of propene) of low vapor pressure and high octane rating [15]. The reaction catalyst is either hydrofluoric acid, which gives octane ratings of 90 – 92, or sulfuric acid, which gives octane ratings of 88 – 91 [16]. If *hydrofluoric acid* is used as catalyst a second reaction is promoted in which hydrogen is transferred from isobutane to propene to yield propane and isobutene. The latter then alkylates isobutane to produce trimethylpentanes in parallel with dimethylpentane from the direct alkylation. The choice of catalyst depends on feedstock and process conditions as well as octane rating.

6.3. Chemical Uses

The principal chemical uses of propene are in the manufacture of polypropylene, acrylonitrile, oxo alcohols, propylene oxide, butanal, cumene, and propene oligomers [13]. Polypropylene is the leading global consumer of propene (see Table 5). Other uses include acrylic acid derivatives and ethylene – propene rubbers.

Polypropylene (→ Polyolefins, Chap. 2.). Isotactic polypropylene is of great commercial importance. Polymerization of chemical-grade or polymer-grade propene is carried out by using Ziegler – Natta catalysts [17], in either a liquid-phase or a gas-phase process. Polypropylene has a high melting point, good rigidity, and good chemical and abrasion resistance. Weaknesses include poor thermal and UV stability, but these can be improved by appropriate stabilizers. Polypropylene is used widely, and global production exceeded 13×10^6 t in 1990.

Acrylonitrile. Acrylonitrile is obtained by catalytic oxidation of chemical-grade propene in the presence of ammonia (ammoxidation).

A number of commercial processes are available but the most significant is the Sohio process (→ Acrylonitrile) [18] in which a stoichiometric mixture of propene, ammonia, and oxygen is reacted over a catalyst at 400 – 500 °C. World consumption of acrylonitrile in 1990 was 3.7×10^6 t.

Propylene Oxide. Propene (polymer grade) can be converted into the ether propylene oxide either by hydrochlorination and epoxidation, or by reaction with an organic hydroperoxide [20], [21] (→ Propylene Oxide).

Propylene oxide is used as a raw material in the production of poly(propylene glycols), propanolamines, and polyether polyols, which themselves are intermediates for flexible and rigid polyurethane foams. Propylene oxide production in 1990 was 2.75×10^6 t, consuming some 2.2×10^6 t of propene.

2-Propanol. 2-Propanol is produced by the hydration of chemical-grade propene either in the presence of sulfuric acid or directly as a vapor-phase reaction in the presence of a catalyst at 180 – 270 °C [22]. Yields $> 70\,\%$ based on propene are achieved at high selectivity (e.g., $> 90\,\%$). 2-Propanol is used extensively as a solvent in paint, cosmetics, and pharmaceuticals. Global production of 2-propanol in 1990 was 1.4×10^6 t, representing 1.2×10^6 t of propene.

Table 5. Consumption of propene (in wt %)

Product	1980	1982	1984	1986	1988	1990	1995 *
Polypropylene	30	33	36	38	40	44	50
Acrylonitrile	19	18	18	17	15	14	12
Oxo alcohols	19	18	16	16	16	15	4
Propylene oxide	9	9	8	9	9	9	8
Cumene	10	9	9	8	8	8	7
Oligomers	8	7	6	5	5	4	3
Others	5	6	7	7	7	6	7

[19]

Cumene. Propene can be used to alkylate benzene to produce 1-methylethylbenzene (cumene) [23]. The reaction is usually undertaken with refinery-grade propene in the presence of phosphoric acid catalyst. Cumene is then oxidized to cumene hydroperoxide, which decomposes to phenol and acetone (→ Phenol, Chap. 4.1.). Global production of cumene was 2.3×10^6 t in 1990, representing 0.9×10^6 t of propene.

Oligomers. Propene oligomers are liquid polymers of propene with a low degree of polymerization. They are intermediates between polygasoline and the high polymer polypropylenes, and contain ten or more carbon atoms per molecule. Propene oligomers are prepared by acid-catalyzed polymerization. Depending on the degree of polymerization the products are used as lube oil additives or as detergent intermediates.

Hydroformylation. Propene reacts with carbon monoxide and hydrogen (synthesis gas) at elevated temperature and pressure in the presence of a transition-metal catalyst (e.g., a rhodium carbonyl catalyst) to give butanals (oxo process) [24]. *n*-Butanal is converted into the aldol and hydrogenated to produce 2-ethylhexanol.

Allylics. Propene can undergo many reactions to produce allylic compounds. In these reactions a methyl hydrogen atom is substituted and the carbon–carbon double bond is preserved (see ammoxidation page 12). The more important allylic derivatives include allyl chloride (3-chloropropene) [25] and acrolein (allyl aldehyde) [22]. Propene can be reacted with *acetic acid* in an acidic medium to produce allyl acetate (i.e., the double bond is retained in the molecule), or acetic acid is added to the carbon–carbon double bond to yield isopropyl acetate.

7. Economic Aspects

Propene is a byproduct of the production of ethylene by steam cracking (Section 4.1) or the production of motor gasoline by catalytic cracking (Section 4.2), so no direct manufacturing costs are attributed to these sources. Propene demand exceeds that from steam cracker sources, and marginal supplies are taken from refinery sources. *Propene economics* are based on its value in marginal applications: generally dimerization (Section 6.2.1) or alkylation (Section 6.2.2) for gasoline. The motor gasoline valuation of propene is set by the octane enhancement of motor gasoline by alkylate or Dimersol, both obtained from refinery-grade propene. Generally, motor gasoline absorbs incremental propene supplies, and all other uses must pay the equivalent octane value plus the costs of distilling refinery grade to the purity required. When propene demand in motor gasoline is weak, surplus refinery propene is returned to LPG uses or to refinery fuel (Section 6.1). When demand for propene is very high the value of the marginal supply rises above the octane value to reflect the new marginal use (e.g., in polypropylene).

Propene from the *dehydrogenation of propane* does have a direct manufacturing cost structure based on feedstock, utility, plant fixed costs, etc. In an environment in which propene demand exceeds the supply from steam crackers and catalytic crackers, propene values should cycle between a minimum of the cash cost of propane dehydrogenation and either a level necessary to justify the investment in further dehydrogenation plants or a higher level set by the demand for derivatives.

8. References

General References

1. F. Andreas, K. Gröbe: *Propylenchemie,* Akademie-Verlag, Berlin 1969.
2. F. Asinger: *Die Petrolchemische Industrie,* parts I and II, Akademie-Verlag, Berlin 1971.
3. L. F. Albright, B. L. Crynes: *Industrial and Laboratory Pyrolyses,* American Chemical Society, Washington 1976.

Specific References

4. C. L. Yaws, *Chem. Eng. (Int. Ed.)* **82** (1975) no. 7, 101 – 109; **83** (1976) no. 17, 79 – 87; **83** (1976) no. 23, 127 – 135; **83** (1976) no. 25, 153 – 162.
5. M. Stam: *Western European Petrochemicals into the 1980's* (unpublished, Euroeconomics, July 31st 1978).
6. E. J. Debreczeni, *Chem. Eng. (Int. Ed.)* **84** (1977) no. 12, 135 – 141.
7. D. S. Sanders, D. M. Allen, W. T. Sappenfield, *Chem. Eng. Prog.* **73** (1977) no. 7, 40 – 45.
8. P. H. Spitz, *Hydrocarbon Process.* **55** (1976) no. 7, 131 – 135.
9. P. R. Pujado, B. V. Vora, *Hydrocarbon Process.* **69** (1990) no. 3, 65.
10. R. G. Craig, T. J. Delaney, J. M. Duffalo: *Catalytic Dehydrogenation Performance of the Catofin Process,* DeWitt Petrochemical Review, Houston 1990.
11. R. O. Dunn et al.: *The Phillips Steam Active Reforming (STAR)* Process of C_3, C_4, and C_5 Paraffins for the Dehydrogenation, DeWitt Petrochemical Review, Houston 1992.
12. *Eur. Chem. News* (1992) March 9, 19.
13. World Petrochemical Industry Survey: *Olefins,* Parpinelli Tecnon srl, Milan 1991.
14. P. M. Kohn, *Chem. Eng.* **114** (1977) May 23.
15. L. F. Albright, *Oil Gas J.* (1990) Nov., 79.
16. L. E. Chapin, G. C. Lilios, T. M. Robertson, *Hydrocarbon Process.* **64** (1985) no. 9, 67.
17. G. Natta et al., *J. Am. Chem. Soc.* **77** (1955) 1708.
18. J. L. Callahan, R. K. Grasselli, E. C. Milberger, H. A. Strecker, *Ind. Eng. Chem. Prod. Res. Dev.* **9** (1970) 134.
19. K. Weissermel, H.-J. Arpo: *Industrielle Organische Chemie,* Wiley-VCH, Weinheim 1998. p. 291.
20. R. B. Stobaugh et al., *Hydrocarbon Process.* **52** (1973) no. 1, 102.
21. J. Poloczek, J. Bobinski, *Int. Chem. Eng.* **11** (1971) no. 1, 87.
22. E. Wilhelan, R. Battino, R. T. Wilcock, *Chem. Rev.* **77** (1977) no. 2.
23. P. R. Pujado, J. R. Salazar, C. V. Berger, *Hydrocarbon Process.* **55** (1976) no. 3, 91.
24. P. Pino, F. Piacenti, M. Bianchi: *Organic Synthesis via Metal Carbonyls,* **vol. 2,** J. Wiley and Sons, Inc., New York 1977, p. 43.
25. A. W. Fairbairn, H. A. Cheney, A. J. Chemiavsky, *Chem. Eng. Prog.* **43** (1947) 280.

Propiolactone → Hydroxycarboxylic Acids, Aliphatic
Propiolic Acid → Carboxylic Acids, Aliphatic
Propionaldehyde → Propanal

Propionic Acid and Derivatives

Ulf-Rainer Samel, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 7, 9, Sections 8.2, 11.1 – 11.3)

Walter Kohler, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 8.1)

Armin Otto Gamer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 10)

Ullrich Keuser, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 11.4)

1. Introduction 2
2. Physical Properties 2
3. Chemical Properties 3
4. Production 4
4.1. Carbonylation of Ethylene (BASF Process) 4
4.2. Oxidation of Propanal 5
4.3. Direct Oxidation of Hydrocarbons 5
4.4. Construction Materials 6
5. Storage and Transportation . . . 6
6. Quality Specifications 7
7. Environmental Protection 7
8. Uses 8
8.1. Use as Food and Feed Preservative 8
8.2. Other Uses 10
9. Economic Aspects 10
10. Toxicology 10
11. Derivatives 11
11.1. Salts 11
11.2. Propionic Anhydride and Propionyl Chloride 12
11.3. Esters 12
11.4. Chloropropionic Acid and Derivatives 12
11.4.1. Properties and Production 13
11.4.1.1. 2-Chloropropionic Acid and Derivatives 13
11.4.1.2. 3-Chloropropionic Acid and Derivatives 14
11.4.1.3. 2,2-Dichloropropionic Acid 14
11.4.2. Uses 14
11.4.2.1. 2-Chloropropionic Acid and Derivatives 14
11.4.2.2. 3-Chloropropionic Acid 15
11.4.2.3. 2,2-Dichloropropionic Acid 15
11.4.3. Toxicology 15
12. References 16

1. Introduction

Propionic acid [*79-09-4*], CH_3CH_2COOH, M_r 74.08, was first described by J. Gottlieb in 1844. In 1848 J. J. Dumas recognized that it belonged to the group of compounds known as fatty acids. Since it is the first fatty acid which can be salted out of aqueous solution, it was named propionic acid ($\pi\varrho\omega\tau o\sigma$ = first, $\pi\acute{\imath}\omega\nu$ = fat).

Propionic acid occurs naturally, predominantly in the form of its esters in some essential oils. It is also formed as the free acid in various enzymatic and fermentation processes, and is produced as a result of anaerobic carbohydrate fermentation in the stomachs of ruminants. Propionic acid derivatives are also important intermediates in various physiological cycles.

Industrially, propionic acid is used in the production of cellulose esters, plastic dispersions, and herbicides. It is also used to a limited extent in pharmaceuticals and in flavors and fragrances. Propionic acid is gaining increasing importance for the preservation of forage cereals and animal feeds because many putrefying and mold-forming microorganisms cannot survive in its presence. This fungicidal and bactericidal action is exploited to protect bakery products and cheese from attack by bacteria and mold.

Propionic acid is produced specifically by carbonylation of ethylene and oxidation of propionaldehyde; in addition, it is obtained in large quantities as a byproduct in the oxidation of hydrocarbons.

2. Physical Properties

Propionic acid, *mp* −20.8 °C, *bp* 141.3 °C (at 101.3 kPa) is a clear, colorless liquid with a

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_223

Table 1. Vapor pressure of pure propionic acid [16]

t, °C	p, kPa	t, °C	p, kPa
0.7	0.1	100.9	25
13.5	0.25	120.1	50
24.0	0.5	140.9	100
35.0	1.0	159.5	200
50.8	2.5	185.7	500
64.5	5.0	203.0	1000
79.3	10.0	219.7	2000

Table 2. Change in density of liquid propionic acid with temperature [17]

t, °C	ϱ, g/cm^3	t, °C	ϱ, g/cm^3
−20	1.038	80	0.9286
−10	1.027	90	0.9176
0	1.016	100	0.9066
10	1.005	110	0.8954
20	0.9935	120	0.8841
30	0.9827	130	0.8726
40	0.9719	140	0.8609
50	0.9611	150	0.8490
60	0.9503	160	0.8367
70	0.9395	170	0.8241

slightly pungent odor (odor threshold in air 0.03 vol %). Solid propionic acid forms monoclinic crystals and undergoes an increase in volume on melting of 12.2 % [1]. In the liquid state and in the gas phase a considerable portion of the molecules of propionic acid are dimerized as is the case with its lighter homologues acetic [*64-19-7*] and formic [*64-18-6*] acids [2–5]. Propionic acid is miscible with water and most organic solvents in all proportions. Propionic acid – water mixtures show a positive deviation from Raoult's law [3]. The heats of mixing of various binary and ternary mixtures of propionic acid are given in [6], [7].

Propionic acid forms azeotropic mixtures with many liquids [8–11], for example with water at 82.3 wt % water [*bp* (101.3 kPa) 99.98 °C]. The liquid – vapor equilibria of the system propionic acid – water at various pressures have been investigated [12]. Further data on the liquid – vapor system propionic acid – water and also on other binary mixtures of propionic acid are given in [13]. The solubility of propionic acid in ternary mixtures is described in [14], [15].

The temperature-dependence of the vapor pressure of pure propionic acid is shown in Table 1 and that of the density of the pure acid in Table 2. Table 3 shows the effect of temperature on the dynamic viscosity.

Table 3. Dynamic viscosity of pure propionic acid as a function of temperature [16]

t, °C	η, mPa · s
0	154.1
20	109.4
50	73.8
100	45.0
150	31.0

Other physical data are as follows:

Refractive index n_D^{20}	1.3865 [18]
Surface tension toward air	
(at 20 °C)	26.7×10^{-3} N/m [19]
(at 90 °C)	19.7×10^{-3} N/m
Thermal conductivity	
(at 0 °C)	153 W m^{-1} K^{-1} [20]
(at 20 °C)	150 W m^{-1} K^{-1}
(at 50 °C)	144 W m^{-1} K^{-1}
(at 100 °C)	136 W m^{-1} K^{-1}
Electrical conductivity	
(at 2 °C)	10^{-4} µS/cm
Dynamic dielectric constant	
(at 10 °C)	3.30 [9]
(at 40 °C)	3.44
Electric dipole moment, μ	1.75 D ± 5 % [9]
Dissociation constant in H_2O	
(at 0 °C)	1.274×10^{-5} mol/L [21]
(at 20 °C)	1.338×10^{-5} mol/L [22]
(at 40 °C)	1.284×10^{-5} mol/L
(at 60 °C)	1.160×10^{-5} mol/L
Heat of neutralization with NaOH	
(at 25 °C)	56.82 ± 0.08 kJ/mol [23]

Ionization potential	10.24 ± 0.03 eV [9]
Heat of fusion	101.65 kJ/kg [24]
Heat of vaporization	418.7 kJ/kg [25]
Heat of formation	− 164.5 kJ/mol [25]
Heat of combustion	− 1528.3 kJ/mol [9]
Specific heat (liquid), c_p	
(at 0 °C)	2.077 kJ kg^{-1} K^{-1} [16]
(at 50 °C)	2.299 kJ kg^{-1} K^{-1}
(at 100 °C)	2.516 kJ kg^{-1} K^{-1}
Specific heat (vapor), c_p	
(at 100 °C)	2.823 kJ kg^{-1} K^{-1}
(at 200 °C)	3.245 kJ kg^{-1} K^{-1}
(at 300 °C)	3.647 kJ kg^{-1} K^{-1}
Isentropic compressibility	
(at 20 °C)	0.750 GPa [26]
Critical data	
t_{crit}	338.9 °C [16]
p_{crit}	5.37 MPa [16]
V_{crit}	0.322 kg/L [16]
Flash point	50 °C [27]
Ignition temperature	485 °C
Explosion limits in air	
(at 20 °C)	2.1 – 12.0 vol %

The specific heat of aqueous propionic acid solutions has been measured by ACKERMANN [28]. The values increase with increasing temperature, but decrease with increasing propionic acid concentration.

A graphical presentation of the temperature dependence of vapor pressure, heat of vaporization, specific heat, density, viscosity, surface tension, and thermal conductivity of gaseous and liquid propionic acid is given in [29].

3. Chemical Properties

As the third member of the homologous series of aliphatic monocarboxylic acids, propionic acid is still quite reactive and undergoes all reactions typical of these acids. It readily forms salts, amides, acid halides, and an anhydride (see Chap. 11). Esters are formed with alcohols and olefins, and with acetylenes the corresponding vinyl esters (→ Vinyl Esters).

The reactions of propionic acid differ from those of acetic acid because of the inductive effect of the methylene group. Thus, propionic acid is attacked by alkaline permanganate solution more readily than acetic acid, whereby it is degraded nonselectively into carbon dioxide, oxalic acid, and lactic acid [30].

Ionic substitution reactions also preferably occur at the methylene group. Chlorination in the presence of halogen-transfer reagents (e.g., PCl_3) gives α-chloropropionic acid. Radical chlorination (presence of peroxides, UV radiation) yields β-chloropropionic acid [31], [32]. Both acids are important intermediates in the production of herbicides and pharmaceuticals (see Section 11.4).

Thermal decomposition of calcium propionate or passage of propionic acid vapor over metal oxides at 300 – 400 °C gives diethyl ketone, which is used in fairly large quantities as an industrial solvent.

4. Production

Industrially, propionic acid is currently produced almost exclusively by three different processes:

1) Carbonylation of ethylene with carbon monoxide and water
2) Oxidation of propanal
3) Direct oxidation of hydrocarbons

Processes such as production of propionic acid as a byproduct in the synthesis of hydroxylammonium salts from 1-nitropropane [*108-03-2*] [33] (→ Hydroxylamine, Chap. 4.4.), by wood distillation, or by nitric acid oxidation of 1-propanol [*62309-51-7*] have become obsolete, although the latter gives propionic acid in 90 % yield [34]. Synthesis of propionic acid via the alkali melt of an *n*- and isopropanol mixture could never compete successfully with other processes, although propionic acid was obtained with > 98 % yield [35], [36]. The carbonylation of ethanol [*64-17-5*] [37] and acetic acid [*64-19-7*] [38] have also not yet been carried out industrially.

The *Koch synthesis* has been investigated intensively [39], [40]. As a carbonylation reaction of ethylene [*74-85-1*] in a strongly acidic medium, it is a variant of the Reppe synthesis. However, compared with the latter it never achieved much industrial significance and was carried out with little success.

Other possible sources of propionic acid, which are not used industrially for economic reasons, are its formation as a byproduct in the high- and low-pressure carbonylation of methanol [*67-56-1*] to give acetic acid (2 % formation of propionic acid), the atmospheric oxidation of

n-butene [*25167-67-3*] (4 % formation of propionic acid), and the direct reaction of ethylene, carbon monoxide, and water over noble-metal catalysts [41–49].

Processes for making propionic acid accessible specifically by C_1-*chemistry* have yet to be assessed for their future importance. Two-step reactions have been described, in which propionic acid is obtained directly from synthesis gas (20 – 60 bar, 150 – 160 °C, Rh catalysts) [50], [51]. With the current availability of ethylene and naphtha these processes are, however, not yet competitive.

It is sometimes desirable, to use propionic acid produced by *natural methods*, particularly for the use of propionic acid in flavors and fragrances. Appropriate microbiological and enzymatic processes have been developed, which are usually based on the anaerobic fermentation of starch or sugars [52–54]. However, the expensive production of this "natural" propionic acid limits its use to special areas of application.

4.1. Carbonylation of Ethylene (BASF Process)

In the Reppe synthesis (see → Carbonylation, Chap. 4.1.) ethylene is reacted with carbon monoxide [*630-08-0*] and water in the presence of $Ni(CO)_4$ [*13463-39-3*] according to the following equation:

$$CH_2 = CH_2 + CO + H_2O \xrightarrow{Ni(CO)_4} CH_3CH_2COOH \qquad \Delta H = -164.5\ \text{kJ/mol}$$

The reaction takes place at high pressure and is characterized by low raw material costs, high conversion, high yield, and a simple workup.

Process Description (Fig. 1) [25]. Ethylene and carbon monoxide are compressed and continuously pumped into the high-pressure reactor (a) together with feed solution. The crude propionic acid formed at 100 – 300 bar and 250 – 320 °C is drawn off at the head of the reactor and cooled in a heat exchanger (b) with production of steam. Part of the cooled reaction product is recycled to the reactor for temperature regulation (c), the main quantity is allowed to expand and is separated into an off-gas and a crude acid stream (d). Nickel is recovered from the off-gas and led back into the reactor. The off-gas is incinerated with recovery of heat. The crude acid stream is subsequently dehydrated and worked up by distillation in several columns (e). The nickel salts thus formed are recycled into the process. The pure propionic acid is finally obtained by distillation. The product residue is channeled out of the process.

Variants of the Reppe and Koch syntheses are described in [40], the effect of inorganic salts on the reaction conditions and yield is discussed. Boric acid is reported to accelerate the reaction and hinder the precipitation of catalyst salts [55].

4.2. Oxidation of Propanal

The oxidation of propanal [*123-38-6*] is an important route to propionic acid. This route is economically attractive although two steps (production of propanal and subsequent oxidation) are required: (1) Propanal is formed in large quantities as an intermediate in the production of *n*-propanol [*62309-51-7*] by hydroformylation of ethylene, and (2) the oxidation takes place in plants in which other aliphatic carboxylic acids (*n*-butyric and isobutyric acids, *n*-valeric and isovaleric acids) are also produced. A high utilization of available capacity is therefore guaranteed in both steps.

Propanal Production. The production of propanal is carried out by the hydroformylation of ethylene (→ Propanal, Chap. 4.). It is favored because, unlike the case of the higher aldehydes, *n*/*iso*mixtures cannot be formed.

Two syntheses compete: the classical cobalt-catalyzed high-pressure carbonylation at 200 – 280 bar, and 130 – 150 °C [56], and the rhodium- or iridium-catalyzed low-pressure carbonylation at ca. 20 bar, and ca. 100 °C (→ Carbonylation, Chap. 4.→ Carbonylation, Chap. 4.1.) [57–59].

In the *high-pressure synthesis* the yield is generally impaired by partial hydrogenation of propanal to give propanol. The isolation of the aldehyde takes place, after removal of cobalt, by distillation of an azeotrope with ca. 98 % aldehyde content [56].

In the *low-pressure synthesis* (Union Carbide) [60], the aldehyde can be distilled directly from the reaction mixture in 99 % purity.

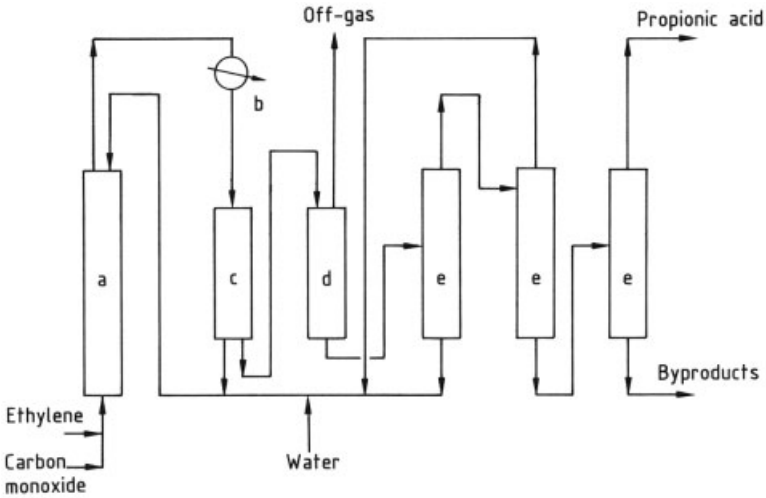

Figure 1. Production of propionic acid by the BASF process
a) High-pressure reactor; b) Heat exchanger; c) Separator; d) Expansion vessel; e) Distillation column

Oxidation. Propanal is subsequently oxidized under very mild conditions at 40 – 50 °C to propionic acid with high selectivity (→ Aldehydes, Aliphatic and Araliphatic).

In the United States propionic acid is produced by the oxidation of propanal by Union Carbide and Eastman Kodak (see Chap. 9).

4.3. Direct Oxidation of Hydrocarbons

A large quantity of propionic acid is obtained by the direct oxidation of hydrocarbons, predominantly naphtha. In this process, which is principally used for acetic acid production (→ Acetic Acid, Chap. 4.2.), formic acid [*64-18-6*], propionic acid, and an isomeric mixture of butyric acids are formed as byproducts. Whether this process is considered economic for propionic acid production is principally a question of the market evaluation of the different products. The process has favorable raw material costs, but requires a relatively complex workup of the product mixture [61], [62].

The composition of the acid mixture formed in the oxidation can be affected by the reaction conditions (pressure, temperature), the type of reactor (tubular reactor, column reactor), the raw material (naphtha, liquefied petroleum gas), and by the catalyst [61], [62]. If naphtha is used, the proportion of propionic acid in the mixture is ca. 10 – 15 %. In other cases, e.g., with 2-methylpentane as raw material, the proportion of propionic acid can increase to 31 % [63]. The total yield of acids does, however, vary considerably with the starting material [64–66].

BP Chemicals Process. (Fig. 2) [61], [67]. Naphtha is preheated to 170 °C and oxidized with air at 40 – 45 bar in several reactors (a) in series. The heat of reaction is used for steam generation. The cooled discharge from the reactors is separated from the reaction off-gas in a separator (b). Entrained liquid is recovered from the off-gas and recycled to the reactor; the off-gas is then incinerated. The liquid reactor discharge is separated into an organic phase, which contains unreacted hydrocarbons, and an aqueous phase which contains the product mixture (c). The unreacted hydrocarbons are recycled to the reactor. The low- and high-boilers are separated from the aqueous phase which then gives the crude acid (d). From this the C_1- to C_4-acids are obtained by extractive dehydration (e) followed by fractional distillation (g) – (i).

Propionic acid is produced by direct oxidation at BP Chemicals (UK), Hoechst – Celanese (United States), and Daicel (Japan) (see Chap. 9).

4.4. Construction Materials

The corrosiveness of propionic acid is not only determined by the water content, pressure, and

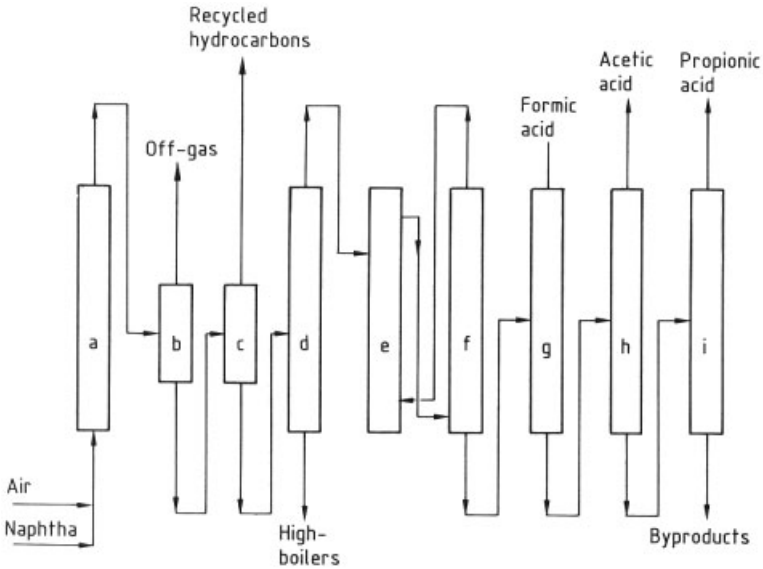

Figure 2. Production of propionic acid by the BP Chemicals process
a) Reactor; b) Gas – liquid separator; c) Liquid – liquid separator; d) Distillation column; e) Extraction; f) Separation of extracting agent; g) Formic acid distillation; h) Acetic acid distillation; i) Propionic acid distillation

temperature, but also by the degree of purity. The corrosion resistance of the materials used is therefore not always predictable and is frequently dependent on individual conditions.

Ordinary steel is totally unsuitable for the handling of propionic acid. Aluminum is only resistant at room temperature and to anhydrous, concentrated propionic acid; at higher temperature corrosiveness toward aluminum varies with the concentration of acid. Thus the corrosion maximum for 50 °C lies at 75 % acid, whereas there are two maxima at the boiling point, a weaker one at 1 % and a strong one at 99.8 % acid [68].

Copper and copper alloys are stable toward propionic acid up to its boiling point, but only if the solutions are free from air or oxidizing substances [68].

Of the stainless steels, the ferritic chromium steels (13 – 17 % Cr content, UNS no.: S 40 000) are unsuitable because propionic acid tends to cause pitting. Generally, austenitic CrNi steels and CrNiMo steels are used for handling propionic acid. At ambient temperature the CrNi steel X 10 CrNiTi 189 (DIN Materials no. 1.4541, UNS no.: S 30 400) is suitable. The CrNiMo steel X 10 CrNiMoTi 1810 (DIN no. 1.4571, UNS no.: S 31 600) is stable toward pure propionic acid up to its boiling point. However, impurities in propionic acid can make the use of titanium or Hastelloy (e.g., Hastelloy C NiMo 16 Cr 15 W, DIN no. 2.4819, UNS no.: N 10 276) equipment necessary, even below the boiling point [68].

Above the boiling point, up to ca. 230 °C Hastelloy C has shown excellent resistance even under oxidizing and reducing conditions. Above 230 °C only silver has proved to be a suitable material [25].

5. Storage and Transportation

Containers made of aluminum with a purity of 99.5 % (DIN no. 3.0255) or alloyed steels (DIN no. 1.4541/UNS no.: S 30 400 and 1.4571/S 31 600) are suitable for storing pure propionic acid. Aluminum is unstable toward aqueous propionic acid. Polyethylene containers can be used for temporary storage of propionic acid and as small packing drums. Plastics are not recommended for long-term storage. Glass-reinforced plastics are unsuitable even for short-term storage.

According to GefStoffV, propionic acid is classified as “corrosive”, R-phrase 34, and S-

phrase 2-23-26. Transport classification is as follows:

RID/ADR	class 8, no. 32 c
GGVE/GGVS	class 8, no. 32 c
ADNR	class 8, no. 32 c
IMDG Code (D-GGVSee)	class 8
IATA – DGR	class 8 1848
UN no.	1848

6. Quality Specifications

Propionic acid is only sold in pure form with a content of 99.5 wt %, determined acidimetrically. The maximum water content is 0.1 wt %; the sulfate, chloride, and heavy metal content is generally < 1 ppm. The color index is normally below 5 APHA. Oxidizable substances (ketones, aldehydes, esters) can be determined from the reaction time in the $KMnO_4$ test (DAB 7). Their proportions should be < 0.05 wt %.

The standard method for quality control is gas chromatography, e.g., with Chromosorb AW/DM 6 S as the carrier and LAC 446 + H_3PO_4 as the stationary phase. Detection is carried out with a flame ionization detector (FID).

The acid corresponding to the requirements of the Food Chemical Codex (FCC) is given in Table 4 [69].

Table 4. Sales specification of propionic acid [70]

Specification	Limits	Analytical method
Propionic acid content, wt %	99.5	gas chromatography
Water content, wt %	0.1	DIN 51 777
Aldehydes, wt %	0.05	gas chromatography
Readily oxidizable substances, wt %	0.05	DAB 7
Evaporation residue, wt %	0.01	DIN 53 172
Iron, ppm (mg/kg)	1	
Other heavy metals, ppm (mg/kg)	10	
Platinum – cobalt color, APHA	10	DIN 53 409
Density (20 °C), g/cm^3	0.990 – 0.998	DIN 51 757
bp, °C	140.7 – 141.6	DIN 51 751

7. Environmental Protection

Like most lower carboxylic acids, propionic acid is rapidly degraded biologically. The biodegradability in the Zahn – Wellens test is 95 % [71]. The biological oxygen demand for degradation after 5 days (BOD_5) is 1.3 g/g [72] which is 86 % of the total oxygen demand (TOD) of 1.513 g/g. The chemical oxygen demand (COD) is 1.4 g/g [72].

The fish toxicity test with the golden orfe resulted in no mortality with 4000 mg/L propionic acid in 48 h [73]. For the blue gill sunfish (*Leponius macrochirus*) an LD_{50} of 188 mg/L was determined in 24 h [72], for *Daphnia magna* an LD_{50} of 130 mg/L in 24 h and an LD_{50} of 50 mg/L in 48 h [72].

The limiting concentration of bacterial toxicity (*Photobacterium phosphoricum*) is 1240 mg/L [74]. For rats an oral LD_{50} of 3500 mg/kg was determined [73].

On the basis of the toxicity data for the rat, bacteria, and golden orfe, propionic acid is classified in Germany as mildly hazardous to water (WGK 1 = water endangerment class 1) [74].

The growth-inhibiting effect of the non-neutralized propionic acid (DIN 38 412/9) on *Scenedesmus suspicatus* was determined in 96 h as follows (EC = effective concentration and 0, 50, and 90 denote no, 50 %, and 90 % effect, respectively) [75]:

EC 0 = 12 mg/L
EC 50 = 43 mg/L
EC 90 = 79 mg/L

The bioaccumulation in an *n*-octanol – water mixture is log P_{OW} = 0.025/0.33 [72].

In the TA Luft propionic acid is classified in class 2 [76], i.e., if a chemical plant emits a quantity stream of ≥ 2 kg/h, the propionic acid concentration may not exceed 0.1 g/m^3.

The maximum workplace concentrations for propionic acid in Germany (MAK) and in the United States (TLV) in 1992 were 10 mL/m^3 ($\hat{=}$ 30 mg/m^3).

8. Uses

8.1. Use as Food and Feed Preservative

Propionic acid and its salts are broad-spectrum preservatives because of their bactericidal, fungicidal, insecticidal, and antiviral effects, and their acaricidal effect which is found at higher concentrations. Effects at lower inclusion rates

are more of a static nature. Minimum inhibition concentration (MIC) against fungi is 0.05 – 0.50 wt %, against bacteria 0.25 – 0.50 wt %, and against yeasts 0.10 – 1.0 wt % [77]. Efficacy of propionic acid against molds which cause spoilage of organic material such as foods and feeds is pronounced and of particular interest to industry. Gram-negative microorganisms such as *Escherichia coli* and *Salmonella ssp.* are more sensitive to propionic acid than gram-positive microorganisms. Minimum inhibition concentration against *Salmonella typhimurium*, *S. eimsbüttel*, *S. anatum*, *S. enteritidis*, and *S. montevideo* is 1.0 wt % after three days of action [78]. Propionates are ineffective against Salmonella in petri dishes but ammonium propionate and calcium propionate were successfully used to decontaminate Salmonella in meat meal at a 4.5 % inclusion rate with monitoring after four weeks of application [79]. Vaccinia virus was inactivated within 1 h by 1.0 wt % and within 7.5 min by 2.0 wt % propionic acid [80].

Mites in cereals (*Acarus siro*) are killed by 1.0 wt %. A repellent effect is achieved by adding only 0.5 wt % propionic acid. Corn beetles (*Sitophilus granarius*) are killed by 1.0 wt % propionic acid. Development of their eggs is impaired by 0.7 wt %.

Although the mode of action of propionic acid is not yet fully understood, the preservative effect is certainly based on the undissociated acid. At pH 4.88 the undissociated form is present to 50 % and at pH 6.0 to 7 %. This pH range is relevant for practical food and feed preservation. The undissociated acid penetrates into the living cell like a nutrient. Propionic acid inhibits intracellular enzymes which are essential for carbohydrate metabolism [81]. Morphological changes in Salmonella bacteria after treatment with 1 – 3 wt % propionic acid are lysis of the bacteria and loss of flagellum.

The broad efficacy spectrum of propionic acid forms the basis for a wide variety of applications. Its salts must be applied at concentrations equivalent to that of the acid but their effect is slower and not as long lasting. This is because the salts must be hydrolyzed to release the acid. Hydrolysis is incomplete and slow under practical conditions in feeds and foods, e.g., at 15 % moisture content and a pH value between 4 and 6.

Table 5. Application of propionic acid (PA) and its salts as preservatives for foods and feeds ∗

Type of material	Application rate of PA or its salts	Purpose/effect
Bread, bakery products	0.3 % Ca or Na salt; higher concentrations inhibit yeast fermentation	prevention of molds and formation of mycotoxins
Feed grain including corn	0.50 % PA at 16 % moisture 0.65 % PA at 20 % moisture 0.95 % PA at 26 % moisture 1.25 % PA at 32 % moisture 1.50 % PA at 36 % moisture	preservation for 6 months storage; reduction of microorganisms and nutrient losses; inhibition of mycotoxin formation and liberation of free fatty acids; protection from lumpiness and mold
Single feeds and complete feeds	0.2 % PA or 0.4 % salt at moisture level up to 14 %; higher rates at unfavorable conditions	increase of storage time up to 2 months; improvement of weight gain and feed conversion with all types of animal
Feeds of animal origin and complete feeds	0.5 % PA or 0.7 – 0.8 % Ca- or NH_4-salt	reduction of pathogenic flora, e.g., Clostridia, Enterobacteriacea
Feeds of animal origin	1.0 % PA or 2.0 – 3.0 % Ca- or NH_4-salt 2.5 – 3.5 % PA or 4.0 – 6.0 % Ca- or NH_4-salt	protection against recontamination with Salmonella after sterile manufacturing decontamination of heavily infected material
Brewer's yeast, liquid (ca. 10 % dry matter)	1.0 % PA	inactivation of yeast cells; direct feeding is possible after this treatment
Whey, liquid	0.3 % PA	increase of storage time by 2 days
	0.8 – 1.0 % PA	increase of storage time up to 1 week
Silage of grass, corn, and other forage	0.4 % PA for treatment of all material 1.5 – 2.0 L PA/m^2 silage surface, diluted in water by 1 : 4	improvement of lactic acid fermentation; inhibition of butyric and acetic acid producing bacteria; reduction of losses and mold on surface; prevention of nutrient losses

∗ Regulatory status in different countries for foods and feeds must be observed.

Calcium, sodium, and ammonium salts of propionic acid have the advantages of an almost neutral taste and smell. Therefore they can also be applied in food preservation whereas propionic acid is not suitable in this field because its odor is pungent and not acceptable to humans. Most animals prefer the acid taste of feed preserved with propionic acid and increase

their feed intake [82]. A feeding concept for pigs was introduced [83] in which organic acids such as propionic acid are of major importance to control digestion. The organic acids reduce the buffering capacity of feeds and influence the microflora content in the stomach and digestion parameters.

The various uses of propionic acid and its salts are listed in Table 5. Seeds and barley for malting process should not be treated with propionic acid because it inhibits germination.

All aspects regarding application technique, equipment, corrosive effects of propionic acid and propionates, storage and appropriate storage material, dosage, and efficacy are reviewed in [84]. The described effects can only be achieved by homogeneous distribution of the preservative in foods or feeds.

Table 6. Worldwide production capacities in 1989

Producer	Country	Capacity, t/a
BASF	FRG	60 000
Union Carbide	USA	68 000 [a]
Eastman – Kodak	USA	25 000 [a]
BP Chemicals	UK	30 000
Hoechst – Celanese	USA	7000
Daicel	Japan	2000
Total		192 000

[a] Used for production of various carboxylic acids.

Regulatory status of propionic acid and its ammonium, sodium and calcium salts is as follows:

FAO/WHO : ADI is not limited according to WHO Food Additives Series 1974, no. 5.

United States : GRAS status was reconfirmed in 1984 for propionic acid as a preservative for various foods, e.g., bread and other bakery products. The same applies to calcium and sodium propionates. Maximum inclusion rates are limited only by GMP (good manufacturing practice) according to CFR 21, §§ 184.1081, 184.1221, 184.1784 of April 1, 1989.

EC : Approved without upper limit and without specifying any fields of application according to EC Directive of Council 64/54 of Nov. 11, 1963.

Federal Republic of Germany : Approval for food preservation was withdrawn on March 31, 1988; a revision due to EC legislation is expected. According to the German feeds regulations there is no restriction or limitation for using propionic acid and its salts in all kind of feeds (feeds law 1989).

Other countries : There is no other country known in which propionic acid or its salts are not approved for food and feeds.

8.2. Other Uses

Zinc, cadmium, lead, and mercury propionates can be used as vulcanization regulators in rubber production [32].

Propionic acid is used as an intermediate, frequently in the form of its anhydride, for the production of esters. The most important esters are cellulose acetate propionate [*9004-39-1*] (CAP, → Cellulose Esters, Chap. 2.2.), from which thermoplastics are produced, and vinyl propionate (→ Vinyl Esters) which is used as a basic monomer for dispersions [→ Poly(Vinyl Esters)].

Methyl, ethyl, propyl, and butyl propionates are used as solvents for resins and paints [32]. Because of their characteristic fruity aromas various propionate esters are also used in flavors and fragrances.

Small quantities of propionic acid are also further processed to propionyl chloride, which is used as a reactive intermediate for the introduction of the propionyl group in syntheses (e.g., in the production of pharmaceuticals).

Until the end of the 1980s a considerable proportion of propionic acid was chlorinated and further processed to form herbicides (see Chloropropionic acids, Section 11.4). With the arrival of more potent, optically active herbicides and of products with a broader spectrum of activity, the importance of racemic herbicides based on propionic acid has decreased considerably.

9. Economic Aspects

The production capacity in 1989 for propionic acid was 190 000 t/a worldwide. Of this, however, ca. 90 000 t/a were only partially usable for propionic acid, because these plants were also used for the production of other acids. Details

of the distribution of the capacities in 1989 are given in Table 6.

Since 1984 the consumption of propionic acid has been stationary at ca. 130 000 t/a. No change is expected in the near future.

The areas of use of propionic acid and its distribution in 1989 are listed below.

Preserving animal feeds	45 %
Preserving bakery products	10 %
Herbicides	20 %
Plastics	10 %
Pharmaceuticals	10 %
Other	5 %

In the medium-term an increase in its use as a preservative for animal feeds is predicted. In contrast, the use in herbicides will decrease because of the increasing use of optically active substances (see Section 11.4.2.1).

10. Toxicology

Biochemistry, Biokinetics, and Metabolism. Propionic acid, or its metabolically activated form propionyl-CoA, occurs in mammals as a metabolic intermediate in the degradation of odd-numbered fatty acids, the side-chain oxidation of cholesterol, or in the oxidative degradation of certain amino acids (valine, isoleucine, threonine) [85], [86]. The proportion of propionic acid in the volatile fatty acids in the blood (total, contents 0.18 – 1.6 mmol/L) is up to 5 % [85]. Orally administered propionic acid or propionate is rapidly absorbed [87]. This can also be assumed for inhaled propionic acid.

Propionate is rapidly and completely metabolized after activation to propionyl-CoA by means of carboxylation to methylmalonyl-CoA and isomerization to succinate, which then enters the tricarboxylic acid cycle [85], [86]. Thus, conversion rates of 4.5 g propionic acid per hour for the liver of a 70 kg human can be calculated from in vitro investigations on slices of liver [85].

Acute Toxicity. The acute oral and inhalatory toxicity of propionic acid is low (LD_{50} oral rat: 2600 – 4290 mg per kilogram body weight, LC_{50} rat inhalatory $> 19.7\,mg\,L^{-1}\,4\,h^{-1}$). The only study on dermal application showed a higher toxicity (LD_{50} dermal rabbit: 500 mg per kilogram body weight), which may be caused by the strong irritant effect of the substance. Sodium and calcium propionates show a lower oral and a clearly lower dermal action in comparison with the acid [87–91].

Even after intravenous and subcutaneous administration there was no toxicity, but only a sedative effect shortly after the injection [85].

Propionic acid has an irritant/corrosive effect on the skin and an irritant effect on the rabbit eye [88], [89]. Its calcium salt, however, does not show these effects [88].

Subacute, Subchronic, and Chronic Toxicity. When $\leq 5\,\%$ propionic acid or sodium propionate was administered to rats with their feed (estimated substance uptake ≤ 4000 mg per kilogram body weight and day) from 28 days up to the total lifetime of the animals no toxic effects as a whole were found [88], [92–97]. However, with acid concentrations of ca. 0.4 % and higher, local changes in the epithelium of the forestomach occurred, caused by cell proliferation (increase in the basal cell and spindle cell layer). On longer administration, wart-like manifestations and ulcers developed [95], [96], [98]. On administration of 4 % propionic acid in animal feed for 21 days the cell division rate of the forestomach epithelium increased, as measured by ^{3}H-thymidine incorporation [99]. Similar local effects are also known from other short-chain fatty acids such as acetic, butyric, valeric, or caproic acids [92], [95]. Since the forestomach of the rat is a special feature of this species, 90-day investigations on beagle dogs, which have no forestomach, were carried out. Here propionic acid produced a proliferation in the epithelium of the gullet (esophagus), with a no-effect concentration in the feeds of 1 wt %. On feeding with calcium propionate no such effect was found up to the highest concentration used of 4.35 wt % [88].

Genotoxicity (Mutagenicity). In many in vitro investigations with bacteria, yeasts, and cell cultures no indication of a mutagenic effect of propionic acid was found [100–103]. The same applies to sodium [104], [105] and potassium [104] propionates. In vivo investigations on the induction of chromosomal changes by propionic acid in the Chinese hamster [101] or by

sodium propionate in the bone marrow of the rat [105] also gave negative results.

Teratogenicity. Investigations on developmental toxicity of calcium propionate in rats, mice, hamsters, and rabbits (300 and 400 mg per kilogram body weight and day) for 5 – 13 days (duration of test depending on animal species) [90] showed no effects.

Effects on Humans. Taking 6 g sodium propionate orally over several days gave slightly alkaline urine without any further effects [85]. Sodium propionate solutions (10 wt %, pH 7.2) have been used for the treatment of eye infections without irritant effects; 20 wt % solutions (pH 7 – 8.5) did not produce irritation or sensitization of the skin [106].

11. Derivatives

11.1. Salts

Ammonium propionate [*17496-08-1*], $C_3H_9NO_2$, M_r 105.11, is a strongly hygroscopic salt, which melts with absorption of water at 45 °C [107] (*bp* 70 – 75 °C at 1.3 kPa) [108]. Ammonium propionate is being increasingly used in the form of an aqueous solution as a preservative for animal feeds and cut grass.

Sodium propionate [*137-40-6*], $C_3H_5O_2Na$, M_r 96.06, melts at 281 – 284 °C to give an anisotropic liquid [109]. For its production and use see calcium propionate.

Calcium propionate [*4075-81-4*], $(C_3H_5O_2)_2$ Ca, M_r 186.18, crystallizes as the monohydrate in monoclinic plates and as the trihydrate. The anhydrous salt dissolves up to 41.7 wt % in water and is insoluble in ethanol [110]. The aqueous solutions can cause inflammation [111].

Production. The industrial production of calcium and sodium propionates is carried out by the neutralization of propionic acid with the corresponding hydroxides and subsequent spray-drying of the concentrated aqueous solutions.

Uses. Calcium and sodium propionate have fungistatic action. They are added to bread and other bakery products to prevent rope of bread caused by bacteria (Na propionate) and mold formation (Ca propionate). They are used as preservatives for animal feeds and for treatment of ketosis in domestic pets. Mixtures of sodium and calcium salts in an aqueous solution containing 2.5 % propionic acid are recommended for external use in treating mycosis [112].

11.2. Propionic Anhydride and Propionyl Chloride

Propionic anhydride [*123-62-6*], $(CH_3CH_2CO)_2O$, M_r 130.14, is a colorless liquid with an unpleasantly pungent odor, which severely irritates the eyes and mucous membranes. Physical properties are as follows [113]:

bp	167 °C
mp	– 45 °C
Flash point (open cup)	74 °C
Ignition temperature	315 °C
Density at 0 °C	1.0336 g/cm^3
at 50 °C	0.9791 g/cm^3
Kinematic viscosity (20 °C)	1.144 mPa · s
(50 °C)	0.7511 mPa · s

Toxicological data are as follows [114]:

LD_{50} (oral in rats)	2.36 g/kg
Skin (rabbit), irritation dose	500 mg (open, moderate)
Eye (rabbit), irritation dose	750 µg (severe)

Production. Propionic anhydride is produced industrially either by heating propionic acid at 8 bar to 235 °C and distilling off the water formed [115], or by heating propionic acid with propionyl chloride in an inert solvent to 160 °C [116].

Other possible routes involve the reaction of ethylene, carbon monoxide, and propionic acid in the presence of nickel carbonyl (Reppe synthesis [117–120]) or cobalt or nickel iodide [121].

Uses. Propionic anhydride is used as an esterifying agent for fatty acids and cellulose, for the production of alkyd resins and perfume oils, and for the introduction of the propionyl group into dyes and pharmaceuticals.

Propionyl chloride [*79-03-8*], CH_3CH_2CO Cl, M_r 92.53, is a colorless, volatile, and readily flammable liquid, whose vapors hydrolyze in moist air. It is corrosive, lachrymatory, and has a strongly pungent odor. With water and lower alcohols it reacts vigorously under solvolysis to give the acid or ester. Important physical properties are:

bp	80 °C
mp	− 94 °C
Flash point (open cup)	12 °C
d_4^{20}	1.0646

Production. Propionyl chloride is produced industrially by treatment of propionic acid with phosgene [*75-44-5*] or thionyl chloride [*7719-09-7*]. The reaction occurs at ca. 50 °C in the liquid phase in the presence of dialkylformamides. The product is then separated by distillation. Propionyl chloride can also be obtained by the reaction of propionic acid with PCl_5 [*10026-13-8*] [122].

Uses. Propionyl chloride is used for the introduction of the propionyl group and for the synthesis of propionate esters because of its high reactivity.

11.3. Esters

The esters of monohydric alcohols with propionic acid are formed in yields of > 90 % in the reaction of ethylene with carbon monoxide and the corresponding anhydrous alcohols in the presence of $Ni(CO)_4$ at 160 – 200 °C and 50 bar [123], [124].

The *ethyl ester* is also obtained in very good yields by treatment of ethanol with methylketene [*463-51-4*] [125].

Uses. The esters of the $C_1 - C_4$ alcohols are used as solvents for resins and paints [32]. Because of their specific, fruity aromas propionate esters are used as flavorings for foods and in the perfume industry [126]. Some esters and their properties are listed in Table 7.

11.4. Chloropropionic Acid and Derivatives

The 2- chloro- and 3- chloropropionic acids and their derivatives are still important industrially, while 2,2-dichloropropionic acid is virtually no longer used.

2-Chloropropionic acid is produced by direct chlorination of propionic acid, whereas 3-chloropropionic acid predominantly by addition of hydrogen chloride to acrylic acid.

11.4.1. Properties and Production

11.4.1.1. 2-Chloropropionic Acid and Derivatives

2-Chloropropionic Acid. Racemic 2-chloropropionic acid [*598-78-7*], $C_3H_5ClO_2$, is a colorless liquid that is miscible with water and most common organic solvents. Important *properties* are as follows : *bp* (101.3 kPa) 186 °C, *bp* (1.5 kPa) 84 °C; *mp* ca. − 12 °C; vapor pressure at 60 °C 0.4 kPa, at 100 °C 3.0 kPa; density at 20 °C (DIN 51 757) 1.27 g/cm^3 ; flash point (DIN 51 758) ca. 107 °C; ignition temperature (DIN 51 794) > 200 °C; explosion limits 3.7 – 14.3 vol %; water hazard class (WGK) 1 [127]; degradability (DIN 38 412 part 25) > 70 % [127]; TA Luft class I [127].

Commercial grade acid is ca. 95 % pure. The impurities are 2,2-dichloro- and 2,3-dichloropropionic acids, 2- chloropropionic anhydride, and unreacted propionic acid.

Production. 2-Chloropropionic acid is produced by reaction of propionic acid with chlorine at 110 – 120 °C in the presence of acid catalysts such as phosphorus trichloride, thionyl chloride, chlorosulfuric acid, or propionyl chloride [128], [129]. Since the reaction mixture is highly corrosive enamel, silver, glass (or leaded steel) are used as construction materials. One molar equivalent of hydrogen chloride is formed in the reaction and appears in the waste gas. From there it must be isolated and recycled on account of the economy of the process and disposal problems.

Methyl 2- chloropropionate [*17639-93-9*], $C_4H_7ClO_2$, is a colorless liquid that is miscible with almost all common organic solvents. Important *properties* are as follows: *bp* (101.3 kPa) 127 – 129 °C; *mp* < − 10 °C; vapor pressure at 20 °C 0.9 kPa, at 50 °C 4.5 kPa; density at 20 °C (DIN 51 757) 1.14 g/cm^3 ; flash

Table 7. Properties of some propionate esters

Ester	CAS registry no.	Molecular formula	*bp*, °C	Flash point, °C	Aroma/use
Methyl	[*554-12-1*]	$C_4H_8O_2$	79.8	6	rum-like/flavoring, fragrance
Ethyl	[*105-37-3*]	$C_5H_{10}O_2$	99.2	12	butter- or rum-like/flavoring, fragrance
Propyl	[*106-36-5*]	$C_6H_{12}O_2$	121.1	19	pear-like/flavoring, fragrance, solvent
Butyl	[*590-01-2*]	$C_7H_{14}O_2$	145.5	32	Solvent
Isopentyl	[*2438-20-2*]	$C_8H_{16}O_2$	160	41	apricot- or pineapple-like/flavoring, fragrance
Benzyl	[*122-63-4*]	$C_{10}H_{12}O_2$	222	94	jasmine-like/fragrance

point (DIN 51 755) ca. 36 °C; ignition temperature (DIN 51 794) 450 °C; explosion limits 2.8 – 13.9 vol %; water solubility at 20 °C: 22.9 g/L [130]; water hazard class (WGK) 1 [130]; degradability > 60 % [130]; TA Luft class III [130]; LD_{50} (oral, rat) 1080 mg/kg [130]; LC_{50} (by inhalation, rat) 15.3 mg L^{-1} 4 h^{-1} [130].

Methyl 2-chloropropionate is *produced* by esterification of 2-chloropropionic acid with methanol at 40 – 50 °C in the presence of acid catalysts such as sulfuric or hydrochloric acid.

2-Chloropropionyl chloride [*7623-09-8*], $C_3H_4Cl_2O$, is a colorless to pale yellow liquid that is miscible with almost all common aprotic solvents. *Physical properties*: *bp* (101.3 kPa) ca. 110 °C; *bp* (10 kPa) 45 °C; vapor pressure at 20 °C 3.0 kPa, at 60 °C 11.5 kPa; density at 20 °C (DIN 51 757) 1.27 g/cm^3; flash point (DIN 51 755) 34.5 °C; ignition temperature (DIN 51 794) 510 °C.

2-Chloropropionyl chloride is *produced* by reaction of propionyl chloride with chlorine at 70 – 90 °C in the presence of acid catalysts such as sulfuric or chlorosulfuric acid; or by reaction of 2-chloropropionic acid with, for example, thionyl chloride at 60 – 70 °C or with phosgene.

11.4.1.2. 3-Chloropropionic Acid and Derivatives

3-Chloropropionic acid [*107-94-8*], $C_3H_5ClO_2$, forms colorless plates that are readily soluble in water and alcohol; *mp* 41 – 42 °C; *bp* (101.3 kPa) 202 – 205 °C (decomp.); *bp* (2.1 kPa) 108 °C.

3-Chloropropionic acid is *produced* industrially by reaction of acrylic acid with hydrogen chloride at 40 – 45 °C or by reaction of propionic acid with chlorine under UV light at 20 °C [131].

3-Chloropropionyl chloride [*625-36-5*], $C_3H_4Cl_2O$, is a pale yellow liquid with *bp* 144 °C.

3-Chloropropionyl chloride is *produced* by reaction of acrylic acid with hydrogen chloride and phosgene in the presence of, for example, dimethylformamide as catalyst [132], or by reaction of propiolactone with thionyl chloride [133].

11.4.1.3. 2,2-Dichloropropionic Acid

2,2-Dichloropropionic acid [*75-99-0*], $C_3H_3Cl_2O_2$, is a colorless liquid that is miscible with water and alcohols; *bp* (101.3 kPa) 185 – 190 °C; *bp* (1.9 kPa) 90 – 94 °C; *mp* 13 – 15 °C; density at 20 °C (DIN 51 757) 1.40 g/cm^3.

2,2-Dichloropropionic acid is *produced* by reaction of propionic acid or 2-chloropropionic acid with chlorine at $\leq$ 180 °C in the presence of, e.g., phosphorus trichloride [134], thionyl chloride, organic acid chlorides [135], or sulfamic acid [136].

11.4.2. Uses

Chloropropionic acids and their derivatives mainly serve as intermediates. They are used to introduce the propionyl group in the production of biocides, dyes, and pharmaceuticals.

These alkylation reactions take place with elimination of chlorine. In the production of cosmetics and plastics additives the chlorine group in the chloropropionic acid is replaced by other active groups, such as thiol groups.

11.4.2.1. 2-Chloropropionic Acid and Derivatives

Racemic 2-chloropropionic acid and 2,2-dichloropropionic acid have been used predominantly for the production of herbicides. The demand for both acids is decreasing sharply, partly as a result of the development of new herbicides with different chemical structures. 2,2-Dichloropropionic acid is seldom used. Racemic 2-chloropropionic acid is being replaced by the optically active L-enantiomer in the herbicide field.

By far the greatest quantities of 2-chloropropionic acid and its derivatives are still used for the production of *phenoxypropionic acids* (**1**, **2**). Like the phenoxyacetic acids they are highly active herbicides of the growth regulator type [137]. Other derivatives with herbicidal action are phenoxyphenoxypropionates. (**3**) [138], naphthoxypropionamides (**4**) [139] and phenylaminopropionates (**5**) [140].

2-Chloropropionic acid, produced as described in Section 11.4.1, is racemic. Of the phenoxypropionic acids only the D-(+)-isomer shows herbicidal action. Its synthetic precursor is L-(−)-2-chloropropionic acid. Since the 1980s there has been a tendency to produce this class of herbicides from pure L-(−)-2-chloropropionic acid which is available, for example, in the form of its esters. These esters are produced, for example, by chlorination of the corresponding D-lactate esters with thionyl chloride. D-Lactic acid is obtained by a biotechnological process (→ Lactic Acid, Chap. 3.2.).

Activity. Phenoxypropionic acids are predominantly used to combat broad-leaved weeds in grasses, e.g., cereals. The spectrum of activity can, however, be shifted by appropriate substitution. The phenyl ethers of phenoxypropionate esters exhibit selective action towards damaging grasses and do not attack cultivated ones [138].

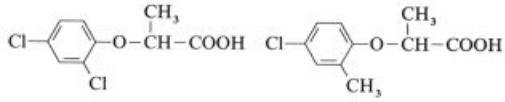

Dichlorprop or 2.4 DP (**1**) Mecoprop or MCPP (**2**)

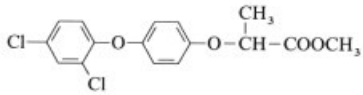

Illoxan (**3**)

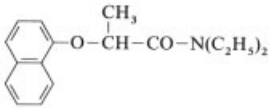

Devrinol (**4**)

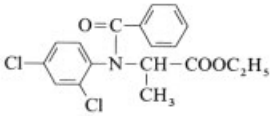

Suffix (**5**)

Trade Names of Herbicides based on Propionic Acid. U-46 range, racemates (**1**, **2**) (BASF), Duplosan range, optically active compounds (**1**, **2**) (BASF), Illoxan (**3**) (Hoechst), Devrinol (**4**) (Stauffer/ICI), Suffix (**5**) (Shell), Fusilade (**6**) (ICI), Verdict (**7**) (Dow).

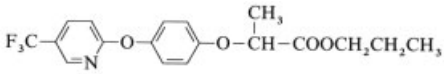

Fusilade (**6**)

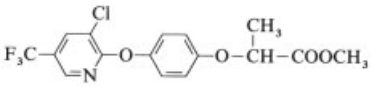

Verdict (**7**)

Esters of 2-Chloropropionic Acid. Methyl 2-chloropropionate is a building block of the fungicide Ridomil (**8**) (Ciba–Geigy) used in plant protection.

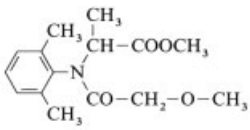

Ridomil (**8**)

2-Chloropropionyl Chloride. The Friedel–Crafts reaction of isobutylbenzene with chloropropionyl chloride provides a synthetic route for the antiphlogistic and antirheumatic ibuprofen (**9**).

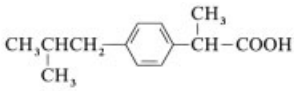

Ibuprofen (**9**)

11.4.2.2. 3-Chloropropionic Acid

Both 3-chloropropionic acid and its chloride are building blocks for the production of pharmaceuticals such as the antiepileptic beclamide (**10**) and the antitussive oxolamine (**11**) [141]. The acid is also used for the production of 3-mercaptopropionate esters, which are employed as stabilizers, for example in plastics and coatings.

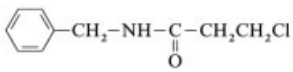

Beclamide (**10**)

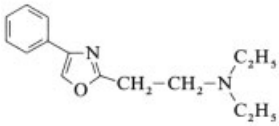

Oxolamine (**11**)

11.4.2.3. 2,2-Dichloropropionic Acid

The sodium salt of this acid was formerly an important herbicide (generic name delapon). It is now no longer used in Western Europe and the United States.

Trade Names. Basfapon (BASF), Dowpon (Dow).

11.4.3. Toxicology

Biochemistry, Biokinetics, and Metabolism. The sodium salt of 2-chloropropionic acid is an activator of the pyruvate dehydrogenase complex in both rats and dogs, which results in a lower blood glucose (in rats only), lactate, and pyruvate level. The sodium salt stimulates leucine metabolism in the perfused rat heart [142].

Acute Toxicity. The oral LD_{50} values of 2-chloropropionic acid in rats are in the range 500 – 800 mg/kg [142–144]. The LD_{50} in the mouse (i.p.) is 102 mg/kg [143], and the LC_{50} (rat, inhalation) > 3.38 mg/L at 4 h exposure [143]. Exposure of rats to an enriched or saturated atmosphere of chloropropionic acid (20 °C) leads to their death after an exposure time of more than 2 h [143]. 2-Chloropropionic acid has a corrosive effect on rabbit's skin and an irritant effect on the rabbit's eye [143].

The sodium salt shows a somewhat lower oral and inhalatory toxicity. Its LD_{50} (rat, dermal) is > 2000 mg/kg [143], [145], [146]. The sodium salt is not irritant to the skin but irritates the eye [143]. No sensitizing effects were observed in guinea pigs [143].

Toxicity on Long-Term Exposure. No clinical effects were observed if 0.1 wt % (2-chloropropionic acid) was administered to rats with their feed (estimated substance uptake up to 78 mg/kg per day) over 42 days. Higher concentrations (0.25, 0.5, and 1 wt %) lead to an increased sensitivity and ataxia after the sixth day of administration. Ataxia – reversible in some cases during a recovery period of 36 days – and tremor were observed in the groups that received the higher concentrations. If the *neutralized acid* (neutralizing agent NaOH) was administered at a concentration of 0.04 mol/kg (ca. 0.5 wt %) for 12 weeks, the same symptoms occurred, however, with temporary delay. Foci of necrotic cells in the inner granular cell layer of the cerebellum (concentration 0.25 – 1.0 wt %), atrophy of the testicular germinal epithelium (concentration 0.1 – 1.0 wt %), and thymic lymphoid necrosis (concentration 1.0 wt %) were observed [142]. In another experiment involving rats (concentration in feed 100, 800, and 4000 ppm over 3 months) with the *sodium salt* the same toxic effects occurred. Neurotoxicity and testicular changes were present in the group administered 4000 ppm. At a concentration of 800 ppm only marginal clinical signs of neurotoxicity remained and at 100 ppm no toxic symptoms were observed [143].

Mutagenicity. 2-Chloropropionic acid and its sodium salt showed no mutagenic activity in the Ames test [143], [147].

12. References

1. H. Sackmann, F. Sauerwald, *Z. Phys. Chemie (Leipzig)* **195** (1950) 295 – 312.
2. C. H. D. Calis-van Ginkel et al., *J. Chem. Thermodyn.* **10** (1978) 1083 – 1088.
3. A. Apelblat, E. Manzurola, *Fluid-Phase Equilib.* **32** (1987) 163 – 193.
4. M. Blinc, R. Blinc, *J. Polym. Sci.* **32** (1958) 506 – 508.
5. K. Tyuzyo, *Bl. Chem. Soc. Jpn.* **30** (1957) 782 – 789, 851 – 856.
6. J. J. Christensen, R. W. Hanks, R. M. Izatt: *Handbook of Heats and Mixing,* Wiley & Sons, New York 1982, pp. 1173 – 1175, 1449.
7. [6] Supplementary vol. 1988, pp. 858 – 860.
8. J. Timmermans: *The Physico-Chemical Constants of Binary Systems in Concentrated Solution,* Interscience, vol. 2, New York 1959, p. 941.
9. D. R. Linde: *Handbook of Chemistry and Physics,* 70th ed., CRC Press, Boca Raton, 1989, D – 15.
10. G. Glaxton: *Physical and Azeotropic Data,* National Benzol and Allied Product Association, W. Heffer & Sons, Cambridge 1958.
11. L. H. Horsley: "Azeotropic Data III," *Adv. Chem. Ser.* **116** (1973) p. 4 ff.
12. T. Ito, F. Yoshida, *J. Chem. Eng. Data* **8** (1963) 315 – 320.
13. J. Gmehling, U. Onken, J. R. Rarey-Nies: "Vapor – Liquid – Equilibrium Data Collection," *Chemistry Data Series,* **vol. I, part 1 (1977), 1 a (1981), 1 b (1981), 2 d (1982), 3 + 4 (1979), 5 (1982).**
14. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* vol. 1, part 2, Pergamon Press, London 1963.
15. [14] vol. 2, part 2 (1964).
16. *VDI-Wärmeatlas,* 5th ed., Düsseldorf 1988.
17. The Texas A & M University System, Table 23-2-1-(1.1210)-d, p. 1, College Station 1984.
18. R. R. Dreisbach, R. A. Martin, *Ind. Eng. Chem.* **41** (1949) 2875.
19. J. J. Jasper, *J. Phys. Chem. Ref. Data* **1** (1972) 851.
20. W. Jobst, *Int. J. Heat Mass Transfer* **7** (1964) 725.
21. G. Kortüm, W. Vogel, K. Andrussow, *Pure Appl. Chem.* **1** (1961) 243.
22. D. D. Perrin, *J. Chem. Soc.* 1959, 1710, 1714.
23. W. J. Canady, H. M. Papée, K. J. Laidler, *Trans. Faraday. Soc.* **54** (1958) 502.
24. *Landolt* – Börnstein, 6th. ed., vol. IV/4 a.
25. *Ullmann,* 4th ed., **19,** 454 ff.
26. M. A. Goodman, S. L. Whittenburg, *J. Chem. Eng. Data* **28** (1983) 350 – 351.
27. BASF AG Ludwigshafen, Result of Techn. Lab., Ludwigshafen 1969.
28. T. Ackermann, F. Schreiner, *Z. Elektrochem.* **62** (1958) 1143 – 1151.
29. R. W. Gallant, *Hydrocarbon Process* **47** (1968) no. 6, 139 – 148.
30. E. Przewalski, *J. Prakt. Chem.* **88** (1913) 500.
31. *Kirk-Othmer,* 2nd ed., **16,** 554.
32. *Ullmann,* 3rd ed., **14,** 383.
33. *Ullmann,* 3rd ed., **8,** 744.
34. Celanese, BP 771 583, 1957.
35. SRI-Process Economics Program, no. 42, A 1, Menlo Park 1975, pp. 18.
36. Dow Chemical, US 1 926 068, 1933 (C. J. Strosacker, C. C. Kennedy, E. L. Pelton).
37. *Ullmann,* 4th ed., **9,** 161 ff.
38. Texaco Dev., DE 3 124 720, 1981 (J. F. Knifton).
39. M. Sittig, *Chem. Proc. Monogr.* no. 8 (1965) p.76.
40. [39] no. 23 (1966) pp. 105.
41. Monsanto, BE 793 203, 1971 (M. D. Forster, D. E. Morris).
42. Monsanto, US 3 989 748, 1968 (F. E. Paulik, A. Hershman, J. F. Roth).
43. Monsanto, US 3 989 747, 1968 (J. H. Craddock, J. F. Roth, A. Hershman, F. E. Paulik).
44. BP Chemicals, DE-OS 2 101 909, 1970 (G. E. Foster, J. R. Bethell).
45. Monsanto, BE 796 294, 1972 (D. E. Morris).
46. BP Chemicals, GB 1 363 961, 1972 (M. J. Wriglesworth, D. J. Westlake).
47. Monsanto, US 3 944 603, 1972 (D. E. Morris).
48. Monsanto, DE-AS 2 263 442, 1972 (D. Forster, D. E. Morris).
49. BASF, DE-AS 2 604 545, 1976 (K. Schwirten, W. Disteldorf, W. Eisfeld, R. Kummer).
50. E. Drent, *ACS Symp. Ser.* **328** (1987) 154 – 175.
51. Texaco, US 4 362 822, 1981 (J. F. Knifton).
52. P. Blanc, G. Goma, *Biotechnol. Lett.* **11** (1989) 189 –194.
53. M. J. T. Carrondo, J. P. S. G. Crespo, M. J. Moura, *Appl. Biochem. Biotechnol.* **17** (1988) 295 – 312.

54. H. Dellweg: *Biotechnology,* **vol. 3,** Verlag Chemie, Weinheim 1983, p. 472.
55. Dow Chemical, US 3 151 155, 1964 (J. W. McKoy, N. Swanson).
56. *Ullmann,* 4th ed., **7,** 123 ff.
57. D. Powergas, J. Matthey, *Chem. Eng. (N.Y.)* **84** (1977) Dec. 5, 110 – 115.
58. D. Forster, A. Hershman, D. E. Morris, *Catal. Rev. Sci. Eng.* **23** (1981) 89 – 105.
59. B. Cornils, R. Payer, K. C. Traenckner, *Hydrocarbon Process.* **54** (1975) no. 6, 83 – 91.
60. *Hydrocarbon Process* **54** (1975) no. 11, 87.
61. SRI-Process Economic Program no. 37, 1968, pp. 83 ff.
62. SRI-Process Economic Program no. 37 a, 1973, pp. 33 ff.
63. Distillers, GB 771 991, 1957 (A. Elce, I. K. M. Robson, D. P. Young).
64. BP Oil, GB 743 989, 1956 (J. Habeshaw et al.).
65. Distillers, GB 743 991, 1956 (A. Elce, I. K. Miles, D. P. Young).
66. Distillers, FR 1 091 333, 1954 (A. Elce, I. K. Miles, D. P. Young).
67. *Hydrocarbon Process* **54** (1975) no. 11, 102.
68. G. B. Elder in B. J. Monitz, W. I. Pollock (eds.): *Process Ind. Corros.,* NACE Houston Tex., 1986, pp. 287 – 296.
69. *Food and Chemical Codex,* 3rd ed., National Academy Press, Washington, DC, 1981.
70. BASF, data sheet D 030 d, e, Ludwigshafen, Aug. 1988.
71. OECD Guidelines for Testing of Chemicals, 302 B, Paris 1981.
72. K. Verschueren: *Handbook of Environmental Data of Organic Chemicals,* 2nd ed., 1983, p. 1023.
73. BASF, Entwurf zum Sicherheitsdatenblatt der BASF, Ludwigshafen 1988.
74. Data sheet no. 140, Katalog wassergefährdender Stoffe, Gemeinsames Ministerialblatt, Bonn 1985.
75. BASF AG, unpublished results.
76. T. A. Luft, Gemeinsames Ministerialblatt, ed. A, no. 7, Bonn, Feb. 1986.
77. BASF, Periodical Animal Nutrition in Research and Practice, vol. 2, no. 4 (1973), no. 5 (1974), no. 7 (1975), no. 11 (1979).
78. I. Brglez, *Prax. Vet. Br.* 1977 no. 1, 17 – 22.
79. V. Haßling: *Rekontaminationsprobleme von Tiermehlen gegen Salmonellen durch Propionsäure, Propionate und Ameisensäure,* Dissertation, Universität Gießen 1985.
80. I. Wekerle, *Tierärztl. Umsch.* **43** (1988) 646 – 654.
81. K.-H. Bässler, *Z. Lebensm. Unters. Forsch.* **110** (1959) 28 – 42.
82. H. Giessler, S. Dammert, *Landwirtsch. Forsch. Sonderh.* **38** (1981) 484 – 489.
83. G. Bolduan et al., *J. Anim. Physiol.* **59** (1988) 72 – 78.
84. G. Gaus: BASF Documentation MEA 3310 AD, Oct. 89, Technical Symposium Aug. 9, 1989, Whitestone, S.C., USA.
85. K. H. Bässler, *Z. Lebensm. Unters. Forsch.* **110** (1959) 28 – 42.
86. P. Karlson: *Kurzes Lehrbuch der Biochemie,* Thieme Verlag, Stuttgart 1987.
87. WHO Food Additives Series no. 5, Genf 1974, pp. 110 – 118.
88. BASF AG, Department of Toxicology, unpublished results.
89. H. F. Smyth et al., *Am. Ind. Hyg. Assoc. J.* **23** (1962) 95 – 107.
90. *Patty's* 2nd, pp. 1779 – 1780 and 3rd ed., pp. 4911 –4913.
91. NIOSH, Registry of Toxic Effects of Chemicals, CD-Rom Version, update 9001.
92. K. Mori, *GANN* **44** (1953) 421 – 427.
93. S. Imai et al., *Nara Igaku Zasshi* **32** (1981) 715 – 722.
94. W. D. Graham et al., *J. Pharm. Pharmacol.* **6** (1954) 534 – 545.
95. H.-J. Altmann, W. Grunow, unpublished report BGA (1988).
96. W. Griem, *Bundesgesundheitsbl.* **28** (1985) 322 – 327.
97. *Bundesgesundheitsbl.* no. 30, Bonn 1987, 370 – 371.
98. BP International, unpublished results.
99. C. Rodriguez et al., *Toxicology* **38** (1986) 103 – 117.
100. Report of Litton Bionetics for FDA, *Nat. Tech. Inf. Serv.* USA PB-266 897 (1976).
101. A. Basler et al., *Food Chem. Toxic.* **25** (1987) 287 –290.
102. W. v.d. Hude et al., *Mutat. Res.* **203** (1988) 81 – 94.
103. W. Szybalski, *Ann. N.Y. Acad. Sci.* **76** (1958) 475 –489.
104. M. Ishidate et al., *Food Chem. Toxic.* **22** (1984) 623 –636.
105. T. Kawachi et al. in R. Montesano et al. (eds.): *IARC Sci. Publ.* **27** (1980) 323 – 330.
106. W. W. Heseltine, *J. Pharm. Pharmacol.* **4** (1952) 120 –122.
107. S. Zuffanti, *J. Am. Chem. Soc.* **63** (1941) 3124.
108. R. Escales, H. Koepke, *J. Prakt. Chem.* **87** (1913) 264.

109. G. E. Symons, A. M. Buswell, *Ind. Eng. Chem.* **24** (1932) 460.
110. A. Renard, *C. R.* **104** (1887) 914.
111. B. Behrens, J. Wajzer, *Biochem. Z.* **264** (1933) 124.
112. *Index Merck,* 9th ed., Verlag Chemie, Weinheim 1961, p. 633.
113. *The Merck Index,* 10th ed., Merck & Co Inc., Rahway 1983, p. 1127.
114. NIOSH, Registry of Toxic Effect of Chemical Substances, 1980.
115. Knapsack AG, DE 951 809, 1954 (W. Vogt).
116. G. Leiderer, DE-AS 1 932 303, 1969.
117. W. Reppe, H. Kröper, *Ann.* **582** (1953) 47, 61.
118. R. E. Brooks, W. F. Gresham, J. v. Hardy, J. M. Lupton, *Ind. Eng. Chem.* **49** (1957) 2004.
119. BASF, DE 848 355, 1943 (W. Reppe, H. Kröper).
120. BASF, DE 863 194, 1943 (W. Reppe, H. Kröper).
121. Monsanto, US 3 989 751, 1974 (D. Forster, A. Hershman).
122. J. D. Roberts, *J. Am. Chem. Soc.* **72** (1950) 4237.
123. IG Farben, DE 765 969, 1940 (W. Reppe, H. Kröper).
124. BASF, DE 915 567, 1953 (W. Reppe, H. Kröper).
125. Rohm & Haas, US 1 898 687, 1931 (F. O. Rice).
126. *Kirk-Othmer,* 3rd ed., **10,** 456 ff.
127. BASF AG, Company data from Aug. 15, 1991.
128. Dow Chem., US 2 010 685, 1932.
129. Dow Chem., US 1 993 713, 1935.
130. BASF AG, Company data from Aug. 15, 1991.
131. 3 M, US 2 682 504 , 1954.
132. BASF, DE 1 140 185, 1963.
133. Goodrich, US 2 548 161, 1951.
134. Dow Chem., US 2 809 992, 1957.
135. Brit. Celanese, GB 892 564, 1962.
136. BASF, DE 1 133 710, 1963.
137. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. V,** Springer Verlag, Berlin 1977, p. 188 ff.
138. Hoechst, DE-OS 2 223 894, 1973.
139. Stauffer Chem., GB 1 066 606, 1967.
140. Shell Int., GB 1 164 160, 1969.
141. A. Kleemann, I. Engel: *Pharmazeutische Wirkstoffe,* Thieme Verlag, Stuttgart 1978.
142. J. L. O'Donoghue, *Neurotoxicity of Industrial and Commercial Chemicals,* **vol. II,** CRC Press, Inc., Boca Raton, Florida 1985, pp. 115 – 116.
143. BASF AG, Department of Toxicology, unpublished results, Ludwigshafen 1960, 1988, 1990, 1992.
144. RTECS, update 8910.
145. J. L. Morrison, *J. Pharmacol. Exp. Pharmaceut.* **86** (1946) 336 – 338.
146. E. Yount et al., *J. Pharmacol. Exp. Therap.* **222** (1982) 501 – 508.
147. Japan Chemical Industry Ecology – Toxicology & Information Center (JETOC), *Newsletter* **4** (1985) 15 (14 – 20).

Propiophenone → Ketones
Propylene → Propene
Propylene Chlorohydrin → Chlorohydrins
Propylene Glycol → Propanediols

Propylene Oxide

Dietmar Kahlich, Dow Deutschland Inc., Stade, Federal Republic of Germany (Chaps. 1 – 3, 4.4, 8 – 10)

Uwe Wiechern, Dow Deutschland Inc., Stade, Federal Republic of Germany (Chaps. 4.1, 4.4, 5 – 7)

Jörg Lindner, Dow Deutschland Inc., Stade, Federal Republic of Germany (Chaps. 4.2 – 4.4)

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production and Raw Materials . . . 4
4.1. Production Route via Propylene Chlorohydrin 4
4.1.1. Chlorohydrin Process 4
4.1.2. Other Chlorohydrin Technologies . . . 9
4.2. Indirect Oxidation Routes 11
4.3. Direct Oxidation Routes 14
4.4. Comparison of the Various Propylene Oxide Technologies 15
5. Environmental Protection and Ecotoxicology 15
6. Quality Specifications and Analysis 16
7. Handling, Storage, and Transportation 16
8. Uses 19
9. Economic Aspects 19
10. Toxicology and Occupational Health 20
11. References 22

1. Introduction

Propylene oxide (PO) [*75-56-9*], 1,2-epoxypropane, methyloxirane, is a very reactive substance and one of the most important chemical intermediates.

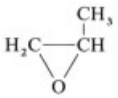

Since the early 1950s it has become increasingly important to the chemical industry. The demand for PO continues to grow throughout the world; annual worldwide production capacity in 1991 was ca. 3.9×10^6 t and in 1993 it was ca. 4.0×10^6 t.

Today PO is a commodity that can be produced economically only in world-scale plants with capacities of 100 000 t/a or above. It is the starting material for a broad spectrum of products, including polymers (polyurethanes, polyesters), oxygenated solvents (propylene glycol ethers), and industrial fluids (monopropylene glycol and polyglycols).

2. Physical Properties

Propylene oxide, C_3H_6O, M_r 58.081, is a colorless, low-boiling liquid with a sweet, ethereal odor. It is also reactive and highly flammable, with a wide explosive range. Flammable concentrations of vapor form readily at room temperature. Important physical constants are as follows [7–20]:

mp (101.3 kPa)	−111.93 °C
bp (101.3 kPa)	34.23 °C
Flash point	− 37 °C
Critical temperature	209.1 °C
Critical pressure	4920 kPa
Critical density	312 kg/m^3
Critical compressibility factor	0.2284
Autoignition temperature in air at 101.3 kPa	430 °C
Explosive limits in air (STP)	
lower	2.3 vol %
upper	36.0 vol %
Heat of combustion (25 °C, 101.3 kPa)	−33 035 kJ/kg
Heat of polymerization	−1500 kJ/kg
Heat of fusion	112.6 kJ/kg
Heat of solution in water at 25 °C	−45 kJ/kg
Heat of formation of the ideal gas (25 °C)	−1600 kJ/kg
Heat of formation of liquid PO (25 °C)	−2080 kJ/kg
Standard enthalpy (298.15 K)	248 kJ/kg
Standard entropy (298.15 K, 1 atm)	4.94 kJ kg^{-1} K^{-1}
Free energy of formation (25 °C, 101.3 kPa)	459 kJ/kg
Cubic expansion coefficient at 20 °C	0.00151 1/K
Solubility of PO in water at 20 °C	40.5 wt %
Solubility of water in PO at 20 °C	12.8 wt %

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_239

Dipole moment	2.010 Debyes
Evaporation rate (butyl acetate = 1)	33.7
Molecular diffusion volume	4.937×10^{-5} m^3/mol
n_D^{20}	1.36610
n_D^{25}	1.36335
n_D^{30}	1.36060
Henry constant at 25 °C	8.65×10^{-5} Pa m^3/mol

Table 1 provides temperature-dependent physical properties for both gaseous and liquid PO.

PO has unlimited miscibility with most common organic solvents, but it is subject to a miscibility gap with water (Fig. 1). The binary system PO – water is characterized in Figure 2. This and other binary and ternary vapor-liquid equilibria systems are described in [22].

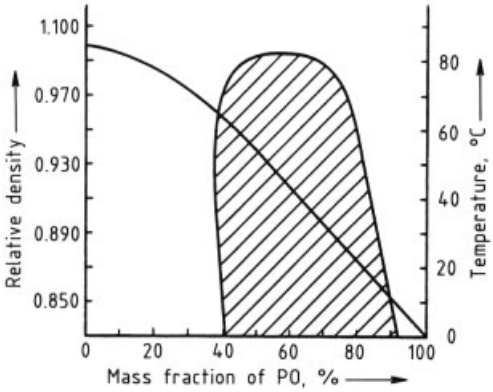

Figure 1. Miscibility gap and density of the propylene oxide – water system [19]

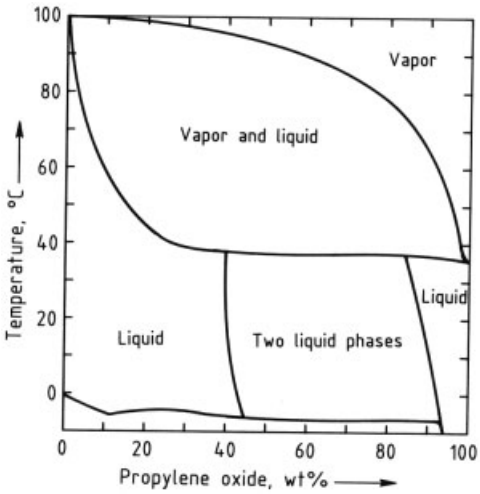

Figure 2. Phase diagram for the propylene oxide – water system [20], [21]

PO forms azeotropes with dichloromethane, diethyl ether, cyclopentene, isoprene, cyclopentane, 2-methyl-1-butene, pentene, pentane, cyclohexane, hexene, and hexane [22], [23].

PO exists in two optical isomers [24], [25], but commercial PO is a racemic mixture, and it is to this mixture that the listed properties refer.

3. Chemical Properties

PO is a highly reactive, versatile chemical intermediate. Its polarity and the strained three-membered epoxide ring cause it to be opened easily by reaction with a wide variety of substances. In this respect it is analogous to ethylene oxide (→Ethylene Oxide), and the number of possible derivatives is very large.

Most industrial reactions of PO are catalyzed, although PO alone reacts readily with all compounds containing active hydrogen atoms (including the hydrogen halides), as well as chlorine and ammonia. Alkaline catalysts are preferred (e.g., sodium hydroxide, sodium acetate, sodium alkoxide, or potassium alkoxide). Acidic materials, such as sulfuric acid, boron trifluoride, stannic chloride, or acid clays (which are sometimes used at lower reaction temperatures) often form byproducts by converting oxides to cyclic ethers, or by isomerizing them to aldehydes.

The most important type of reaction is the addition of PO to compounds containing labile hydrogen, but it also condenses with a number of substances that do not contain reactive hydrogen.

The most important reaction of PO from a commercial standpoint is its violent polymerization to form poly(ether polyols) in the presence of catalysts such as bases, acids, or salts. Chain initiators are substances containing active hydrogen atoms (e.g., water, glycols, phenols, amines, or carboxylic acids). PO forms mixed polyols with ethylene oxide, tetrahydrofuran, 3,4-epoxy-1-butene (butadiene monoxide), and carbon dioxide.

PO is hydrolyzed by water in an uncatalyzed, liquid-phase reaction to give monopropylene glycol, dipropylene glycol, tripropylene glycol, and polyglycols. PO reacts with ketones or aldehydes to form cyclic ketals or acetals, and with carbon dioxide to form propylene carbonate. Hydrogen halides convert PO to propylene halo-

Table 1. Physical properties of propylene oxide at various temperatures [18]

Temperature, °C	Density		Heat of vaporization,	Vapor pressure	Specific heat, kJ kg^{-1} K^{-1}		Thermal conductivity, 10^{-4} W cm^{-1} K^{-1}		Surfacetension, mN/m	Dynamic viscosity, mPa · s	
	Vapor, g/L	Liquid, kg/m^3	kJ/kg	kPa	Vapor	Liquid	Vapor	Liquid		Vapor	Liquid
−20	0.24	877	522	8.6	1.10	1.95	0.78	16.6	(29.8)	0.0076	0.487
0	0.63	854	504	24.4	1.18	2.00	0.91	15.8	(26.8)	0.0082	0.388
20	1.44	830	485	58.8	1.27	2.06	1.05	15.0	23.9	0.0087	0.319
40	2.90	804	465	124.5	1.37	2.13	1.21	14.2	21.1	0.0093	0.269
60	5.33	778	444	238.0	1.48	2.20	1.39	13.4	18.3	0.0099	0.232
80	9.14	749	421	418.0	1.61	2.28	1.60	12.5	15.5	0.0106	0.203
100	14.81	719	396	686.0	1.75	(2.37)	1.84	11.7	12.8	0.0112	0.180
120	23 .06	686	368	1068.0	1.93	(2.47)	2.11	10 .8	10.2	0.0120	0.162
140	34 .93	650	336	1590.0	2.15	(2.61)	2.42	(9 .8)	(7.6)	0.0128	0.147
160	52 .24	609	(298)	2280.0	2.47	(2.84)	(2.79)	(8.9)	(5.2)	0.0137	0.135
180	78.81	599	(250)	3175.0	3.07	(3 .27)	(3.25)	(7.9)	(2.9)	0.0149	0.125
200	126.5	489	(180)	4325.0	5.08	(4 .56)	(3.90)	(6.9)	(1.0)	0.0170	0.116

Extrapolated values are given in parentheses.

hydrins, and ammonia gives mono-, di-, and triisopropanolamine.

Over alumina PO isomerizes to propanal and acetone, and over Li_3PO_4 catalysts to allyl alcohol. Hydrogenation over nickel leads to 1-propanol. Many other reactions are also possible, including reactions with natural products like starch or cellulose. Table 2 summarizes the most important PO reactions.

All the reactions involving PO may become violent if not properly controlled, and possible applications should be investigated very thoroughly.

Under pyrolytic conditions PO isomerizes to propanal, acetone, methyl vinyl ether, and allyl alcohol [26]. Decomposition of PO does not result in hazardous products.

PO is a stable compound under moderate conditions, but polymerization may occur on contact with highly active catalytic surfaces such as iron, tin, or aluminum chlorides, peroxides of iron and aluminum, or alkali-metal hydroxides and clay-based absorbent materials. Anything that destroys the neutrality of PO may also cause polymerization. The chemical stability of PO and its polymerization under storage conditions are discussed in [27].

4. Production and Raw Materials

The selection of production routes is decisively influenced by the application and market potential of coproducts, as well as by availability of raw materials and possibilities for byproduct management [28], [29].

Technologies developed up to this point can be divided into:

1) Chlorohydrin processes (Section 4.1)
2) Indirect oxidation processes (Section 4.2)
3) Direct oxidation processes (Section 4.3)

Only the chlorohydrin and the indirect oxidation processes are practiced currently on an industrial scale. The following sections highlight the various processes and describe in more detail those currently in use (see Fig. 3).

It has so far not been possible to produce PO in a technically and economically satisfactory manner by direct gas-phase oxidation (with silver catalysts) in analogy with the process used for ethylene oxide. All PO processes generate significant amounts of coproducts. For the conventional chlorohydrin process this means up to two moles of salt (sodium or calcium chloride) in the form of dilute brine solution per mole of PO. In the case of the direct oxidation methods, complex organic mixtures arise, which must then be separated via distillation and either sold, further converted, or incinerated. Indirect oxidation technologies also lead to PO and coproducts.

4.1. Production Route via Propylene Chlorohydrin

See also →Chlorohydrins.

Table 2. Summary of the most important PO reactions

Reactant	Product(s)	Structural formula
Polymerization (ROH as initiator)	polyether polyols	$R[-O-\underset{\underset{CH_3}{\vert}}{CH}-CH_2]_n-OH$
Water	monopropylene glycol ($n = 0$) dipropylene glycol ($n = 1$) tripropylene glycol ($n = 2$)	$CH_3-\underset{\underset{OH}{\vert}}{CH}[-CH_2-O-\underset{\underset{CH_3}{\vert}}{CH}]_n-\underset{\underset{OH}{\vert}}{CH_2}$ (and isomers)
Alcohols and phenols (ROH)	propylene glycol ethers	$CH_3-\underset{\underset{OH}{\vert}}{CH}-\underset{\underset{OR}{\vert}}{CH_2}$ (and isomers/highers)
Ammonia	isopropanolamine	$CH_3-\underset{\underset{OH}{\vert}}{CH}-\underset{\underset{NH_2}{\vert}}{CH_2}$ (and isomers/highers)
Carbon dioxide	propylene carbonate	$CH_3-\underset{\underset{O}{\vert}}{CH}-\underset{\underset{O}{\vert}}{CH_2}$ $C=O$
Isomerization	allyl alcohol acetone propanal	$CH_2=CHCH_2OH$ CH_3COCH_3 CH_3CH_2CHO

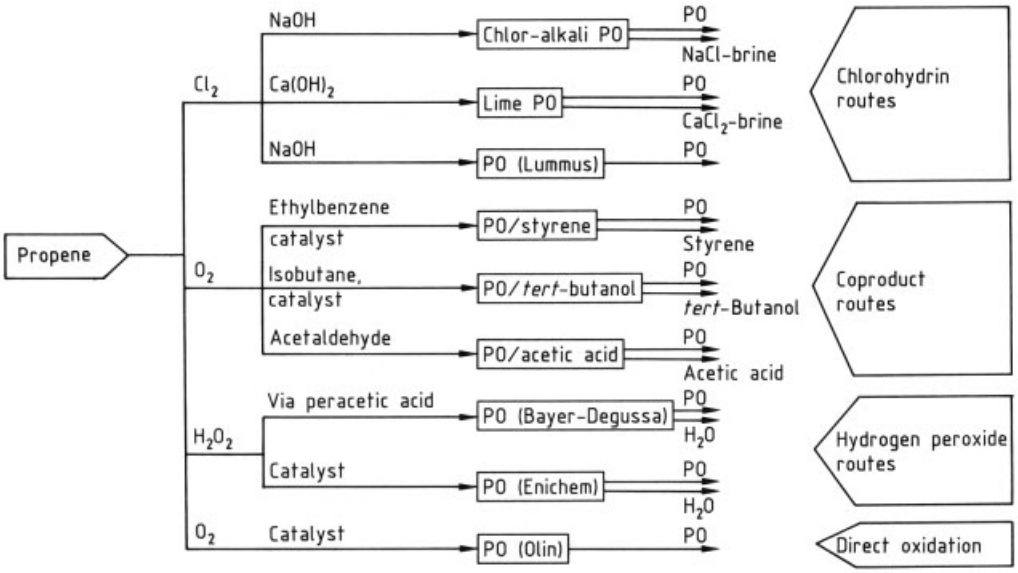

Figure 3. Summary of propylene oxide production routes

4.1.1. Chlorohydrin Process

Conventional production of PO is based on the dehydrochlorination of propylene chlorohydrin [*78-89-7*] with a base [30]. Technically, such units are similar to the chlorohydrin units used to produce ethylene oxide (EO) prior to development of the direct oxidation process. Because the demand for PO has increased rapidly, old plants have been steadily improved and new state-of-the-art, grass-roots PO plants have been built throughout the world.

The present commercial chlorohydrin process is carried out in two main steps: synthesis of

propylene chlorohydrin (PCH), and subsequent dehydrochlorination of PCH to PO. These steps are followed by PO purification and wastewater treatment.

Chlorohydrination. In the PCH synthesis reaction (chlorohydrination), propene and chlorine are mixed in an aqueous solution. A propene – chloronium complex is formed as an intermediate (Eq. 1) [31].

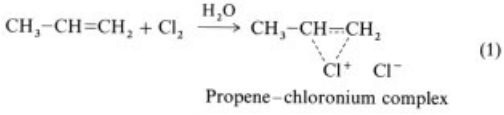

The chloronium ion then reacts with water to form hydrochloric acid and the propylene chlorohydrin isomers 1-chloro-2-propanol (90 %) and 2-chloro-1-propanol (10 %) (Eq. 2).

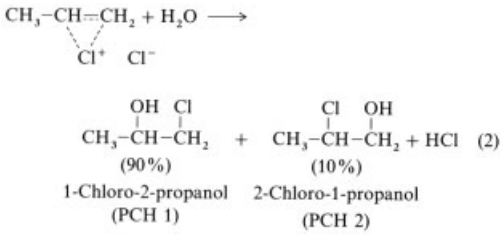

Formation of the chloronium complex is accompanied by two side reactions of chlorine and propene. In the gas phase they react to form the main byproduct, 1,2-dichloropropane (DCP) (Eq. 1.1).

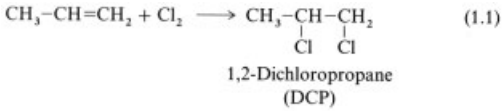

Propene and chlorine also react via 1-chloropropene (allyl chloride) (Eq. 1.2) to give two dichlorohydrin isomers 1,3-dichloro-2-propanol and 2,3-dichloro-1-propanol (DCH 1 and 2, Eq. 1.3).

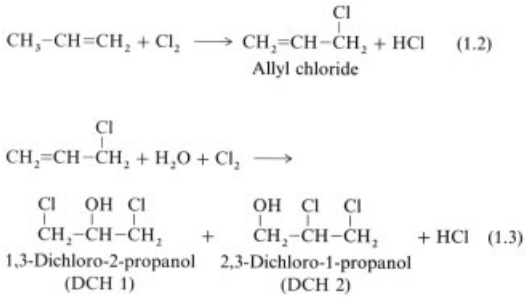

A simplified overview of all propene-chlorine reactions is thus:

Propene + Chlorine	→ Chlorium ion	→ PCH 1 + PCH 2
	→ DCP	
	→ Allyl chloride	→ DCH 1 + DCH 2

In aqueous solution the chloronium ion can also undergo other reactions, leading to formation of DCP with chloride ions alone (Eq. 2.1) and formation of 2,2′-dichlorodiisopropyl ether (DCIPE) in the presence of PCH (Eq. 2.2).

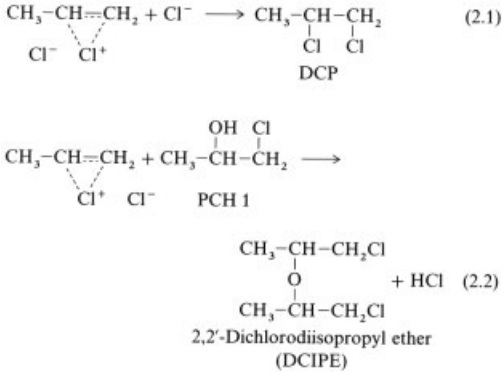

In a simplified overview, the chloronium ion is therefore responsible for the following reactions:

Chlorium ion	→ + H_2O	→ PCH 1 + PCH 2 + HCl
	→ + Cl_-	→ DCP
	→ + PCH	→ DCIPE

Other chlorinated compounds (e.g., monochloroacetone) are produced in very small amounts. Economic considerations require that coproduct formation be reduced as far as possible by using the best available technology. Based on propene, the most expensive raw material for

the chlorohydrin PO process, modern industrial-scale plants achieve the following yields:

PCH	88 – 96 %
DCP	3 – 10 %
DCH	0.3 – 1.2 %
DCIPE	0.2 – 0.8 %
Others	0 – 1.0 %

Besides the design of the reactors, the amount of water entering into the chlorohydrination has a significant impact on coproduct formation, especially DCP and DCIPE. Excess water has two major effects on side reactions. Firstly, it reduces the propylene chlorohydrin and chloride ion concentrations in the reactor, resulting in less DCP and DCIPE. Secondly, excess water prevents the formation of an organic liquid phase in which chlorine and propene react to form more DCP [32].

Dehydrochlorination. In the second reaction step, dehydrochlorination (also called epoxidation/saponification), the aqueous PCH stream is treated with a base – either lime, $Ca(OH)_2$, or caustic soda (as chlorine cell liquor) – to form crude PO and a dilute $CaCl_2$–NaCl brine stream (Eq. 3).

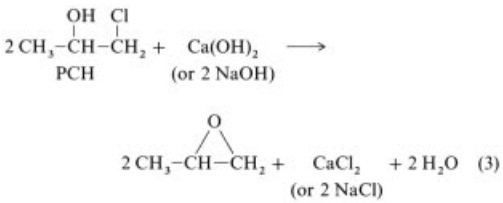

(3)

In a subsequent reaction a minor part of the resulting PO is hydrolyzed to 1,2-propanediol (monopropylene glycol, MPG) (Eq. 3.1).

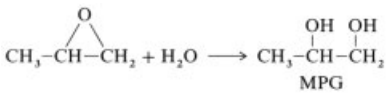

(3.1)

Other byproducts formed in the epoxidation step are 1-chloro-2,3-epoxypropane (epichlorohydrin, Eq. 3.2) and its reaction products 3-chloropropane-1,2-diol (glycerol monochlorohydrin, GMCH), 1,2,3-propanetriol (glycerol), propanal, acetone, and 1-hydroxypropanone (acetol).

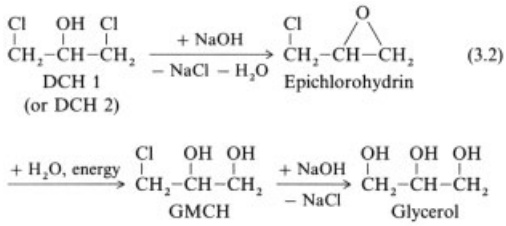

(3.2)

Chlorohydrination. In a typical commercial chlorohydrin PO process, gaseous propene (either chemical grade, ca. 8 % propane, or polymer grade, ca. 0.3 % propane) and gaseous chlorine are mixed in roughly equimolar amounts with an excess of water to form a dilute solution of PCH (ca. 4 wt %), aqueous HCl, and minor amounts of chlorinated hydrocarbon byproducts, mainly DCP.

Since the reactor design (Fig. 4) is based upon propene and chlorine vapor dispersion into water, and because all reaction products remain in aqueous solution, the role of water is very important. It is used not only as reactant (one mole per mole of propene) and direct cooling medium, but also as a diluent to minimize the formation of byproducts. Numerous patents claim technology for boosting chlorohydrin yields to 90 – 95 % [32–38].

Depending on the capacity of the plant, one or several parallel reactors (a) are used. The reaction temperature varies from 45 to 90 °C, depending on the reactor concept. The exothermic reaction itself raises the temperature by 20 – 30 °C. The operating pressure is usually slightly above atmospheric pressure (1.1 – 1.9 bar). The chlorine : propene ratio must be closely controlled to maintain the proper amount and composition of vent gas. In the separator (b) the reactor vent gas (a mixture of propane, excess propene or chlorine, oxygen, nitrogen, hydrogen, carbon dioxide, PCH, DCP, and other reaction products) is separated from the PCH solution and then fed to a scrubber (c). Vent gas from the scrubber can be sent to a thermal oxidizer to protect the environment and recover energy.

Dehydrochlorination (Saponification). The bottoms from the separator (b) are fed to a saponifier (d) where the PCH is mixed with

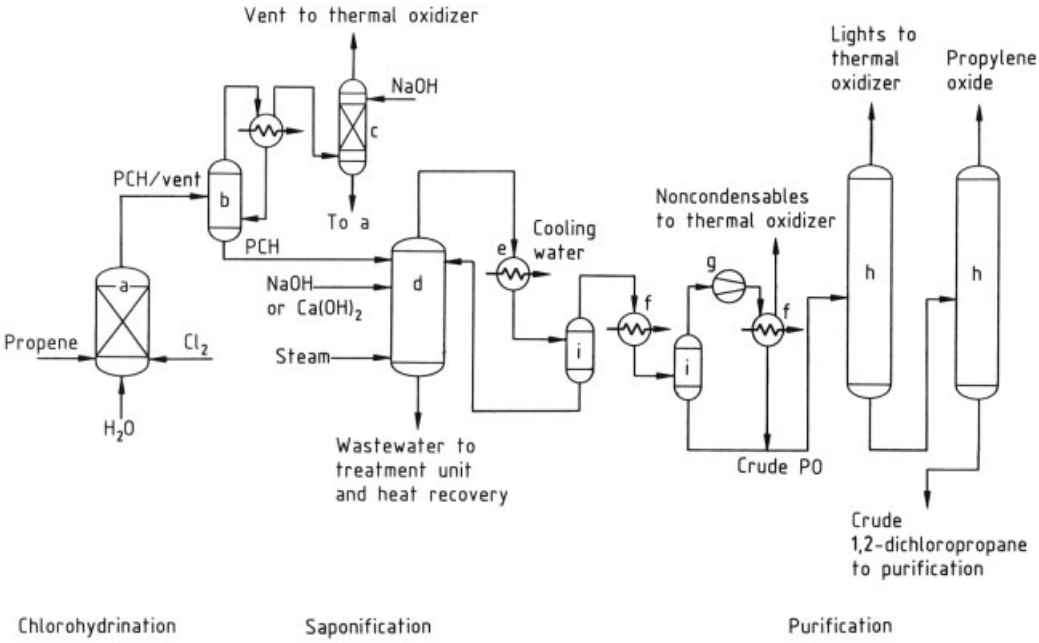

Figure 4. Typical arrangement for a propylene oxide chlorohydrin process
a) Propylene chlorohydrin reactor; b) Separator; c) Vent gas scrubber; d) Saponifier; e) Partial condenser; f) Cross exchanger; g) Compressor; h) Propylene oxide purification train; i) Drums

NaOH or lime. The dehydrochlorination of PCH to PO is rapid. Optimum PCH conversion requires a slight excess of alkalinity. Half the base is consumed by neutralization of HCl produced in the chlorohydrination step.

The saponifier has two main functions: (1) conversion of PCH to PO and (2) stripping of organic impurities from the aqueous brine before final treatment in a biological oxidation plant. Crude PO leaves the saponifier as an overhead stream that also contains organic byproducts and water. If lime is used, solids enter the saponifier, so the column should be equipped with trays to minimize fouling. If caustic soda is used, either trays or packing are acceptable. As far as reaction kinetics are concerned [39], an ideal saponifier would be based on the following design criteria:

1) Short residence time in the top to flash PO as quickly as possible, thereby minimizing the formation of MPG
2) A stripping section to minimize PO losses
3) Long residence time to complete the conversion of DCH 2 to epichlorohydrin (DCH 1 conversion is as fast as the PCH conversion)
4) A stripping section to recover epichlorohydrin

These measures ensure that the formation of volatile products is optimized, and that such compounds are removed before subsequent products (e.g., GMCH, glycerol) are formed and lost in the diluted brine. An overall unit-ratio balance is shown in Figure 5.

Purification. After condensation (e – g) the crude PO is sent to distillation towers (h), in the first of which the lights are removed. Bottoms are sent to the finishing column, where PO product is taken off overhead. DCP, DCIPE, epichlorohydrin, and water remain in the bottom for further treatment which depends on the final utilization of coproducts.

Construction Materials. Since chlorine in aqueous solution as well as generated HCl and PCH are extremely corrosive, the construction materials must be selected very carefully[40–42]. The reactor (a), including the last segments of feed lines, separator (b), vent condensers, scrubber (c), saponifier (d), and all related piping are made of corrosion-resistant materials.

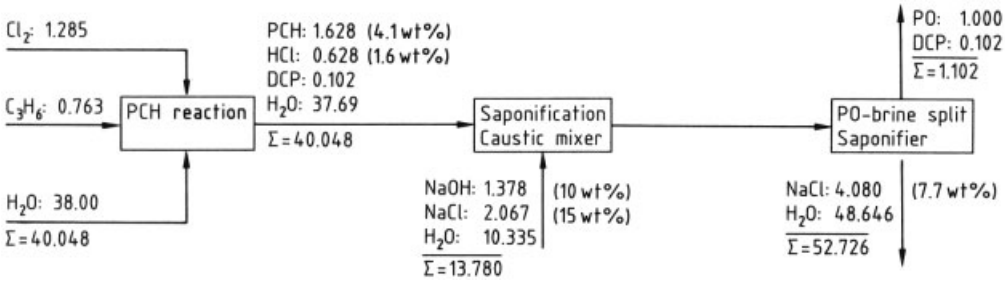

Figure 5. Unit ratios for a propylene oxide chlorohydrin process assuming a water – propene ratio of 50 : 1, caustic soda as cell liquor, DCP as the only byproduct, a DCP yield of 5 %, and all steps stoichiometric.

Adequate for the purpose are poly(vinylidene fluoride) fiber reinforced plastics, graphite, rubber- or brick-lined carbon steel, titanium and titanium – palladium alloys; Inconel, Monel, and Hastelloy.

Wastewater. The brine effluents of modern PO plants contain:

1) Ca. 3 – 5 wt % calcium chloride (lime) or ca. 8 wt % sodium chloride (NaOH)
2) Slight excess alkalinity
3) 100 – 400 ppm total organic carbon (TOC)
4) 30 – 60 ppm adsorbable organic halides (AOX)

The main components are MPG and GMCH–glycerol, which are biodegradable. The amount of brine effluent is 30 – 60 kg/kg PO. Downstream final biological treatment systems achieve an AOX reduction of > 85 % (see Chap. 5 for wastewater treatment).

Lime versus Caustic Soda as Alkalinity Source. The chlorohydrin PO process can be operated with caustic soda or slaked lime. Slaked lime is primarily used when chlorine is produced by the mercury or membrane processes (→ Chlorine, Chap. 5.,→ Chlorine, Chap. 7.) to avoid "downgrading" high-purity caustic (Germany, Japan, former Czechoslovakia, and Poland). Caustic is usually used in the form of cell effluent in combination with the diaphragm chlor-alkali process (→ Chlorine, Chap. 6.) since the low-concentration catholyte from such a chlorine cell can be employed directly (United States, Germany). Although this process option in principle allows the spent PO brine to be recycled to the chlorine plant after resaturation, such a recycle system has never been applied at full technical scale. Major hurdles not yet overcome include hydraulic imbalances, as well as sensitivity of the chlor-alkali process to organic impurities. If an integrated diaphragm chlor-alkali site becomes short of NaOH, it is in principle possible to convert the PO process from cell effluent to lime.

Table 3. Comparison of use of lime and sodium hydroxide in the chlorohydrin process [a]

Lime	NaOH (chlor-alkali/PO integrated complex)
+ Frees NaOH for sales (1.4 t/t PO)	+ no caustic evaporator capacity needed ($NaOH/Cl_2 = 1.08$)
− Higher capital investment; additional caustic evaporation capacity, handling, and infrastructure are needed; unloading of CaO, storage, slaker, thickeners, clarifiers, neutralizer, and centrifuges/filter are required	+ lower capital investment and operating cost (no solids handling required)
+ Up to 50 % less discharge of chloride (NaOH as cell liquid containing up to 15 wt % NaCl, wastewater with 5 wt % calcium chloride)	− higher discharge of chloride (wastewater with 8 wt % sodium chloride)
	+/− brine recycle theoretically possible, ca. 15 % more wastewater
− Disposal of separated solids, uncontaminated and contaminated with organics depending on CaO quality	
− More excess alkalinity in wastewater	
− Lime biox technology, dedicated or diluted (problems with $CaCO_3/CaSO_4$ precipitation)	

[a] − denotes disadvantage, + denotes advantage.

A final detailed evaluation of the use of lime relative to caustic soda cannot be made in this article because the parameters to be considered are too diverse, and they are subject to frequent change (e.g., the global alkalinity sup-

ply/demand situation). However, the key advantages and disadvantages are summarized in Table 3 to provide a qualitative overview with respect to economic, environmental, technological, site, and site-integration issues.

Figure 6 provides a block diagram of a PO–lime process.

One very important issue is the quality of the calcium oxide used for the lime process. Normally, it has a purity of 85 – 95 wt %. The MgO and SiO_2 contents should be kept as low as possible to minimize solids handling.

4.1.2. Other Chlorohydrin Technologies

Modified chlorohydrin technologies have been proposed [29], but they are not currently applied on an industrial scale:

1) The formation of propylene chlorohydrin in two process steps using *tert*-butyl hypochlorite as a chloride-free chlorine-transfer agent (Lummus process)
2) Electrochemical production of PO–the formation of propylene chlorohydrin and subsequent dehydrochlorination to the oxide take place in chlor-alkali electrolysis cells
3) Electrochemical production of PO, in which the formation of propylene chlorohydrin and dehydrochlorination to the oxide occur outside chlor-alkali membrane electrolysis cells

All three process modifications are designed to permit complete recycle of PO brine to the chlor-alkali cells at constant or even reduced DCP byproduct levels. Achieving a balance of the requisite countercurrent conditions leads to (1) increased complexity of the unit operations (e.g., large, complex cascade reactors) and/or (2) high operating costs (e.g., energy-intensive distillation due to a low PO concentration in the catholyte). Furthermore, the selectivity with respect to DCP is at best comparable to that obtained with existing chlorohydrin technology. In fact, DCP formation is likely to be increased, and other chlorinated byproducts (e.g., chloroacetone) may be formed at the expense of DCP. Finally, the sensitivity of chlorine cells to organic contaminants in the concentrated recycle brine must be taken into account. The disadvantages of these process modifications are high in terms of both investment and energy costs.

Propylene Chlorohydrin with Recycle (Lummus Process). One chlorohydrin PO process not currently operated on an industrial scale uses *tert*-butyl hypochlorite instead of chlorine. This allows recycling of the major portion of the aqueous process streams. Process water from the hypochlorination can be recycled to the propylene chlorohydrin reactor. Brine, produced during saponification, can be appropriately treated and recycled to the chlorine cells. Resaturation is not required.

This integrated chlor-alkali/PO process is balanced on water; i.e., kg water in = kg water out. The level of organic byproducts is comparable to that of standard chlorohydrin technology. Lummus has several patents that disclose the technology of such a closed-loop process [43–45].

tert-Butyl hypochlorite is prepared from cell effluent, chlorine, and *tert*-butanol (Eq. 4). The key difference relative to established chlorohydrin processes results from the fact that *tert*-butyl hypochlorite is insoluble in water. It therefore separates as an organic phase from the aqueous brine solution and can be isolated by means of a phase separator. Consequently, propene can be hypochlorinated with *tert*-butyl hypochlorite in an aqueous system, thereby avoiding the presence of chloride ions (Eq. 5).

$$\mathrm{Cl_2 + NaOH + \textit{tert}\text{-}C_4H_9OH} \rightarrow \mathrm{\textit{tert}\text{-}C_4H_9OCl + NaCl + H_2O} \tag{4}$$

$$\mathrm{\textit{tert}\text{-}C_4H_9OCl + H_2O + C_3H_6} \rightarrow \mathrm{C_3H_6OHCl + \textit{tert}\text{-}C_4H_9OH} \tag{5}$$

$$\mathrm{C_3H_6OHCl + NaOH \rightarrow C_3H_6O + NaCl + H_2O} \tag{6}$$

No contact occurs between the chloronium complex and chloride ions (see page 5), so the ratio of water to propene in the process can be reduced without a concurrent increase in the level of byproducts. Furthermore it is possible to recycle a major portion of the nearly chloride-free process water to the front end of the chlorohydrin reactor. Because of the lower water consumption, the two brine solutions–one resulting from *tert*-butyl hypochlorite formation step (Eq. 4) and the other from the hydrolysis of propylene chlorohydrin to PO (Eq. 6)–are close to saturation. They

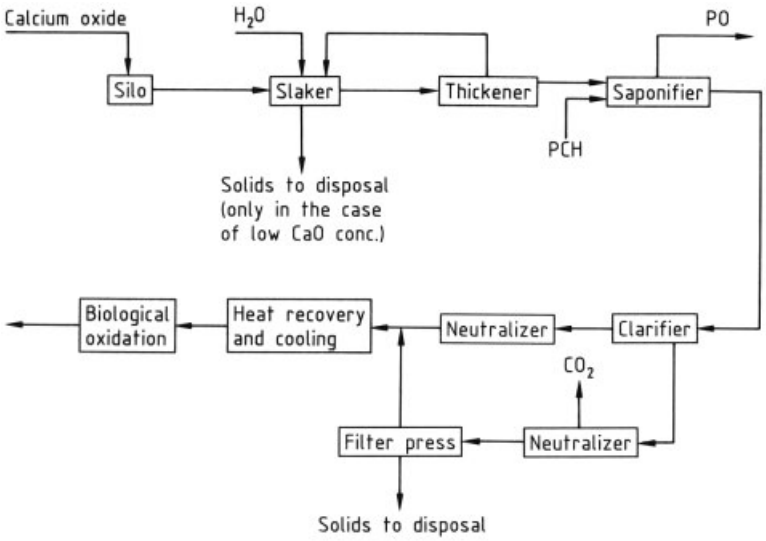

Figure 6. Block diagram for the PO/lime process

can be recycled to a chlorine cell after removal of organic impurities by chlorinolysis.

By contrast, in the conventional chlorohydrin PO process it is essential to minimize reaction between the chloronium complex and chloride ions, because this leads to the formation of DCP. The high ratio of water to propene required to suppress this undesirable side reaction means that the aqueous brine streams have salt concentrations too low to permit complete recycle to the chlorine cells (see Section 4.1.1). A major disadvantage of the Lummus process is the slowness of the reaction between *tert*-butyl hypochlorite and propene. Huge, titanium-clad cascade reactors are required. The presence of an additional component (*tert*-butyl alcohol) also leads to a more complex byproduct spectrum. The treatment of recycle streams is very energy intensive, and investment costs for the process are high.

4.2. Indirect Oxidation Routes

Several reaction routes and process alternatives for the production of PO by indirect oxidation have been reported in the patent literature, but only a few are practiced on an industrial scale. Indirect oxidation is a two-step process:

1) Formation of hydrogen peroxide or an organic peroxide from a suitable alkane, aldehyde, or acid
2) Conversion of the peroxide to water or the corresponding alcohol or acid by epoxidation of propene to PO

Epoxidation of Propene with Hydrogen Peroxide. The epoxidation of propene can be accomplished with hydrogen peroxide in the liquid phase and in the presence of, for example, a molybdenum-based catalyst. Subsequent reactions can lead to formation of considerable amounts of glycols and glycol esters, and thus to separation problems [46–53]. One process modification, published by Enichem [54], makes use of a titanium silicate catalyst in a methanol – water mixture at 40 – 50 °C and a propene pressure of 4 bar. PO selectivities of 75 – 97 % are reported. Disadvantages of this process are the considerable hazard potential associated with concentrated H_2O_2 solutions and high H_2O_2 production costs.

Epoxidation of Propene with Peroxycarboxylic Acids. The required peroxy acids can be prepared by different routes. For example, peracetic acid can be formed by oxidation of acetaldehyde with air or oxygen, or by reaction of acetic acid with H_2O_2. See also →Peroxy Compounds, Organic, Chap. 4.2.

Organic peracids such as performic acid, perphthalic acid, permaleic acid, perpropionic acid,

and peracetic acid react with propene to give PO and the corresponding carboxylic acid. The rate of epoxidation depends on the type of solvent used. Peracetic and perpropionic acids are the preferred epoxidizing agents, and no catalyst is required.

If the starting material for production of peroxyacetic acid is acetaldehyde, the coproduct acetic acid must be removed from the production loop. Process economics depend strongly on the demand for acetic acid and its value to industry. Process experience with a 12 500 t/a plant is available at Daicel; the yield of PO is 90 %. Typically 1.3 kg of acetic acid is produced per kg of PO [55–69]. Scaleup of this technology to industrial capacities ($\geq$ 100 000 t/a) is expensive and difficult, and the value of acetic acid is not sufficiently high to justify the process on a large scale.

The Bayer – Degussa process starts from acetic or propionic acid and H_2O_2, and proceeds via peracetic or perpropionic acid, respectively. The acid coproduct can be recycled to the front end of the process. PO selectivities $\geq$ 95 % can be reached.

A common disadvantage of all processes based on H_2O_2 is that safety considerations require H_2O_2 production facilities to be restricted in size. Economic large-scale production of H_2O_2 for the production of PO is not possible with today's processes.

Coproduct Processes. Oxidation of propene with organic hydroperoxides in the presence of catalysts has developed into a successful indirect oxidation route to PO. Half the world PO production is now based on this route [70–101]. Various organic hydroperoxides can be used (Table 4).

Organic hydroperoxides are generated by catalytic oxidation of appropriate substrates with air or oxygen. Because of coproduct value and market demand, the *tert*-butyl hydroperoxide and ethylbenzene hydroperoxide routes are those practiced in industry.

The production of PO by indirect oxidation with organic hydroperoxides requires two steps. The *tert*-butyl alcohol or α-phenylethanol coproducts formed in parallel with PO are of considerable economic value, because they can be converted to methyl *tert*-butyl ether [*1634-04-4*] (MTBE) or styrene [*100-42-5*], respectively, in subsequent reaction steps. When styrene or MTBE is in great demand, coproduct process economics are competitive with those of the alternative chlorohydrin production routes. The coproducts are always formed in larger amounts than PO itself. In the case of the process with *tert*-butyl hydroperoxide, 2.5 – 3.5 kg of *tert*-butyl alcohol are formed per kilogram of PO. The ethylbenzene process produces 2.2 – 2.5 kg of styrene per kilogram of PO. Obviously, the presence of these coproducts can also become a disadvantage if the demands for PO and the respective coproduct are not properly balanced, requiring that PO production capacity be restricted as demand for the coproduct weakens. No such considerations apply to the chlorohydrin processes. A regional limitation of the coproduct processes results from raw material logistics. Especially for the *tert*-butyl hydroperoxide process, economic operation requires integration into a refinery complex, where mixed butanes and ethylbenzene are readily available.

Process with tert-Butyl Hydroperoxide. The main reactions in this process are:

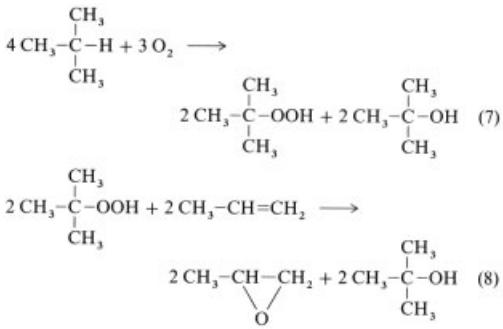

Figure 7 shows a simplified flow scheme for the PO – *tert*-butyl alcohol process. Isobutane is generated by isomerization of a mixed butane fraction. This is then converted with oxygen at 120 – 140 °C and 25 – 35 bar in an epoxidation reactor to *tert*-butyl hydroperoxide, which partially decomposes to *tert*-butyl alcohol [28]. This reaction takes place in the liquid phase and does not require a catalyst; however, additives can be used to optimize reaction conditions. Isobutane conversion is ca. 48 % at a selectivity of 50 % relative to hydroperoxide and 46 % relative to *tert*-butyl alcohol.

The crude peroxide solution is used for the epoxidation of propene. It consists of a mixture

Table 4. Reaction systems for PO – coproduct synthesis [70]

Raw material	Peroxide	Intermediate	Coproduct
Acetaldehyde	peracetic acid		acetic acid
2-Propanol	peracetic acid	acetone	2-propanol
Isobutane	*tert*-butyl hydroperoxide	*tert*-butyl alcohol	isobutene
Isopentane	*tert*-pentyl hydroperoxide	*tert*-pentyl alcohol	isoprene
Ethylbenzene	ethylbenzene hydroperoxide	α-phenylethanol	styrene
Cumene	cumene hydroperoxide	dimethylphenylmethanol	α-methylstyrene
Cyclohexene	cyclohexene peroxide	cyclohexanol	cyclohexanone

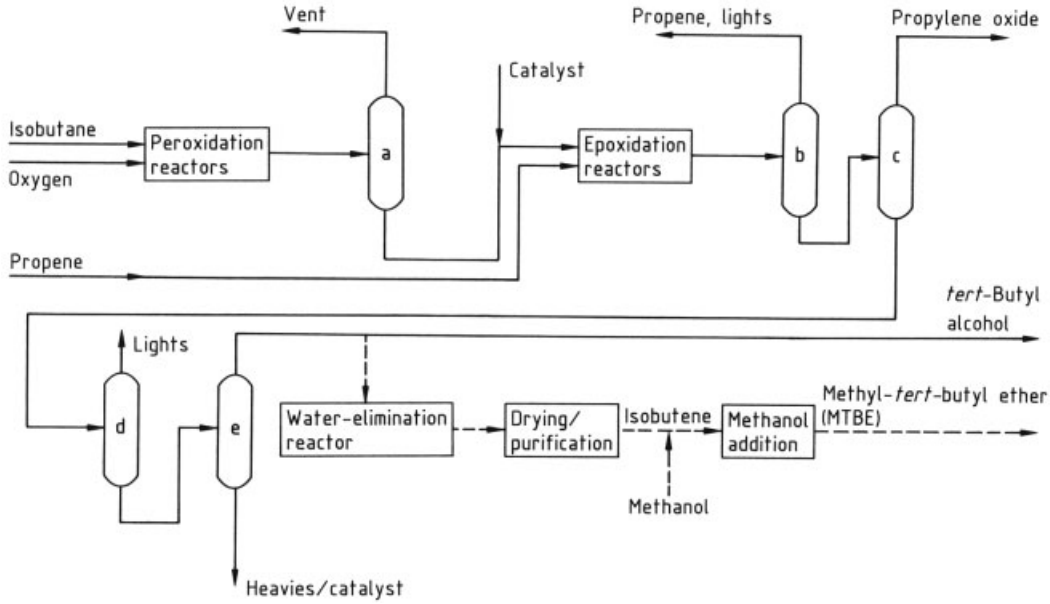

Figure 7. Flow scheme for PO – *tert* butyl alcohol
a) Vent column; b) Lights scrubber; c) PO column; d) *tert*-butyl alcohol lights column; e) *tert*-butyl alcohol column

of *tert*-butyl hydroperoxide, *tert*-butyl alcohol, and smaller amounts of aldehydes and ketones. This is mixed with a homogeneous catalyst, dissolved in toluene; propene is added, and the mixture is fed to the epoxidation reactor system. The preferred catalyst is a molybdenum salt. Epoxidation is accomplished in a staged reaction system consisting of up to five consecutive reactors. The liquid-phase reaction is carried out at 110 °C at ca. 40 bar in the first part of the reactor system. In the second part the temperature is elevated to 120 °C to complete conversion of the hydroperoxide. The selectivity of the converted peroxide to PO is ca. 80 % with a 10-fold excess of propene; i.e., not all the peroxide reacts to form PO. It may decompose without transferring oxygen to propene. Propene conversion reaches ca. 9 % per pass, of which 98 % is converted to PO. Crude PO is then distilled from the reaction mixture and finished. The bottoms of this distillation consist of a mixture of *tert*-butyl alcohol, catalyst, and medium- and high-boiling components.

The crude PO contains hydrocarbons, carbonyl compounds, and ethylene oxide as impurities, which must be removed by rectification and extractive distillation. For further purification of the carbonyl-free PO rectification with *n*-octane is suggested; the overhead product is pure PO. The hydrocarbons present in the crude PO are successively removed from the *n*-octane, which is then reused.

The *tert*-butyl alcohol can be dehydrated to isobutene and then converted to methyl *tert*-butyl ether (MTBE), which is of significant use as a fuel additive for lead-free gasoline. See also

→ Methyl Tert-Butyl Ether. Educt – product ratios are shown below:

Educts:	2.90 kg isobutane
	0.88 kg propene
Products:	2.98 kg *tert*-butyl alcohol
	0.15 kg acetone
	1.00 kg PO

Process with Ethylbenzene Hydroperoxide. The main reactions of this process are:

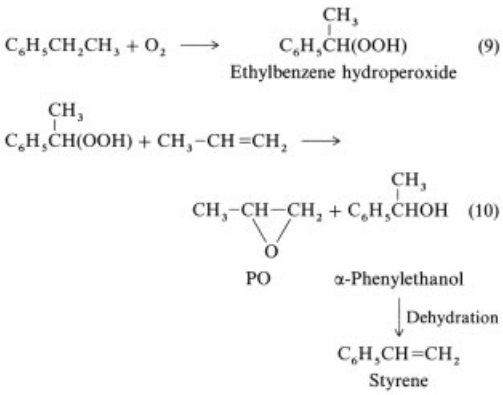

Figure 8 shows a simplified flow sheet of the process, based on patent data. A more detailed flow scheme has been published by Chem Systems Inc. [5].

In the first process step ethylbenzene is reacted with air at 146 °C and 2 bar to give a 12 – 14 wt % solution of ethylbenzene hydroperoxide in ethylbenzene. The reaction takes place in the liquid phase. For safety reasons conversion is restricted to 10 %. The reactor configuration suggested in [102] consists of up to nine separate reaction zones, over which the temperature is reduced from 146 °C to 132 °C by controlled recycle gas addition to avoid decomposition of the peroxide. The resulting peroxide solution in ethylbenzene is mixed with a homogeneous molybdenum catalyst (Halcon) and propene [103]. Alternatively, a heterogeneous titanium-based catalyst can be used (Shell). The latter offers a similar performance, but may be easier to separate from the product stream [104].

For the Halcon route the reaction mixture consists of 100 parts peroxide solution, 2 parts molybdenum naphthenate, and 30 parts propene. Epoxidation of propene is carried out at 100 °C and 35 bar. The crude product stream contains PO, α-phenylethanol, acetophenone, propene, and other organic impurities. Catalyst removal requires a wash with aqueous alkali and phase separation. Propene can be recycled after distillation. Crude PO is purified by distillation. The finishing is similar to that in the *tert*-butyl hydroperoxide route. Ethylbenzene is recovered for recycle from the bottoms of the PO distillation. The remaining α-phenylethanol-rich mixture, which still contains up to 15 % acetophenone, is fed to a dehydration reactor and mixed with an alumina catalyst and triphenylmethane as process solvent. Thermal treatment at 270 °C and 0.35 bar leads to dehydration. Styrene is obtained with a selectivity of 98 %.

α-Phenylethanol and acetophenone are the main byproducts. Educt – product ratios for this process have been reported by Chem Systems Inc. [5]:

Educts:	2.11 kg benzene
	0.76 kg ethylene
	0.88 kg propene
Products:	2.54 kg styrene
	1.00 kg PO

4.3. Direct Oxidation Routes

There is no direct oxidation process for producing PO that is close to commercialization. Difficulties in controlling temperature variations over fixed catalyst beds are among the several hurdles preventing satisfactory selectivity for PO. Nevertheless, construction of a pilot plant has been announced for a new direct oxidation process involving a molten nitrate salt as catalyst [105]. An advantage of this particular direct oxidation variant may be the potential for maintaining a constant reactor temperature (454 °C) in the molten salt bed. However, in addition to PO, several byproducts arise in substantial amounts, including methanol, acetaldehyde, carbon dioxide, ethylene, and formaldehyde.

The breakthrough in direct oxidation of propene to PO has not yet been achieved. For commercial PO production during the next ten years, the only alternatives are the chlorohydrin process or coproduct technologies.

4.4. Comparison of the Various Propylene Oxide Technologies

Direct comparison of the chlorohydrin PO process with the PO/*tert*-butyl alcohol or

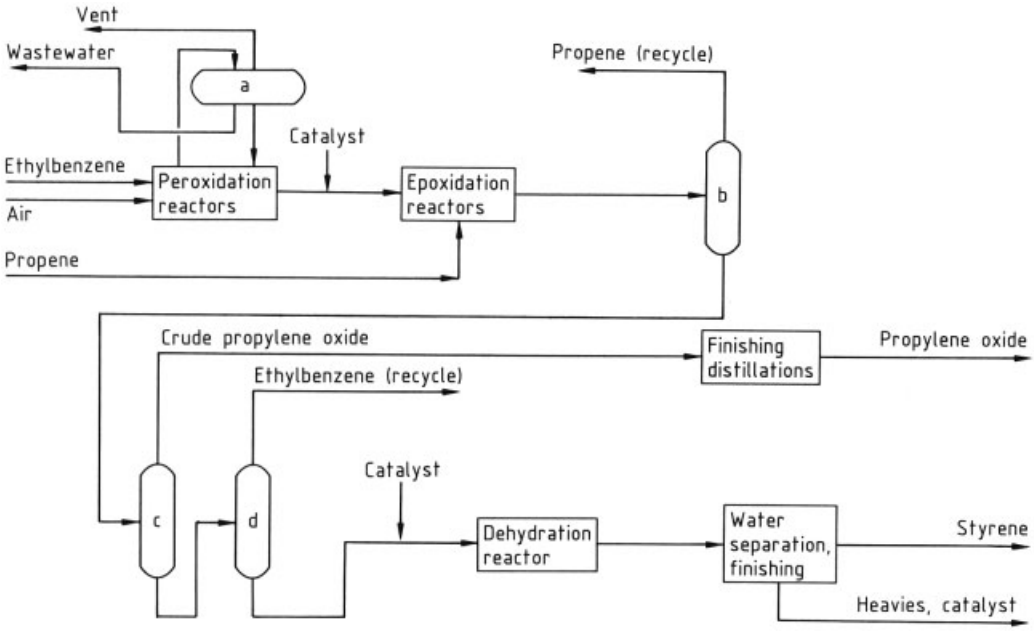

Figure 8. Flow scheme for PO – styrene process
a) Separator; b) Recycle column; c) Crude PO column; d) Ethylbenzene recycle column

PO/styrene coproduct process is very difficult, as the technologies are too diverse and have different objectives. For a PO producer with no market for the coproduct it is obvious that there is currently no alternative to the chlorohydrin PO process. A producer able to include MTBE or styrene in the product portfolio must decide whether a coproduct process or two independent plants represents the preferred solution with respect to economics, demand balances, and geographic flexibility. The fact that stand-alone units for chlorohydrin PO, styrene, and MTBE are still being announced and built indicates that the PO coproduct process is not always economically superior. A comparative economic model of chlorohydrin versus coproduct PO processes is presented in [5]. A decision in favor of one PO production route or the other also depends on the infrastructure, location, and integration of a particular site. Thus, a chlorohydrin PO process can be operated most economically if it is integrated within a chlor-alkali plant. The MTBE process requires integration into a refinery complex. The effluent load of a chlorohydrin PO plant is larger than that of a coproduct plant, so extensive water treatment facilities are required, and the salt load means that an environmentally acceptable geographical location (close to the sea) is desirable. On the other hand, the fire and explosion risk with a PO coproduct plant is significantly greater than in the case of a chlorohydrin facility, and product flexibility is considerably more limited with a coproduct plant than with a stand-alone process.

5. Environmental Protection and Ecotoxicology

Propylene oxide is recognized as a toxic substance (Chap. 10), so specific measures must be taken to avoid environmental pollution during PO production. Critical issues associated with water discharge, air emissions, and handling of PO are discussed briefly below.

Wastewater. State-of-the-art technology for wastewater treatment requires that chlorinated organic compounds first be separated via rectification, stripping, or extraction. The stream is then sent to a biological oxidation unit for eliminating biodegradable compounds like propylene glycol.

In the past, large concrete ponds were often used as aeration basins, but more and more companies are installing biotowers. Significant advantages of such towers include:

1) Higher efficiency
2) Greatly reduced space requirements
3) Less danger of ground water contamination
4) Easier collection and treatment of the vent

The most critical parameters related to water discharge are:

1) Salt discharge at "upstream sweetwater" sites (chlorohydrin process only)
2) Total organic carbon (TOC)
3) Chemical oxygen demand (COD)
4) Adsorbable organic halides (AOX)
5) Temperature
6) Solids disposal (separated as filter cake)

Off-Gas. Environmental legislation varies considerably from country to country, one of the strictest being that in Germany. Discharge to the atmosphere of vent streams with a mass flow $\geq$25 g PO/h is permitted in Germany for PO concentrations up to 5 mg/m^3 [106], [107]. Normally, however, such streams are incinerated with state-of-the-art thermal oxidizers.

For a chlorohydrin PO process, thermal oxidizer treatment is required for the reactor vent, noncondensables in the crude PO, lights from the purification, and tank vent streams. Fugitive emissions represent the only remaining emission source.

In the coproduct processes it is advisable to use oxygen instead of air, since this reduces the volume of vent gas subject to treatment.

Fugitive Emissions. Fugitive emissions are caused mainly by: leaking mechanical seals of pumps, gaskets, and stuffing boxes of valves.

PO is not known to occur naturally. It has been suggested that, apart from PO production facilities, PO may be introduced into the atmosphere from combustion exhausts of sources that burn hydrocarbons [108], [109]. In the atmosphere PO reacts with photochemically produced hydroxyl radicals and has an estimated half-life of 19.3 d. PO is not expected to contribute to ozone depletion [110], [111]. In water, PO is hydrolyzed to propylene glycol with an estimated half-life of 11.6 d at 25 °C and pH 7. The chloride ions in salt water accelerate the chemical degradation to a half-life of 4.1 d [109]. PO is practically nontoxic to fish on a static acute basis (LC_{50} >100 mg/L) [112–116]. Biodegradation of PO under aerobic static laboratory conditions is high (BOD_{20}/ThOD = 74.7 %, where ThOD means theoretical oxygen demand). The biochemical oxygen demand (BOD) of PO is 0.35 g O_2/g PO and 1.65 g O_2/g PO for 5 d and 20 d, respectively [107].

PO evaporates relatively rapidly from dry soil surfaces, and it is moderately volatile from water and wet soils, but the evaporation rate is diminished by leaching. PO is expected to be very mobile in soil. In moist soils hydrolysis is the most significant degradation process. No appreciable concentrations of PO are expected to be found in the environment or aquatic organisms [118].

6. Quality Specifications and Analysis

PO obtained from different processes (Chap. 4) displays consistently high purity (>99.9 %; see Table 5). The quality of the propene feed contributes significantly to the impurity content in the final product. Polymer-grade propene is preferable to reduce formation of acetaldehyde and ethylene oxide. Acetaldehyde causes problems in polyol applications, whereas ethylene oxide leads to undesirable contaminants in propylene glycol.

Other impurities may also be present depending on the nature of the production process (e.g., chlorinated organic compounds, hydrocarbons, ethers, aldehydes, acetone, methanol, butanol, *tert*-butyl alcohol, or furan).

Quality specifications are becoming increasingly strict, and customer satisfaction can best be assured by instituting statistical quality tools.

PO and its process-specific coproduct spectrum (e.g., aldehydes, chlorinated organic compounds, methanol, acetone, hydrocarbons, and ethers) can be determined by gas chromatography. Water can be analyzed by the Karl-Fischer method (DIN 51 777, ASTM D 203). The platinum – cobalt standard is used as a color reference (DIN 53 409, ASTM D 1209).

Table 5. Typical specifications for PO in urethane applications

Parameter	Chlorohydrin process	PO/*tert*-butyl alcohol process	PO/styrene process
Purity, %	99.9	99.9	99.9
Color (APHA)	10	10	10
Water, mg/kg	100	100	100
Aldehydes, mg/kg	30	30	30
Total chlorides, mg/kg	30	30	30
Methanol, mg/kg	not present	50	not present
Furan, mg/kg	not present	not present	50

Processes are monitored with on-line gas chromatographs and, more recently, with IR systems.

7. Handling, Storage, and Transportation

Detailed attention is required on issues of safety, health, and the environment. Product stewardship policies provide the basic tools for managing these vital business activities. For example, in 1988, the U.S. Chemical Manufacturers Association (CMA) adopted a "Responsible Care" program; details are provided in [119], [120]. Based on these principles, many companies have developed new product stewardship policies; e.g., [121].

Handling. Safety precautions related to the handling of PO (e.g., protective clothing, sampling, vessel entry, loading, unloading, and general issues) are described in detail in data sheets provided by the manufacturers [122], [123] and elsewhere [124]. Therefore, only the most important issues are discussed in this article.

Personnel handling PO must receive appropriate training. PO reacts readily with all compounds containing active hydrogen. Contact with materials such as acids, bases, oxidizing materials, and clay-based absorbents must be strictly avoided.

Spills, Leaks, and Fire. In case of a spill, leak, or fire the following measures should be taken:

1) Shut off all potential ignition sources.
2) Dilute liquid spills with copious amounts of water (H_2O : PO ratios $\geq$200 : 1). Use water spray to reduce the extent of vapor. The effectiveness of such a spray can be checked by testing the area with an explosimeter.
3) Maintain the water spray, and shut off the flow of PO to the leak.
4) Do not shut off the water until it is certain that the leak has stopped and that flammable concentrations have dissipated.
5) Prevent runoff from entering sewers and/or natural waters.
6) Adequate fire extinguishing media include water, alcohol foam, CO_2, and dry powder.
7) Small fires (e.g., at a pump-seal leak) can be extinguished with a fire hose, deluge gun, CO_2, or dry powder.
8) Large fires (e.g., from a blown gasket) present a different and more serious problem. Do not attempt to extinguish the external fire, or to shut off the feed to the fire. Instead, change the nature of the feed to the fire. Start filling the container with water or nitrogen and continue to fill at a rate adequate to maintain pressure on the leak. With sufficient water or nitrogen dilution, fuel to the fire will be cut off from the inside, or the fire will be extinguished as a result of creating a nonflammable mixture. Water should be sprayed on all adjacent oxide- or flammable-containing equipment to keep it cool. Once the fire is out, the entire area should be sprayed with water to reduce the chance of reflash.

Storage. Storage tanks for PO are usually made of ASTM A-516 grade steel or stainless steel. Because of the tendency of PO to polymerize, all equipment should be absolutely free of impurities. Therefore, PO tanks are constructed in such a manner that they can be completely drained and cleaned.

Tanks and piping systems must be purged with nitrogen to reduce their oxygen content to < 0.5 vol % [automatic N_2 pad and depad system (N_2 inertization)]. Pure N_2 should be used (max. 100 ppm O_2).

Fragile devices (e.g., glass or plastic sight and gauge glasses as well as expansion joints and flexible or screwed connections) should be avoided in PO service. Teflon spiral-wound stainless steel gaskets are preferred. Flat graphite and/or EPR rubber gaskets are also acceptable.

All equipment containing liquid PO should be protected with automatic deluge or sprinkler systems. The maximum storage temperature should be 35 °C.

Tanks and facilities for loading and unloading PO should be equipped with spill-retention walls or dikes (remotely drained) to direct spills into containment areas. Such catch basins should suffice for 110 % of the tank capacity, or be large enough to accomodate an adequate volume of deluge water.

The use of insulated tanks should be avoided because experience shows that corrosion underneath the insulation creates the potential for a spill. Other relevant safety considerations (e.g., minimum instrumentation) are described in [122], [123].

Transportation. PO is transported in gravity or gas ships, railcars, tank trucks, and containers. Hazard classifications and transport information are listed in Table 6.

Emergency response organizations, e.g., the "TUIS" (Transport Unfall Informationssystem) in Germany, have been established to provide local authorities with adequate information and help when requested. Special emergency response and transportation equipment data sheets (Tremcards) are available [125].

The European Chemical Industry Council (CEFIC) has approved a project called "International Chemical Environment" (ICE) in which emergency response will be further improved through close cooperation between national chemical federations, fire brigades, local authorities, and national emergency response centers. The goal is to provide professional assistance all over Europe in case of distribution incidents involving chemicals.

In order to achieve maximum conveyance safety the following minimum requirements for transportation equipment are recommended (Dow Chemical).

Sea Barges and Gravity Gas Vessels. For detailed information, see [126], [127]:

1) If a vessel under consideration is in nondedicated service, tanks and associated piping should be effectively cleaned to remove all traces of previous cargos and rust before PO is loaded. All tanks must be entered for final inspection.
2) Documentation must be provided for the three most recent cargos. No materials known to catalyze reactions of PO (e.g., acids, alkalis, amines, or ammonia) are permitted among the previous three cargos.
3) Independent vapor return should be used during loading and unloading.
4) Cargo tanks (on gravity ships, for example) with a design pressure of less than 0.6 bar (gauge) should be equipped with a cooling system to maintain PO far below its boiling point. Use of such a cooling unit is not required in colder regions of the world and on journeys of less than seven days.
5) Piping systems for tanks that are to be loaded with PO should be completely segregated from piping systems for all other tanks (removal of spool pieces and installation of blind flanges).
6) For other important points see the Section above on Storage.

Railcars. Design according to RID (Europe), AAR (USA), and GGVE (Germany). Top-loading/unloading nozzles are preferred, protected by a thick (2.5 cm) heavy-duty dome in the event of a roll over.

Tank Trucks. Design according to ADR (Europe), DOT MC 330/331 (USA), GGVS (Germany). Valves should be located at one side and in front of the rear wheels. Heavy-duty protection is recommended for the bottom line and valves.

Containers. Type 5 vessels (liquefied gas under pressure) are recommended for PO service.

Table 6. Hazard classifications and transport information for propylene oxide [a]

Transport mode	Codes according to legal requirements				Danger class	Label
	International	Europe	United States	Germany		
Sea	IMO/IMDG			GGVSee	3.1 UN no. 1280 page no. 3047	3
Barge				ADNR		
Rail		RID [b]	AAR	GGVE	3-2(A)	3
Road		ADR [b]	DOT	GGVS	3-2(A)	3
Air	ICAO/IATA				3 UN no. 1280	LF

[a] Sample shipment not allowed by post.
[b] ADR/RID hazard identification: (Kemler Code) 33-1280 (UN no.).

Valves should be located at the bottom (both sides), protected by an adequate safety system.

Other important criteria for railcars, tank trucks, and containers are that the maximum allowable working pressure should be at least 5 bar, design pressure at least 10 bar. Recommended minimum wall thickness is 6 mm.

8. Uses

PO is an important basic chemical intermediate. Virtually all the PO produced is converted into derivatives, often for applications similar to those of ethylene oxide (EO) derivatives (→Ethylene Oxide, Chap. 9.).

PO is used primarily to produce polyether polyols, propylene glycols, and propylene glycol ethers. The distribution of PO applications (derivatives) in the U.S. market is shown below, and is typical for the world market [128]:

Polyether polyols comprising	
Flexible foams	51 %
Rigid foams	10 %
Nonfoam uses	7 %
Propylene glycols	22 %
Glycol ethers	4 %
Miscellaneous	6 %

Most of the PO produced (65 – 70 %) is used as an intermediate for *polyether polyols* that are mainly consumed in the manufacture of polyurethanes. Polyurethanes can be prepared with a wide range of hardness, rigidity, and density characteristics (→Polyurethanes). Flexible polyurethane foams are important in furniture and automobile seating, bedding, and carpet underlay. They are made from polyols with molecular masses >3000. Polyols with lower molecular masses lead to rigid foams for such applications as thermal insulation. Polyether polyols used in polyurethanes are based on di-, tri-, or polyhydric alcohols. They are usually PO/EO-copolymerized, with a PO content of ca. 70 wt % for rigid and 90 wt % for flexible foams.

Homo- or copolymerized polyethers are also used as surface-active agents in detergents, textiles, defoamers, hair-care preparations, brake fluids, and lubricants. Increasing the PO/EO ratio causes a surfactant to become more hydrophobic. Polypropylene glycols with molecular masses in the range 400 – 4000 are produced by polyaddition of PO and a difunctional alcohol. Polypropylene glycols become less hydrophilic with increasing molecular mass. Nonpolyurethane applications include lubricants for rubber; in metal rolling, drawing, and machining; antifoaming agents; in deicing formulations for gasoline; and in hydraulic fluids.

Propylene glycols constitute the second largest application for PO. Monopropylene glycol (MPG) is the direct reaction product of PO with water (→Propanediols). Di-, tri-, and higher propylene glycols are coproduced by the reaction of MPG with PO. Propylene glycol is consumed mainly as a raw material for unsaturated polyester resins, especially for the textile and construction industries. It is also used as a humectant, and as a solvent and emollient in food, drugs, and cosmetics. Further uses relate to coatings, plasticizers, heat transfer and hydraulic fluids, antifreezes, and aircraft deicing fluid. Applications of propylene glycol have benefitted from its lower toxicity relative to ethylene glycol. Di- and tripropylene glycols are

used in hydraulic and drilling fluids, in cutting oils, and as lubricants. The main advantage of the polyglycols is their biodegradability.

Propylene glycol ethers are formed by the reaction of PO with alcohols, usually methanol, ethanol, propanol, or butanol. Certain traditional ethylene glycol ethers and their acetates are being replaced by their PO-based analogues. The demand for PO-based glycol ethers and acetates is therefore growing rapidly, and now constitutes the third largest market for PO (ca. 4 %). Typical applications include solvent use in coatings, paints, inks, resins, cleaners, waxes, and electronic circuit board lamination. The glycol ethers are also found in heat-transfer fluids and anti-icing agents for jet fuel.

Various *specialty organic compounds* can be derived from PO, including allyl alcohol (glycerol synthesis), propylene carbonate (a specialty solvent for organic and some inorganic compounds), mono-, di-, and triisopropanolamines (detergent raw material), and hydroxypropylated cellulose.

9. Economic Aspects

PO is a raw material for major consumer goods in virtually all industrialized countries, and it occupies third place in the consumption of total propene (ca. 10 %) after polypropylene and acrylonitrile. Because of continuing expansion of its field of application, PO production was forecast to increase (see Table 7).

Until 1969 essentially all the PO was produced by the chlorohydrin process. A plant using a version of the peroxidation process (indirect oxidation) was started by Arco in the United States in 1969. In 1991 ca. 50 % of world PO-production capacity was based on the chlorohydrin process, with the remainder coming from peroxidation (Table 7). Dow Chemical and Arco Chemical are dominant in PO production, representing more than 80 % of world capacity. World production of PO in 1992 was ca. 3.2×10^6 t/a, with an expected annual growth rate of ca. 4 %.

Table 8 provides an overview of PO production capacities. The technical, economic, and environmental aspects of large-scale PO production are discussed in [5], [29].

10. Toxicology and Occupational Health

Extensive experience has shown that PO can be handled safely provided appropriate precautions are taken. However, because of its toxicity, PO should be handled only by persons familiar with the hazards involved and willing to observe appropriate precautions.

In general, the health hazards of handling PO are associated with vapor inhalation and eye and skin contact.

Oral Toxicity. The acute oral toxicity of PO is moderate. LD_{50} values have been reported for rats, mice, and guinea pigs in the range 300 – 1000 mg/kg [129], [130].

Skin Contact. PO has only a low to moderate potential for absorption through the skin in toxic amounts. The dermal LD_{50} for skin absorption of PO in rabbits and guinea pigs is 1240 mg/kg [130] and 7200 mg/kg [131], respectively. PO is not irritating to the skin if allowed to evaporate freely, but contact for several minutes may result in mild local redness [132]. Aqueous solutions are more irritating than undiluted PO. More sustained contact results in the development of chemical burns, blistering, and swelling [133], [134].

Eye Contact. High vapor concentrations of PO are irritating to the eye. Liquid PO has caused serious local injury to the eyes of laboratory rabbits [135]. Cases of corneal burns have been reported in humans [129]. Permanent impairment of vision, even blindness, may result from such contact. Affected eyes should be immediately flushed with flowing water for at least 15 min.

Inhalation. Exposure to PO at several hundred ppm may result in headache, dizziness, drowsiness, nausea, chest discomfort, and cough [134]. Higher concentrations (>1000 ppm) may lead to edema, irritation to the respiratory tract, and loss of consciousness [136], [137]. One case of human poisoning resulted from vapor exposure to 1500 ppm for 10 min. Recovery from this poisoning incident occurred the following day [138].

Inhalation studies with laboratory animals suggest a "no-adverse-effect level" for all

Table 7. Comparison of global PO production capacities, consumption, and process shares

PO production	1986	1991	1996 [a]
Chlorohydrin process, %	58	55	49
PO/styrene process, %	18	19	24
PO/*tert*-butyl alcohol process, %	24	26	24
Direct oxidation, %	0	0	2
World PO capacity, 10^3 t/a	3300	3900	4900
World PO consumption, 10^3 t/a	2600	3250	3900

[a] Forecast.

Table 8. World PO capacities in 1991

Country	Producer	Location	Process [a]	Capacity, 10^3 t/a
America				
Brazil	Dow Quimica	Aratu	CH	150
Canada	Dow	Sarnia, Ont	CH-NaOH	60
United States	Arco Chemical	Channelview, TX	styrene	270
	Arco Chemical	Bayport, TX	TBA	550
	Dow Chemical	Freeport, TX	CH-NaOH	500
	Dow Chemical	Plaquemine, LA	CH-NaOH	200
	Olin	Lake Charles, LA	DO	5
Total				*1735*
Asia				
China	Zhangdian Petrochem.	Zibo/Shandong	CH	10
	Jinling Petrochem. Corp.	Zhongshan	CH	10
	Tianjing Dagu	Tianjing	CH	16
	Gao Qiao	Shanghai	CH	16
	NJ Zhongshan	Guang Dong	CH	16
India	Manali Petrochem.	Baroda, Madras	CH	12
	UB Petroproducts	Manali, Madras	CH	12
Japan	Sumi-Arco	Chiba	styrene	140
	Tokuyama Soda	Tokuyama City	CH	50
	Showa Denko	Kawasaki	CH	30
	Mitsui Toatsu	Nagoya	CH	36
	Asahi Glass Co.	Kashima	CH	100
Korea (South)	Yukong Arco Chemical	Ulsan, Yukong	styrene	100
Taiwan	Chiung Long Petrochem.	Linyuan/Kaohsiung	CH	15
Total				*563*
Europe				
Bulgaria	Neftochim	Burgas	CH	12
Czechoslovakia	Novacke Chemicke Zavody	Novaky	CH	5
France	Arco Chimie France	Fos-sur-Mer	TBA	200
Germany	Dow Deutschland Inc.	Stade	CH-NaOH	420
	Erdoelchemie	Koeln-Worringen	CH	150
	Bunawerke	Schkopau	CH	50
	BASF	Ludwigshafen	CH	90
Italy	Enichem	Priolo, Sicily	CH	60
Netherlands	Arco Chemie Nederland	Botleck	TBA	250
	Shell Netherland Chemie	Moerdijk	styrene	140
Poland	State	Rokita	CH	20
Romania	Oltchim	Rimnicu Vilcea	CH	12
	State	Midia	CH	60
Spain	Repsol Quimica/Arco	Puertollano	styrene	50
Former Soviet Union	State	Sumgait	CH	25
	State	Nizhnekamsk	styrene	50
Yugoslavia	Sodaso	Tuzia	CH	20
Total				*1614*
Total worldwide				*3912*

[a] CH = chlorohydrin PO process (using lime unless stated otherwise); DO = direct oxidation of propene; styrene = styrene/PO process; TBA = *tert*-butyl alcohol/PO process.

species of 150 ppm [129]. No evidence of neurotoxicity was detected in male rats exposed to PO by inhalation for 24 weeks at a level of 300 ppm [17].

At exposure levels ≥4000 ppm PO is considered to have mild central nervous system effects in animals, characterized by ataxia, uncoordination, and general depression [133], [139].

The odor threshold in air is 200 ppm (95 % confidence limit: 114 – 353 ppm) [137], a value too high to prevent accidental overexposure. Nevertheless, the sweet, penetrating odor may be sufficient to signal the danger of acute exposure.

Teratogenicity. PO does not produce teratogenic effects in laboratory animals before or during gestation. In rats, fetotoxicity has been seen only at doses high enough to cause toxic effects to the mother. In rabbits, exposure has not proven substantially toxic either maternally or with respect to the fetus [140].

Mutagenicity. PO directly alkylates proteins and DNA, induces gene mutations in all microorganisms tested, and induces chromosomal aberrations in mammalian cells in vitro [141–147]. In vivo, an increased frequency of micronuclei was induced in mice after intraperitoneal PO exposure, but not after oral administration, even at near-lethal doses. No clear dominant lethal effects were found in rodents or monkeys [142], [148]. No chromosomal abnormalities or effects on the sister chromatid exchanges were observed in monkeys exposed for two years [139], [149], [150].

Carcinogenicity. Lifetime inhalation studies on laboratory rats and mice suggest a weak carcinogenic effect of PO in animals at high dose levels [151–156], but the evidence for carcinogenity in humans is inconclusive [157–159]. Evaluations of toxicity to warm-blooded species are described in [160]. PO has been classified by the U.S. Environmental Protection Agency as a class B-2 substance (probable human carcinogen) [159], and it is listed as a group III carcinogen in the German TA Luft [107].

Permissible Exposure Limits. PO is recognized as a toxic substance by the International Agency for Research on Cancer (IARC), and is classified as a class 2A carcinogen (TLV-TWA = 48 mg/m^3 = 20 ppm) [139]. The German MAK commission has listed PO as a group III A2 carcinogen (TRK = 6 mg/m^3 = 2.5 ppm) (at 20 °C/1013 hPa: 1 ppm = 2.42 mg/m^3).

11. References

General References

1. *Beilstein,* 4th ed., Propylene Oxide.
2. *Kirk Othmer,* 3rd ed., **19,** p. 246 ff.
3. *Winnacker-Küchler,* 4th ed., **2,** p. 565 ff.
4. K. Weissermel, J. J. Arpe: *Industrial Organic Chemistry,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1992.
5. Chem SystemsPERP "Process Evaluation Research Planning," Propylene Oxide, Tarrytown, New York, Nov. 1990.
6. SRI 1985PEP "Process Economics Program" EO/PO Edition 2D. Specific References

Specific References

7. F. L. Ötting: "Low Temperature Heat Capacity and Related Thermodynamic Functions of Propylene Oxide," *Journal Chem. Phys.* **41** (1964) 149 – 153.
8. R. A. McDonald, S. A. Shrader, D. R. Stull: "Vapor Pressures and Freezing Points of 30 Organics," *J. Chem. Eng. Data* **4** (1959) 311 – 323.
9. K. H. Simmrock et al.: "Critical Data of Pure Substances," *DECHEMA Chemistry Data Series,* vol. 2, Frankfurt 1986.
10. C. L. Yaws, M. P. Rackley, *Chem. Eng. (N.Y.)* **83** (1976) 129.
11. Dow Chemical Co., *Alkylene Oxides,* Product Bulletin, Form no. 110-551-77R, Midland, Mich. 1977.
12. K. A. Kobe, A. E. Ravicz, S. P. Vohra, *J. Chem. Eng. Data,* **1** (1956) 50 – 56.
13. R. W. Gallant: "Physical Properties of Hydrocarbons," *Hydrocarbon Process.* **46** (1967) 143 – 150.
14. Dow Chemical Europe, Material Safety Data Sheet, Propylene Oxide, Horgen, Switzerland, 1987.
15. B. G. Witt, M. W. Chase, *Experimental and Calculated Flammability Limits of 30 Chemicals,* NCT-577, 1975.
16. U.S. Coast Guard, Department of Transportation, *CHRIS–Hazardous Chemical Data,* **vol. II,** Washington D.C., U.S. Government Printing Office, 1984 – 1985.

17. The Dow Chemical Company, unpublished data.
18. P. H. Howard et. al.: *Handbook of Environmental Fate & Exposure Data for Organic Chem.,* **vol. I,** Lewis Publishers 1989, pp. 483 – 489.
19. *Ullmann,* 4th ed., **19,** 471 – 481.
20. J. N. Wickert, W. S. Tamplin, R. L. Shank, *Chem. Eng. Prog. Symp. Ser.* **48** (1952) no. 2, 92.
21. G. O. Curme, F. Johnston: *Glycols, ACS Monograph No. 114,* Chap. 11, Reinhold Publishing Corp., New York 1952.
22. J. Gmehling et al., *"Vapor-Liquid Equilibrium Data Collection," DECHEMA Chemistry Data Series,* **vol. 1,** Frankfurt 1977 – 1984.
23. L. H. Horsley et al.: "Azeotropic Data," *Adv. Chem. Ser.* **6** (1952); **35** (1962).
24. Abderhalden, Eichwald, *Ber. Dtsch. Chem. Ges.* **51** (1918) 1312.
25. Schurig et al., *Angew. Chem.* **90** (1978) 993 – 995.
26. M. C. Flowers: "Kinetics of the Thermal Gas-Phase Decomposition of 1,2-Epoxypropane, "*J. Chem. Soc. Faraday Trans. 1* **73** (1977) 1927 – 1935.
27. M. Kwasny, M. Syczewski, *Przem. Chem.* **67** (1988) no. 5, 228 – 231.
28. K. H. Simmrock, *Chem. Ing. Tech.* **48** (1976) no. 12, 1085 – 1096.
29. K. H. Simmrock, *Hydrocarbon Process.* **57** (1978) no. 11, 105 – 113.
30. B. Oser, *Bull. Soc. Chim. Fr.* 1860, 235.
31. R. Elm: "Die Kinetik der Reaktion zwischen gasförmigen Propylen und wässerigen Chlorlösungen," Dissertation, Universität Dortmund 1977.
32. Shell, BE 630 446, 1963.
33. Bayer, FR 1 357 443, 1964.
34. Naphthachimie, DE-OS 2 101 119, 1971 (R. Bouchet).
35. BASF, DE-OS 2 022 819, 1971 (E. Bartholomé).
36. G. Mikula, CS 152 702, 1974; *Chem. Abstr.* **81** (1974) 63 135 r.
37. Petrocarbon Developments, FR 2 194 673, 1974.
38. Mitsui Toatsu Chemicals, JP-Kokai 77 48 606, 1977 (Y. Watanabe et al.).
39. S. Carra, M. Morbidelli, *Chem. Eng. Sci.* **34** (1979) 1123 – 1232.
40. G. N. Kirby, *Chem. Eng.* (1985) Febr. 4, 81 – 83.
41. D. H. DeClerk, A. J. Patarcity, *Chem. Eng.* (1986) Nov. 24, 46 – 63.
42. H. Busse, H. Schindler, *Chem. Ing. Tech.* (1990) no. 4, 271 – 277.
43. Lummus, US 4 496 752, 1985 (A. B. Gelbein et al.).
44. Lummus, DE 3 016 668, 1980 (A. B. Gelbein et al.).
45. Lummus, US 4 008 133, 1977 (A. B. Gelbein et al.).
46. Institut Français du Petrole, DT 1 815 998, 1968 (H. Mimoun et al.).
47. Institut Français du Petrole, FR 2 114 752, 1970.
48. Sumitomo GB 1 302 441, 1969 (J. Watanabe et al.); DT 2 012 049, 1969 (J. Watanabe et al.).
49. Naphtachimie US 3 778 451, 1966 (M. Poite); DE 1 668 500, 1966 (M. Poite).
50. Shell, DE-OS 1 940 205, 1969 (H. Fernholz).
51. Ugine Kuhlmann, DE-OS 2 446 830, 1973 (J. P. Schirmann et al.).
52. UCC, DE-OS 2 607 768, 1975 (Ch. McMullen).
53. Union Oil, US 3 293 269, 1964 (L. G. Wolgemuth).
54. *J. Catal.* **129** (1991) 159 – 167.
55. Institut Français du Petrole, FR 1 549 184, 1964 (H. Mimoun et al.); DT 1 668 312, 1964 (H. Mimoun et al.).
56. Laporte Chem., FR 1 519 147, 1966 (E. C. Kurby et al.); DT 1 618 625, 1966 (E. C. Kurby et al.).
57. Knapsack, GB 1 076 288, 1946 (K. Sennewald et al.); DT 1 240 515, 1964 (K. Sennewald et al.).
58. Celanese, US 3 341 556, 1959 (A. L. Stauzenberger et al.).
59. Petrocarbon Development, DT 1 917 031, 1968 (J. V. Fletscher et al.).
60. Petrocarbon Development, DT 1 925 378, 1968 (J. V. Fletscher et al.).
61. Petrocarbon Development, GB 1 318 524, 1969 (J. V. Fletscher et al.); DT 1 066 920, 1969 (J. V. Fletscher et al.).
62. Daicel, DT 1 923 392, 1968 (K. Jamagishi et al.).
63. Daicel, US 3 663 574, 1968 (K. Jamagishi et al.).
64. Daicel, US 3 838 020, 1971 (O. Kageyama et al.); DT 2 258 521, 1971 (O. Kageyama et al.).
65. Asahi Chem. JP 49 43 926, 1969 (S. Imanura, Y. Wakasa, K. Kotaoka).
66. Metallgesellschaft, DE-OS 2 334 982, 1973 (Th. Simo).
67. Interox, DE-OS 2 502 776, 1977 (A. Hildonk, P. Greenhalgh).

68. Bayer, DE-OS 2 519 297, 1975 (G. Prescher et al.).
69. Olin, NL 7 601 976, 1975.
70. E. J. Mistrik, *Chem. Tech. (Leipzig)* **27,** no. 1, 23 (1975).
71. Sun Oil, GB 1 261 617, 1968.
72. Sun Oil, BE 729 193, 1968 (K. M. Brownstein et al.); DE 1 910 258, 1968 (K. M. Brownstein et al.).
73. Atlantic Richfield, FR 1 500 728, 1966 (H. R. Grane et al.); DT 2 159 605, 1971 (H. R. Grane et al.).
74. Atlantic Richfield, GB 1 345 900, 1971; DT 2 159 605, 1971.
75. Atlantic Richfield, US 3 907 902, 1966; DT 1 643 586, 1966.
76. Halcon Int., NL 6 515 037, 1964.
77. Halcon Int., BE 702 236, 1966 (J. P. Schmidt et al.); DT 1 668 204, 1966 (J. P. Schmidt et al.).
78. Halcon Int., GB 1 308 340, 1969 (St. Herzog); DT 2 017 068, 1969 (St. Herzog).
79. Akzo, DE-OS 2 159 764, 1971 (A. Wegerhoff et al.).
80. Mitsui Petrochem., US 3 592 857, 1967 (Y. Shinohara); DE 1 668 480, 1967 (Y. Shinohara).
81. Atlantic Richfield, GB 1 146 202, 1965 (M. N. Sheng et al.); DT 1 568 002, 1965 (M. N. Sheng et al.).
82. Atlantic Richfield, GB 1 298 253, 1969.
83. Atlantic Richfield, GB 1 339 296, 1969.
84. Halcon Int., BE 665 240, 1964; DT 1 518 996, 1 543 002, 1964.
85. Halcon Int., GB 1 074 330, 1962 (J. Kollar); DT 1 468 012, 1962 (J. Kollar).
86. Halcon Int., GB 1 122 702, 1965.
87. Halcon Int., US 3 360 584, 1965 (J. Kollar); DT 1 568 764, 1965 (J. Kollar).
88. Halcon Int., US 3 360 585, 1965 (Ch. N. Winnick); DT 1 568 763, 1965 (Ch. N. Winnick).
89. Halcon Int., US 3 449 219, 1966 (J. P. Schmidt).
90. Halcon Int., US 3 458 534, 1964 (Ch. N. Winnick et al.); DT 1 543 029, 1964 (Ch. N. Winnick).
91. Halcon Int., US 3 459 810, 1965 (Ch. Y. Choo et al.); DT 1 568 808, 1965 (Ch. Y. Choo et al.).
92. Halcon Int., GB 1 218 560, 1966 (Th. W. Stein et al.); DT 1 668 234, 1966 (Th. W. Stein et al.).
93. Institut Français du Petrole, FR 1 506 286, 1966 (I. Seree de Roch et al.); DT 1 618 532, 1966 (I. Seree de Roch et al.).
94. Shell Int., GB 1 364 367, 1972 (F. Wattimena et al.); DE 2 334 315, 1972 (F. Wattimena et al.).
95. Naphthachimie, FR 1 550 166, 1967 (M. Poite).
96. Rhone-Poulenc, GB 1 156 531, 1965 (J. C. Brunie et al.); DT 1 593 306, 1965 (J. C. Brunie et al.).
97. Mitsui Petrochem, JA 45-35 041, 1965.
98. Sumitomo Chem., GB 1 141 725, 1966 (I. T. Yamachara et al.); DT 1 618 907, 1966 (I. T. Yamachara et al.).
99. R. W. Serebryakov et al., SU 362 828, 1970.
100. Kh. E. Khcheyan et al., SU 381 667, 1971.
101. Kh. E. Khcheyan et al., SU 387 985, 1971.
102. Halcon Int., DE 2 617 432, 1979 (J. P. Schmidt).
103. Halcon Int., DE 1 939 791, 1979 (M. Becker et al.).
104. Shell Oil, US 3 829 392, 1974 (H. P. Wulff).
105. Olin Corp., WO 9 015 053, 1990 (J. L. Meyer et al.).
106. Bundes-Immissionsschutzgesetz (BImSchG) 1991.
107. Technische Anleitung zur Reinhaltung der Luft (TA Luft) 1986.
108. IARC, Allyl Compounds, Aldehydes, Epoxides, and Peroxides 36 : 227, 1985.
109. D. A. Bogyo et al., Investigation of Selected Potential Environmental Contaminants: Epoxides USEPA-560/11-80-005, 1980, p. 201.
110. GEMS, Graphical Exposure Modeling System. Fate of Atmospheric Pollutants (FAP) Data Base. Office of Toxic Substances, USEPA, 1986.
111. L. Cupitt, Fate of Toxic and Hazardous Materials in the Air Evironment, USEPA-600/S3-80-084, 1980, p. 7.
112. Summary of Environmental Response Evaluation of Propylene Oxide. NTIS/OTS, 0509917 Doc, Dow, 1978.
113. R. C. Crews, Effects of PO on Selected Species of Fishes, AD/A-003-637, 1974, in SPLAT (00 033 21A).
114. J. W. Deneer et al., *Quant. SAR for Acute Tox. of some Epoxy Compounds to the Guppy,* **aq. tox. 13 (3), 195 – 204,** 1988.
115. K. Verschueren, *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., van Nostrand, Reinhold 1983.
116. A. L. Bridie et al.: "Acute Toxicity of some Petrochemicals to Goldfish," *Water Res.* **13** (1974) 623 – 626.
117. Beratergremium für umweltrelevante Altstoffe: "1,2-Epoxypropan," *BUA-Stoffbericht,* Umweltbundesamt UBA, Berlin, in print.

118. P. H. Howard (ed.): *Handbook of Environmental Fate & Exposure Data for Organic Chem.,* **vol. I** Lewis Publishers, 1989, pp. 483 – 489.
119. *A Guide to the Distribution Code of Management Practices,* Chemical Manufacturers' Association, Washington, D.C., 1991.
120. *Distribution Code Risk Management Implementation Aid,* Chemical Manufacturers' Association, Washington D.C., 1991.
121. Guideline No. 30, *The Distribution Safety Review Process,* 1st ed., Dow Europe, 1990 (only for internal use).
122. *Safe Handling and Storage of Propylene Oxide,* Dow Chemical Company, 1988.
123. *PO Product Stewardship, a Guide for Safe Handling,* Dow Europe, 1987.
124. Hommel: *Handbuch der gefährlichen Güter* Springer Verlag, Heidelberg 1987, p. 172.
125. *Emergency Response and Transportation Equipment Data Sheet for PO,* Dow Chemical 1989.
126. IMO, *International Code for the Construction and Equipment of Ships Carrying Dangerous Chemicals in Bulk,* 1983.
127. Guidelines for the Marine Chartering and Handling of Alkylene Oxides, Dow Chemical Company, April 1988.
128. A. R. Kavaler, *Chem. Mark. Rep.* **237** (1990) no. 2, 42.
129. Smyth et al., *Am. Ind. Hyg. Assoc. J.* **30** (1969) no. 5, 470 – 476.
130. V. K. Rowe et al., *AMA Arch. Ind. Health* **13** (1956) 228 – 236.
131. Smyth et al., *J. Ind. Hyg. Toxicol.* **30** (1948) 63 – 68.
132. C. H. Hine et al.: *Industrial Hygiene and Toxicology,* 2nd revised ed. **vol. II,** Interscience, New York 1958, p. 1642.
133. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A, 2 B, 2 C, Toxicology,** J. Wiley & Sons, New York 1981, p. 2187.
134. *Encyclopedia of Occupational Health and Safety,* vol. I & II, Geneva, Switzerland, International Labor Office, 1983.
135. W. M. Grant: *Toxicology of the Eye,* 3rd ed., Charles C. Thomas Publisher, Springfield, IL, 1986.
136. N. H. Proctor, J. P. Hughes: *Chemical Hazard of the Workplace,* J. B. Lippincott Company, Philadelphia, PA, 1978, p. 430.
137. K. H. Jacobsen et al., "The Toxicity of Inhaled Ethylene Oxide and Propylene Oxide Vapors," *Arch. Ind. Health* **13** (1956) 237 – 244.
138. R. E. Gosselin et al.: *Clinical Toxicology of Commercial Products,* 5th ed., Williams and Wilkins, Baltimore 1984, p. II – 97.
139. *Documentation of the Threshold Limit Values and Biological Exposure Indices,* American Conference of Governmental Industrial Hygienists, Cincinnati, Ohio, 1989.
140. Hackett, et al., Report to NIOSH, Battelle Pacific Northwest, 1982.
141. R. W. Pero et al., *Cell. Boil. & Tox.* **1** (1985) no. 4, 309 – 314.
142. D. R. Wade et al., *Mutat. Res.* **58** (1978) 217 – 223.
143. J. Bootmann et al., *Mutat. Res.* **67** (1979) no. 2, 101 –112.
144. R. E. McMahon, *Cancer Res.* **39** (1979) 682 – 693.
145. E. H. Pfeiffer, *Food Cosmet. Toxicol.* **18** (1980) no. 2, 115 – 118.
146. J. C. Cline et al., *9th Annual Meeting of Environmental Mutagen Society,* San Francisco, March 9 – 13, 1978.
147. K. Hemminki et al., *Chem. Biol. Interact.* **30** (1980) 259.
148. FAAT 10, *2-Generation Repro. Study in Rats,* 1988, pp. 82 – 88.
149. Environmental Health Research and Testing Inc., 1981, published in *Toxicol. Appl. Pharmacol.* **76** (1984) 85 – 95.
150. Environmental Health Research and Testing Inc.: *Toxic and Mutagenic Effects of PO and EO on the Spermatogenic Functions in Cynomolgus Monkeys,* 1982.
151. A. R. Sellakuman, *J. Natl. Cancer Inst.* **79** (1987) no. 2, 285 – 289.
152. C. F. Kupper et al., *Food Chemical Toxicol.* **26** (1988) 159 – 167.
153. NTP Bioassay, Technical Report 267, 1983.
154. D. Lynch et al., *Toxicol. Appl. Pharmacol.* **76** (1984) 69 – 84.
155. A. L. Walpole, *Ann. N.Y. Acad. Sci.* **68** (1958) 750.
156. H. Dunkelberg, *Br. J. Cancer* **39** (1979) 588.
157. International Agency for Research on Cancer: "Propylene Oxide (Group 2 a)," in *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Overall Evaluations of Carcinogenicity, An Updating of IARC Monographs,* vol. 1 to 42, Supplement 7, 1987, pp. 328 – 329.
158. Propylene Oxide, Environmental Health Crit. 56, 1 – 53, World Health Organization, Geneva 1985.

159. *IARC Monogr. Eval. Carcinog. Risk Chem. Man,* (1985) p. V26 238.
160. Deutsche Forschungsgemeinschaft: *1,2-Epoxypropan, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* VCH Verlagsgesellschaft, Weinheim 1989.

Propylene Polymers → Polyolefins
Propyne → Acetylene

Prostaglandins

BERND BUCHMANN, Medicinal Chemistry, Bayer Schering Pharma AG, Berlin, Federal Republic of Germany
HARTMUT REHWINKEL, Medicinal Chemistry, Bayer Schering Pharma AG, Berlin, Federal Republic of Germany

1. History 1
2. Structure, Nomenclature, and Physical Properties 1
3. Occurrence and Biosynthesis 4
4. Pharmacological Effects and Uses 7
5. Syntheses 12
6. References 20

Prostaglandins are formed from *arachidonic acid*. Those and all other derivatives of arachidonic acid, e.g., prostacyclins, thromboxanes, leukotrienes, and lipoxines are called eicosanoids. Most eicosanoids are chemically and metabolically unstable. Therefore, many pharmaceutical companies have tried to improve the stability by modifying the original structures. Several of these analogues have already reached the market in different therapeutic indications or show promising results in clinical trials. This article describes the chemical and pharmacological properties of the natural derivatives and their analogues as well as the synthetic access to some of the natural compounds.

1. History

Prostaglandins (PG) are a group of compounds formed in animals and humans that are derived from unsaturated C_{20} carboxylic acids. The physiological action of these compounds was described independently by U. S. VON EULER and M. W. GOLDBLATT as early as 1933/34 [37]. However, it was not until 1960 that S. BERGSTRÖM isolated pure $PGF_{1\alpha}$ and PGE_1 [38] and, a few years later, elucidated their structures [39]. Prostaglandins of the D series (PGD) were first characterized in 1966 [40]. The prostaglandin endoperoxides (PGG_2, PGH_2) [41] (1974, B. SAMUELSSON), thromboxane (TXA_2) [42] (1975, B. SAMUELSSON), and prostacyclin (PGI_2) [43] (1976, J. R. VANE) were presumably the last important derivatives of the cyclooxygenase pathway to be found.

In 1979, B. SAMUELSSON discovered another biosynthetic pathway that started from unsaturated C_{20} carboxylic acids, the lipoxygenase pathway [44]. The derivatives of the 5-lipoxygenase pathway are called leukotrienes (A–F). The physiological action of a mixture of some of these substances (then called SRS = slow reacting substances) was described in 1938 by W. FELDBERG and C. H. KELLAWAY [45]. Their absolute configurations were finally assigned by E. J. COREY and B. SAMUELSSON in 1980 [46].

In 1984, the structurally very closely related lipoxines were discovered by B. SAMUELSSON [47]. S. BERGSTRÖM, B. SAMUELSSON, and J. R. VANE received the Nobel Prize in 1982 for their extensive work.

2. Structure, Nomenclature, and Physical Properties

Prostaglandins are formed from *arachidonic acid.* All derivatives of arachidonic acid, e.g., prostaglandins, thromboxanes, leukotrienes, and lipoxines are called *eicosanoids.*

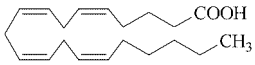

Arachidonic acid

The nomenclature of the *prostaglandins* (PG) is based on the hypothetical compound prostanoic acid [48]; the numbering of the carbon atoms is continuous and starts at the carboxyl group:

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_261.pub2

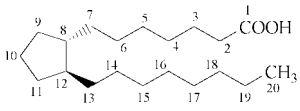

Prostanoic acid

The prostaglandins are classified (A–J) on the basis of the functional groups attached to the cyclopentane ring:

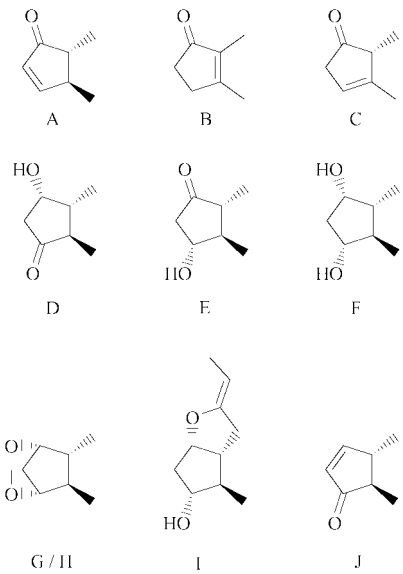

Additional numerals 1, 2, or 3 (subscripts) refer to the number, position, and stereochemistry of the double bonds in the side chains of the natural prostanes. Prostaglandins of the 1 series possess a double bond with an *E* configuration between C-13 and C-14. Those belonging to the 2 series have an additional double bond with a *Z* configuration between C-5 and C-6. Those of the 3 series possess a further double bond with a *Z* configuration between C-17 and C-18 (Table 1).

The subscript α denotes ring substituents on the same side as the carboxyl chain and β, on the opposite side. The *thromboxanes* (TX) have a tetrahydropyran ring instead of the cyclopentane ring of the prostaglandins.

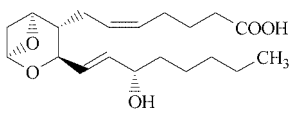

Thromboxane A_2

[*57576-52-0*] (TXA_2)

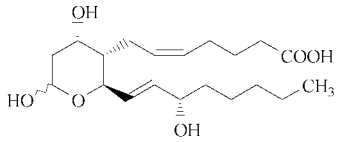

Thromboxane B_2

[*54397-85-2*] (TXB_2)

No uniform basic structure exists for the closely related *leukotrienes* (LT) and *lipoxines* (LX). These compounds are all polyunsaturated, open-chain carboxylic acids. Similar to the prostanes, the numbering of the carbon atoms starts at the carboxyl group. The letter in the name of these compounds indicates the configuration and the position of the double bonds. The numerals refer to the number of double bonds. The 14, 15 double bond with the *Z* configuration is missing in the leukotrienes of the 3 series, and an additional double bond with a *Z* configuration in position 17 exists in compounds of the 5 series.

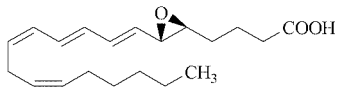

Leukotriene A_4 [*72059-45-1*] (LTA_4)

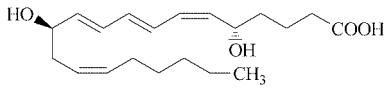

Leukotriene B_4 [*71160-24-2*] (LTB_4)

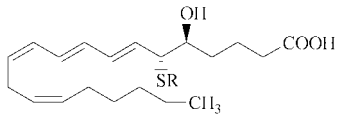

Leukotriene C_4–F_4 R =–Cys(γ-Glu)–Gly–OH, LTC_4 [*72025-60-6*]; –Cys–Gly–OH, LTD_4 [*73836-78-9*]; –Cys–OH, LTE_4 [*75715-89-8*]; –Cys(γ-Glu)–OH, LTF_4 [*83851-42-7*]

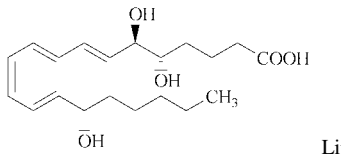

Lipoxine A_4

[*89663-86-5*] (LXA_4)

Table 1. Structure and nomenclature of some naturally occurring prostaglandins

Formula	Trivial name	Chemical name, CAS registry no.
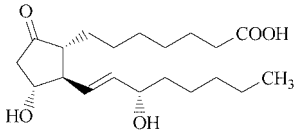	PGE_1	11α,15α-dihydroxy-9-oxo-13*E*-prostenoic acid [*745-65-3*]
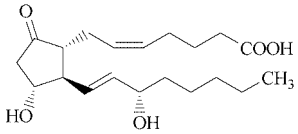	PGE_2	11α,15α-dihydroxy-9-oxo-5*Z*,13*E*-prostadienoic acid [*363-24-6*]
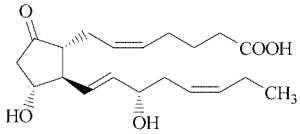	PGE_3	11α,15α-dihydroxy-9-oxo-5*Z*,13*E*,17*Z*-prostatrienoic acid [*802-31-3*]
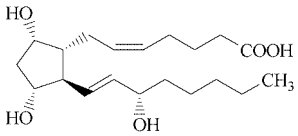	$PGF_{2\alpha}$	9α,11α,15α-trihydroxy-5*Z*,13*E*-prostadienoic acid [*551-11-1*]
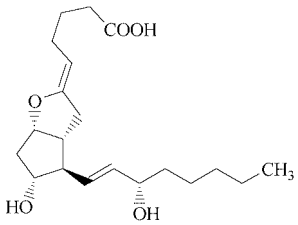	PGI_2	11α,15α-dihydroxy-6,9α-epoxy-5*Z*,13*E*-prostadienoic acid [*35121-78-9*]

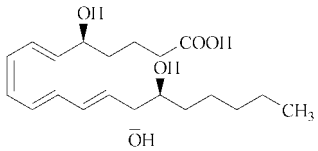

Lipoxine B_4 [*98049-69-5*] (LXB_4)

Physical Properties. Many of the natural prostaglandins, leukotrienes, and lipoxines are oils or solids with low melting points. Thus, they have to be characterized by spectroscopic methods (^{1}H- and ^{13}C-NMR, UV, and IR spectroscopy; mass spectrometry). Radioimmunoassays (RIA) based on antigen–antibody binding are most suitable for the determination of several prostanes and leukotrienes in biological media, because of their high sensitivity and selectivity. The stable metabolites of the prostanes are normally measured to obtain reproducible results.

PGE_1, $C_{20}H_{34}O_5$, M_r 354.49, *fp* 115–116 °C, $[\alpha]_D = -55°$ (tetrahydrofuran), UV spectrum after base treatment: λ_{max} (methanol) 278 nm, ε = 25 350 L mol^{-1} cm^{-1}.

PGE_2, $C_{20}H_{32}O_5$, M_r 352.47, *fp* 64.5–65.5 °C, $[\alpha]_D = -61°$ (tetrahydrofuran).

$PGF_{1\alpha}$, $C_{20}H_{36}O_5$, M_r 356.51, *fp* 101–103 °C, $[\alpha]_D = +30°$ (ethanol).

$PGF_{2\alpha}$, $C_{20}H_{34}O_5$, M_r 354.49, *fp* 30–35 °C, $[\alpha]_D = +24°$ (tetrahydrofuran).

PGA_2, $C_{20}H_{30}O_4$, M_r 334.46, viscous oil, $[\alpha]_D = +140°$ (chloroform).

PGD_2, $C_{20}H_{32}O_5$, M_r 352.47, *fp* 58–59 °C, $[\alpha]_D = +9°$ (tetrahydrofuran).

PGI_2 sodium salt, $C_{20}H_{31}O_5Na$, M_r 374.45, *fp* 116–124 °C (capillary), *fp* 166–168 °C (block), $[\alpha]_D = -124°$ (95 % ethanol), hygroscopic.

LTB_4, $C_{20}H_{32}O_4$, M_r 336.47, *fp* 25–28 °C, $[\alpha]_D = +13.1°$ (deuterochloroform), UV spec-

trum: λ_{max} (methanol) 260, 269, 281 nm, ε = 39 000, 53 000, 43 000 L mol^{-1} cm^{-1}.

LXA_4 methyl ester, $C_{21}H_{34}O_5$, M_r 366.50, $[\alpha]D = +38.2°$ (methanol), UV spectrum: λ_{max} (methanol) 284, 296, 309 nm.

LXB_4 methyl ester $C_{21}H_{34}O_5$, M_r 366.50, colorless waxy solid, $[\alpha]D = +19.2°$ (chloroform), UV spectrum: λ_{max} (methanol) 275, 287, 301, 315 nm.

Most eicosanoids have a very low chemical and metabolic stability (metabolic instability means that the biological degradation is high). For instance, PGE_2 is only stable at pH 7; below pH 7 it is converted to PGA_2, above pH 7 it is converted to PGB_2. Pure PGI_2 has a half-life of a few minutes and is hydrolyzed to 6-oxo-$PGF_{1\alpha}$. The sodium salt of PGI_2 is, however, sufficiently stable for transportation at low temperatures.

While $PGF_{2\alpha}$ is stable enough, it is advisable to store and transport the leukotrienes and lipoxines in alcoholic solution, in a cool place, and as far as possible under exclusion of oxygen.

TXA_2 decomposes within a few minutes to give TXB_2. Hence, only stable TXA_2 analogues are commercially available.

3. Occurrence and Biosynthesis

Eicosanoids occur almost ubiquitously in animals and humans. The concentrations depend on the type of tissue and are generally very low (ca. 1 µg/g wet weight). Prostaglandin concentrations in human seminal fluid are especially high (300 µg/mL).

The occurrence of eicosanoids in corals is of interest. *Plexaura homomalla,* a marine soft coral found in the Carribean, contains the 1-methyl ester of 15*R*-PGA_2-15-acetate in relatively large amounts (1–3 % of the dry weight) and the corresponding 15-epimer in smaller amounts [49]. Furthermore, new prostanoid structures called clavulones and punaglandins have been isolated from *Clavularia viridis* and *Telesto riisei*, respectively [50] (see Table 2). These compounds are of interest because they exhibit antitumor activity.

Prostaglandins also occur in plants [51]. The ubiquitous presence of these arachidonic acid derivatives in living organisms, their wide spectrum of activity, and high potency indicate that they play a central physiological role.

The enzyme phospholipase A_2 (PLA_2) releases arachidonic acid from reserves (mainly phosphatidyl choline) in the lipids of the cell membranes. The biosynthesis of prostaglandins and thromboxanes of the 2 series proceeds via PGH_2, which is formed from arachidonic acid by the enzyme cyclooxygenase (see Fig. 1).

The prostanes of the 1 and 3 series are synthesized from the corresponding unsaturated C_{20} carboxylic acids (dihomo-γ-linolenic acid or eicosapentaenoic acid).

Since 1990 a novel class of prostaglandin-like compounds, the isoprostanes, is known, the members of which may play a role as mediators of oxidant injury in vivo. Newer results indicate that they may be responsible for important biological effects in the lung and in the kidney as well. Isoprostanes are produced by a free radical process and in contrast to prostanes the stereochemical arrangement of the two side chains at the cyclopentane ring is *cis* [52].

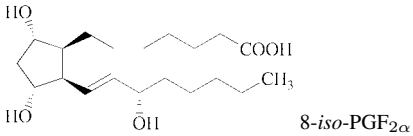

8-*iso*-$PGF_{2\alpha}$

[*27415-26-5*]

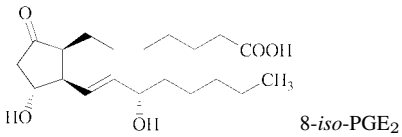

8-*iso*-PGE_2

[*27415-25-4*]

The key compound in the biosynthesis of leukotrienes is LTA_4, which is formed from arachidonic acid by the enzyme 5-lipoxygenase. LTA_4 is then converted to LTB_4 by a hydrolase and to the leukotrienes C_4–E_4 by glutathione-*S*-transferase and further degradation (Fig. 2).

Arachidonic acid can be converted to lipoxines [53 – 55] by the combined action of 15- and 5-lipoxygenase (see Fig. 2).

The eicosanoids formed locally from the released arachidonic acid exert their specific, selective action via receptors [56], causing intracellular changes in the concentration of cyclic AMP and inositol triphosphate. Using the methods of molecular biology several eicosanoid receptors and their subtypes have been charac-

Table 2. Clavulones and punaglandins

Formula	Trivial name	CAS registry no.
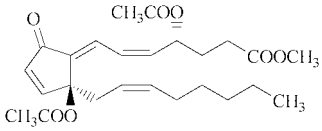	Clavulone I (claviridenone d)	[*85700-42-1*]
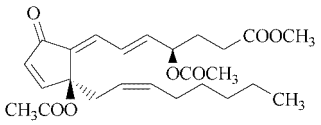	Clavulone II (claviridenone c)	[*85700-43-2*]
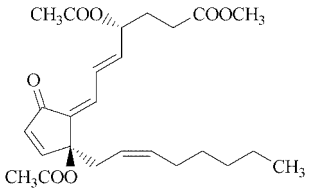	Clavulone III (claviridenone b)	[*85700-44-3*]
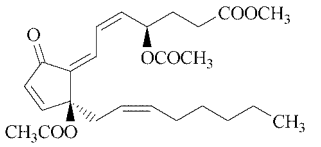	Clavulone IV (claviridenone a)	[*85611-86-5*]
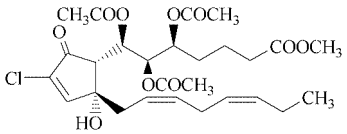	Punaglandin 1	[*96055-63-9*]
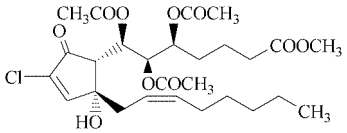	Punaglandin 2	[*96055-64-0*]
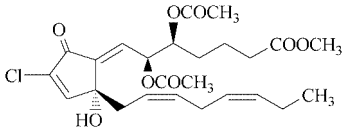	Punaglandin 3	[*96055-65-1*]
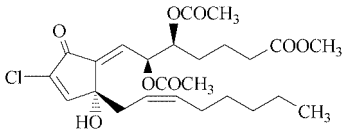	Punaglandin 4	[*96055-66-2*]

terized and some of them have been cloned (see Table 3). The major importance of natural prostaglandins lies in the autocrine and paracrine regulation of physiological and pathological processes.

The biosynthesis of prostanes, leukotrienes, and lipoxines can be influenced by inhibitors of cyclooxygenase (nonsteroidal anti-inflammatory drugs, NSAIDs), and lipoxygenase [57, 58]. Aspirin, for example, causes irreversible inhibition of cyclooxygenase and ibuprofen reversible inhibition. This inhibition is responsible for the antiphlogistic activity of these two compounds. See also, → Anti-inflammatory–Antirheumatic Drugs.

Since 1991 it has been known that cyclooxygenase exists in two isoforms COX-1 and COX-2 [59 – 62]. Whereas COX-1 is responsi-

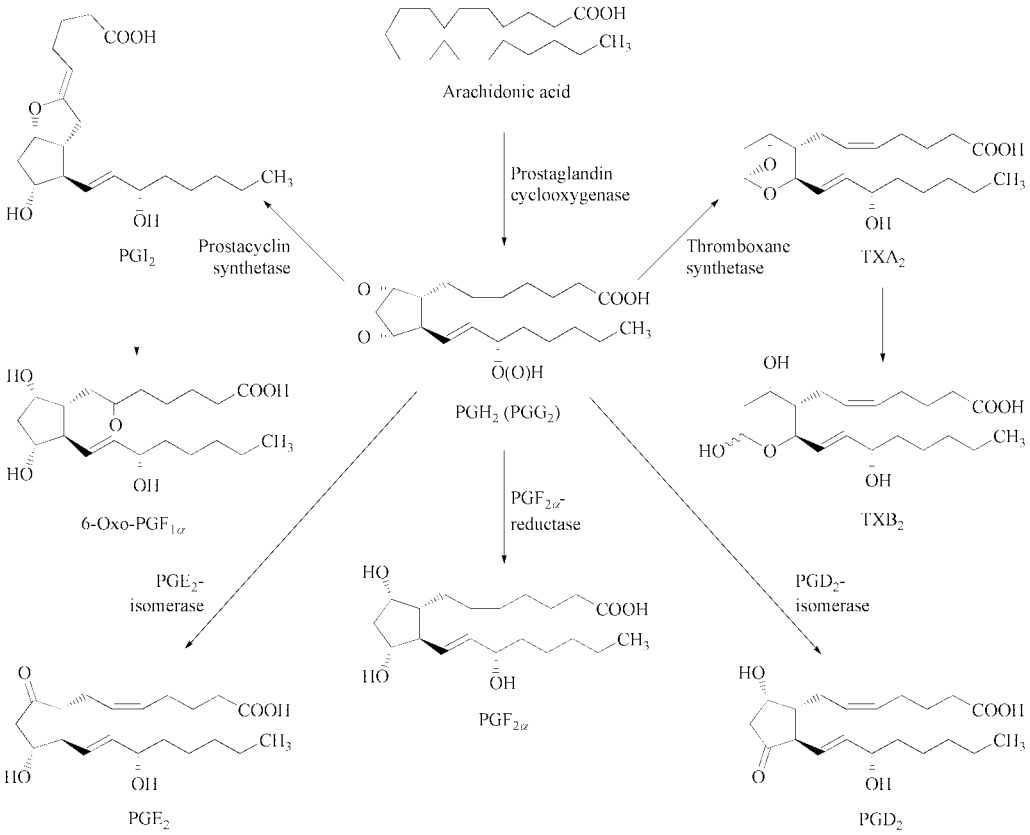

Figure 1. Biosynthesis of prostaglandins and thromboxanes from arachidonic acid

ble for the biosynthesis of prostaglandins under normal physiological conditions, COX-2 is expressed mainly in inflammatory tissue, causing the synthesis of prostaglandins which support this inflammation.

Under physiological conditions COX-2 is present for example in the brain and in the kidney but not in the gastric mucosa. Based on these findings three novel COX-2-selective NSAIDs have been introduced on the market as antirheumatic drugs with lower side effects: Celebrex (celecoxib, Pfizer) [*169590-42-5*], Vioxx (rofecoxib, Merck) [*162011-90-7*], and Bextra (valdecoxib, Pfizer) [*181695-72-7*]. However, in 2004 Merck withdraw Vioxx voluntarily because of safety concerns. In 2005 Pfizer withdraw Bextra because of cardiovascular problems and allergic skin reactions.

Nevertheless, a discussion is ongoing about the return of the drugs to the market.

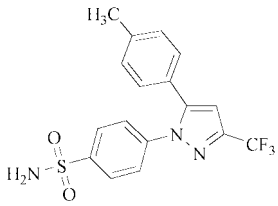

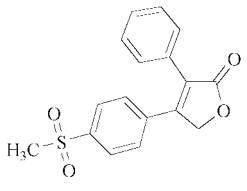

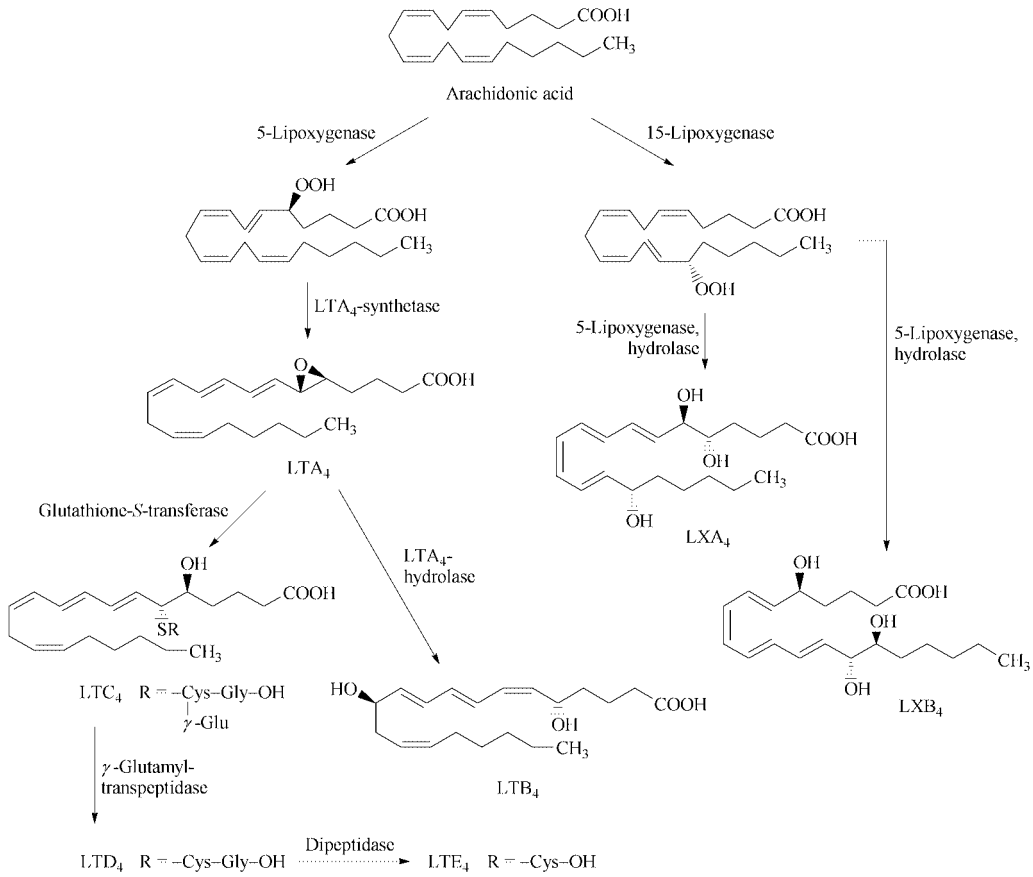

Figure 2. Biosynthesis of leukotrienes and lipoxines from arachidonic acid

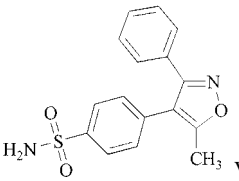

Valdecoxib

The 5-lipoxygenase inhibitor *zileuton* [*111406-87-2*] (trade name Zyflo) was introduced by Abbott as an anti-asthmatic and anti-inflammatory drug (→ Antiasthmatic Agents, Chap. 5).

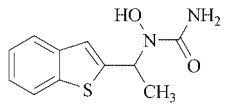

Zileuton

4. Pharmacological Effects and Uses

Eicosanoids exert different pharmacological effects on different organs. The high intrinsic activity at the site is accompanied by rapid metabolic inactivation (e.g., by 15-, β-, or ω-oxidation). The only effects of the structures in the cyclooxygenase and lipoxygenase pathways that are discussed here are those that are of possible therapeutic interest.

Reproduction. Prostaglandins of the E and F series contract uterine smooth muscle and are therefore suitable agents for the induction of parturition or abortion. As a result of their effects on other smooth muscle organs and consequent side effects (headache, nausea, vomiting, diarrhea),

Table 3. Classification of prostanoid receptors

Receptor/subtype	Selective agonist, CAS registry no.	Selective antagonist, CAS registry no.
FP	cloprostenol[*40665-92-7*] fluprostenol[*40666-16-8*] prostalene[*54120-61-5*]	AL-8810[*246246-19-5*]
EP		
EP1	17-phenyl-18,19,20-trinor-PGE_2[*38315-43-4*] sulprostone (also EP3 receptor agonist)[*60325-46-4*] iloprost (also IP receptor agonist)[*78919-13-8*]	AH 6809 (also EP2 and DP receptor antagonist)[*33458-93-4*] SC-19220[*19395-87-0*]
EP2	butaprost[*69648-38-0*]AH 13205[*148436-63-9*] misoprostol (also EP3 receptor agonist)[*59122-46-2*]ONO-AE1-257[*443796-56-3*]	AH 6809 (also EP1 and EP2 receptor antagonist)[*33458-93-4*]
EP3	enprostil[*73121-56-9*] GR 63799[*183023-82-7*] sulprostone (also EP1 receptor agonist)[*60325-46-4*] misoprostol (also EP2 receptor agonist)[*59122-46-2*]	DG041[*861238-35-9*]2-[2-[5-[2-(phenylmethoxy)phenyl]-3-thienyl]phenyl]-cyclopropanecarboxylic acid[*637327-96-9*]
EP4	11-deoxy-PGE1 (also EP2 and EP3 agonist)[*37786-00-8*] ONO-AE1-734[*256382-08-8*]	AH 22921 (also TP receptor antagonist)[*74480-23-2*] AH 23848 (also TP receptor antagonist)[*81443-73-4*] ONO-AE3-208[*402473-54-5*]
DP		
DP	BW 245C[*72814-32-5*] ZK 110841[*105595-17-3*]	BW A868C[*118675-50-6*]AH 6809 (also EP1 and EP2 receptor antagonist)[*33458-93-4*]
CRTH2	13,14-dihydro-15-oxo-PGD2[*59894-07-4*] 15-deoxy-Δ12,14-PGJ2[*87893-55-8*]	ramatroban (also TP receptor antagonist)[*116649-85-5*]TM30089[*844639-57-2*]
IP	cicaprost[*94079-80-8*] iloprost[*78919-13-8*]octimibate[*89838-96-0*]	N-[[[4-[(1H-indol-4-yloxy)methyl]phenyl]methoxy]carbonyl]-phenylalanine[*361457-89-8*]4,5-dihydro-N-[4-[[4-(1-methylethoxy)phenyl]methyl]phenyl]-1H-Imidazol-2-amine[*221529-58-4*]
TP	U44069[*56985-32-1*]U46619[*56985-40-1*]SQ 26655[*82337-14-2*] EP 011[*75010-43-4*]	AH 23848 (also EP4 receptor antagonist)[*81443-73-4*]ramatroban (also CRTH2 receptor antagonist)[*116649-85-5*]GR 32191[*87248-13-3*]EP 092[*81806-67-9*]SQ 29548[*98672-91-4*]S 145[*115266-92-7*]

these compounds are seldom given systemically to induce labor. Even local application (intraamniotic, vaginal) is accompanied by side effects.

A combination of prostaglandin and antigestagen has proved useful for inducing abortions for medical reasons because it is tolerated better than an antigestagen alone [63].

Another economically important application of prostaglandins (mainly of the F series) is in veterinary medicine. Their luteolytic activity (the ability to induce the regression of the corpus luteum) is exploited for the induction of ovulation and synchronization of the menstrual cycle in cows, pigs, sheep, and horses.

Lungs. Prostaglandins of the E and I series administered as an aerosol dilate the smooth muscles of the bronchial tract and antagonize bronchoconstriction caused by histamine. In spite of these effects, the development of a prostaglandin agent for treating asthma has not been possible so far.

However, it was found that prostacyclin analogues are suitable for the treatment of pulmonary hypertension [87]. This disease is one of the most serious chronic disorders of pulmonary circulation. It is characterized by constriction of the pulmonary blood vessels, leading to an increased blood pressure in these vessels. Since the pulmonary arteries narrow progressively in the course of the disease, the right heart is forced to pump even harder against this resistance. The increasing demand on the heart muscle eventually leads to right heart failure. The first inhaled stable prostacyclin analogue on the market for the therapy of this disease was iloprost (Ventavis).

The healthy lung has almost no 5-lipoxygenase activity, whereas formation of leukotrienes and thromboxane A_2 is increased in the asthmatic lung. The bronchoconstricting effects of LTC_4 and LTD_4 are 1000 times that of histamine. The development of selective LTC_4/D_4 antagonists or inhibitors of the 5-lipoxygenase pathway [64] have therefore proven useful in asthma therapy and some of these compounds have reached the market (→ Antiasthmatic Agents, Chap. 5). Antagonists of LTB_4 and TXA_2 have also been tested for this indication. Newer results indicate that there may be a potential to treat also the inflammatory component of cardiovascular disease with leukotriene modulators [88].

Stomach and Intestine. The outstanding significance of the prostanes (E, A, and I types) in the gastrointestinal tract is based on their inhibition of gastric acid secretion and their cytoprotective effect [65]. Cytoprotection here refers to the ability of prostaglandins, in very low concentrations, to protect the gastrointestinal tract against harmful noxious agents, e.g., nonsteroidal anti-inflammatory agents (aspirin etc.). This property is probably based on an increased formation of mucous membranes and improvement of circulation.

One of the main applications of prostaglandin derivatives is in ulcer therapy. Faster healing of peptic ulcers has been observed after the administration of PGE analogues [66] (→ Antiulcer Drugs, Section 3.4.1). In contrast to cytoprotection, relatively high doses are required to inhibit gastric acid production and thus achieve healing. Diarrhea is an undesirable side effect.

In patients suffering from ulcerative colitis, raised levels of LTB_4 and, to a lesser extent, PGE_2 and LTC_4 have been observed [67]. This indicates a possible therapeutic use of LTB_4 antagonists.

Heart and Circulation. (See also, → Cardiovascular Drugs, Section 5.3) Prostaglandins of the E and A series lower blood pressure (hypotensive effect) in humans. In contrast, $PGF_{2\alpha}$ and TXA_2 increase blood pressure because they constrict blood vessels.

Prostacyclin, the natural antagonist of thromboxane A_2, also exerts a hypotensive effect, but its inhibition of platelet aggregation is more important.

The use of prostacyclin mimetics for thrombosis prophylaxis or thrombolysis [68] (e.g., peripheral circulatory disorders, Raynaud's syndrome) is based on this effect. Although less active than PGI_2, PGE_1 is also used for the same indication. Side effects include headache, nausea, flush, and diarrhea.

Despite there has been a lot enthusiasm for the TXA_2 antagonists [69] only few of the development candidates have reached the market so far. The discovery of novel compounds acting as TXA_2-antagonist while activating the PGI2 receptor [89] may be an interesting development.

Analogous to cytoprotection in the stomach, a cardioprotective effect of PGE_1 and prostacyclin has been proposed. This could lead to the application of these compounds for prophylaxis of heart attacks.

PGI_2 mimetics may prove useful in alleviating changes produced by arteriosclerosis because the level of prostacyclin synthetase in healthy vessel walls decreases with progressing arteriosclerosis [70].

Kidneys. PGI_2 and PGE_2 cause renal vasodilation and stimulate kidney circulation [71]. In addition, prostaglandins are involved in the excretion of salt and water. Their influence on other hormones (e.g., angiotensin II) is also of significance.

In the kidney, arachidonic acid is also metabolized by another enzyme system (epoxygenase pathway). The importance of the resulting metabolites in renal circulation has not yet been clarified [72].

Cerebral Effects. TXA_2 and $PGF_{2\alpha}$ contract brain vessels, whereas PGI_2 has a dilatory effect. The use of PGI_2 mimetics [73] and TXA_2 antagonists [74] in cerebral infarctions has therefore been discussed.

With the development of new neuroprotective prostacyclin analogs, e.g., (15R)-16-*m*-tolyl-17,18,19,20-tetranorisocarbacyclin (called 15R-TIC) [90], it was possible to differentiate between the known prostacyclin receptor (now named IP1) and a potential novel prostacyclin receptor subtype named IP2. Iloprost demonstrates high-affinity binding to IP1 in the caudal medulla (nucleus of the solitary tract, NTS) but only weak affinity for the IP2 receptor. Isocarbacyclin showed similar affinity for both IP1 and IP2 subtypes and 15R-TIC (isocarbacyclin with a modified lower side chain) binds only to the novel CNS specific IP2 receptor subtype expressed in the rostral brain region. Such new type of prostacyclins may have an impact in the field of neuroprotection.

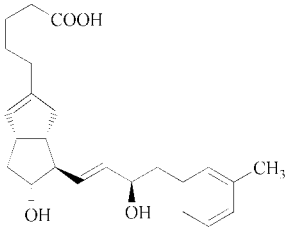

15R-TIC [*175169-68-3*]

Much higher concentrations of PGD_2 are found in the brain than in other tissues. The equilibrium between PGD_2 and PGE_2 in brain tissue apparently plays an important role in the sleep–wake cycle [75].

Miscellaneous Effects. A higher capacity for prostaglandin synthesis is found in human tumors than in normal tissue. Prostanes are involved in tumor growth and metastasis. However, PGI_2 and mimetic agents exert an antimetastatic effect on melanoma cells, probably due to their strong antiaggregation effect [76]. Although the mechanisms of action are still unclear, there is hope that a therapeutic effect will be obtained by influencing the arachidonic acid cascade [77].

An interesting discovery is the fact that prostaglandins of the F and D series are capable of reducing intra-ocular pressure, and thus represent additional therapeutics for glaucoma [78].

In 2001 a novel PGD-type prostanoid receptor named CRTH2 (chemoattractant receptor-homologous molecule expressed on TH2 cells, sometimes also named as "DP2" receptor) was identified, which displays different functions than the classical DP receptor [91]. Pharmacological characterization suggests an important role for the CRTH2 in allergic diseases and other pathophysiological conditions. Ramatroban, which has been originally developed as a TXA_2 antagonist and is marketed for seasonal allergic rhinitis, has a mixed receptor profile and is also an antagonist for CRTH2. Several companies are now looking for selective CRTH2 antagonists and their potential for antiallergic drugs.

In psoriasis, increased levels of LTB_4 have been found in inflamed areas of the skin [79]. LTB_4 seems to be involved in other inflammatory processes as well (e.g., rheumatic diseases) [80].

LTB_4 is assumed to be an important mediator of inflammation. It is a very effective chemotactic agent for neutrophilic leukocytes, i.e., the white blood cells move in the direction of increasing LTB_4 concentrations and release lysosomal enzymes which attack not only the inflamed, but also neighboring healthy tissue. LTB_4 antagonists may therefore be therapeutically useful as new antiphlogistic agents.

In contrast to the leukotriene antagonists or enzyme inhibitors the lipoxins [92] exert their anti-inflammatory potential by activating the endogenous signalling pathways by which inflammation is normally resolved. Novel lipoxins analogs such as 15-epi-17,18,19,20-tetranor-17-*p*-fluorophenoxy-LXA_4 have shown their anti-inflammatory activity in several animal model related to inflammatory bowel disease, psoriasis, and asthma. Nevertheless, none of these compounds has been tested in human studies so far. Therefore the future will show the potential of such compounds in the aforementioned diseases.

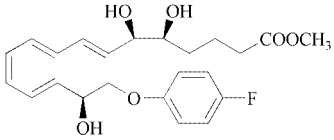

15-epi-17,18,19,20-tetranor-17-*p*-fluorophenoxy-LXA_4 [*205176-29-0*]

The natural eicosanoids are normally not considered for therapeutic applications because of their rapid metabolic degradation and their wide spectrum of activity. Modified derivatives are therefore synthesized to improve not only the selectivity, but also the metabolic stability of these compounds.

A selection of prostaglandin, prostacyclin, thromboxane, and leukotriene derivatives is presented in Tables 4 5 6 7 8, respectively. They include therapeutic agents that have already been introduced to the market and compounds that are still the subject of research and development.

Economic Aspects. The economically most important applications of eicosanoids are the treatment of cardiovascular, gastrointestinal, pulmonary, ophthalmic and gynaecological diseases.

For example, 5 % of men over 50 suffer from peripheral circulatory disorders. Of these 10 % develop a critical phase, i.e., 25 % will lose a limb and 20 % will die within one year. These patients strongly benefit from treatment with commercially available prostanes (e.g., Alprostadil, Epoprostenol, Iloprost, Beraprost).

Another example is the use of prostaglandins in patients with arthritis who must be treated with NSAIDs over a long period. About a quarter of them suffer from gastric ulcers which develop as a result of the daily intake of these drugs.

Table 4. Uses of prostaglandins in human and veterinary medicine

Structure	Generic name, CAS registry no.	Originator	Trade name or phase of development	(Potential) indication
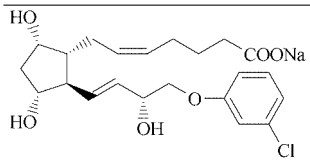	cloprostenol sodium [*55028-72-3*]	Zeneca	Estrumate	synchronization of menstrual cycle in cows
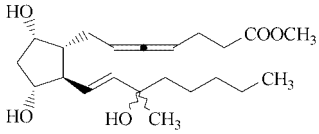	prostalene [*54120-61-5*]	Syntex	Synchrocept	synchronization of menstrual cycle in domestic animals
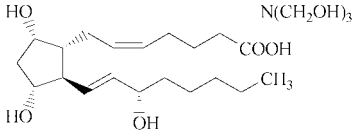	dinoprost [*38562-01-5*]	Pharmacia & Upjohn	Minprostin $F_{2\alpha}$	termination of pregnancy
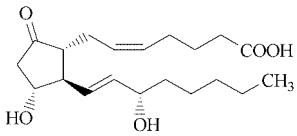	dinoprostone [*363-24-6*]	Pharmacia & Upjohn	Minprostin E_2	termination of pregnancy
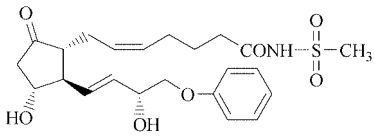	sulprostone [*60325-46-4*]	Schering AG	Nalador	termination of pregnancy
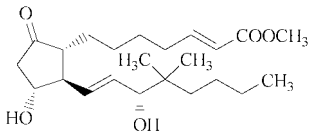	gemeprost [*64318-79-2*]	ONO	Gervagem	termination of pregnancy, initiation of menstruation
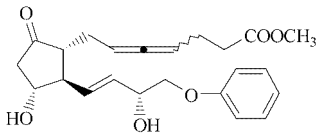	enprostil [*73121-56-9*]	Syntex	Gardrin	antiulcer
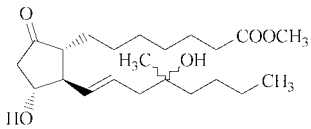	misoprostol [*59122-46-2*]	Searle	Cytotec	antiulcer
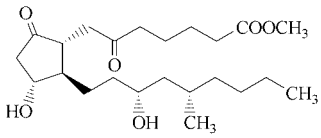	ornoprostil [*70667-26-4*]	ONO	Ronok	antiulcer
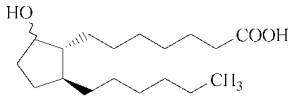	rosaprostol [*56695-65-9*]	IBI	Rosal	antiulcer

Table 4. Continued

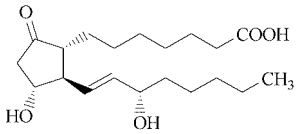	alprostadil [*745-65-3*]	Schwarz Pharma, ONO	Prostandin, Prostavasin	arterial circulatory disorders
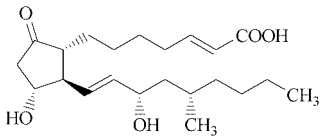	limaprost [*74397-12-9*]	ONO	Opalmon	peripheral circulatory disorders
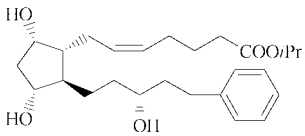	latanoprost [*130209-82-4*]	Pharmacia & Upjohn	Xalatan	glaucoma
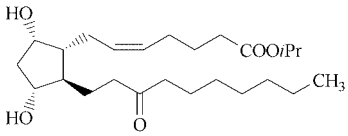	unoprostone isopropylester [*120373-24-2*]	Ueno	Rescula	glaucoma
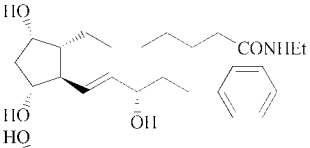	bimatoprost [*155206-00-1*]	Allergan	Lumigan	glaucoma
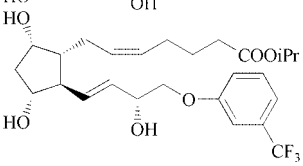	travoprost [*157283-68-6*]	Alcon	Travatan	glaucoma

Prostaglandins (e.g., Misoprostol or Enprostil) prevent these injuries effectively.

In the future additional eicosanoid derivatives may be suitable for treating certain cancers [93] and inflammation diseases (e.g., psoriasis).

5. Syntheses

Many different syntheses exist for the compounds of the cyclooxygenase and lipoxygenase pathways (Chap. 2). A large number of these syntheses are described in numerous review articles and monographs [1 — 32]. Three basic strategies of synthesis that are exemplary for the prostaglandin field are presented here, followed by the first synthesis of the natural substances PGI_2, TXA_2, and LTB_4. All these syntheses yield not only the natural substances, but also allow the synthesis of analogues. Some of these methods are now used for industrial production.

Synthesis of $PGF_{2\alpha}$ According to E. J. COREY (Fig. 3). E. J. COREY was the first to develop a method for a stereocontrolled access to natural prostaglandins [81]. The cyclopentadiene derivative **1** is converted in a Diels–Alder reaction with α-chloroacrylic acid chloride to the bicyclic compound **2**. Curtius rearrangement with subsequent hydrolysis gives the ketone **3**, which is converted to the lactone **4** in a Baeyer–Villiger oxidation. After saponification of the lactone, the acid is subjected to a resolution of the enantiomers with an optically active amine (e.g., amphetamine). Iodine lactonization of the acid **5** and benzoylation give the benzoate **6**. Reductive elimination of the iodine with tributyltin hydride and subsequent hydrogenolysis of the benzyl ether yield the Corey alcohol **7**. Oxidation to the aldehyde followed

Table 5. Uses of prostacyclins in human medicine

Structure	Generic name, CAS registry no.	Originator	Trade name or phase of development	(Potential) indication
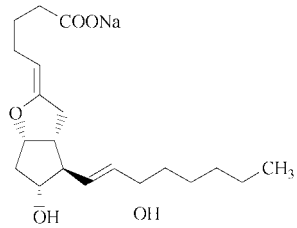	epoprostenol sodium [*61849-14-7*] Na salt, [*35121-78-9*] free acid	Glaxo Wellcome, Pharmacia & Upjohn	Flolan, Cyclo-Prostin	circulatory disorders
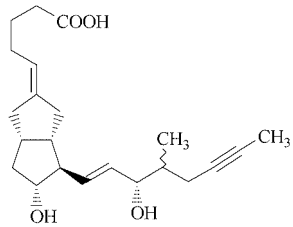	iloprost [*78919-13-8*]	Schering AG	Ilomedin	peripheral circulatory disorders, pulmonary hypertension
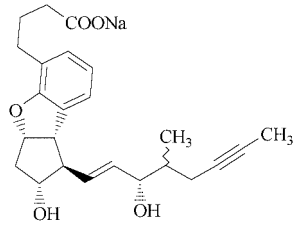	beraprost-sodium [*88475-69-8*]	Toray-Kaken	Dorner, Procyclin	peripheral circulatory disorders
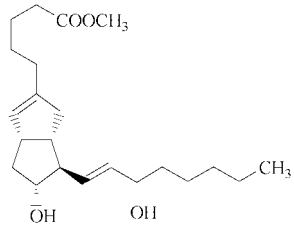	clinprost [*88931-51-5*] methyl ester, [*88911-35-7*] free acid	Teijin	clinical trials, no development reported	peripheral circulatory disorders, cerebral thrombosis
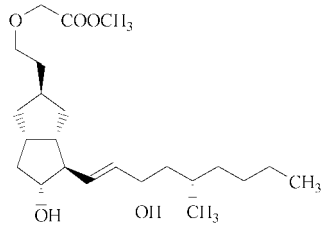	pimilprost [*139403-31-9*] methyl ester, [*157318-92-8*] free acid	Sumitomo	Skipron	peripheral circulatory disorders
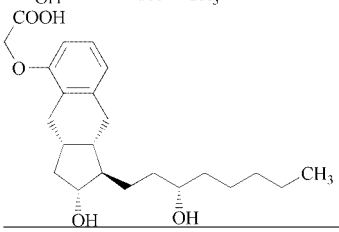	uniprost [*081846-19-7*]	United Therapeutics	Remodulin	peripheral vascular disease, pulmonary hypertension

Table 6. Uses of thromboxane antagonists in human medicine

Structure	Generic name, CAS registry no.	Originator	Trade name or phase of development	(Potential) indication
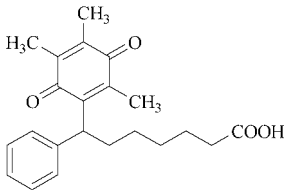	seratrodast [*112665-43-7*]	Takeda	Bronica	asthma, allergic rhinitis
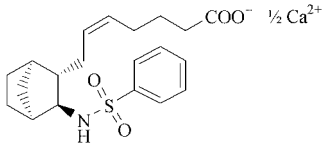	domitroban-calciumhydrate [*132747-47-8*]	Shionogi	Anboxan	asthma, rhinitis, antithrombotic
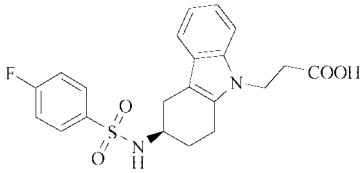	ramatroban [*116649-85-5*]	Bayer	Baynas	asthma, allergic rhinitis, antithrombotic angina
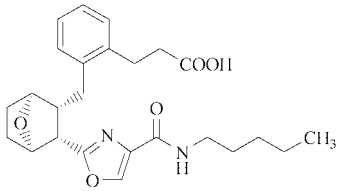	ifetroban [*143443-90-7*]	Bristol-Myers Squibb	discontinued	antithrombotic, anti-ischaemic, topical for pressure-ulcers
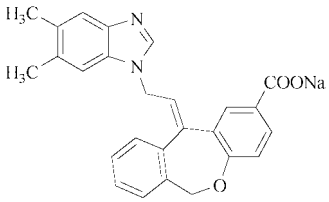	KW-3635 [*127166-41-0*]	Kyowa Hakko	discontinued	asthma, antithrombotic
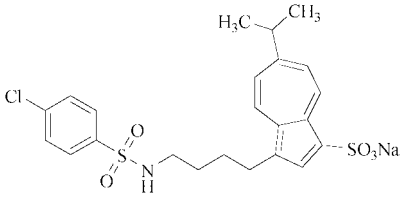	KT2–962 [*129648-96-0*]	Kotobuki Seiyaku	clinical trials, no development reported	cardiovascular, renal diseases
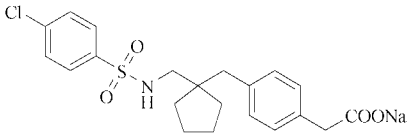	LCB-2853 [*141335-11-7*]	Merck KGaA	clinical trials, no development reported	antithrombotic

by a Horner–Emmons reaction gives the α,β-unsaturated ketone **8**. After reduction with zinc borohydride, the epimeric allyl alcohols are separated, giving the desired alcohol **9**. Cleavage of the benzoate and etherification of the two hydroxyl groups with dihydropyran yield the bis(tetrahydropyranyl) ether **10**. The reduction of the lactone to lactol with diisobutylaluminum

Table 7. Uses of peptido leukotriene antagonists in human medicine

Structure	Generic name, CAS registry no.	Originator	Trade name or phase of development	(Potential) indication
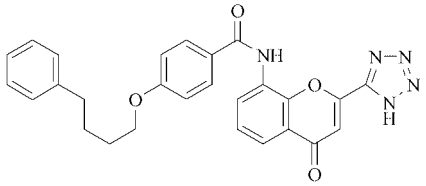	pranlukast [*103177-37-3*]	ONO	Onon, Ultair	asthma, allergy
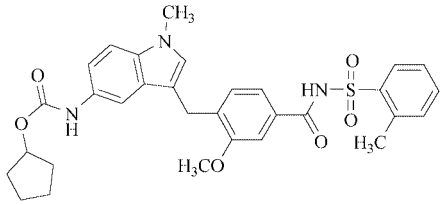	zafirlukast [*107753-78-6*]	Zeneca	Accolate, Vanticon	asthma, allergy
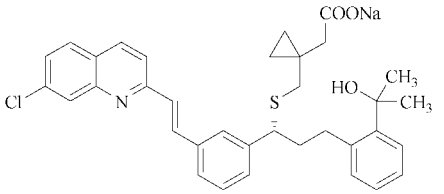	montelukast [*151767-02-1*]	Merck & Co.	Singulair	asthma
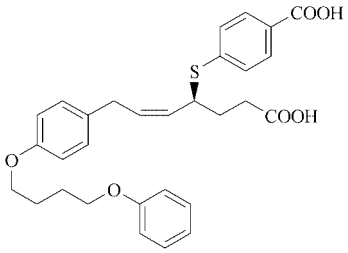	Bay x7195 [*143538-27-6*]	Bayer	discountinued	asthma
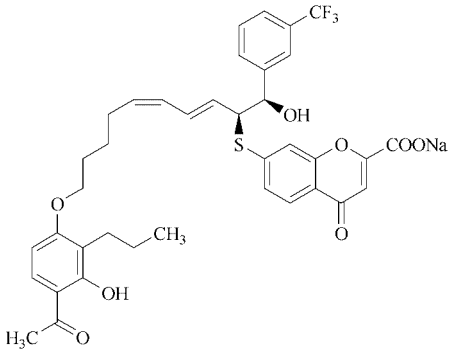	iralukast sodium [*125617-94-9*]	Novartis	clinical trials, no development reported	asthma
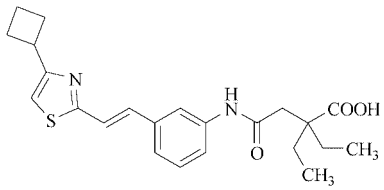	cinalukast [*128312-51-6*]	Hoffmann–La Roche	clinical trials, no development reported	asthma

Table 8. Uses of leukotriene B_4 antagonists in human medicine

Structure	Generic name, CAS registry no.	Originator	Trade name or phase of development	(Potential) indication
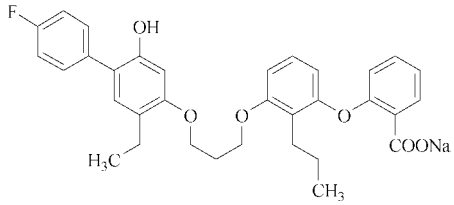	etalocib, VML-295, LY 293111 [*152608-41-8*] Na salt, [*161172-51-6*] free acid	Eli Lilly	clinical trials	asthma, pancreas tumor
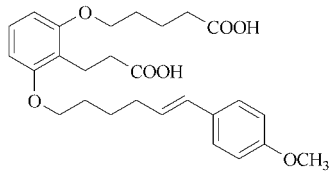	ONO LB-457, ONO 4057 [*134578-96-4*]	ONO	clinical trials, no development reported	atopic dermatitis, ulcerative colitis, Behcet's disease
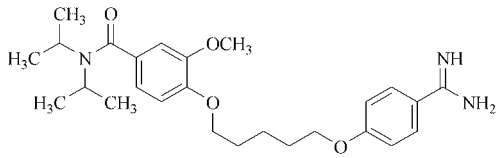	moxilubant, CGS 25019C [*147398-01-4*]	Novartis	clinical trials, no development reported	psoriasis
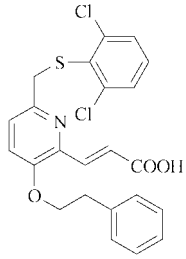	ticolubant, SB 209247 [*154413-16-3*]	SmithKline Beecham	discountinued	eczema
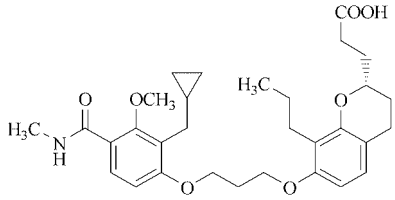	SC-53228 [*153633-01-3*]	Searle	discountinued	psoriasis, inflammatory bowel disease
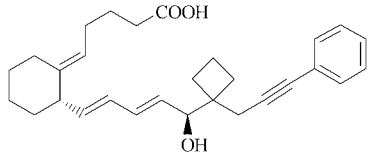	ZK 158252 [*162362-36-9*]	Schering AG	discountinued	psoriasis

hydride (DIBAH) and a subsequent Wittig reaction yield the acid **11**, which gives natural $PGF_{2\alpha}$ after cleavage of the protecting groups.

Synthesis of PGE_2 via the Cyclopentenone Method According to C. J. SIH (Fig. 4). As an example of this synthesis strategy, a method for the production of natural PGE_2 will be described [82]. Optical activity is introduced by a microbiological reaction. The triketone **13** can be obtained in several steps from the acetoacetic ester **12**. The chiral alcohol **14** is made by a regio- and enantioselective, microbiological reduction of the racemic triketone **13**. Conversion to an enol sulfonate, reduction, and protection of the hy-

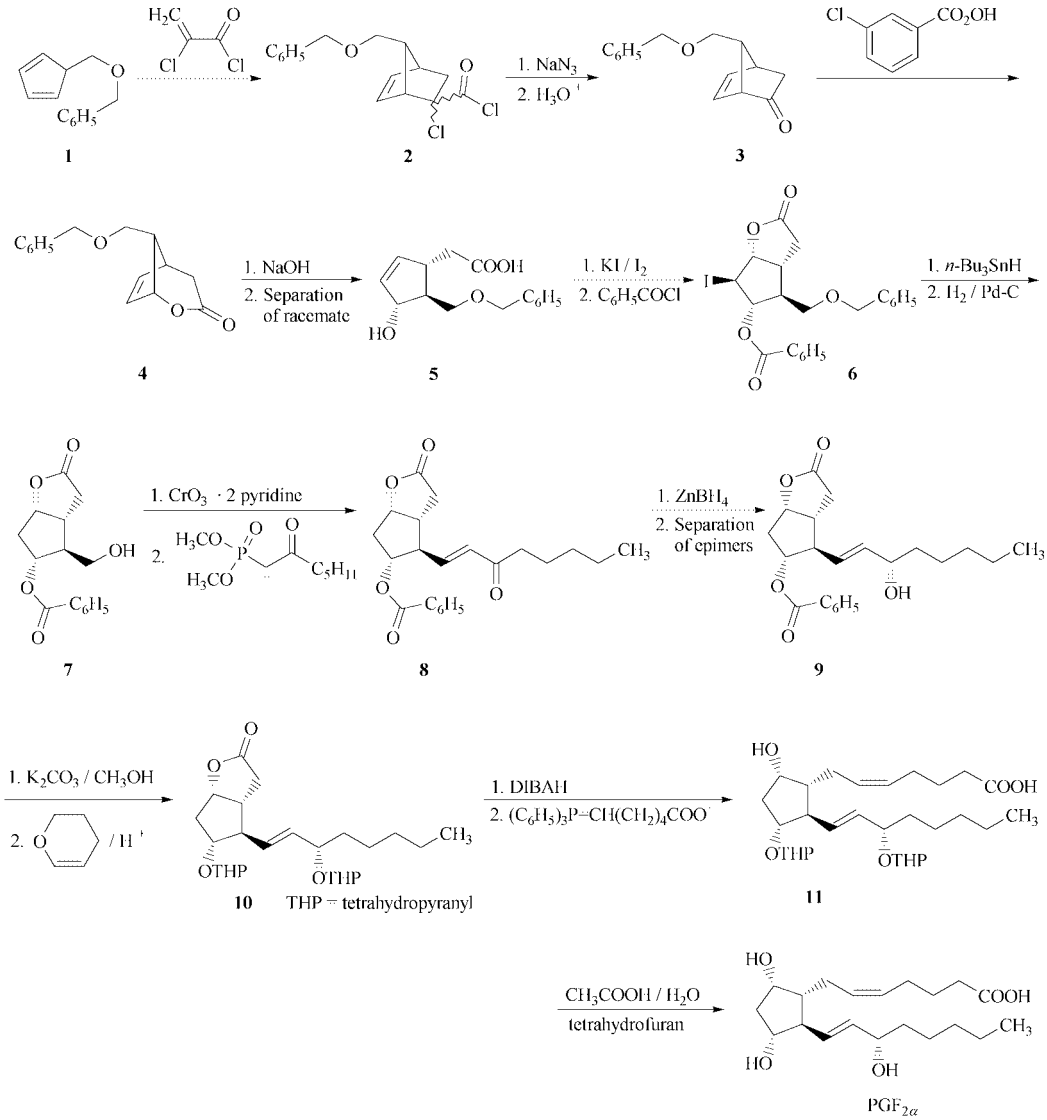

Figure 3. Synthesis of $PGF_{2\alpha}$ according to E. J. COREY

droxyl group yield the cyclopentenone **15**. The stereospecific 1,4-addition of the cuprate **16** to the α,β-unsaturated ketone **15** and cleavage of the protecting groups give the methyl ester of PGE_2 **17**, which yields natural PGE_2 by microbiological saponification. Saponification with a base is not possible because of the labile β-hydroxyketone moiety.

Synthesis of PGE_2 via the Three-Component Method According to R. NOYORI [83] (Fig. 5). This is a "one-pot" process for the introduction of the α- and ω-chain. The optically active 3-hydroxycyclopentenone derivative **18** reacts with the organolithium compound **19** in a copper-catalyzed 1,4-addition. Then the resulting enolate reacts with the allyl iodide **20** via an organotin intermediate to give the PGE_2 derivative **21**. Cleavage of the protecting groups yields PGE_2.

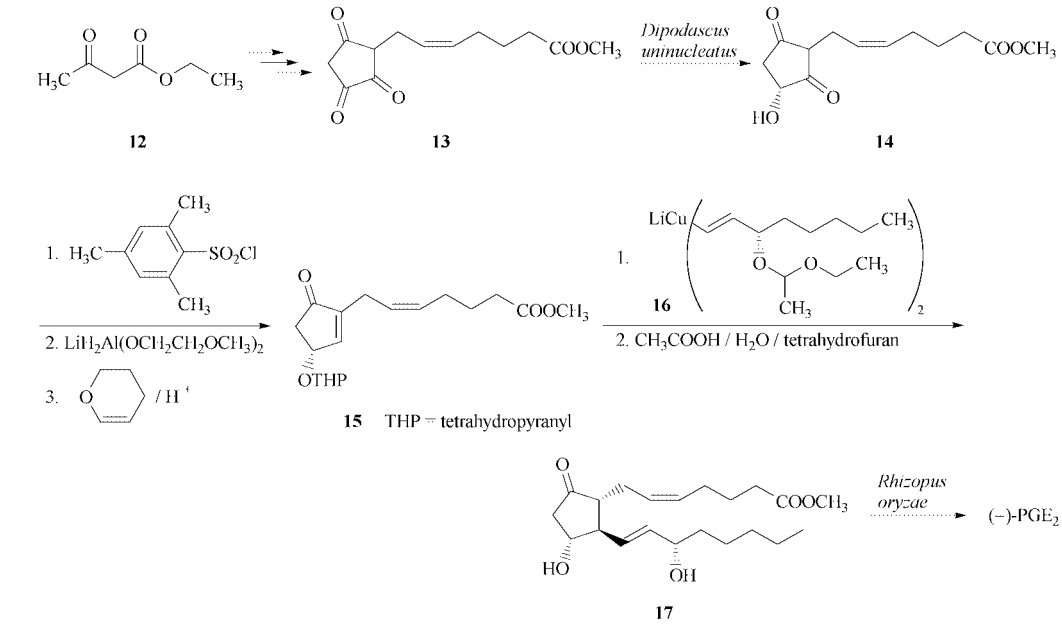

Figure 4. Synthesis of PGE_2 via the cyclopentenone method according to C. J. SIH

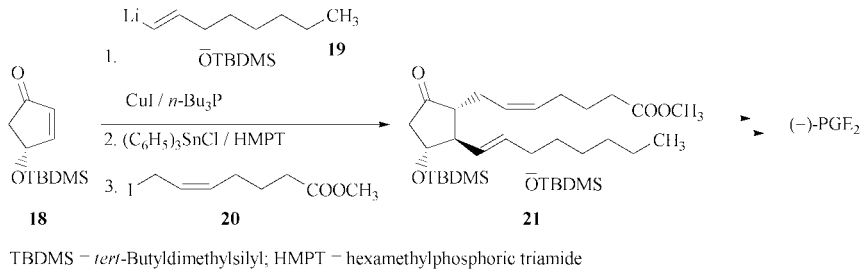

Figure 5. Synthesis of PGE_2 via the three-component method according to R. NOYORI

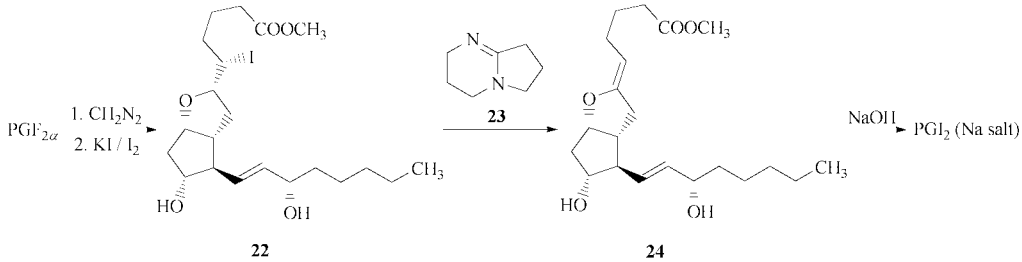

Figure 6. Synthesis of prostacyclin (PGI_2) from $PGF_{2\alpha}$

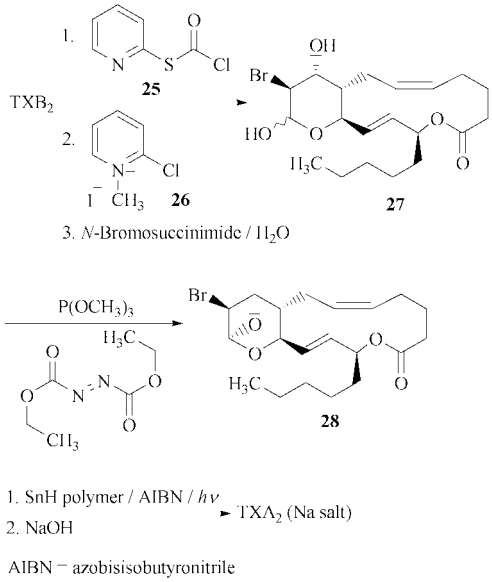

Figure 7. Synthesis of TXA_2 according to W. C. STILL

Synthesis of PGI_2 from $PGF_{2\alpha}$ (Fig. 6). The synthesis of prostacyclin (PGI_2) from $PGF_{2\alpha}$ was described almost at the same time by five different teams independently of each other [84]. Esterification of natural $PGF_{2\alpha}$ with diazomethane and subsequent iodine lactonization yield the PGI_1 derivative **22**. Basic elimination with diazabicyclononene (**23**) gives the enol ether **24**, which, in turn, yields the sodium salt of PGI_2 after saponification. The free acid is normally not made because it is very unstable.

Synthesis of TXA_2 According to W. C. STILL [85] (Fig. 7). Starting with natural thromboxane B_2, a 1,15-macrolactone is synthesized as an intermediate with the help of the pyridine derivative **25**. Elimination of the 11-hydroxyl group with the Mukaiyama reagent **26** followed by a HOBr addition yields the bromohydrin **27**. An intramolecular Mitsunobu reaction gives the strained oxetane system **28**. Reductive elimination of the bromine with the help of polymer-bound tin hydride and subsequent saponification

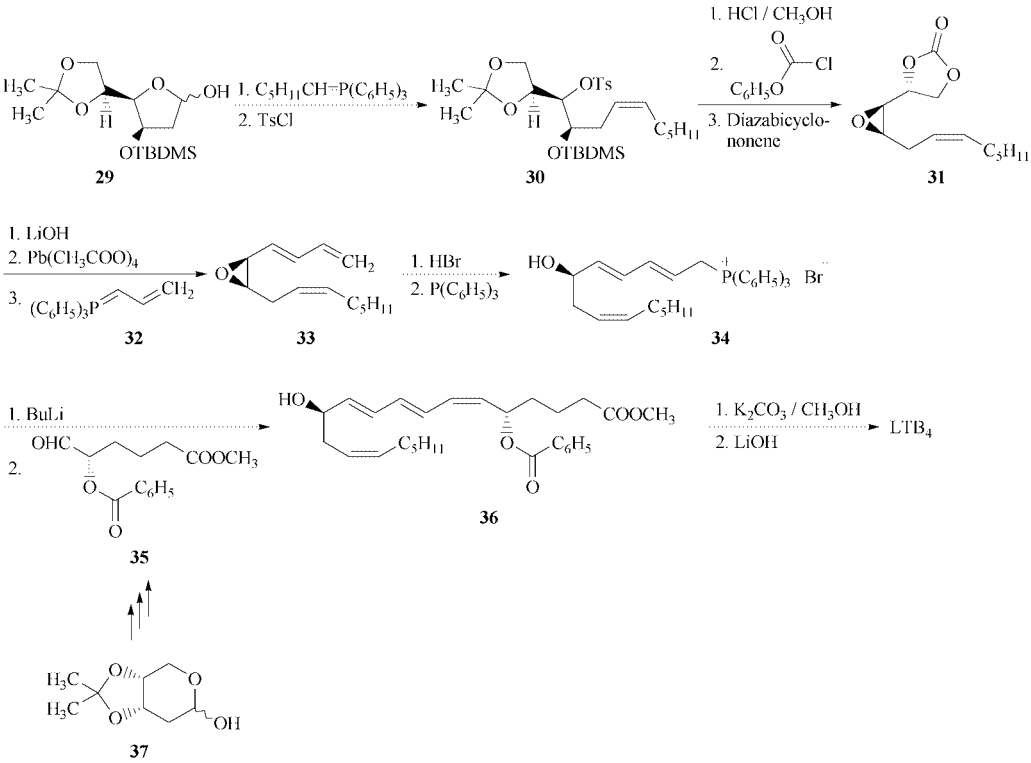

Figure 8. Synthesis of LTB_4 according to E. J. COREY

furnish the sodium salt of TXA_2. It has not yet been possible to synthesize free TXA_2 because of its great instability.

Synthesis of LTB_4 According to E. J. COREY [86] (Fig. 8). The lactol **29** can be made from D-(+)-mannose in a few steps and is then converted to the olefin **30** in a Wittig reaction with subsequent tosylation. Methanolic HCl simultaneously cleaves the acetonide and the silyl ether. The vicinal hydroxyl groups in the resulting triol are converted to the carbonate with phenyl chloroformate. Nucleophilic substitution of the tosylate by the remaining secondary hydroxyl group yields the epoxide **31**. The epoxy aldehyde obtained after saponification and glycol cleavage with lead tetraacetate is converted to the triene **33** in a Wittig reaction with the phosphorane **32**. The epoxide opening with HBr proceeds with simultaneous displacement of the diene system. The resulting primary bromide is converted to the Wittig salt **34** with triphenylphosphine. Deprotonation with butyllithium and reaction of the resulting ylide with the aldehyde **35** give the protected LTB_4 derivative **36**. Stepwise saponification produces natural LTB_4. The aldehyde **35** is synthesized in several steps from the acetonide of the deoxyribose **37**.

Synthesis of LXA_4 According to J. ROKACH [94] (Fig. 9). The enantiomerically pure carbonate **38** can be made in five steps from commercially available 2-deoxy-D-ribose. Swern oxidation followed by a Wittig reaction and subsequent isomerization provides the diene aldehyde **39**. A second Wittig reaction with the known optically active ylide **40** yields the 1:1 E/Z-mixture **41**. Removal of the silyl group followed by E/Z separation and final saponification affords the desired lipoxin A_4.

6. References

General References

1. J. S. Bindra, R. Bindra: *Prostaglandin Synthesis*, Academic Press, New York 1977.
2. A. Mitra: *The Synthesis of Prostaglandins*, Wiley-Interscience, New York 1977.
3. S. M. M. Karim (ed.): *The Prostaglandins*, Medical and Technical Publishing, Oxford 1972.
4. E. W. Horton: *Prostaglandins*, Springer Verlag, Berlin 1972.
5. P. W. Ramwell (ed.): *The Prostaglandins*, **2 vols.**, Plenum Press, New York 1974.
6. J. S. Bindra, R. Bindra, *Prog. Drug Res.* **17** (1973) 410.
7. T. O. Oesterling, W. Morozowich, T. J. Roseman, *J. Pharm. Sci.* **61** (1972) 1861.
8. K. C. Nicolaou, G. P. Gasic, W. E. Barnette, *Angew. Chem.* **90** (1978) 360; *Angew. Chem. Int. Ed. Engl.* **17** (1978) 293.
9. P. R. Marsham: *Aliphatic and Related Natural Product Chemistry (Specialist Periodical Report)*, vol. **1**, The Chemical Society, London 1979, p. 170.
10. W. Bartmann, *Angew. Chem.* **87** (1975) 143; *Angew. Chem. Int. Ed. Engl.* **14** (1975) 337.
11. W. Bartmann, G. Beck, *Angew. Chem* **94** (1982) 767; *Angew. Chem. Int. Ed. Engl.* **21** (1982) 751.
12. B. Samuelsson, *Angew. Chem.* **94** (1982) 881; *Angew. Chem. Int. Ed. Engl.* **21** (1982) 902.

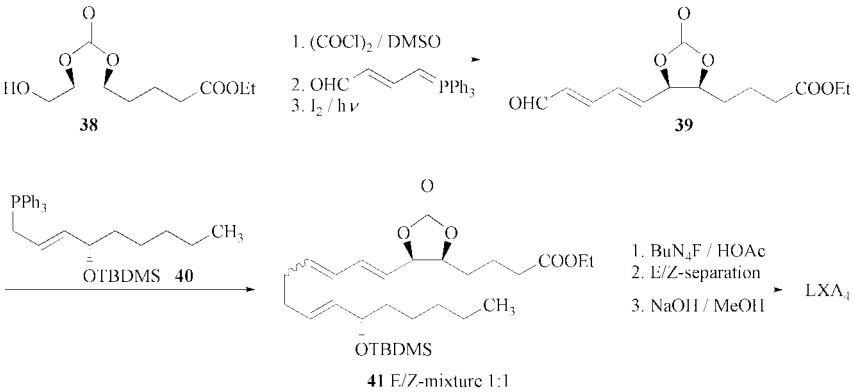

Figure 9. Synthesis of LXA_4 according to J. ROKACH

13. B. Samuelsson, *Angew. Chem.* **95** (1983) 854; *Angew. Chem. Int. Ed. Engl.* **22** (1983) 805.
14. S. Bergström, *Angew. Chem.* **95** (1983) 865; *Angew. Chem. Int. Ed. Engl.* **22** (1983) 858.
15. J. R. Vane, *Angew. Chem.* **95** (1983) 782; *Angew. Chem. Int. Ed. Engl.* **22** (1983) 741.
16. R. H. Green, P. F. Lambeth, *Tetrahedron* **39** (1983) 1687.
17. K. Schrör: *Prostaglandine und verwandte Verbindungen*, Thieme Verlag, Stuttgart 1984.
18. R. C. Nickolson, M. H. Town, H. Vorbrüggen, *Med. Res. Rev.* **5** (1985) 1.
19. P. W. Collins, S. W. Djuric, *Chem. Rev.* **93** (1993) 1533.
20. J. R. Vane, J. O'Grady (eds.): *Therapeutic Applications of Prostaglandins*, Edward Arnold, London–Boston–Melbourne–Auckland 1993.
21. F. Scheinmann, J. Ackroyd: *Leukotriene Syntheses*, Raven Press, New York 1984.
22. L. W. Chakrin, D. M. Bailey (eds.): *The Leukotrienes*, Academic Press, Orlando 1984.
23. J. E. Pike, D. R. Morton (eds.): *Advances in Prostaglandin, Thromboxane, and Leukotriene Research,* vol. 14, Raven Press, New York 1985.
24. G. Beck: *Arzneimittel, Fortschritte 1972–1985*, VCH Verlagsgesellschaft, Weinheim, Germany 1987, p. 835.
25. A. L. Willis (ed.): *Handbook of Eicosanoids: Prostaglandins and Related Lipids*, 2 vols., CRC Press, Boca Raton, Florida 1987.
26. T. W. Hart, *Nat. Prod. Rep.* **5** (1988) 1.
27. J. Rokach (ed.): *Leukotrienes and Lipoxygenases*, Elsevier, Amsterdam 1989.
28. W. T. Jackson, J. H. Fleisch, *Prog. Drug Res.* **46** (1996) 115.
29. C. D. W. Brooks, J. B. Summers, *J. Med. Chem.* **39** (1996) 2629.
30. S. L. Spector, *Annals of Allergy, Asthma, Immunol.* **75** (1995) 463.
31. E. Aedelroth et al., *J. Allergy Clin. Immunol.* **99** (1997) 210.
32. S. E. Dahlén et al. (eds.): *Leukotrienes as Mediators of Asthma and Inflammation: Basic and Applied Research*, Raven Press, New York, 1994.
R. De Caterina, A. Zampolli, *N. Engl. J. Med.* **350** (2004) 4.
33. J.-M. Dogné, J. Hanson, D. Pratico, *Trends Pharmacol. Sci.* **26** (2005) 639.
34. J.-M. Dogné et al., *Curr. Med. Chem.* **11** (2004) 1223.
35. X. de Leval et al., *Curr. Med. Chem.* **11** (2004) 1243.
36. M. Kapoor, O. Shaw, I. Appleton, *Curr. Opin. Invest. Drugs* **6** (2005) 461.

Specific References

37. U. S. von Euler, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **175** (1934) 78.
M. W. Goldblatt, *Chem. Ind. (London)* **52** (1933) 1056.
38. S. Bergström, J. Sjövall, *Acta Chem. Scand.* **14** (1960) 1693.
S. Bergström, J. Sjövall, *Acta Chem. Scand.* **11** (1957) 1086.
S. Bergström, J. Sjövall, *Acta Chem. Scand.* **14** (1960) 1701.
39. D. H. Nugteren et al., *Nature (London)* **212** (1966) 38.
S. Bergström, R. Ryhage, B. Samuelsson, J. Sjövall, *Acta Chem. Scand.* **16** (1962) 501.
S. Bergström, R. Ryhage, B. Samuelsson, J. Sjövall, *J. Biol. Chem.* **238** (1963) 3555.
40. M. Hamberg, B. Samuelsson, *J. Am. Chem. Soc.* **88** (1966) 2349.
D. H. Nugteren, R. K. Beerthuis, D. A. van Dorp, *Rec. Trav. Chim. Pays-Bas* **85** (1966) 405.
41. M. Hamberg, B. Samuelsson, *Proc. Natl. Acad. Sci. USA* **71** (1974) 3400.
M. Hamberg, J. Svensson, T. Wakabayashi, B. Samuelsson, *Proc. Natl. Acad. Sci. USA* **71** (1974) 345.
42. M. Hamberg, J. Svensson, B. Samuelsson, *Proc. Natl. Acad. Sci. USA* **72** (1975) 2994.
43. S. Bunting, R. Gryglewski, S. Moncada, J. R. Vane, *Prostaglandins* **12** (1976) 897.
44. P. Borgeat, B. Samuelsson, *Proc. Natl. Acad. Sci. USA* **76** (1979) 2148.
45. W. Feldberg, C. H. Kellaway, *J. Physiol. (London)* **94** (1938) 187.
46. S. Hammarström et al., *Biochem. Biophys. Res. Commun.* **92** (1980) 946.
47. C. N. Serhan, M. Hamberg, B. Samuelsson, *Biochem. Biophys. Res. Commun.* **118** (1984) 943.
C. N. Serhan, M. Hamberg, B. Samuelsson, *Proc. Natl. Acad. Sci. USA* **81** (1984) 5335.
48. N. A. Nelson, *J. Med. Chem.* **17** (1974) 911.
R. A. Johnson, D. R. Morton, N. A. Nelson, *Prostaglandins* **15** (1978) 737.
49. R. J. Light, B. Samuelsson, *Eur. J. Biochem.* **28** (1972) 232.
50. M. Fukushima, T. Kato, *Adv. Prostaglandin Thromboxane Leukotriene Res.* **15** (1985) 415.
A. N. Grechkin, *J. Lipid Mediat. Cell Signalling* **11** (1995) 205.

51. E. D. Levin, V. E. Cherepanova, I. A. Zimovtseva, T. O. Sedlova, *Phytochemistry* **27** (1988) 3241.
52. M. J. Müller, *Chemistry & Biology* **5** (1998) R323.
53. K. C. Nicolaou, J. Y. Ramphal, N. A. Petasis, C. N. Serhan, *Angew. Chem.* **103** (1991) 1119; *Angew. Chem. Int. Ed. Engl.* **30** (1991) 1100.
L. J. Janssen, *Am. J. Physiol. Lung Cell Mol. Physiol.* **280** (2001) L1067.
K. F. Badr, T. E. Abi-Antoun, *Antioxid. Redox Signal* **7** (2005) 236.
54. C. N. Serhan, *Biochimica et Biophysica Acta* **1212** (1994) 1.
55. C. N. Serhan, *Prostaglandins* **53** (1997) 107.
56. I. Kennedy, R. A. Coleman, P. P. A. Humphrey, P. Lumley, *Adv. Prostaglandin Thromboxane Leukotriene Res.* **11** (1983) 327.
K.-H. Thierauch, H. Dinter, G. Stock, *J. Hypertens.* **11** (1993) 1315.
K.-H. Thierauch, H. Dinter, G. Stock, *J. Hypertens.* **12** (1994) 1.
M. Ushikubi, M. Hirata, S. Narumiya, *J. Lipid Mediators Cell Signalling* **12** (1995) 343.
M. Hirata, F. Ushikubi, S. Narumiya, *J. Lipid Mediators Cell Signalling* **12** (1995) 393.
R. R. A. Coleman, W. L. Smith, S. Narumiya, *Pharmacol. Rev.* **46** (1994) 205.
T. Yokomizo et al., *Nature* **387** (1997) 620.
57. G. Weissmann, *Sci. Am.* **264** (1991) 58.
58. J. A. Salmon, *Adv. Drug Res.* **15** (1986) 111.
59. M. G. Baumgärtner, *Deutsche Apotheker-Zeitung* **137** (1997) 2157.
60. L. J. Marnett, A. S. Kalgutkar, *Curr. Opin. Chem. Biol.* **2** (1998) 482.
61. D. O. Stichtenoth, H. Zeidler, J. C. Fröhlich, *Med. Klin.* **93** (1998) 407.
62. J. R. Vane, R. M. Botting, *Inflamm. Res.* **47** (1998) Suppl. 2, S78.
63. W. Elger et al., *J. Steroid Biochem.* **25** (1986) 835.
64. G. Anderson, *Trends Pharmacol. Sci.* **11** (1990) 348.
B. J. Lipworth, *Lancet* **353** (1999) 57.
S.-E. Dahlén, *Eur. J. Pharmacol.* **533** (2006) 40.
65. S. J. Konturek, *Scand. J. Gastroenterol.* **25** (1990) Suppl. 174, 15.
66. A. Aly, *Scand. J. Gastroenterol.* **22** (1987) Suppl. 137, 43.
67. W. F. Stenson, *Falk Symp.* **46** (1988) 143 (CA 109, 52 485 a).
68. P. G. Adaikan, S. R. Kottegoda, *Drugs Future* **10** (1985) 765.
69. S. E. Hall, *Med. Res. Rev.* **11** (1991) 503.
70. A. L. Willis, D. L. Smith, *Eicosanoids* **2** (1989) 69.
71. C. J. Lote, J. Haylor, *Prostaglandins Leukotrienes Essential Fatty Acids* **36** (1989) 203.
72. F. A. Fitzpatrick, R. C. Murphy, *Pharmacol. Rev.* **40** (1988) 229.
73. K. Hoshi, Y. Mizushima, *Prostaglandins* **40** (1990) 155.
74. H. Yamazaki et al., *Adv. Prostaglandin Thromboxane Leukotriene Res.* **19** (1989) 289.
H. Shirahase et al., *J. Cardiovasc. Pharmacol.* **10** (1987) 517.
75. O. Hayaishi, *J. Biol. Chem.* **263** (1988) 14 593.
76. K. V. Honn, B. Cicone, A. Skoff, *Science (Washington D.C.)* **212** (1981) 1270.
77. A. M. Fulton, *Prostaglandins Leukotrienes Essential Fatty Acids* **34** (1988) 229.
M. R. Schneider, M. Schirner, *Drugs Future* **18** (1993) 29.
M. Schirner, *Wien. Klin. Wochenschr.* **107** (1995) 261.
M. R. Schneider, D. G. Tang, M. Schirner, K. V. Honn, *Cancer Metastasis Rev.* **13** (1994) 349.
Drugs & Therapy Perspectives **8** (1996) 1.
M. Hejna, M. Raderer, C. C. Zielinski, *J. Natl. Cancer Inst.* **91** (1999) 22.
78. A. Alm, *Curr. Opin. Ophthalmol.* **4** (1993) no. 2, 44.
79. F. Meier, E. Gross, K.-M. Klotz, T. Ruzicka, *Skin Pharmacol.* **2** (1989) 61.
80. E. Moilanen et al., *Agents Actions* **28** (1989) 290.
81. E. J. Corey, J. Vlattas, K. Harding, *J. Am. Chem. Soc.* **91** (1969) 535.
E. J. Corey et al., *J. Am. Chem. Soc.* **93** (1971) 1491.
82. C. J. Sih et al., *J. Am. Chem. Soc.* **97** (1975) 865.
J. B. Heather et al., *Tetrahedron Lett.* 1973, 2313.
83. R. Noyori, M. Suzuki, *Chemtracts. Org. Chem.* **3** (1990) 173.
M. Suzuki, A. Yanagisawa, R. Noyori, *J. Am. Chem. Soc.* **110** (1988) 4718.
84. E. J. Corey, G. E. Keck, I. Székely, *J. Am. Chem. Soc.* **99** (1977) 2006.
R. A. Johnson et al., *J. Am. Chem. Soc.* **99** (1977) 4182.
N. Whittaker, *Tetrahedron Lett.* 1977, 2805.
I. Tömösközi, G. Galambos, V. Simonidesz, G. Kovács, *Tetrahedron Lett.* 1977, 2627.
K. C. Nicolaou et al., *J. Chem. Soc. Chem. Commun.* 1977, 630.

85. S. S. Bhagwat, P. R. Hamann, W. C. Still, *J. Am. Chem. Soc.* **107** (1985) 6372.
86. E. J. Corey, A. Marfat, G. Goto, F. Brion, *J. Am. Chem. Soc.* **102** (1980) 7984.
87. N. Galie, A. Manes, A. Branzi, *Clin. Chest Med.* **22** (2001) 529.
88. C. D. Funk, *Nat. Rev. Drug Discovery* **4** (2005) 664.
89. M. Ohno et al., *J. Med. Chem.* **48** (2005) 5279.
90. M. Suzuki, R. Noyori, B. Langström, Y. Watanabe, *Bull. Chem. Soc. Jpn.* **73** (2000) 1053.
M. Suzuki et al., *Tetrahedron* **56** (2000) 8263.
91. K. Nagata, H. Hirai *Prostaglandins, Leukotrienes and Essential Fatty Acids* **69** (2003) 169.
T. W. Ly, K. B. Bacon *Expert Opin. Investig. Drugs* **14** (2005) 769.
92. B. McMahon, C. Godson, *Am. J. Physiol. Renal Physiol.* **286** (2004) F189.
A. T. Gewirtz, *Curr. Opinion Invest. Drugs* **6** (2005) 1112.
M. J. Mangino et al., *Prostaglandins & Other Lipid Mediators* **79** (2006) 84.
Prostaglandins, Leukotrienes and Essential Fatty Acids **73** (2005) issues 3-4, 139–322.
93. D. Wang, R. N. DuBois, *Gut* **55** (2006) 115.
94. J. Adams, B. J. Fitzsimmons, Y. Girard, Y. Leblanc, J. F. Evans, J. Rokach, *J. Am. Chem. Soc.* **107** (1985) 464.

Protective Gases

Wladimir Toman, Air Liquide, Düsseldorf, Federal Republic of Germany
Friedrich Grelle, Air Liquide, Düsseldorf, Federal Republic of Germany
Heike Menges, Air Liquide, Düsseldorf, Federal Republic of Germany
Stefan König, Air Liquide, Düsseldorf, Federal Republic of Germany

1. Introduction 1
2. Protective Gases in Safety Technology 1
3. Protective Gases in Food Technology 4
4. Protective Gases in Metallurgy and Smelting 5
4.1. Introduction of Gases through Permeable Elements in Metallurgical Vessels 5
4.2. Protection of Molten Metals from the Ambient Atmosphere 8
4.3. Use of Protective Gases in the Converter 8
4.4. Use of Protective Gases in Heat Treatment of Metallic Components 9
5. Protective Gases for Gas-Shielded Arc Welding Processes 9
6. References 14

1. Introduction

The term protective gas describes a gas that does not react with its environment (i.e., remains inert). This property is often utilized to protect materials from the ambient atmosphere, usually from the effects of oxygen. Protective gases are also used as carriers for reactive substances. Argon, nitrogen, or carbon dioxide can be used. Argon can be considered the best protective gas because it is completely inert and does not react with or form compounds with other elements. However, argon is considerably more expensive than nitrogen or carbon dioxide, and is therefore used only where materials could react with these gases or where the amounts of gas needed are so small that price is not a determining factor in the choice of protective gas.

Users of protective gases either produce them in gas generators or have them delivered by a manufacturer. Depending on the type of use, protective gases are supplied in high-pressure bottles on mobile pallets, in bundles, or for larger volume requirements, in liquid form via mobile tankers.

For very large consumption, nitrogen can be produced by membrane separation or pressure-swing absorption (PSA) plants installed in the user's factory and operated by the gas supplier. When the supplier delivers gases or operates the production plant, he assumes responsibility for quality, for matching supply to demand, and for the composition of the protective gas.

2. Protective Gases in Safety Technology

In safety technology, protective gases are used in containers, vessels, silos, or pipelines that contain flammable gases, liquids, or dusts, to lower the oxygen content of the atmosphere until it is below the explosion limits for the application in question, so that fires and explosions are prevented [1–3].

Explosive limits for various substances are listed in Tables 1 and 2.

Gaseous nitrogen or carbon dioxide is used, the choice depending on the type of application environment and the flammable material in question.

The type of inerting equipment used also depends on the character of the flammable material and the conditions of the gas space to be inerted.

A carbon dioxide inerting installation for silos is shown in Figure 1. Nitrogen is used in large quantities in mines for preventing and fighting underground fires. Here, liquid nitrogen is

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_275

Table 1. Upper and lower explosion limit, flash point, and ignition temperature of gas – air mixtures [1]

Compound	Formula	Explosion limit, vol %		Flash point, °C	Ignition
		Lower	Upper		temperature, °C
Acetaldehyde	CH_3CHO	4.0	57.0	−27	140
Acetone	CH_3COCH_3	2.5	13.0	−19	540
Acetylene	C_2H_2	2.3	82.0		305
Acrolein	CH_2CHCHO	2.8	31.0		
Ammonia	NH_3	15.0	28.0		630
Benzene	C_6H_6	1.2	8.0	−11	555
Blast furnace gas		36.0	72.0		825
1,3-Butadiene	C_4H_4	1.1	12.5		415
Butane	C_4H_{10}	1.8	9.0		410
n-Butanol	C_4H_9OH	1.4	11.3	+35	340
Carbon disulfide	CS_2	1.0	50.0		<156
Carbon monoxide	CO	12.5	74.0		560
Diethyl ether	$(C_2H_5)_2O$	1.7	48.0	−40	170
Ethane	C_2H_6	3.0	15.5		515
Ethanol	C_2H_5OH	3.5	15.0	+12	425
Ethylene	C_2H_4	2.7	34.0		425
Ethylene oxide	C_2H_4O	2.6	100		440
Ethyl nitrate	$C_2H_5NO_3$	4.0	100	+10	85
Fuel oil		0.6	6.5	+55	220
Gasoline		1.0	7.0		250
n-Hexane	C_6H_{14}	1.2	7.4	−20	240
Hydrogen	H_2	4.0	75.6		560
Hydrogen sulfide	H_2S	4.3	46.0		290
Isopentene	C_9H_{10}	1.4	8.7		
Isopropanol	C_9H_7OH	2.0	12.0	+12	425
Methane	CH_4	5.0	15.0		595
Methanol	CH_3OH	5.5	40	+11	455
Naphthalene	$C_{10}H_8$	0.9	5.9	+80	520
n-Pentane	C_9H_{12}	1.43	8.0		476
Natural gas		5.0	15.0		645
Phenol	C_6H_5OH			+79	605
Propane	C_3H_8	2.1	9.5		470
Propene	C_3H_6	2.2	9.7		
Styrene	$C_4H_5CHCH_2$			+32	490
Toluene	$C_6H_3CH_3$	1.2	7.0	+ 6	535
Town gas and grid gas		6	36		560
Water gas	$CO+H_2+N_2$	7	70		ca. 540
o-Xylene	$C_6H_4(CH_3)_2$	1.0	7.6	+30	465

brought to the fire and evaporated in situ. If an inert gas is to be stored where it will be used in an emergency, the storage loss rates for different gases in various quantities must be taken into account:

1) Nitrogen can be stored without loss only if it is kept in cylinders or cylinder bundles
2) Carbon dioxide is the most suitable gas for large installations because it can be stored without loss even in large quantities due to the nature of CO_2 storage equipment

Nitrogen can be taken in gaseous form directly from the cylinders. However, the usage rate for gaseous carbon dioxide is restricted because it is present in the cylinder in liquid form, the rate of removal of the gas being limited by the pressure drop in the bottle during vaporization of the liquid. If the pressure falls below 5.18 bar (0.518 MPa), CO_2 can exist only in solid or gaseous form. If CO_2 is removed as the liquid phase to increase the rate, evaporation of liquid CO_2 produces dry ice (solid CO_2). The solid particles can generate static electricity and cause ignition of the flammable or explosive atmosphere. Therefore, only gaseous carbon dioxide should be used for inerting purposes.

If a protective gas is to be used continuously in large amounts (e.g., in pneumatic conveying equipment or for blanketing process plant), PSA or membrane plants can be used to lower the oxygen content of the ambient atmosphere to the required concentration.

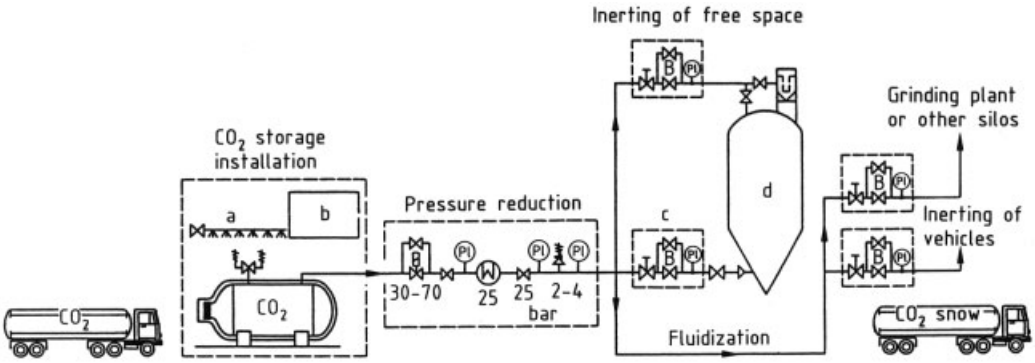

Figure 1. Inerting with carbon dioxide [4]
a) Sprinkler; b) Control unit c) Metering unit; d) Silo

Inerting should be controlled by monitoring oxygen concentration to ensure that it is kept below the lower explosion limit.

Table 2. Limiting oxygen concentration for inerting various air – powder mixtures with nitrogen at ca. 20 °C and 0.1 MPa total pressure in a 1-m^3 vessel with an explosion energy of 10 kJ [2]

Type of dust	Mean particle size, µm	Limiting oxygen concentration, vol %
Aluminum	22	5
Barium stearate	< 63	13
Bisphenol A	34	9
Cadmium laurate	< 63	14
Cadmium stearate	< 63	12
Calcium stearate	< 63	12
Cellulose	22	9
Coal	17	14
Dibenzoyl peroxide	59	10
β-Naphthol	< 30	9
Herbicide	10	12
Maize starch	17	9
Methyl cellulose	70	10
Organic pigment	< 10	12
Paraformaldehyde	23	6
Pea flour	25	15
Polyacrylonitrile	26	11
Wood	27	10
Polyethylene (high density)	26	10
Rye flour	29	13
Soot	13	12
Synthetic resin	63	10
Wettable sulfur with 10 % lignosulfate	30	7
Wheat flour	60	11

The method of controlling the flow rate of protective gas varies with the particular application. Smoldering combustion in coal dust silos is best detected by continuously measuring the CO content in the free space at the top of the silo. Temperature monitoring is often used in this application, but the difficulty is to be sure that the measurement point is correctly located.

In mines, control is based on the determination of methane gas.

Inerting can be effected by various methods: mixing, displacement, and pressure change. In the last of these processes, a vessel can be pressurized with inert gas, the pressure then being reduced to atmospheric, or the vessel can be partially evacuated and then refilled with inert gas. The vessel must be capable of withstanding the applied pressures.

The process of inerting by the mixing and displacement methods depends on the type of gas space to be treated and the nature of the explosive atmosphere.

The displacement method requires the largest possible density difference between the gas to be displaced and the protective gas.

Investigations have shown that silos containing pulverized brown coal can be inerted with CO_2 added via fluidizing equipment. To displace oxygen from the charge, the free space should be inerted first [5].

According to TRD 413 (German regulation for pressurized vessels), explosive dust – air mixtures in brown coal silos are prevented if the oxygen concentration is < 12 vol %. Experiments have shown that this may be achieved by the addition of 2 kg of CO_2 per cubic meter of free space in the silo. When choosing an inert gas, the change in the limiting oxygen concentration with increasing temperature must be considered. Investigations with brown coal have shown that this limit is reduced by ca. 1.3 vol % per

100 °C temperature increase when inerting with nitrogen [6]. Carbon dioxide has been shown to be the more effective inerting gas because it gives a reduction in limiting oxygen concentration of only 0.5 vol % per 100 °C temperature increase.

Table 3. Use of protective gases in the food industry

Process	Gas used			
	N_2	CO_2	O_2	Mixture
Packaging				
Dairy products	+	+	–	N_2–CO_2
Fresh meat	–	–	–	N_2–CO_2–O_2
Sausage and meat products	–	–	–	N_2–CO_2
Fish and seafood	–	–	–	N_2–CO_2
Bakery products	–	+	–	N_2–CO_2
Farinaceous products (pasta)	–	–	–	N_2–CO_2
Vegetables	+	+	+	N_2–CO_2
Ready-to-serve meals	–	–	–	N_2–CO_2
Dried food	+	+	–	N_2–CO_2
Cold grinding (e.g., spices)	+	+	–	–
High-pressure extraction (e.g., coffee)	–	+	–	–
Controlled-atmosphere storage (e.g., fruit, cheese)	+	+	–	–
Silo storage (e.g., cereal, flour)	+	–	–	–
Inerting (e.g., storage tanks and pressure tanks)	+	+	–	–
Sparging (e.g., fruit juice)	+	–	–	–
Carbonated beverages (e.g., beer, sparkling wine)	–	+	–	–
Tapping/dispensing (e.g., beer)	+	+	–	N_2–CO_2

3. Protective Gases in Food Technology

Carbon dioxide, nitrogen, and in special cases, oxygen, or mixtures thereof, are used as protective gases in the preparation, packing, storage, and transport of solid and liquid foodstuffs (see Table 3). The gases inhibit or eliminate various changes that can take place in foods. These changes are:

1) Chemical (oxidation and discoloration)
2) Biological (enzymes already present or generated by microorganisms)
3) Microbial (bacteria, yeasts, molds)
4) Physical (changes in aroma, volume instability, moisture pickup)

These effects are very important in products of this type because their characteristic properties (fat content, aroma, color, and consistency) are very sensitive to oxidation [7–9].

The protective gases are odorless, tasteless, and not harmful to health, and need not be declared.

Carbon dioxide is soluble in water and fat, has bacteriostatic and fungistatic properties, and displaces oxygen. At low temperature and high pressure, it reacts readily with water to form H_2CO_3, which has bactericidal and antimicrobial activity due to its acidity. When moist products are packaged in an atmosphere of pure CO_2, a partial vacuum can be formed by combination of CO_2 with the water present [7], [10], [11].

Nitrogen is used to replace oxygen, and is chemically and biologically inert, although it does have some microbiological activity, mildly inhibiting the metabolism of aerobic bacteria.

Oxygen, mixed with nitrogen and CO_2, is used in the packaging of fresh meat to conserve oxymyoglobin. This is permitted by the German Meat Regulations (Appendix 1) and need not be declared.

Carbon dioxide and nitrogen are also used as cooling media in cryogenic chilling and freezing plants, maintaining an oxygen-free and virtually sterile atmosphere.

The two processes for packaging under a protective gas (Fig. 2) are

1) Flushing with gas in vertical or horizontal continuous-tube bagging machines
2) Replacement of atmospheric air in deep-drawing or chamber machines

In the gas-flushing process,the air in the packaging is replaced continuously by flushing with the protective gas, the gas consumption being

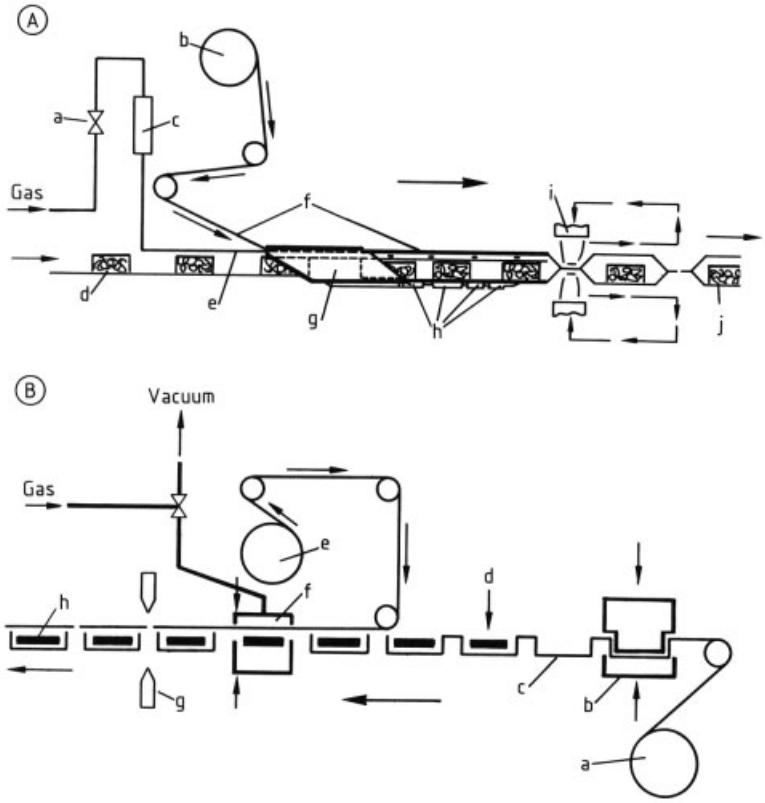

Figure 2. Processing for packaging under protective gas
A) Gas flushing (continuous-tube bagging machine)
a) Valve; b) Roll of plastic film; c) Flow meter; d) Product; e) Gas tube; f) Packaging film; g) Forming channel; h) Longitudinal sealing rolls; i) Transverse sealing bars; j) Product, packed under gas
B) Compensated vacuum (thermoforming machine)
a) Lower, thermoformable film; b) Thermoforming mold; c) Container; d) Addition of product; e) Upper film; f) Vacuum chamber; g) Cutting tool to separate containers; h) Product

three to five times the bag volume. The machine forms a tube of plastic film into which the gas is fed through a nozzle. The end of the gas nozzle is near the point at which the hot welding is carried out. A cushion-shaped package is formed.

In the air replacement system in chamber machines, the product is placed in a bag, evacuated in a chamber, refilled with the gas, and finally heat-sealed. In chamber machines, gas consumption is 0.5 times the package volume. Thermoforming machines for making semirigid boxes are provided with two rolls of plastic sheet. The lower, thermoplastic sheet is drawn to form a container in which the product is placed. The package is evacuated in the vacuum chamber; the gas is injected; and the package is hermetically sealed, the upper sheet forming a closed lid. Gas consumption is 0.2 – 0.4 times the package volume [11].

When large amounts of dissolved oxygen are present in the product (e.g., in granular products, products consisting of porous pieces, or fruit juices with a high fruit content), special equipment and processes are necessary: for example, injection equipment linked to the bagging machines, gas purging of filling nozzles, or sparging.

4. Protective Gases in Metallurgy and Smelting

4.1. Introduction of Gases through Permeable Elements in Metallurgical Vessels

Inert gases are blown into metal melts to produce a physical effect (degassing) or mechanical effects (agitation and purging).

Degassing. The partial pressure of the gases dissolved in the melt is reduced by blowing an inert gas into the melt. The dissolved gases diffuse into the rising bubbles of the sparging gas and are removed from the melt.

Purging. The inert gases rising through the melt cause flotation of nonmetallic inclusions such as oxide or slag particles. The bubbles are extremely small and adhere to the suspended particles, carrying them to the surface of the bath where they are removed as slag or dross.

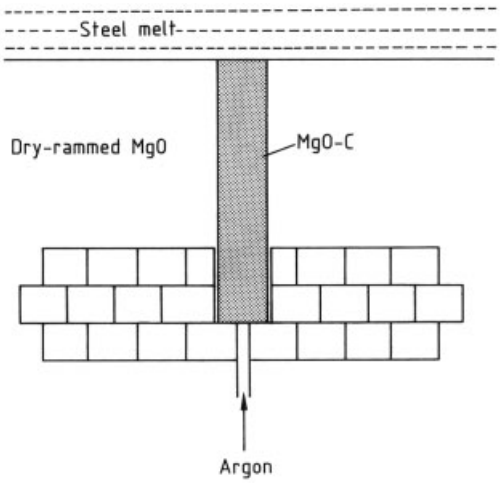

Figure 3. Permeable stirring element in a vessel or furnace bottom

Agitation. Intensive and thorough mixing of the melt is required when metal treatment agents or alloying elements are added that must react with the melt, or to homogenize the temperature distribution and chemical composition of the melt.

Sparging aluminum melts with inert gases is used mainly to lower the hydrogen content, thereby decreasing the porosity of the casting.

When aluminum is melted in foundries in crucible furnaces it is sparged with argon or nitrogen introduced through permeable lances immersed in the melt. The sparging time is ca. 10 min, and 0.2 – 0.5 m^3 of inert gas is used per tonne of aluminum.

The spinning nozzle inert flotation (SNIF) refining process is used for sparging larger aluminum melts. After being melted in a smelting furnace, molten aluminum is transferred to the SNIF furnace where hydrogen and nonmetallic impurities are removed. The melt in the furnace is kept in motion by one or two rotating nozzles through which argon is injected, causing degassing. Argon consumption is ca. 0.4 m^3 per tonne of aluminum.

The SIGMA process is also used for sparging aluminum melts. Here, inert gases are introduced into the reactor through a bottom plate made of a permeable refractory material. A baffle plate in the reactor causes intensive mixing of the metal and the gas, and increases the dwell time of the aluminum [12].

If nitrogen is used, aluminum nitride can form, and argon is therefore preferred. The argon consumption is ca. 1.5 m^3 per tonne of aluminum. To remove lithium from aluminum melts, the sparging gas consists of a mixture of nitrogen or argon with 25 – 50 % sulfur hexafluoride as the reactive gas.

The degassing of steel melts is generally carried out under vacuum. The ladle is filled with molten steel and placed in a vacuum chamber. The inert gas is introduced through a permeable ceramic plate in the base of the ladle, and acts as both a sparging and an agitating gas.

Alternatively, a porous ceramic plug can be built into the floor of small induction furnaces (0.5 – 5 t), which are used to produce alloy and high-alloy casting steels. Stirring with an inert gas removes dissolved impurities, such as deoxidation products, from the steel melt [13]. Sparged steels show a considerable increase in linear extension and necking. The quantity of impurities detectable by micrography can be reduced by > 70 %, so that the weldability of these steels is improved considerably. The stirring time is 15 – 20 min with an argon consumption of 1 –1.5 m^3 per tonne of steel.

Porous plugs are also used in electric-arc steel melting furnaces [14]. Stirring of the melt leads to

1) Acceleration of the reaction between the metal and the slag
2) Homogenization of the temperature and chemical composition of the melt
3) Better heat transfer between the arc and the melt

The stirring elements (see Fig. 3) are usually built into the bottom of the electric-arc furnace. These are carbon – magnesite elements with several small steel tubes. The argon consumption is ca. 0.2 m^3 per tonne of steel. The sparging process leads to energy savings of 10 – 15 kW per tonne of steel and a reduction in the consumption of steel scrap of ca. 5 kg per tonne of steel. An improvement in purity also occurs [14].

In steelworks, both before and during pouring of the steel an inert gas is injected through permeable blocks in the bottom of the ladle to agitate the steel and homogenize the melt. Temperature homogenization is particularly important for the continuous casting process because temperature variations can lead to problems in pouring the melt and to tearing of the metal strand. In continuous casting plants, the tundish offers the last possible opportunity for removal of inclusions from the melt before it solidifies; gas injectors can also be built into the base of the tundish. Argon consumption is 0.05 – 0.2 m^3/t. Two injectors are used for each tundish (Fig. 4) [15].

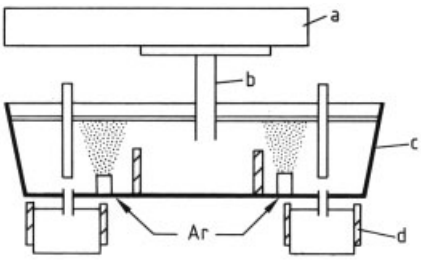

Figure 4. Tundish with injectors
a) Ladle; b) Shroud; c) Tundish; d) Mold

In the production of nodular cast iron the melt can be agitated with a powerful jet of nitrogen that is used to mix in calcium carbide, which reacts with sulfur to form calcium sulfide, thereby desulfurizing the melt [16]. Agitation by a gas stream can also be used for carburizing, magnesium treatment, or inoculation, etc., which can be carried out consecutively in the same ladle. This process can also be used for continuous desulfurization in a cupola furnace. A permeable block is built into a treatment vessel located between the cupola and the holding furnace.

4.2. Protection of Molten Metals from the Ambient Atmosphere

When a melt is poured from one metallurgical vessel into another, a large amount of the metal's surface area comes in contact with the surrounding atmosphere. The metal stream absorbs considerable quantities of oxygen by chemical reaction, so the purity of the metal and its working properties are impaired. In the continuous casting of steel, contamination of the molten metal is prevented by fitting a shroud, swept by inert gas, between the tundish and the mold. Another system consists of a movable shielding device fitted underneath the tundish, surrounding the metal stream for almost its entire length. A stream of inert gas flows between the device and the metal stream. The metal stream between the ladle and the tundish can be protected by two methods: a ceramic shroud filled with inert gas or a hood. Shrouding the pouring operation prevents oxidation of the metal stream in each case. Some steelworks prevent oxidation during pouring by providing the tundish with a cover. The free space between tundish and cover is filled with argon, enabling the metal to be poured into the tundish in a neutral atmosphere. If some degree of nitriding is acceptable, nitrogen can be used as the protective gas.

Purging of vessels to remove air before filling with metal can also be carried out by using a liquefied inert gas. The liquefied gas settles at the bottom of the vessel where it vaporizes and displaces the air in the vessel. Vaporization proceeds relatively slowly, due to the formation of a gas film between the hot refractory material and the liquefied gas, which acts as an insulating layer to prevent rapid vaporization of the liquefied gas. The metal is then poured into the vessel filled with inert gas. In steel foundries, induction furnaces for the production of alloy and high-alloy steels are inerted with a blanket of liquefied argon during melting [17], [18]. Liquefied argon vaporizes on contact with the hot metal, producing a protective gas atmosphere that blankets the hot metal and prevents oxidation.

In powder metallurgy, liquid metals are powdered by atomization with argon or nitrogen. The use of inert gases ensures that the metal surface is not oxidized. This is not the case when water is used.

Minimization of oxidation can also be used to reduce dust emission. Procedures aimed at reducing smoke generation are being tested in blast furnaces [18]. So-called brown smoke, mainly vaporized iron oxide, is produced when a large surface of molten iron is exposed to oxidation by atmospheric oxygen during teeming or at the pig iron spout. In these cases, oxidation can be prevented by a blanket of inert gas (carbon dioxide or nitrogen). Carbon dioxide is added to the molten iron in the form of dry ice. Due to its higher density, dry ice resists the buoyancy of the hot gases better than nitrogen, comes in closer contact with the metal surface, and isolates it from the ambient atmosphere. Furthermore, nitrogen can lead to nitride formation in the pig iron, which can impair the quality of the steel. The route of the raw iron from the tapping hole to the filling point of the transport container must be enclosed by a cover that is as leakproof as possible [19]. When protective gases are used to reduce smoke production, the dedusting equipment can be of smaller dimensions or can be eliminated entirely, with corresponding reductions in the cost of disposing of the metal oxide dust.

If air is allowed free access to molten iron or steel, nitrogenization of the metal can occur due to reaction with atmospheric nitrogen. For certain grades of steel, the nitrogen content must be kept low to prevent impairment of steel quality. The largest exposure of the metal surface to the atmosphere occurs during teeming or on tapping of the converter or electric-arc furnace into the transporting ladle. The metal stream can be protected by using CO_2 snow, which is produced by special snow-generating equipment [20].

4.3. Use of Protective Gases in the Converter

In steel production, molten pig iron is refined by removing carbon by treatment with ca. 50 – 60 m^3 of oxygen per tonne of iron. The oxygen must be evenly distributed in the melt with a high mixing energy. Injectors or bottom blowers can be located in the floor of the converter. Nitrogen or argon passes through these and causes vigorous mixing of the melt (Fig. 5), homogenizing its temperature and chemical composition. Inert gas consumption is 1 – 3 m^3 per tonne of steel. After this treatment, the melt is tapped into a pouring ladle, where metallurgical treatment is carried out. Here, inert gases are also used as carriers for the addition of alloying materials via a blowing lance, or for mixing and homogenizing the steel melt by sparging. Large quantities of inert gas are used in the production of stainless and heat-resistant steels in the argon – oxygen decarburization (AOD), Klöckner combined blowing (KCB), and metal refining process lance (MRP-L) methods. Inert gases (argon and nitrogen) are mixed with the blowing oxygen to lower the partial pressure of carbon monoxide. The resulting thermodynamic conditions promote the oxidation of carbon in preference to oxidation of chromium [21]. The amounts of argon depend on the types of steel produced (15 – 25 m^3 per tonne of steel).

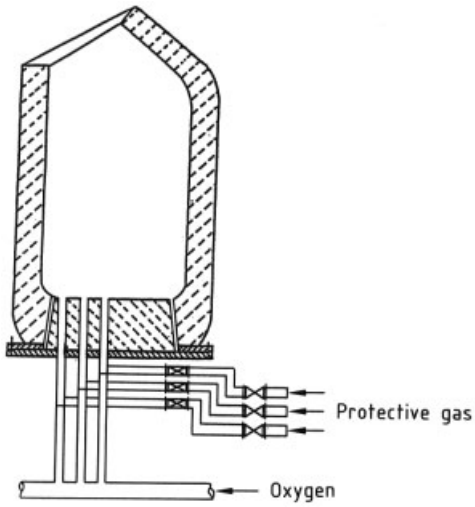

Figure 5. Bottom blown quick basic oxygen process (Q-BOP) converter

4.4. Use of Protective Gases in Heat Treatment of Metallic Components

Heat treatment is used to modify the properties of solid metallic components by changes in temperature and by the action of the ambient atmosphere.

Table 4. Protective gases for heat treatment of metals [22] *

Material	Breaking vacuum	Bright annealing	Decarburizing annealing	Carbon-neutral annealing	Carburizing	Nitriding	Carbonitriding	Sintering	Tempering	Brazing	Conversion of residual austenite
Steel $<0.3\,\%$ C		N_2 N_2+H_2 N_2+HC	N_2+H_2 $(+H_2O)$	N_2 N_2+HC	N_2+HC $(+O_2)$	N_2+NH_3	N_2+HC $+NH_3$	N_2 N_2+H_2		N_2 N_2+H_2	
Steel $>0.3\,\%$ C	N_2 Ar	N_2 N_2+H_2 N_2+HC	N_2 N_2+H_2 $(+H_2O)$	N_2 N_2+HC	N_2+HC $(+O_2)$	N_2+NH_3	N_2+HC $+NH_3$	N_2 N_2+H_2 N_2+HC		N_2 N_2+H_2	liquid N_2
Cr and Mn steels	N_2 Ar	N_2 N_2+H_2 N_2+HC	N_2+H_2 H_2 $(+H_2O)$	N_2 N_2+HC	N_2+HC N_2+HC $(+O_2)$	N_2+NH_3	N_2+HC $+NH_3$			N_2+H_2 $Ar+H_2$	liquid N_2
Stainless and heat-resistant steels	N_2 Ar	N_2+H_2 H_2 Ar	N_2+H_2 H_2, $Ar+H_2$ $(+H_2O)$	N_2 Ar		N_2+NH_3		N_2+H_2 $Ar+H_2$ H_2		N_2+H_2 $Ar+H_2$ H_2	liquid N_2
SG iron		N_2		N_2+HC							
Blackheart malleable iron		N_2		N_2+HC					N_2		
Aluminum and its alloys		N_2 N_2+H_2						N_2+H_2 H_2			
Copper, nickel		N_2 N_2+H_2						N_2 N_2+H_2		N_2 N_2+H_2	
Bronze, brass, nickel, silver		N_2 N_2+H_2						N_2+H_2 H_2		N_2 N_2+H_2	
Noble metals	N_2 Ar	N_2 N_2+H_2								N_2 N_2+H_2	

* HC = hydrocarbon.

The protective gas atmosphere not only protects metallic products from oxidation (heat treatment, sintering, hardening), but also changes the chemical composition of the surfaces by diffusion (e.g., of nitrogen and carbon in nitriding, carbonitriding, and carburizing processes). To achieve this, the carrier gas (nitrogen) is mixed with hydrocarbons and ammonia. Moisture and oxygen from the components being treated and from leaks in the heat treatment furnace enter the furnace atmosphere. The resulting change in the composition of the furnace atmosphere could lead to oxidation of the components. Therefore, a proportion of the protective gas atmosphere must have a reducing effect. This is achieved by adding hydrogen. Various gases are used in protective gas atmospheres, depending on the desired properties of the components (see Table 4).

Protective gases and carrier gases produced from hydrocarbon – air mixtures in generators, or from ammonia in ammonia crackers, are increasingly being replaced by protective gases from gas suppliers.

The protective gas atmosphere is mixed by the user to give the desired composition and introduced into the heat treatment furnace. Substitutes for some generator gases can be produced by catalytic reaction of a fuel gas – air mixture and mixing with nitrogen either inside or outside the heat treatment furnace to form a protective gas atmosphere.

5. Protective Gases for Gas-Shielded Arc Welding Processes

All gas-shielded welding processes have one aspect in common: they require a shielding gas to protect the molten welding pool from the atmosphere. Various welding processes utilize inert, oxidizing, reducing, and quasi-inert protective gases, or mixtures of these, depending on the process and the metal being welded (Table 5) [23]. Optimized mixtures of protective gases offer additional advantages and often improve both quality and productivity [24]. Some welding processes do not allow the use of active protective gases, for example, the tungsten – inert gas welding process (TIG), plasma arc welding (PAW), and the infrequently used tungsten – hydrogen welding process (THG), because an active component (CO_2, O_2) would lead to melting and corrosion of the arc-producing electrodes, and inclusions of tungsten and thorium in the weld. In processes involving the melting of wire electrodes [subdivided into metal – inert gas (MIG) and metal–active gas (MAG) processes, depending on the protective gas used] the choice of gas depends mainly on the material to be welded [23]. Various protective gases, inert or active, may be used for the same material in some cases. However, the selection of a suitable or particularly advantageous protective gas requires a knowledge of the effects of individual gases on the welding process. These effects are principally a function of the physical properties of the gas.

Table 5. Purity, dew point, and reactive behavior of protective gases (DIN 32 526)

Gas	Purity, vol % (min.)	Dew point at 1 bar, °C (max.)	Chemical behavior on welding
Argon	99.99	−50	inert
Helium	99.99	−50	inert
Carbon dioxide	99.7	−35	oxidizing
Oxygen	99.5	−35	oxidizing
Nitrogen	99.5	−50	unreactive
Hydrogen	99.5	−50	reducing ∗

∗ Protective gas mixtures with > 10 % H_2 are an explosion hazard and should be flared off.

The welding processes mentioned are not described in detail here; further information can be found in [25–27] (see also → Welding and Cutting).

The gases are most usefully characterized by the following physical properties (see Table 6 and Fig. 6): gas density, thermal conductivity, and ionization energy.

These properties are often temperature dependent. The important temperature range for gas-shielded arc welding lies between ca. 2000 and 8000 °C. Only the plasma arc welding process involves higher temperatures [28].

Gas density has a direct influence on the quantity of gas required for effective protection. A dense protective gas provides better protection than a lighter gas at a given flow rate.

High thermal conductivity has a beneficial effect on the shape of the bead and on the metal in the weld pool as well as in the heat flow zone. The arc column becomes larger, and the bead formed is therefore wider and has good surface

Table 6. Some physical properties of protective gases [28], [29]

Gas	Density at 15 °C, 1 bar, kg/m^3	Density relative to air at 15 °C, 1 bar	Dissociation and ionization energies, eV [a]	Content in air, vol %	Boiling point at 101.3 kPa, °C
Ar	1.669	1.37	15.76 27,50	0.934	−185.9
He	0.167	0.14	24.56 54.10	5.2×10^{-4} [b]	−268.9
CO_2	1.849	1.44	various reactions	0.033 [b]	−78.5 [c]
O_2	1.337	1.04	5.08 13.62 35.20	20.946	−183.0
N_2	1.17	0.91	9.76 14.55 29.60	78.084	−195.8
H_2	0.085	0.06	4.48 13.59	0.5×10^{-4} [b]	−252.9

[a] 1 eV = 1.6×10^{-9} J.
[b] Not obtained from air.
[c] Sublimation temperature.

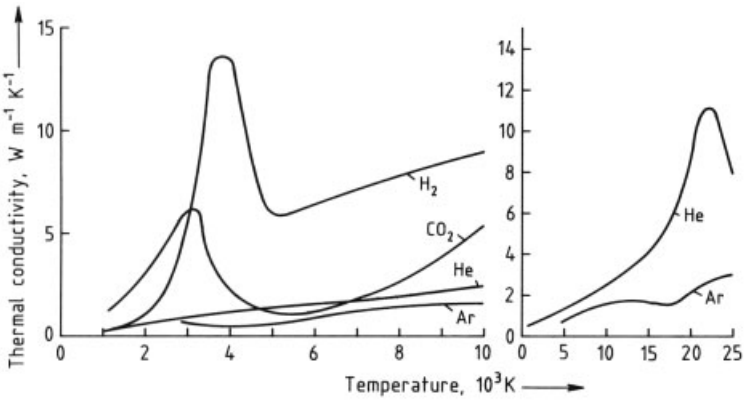

Figure 6. Thermal conductivity of various protective gases [28]

properties [24, p. 152]. However, the thermal conductivity does not vary linearly with temperature, being the result of individual contributions determined by dissociation – ionization processes [30].

The ionization energy of a gas is very important for the electric arc. If only a small amount of energy is required to ionize a gas (i.e., to make it conduct, see Table 6), the arc is easier to strike.

Apart from these important gas properties, the relationship between the electrical conductivity of a plasma and the resistance and voltage of the arc is also significant [28].

Exploitation of these physical properties in welding processes leads to the following application for the various protective gases:

Argon has very favorable physical properties, is produced economically in air separation plants, and is the most widely used protective gas in welding technology. It is odorless, tasteless, and colorless; is available with a purity of 99.996 vol % (4.6 quality); and is used in the welding of high-alloy Cr – Ni steels and Al-, Cu-, and Ni-base alloys by the TIG and PAW welding processes, and also for MIG welding of metals suitable for this process.

Argon is the base gas for the MAG welding of high-alloy Cr – Ni steels, and is also very suitable for unalloyed and low-alloy steels [28]. It has a relatively high density, giving a stable covering of protective gas, and is therefore used in TIG welding of materials that are sensitive to oxygen or nitrogen, such as titanium,

Table 7. Standardized gases and gas mixtures for gas-shielded arc welding [33]

Process	Materials	Gas designation		Composition or purity, wt %	DIN 32526
		AL international	Germany		
MAG (metal–active gas welding)	all unalloyed and low-alloy steels (e.g., structural steels, fine-grained structural steels, tube steels, shipbuilding steels, heat-resistant steels)		Kohlendioxid	CO_2 99.95%	C 1
		Atal 6	Inarc K 18	Ar 82% + CO_2 18%	M 21
		Cargal 4	Inarc S 8	Ar 92% + O_2 8%	M 23
		Teral 23	Inarc KS 5·5	Ar 90% + CO_2 5% + O_2 5%	M 32
		Teral 25	Inarc KS 14·4	Ar 82% + CO_2 14% + O_2 4%	
		Eloxal 35	Inarc KHe 13·5	Ar 82% + CO_2 13% + He 5%	M 13
	high-alloy stainless, acid- and heat-resistant steels, highly heat resistant and cold ductile steels	Cargal 1	Inarc S 1	Ar 99% + O_2 1%	M 11
		Cargal 2	Inarc S 3	Ar 97% + O_2 3%	
		Atal 2	Inarc K 2	Ar 98% + CO_2 2%	M 12
MIG (metal–inert gas welding)	all weldable metals		Argon R	Ar 99.996%	I 1
	aluminum, copper, and their alloys	Inarc 6	Inarc He 30	Ar 70% + He 30%	I 3
		Inarc 9	Inarc He 50	Ar 50% + He 50%	
TIG (tungsten–inert gas welding)	all weldable metals		Argon R	Ar 99.996%	I 1
	gas-sensitive metals		Argon HR	Ar 99.998%	
	high-alloy steels, nickel and nickel alloys	Noxal 2	Inarc H 2	Ar 98% + H_2 2%	R 2
		Noxal 3	Inarc H 5	Ar 95% + H_2 5%	
		Noxal 4	Inarc H 10	Ar 90% + H_2 10%	
			Helium R	He 99.995%	I 2
			Argon R	Ar 99.996%	I 1
			Helium R	He 99.995%	I 2
	aluminum, aluminum alloys, copper, copper alloys	Inarc 6	Inarc He 30	Ar 70% + He 30%	I 3
		Inarc 9	Inarc He 50	Ar 50% + He 50%	
Plasma arc welding	all unalloyed, low-alloy, and high-alloy steels		Argon HR	Ar 99.998%	I 1
			Helium R	He 99.995%	I 2
	high-alloy steels, nickel, and nickel alloys	Noxal 3	Inarc H 5	Ar 95% + H_2 5%	R 2
		Noxal 4	Inarc H 10	Ar 90% + H_2 10%	
	nickel, aluminum, copper, structural steels	Inarc 6	Inarc He 30	Ar 70% + He 30%	I 3
		Inarc 9	Inarc He 50	Ar 50% + He 50%	
Back shielding (root protection)	all unalloyed, low-alloy, and high-alloy steels		Formiergas 10	N_2 90% + H_2 10%	F 2
			Formiergas 20	N_2 80% + H_2 20%	

tantalum, and niobium. Very high-purity argon (99.998 vol %, or 4.8 quality) is recommended for this application because the metals have a high affinity for N_2 and O_2. Gas suppliers also add other gaseous components (see Table 7), of-fering the user standardized two- or three-component gas mixtures that ensure problem-free, cost-effective welding.

Helium has for a long time played only a minor role in welding technology because of its high cost. However, helium is now being used in increasing quantities in certain welding applications where its high price is compensated by its special physical properties. Mixtures of 30 – 50 % helium with argon are used for welding materials of high thermal conductivity (e.g., Al, Cu, and their alloys) by the TIG, MIG, and PAW processes. Because of helium's high thermal conductivity, the need to preheat the workpieces, especially those of heavy section, is reduced and sometimes completely eliminated. This property also produces a larger, hotter fusion zone, which means that more time is available for degassing, and porosity is therefore considerably reduced. The high ionization energy of helium causes problems in striking the arc, but also leads to a high energy of recombination, which is largely returned to the welding pool. This results in deeper penetration and increased welding speed and deposition rate. Helium is also used in the MAG welding process. Addition of 50 % helium to the standard mixture of 2 % CO_2 in argon has been found to be suitable for high-alloy Cr – Ni steels, and addition of 5 % to the standard mixture of argon with 18 % CO_2 is used in flux- cored wire welding of unalloyed mild steels [28], [31]. Pure helium is used in automated TIG welding of aluminum, copper, and nickel components, particularly if preheating cannot be carried out. Since the density of helium or helium-rich mixtures is lower than that of argon, higher gas flow rates are required.

Oxygen is used as a minor oxidizing additive to argon mixtures to improve welding by the MAG process. It leads to the formation of metallic oxides, which lower the surface tension of the molten weld pool. Smaller droplets of weld metal are produced that become detached quickly and smoothly, leading to reduced spattering. In practice, a 1 – 3 % oxygen content for welding Cr – Ni steels is suitable, or 8 % oxygen in argon for welding structural steel [32].

Carbon dioxide has a high density, good thermal conductivity, and energy of recombination originating from the dissociation reaction that takes place in the electric arc ($CO_2 \rightleftharpoons CO + O$). Hence, it is the most important oxidizing component of mixtures for the MAG process for unalloyed and low-alloy steels, as well as for high-alloy Cr – Ni steels. Some standard mixtures are listed in Table 7.

Hydrogen is colorless, odorless, nontoxic, and flammable. It has favorable physical properties (e.g., high enthalpy, high thermal conductivity) and a high energy of recombination, leading to its use in the TIG and PAW welding processes. For welding high-alloy Cr – Ni steels, nickel, and nickel alloys, 5 % H_2 is added to argon. Hydrogen is also used as a back shielding gas. As a rule, ca. 10 % hydrogen is added to nitrogen. The reducing action of the hydrogen protects the root zone of welds (mainly on Cr – Ni steels) from undesirable oxidation.

Nitrogen is the main component of back shielding gas. Due to its very low chemical reactivity little danger exists of nitride formation in the relatively cool interior of the root zone. It is not used as a protective gas component in the welding processes considered here.

6. References

1. *Ullmann,* 4. ed., **20,** 698.
2. Internationale Sektion "Maschinenschutz" der IVSS f. d. Verhütung v. Arbeitsunfällen und Berufskrankheiten, Arbeitskreis 6: Staubexplosionen der Sektion "Maschinenschutz" in Zusammenarbeit mit der Sektion "Chemie": REGELN zum Staubexplosionsschutz an Maschinen und Apparaten – Vorbeugende und Konstruktive Schutzmaßnahmen --, Berufsgenossenschaft Nahrungsmittel Mannheim, 1987, pp. 11 14.
3. Internationale Sektion für die Verhütung von Arbeitsunfällen und Berufskrankheiten, Sektion für die chemische Industrie: Staubexplosionen, 2nd ed. IVSS, Heidelberg 1987.

4. Air Liquide GmbH – Agefko, Prospekt W 163.
5. E. Scherrer: "Inertisierungsmaßnahmen an Braun-Kohlenstaub-Silos," *VDI-Ber.* 1984, no. 494, 99 – 103.
6. W. Wiemann: "Einfluß der Temperatur auf Explosionskenngrößen und Sauerstoffgrenzkonzentration," *VDI-Ber.* 1984, no. 494, 89 – 97.
7. W. Weber: "Schutzgase für die Verpackung von Lebensmitteln," *Lebensmitteltechnik* **21** (1989) 411.
8. R. Wohlgemuth: "Vorratsschädlinge und ihre Bekämpfung," *Gordian* **86** (1986) no. 1 – 2, 18.
9. K. D. Krinninger, Die Eigenschaften der Kohlensäure und ihre besondere Bedeutung in der Getränkewirtschaft, *Brauerei-Forum* **3,** 1987, 17.
10. J. Rüdiger: "Schutzgas – Wie wirken die Gase?" *Fleischmagazin* 1988, no. 4, 70 – 71.
11. M. J. Ansoborlo: "Konservierung unter modifizierter Atmosphäre," *Neue Verpackung* 1988, no. 6, 32 – 33.
12. L. Gluns, D. Saha: "Sigma – Ein neues Verfahren zum Spülen von Aluminium," *Erste Duisburger Industriegas-Tage,* 14.–15. Feb. 1990, [1], p. 27.
13. W. Stecher: "Beitrag zur Metallurgie des Schmelzens hochlegierter Stähle im Induktionsofen, "*Gießerei* **72** (1985) no. 6, 133.
14. R. Szezesny et al.: "Bottom Inert Gas Blowing into an Electric Arc Furnace at Vallourec," *3rd European Electric Steel Congress,* Bournemouth, 1. – 4. Oct. 1989.
15. L'Air Liquide, FR 89 08 008, 1989 (F. Weisang, P. Charlery-Adele).
16. J. Galey, J. Foulard, N. Lutgen, W. Toman: "Treatment of Molten Metal by Agitation with Neutral Purge Gases (GAZAL-Process)," *AFS Int. Cast Met. J.* (1980) September, 16 – 21.
17. S. H. Anderson, J. W. Foss, R. M. Nagan, B. S. Shala: "SPAL for the Investment Casters," *AFS Transactions,* **123** (1989) 709 – 722.
18. D. Fünders, DE 3 904 415, 1990.
19. K. Grützmacher et al.: "Staubunterdrückung in Hochofengießhallen," *Stahl Eisen* **111** (1991) no. 3, 51 – 56.
20. L'Air Liquide, FR 196 952, 1986 (J. M. Naud).
21. M. Velikonja: "Neue Techniken der Stahlraffination durch Sauerstoff und Inertgasanwendung," *Erste Duisburger Industriegas-Tage,* 14 – 15 Feb. 1990.
22. *Ullmann,* 4th ed., **20,** 701.
23. H. U. Pomaska: "Schutzgase für das Lichtbogenschweißen und -schneiden," *DVS Ber.* **43** (1977) 42.
24. H. Golke: "Technische Gase in der Schweißtechnik zur Förderung neuer Technologien," *DVS Ber.* **74** (1982) 151.
25. H. J. Fahrenwaldt: *Schweißtechnik – Verfahren und Werkstoffe,* Vieweg & Sohn, Braunschweig 1988.
26. L. Baum, V. Fichter: "*Der Schutzgasschweißer,*" part II: "MIG/MAG-Schweißen," *Die Schweißtechnische Praxis,* vol. 12, DVS-Verlag, Düsseldorf 1990.
27. F. Eichhorn: *Schweißtechnische Fertigungsverfahren,* VDI-Verlag, Düsseldorf 1983.
28. D. Böhme: Welding Gases – Physical Properties, the Basis for the Development and Optimum Application of Shielding Gases and Gas Mixtures, IIW Doc. XII-1197-90.
29. L'Air Liquide: *Gas Encyclopaedia,* Elsevier, Amsterdam 1976.
30. U. Dilthey: "Einfluß von Schutzgasen auf den Werkstoffübergang bei Metall-Schutzgasschweißverfahren," *DVS Ber.* **42** (1976) 160.
31. L'Air Liquide, C. T. A. S.: Recherche d'une Atmosphère de Protection pour Fils Fourrés, Rapport Technique no. 1203, 1986.
32. A. Farwer: "Neue Entwicklungen bei den Schutzgasen zum WIG-, MIG- und MAG-Schweißen," *DVS Ber.* **112** (1988) reprint.
33. Air Liquide GmbH, Prospekt OR 217.

Proteins

Henning Klostermeyer, Forschungszentrum für Milch und Lebensmittel Weihenstephan, Institut für Chemie und Physik, Freising-Weihenstephan, Federal Republic of Germany (Chap. 1)

Horst Schmandke, Prof. Dr. H. Schmandke & Partner GbR, Potsdam, Federal Republic of Germany (Sections 2.1–2.4)

Carl J. Soeder, Forschungszentrum Jülich, Institut für Biotechnologie, Jülich, Federal Republic of Germany (Section 2.5)

Wolfgang Schreiber, Bundesforschungsanstalt für Fischerei, Institut für Biochemie und Technologie, Hamburg, Federal Republic of Germany (Section 3.1)

Jörg Oehlenschläger, Bundesforschungsanstalt für Fischerei, Institut für Biochemie und Technologie, Hamburg, Federal Republic of Germany (Section 3.1)

Siegfried Scholtyssek, Universität Hohenheim, Institut für Tierhaltung und Tierzüchtung, Stuttgart-Hohenheim, Federal Republic of Germany (Section 3.2)

Margarete Kobald, Fraunhofer Institut für Lebensmitteltechnologie und Verpackung, München, Federal Republic of Germany (Section 3.3)

Andreas Sander, Grünau Illertissen GmbH, Illertissen, Federal Republic of Germany (Chap. 4)

Eberhard Eilers, Grünau Illertissen GmbH, Illertissen, Federal Republic of Germany (Chap. 4)

Edith von Kries, Grünau Illertissen GmbH, Illertissen, Federal Republic of Germany (Chap. 4)

1. Introduction 2
2. Plant Protein Products 4
2.1. Seed Protein Products 4
2.1.1. Cereal Protein Products 4
2.1.2. Oilseed Protein Products 8
2.1.3. Legume Protein Products 12
2.2. Protein Products from Roots and Tubers 13
2.3. Leaf Protein Products 15
2.4. Health Aspects 17
2.4.1. Antinutritive Factors in Plant Protein Raw Materials 17
2.4.2. Minimization of Antinutritive Factors in Processing Selected Plant Protein Raw Materials 17
2.4.3. Food Laws 20
2.5. Proteinaceous Products from Microalgae 21
3. Animal Proteins 24
3.1. Fish Protein Concentrates (FPC) 24
3.1.1. Raw Materials 25
3.1.2. Production 25
3.1.3. Nutritive Value 29
3.1.4. Product Properties and Uses 30
3.1.5. Completely Soluble Hydrolysates from Fish Protein 30
3.2. Egg Proteins 30
3.2.1. Structure and Composition of the Egg 30
3.2.2. Proteins in Egg 31
3.2.2.1. Proteins in Egg White 33
3.2.2.2. Proteins in Egg Yolk 35
3.2.3. Egg Products 36
3.3. Proteins from Blood 37
3.3.1. Physical Properties of Blood 37
3.3.2. Production 37
3.3.3. Functional Properties of Blood Proteins 40
3.3.4. Nutritional Value of Blood Proteins 41
3.3.5. Uses 41
3.3.6. Environmental and Legal Aspects . 41
3.3.7. Economic Aspects 41
4. Protein Hydrolysates and Derivatives 42
4.1. Protein Hydrolysates 42
4.1.1. Raw Materials 42
4.1.2. Production Processes 43
4.1.3. Uses 44
4.1.3.1. Nutrition 44
4.1.3.2. Cosmetics 44
4.1.3.3. Technology 45
4.2. Derivatives of Protein Hydrolysates 45
4.2.1. Acylated Products 45
4.2.2. Alkylated Products 46

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_289

4.2.3. Esters 47
4.3. Product Properties 47
4.4. Analysis 47
4.5. Storage and Transportation 47
4.6. Economic Aspects 47
4.7. Toxicology and Ecology 48
5. References 48

1. Introduction

Different values for the average daily requirement of food protein can be found in the literature. Values between 18 and 28 g per 1000 kcal of food intake should be appropriate. The lower the uptake of food energy, the higher must be the percentage of protein in food [1]. Between 1987 and 1989, 70.4 g of edible protein was statistically available daily *per capita* of the world population. About 65 % of these proteins were plant proteins [1961 – 1963: 62.2 g (68 %), 1969 – 1971: 64.4 g (66 %), 1978 – 1981: 67.5 g (66 %)] [2]. Thus, in spite of the rapidly increasing world population, the global supply of proteins has improved rather than deteriorated.

Until 1990, this also applied to the entire nutritional situation. A daily nutrient intake of 2600 kcal per person, which was judged to be adequate and was strived for, was achieved at the end of the 1980s even by China and the densely populated countries of Southeast Asia. Thus, more than half the population of developing countries is now considered to be adequately supplied. At the same time, however, the situation in some countries, especially in sub-Saharan Africa, has substantially deteriorated. In 1990, 220 million people in the poorest countries lived on less than 2000 kcal per day per person. The index for food production *per capita* in 1991 was 119.22 for Asia and 93.55 for Africa and, in comparison, 104.94 for Europe (basis: 1979 – 1981 = 100) [3]. On the other hand, the index for *total food production* in 1990 was 143.68 for Asia, a substantial 125.83 for Africa, and 108.22 for Europe. In other words, in contrast to Asia, the population growth in Africa is increasing faster than food production [4].

Because of the collapse of the former USSR and in the wake of wars, the global nutritional situation clearly deteriorated at the start of the 1990s. In September 1992, the FAO estimated the number of children on protein-deficient nutrition at 116 million.

By applying these statistics, differences in the biological value of individual food proteins do not come into play and a considerable portion of mostly plant proteins is assessed incorrectly. These are primary products used as fodder to obtain higher-quality animal proteins. If the feeding of grain and oilseeds is taken into account, only about 55 % of the originally produced food proteins are available for human nutrition. This proportion has a downward ten-dency (i.e., the increase in plant protein production benefits, in particular, the enhancement of animal proteins in developed countries).

An improvement in the protein supply of those in need could be expected primarily from increasing the cultivation of protein-rich leguminous plants and types of grain in needy countries. This can be achieved by releasing high-grade food proteins by the provision of inexpensive protein for animal feed, and from the development of low- cost plant protein concentrates that are of a high quality from a nutritional, physiological, and functional viewpoint. A gentler approach to natural resources is also required. Until 1989, for instance, worldwide fishing increased continuously to 99.5×10^6 t/a, but since then it has decreased rapidly due to overfishing (1990: 95.5×10^6 t). About 30 % of the fish caught are still processed into feed.

In the past, the development of *unconventional sources of protein* has not been as successful as expected. The marketing of new developments appears to be directed much more at replacing conventional protein sources in regions with high purchasing power than at reaching the needy classes of the world population. The largest financial scope for the cost of improvements exists in the case of expensive egg proteins, followed by meat and fish, while milk protein is by far the cheapest animal protein. However, since the functional imitation of milk protein is relatively easy and cost effective because of a lack of texture and sensory properties, strong competition exists to replace milk proteins.

Milk production represents the largest production reserves for nutritionally and physiolog-

ically valuable animal protein. About 60 % of the agriculturally utilizable land can be optimally cultivated as grassland. The conversion of irradiated solar energy via grassland → ruminants → milk, with meat as a side product, is much more efficient than any other path to animal protein. In fattening animals, however, considerably more high-quality protein is regularly fed to animals than recovered from them. The biological potential of the oceans to produce high-grade food protein is globally limited and, unlike exploitation of the land, can hardly be increased any further.

The *cultivation of grain* represents the most important global source of food protein. Whether cereals can in the long term play this role will depend on whether fossil energy – required for the production of inexpensive commercial fertilizers – is available for intensive cultivation to the same extent as before. Looked at from this standpoint, the cultivation of *leguminous plants* that do not require nitrogen fertilizer should be of steadily increasing importance. The largest increase in production is found in the case of soy protein. Until now, only about one-tenth of the harvest was employed directly for human nutrition, the remainder was used as feed. For climatic reasons, the cultivation of soybeans is restricted to a few regions of the earth. A search is still on for leguminous plants (e.g., lupines) that can improve the protein supply especially in needy regions.

In publications of the United Nations's Protein Advisory Group (PAG) (FAO/WHO/UNICEF/UNIDO), the main opinion is that the protein problem in undersupplied regions can be most effectively solved in the short term by the production of leguminous plants that thrive locally. These are then used to increase the value of the diet, which is normally rich in carbohydrate but low in protein – especially essential amino acids (particularly lysine and methionine) [5]. The development of protein-rich, lysine-enriched types of grain for suitable climates associated with the "green revolution" greatly contributed to the improvement in the world nutritional situation in the 1980s. With the help of genetic engineering, plants that are especially resistant to pests and agrochemicals should be developed in order to reduce harvest losses.

To what extent industrially produced *single-cell proteins* (yeast, bacteria, → Single Cell Proteins ; → Yeast), *algae,* and *leaf proteins* will gain importance for the human protein supply will depend, among other factors, on how strongly increases in the price of fossil energy and nonregenerable raw materials will affect production costs. The energy crisis of the 1980s had a very negative influence on developments. Moreover, the products should be absolutely harmless – they should contain no carcinogenic hydrocarbons and few nucleic acids and should be easily digestible. They should also be designed to possess attractive functional properties. If industrially produced proteins are used for animal food, they must compete with the cheapest protein carriers. Under these circumstances, predictions made about the future of industrial protein production seem daring, although these processes are extremely interesting from a process engineering point of view. Locally, the industrial production of single- cell proteins could become economically interesting if it can be coupled with measures for environmental protection and waste removal.

Table 1 presents a global survey of the protein supply. The figures were calculated on the basis of FAO statistics on average protein concentrations, without taking differences in quality into account [2, 3]. To compensate for fluctuations in annual harvests, the average of three-year production periods was used.

2. Plant Protein Products

2.1. Seed Protein Products

2.1.1. Cereal Protein Products (see also → Cereals and Cereal Products)

Cereal Producing Countries. The principal countries in which the most important types of cereals are grown are listed in Table 2, which also gives the annual production of each crop. As can be seen from Table 2, wheat, rice, and maize account for more than 80 % of the grain produced in the world.

Main Nutrient Content and Protein Nutritional Value of Cereals. Cereals are a source of both protein and starch (see Table 3). The protein content is ca. 10 wt %. The protein nutritional

Table 1. Annual world production of food protein 1985 – 1990

Food	Total in 1000 t		Change, %	Percentage	
	1985 – 1987	1988 – 1990		1985 – 1987	1988 – 1990
Animal Origin					
Total meat	27 320	30 037	9.9	42.4	43.4
Beef and veal	7877	9073	15.2	12.2	13.1
Pork	7114	9289	30.6	11.0	13.4
Poultry	5008	7290	45.6	7.8	10.5
Mutton and goat's meat	1302	1597	22.7	2.0	2.3
Horsemeat	101	98	−3.0	0.2	0.1
Fish and sea animals ∗	15 732	16 905	7.5	24.4	24.4
Total eggs	4321	4677	8.2	6.7	6.8
Hen's eggs	4245	4594	8.2	6.6	6.6
Total milk	17 084	17 547	2.7	26.5	25.4
Cow's milk	13 480	15 206	12.8	20.9	22.0
Buffalo's milk	1053	1492	41.7	1.6	2.2
Sheep's milk	418	449	7.4	0.6	0.6
Goat's milk	270	313	15.9	0.4	0.5
Total	64 457	69 166	7.3	100	100
Plant origin					
Total grain	155 272	158 032	1.8	63.2	62.6
Wheat	47 713	50 630	6.1	19.4	20.1
Maize	26 031	24 773	−4.8	10.6	9.8
Rice	26 054	28 244	8.4	10.6	11.2
Barley	14 822	14 198	−4.2	6.0	5.6
Oilseeds	60 237	63 697	5.7	24.5	25.2
Roots and tubers	9368	9475	1.1	3.8	3.8
Legumes	9007	9115	1.2	3.7	3.6
Potatoes	4318	4111	−4.8	1.8	1.6
Vegetables	5056	5242	3.7	2.1	2.1
Fruit and melons	2121	2222	4.8	0.9	0.9
Nuts	473	523	10.6	0.2	0.2
Total	245 852	252 417	2.6	100	100
Total food protein	310 309	321 583	3.6	100	100

∗ Excluding mammals, ca. 13.5 % from freshwater.

Table 2. Cereal production (in 1000 t) and principal cereal producing countries 1989 [6]

Type of cereal	Production	Countries
Wheat	538 056	China, Former USSR, USA, India, France, Canada
Rice	506 291	China, India, Indonesia, Bangladesh, Thailand, Vietnam
Maize	470 318	USA, China, Brazil, Former USSR, France, Romania
Barley	168 964	Former USSR, Germany, Canada, Spain, USA, UK
Sorghum	57 976	USA, India, China, Nigeria, Mexico
Oats	42 197	Former USSR, USA, Canada, Poland, Germany, Finland
Rye	34 893	Former USSR, Poland, Germany, China, Canada, Czechoslovakia
Millet	30 512	India, China, Former USSR, Nigeria, Niger, North Korea
Total	1 864 852	

Table 3. Approximate nutrient content and biological value (BV) of cereal proteins [7–11]

Type of cereal	Protein content, wt %	BV	Fat, wt %	Starch, wt %
Wheat	12 (80) [a]	54	1.9	72
Rice	8	70	1.2	76
Maize	10	65	4.6	80
Barley	11	76	2.3	74
Rye	12	74	1.7	72
Sorghum	11	56	3.5	65
Oats	11	76	5.8	56
Millet	10	68	4.5	59
Buckwheat	11	2.4 [b] [13]	2.4	64
Triticale [12]	13	1.3 [c]	2.6	

[a] Gluten proteins.
[b] Grams of essential amino acids per gram of total nitrogen (egg protein: 3.2).
[c] PER (protein efficiency ratio).

value is in the range of average to good, the first limiting amino acid being lysine [*6899-05-4*] (see → Amino Acids, Chap. 5.1.). Therefore, mixtures of grain proteins with a protein having a high lysine content increase the protein nutritional value.

Isolation and Use of Maize and Wheat Proteins. The isolation of maize and wheat proteins is coupled primarily to the production of starch (→ Starch). In 1988, 70 % of the starch produced was obtained from maize and 5 % from wheat. The resulting *maize gluten meal* [*8002-80-0*] (obtained by wet grinding) has a protein content of 60 to 70 % and is used primarily in the feed industry (see Fig. 1). Only smaller amounts are employed in the production of protein hydrolysates and glutamic acid.

Zein (maize prolamine) [*9010-66-6*] is extracted from maize gluten by treatment with an aqueous solution of 88 vol % isopropanol and 0.25 wt % sodium hydroxide at 60 °C, and subsequent precipitation at −15 °C. Zein (producer: Freeman Industries, United States) is used as a surface coating for nuts, candies, tablets, and textile fibers.

Maize Germ Meal. In the isolation of maize germ meal, the maize germ is obtained by wet grinding. Most of the oil is then removed by pressing at 120 °C. The process is concluded by extraction of the remaining oil with hexane.

In the meantime, the separation of maize germ by dry grinding has also proved useful (e.g., Lauhoff Grain, Illinois Cereal Mills, Archer Daniels Midland, United States). In this case, the oil is extracted only with hexane.

The maize germ meals obtained after drying can be used directly for the processing of food or, in principle, as a raw material in the production of protein concentrates. The composition of these meals is given in Table 4.

A *protein concentrate* (i.e., a protein-enriched product obtained by concentration of the proteins in a protein-containing raw material) has been successfully prepared from deoiled maize germ (dry grinding) by *extracting* the protein with a mixture of aqueous solutions of 70 vol % ethanol and 0.7 wt % sodium hydrox-ide (11 : 1) (protein content ca. 60 wt %). The proteins of maize germ are very soluble in water: 20 % at the isoelectric point (≈ 5). The emulsion activity and emulsion stability amount to approximately 50 % (ca. 10 % at the isoelectric point). The fat absorption of maize germ proteins is considerable, being 549 mL of oil per gram of protein. Therefore, maize germ meal and the corresponding protein concentrate are employed in the making of bread and other baked goods, meat products (sausages), and sauces. Textured maize germ meal can be produced by extrusion. It is ideally suitable for use in ground-meat-like products.

The protein enrichment of wheat meal *by air classification* is not very common because the energy required for pin-mill grinding is high. However, factories employing this process do exist in the United States, France, and Germany. The protein content of the wheat meal (at least 20 %) is increased to 25 % by means of air classification. Meal of this type is used primarily to prepare bread flour. It improves both the dough making and baking process as well as the quality of the bread.

Vital Gluten. The vital gluten obtained on separating starch and gluten is often considered the primary product in wheat-starch production. In such cases, instead of whole grain, wheat meal is used as the starting material because it allows an especially good separation of gluten. The separation of wheat vital gluten is shown in Figure 2. The starch is washed out of the wheat dough in two steps (Martin process); the resulting wet vital gluten contains about 11.5 wt % protein, 2 wt % starch, and 0.9 wt % fat. Since the beginning of the 1980s, the soluble wheat albumins and globulins have been isolated from the wash water by heat coagulation, according to the Raiso process (Finland). The resulting product is a protein-enriched flour.

Gluten is a protein complex whose minimum solubility in water is 20 – 30 wt % between pH 6.5 and 9.0. The solubility is increased to about 80 wt % in the acidic or alkaline pH range. To increase the solubility in the isoelectric range, wheat gluten is partially modified by enzymatic hydrolysis. In this way, proteins that have good emulsifying and foam producing properties are formed. The water absorption capacity of vital gluten is remarkable; it can absorb water up to 65 % of its mass. The wet vital gluten thus formed has pronounced viscoelastic properties and forms a firm gel when heated to 85 °C.

These functional properties explain the fact that at present, ca. 80 % of the wheat vital gluten

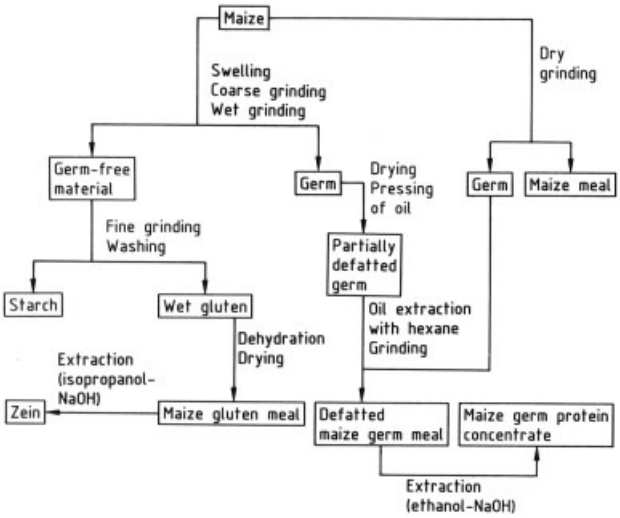

Figure 1. Isolation of maize gluten meal and maize germ meal and of the corresponding protein concentrates

Table 4. Composition and protein nutritional value of differently treated maize germ [14–16]

	Protein content, wt %	Oil content, wt %	PER ∗
Maize germ, wet grinding			
Without oil extraction	18	40	2.6
After oil extraction by pressing (120 °C)	27	10	1.5
After oil extraction with hexane (60 °C)	29	1	
Maize germ, dry grinding			
Without oil extraction	14	20	2.6
After oil extraction with hexane (60 °C)	17	1	2.3

∗ Protein efficiency ratio.

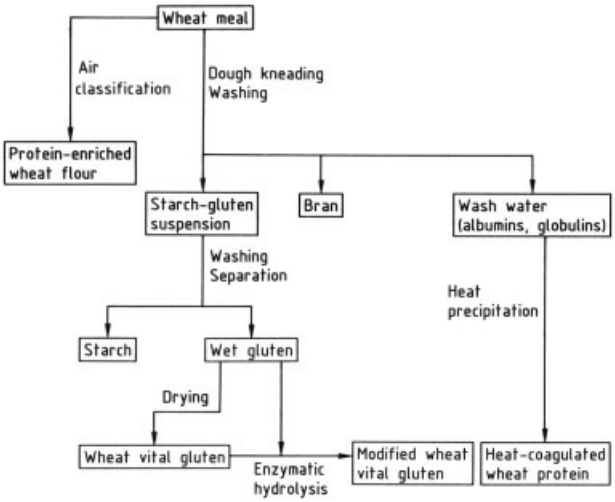

Figure 2. Isolation of protein-enriched flour and protein concentrates from wheat

produced is used in making bread and baked products. It is employed primarily in special breads and breakfast cereals. In the meat and sausage industry, vital gluten has proved useful as a binding agent for sausages and ground-meat products and in the production of sauces. By extrusion of vital gluten, excellent meat extenders can be obtained, which can partially replace the meat in ground-meat products. Gel-like meat analogues made from vital gluten are common in Japan.

As a result of its high biological value compared to gluten, *heat- coagulated* wheat protein is used primarily to improve the value of protein. The composition of wheat protein products is given in Table 5.

Producers and Trade Names of Wheat Gluten. Important producers of wheat gluten and a selection of the trade names of their products are listed below:

Ajinomoto, Japan	A-Glu
Allied Mills Industries, Australia	
Amylum, Belgium	Biogluten (vital gluten) SWP (sodium salt of a soluble wheat protein isolate)
Archer Daniels Midland, United States	Do-Pep Vital
Beldum, Belgium	
Cerestar, France	
Crespel & Deiters, Germany	
Dutch Protein and Services, Netherlands	
G.M.F.F., Minoteries à Féves, France	Fabaten (mixture of gluten and faba bean flour)
George Weston Foods, Australia	
Manildra Milling Corporation, United States	
Molinos Juan Semino, Argentina	
Nidefo, Netherlands	HPF 50 HPM (enzymatically modified vital gluten)
Nisshin Mill	Sun Meat (textured gluten, meatlike, protein content: 28 %)
Raision Tehtaat, Finland	
Roquette, France	Viten (vital gluten) Deviten (denatured)
Rovita, Germany	Rovita
Shiuslim, Japan	Hi Pro-P
Tenstar Aquitaine, France	
Tenstar Products, United Kingdom	
Wheat Industries, Iceland	

2.1.2. Oilseed Protein Products

Oilseed Producing Countries. Details about the annual production of important oilseeds and the main oilseed producing countries in 1989 are given in Table 6. Information on lupine is not available. Oilseeds represent important sources of protein with a high biological value, the most significant being soybean.

Table 6. Production (in 1000 t) and principal oilseed producing countries 1989 [22]

Oil plant	Production	Countries
Soybean	107 350	USA, Brazil, China, Argentina
Cotton	30 703	China, Former USSR, USA, Pakistan, India, Brazil
Peanut	22 594	India, China, USA
Rapeseed	22 302	China, India, Canada, France, Germany, Poland
Sunflower	21 867	Former USSR, Argentina, France, Turkey, Romania, China, USA
Sesame	2352	India, China, Sudan, Myanmar
Linseed	2121	Canada, Argentina, India, Former USSR, China
Castor bean	1155	India, China, Brazil
Safflower	908	India, USA, Mexico, Ethiopia

Main Nutrient Content and Protein Nutritional Value of Oilseeds. Table 7 shows that oilseeds are an equally good source of both protein and oil. The biological value of the proteins indicates that they are generally high grade. This applies particularly to soybean proteins [*9010-10-0*].

Isolation and Use of Soybean Protein Products [30]. *Soybean Flour. Whole-fat flour* is obtained by heat-treating soybeans in a moist atmosphere and subsequently shelling and grinding them. To produce *defatted flour,* steam-treated shelled beans are ground coarsely. The water content is then adjusted to ca. 10 wt % by tempering to enable the subsequent extraction of soybean oil [*8001-22-7*] with hexane, for example. Depending on the destruction degree of antinutritive factors, the extracted cake is then subjected to further heat treatment (toasting) and subsequently ground. Flour with a distinct fat and lecithin content is made by the addition of oil or lecithin to defatted flour (Table 8).

Table 5. Composition of differently isolated wheat protein concentrates [17–21]

	Protein-enriched flour (air classification)	Vital gluten	Enzymatically modified vital gluten	Soluble wheat protein	Heat-coagulated wheat protein (albumins + globulins)
Protein, wt %	25	80 – 85	85 – 90	85	40
Fat, wt %		0.5 – 2	0.5 – 2	8	
Carbohydrate, wt %		3 – 5	3 – 5	1	
PER (NPU) ∗	1.5 (64)	0.7 – 1.0 (58)			2.2

∗ Protein efficiency ratio (net protein utilization).

Table 7. Approximate nutrient content and biological value (BV) of oilseed proteins

Oilseeds	Protein content		BV [29]	Fat, wt %
	Original, wt %	Shelled defatted, wt %		
Soybean [28]	38	40 – 52	60 – 75	17 – 23
Cotton [28]	20	33 – 50	55 – 57	20 – 50
Peanut [28]	26	45 – 60	50 – 75	45 – 55
Rapeseed [28]	20	35 – 42	70	30 – 48
Sunflower [28]	20	40 – 62	60 – 71	30 – 52
Sesame	20	60	62	43 – 59 [23]
Coconut meat	4	24	64	15 – 35
Lupine (*Lupinus mutabilis*)	43 [24]	52 [25]	39 [26]	14 – 22 [24]
Safflower			1.3 ∗	
Linseed				30 – 40
Mustard	23 [27]	42 – 47 [27]		30 – 35

∗ PER (protein efficiency ratio).

Table 8. Average composition (wt %) of soybean flour [31]

	Whole-fat flour	Defatted flour	Flour with a low fat content	Lecithin-containing flour
Water	5	5	6	6
Protein	41	53	46	45
Fat	21	1	6	16
Raw fiber	2	3	3	2
Ash	5	6	6	5
Phytic acid	1 – 1.5			
Saccharose		7		
Stachyose		6		
Raffinose		1		
Starch		≈ 15		

The functional properties of soybean flour can be determined to a considerable extent by the heat treatment process. For instance, flour of varying solubility, with protein solubilities between 10 and 90 wt % in water can be produced. Defatted types of flour, with and without added lecithin, have gained acceptance in the market. Lecithin- containing flour is used in the making of bread and baked products, mayonnaise, dressings, and sauces. Soybean flour without lecithin is employed primarily in the production of meat products, bread, and baked goods. Soybean flour can also be textured for use as a partial meat substitute (e.g., by extrusion).

Soybean protein concentrates commonly have a minimum protein content for the dry products of 60 – 65 wt % (determined as nitrogen content, N×6.25).

Soybean protein concentrates can be isolated to obtain the protein concentrates II, III, and IV as shown in Figure 3. The composition of these concentrates is listed in Table 9.

Table 9. Composition and properties of differently produced soybean concentrates [31]

Composition and properties	Extraction of soluble components with		
	Water (concentrate II)	Ethanol (concentrate III)	Acid (concentrate IV)
Water, wt %	3	7	5
Protein, wt %	70	66	67
Fat, wt %	1	0.3	0.3
Raw fiber, wt %	4	4	3
Ash, wt %	4	6	5
pH	6.9	6.9	6.6
Protein solubility, %	3	5	70

Generally, these products exhibit low water solubility but have good water- (3.5 g/g) and fat-binding (2.5 g/g) properties. They are used primarily in the production of meat products

Table 10. Composition of soybean protein isolates [31]

Protein, wt %	Water, wt %	Raw fiber, wt %	Ash, wt %	pH (aqueous solution)
88 – 95	5 – 10	0.1 – 0.2	1.2 – 2.5	4.6 – 5
85 – 93	5 – 8.5	0.2	3 – 7.5	6.9 – 7.2

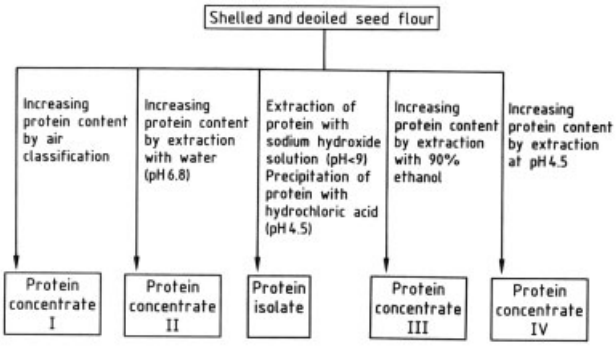

Figure 3. Isolation of protein concentrates and isolates

(sausages to be heated or cooked), pasta, and bread, and baked goods.

In a cooking extruder, concentrates of this type can be textured in a heat – pressure process to give extrusion products. The same result can be achieved with the corresponding raw material flour. A fibrous microstructure is obtained in this process. With soybean protein concentrates, extrusion reduces the water solubility by two-thirds. The protein nutritional value, as expressed by the protein efficiency ratio (PER), is also reduced (from ≈ 2.4 to 2.1). Extrudates of this type are especially suitable for partial meat substitution. Soybean protein concentrates with high solubility, gelling, and emulsion stability can now be prepared by the gentle processing of soybeans.

Isolates of soybean proteins (glycinin [*9007-93-6*], conglycinin [*104042-07-1*]) are made from shelled soybeans according to the procedure shown in Figure 3. Their composition is given in Table 10. They are relatively soluble and exhibit good swelling (10 – 20 mL water/g), gelling, emulsifying, water-binding (≈ 400 wt %), fat-binding (≈ 150 wt %), and foaming properties (250 % increase in volume). Soybean protein isolates produce foams with a high stability. These isolates are used in the production of nearly all groups of food, especially meat products, pasta, bread, and baked goods, milk products, nonalcoholic beverages, and baby food. They are also employed in the making of new foods (spinning of fibers, texturization, and simulation), which include meat (on the basis of protein fibers) and milk products. The protein values of these products are given in Table 11.

Table 11. Comparison of protein nutritional value of natural and simulated meat and milk products [32]

Products	BV [a]
Cow's milk	87
Cottage cheese	97 [b]
Soybean milk	83 – 95
Soybean curd cheese	64
Soybean cheese	2.7 [c]
Beef	76 [d]
Pork	46 [d]
Simulated meat from soybean protein	81
Simulated meat from faba bean protein/casein (1 : 1)	80 [d]

[a] Biological value.
[b] NPU.
[c] PER.
[d] Test on rats.

Producers and Trade Names of Soybean Protein Products. Table 13 presents a selection of the producers and trade names of soybean products. The compositions, functional properties, and uses are listed in the product classes (flour, protein concentrates, and isolates).

Table 12. Producers and trade names of flour, grits, and flakes (F) including textured (TF); protein concentrates (PC) including textured (TPC) and functional (FPC); and protein isolates (PI) *

Producers	Soybean products					
	F	TF	PC	TPC	FPC	PI
Aarhus Olie, Denmark				Danprotex		
Dantex	Danpro					
Archer Daniels	Nutrisoy	TVP	Arcon	Arcon T		Ardex
Midland-Société Industrielle de Oleagineux, USA, France	Soylec		Calfpro			Pro-Fam
Central Soya, Chemurgy Division, USA, Netherlands	Soyarich	Miratex		Response	Promine-HV/DS	
	Soyafluff	Centex	Cenpro Promosoy 100 Procon 2000	Textured Procon Promoveal Promocaf	Promine-70/70 L Promine 1 : 4 : 4 C	
Dutch Protein & Services, Netherlands	D.P.S.					
Fuji, Japan						x
G.M.F.F., Minoteries à Fefes, France		x	x	x		
GMB Proteins, United Kingdom		Bontrae				
Lucas Meyer-Edelsoya, Germany	Soyapan Nurupan Nurulat Nuruflakes Nurugran Soypur	Textratein	Soyamin 50 Soyamin 70		Vegafoom D	Soyamin 90
Loders Croklean-Unilever, Netherlands		Unico Soycomil				

Table 12. (Continued)

Producers	Soybean products							
Ralston Purina-Protein Technologies				Solcaf				Supro
International, USA, Belgium				Proteimax				x
Sanbra, Brazil		Soybar	Soytex	Solcon	Contex	Solpro		
Solbarit, Hatzor, Israel		Solnuts					Mira-Foam	Mira-Pro
Solnuts, Netherlands							Versa-WHIP	
Staley and Gunther Products, USA								
	F		TF	PC	TPC	FPC		PI
Vamo, Mills, Belgium	Provamix							
	Provagrits							
	Provaflor							
	Provabis							

* x = Product available, no special trade name.

No official information is available on the production volume of these products. The total production of soybean protein concentrate and isolate is estimated to amount to ca. 130 000 t and 80 000 t, respectively. A substantial portion of protein isolates is used in the paper industry and in the production of baby food. In the meat industry (e.g., in Europe, Japan, and the Republic of South Africa), soybean protein isolates and concentrates are used in the production of frankfurter-type sausages. In addition, about 30 000 t of soybean protein concentrate is used in the United States for ground-meat products for school lunches and for the army.

2.1.3. Legume Protein Products

Legume-Growing Countries. The legume-growing countries and their production volumes are listed in Table 13. As can be seen from Table 12, the *Phaseolus* bean and the pea have the highest production volumes.

Table 13. Production (in 1000 t) and principal legume (dry) producing countries, 1989 [33]

Plant	Production	Countries
Peas	16 447	Former USSR, France, China, India
		Hungary, UK
Phaseolus beans	15 872	India, Brazil, China, USA, Mexico
Chick-peas (garbanzos)	7429	India, Turkey, Pakistan,
Faba beans	4058	China, Egypt, Ethiopia,
		Morocco, Germany, France,
		Italy
Lentils	2242	India, Turkey, Bangladesh,
		China
Total	57 985	

Main Nutrient Content and Protein Nutritional Value of Legumes. As shown in Table 14, legumes are sources of both protein and starch. Except for the proteins of lentils, which have a lower biological value, the other legume proteins (vicilin and legumin [*102426-92-6*]) have an average biological value. This is due to their relatively low content of sulfur- containing

Table 14. Approximate nutrient content and biological value (BV) of proteins of legumes

Plant	Protein content		BV	Fat	Starch
	Original, wt %	Flour, wt %			
Phaseolus beans [37]	26	30	58 [39]	1 – 3	50 – 55
Peas [36]	23	14 – 28	64 [39]	1.2 – 4.1	44 – 48
Faba beans [34]	29	28 – 32	48 – 63	0.5 – 1.2	44 – 46
Chick peas [35]	20	21 – 26	57 [39]	3.4	50
Lentils [38]	21		45 [39]	0.9 – 1.2	50 – 54

amino acids. However, since lysine concentrations are high, biologically high-grade products are obtained by mixing grain proteins with leguminous proteins.

Isolation and Uses of Legume Protein Products [40–48].

Faba Bean, Pea, and Chick-Pea Flour. Faba beans, peas, or chick-peas are subjected to heat treatment in a moist atmosphere, shelled, and then ground to obtain flour I (see Fig. 4). Antinutritive components are partially eliminated by aqueous extraction at pH 4.2 to obtain flour II. In contrast to soybean flour, protein concentrates cannot be isolated by aqueous extraction (i.e., the protein content of the flour does not increase).

The composition of these types of faba bean flour is given in Table 15; their functional properties are listed in Table 16. These properties are determined primarily by the main components globulins and starch and also by the highly functional phosphatides. Flour made from faba beans, peas, or chick-peas can be used for the same applications as soybean flour, and the high proportion of starch can be an advantage over soybean flour in these applications.

Faba Bean and Pea Protein Concentrates. According to Figure 4, faba bean and pea flour can be enriched in protein by air classification to yield protein concentrates. The composition and functional properties are given in Tables 15 and 16 for faba beans. The simultaneous enrichment of antinutritive components is a disadvantage of this process. Since the starch content of faba bean protein concentrate amounts to only 1 wt %, the fields of application for protein concentrates from faba bean and soybean are identical.

Faba Bean and Pea Protein Isolates. Protein isolates are obtained from the corresponding flour by extraction with water. A pH of 6.5 – 7.0 is established in this process. As in the case of soybean protein isolates, subsequent precipitation from aqueous solution occurs at the isoelectric point (i.e., pH 4.2).

The protein isolates have remarkable emulsifying and foam producing as well as emulsion and foam stabilizing properties. Special purposes can be served by modifications of the protein isolates. For instance, high emulsifying and stabilizing properties can be generated by the *acetylation* of faba bean protein isolates. Both the foam producing capacity and the foam stability can be improved by *heat treatment* of faba bean protein isolates. Similar effects can also be achieved by *enzymatic partial hydrolysis.*

With faba bean protein isolates as an example, the composition and functional properties of these products are presented in Tables 15 and 16. The applications of these protein isolates correspond to those of soybean isolates.

Producers and Trade Names of Leguminous Protein Products. Different types of faba bean and chick-pea flour are commercially available (see Table 17). This applies also to textured faba bean flour, textured faba bean protein concentrate, and pea protein isolate.

2.2. Protein Products from Roots and Tubers

Root and Tuber Producing Countries. The production of roots and tubers in 1989 and the principal countries involved are shown in Table 18.

Main Nutrient Content and Protein Nutritional Value of Roots and Tubers. The protein content of the roots and tubers listed in Table 19 is less than 2.5 wt % (about 10 wt % of

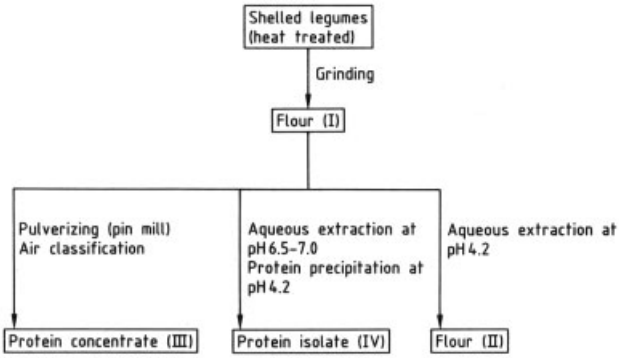

Figure 4. Isolation of flour, protein concentrates, and isolates from leguminous plants

Table 15. Composition (in wt % of dry weight) of flour, protein concentrates and isolates from faba beans

	Flour I	Flour II	Protein concentrate III	Protein isolate IV
Protein	30	28	70	80 ∗
Starch	45	43	1.7	0
Lipid	2.3	1.8	3.4	5
Raffinose	0.2	0	0.6	0.004
Stachyose	0.5	0.02	1.4	0
Verbascose	2.0	0	3.2	0
Vicine	0.7		1.5	0.1
Convicine	0.3		0.7	0.06
Phytic acid	1.8 (3.1) ∗∗	(3.6) ∗∗	4.2	(6.8) ∗∗
Trypsin inhibitor	4.8	3.7	15	2.1

∗ N×5.9.
∗∗ Milligrams of phytic acid phosphorus per gram.

Table 16. Selected functional properties of flour, protein concentrates, and isolates from faba beans

	pH	NSI ∗, %	Absorption of		Degree of gelatinization, %
			Water, wt %	Fat, wt %	
Flour I	6.5	84	65	79	62
Flour (pin-mill ground)	6.6	39	58	48	80
Protein concentrate III	6.5	43	104	90	94
Protein isolate IV	7.0	99	242	171	100

∗ NSI = Nitrogen solubility index.

Table 17. Trade names and producers of flour (F) including textured (TF); textured protein concentrates (TPC); and protein isolates (PI)

Producers	Products from faba bean (FB), pea (P), and chick-pea (C)			
	F	TF	TPC	PI
G.M.F.F., Minoteries à Fefes, France	Fabafroid (FB) C	Fabapro (FB)	Fabatex (FB)	
Cosucra, Belgium				P
Nutrio-Grindsted, Denmark				P-Pro 2000 (P)

the dry weight). The biological value of the proteins is generally good. In normal nutrition, roots and tubers really represent a source of starch. They become simultaneous suppliers of protein on their large-scale processing to starch. This applies especially to potatoes (see also → Starch). The potato protein water ob-tained during processing to starch can represent a real wastewater problem because it enhances microbial growth. For this reason, it is used for the isolation of protein in modern starch production plants.

Isolation and Uses of Potato Protein Products [56] – [2.57 a]. Protein is normally precipitated from potato protein water (protein content 2 – 4 wt %) at pH 4.7 – 6.0 by steam injection at 80 – 120 °C. The separated coagulate is washed with water and spray-dried. Potato protein concentrates of this type contain 78 – 89 wt % of protein and 7 – 8 wt % of fat. This product is brownish, tastes like raw potatoes, and has no favorable functional properties. It is used as feed in calf breeding because of its good protein nutritional value [net protein utilization (NPU) = 73].

Table 18. Production (in 1000 t) and principal root and tuber producing countries 1989 [49]

Plant	Production	Countries
Potato	276 740	Former USSR, Poland, China, Germany, UK, Romania, Spain
Sweet potatoes	133 234	China, Indonesia, Vietnam, Uganda, India
Cassava (manioc)	147 500	Thailand, Brazil, Indonesia, Nigeria, Zaire
Yam	23 459	Nigeria, Côte D'Ivoire, Benin
Taro	5814	Nigeria, China, Ghana
Total	590 176	

Table 19. Approximate nutrient content and biological value (BV) of proteins of roots and tubers [50]

	Protein content, wt %	BV	Starch, wt %
Potato	2.1	48 [a] [51], 62 [b] [53], 59 [c] [11]	18.5
Sweet potato	1.4	2.6 [d] [52], 60 [b] [53]	27.4
Cassava	1.1	52 [b] [53], 50 [c] [55]	35.2
Yam	2.2	66 [b] [53], 60 [c] [55]	24.2
Taro	1.9		35.2

[a] Isolated potato protein, BV = 88 [54].
[b] EAAI (essential amino acid index).
[c] NPU (net protein utilization).
[d] PER.

In principle, potato protein represents a source of protein for direct use in human nutrition. The necessary work to achieve this can essentially be regarded as concluded. This refers primarily to the inactivation of phenol oxidases, separation of lipoproteins in potato protein water, precipitation of proteins at 70 – 80 °C, and lipid extraction of the protein coagulate. In this process, protein isolates with protein content of > 90 wt % and fat content of < 1 wt % can be obtained. At the same time, the functional properties (including taste and color) are improved to such an extent that the use of these proteins in the production of food should be possible.

2.3. Leaf Protein Products

Protein Content and Protein Nutritional Value of Leaves. The selection of plants presented in Table 20 shows that in principle the leaves of all agricultural plants can be used for the isolation of proteins. The possible use as a protein raw material is determined finally by the following factors: protein content, extractable portion of proteins, protein nutritional value, and possibility of eliminating plant components that present a health hazard.

Table 20. Protein content and extractable protein (% of dry weight) of leaves [64] and biological value of protein [54]

	Protein content	Extractable protein	Biological value
Castor bean	41.3		
Chick-pea	31.8		
Cassava	31.6 (63) [a]	10.2	−6.9 [b] (19.2) [c] [71]
Banana	20.9	11.3 [66]	
Coffee	19.5		
Tobacco [65]		0.8	PER of casein
Lucerne	21	10.4	32 – 49 [b] [70]
Clover	31.6 [69]	19.5 [68]	1.3 – 2.6 [d] [67,68]
Sorghum-sudan grass		10.6	
Pondweed		5.4	
Water hyacinth		1.8	
Water lettuce		0.9	

[a] Percent of fresh weight.
[b] NPU.
[c] Cassava leaf protein concentrate + 0.6 % D,L-methionine.
[d] PER.

Isolation and Use of Leaf Protein Concentrates from Lucerne and Clover, Trade Names. The possibilities for the preparation of leaf protein concentrates are summarized in Figure 5. The proteins can be obtained from the leaf extract ("green juice") by heat or acid precipitation, by formation of insoluble complexes with pectin or carboxymethyl cellulose or by ultra- or diafiltration.

Until now, most research has been devoted to the isolation of proteins from lucerne and red

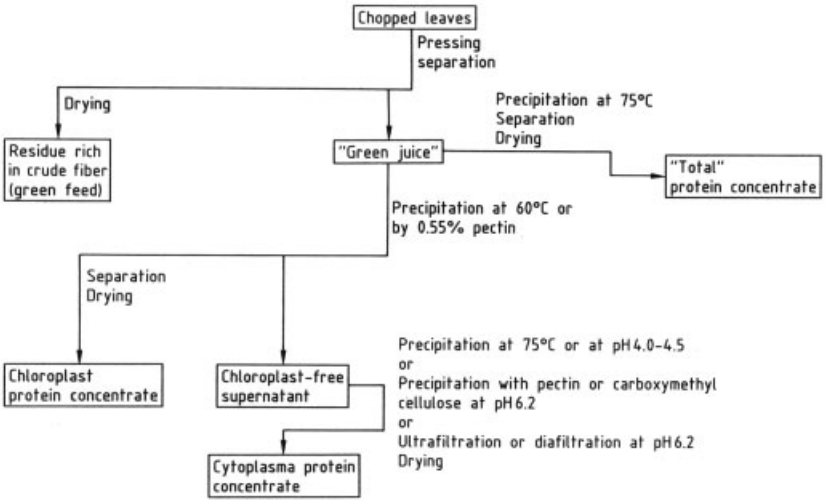

Figure 5. Isolation process for leaf proteins [54, 67]

Table 21. Composition of differently isolated leaf protein concentrates [69, 72–76]

	Protein, wt %	Fat, wt %	Carbohydrate, wt %
Lucerne			
"Total concentrate" (Pro-Xan, Vepex)	50 – 60	9	
Chloroplast concentrate			
By heat precipitation (Pro-Xan II)	45	13	
By precipitation with pectin	32		26
Cytoplasma concentrate			
Acid precipitation at pH 4.8	74	0.5	18
Heat precipitation at 75 °C (Pro-Xan II)	89	0.6	
Diafiltration	88		8
Clover			
"Total concentrate"	68	10	
Cytoplasma concentrate			
Acid precipitation at pH 4.8	68	0.5	15
Precipitation with pectin at pH 6.2	20		55
Precipitation with carboxymethyl cellulose-Na at pH 6.2	22		51

and white clover. The composition of the corresponding protein concentrates is presented in Table 21.

Leaf proteins with good functional properties are obtained by the ultra- or diafiltration of chloroplast-free extracts. Colorless lucerne protein concentrates isolated in this manner are water soluble at pH 2.0 – 2.5 and above 5.0. The solubility at pH 6.2 (solubility 95 wt %) is reduced by addition of NaCl; 2.0 mol/L NaCl decreases the solubility to 70 wt %. The proteins are heat stable up to 50 °C.

In principle, leaf protein concentrates are suitable for human nutrition.

Trade Names. Several production plants exist that are based on lucerne as raw material. In the United States, Pro-Xan (total lucerne protein concentrate) and the Pro-Xan II product series are produced. The latter include a green lucerne protein concentrate (chloroplast protein concentrate) and a colorless lucerne protein concentrate (cytoplasma protein concentrate).

The product Vepex (total lucerne protein concentrate) was produced in Hungary. This concentrate was also available as a mixture with 30 wt % skim milk powder (Velac). "Total" Proxan as well as "green" and "colorless" Proxan contain xanthophyll and are used as special animal feed for poultry.

The water solubility of the proteins contained in Proxan and Vepex at pH 2 – 10 is ca. 5 wt %.

2.4. Health Aspects

2.4.1. Antinutritive Factors in Plant Protein Raw Materials

The important antinutritive factors are listed in Tables 22, 23, 24, 25, 26 for cereals, legumes, roots, and tubers. Some antinutritive factors are found in almost all plant protein raw material. This applies to enzyme inhibitors (inhibition of the digestive enzymes trypsin, chymotrypsin, and α-amylase), lipoxygenases [*9029-60-1*] (off-flavor formation from unsaturated fatty acids), isoflavones [*4737-27-3*] (estrogenic action), hemagglutinins (lectins [*103938-95-0*], blood coagulation), tannins (bitter taste), phytic acid [*83-86-3*] (reduction of the bioavailability of zinc and iron ions), phenol derivatives (e.g., sinapine [*18696-26-9*] and chlorogenic acid [*906-33-2*], reduction of the biological value of the protein), nitriles (formation of hydrogen cyanide), and the oligosaccharides raffinose [*512-69-6*], stachyose [*470-55-3*], and verbascose [*546-62-3*] (flatulence producing).

Specific antinutritive factors are of significance for particular plant protein raw materials. This applies to gossypol [*303-45-7*] in *cotton seeds* (inhibition of spermatogenesis), to 5-vinyloxazolidine-2-thione [*13190-34-6*] (thyrotoxic) formed from glucosinolates

(glucobrassicin [*4356-52-9*]) in *rapeseeds*, to 1-amino-D-proline [*10139-05-6*] (antipyridoxin effect) formed from the peptide linatine [*10139-06-7*] in *linseed,* to the favism factors vicine [*152-93-2*] and convicine [*19286-37-4*] (hemolytic anemia in case of glucose-6-phosphate dehydrogenase deficiency) in *faba bean,* to lupanine [*550-90-3*] and sparteine [*90-39-1*] in *lupines* (CNS excitation, paralysis, muscular stimulation, bitter taste), to α-solanine [*20562-02-1*] and α- chaconine [*20562-03-2*] in *potatoes* (diarrhea with vomiting, nephritis, cerebral edema, cramps, coma, bitter taste, and off-flavor), to dioscorine [*3329-91-7*] in *yam*, as well as to the bitter substance 2-hydroxyarctiin [*20362-31-6*] and to the laxative 1-matairesinoside [*23202-85-9*] in safflower.

In principle, lysinoalanine [*18810-04-3*] (nephrocytomegaly) can be formed in the processing of proteins, especially in an alkaline environment.

The formation of highly toxic metabolic products of mold fungi – mycotoxins – must also be considered for a series of plant protein raw materials. For instance, the real toxicological problem with peanuts is the possible aflatoxin content (liver damage).

All the plant protein raw materials listed here are used to a greater or lesser extent in human nutrition; this also applies to proteins made from these materials. The problems mentioned above can be minimized or even eliminated by the proper storage of raw materials and by the use of processing principles adjusted to the antinutritive factors (inhibition of their formation, removal by degradation, or extraction).

2.4.2. Minimization of Antinutritive Factors in Processing Selected Plant Protein Raw Materials

Cereals. Although protease and amylase inhibitors occur in cereals, their concentrations are relatively low. Cereals contain only ca. 1 % of the amount of trypsin inhibitors found in soybean. Moreover, a large portion of these inhibitors is heat sensitive. Therefore, these raw materials do not present a health risk in human nutrition. Even the phytic acid concentration is tolerable, as shown by daily nutrition.

However, the fact that some people are sensitive to cereal proteins is of great practical importance. The mucosa of the upper small intestine is flattened by the prolamins in the gluten of wheat (gliadin [*9007-90-3*]), rye (secalin), barley

(hordeine [*99550-00-2*]), and oats (avenin). This results in a reduced resorption capacity of the upper small intestine and malabsorption (i.e., in celiac disease). The prolamines of rice, maize (zein), millet, sorghum (kafirin), and buckwheat do not cause celiac disease.

Soybeans. Raw soybeans and untreated soybean flour (fatty and defatted) contain a series of antinutritive factors. During the production and processing of soybean products, these factors can be inactivated by destruction, extraction, or addition of inhibitors, which either directly inhibit the factors or neutralize their activity. Thus, *heat treatment* is used to substantially reduce the activity of hemagglutinins (in fatty soybean flour from 330 to 13 units per milligram of protein)

Table 22. Occurrence of antinutritive factors in cereals (F=flour; C=protein concentrate; I=protein isolate) [77–81]

	Wheat		Rice		Maize	Barley	Sorghum	Oats	Rye
	F	I	F	C					
Lipoxygenases	det. [a]				det. [a]				
Phytic acid, wt %	0.9	1.9 [b]	0.1	3.0	1.0	1.0		1.0	
Protease inhibitors (trypsin + chymotrypsin)		det. [a]		det. [a]	det. [a]	det. [a]		det. [a]	det. [a]
α-Amylase inhibitors		det. [a]				det. [a]	det. [a]		
Hydrogen cyanide, mg/100 g					det. [a]		250 [c]		
Isoflavones, wt %		0.1		det. [a]	det. [a]	det. [a]		det. [a]	
Lysinoalanine, wt %		0.6							

[a] det. = detectable.
[b] Gluten.
[c] Whole plant.

Table 23. Occurrence of antinutritive factors in oilseeds (F = flour, C = protein concentrate, I = protein isolate) [78–80, 82–87]

	Soybean			Cottonseed		Peanut	Rapeseed			Sunflower		
	F	C	I	F	C		F	C	I	F	C	I
Gossypol, wt %				1.7	0.02							
Glucosinolates, μM/g							142	6	0			
VOT $*$, μM/g							0.6	<0.05	0			
Nitrile, μM/g							1.9	<0.1				
Isothiocyanates, μM/g							0.3	<0.1	0			
Phytic acid, wt %	1.4	1.8	0.8	3		1.7	3.4	0	0			0.9
Hemagglutinins, units/mg	330		13									
Protease inhibitors, units/g (trypsin + chymotrypsin)	86		21			det. $**$						
Raffinose, wt %	1.1	0.5				1						
Stachyose, wt %	5.5	1.8				1						
Verbascose, wt %						0.5						
Sinapine, wt %	0.02			0.05			2.6	0.2		0.3		
Chlorogenic acid, wt %										1.5	0.4	0.1
Lysinoalanine, wt %			0.4									0.7
Isoflavones, mg/g	2.6			det. $**$		det. $**$						
Lipoxygenases, units/g	100											

$*$ VOT = 5-vinyloxazolidine-2-thione.
$**$ det. = detectable.

Table 24. Occurrence of antinutritive factors in oilseeds (continuation of Table) (F = Flour; C = protein concentrate; I = protein isolate) [78–80, 82–87]

	Sesame	Linseed	Castor bean	Safflower	Lupine		
					F	C	I
Linatine		det. [a]					
Nitriles		det.					
Phytic acid, wt %	5						
Hemagglutinins			det. [a]				
Oxalic acid, wt %	0.3						
Isoflavones		det. [a]		det. [a]			
Alkaloids, wt % (lupanine + sparteine)					3.7	0.06	0.2
2-Hydroxy-arctiin [b]				det. [a]			
1-Matairesinoside [c]				det. [a]			

[a] det. = detectable.
[b] Bitter substance.
[c] Laxative.

Table 25. Occurrence of antinutritive factors in legumes (F=flour, C=protein concentrate, I=protein isolate) [78–80, 88]

	Peas	*Phaseolus* beans	Chick-peas	Faba beans			Lentils
				F	C	I	
Vicine, wt %				0.7	1.5	0.1	
Convicine, wt %				0.3	0.7	0.06	
Phytic acid, wt %	1.0	0.6	0.4	1.8	4.2	2.5	
Hemagglutinins, units/mg	det. ∗	det. ∗		11	200	6	det. ∗
Trypsin inhibitor, units/mg		det. ∗	det. ∗	5	15	2	det. ∗
Raffinose, wt %	det. ∗	0.3		0.6	1.1	0.004	
Stachyose, wt %	det. ∗	3.2		1.0	2.9	n. det. ∗∗	
Verbascose, wt %	det. ∗	2.1		1.4	1.9	n. det. ∗∗	
Hydrogen cyanide, mg/100 g	2.3	2.0		1.2			
Sinapine, wt %				0.004			
Isoflavones				det.			
Lipoxygenases, units/g	det. ∗			50			

∗ det.=detectable.
∗∗ n. det.=not detectable.

Table 26. Occurrence of antinutritive factors in roots and tubers [78][79][80, 89][90][91]

	Potato	Sweet potato	Cassava	Yam	Taro
Alkaloids, mg/100 g	6 (α-solanine, α-chaconine)			det. ∗ (dioscorine)	
Chlorogenic acid	det. ∗			det. ∗	
Hemagglutinins	det. ∗			det. ∗	
Protease inhibitors (trypsin + chymotrypsin)	det. ∗	det. ∗		det. ∗	det. ∗
Hydrogen cyanide, mg/100 g			55		
Tannic acid, µg/100 g			50		

∗ det. = detectable.

and trypsin inhibitor (in defatted soybean flour from 80 to 10 units per milligram of protein). The goitrogenic activity (low molecular oligo- or glycopeptides), the antivitaminogenic activity (e.g., rachitogenic activity), and the mineral-binding capacity of phytate (results have been obtained for calcium, phosphorus, zinc, iron, manganese, copper, and molybdenum) are partially reduced by heat treatment. At the same time the available lysine is reduced (from 5 to 2.5 g/16 g of nitrogen), and the nutritional value of the protein is increased (for defatted soybean flour the digestibility is increased from 74 to 84, the biological value from 52 to 65, the PER from 0.5 to 1.9, and the cystine [*56-89-3*] availability from 45 to 75 %). Thus, pancreatic hypertrophy (see below) is prevented. In contrast to the lectins of other leguminous plants, the hemagglutinin activity of soybean has no connection with its protein value. The improvement in the nutritional value of the proteins resulting from the heat treatment of soybeans or their products is due to the destruction of trypsin inhibitors and denaturation of the proteins.

Pancreatic hypertrophy is caused by hypersecretion, which is a result of a reduction of active trypsin in the intestinal tract due to the formation of a trypsin – trypsin inhibitor complex in the presence of undigested protein. The goitrogenic residual activity of soybean protein after heating can be compensated for by iodine supplementation (potassium iodide). Supplementation with vitamin D or minerals is possible in the case of any remaining rachitogenic activity or the limited availability of minerals. No effect is achieved by iron supplementation with iron sulfate. On the other hand, iron pyrophosphate can be used to increase iron availability up to 90 %.

For the use of soybeans and soybean products in human nutrition, the fact that undigested oligosaccharides, primarily stachyose but also raffinose, cause increased gas production (flatulence) due to anaerobic fermentation is still important. Although soybean flour (defatted flour: 5.5 wt % stachyose, 1.1 wt % raffinose) still exhibits a marked effect, this is no longer the case with concentrates (1.8 wt % stachyose, 0.5 wt %

raffinose) after extraction of the active oligosaccharides.

Casein, lactalbumin, ovalbumin, beef or pork proteins can be shown to be more cholesterolemic and atherogenic than soybean, peanut, or wheat proteins, for example. However, the activity of these plant proteins is considerably influenced by other food components. While the lower effect of soybean proteins is retained on interaction with saturated fat, dextrose, or cellulose, adaptation to the higher activity of animal proteins occurs with potato starch or lucerne. It has been found that the cholesterolemic and atherogenic effect correlates with a high lysine : arginine ratio (casein 2.0, lactalbumin 2.6, soybean protein 0.9, and wheat proteins 0.4). Since lysine inhibits arginase activity in the liver, a hypothesis has been proposed that due to this inhibition a high lysine content results in greater arginine availability for the synthesis of an arginine-rich apoprotein, which is a component of a lipoprotein with atherogenic activity.

Long-term studies on humans have shown that when used as the main or even sole protein carrier, soybean proteins are, in practice, equally suitable for infants, children, young people, and adults. If the products are made carefully, supplementation with sulfur-containing amino acids is not necessary. Supplementation with methionine [*63-68-3*] is recommended only for newborns as a precautionary measure.

Faba Beans and Peas. As described for soybeans, the antinutritive factors of faba beans and peas can be minimized or eliminated by heat treatment and separation processes (extraction, precipitation, membrane separation). This also applies to vicine and convicine from faba beans. However, if protein and starch are enriched by air classification, an enrichment of antinutritive factors and off-flavor also occurs.

Potatoes. The toxic factors contained in potatoes are glycoalkaloids (especially α-solanine and α-chaconine), hemagglutinins (lectins), and trypsin and chymotrypsin inhibitors. The glycoalkaloids are found primarily in the skin and are, consequently, partially eliminated by peeling the potatoes. An acceptable glycoalkaloid content is 6 – 20 mg per 100 g of fresh potatoes, but the average concentration in cultivated potatoes seldom exceeds 10 mg/100 g. The hemagglutinins and protease inhibitors are almost completely destroyed by heating. Thus the toxic components of potato are largely eliminated during the production of potato protein concentrates (precipitation from solution and heat treatment).

Lucerne and Clover. White clover contains linamarin (2-hydroxyisobutyronitrile-β-D-glucopyranoside) [*554-35-8*] and (*R*)-lotaustralin (2-methyllinamarin) [*534-67-8*]. During the destruction of leaf tissue, the glycoside is cleaved by the endogenous enzyme linamarase [*9031-18-9*] (EC 3.2.1.21) and hydrogen cyanide can be formed. Hydrogen cyanide-forming capacities of between 3.2 and 3.52 ng per gram of fresh weight have been found in clover. The toxicity is reduced considerably by the destruction of the leaves and heating. In human beings, poisoning can occur at concentrations between 0.5 and 3.5 mg of hydrogen cyanide per kilogram of body weight.

L-Canavanine (2-amino-4-guanidinoxybutyric acid) [*543-38-4*] has been found in lucerne. It acts as an arginine antagonist and interferes with the metabolism of ribo- and deoxyribonucleic acids. The toxic effect of this component is minimized by heating.

Lucerne and red clover also contain estrogenically active isoflavanoids, especially coumestrol [*479-13-0*] (lucerne: 57 ppm, clover: 49 ppm) and formononetin [*485-72-3*] (lucerne: 14 ppm, clover: 38 ppm). These phytoestrogens are heat resistant. The exposure of humans to these exogenous estrogens is very slight (about 1/1000 of the hormone equivalent of conventional oral contraceptives). These toxic components are most likely minimized considerably by the precipitation of lucerne and clover proteins. Therefore, such protein concentrates have been employed successfully in human nutrition.

2.4.3. Food Laws

During its 18th meeting in Geneva (July 3 – 14, 1989), the Codex Alimentarius Commission approved "Guidelines for the Use of Non-Meat Protein Products in Processed Meat and Poultry Products." According to these guidelines, standardized nonmeat protein products can be employed in principle for functional and other pur-

poses and for the partial substitution of meat proteins [92].

In this connection, a general standard for plant protein products and recommendations for international standards for soybean protein products and wheat gluten have been elaborated. In fact, the following minimum values with regard to the raw protein content (N×6.25) have been stipulated [93].

Plant protein product	40 wt %
Soybean protein flour	50 wt %
Soybean protein concentrate	65 wt %
Soybean protein isolate	90 wt %
Wheat gluten	80 wt %

National legal regulations for the use of plant protein products in foods vary considerably. A compilation is contained in the report "Plant Protein Production and Its Use in Human Nutrition" of a working party at the fifth meeting of the Codex committee for plant proteins in Ottawa (February 6 – 10, 1989) [94, 95].

2.5. Proteinaceous Products from Microalgae

The exploitation of natural stocks or of mass cultures of larger marine algae (seaweeds) serves as a basis for important industries (e.g., in California and Japan) that produce mainly alginates, carrageenates, or agar (→ Polysaccharides). In addition *red* and *brown marine algae* are highly valued in eastern Asia and considered a delicacy [96], stringy *green algae* are consumed, e.g., in southern Asia. Red and brown algae contain little protein in comparison to edible green algae [97]. The masses of the *blue-green alga Spirulina* occurring in Lake Tchad are used as a local supply of protein [98].

Of the thousands of species of microalgae that occur in inland waters or seas, only a few have been produced in mass cultures until now [99]. The yields per unit area depend on the solar irradiation and the temperature. Year-round production is possible only in warm countries.

Dried microalgae from mass cultures are internationally traded and sold as specialities ("natural pharmaceuticals," etc.) in the form of protein-rich pills or, more often, powders. Despite the relatively high protein quality of microalgae meals, prices of at least $ 14/kg preclude wider application. Nevertheless, protein-rich microalgae are of considerable importance because they are also produced in wastewater (e.g., in Southeast Asia) and are used on the spot to produce fish or animal feed, without being put on the market.

The algae – bacteria biomass obtained from wastewater until now has been used systematically as feed only in individual cases [100]. It contains about 10 – 40 wt % of bacteria, flagellates, or other nonalgal components.

Properties. *Physical Properties.* Microalgae meals (residual water 5 ± 2 wt %) [99] have a more or less intensive green color because of their chlorophyll content. Meals from blue-green algae are more bluish green because they contain the water-soluble pigment phytocyanin. The microalgae biomass produced for the isolation of carotinoids has a bright orange color in the case of *Dunaliella* species (β- carotene) and a brick-red color in the case of *Haematococcus pluvialis.* The consistency of microalgae meals is finely powdered when the raw mass has been freeze dried or spray dried and flaky after drum drying. With drum-dried material, the initially very low bulk density can be doubled by mechanical compacting (Table 27).

The water-binding ability of microalgae meals is relatively high and is influenced by the type of drying. Good nonoverheated products bind approximately 5 L of water per kilogram of dry substance [101].

Chemical Composition. The composition of microalgae products is variable and is influenced by the culture conditions and the climate (e.g., Table 9.1. in [102]).

Although seaweed dry matter sometimes contains relatively little protein (Table 28), microalgae are among the most protein-rich plant raw products (Table 29). Blue-green algal substance differs from the comparable green algal substance in that it has a lower lipid content.

As calculated from the sum of the amino acid concentrations, protein nitrogen accounts for 87 ± 2 % of the total nitrogen of *Spirulina* and 89 ± 2 % of the total nitrogen of *Scenedesmus.*

Due to the analytical techniques employed, data on the amino acid content are subject to uncertainties in measurement (Table 30), e.g., in the case of the important amino acid methionine. For this reason, the protein value determined nutritionally and physiologically might

Table 27. Bulk densities of some microalgae meals (original values)

Type of product	Type of drying	Bulk density, g/L	
		Original values	Mechanically compacted
Blue-green algae meal (*Spirulina platensis*)	spray drying	194	250
Green algae meal (*Scenedesmus obliquus*)	drum film drying	83	170
Algae – bacteria biomass (main component: *Scenedesmus falcatus*)	drum film drying	88	180

Table 28. General chemical composition of seaweed (marine algae) (values in percent g dry matter [103])

Species of algae	Raw protein	Raw fat	Carbohydrate	Ash
Brown algae				
Asophyllum nodosum	3 – 10	3 – 4	56 – 58	16 – 25
Laminaria angustata	6 – 8	1 – 3	60 – 70	17 – 32
Macrocystis meal (USA)	6	0.4	45	43
Green algae				
Enteromorpha linza	23	2	55	23
Undaria pinnatifida	14 – 18	0.5	46 – 68	30 – 38
Red algae				
Porphyra tenera	30 – 36	1	48	13
Gracilaria coronopifolia	9	0.3	70	20

Table 29. General chemical composition of microalgae meals compared to commercially produced yeast powder (*Candida lipolytica*, made by BP) and soybean seeds (values in percent of dry matter, according to [104, 105])

Material	Raw protein	Raw fat	Carbohydrate	Ash	Nucleic acids
Green algae meal (*Scenedesmus obliquus*)	50 – 60	12 – 14	10 – 17	6 – 10	3 – 4
Blue-green algae meal (*Spirulina maxima;* Mexico)	60 – 71	6 – 7	13 – 17	6 – 9	3 – 4
Yeast powder (*Candida lipolytica*)	64	9	15	7	8
Soybean seeds	34 – 40	16 – 20	19 – 35	4 – 5	?

be a more reliable criterion of quality. Microalgae sometimes contain small amounts of rare amino acids [101] that are apparently lacking in higher plants.

Details on the composition of the lipid fraction (fatty acids, etc.) and the high vitamin concentrations in microalgae meals are given in [99]. The β- carotene concentrations in *Dunaliella salina* reach 14 % under salt stress [106]. *Scenedesmus* and *Spirulina* meals contain 1.3 –3.3 and 0.6 – 0.8 % chlorophyll and 0.3 – 0.7 and 0.3 – 0.4 % total carotenoids (based on dry matter), respectively. The xanthophylls, which may cause problems in the feeding of green algae to poultry, can largely be destroyed by thermal pretreatment of the raw product.

Production. Like all photosynthesizing plants, algae require light, carbon dioxide, and minerals for growth. *Mesophilic algae* grow best up to ca. 30 °C, whereas the optimum temperature for *thermophilic algae* is 35 – 40 °C. Algae that *grow autotrophically* utilize carbon dioxide as the almost exclusive source of carbon on exposure to light. Many microalgae are also capable of *heterotrophic growth* (i.e., they can utilize organic compounds in the dark). Several microalgae also show *mixotrophy* (i.e., on exposure to light, they utilize both CO_2 and dissolved organic substances). The culture solution must be cir-culated to attain good yields and prevent sedimentation.

Clean methods for the production of microalgae are characterized by the fact that the composition of the culture solution is largely defined. This ensures the production of pure species and high-quality products. Microalgae are frequently produced in batch culture. The culture solution is based on fresh water (for *Chlorella*), on alkaline brackish water (for *Spirulina*) or on seawater (for *Dunaliella*).

In the *continuous wastewater processes* for the production of microalgae, liquid waste or wastewater is the source of nutrients. The algae – bacteria biomass that accumulates as a byprod-

Table 30. Concentrations of the most important amino acids in green algae meal (*Scenedesmus obliquus*) and blue-green algae meal (*Spirulina maxima*) compared to average values for fodder yeast (*Candida lipolytica* made by BP), extracted soybean meal, and fish meal (values in g/16 g of total nitrogen, according to [110] and [111])

Amino acids	*Scenedesmus*	*Spirulina*	BP yeast	Soybean meal	Fish meal
Isoleucine	3.7 – 4.4	3.7 – 4.1	5.1	4.6	5.4
Leucine	6.6 – 9.3	5.6 – 5.8	7.4	7.3	7.7
Phenylalanine	3.6 – 5.0	2.8 – 4.0	4.3	4.0	5.1
Tyrosine	2.8 – 3.6	4.0	3.6	2.9	2.7
Threonine	4.8 – 5.8	3.2 – 4.2	4.9	4.2	4.0
Tryptophan	1.2 – 1.4	0.8 – 1.1	1.4	1.2	1.5
Valine	5.5 – 7.2	4.2 – 6.0	5.9	5.2	5.0
Arginine	5.3 – 6.3	4.5 – 6.0	5.1	5.0	7.7
Histidine	1.5 – 1.7	0.9 – 1.1	2.1	2.3	2.4
Lysine	5.0 – 5.7	3.0 – 4.0	7.4	7.0	6.5
Cystine/Cysteine	0.8 – 1.4	0.6 – 0.7	1.1	1.0	1.4
Methionine	1.2 – 2.1	1.6 – 2.2	1.8	1.7	1.4

uct of wastewater treatment can be used as a component of animal feed [107].

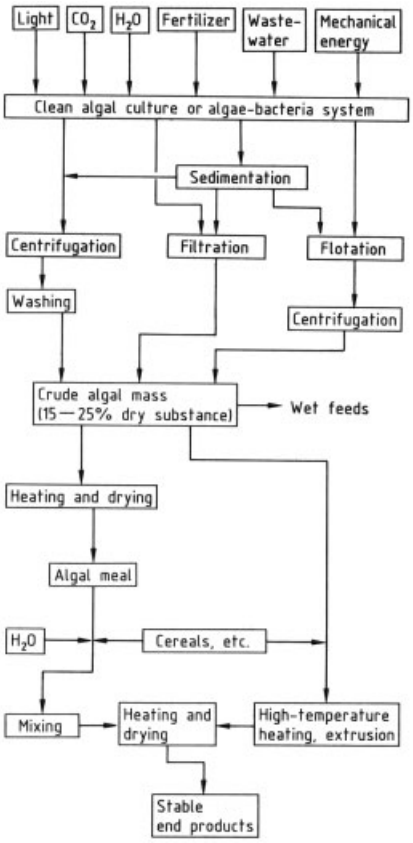

Figure 6. Production of algae meal from microalgae

Culture tanks for outdoor operation are available in many different designs [99, 108]. Horizontal raceway ponds made of concrete with a paddle-wheel circulation are usually preferred. The most advantageous depth of water is 10 – 15 cm for autotrophic "clean" cultures and 40 – 60 cm for wastewater systems. Meandering channels on inclined surfaces [109] presumably require less mechanical energy per hectare than horizontal troughs. Closed culture units consisting of meandering tubings or sandwich structures offer the possibility to produce very pure microalgae biomass at cell concentrations that are at least ten times higher than in open basins. For basins used in the wastewater process, the earth can simply be excavated and the resulting basin sealed with a layer of clay.

The pure carbon dioxide used for autotrophic cultures contributes considerably to production costs. It can, however, be replaced by CO_2-rich waste gas (e.g., from the production of ethanol [110]). A flow diagram for the production of algae meal from microalgae is shown in Figure 6. Details can be found in [99, 107, 108]. Cultivation is followed by separation and processing of the raw algae mass into the particular final product. During processing, the raw mass is sterilized by flash pasteurization. In the case of green algae, this step is absolutely necessary to make the product sufficiently digestible.

Separation is especially laborious because the cell suspensions contain very little solid matter (0.04 – 0.1 %). Before further processing, the proportion of solid matter must be directly or gradually increased to at least 15 % and, if possible, to more than 25 % [110]. Coarse, particulate algae (e.g., *Spirulina*) can be separated by simple filtration. Flocculation and flotation are

suitable first steps in the separation of the algae – bacteria biomass from treated wastewater [111].

Quality Specifications. An algal product must pass a nutritional physiological and toxicological safety test before it can be put on the market as food or feed. International recommendations exist for these tests [101]. Special safety aspects apply to the use of algae – bacteria biomass from wastewater for animal nutrition [112].

Producers. Among the numerous East Asian producers of *Chlorella* are Sunpack Co. Ltd., Tokyo, Japan; China Chlorella Industry Co. Ltd., Changhua, Taiwan.

Important producers of *Spirulina* meal are SOSA Texcaco SA, Mexico and Siam Algae Ltd., Bangkok, Thailand.

The German importers of microalgal products are Dr. Winfried Behr, Bonn, and Sanatur GmbH, Engen, Federal Republic of Germany.

Storage. Fresh microalgae meals have a weak, vegetable, largely neutral taste that is lost within a few days on exposure to air because, especially in the case of green algae meals, they quickly become rancid due to autoxidation [113]. Therefore, the products should be vacuum packed or stored under an atmosphere of protective gas. Decay is prevented if fresh algae meal or heat-digested algae mass is processed into pellets or pasta. Final products of this type retain a perfect taste quality even after several years of storage.

Uses. Approximately 3000 t of microalgae meal are put on the market each year. Roughly half of this amount is processed into natural pharmaceuticals in Japan, Korea, Taiwan, Europe, and the United States (*Chlorella* tablets, powder, and extract; *Dunaliella* tablets for cancer prophylaxis; and *Spirulina* tablets). The other half (including ca. 1500 t/a of *Spirulina* meal from Mexico) is used for aquaculture (cultivation of fish, mussel, and shrimp larvae, addition to aquarium-fish feed, etc.). In many places in Southeast Asia, algae – bacteria suspensions from wastewater ponds are added to pig, fish, or poultry feed. The amount of algae used for these applications is probably several times larger than the total volume of commercially produced microalgae meal. The addition of algae suspensions to the water of cattle, which has been practiced in southern Russia for decades, promotes a weight increase in the animals and stimulates milk production [114].

Nutrition Physiology, Toxicology. All the microalgae meals tested are good sources of protein for rats, pigs, carp, grass carp, and even humans [101, 107, 108]. In the feeding of chicken, however, not more than half of the conventional protein carriers can be replaced by microalgae protein without causing a definite reduction in growth rate.

The microalgae products that have been toxicologically tested more carefully (especially *Chlorella, Scenedesmus,* and *Spirulina*) until now have been considered toxicologically safe, provided that no harmful substances from the fertilizer or the environment have accumulated in them [101]. If microalgae are used to a significant extent in human nutrition, the amounts must not exceed certain safety limits because of their high nucleic acid content (ca. 4 % based on dry substance).

Economic Aspects. Profitability calculations and cost analyses of the various microalgae processes are given in [99, 107], and [108]. Pure microalgae meal costs about $ 14 – 20/kg. To sell substantial amounts of algae meal on the feed market, the present retail price must drop to less than $ 0.60/kg. This would be possible, if at all, only for production on a wastewater basis.

3. Animal Proteins

3.1. Fish Protein Concentrates (FPC) [115–119]

More than 25 % of the world's fish catch is converted into animal feed [120]. Fish proteins are among those with high nutritional value. That is why processes are intensively being investigated that might be suitable for direct production of human food from sources that are normally used to prepare fish meal. The different protein classes contained in cod (in wt % of crude protein) are listed as an example in what follows [121]:

Sarcoplasmic proteins	20 – 30
Myofibrillar proteins	65 – 75
Myosin	40
Actin	15 – 20
Tropomyosin	3
Collagen	2 – 5
Nonprotein nitrogen- containing compounds	13 – 14

To achieve a satisfactory shelf life for fish-based products, most of the fat and water must be extracted from the raw material. Isolation of certain protein classes is not a goal. Depending on the process used, nonprotein nitrogen (NPN) substances and varying proportions of the sarcoplasmic proteins are removed. Proteins may also be hydrolyzed or modified by chemical reactions.

The dewatered and defatted fish protein concentrates (FPCs) are powdery and vary in color from white to brown; they either are odorless or exhibit fishy, rancid, or oily notes. According to the Protein Advisory Group of the FAO, three types of solid FPCs are defined (see Table 31); liquid concentrates are of no importance. The term fish protein concentrate should not be used for fish sauces produced in the Far East. Also dried, salted, or smoked products, which still exhibit anatomical macrostructures (such as klipfish, stockfish, and other dry fish products) are not included in this term (even though these products have an increased protein content). In this section, only those products are treated whose technological or quasi-technological production data has been published.

Table 31. FAO requirements for fish protein concentrates [122]

	Type A, %	Type B, %	Type C (fish meal)
Lipids	< 0.5	*	
Moisture	< 10	< 10	no requirements
Protein	> 75	> 60	
Ash (acid soluble)	< 0.5	< 0.5	
Fluoride	< 250 mg/kg	< 250 mg/kg **	
Available lysine	> 6.5	> 6.5	

* No limitation.

** The FDA in the United States requires an upper limit of 100 mg/kg; the Food and Drug Directorate in Canada, an upper limit of 150 mg/kg [115].

3.1.1. Raw Materials (see Table 32)

The first attempts to produce FPC were performed with lean fish species such as cod [123] or less fatty fish such as hake [124]. Currently these valuable fish species are converted aboard trawlers into quick-frozen (QF) fish fillet products and therefore are not used for FPC. Species that are underused or not used for direct human consumption and thus are a possible source of FPC include a number of small- to medium-sized pelagic, mostly fatty fish (scombroids, anchovies, sardines, menhaden, capelin, sand eels), deep-sea fish (*Myctophidae*), and antarctic krill [131]. Possible catches of these species can be estimated only roughly:deep-sea fish up to 10^8 t/a, squid 10^7–10^8 t/a, and krill ca. 10^6 t/a (the world catch of aquatic animals in 1989 amounted to 99 534 600 t, including 13 776 700 t from aquaculture and inland fisheries [120]).

Use of offal from fillet production lines is of no importance because of the low amounts available and the undesirable composition of the resulting FPC (high content of fluoride and phosphates, off- color). The use of freshwater fish as raw material has also been described [132, 133].

In the United States and Canada legal restrictions govern which species may be processed into FPC: in the United States, herring-type fish (clupeids), hake species, menhaden, and California anchovy; in Canada, osmerids such as capelin and smelt, as well.

Fish stocks that may be used for the production of FPC must be accessible for the entire year to keep production capacities as constant as possible and must be available in large quantities to guarantee homogeneous raw material. A single production plant should process at least 500 t of raw material daily in order to be economical. A program for calculating the optimal design of the technological process and the process parameters is available [134].

3.1.2. Production [116, 135]

Many different processes have been proposed for the preparation of *low-fat FPC* (*type A,* Table 31). Despite differences in details, these processes are all based on three common principles:

1) Enrichment of protein content
2) Reduction of microbial growth by decreasing water content
3) Stabilization of the product against alterations due to chemical reactions of lipids by removing most of these

Table 32. Composition of marine raw material (wt % based on wet weight)

	Crude protein (N×6.25)		Edible part		Reference
	Whole fish	Edible part	Fat, wt %	Ash, wt %	
Lean fish		13 – 24	0.1 – 0.9	0.7 – 1.4	[121, 125]
10 species	15.3 ∗ (12.7 – 17.2) ∗∗	17.3 ∗ (15.2 – 18.9) ∗∗			[116]
Fatty fish		17 – 22	9 – 18	0.7 – 1.4	[121, 125]
Mollusks		9 – 20	0.3 – 4.8	1.6	[121, 125]
Squid		16 – 20	0.8 – 2.2	1.2 – 1.7	[121, 125], [126–128]
Crustaceans		7 – 23	0.2 – 4.5	0.5 – 1.3	[121, 125], [129]
Krill		13.4	1.6 – 7.0	3.2	[130]

∗ Average. ∗∗ Variation.

The fat and water content can be reduced by different combinations of heating, extraction, application of pressure, and drying. The flow diagram given in Figure 7 as an example may be modified by leaving out some steps or inserting others.

The raw material is highly perishable and should be processed immediately after landing; alternatively, it may be stored chilled. If extraction is used as the first processing step the raw material can be stored as a mixture with the extraction solvent [136]. Removal of bones from the raw material by bone separators is recommended; this lowers the ash content of the resulting FPC from 14 – 20 wt % to 3 – 6 wt % and the fluoride content from a maximum of 760 mg/kg to 60 – 160 mg/kg. Simultaneously, its protein content increases from 80 to 90 wt % [135, 137]. Separation of bones from the final FPC is ineffective and results in lower yields.

Decanting. After coagulation of protein by heating at 80 – 100 °C, part of the water and fat may be removed by decanting the liquid components in centrifuges or in screw presses. This procedure is identical to that used in the production of fish meal.

Extraction. Alternatively, most of the *fat and water* may be removed by extraction with 2-propanol, 1,2-dichloroethane, xylene, *n*-octane, *n*-butanol, or *tert*-butanol. However, only the first two solvents have been used on a technical scale [138, 139]. The protein quality reportedly deteriorates when 1,2-dichloroethane is used as extractant [137]. Processes involving two solvents or the removal of water by azeotropic distillation are more costly and have not been inves-tigated widely. Water and fat can be extracted from dried raw material such as fish meal by 95 % ethanol [140]. A preextraction with *n*-hexane which improves the product has been reported [141].

If residual water and fat are to be removed from the press cake or if unpressed material is to be processed, only solvents that are at least partly miscible with water [i.e., acetone, 2-propanol, *tert*-butanol, methanol, ethanol, glycol monomethyl ether (Cellosolve), and dioxane] can be used for extraction. Of these, only 2-propanol and ethanol have been applied successfully. The other solvents lead to residues in the extracted material; acetone contributes a permanent off-odor. 2-Propanol cannot be efficiently removed from the extracted material simply by drying; however, its content could be lowered from 10 000 mg/kg to 30 mg/kg by addition of water to the material and repeated drying or by steam distillation. Countercurrent extraction and dewatering have been applied successfully [142].

Detergents such as dodecyl or lauryl sulfate remove *fat* at 65 °C; the high content of detergent (up to 40 %) that remains in the FPC is a disadvantage [143].

A similar problem is encountered when the raw material is suspended in fish oil (as heat carrier) and water is removed at 66 °C in a vacuum thin-layer evaporator. The product still contains > 12 wt % fat after pressing and must be defatted by an additional extraction [144]. The advantage of this process is that it can be run at reasonable costs and leads to an FPC that contains the water-soluble constituents of fish; however, these constituents may lead to storage problems because of microbial attack or devel-

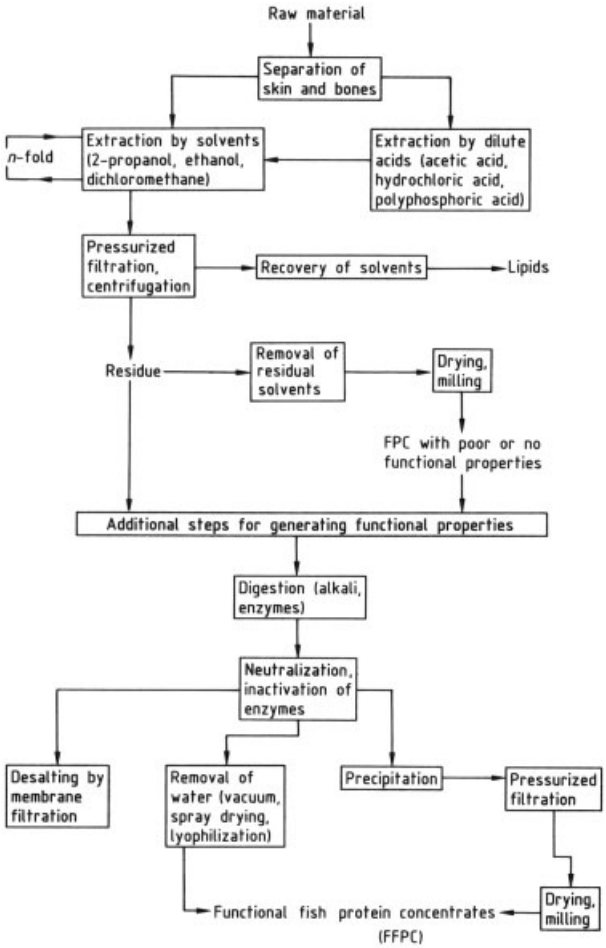

Figure 7. Flow sheet for the preparation of fish protein concentrates (FPC and FFPC)

opment of off-odors [145]. In some processes, these substances are extracted by dilute acids (hot acetic acid, phosphoric acid, or cold hydrochloric acid with sodium chloride) prior to fat extraction [123, 146, 147].

Flocculation. Flocculation of the major part of lipids as lipoprotein complexes by solubilizing the raw material with alkali and acidifying the solution to pH ≤ 5.5 has also been reported [148]. This method can be used only if completely soluble protein hydrolysates are produced. They are then obtained by evaporation of the residual solution. Only a low yield of proteins is obtained by further acidification of the solution after flocculation. By adding sodium hexametaphosphate at pH ≈ 4 a phosphate-containing protein precipitate can be separated from the oily – watery supernatant by centrifugation (fish contains ca. 80 % oil and water). Washing the centrifuge residue with an azeotropic mixture of water and propanol yields a satisfactory protein concentrate, with a phosphate content exceeding by only 1 % that of FPC prepared by extraction [149].

Bleaching, Drying, and Final Processing. The fish protein concentrates can be bleached with hydrogen peroxide at pH 8 to 9.5 [148, 150–152]. Drying may take up to 2 h at up to 120 °C without substantial deterioration of the nutritive value of the protein [153, 154]. The water-solubility and the emulsion stabilizing properties of the protein decrease when drying temperatures increase from 90 – 100 to 110 – 120 °C [155]. The unavoidable removal of solvent residues by water vapor also leads to a remarkable decrease of these two functional prop-

erties. Recovery of the solvent is essential for process economy: otherwise up to 10 % of the solvent used may remain in the extracted product. The construction of drying facilities must guarantee good hygienic conditions; otherwise the plate counts of aerobic microorganisms in the product will increase rapidly. If countercurrent extraction with 2-propanol is used the bulk of the fish oil can be recovered by centrifuging. For this purpose the extract from the last extraction step, which is hot and rich in oil, is mixed with the extraction mixture from the first step, which contains the cold fish. The two phases that form are separated in a centrifuge. When 2-propanol is recovered by distillation, phosphoric acid must be added to the mixture to be distilled to neutralize the basic compounds contained in fish (di- and trimethylamine). Milling of the resulting FPC to 200 mesh (74 µm) is necessary for many purposes to avoid a sandy mouthfeel.

Functional Fish Protein Concentrates (FFPC). The term functional fish protein concentrate applies to an FPC prepared in such a way that it still possesses some properties that are valuable for its technological use and are characteristic of native proteins. Examples are foam formation, coagulation, formation and stabilization of emulsions, and water-binding capacity. An FFPC with good rehydration and emulsification properties can be produced if all the processing steps described above are performed below 52 °C [150] and 2-propanol is removed by warm air saturated with water vapor [156]. The FPC produced by extraction of raw material with ethanol – hexane at pH 2.5 in the presence of 5 wt % NaCl also exhibits emulsification and rehydration capacities [157]. An FFPC has been prepared from krill meat by extracting the meat twice with a fourfold amount by volume of ethanol at 5 – 10 °C in the presence of 1 – 2 % NaCl at pH 7.4 – 7.6 (addition of $NaHCO_3$). A completely soluble FFPC has also been prepared from minced krill meat by extraction with 2-propanol, alkaline digestion at 37 °C and pH 10.5, and subsequent freeze drying: this material exhibits good emulsion and foam-forming properties [158].

Another approach for preparing FFPC uses chemical reactions to reintroduce functional properties into the protein that has been denatured by the normal preparation of FPC (see also Fig. 7). Addition of alkali [123] or proteolytic enzymes [140, 141, 159, 160], as well as acetylation and succinylation of the nucleophilic side chains of proteins [161–163], have been described among others. When alkali is used the formation of toxicologically hazardous lysinoalanine [164] must be monitored carefully.

A singular method for reintroducing functional properties to FPC is milling the product in a ball mill for 100 h [165]; the product is reportedly suitable for forming a resilient *kamaboko* (a Japanese product usually prepared from minced washed fish by grinding and adding NaCl and starch) [166]. Many reports state that the functional properties of FFPC decrease with time [167]: to some extent this can be prevented by the addition of carbohydrates (corn syrup, glucose, sorbitol) prior to drying the FFPC. In contrast, an unchanged quality has been reported when FFPC from filefish [168] and shark [169] was stored for three months at 25 °C, as well as for storage of powdered soup from anchovy and mussels [170].

Production of Fish Protein Concentrates with a Higher Fat Content (Type B). The high degree of unsaturation in fish lipids [171] is responsible for the very limited shelf life of fat-containing fish protein concentrates at normal temperature in air. This is one of the main reasons why type B products have not been investigated with respect to production, properties, etc. Another reason is that consumers in most industrialized countries dislike products with an intense taste or even a slight rancidity. A North American Air Defense Command (NORAD) study demonstrated, however, that people in other parts of the world think differently [172]. An FPC of type B can be obtained rather simply by heating the fish to 95 °C, followed by separation of the protein in a decanting centrifuge or pressing device [173]. This product contains 7.1 –8.8 wt % fat and is storable up to two years. Fish protein that is to be mixed with starch- containing foods is preferably dried in a drum dryer (20 s at 150 °C surface temperature) because this method gives a product with good swelling properties [145]. The use of FPC type B in Africa, Asia, the Far East, and the Caribbean has been described [172].

If the raw meat is kneaded with the addition of NaCl (ca. 1 %) and extruded into ethanol at

Table 33. Production of fish protein concentrate

Country	Years of production	Amount, t/a	Type	Solvent	Reference
Germany	1936 – 1948	700	A	ethanol (acetic acid, NaOH)	[123]
Norway	1972 – 1992	60 000 ∗	B		[176, 177]
RSA	1956 – 1958	400	A	ethanol	[140]
Morocco	1962 – 1973	1400	A	ethanol, hexane	[118]
Chile	1961 – 1964	not known	A	hexane, ethanol	[118]
United States	1969 – 1971	300 ∗	A	dichloroethane, 2-propanol	[178]
United States	1971 – 1972	1500	A	2-propanol	[178]
Sweden	1971 – 1977	9000	A	2-propanol	[150]

∗ Figures not known exactly.

Table 34. Chemical composition of fish protein concentrates [116, 150, 176, 178, 179, 181, 182]

Main components, wt % (dry weight)	Type A	Type B
Crude protein, (N × 6.25)	80 – 93	70 – 75
Ash	11 – 22	22.3
Lipids	0.07 – 0.55	7.1 – 8.8
Volatiles (H_2O)	2 – 8	2.3
Inorganic components		
Calcium, wt %	2.8 – 5.1	1.8
Phosphorus, wt %	1.9 – 3.4	1.7
Potassium, wt %	0.6 – 1.1	1.3
Sodium, wt %	0.34 – 0.44	0.5 – 1.2
Magnesium, wt %	0.16 – 0.28	0.13
Aluminum, mg/kg	67 – 775	
Iron, mg/kg	147 – 770	14 – 108
Manganese, mg/kg	32 – 129	
Copper, mg/kg	6 – 21	1.2 – 3.0
Zinc, mg/kg	112 – 163	65 – 180
Fluorine, mg/kg	58 – 162	100, 270
Selenium, mg/kg	1.8 – 3.8	1.3 – 5.3
Essential amino acids, (wt % of crude protein)		
Valine	4.4 – 6.3	5.2
Isoleucine	3.9 – 5.4	4.4
Leucine	6.5 – 8.4	7.5
Threonine	4.0 – 5.0	4.4
Lysine	7.5 – 10.7	7.8
Available lysine	7.7 – 8.2	6.9
Tryptophan	0.8 – 1.5	1.2
Cysteine + methionine	3.2 – 4.1	4.0
Tyrosine + phenylalanine	6.2 – 8.8	7.3
PER (protein efficiency ratio based on casein)	106 – 111	
Vitamins, mg/kg		
A		4.2
B_1		1.0
B_2		7.3
B_6		5.7
B_{12}		0.25
E		33.6

5 – 30 °C, a product with a defined texture can be obtained, which is suitable for replacing part of the beef in hamburgers [174, 175]. Extraction with ethanol at − 20 °C, followed by extraction with hexane at − 20 °C, reportedly optimizes the functional properties [167].

Commercial Products. Norsildmel (Bergen, Norway) sells a fish protein concentrate type B (Norse Fish Powder) with 67 – 80 wt % raw protein, a maximum of 12 wt % fat, and ca. 100 mg/kg fluoride [176, 177]. All other production plants (some are cited in Table 33 that had been funded by FAO or by national governments ceased production after some time.

3.1.3. Nutritive Value

Fish protein concentrates have high nutritional value because of their high protein content and the protein's well-balanced composition of essential amino acids [121, 125, 179, 180]. The high lysine content of marine protein makes such products particularly well suited for fortifying a cereal-based diet, poor in lysine. The composition of fish protein concentrates is shown in Table 34. Both types, FPC A and B, not only are protein-rich foods of high value but also can contribute considerably to the daily mineral intake in the diet [180, 181, 183]. Fish protein concentrate type B provides a considerable supplement of zinc, calcium, iron, fluoride, iodine, and selenium, and is a good source of *n*-3 or ω-3 fatty acids (i.e., unsaturated fatty acids with double bond at C_3 of the carbon chain when counting from the methyl terminal group of the carbon chain), carnitine [183], and vitamins A, E, and the B group. The PER of an FPC from anchovy equals that of milk whey, whole egg, and mechanically deboned turkey meat, and is superior to egg white and different cereal flours [180]. In growth experiments with rats a high body weight gain has been demonstrated [140, 184].

3.1.4. Product Properties and Uses

The FPCs produced hitherto are not foods with a character of their own, but are added to other foods to reduce the content of more expensive protein sources or to fortify foods with protein. The use of functional FPC for introducing technological properties (e.g., froth forming, emulsifying, and swelling properties) has been demonstrated on an experimental, but not a commercial, scale. An overview of applications of FPC and FFPC is given in Table 35; other examples are provided in [185]. If FPC is used as a meat substitute in hamburgers, the increasing replacement of meat by FPC is often paralleled by a decreasing sensory rating [175, 186]. The substitution of a considerable part of beef in hamburgers (up to 50 %) by an FPC with limited functional properties from krill has also been described [193]; results of a sensory evaluation of the products are lacking.

3.1.5. Completely Soluble Hydrolysates from Fish Protein

The extensive hydrolysis of fish protein by different proteolytic enzymes has been reported frequently [196–206], although whether such a process has been performed on a commercial scale is unknown. The same holds true for alkaline hydrolysis. The only commercially available product of this type was the Wiking-Eiweiß [123], which was produced in Germany in the 1940s by alkaline digestion of cod fillet, followed by an extraction step with ethanol and dilute acetic acid. A slightly gray, odorless, easily digestible powder was obtained, whose taste was reminiscent of hen egg white. On heating, no coagulation of protein resulted, this product was used in mayonnaise, baked goods, sweets, ice cream, and pharmaceutical preparations.

3.2. Egg Proteins [207]

3.2.1. Structure and Composition of the Egg

Even in ancient times, egg was a popular food in all regions and religions of the world. Man almost instinctively chose a product that possesses excellent nutritive qualities. Being the germ for the development of a "new life" (embryo), the egg has an optimal composition and provides essential nutrients and active agents in an easily digestible form. It is also largely shielded from external influences because it is packed in its own shell, which accounts for about 10 % of the total weight. Although eggs from all types of poultry are edible and have all the qualities for human nutrition mentioned above, the word eggs usually refers to hen's eggs. These hens are bred and kept especially for laying and can produce an average of 300 eggs per year per animal.

The structure of an egg is described in Figure 8, the composition of various types of egg is given in Table 36. The average weight of a hen's egg is 60 g, the white and yolk contributing 58 and 32 %, respectively. Heavy eggs contain relatively more egg white than light eggs. Eggs are sold as fresh or shell eggs, either loose or packed in special boxes. In the EC, eggs are available in three grades of quality and seven grades of weight, according to quality regulations [208]. In addition, the three egg fractions white, yolk, and whole egg (i.e., total content of the egg or a mixture of egg white: egg yolk = 1.8 : 1) are available in various forms, namely, liquid, frozen, or dried and powdered. Additives such as salt or sugar are sometimes used. Egg products are subject to strict hygienic controls [209]. The contents of opened eggs must be pasteurized before they are processed further and can be put on the market.

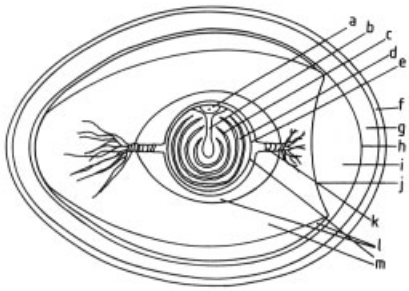

Figure 8. The structure of a hen's egg
Yolk:
a) Germ disk and vesicle; b) Formative yolk; c) White yolk; d) Yellow yolk; e) Yolk membrane Shell; f) Cuticula; g) Egg shell with pores; h) Shell skin; i) Air space; j) Shell membrane; Egg white; k) Chalaza; l) Thin white; m) Thick white

Table 35. Examples of use of fish protein concentrates in other foods

Type	Accepted amount, wt % of final product	Country	Product type	Reference
A	7	Germany	whipped egg white	[123]
A	30 % of protein	United States	angel cake	[163]
A	7	Sweden	emergency rations	[187]
A	70 % of protein	Sweden	baby food	[150]
A	4 – 5	United States	milk substitute	[163, 189]
A	4	Sweden	wheat meal	[190]
A	5 – 10	India	bread, puri	[191, 192]
A	10	Chile	roasted bread	[193]
A	3 – 9	Chile	bread	[193]
A	2 – 4	RSA	brown bread	[140]
A	10	United States	tortilla	[189]
A	3 – 6	United States	pasta	[189]
A	4	United States	cracker	[189]
A	10	United States	cake	[189]
A	5 – 10	United States	pea soup	[189]
A	10 % of meat	United States	frankfurter sausage	[163]
A	10 – 20	United States	hamburger	[145, 188]
A	16 – 32	Japan	hamburger	[175]
A	5 – 7.5	India	bread	[194]
A	2 – 10	India	chapati	[169]
A	20 – 60 % of meat	Greece	hamburger, meat balls	[186]
A	20 – 30 % of meat	Canada	hamburger, sausage	[195]
A	50 % of meat	Korea	hamburger, meat balls	[168]
A	5 – 6	Egypt	biscuit	[133]
A	10	Malaysia	cracker	[160]
B	2 – 38	Africa, Middle East, and Far East	local meals	[172, 182]

Table 36. The composition of eggs of various types of poultry

	Goose	Turkey	Duck	Hen	Guinea fowl	Pigeon	Quail
Average weight, g	160	86	70	60	43	20	10
Yolk, wt %	36	33	36	32	37	19	34
Egg white, wt %	52	56	54	58	48	71	57
Shell, wt %	12	11	10	10	15	10	9
Solids, wt %	29.4	26.3	30.3	26.4	27.2		26.5
Organic substance, wt %	28.2	25.5	29.3	25.6	26.3		25.7
Protein, wt %	14.0	13.1	13.7	12.8	13.5		12.7
Fat, wt %	13.0	11.7	14.4	11.8	12.0		12.0
Carbohydrate, wt %	1.2	0.7	1.2	1.0	0.8		1.0
Ash, wt %	1.2	0.8	1.0	0.8	0.9		0.8

The most important quality criteria of eggs include size, shape, color, shell stability, and inner properties (the latter quality criteria also apply to egg products). Inner properties include degree of freshness, utilization or functional properties (e.g., viscosity, foaming ability, and emulsion suitability), sensory properties (odor, taste, color) and nutrient content [210]. The nutrient composition of eggs, broken down into the major nutrients, minerals, vitamins, lipids, and amino acids, is listed in Table 37. This overview shows the impressive number of essential nutri-ents present in egg. That is why the protein composition of egg has been set as the basis for the optimum biological value = 100. A comparison of eggs with other foods is presented in Table 38.

3.2.2. Proteins in Egg

The chemical binding forms of egg proteins determine the material structure and especially the functional properties of eggs or egg fractions. Amino acids and other substances are united into large molecules, the proteins. Some of these

Table 37. Nutrients in hen's egg ∗ [211]

Nutrient	Amount in 100 g of egg substance		
	White	Yolk	Whole egg
Major nutrients			
Water, g	88.07	48.76	74.57
Calorific value, kJ	204	1546	661
Protein (N×6.25), g	10.14	16.40	12.14
Fat, g	traces	32.93	11.15
Carbohydrate, g	1.23	0.21	1.20
Crude fiber, g	0	0	0
Ash, g	0.56	1.70	0.94
Minerals			
Calcium, mg	11	152	56
Iron, mg	0.03	5.58	2.09
Magnesium, mg	9	15	12
Phosphorus, mg	11	508	180
Potassium, mg	137	90	130
Sodium, mg	152	49	138
Zinc, mg	0.02	3.38	1.44
Vitamins			
Vitamin C (ascorbic acid), mg	0	0	0
Vitamin B_1 (thiamine), mg	0.005	0.254	0.087
Vitamin B_2 (riboflavine), mg	0.285	0.436	0.301
Niacin, mg	0.089	0.069	0.062
Pantothenic acid, mg	0.241	4.429	1.727
Vitamin B_6 (pyridoxine), mg	0.003	0.310	0.120
Folic acid, µg	16	152	65
Vitamin B_{12}, µg	0.065	3.803	1.547
Vitamin A (retinol) (IE)	0	1839	520
Lipids			
Saturated fatty acids	0	9.89	3.35
Myristic acid	0	0.09	0.03
Palmitic acid	0	7.27	2.46
Stearic acid	0	2.53	0.86
Monounsaturated fatty acids	0	13.16	4.46
Palmitoleic acid	0	1.10	0.37
Oleic acid	0	12.06	4.08
Polyunsaturated fatty acids	0	4.28	1.45
Linoleic acid	0	3.66	1.24
Linolenic acid	0	0.09	0.03
Arachidonic acid	0	0.28	0.09
Cholesterol, mg	0	1602	548
Amino acids			
Tryptophan, g	0.156	0.241	0.194
Threonine, g	0.451	0.890	0.596
Isoleucine, g	0.618	0.939	0.759
Leucine, g	0.883	1.396	1.066
Lysine, g	0.625	1.110	0.820
Methionine, g	0.394	0.417	0.392
Cystine, g	0.251	0.291	0.289
Phenylalanine, g	0.638	0.714	0.686
Tyrosine, g	0.407	0.706	0.505
Valine, g	0.759	1.000	0.874
Arginine, g	0.592	1.136	0.777
Histidine, g	0.230	0.394	0.293
Alanine, g	0.654	0.821	0.709
Aspartic acid, g	0.897	1.370	1.204
Glutamic acid, g	1.415	2.007	1.546
Glycine, g	0.380	0.496	0.404
Proline, g	0.381	0.682	0.482
Serine, g	0.748	1.359	0.923

∗ The analytical data for eggs from other types of poultry should be similar.

form complexes with each other; the stability of these complexes depends on external conditions. This explains why the ratio of thick to thin egg white shifts constantly, the formation and stability of foam vary greatly, and gel formation and agglutination are subject to changes. The same applies to emulsifiability, which is mainly a characteristic of the yolk. The emulsifiability is a property of lipoproteins (i.e., compounds formed from proteins and fats).

Table 38. Biological value of food

Food	Essential amino acids	Essential fatty acids
Whole egg	100	100
Milk	92	5
Cheese		40
Meat	85	
Veal		10
Beef		15
Pork		65
Mutton		55
Chicken		60
Fish	81	
Cartilage	31	
Maize	72	
Soybean	72	
Rice	71	
Wheat	59	

Variations of functional properties are mainly caused by changes in temperature, pH, time (duration of storage), and their mutual interactions. With increasing storage time, the pH of the egg increases through CO_2 formation and its subsequent permeation through the shell, the solid matter and the proportion of yolk increase, and foam formation is improved. In contrast, the viscosity, foam stability, and proportion of egg white decrease. These processes are accelerated if eggs are stored above 8 °C [212]. The changes produced in liquid eggs due to pasteurization at high temperature are especially problematic. The difficulty is to choose an adequate temperature for eliminating microorganisms (especially *Salmonella*) while avoiding temperatures that could deteriorate functional properties. The instructions for pasteurization must be adhered to strictly and are checked mainly on the basis of temperature and holding time. An example is given in Table 39 [213].

The original properties of these fractions can be maintained largely by addition of organic acids, metal ions, or salt. With viscosity as an example, Figure 9 shows how final products can differ from each other depending on their pretreatment. In a larger experiment with exactly defined egg material, the variations (in %) given in Table 40 were measured for some quality criteria [213].

Table 39. Pasteurization conditions for eggs

	Pasteurization temperature, °C	Pasteurization time, min	Heat holding time, s ∗
Egg white	57	1.75 – 3.5	840
Egg yolk ∗∗	65	1.75 – 3.5	1080
Whole egg ∗∗	65	1.75 – 3.5	540

∗ In heat-holding container before being cooled to 2 °C.
∗∗ Can be increased to 70 °C if salt is added to the yolk and whole egg.

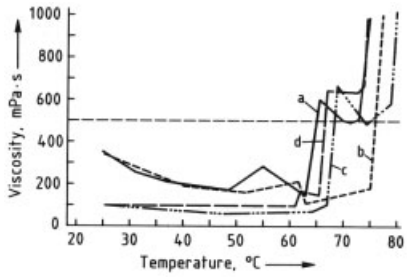

Figure 9. The viscosity of egg white after various pretreatments
a) Native egg white (control); b) Fermented egg white; c) Pasteurized egg white; d) Homogenized egg white

3.2.2.1. Proteins in Egg White

Due to improved electrophoretic techniques, the identification of proteins in egg white has made substantial progress [213]. Initially only the most important fractions could be identified; currently, 12 or more bands (i.e., stripes of ranging intensity on the test substance, see Fig. 10) can be distinguished on the pherograms. The number of investigations has also increased considerably [214] so that the results are becoming more significant all the time. Finally, instead of evaluating the colors via several experimental steps [214] they are nowadays measured directly by photometry. The most frequently used meth-

Table 40. Variation of some quality criteria of egg (%)

Cause of variation	pH	Viscosity	Conductance	Foam formation	Foam stability
Age of hen	30.1	2.2	0.1	10.9	11.9
Age of egg	14.0	4.0	0.1	2.0	3.5
Technological treatment	13.0	89.7	99.6	85.1	79.5

ods today are polyacrylamide gel electrophoresis (PAGE) and sodium dodecyl sulfate – PAGE (SDS – PAGE), which is used mainly for the analysis of whole eggs [213, 215, 216].

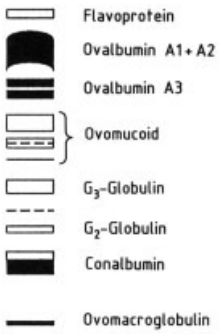

Figure 10. Schematic drawing of the bands on a pherogram of egg white separated by electrophoresis

In the following description of egg white proteins, the percentage of respective proteins in total egg white, the isoelectric point (IEP given as pH), and the molecular mass (M_r in daltons) are given for each case. This compilation is based on several publications (mainly [213] and [217]).

Ovalbumin, 54 wt %, IEP 4.5 – 4.8, M_r 46 000, is a glycoprotein that can be divided into three fractions A 1, A 2, and A 3 (fraction A 1 contains two phosphate units per molecule, A 2 one, and A 3 no phosphate unit per molecule). They differ in their electrophoretic mobility (1 < 3) and possess two, one, or no phosphate groups. This protein is stable up to 62 °C at pH 7. Further, it contains free sulfhydryl groups. On heating, S-ovalbumin is formed, which has a coagulation temperature of 92.5 °C [217].

Conalbumin, 12 – 13 wt %, IEP 6.05 – 6.6, M_r 76 600 – 86 000, is a glycoprotein that corresponds to serum albumin (ovotransferrin) of the chicken. It is heat labile and contributes to good foam formation, as can be seen in the extinction curves from pherograms of egg white from various types of poultry (see Fig. 11). In the case of water fowl (duck and goose) the conalbumin peak (VI) is specially small compared with that of land fowl (hen, turkey, and quail). Duck and goose egg whites are extremely difficult to foam.

Changes in the activity of conalbumin depend not only on temperature but also on pH. The aforementioned stabilizing additives (metal ions and organic acids, see 3.2.2) are added during pasteurization to maintain the functional activity of conalbumin.

Ovomucoid, 11 wt %, IEP 3.9 – 4.5, M_r 28 000, is a protein that contains disulfide bridges, oligosaccharides, and other carbohydrates. Due to its structure, this protein is very heat stable and is considered to be an enzyme inhibitor (trypsin inhibition).

Globulin, G 2: 4 wt %, IEP 5.5; G 3: 4 wt %, IEP 5.8, M_r of both 36 000 – 45 000. The two bands of globulin appear at positions 12 and 14 in the PAGE determination. Little is known about this protein. It seems to be heat stable and contributes to foam formation by increasing the volume of foam.

Lysozyme, 3.5 wt %, IEP 5.5, M_r 14 300 – 17 000 (also called muramidase), is often referred to as globulin G 1. It acts as a preservative and is responsible for the bactericidal effect of egg white. It forms complexes with other proteins (e.g., with ovalbumin, conalbumin, and ovomucin).

Ovomucin, 1.5 – 2.9 wt %, IEP 4.5 – 5.0, M_r 10 000, consists of glycoprotein chains (fibrils) that are linked through disulfide bonds. The gel structure of thick egg white is attributed to the

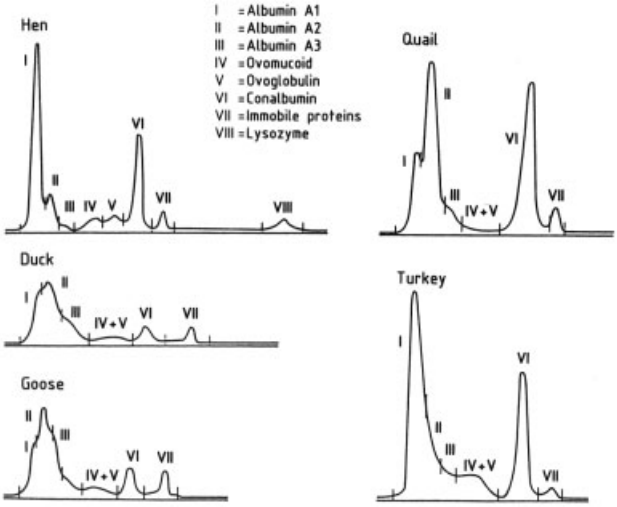

Figure 11. Extinction curves of various egg white samples after electrophoretic differentiation [214]

action of ovomucin, because only 1/4 of ovomucin is found in thin egg white.

The other electrophoretically identifiable proteins occur in egg white only in small amounts.

Ovoinhibitor, 0.1 – 0.5 wt %, IEP 5.1 – 5.2, M_r 46 000 – 49 000, is a moderately heat-stable protein that is stable up to 78 °C in the strongly acidic range. Trypsins are inactivated by ovoinhibitor.

Ovoglycoprotein, 0.5 – 1.0 wt %, IEP 3.9, M_r 24 400,contains hexose,glucosamine,and acetylmuramic acid among other compounds.

Flavoprotein, 0.8 wt %, IEP 3.9 – 4.1, M_r 32 000 – 36 000, or ovoflavoprotein binds riboflavin and is relatively heat stable.

Ovomacroglobulin, 0.5 wt %, IEP 4.5 – 4.7, M_r 760 000 – 900 000, the egg white protein with the highest molecular mass, was identified only in 1966 and is relatively heat stable.

Avidin, 0.05 wt %, IEP 9.5 – 10.0, M_r 68 300, is a basic protein that binds biotin. It is very heat stable and exerts an antibacterial effect.

3.2.2.2. Proteins in Egg Yolk

If egg yolk is centrifuged, the solids separate into yolk droplets (i.e., the plasma) and the granulose fraction [217]. Proteins are found in both fractions (see Fig. 12). Along with the readily water-soluble parts, the yolk contains 48.7 % water, 16.6 % protein and 32.6 % fat and carbohydrates. Most of the fat is bound to protein [217].

Plasma Proteins. The *low-density lipoproteins* (LDLs) are found in the plasma. Low-density lipoproteins account for 33.7 wt % of the yolk. The fat part of the lipoproteins consists of neutral fats and phospholipids. The protein part is involved mainly in emulsion ability; it is thermostable up to 75 °C and is active especially at the boundary layers.

Livetins, 5.5 wt % of the yolk, are present in three forms as water-soluble proteins:

α-Livetin, 1 %, IEP 4.8 – 5.0, M_r 80 000, serum protein
β-Livetin, 3 %, IEP 4.8 – 5.0, M_r 45 000, glycoprotein
γ-Livetin, 1.6 %, IEP 4.8 – 5.0, M_r 150 000, serum protein

The distribution of the three corresponds approximately to the ratio 2 : 5 : 3.

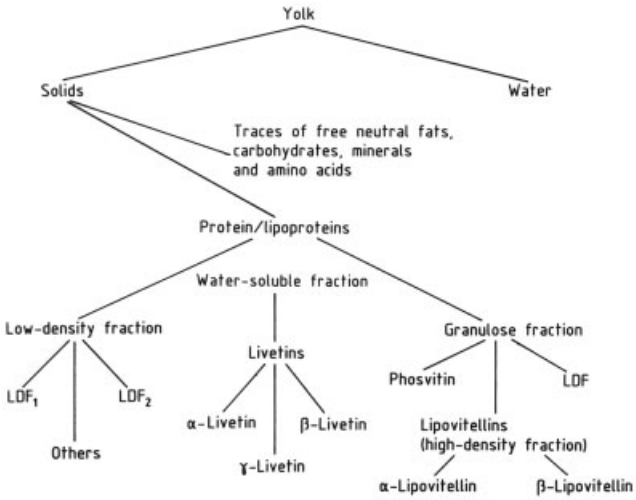

Figure 12. Composition of hen egg yolk [213]

Granulose Fraction. The granulose fraction (13 wt % of the yolk) contains mainly *high-density lipoproteins* (*HDL*s), which are called *lipovitellins.* The lipid part of these compounds (ca. 30 wt %) includes phospholipids and cholesterol. The glycoprotein part contains galactose, mannose, hexosamine, and sialic acid, among other substances; it is also involved in emulsion ability. α- and β-Lipovitellins differ in their mobility and in the proportion of phosphate bonds.

Phosvitin, approximately 2.1 wt % of the yolk, M_r 36 000 – 40 000, contains chiefly the amino acid serine. The phosphorus content is high (10 wt %). Two types of phosvitin occur:

α-Phosvitin, M_r 160 000 (15 – 22 wt % fat and 0.5 – 1.2 wt % protein)
β-Phosvitin, M_r 190 000 (17 – 22 wt % fat and 0.3 – 0.5 wt % protein).

Phosvitin exerts antioxidative effects [217], and binds iron and copper.

Emulsion Ability. Since the yolk proteins primarily determine the emulsion ability, attention must be drawn to the fact that emulsifiability can be evaluated in many ways. In general, definite amounts of fat, water, and yolk are mixed, in accordance with the "mayonnaise method," and the drainage is measured after a defined period of time [213].

3.2.3. Egg Products

In 1990, 35 160 t of egg products were traded in Germany in the following forms

49.5 %	whole egg (liquid/frozen)
17.1 %	egg yolk (liquid)
12.5 %	egg white (liquid/frozen)
6.5 %	whole egg (dried)
5.8 %	egg white (dried)
4.6 %	egg yolk (frozen)
2.2 %	egg rolls (precooked long egg)
1.8 %	egg yolk (dried)

The products are bought by food producers (bakeries, butchers, and meat processors) but also for the production of hair-care products, by the film industry (for coating film rolls), and tanneries.

Finished and semifinished products made from eggs include egg powder; liquids or pastes for making scrambled eggs with various ingredients (ham, mushrooms etc.); spreads as a supplement or substitute for butter, margarine, or cold meats; egg sausage in the form of solid egg white and yolk with the same cross-section as that of a sliced hardboiled egg; egg-based alcoholic and nonalcoholic beverages; and analogous low-cholesterol products.

Lysozyme is the most important substance isolated from egg white. It has a bactericidal effect and is used as a preservative in milk products, meat products, fresh fruit, and baby food [218].

3.3. Proteins from Blood

Blood is the most important byproduct of slaughtering it consists predominantly of protein and water, and is a valuable protein resource. The protein content in blood from domestic meat animals is equivalent to 6 – 7 % of the lean meat content in the carcass [219]. Because of the nutritionally valuable composition of blood and the large amounts of blood available from slaughterhouses, utilization of such blood seems reasonable. A requirement for blood that is utilized in human consumption is its hygienic collection.

3.3.1. Physical Properties of Blood

Blood consists of the intercellular substance, plasma, and suspended cells, which include erythrocytes (red cells), leukocytes (white cells), and thrombocytes (blood platelets). Cattle blood has a packed cell volume of 40 %; swine blood 42 %. Bovine blood has a density of 1.060 g/cm^3. The corresponding value for pig blood is1.040 g/cm^3. The freezing point of plasma is −0.5 to −0.6 °C. Blood is sometimes called "liquid meat" because its composition is similar to that of lean meat (Table 41) [220].

Table 41. Blood composition

	Cattle blood	Swine blood	Lean meat
Water, wt %	80.5	79.2	71.6
Protein, wt %	17.8	18.5	20.0
Fat, wt %	0.1	0.1	4.1
Carbohydrate, wt %	0.1	0.1	
Salts, wt %	0.9	1.0	1.1

3.3.2. Production

If blood is collected hygienically, it can be used for human consumption. Blood that is not hygienically recovered is converted into blood meal.

Collection of Edible Blood. Special precautions must be taken to minimize the contamination of blood with microorganisms during sticking of animals and subsequent storage of blood. The blood of healthy animals is normally sterile but can easily be contaminated in the slaughtering operation. There are two general categories of blood collection – closed and open systems. Closed systems have obvious advantages in preventing contamination. If good hygiene is maintained with cleaning-in-place installations, the microbiological quality of the blood should be less than 10^4 total plate- count of microorganisms per milliliter of blood [221].

The usual methods for blood collection involve the use of specially designed hollow knives discharging into cleanable receivers. The collected blood should be cooled as soon as possible to obtain preparations with low bacterial counts. Usually, blood from small numbers of animals is collected at a time and the result of veterinary inspection is awaited before the blood is mixed into a larger bulk.

Separation Process. Figure 13 shows the main fractions into which whole blood may be separated and Table 42 shows the composition of blood fractions. Unless anticoagulant is added, freshly drawn blood will coagulate or clot. Coagulation is prevented by the addition of an *anticoagulant solution* (sodium citrate or phosphate solution). When completely mixed with blood, this prevents the transformation of fibrinogen to fibrin and keeps the blood fluid. Addition of 3 g of trisodium citrate per kilogram of blood is sufficient as an anticoagulant. Increasing concentrations and amounts of trisodium citrate result in a linear increase of the plasma yield and a corresponding decrease of the protein concentration because the added citrate remains mainly in the plasma fraction [220].

Centrifugation of treated blood yields a red cell fraction (30 – 40 wt %) and a light-colored plasma fraction (60 – 70 wt %). The blood cells (thick blood) consist of cell wall material and proteinaceous material (almost entirely hemoglobin).

Continuous agitation also prevents coagulation by causing the fibrinogen to precipitate as fibrin. After separation, a red cell fraction and a serum fraction are obtained.

Blood Plasma Concentration and Preservation. After separation the plasma fraction can be stored at 4 °C for a maximum of 3 – 4 days. Often liquid or frozen plasma can be utilized "in-house" in meat food products. Alternatively, the

Table 42. Composition of blood fractions

	Blood	Blood plasma	Red cell concentrate	Dried plasma	Dried globin
Dry solids, wt %	18 – 20	7 – 9	28 – 37	96 – 97.5	96.5
Protein, wt %	16 – 18	6 – 9	28 – 38	70 – 96	91 – 95.4
Albumin	2.2	3.3			
Globulin	2.8	4.2			
Fibrinogen	0.3	0.4			
Hemoglobin	10.0		30.0		
Stroma	1.7		5.1		
Fat, wt %	$<$ 1	0.1 – 1	1	0 – 1.5	0
Carbohydrate, wt %	$<$ 1	$<$ 1	$<$ 1		
Salts $*$, wt %	2	1.5	1 – 3	$\leq$ 2	1 – 6
Water, wt %	80 – 82	90 – 91	61 – 63	2.5 – 7.0	3.5

$*$ Contents depend on quantity and type of anticoagulant used.

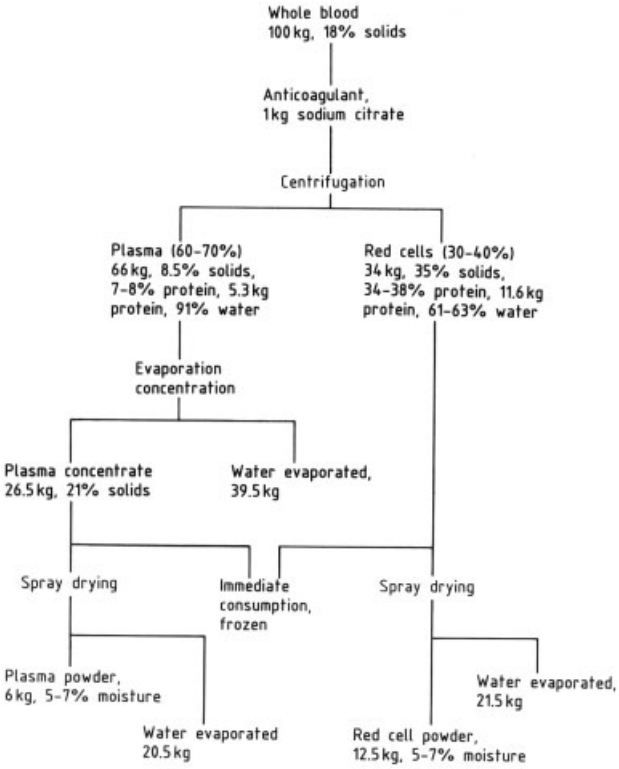

Figure 13. Separation of blood [222]

plasma is dried to remove water and thus to facilitate storage and also to reduce the costs of transportation.

When blood fractions are dried, great care must be taken to minimize denaturation. Plasma proteins begin to coagulate and denature at temperatures as low as 40 – 50 °C. Concentrating the blood plasma prior to drying is usually more economical. Concentration is performed by evaporation (falling-film evaporator or centri-term evaporator) or membrane filtration. The high salt content of plasma, which is a result of the added anticoagulant, can be lowered by ultrafiltration of the plasma. Normally the next step is spray drying of the concentrated plasma. Combination of ultrafiltration and spray drying gives a dried blood plasma protein isolate with 96.5 % protein and 2.5 % moisture [222].

Decolorization of Red Cell Concentrate (Thick Blood). When blood plasma is separated from the erythrocytes, the major portion (60–70 wt %) of the protein content remains in the erythrocyte fraction as hemoglobin. This hemoglobin imparts an undesirable dark red color to the product. Hemoglobin is often considered a very good catalyst for the oxidation of unsatured fatty acids; however, at high concentrations, it appears to have an inhibitory effect on fat oxidation [223]. The color can be lightened by emulsification of blood with fat and subsequent treatment of the blood – fat emulsion with a homogenizer [224]. The effect of this process is due to the light scattering of the dispersed tiny fat droplets. With regard to functional and nutritional properties of unaltered hemoglobin the emulsification process is advantageous but is not able to completely mask the heme pigment color when the product is cooked.

Several processes exist for *decolorizing* hemoglobin. The decolorized hemoglobin protein is generally referred to simply as "globin" in the literature. Traditionally, preparation of colorless globin is based on acid cleavage of hemoglobin and separation of globin from heme by organic solvents. Acidified cold acetone [225], methyl ethyl ketone [226], and ethanol – methanol [227] have been used. Tybor [228] converted hemoglobin to choleglobin by aerating a hemoglobin solution at pH 4 in the presence of ascorbic acid prior to the treatment with acidified acetone and developed a continuous process on a pilot scale. A serious problem with acetone-isolated globin is an off-flavor component that adheres to the protein molecule. All these extraction methods are not feasible economically. Alternatively, methods have been developed in which heme is removed from globin by adsorption on a medium (e.g., alginate) that is in contact with the acidified hemoglobin solution [229]. A chromatographic method has been presented in which heme and globin were separated by using a carboxymethyl cellulose (CMC) column in the acid pH range [230]. This method has been improved and developed further for industrial application [231].

Degradation of hemoglobin by proteolytic enzymes is an alternative method for avoiding dark pigmentation. The amino acids and peptides from the hydrolyzed protein can be separated from the heme by ultrafiltration or centrifugation. Olsen [232] and Hald-Christensen [233] used alcalase, whereas Drepper [234] used an acid proteinase. A degree of hydrolysis between 18 and 20 % is recommended, to give a yield of about 78 – 85 % [232]. Below a degree of 18 % hydrolysis the yield is unsatisfactory, and above 20 % a bitter taste becomes dominant in the product. Therefore a supplementary treatment with active carbon is recommended.

Hemoglobin can also be *lightened by bleaching* using hydrogen peroxide or ozone [235, 236]. The protein obtained is completely insoluble in water and has no functional properties.

Another method for decolorizing hemoglobin is acid-catalyzed *polymerization of released free heme.* Polymerized heme is separated by centrifugation and globin remains in solution. This polymerization is dependent on the concentration, type of hemoglobin derivative, pH, and temperature [237]. A second stage is necessary to remove residual heme [235, 237].

Blood Meal. Blood meal is a dark-brown granular product with 5 – 8 wt % moisture, produced by drying whole or thick blood. Blood is coagulated and then decanted and separated into blood meal and blood water. The blood meal is cooked, pressed to remove moisture, and dried or baked. The dried product is ground and used as feed or fertilizer.

To conserve the energy required for evaporation, newer techniques use a concentration process prior to drying the blood. The strained blood is preheated to 55 – 58 °C in a storage tank and pumped to a live steam coagulator where complete coagulation of blood proteins occurs. The coagulated blood is then dewatered by decanting, which removes 75 % of the water. The fine-crumb dewatered blood protein is then led to a cooker dryer for final drying (Fig. 14).

High temperatures and extended periods of drying decrease the nutritional value of blood meals. Blood can be dried by spray drying, freeze drying, ball drying, or roller drying. It should not be dried above 80 °C to maintain its digestibility. Spray drying is the most popular method of drying blood. Ball drying is accomplished with 20 – 30-mm balls, which are coated with a thin layer of blood and circulate through a drying chamber [239].

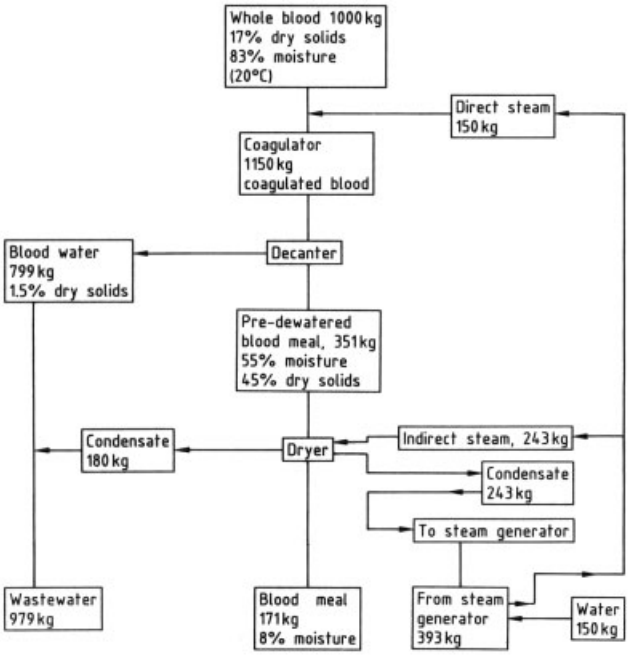

Figure 14. Blood meal production (Westfalia Separator Company [238])

3.3.3. Functional Properties of Blood Proteins

Blood is quite functional in food products primarily because of its albumin content [240]. On heating, albumin gels and retains fat and water. *Plasma,* whose protein content consists of 50 % albumin, is the best water and fat binder of the blood fractions. The firmness of plasma gel, which depends on concentration, temperature, pH, and ionic strength, is often used as an indicator of functionality. If drying is performed properly, plasma is highly soluble in water. The most unique functional property associated with plasma proteins is their exceptional binding ability when used in meat product formulations.

Blood proteins have excellent foaming capacity. For this reason, they can be used as a substitute for egg whites in the baking industry [240]. *Globin* has a high degree of surface activity as reflected in foaming properties and emulsifying activity. Table 43 summarizes the functional properties of blood plasma and globin.

Hemoglobin is a highly water-soluble protein with a high capacity to emulsify fats and develop stable foams. Removal of the heme group greatly reduces the stability of the globin protein molecule, and consequently the protein is much more sensitive to denaturing agents and heat. The solubility of globin is reduced in the pH range 6.5 – 8.5. The method of isolating globin significantly affects the functional properties of the final product. Acetone-treated globin prepared by the method of Tybor [228] possesses good water-binding capacity and emulsifying properties at pH ≤ 6 but lacks the gelation property [242]. Hayakawa [243] was the first to report that globin prepared by CMC – column chromatography forms a firm gel after being heated in the narrow pH range 5 – 6. He later used globin as a model protein for studying the process of gel formation [244].

Table 43. Functional properties of blood proteins [241]

Property	Plasma	Globin
Water solubility	good	good at pH < 6
Swelling with water	poor	very good
Viscosity	low	fair, high after heating
Emulsifying capacity	fair	good
Emulsifying stability	good, poorer with salt	good at pH < 6, better with salt
Foaming	good	very good
Gel formation	very good, sets $> 70\,^{\circ}$C, strength increased by salt	none, forms a thick cream

3.3.4. Nutritional Value of Blood Proteins

Blood is rich in heme iron, which is biologically the most available form of iron. The amino acid composition of whole blood plasma and globin proteins is shown in Table 44 and compared with the ideal values for a well-balanced diet as recommended by the FAO [241]. The low chemical scores of blood or globin indicate that they should be used to complement other proteins.

The amino acid profile of globin shows that its high lysine content and lack of isoleucine should make it an effective complement to certain vegetable proteins that have a high isoleucine content but are low in lysine.

3.3.5. Uses

Human Food. Blood proteins can be added to most food products. The most widely accepted food use of blood proteins is in meat products to increase protein levels and water-holding capacity. Because of its color the main use of *hemoglobin* is confined to products that traditionally have a dark color, such as black sausage, black pudding, blood breads, and blood cookies. Other possible uses for blood proteins are in the baking industry because of their foaming abilities and as nutritional supplements to any cereal food because of the high lysine content. *Blood plasma* is used in comminuted meat products. The addition of plasma to sausages of this type improves their water retention and reduces fat losses. *Globin* is used as a texture modifier and meat extender in ground-meat products [245].

Animal Feed. Blood meal is a basic feed product made from blood. Blood is also used in pet foods. Here its functional properties can be applied (e.g., in the formation of meatlike chunks).

Nonfood Uses. Blood can be used as nitrogenous fertilizer, glue, and fire-extinguishing foam [241].

Special fractions of animal blood processed under strict hygienic conditions are utilized in human medicine. Blood contains all the vital substances for human activity (i.e., proteins, mineral salts, vitamins, and enzymes). The enzyme orgotein (superoxide dismutase) is extracted from red cells. It is an antiinflammatory that detoxifies oxygen free radicals in the body. Purified thrombocytes are also being tested for their suitability in medicine. Bovine-derived blood substitutes have good possibilities for medicinal uses [246]. Apart from pharmaceuticals, animal blood fractions are used for biochemicals, diagnostics, and cosmetics (albumin, serum, plasma, hydrolysates) [246]. In the laboratory the most common uses of blood are as a nutrient for tissue culture media (vaccine production, monoclonal antibodies for mammalian tissue culture) and as a necessary ingredient in some agar cultures (transferrin, fibronectin, albumin).

3.3.6. Environmental and Legal Aspects

Blood is the main organic component in slaughterhouse wastewater [247] and the primary cause of slaughterhouse off-odors [248]. Slaughterhouses must collect the blood as completely as possible. Otherwise blood with a five-day biological oxygen demand (BOD_5) of 140 000 mg of O_2 per liter increases the organic loading of wastewater dramatically [247]. Even a BOD_5 of 250 000 mg of O_2 per liter of blood has been reported [249].

In most European countries, release of blood into wastewater is strictly forbidden. Nevertheless, nearly 15 % of the blood remains in the animal carcass and enters wastewater during the washing and splitting of the carcass [250].

3.3.7. Economic Aspects

Because of the high meat consumption and growing population, world meat production is increasing as are the resulting amounts of blood. Cattle and swine are the main sources of meat. The world meat production in 1990 [251] and calculated blood quantities are given in Table 45.

If the average yield of blood is 2.5 L per pig and 13 L per cow [250], this gives a protein source of 780 000 t/a, which is underutilized worldwide for religious, aesthetic, or social reasons. Scandinavian counries use about 70 % of the blood collected for human consumption; in Germany, only ca. 40 % is collected hygienically [253]. In Sweden, some 80 % of the available

Table 44. Essential amino acids in blood proteins (all figures in g per 100 g protein)

Amino acid	Whole blood	Plasma	Globin	FAO provisional pattern
Lysine	10.2	9.2	10.5	4.2
Threonine	4.3	6.3	3.8	2.8
Methionine	1.6	1.0	1.7	2.2
Valine	8.9	7.0	9.4	4.2
Phenylalanine	7.4	5.6	7.9	2.8
Leucine	13.0	10.1	13.8	4.8
Isoleucine	0.8	2.9	0.2	4.2
Tryptophan	2.0	1.9	2.0	1.4
Histidine	6.9	3.5	7.8	
Chemical score	19 %	45 %	5 %	100 %

Table 45. World meat production in 1990 [251] and calculated blood quantities

Country	Slaughter animal	Carcass weight, 10^6 t	Number of animals [a] $\times 10^{-6}$	Blood quantity [b], 10^3 t	Protein quantity [c], 10^3 t
United States	cattle	10.1	33.2	432.0	77.8
	swine	6.9	84.1	210.0	37.8
Former USSR	cattle	8.7	28.6	372.0	67.0
	swine	6.6	80.7	202.0	36.3
Eastern Europe	cattle	1.8	5.9	77.0	13.9
	swine	4.9	59.3	148.0	26.7
Western Europe	cattle	9.3	30.8	400.0	73.0
	swine	16.9	205.0	513.0	92.5
China	cattle	1.2	4.0	51.7	9.3
	swine	23.6	288.0	719.0	129.0
World	cattle	51.8	174.0	2260.0	406.0
	swine	69.5	847.0	2120.0	381.0

Assumptions used: [a] 304 kg/cattle and 82 kg/swine [252]. [b] 13 L/cattle and 2.5 L/swine [250]. [c] 18 % protein/L blood.

blood is said to be used directly as human food, 30 % of this as whole blood and 70 % after processing into plasma [241]; in Germany, up to 10 % of the blood plasma may be incorporated in cooked sausages.

More than 80 000 t of blood were collected and further processed into food products in the Soviet Union in 1970 and this amounted to only 21 % of all blood collected in that country. In the Soviet Union blood proteins are used as ingredient for various types of sausages and canned meat products.

The edible blood collected in the United States is not used in food products, but is fractionated to be applied for pharmaceutical purposes. Blood collected in accordance with the severe regulations issued by the Department of Agriculture (FSIS) may be used only in nonspecific meat products [240]. (Because of cultural taboos the meat industry in the United States does not utilize a valuable raw ingredient of abundant supply.)

4. Protein Hydrolysates and Derivatives

Protein hydrolysates are mixtures of peptides prepared by the chemically or enzymatically catalyzed hydrolysis of naturally occurring proteins. Protein hydrolysates are employed mainly in the food industry, in the technical sector, and as a raw material in cosmetics. Derivatives of protein hydrolysates are used widely in cosmetics and for technical applications. Total hydrolysates (see Section 3.1.5) and hydrolysates used in seasoning (e.g., soy sauce) [254] are not discussed here.

4.1. Protein Hydrolysates

4.1.1. Raw Materials

Being one of the fundamental building blocks of all living things, proteins are readily available in

nature. Examples of important protein sources for hydrolysates with technical applications are

1) **Animal Sources.** Slaughterhouse offal (skin, hair, blood, and nuchal ligament), milk, meat, fish, egg, silk, leather
2) **Plant Sources.** Press cakes from processing of oil seeds (soybean, cotton, almond, sunflower), protein from wheat, rice, potato, and peas

Some protein raw materials are available in different stages of enrichment (e.g., in the case of soybean as meal, concentrate, and isolate). The selection of a raw material depends on the particular application of the protein hydrolysate, the availability, and the cost of the protein.

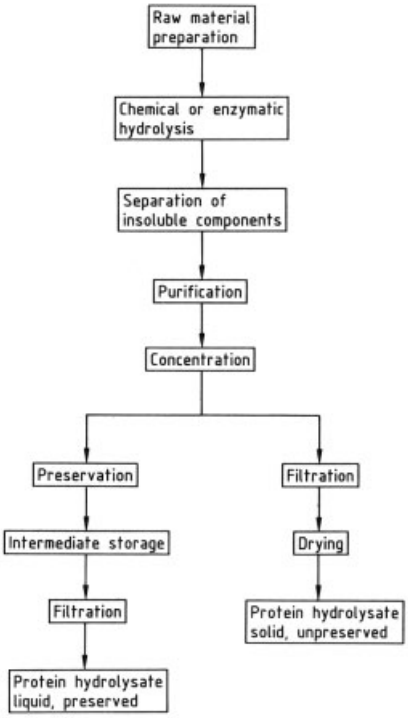

Figure 15. Flow-sheet of the partial hydrolysis of proteins

4.1.2. Production Processes

In principle, proteins can be partially hydrolyzed by using the following processes:

1) Chemical hydrolysis
   Acid catalyzed
   Alkali catalyzed
2) Enzymatic hydrolysis

The general flow sheet for the partial hydrolysis of proteins is presented in Figure 15.

Raw Material Preparation. Depending on the quality of the raw material and the intended application of the protein hydrolysate, preparation consists of measures to reduce ash content, improve odor and color, and support the subsequent hydrolysis (oxidative or reductive pretreatment).

Chemical Hydrolysis. Chemical hydrolysis can be either acid or alkali catalyzed. In the case of acid hydrolysis, sulfuric or hydrochloric acids are mainly used, whereas $Ca(OH)_2$ or MgO is chiefly employed in alkaline cleavage.

The reaction is performed at 80 – 100 °C, and pH < 1 or > 12. The duration of hydrolysis is 6 – 18 h. Yields of at least 80 %, based on nitrogen, are obtained.

Enzymatic Hydrolysis. Enzymatic hydrolysis is performed with proteases that cleave interior peptide bonds of the polymer chains (endopeptidases). In addition, exopeptidases are used, which cleave single amino acids or dipeptides (from the terminal carboxyl and amino groups). The reaction proceeds under considerably gentler conditions at a maximum of 65 °C and a pH range of $\approx 2 - 11$. The duration of hydrolysis is 0.25 –12 h. Color and odor of the products obtained by enzymatic hydrolysis are often better than those obtained by chemical hydrolysis. Enzymatic hydrolysis has the advantage of potential stereochemical specificity. This permits controlled processes and fewer side reactions. The amount of chlorohydrins and salt in the hydrolysate is considered to be smaller than in hydrolysates prepared by acid hydrolysis. The nitrogen yields are at least 75 %.

Purification includes treatment with activated carbon for improvement of color and odor, precipitation of cations for reduction of the ash content [302], separation of bitter substances by adsorption, and fractionation of the hydrolysates by membranes.

In the production of *liquid protein hydrolysates,* the concentrates are preserved and clarified after an intermediate storage period of several days to a few weeks. Unpreserved protein hydrolysate powders can be obtained by spray drying of clarified and sterile filtered, unpreserved solutions.

4.1.3. Uses

4.1.3.1. Nutrition

Human Nutrition. An enzymatic or chemical hydrolysis step is associated with the production of many staple and luxury foods. Examples are the coagulation of casein by the enzyme rennin in cheese making (→ Cheese, Processed Cheese and Whey, Chap. 3.2.) and the decomposition of chill haze by the enzyme papain in the clarification of beer [255].

Apart from the protein hydrolysis inherent in many production processes, protein hydrolysates are deliberately added to foods to influence product properties. These hydrolysates are obtained mainly with enzymes that specifically alter the typical functional properties of proteins [303], such as

1) Solubility by alteration of the isoelectric point
2) Water-binding capacity
3) Whipping properties
4) Emulsifying capacity
5) Foam formation and stabilization
6) Gel formation
7) Fat and flavor binding

The nutritive value of the native protein is retained. The change in the functionality of the protein hydrolysate compared to that of the original protein depends on the enzyme used and the degree of hydrolysis [256]. Changes in the functional properties of soybean, bean, potato, and whey proteins are described in [257–260].

Protein hydrolysates frequently have a bitter taste that is caused, e.g., by the amino acid composition of the protein, the enzymes used, and the degree of hydrolysis. The bitterness restricts the use in foods [256, 261, 304].

As a result of the favorable physiological effects of protein hydrolysates, such as easy digestibility or low allergizing potential, acid-soluble soybean protein hydrolysates [*68607-88-5*] are added to dietetic drinks for athletes [262]. For the same reason, special hydrolysates are used to feed patients with gastroenterologic diseases [256, 263]. Protein hydrolysates are increasingly employed in treatment of, e.g., pancreatitis, short bowel syndrome, Crohn's disease, and food allergies [304, 305]. Hypoallergenic hydrolysates are employed in diets for babies [306, 307], who are allergic to whey proteins. The German legal regulations governing the use of proteins or protein hydrolysates in food are described in [264].

Animal Nutrition. Feed additives based on fish protein hydrolysates [*91079-91-3*] (see Section 3.1.5) serve as a source of nitrogen in milk replacers used for calf breeding [265]. Offal hydrolysates are used as animal feed [266].

4.1.3.2. Cosmetics

Protein hydrolysates are used in cosmetics because of the following properties:

Hydrophilicity
Amphoteric character
Pronounced buffering effect
High substantivity
Good film formation
Action as protective colloid
High water binding capacity

Hair Cosmetics. In hair dyeing, the addition of protein hydrolysates not only produces a more uniform adsorption of dyes on the keratin fibers, but also results in less loss of color during subsequent rinsing [267]. The film formation of protein hydrolysates causes an improvement in the luster of the hair and a better dry and wet combability [308], [309]. Protein hydrolysates in shampoos or hair repair products penetrate through the cuticula into the cortex of the hair and, thus, are capable of repairing split ends [268].

The negative effects of bleaching and cold waving processes on the hair can be reduced by the addition of protein hydrolysates [269]. Especially in the latter process, the increased alkali solubility of keratin can be lowered by the protective action of protein hydrolysates [270].

Skin Cosmetics. The use of skin cosmetics containing protein hydrolysates not only results in a soft skin but also improves the hydration state of the skin. These properties of protein hydrolysates are based on their hydrophilic and amphoteric character [271–273]. The rough skin phenomenon often observed when strongly degreasing surfactants, e.g., bulk surfactants, are

used can be drastically improved by the addition of protein hydrolysates because of their protective colloidal action [274].

Dermatological Effect. The most important property of protein hydrolysates is reduction of the irritant action of surfactants on the skin and the mucous membranes. These properties are based on both the amphoteric character and the substantivity of protein hydrolysates, and are especially pronounced in the case of collagen hydrolysates [*9015-54-7*] because of their nonexistent antigenic action [275,276].

4.1.3.3. Technology

Technical uses of protein hydrolysates are based on their foam producing, dispersive, and wetting properties. Keratin hydrolysates [*69430-36-0*] are used as basic protein foaming agents in the formulation of extinguishing foams [277]. Collagen hydrolysate can be employed in the production of waterproof concrete [278]. The overall prolongating influence exerted by protein hydrolysates on the setting time of gypsum and the specific effect of side chains of the amino acids they contain are described in [279]. Protein hydrolysates are used as textile auxiliaries in the mercerization of cotton [280], for wool protection in the dyeing of wool [281], and as dispersing agents in the formulation of dyes [282].

Soybean protein hydrolysates are used as binding agents for paper [283]. Protein hydrolysates, e.g., fish protein hydrolysate [*91079-91-3*] [310], have been proposed as sources of nitrogen for the nutrient media for microorganisms [284]. In the fermentation of carbohydrates to lactic acid, increases in yield are achieved by the addition of protein hydrolysates from wheat or fish protein [285,286].Collagen hydrolysates [*9015-54-7*] can be used as feeds and fertilizers [311].

4.2. Derivatives of Protein Hydrolysates

Raw Materials. The derivatives are directly synthesized from protein hydrolysates, which are produced as described in Section 4.1.2.

4.2.1. Acylated Products

The acylation of protein hydrolysates with acylating agents (see Table 46) can occur not only on the amino functions that are released by hydrolytic cleavage, but also on nucleophilic groups in the side chains of the protein hydrolysates ($-NH_2$, $-OH$, $-SH$) [312].

Synthesis. Hydrolysates are normally acylated in an aqueous alkaline medium according to Schotten – Baumann. In preparative applications, only acid chlorides are used as acylating agents (Fig. 16), esters have been of theoretical interest only until now [287]. The residue R^1 (see Table 46) of the acylating agent is a long-chain alkyl group in the case of fatty acids, or a tricyclic alkyl group in the case of resin acids (abietic acid).

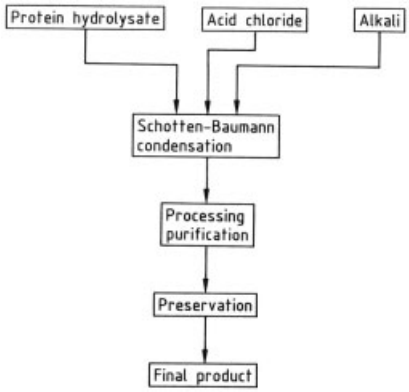

Figure 16. Acylation of protein hydrolysates

Schotten – Baumann condensation proceeds at 60 °C (max.) under pH control. Intensive stirring is required because of the two immiscible liquid phases. Higher temperature favors the formation of soaps by hydrolysis of the acid chlorides [288].

The condensation product is normally purified for cosmetic applications. The acylation products are precipitated from the aqueous solution by lowering the pH (< 3) and are then washed to eliminate byproducts [289]. A water-soluble salt of the acylation product is reformed by subsequent neutralization with a base that is allowed in cosmetics, such as KOH or triethanolamine. The product can also be processed by drying to give an unpreserved powder.

Table 46. Derivatization scheme of protein hydrolysates *

Educt	Reagent	Product	Substituent
$R-NH_2$	$R^1-C(=O)-X$	$R-N(H)-C(=O)-R^1$	R^1 = alkyl-, cycloalkyl-
	$CH_2(O)-CH-CH_2-N^+(CH_3)_2-R^2$	$R-N(H)-CH_2-CH(OH)-CH_2-N^+(CH_3)_2-R^2$	$R^2 = -CH_3$, *n*-alkyl-
	$Cl-CH_2-CO_2^-Na^+$	$R-N(H)-CH_2-CO_2^-Na^+$	
$R-CO_2H$	R^3-OH	$R-C(=O)-OR^3$	R^3 = alkyl-

Applications. Protein hydrolysates acylated with long-chain fatty acids [*68920-65-0*] (R^1 = *n*-alkyl, see Table 46) or resin acids [*68918-77-4*], [*147993-64-4*] (R^1 = cycloalkyl) are anionic surfactants. They possess amphoteric character in their hydrophilic part and therefore show an especially low irritant potential compared to other surfactants [276]. Their foaming and washing ability is roughly comparable to that of bulk surfactants (e.g., alkylether sulfates). Therefore they are used in cosmetics as surfactants with excellent dermatological compatibility. Detailed experimental studies are available that compare the skin and mucous membrane compatibility of various surfactants with that of acylated protein hydrolysates [294–296, 313]. The skin compatibility of acylated protein hydrolysates can be improved by increasing the molecular mass of the peptide part of the acylated protein hydrolysate [290–293]. In the formulation of cosmetics or detergents, addition of small amounts of acylated protein hydrolysates to the more strongly irritating bulk surfactants results in a more than proportional improvement of their compatibility with skin and mucous membranes [275], [314].

As a special cosmetic effect the refatening of washed hair is significantly delayed by the addition of protein hydrolysate – abietic acid condensates [*68918-77-4*], [*147993-64-4*] (R^1 = cycloalkyl-, see Table 46) to shampoo [297, 298]. In anti-dandruff shampoos, protein hydrolysate – undecylenic acid condensate [*68951-92-8*] has a favorable effect [315].

Acylated protein hydrolysates are used as dispersing and complexing agents in the oxidative treatment of textile fibers [281].

4.2.2. Alkylated Products

The reactive groups of protein hydrolysates can be alkylated (see Table 46). Only reaction products with especially reactive alkylating reagents such as α-epoxides or their precursors or with chloroacetic acid are of technical importance.

Quats. In the production of "quats", chlorohydrins (e.g., 3-chloro-2-hydroxypropyl-*N,N,N*-dimethylalkylammonium chloride) are used as alkylating agents that are converted to the corresponding α-epoxides by the addition of alkali. These are the real alkylating agents. In this way, a group bearing a quaternary nitrogen is transferred to the protein hydrolysate, resulting in the formation of a cationic derivative that is positively charged over the entire pH range. The reaction is base catalyzed and proceeds in an aqueous medium at $\leq 60\,°C$.

Carboxymethylated Products. The reaction of protein hydrolysates with salts of chloroacetic acid gives carboxymethylated protein hydrolysates. This reaction is base catalyzed and is carried out at ca. 80 °C.

Applications. In comparison with the hydrolysates, the cationic protein hydrolysates exhibit a considerably stronger substantivity to hair

Table 47. Suppliers of protein hydrolysates and protein derivatives (selection)

	Firm	Protein raw materials used		Producers of protein		Use ∗
		Animal	Plant	Hydrolysates	Derivatives	
Europe	Croda	+	+	+	+	C
	Dt. Gelatine Fabriken (DGF)	+		+	+	C, F
	Diamalt	+	+	+		C, F
	Grünau/Henkel	+	+	+	+	C, F, T
	Sicit	+		+		T
	Seppic	+	+		+	C
	Kelisema		+	+		C
United States	Hormel	+		+		C
	Maybrook	+		+	+	C
	Brooks	+	+	+	+	C
	Sheffield	+		+		F
	Vegetech		+	+		C
Asia	Seiwa Kasei	+	+	+		C

∗ C = cosmetics, F = food/feed, T = technology.

or skin and are also absorbed on the surface of metals, glass, etc. For this reason, they are being increasingly used in cosmetics [299, 316]. In addition the molecules obtain a lipophilic character through the R^2 residue (see Table 46), which can be varied within wide limits. The molecules thus exhibit a weak surfactant action. Carboxymethylated protein hydrolysates are used as foam-free stabilizers for peroxide in the oxidative treatment of textile fibers.

4.2.3. Esters

Esterification of the carboxyl groups in protein hydrolysates and derivatives is possible. However, technical applications for these substances do not exist at present.

4.3. Product Properties

Commercially available protein hydrolysates and derivatives are powders having a light-beige to brown color and a characteristic odor. According to producers, the molecular masses are in the range of 400 to 100 000. Protein hydrolysates are very soluble in water and slightly soluble in polar organic solvents. Hydrolysates obtained by enzymatic degradation are optically active, products obtained by acid or alkaline hydrolysis are largely racemic.

4.4. Analysis

Protein hydrolysates and derivatives are characterized by determining the ash content, pH, amount of dry substance, total nitrogen, molecular mass distribution, and total microorganism count. Other product properties can be determined by using standard analytical methods for proteins [317].

4.5. Storage and Transportation

The austenitic stainless steels 1.4571 and 1.4439 are employed for the storage and transportation of protein hydrolysates and derivatives. The steel 1.4541 is only partially usable for hydrolysates and condensates because it is less resistant to corrosion. The stability to corrosion of the material used should be checked in each individual case because of the different salt concentrations of the commercial hydrolysates and derivatives.

The nonmetallic materials suitable for hydrolysates are low-density polyethylene, polypropylene, and polyester materials. Polyester and low-density polyethylene materials are suitable for protein – fatty acid condensates.

4.6. Economic Aspects

Reliable details about the amounts of protein hydrolysates and derivatives produced cannot be

given because the required statistics are unavailable. However, an estimated 40 000 t/a is produced worldwide for the main areas of application:

Cosmetics
Food and feeds
Technology

Selected suppliers of protein hydrolysates and protein derivatives are listed in Table 47.

4.7. Toxicology and Ecology

Toxicology. The American College of Toxicology has published conclusive reports on safety assessment of collagen hydrolysates [300] and acylated collagen hydrolysates [301, 312]. According to these reports, the two product classes are virtually nontoxic, minimally irritating to the rabbit eye, and not sensitizing in the guinea pig. These results are supported by decades of experience gained by producers and users and, according to present findings, can also be applied to other protein hydrolysates and their condensates.

Ecology. Protein hydrolysates and derivatives are based on renewable natural resources. This explains their biodegradability, which ranges from good to very good, and their easy metabolism. According to German law, protein hydrolysates belong to the water hazard class (WGK) 0; the condensates, to class 1; and the quats, to class 2.

5. References

1. R. E. Olson (ed.): *Protein-Calorie Malnutrition,* Academic Press, New York 1975.
2. FAO, The State of Food and Agriculture 1991, Rome 1992.
3. FAO, Production Yearbook 1990, vol. 44, Rome 1992;FAO Trade Yearbook; FAO World Agricultural Statistics; FAO Yearbook Fishery Statistics.
4. P. Foster: *The World Food Problem, Adamantine Studies in Politics, Economy, and Development,* vol. 7, Adamantine Press, London 1992.
5. PAG Bulletin, Reports, Statements, Guidelines issued by PAG, Rome 1971 ff.
6. *FAO-Production Yearbook 1989,* Rome 1990, p. 113.
7. S. A. Watson, P. E. Ramstad: *Corn-Chemistry and Technology,* American Association of Cereal Chemists, St. Paul 1987, p. 450.
8. R. Lasztity: *The Chemistry of Cereal Proteins,* CRC Press, Boca Raton 1984, p. 4.
9. Y. Pomeranz: *Advances in Cereal Science and Technology,* **vol. 8,** American Association of Cereal Chemists, St. Paul 1986, p. 196.
10. Y. Pomeranz: *Wheat-Chemistry and Technology,* **vol. 2,** American Association of Cereal Chemists, St. Paul 1988, pp. 91 and 167.
11. R. Lasztity, M. Hidvegi: *Amino acid composition and biological value of cereal proteins,* Akadémiai Kiado, Budapest 1985, p. 82 "Triticale," p. 147 "Maize," p. 487 "Rice," pp. 373, 501.
12. K. L. Sehgal, S. Bajaj, K. Sekhon: "Studies on the Composition, Quality and Processing of Triticale," Part 1, "Physico-chemical Characteristics," *Nahrung* **27** (1983) 31 – 37.
13. Y. Pomeranz: "Proteins and Amino Acids of Barley, Oats and Buckwheat," in M. Friedman (ed.): *Protein Nutritional Quality of Foods and Feeds,* Marcel Dekker, New York 1975, p. 64.
14. G. Tegge: *Stärke und Stärkederivate,* 2nd ed. Behr's Verlag, Hamburg 1988, p. 311.
15. S. A. Watson, P. E. Ramstad: *Corn-Chemistry and Technology,* American Association of Cereal Chemists, St. Paul 1987, pp. 305, 370, 390, 450.
16. P. Cerletti, P. Restani: "Maize Germ Proteins, their Composition, Nutritive Value and Functional Properties," in R. Lasztity, M. Hidvegi (eds.): *Amino Acid Composition and Biological Value of Cereal Proteins,* Akadémiai Kiado, Budapest 1985, pp. 467 – 480.
17. Y. Pomeranz: *Wheat-Chemistry and Technology,* 3rd ed., **vol. 2,** American Association of Cereal Chemists, St. Paul 1988, p. 40.
18. Company brochures of Roquette (FR), Amylum (BE) and Nidefo (NL).
19. [8], p. 87.
20. [7], pp. 305, 385, 389, 450.
21. M.-L. Sarkki, E.-R. Saarinen: "Wheat Protein Products in Nutrition," in R. Lasztity, M. Hidvegi (eds.): *Amino Acid Composition and Biological Value of Cereal Proteins,* Akadémiai Kiado, Budapest 1985, pp. 543 – 550.

22. [6], p. 155.
23. T. Tashiro, Y. Fukuda, T. Osawa, M. Namiki: "Oil and Minor Components of Sesam Strains," *J. Am. Oil. Chem. Soc.* **67** (1990) 508 – 511.
24. J. H. Lope, J. L. Peset, P. A. Angulo, F. R. Berbel: "El altramuz (lupino): su composición y aprovechamiento como nueva fuente de proteinas y aceite," *Grasas y Aceites* (Sevilla) 36 (1985) 5 – 13.
25. T. Hatzold: "Chemische und chemisch-technische Untersuchungen zur Beurteilung von Lupinen (*L. mutabilis)* als Nahrungsmittel für den Menschen," Dissertation, Justus-Liebig-Universität Giessen 1982, pp. 134 and 165.
26. H. Schoeneberger et al., *Nahrung* **25** (1981) 667 –674.
27. V. D. Burrows et al. : *Food Protein from Grains and Oilseeds,* Office of the Minister Responsible for the Canadian Wheat Board, Ottawa 1972, pp. 41, 76.
28. B. Gassmann: "Preparation and Application of Vegetable Proteins, especially Proteins from Sunflower Seed, for Human Consumption. An Approach," *Nahrung* **27** (1983) 351 – 369.
29. H. Schmandke in H.-A. Ketz, F. Baum (eds.): *Ernährungslexikon,* Fachbuchverlag, Leipzig 1986.
30. A. Visser, A. Thomas: "Soya Protein Products –their Processing, Functionality, and Application Aspects," *Food Rev. Int.* **3** (1987) 1 – 32.
31. [29], pp. 486 – 488.
32. [29], pp. 483.
33. [22], p. 145.
34. H. Schmandke: *Die Ackerbohne (Vicia faba)* als Lebensmittelrohstoff, Akademie Verlag, Berlin 1988.
35. K. H. Mc Watters: "Use of Peanut and Cowpea Flours in Selected Fried and Baked Foods," in R. L. Ory (ed.): *Plant Proteins-applications, Biological Effects and Chemistry, ACS Symp. Ser.* **312** (1985) 11, 13.
36. B. P. Klein, K. A. Raidl: "Use of Field-Pea Flours as Protein Supplements in Foods," in R. L. Ory (ed.): *Plant Proteins-Applications, Biological Effects and chemistry, ACS Symp. Ser.* **312** (1986), 20, 22.
37. E. W. Lusas, K. C. Rhee: "Applications of Vegetable Food Proteins in Traditional Foods," in R. L. Ory (ed.): *Plant Proteins-Applications, Biological Effects and Chemistry, ACS Symp. Ser.* **312** (1985) 37.
38. G. Sarwar, F. W. Sosulski, N. W. Holt: "Protein Nutritive Value of Legume-Cereal Blends," *Can. Inst. Food. Sci. Technol. J.* **8** (1975) 170 – 174.
39. H. K. Burr: "Pulse proteins," in M. Friedman (ed.): *Protein Nutritional Quality of Foods and Feeds,* Marcel Dekker, New York 1975, p. 125.
40. P. Colonna, D. Gallant, C. Mercier: "Pisum Sativum and Vicia Faba Carbohydrates-Studies of Fractions obtained after dry and wet Protein Extraction," *J. Food. Sci.* **45** (1980) 1629.
41. K. Elkowicz, F. W. Sosulski: "Antinutritive Factors in Eleven Legumes and their Air Classified Protein and Starch Fractions," *J. Food Sci.* **47** (1982) 1301 – 1304.
42. T. Tyler, C. G. Youngs, F. W. Sosulski: "Air Classification of Legumes: I. Separation Efficiency, Yield and Composition of the Starch and Protein Fractions," *Cereal Chem.* **58** (1981) 144 – 148.
43. Ch. Schneider, M. Schultz, H. Schmandke: "Preparation of Broad Bean (Vicia Faba L. minor) Products: Part I. Broad Bean Protein Isolates from Seed Flour and their Acetylation in the Neutral pH Range," *Nahrung* **29** (1985) 785 – 791.
44. J. Gueguen, B. Quenemener, P. Valdebouze: "Elemination des facteurs antinutritionnels de la fèverole Vicia faba L. et du pois Pisum sativum L. au cours de la préparation de isolats protéiques," *Lebensm. Wiss. Technol.* **14** (1980) 72 – 79.
45. H. S. Olsen: "Continuous Pilot Plant Production of Bean Protein by Extraction, Centifugation, Ultrafiltration and Spray-drying," *Lebensm. Wiss. Technol.* **11** (1978) 57 – 64.
46. [34], p. 48.
47. F. W. Sosulski, K. Sosulski: "Composition and Functionality of Protein, Starch and Fiber from wet and dry Processing of Grain Legumes," in R. L. Ory (ed.): *Plant Proteins-Applications, Biological Effects and Chemistry, ACS Symp. Ser.* **312** (1986), 176 – 189.
48. Y. J. Owusu, S. M. Mc Curdy: "Pea Proteins: A Review of Chemistry, Technology of Production, and Utilization," *Food Rev. Intern.* **7** (1991) 103 –134.
49. [22], p. 134.
50. J. A. Woolfe: *The Potato in the Human Diet,* Cambridge University Press, Cambridge 1987, p. 24.
51. [50], p. 151.
52. W. M. Walter, A. E. Purcell: "Protein of the Sweet Potato," in R. L. Ory (ed.): *Plant*

Proteins: Applications, Biological Effects and Chemistry, ACS Symp. Ser. **312** (1986) 244.
53. L. R. Hackler: "Cereal Proteins in Human Nutrition," in R. Lasztity, M. Hidvegi (eds.): *Amino Acid Composition and Biological Value of Cereal Proteins,* Akadémiai Kiado, Budapest 1985, p. 93.
54. L. D. Satterlee: "Proteins for Use in Foods," *Food Technology* **35,** June 1981, p. 53 – 70.
55. P. R. Payne: "Human Protein Requirements," in G. Norton (ed.): *Plant Properties,* Butterworth, London 1978, p. 259.
56. E. Wilhelm, W. Kempf: "Neue Erkenntnisse über die Gewinnung von Kartoffelproteinen für die menschliche Ernährung," *Stärke* **29** (1977) 376 –380.
57. E. Wilhelm, W. Kempf: "Fortschritte bei der Separation und Stabilisierung von Kartoffelproteinen," *Starch Stärke* **33** (1981) 338 – 342.
58. D. Knorr: "Protein Recovery from Waste Effluents of Potato Processing Plants," *J. Food Technol.* **12** (1977) 563 – 580.
59. D. Knorr: "Protein Quality of the Potato and Potato Protein Concentrates," *Lebensm. Wiss. Technol.* **11** (1978) 109 – 115.
60. D. Knorr: "Fortification of Bread with Potato Products," *Starch Stärke* **31** (1979) 242 – 246.
61. D. Knorr: "Effect of Recovery Methods on Yield, Quality and Functional Properties of Potato Protein Concentrates," *J. Food Sci.* **45** (1980) 1183 –1186.
62. [50], pp. 162 – 186.
63. G. Lisinska, W. Leszczynski: *Potato Sciences and Technology,* Elsevier Science Publishers, London 1989.
64. R. E. Berry: "Tropical Seeds, Legumes, Fruits and Leaves as Sources of Protein," in R. L. Ory (ed.): *Plant Proteins: Applications, Biological Effects and Chemistry, ACS Symp. Ser.* **312** (1986) 220 – 233.
65. T. C. Tso, S. D. Kung: "Soluble Proteins in Tobacco and their Potential Use," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, pp. 117 – 132.
66. O. L. Oke: "Leaf Protein Research in Nigeria," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, p. 742.
67. H. T. Ostrowski-Meissner: "Protein Extraction from Grassland," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, p. 33.
68. T. Horigome: "Leaf Protein Research and Development in Japan, in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, pp. 731 – 738.
69. H. A. El-Alaily, H. S. Soliman: "Experimental Studies on Leaf Protein in Egypt," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, p. 670.
70. A. A. Woodham: "The Nutritional Evaluation of Leaf Protein Concentrates," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, p. 423.
71. E. C. Vieira: "Leaf Protein Research in Brazil," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, p. 666.
72. L. D. Satterlee: "Proteins for use in foods," *Food Technol.* **35** (1981, June) 54.
73. Yu. A. Antonov, V. B. Tolstoguzov: "Food Protein from Green Plant Leaves," *Nahrung* **34** (1990) 125 – 134.
74. N. W. Piric: *Leaf Protein and its By-products in Human and Animal Nutrition,* Cambridge University Press, Cambridge 1987.
75. G. O. Kohler, R. Edwards, D. de Fremery: "LPC for Feeds and Foods: The Pro-Xan Process," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, pp. 508 –524.
76. L. Koch: "The Vepex Process," in L. Telek, H. D. Graham (eds.): *Leaf Protein Concentrates,* AVI Publishing, Westport 1983, pp. 601 – 632.
77. L. Lasztity, R. Lasztity, A. Salgo: "Protein-Phytic Acid-Metallic Ion Interactions," in K. D. Schwenke, B. Raab (eds.): *Interactions in Protein Systems,* Akademie Verlag, Berlin 1989, pp. 243 – 245.
78. I. E. Liener, M. L. Kakade: "Protease inhibitors," in I. E. Liener (ed.): *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York 1969, pp. 8 – 53.
79. I. E. Liener: "Miscellaneous Toxic Factors," in I. E. Liener (ed.): *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York 1969, pp. 410 –437.
80. R. D. Montgomery: "Cyanogens," in I. E. Liener (ed.): *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York 1969, pp. 143 – 155.
81. [8], pp. 76 and 148.

82. R. Machholz, H.-J. Lewerenz: *Lebensmitteltoxikologie,* Akademie Verlag, Berlin 1989, p. 255.
83. H.-A. Ketz, F. Baum: *Ernährungslexikon,* Fachbuchverlag, Leipzig 1986, pp. 55 – 58, pp. 449 – 451, p. 464, pp. 486 – 489, pp. 490 – 493.
84. J. Kroll: "Die Ultrafiltration als Trennungs- und Reinigungsverfahren zur Gewinnung und Behandlung von Lebensmittelroh- und zusatzstoffen," *Int. Z. Lebensm.-Techn. Marketing, Verpackung, Analytik* **41** (1990) 720 – 730.
85. K. D. Schwenkel et al.: "Preparation of Detoxified High Functional Rapeseed Flours," *J. Sci. Food Agric.* **51** (1990) 391 – 405.
86. M. Sternberg, C. Y. Kun, F. J. Schwende: "Lysino-alanine-Presence of Foods and Food Ingredients," *Science* (Washington D.C.) 190 (1975) 992 – 994.
87. J. H. Lope, J. L. Peset, P. A. Angulo: "El altramuz (lupino): Su composición y aprovechamiento como fuente de proteinas y de aceites. III. Obtención de aislados proteinicos y de aceite con bajos contenidos en alcaloides," *Grasas y Aceites (Sevilla)* **36** (1985) 280 – 287.
88. [34], pp. 11 – 47.
89. International Institute of Tropical Agriculture, "Unpeeled Cassave Tubers for 'Gari'," in *Annual report and research highlights,* Ibadan, Nigeria 1984, pp. 80 – 81.
90. "Jamswurzel," in *Lebensmittellexikon,* Fachbuchverlag, Leipzig 1979, p. 405.
91. Y. Birk: "Saponins," in I. E. Liener (ed.): *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York 1969, pp. 169 – 203.
92. Codex Alimentarius Commission, ALINORM 89/16, pp. 37 – 38.
93. Codex Alimentarius Commission, ALINORM 87/30, pp. 44 – 54.
94. Codex Alimentarius Commission, CX/VP 89/3, pp. 12 – 34.
95. P. L. Pellett, V. R. Young: Commentary: Joint FAO/WHO Expert Consultation on Protein Quality Evaluation Bethesda, MD, USA, 4 – 8 Dec. 1989, *Ecol. Food Nutr.* **24** (1990) 297 – 303.
96. T. Levring, H. A. Hoppe, O. J. Schmid, *Marine Algae,* Cram, de Gruyter & Co., Hamburg 1969, pp. 288 – 368.
97. C. J. Soeder: in [114], pp. 25 – 41.
98. A. Richmond: in [105] , pp. 85 – 121.
99. A. Borowitzka, L. J. Borowitzka, *Micro-algal Biotechnology,* Cambridge University Press, Cambridge 1988.
100. E. P. Lincoln, D. T. Hill: in [113], pp. 229 – 244.
101. E. W. Becker: In [105] , pp. 222 – 225.
102. C. J. Soeder, M. Subardja, *DECHEMA-Monogr.* **70** (1972) 55 – 70.
103. V. J. Chapman: *Seaweeds and their Uses,* 2nd ed., Methuen & Co., London 1970, pp. 69, 111.
104. H. Durand-Chastel: In [113], pp. 51 – 64.
105. C. J. Soeder, *Fette Seifen Anstrichm.* **81** (1979) 352 –357.
106. M. A. Borowitzka: in [105], pp. 153 – 196.
107. G. Shelef, C. J. Soeder (eds.): *Algae Biomass Production and Use,* Elsevier/North-Holland, Biomedical Press, Amsterdam 1980.
108. A. Richmond (ed.): *Handbook of Microalgal Mass Culture,* CRC Press, Boca Raton 1986.
109. P. S. Heussler, J. M. Castillo, F. Merino, V. Vasquez, *Ergeb. Limnol.* **13** (1980) 796; **11** (1978) 254 – 257.
110. F. H. Mohn: in [105], pp. 396 – 414.
111. W. J. Oswald: in [105], pp. 305 – 328.
112. S. Yannai, S. Mokady, K. Sachs, Z. Berk, *Ergeb. Limnol.* **11** (1978) 139 – 149.
113. K. Kawaguchi: in [2.107], pp. 25 – 34.
114. N. Cherekezov et al., *Zhivodnovud. Nauki* **11** (1974) 17 – 21; C. Beremski et al., *Zhivodnovud. Nauki* **10** (1973) 35 – 40, 99 – 105; K. H. Ganovski et al., *Zhivodnovud. Nauki* **12** (1975) 77 – 83.
115. S. R. Tannenbaum, B. R. Stillings, N. S. Scrimshaw (eds.): *The Economics, Marketing and Technology of Fish Protein Concentrate,* MIT Press, Cambridge, Mass., 1974.
116. R. Finch, *CRC Crit. Rev. Food Technol.* **1** (1970) 519 – 580.
117. R. Kreuzer (ed.): *Fishery Products,* Fishing News (Books), Survey 1974.
118. E. R. Pariser, M. B. Wallerstein, C. J. Corkery, N. Brown: *Fish Protein Concentrate,* MIT Press, Cambridge, Mass., 1978.
119. Z. E. Sikorski, *CRC Crit. Rev. Food Sci. Nutr.* **14** (1981) 201 – 230.
120. FAO: *Yearbook of Fishery Statistics,* vol. 68, FAO, Rome 1989.
121. J. Murray, J. R. Burt: "The Composition of Fish," *Torry Advisory Note,* no. 38, HMSO Press, Edinburgh 1975.
122. *Protein Advisory Group Bull.* **1** (1971) no. 12, 13 –15.
123. F. S. Shenstone, *Food Preserv. Q.* **13** (1952) 45 – 50.
124. E. Yañez et al., *Food Technol. (Chicago)* **21** (1967) 1604 – 1610.

125. V. D. Sidwell, *NOAA Techn. Mem. NMFS F/SEC* 1981, no. 11, 430.
126. C. M. Lee, T.-C. Lee, C. O. Chichester: "The Potential of Squid as a Protein Resource," in [117], pp. 243 – 244.
127. A. R. Pandit, N. G. Magar, *Fish. Technol.* **9** (1972) 122 – 125.
128. R. Shirasaka, J. A. de Shiosaka, *An. Cient.* **11** (1973) 127 – 141.
129. R. V. Krishnamoothy et al., *J. Agric. Food Chem.* **27** (1979) 1125 – 1127.
130. N. Roschke, W. Schreiber, *Arch. Fischereiwiss.* **28** (1977) 135 – 141.
131. J. Arnold, L. M. Sprague: "FPC: Present and Future Status of Available Oceanic Resources," in [115], pp. 8 – 32.
132. G. N. Zein, A. E. I.-F. El-Bedawey, A. M. El-Sherbiney, F. M. A. Dawoud, *Nahrung* **29** (1985) 523 –532.
133. A. E. I.-F. El-Bedawey, G. N. Zein, A. M. El-Sherbiney, F. M. A. Dawoud, *Nahrung* **30** (1986) 19 –24.
134. K. K. Almenase et al.: *NOAA Techn. Rep. NMFS Circ.* 1972, no. 367.
135. R. Finch: "FPC Processes," in [115], pp. 57 – 101.
136. D. Dubrow, O. Hammerle, *Food Technol. (Chicago)* **23** (1969) 254 – 256.
137. B. R. Stillings: "Nutritional and Safety of FPC," in [115], pp. 165 – 211.
138. H. E. Power, *J. Fish. Res. Board Can.* **19** (1962) 1039 – 1045.
139. E. Levin, R. K. Flinn, *Chem. Eng. Prog.* **51** (1955) 223 – 226.
140. G. M. Dreosti: "Technological Developments in South Africa," in [117], pp. 425 – 431.
141. E. Yañez et al., *J. Food Sci.* **41** (1976) 1289 – 1292. G. M. Pigott, G. O. Bucove, J. G. Ostrander, *J. Food Process. Preserv.* **2** (1978) 33 – 54.
142. Marine Protein Concentrate, Fishery Leaflet 584, U.S. Dept. of Commerce, Bureau of Commercial Fisheries, College Park, Md., 1966.
143. J. J. Connell, *Food Technol. (Chicago)* **23** (1969) 206 – 207.
144. C. Greenfield, US 2 651 647, 1953; *Chem. Abstr.* **48** (1954) 390.
145. J. Spinelli, B. Koury, H. Groninger, R. Miller, *Food Technol. (Chicago)* **31** (1977) 184 – 187.
146. A. Guttmann, F. A. Vandenheuvel: "Progress Reports Atlantic Coast Station," *Fish Res. Board Can.* **67** (1957) 29 – 31.
147. Ch.-L. Chu, G. M. Pigott, *Trans. ASAE* 1973, 949 – 952.
148. C. M. Libenson, I. Pirosky, GB 1 108 188, 1968; *Chem. Abstr.* **69** (1968) 42 906 e.
149. J. Spinelli: "Aqueous-solvent Processes," in [115], pp. 111 – 121.
150. E. von Jan, *Fette Seifen Anstrichm.* **79** (1977) 414 –432.
151. J. Rasekh, B. R. Stillings, V. Sidwell, *J. Food Sci.* **37** (1972) 423 – 425.
152. L. B. Sjöberg, S. L. Boström, *Br. J. Nutr.* **38** (1977) 189 – 205.
153. D. L. Dubrow, B. R. Stillings, *J. Food Sci.* **35** (1970) 677 – 680.
154. O. Myklestad: "Fish Protein Processing, Efficiency of Drying Equipment," in [117], pp. 111 – 112.
155. D. L. Dubrow, *Fish. Bull.* **71** (1973) 99 – 104.
156. Astra Protein Products, SW 381 977, 1976.
157. B. F. Cobb, K. Hyder, *J. Food Sci.* **37** (1972) 743 –750.
158. J. Oehlenschläger, W. Schreiber, *Z. Lebensm. Unters. Forsch.* **172** (1981) 393 – 398.
159. H. Yuguchi, *Jpn. J. Dairy Food Sci.* **30** (1981) 95 –104.
160. S. Y. Yu, L. K. Tan, *Int. J. Food Sci. Technol.* **25** (1990) 204 – 208.
161. H. S. Groninger, *J. Food Sci.* **41** (1976) 268 – 273.
162. H. S. Groninger, R. Miller, *J. Food Sci.* **40** (1975) 327 – 330.
163. J. Spinelli, H. Groninger, B. Koury, R. Miller, *Process Biochem.* **42** (1975) 31 – 36.
164. M. Fujimaki et al., *Agric. Biol. Chem.* **44** (1980) 1911 – 1916.
165. J. Kroll, B. Gassmann, S. Cifuentes, W. Koenig, *Lebensmittelindustrie* **30** (1983) 113 – 117.
166. H. Niki, E. Deya, T. Kato, S. Igarashi, *Bull. Jpn. Soc. Sci. Fish.* **48** (1982) 999 – 1004.
167. W. M. Perez, S. Hayasni, Z. Oosniro, H. Omura, *Alimentos* **11** (1986) no. 2, 21 – 26.
168. E. H. Lee et al., *Korean J. Food Sci. Technol.* **14** (1982) 43 – 48.
169. S. T. Chari, A. Sreenivasan, *Fish. Technol.* **17** (1980) 115 – 117.
170. E. H. Lee et al., *Bull. Korean Fish. Soc.* **17** (1984) 299 – 305.
171. R. G. Ackman: "Fish Lipids," in J. J. Connell (ed.): *Advances in Fish Science and Technology,* Part 1, Fishing News Books, Surrey, pp. 86 – 102.
172. S. Valand, *Food Nutr. (FAO)* **5** (1979) no. 2, 24 – 30.
173. F. M. Orejana, J. Espejo-Hermes, C. M. Bigueras, J. B. Gamboa III, *FAO Fish. Rep.* 1985, no. 317 (Suppl.) 418 – 427.

174. E. Okazaki, K. Kanna, T. Suzuki, *Bull. Jpn. Soc. Sci. Fish.* **46** (1980) 727 – 732.
175. E. H. Lee, Sudibjono, S. K. Kim, *Korean J. Food Sci. Technol.* **11** (1979) 232 – 241.
176. *Fish. News Int.* **16** (1977) no. 12, 33 – 34.
177. *Int. FPC News* **1** (1979) no. 1, 2.
178. D. L. Dubrow, N. L. Brown, E. R. Pariser, H. Miller, *Commer. Fish. Rev.* **32** (1970) no. 12, 25 – 31.
179. B. Phitakpol, P. Vittanoonta, V. Nathalang, *FAO Fish. Rep.* 1985, no. 317 (Suppl.) 331 – 337.
180. R. F. Di Cesare, *Lebensmittel-Wiss. Technol.* **18** (1985) 374 – 378.
181. F. H. Hoskins, J. Loustaunau, *Food Technol. (Chicago)* **28** (1974) no. 3, 58 – 62.
182. B. Koury, J. Spinelli, *J. Food Sci.* **40** (1975) 58 – 61.
183. *Int. FPC News* **16** (1986) 3 – 6.
184. A. Arafah, M. Abassy, S. Morcos, L. Hussein, *Cereal Chem.* **57** (1980) 35 – 39.
185. J. H. Hulse: "Fish Proteins," in A. M. Altschul (ed.): *New Protein Foods,* vol. 1 A, Academic Press, New York 1974, pp. 191 – 194.
186. K. Vareltzis, N. Soultos, F. Zetou, I. Tsiaras, *Lebensm. Wiss. Technol.* **23** (1990) 112 – 116.
187. C. Abrahamsson, L. Hambraeus, B. Vahlquist, *Protein Advisory Group Bull.* **4** (1974) 26 – 30.
188. S. R. Tannenbaum, R. P. Bates, L. Brodfeld, *Food Technol. (Chicago)* **24** (1970) 607 – 609.
189. V. D. Sidwell, B. R. Stillings, G. M. Knobl Jr., *Food Technol. (Chicago)* **24** (1970) 876 – 882.
190. W. Krone, FAO, personal communication, 1979.
191. E. M. Nikkila, S. M. Constantinides, T. L. Meade, *J. Agric. Food Chem.* **24** (1976) 1144 – 1147.
192. K. C. S. Sikka et al., *J. Agric. Food Chem.* **27** (1979) 946 – 949.
193. T. Suzuki et al., *Bull. Jpn. Soc. Sci. Fish.* **48** (1982) 105 – 111.
194. N. S. Sudhakara, T. M. R. Setty, A. S. Nagaraj, K. S. Udupa, *Indian J. Nutr. Diet.* **25** (1988) 44 – 49.
195. C. Parsons, A. T. T. Pham, N. F. Haard, *Can. Inst. Food Sci. Technol. J.* **17** (1984) 274 – 277.
196. B. D. Rebeca, M. T. Peña-Vera, M. Diaz-Castañeda, *J. Food Sci.* **56** (1991) 309 – 314.
197. T. Suzuki: *Fish and Krill Protein: Processing Technology,* Appl. Sci. Publ., London, 1981, p. 260.
198. B. Hale: "Making Fish Protein Concentrate by Enzymatic Hydrolysis," *NOAA Techn. Rep., NMFS SSRF* 1972, p. 657.
199. M. Mackie, *Process Biochem.* **9** (1974) no. 10, 12 – 14.
200. M. C. Archer, J. O. Ragnarsson, S. R. Tannenbaum, D. I. C. Wang, *Biotechnol. Bioeng.* **15** (1973) 181 – 196.
201. S. Iseki, T. Watanabe, T. Kinumaki, *Tokai-ku Suisan Kenkyusho Kenkyo Hokoku* **59** (1968) 81 – 89.
202. J. P. H. Wessels, A. Atkinson, *FIRI Ann. Rep.* **27** (1973) 15 – 18.
203. A. Atkinson, *FIRI Ann. Rep.* **28** (1974) 50.
204. L. R. Beuchat, *J. Milk Food Technol.* **37** (1974) 277 – 281.
205. G. Lalasidis, S. Boström, L.-N. Sjöberg, *J. Agric. Food Chem.* **26** (1978) 751 – 756.
206. J. J. Connell: "Fish Muscle Protein and Some Effects of Processing on them," in H. Schultz, A. Anglemeir (eds.): *Proteins and their Reactions,* Avi Publ., Westport, Conn., 1964, pp. 255 – 293.
207. S. Scholtyssek: *Geflügel,* Ulmer, Stuttgart 1987.
208. *Jahrbuch für die Geflügelwirtschaft,* Ulmer, Stuttgart 1992.
209. Verordnung (EWG) Nr. 2772/75 des Rates v. 29. 10. 1975.
210. E. Kallweit, G. Kielwein, R. Fries, S. Scholtyssek: *Qualität tierischer Nahrungsmittel,* UTB 1484, Ulmer, Stuttgart 1988.
211. *Kirk Othmer,* 3rd ed., **8.**
212. S. Scholtyssek, *Eier Wild Geflügelmarkt* **17** (1988); **16** (1990).
213. R. Knorr: "Qualitätsbeurteilung von Eiprodukten," Dissertation, Universität Hohenheim 1991.
214. J. Sahrai-Sani: "Elektrophoretische Untersuchungen mit Eiklar von Hühnern und Wachteln," Dissertation, Universität Hohenheim 1975.
215. S. Scholtyssek, T. Trziszka, *Arch. Geflügelkd.* **49** (1985) 228 – 236.
216. T. Trziszka, S. Scholtyssek, *Zentr. Bl. f. Lebensmittelt.* **38** (1987) 491 – 493.
217. W. Ternes: *Naturwissenschaftliche Grundlagen der Lebensmittelzubereitung,* Behrs Verl., Hamburg 1990.
218. F. E. Cunningham: "Egg Product Pasteurization," *Egg Science* 2. ed., Stadmann and Coterill Avi Publishing Co, Inc., Westport, Conn. 1997, pp. 161 – 188.
219. J. Wismer-Pedersen, *Meat Sci.* **24** (1988) 31 – 45.

220. A. Stiebing: "Blutplasma-Gewinnung, Zusammensetzung, Lagerfähigkeit und technologische Wirkung bei Brühwurst," Dissertation, Technische Universität, Berlin 1985.
221. J. Wismer-Pedersen, K. Nonnemann, *Fleischwirtschaft* **60** (1980) 231 – 235.
222. H. W. Ockerman, C. L. Hansen: *Animal By-Product Processing* VCH Verlagsgesellschaft, Weinheim 1988.
223. J. Wismer-Pedersen, *Fleischwirtschaft* **61** (1981) no. 9, 1272 – 1273.
224. J. Wismer-Pedersen: "Utilization of Blood in Comminuted Meat Products," Proc. Europ. Meeting Meat Res. Workers no. 24, H 4 : 1 – H 4 : 10 Kulmbach 1978.
225. A. Rossi-Fanelli, E. Antonini, A. Caputo, *Biochim. Biophys. Acta* **30** (1958) 608 – 615.
226. F. W. J. Teale, *Biochim. Biophys. Acta* **35** (1959) 543.
227. Unilever, EP 0 013 055, 1979 (J. H. Luijerink).
228. P. T. Tybor, C. W. Dill, W. A. Landmann, *J. Food Sci.* **40** (1975) 155 – 159.
229. Y. Z. Lee, R. M. Wang, S. Nakai, *J. Food Sci.* **55** (1990) no. 2, 577 – 578.
230. Y. Sato, S. Hayakawa, M. Hayakawa, *J. Food Technol.* **16** (1981) no. 1, 81 – 91.
231. K. Autio: "A new Method for Globin and Heme Preparation from Blood Corpuscle Concentrate," Proc. Europ. Meeting Meat Res. Workers no. 29, vol. 2, G 1 Salsomaggiore (1983) 765 – 770.
232. H. S. Olsen, *Intern. Z. Lebensmittel-Technologie und Verfahrenstechnik* **34** (1983) no. 5, 406 – 412.
233. V. Hald-Christensen: "Decolourisation of Slaughterblood by Partial Enzymatic Hydrolysis," Proc. Europ. Meeting Meat Res. Workers no. 24, Kulmbach 1978, H : 5.
234. G. Drepper: "Verfahren zur Herstellung von Globinhydrolysaten aus Blutkörperchenkonzentrat für die menschliche Ernährung und die Isolierung und Identifizierung einiger Peptide," Dissertation, Christian-Albrechts-Universität, Kiel 1982.
235. J. Wismer-Pedersen, "Decoloration of Blood by Heme Oxidation," Proc. Europ. Meeting Meat Res. Workers, no. 33, vol. I, 3 : 6, Helsinki 1987, 119 – 123.
236. G. Pruss, DE-OS 3 243 205, 1984.
237. M. Kobald, "Verfahrensentwurf für die Herstellung einer entfärbten Globinfraktion aus Dickblut," Dissertation, Technische Universität, München 1993.
238. Westfalia Separator Company: Separatoren, Dekanter und Anlagen zur Verarbeitung von Schlachttierblut, Oelde 1992.
239. Atmos-Alfa Laval Group 1989, personal communication.
240. C. Lynn Knipe in A. M. Pearson, T. R. Dutson (eds.): *Edible Meat By-Products. Advances in Meat Research,* vol. 5, Elsevier Applied Science, London 1988.
241. M. D. Ranken in R. A. Grant (ed.): *Applied Protein Chemistry,* Applied Science Publishers LTD, London 1980.
242. A. M. Hermansson, E. Tornberg, "Functional Properties of some Protein Preparations from Blood," Proc. Europ. Meeting Meat Res. Workers no. 22 (1976), 11 : 1.
243. S. Hayakawa, T. Ogawa, Y. Sato, *J. Food Sci.* **47** (1982) no. 5, 1415 – 1418.
244. K. Autio, H. Lyytikäinen, Y. Mälkki, S. Kanko, *J. Food Sci.* **50** (1985) 615 – 617.
245. K. Autio, Personal communication, Technical Research Centre of Finland 1991.
246. T. Lüssman, Personal communication, Kraeber Company Hamburg 1992.
247. W. Jäppelt, H. Neumann in Abwassertechnische Vereinigung (eds.): *Lehr- und Handbuch der Abwassertechnik,* 3rd ed., **vol. 5,** Verlag Ernst & Sohn, Berlin 1985, p. 335.
248. VDI, 2596 Entwurf Emissionsminderung Schlachthöfe 1988.
249. P. Filstrup in R. A. Grant (ed.): *Applied Protein Chemistry,* Applied Science Publishers LTD, London 1980.
250. C. Große Frie: "Absatz- und Verwertungsmöglichkeiten für Schlachtnebenprodukte und Schlachtabfälle in der Bundesrepublik Deutschland," Dissertation, Friedrich-Wilhelms-Universität, Bonn 1984.
251. Institut für landwirtschaftliche Marktforschung der Bundesforschungsanstalt für Landwirtschaft, *Agrarwirtschaft* **40** (1991) 423 – 430.
252. Bundesministerium für Ernährung, Landwirtschaft und Forsten (eds.): *Statistisches Jahrbuch über Ernährung, Landwirtschaft und Forsten der BRD, 1991,* Landwirtschaftsverlag, Münster-Hiltrup 1991.
253. Atmos-Alfa Laval Company 1991 personal communication.
254. G. S. D. Weir in B. J. F. Hudson (ed.): *Developments in Food Proteins,* **vol. 4,** Elsevier Applied Science Publishers, New York 1986, pp. 175 – 217.

255. T. Godfrey in T. Godfrey, J. Reichelt (eds.): *Industrial Enzymology,* The Application of Enzymes in Industry, MacMillan, London 1983, pp. 221 – 259.
256. J. Adler-Nissen: *Enzymatic Hydrolysis of Food Proteins,* Elsevier Applied Science Publishers, New York 1986.
257. J. Adler-Nissen, S. Eriksen, H. S. Olsen, *Qual. Plant. Plant Foods Hum. Nutr.* **32** (1983) 411 – 423.
258. B. G. Venter, A. E. J. McGill, S. H. Lombard, *South Afr. J. Dairy Sci.* **21** (1989) no. 4, 75 – 82.
259. S. Eriksen in T. H. Applewhite (ed): *Vegetable Protein Utilization in Human Foods and Animal Feedstuffs,* Am. Oil Chemists' Society, Champaign 1989, pp. 78 – 83.
260. A. Kilara, *Process Biochem.* **20** (1985) 149 – 157.
261. H.-D. Belitz, H. Wieser, *Food Rev. Int.* **1** (1985) no. 2, 271 – 354.
262. H. Sugimoto, J. P. van Buren, W. B. Robinson, *J. Food Sci.* **36** (1971) 729 – 731.
263. W. Druml in R. Wenger, B. Brandstetter (eds.): *Eiweiß in Nahrung und Ernährung des Menschen,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1988, pp. 165 – 171.
264. W. Pölert, in [263], pp. 119 – 123.
265. I. M. Mackie, *Anim. Feed Sci. Technol.* **7** (1982) 113 – 124.
266. Röhm GmbH, GB 2 011 910, 1979 (Z. Eckmayer, A. Berg, R. Monsheimer, E. Pfleiderer).
267. R. J. Bouthilet, A. Karler, V. Johnsen, *Proc. Sci. Sect. Toilet Goods Assoc.* **45** (1966) 27 – 29.
268. A. Domsch, H. Pospischil, G. Schuster, H. Tronnier, *Parfüm Kosmet.* **61** (1980) 325 – 330.
269. R. J. Bouthilet, A. Karler, *Proc. Sci. Sect. Toilet Goods Assoc.* **44** (1965) 27 – 28.
270. R. Riso, *Am. Perfum. Cosmet.* **85** (1970) no. 9, 129 – 132.
271. R. Riso, *Soap Chem. Spec.* **39** (1963) no. 82, 151.
272. R. Riso, *Proc. Sci. Sect. Toilet Goods Assoc.* **42** (1964) 36 – 39.
273. E. S. Cooperman, *Am. Cosmet. Perfum.* **87** (1972) no. 3, 65 – 67.
274. Colgate-Palmolive Co., US 3 548 056, 1970 (E. Eigen, S. Weiss).
275. G. Schuster, A. Domsch, *Seifen Öle Fette Wachse* **108** (1982) 177 – 184, 248 – 250.
276. A. Domsch, G. Schuster, *Ärztl. Kosmetologie* **13** (1983) 524 – 530.
277. Ullmann, 4th ed., **11,** 572.
278. Chemische Fabrik Grünau, "Tricosal Normal Pulver," no. F 588/4, Illertissen 1988.
279. R. A. Kuntze, *Nature (London)* **211** (1966) 406 –407.
280. Chemische Fabrik Grünau, DT 669 554, 1938; DT 711 652, 1941 (G. Wiegand).
281. H. G. Fröhlich, M. Hackauf, C. Chwala in A. Chwala, V. Anger (eds.): *Handbuch der Textilhilfsmittel,* 2nd ed., Verlag Chemie, Weinheim 1977, pp. 243 – 245.
282. BASF, DE 3 903 362, 1989 (R. Brueckmann, J. P. Dix, M. Herrmann).
283. H. Bjorklund in L. Placzek (ed.): *Produkte der Papierchemie,* Keppler Verlag, Heusenstamm 1973, pp. 68 – 74.
284. K. H. Kludas et al., *Arch. Exp. Veterinärmed.* **44** (1990) 289 – 294.
285. M. B. Lek, M. Charles, *J. Ind. Microbiol.* **4** (1989) 71 – 75, 77 – 80.
286. Rhone-Poulenc, EP 266 258, 1988 (A. Jarry, H. Lamonerie).
287. O. J. Muscio, B. Cole, T. Mc Carty, V. Sandu, *J. Am. Oil Chem. Soc.* **59** (1982) 217 – 221.
288. Akademie der Wissenschaften der DDR, DD 273 264 A 1, 1989 (K. Haage et al.).
289. Chemische Fabrik Grünau, DT 691 411, 1935.
290. G. Schuster, H. Modde, E. Scheld, *Seifen Öle Fette Wachse* **91** (1965) 477 – 482.
291. G. Schuster, H. Modde, *Parfüm. Kosmet.* **45** (1964) 337 – 340.
292. G. Schuster, H. Modde, *Parfüm. Kosmet.* **46** (1965) 1 – 3.
293. Hoechst, Deutsche Gelatine Fabriken, DE 3 929 740 A 1, 1991 (A. Turowski et al.).
294. H. Modde, G. Schuster, H. Tronnier, *Tenside* **2** (1965) 368 – 373.
295. H. Tronnier, W. Schneider, G. Schuster, H. Modde, *Arch. Klin. Exp. Dermatol.* **229** (1967) 40 – 53.
296. H. Tronnier, G. Schuster, H. Modde, *Arch. Klin. Exp. Dermatol.* **221** (1965) 232 – 249.
297. G. Proserpio, S. Gatti, *Riv. Ital. Essenze Profumi Piante Off. Aromat. Syndets Saponi Cosmet. Aerosols* **61** (1979) no. 2, 53 – 62.
298. P. Busch, H. Hensen, D. Fischer, A. Rulinke, J. Franklin *Seifen Öle Fette Wachse* **120** (1994) no. 6, 339 – 344.
299. A. Domsch, P. Busch, H. Hensen, *Seifen Öle Fette Wachse* **114** (1988) no. 17, 695 – 702.
300. *J. Am. Coll. Toxicol.* **4** (1985) no. 5, 199 – 221.
301. *J. Am. Coll. Toxicol.* **2** (1983) no. 7, 75 – 86.

302. F. Aurich, R. Brehme, H. P. Welzel, K. Haage, *Tenside Surfactants Deterg.* **29** (1992) no. 6, 389 – 395.
303. S. Damodaran in S. Nakai, H. W. Modler (eds.): *Food Proteins - Properties and Characterisation,* VCH, New York 1996,pp. 167 – 234.
304. A. M. Mahmoud, S. L. Taylor, *Food Technol.* **10** (1994) 67 – 98.
305. P. J. Milla, *Eur. J. Clin. Nutr.* **45** (1991) no. 1, 23 – 28.
306. A. D. Siemensma, W. J. Weijer, H. J. Bak, *Trends in Food Science & Technology* **1** (1993) no. 4, 16 – 21.
307. T. Nakamura, H. Sado, Y. Syukunobe, T. Hirota, *Milchwissenschaft* **48** (1993) no. 12, 667 – 670.
308. A. Teglia, G. Mazzola, G. Secchi, *Cosmet. Toiletries* **108** (1993) no. 11, 56 – 65.
309. G. R. Mintz, G. M. Reinhart, B. Lent, *J. Soc. Cosmet. Chem.* **42**(1991)no. 1, 35 – 44.
310. A. Gildberg, *Process Biochem.* **28**(1993)1 – 15.
311. G. Reich, *Das Leder* **1** (1995) 2 – 11.
312. A. Sander, E. Eilers, A. Heilemann, E. von Kries, *Fett/Lipid* **99** (1997) no. 4, 115 – 120.
313. A. Domsch, B. Irrgang, C. Möller, *Seifen Öle Fette Wachse* **122** (1996) no. 6, 353 – 369.
314. H. W. Steisslinger, *Seifen Öle Fette Wachse* **118** (1992) no. 15, 908 – 918.
315. Henkel Cospha, Cospha Protein Hydrolysate und -Derivate, company brochure, 1992.
316. Henkel High Care, Gluadin - Proteinprodukte der Zukunft, company brochure,1996.
317. R. Y. Yada, R. L. Jackman, J. L. Smith, A. G. Marangoni, in S. Nakai, H. W. Modler (eds.): *Food Proteins - Properties and Characterisation,* VCH, New York 1996, pp. 333 – 403.

Protozoal Infections → Chemotherapeutics
Protozoal Infections → Veterinary Drugs

Psychopharmacological Agents

Marie-Christine Dubroeucq, Rhône-Poulenc Rorer, Vitry-Sur-Seine, France

Jean-Marie Stutzmann, Rhône-Poulenc Rorer, Vitry-Sur-Seine, France

Franco Manfré, Rhône-Poulenc Rorer, Vitry-Sur-Seine, France

Marc Capet, Rhône-Poulenc Rorer, Vitry-Sur-Seine, France

Georg Andrees Böhme, Rhône-Poulenc Rorer, Vitry-Sur-Seine, France

1. Introduction 3
2. Neuroleptics 3
2.1. Phenothiazines 4
2.1.1. 10-Aminoalkyl-Substituted Phenothiazines 4
2.1.2. 10-Piperidylalkyl-Substituted Phenothiazines 5
2.1.2.1. Piperidylethyl-Substituted Phenothiazines 6
2.1.2.2. Piperidinopropyl-Substituted Phenothiazines 6
2.1.3. 1-Piperazinylpropyl-Substituted Phenothiazines 7
2.2. Thioxanthenes 8
2.3. Piperidinobutyrophenones 9
2.4. Diphenylbutylpiperidines 11
2.5. Dibenzazepines and Analogues . . 12
2.5.1. Dibenzazepines 12
2.5.2. Piperazinyl Derivatives 13
2.6. Benzamides 14
2.7. Miscellaneous Neuroleptics 16
2.8. Neuroleptics Under Development 16
3. Antidepressants 17
3.1. Tricyclic Antidepressants 17
3.1.1. Inhibitors of Noradrenaline and/or Serotonin Uptake 18
3.1.2. Enhancers of Uptake 22
3.1.3. Inhibitors of Dopamine Uptake . . . 22
3.1.4. Other Tricyclic Antidepressants . . 22
3.2. Tetracyclic Antidepressants 24
3.3. Monoamine Oxidase Inhibitors (MAOI) 25
3.3.1. Hydrazine Derivatives 25
3.3.2. Nonhydrazine Derivatives 26
3.3.3. Reversible MAO-A Inhibitors 27
3.4. Inhibitors of Serotonin Uptake . . 27
3.5. Lithium 30
3.6. Miscellaneous Antidepressants . . 30
3.7. Antidepressants Under Development 32
4. Anxiolytics 32
4.1. Benzodiazepines 33
4.1.1. 1,4-Benzodiazepines 33
4.1.2. Other Benzodiazepines 38
4.2. Carbamates 39
4.3. Other Anxiolytics 40
4.4. Anxiolytics Under Development . 41
5. Geriatric Drugs 42
5.1. Nootropics 42
5.2. Cholinergic Drugs 43
5.2.1. Choline Analogues 43
5.2.2. Cholinesterase Inhibitors 44
5.2.3. Thyroliberin Analogues 44
5.3. Metabolically Active Drugs 45
5.4. Antidepressants 47
5.5. Miscellaneous Geriatric Drugs . . 47
5.6. Geriatric Drugs Under Development 48
6. References 48

1. Introduction

The term "psychopharmacological agents" refers to the wide variety of substances that affect the mind. These include not only agents that are used in clinical practice to alleviate or correct symptoms of psychiatric illnesses, but also drugs of abuse that are self-administered by humans for their subjective effects on the mood. The latter generally originate from natural preparations and have been known for a long time (e.g., caffeine, opiates, nicotine, cocaine). The

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_341

present article focuses on psychoactive drugs that are of therapeutic significance in psychiatry. Neurological medications such as hypnotics (→ Hypnotics), analgesics (→ Analgesics and Antipyretics), and sedatives (→ Sedatives) are treated in other chapters of this book, see also → Neuropharmacology.

The beginning of the modern era of psychopharmacology began in the early 1950s with the discovery of the antipsychotic properties of chlorpromazine. Since then, the therapeutic armament has successively benefited from the discovery of antidepressive drugs, anxiolytics or minor tranquilizers, and cognition enhancers that are believed to compensate for age-associated memory impairments and dementias. The psychopharmacological agents will be reviewed following this chronology.

Several classifications based on chemical structures, pharmacological properties, or clinical efficacy have been proposed for psychopharmacological agents, but none appears entirely satisfactory. Currently available drugs work symptomatologically, but do not alleviate the causes of the illnesses. The classification used here is based on the medical use of the psychopharmacological agents. The individual compounds are divided into categories with a common structural feature or a similar mode of action; they have been selected on the basis of their historical importance and their present world market. Several substances undergoing clinical development have also been reviewed. Lists of trade names are not exhaustive.

Some background of neurobiology has been included when necessary for the understanding of the mechanism of action of the drugs. Since the early 1970s knowledge of the central nervous system has become increasingly complicated by the discovery of brain processes with which psychoactive drugs may interact. A comprehensive account of the present state of the art in this field can be found elsewhere(→ Neuropharmacology).

2. Neuroleptics

Psychiatric disorders are generally divided into two classes. *Psychoses* (e.g., schizophrenia, mania, and depression) are severe diseases believed to result from the dysfunction of some brain processes. *Neuroses* refer to less debilitating affections related to exagerated reactions to life events otherwise considered as normal (e.g., anxiety).

The *neuroleptics,* also called *major tranquilizers* or *antipsychotics,* act on schizophrenia, the most prominent form of psychosis (1 % of the population is affected). Until the twentieth century, no treatment was available for those suffering from psychiatric disorders, particularly schizophrenia. Early medications are of little significance today (e.g., extracts of ergot or *Rauwolfia serpentina*). The observation of the antipsychotic properties of chlorpromazine and its introduction in clinical practice by Delay and Deniker in 1952 constituted a major breakthrough in the treatment of psychotic conditions [1]. This molecule, initially tested for its sedative properties, gave rise to the phenothiazine family (→ Phenothiazine and Derivatives). Several other families of neuroleptics were discovered soon after, including the thioxanthenes, butyrophenones, dibenzazepines, their piperazinyl derivatives and benzamides. The symptoms of schizophrenia include positive signs, such as psychomotor agitation, delusions, hallucinations, and negative symptoms including loss of social interaction and affective flattening. The observation that clinically effective neuroleptics have dopamine antagonist properties has led to the widely accepted hypothesis that an overactivity of this neurotransmitter system is involved in the disease [2–5]. Neuroleptics have a high affinity for D_1 and D_2 dopamine receptors, the D_2 receptors being of particular importance for their clinical effectiveness. The discovery of additional dopamine receptor subtypes may underlay the activity of several atypic neuroleptics [6]. Moreover, interaction with serotoninrgic processes in the brain may be involved in the action of certain antipsychotics. The experimental pharmacology of classical neuroleptics is characterized by their capacity to induce catalepsy and to antagonize the behavioral effects of dopaminomimetic agents (e.g., apomorphine). They also decrease body temperature, potentiate the activity of hypnotic and analgesic drugs, and impair conditioned avoidance responses. The dopamine antagonist properties are also related to the neurological side effects of these compounds (akathisia, dystonia, tardive dyskinesia). Other side effects arise from inter-

ference with dopaminergic systems in the hypothalamus or pituitary (e.g., increased prolactin levels). Neuroleptics can also be used against nausea and vomiting, and some are useful in the treatment of neuropsychiatric syndromes (e.g., Huntington's disease).

2.1. Phenothiazines

Phenothiazine itself was first synthesized by the end of the ninteenth century and used as a dye.

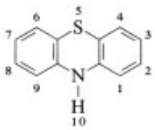

Its insecticidal and anthelmintic properties were described in the mid 1930s. The observation of the strong sedative properties of promethazine, a close analog with antihistaminic activity, prompted the synthesis of other centrally acting phenothiazines. This project resulted in the discovery of chlorpromazine, the prototype antipsychotic drug [1]. For a detailed description of phenothiazines, see → Phenothiazine and Derivatives.

The nitrogen atom of phenothiazine is substituted by aminoalkyl, piperidyl, or piperazinylalkyl side chains. The length of the side chain is critical: neuroleptic derivatives have three carbon atoms between the heterocyclic nitrogen and the tertiary nitrogen of the side chain. Compounds with only two carbon atoms show strong antihistaminic or anticholinergic activities.

The neuroleptic phenothiazines are usually synthesized by condensing a 2-substituted phenothiazine with halogenoalkylamines in the presence of a base:

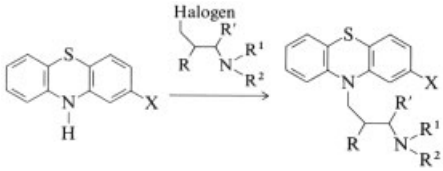

2.1.1. 10-Aminoalkyl-Substituted Phenothiazines

The aliphatic side-chain derivatives are sedatives and act on cholinergic, histaminergic, dopaminergic, and adrenergic receptors [7].

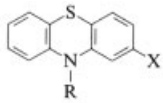

$R = (CH_2)_3N(CH_3)_2$: promazine (X = H); chlorpromazine (X = Cl); triflupromazine ($X = CF_3$); acepromazine ($X = COCH_3$)

$R = (CH_2)CH(CH_3)CH_2N(CH_3)_2$: alimemazine (X = H); levomepromazine ($X = OCH_3$); cyamemazine (X = CN)

Promazine [*58-40-2*], 10-(3-dimethylaminopropyl)-10*H*-phenothiazine, $C_{17}H_{20}N_2S$, M_r 284.41, *bp* (0.04 kPa) 203 – 210 °C, for preparation see [8].

Trade Names. Atarzine (Saunders, Canada), Calmotal (SIT, Italy), Centractyl (Astra), Eliranol (Wyeth), Neuroplegil (Gentili, Italy), Prazine (Wyeth), Prozine (Hauck, USA), Sediston (Serono), Talofen (Pierrel, Italy), Tranquazine (Anthony, USA).

Chlorpromazine [*50-53-3*], 2-chloro-10-(3-dimethylaminopropyl)-10*H*-phenothiazine, $C_{17}H_{19}ClN_2S$, M_r 318.88, *bp* (0.106 kPa) 200 – 205 °C, for preparation see [9].

Trade Names. Elmarine (Marion, USA), Esmind (Otsuka, Japan), Largactil (Rhône-Poulenc), Megaphen (Bayer), Promactil (Wassermann, Spain), Promapar (Parke Davis), Promexin (Meiji, Japan), Propaphenin (Deutsches Hydrierwerk, Germany), Prozin (Lusofarmaco, Italy), Psychozine (Forest, USA), Thorazine (SKF).

Triflupromazine [*146-54-3*], 10-(3-dimethylaminopropyl)-2-trifluoromethyl-10*H*-phenothiazine, $C_{18}H_{19}F_3N_2S$, M_r 352.49, *bp* (0.053 kPa) 162 – 164 °C, for preparation see [10].

Trade Names. Psyquil (Squibb, France), Siquil (Squibb, Belgium), Vesprin (Squibb, Canada).

Acepromazine [*61-00-7*], 2-acetyl-10-(3-dimethylaminpropyl)-10*H*-phenothiazine, $C_{19}H_{22}N_2OS$, M_r 326.47, *bp* (0.066k Pa) 220–240 °C, for preparation see [11].

Trade Names. ACP C-VET (Arovet, Switzerland), Calmivet (Vetoquinol, France), Neurotranq (Alfasan, The Netherlands), Sedalin (Chassot, Switzerland).

Alimemazine [*84-96-8*],10-(3-dimethylamino-2-methylpropyl)-10*H*-phenothiazine, $C_{18}H_{22}N_2S$, M_r 298.44, *mp* 68 °C, for preparation see [12].

Trade Names. Panectyl (Specia, France), Spansule (Herbert, USA), Temaril (SKF), Théralène (Rhône-Poulenc), Vallergan (May & Baker).

Levomepromazine [*60-99-1*], 2-methoxy-10-(3-dimethylamino-2-methylpropyl)-10*H*-phenothiazine, $C_{19}H_{24}N_2OS$, M_r 328.46, *mp* 190 °C, for preparation see [12].

Trade Names. Levomezine (Toho, Japan), Levomeprome (Lederle), Levonormal (Sawai, Japan), Minozinan (Specia), Neurocil (Bayer), Nozinan (Specia), Procrazine (Toyo Pharmar, Japan), Sofmin (Dainippon, Japan), Veractil (Rhône-Poulenc), Sinogan (Rhone, Spain).

Cyamemazine [*3546-03-0*], 2-cyano-10-(3-dimethylamino-2-methylpropyl)-10*H*-phenothiazine, $C_{19}H_{21}N_3S$, M_r 323.47, *mp* 89–96 °C, for preparation see [13].

Trade Name. Tercian (Théraplix, France).

2.1.2. 10-Piperidylalkyl-Substituted Phenothiazines

The phenothiazines containing piperidyl side chains have a relatively high neuroleptic potency; some induce less extrapyramidal side effects than chlorpromazine. They show anticholinergic and adrenolytic activity.

2.1.2.1. Piperidylethyl-Substituted Phenothiazines

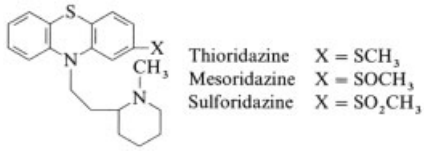

Thioridazine [*50-52-2*], 2-methylmercapto-10-[2-(1-methyl-2-piperidyl)ethyl]-10*H*-pheno-thiazine, $C_{21}H_{26}N_2S_2$, M_r 370.56, *mp* 72–74 °C, for preparation see [14].

Trade Names. Mallorol, Meleril, Mellaril, Melleril (Sandoz), Novoridazine (Novopharm, Canada), Orsanil (Orion, Finland), Ridazin (Taro, Israel), Stalleril (Pharmacal, Finland), Thioril (ICN, Canada).

Mesoridazine [*5588-33-0*], 2-methylmercapto-10-[2-(1-methyl-2-piperidyl)ethyl]-10*H*-phenothiazine-2-oxide, $C_{21}H_{26}N_2OS_2$, M_r 386.59, *mp* 115–120 °C, for preparation see [15].

Trade Names. Calodal (Heyden, Germany), Lidanil (Sandoz), Serentil (Sandoz).

Sulforidazine [*14759-06-9*], 2-methylsulfonyl-10-[2-(1-methyl-2-piperidyl)ethyl]-10*H*-phenothiazine, $C_{21}H_{26}N_2O_2S_2$, M_r 402.57, *mp* 121–123 °C, for preparation see [16].

Trade Name. Inofal (Sandoz, Germany).

2.1.2.2. Piperidinopropyl-Substituted Phenothiazines

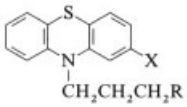

Pipamazine	X = Cl	R = N–CONH$_2$
Pericyazine	X = CN	R = N–OH
Piperacetazine	X = COCH$_3$	R = N–(CH$_2$)$_2$OH
Pipotiazine	X = SO$_2$N(CH$_3$)$_2$	R = N–(CH$_2$)$_2$OH
Spiclomazine	X = Cl	R = N–S–N–O, H

Pipamazine [*84-04-8*], 10-[3-(4-carbamoylpiperidino)propyl]-2-chloro-10*H*-phenothiazine, $C_{21}H_{24}ClN_3OS$, M_r 401.97, *mp* 139 °C, for preparation see [17]. Pipamazine is also used as an antiemetic.

Trade Names. Nometine (Searle), Nansidol, Mornitine.

Pericyazine [*2622-26-6*], 2-cyano-10-[3-(4-hydroxypiperidino)propyl]-10*H*-phenothiazine, $C_{21}H_{23}N_3OS$, M_r 365.50, *mp* 116–117 °C, for preparation see [18].

Trade Names. Amplan (Taisho, Japan), Aolept (Bayer), Apamin (Yoshitomi, Japan), Nemactil (Specia), Neulactil (May & Baker), Neuleptil (Rhône-Poulenc).

Piperacetazine [*3819-00-9*], 2-acetyl-10-[3-[4-(2-hydroxyethyl)piperidino]propyl]-10*H*-phenothiazine, $C_{24}H_{30}N_2O_2S$, M_r 410.59, *mp* 100–110 °C, for preparation see [19].

Trade Name. Quide (Merrell Dow).

Pipotiazine [*39860-99-6*], 10-{3-[4-(2-hydroxyethyl)piperidino]propyl}-*N,N*-dimethyl-10*H*-phenothiazine-2-sulfonamide, $C_{24}H_{33}N_3O_3S_2$, M_r 475.66, *mp* 58–59 °C, for preparation see [20].

Trade Names. Piportil (Rhodia, Brazil), Piportil Depot (May & Baker), Piportil Longum (Specia), Piportyl (Rhône-Poulenc).

Spiclomazine [*24527-27-3*], 8-[3-(2-chloro-10*H*-phenothiazin-10-yl)propyl]-1-thia-4,8-diazaspirol[4, 5]decan-3-one, $C_{22}H_{24}ClN_3OS_2$, M_r 446.02, *mp* 154–156 °C, for preparation see [21].

Trade Name. Diceplon (Yoshitomi, Japan).

2.1.3. 1-Piperazinylpropyl-Substituted Phenothiazines

The piperazine derivatives are the most potent phenothiazine neuroleptics. They have only mild sedative effects, induce strong extrapyramidal side effects, and have a lower propensity for inducing autonomic side effects due to the lack of high-affinity binding to cholinergic and adrenergic receptors.

Perazine [*84-97-9*], 10-[3-(4-methyl-1-piperazinyl)propyl]-10*H*-phenothiazine, $C_{20}H_{25}N_3S$, M_r 339.49, *mp* 51–53 °C, for preparation see [22].

Trade Name. Taxilan (Promonta).

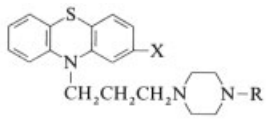

Perazine	X = H	R = CH$_3$
Prochlorperazine	X = Cl	R = CH$_3$
Trifluoperazine	X = CF$_3$	R = CH$_3$
Thioproperazine	X = SO$_2$N(CH$_3$)$_2$	R = CH$_3$
Butaperazine	X = CO(CH$_2$)$_2$CH$_3$	R = CH$_3$
Perphenazine	X = Cl	R = CH$_2$CH$_2$OH
Fluphenazine	X = CF$_3$	R = CH$_2$CH$_2$OH
Acetophenazine	X = COCH$_3$	R = CH$_2$CH$_2$OH
Carphenazine	X = COCH$_2$CH$_3$	R = CH$_2$CH$_2$OH

Prochlorperazine [*58-38-8*], 2-chloro-10-[3-(4-methyl-1-piperazinyl)propyl]-10*H*-phenothi-azine, $C_{20}H_{24}ClN_3S$, M_r 373.94, *mp* 228 °C (dimaleate), for preparation see [23].

Trade Names. Anti-Naus (Protea, Australia), Bucastem (Reckitt & Colman, UK), Compazine (SKF), Nipodal (Bayer), Normalmin (Sawai, Japan), Stemetil (Farmitalia, Italy), Tementil (Specia), Vertigon (SKF).

Trifluoperazine [*117-89-5*], 2-trifluoromethyl-10-[3-(4-methyl-1-piperazinyl)propyl]-10*H*-phenothiazine, $C_{21}H_{24}F_3N_3S$, M_r 407.49, *bp* (0.08 kPa) 202 – 210 °C, for preparation see [24].

Trade Names. Eskazine (SKF), Jatroneural (Röhm Pharma, Germany), Modalina (Maggioni-Winthrop, Italy), Novolurazine (Novapharm, Canada), Pentazine (Pentagone, Canada), Solazine (Horner, Canada), Stelazine (SKF), Telazin (SKF), Terfluzine (Theraplix, France), Trifluoper-Ez-Ets (Barlowe Cote, Canada).

Thioproperazine [*316-81-4*], 10-[3-(4-methyl-1-piperazinyl)propyl]-*N,N*-dimethyl-10*H*-phenothiazine-2-sulfonamide, $C_{22}H_{30}N_4O_2S_2$, M_r 446.64, *mp* 140 °C, for preparation see [25].

Trade Names. Majeptil (Specia, France), Mayeptil (Rhodia, Argentina).

Butaperazine [*653-03-2*], 2-butyryl-10-[3-(4-methyl-1-piperazinyl)propyl]-10*H*-phenothi-azine, $C_{24}H_{31}N_3OS$, M_r 409.60, *bp* (0.006 kPa) 270 – 280 °C, for preparation see [26].

Trade Names. Randolectil (Bayer, Argentina), Repoise (Robins, USA).

Perphenazine [*58-39-9*],2-chloro-10-{3-[4-(2-hydroxyethyl)-1-piperazinyl]propyl}-10*H*-phe-nothiazine, $C_{21}H_{21}ClN_3OS$, M_r 403.97, *mp* 94 –100 °C, for preparation see [27].

Trade Names. Decetan (Merck), Etaperazin (Medexport, IEC), Fentazin (Allen & Hanburys, UK), F-Mon (Kuroishi, Japan), Phenazine (ICN, Canada), Trilafon (Schering), Trilifan (Cétrane, France).

Metofenazate [*388-51-2*], $C_{31}H_{36}ClN_3O_5S$, M_r 598.18, *mp* 102 – 107 °C, is the 3,4,5-trimethoxybenzoate of perphenazine; for preparation see [28].

Trade Name. Frenolon (Egis, Hungary).

Thiopropazate [*84-06-0*], $C_{23}H_{28}ClN_3O_2S$, M_r 446.00, *bp* (0.1 kPa) 214 – 218 °C, is the acetate of perphenazine; for preparation see [29].

Trade Names. Dartal, Dartalan (Searle).

Fluphenazine [*69-23-8*], 10-{3-[4-(2-hydroxyethyl)-1-piperazinyl]propyl}-2-trifluoromethyl-10*H*-phenothiazine, $C_{22}H_{26}F_3N_3OS$, M_r 437.52, *bp* (0.066 kPa) 268 – 274 °C, for preparation see [30].

Trade Names. Anatensol, Calmansial, Dapotum, Moditen, Prolixin (Squibb); Eutimox (Soc. Gen. de Pharmacia, Switzerland); Lyorodin (Deutsches Hydrierwerk, Germany); Omca (Heyden, Germany); Pacinol, Permitil, Trancin (Schering); Siqualine (Iquinosa, Spain).

Acetophenazine [*2751-68-0*], 2-acetyl-10-{3-[4-(2-hydroxyethyl)-1-piperazinyl]propyl}-10*H*-phenothiazine, $C_{23}H_{29}N_3O_2S$, M_r 411.57, *mp* 167 – 168 °C, for preparation see [31].

Trade Name. Tindal (Schering, UK).

Carphenazine [*2622-30-22*], 10-[3-[4-(2-hydroxyethyl)-1-piperazinyl]propyl]-2-propionyl-10*H*-phenothiazine, $C_{24}H_{31}N_3O_2S$, M_r 425.6, *mp* 175 – 177 °C, for preparation see [32]

Trade Name. Proketazine (Wyeth, USA).

Two other compounds are analogous to the phenothiazines described above:

Homofenazine [3833-99-6], 10-{3-[hexahydro-4-(2-hydroxyethyl)-1,4-diazepine-1-yl]propyl}-2-trifluoromethyl-10*H*-phenothiazine, $C_{23}H_{28}F_3N_3OS$, M_r451.52, *bp* (0.133 kPa) 230 – 240 °C, is homologous to fluphenazine; for preparation see [33].

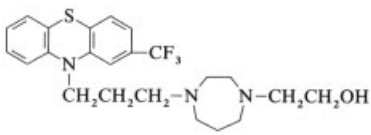

Trade Name. Pasaden (Homburg).

Prothipendyl [*1225-65-6*], 10-(3-dimethylaminopropyl)-10*H*-pyrido[3,2-*b*][1, 4]benzothiazine, $C_{16}H_{19}N_3S$,M_r285.42, *bp* (0.093 kPa) 217 –219 °C, for preparation see [34], is analogous to promazine (azaphenothiazine) with general properties similar to those of chlorpromazine.

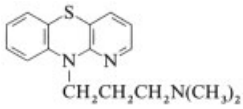

Trade Names. Dominal (Astra, Germany), Tolnate (SKF, UK).

2.2. Thioxanthenes

The thioxanthene neuroleptics are chemically and pharmacologically related to the phenothiazines. Activity resides with the *cis* isomers. The aminoalkyl and piperazinylalkyl side chains are similar to those of the phenothiazines, producing comparable potencies and side effects. The thioxanthenes are prepared by condensing the tricyclic ketone with the Grignard reagent of the desired aminoalkyl chain. The resulting alcohol is then dehydrated in an acidic medium.

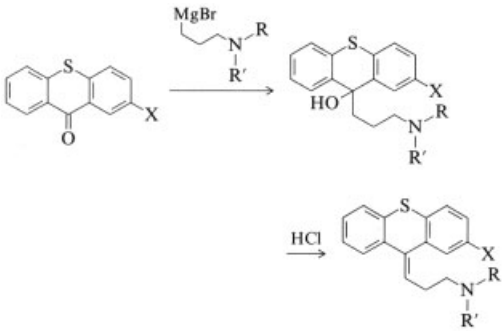

Chlorprothixene [*113-59-7*], 2-chloro-9-(3-dimethylaminopropylidene)-9*H*-thioxanthene, $C_{18}H_{18}ClNS$, M_r 315.86, *mp* 97 – 98 °C, for preparation see [35].

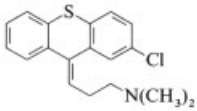

Trade Names. Cloxan (Orion, Finland); Paxyl (Ikepharm); Tactaran (Roche), Taractan, Tarasan, Taratan (Roche); Truxal, Truxaletten (Lundbeck).

Thiothixene [*3313-26-6*], *cis*-9-[3-(4-methyl-piperazinyl)propylidene]-*N*,*N*-dimethyl-9*H*- thioxanthene-2-sulfonamide $C_{23}H_{29}N_3O_2S_2$, M_r 443.63, *mp* 147 – 149 °C, for preparation see [36].

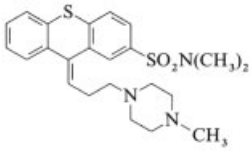

Trade Names. Navane, Navaron, Orbinamon (Pfizer).

Flupentixol [*2709-56-0*], 9-{3-[4-(2-hydroxyethyl)-1-piperazinyl]propylidene}-2-trifluoromethyl-9*H*-thioxanthene, $C_{23}H_{25}F_3N_2OS$, M_r 434.54, *mp* 236 – 240 °C, is prepared with the alcoholic function protected as a benzyl ether. This protection is removed by catalytic hydrogenation at the end of the synthesis [37].

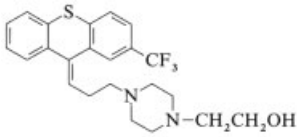

Trade Names. Depixol, Fluanxol (Lundbeck, Denmark), Emergil (Labaz), Metamin (Takeda, Japan), Phrenixol (Ikepharm, Israel), Siplarol (Farmitalia, Italy).

Clopenthixol [*982-24-1*], 2-chloro-9-{3-[4-(2-hydroxyethyl)-1-piperazinyl]propylidene}-9*H*-thioxanthene, $C_{22}H_{25}ClN_2OS$, M_r 400.99, *mp* 250–260 °C (dihydrochloride), for preparation see [38].

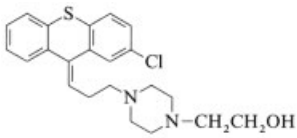

Trade Names. Ciatyl, Sordena, Sordinol (Kefalas).

Zuclopenthixol [*53772-83-1*], $C_{22}H_{25}ClN_2OS$, M_r 400.99, *mp* 84–85 °C, is the *cis* isomer of clopenthixol [38].

Trade Names. Cisodinol, Clopixol (Lundbeck, Denmark); Sedanxol (Tropon, Germany).

2.3. Piperidinobutyrophenones

The piperidinobutyrophenones are the most potent neuroleptics (benperidol, trifluperidol). They have a higher binding selectivity for D_2 than for D_1 receptors, lack affinity for cholinergic receptors, and show less sedative side effects [7]. The butyrophenones are prepared by condensing the relevant piperidine derivative with 4-chloro-1-(4-fluorophenyl)-1-butanone.

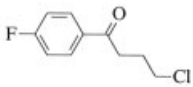

4-Chloro-1-(4-fluorophenyl)-1-butanone

Haloperidol [*52-86-8*], 4-[4-(4-chlorophenyl)-4-hydroxypiperidino]-4′-fluorobutyrophenone, $C_{21}H_{23}ClFNO_2$, M_r 375.88, *mp* 148–149 °C, for preparation see [39]. Haloperidol is also used in the management of Gilles de la Tourette's syndrome.

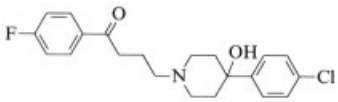

Trade Names. Brotopon (Pfizer), Duraperidol (Durachemie, Germany), Haldol (Janssen), Halojust (Horita, Japan), Halopidol (Johnson & Johnson), Halosten (Shionogi, Japan), Keselan (Sumitomo, Japan), Linton (Yoshitomi, Japan), Novoperidol (Novopharm, Canada), Selezyme (Sawai, Japan), Serenace (Searle).

Bromperidol [*10457-90-6*], 4-[4-(4-bromophenyl)-4-hydroxypiperidino]-4′-fluorobutyrophenone, $C_{21}H_{23}BrFNO_2$, M_r 420.33, *mp* 155–158 °C [39].

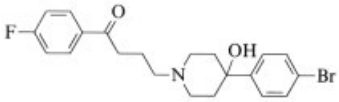

Trade Names. Azuren, Bromidol, Impromen (Janssen), Tesoprel (Organon, Germany).

Moperone [*3871-82-7*], 4′-fluoro-4-[4-hydroxy-4-(4-methylphenyl)piperidino]butyrophenone, $C_{22}H_{26}FNO_2$, M_r 355.46, *mp* 118–119 °C, for preparation see [40].

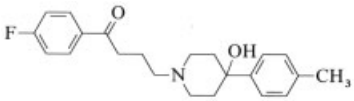

Trade Names. Luvatren, Luvatrena, Luvatrene (Cilag-Chemie).

Fluanisone [*1480-19-9*], 4′-fluoro-4-[4-(2-methoxyphenyl)-1-piperazinyl]butyrophenone, $C_{21}H_{25}FN_2O_2$, M_r 356.45, *mp* 205 °C, for preparation see [41].

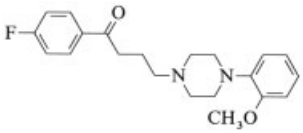

Trade Name. Sédaland (Delalande, France).

Trifluperidol [*749-13-3*], 4′-fluoro-4-[4-hydroxy-4-(3-trifluoromethylphenyl)piperidino]-butyrophenone, $C_{22}H_{23}F_4NO_2$, M_r 409.43, *mp* 200 – 201 °C, for preparation see [42].

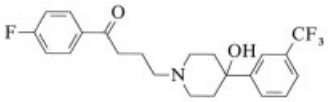

Trade Name. Triperidol (Lagap, UK).

Pipamperone [*1893-33-0*], 4-[(4-carbamoyl-4-piperidino)-4′-fluorobutyrophenone, $C_{21}H_{30}FN_3O_2$, M_r 375.49, *mp* 124 – 126 °C (dihydrochloride), for preparation see [43].

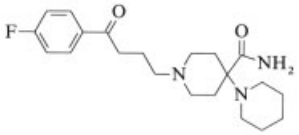

Trade Names. Dipiperal (Johnson & Johnson), Dipiperon (Janssen, France), Piperonil (Lusofarmaco, Italy), Propitan (Esai, Japan).

Spiperone [*749-02-0*], 8-[3-(4-fluorobenzoylpropyl]-4-oxo-1-phenyl-1,3,8-triazaspiro[4,5]-decane, $C_{23}H_{26}FN_3O_2$, M_r 395.48, *mp* 190 – 194 °C, for preparation see [44].

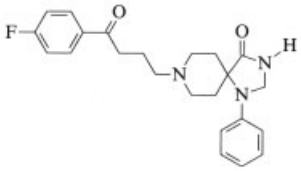

Trade Name. Spiropitan (Japan).

Benperidol [*2062-84-2*], 1-{1-[3-(4-fluorobenzoyl)propyl]-4-piperidyl}-2-benzimidazolinone, $C_{22}H_{24}FN_3O_2$, M_r 381.45, *mp* 170 – 172 °C, for preparation see [45].

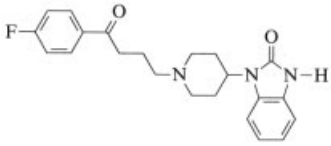

Trade Names. Anquil (Janssen, UK), Frenactil (Janssen, Belgium), Clianimon (Tropon, Germany), Psicoben (Ravizza, Italy).

Timiperone [*57648-21-2*], 1-{1-[3-(4-fluorobenzoyl)propyl]-4-piperidyl}-2-mercaptobenzimidazole, $C_{22}H_{24}FN_3OS$, M_r 397.5, *mp* 201 –203 °C, for preparation see [46].

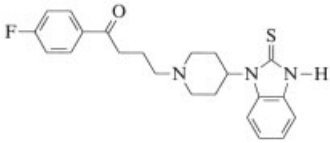

Trade Name. Tolopelon (Daichi, Japan).

Droperidol [*548-73-2*], 1-{1-[3-(4-fluorobenzoyl)propyl]-1,2,3,6-tetrahydro-4-pyridyl}-2-benz-imidazolinone, $C_{22}H_{22}FN_3O_2$, M_r 379.4, *mp* 145 – 146 °C, for preparation see [47]. Droperidol shows a long duration of action. It is used in conjunction with an analgesic (e.g., fentanyl) to induce neuroleptanalgesia in anesthesiology.

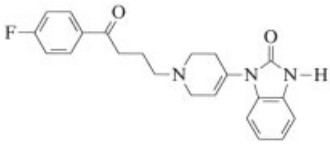

Trade Names. Dehydrobenperidol, Droleptan (Janssen), Inapsine (McNeil, USA), Sintodian (Farmitalia Carlo Erba).

Melperone [*3575-80-2*], 4′-fluoro-4-(4-methylpiperidino)butyrophenone, $C_{16}H_{22}FNO$, M_r 263.37, *mp* 209 – 211 °C (hydrochloride), for preparation see [48]. Unlike the other butyrophenones, melperone shows only weak affinity for the dopamine receptors. It seems to act

through the serotonin 5-HT_2 receptor for serotonin (5-hydroxytryptamine, 5-HT). Melperone is one of the noncataleptogenic neuroleptics with a low tendency for inducing extrapyramidal side effects.

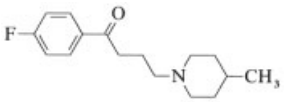

Trade Names. Buronil (Ferrosan, Italy), Eunerpan (Nordmark, Germany).

2.4. Diphenylbutylpiperidines

The diphenylbutylpiperidines are structurally related to the piperidinobutyrophenones. They are relatively selective for the D_2 receptors and do not interact with histaminergic receptors [7]. In addition, these neuroleptics are able to block voltage-dependent calcium channels [49].

Fluspirilene [*1841-19-6*], 8-[4,4-bis(4-fluorophenyl)butyl]-1-phenyl-1,3,8-triazaspiro[4.5]-decan-4-one, $C_{29}H_{31}F_2N_3O$, M_r 475.59, *mp* 187 – 190 °C is prepared by condensing a spirobicyclic amine with the relevant brominated diphenylbutane derivative. The spiro compound is prepared from cyclohexanone via an aminonitrile (Strecker synthesis) [50].

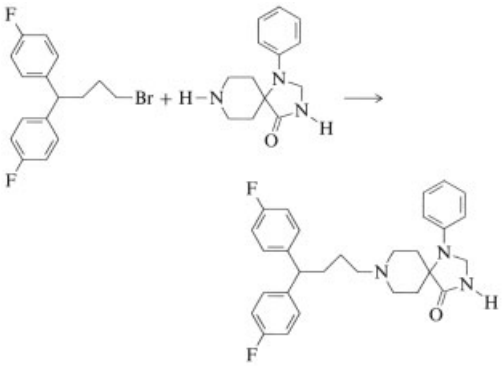

Fluspirilene bears the same side chain as spiperone (see Section 2.3). The deep intramuscular injection of a microcrystalline suspension of fluspirilene results in a long duration of action (ca. one week).

Trade Names. Imap (Janssen), Redeptin (SKF, UK).

Penfluridol [*26864-56-2*], 1-[4,4-bis(4-fluorophenyl)butyl]-4-(4-chloro-3-trifluoromethylphenyl)-4-hydroxypiperidine, $C_{28}H_{27}ClF_5NO$, M_r 523.99, *mp* 105 – 107 °C, is obtained by reducing the ketone of 3-(4-fluorobenzoyl)propionic acid into a lactone. Alkylation of fluorobenzene with this lactone under Friedel – Crafts conditions yields a diarylbutyric acid, which is converted into an amide and reduced to the amine; for details of synthesis see [51].

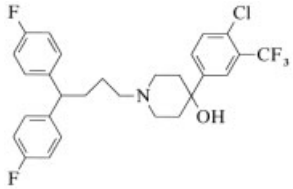

Trade Names. Cyperon (Esteve, Spain), Fluidol (Janssen), Longoran (Isis, Yugoslavia), Micefal (Spofa, Czechoslovakia).

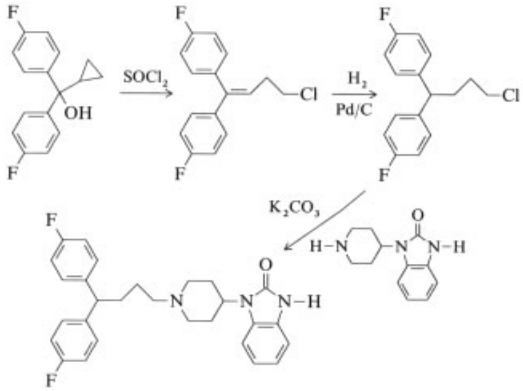

Pimozide [*2062-78-4*], 1-{1-[4,4-bis(4-fluorophenyl)butyl]-4-piperidinyl}-2-benzimidazolinone, $C_{28}H_{29}F_2N_3O$, M_r 461.56, *mp* 214 – 218 °C. Ring opening of diphenylcyclopropylcarbinol gives a homoallylic chloride

which is hydrogenated and then condensed with the relevant benzimidazolinone derivative [52]. The side chain is identical to that of benperidol (see page 10).

Pimozide is used for the treatment of acute and chronic schizophrenia and also in management of Gilles de la Tourette's syndrome.

Trade Names. Antalon (Arzneimittelwerk Dresden, Germany), Opiran (Roussel, France), Oralep (Abic, Israel), Orap (Janssen), Pimotid (Medica, Finland), Pirium (Janssen)

2.5. Dibenzazepines and Analogues

This class of drugs includes tricyclic compounds with seven-membered central rings. Clozapine was the first neuroleptic that seldom showed extrapyramidal side effects and was therefore termed an "atypical" neuroleptic.

Two subclasses can be distinguished:

1) The seven-membered ring is an azepine having a piperidinoalkyl chain on its nitrogen atom (carpipramine and clocapramine)
2) The basic substituent (a *N*-methylpiperazine) is attached to the imino bridge

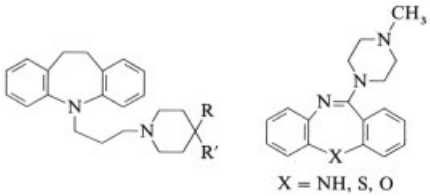

Structural variations in this series of compounds led to the development of classical neuroleptics (clothiapine, loxapine, and clozapine) and hypnotics (perlapine).

2.5.1. Dibenzazepines

The dibenzazepine derivatives are prepared by the following general method: the tricyclic nitrogen atom is alkylated with a 1,3-dihalogenated propane, the resulting monohalogeno compound is then condensed with the piperidine derivative:

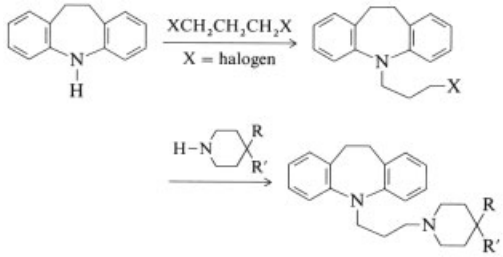

Carpipramine [*5942-95-0*], 5-[3-(4-piperidino-4- carbamoylpiperidino)propyl]-10,11-dihydro-5*H*-dibenz[*b,f*]azepine, $C_{28}H_{38}N_4O$, M_r 446.64, *mp* 260 °C, for preparation see [53]. The compound is a 5-HT_2 antagonist. This could explain its clinical efficacy in symptoms like asthenia, anxiety, and social withdrawal.

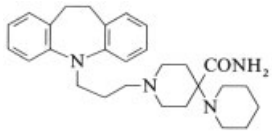

Trade Names. Defekton (Yoshitomi, Japan), Prazinil (Fournier, France).

Clocapramine [*47739-98-0*], 5-{[3-(4-carbamoyl-4-piperidino]propyl}-3-chloro-10,11-dihydro-5*H*-dibenz[*b,f*]azepine dihydrochloride monohydrate, $C_{28}H_{41}Cl_3N_4O_2$, M_r 572.0, *mp* 267 °C is a chlorinated derivative of carpipramine with a bitter taste; for preparation see [53].

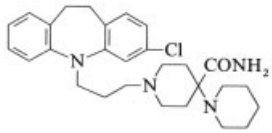

Trade Name. Clofekton (Yoshitomi, Japan).

2.5.2. Piperazinyl Derivatives

Dibenzazepine analogues are prepared by two main routes (Fig. 1). In the first one, the lactam function of the tricycle is activated as an imino chloride (A) or a thioether (B), which is then condensed with *N*-methyl piperazine. In the second

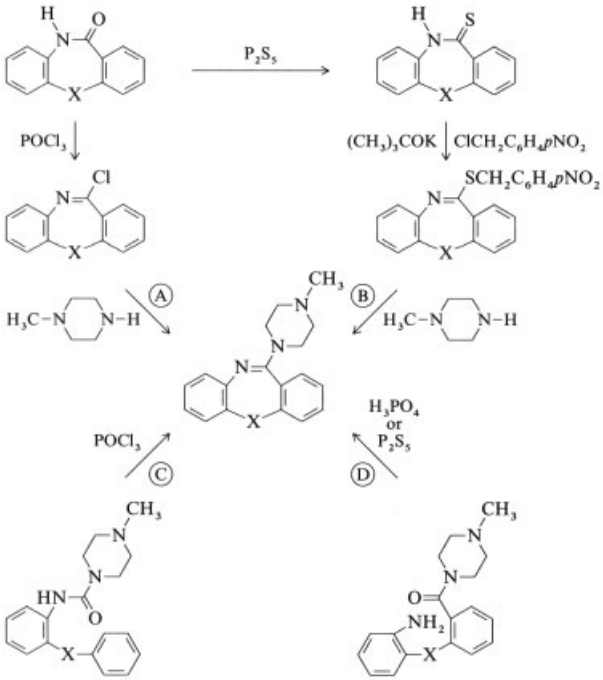

Figure 1. Synthesis of benzazepine analogues

route, the imino bridge of the tricycle is formed by dehydration of an urea function (C) or condensation of an amine and an amide function (D).

Clothiapine [*2058-52-8*], 2-chloro-11-(4-methyl-1-piperazinyl)-dibenzo[*b,f*][1, 4]thiazepine, $C_{18}H_{18}ClN_3S$, M_r 343.89, *mp* 118–120 °C, for preparation see [54].

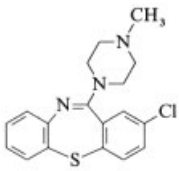

Trade Names. Deliton (Dainippon, Japan), Entumine, Etomine, Etumina, Etumine (Sandoz Wander).

Clozapine [*5786-21-0*], 8-chloro-11-(4-methyl-1-piperazinyl)-5*H*-dibenzo[*b, e*][1, 4]diazepine, $C_{18}H_{19}ClN_4$, M_r 326.83, *mp* 183–184 °C, for preparation see [55].

Despite severe hematologic side effects (agranulocytosis), this compound is used to treat otherwise irresponsive patients. Clozapine has one order of magnitude higher affinity for the D_4 receptors than for D_2 receptors [56].

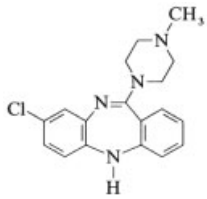

Trade Names. Alemoxan (Arzneimittelwerk Dresden, Germany), Clozaril (Sandoz Wander, UK), Leponex (Sandoz Wander).

Loxapine [*1977-10-2*], 2-chloro-11-(4-methyl-1-piperazinyl)-dibenz[*b,f*][1, 4]oxazepine, $C_{18}H_{18}ClN_3O$, M_r 327.81, *mp* 109–110 °C, for preparation see [57].

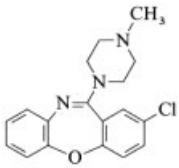

Trade Names. Daxolin (Dome, USA); Desconex (Lafarquim, Spain); Loxapac, Loxitane (Lederle).

2.6. Benzamides

The benzamide neuroleptics are selective blockers of D_2 receptors and are therefore free of the side effects arising from interactions with other receptors. Sulpiride, the most important member of this class, was recognized as an "atypical" neuroleptic [58]. It exhibits a specific action on the mesolimbic dopamine system [59].

Sulpiride [*15676-16-1*], *N*-[(1-ethyl-2-pyrrolidinyl)methyl]-2-methoxy-5-sulfamoylbenzamide, $C_{15}H_{23}N_3O_4S$, M_r 341.43, *mp* 175 – 182 °C, is prepared by reaction of 2-aminomethyl-1-ethyl-pyrrolidine with phosphorus trichloride, followed by the condensation of the intermediate with 5-sulfamoyl-2-methoxybenzoic acid [60].

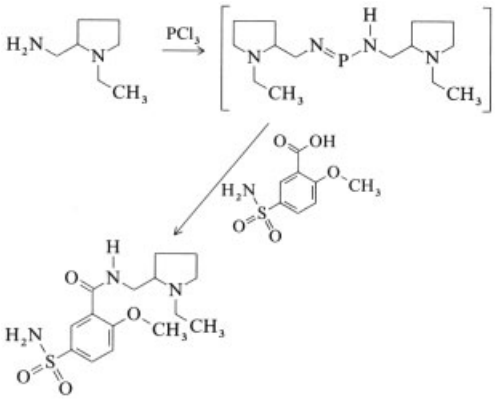

Trade Names. Abilit (Sumitomo, Japan), Aiglonyl (Fu-mouze, France), Arminol (Krewel, Germany), Azugastan (Kobayashi, Japan), Betamac, Dixibon (Sandoz, Switzerland), Dobren (Ravizza, Italy), Dogmatil (Delagrange, France), Dolmatil (Squibb), Equilid (Merrell Dow), Levopraid (Ravizza, Italy), Miradol (Mitsui, Japan), Mirbanil (Boehringer, Germany), Neuromyfar (Emifar, Spain), Sulpril (Astra), Tohpiride (Toyo Pharma, Japan).

Sultopride [*53583-79-2*], *N*-[(1-ethyl-2-pyrro-lidinyl)methyl]-5-(ethylsulfonyl)-2-methoxybenz-amide, $C_{17}H_{26}N_2O_4S$, M_r 354.47, *mp* 181 –182 °C, is related to sulpiride and is prepared analogously [60].

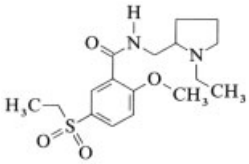

Trade Names. Barnetil (Delagrange, France), Barnotil (Vita, Italy), Topral (Alkaloid, Yugoslavia).

Tiapride [*51012-32-9*], *N*-[2-(diethylamino)-ethyl]-2-methoxy-5-(methylsulfonyl)benzamide, $C_{15}H_{24}N_2O_4S$, M_r 328.43, *mp* 123 – 125 °C, is prepared by activating 2-methoxy-5-(methylsulfonyl)benzoic acid with ethyl chloroformate and reacting the mixed anhydride with 2-(diethylamino)ethylamine [61].

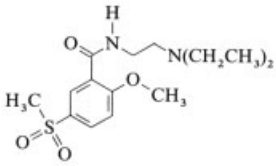

Trade Names. Delpral (Bender, Austria), Equilum (Fumouze, France), Gramalil (Fujisawa, Japan), Luxoben (Chinoin, Italy), Porfanil (Prodes, Spain), Tiaprizal (Delagrange, France).

Amisulpride [*71675-85-9*], 4-amino-*N*-[(1-ethyl-2-pyrrolidinyl)methyl]-5-(ethylsulfonyl)-2-methoxybenzamide, $C_{17}H_{27}N_3O_4S$, M_r 369.48, *mp* 126 – 127 °C, is prepared in an analogous way to tiaprides. 2-Aminomethyl-1-ethylpyrrolidine is condensed with the appropriate benzoic acid derivative in the presence of

ethyl chloroformate [62] or with a benzoyl chloride derivative. In the latter case, the aromatic ring bears a nitro group which is then reduced [63]. Amisulpride is used against both negative and productive symptoms of schizophrenia.

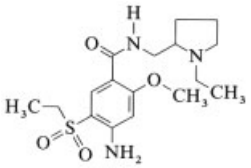

Trade Name. Solian (Delagrange, France).

Remoxipride [*80125-14-0*], (*S*)-3-bromo-2,6-dimethoxy-*N*-(1-ethyl-2-pyrrolidinyl)methyl]-benzamide, $C_{16}H_{23}BrN_2O_3$, M_r 371.23, *mp* 184–185 °C, is prepared by condensing the appropriate benzoyl chloride derivative with the optically active 1-ethyl-2-(aminomethyl)pyrrolidine [64]. Like sulpiride, remoxipride seems to have a specific action on the mesolimbic dopaminergic system and a lower impact on the striatal dopaminergic system, therefore generating less extrapyramidal side effects [65].

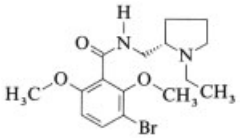

Trade Name. Roxiam (Astra, Denmark).

Clebopride [*55905-53-8*], 4-amino-*N*-(1-benzyl-4-piperidyl)-5- chloro-2-methoxybenzamide, $C_{20}H_{24}ClN_3O_2$, M_r 389.88, *mp* 194–195 °C, for preparation see [66].

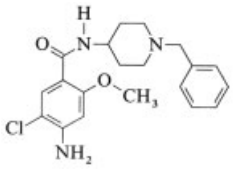

Trade Names. Clanzol (Orfi, Spain), Clebon (Wyeth), Cleboril (Almirall, Spain), Vuxolin (Carlo Erba).

Metoclopramide [*364-62-5*], 4-amino-5-chloro- *N*-[2-(diethylamino)ethyl]-2-methoxybenzamide, $C_{14}H_{22}ClN_3O_2$, M_r 299.81, *mp* 146–148 °C, is widely used as an antiemetic; for preparation see [67].

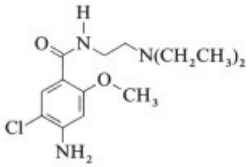

Trade Names. Cerucal (Arzneimittelwerk Dresden, Germany), Duraclamid (Durachemie, Germany), Elieten (Nippon Kagaku, Japan), Emex (Beecham), Gastrese LA (Robins, USA), Gastromax (Farmitalia, Italy), Maxeran (Delagrange, France), Maxolon (Beecham), Meclopran (Lagap, Switzerland), Metramid (Nicholas), Moriperan (Morishita, Japan).

Veralipride [*66644-81-3*], *N*-[(1-allyl-2- pyrrolidinyl)methyl]-5-sulfamoyl-5-veratramide, $C_{17}H_{25}N_3O_5S$, M_r 383.5, for preparation see [68]. It is also used to treat menopausal disorders.

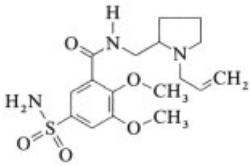

Trade Names. Accional (Fides, Spain), Agradil (Vita, Italy), Agréal (Delagrange, France), Faltium (Prodes, Spain), Verapril (Midy, Italy).

2.7. Miscellaneous Neuroleptics

Molindone [*7416-34-4*], 3-ethyl-2-methyl-5-(morpholinomethyl)-4,5,6,7-tetrahydro-1*H*-indol-4-one, $C_{16}H_{24}N_2O_2$, M_r 276.37, *mp* 180–181 °C, is prepared in two steps. Condensation

of cyclohexane-1,2-dione with a keto-oxime affords the indolic ring system. Mannich condensation with morpholine completes the synthesis [69].

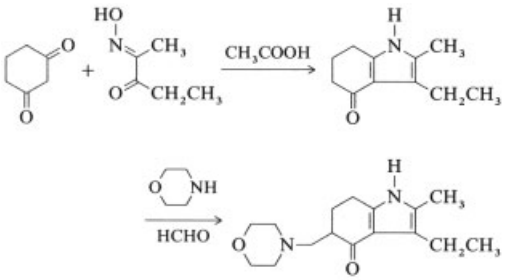

Trade Names. Lidone (Abbott), Moban (Endo, USA).

Oxipertine [*153-87-7*], 5,6-dimethoxy-2-methyl-3-[2-(4-phenyl-1-piperazinyl)ethyl]indole, $C_{23}H_{29}N_3O_2$, M_r 379.5, *mp* 250 – 252 °C, is prepared by condensation of an indoleacetic acid derivative with *N*-phenyl-piperazine, followed by reduction [70].

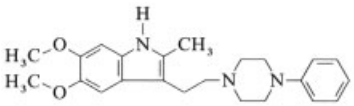

Trade Names. Equipertine (Sterling, Belgium), Forit(Winthrop, Switzerland), Integrin (Sterling, UK), Opertil (Winthrop, Denmark).

2.8. Neuroleptics Under Development

Novel strategies for the treatment of schizophrenia have arisen from the recognition of the possible benefits from additional antagonism at neurotransmitter receptors other than those for dopamine. Thus, several drugs under development act at multiple receptors [71]. *Risperidone* [*106266-06-2*] (R-64 766, Janssen) is a benzisoxazole derivative with mixed D_2/5-HT_2 blocking properties; it is reported to improve negative signs of schizophrenia accompanied with only few motor side effects. *Amperozide* [*75558-90-6*] (FG-5606, Kabi), *ICI 204,636* [*111374-72-2*], and *HP 873* [*133454-47-4*] (Hoechst) are other drugs showing high affinity for 5-HT_2 receptors subjected to clinical investigation for schizophrenia. *Org-5222* [*85650-56-2*] (Organon) is active on a range of receptors including D_1, D_2, α_1, α_2, 5-HT_2, and H_1 receptors; it also antagonizes dopaminomimetics in rodents [72]. Drugs acting at 5-HT_3 receptors have potential interest as antipsychotics [73], e.g., *drolasetron* [*115956-13-3*] (MDL-73147, Marion Merrell Dow) and *zacopride* [*90182-92-6*] (AHR-11190, American Home Products). Most neurotransmitters are subject to a negative-feedback control by presynaptic autoreceptors that prevent further release by activating K^+ channels and/or inhibiting Ca^{2+} channels. Thus the stimulation of presynaptic receptors induces inhibition of transmitter release. Dopaminergic activity can be reduced not only by postsynaptic receptor antagonism, but also by activation of presynaptic autoreceptors. *Roxindole* [*112192-04-8*] (EMD-49980, E. Merck) and *CGS-15873* [*118929-49-0*] (Ciba-Geigy) [74] mimick this feedback inhibition mechanism for endogenous dopamine; they have an antipsychotic-like pharmacological profile. Other antipsychotics under development include antiaggression drugs such as *eltoprazine* [*98224-03-4*] (DU-28853, Solvay), and the glycine prodrug *milacemide* [*76990-56-2*] (CP-15525, Monsanto), which also has antidepressant and cognitive-enhancing potential. Other compounds act as specific D_2- or D_1-receptor antagonists (for a review, see [75]).

3. Antidepressants

Depression corresponds to major modifications of mood and is characterized by affective disorders that extend outside acceptable limits and impair thinking and perception. Most depressive patients suffer only from unipolar or major depression. In others, depression is associated with intermittent episodes of mania (manic-depressive illness). Major depression is characterized by sadness, despair, self-depreciation, insomnia, anorexia, and also sometimes suicidal behavior. Manic-depressive illness is characterized by dysphoria, irritability, marked insomnia,

impaired judgement, or behavior. Major depression and manic-depressive illness appear to react to different treatments.

Treatment of depression was first based on the sedative effect of opium (laudanum) or on the psychostimulant activity of amphetamines. At the end of the 1950s, KUHN discovered that imipramine, a dibenzazepine derivative first tested for its antipsychotic and antihistaminic properties, was a powerful antidepressive agent in clinical trials [76]. Simultaneously, KLINE successfully used iproniazide, an isoniazide derivative used as a drug against tuberculosis, in the treatment of depression [77]. The discovery of imipramine stimulated the search for related compounds based on the tricyclic dibenzazepine structure. Iproniazide was found to be an irreversible inhibitor of monoamine oxidase, a result which also stimulated further research.

3.1. Tricyclic Antidepressants

Tricyclic antidepressants have been widely used as the drugs of choice for the treatment of major depression. In behavioral studies in rodents, tricyclics depress spontaneous motor activity and impair the acquisition and performance of a conditioned avoidance test. Their best recognized psychopharmacological characteristic corresponds to their interaction with reserpine and tetrabenazine which deplete biogenic amines in the brain. The tests used most frequently to predict the antidepressant effect of a drug are based on the inhibition of sedation, catalepsy, hypothermia, or ptosis in rodents pretreated with reserpine [78]. In electroencephalographic (EEG) studies, low doses induce synchronization which reflects sedative activity; high doses produce a desynchronized EEG and are sometimes able to induce EEG epileptogenic events [79, 80]. Moreover, in sleep–wakefulness studies, these drugs often increase slow-wave sleep and decrease paradoxical sleep.

Under normal conditions neurotransmitters in the synapse are taken up by the cell. Uptake inhibitors bind to specific sites and impede physiological uptake; neurotransmitters (NA, DA, or 5-HT) thus show higher concentrations for longer in the synaptic cleft and induce more intense neuronal activity. In contrast an enhancer of 5-HT uptake increases the physiological uptake of the transmitter and thus decreases its synaptic concentration.

Tricyclic antidepressants potentiate the action of biogenic amines [mainly noradrenaline (NA) and serotonin (5-hydroxytryptamine, 5-HT)] by inhibiting their presynaptic uptake after release which leads to increased concentrations of these neurotransmitters at the synapse. None of the compounds in this series are purely selective inhibitors for NA or 5-HT uptake [81, 82]. For example, desipramine is most active in blocking NA uptake and has a low efficacy on 5-HT uptake, while amitriptyline is equipotent in blocking both NA and 5-HT uptake.

3.1.1. Inhibitors of Noradrenaline and/or Serotonin Uptake

Amitriptyline [*50-48-6*], 10,11-dihydro-5-(3-dimethylaminopropylidene)-5*H*-dibenzo[*a,d*]cycloheptene, $C_{20}H_{23}N$, M_r 277.39, hydrochloride *mp* 196 – 197 °C; this compound can be prepared by reaction of 5-oxo-10,11-dihydro-5*H*-dibenzo[*a,d*]cycloheptene with the Grignard reagent of 3-dimethylaminopropyl chloride followed by dehydration of the resulting tertiary alcohol in a strongly acidic medium [83–85]. Amitriptyline and other tricyclic antidepressants are particularly effective in the treatment of endogeneous depressions, but are less effective in reactive depression.

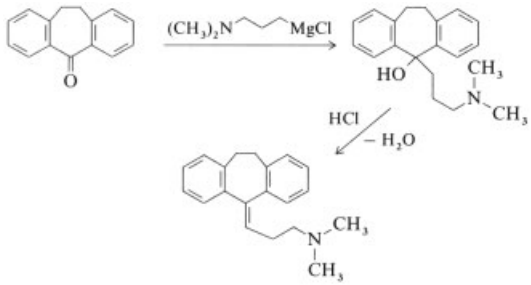

Trade Names. (Hydrochloride) Adepril (Merrell Dow Lepetit), Amitid (Squibb), Amitril (Warner-Chilcott), Endep, Laroxyl (Roche), Lentizol (Warner – Lambert, Parke Davis), Saroten (Lundbeck, Denmark, Tropon, Germany), Sylvemid (Weiner, Germany), Elavil, Tryptanol, Tryptizol (Merck Sharp & Dohme).

Amoxapine [*14028-44-5*], 2-chloro-11-(1-piperazinyc)dibenz[*b,f*][1, 4]oxazepine, $C_{17}H_{16}$-ClN_3O, M_r 313.79, *mp* 175 – 176 °C. Treatment of the starting lactam with phosphorus oxychloride, followed by condensation of the corresponding imidochloride with piperazine affords amoxapine [57].

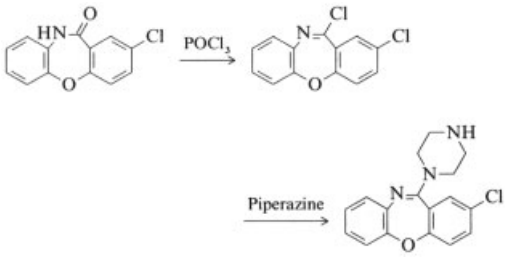

Amoxapine is a tricyclic antidepressant with actions and uses similar to those of amitriptyline [86].

Trade Names. Asendin, Asendis, Defenyl, Demolox, Moxadil, Omnipress (Lederle).

Clomipramine [*303-49-1*], 3-chloro-10,11-dihydro-5-(3-dimethylaminopropyl)-5*H*-dibenz-[*b,f*]azepine, $C_{19}H_{23}ClN_2$, M_r 314.87, *bp* (0.04 kPa) 160 – 170 °C, hydrochloride *mp* 189 –190 °C. The synthesis of clomipramine is as follows [87]:

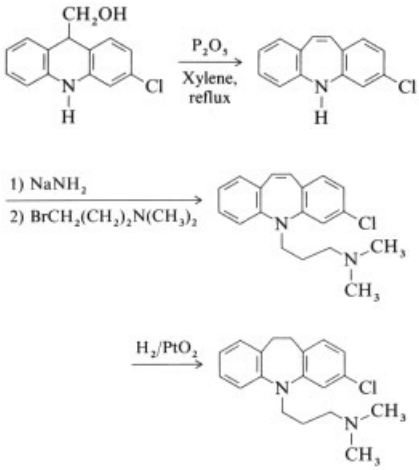

Clomipramine is particulary effective in patients with obsessive – compulsive disorders.

Trade Names. (Hydrochloride) Anafranil (Geigy), Hydiphen (Arzneimittelwerk Dresden, Germany), Maronil (Unipharm, Israel).

Dosulepin [*113-53-1*], Dothiepin, 6,11-dihydro-11-(3-dimethylaminopropylidene)dibenzo-[*b,e*]thiepin, $C_{19}H_{21}NS$, M_r 295.45, *bp* (0.006 kPa) 171 – 172 °C, *mp* 55 – 57 °C, hydrochloride *mp* 218 – 221 °C. The synthesis of dosulepin follows [88]:

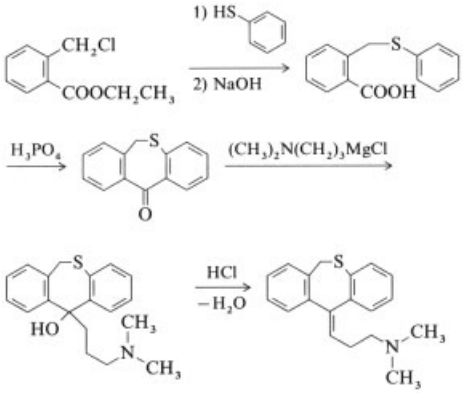

Trade Names. (Hydrochloride) Idom (Kanoldt, Germany), Prothiaden (Boots; Boots Dacour, France; Spofa, Czechoslovakia), Xerenal (Kwizda, Austria).

Doxepin [*1668-19-5*], 6,11-dihydro-11-(3-dimethylaminopropylidene)dibenz[*b,e*]oxepin, $C_{19}H_{21}NO$, M_r 279.37, *bp* (0.026 kPa) 260 – 270 °C, hydrochloride *mp* 188 – 189 °C. This dibenz[*b,e*]oxepin can be considered as an analogue of amitriptyline (replacement of methylene group by an oxygen atom). The synthesis of doxepin is summarized at next side [89].

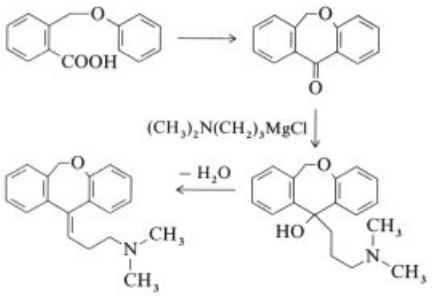

Trade Names. (Hydrochloride) Adapin (Pennwalt, USA), Aponal (Galenus, Germany), Novoxapin (Ester, Spain), Quitaxon (Boehringer Mannheim, Erco Denmark), Sinequan (Pfizer-Roerig).

Imipramine [*50-49-7*], 10,11-dihydro-5-(3-dimethylaminopropyl)-5*H*-dibenz[*b,f*]azepine, $C_{19}H_{24}N_2$, M_r 280.40, *bp* (0.013 kPa) 160 °C, hydrochloride *mp* 174 – 75 °C. Imipramine is prepared by alkylation of 10,11-dihydro-5*H*-dibenz[*b,f*]azepine, which is the tricyclic parent structure for a range of psychopharmaceuticals, with 3-dimethylaminopropyl chloride in the presence of sodamide [90]:

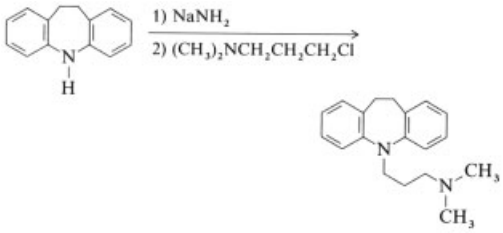

The starting material can be obtained by oxidative dimerization of 2-nitrotoluene to give 1,2-bis-(2-nitrophenyl)ethane which is hydrogenated to give the corresponding diamine. Ring closure takes place on heating with elimination of ammonia to afford the parent dibenzazepine.

Imipramine is a tricyclic antidepressant with a similar mode of action and similar uses as amitriptyline but with less marked sedative properties.

Trade Names. (Hydrochloride) Chrytemin (Fujinaga, Japan), Deprinol (Dumex, Denmark), Efuranol (Pfizer), Feinalmin (Sanko, Japan), Imavate (Robins, USA), Janimine (Abbott), Melipramin (Egis, Hungary), Presamine (USV, USA).

Desipramine [*50-47-5*], 10,11-dihydro-5-[3-(methylaminopropyl)-5*H*-dibenz[*b,f*]azepine, $C_{18}H_{22}N_2$, M_r 266.37, *bp* (0.002 kPa) 172 – 174 °C, hydrochloride *mp* 214 – 218 °C. Treatment of imipramine with ethyl chloroformate followed by alkaline hydrolysis affords desipramine in high yield [91].

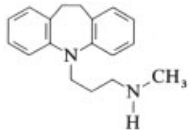

Trade Names. (Hydrochloride) Norpolake (Lakeside, USA), Norpramin (Merrell Dow Pharmaceuticals), Nortimil (Chiesi, Italy), Pertofran, Sertofren (Geigy).

Lofepramine [*23047-25-8*], *N*-methyl-*N*-(4- chlorobenzoylmethyl)-3-(10,11-dihydro-5*H*-di-benzo[*b,f*]azepin-5-yl)propylamine, $C_{26}H_{27}ClN_2O$, M_r 418.97, *mp* 104 – 106 °C, hydrochloride *mp* 152 – 154 °C. Lofepramine is obtained by reaction of desipramine with 4′-chloro-bromoacetophenone [91].

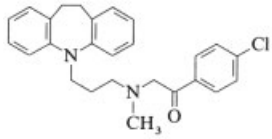

Trade Names. (Hydrochloride) Deftan, Deprimil, Emdalen, Gamanil, Gamonil (E. Merck, Germany), Tymelyt (Leo, Sweden).

Metapramine [*21730-16-5*], 10,11-dihydro-5-methyl-10-(methylamino)-5*H*-dibenz[*b,f*]azepine, $C_{16}H_{18}N_2$, M_r 238.33, hydrochloride *mp* 238 –240 °C. This compound can be prepared in two ways [92]:

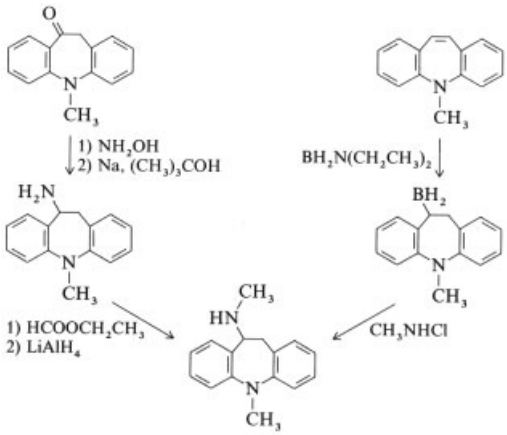

Trade Name. Timaxel (Specia, France).

Nortriptyline [*72-69-5*], 10,11-dihydro-5-(3-methylaminopropylidene)-5*H*-dibenzo[*a,d*]cycloheptene, $C_{19}H_{21}N$, M_r 263.37, hydrochloride *mp* 213 – 215 °C. This compound is prepared as follows [93].

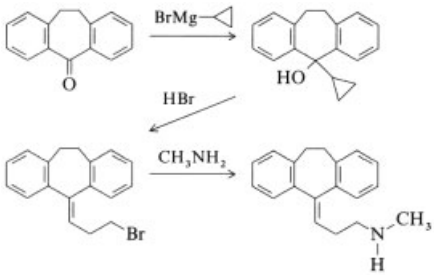

Trade Names. (Hydrochloride) Acetexa (Lilly), Allegron (Dista), Altilev (Squibb), Nortrilen (Lundbeck, Denmark, Tropon, Germany), Pamelor (Lilly, Sandoz-Wander), Sensival (Pharmacia, Sweden), Vividyl (Lilly).

Protriptyline [*438-60-8*], 5-(3-methylaminopropyl)-5*H*-dibenzo[*a,d*]cycloheptene, $C_{19}H_{21}N$, M_r 263.37, hydrochloride *mp* 169 – 171 °C. Protriptyline can be prepared as follows [94]:

Protriptyline has considerably less marked sedative properties than amitriptyline, and may have stimulant effects.

Trade Names. (Hydrochloride) Concordin, Vivactil (Merck Sharp & Dohme), Maximed (Frosst, Germany, Merck Sharp & Dohme), Triptil (Merck-Frosst, Pointe Claire Dorval, Canada, Merck Sharp & Dohme).

Butriptyline [*35941-65-2*], 10,11-dihydro-5-(3-dimethylamino-2-methylpropyl)-5*H*-dibenzo-[*a,d*]cycloheptene, $C_{21}H_{27}N$, M_r 293.46, *bp* (0.133 kPa) 180 – 185 °C, hydrochloride *mp* 188 – 190 °C. The intermediate olefinic compound is prepared analogously to the amitriptyline parent olefin (see page 18), and finally reduced with $LiAlH_4$ [95].

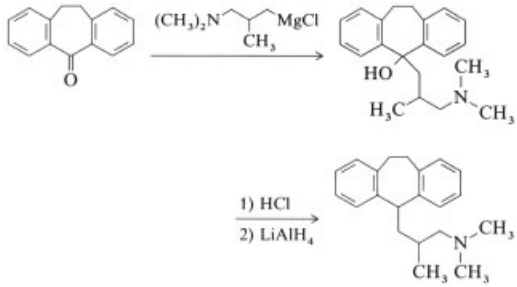

Trade Names. Centrelyse, Evadene (Ayerst), Evadyne (Ayerst), Evasidol (Arcana, Austria).

Quinupramine [*31721-17-2*],10,11-dihydro-5-(3-quinuclidinyl)-5*H*-dibenz[*b,f*]azepine, $C_{21}H_{24}N_2$, M_r 304.43, hydrochloride *mp* 150 °C, is prepared as follows [96]:

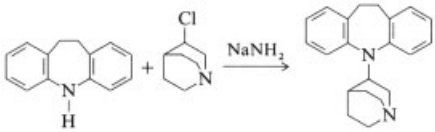

Trade Names. Kevopril (Rhône-Poulenc), Kinupril (Pharmuka, France), Quinuprine (Rhone, Spain).

3.1.2. Enhancers of Uptake

Tianeptine [*66981-73-5*], 7-[(3-chloro-6,11-dihydro-6-methyldibenzo[*c,f*][1, 2]thiazepin-11-yl)amino]heptanoic acid *S,S*-dioxide, $C_{21}H_{25}ClN_2O_4S$, M_r 436.95, sodium salt *mp* 180 °C. Tianeptine is prepared as follows [97]:

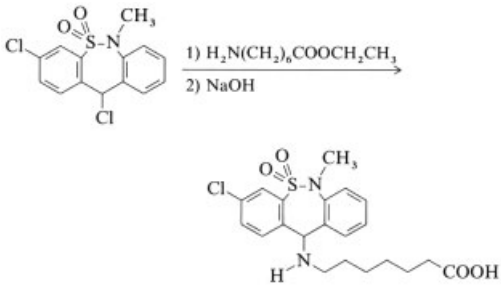

Tianeptine is used in the treatment of reactive and neurotic depression, as well as in the treatment of anxious states.

Trade Name. Stablon (Ardix, France).

3.1.3. Inhibitors of Dopamine Uptake

Amineptine [*57574-09-1*], 7-[(10,11-dihydro-5*H*-dibenzo[*a,d*]cyclohepten-5-yl)amino]heptanoic acid, $C_{22}H_{27}NO_2$, M_r 337.47, hydrochloride *mp* 226 – 230 °C, is prepared as follows [98]:

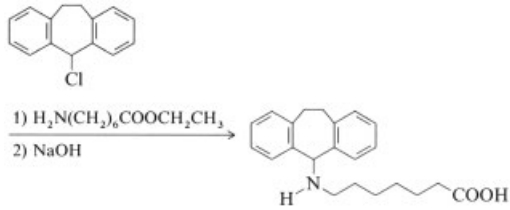

Trade Names. Directim, Neolior, Viaspera (Servier, France), Maneon (Poli, Italy), Provector (Servier, Euthé-rapie, France).

3.1.4. Other Tricyclic Antidepressants

Carpipramine [*5942-95-0*], 5-{3-[4-carbamoyl-4-piperidino)piperidino]propyl}-10,11-dihydro-5*H*-dibenz[*b,f*]azepine, also has neuroleptic properties (see Section 2.5.1).

Demexiptiline [*24701-51-7*], 5*H*-dibenzo[*a,d*]cyclohepten-5-one, *O*-[2-(methylamino)-ethyl]oxime, $C_{18}H_{18}N_2O$, M_r 278.35, hydrochloride *mp* 232 – 233 °C. Demexiptiline is prepared as follows [99]:

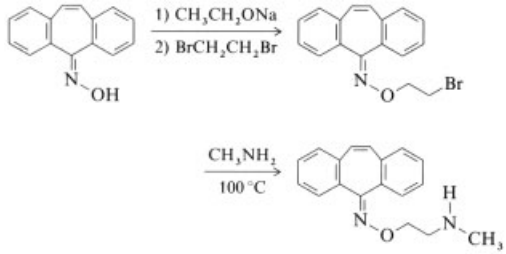

Trade Name. Déparon (Aron, France).

Dibenzepin [*4498-32-2*], 10,11-dihydro-10-(2-dimethylaminoethyl)-5-methyl-11-oxodibenzo[*b,e*][1, 4]diazepine, $C_{18}H_{21}N_3O$, M_r 295.37 *mp* 116 – 117 °C, hydrochloride *mp* 238 °C. Dibenzepin is prepared as follows [100]:

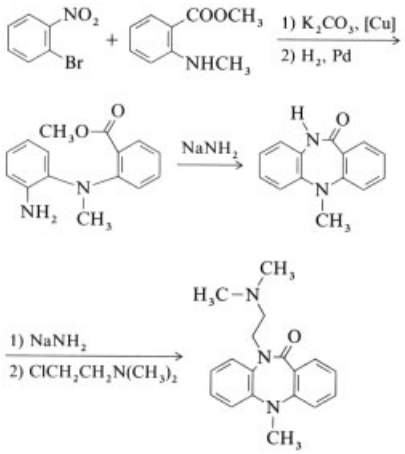

Trade Names. Ecatril, Neodit (Sandoz-Wander).

Noxiptilin [*3362-45-6*], 10,11-dihydro-5*H*-dibenzo[*a,d*]cyclohepten-5-one, *O*-[2-(dimethylamino)ethyl]oxime, $C_{19}H_{22}N_2O$, M_r 294.4, *bp* (0.006 kPa) 160 – 164 °C, hydrochloride *mp* 185 –187 °C. Noxiptilin is prepared as follows [101]:

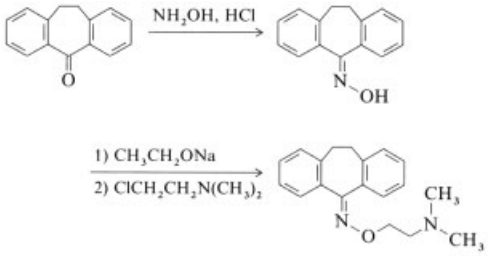

Trade Names. Agedal (Bayer), Nogedal (Théraplix, France), Sipcar (Bernabo, Argentina).

Opipramol [*315-72-0*], 4-[3-(5*H*-dibenz[*b,f*]- azepin-5-yl)propyl]-1-(2-hydroxyethyl)piperazine, $C_{23}H_{29}N_3O$, M_r 363.49, *mp* 100 – 101 °C, dihydrochloride *mp* 228 – 230 °C. Opipramol is prepared as follows [102]:

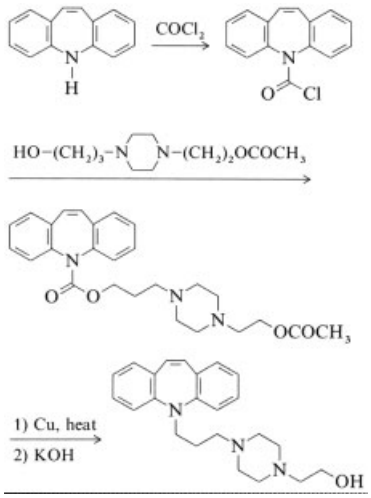

Trade Names. Deprenil (Yurtoglu, Turkey), Dinsidon, Ensidon, Insidon, Nisidana (Geigy).

Propizepine [*10321-12-7*], 6,11-dihydro-6-[2-(dimethylamino)propyl-5*H*-pyrido[2,3-*b*][1,5]-benzodiazepin-5-one, $C_{17}H_{20}N_4O$, M_r 296.38, *mp* 122 °C, hydrochloride, *mp* 235 °C [103]

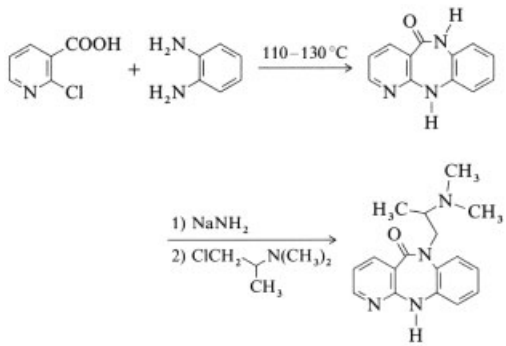

Trade Names. Depressin, Vagran (UPSA, France).

Trimipramine [*739-71-9*], 10,11-dihydro-5-[3-(dimethylamino)-2-methylpropyl]-5*H*-dibenz[*b,f*]azepine, $C_{20}H_{26}N_2$, M_r 294.42, *mp* 45 °C, maleate salt *mp* 142 °C. The shortest synthesis of trimipramine is shown below [104]:

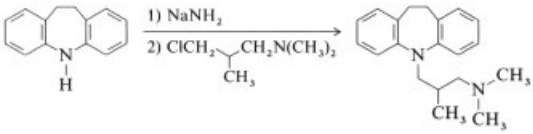

Trade Names. Surmontil (Specia, May & Baker, Farmitalia Carlo Erba), Stangyl (Rhône-Poulenc).

3.2. Tetracyclic Antidepressants

Maprotiline and mianserin constitute a second generation of antidepressants that are chemically characterized by a tetracyclic structure. In animal models, *maprotiline*, a dibenzobicyclooctadiene derivative, has a sedative and tranquilizing activity [105]. Clinical studies have demonstrated its activity to be equivalent to that of the tricyclic antidepressants [106], but with less cardiac side effects. Maprotiline has a potent activity in blocking NA uptake without effect on 5-HT uptake [105].

Mianserin is a tetracyclic piperazinoazepine first developed as an antihistamine and anti-5-HT drug. It has antidepressant activity in humans and induces drowsiness during the first weeks of treatment; its possible cardiotoxic effects are less marked than those observed with tricyclic antidepressants [107]. Mianserin is a potent inhibitor of H_2-receptor-mediated adenylate cyclase formation in brain [108] and has a presynaptic α-adrenoreceptor blocking effect [109]. Its mechanism of action appears to involve release of NA via cortical α_2-adrenergic autoreceptor blockade.

Maprotiline [*10262-69-8*], 1-(3-methylaminopropyl)dibenzo[*b,e*]bicyclo[2.2.2]octadiene, $C_{20}H_{23}N$, M_r 277.41, *mp* 92 – 94 °C, hydrochloride *mp* 230 – 232 °C. Maprotiline is prepared as follows [110]:

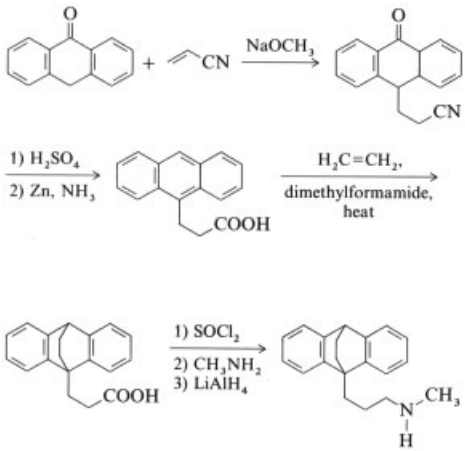

Maprotiline differs structurally from tricyclics due to the presence of an additional alicyclic bridge. It resembles amitriptyline with regard to both efficacy and side effects. It may have proconvulsant properties.

Trade Name. Ludiomil (Ciba).

Mianserin [*24219-97-4*], 2-methyl-1,2,3,4,10, 14*b*-hexahydro-2*H*-pyrazino[1,2-*f*]morphanthridine, $C_{18}H_{20}N_2$, M_r 264.37, hydrochloride *mp* 282 – 284 °C. Mianserin is prepared as follows [111]:

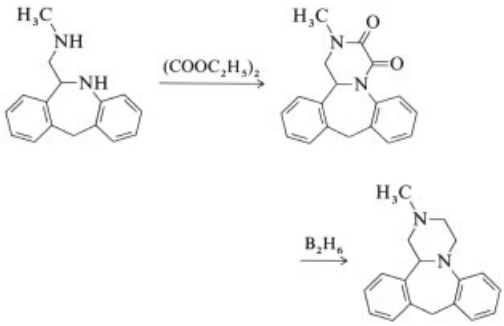

Mianserin does not display significant antimuscarinic properties, but has a marked sedative action.

Trade Names. Athymil, Bolvidon, Lantanon, Tetramide, Tolvin, Tolvon (Organon).

3.3. Monoamine Oxidase Inhibitors (MAOI)

Low single doses of MAOI produce only minor or negligible changes in animal behavioral pharmacology even when a significant increase in biogenic amines is measured in the brain. Nevertheless, combined with other agents such as reserpine or tetrabenazine (which induce biogenic amine depletion), they tend to produce excitement rather than sedation in rodents. MAOI have only slight effects on EEG recordings, with the exception of tranylcypromine which is a psychostimulant. They are amongst the most effective drugs in suppressing paradoxical (rapid eye movement, REM) sleep in humans and are used in narcolepsy for this reason.

Monoamine oxidase (MAO) is a complex enzyme which is responsible for the metabolic degradation of biogenic amines. Drugs which inhibit this enzyme system therefore induce an increase in the concentration of biogenic amines such as adrenaline, NA and 5-HT which explains their antidepressant effect [112]. MAO occurs in two forms that differ in their preferences for substrates and sensitivities to selective inhibitors: MAO-A preferentially acts on NA and 5-HT in humans and MAO-B acts on phenethylamine, methylhistamine, and dopamine (DA). Irreversible MAOIs bind covalently to the enzyme, involving noncompetitive inhibition. Recovery of MAO activity is only achieved by arrival of newly synthesized enzyme. Reversible MAOIs form a less stable bond with the enzyme and they can be removed from their binding site, MAO activity therefore recovers.

3.3.1. Hydrazine Derivatives

Benmoxine [*7654-03-7*], benzoic acid 2-(1-phenylethyl)hyrazide, $C_{15}H_{16}N_2O$, M_r 240.29, *mp* 93 – 94 °C. Hydrogenation of the hydrazone obtained by condensation of acetophenone with benzoylhydrazine affords benmoxine [113].

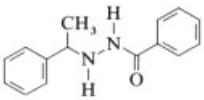

Trade Names. Nerusil (Iquinosa, Spain); Neuralex (Millot, France).

Iproclozide [*3544-35-2*], (4-chlorophenoxy)-acetic acid 2-isopropylhydrazide, $C_{11}H_{15}ClN_2O_2$, M_r 242.72, *mp* 93 – 94 °C. Iproclozide is prepared as follows [114]:

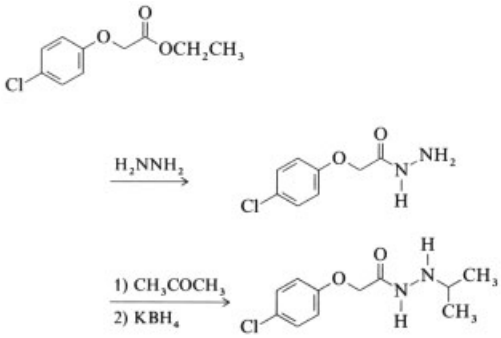

Trade Name. Sursum (Merrell Dow Pharmaceuticals, Ibsa, Switzerland).

Iproniazid [*54-92-2*], isonicotinic2-isopropylhydrazide, $C_9H_{13}N_3O$, M_r 179.22, *mp* 112.5 –113.5 °C, can be prepared using the synthetic pathway described for iproclozide [115]. An alternative synthesis is described in [116].

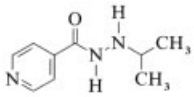

Trade Names. Ipronid (Afi, Norway), Marsilid (Roche).

Isocarboxazid [*59-63-2*], 5-methyl-3-isoxazolecarboxylic acid 2-benzylhydrazide, $C_{12}H_{13}N_3O_2$, M_r 231.25, *mp* 105 – 106 °C. Isocarboxazid is prepared according to the following scheme [117]:

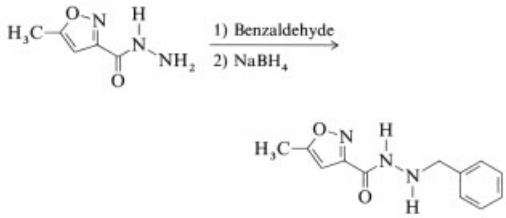

Trade Names. Marplan, Marplon (Roche).

Nialamide [*51-12-7*], isonicotinic acid 2-(2-benzylcarbamoylethyl)hydrazide, $C_{16}H_{18}N_4O_2$, M_r 298.34, *mp* 152 – 153 °C [118].

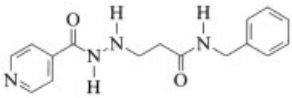

Trade Names. Espril (Saba, Turkey), Niamide (Pfizer-Roerig), Nuredal (Egis, Hungary), Surgex (Firma, Italy).

Phenelzine [*51-71-8*], phenethylhydrazine, $C_8H_{12}N_2$, M_r 136.19, *bp* (0.013 kPa) 74 °C, hydrochloride *mp* 174 °C, for preparation see [119].

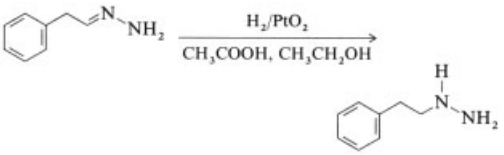

Trade Names. (Acid sulfate) Nardelzine (Substancia, the Netherlands, Substancia-Parke Davis, Spain, Warner-Chilcott), Nardil (Warner-Chilcott, Parke-Davis).

3.3.2. Nonhydrazine Derivatives

Caroxazone [*18464-39-6*], 2-oxo-1,3-benzoxazine-3(4H)-acetamide, $C_{10}H_{10}N_2O_3$, M_r 206.20, *mp* 203 – 205 °C. Caroxazone is prepared as follows [120]:

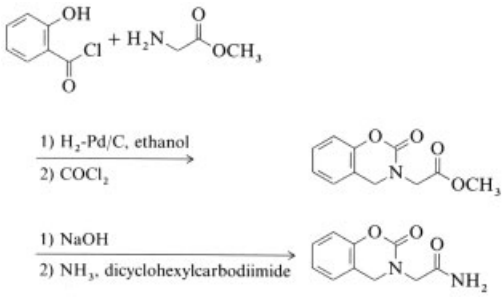

Trade Name. Timostenil(Farmitalia Carlo Erba, Italy).

Milacemide [*76990-56-2*], 2-pentylaminoacetamide, $C_7H_{16}N_2O$, M_r 144.14, *mp* 46 °C, is prepared by condensation of pentylamine with chloroacetamide [121].

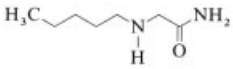

Trade Names. Glyzac, Glyzan (Searle).

Pargyline [*555-57-7*], *N*-methyl-*N*-2-propynylbenzylamine, $C_{11}H_{13}N$, M_r 159.22, hydrochloride *mp* 154 – 155 °C, is prepared by condensation of propynyl bromide with benzylmethylamine [122].

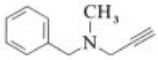

Pargyline is also an antihypertensive agent.

Trade Names. Eudatine, Eutonyl (Abbott).

Selegiline [*14611-51-9*], (*R*)-*N*,α-dimethyl-*N*-2-propynylphenethylamine, $C_{13}H_{17}N$, M_r 187.14, *bp* (0.08 kPa) 92 – 93 °C, hydrochloride *mp* 141.0 – 141.5 °C. Selegiline can be prepared according to various schemes, the shortest is shown below [123]:

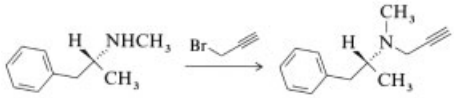

Selegiline is used in the treatment of Parkinson's disease.

Trade Names. Eldeprine (Unicet, France), Jumex (Chinoin, Hungary, Chiesi, Italy), Jumexal (Labatec, Switzerland), Movergan (Asta-Degussa, Germany, Astra).

Tranylcypromine [*155-09-9*], *trans*-(±)-2-phenylcyclopropylamine, $C_9H_{11}N$, M_r 133.19, *bp* (0.20 – 0.213 kPa) 79 – 80 °C, hydrochloride *mp* 164 – 166 °C; for preparation, see [124].

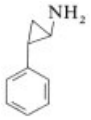

Tranylcypromine is used for the treatment of adult outpatients with reactive depression.

Trade Names. (Sulfate salt) Parnate (Smith Kline & French, Röhm Pharma, Germany), Transapin (Medexport, Independent Estates Community), Tylciprine (Théraplix, France).

3.3.3. Reversible MAO-A Inhibitors

Metralindole [*54188-38-4*], 3-methyl-8-methoxy-3*H*-1,2,5,6-tetrahydropyrazino[1,2,3-*ab*]-*β*-carboline, $C_{15}H_{17}N_3O$, M_r 255.14, *mp* 164 –166 °C. Metralindole is prepared as follows [125]:

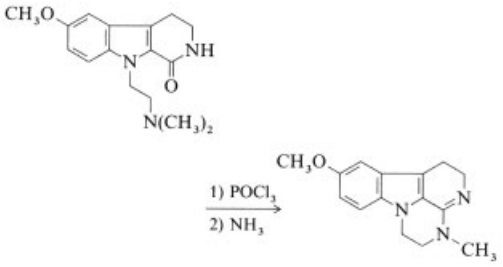

Trade Name. (Sulfate salt) Incazan (Independent Estates Community).

Moclobemide [*71320-77-9*], 4-chloro-*N*-(2-morpholinoethyl)benzamide, $C_{13}H_{17}ClN_2O_2$, M_r 268.7, *mp* 137 °C. Various syntheses are described for moclobemide, the simplest and shortest being the reaction of 4-chlorobenzoyl chloride with *N*-(2-aminoethyl)morpholine in pyridine [126].

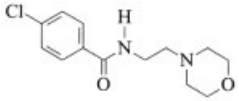

Trade Name. Auroxix (Roche, Switzerland).

Pirlindole [*60762-57-4*], 8-methyl-1,2,3,4-tetrahydro-1,10-trimethylenepyrazino[1,2-a]indole, $C_{15}H_{18}N_2$, M_r 226.14, hydrochloride, *mp* 260 –261 °C. It can be prepared in two ways (Fig. 2) [127].

Trade Names. Lifril (Casen Fisons, Spain), Pyrazidol (Independent Estates Community).

Toloxatone [*29218-27-7*], 5-(hydroxy-methyl)-3-m-tolyl-2-oxazolidinone, $C_{11}H_{13}NO_3$, M_r 207.2, *mp* 76 °C is prepared by the reaction of 3-methylaniline with glycidol followed by condensation of the intermediate (an aminoalcohol) with diethyl carbonate [128].

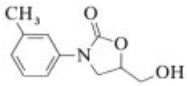

Trade Name. Humoryl (Delalande, France).

3.4. Inhibitors of Serotonin Uptake

During the 1980s new drugs were developed that differed from tricyclics because they were highly selective serotonin (5-HT) uptake inhibitors, (e.g., fluvoxamine, paroxetine, fluoxetine, or sertraline). They proved to be effective antidepressants in humans [129].

Fluoxetine enhances behavioral patterns induced in rodents by pharmacological facilitation of 5-HT transmission [130]. It potentiates 5-HT-induced head twitches in mice pretreated with

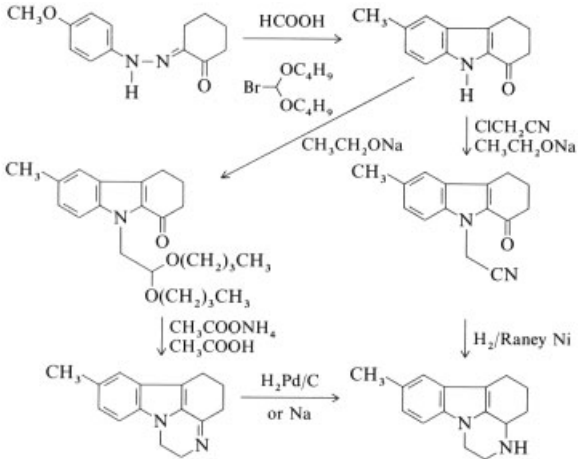

Figure 2. Synthesis of pirlindole

pargyline, and has also been used as a tool in assessing 5-HT transmission in several behavioral paradigms (self-stimulation test). Nevertheless, in the behavioral despair test, where rats are immobile after a short period of swimming in a restricted place, fluoxetine does not significantly reduce the duration of immobility as is the case following administration of antidepressants that inhibit biogenic amine uptake [106, 131]. Fluoxetine and related compounds are specific 5-HT uptake inhibitors. They have no affinity for receptors such as cholinergic (muscarinic), α-adrenergic, or histaminergic, and they possess lower, cardiovascular and sedative side effects than the tricyclic antidepressants. Trazodone is an inhibitor of synaptosomal 5-HT uptake but is also a 5-HT antagonist (see Section 3.6) [132].

The compounds described below are specific inhibitors of 5-HT uptake. An exception is oxaflozane, which also inhibits NA uptake.

Citalopram [*59729-33-8*], 1,3-dihydro-1-[3-(dimethylamino)propyl]-1-(4-fluorophenyl)-5-iso-benzofurancarbonitrile, $C_{20}H_{21}FN_2O$, M_r 324.40, *bp* (0.004 kPa) 175 – 181 °C, hydrobromide *mp* 182 – 183 °C. Citalopram is prepared as follows [133]:

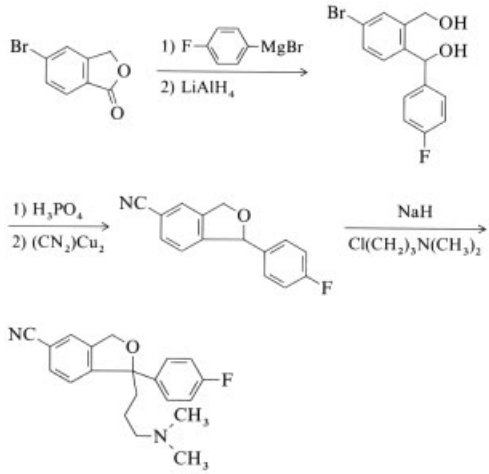

Citalopram is also used as a geriatric drug [134].

Trade Names. Cipramil (Lundbeck, U.K.), Seropam (Zeria).

Femoxetine [*59859-58-4*], (3*R-trans*)-3-[4-(methoxy)phenoxymethyl]-1-methyl-4-phenylpiperidine, $C_{20}H_{25}NO_2$, M_r 311.43. This compound is prepared as follows [135]:

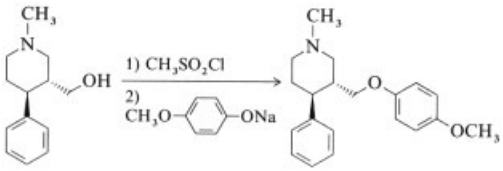

This 5-HT uptake inhibitor is as effective as amitriptyline in the treatment of depression, but displays fewer side effects.

Trade Name. Malexil (Ferrosan, Bristol-Myers, Knoll).

Fluoxetine [*54910-89-3*], *N*-methyl-3-[4-trifluoromethylphenoxy-3-phenylpropylamine, $C_{17}H_{18}F_3NO$, M_r 309.33, oxalate *mp* 179–182 °C (decomp.). Fluoxetine is prepared as follows [136]:

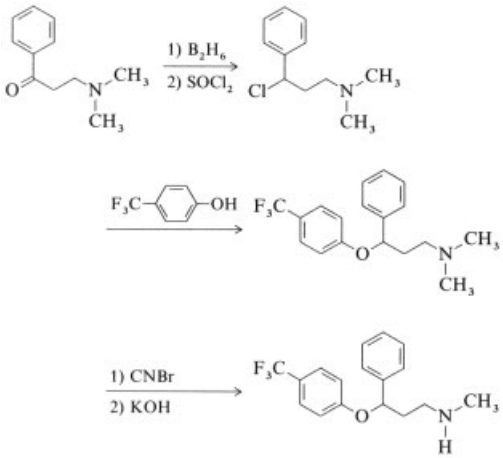

This 5-HT uptake inhibitor is reported to display fewer antimuscarinic side effects than tricyclic antidepressants.

Trade Names. Adofen (Ferrer, Spain), Fluoxeren (Menerini, Italy), Fontex (Lilly), Reneuron (Juste, Spain), Prozac (Lilly).

Fluvoxamine [*54739-18-3*], 5-methoxy-4′-(tri-fluoromethyl)valerophenone(E)-O-(2-aminoethyl)-oxime, $C_{15}H_{21}F_3N_2O_2$, M_r 318.35, maleate *mp* 120–121.5 °C. Fluvoxamine is prepared as follows [137]:

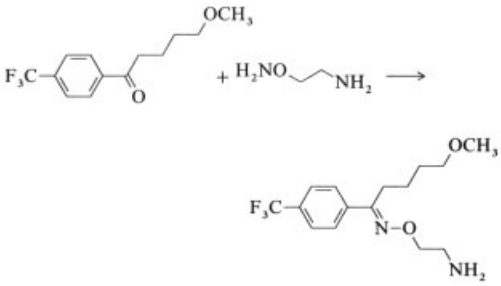

The compound has a good selectivity for 5-HT uptake over noradrenaline uptake.

Trade Names. Avoxin (Krka, Yugoslavia), Dumirox (Kalifarma, Spain), Faverin, Fevarin, Floxyfral, Floxytral, Myroxim (Duphar).

Oxaflozane [*26629-87-8*], 4-isopropyl-2-[3-(tri-fluoromethyl)-phenyl]morpholine, $C_{14}H_{18}F_3NO$, M_r 273.29, *bp* (0.0006 kPa) 52 °C, hydrochloride *mp* 164 °C. Oxaflozane is prepared as follows [138]:

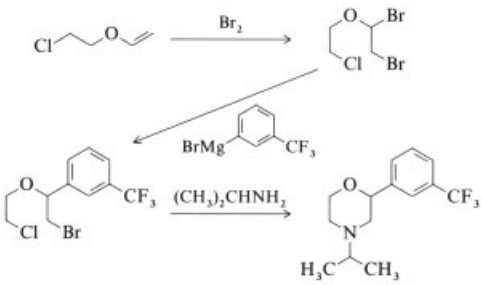

Trade Name. Conflictan (LTM Sarbach, France).

Paroxetine [*61869-08-7*], (−)-*trans*-4-(4-fluorophenyl)-3-{[3,4-(methylenedioxy)phenoxy]-methyl}-piperidine, $C_{19}H_{20}FNO_3$, M_r 329.37, hydrochloride *mp* 118 °C. This piperidine derivative is prepared by condensation of 4-(4-fluorophenyl)-3-(hydroxymethyl)piperidine with 3,4-methylenedioxyphenol in the presence of dicyclohexylcarbodiimide [139].

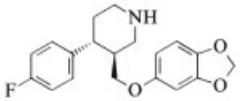

Trade Names. Aropax (Beecham Laboratories, USA), Seroxat (Ferrosan).

Sertraline [*79617-96-2*], (1*S*,4*S*)-4-(3,4-dichlorophenyl)-1,2,3,4-tetrahydro-*N*-methyl-1-naphthylamine, $C_{17}H_{17}Cl_2N$, M_r 306.23, hydrochloride, *mp* 243 – 245 °C. For the synthesis of sertraline, see [140].

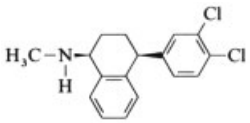

Trade Name. Lustral (Invicta and Pfizer, UK).

3.5. Lithium

Despite the discovery in 1949 of the value of lithium salts, their use as a therapy for mania only became current at the end of the 1950s in Europe, and only in the early 1970s in the United States. Lithium salts have proven to be active in mania and to reduce the frequency of attacks of manic depressive illness [141, 142]. In experimental psychopharmacology lithium salts only show weak activity. A sedative effect may be seen at high doses, but is less evident than with antipsychotics. Lithium salts do not antagonize stereotypies induced by amphetamine or apomorphine (tests for neuroleptic activity) nor the effects of reserpine or tetrabenazine (tests for antidepressant activity). In manic patients, lithium salts improve slow-wave sleep but depress REM sleep; lithium induces slow-wave activity, sometimes with superimposed fast frequencies [79].

The mechanism of action of lithium salts is poorly understood. Lithium acts on biological membranes and reduces excitability at therapeutic concentrations (1 m Eq/L). Lithium has similar ionic properties to sodium but is not the ideal substrate for the transmembrane sodium pump and thus cannot maintain membrane potentials. This leads to a decrease in excitability, which could correspond to the reduction of the psychic reactivity of the patient. Nevertheless biochemical studies have demonstrated that lithium (1 – 10 m Eq/L) inhibits release of NA and DA but not that of 5-HT [143].

Second messengers, such as inositol phosphates, are released by hydrolysis of membrane phosphatidylinositides after activation of the neurotransmitter receptors; lithium can increase this release. Depletion of phosphatidyinositides reduces neuron reactivity to muscarinic, α, – adrenergic, or other stimuli where phosphatidylinositol is the second messenger. This suggests that lithium could preferentially act on hyperactive cells which induce the manic depressive illness [144] and the noradrenergic system which abnormally increases brain electrical activity.

Finally, studies have shown that lithium can interact with other antidepressants and enhance their efficacy by acting on 5-HT neurotransmission. Studies in rats and cats have shown that tryptophan concentration and uptake are increased after short- or long-term lithium treatment. Short-term treatment with lithium has no effect on basal 5-HT release in the hippocampus or frontal cortex. Studies with lithium salts on 5-HT function indicate that they act at various levels, including reuptake, release, synthesis, storage, or receptor interactions; many of these effects are time, region, or species dependent.

Trade Names. (Carbonate) Teralithe (Rhône-Poulenc Rorer, France), (gluconate) Neurolithium (Labcatal, France).

3.6. Miscellaneous Antidepressants

Amfebutamone [*34911-55-2*], 2-(*tert*-butylamino)-3′-chloropropiophenone, $C_{13}H_{18}ClNO$, M_r 293.74, *bp* (0.0006 kPa) 52 °C, hydrochloride *mp* 233 – 234 °C. Amfebutamone can be prepared as follows [145]:

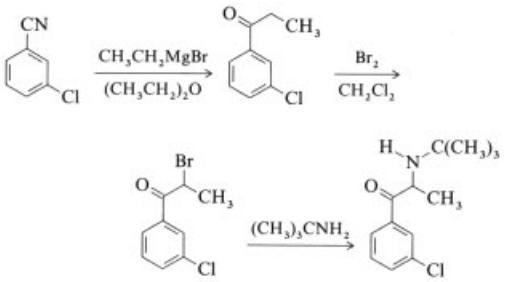

Amfebutamone is an inhibitor of dopamine uptake.

Trade Names. Wellbatrin, Wellbutrin (Burroughs Wellcome).

Medifoxamine [*32359-34-5*], *N,N*-dimethyl-2,2-diphenoxyethylamine, $C_{16}H_{19}NO_2$, M_r 257.33, fumarate *mp* 128.5 °C. Medifoxamine is prepared as follows [146]:

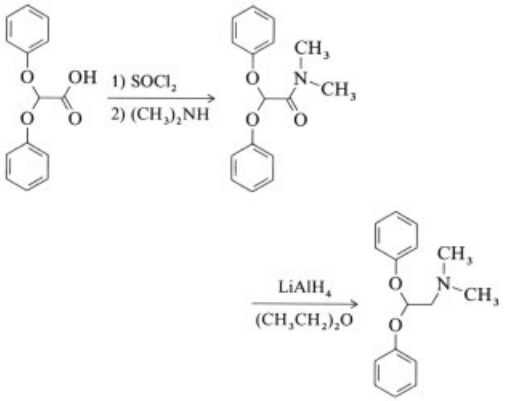

The mechanism of action of medifoxamine is analogous to that of amfebutamone.

Trade Names. Clédial (Anphar-Rolland, France), Gerdaxyl (Gerda, France; Promesa, Spain).

Trazodone [*19794-93-5*], 2-{3-[4-(3-chlorophenyl)-1-piperazinyl]propyl}-1,2,4-triazolo[4,3-*a*]pyridin-3(2*H*)-one, $C_{19}H_{22}ClN_5O$, M_r 371.88, *mp* 86 – 87 °C, hydrochloride *mp* 223 °C. The bicyclic starting material is prepared by condensation of 2-chloropyridine with semicarbazide [147]:

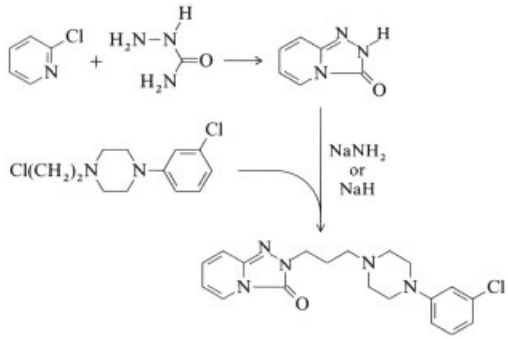

The triazolopyridine derivative trazodone is special because it is not active in classical animal models, such as reserpine antagonism [148], but potentiates the behavioral changes induced by the 5-HT precursor 5-hydroxytryptophan. Nevertheless its efficacy as an antidepressant drug has been proved in clinical trials, with fewer side effects (e.g., cardiotoxicity) and a larger safety margin than the tricyclics. Drowsiness is, however, common [106]. Trazodone is a selective inhibitor of central 5-HT uptake and decreases peripheral 5-HT uptake. It also increases the turnover of brain dopamine.

Trade Names. Beneficat (Nemi, Argentina), Depyrel (Abic, Israel), Desyrel (Angelini, Italy, Bristol-Meyers, Mead-Johnson), Molipaxin (Roussel), Pragmazone (UPSA, France), Taxagon (Rhodia, Argentina), Tombran (Boehringer, Ingelheim), Trazodil (Unipharm, Israel).

Viloxazine [*46817-91-8*], 2-[(2-ethoxyphenoxy)methyl]tetrahydro-1,4-oxazine, $C_{13}H_{19}NO_3$, M_r 237.3, hydrochloride, *mp* 185 – 186 °C. For preparation, see [149].

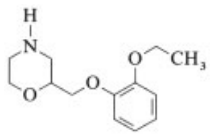

This derivative is devoid of cholinergic-type side effects and is as potent as the standard tricyclic antidepressants. Viloxazine is particularly useful for treating depression in the elderly.

Trade Names. Catarol, Vivalan, Vivarint (ICI).

3.7. Antidepressants Under Development

Adinazolam [*37115-32-5*] is a triazolobenzodiazepine under development (Upjohn) as an antidepressant and anxiolytic (preregistration). It has a rapid onset of action with minimal anticholinergic side effects [150]. *Brofaromine* [*63638-91-5*], (a benzofuranylpiperidine derivative, (Ciba-Geigy), is a selective, competitive, reversible MAO-A inhibitor (phase III clinical trials). Side effects were less frequent than with imipramine or tranylcypromine [151]. *Cericlamine* [*112922-55-1*] (Jouveinal), a potent and selective monocyclic 5-HT uptake inhibitor, is claimed to exhibit a better safety profile than other new-generation antidepressants. Chemically related to buspirone, *gepirone* [*83928-76-1*] is a partial 5-HT-1 A agonist under development by Bristol-Myers Squibb [152]. *Milnacipran* [*92623-85-3*] is a cyclopropanecarboxamide derivative which behaves as a 5-HT and NA uptake inhibitor (preregistration, Pierre Fabre). Wyeth-Ayerst (American Home Products) has developed *venlafaxine* [*99300-78-4*], which has a binding profile identical with that of milnacipran (phase III clinical trials). *Mirtazapine* [*85650-52-8*] (Organon), an analog of mianserin, blocks presynaptic α-receptors (preregistration). EEG studies on humans have confirmed its antidepressant potential [153]. The nontricyclic triazolone derivative *nefazodone* [*83366-66-9*] is an antidepressant under development by Bristol-Myers Squibb (preregistration). Farmitalia Carlo Erba (Erbamont) is developing *reboxetine* [*71620-89-8*], which inhibits NA reuptake and blocks α_2-receptors (phase III clinical trials). *Ritanserin* [*87051-43-2*], a thiazolopyrimidine is a selective 5-HT_2 antagonist. The compound is under development by Janssen (phase III clinical trials) and is effective in mild to moderate depression, possibly associated with anxiety. Inhibitors of brain type II phosphodiesterase, such as *rolipram* [*61413-54-5*] (Schering AG), also have antidepressant activity. They possibly act by enhancing the second messenger concentrations involved in β-adrenoreceptor activation [154]. Rolipram was as effective as amitriptyline in treating patients with manic-depressive psychosis [155].

Efforts are continuing to elucidate the mechanism of action of antidepressants, and to develop new behavioral models with better predictability. Potentially interesting are compounds with combined properties (e.g., MAO inhibitor with uptake blocking activity).

4. Anxiolytics

Anxiety can be considered as a normal psychophysiological function which, in certain situations, may present a pathological feature due to its intensity and the excessive responses that it produces. Anxiety may be acute or chronic, and may take many forms such as generalized anxiety, phobias, anguish, or panic attacks. The aim of the anxiolytics is to treat the psychological and physiological effects of anxiety.

Alcohol and opium were probably the first substances used in the ancient times for their anxiolytic and hypnosedative properties, despite their undesirable secondary effects (habituation and dependence). It was not until the nineteenth century that specific synthetic products became available for the control of sleep and anxiety, notably bromine, with its frequent toxic effects (bromism), and chloral. At the beginning of the twentieth century, an important milestone was the discovery of the hypnosedative *barbiturates*.

These derivatives of barbituric acid have since been extensively used for the treatment of insomnia (→ Hypnotics, Chap. 5.1.) and, at lower doses, for anxiety. Some members of this family have potent sedative properties and are still employed as general anesthetics (→ Anesthetics, General). Minor modifications of the barbiturate structure gave the *piperidinedione derivatives* which possess powerful sedative properties (→ Hypnotics, Chap. 5.2.). A further step was made in the mid 1950s with the clinical use of meprobamate, a *propanediol derivative*, for the specific treatment of anxiety. Up until this time, the treatment of sleep problems and anxiety were often confused. Chlordiazepoxide, the first *benzodiazepine* (BZD), was discovered in 1957 and led to a major advance in the treatment of anxiety disorders. The BZDs, which possess power-

ful anxiolytic and sedative properties (see also → Hypnotics, Chap. 5.4.) have superseded the barbiturates, thanks to their greater safety margins and their lower toxicity [156, 157]. Some benzodiazepines (e.g., chlordiazepoxide, clorazepate, or alprazolam) are almost exclusively prescribed as anxiolytics while others (e.g., nitrazepam, flunitrazepam, and triazolam) are exploited for their sedative hypnotic effects. Midazolam is an exception, being used in anesthesiology to induce or maintain narcosis, as premedication before anaesthesia, or during painful or disagreeable procedures which do not merit general anesthesia (endoscopy, biopsies). The consumption of BZDs as anxiolytics or sedatives continued to grow until the end of the 1970s when tolerance and dependence phenomena were reported after prolonged administration. Consumption is now judged to be excessive.

The discovery of compounds with chemical structures different from those of the BZDs, but with anxiolytic or hypnotic potential, was a further important step forward. Anxiolytics normally possess both a sedative and hypnotic action, the latter may be considered as an undesirable secondary effect, especially if diurnal somnolence results. More recent progress offers clinicians anxiolytics without sedative activity. Examples are the BZDs (bretazenil) or the cyclopyrrolones (RP 59037) [158]. Buspirone (an azaspirodecanedione) is a case apart [159] and is already available as a nonsedative anxiolytic (see page 41).

Animal tests for anxiety can be classified in three groups:

1) Tests based on conflict or conditioned fear: since the first experiments of GELLER and SEIFTER in 1960 [160], anxiolytics have been administered to animals to restore certain behavioral responses previously suppressed by punishment
2) Tests based on anxiety generated by novel environments or situations: the social interaction test and the elevated plus-maze test in which BZDs are active as anxiolytic compounds
3) Tests in which anxiety is induced by chemical agents: BZDs antagonize the anxiogenic effects of drugs such as β- carbolines

4.1. Benzodiazepines

Since the early 1960s the study of BZDs has revealed five main pharmacological properties that can be considered characteristic of these anxiolytic – hypnotic drugs: anticonvulsant, myorelaxant, antiaggressive, anxiolytic, and sedative – hypnotic effects [156]. The potency of these pharmacological effects depends on the pharmacokinetics and on the receptor affinity of the tested compound.

Behavioral, biochemical, and electrophysiological studies indicate that these tranquilizing and sedative – hypnotic substances interact with neurotransmission mediated by γ-aminobutyric acid (GABA). BZDs increase the amplitude of the cat dorsal root potential which is mediated by GABA. This increase in amplitude reflects the intensified presynaptic inhibition in the spinal cord [161, 162]. Neurobiochemical research in the domain of the sedative hypnotics has been stimulated by the discovery of specific receptors in the CNS of rats and humans that recognize BZDs with tranquilizing or sedative – hypnotic activity [163, 164]. The so- called BZD receptor consists of at least three different subunits that form a pentameric macromolecular complex, the GABA receptor complex, which constitutes a chloride-permeable ion channel [165–167]. Electrophysiological studies have shown that BZDs increase the frequency at which the chloride channels open and consequently the flux of chloride ions at a given concentration of GABA [157].

4.1.1. 1,4-Benzodiazepines

Chlordiazepoxide [*58-25-3*], 7- chloro-2-methylamino-5-phenyl-3*H*-1,4-benzodiazepine-4-oxide, $C_{16}H_{14}ClN_3O$, M_r 299.75, *mp* 236 –236.5 °C. This first pharmacologically active member of the 1,4-benzodiazepine series is prepared as follows [168]:

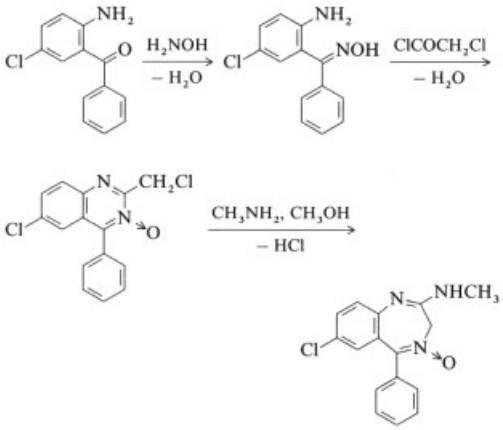

Chlordiazepoxide is used in the treatment of anxiety disorders, for the short-term relief of anxiety symptoms, for symptoms of acute alcohol withdrawal, or preoperative apprehension.

Trade Names. Librelease, Libritabs, Librium, Librizan, Tabrium (Roche), Retcol (Towa Yakuhin, Japan), Serendyl (Nezel, Spain), Servium (Teve, Israel), Zeisin (ICN, USA).

Diazepam [*439-14-5*], 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2*H*-1,4-benzodiazepin-2-one, $C_{16}H_{13}ClN_2O$, M_r 284.76, *mp* 125–126 °C. Diazepam is prepared as follows [169]:

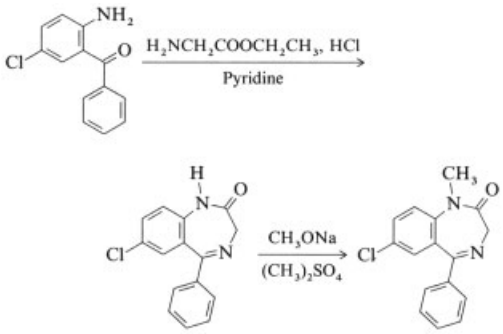

Use of diazepam is analogous to that of chlordiazepoxide.

Trade Names. Ansiolin (Scharper, Italy), Apozepam (A. L., Norway), Ceregulart (Kaken, Japan), Eurosan (Mepha, Switzerland), Levium (Sodelco, The Netherlands), Paxel (Marion, USA), Stesolid (Dumex, Denmark, CP Pharmaceuticals, UK), Valiquid, Valium, Valrelease (Roche).

Bromazepam [*1812-30-2*], 7-bromo-1,3-dihydro-5-(2-pyridinyl)-2*H*-1,4-benzodiazepin-2-one, $C_{14}H_{10}BrN_3O$, M_r 316.16, *mp* 237–238.5 °C (decomp.). This compound can be synthesized from 2-(2-aminobenzoyl)pyridine in three steps [170]:

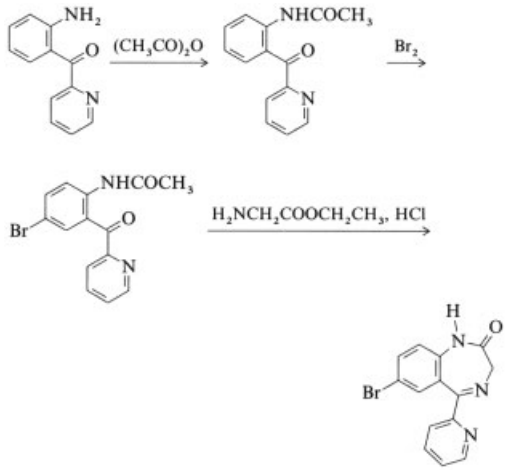

Indications are analogous to those of diazepam. Bromazepam, however, seems to be more efficient in the treatment of neurotic anxiety disorders.

Trade Names. Bromazenil (Hexal, Germany), Compendium (Polifarma, Italy), Creosedin (Osiris, Argentina), Durazanil (Durachemie, Germany), Lectopam, Lexomil, Lexotan, Lexotanil (Roche), Normoc (Merckle, Germany).

Lorazepam [*846-49-1*], 7-chloro-5-(2-chlorophenyl)-1,3-dihydro-3-hydroxy-2*H*-1,4-benzodiazepin-2-one, $C_{15}H_{10}Cl_2N_2O_2$, M_r 321.16, *mp* 166–168 °C. Lorazepam is prepared as follows [171].

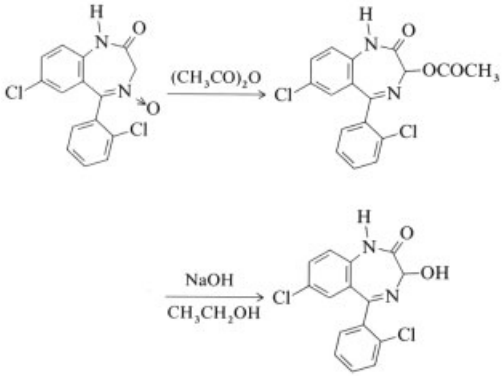

This short-acting compound is prescribed for management of anxiety disorders or anxiety associated with depressive symptoms.

Trade Names. Almazine (Steinhard, UK), Ativan (Wyeth), Emotival (Armstrong, Argentina), Lorasolid (Dumex, Denmark), Lorax (Wyeth), Pro Dorm (Schürholz, Germany), Tavor (Wyeth), Temesta (Ferrosan, Germany; Wyeth), Wypax (Yamanouchi, Japan).

Oxazepam [*604-75-1*], 7-chloro-1,3-dihydro-3-hydroxy-5-phenyl-2*H*-1,4-benzodiazepin-2-one, $C_{15}H_{11}ClN_2O_2$, M_r 286.74, *mp* 205 – 206 °C. A synthetic pathway analogous to that of lorazepam leads to oxazepam [172].

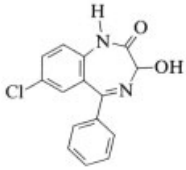

This short-acting compound is prescribed for the short-term relief of anxiety symptoms. Anxiety associated with depression as well as management of anxiety, tension, agitation, and irritability in the elderly are also responsive to oxazepam therapy. This compound is sometimes used as an hypnotic.

Trade Names. Limbial (Chiesi, Italy), Noctazepam (Brenner, Germany), Praxiten, Serax, Serenid, Seresta (Wyeth), Propax (Cipan, Portugal), Serepax (Wyeth, Ferrosan, Denmark), Sobril (Kabi Vitrum, Sweden), Zaxopam (Quantum, USA).

Clorazepate [*57109-90-7*], 7-chloro-2,3-dihydro-2,2-dihydroxy-5-phenyl-1*H*-1,4-benzodiazepine-3-carboxylic acid, dipotassium salt, $C_{16}H_{11}ClK_2N_2O_4$, M_r 408.93. This compound (used as its dipotassium salt) is prepared as follows [173]:

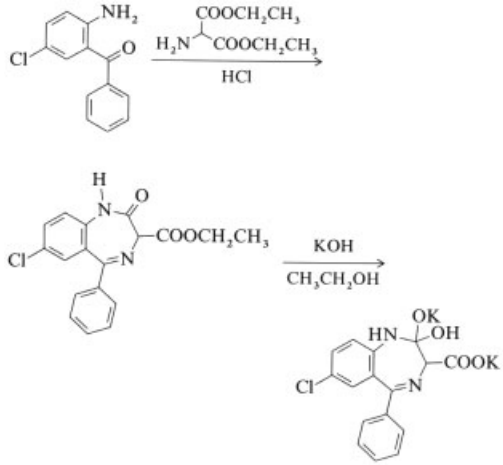

It is suitable for the symptomatic relief of acute alcohol withdrawal or, as adjunctive therapy, in the management of partial seizures.

Trade Names. Belseren (Mead-Johnson), Mendon (Dainippon, Japan), Transene, Tranxilium (Clin-Midy, France), Tranxène (Abbott, USA; Boehringer Ingelheim, Germany; Clin-Comar-Byla, France), Tranxilène (Clin-Comar-Byla, France).

Prazepam [*2955-38-6*], 7-chloro-1-(cyclopropylmethyl)-1,3-dihydro-5-phenyl-2*H*-1,4-benzodiazepin-2-one, $C_{19}H_{17}ClN_2O$, M_r 324.83, *mp* 145 – 146 °C. Alkylation of desmethyldiazepam using sodium hydride and cyclopropylmethyl bromide affords prazepam [174].

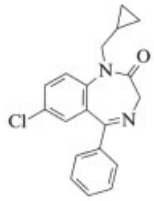

Analogously to diazepam, prazepam is suitable for treating anxiety disorders or short-term relief of anxiety symptoms.

Trade Names. Centrax (Parke Davis), Demetrin (Gödecke, Germany, Substantia, Austria, Parke Davis), Lysanxia (Substantia, France), Prazene (Parke Davis), Trepidan (Sigma Tau, Italy), Vestran (Warner-Chilcott).

Camazepam [*36104-80-0*], 7-chloro-1,3-dihydro-3-(*N,N*-dimethylcarbamoyl)-1-methyl-5-phenyl-2*H*-1,4-benzodiazepinone, $C_{19}H_{18}ClN_3O_3$, M_r 371.82, *mp* 173–174 °C. Camazepam is prepared as follows [172,175]:

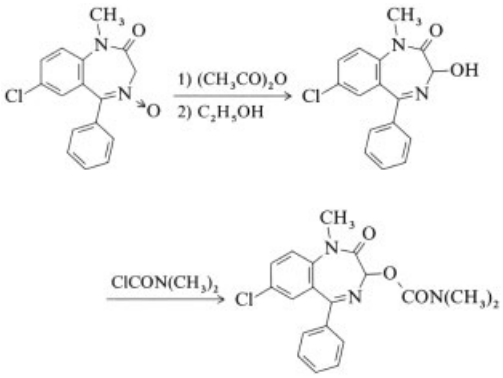

Trade Names. Albego (Boehringer Ingelheim; Farmasimes, Spain; Inpharzam, Switzerland), Limpidon, Nebolan.

Etizolam [*40054-69-1*], 1-methyl-6-(2-chlorophenyl)-8-ethyl-4*H*-*s*-triazolo[3,4-*c*]thieno-[2,3-*e*]-1,4-diazepine, $C_{17}H_{15}ClN_4S$, M_r 342.85, *mp* 147–148 °C. Treatment of the starting thienodiazepin-2-one with phosphorus pentasulfide in pyridine affords the 2-thione derivative which, by subsequent reaction with hydrazine, yields the 2-hydrazino compound. Final cyclization with azeotropic removal of water gives etizolam [176].

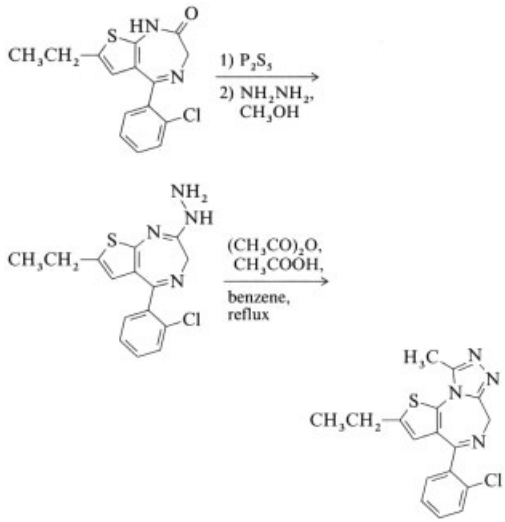

Trade Names. Depas (Yoshitomi, Japan), Pasaden (Farmades, Italy).

Adinazolam [*37115-32-5*], 8-chloro-1-(dimethylamino)methyl-6-phenyl-4*H*-*s*-triazolo-[4,3-*a*]-1,4-benzodiazepine, $C_{19}H_{18}ClN_5$, M_r 351.84, *mp* 165–166 °C. A synthesis of adinazolam starting from nonbenzodiazepine intermediates has been developed. The key step is the conversion of an aminoquinazoline into a benzophenone bearing a potential triazole ring [177].

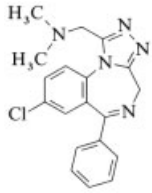

Trade Name. Deracyn (Upjohn).

Alprazolam [*28981-97-7*], 8-chloro-1-methyl-6-phenyl-4*H*-*s*-triazolo[4,3-*a*]-1,4-benzodiazepine, $C_{17}H_{13}ClN_4$, M_r 308.77, *mp* 228–228.5 °C. Alprazolam can be prepared as follows [178]:

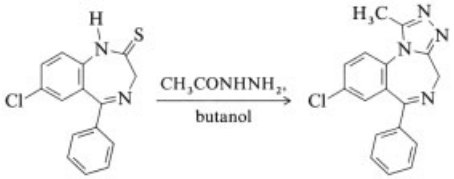

This anxiolytic is less sedative than diazepam, and is useful in the treatment of anxiety associated with depression.

Trade Names. Constan (Takeda, Japan), Frontal, Solanax, Tafil, Trankimazin, Xanax, Xanor (Upjohn), Valeans (Valeas, Italy).

Ketazolam [*27223-35-4*], 11-chloro-8,12*b*-dihydro-2,8-dimethyl-12*b*-phenyl-4*H*-1,3-oxazino[3,2-*d*][1, 4]benzodiazepine-4,7(6*H*)-dione, $C_{20}H_{17}ClN_2O_3$, M_r 368.82, *mp* 182–183.5 °C. Ketazolam is prepared in a one-step synthesis [179]:

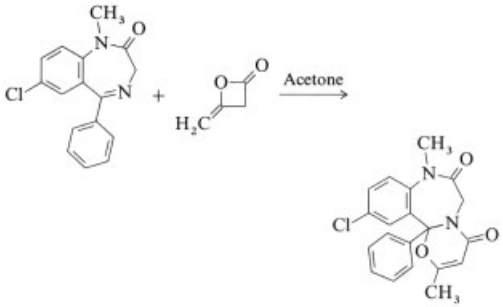

This compound is used as an anxiolytic with a long-term effect.

Trade Names. Anseren (Ciba-Geigy, Switzerland), Ansieten (Exa, Argentina), Anxon, Loftram, Sedotime (Beecham), Marcen, Parcil (Antibioticos, Spain), Unakalm (Upjohn).

Clotiazepam [*33671-46-4*], 5-(2-chlorophenyl)-7-ethyl-1,3-dihydro-1-methyl-2*H*-thieno[2,3-*e*]-1,4-diazepin-2-one, $C_{16}H_{15}ClN_2OS$, M_r 318.82, *mp* 105–106 °C. 2-Amino-3-(2-chlorobenzoyl)-5-ethylthiophene, prepared by reaction of butyraldehyde with 2-chloro-1-cyanoacetophenone and sulfur in the presence of triethylamine as catalyst, can be converted to the corresponding aminoacetamide derivative via two methods (Fig. 3) [180]:

This derivative is an anxiolytic thienodiazepine with apharmacological profile analogous to diazepam.

Trade Names. Clozan (Roerig, Belgium), Distensan (Esteve, Spain), Rize (Yoshitomi, Japan), Rizen (Puropharma, Italy), Tienor (Farmaka, Italy), Trecalmo (Tropon, Germany), Veretran (Latéma, France).

Bentazepam [*29462-18-8*], 1,3,6,7,8,9-hexahydro-5-phenyl-2*H*-1-benzothieno[2,3-*e*]-1,4-diazepin-2-one, $C_{17}H_{16}N_2OS$, M_r 296.38, *mp* 251–253 °C, can be prepared as follows [181]:

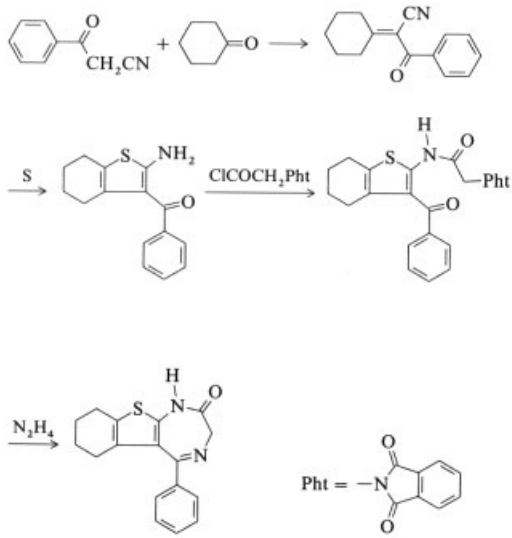

Trade Name. Thiadipone (C.S.I.C., Spain).

Halazepam [*23092-17-3*], 7-chloro-1,3-dihydro-5-phenyl-1-(2,2,2-trifluoroethyl)-2*H*-1,4-benzodiazepin-2-one, $C_{17}H_{12}ClF_3N_2O$, M_r 352.75, *mp* 164–166 °C. Halazepam is prepared as shown in Figure 4 [182].

This polyfluoroalkyl analogue of diazepam exhibits anxiolytic activity through its metabolite *N*-desmethyldiazepam. This benzodiazepine has a long-lasting effect and is prescribed for treatment of anxiety disorders.

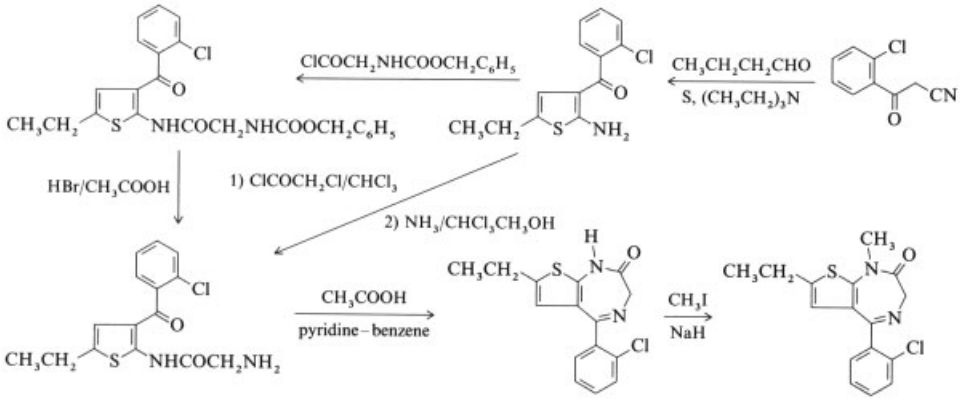

Figure 3. Synthesis of clotiazepam

Trade Names. Alapryl (Menarini, Spain), Pacinone (Delagrange, Belgium), Paxipam (Schering Corp., UK).

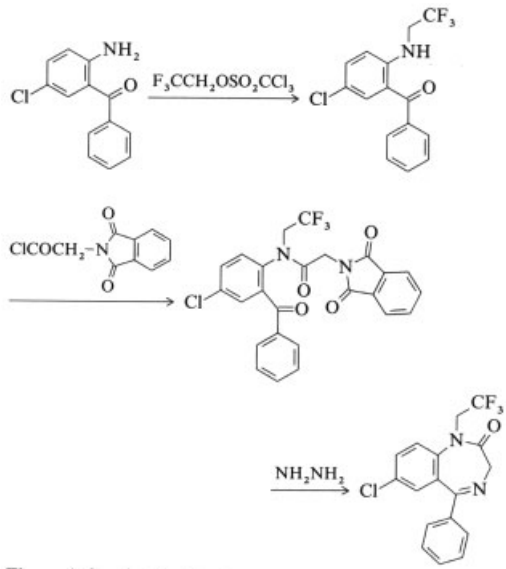

Figure 4. Synthesis of halazepam

4.1.2. Other Benzodiazepines

Tofisopam [*22345-47-7*], 7,8-dimethoxy-1-(3,4-dimethoxyphenyl)-5-ethyl-4-methyl-5*H*-2,3-benzodiazepine, $C_{22}H_{26}N_2O_4$, M_r 382.46, *mp* 156 – 157 °C. This 2,3-benzodiazepine is prepared as follows [183]:

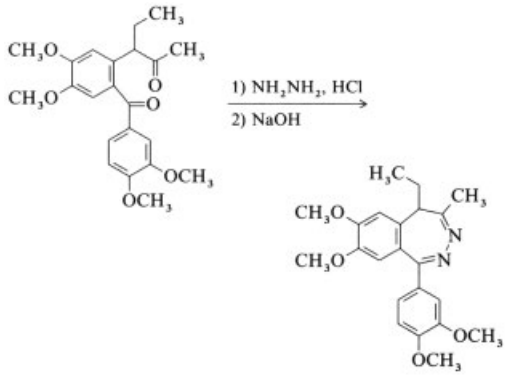

Tofisopam exhibits anxiolytic activity without the classical side effects of 1,4-benzodiazepines (hypnotic, anticonvulsant, myorelaxant effects) and does not affect the psychological faculties.

Trade Names. Grandaxin (Egis, Hungary; Ozothine, France), Sériel (Fabre, France), Tavor (Gerardo Ramon, Argentina).

Clobazam [*22316-47-8*], 7-chloro-1-methyl-5-phenyl-1*H*-1,5-benzodiazepine-2,4(3*H*,5*H*)-dione, $C_{16}H_{13}ClN_2O_2$, M_r 300.74, *mp* 180 – 182 °C. Clobazam is prepared as follows [184]:

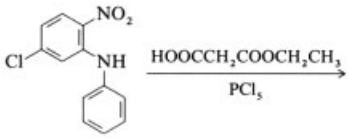

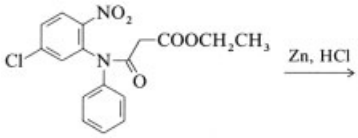

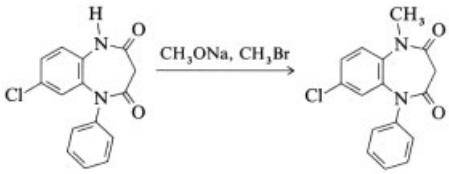

Trade Names. Clarmyl (Roussel-Iberica, Spain), Frisin, Frisium, Karidium, Noiafren, Sentil (Hoechst), Urbadan, Urbanol (Roussel), Urbanyl (Diamant, France, Roussel).

4.2. Carbamates

Meprobamate [*57-53-4*], 2-methyl-2-propyl-1,3-propyldicarbamate, $C_9H_{18}N_2O_4$, M_r 218.25, *mp* 104 – 106 °C. Meprobamate is prepared as follows [185]:

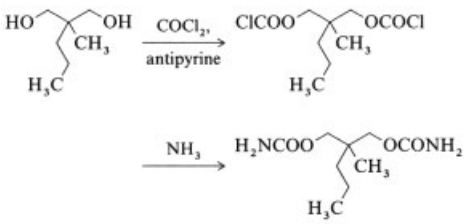

Meprobamate is an anxiolytic agent which has anticonvulsant, myorelaxant, and sedative properties [160], but it does not directly interfere with the GABA-receptor complex. The potentiation of the release of adenosine by meprobamate may account for its central effects [186]. This carbamate derivative has been used for the management of anxiety disorders and tension states but has been superseded by benzodiazepines.

Trade Names. Amosene (Ferndale, USA), Artolon (Roter, The Netherlands), Cyrpon (Tropon, Germany), Equanil (Wyeth, Clin-Comar-Byla, France), Miltaun (Mack, Germany), Perequil (Merrell Dow Lepetit), Aneural, Quaname, Quanil (Wyeth), Tranquilin (SAM-Parke-Davis, Belgium).

Phenprobamate [*673-31-4*], 3-phenylpropyl carbamate, $C_{10}H_{13}NO_2$, M_r 179.21, *mp* 101 –104 °C, for preparation see [187].

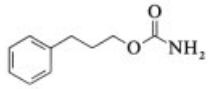

It has been given as an anxiolytic and myorelaxant.

Trade Names. Actiphan (Teikoku, Nagase, Japan), Ansepron (Fuso, Japan), Gamaquil (Siegfried, Switzerland; Leo, Mekos, Pharmacia, Sweden), Kuilil (Embil, Turkey), Nelaxan (Toho, Japan), Paraquick (Ohta, Japan), Phencol (Towa Yakuhin, Japan), Spantol (Nippon Chemiphar, Japan).

Tybamate [*4268-36-4*], *N*-butyl-2-methyl-2-propyl-1,3-propyldicarbamate, $C_{13}H_{26}N_2O_4$, M_r 274.35, *mp* 49 – 51 °C, is prepared by reaction of 2-methyl-2-propyl-3-hydroxypropyl carbamate with butyl isocyanate [188].

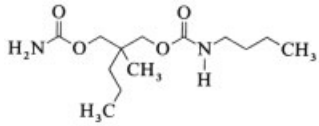

Trade Name. Tybatran (Robins, USA).

Hydroxyphenamate [*50-19-1*], 2-hydroxy-2-phenylbutyl carbamate, $C_{11}H_{15}NO_3$, M_r 209.24, *mp* 55 – 56.5 °C. This carbamate is prepared from β-ethyl-β-hydroxyphenethyl alcohol and ethyl chloroformate followed by reaction with ammonia [189].

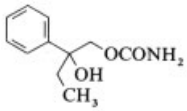

Trade Name. Listica (Armour).

4.3. Other Anxiolytics

Hydroxyzine [*68-88-2*], 2-{2-{4-[(4-chlorophenyl)phenylmethyl]-1-piperazinyl}ethoxy}-ethanol, $C_{21}H_{27}ClN_2O_2$, M_r 374.92, *mp* 193 °C. Hydroxyzine is produced as follows [190]:

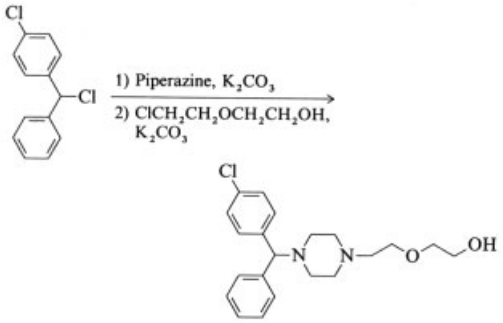

Hydroxyzine is an antihistaminic diarylalkylpiperazine which is used as an anxiolytic. It has no cortical depressant activity, its anxiolytic action could be explained by its suppressive effect on neuronal activity in subcortical areas. A myorelaxant activity has also been demonstrated experimentally.

Trade Names. (Dihydrochloride) Alamon (Grelan, Japan), Atarax (Pfizer, Belgium; UCB-Fraysse, France), Aterax (UCB, Belgium), Multipax (Rorer, USA), Orgatax (Organon), Ulgrax (UCB, Belgium), Vistaril Parenteral (Pfizer), (Pamoate) Equipose (Pfizer).

Etifoxine [*21715-46-8*], 6-chloro-2-ethylamino-4-methyl-4-phenyl-4*H*-3,1-benzoxazine, $C_{17}H_{17}N_2O$, M_r 300.79, *mp* 90 °C, hydrochloride, *mp* 150 °C. Etifoxine is produced as follows [191]:

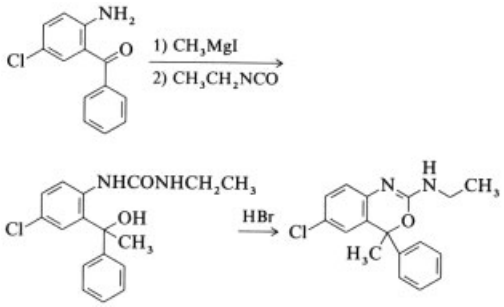

In rodents, etifoxine proved to be an anxiolytic. It possesses sedative, antiaggressive, anticonvulsant, and anticonflict effects [192]. Etifoxine enhances GABA-mediated neurotransmission.

Trade Name. (Chloride) Stresam (Beaufour, France).

Alpidem [*82626-01-5*], 6-chloro-2-(4-chlorophenyl)-*N,N*-dipropylimidazo[1,2-*a*]pyridine-3-acetamide, $C_{21}H_{23}Cl_2N_3O$, M_r 404.34, *mp* 140–141 °C. Alpidem is produced as follows [193]:

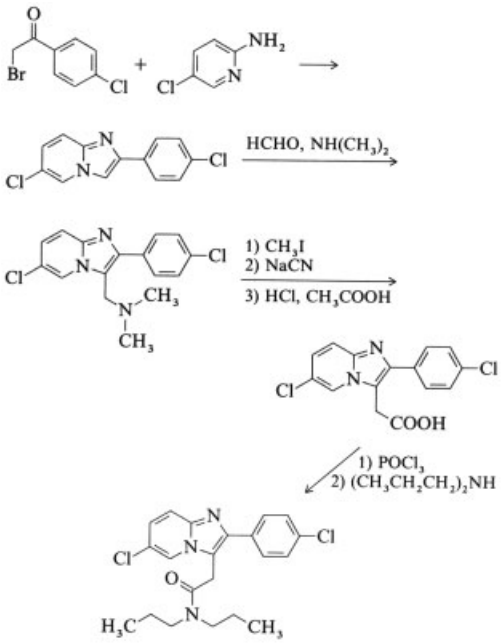

Alpidem is an imidazopyridine derivative which produces anxiolytic-like effects in some but not all models and anticonvulsant properties without myorelaxant and sedative effects [194].

Whereas BZDs are largely nonselective for the subtypes of the GABA receptor complex, alpidem displays a preferential affinity for the BZD_1 modulatory site [195]. Moreover, relatively high doses of alpidem antagonize the myorelaxant and sedative – hypnotic activities of diazepam.

Trade Name. Ananxyl (Synthelabo, France).

Buspirone [*36505-84-7*], 8-[4-[4-(2-pyrimidinyl)-1-piperazinyl]butyl]-8-azaspiro[4.5]decane-7,9-dione, $C_{21}H_{31}N_5O_2$, M_r 385.51, hydrochloride, *mp* 202 °C, is prepared as follows [196]:

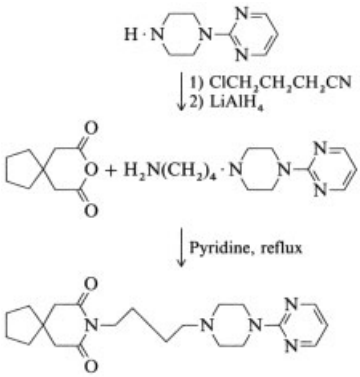

In tests for predicting clinical anxiolytic activity, buspirone, an azaspirodecanedione derivative, attenuates conflict behavior in pigeons, rodents, and monkeys. Buspirone is less sedative than the classical BZDs as shown in experimental and clinical trials. It is ineffective as an anticonvulsant against convulsions induced by pentylenetetrazol, bicuculline, picrotoxine, or supramaximal electroshock. Similarly, no motor incoordination or myorelaxant activity occurs in buspirone-treated animals [195].

Buspirone does not bind to the GABA receptor complex and does not affect BZD binding. Its most evident interaction occurs at the $5\text{-}HT_{1A}$ receptor. This action is supported by binding studies, anatomical localization, neurochemistry, electrophysiology, and animal behavior [197].

Trade Names. Anisal (Vita, Spain), Ansiced (Abello, Spain), Bespar (Bristol), Buspar (Astra, Bristol-Myers, Mead-Johnson), Buspisal (Lesvi, Spain), Buspon (Deva, Turkey), Censpar (Bristol-Myers), Noral (Almirall, Spain).

4.4. Anxiolytics Under Development

New agents have been reported which interact with the central BZD receptor complex, but do not have a simple benzodiazepine structure. They appear to be selective in their anti-anxiety action. Examples include bretazenil, abecarnil, FG-7516, CGS 20625, and dipavlon. *Bretazenil* [*84379-13-5*] (Ro 166028-Roche) exhibits anticonflict and anticonvulsant properties. Only mild sedation appears at doses needed to produce anticonvulsant or anxiolytic effects, and potentiation of ethanol-induced sedation is less pronounced than with diazepam. Unlike full agonists, physical dependence could not be induced in monkeys after administration of bretazenil, even at a very high dose. Bretazenil is a partial agonist at the GABA-A receptor complex. It is in clinical trials (phase III) for generalized anxiety disorders, panic attacks, and schizophrenia [198]. *Abecarnil* [*111841-85-1*] [199] (ZK 112119, Schering) belongs to the β-carboline family and is in phase III clinical trials. It is active as an anxiolytic but less sedative or myorelaxant than diazepam. Like bretazenil, abecarnil must be considered as a partial agonist at the GABA-A receptor complex. *FG-7516* [*83910-44-5*] (Novo-Nordisk) is structurally related to abecarnil and is also in phase III clinical trials [200].

CGS-20625 [*111205-55-1*] is undergoing phase I clinical trials as a potent, selective inhibitor at the central benzodiazepine receptor. It does not appear to have the side effects of diazepam. This analogue of CGS-9895 (suspended) belongs to the cyclohepta[*b*]pyrazolo[3,4-*d*]pyridin-3(2*H*)-one family. *Dipavlon* [*90808-12-1*] (RU 32698, Roussel-Uclaf) is an imidazo[1,2-*a*]pyrimidine with anticonvulsant and anxiolytic properties but with few effects in tests used to predict sedation or motor impairment in rodents. Anxiolytic

and anticonvulsant effects are antagonized by flumazenil, the BZD antagonist. Dipavlon acts as a partial agonist at the GABA-A receptor complex and is in phase II clinical trials.

Suriclone [*53813-83-5*] (Rhône-Poulenc Rorer) is a cyclopyrrolone derivative which acts at the GABA-A receptor complex. It does not induce physical dependence in mice even at a high dose after a chronic treatment. Binding studies show that suriclone interacts with a different binding domain to anxiolytic BZD. Clinical studies have shown that suriclone is an anxiolytic with a low incidence of drowsiness and a low dependence potential compared with BZD.

The discovery of the anxiolytic properties of the 5-HT_{1A} partial agonist buspirone has given rise to a second generation of serotoninrgic agents (e.g., gepirone, ipsapirone, tandospirone, umespirone, zalospirone, and metanopirone). All these agents are, however, of low efficacy in patients having been treated with BZDs (lack of muscle relaxant and sedative properties). Other serotoninrgic agents such as ritanserin (5-HT_2 antagonist) or ondansetron (5-HT_3 antagonist) are effective in the treatment of generalized anxiety disorders [201].

Among newer molecular approaches to anxiety, selective antagonists such CI 988 or L 365,260, for the brain cholecystokinin receptors as well as the angiotensin II antagonist DuP 753 have been reported to be potential anxiolytics [202, 203].

5. Geriatric Drugs

Geriatric drugs include substances with cognition-enhancing properties that are used to treat the mental manifestations of aging. The continual increase in the expected lifespan in developed countries enhances the prevalence of conditions ranging from mild age-associated memory impairments (AAMI) to senile dementias (SD). Neurodegenerative processes and atherosclerosis have been identified as possible etiologic causes for certain types of dementias such as senile dementias of the Alzheimer's type (SDAT) and multi-infarct dementia.

The search for geriatric drugs is characterized by the absence of unequivocal clinically effective reference compounds. Moreover, inability to identify neurobiological processes that are specifically involved in memory formation in the brain makes the preclinical evaluation of these compounds difficult. Behavioral models of learning based on cognitive interpretations of rodent locomotor performances often remain the only available testing procedure [204].

Although no currently available drug can stop the evolution of cognitive disorders, this chapter focuses on important drugs that alleviate or correct some of the symptoms of the disease.

5.1. Nootropics

The term nootropic denotes substances that do not act through a given neurotransmitter system, that are neither psychostimulant nor sedative, and that improve learning performance or compensate for learning deficits induced by amnesic manoeuvres in behavioral models. Although a number of geriatric drugs fall within this rather wide unrestrictive definition, the classical nootropics are chemically related to their prototype piracetam (2-pyrrolidinone derivatives). Their mechanism of action is not known, although they are believed to act in the telencephalon [205]. The toxicity of nootropics is generally low.

Piracetam [*7491-74-9*], 2-oxo-1-pyrrolidineacetamide, $C_6H_{10}N_2O_2$, M_r 142.16, *mp* 151 – 152 °C is prepared by condensing 2-pyrrolidinone with ethyl chloroacetate in the presence of a metal hydride and then converting the ester into an amide with ammonia [206].

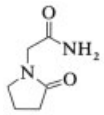

Piracetam protects the brain from hypoxic insults and facilitates transcallosal synaptic transmission in animals. In addition to senile dementia, it is proposed for the treatment of dyslexia in children and of cerebrovascular stroke.

Trade Names. Agivilen (Efeka, Germany), Ciclofalina (Almirall, Spain), Durapitrop (Durachemie, Switzerland), Noostan (UCB, Belgium), Nootrop (UCB, Belgium), Novocetam

(Beiersdorf, Germany), Piracebral (Hexal, Germany), Stimucortex (Kalifarma, Spain).

Aniracetam [*72432-10-1*], 1-(4-methoxybenzoyl)-2-pyrrolidone, $C_{12}H_{13}NO_3$, M_r 219.24, *mp* 121 – 122 °C is prepared by condensing 2-pyrrolidone with 4-methoxybenzoyl chloride [207].

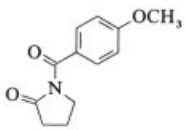

This piracetam-related cognition enhancer reduces glutamate receptor desensitization [208].

Trade Names. Draganon (Roche, Switzerland), Sarpul (Toyama, Japan).

Oxiracetam [*62613-82-5*], 4-hydroxy-2-oxo-1-pyrrolidineacetamide, $C_6H_{10}N_2O_3$, M_r 158.16, *mp* 165 – 168 °C. Oxiracetam can be prepared as follows [209]:

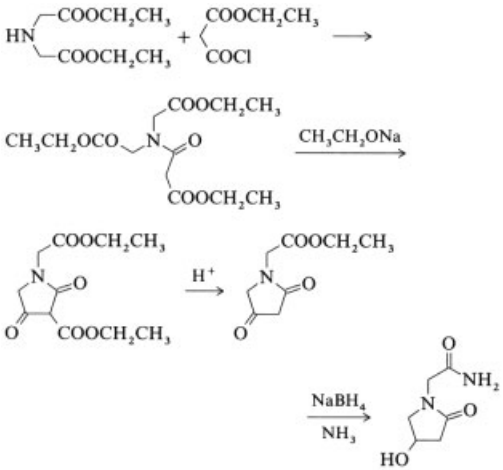

A second route starts with 4-amino-3-hydroxybutyric acid which is *O*-silylated and dehydrated to give 2-oxo-3-trimethylsilyloxypyrrolidine. Alkylation with ethyl bromoacetate, deprotection of the alcohol, and ammonolysis afford oxiracetam.

Oxiracetam is more hydrophilic than aniracetam and is claimed to have a wider therapeutic spectrum.

Trade Names. Neupan (SmithKline Beecham, Portugal), Neuractiv (Ciba-Geigy, Switzerland), Neuromet (ISF, Italy).

5.2. Cholinergic Drugs

Cholinergic drugs are used in cognitive disorders on the assumption that they compensate for the deficiency of central cholinergic processing that is thought to underly dementia [210]. This cholinergic hypothesis is supported by results from learning experiments which show that antagonists at muscarinic receptors (e.g., scopolamine) induce impairment in animals and humans. Destruction of the principal cholinergic projection system in the rodent brain produces similar effects, and a loss of cholinergic terminals has been observed in the brains of patients suffering from Alzheimer's disease.

5.2.1. Choline Analogues

Choline analogues or choline-containing precursors are used in an attempt to increase the rate of brain acetylcholine synthesis. These compounds are expected to be more effective than simple dietary supplementation with choline or choline-containing lecithin.

Deanol [*108-01-0*], 2-(dimethylamino)ethanol, $C_4H_{11}NO$, M_r 89.14, *bp* 135 °C, is prepared from dimethylamine and ethylene oxide [211]. It also exists as the salts of acetylglutamic, 4-acetamidobenzoic, succinic, and tartaric acids.

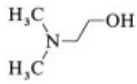

Deanol is metabolized and incorporated into the brain choline pool, thereby increasing choline availability [212]. This compound may also have an effect on brain phospholipid metabolism.

Trade Names. Acumen (Menarini, Spain), Bimanol (Polfa, Poland), Cervoxan (S.M.B., Belgium), Deaner (Riker), Diforène (Choay, France), Pabenol (Gentili, Italy), Tonibral (Bouchara, France), Varesal (Arzneimittelwerk, Germany).

Meclofenoxate [*51-68-3*], dimethylaminoethyl 4- chlorophenoxyacetate, $C_{12}H_{16}Cl_2NO_3$, M_r 293.22, *mp* 135 – 139 °C. Meclofenoxate is obtained by condensing 4- chlorophenoxyacetyl chloride with 2-(dimethylamino)ethanol [213].

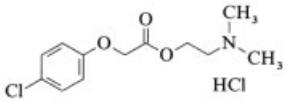

Trade Names. Analux (Polfa, Poland), Brenal (Tanabe, Japan), Cerutil (Arzneimittelwerk, Germany), Helfergin (Promonta, Germany), Marucotol (Maruko, Japan), Mecloxate (Sawai, Japan), Proseryl (Funai, Japan), Ropoxyl (Nippon Chemiphar, Japan).

Citicoline [*987-78-0*], cytidine 5′-sodium diphosphate choline ester, $C_{14}H_{25}N_4NaO_{11}P_2$, M_r 456.33, is a nucleotide coenzyme that plays an important role in the metabolism of phospholipids. It is prepared by reacting choline phosphate and cytidine 5′-phosphate with a condensation reagent such as *N,N′*-dicyclohexylcarbodiimide [214] or tosyl chloride [215].

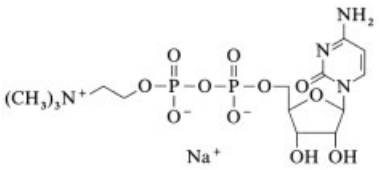

Trade Names. Alaton (Zambon, Italy), Colite (Nippon Chemiphar, Japan), Corenalin (Kaken, Japan), Ensign (Yamanouchi, Japan), Eumetabol (Llano, Spain), Nicholin (Takeda, Japan; Byk Procienx, Brazil), Nicolsint (Von Boch, Italy), Nicolin (Morishita, Japan), Startonyl (American Cyanamid, USA).

5.2.2. Cholinesterase Inhibitors

Cholinesterase inhibitors are used with the rationale that they can increase the efficiency of the remaining cholinergic terminals by inhibiting acetylcholine depletion.

Physostigmine [*57-47-6*], (3*aS*- *cis*)-1,2,3,3*a*, 8,8*a*-hexahydro-1,3*a*,8-trimethylpyrrolo[2,3-*b*]-indol-5-ol methyl carbamate, $C_{15}H_{21}N_3O_2$, M_r 275.34, *mp* 105 – 106 °C, is extracted from calabar beans [216].

Physostigmine restores learning performance in animals with lesions of the nucleus basalis magnocellularis [217].

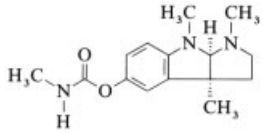

Trade Names. Anticholium (Köhler, Germany), Antilirium (Forest, USA), Fiostin (Tubi Lux Farma, Italy).

Tacrine [*321-64-2*], 9-amino-1,2,3,4-tetrahydroacridine, $C_{13}H_{14}N_2$, M_r 198.27, *mp* 183 – 184 °C, can be prepared by different methods, the shortest is condensation of 2-aminobenzonitrile with cyclohexanone [218].

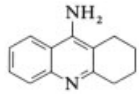

This product is a known curare antidote. It has a longer half-life than physostigmine and has been used with some success by oral route in Alzheimer's patients [219]. Tacrine competitively inhibits acetylcholinesterase and also modifies the resting conductance and responses of neurons in the hippocampus.

Trade Name. Cognex (Warner Lambert).

5.2.3. Thyroliberin Analogues

Thyroliberin, also known as thyrotropin releasing hormone (TRH), is a hypothalamic regu-

latory hormone. It also acts on cholinergic neurons to enhance high affinity choline uptake and stimulate acetylcholine synthesis and release. See also → Peptide and Protein Hormones, Chap. 2.2. Thyroliberin is of possible interest for the treatment of neurodegenerative processes, such as Alzheimer's disease and amyotrophic lateral sclerosis.

Protirelin [*24305-27-9*], 5-oxo-L-prolyl-L-histidyl-L-prolinamide, $C_{16}H_{22}N_6O_4$, M_r 362.39, the natural TRH tripeptide is prepared by conventional peptide synthesis [220].

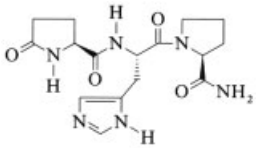

Trade Name. Hirtonin (Hoechst).

Posatirelin [*78664-73-0*], L-pyro-2-aminoadipyl-L-leucyl-L-prolinamide, $C_{17}H_{28}N_4O_4$, M_r 352.44, *mp* 214–216 °C, is prepared by sequential coupling, amino acids being activated as their pentafluoro esters. Prolinamide is condensed with *N*-carboxybenzylleucine. The dipeptide is then deprotected and coupled with L-pyro-2-aminoadipic acid [221].

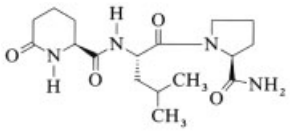

Posatirelin is a pseudopeptide which improves sequelae of head injury and subarachnoid hemorrhage.

Trade Name. Pyladox (Dainippon).

5.3. Metabolically Active Drugs

Metabolically active drugs are heterogeneous compounds with different proposed mechanisms of action. They help to restore deficient brain energy metabolism. Some of the most prescribed cognition enhancers are members of this class.

Idebenone [*58186-27-9*], 2,3-dimethoxy-6-(10-hydroxydecyl)-5-methyl-1,4-benzoquinone, $C_{19}H_{30}O_5$, M_r 338.45, *mp* 46–50 °C, is prepared in five steps. Friedel–Crafts acylation of 3,4,5-trimethoxytoluene with 1,10-decanedioic acid semiester affords a partially demethylated keto acid which is esterified and then reduced twice. The phenolic ring is oxidized with Fremy's salt to give the quinone [222].

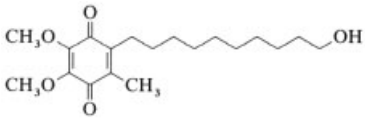

Idebenone is a benzoquinone that is derived from the widely distributed ubiquinones. It accelerates ATP formation by activating the electron-transfer system. It has free-radical scavenger properties, stimulates mitochondrial respiration, and reduces membrane lipid peroxidation.

Trade Names. Aban, Avan (Takeda, Japan).

Nicergoline [*27848-84-6*], 5-bromo-3-pyridinecarboxylate (8*β*)-1,6-dimethyl-10-methoxyergoline-8-methanol, $C_{24}H_{26}BrN_3O_3$, M_r 484.40, *mp* 136–138 °C, is prepared by reducing a lumilysergic acid derivative with $LiAlH_4$ and acylating the alcohol thus obtained with 5-bromo-3-pyridinecarboxylic acid chloride [223].

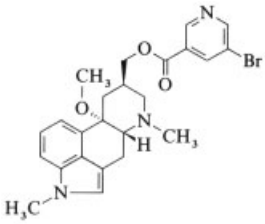

Trade Names. Duracebrol (Durachemie, Germany), Ergobel (Hormosan, Germany), Memoq (Gödeke, Germany), Nicergolyn

(Farnex, Italy), NicerHexal (Hexal, Germany), Sermion (Farmitalia Carlo Erba; Specia), Varson (Almirall, Spain).

Ergoloid mesylates, also called dihydroergotoxine or co-dergocrine, is a mixture of the hydrogenated ergot alkaloids dihydroergocornine [*25447-65-8*], dihydroergocristine [*17479-19-5*], and dihydro-α-ergocryptine [*25447-66-9*], and dihydro-β-ergocryptine [*77528-01-9*].

Dihydroergotoxine is among the most widely prescribed geriatric drugs. It has vasoactive properties and increases brain oxygen tension and electrical activity [224].

Trade Names. Circanol (Riker), Deapril-ST (Mead Johnson), Hydergine (Sandoz), Niloric (Ascher), Trigot (Squibb).

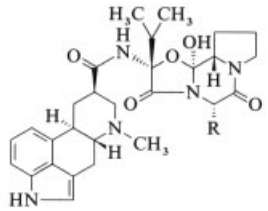

Dihydroergocornine	R = $CH(CH_3)_2$
Dihydroergocristine	R = $CH_2C_6H_5$
Dihydro-α-ergocryptine	R = $CH_2CH(CH_3)_2$
Dihydro-β-ergocryptine	R = $CH(CH_3)CH_2CH_3$

Bifemelane [*90293-01-9*], 2-(4-methylaminobutoxy)diphenylmethane, $C_{18}H_{23}NO$, M_r 269.39, *mp* 117 – 121 °C, is prepared by alkylation of 2-hydroxydiphenylmethane with 1,4-dibromobutane and subsequent condensation with methylamine [225].

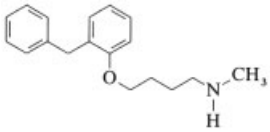

Bifemelane has antianoxic and anti-ischemic properties. It also exerts an indirect cholinomimetic effect by increasing acetylcholine release [226].

Trade Names. Alnert (Mitsubishi, Japan), Celeport (Esai, Japan).

Vinpocetine [*42971-09-5*], $C_{22}H_{26}N_2O_2$, M_r 350.46, *mp* 147 – 153 °C is prepared by esterification of vincaminic acid [227].

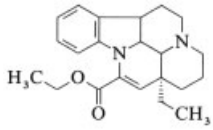

Vinpocetine is a Vinca alkaloid. It increases cerebral blood flow and glucose utilization, and inhibits brain edema formation produced by cerebral metabolic inhibitors [224].

Trade Names. Cavinton, Ceractin (Ayerst), Eusenium (Thiemann, Germany).

Propentofylline [*55242-55-2*], 3,7-dihydro-3-methyl-1-(5-oxohexyl)-7-propyl-1*H*-purine-2,6-dione, $C_{15}H_{22}N_4O_3$, M_r 306.37, *mp* 69 – 70 °C, is prepared by reacting 3-methyl-7-propylxanthine with 6-bromo-2-hexanone in refluxing sodium hydroxide [228].

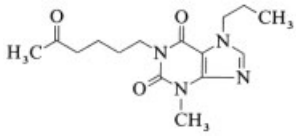

This alkylxanthine is structurally related to the naturally occurring psychostimulant caffeine (methylxanthine). Propentofylline inhibits cyclic nucleotide phosphodiesterases, thereby increasing the levels of intracellular second messengers [229].

Trade Name. Hextol (Hoechst, Japan).

Pyritinol [*1098-97-1*], bis(4-hydroxymethyl-5-hydroxy-6-methyl-3-pyridylmethyl)-disulfide, $C_{16}H_{20}N_2O_4S_2$, M_r 368.48, *mp* 218 – 220 °C, is prepared by reacting a salt of 3,4-bis(bromomethyl)-5-hydroxy-6-methylpyridine with potassium xanthogenate and then hydrolyzing the intermediate with alcoholic ammonia [230].

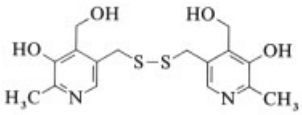

Pyritinol consists of two pyridoxine (vitamin B_6) molecules linked by a disulfide bridge. It is devoid of vitamin B_6-like activity and improves glucose metabolism in patients with organic brain disorders based on disturbed cerebral glucose metabolism [231].

Trade Names. Bonifen (Merck, Spain), Dinerfene (Azevedos, Spain), Divalvon-D (Nippon Kagaku, Japan), Enbol (Merck, Japan), Encefavol (Bracco, Italy), Neuroxin (Yamanouchi, Japan), Sawaxin (Sawai, Japan).

5.4. Antidepressants

Several atypic antidepressants that lack the anticholinergic activity of classical tricyclic antidepressants are successfully used in the treatment of cognitive disorders. Their activity may also be due to the alleviation of depression which often contributes to a decline in cognitive performance in the elderly patient.

Citalopram (see Section 3.4) inhibits 5-HT reuptake and significantly improves cognitive performance in Alzheimer patients [134].

Exifone [*52479-85-3*], 2,3,3′,4,4′,5′-hexahydroxybenzophenone, $C_{13}H_{10}O_7$, M_r 278.22, *mp* 270 °C, is prepared by the Friedel–Crafts acylation of pyrogallol with 3,4,5-trihydroxybenzoic acid [232].

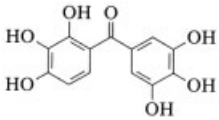

Exifone shows an atypical antidepressant profile of activity in a range of standard psychopharmacological tests [233] and is also claimed to protect against hypoxia. This metabolic enhancer has no anticholinergic properties and thus lacks associated side effects.

Trade Name. Adlone (Pharmascience, France).

Indeloxazine [*65043-22-3*], 2-[(inden-7-yloxy)methyl]morpholine hydrochloride, $C_{14}H_{18}ClNO_2$, M_r 267.76, *mp* 155–156 °C, is prepared as follows [234]:

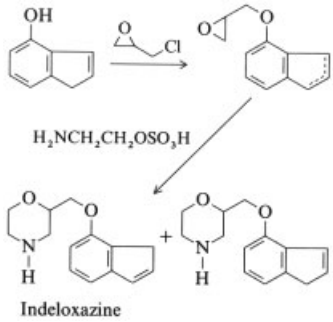

The two isomers are separated by fractional crystallization [234].

This antidepressant agent inhibits 5-HT and noradrenaline reuptake but has only weak effects on neurotransmitter receptors. It is effective in treating neuropsychological symptoms, the major side effect being gastrointestinal disorders.

Trade Names. Elen, Noin (Essex Nippon, Yamanouchi, Japan).

5.5. Miscellaneous Geriatric Drugs

Phosphatidylserine is extracted from bovine cerebral cortex. It is believed to act on neuronal membrane fluidity and acetylcholine release. It improves learning performance in aged rats [235].

Trade Name. Bros (Fidia).

Acetylcarnitine [*5080-50-2*], $C_9H_{18}ClNO_4$, M_r 239.70, *mp* 187 °C, is an endogeneous substance. It is prepared by acetylation of L-carnitine [236].

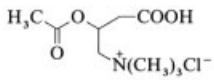

Acetyl-L-carnitine stereospecifically facilitates neuronal electrophysiological responses to acetylcholine and 5-HT [237].

Trade Names. Branigan (Glaxo, Italy), Nicetile (Sigma-Tau, Italy).

Nimodipine [*66085-59-4*], isopropyl 2-methoxyethyl 1,4-dihydro-2,6-dimethyl-4-(3-nitrophenyl)-3,5-pyridinedicarboxylate, $C_{21}H_{26}N_2O_7$, M_r 418.45, *mp* 125 °C, is prepared as follows [238]:

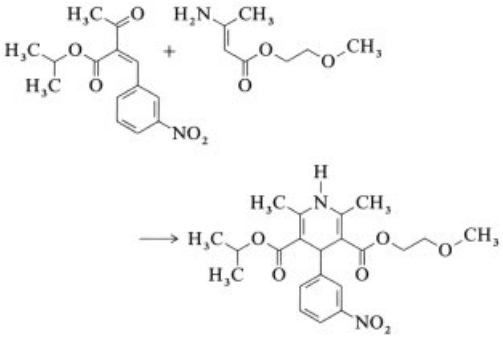

Nimodipine is a calcium channel antagonist used for its vasodilatatory and hemodynamic effects on cerebral blood vessels [239].

Trade Name. Nimotop (Bayer).

5.6. Geriatric Drugs Under Development

The search for cognition enhancers is a rapidly growing field and numerous substances are being investigated. The following are in the phase III stage of development. *Sabeluzole* [*104592-54-3*] (R 58735) is a benzothiazole derivative under investigation by Janssen. It has antihypoxic and anticonvulsant properties [240]. *Denbufylline* [*57076-71-8*] (BRL-30892, Smith K-line Beecham) is an alkylxanthene derivative which shares the phosphodiesterase-inhibiting properties of this class of drugs [224]. It is under investigation in multi-infarct dementia. *Pramiracetam* [*68497-62-1*] (CI-879) is a nootropic closely related to piracetam under development by Parke-Davis for memory and cognitive deficits following electroconvulsive shock therapy. *Teniloxazine* [*62473-79-4*] (Y-8894) is a nonpyrrolidine nootropic compound developed by Yoshitomi which has also antiulcer properties in the gastrointestinal tract. *Velnacrine* [*118909-22-1*] (HP-029, Hoechst-Roussel) is an analogue of tacrine with cholinesterase-inhibiting properties. It is under clinical investigation in Alzheimer's patients. *Vinconate* [*70704-03-9*] (OC-340, Omnichem) increases brain oxygen levels and delays the onset of EEG abnormalities during hypobaric hypoxia in humans.

6. References

1. J. C. Blanchard, L. Julou in J. P. Olie, D. Ginestet, G. Jolles (eds.): *Histoire d'Une Découverte en Psychiatrie: 40 Années de Chimiothérapie,* Douin, Paris 1992.
2. M. Goldstein, A. Y. Deutsch, *FASEB J.* **6** (1992) 2413.
3. R. J. Baldessarini in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 383.
4. H. P. Rang, M. M. Dale: *Pharmacology,* Churchill Livingstone, London 1987.
5. B. B. Leonard: *Fundamentals of Psychopharmacology,* J. Wiley & Sons, Chichester 1992.
6. J. C. Schwartz, B. Giros, M. P. Martres, P. Sokoloff in *Proceedings of the International Academy for Biomedical and Drug Research: Workshop on New Generation of Antipsychotic Drugs: Novel Mechanisms of Action,* Karger, Basel 1992, in press.
7. G. P. Reynolds, *Trends Pharmacol. Sci.* **13** (1992) 116.
8. Rhône-Poulenc, US 2 519 886, 1950 (P. Charpentier).
9. Rhône-Poulenc, US 2 645 640, 1953 (P. Charpentier).
10. Smith, Kline and French Laboratories, GB 813 861, 1959.
11. J. Schmitt, *Bull. Soc. Chim. Fr.* 1957, 938.
12. Rhône-Poulenc, US 2 837 518, 1958 (R. M. Jacob, J. G. Robert).
13. Rhône-Poulenc, US 2 877 224, 1959 (R. M. Jacob, J. G. Robert).
14. J.-P. Bourquin et al., *Helv. Chim. Acta* **41** (1958) 1072.
15. Sandoz, US 3 084 161, 1963 (J. Renz, J. P. Bourquin, G. Schwab).

16. Sandoz, FR 1 363 683, 1 459 476, 1964 (J. Renz et al.).
17. Searle, US 2 957 870, 1960 (J. W. Cusic, H. W. Sause).
18. Rhône-Poulenc, FR 1 212 031, 1961 (M. J. Robert, G. R. Jacques).
19. Searle, GB 861 807, 1958.
20. Rhône-Poulenc, FR M 7835, 1970.
21. Yoshitomi Pharmaceutical Ind., US 3 574 204, 1966.
22. Rhône-Poulenc, GB 780 193, 1957 (R. J. Horclois).
23. Rhône-Poulenc, US 2 902 484, 1959 (R. J. Horclois).
24. SKF, US 2 921 069, 1960 (G. E. Ullyot).
25. Rhône-Poulenc, GB 814 512, 1959.
26. Farbenfabrik Bayer, DE 1 120 451, 1956.
27. Schering Corp., US 2 860 138, 1958 (M. H. Sherlock, N. Sperber).
28. A. Toldy et al., *Acta Chim. Acad. Sci. Hung.* **42** (1964) 351.
29. J. W. Cusic, US 2 766 235, 1956.
30. Smith, Kline and French, GB 829 246, 1960.
31. Schering Corp., US 2 985 654, 1956 (M. H. Sherlock, N. Sperber).
32. American Home Products, US 3 023 146, 1962 (R. F. Tislow, W. F. Bruce, J. A. Page).
33. Deutsche Gold und Silber, DE 1 445 648, 1962 (H. Beschke, W. A. Schuler).
34. Olin Mathieson, US 2 943 086, 1960 (H. L. Yale, J. Bernstein).
35. Merck, GB 829 763, 1960.
36. J. F. Muren, B. M. Bloom, *J. Med. Chem.* **13** (1970) 17.
37. Smith, Kline and French, GB 925 538, 1963.
38. Kefalas, BE 816 855, 1973 (N. Lasson).
39. P. A. J. Janssen, BE 577 977, 1959.
40. P. A. J. Janssen et al., *J. Med. Pharm. Chem.* **1** (1959) 281.
41. P. A. J. Janssen, US 5 967 472, 1959.
42. Janssen, GB 895 309, 1962 (P. A. J. Janssen).
43. Janssen, BE 610 830, 1962 (P. A. J. Janssen).
44. Janssen, US 3 155 669, 1962.
45. Janssen, BE 626 307, 1963.
46. M. Sato et al., *J. Med. Chem.* **21** (1978) 1116.
47. Janssen, US 3 141 823, 1964 (P. A. J. Janssen, J. F. Gardocki).
48. R. J. Gould et al., *Proc. Natl. Acad. Sci. U.S.A.* **80** (1983) 5122.
49. Ferrosan, BE 651 144, 1964.
50. Janssen, BE 633 914, 1963.
51. K. Sindelár et al., *Collect. Czech. Chem. Commun.* **38** (1973) 3879.
52. Janssen, FR M 3695, 1965 (P. A. J. Janssen).
53. M. Nakanishi, *J. Med. Chem.* **13** (1970) 644.
54. Dr A. Wander, FR CAM 51, 1964.
55. F. Hunziker, *Helv. Chem. Acta* **50** (1967) 1588.
56. H. H. M. Van Tol et al., *Nature* **350** (1991) 610.
57. J. Schmutz, *Helv. Chim. Acta* **50** (1967) 245.
58. P. Jenner, C. Marsden, *Life Sci.* **25** (1979) 479.
59. F. J. White, R. Y. Wang, *Science (Washington D.C.)* **221** (1983) 1054.
60. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, FR 2 187 309, 1971.
61. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, DE 2 327 192, 1972.
62. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, BE 872 585, 1978 (M. Thominet, J. Acker, J. C. Monier).
63. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, FR 2 480 930, 1980 (A. P. G. Coubes, R. E. M. Poubes).
64. L. Florvall, S. O. Ogren, *J. Med. Chem.* **25** (1982) 1290.
65. Gallardo Antonio, DE 2 513 136, 1975 (R. G. Spickett et al.).
66. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, BE 620 543, 1962.
67. Société d'Etudes Scientifiques et Industrielles de l'Ile de France, NL 7 707 982, 1976.
68. F. T. Bruderlein, L. G. Humber, *J. Med. Chem.* **18** (1975) 185; *Can. J. Chem.* **52** (1974) 2119.
69. Endo Laboratories, BE 670 798, 1966.
70. S. Archer et al., *J. Am. Chem. Soc.* **84** (1962) 1306.
71. G. P. Reynolds, *Trends Pharmacol. Sci.* **13** (1992) 116.
72. C. L. Broekkamp et al., *Arzneim. Forsch.* **40** (1990) 544.
73. M. Tricklebank, *Trends Pharmacol. Sci.* **10** (1989) 127.
74. B. S. Glaeser et al., *Drug Dev. Res.* **18** (1989) 191.
75. L. D. Wise, T. G. Heffner, *Ann. Rep. Med. Chem.* **26** (1991) 53.
76. R. Kuhn, *Am. J. Psychiatry* **115** (1958) 459.
77. N. S. Kline, *J. Clin. Exp. Psychopathol Q. Rev. Psychiatry Neurol.* **19** (1958) 72.
78. M. Bourin, *Fundam. Clin. Pharmacol.* **4** (1990) 49.
79. T. M. Itil in W. G. Clark, J. Del Guidice (eds.): *Principles of Psychopharmacology,* Academic Press, New York 1978, pp. 261 – 277.
80. V. G. Longo in W. G. Clark, J. Del Guidice (eds.): *Principles of Psychopharmacology,* Academic Press, New York 1978, pp. 247 – 260.
81. S. Ankier in G. P. Ellis, G. B. West (eds.): *Progress in Medicinal Chemistry,* **vol. 23,** Elsevier, Amsterdam 1986, pp. 121 – 185.

82. R. J. Baldessarini in A. G. Goodman et al. (eds.): *The Pharmacological Basis of Therapeutics,* 7th ed., Chap. 19, McMillan, New York 1985, pp. 387 – 445.
83. A. C. Cope, S. W. Fenton, *J. Am. Chem. Soc.* **73** (1951) 1673.
84. Hoffmann-La Roche, GB 858 186, 1958.
85. Hoffmann-La Roche, DT 1 109 166, GB 858 187, 1961.
86. S. G. Jue et al., *Drugs* **24** (1982) 1.
87. P. N. Craig et al., *J. Org. Chem.* **26** (1961) 135.
88. M. Protiva et al., *Experientia* **18** (1962) 326. F. Gadient et al., *Helv. Chim. Acta* **45** (1962) 1860.
89. K. Stach, F. Bickelhaupt, *Monatsh. Chem.* **93** (1962) 896. F. Bickelhaupt et al., *Monatsh. Chem.* **95** (1964) 485.
90. W. Schindler, F. Häfliger, *Helv. Chim. Acta* **37** (1954) 472.
91. E. Eriksoo, O. Rohte, *Arzneim. Forsch.* **20** (1970) 1561.
92. *Drugs Future* **8** (1981) 479.
93. R. D. Hoffsommer et al., *J. Org. Chem.* **27** (1962) 4134.
94. E. L. Engelhardt et al., *J. Med. Chem.* **11** (1968) 325.
95. S. O. Winthrop et al., *J. Org. Chem.* **27** (1962) 230.
96. Société General de Rech. et d'Appl. (Sogeras), DE 2 030 492, 1969 (C. Gueremy, P. C. Wirth); *Chem. Abstr.* **74** (1971) 141 581 e.
97. Science Union et Cie Soc., FR 2 037 266, 2 104 728, 2 104 729, 1969 (C. Malen et al.).
98. Science Union et Cie Soc., DE 2 011 806, 1969 (C. Malen et al.).
99. G. Aichinger et al., *Arzneim. Forsch.* **19** (1969) 838.
100. F. Hunziker et al., *Arzneim. Forsch.* **13** (1963) 324.
101. H. Engelhard et al., DE 1 198 353, 1964.
102. Rhône-Poulenc S.A., FR M 209, 1961; *Chem. Abstr.* **58** (1963) 3442 f. Geigy, DT 1 133 729, 1962;CH 359 143, 360 061, 1962 (W. Schindler).
103. C. Hoffmann, A. Faure, *Bull. Soc. Chim. Fr.* 1966, 2316.
104. R. M. Jacob, M. Messer, *C. R. Hebd. Seances Acad. Sci.* **252** (1961) 2117.
105. D. J. King, *Irish Med. J.* **73** (1980) 366.
106. M. V. Rudorfer, W. Z. Potter, *Drugs* **37** (1989) 713.
107. P. J. McGrath et al., *J. Clin. Psychopharmacol.* **7** (1987) 335.
108. J. Maj et al., *Psychopharmacology* **59** (1978) 79.
109. J. M. Fludder, B. E. Leonard, *Psychopharmacology* **64** (1979) 329.
110. M. Wilhelm, P. Schmidt, *Helv. Chim. Acta* **52** (1969) 1385.
111. W. J. Van der Burg et al., *J. Med. Chem.* **13** (1970) 35.
112. J. P. M. Findberg, M. B. H. Youdim, *Neuropharmacology* **22** (1983) 441.
113. G. F. Bettinetti, *Farmaco Ed. Sci.* **16** (1961) 823; *Chem. Abstr.* **59** (1963) 2716 a.
114. D. Libermann, *Bull. Soc. Chim. Fr.* 1961, 1952.
115. F. H. McMillan et al., *J. Am. Pharm. Assoc. Sci. Ed.* **42** (1953) 457.
116. H. H. Fox, J. T. Gibas, *J. Org. Chem.* **18** (1953) 994.
117. Hoffmann-La Roche, US 2 908 688, 1959 (T. S. Gardner, E. Wenis). T. S. Gardner et al., *J. Med Pharm. Chem.* **2** (1960) 133.
118. Pfizer, US 2 894 972, 1959; 3 040 061, 1962 (B. M. Bloom, R. E. Carnahan).
119. E. Votocek, O. Leminger, *Collect. Czech. Chem. Commun.* **4** (1932) 271. J. H. Biel et al., *J. Am. Chem. Soc.* **81** (1959) 2805.
120. L. Bernardi et al., *Experientia* **24** (1968) 774.
121. *Drug future* **9** (1984) 587.
122. Abbott Labs., GB 906 245, 1959; US 3 155 584, 1962.
123. J. Castañer, *Drugs Future* **4** (1979) 128.
124. A. Burger, W. L. Yost, *J. Am. Chem. Soc.* **70** (1948) 2198.
125. *Drugs Future* **9** (1984) 195.
126. M. Grau, *Drugs Future* **8** (1983) 124.
127. P. J. Thorpe, *Drugs Future* **5** (1980) 39.
128. C. Fauran et al., *Chim. Ther.* **8** (1973) 324.
129. J. Mendels, *J. Clin. Psychiatry Suppl.* **48** (1987) 26.
130. P. Benfield et al., *Drugs* **32** (1986) 481.
131. S. A. Montgomery, *Acta Psychiatr. Scand. Suppl.* **350** (1989) 107.
132. J. Zohar et al., *Arch. Gen. Psychiatry* **45** (1988) 167.
133. A. J. Bigler et al., *Eur. J. Med. Chem. Chim. Ther.* **12** (1977) 289; *Drugs Future* **4** (1979) 407.
134. A.-L. Nyth et al., *Psychopharmacology Suppl.* **96** (1988) 268.
135. Ferrosan, BE 810 310, 1973 (J. A. Christensen, R. F. Squires).
136. Eli Lilly, DE 2 500 110, 1974 (B. B. Malloy, K. K. Schmiegel). J. Castaner, *Drugs Future* **2** (1977) 27.
137. Philips Gloeilampenfabriken, US 4 085 225, 1975 (H. B. A. Welle, V. Classen).

138. M. N. Serradell, P. Blancafort, *Drugs Future* **3** (1978) 667. N. Busch et al., *Eur. J. Med. Chem. Chim. Ther.* **11** (1976) 201.
139. *Drugs Future* **11** (1986) 112. Ferrosan, DE 2 404 113, 1973 (J. A. Christensen, R. F. Squires).
140. *Drugs Future* **9** (1984) 277. W. M. Welch et al., *J. Med. Chem.* **27** (1984) 1508.
141. M. J. Schou, *J. Psychiatric. Res.* **6** (1968) 67.
142. J. W. Jefferson et al., *Am. J. Psychiatry* **139** (1982) 1593.
143. S. L. Treiser et al., *Science (Washington D.C.)* **213** (1981) 1529.
144. M. J. Berridge et al., *Biochem J.* **206** (1982) 587.
145. Wellcome Foundation, DE 2 059 618, 1969 (N. B. Mehta, D. A. Yeowell).
146. A. Brunet et al., *Bull Soc. Chim. Fr.* 1967, no. 6, 2000.
147. Aziende Chimiche Riunite, US 3 381 009, 1966 (A. Francesco).
148. M. V. Rudorfer et al., *Biol. Psychiatry* **19** (1984) 179.
149. ICI, US 3 714 161, 1966.
150. D. Dunner et al., *J. Clin. Psychopharmacol.* **7** (1987) 170.
151. *Proceedings of the 17th Collegium Internationale Neuropsychopharmacologicum Congress,* Kyoto 1990, p. 200.
152. J. M. Cott et al., *Psychopharmacol. Bull.* **24** (1988) 164.
153. M. Fink, P. Irvin, *Psychopharmacology* **78,** (1982) 44.
154. G. F. Hebenstriet et al., *Pharmacopsychiatry* **22** (1989) 156.
155. F. Eckmann et al., *Curr. Ther. Res.* **43** (1988) 291.
156. G. Zbinden, L. O. Randall, *Adv. Pharmacol.* **5** (1967) 213.
157. W. Haefely, P. Polc in R. W. Olsen, J. C. Venter (eds.) : *Benzodiazepine/GABA Receptors and Chloride Channels: Structural and Functional Properties,* A. R. Liss, New York 1986, pp. 99 – 133.
158. J. M. Stutzmann et al.: *Poster in Neurosciences,* 1992.
159. M. Briley, P. Chopin in M. Sandler, A. Coppen and S. Harnett (eds.): *5-Hydroxytryptamine in Psychiatry,* Medical Publ., Oxford 1991, pp. 177 – 197.
160. I. Geller, J. Seifter, *Psychopharmacologia* **1** (1960) 482.
161. A. Boireau et al., *Eur. J. Pharmacol.* **104** (1984) 139.
162. P. Polc et al., *Naunyn Schmiedebergs Arch. Pharmacol.* 1974, 284.
163. H. Mohler, T. Okada, *Science (Washington D.C.)* **198** (1977) 849.
164. R. F. Squires, C. Braestrup, *Nature* **266** (1977) 732.
165. E. S. Levitan et al., *Nature* **335** (1988) 76.
166. P. R. Schofield et al., *Nature* **328** (1987) 221.
167. D. B. Pritchett, P. H. Seeburg, *J. Neurochem.* **54** (1990) 1802.
168. Allied Chemical Corp., US 2 893 992, 1975 (C. Swayambu). L. H. Sternbach, E. Reeder, *J. Org. Chem.* **26** (1961) 1111.
169. L. H. Sternbach, E. Reeder, *J. Org. Chem.* **26** (1961) 4936. Hoffmann-La Roche, US 3 371 085, 1960 (L. H. Sternbach, E. Reeder).
170. Hoffmann-La Roche, BE 619 101, 1961 (L. Berger et al.); US 3 100 770, 1961 (R. I. Fryer et al.).
171. American Home Products, US 3 296 249, 1963 (S. C. Bell). S. C. Bell, *J. Med. Chem.* **11** (1968) 457.
172. S. C. Bell, S. J. Childress, *J. Org. Chem.* **27** (1962) 1691. Hoffmann-La Roche, US 3 109 843; US 3 340 253, 1962 (E. Reeder, L. H. Sternbach).
173. Clin-Byla, FR AD 91 403, 1964 (J. Schmitt); US 3 516 988, 1965.J. Schmitt et al., *Chim. Ther.* **4** (1969) 239.
174. Parke Davis, US 3 192 199, 1964 (F. H. McMillan, I. Pattison). S. Inaba et al., *Chem. Pharm. Bull.* **17** (1969) 1263.
175. Siphar, DE 2 142 181, 1970 (G. Ferrari, C. Casagrande), US 3 799 920, 1970.
176. Yoshitomi Pharmaceutical Ind., US 3 904 641, 1971 (M. Nakanishi et al.). M. T. Tahara et al., *Arzneim. Forsch.* **28** (1978) 1153.
177. Upjohn, US 4 250 094, 1979 (J. B. Hester Jr.). J. B. Hester Jr., *J. Heterocycl. Chem.* **17** (1980) 575.
178. Upjohn, US 3 987 052, 1969 (J. B. Hester Jr.). J. B. Hester Jr. et al., *Tetrahedron Lett.* 1971, 1609. A. Walser, G. Zenchoff, *J. Med. Chem.* **20** (1977) 1694.
179. Upjohn, US 3 573 282, 1969 (J. Szmuszkovicz). J. Szmuszkovicz et al., *Tetrahedron Lett.* 1971, 3665.
180. Yoshitomi Pharmaceutical Ind., US 3 849 405, 1970 (M. Nakanishi et al.). M. Nakanishi et al., *J. Med. Chem.* **16** (1973) 214.
181. F. J. Tinney et al., *J. Med. Chem.* **17** (1974) 624.
182. Schering Corp., US 3 429 874, 1967 (G. J. Topliss). M. Steinman et al., *J. Med. Chem.* **16** (1973) 1354.

183. Gyogyszerkutato Intezet, HU 155 572, 1969 (J. Korosi et al.). ; Chem. Abstr. 70 (1969) 115 026 a. J. Korosi, T. Lang, *Chem. Ber.* **107** (1974) 3883.
184. Boehringer Sohn CH, US 3 684 798, 1968 (K.-H. Weber et al.). S. Rossi et al., *Chim. Ind. (Milan)* **51** (1969) 479. K.-H. Weber et al., *Liebigs Ann.* **756** (1972) 128.
185. B. J. Ludwig, E. C. Piech, *J. Am. Chem. Soc.* **73** (1951) 5779. Carter Products, US 2 724 720, 1960 (F. M. Berger, B. J. Ludwig).
186. J. W. Phillis, R. E. Delong, *Eur. J. Pharmacol.* **101** (1984) 295.
187. Armour-Pharmaceutical, GB 837 718, 1962 (R. H. Sifferd, L. D. Braitberg).
188. Carter Products, US 2 937 119, 1960 (F. M. Berger, B. J. Ludwig).
189. Armour-Pharmaceutical, US 3 066 164, 1962 (R. H. Sifferd, L. D. Braitberg).
190. H. G. Morren, BE 523 899, 1954.
191. M. Neuman, *Drugs Future* **6** (1981) 550. Hoechst, US 3 725 404, 1966 (H. Kuch et al.).
192. H. J. Kruse, H. Kuch, *Arzneim. Forsch.* **35** (1985) 133.
193. M. Dimsdale et al., *Drugs Future* **13** (1988) 106. Synthelabo, EP 50 563, 1980;US 4 382 938, 1980 (P. Kaplan, P. George).
194. B. Zivkovic et al., *Eur. Neuropsychopharmacol.* **1** (1991) no. 1, 202.
195. K. L. Goa, A. Ward, *Drugs* **32** (1986) 114.
196. Y. H. Wu et al., *J. Med. Chem.* **15** (1972) 477. Bristol-Myers, DE 2 057 845, 1969;US 3 717 634, 1969 (Y. H. Wu, J. W. Rayburn).
197. D. P. Taylor, *FASEB J.* **2** (1989) 2445.
198. W. Haefely et al., *Trends Pharmacol. Sci.* **11** (1990) 452.
199. *Scrip* **1655** (1991) 26.
200. *Scrip* **1489** (1990) 28.
201. M. Williams in B. S. Meldrum, M. Williams (eds.): *Current and Future Trends in Anticonvulsant, Anxiety and Stroke Therapy,* Wiley-Liss, New York 1990, p. 131.
202. J. Hughes et al., *Proc. Natl. Acad. Sci. U.S.A.* **87** (1990) 6728. D. C. Horwell et al., *J. Med. Chem.* **34** (1991) 404.
203. N. M. Barnes, *Neuroreport* **1** (1990) 20.
204. R. D. Porsolt, S. Roux, A. Lenègre in : G. Racagni, J. Mendlewicz (eds.): *Treatment of Age-Related Cognitive Dysfunction: Pharmacological and Clinical Evaluation.* Karger, Basel 1992, p. 50.
205. C. E. Giurgea, *Drug Dev. Res.* **2** (1982) 441.
206. UCB, NL 6 509 994, 1964, US 3 459 738, 1964 (H. Morren).
207. Hoffmann-La Roche, US 4 369 139, 1978 (E. Kyburz, W. Aschwanden).
208. J. S. Isaacson, R. A. Nicoll, *Proc. Natl. Acad. Sci. U.S.A.* **88** (1991) 10 936.
209. I.S.F., DE 2 635 853, 1975 (G. Pifferi); DE 2 758 937, 1977 (S. Banfi et al.).
210. L. L. Iversen, *Trends Pharmacol. Sci.* **7** (1986) 44.
211. L. Knorr, *Ber. Dtsch. Chem. Ges.* **37** (1904) 3507.
212. Y. Dormard et al., *Arzneim. Forsch.* **25** (1975) 194.
213. G. Thuillier, P. Rumpf, *Bull. Soc. Chim. Fr.* 1960, 1786.
214. E. P. Kennedy, *J. Biol. Chem.* **222** (1956) 185.
215. K. Kikugawa, M. Ichino, *Chem. Pharm. Bull.* **19** (1971) 1011.
216. F. Chemnitius, *J. Prakt. Chem.* **116** (1927) 59.
217. C. L. Murray, H. C. Fibiger, *Neuroscience (Oxford)* **14** (1985) 1025.
218. J. A. Moore, L. D. Kornreich, *Tetrahedron Lett.* 1963 1277.
219. W. K. Summers et al., *New Engl. J. Med.* **315** (1986) 1241.
220. F. Matsuda et al., *Tetrahedron* **41** (1985) 3625.
221. T. Szirtes et al., *J. Med. Chem.* **29** (1986) 1654.
222. K. Okamoto, *Chem. Pharm. Bull.* **30** (1982) 2797.
223. Soc. Farmitalia Italia, DE 2 112 273, 1970.
224. C. D. Nicholson, *Psychopharmacology* **101** (1990) 147.
225. Mitsubishi, DE 2 627 227, 1976 (R. Kikumoto et al.).
226. K. I. Saito et al., *Jpn. J. Pharmacol.* **39** (1985) 406.
227. C. Loerincz et al., *Arzneim. Forsch.* **26** (1976) 1907.
228. Farbwerke Hoechst, DE 2 330 742, 1973.
229. V. Stephanovich, *Drug Dev. Res.* **6** (1985) 327.
230. Merck, US 3 010 966, 1961 (O. Zima, G. Schorre).
231. S. Hoyer et al., *Arzneim. Forsch.* **27** (1977) 671.
232. Pharmascience, DE 2 501 443, 1974 (J. M. Gazave et al.).
233. R. D. Porsolt et al., *Arzneim. Forsch.* **37** (1987) 388.
234. T. Kojima et al., *Chem. Pharm. Bull.* **33** (1985) 3766.
235. G. Calderini et al., *Ann. N. Y. Acad. Sci.* **444** (1985) 504.
236. R. Krimberg, W. Wittandt, *Biochem. Z.* **251** (1932) 229.
237. E. Tempesta et al., *Neuropharmacology* **24** (1985) 41.

238. H. Meyer et al., *Arzneim. Forsch.* **31** (1981) 407.
239. A. Scriabine et al., *FASEB J.* **3** (1989) 1799.
240. A. Wauquier et al., *Drug. Dev. Res.* **8** (1986) 373.

Pulp → Paper and Pulp
Pumice → Silicates
Pumps → High-Pressure Technology
Pumps → Vacuum Technology

Purine Derivatives

HANS H. LENZ, Knoll AG, Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **Properties** 1
3. **Occurrence and Production** 2
4. **Quality Specifications and Analysis** . . . 4
5. **Uses** 4
6. **Economic Aspects** 5
7. **References** 5

1. Introduction

Purine [*120-73-0*], 7*H*-imidazo[4,5-*d*]pyrimidine, M_r 120.12, is a bicyclic molecule composed of the fused heterocycles imidazole and pyrimidine.

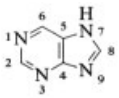

The name purine (from "purum uricum") as well as the ring-numbering technique still practiced today can be attributed to EMIL FISCHER. Oxygen-containing purine derivatives are often represented in the isomeric lactam form, as with xanthine (**1**):

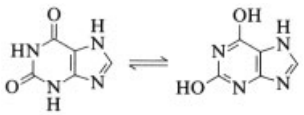

1

Infrared spectra reveal that oxygen-containing purine derivatives (including hypoxanthine and uric acid) exist mainly as lactams whereas free amine groups can be detected with aminopurines (e.g., adenine) [1]. Purine derivatives with an *N*-methylated pyrimidine ring (e.g., from compounds like xanthine) are also based on the lactam form.

Purines were among the first natural products identified. Thus, in 1776 SCHEELE extracted a substance from urinary calculi which FOURCROY named uric acid (*acide ourique*) in 1793. In 1834, LIEBIG and MITSCHERLICH established its correct empirical formula as $C_5H_4N_4O_3$. RUNGE isolated caffeine from coffee beans in 1820, and in 1844 UNGER isolated guanine from guano. These and other studies led MEDICUS in 1875 to suggest a bicyclic structure for uric acid and its derivatives. Thorough investigations by EMIL FISCHER (1882 – 1897) clarified the relationships linking these compounds, and in 1884 the current name was assigned to the series of substances and the Medicus structure was confirmed. At about the same time A. KOSSEL isolated guanine (1883) and adenine (1885) from nucleic acids. Finally, around the turn of the century, this initial research era was brought to a close with a series of total syntheses by W. TRAUBE [2]. Extensive work was later resumed by BREDERECK [3].

2. Properties

Physical Properties. Purine and its *C*-alkyl derivatives are very soluble in water. Purines containing hydroxyl groups show only poor solubility in water and organic solvents, whereas aminopurines are fairly soluble. With the exception of uric acid, the compounds are amphoteric, yielding water-soluble salts with both acids and bases. The acidic character of purine derivatives increases with the number of hydroxyl groups in the molecule. The dissociation constant of uric acid (pK_a 5.7) is almost comparable to that of aliphatic carboxylic acids; aminopurines, on the other hand, are more basic. The melting points of purines are often very high, and they are of limited use for characterization. Melting points of several purines are listed in Table 1.

Chemical Properties. Purine derivatives undergo nucleophilic substitution at the 2-, 6-, and 8-positions due to delocalization of the pyrimidine and imidazole π-electrons. Appropri-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_383

Table 1. Melting points of representative purines

Common name	Systematic name	CAS registry no.	M_r	*mp*, °C
Purine	7*H*-imidazo[4,5*d*]pyrimidine	[*120-73-0*]	120.1	217
Adenine	6-aminopurine	[*73-24-5*]	135.1	360 – 365 (decomp.)
Guanine	2-amino-6-hydroxypurine	[*73-40-5*]	151.1	365 (decomp.)
Hypoxanthine	6-hydroxypurine	[*68-94-0*]	136.1	
Mercaptopurine	6-mercaptopurine	[*50-44-2*]	152.2	313 – 314 (decomp.)
Thioguanine	2-amino-6-mercaptopurine	[*154-42-7*]	167.2	>360
Xanthine	2,6-dihydroxypurine	[*69-89-6*]	152.1	
3-Methylxanthine		[*1076-22-8*]	165.1	380 (decomp.)
Theophylline	1,3-dimethylxanthine	[*58-55-9*]	179.2	269 – 272
monohydrate		[*5967-84-0*]	197.2	
Theobromine	3,7-dimethylxanthine	[*83-67-0*]	179.2	351
Caffeine	1,3,7-trimethylxanthine	[*58-08-2*]	193.2	234 – 235
monohydrate		[*5743-12-4*]	211.2	
Uric acid	2,6,8-trihydroxypurine	[*69-93-2*]	168.1	

ately activated derivatives are also subject to electrophilic substitution at position 8. Derivatives free of substituents in the 7-position readily yield salts under alkaline conditions (e.g., sodium theophylline). These salts remain very stable to hydrolytic ring cleavage. On the other hand, the pyrimidine ring in 1,3,7-peralkylated xanthines (e.g., caffeine) is highly susceptible to basic cleavage. For more information see [4–6].

3. Occurrence and Production

Purine derivatives are widespread in nature. Adenine and guanine, as components of nucleic acids and nucleosides, are fundamental building blocks of life (→ Nucleic Acids). Purine itself is a component of the *N*-nucleoside nebularine [*550-33-4*] [7], which displays antibiotic properties (→ Antibiotics). *N*-Alkylpurinediones (*N*-alkylxanthines) are also abundant in nature; the most common is caffeine, which occurs in concentrations of 1 – 1.5 % in coffee beans and up to 5 % in black tea leaves, as well as in a number of other plants. Caffeine is isolated by solvent extraction from green coffee beans and tea leaves. The Zosel high-pressure extraction process is a more sophisticated procedure involving supercritical carbon dioxide [8], resulting in mild, selective, and toxicologically safe decaffeination (→ Coffee, Chap. 4.). Another purine, theobromine, is found in cacoa hulls at concentrations of 1.5 – 3 %.

Purine biosynthesis is a multistep enzyme-catalyzed process based on simple building blocks. 5′-Phosphoribosylglycinamide (**2**) is first produced from α-D-ribose and glycine. Subsequent steps lead to the formation of 5′-phosphoribosyl-5-aminoimidazole (**3**), onto which the pyrimidine ring is added, again in a multistep sequence.

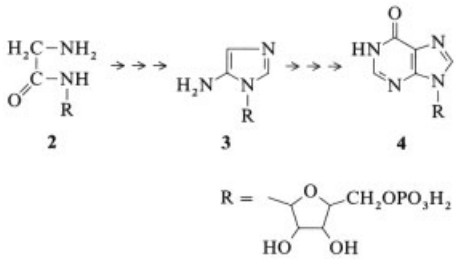

The product inosinic acid (**4**) [*131-99-7*] is a starting material for the synthesis of other purine compounds [9].

The final product of human purine metabolism is uric acid (**5**). Thus, adenine (**6**) is converted to xanthine (**1**) via hypoxanthine (**7**).

Guanine (**8**) is converted to **1** directly. Xanthine oxidase then catalyzes oxidation of xanthine to uric acid.

Physical disorders related to purine metabolism include xanthinuria, an inherited metabolic disease caused by the absence of xanthine oxidase, and gout, which leads to deposits in the joints as a result of excessive uric acid concentration in the blood (→ Gout Remedies, Chap. 1.→ Gout Remedies, Chap. 2.→ Gout Remedies, Chap. 3.).

Other organisms are capable of degrading uric acid further to allantoin(most mammals and reptiles),allantoic acid,and ultimately urea(e.g.,fish).

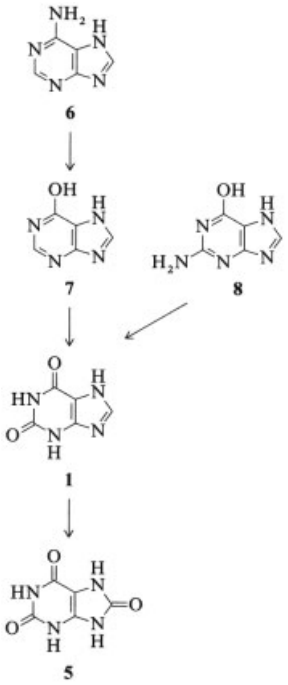

A number of synthetic pathways for purines have been described with such starting materials as hydrocyanic acid or formamide [10], but only moderate yields have been achieved. A prebiotic synthesis based on hydrocyanic acid has also been considered (e.g., for adenine, which is formally a pentamer of HCN) [11]. Industrial processes begin with simple starting materials. A stepwise approach is used to construct either the pyrimidine or imidazole ring, onto which is subsequently fused the second heterocycle. A synthetic approach originally developed by TRAUBE [2] and later subjected to several improvements is well-adapted to the production of the commercially important xanthine derivatives caffeine, theophylline, and theobromine. A series of steps starting with *N,N′*-dimethylurea [12, 13] and cyanoacetic acid leads to theophylline (**9**), which can be methylated to give caffeine (**10**):

The same reaction path starting with monomethylurea [12] yields 3-methylxanthine (**11**), which undergoes preferential methylation to theobromine (**12**).

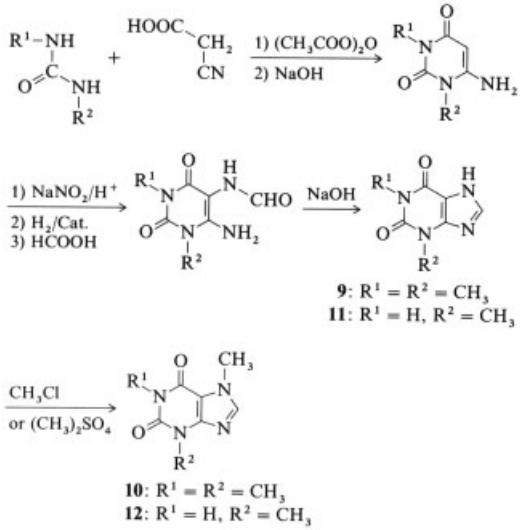

Other known routes to xanthine derivatives are not utilized commercially, either because the starting materials are expensive or because overall yields are less satisfactory (e.g., syntheses starting from imidazole derivatives, or the direct cyclization of 1,3-dimethyl-6-methylamino-5-nitrosouracil [14]).

Guanine and its derivatives can also be prepared from suitably substituted pyrimidine derivatives [15]. Guanine, adenine, xanthine, and hypoxanthine can be made from appropriate imidazole derivatives [16]:

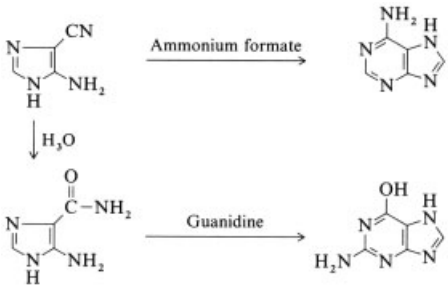

Purine derivatives are also accessible from arylazomalonic acid derivatives [17]:

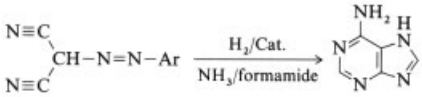

4. Quality Specifications and Analysis

Since purine derivatives find their most important use as pharmacological agents, quality standards and analytical practices are subject to the regulations of pharmacopoeia (USP, BP, Ph. Eur., DAB, JP, etc.). The food industry also generally employs materials that comply with the specifications of the pharmacopoeia (e.g., caffeine for soft drinks). Purine derivatives employed as pharmacological agents must be prepared according to current good manufacturing practice (GMP) specifications. Product control must always comply with relevant national laws and international agreements (e.g., the Pharmaceutical Inspection Convention). Purity control by thin layer chromatography, customary in the past, is increasingly being replaced by high-pressure liquid chromatography.

5. Uses

Due to their broad spectrum of biological activity, many purine derivatives find application in medicine. Theophylline was previously utilized as a diuretic, but it no longer plays a role in this type of therapy. The most important use for theophylline and certain of its salts (e.g., theophylline ethylenediamine [*317-34-0*], theophylline cholinate [*4499-40-5*]), and such derivatives as diprophylline (**13**) [*479-18-5*] [18], proxyphylline (**14**) [*603-00-9*] [19], etofylline (**15**) [*519-37-9*], [18], or bamifylline (**16**) [*2016-63-9*] [20] involves their bronchodilating antiasthmatic properties (→ Antiasthmatic Agents, Chap. 2.2.). More recent pharmaceutical developments have resulted in a trend toward increasing use of theophylline itself rather than its derivatives.

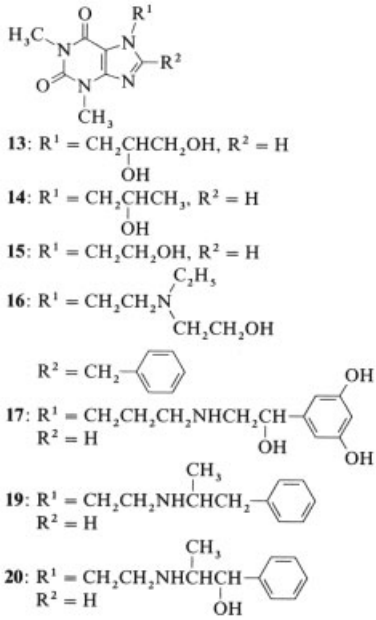

Another theophylline derivative, reproterol (**17**) [*54063-54-6*] [21], in which the xanthine skeleton is combined with a sympathomimetically active structure, shows only sympathomimetic action (→ Antiasthmatic Agents). Pentoxifylline (**18**) [*6493-05-6*], a theobromine derivative, is employed as a vasodilator [22] (→ Cardiovascular Drugs).

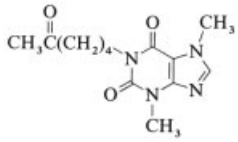

18

Fenetylline (**19**) [*3736-08-1*] is used as a CNS stimulant [23, 24]; cafedrine (**20**) [*58166-83-9*] as an analeptic agent [24, 25]. Dimenhydrinate [*523-87-5*], the salt of 8-chlorotheophylline [*85-18-7*] [26] and diphenhydramine [*58-73-1*] [26], is a commonly employed antiemetic (→ Antiemetics).

A number of guanine and adenine derivatives are employed as chemotherapeutic agents in treatment of the herpes virus, including acyclovir [*59277-89-3*] [27] and vidarabin [*24356-66-9*] [28] (→ Chemotherapeutics,

Chap. 4.6.1.). 6-Mercaptopurine [29] and 6-thioguanine [30] are applied in the treatment of leukemia (→ Cancer Chemotherapy, Chap. 2.7.).

Caffeine is the purine derivative produced industrially in the greatest quantity. Caffeine stimulates the central nervous system, causing increases in cardiac activity, metabolism, respiration, and blood pressure. It is frequently utilized in combination with analgesics (e.g., acetylsalicylic acid), thereby enhancing the effect of the analgesic [31].

The cosmetic industry exploits caffeine's ability to increase the flow of blood (e.g., in skin cosmetics). Another industrial use relates to the manufacture of copying papers (diazo paper). More important economically is caffeine's use in the soft drink industry. Cola beverages typically contain the compound at a level of about 100 mg/L (→ Beverages, Nonalcoholic, Chap. 4.1.2.→ Beverages, Nonalcoholic, Chap. 4.2.2.). An extensive discussion and evaluation of caffeine metabolism and of potential health hazards associated with its use in beverages and pharmaceuticals is given in [32].

6. Economic Aspects

The worldwide demand for caffeine is ca. 10 000 t/a. Approximately one-third of the total demand is covered by naturally occurring caffeine, the rest by synthetic material. The major share of synthetic caffeine is produced in Germany, with smaller amounts in the United States, Mexico, and the People's Republic of China.

The second most widely distributed purine derivative, theophylline, is entirely synthetic. By far the major portion of the annual demand (ca. 2500 t) is supplied by Germany. The major suppliers of caffeine and theophylline are Knoll and Boehringer Ingelheim.

Theobromine is used mainly in the production of pentoxifylline. The major demand is met by synthetic material and a small part by extraction of cacoa hulls.

7. References

1. C. L. Angell, *J. Chem. Soc.* 1961, 504.
2. W. Traube, *Ber. Dtsch. Chem. Ges.* **33** (1900) 3035.
3. H. Bredereck, H. G. v. Schuh, A. Martini, *Chem. Ber.* **83** (1950) 201. H. Bredereck, F. Effenberger, H. G. österlin, *Chem. Ber.* **100** (1967) 2280.
4. J. H. Lister in D. J. Brown (ed.): *Heterocyclic Chemistry, Fused Pyrimidines,* Part II, Wiley-Interscience, New York 1971.
5. G. Shaw in S. Coffey (ed.): *Rodd's Chemistry of Carbon Compounds,* 2nd ed., **vol. 4,** part L, Elsevier, Amsterdam 1980, pp. 1 – 100.
6. A. R. Katritzky, C. W. Rees in K. T. Potts (ed.): *Comprehensive Heterocyclic Chemistry,* **vol. 5,** part 4 A, Pergamon Press, London 1984, pp. 499 – 605.
7. N. M. Löfgren, B. Lüning, *Acta Chem. Scand* **7** (1953) 15, 225; **8** (1954) 670. K. Isono, S. Suzuki, *J. Antibiot. Ser. A* **13** (1960) 270.
8. Studiengesellschaft Kohle, DE 2 005 293, 1970 (K. Zosel). HAG, DE 2 119 678, 1971 (W. Roselius). HAG, DE 2 127 642, 1971 (O. Vitzhum). Studiengesellschaft Kohle, AT 4 4003-71, 1971 (K. Zosel).
9. J. H. Lister, *Rev. Pure Appl. Chem.* **13** (1963) 30.
10. Ethyl Corporation, US 4 511 716, 1985 (K. H. Shin).
11. J. Oro, A. P. Kimball, *Arch. Biochem. Biophys.* **94** (1961) 217. A. W. Schwartz, C. G. Bakker, *Science (Washington D.C.)* **245** (1989) 1102.
12. Knoll AG, DE 896 640, 1942 (K. Kraft, L. Suranyi).
13. BASF, GB 750 549, 1956.
14. H. Goldner, G. Dietz, E. Carstens, *Justus Liebigs Ann. Chem.* **691** (1965) 142.
15. Merck, DE 162 336, 1904. Hüls Troisdorf, DE 3 729 471, 1989 (M. Feld, H. a. d. Fünten, W. Voigt). C. Parkanyi, H. L. Yuan, *J. Heterocycl. Chem.* **27** (1990) 1409.
16. Sagami Chemical Research Center, DE 2 166 506, 1971 (N. Asai).
17. Kohin Co, EP 45 503, 1981 (J. Suzuki). Merck & Co., US 4 092 314, 1978 (M. L. Vander Zwan, D. F. Reinhold).
18. H. J. Roth, *Arch. Pharm. (Weinheim Ger.)* **292** (1959) 234.
19. Ganes Chem. Works, US 2 715 125, 1955 (R. V. Rice).
20. A. Christiaens, BE 602 888, 1961 (R. de Ridder).
21. K. H. Klingler, *Arzneim. Forsch.* **27** (1977) 4.
22. Chemische Werke Albert, DE 1 235 320, 1964 (W. Mohler, M. Reiser, K. Popendiker); DE 2 234 202, 1972 (A. Söder); 2 302 772,

1973 (G. Nesemann, A. Söder, H. Thurm); 2 330 741, 1973 (A. Söder).
23. Degussa, DE 1 123 329, 1958 (E. Kohlstaedt, K. H. Klingler).
24. Degussa, US 3 029 239, 1962 (E. Kohlstaedt, K. H. Klingler).
25. Degussa, DE 1 095 285, 1956 (E. Kohlstaedt, K. H. Klingler).
26. Searle, US 2 499 058, 1950; 2 534 813, 1950 (J. W. Cusic).
27. H. J. Schaeffer et al., *Nature (London)* **272** (1978) 583. Wellcome Foundation, BE 833 006, 1975.
28. W. W. Lee et al., *J. Am. Chem. Soc.* **82** (1960) 2648. E. J. Reist et al., *J. Org. Chem.* **27** (1962) 3274; **29** (1964) 3725.
29. G. B. Elion, E. Burgi, G. H. Hitchings, *J. Am. Chem. Soc.* **74** (1952) 411.
30. G. B. Elion, G. H. Hitchings, *J. Am. Chem. Soc.* **77** (1955) 1676.
31. W. T. Beaver: "Combination Analgesics," *Am. J. Med.* **77** (1984) no. 3 A, 38 – 53. E. M. Laska et al.: "Caffeine as an Analgesic Adjuvant," *JAMA J. Am. Med. Assoc.* **251** (1984) 13.
32. P. B. Dews: *Caffeine* : Perspectives from Recent Research, Springer Verlag, Berlin 1984.

Pyranthrone → Anthraquinone Dyes and Intermediates

Pyrazolone Derivatives

Hartmut Heise, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

Manfred Hintzmann, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Properties 2
2.1. Reactions 2
3. Production 3
4. Environmental Protection 5
5. Quality Specifications 6
6. Analysis 6
7. Storage and Transport 6
8. Uses 6
8.1. Pharmaceuticals 6
8.2. Dyes and Pigments 6
8.3. Crop Protection 7
9. Economic Aspects 7
10. Derivatives 7
11. Toxicology 10
12. References 10

1. Introduction

Pyrazolones (**3**) are defined as ketones of pyrazolines (**2**) which are, in turn, dihydro derivatives of pyrazole (**1**):

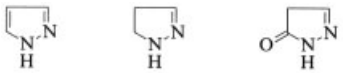

1 Pyrazole [*288-13-1*]
2 Pyrazoline [*109-98-8*]
3 5-Pyrazolone [*137-44-0*]

1-Phenyl-3-methyl-5-pyrazolone (**4**) [*89-25-8*] is industrially the most important of the various derivatives, being the basis of many dye precursors and a range of pharmaceuticals:

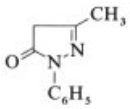

4

Replacement of the methyl group by a COOH or $COOC_2H_5$ group gives the corresponding 3-carboxy- or 3-carbethoxypyrazolones, which are important dye precursors.

Pyrazolone chemistry began in 1883 when Ludwig Knorr first reacted phenylhydrazine with acetoacetic ester [2]. He thought that he had found a new quinoline derivative, but only later did he realise that his new compound was a pyrazolone (**4**). Since pharmacological tests on this substance showed relatively weak potency, Knorr reacted the pyrazolone with methyl iodide in methanol, at the suggestion of the pharmacologist Wilhelm Filehne [3]. This new pyrazolone, named Antipyrin, was an effective analgesic and antipyretic. Thus Antipyrin formed the basis of a range of pharmaceuticals with similar effects such as Pyramidon, Novalgin, and Butazolidin.

When pyrazolones were discovered as coupling components for azo dyes in the late 1800s [4], they rapidly increased in importance. Pyrazolones are still economically important precursors for dyes and pharmaceuticals today.

2. Properties

Most pyrazolones are colorless to yellow solids with melting points of over 100°C, although some of them can be distilled. They are generally insoluble in nonpolar solvents, but soluble in polar ones. Some pyrazolones, particularly those with lower molecular masses, can be recrystallized from hot water. The UV and IR spectra of pyrazolones are very complex and not very meaningful because of tautomerism. Pyrazolones possess both basic and acidic character. The weakly acidic character generally predominates, so they can be titrated with strong bases. However, some pyrazolones can also be titrated with perchloric acid in glacial acetic acid. Most pyrazolones are readily soluble in aqueous alkali.

1-Phenyl-3-methyl-5-pyrazolone (**4**) can exist in 3 tautomeric forms:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_389

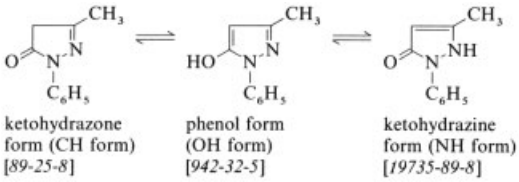

KATRITZKY [5], ELGUERO [6], and DORN [7] have carried out extensive investigations into the tautomerism of various pyrazolones. By means of UV, IR, and NMR spectroscopy and measurement of p*K* values it was shown that for 1,3-disubstituted 5-pyrazolones in aqueous solution 90 % of the NH form is in equilibrium with 10 % of the OH form, while in nonpolar solvents such as carbon tetrachloride or cyclohexane the CH form predominates. The C-unsubstituted 5-pyrazolones exist exclusively in the OH form in dimethyl sulfoxide. Electron donors on C-3 shift the equilibrium towards the CH form.

2.1. Reactions

Because of the keto – enol tautomerism the 4-position in the pyrazolone ring is particularly reactive. It exhibits the behavior typical of an active methylene group and can thus undergo condensation, alkylation, acylation, coupling, and dimerization reactions.

Alkylation. Because of the dependence of the equilibrium on the solvent and pH, 1,3-disubstituted 5-pyrazolones can be alkylated under appropriate reaction conditions at C-4 (CH form), at N-2 (NH form), and at the O atom (OH form). One of the longest known reactions of this type is the alkylation at N-2, first carried out in 1884 by KNORR, who treated **5** with methyl iodide in methanol at 100°C under pressure and obtained 1-phenyl-2,3-dimethyl-5-pyrazolone (**6**) [*60-80-0*] (Antipyrin):

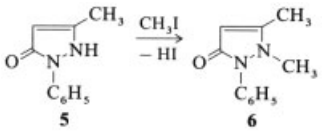

Today this methylation is carried out with methyl chloride or dimethyl sulfate. If dimethyl sulfate in an alkaline medium or methyl iodide and sodium methoxide are used [8], alkylation occurs predominantly at C-4, giving 4,4-dimethylpyrazolones. However, *N*- and *O*-alkylation are observed as side reactions.

Exclusive *O*-alkylation can be achieved by treatment with diazomethane in methanol. In this way PECHMANN [9] succeeded in producing 5-methoxy-3-methyl-1-phenylpyrazole [*27349-35-5*], which rearranges to Antipyrin (**6**) on heating to 250°C [10].

Acylation. Reaction of 1,3-disubstituted 5-pyrazolones with acetic anhydride and sodium acetate [11] or with acetyl chloride and calcium hydroxide in dioxane [12] gives the 4-acetyl compounds.

O-Acetyl compounds are obtained by treatment with acetyl chloride in benzene. If the 1-position is unsubstituted, acetylation takes place predominantly at N-1 [13].

Halogenation occurs almost exclusively at the 4-position. Passing chlorine into an acetic acid solution of a pyrazolone gives the 4,4-dichloropyrazolone. The same product is formed on treatment with phosphorus pentachloride or sodium hypochlorite. 4-Chloropyrazolones can be most easily produced by reduction of 4,4-dichloropyrazolones with hydriodic acid at room temperature.

The corresponding mono- and dibromo compounds are obtained by bromination in glacial acetic acid. The monobromo compound is formed by using an equimolar quantity of bromine. Treatment with phosphorus tribromide gives exclusively 4,4-dibromopyrazolones.

The corresponding iodo compounds can be produced with potassium iodide in aqueous alkaline solution. However, as these compounds are very unstable and light sensitive, they are difficult to obtain pure [14]. When pyrazolones are treated with phosphorus oxychloride under pressure at 150 °C, C=O cleavage occurs to give 5-chloropyrazoles [15].

Nitration. 4-Nitropyrazolones are obtained either by direct nitration under mild conditions or by oxidation of the corresponding nitroso compounds, which are readily accessible by treatment with nitrous acid.

Oxidation. Mild oxidizing agents such as arylhydrazines react with 1-aryl-3-methyl-5-pyrazolones to give 4,4′-bisarylmethylpyrazolones:

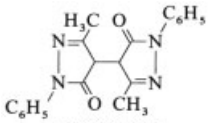

7 [*7477-67-0*]

These compounds can also be formed during pyrazolone synthesis if the arylhydrazine is present in excess.

Stronger oxidizing agents such as iron(III) chloride give the indigo dye, Pyrazole Blue (**8**).

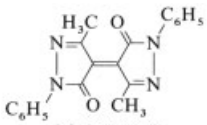

8 [*6334-24-3*]

Oxidation in alkaline solution often gives the corresponding 4-hydroxy compounds. Very strong oxidizing agents such as potassium permanganate cleave the pyrazolone ring.

Condensation Reactions. Like all compounds with a reactive methylene group, pyrazolones react at C-4 with aldehydes and ketones. Depending on the reaction conditions both mono- and bispyrazolone derivatives can be formed. Reaction with formaldehyde leads to introduction of a hydroxymethyl group:

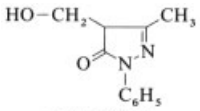

[*86467-82-5*]

Coupling Reactions. Besides alkylation, coupling reactions between pyrazolones and aromatic diazonium salts for the production of azo pigments and dyes have achieved the greatest industrial importance:

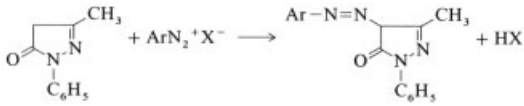

3. Production

Industrially 1-aryl-3-methyl-5-pyrazolones are produced by condensation of arylhydrazines with acetoacetic esters or acetoacetamide [*5977-14-0*]:

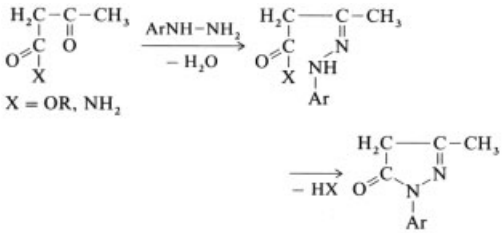

In this process, in which instead of acetoacetamide, diketene in the presence of ammonia can be used [16], the hydrazone formed initially is not isolated. Arylmethylpyrazolones are now generally produced from the corresponding amines, without isolating the intermediates:

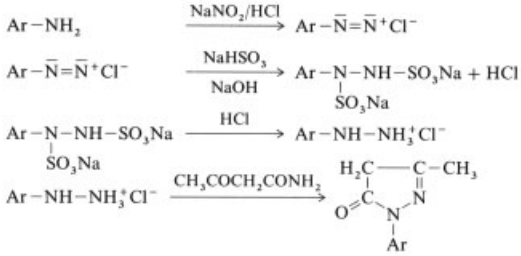

1-Aryl-5-Pyrazolone-3-Carboxylic Acids. Besides acetoacetic ester and acetoacetamide, which give methylpyrazolones, diethyl oxalacetate [*63277-17-8*] has become important for the production of 3-carboxy- and 3-carbethoxypyrazolones:

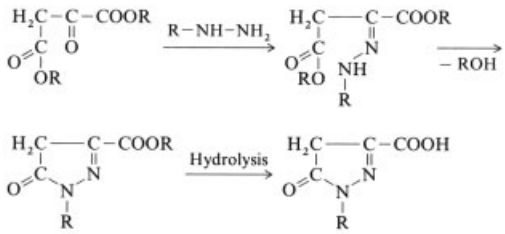

The latter can also be obtained by direct coupling of diazonium salts with esters of acetylsuccinic acid. The azo compounds formed as intermediates rearrange with cleavage of the acyl group to give the corresponding hydrazone derivatives:

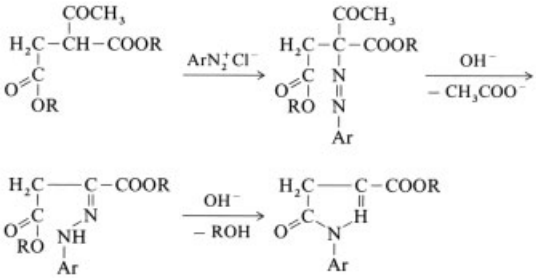

2-Aryl-3-Methyl-5-Pyrazolones. If equimolar quantities of diketene are treated with arylhydrazines in dilute acetic acid, 2-aryl-3-methyl-5-pyrazolones are obtained as the main products [17]. These have not, however, yet achieved industrial importance.

3-Aryl-1-Methyl-5-Pyrazolones (Isopyrazolones). Isopyrazolones are defined as compounds in which the positions of the aryl and methyl groups of 1-aryl-3-methyl-5-pyrazolones are exchanged. 1-Methyl-3-phenyl-5-pyrazolone [*41927-50-8*] (**9**), which melts at 214 °C, is, however, very similar to its isomer (**4**) (*mp* 128–129 °C) in its other properties. It can be obtained, for example, by warming ethyl 3-oxo-3-phenylpropionate [*94-02-0*] with methylhydrazine [*60-34-4*] [18]:

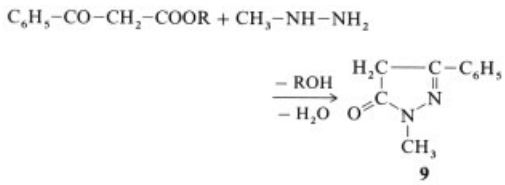

4-Amino-1-Aryl-3-Methyl-5-Pyrazolones. Arylmethylpyrazolones react with nitrous acid to give 4-isonitroso compounds, and with nitric acid to give 4-nitro compounds. Both types of compound can be reduced with tin–hydrochloric acid, zinc–glacial acetic acid, sulfite, or catalytically with hydrogen to give the corresponding 4-aminopyrazolones. The 4-amino compound can also be produced by reductive cleavage of azo dyes derived from pyrazolones. 4-Aminopyrazolones are only stable in the form of their salts. The free bases are very sensitive to oxidation. Thus, for example, 4-amino-1-phenyl-3-methyl-5-pyrazolone [*5142-72-3*] reacts further to give the red dye Rubazonic acid (**10**):

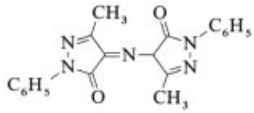

10 [*909-59-1*]

5-Iminopyrazolones. As spectroscopic investigations have shown, these compounds are actually 5-aminopyrazoles. Nevertheless they are mostly described in the literature as imino compounds in analogy to the pyrazolones. 1-Phenyl-3-methyl-5-iminopyrazolone (**11**) is formed on treatment of butadienenitrile [*1001-56-5*] with phenylhydrazine [*100-63-0*] in methanol and subsequent warming with hydrochloric acid [19]:

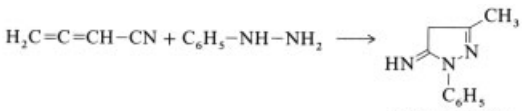

11 [*6401-97-4*]

Another method involves the reaction of the corresponding 5-chloropyrazolium chlorides with ammonia or amines. For example, the reaction of phosphorus oxychloride with Antipyrin (**6**) gives 5-chloro-2,3-dimethyl-1-phenylpyrazolium chloride. The latter can be converted into the imino compound **12** by heating with ammonia or amines at moderate temperatures. At higher temperatures methyl-group cleavage occurs to give the imino compound **13** [20].

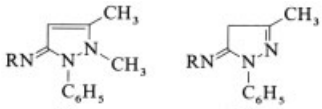

12 **13**

3,5-Dioxypyrazolidines. 1,2-Diphenyl-3,5-pyrazolidinedione [*2652-77-9*] is obtained by condensation of diethyl maleate with hydrazobenzene in the presence of sodium ethoxide [21].

4-Butyl-1,2-diphenyl-3,5-pyrazolidinedione (phenylbutazone) (**14**) and 4-butyl-1-(4-hydroxyphenyl)-2-phenyl-3,5-pyrazolidinedione (oxyphenbutazone) (**15**) also belong to the pyrazolidinedione group. As Butazolidin [*50-33-9*] and Tanderil [*129-20-4*] they have become important in the treatment of rheumatism.

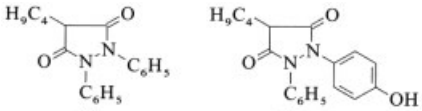

14 **15**

4. Environmental Protection

Pyrazolones are widely produced by the reaction of the corresponding amines with acetoacetic esters, acetoacetamides, or esters of acetylsuccinic acid, without isolation of the intermediate diazonium salts, hydrazines, and hydrazones. Wastewater is only produced in the final stage during the isolation of the pyrazolones. Depending on the product the wastewater has a COD of 1 – 50 g O_2/L. The COD elimination is ca. 50 % for most pyrazolones; pollution of the wastewater by hydrogen sulfate and sodium chloride poses additional problems. The biodegradability of pyrazolones is generally very poor. Some pyrazolones are even classified as toxic to bacteria and fish (EC_{50} 200 – 1000 mg/L, 3 – 6 h; LC_{50} 10 – 500 mg/L, 48 h) [22], so that special wastewater-treatment methods are necessary. For example, the mother liquors of nonbiodegradable pyrazolones can be evaporated and the residue incinerated. Oxidative cleavage of the pyrazolone ring into more readily biodegradable fragments or charcoal adsorption with subsequent incineration are conceivable. For all these processes the high salt content of the wastewater is problematic.

5. Quality Specifications

Phenylmethylpyrazolone (**4**), which is used as a starting material for various pharmaceuticals, should not contain more than 0.1 % bispyrazolone [*7477-67-0*]. The formation of this byproduct must therefore be hindered by avoiding an excess of phenylhydrazine during production. Furthermore, special quality specifications apply for various pyrazolones depending on the use. For pyrazolones which are used as precursors in dye production, a pure material content of 98 – 99 % is sufficient. Varying quantities of inorganic salts do not interfere here. Pyrazolones are further processed when both wet (water content 25 – 35 %) and dry (water content <0.5 %).

6. Analysis

The pyrazolone content can be determined by coupling with diazotized *m*-nitroaniline solution in weakly acidic, acetate-buffered solution. The end point of the titration can be recognized by spotting against H-acid solution (red coloration). The content can also be determined, by potentiometric titration with an 0.1N KBr –$KBrO_3$ solution in a hydrochloric acid medium. Other components which may be present are determined by HPLC and TLC.

7. Storage and Transport

Pyrazolones are sold as both wet and dry products. The wet product is packed in metal drums with polyethylene liners (100 or 130 kg), and the dry product also in paper sacks (25 kg) or big

bags (200 – 250 kg). Dry pyrazolones can generally be stored for an unlimited time under appropriate conditions (at least 2 years). Wet pyrazolones can be stored for at least 6 months. Pyrazolones are nonhazardous materials, so no special international regulations and classifications for ship, rail, or road transport have been imposed. However, national transport regulations, which may differ from these, must be adhered to.

8. Uses

Pyrazolones are used almost exclusively as precursors for pharmaceuticals and dyes.

8.1. Pharmaceuticals

Among the pyrazolone derivatives there are a large number of compounds with antipyretic and analgesic action. The following are some of the most important ones: phenazone (Antipyrin), aminopyrin (Pyramidon), propyphenazone (Isopropylphenazone), noraminopyrine methane-sulfonate sodium (Novalgin), phenylbutazone (Butazolidin), nifenazone (Nicopyron), oxyphenbutazone (Tanderil), morazone (Tarugan), isopyrin, and piperylone (→ Analgetics and Antipyretics, Chap. 3.). The parent substance of almost all these compounds is 1-phenyl-3-methyl-5-pyrazolone methylated in the 2-position (Antipyrin **6**).

8.2. Dyes and Pigments

As early as one year after the discovery of pyrazolone, Tartrazine, the first dye based on pyrazolone, was produced by Ziegler and Lochner [4]. Tartrazine is still used as a dye for foodstuffs. The industrially important pyrazolone dyes are almost exclusively azo dyes. The pyrazolone molecule couples with diazonium compounds in the 4-position to give azo dyes and pigments that exhibit good color strength, lightfastness, and other favorable properties.

Besides their use in leather and wool dyeing (→ Azo Dyes) pyrazolones have also found use as precursors for cotton dyes and for pigments and dyes for synthetic fibers and plastics. Pyrazolones are thus found among nearly all types of dye. For wool dyes, solubilizing sulfo and carboxyl groups are used which can be substituents both in the pyrazolone and in the coupling component. The capacity of pyrazolone derivatives to form complexes with chromium is exploited in the production of chrome dyes.

By introducing so-called reactive components into pyrazolones or into the coupling component, reactive dyes for cellulose fibers are obtained (→ Reactive Dyes). Pyrazolone derivatives without solubilizing groups are suitable as precursors for pigments. For coloring synthetic fibers, either pyrazolone pigments are used or pyrazolone dyes are dissolved in the spinning solution.

Pyrazolones are gaining increasing importance as photographic dyes, particularly in sensitizer dyes and as precursors for color couplers and color filters (→ Photography). In addition pyrazolones and related compounds have proved to be good developing agents for black and white and Polaroid photography.

8.3. Crop Protection

Pyrazolones with fluorine- or trifluoromethyl-substituted aryl groups in the 1-position have become important as herbicides and fungicides [23].

9. Economic Aspects

Pyrazolones are used almost exclusively as precursors for pharmaceuticals and dyes. No data are available on production capacity, but it can be estimated at ca. 20 000 t worldwide. About 16 000 – 17 000 t of pyrazolones are produced annually, of which ca. 9000 t phenylmethylpyrazolone (**4**) are used for pharmaceuticals. The rest is divided between about 10 – 15 other pyrazolones, which are required for dye and pigment production including ca. 3000 t of carboxypyrazole acid **4** (see page 8) for the production of Tartrazine. The most important producing countries are Germany, Switzerland, and the United States.

10. Derivatives

1-Phenyl-3-methyl-5-pyrazolone [*89-25-8*] (**4**), $C_{10}H_{10}N_2O$, M_r 174.20, *mp* 128–129 °C, bp_{17mbar} 191–192 °C, forms colorless crystals that are almost insoluble in cold water, ether, and petroleum ether, but soluble in alcohol, acids, and alkalis. With heavy metal ions, salts are formed, which are mostly deeply colored [24]. Phenylmethylpyrazolone reduces ammoniacal silver nitrate, reacts with nitrous acid to give a 4-oximino compound [25], and gives characteristic color reactions with benzotrichloride [26]. On methylation, 2,3-dimethyl- [*60-80-0*], 3,4-dimethyl- [*17900-68-4*], 2,3,4-trimethyl- [*5677-84-9*], and 3,4,4-trimethyl-1-phenyl-5-pyrazolone [*947-82-0*] and 5-methoxy-3-methyl-1-phenyl-1*H*-pyrazole [*27349-35-5*] are formed, depending on the reaction conditions [27]. The industrial production of this pyrazolone, which is produced in large quantities, starts with ethyl acetoacetate and phenylhydrazine, which are converted into the desired pyrazolone in one reaction step. Another method involves the reaction of acetoacetamide solution with aniline without isolation of the intermediates [28].

Phenylmethylpyrazolone is industrially the most important pyrazolone. It is used in large quantities for the production of Antipyrin (**6**) and related pharmaceuticals. In the colorant industry it is used as a coupling component for azo pigments and dyes. With diazotized dichlorobenzidine, for example, it is coupled to give Pigment Orange 13, which is used in the rubber and plastics industry. Phenylmethylpyrazolone is also a coupling component in Solvent Yellow 16, an oil-soluble dye, and in Mordant Yellow 30, a chrome dye.

***p*-Tolylmethylpyrazolone** [*86-92-0*], 1-(*p*-tolyl)-3-methyl-5-pyrazolone, $C_{11}H_{12}N_2O$, M_r 188.20, *mp* 137–138 °C, forms colorless crystals from water or cyclohexane. It is readily soluble in alcohol, glacial acetic acid, and benzene, but sparingly soluble in cold water [29].

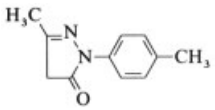

It is produced from *p*-tolylhydrazine and acetoacetamide in acidic solution or directly from *p*-toluidine without isolating the intermediates. Tolylmethylpyrazolone is used for the production of pigments. Coupling with diazotized *o*-dianisidine gives Pigment Red 37, which is used in the plastics and rubber industry. It is also the coupling component in Pigment Orange 34, which is used in textile printing and paints.

Pyrazole acid 4 [*89-36-1*], 1-(4-sulfophenyl)-3-methyl-5-pyrazolone, $C_{10}H_{10}N_2O_4S$, M_r 254.26, *mp* 320 °C (dec.), forms colorless crystals which are sparingly soluble in alcohol, glacial acetic acid, and cold water, but readily soluble in aqueous alkali. It gives an intense red coloration with iron(III) chloride solution [30].

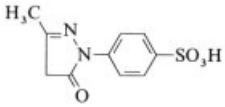

Pyrazole acid 4 is produced either from phenylhydrazine-4-sulfonic acid or from sulfanilic acid with acetoacetamide solution, without isolation of the intermediates.

Pyrazole acid 4 is used as a coupling component for azo dyes in many areas of the textile industry. For example, Acid Yellow 11, Acid Yellow 25, and Acid Yellow 117 are used for wool dyeing; and Direct Orange 1, Direct Yellow 16, and the reactive dyes Reactive Yellow 15, Reactive Yellow 17, and Reactive Yellow 24, for cotton dyeing. Mordant Yellow 8 and Mordant Orange 35 are chrome dyes.

Oxyethylsulfonepyrazolone [*21951-34-8*], 1-(4-β-hydroxyethylsulfonylphenyl)-3-methyl-5-pyrazolone, $C_{12}H_{14}N_2O_4S$, M_r 282.25, *mp* 172 °C (dec.), forms reddish crystals which are readily soluble in hot water and in aqueous alkalis.

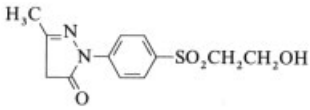

The production is carried out by the usual process starting from the corresponding hydrazine

and acetoacetic ester or acetoacetamide. This pyrazolone is used as the reactive dye component in Reactive Yellow 13, Reactive Yellow 31, Reactive Yellow 37, Reactive Yellow 96, and Reactive Orange 7.

Carboxypyrazole acid 4 [*118-47-8*], 1-(4-sulfophenyl)-5-pyrazolone-3-carboxylic acid, $C_{10}H_8N_2O_6S$, M_r 284.23, is readily soluble in hot water and aqueous alkalis, crystallizes with 2 mol water, and gives a violet coloration with iron(III) chloride solution [31].

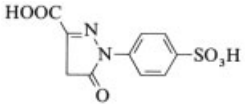

It is produced by treating the sodium salt of ethyl oxaloacetate with phenylhydrazine-4-sulfonic acid and warming the reaction product with aqueous sodium hydroxide, which results in ring-closure and simultaneous hydrolysis of the ester group to give the free carboxylic acid. This pyrazolone can also be produced by treatment of diazotized sulfanilic acid with diethyl acetylsuccinate. The carboxylic acid ester obtained is subsequently hydrolyzed with sodium hydroxide [32].

Carboxypyrazole acid 4 is used extensively in the production of Tartrazine (Food Yellow 4), the oldest pyrazolone dye, which is used to dye foodstuffs yellow. This pyrazolone is also used in the production of reactive dyes and is, for example, the coupling component in Reactive Yellow 23, Reactive Yellow 42, Reactive Orange 74, and Reactive Brown 16. It is present as a copper pigment in Reactive Black 31.

Carbethoxypyrazolone [*89-33-8*], ethyl-1-phenyl-5-pyrazolone-3-carboxylate, $C_{12}H_{12}N_2O_3$, M_r 232.22, *mp* 185 – 186 °C, forms colorless crystals which are sparingly soluble in water and ether, but soluble in alcohol [33].

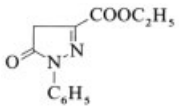

It is produced by reacting the sodium salt of ethyl oxaloacetate with phenylhydrazine or by reaction of diethyl acetylsuccinate with diazotized aniline [34]. It is principally used as a component of Pigment Red 38 in the rubber and plastics sector, in particular for polystyrene and polyethylene.

Iminopyrazolone [*6401-97-4*], 1-phenyl-3-methyl-5-iminopyrazolone (**11**), $C_{10}H_{11}N_3$, M_r 173.21, *mp* 117 – 119 °C, is readily soluble in hot alcohol and chloroform. The chlorohydrate [*20737-88-6*] is a crystalline powder, *mp* 199 – 200 °C, which is readily soluble in water and alcohol [35]. Production involves treatment of butadienenitrile with phenylhydrazine in methanol and warming the reaction mixture with aqueous hydrochloric acid [19], or refluxing 5-methylisoxazole-3-carboxylic acid [*3405-77-4*] with phenylhydrazine [36]. Iminopyrazolone is used in cotton dyeing as a coupling component in Direct Red 180 in the form of the copper complex.

Phenazone [*60-80-0*], 1-phenyl-2,3-dimethyl-5-pyrazolone, Antipyrin (**6**), $C_{11}H_{12}N_2O$, M_r 188.22, *mp* 113 – 114 °C, forms colorless crystals which are readily soluble in water, alcohol, and chloroform, and sparingly soluble in cold toluene and diethyl ether [37]. Production involves methylation of phenylmethylpyrazolone (**4**) with dimethyl sulfate in the melt [38]. Besides its major industrial importance as a starting material for Pyramidon and Novalgin, Antipyrin is used almost exclusively in the area of pharmacology as the analgesic component in combination preparations or in the form of its salts (→ Analgesics and Antipyretics, Chap. 3.).

Aminopyrin [*58-15-1*], 1-phenyl-2,3-dimethyl-4-dimethylamino-5-pyrazolone, Pyramidon, $C_{13}H_{17}N_3O$, M_r 231.29, *mp* 107 – 109 °C, colorless crystals, readily soluble in alcohol and chloroform, moderately soluble in benzene, ether, and cold water [39].

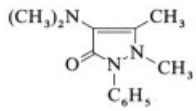

In the production process Antipyrin is nitrosated in the 4-position, reduced to the 4-amino compound, and subsequently converted into the

4-dimethylamino compound by refluxing with formic acid and formaldehyde [40].

Pyramidon has a broad spectrum of activity and surpasses Antipyrin in its analgesic and antipyretic potency. In animals and humans the compound undergoes extensive metabolization. Unchanged Pyramidon has not been detected in excreta. Up to 3 % is excreted in the urine as the red dye, Rubazonic acid [*909-59-1*]. Because its toxicity is 4 – 5 times greater than that of Antipyrin, Pyramidon has now mostly disappeared from the pharmaceuticals market. However, it is still used as a double compound with barbiturates, which results in detoxification.

Propyphenazone [*479-92-5*], 1-phenyl-2,3-dimethyl-4-isopropyl-5-pyrazolone, $C_{14}H_{18}N_2O$, M_r 230.30, *mp* 103 °C, forms colorless crystals which are readily soluble in alcohol and sparingly soluble in cold water [41].

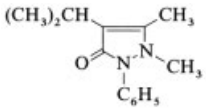

It is produced by condensation of phenylmethylpyrazolone (**4**) with acetone in the 4-position to give the isopropylidene compound, which is then converted into the 4-isopropyl compound by catalytic hydrogenation over a nickel catalyst. The subsequent methylation at N-2 is carried out with dimethyl sulfate at 125°C [42]. As an effective analgesic and antipyretic propyphenazone is a component of many mixed preparations, of which Melabon and Saridon are the best known.

Dipyrone [*68-89-3*], sodium 1-phenyl-2,3-dimethyl-5-pyrazolone-4-methylamino-methylsulfonate, Novalgin, Metamizol, $C_{13}H_{16}N_3O_4SNa \cdot H_2O$, M_r 351.35, *mp* 223 – 224 °C (dec.), is a colorless substance that forms a monohydrate. It is readily soluble in water and methanol, and practically insoluble in ether and chloroform.

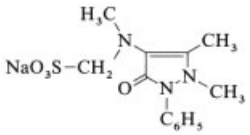

The aqueous solution is neutral and becomes yellow with time, although there is no change in activity. On addition of nitrous acid the solution becomes blue and then green [43]. Dipyrone is produced from 4-aminoantipyrin, an intermediate in the synthesis of Pyramidon. This is reacted with benzaldehyde and then with dimethyl sulfate to give 4-methylaminoantipyrin. The latter is warmed with sodium hydrogensulfite and formaldehyde to give dipyrone [44]. The free acid [*50567-35-6*] can be produced by passing sulfur dioxide through an alcoholic solution of 4-aminoantipyrin and formaldehyde [45]. With inorganic and organic bases this acid forms salts, of which the calcium and quinine salts have proved to be particularly useful as components of Novalgin – quinine in the treatment of acute infections [46].

11. Toxicology

Pyrazolones are of low toxicity, so they do not present a direct health hazard. For all the pyrazolones investigated so far the acute oral toxicity (LD_{50}, rat) is >2000 mg/kg; for many pyrazolones the values are considerably higher, e.g., carboxypyrazole acid 4, 15 000 mg/kg; *p*-tolylmethylpyrazolone, 7450 mg/kg; and 2,5-dichloropyrazole acid 4, 17 227 mg/kg [47]. However, skin and eye contact should be avoided. When handling dry pyrazolones a dust mask should be worn to prevent inhalation. Contaminated clothing should be changed immediately. MAK values have not been set for pyrazolones.

12. References

General References

1. R. H. Wiley, P. Wiley: *The Chemistry of Heterocyclic Compounds; Pyrazolones, Pyrazolidones, and Derivatives,* J. Wiley & Sons, New York 1964. G. Ehrhart, H. Ruschig: *Arzneimittel,* **vol. 1,** Verlag Chemie, Weinheim 1972, pp. 148 – 162. *Kirk-Othmer,* **vol. 19,** 445 – 453. K. Brune, R. Lanz: *100 Jahre Pyrazolone,* Urban & Schwarzenberg, München 1985. A. R. Katritzky: *Comprehensive Heterocyclic Chemistry,* **vol. 5,** Pergamon Press, Oxford 1984. S. Coffey, M. F. Ansell: *Rodd's Chemistry of Carbon Compounds,* **vol. IV C,** Elsevier,

Amsterdam 1986. W. Krohs, O. Hensel: *Pyrazolone und Dioxopyrazolidine, Arzneimittel-Forschung,* 11. Beiheft, Editio Cantor, Aulendorf/Württ. 1961. N. N. Woroshzow: *Grundlagen der Synthese von Zwischenprodukten und Farbstoffen,* Akademie-Verlag, Berlin 1966. H. R. Schweizer: *Künstliche Organische Farbstoffe und ihre Zwischenprodukte,* Springer Verlag, Berlin-Göttingen-Heidelberg 1964. H. E. Fierz-David, L. Blangey: *Farbenchemie,* 4th ed., Verlag J. Springer, Wien 1938. W. Herbst, K. Hunger: *Industrielle Organische Pigmente,* VCH Verlagsgesellschaft, Weinheim 1987.

Specific References

2. L. Knorr, *Ber. Dtsch. Chem. Ges.* **16** (1883) 2597.
3. Hoechst, DE 26 429, 1883 (L. Knorr). *Friedländer* **1,** 208.
4. J. H. Ziegler, M. Locher, *Ber. Dtsch. Chem. Ges.* **20** (1887) 834.
5. A. R. Katritzky, F. W. Maine, *Tetrahedron* **20** (1964) 299, 315. A. R. Katritzky, F. W. Maine, S. Golding, *Tetrahedron* **21** (1965) 1693.
6. J. Elguero, R. Jaquier, G. Tarrago, *Bull. Soc. Chim. Fr.* 1967, 3772.
7. H. Dorn, *J. Prakt. Chem.* **315** (1973) 382.
8. L. Knorr, *Justus Liebigs Ann. Chem.* **238** (1887) 137, 165.
9. H. v. Pechmann, *Ber. Dtsch. Chem. Ges.* **28** (1895) 1624.
10. DE 95 643, 1897 (F. Stolz). *Friedländer* **5,** 779.
11. F. Stolz, *J. Prakt. Chem. [2]* **55** (1897) 154.
12. B. S. Jensen, *Acta Chem. Scand.* **13** (1959) 1668.
13. T. Curtius, *J. Prakt. Chem. [2]* **50** (1894) 511.
14. G. Westöö, *Acta Chem. Scand.* **6** (1952) 1499.
15. A. Michaelis, H. Römer, *Ber. Dtsch. Chem. Ges.* **31** (1898) 2908.
16. Hoechst, DE 3 416 205, 1984 (R. Berthold).
17. I.G. Farbenind., DE 740 249, 1938 (H. Vollmann, W. Ostrowski, F. Schloffer). *Friedländer* **VI/2,** 2424.
18. A. Carpino, *J. Am. Chem. Soc.* **80** (1958) 5796, 5798.
19. P. Kurz, H. Gold, H. Disselnkötter, *Justus Liebigs Ann. Chem.* **624** (1959) no. 1, 20.
20. A. Michaelis, R. Pasternak, *Ber. Dtsch. Chem. Ges.* **32** (1899) 2398; **34** (1901) 723.
21. H. Ruhkopf, *Ber. Dtsch. Chem. Ges.* **73** (1940) 820.
22. NIOSH, *Registry of Toxic Effects of Chemical Substances,* Washington, D.C., 1982.
23. Bayer, DE 3 728 278, 1986 (R. Gehring et al.); DE 3 617 554, 1986 (R. Gehring et al.). Roussel-UCLAF, FR 2 604 174, 1986 (J. Tessier, P. Girault).
24. *Beilstein* **24,** H 20, E I 190, E II 9, E III/IV 71, E V 317.
25. L. Knorr, *Ber. Dtsch. Chem. Ges.* **17** (1884) 2034.
26. L. Knorr, *Ber. Dtsch. Chem. Ges.* **17** (1884) 548.
27. H. Henecka: *Chemie der β-Dicarbonyl-Verbindungen,* Springer Verlag, Berlin – Göttingen – Heidelberg 1950, p. 336.
28. I.G. Farbenind., DE 747 734, 1939. *Friedländer* **VI/2,** 2425.
29. *Beilstein,* **24,** H 39, E I 205, E III/IV 85, E V 320.
30. *Beilstein,* **24,** H 44, E I 210, E II 20, E III/IV 93, E V 324.
31. *Beilstein,* **25,** H 207, E I 568, E III/IV 1553.
32. Ilford Ltd., US 2 459 226, 1949 (J. D. Kendall, D. J. Fry).
33. *Beilstein,* **25,** H 207, E III/IV 1549.
34. E. Koike et al., *J. Chem. Soc. Japan Ind. Chem. Sect.* **57** (1954) 123.
35. *Beilstein,* **24,** H 25, E I 191; 25, E IV 2034.
36. S. Cusmano, *Gazz. Chim. Ital.* **70** (1940) 86, 88.
37. *Beilstein,* **24,** H 27, E I 194, E II 11, E III/IV 75, E V 330.
38. FIAT, *Final Report* **1023,** 44.
39. *Beilstein,* **25,** H 452, E I 672, E II 364, E III/IV 3555.
40. FIAT, *Final Report* **1023,** 47.
41. *Beilstein,* **24,** E III/IV 141, E V 408.
42. Hoffmann-La Roche, DE 558 473, 1931. *Friedländer* **19,** 1186.
43. *Beilstein,* **25,** E II 369, E III/IV 3569.
44. I.G. Farbenind., DE 476 663, 1922 (M. Bockmühl, A. Schwarz). *Friedländer* **16,** 2449.
45. I.G. Farbenind., DE 617 237, 1934 (M. Bockmühl, L. Stein). *Friedländer* **22,** 551.
46. I.G. Farbenind., DE 644 647, 1934 (M. Bockmühl, L. Stein). *Friedländer* **22,** 553.
47. Hoechst AG, Product Information, Sales Department Fine Chemicals, Frankfurt/Main, 1989.

Pyrene → Hydrocarbons
Pyrethroids → Insect Control

Pyridine and Pyridine Derivatives

Shinkichi Shimizu, Koei Chemical Co. Ltd., Osaka 541, Japan
Nanao Watanabe, Koei Chemical Co. Ltd., Osaka 541, Japan
Toshiaki Kataoka, Koei Chemical Co. Ltd., Osaka 541, Japan
Takayuki Shoji, Koei Chemical Co. Ltd., Osaka 541, Japan
Nobuyuki Abe, Koei Chemical Co. Ltd., Osaka 541, Japan
Sinji Morishita, Koei Chemical Co. Ltd., Osaka 541, Japan
Hisao Ichimura, Koei Chemical Co. Ltd., Osaka 541, Japan

1. Introduction 2
2. Pyridine and Alkylpyridines 3
2.1. Properties 3
2.2. Production 4
2.2.1. Separation from Tar 4
2.2.2. Synthesis from Aldehydes or Ketones with Ammonia 4
2.2.3. Synthesis from Acrylonitrile and Ketones 7
2.2.4. Synthesis from Dinitriles 7
2.2.5. Dealkylation of Alkylpyridines 8
2.2.6. Synthesis of 5-Ethyl-2-Methylpyridine from Paraldehyde and Ammonia . 8
2.2.7. Synthesis from Nitriles and Acetylene 8
2.2.8. Other Synthetic Methods 8
2.3. Quality Specifications, Storage, and Transportation 8
2.4. Uses 10
2.5. Economic Aspects 10
3. Pyridine Derivatives 10
3.1. Vinylpyridines 10
3.2. Bipyridines 12
3.3. Quaternary Pyridinium Salts 12
3.4. Pyridine *N*-Oxides 13
3.5. Piperidines 15
3.6. Halopyridines 15
3.7. Pyridinecarbonitriles, Carboxylic Acids, and Carboxamides 18
3.8. Aminopyridines 20
3.8.1. 2-Aminopyridine 20
3.8.2. Other Aminopyridines 21
3.9. Pyridinols 22
3.10. Pyridyl Alcohols 23
3.11. Pyridinecarbaldehydes 24
3.12. Pharmaceuticals and Agrochemicals 25
4. Toxicology 25
4.1. Acute Toxicity 25
4.2. Subacute and Chronic Toxicity . . . 31
4.3. Mutagenicity and Ecotoxicity 31
5. References 31

1. Introduction

Pyridine, C_5H_5N, is a six-membered heterocyclic compound containing one nitrogen atom. Pyridine and its homologues are commonly called pyridine bases. The first pyridine derivative, 2-methylpyridine (α-picoline; *pix*, Latin = pitch) was isolated from coal tar in 1846 by Anderson [1]. In 1851, Anderson obtained pyridine (*pyros*, Greek = fire) and dimethylpyridine (lutidine; *lutum*, Latin = dirt) from bone oil [2].

Ramsay synthesized pyridine in 1876 by passing a mixture of acetylene and hydrogen cyanide through a red-hot tube [3]. Typical syntheses of pyridine derivatives are based on the work of Hantzsch (1882) [4] and Chichibabin (1906) [5]. The method of the latter is especially suitable for mass production and it is still an important industrial process.

Compounds containing a pyridine ring, such as vitamin B_6 (pyridoxine) [6], nicotinamide [7], nicotinic acid, the coenzymes nicotinamide adeninedinucleotide (NAD) and reduced NAD (NADH), and many alkaloids, play important roles in metabolism. Pyridine bases are widely used in pharmaceuticals including nicotinamide and nicotinic acid. Similarly, pyridine derivatives are important insecticides and herbicides due to their high bioactivity. Further, they are used as adhesives for textiles and as chemicals, solvents, and catalysts.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_399

Table 1. Physical properties of pyridine and alkylpyridines

Compound	CAS registry no.	M_r	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}	pK_a (H_2O, 25 °C)	Solubility in H_2O at 20 °C, g/100 g	Azeotrope with H_2O: wt % H_2O	Azeotrope with H_2O: *bp*, °C	Ignition temperature, °C	Explosion composition, %
Pyridine	[*110-86-1*]	79.10	−41.7	115.3	0.9819	1.5102	5.22	miscible	41.3	93.6	550	1.7–10.6
2-Methylpyridine	[*109-06-8*]	93.13	−66.7	129.4	0.9455	1.5010	5.96	miscible	48	93.5	535	1.4–8.6
3-Methylpyridine	[*108-99-6*]	93.13	−18.2	144.1	0.9564	1.5043	5.63	miscible	63	97		1.3–8.7
4-Methylpyridine	[*108-89-4*]	93.13	3.6	145.4	0.9546	1.5058	5.98	miscible	63.5	97.35	500	1.3–8.7
2,6-Dimethylpyridine	[*108-48-5*]	107.16	− 6.1	144.5	0.9237	1.4977	6.72	27.2 (45.3 °C)	51.5	96.02		
3,5-Dimethylpyridine	[*591-22-0*]	107.16	− 6.5	171.9	0.944	1.5049	6.15	3.3				
5-Ethyl-2-methylpyridine	[*104-90-5*]	121.18	−70.3	178.3	0.9208	1.4974		1.2	72	98.4		

2. Pyridine and Alkylpyridines

Pyridine and alkylpyridines are produced commercially by synthesis as well as by isolation from natural sources such as coal tar. Commercially important compounds are pyridine, 2-methylpyridine, 3-methylpyridine, 4-methylpyridine, 2,6-dimethylpyridine, 3,5-dimethylpyridine, and 5-ethyl-2-methylpyridine.

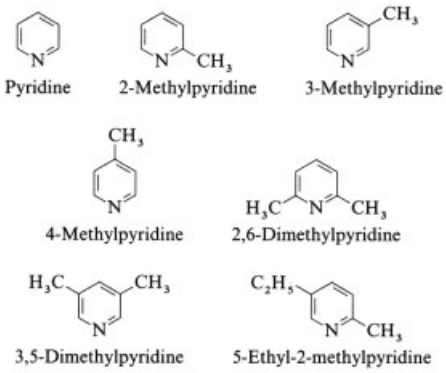

2.1. Properties

Pyridine is miscible in all proportions with water and most common organic solvents and has a boiling point 35 °C higher than benzene.

Physical data of pyridine and alkylpyridines are summarized in Table 1.

Since the pyridine ring has three double bonds, six π-electrons exist, which are sufficient for aromatic ring formation without involving the lone pair electrons of the nitrogen atom. Since the lone pair electrons remains free, quaternary salts retain the aromaticity. However, the nitrogen atom has a higher electronegativity than the carbon atoms and shows an electron-withdrawing effect. This is represented by the resonance hybrids:

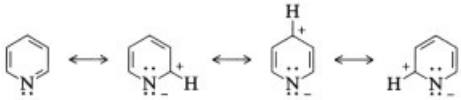

Therefore the electron densities at the 2- and 4-positions are low, and the ring is regarded as a

π-electron-deficient aromatic ring. The weak basicity of pyridine and alkylpyridines is due to the lone pair of electrons on the ring nitrogen atom. The basicity of pyridine derivatives is increased by electron-donating substituents and decreased by electron-withdrawing substituents.

In oxidation and reduction reactions, the pyridine ring exhibits properties characteristic of π-electron-deficient aromatic rings: resistance to oxidation and facile reduction. In the oxidation of alkylpyridines with alkaline $KMnO_4$, the pyridine ring is not oxidized; instead the corresponding carboxylic acids are formed [8]. Furthermore in the oxidation of quinoline with alkaline $KMnO_4$, the main product is pyridine-2,3-dicarboxylic acid. This shows that the pyridine ring is more stable to oxidation than the benzene ring [9].

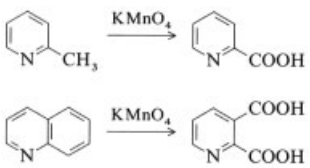

On reduction with hydrogen in the presence of catalyst, quinoline and 2-phenylpyridine are reduced preferentially at the pyridine ring [10].

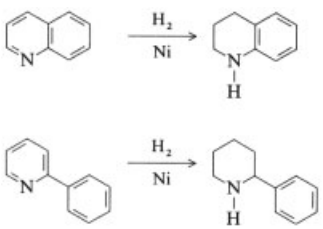

Because of the low π-electron density at the ring carbon atoms, electrophilic reactions rarely occur on the pyridine ring, and they occur at the 3-position only under drastic conditions. Although nitration of pyridine bases is difficult [11], sulfonation occurs more readily [12].

Friedel – Crafts reactions and reactions catalyzed by $PdCl_2$ such as oxychlorination of benzene, which are useful for benzene derivatives, do not take place because of the coordination of the nitrogen atom to $AlCl_3$ or $PdCl_2$. Electron-donating substituents such as hydroxyl, amino, or alkyl groups attached to the pyridine ring facilitate electrophilic reactions because of the increased electron density in the ring.

In contrast, nucleophilic substitution reactions occur readily at the 2-, 4-, and 6-positions. Treatment of pyridine with sodium amide gives 2-aminopyridine

(Chichibabin reaction) [13]. Reaction of pyridine bases with Grignard reagent, alkyllithium, or aryllithium gives 2-substituted pyridines [14]. Treatment of pyridine with methanol and hydrogen in the presence of a nickel catalyst gives 2-methylpyridine [15]. Substitution reactions of 2- or 4-halopyridines with nucleophiles, such as alkoxides, thiolates, amines, or carbanions, occur easily and provide important synthetic methods.

Radical reactions of pyridine – for example, the phenylation of pyridine with benzene diazonium chloride – give substituted products in the decreasing order 2- $>$ 3- $>$ 4- [16], and chlorination with chlorine above 270 °C or under UV irradiation gives a mixture of 2-, 4-, and 6- chloropyridines.

In the reaction of substituted groups in the pyridine ring, dealkylation and decarboxylation occur mainly at the 2-position [17], whereas hydroxymethylation of methyl groups with formaldehyde occurs at the 2- and 4-positions [18].

Like aliphatic tertiary amines, pyridine bases give *N*-oxides with hydrogen peroxide and peroxy acids, and form quaternary ammonium salts with alkyl halides.

2.2. Production

2.2.1. Separation from Tar

Pyridine bases are a constituent of tars. They were isolated from coal tar or coal gas before synthetic manufacturing processes became established. The amounts contained in coal tar and coal gas are small, and the pyridine bases isolated from them are a mixture of many components. Thus, with a few exceptions, isolation of pure pyridine bases was expensive. Today, almost all pyridine bases are produced by synthesis.

2.2.2. Synthesis from Aldehydes or Ketones with Ammonia

The reaction of aldehydes or ketones with ammonia is the most general synthetic reaction for the manufacture of pyridine bases and allows the preparation of various pyridines. This reaction was first studied in detail by CHICHIBABIN in 1924 [19] and since then been studied extensively for industrial manufacturing because of cheap access to raw materials. The reaction is usually carried out at 350 – 550 °C and a space velocity of 500 – 1000 h^{-1} in the presence of a solid acid catalyst (e.g., silica – alumina).

Aldehydes react with ammonia as follows:

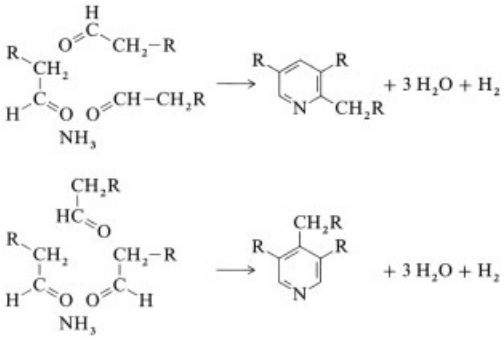

For example, acetaldehyde and ammonia give 2-methylpyridine (Eq. 1) and 4-methylpyridine (Eq. 2). Some examples of the synthesis are given in Table 2.

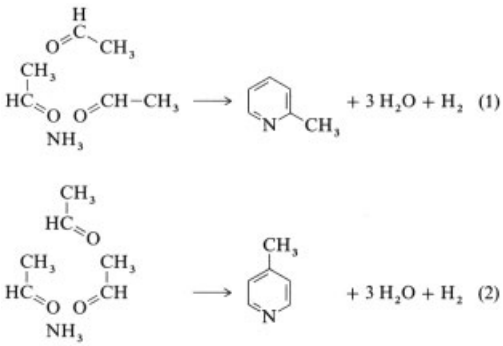

With α, β-unsaturated aldehydes the reaction occurs according to the following schemes:

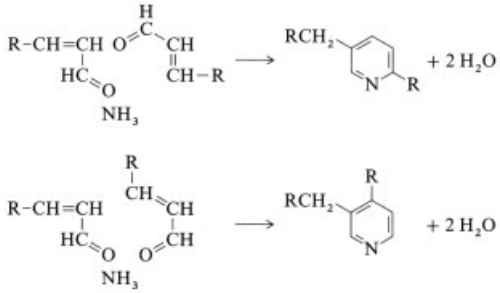

For example, acrolein and ammonia give 3-methylpyridine (Eq. 3); pyridine is simultaneously formed by demethylation. Examples of this synthesis are given in Table 3.

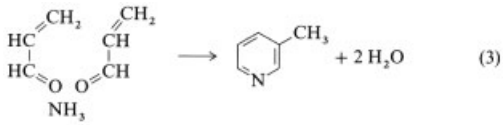

Acrolein and acetaldehyde react with ammonia mainly to form pyridine:

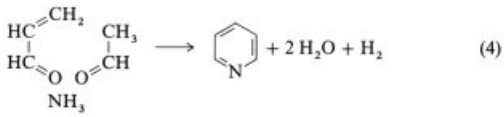

Acrolein and propionaldehyde react with ammonia to give primarily 3-methylpyridine (Eq. 5).

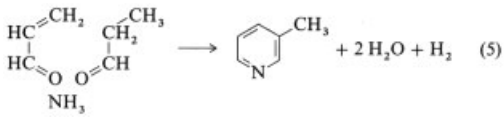

Acetaldehyde and formaldehyde react with ammonia to give mainly pyridine (Eq. 6): they appear to first form acrolein (Eq. 7), and then acrolein and formaldehyde react with ammonia to give pyridine. Simultaneously, 2-, 3-, and 4-methylpyridines are formed, as shown in Equations (1) – (4). This method is one of the most widely used for pyridine production. Table 4 lists some examples of this process, and Figure 1 illustrates the flow sheet of the plant.

A preheated gaseous mixture of acetaldehyde, 36 % formaldehyde, and ammonia is

Table 2. Synthesis of 2- and 4-methylpyridine from acetaldehyde and ammonia

Company	Catalyst	Yield, %		Ref.
		2-Methylpyridine	4-Methylpyridine	
Koei Chemical	$Co_3Al_3(PO_4)_5^-$	45	9	[20]
Nippon Kayaku	$Al_2O_3-SiO_2-CdCl_2$	35	44	[21]

Table 3. Synthesis of 3-methylpyridine and pyridine from acrolein and ammonia

Company	Catalyst	Yield, %		Ref.
		Pyridine	3-Methylpyridine	
Degussa	$Al_2O_3-MgF_2$	25	49	[22]
ICI	$SiO_2-Al_2O_3-H_2SiF_2$	62	15	[23]
Nippon Kayaku	$SiO_2-Al_2O_3-CdF_2$	26	56	[24]
Koei Chemical	$SiO_2-Al_2O_3-MnF_2$	20	45	[25]
Daicel Chemical	$SiO_2-Al_2O_3$	22	49	[26]

Table 4. Synthesis of pyridine and 3-methylpyridine from acetaldehyde and formaldehyde with ammonia

Company	Catalyst	Yield, %		Ref.
		Pyridine	3-Methylpyridine	
ICI	$SiO_2-Al_2O_3-$coke	38	25	[27]
Rütgerswerk	$SiO_2-Al_2O_3-CdF_2$	57	29	[28]
Nepera	ZSM-5 *	54	28	[29]
Koei Chemical	Tl–ZSM-5	63	9	[30]

* Zeolite.

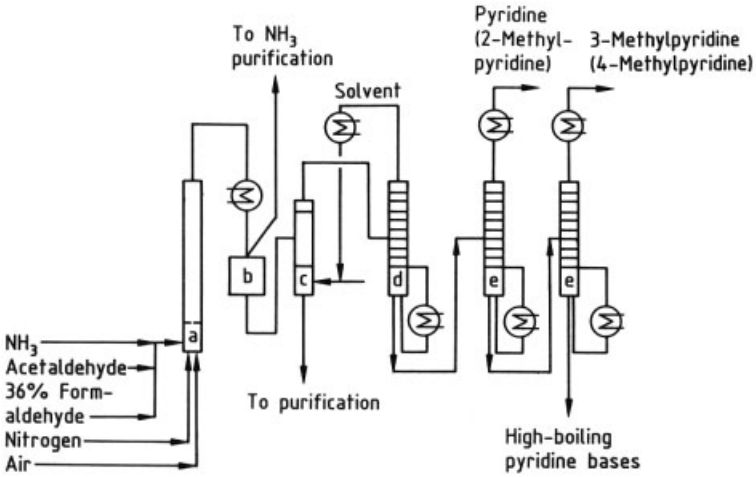

Figure 1. Flow sheet of pyridine and methylpyridine production from acetaldehyde and formaldehyde with ammonia
a) Reactor; b) Collector; c) Extraction; d) Solvent distillation; e) Distillation

passed through the reactor (a) packed with the catalyst at 400 – 450 °C. The reaction mixture is separated from ammonia and hydrogen by a collector (b) and extracted with solvent, e.g., benzene (c). The solvent is removed from the extract (d) and pyridine and 3-methylpyridine are isolated in continuous distillation columns (e). The catalyst is periodically regenerated by air.

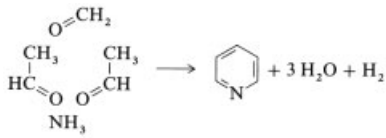

(6)

$$CH_3CHO + HCHO \longrightarrow CH_2{=}CHCHO + H_2O \quad (7)$$

Propionaldehyde and formaldehyde react with ammonia to give 3,5-dimethylpyridine (Eq. 8) [31]; benzaldehyde and acetaldehyde give 2-phenylpyridine [*1008-89-5*] (Eq. 9) and 4-phenylpyridine [*939-23-1*] (Eq. 10) [32].

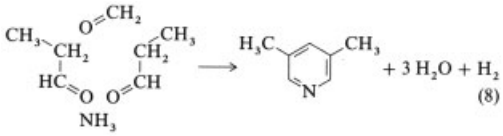

(8)

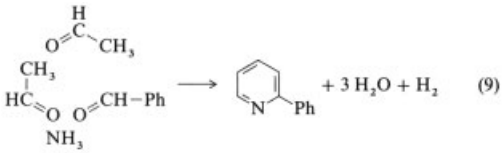

(9)

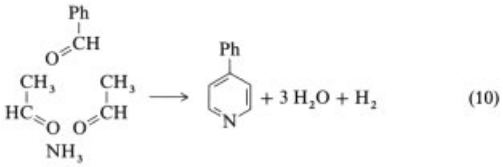

(10)

Ketones and aldehydes react with ammonia according to the following general scheme:

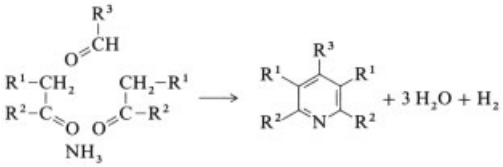

Typically, acetone [*67-64-1*] and formaldehyde with ammonia give 2,6-dimethylpyridine (Eq. 11) [33]:

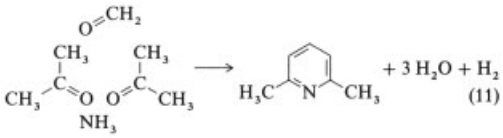

(11)

α,β-Unsaturated ketones or aldehydes react with ammonia according to the following scheme:

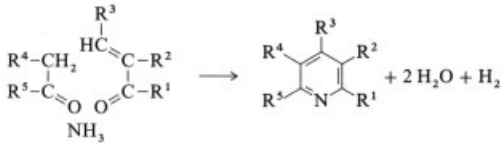

For example, acrolein and acetone react with ammonia to give 2-methylpyridine (Eq. 12) [34]:

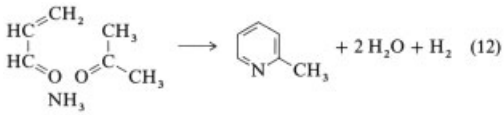

(12)

As a variant, acetone with ammonia gives 2,4,6-trimethylpyridine [*108-75-8*] with simultaneous demethylation (Eq. 13) [35]. Cyclopentanone [*120-92-3*] and acrolein with ammonia give 2,3-cyclopentenopyridine [*533-37-9*] (Eq. 14) [36]. Using aniline instead of ammonia results in formation of quinoline (Eq. 15):

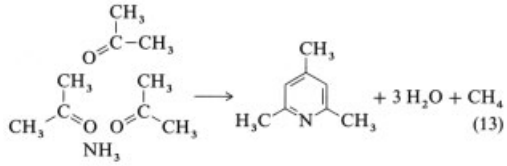

(13)

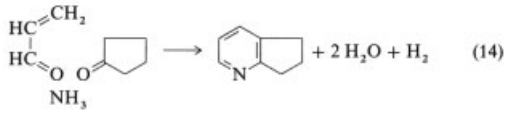

(14)

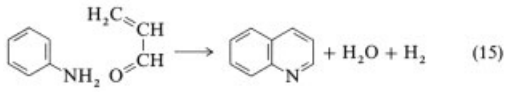

(15)

As shown above, various pyridines can be obtained by using different combinations of aldehydes, ketones, ammonia, and amines.

2.2.3. Synthesis from Acrylonitrile and Ketones

Synthesis from acrylonitrile and acetone gives 2-methylpyridine selectively, in contrast to the process using acetaldehyde and ammonia, which gives 4-methylpyridine as a byproduct. First, the reaction of acrylonitrile and acetone, catalyzed by a primary aliphatic amine such

as isopropylamine and a weak acid such as benzoic acid [*65-85-0*], occurs in the liquid phase at 180 °C and 2.2 MPa to give 5-oxohexanenitrile [*10412-98-3*], with 91 % selectivity (Eq. 16). The acrylonitrile conversion is 86 % [37]. Then cyclization and dehydration of the initial product are carried out in the gas phase in the presence of hydrogen over a palladium, nickel, or cobalt-containing catalyst at ca. 240 °C to give 2-methylpyridine in 84 % yield (Eq. 17) [38]. 4-Methyl-5-oxohexanenitrile [*10413-01-1*], formed from acrylonitrile and 2-butanone, gives 2,3-dimethylpyridine [*583-61-9*] in 89 % yield [38].

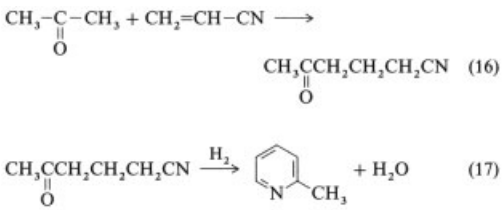

2.2.4. Synthesis from Dinitriles

In a vapor-phase reaction over a nickel-containing catalyst in the presence of hydrogen, 2-methylglutaronitrile [*4553-62-2*] gives 3-methylpiperidine [*626-56-2*], which then undergoes dehydrogenation over palladium – alumina to give 3-methylpyridine [39]:

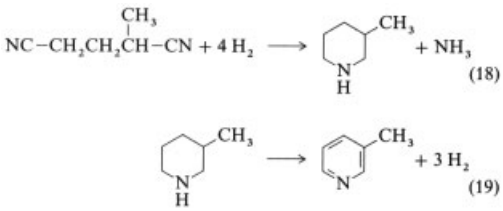

A one-step gas-phase reaction over a palladium-containing catalyst is reported to give 3-methylpyridine in 50 % yield [40].

2.2.5. Dealkylation of Alkylpyridines

Alkylpyridines of low commercial value, obtained as byproducts of pyridine base synthesis, are occasionally converted into useful pyridine bases by dealkylation. The methods for dealkylation involve oxidative dealkylation by air over a vanadium oxide catalyst [23], steam dealkylation over a nickel catalyst [24, 25], and hydrodealkylation over a silver or platinum catalyst [44]. Examples are listed in Table 5.

2.2.6. Synthesis of 5-Ethyl-2-Methylpyridine from Paraldehyde and Ammonia

Reaction of paraldehyde with aqueous ammonia in the liquid phase is carried out at 200 – 300 °C and 12 – 13 MPa in the presence of an ammonium salt (e.g., ammonium phosphate) to give 5-ethyl-2-methylpyridine (MEP) in about 70 % yield (Eq. 20) [45]. Figure 2 shows the reaction route, and Figure 3 illustrates the manufacture of MEP by the Montecatini – Edison process.

Paraldehyde, produced from acetaldehyde and sulfuric acid, is reacted with 30 – 40 % aqueous ammonia and acetic acid at 220 – 230 °C and 10 – 20 MPa. The reaction mixture is separated into two phases in a separator (c). Ammonia is recovered from the aqueous layer by a stripper (d). MEP, 2-methylpyridine, and 4-methylpyridine are isolated from the organic layer by distillation.

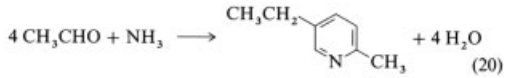

2.2.7. Synthesis from Nitriles and Acetylene

Liquid-phase reaction of nitriles with acetylene is carried out at 120 – 180 °C and 0.8 – 2.5 MPa in the presence of an organocobalt catalyst and gives 2-substituted pyridines [46]:

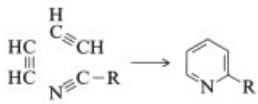

For example, acetonitrile [*75-05-8*] and acetylene react in the presence of cobaltocene as catalyst to give 2-methylpyridine in 76 % yield [47]. Acrylonitrile and acetylene react in the

Table 5. Dealkylation of alkylpyridines

Starting material	Catalyst	Additives	Yield of pyridine, %	Ref.
3-Methylpyridine	$V/Cr/Ag-Al_2O_3$	air, H_2O	82	[41]
2-Methylpyridine	$Ni-SiO_2$	H_2, H_2O	93	[42]
2-Methylpyridine	$Ni-ZrO_2$	H_2O	50	[43]
Alkylpyridine	Ag	H_2	58	[44]

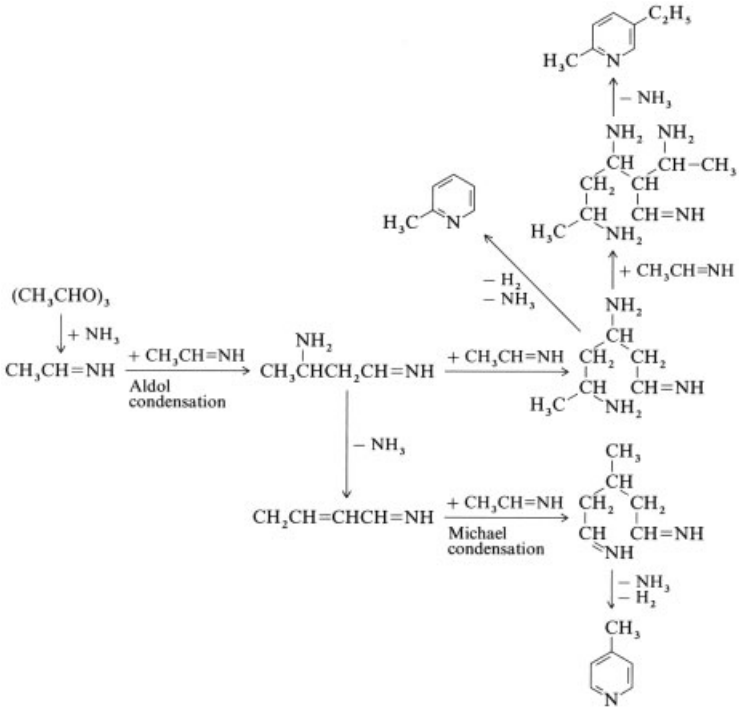

Figure 2. Mechanism of 5-ethyl-2-methylpyridine formation [45]

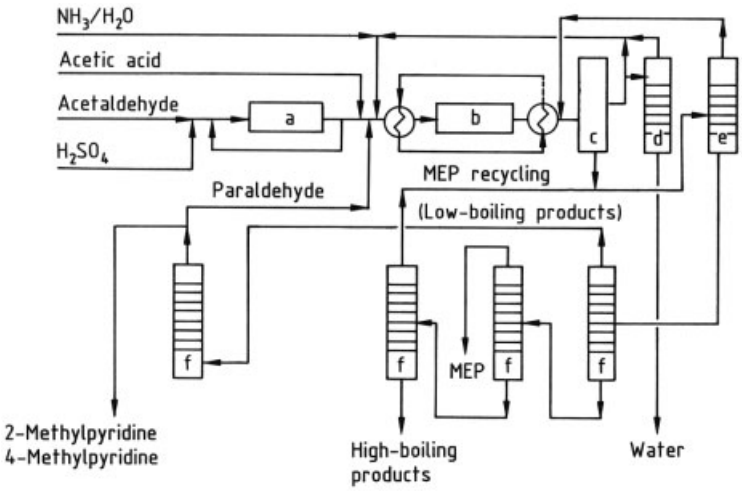

Figure 3. Flow sheet of 5-ethyl-2-methylpyridine (MEP) production by Montecatini – Edison process
a) Paraldehyde production; b) Pyridine reactor; c) Separator; d) Stripper; e) Dewatering column; f) Fractionating columns

presence of cyclopentadienylcobalt – cycloocta-1,5-diene catalyst to give 2-vinylpyridine with 93 % selectivity [48].

2.2.8. Other Synthetic Methods

Ethylene [*74-85-1*] and ammonia react in the presence of a palladium complex catalyst to give 2-methylpyridine and MEP [49]. Pyridine can be prepared from cyclopentadiene by ammoxidation [50], or from 2-pentenenitrile [*13284-42-9*] by cyclization and dehydrogenation [51]. Furfuryl alcohol or furfural reacts with ammonia in the gas phase to give pyridine [52]. 2-Methylpyridine is also prepared from aniline [53].

2.3. Quality Specifications, Storage, and Transportation

Specifications of pyridine vary according to country but are usually > 99.8 % purity by gas chromatographic analysis. Table 6 lists the standard specification for refined pyridine (ASTM) in the United States [54].

Table 6. Standard specifications of refined pyridine (ASTM) [54]

Appearance	clear liquid, free of extraneous matter and sediment
Odor	pyridine, characteristic
$d^{15.56}$	0.985 – 0.990
Color	not darker than no. 20 on platinum – cobalt scale
Distillation range at atmospheric pressure	
Total distillation range	$\leq 2\,^{\circ}C$
Initial distillation temperature (first drop)	$\geq 114.0\,^{\circ}C$
End point (dry point)	$\leq 117.0\,^{\circ}C$
Water	≤ 0.20 wt %
Water solubility	clear solution, no turbidity or oil film

Pyridine bases should generally be stored under dark, cool conditions. They are transported in drums, tank cars, and bulk containers in accordance with the following regulations:

Road:	GGVS/ADR	Class 3, No. 15 b
Rail:	GGVE/RID	Class 3, No. 15
Sea:	GGVSee/IMDG	Class 3.2
Air:		Class 3
	IATA-DGR	
UN no.:	1282	

2.4. Uses

Pyridine is an excellent solvent, especially for dehydrochlorination reactions and extraction of antibiotics. Large amounts of pyridine are used as starting material for pharmaceuticals and agrochemicals: for example, herbicides such as diquat and paraquat, insecticides such as chlorpyrifos, and fungicides such as pyrithione (see Section 3.12).

2-Methylpyridine. The major use of 2-methylpyridine is as a precursor of 2-vinylpyridine. The terpolymer of 2-vinylpyridine with butadiene and styrene is used as an adhesive for textile tire cord. 2-Methylpyridine is also used as a material for a variety of pharmaceuticals and agrochemicals: for example, chemicals such as nitrapyrin to prevent loss of ammonia from fertilizers, herbicides such as picloram, and coccidiostats such as amprolium (see Section 3.12).

3-Methylpyridine. A considerable amount of 3-methylpyridine is used as a starting material for pharmaceuticals and agrochemicals: for example, insecticides such as chlorpyrifos, feed additives such as nicotinic acid and nicotine carboxamide, and herbicides such as fluazifop-butyl (see Section 3.12).

4-Methylpyridine. The primary use of 4-methylpyridine is in the production of the antituberculosis agent isonicotinic hydrazide. Polymers containing 4-vinylpyridine, obtained from 4-methylpyridine, are used as anion exchangers (see Section 3.12).

Polyalkylpyridines. Large amounts of MEP are used as a starting material for nicotinic acid. 2,6-Dimethylpyridine is used for the antiarteriosclerotic pyridyl carbamate, while 3,5-dimethylpyridine is used for producing the antiulcer medication omeprazole (see Section 3.12).

2.5. Economic Aspects

The amounts of pyridine bases produced worldwide are estimated roughly as follows: pyridine, ca. 26 000 t/a (1989); 2-methylpyridine, ca. 8000 t/a (1989); 3-methylpyridine, ca. 9000 t/a

(1989); 4-methylpyridine, ca. 1500 t/a (1989); and MEP, ca. 8000 t/a (1989). Table 7 lists the primary manufacturers of pyridine and alkylpyridines.

Table 7. Manufacturers of pyridine and alkylpyridines

Country	Company	Products
United States	Reilly Industry	pyridine; 2-, 3- and 4-methylpyridine; dimethylpyridine; trimethylpyridine
	Nepera Chemical Co.	pyridine; 2-, 3- and 4-methylpyridine
	Kopper Co.	coal-tar-derived pyridine
Japan	Koei Chemical Co.	pyridine; 2-, 3- and 4-methylpyridine; dimethylpyridine; trimethylpyridine
	Daicel Chemical	pyridine; 3-methylpyridine
	Nippon Steel	coal-tar-derived pyridine
Belgium	Reilly Industry	pyridine; 2-, 3- and 4-methylpyridine
Netherlands	DSM	2-methylpyridine
Switzerland	Lonza	2-methyl-5-ethylpyridine

Table 8. Physical properties of vinylpyridines

	2-Vinylpyridine	4-Vinylpyridine
CAS registry no.	[*100-69-6*]	[*100-43-6*]
Density (20 °C), g/cm^3	0.977	0.988
bp, °C (20 kPa)	110	120
(4 kPa)	70	
(2 kPa)		65
n_D^{20}	1.5509	1.5525
Solubility in water (20 °C), g/L	27.5	29
Viscosity (20 °C), mPa · s	1.17	
pK_a	4.98	5.5

3. Pyridine Derivatives

3.1. Vinylpyridines

Vinylpyridines have the following general formula:

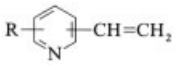

2-Vinylpyridine was first synthesized in 1887, and 4-vinylpyridine in 1920. However, not until 1950 did vinylpyridines begin to be used for various commercial purposes. Only 2- and 4-vinylpyridines are of industrial importance; Table 8 lists some physical properties.

Due to the electron-withdrawing effect of the ring nitrogen atom, 2- and 4-vinylpyridines act as electrophiles. Nucleophiles such as methoxide, cyanide, hydrogen sulfide, and others add to 2- and 4-vinylpyridine at the vinylic site to give addition products. The reactions shown below have been carried out on a commercial scale [55–57]:

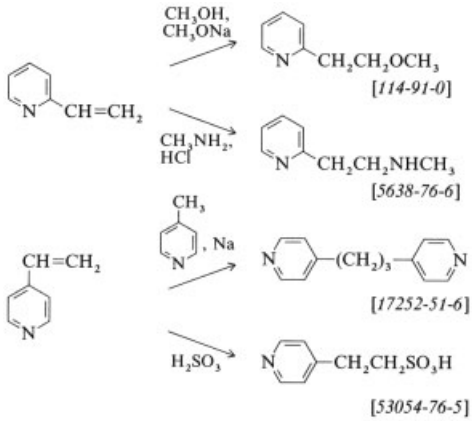

In addition to these Michael-type reactions, vinylpyridines are reduced at the side chain to give ethylpyridines [58]. Addition of chlorine to the vinyl group leads to (1,2-dichloroethyl)-pyridines [59].

Vinylpyridines are readily polymerized or copolymerized with styrene, butadiene, isobutylene, methyl methacrylate, and other compounds in the presence of radical, cationic, and anionic initiators. The homopolymer is soluble in organic solvents such as methanol and acetone, whereas cross-linked copolymers are insoluble in organic solvents.

Preparation. Industrially, 2- and 4-vinylpyridines are manufactured by treatment of 2- or 4-methylpyridine with aqueous formaldehyde, followed by dehydration of the resulting intermediate alcohol [60]:

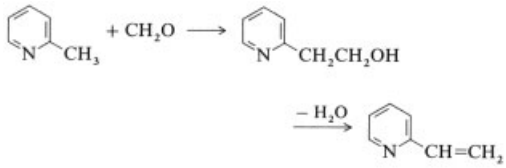

For the manufacture of 2-vinylpyridine, the reaction is carried out at 150 – 200 °C in an au-

toclave. The conversion must be kept relatively low with short reaction time to suppress the formation of byproducts. After removal of unreacted 2-methylpyridine by distillation, concentrated aqueous sodium hydroxide is added to the residue and the resultant mixture is distilled under reduced pressure. During distillation, the dehydration of 2-(2-pyridyl)ethanol occurs to give 2-vinylpyridine as a distillate, which can be purified further by fractional distillation under reduced pressure in the presence of an inhibitor such as 4-*tert*-butylcatecohol. 4-Vinylpyridine is manufactured by a similar method.

Uses. Among vinylpyridines, only 2-vinylpyridine is in large demand, as much as 3000 t annually. The polymer of 2-methyl-5-vinylpyridine [*140-76-1*] was formerly used as a coating material for medicine tablets; today, it is not used at all. Although 4-vinylpyridine has been used for various purposes, its total annual demand is presumed to be only several hundred tons.

Most of the 2-vinylpyridine is used in the production of a latex terpolymer of 2-vinylpyridine, styrene, and butadiene, used as a tire-cord binder.

The cross-linked resin made from 4-vinylpyridine and divinylbenzene has been used to remove poisonous hexavalent chromium ions from wastewater [61] by formation of complexes with the nitrogen atoms of the pendant pyridine rings. Similarly, this type of resin can be used to remove or recover phenol from wastewater [62]. Furthermore, the poly(4-vinylpyridines) are used as ion-exchange membranes and electrodes for reversible cells.

The addition product of methanol to 2-vinylpyridine, 2-(2-methoxyethyl)pyridine is a veterinary anthelmintic (methypidine) [63]. 4,4′-(1,3-Propanediyl)bispiperidine, made from 4-vinylpyridine, is used as a raw material for polyamide resins [64], and 2-(4-pyridyl)ethylsulfonic acid from 4-vinylpyridine is a coagulation accelerator for the gelatin layer of photographic plates [65].

3.2. Bipyridines

Bipyridines have become increasingly important as intermediates for herbicides since the 1960s. A significant part of pyridine is used for the manufacture of bipyridines (i.e., 2,2′- and 4,4′-bipyridine). The herbicides diquat [*85-00-7*] and paraquat (dichloride [*1910-42-5*]; bis(methylsulfate) [*2074-50-2*]) are produced by quaternization of 2,2′- and 4,4′-bipyridines, respectively. Their physical properties are listed in Table 9.

2,2′-Bipyridine is synthesized by the dimerization of pyridine in the presence of an oxidizing agent or catalyst, such as iron(III) chloride, iodine, or nickel – aluminum, or by the reaction of 2-bromopyridine with copper. For industrial production, Raney nickel is used, and the amount of catalyst employed is an indicator of the economy of the process. In Table 10, the historical development of the process is demonstrated in terms of the decrease in amount of catalyst employed.

Table 10. Amount of Raney nickel employed in synthesis of 2,2′-bipyridine

Method	Raney nickel, g/100 g 2,2′-bipyridine
[66]	595
[67]	143
[68]	25
[69]	17
[70]	14
[71]	8

Production of 4,4′-Bipyridine. As early as 1870, pyridine was known to react with sodium to give 4,4′-bipyridine after oxidation of the initial product [72]. The reduction of pyridine with zinc powder and acetic acid followed by oxidation was also known. Many patents have been applied for concerning the improvement of this procedure with respect to reaction temperature, choice of solvent, and isolation of the product. The ICI process, which seems at present to be the best, is described below [73].

Sodium is dissolved in liquid ammonia at −45 °C, and the solution is successively diluted with pyridine and *N,N*-dimethylformamide. The resulting mixture is poured into *N,N*-dimethylformamide at −25 °C, while air is passed through the solution. The mixture is then warmed to room temperature, and ammonia is allowed to evaporate. After work-up, 4,4′-bipyridine is obtained in 84 % yield.

Other variants are described in [74–77].

Table 9. Physical properties of bipyridines (M_r 156.18)

Compound	CAS registry no.	*mp*, °C	*bp*, °C	Solubility in water
2,2′-Bipyridine	[*366-18-7*]	70.1	272 – 273	sparingly soluble
2,3′-Bipyridine	[*581-50-0*]		295 – 296	practically insoluble
2,4′-Bipyridine	[*581-47-5*]	61.1 – 61.5	280 – 282	practically insoluble
3,3′-Bipyridine	[*581-46-4*]	68	291 – 292	readily soluble
3,4′-Bipyridine	[*4394-11-0*]	61	297	soluble in hot water
4,4′-Bipyridine	[*553-26-4*]	114	304.8	soluble in hot water

3.3. Quaternary Pyridinium Salts

Properties. Pyridine and some of its derivatives are readily converted to quaternary salts by alkylating agents such as alkyl halides. Quaternary salts can be regarded as being formed by the neutralization of strong bases (i.e., pyridinium hydroxides) with strong acids (e.g., hydrogen halides). Therefore, these salts are practically neutral when dissolved in water. Although stable under normal conditions, they are degraded into pyridine hydrohalides and alkenes on intense heating [78], or they undergo the Ladenburg rearrangement in the presence of copper as a catalyst [79]:

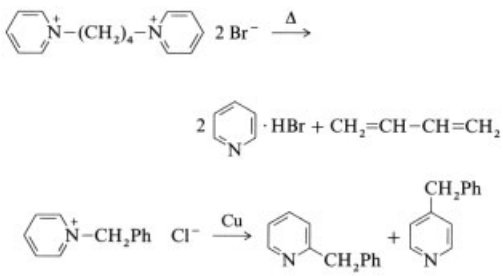

The pyridine ring of these salts is susceptible to nucleophilic attack as a result of the quaterniza-tion of the nitrogen atom. The reaction of 1-methylpyridinium salts with alkaline ferricyanide gives *N*-methylpyridone via oxidation of a pseudobase, which is considered to exist at low concentration in equilibrium with the pyridinium hydroxide [80, 81]:

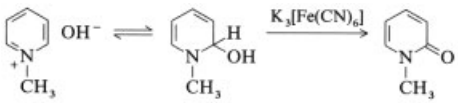

1-Pyridylpyridinium dichloride [*5421-92-1*], obtained from pyridine and thionyl chloride or chlorine, is particularly useful as a synthetic intermediate [82, 83]. It affords 4-amino- and 4-hydroxypyridine on ammonolysis and hydrolysis, respectively. Treatment with phenol and sodium phenoxide gives 4-phenoxypyridine, and reaction with hydrogen sulfide in pyridine gives 4-mercaptopyridine [84–87].

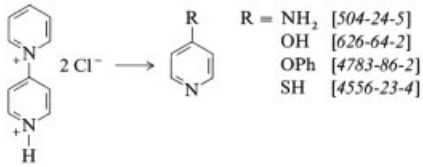

Uses. The major use of quaternary pyridinium salts is in the manufacture of the herbicides paraquat and diquat. These compounds are produced by the quaternization of 4,4′-bipyridyl and 2,2′-bipyridyl with methyl chloride and dibromoethane, respectively. Higher alkylpyridinium salts are used in the textile industry as dye auxiliaries and spin bath additives (antistatic agents and softeners). The higher alkylpyridinium salts also exhibit antimicrobial activity. Hexadecylpyridinium chloride [*123-03-5*] is a topical antiseptic, and amprolium [*121-25-5*], a quaternary salt of 2-methylpyridine, is a veterinary coccidiostat.

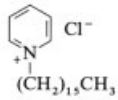

Hexadecylpyridinium chloride

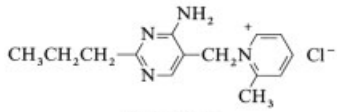

Amprolium

Of the many quaternary salts, 1-butylpyridinium bromide [*1124-64-7*] and other lower 1-

alkyl homologues are of interest. Although each component is solid at ambient temperature, the mixture of these salts with aluminum chloride leads to melts that can exist as liquids below room temperature in fairly wide proportions.

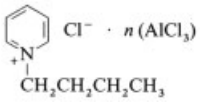

The molten salt with $n = 2$ exhibits a specific conductivity of ca. 7 mS/cm at 25 °C [88]. The utilization of these molten salts for battery electrolytes [89] and electroplating baths for aluminum has been proposed [90]. These binary salts are reportedly excellent solvents and catalysts for Friedel–Crafts reactions [91] and for the formylation of toluene by carbon monoxide [92].

3.4. Pyridine *N*-Oxides

Properties. Some physical properties of pyridine *N*-oxides are listed in Table 11.

Resonance structures of pyridine *N*-oxide are as follows:

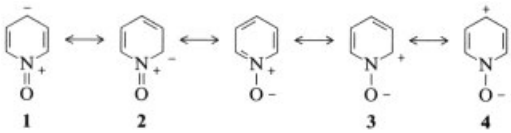

The *N*-oxide group in pyridine *N*-oxide has both electron-withdrawing and electron-donating effects. Consequently, pyridine *N*-oxide reacts with both electrophiles and nucleophiles, resulting in a more versatile reactivity of the pyridine ring compared to pyridine itself.

Nitration of pyridine *N*-oxide is a typical electrophilic reaction of the *N*-oxide. Pyridine *N*-oxide is nitrated under relatively mild conditions due to the contribution of structures **1** and **2**, giving 4-nitropyridine *N*-oxide in good yield [93, 94], in contrast to the nitration of pyridine, which requires drastic conditions and gives 3-nitropyridine only in poor yield. The nitro group of 4-nitropyridine *N*-oxide can be reduced to an amino group or displaced by various nucleophiles such as halides and alkoxides to give many useful 4-substituted pyridines. The oxygen of pyridine *N*-oxides can usually be removed by phosphorus trichloride.

Owing to the contribution of resonance structures **3** and **4**, the oxygen atom of the *N*-oxide group readily undergoes protonation, acylation, allylation, etc. Reaction with nucleophiles proceeds via quaternized intermediates, an example being the formation of 2-acetoxypyridine [*3847-19-6*] from the reaction of pyridine *N*-oxide with acetic anhydride [95]:

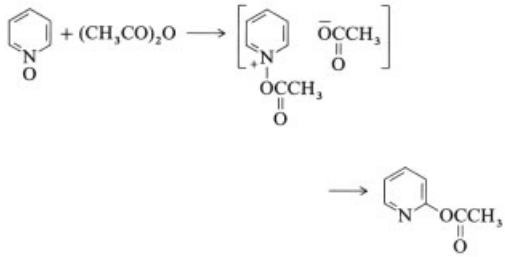

Trimethylsilyl cyanide, formed in situ from trimethylsilyl chloride and sodium cyanide, reacts with pyridine *N*-oxide to give almost exclusively 2-pyridinecarbonitrile [*100-70-9*] in 80 % yield [96]. Although the chlorination of pyridine *N*-oxide with sulfuryl chloride gives a mixture of 2- [*109-09-1*] and 4- chloropyridines [*626-61-9*] [97], treatment of 3-methylpyridine *N*-oxide with phosphoryl chloride in the presence of diisopropylamine yields predominantly 2- chloro-5-methylpyridine [*18368-64-4*] [98]. In the chlorination of 3-pyridinecarbonitrile *N*-oxide and nicotinic acid *N*-oxide, 2- chloro-3-pyridinecarbonitrile [*6602-54-6*] [99] and 2-chloronicotinic acid [*2942-59-8*] [100] are obtained as the main products.

The *N*-oxide of 2-methylpyridine reacts with acetic anhydride at the methyl group to give 2-acetoxymethylpyridine [*1007-49-4*] [101], which on hydrolysis leads to 2-pyridylmethanol [*586-98-1*].

Preparation. Pyridine *N*-oxide can be prepared by treatment of pyridine with 30 % hydrogen peroxide in acetic acid [102]. A similar and easier method is also known [103], in which molybdenum trioxide is used as catalyst and water as solvent. Pyridine *N*-oxide and its alkyl derivatives are isolated by distillation after complete decomposition of unreacted peroxides.

Table 11. Physical properties of pyridine N-oxides

Compound	CAS registry no.	*mp*, °C	*bp*, °C (kPa)	pK_a
Pyridine *N*-oxide	[*694-59-7*]	67	122 – 124 (0.67)	0.79 (24 °C)
2-Methylpyridine *N*-oxide	[*931-19-1*]		123 – 124 (2.0)	
3-Methylpyridine *N*-oxide	[*1003-73-2*]	37 – 38	146 – 149 (2.0)	1.08 (25°C)
4-Methylpyridine *N*-oxide	[*1003-67-4*]	186 – 188		1.29 (25°C)
Nicotinic acid *N*-oxide	[*2398-81-4*]	258		
3-Pyridinecarbonitrile *N*-oxide	[*14906-64-0*]	178		

Uses. Pyridine *N*-oxides are important as synthetic intermediates in the manufacture of pharmaceuticals and agrochemicals. The antiulcer agent omeprazole [*73590-58-6*] is produced from 2,3,5-trimethylpyridine *N*-oxide [*74409-42-0*]. Niflumic acid [*4394-00-7*] and pranoprofen [*52549-17-4*] are analgesics and anti-inflammatories, which are manufactured from nicotinic acid *N*-oxide, obtained either by *N*-oxidation of nicotinic acid or by hydrolysis of 3-pyridinecarbonitrile *N*-oxide. Zinc pyrithione [*13463-41-7*], the zinc salt of 2-pyridinethiol *N*-oxide [*1121-31-9*], is a fungicide derived from 2- chloropyridine *N*-oxide [*2402-95-1*].

3.5. Piperidines (see also → Amines, Aliphatic)

Physical properties of common piperidines are listed in Table 12.

Piperidines react with alkyl halides and acid anhydrides to give *N*-alkylpiperidines and amides, respectively.

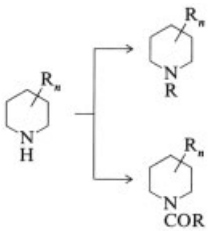

Piperidine and most of its derivatives are easily produced by hydrogenation of the corresponding pyridine derivatives at elevated temperature and pressure over nickel, palladium, or ruthenium catalysts [104–108].

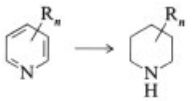

A major use of piperidine is in the production of dithiuram tetrasulfide [*120-54-7*], which is used as a vulcanization accelerator in the rubber industry. Other uses of piperidine are in the production of vasodilators such as dipyridamole and minoxidil , diuretics such as etozolin , and fungicides such as piperalin . Piperidine is also used as a solvent. 2-Methylpiperidine is used for the herbicide piperophos , and *cis*-2,6-dimethylpiperidine for the antiarrhythmic pirmenol . 4-Benzylpyridine is quaternized and hydrogenated to give a 4-benzylpiperidine derivative that is used as a cerebral vasodilator (ifenprodil tartrate) .

3.6. Halopyridines

Physical properties of common halopyridines are listed in Table 13.

The 2- and 4- chloropyridines react with various nucleophiles to give alkyl ether [109], alkyl thioether [110], and alkylamine derivatives [111]. Furthermore, pyridylation of phenylacetonitrile can be achieved by using strong base [112].

Because of its relatively low reactivity, 3-chloropyridine does not undergo nucleophilic attack easily.

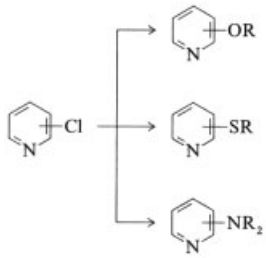

Haloalkylpyridines are useful intermediates for pyridylacetonitriles [113] or benzimidazolylthiomethylpyridines [114, 115].

Table 12. Physical properties of piperidines

Compound	CAS registry no.	M_r	*bp*, °C (kPa)	*mp*, °C	n_D^{20}	d_4^{20}	Flash point, °C
Piperidine	[*110-89-4*]	85.15	106		1.4525	0.861	4
2-Methylpiperidine	[*109-05-7*]	99.18	119		1.4459	0.844	8
3-Methylpiperidine	[*626-56-2*]	99.18	125		1.4470	0.845	17
4-Methylpiperidine	[*626-58-4*]	99.18	124		1.4458	0.838	7
cis-2,6-Dimethylpiperidine	[*766-17-6*]	113.20	126		1.4394	0.840	11
2-Piperidinecarboxylic acid	[*4043-87-2*]	129.16		282			
3-Piperidinecarboxylic acid	[*498-95-3*]	129.16		261			
4-Piperidinecarboxylic acid	[*498-94-2*]	129.16		> 300			
4-Piperidinol	[*5382-16-1*]	101.15	108 – 114 (1.3)				107
4-Benzylpiperidine	[*31252-42-3*]	175.28	279	6 – 7	1.5370	0.997	
2-Piperidylmethanol	[*3433-37-2*]	115.18		68 – 70			
2-(2-Piperidyl)ethanol	[*1484-84-0*]	129.20	234	38 – 40		1.010	102
1-Methylpiperidine	[*626-67-5*]	99.18	106 – 107		1.4378	0.816	3
1-Methyl-2-piperidylmethanol	[*20845-34-5*]	129.20	79 – 80 (0.9)		1.4823	0.984	81

Table 13. Physical properties of halopyridines

Compound	CAS registry no.	M_r	*bp*, °C (kPa)	*mp*, °C	n_D^{20}	d_4^{20}	Flash point, °C
2-Chloropyridine	[*109-09-1*]	113.55	166 (95)	− 46.5	1.5320	1.200	65
3-Chloropyridine	[*626-60-8*]	113.55	148		1.5330	1.194	65
4-Chloropyridine hydrochloride	[*7379-35-3*]	150.01		210			
2-Bromopyridine	[*109-04-6*]	158.00	192 – 194		1.5720	1.657	54
3-Bromopyridine	[*626-55-1*]	158.00	173		1.5700	1.640	51
4-Bromopyridine hydrochloride	[*19524-06-2*]	194-46		270			
2-Fluoropyridine	[*372-48-5*]	97.09	126 (100)		1.4660	1.128	28
2-Chloro-3-methylpyridine	[*18368-76-8*]	127.57	192 – 193 (100)				
2-Chloro-6-methylpyridine	[*23468-31-7*]	127.57	184/100		1.5270	1.167	73
2,3-Dichloropyridine	[*2402-77-9*]	147.99	203 – 204	65 – 67			
2,5-Dichloropyridine	[*16110-09-1*]	147.99	193 – 194	59 – 62			
2,6-Dichloropyridine	[*2402-78-0*]	147.99	211 – 212	86 – 89			
3,5-Dichloropyridine	[*2457-47-8*]	147.99	178 – 179	65 – 67			
2-Chloromethylpyridine hydrochloride	[*6959-47-3*]	164.04		125 – 129			
3-Chloromethylpyridine hydrochloride	[*6959-48-4*]	164.04		137 – 143			
4-Chloromethylpyridine hydrochloride	[*1822-51-1*]	164.04		166 – 173			
2,3,5,6-Tetrachloropyridine	[*2402-79-1*]	216.88	251 – 252	91 – 92			
Pentachloropyridine	[*2176-62-7*]	251.33	279 – 280	124 – 126			
2-Chloro-5-trifluoro-methylpyridine	[*52334-81-3*]	181.54	152 – 153	30 – 32			
2-Chloro-3-pyridinecarboxylic acid	[*2942-59-8*]	157.56		175			
6-Chloro-3-pyridinecarboxylic acid	[*5326-23-8*]	157.56		190			

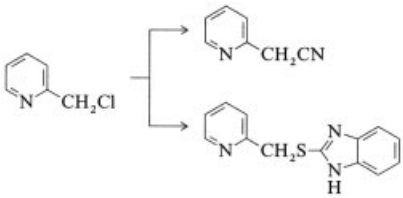

Halogenation of pyridines generally gives a mixture of chlorinated pyridines. For instance, direct chlorination of pyridine with molecular chlorine can be achieved above 270 °C to give 2-chloropyridine and 2,6-dichloropyridine [116–118].

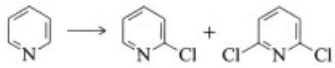

In the chlorination of methylpyridines, reaction occurs first at the side chain and then at a

ring position. Liquid-phase chlorination generally leads to chloromethylpyridines [119].

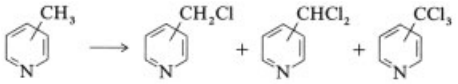

In contrast, gas-phase chlorination tends to give a mixture of trichloromethylchloropyridines [120–122]:

Exhaustive chlorination of pyridines gives pentachloropyridine [123, 124]:

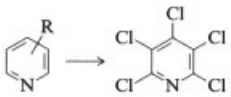

Pentachloropyridine can be reduced by zinc metal to 2,3,5,6-tetrachloropyridine because of enhanced reactivity at the 4-position [125, 126].

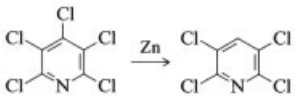

5-Trifluoromethyl-2- chloropyridine is produced directly from 3-methylpyridine by the combined action of chlorine and hydrogen fluoride [127]:

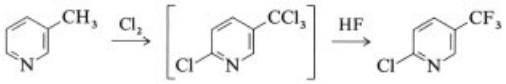

The preparation of 2- and 4- chloropyridines from pyridinols is achieved by the use of halogenating reagents such as phosphoryl chloride [128–130].

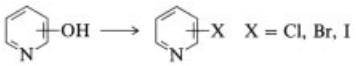

Pyridyl alcohols react with thionyl chloride or phosphorus halide reagents, such as phosphoryl chloride, phosphorus trihalide, or phosphorus pentahalide, to give the corresponding haloalkylpyridines [114, 115, 131].

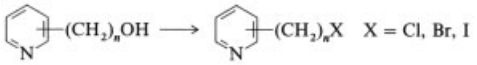

Diazotization of aminopyridines in the presence of halide ions is used for the preparation of halopyridines [132–135].

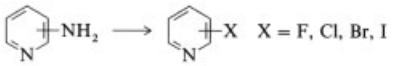

Chlorination of pyridine *N*-oxides leads to halopyridines; usually, 2- and 4-isomers are formed by the action of phosphorus halide reagents [98–100] (see Section 3.4):

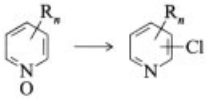

In industry, vapor-phase chlorination products are important intermediates for pharmaceuticals and agrochemicals.

A major use of 2- chloropyridine is the production of pyrithione, which is used widely as a fungicide. Another use is in the production of insecticides such as pyripropoxyfen. Pyridylation products of phenylacetonitrile lead to antihistamines such as chloropheniramine and antiarrhythmics such as disopyramide . Reaction of 4-chloropyridine with mercaptoacetic acid gives pyridylmercaptoacetic acid, which is a precursor for cephalosporin antibiotics (e.g., cephapirin sodium salt) . 2,6-Dichloropyridine is used for quinoline antibiotics such as enoxacin.

2-Chloro-6-trichloromethylpyridine is used to prepare nitrapyrin, which prevents the loss of ammonia from fertilizers. 2,5-Dichloro-6-trichloromethylpyridine is converted to the herbicide clopyralid . 2,3,4,5-Tetrachloro-6-trichloromethylpyridine is a precursor of the herbicide picloram . 2,3,5,6-Tetrachloropyridine is used for producing the insecticide chlorpyrifos, and 2- chloro-5-trifluoromethylpyridine for the herbicide fluazifop-butyl .

Uses of liquid-phase chlorination products are limited at present. Major uses of 2-chloropyridine-3- carboxylic acid are for the production of herbicides such as diflufenican

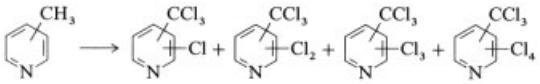

and anti-inflammatory agents such as niflumic acid and pranoprofen.

3.7. Pyridinecarbonitriles, Carboxylic Acids, and Carboxamides

The 3-carbonitrile, 3-carboxylic acid, and 3-carboxamide are important from both a physiological and an industrial point of view.

3-Pyridinecarboxylic acid and 3-pyridinecarboxamide are known as niacin. Niacin is widely distributed in plants and animals. It forms coenzymes, nicotinamide – adenine dinucleotide (NAD) and nicotinamide – adenine dinucleotide phosphate (NADP), which participate in oxidation – reduction cycles in living cells. Since 3-pyridinecarboxylic acid is obtained by oxidation of nicotine with nitric acid, it is commonly called nicotinic acid. 3-Pyridinecarbonitrile and 3-pyridinecarboxamide are referred to as nicotinonitrile and nicotinamide. The prefix isonicotin- is often used to denote the 4-position of pyridinecarbonitrile, - carboxylic acid, and - carboxamide.

Physical data for pyridinecarbonitriles, pyridinecarboxylic acids, and pyridinecarboxamides are listed in Table 14.

Pyridinecarbonitriles. Pyridinecarbonitriles are important intermediates for a variety of pyridine derivatives [136–141]:

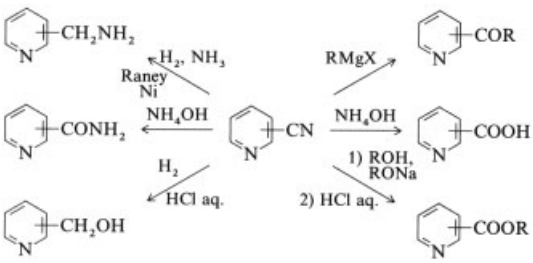

Pyridinecarbonitriles are usually manufactured by catalytic vapor-phase ammoxidation of alkylpyridines :

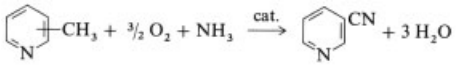

Generally, 1 to 20 mol of ammonia and 2 to 20 mol of oxygen are used per mole of alkylpyridine. Reaction temperatures range between 280 and 500 °C. Catalysts containing vanadium oxide are commonly used. Examples of the process are given in Table 15, and a flow sheet is illustrated in Figure 4.

3-Methylpyridine is ammoxidized to 3-cyanopyridine in the reactor (a). The reaction gas is quenched with water in the absorber (b), and the condensed mixture is extracted with a solvent in the extraction column (c). 3-Cyanopyridine is separated from the solvent and 3-methylpyridine by three-stage distillation.

3-Pyridinecarbonitrile is used for the production of 3-pyridinecarboxamide (vitamin complex), and 4-pyridinecarbonitrile for antituberculosis agents such as isoniazid [148].

Pyridinecarboxylic Acids. Pyridinecarboxylic acids behave like ordinary carboxylic acids. Some important reactions are shown below [149, 150]:

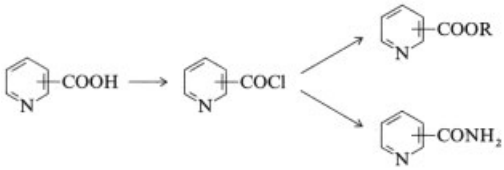

Two basic methods are used for the production of pyridinecarboxylic acids: hydrolysis of pyridinecarbonitriles and nitric acid oxidation of alkylpyridines.

Pyridinecarbonitriles are hydrolyzed to pyridinecarboxamides under basic conditions at 0 – 100 °C; subsequent hydrolysis to pyridinecarboxylic acids is carried out under more severe conditions [151, 152].

Table 14. Physical data of pyridinecarbonitriles, pyridinecarboxamides, and pyridinecarboxylic acids

Compound	CAS registry no.	M_r	*mp*, °C	*bp*, °C (kPa)
Nitriles				
2-Pyridinecarbonitrile	[*100-70-9*]	104.11	26	216 – 226
3-Pyridinecarbonitrile	[*100-54-9*]	104.11	51	198 – 202
4-Pyridinecarbonitrile	[*100-48-1*]	104.11	76 – 78	190 – 226
6-Methyl-2-pyridinecarbonitrile	[*1620-75-3*]	118.15	71 – 71.5	112 – 114 (2.0)
2,6-Pyridinedicarbonitrile	[*1452-77-3*]	129.12	104 – 105	143 (2.67)
Amides				
2-Pyridinecarboxamide	[*2893-33-6*]	122.13	126 – 127	
3-Pyridinecarboxamide	[*98-92-0*]	122.13	128 – 131	
4-Pyridinecarboxamide	[*1453-82-3*]	122.13	155 – 156	
Acids				
2-Pyridinecarboxylic acid	[*98-98-6*]	123.11	137 – 138	
3-Pyridinecarboxylic acid	[*59-67-6*]	123.11	236	
4-Pyridinecarboxylic acid	[*55-22-1*]	123.11	317	
2,3-Pyridinedicarboxylic acid	[*89-00-9*]	167.12	229 – 230	
2,4-Pyridinedicarboxylic acid	[*499-80-9*]	167.12	242 – 243	
2,6-Pyridinedicarboxylic acid	[*499-83-2*]	167.12	228	

Table 15. Synthesis of pyridinecarbonitriles

Company	Catalyst	Starting material	Conversion, %	Yield, %	Ref.
Lonza	V_2O_5	2-methylpyridine	62.7	44.2	[142]
	V_2O_5	3-methylpyridine	89.3	83.5	
	V_2O_5	4-methylpyridine	98.4	81.3	
Degussa	$Sb_2O_5 - V_2O_5 - TiO_2 -$ montmorillonite $- SiO_2$	3-methylpyridine	94	90	[143]
Yuki Gousei	$MoO_3 - V_2O_5$	3-methylpyridine	96.4	83.0	[144]
	$MoO_3 - Cr_2O_3 - Al_2O_3$	4-methylpyridine	97.3	82.3	
Takeda Chemical	$V_2O_5 - Sb_2O_5 - Cr_2O_3 - TiO_2$	3-methylpyridine	100	98.6	[145]
Koei Chemical	$V_2O_5 - P_2O_5 - SiO_2$	2-methylpyridine	97.6	79.3	[146]
	$V_2O_5 - P_2O_5 - SiO_2$	3-methylpyridine	96.1	82.3	
	$V_2O_5 - P_2O_5$	4-methylpyridine	99.5	94.0	
Nippon Shokubai	$V_2O_5 - Sb_2O_5 - TiO_2 - SiO_2 - SiC$	2-methylpyridine		73	[147]
	$V_2O_5 - Sb_2O_5 - TiO_2 - SiO_2 - SiC$	3-methylpyridine		85	
	$V_2O_5 - Sb_2O_5 - TiO_2 - SiO_2 - SiC$	4-methylpyridine		97	

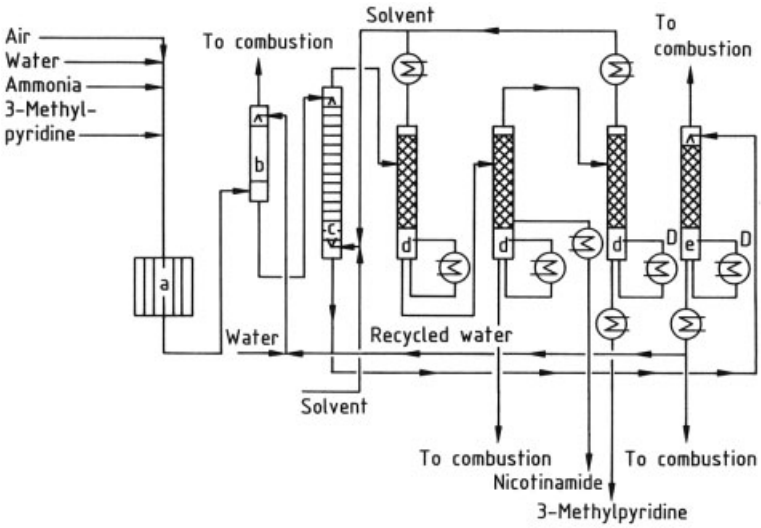

Figure 4. Ammoxidation of 3-methylpyridine
a) Multitubular reactor; b) Absorber; c) Extraction column; d) Fractionating columns; e) Treatment of process water

Nitric acid oxidation of alkylpyridines (e.g., 5-ethyl-2-methylpyridine to 3-pyridinecarboxylic acid) has been developed commercially by Lonza.

The flow diagram of the process is shown in Figure 5 [153]. 5-Ethyl-2-methylpyridine (MEP) is mixed with nitric acid and fed to a tubular reactor (b). The reaction to form 3-pyridine carboxylic acid nitrate is carried out at 180–370 °C and 2–50 MPa. Nitrogen oxides gases are recovered by the absorber (d) and recycled. After removal of water (e), crystallization of the nitrate (f), and neutralization with MEP (i), the free acid is crystallized (j) and dried. Unreacted nitric acid from the separators (g, k) is recycled to the reactor.

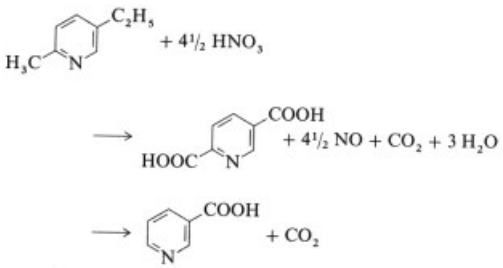

3-Pyridinecarboxylic acid has been used mainly as vitamin B_3 or vitamin PP for treating pellagra. Furthermore, it is used as an intermediate for the vasodilator nicorandil . Other uses are for plant growth regulators (inabenfide); antihistamines (terfenadine) ; and antidepressants (nialamide) . 2-Piperidinecarboxylic acid, derived from 2-pyridinecarboxylic acid, is used as intermediate for local anesthetics such as mepivacaine hydrochloride and bupivacaine hydrochloride . 2,3-Pyridinedicarboxylic acid is used as an intermediate for the herbicide imazapyr .

Pyridinecarboxamides. Pyridinecarboxamides are used for the production of pyridinecarboxylic acids [152] and aminopyridines [154].

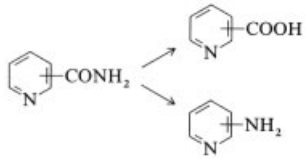

Pyridinecarboxamides are prepared by hydrolysis of pyridinecarbonitriles (see p. 415) [155].

3-Pyridinecarboxamide and 3-pyridinecarboxylic acid are important as vitamin B_3 [156]. 4-Pyridinecarboxamide is used for producing antibiotics (e.g., cefsulodin sodium) [157].

3.8. Aminopyridines

Table 16 lists commercially available aminopyridines and their physical properties.

3.8.1. 2-Aminopyridine

2-Aminopyridine [*504-29-0*] exists as tautomeric amino and imino forms. The amino form predominates over the imino form, and their ratio is generally 1000: 1.

2-Aminopyridine generally reacts with alkylating agents at the ring nitrogen to give derivatives of type **5**. In the presence of sodium amide or sodium methoxide, however, it reacts with alkylating agents at the exocyclic nitrogen. Such reactions are used to produce antihistamines of type **6**.

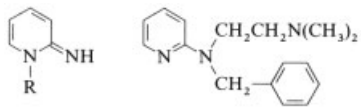

5 **6**

Electrophilic reactions such as halogenation and nitration occur at the positions *para* and *ortho* to the amino group [158–160]:

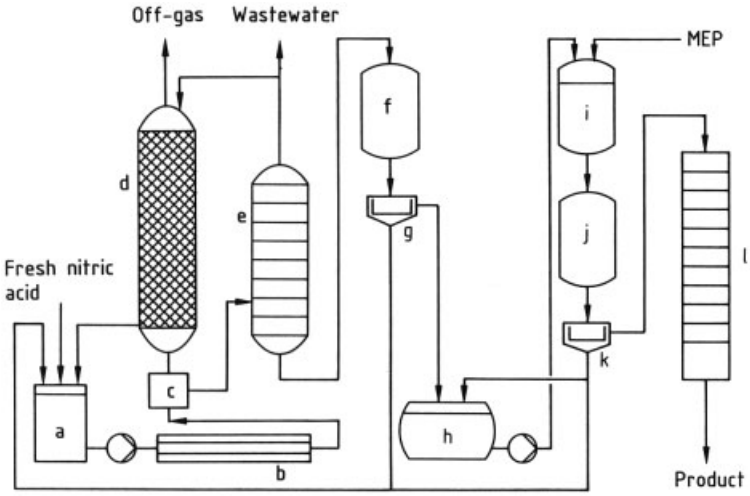

Figure 5. Lonza process for 3-pyridinecarboxylic acid
a) Mixing; b) Reactor; c) Pressure relief; d) Absorption column; e) Distillation column; f) Crystallization; g) Separator; h) Dissolution; i) Neutralization; j) Crystallization; k) Separator; l) Dryer

Table 16. Physical properties of aminopyridines

Compound	CAS registry no.	M_r	*mp*, °C	*bp*, °C (kPa)
2-Aminopyridine	[*504-29-0*]	94.12	59 – 60	210
3-Aminopyridine	[*462-08-8*]	94.12	64 – 65	260
4-Aminopyridine	[*504-24-5*]	94.12	159	273
2-Amino-6-methylpyridine	[*1824-81-3*]	108.14	40	208 – 209
2,6-Diaminopyridine	[*141-86-6*]	109.13	121 – 122	170 (4.0)
N,N-Dimethyl-4-pyridinamine	[*1122-58-3*]	122.17	112 – 113	145 – 150 (4.0)

Table 17. Physical properties of pyridinols

Compound	CAS registry no.	M_r	*bp*, °C (kPa)	*mp*, °C
2-Pyridinol	[*142-08-5*]	95.10	280 – 281	105 – 107
3-Pyridinol	[*109-00-2*]	95.10	151 – 153 (0.4)	126 – 129
4-Pyridinol	[*626-64-2*]	95.10	230 – 235 (1.6)	146 – 149
6-Methyl-2-pyridinol	[*3279-76-3*]	109.13		158 – 160
6-Methyl-3-pyridinol	[*1121-78-4*]	109.13		168 – 170
5-Chloro-2-pyridinol	[*4214-79-3*]	129.55		163 – 165
2-Nitro-3-pyridinol	[*15128-82-2*]	140.10		69 – 71

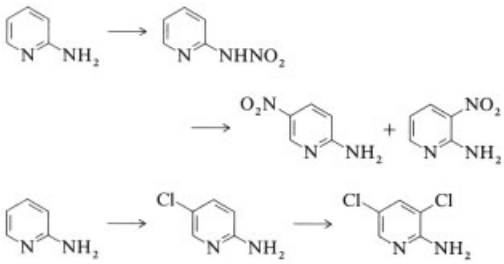

The most important method for the manufacture of 2-aminopyridine is the reaction of pyridine with sodium amide (Chichibabin amination). Upon hydrolysis of the intermediate sodium salt, 2-aminopyridine is obtained in high yield [161–164]:

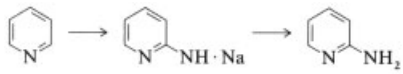

The major uses of 2-aminopyridine are as an intermediate for pharmaceuticals such as sulfapyridine [*144-83-2*], tripelennamine [*91-81-6*], piroxicam [*36322-90-4*], and tenoxicam [*59804-37-4*], as well as a variety of agrochemicals.

3.8.2. Other Aminopyridines

2-Amino-6-methylpyridine. 2-Amino-6-methylpyridine [*1824-81-3*] is produced from 2-methylpyridine and sodium amide. It is converted to nalidixic acid [*389-08-2*], an antibacterial agent [165].

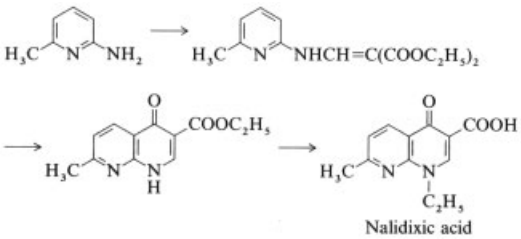

2,6-Diaminopyridine. Amination of pyridine or 2-aminopyridine under severe conditions gives 2,6-diaminopyridine [*141-86-6*], the coupling of which with benzenediazonium salts gives the antiseptic phenazopyridine [*136-40-3*] [166]. 2,6-Diaminopyridine is also utilized for the production of polyamides.

3-Aminopyridine and 4-Aminopyridine. 3-Aminopyridine [*462-08-8*] and 4-aminopyridine [*504-24-5*] are produced from the corresponding pyridinecarboxamides and sodium hypochlorite in alkaline solution (Hofmann reaction).

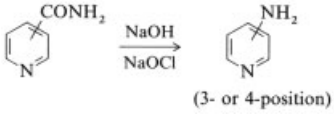

These aminopyridines are used as intermediates for pharmaceuticals such as pinacidil [*85371-64-8*] and for agrochemicals.

***N,N*-Dimethyl-4-pyridinamine** [*1122-58-3*] is widely employed as a supernucleophilic catalyst for many organic reactions. Two processes are used for its production:

1) *N*-(4-Pyridyl)pyridinium halide, prepared from pyridine and halogenating agents, is reacted with dimethylamine or *N,N*-dimethylformamide to give *N,N*-dimethyl-4-pyridinamine [167–170].

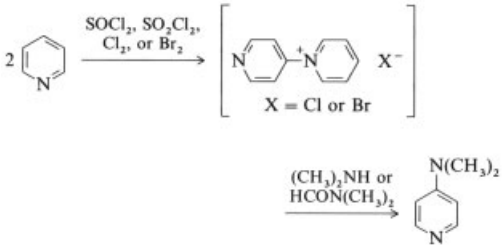

2) 1-[2-(2-Pyridyl)ethyl]-4- cyanopyridinium chloride, prepared from 4-pyridinecarbonitrile, 2-vinylpyridine, and hydrogen halide, gives *N,N*-dimethyl-4-pyridinamine by a reaction similar to that described above [171].

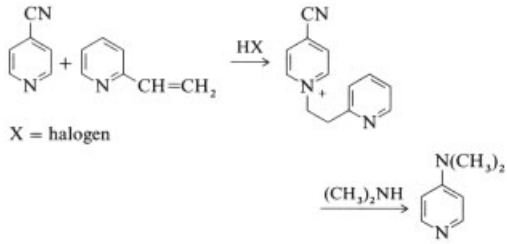

The major uses are as a catalyst for acylation, alkylation, halogenation, cyanation, and silylation, and as an accelerator in the manufacture of polyurethanes.

3.9. Pyridinols

Physical properties of typical pyridinols are listed in Table 17.

Tautomerization of pyridinols is discussed in [172]; both 2- and 4-pyridiols prefer the ketonic pyridone tautomer to the hydroxylic tautomer.

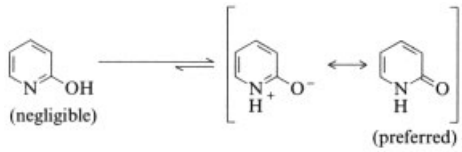

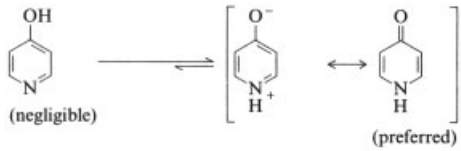

In contrast, 3-pyridinol is a phenolic compound because the ketonic tautomer does not exist.

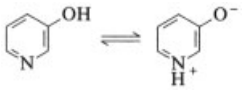

Chlorination of 2- and 4-pyridinols with phosphorus halides gives halopyridines [126–128].

Electrophilic substitution of pyridinols is facile because of the electron-donating nature of the hydroxyl group. Reaction occurs at positions *ortho* or *para* to the hydroxyl group, similar to phenols [173–176].

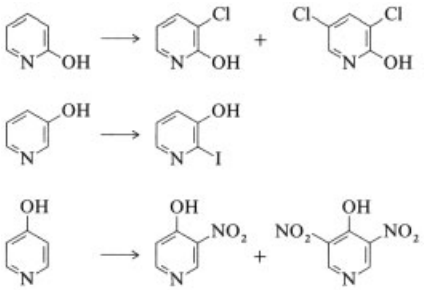

When 2- and 4-pyridinols are alkylated with alkyl halides, they give a mixture of the *O*- and *N*-alkylated products [177], but 3-pyridinol gives only the *O*-alkylated product [178].

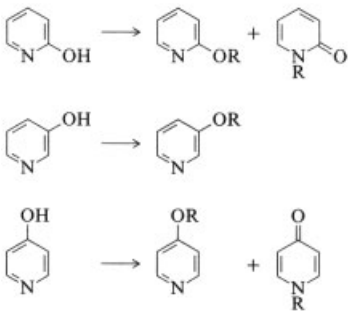

Diazotization of aminopyridines is used for the production of pyridinols [179–182].

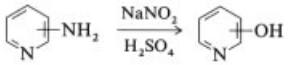

2-Pyridinol is also obtained from pyridine *N*-oxide by reaction with acid anhydrides and subsequent hydrolysis [183, 184].

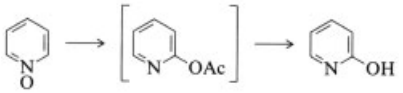

2-Alkyl-3-pyridinols are reported to be formed from alkanoylfurans and ammonia under pressure [185–187].

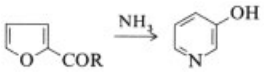

2-Pyridinol is a synthetic intermediate for the tranquilizer amphenidone . 3-Pyridinol is used for synthesizing the cholinesterase inhibitors pyridostigmine bromide and distigmine bromide . 4-Pyridinol is a precursor of the X-ray contrast agent propyliodone. It is hydrogenated to give 4-piperidinol, which is used to produce the psychotropic agent pericyazine.

3.10. Pyridyl Alcohols

Physical properties of common pyridyl alcohols are listed in Table 18.

Reactivities of pyridyl alcohols are similar to ordinary alcohols; for example, chloroalkylpyridines are formed by the action of thionyl chloride or phosphorus halide reagents [130, 131], and dehydration gives alkenylpyridines.

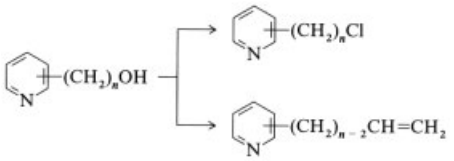

Pyridylmethanols. The simple pyridylmethanols are produced by catalytic hydrogenation of pyridinecarbonitriles in acidic media [188].

Table 18. Physical properties of pyridyl alcohols

Compound	CAS registry no.	M_r	*bp*, °C (kPa)	*mp*, °C	n_D^{20}	d_4^{20}	Flash point, °C
2-Pyridylmethanol	[*586-98-1*]	109.13	112 – 113 (0.5)		1.5440	1.131	110
3-Pyridylmethanol	[*100-55-0*]	109.13	154 (2.9)		1.5460	1.124	130
4-Pyridylmethanol	[*586-95-8*]	109.13	107 – 110 (0.1)	57 – 59			112
2,6-Pyridyldimethanol	[*1195-59-1*]	139.15	185 (2.0)	114 – 116			
2-(2-Pyridyl)ethanol	[*103-74-2*]	123.16	114 – 116 (1.2)		1.5370	1.093	92
2-(4-Pyridyl)ethanol	[*5344-27-4*]	123.16	262				160

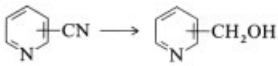

Substituted 2-pyridylmethanols can be produced from 2-methylpyridine *N*-oxides by the action of acetic anhydride, followed by hydrolysis [130, 131, 189].

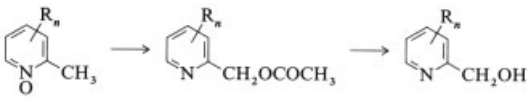

2-Pyridylmethanol is used in the preparation of the anti-inflammatory ibuprofen piconol . 3-Pyridylmethanol is used as an intermediate for the vasodilator nicotinic alcohol. 2,6-Pyridyldimethanol is carbamated to give the antiarteriosclerotic agent pyridinol carbamate.

Pyridylethanols. Condensation of 2- or 4-methylpyridines with formaldehyde or other aliphatic aldehydes gives the corresponding pyridylethanols due to the high acidity of the methyl group in the 2- or 4-position [18, 190–193].

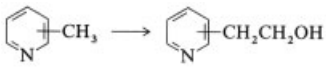

Pyridylethanols are used mainly for the manufacture of vinylpyridines (see Section 3.1). Other uses of 2-(2-pyridyl)ethanol are in the production of a psychotropic agent thioridazine and the antispasmodic tiquizium bromide .

3.11. Pyridinecarbaldehydes

Physical properties of typical pyridinecarbaldehydes are listed in Table 19.

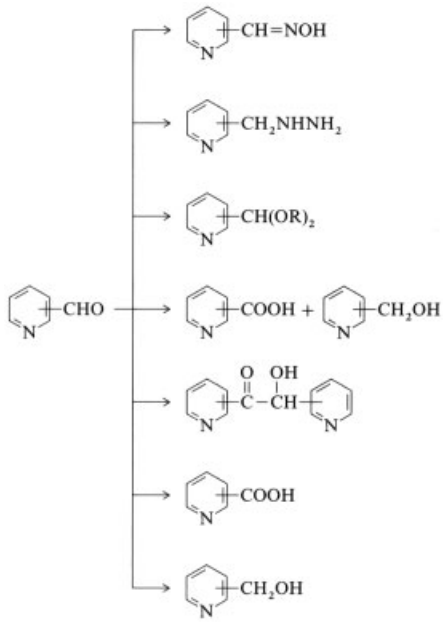

Reactions of pyridinecarbaldehydes are similar to those of aromatic aldehydes: benzoin condensation [194], oxidation to pyridinecarboxylic acids [195], reduction to pyridylmethanols [196], oximation [197], and hydrazone formation [198].

Pyridinecarbaldehydes are prepared from pyridylmethanols by oxidation with manganese dioxide [199].

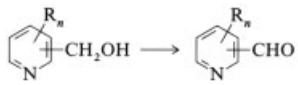

The oxidation of methylpyridines with selenium dioxide gives pyridinecarbaldehydes [200], which are also obtained by vapor-phase oxidation of methylpyridines over avanadium – molybdenum – titanium oxide catalyst [201].

Table 19. Physical properties of pyridylcarbaldehydes

Compound	CAS registry no.	M_r	*bp*, °C (kPa)	*mp*, °C	n_D^{20}	d_4^{20}	Flash point, °C
2-Pyridinecarbaldehyde	[*1121-60-4*]	107.11	181		1.5370	1.126	54
3-Pyridinecarbaldehyde	[*500-22-1*]	107.11	95 – 97 (0.9)		1.5490	1.135	60
4-Pyridinecarbaldehyde	[*872-85-5*]	107.11	77 – 78 (1.6)		1.5440	1.122	54
2-Pyridine aldoxime	[*873-69-8*]	122.13		110 – 112			
3-Pyridine aldoxime	[*1193-92-6*]	122.13		150 – 153			
4-Pyridine aldoxime	[*696-54-8*]	122.13		130 – 133			

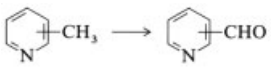

The catalytic partial reduction of 3-pyridinecarbonitrile with hydrogen over a palladium catalyst in acidic media gives 3-pyridinecarbaldehyde [202].

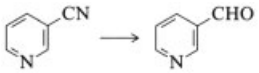

Quaternary salts of pyridinecarbaldehyde oximes, particularly pralidoxime methiodide, are used as antidotes for poisoning by organophosphate acetylcholinesterase inhibitors. Another use of 2-pyridinecarbaldehyde is in the production of bisacodyl , a laxative.

3.12. Pharmaceuticals and Agrochemicals

Typical pharmaceuticals and agrochemicals derived from pyridine are listed in Table 20.

4. Toxicology

4.1. Acute Toxicity

The acute toxicity of pyridine and its common derivatives is listed in Table 21.

Pyridine has a narcotic action. Toxic doses cause weakness, ataxia, unconsciousness, and salivation. In humans, pyridine is readily absorbed through the lungs, gastrointestinal tract, and skin. In 1893 a human fatality resulted from the accidental ingestion of half a cupful of commercial pyridine [203]. OSHA states that a vapor concentration of 3600 ppm in air is immediately dangerous to life and health [204]. The LDLo (oral) in humans is 500 mg/kg [205].

The acute toxicity of monoalkylpyridines is similar to that of pyridine, but some dimethyl- and trimethylpyridines are several times more toxic than pyridine. 2-Methylpyridine is one of the most important commercially produced alkylpyridines; its toxicity and safety data are summarized in [206].

Of aminopyridines, the toxicity of 4-aminopyridine has been thoroughly investigated because of its high toxicity and its use as a bird repellent. The acute toxicity of 4-aminopyridine has been determined for 41 species of birds and mammals. The oral LD_{50} values are generally < 10 mg/kg [207], and the LDLo (oral) in humans is 0.59 mg/kg. Most of the pharmacological effects of 4-aminopyridine can be attributed to its neurotoxicity [208].

The calculated fatal dose of 2-aminopyridine for a man of 70-kg weight is ca. 5 g [209]. A fatal accident involving 2-aminopyridine was described as follows: an employee died 1.5 h after spilling 2-aminopyridine on his clothes during its distillation and continuing to work. Aminopyridines are toxic if inhaled, swallowed, or absorbed through skin [209].

The acute toxicity of 13 chlorinated pyridines has been investigated [210], including monochloro to pentachloro derivatives. In the comparison of relative toxicity to mice by intraperitoneal injection, 2,6-dichloropyridine shows the highest toxicity ($LD_{50} = 115$ mg/kg) and 2,5-dichloropyridine the lowest ($LD_{50} = 1690$ mg/kg).

The toxicity of 2-chloropyridine to rabbits after dermal application ($LD_{50} = 64$ mg/kg) and by intraperitoneal injection ($LD_{50} = 48$ mg/kg) indicates that 2-chloropyridine is readily absorbed through the skin.

Table 20. Pharmaceuticals and agrochemicals derived from pyridine

Compound	Common name	Developer	Use
Pharmaceuticals			
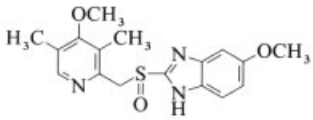	omeprazole [*73590-58-6*]	Astra	antiulcer
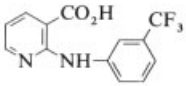	niflumic acid [*4394-00-7*]	Heydon	anti-inflammatory
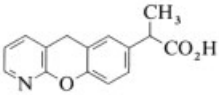	pranoprofen [*52549-17-4*]	Yoshitomi	anti-inflammatory
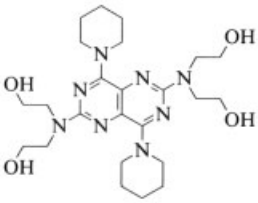	dipyridamole [*58-32-2*]	Thomae	vasodilator
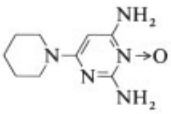	minoxidil [*38304-91-5*]	Upjohn	vasodilator
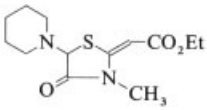	etozolin [*73-09-6*]	Warner-Lambert	diuretic
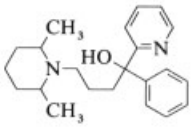	pirmenol [*61477-94-9*]	Parke, Davis	antiarrhythmic
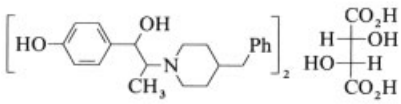	ifenprodil tartrate [*23210-56-2*]	Funai	cardiovascular
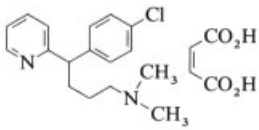	chloropheniramine maleate [*2438-32-6*]	Schering	antihistamine
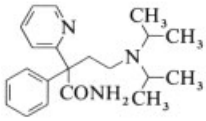	disopyramide [*3737-09-5*]	Roussel Uclaf	antiarrhythmic
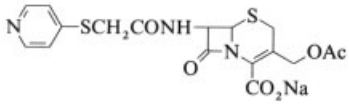	cephapirin sodium [*24356-60-3*]	Bristol-Myers	antibiotic

Table 20. Continued

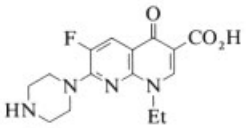	enoxacin [*74011-58-8*]	Dainippon	antibiotic
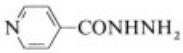	isoniazid [*54-85-3*]	Bayer	antituberculous
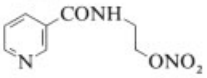	nicorandil [*65141-46-0*]	Tyugai	vasodilator
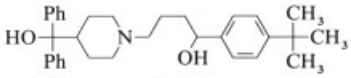	terfenadine [*50679-08-8*]	Merrell Dow	antihistamine
	nialamide [*51-12-7*]	Pfizer	monoamine oxidase
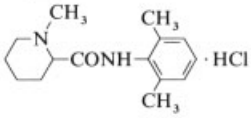	mepivacaine hydrochloride [*1722-62-9*]	Astra	local anesthetic
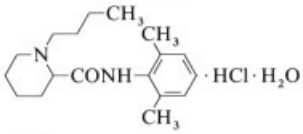	bupivacaine hydrochloride [*14252-80-3*]	Astra	local anesthetic
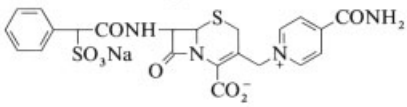	cefsulodin sodium [*52152-93-9*]	Takeda	antibiotic
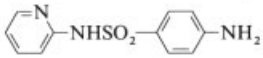	sulfapyridine [*144-83-2*]	Visuvia	chemotherapeutic
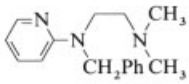	tripelennamine [*91-81-6*]	Ciba-Geigy	antihistamine
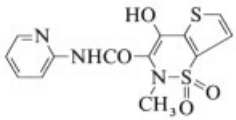	tenoxicam [*59804-37-4*]	Roche	anti-inflammatory
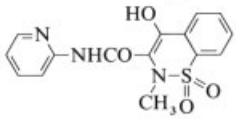	piroxicam [*36322-90-4*]	Pfizer	anti-inflammatory
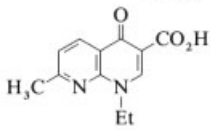	nalidixic acid [*389-08-2*]	Winthrop	chemotherapeutic
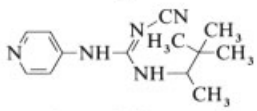	pinacidil [*85371-64-8*]	Leo Denmark	antihypertensive
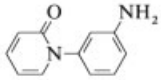	amphenidone [*134-37-2*]	Wallace & Tiernan	tranquilizer

Table 20. Continued

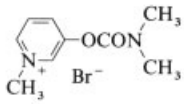	pyridostigmine bromide [*101-26-8*]	Roche	autonomic anticholinesterase
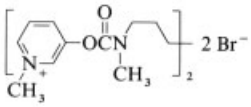	distigmine bromide [*15876-67-2*]	Chemie Linz	autonomic anticholinesterase
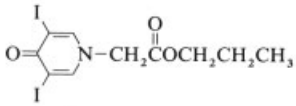	propyliodone [*587-61-1*]	Glaxo	X-ray contrast
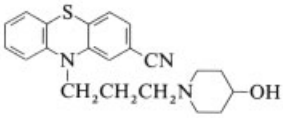	pericyazine [*2622-26-6*]	Rhône-Poulenc	psychotropic
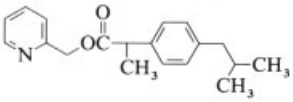	ibuprofen piconol [*112017-99-9*]	Hisamitu/Torii	anti-inflammatory
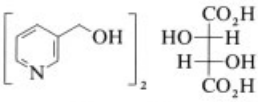	nicotinyl tartrate [*100-55-0*]	Roche	vasodilator
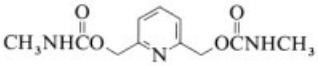	pyridinol carbamate [*1882-26-4*]	Banyu	antiarteriosclerotic
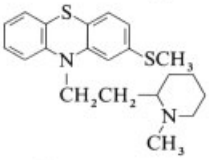	thioridazine [*50-52-2*]	Sandoz	psychotropic
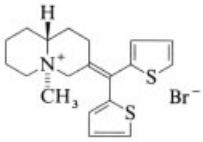	tiquizium bromide [*71731-58-3*]	Hokuriku	antispasmodic
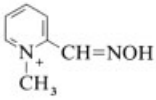	pralidoxime iodide [*51-15-0*]	Wyeth-Ayerst	antidote
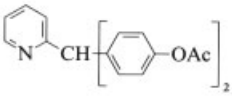	bisacodyl [*603-50-9*]	Thomae	laxative
Agrochemicals			
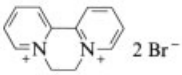	diquat dibromide [*85-00-7*]	ICI	herbicide
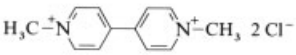	paraquat dichloride [*1910-42-5*]	ICI	herbicide

Table 20. Continued

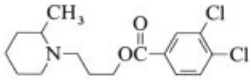	piperalin [*3478-94-2*]	Eli Lilly	fungicide
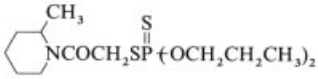	piperophos [*24151-93-7*]	Ciba-Geigy	herbicide
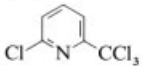	nitrapyrin [*1929-82-4*]	Dow	prevents loss of fertilizer
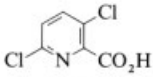	clopyralid [*1702-17-6*]	Dow	herbicide
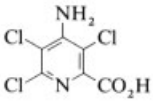	picloram [*1918-02-1*]	Dow	herbicide
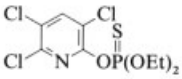	chlorpyrifos [*2921-88-2*]	Dow	insecticide
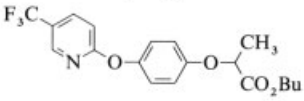	fluazifop-butyl [*69806-50-4*]	Ishihara	herbicide
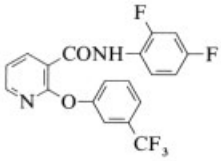	diflufenican [*83164-33-4*]	May & Baker	herbicide
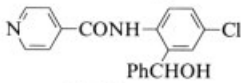	inabenfide [*82211-24-3*]	Tyugai	growth regulator
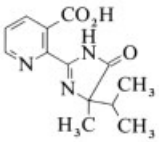	imazapyr [*81334-34-1*]	ACC	herbicide
Others			
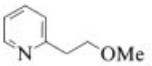	metyridine [*114-91-0*]	ICI	anthelmintic
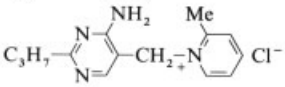	amprolium [*121-25-5*]	Merck	coccidiostat
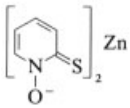	zinc pyrithione [*13463-41-7*]	Olin	fungicide
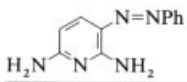	phenazopyridine [*136-40-3*]	Gödecke	antiseptic

Table 21. Acute toxicity of pyridine and derivatives∗

Compound	CAS registry no.	LD_{50} (oral), mg/kg	LCLo (inhalation), ppm/h	LD_{50} (i.p.), mg/kg	LD_{50} (i.v.), mg/kg
Pyridine	[*110-86-1*]	R: 891	R (LC_{50}) : 9010/1	R: 866	
Pyridine HCl	[*628-13-7*]	R (LDLo): 1600		R: 800	
2-Methylpyridine	[*109-06-8*]	R: 790	R: 4000/4	M: 529	
3-Methylpyridine	[*108-99-6*]	R: 400	R: 8700/2	M: 596	M: 298
4-Methylpyridine	[*108-89-4*]	R: 1290	R: 100/4	M: 335	
2,4-Dimethylpyridine	[*108-47-4*]	R: 200			
2,6-Dimethylpyridine	[*108-48-5*]	R: 400	R: 7500/1		
2-Methyl-5-ethylpyridine	[*104-90-5*]	R: 368	R: 1000/4		
2,4,6-Trimethylpyridine	[*108-75-8*]	R: 400	R: 2500/2		
2-Chloropyridine	[*109-09-1*]	M: 110	R: 100/4	M: 130	
2,3-Dichloropyridine	[*2402-77-9*]			M: 135	
2,5-Dichloropyridine	[*16110-09-1*]			M: 1690	
2,6-Dichloropyridine	[*2402-78-0*]	M: 176		M: 115	
Pentachloropyridine	[*2176-62-7*]			M: 235	
2-Chloro-6-(trichloromethyl)-pyridine	[*1929-82-4*]	R: 940, M: 710			
2-Chloromethylpyridine HCl	[*6959-47-3*]	R: 316, M: 316			
3-Chloromethylpyridine	[*6959-48-4*]	R: 316, M: 316			
2-Aminopyridine	[*504-29-0*]	R: 200	Hum (TCLo): 5/5	M: 28	R: 29
3-Aminopyridine	[*462-08-8*]	Q: 178		M: 28	M: 24
4-Aminopyridine	[*504-24-5*]	R: 21, Hum (LDLo): 0.59		R: 6.5	
2,3-Diaminopyridine	[*452-58-4*]			M: 25	
2,6-Diaminopyridine	[*141-86-6*]			M: 100	M: 56
4-Dimethylaminopyridine	[*1122-58-3*]	M (LDLo): 470			M: 56
2-Aminomethylpyridine	[*3731-51-9*]	Q: 750			M: 340
2-Vinylpyridine	[*100-69-6*]	R: 100	R: 5500/1.5		
4-Vinylpyridine	[*100-43-6*]	R: 100	R: 2000/2		
3-Cyanopyridine	[*100-54-9*]	R: 1185		M (LD_{25}): 800	
Picolinic acid	[*98-98-6*]	Q: 562		M: 360	M: 487
Nicotinic acid	[*59-67-6*]	R: 7000		M: 358	M: 5000
Isonicotinic acid	[*55-22-1*]	R: 5000, M: 3123			M: 5000
Methyl nicotinate	[*93-60-7*]			M (LDLo): 2000	
Ethyl isonicotinate	[*1570-45-2*]				M: 56
Nicotinamide	[*98-92-0*]	R: 3500		M: 2050	R: 2200
Isonicotinic acid hydrazide	[*54-85-3*]	R: 1250		M: 100	M: 149
Pyridine 1-oxide	[*694-59-7*]	Q: 1000		M: 1425	M: 180
2-Pyridinemethanol	[*586-98-1*]	Q: 1000			M: 1000
2-Acetylpyridine	[*1122-62-9*]	R: 2280			
2-Benzoylpyridine	[*91-02-1*]			M: 475	
Nicotinaldehyde	[*500-22-1*]			M: 720	
2-Pyridinaldoxime	[*873-69-8*]			M: 200	
2-Hydroxypyridine	[*142-08-5*]			M: 410	M: 750
2-Pyridinethione	[*2637-34-5*]	M: 533		M: 250	M: 250
2,2′-Bipyridine	[*366-18-7*]	R: 100		M: 200	
Diquat dibromide	[*2764-72-9*]	R: 215			
4,4′-Bipyridine	[*553-26-4*]	R: 172			
Paraquat dichloride	[*1910-42-5*]	R: 57, Hum (LDLo): 43		R: 26	R: 21
4-Phenylpyridine	[*939-23-1*]				M: 89
Hexadecylpyridinium chloride	[*6004-24-6*]	R: 200		M (LDLo): 3	
4-Nitropyridine *N*-oxide	[*1124-33-0*]	R: 107			
Nicotine	[*54-11-5*]	R: 50, Hum (LDLo): 0.88			
Niflumic acid	[*4394-00-7*]	R: 250, M: 375		M: 196	M: 152
Piperidine	[*110-89-4*]	R: 400, M: 30	R: 4000/4	M: 50	

∗ R = rat; M = mouse; Q = quail; Hum = human.

3-Acetylpyridine causes central neurotoxic effects upon intraperitoneal injection of 70–80 mg/kg in rats [211]. The zinc salt of 1-hydroxypyridine-2-thione (HPT), known as zinc pyrithione (ZnPT), is used as an antidandruff agent. The oral LD_{50} of HPT in mice is 533 mg/kg. The metal complexes (Zn, Cd) are more toxic than HPT itself. Percutaneous and eye toxicity is also discussed in [212].

Paraquat (1,1′-dimethyl-4,4′-bipyridinium dichloride), which consumes the largest amount of pyridine as starting material, exhibits very high toxicity. The oral LDLo in humans is 43 mg/kg. Nicotine is one of the well-known pyridine derivatives from natural sources, and its toxicity has been investigated in detail. The LDLo in humans is 0.882 mg/kg. Reports on the toxicity of paraquat and nicotine can be found in [213].

Vinylpyridines irritate the skin intensely, and contact with the vapor or liquid generally causes inflammation or blisters. They also show high toxicity by inhalation.

4.2. Subacute and Chronic Toxicity

Few long-term studies can be found on the toxicity of pyridine to animals. Six rats were fed a basal diet containing 0.1 % pyridine. Five rats died during the 14th to 32nd days. The livers and kidneys revealed acute lesions and some of the livers exhibited pronounced cirrhosis [214].

When humans are exposed to pyridine vapor, the toxicity affects mainly the central nervous system and the gastrointestinal tract. Symptoms are headache, dizziness, nausea, insomnia, anorexia, and weakness of the limbs. As examples, the cases of a chemist who had worked with pyridine for six months and of a man who had worked with it for 13 years are reported in [215]. Formerly, pyridine was used as an anticonvulsant, and an epileptic who received 1.8–2.5 mL/d of pyridine for one month died as a result of liver and kidney damage [216].

In 1983 Langston first reported that 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) causes acute symptoms of parkinsonism [217]; more than 1000 papers have been published, mainly in the United States, in the following years [218].

4.3. Mutagenicity and Ecotoxicity

Mutagenic pyridine derivatives include 2,2′-bipyridine, 4-nitropyridine *N*-oxide derivatives, chloropyridine derivatives,chloromethylpyridine derivatives, 2,6-diaminopyridine, 4-phenylpyridine, 2-hydroxy-3-nitropyridine, isonicotinic acid hydrazide, and 2-pyridine aldoxime.

Of these compounds, chloromethylpyridine derivatives [219] and 4-nitropyridine *N*-oxide derivatives [220] have been investigated with respect to their carcinogenicity.

Pyridine has a characteristic disagreeable, nauseating odor, detectable at < 1 ppm. The odor threshold is 0.021 ppm [221]. TLV and MAK values (in ppm) in several countries are as follows:

Australia	5
Germany	5
Italy	2
Japan	5
CIS	1.7
United States	5

The acute toxicity of pyridine for *Daphnia magna* is 240 mg/L, and the 24-h LC_{50} is 170 mg/L for newborns and 944 mg/L for adults. Reported acute toxic concentrations in fish vary widely and range from 26 to 1300 mg/L [204].

5. References

1. T. Anderson, *Justus Liebigs Ann. Chem.* **60** (1846) 86.
2. T. Anderson, *Justus Liebigs Ann. Chem.* **80** (1851) 44.
3. W. Ramsay, *Ber. Dtsch. Chem. Ges.* **10** (1877) 736.
4. A. Hanzsch, *Justus Liebigs Ann. Chem.* **215** (1882) 72.
5. A. E. Chichibabin, *Zh. Russ. Fiz. Khim. O-va.* **37** (1905) 1229.
6. R. Kuhn, G. Wendt, *Ber. Dtsch. Chem. Ges. B* **71** (1938) 1118.
7. P. Karrer, H. Heller, *Helv. Chim. Acta* **22** (1939) 1292.
8. A. W. Singer, S. M. McElvain, *Org. Synth.,* **coll. vol. 3** (1955) 740.
9. C. F. Koelsch, A. F. Steinhauer, *J. Org. Chem.* **18** (1953) 1516.
10. J. Braun, A. Pelgold, J. Zeemann, *Ber. Dtsch. Chem. Ges.* **55** (1922) 3783.

11. H. J. den Hertog, J. Overhoff, *Recl. Trav. Chim. Pays-Bas* **49** (1930) 552.
12. H. Meyer, W. Ritter, *Monatsh. Chem.* **35** (1914) 765.
13. A. E. Chichibabin, A. Zeide, *Zh. Russ. Fiz. Khim. O-va.* **46** (1914) 1216.
14. R. A. Abramovich, F. Helmer, J. G. Saha, *Can. J. Chem.* **43** (1965) 725. R. A. Abramovich, J. G. Saha, *Adv. Heterocycl. Chem.* **6** (1966) 229.
15. M. G. Reinecke, L. R. Kray, *J. Am. Chem. Soc.* **86** (1964) 5355.
16. P. Riehm, *Justus Liebigs Ann. Chem.* **238** (1887) 1.
17. J. Herzenberg, R. Covini, M. Pieroni, A. Nenz, *Ind. Eng. Chem. Prod. Res. Dev.* **6** (1967) no. 3, 195 – 197.
18. Allied Chem. & Dye Co., US 2 556 845, 1947.
19. A. E. Chichibabin, *J. Prakt. Chem.* **107** (1924) 122.
20. Koei Chemical Co., DE 1 770 870, 1967.
21. Nippon Kayaku, JP 71 39 873, 1969; *Chem. Abstr.* **76** (1972) 34112x.
22. H. Beschke, H. Friedrich, *Chem. Ztg.* **101** (1977) 377 – 384.
23. ICI, DE-AS 1 917 037, 1968.
24. Nippon Kayaku, JP 70 39 545, 1967; *Chem. Abstr.* **74** (1971) 141558c.
25. Koei Chemical Co., BE 758 201, 1969.
26. Daicel Chemical, JP 86 14 859, 1979.
27. ICI, BE 845 405, 1975.
28. Rütgerswerke, DE 2 203 384, 1972.
29. Nepera Chemical Co. Inc., US 512 834, 1983.
30. Koei Chemical Co., EP-A 232 182, 1987.
31. Koei Chemical Co., JP-Kokai 61 53 265, 1984.
32. I. Liepina et al., *Latv. PSR Zinat. Akad. Vestis, Kim. Ser.* 1989, no. 3, 334 – 338; *Chem. Abstr.* **111** (1989) 232533q.
33. Koei Chemical Co., DE-OS 2 064 397, 1984.
34. Degussa, DE-OS 2 703 069, 1978.
35. Distillers, GB 817 038, 1957.
36. Degussa, DE-OS 2 639 702, 1976.
37. Stamicarbon N.V., NL 7 013 453, 1970.
38. Stamicarbon N.V., NL 7 809 552, 1978.
39. Dynamit Nobel, DE-OS 2 514 004, 1975. Dynamit Nobel, DE-OS 2 519 529, 1975. The Rummus Co.; Dynamit Nobel, DE-OS 2 729 072, 1977.
40. BASF, DE 3 104 765, 1981.
41. Daicel Chemical, JP-Kokai 81 127 357, 1980; *Chem. Abstr.* **96** (1982) 35109b.
42. Stamicarbon N.V., NL 7 017 718, 1970.
43. Agency of Industrial Sciences and Technology, Japan Steel Works Ltd., JP-Kokai 89 70 464, 1987; *Chem. Abstr.* **111** (1989) 77861z.
44. F. Mensch, *Erdoel Kohle Erdgas Petrochem.* **22** (1969) no. 2, 67 – 71.
45. A. Nenz, M. Pieroni, *Hydrocarbon Process.* **47** (1968) no. 11, 139 – 144. R. L. Frank, F. J. Pilgrim, E. F. Riener, *Org. Synth.,* **coll. vol. 4** (1963) 451. T. Shoji, N. Abe, *Petrotech (Tokyo)* **7** (1984) 195; *Chem. Abstr.* **101** (1984) 56791e.
46. H. Boenemann, *Angew. Chem.* **90** (1978) 517 – 526; *Angew. Chem. Int. Ed. Engl.* **17** (1978) 505.
47. Lonza, US 4 212 978, 1978.
48. Studienges. Kohle mbH, DE 2 840 460, 1978.
49. Y. Kusunoki, H. Okazaki, *Hydrocarbon Process.* **53** (1974) no. 11, 129 – 131.
50. Hoechst, DE-OS 3 244 032, 1984.
51. Du Pont, US 655 087, 1984.
52. U. D. Nazirova et al., *Dokl. Akad. Nauk UzSSR* 1988, no. 2, 40 – 42; *Chem. Abstr.* **109** (1988) 230746w.
53. Mobil Oil Corp., EP-A 82 613, 1983.
54. *Reagent Chemicals,* American Chemical Society, Washington, D.C., 1977, p. 724.
55. H. E. Reich et al., *J. Am. Chem. Soc.* **77** (1955) 5434.
56. L. M. Jampolsky et al., *J. Am. Chem. Soc.* **74** (1952) 5222.
57. Koei Chemical, JP-Kokai 60 178 863, 1984; *Chem. Abstr.* **104** (1986) 68760u.
58. R. L. Frank et al., *J. Am. Chem. Soc.* **71** (1949) 2804.
59. Eli Lilly, US 3 180 872, 1963.
60. Allied Chemical & Dye, US 2 556 845, 1947.
61. Sumitomo Chemical, JP-Kokai 75 73 898, 1973; *Chem. Abstr.* **83** (1975) 209277d.
62. Koei Chemical, JP-Kokai 59 95 228, 1982; *Chem. Abstr.* **101** (1984) 193022u.
63. A. M. J. Broome et al., *Nature (London)* **189** (1961) 59.
64. Schering, DE-OS 2 942 680, 1979.
65. Agfa-Gevaert, DE-OS 2 439 551, 1974.
66. W. H. F. Sasse, *Org. Synth.* **46** (1966) 5 – 10.
67. ICI, GB 960 176, 1962 (G. H. Lang et al.).
68. ICI, DE 1 445 079, 1960.
69. M. Cieslak, M. Kopepska, *Pr. Inst. Przem. Org.* **2** (1970) 21.
70. ICI, DE-OS 2 230 560, 1971.
71. ICI, GB 1 202 711, 1967.
72. T. Anderson, *Justus Liebigs Ann. Chem.* **154** (1870) 270.
73. ICI, NL-A 6 603 415, 1965.
74. ICI, NL-A 6 512 461, 1964.
75. ICI, FR 1 380 806, 1962.
76. ICI, BE 639 705, 1962.
77. ICI, GB 957 098, 1961.

78. R. E. Lyle et al., *J. Am. Chem. Soc.* **77** (1955) 1291.
79. C. H. G. Hands et al., *J. Soc. Chem. Ind.* **66** (1947) 407.
80. A. H. Blatt, *Org. Synth.,* coll. vol. 2 (1943) 419.
81. R. A. Abramovitch et al., *J. Chem. Soc. B* 1971, 131.
82. B. C. McKusick, *Org. Synth.* **43** (1963) 97.
83. Nippon Soda Co., JP-Kokai 80 59 167, 1978; *Chem. Abstr.* **93** (1980) 168139w.
84. J. P. Wibaut et al., *Recl. Trav. Chim. Pays-Bas* **73** (1954) 140.
85. K. Bowden, P. N. Green, *J. Chem. Soc.* 1954, 1795.
86. R. R. Renshaw et al., *J. Am. Chem. Soc.* **59** (1937) 297.
87. D. Jerchel, H. Fisher, K. Thomas, *Chem. Ber.* **89** (1956) 2921.
88. R. A. Carpio et al., *J. Electrochem. Soc.* **126** (1979) 1644.
89. US Dept. of the Air Force, US 4 122 245, 1977.
90. Nisshin Steel, JP-Kokai 62 70 593, 1985; *Chem. Abstr.* **107** (1987) 224963g.
91. J. A. Boon et al., *J. Org. Chem.* **51** (1986) 480.
92. Texaco, US 4 554 383, 1984.
93. E. Ochiai et al., *J. Pharm. Soc. Japan* **67** (1947) 79; *Chem. Abstr.* **45** (1951) 9538a.
94. A. R. Katritzky et al., *J. Chem. Soc.* 1957, 1769.
95. M. Katada, *Yakugaku Zasshi* **67** (1947) 51; *Chem. Abstr.* **45** (1945) 9536d. J. H. Markgraf, *J. Am. Chem. Soc.* **85** (1963) 958.
96. H. Vorbrugger et al., *Synthesis* 1983, 316.
97. B. Bobranski et al., *Chem. Ber.* **71** (1938) 2385.
98. Bayer, DE-OS 3 800 179, 1988.
99. Koei Chemical, JP-Kokai 81 169 672, 1980; *Chem. Abstr.* **96** (1982) 162549v.
100. Lonza, DE-OS 2 713 316, 1976.
101. V. Boekelheide et al., *J. Am. Chem. Soc.* **76** (1954) 1286.
102. E. Ochiai, *J. Org. Chem.* **18** (1953) 534.
103. E. G. Novikov et al., *Khim. Prod. Koksovaniya Uglei Vostoka SSSR* (1970) 170; *Chem. Abstr.* **76** (1972) 153512n.
104. H. Adkins et al., *J. Am. Chem. Soc.* **56** (1934) 2425.
105. M. Freifelder et al., *J. Org. Chem.* **26** (1961) 3805.
106. M. Freifelder et al., *J. Org. Chem.* **27** (1962) 287.
107. O. Hibino et al., *Kogyo Kagaku Zasshi* **68** (1965) 1703; *Chem. Abstr.* **64** (1966) 3472h.
108. H. K. Hall, Jr., *J. Am. Chem. Soc.* **80** (1958) 6413.
109. M. Freifelder et al., *J. Org. Chem.* **30** (1965) 1319.
110. A. Michaelis et al., *Justus Liebigs Ann. Chem.* **331** (1904) 245.
111. Dow, BE 628 487, 1963.
112. G. D. Searle & Co., US 3 225 054, 1965.
113. G. M. Badger et al., *J. Chem. Soc.* 1956, 616.
114. Haessel A.G., EP 103 553, 1982.
115. Takeda Chemical Ind., EP 174 726, 1984.
116. E. Klingsberg: *Pyridine and Its Derivatives,* **vol. 2,** Interscience Publishers, New York 1961, p. 299.
117. J. J. Eisch, *Adv. Heterocycl. Chem.* **7** (1966) 1.
118. Degussa, US 3 920 657, 1975.
119. W. Mathe et al., *Angew. Chem. Int. Ed. Engl.* **2** (1963) 144.
120. Dow, US 3 420 833, 1969.
121. Dow, US 3 424 754, 1966.
122. Dow, US 3 732 230, 1973.
123. ICI, EP 65 358, 1981.
124. Dow, US 4 256 894, 1978.
125. Dow, US 3 694 332, 1972.
126. Dow, US 3 993 654, 1976.
127. Ishihara Sangyo, JP-Kokai 55 85 564, 1980; *Chem. Abstr.* **94** (1981) 47146h.
128. E. Koenig et al., *Chem. Ber.* **54** (1921) 1360.
129. M. Conrad et al., *Chem. Ber.* **20** (1896) 1679.
130. J. A. Leben, *Chem. Ber.* **29** (1896) 1679.
131. T. Itai et al., *Yukagaku Zasshi* **75** (1955) 296; *Chem. Abstr.* **50** (1956) 1810g.
132. C. F. H. Allen et al., *Org. Synth.,* **coll. vol. 3** (1955) 136.
133. C. Rath, *Justus Liebigs Ann. Chem.* **486** (1931) 95.
134. O. Seide, *Chem. Zentralbl.* **3** (1923) 1022.
135. H. J. Hertog, *Recl. Trav. Chim. Pays-Bas* **69** (1950) 673.
136. Koei Chemical Co., JP-Kokai 86 251 663, 1985; *Chem. Abstr.* **106** (1987) 176178k.
137. Lummus Co., US 40 082 441, 1975.
138. Roche Product Ltd., GB 717 172, 1954.
139. R. L. Frank et al., *J. Am. Chem. Soc.* **70** (1948) 3482.
140. B. Lipka et al., PL 50 077, 1965; *Chem. Abstr.* **65** (1966) 5446d.
141. Koei Chemical Co., JP-Kokai 89 175 968, 1987; *Chem. Abstr.* **112** (1990) 118662b.
142. Lonza, DE-OS 2 435 344, 1974.
143. Degussa, DE-OS 3 107 755, 1982.
144. Yuki Gosei Kogyo, JP 70 13 572, 1965; *Chem. Abstr.* **73** (1970) 77059r.
145. Takeda Chemical Ind., EP 290 996, 1988.
146. Koei Chemical Co., US 4 778 890, 1987.

147. Nippon Shokubai Kagaku Kogyo Co., EP 290 996, 1988.
148. T. P. Sycheva et al., *Khim. Farm. Zh.* **6** (1972) 6 – 8.
149. H. Spitzer et al., *Chem. Ber.* **59B** (1926) 1477 – 1486.
150. S. M. McElvan, R. Adams, *J. Am. Chem. Soc.* **458** (1922) 2744.
151. Degussa, DE 2 517 054, 1975; *Chem. Abstr.* **86** (1977) 43576k. J. E. Paustian et al., *CHEMTECH* 1981, 174 – 179.
152. C. B. Rosas, G. B. Smith, *Chem. Eng. Sci.* **35** (1980) 330 – 337; *Chem. Abstr.* **93** (1980) 79696.
153. R. A. Bergamin, *Chimia* **44** (1990) 255 – 257.
154. C. F. H. Allen, C. N. Wolf, *Org. Synth.,* **coll. vol. 4** (1963) 45.
155. Degussa, DE-OS 2 517 054, 1942.
156. H. Friedrich: *Kirk-Othmer,* 3rd ed., **vol. 24,** Wiley-Interscience, New York 1983, pp. 59 – 93.
157. Takeda Chemical Ind., DE-OS 2 619 243, 1976.
158. W. Deady, M. R. Grimett, C. H. Potts, *Tetrahedron* **35** (1979) 2895.
159. E. Klingsberg, R. H. Mizzoni: *Pyridine and Its Derivatives,* **vol. 2,** Interscience Publishers, New York 1961, p. 469.
160. R. A. Abramovitch, R. H. Mizzoni: *Pyridine and Its Derivatives,* **vol. 3,** John Wiley & Sons, New York 1974, p. 1.
161. M. T. Leffler, *Org. React. (N.Y.)* **1** (1942) 91.
162. A. F. Pozharskii, A. M. Simonov, V. H. Doron'kin, *Russ. Chem. Rev. (Engl. Transl.)* **47** (1978) 1042.
163. E. Klingsberg, A. S. Tomcufcik, L. N. Starker: *Pyridine and Its Derivatives,* **vol. 3,** Interscience Publishers, New York 1962, p. 1.
164. R. A. Abramovich, C. S. Giam: *Pyridine and Its Derivatives,* **vol. 3,** John Wiley & Sons, New York 1974, p. 41.
165. Sterling Drug Inc., US 3 149 104, 1961.
166. Pyridium Corporation, US 1 680 108, 1928. Pyridium Corporation, US 1 680 109, 1928. Pyridium Corporation, US 1 608 111, 1928.
167. R. F. Evans, H. C. Brown, H. C. van der Plas, *Org. Synth.* **43** (1963) 97.
168. Nippon Soda Co., JP-Kokai 82 72 961, 1980. Nippon Soda Co., JP-Kokai 82 108 068, 1980. Nippon Soda Co., JP-Kokai 82 108 069, 1980. Nippon Soda Co., JP-Kokai 82 102 864, 1980.
169. V. G. Hofle, W. Steglich, H. Vorbruggen, *Angew. Chem.* **90** (1978) 602.
170. Schering AG, DE 2 517 774, 1975.
171. Reilly Tar, US 4 158 093, 1977.
172. R. C. Elderfield: *Heterocyclic Compounds,* **vol. 1,** Wiley-Interscience, New York 1950, p. 435.
173. C. A. Salemink et al., *Recl. Trav. Chim. Pays-Bas* **68** (1949) 1013.
174. O. v. Schickh et al., *Chem. Ber.* **69** (1936) 2593.
175. E. Koenigs et al., *Chem. Ber.* **57** (1924) 1187.
176. P. Nantka-Namirski et al., *Acta Pol. Pharm.* **24** (1967) 228; *Chem. Abstr.* **69** (1968) 2827.
177. L. Ruzicka et al., *Helv. Chim. Acta* **3** (1920) 806.
178. E. Koenigs et al., *Chem. Ber.* **61** (1928) 1022.
179. F. Freidel, *Chem. Ber.* **45** (1912) 428.
180. M. Brash et al., *J. Chem. Soc.* 1959, 3530.
181. W. Herz et al., *J. Org. Chem.* **26** (1961) 122.
182. D. E. Parker, W. Shive, *J. Am. Chem. Soc.* **69** (1947) 62.
183. S. Oae et al., *Tetrahedron* **21** (1965) 1971.
184. B. M. Bain et al., *Chem. Ind. (London)* 1960, 402.
185. B. R. Baker et al., *J. Org. Chem.* **20** (1955) 116.
186. H. Ledetschke, *Chem. Ber.* **85** (1952) 202.
187. H. Ledetschke, *Chem. Ber.* **86** (1953) 123.
188. Koei Chemical Co., JP 60 132 959, 1983; *Chem. Abstr.* **103** (1985) 215195y.
189. P. W. Ford et al., *Aust. J. Chem.* **18** (1968) 867.
190. W. Koenigs et al., *Chem. Ber.* **35** (1902) 1343.
191. K. Loffer et al., *Chem. Ber.* **40** (1907) 1325.
192. R. L. Frank et al., *J. Am. Chem. Soc.* **68** (1946) 1368.
193. K. Winterfeld et al., *Justus Liebigs Ann. Chem.* **573** (1951) 85.
194. W. Mathes et al., *Chem. Ztg.* **80** (1956) 475.
195. W. Mathes et al., *Chem. Ber.* **84** (1951) 452.
196. C. Harries et al., *Justus Liebigs Ann. Chem.* **410** (1915) 95.
197. S. Gisburg et al., *J. Am. Chem. Soc.* **79** (1957) 481.
198. Ciba Geigy, DE-OS 2 262 780, 1972.
199. E. P. Papodopoulous et al., *J. Org. Chem.* **31** (1966) 615.
200. W. Bosche et al., *Chem. Ber.* **73** (1940) 839.
201. Raschig AG, BE 710 192, 1951.
202. Nepera, US 3 274 206, 1964.
203. E. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, Amsterdam 1965, pp. 304 – 309.
204. Occupational Safety & Health Administration, Material Safety Data Sheets from the Occupational Health Services Database, OSHA, Washington, D.C., 1985.
205. A. Jori et al., *Ecotoxicol. Environ. Saf.* **7** (1983) no. 3, 251 – 275.

206. *Dangerous Prop. Ind. Mater. Rep.* **7** (1987) no. 4, 101 – 104.
207. E. W. Schafer Jr., P. B. Brunton, D. J. Cunningham, *Toxicol. Appl. Pharmacol.* **26** (1973) 532.
208. D. A. Spyker, C. Lynch, J. Shabanowitz, J. A. Sinn, *Clin. Toxicol.* **16** (1980) no. 4, 487 – 497.
209. *Dangerous Prop. Ind. Mater. Rep.* **5** (1985) no. 5, 39 – 41.
210. P. J. Gehring, T. R. Torkelson, F. Oyen, *Toxicol. Appl. Pharmacol.* **11** (1967) 361.
211. C. D. Balaban, *Brain Res. Rev.* **9** (1985) no. 1, 21 –42.
212. J. G. Black, D. Howes, *Toxicol. Annu.* **3** (1979) 1 – 26.
213. NIOSH, Registry of Toxic Effect of Chemical Substances (RTECS), Dialog Information Service Inc., File 336, July 1990.
214. J. H. Baxter, M. F. Mason, *J. Pharmacol. Exp. Ther.* **91** (1947) 350.
215. Comm. Eur. Communities, (Rep.) no. Eur 11553, Solvents in Common Use 1988, 215 – 234.
216. L. J. Pollock, I. Finkelman, A. J. Arieff, *Arch. Intern. Med.* **71** (1943) 95 – 106.
217. J. W. Langston et al., *Science (Washington, D.C.)* **219** (1983) 979.
218. T. P. Singer, R. R. Ramsay, K. Mckeown, A. Trevor, *Toxicology* **49** (1988) no. 1, 17 – 23. H. Imai, *Brain Nerve* **40** (1988) no. 11, 1011 – 1024.
219. *Carcinog. Tech. Rep. Ser. U.S. Nat. Cancer Inst.,* NCI-TR-95, 78.
220. K. Takahashi, G. F. Huang, M. Araki, Y. Kawazoe, *Gann* **70** (1979) 799 – 806.
221. O. Tada *Kogai Bunseki Shishin* **3** (1972) 49 – 62.

Pyridoxine → Vitamins

Pyrimidine and Pyrimidine Derivatives

ALAN P. CHORLTON, ICI Specialties, Blackley, Manchester, United Kingdom

1. **Introduction** 1
2. **Chemical Properties** 1
3. **Production** 2
4. **Important Naturally Occurring Derivatives of Pyrimidine** 3
5. **Industrially Important Pyrimidines** . 4
5.1. **Pharmaceuticals** 4
5.2. **Agrochemicals** 4
5.3. **Dyes** 5
6. **References** 5

1. Introduction

Pyrimidine was first isolated by GABRIEL and COLMAN in 1899 [1–4].

Pyrimidine [*289-95-2*], $C_4H_4N_2$, M_r 80.09, *mp* 20 – 22 °C, *bp* 123 – 124 °C, UV max (water): 240 nm (ε 2400) is a hygroscopic, water-soluble solid with a penetrating odor. It is also a weak base ($pK_a = 1.23$) that forms a number of solid derivatives (hydrochloride: *mp* 162 °C; picrate: yellow needles, *mp* 156 °C; methiodide: yellow plates, *mp* 136 – 137 °C). There are many synthetic approaches to the preparation of pyrimidine [4]. The most effective way to prepare pyrimidine in large quantity is to react 1,1,3,3-tetraethoxypropane with formamide over alumina at 200 °C [5].

2. Chemical Properties

The reactivity of the pyrimidine ring is approximately equivalent to that of 1,3-dinitrobenzene or 3-nitropyridine. This is because a double-bonded ring nitrogen is comparable in its electron-withdrawing ability to a nitro group.

Electrophilic substitution at the ring positions occurs only in the presence of one or more activating substituents such as amino or hydroxyl groups. Attack invariably proceeds at a vacant 5-position. Pyrimidines with a single activating substituent are subject only to halogenation; 2-and 4-aminopyrimidines all give 5-bromo derivatives.

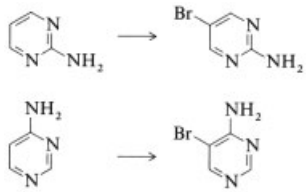

Pyrimidines bearing two activating substituents (e.g., pyrimidine-4,6-dione) readily undergo electrophilic substitution at the 5-position giving a series of compounds with the general formula (**1**).

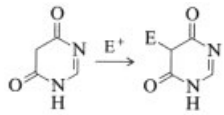

1

E = NO, NO_2, Ph –N=N, halogen

Nucleophilic attack occurs readily at positions 2, 4, and 6. In general, the 4-position is the most reactive. For example, the reaction of phenyl magnesium bromide or phenyl lithium with pyrimidine leads to 6-phenyl-1,6-dihydropyrimidine, which on oxidation gives 4-phenylpyrimidine.

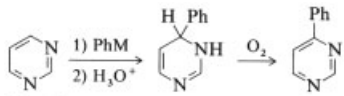

M = Li, MgBr

The reactivity of pyrimidine and its derivatives has been comprehensively reviewed [4].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_431

3. Production

The most general types of ring closure applicable to the preparation of pyrimidines are illustrated below. Other approaches are possible but are rarely utilized.

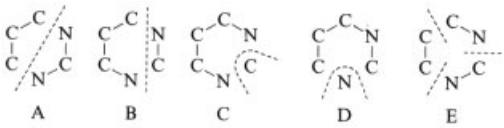

Type A syntheses are usually the most successful, and over 80 % of all known pyrimidines have been prepared in this way. A typical example is the condensation of urea with diethyl malonate to give barbituric acid. For information regarding other barbiturates see →Antiepileptics, Chap. 2.1.; →Hypnotics, Chap. 5.1.

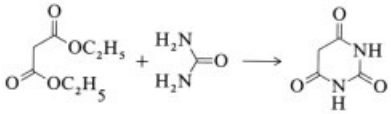

The synthetic utility of this reaction is greatly enhanced by the fact that any compound containing the group $H_2N{-}C{-}NH_2$ can take the place of urea. Thus, guanidines are precursors to 2-aminopyrimidines, amidines give 2-alkyl or aryl-substituted pyrimidines, and thioureas lead to pyrimidine-2-thiones. The ester group in diethyl malonate can also be replaced by other groups (e.g., keto, aldehyde, nitrile, acid chloride and imino moieties) opening the way to a wide variety of pyrimidine derivatives.

The most common example of a type B synthesis is the reaction of an isocyanate with a 3-aminoacrylate derivative to yield a uracil.

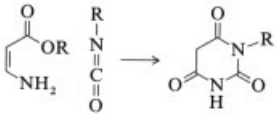

2-Alkyl pyrimidines are readily accessible by type C syntheses. Condensation of 1,3-diaminopropane with the appropriate carboxylic acid, followed by cyclization, results in the dihydropyrimidine **2**, which on dehydrogenation with palladium/alumina affords the 2-substituted pyrimidine **3** [6].

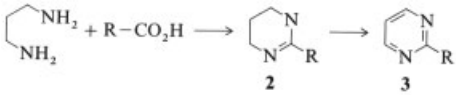

In a similar sequence diamines with a higher oxidation level lead to pyrimidines directly. Thus, malonodiamide condenses with ethyl formate to give 4,6-dihydroxypyridimines [7], and with diethyl carbonate to give 4,6-diaminopyrimidine-2-ones [8].

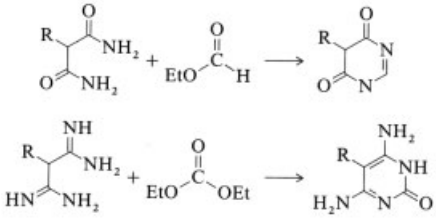

The Shaw synthesis is an example of a type D synthesis, and is used to prepare many uracil and thiouracil derivatives [9].

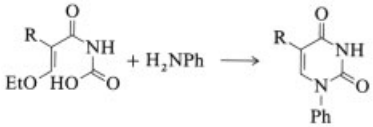

Base-catalyzed trimerization of acetonitrile is an example of a type E synthesis, and constitutes a very efficient way of preparing certain pyrimidines (e.g., 2,6-dimethyl-4-aminopyrimidine from acetonitrile) [10].

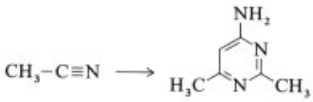

Other approaches to pyrimidines have been reviewed elsewhere [4].

4. Important Naturally Occurring Derivatives of Pyrimidine

Ortic acid [*65-86-1*], $C_5H_4N_2O_4$, M_r 156.10, *mp* 345 – 346 °C, is thought to be the key precursor in the biosynthesis of all naturally occurring pyrimidines. It can be synthesized directly from urea and oxalacetic acid ester [11]. It has been used in combination with 4-amino-5-imidazole carboxamide for the treatment of liver disease [12].

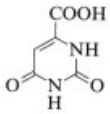

Uracil [*66-22-8*], $C_4H_4N_2O_2$, M_r 112.09, *mp* 335 °C, is widely distributed in nature, since it is a constituent of nucleic acids. Synthesis: reaction of malic acid and urea in concentrated sulfuric acid [13]. Uracil is an excellent starting material for the preparation of pyrimidines, pteridines, and purines.

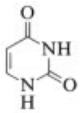

Thymine [*65-71-4*], $C_5H_6N_2O_2$, M_r 126.11, *mp* 335 – 340 °C, unlike uracil, is not found in ribonucleic acid but in deoxyribonucleic acid. Synthesis: treatment of 3-methylmalic acid with fuming sulfuric acid to yield 2-formylpropionic acid, which in turn condenses with urea to give thymine [14].

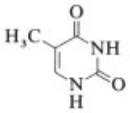

Alloxan [*50-71-5*], $C_4H_2N_2O_4$, M_r 142.07, *mp* 256 °C (decomp.), turns pink at 230 °C and forms mono- and tetrahydrates which lose water at 150 °C. Alloxan induces permanent diabetes in many animals, but not in humans. The best synthetic route to alloxan is direct oxidation of barbituric acid with chromium trioxide [15].

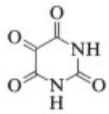

Cytosine [*71-30-7*], $C_4H_5N_3O$, M_r 111.10, *mp* 320 – 325 °C (decomp.), is widely distributed in nature. Since it is a constituent of nucleic acids it can be isolated by hydrolyzing thymus nucleic acids. The principal synthesis is from 3,3-dihydroxypropionitrile and urea [16].

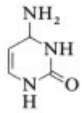

Thiamine [*59-43-8*], vitamin B_1 (→Vitamins), is a very important pyrimidine, manufactured at a level of about 2000 t/a. Synthetic thiamine is prepared in a multistage process. The starting material is acetamidine, which is condensed with the formyl derivative of ethyl 2-ethoxypropionate to give 5-ethoxymethyl-4-hydroxy-2-methylpyrimidine. The heterocyclic thiazole component is introduced in the form of 5-(2′-hydroxy)-4-methylthiazole [17].

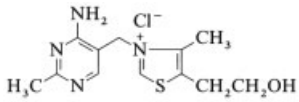

Numerous bicyclic heterocycles contain the pyrimidine nucleus, including the purines adenine and guanine, which are important basic constituents of nucleic acids (→Nucleic Acids, →Purine Derivatives).

5. Industrially Important Pyrimidines [18, 19]

5.1. Pharmaceuticals

Pyrimidine derivatives constitute a major class of important pharmaceuticals and have been comprehensively covered elsewhere:

→Antibiotics (e.g., polyoxines); →Antidiabetic Drugs (e.g., sulfonamido pyrimidines); →Antihypertensives, Chap. 5.2.3. (e.g., minoxidil); →Cancer Chemotherapy, Chap. 2.2. (e.g., fluoropyrimidines); →Chemotherapeutics, Chap. 2.7.3.7. (e.g., sulfamethoxazole, trimethoprim, sulfisoxazole, sulfadiazine, sulfacytine), →Chemotherapeutics, Chap. 2.7.3.8. (e.g., sulfadoxine), →Chemotherapeutics (e.g., pyrimethamine), →Chemotherapeutics, Chap. 4.6.1. (e.g., pyrimidine nucleosides and nucleotides). The articles →Anesthetics, General, Chap. 3.1., →Antiepileptics, Chap. 2.1., and →Hypnotics, Chap. 5.1., all deal with barbiturates. Important pyrimidines that do not fall within these categories are summarized below.

Thonzylamine hydrochloride [*63-56-9*], $C_{16}H_{23}ClN_4O$, M_r 322.83, *mp* 173 – 176 °C, is used as an antihistaminic drug. It is prepared by treating the sodium salt of 2-(4-methoxybenzyl)aminopyrimidine with *N,N*-dimethyl-2-chloroethylamine [20].

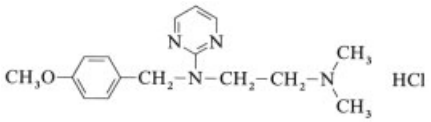

Piribedil [*3605-01-4*], $C_{16}H_{18}N_4O_2$, M_r 298.35, *mp* 98 °C, is a central dopaminergic agonist used as a peripheral vasodilator. It is prepared by treating 1-(3′,4′-methylenedioxybenzyl) piperazine with 2-chloropyrimidine in the presence of a base [21].

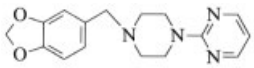

Citicoline [*987-78-0*], $C_{14}H_{26}N_4O_{11}P_2$, M_r 488.33, is a coenzyme in one of the biosynthetic pathways leading to lecithin. It can be isolated from liver and yeast. Citicoline is of clinical use as a central circulation stimulant. It is produced industrially by biochemical action of *Brevibacterium ammoniagenes* on cytidine 5′-monophosphate and choline [22].

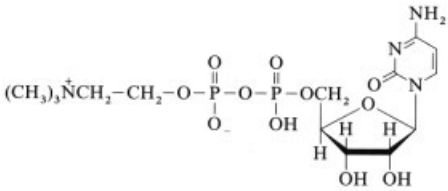

Cytarabin [*69-74-9*], $C_9H_{13}N_3O_5$, M_r 243.22, *mp* 212 – 215 °C, is used for the treatment of leukemia and is also active against herpes. Cytarabin is synthesized by treatment of 2,4-dimethoxypyrimidine with 2,3,5-tri-2-benzyl-D-arabinofuranosyl chloride to give the 4-methoxypyrimidone (**4**: R = CH_2Ph, X = OCH_3Me), which undergoes aminolysis with ammonia to the corresponding amine (**4**: R = CH_2Ph, X = NH_2). Subsequent hydrogenolysis leads to cytarabin (**4**: R = H, X = NH_2) [23].

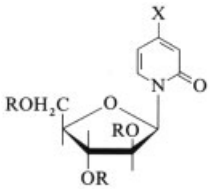

4

5.2. Agrochemicals

Certain pyrimidines are also important agrochemicals: →Fungicides, Chap. 8.7. (e.g., ethirimol and bupirimate); →Insect Control, Chap. 6.6. (e.g., diazinon, pirimiphos), →Insect Control, Chap. 7.1. (e.g., pirimicarb); →Nematocides, Chap. 4.3. (e.g., diazinon).

Pyrimidines are also being used in the newer generation of agrochemicals. Examples follow.

Bensulfuron-methyl [*83055-99-6*], $C_{16}H_{18}N_4O_7S$, M_r 410.4, *mp* 185 – 188 °C, is an example of a sulfonyl urea, a new class of selective systemic herbicides that inhibit biosynthesis of the essential amino acids valine and isoleucine, thereby interrupting cell division and plant growth. Bensulfuron-methyl is used

in the control of annual and perennial weeds in rice [24].

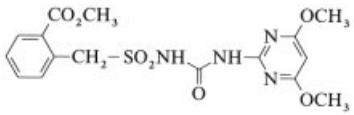

Nuarimol [*63284-71-9*], $C_{17}H_{12}ClFN_2O$, M_r 314.7, *mp* 126 – 127 °C, is a systemic fungicide that acts by inhibiting ergosterol biosynthesis. It is used to control a wide range of pathogenic fungi, such as powdery mildews on fruit [25].

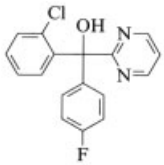

5.3. Dyes

Pyrimidines are also used in the dye industry [26]. The most important examples are the isoindoline pigments, *Pigment yellow 139* [*36888-99-0*], $C_{16}H_9N_5O_6$, M_r 367.27, is used for the coloration of paints and plastics. It is synthesized from 1,3-diiminoisoindoline and barbituric acid [27].

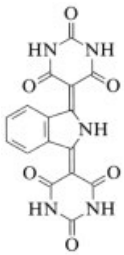

Pyrimidines of lesser industrial importance are discussed in [18, 19, 26].

6. References

1. G. W. Kenner, A. Todd: "Pyrimidine and its Derivative," in R. C. Elderfield (ed.): *Heterocyclic Compounds,* vol. 6, Wiley-Interscience, New York 1957, pp. 234 – 324.
2. D. J. Brown: "The Pyrimidines," in A. Weissberger (ed.): *The Chemistry of Heterocyclic Compounds,* vol. 16, Wiley-Interscience, New York 1962.
3. D. J. Brown: "The Pyrimidines," Wiley-Interscience, New York 1970, [2] Suppl. I .
4. D. J. Brown: "Pyrimidines and Their Benzo Derivatives," in A. J. Boulton, A. McKillop (eds.): *Comprehensive Heterocyclic Chemistry,* vol. 3, Pergamon Press, New York 1984, pp. 57 – 155.
5. H. Bredereck, R. Gompper, H. Herlinger, *Chem. Ber.* **91** (1958) 2832.
6. G. E. Hibert, *J. Am. Chem. Soc.* **54** (1932) 2076 – 2083.
7. Z. Budensinsky, F. Roubinek, E. Svatek, *Collect. Czech. Chem. Commun.* 1965, 3730.
8. G. A. Howard, B. Lythgoe, A. R. Todd, *J. Chem. Soc.* 1944, 476 – 477.
9. M. R. Atkinson, G. Shaw, R. N. Warrener, *J. Chem. Soc.* 1956, 4118 – 4123.
10. E. C. Horning (ed.): *Org. Syn. Coll.,* **vol. III,** Wiley-Interscience, New York 1955, p. 71.
11. R. Muller, *J. Prakt. Chem.* **56** (1897) 488 – 495.
12. R. Haroka, T. Kamiya, US 3 271 398, 1966.
13. O. Baudisch, D. Davidson, *J. Am. Chem. Soc.* **28** (1926) 2379 – 2383.
14. H. W. Scherp, *J. Am. Chem. Soc.* **68** (1946) 912 – 913.
15. E. C Horning (ed.): *Org. Syn. Coll.,* **vol. III,** Wiley-Interscience, New York 1955, p. 39.
16. J. A. Hill, W. J. Le Quesne, *J. Chem. Soc.* 1965, 1515 –1516.
17. H. G. Frank: *Industrial Aromatic Chemistry,* Springer Verlag, Berlin 1988, p. 411.
18. M. Sittig: *Pharmaceutical Manufacturing Encyclopedia,* vols. 1 and 2, Noyes Data Corp., Park Ridge, NJ, 1988.
19. C. R. Worthing (ed.): *The Pesticide Manual,* 9th ed., British Crop. Protection Council, Farnham 1991.
20. H. L. Friedman, A. V. Tolstouhov, US 2 465 865, 1949.
21. G. Negnier, R. Canevari, M. Laubic, US 3 299 067, 1967.
22. Takeda Chemical, US 3 687 932, 1972 (J. Nakamachi).
23. Upjohn, US 3 116 282, 1963 (J. Hunter).

24. Du Pont, US 4 420 325, 1984 (R. Saurs).
25. Eli Lilley, GB 1 218 623, 1967 (J. Davenport).
26. *The Color Index,* 3rd ed., The Society of Dyers and Colorists, Bradford 1982.
27. Bayer, DE, 3 022 839, 1980 (M. Lorenz).

Pyrocatechol → Phenol Derivatives
Pyrogallol → Phenol Derivatives
Pyromellitic Acid → Carboxylic Acids, Aromatic
Pyrophosphates → Phosphoric Acid and Phosphates
Pyrophyllite → Talc

Pyrotechnics

Explosives and Propellants are separate keywords.

Uwe Krone, Nico-Pyrotechnik, Hanns-Jürgen Diederichs GmbH & Co. KG, Trittau, Federal Republic of Germany (Chaps. 1–3, 5–7)

Ronald Lancaster, Kimbolton Fireworks Ltd., Kimbolton, Huntingdon, United Kingdom (Chap. 4)

1. Introduction 1
2. Pyrotechnic Compositions 2
3. Pyrotechnic Effects 3
4. Fireworks 4
4.1. Shop Fireworks 5
4.1.1. Toy Fireworks 5
4.1.2. Garden Fireworks 6
4.2. Display Fireworks 8
5. Pyrotechnics for Industrial and Military Use 12
6. Production 14
7. Legal Aspects; Storage and Transportation 15
8. References 17

1. Introduction

The word "pyrotechnics" is derived from a Greek term *pyros techneion* roughly meaning "the art of working with fire." A comprehensive definition was devised in the 1970s by a German working group [1]: "Pyrotechnics is the science of the manufacture and use of articles containing single substances or mixtures of substances whose exothermic chemical reaction is designed to produce desired acoustic, incendiary, thermal, illuminative or luminous, fog, smoke, irritant, pesticidal, propulsive, and/or igniting effects." Ellern has defined pyrotechnics as the art and science of creating and utilizing the heat effects and products from exothermally reacting, predominantly solid mixtures or compounds when the reaction is, with some exceptions, nonexplosive and relatively slow, self-sustaining, and self-contained [2].

The use of mixtures of energy-rich chemicals for the production of heat, light, smoke, or sound is a tradition extending over millennia. Sources assert that the first applications of pyrotechnic mixtures were associated with natural deposits of saltpeter (niter) in China and India [3]. Mixtures of saltpeter and sulfur with wood tar or natural asphalt were probably constituents of the first military pyrotechnics, including the "Greek fire" mentioned in the 7th century. Powder recipes reminiscent of types of black powder still in use today were described in Greece during the 8th century and in China around A.D. 1000–1200 [4]. Fireworks designed for entertainment purposes first appeared in Europe and Japan at the beginning of the Middle Ages, long after the destructive and propulsive power of black powder (→ Explosives, Chap. 1.) had already been invoked for numerous explosive applications in mining and quarrying, and in propellant charges for cannons, muskets, and rockets. Fireworks as a source of amusement also had their origin in China. Historical sketches of the introduction, manufacture, and use of fireworks in Europe (especially in Italy, Germany, England, and France), are provided in [5–7].

Numerous national and international pyrotechnic industry associations, trade associations (Hong Kong), and scientific societies are dedicated to promoting the national and international transfer of technical and scientific pyrotechnic information. The principal forum for this exchange are regular international meetings, seminars, and special publications (e.g., the journal *Propellants, Explosives, and Pyrotechnics*). Worthy of particular mention are the international seminars of the International Pyrotechnics Society, IPS (USA); the American Defense Preparedness Association, ADPA; the Fraunhofer Institut für Chemische Technologie, ICT (Germany); the Groupe de Travail de Pyrotechnic Spatiale, GTPS (France); the Royal Armament Research and Development Establishment, RARDE (UK); the Royal Swedish Academy of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_437

Sciences, Section for Detonation and Combustion (Sweden); and the China Ordnance Society (Peoples' Republic of China). Colloquia dealing with pyrotechnic topics are sponsored by national institutes and societies, e.g., the German Society for Defense Technology (DWT) and the Franklin Research Center (Maryland, USA). Conference proceedings issued by these societies, along with the patent literature, remain among the most vital and current sources of information. Surveys of the entire subject of pyrotechnics, including both civilian and military – industrial applications, are given in [2, 3, 6, 8, 9].

2. Pyrotechnic Compositions

The energy sources for pyrotechnic charges, pyrotechnic munitions, and fireworks are reactive chemical mixtures known as *pyrotechnic compositions*, whose principal components are oxidizing agents, reducing agents, and binders. Formulation aids as anticaking or friction-reducing substances (lubricants) are also included, and in some cases, catalysts.

Oxidizing agents: nitrates [KNO_3, $NaNO_3$, $Ba(NO_3)_2$, $Sr(NO_3)_2$, $NH{=}C(NH_2)_2 \cdot HNO_3$]; perchlorates ($KClO_4$, NH_4ClO_4); chlorates [$KClO_3$, $Ba(ClO_3)_2 \cdot H_2O$]; chromates [$BaCrO_4$, $K_2Cr_2O_7$, $(NH_4)_2Cr_2O_7$, $PbCrO_4$]; peroxides and oxides (BaO_2, SrO_2, PbO_2, Pb_3O_4, MnO_2, ZnO, Fe_2O_3, Cu_2O, CuO); halogen-containing substances [C_2Cl_6, C_6Cl_6, $(C_{2n}F_{4n})_x$].

Reducing agents : metals (Al, Al/Mg, Ca/Si, Cr, Fe, Mg, Sb, W, Zn, Zr/Ni); nonmetals [B, C (graphite, charcoal, activated carbon), P, S, Si]; carbohydrates (sucrose, lactose); inorganic compounds (NaN_3, Sb_2S_3, ZrH_x); organic compounds (sodium salicylate, potassium benzoate, sorbitol, metaldehyde, hexamethylenetetramine, oxalates, tetrazoles, azodicarbonamide, dicyandiamide).

Binders : natural resins (gum acaroid, gum arabic, gum tragacanth, shellac); synthetic resins [polyethylene, poly-(vinyl chloride), chlorinated rubber, polyester, polyurethane]; plasticizers; dextrin; starch; inorganic binders ($CaSO_4 \cdot H_2O$, sodium silicate, bentonite).

Formulation aids : anticaking agents (pyrogenic SiO_2, pyrogenic Al_2O_3, $CaCO_3$, $MgCO_3$); lubricants (graphite, talc, wax, polytetrafluoroethylene).

Catalysts : Fe_2O_3, $CuO \cdot CrO_2$, V_2O_5, MnO_2.

Pyrotechnic compositions usually contain three to six components. Grain sizes vary from 40 – 400 µm for the components, and from 250 µm to ca. 5 mm for the pyrotechnic compositions themselves. The compositions are utilized in loose form (flash compositions) or compressed into granules or pellets to improve meterability. Most pyrotechnic compositions are then pressed into tubes made of cardboard, plastic, or metal. Special shapes are often conferred on the surfaces of both pressed and granulated compositions to ensure specific burn characteristics. With pressed compositions, the degree of compaction depends largely on the level of stress the charge is expected to experience. A satisfactory density is normally attained at 300 bar, but pressures up to 1200 bar are required for severe application stresses (e.g., in tracer charges, which must survive firing).

Ignition introduces a series of complex solid-state redox reactions in which the activation energies of surface reactions play a crucial role [7]. The main energetic product from the pyrochemical reaction of a pyrotechnic composition is electromagnetic radiation in the infrared region ($\lambda = 1 - 5$ µm). The resulting high temperature may be desirable and even useful (as in priming compositions, rocket-propellant compositions, illuminating and signal-flare compositions), but in certain cases the temperature must be controlled to ensure that other components are only vaporized and not degraded (e.g., colored smoke compositions). Stars for display fireworks need not burn at very high temperatures. The deepness and effect of colors are better achieved with temperatures of only 800 – 1000 °C. In an optimized flare composition, little more than 10 % of the radiation output constitutes visible light useful for illumination and signaling purposes; the ultraviolet contribution is only < 0.1 % of total emission. The most effective *illuminating compositions* are mixtures in which magnesium and $NaNO_3$ are combined with polyester or fluorocarbon binders (55 % magnesium, 40 % sodium nitrate, 5 % binder). Along with the oxidation of magnesium by $NaNO_3$ (reaction temperature > 3000 °C), stoichiometric excesses of magnesium (*bp* 1107 °C) and sodium (*bp* 892 °C) undergo combustion in the vapor state with at-

mospheric oxygen, resulting in a large amount of additional energy. The radiation produced in such flares is distributed mainly by MgO and sodium vapor.

Screening smoke compositions are used particularly by the military for concealment and deception. Classic smoke compositions include the hexachloroethane/Zn/ZnO/Al (HC) system, systems based on red phosphorus (RP) and its derivatives, and related formulations that all release or vaporize hygroscopic substances through pyrochemical reactions. The hygroscopic material interacts with atmospheric humidity to form droplet aerosols that are impenetrable to visible light. More recently developed systems utilize pyrotechnic compositions producing solid particle aerosols based on some form of carbon (carbon black, graphite), which extend the concealing effect into the infrared and far infrared (wavelengths in the 3 – 5 and 8 – 14 μm range).

Smoke compositions likewise serve to produce aerosols. For example, the heat of reaction generated in a bed of $KClO_3$ and lactose hydrate can be used to drive off such sublimable substances as dyes (anthraquinones), NH_4Cl, tear gases (CN, CS), or insecticides (lindane).

Gas compositions are widely used in gas generators, propellant charges, rocket motors, and actuators. (→ Propellants, Chap. 5.3.). These compositions are designed to produce a predetermined large volume of gas over a short period of time. Pyrotechnic gas compositions based on NaN_3 and KNO_3 together with appropriate slag formers have found unique application in automobile air bags (active restraint systems).

Delay compositions molded into small diameter tubes are used for the transmission of detonating impulses. They delay the onset of a device's main function by a precisely adjustable time interval which is equivalent to the burning time of the composition. Most compositions of this type contain substances such as metal powders and oxides, or chromates that react without the production of gas, although delay charges for fireworks and simple pyrotechnic munitions may also be made from compressed black powder.

3. Pyrotechnic Effects

Pyrotechnic effects achievable through the high chemical energies of pyrotechnic compositions can all be attributed to a few basic phenomena. A pyrochemical reaction is initiated when the composition is ignited. What follows is an extremely complicated, multicomponent, solid-state process governed by parameters subject to wide variation. The most important factors are the masses and physical states (allotropic form, density, grain size, internal and external surface area, and purity) of the reactive components. In addition, key macroscopic physical parameters include temperature, extent of compaction of the composition, degree to which the components are mixed, and external configuration of the charge (enclosed or bare, burning on all sides or only on one end, or provided with a specially-shaped burning surface). All these factors are exploited for particular effects, and are used to varying degrees depending upon the intended purpose of a given pyrotechnic device, firework, or pyrotechnic munition.

Hot Slags with High Heat Capacity. Compositions that burn with little gas evolution, such as delay and thermal compositions, but also IR-emitters, produce hot slags with a high heat capacity. Particles and slags with prolonged glow times are used in firework compositions to obtain special comet or fountain effects. Slags formed in the heating-bed components of smoke compositions may also serve to vaporize agents that are to be dispersed. Black powder or boron/potassium nitrate pellets may be used as igniter or booster charges with cannon or solid rocket propellants. Their formation of hot spots (slag droplets) improves and increases the effect of the instantaneous ignition of the propellants.

Light Emission. The oxidation of such light metals as magnesium or aluminum, but also titanium and zirconium, is used to produce light. Oxides that result from the combustion reaction (e.g., MgO, Al_2O_3, TiO_2, or ZrO_2) are the sources responsible for the radiation. Silver, waterfall, and fountain effects in display fireworks are achieved in this way, as is the light emission from flash and illuminating compositions. The color of the resulting light

can be influenced by the choice of the oxidizing agent: $Ba(NO_3)_2$ or $Ba(ClO_3)_2 \cdot H_2O$ for green, $Sr(NO_3)_2$ or $KClO_3/KNO_3$ together with $Sr(COO)_2 \cdot H_2O$ for red, $NaNO_3$ or $Na_2(COO)_2$ for yellow, and copper salts for blue. Star and Bengal-illumination firework compositions seldom contain powdered metals; instead, the reducing agents are often natural resins, such as gum acaroid (red gum) or shellac. The resulting light is an order of magnitude less intense, but the colors are stronger and more impressive.

Gas Emission. Compositions that react rapidly, such as loose black powder, are useful for propulsion, ejection, and explosion (as in firecrackers). On the other hand, compositions that burn evenly over well-defined surfaces are important in rockets and gas generators. Colored illuminating compositions that burn with the evolution of large amounts of gas confer an additional illuminating effect on firework rocket exhaust. Such compositions are often included in wheels, suns, ground spinners, and fountains, where they lead to copious emissions. The thermal compositions in smoke-forming and screening smoke charges disperse clouds of vapor that condense in the air. The smoke-forming charges rely on colored aerosols of sublimable, oxidation-resistant dyes, whereas highly hygroscopic $ZnCl_2$ or P_2O_5 from the latter combine with atmospheric water vapor to form dense masses of fog (screening smoke).

Oscillating Combustion. Whistling and continuously flashing compositions take advantage of intermittent burning. With whistling charges, competing reactions withdraw so much energy from their immediate surroundings that they nearly quench the main reaction (e.g., $KClO_4$/potassium benzoate). In flashing charges, the main reaction is interrupted by one or more "dark reactions" in which one of the principal components is first converted to a pyrophoric, partially oxidized slag [10]. This slag then undergoes further reaction, generating a burst of light, after which the original dark reaction resumes its normal course (e.g., Mg/Al-alloys in $Ba(NO_3)_2$- and $Sr(NO_3)_2$-based compositions).

4. Fireworks

Fireworks have always been popular because of their capacity to bring pleasure and beauty, permitting gunpowder (black powder) to make some amends for its evil reputation as a source of death and destruction. Fireworks have been fabricated since the earliest of times, but prior to the nineteenth century and the advent of modern chemistry they must have been relatively dull and unexciting. During this time the golden color produced by a combination of gunpowder and charcoal with various fats, waxes, oils, and natural resins stimulated the production of rockets, cannon shots, and various types of flames. Our forebears also made some use of iron and antimony, but they would have been stunned by the range of effects available today.

The discovery by BERTHOLLET in 1786 of the oxidizing potential of $KClO_3$ led to a new era in fireworks some forty years later through the introduction of colors (e.g., red, blue, and green). $KClO_3$ alone with fuels produces a weak purple color. It is the salts of copper, barium, strontium, or sodium that provide the intense color, since $KClO_3$, for example, converts the strontium salt to SrCl in the flame. However, $KClO_3$ was also a source of many problems in later years because of its inherent instability in pyrotechnic mixtures and the undesired simultaneous presence in early preparations of highly unstable $Ca(ClO_3)_2$. Whenever possible, $KClO_4$ is now utilized instead.

The next great leap forward came with the development of electrochemistry and subsequent isolation of the metals aluminum and magnesium. Aluminum was a great gift to firework makers, in that it added a brilliant silver dimension to the available effects. Magnesium was equally important, but since corrosion does not lead to a protective oxide layer, unlike aluminum, corrosion continues unabated.

In recent times nothing has been added that was desperately new and eye-catching. Titanium's availability as a noncorrosive metal brought only another silver effect, and while it offers certain commercial advantages, these are offset by the fact that titanium's hardness raises friction problems; sparks created by hard metal particles rubbing against each other can ignite the whole mass.

Advances in chemical technology have also provided a variety of interesting complex compounds, many of which have contributed to a gradually maturing understanding of pyrotechnic reactions. For example, such halogen-donor additives as poly(vinyl chloride), chlorinated rubbers, and chlorinated hydrocarbons have been used to improve color; metal halides are then responsible for the flame color by producing unstable molecules such as SrCl, BaCl, or CuCl. The introduction of polymerizing resins was another important development. Experiments with these resins attempted to solidify pyrotechnic compositions into a hard mass. The solids were mixed with liquid resins and then cast into molds; this avoided the traditional use of hydraulic pressure. SOBRERO's discovery of the explosive properties of nitroglycerine and its extensive commercialization by NOBEL inevitably signed the death warrant for large-scale use of black powder, although it remains an essential ingredient in fireworks, both as an explosive and as a propellant. Europe and North America began importing large quantities of Chinese fireworks during the 1960s. Many assume that fireworks actually originated in the East, but this is not the case. Although the Chinese may be credited for pioneer work with black powder and firecrackers, modern fireworks are very much a western development, and western technology has been transported back to the East only during the twentieth century. Interestingly, since World War II the Chinese have pursued their own independent course, with the result that native compositions, ingenuity, and cheap labor have given rise to a new approach. For many years the West has thus had access to ingenious, intricate, little fireworks based on bright magnalium flames that would have been very expensive to produce domestically.

4.1. Shop Fireworks

Shop fireworks are sold directly to the public throughout the world, usually in conjunction with some holiday or special event: New Year's Eve, a national day like July 4th (Independence Day in the United States), or a special occasion like November 5th in England. Age restrictions are increasingly being applied to the sale of shop fireworks (e.g., 16 in the United Kingdom and 18 in Germany), as are limits on the types of fireworks available for purchase, their size (caliber), and their net explosive content (Chap. 7).

Shop fireworks are also usually subject to a restricted sales period, but this varies widely. For example, the authorized period of sale before New Year's Day in Germany is only two and a half days, but it is three weeks in the United Kingdom prior to the November 5th celebration. Shop fireworks consist mainly of *toy fireworks* and *garden fireworks* (Class 1 and 2 fireworks in Germany, France, and other European countries; see Chap. 7).

4.1.1. Toy Fireworks

Toy fireworks are generally indoor or hand-held devices, such as sparklers.

Amorces and Pistol Caps. Although these items are widely distributed, they are made with the dangerous combination of $KClO_3$ and amorphous red phosphorus. The mixture has to be handled as a wet slurry, and is explosive at low concentrations of water. As the mixture is extremely sensitive to impact and friction only a tiny spot is placed between two layers of paper or into a tiny cup and then dried. It is also used in plastic rings for repetitive fire in toy pistols.

Sparklers. These popular fireworks, known as "Wunderkerzen" in Germany, release a prolonged shower of relatively harmless sparks upon ignition. They are made by dipping wires into a slurry of $Ba(NO_3)_2$, powdered aluminum, and powdered iron. The best sparklers are fabricated by a process similar to matchmaking (→ Matches, Chap. 3.), employing essentially the same adhesives and high temperatures, but this approach is too expensive for the manufacture of cheap commercial products. Good sparklers must not drop hot slag, and the wires must not bend during use.

Snakes. This German discovery ("Schlangen") is unique and unpleasant to prepare as it involves the use of fuming nitric acid and produces large quantities of NO_2 gas. The choice of pitch, and the strength of the acid which nitrates it are important considerations. The snake is formed when the nitrated pitch

has burnt and the liquid ash of the reaction is blown up by gaseous decomposition products. The results are extraordinary – a 15 mm pellet produces a snake more than a meter long.

Table bombs utilize nitrocellulose made from cotton wool to eject their contents. They usually consist of a paper tube with a fuse at the base which connects directly with the nitrocellulose charge. A disk on top of the charge has the dual purpose of lifting the toys or other items and preventing them from being scorched. *Indoor fountains* are also made with powdered nitrated cellulose.

Throwdown crackers ("Knallerbsen" in Germany) are made from silver fulminate (AgOCN). A grain of this material is placed in a twist of paper with sharp sand; a sharp crack is produced when it is thrown onto the ground. Packing for transportation needs careful control.

4.1.2. Garden Fireworks

Garden fireworks are regarded as Class 2 fireworks in Europe. Most are smaller versions of large display fireworks, and the restrictions to which they are subject vary widely. Rockets for sale in shops are identical to display rockets (Fig. 1) except for being smaller. The powder content of Class 2 rockets is limited (e.g., 20 g in Germany). The general public in France, Switzerland, and the United Kingdom is still permitted to purchase relatively large rockets and even star shells.

Bengal illuminations are an important feature of many firework displays, particularly effective for illuminating parks and buildings. They can be produced in any color, but traditionally they are red or green. They are invariably made from $Ba(NO_3)_2$ or $Sr(NO_3)_2$ with added $KClO_3$ and a suitable fuel such as acaroid resin. As they burn slowly, a large quantity of ash is produced. It is therefore necessary to burn them horizontally. The reflected light produced by these devices is much more attractive than the direct view into a flame, so they are frequently placed behind a barrier.

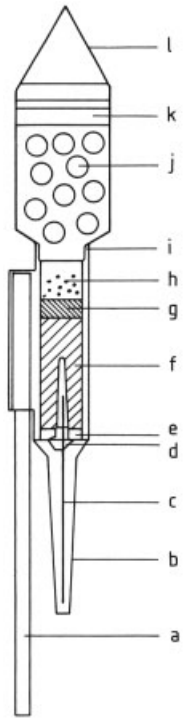

Figure 1. Rocket
a) Stabilizing stick; b) Fuse protection; c) Fuse; d) First fire; e) Nozzle; f) Propellant; g) Delay; h) Expelling charge; i) Rocket tube; j) Stars; k) Sealing material; l) Cap

Firecrackers. Popular with young people as a way of producing a loud noise, firecrackers (or "bangers") have been a source of considerable criticism of the firework industry. In former times they were prepared from $KClO_4$ and aluminum, but most now utilize black powder in an attempt to restrict the resulting noise and flash levels. Every country sets its own limits on the permissible size and black powder content of firecrackers. In the United Kingdom the limit is as little as 1.6 g, whereas in Germany it is 10 g. The burning time of the fuse is a critical parameter in all firecrackers. It must obviously not be too short, but it should also not be too long, because the user may wrongly conclude that the device has failed. A delay of 3 – 6 s is generally regarded as optimal. Methods of igniting firecrackers vary from country to country. In Germany they are often ignited with either a matchtype friction igniter or a black powder fuse, while the United Kingdom displays a pref-

erence for blue igniter paper soaked in KNO_3 (blue touch paper).

Fountains play a central role in garden fireworks, in part because they exist in so many different forms. For children, they may take the form of *golden* or *silver rains*. These are composed of black powder and charcoal, with aluminum or titanium added for a silver effect and iron added for a golden effect. Brilliant *silver jets* utilize $Ba(NO_3)_2$ and aluminum, or black powder and titanium; these are highly prized in France, where they are called "vif argent". Some fountains generate a colored fire made more dramatic by the presence of magnesium–aluminum (magnalium) alloy, which results in large silver crackles and flickering ("strobe") effects. Fountains were formerly made in the United Kingdom from a black-powder mixture containing lampblack and As_2S_3, which formed beautiful and distinctive golden corruscations. However, since lampblack is too dirty to handle and As_2S_3 unobtainable and toxic, these components are now no longer used. Iron powder that has been suitably coated with linseed oil or acaroid resin to prevent corrosion also leads to golden effects. Most garden fountains produce relatively gentle performances, unlike the fierce jets featured in displays.

Display fountains have internal diameters ranging from 18 to 30 mm. They contain large quantities of black powder to produce columns of fire whose height may exceed 2 m. Beautiful patterns may result when several such fountains are carefully fastened to wooden posts and then ignited simultaneously. Similar or slightly modified jets can also be used to turn wheel pieces. In some parts of the world, fountains are prepared with diameters as large as 75 mm, producing high, roaring jets that resemble geysers.

Volcanoes are fountains with cone-shaped tubes that gradually increase the height of the fire and its output of sparkles.

Whistles (Fig. 2) never cease to fascinate people, and when carefully constructed they can provide a wide variety of sounds. It was once thought that the sound emitted had an essentially mechanical origin, but it is now recognized that it results from gas evolved at the base of an empty tube that acts like the resonator of an organ pipe. This gas is generated in an oscillating fashion, with a frequency of ca. 3000–4000 Hz. Differences in tone result from modifications in the diameter and length of the tube, from appropriate design of the surface of the pressed pyrotechnic composition, or the fact that the resulting gas is forced through a nozzle. Special effects are also achieved by burning a whistle at both ends of the tube, but with different lengths of open tubing.

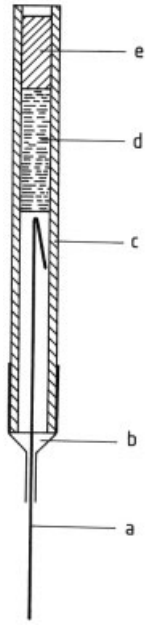

Figure 2. Whistle
a) Fuse (quickmatch); b) Paper or plastic foil; c) Paper tube; d) Whistle composition; e) Clay seal

The chemicals utilized as fuels in whistling systems are aromatic carboxylic acids (e.g., 3,4,5-trihydroxybenzoic acid, benzoic acid salts, salts of salicylic acid, phthalic acid, terephthalic acid and derivatives). The earliest whistling compositions contained potassium and sodium picrate, but these proved to be very dangerous. One of the first whistling mixtures prepared on an industrial scale was made from $KClO_3$ and gallic acid (3,4,5-trihydroxybenzoic acid). Extremely sensitive to impact and friction, this mixture has been the source of many accidents, and is no longer used. Most mixtures today are composed of $KClO_4$ and potassium benzoate, sodium benzoate, sodium salicylate, or a mixture of these fuels. Chinese firework manufacturers now employ monopotassium terephthal-

ate. Each whistling formulation must be manipulated in a different way. All whistling compositions are hazardous and must be handled with great care.

4.2. Display Fireworks

Rockets are display fireworks which have their own means of propulsion in the form of a gas stream; their flight is invariably stabilized with sticks, fins, or other more sophisticated devices. *Shells* are actually bombs launched from mortar tubes. A third type of display firework is known as a *mortar mine*. The firework effect discharges from a mortar tube at a height of 3 m to 20 m. A column of stars, crackers, and noisemakers is ejected from ground level to this height. The calibers vary from 40 mm to 100 mm. All three fireworks produce very similar aerial effects.

Rockets (Fig. 1). The first firework rockets utilized as their propellants modified gunpowder mixtures (typically 61 % KNO_3, 5 % sulfur, and 34 % charcoal), but the proportions of the ingredients can vary considerably. The burning rate of the powder must be controlled to ensure consistency, and this requires some technical skill. Thus, the manufacturer must carefully regulate the amount of gas issuing from the rocket venturi (nozzle) by adjusting the burning surface area of the composition. Firework rocket motors traditionally consisted of a paper tube partially closed at one end by a clay washer. The hole in the washer may be as much as half the diameter of the washer itself. The center of the rocket tube is occupied by a spindle-shaped space offering sufficient burning area to produce a large quantity of gas. In practice, the composition is pressed around a metal spindle, which is subsequently withdrawn. The process is completed by pressing a diameter of powder onto the top of the spindle to create a delay sufficient to permit the rocket to reach its maximum height before stars are ejected.

Firework rockets are fitted with a very lightweight cap filled with colored stars or noise units, which are ejected by the detonation of small bursting charges. Locating the rocket's center of gravity immediately below the venturi ensures that the device will rise vertically, particularly if the stick is a bit heavy. During its upward climb a rocket tends to turn toward the wind because the stick is forced into a downwind position.

Traditional methods of rocket production are subject to many technical problems because of the large burning surface area and the fact that a three component mixture can vary in its burning speed. Modern rockets are formulated with black powder containing a small amount of graphite as this increases the control of the burning speed and aids the pressing operation. The nozzle has become smaller in diameter due to the increased burning time resulting in a smaller burning surface area, and only a slight indentation is made into the powder to increase the initial thrust. Rockets of this type are not as attractive as traditional ones, since both the golden trail and the maximum height are reduced, but they are easier and cheaper to produce, and they are also more reliable. The replacement of paper tubes by aluminum or hard cardboard also facilitates the use of molds for large-scale pressing of the charge.

Rockets are a special feature of Spanish firework displays, which often achieve the most magnificent effects from hundreds of 14-mm rockets fired in a single volley. The Portuguese also utilize enormous rockets (with sticks longer than 2 m) featuring multiple effects.

Roman Candles (Fig. 3). The Roman candle is thought to have been English in origin, and this name is used worldwide. Essentially it is a long, narrow tube, and is a stationary device. Stars, noise units, or little bombs are ejected at short intervals by positioning some kind of delay mechanism between each item. A small charge of blackpowder is placed in the bottom of the closed tube, followed by the projectile star and then a small amount of blackpowder to ease the ejection. The delay is placed on top of the blackpowder. In cheap items it is usually a modified black powder mixture containing black powder mixed with KNO_3, sulfur, and charcoal, which burns slowly. This arrangement is repeated to produce one to ten shots over a period of 20 or 30 s. Well-made items eject 16 – 18 mm stars regularly to the same height, but cheaply-produced oriental items can produce as many as 30 shots with poor consistency.

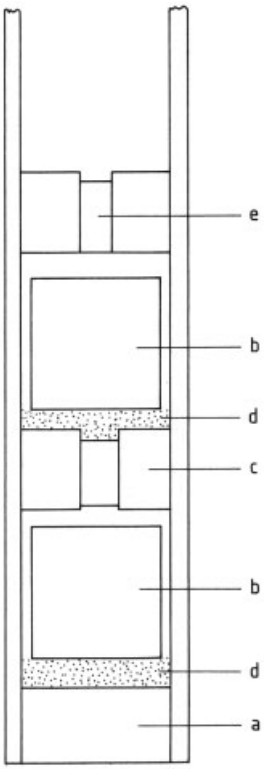

Figure 3. Section of a 30 mm candle
a) Plug; b) Star; c) Felt wad; d) Grain gunpowder; e) Delay

Commercial Roman candles would normally have a caliber between 10 and 20 mm; however, military thinking has considerably advanced the Roman candle technique during the 1970s and 1980s, with the result that excellent items are made in much larger bores from 28 up to 60 mm. In such large diameters black powder-type delay compositions are unreliable and produce too much flame and smoke, to say nothing of the high cost and elevated net explosive content. This problem has been overcome by replacing the bulky powder delay with a pure wool-felt wad that is forced tightly into place. A much smaller delay mechanism in the center of the felt is responsible for transferring the fire. Some manufacturers have devised plastic inserts, and although these are reliable substitutes, they lack the flexibility of felt.

Bombs and Shells. It may be argued that shells are the most spectacular of all fireworks, and that they offer the greatest variety. In the earliest times military personnel managed most civilian firework displays, and bombs and shells for both military and civilian use therefore developed in tandem. The principle underlying these devices is propulsion of a round or cylindrical projectile from a mortar or gun. For civilian applications projectiles had to be made of paper, although paper has been replaced at least in part by plastic.

It was once assumed that bombs should be spherical (Fig. 4), and for many years they consisted of strong paper hemispheres glued together with extra paper pasted on the outside. Firework stars and gunpowder filled the inside, and the sphere was fitted with a time fuse to ensure ignition at maximum height. The contents were invariably dispersed in a disorderly way into the air, with the two paper hemispheres falling to the ground. Some of the fragments could be quite large, since displays often included balls 250 or 300 mm in diameter. Charges of black powder were used to project the bombs.

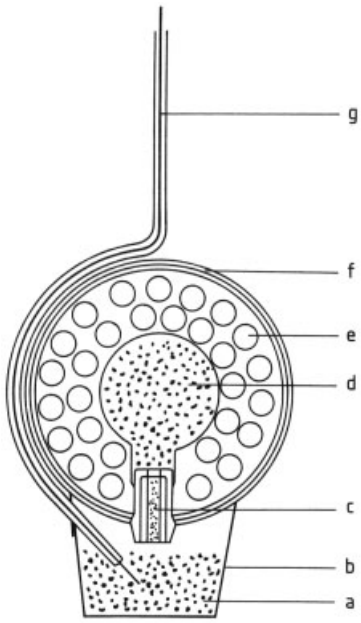

Figure 4. Spherical shell (bomb)
a) Propellant charge (black powder); b) Casing; c) Delay; d) Bursting charge; e) Effect units (stars, etc.); f) Shell case; g) Quickmatch

A sphere is clearly not the optimum ballistic shape, and inevitably there was a move from spherical bombs toward cylindrical shells

(Fig. 5). Originally pioneered by the Italians, many large, beautiful, cylindrical shells are still produced today. The shell body is not particularly strong; it consists primarily of several layers of paper rolled around a cylindrical former. The interior is very carefully but completely filled with gunpowder and stars, noise units, or other effect items to form a solid projectile. The ends are then closed and the device is fitted with a time fuse. Binding the cylinder with many horizontal and vertical turns of string provides additional density and integrity, creating a durable projectile more resistant to the shock of the propelling charge. For further protection during the firing stage, the cylinder's thin paper walls are covered with additional paper. The debris from modern cylindrical shells is less dangerous than that from the earlier spheres, because it consists mostly of bits of paper and string.

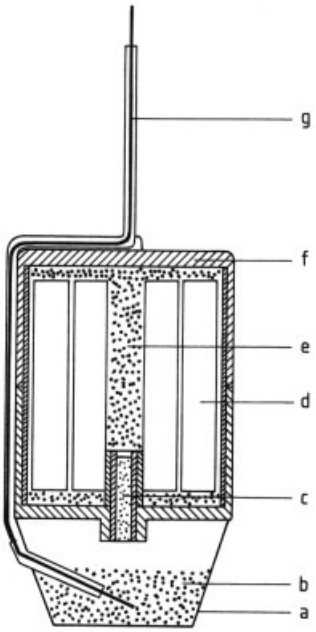

Figure 5. Cylindrical shell
a) Propellant charge casing; b) Propellant charge (black powder); c) Delay; d) Effect-carrying subunits; e) Bursting charge; f) Tube; g) Quickmatch

Several shells can be placed inside one another as a way of producing several pyrotechnic sequences from a single firing. Cylindrical shells may be as large as 25 cm in diameter and 75 cm long, but also as small as 50 mm in diameter. Because of the handwork involved, good Italian shells are expensive, yet a joy to view, especially the large specimens fired individually at the beginning of a display as encouragement for the audience to savor the skill of the pyrotechnist.

The next stage in shell development originated in Japan. Japanese fireworks utilized only black powder until 1880, when $KClO_3$ was first introduced from Europe. However, the distinctive Japanese "warimomo" (chrysanthemum) shell first appeared in 1926. These elegantly patterned devices are spherical and are surrounded by thick layers of pasted paper to provide perfectly balanced resistance against the strong, bursting charge in the center of the shell. Preparing the pasted outer layer is time-consuming and expensive, and requires considerable skill; achieving a given number of equivalent pasted-paper layers on the outside of a sphere using only paper strips is no easy task!

The Japanese also developed spherical stars, which are made by layering one star color over another to provide two or three changes of color. Great manufacturing skill is required to create the impression that all the stars from a given shell change color simultaneously. Methods have also been devised for creating patterns from such shells, causing letters, star-like shapes, spirals, interlocking circles, and various other novelties to appear. Nonetheless, there are limitations; many of the most dramatic effects require observation in a specific viewing plane, which means that the shell must burst in a particular way to prevent spectator disappointment. Ropes and weights have sometimes been attached to the shells to achieve this. Patterns can be produced with Japanese shells by working with stars of different diameters, since larger, heavier stars tend to be projected farther than the lighter, smaller ones. Beautiful flower effects have been achieved in this way, some featuring central pistils of contrasting colors.

The inevitable rise in the cost of manual labor has caused manufacturers to seek new ways of reducing the expenses involved in shell production. Some look to third-world countries for their supplies, but others fabricate their shell bodies from high-density polyethylene, which simplifies the manufacturing process regardless of whether the device is spherical or cylindrical. On the other hand, debris represents a potential problem with plastic shells, and in some coun-

tries there is talk of prohibiting such materials. Manufacturers are therefore looking forward to the discovery of a suitable biodegradable material.

Countries vary in their attitudes toward the handling of shells. In France and the United Kingdom, for example, the public is permitted to purchase and use shells up to 100 mm in caliber, whereas in Germany such large shells are restricted to professionals. There are also differing opinions regarding the "safe distance" for firing shells in public. The United Kingdom issues only general recommendations, but other countries, (e.g., the United States) are attempting to establish a uniform code (e.g., a surrounding radius of 21 m for each 25 mm of shell diameter, so that a 75-mm shell would require a display site of 63 m, a 100-mm shell would require 84 m, etc.). However, these guidelines are only approximate as so much depends on the nature of the site.

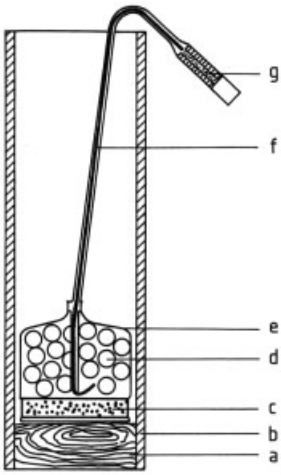

Figure 6. Mortar mine (in mortar)
a) Wooden seal; b) Mortar tube (cardboard); c) Propellant charge (black powder); d) Effect subunits (stars, etc.); e) Paper bag; f) Quickmatch; g) Delay (connecting fuse)

Mines (mortar mines) are used in combination with other fireworks for public displays. A mine is essentially a simplified form of a bomb (Fig. 6). A launching tube is required just as with a bomb, but the mine bursts within the launching tube rather than being first projected into the air. This produces a long column of sparks, colored stars, and noise starting at ground level and extending to a height of 25 m. The effect is very dramatic, because the visual impact is much closer to the spectators than in the case of a bomb or shell. Mines are typically designed for mortar tubes 75 – 125 mm in diameter, but 30 – 40-mm mines are also produced for direct sale to the public.

Combinations. Many of the fireworks described so far lend themselves to combinations, in which a rapid-firing fuse causes the various units to ignite almost simultaneously. Such a fuse is called a "quickmatch," and consists of long strands of cotton that have been passed through a slurry of black powder. When dry, the fuse burns quite steadily in air, but placed inside a narrow paper tube it burns almost instantly as a result of the increased temperature and pressure. *Wheels* are created by tying fast-burning fountains onto pieces of wood pivoted at their centers so they will spin. Fountains can also be attached to stationary wooden structures in conjunction with colored lights and small wheels to produce complex fire patterns.

Water fireworks rely on fountains that have been floated on pieces of cork. Shells are sometimes fired on the water to add dramatic effects, in which case waterproof fuses must be used. Larger shells are often encased in polystyrene foam to help them float.

Fire pictures are created from small, colored lights called lances (each about the size of a cigarette) that burn for almost a minute. Lances are made from $KClO_4$ with the addition of fuel, coloring agents, and intensifiers. When spaced along a piece of wood or cane the tiny pinpoints of light from these small fireworks merge to form an outline, although the projected outlines must be kept quite simple for maximum effectiveness.

The ever-popular *waterfalls* are made from brilliantly-burning aluminum incendiary compositions in combination with barium nitrate or $KClO_4$.

5. Pyrotechnics for Industrial and Military Use

Pyrotechnic devices for industrial and military use include munitions for illumination, communication, and emergency signaling; gas generators; actuators; detonator subassemblies; tracers for many conventional munitions; and smoke

devices. In these applications, chemical energy stored within a pyrotechnic composition is released selectively in order to perform useful work (gas generators, actuators), convey information (signaling and emergency munitions, colored smoke generators), mark the trajectory of a projectile (tracers), transfer a firing impulse (delay fuses, priming charges, intensifying charges), or provide concealment (smoke generators, smoke munitions). Smoke-generating compositions find special application in cartridges designed to release insecticides or irritants, especially in police tear-gas munitions.

A trigger mechanism is used to activate these pyrotechnic devices. Generally, this is a primer cap which is ignited through the percussion of a firing pin (percussion cap) or by friction (friction igniter). Other triggering mechanisms incorporate electrical bridge initiators, striking heads, or black powder priming fuses.

The method used to transmit and intensify the initiating impulse varies. For munitions in cartridge form, a propellant charge is ignited immediately after primary ignition to convey the effective or working part of the pyrotechnic munition to the desired point of application. For munitions designed to perform their function at the site of initiation (signaling and illuminating flares, smoke generators), the initiating impulse is transmitted directly to the pyrotechnic illuminating, signaling, or smoke composition by a booster or transfer charge. In some cases safe use of a devide dictates the incorporation of a pyrotechnic delay column, whereby activation of the effect charge occurs only after the time-delay element has functioned. Delay columns are also utilized when active components (bombettes) within a pyrotechnic munition in cartridge form must be ignited some time after the device has been discharged from the firing weapon (Fig. 7). Use of an auxiliary delay element (or a rocket propellant composition) permits a signal to be displaced a considerable distance from the point of release.

The intended locus of action for such a signal has a significant influence on the design of the device. *Items intended for stationary operation*, such as signal flares, alarm flares (Fig. 8), alarm mines, smoke-signal generators (Fig. 9), floating smoke signals, and illuminating or marking devices for release from aircraft, always have a simple, robust construction. A priming element and an intensifying charge (booster and intermediate compositions) may cause the main charge to be ignited directly, but safety considerations may force ignition to be delayed as described above.

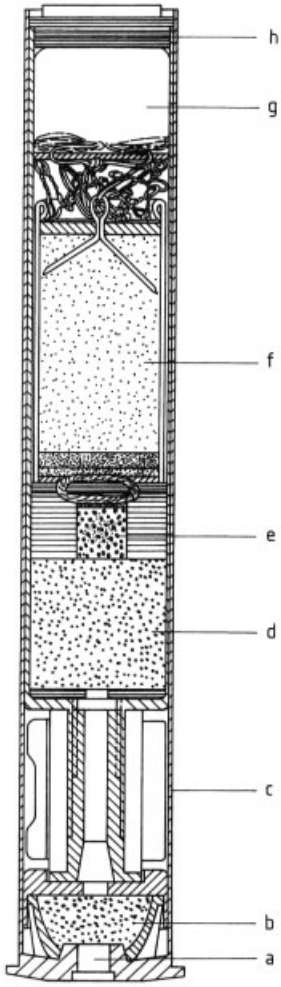

Figure 7. Illuminating cartridge with rocket and parachute a) Percussion cap; b) Propellant charge; c) Cartridge tube; d) Propellant (rocket); e) Delay charge; f) Illuminating star; g) Parachute; h) Sealing disks

Pyrotechnic munitions in cartridge form, with propellant charges, (Fig. 7) often differ significantly from devices for stationary use. Once such a munition has been activated, the component containing the pyrotechnic element is accelerated by combustion gases released from the propellant powder, permitting it to escape from

both the propellant casing and the launching device (e.g., a signal pistol).

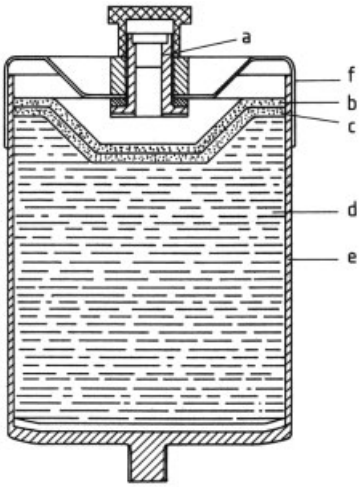

Figure 8. Illuminating alarm flare
a) Percussion cap; b) Starter composition; c) Transition charge; d) Illuminating charge; e) Case; f) Sealing cap

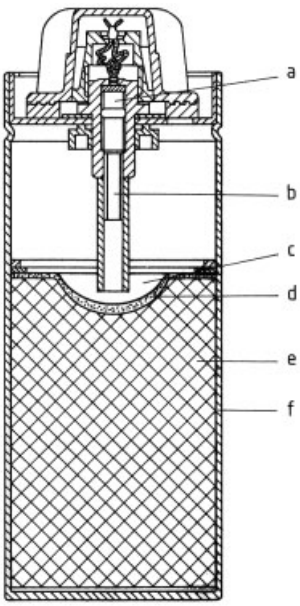

Figure 9. Smoke grenade with friction igniter
a) Friction igniter; b) Delay column; c) First fire; d) Transition charge; e) Smoke composition; f) Aluminium case

The *"non-self-propelled single-star munitions*" utilize the igniting impulse from the propellant charge or primer cap (Fig. 10) directly for activation of the pyrotechnic composition. Shortly after the device leaves the muzzle of the launcher, the effect composition (e.g., a red star) is ignited by a transition composition, which may consist of both a booster and a transition charge. The resulting signal is seen as a red light travelling away from the firing point. A *pyrotechnic tracer* (Fig. 11) incorporated into a conventional projectile operates in a similar way. If a pyrotechnic signal is to be activated at some predetermined distance from the firing point (e.g., to provide better or extended visibility, or as a way of avoiding betrayal of the shooter's location), the propellant charge first ignites a delay element or a rocket-propellant composition.

Pyrotechnic signal munitions in cartridge form, and incorporating the design features described (Fig. 7 and 10), are most frequently encountered in Europe in the following metric calibers: 15, 16, 19, 26.5, 28, and 40 mm. Typical sizes in Anglo-Saxon countries, and especially the United States, include 1 in and 1.5 in. The 15-mm devices, especially in Germany, are in- creasingly taking on the character of fireworks, and are mainly produced for this purpose. Larger caliber devices are used by the military and for emergency signaling in both maritime- and mountain-rescue operations.

Two classes of pyrotechnic munitions, training munitions and simulation devices, are reserved almost exclusively for military purposes. Pyrotechnic *training munitions* combine the potential for realistic training with the advantage of low cost. The associated aiming and firing operations are identical to those for live munitions, but the appropriate sound effects (report) and the trajectory in the case of a projectile or rocket are reproduced or simulated pyrotechnically.

Simulation devices are employed in training maneuvers as a way of evoking the sense of a field of battle, for example, or the effects generated by various types of weapons. Pyrotechnic charges are thus used to simulate the firing and impact of grenades (Fig. 12), artillery strike points, the elimination of a target (e.g., a tank), or the use of a weapon requiring specific countermeasures on the part of the troops.

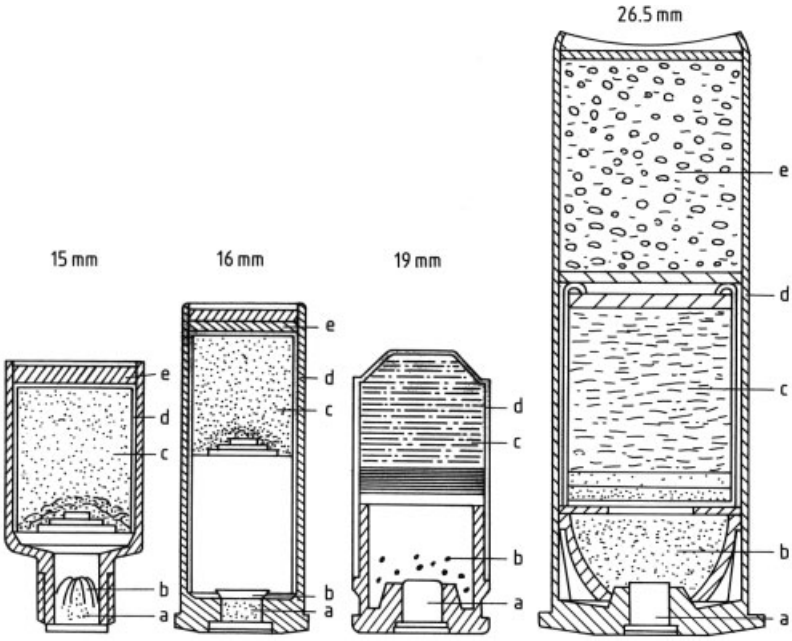

Figure 10. Single-star signal cartridges
a) Percussion cap; b) Propellant; c) Pyrotechnic star (signalling or illuminating) with starter mixture (booster), transition charge, and working composition pressed in aluminum tube; d) Cartridge; e) Filling material

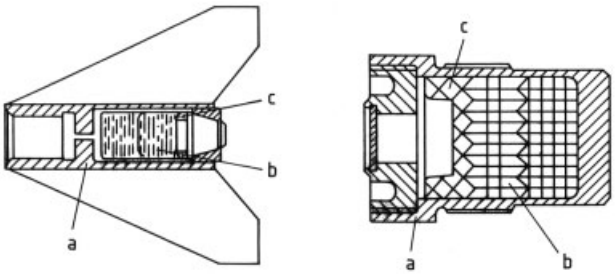

Figure 11. Tracer for ammunition
a) Case; b) Tracer composition; c) Starter composition

6. Production

Most of the larger pyrotechnic producers are affiliated with the chemical industry. The majority originated as family enterprises, and some remain family operations or combinations of such ventures. These firms have devoted their efforts primarily to the fabrication and presentation of public firework displays, but also to the production and marketing of small fireworks. Larger firms in central and southern Europe, the United States, and Japan have long since adopted industrial-scale production techniques. This applies especially to operations in which fireworks manufacture is linked to industrial and military product lines, where the latter often account for most of the sales. Middle-sized operations with 10 – 300 employees continue to typify the pyrotechnic industry. Most are purely family ventures (especially around the Mediterranean), but some operate as independent divisions of larger corporations, particularly as specialized branches of armaments manufacturers [for example, COMET (Diehl group) and NICO (Rheinmetall group)]. Smaller "fireworks-only" firms produce display fireworks (bombs, fountains, rockets, and garden displays) almost entirely by hand, relying on traditional recipes and techniques (see Chap. 4).

Although a highly specialized and traditional craft, professional training in the field of pyrotechnics is nonexistent in virtually all the producer countries. Workers often enter the trade from a metalworking background, or after experience with the stage, theater, or some other aspect of show business. Even within the pyrotechnic industry itself no formal apprenticeship programs are offered. Personnel with metalworking skills, including machinists and toolmakers, together with others conversant with the natural sciences, constitute the working body of pyrotechnists. A typical fireworks manufacturer, which may be a division of an arms supplier, supports 200 – 500 employees and has annual sales of $30 - 100 \times 10^6$ DM (1990).

These enterprises manufacture small fireworks and pyrotechnic devices for maritime signaling; for military illumination, signaling, simulation, training, and smoke concealment; and for industrial applications (e.g., gas generators in air bags and belt tighteners, actuators, and starter mixtures for polymerizations in the chemical industry). Other products include pyrotechnic components for primers and munitions (delay elements, ejectors, bursting charges, and tracers) and munitions designed specifically for police use (e.g., stun grenades and tear gas cartridges).

As a result of political developments during the 1980s, American and European pyrotechnic companies with military product lines have also become engaged in the dismantling of enormous quantities of munitions, including pyrotechnic devices. Disposing of these materials in accordance with environmental regulations imposes severe technical demands on the pyrotechnic industry, especially with respect to technologies for the incineration of explosive wastes and associated off-gas control.

7. Legal Aspects; Storage and Transportation

The dangers inherent in all aspects of pyrotechnics – including production, storage, transport, sale, and use – have resulted in a wide range of legal restrictions throughout the world.

Pyrotechnics manufacturers as well as those engaged in the sale of pyrotechnic products must obtain government licenses for both plants and warehouses. Provisions regarding the distribution of fireworks vary greatly from country to country, and will not be covered here in detail. All share the common goal of exerting some measure of control over the way fireworks are sold and utilized to minimize potential danger and disturbance of the general public. Sales to young people are generally prohibited, and even adults are often permitted to acquire only certain types of fireworks for display at certain times.

The purchase of pyrotechnic products for industrial purposes and of pyrotechnic munitions (e.g., emergency flares) is similarly governed by restrictions.

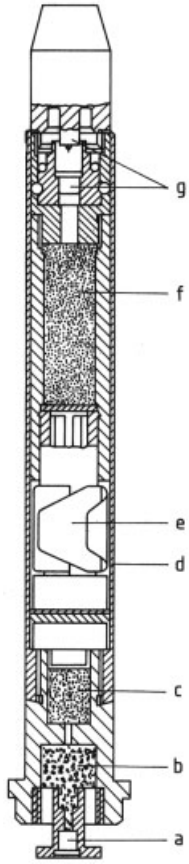

Figure 12. Subcaliber demonstration cartridge for a mortar a) Priming cap; b) Sabot "dummy" ejection charge; c) Propellant charge; d) Cartridge; e) Stabilizing fins; f) Flash/smoke charge; g) Impact fuse

The transport of dangerous goods has long been the subject of a consistent set of inter-

national regulations. Most pyrotechnic goods fall into Class 1 as defined by recommendations from the Committee of Experts on the Transport of Dangerous Goods of the United Nations Economic and Social Council (ECOSOC). These recommendations in turn form the basis for nearly all important national and international transport regulations [11]. Class 1 includes all explosives, as well as items containing explosives, and thus essentially all pyrotechnic materials. Class 1 is subdivided into five divisions based on the nature of the potential hazard posed by a particular substance or product. The classification criteria themselves were also defined by the United Nations, and have been recommended for international use [12]. The five categories of explosives follow [11]:

Division 1.1. Substances and articles that create a mass explosion hazard (where a mass explosion is one that affects almost the entire load virtually instantaneously).

Division 1.2. Substances and articles that constitute a projection hazard, but not a mass explosion hazard.

Division 1.3. Substances and articles that present a fire hazard and either a minor blast hazard or a minor projection hazard or both, but not a mass explosion hazard. This division comprises substances and articles which (1) give rise to considerable heat, or (2) burn one after another, producing minor blast and/or projection effects.

Division 1.4. Substances and articles which present no significant hazard. This division comprises substances and articles presenting only a small hazard in the event of ignition or initiation during transport. The effects are largely confined to the package, and no projection of fragments of appreciable size or range is to be expected. An external fire must not cause virtually instantaneous explosion of almost the entire contents of the package.

Division 1.5. Very insensitive substances that nevertheless present a mass explosion hazard (slurry explosives).

Because some substances and products are not mutually compatible from a safety or chemical standpoint, Class 1 is also divided into 12 "compatibility groups" to avoid problems that might arise through inappropriate packaging. These compatibility groups are identified by cap-ital letters. Black powder is listed among the explosives in Division 1.1 and has been allocated to compatibility group D. Most other pyrotechnic items fall into compatibility group G of Division 1.1 (e.g., display firework shells with flash compositions or large calibers), Division 1.3 (display fireworks of minor caliber, flares, and illuminating cartridges), and Division 1.4 (Category 2 fireworks, signal ammunition, smoke grenades).

Another hazard category with the compatibility designation S in Division 1.4 is defined by United Nation recommendations as follows: "Substances and articles of this division are in compatibility group S if they are so packaged or designed that any hazardous effects arising from accidental functioning are confined within the package unless the package has been degraded by fire, in which case all blast or projection effects are limited to the extent that they do not significantly hinder fire-fighting aids or other emergency response efforts in the immediate vicinity of the package." (For example, Category 1.4 S includes Category 1 fireworks, caps, and igniters.)

Assigning explosive substances and articles to explosion hazard groups has resulted in a broad international consensus with regard to assessment and classification, which now extends far beyond the transport sector alone. All Class 1 substances and articles now not only belong to designated hazard and compatibility groups, but they also carry internationally binding code numbers referred to as "UN numbers." For example, the designation "0333, Fireworks, 1.1 G" unambiguously identifies an article as fireworks that present a mass explosion hazard, and "0336, Fireworks, 1.4 G" indicates fireworks that present no special hazard. The codes can be used to find and classify the article in relevant transport regulations applicable to carriers of all types [13–16].

Efforts are being made to expand the application of these internationally accepted classifications and hazard descriptions to other aspects of handling explosive materials and devices, including production and storage. Thus, this system serves as a basis for classifying pyrotechnic

compositions, semifinished products, and articles under the accident prevention regulations of the German trade associations [17, 18]. It also applies to the assessment of hazards posed by the storage of explosives, devices containing explosives, and munitions [19]. The definitions are equivalent and often identical to those proposed by the United Nations.

A European project foresees complete uniformity in the description and distribution of pyrotechnic devices, specifically fireworks, and is an important step in the standardization of safety evaluation and technical characterization [20]. A similar standardization effort is ongoing in the United States and Canada. Virtually every European country classifies fireworks according to both structure and intended use, but with differing restrictions and regulations. Efforts toward standardization were initiated in 1990 in an attempt to establish a single set of generally applicable category definitions [21]:

Category 1. Fireworks which present a very low hazard and which are intended for use in confined areas, including fireworks which are intended for use inside domestic buildings

Category 2. Fireworks which present a low hazard and which are intended for outdoor use in confined areas

Category 3. Fireworks which present a medium hazard and which are intended for outdoor use in large, open areas

Category 4. Fireworks which present a high hazard or are unfinished, and which are intended for professional use only.

8. References

1. H. Treumann, "Ein Ordnungs- und Dokumentationssystem für die Pyrotechnik." *Proc. ICT Meeting,* Karlsruhe 1975, 407.
2. H. Ellern: *Military and Civilian Pyrotechnics,* Chemical Publishing Co. Inc., New York 1968.
3. J. A. Conkling: *Chemistry of Pyrotechnics,* Marcel Dekker, New York 1985, p. 3.
4. J. R. Partington: *A History of Greek Fire and Gunpowder,* W. Heffer & Sons Ltd., Cambridge, 1960.
5. A. Lotz: *Das Feuerwerk, seine Geschichte und Bibliographie,* Edition Olms, Zürich 1978.
6. Revd. R. Lancaster: *Fireworks, Principle and Practice,* 2nd ed. Chemical Publishing Co. Inc., New York 1992, p. 1 – 23.
7. *Das Buch der Feuerwerkskunst,* Greno Verlagsgesellschaft, Nördlingen 1987.
8. J. H. McLain: *Pyrotechnics from the Viewpoint of Solid State Chemistry,* The Franklin Institute Press, Philadelphia, Pa., 1980.
9. T. Shimizu: *Fireworks, The Art, Science and Technique,* Maruzen Co., Tokyo 1981.
10. U. Krone, "Oscillierend abbrennende pyrotechnische Sätze," *Proc. ICT Meeting,* Karlsruhe 1975, 225.
11. Transport of Dangerous Goods, Recommendations of the Committee of Experts on the Transport of Dangerous Goods, United Nations, New York 1984,ST/SG/AC.10/1/Rev. 3.
12. Recommendations on the Transport of Dangerous Goods–Tests and Criteria, 1st ed., United Nations, New York 1986, ST/SG/AC.10/11.
13. ADR, Accord européen relatif au transports des marchandises dangereuses par route. BGBl. II, p. 1491, 21st Aug. 1969 (Fassung der Anlagen A und B siehe Anlagenband zum BGBl. I, Nr. 62, 17. 11. 1990).
14. RID, Règlement international des marchandises dangereuses par chemin de fer. übereinkommen über den internationalen Eisenbahnverkehr (COTIF), appendix B, Bern 1990.
15. IMDG Code, International Maritime Dangerous Goods Code, International Maritime Organization (IMO), London, Amendment 25, 1990.
16. IATA-DGR, International Air Transport Association, Dangerous Goods Regulations Montreal – Geneva, 32nd ed., 1991.
17. Unfallverhütungsvorschrift "Explosivstoffe und Gegenstände mit Explosivstoff", Allgemeine Vorschrift (VBG 55 a), BG-Chemie, Aug. 1978.
18. Unfallverhütungsvorschrift "Herstellen pyrotechnischer Gegenstände" (VBG 55k), BG-Chemie, April 1981.
19. 2. Verordnung zum Sprengstoffgesetz (2. SprengV), 05.09.1989, BGBl. I, pp. 1620, 2458.
20. European Committee for Standardisation (CEN), Technical Committee 212, Fireworks.
21. Resolution 9, Document CEN/TC 212-N 54, taken at 1st Plenary session, London, June 18, 1990.

Pyrrole

Albrecht Ludwig Harreus, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. **Pyrrole** 1
2. **Pyrrole Derivatives** 2
3. **Toxicology and Occupational Health** . . 3
4. **References** 3

1. Pyrrole

Pyrrole[*109-97-7*],1*H*-pyrrole, azole, C_4H_5N, M_r 67.09, with its five-membered heterocyclic ring system has a fundamental role in the living world.

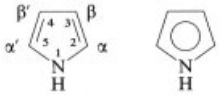

1 **2**

Pyrrole derivatives (cyclic tetrapyrroles) assume their most important role in respiration (hemoglobin, cytochromes) and photosynthesis (chlorophyll). Bile pigments are examples of natural linear tetrapyrroles. Low molecular mass pyrroles with antibiotic action have been isolated from microorganisms (pyrrolnitrin, → Antibiotics).

F. F. Runge detected pyrrole in 1834 as a constituent of coal tar. The very nonspecific reaction used for detection (red coloration of a spruce splint dampened with hydrochloric acid) gave pyrrole its name (Greek pyrros = fiery).

Physical Properties. Pyrrole, *mp* −23.4 °C, *bp* 130 °C (101.3 kPa), when freshly distilled, is a colorless liquid with a chloroform-like odor, which rapidly becomes brown in air and gradually resinifies. Pyrrole is hygroscopic, not very soluble in water (ca. 5 %), and miscible with most organic solvents. Further physical data are as follows:

Density d_4^{20}	0.9698
Refractive index n_D^{20}	1.5085
Viscosity (20 °C)	1.31 mPa · s
Flash point (DIN 51 755)	33 °C
Ignition temperature (DIN 51 794)	550 °C
Explosion limits	
lower	3.1 vol %
upper	14.8 vol %

The most important physical properties and spectroscopic data for pyrroles are summarized in [5].

The pronounced aromatic character of pyrrole is not consistent with the classical constitutional formula (**1**). The direction of the dipole moment, the bond lengths, and a very low basicity show that the free electron pair on nitrogen participates in the formation of an aromatic 6π-electron system, which is conveyed schematically by (**2**), whereas (**1**) represents only one of several possible mesomeric structures.

Chemical Properties. Pyrrole is unstable toward mineral acids, and is protonated in the 2-position. The resulting pyrrylium cation polymerizes very readily to give high molecular mass pyrrole resins (pyrrole red).

Because of its high π-electron density, the characteristic reactions of pyrrole are electrophilic substitutions. Substitution occurs preferentially in the 2-position. The reactions proceed under mild conditions, but must be adapted to the compound's sensitivity toward acid. Examples of electrophilic substitution include acetylation with acetic anhydride (which does not require a Friedel – Crafts catalyst), sulfonation with the SO_3– pyridine complex, nitration with acetyl nitrate (nitric acid/acetic anhydride), and halogenation with bromine, which leads directly to a tetrasubstituted product.

As a result of delocalization of the electron pair on the nitrogen, pyrrole is weakly acidic (pK_a ca. 17.5). Corresponding salts can be prepared with alkali metals in liquid ammonia or metal hydrides in inert solvents (e.g., tetrahydrofuran). These can be converted into *N*-alkylpyrroles by reaction with alkyl halides. Pyrryl potassium reacts with carbon dioxide to give the potassium salt of pyrrole-2- carboxylic acid.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_453

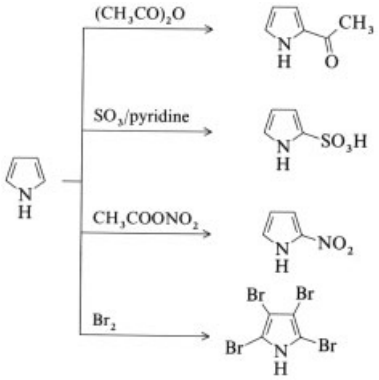

Pyrrole can be reduced catalytically (Pd, Pt, Rh, Ni) to pyrrolidine. With less reactive reducing agents (e.g., zinc in glacial acetic acid) the intermediate 3-pyrroline [*109-96-6*] (2,5-dihydropyrrole) can be obtained. This is the only one of the three dihydropyrroles that is somewhat stable.

Production and Use. Suitable processes for the industrial production of pyrrole include the reaction of furan (→ Furan and Derivatives) with ammonia over SiO_2/Al_2O_3 catalysts (zeolites) [6] and the catalytic dehydrogenation of pyrrolidine [*123-75-1*] [7].

The reaction of 2,5-dimethoxytetrahydrofuran with ammonia over zeolites [8] also leads to pyrrole, as does gas-phase cyclocondensation of butynediol with ammonia over an aluminum oxide/thorium oxide catalyst [9].

Mild steel drums lined with polyethylene are suitable as transport and storage vessels.

BASF and Shell produce pyrrole commercially, although its use remains limited. An electrically conducting polypyrrole is prepared from pyrrole by electrochemical polymerization, → Polymers, Electrically Conducting, Chap. 5.3. [10].

2. Pyrrole Derivatives

There are many possible synthetic routes to substituted pyrroles [11–13]. Of the classical syntheses of *C-substituted pyrroles*, the Knorr pyrrole synthesis (**3**) together with the Paal–Knorr (**4**) and Hantzsch syntheses (**5**) are particularly important. Appropriate α-aminoketones for the Knorr synthesis are usually generated in situ by reduction of the corresponding oximino compound, most frequently with zinc in acetic acid. The methylene group of the carbonyl compound has to be activated, therefore β-dicarbonyl compounds such as β-ketoesters and β-diketones are preferred.

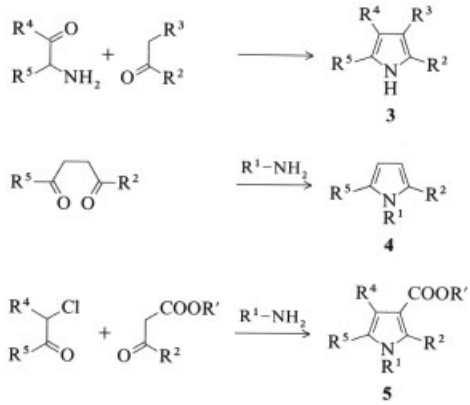

N-Substituted pyrroles can easily be prepared industrially by dehydrogenation of the corresponding pyrrolidines. This reaction is carried out in the gas phase at 250–300 °C over such catalysts as Pd/Al_2O_3 or $Pd/MgO/Al_2O_3$ [7], [14]. Yields of *N*-alkylpyrroles surpass 90 %. Silicates are less suitable as carriers. *N*-Alkylpyrroles can also be produced by reaction of 2-butene-1,4-diol with amines in the gas phase (e.g., copper catalysts) [15].

Many *N*-alkyl and *N*-aryl pyrroles can be produced quite elegantly by reacting appropriate amines with 2,5-dialkoxytetrahydrofurans, in particular 2,5-dimethoxytetrahydrofuran [*696-59-3*][16],which is readily accessible industrially through electrooxidation of furan with subsequent hydrogenation [8], [11], [12], [17].

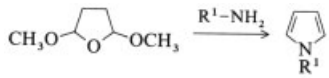

Special methods have been developed for the synthesis of *3-substituted pyrroles* [18].

3-Substituted 4-phenylpyrroles, such as fenpiclonil [*74738-17-3*] (Ciba-Geigy), display fungicidal properties, and are derived from the natural product pyrrolnitrin, → Fungicides, Agricultural [19].

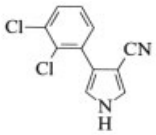

Fenpiclonil

Derivatives of pyrrole-3,4-dicarboxylate esters show herbicidal activity [20], and those of 1-aryl-3- cyanopyrroles have insecticidal effects [21]. *N*-Methyl-2-pyrrylmethyl sulfides can be used as flavorings [22].

N-Methylpyrrole [*96-54-8*], 1-methylpyrrole, 1-methyl-1*H*-pyrrole, C_5H_7N, M_r 81.12, *mp* −57 °C, *bp* 112 – 113 °C (101.3 kPa), d_4^{20} 0.9088, n_D^{20} 1.4889, flash point 16 °C (DIN 51 755), ignition temperature 400 °C (DIN 51 794), is a bright yellow liquid that must be protected from light and air during storage. *N*-Methylpyrrole is miscible with almost all common organic solvents and ca. 1 % soluble in water.

N-Methylpyrrole is added as a stabilizer to chlorinated hydrocarbons [23]. It serves as starting material for the synthesis of 1-methylpyrrole-2-acetic acid and its derivatives, which are building blocks for many anti-inflammatory drugs, including tolmetin (Cilag), → Anti-inflammatory – Antirheumatic Drugs [24].

N-Methylpyrrole is produced commercially by BASF.

3. Toxicology and Occupational Health

Pyrrole is toxic if swallowed. The acute oral LD_{50} in rats is 137 mg/kg [25]. The acute intraperitoneal LD_{50} in mice is 98 mg/kg [26]. The subcutaneous LD_{50} in mice is 61 mg/kg [27]. Pyrrole does not cause skin irritation when applied to intact skin of white rabbits for up to 4 h, but it does cause significant ocular lesions, with a risk of serious permanent damage to the eyes. The LC_{50} is 2.27 mg/L (rat, inhalation, 4 h). Rats exposed to an atmosphere highly enriched or saturated with pyrrole at room temperature died after 30 min. Irritation of eye and nasal mucosa were observed after overexposure [25].

Pyrrole showed no mutagenic activity in the Ames test (with or without metabolic activation) [28–32]. Pyrrole did not induce DNA damage in a test with *Escherichia coli* [32] and also gave a negative result in the DNA-repair test in rat hepatocytes [33], [34]. Pyrrole had positive genotoxic activity in the rec-assay with *Bacillus subtilis*, and exhibited the ability to cleave DNA in lambda phages [35]. After intraperitoneal administration of pyrrole to mice, an epoxide was identified as a toxic metabolite in the lung. The enzyme involved was P450. Emphysema and bronchiolar necrosis were found in the animals [36], [37].

***N*-Methylpyrrole** is harmful if swallowed. The acute oral LD_{50} in rats is about 1400 mg/kg. The dermal LD_{50} in rabbits is > 400 mg/kg. The substance causes irritation of healthy intact or scarified skin in white rabbits when it is applied for up to 2 h, but it does not lead to significant ocular irritation. The LC_{50} in rats after inhalation for 1 and 4 h is greater than 22.4 and 5.6 mg/l, respectively; there was no mortality. Rats exposed to an atmosphere highly enriched or saturated with *N*-methylpyrrole at room temperature died after 1 h. Both severe irritation of the mucosa and narcotic effects were observed [38].

N-Methylpyrrole showed no mutagenic activity in the Ames test with or without addition of S-9 mix [28], [38]. It was positive in the rec-assay with *Bacillus subtilis* (± S-9 mix), and exhibited an ability to cleave DNA in lambda phages [35].

4. References

General References

1. A. Gossauer: *Die Chemie der Pyrrole,* Springer Verlag, Berlin 1974.
2. R. A. Jones, G. P. Bean: "The Chemistry of Pyrroles," *Organic Chemistry,* **vol. 34,** Academic Press, New York 1977.
3. A. H. Jackson in P. G. Sammes (ed.): *Comprehensive Organic Chemistry,* **vol. 4,** Pergamon Press, Oxford –New York 1979, pp. 275.
4. R. A. Jones (ed.): *Pyrroles. Chemistry of Heterocyclic Compounds,* vol. 48 part 1, Wiley Interscience, New York 1990.

Specific References

5. R. A. Jones in A. R. Katritzky, A. J. Boulton (eds.): *Advances in Heterocyclic Chemistry,* **vol. 11,** Academic Press, New York – London 1970, pp. 383.
6. K. Hatada, M. Shimada, K. Fujita, Y. Ono, T. Keii, *Chem. Lett.* **1974,** 439 – 442. Daicel, JP-Kokai 83/090 548, 1982. Asahi, JP-Kokai 89/301 658, 1988 (T. Yamamoto, M. Iwasaki).
7. BASF, EP 67 360, 1981 (N. Goetz, L. Hupfer, W. Franzischka).
8. BASF, EP-A 303 206, 1987 (W. Hoelderich, H. Hesse, H. Siegel).
9. W. Reppe, *Justus Liebigs Ann. Chem.* **596** (1955) 155.
10. H. Naarmann, *Angew. Makromol. Chem.* **162** (1988) 1 – 17.
11. J. M. Patterson, *Synthesis* 1976, 281 – 304.
12. L. N. Sobenina, A. I. Mikhaleva, A. I. Trofimov, *Khim. Geterotsikl. Soedin.* 1989, 291 – 308.
13. L. N. Sobenina, A. I. Mikhaleva, A. I. Trofimov, *Russ. Chem. Rev. (Engl. Transl.)* **58** (1989) 163 –180.
14. Ansul Chem. Company, US 3 008 965, 1959 (R. J. Zellner).
15. BASF, DE-OS 3 309 355, 1983 (H. Menig, M. Fischer, K. Baer).
16. BASF, DE-AS 2 710 420, 1977 (D. Degner, H. Nohe, H. Hannebaum).
17. N. Elmig in R. A. Raphael, E. C. Taylor, H. Wynberg (eds.): *Advances in Organic Chemistry,* **vol. II,** Interscience Publishers, New York 1960, pp. 100 – 102.
18. H. J. Anderson, C. E. Loader, *Synthesis* 1985, 353 –364.
19. H. P. Fischer, *Nachr. Chem. Tech. Lab.* **38** (1990) 732 –740. Nippon Soda, DE 2 927 480, 1979 (K. Ahkuma, H. Takagi, A. Nakata). Bayer, EP-A 392 286, 1989 (D. Wollweber, W. Brandes, S. Dutzmann, G. Haenssler).
20. Monsanto, EP-A 88 743, US 4 461 642, 1982 (R. K. Howe, L. F. Lee). Shell Int. Res., GB 2 214 180, 1988 (K. M. Patel, J. E. Powell).
21. Rhône-Poulenc, EP-A 372 982, 1988 (P. Timmons et al.).
22. Firmenich, US 3 985 906, 1975 (M. Winter et al.).
23. Du Pont, US 2 795 623, 1954 (F. W. Starks). Hooder Chem. Corp., US 2 998 461, 1959 (D. H. Campbell).
24. A. Kleemann, J. Engel: *Pharmazeutische Wirkstoffe,* 2nd ed., Georg Thieme Verlag, Stuttgart 1982.
25. BASF, unpublished results (1987 – 1988).
26. R. Rinaldi, Y. Bernard, *Prog. Biochem. Pharmacol.* **1** (1965) 542 – 549.
27. C. F. Reinhardt, M. R. Britelli (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A,** Wiley-Interscience, New York 1981, pp. 2700 – 2701.
28. H. U. Aeschbacher et al., *Food Chem. Toxicol.* **27** (1989) 227 – 232.
29. L. R. Ferguson, W. A. Denny, D. G. MacPhee, G. Donald, *Mutat. Res.* **157** (1985) 29 – 37.
30. I. Florin, L. Rutberg, M. Curvall, C. R. Enzell, *Toxicology* **15** (1979) 219 – 232.
31. C. H. Ho et al., *Mutat. Res.* **85** (1981) 335 – 345.
32. M. Riebe, K. Westphal, P. Fortnagel, *Mutat. Res.* **101** (1982) 39 – 43.
33. G. M. Williams, H. Mori, J. Hirono, M. Nagao, *Mutat. Res.* **79** (1980) 1 – 5.
34. G. M. Williams, *Regul. Toxicol. Pharmacol.* **5** (1985) 132 – 144. *Food. Addit. Contam.* **1** (1984) 173 – 178.
35. E. H. Kim, K. Shinohara, H. Murakami, H. Omura, *J. Fac. Agric., Kyushu Univ.* **31** (1987) 279 – 285.
36. M. R. Boyd, *CRC Crit. Rev. Toxicol.* **7** (1980) 103 –176.
37. J. W. Bridges, L. F. Chasseaud (eds.): *Progress in Drug Metabolism,* **vol. 9,** Taylor & Francis, London –Philadelphia 1986, p. 83.
38. BASF, unpublished results (1978 – 1990).

Pyrrolidine → Amines, Aliphatic

2-Pyrrolidone

Albrecht Ludwig Harreus, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **2-Pyrrolidone** 1
3. ***N*-Methyl-2-pyrrolidone** 2
4. ***N*-Vinyl-2-pyrrolidone** 4
5. **Other Pyrrolidone Derivatives** 5
6. **Toxicology and Occupational Health** . . 5
7. **References** 7

1. Introduction

Pyrrolidones became available industrially through the fundamental work of W. Reppe on acetylene chemistry. He synthesized butynediol and from it γ-butyrolactone, the starting material for the commercial production of pyrrolidones (→ Butyrolactone).

2. 2-Pyrrolidone

Physical Properties. 2-Pyrrolidone [*616-45-5*], pyrrolidone, 2-pyrrolidinone, γ-butyrolactam, C_4H_7NO, M_r 85.12, *mp* 25.57 °C, *bp* 245 °C (101.3 kPa), is a colorless, hygroscopic liquid above its melting point. It is miscible with water and common organic solvents, but not with aliphatic or cycloaliphatic hydrocarbons. Further physical data follow:

Density (30 °C)	1.103 g/cm^3
Refractive index n_D^{30}	1.484
Viscosity (30 °C)	10.2 mPa · s
Enthalpy of fusion	135 kJ/kg
Enthalpy of evaporation (40 °C)	666 kJ/kg
Dielectric constant (35 °C)	27.1
Dipole moment (30 °C, dioxane)	3.8 D
Flash point (DIN 51 758)	138 °C
Ignition temperature (DIN 51 794)	390 °C

Chemical Properties. 2-Pyrrolidone has amphoteric properties: it forms salts with hydrogen chloride, hydrogen bromide, and alkalis. It exhibits typical properties of primary lactams [1]. On heating with strong aqueous alkalis or mineral acids it is hydrolyzed to 4-aminobutyric acid [2]. The reactions at the lactam nitrogen are industrially the most important, in particular viny-lation (see Chap. 4). Alkali-metal salts of 2-pyrrolidone react with alkyl halides or sulfates to give the corresponding *N*-alkyl-2-pyrrolidones. *N*-Alkylation can also be achieved by treatment of 2-pyrrolidone with alcohols in the gas phase on alumina [3] or copper chromite catalysts [4]. *N*-Acylation can be readily achieved with carboxylic acid anhydrides or chlorides [2]. Noteworthy among the condensations with aldehydes are the reactions with formaldehyde to give *N*-hydroxymethyl-2-pyrrolidone [5] and with acetaldehyde to give *N*-(1-hydroxyethyl)-2-pyrrolidone [6, 7], which can be used to produce *N*-methyl- and *N*-vinyl-2-pyrrolidone, respectively. 2-Pyrrolidone polymerizes to a polyamide (nylon 4) in the presence of alkali and a carboxylic acid chloride catalyst (→ Fibers, 4. Synthetic Organic) [8].

Production. 2-Pyrrolidone can be produced by the following methods:

1) Catalytic [9] or electrochemical reduction [2] of succinimide
2) Carbonylation of allylamine [10]
3) Hydrogenation of succinic acid dinitrile under hydrolytic conditions [11]
4) Reaction of maleic or succinic anhydride in aqueous ammonia with Pd – Ru catalysts [12]

Industrially, 2-pyrrolidone is almost exclusively produced by reacting aqueous γ-butyrolactone [*96-48-0*] with ammonia [13]:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_457

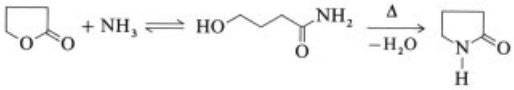

Formation of the undesired 4-(*N*-2-pyrrolidonyl)butyramide can be suppressed if the twostage reaction is carried out in the gas phase on a magnesium silicate catalyst (250–290 °C, 0.4–1.4 MPa) [14].

Quality Specifications and Analysis. Technical-grade 2-pyrrolidone has a minimum content of 99.5 wt % (gas chromatography), and a maximum water content of 0.1 wt % (Karl Fischer titration).

Handling, Storage, and Transportation. 2-Pyrrolidone is transported in steel drums. Other materials suitable for storage and transportation are aluminum and stainless steel. Moisture must be excluded. 2-Pyrrolidone can be kept for ca. 1 year in closed containers. Yellowing does not affect the quality. 2-Pyrrolidone is not self-igniting or explosive, but is flammable.

Uses. Over 95 % of the 2-pyrrolidone produced is processed into *N*-vinyl-2-pyrrolidone. It is also used in the pharmaceutical industry for the production of pyrrolidone nootropics, especially piracetam (see 5). 2-Pyrrolidone is also used as a solvent and reaction medium since it is high boiling, inert, and noncorrosive. It can be used to decolorize kerosene and other hydrocarbons. Aromatic hydrocarbons can be extracted from petroleum with 2-pyrrolidone. In the polish and cleaning materials industry it is added to styrene–acrylate copolymer dispersions as a film-forming agent.

Economic Aspects. Worldwide demand for 2-pyrrolidone was 14 000 t in 1983 [15] and has since risen considerably. Important producers are BASF and GAF.

3. *N*-Methyl-2-pyrrolidone

N-Methyl-2-pyrrolidone (NMP) [*872-50-4*], *N*-methylpyrrolidone, 1-methyl-2-pyrrolidone, C_5H_9NO, M_r 99.13, is an important, versatile solvent and reaction medium for the chemical industry because of its favorable properties [16, 17].

Physical Properties. NMP, *mp* −24.4 °C, *bp* 204.3 °C (101.3 kPa), is a colorless liquid with a weakly aminelike odor; it is completely miscible with water and most common organic solvents. Further physical data follow:

bp (1.3 kPa)	81–82 °C
Density (25 °C)	1.028 g/cm^3
Refractive index n_D^{25}	1.469
Viscosity (20 °C)	1.796 mPa · s
Surface tension (25 °C)	0.041 N/m
$t_{crit.}$	451 °C
$p_{crit.}$	4.78 MPa

Vapor pressure

t, °C	40	60	80	100	150
p, kPa	0.134	0.459	1.33	3.37	21.9

Specific heat capacity

t, °C	0	25	50	100
c_p (liquid), kJ kg^{-1} K^{-1}	1.70	1.78	1.86	2.03

Enthalpy of evaporation (20 °C)	550 kJ/kg
Thermal conductivity (20 °C)	1.8 W m^{-1} K^{-1}
Dielectric constant (23 °C)	32
Dipole moment	4.09 D
Flash point (DIN 51 758)	91 °C
Ignition temperature (DIN 51 794)	245 °C
Lower explosion limit	1.3 vol %
Upper explosion limit	9.5 vol %

Chemical Properties. NMP is a very weak base and forms a solid hydrochloride with anhydrous hydrogen chloride. A 10 % aqueous solution has a pH of 7.7–8. NMP is chemically very stable. The lactam ring is only opened by strong aqueous acids and bases to give 4-methylaminobutyric acid. NMP has limited resistance toward oxygen. It is oxidized at the 5-position, after several intermediate stages *N*-methylsuccinimide is formed. This compound can also be produced directly from NMP, for example by oxidation with ruthenium tetroxide [18]. The carbonyl group of NMP reacts

with phosgene, phosphorus pentachloride, and other chlorinating agents. 2-Chloro-1-methyl-1-pyrrolidinium chloride [*15862-82-5*] can be readily produced with phosgene [19] and reacts with various nucleophiles (amines, alkoxides) [19,20]:

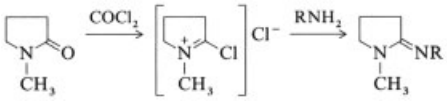

It can be used as a catalyst for the synthesis of carboxylic acid chlorides from carboxylic acids [19].

With strong bases (e.g., lithium diisopropylamide) NMP can be deprotonated in the 3-position to form amide enolates which can react with alkyl halides [21] and aryl bromides [22]. Condensations with esters at the CH_2 group in the 3-position are achieved in the presence of sodium ethoxide [23]. This route is also used in the Späth nicotine synthesis [24]. Numerous catalytic effects can be obtained with NMP as the reaction medium. For example, certain acetylides can only be alkylated in NMP [25]; ethylene glycol can only be produced from synthesis gas on rhodium carbonyls when NMP is used as the reaction medium [26].

Production. Large-scale production of NMP is predominantly carried out by reacting γ-butyrolactone with methylamine in a shaft reactor (high-pressure tube with special baffles) at 200–350 °C and ca. 10 MPa.

Other processes also analogous to those used for pyrrolidone synthesis can also be used, in particular hydrogenation of *N*-methylsuccinimide or mixtures of maleic or succinic anhydride and methylamine [27]. NMP can also be produced by hydrogenation of *N*-hydroxymethyl-2-pyrrolidone (see page 2) [28] or by reaction of acrylonitrile with methylamine in the presence of a peroxide radical initiator [29].

Quality Specifications and Analysis. Technical-grade NMP has the following specifications (gas chromatography): NMP content, min. 99.5 wt %; methylamine, max. 0.02 wt %; water, max. 0.1 wt % (Karl Fischer titration); color index, APHA 50 max. For special areas of use (e.g., electronics) higher purities are supplied with the following specifications: methylamine, max. 0.01 wt %; water, max. 0.05 wt %.

Handling, Storage, and Transportation. NMP is transported in tank cars or trailers and in drums. Transport and storage containers are generally made of mild steel. Stainless steel, nickel, and aluminum are also suitable as container and drum materials. Small quantities can be stored in polyethylene, polypropylene, or clear glass bottles. NMP shows unlimited shelf life in tightly closed containers, slight discoloration does not impair its quality. It is flammable.

Uses. NMP is an important solvent because of its low volatility, thermal stability, high polarity, and aprotic, noncorrosive properties. Its favorable toxicological and ecological properties account for the fact that NMP is replacing other solvents such as chlorinated hydrocarbons.

The most important areas of use are:

1) *Petrochemical processing*: acetylene recovery from cracked gas, extraction of aromatics and butadiene, gas purification (removal of CO_2 and H_2S), lube oil extraction
2) *Engineering plastics*: reaction medium for the production of high-temperature polymers such as polyethersulfones, polyamideimides, and polyaramids
3) *Coatings*: solvent for acrylic and epoxy resins, polyurethane paints, waterborne paints or finishes, printing inks, synthesis/diluent of wire enamels, coalescing agent
4) *Agricultural chemicals*: solvent and/or cosolvent for liquid formulations
5) *Electronics*: cleaning agent for silicon wafers, photoresist stripper, auxiliary in printed circuit board technology
6) *Industrial and domestic cleaning*: component in paint strippers and degreasers e.g., removal of oil, fat, and soot from metal surfaces, carbon deposits and other tarry polymeric residues in combustion engines

Economic Aspects. World production of NMP is currently estimated at 20 000–30 000 t/a. The most important producers are BASF and GAF.

Environmental Aspects. NMP is nontoxic to most aquatic life and shows good biodegradability [16,30]. The most important data are:

BOD_5 1100 mg/g NMP (DIN 38 409 part 51), COD 1600 mg/g NMP (DIN 38 409 part 41), TOC (total organic carbon) 600 mg/g NMP (DIN 38 409 part 3), biodegradability: > 90 % DOC (dissolved organic carbon) degradation (Zahn-Wellens static test, DIN 38 412 part 25).

4. *N*-Vinyl-2-pyrrolidone

Physical Properties. *N*-Vinyl-2-pyrrolidone (NVP) [*88-12-0*], *N*-vinylpyrrolidone, 1-vinyl-2-pyrrolidone, 1-ethenyl-2-pyrrolidinone, C_6H_9NO, M_r 114.14, is a colorless to yellowish liquid when freshly distilled, with a characteristic odor, *mp* 13.5 °C, *bp* 90 – 92 °C (1.3 kPa). It is completely miscible with water and most organic solvents, but only partially miscible with aliphatic hydrocarbons. Further physical constants follow:

Density (20 °C)	1.043 g/cm^3
Refractive index n_D^{20}	1.514
Viscosity (20 °C)	2.4 mPa · s
Vapor pressure (20 °C)	0.012 kPa
Flash point (DIN 51 758)	95 °C
Ignition temperature (DIN 51 794)	240 °C
Lower explosion limit	1.4 vol %
Upper explosion limit	10 vol %

Chemical Properties. NVP is stable toward alkalis at room temperature. Above 0 °C it is cleaved by aqueous mineral acids into acetaldehyde and 2-pyrrolidone; the latter reacts with an excess of NVP to give 1,1′-ethylidene-bis-2-pyrrolidone [31]. Protic compounds such as amides, thiols, alcohols, and phenols add to the double bond according to the Markovnikov rule, for example, *N*-(1-phenoxyethyl)-2-pyrrolidone is formed with phenol [32, 33]. Catalytic hydrogenation produces *N*-ethyl-2-pyrrolidone. Hydroformylation on rhodium catalysts forms 2-*N*-(2-pyrrolidonyl)propanal as the major product [34]. On prolonged standing, particularly in warm conditions, NVP tends to polymerize. Industrially, NVP is converted into polyvinylpyrrolidone using radical initiators.

Production. Industrially, NVP is produced by reacting 2-pyrrolidone with acetylene in high-pressure autoclaves at 130 – 160 °C and pressures of up to 2.6 MPa [32, 33, 35]. Vinylation proceeds in the liquid phase and is catalyzed by 2-pyrrolidone – potassium, which is obtained by dissolving 1 – 5 % potassium hydroxide (the water formed is removed by distillation in vacuo) or potassium in 2-pyrrolidone. To avoid the danger of explosions acetylene – nitrogen mixtures are used instead of acetylene. The crude product obtained by continuous or batch production is purified by vacuum distillation. Leading producers are BASF and GAF.

Other methods such as the elimination of *N*-(1-hydroxyethyl)- (see page 2) and *N*-(1-alkoxyethyl)-2-pyrrolidones are not industrially important [6, 7, 33].

Quality Specifications and Analysis. Composition is determined by gas chromatography and water is titrated with the Karl Fischer reagent. A typical commercial product has the following composition: NVP content, min. 99 wt %; 2-pyrrolidone, max. 0.2 wt %; water, max. 0.1 wt %; color index (freshly filled, APHA), 40 max.; stabilizers: 10 ppm *N,N′*-bis(1-methylpropyl)-1,4-benzenediamine or 0.1 wt % sodium hydroxide.

Handling, Storage, and Transportation. NVP must be protected from heat and direct sunlight. Even in closed containers NVP can only be stored for a limited period because it tends to polymerize. Stabilizers are therefore added. *N,N′*-Bis(1-methylpropyl)-1,4-benzenediamine [*101-96-2*] is an important, NVP-soluble stabilizer that does not interfere with further processing. Use of this stabilizer allows the product to be stored for ca. 300 d at 20 °C, and ca. 150 d at 25 °C. Stabilization with solid sodium hydroxide that can be easily filtered off, is not as efficient. Storage at temperatures close to the melting point results in separation of NVP from the stabilizer due to repeated solidification and melting. This shortens the shelf life considerably and can lead to uncontrolled polymerization. Solidified NVP must therefore be melted carefully in a water bath (max. 40 °C) or at room temperature (max. 30 °C) and homogenized by continuous agitation. NVP is flammable.

Uses. NVP is an important precursor and intermediate for process auxiliaries and additives. The main areas of use are the production of polyvinylpyrrolidone and copolymers (→ Polyvinyl Compounds, Others, Chap. 3.1.), the preferred comonomers being vinyl acetate and methyl acrylate. Ca. 10 – 15 % of the monomer is used in the pharmaceutical industry for the production of a polyvinylpyrrolidone – iodine complex used as a disinfectant. NVP is also used as a reactive solvent for UV-curable resins for the production of printing inks and paints as paper and textile auxiliaries, and as an additive in the cosmetics industry.

5. Other Pyrrolidone Derivatives

For certain areas of application special *N*- or *C*-substituted 2-pyrrolidones have been developed, whose physiochemical properties are especially adapted for a specific use. For example, *N*-ethyl-2-pyrrolidone [*2687-91-4*] is used as a reaction medium for low- and high-temperature reactions; *N*-(2-hydroxyethyl)-2-pyrrolidone [*3445-11-2*] is used in the electronics industry, *N*-octyl- [*2687-94-7*] and *N*-dodecyl-2-pyrrolidone [*2687-96-9*] serve as formulation auxiliaries for triazole fungicides [36]. *N*-(2-Aminoethyl)-2-pyrrolidone [*24935-08-8*] derivatives are used as solubilizing and complexing aids [37]. A range of nootropic pharmaceuticals (e.g., piracetam [*7491-74-9*] **1**) are *N*-substituted 2-pyrrolidones [38]. *N*-1-(3-Trifluoromethylphenyl)-2-pyrrolidone derivatives (e.g., fluorochloridone [*61213-25-0*] **2**) are used as herbicides [39].

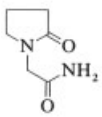

1

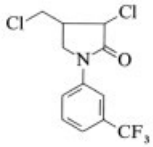

2

6. Toxicology and Occupational Health

2-Pyrrolidone does not have a particularly high acute toxicity: LD_{50} values are > 3.2 g/kg – ca. 8.5 mL/kg (rat, oral); 3.2 g/kg – 5.8 mL/kg (mouse, oral); 6.5 mL/kg (guinea pig, oral); 3 g/kg (rat, i.p.); 3.7 g/kg (mouse, i.p.); 3 g/kg (mouse, i.v.) [40–46]. In rabbits it is only mildly irritating to intact skin, if at all, and not irritating to the eyes [41,42]; in another study clouding of the cornea has, however, been reported [43]. In guinea pigs the compound causes severe skin irritation but no sensitization [40]. It can be absorbed through the skin and causes death in rats (ca. 2 mL/kg) [41]. Inhalation (8 h) of air saturated with 2-pyrrolidone at 30 °C produced no symptoms in rats [41]. In a 13-d study a slight increase in the mean relative kidney weight was observed in rats fed with 10 g/kg [40]. 2-Pyrrolidone did not produce mutagenicity in the Ames test (± S-9 mix) [41] but induced aneuploidy in *Saccharomyces cerevisiae* [46]. It is nonteratogenic in mice (oral, i.p.) and rats (oral) [40,41].

2-Pyrrolidone has been identified as an endogenous component in the blood plasma of rats, mice, and humans; it is metabolized to 4-aminobutanoic acid and other compounds (e.g., succinimide) [47].

2-Pyrrolidone possesses psychotropic properties resembling those of tranquilizing drugs in mice, rabbits, rats, cats, and pigeons. It is said to have irritating properties to human skin [40].

***N*-Methylpyrrolidone** is only slightly toxic if ingested or inhaled. LD_{50} values are 3.5 mL/kg –4.3 g/kg (rat, oral); 7.5 mL/kg (mouse, oral); 3.5 g/kg (rabbit, oral); 2.2 mL/kg (rat, i.v.); 2.4 mL/kg (rat, i.p.); 1.9 mL/kg (mouse, i.p.), > 2.5 g/kg (rat, percutaneous); 4 – 8 g/kg (rabbit, percutaneous). Inhalation (6 – 8 h) of a highly enriched or saturated atmosphere at 20 or 50 °C produced no symptoms in cats, mice, rats, rabbits, or guinea pigs. The LC_{50} value is > 5.1 mg/L (rat, inhalation, aerosol, 4 h) [16,48].

In rabbits, NMP produced slight skin irritation (method according to Draize); undiluted it caused moderate reddening and scaling after short exposure (5 – 15 min), 20-h exposure led to additional edema. In the rabbit eye it caused conjunctivities, redness, and reversible clouding of the cornea. NMP is not a skin sensitizer in guinea pigs [16,48].

Rabbits administered NMP by gavage in doses of 0.4 mL kg^{-1} d^{-1} (5 times/week for

5 weeks) showed no adverse effects, yet death occurred at 1 and 2 mL kg^{-1} d^{-1}. Cats were less sensitive. In rats, 0.25 mL kg^{-1} d^{-1} (4 weeks) caused no effects; 0.5 and 1 mL kg^{-1} d^{-1} led to slightly retarded body weight gain in males. A slight decrease in leucocytes was observed at 1 mL kg^{-1} d^{-1}; 2 mL kg^{-1} d^{-1} retarded body weight gain and led to a decrease in leucocytes and testicular damage. In a 90-d study on rats, 2.5 g/kg feed caused no changes; 5 g/kg increased thyroid gland weight. Dermal application (20 d, 0.4 and 0.8 mL kg^{-1} d^{-1}) to intact or abraded skin of rabbits produced slight local irritation.

Inhalation (17 exposures, 6 h/d, 5 d/week) of a saturated or highly enriched atmosphere produced no symptoms in cats, rabbits, guinea pigs, or rats, but caused death in mice. In a 6-week study (6 h/d, 5 d/week) air saturated at 25 °C caused no symptoms in rats. In a 4-week inhalation study (whole body exposure, 6 h/d, 5 d/week), 0.1 and 0.5 mg/L were tolerated whereas 1 mg/L caused lethality through additional dermal and oral absorption. In another study (head – nose exposure), no symptoms could be detected (0.9 mg/L, 6 h/d, 10 d) [16,48].

NMP was found to be nonteratogenic in rats (oral, dermal, and inhalative) and mice (i.p., oral) in doses that were not maternally toxic[16, 48]. In several mutagenicity tests NMP did not demonstrate a genotoxic potential. It induced aneuploidy in *Saccharomyces cerevisiae*, but neither spindle disfunction nor chromosome damage was found in mice [16]. NMP was not carcinogenic in a 2-year inhalation study on rats (0.04 and 0.4 mg/L, 6 h/d, 5 d/week) [16].

NMP is rapidly absorbed and is mainly excreted in the urine within 24 h (rats, i.v.), excretory half-life 2.3 h. It is metabolized into three compounds, mainly 4-methylaminobutanoic acid [49].

The MAK value is 100 ppm (8 h) [48]. Additional information on toxicity is given in [16, 48].

***N*-Vinyl-2-pyrrolidone** exhibits moderate acute toxicity if ingested, inhaled, or administered intraperitoneally or intravenously. The LD_{50} values are > 0.8 – 2.5 g/kg (rat, oral); 0.9 g/kg (mouse, oral); ca. 0.3 – 0.47 g/kg (mouse, i.v.); > 0.25 – 0.605 g/kg (mouse, i.p.); 1 – 4 mL/kg (rat, dermal). The LC_{50} is > 3.7 mg/L (rat, cat, rabbit, mouse, guinea pig, inhalation, 6 h) [41,50]. NVP induces narcotic symptoms, irritation of the skin and eye, and damage to the liver, spleen, and kidneys. In rabbits, it is a mild irritant to intact skin; the undiluted compound causes redness, edema, and clouding of the cornea that lasts for more than 8 days. Inhalation (8 h) of air saturated with NVP at room temperature produced slight irritation of the mucosa. Repeated oral administration of NVP (drinking water or by gavage) caused liver damage in rats. Signs of liver damage were recorded for rabbits and cats. Intraperitoneal administration (5 d, 0.1, 0.25 mL/kg) caused death in mice but not in rats [41,50].

Inhalation of NVP vapor (up to 15 days, 6 h/d, 180 ppm) caused death of mice, cats, and rabbits but not of rats and guinea pigs. Damage to the liver was seen in mice and cats. Inhalative exposure of rats to the aerosol (0.3 – 0.75 mL/kg, 4 h/d, 5 d/week for 4 weeks) or vapor (1 – 120 ppm, 6 h/d, 5 d/week for 3 months) caused damage to the nasal mucous membranes and adenoids above 0.75 mL/L (5 ppm) and foci and areas of cellular alterations in the liver above 15 ppm. 120 ppm caused premature death. In the Syrian gold hamster (45 ppm, 6 h/d, 5 d/week for 7 weeks to 3 months) symptoms of irritation were seen in the upper respiratory tract [41, 50].

Slight signs of liver toxicity but no carcinogenic impairment of the liver was noted. In several tests NVP did not show mutagenic or clastogenic activity. In a 2-year inhalation study in rats (5, 10, and 20 ppm, 6 h/d, 5 d/week) carcinomas of the liver, adenomas, and adenocarcinomas of the nasal cavity were observed. Squamatous carcinomas of the larynx were reported at 20 ppm [41,50].

A single intravenous dose of ^{14}C-NVP is quickly metabolized and eliminated (mainly in the urine) in the rat. The plasma half-life is 1.9 h [51].

Epidemiological studies in humans revealed no adverse effects [52]. An MAK value cannot be defined on account of the carcinogenic potential. Additional information on toxicity is given in [50].

7. References

1. BASF, *Pyrrolidon,* technical bulletin, Ludwigshafen 1987.
2. J. Tafel, M. Stern, *Ber. Dtsch. Chem. Ges.* **33** (1900) 2224 – 2231.
3. BASF, DE 830 194, 1950 (C. Schuster).
4. Mitsubishi Petrochem., JP-Kokai 74 117 459, 1973 (T. Ayusawa, S. Fukami).
5. W. Reppe, *Justus Liebigs Ann. Chem.* **596** (1955) 210.
6. J. Falbe, H.-J. Schulze-Steinen, *Brennst. Chem.* **48** (1967) 136 – 139.
7. Mitsubishi Chemical, DE-OS 3 544 134, 1985; US 4 837 337, 1988 (Y. Murao, M. Miyake).
8. K. Dachs, E. Schwartz, *Angew. Chem.* **74** (1962) 540 – 545.
9. Du Pont, US 2 187 745, 1934 (W. A. Lazier).
10. J. Falbe, F. Korte, *Chem. Ber.* **98** (1965) 1928 – 1937. J. F. Knifton, *J. Organomet. Chem.* **188** (1980) 223 – 236.
11. Chevron, US 4 181 662, 1978 (W. A. Sweeney). Standard Oil, US 4 036 836, 1975 (J. L. Greene).
12. Phillips Petroleum, US 4 800 227, 1984 (M. S. Matson).
13. E. Späth, J. Lintner, *Ber. Dtsch. Chem. Ges.* **69** (1936) 2727 – 2731. BASF, DE-AS 1 765 007, 1968 (W. Himmele, E. Hofmann, H. Hoffmann, R. Plaß).
14. GAF, US 4 824 967, 1988 (K.-C. Liu, P. D. Taylor).
15. *Inf. Chim.* **239** (1983) 65.
16. BASF: *N-Methylpyrrolidone,* product brochure, Ludwigshafen 1990.
17. GAF, *N-Methyl-2-pyrrolidone, Properties and Chemical Reactions,* 1972.
18. S. Yoshifuji, Y. Arakawa, Y. Nitta, *Chem. Pharm. Bull.* **35** (1987) 357 – 363.
19. H. Eilingsfeld, M. Seefelder, H. Weidinger, *Angew. Chem.* **72** (1960) 836 – 845; *Chem. Ber.* **96** (1963) 2671 – 2690.
20. General Electric, US 4 353 830, 1979 (V. Mark). T. Jen et al., *J. Med. Chem.* **18** (1975) 90 – 99. McNeil, DE-OS 1 770 752, 1968 (G. I. Poos).
21. P. Hullot, T. Cuvigny, M. Larchevêque, H. Normant, *Can. J. Chem.* **54** (1976) 1098 – 1104.
22. J. D. Stewart, S. C. Fields, K. S. Kochhar, H. W. Pinnick, *J. Org. Chem.* **52** (1987) 2110 – 2113.
23. G. M. Ksander, J. E. McMurry, M. Johnson, *J. Org. Chem.* **42** (1977) 1180 – 1185.
24. E. Späth, H. Bretschneider, *Ber. Dtsch. Chem. Ges.* **61** (1928) 327 – 334.
25. BASF, DE 944 311, 1955 (M. Kracht, H. Pasedach).
26. E. Watanabe et al., *J. Chem. Soc. Chem. Commun.* 1986, 227 – 228.
27. Mitsubishi Chemical, US 4 731 454, 1986 (M. Otake, J. Fukushima). Amoco, US 4 814 464, 1987 (R. J. Olson). Hüls AG, EP-A 252 242, 1986 (H. M. Zur, W. Otte). BASF, EP-A 460 474, 1990 (H.-J. Weyer, R. Fischer, W. Harder).
28. Mitsubishi Chemical, JP-Kokai 87 120 360, 1985 (M. Otake, I. Fukushima, K. Fujita). Stamicarbon, EP-A 37 603, 1980 (P. J. N. Meyer, J. M. Penders).
29. BASF, DE-OS 3 701 297, 1987 (F. Merger, J. Liebe, W. Harder).
30. BASF, *N-Methylpyrrolidone, Biodegradability,* company brochure, Ludwigshafen 1991.
31. E. Senogles, R. A. Thomas, *J. Chem. Soc. Perkin Trans. II* 1980, 825 – 828.
32. W. Reppe, *Justus Liebigs Ann. Chem.* **601** (1956) 135 – 138. I. G. Farbenind., Chem. Zentralbl. 1941, II, 2735.
33. S. A. Miller: *Acetylene. Its Properties, Manufacture and Uses,* **vol. 2,** Ernest Benn, London 1966, pp. 338 –341.
34. Texaco, US 4 831 159, 1986 (J.-L. Lin).
35. GAF, WO 89 09210, 1988 (K. C. Liu, P. D. Taylor); DE-OS 3 215 093, 1982 (R. Parthasarathy, E. V. Hort, P. M. Chakrabarti).
36. Bayer, DE 3 910 921, 1989 (K. Reizlein et al.); EP-A 453 915, 1990 (H. O. Horstmann, K. Wangermann).
37. GAF, US 4 918 198, US 4 924 006, 1989 (L. R. Anderson, M. M. Hashem, R. B. Login).
38. E. R. Gamzu, T. M. Hoover, S. I. Gracon, M. V. Nintemann, *Drug Dev. Res.* **18** (1989) 177 – 189.
39. Stauffer, US 4 110 105, 1976 (E. G. Tech).
40. CRCS Inc., working draft, EPA contract no. 68-01-650, TSCA Interagency Testing Committee, July 15, 1985.
41. BASF AG, unpublished results (1953 – 1990).
42. W. B. Deichmann, H. W. Gerade (eds.): *Toxicology of Drugs and Chemicals,* 4th ed., Academic Press, New York 1969, pp. 508.
43. V. V. Aleshin, *Med. Probl. Okhr. Vneshn. Stredy* 1974, 4; *Chem. Abstr.* **86** 26605b.
44. National Institute for Occupational Safety and Health, RTECS no. 20 387, Jan. 1987.
45. G. H. Cocolas et al., *J. Pharm. Sci.* **72** (1983) 812 –814.
46. V. W. Mayer, C. J. Goin, R. E. Taylor-Mayer, *Environ. Mol. Mutagen.* **11** (1988) 31 – 40.

47. E. F. Bandle et al., *Life Sci.* **35** (1984) 2205 – 2212.
48. Deutsche Forschungsgemeinschaft: "N-Methylpyrrolidon," *Gesundheitsschädliche Arbeitsstoffe,* VCH Verlagsgesellschaft, Weinheim 1978, 1988.
49. D. A. Wells, G. A. Digenis, *Drug. Metab. Dispos.* **16** (1988) 243 – 249.
50. Deutsche Forschungsgemeinschaft: "N-Vinylpyrrolidon," *Gesundheitsschädliche Arbeitsstoffe,* VCH Verlagsgesellschaft, Weinheim 1991.
51. G. A. Digenis, J. S. McClanahan, *The Toxicologist* **2** (1982) 165 – 166. J. S. McClanahan et al., *Drug. Chem. Toxicol.* **7** (1984) 129 – 148.
52. A. Zober et al., to be published in *Proceedings of XIXth International Conference on Occupational Health in Chemical Industry,* Basel (Switzerland), Sept. 17 – 20, 1991, WHO, Copenhagen 1992.

Pyruvic Acid → **Oxocarboxylic Acids**
Quartz Glass → **Glass**

Quinoline and Isoquinoline

GERD COLLIN, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany (Chaps. 1, 2)

HARTMUT HÖKE, Rütgerswerke AG, Frankfurt/Main, Federal Republic of Germany (Chap. 3)

1. Quinoline 1
1.1. Properties 1
1.2. Production 1
1.3. Uses and Economic Aspects 2
1.4. Derivatives 2
2. Isoquinoline 3
2.1. Properties 3
2.2. Production 3
2.3. Uses 3
3. Toxicology 3
4. References 4

1. Quinoline

Quinoline was discovered in coal tar by F. F. RUNGE in 1834.

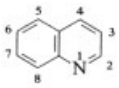

Quinoline

1.1. Properties

Quinoline [*91-22-5*], C_9H_7N, M_r 129.16, *mp* − 15.6 °C, *bp* 238 °C (101.3 kPa), d^{20} 1.0929, n_D^{20} 1.6268, specific heat (17 °C) 1.516 kJ/kg, heat of vaporization 367.6 kJ/kg, heat of combustion (25 °C) 36.4 MJ/kg, is a colorless liquid with characteristic odor. It is sparingly soluble in cold water, readily soluble in hot water, miscible with many organic solvents, and volatile in steam.

Quinoline is a weak tertiary base. With strong acids, it forms readily crystallizable salts; with alkyl halides and dialkyl sulfates, it forms quaternary addition compounds. Halogenation gives preferably the 3-substituted product; nitration, the 5- and 8-substituted products; and sulfonation, the 6- or 8-substituted products, depending on reaction conditions. Quinoline is catalytically hydrogenated to 1,2,3,4- or 5,6,7,8-tetrahydroquinoline, and further to decahydroquinoline. Oxidation leads to quinoline *N*- oxide, 2-quinolinol or 2-quinolinone, quinolinic acid (pyridine-2,3-dicarboxylic acid), or nicotinic acid, depending on the oxidizing agent used and the reaction conditions. Biochemical oxidation yields 2-oxoquinoline as the primary metabolite, which is then hydroxylated in the 6- or 8-position [5]. Reaction with sodium amide or ammonia in the presence of a Raney nickel catalyst gives 2-aminoquinoline.

1.2. Production

High-temperature coal tar contains an average of 0.3 % quinoline. Quinoline, together with isoquinoline and 2-methylquinoline, is recovered from the methylnaphthalene fraction of coal tar by extraction with sulfuric acid, followed by precipitation with ammonia. Then quinoline is separated from isoquinoline (*bp* 6 °C higher) by rectification of the crude base mixture (quinoline bases). It can be purified further by several processes, for example, selective oxidation, treatment with alkali, resinification of the byproducts with formaldehyde, or formation of the hydrate (in contrast to isoquinoline) [6]. Quinoline can also be obtained from the methylnaphthalene fraction by azeotropic distillation with ethylene glycol or diethylene glycol, and subsequent distillation [7].

Quinoline can be synthesized via the Skraup method by heating aniline and glycerol with sulfuric acid in the presence of a dehydrogenating agent (→ Amines, Aromatic, Chap. 2.2.4.), catalytic gas-phase reaction of aniline and acetaldehyde with formaldehyde – methanol [8], dehydrogenative cyclization of *N*-propylaniline in the gas phase [9], or pyrolysis of *N,N′*-diphenyl-1,1-propanediamine [10]. However, these syntheses are not yet industrially important because sufficient quantities of quinoline can be recovered from coal tar.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_465

1.3. Uses and Economic Aspects

World production of quinoline is more than 2000 t/a. The main application of quinoline is the production of 8-quinolinol, which is obtained by alkali fusion of quinoline-8-sulfonic acid [*85-48-3*]. In addition, quinoline is used in the production of pyridine-2,3-dicarboxylic acid [*89-00-9*] (quinolinic acid) by oxidation [11]. Quinolinic acid is used to manufacture the herbicide Assert [*81405-85-8*] (imazapyr) [12]. Quinoline can be used to produce methine dyes (cyanines) (→ Methine Dyes and Pigments, Chap. 7.4.2.) and nicotinic acid [*59-67-6*] (via quinolinic acid).

Hydroquinolines can be employed as intermediates for pharmaceuticals such as antibiotics [13]. From quinoline, 4,5-dihydroimidazo[1,2-*a*]- quinoline-1-acetamides can be prepared as anticonvulsants, sedatives, and anxiolytics [14]. Quinoline can act as catalyst in the preparation of pharmaceuticals such as pyridazinone antihypertensives [15]. Fungicides can be obtained from 2-anilinoquinolines (from 2-aminoquinoline) [16]. Derivatives of 2-quinolinone [*59-31-4*] (carbostyril, prepared by oxidation of quinoline with hypochlorous acid, or selective, enzymatic hydroxylation of quinoline [17]) can be used as pharmaceuticals (e.g., cardiotonic drugs or antihistaminics) [18]. Quinoline alone, or as a mixture with isoquinoline and quinoline homologues, is an excellent solvent and extractant, especially for polycyclic aromatic compounds. Quinoline bases serve as acid-binding agents, corrosion inhibitors, or pickling inhibitors.

1.4. Derivatives

2-Methylquinoline [*91-63-4*], quinaldine, $C_{10}H_9N$, M_r 143.19, *bp* 247.6 °C (101.3 kPa), *mp* −2 °C, d^{20} 1.0582, n_D^{20} 1.6116, is a colorless liq-uid with a weak quinoline-like odor. It is soluble in organic solvents and very slightly soluble in water.

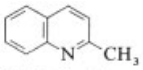

2-Methylquinoline

High-temperature coal tar contains an average of 0.2 % quinaldine. It is recovered from the quinoline base mixture by rectification and hydration. Quinaldine can be synthesized by the Skraup method from aniline and crotonaldehyde with sulfuric acid in the presence of a dehydrogenating agent.

Quinaldine is used for the synthesis of several dyes such as the trimethinecyanine dye pinacyanol [*605-91-4*] and the corresponding vinylogous dye (→ Methine Dyes and Pigments, Chap. 7.2.2.→ Methine Dyes and Pigments, Chap. 7.4.), the quinophthalone dye quinoline yellow [*8004-92-0*], and the indicator dye quinaldine red [*117-92-0*].

4-Methylquinoline [*491-35-0*], lepidine, $C_{10}H_9N$, M_r 143.19, *bp* 265.6 °C (101.3 kPa), *mp* 8 °C, d^{20} 1.0868, n_D^{20} 1.6200, is a colorless liquid with a weak quinoline-like odor. It is slightly soluble in water, miscible with many organic solvents, and volatile in steam.

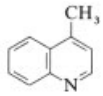

4-Methylquinoline

4-Methylquinoline can be recovered from the highest-boiling fraction of the coal-tar quinoline bases via its *o*- cresol addition compound. It can be synthesized by condensation of aniline and methyl vinyl ketone [19].

4-Methylquinoline is used for the synthesis of several cyanine dyes (→ Methine Dyes and Pigments, Chap. 7.2.2.,→ Methine Dyes and Pigments, Chap. 7.4.4.).

8-Quinolinol [*148-24-3*], 8-hydroxyquinoline, oxine, C_9H_7ON, M_r 145.15, *bp* 266.6 °C (100.2 kPa), *mp* 75.8 °C, forms colorless crystals with a phenolic odor. It is readily soluble in ethanol, benzene, and chloroform; soluble in acetone, alkali, and acid; slightly soluble in diethyl ether; and very slightly soluble in water.

8-Quinolinol

8-Quinolinol is produced by alkali fusion of quinoline-8-sulfonic acid or synthesized via the

Skraup method by heating *o*-aminophenol (via benzoxazolone [20]) with *o*-nitrophenol, glycerol, and sulfuric acid.

8-Quinolinol and its salts with bivalent metals or mineral acids are used as fungicidal, bactericidal, insecticidal, and algicidal agents in crop protection products and disinfectants, as well as in preservatives for leather, textiles, plastics, coatings, paper pulp, and seeds. Several halogenated derivatives of 8-quinolinol show, in addition to a general antiseptic activity (→ Dermatotherapeutic Agents, Chap. 3.4.), a specific activity against amoebas (→ Chemotherapeutics, Chap. 3.5.2.). In azo dyes 8-quinolinol can serve as a heterocyclic coupling component. 8-Quinolinol, its dibrominated derivative Bromoxine [*521-74-4*], and the thiol analogue Thiooxine [*491-33-8*] are analytical reagents for the detection of numerous metals (→ Indicator Reagents).

2. Isoquinoline

Isoquinoline was discovered in coal tar by S. Hoogewerff and W. A. van Dorp in 1885.

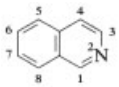

Isoquinoline

2.1. Properties

Isoquinoline [*119-65-3*], C_9H_7N, M_r 129.16, *mp* 26.5 °C, *bp* 243.25 °C (101.3 kPa), d^{20} 1.0986, n_D^{20} 1.6148, specific heat 1.507 kJ/kg, heat of fusion 57.8 kJ/kg, heat of vaporization 379.3 kJ/kg, heat of combustion (25 °C) 36.45 MJ/kg, is a colorless, hygroscopic liquid that forms crystals on solidifying, with an odor resembling that of benzaldehyde. It is very slightly soluble in water, but miscible with many organic solvents.

Isoquinoline is a stronger base than quinoline. It forms readily crystallizable salts and quaternary addition compounds. Catalytic hydrogenation leads to 1,2,3,4- or 5,6,7,8-tetrahydroisoquinoline and further to decahydroisoquinoline. Oxidation in the liquid or vapor phase gives a mixture of nicotinic and isonicotinic acids (via pyridine-3,4-dicarboxylic acid by decarboxylation), whereas oxidation with peroxyacids gives isoquinoline *N*-oxide.

2.2. Production

High-temperature coal tar contains an average of 0.2 % isoquinoline. It is separated by distillation from the lower-boiling quinoline and the higher-boiling 2-methylquinoline of the quinoline base mixture. Further refining is based on the fact that isoquinoline, in contrast to quinoline and 2-methylquinoline, cannot be hydrated but can be crystallized at low temperature.

Isoquinoline can by synthesized, for example, via the Bischler – Napieralski reaction by cyclodehydration of *N*-acyl derivatives of β-phenylethylamine with Lewis acids and subsequent dehydrogenation.

2.3. Uses

Isoquinoline is used predominantly in the production of pharmaceuticals such as the anthelmintic praziquantel [*55268-74-1*] (→ Anthelmintics). Isoquinoline and hydroisoquinolines can also be used as intermediates for fungicides [21] and dyes (e.g., isoquinoline red [*6359-40-6*]).

3. Toxicology

The oral LD_{50} of quinoline in rats is 270 – 460 mg/kg [22], [23]; the dermal LD_{50} is 1400 mg/kg in rats [24], and ca. 600 mg/kg in rabbits [23]. Inhalation of air saturated with quinoline for 7 h caused death of almost all rats [25]. Acute symptoms were unspecific: the rats exposed showed piloerection, ataxia, and lack of motility. Dissection revealed hemorrhaging and cellular edema in the gastrointestinal tract, liver, and lung. The 24-h exposure of the skin of rabbits caused slight to moderate, but reversible, edema and erythema; after exposure of the eyes of rabbits, slight to moderate, but reversible, irritation was found [26]. After resorption in mammals, quinoline is excreted rapidly after hydroxylation and conjugation [27], [28].

Long-term feeding (0.05 – 0.25 % in the diet for 20 – 40 weeks) led to hepatocellular carcinomas and hemangiosarcomas in rats and mice [29], [30], but not in hamsters and guinea pigs [30]. When low doses (1.75 µmol per animal) were injected intraperitoneally into newborn mice on the first, eighth, and fifteenth day after birth, liver tumors were found 52 weeks after treatment [31]. Quinoline showed tumor-initiating activity on mouse skin at an effective total dose of > 2 mg per animal [32], which is ca. 250 – 300 times higher than the dose of benzo[*a*]pyrene required to obtain the same tumor incidence. Quinoline has toxic effects on aquatic life, with fish apparently being more sensitive than the aquatic invertebrates and algae tested; e.g., 5 mg/L was lethal to trout after 14 h and bluegill sunfish after 4 h [33], whereas toxic effects on daphnia were noticeable only above 50 mg/L [34]. The compound is readily degraded under natural conditions.

The oral LD_{50} of 2-methylquinoline (quinaldine) is 1230 mg/kg in rats [23], and the LD_{50} (i.p.) of 4-methylquinoline (lepidine) is 270 mg/kg in rats [35]. 4-Methylquinoline was found to be a tumor initiator on mouse skin, whereas 2-methylquinoline was not [32]. 4-Methylquinoline is tumorigenic when injected into newborn mice [31]. 2-Methylquinoline showed very weak mutagenicity with *Salmonella typhimurium* TA 100 [36], whereas the number of his-revertants was greater for 4-methylquinoline than for quinoline [37].

For 8-quinolinol an oral LD_{50} of 1200 mg/kg in rats and an i.p. LD_{50} of 48 mg/kg in mice are reported [38]. After i.p. injection of 100 mg/kg into mice, many died after several hours, whereas the same dose of quinoline caused no deaths [39]. In plant chromosomal studies, 8-quinolinol was found to act as a spindle inhibitor [40]. In two rodent studies by the U.S. National Toxicological Program, 8-quinolinol was classified as noncarcinogenic [40], [41] but was found to be mutagenic in the Ames test [42], [43], genotoxic in several other in vitro assays, and positive in the micronucleus assay [39].

The oral LD_{50} of isoquinoline is 360 mg/kg [23]. Isoquinoline shows no tumor-initiating activity [44] and proved to be nonmutagenic in the Ames test [37].

4. References

General References

1. *Beilstein,* Quinoline **20** 339; **20** (1) 134; **20** (2) 222; **20** (3/4) 3334; 2-Methylquinoline **20** 387; **20** (1) 148; **20** (2) 238; **20** (3/4) 3454; 4-Methylquinoline **20** 395; **20** (1) 150; **20** (2) 244; **20** (3/4) 3477; 8-Quinolinol **21** 91; **21** (1) 221; **21** (2) 55; **21** (3/4) 1135; Isoquinoline **20** 380; **20** (1) 143; **20** (2) 236; **20** (3/4) 3410.
2. H.-G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Berlin 1968, pp. 13, 23, 26, 62 – 64, 83, 97, 151, 158, 160 – 162.
3. H.-G. Franck, J. W. Stadelhofer: *Industrial Aromatic Chemistry,* Springer Verlag, Berlin 1988, pp. 274, 419 –423.
4. *Kirk-Othmer,* 3rd ed., **19,** 532 – 572.

Specific References

5. O. P. Shukla, *Biol. Mem.* **13** (1987) 115 – 131. G. Schwarz et al., *Hoppe Seyler Biol. Chem.* **370** (1989) 1183 – 1189.
6. R. Oberkobusch, *Brennst. Chem.* **40** (1959) 145 – 151.
7. Rütgerswerke, EP 100 109, 1983 (H. Hörmeyer). Sumikin Coke & Chemicals, JP 86 161 265, 1986 (K. Kageyama, S. Takeya, T. Nakamura). Allied, DE 3 227 492, 1982 (S. E. Belsky, C. T. Mathew).
8. Reilly Tar & Chemical, US 3 020 280-2, 1957 (F. E. Cislak, W. R. Wheeler).
9. ICI, GB 1 184 242, 1967 (W. H. Bell, R. A. C. Rennie).
10. Snam Progetti, DE 2 223 018, 1971.
11. Rütgerswerke, EP 149 857, 1984 (W. Orth, E. Pastorek, W. Fickert). Ruetgers-Nease Chemical, EP 232 118, 1987 (W. Michalowicz). Hilton-Davis Chemical, US 4 537 971, 1985 (R. W. J. Rebhahn, J. E. Kassner, R. E. Werner).
12. American Cyanamid, EP 41 623, 41 624, 1981 (P. L. Orwick, A. R. Templeton).
13. Sumitomo Chemical, JP 87 212 363, 1986 (S. Murahashi). Hoechst, DE 3 706 020, 1987 (R. Lattrell, W. Dürckheimer, R. Kirstetter).
14. Synthelabo, FR 2 593 179, 1986; EP 231 138, 1987 (P. George, D. de Peretti).
15. Diamond Shamrock, US 4 281 125, 1980 (M. F. Depompei, A. Hlynsky). Société Nationale des Poudres et Explosifs, EP 249 556, 1987 (J. P. Senet, G. Sennvey, G. Wooden).
16. BASF, DE 3 716 512, 1987 (J. Schubert et al.).
17. Rütgerswerke AG, DE 3 903 759, 1989 (F. Lingens, R. Bauder, H. Höke).

18. Otsuka Pharmaceutical, JP 82 35 588, 1980; JP 83 88 314, 1981.
19. BASF, DE 3 719 014, 1987 (H. Merkle, G. Reissenweber).
20. Rütgerswerke, DE 3 601 024, 1986 (J. Haase).
21. Nippon Steel, JP 87 294 679, 1986 (S. Watanabe et al.).
22. Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker (BUA): *Chinolin,* BUA-Stoffbericht, VCH Verlagsgesellschaft, Weinheim (to be published).
23. H. F. Smyth jr., C. P. Carpenter, C. S. Weil, *AMA Arch. Ind. Hyg. Occup. Med.* **4** (1951) 119 – 122.
24. Rütgerswerke AG: unpublished test report (1981).
25. Rütgerswerke AG: unpublished test report (1981).
26. Rütgerswerke AG: unpublished test reports (1979).
27. L. Novack, B. B. Brodie, *J. Biol. Chem.* **187** (1950) 787 – 792.
28. J. N. Smith, R. T. Williams, *Biochem. J.* **60** (1955) 284 – 290.
29. K. Hirao et al., *Cancer Res.* **36** (1976) 329 – 335.
30. Y. Shinohara et al., *Gann* **68** (1977) 785 – 796.
31. E. J. La Voie et al. in M. Cooke, A. J. Dennis (eds.): "Polynuclear Aromatic Hydrocarbons," *Proc. 10th Int. Symp. 1985,* Battelle Press, Columbus/Richland 1988, pp. 503 – 518.
32. E. J. La Voie et al., *Cancer Letters (Shannon Irel.)* **22** (1984) 269 – 273.
33. V. C. Applegate et al., *Fish Wildl. Serv. (U.S.) Res. Rep.* **207** (1957) 1 – 157.
34. K. Verschueren (ed.): Handbook of Environmental Data on Organic Chemicals, 2nd ed., Van Nostrand, New York 1983.
35. National Institute for Occupational Safety and Health (NIOSH): Registry of Toxic Effects of Chemical Substances, US Government Printing Office, Washington, D.C. 1987, p. 13841.
36. M. Dong, I. Schmeltz, E. J. La Voie, D. Hoffmann in P. W. Jones, R. I. Freudenthal (eds.): *Carcinogenesis,* **vol. 3,** Raven Press, New York 1978, 97 – 108.
37. M. Nagao et al., *Mutat. Res.* **42** (1977) 335 – 342.
38. in [35] , p. 21549.
39. M. A. Hamoud, T. Ong, M. Petersen, J. Nath, *Teratog. Carcinog. Mutagen.* **9** (1989) 111 – 118.
40. L. S. Gold et al., *EHP Environ. Health Perspect.* **74** (1987) 237 – 239.
41. J. K. Hasemann, J. E. Huff, E. Zeiger, E. E. McConnel, *EHP Environ. Health Perspect.* **74** (1987) 229 – 235.
42. J. L. Eppler et al., *Mutat. Res.* **39** (1977) 285 – 296.
43. R. Talcott, M. Hollstein, E. Wei, *Biochem. Pharmacol.* **25** (1976) 1323 – 1328.
44. M. Dong, I. Schmeltz, E. Jacobs, D. Hoffmann, *Anal. Toxicol.* **2** (1978) 21 – 55.

Quinone Diazides → Diazo Compounds and Diazo Reactions
Quinones → Benzoquinone
Quinophthalone Dyes → Methine Dyes and Pigments

Radiation Chemistry

Fundamental aspects of the interaction between radiation and matter are treated in Radionuclides, Chap. 1.; for other uses of radiation see Radionuclides, Chap. 3. and Radionuclides, Chap. 5.

Reiner Mehnert, Zentralinstitut für Isotopen- und Strahlenforschung, Leipzig, Federal Republic of Germany (Chap. 1)

Klaus W. Bögl, Bundesgesundheitsamt, Berlin, Federal Republic of Germany (Chap. 2)

Norbert Helle, Bundesgesundheitsamt, Berlin, Federal Republic of Germany (Chap. 2)

Georg A. Schreiber, Bundesgesundheitsamt, Berlin, Federal Republic of Germany (Chap. 2)

1. Radiation-Induced Polymerization 2
1.1. Polymerization by Ionizing Radiation 2
1.1.1. Initiation of Polymerization by Ionizing Radiation 2
1.1.2. Free-Radical Polymerization 3
1.1.3. Ionic Polymerization 5
1.1.4. Radiation-Induced Polymerization of Solid Monomers 6
1.2. Action of Ionizing Radiation on Polymers – Polymer Modification 7
1.2.1. Reaction Mechanisms 7
1.2.2. Cross-Linking and Degradation . . 8
1.2.3. Graft Copolymerization 9
1.3. Applications 10
1.3.1. Electron-Beam Curing of Paints, Printing Inks, Adhesives, and Release Agents 10
1.3.1.1. Principle of Radiation Curing 10
1.3.1.2. Chemistry of Electron-Beam Curing 10
1.3.1.3. Industrial Irradiation Equipment, Low-Energy Electron Accelerators 13
1.3.1.4. Industrial Applications of Electron-Beam Curing 14
1.3.2. Radiation-Induced Cross-Linking . 15
1.3.2.1. Cross-Linking of Cable and Wire Insulation 16
1.3.2.2. Vulcanization of Elastomers 17
1.3.2.3. Radiation-Cross-Linked Foam Polyethylene 17
1.3.2.4. Shrink Materials 18
1.3.3. Industrial Application of Radiation-Induced Graft Copolymerization . . 18
2. Irradiation of Foods and Feeds . . 18
2.1. Introduction 18
2.2. Capabilities of Food and Feed Irradiation 19
2.2.1. Prolongation of Shelf Life 20
2.2.2. Quarantine Measures 21
2.2.3. Control of Pathogenic Organisms . 21
2.3. History 21
2.4. Effects of Food Irradiation 22
2.4.1. Induced Radioactivity 22
2.4.2. Chemical and Sensory Changes . . 22
2.4.3. Changes in Nutritive Value 23
2.4.4. Effects on Microorganisms 23
2.4.5. Toxicological Consequences 24
2.5. Regulations on the Irradiation of Foods and Feeds 25
2.5.1. The European Community 25
2.5.2. Federal Republic of Germany 25
2.5.3. Other Countries 25
2.5.4. International Organizations 25
2.5.5. Consumers 26
2.6. Industrial Irradiation of Foods and Feeds 26
2.6.1. Irradiation Plants 26
2.6.2. Costs 26
2.6.3. Present Level of Use 26
2.7. Inspection of Irradiated Foods . . 27
3. References 27

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_471

1. Radiation-Induced Polymerization

1.1. Polymerization by Ionizing Radiation

1.1.1. Initiation of Polymerization by Ionizing Radiation

When electrons, positrons, ions, gamma quanta, or X-rays with typical energies in the keV and MeV range are absorbed in solids, liquids, or gases, secondary electrons are produced in the energy degradation process. These enter into a Coulomb interaction with the atoms or molecules of the absorbers, which results in the formation of ions, electrons, and excited atoms or molecules [23–26].

As a direct consequence of the ionization process, the first species formed are cation–electron pairs with widely varying distances between the partners. The closer together they are, the stronger the Coulomb attraction between cation and electron.

Diffusional motion in liquids prevents complete recombination of the ion pairs. In liquid alkanes, for example, only a few percent of the ions initially formed with a *G* value (i.e., the amount of transformed molecules per 100 eV of absorbed energy) of ca. 4 escape ion pair recombination [27], [28].

When polymerization processes are initiated by ionizing radiation (radiation-induced polymerization), both the action of ions in pairs and that of "free" particles are important.

Another feature of radiation-induced polymerization is the possible simultaneous production of radical ions and free radicals that can initiate polymerization. Elementary reactions preceding the initiation of radiation-induced polymerization have been studied chiefly by pulse radiolysis [25], [28], [29–31]. Some reaction mechanisms are schematically illustrated in Figure 1.

When alkyl chlorides are irradiated (Fig. 1 A), the electron produced in the primary ion pair $RCl^+ \ldots + e_s^-$ is converted to unreactive chloride ion by dissociative electron capture, while the primary radical cation RCl^+ transfers its charge to the monomer M. If the monomer concentration is sufficiently high, the formation of a dimeric cation MM^+ and thus the initiation of the ionic polymerization chain reaction can be observed.

Reactions between solvated electrons e_s^- and monomers can be investigated in ethers such as tetrahydrofuran or dioxane (Fig. 1 B). In this case the primary radical cation RH^+ is very rapidly converted to the unreactive RH_2^+. If a monomer with sufficiently high electron affinity (styrene, acrylate) is added, a radical anion is formed; it is mainly converted by protonation into a free radical $(MH)^\bullet$. The formation of a dimeric anion MH^- is less probable.

Figure 1 C summarizes elementary reactions that are induced by radiation and, in pure monomers, can lead to chain-initiating free radicals and radical ions.

Figure 1 C shows how monomeric radicals can be produced from excited molecules M∗. The triplet states M∗(T), which after hydrogen abstraction can produce radicals of the $(MH)^\bullet$ type, are particularly important.

The majority of chain-initiating radicals and ions are produced some 10^{-8} to 10^{-6} s after the primary ionization event. Chain initiation is immediately preceded by elementary reactions such as energy, charge, and proton transfer, hydrogen abstraction, and recombination. In *liquids* these reactions are diffusion- controlled. The activation energy of the elementary reaction consists only of the diffusional part. Accordingly, almost no activation energy is required to initiate radiation-induced polymerization [32–36].

Because many of the transfer processes shown in Figure 1 C can occur at low temperature even in solids and liquids, and a diffusional approach of the reactants is not required, ionizing radiation is an ideal means of initiating polymerization. Other advantages of polymerization by irradiation of solids include

1) Generation of a homogeneous dose distribution
2) No contamination of the product by catalysts, photoinitiators, and the like
3) Only minimal temperature increases, since it is ionizing radiation that is absorbed
4) The oxygen sensitivity of the polymerization process is decreased

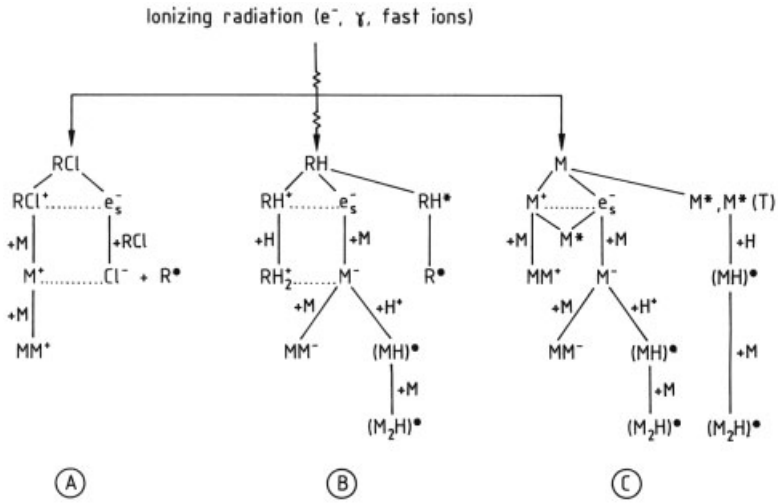

Figure 1. Radiation-induced reactions leading to the production of free radicals (radical ions) that initiate polymerization M = Monomer. M∗ = Excited molecular state of M (singlet, triplet). $RH^+ \ldots e_s^-$ = Ion pair.

1.1.2. Free-Radical Polymerization [32], [34], [37–42]

Free-radical addition polymerization, which is, for example, typical for many vinyl monomers, follows the mechanism of a classical chain reaction even when initiated by ionizing radiation. In many cases, the method of steady-state homogeneous kinetics can be applied to the description of radiation-induced polymerization. If all the elementary reactions that lead to the production of homogeneously distributed chain-initiating radicals $R^\bullet$ [= $(MH)^\bullet$ in Fig. 1] in irradiated monomers M are combined as $M \rightarrow 2\,R^\bullet$ with allowance made for chain propagation and termination, the following reaction scheme is obtained:

Initiation:

$$M \rightarrow 2\,R^\bullet \tag{1}$$

Initiation rate:

$$v_s = k_s\,[M]\,D^\bullet \tag{1a}$$

where v_s is the initiation rate, $D^\bullet$ the dose rate in Gy/s, and k_s [M] the production rate of $R^\bullet$ in mol L^{-1} Gy^{-1}.

Chain propagation:

$$R^\bullet + M \rightarrow RM^\bullet \tag{2a}$$

$$RM_n^\bullet + M \rightarrow RM_{n+1}^\bullet \tag{2b}$$

Rate of chain propagation v_p:

$$v_p = k_p\,[RM^\bullet]\,[M] \tag{2c}$$

Chain termination:

$$RM_n^\bullet + RM_m^\bullet \rightarrow P_{n+m} \tag{3a}$$

$$RM_n^\bullet + RM_m^\bullet \rightarrow P_n + P_m \tag{3b}$$

Rate of chain termination v_t:

$$v_t = k_t\,[RM^\bullet]^2 \tag{3c}$$

In the steady state,

$$v_s = v_t \tag{4a}$$

$$k_s\,[M]\,D^\bullet = k_t\,[RM^\bullet]^2 \tag{4b}$$

and the steady free-radical concentration is

$$[RM]^\bullet = \{(k_s/k_t)\,[M]\,D^\bullet\}^{1/2} \tag{5}$$

If the kinetic chain length is sufficiently large, the overall rate v of a chain reaction can be set equal to the propagation rate:

$$v = k_p\,[RM^\bullet]\,[M] \tag{6a}$$

$$v = k_p\left\{(k_s/k_t)^{1/2}\,D^{\bullet 1/2}\right\}[M]^{3/2} \tag{6b}$$

Table 1. G value for free-radical formation G (R) and activation energy E_a of radiation-induced polymerization for several monomers

Monomer	$G(R)$, molecules/100 eV	E_a, kJ/mol
Styrene	0.69	27.8
Methyl methacrylate	11.5	20.6
Vinyl acetate	12.0	23.9
Acrylonitrile	11.2	12.6 – 54.6
Methyl acrylate	≈15	≈21
Ethylene	4.4	

Given a fixed dose rate, the overall rate of radiation-induced free-radical polymerization depends on the rate constants of the propagation and termination reactions and, via the production rate of primary radicals $R^\bullet$, on their G value (Table 1). The overall polymerization rate (unit mol L^{-1} s^{-1}) is proportional to the square root of the dose rate, at least in the range of low dose rates (Fig. 2, region 1). This, however, applies only to monomers for which no part of the polymer precipitates during the reaction. Exceptions are ethylene, vinyl chloride, acrylic acid, and acrylonitrile. The kinetics of chain propagation and chain initiation are largely the same as those of the chemically-initiated chain reaction. The activation energies for radiation-induced polymerization (Table 1), are primarily the activation energies of the chain propagation step.

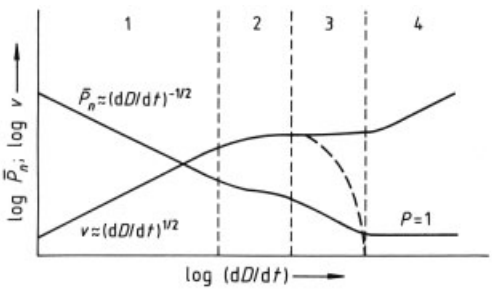

Figure 2. General dependence of polymerization rate v and mean degree of polymerization P_n on dose rate, according to [32]

As the dose rate increases (Fig. 2, region 2), chain termination by primary radicals becomes more important and effectively competes with reactions (3 a) and (3 b):

$$RM_n^\bullet + R^\bullet \rightarrow RM_n R \qquad (7)$$

If the dose rate is increased further (region 3), a state is reached in which all propagating chains are terminated by primary radicals. The molecular mass of the product polymers falls off sharply. The monomer consumption is no longer governed by chain propagation (v shown by dashed curve) but by the radiolysis of the monomer. In region 4, radiolysis of the monomer is the sole process. The degree of polymerization drops to 1.

This unfavorable dose-rate dependence, however, is not a limiting factor in the industrial use of high dose rates (such as those in electron irradiation) for polymerization processes. Two factors ameliorate the detrimental effect of high dose rates: If the polymer precipitates as it is produced, the kinetic relations stated above are no longer valid. The overall polymerization rate is described by dose-rate exponents between 0.5 and 1. In highly viscous media, chain termination is strongly hindered by the low diffusion coefficient, and here the dose-rate exponent can again reach a value of unity.

As is shown in more detail in Section 1.3, electron bombardment even at dose rates of 10^5–10^6 Gy/s is very effective in achieving free-radical polymerization or cross-linking of high-viscosity monomer–prepolymer mixtures such as those used industrially as binders for paints, printing inks, adhesives, etc.

Ionizing radiation can act on both monomer molecules and polymer macromolecules by direct electron mechanisms, producing excitation or ionization. The formation of polymeric free radicals $P^\bullet$ cannot be neglected at conversions as low as ca. 20 %. At high conversions, there are two competing initiation reactions:

$$M \rightsquigarrow 2R^\bullet \qquad (8a)$$

$$P \rightsquigarrow 2P^\bullet \quad \text{or} \quad P^\bullet + R^\bullet \qquad (8b)$$

As the monomer conversion increases, branched-chain polymers (homograft polymers) are obtained. When the monomer concentration has declined to the point that not all the free radicals produced can be scavanged by monomer molecules, radiolysis of the polymer becomes dominant, leading to either cross-linking or chain termination. Figure 3 shows schematically the phases of radical polymerization in a plot of conversion versus dose.

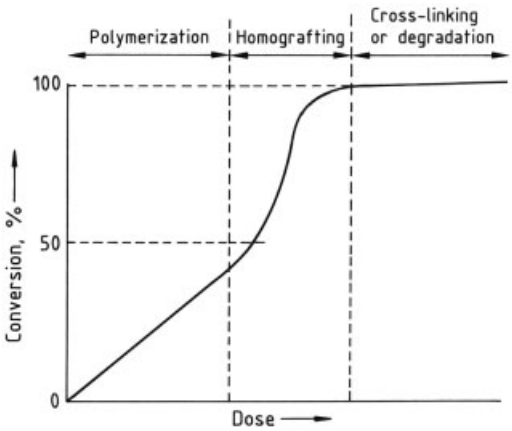

Figure 3. Processes in irradiated monomers versus dose, according to [40]

1.1.3. Ionic Polymerization [32], [34], [41], [43–45]

If monomers, such as those in Table 2, are purified of trace water and irradiated at sufficiently high dose rates, polymerization can be initiated by free ions that have escaped recombination with their partners in the primary ion pair (Fig. 1 C). Cationic polymerization is the principal type observed for the monomers of Table 2. Styrene and α-methylstyrene can, however, also be anionically polymerized. In radiation-induced ionic addition polymerization, chain propagation and chain termination exhibit kinetics similar to that in radical chain polymerization.

Table 2. Rate constants for the chain propagation reaction with superdry monomers, according to [40]

Monomer	t, °C	Rate constant, L mol^{-1} s^{-1}	Activation energy, kJ/mol
Styrene	15	3.5×10^6	0
α-Methylstyrene	0	4×10^6	0
	30	3×10^6	0
Cyclopentadiene	−78	6×10^8	<8
Isobutene	0	1.5×10^8	<8
Isobutyl vinyl ether	30	3×10^5	<21

$$\text{Initiation: } M \rightsquigarrow M^+(M^-) \quad (9)$$

$$\text{Propagation: } M^+ + M \longrightarrow MM^+ \quad (10a)$$

$$M_nM^+ + M \longrightarrow M_{n+1}M^+ \quad (10b)$$

In the presence of water, deprotonation of the radical cation is a competing process:

$$M_nM^+ + H_2O \longrightarrow M_n M^\bullet(-H^+) + H_3O^+ \quad (11)$$

Chain termination:

$$M_n M^+ + M_m M^- \longrightarrow P_1 \quad (12a)$$

$$M_n M^+ + Y^- \longrightarrow P_2 \quad (12b)$$

where Y^- denotes an impurity.

Similarly to radical chain polymerization, the overall rate of radiation-induced ionic polymerization is given by

$$v = k_p\,[M_n M^+]\,[M] \quad (13)$$

$$v = k_p\,(k_s/k_t)^{1/2}\,D^{\bullet 1/2}\,[M]^{3/2} \quad (14)$$

and the mean lifetime τ of charge carriers in the steady state by

$$\tau = \frac{[M_nM^+]}{k_s\,[M]\,D^\bullet} = (1/k_t \cdot k_s\,[M]\,D^\bullet)^{1/2} \quad (15)$$

The kinetic chain length $\hat{L}$ is

$$\hat{L} = k_p\,[M]\cdot\tau \quad (16)$$

The rate constants for chain propagation in radiation-induced ionic polymerization are much larger than those for conventional ionic polymerization, where the active species consist of thermodynamically stable ion pairs. Table 3 compares the propagation rate constants for the addition polymerization of styrene in various processes. Measurements of the molecular mass distribution of radiation-polymerized polystyrene have shown that free-radical and ionic polymerization can occur at the same time if traces of water are present in the monomers. If the dose rate is increased, for example if polymerization is initiated by fast electrons, the effect of impurities on cationic chain propagation can be minimized.

Table 3. Comparison of chain propagation rate constants k_p for addition polymerization of styrene, according to [45]

Initiation by	Solvent	t, °C	k_p, L mol^{-1} s^{-1}
Radiation, free cation	styrene	15	3.5×10^6
$Bu_4N^+CF_3SO_6^-$, free cation	dichloroethane	30	3.0×10^6
Free anion	tetrahydrofuran	25	6.0×10^4
Free radical	styrene	20	35
Cationic ion pair	sulfuric acid	25	7.7

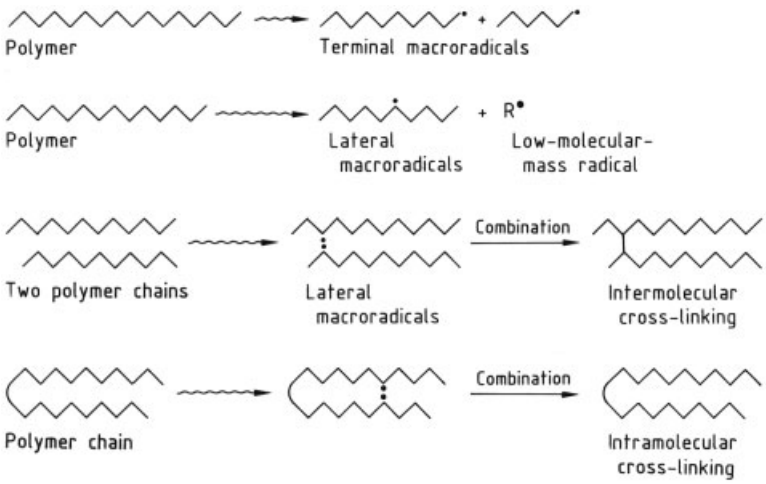

Figure 4. Radical formation processes in irradiated polymers (schematic)

Other monomers that are polymerized by a free-radical process at low dose rates and a cationic process at higher dose rates include α-methylstyrene, isobutylvinyl ether, and isoprene. With butadiene only cationic polymerization is found.

1.1.4. Radiation-Induced Polymerization of Solid Monomers [44], [46–51]

Diffusion processes in the solid phase are orders of magnitude slower than in the liquid phase. Although radiation-induced polymerization of solid monomers is also initiated by free radicals or ions, the much slower diffusion results in some special features of solid-phase polymerization. Chain propagation may slow down, and chain termination by free-radical combination is also strongly hindered. The lifetime of free radicals in the solid phase may be eight to ten orders of magnitude longer than that in the liquid phase. The migration of free radicals in solid monomers and polymers is partly explained by nondiffusion processes such as the migration of radical sites due to a series of hydrogen addition and abstraction reactions.

Differences in polymerization kinetics are observed between crystalline and amorphous monomers. In some cases, the crystal structure of the monomer determines that of the polymer (e.g., 1,3,5-trioxane). Monomers that can be polymerized in the solid phase by ionizing radiation include acrylamide, diketene, trioxane, tetraoxane, pentoxane, styrene, and methacrylates.

In mixtures of monomers, for example acrylamide or acrylic acid with glass-forming matrices, radiation-induced polymerization has been observed to occur in the glassy state.

Methacrylates and acrylates in the glassy state can be polymerized especially well at low temperatures. Since scarcely any three-dimensional contraction occurs during this process, polymers with good optical properties are obtained in this way.

Bioactive substances such as enzymes can be readily introduced into glass-forming acrylate monomers, cooled, and immobilized through polymerization of the matrix. These biofunctional polymers are used in biomedical fields and biochemistry.

1.2. Action of Ionizing Radiation on Polymers – Polymer Modification

1.2.1. Reaction Mechanisms [52–63]

The action of ionizing radiation on polymers commonly results in simultaneous cross-linking and degradation reactions. These processes can be detected through the change in mean molecular mass of the polymers, the formation of low molecular mass reaction products such as H_2,

CH_4, CO, CO_2, hydrogen halides, and changes in the physical properties.

In addition, oxidation reactions occur in the presence of atmospheric oxygen. Even low doses of radiation can markedly change the physical properties of the polymer. Formally, free-radical formation processes can be accounted for by the cleavage of main-chain and side-chain bonds in the polymer (Fig. 4).

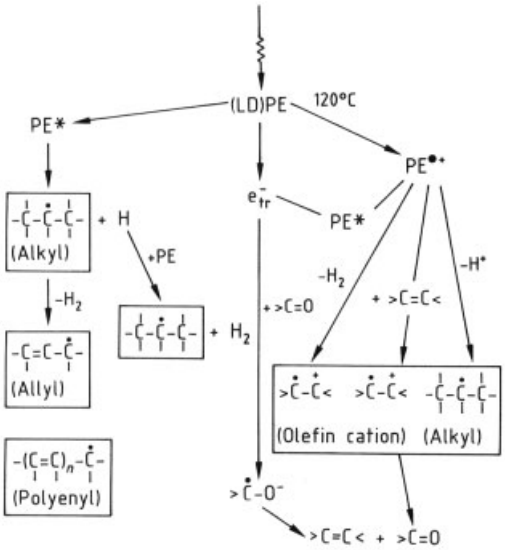

Figure 5. Radical formation processes in polyethylene irradiated at 120 °C, according to [61]

The free-radical formation mechanism in irradiated polymers is generally far more complex and cannot be satisfactorily explained by the assumption of direct bond cleavage. Attempts to use pulse radiolysis to explain free-radical formation in low density polyethylene (LDPE) heated to 120 °C have led to the (simplified) reaction scheme of Figure 5. The radiation-generated primary products in polyethylene (PE) are excited polymer units PE∗, "trapped" electrons e^-_{tr}, and cationic polymer units $PE^{\bullet+}$. Recombination of electrons and cations again leads to excited polymer units, which decompose into alkyl and hydrogen radicals. Alkyl radicals, however, can also be produced by deprotonation of radical cations $PE^{\bullet+}$. Elimination of H_2 competes with deprotonation. Products of H_2 elimination are olefinic cations; these, however, can also be produced by charge transfer to double bonds already present in PE. Electrons are probably stabilized on existing carbonyl groups to form anions. Recombination of these anions with cations leads to radiation-generated double bonds.

In the presence of oxygen, radiation-induced degradation is commonly greatly accelerated. The initiation step is the formation of peroxy radicals:

$$-C-C^{\bullet}-C- + O_2 \longrightarrow -C-\overset{\overset{\displaystyle OO^{\bullet}}{|}}{C}-C- \quad (17)$$

Hydroperoxides are obtained after hydrogen addition; these can decompose to alkoxy and OH radicals

$$-C-\overset{\overset{\displaystyle OOH}{|}}{C}-C \longrightarrow -C-\overset{\overset{\displaystyle O^{\bullet}}{|}}{C}-C- + OH^{\bullet} \quad (18)$$

or can form carbonyl groups:

$$-C-\overset{\overset{\displaystyle OOH}{|}}{C}-C \xrightarrow{+H} -C-\overset{\overset{\displaystyle O}{\|}}{C}-C- + H_2O \quad (19)$$

1.2.2. Cross-Linking and Degradation

When polymers are irradiated, cross-linking and degradation usually occur simultaneously. Cross-linking occurs when two species (e.g., free radicals) produced on neighboring polymer units recombine (Fig. 4). The relative molecular mass of the branched macromolecules produced increases with increasing dose and the melting point rises. At the same time, a temperature range is found in which the cross-linked polymer can possess rubber elasticity and is no longer completely soluble. The minimum dose necessary for gel formation (i.e., for production of cross-links) is called the *gel dose* D_{gel}. If two monomer units form a cross-link and no simultaneous chain degradation occurs, the gel dose (in rad) is given by

$$D_{gel} = \frac{100 \cdot N_A}{2 \cdot 6.24 \times 10^{13} \cdot G(X) \cdot M_r} \quad (20)$$

where $G(X)$ is the number of cross-links produced per 100 eV of absorbed energy, M_r is the number-average relative molecular mass of the uncross-linked polymers, and N_A is the Avogadro constant.

Thus the G value for cross-linking can be determined by measuring the gel dose. If *degradation* takes place at the same time, the gel dose increases; $2\,G(X)$ in Equation (20) is then replaced by $[2\,G(X) - G(S)/2]$, where $G(S)$ is the number of cleaved bonds in the main chain per 100 eV of absorbed energy.

The radiation-induced degradation of a polymer leads to a decrease in molecular mass. Gel formation is possible only if $G(S)/G(X) < 4$. For linear polymers, the degree of cross-linking is proportional to the dose. Its dependence on the dose rate, the chemical structure of the polymer, and the type of ionizing radiation is not significant. The G value for cross-linking is only partially influenced by the molecular mass of the polymer but it can be changed significantly by additives (sensitization or desensitization of cross-linking).

The extraction method is often used to determine the *cross-linking density* q. CHARLESBY and PINNER [64] found the following relation for the soluble (sol) content s,

$$s + \sqrt{s} = p/q + 1/qP_n \cdot D \tag{21}$$

where s is the sol content, q the cross-linking density, p the degradation density, P_n the number average degree of polymerization, and D is the dose.

If $s + \sqrt{s}$ is plotted versus $1/D$, the intersection of the resulting line with the vertical axis gives the ratio p/q of the degradation density to the cross-linking density.

The degree of cross-linking can be also determined by measurement of the swelling capacity and elastic modulus of the network. The measurement of spin – spin relaxation times (T_2) by pulsed NMR yields direct information about network formation in polymers and thus about the degree of cross-linking [65], [66]. Solid-state NMR studies can also supply information about cross-links and new terminal groups produced by irradiation.

The above model of a statistically distributed production of cross-links in polymers with statistically distributed molecular masses has to be modified if a crystalline phase and "chain folding" are present. For example, many experimental indications show that radiation-induced cross-linking in, for example, polyethylene occurs mainly in the amorphous phase [57]. Only at very high doses ($> 10^6$ Gy) can a marked decline in crystallinity be seen.

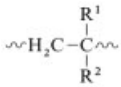

Vinyl polymers having a structure of the type with R^1 and R^2 = alkyl, COOR, OH, or halogens are predominantly degraded. When $R^1 = R^2 = H$, then cross-linking becomes dominant. Aromatic side groups transform much of the absorbed energy without any chemical consequences and thus provide stability against degradation. CO groups in the main chain, on the other hand, promote degradation. Polymers whose monomers exhibit a high enthalpy of polymerization are, as a rule, predominantly cross-linked; polymers that readily form monomers by pyrolysis have a tendency to degrade.

The following are mainly *cross-linked*: polyethylene, polypropylene, polystyrene, poly(vinyl chloride), poly(vinyl alcohol), polyacrylates, polyesters, polyamides, polysiloxanes, natural rubber, synthetic rubber (except butyl rubber), phenol – formaldehyde resins, aminoformaldehyde resins, and polysiloxanes.

The following are mainly *degraded*: polyisobutylene, butyl rubber, poly-α-methylstyrene, poly(vinylidene chloride), polymethacrylates, polymethacrylamides, polytetrafluoroethylene, cellulose and its derivatives, polytrifluoroethylene, and polycarbonates.

1.2.3. Graft Copolymerization [32], [34], [67], [69–78]

The reactive species produced when ionizing radiation acts on polymers or monomers are primarily free radicals and, in a few cases, radical ions. Radiation graft copolymerization offers the following general advantages:

1) Good range of variation for the initiation step
2) Initiation under conditions where conventional methods fail (low temperature, solid-phase polymerization)
3) Freedom from chemical initiators or catalysts ("clean" systems)

However, special conditions must be applied to every system to be grafted, on account of dif-

ferences in the formation mechanisms of reactive species, the *G* values, and the reactivity.

The only process that has gained practical importance is *heterogeneous grafting*. Here the polymer to be activated is in, for example, solid form while the monomer can be in the vapor or liquid state (see Table 8).

The following methods are available for graft copolymerization:

Preirradiation Method. The polymer P_A is irradiated with oxygen exclusion. The free radicals formed are long-lived and react with the monomer M_B to form a side chain:

$$P_A \rightsquigarrow P_A^\bullet + H^\bullet \quad (22\,a)$$
$$H^\bullet + P_A \longrightarrow P_A^\bullet + H_2 \quad (22\,b)$$
$$P_A^\bullet + n\,M_B \longrightarrow P_A\,n\,M_B \quad (22\,c)$$

Preirradiation is well suited to continuous grafting processes, such as the electron irradiation of polyethylene film followed by grafting of acrylic acid.

Peroxide Method. When the polymer P_A is irradiated in air, the active centers formed are chiefly hydroperoxides. These, after heating or UV irradiation if necessary, decompose to give oxygen and OH radicals.

$$P_A \xrightarrow{O_2} P_A O_2^\bullet \xrightarrow{+H} P_A OOH$$
$$\xrightarrow{\Delta,\,UV} P_A O^\bullet + OH^\bullet \quad (23\,a)$$
$$P_A O^\bullet + M_B \longrightarrow P_A OM_B \text{ (graft copolymer)} \quad (23\,b)$$
$$OH^\bullet + n\,M_B \longrightarrow n\,M_B^\bullet OH \text{ (homopolymer)} \quad (23\,c)$$

Homopolymerization can be lowered by the addition of OH scavengers (Fe^{2+}).

Simultaneous Method. The polymer P_A and the monomer (vapor, liquid, or solution) in contact with it are simultaneously irradiated.

$$P_A \xrightarrow{O_2} P_A^\bullet, \quad M_B \xrightarrow{O_2} M_B^\bullet \quad (24\,a)$$
$$P_A^\bullet + M_B \longrightarrow P_A M_B \text{ (graft copolymer)} \quad (24\,b)$$
$$M_B^\bullet + n\,M_B \longrightarrow M_B\,n\,M_B \text{ (homopolymer)} \quad (24\,c)$$

This method is very efficient because of low radical losses (e.g., polyethylene – styrene), but it also yields considerable amounts of homopolymer.

In many cases, radiation-induced graft polymerization is sensitized by additives (polyfunctional monomers, acids, inorganic salts).

1.3. Applications

1.3.1. Electron-Beam Curing of Paints, Printing Inks, Adhesives, and Release Agents (see also → Paints and Coatings, Chap. 3.7.; → Printing Inks, Chap. 10.1.)

1.3.1.1. Principle of Radiation Curing [69], [79–91]

As proposed by the RadTech Conference 1988, radiation curing is defined as "The use of electron beam or ultraviolet radiation as an energy source to induce the rapid conversion of especially formulated 100 % reactive liquids to solids."

The polymerization and cross-linking of predominantly solvent-free coatings and printing inks can be initiated either by ionizing radiation (in industry, mainly fast electrons with energies between 120 and 300 keV) or by UV quanta.

Initiation by *electrons* leads to the primary free-radical polymerization and cross-linking processes described in Section 1.1. Cationic polymerization is only found in rare cases. With *UV excitation*, the radiation quanta are first absorbed by photoinitiators. After excitation, the photoinitiators form active species (free radicals or protons), which initiate the chain reaction of radical or cationic polymerization. In industry, these two types of radiation curing are in competition with thermal curing. The radiation process offers significant advantages over the thermal process:

1) No solvent release
2) Reduction of energy consumption by as much as an order of magnitude
3) Significantly smaller space requirement
4) High production rates
5) Elimination of processing steps; possibility of continuous coating and curing (since the cured products can be stored or forwarded to downstream operations immediately)
6) Generally high quality of coating or printing

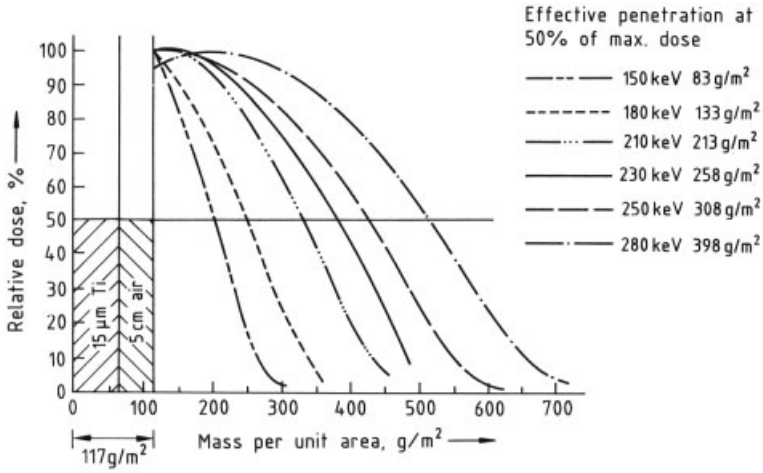

Figure 6. Depth – dose distribution (relative dose versus mass per unit area) for ISO – 280 eV electrons [110]

7) Only slight increase of temperature in curing, which makes it possible to process temperature-sensitive substrates

Electron-beam curing processes have the drawback that raw materials for radiation-cured coatings and inks are more expensive than conventional ones and some contain skin-irritating ingredients. Energetic and promising work is under way to remedy these defects through the development of low- cost feedstocks and formulations with improved service properties [84–90].

The discussion that follows will deal chiefly with the electron-beam curing of solvent-free coatings. The dose distribution in the coating and substrate is particularly important for the curing process.

Typical formulation for electron-beam curing require doses between 10 and 50 kGy. Given the mass per unit area M_A of coating plus substrate, the required electron energy and the overdose (dose inhomogeneity) are determined from the depth – dose distribution (Fig. 6) because the following relation holds:

$$M_A = \frac{m}{A}$$

$$\frac{M_A}{\varrho} = \frac{m}{A} \cdot \frac{A \cdot d}{m} = d$$

where d is the penetration depth. Low-energy electron accelerators suitable for industrial electron-beam curing are on the market, together with raw materials for radiation- curing formulations.

1.3.1.2. Chemistry of Electron-Beam Curing [79], [80], [91–96]

Coatings for electron-beam curing generally consist of the following:

1) Prepolymers, usually with acrylic ($H_2C{=}CH{-}CO{-}O{-}$) double bonds or double bonds in the main chain (polymaleates and polyfumarates)
2) Monomers, usually acrylates, used as reactive thinners to reduce the viscosity and influence the properties of the coating
3) Pigments and dyes; if needed, fillers, flatting agents, and additives to improve the film and surface properties

Acrylated Prepolymers. Many highly reactive prepolymers can be produced by acrylation of oligomers (e.g., epoxy resins, urethanes, polyesters, polyethers, silicones, oligobutadiene, melamine derivatives, cellulose, and starche) with a variety of agents (e.g., acrylic acid, acrylamide, hydroxyethyl acrylate, and glycidyl acrylate). Prepolymers are the principal ingredients of coating formulations and largely determine the basic properties of the coating.

Polyester acrylates, epoxy acrylates, polyurethane acrylates, and silicone acrylates have found wide use in industry (see Fig. 7).

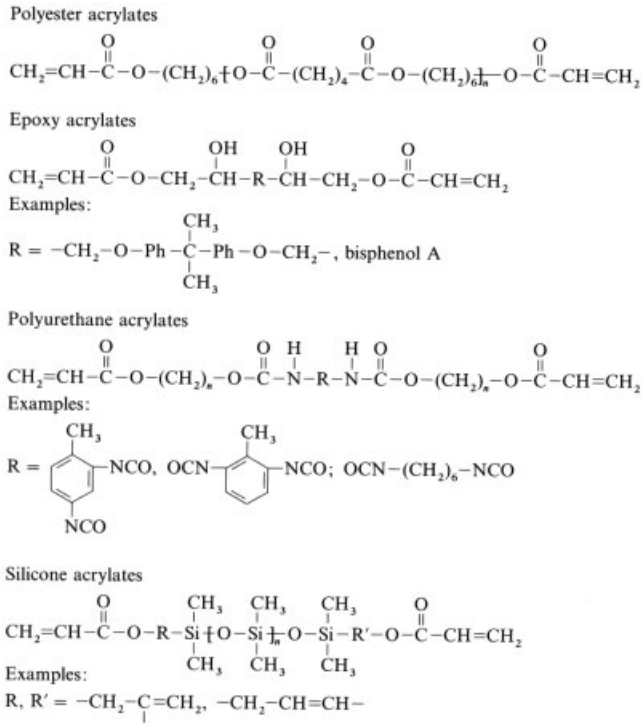

Figure 7. Acrylated prepolymers for radiation-curable binders

Polyester acrylates, even those with molecular masses of 300 – 1000, have low viscosities. They are suitable for coatings with good elasticity and high pigment content.

Epoxy acrylates are produced from bisphenol A, aliphatic and cycloaliphatic expoxides, epoxidated fatty acids (e.g., soybean oil), and novolacs. *Aromatic epoxy acrylates* enable the production of radiation-cured printing inks and coatings for wood and paper. They have a high viscosity and, in combination with suitable monomers, yield hard, scratch-resistant coatings. *Epoxidated and acrylated oils* generally have low viscosities and good cross-linking properties; they are not irritating to the skin but are less reactive. Good thermal and chemical stability is achieved with *acrylated novolacs.*

Urethane acrylates, in general, are highly reactive and are applied in high-gloss coatings with good chemical and scratch resistance. They show good adhesion to a variety of substrates. The hardness and flexibility of the films can be controlled through, for example, the chain length of the oligomer.

Silicone acrylates are chiefly employed for the production of antiadhesive papers, separating foils, and adhesive tape backings, and also for the coating of glass fibers.

Polydimethylsilanes are usually modified with vinyl, acrylate, or thiol groups to lend them adequate reactivity.

Monomers. Acrylate esters have found aceptance in radiation-curing practice because of the range of different products on the market as well as their availability, reactivity, viscosity, and price. The monomers, also called reactive thinners, form a high-molecular-mass network with the oligomer after curing. In order to attain an adequate degree of cross-linking, principally bifunctional and polyfunctional acrylates are employed. Monofunctional acrylates give a less reactive coating; they are not suitable constituents for industrial formulations because of their volatility, odor, and skin-irritating action.

In what follows some bifunctional and polyfunctional acrylates are listed that have found

Table 4. Low-energy electron accelerators used in industry

Manufacturer, model	Cathode configuration	Electron energy, keV	Beam power, kW	Maximum beam width, m	Remarks
Energy Sciences, Inc. (ESI), US, Electrocurtain	linear cathode with control grid	150 – 300		2.3	up to 4 cathode systems in one accelerator head
Electrocure [98–101], [103]	multicathode		200		used in printing industry
Radiation Polymer Co. (RPC), US [104–106]	multicathode with 2 grids	150 – 300	200	2.3	
Broad Beam Dual Accelerator	multicathode	300	980	2.3	consists of 2 accelerators
Polymer Physik (PP), Germany. Scanner [108–111]	needle cathode	150 – 280	100	2.0	2 cathode systems possible
LEA [112]	linear cathode with no control grid	120 – 200	60	1.0	used chiefly in printing industry
Nissin-High Voltage Co. (NHV), Japan	linear cathode	150 – 200	200	1.0	
Scanner [107]	needle cathode	300 – 500	65	1.2	

broad use in industrial coatings and printing inks (abbreviations are given in parentheses):

Tripropylene glycol diacrylate	(TPGDA)
1,6-Hexanediol diacrylate	(HDDA)
Dipropylene glycol diacrylate	(DPGDA)
Trimethylolpropane triacrylate	(TMPTA)
Trimethylolpropane ethoxytriacrylate	(TMP(EO)TA)
Trimethylpropane propoxytriacrylate	(TMP(PO)TA)
Pentaerythritol triacrylate	(PETA)
Glyceryl propoxytriacrylate	(GPTA)

Ethoxylation and propoxylation of acrylates have been used to produce monomers which are less skin-irritant but highly reactive. Well-cured coatings contain an extremely low amount of extractable residual monomers or other products that tend to migrate.

Use of Inert Atmospheres. Free-radical polymerization and cross-linking processes are inhibited by oxygen. Oxygen reacts both with the initiating free radicals and with the propagating free-radical chain. The reaction with oxygen either prevents chain propagation or terminates the propagating chain. The coating surfaces remain tacky ("oxygen-inhibited") and exhibit totally inadequate properties. Accordingly, the residual oxygen concentration in the irradiation zone must be lowered to 100 – 1000 ppm (depending on the coating system) by inert-gas purging. At present, virtually the only inert gas used is nitrogen, which is obtained by vaporizing liquid nitrogen (purity $> 99.99\,\%$).

In the electron-beam curing of, e.g., coatings on flexible substrates, the engineering requirement of inert conditions with minimum nitrogen consumption is met by using a directed high-velocity jet of inert gas to "peel off" the air layer on the coating at the inlet to the curing zone. The irradiation chamber is then purged in such a way as to maintain the most laminar flow conditions possible. The inert-gas consumption can thus be reduced to 70 – 100 m^3/h (STP) for a curing zone 1 m wide and a product speed of 300 m/min.

Often, part of the nitrogen required for inertizing is used for convective cooling of the accelerator beam window. The inert gas also prevents the formation of ozone, which cannot be prevented when air is irradiated.

1.3.1.3. Industrial Irradiation Equipment, Low-Energy Electron Accelerators [69], [78], [81–83], [97–112]

Accelerators generating electrons with energies between 120 and 300 keV are employed for the electron-beam curing of coatings and printing inks or rigid and flexible substrates. This energy range is sufficient to cure coatings with mass per unit area between 60 g/m^2 (120 keV) and 400 g/m^2 (300 keV), given a 20 % dose inhomogeneity in the coating (Fig. 6) and it permits the use of simple accelerators with integral radiation shielding. Low-energy accelerators are a mature technology. Under industrial conditions, they can be used as computer-controlled subsystems in coating machines, printing presses, and laminating machines, etc. Their operating

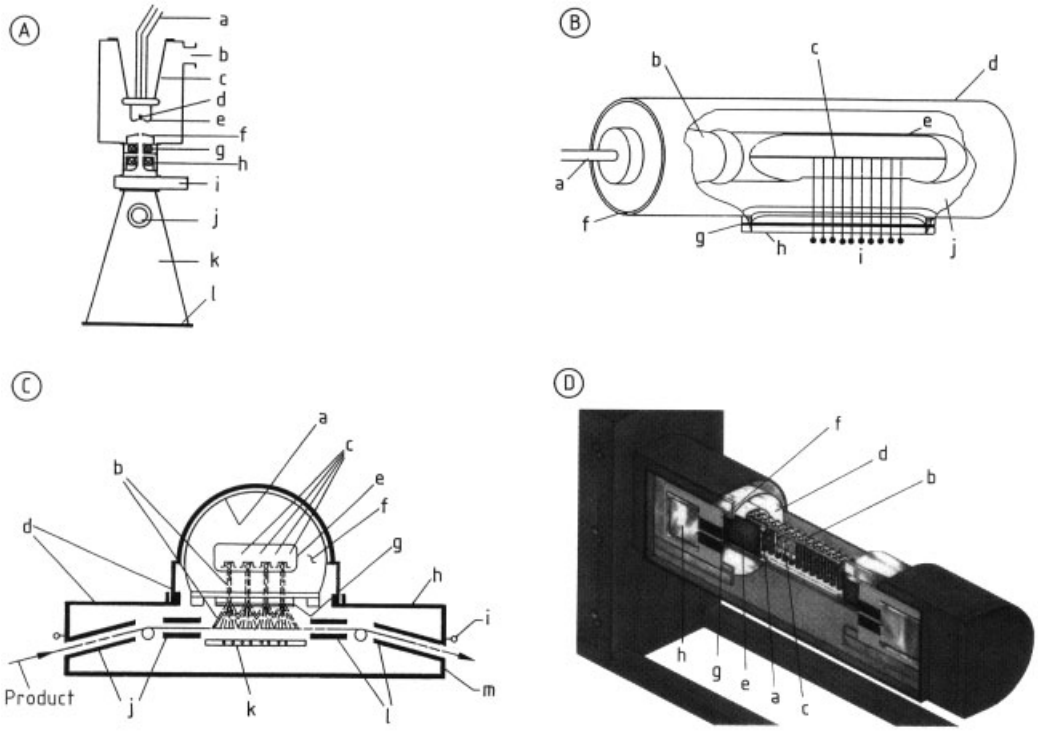

Figure 8. Designs of low-energy electron accelerators used in industry
A) Scanner; a) Accelerating potential; b) Vacuum connection; c) Insulator; d) Cathode; e) Wehnelt cylinder; f) Anode; g) Beam focussing; h) Beam deflection; i) Valve; j) Vacuum connection; k) Scanner; l) Electron exit window; B) Linear cathode with no control grid; a) Cable; b) High-voltage insulator; c) Cathode; d) Acceleration vessel; e) Electron optics; f) Lead radiation shielding; g) Titanium foil; h) Window; i) Electrons; j) Vaccum; C) Linear cathode (1 – 4) with grid; a) Chamber; b) Electron beam; c) Electron guns; d) Shielding; e) Terminal; f) Vacuum; g) Titanium foil (window); h) "Top" selfshield; i) Radiation monitor; j) Infeed collimators; k) Beam collector; e) Outfeed collimators; m) "Bottom" selfshield; D) "Broad Beam" pocessor; a) Tungsten filament; b) Filament Support; c) Control grid; d) Vacuum chamber; e) Screen grid; f) Window support plate; g) Holes; h) Titanium foil window

parameters (electron energy, beam power, width of irradiation zone, homogeneity of dose over the irradiation width) can be precisely matched to the results of the industrial process in which the accelerators are used.

At present, the electron-beam power is limited to around 200 kW, the irradiation width to 2.5 m, and the product speed to some 1500 m/min (for an applied radiation dose of 10 kGy). In graft copolymerization applications, however, a dual accelerator with a 2.3 m irradiation zone width, 300 keV energy, and 980 kW beam power is being built [106]. The estimated number of low-energy electron accelerators in industrial plants in 1992 is ca. 150 units worldwide.

Table 4 gives the parameters and manufactureres of the most widely used modern low-energy accelerators. In all the accelerator types shown in Table 4 the electrons are produced by a hot cathode held at a high voltage. As they travel across the potential difference between cathode and anode (the outer wall of the accelerator beam window), the electrons acquire the energy needed to pass through the titanium foil window. The energy loss of 120 keV electrons in a 15 µm titanium foil is as much as 30 %. This fact sets a practical lower limit on the electron energy. An upper limit on the electron-beam power is given by the maximum possible current per square centimeter of window area which is 0.2 mA/cm^2.

The accelerating voltage limit of 300 keV is dictated, firstly, by the physical capabilities of electron acceleration in a simple cathode – anode system with no intermediate

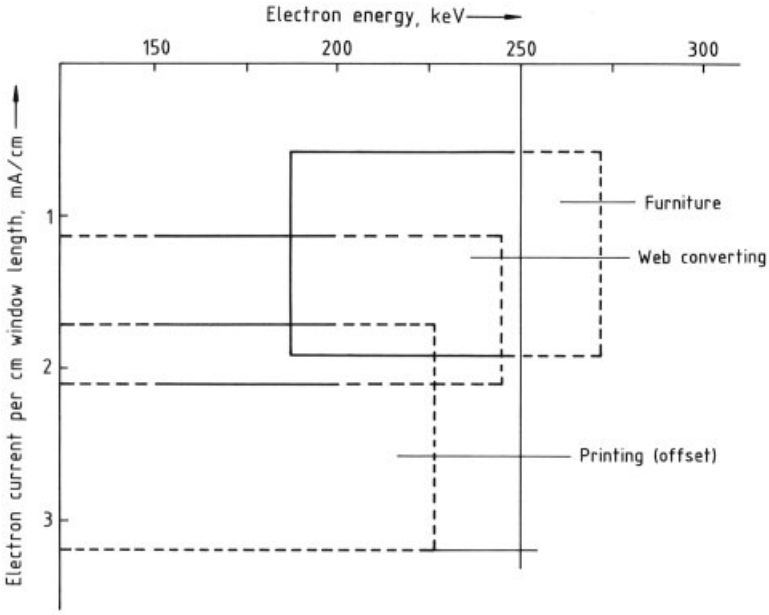

Figure 9. Applications of electron-beam curing versus electron energy and electron current per cm of window length

electrodes and, secondly, by the fact that the lead shielding for higher voltages is not easy to handle and offers little advantage over fixed concrete shielding.

If a point (needle) cathode is used, the roughly pencil-sized electron beam must be deflected by a varying electromagnetic field so as to cover the required irradiation zone. This beam deflection or scanning gives the name scanner to this type of accelerator (Fig. 8 A). Alternatively, a hot wire of appropriate length can also be chosen as the electron emitter; the electrons are then focused by electron optics and accelerated toward the beam window. Systems of this type are shwon in Figure 8 B – D. Several configurations are distinguished: linear cathode with con-trol grid (Electrocurtain, Fig. 8 C), linear cathode without control grid (LEA, Fig. 8 B), and multicathode arrangements with the hot wires transverse to the long axis of the window ("Broad Beam," Fig. 8 D).

Linear-cathode and multicathode accelerators are more compact than scanners, but do not offer such good dose homogeneity over the irradiation zone width.

1.3.1.4. Industrial Applications of Electron-Beam Curing [69–79], [81–90], [102], [113–123]

There are presently three major applications for electron-beam curing: upgrading the surfaces of rigid and flexible materials, and printing. These fields impose different requirements on electron accelerators (Fig. 9). The electron-beam curing of coatings on *rigid substrates* – chiefly furniture, doors, laminated panels, and ledges – requires relatively high electron energies. Some of these coatings have mass per unit area of 200 g/cm^2. When edges and small-area parts cannot be reached by direct electron irradiation, curing is effected by electrons scattered in the air.

Production rates in, for example, the furniture industry are not determined by the curing step but by other operations, such as feeding, discharge, and polishing, etc. Low to moderate beam powers are therefore adequate.

Electron-beam curing of coatings on *flexible* materials often takes place as the substrate passes from roll to roll. Production rates between 100 and 300 m/min are common, at coating weights of 1 – 30 g/m^2. Therefore, low electron

energies and moderate to high beam powers are needed.

In *printing*, especially in roll-fed offset printing, press speeds range from 600–1000 m/min, but printing speeds are somewhat lower. Despite the relatively low curing doses required for printing inks (15–25 kGy), extremely high beam powers with the lowest possible electron energies are required.

Table 5 summarizes important industrial applications of electron-beam curing.

1.3.2. Radiation-Induced Cross-Linking [69], [81–83], [124–137]

The radiation-induced cross-linking of polymers is an established industrial technology, especially in the following areas:

1) Cross-linking of cable insulation, tubes and pipes, and moldings
2) Heat-shrink materials
3) Vulcanization of elastomers
4) Processing of foamed plastics

Radiation-cross-linked products produced in the United States alone had a value of $\$ 1840 \times 10^6$ in 1988.

Industrial irradiation equipment comprises chiefly electron accelerators whose energy, beam power, and irradiation zone width can be opimally matched to the process requirements [138–147]. Only in a few exceptional cases (shrink-wrap films, thin elastomer coatings) can low-energy accelerators be employed. Because of the required electron penetration, accelerators with energies of 0.5–5 MeV and power < 200 kW are mainly used.

Radiation-induced cross-linking is competitive with peroxide cross-linking (initiated by thermal cleavage of peroxides) in the cable insulation sector. Although there is no fundamental difference in the chemistry of peroxide and radiation-induced cross-linking, aside from the initiation process, there are a number of process differences (Table 6).

1.3.2.1. Cross-Linking of Cable and Wire Insulation

The use of electron beams to cross-link cable and wire insulation is an economical and highly productive process. Insulation with a high heat resistance, which is thus stable against Ohmic heating and soldering operations (< 250 °C for brief periods), is easy to produce in this way. In addition to stability at high temperature, the solvent resistance and the mechanical properties of the insulation can be improved by cross-linking. Besides polyethylene, other radiation-cross-linkable polymers are increasingly used for wire insulation. Table 7 lists cross-linkable materials suitable for electrical insulation.

Table 7. Radiation-cross-linked materials for cable and wire insulation [124]

Material	Cross-linking dose, kGy
Polyethylene	200–300
Polyethylene (sensitized)	100–150
EPDM [a]	100–150
PVC (sensitized)	50–80
Sulfonated polyethylene	100–150
Polyvinylidene fluoride	80–160
EPR [b]	50–150
Ethylene–tetrafluoroethylene copolymer	60–100
Silicone rubber	200

[a] Terpolymer of ethylene, propene, and a diene.
[b] Ethylene–propylene rubber.

Thermoplastics and elastomers to be cross-linked often contain additives to either lower the cross-linking dose or to make cross-linking possible at all (PVC). Such additives can be polyfunctional unsaturated compounds, e.g., acrylates, methacrylates, vinyl, allyl, and acetylene compounds, or oligobutadiene, but inorganic, metallic, and organometallic compounds can also be employed.

In the radiation cross-linking of polyethylene insulation layers more than 10 mm thick, the accumulation of charge carriers leads to electrical breakthrough and damage to the insulation layer. Radiation- cross-linked insulation can therefore be used only for low-voltage and medium-voltage cables.

1.3.2.2. Vulcanization of Elastomers

Radiation-cross-linked elastomers (Table 7) are used chiefly for cable and wire insulation and in tire manufacturing [148], [149]. Cross-linking and degradation are again the main processes when elastomers are treated with radiation;

Table 5. Industrial applications of electron-beam curing

Surface treated	Application	Location
Solid materials		
Wood	door finishes	Europe
	finishes on front panels of furniture	Europe, USA
	lamination and lacquer finishes	USA
	coating of wood–cement panels	Europe
	edge curing	Europe
	coating of moldings (on all sides)	Europe
	coating of medium density fiberboard (MDF) panels	Europe
Gypsum	coating of gypsum panels	Japan
Metal	coating of metal panels	Japan
	coating of metal strip (roll-to-roll)	Japan
Flexible materials		
Paper	coating of paper used to decorate wood	Europe, USA, Australia
	silicone coating of paper for labels, protective films, dividers,	Europe, USA, Japan
	high-gloss lacquer coating of gift wrapping papers, etc.	USA, Europe, Australia
	metallizing of paper, coating with base lacquer, or transfer metallizing	Europe, USA
Films	cross-linking of pressuresensitive adhesives	Europe, Japan, USA
	production of films from liquid phase	Europe
	antistatic coatings	USA
	cross-linking of flock adhesives	USA
	antiscratch coating of, e.g., telephone cards	Europe
	magnetic coating on diskettes and video tape	USA, Japan
	lamination of films on plastic, etc.	USA
Printing media		
	web offset printing, wet-on-wet	Australia, Asia, Japan, Europe, USA, South America
	recoating in high-speed gravure printing	USA
	steel engraving	USA

Table 6. Comparison of peroxide and radiation-induced cross-linking

Property	Peroxide cross-linking	Radiation-induced cross-linking
Process temperature	high temperature (polymer melt)	normal temperature; only slight temperature raise by absorption of radiation
Energy consumption, ratio peroxide – radiation	5	1
Cross-linking rate	depends on peroxide concentration	depends on dose rate
Production costs, ratio peroxide – radiation	2.0 – 2.5	1 (1.5 – 2 $/kg)
Range	limited to polymers that can be extruded up to 250 °C	nearly universal
Crystallinity of polymer	diminished	unaffected
Gel content of cross-linked polymer	virtually always > 70 %, little possibility of variation	controllable through radiation dose

which process is dominant will depend on the elastomer, its composition, the type and quantity of additives, and the radiation dose. The required cross-linking dose is a function of the elastomer and the additives. The changes in elastomer properties due to vulcanization by heat treatment and radiation cross-linking are often very similar. Irradiation, however, permits homogeneous cross-linking in a simple way. More intense cross-linking at the surface, which can occur in heat treatment because of poor thermal conductivity of the elastomer, is not a problem with radiation cross-linking if sufficiently high electron energies are employed.

The possibility of using radiation cross-linking for parts of articles, as in tire manufacturing, leads to technological advantages in the production process, and savings of materials. Natural rubber can be replaced by synthetic material with no loss of quality. In addition, rates for production of, e.g., calendered rubber film are an order of magnitude higher than those in thermal vulcanization.

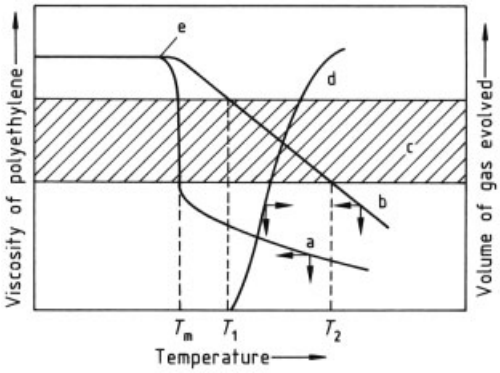

Figure 10. Radiation-cross-linked polyethylene foam; principle of PE expansion, according to [151]
a) PE before cross-linking; b) Cross-linked PE; c) Optimal viscosity; d) Decay of blowing agent; e) Melting point

1.3.2.3. Radiation-Cross-Linked Foam Polyethylene [150], [151] (see also → Foamed Plastics)

More than half of foamed polyethylene is produced with the aid of electron beams for cross-linking. Four processes are available for radiation cross-linking that have been developed by the following companies: Sekisui Chemical, Japan; Toray Industries, Japan; BASF, Germany; and Expanded Rubber and Plastics, United Kingdom.

In the United States, *slightly expanded* polyethylene foam (two-fold expansion) has been used in cable insulation. The compounds are foamed right on the wire and then cross-linked by radiation (Rayfoam L, Cheminax, by Raychem Corp.).

The foaming of polyethylene into a *highly expanded product* has proved difficult for a long time especially as the viscosity of polyethylene drastically decreases above its melting point. The melt viscosity of polyethylene drops abruptly above the metling point T_m, while that of cross-linked polyethylene declines gradually between T_1 and T_2 (Fig. 10). In the cross-linked state, polyethylene can be expanded in the much broader temperature range from T_1 to T_2. The expandability of cross-linked polyethylene at elevated temperature is a function of the dose. At 70 – 130 °C optimal percent expansion is obtained in the range of ca. 50 – 80 kGy. The expansion ratio of polyethylene is highest (15 – 20) for a gel fraction of 20 – 40 %. The expandability (foamability) of polyethylene bubbles also depends on the degree of cross-linking. Good foams are obtained after a cross-linking dose between 30 and 60 kGy when the pressure in the foam bubble is ca. 50 g/cm^2. The size of the individual polyethylene foam cell also depends on the gel fraction. Azodicarbonamide has proven suitable as a blowing agent.

1.3.2.4. Shrink Materials [152–154]

Polyethylene can easily be cross-linked into a product that will become rubber-elastic when heated. The structure of normally entangled polyethylene molecules includes crystalline and amorphous regions. Above the melting point of polyethylene, the crystalline regions vanish. Cross-linking (probably of the amorphous regions) results in the formation of a three-dimensional network. If the cross-linked polyethylene is heated above the melting point, it becomes rubber-elastic and can be mechanically deformed; if it is cooled in the deformed state to a temperature below the melting point, the crystalline portions form again and the material retains its deformation (frozen-in state). If the polyethylene is again heated above the melting point, it recovers from the deformation, taking on its original form (memory effect), i.e., it shrinks when it had previously been expanded.

Production of Shrink Tubes. First a tubular film is made with an extruder or the shaped part is fabricated with suitable pressing dies. The mold is then homogeneously irradiated with electrons (1.0 – 2.0 MeV for polyethylene tubing, 0.5 – 1.0 MeV for polyethylene films or web material) at a dose of 200 – 300 kGy. Excessive doses should not be applied the material properties so as to avoid deterioration of, e.g., poorer aging resistance. The tube is then expanded with heating and subsequently cooled.

1.3.3. Industrial Application of Radiation-Induced Graft Copolymerization [69], [81–83], [155–162]

Radiation-induced grafting is taking up a prominent role in industrial applications. Although large-scale applications are rare, a number of

special processes has been developed, most of them using electron irradiation to initiate grafting. High dose rates, controllable electron penetration depth, and the possibility of initiating the graft reaction at room temperature and atmospheric pressure are advantages of the electron-beam process. In most cases, the preirradiation or simultaneous method is employed. The flowsheet of a tpyical grafting apparatus for flexible polymers is shown in Figure 11. Properties that can be improved through grafting of polymer surfaces include: wettability, adhesion, printability, surface conductivity, biocompatibility, flammability, water absorption, resistance to certain chemicals, thermal stability, mechanical properties, chemical compatibility, hydrophilic and hydrophobic qualities. Table 8 lists some important applications.

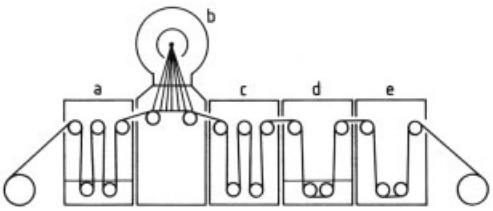

Figure 11. Grafting apparatus (preirradiation or simultaneous method)
a) Impregnation with monomer solution; b) Irradiation apparatus; c) Grafting section; d) Rinsing – washing; e) Drying

2. Irradiation of Foods and Feeds

2.1. Introduction

Many of the traditional food-processing methods employ energy in one form or another. Examples are the use of conventionally generated heat, microwaves, or high-frequency electromagnetic waves in pasteurization and sterilization, and the use of solar radiation in drying. The irradiation of *foods*, utilizes the energy of ionizing radiation. Gamma rays, X-rays, and electron beams are all in practical service or under discussion at present. The Codex General Standard for Irradiated Foods, which has found wide international acceptance, states [163]:
"The following types of ionizing radiation may be used:

a) Gamma rays from the radionuclides ^{60}Co or ^{137}Cs;
b) X-rays generated from machine sources operated at or below an energy level of 5 MeV;
c) Electrons generated from machine sources operated at or below an energy level of 10 MeV."

The standard further provides that: "The overall average dose absorbed by a food subjected to radiation processing should not exceed 10 kGy."

The irradiation of *feeds* is carried out with the same radiation sources, except that doses higher than 10 kGy are permitted.

The list of references includes a selection of monographs [164–174].

2.2. Capabilities of Food and Feed Irradiation

The irradiation of foods and feeds can be practiced

1) To prolong shelf life
2) As a quarantine measure
3) To control pathogenic microorganisms and pathogenic parasites

Table 9 lists a number of important applications along with the radiation doses involved. The use of radiation doses higher than 10 kGy with foods is restricted to just a few cases, such as the irradiation of hospital meals for patients with immune system impairments and the irradiation of spices and dried vegetables up to maximum doses of 30 kGy, which is allowed in the United States.

The radiation treatment causes only a slight temperature rise in the treated product, so this method can be used to kill microorganisms even in frozen foods. In addition, an effective radiation dose can be administered through most packaging materials, so that the risk of recontamination after irradiation is virtually eliminated.

2.2.1. Prolongation of Shelf Life

While extending the shelf life of foods may not appear to be an urgent problem in the industrialized countries, the situation is more dramatic

Table 8. Radiation-induced graft copolymerization of selected polymer–monomer systems

Polymer	Monomer	Properties modified
Polyethylene	acrylic acid	hydrophilicity, wettability, antistatic property
	acrylamide	adhesion, permeability, permselectivity
	styrene	mechanical properties
	acrylates, methacrylates	surface structure, functionalization
Poly(vinyl chloride)	acrylic acid, methacrylic acid, polyfunctional methacrylates	adhesion, abrasion resistance
	styrene	impact strength
	acrylonitrile	thermal stability
Poly(ethylene terephthalate)	acrylic acid	hydrophilicity, surface structure
	acrylamide	biocompatibility
	poly(ethylene glycol) acrylate	antistatic properties
Polyamide	acrylic acid	hydrophilicity, antistatic properties
	acrylonitrile, vinyl acetate, methacrylate	processability (textiles), water vapor permeability
	hydroxyethyl methacrylate (HEMA)	functionalization, biocompatibility
Polyacrylonitrile	acrylic acid	moisture absorption, antistatic properties
Polypropylene	styrene, acrylic acid, methacrylates	functionalization, ion exchange
Cellulose	styrene, acrylic acid, HEMA	functionalization
Polyurethanes	*N*-vinyl-2-pyrrolidone, HEMA, acrylamide	biocompatibility
Fluoropolymers	styrene, *N*,*N*-diethylacrylamide	functionalization, hemocompatibility
	N-vinylpyrrolidone	membranes

Table 9. Purposes of food irradiation according to [169]

Food	Main objective	Means of attaining objective	Dose, kGy
Low-dose treatments (up to about 1 kGy)			
Potatoes, yam tubers, onions, garlic, ginger, chestnuts	extension of storage life	control of sprouting and germination	0.05 – 0.15
Certain fruits and vegetables	improved keeping properties	delay in ripening and ageing	0.25 – 1
Cereal grain, flour, dried fruits and vegetables, nuts, pulses	prevention of losses caused by insects	killing or sexual sterilization of insects	0.2 – 0.7
Fruits	prevention of spreading of pests; quarantine treatment	killing or sexual sterilization of insects	0.2 – 0.7
Pork, beef, fish	prevention of parasitic disease transmitted through food	destruction of parasites (e.g., *Trichinella spiralis*)	0.3 – 0.5
Medium-dose treatments (about 1 – 10 kGy)			
Certain fruits and vegetables, sliced bread	improved keeping properties	reduction of bacteria, molds, and yeasts	1 – 3
Meat, poultry, fish	improved refrigerated storage	reduction of certain microorganisms (growing at refrigeration temperatures)	1 – 5
Meat, poultry, eggs, egg powder, froglegs, frozen seafood, and other foods	prevention of foodborne diseases	destruction of *Salmonella shigella*, *Campylobacter*, *Vibrio*, *Yersinia*, and other nonspore-forming pathogens	3 – 10
Spices, dried vegetables, and other food ingredients	prevention of contamination of food due to the ingredients	reduction of microorganisms in the ingredients	3 – 10
High-dose treatments (about 10 – 45 kGy)			
Meat, poultry	long-term storage without refrigeration	destruction of spoilage organisms and pathogens, including spore-formers	25 – 45
Hospital meals	to supply patients with sterile meals	destruction of spoilage organisms and pathogens, including spore-formers	25 – 45

in some of the developing countries. For example, one African study estimated that losses of foods amounted to more than 35×10^6 t in Africa during the year 1980 [175]. Approximately 1 t of food will nourish one person for one year; hence losses in Africa represent the food needed by at least 35×10^6 persons a year.

According to WHO research, the estimated loss of grains and legumes from storage in the developing countries is at least 10 %. In the case of noncereal products, vegetables, and fruits,

the losses due to microbial contamination and spoilage are estimated at up to 50 %. Products such as dried fish are reported to be subject to 25 % losses due to insect infestation, with an additional 10 % loss by spoilage.

Inhibition of Germination and Sprouting. Treatment with a low radiation dose prevents the germination and sprouting of many products, such as potatoes, yams, onions, garlic, ginger, and chestnuts. Irradiation also prevents germination of barley during malting, with a significant economic benefit; in contrast, very low doses of radiation stimulate the germination of barley.

Disinfestation. Irradiation at a relatively low dose kills or sterilizes all developmental stages of the common organisms that infest grains, including eggs laid in the grain. Dried fruits, vegetables, nuts, and dried fish can also be protected from insect losses by irradiation, even after packaging.

Prolongation of Shelf Life of Easily Spoiled Foods. The shelf life of many fruits and vegetables, as well as those of meat, poultry, fish, and other seafood, can be substantially extended – if not doubled – through irradiation alone, or through a combination of cooling with doses of radiation low enough to leave the flavor and other qualities of the product unchanged. In the case of fruits that are sensitive to high radiation doses (e.g., mangoes), a combination of gentle heat treatment (dipping in hot water) and low-dose irradiation is successfully employed.

Retardation of Ripening and Aging of Fruits and Vegetables. Low radiation doses retard the ripening process and/or the ageing of some fruits and vegetables, thus extending their shelf life. This technology can be used for instance, with mangoes, bananas, mushrooms, and asparagus.

Killing of Parasites. Examples of pathogenic parasites that are inactivated by irradiation are the nematode *Trichinella spiralis*, which occurs in pork and causes trichinosis, the beef and pork tapeworms, the protozoa in pork that cause toxoplasmosis, and various worms that can infest fish. All these organisms can be rendered harmless with relatively low doses of radiation.

2.2.2. Quarantine Measures

Food irradiation may heavily increase the growing trade in tropical and subtropical fruits, such as citrus, mangoes, and papayas. Because irradiation prevents the introduction of harmful insects without leaving residues in the fruits, this treatment offers an alternative to fumigation as a way of complying with quarantine regulations in many countries. Irradiation can even destroy flies and weevils that live in mango stones.

2.2.3. Control of Pathogenic Organisms

According to WHO research, diseases caused by foods represent one of the most widespread threats to human health and a significant constraint on economic productivity. The principal sources of such diseases are foods of animal origin, a high percentage of which are contaminated by pathogenic bacteria. Meat and meat products are the main cause of infections with trichinosis and toxoplasmosis. According to WHO estimates more than 10^9 cases of acute diarrhea occur annually in children younger than five years in the developing countries. Some 5×10^6 children die of this disease, which is largely caused by contaminated foods and contaminated water. This is not solely a problem of the developing countries but also affects the industrialized countries, as shown by the increase of Salmonella-caused infectious enteritis in the Federal Republic of Germany from about 30 500 reported cases in 1985 to about 114 500 in 1991.

Most pathogenic microorganisms and virtually all pathogenic parasites can be eliminated by irradiation. For example, only relatively low radiation doses are required to destroy the nonspore-forming bacteria *Salmonella*, *Campylobacter*, *Listeria*, and *Yersinia*.

A problem may, however, be posed by microorganisms that survive a treatment with a relatively low radiation dose, for example the spores of *Clostridium* and *Bacillus* species. There is a danger that certain applications of food irradiation will suppress spoilage- causing microorganisms to the extent that the extra shelf life of the treated food will allow a species such

as *Clostridium botulinum* to produce a dangerous toxin level. This threat can be averted only through special precautions such as cold storage.

2.3. History

The bactericidal action of ionizing radiation was described around 1900, giving rise to the notion of using such radiation for preservation purposes. The first patent on food irradiation was published in 1905 (British Patent 1609). This patent stated how the quality and shelf life of foods could be improved through the application of ionizing radiation. Concerned chiefly with cereals, the patent stressed that one advantage of using ionizing radiation lies in the replacement of the corresponding chemical treatments. The use of ionizing rays to control trichinae in pork was first discussed in the United States in 1921. A further advance was a French patent issued in 1930 (O. Wüst), which proposed the treatment of foods with ionizing radiation as a method of controlling bacteria. None of these initiatives led to practical applications since no radiation sources with proper dose-rates were available before World War II.

After this period the situation changed when, parallel to the renewed interest in food irradation, the first powerful electron accelerators were built. In a research program on the use of ionizing radiation to preserve foods, the U.S. Atomic Energy Commission also developed powerful γ sources. After it had become increasingly clear that spent fuel assemblies from nuclear reactors are unsuitable for food irradiation, several U.S. universities installed the first ^{60}Co sources at the beginning of the 1960s. Extensive development of food irradiation by the U.S. Army in the 1950s, culminating in the establishment of a large army food irradiation laboratory at Natick, Massachusetts, in 1961 – 1962, also led to early studies along the same lines in many other countries as early as the 1950s.

The first commercial use of food irradiation took place in the Federal Republic of Germany in 1957, when a spice company near Stuttgart began irradiating its products with electrons from a Van de Graaff generator. This plant had to be shut down in 1959 when new legislation banned the treatment of foods with ionizing radiation.

In 1966 the first international symposium on food irradiation was held in Karlsruhe by the IAEA. At that time, the irradiation of a few foods was permitted only in Canada, the United States, and the former Soviet Union, but the food industry was not taking advantage of such approvals. An international project to identify any health risks involved in consuming irradiated foods was started in 1970. More than 20 countries took part in this study which was carried out under the auspices of IAEA, FAO, OECD, and WHO. In November 1980, an FAO – IAEA – WHO committee of experts stated that food irradiation up to doses of 10 kGy is generally unobjectionable from the toxicological standpoint [175]. Given this assessment, it was only logical that the committee did not deem further toxicological studies of irradiated food necessary. Since the recommendation of these international agencies, the number of approvals of irradiated foods has been growing year by year. Accompanying this trend is a steady growth in the use of the method in food technology.

2.4. Effects of Food Irradiation

2.4.1. Induced Radioactivity

High energetic ionizing radiation can initiate nuclear reactions. Such reactions do not occur below a certain threshold energy, and the energy of γ-rays from ^{60}Co and ^{137}Cs sources is too low to produce radioactive substances in foods and feeds. Such an effect cannot be ruled out when higher-energy X-rays and electron beams are employed. The total quantity, however, of induced radioactive substances, given the maximum energies and doses used for food irradiation, is several orders of magnitude less than the quantity of naturally radioactive substances always present in foods. Accordingly, induced radioactivity is not significant for food irradiation.

2.4.2. Chemical and Sensory Changes

Irradiation causes chemical changes in the treated foods and feeds. Much is known about the substances produced and about the factors

that influence the formation of radiolysis products, such as temperature, moisture, and the effect of oxygen. The most important factor is the radiation dose. At the low doses used to control insects in grain (< 0.5 kGy), it is very difficult to detect chemical changes in the irradiated product. On the other hand, many chemical changes become detectable at higher doses in the range of a few kilograys.

An extremely important point should be kept in mind: In isolated form, food constituents such as amino acids, vitamins, and carbohydrates are severely attacked by irradiation; they suffer less damage when treated with ionizing radiation in the complex and protective matrix of an intact food. In addition, most radiolysis products are not radiation-specific and are thus found in the unirradiated foods as well. For example, the irradiation of pure triglycerides generates hundreds of radiolysis products, but the quantity of any one component is so small that no significant toxicological potential can be attributed to these reactions – even though some of the products occurring are exceedingly harmful to health when present in higher concentrations. Cyclobutanones, produced when fats are irradiated, are one of the very few groups of substances now suspected to be formed solely by irradiation.

The chemical changes caused by irradiation can also lead to perceptible effects in flavor, smell, color, and texture. Some foods, for example, and some other dairy products are very sensitive to even low radiation doses. Even doses of ca. 0.1 kGy can cause a negative flavor change in milk, making it unacceptable to the consumer.

The treatment of *meat* with high radiation doses can also give rise to undesirable flavor changes. It appears that these changes occur in lean meat rather than fat. Pork is less subject to these changes than is beef or veal, presumably because of its higher fat content. The negative flavor changes are most marked just after irradiation, diminishing or vanishing after storage or during cooking. The changes are less pronounced at a low irradiation temperature. The color of meat can also change when it is irradiated. For all these reasons, radiation doses applied to meat generally do not exceed a few kilograys.

The upper dose limit for *fruits and vegetables* in practice is set by the strength of the plant tissue. Doses of 1 – 3 kGy are used, depending on the product, and many fruits become soft even at these doses. This is not a direct effect of the radiation but a physiological reaction, the breakdown of the cell membrane by enzymes. The softening is not immediately observed but begins hours and sometimes days after the exposure to radiation.

2.4.3. Changes in Nutritive Value

Detailed studies have shown that the metabolizable energy of macronutrients (carbohydrates, fats, proteins) is not affected by irradiation up to doses of 10 kGy and even far above this value. Essential amino acids, essential fatty acids, minerals, trace elements, and most vitamins suffer no significant degradation, provided the irradiation is restricted to foods suitable for this treatment and is performed in a controlled manner (e.g., sensitive products are irradiated at low temperature with at least partial exclusion of atmospheric oxygen). Some vitamins show great losses even under these conditions. An example is the roughly 30 % loss of vitamin B_1 when chicken meat is irradiated (dose 5 kGy, irradiation temperature − 15 °C). Such losses are not, however, greater than those occurring in other, generally accepted food-processing methods such as preservation by heat.

2.4.4. Effects on Microorganisms

Microorganisms, especially gram-negative bacteria such as *Salmonella*, can be killed by irradiation. The spores of bacteria, however, are killed only at higher doses; this means that botulism, a commonly lethal food-born illness, cannot always be prevented by irradiation.

Regardless of the number of microorganisms in food, a certain fraction of the microbial population is killed by a given radiation dose. It follows therefore that the number of microorganisms still present in the product after irradiation will be larger, the more organisms were present before the treatment. For this reason, irradiation (like any other food preservation technique) cannot replace good hygiene during food production and processing. Once spoilage has begun, irradiation cannot reverse it.

The degree of destruction of microorganisms depends, as do many other radiation effects, on

a number of factors, one of which is the temperature during the radiation treatment. Microorganisms become more sensitive to radiation at elevated temperature. Many organisms also become easier to control if the moisture content in the food is high.

Concerns that irradiation might produce radiation-resistant microorganisms or other dangerous mutations have not been eliminated. All experimental studies indicate that microorganisms surviving radiation treatment are injured and then react more sensitively to factors known to have negative effects on microbial growth. In other words, surviving microorganisms are more easily destroyed in cooking.

Approximate radiation doses (kGy) needed to cut the viable populations of various microorganisms by a factor of a million are listed below [165]:

Gram-negative bacteria	
Escherichia coli	1.5 – 3.0
Salmonella spp.	3.0 – 5.0
Shigella spp.	1.0 – 2.0
Acinetobacter spp.	0.5 – 1.0
Pseudomonas spp.	0.5 – 2.0
Proteus	0.5 – 1.0
Vibrio parahaemolyticus	0.5 – 1.0
Moraxella	5.0 – 7.0
Gram-positive bacteria	
Bacilli (vegetative forms)	2 – 3
Bacillus cereus	20 – 30
Bacillus stearothermophilus	10 – 20
Clostridium botulinum A and B	10 – 30
Clostridium botulinum E	10 – 20
Clostridium perfringens	20 – 30
Leuconostoc	0.5 – 3.0
Lactobacillus	2.0 – 7.5
Micrococcus spp.	2 – 5
Micrococcus radiodurans	30
Micrococcus roseus	30
Staphylococcus aureus	0.5 – 5.0
Streptococcus	5.0 – 7.5
Streptococcus faecalis	10 – 20
Molds and yeasts	
Aspergillus	1.5 – 5.0
Penicillium	0.5 – 2.0
Saccharomyces	5 – 10
Viruses	
Various types	30
Foot-and-mouth disease	10 – 30

2.4.5. Toxicological Consequences

Many toxicological studies of irradiated foods have been carried out since the 1950s. Foods on which toxicological investigations have been performed are listed in the following [165]:

1) *Fruits:* Apples [b]; apricots [a]; bananas [a]; dates [c]; tangerines (mandarins) [b]; mangoes [c]; oranges [a]; papayas [c]; peaches [a]; plums [b]; strawberries [c]
2) *Vegetables:* Carrots [a]; celery root [a]; cauliflower [a]; lettuce [a]; meadow mushrooms (champignons) [b]; onions [c]
3) *Grains:* Corn (maize) [a]; rice [c]; wheat [c]
4) *Legumes:* Beans [b]; lentils [a]; peas [b]
5) *Spices and prepared spices:* Garlic (powder) [a]; mixed spices [b]; onion (powder) [b]; paprika [b]; pepper (black) [b]
6) *Miscellaneous plant-derived foods:* Cacao (cocoa) beans [b]; potatoes [c]; walnuts [a]
7) *Fish and fish products:* Carp [a]; catfish [c]; codfish [c]; fish paté (kamaboko) [c]; flounder [b]; herring [b]; mackerel [c]; haddock [c]; plaice [b]; European trout [a]
8) *Shellfish:* Soft calm (gaper) [b]; shrimp [b]
9) *Meat* [d]*:* Bacon [c]; beef and beef products [c]; prepared foods containing meat [b]; ham [c]; horse meat [b]; Mixed offal [b]; pork [c]; poultry [c]

[a] This food has not been extensively tested.
[b] This food has been less thoroughly tested; normally, no long-term study has been performed.
[c] This food has been thoroughly tested; one or more long-term studies have been performed.
[d] Some of the studies considered were performed for special purposes with very high radiation doses.

While most of these studies do not point to special risks of irradiated foods, some authors have published alarming results. The main problem areas relate to observations on children and to feeding tests with rats, mice, and monkeys, which have indicated a slightly increased incidence of polyploidy in bone marrow cells or peripheral lymphocytes after the children and animals had been fed with freshly irradiated wheat (< 1 kGy) for several weeks. These changes did not, however, occur when the wheat had been stored for about three months after the radiation treatment. Other authors have not confirmed these findings, so that nowadays most toxicologists do not accept the argument that polyploidy results from the consumption of irradiated food.

In contrast, there are many extensive studies that support the conclusion that irradiated foods

are unobjectionable from the health standpoint. In a French study, for example, nine radiolysis products identified in irradiated starch were fed daily to rats in a quantity 800 times the presumed rate of intake given normal daily consumption of irradiated starch. No detrimental effects could be found.

The fraction of irradiated food in most animal feeding experiments has been 30 % of the daily diet. In several studies, multiple generations of animals were raised on an exclusive diet of irradiated feed. Neither carcinogenic nor other toxic effects were observed.

A US study with irradiated chicken meat is important because of the scope of research and the high dose administered (58 kGy). Dogs, mice, and fruit flies were fed on irradiated meat, on heat-sterilized meat, and on enzyme-inactivated meat. Although the dose was far higher than is permitted for the irradiation of chicken meat (max. 10 kGy), no detrimental effects of the radiation treatment were found.

A study in the People's Republic of China, in which volunteers consumed various irradiated foods over a period of 7 – 15 weeks, also showed no negative health effects and no chromosomal damage.

2.5. Regulations on the Irradiation of Foods and Feeds

2.5.1. The European Community

Because the member countries of the European Community have widely varying laws governing the irradiation of foods (only Belgium, Denmark, France, Italy, The Netherlands, Spain, and the United Kingdom permit the irradiation of some foods, and not the same foods), the EC is attempting to create a unified regulation. What compromise will be arrived at cannot yet be foreseen, since the question of permitting irradiated foods is disputed not just between the member countries but also between the Commission of the European Communities in Brussels and the European Parliament. While countries such as the United Kingdom and France, together with the EC Commission, seek a relatively broad approval of irradiated foods, other member countries such as Luxembourg, Denmark, and Germany, with the European Parliament, oppose more liberal approval of irradiation and wish either to allow the irradiation of spices and dried vegetables only, or else to dispense with this type of food processing altogether.

2.5.2. Federal Republic of Germany

As of the end of 1992, the Food and Commodities Act (Lebensmittel- und Bedarfsgegenständegesetz, § 13 LMBG) prohibited the sale of *foods* treated with ionizing radiation. No use has been made of the possibility – allowed for in the Act – of gaining approval for the irradiation of foods either through regulatory proceedings or by the granting of variances (except as allowed for monitoring and measuring purposes by the Food Irradiation Regulation of December 19, 1959, as amended up to May 16, 1975). On the international level the federal government has pressed and continues to press for a complete ban on food irradiation.

The creation of the European common market on January 1, 1993, changed this situation. Under the Act to Modify the Provisions of Legislation on Veterinary Medicine, Foods, and Animal Husbandry (Gesetz zur Änderung veterinärrechtlicher, lebensmittelrechtlicher und tierzuchtrechtlicher Vorschriften) passed on December 18, 1992, irradiated foods lawfully produced in another member country of the European Community and lawfully traded there, or originating in a third country and lawfully traded in a member country of the European Community, may be introduced into Germany and traded there. Any person who whishes to place the irradiated food on the market in Germany must, however, apply to the agency having jurisdiction, the Federal Ministry for Health in Bonn, for permission to do so. In case of an affirmative decision, such decision is announced in the Bundesanzeiger (official gazette); any permission is to be linked with an appropriate labeling requirement.

As in virtually every country in the world, the irradiation of animal *feeds* is permitted in the Federal Republic of Germany.

2.5.3. Other Countries

In addition to the EC member countries listed, some 30 other countries permitted the irradiation

of certain foods in the 1980s. These include Argentina, Brazil, Canada, Chile, the People's Republic of China, Hungary, India, Israel, Japan, Poland, South Africa, Thailand, and the United States.

2.5.4. International Organizations

The WHO, FAO, and the IAEA advocate the irradiation of food. The three international agencies promote the use of food irradiation all over the world; the IAEA, above all, has supported applied research in this field, especially with the participation of developing countries. The WHO also takes an active role in informational efforts aimed at gaining acceptance for food irradiation. One example is the WHO brochure [164]; another comes from WHO's 10 Golden Rules for hygienic food preparation, in which Rule 1, "Do not eat certain foods raw," is followed by the text, "If you have the choice, you should use fresh or frozen poultry that has been treated with ionizing radiation."

2.5.5. Consumers

Food irradiation will find wider use only when products treated in this way are also accepted by the consumer. Predicting consumer reaction, however, is quite difficult, and so market tests have been performed in some countries for many years. Consumers are offered some properly labeled irradiated foods in comparison with unirradiated products. No such large-scale studies have yet been carried out in Germany. One IAEA summary of 1984 – 1990 market research on irradiated foods states that some 40 distinct irradiated products were offered to consumers in 13 countries over this period. Some 7000 t of irradiated products were test-marketed in this way. In most of the studies reported, a majority of consumers either preferred the irradiated products on account of their higher quality or at least expressed no reservations about purchasing them.

Consumer organizations in quite a few countries (including Germany) oppose food irradiation, sometimes vehemently. The cited studies, however, show that these organizations may take a much more critical view of food irradiation than are the individual consumers.

2.6. Industrial Irradiation of Foods and Feeds

2.6.1. Irradiation Plants (see also → Radionuclides, Chap. 5.1.3., → Radionuclides, Chap. 5.1.5.)

Food irradiation plants varies in design and physical operations performed, depending on the intended purpose. There are essentially tow classes, batch and continuous. In systems operating *batchwise*, a certain quantity of food is irradiated for a definite time. The irradiation cell is then unloaded and recharged. In a *continuous irradiation system*, the food is conveyed around the source at a certain speed in order to ensure that all the food receives the same desired dose. Batch systems are simpler in design and more flexible than continuous ones. Continuous systems, on the other hand, can handle large volumes of products, especially when an individual product must be treated with a set dose. The food-processing industry prefers continuous treatment for economic reasons.

Food irradiation ordinarily involves radiation doses of 50 Gy to 10 kGy, depending on the kind of food and the effect desired.

2.6.2. Costs

Costs of food irradiation range from a few cents to around \$ 0.75 per kilogram. This wide variation results from the many factors involved in the irradiation process. These include, among others, the dose administered, the volume and kind of product, and the type and efficiency of the radiation source. Further, the costs depend on whether the system is designed to irradiate a single food or a variety of foods. Shipping and packaging costs are additional. The construction of an industrial system costs ca. \$ 7×10^6.

2.6.3. Present Level of Use

Complete data on the commercial use of food irradiation in the *EC member countries* are unfortunately not available. With respect to quantity, France and The Netherlands were the leading users in 1990, with about 20 000 t each,

followed by Belgium with ca. 10 000 t (1988). The foods irradiated are mainly shrimp, poultry, frogs' legs, and dried foods such as spices, herbs, and powdered eggs. Outside the EC, irradiation is used principally in the *People's Republic of China*, with a total of ca. 40 000 t between 1984 and 1990, the leading product being a liquor made from sweet potatoes and garlic; *Japan*, with irradiation of ca. 15 000 – 20 000 t of potatoes per year; *South Africa*, with more than 20 000 t per year, chiefly fruits, spices, potatoes, dried vegetables, and onions; the *United States*, where the leading item is some thousands of tonnes of spices per year; and the former *Soviet Union* (Odessa, now Ukraine), with ca. 400 000 t of grain per year (disinfestation). In *Germany*, commercial irradiation of foods (except in scientific research) is permitted only for export and is practiced sporadically for this purpose. In the former German Democratic Republic, at least during one season, ca. 5000 t on onions for cooperative marketing were irradiated. The overall trend is toward increasing use.

The irradiation of *feeds* is commercially practiced in some countries, primarily in the breeding of laboratory animals. More exact information is lacking as to the quantities of feed irradiated.

2.7. Inspection of Irradiated Foods

Irradiated products must be monitored and inspected in several areas. Inspection the *irradiation plants* serves to maintain the prescribed treatment conditions, prevent the use of irradiation to replace hygienic practices in production, storage, and shipping, and prevent multiple irradiation. In *trade*, proper documentation either in accompanying papers or by labeling must be verifiable. Methods of detecting irradiated products are needed for this kind of control.

Tetsting and Analysis. Tests aimed at identifying irradiated foods were undertaken as early as the beginning of the 1960s. Truly successful methods, however, were not developed until the early 1980s. Work by the BGA in Berlin, for example, had the goal of detecting irradiated spices and dried vegetables with the aid of chemiluminescence and thermoluminescence tech-niques; this project led to the establishment of the first four official food inspection methods for irradiated products. Germany thus became the first country in the world to have officially acknowledged techniques for the detection of irradiated foods. Another method for irradiated bone-containing meat products was adopted in 1991: electron spin resonance spectroscopy (ESR). Other methods to follow shortly will enable the effective inspection of virtually any food that is a candidate for irradiation. In meat products not containing bone, the detection of certain hydrocarbons in the fat fraction makes it possible to state unambiguously whether the product has been treated with ionizing radiation. The method can also be applied to other foods with sufficiently high fat contents. Examination of fish bones and nutshells (i.e., water-poor constituents of foods) by ESR spectroscopy can reveal the irradiation of these products. Finally, the thermoluminescence technique can be applied to any food contaminated with minerals (e.g., sand) or containing crystalline minerals, such as spices, herbs, fruits, vegetables, and mussels (shells). Intensive work on other techniques is under way, and it can be expected that the range of containing crystalline minerals, such as spices, herbs, fruits, vegetables, and mussels (shells). Intensive work on other techniques is under way, and it can be expected that the range of inspection methods will be largely complete in some two years.

Germany continues to be a world leader in this field. The methods cited have been used in a number of countries to identify irradiated products unlawfully offered for sale.

3. References

General References

1. P. Ausloos (ed.): *Fundamental Processes in Radiation Chemistry,* Interscience, New York 1968. A. Chapiro: *Radiation Chemistry of Polymeric Systems,* Wiley-Interscience, New York 1962.
2. A. Charlesby: *Atomic Radiation and Polymers,* Pergamon Press, Oxford 1960.
3. J. Silverman, A. R. van Dyken (eds.): *Radiation Processing,* Pergamon Press, Oxford 1977.
4. R. O. Bolt, J. G. Carroll: *Radiation Effects on Organic Materials,* Academic Press, New York 1963.

5. M. Dole: *The Radiation Chemistry of Macromolecules,* Academic Press, New York 1972.
6. J. E. Wilson: *Radiation Chemistry of Monomers, Polymers and Plastics,* Marcel Dekker, New York 1974.
7. G. Hughes: *Radiation Chemistry, Oxford Chemistry Series,* Clarendon Press, Oxford 1973.
8. J. Silverman, A. R. van Dyken (eds.): "Radiation Processing, Transactions of the First Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **9** (1977) 1 – 6.
9. J. Silverman (ed.): "Advances in Radiation Processing, Transactions of the Second Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **14** (1979) 1 – 6.
10. J. Silverman (ed.): "Trends in Radiation Processing, Transactions of the Third Internat. Meeting of Radiation Processing," *Radiat. Phys. Chem.* **18** (1981) 1 – 6.
11. V. Markovic (ed.): "Radiation Processing, Transactions of the Fourth Meeting on Radiation Processing," *Radiat. Phys. Chem.* **22** (1983) 1 – 6.
12. S. V. Nablo (ed.): "Progress in Radiation Processing, Proceedings of the 5th Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **25** (1985) 1 – 6.
13. F. M. Fraser (ed.): "Progress in Radiation Processing, Proceedings of the 6th Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **31** (1988) 1 – 6.
14. J. G. Leemhorst, A. Miller (eds.): "Radiation Processing: State of Art, Proceedings of the 7th Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **35** (1990) 1 – 6.
15. Y. Tabata, Y. Katsumura (eds.): "Proceedings of an Internat. Symp. on Radiation Chemistry of Polymers," *Radiat. Phys. Chem.* **37** (1991) 1.
16. D. R. Randell (ed.): *Radiation Curing of Polymers,* The Royal Society of Chemistry, London 1987.
17. R. L. Clough, Sk. W. Skalaby (eds.): *Radiation Effects on Polymers,* American Chemical Society, Washington D.C. 1991.
18. P. K. D. Oldring (ed.): *Chemistry and Technology of UV and EB Formulation for Coatings, Inks and Paints,* **vols. 1 – 4,** SITA Technology, Gardiner House, London 1991.
19. A. K. Pikaev: *Sovremennaya radiazionaya chimia (russ.),* **vols. 1 – 3,**Nauka, Moskau 1987.
20. A. K. Pikaev, S. A. Kabakchi, I. E. Makarov, B. G. Ershov: *Impulsnij radiolis i ero primenenija,* Atomisdat, Moskau 1980.
21. A. Heger (ed.): *Technologie der Strahlenchemie von Polymeren,* Akademie-Verlag, Berlin 1990.
22. A. Singh, J. Silberman (eds.): *Radiation Processing of Polymers,* Hanser Verlag, München 1992.

Specific References

23. P. Ausloos (ed.): *Fundamental Processes in Radiation Chemistry,* Interscience, New York 1968.
24. M. Burton, J. L. Magee (eds.): *Advances in Radiation Chemistry,* Wiley-Interscience, New York 1976.
25. A. Henglein, W. Schnabel, J. Wendenburg: *Einführung in die Strahlenchemie,* Akademie Verlag, Berlin 1969.
26. A. K. Pikaev: *Sovremennaya radiazionaya chimia (russ.),* Nauka, Moskau 1987.
27. J. M. Warman, K.-D. Asmus, R. H. Schuler, *J. Phys. Chem.* **73** (1969) 931.
28. J. H. Baxendale, F. Busi (eds.) : *The Study of Fast Processes and Transient Species by Electron Pulse Radiolysis,* Reidel, Dordrecht 1982.
29. M. Ebert (ed.): *Pulse Radiolysis,* Academic Press, New York 1965.
30. M. S. Matheson, L. M. Dorfman: *Pulse Radiolysis,* The MIT Press, Cambridge 1969.
31. A. K. Pikaev, S. A. Kabakchi, I. E. Makarov, B. G. Ershov: *Impulsnij radiolis i ero prinienenija (russ.),* Atomisdat, Moskau 1980.
32. A. Chapiro: *Radiation Chemistry of Polymeric Systems,* Wiley-Interscience, New York 1962.
33. R. O. Boll, J. G. Carroll: *Radiation Effects on Organic Materials,* Academic Press, New York 1963.
34. J. E. Wilson: *Radiation Chemistry of Monomers, Polymers and Plastics,* Marcel Dekker, New York 1974.
35. M. Dole (ed.): *The Radiation Chemistry of Macromolecules,* Academic Press, New York 1972.
36. Y. Tabata, Y. Katsumura (eds.): "Proceedings of an Internat. Symp. on Radiation Chemistry of Polymers," *Radiat. Phys. Chem.* **37** (1991) 1.
37. E. S. Huyser: *Free Radical Chain Reactions,* Wiley-Interscience, New York 1970.
38. A. Chapiro: in *Encyclopedia of Polymer Science and Technology,* vol. 11, Wiley-Interscience, New York 1969, pp. 702 – 760.

39. A. Chapiro, *Plaste Kautsch.* **11** (1964) 451.
40. A. Chapiro, *Radiat. Phys. Chem.* **14** (1979) 101.
41. K. Hayashi, S. Okamura, *Radiat. Phys. Chem.* **18** (1981) 1133.
42. A. Chapiro: Europäische Konferenz über Elektronenstrahlverfahren, AEG-Telefunken, Hamburg 1980.
43. S. H. Pinner: "Cationic Polymerization Initiated by High Energy Radiation," in P. H. Plesch (ed.): *The Chemistry of Cationic Polymerization,* Pergamon Press, London 1963, p. 611.
44. K. Hayshi: "Radiation-Induced Polymerization by Free Ions," in M. Haissinsky (ed.): *Actions Chimiques et Biologiques des Radiations,* Ser. 15, Masson, Paris 1971, p. 145.
45. K. Hayashi, *Radiat. Phys. Chem.* **18** (1981) 183.
46. A. Henglein, R. Schulz, *Z. Naturforsch. B Anorg. Chem. Org. Chem. Biochem. Biophys. Biol.* **B 9** (1954) 617.
47. R. B. Mesrobian, P. Ander, D. S. Ballantine, *J. Phys. Chem.* **58** (1954) 565.
48. A. Chapiro: "Polymerization en phase solide amourceés par les radiations," *Actions Chim. Biol. Radiat.* **10** (1966) 187.
49. A. Charlesby, *Radiat. Phys. Chem.* **37** (1991) 5.
50. I. Kaetsu et al., *Radiat. Phys. Chem.* **14** (1979) 737.
51. I. Kaetsu, *Radiat. Phys. Chem.* **25** (1985) 517.
52. A. Charlesby, *Proc. R. Soc. London Ser. A* **215** (1952) 187.
53. M. Dole, G. D. Keeling, P. G. Rose, *J. Am. Chem. Soc.* **76** (1954) 4304.
54. V. C. Karpov: *Sessiya Akad. Nauk S.S.S.R., po Mirnomu Ispolzovaniym Atomnoi Energii,* Academic Science of the U.S.S.R., Moscow 1958.
55. G. Ungar, *J. Mater. Sci.* **16** (1981) 2635.
56. Y. Tabata, *Radiat. Phys. Chem.* **14** (1979) 235.
57. A. Keller, G. Ungar: *Radiat. Phys. Chem.* **22** (1983) 155.
58. J. C. Randall, F. I. Zoepfl, J. Silvermann, *Radiat. Phys. Chem.* **22** (1983) 183.
59. M. H. G. Jellinek (ed.): *Degradation and Stabilization of Polymers,* Elsevier, Amsterdam 1983.
60. F. A. Maklis: Radiation Physics and Chemistry of Polymers, Wiley-Interscience, New York 1975.
61. O. Brede (ed.): "Radiation Induced Processes in Polyethylene," *ZfI-Mitt.* **132** (1987).
62. Y. Tabata (ed.): "Proc. of an Int. Symp. on Radiation Chemistry of Polymers," *Radiat. Phys. Chem.* **37** (1991) no. 1.
63. A. Charlesby: *Atomic Radiation and Polymers,* Pergamon Press, Oxford 1960.
64. A. Charlesby, S. H. Pinner, *Proc. R. Soc. London Ser. A* **249** (1959) 367.
65. A. Charlesby, *Radiat. Phys. Chem.* **18** (1981) 59.
66. A. Charlesby: "Radiation Chemistry," *Proceedings of the 5. Thihany Symp. on Radiation Chemistry,* Akademiai Kiado, Budapest 1983, p. 843.
67. B. J. Lyons, *Radiat. Phys. Chem.* **22** (1983) 135.
68. H. Kashiwabara, S. Shimada, Y. Hori, *Radiat. Phys. Chem.* **37** (1991) 43.
69. A. Heger: *Technologie der Strahlenchemie von Polymeren,* Akademie Verlag, Berlin 1990.
70. D. O. Hummel, *J. Polym. Sci. Polym. Symp.* **67** (1980) 159.
71. H. A. J. Battaerd, G. W. Fregear: *Graft Copolymers,* Wiley-Interscience, New York 1967.
72. V. Y. Kabanov, L. P. Sidorova, V. I. Spitsyn, *Eur. Polym. J.* **10** (1974) 1153.
73. J. L. Garnett, *Radiat. Phys. Chem.* **14** (1979) 79.
74. V. Stannett, *Radiat. Phys. Chem.* **18** (1981) 215.
75. A. Chapiro, *J. Polym. Sci. Polym. Symp.* **56** (1976) 431.
76. G. Ellinghorst, A. Niemöller, D. Vierkotten, *Radiat. Phys. Chem.* **22** (1983) 635.
77. V. Y. Kabanov, *Radiat. Phys. Chem.* **33** (1989) 51.
78. V. T. Stanett, *Radiat. Phys. Chem.* **35** (1990) 82.
79. P. K. D. Oldring (ed.): *Chemistry and Technology of UV and EB Formulation for Coatings, Inks and Paints,* SITA Technology, Gardiner House, London 1991.
80. D. R. Randell (ed.): *Radiation Curing of Polymers,* The Royal Society of Chemistry, London 1987.
81. S. V. Nablo (ed.): "Progress in Radiation Processing, Proceedings of the 5th Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **25** (1985) no. 1 – 6.
82. F. M. Fraser (ed.): "Progress in Radiation Processing, Proceedings of the 6th Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **31** (1988) 1 – 6.
83. J. G. Leemhorst, A. Miller (eds.): "Radiation Processing: State of Art, Proceedings of the 7th

Internat. Meeting on Radiation Processing," *Radiat. Phys. Chem.* **35** (1990) 1 – 6.
84. *Proceedings of the Radcure Europe '83 Conference,* Lausanne 1983.
85. *Proceedings of the Radcure Europe '85 Conference,* Basel 1985.
86. *Proceedings of the Radcure Europe '87 Conference,* Munich 1987.
87. *Proceedings of the RadTech '89 Europe,* France 1989.
88. *Proceedings of the RadTech '91 Europe,* Edinburgh 1991.
89. *RadTech '88,* North America Conference Papers, New Orleans 1988.
90. RadTech '90, North America Conference Papers, Chicago 1990.
91. C. E. Hoyle, J. F. Kinstle (eds.): "Radiation Curing of Polymeric Materials," *ACS Symp. Ser.* **417** (1990).
92. K. J. O'Hara, *beta-gamma* **1** (1991) 13.
93. E. Häring, *Ind. Lackierbetr.* **55** (1987) 51.
94. G. Strohner, *Plaste Kautsch.* **31** (1984) 233.
95. S. Paul: *Surface Coatings, Science and Technology,* Wiley-Interscience, New York 1985.
96. W. Brushwell, *Farbe Lack* **86** (1980) 888.
97. D. R. Bailey, A. Wright: *Electron Beam Curing,* D. Macpherson & Co., Warth Mills 1971.
98. E. P. Tripp, S. V. Nablo, *Radiat. Phys. Chem.* **14** (1979) 481.
99. O. French, B. S. Quintal, S. V. Nablo, *Radiat. Phys. Chem.* **18** (1981) 879.
100. J. N. Aaronson, S. V. Nablo, *Nucl. Instrum. Methods Phys. Res. B* (1985) no. 10/11 998.
101. S. V. Nablo, *Radiat. Phys. Chem.* **35** (1990) 46.
102. U. V. Läuppi, *Radiat. Phys. Chem.* **35** (1990) 30.
103. *Proceedings of the RadTech '91 Europe,* Edinburgh 1991.
104. A. Klein in *Symposium on Magnetic Media Manufacturing Methods,* Honolulu 1982.
105. T. J. Menezes: "Commercial EB Processing Applications, A Progress Report," *Polymers, Laminations and Coating Conference,* Boston 1984.
106. T. J. Menezes, *Radiat. Phys. Chem.* **35** (1990) 52.
107. J. Sakamoto, K. Mizusawa, M. Kashiwage, Y. Hoshi: *Radcure Europe '85,* Basel 1985.
108. P. Holl, *Ind. Lackierbetr.* **91** (1980) 10.
109. P. Holl, *Radiat. Phys. Chem.* **25** (1985) 665.
110. P. Holl, *Optik* **77** (1987) 99.
111. P. Holl, E. Föll, *beta-gamma* **1** (1988) 10.
112. R. Mehnert, P. Klenert, E. Hartmann, *Nucl. Instrum. Methods Phys. Res.* **B 68** (1992) 73.
113. P. Holl, G. Reuter, *Farbe Lack* **90** (1984) 18.
114. W. Baulmann, *Ind. Lackierbetr.* **57** (1989) no. 6 27.
115. I. Mc Bosch, J. D. Le Fors, T. Hetzel, *beta-gamma* **2** (1988) 4.
116. S. Fuijoka, I. Fujikawa, U. Ueno, A. Okamoto, *Radiat. Phys. Chem.* **18** (1981) 865.
117. L. B. Lueck, *Radiat. Phys. Chem.* **25** (1985) 581.
118. S. V. Nablo, *Radiat. Phys. Chem.* **25** (1985) 599.
119. T. O'Brien, *Radiat. Phys. Chem.* **25** (1985) 609.
120. A. M. Rodrigues, W. T. Newcomb, *Radiat. Phys. Chem.* **25** (1985) 617.
121. V. P. Talacker, *Radiat. Phys. Chem.* **35** (1990) 18.
122. Y. Tabata, *Radiat. Phys. Chem.* **35** (1990) 36.
123. S. V. Nablo, *Radiat. Phys. Chem.* **35** (1990) 46.
124. J. H. Bly, Radiation Dynamics Inc., *Techn. Inform.,* Melville August 1976.
125. H. R. Studer, *Radiat. Phys. Chem.* **14** (1979) 809.
126. E. Oda, *Radiat. Phys. Chem.* **18** (1981) 241.
127. R. Bäuerlein, H. D. Bickel, *Radiat. Phys. Chem.* **18** (1981) 837.
128. K. Posselt, *Wiss. Ber. Zentralinst. Festkörperphysik Werkstofforschung* **4** (1975) 1.
129. I. H. Bly, E. S. Brandt, R. G. Burgus, *Radiat. Phys. Chem.* **14** (1979) 931.
130. S. Tada, I. Uda, *Radiat. Phys. Chem.* **22** (1983) 575.
131. L. Wiesner, *Radiat. Phys. Chem.* **22** (1983) 587.
132. I. Uda et al., *Radiat. Phys. Chem.* **25** (1985) 855.
133. K. Ueno, *Radiat. Phys. Chem.* **35** (1990) 126.
134. N. Studer, *Radiat. Phys. Chem.* **35** (1990) 680.
135. N. Studer, *beta-gamma* **1** (1988) 39.
136. L. Wiesner, *beta-gamma* **2** (1988) 14.
137. K. Ueno, I. Uda, S. Tada, *Radiat. Phys. Chem.* **37** (1991) 89.
138. R. C. Becker, M. R. Cleland, I. P. Farrell, *Radiat. Phys. Chem.* **14** (1979) 353.
139. M. R. Cleland, *Radiat. Phys. Chem.* **18** (1981) 301.
140. J. H. Bly, *Radiat. Phys. Chem.* **18** (1981) 1331.
141. J. Mc Keown, *Radiat. Phys. Chem.* **22** (1983) 419.

142. B. J. Albertinsky, A. S. Ivanow, M. P. Svinin, M. T. Fedotov, *Radiat. Phys. Chem.* **22** (1983) 441.
143. M. S. De Wilton, *Radiat. Phys. Chem.* **25** (1985) 643.
144. T. Sadat, *Radiat. Phys. Chem.* **35** (1990) 616.
145. G. E. Hase, *Radiat. Phys. Chem.* **35** (1990) 619.
146. V. L. Auslender, *Radiat. Phys. Chem.* **28** (1986) 515.
147. V. L. Auslender, I. N. Meskkov, *Radiat. Phys. Chem.* **27** (1986) 627.
148. J. D. Hunt, G. Alliger, *Radiat. Phys. Chem.* **14** (1979) 39.
149. D. S. Pearson, *Radiat. Phys. Chem.* **18** (1981) 89.
150. C. J. Benning, *J. Cell. Plast.* **3** (1967) 62.
151. N. Sagane, H. Harayama, *Radiat. Phys. Chem.* **18** (1981) 99.
152. S. Ota, *Radiat. Phys. Chem.* **18** (1981) 81.
153. J. Kammel, L. Finegold, *Radiat. Phys. Chem.* **26** (1985) 685.
154. K. Ueno, *Radiat. Phys. Chem.* **31** (1988) 867.
155. V. T. Stannet, E. P. Stahel, *Rev. Nucl. Sci.* **21** (1971) 397.
156. A. Heger, *Dtsch. Textiltech.* **17** (1967) 307.
157. D. O. Hummel, *J. Polym. Sci. Polym. Symp.* **67** (1980) 159.
158. R. Teichmann, *Faserforsch. Textiltech.* **26** (1978) 66.
159. K. Hayakawa, K. Kawase, H. Yamahita, *Radiat. Phys. Chem.* **18** (1981) 479.
160. J. L. Garnett, S. V. Jankiewiez, R. Levot, *Radiat. Phys. Chem.* **25** (1985) 509.
161. D. Müller-Schulze, H. G. Thomas, *Radiat. Phys. Chem.* **35** (1990) 93.
162. P. A. Dworjanyn, J. L. Garnett: "Radiation Grafting of Monomers on Plastics and Fibers," in A. Sing, J. Silverman (eds.): *Radiation Processing of Polymers,* Hanser Verlag, München 1992.
163. Codex General Standard for Irradiated Foods, CAC/vol. 15, 1st ed., Codex Stan. FAO, WHO, Rome 1984, pp. 106 – 1983.
164. *Food Irradiation. A Technique for Preserving and Improving the Safety of Food,* World Health Organization, Geneva 1988.
165. Food Irradiation, Report of the Scientific Committee for Food (Statement from March 13, 1986), Commission of the European Communities, EUR 10 840 DE, Luxembourg 1989.
166. E. S. Josephson, M. S. Peterson (eds.): *Preservation of Food by Ionizing Radiation,* vols. 1 – 3, CRC Press, Boca Raton, Fla. 1982.
167. W. M. Urbain: *Food Irradiation,* Academic Press, Orlando, Fl. 1986.
168. S. Thorne (ed.): *Food Irradiation,* Elsevier Applied Science, London 1991.
169. J. F. Diehl: *Safety of Irradiated Foods,* Marcel Dekker, New York 1990.
170. D. E. Johnston, M. H. Stevenson (eds.): *Food Irradiation and the Chemist,* Royal Society of Chemistry, Cambridge 1990.
171. D. R. Murray: *Biology of Food Irradiation,* Research Studies Press Ltd., J. Wiley & Sons, Taunton 1990.
172. D. Robins: *The Preservation of Food by Irradiation. A Factual Guide to the Process and its Effect on Food,* IBC Technical Services Ltd., London 1991.
173. P. S. Elias, A. J. Cohen (eds.): *Recent Advances in Food Irradiation,* Elsevier Biomedical Press, Amsterdam 1983.
174. *Food Irradiation now,* Martinus Nijhoff/Dr. W. Junk Publishers, The Hague 1982.
175. Wholesomeness of Irradiated Food, *W.H.O. Tech. Rep. Ser.* **659** (1981).

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Radionuclides

Standard Article

Heribert Luig[1], Cornelius Keller[2], Walter Wolf[3], Jashovam Shani[4], Horst Miska[5], Alfred Zyball[6], Andreas Gervé[7], Alexandru T. Balaban[8], Albrecht M. Kellerer[9], Jürgen Griebel[10]

[1]Universität Göttingen, Abteilung Nuklearmedizin, Göttingen, Federal Republic of Germany

[2]Kernforschungszentrum Karlsruhe, Fortbildunsgzentrum für Technik und Umwelt, Eggenstein – Leopoldshafen, Federal Republic of Germany

[3]University of Southern California, Radiopharmacy Program, Los Angeles, CA 90033, United States

[4]University of Southern California, Radiopharmacy Program, Los Angeles, CA 90033, United States

[5]Ministerium des Innern, Rheinland-Pfalz, Mainz, Federal Republic of Germany

[6]Beta-Gamma-Service (BGS), Wiehl-Bomig, Federal Republic of Germany

[7]Institut für Angewandte Verschleißforschung (IAVF), Karlsruhe, Federal Republic of Germany

[8]Romanian Academy of Science, Bucharest, Romania

[9]Strahlenbiologisches Institut der Universität München, Federal Republic of Germany

[10]Strahlenbiologisches Institut der Universität München, Federal Republic of Germany

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a22_499
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Nucleons
1.2. Timeline of Discoveries in Radioactivity
1.3. Nuclides and Radionuclides
1.4. The Law of Radioactive Decay
1.5. Interaction between Radiation and Matter
1.5.1. Interaction between Gamma Quanta and Matter
1.5.2. Interaction between Electrons and Matter
1.5.3. Interaction between Positrons and Matter
1.5.4. Interaction between Alpha Particles and Matter
1.5.5. Interaction between Neutrons and Matter
1.5.6. Interaction between Neutrinos and Matter
2. Radioactive Elements
2.1. Promethium
2.2. Technetium
2.3. Polonium
2.4. Astatine
2.5. Radon
2.6. Francium
2.7. Radium
2.8. Actinium
2.9. Protactinium
3. Production of Artificial Radionuclides

3.1. Nuclear Reactors
3.2. Cyclotrons and Other Accelerators
3.3. Radionuclide Production from Reactors
3.4. Radionuclide Production from Cyclotrons
3.5. Radionuclide Generators
4. Detection of Ionizing Radiation and Radionuclides
4.1. Types of Ionizing Radiation
4.2. Detection Methods
4.2.1. Interactions Between Ionizing Radiation and Matter
4.2.1.1. Energy Loss of Charged Particles
4.2.1.2. Bremsstrahlung Production
4.2.1.3. Energy Loss of Neutrons and Gamma Quanta
4.2.2. Detector Types
4.2.2.1. Gas Ionization Detectors
4.2.2.2. Scintillation Detectors
4.2.2.3. Semiconductor Detectors
4.2.2.4. Basic Properties of the Detector Types
4.2.2.5. Integrating Detectors
4.3. Measurement of Activity and Dose
4.3.1. Gamma Spectroscopy for Radionuclide Determination
4.3.2. Other Methods of Radionuclide Determination
4.3.3. Determination of Surface Activities
4.3.4. Measurement of Photon Dose Rate
4.3.5. Area Monitoring and Personal Dosimetry
5. Uses
5.1. Industrial Applications of Radionuclides
5.1.1. Introduction
5.1.2. Radionuclides for Industrial Gamma Facilities
5.1.3. Design of Gamma Equipment
5.1.4. Irradiation Using Gamma Rays
5.1.5. Uses of Irradiation
5.1.5.1. Disinfection - Sterilization
5.1.5.2. Modification of Plastics
5.1.6. Wear Measurement with Radionuclides
5.1.7. Other Industrial Applications of Radionuclides
5.2. Uses in Hydrology
5.2.1. Snow and Ice
5.2.2. Transient Surface Water
5.2.3. Lakes and Reservoirs
5.2.4. Sediments in Hydrology
5.3. Uses in Agriculture
5.4. Uses in Environmental Studies
5.5. Uses in Recombinant DNA Research
5.6. Uses in Biochemistry
6. Syntheses of Radiolabeled Compounds
6.1. Introduction
6.2. Principles of Radiolabeling
6.2.1. Introduction
6.2.2. Methods of Radiolabeling
6.2.3. Stability of Radiolabeled Compounds
6.3. Radiolabeling with ^{14}C, ^{3}H, ^{32}P, and ^{35}S
7. The Biological Effects of Radiation and Radionuclides
7.1. Physical Radiation Effects
7.2. Molecular Radiation Effects
7.2.1. Radiation Damage to DNA
7.2.2. Repair of DNA Damage in the Cell
7.3. Cellular Radiation Effects
7.4. Radiation Effects on Tissue
7.5. Radiation Hazards

7.5.1. Radiation Carcinogenesis
7.5.2. Hereditary Radiation Effects
7.5.3. Teratogenic Radiation Effects
7.6. Radionuclides
7.6.1. Radionuclides of Concern in Radiation Protection
7.6.2. Exposure of the Population due to Natural and Artificial Radioisotopes
7.7. Basic Principles of Radiation Protection

[Top of Page]

1. Introduction

Heribert Luig

1.1. Nucleons

The term “radioactivity” refers to the transformation or decay of atomic nuclei, occurring spontaneously (i.e., with no external stimulus and at an unpredictable time) with the emission of radiation (high-energy quanta or particles). This process is also called the radioactive decay of radionuclides. Radionuclides have to acquire the excess energy required for spontaneous decay. This energy comes from one of three sources:

1. As a result of nuclear fusion; the process that has created the nuclei of the elements (natural radioactivity)
2. As a result of activation by natural ambient radiation, especially cosmic rays (natural radioactivity)
3. Through artificial activation or nuclear fission (artificial radioactivity)

Atomic nuclei (nuclides) consist of nucleons. A given type of nucleus consists of a specific number of nucleons. Nucleons are bound together by an attractive force called the strong interaction force (strong force) or nuclear force. This strong force is one of the four fundamental interactions: strong force, weak force, electromagnetic force, and gravitation.

The *strong force* is characterized by a limited range that, logically, corresponds to a nuclear diameter, roughly 10^{-13} cm. Outside this range, the strong force is equal to zero.

There are two types of nucleons; protons (p) and neutrons (n). Protons and neutrons are referred to as elementary particles. According to results from high-energy physics, however, these are not simple but structured systems. Their structural components consist of three quarks. The quarks that make up nucleons exist in two different forms, “up” (u) and “down” (d). Protons and neutrons have the following make-up:

p: (uud)
n: (udd)

The strong force acts directly between quarks, but indirectly between nucleons. Quarks bear electric charges.

The charge of the up quark: + 2/3 (positive) elementary charge. The charge of the down quark: – 1/3 (negative) elementary charge.

On balance, the proton has a single positive unit of charge whereas the neutron appears electrically neutral. The two types of nucleons are therefore similar. If a down quark in a neutron is replaced by an up quark, a proton is obtained. This is what takes place in β decay:

$$\mathrm{d} \rightarrow \mathrm{u} + \mathrm{e}^- + \bar{\nu} \tag{1.1}$$

In the transformation of a down to an up quark, the down quark emits an electron (e^-) and an antineutrino ($\bar{\nu}$).

From the viewpoint of the neutron, or of the nucleus balancing its nucleons, the process can be described as:

$$\mathrm{n} \rightarrow \mathrm{p} + \mathrm{e}^- + \bar{\nu} \tag{1.2}$$

The β decay process is governed by the *weak interaction*.

β Decay, or more correctly, the β transformation of the neutron, is an exothermic process. A free neutron (i.e., one not bound to other nucleons) has a mean lifetime of 930 s. In the terrestrial environment, however, a free neutron scarcely ever reaches the end of its life-time — apart from the fact that there are hardly any free neutrons — because, attracted by the strong force, it quickly attaches itself to an atomic nucleus as an additional nucleon. This process is called *neutron capture*, a name reflecting the viewpoint of the absorbing nucleus.

Nucleons do not form some kind of perpetual motion system. The inverse process, in which a proton changes into a neutron, is not exothermic. The free proton, the hydrogen nucleus, is stable.

Accordingly, the β transformation of the proton can be described as:

$$\mathrm{p} + \mathrm{e}^- \rightarrow \mathrm{n} + \nu \tag{1.3}$$

A proton incorporates an electron, a neutron is formed, and a neutrino is emitted. The above relation describes the β decay of the proton by *electron capture*. A proton in an atomic nucleus preferentially incorporates an orbital electron from the K shell, because K electrons have the highest probability of being present at the location of the nucleus. The β decay of the proton is described as:

$$\mathrm{p} \rightarrow \mathrm{n} + \mathrm{e}^{+} + \upsilon \tag{1.4}$$

Here a proton is transformed into a neutron, emitting a positron (an electron with a single positive charge — the "antiparticle" of the electron) and a neutrino. This process is energetically possible only in a nucleus whose energy is greater than the energy of the product nucleus by at least twice the rest-mass energy of a positron (= 2×511 keV), for according to Dirac's theory, the antimatter particle (the anti-electron) must be raised from the negative to the positive energy state that characterizes the surrounding matter.

The nuclear event governed by the weak interaction, generally called β transformation or β decay, thus occurs in three forms:

1. With emission of an electron (β^{-})
2. With emission of a positron (β^{+})
3. With capture of an electron (β^{+})

General Remarks on Nucleons. Neutrons and protons have approximately the same mass. The neutron is slightly, but crucially, heavier; according to Einstein's equivalence between mass and energy, the mass of the neutron is larger by an amount that is equivalent to the energy difference linked with the spontaneous exothermic conversion of a neutron to a proton.

Mass of proton. 938.2 MeV
Mass of neutron. 939.6 MeV

Both species of nucleons have an intrinsic angular momentum, spin 1/2; they are therefore subject to Pauli's exclusion principle as are the spin –1/2 electrons in the atomic shell; and, like the electrons, move along orbits having well-defined angular momenta, in "shells." The shell model of the atomic nucleus is described in Section Nuclides and Radionuclides. Both the proton and neutron have a magnetic moment; this is expected in the case of the proton and, in the case of the neutron, plausible because of the three-quark structure.

Essentially, a nuclide (an atomic nucleus) is uniquely characterized by the number *Z* of protons (*Z* = atomic number in chemistry) and the number *N* of neutrons. The total of these is *M*, the mass number of the nucleus:

$$M = Z + N \tag{1.5}$$

just as in chemistry an element is characterized by the number *Z* of orbital electrons.

An atom is said to be in an excited state if not all of its electrons are located in the lowest possible shells. Similarly for the nucleus: It is in an excited state if a neutron or proton is in a higher shell, otherwise it is in the most favored lowest-energy state known as the ground state. In addition to single-nucleon excitation, the nucleus can also be in a state of "collective" excitation, which consists of the rotational or vibrational states of all the nucleons collectively.

In an excited atom, if an electron from a higher shell drops into the free orbital position, a photon (X-ray quantum) or another orbital electron (Auger electron) is ejected.

In an excited nucleus, each nucleon that has acquired an "unduly" high energy drops back into its lower-energy position; the energy liberated is again emitted, either in the form of a photon (γ quantum) or an orbital electron (conversion electron). In "internal conversion", K conversion (with a conversion electron from the K shell) is again preferred (similarly to the β transition by electron capture), because K-shell electrons have the highest probability of being present at the location of the nuceus and thus falling victim to the energetic process involved in the restructuring of a nuclide.

The events in a γ transition are governed by the *electromagnetic force*. There are plenty of charged particles in the nucleus, since three quarks form one nucleon.

Usually, nuclides remain in a given excited state for only a short time. A state that is "longer-lived" from the metrological or other viewpoint is said to be metastable. For example, the symbol ^{99m}Tc denotes a nucleus with 43 protons and 56 neutrons in an excited state, in this instance 142 keV above the ground state. It also has a longer lifetime; in this instance, a half-life of 6.03 h.

The term "metastable" now survives only in the "m" of the symbol. The longer-lived state is usually called an *isomer*, by analogy with the term in chemistry: a different chemical structure with same empirical formula corresponds here to a different nuclear structure with identical constituents (*Z* protons and *N* neutrons).

The last process to be described is α *decay*. This type of decay, which happens only to heavy nuclei consisting of at least 144 nucleons (^{144}Nd being the lightest α-emitting nuclide), involves the spontaneous emission of a helium nucleus.

1.2. Timeline of Discoveries in Radioactivity

1896: Discovery of Radioactivity. H. BECQUEREL discovered radioactivity one year after the discovery of X rays. Using photographic plates to study the fluorescence of uranium compounds, BECQUEREL found that uranium blackened plates even after a long period in the dark. Fluorescence, therefore, cannot be the cause of the blackening; the radiation comes spontaneously from the uranium, without prior excitation.

Radio-Thorium. MARIE CURIE made a systematic study of all elements to investigate which ones emit the newly discovered radiation. She discovered that thorium, two atomic numbers away from uranium, is also "radioactive."

First Activity Unit, the Curie (Ci). In quantitative investigations with a galvanometer, M. CURIE found that the intensity of the radiation from uranium (the radioactivity) is proportional to the quantity of uranium.

1898: Polonium. From the pitchblende ore of a uranium deposit, M. CURIE isolated an element with atomic mass 210. She proposed the name "polonium" after her homeland.

1898: Radium. PIERRE and MARIE CURIE found element 88 in pitchblende from Joachimsthal and named it "radium."

1899: Birth of Nuclear Physics. ELSTER and GEITEL stated the hypothesis that radioactivity originates from the transmutation of elements. This hypothesis has ultimately become the basis for nuclear physics.

1899: Alpha and Beta Radiation. By measuring ranges, RUTHERFORD determines that radioactive radiation consists of at least two distinct components, which are designated α and β.

1903: Nobel Prize for Research on Radioactivity. H. BECQUEREL, M. SKLODOWSKA-CURIE, and P. CURIE shared the Nobel Prize for Physics.

1903: Gamma Radiation. The third component of radioactivity, called γ radiation, was first reported in 1903 and later described in 1906 by P. VILLARD.

1905: Gamma Quanta. EINSTEIN postulated the existence of corpuscular light quanta (photons) and explained the photoelectric effect. Einstein's postulate also explains the character of γ quanta.

First Decade of the 20th Century: Identification of Alpha and Beta Particles. The α *particle* was identified in 1903 by DES COUDRES, who determined the ratio of mass to charge of this particle, and in 1908 E. REGENER measured the charge. The α particle is found to have a mass identical to that of the helium nucleus. This result supplements the observation made by RAMSEY and SODDY (1903) that the noble gas helium is generated from "radium emanation"; this observation in turn supports ELSTER and GEITEL's 1899 thesis of transmutation. In 1909, RUTHERFORD also showed by spectroscopic measurements that α particles are helium nuclei.

For β *particles*, the deflection in electric and magnetic fields gives the ratio of charge to mass (KAUFMANN) and showed that the specific charge is the same as that of cathode rays, i.e., electrons.

1906: Equivalence of Mass and Energy. EINSTEIN stated that mass and energy are equivalent and can be transformed into each other. This equivalence is fundamental to all phenomena of radioactivity and nuclear physics.

1907: Three New Radioactive Elements. OTTO HAHN discovered the radioactive elements radiothorium, radioactinium, and "mesothorium" I and II.

1909: Isotopy. F. SODDY reported that lead atoms have different masses. The phenomenon of isotopy had not yet been explained until the discovery of the neutron in 1932.

1911: Second Nobel Prize for M. SKLODOWSKA-CURIE. M. SKLODOWSKA-CURIE received the Nobel Prize in Chemistry for the discovery of radium and polonium.

1911: The Atomic Nucleus. Rutherford's model of the atom has a small, heavy nucleus surrounded by a shell of electrons.

1913: Charge and Size of the Atomic Nucleus. H. GEIGER and E. MARSDEN obtained experimental confirmation of the Rutherford atomic model by measuring the charge and size of nuclei.

1913: The Hydrogen Atom; Planck's Constant. NIELS BOHR described a planetary model of the atom. The spectral lines of the hydrogen atom were calculated with the help of Planck's constant.

1918: Protactinium. O. HAHN and L. MEITNER discovered the radioactive element protactinium.

1919: Discovery of the First Artificial Nuclear Reaction. While studying the range of α particles in nitrogen, RUTHERFORD observed components having a longer range but a lower signal intensity on the scintillation screen. RUTHERFORD interpreted these unexpected components as protons:

$$^{14}N + \alpha \rightarrow {}^{17}O + p$$

1923: Compton Effect. A. H. COMPTON showed that X-ray quanta collide with electrons like discrete particles (1927 Nobel Prize).

1925: Pauli Exclusion Principle. The exclusion principle of W. PAULI, which holds for particles with spin 1/2 and thus in particular for electrons in the atomic shell, explains the periodic system of the elements. It has also accounted for the nuclear systematics in the framework of the shell model. Protons and neutrons are also spin –1/2 particles.

1926: Fermi Statistics. Fermi statistics apply to all spin –1/2 particles ("fermions"), thus in particular to electrons and nucleons.

1927: Chemical Bonding. W. HEITLER and F. LONDON accounted for chemical bonding on a quantum-physical basis.

1927: Betatron Principle. R. WIDERÖE and M. STEENBECK described the betatron principle for accelerating electrons.

1928: Spin and Magnetic Moment of the Electron. P. A. M. DIRAC explained the angular momentum and magnetic moment of the electron by introducing the laws of relativity theory to extend quantum wave mechanics.

1928: Heisenberg's Uncertainty Principle. HEISENBERG stated that in the world of elementary particles, position and momentum cannot both be determined simultaneously.

1930: Prediction of the Positron. In the framework of his extended quantum theory, DIRAC predicted the existence of the positron, which was discovered experimentally in 1932.

1930: Cyclotron. E. LAWRENCE constructed the cyclotron.

1932: Discovery of the Positron. C. D. ANDERSON discovered the positron.

1932: Discovery of the Neutron. JAMES CHADWICK discovered the neutron. Nuclei consist of neutrons and protons.

1932: First Artificially Induced Nuclear Reaction. The first nuclear reaction initiated by artificially accelerated particles (150 keV protons) was accomplished by J. D. COCKROFT and E. T. S. WALTON. They received the Nobel Prize in 1951.

1932: Deuterium. H. C. UREY discovered deuterium (Nobel Prize 1934).

1932: Pair Production. I. CURIE and F. JOLIOT discovered pair production with the postulate that a γ quantum transforms into an electron – positron pair; mass and energy are equivalent (EINSTEIN, 1906).

1934: Annihilation Radiation. THIBAUD and JOLIOT discovered the emission of annihilation radiation when a positron and an electron interact and subsequently produce two γ quanta. As in the case of pair production, mass and energy are equivalent (EINSTEIN, 1906).

1934: Artificial Radioactivity. I. CURIE and F. JOLIOT discovered artificial radioactivity.

1934: Neutron-Induced Nuclear Reactions. E. FERMI initiated nuclear reactions with neutrons.

1934: Neutrino Theory. E. FERMI postulated the existence of the neutrino in order to account for β decay.

1935: Betatron. M. STEENBECK constructed the betatron, an electron accelerator.

1937: The μ Meson. The μ meson, predicted in 1935 by YUKAWA, was discovered in cosmic rays.

1937: Technetium. G. PERRIER and E. SEGRÉ discovered element 43, technetium, as a nuclear reaction product. Because its half-life of 4.2×10^6 a (^{98}Tc) or 210 000 a (^{99}Tc) is short on a cosmological scale, the element has never been found in nature. It is, however, certainly present in nature as a product of the very rare spontaneous fission of uranium.

1938: Neutron-Induced Nuclear Fission. HAHN and STRASSMANN discovered that the uranium nucleus can be induced to fission by neutron bombardment.

1942: Chain Reaction. E. FERMI achieved the first nuclear chain fission reaction.

1945: Synchrotron. E. M. MCMILLAN and V. VEKSLER developed the synchrotron principle.

1950: High-Energy Accelerators. In 1950 there were 10 accelerators producing particles with energies of 100 MeV or higher.

1954: 5 GeV Protons. The "bevatron" in Berkeley accelerated protons to an energy of 5 GeV. The age of elementary-particle physics has now begun.

As this timeline shows, research on the atomic nucleus and radioactivity was far advanced by the 1950s. To list the advances of succeeding decades would go beyond the bounds of an introductory survey. The public has shown more interest in the engineering use of nuclear fission and the so-far unsuccessful attempt to obtain energy from nuclear fusion than in the results of basic research (⟶ Nuclear Technology). The theoretical understanding of nuclear physics has since been extended and is now being supplemented by knowledge of elementary-particle physics (high-energy physics), for example, what has been learned about the strong force that governs the interaction between nucleons, or about the weak force that governs β decay.

1.3. Nuclides and Radionuclides

A nuclide is a structured, dynamic assemblage of nucleons forming a stable combination. A radionuclide is, similarly, a structured, dynamic assemblage of nucleons but has the capability of spontaneous change.

Production of Radioactive Nuclides. The majority of radioactive nuclides on Earth — now more than 1500 — have been produced *artificially* (see Chap. Production of Artificial Radionuclides). The number per se is not really important, since it is subject to almost arbitrary increase. There is a large stock of ions that can be accelerated to high energies and aimed at one of a long list of targets, initiating a nuclear reaction. Many of the nuclei produced in such reactions have extremely short lifetimes, typically, the shorter the lifetime the more the reaction product differs from a stable nucleus.

Without human intervention, radionuclides are produced on Earth by *cosmic-ray species*: high-energy protons, mesons (entities consisting of two quarks), γ quanta, or electrons. In particular, activation of stable elements in the atmosphere and crust yields the radioactive nuclides ^{14}C, ^{3}H, and ^{10}Be. These undergo β^- decay toyield the following products: ^{14}C to ^{14}N (half-life 5730 a), ^{3}H to ^{3}He (12.323 a), and ^{10}Be to ^{10}B (1.6×10^{6} a).

Primordial Radionuclides. These nuclei originated in the early life of our universe and today, under terrestrial conditions, provide evidence of the occurrence of nucleosynthesis.

There are many theories, some competing and some possible supplementing others, about what happens in nucleosynthesis. All these theories, however, agree on one key point: that the heavier nuclei arose from lighter nuclei while the very light nuclei were produced by the fusion of single nucleons. Whether the very light nuclei, such as ^{4}He, were born as a direct result of the “Big Bang” or created in stellar interiors (as is thought to have happened for most of the elements), these controversial questions go beyond the scope of this article.

Figure 1 gives a two-dimensional scheme of all the primordial nuclides still present in the Earth's crust, broken down into the number of protons and number of neutrons. It is striking that the heavier primordial nuclei have more neutrons that protons and that the margin steadily increases. This “neutron excess” will be discussed further on. ^{42}Ca is the heaviest primordial nucleus with no neutron excess.

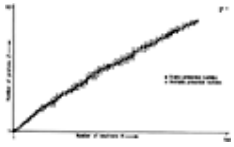

Figure 1. All the primordial nuclides still present in the Earth's crust

Bars and Isobars. The two-dimensional representation of the nuclide chart exaggerates the difference between neutrons and protons. Accordingly, nuclear systematics will first be discussed from the viewpoint of the nucleus, that is, in a one-dimensional treatment based on a parameter called nucleon number (*M*).

In chemistry the term “element 10,” is used with unambiguous reference to the element having atomic number 10. The atomic number is equal to the number of protons. For the neutral element, the atomic number is also equal to the number of electrons in the atomic shell. This element 10, with the name neon, has an atomic mass of 20.179. This value arises from the fact that there are three stable isotopes of element 10, ^{20}Ne (with isotopic abundance 90.51 %), ^{21}Ne (0.27 %), and ^{22}Ne (9.22 %).

By analogy with the identification of chemical element 10, a nucleus consisting of 10 nucleons is called bar 10 (from Greek τό βάρος, which means weight). All nuclei having 10 nucleons are isobaric with one another. The stable bar 10 consists of five neutrons and five protons. It is the nucleus of element 5, boron. There is no other stable isobar 10, although there is a stable boron isotope ^{11}B; fewer isobars than isotopes exist. A given bar, as a rule, has either a single stable form or, in some cases, two stable isobars. This law follows from the possibility of β decay, which enables the bar to change a nucleon of one species into the other species. A bar always has this capability when the transformation is energetically possible; naturally, the conservation of angular momentum also governs any transformation. Thus every bar undergoes transforof angular momentum also governs any transformation. Thus every bar undergoes transformation until it reaches the energetically most favorable situation.

Bars with an Odd Number of Nucleons. Examine bar 99 from this standpoint. The bar has just one stable nuclide, ^{99}Ru, an isotope of element 44. Element 44 has no fewer than seven stable isotopes. Figure 2. lists the radioactive isobaric nuclei of bar 99. From ruthenium to cadmium (atomic number 44 – 48) the proton number increases by 1 and the neutron number decreases by 1 at each step. The ordinate is the mass difference between the respective isobar and the stable isobar (Ru). Since the sum of the masses of the nucleons changes little through an isobaric series, the equivalence of mass and energy means that larger mass corresponds with lower bindingenergy. The mass defect of the lightest nucleus in the series corresponds to a particularly high binding energy.

Figure 2. Mass or energy parabola of the isobars of bar 99

Among the isobaric nuclei of bar 99, why do elements 48 – 45 and 37 – 43 differ in binding energy from element 44? For a given bar with an odd number of nucleons, 99 in the present case, the positive binding energy is governed by the *strong interaction* between each of the 99 nucleons and (especially) its nearest neighbors, since the strong interaction acts over a short range.

The *electrostatic interaction* between protons is a repulsive force that lowers the binding energy. If only the electrostatic interaction is considered, the binding energy increases when a β^+ transformation changes one of its protons into a neutron. This happens to elements 48 to 45 of bar 99 (gain in binding energy). For quantum-mechanical reasons, however, an increase in the number of neutrons also carries an energetic disadvantage, which is described by the "nuclear shell model". In the nuclear field set up by all the nucleons as a whole, both protons and neutrons are located in individual orbits arranged in shells, similar to the configuration of electron shells in the central field set up by the nucleus. Because nucleons, like electrons, have an intrinsic angular momentum, spin 1/2, the Pauli exclusion principle dictates that each nucleon is located in a site that can be singularly occupied: No neutron can match another neutron in all the quantum numbers, and similarly no proton can match another proton.

In this way, every β^+ transformation causes a higher-energy neutron site to be occupied while releasing a lower-energy (energetically more favorable) proton site. On the other hand, such a transformation frees the nucleus of one positive electrical charge. These two effects are balanced for the stable nucleus with proton number 44, the element ruthenium. Elements 37 to 43 of bar 99 release a higher-energy neutron site upon a β^- transformation, but the electrostatic repulsion increases due to the proton that is produced. The masses of the isobars of bar 99 fall approximately on a parabola (mass or energy parabola), as shown in Figure 3. The sole stable isobar is located near the vertex of the curve. Odd-numbered bars have no more than one stable nuclide. The stable bar 99 has 44 protons (an even number) and 55 neutrons (an odd number), and so ^{99}Ru is said to be an even – odd nucleus. All the nuclei of the isobaric series 99 alternate between even – odd and odd – even. This law naturally holds for all odd bars.

Early in the study of stable-nucleus systematics, it was believed that five exceptions had been found to the rule that any odd bar has only one stable nuclide. The existence of these exceptions was later disproved: ^{87}Rb (half-life 4.8×10^{10} a; see Fig. 3), ^{115}In (4.0×10^{14} a), ^{187}Re (5.0×10^{10} a), and ^{113}Cd (9.0×10^{15} a) undergo β^- decay to their stable neighbors ^{87}Sr, ^{115}Sn, ^{187}Os, and ^{113}In. The fifth unconfirmed excep-tion, ^{123}Te, is transformed to stable ^{123}Sb by a β^+ decay with a half-life of 1.24×10^{13} a (Fig. 4). The β instability of these five nuclei was not detectable until the availability of advanced measurement techniques, because a long half-life is equivalent to a low transition probability and frequency. These nuclides are radioactive primordial nuclides, their production by nucleosynthesis, however, has left them in an energetic niche from which they escape only on an almost cosmological time scale. The continuing nucleosynthesis of the stable daughter nuclides ^{87}Sr, ^{115}Sn, ^{187}Os, ^{113}In, and ^{123}Sb can thus be witnessed.

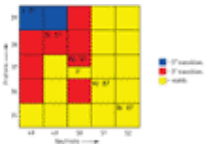

Figure 3. The transformation of ^{87}Rb to ^{87}Sr through β^- decay

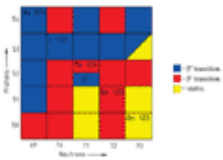

Figure 4. The transformation of ^{123}Te to ^{123}Sb through β^+ decay

Bars with Even Nucleon Numbers. Even bars from M = 36 up have two stable nuclides. This is a consequency of the pairing energy: The formation of very tightly bound neutron – neutron and proton – proton pairs results in a more favorable net nuclear binding energy. It can easily be shown that the isobars of an even-numbered bar alternate between even – even and odd – odd nuclei; the energies fall on two parabolas offset by roughly twice the pairing energy. Figure 5 shows the mass parabola for bar 138. ^{138}Ba and ^{138}Ce are stable, while ^{138}La is a primordial, naturally radioactive, odd – odd nuclide. It undergoes either a β^+ or a β^- transformation into one of its two stable neighbors; the net half-life is 1.35×10^{11} a. Figure 6 is a diagram of the corresponding situation for the primordial nuclide ^{40}K. Most nuclides lying on the mass parabolas in the figures — except the primordial nuclides — have been produced artificially.

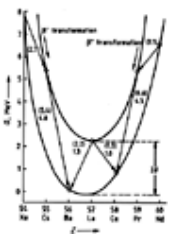

Figure 5. Mass or energy parabola of the isobars of bar 138

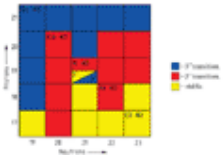

Figure 6. The transformation of the primordial odd – odd nuclide ^{40}K through either β^- decay to the even – even isobar ^{40}Ca or through β^+ decay to the even – even isobar ^{40}Ar

So far, there are two exceptions to the rule that even-numbered bars from *M* = 136 upwards have two stable nuclides. The first is the stable odd – odd nucleus ^{50}V, the third stable species of its bar. It lies between the stable even –even isobars ^{50}Cr and ^{50}Ti. The second exception is bar 82 (Fig. 7), which has a single stable species, the even – even nucleus ^{82}Kr. Up until the 1970s, the primordial even – even neighbor ^{82}Se was regarded as stable for an obvious reason: It is now known that ^{82}Se decays by a double β^- process (simultaneous emission of two β^- particles) with a correspondingly long half-life of 1.0×10^{19} a, yielding the stable nuclide ^{82}Kr, which is located two places away. In other words, ^{82}Kr is another species whose nucleosynthesis is far from finished. The ^{82}Se (2 β^-) ^{82}Kr process has a very low probability of occurrence and is perceived as "against the rules," but its example confirms the trivial statement that the knowledge of nuclear physics and the laws that are regarded as governing it are correct only in the context of present measuring techniques and within a time "window" that is tiny when compared with cosmological spans of time.

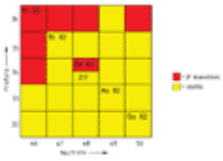

Figure 7. The double β^- transformation of ^{82}Se to ^{82}Kr

Light bars (up to bar 35) have exactly one stable isobar, with two exceptions that have not yet been well explained by nuclear physics: There is no bar 5 at all, and bar 8 exists only in the form of artificially created unstable radionuclides.

α Decay. Within a given bar, there is a tendency to acquire states with the lowest possible energies and the strongest possible binding via the β process. While β decay is a local event in the context of nuclear systematics, governed mainly by the properties of the weak interaction and operating within a single bar, α decay involves the nuclear systematics of all bars and the mechanism of the strong nuclear force. An α decay takes place when a nucleus can become more stable by giving up some nucleons. The preconditions for this event should be understood in the framework of a collective model of all bars: the droplet model (Fig. 8). This model does not, however, include pairing energy; important for understanding the systematics of even bars.

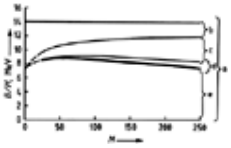

Figure 8. Mean binding energy per nucleon *B/M* according to the droplet model

The droplet model does not hold for *M* < 40The binding energy is the result of the volume energy a, minus b, c, and da) Volume energy; b) Surface energy; c) Coulomb energy; d) Asymmetric energy; e) Binding energy

In the droplet model, the nucleons in the nucleus are treated as analogous to the molecules in a droplet of water. Every nucleon, whether proton or neutron, has the ability to take part in the strong interaction. Since the nuclear force essentially acts between quarks, the only bonds between nucleons are between nearest neighbors. Thus the binding energy of a nucleus is initially proportional to the number *M* of nucleons. In-side the nucleus, every nucleon takes up space, so that it is not incorrect to say that the volume of a nuclide is approximately proportional to the nucleon number (volume energy; see Fig. 8). The nucleons at the surface of a nucleus have, on average, fewer neighbors than those in the interior. Hence the surface (exterior) nucleons contribute less to the nuclear binding energy. Because the area : volume ratio becomes smaller with increasing volume (increasing nucleon number), nuclei with more nucleons have a more favorable area : volume ratio. Thus, in this respect, heavier nuclei are more stable than lighter ones, because the (negative) *surface energy* decreases.

Next, the contribution to binding energy due to the interactions already discussed is evaluated. *Electrostatic repulsion* between protons has a destabilizing effect and, accordingly, lowers the binding energy. Unlike the nuclear force, electrostatic repulsion acts over an unlimited range, so that every proton repels every other proton and is itself repelled. The destabilizing force of electrical repulsion is therefore proportional to $Z(Z-1)$, where Z is the number of protons. For Z "not too small," this is approximately a quadratic dependence, so that stable nuclei (as well as the primordial radiactive nuclei) have an increasing surplus of electrically neutral neutrons ("neutron excess"), which contribute to the stability of the nucleus with their nuclear force.

For the same quantum-mechanical reasons stated in accordance with the nuclear shell model, however, energetic disadvantages relate to the increasing number of neutrons. Additional neutrons occupy higher and higher energy levels, so that it is not possible to achieve stability through an "excessive" neutron excess (asymmetry energy). The contribution to nuclear stability due to an additional neutron grows smaller and smaller until it finally vanishes. The heaviest stable nucleus is ^{209}Bi.

Now if the net binding energy:

Binding energy = volume energy (nuclear force) + surface energy (negative) (nuclear force) + repulsion between protons (negative) (electrostatic) + higher energy of levels for excess neutrons (asymmetry energy) (quantum-mechanical)

is plotted versus the mean binding energy per nucleon bars starting around M = 70 – 80 appear to be less and less tightly bound. If the nucleus, however, could separate itself from a few nucleons, it would be more tightly bound. Some nuclei do take advantage of this possibility and expel a helium nucleus (α particle), which is very tightly bound. The lightest primordial α emitter is ^{144}Nd, which decays to ^{140}Ce and ^{4}He with a half-life of 2.1×10^{15} a. The heaviest primordial α emitter in the neighborhood of stable nuclei is ^{204}Pb which decays to stable ^{200}Hg with a half-life > 1.4×10^{17} a (Fig. 9).

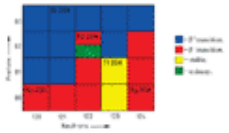

Figure 9. The transformation of ^{204}Pb (82 p, 122 n) to ^{200}Hg (80 p, 120 n) through α decay

The heaviest primordial nuclides are the three isotopes of uranium (^{234}U, ^{235}U and ^{238}U) — whose radiation led to the discovery of radioactivity in 1896 — and ^{232}Th. The decay chains discussed in Chap. Radioactive Elements, Figure 18 begin from these four nuclides; each chain is a branching sequence of α and β^- decays, ending with isotopes of lead. There is also a low probability of the four primordial nuclides decaying by spontaneous *fission*; the products include, in particular, technetium. This element has formerly thought to exist on Earth only as an artificial product, but has reportedly been found in uranium ores.

1.4. The Law of Radioactive Decay

In a tennis game, the ball occasionally lands among the spectators. This is neither the point of the game nor the intent of the players; it is a random event in the dynamics of play. When it will occur cannot be predicted; there is merely a certain probability that it will happen.

Comparable with the dynamics of a tennis game are the internal dynamics of an individual radioactive nuclide, which lives with the permanent risk of losing, e.g., four nucleons (an α particle in this situation) all at once. The nuclide suffers this loss at an unpredictable time and without external stimulation. Radioactive nuclides undergo such changes spontaneously, and radioactive decay must be regarded as a stochastic process.

The behavior of the stochastic occurrence in time is characterized by a decay constant λ. This constant is defined as a probability per unit time, specifically the (infinitesimal) probability dw that a certain individual nucleus will decay in the infinitesimal time dt:

$$\lambda = \mathrm{d}w/\mathrm{d}t \tag{1.6}$$

or

$$\mathrm{d}w = \lambda\,\mathrm{d}t \tag{1.7}$$

The decay constant λ has the following properties:

1. It is equal for all nuclides of a given species
2. It is constant over time; Nuclides do not “age” and do not have any history

If a certain time interval with duration 1T is considered, the probability of observing the decay of a certain nucleus in T can be defined in the following way by analogy with Equation (1.7):

$$w\,(1\,T) = \lambda_{\mathrm{T}} T \tag{1.8}$$

If $\lambda_T T$ is the probability that the nucleus decays in the time interval T under consideration, then

$$w^{\circ}\,(1\,T) = 1 - w\,(1\,T) = 1 - \lambda_{\mathrm{T}} T \tag{1.9}$$

is equal to the probability that the nucleus will not yet have decayed after time 1T (that it will have “survived”). After two intervals of length T have passed,

$$\begin{aligned} w^{\circ}(2\,T) &= w^{\circ}(1\,T)(1-\lambda_{\mathrm{T}}) \\ &= (1-\lambda_{\mathrm{T}} T)(1-\lambda_{\mathrm{T}} T) \\ &= (1-\lambda_{\mathrm{T}} T)^2 \end{aligned} \tag{1.10}$$

after three intervals of length T,

$$\begin{aligned} w^{\circ}(3\,T) &= w^{\circ}(2\,T)(1-\lambda_{\mathrm{T}} T) \\ &= (1-\lambda_{\mathrm{T}} T)^3 \end{aligned} \tag{1.11}$$

and so forth. For any number n of intervals of length T, then:

$$w^{\circ}\,(n\,T) = (1-\lambda_{\mathrm{T}} T)^n \tag{1.12}$$

Equation (1.12) states that the probability w^o that the nucleus has not yet decayed is reduced by the same fraction λ_T in every equal time interval T. This kind of time dependence is described by an exponential function. The amplitudes of the time-dependent probabilities w^o (mT) at discrete times $t_m = mT$ lie exactly on the curve of an exponential:

$$w^o(mT) = e^{-\lambda mT}, m = 0, 1, 2, 3, \cdots, n, \cdots \tag{1.13}$$

If Equations (1.12) and (1.13) are combined, the result is:

$$w^o(mT) = (1 - \lambda_T T)^m = e^{-\lambda mT} \quad m = 0, 1, 2, 3, \cdots, n, \cdots \tag{1.14}$$

What does λ signify in the exponential function of Equation (1.4)? First, change from discrete time to continuous time t and write:

$$t = mT \tag{1.15}$$

It then follows that

$$m = t/T \tag{1.16}$$

so that Equation (1.14) becomes:

$$w^o(mT) = (1 - \lambda_T T)^{t/T} = e^{-\lambda t} \tag{1.17}$$

$$(1 - \lambda_T T)^{t/T} = e^{-\lambda t} \tag{1.18}$$

The next steps are simple mathematical transformations of Equation (1.18), in which the expression are taken to the limit of infinitesimal intervals T ($T \longrightarrow 0$).

Taking logarithms:

$$(t/T) \ln (1 - \lambda_T T) = -\lambda t$$
$$-\ln (1 - \lambda_T T) = -\lambda T$$
$$\ln (1 - \lambda_T T) = \lambda T$$

Inverting the logarithms:

$$(1 - \lambda_T T) = e^{-\lambda T}$$
$$\lambda_T T = 1 - e^{-\lambda T}$$

Series expansion of the exponential function:

$$\lambda_T T = 1 - \left[1 - \lambda T + (1/2)(\lambda T)^2 - \cdots\right]$$

or

$$\lambda_T T = \lambda T - (1/2)(\lambda T)^2 + \cdots$$

Division by T:

$$\lambda_T T = \lambda - (1/2)\lambda^2 T + \cdots \tag{1.19}$$

As the interval T becomes shorter and shorter, Equation (1.19) takes the following form in the limit $T \longrightarrow 0$:

$$\lim_{T \to 0} \lambda_T = \lambda \tag{1.20}$$

This shows that the decay constant λ defined in Equation (1.6) is immediately equal to the coefficient λ in the exponential function of Equations (1.17) and (1.18). Thus, for any time t,

$$\omega^o(t) = e^{-\lambda t} \tag{1.21}$$

Equation (1.21) states that the probability w^o (t) that a particular nucleus has not yet decayed after an “observation time” t will decrease exponentially with the decay constant λ as coefficient.

If now a group of identical radioactive nuclei are observed starting at a time $t = 0$ when the group consists of N individual nuclei, the probability is N (t)/N ($t = 0$) that there are N (t) undecayed nuclei after a time t:

$$N(t)/N(t = 0) = e^{-\lambda t} \tag{1.22}$$

This can be rewritten as:

$$N(t) = N(t = 0) \cdot e^{-\lambda t} \tag{1.23}$$

This exponential function describes the decrease in the number of radioactive nuclei that still remain to decay.

Consider now the time $t = \tau$ at which exactly half of all the nuclei have decayed (half-life):

$$N(\tau) = (1/2)N(t=0) \tag{1.24}$$

By Equation (1.23)

$$1/2 = e^{-\lambda t} \tag{1.25}$$

Manipulation of Equation (1.25) yields the following relations between the half-life τ and the decay constant λ:

$$\lambda - (\ln 2)/\tau \text{ or } \tau = (\ln 2)/\lambda \tag{1.26}$$

The number of radioactive nuclei is not accessible to measurement, but the rate of decaying nuclei, called the activity *A*, is.

Since λ is the decay probability of an individual nucleus in an (infinitesimal) time, if there are *N* radioactive nuclei then the expected number of decay events per unit time is:

$$A(t) = \lambda N(t) \tag{1.27}$$

If λ is expressed in reciprocal seconds (s^{-1}), Equation (1.27) gives the frequency in decay events per second, i.e., the activity *A* of a collection of radioactive nuclei (a "source").

In time-dependent notation, Equation (1.23) gives:

$$A(t) = \lambda N(t) = \lambda N(t=0)e^{-\lambda t} \tag{1.28}$$

$$A(t) = A(t=0)e^{\lambda t} \tag{1.29}$$

Equation (1.29) is the law of radioactive decay in its most frequently used form.

A comparison of Equations (1.21), (1.23), and (1.27):

$$w^o(t) = e^{-\lambda t}$$
$$N(t) = N(t=0)e^{-\lambda t}$$
$$A(t) = A(t=0)e^{-\lambda t}$$

shows that the probability $w^o(t)$ of a specific individual nucleus not yet having decayed (Eq. 1.21), the number $N(t)$ of decayable nuclei (Eq. 1.23), and the activity $A(t)$ (Eq. 1.27) display the same time dependence.

Definition of the Activity A. The physical quantity "activity" is defined as the number of decays per unit time. Since 1985, the unit of activity has been the number of decays per second:

$$A = \text{decays} \cdot s^{-1} \tag{1.30}$$

Its special name since 1985 has been the becquerel (Bq):

$$1\,\text{Bq} = 1\,\text{decay} \cdot s^{-1} \tag{1.31}$$

Up to 1950, the unit of activity was defined in terms of the activity of 1 g of uranium (special name: curie (Ci) after M. CURIE, who discovered the proportionality between the mass of radioactive uranium and the "intensity" of radiation). This definition was made numerically exact in 1950:

$$1\,\text{Ci} = 3.7 \times 10^{10}\,\text{decays} \cdot s^{-1} \tag{1.32}$$

By international convention, the unit curie may no longer be employed, but is, however, still frequently encountered. However, since radioactive decay is a stochastic process, sequential measurements of a radioactive source over time will not yield count rates $I(t)$ that correspond exactly to the exponential law $I(t) = I(0)\,e^{-\lambda t}$. The exponential law describes only the time dependence of the "expected" value $I(t)$. The individual measurements $\hat{I}(t_i)$ are Poisson distributed about the expected value $I(t_i)$. For "not too low" count rates $\hat{I}$ ($\hat{I} > 20$), typical of all technical uses of radioactive sources, the skew-symmetric Poisson distribution can be replaced, to a good approximation, by the symmetric Gaussian normal distribution.

1.5. Interaction between Radiation and Matter

1.5.1. *Interaction between Gamma Quanta and Matter*

γ Quanta (or, in general, photons), the quanta of the electromagnetic interaction, bear no charge and are thus not sensitive to the longrange electrostatic force of the nuclear charge or electrons.

γ Quanta in the energy range of nuclear transitions or the radiation following such transitions (roughly 0.01 – 10 MeV) are detected almost exclusively through three processes:

1. Compton effect on an electron
2. Photoeffect on a bound electron

3. Pair production in the field of a nucleus

Compton Effect. The Compton effect can be regarded as the inelastic scattering of a photon on an electron. In principle, Compton scattering can occur on either a free electron or a bound orbital electron. The photon undergoes a change in direction by an angle θ called the scattering angle and a loss of energy that depends on θ. If the initial photon energy is $h\nu$, the energy $h\nu'$ of the scattered photon is:

$$h\nu' = h\nu/[1 + \alpha(1 - \cos\theta)] \tag{1.33}$$

In Equation (1.33), α is the ratio of the photon energy $h\nu$ (in keV) to the rest-mass energy m_0 of the electron:

$$\alpha = h\nu/m_0c^2 = h\nu/511\ \mathrm{keV} \tag{1.34}$$

In the limit of very small angles ($\theta \longrightarrow 0$), the energy loss vanishes: $h\nu' \longrightarrow h\nu$. The largest energy loss occurs in "backscattering" (θ = 180°):

$$h\nu' = h\nu/(+2\alpha) \tag{1.35}$$

Thus a photon can never lose all of its energy in a Compton event and so can never disappear entirely.

By the law of conservation of energy, the energy lost by the photon is transferred to the electron in the collision:

$$E_{e1} = h\nu - h\nu' \tag{1.36}$$

By the conservation of momentum, the angle ϕ at which the electron is scattered is given by:

$$\cot\phi = (1 + \alpha)\tan(\theta/2) \tag{1.37}$$

The probability of an interaction between a photon and an electron is measured by the cross section of the electron. The cross section turns the interaction probability for some process into a geometrical relationship. The projectile, in this case the photon, is idealized as a body of infinitesimally small extent, i.e., a mathematical point. The other body in the interaction, in this case the electron, is modeled as a sphere. The radius of the sphere is selected so that, for given relative direction and position between the bodies, the process occurs with the same probability as a geometrical mechanical collision. The area of the plane projection through the center of the sphere defines the cross section.

The Compton effect cross section ${}_e\sigma$ of an electron is largest for low photon energies, decreasing very rapidly with increasing photon energy (Fig. 10). The direction distribution of the scattered photons is plotted versus primary photon energy in Figure 11.

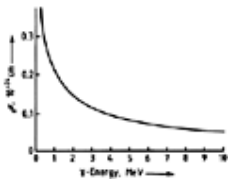

Figure 10. The Compton effect

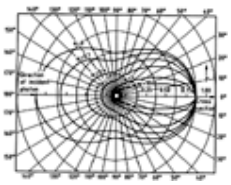

Figure 11. The angular distribution of Compton – scattered photons as a function of primary photon energy where $\alpha = E\ {}^{\sigma}/M_0C^2 = E\ {}^{\sigma}/511$ keV

Photoelectric Effect (Photoeffect). At energies below 0.1 MeV, the dominant interaction between γ quanta and matter having moderate to high atomic numbers is the photoelectric effect or photoeffect. This effect is the interaction with a bound orbital electron, in which the γ quantum gives up all of its energy to the electron in such a way that the photon disappears and the electron escapes from the atomic shell with a kinetic energy T described by the equation:

$$T = h\nu - B_e \tag{1.38}$$

where T is the kinetic energy of the electron and B_e is the electron binding energy.

By the conservation of momentum, a γ quantum cannot be "absorbed" by an unbound electron. Even in the atomic electron shell, the photoeffect in roughly 4/5 of all cases takes place on the most strongly bound electrons, those in the K shell. Of course, the energy of the γ quantum must be greater than the binding energy of the orbital electron for this to happen. This condition leads to a "K edge" in the curves representing the photoeffect cross section versus photon or γ energy for a given material.

If the γ energy drops below the K-shell binding energy of the absorbing material, the cross section decreases abruptly as K-shell electrons cease to be available as partners for the photoeffect (Fig. 12).

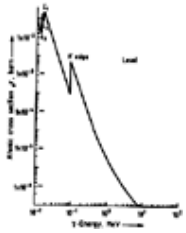

Figure 12. Energy dependence of the photoeffect

The directional distribution of orbital electrons after the absorption of a γ quantum (photoelectrons) depends on the photon energy. At low photon energies (ca. 0.01 MeV), the distribution has a peak nearly perpendicular to the initial photon direction; as the photon energy increases, the forward direction becomes more and more preferred. Figure 13 is a plot of the bipartition angle versus photon energy; this is the angle that exactly divides the frequency distribution of electron scattering angles in half. For example, the bipartition angle is just 20° at a photon energy of 1 MeV; this means that half of all photoelectrons have a direction less than 20° away from the photon or γ incidence direction.

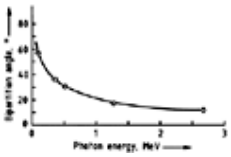

Figure 13. Bipartition angle for photoelectrons as a function of primary photon energy

The conservation of momentum gives rise to an interesting observation. Because the photoelectron, having a rest mass, may absorb a large fraction of the energy of the zero-rest-mass photon in the form of kinetic energy, there is a large mechanical momentum in the direction of the photoelectron, and the nucleus must compensate for this by absorbing a recoil momentum.

The probability that a photon will experience a photoeffect on an orbital electron while pass-ing through a material is a function of the atomic number *Z* of the material and the energy $h\nu$ of the photon. The dependence is not a simple one.

Again, a cross section ${}_a\tau$ serves as a measure of this probability (where *a* denotes atom):

$$ {}_a\tau = \mathrm{const} \cdot Z^n/(h\nu)^m \tag{1.39} $$

where *n* in turn depends on $h\nu$ and *m* depends on both $h\nu$ and *Z*. Thus *n* and *m* are not constants. The complicated nature of the relation suggests that it is not the mathematical reflection of a well-known physical theory but a heuristic numerical expression. The value of *n* does not vary rapidly; roughly speaking, *n* = 4.05 ($h\nu$ = 0.01 MeV) while *n* = 4.6 ($h\nu$ = 3 MeV). Further, *m* is about 3 for $h\nu$ = 100 keV, 2 for $h\nu$ = 400 keV, and approaches unity at higher photon energies. For a constant photon energy, *m* increases slowly as the atomic number *Z* decreases.

Pair Production. At γ energies above 1.02 MeV, a third type of interaction called pair production becomes significant and ultimately dominant.

In pair production, the photon or γ quantum disappears altogether, and an electron – positron pair appears instead. The total energy of the elec-tron and positron is exactly equal to the initial photon energy:

$$ h\nu = (T_- + m_0c^2) + (T_+ + m_0c^2) \tag{1.40} $$

where *T* is the kinetic energy and m_0c^2 is the energy equivalent to the electron or positron rest mass. Both particles have exactly the same rest mass.

Pair production can occur only in the field of a charged particle, preferably of an atomic nucleus but potentially in that of an electron, since either nucleus or electron can absorb momentum and thus ensure is convervation.

The directions of the electron and positron are forward with regard to the photon direction, diverging symmetrically from the photon trajectory. As the photon energy increases, the divergence between the electron and positron paths decreases and the forward tendency becomes more marked.

While the cross sections for the Compton effect ${}_e\sigma$ and photoeffect ${}_a\tau$ decrease continuously with increasing photon energy, the cross section ${}_\alpha\kappa$ for pair production increases rapidly. At photon energies of a few MeV, pair production is the dominant process.

Linear Attenuation Coefficient and Exponential Attenuation Law for Photons. The linear attenuation coefficient is useful to describe the fate in macroscopic terms of photons as they pass through matter. This coefficient is nothing more than the three-dimensional cross-section density in the material.

The interaction partner of the photon in the Compton effect is a single electron. Consider the material having an atom density of *N* atoms per cubic centimeter and the atom having *Z* electrons. The cross-section density is then *NZ* and the

linear attenuation coefficient for the Compton effect is:

$$\sigma = {}_e\sigma\, NZ \tag{1.41}$$

For the photoeffect, the cross section is stated for a single atom. Accordingly,

$$\tau = {}_a\tau\, N \tag{1.42}$$

and likewise for pair production

$$\chi = {}_\alpha\kappa N \tag{1.43}$$

In the cross-section model, a three-dimensional element of the absorber having a volume of 1 cm^3 is thus filled by

$$NZ + N + N \tag{1.44}$$

spheres having diameters of ${}_a\tau$, ${}_e\sigma$, and ${}_\alpha\kappa$, respectively.

As the photon travels along its straight-line path, it does not “see” spherical obstacles but disks blocking its way. These disks have a total area μ per unit volume given by:

$$\mu = \sigma + \tau + \chi \tag{1.45}$$

Now μ is the total linear attenuation coefficient.

The probability d*W* that a photon will interact (hit a disk) while traversing an infinitesimal path element d*x* is equal to μ d*x*— the area covered by cross sections (perpendicular to the photon direction) — per unit area (1 cm^2):

$$\mathrm{d}W = \mu\,\mathrm{d}x/1 = \mu\mathrm{d}x \tag{1.46}$$

The probability $\mathrm{d}W_0$ that the photon travels a distance d*x* without a collision is:

$$\mathrm{d}W_o = 1 - \mathrm{d}W = 1 - \mu\,\mathrm{d}x \tag{1.47}$$

By analogy with the derivation leading to the exponential law of radioactive decay (see Section The Law of Radioactive Decay), the attenuation of a beam of initially *I* γ quanta over a path of length *x* will also be described by an exponential function:

$$I(x) = I(x = 0)\exp(-\mu x) \tag{1.48}$$

The exponential law applies only to a collimated beam, i.e., a collective of γ quanta having the same initial direction. Equation (1.48), in particular, makes no distinction between quanta that are completely lost by photoeffect or pair production and quanta that are merely removed from the primary ray bundle by Compton scattering. In this way, all γ quanta still considered to belong to the collimated beam still have their initial energy, so that the attenuation coefficient remains constant. This is an important prerequisite for the derivation of the exponential law. Equation (1.48) thus may describe an observation associated with a certain experimental setup rather than an especially significant physical law.

As will be shown in the following sections, the exponential attenuation law does not apply to charged particles (α, β). If one argued that α and β particles, as charged particles, were subject to the “hindering” effect of the electrostatic interaction with nuclear and orbital electron charges, this explanation would wrongly mix concepts from classical physics and quantum mechanics, and what is more it would presumably miss the point: γ Quanta have a very much lower interaction probability than α or β particles. If the volume were packed full of cross sections for γ quanta, they would not travel far; the idealization of exponential attenuation would never have been achieved. Thus the practical consequences are that shielding for γ quanta can never be as complete as for α or β particles and that the law of exponential attenuation is in many cases nothing but a good approximation.

1.5.2. *Interaction between Electrons and Matter*

The observed interactions are essentially electromagnetic in nature. The three most frequent types of interaction are as follows:

1. An inelastic collision takes place between the electron and a bound orbital electron. This is the dominant process; it results in the ionization or excitation of the atom to which the orbital electron belongs.
2. The electron collides inelastically with a nucleus. A bremsstrahlung quantum is emitted in this process.
3. Elastic scattering occurs on a nucleus. The electron undergoes a change in direction, no bremsstrahlung is emitted, but the electron loses energy so as to conserve momentum.

In an absorbing material, a fast electron is slowed and eventually comes to rest. It is a matter of change which of the above three processes occurs and where.

Because of the stochastic character of the processes, the individual members of an electron beam, even a beam of monoenergetic electrons, have quite different fates. The depth of penetration into the absorber can be expressed only as a distribution about a mean depth, called the mean range.

A quantity to be distinguished from the range is the *path length S* traveled by an electron in an absorber before the electron comes to “thermal” rest. The mean path length is several times the mean range (Fig. 14). As it passes through the absorber, the electron generates other high-energy electrons (δ electrons) by ionization.

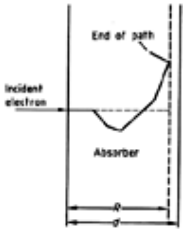

Figure 14. Total path length *S* and range *R* of an electron traversing an absorber of thickness *d*

1.5.3. *Interaction between Positrons and Matter*

Positrons or β^+ particles are rare in natural terrestrial radioactivity. A *high-energy* (“*fast*”) *positron* interacting with matter initially experiences the same fate as an electron. When, however, the positron has lost its kinetic energy and comes to take part only in the thermal motion of its surroundings, its special property of being the antiparticle of the electron comes into play.

At first, the positron seeks as a partner an electron, with which it forms *positronium* (a hydrogen-like system in which the proton is replaced by the positron) for a vanishingly short time. The next event is the mutual annihilation of the electron and positron; the rest mass of both is converted to photons of corresponding energy, which are called “annihilation radiation.”

In virtually every case, positronium has total angular momentum 0 (S state, “parapositronium”). Accordingly, the annihilation of positronium yields two quanta of equal energy (511 keV each), which are emitted in exactly opposite directions. The annihilation of the very rare *orthopositronium* gives rise to three annihilation quanta.

1.5.4. *Interaction between Alpha Particles and Matter*

When fast heavy particles (i.e., particles whose mass is large compared to that of the electron, as is true for α particles) interact with matter (atomic nuclei and their electron shells; also in chemical compounds), the principal result is the ionization or excitation of the atoms. Elastic scattering on the nucleus and bremsstrahlung are negligible.

Naturally, α particles are more and more slowed down as they travel through matter. While in the initial portion of their path they chiefly ionize or excite the atoms present, at lower velocities an exchange of electrons takes place. The doubly positively charged helium atom can capture either one or two electrons, lose these electrons again, and also excite or ionize its surroundings.

The paths of α particles through matter are approximately straight lines. As a result, the energy-dependent range of α particles is a better-defined quantity than the range of electrons (Fig. 15).

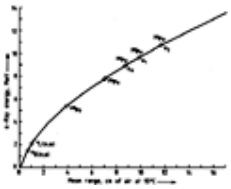

Figure 15. The energy-dependent range of α particles

1.5.5. *Interaction between Neutrons and Matter*

Neutrons “almost never” occur in natural radioactivity. 1 g radium emits 1.7×10^{10} α particles a second, while a spontaneous fission releasing a neutron occurs just once every 10 min. For the sake of completeness (“almost never” is not “never”), some features of the interaction between neutrons and matter should be briefly outlined.

Neutrons are sensitive to the strong interaction but not to electric charges. Accordingly, neutrons passing through matter become involved in interactions when they come so close to a nucleus that the nuclear force comes into play. The neutron attaches itself to the collection of nucleons, thus becoming a new member of the collection. From the viewpoint of the nucleus, the nucleus captures the neutron. The cross sections for this process are accordingly called neutron capture cross sections.

The capture cross section depends on the atomic mass of the capturing nucleus: A greater atomic mass implies a larger nuclear volume and a higher probability that the neutron will hit the nucleus (there is no potential threshold), and hence a larger capture cross section. Further, the cross section depends on the neutron energy: The higher the energy, the faster the neutron moves, hence the shorter the interaction time and the smaller the cross section.

In the range of low neutron energies, the interaction probability can be described well by the statement that the capture cross section σ_n is inversely proportional to the neutron velocity v:

$$\sigma_n \sim 1/v \tag{1.49}$$

as Figure 16 shows. Resonances are super posed on this behavior if the neutron excites the nucleus to preferred higher-energy states.

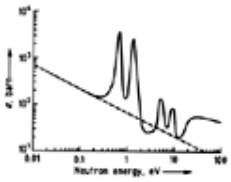

Figure 16. Neutron capture cross section σ as a function of the neutron energy E_n for a natural isotope mixture of iridium (37.3 % ^{191}Ir, 62.7 % ^{193}Ir)

The nucleus, now with one more neutron, can become stable after γ transitions to the ground state; the process is then identified as an (n, γ) process. It may alternatively act as a "compound" or "intermediate" nucleus, undergoing β^- decay (typical of light and moderately heavy nuclei), α decay, or fission. The fission process is described by a separate fission cross section (⟶ Nuclear Technology).

These (n, γ) processes are naturally probable in elements having many closely grouped stable isotopes. Thus cadmium is as important for shielding of neutrons as lead is for shielding of γ quanta.

Along with neutron capture, *neutron scattering* (moderation) also occurs on nuclei; this process, sometimes going by many pathways, slows the neutron until its motion is the thermal motion of its surroundings (⟶ Nuclear Technology – Moderation of Neutrons). The neutron is now said to be thermalized.

1.5.6. *Interaction between Neutrinos and Matter*

In contrast to neutrons, neutrinos are abundant in natural terrestrial radioactivity; more precisely, antineutrinos produced in β^- decay are "naturally" abundant.

Neutrinos are thought to have no rest mass—at least experimental results up to the present show that their rest mass, if any, is vanishingly small. They do not take part in either the electromagnetic or the strong interaction, but only in the weak interaction. Thus a neutrino has a high probability of traveling all the way through the Earth without interacting. The interaction probability of neutrinos with detectors is correspondingly slight; while they are not uninteresting, they are hard to detect.

[Top of Page]

2. Radioactive Elements

Cornelius Keller

Elements having only radioactive isotopes are treated in this article. Plutonium, Thorium, Uranium, and Transuranium Elements are separate keywords.

In 1993, 109 chemical elements have been identified (Fig. 17). They can be divided into three groups:

1. 81 elements with one or more stable isotopes
2. 10 naturally occurring elements composed solely of radioactive isotopes
3. 18 elements composed of radioactive isotopes that do not occur in nature

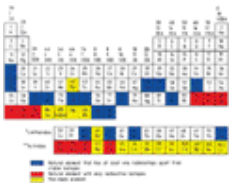

Figure 17. Periodic table of elements with characterization of the different types of radioelements

81 elements possess one or more stable isotopes. The heaviest element with one stable isotope is bismuth (Z = 83). Several stable elements contain naturally occurring long-lived radioisotopes, usually in low concentrations, for example, ^{40}K with $t_{1/2}$ = 1.28×10^9 a and 0.0117 % abundance. However, in the case of the elements indium and rhenium, the most abundant isotopes are radioactive: ^{115}In ($t_{1/2}$ = 8×10^{14} a, 95.7 % abundance), ^{187}Re ($t_{1/2}$ = 5×10^{10} a, 62.6 % abundance).

10 elements occur naturally but are composed only of radioactive isotopes. Besides the primordial elements uranium and thorium with their longest-lived isotopes ^{238}U ($t_{1/2}$ = 4.486× 10^9 a) and ^{232}Th ($t_{1/2}$ = 1.405×10^{10} a), this group comprises the heavy elements protactinium, actinium, radium, francium, radon, astatine, and polonium. Although most of them are only composed of short-lived isotopes, they are found in nature as the decay products of the primordial long-lived radionuclides ^{238}U, ^{235}U, and ^{232}Th. The naturally occurring isotopes are summa-rized in Figure 18 which also shows their formation and decay properties.

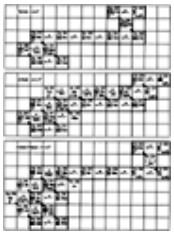

Figure 18. Natural radioactive families

[a] Since the mass numbers within a series can only change by 0 or 4, the parent element and the decay products of the three series must have the atomic masses 4*n*+0, 4*n*+2, and 4*n*+3, where *n* is an integer (the series 4*n*+1 no longer occurs in nature). [b] M=min. [c] UX_2, 1.2 min, β^- 99.85 %, IT γ 0.15 % (IT=isomeric transition). [d] UZ 6.7 h, β^-.

Assuming that the Earth's outer crust (ca. 20 km in depth) contains 4×10^{12} t uranium and about five times the amount of thorium, the following quantities of the longest-lived isotopes of these elements formed by uranium and thorium decay can be calculated.

Polonium	(^{210}Po)	3.3×10^{9} g
Astatine	(^{218}At)	0.1 g
Radon	(^{222}Rn)	9.2×10^{7} g
Francium	(^{233}Fr)	2.1×10^{2} g
Radium	(^{226}Ra)	1.4×10^{13} g
Actinium	(^{227}Ac)	9.1×10^{9} g
Protactinium	(^{231}Pa)	1.4×10^{13} g

The amount of shorter-lived isotopes such as ^{228}Ac, ^{224}Ra, ^{220}Rn, or ^{218}Po is considerably lower.

Because primordial plutonium (^{244}Pu) was detected as early as 1970 and since ^{239}Pu is constantly being produced in uranium deposits via *neutron capture* of ^{238}U, as in nuclear reactors, (^{239}Pu : ^{238}U $\approx 10^{-10} - 10^{-12}$), plutonium can be considered as a naturally occurring element.

*18 elements*have only radioactive isotopes and do not occur in nature. Besides the lighter radioelements technetium (Z = 43) and promethium (Z = 61), this group comprises the transuranium element neptunium (Z = 93) and all elements from Z = 95 (Am) to Z = 109 (Mt = meitnerium).

The elements up to Z = 100 (fermium; longest-lived isotope ^{257}Fm, $t_{1/2}$ = 100 d) are obtained via neutron capture in a nuclear reactor. They can be produced in amounts ranging from kilogram quantities (e.g., neptunium and americium, and also the fission products technetium and promethium) to gram and microgram amounts (e.g., berkelium and californium) down to nanogram quantities (^{257}Fm, the heaviest isotope obtainable via neutron capture).

All isotopes of elements Z > 100 must be produced by *accelerator-based* nuclear reactions with charged particles. Except for rare exotic nuclei (with even shorter half-lifes), they have short half-lifes of $t_{1/2}$ < 1 h. They can be obtained only in tracer quantities (e.g., 10^3 to 10^5 atoms for Z = 101 to Z = 103) or even in form of single atoms for elements $Z \geq 106$.

Superheavy elements — elements with atomic numbers close to the next closed proton shell with Z = 114 — could not yet be identified either in nature or by nuclear reactions, even if some of them may have relatively long half-lifes.

There are few institutions which are able to work with these highly radioactive substances. As a result, the number of publications of general interest, especially review articles, are very scarce; sometimes these articles are published in journals and reports that are not easily available.

Therefore, the interested reader is advised to consult the Gmelin Handbooks of these elements which were published in the 1970s and 1980s. They contain all information in a critically revised form because all chapters were written by experts from all over the world.

Due to decreasing interest in these rare and difficult-to-handle radioelements, no more investigations of special importance — except of ^{99m}Tc for nuclear medicine — have been done in the last decade and, therefore, no review papers have been published recently.

2.1. Promethium

Promethium (Z = 61) was discovered after 1945 as a product of ^{235}U fission by neutrons (J. A. MARINSKY, L. E. GLENDENIN, C. D. CORYELL). The isotopes ^{130}Pm to ^{158}Pm are known, but only β^--active ^{147}Pm [14380-75-7] ($t_{1/2}$ = 2.62 a) has found some industrial importance. ^{145}Pm [15706-44-2] ($t_{1/2}$ = 17.7 a) is the longest-lived promethium isotope but it can only be produced by nuclear reactions involving accelerators and is, therefore, only available in tracer quantities.

^{147}Pm is obtained from high active waste solutions of the PUREX process, together with other fission lanthanides. The waste solution is stored for at least several years and then extracted by a solution of di-(2-ethylhexyl)phosphoric acid in kerosene. Pure ^{147}Pm is subsequently separated by adsorption to a cation exchanger followed by elution with solutions of complexing agents such as nitrilotriacetic acid or ethylene diaminotetraacetic acid.

In its chemical and physical behavior, promethium (electronic configuration [Xe]4 $f^{14}6s^2 6p$) is a typical trivalent rare earth element.

Promethium crystallizes in the double-hexagonal close-packed (dhcp) structure (a = 365 pm, c = 1165 pm) and has a melting point of 1168 °C. It is produced either by reduction of PmF_3 using calcium as a reductant

$$2\,PmF_3 + 3\,Ca \rightarrow 2\,Pm + 3\,CaF_2$$

or in purer form by reduction of Pm_2O_3 with metallic thorium and subsequent sublimation and condensation.

$$2\,Pm_2O_3 + 3\,Th \xrightarrow[10^{-4}\,Pa]{1600\,°C} 4\,Pm\uparrow + 3\,ThO_2$$

Uses. All uses of ^{147}Pm are based on its inherent radioactivity. The heat released on ^{147}Pm decay (0.3 W/g ^{147}Pm) can be transformed into electric energy in isotope batteries via the thermoelectric effect. Backscattering of the electrons emitted by ^{147}Pm is applied for determination of the thickness of material that is composed of elements with low Z, for example a plastic cover on a metallic substrate.

Intimate mixtures of $^{147}Pm_2O_3$ and phosphors such as ZnS(Cu) are widely applied as luminescence materials for watches and space modules.

Handling. ^{147}Pm is a soft β emitter ($E\,\beta_{(max)}$ = 225 keV) so no special radiation shielding measures are necessary. However, when handling solid compounds in powder form tight glove boxes must be used to avoid contamination.

2.2. Technetium

Technetium (Tc, Z = 43), the lightest radioelement was discovered by C. PERRIER and E. SEGRÈ in 1937 after irradiation of molybdenum with deuterons at the Berkeley cyclotron (United States). SEGRÈ took the irradiated sample to Palermo (Italy) where he and PERRIER probably separated the isotope ^{97m}Tc ($t_{1/2}$ = 90 d).

In 1992, all isotopes between mass numbers M = 89 [prepared by the reaction ^{60}Ni(^{32}S, p2n), with $t_{1/2}$ = 12.8 s] and M = 111 (prepared by fission of ^{238}U after bombardment with 20 MeV protons, $t_{1/2}$ = 0.30 s) were identified. ^{98}Tc [32025-58-4] ($t_{1/2} = 4.2\times10^6$ a) is the longest-lived isotope but it cannot be obtained in weighable amounts. The isotope ^{99}Tc [14133-76-7] ($t_{1/2} = 2.14\times10^5$ a) is used for chemical and physical studies of technetium and obtained by fission of ^{235}U with thermal neutrons in ca. 6 % yield. This corresponds to formation of 2.8 g technetium per day and 100 MW_{th} power of a nuclear reactor (i.e., about 30 kg/a in a modern 1300 MW_{el} nuclear power reactor). ^{99}Tc can be recovered from high active waste solutions after reprocessing of spent fuel by the PUREX process.

The most important technetium isotope, however, is ^{99m}Tc ($t_{1/2}$ = 6.0 h) which is the most frequently used radionuclide in nuclear medicine. Due to its short half-life, it must be separated from its mother nuclide ^{99}Mo ($t_{1/2}$ = 66 h) by the use of a generator at the place of application.

Several types of ^{99}Mo/^{99m}Tc generators exist which can supply sterile ^{99m}Tc in high purity and good yield, e.g., by elution of technetium with NaCl solution from an alumina column loaded with ^{99}Mo (Fig. 19).

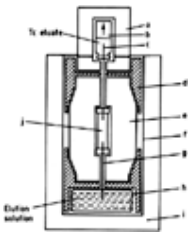

Figure 19. ^{99m}Tc generator "Tecegen" (Hoechst)

a) Eluate shielding (lead glass); b) Evacuated eluate bottle, 13 mL, 5mL calibration; c) Canula for eluate; d) Adapter for eluate (plastic); e) Transport shielding; f) Transport box; g) Canula for elution solution; h) Box containing elution solution in plastic container; i) Additional shielding; j) Generator column for fission Mo

^{99}Mo [14119-15-4] itself can be produced by two processes (see Section Radionuclide Production from Reactors):

1. Irradiation of ^{98}Mo-enriched molybdenum in a nuclear reactor via neutron capture ^{98}Mo(n, γ)^{99}Mo
2. Irradiation of highly ^{235}U-enriched (≥ 20 %) uranium in a nuclear reactor and separation of the fission product ^{99}Mo (fission yield ≈ 6 %)

The second type of production is preferred because it yields ^{99}Mo with a high specific activity so that the eluted ^{99m}Tc is obtained in high concentration. In contrast, the product of ^{98}Mo irradiation only delivers ^{99}Mo with a much lower specific activity and, therefore, requires larger generators.

Nowadays, an adequate supply of ^{99m}Tc is provided by the so-called ^{99m}Tc generators. These generators are efficient and reliable and differ principally from common generators in the technique of separating technetium from molybdenum. Preferred separation methods are the adsorption of $^{99}MoO_4^{2-}$ onto a solid medium and elution of $^{99m}TcO_4^-$ with saline solution, selective extraction of $^{99m}TcO_4^-$ into organic solvents, and sublimation of ditechnetium heptaoxide at high temperatures.

Technetium has the ground state electronic configuration of [Kr]4*d* ^{6} 5*s*. Its chemical properties are more similar to rhenium than to manganese. Valency states between – 1 and +7 are known, the most stable ones being Tc(IV) and Tc (VII). The most important redox potentials in acidic solution at 298.2 K are:

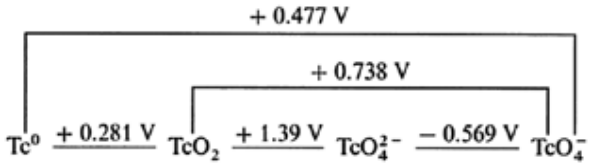

Metallic technetium can be easily prepared either by reduction of NH_4TcO_4 or TcO_2 with hydrogen at 700 – 900 °C or by electrodeposition from a 1 M solution of NH_4TcO_4 in sulfuric acid. Its most important *physical properties* are:

mp	2250 ± 50 °C
Density	11.47 g/cm^3
Crystal structure	hexagonal close-packed
Lattice dimensions (25 °C)	*a* = 274.07 pm
	c = 439.80 pm
Critical temperature (superconducting)	7.73 ± 0.02 K
Debye temperature	454 K
Standard entropy of formation	31 ± 0.8 J mol^{-1} K^{-1}
Magnetic susceptibility (25 °C)	270×10^{-6}
Specific heat C_p (1000 K)	0.29 J g^{-1} K^{-1}

Chemical Properties. Solid technetium is resistant to oxidation, but in sponge or powder form it is readily oxidized to the volatile heptaoxide when heated in air. The metal dissolves in dilute or concentrated nitric acid, concentrated sulfuric acid, and bromine water, but not in hydrochloric acid. Technetium also dissolves in aqueous acid hydrogen peroxide solutions. It is oxidized by these solvents to pertechnetate, but in the case of bromine water $[TcBr_6]^{2-}$ is also formed.

TcO_2 reacts with fluorine to yield TcO_3F. Contrary to rhenium, no heptafluoride of technetium is known. Only TcF_5 and TcF_6 are obtained, for example, by fluorination of technetium metal at 400 °C. The corresponding reaction with chlorine or bromine at 550 °C produces the tetrahalides $TcCl_4$ and $TcBr_4$. These halides can form hexahalide complexes with an octahedral $TcCl_6^{2-}$ arrangement that are more stable than the corresponding tetrahalides, for example K_2TcCl_6. $TcOCl_4$ and TcO_3Cl are formed if technetium reacts with chlorine that contains oxygen. Sulfur reacts with technetium at elevated temperature to give the disulfide, and carbon to give TcC. A technetium hydride of the composition $TcH_{0.73}$ is formed at 300 °C under a hydrogen pressure of ca. 0.2 GPa.

Only a few compounds have been prepared so far, with Tc in low oxidation states, including:

Tc(–1) $HTc(CO)_5$

Tc(0) $Tc_2(CO)_{10}$, $CoTc(CO)_9$

Tc(+1) $K_5Tc(CN)_6$, $(C_6H_5)_2TcH$, $Tc(CO)_5Cl$

Tc(+2) $[(C_5H_5)_2Tc]_2$

Tc(+3) [Tc(*o*-phenylene-bis-dimethylarsine)$_2Cl_2$]Cl

Tc(+4) TcO_2, M_2TcX_6 (X = F, Cl, Br, I, SCN)

Oxidation of technetium metal by oxygen at 500 °C yields pale yellow Tc_2O_7 (*mp* 119.5 °C, *bp* 310.6 °C) which reacts with

water vapor to form dark red $HTcO_4$. Many pertechnetates such as $LiTcO_4$ or the sparingly soluble $CsTcO_4$ have been prepared and characterized.

Reduction of Tc(VII) solutions or thermal decomposition of NH_4TcO_4 at 450 °C in vacuum gives TcO_2 which has a distorted rutile structure. Brown-black Tc_2S_7 is precipitated from TcO_4^- solution by gaseous H_2S.

Uses. Technetium itself and some of its compounds and ions have properties that should make them suitable for broad industrial applications. However, their inherent radioactivity prevents their general use except in special applications, for example, ^{99m}Tc in nuclear medicine.

Metallic technetium supported on Al_2O_3 or SiO_2 is an excellent catalyst for hydrogenation, dehydrogenation, dehydrocyclization, and cracking reactions, particularly of C_6 and C_7 species. The reforming of naphtha to gasoline with higher aromatic content is catalyzed by technetium-promoted Cr_2O_3 catalysts, at a lower reaction temperature than other promoted catalysts. Technetium can, therefore, partly substitute for expensive platinum catalysts, especially as technetium catalysts are highly resistant to sulfur poisoning.

Metallic technetium is a weak-coupling type II superconductor with a high critical temperature (T_c = 7.73 K). Some technetium alloys have still higher critical temperatures up to $T_c \approx 15.0$ K for a technetium – molybdenum alloy with 60 atom % technetium.

TcO_4^- ions are excellent corrosion inhibitors even in the presence of corrosive media and at temperatures ≤ 250 °C. At pH ≈ 6 and 10 ppm Cl^-, a solution of 50 ppm technetium (≈ $5 \cdot 10^{-4}$ M Tc) suppresses corrosion of a 0.1 % carbon steel even after 20 years. The corrosion rate is at least four orders of magnitude less than in the absence of TcO_4^- ions. The surface concentration of technetium on the steel corresponds to about a tenth of a monolayer. Furthermore, technetium-coated stainless steel does not corrode in seawater.

^{99}Tc is used very often as a standard source of β radiation for radiometry and dosimetry. Such sources are safe and convenient to handle because of the lack of γ radiation.

In nuclear medicine, chemical compounds of ^{99m}Tc are widely used for radionuclidic imaging of the brain, liver, lung, and skeleton, and to a lesser extent in thyroid scintigraphy. ^{99m}Tc-labeled radiopharmaceuticals have been developed that are suitable for both static and dynamic evaluation of renal and cardiac pathology (see Chap. Production of Artificial Radionuclides). Sales of ^{99m}Tc-labeled products account for about 75 % of the total sale of radiopharmaceuticals in the United States; similar trends are observed in Europe.

Handling. ^{99}Tc is a soft β emitter so no special radiation protection measures are necessary. Tight glove boxes must be used to avoid contamination of work places and incorporation when handling powdered samples of ^{99}Tc compounds. Lead shielding is necessary for $^{99}Mo/Tc$ generators because ^{99}Mo also emits γ-rays.

2.3. Polonium

Element 84 was discovered in 1894 by P. and M. CURIE. Its discovery was possible because of its inherent radioactivity.

In 1993 all isotopes between mass numbers M = 193 and M = 218 are known including many relatively long-lived isomers. The longest-lived isotope ^{209}Po [15735-81-6] ($t_{1/2}$ = 102 a) is only available through charged-particle nuclear reactions and therefore not produced in weighable amounts. Nearly all the chemical and physical studies on polonium were performed with the pure α emitter ^{210}Po ($t_{1/2}$ = 138 d).

Even though polonium was first discovered by the separation of ^{210}Po from uranium ores, due to the very small amount of ^{210}Po (equilibrium concentration 7.4×10^{-11} g ^{210}Po/g ^{238}U) in ores, nowadays most ^{210}Po is isolated in a different way. It is either separated from "old" radium samples or more recently by irradiation of very pure bismuth with neutrons in a nuclear reactor, even if the neutron capture cross section σ is low:

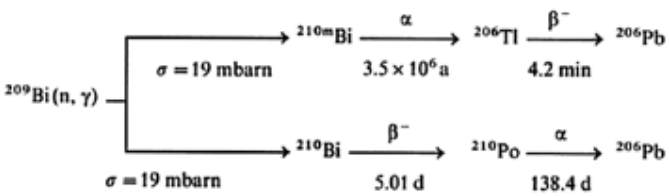

One year of irradiation of 1 kg bismuth at a neutron flux of 10^{14} n cm^{-2} s^{-1} gives 25 mg ^{210}Po (ca. 4×10^{12} Bq). The isolation of such quantities from natural sources is practically impossible. Following irradiation of bismuth, ^{210}Po is separated from molten bismuth by vacuum distillation. Final purification of polonium generally is accomplished by its spontaneous deposition on silver or copper powders. *Metallic polonium* is then easily prepared by sublimation of this deposit at about 700 °C. It is also obtained by reduction of polonium sulfide with hydrogen at 500 °C. Polonium sulfide itself is precipitated from an acid solution of polonium ion by gaseous H_2S.

Polonium with electronic configuration [Xe]4 $f^{14}5d^{10}6s^2 6p^4$ is the heavy homologue to tellurium. The oxidation states – 2, +2, +4, and +6 are known, but Po(IV) (ionic radius 102 pm) is the most stable one. Po(VI) (ionic radius 67 pm) is easily reduced to a lower oxidation state by its own radiation.

Important *physical properties* of polonium metal are:

mp	≈ 250 °C
bp	962 °C
Atomic radius	164 pm
Vapor pressure	
(438 – 745 °C)	
log (*p*/Pa) =	(964.36 ± 0.9064) – (716860 ± 893)/ *T* (°C)
Structure	
α Po (<18 °C)	cubic, *a* = 335.9 pm
β Po (>54 °C)	rhombohedral, *a* = 336.9 pm, α = 98.23°
Density	
αPo	9.142 g/cm^3
β Po	9.352 g/cm^3

Chemical Properties. Polonium is soluble in HCl and forms Po(II), in HNO_3 and aqua regia Po(IV) is produced.

Po(VI) solutions are obtained by the oxidation of Po(IV) with Cl_2, $K_2Cr_2O_7$, or Ce(IV) – HNO_3. SO_2 and N_2H_4 reduce Po (IV) to Po(II), by reduction with $SnCl_2$, Ti(III), and $NaHSO_3$, a precipitate of finely divided metallic polonium is formed.

Compounds. Polonides of several metals with the composition MPo (e.g., M = Mg, Mn, Y, rare earth element) are prepared by the interaction of polonium vapor with powdered metals. Most of these polonides dissociate at ca. 700 °C, only rare earth polonides are stable up to 1000 °C.

Polonium monoxide, PoO [19268-61-2], is obtained by thermal decomposition of polonium selenide. The more stable PoO_2 [7446-06-2], is formed by thermal decomposition of, for example, $PoO(OH)_2$ or Po(IV) nitrate, or by reaction of metallic polonium and oxygen. It can be sublimed in an oxygen stream at 885 °C but decomposes in vacuum at 500 °C. PoO_2 vapor reacts with rare earth oxides to form compounds such as $Pr_2O_3 \cdot PoO_2$ and $Dy_2O_3 \cdot PoO_2$.

Halides such as $PoCl_4$ [10026-02-5] and $PoBr_4$ [60996-98-7] are formed when polonium metal reacts with chlorine or bromine at ca. 100 °C. At 200 °C these halides decompose to the corresponding polonium dihalides. Complex halides such as $M_2^{I}PoX_6$ (M = K, Rb, Cs; X = Cl, Br, I) of Po(IV) and $MPoCl_3$ and M_2PoCl_4 of Po(II) are also known.

Due to the high radioactivity of ^{210}Po, relatively unstable compounds such as PoH_2, PoF_6, PoO_3, and organometallic compounds could be identified only in tracer amounts but not in quantities that are needed for detailed investigations of their properties.

In aqueous solution, polonium ions are easily hydrolyzed even at pH ≈ 1, if not complexed by, for example, oxalic, citric, or tartaric acids. Po(IV) in particular, is readily extracted from aqueous solutions by, e.g., tri-*n*-butyl phosphate, dithizone – $CHCl_3$, and amines. Polonium is adsorbed as complex ions by anion exchangers in strong acid solution — elution must be performed rapidly due to the fast decomposition of the ion exchanger by α activity.

Uses. ^{210}Po is the only isotope that finds broad application. It is used as an energy source for radioisotope batteries. This isotope has the advantage that it does not emit γ-rays therefore heavy weight shielding of the batteries can be avoided. The disadvantage is the short half-life of ^{210}Po. ^{210}Po (140 W_{th}/g) was especially used as an energy source in Soviet space missions, e.g., in Lunochod-1 and Lunochod-2. About 10^{15} Bq ^{210}Po are needed for a radioisotope battery that delivers 1 kW electric power. The polonium is mostly provided in the form of rare earth polonides.

^{210}Po is used also as an α emitter for (α, n) neutron sources. A BeO – ^{210}Po source emits up to 93 neutrons per 10^6 α particles.

Handling. Tracer amounts of ^{210}Po can be handled in well-working fume hoods but even microgram quantities of ^{210}Po require handling in tight glove boxes due to unavoidable α recoil and the consequent contamination of the working area.

2.4. Astatine

Astatine, the heaviest halogen element (Z = 85) was discovered in 1940 by D. R. CORSON, K. R. MCKENZIE, and E. SEGRÈ as a product of bombardment of bismuth by 20 MeV α particles via the reaction ^{209}Bi (α, 2n)^{211}At. Minute amounts of short-lived ^{215}At, ^{218}At, and ^{219}At are intermediates in the natural ^{235}U- and ^{238}U-decay chains.

All astatine isotopes of mass numbers between M = 196 and M = 223 have been identified. Only the easily obtainable isotopes ^{209}At [5755-38-1] ($t_{1/2}$ = 5.5 h), ^{210}At [18830-37-0] ($t_{1/2}$ = 8.1 h), and ^{211}At [15755-39-2] ($t_{1/2}$ = 7.2 h) are of importance, especially for use as radiopharmaceuticals in nuclear medicine. Irradiation of *bismuth* with a l-μA beam of 43 MeV α particles for 1 h yields ca. 1.5×10^9 Bq ^{210}At plus 1.1×10^9 Bq ^{211}At, corresponding to ca. 2 ng of the element. To separate the produced astatine, the irradiated bismuth is heated to 660 °C at 0.1 Pa. The volatilized astatine can be condensed in a trap cooled by liquid nitrogen. A mixture of all three longer-lived astatine isotopes is formed by spallation reactions when *thorium metal* is bombarded with 600 MeV protons. This method gives higher yields of astatine [2×10^7 Bq At μA^{-1} h^{-1} $g(Th)^{-1}$] and is used as a routine procedure at Dubna in the CIS.

Because of its high specific activity, astatine (electronic configuration [Xe]$4f^{14}5d^{10}6s^2 6p^5$) can only be handled in tracer quantities. Properties of the element (such as *mp* 244 °C and *bp* 309 °C) are only estimated figures.

Chemical Properties. Tracer studies have given, however, a good indication of the properties of the elements and its ions, especially in solution. Astatine occurs in oxidation states –1, 0, +1, +3, +5, and +7 and it has also been shown that astatine is more electropositive than iodine [redox potentials X^-/X_2(aq) are 0.62 V for iodine and 0.3 V for astatine]. The molecule At_2 has never been observed, however, because the concentration and thus chance of impurities to react with an astatine atom is generally much higher than the concentration of astatine itself.

Therefore, At(0) was only identified in the form of AtI. It is easily adsorbed by metal surfaces and extracted from aqueous solutions by, for example, benzene or CCl_4. Strong reducing agents form At^-, oxidation by $K_2Cr_2O_7$, for example, yields At^+ in the form of its aquo complex. At(V) is formed by oxidation of lower valency states of astatine by $(NH_4)_2S_2O_8$ or IO_3^-/IO_4^-, At(VII) by oxidation with YeF_2 in warm alkaline solution.

Organometallic compounds such as C_2H_5At or C_6H_5At were synthesized in tracer quantities by using iodine as nonisotopic carrier as well as, for example, astatine acetic acid ($AtCH_2COOH$) or astatine-labeled radiopharmaceuticals (γ-globulin, albumin, etc.). In most cases isotope exchange reactions with corresponding iodine compounds are used for their synthesis.

Uses. Astatine-labeled organic compounds have high potential as radiopharmaceuticals, both for diagnostics and therapy. These compounds can be, for example fixed onto antibodies for immunodiagnostic studies. It is very important, however, to synthesize compounds with a strong fixation of astatine. Otherwise free astatine in ionic form is accumulated in the thyroid similar to iodine and colloidal astatine is accumulated in the liver, resulting in unwanted radiation exposure to these organs.

Handling. In tracer amounts, astatine can be handled safely in well-aerated fume hoods; incorporation of astatine has to be strictly avoided.

2.5. Radon

Radon, the heaviest noble gas, was discovered in 1900 as a decay product of radium by F. E. DORN.

Radon isotopes with atomic masses between 199 and 226 are known, of which ^{222}Rn [14859-67-7] ($t_{1/2}$ = 3.82 d), ^{220}Rn [22481-48-7] ($t_{1/2}$ = 55.6 s), and ^{219}Rn [14835-02-0] ($t_{1/2}$ = 3.96 s) are members of natural decay series. On average, they contribute to about half of the natural radiation dose the population is exposed to.

Despite the fact that the chemical behavior of radon is only known from tracer experiments with the longest-lived ^{222}Rn, there is no doubt that radon is the most reactive noble gas. Reaction of radon with fluorine or halogen fluorides, such as BrF_3, is likely to produce RnF_2 which can be reduced to radon by hydrogen only at temperatures > 500 °C. Radon fluoride can be sublimed at about 250 °C. It also seems to be possible to oxidize radon with O_3.

There are no known uses of radon, except application of ^{222}Rn capsules for the treatment of cancer. The release of radon from solid substances (“emanation method” by O. HAHN) is often used to study structural changes of solids.

SbF_5 in the presence of ClF_3 and $N_2F_2Sb_2F_{11}$ might be suitable for the removal of radon from the air in uranium mines in order to reduce radiation dose to workers by the formation of RnF.

2.6. Francium

Element 87 is the most unstable of all naturally occurring elements. It was discovered in 1939 by M. PEREY. ^{223}Fr [51712-59-5] ($t_{1/2}$ = 21.8 min) is the most stable of all 27 isotopes.

Relatively pure ^{223}Fr solutions are obtained by its elution with NH_4Cl – CrO_3 from a cation exchanger loaded with ^{227}Ac. The solution is further purified by passing it through a SiO_2 column loaded with $BaSO_4$.

Francium is only monovalent and shows a very similar behavior to the other heavy alkali elements. Francium has the most negative standard potential of all elements.

No industrial applications or uses in nuclear medicine have yet been found.

2.7. Radium

Radium (Z = 88) was discovered in 1898 by P. and M. CURIE and G. BÉMONT. ^{226}Ra [13982-63-3] ($t_{1/2}$ = 1600 a) is the longest-lived of all the 25 radium isotopes known to date. Its concentration in uranium ores is 0.36 µg per gram uranium. ^{223}Ra [15623-45-7], ^{224}Ra [13233-32-4], and ^{228}Ra [15262-20-1] also occur in nature.

According to its electronic configuration [Rn]$7s^2$, radium shows the typical behavior of an alkaline earth element with a large ionic radius. It is obtained from uranium ores by a mixed precipitation – ion exchange procedure.

Radium metal is prepared by reduction of the oxide by aluminum at 1200 °C in vacuum. The pure metal has a cubic body-centered structure (a = 515 pm) such as barium and a *mp* of 700 °C. The known compounds, including the halides RaX_2 (X = F, Cl, Br, I), the oxide RaO, the carbonate $RaCO_3$, and the ternary oxides $RaCeO_3$ and $RaMO_4$ (M = Cr, Mo, W) are isostructural with corresponding barium compounds. Because of its large ionic radius, radium forms weak complexes and is not well extracted from aqueous solutions except at high pH values.

Uses. The use of ^{226}Ra as a γ emitter in nuclear medicine for cancer therapy lost its importance when larger and cheaper ^{60}Co sources became available. Also most (α, n)-neutron sources based on ^{226}Ra were replaced by ^{241}Am – Be sources even if the number of emitted neutrons in ^{226}Ra sources are higher (for $RaBeF_4$: 1.84×10^6 neutrons per second and per gram ^{226}Ra).

Radium is mainly used to prepare ^{227}Ac via neutron capture in a nuclear reactor.

Handling. ^{226}Ra requires special precautions to work with because it is considered the most toxic radioelement. Handling of ^{226}Ra is possible only in tight glove boxes where the circulating air stream must also be treated to avoid the escape of decay product ^{222}Rn to the environment. Special care must be taken when opening disused radium ampoules because of possible overpressure formed by the decomposition of water.

2.8. Actinium

The element with Z = 89 was discovered in 1899 by A. DEBIERNE, but F. O. GIESEL in 1902 was the first to have prepared pure actinium. Of all 25 known isotopes, only the longest-lived ^{227}Ac [51712-63-1] ($t_{1/2}$ = 21.77 a) has some importance. In uranium ores its amount is too low to be recovered (2×10^{-10} g per gram uranium), so nearly all ^{227}Ac is obtained by irradiation of ^{226}Ra. Three months irradiation at a neutron flux of 5×10^{14} n cm^{-2} s^{-1} transforms 2 % of ^{226}Ra to ^{227}Ac. In the 1970s, about 18 g $^{227}Ac_2O_3$ were prepared by this method. The ^{227}Ac obtained was separated from ^{226}Ra and decay products by a precipitation – ion exchange process.

Because of its electronic configuration [Rn]$6d\,7s^2$ and ionic radius (r = 111 pm for Ac^{3+}), actinium closely resembles lanthanum, but only few compounds [e.g., the oxide Ac_2O_3, the sulfide Ac_2S_3, the phosphate $AcPO_4 \cdot 0.5\,H_2O$, the halides AcX_3, and oxide halides AcOX (X = F, Cl, Br, I)] have been synthesized up to now.

Actinium metal is obtained by reduction of Ac_2O_3 with thorium at 1750 °C and 10^{-4} Pa. It crystallizes in the cubic β-lanthanum type structure with a lattice constant of a = 531.5 pm. In solution only Ac^{3+} occurs. The stability of Ac^{3+} complexes is low due to the large ionic radius. Because of this, actinium can only be extracted from aqueous solutions at relatively high pH values with low distribution coefficients. Most solution experiments were performed by using ^{228}Ac ($t_{1/2}$ = 6.13 h), separated from ^{228}Ra ($t_{1/2}$ = 5.7 a).

Uses. ^{227}Ac is a radionuclide with high specific power (14.3 W_{th}/g) and besides ^{244}Cm, is the only radioisotope used for thermionic transformation of the heat of decay into electricity. For a thermionic battery that transforms 12 % of heat into electricity, about 17 g $^{227}Ac_2O_3$ (equivalent to about 5×10^{13} Bq) are needed to obtain 250 W_{th}. Due to the instability of thermionic batteries compared to thermoelectric devices, the interest in ^{227}Ac is currently decreasing.

Handling. Handling of ^{227}Ac is possible only in tight glove boxes, gram-amounts require hot cells with γ shielding. Well-aerated fume hoods are sufficient for the handling of trace quantities of all actinium isotopes.

2.9. Protactinium

The element with Z = 91 was discovered in 1913. K. FAJANS and O. GÖHRING were probably the first to have separated pure protactinium.

^{231}Pa [14331-85-2] ($t_{1/2}$ = 32 480 a) is the most important of all 24 known isotopes, even though ^{233}Pa [13981-14-1] is of special importance because it is formed as an intermediate in the production of fissile ^{233}U from nonfissile ^{232}Th by

neutron capture

$$^{232}\mathrm{Th}(n,\gamma)\,^{233}\mathrm{Th}\xrightarrow[22.3\ \mathrm{min}]{\beta^-}\,^{233}\mathrm{Pa}\xrightarrow[27.0\ \mathrm{d}]{\beta^-}\,^{233}\mathrm{U}$$

$$\sigma = 7.4\ \mathrm{barn}$$

This process is used in thorium high temperature reactors (THTR, see ⟶ Nuclear Technology) which utilize the fivefold occurrence of thorium in nature compared to uranium.

The separation of ^{231}Pa from residues of uranium production is a difficult process. Nevertheless, around 1960 ca. 100 g pure $^{231}Pa_2O_5$ were isolated from residues of the uranium production at Springfield, United Kingdom, that contained 3.7 ppm ^{231}Pa. In future, larger quantities of ^{231}Pa might be recovered from nuclear waste originating from reprocessing of thorium fuels, because it is formed via

$$^{232}\mathrm{Th}(n,2n)\,^{231}\mathrm{Th}\xrightarrow[25.6\ \mathrm{h}]{\beta^-}\,^{231}\mathrm{Pa}$$

by the fast neutrons in a THTR nuclear reactor.

With its electronic configuration [Rn]$5f^2 6d^1 7s^2$, protactinium is the first element to posses 5 *f* electrons in the ground state. Both in the solid state and in solution protactinium can occur in the oxidation state Pa^{4+} (r = 96 pm) and Pa^{5+} (r = 90 pm). In its chemical and physical behavior protactinium is much more similar to tantalum than to typical 5 *f* elements.

Protactinium metal is best prepared by the Van Arkel method using a PaC – I_2 mixture (⟶ Hafnium and Hafnium Compounds – Refining). α Protactinium is tetragonal-body centered (a = 393.1 pm, c = 323.6 pm), β protactinium is cubic face-centered (a = 501.9 pm). The melting point of protactinium is 1572 ± 20 °C.

Compounds. The solid state chemistry of protactinium is quite well known. Binary compounds exist such as the tetrahalides PaX_4 and pentahalides PaX_5 and complex compounds including the sparingly soluble K_2PaF_7. The very stable white Pa_2O_5 occurs in at least five modifications. The reduction of Pa_2O_5 with pure hydrogen at 1550 °C yields black PaO_2. A protactinium peroxide of still unknown composition can be precipitated from diluted acids. This method is used to separate protactinium from tantalum and niobium which form soluble peroxide complexes.

Pa(V) is hydrolyzed in aqueous solution even in strong acids, if not complexed, for example, by F^- ions. No Pa^{5+} exists in solution; only partly hydrolyzed species with two positive charges (in ca. 1 M HCl) or one positive charge (in 10^{-3} – 10^{-1} M H^+) could be identified. Pa(V) is readily adsorbed onto precipitates, surfaces (even metal surfaces), graphite, quartz, etc. The partly hydrolyzed protactinium species form very stable complexes with most organic ligands. Pa(V) can be extracted from aqueous solution by amines and most neutral and acid phosphoric and phosphonic extraction agents. The distribution coefficients are high.

Pa(IV) solution is obtained by reduction of Pa(V) solutions using chemical (e.g., zinc) or electrochemical methods. The behavior of Pa(IV) is very similar to all other actinides of oxidation state IV, however, its tendency to hydrolyse is much higher.

Due to the unavoidable hydrolysis and formation of radiocolloids, solutions of protactinium in various oxidation states must be freshly prepared before use. The large number of unexplainable results found in the literature are mostly due to using degraded protactinium solutions with partly hydrolyzed, nonspecified protactinium species that change composition by time, even in 6 M HCl.

Uses. Protactinium plays an important role in the THTR fuel cycle, especially because of the high neutron absorption cross section of ^{233}Pa ($t_{1/2}$ = 27.0 d) that lowers the amount of fissile ^{233}U formed. Therefore, several processes, for example, for a “molten salt reactor”, have been developed to separate protactinium continuously during reactor operation with the consequence that ^{233}Pa will decay outside the reactor core and, therefore, not absorb neutrons. However, many technological problems have to be solved before practical application.

Handling. ^{231}Pa must be handled in tight glove boxes, whereas tracer chemistry of protactinium (e.g., with ^{233}Pa) can be performed in well-aerated fume hoods.

[Top of Page]

3. Production of Artificial Radionuclides

Walter Wolf and Jashovam Shani

Introduction (see Chap. Introduction). Artificial radionuclides can be produced when nuclei of stable elements are modified as a consequence of their interaction with high energy particles. Such particles may be neutral, such as neutrons, or charged, either positively (e.g., protons, deuterons, etc.) or negatively. The stable nuclide that is subject to particle bombardment is called the *target*, and the bombarding particles can arise either as the result of a nuclear decomposition

(such as neutrons in a reactor), or through acceleration of a charged particle to the desired energy level. Theoretically, any elemental particle or nucleus can be accelerated, but radionuclide production for biomedical and chemical purposes (as opposed to the study of nuclear properties) is limited to a finite number of particles and energy ranges.

A nuclear reaction leading to an artificial radionuclide can be formulated as follows:

$$ {}^{M}_{A}\mathrm{T} + {}^{M'}_{A'}\mathrm{P} \longrightarrow {}^{M+M'}_{A+A'}\mathrm{CN} \longrightarrow {}^{M''}_{A''}\mathrm{RN} + {}^{M'''}_{A'''}\mathrm{P}_1 + {}^{M''''}_{A''''}\mathrm{P}_2 \quad (3.1) $$

where T is the target nuclide of atomic number A and mass *M*, and where the bombarding particle P has mass *M′* and atomic number *A ′*. An absorptive collision will result in a nucleus in an activated transition state that is called a compound nucleus (CN) [4], usually so unstable that it will decompose into two or more products: a radionuclide RN of mass *M″* and atomic number *A″*, and one (P_1) or more (P_2, . . .) additional particles that carry the excess mass. Finally, an energy term *Q* is included to complete the balance of Equation (3.1). Its value will be positive if energy is released (an exothermic reaction), and negative if energy must be provided (an endothermic reaction).

The following are examples of some of the more common nuclear reactions used in the production of radionuclides. The first illustrates a neutron activation reaction:

$$ {}^{98}_{42}\mathrm{Mo} + {}^{1}_{0}\mathrm{n} \longrightarrow {}^{99}_{42}\mathrm{Mo}^{**} \longrightarrow {}^{99}_{42}\mathrm{Mo} + E(\gamma) \qquad (3.2) $$

where excess energy is released as two primary ϒ quanta of 6.92 and 0.78 MeV. Such a reaction is normally abbreviated in nuclear chemical notation as

$$ {}^{98}\mathrm{Mo}(\mathrm{n},\gamma){}^{99}\mathrm{Mo} \qquad (3.3) $$

where the expression in parentheses indicates first the bombarding particle followed by the emitted particle(s) or photons. When mass particles (P_1, P_2) are emitted, primary photons are usually not listed.

A simple charged particle reaction is illustrated by one of the most common reactions used in the production of ^{18}F:

$$ {}^{18}\mathrm{O}(\mathrm{p},\mathrm{n}){}^{18}\mathrm{F} \qquad (3.4) $$

Because the entering particle — the positively charged proton — must overcome the coulombic barrier of the oxygen nucleus (which has a charge of +8), there is a minimum threshold energy required for the bombarding particle to allow this reaction to occur. While most of the excess energy of the compound nucleus is dissipated by energy carried by the emitted particle and by primary ϒ quanta released in the reaction, significant energy remains as kinetic energy in the radionuclide itself, a feature responsible for the “hot atom” [5] characteristics of most radionuclides produced in situ.

Nuclear reactors are the most common and efficient devices to produce neutrons required in this type of radionuclide production, while charged particles, such as protons or deuterons (i.e., nuclei of ^{1}H or ^{2}H), must be accelerated to provide them with the energy required by the nuclear reaction in question. Devices for this purpose are called charged-particle accelerators. Both sets of systems were initially developed during the 1930s and then adapted to a large scale during World War II as part of the Manhattan Project [6], [7].

Several key features distinguish the production of radionuclides in a reactor from that in charged-particle accelerators: (1) absorption of *neutrons* usually results in a nucleus with an excess of neutrons, which will therefore disintegrate by emission of an electron (β decay). Absorption of a *charged particle* will result in a nucleus that is neutron-deficient, which will decay by either positron (β^+) emission or electron capture (EC); (2) *thermal neutrons* (E_{av} = 0.025 eV) can penetrate into a nucleus, whereas a significant (threshold) energy is required for *charged particles* to overcome the coulombic barrier; (3) a reactor can usually accommodate many targets for simultaneous irradiation, sometimes over weeks or months, whereas a cyclotron or other charged-particle accelerator can usually accommodate only one or, at most, a small number of targets to be irradiated in some interleaved manner, and then, usually, only for a short period of time.

Another important consideration is that the nature of the nuclear reaction (hence, radionuclide production) may be very much a function of the energy of the accelerated particle. As an example, consider two types of interactions of protons with iodine. When the bombarding protons have low energy (5 MeV < E_p< 20 MeV) the excess energy in the nucleus results only in the ejection of one neutron (Eq. 3.5), whereas at much higher bombarding energies (> 60 MeV) the excess energy leads to ejection of several nucleons (5 neutrons in Eq. 3.6):

$$ {}^{127}\mathrm{I}(\mathrm{p},\mathrm{n}){}^{127}\mathrm{Xe}\ 5\ \mathrm{Mev} < E_{\mathrm{P}} < 20\ \mathrm{MeV} \qquad (3.5) $$

$$ {}^{127}\mathrm{I}(\mathrm{p},5\,\mathrm{n}){}^{123}\mathrm{Xe}\ 60\ \mathrm{MeV} < E_{\mathrm{P}} < 80\ \mathrm{MeV} \qquad (3.6) $$

Other reactions may occur in this system at intermediate proton energies (e.g., p, 2n; p, 3n; p, pn). A plot of production yields for various nuclear reactions versus energy of the incident particle is called an excitation function. However, when the energy of the bombarding particle exceeds 100 MeV, then the stability and organization of the compound nucleus is so severely modified that various nuclear fragments are emitted. Such reactions are called *spallation reactions*, and will be discussed in Section Radionuclide Production from Reactors.

Finally, absorption of a neutron by ^{235}U literally represents the drop that spills the glass: the resulting compound nucleus is so unstable that it breaks up into a large number of nuclides and several neutrons, and a considerable amount of energy is

released. This process is known as *nuclear fission* (see also ⟶ Nuclear Technology – Physical Processes), and has been used for the production of a number of radionuclides, including ^{60}Co, ^{90}Sr, ^{99}Mo, and ^{131}I.

Probably the major consideration in choosing a method for the production of a given radionuclide is the expected yield. With Equation (3.7) the yield (in curies) of a given radionuclide can be calculated at the end of neutron bombardment (EOB). Similar equations can be written for radionuclide production in a charged particle accelerator:

$$A = N \cdot \sigma \cdot F \cdot S \tag{3.7}$$

where A is the activity, in curies, at time t; N the number of atoms of target nuclei in the sample; σ the production cross section in cm^2 (1 barn = 10^{-24} cm^2); F the particle flux in neutrons cm^{-2} (n cm^{-2} s^{-1}); S the saturation factor = 1 – exp (– 0.693 · $t/t_{1/2}$); $t_{1/2}$ the physical half-life of the radionuclide produced; and t the time of bombardment or irradiation.

The activity remaining at time d will be $A_d = A_t \cdot D$, where D, the decay factor, is $D = \exp(-0.693 \cdot t/t_{1/2})$.

Several of the factors in the above equation depend on physical properties of the elements, such as the cross section σ and the $t_{1/2}$ of the resulting radionuclide. The higher the cross section σ, the greater the yield. Other parameters in the above equation can be adjusted to the reaction in question and required yield. The number of atoms undergoing transmutation is $N = (6.02\times10^{23} \cdot \text{mass} \cdot \%\text{isotopic abundance})/ A_r$. If the isotopic abundance of the target nuclide is less than 1, then the yield of radionuclide production can be increased by using a target enriched in the desired nuclide. This not only increases the yield of radionuclide production, but may also increase the purity of the desired radionuclide by minimizing the production of unwanted radionuclides. Another parameter that can be adjusted is the flux of neutrons. The higher the flux, the greater the yield. Finally, the time of bombardment is important: the combined rates of formation and decay eventually lead to a steady-state condition. For this reason, many radionuclide production yields are expressed as yields at half-saturation. Similar considerations apply to production in a charged-particle accelerator.

The highest capital investments are associated with the machines that produce the bombarding particles (nuclear reactors or charged particle accelerators). Some of the characteristics of these devices are examined, followed by examples of how a number of key radionuclides have been produced, and how the production of each has been optimized in the system of choice.

3.1. Nuclear Reactors

A reactor intended for radionuclide production is a device that produces neutrons as a consequence of nuclear fission. An average of 2.2 neutrons is produced per fission of a ^{235}U nucleus induced by slow neutrons, and the energies of the generated neutrons range from 14 MeV to a fraction of an electron volt. Operationally, such neutrons are classified as fast (E > 1 MeV), thermal or slow (E_{av} = 0.025 eV), or epithermal (of intermediate energies). Although the various types of neutrons can lead to different radionuclides, there is no way to separate the neutrons, so one must be well aware of the mix available at various positions in each reactor. However, most radionuclide production is achieved using thermal neutrons. Insertion of targets to and their removal from the reactor is accomplished with hydraulic systems (“rabbits”) that enable operators to introduce or withdraw irradiated samples at any time.

Reactors vary in size, according to their neutron flux; the fuel elements used (natural or enriched uranium); and type of cooling system (distilled or heavy water). Some of the smaller reactors, such as the Slowpike reactor manufactured by Atomic Energy of Canada Limited (AECL) or the TRIGA reactors from General Atomics, are of the “swimming-pool” type. They use uranium as fuel enriched to 20 % (uranium – zirconium hydride), and they supply fluxes of 10^{12} n cm^{-2} s^{-1} (during 4 h) or 2×10^{11} n cm^{-2} s^{-1} (during 24 h). Such a *low-flux nuclear reactor* yields relatively small amounts of radionuclides (in the order of a few millicuries), usually with low specific activities (under 100 mCi/mmol). These reactors operate at up to 2 MW power. At the other end of the spectrum are the *high-flux reactors*, which operate at about a 100 MW power level, and have thermal neutron fluxes of 2 – 5×10^{15} n cm^{-2} s^{-1}. There are very few such reactors available for radionuclide production. Most of those commonly used are of the *intermediate type*, with fluxes in the 10^{13} – 10^{14} n cm^{-2} s^{-1} range and operating at 10 – 40 MW power. They are water-moderated and cooled, and use 10 – 93 % enriched ^{235}U – Al or UO_2 pellets as fuel. Any listing of such facilities would very rapidly become outdated given the environmental and political concerns associated with nuclear reactors. Current information is readily available from the IAEA (Vienna, Austria) or from the national authority in any country that regulates the use and possession of radioactive materials.

In a typical reactor, natural uranium enriched with ^{235}U or ^{233}U undergoes fission and produces heat, fission products, and neutrons for maintaining the chain reaction. Because there is no coulombic barrier to neutrons, neutrons of very low kinetic energies can interact with target nuclei and cause nuclear reactions. The most frequent reaction is of the (n, γ) type, which yields “neutron-rich” radionuclides. Two other common nuclear reactions are (n, p) and (n, α), which may be induced by both slow (thermal) and fast neutrons. In the first of these reactions (n, γ), the radionuclide produced is usually identical to the chemical element that has been bombarded. Hence, this process is also called “activation”; one of its limitations is that the specific activity of the resulting radionuclide will be low due to “dilution” with the original isotope. In the latter two reactions, however, the radionuclide that is produced is chemically different from the target element, and a chemical separation can (and must) be carried out [8]. Other common processes involve isolating radionuclides produced during the fission of uranium, or following successive neutron capture and Szilard – Chalmers [5] recoil-enrichment reactions (see Nuclear Reactors).

The materials most commonly used as targets for nuclear reactions are densely-packed solids, such as metal wires, foils, and powders, so that the maximum number of neutrons can be absorbed within the exposed volume. Some chemical

compounds of the metals (e.g., oxides, carbonates, and nitrates) may also be used, but such targets are likely to undergo both Szilard – Chalmers and thermal decompositions during irradiation, thus necessitating more complex purification at the end of bombardment. The most common containers for irradiations are aluminum, quartz, or plastic tubing, as well as aluminum foil packets, which are sealed in an aluminum capsule. The heat generated in exothermic nuclear reactions (n, γ and n, fission) limits the mass of the irradiated target to about 1 g, because the amount of heat absorbed during radiation increases with mass and specific gravity. The (n, γ) reaction usually produces only one radionuclide, so the final step in its production is separation from the container and dissolution or extraction in a suitable solvent. Some of the most common radionuclides produced in nuclear reactors are listed in Table 1.

Table 1. Examples of radionuclides produced in a nuclear reactor (this tabulation is based on a more extensive table in [10])

Radionuclide produced	$t_{1/2}$	Mode of decay	Main γ energies emitted, keV (% abundance)[d]	Selected mode of production
^{18}F	110 min	β^+, EC [a]	511 (194)	^{16}O (^3H, n) [b]
^{32}P	14.3 d	β^-		^{31}P (n, γ)
^{24}Na	14.9 h	β^-	1369 (100); 2754 (100)	^{23}Na (n, γ)
^{35}S	87.9 d	β^-		^{35}Cl (n, p)
^{36}Cl	3.1×10^5 a	β^-		^{35}Cl (n, γ)
^{42}K	12.3 d	β^-	1524 (18)	^{41}K (n, γ)
^{47}Ca	4.5 d	β^-	1308 (74)	^{46}Ca (n, γ)
^{51}Cr	27.8 d	EC	3.2 (9)	^{50}Cr (n, γ)
^{58}Co	71.3 d	β^+	511 (30); 810 (99)	^{58}Ni (n, p)
^{59}Fe	45.6 d	β^-	1095 (56); 1292 (44)	^{58}Fe (n, γ)
^{64}Cu	12.8 h	β^- β^+, EC [a]	511 (38)	^{63}Cu (n, γ)
^{69m}Zn	13.8 h	IT	439 (96)	^{68}Zn (n, γ)
^{85}Kr	10.7 a	β^-	514 (4)	^{235}U (n, fission)
^{85}Sr	64 d	EC [a]	514 (100)	^{84}Sr (n, γ)
^{86}Rb	18.7 d	β^-	1078 (9)	^{85}Rb (n, γ)
^{90}Y	64 h	β^-		^{89}Y (n, γ)
^{95}Nb	35 d	β^-	765 (100)	^{94}Zr (n, γ)
^{99}Mo	67 h	β^-	740 (12)	^{235}U (n, fission)
^{113}Sn	115 d	EC [a]	255 (18)	^{112}Sn (n, γ)
^{125}I	60 d	EC [a]	27 (138); 35 (7)	^{124}Xe (n, γ)
^{131}I	8.1 d	β^-	364 (82)	^{235}U (n, fission)
^{131}Cs	9.7 d	EC [a]	Xe X rays	^{130}Ba (n, γ)
^{140}Ba	12.8 d	β^-	537 (34)	^{235}U (n, fission)
^{153}Gd	242 d	EC [a]	Eu X rays	^{152}Gd (n, γ)
^{195m}Pt	4.1 d	IT [c]	67 (84); 99 (12)	^{194}Pt (n, γ)
^{198}Au	2.7 d	β^-	412 (95)	^{197}Au (n, γ)

[a] EC = electron capture.
[b] This is the result of a double reaction: the first reaction is the activation of lithium: ^6Li (n, ^3H)^4He. This reaction generates ^3H, which then reacts with the ^{16}O present.
[c] IT = isomeric transition (transition to a lower nuclear energy level without a change in mass).
[d] 100 % corresponds to emission of one photon per nuclear decomposition.

3.2. Cyclotrons and Other Accelerators

Charged-particle accelerators are high-voltage machines capable of accelerating charged particles to the desired energy levels. For biomedical and chemical purposes, the machines used need only accelerate charged particles to less than 100 MeV, although radionuclide production has also been accomplished in machines that accelerate charged particles to higher energies. However interesting such high energy machines may be, they were built originally for purposes other than radionuclide production, and their availability is thus very limited.

Cyclotrons. The most commonly used system is the cyclotron [9], which is an accelerator whose path has been shaped into a spiral by means of a magnetic field applied perpendicularly to the plane of motion of the ions. It was originally invented by ERNEST O. LAWRENCE in 1930. The particles most commonly accelerated in a cyclotron are protons, deuterons, and ^3He and ^4He (i.e., α) particles. During operation, a charged particle in a uniform magnetic field moves

along a circular orbit with a period of revolution independent of its velocity. The magnetic field causes a bending of the particle's trajectory and guides it along a circular path while an electrostatic potential difference between the electrodes supplies energy to the particle itself, progressively taking it into orbits of greater and greater radius without any change in the frequency of revolution. Thus, particles follow a spiral path from the center of the machine (where the ion source is located) to the maximum radius where they are either used to strike a target directly (internal target) or extracted from the cyclotron for use outside (external target) [11].

The radionuclides formed in charged-particle reactions are all neutron-deficient. Most are short-lived, and many are positron emitters (see ⟶ Isotopes, Natural – Stable Isotopes as Radioisotope Precursors), which makes them of particular interest for biomedical studies. Indeed, it is worth recalling here that the first biological studies with radiocarbon were done with ^{11}C, the positron-emitting isotope of carbon ($t_{1/2}$ = 20.4 min) made at LAWRENCE's cyclotron, before M. D. KAMEN produced the much longer-lived (and reactor-produced) ^{14}C [12].

The most important field of application for particle accelerators is in biomedicine. This includes radionuclide production, cancer treat-ment with fast neutrons and proton beams, and radiological studies. Use of such radionuclides requires a more complex environment than that for handling the radionuclides produced in reactors, inasmuch as it is generally necessary to integrate radionuclide production with its utilization. This is especially true of the “physiological” radionuclides (^{11}C, ^{13}N, and ^{15}O), which must be produced at or extremely near the site of utilization. However, a much greater variety of radionuclides can be prepared with the aid of a cyclotron, and the resulting specific activities are much higher than those of materials from reactors.

Cyclotrons can be grouped into four main categories, classed according to the maximum energies of the accelerated protons: < 10 MeV, 10 – 18 MeV, 18 – 40 MeV, and > 40 MeV. While the smaller (“baby”) cyclotrons are usually single-beam machines, all the others can accelerate both protons and deuterons (the latter at half-proton energy). Acceleration of ^{3}He and ^{4}He particles requires energies > 18 MeV. Higher-energy cyclotrons (> 40 MeV) are required for the production of fast-neutron and proton beams suitable for cancer chemotherapy. In addition, and with sufficient knowledge of the excitation function, higher beam energies are often degraded to the desired level at the entry to the target. The thickness of the target is also crucial, to ensure that energy deposited inside the target falls within the desired range.

Charged-particle accelerators are utilized extensively for the production of several neutron-deficient radionuclides that are applied in the synthesis of radiopharmaceuticals, mainly involving the following four positron emitters, whose $t_{1/2}$ values are given in parentheses: ^{11}C (20.4 min), ^{13}N (10 min), ^{15}O (2 min), and ^{18}F (110 min). One of the main applications of such positron-emitting radionuclides takes advantage of the two 511-keV annihilation photons emitted in coincidence with an angular correlation of 180°, the positron – emission tomography (PET) technique [13]. Cyclotrons used for such purposes are usually two-particle machines, yielding proton energies ≤ 20 MeV and deuteron energies ≤ 10 MeV, with a maximum beam current of 60 μA.

Many other nuclear reactions can be accomplished with cyclotrons, including the (p, α), (p, d), (p, γ), (p, *x*n), (p, *x*p*x*n), (d, α), (d, p), (d, *x*n), (d, *x*p*x*n), (^{3}He, p), (^{3}He, *x*n), (^{3}He, *x*p*x*n), (α, p), (α, *x*n), and (α, *x*p*x*n) reactions, where *x* is an integer, usually between 1 and 6. The nuclear reactions occurring in a cyclotron are varied, and each nuclear reaction induced by a charged particle has an optimum energy range for the bombarding particle to produce the best results in terms of yield and purity of the desired product. The choice of nuclear reaction, knowledge of the excitation functions, proper timing, and the use of appropriate radiochemical methods determines the quality of radionuclide preparation. When radionuclides are produced in accelerators they are generally obtained as *no-carrier-added* materials [14], thereby assuring very high specific activity. (The term “no-carrier-added” expresses the fact that no stable isotope of the radionuclide under production has been introduced during any of the steps in the generation of that radionuclide.)

Choice of a suitable *target* is perhaps the most important single factor in selecting a reaction for the production of a radionuclide — or for that matter, of a radiopharmaceutical — in a cyclotron. Because of intense heating caused in the target and its containment as a result of beam-energy degradation, target materials must have high melting points, high thermal conductivities, and they must be chemically inert. In order to minimize damage to the target, and in order to obtain a uniform production of radionuclides, the target should be of the highest purity attainable, and it must be properly cooled [15]. A major concern with respect to operating a cyclotron efficiently is removing the heat generated on the target during its bombardment, with temperatures that may exceed 1000 °C. This is accomplished by circulating cold deionized water or air around the target, and by using stable, heat-conducting target materials.

Target holders are usually made of aluminum. This metal is easy to work with, leads to minimum build-up of induced radioactivity, and displays good heat transfer. Copper is an alternative; although it has good mechanical and thermal properties, it is subject to increased build-up of radioactivity following target bombardment. The target's windows are usually made of thin foils of aluminum, Mylar (PETP film), Havar (Co-based alloy), tungsten, or tantalum. Windows must be as thin as possible to avoid beam degradation, yet strong enough to resist internal target pressure. Target materials may be solids, liquids, or gases. To facilitate the separation of radionuclides generated during activation, radiochemists have devised gaseous targets that yield solid radionuclides and solid targets for producing gaseous radionuclides. Developed mainly for the production of the PET radionuclides ^{15}O, ^{13}N, ^{11}C, and ^{18}F, gas targets are also used for the preparation of ^{123}I (from ^{124}Xe), ^{81}Rb (from ^{82}Kr), and other species. A discussion of gas target techniques for radionuclide production with small cyclotrons is given in [16]. Sometimes a stack of different metal foils or a double chamber can be used to produce several radionuclides simultaneously. Sometimes it is necessary to recycle a target material after processing, especially if it is expensive, but most of the common targets support continuous production of specific radionuclides, with automatic separation of the products.

One problem sometimes encountered is the fact that the radionuclide formed may react with the walls of the target. Thus, in the original scheme for production of ^{18}F using the ^{20}Ne(d, α) reaction, most of the (hot) ^{18}F generated reacted with the

walls of the gas target. This problem was solved by use of a nickel target, the walls of which had been passivated with fluorine gas, together with addition of a small amount of cold $^{19}F_2$ to maintain the ^{18}F as ^{18}F - F_2 [17].

Even when the target is of high purity, isolation and purification of radionuclides is usually required subsequent to their production. The most common chemical separation methods are precipitation, co-precipitation, solvent extraction, chromatography, electrolysis, and ion exchange. In addition, physical methods are used (i.e., mass separation, vacuum distillation, vacuum deposition, or phase changes). Gas targets are obviously the most desirable, because they allow rapid and efficient recovery of products with high specific activity and good radiochemical and radionuclidic purity [15].

Van de Graaff Accelerators. *Single-stage Van de Graaff* accelerators have been used effectively for the preparation of radionuclides due to their very high beam currents [18]. *Tandem Van de Graaff accelerators* with higher energies display lower beam-current capacity. One of the advantages of Van de Graaff accelerators is that their deuteron and proton energies are equivalent, whereas with a cyclotron the deuteron energy is only half the proton energy. A Tandem Van de Graaff accelerator can supply a PET center with useful quantities of a variety of radionuclides, including ^{11}C, ^{13}N, ^{15}O, and ^{18}F. All of these can be prepared with (p, n) or (d, n) reactions, and the reactions themselves have cross sections of at least 100 millibarn over most of the energy ranges of interest. With such cross sections, the thick-target yields for single half-life bombardment with 1 μA (a beam readily available in nearly all Van de Graaff accelerators) are about 30 mCi for 6 MeV incident energy and 100 mCi for 12 MeV. This is within the order of magnitude required for medical applications (taking into account losses accompanying synthesis) [19]. Historically, cyclotrons have been the accelerators of choice because of their high beam currents. For long-lived radionuclides this is always advantageous, since products not used immediately can be stored for later use. For short-lived radionuclides, however, a Tandem Van de Graaff accelerator is preferable, as storage of the products is clearly impossible, and the Van de Graaff accelerator can generate as much activity as required for immediate use.

Electron Accelerators. An electron accelerator is another type of device that can and has been used to prepare radionuclides via photonuclear reactions. Its advantage over the (positively) charged particle accelerators is that both neutron-deficient and neutron-rich radionuclides can be produced. Also, bremsstrahlung causes less heating of the target than does charged-particle bombardment, simplifying both target preparation and heat removal. Sometimes a higher radionuclidic purity can be achieved through a photonuclear reaction than with a reaction involving a charged particle. For example, the yield of ^{155}Tb in a (γ, n) reaction after irradiation with an electron beam of 25 MeV (beam current 10 μA) for 20 h is about 3.0 μCi μA^{-1} h^{-1} g^{-1}. The purity of the product from such a reaction is usually very high. Some powerful electron linear accelerators achieve E_{max} values of 500 MeV and maximal beam power of 250 kW [20].

Another example of a photonuclear reaction for producing a carrier-free radionuclide of high purity is the preparation of ^{123}I free of any ^{124}I contamination (which is very detrimental to its medical use) via the sequence of reactions:

$$^{124}Xe(\gamma, n)^{123}Xe\ (EC, \beta^+)\ ^{123}I \tag{3.8}$$

The reaction threshold here is 10.5 MeV, and the energy for maximum reaction cross section is 15.5 MeV. An overenergetic beam can generate the undesirable ^{124}I via the ^{126}Xe (γ, n)^{124}I reac-tion with a threshold of 17.3 MeV. Therefore, bremsstrahlung beam energies higher than 30 MeV may be detrimental. However, by reducing the proportion of ^{126}Xe in the enriched ^{124}Xe, it is possible to increase the electron energy, and hence the yield of ^{123}I, while retaining high radionuclidic purity [21]. Other reactions for producing ^{123}I are described in Section Radionuclide Production from Cyclotrons.

Neutron Generator. Another powerful application of neutron beams takes advantage of a “neutron generator,” which can be utilized for the preparation of certain short-lived radionuclides, especially in laboratories without access to a reactor or a cyclotron. A neutron generator is a source of high-energy fast neutrons, and is based on a reaction between deuterium and tritium:

$$^{3}H(d, n)^{4}He \tag{3.9}$$

This reaction is exothermic and produces high-energy neutrons with a kinetic energy of about 14 MeV. A practical generator consists of an ion source, a target, vacuum control, a cooling system, and various power circuits for the production of an accelerated deuteron beam which is caused to impinge on a tritium target. Two types of such generators are available: the first operates with a demountable target, making it possible to change targets, while the second operates with a sealed tube. The former is used mainly in research, where the ability to change targets and study other nuclear reactions is advantageous. The latter is used where only a fixed flux of neutrons is required. An example of sealed tube device is the Kanan generator with an output of 10^{11} neutrons per second. The maximum neutron flux available for irradiating small samples is 2×10^{10} n cm^{-2} s^{-1}. Because of the high energy of the resulting neutrons (14 MeV), it is possible with this device to conduct (n, 2n), (n, p), and (n, α) reactions. Due to the fact that the cross section for 14 MeV reactions is relatively low, and the fluxes available are several orders of magnitude lower than with nuclear reactors and cyclotrons, large quantities of material must be activated in order to prepare useful quantities of radionuclides. For this reason the main applications of this system are in activation analysis and neutron-beam therapy [22].

Worldwide Availability of Accelerators. Many small (< 10 MeV protons), medium (10 –40 MeV), and larger (40 – 70 MeV) cyclotrons are now available in both industry and in a number of developing nations. In addition, many 70 MeV proton accelerators are now available in a number of countries where they are used for commercial and research purposes, yielding beam intensities ≤ 200 μA. Proton beams have a greater range in matter than either deuterons or α particles, so their efficiency in the production of radionuclides is relatively high. One of the reasons for promoting cyclotrons in developing countries is to provide alternatives to reactor-produced technologies. Overall yields are usually higher from a reactor than with a simple accelerator, but the cost of installing, operating, and ultimately decommissioning a

reactor and its complex ancillary facilities is presumed to be considerably higher than that of installing and operating a modern accelerator and radiochemistry facility [23].

Descriptions of some of the larger accelerators follow, demonstrating variability in terms of beam strength and production capability.

Los Alamos Meson Physics Facility (LAMPF) has a powerful (800 MeV, 300 – 550 μA) proton beam. This beam is accelerated through a straight tunnel about 1 km long on a collision course with nine isotope targets. The beam accelerates to 252 000 km/s, 84 % of the velocity of light, and at the end of the line crashes into stacked targets [24].

Linear Accelerator, Brookhaven National Laboratory (LINAC) is another high-energy, high-current linear accelerator with a 200 MeV, 40 – 70 μA proton beam, a part of which is deflected to targets in the Brookhaven LINAC Isotope Production (BLIP). The main portion of the beam, however, goes to the alternating gradient synchrotron, where the particles are accelerated in a 0.5 mile (0.804 km) orbit for high-energy physics experiments [24].

Tri-University Meson Facility of the Atomic Energy Commission of Canada (TRIUMF –AECL) complex is housed at the University of British Columbia. It contains a high-energy, negative-ion separated-sector cyclotron, which delivers proton beams ≤ 500 MeV and beam currents of 140 μA.

Another powerful medical cyclotron is the sector-focusing frequency-modulated cyclotron of the *Gustaf Werner Institute* in Uppsala, Sweden, which permits acceleration of protons up to 200 MeV and other light ions to corresponding energies, as well as spallation neutrons. Its max-imum external proton current in the high-energy range is about 10 μA, and ≤ 40 μA as an external beam for the lower-energy range [25].

One of the most powerful European production facilities is the *Swiss Institute of Nuclear Research* (SIN) in Villigen. This facility attempts to utilize a > 200 MeV proton beam to induce spallation reactions as a means of high-level radionuclide production. Such spallation reactions usually do not yield radionuclides with the quality standards already established for accelerator-produced radionuclides [15].

Other important large accelerators are the *Hammersmith cyclotron* in London, the *Variable Energy Cyclotron* (VEC) of Harwell in Didcot, United Kingdom, the 86-inch cyclotron at the *Oak Ridge National Laboratory* (ORNL), the 600 MeV proton cyclotron of the *CERN synchrocyclotron,* and the 400 MeV proton beam of the *Fermi National Lab* in Batavia, Illinois. More detailed lists of available cyclotrons are provided in [26] and [27].

3.3. Radionuclide Production from Reactors

Methods for radionuclide production using reactors have been described extensively [28], [29]. Basically, a suitable target is placed inside the reactor for a preset period of time. The irradiation time can vary from 2 – 4 half-lifes for radionuclides with half-lifes of less than 2 weeks to weeks or months for longer-lived radionuclides. The target is removed from the reactor at the EOB and then processed. Typical irradiation times range from days for most radionuclides (e.g., 5 – 7 d to activate ^{98}Mo to ^{99}Mo at a flux of 2×10^{14} n cm^{-2} s^{-1}) to weeks (two weeks to activate ^{194}Pt to ^{195m}Pt at a similar flux, 30 – 60 d to activate ^{124}Xe to ^{125}Xe, followed by another 60 d to allow the ^{125}Xe to decay to ^{125}I), months, or years (e.g., to produce ^{3}H, $t_{1/2}$ = 12.5 a), and ^{14}C ($t_{1/2}$ = 5640 a). Given so much variability, a common way of expressing practical yields is to define the activity that can be achieved at half-saturation. For example, the activation of ^{152}Sm to form ^{153}Sm has a cross section of 200 barn. The isotopic abundance of ^{152}Sm is only 26.6 %, and activation of natural samarium would lead to a mixture of radioisotopes. Approximately 370 GBq of ^{153}Sm can be produced by the irradiation of 10 mg of Sm_2O_3 (made from > 99 % enriched ^{152}Sm) at a flux of 10^{14} n cm^{-2} s^{-1} for 46 h (half-saturation). Production of ^{153}Sm with higher specific activities is possible only in a high-flux reactor (flux $\geq 2\times10^{15}$ n cm^{-2} s^{-1}), which should be sufficient for use in clinical protocols requiring administration of minute amounts (< 10 mg) of labeled proteins for therapeutic applications [29]. In another example, irradiation of 1 g of natural palladium for 13 h in a flux of 10^{14} n cm^{-2} s^{-1} resulted in 20 mCi of ^{109}Pd that still required purification. A much purer radionuclide was obtained by the irradiation of enriched ^{108}Pd, leading to a ^{109}Pd with specific activity of about 2500 Ci/mol [29].

Thus, the *direct* (n, γ) *activation* of enriched target molecules with σ > 100 barn in reactors with high thermal neutron fluxes should provide specific activities sufficient for many therapeutic radionuclide applications. It is clear from the above examples that abundant quantities of radionuclides can be prepared in reactors with relatively modest fluxes if enriched targets (σ > 10 barn) are used. Examples of β-emitting radionuclides with $t_{1/2}$ values of < 15 d produced in nuclear reactors include ^{165}Dy, ^{109}Pd, ^{188}Re, ^{166}Ho, ^{153}Sm, ^{90}Y, ^{186}Re, ^{177}Lu, ^{169}Er, and ^{32}P (see also Table 1) [10].

Alternatively, it is possible to achieve high activities (both total and specific) of *no-carrier-added reactor-produced radionuclides* through indirect (n, γ) sets of reactions, e.g.:

$$^{124}\text{Xe (n, }\gamma)^{125}\text{Xe} \longrightarrow {}^{125}\text{I}$$
$$^{104}\text{Ru (n, }\gamma)^{105}\text{Ru} \longrightarrow {}^{105}\text{Rh}$$
$$^{130}\text{Te (n, }\gamma)^{131}\text{Te} \longrightarrow {}^{131}\text{I}$$
$$^{198}\text{Pt (n, }\gamma)^{199}\text{Pt} \longrightarrow {}^{199}\text{Au}$$

One advantage of this approach is that the reaction yields need not be high, since high activities can be achieved by

irradiating large quantities of target nuclei, followed by chemical purification of the desired radionuclide [7]. In the irradiation of ^{130}Te, the intermediate ^{131}Te ($t_{1/2}$ = 25 min) decays to ^{131}I ($t_{1/2}$ = 8 d). Because this radionuclide is nonisotopic with the target, it is, in theory, "carrier-free" (or — more accurately — "no-carrier-added"). Specific activities > 10^6 Ci/mol can be obtained by this reaction [8]. Other β-emitting radionuclides produced by indirect activation are ^{77}As, ^{149}Pm, ^{199}Au, ^{177}Lu, and ^{111}Ag (see also Table 1).

Fission. The energy liberated in a fission process is much higher than that liberated in β decay: part of it goes to the newly formed fragments, and the rest goes to the accompanying free neutrons. Since heavy nuclei have a large proportion of neutrons, fission fragments tend to contain more neutrons than necessary for their stability, and the nuclei stabilize themselves by β decay. When ^{235}U absorbs a thermal neutron, it fragments into a large number of nuclei, distributed around two peaks of mass 100 and mass 130. Several of the radionuclides formed are of significant medical, industrial, and radiochemical interest. Some of the key fission products from ^{235}U and their cumulative yields are:

^{90}Sr as a "pure beta" source	5.89 %
^{137}Cs as a γ source	6.18 %
^{99}Mo as a generator parent	6.14 %
^{131}I for labeling and diagnosis	2.84 %

While long-lived fission-product nuclides such as ^{137}Cs and ^{90}Sr can be conveniently recovered from fuel-reprocessing wastes, medium- and short-lived radionuclides are usually produced by irradiation of a uranium (natural or enriched) target in the reactor. Since the fission product gives rise to a mixture of radionuclides, some of which are highly toxic (e.g., ^{90}Sr and ^{137}Cs), stringent purification procedures must be employed in order to recover radionuclides of interest with the requisite purity [30]. Fission reactions may also be regarded as a type of neutron activation, because they are generally induced by neutron capture.

Spallation is another process yielding large numbers of neutrons. It is a two-step process that occurs as a consequence of the impact of high energetic nucleons onto a nucleus. In the first stage of events, called the "intra-nuclear cascade", highly energetic nucleons escape from the nucleus and trigger a similar cascade in another nucleus. The nucleus left behind by the cascade particles is in a highly-excited state and loses most of its energy by evaporating off other particles: mainly neutrons, but also to a rapidly decreasing extent protons, deuterons, tritons, and α particles. The result of this process is generally an active nucleus that decays further through γ emission and β^+ decay. The spallation reaction is basically endothermic [31], and it can only occur when the incident particles have energies > 100 MeV. Production of radionuclides at very large accelerators (LAMPF, BLIP) occurs by spallation.

Successive Neutron Capture. Successive capture of two or more neutrons is of interest for the production of certain radionuclides, such as ^{38}S, which can be used as a parent for ^{38}Cl in a nuclide generator.

$$^{36}S(n,\gamma)^{37}S(n,\gamma)^{38}S \rightarrow {}^{38}Cl$$

$$^{194}Pt(n,\gamma)^{195m}Pt(n,\gamma)^{196}Pt$$

In some instances, however, successive neutron-capture reactions are a source of radionuclide impurities (e.g., formation of ^{126}I from ^{125}I and ^{199}Au from ^{198}Au [7], or the burnup of ^{195m}Pt ($t_{1/2}$ = 4 d) to ^{196}Pt, where the σ for production of ^{195m}Pt is 0.1 barn, but that of its burn up is 20 000 barn!) [32].

Szilard – Chalmers Recoil-Enrichment Processes. When stable compounds of certain elements are irradiated with thermal neutrons in an (n, γ) reaction, a significant fraction of the radionuclide produced appears in a chemical form different from that of the target because chemical bonds are cleaved due to the high energy of the radionuclide formed. This fraction is known as the "recoil chemical form," and it can be readily separated from the target by simple radiochemical techniques. The fraction of the total activity of the radionuclide produced that appears in the recoiled chemical form is known as the Szilard – Chalmers recoil enrichment recovery. For example, free ^{82}Br can be obtained from a $KBrO_3$ target with a 91 % yield and 2×10^4 enrichment [7].

Special Radionuclides Produced in Nuclear Reactors. The three most widely used radionuclides in nuclear medicine are ^{99m}Tc, obtained from the β decay of ^{99}Mo and used in 85 % of all in vivo nuclear medicine studies; and ^{125}I and ^{131}I, where ^{125}I accounts for 95 % of all in vitro nuclear medicine procedures, while ^{131}I remains the second most widely used radionuclide for in vivo studies. All three radionuclides are produced, either directly or indirectly, in nuclear reactors.

Production of Molybdenum-99. Two major methods exist for the production of ^{99}Mo at high specific activity: direct irradiation of molybdenum trioxide and fission of uranium, either enriched or natural.

When MoO_3 is irradiated at fluxes exceeding 2×10^{14} n cm^{-2} s^{-1} for 5 – 7 d and then extracted in aqueous ammonia, specific activities of > 5 Ci/g $^{99}MoO_3$ of > 99.5 % radionuclidic purity are obtainable. To ensure even higher specific activities, enriched $^{98}MoO_3$ can be used.

In the *uranium fission process*, natural uranium or enriched ^{235}U can be employed in the form of metal, U – Al alloy, or uranium oxide. If 90 %-enriched ^{235}U is irradiated at a neutron flux of 2×10^{13} n cm^{-2} s^{-1} for 3 – 7 d, the yield of ^{99}Mo is about 15 Ci/g. The ^{99}Mo can be extracted in HNO_3 and passed through an alumina column. ^{99}Mo is adsorbed onto the column, which is than washed to remove radionuclidic contamination, resulting in ^{99}Mo obtained of specific activity > 50 Ci/mg and > 99.9 % radionuclidic purity [7].

Production of Iodine-125. For the production of ^{125}I, natural xenon is used containing about 0.1 % ^{124}Xe in the target. It is filled under pressure into an capsule and irradiated at a neutron flux of 2×10^{14} n cm^{-2} s^{-1} for 30 – 60 d. The yield of ^{125}I is approximately 0.2 Ci/g. This material should be allowed to stand for 60 d before processing to reduce ^{126}I levels to < 2 %. Then the xenon gas is sublimed out of the cooled ampule, and ^{125}I is extracted with dilute NaOH. The final specific activity is > 15 Ci/mg and product purity is > 99.9 %.

Production of Iodine-131. For ^{131}I production a TeO_2 target is irradiated, from which ^{131}I is obtained by wet or dry distillation [7].

3.4. Radionuclide Production from Cyclotrons

Radionuclides of major biomedical interest that are produced in cyclotrons include the physiological radionuclides (^{11}C, ^{13}N, ^{15}O, and ^{18}F) as well as a number of other key neutron-deficient nuclides. Table 1 summarizes some of the key cyclotron-produced radionuclides, based on [10]. An analysis of the relative merits of various nuclear processes using a cyclotron to produce radionuclides for generators is presented in [33].

Table 2. Examples of radionuclides produced in cyclotrons (this tabulation is based on a more extensive table in [10]) [a]

Radionuclide produced	$t_{1/2}$	Mode of decay	Main γ energies emitted, keV (% abundance)	Selected mode of production
^{11}C	20.3 min	β+	511 (200)	^{11}B (p, n)
^{13}N	9.9 min	β+	511 (200)	^{12}C (d, n)
^{15}O	123 s	β+	511 (200)	^{14}N (d, n)
^{18}F	110 min	β+	511 (200)	^{20}Ne (d, α)
^{22}Na	2.62 a	β+, EC	511 (180); 1275 (100)	^{24}Mg (d, α)
^{28}Mg	21.2 h	β−	31 (90); 400 (30); 950 (30)	^{26}Mg (α, 2p)
^{30}P	2.5 min	β+	511 (200); 2230 (0.5)	^{27}Al (α, n)
^{38}K	7.7 min	β+	511 (200); 2170 (100)	^{35}Cl (α, n)
^{48}Cr	23 h	EC	116 (98); 310 (99)	^{50}Cr (d, 5n)
^{51}Mn	45 min	β+, EC	511 (194); 1560 (34)	^{52}Cr (p, 2n)
^{52}Fe	8.2 h	β+, EC	165 (100); 511 (112)	^{50}Cr (α, 2n)
^{58}Co	71.3 d	β+, EC	511 (30); 810 (99)	^{58}Ni (n, p)
^{62}Zn	9.1 h	β+, EC	511 (186); 669 (81)	^{60}Ni (α, 2n)
^{64}Cu	12.8 h	β+, EC, β−	511 (38)	^{65}Cu (p, pn)
^{74}As	18 d	β+, EC, β−	511 (59); 596 (61); 635 (14)	^{71}Ga (α, n)
^{75}Br	1.7 h	β+, EC	285 (92); 511 (180)	^{75}As (α, 4n)
^{77}Br	57 h	β+, EC	240 (30); 511 (1.5)	^{75}As (α, 2n)
^{81}Rb	4.7 h	β+, EC	511 (26); 1110 (42)	^{82}Kr (p, 2n)
^{85m}Kr	4.2 h	β−, IT	150 (75); 305 (13)	^{84}Kr (d, p)
^{85}Sr	64 d	EC	514 (100)	^{86}Rb (p, n)
^{97}Ru	2.9 d	EC	215 (91); 324 (8)	^{95}Mo (α, 2n)
^{99}Mo	67 h	β−	740 (12)	^{98}Mg (d, n)
^{111}In	2.8 d	EC	173 (89); 247 (94)	^{111}Cd (p, n)
^{123}I	13.3 h	EC	159 (83)	^{127}I (p, 5n)
^{127}Xe	36.4 d	EC	172 (22); 203 (65); 375 (20)	^{133}Cs (p, 2p,5n)
^{201}Tl	74 h	EC	Hg X-rays; 81 (98); 168 (8)	^{200}Hg (d, n)

[a] For abbreviations see footnote to Table 3.

Production of Oxygen-15 and Fluorine-18. ^{15}O and ^{18}F can be prepared by a variety of methods. One involves their

simultaneous production using 30 MeV protons, in which a double liquid-chamber target is irradiated [34]. Such a system provides the option of producing $H_2^{15}O$ and either ^{13}N or ^{18}F with a single proton beam. Irradiating natural water in the front chamber leads to $H_2^{15}O$ by the ^{16}O (p, pn) reaction, while the rear silver-liquid chamber efficiently generates either ^{13}N by the ^{16}O (p, α) reaction or ^{18}F by the ^{18}O (p, n) reaction. This combination target ensures rapid and high-volume production of the three most fully-validated and widely-used radiopharmaceuticals in clinical PET. Other useful reactions for ^{15}O production include ^{14}N(d, n), ^{16}O (p, d), and ^{15}N (p, n), where the former is most widely practiced. The radiopharmaceuticals produced on-line by these reactions are $H_2^{15}O$, $C^{15}O_2$, and $C^{15}O$, respectively [35]. Two additional methods are applied for no-carrier-added ^{18}F production: proton irradiation of a natural magnesium target or of enriched ^{27}Al. ^{27}Al produces ^{18}Ne via Al (p, 2 α2n)^{18}Ne, the latter decays ($t_{1/2}$ = 1.7 s) to ^{18}F and can be used in a generator system [36], [37]. A very useful method for preparing electrophilic ^{18}F is the (d, α) irradiation of natural neon in a gas target [17].

Production of Carbon-11. ^{11}C can be made using the ^{12}C (^3He, α), ^{10}B (d, n), and ^{14}N (p, α) reactions [38]. In the latter case a gas target is irradiated with 18 MeV protons and then combined with traces of O_2 to yield $^{11}CO_2$. This target can also be used as a starting material for $^{11}CH_3I$, a suitable precursor for the methylation of organic molecules [39]. ^{11}C can also be profrom a linear accelerator, with CO_2, liquid cyclohexane, or liquid glacial acetic acid as target materials. The specific activities achieved in such a production are reported to be suitable for clinical duced by using a 26 MeV, 100 μA electron beam applications [40].

Production of Molybdenum-99. ^{99}Mo was produced by a cyclotron long before it was used as a diagnostic radiopharmaceutical. The direct production of curie quantities of ^{99m}Tc and 100 mCi quantities of ^{99}Mo on a cyclotron were first suggested by BEAVER and HUPF [41], using an isotopically-enriched ^{100}Mo target. RANDA and SVOBODA [42] used the ^{98}Mo (d, p) ^{99}Mo reaction for the same purpose. The use of natural molybdenum targets was later studied by ALMEIDA and HELUS [43] applying a maximum proton energy of 25 MeV. Thin foils of metallic molybdenum (purity > 99.95 %) were stacked and used as targets in a water-cooled aluminum holder. Despite the relatively low abundance of ^{100}Mo in the target (9.6 %), the use of natural molybdenum is appropriate for production of no-carrier-added ^{99}Mo by one of the following cyclotron-reactions, in addition to the neutron activation reactions discussed above:

$$^{100}\mathrm{Mo(p,2p)}^{99}\mathrm{Nb(15s)} \rightarrow {}^{99}\mathrm{Mo}$$

$$^{100}\mathrm{Mo(p,pn)}^{99}\mathrm{Mo}$$

$$^{100}\mathrm{Mo(n,2n)}^{99}\mathrm{Mo}$$

with such natural molybdenum targets the production yields are about 20 mCi μA^{-1} h^{-1} [23].

Production of Iodine-123. ^{123}I of high purity can be obtained either directly or indirectly from a cyclotron. In the *direct approach* ^{123}I is produced by 25-MeV-proton irradiation of enriched (> 95 %) ^{124}TeO, but contamination with ^{124}I (up to 5 %) is unavoidable. When used in medicine the 386 – 665 keV γ radiation of ^{124}I causes a degradation of scintigraphic images and contribute to the radiation burden of the patient [43]. In the *indirect method,* 60-MeV protons irradiate elemental iodine, NaI, or KI to produce ^{123}Xe, or else they are used to bombard a ^{124}Xe target:

$$^{127}\mathrm{I(p,5n)}^{123}\mathrm{Xe} \rightarrow {}^{123}\mathrm{I}$$

$$^{127}\mathrm{I(d,6n)}^{123}\mathrm{Xe} \rightarrow {}^{123}\mathrm{I}$$

$$^{124}\mathrm{Xe(p,2n)}^{123}\mathrm{Cs} \rightarrow {}^{123}\mathrm{Xe} \rightarrow {}^{127}\mathrm{I}$$

$$^{124}\mathrm{Xe(p,pn)}^{123}\mathrm{Xe} \rightarrow {}^{123}\mathrm{I}$$

The advantages of using ^{127}I as a target are its 100 % isotopic abundance and low cost, but a significant disadvantage is that it requires a high-energy cyclotron. The advantage of using a ^{124}Xe target is that it utilizes protons of < 42 MeV, but a disadvantage is the high cost of ^{124}Xe gas [44]. In another unusual procedure, ^{123}I is produced by a *beam of 25 MeV electrons* in a 10-h irradiation of a tantalum target enriched with ^{124}Xe. About 200 mCi of ^{123}I accumulate through a photofission reaction, and this material can be used for medical purposes after appropriate extraction [45]. A powerful linear accelerator is capable of producing high-purity ^{123}I by the *spallation process.* One such device is the Canadian TRIUMF, which can generate 1-Ci quantities. The spallation yield of this reaction can compete with yields from low-energy cyclotrons by the use of very thick spallation targets [46].

Production of Thallium-201. ^{201}Tl is one of the most useful neutron-deficient radionuclides produced by a cyclotron. It is used extensively as a tracer in cardiac perfusion studies, although it is a relatively expensive radiopharmaceutical. The most common production method involves the ^{203}Tl (p, 3n)$^{201}Pb \longrightarrow {}^{201}Tl$ reaction [3.44]. The material is produced by bombarding enriched (95 %) ^{203}Tl in the form of the sulfate, yielding 2.2. mCi μA^{-1} h^{-1} with a beam energy of 22 MeV. Other production reactions entail proton bombardment of enriched (> 99 %) ^{205}Tl for the ^{205}Tl (p, 5n)$^{201}Pb \longrightarrow {}^{201}Tl$ reaction or enriched (95 %) ^{206}Pb. Yields and impurities in the three reactions are very similar (see also Section Radionuclide Generators).

Production of Selenium-73. ^{73}Se is increasingly replacing ^{75}Se ($t_{1/2}$ = 120 d), which has been the most widely-used radionuclide of selenium for radiotracer and clinical studies. The long physical $t_{1/2}$ and decay characteristics of ^{75}Se provide a highly absorbed radiation dose to the patient and lead to images of poor quality. ^{73}Se is a more suitable tracer from the standpoint of radiation dosimetry, with a half-life of 7.1 h and applicability to both PET and single photon emission computer tomography (SPECT) studies. It emits both β^+- and a 361-keV γ radiation of 97 % abundance. The reactions ^{70}Ge(^{4}He, *x*n) and ^{75}As (p, 3n) have been used for its production: starting with the 26-MeV α particles available from compact cyclotrons, the reported production yield is 1 mCi μA^{-1} h^{-1}, and with a 55-MeV proton beam the reaction yields 20 mCi μA^{-1} h^{-1}. The radiochemical yield of ^{73}Se is > 75 % in less than 90 minutes [48].

Production of Cobalt-55. ^{55}Co-bleomycin has been used for scanning malignant tumors (e.g., lung cancer and brain metastases), and is considered superior to ^{67}Ga citrate for that purpose. ^{55}Co ($t_{1/2}$ = 18 h) is practical for PET studies, because it decays mainly (81 %) by positron emission. A relatively simple method for preparing it, is by applying a 12 MeV indirect deuteron energy for a ^{54}Fe(d, n) reaction, yielding 1.7 mCi μA^{-1} h^{-1} of ^{55}Co. Other applicable reactions are ^{56}Fe (p, 2n) (natural iron target with 40 MeV incident proton energies) and ^{58}Ni(p, α) (natural nickel foil with 20 MeV protons), followed by separation of the ^{55}Co on an anion-exchange column [49].

Production of Radioactive Gallium Isotopes. ^{66}Ga has replaced other radiogalliums in clinical studies. Accumulation of ^{67}Ga citrate in soft-tumor tissues has promoted its use for the detection of tumors and inflammatory diseases. In addition, ^{68}Ga, a positron emitter with a $t_{1/2}$ of 68 min, has been used in conjunction with tomography after chelation to a number of pharmaceuticals. However, the short half-life of ^{68}Ga limits its suitability in lengthy procedures, so that longer-lived PET gallium radionuclides have been sought. ^{66}Ga has a $t_{1/2}$ of 9.4 h, and 56.5 % of its decay is by positron emission. It can be produced by α particle bombardment of natural copper via the ^{63}Cu(α, n) reaction, with a maximum cross section of 667 mbarn at 16.6 MeV. The extent of long-lived ^{67}Ga contamination at EOB is 0.25 %. If the incident energy is increased beyond 18 MeV, the contamination by ^{67}Ga increases with no significant gain in ^{66}Ga yield. The yield of ^{66}Ga at 18 MeV incident energy is about 700 μCi μA^{-1} h^{-1} [50].

Production of Bromine-77. ^{77}Br is produced by proton irradiation of molybdenum metal. It is separated from the target by co-distillation of the radiobromine species with 0.5 – 1.0 mg of added stable chloride carrier. The solids are trapped in aqueous silver nitrate, the silver is subsequently removed with a cation-exchange column. In a typical ^{77}Br preparation, 100 g of molybdenum metal foil is irradiated for a period of 4 – 7 d, with a chemical yield of 80 – 90 %; the amount of ^{77}Br removed is 1 – 2 Ci [26].

Production of Copper-67. ^{67}Cu is a pure β emitter, the absorbed dose of its radiation in a tumor of 30 – 40 cell diameters is calculated to be uniform. It is therefore used for radiotherapy, especially when chelated to tumor-associated monoclonal antibodies. The suggested nuclear reactions for its production, in quantities relevant to nuclear medicine, include ^{68}Cu (p, 2p), ^{68}Zn (d, 2pn),^{67}Zn (d, 2p),^{64}Ni (α, p),^{67}Zn (n, p), and As (p, spallation), utilizing a natural zinc target. After irradiation with 200 MeV protons, ^{67}Cu is purified via electrodeposition, ion-exchange, dry distillation, and extraction. The following reactions can also be used: ^{68}Zn (p, 2p), ^{70}Zn (p, α), and ^{67}Zn (n, p). These methods provide sufficient quantities of ^{67}Cu with high specific activity and adequate radioactive purity. The production yield of ^{67}Cu at 200 MeV proton energy is 4.2 μCi μA^{-1} h^{-1} [51]. In a 3-d irradiation with 200 MeV protons, 100 mCi can be produced from a 4-g zinc target. The ^{67}Cu can be then separated by electrolysis, which results in a separation factor of 1×10^6. Up to 10 Ci can be produced with irradiation by 800-MeV protons for 2 weeks [7].

Production of Rubidium-81. ^{81}Rb is the parent radionuclide for ^{81m}Kr, used in cardiac studies. Its production represents an example in which the yield often exceeds local requirements even with nonoptimized targetry and irradiation conditions. The method generally employed involves proton or deuteron irradiation of krypton gas of natural isotope composition to yield ^{82}Kr (p, 2n)^{81}Rb, ^{83}Kr (p, 3n)^{81}Rb, ^{84}Kr (p, 4n)^{81}Rb, and ^{82}Kr (d, 3n)^{81}Rb. Other possibilities are ^{85}Rb (p, 5n)^{81}Sr $\longrightarrow$ ^{81}Rb, ^{79}Br (α, 2n), and ^{81}Br (^{3}He, 3n). All methods produce ^{81}Rb in overwhelming quantities. It has also been shown that nickel and aluminum targets are satisfactory for ^{81}Rb production: the recovery yield of ^{81}Rb on an aluminum target is about 3.5 mCi μA^{-1} h^{-1}, and that on a nickel target is about 6.0 mCi μA^{-1} h^{-1}, both with proton irradiation at 20 – 40 μA · h. Nickel seems to be superior to aluminum for the cyclotron targetry utilized in ^{81}Rb production, giving consistent and satisfactory yields [52].

3.5. Radionuclide Generators

Generator systems provide a common and convenient method for obtaining a variety of γ- and β-emitting radionuclides at a patient's bedside. From the practical perspective they offer a desirable alternative to cyclotrons in the case of short-lived radionuclides. In a generator system, a short-lived radionuclide (a “daughter”) is obtained from a longer-lived “parent”, which is adsorbed onto a solid-phase carrier, usually an alumina column, inside an appropriate shield [53], [54]. Efficient separation of the “daughter” from the “parent” is a basic requirement for such a system. As soon as the daughter radionuclide is formed it can be eluted from the column, leaving the parent behind to generate more of the desired radionuclide. Nuclear reactors are the major source for generator “parent” compounds, due to their high neutron flux and relative ease of target preparation. Since most of the labeling procedures involving low protein concentrations are performed with a single conjugated ligand, it is imperative that a daughter radionuclide eluted from the generator has a high specific activity and is free of contamination from other trace metals that may compete with the major radionuclide on a limiting number of chelating agents [29].

One of the prerequisites for a generator system is that there be no (or minimal) breakthrough of the parent radionuclide or the column chemical into the eluent. Breakthrough is defined as the extent of contamination of the eluate by the parent radionuclide, and depends on the distribution coefficient of the parent, on the support material, and on the chemical stability of the support column under high radiation doses [55]. The maximum permissible parent concentration in the eluate depends on its toxicity, half-life, and dose as delivered to the patient in a medical application. It is usually less than 10^{-3} % per elution [56].

Separation of a daughter radionuclide from its parent can be accomplished by a continuous (steady-state) or discontinuous (bolus) method. The *continuous method* involves elution of the daughter radionuclide from the generator as it is formed, and its direct use. The *bolus method* involves one-time elution of all the available daughter activity. Such discontinuous elution can be repeated after an appropriate intervening build-up time. This method is employed when the daughter is *not* a short-lived radionuclide, as in the case of ^{99m}Tc. The theoretical maximum daughter activity continuously available from a generator is a function of the half-life of the daughter radionuclide: for ^{81m}Kr ($t_{1/2}$ = 13 s) the value is 533 μCi/s, while for ^{99m}Tc ($t_{1/2}$ = 21 600 s) it is only 0.32 μCi/s [55]. Several of the more important generator systems are described below, and others are characterized in Table 3.

Table 3. List of selected generator-produced radionuclides (this tabulation is based on a more extensive table in [10]) [a]

Parent radionuclide	$t_{1/2}$	Decay	Mode of production	Daughter radionuclide	$t_{1/2}$	Decay	Major photon energies, keV, (% abundance)
^{28}Mg	21.2 h	β−	^{26}Mg (^{3}H, p)	^{28}Al	2.3 min	β−	1780 (100)
^{38}S	2.9 h	β−	^{37}Cl (α, 3p)	^{38}Cl	37.3 min	β−	2170 (47)
^{44}Ti	46 a	EC	^{45}Sc (p, 2n)	^{44}Sc	3.9 h	β+,EC	511 (188); 1159 (100)
^{52}Fe	8.2 h	β+, EC	^{50}Cr (α, 2n)	^{52m}Mg	21.1 min	β+, IT	511 (196); 1430 (100)
^{62}Zn	9.1 h	β+, EC	^{63}Cu (p, 2n)	^{62}Cu	9.8 min	β+	511 (194)
^{72}Se	8.4 d	EC	^{75}As (d, 5n)	^{72}As	1.1 d	β+, EC	511 (150); 835 (78)
^{81}Rb	4.7 h	β+, EC	^{79}Br (α, n)	^{81m}Kr	13 s	IT	190 (80)
^{83}Rb	83 d	EC	^{81}Br (α, 2n)	^{83m}Kr	1.86 h	IT	388 (80)
^{87}Y	3.3 d	β+, EC	^{81}Sr (d, n)	^{87m}Sr	2.83 h	IT	1780 (100)
^{89}Zr	3.3 d	β+, EC	^{89}Y (p, n)	^{89m}Y	16.1 s	IT	910 (99)
^{90}Mo	5.7 h	β+, EC	^{93}Nb (p, 4n)	^{90m}Nb	24 s	IT	122 (71)
^{98}Mo	67 h	β−	^{98}Mo (n, γ)	^{99m}Tc	6 h	β−	141 (99)
^{103}Pd	17 d	EC	^{103}Rh (p, n)	^{103m}Rh	2.5 s	IT	208 (43)
^{109}Cd	1.24 a	EC	^{108}Cd (n, γ)	^{109}Ag	39.2 s	IT	88 (5)
^{113}Sn	115 d	EC	^{112}Sn (n, γ)	^{113m}In	100 min	IT	393 (64)
^{115}Cd	53.5 h	β−	^{115}Cd (n, γ)	^{115m}In	4.5 h	β−	335 (50)
^{118}Te	6 d	EC	^{121}Sb (n, 5n)	^{118m}Sb	3.5 min	β+, EC	511 (150)
^{122}Xe	20.1 h	EC	^{127}I (p, 6n)	^{122}I	3.5 min	β+, EC	511 (130)
^{128}Ba	2.4 d	EC	^{133}Cs (p, 6n)	^{128}Cs	3.8 min	β+, EC	511 (102); 441 (27)
^{132}Te	20.1 h	EC	^{127}I (p, 6n)	^{132}I	2.3 h	β−, EC	670 (144); 1398 (14)
^{137}Cs	30 a	β−	^{235}U (fiss)	^{137m}Ba	2.6 min	IT	662 (89)
^{134}Ce	2.0 d	EC	^{139}La (p, 6n)	^{134}La	6.8 min	β+, EC	511 (124); 610 (61)
^{183}Ta	5.0 d	β−	^{181}Ta (2n, γ)	^{183m}W	5.3 s	IT	108 (19)

[a] For abbreviations see footnote to Table 3.

The $^{99}Mo \longrightarrow {}^{99m}Tc \longrightarrow {}^{99}Tc \longrightarrow {}^{99}Ru$ Generator. ^{99}Mo is produced commercially in a reactor by a ^{98}Mo(n, γ) reaction, by fission of ^{235}U(n, fission), or by neutron-activation of natural molybdenum (see Fig. 19 and The ^{99}Mo ⟶ ^{99m}Tc ⟶ ^{99}Tc ⟶ ^{99}Ru Generator.). Practical considerations usually limit the irradiation time of the parent to three daughter half-lifes (i.e., ca. 7 d). Under these conditions about 53 % of the ^{99}Mo and more than 92 % of the ^{99m}Tc decays before the irradiation is complete. At the end of bombardment, the ^{99}Tc: ^{99m}Tc ratio in the irradiated target is about 14: 1. Most of the ^{99}Tc is washed out during generator manufacture. Because of the branching decay of ^{99}Mo and the decay of ^{99m}Tc by isomeric transition, ^{99m}Tc solutions always contain some ^{99}Tc. The amount varies considerably depending on the production method, general performance, and the time between production and use. The sensitivity of particular kits to ^{99}Tc depends on the ligand, the amount of usable Sn^{2+}, and the ratio of technetium to ligand. If elution or extraction efficiencies are high, any ^{99}Tc is reduced within about two elutions to an insignificant level that is maintained in subsequent elutions. With efficiencies < 40 %, however, the ratio of ^{99}Tc to ^{99m}Tc increases with each elution. Although the activity of ^{99}Tc is negligible compared with that of ^{99m}Tc, even when present in amounts that are 10 or 100 times higher, the two isomers react in the same way with the components of a radiopharmaceutical kit. The labeling and imaging of some radiopharmaceuticals are adversely affected by a high atomic ratio of ^{99}Tc to ^{99m}Tc [57].

The $^{81}Rb \longrightarrow {}^{81m}Kr \longrightarrow {}^{81}Kr \longrightarrow {}^{81}Br$ Generator. The radionuclidic production process for this generator used in pulmonary studies involves irradiation of natural krypton gas in a target chamber by the ^{82}Kr (p, 2n) reaction and dissolution of the ^{81}Rb formed with water, steam, or saline solution. The resulting solution can then be transferred into a hot cell for loading into the generator. ^{81m}Kr is washed out of the generator by continuous elution with humidified air, and is used to produce a sequence of images from multiple breaths of the patient, enabling reconstruction of time – activity images. A systematic evaluation has been carried out of the excitation functions for formation of several rubidium radionuclides in proton-induced nuclear reactions on natural krypton, ^{82}Kr, and ^{83}Kr, with special reference to the production of ^{81}Rb for generating ^{81m}Kr. The method of choice for the preparation of ^{81}Rb ⟶ ^{81m}Kr generator system is the ^{82}Kr (p, 2n) process on highly enriched ^{82}Kr. The optimum energy range for production is 19 –27 MeV, and the thick-target yield of ^{81}Rb amounts to 48 mCi μA^{-1} h^{-1}. Due to the high yield of the proton process sufficient quantities of ^{81}Rb can be prepared even with a cyclotron in the energy range of 20 MeV [58], [59].

The $^{52}Fe \longrightarrow {}^{52m}Mn \longrightarrow {}^{52}Mn \longrightarrow {}^{52}Cr$ Generator. *^{52m}Mn* has shown potential as a myocardial imaging agent and for evaluating the efficiency of therapeutic intervention on ischemic perfusion. ^{52m}Mn decays by isomeric transition and positron emission to a much longer-lived radionuclide. Its high-flux 511 keV annihilation γ-rays make it ideal for imaging in emission-computed tomography. *^{52}Fe* itself has nuclear-medicine merit as an erythropoietic marrow agent, because it is the only radionuclide of iron with decay properties suitable for imaging applications. It can be produced by α, ^{3}He, and proton reactions on ^{50}Cr, ^{52}Cr, or ^{55}Mn. Using particle energy ≤ 70 MeV the maximum thicktarget yield is 160 μCi μA^{-1} h^{-1}. Higher particle energies can yield up to 3.3 mCi μA^{-1} h^{-1} at a beam current of 70 μA. A stack of foils irradiated and processed at the LAMPF can produce as much as 1 Ci at EOB after 16 h irradiation. This allows production of multicurie generators for hospital use. A similar method for ^{52}Fe production is based on a nickel target, which is simpler to use [27], [60].

The $^{113}Sn \longrightarrow {}^{113m}In \longrightarrow {}^{113}In$ Generator. ^{113m}In has found considerable application in nuclear medicine with its convenient half-life of about 100 min, a monoenergetic γ radiation of 393 keV, and the long half-life of the parent ^{113}Sn (118 d). Only 2 – 3 generators of this pair need be purchased per year to obtain a continuous supply of the daughter, compared to weekly procurement of a ^{99m}Tc generator. ^{113m}In has been used in various radiopharmaceuticals in scanning the brain, blood-pool, reticuloendothelial system, and kidney. The parent ^{113}Sn may be produced either by ^{113}In (p, n) or by ^{112}Sn(n, γ) reactions and is obtained in relatively high purity [54].

The $^{82}Sr \longrightarrow {}^{82}Rb \longrightarrow {}^{82}Kr$ Generator. This generator, used for cardiac PET imaging, contains about 100 mCi of ^{82}Sr ($t_{1/2}$ = 25 d) and lasts 4 – 6 weeks. It is the single most important cardiovascular imaging generator, because ^{82}Rb is useful for perfusion imaging in diagnosing coronary artery disease, acute myocardial infarction, and ischemia. The major production methods of ^{82}Sn are Mo (p, spallation), Rb (p, *x*n), and Kr (α, *x*n). Because of the short $t_{1/2}$ of the daughter radionuclide (76 s), serial PET studies can be conducted at 5 – 10 min intervals. If provided in 100 – 200 mCi amounts, the generator can provide adequate ^{82}Rb elution yields for three months. If ^{82}Sr is produced by spallation of a molybdenum target with 800 MeV protons (as in LAMPF), it may contain a 1 : 1 ratio of isotope ^{85}Sr ($t_{1/2}$ = 65 d), so for practical purposes this should be considered as a $^{82/85}$Sr column. The radiation doses to the skeleton of a 70 kg person from 1 μCi of ^{82}Sr/^{82}Rb and ^{85}Sr are 71 and 7.6 mrad, respectively [24], [61].

The $^{68}Ge \longrightarrow {}^{68}Ga \longrightarrow {}^{68}Zn$ Generator. ^{68}Ge is obtained by the ^{69}Ga (p, 2n) reaction in high yield. The practicality of ^{68}Ga emerges from its ability to form chelates with proteins and other compounds, and from its long $t_{1/2}$, which permits its continued use for about a year. Medical interest in this radionuclide developed from its potential for imaging thrombi formation with labeled platelets and imaging atheromas with bifunctional chelate-labeled lipoproteins. Since ^{68}Ga is obtained from the generator as a chelate of EDTA,an extra step must be incorporated into its preparation if ionic gallium is required [136].

The $^{191}Os \longrightarrow {}^{191m}Ir \longrightarrow {}^{191}Ir$ Generator. The radionuclide ^{191m}Ir ($t_{1/2}$ = 5 s) is suitable for pediatric radionuclide angiography, for the quantitation of shunts and ejection fractions, and for the diagnosis of congenital cardiovascular

malformations. Thermal neutron irradiation of enriched ^{190}Os for production of ^{191}Os is feasible at a flux as low as 4×10^4 n cm^{-2} s^{-1}. ^{191}Os can then be purified by conversion to OsO_4 and isolation by distillation and solvent extraction. Contamination with ^{192}Ir is produced, and this must be eliminated prior to medical use. Multicurie amounts can be prepared by irradiating enriched ^{190}Os in a high-flux reactor for 3 – 4 d. Long periods of irradiation may result in increasing levels of induced radionuclide impurities and decreasing yields of ^{191}Os due to neutron-capture (burn up) of this radionuclide [62], [63].

The $^{195m}Hg \longrightarrow {}^{195m}Au \longrightarrow {}^{195}Au \longrightarrow {}^{195}Pt$ Generator. This generator is suitable for first-pass and steady-state radionuclide cardiac angiography. Its attractiveness stems from its availability as a carrier-free single-photon emitter in a biologically-inert state, suitable for vascular imaging. Dynamic functional studies, especially cerebral blood-flow measurements, require repeated patient investigation at short intervals, and the advantage of ^{195m}Au over ^{81m}Kr is the fact that it is not quickly eliminated by the lungs after IV bolus. Also, ^{195m}Au emits predominantly 262-keV γ quanta and gives a clean spectrum for analysis. Typical production yield at a proton energy of 32 MeV is 5.4 mCi μA^{-1} h^{-1}. For medical use, such a preparation should be relatively free of the high-energy contaminants ^{195m}Hg (560 keV) and ^{195}Hg (1.17 MeV, $t_{1/2}$ = 9.5 h) [64], [65].

The $^{224}Ra \longrightarrow {}^{212}Pb + {}^{212}Bi \longrightarrow {}^{208}Pb$ Generator. Increased awareness of the therapeutic potential of α emitters, especially short-lived ones, has prompted the development of generators for such radionuclides. ^{212}Pb and ^{212}Bi are suited to use as such radiotherapeutic agents due to their relatively short half-lifes and appropriate α-particle emissions. The parent ^{224}Ra is available from a long-lived parent ^{228}Th, and can be isolated on a generator that produces the daughters in good yield and low breakthrough. ^{212}Bi, which has a half-life of 61 min, and is conveniently available in large quantities from this generator, may be bound to proteins via a bifunctional chelate. It is retained in various tumors, and a 25 mCi generator suffices for clinical use. ^{212}Bi alone delivers 24 500 g · rad/mCi, and the ^{212}Pb 250 000 g · rad/mCi [66].

Another useful way of obtaining the *radiolead isotope ^{211}Pb* is from a gaseous generator daughter via the $^{227}Ac \longrightarrow {}^{223}Ra \longrightarrow {}^{219}Rn \longrightarrow {}^{211}Pb$ reaction. In this reaction ^{211}Pb is produced in a generator system that utilizes distillation of the intermediate daughter ^{219}Rn from ^{223}Ra. The ^{223}Ra is precipitated as a stearate to permit its isolation from ^{227}Ac by cation exchange. While the yield of the system is only about 10 %, the radionuclidic purity is very high [67].

The $^{109}Cd \longrightarrow {}^{109m}Ag \longrightarrow {}^{109}Ag$ Generator. This generator is used primarily for pediatric imaging and blood flow studies. Special interest in it stems from the fact that a hospital can obtain several years of use from a single generator because the physical half-life of the parent radionuclide is 453 d. The daughter ^{109m}Ag produces low-energy γ-rays (88 keV) and has a 4-s half-life [54].

The $^{72}Se \longrightarrow {}^{72}As$ Generator. The product of this generator is a positron emitter, and may be applicable to the labeling of monoclonal antibodies and for toxicological studies. It is mentioned here because of the rich organic chemistry of arsenic, which can easily be incorporated into various chemical structures. Due to the intermediate half-life of ^{72}As ($t_{1/2}$ = 26 h), it can also be used for brain imaging, while the longer half-life of the parent ($t_{1/2}$ = 8 d) permits such a generator to last for 3 – 4 weeks [24].

The $^{115}Cd \longrightarrow {}^{115m}In \longrightarrow {}^{115}In$ Generator. ^{115m}In emits 300-keV conversion electrons and has a range of approximately 1 mm in soft tissue. This short range and a $t_{1/2}$ of 4.5 h makes it advantageous for radiation synovectomy.

The $^{201}Pb \longrightarrow {}^{201}Tl \longrightarrow {}^{201}Hg$ Generator. The ^{201}Tl radionuclide is accelerator-produced and used for evaluating regional myocardial metabolism (see also Section Radionuclide Production from Cyclotrons). Because of significant decay losses of ^{201}Tl during transportation an unusual generator was designed — unusual in the sense that the $t_{1/2}$ of the parent (9.4 h) is much shorter than that of the daughter (73.5 h). In a typicial generator, separation of the ^{201}Tl on an ion-exchange Dowex resin yields 400 mCi of ^{201}Tl from 4 Ci of ^{201}Pb, in a small extraction volume, with no detectable radioactive lead breakthrough and < 2 µg/mL of carrier thallium [68].

The $^{128}Ba \longrightarrow {}^{128}Cs \longrightarrow {}^{128}Xe$ Generator. Production of ^{128}Cs from spallation reactions, using a 85 MeV proton bombardment on a cesium chloride target, was reported as early as 1950. Similarly, production of ^{128}Cs by high-energy proton beams (> 200 MeV) has also been reported, using a $CsNO_3$ target. ^{128}Cs has potential applications for positron tomography, particularly for imaging the myocardium (where it is being considered as a potassium analogue for its trapping mechanism), and for measuring regional blood flow. Its $t_{1/2}$ is short enough to allow its use in sequential studies with PET systems. Cyclotron production of the pair Ba ⟶ Cs in transient equilibrium has yielded 3.1 mCi μA^{-1} h^{-1} with a 2.3 g/cm^2 thick CsCl target and a proton energy of 67 MeV. Ion exchange can be used for preparing highly-purified ^{128}Cs for the production of this generator [69].

[Top of Page]

4. Detection of Ionizing Radiation and Radionuclides

Horst Miska

The basis for using radioactive materials in research, medicine, and industry is the fact that these substances can be

detected easily and with high sensitivity through the radiation they emit. This nuclear radiation has the ability to ionize atoms and molecules, so that even the tiniest quantities of a radioactive material can be detected by physical methods. These detection limits cannot be achieved with chemical methods.

Chapter Detection of Ionizing Radiation and Radionuclides begins with a brief description of the types of ionizing radiation and its interactions with matter, then proceeds on to a discussion of detection methods and instruments. A detailed discussion of ionizing radiation is given in [70-74]; details on instruments can be found in manufacturers' catalogs and operating instructions [75-84] contains a summary of European manufacturers and suppliers of detectors.

4.1. Types of Ionizing Radiation

As explained in Chapter Introduction, radioactive materials can emit alpha (α), beta (β), or gamma (γ) radiation. The neutrino radiation (ν) emitted in every β decay event is of no importance for practice; it is extremely difficult to detect and is therefore ignored in the discussion that follows.

In α *decay*, the mass of a heavy nucleus decreases by four atomic mass units and an α particle (^{4}He nucleus) is emitted. It has a mass of M = 4 and a positive charge of two. The energy released in the decay event is precisely fixed for every nuclide, so α particles have a discrete energy distribution. In β^+ *and* β^- *decay* a proton is converted to a neutron or a neutron to a proton and a positron or electron is emitted. β particles thus have a charge of one unit, positive or negative; the mass is ca. 1/1836 times that of a proton or neutron. Although the energy available in a β decay event is fixed for each nuclide, β particles have a continuous energy distribution because the neutrino that is emitted simultaneously takes up part of the energy.

These types of nuclear radiation are directly ionizing by virtue of their charge. The *gamma radiation* that usually follows α or β decay is a purely electromagnetic phenomenon; it has no charge and no (rest) mass and thus does not directly cause ionization. Because the excited state of the nucleus emitting γ radiation is exactly defined, γ radiation has a discrete energy spectrum. In a similar way, the *neutrons* released in nuclear fission, nuclear fusion, and other nuclear reactions cannot directly ionize matter since they are uncharged. Again, the energies available to the neutron are well-defined for every event, but the fact that several particles are involved in neutron-generating reactions means that neutrons do not generally exhibit a discrete energy spectrum.

The types of radiation discussed up to this point originate in radioactive materials or may, as in the neutron case, come from mixed sources (e.g., polonium/beryllium). *X-rays*, used chiefly in medical diagnosis, are emitted as bremsstrahlung when accelerated electrons are stopped in matter; this is a purely electromagnetic radiation like γ radiation, but because of the way in which it is produced, its spectrum is continuous with a maximum energy corresponding to the kinetic energy of the accelerated electrons. These bremsstrahlung X-rays are accompanied by characteristic X radiation having discrete energies, which depend on the anode material used. Like visible light, this characteristic radiation originates in the atomic electron shell (excess binding energy is released when electrons drop into the vacancies that have been created in the inner shells by electron bombardment).

Electromagnetic radiation generated by electron accelerators with final energies of several MeV is no longer called X radiation but *bremsstrahlung.* Its maximum energy equals the electron energy, and its spectrum is continuous like that of X-rays.

Other forms of ionizing radiation are produced by *high-energy particle accelerators*, either as desired radiation for use in fundamental research, industry, and medicine (therapy), or as an undesired side effect when the particle beam is stopped. The interactions and detection of these types of radiation are not discussed here, although the techniques are similar in principle to those covered.

Finally, part of our natural radiation exposure is due to *cosmic rays*, which originate when energy is transformed by nuclear fusion in the sun and stars. Direct cosmic radiation comprises the rays that penetrate to the Earth's surface. Its principal constituents are muons (μ), light charged particles similar to electrons and detected in the same way. In addition, cosmic radiation produces many varieties of secondary radiation, which are of interest here only because they make up part of the background radiation measured with detecting instruments.

4.2. Detection Methods

When radiation passes through matter, the particles of radiation are subject to interactions whereby they give up energy to the matter and are either stopped or absorbed. Because of the electromagnetic interaction of *charged particles*, these particles are persistently slowed by ionization and excitation of the atoms and molecules in the material involving a large number of single events. *Uncharged* (*neutral*) *radiation*, in contrast, is subject to “all or nothing” interactions; in only one event the primary radiation particle vanishes, i.e., it gives up its total kinetic energy (or at least a substantial part of the energy) all at once.

The instruments and methods used to detect radiation depend on the types of interaction involved and the objectives pursued.

4.2.1. *Interactions Between Ionizing Radiation and Matter*

4.2.1.1. *Energy Loss of Charged Particles*

The energy loss suffered by charged particles as a result of ionization and excitation heavily depends on the properties of the particles and only slightly on those of the irradiated matter. The energy loss dE over a path length dx can be approximated to:

$$\frac{dE}{dx} = -\varrho C \frac{Z}{A_r} \frac{z^2}{v^2} \left[\log \frac{m_e v^2}{I(1-\beta^2)} - \beta^2 \right] \quad (4.1)$$

Here, C is a constant, ϱ is the density of the material, Z is its atomic number, and A_r its relative atomic mass; z and v are the charge and velocity of the particle, respectively. The expression shown in square brackets varies slowly with velocity (as the logarithm) and describes the increase of energy loss at relativistic energies.

If the velocity v of the particle is expressed in terms of the mass m and the kinetic energy $T_p = 1/2\, mv^2$, the following expression results. (It has been further simplified, includes a new constant C', and is normalized to the density of the material):

$$\frac{dE}{\rho dx} = -C' \frac{Z}{A_r} \frac{mz^2}{T_p} \left[\log \frac{m_e v^2}{I(1-\beta^2)} - \beta^2 \right] \quad (4.2)$$

Particles having kinetic energy of the same order as their rest mass energy (such as electrons at a few MeV) have a minimal ionizing effect. The differential energy loss $dE/\varrho\, dx$ for such particles is almost 2 MeV/(g/cm^2). The loss increases sharply at lower energies and is proportional to the z^2 and m of the particle. For α particles with a given kinetic energy, the energy loss is therefore roughly 10^4 times larger than for slow electrons (β particles) having the same kinetic energy because the following holds: $(z_\alpha/z_\beta)^2 = 4$ and $m_\alpha/m_\beta \approx 8000$ (4×1836).

The differential energy loss by ionization and excitation is shown in Figure 20.

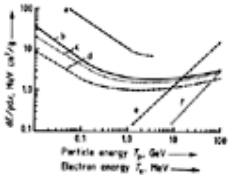

Figure 20. Energy loss of charged particles

Ionization losses: a) α Particles in silicon; b) Protons in silicon; c) Electrons in air; d) Electrons in lead Radiative losses: e) Electrons in lead; f) Electrons in airDifferent energy scales are employed for electrons (MeV) and for heavy particles (GeV)

Because of the persistent stopping by ionization and excitation, charged particles have a well-defined maximum range, which in principle depends on their energy alone. The differential energy loss is greatest toward the end of the range; the ionization here is most intense. The range for β particles in air is ca. 3 m, whereas that for α particles is ca. 10 cm. The corresponding ranges in water or tissue are less than 1 cm and 1/10 mm, respectively. Because α particles are completely stopped by a layer corresponding in thickness to a sheet of paper, the instrument used to detect α radiation must have an extremely thin entrance window.

4.2.1.2. *Bremsstrahlung Production*

For light charged particles having a kinetic energy T_p of the same order as their rest mass m_0c^2 (or higher), energy loss due to bremsstrahlung production also plays a major role. This energy loss, normalized to the material density, is

$$\frac{dE}{\rho dx} = -C'' \frac{Z^2}{A_r} \frac{T_P}{m^2} \left(\text{for } T_p > m_0 c^2\right) \quad (4.3)$$

Z^2 represents a strong material dependence, and the radiative loss also decreases as $1/m^2$, so that this energy loss chiefly affects electrons. The bremsstrahlung loss for cosmic-ray muons, which have a mass roughly equal to 200 electron masses, is therefore just 1/40 000 of the loss for electrons.

The “radiation length” t_0 has been introduced in order to describe the energy loss of electrons. This material-dependent quantity is the distance over which bremsstrahlung losses reduce the energy of an electron to $1/e$ times its initial value. The radiation length is a function of the mean Z of the material. It is about 36 cm for water, 9 cm for aluminum, and only 0.5 cm for lead. Thus fast electrons must be stopped by a material with a long radiation length (i.e., small Z) if the production of bremsstrahlung radiation is to be minimized or if all the energy transferred to the material (detector) is to be determined in particle spectroscopy.

4.2.1.3. *Energy Loss of Neutrons and Gamma Quanta*

Neutrons and γ quanta do not experience energy loss by ionization, excitation, or by bremsstrahlung since they have no charge.

Neutrons. Neutrons transfer only part of their energy to nuclei in elastic or inelastic collisions. An elastic collision imparts only kinetic energy to the nucleus; in an inelastic collision, the nucleus is also excited, and this energy is later released as γ radiation. The energy transferred to the recoil nucleus is greater the smaller its mass. In the extreme case, a neutron can give up all of its energy to a hydrogen nucleus ($A_r = 1$); on average, a neutron loses half of its kinetic energy in a collision with a proton. Materials with low nuclear mass, i.e., with a high hydrogen content (e.g., paraffin or polyethylene), are therefore used to stop and detect neutrons. The kinetic energy of the recoiling proton or nucleus is transferred by ionization in the material over a short distance, since these particles have high ionizing power by virtue of their large mass and charge (see Eq. 4.2).

Neutrons decelerated in this way, now having low energies, are then absorbed by nuclei because the cross sections are generally proportional to 1/*v*. They initiate a variety of reactions:

1. In (n, γ) reactions, the binding energy of the captured neutron is subsequently emitted in the form of γ radiation
2. In (n, p) and (n, α) reactions, a proton (p) or α particle is emitted after the capture of the neutron
3. In the (n, f) reaction, which possibly occurs only with some superheavy nuclei, a nuclear fission takes place after the neutron capture

The (n, p) and (n, α) reactions are, therefore, preferred for the detection of slow neutrons, since protons and α particles are easy to detect. The (n, f) reaction is suitable for extremely sensitive neutron detection because of the amplifying action of nuclear fission events (neutron multiplication, high-energy fission fragments).

Gamma Radiation and Bremsstrahlung. The main interactions of γ and bremsstrahlung radiation with matter are the photoeffect, Compton scattering, and pair production.

The *photoeffect* (photoelectric effect) involves the complete absorption of a γ quantum, which transfers its energy to a stronlgy bound electron (usually a K electron) in the material which is then expelled. In order to conserve momentum and energy, the nucleus must also acquire recoil energy. The vacancy in the K shell is then refilled, either with the emission of characteristic X radiation or Auger electrons. If this X radiation is also absorbed in another photoeffect in the material (of the detector), the γ quantum has then given up its total energy to charged particles in the material. The cross section for the photoeffect is proportional to Z^4 to Z^5 of the material and also to 1/*E* to $1/E^{3.5}$ of the radiation.

In *Compton scattering*, a γ quantum gives up part of its energy to a (free) electron; the scattered γ escapes in a direction different from its original one. The energy transfer reaches maximum in backscattering (180°), where the electron acquires an energy T_e:

$$T_e(\theta = 180\,^\circ) = \frac{E_\gamma}{1 + 511\text{keV}/2E_\gamma} \tag{4.4}$$

The Compton-scattering cross section is proportional to *Z* (i.e., to the number of electrons in the material) and $1/E_\gamma$.

Pair production is the inverse process to bremsstrahlung. An electron – positron pair is created in the material. A nucleus must be involved in the process as a collision partner, which acquires recoil energy. The formation of the electron and positron requires 511 keV each; the remainder is imparted to these as kinetic energy. Above the threshold of $2\,m_e$ = 1022 keV, the pair production cross section is proportional to Z^2 and to log E_γ. The distance t_P over which the number of γ quanta decreases to 1/*e* by pair production is $t_P = (9/7)t_0$. This result makes it clear that the process is related to bremsstrahlung production.

The electron and positron are stopped by ionization in the material. Together with an electron, the positron forms positronium, a species similar to a hydrogen atom (see Section Interaction between Positrons and Matter), which rapidly transforms its rest mass energy to radiation (annihilation radiation). Usually the results is a pair of γ quanta, each with energy 511 keV, propagating in opposite directions. Three γ quanta are formed in rare cases, the sum of their energies being 1022 keV.

With coordinates *Z* versus E_γ, Figure 21 shows the regions in which the several effects are dominant. The lines do not represent sharp boundaries but rather indicate where the reaction cross sections are just equal. The photoeffect dominates at low γ energy and high atomic number of the material, while pair production is dominant at high energy and high *Z*. The Compton effect is the most frequent interaction at small *Z* and moderate energy. High-*Z* materials are thus needed for the shielding and detection of γ radiation, because the γ quanta should be absorbed by the photoeffect rather then scattered by the Compton effect.

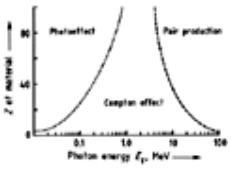

Figure 21. Interaction of γ-rays with matter

Cross sections for the photoeffect and Compton effect are equal (. . . .); cross sections for the Compton effect and pair production (- - -) are equal

Because they do not undergo persistent stopping by ionization, no well-defined range can be identified for γ and neutron radition. Instead, a half-value layer or tenth-value layer is defined as the thickness of material after which only half or a tenth of the initial particles are still present. The quantity usually cited in radiation protection is the material thickness behind which the dose rate is only a tenth of the original value.

4.2.2. *Detector Types*

Ionizing radiation can be detected by different techniques: collecting and registering the charge carriers liberated in the material by ionization, or studying other defects produced by the radiation. Common to all methods, however, is an amplification step in order to detect the energy deposited in the material by a particle or quantum.

Detector types are classified by material: gas ionization detectors, scintillators, semiconductor detectors, and solid-state detectors. Which one is selected will depend on what is required; compromises are often necessary, because generally

one type cannot meet all the requirements.

4.2.2.1. *Gas Ionization Detectors*

The simplest detector is the gas ionization type, usually called counter tube (see Fig. 22).

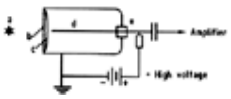

Figure 22. Schematic of a counter tube

a) Source; b) Protective screen; c) Window; d) Counter wire; e) Insulator

The radiation enters the active volume of the counter through a window (c). The chamber is most often filled with a special counter gas (filling gas). The case of the counter tube is at ground potential, whereas a positive voltage is imposed on the counter wire (d) in order to attract the electrons liberated in the gas. If the voltage on the wire is so low that just those electrons that are liberated by radiation in the chamber reach the wire, the device is referred to as an *ionization chamber.* The electron current flowing to the wire is proportional to the energy deposited in the chamber by radiation. The current is extremely small and must be electronically amplified before detection.

If the wire is held at a higher voltage, the electrons are accelerated until they can produce further ionization, so that the number of electrons reaching the wire increases in an avalanche-like fashion when propagating to the wire; that is, the electrons undergo gas multiplication. In Figure 23, the gas multiplication factor (gas amplification), that is, the number of electrons reaching the wire divided by the number of electrons liberated in the active volume by the radiation, is plotted versus the applied voltage.

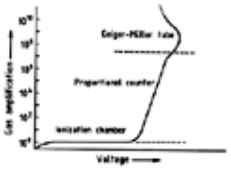

Figure 23. Response of counter tubes

The gas multiplication factor at a given voltage is constant in the *proportional counter region*; typical values range from $10^6 – 10^8$. The number of electrons detected is proportional to the energy deposited in the gas by the radiation. The pulse height or pulse shape can thus be used to distinguish pulses initiated by α particles from those initiated by β particles or γ quanta. The particle energy can also be determined, so the proportional counter can also be employed in spectroscopy.

The voltage can be further increased until enough photons are formed in the gas to produce photoelectrons in the gas and the walls. Ultimately, the entire volume of gas is ionized in a counter pulse (similar to what happens in a fluorescent tube). Regardless of the energy transferred from the particle to the gas, an extremely large pulse is generated. This pulse is followed by a long detector dead time. A device operated in this mode is called a *Geiger – Müller counter.*

Ionization chambers are often used to measure a local dose rate; when in the form of pen dosimeters, they can also measure individual dose. *Proportional counters* find use in contamination detectors and in the measurement of activities in air, for example, that of radioactive inert gases. Special filling gases such as BF_3 also permit the detection of neutrons, in this case via the (n, α) process. *Geiger – Müller counter tubes* are employed chiefly in dose-rate meters because of their robustness and small size.

4.2.2.2. *Scintillation Detectors*

It was discovered early on that α particles initiate flashes of light when they impact on certain materials such as zinc sulfide. Nowadays, scintillation detectors are made from special transparent materials, in which a large fraction of the energy deposited in the material by the particles is emitted in the form of light. Traces of other substances are generally added in order to enhance the light output. Figure 24 shows a schematic diagram of a scintillation detector.

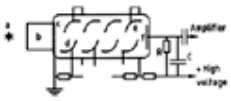

Figure 24. Schematic of a scintillation counter

a) Source; b) Scintillator; c) Photocathode; d) First and e) last dynodes in photomultiplier; f) Anode. The voltage divider supplying the dynodes and the anode is merely indicated

The intensity of the emitted light is proportional to the energy deposited in the scintillator by the particle. Falling on the photocathode (c) of a photomultiplier, this light sets free electrons, which are accelerated toward the successive dynodes (d – e) by the applied voltage. Every impact of an electron on a dynode generates several electrons. In this way, electron multiplication factors of 10^6 or more can be achieved, depending on the number of dynodes and the high voltage imposed on the tube. The total number of electrons arriving at the anode is thus a measure of the energy given up by the particle.

Charged-particle spectroscopy is usually performed with scintillators made of organic compounds (e.g., naphthalene or anthracene), called plastic scintillators. Because of their high hydrogen content, these devices can also detect neutrons via their recoil protons.

Gamma Radiation. The detection of γ radiation, however, requires a material of high density and high atomic number. Scintillators made of sodium iodide or cesium iodide are used for this purpose; a new development are oxides, e.g.,

bismuth germanate ($Bi_4Ge_3O_{12}$, BGO) having a high mean *Z* and a high density. The detector efficiency depends heavily on the size of the scintillator. In the laboratory, scintillators measuring up to 3×3 inch (ca. 76 mm in diameter and height) are common, but scintillators ranging up to several liters in volume are available for special applications, such as the detection of high-energy bremsstrahlung.

The variety of interactions between γ radiation and matter means that a γ detector will have a complex energy spectrum. If the γ quantum gives up its total energy to the detector in a single event (photoeffect) or in several events over a short time, the event will register in the total absorption line or photopeak. On the other hand, a high-energy quantum may impart its energy to the detector by pair production, and one or both of the 511 keV quanta emitted afterwards as annihilation radiation may escape detection. Since the energy of the escaping quanta is not registered in the detector, such an event is recorded under the single-escape or double-escape peak, respectively; these lines lie 511 and 1022 keV below the photopeak. A Compton continuum is present in addition to these lines; it arises because quanta escape from the detector after a Compton scattering event and their energies are not detected. What is more, the experimental geometry may result in other lines produced by backscattering. All in all, the interpretation of γ spectra is very complex, especially if the source exhibits more than one γ line.

Liquid scintillators (LSC) are used to detect low-energy β radiation such as that from ^3H or ^{14}C. Such a scintillator consists of organic compounds to which the specimen material (e.g., biological or medical samples with ^3H or ^{14}C as tracers) is added. This configuration ensures that the total β energy will be absorbed in the scintillator. The specimen – scintillator liquid is observed by two or more photomultipliers connected in coincidence, in order to reduce the background rate due to multiplier noise.

4.2.2.3. *Semiconductor Detectors*

Since researchers have succeeded to fabricate large-volume semiconductor crystals, these materials have gained importance in the spectroscopy of ionizing radiation. In semiconductor detectors, the *n* layer is doped much higher than the *p* layer, so that the thicknesses of the respective depletion layers are drastically different. The highly doped *n* layer then forms a thin entry window for the radiation, while the depth of the depletion region in the *p* layer can be increased by applying a reverse bias (similar to that carried out in an electrical semiconductor diode). The depletion region forms a zone almost free of charge carriers, which then becomes a “solid-state ionization chamber.” The charges liberated in this volume by ionization are delivered to a charge-sensitive preamplifier, whose output signal is then proportional to the energy deposited in the semiconductor by the particle.

Silicon surface-barrier detectors (SSBDs) are employed for spectroscopy involving *charged particles*, such as electrons, protons, deuterons, or α particles. If it is desired to distinguish between particle species, a telescope consisting of one or more thin detectors with a thick detector mounted behind is used. The energy loss of the particle is measured in the thin detector(s), while the sum of all pulses is proportional to the total particle energy. Since the energy loss depends on the particle's charge and mass (Eq. 4.2), it can be identified once the loss is known.

Germanium detectors are used to detect γ *radiation.* Earlier these devices were doped with lithium to form Ge(Li) detectors, but these detectors are being replaced by high-purity germanium (HP – Ge) detectors that are now available and easier to use. Because the lithium diffuses out of the Ge(Li) at room temperature, such devices must be continuously cooled with liquid nitrogen; HP – Ge diodes, on the other hand, are cooled only during measurement and solely to reduce noise and improve the resolution. These detectors make high-resolution γ spectroscopy possible, so that the setup illustrated in Figure 25 permits nuclide identification based on the emitted γ radiation.

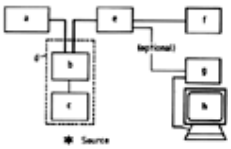

Figure 25. Block diagram of γ spectroscopy setup (nuclide identification)

Data analysis can be performed with a multichannel analyzer (MCA), or the analog pulses from the amplifier can be fed to a converter (ADC), digitized, and then input to a computer system for analysisa) Bias supply; b) Preamplifier; c) High-purity germanium; d) Dewar vessel with liquid nitrogen; e) Amplifier; f) Multichannel analyzer; g) Analog – digital converter; h) Personal computer

4.2.2.4. *Basic Properties of the Detector Types*

Along with ease of use and cost, the following properties must be considered when choosing a detector type: detector efficiency, energy resolution, and (for many applications) promptness of signal, i.e., pulse shape.

Resolving Power. The resolving power depends on several parameters but chiefly on stochastic fluctuations in the number of charge carriers detected. The relative error in this number is larger, the fewer the charge carriers originally produced. It takes about 3 eV to generate a carrier pair in a *semiconductor.* Thus, if the 661 keV γ quantum emitted after the β decay of ^{137}Cs is absorbed in the semiconductor through the photoeffect

$$N = \frac{661 \times 10^3}{3} = 2.2 \times 10^5 \tag{4.5}$$

charge carriers are produced. The relative statistical error of this number is

$$\Delta N/N = \sqrt{N}/N = 0.21\ \%\ ;$$

if a normal distribution is assumed, the full width at half maximum (FWHM) of the line is about 0.5 % for semiconductor detectors.

With a *scintillation detector*, an energy of ca. 300 eV is required to expel an electron from the photocathode of the

multiplier tube (with allowance for light losses, etc.). Hence, the statistical uncertainty for the same line is $\Delta N/N$ = 2.1 %, corresponding to an FWHM of 5 %. The formation of an ion pair in a *proportional counter tube* takes ca. 30 eV, so that the line width calculated in a similar way would then be about 1.6 %. The fundamental limitations described here cannot be greatly improved by experimental technique, so these estimates are generally valid.

Efficiency. The detectors described up to this point have near 100 % efficiency for charged particles, provided the particles can enter the active volume. For γ radiation and neutrons, the detector efficiency depends on the density and atomic number of the detector material and its volume, as well as the energy of the radiation. Gas ionization detectors are not well-suited to detecting γ radiation except at extremely low energies, when xenon-filled proportional counters can be employed. The efficiency of a 3×3 inch sodium iodide detector located 60 mm away from the source is 4.5 % for 661 keV γ-rays, decreasing to ca. 3.5 % at 1.461 MeV (^{40}K). The proportion of events registered in the photopeak is only 45 % in the first case and 27 % in the second. The activity must be calculated from the area under this line.

Promptness of Signal. The time-domain properties of the pulses from these detector types may somewhat limit their use. Because the charge-carrier avalanche propagates throughout the tube volume, a Geiger – Müller counter has a long dead time and is thus poorly suited to the measurement of high pulse rates.

Short pulse rise time is important, along with pulse length, in coincidence applications. The behavior of the pulse over time may also depend on the type of particle detected or the place where it is absorbed in the detector; this information is also utilized in some applications.

4.2.2.5. *Integrating Detectors*

Besides the detectors described in the preceding sections, which are operated chiefly in on-line mode (signals interpreted as they are received), radiation can be detected by other methods.

It has long been known that ionizing radiation will blacken *photographic film.* The effect has been studied qualitatively and quantitatively and is highly reproducible; for these reasons, it is frequently utilized in official individual dosimetry. In order to gain information about the energy of γ and X-rays involved, the film is placed in a cassette with filters that attenuate the radiation in an energy-dependent fashion.

Ionizing radiation produces defects in special *glasses* so that part of the incident energy is permanently stored in the glass. If such a material is illuminated, it luminesces, i.e., emits light of a different wavelength than the incident light. The luminescent intensity is proportional to the irridiation dose previously received by the glass. The device using this effect, called the *photoluminescence dosimeter* (*PLD*), is distinguished by high long-term stability, low background, and low fading. The stored information about the dose can be erased by heating; optical reading does not affect the stored information.

In *thermoluminescence dosimeters* (*TLD*), the energy stored by radiation in a crystal, for example LiF, is emitted as light when the crystal is heated to 200 – 400 °C. The information is, however, erased when it is read out in this way; the information can be read only once. If the crystal is enriched with ^{6}Li, the dosimeter will be sensitive chiefly to neutrons because of the large (n, α) cross section of this nuclide; ^{7}Li-enriched crystals find use in γ dosimetry.

4.3. Measurement of Activity and Dose

As stated in Chapter Introduction, *activity* is always referred to a given nuclide or group of nuclides. If the nuclide is not known, it must be determined before an activity can be measured.

With a few exceptions, all radionuclides emit γ radiation after a decay event, and the γ spectrum can be used for nuclide determination.

The measurement of a *dose* or *dose rate* is generally referred to the photon dose rate, since (as discussed in Chap. The Biological Effects of Radiation and Radionuclides) this is the only form of external radiation — aside from the β skin dose — that can cause damage. In most countries, instruments for measuring the γ dose are subject to a calibration requirement; in Germany, only instruments of desaster response units are exempt from this requirement. On the other hand, no calibratable instruments are now available for the determination of the beta skin dose. This dose is derived from the measured activity per unit area.

4.3.1. *Gamma Spectroscopy for Radionuclide Determination*

Figure 26 shows a typical setup for high-resolution γ spectroscopy with a high-purity germanium detector. The detector and the preamplifier are mounted in a Dewar to keep them cold. Nowadays, the multichannel analyzer is usually replaced by an analog – digital converter connected to a personal computer, so that the spectrum can be plotted, the data can later be analyzed in the form of an energy and intensity calibration, and the activity can be determined from the results.

The spectrum in Figure 26 (curve b) shows the result of one such measurement. The resolution is higher than that of the sodium iodide detector (curve a); this gain is offset by a loss of detector efficiency and hence a much longer measurement time. Whereas the high-energy lines stand above a low continuous background, this background rises sharply at lower energies because of the Compton continuum of the higher-energy lines. The analysis of such complex spectra consumes a large amount of machine time. It involves fitting analytic functions to the lines, from high to low energies, determining the area enclosed, and calculating the Compton component.

Figure 26. Typical spectrum output by the system of Figure 25

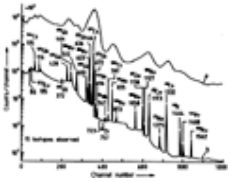

Reproduced with the kind permission of [77].

a) NaI(Tl) detector, 3+5 inch; b) High-purity germanium detector

Manufacturers offer complete systems with extensive programs and libraries. On the basis of the energy calibration, these programs assign the lines to the respective radionuclides; using the intensity calibration and the known emission rate, they can also calculate the number of decays.

The result of the analysis is reported as the absolute activity of the specimen or may be converted to a specific activity (in, e.g., Bq/kg) or an activity concentration (in Bq/m^3).

If a high detector efficiency is required and the expected spectrum is less complex, so that less resolving power will be adequate, the energy, intensity, and thus the activity can also be measured with a sodium iodide scintillator. This detector is more robust and does not need to be cooled; what is more, the electronic hardware has a somewhat simpler design. The interpretation technique is, however, similar in principle to that employed with the HP – Ge detector.

4.3.2. *Other Methods of Radionuclide Determination*

For α *emitters that show no* γ *lines* or only at very low energies, the nuclide and its activity can also be determined on the basis of α spec-troscopy. The high energy loss of α radiation means, however, that a very thin specimen must be prepared so that the energy loss in the specimen will not interfere with the measurements. The specimen-to-detector distance must also be small, or the volume in between must be evacuated. The detectors employed are high-resolution silicon surface-barrier detectors or the easier-to-use plastic scintillators, though the latter do not offer the same resolving power as the semiconductors.

Since the detection efficiency for charged particles is close to 100 %; the minimal size of the detector is governed solely by the maximum range of the particles to ensure total energy absorption. The interpretation of the spectra is similar to that of the γ spectra, except that here the Compton background is not present.

The same detectors can also be used for *electrons.* Because β spectra are continuous, they cannot serve for nuclide identification unless a chemical separation has been performed first. If no γ spectrometer is available, then some of the radionuclides that decay by *electron capture*, such as ^{207}Bi, can be detected by conversion electron spectroscopy since the conversion electron spectrum is discrete. ^{207}Bi preparations are often used for the testing and calibration of particle detectors.

4.3.3. *Determination of Surface Activities*

In radionuclide laboratories, surfaces are often examined with hand-held probes using proportional counter tubes or plastic scintillators. In either case, the probe detects the charged particles but does not determine their energy, so that nuclide identification is not possible. The surface activity can be determined from the known detector efficiency and the known sensitive area of the detector, provided it is known what nuclide is responsible for the contamination.

Examples of contamination monitors employing counter tubes are shown in Figures 27 and 28 (left-hand instrument). The device in Figure 27 is an supplementary probe for use with series AD dose-rate meters (Automess, Germany) [75]. As in most modern hand-held meters, the digital display is controlled by a process control computer, which also performs functions such as count rate correction, averaging, timedomain integration, automatic range switching, and comparison with alarm thresholds. The instrument can distinguish between different types of radiation in the following way. The counter tube has an α-transparent window and is permanently gas-filled. If a low voltage range is selected, the counter tube is sensitive only to intensely ionizing α particles. At high counter tube voltages, where β and γ radiation is detected, β particles may be blocked by a steel plate (shown clamped to the detector) to measure the γ radiation alone.

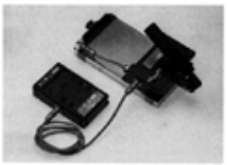

Figure 27. Model AD-k large-area proportional counter (right) for contamination measurement, shown as an accessory to the AD-2 dose-rate meter [75]

The AD-2 can also be mounted directly on the AD-k

Figure 28. Contamat model FHT 111 M large-area proportional counter for contamination measurement (left); FH 40 F 1 dose-rate meter (center); FAG FH 41 F dose and dose-rate alarm (right) [78]

The instrument in Figure 28 [78] can use a permanently xenon-filled counter tube, which is suitable for detecting low-

energy γ radiation but has a thick window that blocks α particles. For α detection, a flow-through counter tube is used; this must be purged with butane or methane at regular intervals. α Radiation is identified by pulse shape allowing α and β radiation to be detected at the same time and displayed one after the other. The instrument offers functions similar to those just described. Calibration factors can also be programmed-in for various nuclides, permitting the direct display of a surface activity in Bq/cm^2. This figure will be correct only if the stated nuclide is actually responsible and, only if it is present in the form of a thin film on a smooth surface.

4.3.4. *Measurement of Photon Dose Rate*

A widely used instrument for measuring the photon equivalent dose *H* or photon equivalent dose rate $\dot{H}$ is the *Geiger – Müller tube*, which directly shows a dose rate in Sv/h, mSv/h, or even μSv/h. Modern instruments such as those shown in Figures 27 and 28 feature automatic range selection and can serve as front ends for other probes. Such a device can be supplemented by a high-dose counter tube on a telescoping arm for measurements in hot places or those with restricted accessibility.

Because of their long dead time, Geiger – Müller counters should be used with care when measuring high dose rates from pulsed radiation sources (accelerators, X-ray tubes). The manufacturer usually indicates the range in which reliable measurements are possible with pulsed sources. Geiger – Müller counters may give low readings on very low-energy γ radiation too; the permissible energy range is indicated with each instrument.

Dose-rate meters based on *ionization chambers* are free from these drawbacks but are less convenient and more fragile. Their sensitive electronics must be calibrated and checked more frequently, and the instruments are more costly.

In addition, *scintillation dose-rate meters* are on the market. They are similar to ionization chambers in operation and can be employed for measuring extremely low dose rates. Because of their complex design, they are delicate and require care in use.

4.3.5. *Area Monitoring and Personal Dosimetry*

International recommendations as applied in most countries require measurement of the dose to persons who work in controlled access areas. In most such facilities, the *dose* (*local dose*) or local dose rate at certain places must be measured and recorded, and alarms must be given if prescribed threshold values are exceeded. Moreover, many countries operate wide-area systems, especially around nuclear power plants, to monitor the dose rate. These fixed monitoring systems generally employ proportional or Geiger – Müller counter tubes.

In most cases, two independent dosimeters are used to measure the *personal* (*individual*) *dose;* one is read by the wearer, for example, at the end of a working day, whereas the second must be evaluated by designated agencies. Dosimetric films and PLDs are permitted for this use in many countries.

Pen dosimeters, the size of a fountain pen and incorporating an ionization chamber, used to be employed as self-readable dosimeters. An internal capacitor was charged to a certain voltage, which was also imposed on the wire of the ionization chamber. The charge carriers produced by radiation then discharged the capacitor. The decrease in voltage was directly readable on a built-in electrometer or an external instrument. Dosimeters of this type were not, however, very accurate; they had a high background and were also sensitive to external interference. Most of the personal dosimeters now on the market are based on Geiger – Müller tubes (Fig. 28). Miniaturization techniques make it possible to fit such a device into the volume of a fountain pen. A built-in microprocessor detects and counts the pulses from the tube, which can be output as a dose figure on an LCD display. Many instruments have an integrated dose-rate alarm, which beeps when a predetermined value is exceeded.

In certain operations, a *partial body dose*, for example, a hand dose, must also be determined. Small LiF crystals are employed as TLDs for this purpose; they can be worn as ring dosimeters or applied directly to the skin with tape. Depending on the ^7Li – ^6Li ratio, they can be applied to measure either the γ or neutron doses. The range of possible shapes for TLD crystals means that they have a wide variety of applications; automated readers are available (e.g., [82]).

Depending on their chemical form and metabolism, *incorporated radionuclides* can be detected by excretion analysis (e.g., ^3H in the urine). The activity of incorporated γ emitters can be measured with body counters (whole-body counters). Most such instruments feature large-volume scintillators, which make it possible to detect even low activities such as that of ^{137}Cs in the body.

A summary of monitoring techniques and instruments is given in Table 4.

Table 4. Summary of monitoring techniques and instruments

Method	Instrument	Application/detection of	Remarks
Counter tube	ionization chamber	photon dose rate/charged particles	also suitable for pulsed radiation sources
	proportional counter	contamination measurement, dose rate/charged particles, neutrons (BF_3 filling), low-energy photons (Xe filling)	moderate-resolution spectroscopy possible; particle discrimination possible
	Geiger –	dose-rate measurement	sturdy; no particle

	Müller counter		discrimination; caution when using with pulsed sources
Scintillator	organic (plastic) scintillator	dose-rate measurement/charged particles, neutrons	low resolution for spectroscopy
	NaI, CsI, BGO	gamma spectroscopy, body counter	high detector efficiency, low resolution
Semiconductor	SSBD	charged particles	high resolution; telescope for particle identification
	Ge(Li), HP – Ge	gamma spectroscopy	high resolution, low detector efficiency
Solid-state	film	dosimetry	official dosimeters
	PLD	dosimetry	high constancy, long-time measurements, official dosimeters
	TLD	^{6}LiF: neutron dosimetry	long-term constancy not so good;
		^{7}LiF: photon dosimetry	small, therefore good partial body dosimeters

[Top of Page]

5. Uses

Alfred Zyball, Andreas Gervé, Alexandru T. Balaban

5.1. Industrial Applications of Radionuclides

5.1.1. *Introduction*

The use of radiation in industry had its starting point in fundamental research in the 1940s and 1950s. The equipment used includes:

1. Electron accelerators with maximum energies up to 10 MeV
2. Gamma facilities with suitable radionuclides

The energy restriction arises from the process requirement that the irradiated products must not become activated. Thus nuclear reactions are not initiated in the products, nor are radioactive substances produced. The products can be utilized immediately after irradiation.

This chapter deals with the industrial applications of radionuclides (for application of electron accelerators, see ⟶ Radiation Chemistry). The γ radiation emitted by the radionuclides is utilized.

The output of *electron accelerators*, with energy ranges up to 10 MeV, is called “machine radiation” or particle radiation, because the electrons emitted by glowing cathodes are accelerated in a high-voltage field and used as radiation. In comparison, γ-*rays* have the advantage that their penetration is greater in the irradiated product than that of accelerated electrons.

The use of γ-rays enables the irradiation of larger units of products, such as products stacked on a pallet. Electron accelerator systems can only uniformly irradiate smaller units of specified size, such as single packages.

With regard to radiation protection, electron accelerators have the advantage that the radiation can be stopped by cutting off the power supply to the machine. Personnel can then enter the irradiation cell provided the ozone concentration is monitored. In a gamma facility the radiation can only be shielded but not turned off. This fact is crucial in site selection (earthquake hazard), in the design and construction of the equipment and safety systems, and in the organization of operation.

Work is now under way aimed at combining the advantages of γ irradiation (greater penetration) with those of the electron accelerator [85], [86]. If accelerated electrons strike a target, secondary radiation (bremsstrahlung or X-rays) is produced; this has a greater penetration which is dependent on the radiation energy. The research objective is to improve the efficiency of conver-sion and to study the energy spectrum of the secondary radiation with a view to industrial application.

5.1.2. *Radionuclides for Industrial Gamma Facilities*

About 1300 man-made radionuclides are known but only two of them account for virtually all industrial uses:

Cobalt-60 (^{60}Co)
Cesium-137 (^{137}Cs)

The more commonly used radionuclide, ^{60}Co, is generated in nuclear reactors when the stable isotope ^{59}Co captures thermal neutrons:

$^{59}Co(n,\gamma)^{60}Co$

^{60}Co is produced in reactors with thermal neutron fluxes of 10^{12} – 10^{15} n cm^{-2} s^{-1}. The conversion of ^{59}Co to ^{60}Co is 99 % [87]. The maximum specific activity produced is about 3.7×10^{13} Bq/g (1000 Ci/g). Commercial ^{60}Co sources take the form of rods with double metal cladding (Fig. 29). During fabrication, the metal capsules are leak-tested in order to prevent contamination in service. The individual ^{60}Co sources fabricated in this way generally have an activity of ca. 2×10^{14} – 6×10^{14} Bq (6 – 15 kCi). The activity of the strongest individual source is about 4×10^{15} Bq (10^5 Ci).

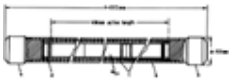

Figure 29. C-188 ^{60}Co radiation source at Nordion International Inc., Kanata, Ontario, Canada [88]

a) End cap; b) Body tube, *d* = 9.65 mm; c) Inner capsule; d) Spacer (as required); e) End cap

Around 1970, production of ^{60}Co was ca. 0.15×10^{18} Bq/a (4 MCi/a). A yearly output of 2×10^{18} – 3×10^{18} Bq (50 – 80 MCi) is exptected for the first part of the 1990s [89].

In 1991, 170 gamma-irradiation systems having a total activity of ca. 6×10^{18} Bq (150 MCi) were in use in 45 countries [90]. About 30 % of these systems are operated in service centers [91], [92].

Cobalt-60 Sources. Figure 30 shows a simplified decay scheme for ^{60}Co. The 314 keV β radiation is absorbed in the metal cladding around the ^{60}Co source and is not available for irradiation. The gamma quanta emitted by ^{60}Ni have energies of 1.173 MeV and 1.332 MeV; thus every decay results in the emission of two γ quanta with an average energy of 1.25 MeV. The half-life is 5.27 years (Table 5). The decline in activity is around one percent per month in practice. In industrial facilities, the source is generally in service for three half-lifes, i.e., 15 years. After this time, the source is either installed in a facility which requires irradiation with only a small activity, or it may be disposed of in the same way as radioactive waste from nuclear power plants. After ca. 200 years, by virtue of the short half-life, the activity will have decayed to a point where radiation exposure due to the source is at the natural background level.

Table 5. Characteristics of the radionuclides ^{60}Co and ^{137}Cs [93], [94]

Radionuclide	^{60}Co	^{137}Cs
Form	Metal (solid)	CsCl (water-soluble)
Half-life, a	5.27	30.17
Decrease in activity per year, %	12.39	2.28
γ energy, MeV	1.173 (1.0)	0.662 (0.858)
(emission probability)	1.337 (1.0)	
Half-value layer in water, mm	110	70
Self-absorption of energy, %	≈ 10	≈ 30
Radiation power, kW/10^{16} Bq (kW/MCi)	4.01 (14.84)	0.90 (3.32)

Figure 30. Simplified decay of ^{60}Co

Cesium-137 Sources. ^{137}Cs originates as a fission product of ^{235}U in nuclear reactors. In industrial irradiation systems, it is used in encapsulated form as the sulfate, cesium glass, or often as the chloride. According to [93], ^{137}Cs is one of the most abundant fission products of ^{235}U (55 % of waste radioactivity, 50 % of energy content). The amount of ^{137}Cs produced by nuclear reactors in the United States in 1987 was reported to be 3.5×10^{18} Bq (90 MCi). Despite the availability of this high activity, ^{137}Cs is seldom used in industrial gamma facilities. The reasons are the low energy of 0.662 MeV and the long half-life of the nuclide (see Fig. 31 and Table 6). The low energy of the emitted γ-rays also means that self-absorption of the radiation in the cladding of the radioactive material is high (30 –50 %). Even though the activity is high, the output of a ^{137}Cs gamma equipment is low. The most frequent application of ^{137}Cs is in pilot plants for the disinfection treatment of sludge [94].

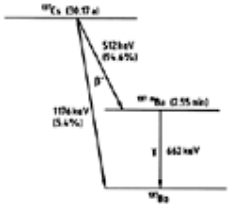

Figure 31. Simplified decay of ^{137}Cs

5.1.3. *Design of Gamma Equipment*

The initial planning of an irradiation facility that will utilize radionuclide sources must deal with two fundamental problems:

1. How to design the irradiation equipment so as to ensure safe operation, especially in compliance with radiation safety regulations
2. How to design and lay out the entire equipment in order to achieve a technically and economically optimal handling of the product to be irradiated

An industrial irradiation system includes the following basic devices (Fig. 32).

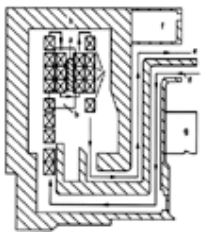

Figure 32. Gamma facility for irradiation of products on pallets

a) ^{60}Co sources; b) Water basin; c) Shipping packages; d) Entrance labyrinth; e) Exit labyrinth; f) Water treatment; g) Control room; h) Concrete shielding

Radiation Source. Sources are available in point, line, plate, or cylindrical shape; cylinders and plates being the most common (Fig. 32). In every source the source rods (Fig. 29) are assembled in such a way that the radiation field produced in the irradiation cell complies with the requirements on the irradiated product. The positions of the rods in the radiation source are calculated by computer.

Radiation Shielding. Shielding and other types of radiation safety equipment must be designed to enable safe operation and to prevent the radiation source from endangering the plant personnel or the environment. For this reason, the radiation source is enclosed in concrete walls that are 1.5 – 3.0 m thick. The walls and the labyrinth through which the product is conveyed must be designed so that the dose rate outside the irradiation cell (monitored area of the plant) is < 2.5 μSv/h (0.25 mrem/h).

Source Storage. The radiation source can be safely stored when personnel have to carry out necessary maintenance and repairs or when source rods need be replaced or repositioned inside the source. For storage purposes, the irradiation source can either be completely submerged in water to permit safe entry into the cell or in some plants, the source is securely stored by means of a radiation-proof cover (dry storage) [95].

Conveyor System. The conveyor system transports the product into the cell for irradiation and places it in the correct irradiation positions. Both functions must be efficiently carried out as the effect produced by irradiation in the product is largely dependent on them.

Ventilation System. The cell has a forced-air ventilation system to remove gases produced during irradiation (e.g., ozone, nitrogen oxides).

Cooling System. Because most systems built today have high activities (45×10^{18} Bq, i.e., 12 MCi), the source storage facility must be cooled. The irradiation cell itself must also be cooled in plants handling frozen products.

Monitoring and Control System. The first task of the monitoring system is radiation protection; this equipment ensures that the specified local dose rates are not exceeded; unauthorized access to the irradiation cell is prevented until the radiation source has reached its final position in storage; and that contamination does not occur as a result of any leakage from the source rods. The control system, which employs a microprocessor in new plants, monitors, controls, and documents the irradiation process.

Storage Facilities. Product storage facilities are associated with any industrial irradiation system. In order to clearly distinguish irradiated and nonirradiated products in the storage area, the two are stored separately.

5.1.4. *Irradiation Using Gamma Rays*

The changes in the product properties that occur during irradiation depend to some degree on the energy dose and are also influenced by the dose rate. The dose rate becomes a factor, however, only when a product is irradiated with a high-dose-rate electron accelerator rather than a low-dose-rate gamma system.

Achieving specified changes requires delivering a *minimum dose* to the product; there is also a *maximum dose* that must

not be exceeded in order to prevent undesirable damage to the product. For a product with homogeneous density, the *absorption* of radiation and thus the dose exponential relationship (Fig. 33). The shape of the curve is determined by the density of the absorbing material. As the figure shows, products are often irradiated from at least two sides in order to optimize the maximum/minimum ratio.

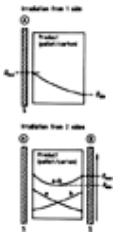

Figure 33. Dose distribution after irradiation from one and two sides

A and B positions of source relative to product

a) Dose distribution after irradiation from position A; b) Dose distribution after irradiation from position B; a + b) Resulting dose after irradiation from two sides

D_{max} = maximum dose; D_{min} = minimum dose; *S* = source; a + b=total dose

The following questions must be answered in the development of a facility.

1. What products will be irradiated in the gamma facility, and in what state will they be?
2. What are the minimum doses? What maximum values must not be exceeded?
3. What are the nominal doses for product irradiation?
4. What equipment efficiency is possible given the above requirements?

If granular or free-flowing products are irradiated in a system, 100 % efficiency is possible, because the product can be configured in accordance with the spatial pattern of γ emission from the radiation source.

The product to be irradiated is supplied in containers, cartons, or other packaging units and is transported past the radiation source by conveying systems. The efficiency of such installations is < 40 %. In the past, a number of concepts were devised and designed with the objective of improving the efficiency. The essential requirement is to surround the radiation source with products so that they correspond to the spatial γ emission pattern.

A few concepts are described in Figure 34. Further information can be found in [96-104]. Figure 34 A shows a cylindrical source with eight irradiation positions. The concept illustrated in Figure 34 B has four irradiation positions around the cylindrical source; the product must be irradiated in at least two configurations to ensure homogeneity of dose. Figure 34 C shows a plate-type source with ten irradiation positions. Again, the product must pass through different positions.

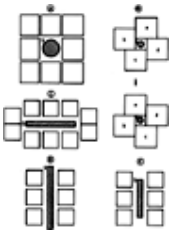

Figure 34. Configuration of sources relative to product

A) Cylindrical source with 8 irradiation positions; B) Cylindrical source with 4 irradiation positions; C) Plate-type source with 10 irradiation positions; D) Source overlapping product; E) Product overlapping source

Basic arrangements of radiation source and product are shown in Figure 34 D and 5.6 E. In Figure 34 D, the radiation source overlaps the product. This results in a homogeneous dose distribution but the efficiency is poor. The arrangement in Figure 34 E offers an optimized efficiency. The product overlaps the source. A complicated handling system is required in the irradiation cell to obtain optimal dose distribution.

A gamma facility should also exhibit flexibility or versatility. The unit must be able to irradiate products that vary in density (with specified maximum/minimum dose ratios), and it must deliver a range of doses with no loss of efficiency. There are two possibilities: First, the maximum/minimum ratio can be improved by reducing the irradiation depth. The dose distribution for a constant density is determined by the depth of the product in the irradiation direction (see Fig. 34). The key parameter for the maximum/minimum ratio is the product of density and size of the product. The depth of dose distribution is often referred to as the mass or weight per unit area.

The second possibility is to apply the total dose step by step, using an appropriately designed conveyor system inside the irradiation source. The total dose is then a multiple of the dose that the product absorbs while passing through the cell once.

5.1.5. *Uses of Irradiation*

Industrial applications of irradiation with radionuclides include:

1. Disinfection – sterilization
2. Modification of plastics (cross-linking – degradation)

5.1.5.1. *Disinfection – Sterilization*

Excited molecules, ions, and/or free radicals are formed in an irradiated product as a result of energy absorption. In living cells, e.g., microorganisms, this energy absorption leads to defects in the DNA. A repair mechanism for DNA exists, but if the number of defects highly exceeds the DNA repair capability, the microorganism will die. The rate of destruction is an

exponential relationship of dose if other environmental conditions remain constant. It is possible, therefore, if the microbial contamination of the product is known, to determine what dose is necessary to achieve an acceptable population level of microorganisms or to obtain a sterile product.

“Sterile” products are required in the medical field. Although this term is defined as “free of any microbial forms capable of reproducing, and also of viruses and bacterial spores,” all that is possible is a probabilistic statement of the product condition. Most countries therefore consider a product sterile if there is less than a 10^{-6} probability that a single viable microorganism is present.

The history of radiation sterilization since the 1950s has been described [105]. Intensive research on the resistance and inactivation of microorganisms has led to the recognition that a product fabricated in compliance with good manufacturing practise (GMP) standards and irradiated with a dose of 25 kGy is sterile as defined above. Because irradiation may also change the properties of the product or its packaging, statistical procedures have been devised so that products can be irradiated with doses of less than 25 kGy and still be reliably sterilized at the 10^{-6} level [106].

Radiation sterilization is employed with products that are thermolabile and thus cannot be sterilized with superheated steam. These include products made from plastics. Radiation sterilization offers other advantages in this area:

1. The product is sterilized in the final packaging, so there is no danger of subsequent recontamination.
2. The product can be utilized immediately after sterilization.
3. The sterilization process is easy to monitor. The destruction rate of the microorganisms depends on the radiation dose, a physically measurable quantity. The duration of the product in the irradiation equipment is the deciding factor in this process.
4. Radiation sterilization is no more expensive than sterilization with ethylene oxide (⟶ Ethylene Oxide – Uses). It may even be less expensive if safety standards and regulations in gas sterilization with ethylene oxide are taken into account.

About 95 % of installed ^{60}Co activity in 1990, i.e., 90 % of the total volumetric radiation capacity was utilized for the radiation sterilization of medical devices. Thus, in 1990, 45 % of medical devices were being sterilized with radiation instead of ethylene oxide; in 1977 this fraction was a mere 5 % [107].

The second most important application of irradiation with radionuclides is the treatment of food products (⟶ Radiation Chemistry – Regulations on the Irradiation of Foods and Feeds, ⟶ Radiation Chemistry – Industrial Irradiation of Foods and Feeds). In 1990, 29 gamma facilities were in service in 22 countries for this purpose [108]. Projections for the early 1990s are that 55 units will be in operation. Irradiation of food has been heavily promoted by the FAO, IAEA, and WHO. In 1980, an expert committee concluded that food irradiated up to an overall absorbed dose of 10 kGy, presents no toxicological hazard and introduces no special nutritional or microbiological problems [109].

The goals of food irradiation are improved shelf life, sterilization, decontamination, disinfestation, insect control, and delayed ripening.

Other applications of irradiation sterilization include the following:

Packaging Industry. Packaging for beverages, milk products, pharmaceuticals, or cosmetics must be sterile in order to prevent contamination of the products.

Pharmaceutical and Cosmetics Industries. Irradiation can be employed to sterilize either the raw material or the end product.

Sludge and Wastewater Treatment. Clarifier sludge containing no heavy metals or toxic chemicals can be utilized as an organic fertilizer, provided no pathogens are present. This can be ensured by irradiation. A few plants using this technology have already been in operation in various countries. Studies are now being performed to determine to what extent wastes generated in airports, seaports, and hospitals can be irradiated before downstream treatment.

Treatment of Wood. The purpose of irradiating wood is to kill harmful organisms that might cause major damage if exported to other regions.

5.1.5.2. *Modification of Plastics (see also ⟶ Radiation Chemistry)*

The following reactions can be initiated by irradiation of plastics:

1. Cross-linking
2. Degradation
3. Grafting

These reactions are generally accomplished by irradiating products in electron accelerators and have led to a number of improvements in materials as well as new applications for plastics. *Cross-linking* improves the mechanical, chemical, and thermal properties of a plastic. Areas of application include, e.g., the cable industry, the manufacturing of pipes (plastic pipes for floor heating and hot water supply systems), and shrinkable products. *Degradation* (reduction in molecular mass) is employed in order to produce low-molecular-mass polytetrafluoroethylene (for use as a lubricant or dye additive) and polypropylene (as a seeding agent for crystallization of injection-molded articles). This technology is occasionally used in industry for the grafting of plastics to fabricate membranes or fibers with well-defined properties.

These technologies are being adapted to permit irradiation in gamma facilities because the greater penetration of ϒ-rays enables irradiation even of large plastic components at low cost. The results achieved with an electron accelerator cannot be simply extended to the gamma facility, because the latter features a dose rate several orders of magnitude lower. The radicals produced in the irradiated products prefer to react with oxygen. If the irradiation occurs in low dose rates, atmospheric oxygen diffuses into the irradiated product (paricularly at and near the surface) and other processes, e.g., oxidation can occur. Plastic components irradiated in a gamma system therefore still require testing.

5.1.6. *Wear Measurement with Radionuclides [110]*

The use of radionuclides for wear measurement is an important tool for improving tribological systems. This technique offers the following advantages:

1. Continuous measurement of wear versus time on machine parts in service
2. Correlation of wear performance with operating conditions
3. High sensitivity
4. Problem-oriented selection and definition of wear regions for study
5. No interference with operating conditions or with the material of the machine part under study
6. Cost savings
7. Time savings over conventional methods

Continuous wear measurement with radionuclides comprises two methods; the differential method and the concentration method. Each permits continuous observation of the progress of wear without having to stop or disassemble the machine. The time dependence of wear is determined from the radiation emitted by the radioactive material.

Differential Method [111], [112]. In the differential or thin-layer differential method, the condition of a wearing part is ascertained by activating a thin layer of the part's surface and measuring its residual activity. During the trial, a probe comprising a detector and an integrated amplifier monitors a tribological system through the housing wall, measuring the decline in radioactivity remaining on the part and thus the increase in wear. The sensitivity of this form of wear measurement is ca. 10 nm/h.

Thin-layer activation of machine parts is performed with charged particles (e.g., protons, deuterons, α particles) in an accelerator such as a cyclotron [113], [114].

The thin-layer activation technique yields regions of constant activity between 20 and 500 μm. The total activity of the machine part is ca. 370 kBq.

Solving a tribological problem entails converting tribological data (material composition, sensitivity required, test duration, etc.) into activation data (type of bombarding particle, irradiation angle and time, etc.).

Generally, sodium iodide detectors are employed with electronic output coupling to obtain high counting rates. The count pulses can be recorded and classified and converted to wear values (e.g., in nm or μg) with a wear meter; this instrument is easy to operate without special knowledge of nuclear physics and radiometry.

These measurement data must be converted to tribological data, which requires adequate knowledge not only of radionuclide techniques but also of tribology.

Concentration Method [112], [114]. The concentration method involves measuring wear by detecting the radioactivity of the activated wear particles carried away by a medium. The transporting medium can be, for example, a lubricant or a fuel that flushes the wear particles from the activated part, such as oil in an internal-combustion engine. The medium and wear particles are conveyed to a highly sensitive probe, which measures the activity concentration. The wear probe determines the numerical activity values. From this, the volume of carrier medium, the activity per unit mass, and the half-life of the radionuclide can also be calculated. The instrument calculates the wear values and outputs them, in, for example, nm or μg. The concentration method again requires a radioactivity of ca. 370 kBq in the machine part.

The transport medium can be brought to the measuring probe either as discrete samples or in a continuous fashion. With gamma spectrometry, several wear components can be measured simultaneously [115]. Sodium iodide detectors are advantageously used in the concentration method as well.

The concentration method offers higher accuracy and sensitivity than the differential technique (in the nm/h range in typical applications).

If filters are located in the circulation loop of the transport medium, the continuous-flow method can be supplemented by a filter probe. Such a wear measuring system thus determines wear particles in the filter, wear particles not retained on the filter, and total wear.

Uses [116-120]. Radionuclide methods of wear measurement can be used to analyze the wear behavior of tribological systems. The results (wear as a function of the operating conditions of the machine or system under study) can aid in optimizing material pairs, design (macroscopic and microscopic form of wearing partners), lubrication, and operation in tribological systems. Another important aspect is the measurement of the anti-wear actions of various lubricants in order to optimize tribological systems. Radionuclide-based techniques can also be used for corrosion measurements.

Radionuclide methods are especially suitable for wear analysis on *internal-combustion engine parts.* Parts that have been examined in the running motor include: piston ring, cylinder liner, piston, piston ring groove, main bearing, connecting-rod bearing, crankshaft, valve seat, valve seat rings, valve stem guide, rocker arm, camshaft, cam follower, and chain wheel.

Other suitable applications include *drive-train technology* (wear of gears, roller bearings, free-wheels, steering joints), *railroad engineering* (wheel and rail wear), *pumps* (cavitation at pump impellers), and *machining* (tool wear). Radionuclides find use for wear measurement chiefly in research and development departments of automobile producers and in research institutes.

The highly sensitive continuous techniques described here yield on-line results. Extensive new information can thus be obtained on the wear behavior of tribosystems and the anti-wear performance of lubricants and additives. Only these wear measurement methods allow practical wear problems to be solved more quickly and completely; only with these can significant advances be achieved in wear research.

5.1.7. *Other Industrial Applications of Radionuclides [121], [122]*

In the *iron* and *steel industries*, radioisotopes are used for level gauges to control the feed of blast furnaces. Radioactive sources at various depths in the refractory brick walls of blast furnaces monitor the rate of refractory disappearance.

In the *aluminum industry*, the flow of material in the dissolution of bauxite in aqueous sodium hydroxide can be monitored by ^{32}P or ^{59}Fe labeling of particles. During the dehydration stage in the production of alumina, residence times in the fluidized-bed reactor can be measured by similar labeling. The movement of molten aluminum in the electrolytic reduction cells can be followed with titanium capsules containing a radioactive isotope.

In the manufacture of finished *basic metal products*, autoradiography gives an image of the distribution of various radioelements and of the crystallization process in metals.

Radioisotope Applications in Dairy Technology. These applications include automation gauges with ^{90}Sr or ^{137}Cs for filling containers with predetermined volumes of milk, cheese, etc. Also, milk fermentation and cheese ripening have been followed by ^{14}C- and ^{35}S-labeling of saccharides and methionine, respectively. Neutron activation allows the detection of several antibiotics in milk, which may have been administered to cattle, while dosimetry of milk and dairy products can detect harmful radionuclides originating from fall-out or nuclear events such as the Chernobyl accident.

Applications of Radionuclides in the Chemical Industry. The main *applications of* γ *radiation* are described in Sections Uses of Irradiation and Wear Measurement with Radionuclides and in ⟶ Radiation Chemistry.

In addition, radionuclides are used in the disinfection of solid waste, sewage sludge, and polluted water.

Near Munich, a plant treated 1000 m^3 of wastewater per day with ^{60}Co γ radiation for destroying disease-carrying microorganisms. Disinfected water was subsequently applied to farmland.

Applications of *sealed sources* are based mainly on the exponential attenuation of γ-rays. When the density and composition of the medium to be investigated remain constant, its thickness can be determined (e.g., coke deposition in cracking furnaces, or corrosion of pipes). When the distance between the sealed emitter and the detector is constant, the density or composition of the medium can be determined and monitored continuously. For instance, the height of a foam layer in a chemical reactor, the efficiency of a demister, or the amount of carry-over of liquid droplets can be determined as a function of reactor conditions. For packed columns, the presence of voids in catalyst beds can be identified. Density profiling of distillation columns may be carried out routinely for determining faults such as flooding, foaming, or missing or collapsed trays. *Gamma ray back-scattering* can be used for thickness and coating measurements or for detecting build-up of deposits inside reactor walls, boilers, or condenser water boxes.

Neutron source applications include moisture measurements, elemental analysis, and determination of the position of interfaces between immiscible liquids. The best sources for fast neutrons are ^{241}Am – Be.

For *radioactive tracer applications*, the tracer can be in a chemically identical form to the material under study (intrinsic isotopic label), or in a chemically different one: either an isotope of an element present in the material is used or a foreign radionuclide (extrinsic label). The most important tracer application is on-line flow rate measurement for liquids or gases. It is possible to precisely calibrate flow meters, to identify leakages, and to quantify leakage rates. For improved radiological safety, short-lived radionuclides from radioisotope generators are preferably used, e.g., the ^{137}Cs – ^{137m}Ba generator (the half-life of ^{137m}Ba is 2.6 min). Mean residence times and residence-time distributions of materials in full-scale process vessels can be measured. For leak detection in gaseous separation systems, ^{41}Ar can be used; for similar purposes in liquid systems ^{82}Br in a suitable chemical form can be employed.

Gross liquid droplet entrainment can be monitored by γ-ray absorption techniques. For detecting *minor carry-over* (the elimination of which considerably improves the operating conditions), tracer techniques can be used; thus in distillation columns, droplet entrainment is quantitatively measured by injecting a nonvolatile radiolabeled component (e.g., manganese-56 naphthenate) into the feed of the column and by assaying its content in the gaseous overheads stream.

Activable tracers (⟶) can be used when radioactive tracers are not allowed because of safety reasons, or because a radioactive label would interfere with existing radioactive gauges. Unlike other methods which can detect minute amounts of materials (such as dyes, fluorescent compounds, or bacteria), activable elements hardly contaminate the materials and are innocuous before activation; only a small sample is activated. The only drawback is the necessity of using a nuclear reactor.

Some activated elements have a very low limit of detection: it is 10 pg for Eu and Dy, 50 pg for In and Mn, 100 pg for V, Au, Lu, Ho, Ir, 1 ng for I, and 5 ng for Co. For instance, if a leak is suspected in a factory drainage system, any or each of the above elements can be introduced in a suitable chemical form (chelates for the metallic cations, iodide anion for

iodine) into possible sites, and after a certain time samples are taken, activated in a nuclear reactor, and monitored with a multichannel γ analyzer. Neutron activation analysis can also be employed for determining the efficiency of full scale mixing units such as polymer chip blenders or of grinding devices.

5.2. Uses in Hydrology

There are three main techniques for using isotopes and radiation in hydrological studies: (1) Use of natural isotopes; (2) of artificial isotope tracers; and (3) of sealed radiation sources. Table 6 shows some natural radioisotopes of hydrological use.

Table 6. Natural radioisotopes of hydrological use [123]

Nuclide	Half-life	Residence time	Hydrogeological applications
^{3}H	12.3 a	1 – 40	travel time
^{14}C	5770 a	500 – 50 000	origin of carbon compounds
^{36}Cl	3.10^{5} a	$5\times10^{4} – 2\times10^{6}$	infiltration rates after the 1960s
^{222}Rn	3.8 a		presence of groundwater in subsurface water
^{226}Ra	1622 a		tracing water origin

In most cases, radioactivity measurements are simple to perform and can be conducted using portable instruments, unlike measurements of isotopic abundance for stable isotopes (⟶ Isotopes, Natural – Isotope Effects in Nature).

5.2.1. *Snow and Ice*

For dating ice in glaciers, natural radioisotopes with a medium half-life such as tritium or ^{210}Pb can be used when the ice is up to 100 years old; for ice of 100 – 500 years of age, ^{32}Si is suitable and for even older ice only ^{14}C is applicable. Tritium has the drawback that its concentration on earth has increased considerably after 1952, and that contamination with atmospheric water vapor must be completely avoided. ^{14}C also needs exclusion of contact with the atmosphere, and requires correction factors.

Since about 80 % of all land fresh water on earth exists as ice and snow (most of which is concentrated in the polar regions and in glaciers), assessment of global water balance and turnover must consider this ice – snow cover. The density of this cover varies with depth and various local conditions. Nuclear techniques for measuring this *density* use the exponential attenuation of γ radiation emitted by artificial radioisotopes in sealed radiation sources (^{60}Co, ^{90}Sr, or ^{137}Cs) each with characteristic γ energies:

$$\ln(I_0/I) = \mu d$$

where I_0 and I are the intensities of incident and attenuated γ radiations, μ is the linear absorption coefficient of water for the corresponding γ energy, and d is the distance from the γ emitter which is located at a certain depth in the ice pack to the detector. A bore is drilled into the ice thus securing a place for the γ-ray source. The amount of snow deposited in winter time is measured using a ^{60}Co source which is placed on the soil in summer. A detector is placed at a known distance (a few metres) above the ground. During the winter the amount of snow (or snow-containing water at temperatures around 0 °C) between the γ-ray source and the detector can be determined. Similar devices have been used for measuring soil moisture.

5.2.2. *Transient Surface Water*

For measuring the flow rate of a water course, a tracer can be injected at some point and its concentration measured at a point sufficiently far downstream (0.2 – 2 km) for complete lateral mixing to have occurred. The point should not be too close as then mixing would be incomplete, and not too far otherwise excessive dispersion and adsorption on the river bed occurs. Unlike nonisotope tracers (such as dichromate, chloride, sulfate, or nitrite anions, and rhodamine or other dyes which may undergo chemical reactions within the river, especially in the presence of pollutants), isotopic methods are usually safe from such sources of error. Tritiated water is also error-free with respect to adsorption on the river bed or suspended particles, but measurement of the emitted radiation requires more sophisticated equipment than that required for γ emitters, and its half-life is longer than that of radioisotopes such as ^{24}Na, ^{82}Br, ^{32}P, ^{123}I, ^{131}I, and ^{198}Au. Therefore, ease of measurement, accuracy of results, and safety considerations should always be taken into account when selecting the tracer.

There are two main techniques for such measurements:

Constant-Rate Injection Method. Here a tracer with a certain concentration c_0 is injected at a constant rate q into a watercourse with a flow rate Q. By measuring the concentration c after mixing is complete (when c becomes constant downstream) the flow rate can be determined as :

$$Q = q\,c_0/c$$

.

Integration Method. This method makes use of an instantaneous injection of a tracer with activity A_0 into a watercourse. After the mixing distance, the passage of the radioactive cloud is measured at a constant rate over a period, and the mean concentration C is obtained by integration; hence

$$Q = A_0/Ct = A_0V/At,$$

where A is the activity of the sample with volume V.

Most measurements are performed with ^{82}Br which has a large permissible dose in drinking water, or with ^{198}Au, and the activity is determined with scintillation detectors [124], [125].

When using *tritiated water* (HTO) the integration technique is preferred. An ampoule with HTO is crushed under water to avoid contamination: activities of about 150 – 200 GBq must be used for a flow rate of 100 m^3/s. For flow rates larger than 200 m^3/s and velocities in excess of 5 m/s, only radioactive tracers give reliable results [126].

In order to measure transit times for direct surface run-off in hydrological basins, similar techniques can be applied by injecting instantaneously a radiotracer (either as anions such as $^{82}Br^-$ or $^{131}I^-$, or chelates such as $^{51}Cr^{3+}$ – EDTA).

5.2.3. *Lakes and Reservoirs*

The dynamics of water flow, i.e., horizontal or vertical circulation, can be estimated either by using natural isotopic tracers (stable isotopes HDO or $H_2^{18}O$, or radioisotopes HTO, $^{14}CO_2$, ^{222}Rn), or artificial radionuclides. These are either anions such as $^{82}Br^-$, $^{131}I^-$, $^{60}Co(CN)_6^{3-}$, $^{51}CrO_4^{2-}$, $^{198}AuCl_4^-$, or EDTA chelates of transition-metal cations such as ^{46}Sc, ^{51}Cr, or ^{140}La. The radioactivity measurement is performed at various distances, depths, and times from the point where the artificial radiotracer was introduced.

Lake Constance was shown, by means of tritium monitoring, to undergo complete mixing in late winter and spring, whereas in summer its water is layered, these studies were complemented by the monitoring of ^{222}Rn in equilibrium with ^{226}Ra [127].

Leakage of water from reservoirs can be monitored by natural tracers or artificial radionuclides as tracers such as the anions $^{51}CrO_4^{2-}$, $^{82}Br^-$, $^{131}I^-$ or $^{198}AuCl_4^-$. If a higher sensitivity of measurement is necessary, the chelate ^{51}Cr – EDTA can be used which is retained on percolation through active carbon. This facilitates the measurement of very dilute solutions, so that 0.4 GBq of ^{51}Cr can be used for labeling 1×10^6 m^3 water [128].

Vertical or horizontal movements of water in the soil can be traced by injecting tritiated water when little interference of adsorption by soil constituents arises.

When, owing to government regulatory restrictions or to proximity of public or private drinking water supplies, it is not possible to use radioisotopes for tracing the movement of groundwater, neutron activation analysis with a nontoxic neutron-activable element can alternatively be used. The most suitable is ^{79}Br (50 % natural abundance in natural bromine) as bromide anion. By the nuclear reaction (in a nuclear reactor) ^{79}Br (n, γ) ^{80}Br, ^{80}Br with a half-life of 17.6 min is formed whose detection limit is below 1 ppb; in fresh water the typical background concentration is 10 – 100 ppb. For deep groundwater,iodide is the most suitable acturable anion, having very low background concentrations.

For localization of leakage sites in large water reservoirs, a radioactive tracer such as ^{198}Au or ^{113m}In (from a ^{113}Sn – ^{113m}In radioisotope generator) is introduced in the form of a ribbon-like cloud at a constant distance from the bottom of the reservoir (activity about 10 MBq per meter length of the investigated section). When a leak is present, the vertical component of the water flow draws radioactivity to the bottom where it is adsorbed on the bottom surface. After a certain time (e.g., 1 h), an assembly of scintillation detectors with a buoy marker is swept over the bottom; the radioactivity profile over the strip then allows to detect the sites of leaks [129-131].

Ground water in aquifers and underground water can be monitored using natural stable isotopes (^2H and ^{18}O) or radioisotopes (HTO and $^{14}CO_2$). ^{14}C needs corrections for the dilution. Measurements of the radioactive $^{14}CO_2$ content in water trace the contact of the latter with atmosphere, because limestone is practically ^{14}C-free owing to its very old age [132].

By means of natural isotopes (both stable, and radioactive, namely ^3H and ^{14}C) the age of deep aquifers in Western Roumania could be established to be around 30 000 a; in South-East Roumania (between the Danube and the Black Sea), isochronic curves revealed the important part played by recent tectonics, which give rise to conditions for an active, upward exchange of water from deep aquifers [133-136].

5.2.4. *Sediments in Hydrology*

Sediment Movement in Rivers and Estuaries. For labeling sediment, it can be mixed either with a radioactive glass (containing Ta, Sc, La, Ir, or Au, and activated in a nuclear reactor) crushed to a satisfactory particle size, or with a radioisotope-charged exchange resin. Alternatively, sand that is surface-labeled with a radioisotope can be applied. Sand used for the study of sediment movement in Botany Bay (Australia) was surface-labeled with $^{51}Cr_2O_3$ by immersing it under stirring in an ammonium [^{51}Cr]dichromate solution which was evaporated to dryness and then heated at 350 °C for 30 min [137]. Other radioisotopes frequently used for sand labeling are ^{46}Sc, ^{192}Ir, and ^{198}Au [138].

When building dams, it is important to estimate future sedimentation conditions. By γ-ray scattering, the density of sediments can be measured at the bottom of lakes and reservoirs to depths of 6 – 12 m (less for sandy sediments), and to water depths of about 100 m. In Lake Mead (United States), a sealed ^{60}Co source was used for the sediment density measurement.

Harbor construction and enlargement works need to choose dumping areas for dredged material, i.e., to take into account the global movement of sediments. The labeled particles representative of natural sediments are mixed with dredged materials [139].

Suspended sediment flows much faster than the sediment deposited along the bed of the river. Conventional determination methods (filtration or centrifugation, followed by weighing the solids) perturb the flow and cannot be performed continuously. By using a sealed γ source such as ^{109}Cd (γ energy: 22 keV) or ^{241}Am (60 keV) the attentuation caused by the suspended particles can be monitored continuously.

5.3. Uses in Agriculture

The main elements of agricultural interest are carbon, hydrogen, oxygen, nitrogen, and phosphorus. Radiotracers for some of these elements are ^{14}C, ^{3}H, and ^{32}P. Oxygen and nitrogen radioisotopes are very short-lived, therefore stable oxygen (^{18}O) and nitrogen (^{15}N) isotopes are preferred. From the metallic elements, ^{22}Na and ^{45}Ca are also useful in agrochemical studies, but ^{42}K and ^{43}K are too short-lived, and ^{40}K too long-lived (specific activity too low), so that sometimes ^{86}Rb is used as a foreign (extrinsic, or nonisotopic) tracer for potassium.

For determining the ion exchange capacity of soils, $^{36}Cl^{-}$, $^{45}Ca^{2+}$, $^{23}Na^{+}$, or other radioactive cations or anions are used. Weighed amounts of soil are treated with known volumes of radionuclide solutions with calibrated concentration, and after separation by centrifugation, the radioactivity of the supernatant is measured for determining how much radionuclide has been removed by adsorption onto the soil.

The adsorption and movement of phosphate anions (from apatite, superphosphate, or other phosphatic fertilizers) differ markedly from those of other anions. ^{32}P-Labeled superphosphate has been used for determining the optimum depth and timing for applying the phosphate fertilizer as a function of the soil pH [140]. Isotope dilution analysis can be used for determining the amount of phosphate and its adsorption by the soil, using ^{32}P or ^{33}P. The uptake of sulfate and phosphate anions can be easily traced using $^{35}SO_4^{2-}$ and $^{32}PO_4^{3-}$, respectively.

Microelements in the soil can be traced by means of corresponding radionuclides, such as ^{54}Mn, ^{58}Co, ^{60}Co, ^{59}Fe, ^{64}Cu, and ^{65}Zn [141].

Studies on photosynthesis use $^{14}CO_2$ in a confined atmosphere and blue-green algae, cyanobacteria, or halobacteria, and give information on the fate of energy transfer from photons at various extremely short times in the photosynthetic chain.

Similarly, biosyntheses of terpenoids, vitamins, antibiotics, and other phytosynthetic products can be elucidated by using isotopically-labeled precursors (preferably ^{14}C and ^{3}H).

Deforestation occurs as the world population increases, claiming agricultural areas and wood for burning or for building homes. The conversion of forested areas into agricultural land is as old as civilization, but studies on the formation of fertile soil from plant residues, as well as of soil humic constituents, are fairly recent. With ^{14}C-labeled plant material it was possible to monitor the mineralization to humin and humic or fulvic acids in the presence of mycoflora (cellulolytic fungi) [142], [143].

5.4. Uses in Environmental Studies

Agrochemicals, e.g., fertilizers, pesticides (insecticides, herbicides, fungicides, rodenticides, etc.) can pose serious threats to the environment when they are not used rationally. In many countries, it is mandatory to study the fate of agrochemicals when they are registered at a government regulatory agency (e.g., EPA in the United States).

Isotopically-labeled compounds (in many cases the tracer is ^{14}C) are used for following the fate of agrochemicals and of their chemical or biochemical degradation products. The following studies are normally carried out: degradation (e.g., hydrolysis, biodegradability; photodegradation in water,on plants,or on soil);metabolism (aerobic or anaerobic, in soil or water); mobility (leaching, adsorption in soil, desorption, remanence in ground or underground water); and accumulation (e.g., in members of ecological chains such as insects, fish, birds, and mammals).

Radioisotopes for Studying Pesticide Residue Problems. Radioisotope-labeled pesticides are used for studying the distribution, metabolism, and excretion of pesticides administered todomestic animals. Thus, ^{32}P-labeled organophosphorus pesticides used against internal and external cattle parasites (grubs inside the bodies of cattle, lice, ticks, horn flies, and other pests) served to trace the distribution of the pesticides and their metabolism in meat and milk and the excretion via urine and feces [144].

The metabolism and retention of chlorinated insecticides such as aldrin, dieldrin, and related products were studied by means of ^{14}C- and ^{36}Cl-labeled compounds, and contributed to evidencing the accumulation in fatty tissues of such compounds; the legislation restricting the use of such chlorinated products was a direct result of such studies, after similar results had been obtained with DDT and γ-hexachlorocyclohexane.

Isotope dilution methods can be used advantageously for the assay of pesticides in the environment. Higher sensitivity is

attained by double isotope derivative assay methods.

Application of pesticides in larger amounts than actually needed results in unnecessary spending of resources and increased amounts of residues causing undesirable ecological effects. Knowledge of the behavior and biology of the pest in its ecological context, and of the effect of the pesticide, are essential for correct application. Radioisotopes play an indispensable role in the research for less persistent and more potent pesticides, as well as for their efficient application.

5.5. Uses in Recombinant DNA Research (see also ⟶ Genetic Engineering; ⟶ Nucleic Acids)

The revolution in modern biotechnology and in genetic engineering is mainly due to recombinant DNA techniques, which in turn would not have been possible without the high sensitivity of radioactive methods.

Radioactive compounds have been essential in recombinant DNA work owing to their high sensitivity of detection: single-copy human genes can be detected in as little as 5 µg total genomic DNA, i.e., the limit is 1 – 10 pg.

Most studies have been made with ^{32}Plabeled compounds for obtaining nucleic acid precoursors with high specific activity (most commonly α-^{32}P-labeled deoxynucleotide triphosphates dATP or dCTP) [145].

In some cases, it may be advantageous to use ^{35}S or ^{3}H for labeling nucleotide analogues, especially when longer half-life and radiations of lower energy are desirable. In some applications requiring the use of labeled proteins, ^{125}I-labeling is used.

There are two main methods for sequencing and mapping DNA macromolecules, both published in 1977, the *Maxam and Gilbert* and the *Sanger method* (⟶ Genetic Engineering).

Maxam and Gilbert's specific chemical cleavage starts with a DNA which is labeled at the phosphorus atom bound to the 5′-hydroxyl terminus in the presence of polynucleotide kinase. This labeled DNA is then randomly cleaved at one of the four nucleotides: *purines* are methylated by dimethyl sulfate (guanine at N-7, and adenine at N-3), then heating at neutral pH causes breakage of the chain at guanine, whereas heating in dilute acidic solution breaks the polynucleotide chain both at adenine and guanine; for *pyrimidine-specific* cleavage, treatment with hydrazine followed by cleavage with piperidine leads to cleavage at cytosine and thymine, while hydrazinolysis in the presence of 2 N NaCl leads to breakage at cytosine only. Each of these reactions is carried out separately and the four specifically cleaved reaction mixtures are then separated on four lanes by downward electrophoresis on a polyacrylamide gel which resolves DNA molecules differing in their lengths. The resulting fragments are separated on the basis of their size, the shortest chains have the highest electrophoretic mobility. Because the starting DNA fragment is radioactively labeled at the 5′ end the fragment on the gel can be detected by autoradiography. The sequence can be read directly from bottom to top, according to the G; A + G; C + T; and C lanes [146].

Sanger's Dideoxy Sequencing Method (controlled interruption of replication). In this method new radioactive DNA strands of different lengths are synthesized on the DNA template, all strands ending at one of the four different bases. Incubation involves five components: the DNA template (usually single-stranded, produced via bacteriophage m 13), a primer (short single-stranded polynucleotide obtained from a restriction enzyme or chemical synthesis), DNA polymerase (*E. coli* DNA Polymerase I — large fragment or Klenow fragment) ⟶ Enzymes – Klenow Enzyme, all four nucleoside triphosphates (one or several of which are radioactive, usually the deoxyadenosine triphosphate), and one each (for each lane) of the four chain terminators (usually 2′,3′-dideoxynucleotide triphosphates ddATP, ddGTP, ddCTP, ddTTP). The primer is annealed to the template which leads to the elongation of the primer with nonradioactive nucleotides (dGTP, dCTP, dTTP) and with the radioactive derivative of dATP which is usually [α-^{32}P]dATP [2′-deoxyadenosine-5′-[α-^{32}P]-triphosphate (**1**)] or [^{35}S]dATP αS [2′-deoxyadenosine-5′- α[^{35}S]thio-triphosphate (**2**)].

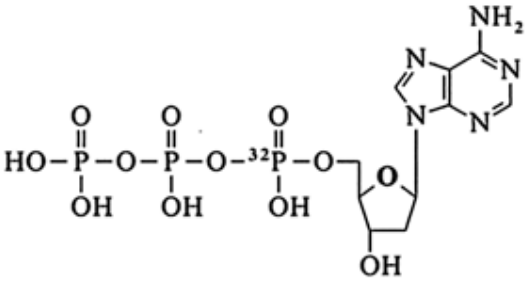

1

[α-^{32}P]dATP

2′-Deoxyadenosine-5′-[α-^{32}P]-triphosphate

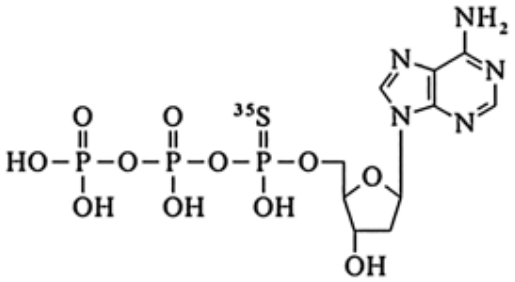

2

[^{35}S]dATPαS

2′-Deoxyadenosine-5′-α[^{35}S]thio-triphosphate

When a chain terminator is incorporated into the chain no phosphodiester can be formed since the chain has no 3′-hydroxy group and chain growth stops. If the concentration of dideoxynucleotides is limiting, then chain terminations will occur at every nucleotide so that all possible chain lengths will be obtained. Electrophoresis on polyacrylamide sequencing gels leads to clean separation of fragments on the basis of increasing size on the four lanes corresponding to the four types of 3′-end bases. Detection of these fragments is easily accomplished by autoradiography, then the sequence can be read directly from the four lanes, starting from the bottom [147], [148].

Sanger's dideoxy sequencing method has been continuously improved and offers many advantages. Accordingly, it has been the preferred method for large sequencing projects and for automation [149]. The Maxam – Gilbert method needs two separate steps, DNA labeling is tedious and requires the presence of suitable restriction sites, and, moreover, a systematic primer walking strategy cannot be applied.

The typical shelf life of ^{32}P-labeled nucleotides is about one month, whereas ^{35}S-labeled nucleotide analogues have shelf lives of at least three months. In addition, the weaker β emission of ^{35}S allows sharper autoradiograms (important in sequencing large DNA fragments) and obviates the need for lead shielding. Moreover, DNA labeled with ^{35}S is more stable than DNA labeled with ^{32}P, probably because of the phosphorothioate linkage which is resistant to many phosphodiesterases and nucleases, and because there is less radiation damage by the weaker β radiation of ^{35}S [150], [151].

Hybridization Probes (⟶ Nucleic Acids – Hybridization Techniques for Nucleic Acid Detection). Radioactively-labeled nucleic acids can function as hybridization probes for obtaining information on specific nucleic acid sequences which are present in a large mixture of unrelated sequences. This means in practice the identification and characterization of specific genes and DNA fragments, and the monitorization of gene expression.

Although the formation of hybrids can be also monitored optically or by fractionation on columns, radioactive methods allow the use of very small amounts of materials, and autoradiography is particularly convenient for detecting hybrids immobilized on solid supports such as nitrocellulose.

A restriction DNA fragment containing a specific base sequence can be identified by hybridizing it with a labeled complementary DNA strand. After separating the restriction fragments by gel electrophoresis, denaturation to form single-stranded DNA, and transfer to a nitrocellulose sheet, a particular sequence can be identified by formation of a hybrid with a ^{32}P-labeled single-stranded DNA probe (⟶ Genetic Engineering – Nucleic Acid Hybridization).

5.6. Uses in Biochemistry (for uses in medicine see Chap. Production of Artificial Radionuclides)

A few landmarks on radioisotope uses in biochemistry since the 1950 are given in the following.

In 1952, A. HERSHEY and M. CHASE [152] labeled T 2 bacteriophage DNA with ^{32}P, and its protein coat with ^{35}S. A sample of *Escherichia coli* infected with the labeled virus (T 2 phage) was separated by high-speed centrifugation leaving the virus in the supernatant and agglomerated the infected bacteria. It was found that most of the phage DNA appeared in the

infected bacteria while most of the phage protein appeared in the supernatant. The infected bacteria conserved their capacity to produce progeny virus; little parental ^{35}S but 30 % of parental ^{32}P was transferred to the progeny. In conclusion, the DNA and not the protein is the genetic information carrying part, confirming what AVERY, MACLEOD, and MCCARTY had inferred eight years earlier.

In 1958, A. KORNBERG used ^{14}C-labeled deoxythymidine in his discovery of DNA polymerase I which catalyzes step-by-step addition of deoxyribonucleotide units to the 3′-end of a primer chain with a free 3′-hydroxy group. These units are complementary to a template DNA molecule [153].

In 1961, BRENNER, JACOB, and MESELSON used both stable (^{15}N, ^{13}C) and radioactive isotopes (^{14}C, ^{32}P, ^{35}S) to prove that ribosomes synthesize at a given time the specific protein that is dictated by the messenger RNA (mRNA) they happen to contain.

Also in 1961, S. SPIEGELMAN used ^{32}P-labeled T 2 phage mRNA and ^{3}H-labeled T 2 DNA for showing that RNA –DNA hybrids are formed when they have complementary sequences [154].

The deciphering of the genetic code by M. NIRENBERG [155] and by H. G. KOHORANA [156] in 1962 – 1966 involved radioactive ^{14}C-labeled amino acids incorporated into proteins by a cell-free protein synthesizing system containing synthetic polynucleotides with known sequences. Such mRNA is synthesized by polynucleotide phosphorylase from given nucleotides.

Radioautography of the DNA of *Drosophila melanogaster* confirmed the existence of very large DNA molecules. A *D. melanogaster* chromosome contains a linear unbranched chain with a molecular mass of 43×10^{9}, i.e., 76 megabase pairs. The incorporation of radioactive labels into the chromosome was carried out over several generations by feeding these fruit flies with radioactive precursors [157].

Protein chemistry, nucleic acid chemistry, immunology, and molecular genetics are highly synergistic; powerful and highly sensitive radioactive methods allow minute amounts of genetic material to be amplified by gene cloning. A gene or cDNA (complementary DNA) can be inserted into an expression vector for infecting a host cell, e.g., an *E. coli* bacterium, and for producing the encoded protein. Alternatively, given a protein, its amino acid sequence can be determined, DNA probes synthesized for hybridization, and a DNA library can be screened by Southern blotting for isolating and cloning the respective gene or cDNA (⟶ Genetic Engineering).

Correspondingly, immunochemical methods can be used, starting with the gene or cDNA via the deduced amino acid sequence and the preparation of synthetic peptides for obtaining an antibody specific for the encoded protein; alternatively, starting from a given protein, a specific antibody can be prepared, and then via Western blotting a DNA library for arriving at the gene or cDNA can be contrived.

Fetal DNA *in utero*, or chromosomal DNA in any human, can be analyzed by restriction enzyme digestion followed by Southern blotting with a ^{32}P-labeled probe for the presence of genetic diseases such as sickle cell anemia, thalassemia, hemophilia, etc.

[Top of Page]

6. Syntheses of Radiolabeled Compounds

Walter Wolf

6.1. Introduction

Radiolabeling is the process whereby a radioactive nuclide is introduced into a chemical compound. Consider a very simple molecule, such as hydrogen (H_2). If every single atom of each hydrogen molecule were tritium (^{3}H) this would be a preparation that is *carrier-free.* While this is theoretically possible, the more likely situation is that a sample of radioactive hydrogen contains a mixture of ^{1}H and ^{3}H. The nonradioactive isotopes are called the *carrier* atoms, and it can be estimated from the *specific activity* of the material how much of each nuclide is present. Specific activities are expressed in activity per unit mass. Thus, carrier-free tritium would have a specific activity of 29.1 Ci/mol (i.e., 3 g), and the mass of tritium gas required to provide 1 Ci of material would be 0.103 mg. Thus, if a sample of tritium has a specific activity of 1 Ci/mg, this means that only 10 % of the hydrogen atoms present are tritium, and 90 % of the hydrogen atoms are ^{1}H (containing, presumably a small proportion of ^{2}H). In practice a much smaller proportion of radioactive atoms is used to radiolabel materials, inasmuch as the radiolabel is usually required only as a tracer — i.e., as a means of detecting the molecular species that contain that atom. The value of the activity – mass relationship can be calculated using the radioactive decay law. The mass (in grams) of a radionuclide that corresponds to 1 mCi is given by:

$$\text{mass (of 1 mCi)} = A_r \cdot t_{1/2} \cdot 8.87 \times 10^{-17}$$

where the half-life ($t_{1/2}$) is expressed in seconds. Radiolabeling with truly carrier-free radionuclides is rarely required, and radionuclides of medium to high specific activities are generally satisfactory. Table 7 lists selected activity – mass relationships of some of the more common elements used in the radiolabeling of organic compounds: This table includes two extremes: ^{14}C, with a $t_{1/2}$ of 5568 a, is probably the longest-lived radionuclide used for radiolabeling purposes, while ^{15}O — whose $t_{1/2}$ is only 2 min (123 s) — is probably one of the shortest of those that can be used for radiolabeling.

Table 7. Selected activity – mass relationships of the most common radionuclides used in the radiolabeling of

organic and biological molecules [a]

Nuclide	$t_{1/2}$	Mass of 1 mCi, mg	Maximum specific activity, mCi/mmol
^{3}H	12.262 a	1.03×10^{-4}	2.91×10^{4}
^{11}C	20.34 min	1.19×10^{-9}	9.24×10^{9}
^{14}C	5568 a	0.224	64.2
^{13}N	9.95 min	6.89×10^{-10}	1.89×10^{10}
^{15}O	123 s	1.64×10^{-10}	9.15×10^{10}
^{18}F	109.7 min	1.05×10^{-8}	1.71×10^{9}
^{32}P	14.22 d	3.49×10^{-6}	9.17×10^{6}
^{35}S	87.2 d	2.33×10^{-5}	1.5×10^{6}
^{123}I	13.3 h	5.22×10^{-7}	2.26×10^{8}
^{125}I	60.2 d	5.77×10^{-5}	2.17×10^{6}
^{131}I	8.05 d	8.08×10^{-6}	1.62×10^{7}

[a] From [158].

Radiolabeling may be performed by substitution or replacement of one of the atoms of a molecule with one of its radioisotopes (*intrinsic labeling*). In other cases, such as in the use of iodine for radiolabeling proteins, the iodine atom is not originally present in the parent molecule, but is added (*extrinsic label*). While intrinsic radiolabeling does not change the chemical nature of the original molecule, extrinsic radiolabeling may do so and it must be checked whether it does or not. The method of radiolabeling may become very critical in determining such chemical changes.

Another problem is where in the molecule the radiolabel will be placed. As an example, consider uracil, one of the key pyrimidines. The ring contains four carbon atoms of which carbon-2 is the easiest to radiolabel, as the 1 – 3 chain is formed from urea during biosynthesis. However, when uracil is catabolized, the first atom that is split off is C-2. To monitor catabolites such as fluoro- β-alanine, it is necessary to radiolabel the carbons at positions 4, 5, or 6.

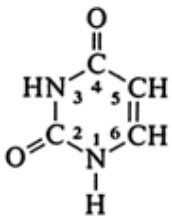

Thus, more than one *radioisomer* exists in this case, and it is important to select the method of radiolabeling that will yield the desired compound with the radionuclide(s) at the desired position. Radiolabeling may be *specific*, i.e., the position of the radionuclide in the molecule is known. It may be *uniform* (U) or *general* (G), which means that a radiolabel is present at all the nonexchangeable positions in the molecule, or it may be *nominal* (n), which means that the location of the radionuclide has not yet been determined. It is very important to follow a uniform system of nomenclature, and to acquire as much information as possible on the position of the radiolabel. The purity of the final product must also be known: *radionuclidic purity* is a measure of the proportion of the desired radionuclide in the total radioactivity present, whereas *radiochemical purity* refers to the chemical identity and purity of the radiolabeled product.

The *stability* of the radiolabeled material is one further consideration that is very critical. Energy is continuously being released within the radiolabeled material, leading to the production of radiolysis products in addition to the decomposition products derived from those molecules in which an atom has decayed. Thus, although tritium has a $t_{1/2}$ of 12.262 a, the shelf life of tritiated compounds is usually measured in weeks (or a few months, at most). It is critical that the user of radiolabeled materials documents the radiochemical purity of the material at the time of use and performs additional purifications if needed.

Many of the labeling techniques (albeit, not the accompanying precautions) described in the following have also been used for the preparation of compounds labeled with stable isotopes of key elements, such as deuterium, ^{13}C, and ^{15}N.

A number of treatises discuss radiolabeling in much greater detail, in books [159-163], in symposia [164-169], in review articles and book chapters [170-174], as well as in a very large number of original papers.

Chemical differences can be observed between compounds labeled with isotopes (see also ⟶ Isotopes, Natural – Chemical Differences). This occurs in all reactions where the mass of the material is key to the rate determining step. Such phenomena are known as *isotope effects* [175]. They are of special importance for hydrogen, whose isotopes have twice (^{2}H) or three times (^{3}H) the mass of the regular hydrogen (^{1}H). Such differences still exist for carbon, where the ^{14}C/

^{12}C mass ratio is 1.16667, and the $^{12}C/^{11}C$ mass ratio is 1.091. It is reduced to 1.03 for iodine ($^{127}I/^{123}I$). Isotope effects need, on a practical basis, be considered important only in reactions involving deuterium and tritium.

6.2. Principles of Radiolabeling

6.2.1. *Introduction*

A number of companies prepare radiolabeled compounds on a routine basis, and investigators should consult their appropriate catalogues to determine what may be available. Most of these companies are also willing to undertake custom syntheses. Whether a radiolabeled compound is prepared "in house" or commercially, certain general principles always apply:

1. The radionuclide of choice must be selected, as well as the position(s) on the molecule to be radiolabeled.
2. The investigator must be aware of how this radiolabled compound will be used, in order to estimate what specific activity is required, how much total activity will be needed, and which atom(s) of the molecule should be radiolabeled.
3. The investigator must also be familiar with all the appropriate safety aspects for handling of the radionuclide. Thus, very different shielding will be required for ^{3}H, a very soft β emitter (E_{av} = 6 keV), than for a positron emitter (511 keV annihilation radiation). In all cases, in particular with biological tritium compounds, absolute care must be taken to avoid ingestion, inhalation, or other internalization of the radiolabeled material.
4. The required equipment for correct (remote and enclosed) handling of radioactive materials must be available, and each investigator must become familiar with such techniques.

6.2.2. *Methods of Radiolabeling*

Methods used for radiolabeling can be divided into chemical and biological methods. *Chemical methods* include: (1) Direct chemical synthesis; (2) exchange labeling, where an existing cold (nonradioactive) atom is replaced by its hot (radioactive) isotope; and (3) recoil labeling, where the primary radioactive atoms that are chemically excited can be used directly. *Biological methods* include both those performed in vivo and in vitro.

The actual synthesis is only the first step. The isolation and purification of the radiolabeled product is sometimes as difficult and time consuming as the original preparation, if not more so. This is because of the amounts of product involved (often in the milligram or microgram range) and of the precautions that need to be taken during the preparation. HPLC, either preparative or analytical, is one of the most widely used purification techniques.

A key consideration in any synthesis of a radiolabeled compound is that the radionuclide should be introduced at the latest stage possible. This is because any "hot" synthesis requires special handling, it must be accomplished within a limited period of time, and low yields obtained may become extremely costly. Time is of particular importance when using some of the very short-lived radionuclides where syntheses must be completed within less than 2 – 4 half-lifes of the radionuclide. For these reasons the synthetic methods used in radiolabeling may differ from those employed on a larger scale and where time is not a primary consideration. As an example, it was required to radiolabel 5-fluoro-uracil (5FU), a key antitumor agent, with ^{18}F. The standard method of synthesis of (cold) 5FU starts with a fluorinated intermediate, and takes place over two days. Such a method is obviously totally impractical for the introduction of the short-lived (109.7 min) ^{18}F into 5FU. The alternative synthesis that was developed allowed the preparation of a clinical dose of ^{18}F-5FU in a 45 % radiochemical yield in 45 min from EOB, including its purification and sterilization for human use [176].

While most conventional syntheses (organic or inorganic) can be adapted to the preparation of radiolabeled compounds the scale of the synthesis may have a significant effect. Syntheses that lead to good yields with gram amounts of starting materials may not scale down well to milligram or microgram levels. Such scaling phenomena are not always well understood, but some of the more common problems are due to changes in heat transfer rates, levels of interfering contaminants, changes in solvents, etc.

Direct syntheses include all those methods of radiolabeling based on standard and well established synthetic methods from organic and inorganic chemistry, where the previously noted constraints of time, scaling, yield, and safe handling must be taken into account. All that is needed is adaptation of a new synthetic pathway to provide the amounts of material available for a radiosynthesis.

Exchange reactions of the type

$$AX^* + BX \rightleftharpoons BX^* + AX$$

allow for the replacement of a given atom by a corresponding radioactive isotope. This technique is most widely used for atoms that form only one single bond, such as the halogens or hydrogen. Yields in such reactions are usually low, but because they only require one step (followed by purification), their time normalized yields may still be acceptable.

Recoil reactions take advantage of the fact that radionuclides, upon production, still have excess (recoil) energy, and that they will react with almost anything present in their environment. If that environment is properly controlled, the radiolabeled intermediates can effectively be used later in a synthesis. Thus, in the production of ^{18}F by the $^{20}Ne(d, {}^{4}He)$ reaction, 0.1 – 1 % cold F_2 is added to the neon gas so that fluorine atoms in F_2 will exchange with the (kinetically) hot $^{18}F^{**}$ produced in the nuclear reaction.

Biological methods make use of whole living systems, tissues removed from such living systems, or isolated enzymes. They are capable of producing and/or transforming very complex molecules quite rapidly, and they permit radionuclides to be introduced at specific positions in a very selective manner. However, these methods are limited to molecules that can

be transformed by enzymatic reactions, and although they are highly selective, yields are usually low.

6.2.3. *Stability of Radiolabeled Compounds*

Another factor to be considered is the stability of the radiolabeled material. Such materials contain, by definition, atoms that decompose, giving rise to both reactive and stable decomposition products. A much larger quantity of reactive products is formed as a consequence of the interaction of radiation with matter: hydroxyl free radicals (OH), hydrogen free radicals (H), and the hydrated electron (e^-_{aq}) result from the interaction of α, β, or γ radiation with water, and free radicals arise from the interaction of radiation with other molecules. While the chemical entities resulting from subsequent reactions with these reactive intermediates may or may not be radiolabeled, they will contribute significantly to an ever-increasing deterioration of the purity of the radialabeled material.

Phosphorus-32, for example, has a $t_{1/2}$ of 14.22 d, which means that the inherent rate of decomposition of a ^{32}P-labeled material is 4.83 %/d. In a product that is seven days old, 32.7 % of the ^{32}P-phosphorylated material will have decomposed. Such impurities become more significant the higher the specific activity. While ^{32}P has a maximum specific activity of 9.17×10^6 Ci/mol, the specific activity of most ^{32}P-labeled materials is in the range 10 –100 mCi/g, which means that only 0.03 % of all phosphorus atoms are radiolabeled, hence only 0.01 % of the total phosphorylated product will have decomposed in one week. However, to this impurity level,the products of the radiolytic reactions must be added. The more chemically sensitive the molecule, or the higher its specific activity, the higher the probability of decomposition. It is therefore important that any laboratory intending to use radiolabeled materials has a system of quality control in place in order to ascertain the purity of the radiolabeled products. A number of methods have been developed for the preservation and purification of radiolabeled compounds [170], [177].

The discussion that follows on specific types of radiolabeling is divided into four sections:

1. ^{14}C, ^{3}H, ^{32}P, and ^{35}S, all pure β emitters
2. Short-lived positron emitters (^{11}C, ^{13}N, and ^{15}O)
3. The halogens, including ^{18}F and the various radio-iodines
4. Radiometals and radiolabeling with other elements

6.3. Radiolabeling with ^{14}C, ^{3}H, ^{32}P, and ^{35}S

^{14}C. A general scheme for the preparation of ^{14}C-labeled materials is presented in Figure 35. The central starting material is $^{14}CO_2$, usually prepared by the addition of sulfuric acid to $Ba^{14}CO_3$. A simple manifold system can be used to handle such radioactive gases and ensure that transfers and reactions occur in a closed environment.

Figure 35. Scheme for the preparation of ^{14}C-labeled materials from $^{14}CO_2$

The yield for each reaction and the reagents reacting with $^{14}CO_2$ are listed above for each of the key intermediates formed. Each of these primary intermediates can then, in turn, be used in a variety of synthetic steps. Some aspects of these reactions are illustrated with the formation of ^{14}C-acetic acid. Actually, there are three radioisomers of ^{14}C-acetic acid: 1-^{14}C-acetic acid, 2-^{14}C-acetic acid, and 1,2-^{14}C-acetic acid, and the following set of reactions illustrates how each these three different compounds can be synthesized.

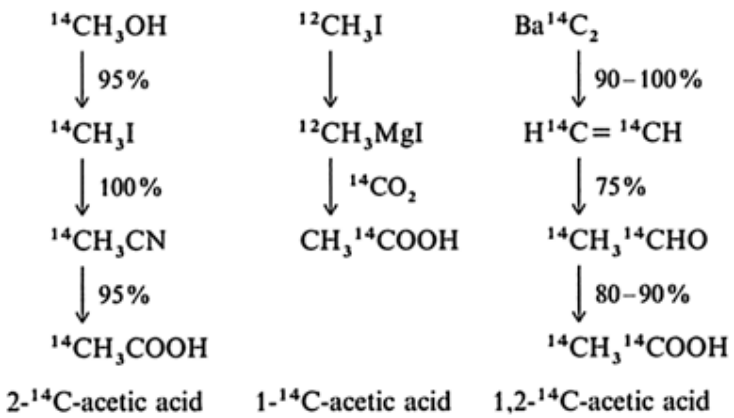

If the purpose of the study is to use ^{14}C-labeled acetic acid to determine the degree of incorporation of its methyl group into an organic or a biological molecule, then either 2-^{14}C-acetic acid or the 1,2-^{14}C-derivative must be used. The first of these is preferred: all its activity is in the methyl group, and any activity detected must come from that function. The doubly labeled acetic acid would be a poorer choice, inasmuch as up to 50 % of the activity in the final product would either be lost or come from some source other than the methyl group. This example illustrates why it is important to understand the purpose of a study before selecting the radiolabel to be used and the method of radiolabeling.

As noted previously, the ultimate starting material for all ^{14}C syntheses is $^{14}CO_2$. Radioactive carbon dioxide can easily be frozen, transferred, and reacted in closed and remotely-controlled manifolds. Because of the long half-life of ^{14}C, radiolabeled compounds containing this radionuclide are widely available and many suppliers are also willing to perform custom syntheses.

Selected carbon-labeled compounds can be and have been prepared by biochemical methods. For example, incubation of ^{14}C-acetic acid with liver tissue leads to ^{14}C-cholesterol as well as its precursors and metabolites, all of which can then be isolated and purified. Radiolabeled alkaloids and other natural products can also be prepared biochemically, and biotechnology is beginning to be used to generate specific novel ^{14}C-radiolabeled materials whose direct synthesis has not yet been achieved.

^{3}H. Tritium is probably the most widely used radionuclide in biochemical studies. Its half-life of 12.262 a makes compounds radiolabeled with this radionuclide readily available from commercial sources and it can be obtained in much higher specific activities than are available with carbon (up to 29 Ci/mol). Chemically, the single bond, between hydrogen and carbon (in organic molecules) makes it relatively easy to achieve radiolabeling, and the very low energy of the β radiation (18 keV E_{max}) makes it very simple to provide shielding from tritiated materials as external emitters. On the other hand, tritiated thymidine — probably one of the most widely used tritiated compounds since it can monitor DNA synthesis — is also one of the most radiotoxic substances when it is internalized, because all its energy is deposited in the DNA of cells at the site of incorporation — which, in humans, would be the gonadal and the hematopoeitic tissues.

Exchange labeling is very widely used for tritiation. Palladium, platinum, and other catalysts can activate tritium (usually furnished as tritiated water) with hydrogen replacement in organic molecules upon heating. Radiolabeling by this method is usually general, but rarely uniform. Radiation-catalyzed exchange reactions (e.g., Wilzbach technique) are also very effective and give high yields, but they have the drawback of requiring very extensive purifications.

Specific tritiations can best be achieved by using chemical reduction methods; e.g., reduction of a halogen compound in the presence of zinc and tritiated water; hydrolysis of a Grignard reagent (ArMgX) by T_2O; reduction of a carbonyl compound with tritiated sodium borohydride or reduction of a halogen compound with tritium gas in the presence of a palladium or platinum catalyst. One caveat: the exact position and specific activity of the final product must be documented, inasmuch as hydrogen migrations may occur during the radiolabeling process.

^{32}P and ^{35}S. Both of these radionuclides are relatively short-lived ($t_{1/2}$ 14.22 d and 87.2 d, respectively), and very few compounds labeled with these radionuclides are commercially available. However, most *radiolabeling with ^{32}P* — obtained either by the activation of ^{31}P or in much higher specific activity by the (n,p) reaction of ^{32}S — is quite simple, inasmuch as most of the compounds studied carry the radiolabel in a phosphate function. Labeling of the terminal phosphate of ATP permits transfer of the radiolabeled function to other phosphates. Other organophosphorus compounds of interest include insecticides, fertilizers, and industrial chemicals.

^{35}S is obtained either by the activation of ^{34}S (low specific activity), or by the (n,p) reaction of ^{35}Cl (high specific activity). Compounds containing either S(–II) (mercaptans) or S(+VI) (sulfates) have been prepared to study drugs and biochemical agents, fungicides, and insecticides. Industrially important sulfur-containing compounds that can be radiolabeled include surface-active agents and rubber additives.

^{11}C, ^{13}N, and ^{15}O. These three radionuclides (together with ^{18}F, see below) have found wide application as nuclides for PET studies. Their half-lifes (20.34 min, 9.95 min, and 123 s, respectively) make syntheses with these radionuclides a very unique and distinct specialty within radiochemistry and radiopharmacy. In practice a synthesis itself may occupy up to 6 half-lifes. Thus, starting with 1 Ci of a ^{11}C-containing reagent, 15 mCi of the original activity will remain after 6 half-lifes (120 min). The main problem is thus to select a synthesis that is fast, even though the absolute yield may be less than in a traditional preparation.

The chemistry used for the preparation of *^{11}C-labeled compounds* is the same as for the preparation of ^{14}C-labeled compounds, except that the time factor must be considered as a key element in the process. In addition, very significant shielding (hence, remote handling) is required. Typical syntheses with ^{11}C use 100 mCi or more of starting material, and most biological and human studies require 1 – 20 mCi per study. Another feature of such a synthesis is that a given preparation is usually limited to a single use, rather than for multiple uses and users as in most other syntheses with radionuclides. A high degree of coordination is also required: the experiment or study in which the ^{11}C compound will be used must be ready to start at the moment the synthesis is completed.

Yet another distinguishing feature is that the method used for preparing the radionuclide must be tailored to the subsequent chemical reaction. Thus, rather than making $^{11}CO_2$ and then generating from it a series of intermediates, as described above for $^{14}CO_2$, the usual procedure is to make the desired intermediate directly in the cyclotron. As an example, irradiation of ^{11}B [a (p,n)reaction] in the presence of NH_3 and K leads directly to $^{11}CN^-$, as does the irradiation of $N_2 – H_2$ mixtures using the ^{14}N (p, α) reaction.

Many compounds radiolabeled with ^{11}C have been prepared, and several are being prepared on a daily basis. These PET radionuclides constitute a unique tool for studying biochemical, physiological, and pathological changes inside living systems using what is fundamentally a noninvasive technique. They have already revealed unique features of processes inside the brain, which segments are activated when we hear, see, or receive other stimuli, and how such physiological

features change with age, disease, and drug treatment.

Syntheses with ^{13}N and ^{15}O are based on the same rules, with the caveat that the time available is much shorter. Thus, the only practical compounds that can be produced with a ^{15}O label are O_2, H_2O, and CO_2, together with their in vivo reaction products (e.g., with hemoglobin).

^{18}F, ^{123}I, and Other Halogens. All halogens have radionuclides that have been used for radiolabeling. Here the discussion is limited to the most widely used cases: ^{18}F and the radioiodines.

Fluorine-18. ^{18}F has a half-life of 109.7 min, and is therefore by far the longest-lived of the PET radionuclides. Taking the six-half-life rule as a guide, studies can be performed with compounds radiolabeled with ^{18}F for more than 600 min (10 h). The time window is thus significantly wider than with any of the other PET radionuclides. ^{18}F-2-fluoro-2-deoxyglucose (^{18}F-FDG) is the most widely used PET radiopharmaceutical, since it makes it possible to measure the metabolic status of various tissues and organs in which that molecule is metabolically trapped [173].

Radiolabeling with ^{18}F is not easy, given the nature of fluorination reactions. As with all halogens, fluorination can be envisioned using either an electrophilic fluorine or a nucleophilic fluoride ion (F^-). Inasmuch as fluorine is the most electronegative element, the possibility of removing an electron to yield a reactive F^+ species has been seriously questioned. In addition, fluorine in aqueous solution is extensively and strongly solvated, and such species must first be transformed into “naked” fluorine before they can be used in electrophilic substitution reactions. Removal of the water is difficult. Hence, even though nucleophilic fluoride is best be made by the (p,n) reaction of ^{18}O in $H_2^{18}O$, production of electrophilic fluorine must be accomplished under strictly anhydrous conditions using reactions such as the (d, α) reaction of ^{20}Ne [166], [173].

Dozens of reactions have been developed for making radiofluorinated species. As an example, ^{18}F-5-fluorouracil has been synthesized by the reaction of ^{18}F-F_2 [produced in the ^{20}Ne (d, α) raction] with uracil, fractional sublimation of the addition product, and purification of the final material. Yields of 45 % have been achieved in 45 min following end-of-synthesis [176]. ^{18}F-FDG is now prepared routinely in many laboratories throughout the world for PET studies; other ^{18}F-labeled compounds are being synthesized and used for measuring receptors, monitoring drug targeting, and, in general, to elucidate the mechanisms involved in the use of fluorinated agents.

Radioiodine was the first radionuclide used for human studies (in 1937), and it acquired a high profile when ^{131}I became readily available from reactors in 1945. Indeed, the first 20 years of application of radionuclides in humans were dominated by ^{131}I, and one of its isotopes, ^{125}I, remains the dominant radionuclide, used in nearly 90 % of all in vitro clinical radionuclide applications. Table 7 lists some of the key radioisotopes of iodine used in the mid 1980s. All share a common and very simple chemistry: iodination of proteins by electrophilic iodine (I^+, iodonium), readily produced by oxidation of iodide ion ($I^- - 2e^- \longrightarrow I^+$). Such an oxidation can be achieved with hydrogen peroxide, chloramine T, or enzymatically using lactoperoxidase. In proteins, iodination occurs at the 3 or 5 position of tyrosine. These positions have increased electron density due to the hydroxyl group on carbon 4, and facilitates electrophilic substitution. Typically, a protein is placed in solution, radioiodine is added (as iodide), and an oxidizing agent is introduced to start the reaction. The final product is then isolated and purified.

Nucleophilic displacement reactions (mostly exchange reactions) are employed when an iodine atom is required at a nonactivated position, such as in hippuran (2-iodohippuric acid). Addition of radioiodide and heating at 170 °C for 30 min readily exchanges the stable ^{127}I for one of its radioactive isotopes (^{123}I, ^{125}I, ^{131}I).

As mentioned above, ^{125}I is used primarily in laboratory studies. Its relatively long half-life (60.2 d) and low photon energies make it very suitable for such purposes. It is readily obtained in nuclear reactors by neutron activation of ^{124}Xe or ^{125}Xe, which then decays to ^{125}I. ^{125}I is not useful for imaging. ^{131}I is obtained mainly from the fission products of ^{235}U, and it is therefore available both in large quantities and at reduced costs. While its high photon energies (364 keV) and short half-life (8.05 d) make it no longer the best radionuclide for human imaging studies (a position now held by ^{99m}Tc), it is still a very valuable and widely used species. ^{123}I on the other hand, is ideally suited for imaging studies: it has a half-life of 13.3 h, and a photonenergy of 159 keV. However, it is much more difficult to make than ^{99m}Tc and hence much more expensive. It is used for special studies in which its chemistry offers unique advantages.

^{59}Fe, ^{111}In, ^{99m}Tc, and Other Radionuclides. Labeling has also been achieved with other elements. Modern nuclear medicine, which is primarily based on the use of products radiolabeled with ^{99m}Tc, has required the development of novel, specialized reactions for dealing with this man-made element, known only since 1937 (see Section Technetium). The reader is referred to specialized literature for further information.

[Top of Page]

7. The Biological Effects of Radiation and Radionuclides

Albrecht M. Kellerer and Jürgen Griebel

The assessment of the health hazards arising from ionizing radiation has undergone a striking change in the course of this century. When RÖNTGEN announced in 1895 that he had discovered a new type of radiation, which he called X-rays, the world reacted with an immense faith in progress. Although harmful side effects of X-rays were soon encountered, no one

thought about hazards or radiation protection. Inflammation similar to that caused by burns was observed on the hands of operators repeatedly testing X-ray pictures and, later, the first cases of radiation-induced skin cancer became known. Only a decade after the discovery of X-rays, an increased frequency of leukemia among radiologists was noticed [178]. This was not regarded as a warning sign but, at best, as an indication of the possible hazards arising from the long-term exposure to high doses of radiation.

The positive expectations that were aroused by this new agent are reflected in the application of X-rays and natural radioactivity (which was discovered just a year later) in medicine within few weeks after the discovery by RÖNTGEN. X-rays were being used extensively in the examination of the human body and first in simple medical diagnoses. Radioactivity was also quickly applied for all kinds of medical therapy and especially radium, with its penetrating γ radiation, was frequently and indiscriminately used at the start of this century in a large number of diseases and for the supposed stimulation of the organism. For instance, the application in the form of radium pillows was common, even in the absence of disease and without medical supervision.

Early fundamental work in radiation biology, e.g., by the young FRIEDRICH DESSAUER [179], showed that higher doses of ionizing radiation destroy living cells. This finding quickly led to the introduction and subsequent improvement of radiation therapy for cancer patients. In spite of the accumulated results on the biological effects of ionizing radiation, the belief in the harmlessness of low doses of radiation persisted for decades.

The fate of hundreds of young women, especially in the United States, who painted the faces of clocks and instruments with luminous paint containing the long-lived, α-emitting ^{226}Ra was one of the first great tragedies resulting from unprotected exposure to radionuclides. Since they had not been warned of possible hazards, the workers sharpened their brushes soaked in luminous paint with their mouths, thus incorporating considerable amounts of activity. Many of them died of bone tumors [180], [181]. A young doctor published his suspicions, that the severe lesions in the jaws of the young women could be caused by radiation. In response, he was called a charlatan by a reader because, as was explained in the reader's letter, it was common knowledge that radiation was dangerous only in massive overdoses over long periods of time. The letter was written by the great scientist Marie Curie, who herself fell victim to radiation exposures later.

These and similar examples illustrate the consequences of the uncritical contact with ionizing radiation which took place in the first half of this century. A fundamental change in assessment of radiation hazards occurred, when the long-term effects of radiation exposure were discovered in the survivors of the atomic bombings of Hiroshima and Nagasaki. The possibility that leukemia might result from a single and relatively small dose of radiation was proposed by LEWIS in the 1950s [182]. He inferred from the increased frequency of leukemia among the survivors of the atomic bombings — about a dozen cases per year — that leukemia could result from somatic mutations, i.e., individual alteration in the DNA of single cells. This finding became the basis for the present philosophy of radiation protection, which holds the principle that unnecessary exposure to radiation should be prevented and unavoidable exposures should be reduced as far as possible.

The recognition that even the smallest doses of radiation can — with correspondingly low probability — cause hereditary defects or cancers and the fact that this possibility was recognized first for radiations, and only later for chemical carcinogens, may have led to the amazing change in the attitude to ionizing radiation. In the past, people had blind faith in the mysterious biopositive effects of radiation, but now the risks, which must indeed be taken seriously, are often overestimated in comparison with other risk factors.

7.1. Physical Radiation Effects (see also ⟶ Nuclear Technology – Radiation Protection and Safety)

Several levels can be distinguished in the effects of ionizing radiation: physical, molecular, cellular, and tissue effects (Fig. 36).

Figure 36. The different levels in the action of ionizing radiation

In spite of considerable differences in the energy transport and interactions with matter (see Chaps. Introduction and Detection of Ionizing Radiation and Radionuclides), all types of ionizing radiation act through the same physical primary processes and, as discussed later, they all act predominantly by damaging the DNA, i.e., the genetic material of the cell. *Indirectly ionizing radiation*, i.e., photons and neutrons, release fast particles in matter. The charged particles are, in turn, ionizing and produce secondary electrons which cause the major part of the ionizations and excitations. In the case of excitation, the excited molecule dissociates into radical or nonradical products or it gives up its energy to the environment without disintegrating. In comparison with ionizations, however, electron excitations are of secondary importance for radiation effects. In the case of *directly ionizing radiation*, such as electrons or α particles, charged particles are present from the start.

The photons of X-rays or γ-rays generated in nuclear decay and energy-rich neutrons have substantial penetrating power in tissue. The high-energy electrons produced in an accelerator can also penetrate deeply into tissue. In contrast the electrons of β radiation, which are produced in radioactive decay, have a range of only a few millimeters in tissue. For this reason, β emitters are active primarily when they are incorporated. The same applies to α emitters. As a result of their short range, they are of importance only in the case of inhalation or ingestion of radionuclides.

Since the primary physical and chemical effects are identical for all types of ionizing radiation, the same quantity can be used to measure the extent of a radiation exposure. This quantity is called the *absorbed dose, D*, or simply the dose (see Chap. Introduction). The unit used, to express the absorbed dose, is J/kg, or gray (Gy). In spite of the identity of the physical primary processes, the various types of radiation have, even at the same absorbed dose, different efficiency in the cell or tissue. The reason for the differences lies in the varying microscopic concentration of the energy transferred along the tracks of the various particles (Fig. 37). *Sparsely ionizing radiations*, such as γ-rays, X-rays, or fast electrons, cause only a few dozen ionizations along their path through the cell. The resulting direct and indirect damage to DNA (see Section Molecular Radiation Effects) can be largely repaired. *Densely ionizing particles*, such as the relatively slow, heavy α particles or recoil nuclei triggered by neutrons, cause thousands of ionizations as they pass through the cell. The clustered DNA damage produced within a small site can often be only imperfectly repaired (see Section Molecular Radiation Effects). The difference between the various types of radiation is quantified by the linear energy transfer (LET), i.e., the rate of energy loss per unit distance.

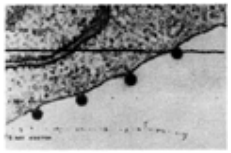

Figure 37. Schematic representation of the tracks of charged particles (from B. Alberts: Molecular Biology of the Cell, 1988)

To bring the frequency and spatial distribution of the ionizations into relationship with submicroscopic cell structures, a part of an electron micrograph of a cell is used as the background. Virus particles can be seen budding outward from the cell membrane. The electron tracks that are shown can also be regarded as typical of the fast or slow electrons triggered by X-rays or γ-rays. The proton tracks also represent the case of neutron irradiation, because neutrons produce predominantly fast protons in the tissue.

In the application of ionizing radiation in *tumor therapy*, the different efficiency of the various types of radiation must be carefully evaluated. Moreover, the effect ratio depends in a complicated manner on the dose and the dose distribution in time. Simplified evaluations of radiation quality are adequate for the purpose of *radiation protection*, where low doses are involved and, furthermore, only summary estimates are required of effects that are possible, but less likely to occur.

Radiation protection summarily takes into account the different efficiency of densely and sparsely ionizing radiation through the *quality factor, Q.* This factor represents the relative biological effectiveness of the different types of radiation with regard to the production, through DNA damage, of stochastic, long-term effects, i.e., hereditary damage or carcinogenesis (see Section Radiation Effects on Tissue). The International Commission for Radiation Protection (ICRP) has defined the quality factor as a function of the LET [183] (Fig. 38). The *dose equivalent, H*, is the product of the absorbed dose, *D*, times the quality factor, *Q*:

$$H = D \cdot Q \tag{7.1}$$

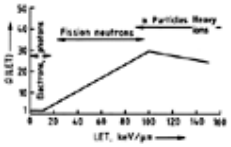

Figure 38. Quality factor as a function of the linear energy transfer, LET [183]

The corresponding ranges of LET are shown for some types of radiation.

To differentiate between the absorbed dose and the dose equivalent, the latter is expressed in the unit sievert (Sv).

7.2. Molecular Radiation Effects

The conversion from the physical to the molecular radiation effects is predominantly due to radicals generated by the ionization of a molecule. These radicals can quickly react with each other or with stable compounds in their environment. Two mechanisms of action can be differentiated. In the *direct radiation effect*, ionization and, consequently, radical formation occur in a biomolecule itself. In the *indirect radiation effect*, the primary radicals formed during the radiolysis of water react with the biomolecules. This mechanism plays the principal role in biological systems with their high water content.

7.2.1. *Radiation Damage to DNA*

DNA and its structural elements, the bases, and the sugar phosphate skeleton, can be chemically changed by direct and indirect radiation effects. One possibility is the formation of a *single-strand break* in the DNA double helix, the complementary strand remaining intact. Another possibility are *double-strand breaks* resulting when both strands are damaged at closely neighboring positions and when, as a result, the DNA molecule is severed. Other alterations, that can occur, include changes in the chemical structure or sequence of the bases in the DNA strand.

There is a range of possible results of the molecular damage to DNA (Fig. 36):

1. Disturbances in replication and transcription
2. Gene mutations
3. Chromosomal aberrations

It is well established by numerous studies in radiation biology that DNA is the main target of the exposure to ionizing radiation. However, other structures may also be affected, the most important being biological membranes.

7.2.2. *Repair of DNA Damage in the Cell*

Damage to DNA can be caused by numerous noxious agents. Ionizing radiation represents only one, although important, example of these agents. Damage to DNA would very quickly lead to a loss of cell function in the absence of effective

repair mechanisms. In fact, the evolution of bacteria and higher cells would not have been possible without complex enzyme-controlled mechanisms for the repair of molecular damage to DNA. For instance, a defective base in a DNA strand can be excised and replaced by the insertion of nucleotides in the correct order, determined by the base sequence of the complementary strand. This repair involves the action of various enzymes.

As mentioned in Section Physical Radiation Effects , the different biological effectiveness of sparsely and densely ionizing radiation is closely related to the ability of the cell to repair DNA damage. The passage of individual sparsely ionizing electrons through the cell nucleus produces only few lesions in DNA, and these can be enzymatically repaired with high efficiency. In contrast, the passage of a single densely ionizing particle through the cell nucleus produces a large number of lesions in the DNA. These lesions are in close proximity and this leads, with high probability, to misrepair. At larger doses of several gray, the differences between both types of radiation play a less important role because a large number of DNA lesions will then also occur with sparsely ionizing radiation.

7.3. Cellular Radiation Effects

Ionizing radiation can cause a large number of changes in the chromosomes by damaging DNA. Depending on whether these cytogenetic changes are visible under the light microscope, they are divided into visible *chromosomal aberrations* (numerical or structural) and invisible *gene mutations* (Fig. 36). Even if gene mutations are not microscopically visible, their consequences can be serious. The division between chromosomal aberrations and gene mutations have become less distinct with the development of new molecular biology techniques (e.g., fluorescence in situ hybridization, see below).

Numerical chromosomal aberrations are deviations from the modal number of chromosomes (in humans 46 = 2×22 autosomal + 2 sex chromosomes). They give rise to aneuploid and polyploid cells. *Aneuploidy* is characterized by the lack or excess of individual or several autosomal or sex chromosomes, as in certain inborn diseases (trisomy 21, Turner or Klinefelter syn-drome). In *polyploidy*, the diploid number of chromosomes is increased by one or several sets of chromosomes. After exposure to ionizing radiation, the formation of cells with numerical chromosomal aberrations is a result of disturbances in the division of the cell nucleus (mitosis) or the cell (cytokinesis). Altered chromosomes, e.g., the dicentric chromosomes (see below), can disturb or prevent regular division of the cell or its nucleus. Radiation can also cause a disturbance of the spindle microtubules and thus interfere with mitosis.

Radiation-induced *structural chromosomal aberrations* are either simple fragment losses (deletions) which lead to the shortening of a chromosome, or structural rearrangements. The latter are produced when several breaks occur in neighboring chromosomes and the ends are joined to give abnormal configurations (Fig. 39). If, for instance, interactions occur between the chromosomes with one break each (interchromosomal exchange), symmetrical or asymmetrical structural rearrangements can occur. An *asymmetric chromosomal exchange* figure that arises from the joining of two fragments with one centromere each is the dicentric chromosome. If chromosome sections without a centromere are exchanged, *symmetric translocations* are produced. In the metaphase of mitosis, dicentric chromosomes are readily detected with conventional staining techniques under the light microscope. Symmetric translocations are more difficult to detect.

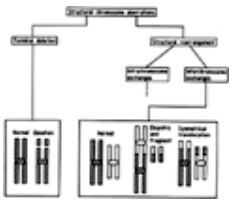

Figure 39. Schematic representation of the principal types of structural chromosome aberrations [184]

Both types of chromosome aberrations can be used as a biological dosimeter to estimate the radiation exposure of humans. The dose received can be determined on the basis of calibration curves by counting the frequency of radiationinduced chromosomal aberrations in the lymphocytes of blood. The dose estimation is usually performed through the determination of dicentric chromosomes, a significant increase being detectable only after a whole body exposure of about 100 mGy of X-rays or γ-rays. Since dicentric chromosomes interfere with the regular course of mitosis, the number of lymphocytes with dicentric chromosomes decreases with time after irradiation. For this reason, dose estimations for exposures that originated a long time ago are not very reliable. On the other hand, the number of cells with symmetric translocations remains largely unchanged for years, making this type of abnormality more suitable for the quantification of radiation exposure that is either chronic or that has taken place in the past. With the aid of *fluorescence in situ hybridization*, individual chromosomes can be specifically labeled over their entire length with chromosome-specific DNA probes. Thus, symmetric translocations with other chromosomes can be recognized. In the experimental procedure DNA samples are hybridized in situ with lymphocyte chromosomes, and labeled DNA sections are detected with a fluorescence-labeled antibody.

Radiation-induced chromosomal lesions can lead either to cell death or, in the case of less drastic damage, to nonlethal changes in the genetic material of the cell, which can culminate in malignant neoplasia (somatic cells) or hereditary disease (germ cells) (Fig. 36).

7.4. Radiation Effects on Tissue

In the past decades, a large amount of information on the effects of radiation doses has been gathered. High doses of radiation lead to severe, irreparable damage to DNA, resulting in cell destruction. If a dose of radiation exceeds a threshold that depends on the type of the tissue and the type of radiation, tissue damage results. The degree of damage, in turn, depends on the dose. No substantial effects are observed at doses below the threshold (*deterministic mechanism* of damage, Fig. 36) [183].

An important example of deterministic damage is the acute radiation sickness that occurs after a large exposure of the whole body. The $LD_{50/60}$, i.e., the dose of radiation that will kill a patient with a 50 % probability within 60 days after whole body exposure, is 3 – 5 Gy in the absence of special medical care [183]. Without treatment, death is possible after doses of 2 Gy. In the therapy for leukemia, whole body doses of 8 –12 Gy — protracted over several hours — are administered to condition the patient for a bone marrow transplantation. At these doses, the patient has no chance of survival without the bone marrow transplantation.

In contrast to deterministic effects — previously known as nonstochastic — *stochastic damage* refers to lesions in individual cells, and changes in their DNA that do not stop cell proliferation and that can lead to hereditary damage or to cancer (Fig. 36). In germ cells, changes of the DNA can cause inherited defects in the descendants of irradiated persons. Malignant diseases, which appear years or decades after exposure, can be induced in somatic cells. The development of a malignant (Section Radiation Carcinogenesis) or hereditary (Section Hereditary Radiation Effects) disease can result from the alteration of a single cell, although the probability of any single cell to cause this process is low. The dose does not determine the severity of the disease, but the number of affected cells and thus the probability of its occurrence. Stochastic effects have no threshold. Even small doses can — with correspondingly low probability — cause stochastic damage.

The absence of a threshold dose is not exclusively a characteristic of ionizing radiations. It is typical of all noxious agents with genotoxic effects. In the case of most chemical carcinogens, there is also no dose or concentration threshold below which harmful effects are impossible, and above which the risk increases. Therefore, the aim of protection against such substances, as well as ionizing radiation, can only be, to limit the risks to a level that can be neglected compared with existing hazards: exposures must be kept as low as reasonably achievable.

Prenatal exposure to radiation can cause developmental disturbances (Section Teratogenic Radiation Effects); for these effects a threshold type of dose dependency can be assumed. However, the sensitivity, especially of the developing central nervous system (CNS) in the early fetal period (primarily between the 8th and 15th week of pregnancy), is so high that the principle of avoiding all unnecessary exposure to radiation must be strictly adhered to.

7.5. Radiation Hazards

7.5.1. *Radiation Carcinogenesis*

An unexpected number of leukemias among the survivors of the atomic bombings initiated the largest epidemiological study that was ever carried out. It is a study of about 100 000 survivors that has not yet been terminated. The enhanced number of leukemias suggested that the frequency of other neoplasias may also be increased. It took more than a decade to confirm this assumption. In an analysis of this universe of atomic bomb survivors up to 1985, about 70 of the 140 deaths due to leukemia and about 350 of the 7000 deaths due to cancer are attributed to the irradiation [185].

In the first analyses of the survivors of the atomic bombings, the number of fatal cases of cancer per year was determined and divided by the total number of persons in the study. From this observed total mortality rate, the spontaneous rate, i.e., the corresponding quotient for nonexposed persons, was subtracted. The additional cancer mortality rate or the excess rate determined in this manner is shown in Figure 40. The numbers given on the ordinate can be understood as additional fatal cancers occurring in a group of 10 000 persons exposed to 1 Sv in the year in question. This simple analysis does not involve a distinction in age at the time of irradiation or in sex. Nevertheless, it shows that the distribution with time is completely different for the radiation-induced leukemias and the solid tumors. Additional cases of leukemia were observed only a few years after the exposure to radiation. The leukemia rates returned toward spontaneous rates 20 – 30 years after the exposure. In the case of solid tumors, however, an increase was observed only after a latent period of about 10 years. This enhancement of the tumor rate has then persisted and it still increases.

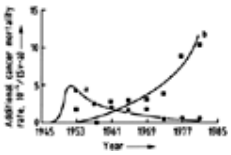

Figure 40. The radiation-induced increase in the cancer mortality rate in the survivors of the atomic bombings of Hiroshima and Nagasaki [186]. Each point corresponds to the data from four years of observation. The interpolating curves are given to facilitate readability. The data are influenced by the age distribution of the survivors of atomic bombings which changes with time; they can, therefore, not be directly linked to risk estimates.

a) All leukemias; b) All types of cancer excluding leukemias

More accurate results are obtained from a detailed analysis which takes into account not only the dose and the time after irradiation, but also the age at irradiation and the sex. Thus, the most important results of the epidemiological study in Hiroshima and Nagasaki can be represented in relatively simple models. The difference between the *model of relative risk*, that seems to apply to most of the solid tumors, and the *model of absolute risk*, which applies to leukemias, is schematically presented in Figure 41. In the model of relative risk (Fig. 41 A), it is assumed that a single exposure — after a certain latent period of 10 – 15 years — the spontaneous and age-dependent tumor rates by a factor that is a function of the dose. But the enhancement factor depends on the age at the time of irradiation. There is a clear trend of increased sensitivity for exposure at a young age. However, the survivors of the atomic bombings who were irradiated as children have not yet reached the age of high spontaneous cancer rates; the total risk estimates are, therefore, still uncertain. The model of absolute risk (Fig. 41 B) is of quite different nature, it contains the assumption that after exposure to radiation a pronounced wave of additional cases appears within a few years.

Figure 41. Diagram of the cancer mortality according to the models of relative (A)

— · · — · · 1.0 Sv; · · · · 0.3 Sv; ——— spontaneous rateand absolute risk (B)

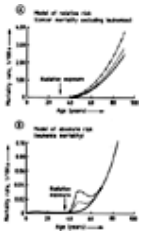

— · · — · 0.65 Sv; · · · · 0.25 Sv; ——— spontaneous rate

The epidemiological observations made in Hiroshima and Nagasaki refer to relatively high doses of radiation and they are still based on limited observation periods. Risk estimates, however, are based on the entire remaining life expectancy after irradiation and on the much smaller doses which are of importance in radiation protection. Therefore, a two-fold extrapolation in time and in dose is required. The absolute risk model for leukemias and the relative risk model for solid tumors, as shown in Figure 41, form the basis for the extrapolation [183], [187], [188]. More recent evaluations do not assume simple proportionality in the extrapolation to low doses, but postulate a further reduction factor of 2 [183]. This reduction is not supported by epidemiological observations on humans, but only by the results of animal experiments. On the other hand, the application of the relative risk model would provide an overestimate of the risk, if the relative increases of tumor mortality do not persist over the entire life. Thus, the risk estimates can be regarded as realistic; the extrapolations to longer periods after radiation exposure could be an overestimation while the assumption of the reduction factor could be an underestimation.

The risk estimates for small doses given by the International Commission for Radiological Protection (ICRP) for the different types of tumors are presented in Table 8 [183]. The total risk associated with cancer mortality, i.e., the additional probability of the individual for dying of cancer, is given as 5×10^{-5}mSv^{-1}.

Table 8. Estimated additional probability of the occurrence of a fatal cancer for various tissues and organs. The estimates include a reduction factor of 2 for the extrapolation of the observations on high doses to low doses [183].

Tissue or organ	Estimated additional probability of the occurrence of a fatal cancer (10^{-7} mSv^{-1})
Bladder	30
Bone marrow	50
Bone surface	5
Breast	20
Colon	85
Liver	15
Lung	85
Esophagus	30
Ovaries	10
Skin	2
Stomach	110
Thyroid	8
Other organs	50
Total	500

The observations on the survivors of the atomic bombings are supplemented, and to a substantial extent confirmed, by observations on patients who were exposed to radiation internally or externally for *medical diagnosis or therapy.*

The radiopaque medium thorotrast, which was used in X-ray diagnostics in Germany and Japan up to the 1950s, had disastrous consequences. The quality of the radiopaque medium was based on the fact that it contained the long-lived radionuclide ^{232}Th an element with a high atomic number. It was not sufficiently accounted for that thorium, an α emitter and precursor of ^{226}Ra, represents a radiation risk. The long-term effects were serious [181], [183]. The rates of tumors, and especially of live tumors in the patients who had received thorotrast injection, increased and continued to increase decades after.

Serious long-term effects were also observed in patients suffering from Morbus Bechterew, an inflammatory process of the bone surface, and in young patients with bone tuberculosis who were injected with high activities of ^{224}Ra, a short lived α emitter, in a German sanatorium after World War II. Of the 800 patients who were observed epidemiologically, more than 50 died of bone tumors which appeared up to 20 – 30 years later [189], [190].

Most of the studies are based on relatively high doses. However, the prenatal X-ray diagnostics that was routinely applied up to about 1960 and that resulted in an exposure of the fetus of 10 – 20 mSv represents an exception. The rates of

leukemia in these children were increased by 20 % above the normal values [191], [192].

Further important observations were made on *uranium miners* who inhaled increased concentrations of radon and its decay products for years and showed greatly increased rates of bronchial cancers. The observations made on the uranium miners were used to estimate the risk from radon in houses. Radon, with its α-emitting decay products, represents the largest, although greatly fluctuating, contribution to natural radiation exposure.

7.5.2. *Hereditary Radiation Effects*

Until now, no statistically significant increase in hereditary abnormalities has been detected in human beings as a result of ionizing radiation. This also applies to the extensive observations on the descendants of those who survived the atomic bombings of Hiroshima and Nagasaki. However, it follows from animal experiments that ionizing radiation causes heritable defects also in humans. In the absence of direct data, the estimation of the genetic risk from radiation in human populations must be obtained by extrapolation from the results of the animal experiments. These experiments which were mostly carried out on mice, indicate that a radiation dose of 1 Sv produces about the same number of genetic defects as would normally appear spontaneously in one generation. This is referred to as a *doubling dose* of 1 Sv [187]. Although no increases could be recognized there, this estimate is consistent with the limits of uncertainty of the observations among the survivors of the atomic bombings of Hiroshima and Nagasaki.

7.5.3. *Teratogenic Radiation Effects*

Prenatal irradiation can lead to developmental disturbances and organ deformities. The type and degree of these *teratogenic defects* depend in a complex manner on the phase of the embryonal or fetal development in which irradiation takes place.

Severe deformities can occur after exposure to high doses during the *embryonal period* (9th to 60th day after conception). Animal studies have indicated that a threshold dose exists for this type of damage which is substantially higher than 0.1 Sv [183]. After irradiation during the *fetal period* (from the 60th day after conception) the development of the fetus can slow down. Observations on children who were subjected to prenatal irradiation as a result of the atomic bombings of Hiroshima and Nagasaki have shown that the CNS is an especially critical organ. Developmental disturbances with pronounced morphological alterations have been observed; the clinical expression can range from a reduction in the motor and intellectual capacities (below-average IQ) to severe mental disability and microcephaly. The developing CNS is most radiation-sensitive between the 8th and 15th week of pregnancy. Even low doses carry an increased risk, but a threshold value in the order of 0.1 Sv is assumed for this phase with regard to severe developmental disturbance in the CNS [183].

Besides this deterministic damage, prenatal irradiation can also increase the risk of stochastic effects, such as leukemia and, possibly, other tumors. The use of diagnostic X-rays during pregnancy is, thus, justifiable only in exceptional cases with especially important indications. On the other hand, if the fetus has been exposed to radiation doses less than 0.1 Sv (which is the case in almost all radiodiagnostic examinations), no eugenic indication exists from a radiobiological point of view for the termination of pregnancy.

7.6. Radionuclides

In contrast to radiation exposure the outside, i.e., due to a source of radiation outside the body, radiation exposure from the inside refers to the condition in which radioactive substances are incorporated with food, water, air, or through the intact skin or wounds. In the case of some radionuclides, e.g., the two isotopes ^{134}Cs and ^{137}Cs, incorporation can lead to approximately homogeneous exposure of the whole body from the inside, because the nuclide is uniformly distributed. Many radionuclides are, however, inhomogeneously distributed, accumulating in certain molecules, in certain organelles of the cell, in certain groups of cells, and in certain organs. Examples are the accumulation of radioiodine in the thyroid, or the accumulation of so-called “bone seekers” such as strontium or radium in the skeleton or on the bone surfaces.

There is, in principle, no difference between the action of ionizing radiation on the body, whether it arises from external sources or from sources that are incorporated. The γ-rays produce their actions through the electrons released in the body, and the potential effect is therefore merely dependent on these electrons which do not differ for external or internal radiation. Furthermore, there is no principal difference between the β-rays released within the body and those electrons released by γ-rays or X-rays.

The radiation effects are, of course, greatly dependent on the dose, and the dose and its distribution in the different tissues and organs of the body can vary greatly with the type and strength of external or internal radiation sources. The incorporation, the metabolic distribution and retention, and the excretion of radionuclides are complex factors that need to be accounted for in the dosimetry of internal radiation sources. A further complexity arises from the fact that heavy radioisotopes, such as americium, plutonium, radium, and the ubiquitous radon and its daughter products are α emitters (short ranges of only a few centimeters in air and only ca. 0.05 mm in tissue, see Section Physical Radiation Effects). Their action is therefore only of concern if they are inhaled, ingested, or otherwise taken up by the body. However, in these cases they are particularly effective because they are densely ionizing (see Section Physical Radiation Effects).

7.6.1. *Radionuclides of Concern in Radiation Protection*

A multitude of radionuclides are used in medicine, in industry, and in research. Typical examples indicate the range of their different characteristics and their possible effects. The effects of the α-emitting radionuclides thorium ($t_{1/2}$: 1.4×10^{10} a) and radium (^{226}Ra $t_{1/2}$: 6000 a; ^{224}Ra $t_{1/2}$: 3.6 d) have already been mentioned (see Section Radiation Carcinogenesis). Even heavier elements have become of concern in the meantime. Examples are the artificially-produced elements americium and plutonium; they, too, are α-emitters. Although these radionuclides have been accidentally incorporated in certain instances, their biological effects, and especially the induction of bone tumors or lung tumors, have been seen only in animal experiments.

Various radioisotopes have been used and are still being used for diagnostic and therapeutic purposes in nuclear medicine. The most important example is the β emitter ^{131}I ($t_{1/2}$: 8 d). In the past it was used for thyroid diagnostics. This produced relatively high doses of about 0.5 Gy to the thyroid, and iodine has therefore been replaced for this purpose by the short-lived ^{99m}Tc (half-life: 6 h, see Chap. Production of Artificial Radionuclides). For the radiotherapy of thyroid cancer the application of high activities of ^{131}I is still the method of choice.

7.6.2. *Exposure of the Population due to Natural and Artificial Radioisotopes*

^{222}Rn and ^{220}Rn are the daughter products of ^{226}Ra and ^{224}Ra, respectively. Depending on the geological characteristics of the soil, these gaseous radionuclides are present in greatly varying concentrations. In underground mines and especially in uranium mines, radon and its α-emitting short-lived daughter products are the major health hazard, and a number of studies [183] have demonstrated strikingly enhanced lung cancer frequencies in underground miners. Radon is also the main contributor to the *natural radiation exposure* of the population, although it affects predominantly only the lung. It is estimated that up to 5 % of all lung cancer cases may be due to radon in houses [187].

A far less important contribution to the natural radiation exposure is that of the extremely long-lived ^{40}K which is present in the human body. Its specific activity is about 60 Bq/kg, and it causes a dose of approximately 0.1 μGy per year [188].

Artificial radioactivity has become notable as a cause of the exposure of large populations due to the atomic bomb tests and reactor accidents. The nuclear bomb tests produced a wide spectrum of fission products, and since these were dispersed at high altitudes and transported over large distances, only the long-lived nuclides were significant contributors to the dose. To indicate the broad spectrum of relevant radionuclides, Table 9 gives the effective doses to the population of the Northern Hemisphere that are caused due to the fallout from weapons by the main fission products.

Table 9. Effective doses to the population of the Northern Hemisphere due to the fall-out of the atomic bomb tests between 1945 and 1980 [193]

Nuclide	Half-life	Inhalation, 10^{-2} mSv	Ingestion, 10^{-2} mSv	External radiation, 10^{-2} mSv	Total, 10^{-2} mSv
^{3}H	12.3 a	0.4	4.7		5.1
^{14}C	5730 a		18		18
^{90}Sr	28.6 a	1.4	17		18.4
^{95}Zr	64 d			29	29
^{103}Ru	39.4 d			2.5	2.5
^{106}Ru	1 a	4.9		8.7	13.6
^{131}I	8 d		4.8		4.8
^{137}Cs	30.2 a		28	60	88
^{140}Ba	12.8 d			3.7	3.7
^{144}Ce	284 d	6.0		2.8	8.8
^{239}Pu	24 100 a	4.0	0.3		4.3
^{240}Pu	6 578 a	2.6	0.2		2.8
^{241}Pu	14.4 a	1.4			1.4
All nuclides		22	78	110	210

The major population exposure due to *nuclear accidents* occurred after the catastrophe of Chernobyl in 1986. In this accident the release of plutonium was largely confined to the close vicinity of the reactor. It was not, therefore, of major concern in the overall picture. The important contributors to the dose were, on the one hand, the short-lived radioiodine and, on the other hand, the long-lived cesium (^{134}Cs: $t_{1/2}$ 2.1 a, ^{137}Cs: $t_{1/2}$ 30 a) and strontium (^{90}Sr: $t_{1/2}$ 28 a).

Radioiodine, and predominantly the isotope ^{131}I, were of major concern immediately after the accident. In the contaminated territories of the former Soviet Union large activities of iodine were ingested with milk and vegetables, and it now appears that more than hundred thyroid cancers in children in Belarus alone have arisen within a few years, beginning from three to four years after the accident. In the Western European countries, the radioiodine has also been of concern after the accident, but the concentrations were lower than in the former Soviet Union and measures had been taken immediately after the accident, to block the use of contaminated milk and vegetables.

The nuclide of lasting consequence after the reactor accident has been ^{137}Cs. With its γ-rays, it acts as an external source of exposure and, both with its γ-rays and β-rays, it contributes to the exposure when it is incorporated through the food chain (see Table 10).

Table 10. Effective doses produced by the ingestion with food of 1 kBq of the long-lived radionuclides that were of main concern after the accident of Chernobyl [193] [a]

Nuclide	Effective dose by ingestion of 1 kBq, 10^{-2} mSv	Total dose in 50 years due to contaminated food, 10^{-2} mSv [b]	
		By deposition of 1 kBq/m^2	By deposition after the Chernobyl accident in the north of Munich
^{137}Cs	1.4	0.9 [d]	59.0
		2.2 [e]	
^{134}Cs	2.0	0.1 [d]	4.0
		0.3 [e]	
^{90}Sr	3.5 [b]	5.4 [d]	1.2
		0.7 [e]	

[a] The doses are given that result over a period of 50 years from the use of products in a region that is contaminated with a surface specific activity of 1 kBq/m^2.
[b] The last column gives the doses that result in this way in a region subject to the same contamination that occurred in the north of Munich after the Chernobyl accident [194].
[c] 39×10^{-2} mSv bone surface; 17×10^{-2} mSv bone marrow.
[d] Food derived from plants.
[e] Food derived from animals.

Strontium is of considerable concern after a reactor accident because it is a bone seeker that can specifically contribute to the exposure of the bone marrow and thereby produce leukemias. It is also difficult to measure; it is a pure β emitter with no accompanying γ-rays. After the accident of Chernobyl the release of strontium has been less important even in the vicinity of the reactor than cesium. However, there have been earlier nuclear contaminations where strontium was the main contributor to the population dose. The most important example is the lasting contamination of the river Techa near the city of Cheliabinsk that occurred when the first nuclear bombs were produced in the former Soviet Union. The follow-up of health effects among the exposed population has shown a definite increase of leukemias [195].

7.7. Basic Principles of Radiation Protection

The basic principle for limiting the risk of stochastic radiation effects, i.e., hereditary damage and cancers, is the reduction of unnecessary radiation exposure to a level as low as reasonably achievable (ALARA). Upper limits for radiation exposure have been given as an additional constraint, but not as values to be freely exhausted. According to the latest recommendations of the ICRP [183], the upper limit for whole body exposure in a period of 5 years is 100 mSv, and a value of 50 mSv must not be exceeded in one calendar year.

Since the body is not always exposed uniformly, the ICRP has introduced the concept of the *effective dose, E.* This concept permits the comparative evaluation of an inhomogeneous irradiation of the body. The effective dose is obtained from the dose equivalents (see Section Physical Radiation Effects) for the different tissues and organs (H_T) of a person who is partially or completely exposed to radiation. These dose equivalents are the mean values for the particular organs or tissue. Since the radiation-induced risk is not the same for different tissues or organs, the corresponding equivalent doses are multiplied by the *tissue weighting factors,* w_T (Table 11) and the products are then summed up for all the tissues and organs of the body:

$$E = \sum_T w_T \cdot H_T \tag{7.2}$$

As in the case of the dose equivalent, the effective dose is expressed in sievert.

Table 11. Tissue weighting factors for the calculation of the effective dose [183]

Tissue or organ	Tissue weighting factor, w_T
Gonads	0.20
Bone marrow	0.12
Colon	0.12
Lung	0.12
Stomach	0.12
Bladder	0.05
Breast	0.05
Liver	0.05
Esophagus	0.05
Thyroid	0.05

Skin	0.01
Bone surface	0.01
Other organs or tissue	0.05

The sum of the weighting factors for all the tissues and organs of the body is 1. This means that in the case of homogeneous exposure of the whole body, in which the equivalent doses for all the tissues and organs are the same, the effective dose is equal to the dose equivalent. If only one organ is irradiated, the effective dose is less than the dose equivalent in this organ. To give an example, if only the thyroid gland is exposed, the effective dose is 1/20 of the dose equivalent in the thyroid (Table 9).

The dose limits mentioned above apply to the effective dose, i.e., there is an upper limit of 100 mSv in a 5 year period and 50 mSv in one year. According to the present law in Germany [196], [197], the limit of the lifetime dose from exposure in an occupation is 400 mSv.

Exposure to low doses is the primary concern of radiation protection. The concept of the effective dose and the tissue weighting factors refers, therefore, to stochastic effects (malignant diseases, hereditary defects in descendants). However, the limiting values for the effective dose are sufficiently low, that even in the case of inhomogeneous irradiation of the body, the dose equivalents for any organ remain below the threshold values for deterministic damage (acute or chronic damage). Such damage can generally be excluded within the limiting values. Exceptions to this rule are the lens of the eye and the skin. Radiation can produce deterministic damage (cataracts) in the lens of the eye. The skin can be irradiated very inhomogeneously, and the limitation of the average dose can, therefore, not exclude deterministic effect from a local over-exposure. For these reasons, it is also required that the annual dose equivalent does not exceed 150 mSv for the lenses of the eye and 500 mSv for the skin (taken as an average over 1 cm^2).

The laws governing radiation protection and X-rays in Germany [196], [197] go beyond the recommendations of the ICRP in several areas, especially in the regulations concerning young people, pregnant women, and nursing mothers. For instance, pregnant women must not work in control areas, and nursing women must not come into contact with open radioactive substances.

To protect the population, the laws on radiation protection stipulate that facilities are to be planned and operated in such a way that they never contribute to any member of the public, more than 0.3 mSv per year to the effective dose through the discharge to air or water of radioac-tive substances. The corresponding value proposed by the ICRP is 1 mSv per year.

For special properties of Pm and Rn: see the corresponding Rare Earths and Noble Gas volumes for specific compounds

[Top of Page]

References

General References

1. *Gmelin*, 8th ed., Technetium, System no. 69, Suppl. vol. **1 and 2** (1982). *Gmelin*, 8th ed., Polonium, System no. 12, Suppl. (1990). *Gmelin*, 8th ed., Astatine, System no. 8 a, Suppl. (1985). *Gmelin*, 8th ed., **Francium, System no. 25 a, Suppl.** (1983). *Gmelin*, 8th ed., Radium, System no. 31, Suppl. vol. **1 and 2** (1977). *Gmelin*, 8th ed., Actinium, System no. 40, Suppl. (1981). *Gmelin*, 8th ed., Protactinium, System no. 51, Suppl. vol. **1 and 2** (1977).
2. E. J. Wheelwright: *Promethium Technology*, American Nuclear Society, Winsdale, Ill., 1977.
3. F. Weigel, *Chem. Ztg.* **102** (1978) 287 – 299.
4. S. Glasstone: *Sourcebook of Atomic Energy*, 3rd ed., Van Nostrand, Princeton 1967, 10.36 – 10.50, pp.353 – 358.
5. G. Harbottle, A. G. Maddock (eds.): *Chemical Effects of Nuclear Transformations in Inorganic Systems*, North Holland, Amsterdam 1979.
6. G. A. Andrews: “History of Radiopharmacy,” in M. Tubis, W. Wolf (eds.): *Radiopharmacy*, John Wiley and Sons, New York 1976.
7. R. S. Mani: “Reactor Produced Radionuclides,” in F. Helus (ed.): *Radionuclides Production*, vol. 2, CRC Press, Boca Raton, Fla., 1983, pp. 2 – 45.
8. F. Helus, G. Wolber: “Activation Techniques,” in F. Helus (ed.): *Radionuclides Production*, vol. 1, CRC Press, Boca Raton, Fla., 1983, pp. 57 – 120.
9. B. Hoop, J. S. Laughlin, R. S. Tilbury: “Cyclotrons in Nuclear Medicine,” in G. J. Hine, J. A. Sorenson (eds.): *Instrumentation in Nuclear Medicine*, Academic Press, New York 1974, pp. 407 – 457.
10. G. Rayudu, *Semin. Nucl. Med.* **20** (1990) 100 – 110. Links
11. J. R. Votaw, R. J. Nickles, *Nucl. Instrum. Methods Phys. Res. Sect. B* **B40/41** (1989) 1093 – 1099.
12. M. D. Kamen: *Radioactive Tracers in Biology*, 2nd ed., Academic Press, New York 1951.
13. *JAMA, J. Am. Med. Assoc.* **259** (1988) 1854 – 1859.
14. W. Wolf, M. Tubis: “Radiopharmaceuticals,” *J. Pharm. Sci.* **56** (1967) 1 – 17. Links
15. M. C. Lagunas-Solar, *Int. J. Nucl. Med. Biol.* **10** (1983) 75 – 84. Links
16. J. Steinbach, *Isotopenpraxis* **26** (1990) 102 – 106.
17. R. M. Lambrecht, R. Neirinckx, A. P. Wolf, *Int. J. Appl. Radiat. Isot.* **29** (1978) 175 – 180.
18. S. C. Jones, G. D. Robinson, E. Mcintyre, *Int. J. Appl. Radiat. Isot.* **35** (1984) 721 – 729. Links
19. B. L. Cohen, T. R. Canada, *Phys. Med. Biol.* **17** (1972) 416 – 417. Links

20. V. I. Levin, A. B. Malinin, I. N. Tronova, *Radiochem. Radioanal. Lett.* **49** (1981) 111 – 118.
21. A. B. Malinin, N. V. Kurenkov, *Radiochem. Radioanal. Lett.* **53** (1982) 311 – 318.
22. A. P. Kushelevsky, Z. B. Alfassi, T. Schlessinger, W. Wolf, *Int. J. Appl. Radiat. Isot.* **30** (1979) 275 –280.
23. M. C. Lagunas-Solar *et al.*, *Int. J. Appl. Radiat. Isot.* **7** (1991) 643 – 657.
24. L. E. Ketchum, *J. Nucl. Med.* **28** (1987) 1371 – 1378. Links
25. B. Larsson, *Radiat. Res.* **104** (1985) S310 – S318.
26. H. A. O'Brien, Jr., et al.: "Development of Short-Lived Radionuclides at Los Alamos," in P. Paras, J. W. Thiessen (eds.): "The Developing Role of Short-Lived Radionuclides in Nuclear Medical Practice," *DOE Symp. Ser.* **56** (1985) 147 – 160.
27. L. F. Mausner, P. Richards: "Short-Lived Radionuclide Production Capability at the Brookshaven Linac Isotope Producer," in P. Paras, J. W. Thiessen (eds.): "The Developing Role of Short-Lived Radionuclides in Nuclear Medical Practice," *DOE Symp. Ser.* **56** (1985) 161 – 170.
28. *Radioisotope Production and Quality Control*, Technical Report no. 128, International Atomic Energy Agency, Vienna 1971.
29. W. A. Volkert, W. F. Geokeler, G. J. Ehrhardt, A. R. Ketring, *J. Nucl. Med.* **32** (1991) 174 – 185. Links
30. M. Mladjenovic: "Nuclear Physics Fundamentals," in F. Helus (ed.): *Radionuclides Production*, vol. 1, CRC Press, Boca Raton, Fla., 1983.
31. G. S. Bauer, *J. Fusion Energy* **8** (1989) 169 – 180.
32. W. Wolf, J. A. Berman, F. K. V. Leh, K. Poggenburg, *Recent Adv. Nucl. Med. Proc. World Congr. Nucl. Med.* **1st** (1974) 944 – 945.
33. S. M. Quaim, *Radiochim. Acta* **41** (1987) 111 – 117.
34. G. K. Mulholland, M. R. Kilbourn, J. J. Moskwa, *Int. J. Appl. Radiat. Isot.* **41** (1990) 1193 – 1199.
35. J. C. Clark *et al.*, *Int. J. Appl. Radiat. Isot.* **38** (1987) 597 – 600.
36. M. C. Lagunas-Solar, F. Omar, *Int. J. Appl. Radiat. Isot.* **41** (1990) 349 – 357.
37. M. C. Lagunas-Solar, O. F. Carvacho, R. R. Cima, *Int. J. Appl. Radiat. Isot.* **39** (1988) 41 – 47.
38. R. D. More, J. H. Troughton, *Int. J. Appl. Radiat. Isot.* **23** (1972) 344 – 345. Links
39. C. Vandecasteele, *IEEE Trans. Nucl. Sci.* **NS-28** (1981) 1928 – 1932.
40. H. V. Pitingsrud, P. J. Robbins, *Med. Phys.* **12** (1985) 613 – 617. Links
41. J. E. Beaver, H. B. Hupf, *J. Nucl. Med.* **12** (1971) 739 – 741. Links
42. Z. Randa, K. Svoboda, *J. Inorg. Nucl. Chem.* **39** (1977) 2121 – 2123.
43. G. L. Almeida, F. Helus, *Radiochem. Radioanal. Lett.* **28** (1977) 205 – 214.
44. P. V. Kulkarni, J. R. Corbett, *Semin. Nucl. Med.* **20** (1990) 119 – 129. Links
45. Z. Alfassi: "Production of Radionuclides by 14 MeV Generators," in F. Helus (ed.): *Radionuclides Production*, vol. 2, CRC Press, Boca Raton, Fla., 1983.
46. J. S. Vincent: "The Iodine-123 Program at the TRIUMPH Laboratory," in P. Paras, J. W. Thiessen (eds.): "The Developing Role of Short-Lived Radionuclides in Nuclear Medical Practice," *DOE Symp. Ser.* **56** (1985) 232 – 240.
47. M. C. Lagunas-Solar, J. A. Jungerman, N. F. Peek, R. M. Theus, *Int. J. Appl. Radiat. Isot.* **29** (1978) 159 – 171.
48. A. Plenevaux, *Int. J. Appl. Radiat. Isot.* **41** (1990) 829 – 838.
49. H. Sharma *et al.*, *Int. J. Appl. Radiat. Isot.* **37** (1986) 105 – 109.
50. J. Zweit, H. Sharma, S. Downey, *Int. J. Appl. Radiat. Isot.* **38** (1987) 499 – 501.
51. S. Mirzadeh, L. F. Mausner, S. C. Srivastava, *Int. J. Appl. Radiat. Isot.* **37** (1986) 29 – 36.
52. D. E. Troutner, *Int. J. Nucl. Med. Biol.* **14** (1987) 171 – 176.
53. Brucer: "One Hundred and Eighteen Radioisotope Cows," *Isot. Radiat. Technol.* **3** (1965) 1 – 36.
54. G. S. Subramanian: "Radionuclide Generators," in M. Tubis, W. Wolf (eds.): *Radiopharmacy*, John Wiley and Sons, New York 1976.
55. M. Guillaume, C. Brihaye, *ACS Symp. Ser.* **241** (1984) 185 – 197.
56. R. S. Mani, *Radiochim. Acta* **41** (1987) 103 – 110. Links
57. P. W. Moore, *J. Nucl. Med.* **25** (1984) 499 – 502. Links
58. R. M. Lambrecht, M. Sajjad, *Int. J. Appl. Radiat. Isot.* **39** (1988) 1081 – 1083.
59. Z. Kovacs *et al.*, *Int. J. Appl. Radiat. Isot.* **42** (1991) 329 – 335.
60. R. W. Atcher, A. M. Friedman, J. R. Huizenga, *Int. J. Nucl. Med. Biol.* **7** (1980) 75 – 78. Links
61. P. L. Horlock, *J. Radioanal. Chem.* **64** (1981) 257 –265.
62. C. Brihaye *et al.*, *Int. J. Appl. Radiat. Isot.* **40** (1989) 183 – 191.
63. T. A. Butler, C. E. Guyer, F. F. Knapp, *Nucl. Med. Biol. Adv. Proc. World Congr.* **3rd** (1983) no. 1, 617 –620.
64. L. Brihaye, *J. Nucl. Med.* **23** (1982) 114 – 1120.
65. I. Mena, C. S. Marcus, R. deJong, W. Wolf, in F. F. Knapp Jr., T. A. Butler (eds.): *Radionuclide Generators: New Systems for Nuclear Medicine Applications*, American Chemical Society, Washington, D.C., 1984, pp. 23 – 33.
66. R. W. Atcher, A. M. Friedman, J. Hines, *Int. J. Appl. Radiat. Isot.* **39** (1988) 283 – 286.
67. R. W. Atcher *et al.*, *J. Radioanal. Nucl. Chem.* **135** (1989) 215 – 221.
68. M. C. Lagunas-Solar, F. E. Little, C. D. Goodart, *Int. J. Appl. Radiat. Isot.* **33** (1982) 1439 – 1443. Links
69. M. C. Lagunas-Solar, F. E. Little, H. A. Moore, *Int. J. Appl. Radiat. Isot.* **33** (1982) 619 – 628. Links
70. G. F. Knoll: *Radiation Detection and Measurements*, J. Wiley & Sons, New York 1989.

71. W. R. Leo: *Techniques for Nuclear and Particle Physics Experiments*, Springer Verlag, Berlin 1987.
72. H. Neuert: *Kernphysikalische Meßverfahren*, Verlag G. Braun, Karlsruhe 1966.
73. E. Sauter: *Grundlagen des Strahlenschutzes*, Thiemig AG, München 1982.
74. E. Segrè: *Nuclei and Particles*, Benjamin Inc., New York 1965.
75. Automess GmbH, company brochure, Ladenburg, FRG.
76. EG & G Berthold, company brochure, Wildbad, FRG.
77. EG & G Ortec, company brochure, München, FRG.
78. FAG Kugelfischer KGaA, company brochure, Erlangen, FRG.
79. Genitron Instruments GmbH, company brochure, Frankfurt/M, FRG.
80. Herfurth GmbH, company brochure, Hamburg, FRG.
81. SILENA Wiss. Instrumente GmbH, company brochure, Hasselroth, FRG.
82. STI SOLON Techn. Instruments (formerly HARSHAW CHEM.), company brochure, Wermelskirchen, FRG.
83. HARSHAW/QS S.T.I., Solon Technologies Inc., company brochure, Solon, Ohio, USA.
84. R. Maushart: *Strahlenschutz-Technik, Strahlenschutz-Meßtechnik*, Fachverband für Strahlenschutz e.V., FS-91–58, Aachen 1991.
85. M. R. Cleland, G. M. Pagean, *Nucl. Instr. and Meth. Phys. Res. B* **24/25** (1987) 967 – 972.
86. M. R. Cleland, C. C. Thompson, M. Strelczyk, D. P. Sloan, *Radiat. Phys. Chem.* **35** (1990) nos. 4 – 6, 632 – 637.
87. J. L. Williams, T. S. Dunn, *Radiat. Phys. Chem.* **14** (1979) nos. 1, 2, 185 – 201.
88. P. Kunstadt, W. Gibson, *Radiat. Phys. Chem.* **35** (1990) nos. 4 – 6, 551 – 559.
89. R. C. Oberth: Cobalt Production and Supply, 4th Nordion Gamma Processing Seminar, Ottawa, 26.–31. 05. 1991.
90. Nordion International Inc.: World List of Industrial Gamma Processing Facilities, Canada, March 1991.
91. L. Wiesner, *Jahrbuch der Atomwirtschaft* **21** (1990) A 47 – A 61.
92. V. Markovic, Bull. Spring IAEA Vienna 1985, pp. 33 – 39.
93. A. Heger: *Technologie der Strahlenchemie von Polymeren*, Hanser Verlag, München 1990.
94. F. M. De Meulemeester, *Radiat. Phys. Chem.* **35** (1990) nos. 4 – 6, 580 – 584.
95. H. G. Schrader, *Radiat. Phys. Chem.* **35** (1990) nos. 4 – 6, 569 – 571.
96. E. W. Beers: “Optimum Irradiator Design with Flexibility,” *4th Nordion Gamma Processing Seminar*, Ottawa, May 26 – 31, 1991.
97. H. G. Gay, B. K. Wilson, J. F. Swinwood, *beta-gamma* **1** (1988) no. 2, 21 – 27.
98. G. R. Malkoske, W. H. Gibson, *beta-gamma* **2** (1990) no. 2, 21 – 24.
99. V. Stenger, P. Hargittai, B. Kalman, M. Sztyevko, *Radiat. Phys. Chem.* **26** (1985) no. 5, 591 – 597.
100. R. G. McKinnon, R. D. H. Chu, *Radiat. Phys. Chem.* **25** (1985) nos. 1 – 3, 141 – 154.
101. J. Cuda, *Radiat. Phys. Chem.* **25** (1985) nos. 1 – 3, 411 – 423.
102. J. Laizier, F. Erhart, R. Vuillemey, *Radiat. Phys. Chem.* **36** (1990) no. 5, 667 – 678.
103. R. A. Harrod, *Radiat. Phys. Chem.* **9** (1977) nos. 1 –3, 91 – 117.
104. E. W. Beers, *Radiat. Phys. Chem.* **35** (1990) nos. 4 –6, 539 – 546.
105. S. W. Dorpema, *Radiat. Phys. Chem.* **35** (1990) nos. 1 – 3, 357 – 360.
106. Guideline for Gamma Radiation Sterilization, AAMI ST 32-R-3/91, Association for the Advancement of Medical Instrumentation, Arlington, March 1991.
107. R. M. Brinston: “Cobalt 60-Demand and Competitiveness,” *4th Nordion Gamma Processing Seminar*, Ottawa, May 26 – 31, 1991.
108. P. Loakaranu, *Radiat. Phys. Chem.* **35** (1990) nos. 1 – 3, 223 – 231.
109. Wholesomeness of Irradiated Food, Report of a Joint FAO/IAEA/HO Expert Committee, Technical Report Series 659, World Health Organization, Geneva 1981.
110. A. Gervé, Radioisotopes in Mechanical Engineering, Fourth United Nations International Conference on the Peaceful Uses of Atomic Energy, AED-Conf. 71–100–55, May 1971.
111. A. Gervé, J. Schüßler, Isotopentechnik im Maschinenbau, Aktivierungsanalytische Spezialmethoden-Dünnschichtdifferenzverfahren, FKM-Forschungsheft 6, Forschungskuratorium Maschinenbau e.V., Frankfurt 1970.
112. B. Kehrwald, RNT-Radionuklid-Verschleißmeß-technik, Lehrgang Nr. 09931/42.040, Verschleiß-messung an der Technischen Akademie Esslingen, 30. 11. – 01. 12. 1987.
113. A. Gervé, H. Kamm, G. Katzenmeier: “Zur Deuteronenaktivierung von Maschinenteilen aus Stahl und Gußeisen,” *Forschungsbericht Heft 77*, Forschungsvereinigung Verbrennungskraftmaschinen e.V., Frankfurt, 1968.
114. A. Gervé, G. Schatz, “Applications of Cyclotrons in Technical and Analytical Studies,” 7. *Intern. Confer. on Cyclotrons and their Applications*, Zürich, Aug. 19 – 22, 1975.
115. A. Gervé: “Die wichtigsten Verschleißmeßmethoden der Isotopentechnik,” *Kerntech. Isotopentech. Chem.* **14** (1972) no. 5, 204 – 209.
116. A. Gervé, B. Kehrwald, W. Mach: “New Strategies for Analyzing and Optimizing Engine Components and Engine Testing Programmes in Applying the Continuous Wear Measurements Technology (Radionuclide Technology),” *Congres International Méthodologic et moyens d'essais: Contribution à la conception du véhicule*, Nov. 20 – 21, 1991, E.S.I.E.E. Paris Marne la Vallée.
117. W. Bockelmann, A. Gervé, B. Kehrwald, O. Willenbockel: *Optimierung des Verschleißverhaltens am Ventiltrieb des Opel 3, 0 l/24 V-Motors, MTZ 52*, Franckh-Kosmos Verlag, Stuttgart 1991.

118. B. Kehrwald, A. Gervé: “Verschleißminderung durch Ionenimplantation, Ergebnisse aus Tribometerversuchen und Prüfstandsversuchen mit Diesel-Vollmotoren, *VDI-Ber.* **852** (1991).
119. W. Kaiser, *MT Z.* **34** (1973) no. 4.
120. G. Katzenmeier, *MT Z.* **34** (1973) no. 6.
121. J. S. Charlton, J. A. Heslop, P. Johnson in: *Industrial Application of Radioisotopes and Radiation Technology*, Internat. Atomic Energy Agency, Vienna 1982, p. 393.
122. G. Földiàk (ed.): *Industrial Application of Radioisotopes*, Elsevier, Amsterdam 1986.
123. S. N. Davis in: *Isotope Techniques in the Study of the Hydrology of Fractured and Fissured Rocks*, Internat. Atomic Energy Agency, Vienna 1989, p. 1.
124. S. W. Smith in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 834.
125. M. Virtanen in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 695.
126. T. Dincer, T. Florkowski, S. Salamba in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 681.
127. M. Weiss et al. in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 43.
128. J. Molinari, J. Guizerix in: *Isotopes in Hydrology*, Internat. Atomic Energy Agency, Vienna 1967, p. 653.
129. E. Makowski in: *Isotopes in Hydrology*, Internat. Atomic Energy Agency, Vienna 1967, p. 601.
130. A. Owczarczyk, S. Szpilowski in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 709.
131. J. F. Easey, G. L. Hanna in: *Isotope Techniques in Water Resources Development*, Internat. Atomic Energy Agency, Vienna 1987, p. 593.
132. A. B. Muller, C. V. Haynes, Jr. in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 843.
133. E. Gaspar, *Modern Trends in Tracer Hydrology*, CRC Press, Boca Raton, Fla., 1987.
134. E. Gaspar, M. Oncescu, *Radioactive Tracers in Hydrology*, Elsevier, Amsterdam 1972.
135. A. Tenu, F. D. Davidescu, A. Slavescu in: *Isotope Techniques in Water Resources Development*, Internat. Atomic Energy Agency, Vienna 1987, p. 439.
136. A. Tenu, F. D. Davidescu in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 249.
137. B. L. Campbell, B. W. Seatonberry in: *Isotopes in Hydrology*, Internat. Atomic Energy Agency, Vienna 1967, p. 227.
138. L. Gonzalez-Hita, M. B. Tienza in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 753.
139. F. Tola et al. in: *Isotope Hydrology 1983*, Internat. Atomic Energy Agency, Vienna 1984, p. 663.
140. M. I. Darmawijaya, W. H. Sisworo in: *Isotopes and Radiation in Research on Soil-Plant Relationships*, Internat. Atomic Energy Agency, Vienna 1979, p. 97.
141. T. J. D'Souza, K. B. Mistry in: *Isotopes and Radiation in Research on Soil-Plant Relationships*, Internat. Atomic Energy Agency, Vienna 1979, p. 407.
142. K. A. Malik, K. Haider in: *Soild Organic Matter Studies*, vol. 1, Internat. Atomic Energy Agency, Vienna 1977, p. 215.
143. F. F. Turenne, J. L. Repaire in: *Isotopes and Radiation in Research on Soil-Plant Relationships*, Internat. Atomic Energy Agency, Vienna 1979, p. 333.
144. W. C. McDuffie in: *Radioisotopes in the Detection of Pesticide Residues*, Internat. Atomic Energy Agency, Vienna 1966, p. 3.
145. R. Hirschberg in : R. R. Muccino (ed.): *Synthesis and Applications of Isotopically Labeled Compounds 1985*, Elsevier, Amsterdam 1986.
146. M. A. Maxam, W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74** (1974) 560. Links
147. F. Sanger, S. Nicklen, A. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74** (1977) 5463. Links
148. F. Sanger, *Science (Washington D.C.)* **214** (1981) 205.
149. G. J. Sauvé, K. E. Lipson, S. Chen, R. Baserga, *DNA Sequence* **1** (1990) 13. Links
150. M. A. Kashdan, D. L. Ornstein in R. R. Muccino (ed.): *Synthesis and Applications of Isotopically Labeled Compounds 1985*, Elsevier, Amsterdam 1986, p. 183.
151. A. Rosenthal *et al.*, *DNA Sequence* **1** (1990) 63. Links
152. A. D. Hershey, M. Chase, *J. Gen. Physiol.* **36** (1952) 39. Links
153. A. Kornberg: *DNA Replication*, W. H. Freeman, New York 1980.
154. W. Sänger: *Principles of Nucleic Acid Structure*, Springer Verlag, Berlin 1984.
155. M. Nirenberg in: *Nobel Lectures: Physiology or Medicine* (*1963 – 1970)*, American Elsevier, New York 1973, p. 372.
156. H. G. Khorana in: *Nobel Lectures: Physiology or Medicine* (*1963 – 1970)*, American Elsevier, New York 1973, p. 341.
157. R. Kavenoff, L. C. Klotz, B. H. Zimm, *Cold Spring Harbor Symp. Quant. Biol.* **38** (1974) 4.
158. M. Tubis, W. Wolf, *Radiopharmacy* 1976, 206 – 207.
159. M. Calvin *et al*: *Isotopic Carbon*. John Wiley and Sons, New York 1949.
160. A. Murray, D. L. Williams: *Organic Syntheses with isotopes*, Interscience, New York 1958.
161. J. R. Catch: *Carbon 14 Compounds*, Butterworths, London 1961.
162. V. F. Raaen, G. A. Ropp, H. P. Raaen: *Carbon-14*, McGraw Hill, New York 1968.
163. E. A. Evans: *Tritium and its Compounds*, 2nd ed., Halsted Press, New York 1974.
164. J. Sirchis (ed.): *Proceedings of the Conference on Methods of Preparing and Storing Marked Molecules. (EUR 1625.e)*, Euratom, Brussels 1964.
165. J. Sirchis (ed.): *Methods of Preparing and Storing Labelled Compounds (EUR 4746 d-f-e)*, Euratom, Brussels 1968.

166. "Radiopharmaceuticals and Labelled Compounds," *Proceedings of an IAEA Symposium* (IAEA-SM-171/30), Vienna 1973.
167. "Radiopharmaceuticals and Labelled Compounds 1984." *Proceedings of an IAEA Symposium* (IAEA/SM), Vienna 1984.
168. R. R. Muccino (ed.): *Synthesis & Applications of Isotopically Labeled Compounds*, Elsevier, Amsterdam 1985.
169. T. A., Ballio (ed.): *Synthesis & Applications of Isotopically Labelled Compounds*, Elsevier, Amsterdam 1988.
170. R. J. Bayly, E. A. Evans, J. S. Glover: "Synthesis of Labeled Compounds," M. Tubis, W. Wolf (eds.): *Radiopharmacy*, Wiley-Interscience, New York 1976.
171. H. E. Mertel: "Synthesis of Isotopically Labeled Compounds", E. H. Garrett, J. L. Hirtz, (eds.): *Drug Fate and Metabolism*, vol. 3, Marcel Dekker, New York 1979.
172. E. A. Evans: "The Synthesis of Radiolabeled Compounds," *J. Radioanal. Chem.* **64** (1981) 9 – 32.
173. J. S. Fowler, A. P. Wolf: *The Synthesis of Carbon-11, Fluorine-18 and Nitrogen-13 Labeled Radiotracers for Biomedical Applications*, NAS-NS-3201, US. Department of Energy, Technical Information Center Springfield 1982.
174. W. Schwarz, J. Welch: "Radiolabeled Antibodies, Introduction and Metal Conjugation Techniques," *Appl. Enzyme Biotechnol.*, IUCCP Symposium, 1991, 1 – 13.
175. L. Melander: *Isotope Effects on Reaction Rates*, Ronald Press, New York 1981.
176. E. N. Vine, D. Young, W. H. Vine, W. Wolf: "An Improved Synthesis of ^{18}F-5-fluorouracil," *Int. J. Appl. Radiation* **30** (1979) 401 – 405.
177. R. J. Wilson (ed.): *The Radiochemical Manual.*, 2nd ed., The Radiochemical Centre, Amersham 1966.
178. N. von Jagle, G. Schwarz, L. von Sienbenrock: "Blutbefunde bei Röntgenologen," *Berl. Klin. Wochenschr.* **48** (1911) 1220 – 1222.
179. F. Dessauer: "Über einige Wirkungen von Strahlen," *Z. Phys.* **12** (1922) 38.
180. H. S. Martland: "The Occurrence of Malignancy in Radioactive Persons: A General Review of Data Gathered in the Study of the Radium Dial Painters, with Special Reference to the Occurrence of Osteogenic Sarcoma and the Interrelationship of Certain Blood Diseases," *Am J. Cancer* **15** (1931) 2435 –2516.
181. D. M. Taylor, C. W. Mays, G. B. Gerber, R. G. Thomas: "Risk from Radium and Thorotrast" *BIR Report* 21, EUR 12088, Brussels 1989.
182. E. B. Lewis: "Leukemia and Ionizing Radiation," *Science (Washington, D.C.)* **125** (1957) 965 – 972.
183. International Commission on Radiological Protection, ICRP Publication 60. 1990 Recommendations of the International Commission on Radiological Protection, Pergamon Press, Oxford 1991.
184. M. Tubiana, J. Dutreix, A. Wambersie: *Introduction to Radiobiology*, Taylor & Francis, London 1990.
185. D. L. Preston, D. A. Pierce: "The Effect of Changes in Dosimetry of Cancer Mortality Risk Estimates in Atomic Bomb Survivors," *Radiat. Res.* **114** (1988) 437 – 466. Links
186. W. K. Sinclair: "Risk, Research, and Radiation Protection," *Radiat. Res.* **112** (1987) 191 – 216. Links
187. National Research Council, BEIR V report, Health Risks of Exposure to Low Levels of Ionizing Radiation, National Academy Press, Washington, D.C., 1990.
188. UNSCEAR: Sources, Effects and Risks of Ionizing Radiation. Annex F. Radiation Carcinogenesis in Man. E.88.IX.7, New York 1988.
189. H. Spiess, C. W. Mays: "Bone Cancer Induced by 224-Ra (ThX) in Children and Adults," *Health Phys.* **19** (1970) 713 – 729. Links
190. H. Spiess, C. W. Mays, D. Chmelevsky: "Malignancies in Patients Injected with Radium-224," in *BIR Report 21*, 7 – 12, 1989.
191. A. Stewart, J. Webb, D. A. Hewitt: "A Survey of Childhood Malignancies," *Br. Med. J.* **1** (1958) 1495 – 1508. Links
192. R. W. Miller, J. D. Boice: "Radiogenic Cancer after Prenatal or Childhood Exposure," in: A. Upton et al.(eds.): *Radiation Carcinogenesis*, Elsevier, New York 1986.
193. UNSCEAR: Ionizing Radiation: Sources and Biological Effects, E.82.IX.8, New York 1982.
194. M. Matthies *et al.*: "Simulation des Transfers von Radionukliden in landwirtschaftlichen Nahrungsketten," *GSF-Bericht S-882*, Neuherberg 1982.
195. M. M. Kossenko, M. O. Degteva, N. A. Petrushova: "Estimate of the Risk of Leukemia to Residents Exposed to Radiation as a Result of a Nuclear Accident in the Southern Urals," *The PSR Quarterly* **2** (1992) 187.
196. R. Kramer, Z. Zerlett: *Strahlenschutzverordnung, Strahlenschutzvorsorgegesetz*, Verlag W. Kohlhammer, Stuttgart 1990.
197. R. Kramer, Z. Zerlett: *Röntgenverordnung*, Verlag W. Kohlhammer, Stuttgart 1991.

[Top of Page]

Related Articles

⟶ Isotopes, Natural; ⟶ Nuclear Technology; ⟶ ; ⟶ Radiation Chemistry; ⟶ Thorium and Thorium Compounds; ⟶ ; ⟶ Uranium, Uranium Alloys, and Uranium Compounds; are separate keywords.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007 John Wiley & Sons, Inc. All Rights Reserved.

Radiopaque Media

Ulrich Speck, Schering Aktiengesellschaft, Berlin, Federal Republic of Germany

Ute Hübner-Steiner, Schering Aktiengesellschaft, Berlin, Federal Republic of Germany

1. Introduction 1
2. Urography, Angiography, and Computerized Tomography 5
3. Cholecystography and Cholangiography 10
4. Bronchography 13
5. Myelography 13
6. Lymphography 14
7. Gastrointestinal Contrast Media 14
8. References 14

1. Introduction

General Principles. The clinical use of X-rays for diagnosis (roentgenography) began immediately after the discovery of X-rays by Conrad Wilhelm Röntgen in 1895. This new tool was first used for the diagnosis of foreign bodies, fractures, and diseases of the thorax. Air-containing lung tissues, soft parts of the body (e.g., the heart), and bones differ from the surrounding tissues in their X-ray absorbing properties. This allows differentiation of tissue types on the basis of differences in the blackening of a photographic film. X-ray contrast media can alter the radiolucency; they may be either "negative" or "positive".

Negative X-ray contrast media are gaseous substances that absorb X-rays less strongly than the surrounding body tissues. In contrast to positive media, which produce shadows, negative media darken the X-ray film.

Positive contrast media include water-soluble and oil-based substances, as well as suspensions. The development of strongly absorbing positive X-ray contrast media has allowed the visualization of organs and organ systems consisting of radiolucent tissue, such as the gastrointestinal tract, the urinary and biliary tracts, the blood vessels (peripheral, cardiac, and encephalic), the lymph nodes, the lymph vessels, and the bronchi (morphological diagnostics). In addition, some physiological and pathological processes can also be recognized, for example renal function, perfusion of blood vessels and organs, and the permeability of capillaries, e.g., in the brain (functional diagnostics). In some cases, the physiological functions of the organism are also utilized, for example, intravenous adminstration of contrast media for imaging the urinary and biliary tracts, and oral administration for imaging the biliary tract. After intravasal injection or oral absorption, X-ray contrast media, are excreted by the liver and/or kidneys.

More recently, the development of X-ray contrast media has led to the synthesis of nonionic agents of low osmolarity. At the same time a great deal of progress has been achieved in technology. One example is the continued development of computerized tomography and subtraction techniques leading to considerable enhancement of contrast.

Physical Principles. Substances that are suitable for use as positive X-ray contrast media (radiopaque media) should absorb X-rays more strongly than the body tissue. This depends primarily on the mass absorption coefficient of an element at the X-ray wavelengths used for diagnosis. Elements with medium atomic numbers of 50 – 60 (iodine, barium, lanthanum, cerium) are especially suitable; those with high atomic numbers (mercury, lead, bismuth, thorium) are less appropriate.

When a parallel beam of X-rays of wavelength λ passes through a layer of material with thickness d cm and density ρ g/cm^3, the absorption of monoenergetic X-rays is described by the attenuation formula $J = J_0 e^{\mu d}$, where J_0 is the intensity of the radiation in front of the layer, J is the intensity behind the layer, and μ is the linear absorption coefficient. The mass attenuation coefficient is defined as μ/ρ; it varies with the energy of the radiation and with the atomic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_593

number of the element in the material through which the radiation passes. The attenuation formula thus becomes $J = J_0 e^{\mu m/\rho}$, where m denotes the mass of material in grams per square centimeter of surface area normal to the propagation direction. Figure 1 shows X-ray absorption as a function of the X-ray energy.

Types of Contrast Media (Fig. 2). Negative contrast media are weakly absorbing substances and include gases that are inert, readily soluble in blood, and quickly eliminated (e.g., NO, CO_2, air, O_2, and He). In cavities such as the abdominal cavity, joint cavities, or cerebral ventricles, they produce a negative contrast. The contrast difference is, however, slight, so that special techniques have to be used to obtain suitable images.

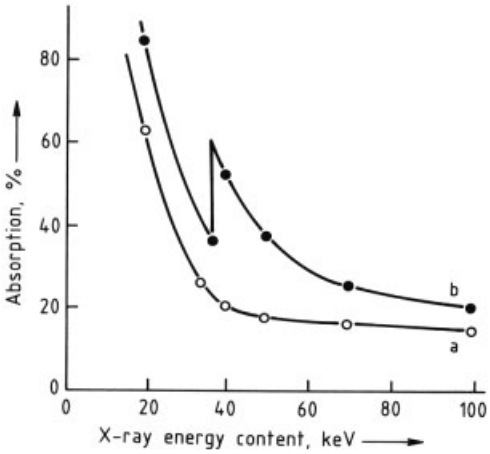

Figure 1. Absorption of X-rays as a function of X-ray energy
a) Absorption by water (corresponds to soft tissues); b) Absorption by an aqueous solution of a contrast medium containing 20 mg I/mLAn X-ray energy of 50 keV corresponds to a tube voltage of ca. 100 kV.

Positive contrast media are divided into two types, according to whether they contain iodine or not. Iodine-substituted organic compounds are more important in practice. Water-soluble and water-insoluble solids, colloids, and oil-based compounds are also employed.

The first X-ray images of the urinary tract and blood vessels were made in 1900 – 1925 with bismuth and thorium compounds, sodium iodide, sodium bromide, lithium iodide, and iodized oils. Because of their high toxicity, the use of these products remained very limited, however. Roentgenography of the urinary and biliary tracts was given a boost when iodinated organic molecules, such as the tetrasodium salt of tetraiodophenolphthalein, were introduced. The iodopyridones, synthesized by Binz and Räth in 1928, represented milestones in the development of iodine-substituted organic contrast media. After clinical investigation by von Lichtenberg and Swick in 1929,
the sodium salt of 2-oxo-5-iodo-1-pyridineacetic acid was brought on the market in 1930 under the trade name Uroselectan. It was considered the "model" of a physiologically compatible, radiopaque renal contrast medium.

Modern iodine-containing contrast media were developed from 1950 onwards for use in urography, angiography, cholecystography, and cholangiography. These compounds contain a benzene ring with three iodine atoms. They also have a salt-forming carboxyl group attached to their nucleus or side chain.

Table 1. Applications of X-ray contrast media

Application	Organs imaged	Administration
Urography (pyelography)	renal pelvis and ureters	intravenous (excretion urography), by urethral catheter (retrograde urography)
Angiography (venography, arteriography, cardiography)	blood vessels	intravenous, intraarterial
Computerized tomography	brain tumors, ventricular system of the brain	intravenous, intraventricular
	spinal cord, nerve roots	intrathecal
Cholecystography, cholangiography	gall bladder and bile ducts	intravenous, oral
Bronchography	bronchi	instillation
Myelography	canal of the spinal cord	intralumbal injection
Lymphography	lymphatic vessels	intralymphatic injection
Gastrography,	gastrointestinal tract	oral, retrograde
Hysterosalpingography	uterus and fallopian tubes	instillation

Contrast media differ in their side chains and in the types of cations present. It was soon realized that many side effects of angiographic contrast media (e.g., pain, vasodilation, hypotension) were due to the high osmolality of these agents. At the beginning of the 1970s, two contrast media of low osmolarity were synthesized:

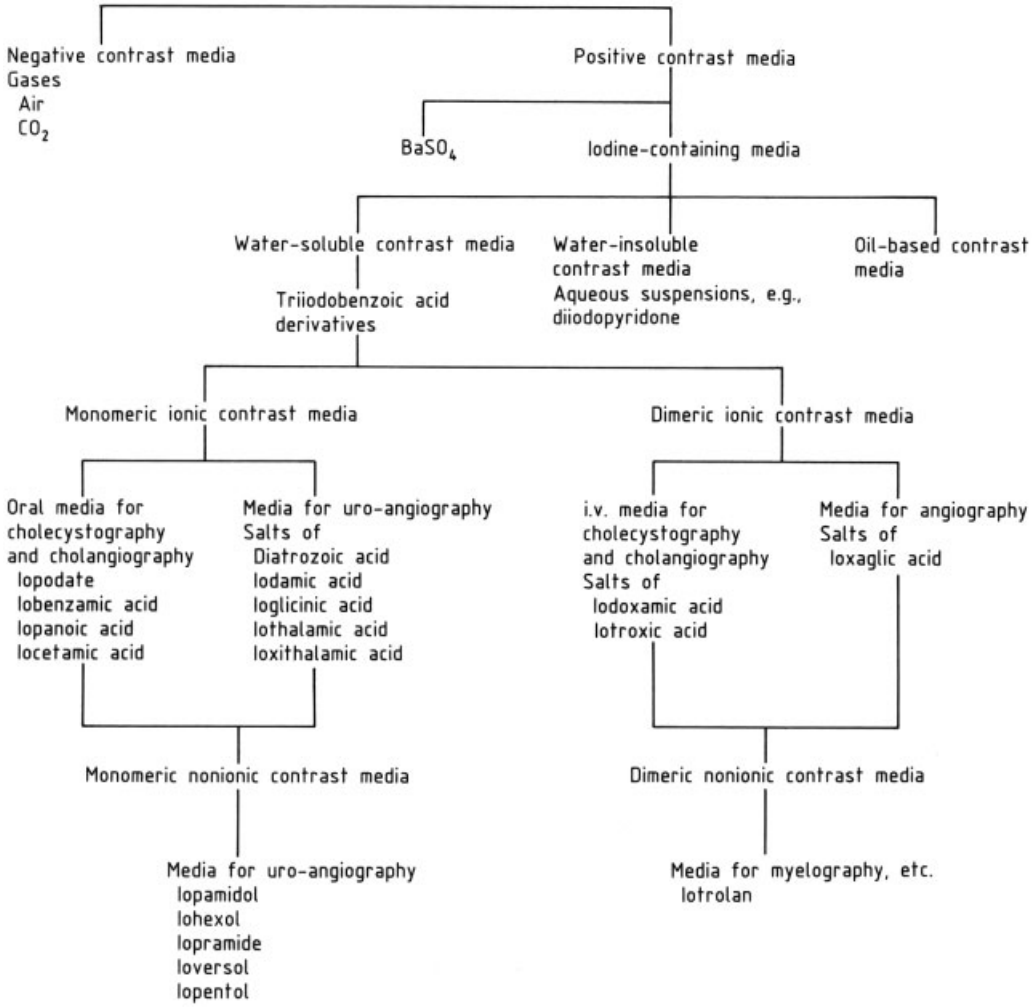

Figure 2. Classification of X-ray contrast media

ioxaglate, a monocarboxylic dimer, and metrizamide, a nonionic monomer. Both of these, however, have considerable drawbacks. Ioxaglate causes a higher incidence of some side effects, such as nausea, vomiting, and urticaria; it is viscous on account of its dimeric structure. Metrizamide is unstable in aqueous solution and must be prepared immediately prior to use.

Further developments led to the nonionic contrast media iopamidol, iohexol, and iopromide, which can all be administered directly as stable solutions.

The first blood-isotonic X-ray contrast medium was developed in the 1980s. Iotrolan, a nonionic dimeric compound, is particularly compatible with nerves and tissues.

Virtually the only iodine-free contrast medium still in use is water-insoluble barium sulfate (in gastrointestinal diagnostics).

Applications. Contrast media have a wide range of uses (Table 1), from the imaging of body cavities to the investigation of parenchymatous structures. Despite extensive research, however, only a limited number of suitable chemical compounds are available.

Requirements of Iodine-Containing X-Ray Contrast Media. The following *physical and chemical requirements* apply:

1) The iodine atoms must be tightly bound to the compound to avoid side effects caused by partially deiodized cleavage products or iodine which has been cleaved off in the body. Substances intended for intravasal administration must also be highly stable to withstand heat sterilization of the solutions.
2) To absorb X-rays efficiently in diagnostics, compounds often have to have a high iodine content (50 – 60 wt %).
3) Compounds for intravenous administration (urography, angiography, cholecystography, and cholangiography) must be highly soluble in water. Commercial preparations contain 300 – 850 mg of the contrast material per milliliter of solution. The iodine content is 140 – 380 mg/mL. Ideally, the solutions should be stable and have a low viscosity, a weak osmotic effect, and a neutral pH.
4) Oil-based contrast media (for lymphography) must have a low viscosity.

Physiological and pharmacological requirements are as follows:

1) X-ray contrast media must have very low general and local toxicities (irritant effect) because they are used in higher dosages than other pharmaceuticals. They must have a minimum inherent pharmacodynamic action (e.g., effect on blood pressure or cardiac output) and produce minimal side effects (vertigo, nausea, vomiting, headache).
2) The contrast medium must accumulate quickly and selectively in the organ to be visualized, and must have an adequate residence time. This means that substances administered intravenously or orally must be appropriately eliminated from the organism, i.e., either via the kidneys (urography) or via the liver (cholecystography and cholangiography). In addition to these techniques that involve physiological, functional contrast enhancement, body cavities and blood vessels can also be filled by the direct administration of X-ray contrast media (angiography, bronchography, myelography, lymphography, and gastrography). The requirement for selective accumulation does not apply to these applications.
3) X-ray contrast media must be excreted from the organism rapidly, with as little modification as possible.

Pharmacological Testing of X-Ray Contrast Media. Primary screening is first carried out to assign the medium to one of the indication areas (urography, angiography, cholecystography, cholangiography). The dependence of its absorption on the mode of administration and its elimination characteristics (ratio of excretion by kidneys and in bile) are determined in pharmacokinetic studies (using tracer-labeled sub-stances in rats). A further criterion is the acute toxicity (LD_{50}) after intravenous administration; an $LD_{50} > 12$ g I/kg body weight indicates suitability for urography and angiography; > 5 g I/kg body weight, for cholecystography and cholangiography (intravenous administration); > 0.3 g I/kg body weight, for cholecystography and cholangiography (oral administration). Studies are performed on dogs to compare the medium under test against standard preparations with regard to the sharpness and density of the radiographic contrast, time and duration of imaging, and reproducibility of results; these tests supplement those performed in screening. Other series of trials are carried out to investigate local compatibility (tissue reactions), hemodynamics, and pharmacokinetics. Further studies are designed to provide a more detailed description of the pharmacology (extended hemodynamics, studies of kidney and liver function, vessel tolerance, cerebral tolerance, effect on the hematopoietic system, effect on endogenous metabolism). Toxicity studies are also performed on two animal species for several weeks as required by the health authorities.

2. Urography, Angiography, and Computerized Tomography

Pharmacological Principles. Roentgenographic imaging of the kidneys (pelvis and calices) and urinary tract (ureters and bladder) requires substances whose nephrotropic properties cause them to be excreted quickly via the kidneys after intravenous administration. Urographic contrast media bind only slightly to

proteins. Due to their small size the molecules are effectively filtered by the glomeruli.

The contrast media used for imaging the kidneys can also be employed for angiography and computerized tomography. Computerized tomography is an X-ray imaging technique which produces transsectional pictures of the body. Since the high flow rate in the blood vessels results in rapid dilution, a high concentration and injection volume of the contrast medium is necessary to obtain adequate opacification. The organ and systemic toxicity must therefore be extremely low, local tissue compatibility must be good, and the medium must be as pharmacodynamically inert as possible. Excretion requirements are the same in angiography as in urography.

Preparations (see Table 2).

Ionic Contrast Media. In 1950, WALLINGFORD and coworkers discovered that the sodium salt of *acetrizoic acid* (3-acetamido-2,4,6-triiodobenzoic acid) had a low toxicity, was quite soluble in water, and was excreted in the urine. The compound was introduced in the United States by Mallinckrodt under the trade name Urokon. In 1954, it was replaced by *amidotrizoic acid* (diatrizoic acid) which was independently developed by Schering and Sterling – Winthrop. It offered a lower toxicity and a better neurotoxic and hemodynamic compatibility. Diatrizoate became the world's most widely used contrast medium for urography and angiography. A number of monomeric ionic compounds are derived from diatrizoic acid (e.g., metrizoic acid, iothalamic acid, iodamide, ioxithalamic acid, and ioglicic acid).

Ioxaglate is a monocarboxylic dimer consisting of two triiodobenzene rings connected by a chain. This contrast medium has become important in peripheral angiography because it causes little vessel pain. Despite its low osmolality, however, it does not have the expected general and neural tolerance as compared to nonionic contrast media. The osmolality is just one of the many factors that elicit undesired reactions to contrast media.

Nonionic Contrast Media. The *monomeric nonionic contrast media* (iopamidol, iohexol, iopromide, ioversol, and iopentol, Table 3) exhibit better neural compatibility because they are hydrophilic, do not bear charges, and do not require the presence of cations (sodium, meglumine) to form solutions. They are also less likely to cause general side effects, such as nausea, vomiting, and allergic reactions. In addition, they are largely free of side effects caused by high osmotic pressure (e.g., pain, endothelial injury, disturbance of the blood – brain barrier, thromboses, bradycardia in cardioangiography, and increased pressure in the pulmonary circulation).

The nonionic contrast media differ in osmolality and chemotoxicity. The term chemotoxicity includes various properties of the molecule, such as the ability to liberate histamine from most cells, to influence the complement and coagulation system, to change the erythrocyte shape, and to bind to proteins.

Iotrolan is the first *dimeric nonionic contrast medium.* Being isotonic with cerebrospinal fluid and blood at high iodine concentrations (ca. 300 mg/mL), it offers excellent compatibility with neural tissue, along with good general tolerance. After subarachnoidal injection, only a very small fraction diffuses into the interstitial space of the brain parenchyma. The residence time there is short, so irritation of neural structures is unlikely. The high iodine content of the molecule means that not only the entire subarachnoidal space but also other body cavities can be imaged (arthrography, hysterosalpingography, and endoscopic retrograde cholangiopancreatography). The application has been extended to angiography, urography, and computerized tomography enhancement as well.

Structure – Activity Relationships. All the contrast media used in urography and angiography are based on 2,4,6-triiodobenzoic acid:

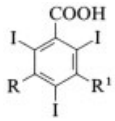

The individual structural groups perform special functions. The three iodines with their symmetrical configuration are responsible for the contrast density (a different configuration would make the substances less stable). The carboxyl group forms salts and therefore provides good solubility in water. The substituents R and R^1 also affect solubility but primarily in-

Table 2. Contrast media for urography and angiography

Formula	Chemical name, CAS registry no.	Empirical formula	Generic name	Trade name	Company	Year of introduction
Monomeric ionic contrast media						
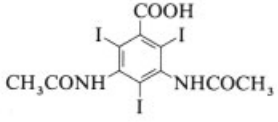	3,5-bis(acetamido)-2,4,6-triiodobenzoic acid [*117-96-4*]	$C_{11}H_9I_3N_2O_4$	amidotrizoic acid (diatrizoic acid)	Urografin, Hypaque	Schering AG	1954
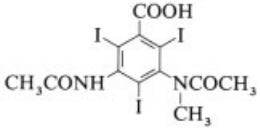	3-acetamido-2,4,6-triiodo-5-(*N*-methylacetamido)benzoic acid [*1949-45-7*]	$C_{12}H_{11}I_3N_2O_4$	metrizoic acid	Isopaque	Nyegaard	1962
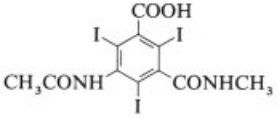	5-acetamido-2,4,6-triiodo-3-methylisophthalamic acid [*2276-90-6*]	$C_{11}H_9I_3N_2O_4$	iothalamic acid	Conray	Mallinckrodt	1962
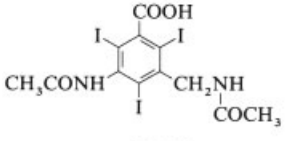	3-acetamido-5-acetamidomethyl-2,4,6-triiodobenzoic acid [*440-58-4*]	$C_{12}H_{11}I_3N_2O_4$	iodamide	Uromiro	Bracco	1965
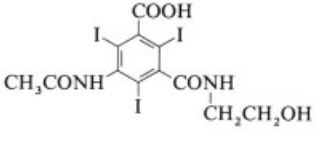	5-acetamido-3-(2-hydroxyethyl)-2,4,6-triiodoisophthalamic acid [*28179-44-4*]	$C_{12}H_{11}I_3N_2O_5$	ioxithalamic acid	Tetebrix	Guerbet	1972
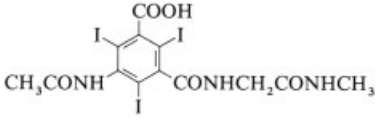	5-acetamido-3-(2-methylamino-2-oxoethyl)-2,4,6-triiodoisophthalamic acid [*49755-67-1*]	$C_{13}H_{12}I_3N_3O_5$	ioglicinic acid	Rayvist	Schering AG	1978

Table 2. (Continued)

Dimeric ionic contrast media						
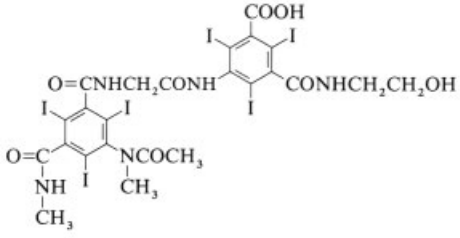	*N*-(2-hydroxyethyl)-2,4,6-triiodo-5-{2-[2,4,6-triiodo-3-*N*-(methylacetamido)-5-(methylcarbamoyl)benzamido]acetamido}-isophthalamic acid [*59017-64-4*]	$C_{24}H_{21}I_6N_5O_8$	ioxaglic acid	Hexabrix	Byk Gulden, Guerbet	1979
Monomeric nonionic contrast media						
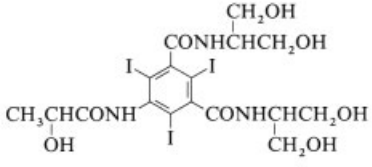	5-(α-hydroxypropionylamino)-2,4,6-tri-iodoisophthalic acid di(1,3-dihydroxyiso-propylamide) [*60166-93-0*]	$C_{17}H_{22}I_3N_3O_3$	iopamidol	Solutrast	Byk Gulden	1981
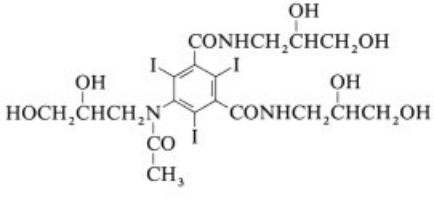	*N,N'*-bis(2,3-dihydroxypropyl)-5-[*N*-(2,3-dihydroxypropyl)acetamido]-2,4,6-tri-iodoisophthalamide [*66108-95-0*]	$C_{19}H_{26}I_3N_3O_9$	iohexol	Omnipaque, Nycamed	Schering AG, Winthrop	1982
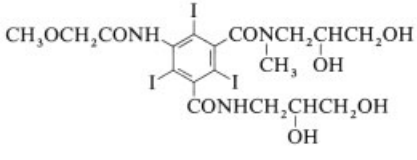	*N,N'*-bis(2,3-dihydroxypropyl)-2,4,6-triiodo-5-(2-methoxyacetamido)-*N*-methylisophthalamide [*73334-07-3*]	$C_{18}H_{24}I_3N_3O_8$	iopromide	Ultravist	Schering AG	1985

Table 2. (Continued)

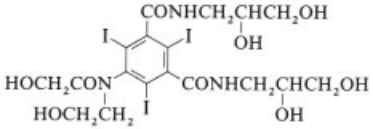	DL-5-[*N*-(2-hydroxyethyl)hydroxy-acetamido]-2,4,6-triiodoisophthalic acid di(2,3-dihydroxypropylamide) [*87771-40-2*]	$C_{18}H_{24}I_3N_3O_9$	ioversol	Optiray	Mallinckrodt	1988
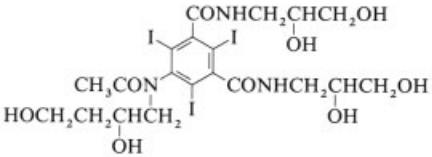	*N,N'*-bis(2,3-dihydroxypropyl)-5-[*N*-(2-hydroxy-3-methoxypropyl)acetamido 2,4,6-triiodoisophthalamide [*89797-00-2*]	$C_{20}H_{13}I_3N_3O_9$	iopentol	Imagopaque	Nycomed	1991
Dimeric nonionic contrast media						
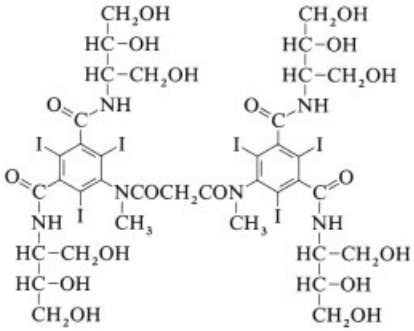	5,5'-[malonylbis(methylimino)]-bis [*N,N'*-bis(2,3-dihydroxy-1-(hydroxymethyl)propyl-2,4,6-triiodoisophthalamide [*79770-24-4*]	$C_{37}H_{48}I_6N_6O_{18}$	iotrolan	Isovist	Schering AG	1988

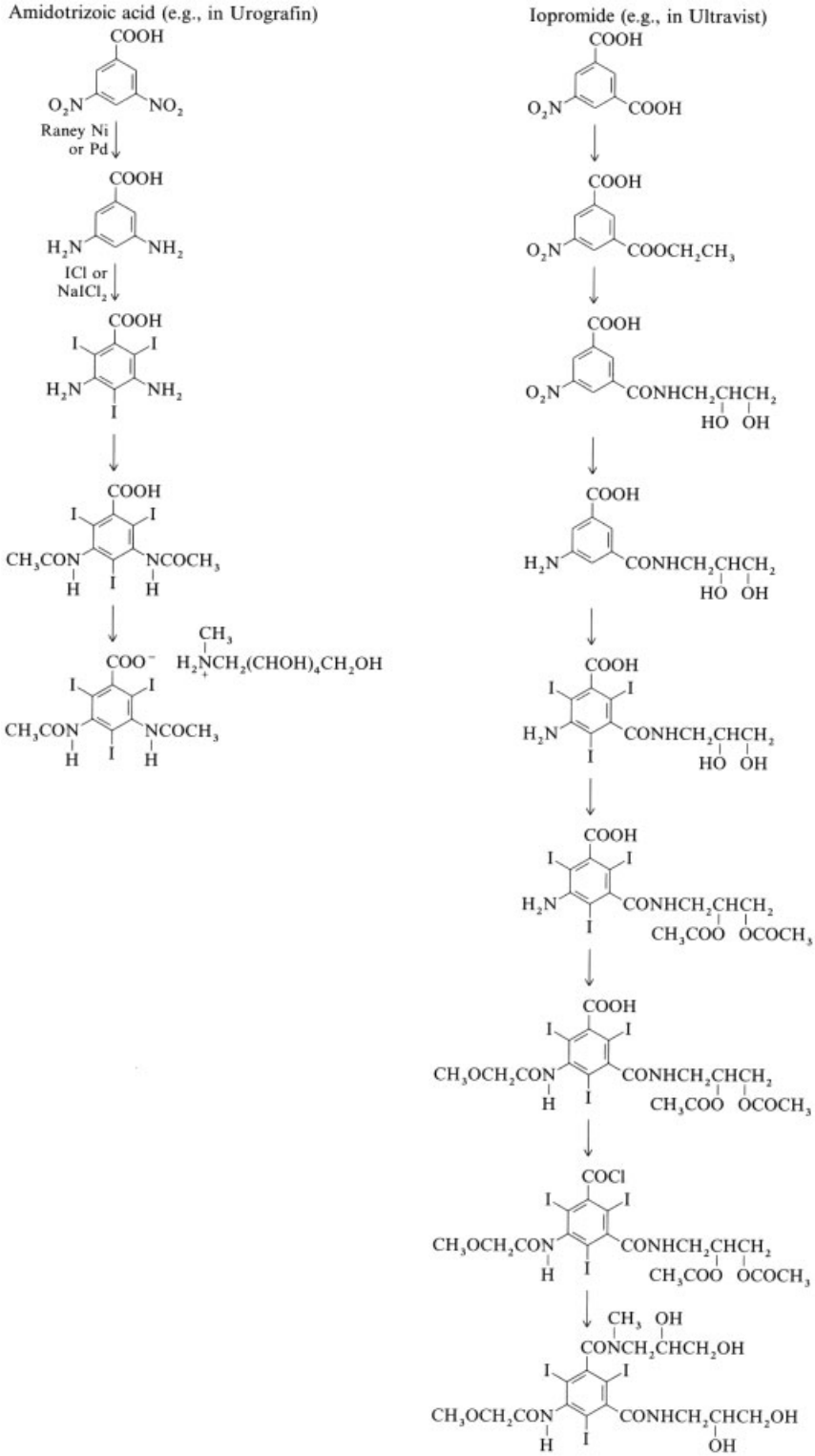

Figure 3. Synthesis routes for an ionic and a nonionic contrast medium

Table 3. Nonionic contrast media for intraversal administration

Contrast medium	Mean osmolality (300 mg I/mL, 37°C), m · osm/kg H_2O	Viscosity at 37°C, mPa · s		Protein binding at 1.2 mg I/mL plasma, %
		300 mg I/mL	350 mg I/mL	
Iopromide	586	4.6	9.5 *	0.9 ± 0.2
Iopamidol	653	4.5	9.5 *	2.9 ± 0.2
Iohexol	667	5.7	10.5	1.5 ± 0.3
Ioversol	661	5.5	9.0	1.6 ± 0.9
Iopentol	683	6.5	12.0	1.9 ± 0.6

* 370 mg I/mL

fluence toxicity, the rate of renal excretion, and physical parameters such as viscosity and osmotic activity. R must always be an acylamino or acylalkylamino group, whereas R^1 can be acylamino, alkylaminocarbonyl, or acylaminomethyl. These requirements set narrow limits on possible chemical variations, especially as higher acyl and alkyl residues have detrimental effects on tolerance, excretion, viscosity, and circulatory system compatibility.

Nonionic, water-soluble contrast media have the same basic structure, but R and R^1 are usually large, hydrophilic side chains and the carboxyl group is reacted with an amine so that the molecules are electrically neutral in solution.

Synthesis. The starting products for the synthesis of water-soluble X-ray contrast media are iodine and derivatives of 3,5-dinitrobenzoic or 5-nitroisophthalic acid. Ionic contrast media can be prepared in fewer synthesis steps than nonionic ones (see Fig. 3).

3. Cholecystography and Cholangiography

The gall bladder and bile ducts can be imaged by oral and intravenous administration of a hepatotropic X-ray contrast medium. When administered orally, the agent is absorbed in the small intestine. In the pH range of the intestine, some of it is in the undissociated, lipophilic form and can therefore pass through the lipid-containing membranes of the intestinal epithelium into the blood. Bound to plasma proteins, it passes through the portal vein into the liver. Oral contrast media for cholangiography and cholecystography are monomers. They initially lack the higher molecular size needed for biliary excretion but gain it after conjugation with glucuronic acid. After biliary excretion and subsequent partial enteric absorption, the contrast medium is eliminated via the kidneys, primarily in the form of a glucuronide (Fig. 4).

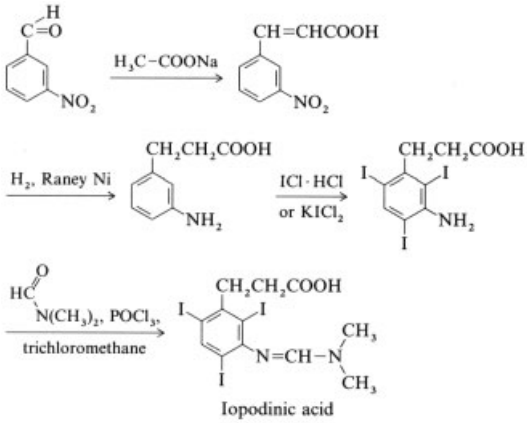

The synthesis of iopodinic acid.

The structural requirements relating to absorption in the small intestine do not apply to contrast media that are administered intravenously. These agents must have good solubility and lower toxicity. Unlike orally administered preparations, an intravenous medium reaches all organs and is then concentrated in the gall bladder by a factor of 30 – 100 after active secretion via the liver cells. From the gall bladder, it passes into the small intestine, and is then eliminated with the feces. There is no enterohepatic circulation because absorption is low. 10 – 20 % of the quantity administered is excreted via the kidneys.

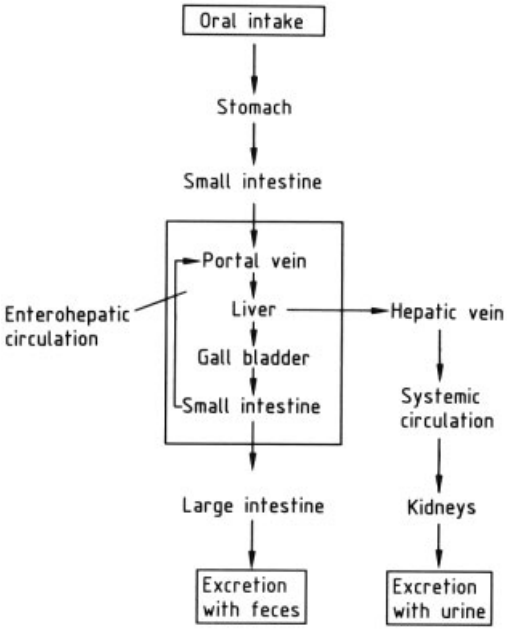

Figure 4. Schematic diagram of absorption and excretion of oral contrast media in cholecystography and cholangiography

Oral Contrast Media. Oral media for cholangiography and cholecystography show a wide range of chemical variation but a high degree of similarity between the individual components. Many of the compounds synthesized can be described by the following formula:

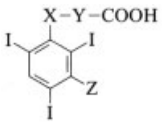

The symmetrical configuration of iodine atoms is responsible for the high contrast density, as in the case of media for uro-angiography (Chap. 2). The carboxyl group aids in the formation of water-soluble salts. The structural element X can be CH_2, O, S, CONH, SO_2, NH, $N(COCH_3)$, H_3C-NHCO, or some other group; this component governs the toxicity and also contributes to the solubility. The performance of these media is determined by the lipophilic aliphatic residue Y. Substituent Z, which can be OH, NH_2, or substituted NH_2, influences the absorption, excretion, and toxicity. Good enteral absorption requires a balance between lipophilic and hydrophilic groups.

Among the many oral contrast media, the following are chiefly used in diagnostics: iopanoic acid, iobenzamic acid, iopodinic acid, and iocetamic acid (Table 4).

Iopanoic acid is excreted with the feces, chiefly in the form of a nonabsorbable glucuronide ester. Iobenzamic acid is excreted mainly with the feces in the form of a metabolite. More than 50 % of iopodinic acid leaves the body via the kidneys, in the form of the glucuronide, whereas iocetamic acid is almost exclusively eliminated via the kidneys.

Intravenous Contrast Media. A key advance in the roentgenographic imaging of the intra- and extrahepatic bile ducts resulted in 1953 when Langecker, Harwart, and Junkmann introduced *iodipamide* [*606-17-7*] [*N,N′*-adipyl-bis(3-amino-2,4,6-triiodobenzoic acid); the methylglucamine salt is marketed under the trade name Biligrafin, Schering AG, Berlin].

Iodipamide

Iodipamide is the prototype of gall bladder and bile-duct contrast media for intravenous administration. It is a dimeric diacid with no side chains in the 5,5′-positions of the triiodobenzoic acid moieties. It is excreted without chemical modification, primarily in the bile.

Iotroxic acid [*51022-74-3*] *and iodoxamic acid* [*31127-82-9*], the modern intravenous contrast media for cholangiography and cholecystography, are less strongly bound to albumin than iodipamide and hence display better tolerance. These contrast media also have a higher affinity for the carrier proteins of the liver cells; they are therefore excreted more quickly into the biliary system while giving a high contrast density.

Table 4. Oral contrast media for cholecystography and cholangiography

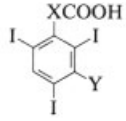

Substituents X	Substituents Y	Chemical name, CAS registry no.	Empirical formula	Generic name	Trade name	Company	Year of introduction	LD_{50} (intravenous, rat), g/kg
$-CH_2CH(C_2H_5)-$	$-NH_2$	2-ethyl-3-(3-amino-2,4,6-triiodophenyl)propionic acid [*96-83-3*]	$C_{11}H_{12}I_3NO_2$	iopanoic acid	Tetepaque	Sterling–Winthrop	1952	ca. 0.39
$-CO-N(C_6H_5)-CH_2CH_2-$	$-NH_2$	*N*-(3-amino-2,4,6-triiodobenzoyl)-*N*-phenyl-β-alanine [*3115-05-7*]	$C_{16}H_{11}I_3N_2O_3$	iobenzamic acid	Osbil	Österreichische Stickstoffwerke	1962	ca. 0.41
$-CH_2-CH_2-$	$-N=CH-N(CH_3)_2$	3-(3-dimethylaminomethyleneamino-2,4,6-triiodophenyl)-propionic acid [*5587-89-3*]	$C_{12}H_{13}I_3N_2O_2$	iopodinic acid	Biloptin	Schering AG	1958	ca. 0.43
$-N(CO-CH_3)-CH_2CH(CH_3)-$	$-NH_2$	3-[*N*-acetyl(3-amino-2,4,6-triiodoanilino-2-methylpropionic acid [*16034-77-8*]	$C_{12}H_{13}I_3N_2O_3$	iocetamic acid	Cholebrin	Dagra	1969	ca. 0.8

4. Bronchography

Bronchography is relatively difficult, time-consuming, and uncomfortable for the patient. After local anesthesia, a special catheter is used to introduce the X-ray contrast medium into the bronchial region to be imaged. An optimal medium, with high contrast density, low diffusibility, good adhesion and miscibility with the bronchial secretion, and rapid, quantitative absorption, has not been developed. Trials of a simple, noninvasive inhalation technique for administering the X-ray contrast medium (inhalation bronchography) have been unsuccessful. The reason is that it has not proved possible to generate a well tolerated and highly concentrated aerosol of contrast media powders or solutions.

Bronchography is becoming less popular. It is being replaced by bronchoscopy and less intrusive procedures, such as computerized axial tomography, that often supply more information.

Preparations. For bronchography with a fiber-optic bronchoscope, the nonionic dimer iotrolan-300 (Table 2) is greatly prefered over the previously used contrast medium, dionosil. Iotrolan can be injected directly through the bronchoscope; it does not have a detrimental effect on the bronchoscopic image or on subsequent studies.

5. Myelography

The nonionic dimeric contrast medium iotrolan has been available since 1988, and is very satisfactory for the visualization of the subarachnoidal space (Table 2). The contrast medium can be injected directly, and mixes well with the cerebrospinal fluid. It is isotonic and consequently better tolerated than hypertonic nonionic contrast media. Iotrolan is completely absorbed from the subarachnoidal space and is excreted via the kidneys in chemically unchanged form. Its high iodine content results in better definition of detailed neural structures. Other nonionic contrast media such as iopamidol and iohexol ase also suitable for myelography.

6. Lymphography

Noninvasive techniques – ultrasound and whole-body computerized tomography, (CT) – have largely replaced lymphography, which has been practiced since 1952. Lymphography can image finer structures, so it tends to find use only when uncertain or negative results are obtained.

In order to visualize the lymph nodes, the contrast medium must be slowly injected into an exposed, cannulated lymph vessel using an infusion pump. The lymph nodes in the thorax become visible after ca. 1 – 2 h. A lymph duct can only be detected after the subcutaneous injection of a lymphotropic dye (e.g., methylene blue). The method of "direct lymphography" (intralymphatic injection) is time- consuming and difficult. Efforts to develop "indirect lymphography," with subcutaneous, intramuscular, intracutaneous, or submucosal injection of an X-ray contrast medium, have only yielded limited success because the media studied were not sufficiently lymphotropic.

Contrast media for lymphography must satisfy the same requirements as other substances for endovasal administration (high iodine content, good local compatibility, low toxicity, and low viscosity) and must also have a well-defined retention time in the lymph node. Lymphographic agents must also be quickly and completely excreted, either as such or after degradation. Temporary storage and retention are desirable in media used for lymphography, as opposed to those used for imaging other vessels. These properties are not, however, essential for X-ray contrast media employed in peripheral lymphangiography.

Preparations. All commercially available contrast media for direct lymphography are oil-based substances such as lipiodol (iodized oil). These meet the requirements for a well-defined retention time and good contrast enhancement. They are not, however, fully satisfactory with regard to viscosity, excretion rate, stability, and compatibility. Pulmonary embolisms can easily occur when lymphangiography is carried out with oil-based contrast media because the lymphatic system is connected to the venous system. Water-soluble contrast media have not yet proved successful because the intralymphatically injected media diffuse out of the lymphatic

vessels too quickly and fail to give adequate contrast enhancement, especially of the lymph nodes.

In special cases, lymphatic vessels and local lymph nodes can also be imaged by intracutaneous injection of the water-soluble, nonionic contrast medium iotrolan.

7. Gastrointestinal Contrast Media

Since the early days of roentgenography, barium sulfate has maintained its position as the optimal medium for imaging the gastrointestinal tract. Numerous preparations are available with standardized particle sizes (0.3 – 1 µm). They contain additives (e.g., tragacanth, methyl cellulose, glycerol, mineral oil) to improve the viscosity, adherence, and wetting power and to reduce the sedimentation rate.

There are, however, some risk groups of patients where barium sulfate should not be used (e.g., risk of aspiration or of perforation). Watersoluble X-ray contrast media are also employed for imaging the stomach, duodenum, and colon. An example is diatrizoate, which is available as a 76 % aqueous solution of a sodium – methylglucamine salt mixture, with flavorings and wetting agents (Gastrografin, Schering). If it leaks into the peritoneum, it is readily absorbed and excreted.

Nonionic, water-soluble contrast media are being increasingly employed in gastrointestinal diagnostics because of their good compatibility. Few side effects are produced if the medium passes into the lung or peritoneal cavity. Since these media have a low osmolality they produce much smaller changes in the water and electrolyte balance than do ionic water-soluble agents.

Iotrolan (Isovist-300) is employed in children and, in special cases, in adults. It shows no dilution due to osmotic effects and is the water-soluble contrast medium with the most pleasant taste.

8. References

1. R. Barke: *Röntgenkontrastmittel, Chemie – Physiologie – Klinik,* VEB Georg Thieme, Leipzig 1970.
2. H. Dorn: "Röntgenkontrastmittel," *Pharmazie* **12** (1957) 315, 415, 499.
3. M. Gutcho: *X-Ray Contrast Agents,* Noyes Data Corp., Park Ridge, N. J. 1971.
4. G. Hecht, C. Gloxhuber: "Kontrastmitteldarstellung; Chemie, Pharmakologie und Toxikologie der gebräuchlichsten Kontrastmittel" in L. Diethelm et al. (eds.): *Handbuch der Medizinischen Radiologie,* Springer Verlag, Berlin 1967, p. 518.
5. J. O. Hoppe: "X-Ray Contrast Media," *Med. Chem. (Wiley)* **6** (1963) 290.
6. P. K. Knoefel: *Radiopaque Diagnostic Agents,* Ch. C. Thomas Publ., Springfield 1961.
7. P. K. Knoefel (ed.): *Radiocontrast Agents,* Pergamon Press, Oxford 1971, vol. 1, 2.
8. X. Perlia, J. Wellauer: "Röntgenkontrastmittel," *Subsidia Med.* **14** (1962) 31, 69.
9. *Ullmann,* 3rd ed., **9,** 139.
10. A. T. Shockman: "Radiologic Diagnostic Agents," *Top. Med. Chem.* **1** (1967) 381.
11. U. Speck: *Röntgenkontrastmittel, übersicht, Anwendung und pharmazeutische Aspekte,* Springer Verlag, Berlin 1991.
12. U. Speck, W. Mutzel, H.-J. Weinmann: "Chemistry, Physiochemistry and Pharmacology of Known and New Contrast Media for Angiography, Urography and CT Enhancement" in V. Taenzer, E. Zeitler (eds.): *Contrast Media in Urography, Angiography and Computerized Tomography,* Thieme, Stuttgart 1983, pp. 2 – 10.
13. W. H. Shehadi, G. Toniolo: "Adverse Reaction to Contrast Media," *Radiology* **137** (1980) 299 – 302.
14. H. Katayama: "Clinical Survey on Adverse Reaction of Iodinated Contrast Media," *International Symposium on Contrast Media,* **section II, part 2,** Tokyo 1982, pp. 144 – 151.

Radiopharmaceuticals → Diagnostic Reagents
Radium → Radionuclides
Radon → Radionuclides
Rapeseed Oil → Fats and Fatty Oils

Rare Earth Elements

Cerium Mischmetal, Cerium Alloys, and Cerium Compounds is a separate keyword

IAN MCGILL, Johnson Matthey Technology Centre, Reading, England

1. Introduction . . . 2
2. Mineralogy, Abundance, Occurrence . . . 2
3. Properties . . . 6
3.1. Properties of the Nuclei . . . 6
3.2. Properties of the Atoms and Ions 7
3.2.1. Electronic Configuration, Position in the Periodic Table . . . 7
3.2.2. Oxidation States, Atomic and Ionic Radii . . . 8
3.2.3. Magnetic and Spectral Properties . 9
3.2.4. Bonding, Coordination Numbers . . 10
3.3. Other Chemical and Physical Properties . . . 10
3.4. Miscibility and Alloying Behavior 10
3.5. Mechanical Workability . . . 12
4. Digestion of Rare Earth Ores . . . 12
4.1. Wet Chemical Digestion; Fusion . 12
4.1.1. Monazite . . . 13
4.1.2. Bastnaesite . . . 13
4.1.3. Other Ores . . . 14
4.2. Direct Chlorination at High Temperature . . . 14
5. Separation . . . 15
5.1. Principles of Separation . . . 15
5.2. Separation by Classical Methods 16
5.2.1. Fractional Crystallization . . . 16
5.2.2. Fractional Precipitation . . . 16
5.2.3. Separations Based on Oxidation State Changes . . . 16
5.3. Separation by Ion Exchange . . . 17
5.3.1. Ion Exchange with Chelating Agents 17
5.3.2. Separation Process . . . 18
5.3.3. Industrial Processes . . . 19
5.3.4. Disadvantages of Ion Exchange . . 19
5.4. Separation by Liquid – Liquid Extraction . . . 19
5.4.1. Theoretical Basis . . . 19
5.4.2. Extractants . . . 21
5.4.2.1. Neutral Extractants . . . 21
5.4.2.2. Acidic Extractants . . . 22
5.4.2.3. Amines and Quaternary Ammonium Salts as Extractants . . . 24
5.4.2.4. Synergistic Effects . . . 25
5.4.3. Industrial Liquid – Liquid Extraction . . . 25
5.4.3.1. Separation of Europium Oxide . . . 26
5.4.3.2. Separation of Yttrium Oxide . . . 26
5.4.3.3. Separation of Scandium . . . 30
6. Production of the Metals . . . 30
6.1. Fused Salt Electrolysis . . . 30
6.2. Metallothermic Reduction . . . 31
6.3. Purification . . . 32
7. Analysis . . . 33
8. Compounds . . . 33
8.1. Hydrides . . . 33
8.2. Oxides, Hydroxides, Peroxides, Salts of Inorganic Oxoacids, Double Salts . . . 34
8.3. Halides . . . 35
8.4. Chalcogenides . . . 36
8.5. Nitrides . . . 36
8.6. Carbides . . . 36
8.7. Complexes; Organic Compounds 37
9. Uses . . . 37
10. Economic Aspects . . . 40
11. Toxicology . . . 41
12. References . . . 42

Based on the article Seltene Erden from Ullmann's 4th edition, written by Jaques Helgorsky, Alain Lévêque, Thomas Petzel, and Klaus Reinhardt.

1. Introduction

The metallic elements known collectively as the rare earths (better: rare earth metals or rare earth elements, abbreviated to RE) are listed in Table 1. General reviews are given in [1–6]. Originally, the name was only used for the sesquioxides, RE_2O_3, which are extraordinarily similar to one other in their chemical and physical properties and are therefore difficult to separate. Within the rare earth group, the elements scandium, yttrium, and lanthanum differ in their atomic structure (see Table 2) from the elements

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_607

cerium to lutetium (the lanthanides, Ln). The term lanthanoids has been proposed by IUPAC to include lanthanum and the lanthanides [7], but this usage is not yet universal. In the fields of mineralogy and geochemistry (see Chap. 2), the term cerium earths is used for the elements La and Ce to Eu, and yttrium earths for Y and Gd to Lu. Scandium occupies a special position with respect to this classification and its other properties, and therefore does not belong to either of these groups. The terms terbium earths for Eu, Gd, and Tb; erbium earths for Dy, Ho, Er, and Tm; and ytterbium earths for Yb and Lu are now hardly used. The rare earth elements always occur in nature in association with each other (see Chap. 2). The isolation of groups of rare earth elements or of individual elements requires costly separation and fractionation processes owing to the great similarity of the chemical and physical properties of their compounds (see Chap. 5), which explains why the history of their discovery has extended over such a long period.

The discovery of the rare earths is summarized in Table 3, with the exception of the radioactive element promethium (see Section 3.1). Further details are given in [8]. The word "rare", when used to describe this group of elements, originates from the fact it was thought that the rare earth elements could only be isolated from very rare minerals. Considering the abundance of the rare earth elements in the Earth's crust (Chap. 2), the term rare is inappropriate.

2. Mineralogy, Abundance, Occurrence

The rare earth elements are lithophilic and are therefore concentrated in oxidic compounds such as carbonates, silicates, titanotantaloniobates, and phosphates, or form minerals of these types on their own [6, 9] (see also [1, parts A 4, A 5]). The formation of the minerals is influenced by variations in ionic radii, crystallochemical factors such as coordination number, basicity, potential for isomorphic replacement, tendency to form complexes, and differences in oxidation states [10], (see also [2, vol. 3, pp. 1 – 80]). These factors lead to the division of the rare earth elements into three groups, according to the mineralization process (REO denotes the rare earth sesquioxides):

1) Minerals that contain lanthanum to neodymium, samarium, and europium, and in which cerium and, in some cases also lanthanum or neodymium, are present as the main components (cerium earths, Ce...). Typical examples of this group include bastnaesite (Ce...)FCO_3 (max. 75 % REO content), monazite (Ce...)PO_4 (max. 65 % REO content), and allanite (Ca, Ce...)$_2$(Fe, Al)$_3$ $(SiO_4)_3$(OH) (max. 28 % REO content).
2) Minerals with gadolinium to lutetium and yttrium as main components (yttrium earths, Y...). Typical examples of this group are xenotime (Y...)PO_4 (max. 62 % REO content) and gadolinite (Y...)$_2FeBe_2Si_2O_{10}$ (max. 48 % REO content). Typical compositions of bastnaesite, monazite, and xenotime from various mineral resources are given in Table 4.
3) Complex minerals in which both yttrium earths and cerium earths can be represented in the same mineral type, either of these groups of rare earths being the main one. Minerals of this type are oxidic ores containing titanium, niobium, tantalum, uranium, and thorium; for example:

Euxenite	(Y..., Ce...)(Nb, Ta, Ti)$_2O_6$
Samarskite	(Y..., Ce...)$_4$(Nb, Ta, Ti)$_2O_6$
Fergusonite	(Y...)(Nb, Ti, Ta)O_4
Betafite	(U, Ca, Y..., Ce...)$_2$(Nb, Ta,Ti)$_2O_6$(OH)

The minerals of the first two groups occur in pegmatites and metamorphic gneisses, as well as hydrothermal and pneumatolytic veins, and in skarns and carbonatites. The minerals of the third group are mainly found in pegmatites.

Useful concentrations (up to 5 %) of rare earth elements occur in apatite Ca_5(F, Cl, OH)$(PO_4)_3$ (see typical composition in Table 4), and up to 10 % in pyrochlor (Na, Ca, Ce...)$_2$(Nb, Ta, Ti)$_2$(O, OH, F)$_7$, and loparite (Na, Ca, Ce...)$_2$(Nb, Ta, Ti)$_2O_6$ in the Kola Peninsula (CIS), and also in most uranium minerals as trace substituents.

Scandium occurs in trace amounts in most rare earth minerals along with the true rare earth elements (e.g., as a substituent ion of the same valency). However, the amounts are very small. Thortveitite, $Sc_2Si_2O_7$; sterretite, $ScPO_4 \cdot 2 H_2O$; and kolbeckite (Sc, Be, Ca) $(SiO_4, PO_4) \cdot 2$

Table 1. Atomic numbers, relative atomic masses, and natural nuclides of the rare earth elements

Element	Symbol	Atomic number	M_r	Mass numbers of the nuclides (abundance in %) ∗
Scandium	Sc	21	44.9559	45 (100)
Yttrium	Y	39	88.9059	89 (100)
Lanthanum	La	57	138.9055	138 (0.089; EC, 1.1×10^{11} a), 139 (99.91)
Cerium	Ce	58	140.12	139 (0.139), 138 (0.250), 140 (88.48), 142 (11.07; β^-, 5×10^{15} a)
Praseodymium	Pr	59	140.9077	141 (100)
Neodymium	Nd	60	144.24	142 (27.11), 143 (12.17), 144 (23.85; α, 1×10^{15} a), 145 (8.30), 146 (17.22), 148 (5.73), 150 (5.62)
Promethium	Pm	61	(145)	only radioactive isotopes
Samarium	Sm	62	150.4	144 (3.09), 147 (14.97; α, 1.06×10^{11} a), 148 (11.24; α, 1.2×10^{13} a), 149 (13.83; α, 4×10^{14} a), 150 (7.44), 152 (76.72), 154 (22.71)
Europium	Eu	63	151.96	151 (47.82), 153 (52.18)
Gadolinium	Gd	64	157.25	152 (0.20), 154 (2.15), 155 (14.73), 156 (20.47), 157 (15.68), 158 (24.87), 160 (21.90)
Terbium	Tb	65	158.9254	159 (100)
Dysprosium	Dy	66	162.50	156 (0.052), 158 (0.090), 160 (2.29), 161 (18.88), 162 (25.53), 163 (24.97), 164 (28.18)
Holmium	Ho	67	164.9304	165 (100)
Erbium	Er	68	167.26	162 (0.136), 164 (1.56), 166 (33.41), 167 (22.94), 168 (27.07), 170 (14.88)
Thulium	Tm	69	168.9342	169 (100)
Ytterbium	Yb	70	173.04	168 (0.135), 170 (3.03), 171 (14.31), 172 (21.82), 173 (16.13), 174 (31.84), 176 (12.73)
Lutetium	Lu	71	174.97	175 (97.41), 176 (2.59; β^-, 2.1×10^{10} a)

∗ For the radioactive nuclides, the decay mode (EC = electron capture) and half life is given in brackets as well as the abundance.

Table 2. Electronic configurations and spectroscopic terms of the gaseous rare earth elements in the ground state

Symbol	RE^0			RE^{2+}			RE^{3+}	
Sc	[Ar]	$3d^1$	$4s^2$ ($^2D_{3/2}$)	[Ar]	$3d^1$	($^2D_{3/2}$)	[Ar]	($^1S_0$)
Y	[Kr]	$4d^1$	$5s^2$ ($^2D_{3/2}$)	[Kr]	$4d^1$	($^2D_{3/2}$)	[Kr]	($^1S_0$)
La	[Xe]	$5d^1$	$6s^2$ ($^2D_{3/2}$)	[Xe]	$5d^1$	($^2D_{3/2}$)	[Xe] $4f^0$	($^1S_0$)
Ce	[Xe] $4f^1$	$5d^1$	$6s^2$ ($^1G_4$)	[Xe] $4f^2$		($^3H_4$)	[Xe] $4f^1$	($^2F_{5/2}$)
Pr	[Xe] $4f^3$		$6s^2$ ($^4I_{9/2}$)	[Xe] $4f^3$		($^4I_{9/2}$)	[Xe] $4f^2$	($^3H_4$)
Nd	[Xe] $4f^4$		$6s^2$ ($^5I_4$)	[Xe] $4f^4$		($^5I_4$)	[Xe] $4f^3$	($^4I_{9/2}$)
Pm	[Xe] $4f^5$		$6s^2$ ($^6H_{5/2}$)	[Xe] $4f^5$		($^6H_{5/2}$)	[Xe] $4f^4$	($^5I_4$)
Sm	[Xe] $4f^6$		$6s^2$ ($^7F_0$)	[Xe] $4f^6$		($^7F_0$)	[Xe] $4f^5$	($^6H_{5/2}$)
Eu	[Xe] $4f^7$		$6s^2$ ($^8S_{7/2}$)	[Xe] $4f^7$		($^8S_{7/2}$)	[Xe] $4f^6$	($^7F_0$)
Gd	[Xe] $4f^7$	$5d^1$	$6s^2$ ($^9D_2$)	[Xe] $4f^7$	$5d^1$	($^9D_2$)	[Xe] $4f^7$	($^8S_{7/2}$)
Tb ∗	[Xe] $4f^9$		$6s^2$ ($^6H_{15/2}$) ∗	[Xe] $4f^9$		($^6H_{15/2}$)	[Xe] $4f^8$	($^7F_6$)
Dy	[Xe] $4f^{10}$		$6s^2$ ($^5I_8$)	[Xe] $4f^{10}$		($^5I_8$)	[Xe] $4f^9$	($^6H_{15/2}$)
Ho	[Xe] $4f^{11}$		$6s^2$ ($^4I_{15/2}$)	[Xe] $4f^{11}$		($^4I_{15/2}$)	[Xe] $4f^{10}$ ($^5I_8$)	
Er	[Xe] $4f^{12}$		$6s^2$ ($^3H_6$)	[Xe] $4f^{12}$		($^3H_6$)	[Xe] $4f^{11}$ ($^4I_{15/2}$)	
Tm	[Xe] $4f^{13}$		$6s^2$ ($^2F_{7/2}$)	[Xe] $4f^{13}$		($^2F_{7/2}$)	[Xe] $4f^{12}$ ($^3H_6$)	
Yb	[Xe] $4f^{14}$		$6s^2$ ($^1S_0$)	[Xe] $4f^{14}$		($^1S_0$)	[Xe] $4f^{13}$ ($^2F_{7/2}$)	
Lu	[Xe] $4f^{14}$	$5d^1$	$6s^2$ ($^2D_{3/2}$)	[Xe] $4f^{14}$	$6s^1$	($^2S_{1/2}$)	[Xe] $4f^{14}$ ($^1S_0$)	

∗ For Tb^0, possibly also $4f^8\,5d^1\,6s^2$.

H_2O are genuine scandium minerals. Apart from scandium, thortveitite also contains yttrium, ytterbium, and lutetium.

In many minerals, scandium is present in a dispersed state. Wolframite and cassiterite can contain up to 1 % scandium, so that scandium is a byproduct of the production of tungsten and tin. Uranium minerals contain much smaller amounts of scandium, but, since uranium is produced in relatively large quantities, scandium is produced in appreciable quantities also.

The abundance of the rare earth elements taken together is quite considerable. Cerium, the most common rare earth, is more abundant than, e.g., Co. Yttrium is more abundant than, e.g., Pb, whereas genuinely rare REs such as Lu and Tm are as abundant as Sb, Hg, Bi, and Ag. Promethium in rare earth minerals is present only in amounts of $< 10^{-19}$ % as a result of nuclear re-

Table 3. History of the discovery of the rare earth elements [6, 8]

Element	Year	Discoverer	Notes
Scandium	1879	Nilson	from ytterbium earths (De Marignac, 1878)
Yttrium	1843	Mosander	from yttrium earths (Gadolin, 1794)
Lanthanum	1839	Mosander	from cerium earths (Klaproth, Berzelius, Hisinger, 1803)
Cerium	1841		
Praseodymium	1885	Auer v. Welsbach	from didymium (Mosander, 1839 – 1841)
Neodymium			
Samarium	1901	Demarcay	from the samarium fraction (Lecoq de Boisbaudran, 1879)
Europium			of didymium (Mosander, 1839 – 1841)
Gadolinium	1886	Lecoq de Boisbaudran	from didymium (Mosander, 1839 – 1841)
	1880	de Marignac	independently of each other
Terbium	1878	de Marignac	from terbium earth, (Delafontaine, 1878), originally named
		Delafontaine	terbium earth (Mosander, 1843)
Dysprosium	1886	Lecoq de Boisbaudran	from the holmium fraction (Cleve, 1879) of erbium earth
			(de Marignac, 1878)
Holmium	1879	Cleve	
Erbium	1879	Cleve	from erbium earth (De Marignac, 1878)
Thulium	1879	Cleve	from the holmium fraction (Cleve, 1879) of erbium earth
			(de Marignac, 1878)
Ytterbium	1878	de Marignac	from ytterbium earth (de Marignac, 1878)
Lutetium	1907	Urbain,	from ytterbium earth, independently of each other
		Auer v. Welsbach	(de Marignac, 1878)
		James	

Table 4. Rare earth content (% of total REO) in typical minerals

	Monasite		Bastnaesite		Xenotime	Apatite
	Australia	China	United States	China	Malaysia	CIS
Cerium earths	94.940	92.090	99.547	98.600	10.500	90.100
La_2O_3	23.900	23.350	33.200	23.000	0.500	25.100
CeO_2	46.030	45.690	49.100	50.000	5.000	45.000
Pr_6O_{11}	5.050	4.160	4.340	6.200	0.700	3.900
Nd_2O_3	17.380	15.740	12.000	18.500	2.200	14.000
Sm_2O_3	2.530	3.050	0.789	0.800	1.900	1.600
Eu_2O_3	0.050	0.100	0.118	0.200	0.200	0.500
Yttrium earths	5.060	7.910	0.315	1.400	89.500	7.250
Gd_2O_3	1.490	2.030	0.166	0.700	4.000	1.500
Tb_4O_7	0.040	0.100	0.016	0.100	1.000	0.100
Dy_2O_3	0.690	1.020	0.031	0.100	8.700	1.000
Ho_2O_3	0.050	0.100	0.005	trace	2.100	0.100
Er_2O_3	0.210	0.510	0.004	trace	5.400	0.150
Tm_2O_3	0.010	0.510	0.001	trace	0.900	0.020
Yb_2O_3	0.120	0.510	0.001	trace	6.200	0.080
Lu_2O_3	0.040	0.100	trace	trace	0.400	trace
Y_2O_3	2.410	3.030	0.091	0.500	60.800	4.300
Total	100.000	100.000	99.862	100.000	100.000	97.350

actions with cosmic rays or the spontaneous fission of ^{238}U.

The rare earth minerals of the three groups referred to above occur in distinct deposits (Fig. 1) [11]. Perhaps the most important mined rare earth deposit is at the Mountain Pass Mine in California (carbonatite), where up to 40 000 t/a bastnaesite ore concentrate (70 % REO) is produced by ore beneficiation. Other important bastnaesite deposits are in Burundi (carbonatite), Madagascar (alkali rock), and in Bayan Obo, near the town of Baotou in Inner Mongolia in the People's Republic of China. The bastnaesite, with monazite, is associated with magnetite – hematite – fluorspar.

Monazite mainly occurs in secondary deposits in the heavy minerals of coastal sands. The extraction of monazite is closely linked with the extraction of rutile, ilmenite, and zircon in Australia, Brazil, India, and the United States, and

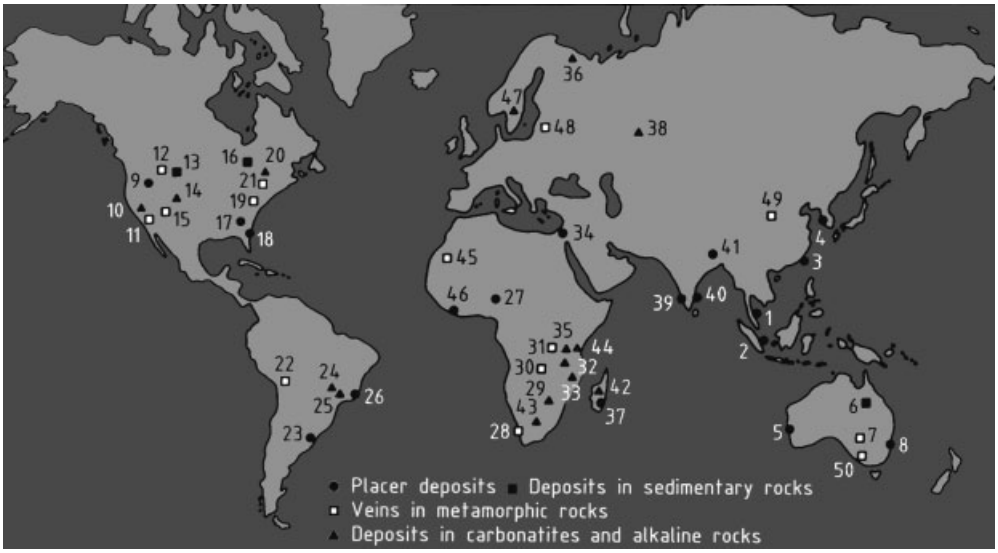

Figure 1. Important rare earth deposits [11]
1) Malaysia; 2) Singkep, Billit and Bangka, Indonesia; 3) Taiwan; 4) Korea; 5) Western Australia; 6) Mary Kathleen Mine, Australia; 7) Radium Hill, Australia; 8) New South Wales, Australia; 9) Bear Valley, Idaho/Montana; 10) Mountain Pass, California; 11) Music Valley, California; 12) Mineral Hill/Lenhi Pass, Idaho/Montana; 13) Bald Mountain, Wyoming; 14) Powderhorn and Wet Mountains, Colorado; 15) Gallinas Mountains, New Mexico; 16) Blind River/Elliot Lake, Ontario; 17) Piedmont, Georgia; 18) Atlantic Coast placer deposits; 19) Dover, New Jersey; 20) Oka, Quebec; 21) Mineville, New York; 22) Llallagua, Bolivia; 23) Atlanida, Uruguay; 24) Araxa, Brazil; 25) Morro do Ferro, Brazil; 26) Espirito Santo, Brazil; 27) Jos Plateau, Nigeria; 28) Steenkampskraal, Republic of South Africa; 29) Glenover, Republic of South Africa; 30) Shinkolobwe, Zaire; 31) Karonge, Burundi; 32) Panda Hill, Tanzania; 33) Kangakunde Hill, Malawi; 34) Nile Delta, Egypt; 35) Mrima, Kenya; 36) Kola Peninsula, CIS; 37) Madagascar; 38) Vishnevye Mountains, CIS; 39) Kerala, India; 40) Sri Lanka; 41) Bihar and Bengal; 42) Itremo, Madagascar; 43) Pilanesberg, Boputhatswana; 44) Wigu Hill, Tanzania; 45) Bon Nage Mauretania; 46) Monrovia, Liberia; 47) Fen, Norway; 48) Kangasala, Finland; 49) Bayan Obo, Inner Mongolia; 50) Roxby Downs, Australia

with the extraction of cassiterite in Indonesia, Malaysia, Thailand, Nigeria, and Zaire. However, this also means that the amount of monazite available depends on the demand for the main product. These monazites contain up to ca. 10 % thorium. Xenotime is found in similar deposits, Malaysian monazite having particularly high xenotime contents.

Primary monazite deposits with high thorium contents are found in Brazil, and particularly in the Republic of South Africa, though these are not extracted at present. Low-thorium monazites occur in primary carbonatitic deposits in Malawi and Burundi.

The minerals of the third group are extracted mainly for their Nb, Ta, and U contents.

In 1990, world reserves of rare earth minerals were estimated at ca. 84×10^6 t REO. China has estimated reserves of 43×10^6 t REO (ca. 50 % of the world reserves). The Baiyunebo deposit near Baotou in Inner Mongolia is the largest source of rare earth minerals in the world and accounts for over half of the Chinese production. The bastnaesite deposit at Mountain Pass, California, contains an estimated reserve of 4.3×10^6 t REO representing approximately 5 % of world reserves. Total reserves within the United States amount to 12.6×10^6 t REO.

Reserves elsewhere have been estimated as follows: Australia: 5.1×10^6 t REO, India: 2.3×10^6 t REO, Canada: 0.95×10^6 t REO, CIS: 0.45×10^6 t REO, Brazil: 0.28×10^6 t REO, and others, excluding the United States and China, 19.63×10^6 t REO.

The production of bastnaesite concentrate at the Mountain Pass Mine, California reached a peak in 1984 of 25 312 t REO. In 1990, production totalled 22 829 t REO, representing approximately 27 % of world production [12]. Estimated monazite production expressed as rare earth oxide (REO) in 1990 was as follows: Australia: 6500 t, Brazil: 1900 t, India: 2030 t, Malaysia: 1740 t, Republic of South Africa: 1000 t, United States: 2000 t, and the CIS: 1500 t.

Raw material prices in early 1990 for bastnaesite from the United States with 70 % REO was \$ 1.3/lb, reflecting an increased use of cerium oxide in autocatalysts, glass additives, and polishing compounds. From 1988 to 1991, the price of Australian monazite with > 55 % REO was stable between A\$ 800 – 900/t; Malaysian xenotime as 60 % yttrium concentrate costed \$ 32 – 33/kg.

3. Properties

3.1. Properties of the Nuclei

Some naturally occurring nuclides of the rare earth elements are radioactive. Data on types of radioactive decay and half lives are given in Table 1. Radioactive properties of artificial rare earth nuclides are given in [13]. Promethium, Pm, whose occurrence in nature was sought for a long time without success, was first identified with certainty in 1945 by MARINSKY et al. as a radioactive fission product of uranium [14].

So far, 25 relatively short-lived isotopes of Pm are known (longest half life: 17.7 years for $^{145}_{61}$Pm). ERÄMETSÄ [15] isolated traces of $^{145}_{61}$Pm from the rare earth concentrate of an apatite in 1965, and deduced that the abundance of Pm in nature is $< 10^{-19}$ %. The question whether this is produced by natural uranium fission or by the action of cosmic rays on $^{146}_{60}$Nd is the subject of debate.

The very high thermal neutron absorption cross section of several nuclides of Sm, Eu, and Gd is noteworthy.

3.2. Properties of the Atoms and Ions

3.2.1. Electronic Configuration, Position in the Periodic Table

The electronic configurations of neutral, isolated atoms of the rare earth elements in the ground state (Table 2) lead to the following position of the rare earth elements in the periodic table: the elements Sc, Y, and La are the first members of a series of transition elements (d-elements) with the valence electron configuration $ns^2\,(n-1)\,d^{\,1}$ ($n = 4, 5$, or 6). The 14 elements, Ce – Lu, constitute a series of inner transition elements ("f-elements") which must be inserted between La and Hf. These are the lanthanides, with valence electron configurations of the type $6s^{\,2}\,5d^{\,1}\,4f^{\mathrm{n}-1}$ or $6s^2\,4f^{\mathrm{n}}$. Thus, all the rare earth elements belong formally to Group 3 of the Periodic System.

Table 5. Atomic radii and effective ionic radii of the rare earth elements (in nm)

Symbol	Atom[a]	RE^{2+} [b]	RE^{3+} [c]	RE^{4+} [d]
Sc	0.1654		0.0730	
Y	0.1824		0.0892	
La	0.1884		0.1061	
Ce	0.1825		0.1034	0.096
Pr	0.1836		0.1013	0.095
Nd	0.1829		0.0995	
Pm	0.1825		0.0979	
Sm	0.1814	0.1232	0.0964	
Eu	0.1984	0.1220	0.0950	
Gd	0.1817		0.0938	
Tb	0.1803		0.0923	0.088
Dy	0.1796		0.0908	
Ho	0.1789		0.0894	
Er	0.1780		0.0881	
Tm	0.1769	0.1127[e]	0.0869	
Yb	0.1932	0.1115	0.0858	
Lu	01760		0.0848	

[a] Half the interatomic distance for the modification stable at room temperature, calculated from lattice constants [17, 18].
[b] For coordination number 8 in difluorides with $r_{F^-} = 0.132$ nm [20].
[c] For coordination number 6 [19].
[d] For coordination number 8 in dioxides with $r_{O^{2-}} = 0.138$ nm.
[e] Estimated.

In the neutral atoms, the $5d$ and $4f$ electrons have similar energies. This fact explains the occurrence of the two typical electronic configurations referred to above within the lanthanide series (see Table 2), whereby the configurations $4f^{\,7}$ (half-filled orbital) and $4f^{\,14}$ (completely filled orbital) are preferred owing to the high thermodynamic stability of such states. The irregularity of the configurations of the two elements immediately before the middle and the end of the lanthanide series, Eu and Yb, are characteristic of the neutral atoms and the mono- and divalent ions. However, in the case of the trivalent ions, the $4f$ orbital is filled in a completely regular manner. The energy of the $4f$ orbital decreases with increasing atomic number and increasing ionic charge, so that when the valency is 3+, which is characteristic of the rare earth elements, the $4f$ electrons represent well-defined "inner electrons" that take virtually

no part in chemical bonding due to their proximity to the nucleus and shielding by the $5s^2$ $5p^6$ octet. Therefore, the regular filling of the $4f$ orbital from Ce^{3+} to Lu^{3+} determines the chemical properties, as the resulting electronic configurations resemble those of the rare gases, and lead to the great similarity of the Ln^{3+} compounds to each other, to the corresponding compounds of La^{3+}, Y^{3+}, and, to a limited extent, to those of Sc^{3+}.

The regular filling of the $4f$ orbital is accompanied by a steady reduction in the Ln^{3+} ionic radii (see Table 5, [17–20], and Figure 2) – the so-called lanthanide contraction. This steady shrinkage of the inner electron shell is mainly caused by the incomplete mutual shielding of the $4f$ electrons against the increasing attraction of the nucleus.

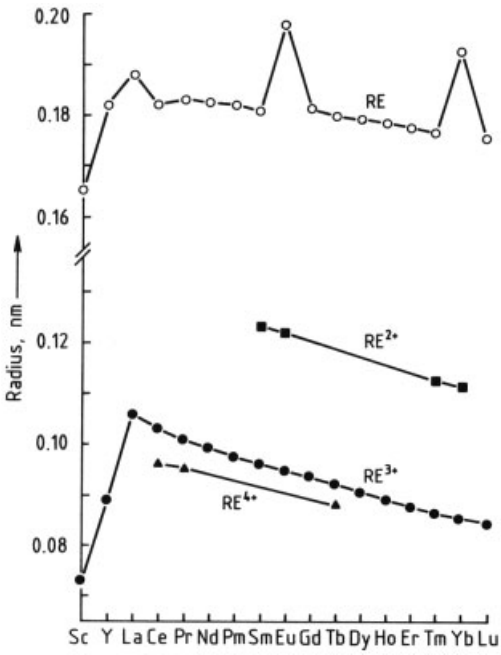

Figure 2. Atomic and ionic radii of the rare earth elements

The peculiarities of the electronic structure of the rare earth elements described here lead to periodic changes in some physical and chemical properties of these elements and their compounds in the series La to Lu, while other properties show a steady or nonperiodic change. The typical similarity of the rare earth elements to each other, which is often emphasized, is in essence only exhibited for the nonperiodic properties. Thus, the regular change in the RE^{3+} radii leads to largely nonperiodic changes in properties that depend on ion size, such as solubility, basicity, normal potential, and enthalpy of hydration. Periodic variation in properties can have various causes and cannot be interpreted with certainty in all cases. Typically, maxima or minima occur at Eu and Yb, because they are situated immediately before the $4f$ shell is half filled ($4f^7$) or completely filled ($4f^{14}$) and attempt to achieve these stable configurations. Periodic variations are observed for atomic radii, molecular volumes, melting points, boiling points, and sublimation enthalpies of the elements. A similar periodicity is also observed in some dissociation enthalpies of gaseous molecules REX (X=chalcogenide) due to cer-tain peculiarities in the electronic configurations [16]. Other types of pronounced periodicity occur, e.g., in magnetic moments and colors of the RE^{3+} ions (see Section 3.2.3).

3.2.2. Oxidation States, Atomic and Ionic Radii

The oxidation state 3+ is characteristic for all rare earth elements, both for solid compounds and for solvated ions in various solvents. The oxidation state 2+ can occur in pure solid compounds of Eu, and, to a slightly lesser extent, Yb, due to the higher stability of the configurations $4f^7$ and $4f^{14}$. It is also observed to a very limited extent with Sm and Tm. In aqueous solutions, Eu^{2+} ions are metastable. All the other Ln^{2+} ions react in acidic media as reducing agents:

$$Ln^{2+}\,(aq) + H^+\,(aq) \rightarrow Ln^{3+}\,(aq) + 1/2\,H_2\,(g)$$

The oxidation state 4+ can occur for Ce and Tb, which follow La and Gd respectively, and also to a very limited extent for Pr and Nd. Only the Ce^{4+} ion is metastable in aqueous solution; The other Ln^{4+} ions oxidize water as follows:

$$Ln^{4+}\,(aq) + 1/2\,H_2O\,(liq) \rightarrow Ln^{3+}\,(aq) + H^+\,(aq) + 1/4\,O_2\,(g)$$

The general preference for the oxidation state 3+ in aqueous solution is probably a consequence of the similarity between the hydration

enthalpy and ionization enthalpy of the ions. The standard electrode potentials listed in Table 6 [8] illustrate the behavior of the various RE^{n+} ions in aqueous solution, and show that the rare earths are a group of highly nonnoble elements. Figure 2 shows the atomic radii (in the metallic state) and ionic radii for the various typical valencies. The periodicity of the atomic radii described above (maxima for Eu and Yb) shows that in the metallic state these elements contribute only two electrons to the bond, so that their atomic volumes are similar to those of the alkaline earth metals. The radii of the Ln^{2+} ions, plus the radius of the Ba^{2+} ion, give a curve resembling the lanthanide contraction curve, and the term baride contraction has been suggested to describe this [21]. The radii of Eu^{2+} and Sr^{2+} are similar, as are the radii of Yb^{2+} and Ca^{2+}, which explains the commonly observed isomorphic replacement of alkaline earth ions by Ln^{2+} ions of similar size.

Table 6. Standard electrode potentials E^0_{298} of the rare earth elements in volts [8] (much of the data is estimated)

Element	RE (s)/ RE^{3+} (aq)	RE^{2+} (aq)/ RE^{3+} (aq)	RE^{3+} (aq)/ RE^{4+} (aq)
Sc	+2.077		
Y	+2.372		
La	+2.522		
Ce	+2.483		−1.74
Pr	+2.462	ca.	−2.86
Nd	+2.431		
Pm	+2.423		
Sm	+2.414	+1.55	
Eu	+2.407	+0.43	
Gd	+2.397		
Tb	+2.391		−1.28
Dy	+2.353		
Ho	+2.319		
Er	+2.296		
Tm	+2.278		
Yb	+2.267	+1.15	
Lu	+2.255		

The similarity of the chemical properties of Ln^{3+} compounds to those of the corresponding compounds of La^{3+} and Y^{3+} can be explained by the similarities between the sizes of these ions. Therefore, as expected, in the case of those properties that are mainly determined by ion size, La should be placed near Ce, and Y between Ho and Er. However, for scandium, which has a comparatively small radius ($r_{Sc^{3+}} = 0.0730$ nm), there is a greater similarity to the chemistry of aluminum ($r_{Al^{3+}} = 0.0530$ nm) than to that of lutetium ($r_{Lu^{3+}} = 0.0848$ nm).

3.2.3. Magnetic and Spectral Properties

The electronic configurations and spectroscopic terms of the Ln^{3+} ions are largely unaffected by the type of compound in which they are situated, owing to the shielding of the $4f$ electrons from the chemical environment by the $5s^2 p^6$ octet. For these ions, with the exception of Sm^{3+} and Eu^{3+}, it is characteristic that the values of the total spins J applicable to the spectroscopic term of the ground state differ from each other fairly widely (difference at room temperature $> kT$) so that in the ground state practically only one J value is occupied. This peculiarity has interesting consequences for the magnetic properties of these ions. While Sc^{3+}, Y^{3+}, La^{3+}, and Lu^{3+} are diamagnetic due to their closed shells, the remaining Ln^{3+} ions are paramagnetic, (see Figure 3). The double maximum in the curve of permanent magnetic moments can be explained theoretically by the additive behaviour of the spin and orbital moments [3, vol. 1, pp. 310 – 350], [22]. These moments are in opposition for the elements before gadolinium, but additive for the elements that follow it. The magnetic moment of Gd^{3+} is due entirely to the electron spin. The magnetic moments of the Ln^{2+} and Ln^{4+} ions mainly correspond to those of the isoelectronic Ln^{3+}ions. Also, there is a noticeable difference between the magnetic properties of the lanthanides and those of the $3d$ transition elements, which are characterized by paramagnetism mainly due to spin.

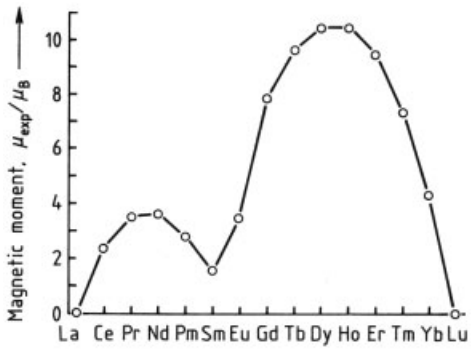

Figure 3. Magnetic moments of the RE^{3+} ions at 300 K

The Ln^{3+} ions, with the exception of Ce^{3+} and Yb^{3+}, have very sharp absorption bands

in the visible region and partly in the ultraviolet (Table 7). These bands are caused by $f \rightarrow$ f transitions. Their sharpness is due to the above-mentioned shielding of the $4f$ inner shell, which greatly hinders splitting of the spectroscopic terms by external fields. The colors of the Ln^{3+} ions are given in Table 7 [8]. The sequence of colors in the series La to Gd is almost exactly repeated in the opposite direction in the series Gd to Lu. The ions Ce^{3+} and Yb^{3+} do not absorb in the visible region because an $f \rightarrow$ f transition is not possible for the configurations $4\,f^{\,1}$ and $4\,f^{\,13}$. The broad absorption bands of these ions in the ultraviolet region are due to configuration transitions of the type $4f^{\,n} \rightarrow 4\,f^{\,n-1}\,5\,d$.

In aqueous solution, Eu^{2+} is pale yellow, Sm^{2+} is deep red, and Yb^{2+} is yellow.

Table 7. Colors and main absorption bands of RE^{3+} ions in the range 200 – 1000 nm [8]

Ion	Wavelength, nm	Color
La^{3+}		colorless
Ce^{3+}	210.5, 222.0, 238.0, 252.0	colorless
Pr^{3+}	444.5, 469.0, 482.2, 588.5	yellow-green
Nd^{3+}	354.0, 521.8, 574.5, 739.5, 742.0, 797.5, 803.0, 868.0	red-violet
Pm^{3+}	548.5, 568.0, 702.5, 735.5	pink
Sm^{3+}	362.5, 374.5, 402.0	yellow
Eu^{3+}	375.5, 394.1	colorless
Gd^{3+}	272.9, 273.3, 275.4, 275.6	colorless
Tb^{3+}	284.4, 350.3, 367.7, 487.2	very pale pink
Dy^{3+}	350.4, 365.0, 910.0	pale yellow-green
Ho^{3+}	287.0, 361.1, 416.1, 450.8, 537.0, 641.0	yellow
Er^{3+}	364.2, 379.2, 487.0, 522.8, 652.5	pink
Tm^{3+}	360.0, 682.5, 780.0	pale green
Yb^{3+}	975.0	colorless
Lu^{3+}		colorless

3.2.4. Bonding, Coordination Numbers

The physical, chemical, and above all structural properties of the rare earth compounds indicate that these elements exhibit mainly ionic bonding.

In solution and in crystalline compounds, the RE^{x+} ions (except Sc^{3+}) generally have coordination numbers > 6 [8]. For scandium compounds, which can also be regarded as homologues of aluminum, coordination numbers > 6 do not occur. In general, the tendency of the lanthanides, including Y and La, to form complexes is smaller than for the d-transition elements, as the shielded $4f$ orbitals are not available for forming the hybrid orbitals that form covalent bonds. Also, the very large size of the Ln^{3+} ions compared with other triply charged cations leads to smaller electrostatic forces of attraction.

3.3. Other Chemical and Physical Properties [1, part B 6], [23]

Reactivity. Apart from Eu, which is very readily oxidized under all conditions, the metals react at room temperature with oxygen or dry air relatively slowly, and in moist air rapidly. At elevated temperatures, all rare earth metals ignite in air and react with most nonmetallic elements. In dilute mineral acids, rapid dissolution takes place with formation of RE^{3+} ions. Like the alkaline earth metals, europium and ytterbium form deep blue, strongly reducing solutions in liquid ammonia.

Crystal Structure. Most of the rare earth metals are polymorphic. At room temperature, Sc, Y, Gd to Tm, and Lu crystallize in the hexagonal close-packed system, and La, Pr, and Nd in the double hexagonal close-packed (La type) system. The metals Ce and Yb crystallize in the face-centered cubic system. Sm is rhombohedral, and Eu body-centered cubic. Whenever the body-centered cubic modification occurs, it is always the stable high-temperature form up to the melting point.

Melting Point. The melting points are in the range 816 °C (Yb) to 1663 °C (Lu), and the boiling points lie between 1194 °C (Yb) and 3512 °C (Pr). Eu and Yb have a certain resemblance to the heavy alkaline earth metals in their melting points, boiling points, and molar volumes. Other physical data are listed in Table 8, 9, [1,2,4].

3.4. Miscibility and Alloying Behavior

If trivalent Eu and Yb and tetravalent Ce and Tb are excluded, the atomic radii from La to Lu differ by only 8 % for coordination number 12, so that in all alloy systems of adjacent rare earth

Table 8. Properties of the rare earth elements [1,2,4]

	Abundance in Earth's crust, ppm	*mp*, °C	Heat of fusion kJ/mol	*bp*, °C	Specific heat capacity at 298 K, $J\,mol^{-1}\,K^{-1}$	Linear coefficient of expansion, $10^{-6}\,K^{-1}$	Thermal conductivity, $W \cdot cm^{-1} \cdot K$	Vickers hardness	Modulus of elasticity, $10^{10}\,N/m^2$	Shear modulus, $10^{10}\,N/m^2$	Compression modulus, $10^{10}\,N/m^2$
Scandium	5 – 10	1540	15.5	2832	25.5	10.0	0.158	85 [a]	7.55	2.88	6.65
Yttrium	28 – 70	1525	11.4	3337	26.5	10.8	0.172	38	6.36	2.55	4.15
Lanthan	5 – 18	920	6.2	3457	26.2	4.9	0.135	37	3.84	1.49	3.03
Cerium	20 – 46	798	5.5	3427	27.0	8.5	0.114	24 [b]	3.00	1.20	2.00
Praseodymium	3.5 – 5.5	931	6.9	3512	27.0	6.79	0.125	37 [b]	3.52	1.35	2.99
Neodymium	12 – 24	1016	7.1	3067	27.5	9.98	0.165	35	3.79	1.45	3.27
Promethium	$< 10^{-15}$	1042	8	ca. 3000	24.3		0.179		2.14	0.87	1.32
Samarium	4.5 – 7	1073	8.6	1791	29.6	10.4	0.133	45	3.41	1.26	3.87
Europium	0.14 – 1.1	822	9.2	1597	27.7	33.1	(0.139)	17 [b]			
Gadolinium	4.5 – 6.4	1312	10.1	3267	37.1	8.28	0.105	57	5.62	2.23	3.78
Terbium	0.7 – 1	1357	10.8	3222	28.9	10.3	0.111	46	5.75	2.28	4.01
Dysprosium	4.5 – 7.5	1409	10.8	2562	28.2	10.0	0.107	42	6.31	2.54	4.11
Holmium	0.7 – 1.2	1470	12.2	2695	27.2	10.7	0.162	42	6.71	2.67	4.01
Erbium	2.5 – 6.5	1522	19.9	2862	28.1	12.3	0.143	44	7.33	2.96	4.55
Thulium	0.2 – 1	1545	16.9	1927	27.0	13.3	0.169	48			
Ytterbium	2.7 – 8	816	7.7	1194	26.7	25.0	(0.4)	21 [b]	1.78	0.70	1.30
Lutetium	0.8 – 1.7	1663	14.0	3397	26.9	12.5	0.164	77	6.846	2.715	4.766

[a] Rockwell hardness. [b] Heated to red heat.

Table 9. Properties of the rare earth elements [1,2,4]

	Poisson's ratio	Grüneisen constant for temperature dependence of electrical resistivity	Electrical resistivity, $\mu\Omega \cdot cm$	Electron emission, eV	Crystal structure of stable phase at room temperature (high-temperature polymorphs) [c]	Lattice constants, nm *a*	Lattice constants, nm *c*	Density, g/cm^3
Scandium	0.309	1.03	52	3.5	a (d)	0.33085	0.52683	2.989
Yttrium	0.265	1.00	60	3.1	a (d)	0.36482	0.57318	4.469
Lanthan	0.284	0.66	61	2.96	b (c, d)	0.3774	1.2159	6.145
Cerium	0.248	0.47	77	2.97	c (b, d)	0.51610		6.770
Praseodymium	0.305	0.49	71	2.96	b (d)	0.36721	1.18326	6.773
Neodymium	0.306	0.74	64	3.0	b (d)	0.36582	1.17926	7.007
Promethium		(0.72)	54	2.85	b (d)	0.365	1.165	7.26
Samarium	0.352	0.63	94	2.7 – 2.8	(a, d)	0.3621	2.625	7.536
Europium		1.64	90	3.20	d	0.45824		5.243
Gadolinium	0.259	0.52	131		a (d)	0.36360	0.57810	7.900
Terbium	0.261	0.83	124	3.25	a (d)	0.36055	0.56966	8.229
Dysprosium	0.243	0.78	93	3.15	a (d)	0.35915	0.56501	8.550
Holmium	0.255	0.89	94	3.18	a	0.35778	0.56178	8.795
Erbium	0.238	1.01	90	3.10	a	0.35592	0.55850	9.066
Thulium		1.08	72	2.59	a	0.35375	0.55540	9.321
Ytterbium	0.284	0.98	25	3.35	c (d)	0.54862		6.965
Lutetium		0.66	59		a	0.35052	0.55494	9.840

[c] a = hexagonal close packed (hcp); b = double hexagonal close packed (dhcp); c = face-centered cubic; d = body-centered cubic; r = rhombohedral.

metals there is complete gap-free miscibility in the liquid state and also in a temperature region close to the solidus curve. The liquidus curves are always continuous, corresponding to the behavior of almost ideal mixtures [24–26].

The mutual effect of atomic radius and electronegativity on the formation of mixed crystals between the rare earth metals and other metals is shown in Figure 4, [27], in which La and Lu are given as examples. The nearby elements (within the inner ellipse), exhibit extensive miscibility in the solid state. At a greater distance (within the outer ellipse), they have limited miscibility with the rare earth elements. Elements outside the ellipses have such large differences that only slight solubility in the solid state can be expected. Thus, all transition elements of the adjacent groups 4 – 8 are soluble to a maximum extent of 1 % in the solid state.

In Figure 5 [2], the compositions of intermetallic compounds between various elements and the rare earth metals are shown. The vertical lines show the region in which intermetallic phases exist, while the horizontal lines show the phases themselves. The most important of these are the phases REX_2, REX, and REX_3. The phase containing the highest proportion of rare earth metal is RE_3X. There is no lower limit; for example, a compound REB_{66} is known (see also [28–31]).

3.5. Mechanical Workability

When casting rare earth metals in molds, they must be melted in crucibles made of materials with which they do not react. Depending on the purity required, tantalum, molybdenum, clay/graphite, graphite, or boron nitride, etc. can be used. Inert gas atmospheres, vacuum, or sometimes only a covering salt may also be required. The castings are easily machined by turning, milling, drilling, or sawing, though care must be taken that the pyrophoric character of the metals does not lead to spontaneous ignition of the turnings or powdered material produced. Rare earth metals can easily be extruded at temperatures just below the melting point. The presence of impurities, especially oxygen, has a detrimental effect on cold and hot rolling properties.

4. Digestion of Rare Earth Ores

4.1. Wet Chemical Digestion; Fusion

The most important rare earth minerals monazite, bastnaesite, and xenotime are generally concentrated by physical processes such as heavy medium separation, flotation, and magnetic separation [1, part B 1, pp. 60 – 67]. They are ground to 50 µm – 1 mm and then digested with acids or alkalis. Treatment with alkalies produces the hydroxides of the rare earth metals and thorium, which can then be dissolved in concentrated hydrochloric or nitric acid. Treatment of the ores with sulfuric or hydrochloric acid converts the rare earth metals into soluble sulfates or chlorides [1, part B 1, pp. 67 – 104].

4.1.1. Monazite

Digestion by Alkalis. The reaction of rare earth phosphates with hot, concentrated alkali (50 – 75 %) proceeds as follows:

$$REPO_4 + 3\,NaOH \rightleftharpoons RE(OH)_3 + Na_3PO_4$$
$$Th_3(PO_4)_4 + 12\,NaOH \rightleftharpoons 3\,Th(OH)_4 + 4\,Na_3PO_4$$

This can be carried out at high temperatures in autoclaves (60 % alkali, mass ratio 1 : 1) [32], or with more concentrated alkali at lower temperature (120 °C) and normal pressures [33].

The hydroxides produced are treated with hot water. The trisodium phosphate goes into solution, and the hydroxides are filtered off. The trisodium phosphate, a saleable byproduct, crystallizes out. The washed hydroxides are then dissolved in hydrochloric or nitric acid. If the acid concentration is held at pH 4, a partial separation results, impure thorium hydroxide remaining behind while the rare earth hydroxides dissolve.

Sulfuric Acid Digestion. Monazite can be digested with 98 % sulfuric acid at 200 – 230 °C. The rare earth sulfates formed are then dissolved out of the crystalline, hygroscopic reaction product with cold water. The thorium sulfate either remains undissolved or also goes into solution, depending on the reaction conditions [34].

The dissolved thorium, which was at one time the most important product, was formerly precipitated as sulfate. However, the sharpness of

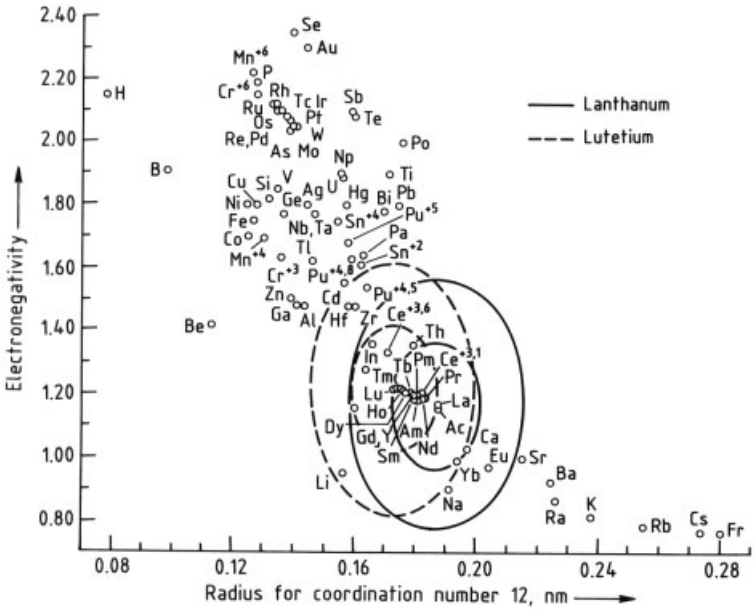

Figure 4. Darken – Gurry diagram for lanthanum and lutetium [27]

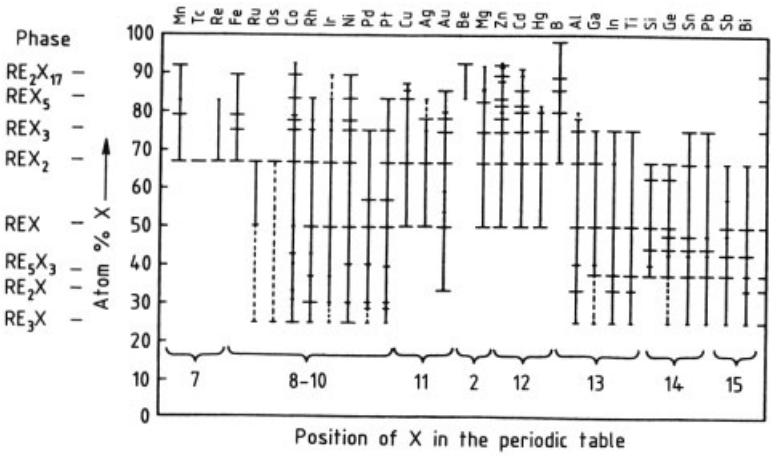

Figure 5. Compositions of binary rare earth alloys [2, vol. 2, chap. 13, pp. 1 – 54]

separation from the rare earth elements was poor. The method now used is to dissolve all the sulfates and separate the thorium by one of the following more effective methods:

1) Precipitation of ThF_4.
2) Precipitation of thorium phosphate by increasing the pH or by dilution.
3) Precipitation of the sodium/cerium earth double sulfates, while the corresponding much more readily soluble salts of thorium and the yttrium earths remain in solution. The thorium is then precipitated by adding oxalic acid, the solubility of thorium oxalate being much lower than that of the oxalates of the yttrium earths. The sparingly soluble double sulfates of the cerium earths are treated with boiling, concentrated alkali solution, forming the hydroxides, which are then dissolved in acid for subsequent fractionation.

4.1.2. Bastnaesite

Digestion with Acids. Many processes have been described for digesting bastnaesite with sulfuric acid. In one of these, the mineral is calcined to decompose the carbonates and then treated with 6 N sulfuric acid to dissolve the rare earth elements as sulfates [35].

In another process, the mineral is slurried with concentrated sulfuric acid and heated to 500 °C. The fluorine is driven off as hydrogen fluoride together with CO_2 and SO_2, and the rare earth elements remain behind as their anhydrous sulfates [36]. These can then be treated like the products from the treatment of monazite with sulfuric acid.

In another process, the mineral is calcined at > 600 °C, and then treated with 16 N nitric acid, which is considerably more effective than 12 N hydrochloric acid or 18 N sulfuric acid [37].

In the Molycorp process , the mineral is concentrated to 60 % by flotation and then calcined, converting the cerium to the tetravalent state. It is then treated with hydrochloric acid, which causes only the trivalent rare earth elements to go into solution, leaving behind 65–80 % CeO_2, which can be converted directly to a glass-polishing material by further calcination [38].

In another process, the carbonates are decomposed by hydrochloric acid. A fluoride residue is obtained, which is treated with alkali. The rare earth hydroxides obtained in this way are used to neutralize the excess acid from the chloride solution [39].

Digestion with Bases. Bastnaesite can be treated with concentrated alkali at ca. 200 °C to form rare earth hydroxides, which are then dissolved in acid [40].

4.1.3. Other Ores

Xenotime is more difficult to digest than monazite. It is usually treated with concentrated alkali, like monazite, but under much more rigorous conditions. Rare earth silicate minerals are best digested with sulfuric acid at fairly high temperatures. Various methods have been described for the extraction of the rare earth elements from apatites during wet production of phosphoric acid [41,42]. A process for the direct treatment of apatite with $NOCl/NO_2/N_2O_4$ at high temperatures is proposed in [43].

In the field of uranium production, only Denisson (Ontario) has so far produced an yttrium concentrate by solvent extraction of a sulfate solution.

4.2. Direct Chlorination at High Temperature

The direct chlorination of rare earth ores in a shaft furnace at 1000–1200 °C produces two groups of chlorides, depending on the impurities present. These are firstly the chlorides volatile at this temperature such as $AlCl_3$, $FeCl_3$, $POCl_3$, $SiCl_4$, $ThCl_4$, $TiCl_4$, $NbOCl_3$, $NbCl_5$, and $TaCl_5$, and secondly the nonvolatile chlorides of the alkali and alkaline earth metals as well as the rare earth chlorides, which collect at the bottom of the chlorinating furnace and are tapped off. The anhydrous rare earth chlorides can be treated, without further drying, by molten salt electrolysis to produce cerium mischmetal or they can be dissolved for further wet chemical processing.

The reaction of bastnaesite or monazite proceeds according to the following equations:

$$3\,REFCO_3 + 3\,C + 3\,Cl_2 \rightarrow 2\,RECl_3 + REF_3 + 3\,CO_2 + 3\,CO$$

$$REPO_4 + 3\,C + 3\,Cl_2 \rightarrow RECl_3 + POCl_3 + 3\,CO$$

The rare earth ore, with a grain size of < 0.2 mm, is pelletized with charcoal powder and a binder. It is then dried and fed into the chlorination furnace (Fig. 6). This consists essentially of a vessel lined with refractory bricks, inside which is a carbon cylinder that is the actual reaction space. A perforated graphite tube located inside this carbon cylinder contains a bed of coke through which the chlorine gas is passed to ensure even distribution before it passes through the solid feed material. The reaction is energetically self-sustaining, and only requires sufficient resistance heating to compensate for radiation losses. The rare earth chlorides collect in the melt space and are tapped off from time to time. A total of ca. 400 t rare earth chlorides can be produced during the working life of the carbon cylinder. The cylinder can be

replaced without difficulty while hot. The output of a chlorination furnace of this type is ca. 100 t/month. The lining of the chlorination furnace has a lifetime corresponding to a total output of 1500 t rare earth chloride.

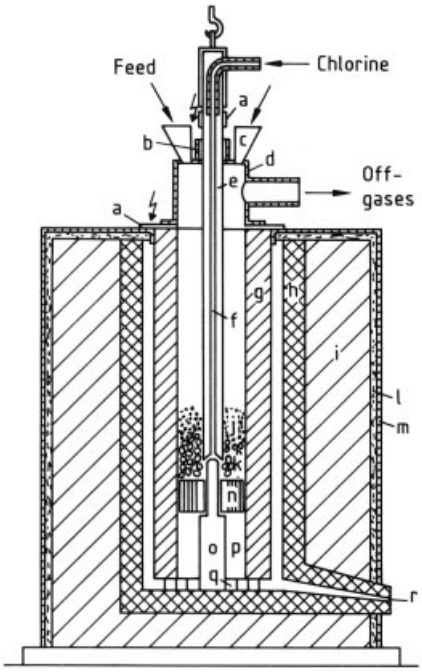

Figure 6. Diagram of a chlorination furnace [44,45]
a) Electrical connection; b) Insulation and seal; c) Charging tunnel; d) Furnace hood; e) Central electrode; f) Central hole for chlorine feed; g) Insert; h) Inner refractory lining; i) Outer refractory lining; j) Charge materials; k) Layer of coke; l) Slag wool; m) Iron casing; n) Sieve plate; o) Sieve plate supports; p) Space for molten material; q) Exit holes in insert; r) Tapping hole

The chlorination furnace operates at a slight underpressure. The waste gases are passed into a combustion chamber, oxidized, and washed with alkali to remove residual chlorine. In principle, it is possible to recover the volatile chlorides of valuable elements from the off-gas stream (e.g., $POCl_3$, $NbCl_5$, $TiCl_4$, $TaCl_5$, $ThCl_4$, UCl_4).

5. Separation

The separation of the rare earth elements poses one of the most difficult problems in inorganic chemistry. Due to the great similarity of the chemical properties, the methods are generally not very selective. In a few cases, much better separation can be achieved by conversion to another oxidation state of adequate stability ($Ce^{3+} \rightarrow Ce^{4+}$, $Eu^{3+} \rightarrow Eu^{2+}$).

A further difficulty is presented by the unfavorable distribution of concentrations of individual rare earth metals in the common minerals. The method of separation must therefore be selected to suit the existing ratios. The first step is a preliminary separation into groups of elements that resemble each other very closely.

5.1. Principles of Separation

Two principal types of process are used for the extraction of rare earth elements:

1) Solid – liquid systems using fractional precipitation or crystallization, or ion exchange. Ion exchange processes are used in the production of small quantities of higher value heavy rare earth elements.
2) Liquid – liquid systems using solvent extraction. This process is the most commonly used commercial process for the extraction of rare earth elements.

The distribution of two rare earth elements RE_A and RE_B between the two phases is given by the equilibrium constants k_A and k_B:

$$k_A = \frac{[RE_A]}{[RE_A]} \text{ and } k_B = \frac{[RE_B]}{[RE_B]}$$

where [RE] and [RE] represent the concentrations in the aqueous and nonaqueous phase, respectively. The selectivity of this separation system is given by the separation factor α_{RE_B/RE_A}, where

$$\alpha_{REB/RE_A} = k_B/k_A$$

More correctly, activities should be used instead of concentrations. In practice, apparent equilibrium constants based on concentrations are adequate. In the following discussion, these apparent constants are used throughout.

The processes of fractional crystallization or fractional precipitation are now mainly of historic interest only. The separation factor corresponds to the solubility ratio of the two rare earth elements in the aqueous phase.

In the case of ion exchange or liquid – liquid extraction, the ratio of the distribution coefficients is a measure of the different affinities of the rare earth metals for the two phases.

For the trivalent ions of neighboring rare earth elements, the separation factor is low for all methods. For precipitation or crystallization, $\alpha = 1.1$ to 5; for separation by ion exchange with addition of complexing agents, $\alpha = 1.1$ to 10; for liquid – liquid extraction, $\alpha = 1.1$ to 5.

For an effective separation, many repetitions of a single separation operation are necessary, which can be carried out batchwise or continuously.

With ion exchange separation, although the establishment of equilibrium is continually repeated, the overall process is discontinuous, since continuous feeding is not possible. However, it has the advantage that several rare earth elements can be separated at the same time.

A fully continuous liquid – liquid extraction process in countercurrent flow with partial recycle enables separation to be achieved into two groups or two individual rare earth elements of high purity. The production of three pure rare earth elements requires two batteries of separation equipment. This limitation is, however, compensated for by the advantage of continuous operation.

A further characteristic of a separation process is the size of the apparatus required. This depends on the concentration of the rare earth elements in the two phases. Cationic ion exchangers have a high capacity (ca. 150 g REO/L), but elution in the presence of complexing agents requires very dilute solutions (ca. 5 – 10 g REO/L). These separation plants are therefore very bulky.

In the liquid – liquid extraction process, the capacity of the solvent can be very high (170 – 190 g REO/L), and it is therefore often possible to work with aqueous solutions with concentrations of 100 – 140 g REO/L, i.e., approaching saturation. The equipment is therefore compact.

5.2. Separation by Classical Methods [1, part B 2], [3, vol. 1, pp. 30 – 61], [6]

5.2.1. Fractional Crystallization

Fractional crystallization was the first process for separating the rare earth elements, both in the laboratory and on the industrial scale. It is no longer of importance.

The purification effect decreases with increasing purity of the crystals. The achievable purity is therefore limited. Also, many intermediate fractions are produced that must be reprocessed.

In general, this process is more problematical for the yttrium earths than the cerium earths. The cerium earths were therefore separated by fractional crystallization and the yttrium earths by ion exchange. The cerium earths were converted to nitrates after oxidation of the cerium to the tetravalent state, and the double nitrates $NH_4NO_3 \cdot RE(NO_3)_3 \cdot 4\,H_2O$ were crystallized. Pechiney Saint Gobain produced 99.9 % pure lanthanum in five crystallization stages, this salt being the most soluble. The yttrium was first precipitated as the double chromate and then purified by ion exchange.

5.2.2. Fractional Precipitation

Fractional precipitation was only used for separating the rare earth elements into groups, due to filtration difficulties and the slow establishment of equilibrium.

The best known example is the precipitation of the hydroxides, which enables separation into three groups: (1) yttrium, (2) praseodymium, neodymium, and samarium, and (3) a fraction enriched in lanthanum and cerium.

Another process is the precipitation of double sulfates, $Na_2SO_4 \cdot RE_2(SO_4)_3$. The greater solubility of the yttrium earths enables a crude separation from the cerium earths to be carried out.

5.2.3. Separations Based on Oxidation State Changes

Cerium (IV) or europium (II) can be economically separated from the other rare earth metals by precipitation, ion exchange, or liquid – liquid extraction.

Cerium can be oxidized by oxidizing agents such as hypochlorite, H_2O_2, atmospheric oxygen, or electrolytically at a suitable pH, precipitating it as cerium(IV) oxide hydrate. By adding chlorine and sodium hydroxide solution to a

solution of cerium earth chlorides, cerium(IV) oxide hydrate is precipitated. This can be converted to a 90 % pure cerium oxide that is suitable for use as a polishing material. The separation of tetravalent cerium from the trivalent rare earth elements by liquid – liquid extraction is described in 5.4.2.2.

Divalent europium has properties similar to those of the alkaline earths, especially strontium (insoluble sulfate, soluble hydroxide, etc.). The rare earth metals are usually reduced in hydrochloric acid solution by zinc amalgam or sodium amalgam. The subsequent precipitation of europium sulfate from the very dilute solution can be improved by coprecipitation of strontium or barium sulfate. The europium can be further purified by precipitating trivalent elements as their hydroxides by using ammonia. Under these conditions, Eu^{2+} remains in solution [46]. The separation of divalent europium from the trivalent rare earth elements by liquid – liquid extraction is described in 5.4.3.1.

5.3. Separation by Ion Exchange

This process is based on work carried out on the separation of rare earth elements produced by uranium fission, and formed part of the Manhattan Project (1943 – 1947). The principles of separation by elution were developed by Boyd and coworkers in the Oak Ridge National Laboratory. The Ames Laboratory, Iowa State University, under the leadership of Spedding, investigated the production of rare earth metals on a large scale by displacement elution. Spedding and Powell have provided review of the developments up to 1950 [5, pp. 55 – 73], [47]. Attempts to develop ion exchange column chromatography into a continuous process was realized with the development of the continuous annular chromatograph (CAC) by researchers at Oak Ridge Laboratory in the mid 1970s. However, the process was designed for analytical requirements with low loading of the columns and was therefore not commercialized [48–50]. A more recent development, known as continuous displacement chromatography (CDC) [51], shows greater promise for surpassing present industrial, batch, fixed-column displacement ion exchange processes.

5.3.1. Ion Exchange with Chelating Agents

Sulfonated polystyrene ion exchange resins, which possess high chemical stability, can combine with 0.5 – 0.8 mol of rare earth elements per 1 mol resin. However, differences in the affinities of the rare earth elements for the resin are very small, and the separation factor for two neighboring rare earth elements is only ca. 1.08 [52]. However, this can be considerably improved by the addition of chelating agents.

The separation factor $F_{\mathrm{RE_B/RE_A}}$ for two rare earth elements $\mathrm{RE_A}$ and $\mathrm{RE_B}$ between an aqueous solution and the cation exchanger (concentrations marked with a bar) is

$$F_{\mathrm{REB/RE_A}} = \frac{\left[\overline{\mathrm{RE_B^{3+}}}\right] \cdot \left[\mathrm{RE_A^{3+}}\right]}{\left[\overline{\mathrm{RE_A^{3+}}}\right] \cdot \left[\mathrm{RE_B^{3+}}\right]}$$

On the addition of a chelating agent L, the equilibrium constants in solution are given by:

$$K_{\mathrm{A}} = \frac{[\mathrm{RE_A L}]}{[\mathrm{RE_A L}] \cdot [\mathrm{L}]} \quad K_{\mathrm{B}} = \frac{[\mathrm{RE_B L}]}{[\mathrm{RE_B L}] \cdot [\mathrm{L}]}$$

The constants depend on conditions in the medium, especially the pH of the aqueous phase. For large values of K_{B} and K_{A} and an excess of chelating agent (the concentration of the unchelated rare earth metals then being negligible), a new separation factor $F*_{\mathrm{RE_B/RE_A}}$ can be defined:

$$F^*_{\mathrm{REB/RE_A}} \approx F \cdot \frac{K_{\mathrm{B}}}{K_{\mathrm{A}}}$$

Separation can therefore be much improved by using a chelating agent whose equilibrium constants K_{A} and K_{B} are very different from each other.

Certain carboxylic acids form 1 : 1 chelates over a large pH range with K values that differ considerably for neighboring rare earth elements and which are sufficiently soluble in water. Citric acid is only of historical interest; it was used in the Ames Laboratory to prepare rare earth oxides in kilogram quantities with a purity of 99.9 % in 1947.

Later, polycarboxylic acids were replaced by aminocarboxylic acids, which form chelates with considerably improved stability, selectivity, and solubility.

Figure 7 shows the K values as a function of the atomic numbers of the rare earth elements. They vary between 10^{10} and 10^{20}. The selectivity F_L also varies widely, and can have values of between 1 and 10 for neighboring elements. For ethylenediaminetetraacetic acid (EDTA) and 1,2-diaminocyclohexanetetraacetic acid (DCTA), the selectivity for the cerium earths is much higher than for the yttrium earths. The position of yttrium (dashed lines) varies with the chelating agent over a range from Pr to Dy.

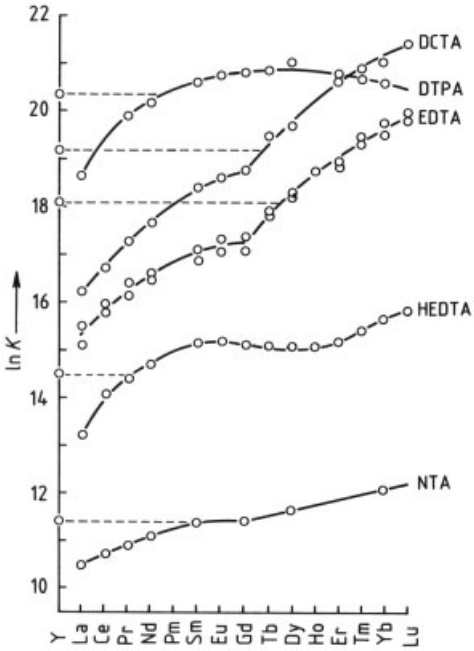

Figure 7. Stability constants K of selected rare earth chelates
– – – – Position of yttrium.

The disadvantage of using chelates is their lower solubility. The most soluble chelates such as those of N-(hydroxyethyl)ethylenediaminetriacetic acid (HEDTA) or diethylenetriaminepentaacetic acid (DTPA) reach concentrations of ca. $(2-2.5) \times 10^{-2}$ mol/L at 25 °C.

5.3.2. Separation Process

The ion exchange resin is initially loaded with NH_4^+. A solution of the salts of the rare earth elements to be separated is then passed through the column. The cations are absorbed by ion exchange onto the upper part of the column. The column is developed with a solution of the chelating agent, e.g., the triammonium salt of ethylenediaminetetraacetic acid.

In partition elution, the rare earth elements that form the most stable chelates are eluted first. The zones of the various rare earth elements are clearly separated from one other, but the eluted solutions are very dilute. This process is therefore used only for analytical purposes.

In displacement elution, a preliminary charge of a foreign ion (the so-called retarding ion) precedes the band of rare earth elements. This ion (denoted by R^{Y+}) has a lower affinity for the exchange resin than the rare earth elements. Displacement is now carried out using an ion that has a greater affinity for the resin than the rare earth elements. In some circumstances this ion can be NH_4^+. The sequence of affinities for the resin is then given by $NH_4^+ > RE_B > RE_A > R^{Y+}$. NH_4^+ displaces the RE_B ions almost quantitatively, while the RE_A ions displace the R^{Y+} ions. A band of constant length containing the rare earth elements moves through the column, and the concentration of the rare earth elements in the band remains constant. Between the zone containing pure RE_A and the zone containing pure RE_B there is a narrow zone containing both RE_A and RE_B.

By this method, more highly concentrated solutions of separated rare earth elements can be obtained than with partition elution. The same amount of resin can therefore be used to produce considerably larger quanities of rare earth elements.

The theory of displacement elution is described in [53]. Spedding and coworkers [54] have shown that, if the separation factor F_L is large enough, the band requires only a minimum movement ϑ along the column to achieve a constant equilibrium state. This relationship can be described as follows:

$$\vartheta = \frac{1 + (F_L + 1)\, x_i}{(F_L - 1)}$$

where x_i denotes the mole fraction of the rare earth element in the original mixture that has the lowest affinity for the resin and, therefore, the greatest chelating ability. If $x_i = 0.5$ (equimolar mixture of two rare earth elements) and $F_L = 2$,

the band must be moved along the column only 1.5 times its length to achieve equilibrium.

The cations in the equilibrium follow a logarithmic law. The molar fraction x_A of rare earth element A is a function of the number of theoretical plates n required to establish equilibrium:

$$x_A = \frac{F_L^{2n}}{1+F_L^{2n}}$$

A refinement of the ion exchange model is given in [55,56]. However, the equations given above are generally adequate for calculating the separation of the rare earth elements.

The main difficulty is the extrapolation from laboratory investigations to the industrial scale. The height equivalent to a theoretical plate (HETP) often becomes much greater (see Section 5.3.4).

Selection of Retarding Ions. The chelating ability of the weak aminopolycarboxylic acids increases with increasing pH. The H^+ ions can function as retarding ions:

$$REL^- + \overline{4H^+} \rightleftharpoons H_4L + \overline{RE^{3+}}$$

where REL^- denotes the dissolved ionic chelate of the rare earth metal. Owing to their low solubility, the acids tend to crystallize in the column. Only DTPA and HEDTA are sufficiently soluble at normal temperatures to enable H^+ to be used as the retarding ion for the separation of rare earth elements. EDTA can be used at 95 °C [57].

Many cations, e.g., Fe^{3+}, Cu^{2+}, Ni^{2+}, Zn^{2+}, and Pb^{2+} have been investigated as possible retarding ions [58]. Of these, Cu^{2+} has been most widely used.

5.3.3. Industrial Processes

Ion exchange was first used for separating rare earth elements on an industrial scale in the late 1950s. However, it was largely replaced by liquid–liquid extraction during the 1960s, and only used if no suitable liquid–liquid extraction process was known. An example was the production of yttrium oxide [59–62], which, with europium oxide, is important as a red phosphor for TV tubes. However, yttrium oxide is now produced more economically by liquid–liquid extraction (see 5.4.3.2).

Since 1970, small amounts of rare earth elements have continued to be separated by ion exchange [e.g., those required in high purity (> 99.9999 %) for some electronic applications] or for the production of rare earth compounds with a very small market and which are only occasionally needed. A new phenol-based ion exchange resin was introduced by Unitika Ltd. of Osaka in 1989. The resin reportedly separates the elements to a purity approaching 99.999999 % and at a faster rate than previous resins.

DTPA and HEDTA have proved to be the most suitable chelating agents for the production of high-purity rare earth compounds [63]. One process for producing small amounts of very pure rare earth compounds makes use of the temperature effect, which not only enables much more concentrated solutions of chelating agents to be used, but also enables the separation factor to be modified [64,65].

New advances in continuous displacement chromatography (CDC) are mentioned in Section 5.3.

5.3.4. Disadvantages of Ion Exchange

Chelating agents are expensive and must be recovered after use, but recovery is difficult and requires a large number of batch operations. Al-so, the low solubility leads to very dilute solutions of the rare earth elements (2–10 g REO/L), whereas with liquid–liquid extraction 10–50 times more concentrated solutions may be used. These high dilutions lead to bulky equipment and handling losses.

On the laboratory scale, theoretical plate heights of the order of millimeters are often possible. On conversion to the industrial scale, the HETP increases considerably, and can reach 10–20 cm in columns 230 cm high and 63 cm in diameter.

For reasons as yet unexplained, the purifying effect decreases with increasing purity. This effect is less significant if rare earth elements are required with impurity contents of ca. 50 ppm, but it makes it almost impossible to reduce the impurity content to, e.g., 5 ppm.

5.4. Separation by Liquid – Liquid Extraction

5.4.1. Theoretical Basis

The liquid – liquid extraction process for the separation of the rare earth elements was discovered by FISCHER et al. who showed that extraction of rare earth metal solutions in hydrochloric acid with an alcohol, ether, or ketone gives separation factors of up to 1.5 [66]. The principles of this separation technique are described in →6. Liquid – Liquid Extraction and [1, part B 2], [67, 68].

In the liquid – liquid extraction process used today, the organic phase usually consists of two or more substances. The active extractant has chelating properties and is mainly responsible for the transfer of the dissolved rare earth elements from the aqueous to the organic phase. In some cases, active extractants can be used neat, but they are usually highly viscous or even solid and must be dissolved in a solvent to ensure good contact between the two phases. Suitable solvents are kerosene and aromatics. Usually, a modifying agent is added to the organic phase to improve the hydrodynamics and sometimes also alter the separation factor.

Distribution Coefficient and Separation Factor. The distribution coefficient is the fundamental parameter in liquid – liquid extraction. It is strongly dependent on concentration and is given by;

$$k = \frac{[\mathrm{RE}]_{\mathrm{organic\,phase}}}{[\mathrm{RE}]_{\mathrm{aqueous\,phase}}} = \frac{\overline{[\mathrm{RE}]}}{[\mathrm{RE}]}$$

The separation factor for two rare earth elements $\mathrm{RE_A}$ and $\mathrm{RE_B}$ is thus

$$\alpha_{\mathrm{REB/RE_A}} = \frac{k_\mathrm{B}}{k_\mathrm{A}} \frac{\overline{[\mathrm{RE_B}]}\,[\mathrm{RE_A}]}{[\mathrm{RE_B}]\,\overline{[\mathrm{RE_A}]}}$$

Method of Operation. (The technology of liquid – liquid extraction is described in →6. Liquid – Liquid Extraction).

The flow diagram of a countercurrent extraction plant for the continuous separation of two rare earth elements or two groups of rare earth elements is shown in Figure 8 [69]. The mixture of rare earths that is to be separated is fed into an intermediate stage of the liquid – liquid extraction plant operating in countercurrent flow. The extraction medium becomes preferentially charged with the rare earth metals (extract) that form the more stable chelates, while the less stable rare earth chelates remain preferentially in the aqueous phase (raffinate).

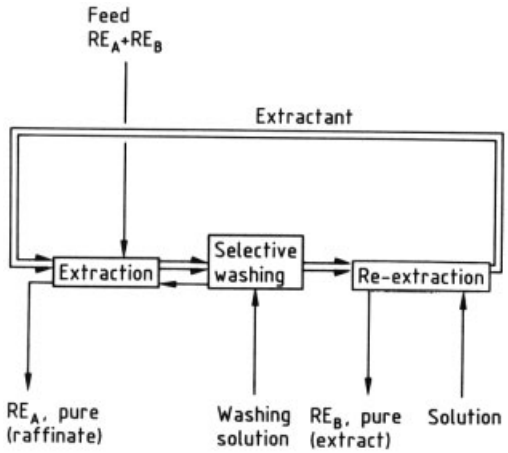

Figure 8. Principle of countercurrent liquid – liquid extraction

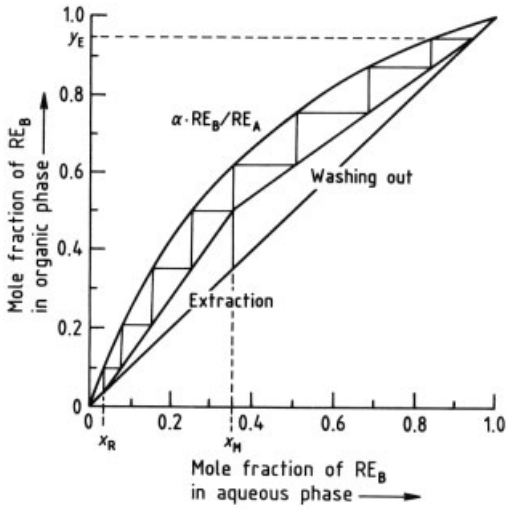

Figure 9. Mole fraction $\mathrm{RE_B}$ in the aqueous phase

This phase flows countercurrent to the extract, leading to further separation and preferential concentration of the more stable rare earth chelates in the extract (selective wash-out). The

rare earth metals are re-extracted for further treatment and the extraction medium is treated for re-use. At both ends of the plant, partial recycle takes place. If the recycle was total, separation into individual rare earth elements would take place, but this method is not practical, as it can only be operated batchwise, and a very long time is necessary for the establishment of equilibrium. The separation effect can be calculated by the McCabe – Thiele method in simple cases (Fig. 9).

5.4.2. Extractants

5.4.2.1. Neutral Extractants

The most effective extractant tri-*n*-butyl phosphate (TBP), with which the nitrates of the rare earths are usually extracted by the following generalized process:

$$RE(NO_3)_3 \cdot xH_2O + \overline{3\,TBP} \rightleftharpoons \overline{RE(NO_3)_3 \cdot 3\,TBP} + xH_2O$$

Separation factors between various rare earths and between the rare earths and yttrium are listed in Table 10 [70].

Concentrated nitric acid improves the selectivity. However, the strong acid solutions are difficult to handle, and the solubility of the rare earth nitrates in these is low. Furthermore, hydrolysis of the solvent takes place. A similar effect to that obtained with nitric acid can be achieved by adding nitrates that are not extracted by TBP, such as $Al(NO_3)_3$, alkali metal nitrates, or alkaline earth nitrates.

Pure TBP can be loaded with rare earth nitrates up to concentrations of 0.5 mol/L, corresponding to ca. 170 – 190 g REO/L. Temperature has a relatively small effect. The early development of mathematical models is described in [71]. A review of the many extraction processes with TBP is given in [72, 73]. Other salts of rare earth elements, e.g., perchlorates, chlorides, and thiocyanates have poor distribution coefficients or the selectivities are inadequate. The effect of low temperatures is described in [74].

Sulfoxides, RR′SO, behave analogously to TBP (Fig. 10), [75]. Alcohols and ethers, which were used in the early days, are no longer of importance.

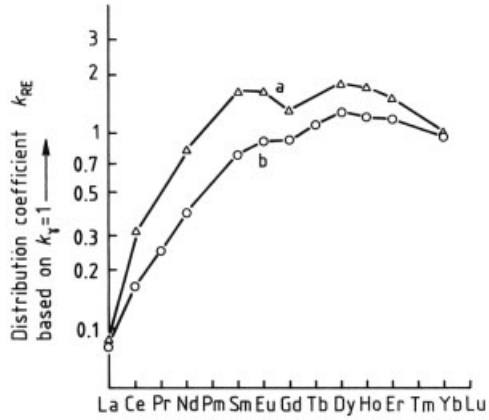

Figure 10. Distribution coefficients in the extraction of rare earth nitrates (480 g REO/L)
a) TBP with 50 % kerosene [75]; b) Di-*n*-heptyl sulfoxide (1 mol/L) in 1,2-dichlorobenzene

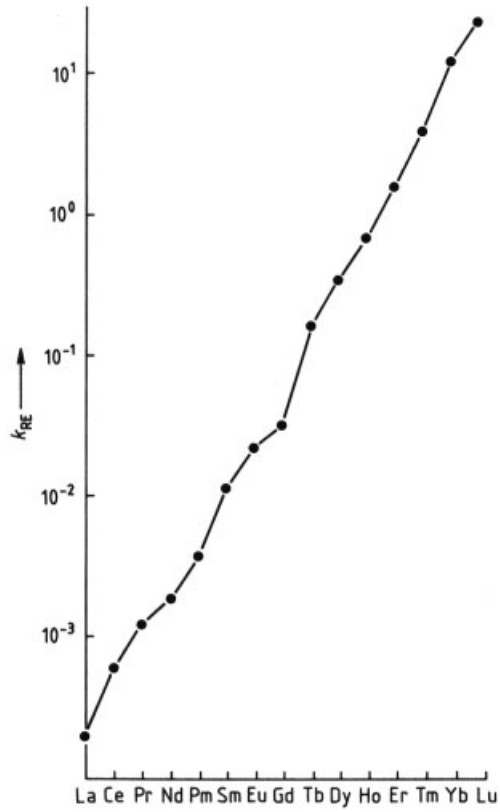

Figure 11. Distribution coefficients k_{RE} (relative to $k_y = 1$) in the system DEHPA/perchloric acid [78]

Table 10. Separation factors with rare earth concentration, nitrate media, with 50 vol % TBP in Shellsol A [70]

Aqueous phase REO	Separation factors – yttrium/rare earths							
concentration, g/L	Ce/Y	Nd/Y	Sm/Y	Gd/Y	Dy/Y	Ho/Y	Er/Y	Yb/Y
460		0.39	0.88	0.89	1.30	1.20	1.15	0.93
430		0.50		1.14	1.37	1.31	0.85	
310		0.69	1.41	1.51	1.76	1.65	1.35	
220	0.60	1.30	2.02	1.99	2.15	1.91	1.48	0.83
125	0.75	1.77	2.79	2.29	2.10	1.75	1.25	0.83
60		4.08	5.71	4.44	3.96	3.03	2.13	
	Separation factors – adjacent rare earth elements							
	Sm/Nd	Gd/Sm	Dy/Gd	Ho/Dy	Er/Ho	Yb/Er		
460	2.26	1.01	1.45	0.92	0.96	0.81		
430			1.20	0.96	0.65			
310	2.04	1.07	1.17	0.94	0.82			
220	1.55	0.99	1.08	0.89	0.78			
125	1.58	0.82	0.92	0.83	0.72			
60	1.40	0.18	0.89	0.77	0.70	0.63		

5.4.2.2. Acidic Extractants

Organophosphoric acids are the second most important extraction media after TBP. They react with the rare earth elements in what is considered to be a cation-exchange extraction process, e.g., as follows:

$$RE^{3+} + \overline{3\,HL} \rightleftharpoons \overline{REL_3} + 3\,H^+$$

The most fully investigated and most readily available is di(2-ethylhexyl)phosphoric acid (DEHPA). In the hydrocarbons that are used as diluents, the acid is dimeric. In the case of the rare earth elements, it has been found that the exchange reaction depends on the diluent [76]. In aromatic diluents, the reaction is

$$RE^{3+} + \overline{3\,H_2L_2} \rightleftharpoons \overline{RE(HL_2)_3} + 3\,H^+$$

In aliphatic diluents it is

$$RE^{3+} + \overline{2.5\,H_2L_2} \rightleftharpoons \overline{REL(HL_2)_2} + 3\,H^+$$

The dimeric phosphoric acid is represented by H_2L_2. The distribution coefficient depends on the diluent, decreasing in the following order: kerosene > cyclohexane > CCl_4 > xylene > benzene. In accordance with the above equations, it is strongly affected by the strength, being inversely proportional to the third power of the concentration of H^+ ions for low concentrations of the rare earth elements. This effect, which is very important in nitric or hydrochloric acid solution, leads to a minimum value of the distribution coefficient at an acid concentration of 5 – 6 N [77].

The distribution coefficient increases with increasing atomic number of the rare earth element. Yttrium comes between Ho and Er, in accordance with its ionic radius. DEHPA is a very selective extraction medium, with a separation factor of ca. 2.4 for neighboring elements. The α values in $HClO_4$medium are listed in Table 11, [79].

Figure 11 shows the distribution coefficients as a function of atomic number. The high selectivity of DEHPA is clearly shown. Also, the "tetrad effect" can be seen, shown by a discontinuity after every fourth element. This effect also occurs in other properties of the rare earth elements and is associated with their electronic structure.

The influence of the anions of the rare earth salts on the separation effect is fairly small, although it does play a role in some cases. Thus, for example, chlorides or sulfates are suitable for the separation of Nd and Sm. Sulfates also appear to be somewhat more selective for the sepa-ration of the cerium earths from the other rare earth metals.

In general, the total concentration of the rare earth elements and the relative concentrations of the various elements do not greatly affect the separation factor.

Table 11. Separation factors in the system DEHPA/perchloric acid [78]

Cerium	Praseodymium	Neodymium	Promethium	Samarium	Europium	Gadolinium	Terbium	Dysprosium	Holmium	Yttrium	Erbium	Thulium	Ytterbium	Lutetium	
2.98	6.09	8.42	1.82×10^1	5.54×10^1	1.05×10^2	1.51×10^2	7.43×10^2	1.56×10^3	3.03×10^3	5.00×10^3	6.84×10^3	1.70×10^4	5.25×10^4	9.78×10^4	Lanthanum
	2.05	2.83	6.11	1.86×10^1	3.54×10^1	5.07×10^1	2.50×10^2	5.26×10^2	1.02×10^3	1.68×10^3	2.30×10^3	5.71×10^3	1.76×10^4	3.29×10^4	Cerium
		1.38	2.98	9.11	1.73×10^1	2.48×10^1	1.22×10^2	2.57×10^2	4.98×10^2	8.22×10^2	1.12×10^3	2.79×10^3	8.62×10^3	1.61×10^4	Praseodymium
			2.16	6.58	1.25×10^1	1.79×10^1	8.82×10^1	1.86×10^2	3.60×10^2	5.49×10^2	8.12×10^2	2.02×10^3	6.23×10^3	1.16×10^4	Neodymium
				3.05	5.80	8.30	4.09×10^1	8.61×10^1	1.67×10^2	2.75×10^2	3.76×10^2	9.36×10^2	2.89×10^3	5.38×10^3	Promethium
					1.90	2.72	1.34×10^1	2.82×10^1	5.47×10^1	9.02×10^1	1.23×10^2	3.07×10^2	9.46×10^2	1.76×10^3	Samarium
						1.43	7.04	1.48×10^1	2.88×10^1	4.74×10^1	6.48×10^1	1.61×10^2	4.98×10^2	9.28×10^2	Europium
							4.93	1.04×10^1	2.01×10^1	3.32×10^1	4.53×10^1	1.13×10^2	3.48×10^2	6.49×10^2	Gadolinium
								2.11	4.08	6.73	9.21	2.29×10^1	7.06×10^1	1.32×10^2	Terbium
									1.94	3.20	4.37	1.09×10^1	3.35×10^1	6.25×10^1	Dysprosium
										1.65	2.25	5.61	1.73×10^1	3.22×10^1	Holmium
											1.37	3.40	1.05×10^1	1.96×10^1	Yttrium
												2.49	7.67	1.43×10^1	Erbium
													3.09	5.75	Thulium
														1.86	Ytterbium

Increasing the temperature tends to decrease the distribution coefficients and increases slightly the selectivity for the cerium earths, while that for the yttrium earths decreases.

Computer models for separation are described in [79,80]. Cerium(IV) is much more easily extracted than is the trivalent ion, and this effect has been utilized in the separation of cerium from fission products [81]. On the other hand, Eu^{2+} is much more difficult to extract than the RE^{3+} cations, so that an effective separation can be carried out from chloride solutions using HDEHP. However, ensuring that the europium remains in the divalent state during the extraction process is a problem [82].

Thermodynamic models for the prediction of distribution coefficients have been developed. In particular, a model has been presented for the extraction mechanism and chloride complexation when Sm, Eu, Gd, Tb, Dy, and Ho are extracted from acidic chloride solutions by DEHPA and 2-ethylhexylphosphonic acid mono-2-ethylhexyl ester in kerosene [83].

Other Phosphorus-Containing Extractants. A large number of phosphorus compounds have been investigated [84]. Di-*n*-octyl-, di-*n*-butyl-, and di(octylphenyl)phosphoric acids are much more powerful extractants than DEHPA, but much less selective. Substituted phosphonic acids of the type (RO)R′PO(OH) are less effective than the corresponding phosphoric acids, and the substituted phosphinic acids of the type (R)(R′)PO(OH) are even less so. The general reaction is of the form

$$RE^{3+} + \overline{3\,H_2A_2} \rightleftharpoons \overline{RE(HA_2)_3} + 3\,H^+$$

Monoalkylphosphoric acids of the type $(RO)PO(OH)_2$, such as 2-ethylhexylphosphoric acid, have very low selectivity and fairly high solubility in the aqueous phase.

Very generally, the extraction ability of the organophosphoric acids decreases if the degree

of branching of the groups R or RO increases, if an RO group is replaced by the corresponding R group, or if OH is replaced by an R or RO group [85].

Carboxylic acids ($C_7 - C_{15}$) are soluble in hydrocarbons, and can be used as extraction agents. The extraction according to

$$\overline{3\,R_3CCOOH} + RE^{3+} \rightleftharpoons \overline{(R_3CCOO)_3RE} + 3\,H^+$$

is carried out in neutral or weakly acidic medium. The extraction properties of naphthenic acids are described in [86].

The most widely used acids are the Versatic acids (Shell) $R^1R^2(CH_3)C\text{-}COOH$, in which R^1 and R^2 are branched-chain alkyl groups. The most commonly used products are Versatic 911 with 9 – 11 C atoms and Versatic 10 with 10 C atoms.

In contrast to systems using TBA and DEHPA, which favour the extraction of nitrate and chloride, respectively, Versatic acid exhibits similar extraction behaviour for both chloride and nitrate anion. Heavy rare earths are also strongly extracted by this reagent. Separation factors for Versatic 911 are listed in Table 12 [70].

Investigations with complex-forming aminopolycarboxylic acids such as EDTA or DTPA are reported in [87].

5.4.2.3. Amines and Quaternary Ammonium Salts as Extractants

Long-chain amines will extract a large number of metall ions in the form of anionic complexes [88, 89].

Primary and secondary amines are relatively weak extractants for rare earth elements, so that high concentrations of salts or acids are necessary to achieve a useful separation effect. They preferentially extract rare earth sulfates. The poor selectivity can be considerably improved by the presence of EDTA and DTPA. In all cases, results are strongly dependent on pH and concentration, which can lead to problems in process control.

Tertiary amines are poor extractants. On the other hand, long-chain quaternary ammonium salts containing the cation $R^1R^2R^3(CH_3)N^+$ in which R^1, R^2, and R^3 represent $C_8 - C_{12}$ groups are effective, especially in the form of nitrates or thiocyanates [90, 91]. The rare earth elements are extracted from nitric acid solution and form hydrate complexes of the type$(R_4N)_2RE(NO_3)_5 \cdot x\,H_2O$. If a certain concentration in the organic phase is exceeded (ca. 35 g REO/L), separation of the complex from the diluent takes place due to its poor solubility. This phenomenon can be prevented by adding a modifying agent (e.g., a higher alcohol).

The distribution coefficients and selectivity depend on the anion (Fig. 12) [92]. In nitrate solutions, the light elements are more effectively extracted, but xttrium accompanies Tm. In thiocyanate solutions, complex formation in the aqueous phase plays an important role. The degree of extraction increases with atomic number, yttrium occupying a position between Nd and Sm [92]. Separation using quaternary ammonium salts in the presence of chelating agents such as oxalic acid, EDTA, DTPA, and HEDTA is described in [93, 94].

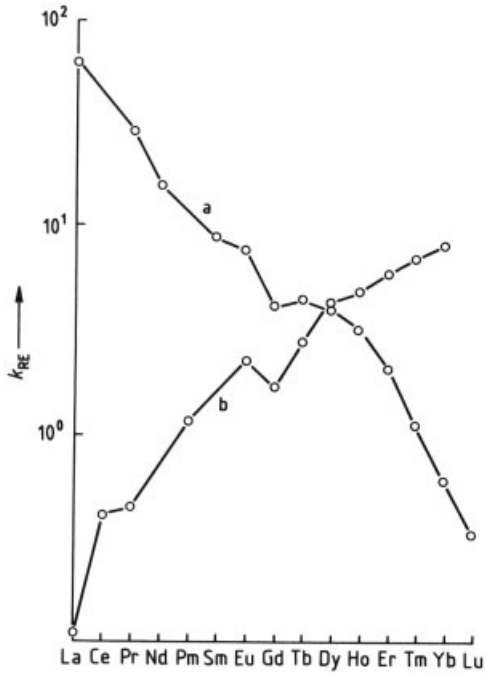

Figure 12. Distribution coefficients k_{RE} (relative to $k_y = 1$) for systems containing quaternary ammonium salts
a) Aliquat 336 in the solvesso/nitrate system; b) Aliquat 336 in the system xylene/thiocyanate

Table 12. Separation factors between rare earths in the Versatic 911/Shellsol A system (nitrate or chloride) [70]

Adjacent	Ce/La	Pr/Ce	Nd/Pr	Sm/Nd	Gd/Sm	Dy/Gd	Ho/Dy	Er/Ho	Yb/Er	
elements	3.00	1.60	1.32	2.20	1.96	1.94	1.17	1.28	1.49	
Versus	La/Y	Ce/Y	Pr/Y	Nd/Y	Sm/Y	Gd/Y	Dy/Y	Ho/Y	Yb/Y	Er/Y
yttrium	0.028	0.085	0.14	0.18	0.39	0.77	1.50	1.75	2.25	3.35

5.4.2.4. Synergistic Effects

If a mixture of extractants is used, the distribution coefficient can be considerably greater than the sum of the coefficients of the individual extraction agents, although the opposite effect is also known. Several systems have been described but have not achieved industrial importance [95].

5.4.3. Industrial Liquid – Liquid Extraction

The first step is the selection of the extractant and the aqueous phase. An important cost question is the consumption of reactive chemicals, and depends on the nature of the extraction medium. When the rare earth salts are extracted by neutral extraction media or anionic complexes, only water is required. Separation using cationic complexes, however, requires a base for removal from the aqueous phase, and an acid for selective washing and re-extraction from the organic phase. Cationic extraction is therefore considerably more costly than anionic or molecular extraction, and is only used where its high selectivity is of major advantage (e.g., if a product of very high purity is required).

When choosing the extraction medium, the only decisive factor is selectivity, although a less selective solvent may be better if it has higher dissolving capacity. The number of extraction systems is so great that the difficulty does not usually lie in finding an effective system, but in finding a system that offers the best compromise between a number of criteria. Thus, each separation problem has its own particular solution.

In industry, the preferred extraction apparatus is the mixer – settler. The first step in liquid – liquid extraction can be either the separation of the rare earth elements from other elements or the concentration of the rare earth elements. In the first case, a solvent is chosen whose distribution coefficient for the rare earth elements as a group is markedly different from that of the other elements. In the second case, a solvent is required with high extraction ability for the rare earth elements, enabling it to extract them from dilute, and often acidic solutions. DEHPA is very suitable for this. A practical example is the process of Denison Mines, Elliot Lake (Ontario), in which a concentrate with ca. 60 % REO, very rich in yttrium (Y_2O_3/REO = 1 : 1), is produced from solutions containing 150 mg Y_2O_3/L acidified with sulfuric acid, from which the uranium has first been removed (Fig. 13) [96].

The next task is the separation of the rare earth metals from each other. A sequence of separation stages is established such that the required products are obtained. The process of separation depends as much on this as on the great diversity of starting materials.

Rhône-Poulenc can produce all the rare earth elements in a purity of > 99.999 % almost entirely by solvent extraction, except for a few elements for which this process is very expensive, and which are still produced in kilogram quantities by ion exchange. The principle of the separation is shown in Figure 14.

The mineral is first digested with NaOH, and two solutions–one of rare earth chlorides and another of rare earth nitrates–are produced. The next step in both cases is the removal of other, unwanted elements. The cerium earths and yttrium earths (to the Nd/Sm level) are then separated from each other. Purification is then carried out using all the types of extraction agents described above (amines, carboxylic acids, acidic and neutral organophosphorus compounds). The choice of the extraction agent depends not only on the separation factor, but also on economic considerations. Rhône-Poulenc operates with more than 1000 separation units, some difficult separations alone requiring 60 such units. The byproducts from the intermediate fractions are usually stored or separated by ion exchangers.

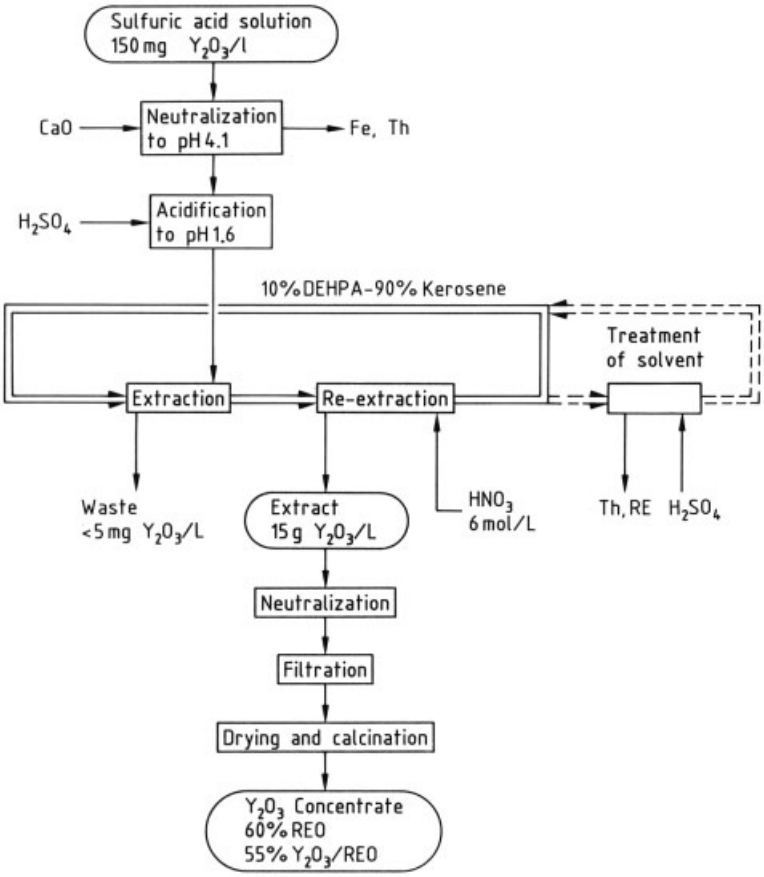

Figure 13. Denison Mines process for yttrium-rich rare earth concentrates [96]

The production capacity of Rhône-Poulenc is ca. 8000 t/a REO. Other companies producing rare earth compounds have mainly specialized in yttrium and europium, from which the red luminophors for television tubes are produced.

The Yao Lung Chemical Plant in Shanghai came on stream in 1964 and operates a similar process for cracking monazite as Rhône-Poulenc. The simplified flow sheet is described in [97]. Little is known of the Batou process for the beneficiation of the Baiyunebo ore in Inner Mongolia. The ore is principally mined for iron. A simplified flow sheet and principle of rare earth extraction and separation are described in [98].

5.4.3.1. Separation of Europium Oxide

The main problem in separating europium is its low content in the ores compared with the other rare earth metals. The classical minerals such as bastnaesite contain only 0.12 % Eu_2O_3 based on total REO content. Therefore a strongly enriched solution must first be prepared, from which the europium can be precipitated as its sulfate after first converting it to the divalent state.

The process used by Molycorp, Inc. is shown in Figure 15 [99]. The extraction consists of two consecutive cycles with several separation units operating with DEHPA. These concentrate the europium by a factor of 100 before it is precipitated.

5.4.3.2. Separation of Yttrium Oxide

The separation of yttrium oxide exploits the fact that yttrium is unique among the rare earth elements in that its position in the series of elements

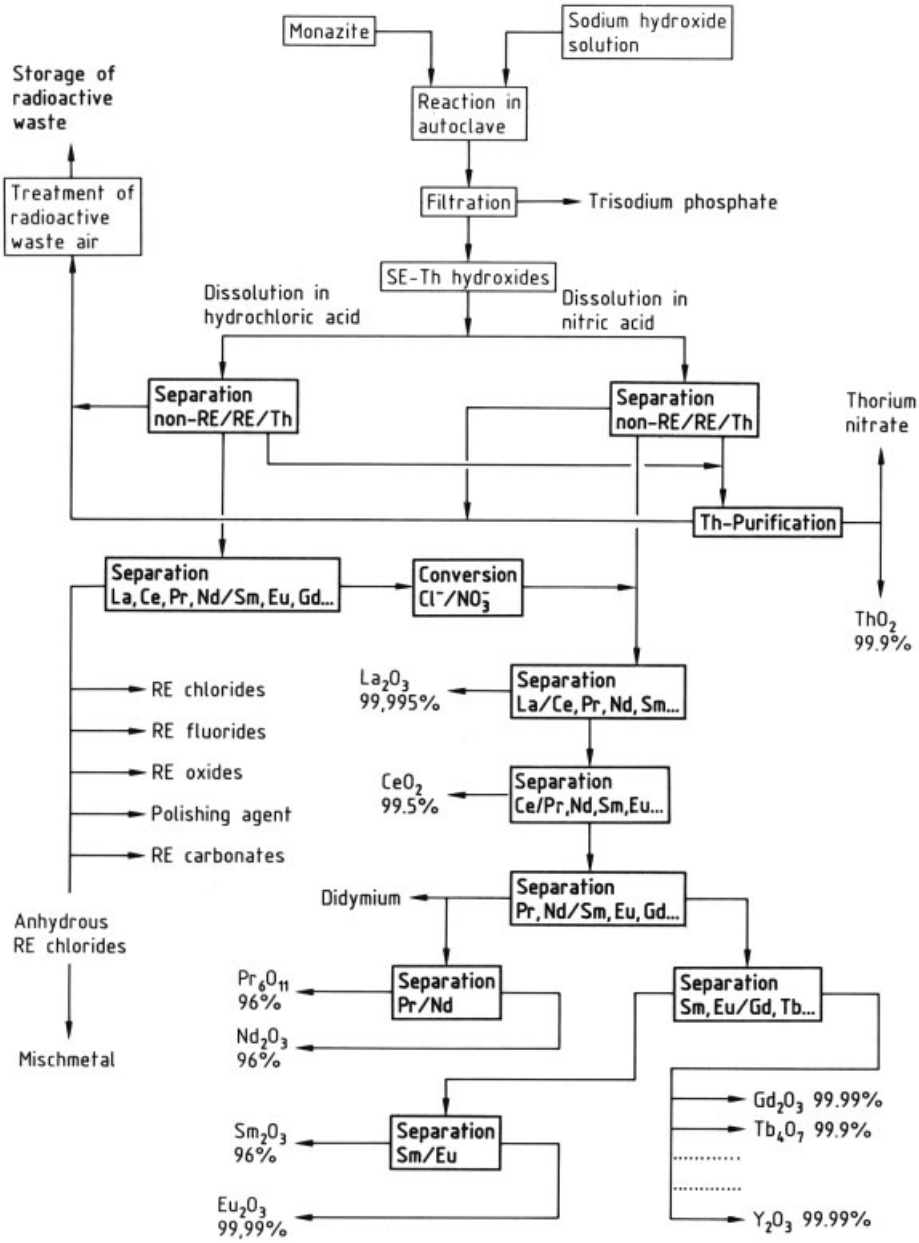

Figure 14. Rhône-Poulenc liquid – liquid extraction process for separation of the rare earth elements

with respect to separation operations is not constant. The company As Megon has developed a process for producing high-purity yttrium starting from a xenotime concentrate [100]. The three extraction cycles are shown in Figure 16.

The first cycle consists of a selective extraction by DEHPA, followed by three re-extraction units. Four fractions are obtained, in one of which the yttrium is concentrated to 75 %.

The second cycle uses the nitrate of tricaprylmethylamine as extractant. The yttrium accompanies Tm, while the cerium earths are extracted. A Y – Tm – Yb – Lu mixture containing 95 % Y is obtained.

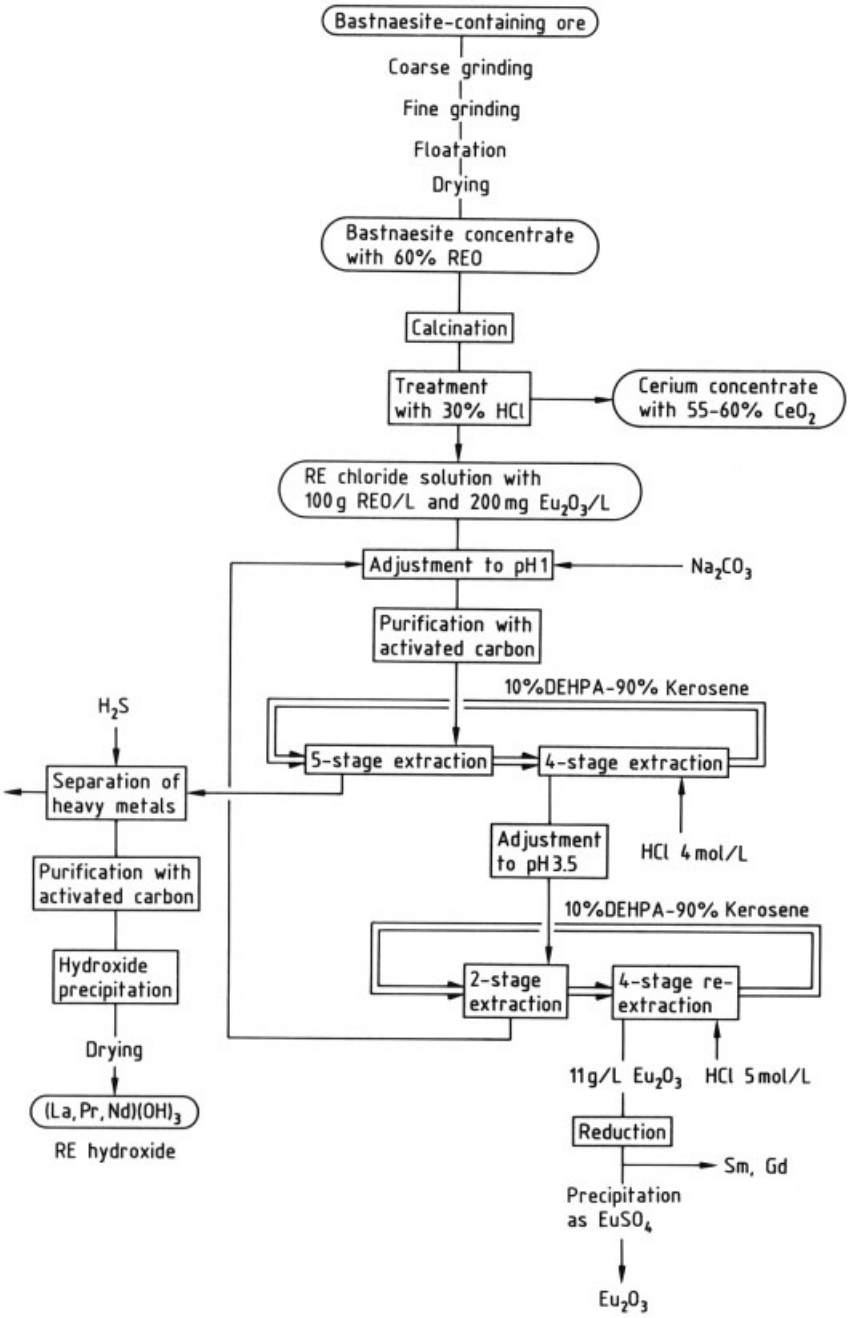

Figure 15. Molycorp process for the production of europium oxide [99]

The last cycle extracts Tm, Yb, and Lu by means of tricaprylmethylammonium thio-cyanate. Here, yttrium behaves like a cerium earth element. The only impurities in the yttrium oxide end product that exceed 1 ppm are Yb (2.5 ppm) and Er (3 ppm). From this product, 99.9999 % Y_2O_3 is obtained.

Molycorp starts from a concentrate containing 52 % Y_2O_3 in the REO mixture, and produces yttrium in two consecutive cycles

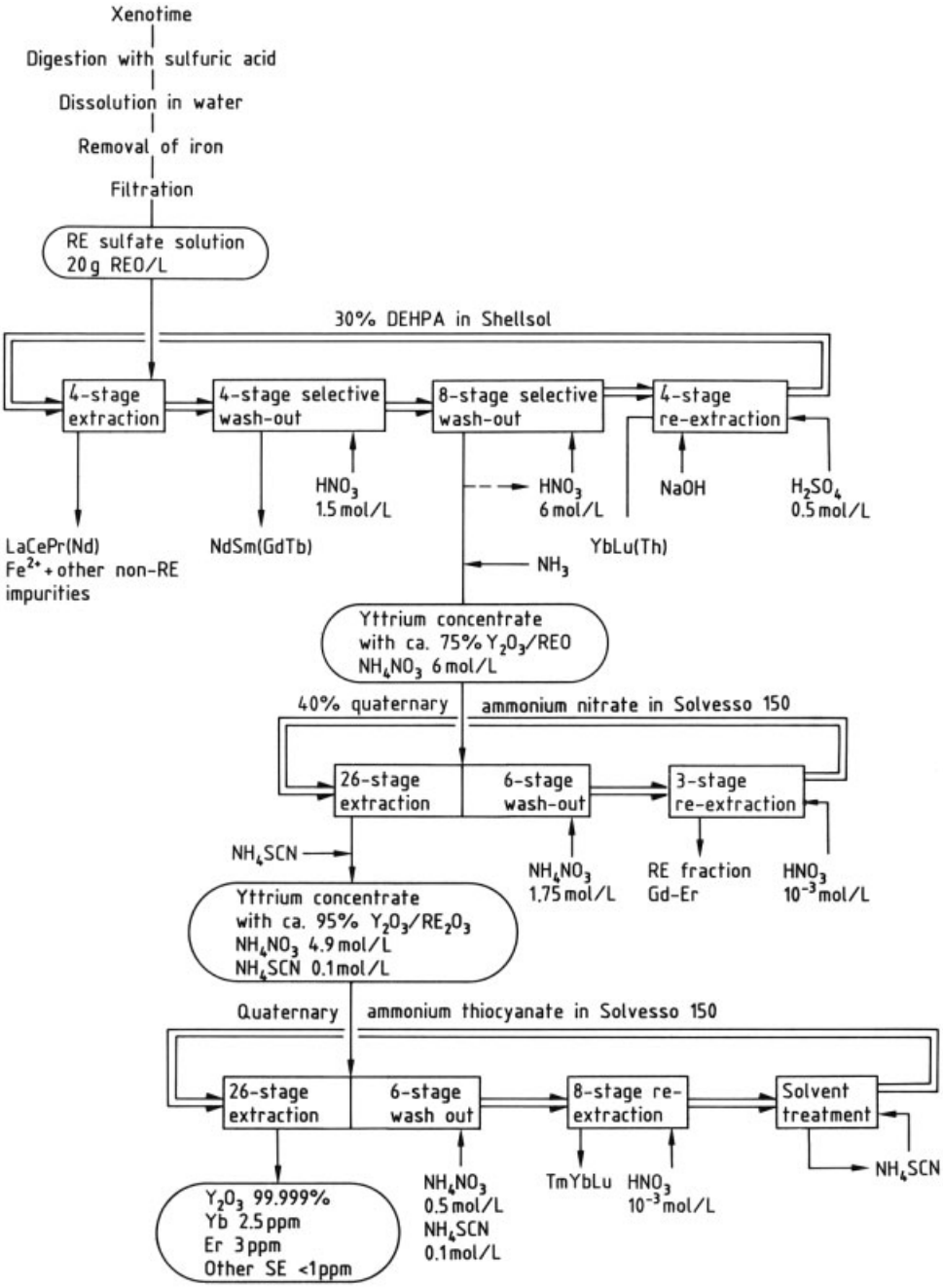

Figure 16. AS Megon process for high-purity yttrium oxide [100]

(Fig. 17), [101]. La – Er/Y – Lu are separated by Aliquat – NO_3^- with recycle, using the displacement effect of magnesium nitrate. Final purification of the yttrium is carried out with a Versatic acid, including a double recycle.

Thorium Ltd. developed the technique of total recycle in two consecutive systems. The first, 25-stage process operated with Versatic acid, and the second, 50-stage process with TBP. The product was 99.999 % Y_2O_3, starting from a

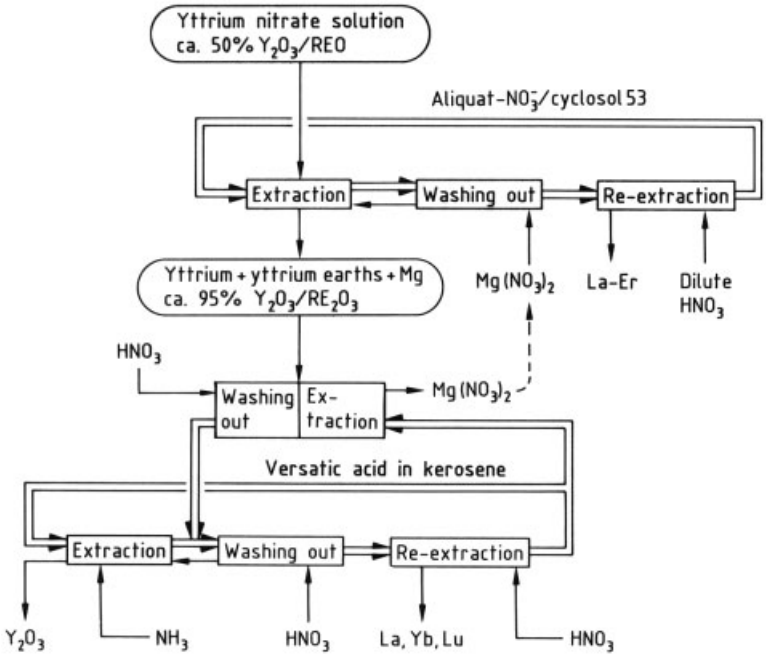

Figure 17. Molycorp process for production of pure yttrium oxide [101]

47 % concentrate [70, 75]. Details of other separation processes operated at one time by Thorium Ltd. have been published in [70].

5.4.3.3. Separation of Scandium

Scandium is present in all rare earth minerals, although only in trace amounts. Because it has similar chemical properties to the other rare earth elements, it cannot easily be obtained by selective precipitation as the oxalate or a double salt. The most effective separation is by extraction, scandium behaving quite differently from the other rare earth metals. For example, it is very readily extracted by diethyl ether from thiocyanate solution. TBP is an effective extractant from chloride solutions, while the other rare earth elements are hardly extracted. Organophosphoric acids such as DEHPA are much more effective extractants for scandium than for the other rare earth elements.

6. Production of the Metals

The electropositive character of the rare earth elements prevents their deposition from aqueous solution. Neither electrolytic processes in aqueous or organic media nor amalgam formation have never been very successful. However, they have achieved some importance for the production of high-purity compounds of samarium, europium, and ytterbium [102].

The rare earth metals are produced, both on the laboratory and industrial scale, by molten salt electrolysis and metallothermic reduction [1, part B 2], [2, vol 1, chap. 2, 173 – 232], [5, 6, 103].

6.1. Fused Salt Electrolysis

La, Ce, Pr, Nd, and the mixture of Nd and Pr known as didymium can be produced, like cerium mischmetal, by molten salt electrolysis of a mixture of anhydrous rare earth chlorides and fluorides with alkali and alkaline earth chlorides and fluorides [104] (see → Cerium, Cerium Alloys, and Cerium Mischmetal, Chap. 2.2.).

This process is made possible by the lower melting points of the cerium earth metals, whereas the yttrium earth metals have higher melting points and therefore do not melt during the electrolysis process.

A lower melting point can be achieved by using a cathode (e.g., Cd, Zn) that forms a low-melting alloy with the rare earth metal. This alloy can also have a lower density than the fused salt (e.g., Mg), so that it rises to the surface and can be removed from the fused salt [105]. The alloying element can then be removed by distillation.

By using molten Cd or Zn as cathode, Sm, Eu, and Yb can also be produced. Under the conditions used for cerium mischmetal, these elements are only reduced to the divalent state.

The production of rare earth metals and their alloys with 3*d* metals by molten salt electrolysis of the oxides dissolved in a fluoride melt has been described. The cerium earth metals are particularly suitable due to their low melting points [106].

The reactivity of the rare earth metals causes problems in the choice of construction materials of the electrolysis furnace. High-purity metals can be produced by using molybdenum, tungsten, or tantalum as the crucible and cathode materials. For industrial manufacture, iron crucibles with ceramic or graphite linings are used. If halogens are liberated during the electrolysis, carbon is used as the anode material. Molybdenum is also used in the electrolytic decomposition of the oxides in a fluoride melt.

6.2. Metallothermic Reduction

Metallothermic reduction of the rare earth oxides and anhydrous rare earth chlorides and fluorides can be used to produce high-purity rare earth metals, especially Gd to Lu, including Y. Metallothermic reduction is also used to produce rare earth mixtures with precisely controlled compositions that cannot be obtained by molten salt electrolysis. Alloys can also be produced by this process.

Alkali metals, alkaline earth metals, and aluminum are suitable reducing agents, as are alloys of these elements with each other. Lithium, which forms low-melting LiF, is of special importance, as is calcium. The use of Mg or Zn has the additional advantage of producing a low-melting alloy with the rare earth metal. The alloying elements can be removed by distillation, yielding the pure rare earth metal.

For the production of La, Ce, Pr, and Nd, the metallothermic reduction of the anhydrous rare earth chlorides is preferred. The reaction is carried out in crucibles lined with MgO at temperatures up to 1100 °C. At higher temperatures, reaction takes place between the rare earth metal and the MgO, and the rare earth chlorides vaporize. The process is not suitable for the production of Sm, Eu and Yb, which are merely reduced to the divalent state.

Gd to Lu and Sc, which have higher melting points, are obtained by reduction of the fluorides with Ca at 1500 – 1600 °C. The reduction is carried out in tantalum crucibles under a protective gas or in vacuum. The reaction temperature can be reduced by adding a "booster" such as iodine, which gives a slag with good flow properties that separates cleanly from the metal. Praseodymium is produced by reduction of the fluoride with lithium.

The Carlson – Schmidt apparatus is shown in Figure 18 [107]. This was used between 1957 and 1959 to produce high-purity yttrium metal in 50 kg batches.

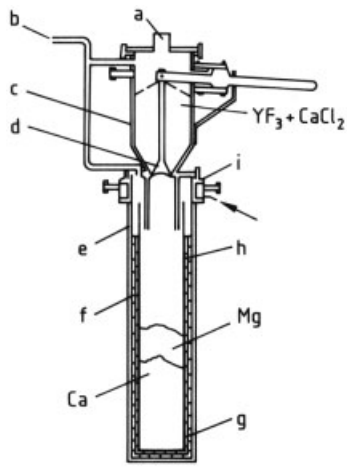

Figure 18. Reduction apparatus for the production of yttrium [107]
a) Sight glass; b) Vacuum connection and helium inlet; c) Charging tunnel; d) Charging shut-off mechanism; e) Steel reaction chamber; f) Titanium reaction crucible; g) Graphite insulation; h) Titanium or steel supporting crucible; i) Water cooling

Sm, Eu, and Yb can be produced by reduction of the oxides with La or the cheaper cerium mischmetal at 1000 – 1300 °C. The rare earth oxides and the reducing metal are used in the form of pellets, prepared from chippings or thin disks.

Sm, Eu, and Yb are volatile at the reaction temperature under vacuum ($< 10^{-4}$ bar), and can be distilled from the reaction space during the reaction and condensed on coolers. Thus, these metals can be separated from rare earth metals that are not volatile under these conditions and obtained in a pure state. Hence, the starting materials can consist of rare earth oxides in which Sm, Eu, and Yb have merely been concentrated. The principle of the reduction – distillation apparatus is illustrated in Figure 19 [108]. Further purification can be carried out by a second distillation.

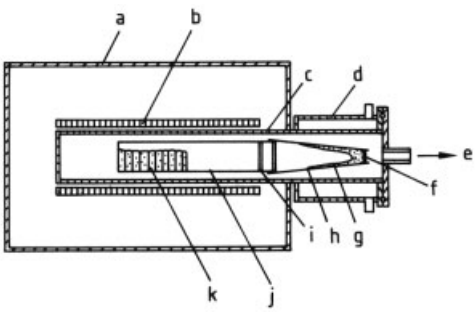

Figure 19. Reduction – distillation apparatus for the production of samarium, europium, and ytterbium [108]
a) Furnace; b) Heating element; c) Stainless steel container; d) Water cooling; e) Vacuum; f) Baffle; g) RE metal; h) Molybdenum condenser; i) Heat shield; j) Molybdenum crucible; k) Pelletized reactants

The calciothermic reduction of the rare earth oxides can be carried out such that a metal powder with a defined grain size distribution is produced. In this way, it is possible to produce alloys of rare earth metals with 3*d* metals that can be used to manufacture permanent magnets.

Methods of producing such alloys include, e.g., the coreduction process carried out under vacuum at 1000 °C for 3 h [109]:

$$RE_2O_3 + n\,Co_3O_4 + (10 - 3n)Co + (4n + 3)Ca \rightarrow 2\,RECo_5 + (4n - 3)CaO$$

where ($n = 0 - 10/3$), or the reduction – diffusion process carried out at 850 – 1150 °C for 5 h [110]:

$$RE_2O_3 + 10\,Co + 3\,CaH_2 \rightarrow 2\,RECo_5 + 3\,CaO + 3\,H_2$$

The CaO formed can be removed without affecting the properties of the metal powder.

If C and Si are used as reducing agents, rare earth carbides or silicides are formed. Prealloys or inoculation alloys containing rare earth metals are produced by reducing rare earth oxide mixtures or rare earth ore concentrates with CaSi or FeSi in an electric arc furnace.

6.3. Purification

Production of the pure rare earth metals necessitates the removal of products of the reaction of the metals with the atmosphere, crucible materials, and coreactants. Suitable methods inlcude melting under a protective gas or in a vacuum.

The high-boiling rare earth metals can also be purified by distillation. Tungsten powder is added to bind interstitial impurities (mainly C, N, O) as involatile compounds. Table 13 lists the temperatures at which the rare earth metals have vapor pressures of 1.3×10^{-3} to 1.3×10^{-5} bar [111].

Table 13. Vapor pressure and distillation rate of the rare earth metals [111]

RE element	Temperature, °C 1.3×10^{-5} bar	1.3×10^{-3} bar	Distillation rate at 1.3×10^{-3} bar, g cm^{-2} h^{-1}
Sc	1397	1773	33
Y	1637	2082	43
La	1754	2217	53
Ce	1744	2174	53
Pr	1523	1968	56
Nd	1341	1759	60
Sm	722	964	83
Eu	613	837	90
Gd	1583	2022	59
Tb	1524	1939	60
Dy	1121	1439	71
Ho	1179	1526	69
Er	1271	1609	68
Tm	850	1095	83
Yb	471	651	108
Lu	1657	2098	61

Alloying elements and impurities such as Mg, Cd, Zn, and Ca are distilled off under vacuum. The remaining rare earth metal is then in the form of a sponge, which can be consolidated by fusion in an electric arc furnace.

Zone melting, melt extraction, and melt filtration, e.g., through tungsten powder, degassing in a high vacuum, etc., are other recommended purification methods.

7. Analysis

Analytical methods for the determination of the rare earth elements are reviewed in [5, pp. 570–593], [112]. Modern analytical methods are reported in [2, vol. 4], [113, 114].

Analysis may be divided into determination of individual rare earth elements, and group determination, in which the rare earth elements are coprecipitated as oxalates, hydroxides, or carbonates, and are then determined quantitatively after calcination to give the oxides. Complexometric titrations are also used for group analysis.

The ready conversion of cerium to the tetravalent state and of europium to the divalent state offer simple methods of titration for the single elements.

The determination of individual colored rare earth ions is normally carried out by spectrophotometry, emission spectroscopy, and spectroscopy in flames. Emission, absorption, or fluorescence methods are used. The technique and application of laser source mass spectrometry is discussed in [115].

X-ray fluorescence analysis can be used to determine concentrations in the range between 90 % and ppm quantities with moderate sensitivity. It is useful to concentrate the rare earth elements (sometimes with the aid of a trace collector), and to determine the individual rare earth metals in the separated oxide, since a direct determination on the matrix can lead to erroneous results due to the presence of foreign elements or to the physical or chemical condition of the sample. For the determination of yttrium, the K lines are preferably used, although the L lines are preferred for the other rare earths due to the disturbance of the K lines by bremsstrahlung spectra. Inner standards such as strontium or zirconium with yttrium are used, or standard samples that should be as similar as possible to the sample being tested. The spectra can then be used for quantitative determinations. The samples are prepared by compressing the material for analysis with borax.

8. Compounds

8.1. Hydrides [2, vol. 3, chap. 26, pp. 299–336]

All the rare earth metals react with molecular hydrogen to form the dihydrides, REH_2. Most rare earth metals will also react with hydrogen to produce the trihydrides, REH_3. The rate of reaction depends on temperature, hydrogen pressure, and the surface condition of the metal. Both forms of the hydride are nonstoichiometric, showing wide ranges of homogeneity. The dihydrides of La, Ce, Pr, and Nd have ranges of homogeneity with a composition approaching REH_3. The dihydrides of Sm, Gd to Tm, Lu, and Y have narrower ranges of existence and form a trihydride phase that is also nonstoichiometric. ScH_2, EuH_2, and YbH_2 are almost stoichiometric, but YbH_2 takes up additional hydrogen at very high pressures, up to the composition $YbH_{2.55}$. With the exception of EuH_2 and YbH_2, which form orthorhombic crystal structures similar to the alkaline earth hydrides, the dihydrides are cubic (fluorite structure). The trihydrides crystallize with a hexagonal structure.

Rare earth hydrides are brittle compounds, sometimes with a metallic luster. Similar to the transition metal hydrides, the rare earth dihydrides (except for YH_2 and EH_2) are metallic conductors. In particular, hydrogen-deficient dihydrides are better electronic conductors than their parent metals. In general, as the trihydride composition is approached the materials begin to show semiconducting properties.

Hydrides formed by the hydrogenation of binary rare earth solid solution alloys are usually considered to be solid solutions of the two parent rare earth hydrides, and as such their properties can be predicted with reasonable confidence from those of the constituent hydrides. In the case of binary alloys REM, where M is a non-rare- earth metal, the properties of the intermetallic hydride formed by hydrogenation of the alloy bear no relationship to those of the parent hydrides. They are therefore generally viewed as pseudobinary metal hydrides. Some of these materials can reversibly and efficiently absorb and desorb considerable amounts of hydrogen, and are therefore of commercial interest as a means

of storing hydrogen (see 8.1). A review of hydrogen absorption in intermetallic compounds is provided in [116].

8.2. Oxides, Hydroxides, Peroxides, Salts of Inorganic Oxoacids, Double Salts

Aspects of the transformation and defect structures in rare earth sesquioxides is provided in detail in [117]. The sesquioxides, RE_2O_3, of the rare earth elements, with the exception of Ce, Pr, and Tb, are obtained by oxidation of the metals with oxygen at elevated temperatures or by thermal or pyrohydrolytic decomposition of the hydroxides, halides, and certain oxosalts. In the case of Ce, Pr, and Tb, similar conditions yield the compounds CeO_2, Pr_6O_{11}, and Tb_4O_7, which can be reduced to the sesquioxides by hydrogen at elevated temperature. The sesquioxides melt at very high temperatures (in the range of 2200–2400 °C) and have very high enthalpies of formation [−1907.9 kJ/mol (Sc_2O_3) to −1661 kJ/mol (Eu_2O_3)].

They crystallize in the hexagonal system (type A), monoclinic system (type B) or cubic system (type C), depending on the temperature and the radius of the cation. At temperatures near the melting point, high-temperature modifications, types H and X, are also formed. Most sesquioxides are polymorphic, not including the two high-temperature forms. La_2O_3 forms type A crystals; Ce_2O_3 to Nd_2O_3 form types A and C; Sm_2O_3 to Ho_2O_3 form types C and B; and Er_2O_3 to Lu_2O, Y_2O_3, and Sc_2O_3 form only type C crystals. Whether the observed modifications are thermodynamically stable or metastable has so far been established in only a few cases.

Cubic crystals (fluorite type) are formed both by the dioxides of Ce, Pr, and Tb and by their intermediate oxides, in which cations in the two oxidation states 3+ and 4+ are present together. With the exception of CeO_2, which can be produced by air oxidation, the dioxides can only be obtained under strongly oxidizing conditions (e.g., oxygen at high pressure). The nonstoichiometric intermediate oxides have been extensively investigated in the system Pr–O [118]. Here, phases exist that are members of a series with the general composition Pr_nO_{2n-2} ($n = 7$, 9, 10, 11, 12).

The compound EuO is fairly easy to obtain by reduction of Eu_2O_3, e.g., with EuH_2 or Eu (g). With this exception, the condensed monoxides of the rare earth elements have only been observed as metastable compounds or as compounds stable under extreme conditions such as high pressure. For example, gaseous monoxides of all the rare earth elements have been observed at very high temperatures. EuO is a violet-black, cubic (NaCl type), ferromagnetic substance ($T_C = 70$ K). Eu also forms the oxide Eu_3O_4, in which both Eu^{2+} and Eu^{3+} ions are present.

The hydroxides, $RE(OH)_3$, are obtained as hydrates by the precipitation of RE^{3+} ions with ammonia or other bases. The hydroxides and oxide hydroxides, REOOH, can be prepared in crystalline form by hydrothermal synthesis. The hydroxides are sparingly soluble in water. Their basicity and solubility decrease with increasing atomic number. This progressive change in properties is important both for the separation of the rare earth elements from one other (see Section 5) and for chemical analysis. The hydroxides react with carbon dioxide to form basic carbonates, the rate of reaction depending on their basicity.

So far, very little is known of the existence of peroxides of the rare earth elements. Cerium peroxide forms as a deep brown, hydrated compound when Ce^{3+} ions are precipitated with ammonia and hydrogen peroxide. This intense colour can be used for the detection of Ce in the presence of other rare earth elements.

Of the group of compounds of rare earths with inorganic oxoacids, the carbonates, sulfates, nitrates, and certain double salts are of importance for analytical and preparative chemistry. Precipitation of RE^{3+} ions by solutions of alkali metal carbonates yields the sparingly soluble hydrated or basic carbonates. These are readily decomposed by acids or on heating, liberating CO_2. Apart from lanthanum sulfate, which forms a nonahydrate, the sulfates crystallize from aqueous solution as octahydrates. The hydrated sulfates can be thermally decomposed to the anhydrous form. They are fairly soluble, having a negative temperature coefficient of solubility, and are of industrial importance since sulfuric acid is used for digesting certain rare earth

ores. Double sulfates, e.g., salts of the type $RE_2(SO_4)_3 \cdot Na_2SO_4 \cdot n\,H_2O$ are important in the separation of the cerium earths from the yttrium earths.

The rare earth nitrates crystallize from solution in dilute nitric acid in the form of hydrates, $RE(NO_3)_3 \cdot n\,H_2O$ ($n \leq 6$). These salts are very soluble in water and moderately soluble in various organic solvents (alcohols, ketones, esters), and are important in the separation of the rare earth elements from each other by liquid – liquid extraction (see 5.4.2.1). Certain double nitrates, e.g., of the type $2\,RE(NO_3)_3 \cdot 3\,Mg(NO_3)_2 \cdot 24\,H_2O$ were at one time important in the separation of the rare earth elements by fractional crystallization. A comprehensive review of inorganic complex compounds is provided in [119, 120].

8.3. Halides

The rare earth elements, except for europium, form trihalides, REX_3, where X = F, Cl, Br, I. The trifluorides are sparingly soluble and are obtained from aqueous solution by precipitation with hydrogen fluoride. If these precipitates are dried at elevated temperatures, partial hydrolysis takes place, particularly of the more weakly basic elements. This can be effectively prevented by adding an excess of NH_4F or HF(g) during the dehydration process. With the exception of ScF_3, the trifluorides crystallize in the hexagonal LaF_3 structure ("tysonite structure") or the orthorhombic YF_3 structure, depending on the radius of the cation and the temperature range. ScF_3 crystallizes in the cubic system (REO_3 type). When dimorphism occurs, the YF_3 type is the low-temperature, and the LaF_3 type, the high-temperature modification. The trifluorides are high-melting compounds [1140 °C (HoF_3) to 1552 °C (ScF_3)]. EuF_3 and, to a very limited extent, YbF_3, tend to liberate fluorine at high temperatures. A review of the rare earth fluorides is given in [117].

The trichlorides can be obtained as their hexa- or heptahydrates by careful evaporation of their solutions in hydrochloric acid. The dehydration of the hydrates must be carried out in a stream of HCl and/or in the presence of an excess of NH_4Cl to prevent hydrolysis. On an industrial scale, the anhydrous trichlorides are obtained by reacting oxidic ores with chlorine and carbon (see 5.4.3). They crystallize in the hexagonal UCl_3 type structure ($LaCl_3$ to $GdCl_3$) or the monoclinic YCl_3 type structure ($DyCl_3$ to $LuCl_3$, YCl_3, $ScCl_3$). They are very soluble in water and in certain organic solvents such as ethanol and pyridine. Their melting points are in the range 588 ($TbCl_3$) to 892 °C ($LuCl_3$). With the exceptions of $EuCl_3$ and $YbCl_3$, which decompose on heating with liberation of chlorine, all the other rare earth chlorides vaporize without decomposition or sublime ($ScCl_3$).

The tribromides and triiodides are obtained in the same way as the trichlorides. The tribromides, and more so the triiodides of Eu, Yb, and Sm, have a marked tendency to lose halogen on heating, since the 2+ oxidation state of the cation becomes more stable with increasing size of the anion. The melting points of the tribromides are in the range 664 ($SmBr_3$) to 960 °C ($LuBr_3$), and those of the triiodides in the range 738 (PrI_3) to 1045 °C (LuI_3). The trihalides react with the sesquioxides to form oxide halides of various compositions.

The existence of tetrahalides is limited to the tetrafluorides of Ce, Pr, and Tb. CeF_4 and TbF_4 are obtained by the reaction of the elements or the trifluorides with fluorine. PrF_4 can only be obtained by the extractive removal of NaF from the readily prepared Na_2PrF_6 by treating it with liquid hydrogen fluoride in a fluorine atmosphere. The tetrafluorides are white solids which crystallize in the monoclinic system (UF_4 or ThF_4 type). They are slightly soluble in water but decompose to the trifluorides and are thermally unstable. Cerium tetrafluoride is appreciably more stable than the tetrafluorides of praseodymium and terbium.

The dihalides may be divided into two groups: the mainly salt-like compounds, in which the cation has a valency of 2+, and those compounds whose properties (e.g., electrical conductivity) indicate that the bond has some metallic character, and can be represented, in a much simplified way, by the general formula $Ln^{3+}(e^-)X_2$. Important methods of preparation include reduction of the trihalides by hydrogen at elevated temperatures (dichlorides, dibromides, and diiodides of Eu, Yb, and Sm, as well as EuF_2), or by solid or gaseous rare earth metals.

Thermodynamically stable difluorides are known only for Eu, Sm, and Yb. However, the

difluorides of almost all the other rare earth elements can be prepared as metastable compounds which crystallize in the cubic system (fluorite type). EuF_2, YbF_2, and SmF_2 can take up considerable quantities of the corresponding trifluorides in solid solution. Other dihalides of many rare earth elements can be prepared, but it is difficult to give definite trends for the occurrence of salt-like or metallic properties. However, it can be stated that all the dihalides of Sm, Eu, and Yb are of a mainly salt-like nature.

8.4. Chalcogenides

The rare earth elements form chalcogenides with the general formula REZ, RE_3Z_4, RE_2Z_3 (RE_5Z_7), and RE_2Z_4 (where Z = S or Se) [3, vol. 3, pp. 149 – 248], [121]. With Te, compounds are known with the formulae RETe, RE_3Te_4, RE_2Te_3, RE_4Te_7, $RETe_2$, RE_2Te_5, and $RETe_3$, of which some must be regarded as intermetallic compounds. The rare earth chalcogenides exhibit a wide range of structural, electrical, and magnetic properties. Most of them are very sensitive to hydrolysis and can therefore only be prepared and handled if moisture is excluded.

The monochalcogenides are mainly prepared by direct synthesis from the elements, but may also be prepared in some cases (YbS, YbSe, SmS, SmSe [20, 122]) by the reduction of higher chalcogenides with the corresponding gaseous rare earth metal. EuS and EuSe are also obtained by the reaction of Eu_2O_3 with H_2S/H_2 or H_2Se/H_2 at ca. 1100 °C. All the rare earth monochalcogenides crystallize in the cubic system (NaCl type) and have high melting and boiling points. Thus, for example, the compound CeS has been recommended as a sulfide ceramic material for coating vessels due to its inertness to many metals [123]. It begins to liquefy above 2100 °C, and melts at 2450 °C. Despite the 1 : 1 stoichiometry, only the monochalcogenides of Eu, Yb, and Sm can (within limits) be regarded as compounds of RE^{2+}. The other compounds have properties that show the simultaneous presence of ionic and metallic bonding.

The sesquichalcogenides, RE_2Z_3 (Z = S, Se, or Te), can also be prepared by direct synthesis from the elements (the only practical method for the tellurides), or by reaction of the sesquioxides or trihalides with the hydrogen chalcogenides at high temperatures. When the oxides are used, very stable oxide chalcogenides such as RE_2O_2Z can be produced as intermediate products. The sesquichalcogenides can exist in several modifications. The cubic modification is nonstoichiometric, with the approximate composition RE_3Z_4 (Th_3P_4 type).

Rare earth polychalcogenides include the polysulfides RE_2S_4. With the exceptions of the polysulfides of Tm and Lu, they are obtained by heating the sesquisulfides with an excess of sulfur. They are interesting in that they are not RE^{4+} compounds, but contain two S^{2-} ions and one S_2 group per formula unit.

The crystal chemistry and phase equilibria of rare earth ternary systems with chalcogen elements is dealt with in [124].

8.5. Nitrides

The rare earth elements react with nitrogen to give nitrides of the type REN [125], and with the elements P, As, Sb, and Bi to give compounds with other stoichiometric compositions. The nitrides are formed in slow reactions of nitrogen with the metals or hydrides in very finely divided form at ca. 1000 °C, or more readily by reaction with ammonia at ca. 700 °C. The nitrides are cubic (NaCl type) and have very high melting points and decomposition temperatures. They have a metallic appearance when in a compact form, and are ionic compounds with metallic conductivity or semiconductors. Apart from ScN, all the rare earth nitrides are very sensitive to hydrolysis, and can therefore only be handled when moisture is excluded.

8.6. Carbides

Of the rare earth compounds with elements of group 14, only the carbides are described here. Compounds with the formulae REC_2, RE_2C_3, RE_2C, and REC_{1-x} are known. Of these, the carbides REC_2 represent the largest group [3, vol. 3, pp. 284 – 342]. The carbides can be synthesized directly from the elements. The high-melting dicarbides are highly involatile, crystallize in the tetragonal system (CaC_2 type), and are metallic conductors with approximately one

conducting electron per C_2 group. Hydrolysis of the carbides under neutral or weakly acidic conditions leads to complex mixtures of unsaturated and saturated hydrocarbons, most of which are C_2 hydrocarbons due to the presence of C_2 groups in the crystal structure.

8.7. Complexes; Organic Compounds

From the point of view of complex-forming properties, the rare earth elements resemble the elements of group 2 more closely than those of the 3*d* transition elements, owing to the peculiarities described in Section 3.2.4. Only very strong ligands, with strong chelate-forming properties, are able to form thermodynamically stable, isolable rare earth complexes.

Many rare earth chelates are very difficult to isolate from aqueous solution, as water is an extremely strong ligand for rare earth ions. Also, ligand exchange reactions take place very rapidly in solution, which makes it impossible to isolate definite isomers. Of the rare earth chelate compounds so far prepared (e.g., [3, vol. 3, pp. 61 – 128], [126]), those useful for liquid – liquid extraction and separation of rare earth elements are of particular interest. The technically important chelating agents include certain ketones, but, above all, tri-*n*-butylphosphate (TBP) [3, vol. 1, pp. 89 – 109]. The chelates of the rare earth elements with 2,4-pentanedione (acetylacetone) and its derivatives are also important, these being characterized by great stability and sometimes considerable volatility [127].

When considering organic compounds [1, part D 1 – D 3], [128], a distinction should be made between true organometallic compounds and salts of organic acids or coordination compounds with organic ligands. The relatively small number of true organic compounds (e.g., the tris(cyclopentadienyl) derivatives [129] and the triphenylene and anionic tetraphenyl complexes [130, 131]) are mostly very reactive, sensitive substances. Of the large group of salts of organic acids, the oxalates deserve mention. These are formed as sparingly soluble precipitates from solutions in moderately concentrated hydrochloric acid and are important for the purification of the rare earth elements, and also for their quantitative determination. Organometallic compounds of the rare earth elements are discussed in [132].

9. Uses [3, vol. 1, pp. 416 – 500; vol. 2, pp. 190 – 297]

The rare earth metals and their compounds are used in numerous areas of industry for a wide range of purposes. The most important of these include metallurgy, catalysts in the chemical industry, coloring of glass and ceramics, the production of magnets, and phosphors.

Metallurgy. The importance of the rare earth metals in metallurgy is highlighted in [133, 134]. In 1990, ca. 22 % by volume of rare earth consumption in the United States was devoted to metallurgical end use. The equivalent volume usage in Japan by contrast was ca. 10 %. One of the most important areas of application, is based on their reactivity with water, hydrogen, nitrogen, sulfur, and the so-called tramp elements such as arsenic, antimony, and bismuth. Sulfur reacts to form rare earth sulfides or oxysulfides in spheroidal form which are not deformed under rolling conditions, and which form solid particles at the temperature of molten steel, thus improving the hot formability and flexibility of microalloyed steels, and reducing anisotropy in the notched toughness test. Rare earths are used in the form of mischmetal containing principally Ce, La, and Nd.

The addition of rare earth metals such as yttrium and cerium, or cerium mischmetal imparts improved oxidation resistance to heating element alloys, to substrate alloys for catalysts, and to superalloys. Rare earth metals improve the precipitation of spheroidal graphite in SG iron casting. Although for economic reasons cerium mischmetal is more often used in these applications than the pure rare earth metals, the pure metals can give better results when used separately.

Addition of rare earths to copper and aluminum alloys, used as electrical conductors, improves the mechanical properties. In China, an yttrium – magnesium – aluminum alloy has been developed for transmission cabling. Yttrium improves tensile strength, heat resistance, vibration resistance and can raise electrical conductivity by 50 %. Creep resistance and tensile strength

at high temperatures are improved by addition of rare earth metal (mainly neodymium, praesodymium, and yttrium) to magnesium alloys for pressure casting and aircraft construction. Significant improvements to high strength Al–Mg and Al–Li alloys are achieved by the addition of scandium. Similar effects can be achieved with titanium and titanium alloys, and with alloys of niobium, tantalum, and vanadium.

The rare earth metals react readily with gaseous elements, and are therefore used as getters. Their pyrophoric properties are utilized in friction igniters ("flints") (→ Cerium, Cerium Alloys, and Cerium Mischmetal) and in military projectiles.

Alloys of the type $LaNi_5$ are capable of absorbing hydrogen, even from mixtures with other gases, forming rare earth hydrides. These are used to store hydrogen reversibly. In these alloys La can be partially replaced by Ce, Pr, and Nd, and Ni by Co, Cr, Cu, Fe, or Mn. They can also be used to separate D_2 from H_2 [135–138].

Rare earth alloys with 3 d transition elements have magnetic properties, [2, vols. 14, 16], [139, 140]. These alloys are of the type REA_5 or RE_2A_{17}, in which the rare earth metals can be cerium earths, especially Sm, and the 3*d* transition metals can be pure Co, or Co partially replaced by Fe, Mn, Cr, or Cu. These alloys give extremely stable magnets, having energy products, high remanence, and high coercive field strengths [141–143] (see also →Magnetic Materials, Chap. 4.2.). Some typical properties are given in Table 14. These high-power magnets are used in small motors, printers, quartz watches, headphones, loudspeakers, magnetic storage, traveling-wave tubes, etc., mainly in miniaturized equipment.

Rare earth alloys also exhibit very large magnetostriction at cryogenic temperatures. $REFe_2$ Laves phase compounds show very large magnetostriction but possess intrinsic magnetocrystalline anisotropy [2, vol. 2, chap. 15, pp. 231–258]. For many applications, high strain is required at low magnetic fields, and therefore anisotropy must be minimized. The most important material to emerge from research activity so far is the ternary alloy $Tb_{0.27}Dy_{0.73}Fe_2$, which shows promise for high power actuators, transducers, and sensors.

Catalysts [117, vol. 5, chap. 43, 217–320]. A further important use for rare earth elements (as their chlorides or nitrates) is in zeolite cracker catalysts for the improvement of gasoline yields and reductions in the formation of coke and light hydrocarbons. These catalyst compositions can correspond either to the natural composition of the cerium earths or can be based mainly on cerium [144].

Rare earth elements are also used as catalysts for alkylation, isomerization, hydrogenation, dehydrogenation, dehydration, reforming of hydrocarbons, polymerization, oxidation of CO and hydrocarbons [145, 146], and reduction of the concentration of nitrogen oxides in automobile exhaust gases [147]. The formation of hydrides enables them to be used as methanation catalysts [148]. Cerium stabilizes γ-Al_2O_3 as a carrier for noble metal catalysts [149, 150]. Rare earth chelates with diketones can replace lead compounds as antiknock agents [151].

Energy Production. Rare earth elements have a possible future use in energy production. $RENi_5$ [152], La_2O_3, Y_2O_3, and CeO_2, with Ni, Co, Cr, or Zn [153], can be used as electrodes in fuel cells, while the solid electrolyte can be ZrO_2 stabilized with Y_2O_3, $ZrO_2-Y_2O_3-Ta_2O_5$, or $ZrO_2-CeO_2-Y_2O_3$ [154, 155].

Water, in the form of a photooxidizable compound with Eu^{2+}, is reduced in the first stage, and in the second stage the Eu^{3+} formed is reduced by a light-sensitive transition metal complex [156]. Water can be oxidized by Ce^{4+} in the presence of the redox catalyst PtO_2 or IrO_2, and the Ce^{3+} produced reconverted photochemically to Ce^{4+}, with liberation of H_2 [157]. The low Curie point (20 °C) of Gd can be utilized to produce a heat pump [158].

In magnetohydrodynamic generators, $LaCrO_3$ [159] and $LaCoO_3$ [160] are used as electrodes, and LaB_6 [161] as an electron emitter.

Glass and Ceramics. Another important use of rare earth elements is in the glass and ceramic industries [162]. Rare earth oxides in their naturally-occurring composition or with an increased CeO_2 content (or pure CeO_2) are used as polishing agents in the glass industry. Other uses of Ce include the chemical decolorization of glass ($Fe^{2+} \rightarrow Fe^{3+}$) [163], reduction of UV

Table 14. Characteristic properties of selected rare earth based permanent magnetic materials

Material	Energy product $(BH)_{max}$, kJ/m^3	Remanence B_r, mT	Coercivity before polarisation H_{cb}, kA/m	Coercivity after polarisation H_{cj}, kA/m
$SmCo_5$	160	920	2500	680
$(Sm_{0.4}Pr_{0.6})Co_5$	200	1050	1200	720
$(Sm_{0.2}M_{0.8})Co_5$	120	800	1000	580
$Sm(Co,Cu,Fe)_5$	110	770	300	290
$Sm(Co,Cu,Fe,Zr)_{6.9}$	240	1130	580	540
$Sm(Co,Cu,Fe,Zr)_{7.5}$	210	1080	800	780
NdFeB	300	1300	1000	700
NdDyFeB	250	1100	2000	880

transparency of glass containers for foodstuffs and medicines, and the prevention decolorization of television screens by X-rays and γ rays.

Nd (blue to wine red), Pr (green), Er (pink), and Ho (blue) can be used both for coloring glass and for color neutralization. Combinations of these with each other and with other elements such as Ti(+Ce: yellow), Se(+Nd: violet), and Ni(+Nd: reddish) give further effects.

In optical glasses for lenses, La improves the refractive index and dispersion. La and Gd also improve chemical and mechanical properties, and enable these types of glass to be used under the demanding conditions of magneto-optical and electro-optical systems [120, chap. 58, pp. 1–90].

In very pure types of glass, Eu^{2+} and Ce^{4+} give a phototropic effect in sunlight.

In the ceramic industry, rare earth oxides are used as yellow pigments (Pr) [164], as opacifiers in glazes and enamels (Ce) [165], for radiation-shielding ceramics (Sm, Eu, Gd, Dy) [166], to improve the light fastness of lead chromate and titanium dioxide pigments, to impart "natural" fluorescence to artificial teeth (Eu, Sm, Ce) [167–169], and especially as stabilizers for cubic ZrO_2 [170]. An interesting application is in so-called lambda probes, in which Y_2O_3-stabilized ZrO_2 is the solid electrolyte in a measurement/control system that optimizes oxidation and reduces the content of harmful gases in automobile exhausts [171]. In a similar application, it is used for determining the O_2 content of liquid steel [172].

CeS and Y_2O_3 can be used for the manufacture of crucibles and as temperature-resistant materials for gas turbines [173]. The mechanical properties of Si_3N_4 are improved by the addition of rare earths to promote sintering [174].

Electronics. Compounds of the type $Pb_{1-x}La_x(Zr_{1-y}Ti_y)O_3$ (PLZT) are used as electro-optical ceramics in displays [175], as electromechanical transducers [176], and for their piezoelectrical properties [177]. Lanthanum doped $BaTiO_3$ is used as a dielectric in capacitors [178], and in thermistors [179].

Compounds of the garnet type such as $Y_3Al_5O_{12}$ (YAG), $Y_3Fe_5O_{12}$ (YIG), and yttrium orthoaluminate $YAlO_3$ (YALO), doped with neodymium, are used as solid state lasers [2, vol. 4, pp. 275–315]. Solid state Nd–YAG lasers (→ Lasers, Chap. 3.3.), operating at 1064 nm wavelength, are most popular for cutting, welding and metal heat treatment. In Japan production is growing at ca. 10 %/a. Ho–YAG and Er–YAG lasers, emitting at 2000 nm and 2900 nm, respectively, have been developed to partly replace CO_2 lasers for use in microsurgery. In 1986 Sumitomo Metal Mining Company produced single crystals of gadolinium aluminum scandium garnet (GASG). These laser materials doped with nickel and chromium are expected to capture some of the tunable and high-power laser markets. $Y_3Fe_5O_{12}$ (YIG) is also used in frequency control applications [180, 181]. $Gd_3Ga_5O_{12}$ (GGG) crystal contains 53.7 % gadolinium oxide and is used for bubble type magnetic information storage [182] as well as in the doped form (ca. 3 atom %Nd) for high-power pulsed lasers. Rare earths used in laser technology are of very high purity (> 99.999 %) and therefore have high added value.

Rare earth–transition metal alloys deposited onto glass or polycarbonate provide an excellent magneto-optical medium for high density data storage and retrieval. Films are generally deposited by sputtering, and the targets used in

the process usually contain terbium, gadolinium, dysprosium, or neodymium together with cobalt and iron and other minor elements. Rare earth halides of Eu and Yb have semiconductor properties [183].

Phosphors [2, vol. 4, pp. 237 – 274] (→ Luminescent Materials). Stimulation by UV, X-rays, or electron beams causes certain rare earth elements to emit light of a definite wavelength. The energy is transferred from the host lattice to an activator. The following can act as host lattices: oxide sulfides and vanadates of Y, La, Gd, and Lu (favorable $4f$ electron configurations), and borates, aluminates, phosphates, silicates, tungstates, fluorides, compounds of the ruby type, etc. Phosphors containing rare earth oxides are used in color cathode ray tubes, X-ray intensifying screens, and fluorescent lamps. The following are used as red emitting phosphors in cathode ray tubes for TV and VDU screens: $YVO_4:Eu^{3+}$ (4.5. %), $Y_2O_3:Eu^{3+}$(3.5 %), and $Y_2O_2S:Eu^{3+}$ (3.65 %). Examples of green-emitting phosphors are $La_2O_2S:Tb^{3+}$, $CaS:Ce^{3+}$, and $SrGa_2S_4:Eu^{2+}$. Blue phosphors include $Sr_5Cl(PO_4)_3:Eu^{2+}$ and $ZnS:Tm^{3+}$. Other rare earth activators are Ce^{3+} (UV, blue), Pr^{3+} (red, green, blue, UV), Nd^{3+} (IR, lasers), Sm^{3+} (orange), Eu^{2+} (blue, green), Tb^{3+} (green), Dy (yellow, white), Er (green, lasers), Ho (green, yellow, lasers), Tm (blue), and Yb (green). Other areas of use include high-sensitivity X-ray films, optical scanning devices, photocopying equipment, and scintillation crystals [184]. Rare earth fluorides and oxides are still added to the carbon of classical carbon arc lamps, e.g., for cinemas, to improve light emission [185]. In 1991, 73 % by volume of the rare earths used in phosphors was associated with yttrium, 5 % with europium, 4 % with terbium, and 18 % with other rare earth elements.

Reactor Technology. Sm, Eu, Gd, and Dy have high thermal neutron capture cross sections. Eu and Dy can take up five neutrons in succession. Therefore, they have a low rate of burn off when used in control rods or in ceramics for radiation protection [166]. Table 15 lists the thermal neutron capture cross sections of these elements.

Other Applications. Rare earth compounds are used in the paint industry as dryers [186]. They are also used in the textile industry [187] and as oxidizing agents for self-cleaning ovens [188, 189].

Europium complexes are used as shift reagents in NMR spectroscopy [190].

Rare earth elements can be used in the third stage of wastewater purification for the precipitation of phosphates [191].

10. Economic Aspects [192]

World production of rare earth minerals totalled ca. 96 500 t (57 000 t REO) in 1990, compared with an estimated output of ca. 107 500 t (64 000 t REO) in 1989. The decrease in mineral output over these two years has been associated with a decrease in Australian and Chinese production. In particular, up to 1988 production and pricing policy of raw materials and separated rare earth elements were essentially stable. Chinese production and export of these products increased dramatically between 1988 and 1990 without a parallel increase in market demand, thus causing some destabilization of world pricing policies. Prices of rare earth materials during this period dropped by between 20 and 70 %, depending on the rare earth element. China's export revenues also suffered a 40 % decrease. Some stability has returned to the world market with China reducing output of rare earth minerals from 43 000 t (25 250 t REO) in 1989 to 27 500 t (16 500 t REO) in 1990, and to 16 000 t REO in 1992. Exports of raw materials also dropped by 50 % in 1990 and by 46 % in 1991.

World rare earth consumption grew by less than 3 % annually between 1980 and 1990 against a predicted growth of 6 %. In 1980 the demand for rare earths expressed as oxide (REO) was ca. 27 000 t. Rare earth consumption in 1990 is estimated at 35 000 t with a market value of ca. $\$400\times10^6$. Ores, concentrates and mixed rare earth compounds account for ca. 75 % of the tonnage consumption but only 25 % of the total market value.

Rare earths in the form of metals and separated high-purity oxide for use in applications such as phosphors, magnets and superconduc-

Table 15. Thermal neutron capture cross sections of some rare earth elements in n, γ reactions

Rare earth element	Number of absorbing isotopes	Relative atomic mass	Abundance in nature, wt %	Thermal neutron capture cross section, 10^{-28} m^2	
				For natural isotope mixture	For single isotopes
Sm	1	149	13.8	5 600	66 000
Eu	5	151	47.8	4 300	9 000
		152			5 000
		153	52.2		420
		154			1 500
		155			13 000
Gd	2	155	14.9	46 000	70 000
		157	15.7		180 000
Dy	5	160	2.3	950	130
		161	18.9		680
		162	25.5		240
		163	24.9		220
		164	28.2		2 780

tors, demand much higher added value and account for nearly 75 % of the market.

Estimates of world rare earth consumption by end use for 1990 and predicted values for 1996 are shown in Table 16. However, as is shown in Table 17, the balance of consumption for end use applications varies significantly between the major user countries. This can reflect differences in technological status, future commercial strategies, and the availability of rare earth stock.

New developments based on the less common rare earth elements or requiring only a single rare earth element can affect the supply/demand situation. A favorable cost/price situation will only exist if there is a good market for all the principal rare earth elements that occur in nature, since rates of production of the elements by the rare earth industry are interdependent.

Table 16. World markets for rare earth consumption by end use [192]

Use	1990		Predicted for 1996	
	% vol	% by value	% vol	% by value
Phosphors	3	37	4	40
Magnets	7	20	15	23
Catalyst	40	16	35	11
Glass	18	12	17	10
Metallurgy	26	4	22	2
Ceramics	2	6	2	8
Other	4	5	5	6
Total	100	100	100	100

Prices of rare earths can vary significantly depending upon the abundance of the rare earth element in the ore, the source of the product, the purity of the element or compound required, and the quantity of material purchased.

Approximate prices of rare earth oxides are given in Table 18. Reviews of the rare earth industry are given in [14, 192–194].

11. Toxicology

The pure rare earth metals and their compounds are regarded as having low toxicities with respect to the Hodge – Sterner classification system [195]. Haley describes the rare earth elements as only slightly toxic [2, vol. 4, pp. 553 – 585], [196]; see also [1, part B 2]. A further extensive review on this subject is provided by Arvela [197].

Several authors investigations into the inhalation of rare earth oxides by humans (smoke from carbon arc lamps in cinemas) have been carried out [198–200]. Chest X-rays show changes to the lungs caused by dust deposits, although these harmful effects are not caused by the rare earth elements, but by radioactive impurities, i.e., products of the radioactive decay of thorium and uranium. Modern methods of rare earth separation together with an increased use of bastnaesite have led to rare earth products that are virtually free from Th and U.

Experiments on animals showed that intravenous injection of rare earths caused impairment of liver function.

Table 17. Rare earth consumption by end use application [192]

Application	United States 1989		Japan 1990		China 1990	
	% vol	tonnes REO	% vol	tonnes REO	% vol	tonnes REO
Glass and ceramics	43	7 524	67	3 639		
Catalysis	42	7350			30	2200
Metallurgy	14	2 450	4	217	56.5	4100
Permanent magnets			18	978	3	216
Agriculture					6.5	480
Phosphors			5.5	299		
Other	1 ∗	175 ∗	5.5 ∗∗	299 ∗∗	4	260
Total	100	17 500	100	5432	100	7256

∗ Includes permanent magnets, phosphors, and electronics.
∗∗ Includes catalysis.

Table 18. Approximate prices of rare earth oxides in 1990 [192]

Rare earth oxide	Rhône-Poulenc		Molycorp	
	Purity, %	$/kg	Purity, %	$/kg
Cerium	99.50	25.75	99.0	19.25
Dysprosium	95.00	132.00	96.00	132.00
Erbium	96.00	190.00	98.00	143.00
Europium	99.99	1960.00	99.99	1639.00
Gadolinium	99.99	136.50	99.99	132.00
Holmium	99.90	510.00		
Lanthanum	99.99	23.00	99.99	19.25
Lutetium	99.99	7000.00		
Neodymium	95.00	20.00	99.90	88.00
Praseodymium	96.00	38.85	96.00	37.00
Samarium	96.00	175.00	96.00	143.00
Terbium	99.90	880.00	99.90	825.00
Thulium	99.90	3600.00		
Ytterbium	99.00	230.00		
Yttrium	99.99	115.50	99.99	115.50

12. References

1. *Gmelin,* 8th ed. Seltene Erden, **A2 – A5, B1 – B7, C1 – C3, C4 a, C5 – C6, C8, D1 – D3.**
2. K. A. Gschneidner, Jr., L. Eyring, (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* vol. 1 – 4, North Holland Publ., Amsterdam 1979.
3. L. Eyring (ed.): *Progress in the Science and Technology of the Rare Earths,* **vol. 1, 1964; vol. 2 1966; vol. 3, 1968,** Pergamon Press, Oxford.
4. J. A. Gibson, G. S. Harvey: "Properties of the Rare Earth Metals and Compounds," *Technical Report AFML-TR-65-430,* Wright Patterson Air Force Base, Ohio, 1966.
5. F. H. Spedding, A. H. Daane: *The Rare Earths,* J. Wiley, New York 1961.
6. F. Trombe, J. Loriers, P. Gaume-Mahn, Ch. H. La Blanchetais: "Scandium-Yttrium-Elements des Terres Rares-Actinium" in P. Pascal (ed.): *Nouveau Traite De Chimie Minerale,* vol. 7, parts 1, 2, Masson, Paris 1959.
7. International Union of Pure and Applied Chemistry (IUPAC): *Nomenclature of Inorganic Chemistry, Definition Rules 1970,* Butterworths, London 1971.
8. Th. Moeller: "The Lanthanides" in I. C. Bailar, H. J. Emeleus, R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorganic Chemistry,* vol. 4, Pergamon Press, Oxford 1971, p. 1.
9. J. W. Adams, E. R. Iberall: "Bibliography of the Geology and Mineralogy of the Rare Earths and Scandium to 1971," *Geological Bulletin 1366,* Washington 1973.
10. J. W. Adams: "Distribution of the Lanthanides in Minerals," *Proc. Rare Earth Res. Conf.,* **7 th** (1968) 28 – 30.
11. J. W. Adams: "Resources" in J. G. Parker, C. T. Baroch (eds.): *The Rare Earth Elements, Yttrium and Thorium, A Materials Survey,* chap. 3, USBM Information Circular 8476, Washington 1973, pp. 22 –39.
12. United States Bureau of Mines: Mineral Commodity Summaries, and Unocal Company Reports, Washington 1991.
13. G. Friedlander, J. W. Kennedy: *Lehrbuch der Kern- und Radiochemie,* Thiemig, München 1962.
14. J. A. Marinsky, L. E. Glendenin, C. D. Coryell, *J. Am. Chem. Soc.* **69** (1947) 2781.

15. O. Erämetsä, *Acta Polytech. Scand. Chem. Ind. Matall. Ser.* **37** (1965).
16. D. A. Johnson, *J. Chem. Educ.* **57** (1980) 475.
17. F. H. Spedding, B. J. Beaudry, *J. Less Common Met.* **25** (1971) 61.
18. K. N. R. Taylor, M. I. Darby: *Physics of Rare Earth Solids,* Chapmann and Hall, London 1972, pp. 1 – 324.
19. R. D. Shannon, C. T. Prewitt, *Acta Crystallogr. Sect. B* **B 25** (1969) 925.
20. T. Petzel, O. Greis, *Z. Anorg. Allg. Chem.* **396** (1973) 95.
21. K. A. Geschneidner, Jr., *J. Less-Common Met.* **17** (1969) 1.
22. J. H. Van Vleck: *The Theory of Electric and Magnetic Susceptibilities.* Oxford University Press, Oxford 1932, pp. 1 – 395.
23. K. N. R. Taylor, *Contemp. Phys.* **11** (1970) 423 – 454.
24. H. R. Kirchmary, *Z. Metallkd.,* **60** (1969) 699 – 708.
25. H. R. Kirchmayr, *Z. Metallkd.,* **60** (1969) 778 – 784.
26. W. E. Wallace: *Rare Earth Intermetallics,* Academic Press, New York 1973.
27. K. A. Gschneidner, Jr.: *Rare Earth Alloys,* Van Nostrand, Princeton 1961.
28. K. H. J. Buschow, *Rep. Prog. Phys.* **42** (1979) 1373 –1477.
29. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* 2nd ed., McGraw-Hill, New York 1958, 1st Supplement R. P. Elliot, 1965, 2nd Supplement T. A. Shunk, 1969.
30. W. G. Moffart: *The Handbook of Binary Phase Diagrams,* General Electric, Schenectady 1978.
31. K. A. Gschneidner, Jr., M. E. Verkade: *Selected Cerium Phase Diagrams, IS-RIC-7,* Rare Earth Information Centre, Iowa State University 1974.
32. Soc. Prod. Chim. Terres Rares, FR 995 112, 1949.
33. A. E. Bearse, G. D. Calkins, J. W. Glegg, R. B. Filbert, *Chem. Eng. Prog. Monogr. Ser.* **50** (1954) 235 –239.
34. M. Smutz, G. L. Bridger, K. G. Shaw, M. E. Whatley, *Chem. Eng. Prog. Symp. Ser.* **50** (1954) 13, 167 – 170.
35. American Potash and Chemical Corp., US 2 900 231, 1952.
36. V. E. Shaw, *Rep. Invest. U.S. Bur. Mines <CFPS TH 7 7>RI* **5474** (1959) 1 – 12.
37. Y. Hirashima, J. Shiokawa, *Nippon Kagaku Kaishi,* 1973 no. 3, 496 – 499; *Chem. Abstr.* **78** (1973) 152030.
38. *Chem. Eng. (N.Y.)* **74** (1967) 2, 122 – 123.
39. P. R. Kruesi, G. Duker, *J. Met.,* **17** (1965) 847 – 849.
40. D. J. Bauer, V. E. Shaw, *Rep. Invest. U.S. Bur. Mines.* **RI 6381** (1964) 1 – 15.
41. B. I. Kogan, I. P. Skripka, *Redk. Elem.* **8** (1973) 3 – 57.
42. H. Richter, A. Krause, *Chem. Tech. (Leipzig)* **17** (1965) 707 – 710.
43. G. Trombe, DE-OS 2 238 665, 1973.
44. W. Brugger, E. Greinacher, *J. Met.* **19** (1967) 32 – 35.
45. Th. Goldschmidt AG, DT 956 993, 1954.
46. Soc. Prod. Chim. Terres Rares, FR 1 209 251, 1958.
47. F. H. Spedding, J. E. Powell in F. C. Nachod, J. Schubert (eds.): *Ion Exchange Technology,* Academic Press, New York 1956, pp. 359 – 390.
48. C. D. Scott, R. D. Spence, W. G. Sisson: "Pressurised Annular Chromarography for Continuous Separation" *J. Chromatogr.* **126** (1976) 381 – 400.
49. J. M. Begovich, W. G. Sisson: "A Rotating Annular Chromatograph for Continuous Metals Separation and Recovery," *Resour. Converv.* **9** (1982) 219 – 229.
50. J. M. Begovich, C. H. Byers, W. G. Sisson: "A High Capacity Pressurised Continuous Chromatograph, " *Sep. Sci. Technol.* **18** (1983) 1167 – 1191.
51. V. T. Taniguchi, A. W. Doty, C. H. Byers: "Large Scale Chromatographic Separations using Continuous Displacement Chromatography (CDC)" in R. G. Bautista, M. M. Wong (eds.): *Rare Earths; Extraction, Preparation and Applications,* The Minerals, Metals and Materials Soc., annual meeting, Las Vegas 1989, 147 – 161.
52. J. P. Surls, Jr., G. R. Choppin, *J. Am. Chem. Soc.* **79** (1957) 855 – 859.
53. S. W. Mayer, E. R. Tompkins, *J. Am. Chem. Soc.* **69** (1947) 2866 – 2874.
54. J. E. Powell, F. H. Spedding, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **215** (1959) 457 – 463.
55. P. Conrard, M. Caude, R. Rosset, *Sep. Sci.* **7** (1972) no. 5, 457 – 463.
56. P. Conrard, M. Caude, R. Rosset, *Sep. Sci.* **9** (1974) no. 4, 269 – 285.
57. R. E. Linstrom, J. O. Wingert, *Rep. Invest. U.S. Bur. Mines* **RI 6131** (1961) 1 – 21.
58. D. B. James, J. E. Powell, F. H. Spedding, *J. Inorg. Nucl. Chem.* **19** (1961) 133 – 141.

59. W. L. Silvernail, M. M. Woyski: "Commercial Ion Exchange Separation of the Rare Erths," *Proc. Rare Earth Res. Conf.* **6th,** Gatlinburg, May 1967, 678–690.
60. Santoku Metals Industry Co., JP 70 28 292, 1966.
61. GTE Sylvania Inc., US 3 482 932, 1969.
62. C. Persiani, J. F. Cosgrove, R. Walters, D. J. Bracco, *Sep. Sci.* **2** (1967) no. 6, 789–796.
63. D. Asher et al., *Ind. Eng. Chem. Process Des. Dev.* **1** (1962) no. 1, 52–56.
64. J. E. Powell, H. R. Burkholder, *J. Chromatogr.* **36** (1968) 99–104.
65. J. E. Powell, H. R. Burkholder, *J. Chromatogr.* **29** (1967) 210–217.
66. W. Fischer, W. Dietz, O. Jubermann, *Naturwissenschaften* **25** (1937) 348.
67. W. Fischer, K. Biesenberger, J. Heppner, U. Neitzel, *Chem. Ing. Tech.* **36** (1964) 85–99.
68. F. Horn, *Chem. Ing. Tech.* **36** (1964) 99–111.
69. N. M. Rice, *Chem. Ind. (London)* 1977, Sept., 718–723.
70. C. G. Brown, L. G. Sherrington: "Solvent Extraction used in Industrial Separation of Rare Earths," *J. Chem. Technol. Biotechnol.* **29** (1979) 193–209.
71. R. F. Sebenik, B. M. Sharp, M. Smutz, *Sep. Sci.* **1** (1966) no. 4, 375–386.
72. E. Hesford, E. E. Jackson, H. A. C. McKay, *J. Inorg. Nucl. Chem.* **9** (1959) 279–289.
73. D. Scargill et al., *J. Inorg. Nucl. Chem.* **4** (1957) 304–314.
74. Th. Goldschmidt AG, DT 1 592 120, 1967.
75. Thorium LTD., GB 1 180 922, 1966.
76. G. W. Mason, D. N. Metta, D. F. Peppard, *J. Inorg. Nucl. Chem.* **38** (1976) 2077.
77. J. Alstad, J. H. Augustson, T. Danielssen, L. Farbu: "Comparative Study of the Rare Earth Elements in Extraction by HDEHP/Shell Sol T from Nitric and Sulphuric Acid Solutions" in G. V. Jeffreys (ed.): *Proc. Int. Solvent Extr. Conf.* **1974** 1083–1102.
78. T. B. Pierce, P. F. Peck, *Analyst (London)* **88** (1963) 217–221.
79. N. E. Thomas, S. Smutz, L. Burkhart, *Ind. Eng. Chem. Fundam.* **10** (1971) 453–458.
80. D. S. Flett, G. W. Cutting, P. Carey, *Proc. Int. Miner. Process. Congr.* **10th** (1973) Paper 23, (Pub. 1974) 1147–1167.
81. D. E. Peppard, G. W. Mason S. W. Moline in R. G. Bautista, M. M. Wong (eds.): *Rare Earths; Extraction, Preparation and Applications,* The Minerals, Metals and Materials Soc., Warrendale 1989.
82. D. E. Peppard, E. P. Horwitz, G. W. Mason in R. G. Bautista and M. M. Wong (eds.): *Rare Earths; Extraction, Preparation and Applications,* The Minerals, Metals and Materials Soc., Warrendale 1989.
83. K. S. Han, K. Tozawa: "Thermodynamic Prediction of Coefficients for the Solvent Extraction of the Rare Earths" in R. G. Bautista, M. M. Wong (eds.): *Rare Earths; Extraction, Preparation and Applications,* The Minerals, Metals and Materials Soc., Warrendale 1989, pp. 115–128.
84. C. F. Baes, Jr., *J. Inorg. Nucl. Chem.* **24** (1962) 707–720.
85. Z. Kolarik, S. Drazanova, V. Chotivka, *J. Inorg. Nucl. Chem.* **33** (1971) 1125–1133.
86. R. A. Alekperov, S. S. Geibatova, *Dokl. Chem. (Engl. Transl.)* **178** (1968) 29.
87. R. D. Korpusova, G. L. Vaks, E. N. Petrusheva, *Russ. J. Inorg. Chem. (Engl. Transl.)* **14** (1969) 1004.
88. F. J. Seelay, D. J. Crouse, *J. Chem. Eng. Data* **11** (1966) 42.
89. F. J. Seelay, D. J. Crouse, *J. Chem. Eng. Data* **16** (1971) 393.
90. US Atomic Energy Commission, US 3 294 494, 1966.
91. J. S. Preston, A. C. DuPreez, *Rep. Mintek* **M 378** (1988).
92. P. Th. Gerontopulis, L. Rigali, P. G. Barbano, *Radiochim. Acta* **4** (1965) 75–78.
93. Government of the United States, US 3 323 857, 1967.
94. US Atomic Energy Commission, US 3 409 415, 1968.
95. T. V. Healey, *J. Inorg. Nucl. Chem.* **19** (1961) 314–327.
96. B. H. Lucas, G. M. Ritcey, *CIM Bull.* **68** (1975) no. 753, 124–130.
97. B. Z. Zhang et al., *Hydrometallurgy* **9** (1982) 205–210.
98. K. A. Gschneidner, Jr, *Rare Earths in the Electronics Industry* **88** (1991) 63–94; Royal Soc. Chem., Special Publication, Fine Chemicals for the Electronics Industry II.
99. H. W. Harrah: *Deco Trefoil,* Denver Equipment Comp., Denver, Colorado, Nov–Dec, 9, 1967.
100. Forskningsgruppe for Sjeldna Jordarter, US 3 751 553, 1971.
101. Molybdenum Corp., US 3 640 678 1970.
102. H. N. McCoy, *J. Am. Chem. Soc.* **63** (1941) 1622.

103. E. Greinacher, K. Reinhardt: "Seltene Erden" in H. Harnisch, R. Steiner, K. Winnacker (eds.): *Chemische Technologie,* 4th ed., vol. 2, "Anorganische Technologie I" Hanser Verlag, München 1982, pp. 678 – 707.
104. E. Morrice, M. M. Wong, *Miner. Sci. Eng.* **11** (1979) 125 – 136.
105. Molybdenum Corp., US 3 729 397 1970.
106. T. A. Henrie, *J. Met.* **16** (1964) 978 – 981.
107. O. N. Carlsson, J. A. Haefling, F. A. Schmidt, F. H. Spedding, *J. Electrochem. Soc.* **107** (1960) 540 – 545.
108. T. T. Campbell, F. E. Block, *J. Met.* **11** (1959) 744 –746.
109. Th. Goldschmidt AG, DT 2 303 697, 1973.
110. R. E. Cech, *J. Met.* **26** (1974) 32.
111. C. E. Habermann, A. H. Daane, P. E. Palmer, *Trans. AIME,* **233** (1965) 1038 – 1042.
112. A. Brusdeylins, *Chem. Ztg.* **97** (1973) 343 – 347.
113. O. B. Michelsen (ed.): *Analysis and Application of Rare Earth Materials,* NATO Advanced Study Institute, Kjeller, Norway 23 – 29 Aug., 1972, Universitets-forleget, Oslo 1973.
114. K. A. Geschneidner, Jr., L. Erying (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* vol. 13, North-Holland Publ., Amsterdam 1991.
115. R. J. Conzemius, S.-K. Zhao, R. S. Houk, H. J. Svec, *Int. J. Mass Spectrom. Ion Processes* **61** (1984) 277.
116. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 6.**, chap. 47, North-Holland Publ., Amsterdam 1984, pp. 1 – 111.
117. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 5,** chap. 44, North-Holland Publ., Amsterdam 1982, pp. 321 – 386.
118. B. G. Hyde, D. J. M. Bevan, L. Eyring, *Phil. Trans. R. Soc. London Ser. A* **259** (1966) no. 1106, 583.
119. K. A. Gschneidner, Jr., L. Eyring (eds.). *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 8.**, chap. 56, part I, North-Holland Publ., Amsterdam 1986, 203 – 334.
120. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 9.**, chap. 59, part II, North-Holland Publ., Amsterdam 1987, pp. 320.
121. A. W. Sleight, C. T. Prewitt, *Inorg. Chem.* **7** (1968) 2282.
122. T. Petzel, *Inorg. Nucl. Chem. Lett.* **10** (1974) 119.
123. E. D. Eastmann et al., *J. Am. Chem. Soc.* **72** (1950) 2248.
124. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 13.**, chap. 89, North-Holland Publ., Amsterdam 1990, pp. 191 – 281.
125. G. V. Samsonov: *High Temperature Compounds of Rare Earths with Nonmetals,* Consultants Bureau, New York 1965.
126. Th. Moeller et al., *Chem. Rev.,* **65** (1965) 1.
127. R. E. Sievers, K. J. Eisentraut, C. S. Springer Jr.: "Volatile Rare Earth Chelates of β-Diketones" in R. F. Gould (ed.): *Adv. Chem. Ser.* **71** (1967) 141 –154.
128. H. Schumann, *Nachr. Chem. Tech. Lab.* **27** (1979) 393.
129. G. Wilkinson, J. M. Birmingham, *J. Am. Chem. Soc.* **76** (1954) 6210.
130. F. A. Hart, M. S. Saran, *Chem. Commun.* 1968 1614.
131. F. A. Hart, A. G. Massey, M. S. Saran, *J. Organomet. Chem.* **21** (1970) 147.
132. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths.* **vol. 7.**, chap. 53, North-Holland Publ., Amsterdam 1985, p. 446.
133. A. Raman, *Z. Metallkd.* **68** (1977) 163 – 172.
134. A. Raman, *Z. Metallkd.* **67** (1976) 780 – 789.
135. H. Zulstra, *Chem. Tech. Technol.* **2** (1972) 280 – 284.
136. K. H. J. Buschow, A. R. Miedema: "Hydrogen Absorption in Rare Earth Intermetallic Conmpounds" in A. F. Andresen, A. J. Maeland (eds.): *Hydrides Energy Storage Proc. Int. Sym.,* **1977,** (Pub. 1978) 235 – 249.
137. C. E. Lundin, R. W. Sullivan: "The Safety Characteristics of Lanthanum-Nickel (LaNi5) Hydrides" in T. N. Veziroglu (ed.): *Hydrogen Energy Proc. Hydrogen Econ. Miami Energy Conf. 1974,* Plenum Publishing, New York 1975, pp. 645 – 658.
138. I. Lambert, A. Percheron-Guegan, J. Montel, *J. Chim. Phys. Phys. Chim. Biol.* **74** (1977) no. 3, 380 –381.
139. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 11,** chap. 76, North-Holland Publ., Amsterdam 1988, pp. 293 – 321.
140. K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths,* **vol. 12,** chap. 83, North-Holland Publ., Amsterdam 1989, pp. 133 – 212.

141. *Goldschmidt Informiert,* **4 (35),** 1 – 90 (1975).
142. *Goldschmidt Informiert,* **2 (48),** 1 – 75 (1979).
143. *Proc. 7th Int. Workshop Rare Earth – Co Permanent Magnets and their Applications,* China Academic, Peking 1983.
144. Socony Mobil Oil Compl. Inc., US 3 173 854, 1961.
145. *Molycorp Application Report,* vol. 1, "A Bibliography. Six Years of Research in Catalysis with Rare Earth Elements," no. 7101, 1964 – 1970, vol. 2, "A Bibliography of the Research in Catalysis with the Rare Earth Elements," no. 7907, 1971 – 1976.
146. M. P. Rosynek, *Catal. Rev. Sci. Eng.* **16** (1977) 111 –154.
147. S. C. Sorenson, J. A. Wrongkiewicz, L. B. Sis, G. P. Wirtz, *Am. Ceram. Soc. Bull.* **53** (1974) 446 – 449.
148. V. T. Coon, W. E. Wallace, R. S. Craig: "Methanation by Rare Earth Intermetallic Catalysts," in J. G. McCarthy, J. J. Rhyne (eds.): "*Rare Earth Research Conf. 13th.*, 1977, Plenum, New York 1978.
149. W. R. Grace, US 3 903 020, 1975.
150. Engelhard, US 4 021 185, 1977.
151. Government of the United States, US 3 794 473, 1972.
152. Allis-Chalmers Manufacturing Comp., US 3 405 008, 1965.
153. K. Shirai, S. Ihara, H. Sato, *Denshi Gijutsu Sogo Kenkyusho Iho* **38** (1974) 378 – 382; *Chem Abstr.* **82** (1975) 173473 c.
154. G. Robert, M. Forestier, C. Deportes, *Journ. Int. Etude Piles Combust. C. R.* **3rd** (1969) 97 – 105.
155. Molycorp Inc.,: "Overview of Application Possibilities," no. 48.
156. P. R. Ryason, US 4 105 517, 1977.
157. J. Kiwi, M. Gratzel, *Angew. Chem.* **91** (1979) 659 –660.
158. B. Wyatt, *Ind. Res. Dev.* **21** (1979) June, 64 – 65.
159. General Refractories Co., US 3 974 108, 1974.
160. Tokai Rika Co., Ltd., JP 76 52 409, 1974.
161. P. H. Schmidt, D. C. Joy, *J. Vac. Sci. Technol.* **15** (1978) 1809 – 1810.
162. T. C. Shutt, M. Martin, C. J. Lewis, J. Drobnick, *J. Can. Ceram. Soc.* **40** (1971) 29 – 37.
163. T. C. Shutt, G. Barlow, *Am. Ceram. Soc. Bull.* **51** (1972) 155 – 157.
164. Harshaw Chem. Co., US 2 992 123, 1959.
165. A. Nedeljkovic, R. L. Cook, *Vitreous Enameller* **26** (1975) 2 – 13.
166. W. Mialki, *Metall (Berlin)* **19** (1965) 16 – 21.
167. Dentists' Supply Comp., US 2 895 050, 1956.
168. National Research Development Corp., GB 1 529 984, 1976.
169. Zahnfabrik Bad Nauheim KG, DT 1 813 163, 1968.
170. L. Tcheichvili, M. N. P. DeMaeques, *Keram. Z.* **30** (1978) no. 3, 138 – 139; **30** (1978) no. 5, 252 – 254.
171. E. Hamann, H. Manger, L. Steinke, *Automot. Appl. Sens. SAE Spec. Publ.* **SP-418** (1977) 53 – 80.
172. D. Janke, W. A. Fischer, *Arch. Huttenw.* **48** (1977) 255 – 260.
173. Molycorp Inc.: "Overview of Application Possibilities," no. 40.
174. I. C. Huseby, G. Petzow, *Powder Metall. Int.* **6** (1974) 17 – 19.
175. J. R. Maldonado: *IEEE Int. Conv. Dig.* 1972, 190 –191.
176. N. V. Philips' Gloeilampenfabrieken, DE-OS 2 037 197, 1969.
177. Matsushita Electric Industrial Co., Ltd., DE-OS 2 055 197, 1969.
178. J. B. Chesney, *J. Am. Ceram. Soc.* **46** (1963) 197 – 202.
179. Matsushita Electric Industrial Co., Ltd., US 3 586 642, 1968.
180. Molycorp Inc.: "Overview of Application Possibilities," no. 21.
181. Hitachi Ltd., JP 74 129 199, 1973.
182. Molycorp Inc.: "Overview of Application Possibilities," no. 54.
183. G. M. Schwab, F. Bohla, *Z. Naturforsch.* **32** (1968) no. 10, 1555 – 1558.
184. Molycorp Inc.: "Overview of Application Possibilities," no. 28.
185. R. E. Harrington, *Environ.* **13** (1967) March, 32 – 33; **13** (1967) June, 32 – 34; **13** (1967) Sept. 30 – 32.
186. Molycorp Inc.: "Overview of Application Possibilities," no. 38.
187. Molycorp Inc.: "Overview of Application Possibilities," no. 46.
188. A. B. Stiles, US 3 266 477, 1966.
189. Fedders Corp., US 3 627 560, 1970.
190. B. C. Mayo, *Chem. Soc. Rev.* **2** (1973) 49.
191. North American Rockwell Corp., DE-OS 1 926 969, 1968.
192. Roskill Information Services: *The Economics of Rare Earths and Yttrium,* 8th ed., London 1991.
193. B. Campana, *Bull. Ver. Schweiz. Pet. Geol. Ing.* **43** (1977) no. 104, 1 – 28.
194. *Inf. Chim.* **190** (1979) 67 – 70.

195. H. C. Hodge, H. S. Sterner, *Am. Ind. Hyg. Assoc. Q.* **10** (1943) 93.
196. Th. A. Haley, *J. Pharm. Sci.* **54** (1965) no. 5, 663 – 670.
197. P. Arvela in H. Grobecker et al. (eds.): *Progress in Pharmacology,* Publ. G. Fischer Verlag, Stuttgart 1979, 69 – 114.
198. R. Hoschek: "Die 'Cer-pneumokoniose' nach Einatmen von natürlichen Seltenen Erden," *Schriftenreihe Arbeitsmedizin-Sozialmedizin-Arbeitshygiene,* vol. 18, Genter, Stuttgart 1968.
199. M. Menz, E. Kaufmann, *Z. Unfallmed. Berufskr.* **65** (1972) no. 1, 62 – 73.
200. J. Sykora, F. Huzl, A. Kubat, M. Vykrrocil, *Zentralbl. Arbeitsmed. Arbeitsschutz,* **16** (1966) no. 1, 3 – 9.

Rare Gases → Noble Gases
Rat Control → Rodenticides

Reactive Dyes

Horst Tappe, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Walter Helmling, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Peter Mischke, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Karl Rebsamen, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Uwe Reiher, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Werner Russ, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Ludwig Schläfer, Hoechst AG, Frankfurt/Main, Federal Republic of Germany
Petra Vermehren, Hoechst AG, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Reactive Systems 2
2.1. Mono-Anchor Dyes 2
2.2. Double-Anchor Dyes 4
2.3. Multiple-Anchor Dyes 5
3. Chromogens 5
3.1. Azo Dyes 5
3.2. Anthraquinone Dyes 6
3.3. Triphenodioxazine Dyes 7
3.4. Formazan Dyes 12
3.5. Phthalocyanine Dyes 12
4. Synthetic Procedures 13
5. Uses 14
5.1. Forms of Supply 14
5.2. Dyeing Processes 15
6. References 15

1. Introduction

The most important distinguishing characteristic of reactive dyes is that they form covalent bonds with the substrate that is to be colored during the application process. Thus, the dye molecule contains specific functional groups that can undergo addition or substitution reactions with the $-OH$, $-SH$, and $-NH_2$ groups present in textile fibers.

Cross and Bevan first succeeded in fixing dyes covalently onto cellulose fibers (in 1895) [1], but their multistep process was too complicated for practical application. Early work by Schröter with sulfonyl chloride-based dyes was unsuccessful [2], but Günther later did succeed in fixing derivatives of isatoic anhydride onto cellulose fibers [3].

An alternative approach is to modify the fibers themselves and then introduce the coloration. For example, cellulose can be reacted with 4-nitrobenzyldimethylphenylammonium hydroxide or 3-nitrobenzyloxymethylpyridinium chloride, followed by reduction, diazotization, and coupling to a dye [4]. Another process, invented by Haller and Heckendorn, involves treating the cellulose with cyanuric chloride in an organic solvent and then reacting the product with an amino group-containing dye [5].

The first industrially important reactive dye systems were developed for wool, and took advantage of the chloroacetylamino [6] and the chloroethanesulfonyl groups [7]. Vinylsulfonyl- and 2-sulfooxyethanesulfonyl groups were found to be applicable to both wool and cellulose. Heyna and Schumacher patented some of the first dyes of this type in the 1940s [8, 9], and vinylsulfone dyes continue to be of great importance.

The dyes described in [8] entered the market for wool in 1952 under the trade name Remalan, followed a year later by the Cibalan dyes from Ciba-Geigy, which contained monochlorotriazinyl reactive groups [10]. In 1953 Rattee and Stephen were able to fix chlorotriazine-containing dyes onto cellulose in basic, aqueous solvents, representing an economic breakthrough for the reactive dye concept [11].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_651

Introduction of the Procion-MX dyes by ICI in 1956 was followed in 1957 by the appearance of the Remazol dyes from Hoechst and the Cibacron dyes from Ciba-Geigy. The year 1960 brought the Levafix dyes from Bayer and the Drimaren range from Sandoz. Other significant developments included the introduction of multiple-anchor dyes (Solidazol, Cassella, 1975; Sumifix-Supra, Sumitomo, 1981), fluorotriazine-containing reactive dyes (Ciba-Geigy, 1978), and fluorotriazine-containing multiple-anchor dyes (Ciba-Geigy, 1988).

The development of reactive dyes has continued to be rapid, and world demand for such materials reached ca. 97 000 t in 1990, valued at ca. € 2×10^9 . Reactive dyes thus constitute the largest of the textile dye classes on a monetary basis, accounting for ca. 17 % of the textile dye market by volume.

2. Reactive Systems

Reactive dyes are often classified according to the number of reactive systems they contain, leading to the terms mono-, double-, and multiple-anchor dyes. On the world market, the first two categories are of greatest importance; multiple-anchor dyes (those with three or more reactive groups) play a more minor role and will be discussed here only briefly (Section 2.3).

2.1. Mono-Anchor Dyes

Most mono-anchor dyes are derivatives of cyanuric chloride (2,4,6-trichloro-1,3,5-triazine [*108-77-0*]), a molecule of wide synthetic potential because the three chlorine atoms on the triazine ring differ in their reactivities [12]. The first chlorine atom exchanges with nucleophiles in water at 0 – 5 °C, the second at 35 – 40 °C, and the third at 80 – 85 °C. A wide variety of triazinyl dyes can thus be prepared by careful selection of the reaction conditions. Condensation of cyanuric chloride with a chromophore ("Chrom.") containing an amino group yields the highly reactive dichlorotriazinyl dyes (e.g., Procion MX) [11]. These very reactive dyes are sensitive to hydrolysis, and a suitable buffer is usually added to the powdered dye to increase its stability [13].

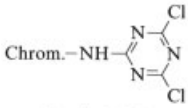

Procion MX

When two of the chlorines are substituted, for example with amino or alkoxy groups, monochlorotriazinyl dyes such as Procion H, Procion P, and Cibacron are obtained. These are considerably less reactive, and thus react with cellulose in the exhaust dyeing process only at relatively high temperature (80 °C). Such dyes are especially advantageous for printing [14].

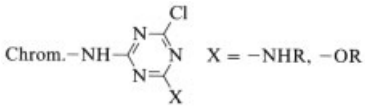

The reactivity of monochlorotriazinyl dyes can be increased by replacing the chlorine with fluorine, permitting the exhaust process to be carried out at 40 °C [15].

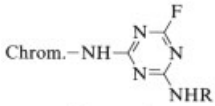

Cibacron F

Reaction of monochlorotriazinyl dyes with tertiary amines also yields highly reactive dyes, such as Kayacelon React and Procion H-EG. One advantage conferred by the ammonium groups in these dyes is high water solubility [16].

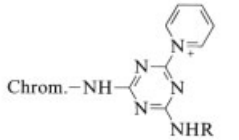

Kayacelon React/Procion H-EG

The reactivity of dyes of this type is strongly influenced by the choice of tertiary amine. Nicotinic acid is preferred because it acts as a good leaving group and introduces comparatively little odor into the dyeing bath. The ammonium residue is also a better leaving group than, for example, chlorine, permitting the exhaust process to be conducted at 40 – 60 °C. In contrast

to halogen-containing triazinyl dyes, these dyes do not require alkali in the application process, so dyeing can be conducted under neutral conditions.

In addition to the 1,3,5-triazines, other classes of heterocycles are important for the dyeing of cotton. Especially noteworthy are the halopyrimidine dyes distributed under the trade name Drimaren [17].

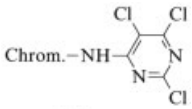

Drimaren

These dyes are less reactive than the triazines because the extra carbon atom reduces the ability of the ring to stabilize a negative charge. The reactivity of the system can be increased by introducing strongly electron withdrawing groups, including the cyano (Reacton) [18], fluoro (Levafix EA) [19], or methylsulfonyl groups (Levafix P) [20]. The pyrimidine ring can also be activated by inserting a carbonyl group between the chromophore and the heterocycle

(Reactophil) [21].

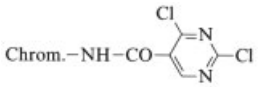

Reactophil

Other diazine moieties are often incorporated into reactive dyes as well, examples being found in the Primazine-P (BASF) [22] and Solidazol (Cassella) dyes [23].

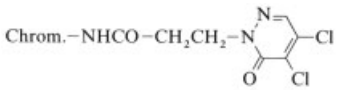

Primazin P

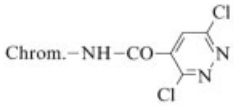

Solidazol

The list of economically attractive reactive heterocyclic intermediates also includes dichloroquinoxalines

(Levafix E) and dichlorophthalazines (Elisian). Anchor systems in both cases are bound to the chromophore via a carbonylamino group, and the reactivities correspond roughly to dichlorotriazinyl dyes.

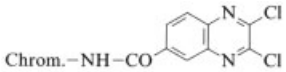

Levafix E

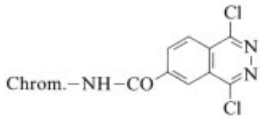

Elisian

Benzothiazole derivatives with good leaving groups at the 2-position of the heterocycle can also be invoked as reactive anchors. An example with chlorine as the leaving group is the Reatex system:

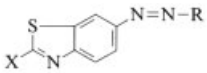

Reatex (X = Cl)

The reactive group that has had the greatest impact on the market is the 2-sulfooxyethylsulfonyl group (Remazol) [9,24,25]. Treatment with alkali in this case causes the elimination of sulfuric acid to form a vinylsulfonyl moiety that reacts with cotton to give a dye – fiber bond [26]. Describing this as an elimination – addition sequence is not meant to rule out the possibility that the nucleophile attacks the β-carbon directly, without intervention of a vinyl intermediate [27]. Reactive vinylsulfones are also prepared from 2-chloroethylsulfonyl derivatives, which lead to the desired intermediates by elimination of hydrogen chloride.

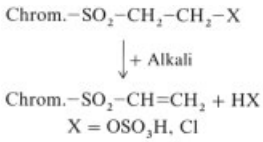

Numerous derivatives of ethylsulfonyl and vinylsulfonyl groups have been prepared,

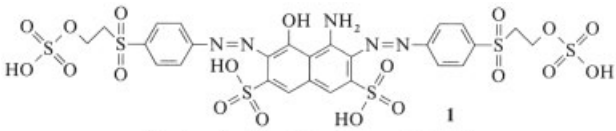

C.I. Reactive Black 5 – Remazol Black B

though none have approached the economic importance of the sulfate esters [28, 29].

Dyes containing phosphoric acid groups constitute a class of reactive dyes especially well-suited for polyester – cotton blends (Procion T) [30].

Chrom. – PO_3H

Esterification of the phosphoric acid group with cellulose leads to dye – fiber bonds that are very stable to both acids and bases. However, the required high fixation temperature (210 – 220 °C) in combination with acidic dyeing conditions considerably diminishes the stability of the cotton fibers. (Production of dyes in this category was suspended in 1987 [31]).

2.2. Double-Anchor Dyes

Since the mid-1980s considerable interest has been shown in highly fixing reactive dyes, and the requisite high fixation values have increasingly been achieved with the aid of double anchors. Double-anchor dyes can be divided into two categories: those containing two equivalent reactive groups, and those with mixed-anchor systems.

The first group includes the Procion HE hot dyes, in which two monochlorotriazine units are connected by a suitable bridge (e.g., a diamine). The synthetic approach to these compounds makes it possible to combine two different chromophores in a single system, opening the way to certain color shades that are not easily accessible via a single chromophore [32].

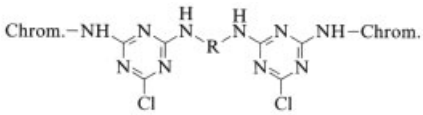

Procion HE

The category of reactive dyes with two equivalent anchors also encompasses products containing two vinylsulfonyl or 2-sulfooxyethylsulfonyl groups. An example is Remazol Black B (Hoechst, **1**), which not only displays a high degree of fixation but is also accessible by a synthetic route that is both simple and economical (see above).

Selective modification of the reactive anchors makes it possible to vary the physical property profile of the dye.

Mixed-anchor systems in reactive dyes were first described by ICI in 1959 [31]. Such "bifunctional" dyes were first marketed by Sumitomo (Sumifix Supra) and Hoechst (Remazol) in the early 1980s. These products are characterized by two anchors with differing reactivities: a more reactive 2-sulfooxyethylsulfonyl group and a less reactive monochlorotriazinyl residue. An example of such a system is Sumifix Supra Yellow 3RF.

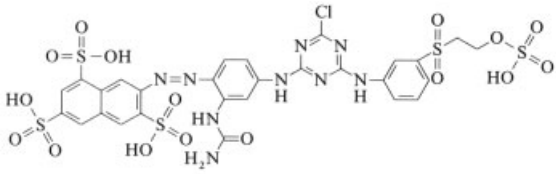

Sumifix Supra Yellow 3RF

Warm- and cold-dyeing double-anchor dyes are prepared by incorporating halotriazinyl and vinylsulfonyl reactive anchors. The bond between the triazine ring and a fiber is stable under basic conditions, whereas that to the vinylsulfonyl group is stable to acid. A combination of the two anchor systems therefore produces a dye with good fastness over a wide pH range. Relative to dyes based on dichlorotriazinyl, dichloroquinoxalinyl, or difluorochloropyrimidyl groups, mixed double-anchor dyes show greater acid fastness and more satisfactory behavior with respect to laundry water containing peroxide detergents. The Cibacron C dyes, introduced in 1988, include products in which a fluorotriazine unit bridges a vinylsulfonyl anchor and a chromophore [33]. Here both reactive systems exhibit similar reactivities.

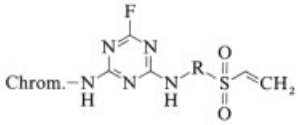

Cibacron C

Other mixed reactive systems have also been patented in addition to the double-anchor dyes so far described, including combinations of monofluorotriazine with difluorochloropyrimidine [34], alkoxychlorotriazine [35], or the 2-(sulfothio)ethylsulfonyl groups [36]. Other examples incorporate a difluorochloropyrimidine ring in conjunction with a vinylsulfonyl group [37].

Despite the growing importance of mixed reactive systems, only the monochlorotriazinyl/vinylsulfonyl and monofluorotriazinyl/vinylsulfonyl dyes described above have become commercially established.

2.3. Multiple-Anchor Dyes

The introduction of more than two anchor groups has only a minor influence on dye fixation characteristics, so multiple-anchor dyes play only a subordinate role in the world market. One example is a multiple-anchor system developed by Cassella, in which cyanuric chloride is bound to an amine with two aliphatic 2-chloroethylsulfonyl chains [38].

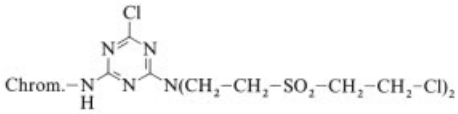

3. Chromogens

Virtually every conceivable chromophore has been used in the synthesis of reactive dyes, including monoazo and disazo species (→Azo Dyes), metal complexes of azo dyes (→Metal-Complex Dyes), formazans (→Metal-Complex Dyes, Chap. 6.), anthraquinones (→Anthraquinone Dyes and Intermediates) triphenodioxazines, and phthalocyanines (→Phthalocyanines). The product lines offered by the major dye producers in most cases feature comparable chromophores, differing primarily in the nature of the reactive systems and the particular substitution patterns adopted.

3.1. Azo Dyes [12, 39]

Most reactive dyes fall in the category of azo dyes. Virtually every hue in the dye spectrum can be achieved by appropriate structural modifications (mono- and disazo dyes, combinations involving either single or multiple aromatic and heterocyclic ring systems).

Monoazo dyes with heterocyclic coupling components, such as pyrazolone or pyridone, are yellow to greenish yellow. Other suitable coupling components include aniline and naphthylamine derivatives. Substituted anilines or aminonaphthalenesulfonic acids are employed as diazo components.

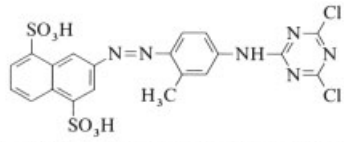

C.I. Reactive Yellow 4 – Procion Yellow MX-R

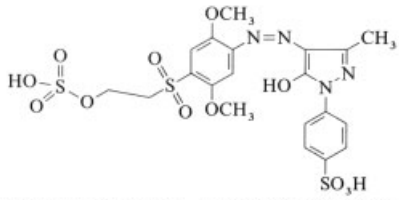

C.I. Reactive Yellow 17 – Remazol Golden Yellow G

A particularly important class of coupling components is the aminohydroxynaphthalenesulfonic acids. Appropriate variation of the disazo component permits the development of shades ranging from orange to black. Orange and scarlet are achieved with I-acid (6-amino-1-hydroxynaphthalene-3-sulfonic acid) and γ-acid (7-amino-1-hydroxynaphthalene-3-sulfonic acid), whereas H-acid (8-amino-1-hydroxynaphtha-lene-3,6-disulfonic acid) and K-acid (8-amino-1-hydroxynaphthalene-3,5-disulfonic acid) derivatives are useful for red to bluish-red hues. Extremely lightfast red shades

are also accessible through disazo dyes ("brown dyes").

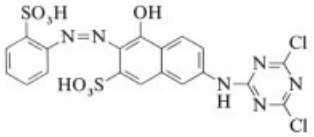

C.I. Reactive Orange 1 – Procion Brilliant Orange MX-GS

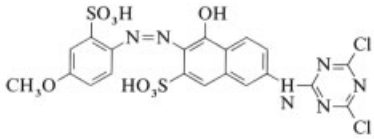

C.I. Reactive Red 8 – Procion Scarlet MX-G

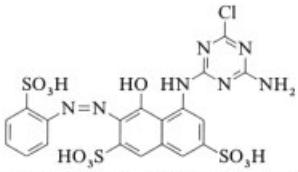

C.I. Reactive Red 12 – Cibacron Brilliant Red B

When H-acid is used as a double coupling component (as in **1**, where reactive groups are present in both diazo functions) the resulting dye color ranges from blue to black.

Exceptionally lightfast colors are obtained with metal-complexed azo dyes. Copper complexes of *o,o′*-disubstituted azo compounds produce a wide range of colors (yellow, ruby, violet, blue, brown, olive, black).

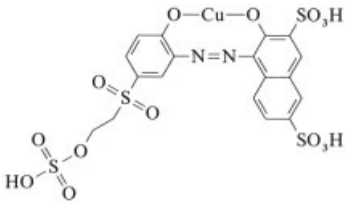

C.I. Reactive Red 23 – Remazol Red 3B

3.2. Anthraquinone Dyes [40]

Anthraquinone-based dyes are significant because of their brilliance, good lightfastness, and chromophore stability under both acidic and basic conditions. Until quite recently they dominated the market for brilliant blue reactive dyes in spite of their relatively low color strength (molar extinction=15 000) and comparatively high cost. The shades of commercial reactive anthraquinone dyes range from violet to blue. Almost all the important dyes are descendants of "bromaminic acid". This is treated with a reactive amine component, wherein the reactive group may be linked to the amino group by way of various aliphatic or, more often, aromatic bridging units.

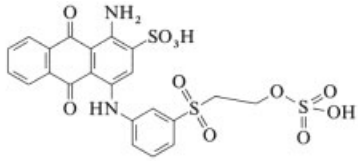

C.I. Reactive Blue 19 – Remazol Brilliant Blue R

3.3. Triphenodioxazine Dyes

Dyes derived from the triphenodioxazine ring system (**2**) have been commercially available since 1928 when KRÄNZLEIN and coworkers (Farbwerke Hoechst) discovered dyes with this basic structure augmented by sulfonic acid groups. The unsubstituted triphenodioxazine (which is of no importance as a colorant) was first obtained by G. FISCHER from the oxidation of 2-aminophenol in 1879 [41], and its structure was elucidated in 1890 [42].

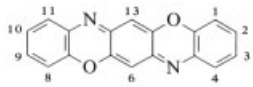

2

The first commercial products of this type were sold as blue direct dyes with unmatched tinctorial strength, brilliance, and high lightfastness. An example is C.I. Direct Blue 106 (**3**), which is still on the market today.

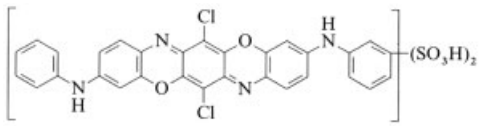

3

The commercial success of these products stimulated considerable research activity, especially by the I.G. Farben AG in Frankfurt/Main. This led to a large number of patent applications and several new direct dyes. After World War II research attention shifted toward disperse dyes, cationic dyes, and pigments. In 1952 appropriate treatment of a water-insoluble dichloro derivative led to C.I. Pigment Violet 23 (**4**) [43], whose structure was only later clarified by X-ray structural analysis [44].

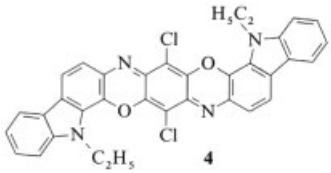

4

The triphenodioxazine chromophore was tested in almost every class of dyes, but its only major commercial application was as a direct dye and pigment. Only recently has it been introduced into reactive dyes. For the synthesis and properties of dioxazine dyes, see → Azine Dyes. Dioxazines were investigated in some of the earliest work on reactive dyes in the late 1950s, and they are included in some of the first patents covering reactive anchors and reactive dyes [45]. In particular, compound **5** was invoked as a chromophore, with reactive anchors attached to the free amino groups of the molecule.

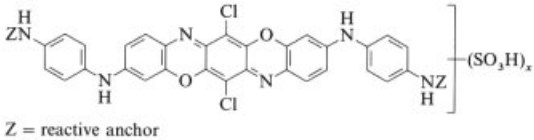

Z = reactive anchor

5

The first mixed azo–triphenodioxazine dye containing a monochloro-*s*-triazine group emerged in 1953 [46], with the triazine residue serving not as a reactive anchor but as a link between the two chromophores. In the years that followed, the dioxazine chromophore was also mentioned in most of the patents related to new reactive systems, although none led to a commercial product. At the beginning of the 1970s, the major activity seemed to be concentrated at ICI, leading to several exclusive patents for triphenodioxazine dyes [47]. Within a few years, the first commercial triphenodioxazine reactive dye was introduced: Procion Blue MX-G (C.I. Reactive Blue 163). Other products followed, all designed to achieve brilliant blue shades that could compete with the then-dominant anthraquinone chromophores in certain market segments [48].

The launching of these new products initiated research by other dye manufacturers. Between 1970 and 1979 only about 20 patents appeared on reactive triphenodioxazines and new reactive systems for dioxazine, compared to over 130 patents published in the period 1980–1988. The latter years saw the following new products brought onto the market: Procion Blue T-G, Procion Blue H-EGN, Cibacron Blue F-GF, Levafix Royal Blue E-FR, and 1989 Remazol-Brilliant Blue FB.

Table 1 provides information on a number of commercial triphenodioxazine products and their anchor systems.

By varying the substituents on the orange parent substance **6**, particularly in the positions *para* to the imino groups, its color can be modified. Red to red-violet shades result when X=OH or OR, and blue colors when X=NH_2 or NHR′. A blue shade also results if the external phenyl groups of the dioxazine system are part of an annulated and highly condensed aromatic ring system, for example, pyrene (C.I. Direct Blue 109). For details concerning the influence of chemical structure on color, particularly of dioxazine pigments, see [49].

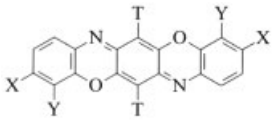

6

At present, only blue dioxazines are used commercially to color cellulose, where R′ is varied to give reddish to greenish-blue tints. Variation of T has only a minor effect on the shade. Table 2 summarizes observed absorption maxima as a function of various substituents. In particular, blue triphenodioxazines have very high molar extinctions ε, comparable to those of bisazo and phthalocyanine dyes. Until recently, anthraquinone dyes ($\varepsilon = 15\,000$) were predominent in most applications requiring brilliant blue dyes,

Table 1. Examples of triphenodioxazine dyes

Commercial name	C.I. Reactive Identification	Anchor	Year of introduction
Procion Blue MX-G	Blue 163	dichlorotriazine	1975
Procion Blue T-G	Blue 172	phosphonic acid	1978
Procion H-EG	Blue 187	triazylammonium	1979
Procion Blue H-EGN	Blue 198	monochlorotriazine	1982
Cibacron Blue F-GF	Blue 204	monofluorotriazine	1983
Levafix Royal Blue E-FR	Blue 224	sulfatoethylsulfone	1986
Remazol Brilliant Blue FB		sulfatoethylsulfone	1989

but the much stronger triphenodioxazine dyes now represent a less expensive alternative in many applications.

Dioxazine chromophores show a sensitivity to acid and base that is more pronounced in dye solutions than in the fixed dyes. Electron-withdrawing substituents enhance this effect, a major reason why red reactive dyes based on triphenodioxazine have not yet been commercialized [50] even though they are described in most dioxazine patents. Much research and development effort has succeeded in suppressing this sensitivity in the blue dye products currently on the market.

Generally, triphenodioxazine chromophores show a high degree of agglomeration and thus increased substantivity because of their planar arrangement. A higher substantivity causes problems during dyeing, such as unsatisfactory leveling in the exhaust process, tailing from dark to light in the padding process, or limited removal of unfixed dye in the rinsing process. Studies have been directed toward overcoming these problems either with new products displaying less critical substantivity or by control of the substantivity through more suitable dyeing conditions.

The synthetic routes (see below) to almost all dioxazines yield products with a symmetrical structure, resulting in at least two reactive anchors. Indeed, the two dyes most recently introduced (Levafix Royal Blue EFR and Remazol Brilliant Blue FB) are highly reactive double-anchor dyes suitable for both the exhaust and the padding processes. Ecological limits for wastewater are satisfied as a result of a high degree of fixation and low salt requirements in the exhaust process, where the latter also has a positive influence on leveling. These factors and the excellent fastness properties of the dioxazine dyes account for the fact that such dyes are expected to overcome the dominance of anthraquinone-based materials in some segments of the brilliant blue market.

The most important industrial process for synthesizing triphenodioxazines for cellulose fibers involves reaction of two equivalents of an aromatic amine with 2,3,5,6-tetrachloro-1,4-benzoquinone [*118-75-2*] (*p*-chloranil) and subsequent oxidative closure of the ring. All commercially available dioxazines thus contain chlorine atoms at the 6- and 13-positions. The process was discovered in 1911 at Hoechst [51], and its simplicity, coupled with excellent availability of raw materials, was the basis for the commercial success of early dioxazines. In the first reaction step the starting materials are condensed in the presence of an acid-binding agent such as soda ash or magnesium oxide to form the 2,5-diarylamino-3,6-dichloro-1,4-benzoquinone (**7**). Ring closure is then induced by heating either with a metal halide or an acid halide catalyst in a high-boiling solvent (e.g., nitrobenzene) or, in the case of sulfonated dyes, by heating with an oxidant in concentrated sulfuric acid.

In general, the cyclization is a relatively sensitive reaction, and very careful choice of starting materials and reaction parameters is necessary to obtain a product with the desired brilliance and tinctorial strength. The use of oleum in the second step is a significant improvement, because oleum serves not only as a solvent but also as an oxidant, permitting cyclization to proceed at or below room temperature.

Ring closure continued to prove difficult with starting materials containing sulfonic acid groups, which are of great importance in reactive dyes, but this problem was overcome by resorting to a mixture of oleum and persulfate compounds [52]. Later patents describe the use

of halogens [53], manganese dioxide [54], and aqueous hydrogen peroxide [55] as alternatives to persulfate.

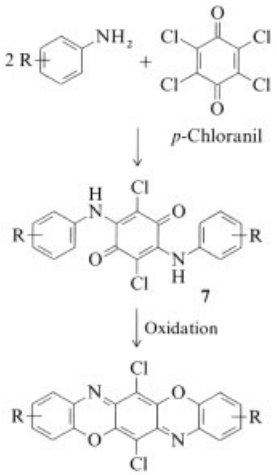

The triphenodioxazine chromophore can be converted to a reactive dye by incorporating anchors that are attached to the dioxazine core either directly or via bridging groups. The first patents, like the first commercial products, were based almost exclusively on the use of bridging groups. The usual starting material is 2-chloro-5-nitrobenzenesulfonic acid (**8**), which is reacted with a diamine (**9**).

Table 2. Absorption maxima of triphenodioxazines (6) as a function of substituent groups

T	Y	X	λ_{max}	ε
Cl	SO_2-alkyl	O-alkyl	530 nm	45 000
Cl	SO_2-alkyl	NH-allyl	570 nm	80 000
Cl	SO_2-alkyl	NH-alkyl	612 nm	80 000
Br	SO_2-alkyl	NH-alkyl	613 nm	80 000
Cl	SO_2 —NH-alkyl	NH-alkyl	620 nm	60 000
Cl	SO_2-alkyl	NH-aryl	623 nm	75 000

Reduction produces the *p*-phenylenediamine derivative (**10**), which is condensed with chloranil, and the resulting diarylamino-3,6-dichloro-1,4-benzoquinone(**11**) is isolated, dried, and then cyclized in oleum containing persulfate to yield the desired dye base. The base is converted to a reactive dye by reacting the two primary amino groups with two equivalents of an acylating agent **12**. Typical diamine bridges include 1,4-phenylenediamine-2-sulfonic acid, ethylenediamine, and 1,3-diaminopropane. The reactive anchors are almost always derivatives of *s*-triazines [56–63].

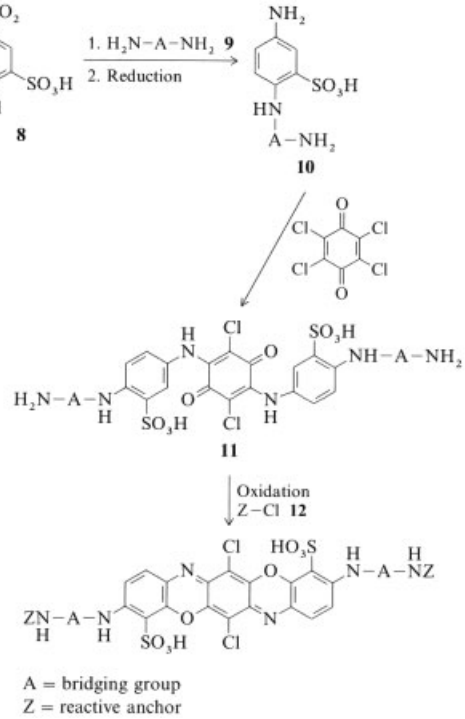

Because of their symmetry, all such triphenodioxazine reactive dyes feature double or even four-fold anchor systems. However, dyes with only one anchor group can be prepared by adding a single equivalent of acylating agent. Most asymmetric products of this type are based on 1,4-phenylenediamine-2-sulfonic acid [64] and characterized by somewhat reddish-blue hues [e.g., **13**, Z=monochloromonomethoxy-*s*-triazine, λ (max.)=560 nm].

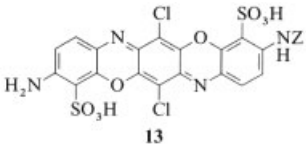

Red dioxazines can be synthesized by substituting an amino alcohol for the diamine derivative [65].

Replacing the sulfonic acid **8** with the appropriate carboxy-, carboxamido-, sulfonyl-, or

sulfonamido-4-nitrochlorobenzene leads to analogues described in [66, 67].

Another alternative is attaching the reactive anchor to the dioxazine ring via sulfonamide or carboxamide groups instead of via amine groups, as in **14** [67, 68].

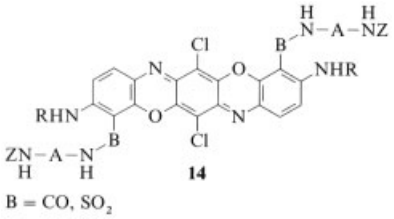

Compound **8** or one of the above-mentioned derivatives can also be reacted with a 2-hydroxyethylsulfonylaniline instead of with a diamine. After reduction of the 4-nitroaniline product **15** followed by condensation with chloranil, the cyclization takes place in oleum where, besides ring closure the 2-hydroxyethylsulfonyl groups are also esterified to β-sulfooxyethylsulfonyl groups [vinylsulfone (VS) groups]. Thus, the reactive anchor is formed concurrently during the cyclization reaction [69, 70].

Other similar fiber-reactive triphenodioxazines are prepared by replacing the sulfonic acid group in **8** with a 2-hydroxyethylsulfonyl group. The starting material here is 4-chloro-3-(2-hydroxyethylsulfonyl) nitrobenzene (**16**), which leads to products containing a VS group bound directly to the dioxazine ring [69, 71] (see top of the next page).

In an analogous reaction, use of a diamine affords dyes with both VS and amino groups. Subsequent acylation of the free amino groups produces the four-anchor system **17** [72].

An entirely different route to reactive dioxazines resembles the synthesis of reactive dyes based on copper phthalocyanine (cf. Sec. 3.5.) [73]. Thus, C.I. Pigment Violet 23 or C.I. Direct Blue 109 is first sulfochlorinated or aminomethylated and the product is then converted into the corresponding reactive dye. However, no compounds based on this approach have yet been commercialized.

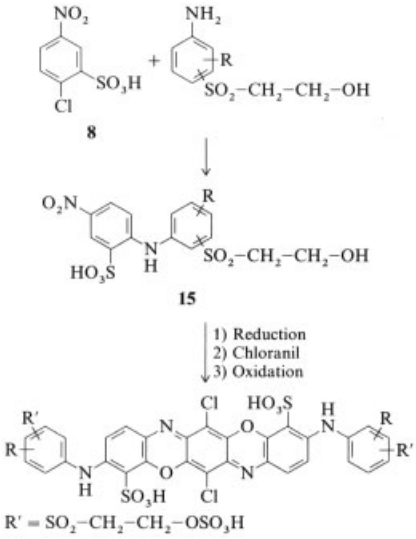

Because of the symmetrical nature of the chloranil condensation product **7**, all the syntheses described so far afford exclusively symmetrical triphenodioxazines; i.e., the two phenyl rings fused onto the dioxazine skeleton bear the same substituents. An elegant synthesis for an asymmetric chloranil condensation product has not yet been developed. One possible approach has been described starting with 2-methoxy-3,5,6-trichloro-1,4-benzoquinone [74], but its feasibility is uncertain [75].

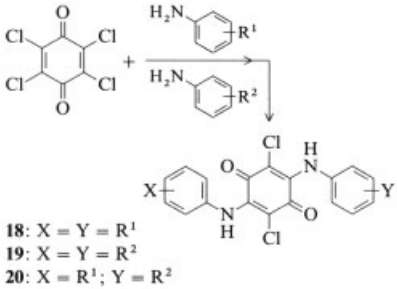

A *mixed chloranil condensation* [37] involves reacting two different amines with chloranil to give a mixture of three condensation products, two of which are symmetric (**18**, **19**) and one asymmetric (**20**). Ring closure then yields a mixture of dyes.

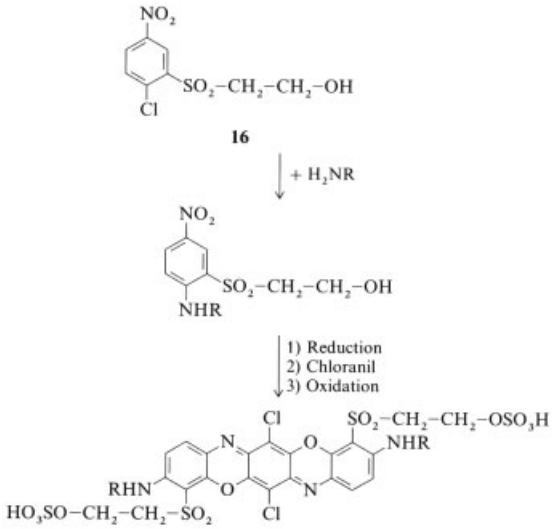

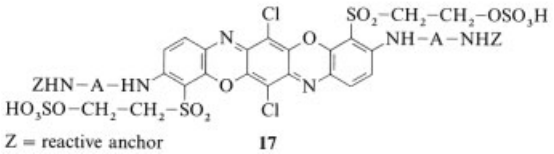

Z = reactive anchor **17**

Almost all of the amines described earlier can be used in this condensation. In the case of dye bases with amino groups, reactive anchors can be subsequently introduced by acylation [76].

3.4. Formazan Dyes [12,77]

Copper complexes of the formazan dye series (→Metal-Complex Dyes, Chap. 6.) are another alternative to reactive anthraquinone dyes, producing red to greenish-blue shades. Like triphenodioxazine dyes, copper formazans exhibit high molar extinctions (ε=25 000–30 000). These materials are derived from 1-(2-hydroxyphenyl)-3-phenyl-5-(2-carboxyphenyl)-formazan **21**, in which all three rings are capable of supporting groups that increase the compound's reactivity and solubility.

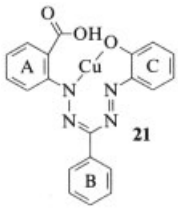

21

3.5. Phthalocyanine Dyes [12,39]

The water-soluble reactive phthalocyanine dyes (→Phthalocyanines) yield brilliant turquoise and green colors not available from any other dye category. The most important reactive phthalocyanine dyes contain copper or nickel as their central atom; they are substituted with sulfonic acid groups and also with reactive groups joined via sulfonamide bridges.

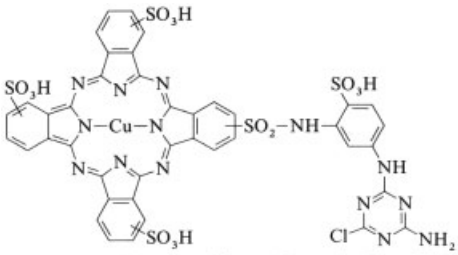

C.I. Reactive Blue 15 – Cibacron Turquoise Blue GF-P

4. Synthetic Procedures

The synthesis of selected reactive dyes is illustrated by the following examples taken from the patent literature.

Example 23 from EP 84,318 (Bayer, German Prior. of 01/15/82). 61.1 g of dye (obtained by coupling diazotized 4-(2-sulfooxyethylsulfonyl)aniline with 1-amino-8-naphthol-3,6-disulfonic acid at pH 1 – 2) is coupled at pH 6 – 7 with a diazonium compound synthesized from 33.7 g of the condensation product derived from 18.8 g of 1,3-diaminobenzene-6-sulfonic acid and 17 g of 2,4,6-trifluoro-5-chloropyrimidine. The crude material is salted out of solution, isolated, dried, and pulverized to a dark powder (**22**). This product gives a blue solution in water and dyes cotton a marine blue to black color [78] (see below).

Example 1 from EP 28,788 (Hoechst, German Prior. of 11/10/79). 32.0 parts of the hydrazone from 2-carboxyphenylhydrazine-4-sulfonic acid and benzaldehyde are suspended in 200 parts of water at 20 – 25 °C. The solid is dissolved by adding 50 % aqueous sodium hydroxide to obtain pH 6.5 –7.0. This solution is added at 5 – 15 °C to the aqueous diazonium salt solution obtained from a typical aqueous diazotization of 37.7 parts 4-(2-sulfooxyethylsulfonyl)-2-aminophenyl-6-sulfonic acid. The mixture is then dripped during 15 min into 100 parts by volume of an aqueous solution of 1 mol/L copper sulfate (at 10 – 15 °C) while the pH is maintained at 5.5 – 6.5 with 18 parts of soda. The reaction mixture is stirred for an additional hour at room temperature to ensure complete coupling, after which the pH is adjusted to about 1 with 30 parts by volume of conc. hydrochloric acid.

The strongly acidic solution is further stirred for one hour and then neutralized with either sodium carbonate or potassium carbonate to pH 5.5. The copper – formazan complex is salted out along with sodium or potassium chloride, filtered, washed with dilute aqueous sodium or potassium chloride solution, and dried at 80 °C. A dark powder results which gives a dark blue solution in water. It consists of an electrolyte-containing powdered sodium or potassium salt of the following acid:

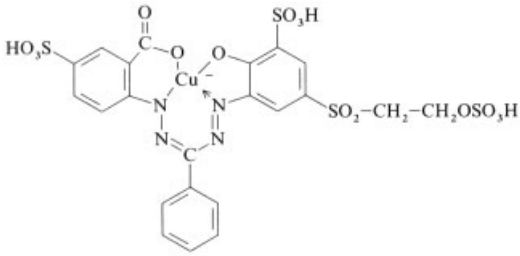

This compound is a very effective dye, rendering cotton and regenerated cellulose fibers a clear blue shade on prolonged treatment in the presence of an acid-binding agent. After a 10 minute soapy water wash the resulting colors prove to be very light- and waterfast [79].

Example 3 from EP 248,493 (Hoechst, Priority from 08/19/86). A triphenodioxazine compound is synthesized starting with 54 parts of aniline derivative **23** dissolved in 500 parts of water at pH 6.

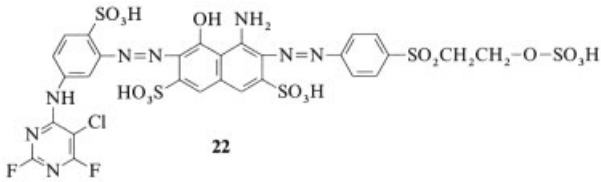

22

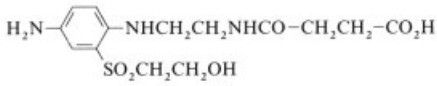

23

This solution is first warmed to 60 °C and then added in portions to 19 parts of chloranil at pH 6 – 6.5. The mixture is stirred for 3 h, after which the precipitated product is filtered, washed with water, and dried at 80 °C under vacuum.

Cyclization of the condensation product to the triphenodioxazine entails slowly adding 20 parts of the starting material at 20 °C to 220 parts of 10 % oleum with stirring until solution is complete. The temperature is carefully maintained at 20 – 25 °C while 11.5 parts of sodium peroxodisulfate are added. After one hour of further stirring the mixture is poured onto ice. The resulting suspension is adjusted to pH 1 – 1.5 with calcium carbonate and treated with sodium carbonate until the pH reaches 5.5. Calcium sulfate is then removed by filtration and washed with water. The combined filtrates are concentrated by evaporation or spray drying to give the desired product.

A sodium salt based on the following structure is obtained as a powder that also contains sodium sulfate:

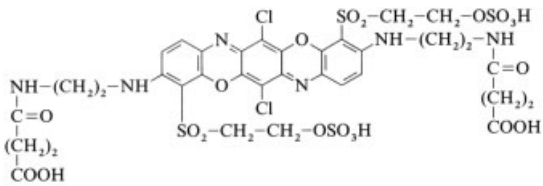

This material offers very good reactive dye properties [$\lambda_{(max)}$=612 nm] with respect to the cited materials, especially cellulose fibers (e.g., cotton), producing a strong blue shade with good fastness properties [80].

Synthesis of 24 [57]**:**

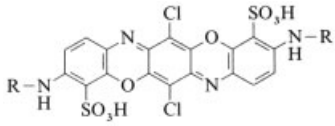

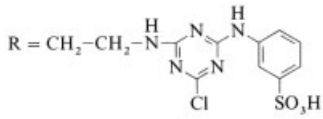

24

Chloranil Condensation Product. To a solution of 16 g of 4-(2′-aminoethylamino)-aniline-3-sulfonic acid (for synthesis, see [59]), in 200 mL of water is added a suspension of 6.1 g of chloranil in 12 mL of water and 5 mL of acetone. The pH is brought to 5.5 with 60 g of 10 % aqueous sodium acetate and the mixture is stirred at pH 5.3 – 5.5 for 2 h at 45 – 50 °C. After cooling to 10 °C, the precipitated brown solid is filtered, washed with 300 mL of acetone, and dried at 60 °C.

Dye Base. The chloranil condensation product (7.5 g) is added to 150 g of 25 % oleum, stirred for 10 min at 50 °C, cooled to 20 °C, and stirred another 30 min. The reaction mixture is then poured onto 500 g of an ice – water mixture, and the precipitate is filtered and washed with 200 mL of cold water. For purification, the crude product is dissolved at pH 11 in 7 g of 40 % sodium hydroxide solution and then reprecipitated with 30 g of a saturated aqueous solution of potassium acetate. The purified product is filtered, washed with 10 % aqueous potassium acetate and 800 mL of ethanol, and then dried at 60 °C.

Reactive Dye. A solution of 3.5 g of aniline-3-sulfonic acid in 50 mL of water containing 10 g of 2 mol/L sodium hydroxide solution is added to a suspension of 4 g of cyanuric chloride in 40 mL of water and 40 mL of acetone at 0 – 5 °C while the pH is kept between 6.0 – 7.0 with more sodium hydroxide. After 1 h the excess cyanuric chloride is removed by filtration and the filtrate added to a solution of 3.5 g of the dye base in 200 mL of water. The resulting mixture is heated for 8 h at 45 – 50 °C with the pH maintained at 9.0 – 9.2. After cooling to room temperature, the product is salted out with 25 wt % sodium chloride solution, filtered, washed with saturated sodium chloride solution, and dried at 60 °C.

5. Uses

5.1. Forms of Supply

In the past, reactive dyes were marketed mainly as powders, but other physical forms of these products have gained in importance. Dyes are now expected to be dedusted in order to minimize hygiene-related problems in the workplace. Moreover, the growing trend toward automated dye works has encouraged the development of more readily meterable products. Reactive dyes should also display good cold water solubility. These demands are all met by liquid dyes–highly concentrated, aqueous solutions–and by solids supplied in the form of pourable granules and finished powders [81–85].

5.2. Dyeing Processes

The most important application of reactive dyes is in the dyeing and printing of cellulose fibers and cellulose fiber-based materials. Wool, silk, and synthetic polyamide fibers also lend themselves to reactive-dye treatment. The process of dyeing cellulose requires the presence of a base (e.g., sodium carbonate or sodium hydroxide) together with an electrolyte, usually sodium chloride or sodium sulfate, which increases the fixation yield. The optimal dyeing temperature and pH is a function of the dye's anchor system. Thus, the optimal temperature for a "cold dye" is ca. 30 – 50 °C, whereas that for a "hot dye" is ca. 70 – 98 °C.

There are three major variants of the dyeing procedure for reactive dyes. The dominant method is still the *discontinuous exhaust dyeing process*, but the *semicontinuous pad-batch process*, and the *continuous process* are gaining in importance.

Improvements in the exhaust process itself and the development of new equipment have led to improved process safety and lower costs. The conventional exhaust dyeing methods are the "all-in" process, which means that all the dyebath components are introduced at room temperature and fixation is initiated by slow heating until the correct dyeing temperature is reached, and the "preconditioning method," in which an electrolyte is added to facilitate absorption of the dye onto the fibers in the absence of alkali. Base is added subsequently in portions at the dye temperature, causing fixation of the color onto the material. In a modern facility, alkali and salt may be metered and added to the bath automatically. This increases the extent of levelling, and it also reduces the amount of salt required and contributes to process safety. Reduced salt load in the process wastewater is important from an ecological standpoint [86–88].

In the *semicontinuous pad-batch process* the fabric is brought in contact with the dye solution and fixation is allowed to occur over several hours before the dyed material is rinsed, washed, and dried [89].

The *pad-steam procedure* is the most important of the continuous methods. Here the fabric is dyed and washed in a single process. The dye solution is padded onto the material, and drying and fixation are accomplished through treatment with hot air or steam at elevated temperature prior to washing.

A reactive dye constitutes only one of several components in the dye paste that is applied to cellulose fibers during a textile printing operation. Such a paste also includes a thickening agent (usually an alginate), urea, alkali, and an oxidant. Fixing is usually accomplished by a steam or dry-heat method. Of increasing importance is the *two-phase* printing process based on alkali-free print pastes. In this case the alkali required for fixation is padded onto the fabric after drying. Fixing is achieved subsequently either through a steam process or by cold dwelling.

In the application of reactive dyes to wool or silk attention must be directed to the sensitivity of these materials to base. Silk can be dyed by the exhaust process in a weakly basic medium at 60 °C, but the dyeing of wool is usually carried out in a weakly acidic bath [90, 91].

6. References

1. C. F. Cross, E. J. Bevan, *J. Res. Cellulose* **34 – 39** (1895 – 1900).
2. G. Schröter, *Ber. Dtsch. Chem. Ges.* **39** (1906) 1559.
3. BASF, GB 259 634, 1924.
4. D. H. Peacock, *J. Soc. Dyers Colour.* **42** (1926) 53.
5. Ciba, DE 560 035, 1930; 583 398, 1932 (R. Haller, A. Heckendorn).

6. I. G. Farbenind., GB 341 461, 1929.
7. H. Schweizer, O. Bayer, *Angew. Chem.* **73** (1961) 343.
8. Hoechst AG, DE 965 902, 1949.
9. J. Heyna, *Angew. Chem.* **74** (1962) 966; *Angew. Chem. Int. Ed. Engl.* **2** (1963) 20.
10. *Ullmann,* 3rd ed., **14,** 615.
11. ICI, DE-AS 1 019 025, 1 062 367, 1955.
12. D. Waring: *The Chemistry and Application of Dyes,* Plenum Press, New York 1990.
13. ICI, GB 838 337, 1957 (R. N. Heslop, W. E. Stephen).
14. ICI, GB 805 562, 1956; Ciba-Geigy, DE-AS 1 105 541, 1956.
15. Bayer, DE-OS 1 644 208, 1967.
16. ICI, GB 937 182, 1963 (G. A. Gamlen, C. Morris, D. F. Scott, J. Twichett). M. Miujumoto, R. Parham, *AATCC Int. Conf. Exhib.* (1986) 153.
17. M. Capponi, E. Metzger, A. Giamara, *Am. Dyestuff Rep.* **50** (1961) 23.
18. ICI, GB 917 780, 1963 (H. F. Andrew, V. D. Poole).
19. Bayer, GB 1 169 254, 1969 (H. S. Bien, E. Klauke).
20. Bayer, US 3 853 840, 1974 (K.-H. Schündehütte, K. Trautner).
21. Geigy, BE 644 495, 1964.
22. BASF, DE-AS 1 193 623, 1960.
23. Cassella, FR 1 403 233 (1963).
24. *Ullmann,* 4th ed., **20,** 113.
25. E. P. Sommer, *Am. Dyestuff Rep.* **47** (1958) 895.
26. H. Zimmermann, *Melliand Textilber.* **39** (1958) 1026.
27. Hermann Rath: *Lehrbuch der Textilchemie,* 3rd ed., Springer Verlag, Berlin 1972.
28. Hoechst, DE 966 651, 1950, US 3 387 914, 1963, DE-AS 1 103 886, 1958.
29. F. Wolf, S. Stefaniak, *Melliand Textilber.* **47** (1966) 767.
30. B. L. McConnell, L. A. Graham, R. A. Swidler, *Text. Res. J.* **49** (1979) 458.
31. A. H. M. Renfrew, J. A. Taylor, *Rev. Prog. Color. Relat. Top.* **20** (1990) 1.
32. ICI, DE-OS 2 244 537, 1973 (D. Waring).
33. Ciba-Geigy, EP 70 806, 1981.
34. Ciba-Geigy, EP 97 119, 1982.
35. Ciba-Geigy, EP 89 923, 1982.
36. Ciba-Geigy, EP 85 654, 1982.
37. Ciba-Geigy, EP 133 843, 134 193, 1983.
38. Cassella, DE-AS 2 614 550, 1976.
39. K. Venkataraman: *The Chemistry of Synthetic Dyes,* vol. VI: *Reactive Dyes,* Academic Press, New York –London 1972.
40. A. H. M. Renfrew, *Rev. Prog. Color. Relat. Top.* **15** (1985) 15.
41. G. Fischer, *J. Prak. Chem.* **19 (2)** (1879) 317.
42. P. Seidel, *Ber. Dtsch. Chem. Ges.* **23** (1890) 182.
43. Hoechst, DE 946 560, 1952.
44. E. Dietz, Hoechst AG, *11. Internationales Farbensymposium,* Montreux 1991.
45. Ciba, CH 358 188, 1957; FR 1 205 384, 1958; DE-AS 1 058 467, 1957; BE 567 435, 1958. Hoechst, BE 566 648, 1958; BE 570 437, 1958. Bayer, BE 569 439, 1958. Sandoz, BE 573 299, 1958; FR 1 233 530, 1958.
46. Ciba, FR 1 116 564, 1954.
47. ICI, GB 1 384 749, 1972; GB 1 368 158, 1972.
48. A. H. M. Renfrew, DE 2 503 611, 1975 *Rev. Progr. Colouration* **15** (1985) 15 – 20.
49. A. Pugin, Offical. Dig., *J. Paint Technol., Eng.,* **37** (486) (1965) 782.
50. J. Wolff, H. Henk, *Textilveredlung* **25 (6),** 1990, 213 –218.
51. Meister, Lucius und Brüning, GE 253 091, 1911.
52. ICI, GB 1 589 915, 1976.
53. Bayer, DE-OS 3 444 888, 1984; EP-A 296 411, 1987 EP-A 141 359, 1984.
54. Bayer, DE-OS 3 510 613, 1985.
55. Hoechst, EP-A 292 906, 1988.
56. ICI, GB 1 368 158, 1972; GB 2 059 985, 1981.
57. ICI, DE 2 503 611, 1975.
58. Ciba, EP 070 807, 1982. Bayer, DE-OS 3 410 236, 1984.
59. Ciba, EP-A 101 665, 1983.
60. Sumitomo, JP-Kokai 62 048 768, 1987; JP 57 14 654, 1980.
61. Sumitomo, EP-A 325 246, 1988. Mitsubishi, JP 12 89 867, 1988.
62. Sumitomo, EP-A 281 799, 1987.
63. ICI, EP 135 381, 1984.
64. Bayer, DE-OS 3 423 581, 1984. A. H. M. Renfrew, *J. Soc. Dyers Colour* **105** (1989) 262 – 4. Sumitomo, EP-A 385 120, 1990.
65. Sumitomo, JP 22 38 064, 1989. Bayer, DE-OS 3 635 312, 1988.
66. ICI, EP-A 279 122, 1987. Bayer, DE-OS 3 336 362, 1985; EP-A 199 036, 1986; EP-A 158 857, 1985; Hoechst, DE-OS 3 544 982, 1985.
67. Bayer, DE-OS 3 409 439, 1984.
68. Bayer, EP-A 205 080, 1986; EP-A 204 245, 1986; EP-A 190 603, 1986.
69. Bayer, EP-A 141 996, 1984.
70. Hoechst, DE-OS 3 625 386, 1986; DE-OS 3 625 347, 1986.

71. Hoechst, DE-OS 3 628 084, 1986; DE-OS 3 627 458, 1986; EP-A 355 735, 1988; DE-OS 3 426 727, 1984; Bayer, EP-A 334 172, 1988; DE-OS 3 439 756, 1984.
72. Bayer, DE-OS 3 439 755, 1984.
73. Sumitomo, DE-OS 3 403 087, 1984; EP-A 234 778, 1987. Bayer, EP-A 053 743, 1981.
74. IG Farben, GB 509 891, 1938; GB 509 893, 1938 ICI EP-A 356 014, 1989.
75. *Houben-Weyl,* 4th ed., **Chinone I,** 434.
76. Bayer, EP-A 299 328, 1987; EP-A 349 848, 1988. Sumitomo, EP-A 385 120, 1989.
77. H. Baumann, A. R. Heusel, *Fortschr. Chem. Forsch.* **7** (1967) 714.
78. Example 23 from EP 84 318 (Bayer AG).
79. Example 1 from EP 28 788 (Hoechst AG).
80. Example 3 from EP 258 493 (Hoechst AG).
81. H.-U. von der Eltz, *Melliand Textilber.* **68** (1987) 132.
82. W. Beckmann, F. Hoffmann, J. Wolff, *Textilveredlung* **24** (1989) 81.
83. D. Link, J. Moreau, *Textilveredlung* **24** (1989) 87.
84. W. Zysset, *Textilveredlung* **24** (1989) 91.
85. B.-T. Gröbel, M. Kunze, U. Mrotzeck, *TPI. Text. Prax. Int.* **10** (1991) 1111.
86. E. Ungermann, *TPI. Text. Prax. Int.* **42** (1987) 411 –414, 419 – 421.
87. D. Hildebrand, B. Renziehausen, D. Heilmann, *Melliand Textilber.* **69** (1988) 895.
88. A. Lübbers, I. Souren, H.-K. Rouette, *Melliand Textilber.* **71** (1990) 371.
89. E. Ungermann, *TPI. Text. Prax. Int.* **39** (1984) 63 – 66, 153 – 157, 262, 269 – 271.
90. H. Putze, *TPI. Text. Prax. Int.* **39** (1984) 1051.
91. H.-U. von der Eltz, *Textilveredlung* **7** (1972) 5, 297.

Reactor Engineering → Reactor Design and Optimization

β-Receptor Blocking Agents

Antiarrhythmic Drugs, Antihypertensives, Coronary Therapeutics, and Ophthalmological Preparations are separate keywords.

Brian N. C. Prichard, Clinical Pharmacology, University College London, London WC1 E6JJ, United Kingdom

Christopher W. I. Owens, Clinical Pharmacology, University College London, London WC1 E6JJ, United Kingdom

Barrie R. Graham, Clinical Pharmacology, University College London, London WC1 E6JJ, United Kingdom

1. Introduction 1
2. Pharmacodynamics 3
3. Pharmacokinetics 5
4. Clinical Uses of β-Adrenoceptor Blocking Drugs 6
5. Side Effects 10
6. Individual β-Blocking Drugs 11
7. References 17

1. Introduction

In 1905, Langley laid the ground work for the idea of differing adrenergic receptors [1–3] and suggested that the excitatory and inhibitory effects of sympathetic nerve stimulation or of giving adrenaline could be explained by the concept that "a cell may make motor or inhibitory receptive substances or both, and that the effect of a new impulse depends upon the proportion of the two kinds of receptive substance which is affected by the impulse" [4]. Sir Henry Dale remarked that the "...cardioaccelerator effect of adrenaline ... is not abolished by doses of ergot sufficient to reverse the effect on the blood pressure," while the noncardiac excitatory actions of adrenaline such as vasoconstriction were blocked by ergot, the inhibitory effects such as smooth muscle relaxation were not [5]. Dale almost anticipated Ahlquist's [6] simple division of catecholamine responses into α- and β-effects [1]. Barger and Dale [7] formulated the problem more in terms of receptors. They suggested that the most notable feature of the catecholamines they studied was "the more or less strict localization of their action to cells innervated by a certain morphological system, on which cells they act when present in very high dilutions. There must evidently be something in these cells, or connected with them and them only, which has a strong affinity for these amines." While acknowledging the possible importance of receptive side-chains they concluded "it by no means follows that the peculiar distribution of the action of nicotine or the sympathomimetic amines depends on the existence of specific chemical receptors in the cells peculiarly sensitive to them, as supposed by Langley" [7]. However, as Black observed: "Dale never referred to receptors again in his writings. Dale's attitude seems to have had a powerful effect in delaying the introduction of the idea of receptors into pharmacological teaching and his impact was still dominant when Ahlquist's paper appeared in 1948."

In 1933, Cannon and Rosenblueth [8] suggested that a transmitter substance was released by sympathetic nerve endings which combined with either excitatory (E) or inhibitory (I) substances to form sympathin (E) or sympathin (I). The complexes formed then acted on tissues or were released into the blood in varying concentrations. This sympathin theory explained sympathetic effects in hormone terms rather than receptors. The ideas of Ahlquist were seen to be revolutionary against this background.

α- and β-Receptors. Ahlquist reported the potency of six sympathomimetic amines in 1948 [6]. The amines fell into two distinct classes, which he explained by the existence of two types of receptor, α and β. The excitatory effects (vasoconstriction, stimulation of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_665

the uterus, nictitating membrane, ureter, and dilator pupillae) and one inhibitory action (intestinal relaxation) were termed *α-effects;* at the α-receptor adrenaline was the most potent followed by noradrenaline. Most of the inhibitory actions (vasodilation, inhibition of uterine and bronchial smooth muscle), and one excitatory action (myocardial stimulation) were called *β-effects;* at the β-receptor isoprenaline was the most active, adrenaline was third, and noradrenaline least effective. AHLQUIST therefore realized that receptors could either be excitatory or inhibitory, i.e., one receptor type did not have just either excitatory or inhibitory effects.

The adrenergic receptors have been further divided into α_1-, α_2-, β_1-, and β_2-receptors according to the properties of stimulant (agonist) drugs and blocking (antagonist) drugs [9, 10]. These receptors mediate various physiological effects, see Table 1.

Table 1. Adrenoceptors

Organ	Receptor	Physiological effect
Heart	$\beta_1 > \beta_2$	stimulation of contractility and heart rate
Smooth muscle of bronchi	β_2	bronchodilatation
Blood vessels smooth muscle	α_1 α_2	vasoconsttriction vasoconstriction
	β_2	vasodilatation
Skeletal muscle	β_2 β_2	glycogenesis tremor
Genitourinary tract smooth muscle	α_1 β_2	muscle contraction muscle relaxation
Fat tissue	α_2 β_1	inhibition of lipolysis stimulation of lipolysis
Platelets	α_2	aggregation
Liver	α_1 β_2	glycogenolysis glycogenolysis glycogenolysis
Pancreas	α_2 β_2	inhibition of insulin release stimulation of insulin release
Sympathetic nerve terminals (presynaptic)	α_2 β_2	inhibition of noradrenaline release stimulation of noradrenaline release

Dichloroisoprenaline (DCI) was found to block the bronchodilator, vasodilator, and the inhibitory actions of adrenaline [11, 12]. MORAN and PERKINS then reported the blockade of the cardiac stimulant effects of adrenaline by DCI [13], and they suggested that DCI acted on the β-receptor described by AHLQUIST. They also proposed that the compounds that blocked the effects of adrenaline be termed according to the receptor for which they had the greatest affinity; they introduced the terms α- or β-adrenergic blocking drugs. Although DCI was shown to block the adrenaline-induced increase in forearm blood flow [14] and control the tachycardia [relatively rapid heart action whether physiological (as after exercise) or pathological] and arrhythmias associated with the removal of pheochromocytoma [15], its clinical potential was not appreciated.

Development of Clinically Useful β-Blocking Drugs. BLACK paid special attention to the problem of coronary heart disease. He thought that rather than trying to increase coronary artery flow, as was then usual, a better idea might be to reduce oxygen consumption. He appreciated the clinical value of developing compounds to inhibit the sympathetic innervation to the heart. He led a team of scientists at ICI who put this idea into practice. This was at a time when the sympathetic system was only regarded as essential for the flight and fight reaction and the value of inhibiting it was not generally appreciated [2].

Research at ICI led to the introduction of pronethalol [16] which was found to be effective in angina pectoris [17–19], arrhythmias [20], pheochromocytoma [21], and in the treatment of hypertension [22]. Pronethalol's evaluation and use was delayed until 1964 because of the development of tumors in mice [23]. Most of the pioneer work with β-adrenoceptor blocking drugs was performed on this drug [24].

2. Pharmacodynamics

Differences in the pharmacological actions of β-adrenoceptor blocking drugs can be used as a basis for classification. They may be nonselective (i.e., they act on all the β-adrenergic receptors), have a selective action on the β_1-receptors, and possess vasodilator or α-receptor blocking properties. Besides the blocking activity they may be further subdivided according to the presence or absence of a stimulatory effect, i.e., intrinsic sympathomimetic activity (ISA) or partial agonist (stimulant) activity. Finally, a membrane stabilizing activity (MSA) or local anesthetic action may also be considered [25–27] (Table 2).

Table 2. Classification of beta-adrenoceptor blocking drugs

	Partial agonist effect (intrinsic sympathomimetic effect)	Membrane stabilizing effect (quinidine effect)
Division I: Nonselective ($\beta_1+\beta_2$) block		
Group I		
Oxprenolol	+	+
Alprenolol		
Group II		
Propranolol	–	+
Group III		
Pindolol	+	–
Carteolol		
Group IV		
Sotalol	–	–
Timolol		
Nadolol		
Division II: β_1 selective		
Group I		
Acebutolol	+	+
Group III		
Practolol	+	–
Celiprolol		
Group IV		
Atenolol	–	–
Betaxolol		
Metoprolol		
Bisoprolol		
Division III: Nonselective block and α block		
Group II		
Labetalol	–	+
Carvedilol		

β_1-Selective drugs were previously termed cardioselective, reflecting their pronounced effect on the heart in preference to other tissues, however as both β_1- and β_2-receptors are found in the heart the term "cardioselective" is inaccurate.

Competitive inhibition is shown by all presently available β-blocking drugs. The receptor occupancy is therefore dependent on the ratio between the concentration of agonist and antagonist. The agonist may be the natural neurotransmitter (noradrenaline), an endogenous stimulant, or an exogenous (administered) β-stimulant such as isoprenaline. It is possible to construct a series of dose – response curves to isoprenaline in the presence of a β-adrenoceptor antagonist. The dose – response curves demonstrate a parallel shift to the right as the dose of antagonist increases, the same maximum response is obtained on increasing the dose of the agonist, isoprenaline [28].

Blockade of Endogenous Sympathetic Activity and Hemodynamic Effects. When sympathetic activity is low (e.g., resting supine or asleep), there is a dose-related fall in heart rate on administering *β-blocking drugs without ISA* (e.g., propranolol) [29]. All β-blockers reduce exercising heart rate when sympathetic tone is high [30]. Acute intravenous administration of propranolol reduces heart rate and cardiac output, but increases peripheral resistance, therefore resting blood pressure is not initially affected [31]. With oral administration, heart rate and cardiac output are reduced [32, 33], but there is also a fall in peripheral resistance and blood pressure falls [32, 34, 35]. Changes in heart rate and cardiac output for non-ISA drugs are similar whether they are β_1-selective or nonselective [36, 37].

β-Blocking drugs with ISA reduce heart rate less than the non-ISA drugs, some have no effect on resting heart rate [38, 39]. At modest levels of exercise, reduction in the increase of heart rate on exercising may be similar with both ISA and non-ISA drugs [38]. At high levels of exercise, however, drugs with ISA (at maximum dosage) produce less of a reduction of exercise tachycardia than non-ISA drugs [30, 40, 41]. The increase in exercise heart rate remaining after a non-ISA drug has been given is largely a result of the withdrawal of the slowing effect of the vagus. Whereas the increase in heart rate may be similar with exercise, the possession of ISA means that the actual heart rate is higher, as the result of the modest sympathetic stimulus from ISA. The changes in cardiac output after the administration of β-blocking drugs usually parallels alterations in heart rate.

When given at low levels of sympathetic activity (i.e., at rest), drugs with ISA induce less reduction in cardiac output, as has been shown with pindolol [39, 41], practolol [36, 39], and oxprenolol [42]. Peripheral resistance increases only slightly and can even fall.

The rise in blood pressure during dynamic exercise is reduced by β-blockade [43]. The rise in blood pressure due to isometric exercise is not greatly modified by β-blocking drugs, although the final pressure may be reduced because of a decline in baseline values [44, 45]. There was

evidence of a smaller rise in diastolic pressure with atenolol than with propranolol, pindolol, or oxprenolol with isometric exercise [46] and less after metoprolol than following propranolol [47].

β_1-Selectivity. The efficacy of β-blocking drugs in inhibiting the effect of isoprenaline, a nonselective (i.e., acts at both β_1- and β_2-receptors) β-stimulatory drug on the heart shows considerable variation. For equivalent inhibition of exercise tachycardia, far less antagonism of the cardiac effects of isoprenaline is seen with the β_1-selective drugs than with the nonselective agents, with or without ISA [48–52]. The reason for this discrepancy appears to be the presence of cardiac β_2-receptors in addition to β_1-receptors [9,51,53]. The fact that both nonselective and β_1-selective drugs are more equivalent in terms of inhibition of exercise tachycardia indicates that exercise tachycardia (i.e., sympathetic nerve stimulation) is mediated via β_1-receptors.

A rise in blood pressure, involving the liberation of adrenaline, occurs with smoking [54] or insulin hypoglycemia [55] and tends to be aggravated when the β_2-dilator action of adrenaline is antagonized. Antagonism of the dilator component of adrenaline by nonselective β-blocking agents converts the action of adrenaline into that of noradrenaline [56]. It therefore might be expected that pressor responses involving the liberation of adrenaline are differently affected by β_1-selective blockade. For example, the pressor responses to a combination of coffee and cigarettes is attenuated by atenolol (a β_1-receptor blocking drug) whereas neither oxprenolol or propranolol, both nonselective (i.e., block β_1- and β_2-receptors) have any effect [57]. There is less increase in airway resistance with β_1-blockade than with nonselective blockade as might be expected from the predominance of β_2-bronchodilator receptors. This is the most important difference of the β_1-selective agents from the nonselective drugs [58–60]. However, even the β_1-selective drugs must be regarded as a potential risk in asthmatic patients [60].

There is also evidence that β_1-selective drugs produce a lower increase in low-density lipoproteins (LDL) and a lower decrease in high-density lipoproteins (HDL) than nonselective agents [61]. Additionally β_1-selective agents, un-like nonselective drugs, do not appear to prolong the action of insulin hypoglycemic drugs [61,62]. Finally β_1-selective agents do not impair the metabolic changes in muscles associated with exercise to the degree seen with other β-blocking drugs [61].

Intrinsic Sympathomimetic Activity (ISA). The most marked property of a β-blocking drug that possesses ISA is its central hemodynamic effect [63]. Such a drug produces a lower fall in resting heart rate [38,64], cardiac ouput [39,65], and peripheral blood flow [36, 66,67] than a drug without ISA. If the level of sympathetic tone is low enough (e.g., during sleep) and the amount of ISA is sufficient, the drug may even increase the heart rate. If the drug possesses ISA at the β_2-receptor (e.g., pindolol), then peripheral vasodilation due to stimulation of β_2-vasodilator receptors may also be important [68]. At high levels of exercise and full doses, a β-blocking drug with ISA gives a lower reduction in heart rate than one without [52,60]. In patients with asthma, the modest, β-stimulant action of an ISA drug on bronchial smooth muscle is not important and is more than offset by β-blockade – these patients are potentially very sensitive to any receptor blockade which worsens their asthma [69]. Chronic stimulation by a β-blocking drug with ISA reduces the number of β-receptors and thus when the drug is stopped abruptly there is no post-β-blocking drug hypersensitivity in contrast to non-ISA agents (which increase the number of β-receptors) [27,70–72]. Finally, there is evidence that drugs with ISA do not interfere as much in metabolic processes, particularly lipid metabolism [61], and the liver metabolism of other drugs [73].

Combination of β- and α-Blockade or Vasodilator Activity. *Labetalol* was the first combined β- and α-receptor blocking drug [28], and it has been extensively evaluated in hypertension [74,75]. *Medroxalol* also possesses both β- and α-blocking properties [76], and probably an additional vasodilator effect mediated by β_2-adrenergic stimulation [77,78]; it is an effective drug in hypertension [79]. *Carvedilol* has α_1-blocking properties and some additional vasodilator action [80–87]. *Celiprolol* has β_1-blocking properties and a vasodilatoy action which seem to be a combination of some β_2-

stimulatory action; also a direct vasodilator action [88, 89].

Several drugs which possess β-blocking and vasodilator activity have been evaluated. *Prizidolol* has been withdrawn because of animal toxicity [90]. *Bucindolol* and similar drugs are still being evaluated [91, 92].

3. Pharmacokinetics

The β-blocking drugs fall into two broad groups in terms of their elimination. *Lipid-soluble drugs* undergo extensive first-pass metabolism in the liver and readily penetrate the brain. *Water-soluble drugs* are excreted unchanged in the urine and have low brain penetration.

Lipid-Soluble Drugs. The liver metabolism of the lipid-soluble β-blocking drugs (e.g., alprenolol, metoprolol, labetalol, oxprenolol, propranolol) differs and their plasma concentration therefore varies between individuals after the same oral dose [93–95]. The relationship between the dose and the area under the plasma concentration curve (AUC) varies at low dosages. For example, at lower doses (e.g., 200 mg alprenolol [96] or 40 mg propranolol [97]) alterations in AUC are nonlinear and increases are lower than at higher dose levels. Oxprenolol, on the other hand, also undergoes hepatic metabolism but shows a linear, dose-related increase in plasma concentration [98]. Although the first-pass effect becomes saturated at high doses, systemic bioavailability does not exceed 50 %, even with larger doses of propranolol [99]. Liver metabolism and therefore steady-state plasma concentration are dependent on hepatic blood flow [94, 100]. Conditions that reduce liver blood flow (e.g., congestive heart failure, cirrhosis) lower hepatic clearance and prolong the drug half-life. Drugs such as propranolol, which reduce cardiac output and hepatic blood flow, slow their own metabolism and that of any other drug metabolized by the liver. Thus, while single doses have a plasma elimination half-life of 2 – 3 h, chronic administration may prolong this to 6 h.

The lipid solubility of pindolol is high but only about 50 % is metabolized in the liver and therefore first-pass effects are less important [101]. In the case of bisoprolol, about half is also eliminated via liver metabolism and half excreted via the kidney [102]. The intrinsic sympathomimetic activity of pindolol and oxprenolol means that they have less effect on cardiac output at rest, and therefore plasma half-lifes should be similar after acute and chronic administration [103].

Water-soluble drugs are eliminated in the urine unchanged (e.g., atenolol and sotalol) or as active metabolites. Half-life increases with renal insufficiency [104, 105]. The water-soluble β-blockers have more predictable plasma concentrations than the lipid-soluble drugs because they are independent of variable liver metabolism. Plasma concentrations of atenolol vary two- to five-fold between individuals [106, 107], much less than, for example, propranolol.

Distribution. The lipid solubility of drugs affects their volume of distribution within the body, which may vary about eight-fold (0.7 and 5.6 L/kg for atenolol and propranolol, respectively) and mostly exceeds the physiological body space [108]. There is a slight reduction of volume of distribution in renal failure or in the elderly (e.g., with sotalol) [109]. Drugs with a volume of distribution greater than the total body water indicate that concentration occurs in certain tissues (particularly lung, liver, and heart) [110, 111]. Drugs with high lipid solubility (e.g., alprenolol and propranolol) show rapid penetration across the blood – brain barrier [108]. The lung is a major organ of deposition for propranolol [110]. Although lipophilic substances have larger volumes of distribution than the water-soluble drugs, they have shorter plasma elimination half-lifes because drug clearance occurs via the liver rather than the kidneys and the liver has a greater elimination capacity [108].

4. Clinical Uses of β-Adrenoceptor Blocking Drugs

The β-adrenoceptor blocking drugs are widely used, particularly in cardiovascular medicine, and have markedly changed therapeutic practice [24].

Ischemic heart disease involves an imbalance between the use and supply of oxygen, this is temporary and reversible in angina pectoris, but sufficient in myocardial infarction to give myocardial necrosis. Angina pectoris was one of the original indications for β-adrenoceptor blocking drugs, and they have become highly effective prophylactics [75], see also → Coronary Therapeutics, Chap. 4. There are two approaches to therapy, vasodilation of coronary vessels (e.g., glyceryl trinitrate) or reduction of oxygen usage by the heart. β-Blocking drugs reduce cardiac oxygen demand, particularly during excercise and other forms of sympathetic stress such as emotion. They reduce heart rate, systolic blood pressure, and ventricular contractility. Although ventricular size increases, thus increasing oxygen consumption, this only partly offsets the benefit. There is an overall fall in coronary flow, presumably due to an autoregulatory effect. β-Blocking drugs redistribute blood flow in ischemia, blood flow is increased in vulnerable areas because the relatively well perfused myocardium uses less oxygen. In addition to these hemodynamic effects, β-blockers lower the free fatty acid concentration which improves myocardial utilization of carbohydrates.

Silent Ischemia [112]. Most episodes of silent ischemia are not associated with pain and are often hemodynamically indistinguishable from overt ischemic events. They may be preceded by an increase in heart rate. β-Blockade reduces the number of episodes.

Myocardial Infarction. β-Blocking drugs are useful in treating myocardial infarction [113]. Acute intravenous administration reduces infarct size by ca. 25 %. Blockade of the β_2-receptors appears to be the important component. The number of patients with threatened infarction who develop frank infarction is also probably significantly reduced [114, 115]. Effects on mortality in acute studies have been less impressive than in late intervention studies. Large-scale studies have indicated a significant reduction in vascular mortality of about 15 % in groups treated with metoprolol [116] and atenolol [117]. The value of intravenous followed by oral β-blockade is probably maximized if intravenous administration occurs within 12 h of the onset of pain; after one year, a significant reduction in mortality is still apparent [117, 118]. The incidences of reinfarction and cardiac arrest are significantly decreased (15 – 18 %) by early intravenous β-blockade. There have been some problems with morbidity or adverse reactions with these early-entry studies, but while bradycardia and hypotension can occur they rarely require active intervention. Cardiogenic shock and heart fail-ure, even in high risk patients, are uncommon [116, 117].

The long-term administration of β-blockers after myocardial infarction (secondary prevention studies) indicates that β-blockers decrease total mortality by about 25 %. This benefit derives from β_1-blockade, and both β-selective and nonselective agents give similar benefits [115]. Pooled analysis showed a 10 % decrease in mortality with β-blockers possessing ISA, significantly less than the 30 % with β-blockers without ISA [115]. A positive relationship between the decrease in resting heart rate on β-blocker treatment and the decrease in mortality has been suggested: β-blockers with moderate to high ISA have little effect on resting heart rate [119].

Hypertension. The β-adrenoceptor blocking drugs are first-choice drugs for treating hypertension [120] and have been in use since the early 1960s (→ Antihypertensives, Chap. 4.2.3.). Blood pressure falls regardless of the presence or absence of β_1-selectivity, intrinsic sympathomimetic activity, and membrane activity. β_1-Blockade is the important factor and a high level of pure β_1-stimulation prevents the fall of blood pressure. The membrane-active, non-β-blocking isomers of β-blocking drugs are without antihypertensive activity [121, 122].

Attempts to explain the hypotensive effect of β-adrenergic blocking drugs include a direct action on the central nervous system (CNS), blocking of adrenergic neurones, antirenin activity, an increase of vasodilator prostaglandin concentrations, and secondary effects resulting from pressor stimuli caused by reduction in cardiac activity consequent to β-blockade. However, the exact mode of action remains unclear [120], and may not be the same in all patients.

The greatest attention has been given to correlation of the effect of β-adrenoceptor blocking drugs on blood pressure and renin levels. Renin acts on angiotensinogen to form an-

giotensin I, a blood pressure-regulating peptide (→ Peptide and Protein Hormones, Chap. 5.1.). Patients with high renin levels responded best to propranolol, whereas patients with normal renin levels had a lower fall in blood pressure [123]. The correlation between renin levels and response of blood pressure with a variety of β-adrenoceptor blocking drugs is described [123, 124]. A relatively low dose of propranolol (160 mg/d) readily lowered the blood pressure in high-renin patients, but higher doses (320–960 mg/d) were needed to lower the blood pressure in low-renin patients, independent of effects on renin levels. However, other investigators have not found a correlation between renin levels and response to β-adrenoceptor blocking drugs (propranolol [34, 125–127], pindolol [127, 128], sotalol [129], timolol [130], or metoprolol [131, 132]).

The efficacy ofβ-adrenoceptor blocking drugs is similar to that of other antihypertensive drugs (diuretics, angiotensin converting enzyme inhibition, and calcium antagonists) but give better control of the exercising blood pressure. The β_1-selective agents have a slightly greater antihypertensive effect than the nonselective agents [24]. It had been hoped that the use of β-blocking drugs in the treatment of hypertension would reduce the incidence of myocardial infarction (primary prevention) because a reduction in further infarction had been demonstrated when β-blocking drugs were given after myocardial infarction. The result of studies, however, is not clear cut. Investigations in younger patients [133–135] suggest β-blockers may have some benefit, but in elderly patients no benefit was seen [136].

Orthostatic Hypotension. There has been some attempt to treat the excessive fall in blood pressure that occurs in the upright position (orthostatic hypotension) with β-receptor blocking drugs, the rationale being that β_2-blockade would allow the unopposed α- constrictor activity of circulating adrenaline to raise the blood pressure. Some reports show good results with pindolol (moderate ISA) [137] or propranolol, but others have not found either drug effective [138]. Some benefits have been obtained with agents with ISA (e.g., prenalterol or xamaterol) [139], which may also give additional benefit by maintaining or increasing cardiac output.

Dissection of the Aorta. The tearing of the wall of the aorta, dissection, is a highly lethal condition: without treatment 90 % of patients die within three months of the event [140]. Medical treatment is aimed at lowering blood pressure, it would also be useful to reduce the speed and force of contraction of the heart. β-Blockade should be administered intravenously. The blood pressure, if high, may be acutely lowered with, for instance, nitroprusside [41]. For distal dissections, long-term medical therapy rather than surgery may be desirable, the goal being good antihypertensive control and a decrease in ventricular ejection velocity [141]. β_1-Blockade is required and thus the choice of β-blocker is unimportant, unless there is a contra-indication to β_2-blockade (e.g., asthma) in which case a β_1-selective agent should be cautiously used. Late survival can be expected to be reduced from 90 % in untreated patients to about 40 – 50 % treated by drugs or surgery [141, 142].

Valvular Disease. In mitral valve prolapse, oral propranolol reduces the fall in blood pressure on standing associated with faintness and syncope [143]. β-Blockade reduces the pressure gradient across the mitral valve in mitral stenosis by increasing the time available for diastolic filling [24]. There have not been many studies on aortic valvular disease: administration of 100 mg atenolol for one week improved exercise tolerance and reduced exercise-induced angina and dyspnea in patients with aortic stenosis.

Cardiac Arrhythmias. The effect of β-blocking drugs varies considerably according to the type of cardiac arrhythmia. Catecholamines accelerate the rate of diastolic depolarization of automatic pacemaker cells and therefore increase the heart rate, accelerate the conduction process, and shorten the refractory period of ventricular cells. They therefore predispose the heart to extrasystoles and lower the threshold to ventricular fibrillation [144]. The β-adrenoceptor blocking drugs inhibit these influences [145], but may also indirectly modify processes which predispose to arrhythmias such as myocardial ischemia or thyrotoxicosis [24]. See also → Antiarrhythmic Drugs.

The direct antiarrhythmic property of β-blockers occurs via β_1-blockade [24]. β_1-

Selective and nonselective agents are equally effective in inhibiting stress-induced arrhythmias. Membrane-stabilizing activity is relevant only at very high dosages [146] (e.g., propranolol > 640 mg/d) where a quinidine-like antiarrhythmic effect might contribute [24]. Sotalol possesses additional antiarrhythmic properties (i.e., it prolongs the duration of the action potential) and so differs from the other β-blockers [147, 148].

Supraventricular Tachycardias. Abnormal (ectopic) atrial tachycardias are frequently associated with increased sympathetic activity and may respond well to a β-blocker [149]. Multifocal atrial tachycardia associated with pulmonary disease is usually unresponsive. Intravenous β-blockade is often effective in paroxysmal (sudden) supraventricular tachycardias that are not due to pre-excitation. Additionally vagal maneuvers are often more successful in the presence of a β-blocker [150]. Chronic oral administration of β-blockers often prevents the onset of these arrhythmias, particularly if the arrhythmia is precipitated by emotion or exercise [144]. The effective refractory period of accessory pathways involved in supraventricular tachycardias arising from preexcitation (e.g., Wolff – Parkinson – White Syndrome, WPW) is only influenced by sotalol. Nevertheless by altering conduction and refractoriness in the atrioventricular (AV) node or suppressing initiating premature beats, intravenous or oral propranolol can be very effective in preventing or suppressing these arrhythmias [151]. The WPW syndrome is an abnormal route of conduction of the cardiac impulse between the atria and ventricles which predisposes to tachyarrhythmias. It should not be treated with β-blockade if it is complicated by atrial flutter and fibrillation. Conversion of atrial fibrillation to sinus rhythm by β-blockade may be seen if the precipitating cause involves a high sympathetic tone (e.g., exercise) [150], long-standing cases rarely convert to sinus rhythm [152]. Likewise atrial flutter, rarely is converted to sinus rhythm with β-blockers, but as a result of increasing the AV block a useful reduction of the heart rate can be achieved [150, 153].

Ventricular Arrhythmias. β-Blockade may be effective in ventricular premature beats which occur in ischemic heart disease, or that are induced by exercise, stress, or emotion [154]. There is the risk of heart failure, and sustained ventricular tachycardia often does not respond to β-blockers. However, paroxysmal ventricular tachycardia responds, particularly if induced by high sympathetic activity [150]. Ventricular tachycardia is a medical emergency and cardioversion should be employed, a β-blocker can sometimes be helpful [24].

Ventricular arrhythmias arising from digitalis toxicity often respond well to β-blockers, but there is a risk of fatal heart block [150]. β-Blockers are the drugs of choice in the prevention of premature ventricular beats, ventricular tachycardia, and ventricular fibrillation associated with a prolapsing mitral valve [24].

Outflow Tract Obstruction. Sufferers from the congenital heart disease known as Fallot's tetralogy may experience attacks of severe cyanosis and syncope, often precipitated by effort, crying, or stress. These episodes are associated with shut-down of the hypertrophic, right ventricular outflow tract and are thought to be the consequence of stimulation of the tract by catecholamines. This has been the rationale for the use of β-blocking drugs. The use of propranolol is probably preferable to palliative shunt surgery in buying time before corrective surgery [155].

In patients with obstruction of the outflow tract from the left ventricle, the gradient across the obstruction may appear only with sympathetic stimulation (e.g., isoprenaline infusion or exercise). Intravenous administration of β-blockers usually diminishes this gradient. Long-term use of β-blockers may abolish or diminish symptoms and may decrease sudden death [156].

Congestive Cardiomyopathy. The cause of congestive cardiomyopathy (COCM) is not known [157]. Catecholamines have toxic myocardial effects and patients with COCM may be hypersensitive to them due to downregulation of cardiac β_1-receptors. β-Blockers may diminish the effect of catecholamines: heart rate is decreased, with a decreased myocardial oxygen requirement and an up-regulation of cardiac β_1-receptors [24]. β-Blockers given initially in low dosage can benefit patients with COCM [158],

but some are precipitated into heart failure and shock.

Portal Hypertension. Cirrhosis of the liver is often the result of prolonged, excessive alcohol consumption, and leads to a rise in blood pressure in the portal venous system. β-Blockade reduces portal venous pressure because cardiac output falls, and possibly also β_2-blockade results in unopposed α-induced vasoconstriction in the hepatic artery. The nonselective drugs are probably more effective than β_1-selective agents [159–162], while labetalol, with α- and β-blocking properties, is ineffective [163].

Pheochromocytoma is a tumor, usually of the medulla of the adrenal gland, and it is characterized by the secretion of high levels of noradrenaline and adrenaline. Severe hypertension due to α-mediated vasoconstriction, tachycardia, and arrhythmias due to cardiac β-stimulation can thus occur. The α-blocker phenoxybenzamine is most often used for treating the hypertension. A β-blocker is often needed to counteract increased β-activity that occurs when phenoxybenzamine has been given. Such a combination provides good preparation for the removal of the tumor, diminishes hypertension and arrhythmias, and reverses hypovolemia associated with peripheral vasoconstriction [164, 165]. Labetalol, a drug that combines α- and β-blocking activity has also been employed [75]. An excessive rise in blood pressure from a β-blocker given alone can occur due to unopposed α-activity [56].

Subarachnoid Hemorrhage. The leaking of blood from cerebral vessels into the subarachnoid space is extremely stressful. It is associated with high levels of catecholamines, electrocardiogram (ECG) abnormalities, and focal myocardial necrosis. Patients with particularly high catecholamine levels have a poor prognosis. Many of the ECG abnormalities can be prevented by the administration of propranolol and prophylactic β-blockade to prevent catecholamine-induced myocardial necrosis [166, 167].

Head injury is associated with cardiovascular and metabolic signs of marked sympathetic overactivity. Ischemia and arrhythmias occur. Atenolol (intravenously then orally) decreases the incidence of supraventricular arrhythmias and prevents catecholamine-induced myocardial necrosis [168].

Anxiety. Patients who suffer from anxiety may experience cardiovascular symptoms such as palpitations, excessive effort dyspnoea, fatigue, precordial discomfort, tremor, and sweating. The symptoms are similar to those of sympathetic overstimulation, β-blockers are therefore used to reduce or remove the symptoms [169–171].

Migraine. Propranolol is an effective prophylactic and reduces the frequency of migraine headaches. β-Blockers with ISA do not appear as effective, while positive studies have been reported with the β_1-selective atenolol [24, 172].

Thyrotoxicosis. The part played by catecholamines in thyrotoxicosis is not clear, but many signs and symptoms are similar to adrenaline overactivity and propranolol is effective in controlling them. It is usually used prior to surgery or in the period the antithyroid drugs take to show their benefit [24].

Glaucoma. Increased intraocular pressure in untreated glaucoma leads to blindness. Many β-blockers lower intraocular pressure, but their local anesthetic effect is a disadvantage because an anesthetized cornea is injury prone. A drug without local anesthetic effects but with sufficient lipid solubility to penetrate the cornea is indicated (e.g., timolol) [24, 173, 174]. β-Blockers may be absorbed from the eye into the systemic circulation in sufficient amounts to cause adverse reactions such as asthma in susceptible patients [174]. See also → Ophthalmological Preparations, Chap. 5.2.

Other Uses of β-Blocking Drugs. β-Blockers have also been tried in several other conditions. β-Blockers, primarily propranolol, have been used in tremors, including the tremor associated with alcohol withdrawal [24].

5. Side Effects

Respiratory Side Effects. The possible precipitation of an asthma attack in susceptible subjects is the most important adverse reaction to β-blockers. The drugs significantly increase airway resistance in asthmatic subjects due to the inhibition of sympathetic tone in bronchial smooth muscle [175]. β_1-Selective blocking drugs have less effect than nonselective agents [59, 176, 177].

Although the effect is small, β-blocking drugs probably increase airway obstruction in patients with chronic obstructive lung disease [178, 179]. They also reduce the respiratory response to carbon dioxide [180].

Heart Failure. Patients with heart failure are crucially dependent on sympathetic activity to the heart to maintain cardiac output and should not normally be given a β-receptor blocking drug without prior administration of digitalis and diuretics to improve cardiac function. If these drugs alleviate the symptoms it is reasonable, with caution, then to administer to the patient a β-blocking drug. Patients who have developed heart failure while taking propranolol have been subsequently able to tolerate the drug after the administration of digoxin and a diuretic [181]. If heart failure remains uncontrolled, β-adrenergic blocking drugs should not be given. β-Blockers with ISA (alprenolol, oxprenolol, penbutolol, and pindolol) can also precipitate heart failure. A study is being undertaken to assess possible benefits of very small doses of a β_1-selective agent (bisoprolol) in heart failure [24].

It had been thought that patients receiving β-blocking drugs might be at special risk of developing acute heart failure if they sustained an intercurrent myocardial infarction. Controlled trials of intravenous, followed by oral administration of propranolol or atenolol show that this is not the case [182, 183]. Prophylactic use in high-risk patients following myocardial infarction does not increase the incidence of heart failure [184, 185]. Larger-scale trials in hypertension have not shown an increased incidence of heart failure [186, 187].

Peripheral Circulation. β-Adrenergic blocking drugs often cause cold extremities [186] as a result of peripheral vasoconstriction due to reduced cardiac output and peripheral β-blockade. There is some evidence that β_1-selective drugs give a lower incidence of cold extremities [188], but can cause a reduction in peripheral flow [189] and the exceptional case of gangrene [190]. This is also the case with nonselective drugs and those with some partial agonist activity [191].

CNS. Sleep disturbances, insomnia, increased dreaming, and tiredness are not uncommon, and there have been reports of hallucinations, depression, delirium, headache, and dystonic reactions [184, 192]. Large-scale studies have not confirmed an increased frequency of depression [184, 186]. Dreams are more frequent with the lipid-soluble drugs [193, 194]. There is some evidence that lipid-insoluble β-blocking drugs (e.g., atenolol [193, 195] or sotalol [196]) produce fewer CNS side effects.

Fatigue. Mechanisms that might explain fatigue caused by β-blocking drugs [197, 198] include reduction of exercise cardiac output, depressed release of lactic acid from muscle, a reduced blood glucose level on exercise, or interference with β_2-mediated cellular sodium efflux and potassium influx [199]. Some observations suggest that β_1-selective drugs (atenolol) interfere less with exercise than nonselective ones (propranolol) [200].

Gastrointestinal Side-Effects. Indigestion with β-blockers often with nausea [186], is not uncommon and can be avoided by taking the drug with meals. An 11 % incidence of gastrointestinal side-effects was found in a survey of 23 published studies on propanolol [192].

Sensitivity Reactions. Practolol was associated with severe sensitivity reactions which only became apparent after it had been generally released, it was then withdrawn [24]. The only other β-blocker that may be associated with a high incidence of positive antinuclear antibodies appears to be acebutolol [201, 202]. The oculomucocutaneous syndrome is rare [198, 203].

Skin reactions, most often psoriasiform or lichenoid, are associated with β-blocking drugs [186, 198]. Retroperitoneal fibrosis has been reported in patients taking β-blocking drugs, but this is thought to be coincidental rather than

causative [198, 204]. Sclerosing peritonitis has been associated with practolol, but probably does not occur with other β-blockers [205].

Withdrawal of β-Blocking Drugs. There is a period of β-receptor hypersensitivity after withdrawal of β-blocking drugs without ISA. This is important in patients with ischemic heart disease and may be associated with an increased incidence of ischemic events. It is of less importance in uncomplicated hypertensive patients, but gradual reduction of dosage is recommended [206]. The occurrence of important adverse reactions to β-adrenergic blocking drugs is relatively uncommon, particularly if cases of heat failure and asthma are avoided. Severe adverse cardiac reactions may occur soon ofter the start of treatment, even with small doses; but once therapy is introduced, a gradual increase in dosage is not likely to precipate adverse reactions.

6. Individual β-Blocking Drugs

More detailed accounts of individual β-adrenoceptor blocking drugs can be found [24, 207]; their chemistry is summarized in [208] and their structure – activity relationship discussed in [209]. The principal β-blocking drugs that have been marketed are described, many other examples have been reported.

Acebutolol (hydrochloride) [*37517-30-9*], 1-(2-acetyl-4-*n*-butyramidophenoxy)-2-hydroxy-3-isopropylaminopropane, β_1-selective, some partial agonist activity (P.A.A.), [210–213]. $C_{18}H_{29}ClN_2O_4$, M_r 372.9, crystals *mp* 141 – 143 °C, octanol – water partition coefficient 0.68.

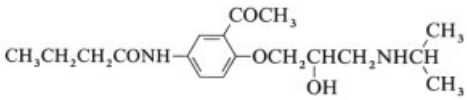

Trade Names. M & B 17 803 A, Sectral, Diasectral, Neptall, Prent.

Pharmacodynamics. Acebutolol is a competitive antagonist at β-receptors, it is selective for the β_1-receptor with less effect at the β_2-receptor. It has some partial agonist activity.

Pharmacokinetics. Acebutolol is fairly well absorbed (> 50 %) with considerable presystemic metabolism; the systemic bioavailability is 50 – 70 %. The peak concentration is reached 2 – 3 h after oral administration, the half-life is 5 – 9 h. Acebutolol is metabolized to diacetolol which is also a β-blocking agent, about 3 – 25 % is excreted unchanged. Acebutolol has a volume of distribution of about 1.2 L/kg, which indicates some accumulation in the tissues.

Uses. Acebutolol has been promoted for the treatment of hypertension, angina pectoris, and in certain cardiac arrhythmias.

Atenolol [*29122-68-7*], 1-*p*-carbamoylmethylphenoxy-3-isopropylamino-2-propanol, β_1-selective, [214–217]. $C_{14}H_{22}N_2O_3$, M_r 266.35, crystals from ethyl acetate *mp* 146 – 148 °C, octanol – water partition coefficient 0.015.

Trade Names. ICI 66 082, Tenormin, Atenolol, AteHexal, Ibinolo, Myocard, Prenormine, Seles Beta, Selobloc, Tenoblock, Tenormin, Uniblock, Uniloc.

Pharmacodynamics. Atenolol is a competitive antagonist at the β_1-receptors but its action at β_2-receptors is very weak. It is without partial agonist activity.

Pharmacokinetics. As with other hydrophilic drugs, oral absorption of atenolol is poor (ca. 44 %). There is no significant hepatic metabolism. The peak plasma concentration occurs 2 – 4 h after an oral dose, plasma half-life is 6 – 8 h. Most is excreted unchanged in the urine. The volume of distribution is low (0.7 L/kg) as atenolol has a low lipid solubility; it is the most hydrophilic β-blocking drug that is generally available.

Uses. Atenolol has been promoted for the treatment of hypertension, angina pectoris, in cardiac arrhythmias, and myocardial infarction.

Betaxolol (hydrochloride)[*63659-19-8*],(±)-1-(isopropylamino)-3-[*p*-(cyclopropylmethoxyethyl)phenoxy]-2-propanol, β_1-selective, [218, 221]. $C_{18}H_{30}ClNO_3$, M_r 343.9, crystals *mp* 116 °C.

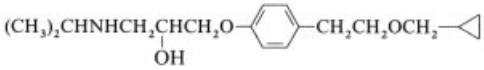

Trade Names. SLD 212 SL-75 212 Kerlone, Betoptic, Betoptima.

Pharmacodynamics. Betaxolol is relatively β_1-selective, a competitive antagonist with weak membrane-stabilizing action, and no partial agonist activity at low doses, although there may be some at high doses.

Pharmacokinetics. Betaxolol is well absorbed, bioavailability is 80 – 90 % and, although lipid solubility is high, first-pass metabolism is minimal. Peak concentration is achieved 2 – 4 h after oral administration, plasma half-life is 13 – 24 h, the volume of distribution is about 6.1 L/kg. Metabolism is slow, ca. 10 – 15 % is excreted unchanged.

Uses. It has been promoted for the treatment of hypertension and angina, and as eye drops for glaucoma.

Bisoprolol (hemifumarate) [*66722-44-9*], (±) -1-[[α-(2-isopropoxyethoxy)-*p*-tolyl]oxy]-3-(isopropylamino)-2-propanol, β_1-selective, [70, 222–226]. $C_{18}N_{31}NO_4$, M_r (free base) 325.45, bisoprolol hemifumarate: $C_{36}H_{64}N_2O_{12}$, crystals *mp* 100 °C, octanol – water partition coefficient 4.8.8

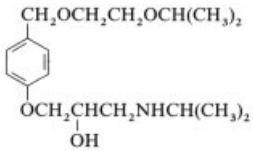

Trade Names. EMD 33.512, Bisobloc, Concor, Emconcor, Emcor, Pacteus, Detensiel, Euradel, Isoten, Maintate, Monocor, Soprol.

Pharmacodynamics. Bisoprolol is a highly selective β_1-receptor antagonist. It has no partial agonist activity, and no membrane activity within the therapeutic range.

Pharmacokinetics. Bisoprolol is well absorbed (> 90 %). Peak concentration is reached in 3 h. The plasma half-life is relatively long (ca. 10 – 12 h). It is modestly lipophilic with an apparent volume of distribution of 3 L/kg. About 30 % of the plasma concentration is bound to protein. About 50 % of bisoprolol is metabolized in the liver, 50 % is excreted unchanged.

Uses. Bisoprolol has been promoted for the treatment hypertension, angina pectoris, and some cardiac arrhythmias, it is also used in the management of hyperthyroidism.

Carteolol (hydrochloride) [*51781-21-6*], 5-[3-(*tert*-butylamino)-2-hydroxypropoxy]-3,4-dihydrocarbostyril, [227–231]. Low selective β-blockade, with partial agonist activity, $C_{16}H_{24}N_2O_3HCl$, M_r 328.8, *mp* 278 °C, octanol – water partition coefficient 0.21.

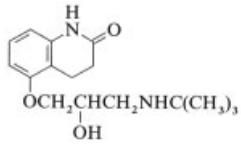

Trade Names. Abbott 43 326, OPC 1085 Arteolol, Arteoptic, Caltridren, Carteol, Cartrol, Endak, Mikelan, Ocupress, Tenalet, Tenalin, Teoptic.

Pharmacodynamics. Carteolol is a competitive antagonist at both the β_1- and β_2-receptors with some partial agonist action. It has a weak membrane-stabilizing (local anesthetic) action.

Pharmacokinetics. Carteolol is well absorbed (ca. 95 %). There is some presystemic (first pass) hepatic metabolism (ca. 15 %) so that systemic bioavailability exceeds 80 %. The peak plasma concentration is reached in about 1.5 h, the plasma half-life averages 5.8 h. Carteolol is excreted mainly unchanged, a small amount is metabolized by the liver. Plasma protein binding is about 15 %, the volume of distribution is about 4 L/kg.

Uses. Carteolol has been used in the treatment of hypertension, angina pectoris, arrhythmias, and in cardiac neurosis. An ocular preparation is available for the treatment of glaucoma.

Carvedilol [*72956-09-3*], (RS)-1-(9*H*-carbazol-4-yloxy)-3-[[2-(2-methoxyphenoxy)ethyl] amino]-2-propanol, nonselective β-blockade, α-receptor blockade, [232–234]. $C_{24}H_{26}N_2O_4$, M_r 406.5, *mp* 114 – 115 °C, octanol – water partition coefficient 3.3.

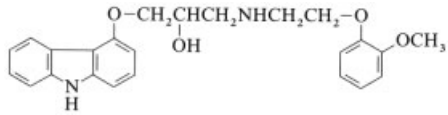

Trade Names. BM 141 90, DQ 2466, Kredex, Dilatrend.

Pharmacodynamics. Carvedilol has a nonselective β-adrenoceptor blocking activity for both β_1- and β_2-receptors. It additionally has a vasodilator activity which seems to be dependent on an α_1-receptor blocking action, it thus inhibits the vasoconstrictor action of adrenaline and noradrenaline in vascular smooth muscle. The vasodilator action results in a reduced peripheral resistance, and represents an additional mode of action for lowering blood pressure.

Pharmacokinetics. Carvedilol is well absorbed (ca. 80 %) but with extensive presystemic first-pass metabolism by the liver. Systemic bioavailability is therefore low (ca. 23 %). The peak plasma concentration is reached 1 – 2 h after the dose. The plasma half-life averages 6.4 h. Carvedilol is eliminated by liver metabolism. Over 95 % is bound to plasma proteins, its volume of distribution is 2.0 L/kg.

Uses. Carvedilol is promoted for the treatment hypertension. Its value in other conditions (e.g., angina) is still being evaluated.

Celiprolol [*56980-93-9*], *N*-[3-acetyl-4-(3′-*tert*-butylamino-2′-hydroxy)propoxy]phenyl-*N*′-diethylurea, β_1-selective, P.A.A., [235–237]. $C_{20}H_{33}N_3O_4$, M_r 379.5, *mp* 110 – 112 °C.

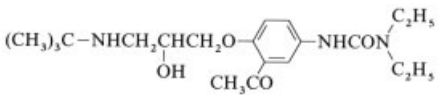

Trade Names. Celectol, Selectol.

Pharmacodynamics. Celiprolol is a competitive β_1-receptor antagonist. It has partial agonist activity which is mediated via β_2-receptors. It has some weak α_2-blocking action but this may not be clinically important.

Pharmacokinetics. Absorption of celiprolol is only about 20 %. Peak levels are reached 2 – 3 h after an oral dose, the plasma half-life is 4 – 5 h. Very little is metabolized, it is almost all excreted unchanged. The volume of distribution is low, celiprolol is markedly hydrophilic.

Uses. Celiprolol has been promoted for the treatment hypertension and angina pectoris.

Labetalol (hydrochloride) [*32780-64-6*], 5-[1-hydroxy-2-[(1-methyl-3-phenylpropyl)amino] ethyl]salicylamide hydrochloride, nonselective β-blockade, some partial agonist action, α-receptor blockade, [75, 238–240]. $C_{19}H_{24}N_2O_3$-HCl, M_r 364.9 (free base 328.4), *mp* 187 – 189 °C, octanol – water partition coefficient 7.08.

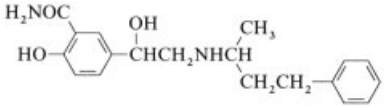

Trade Names. AH 5158, SCH 15 719 W, Amipress, Ibidomide hydrochloride, Ipolab, Labelol, Labrocol, Normadate, Normodyne, Presdate, Pressalolo, Trandate.

Pharmacodynamics. Labetalol is a competitive, nonselective β-antagonist, it blocks both β_1- and β_2-receptors with some β_2-agonist effect. Labetalol also has some α_1 (postsynaptic) blocking activity and thus inhibits the constrictor effect of adrenaline and noradrenaline on vascular smooth muscle. This reduces peripheral resistance, representing an additional mode of action to help lower the blood pressure.

Pharmacokinetics. Labetalol is rapidly absorbed but undergoes variable extensive presystemic metabolism. The systemic bioavailability lies between 11 and 86 %. The peak plasma concentration is reached in 40 min to 1.5 h. The plasma half-life averages 3.3 h. Labetalol is almost entirely metabolized by the liver, < 5 % is excreted unchanged in the urine. The plasma protein binding is ca. 50 %, the volume of distribution varies between 3.4 and 10.7 L/kg.

Uses. Labetalol is used for the treatment of hypertension, angina in patients with hypertension, and in pheochromocytoma.

Metoprolol (tartrate), [*56392-17-7*], (±)-1-(isopropylamino)-3-[*p*-(β-methoxyethyl)phenoxy]-2-propanol, β_1-selective, [241–245]. $C_{34}H_{56}N_2$-O_{12}, M_r 684.8 (free base 267.38), octanol – water partition coefficient 0.5.

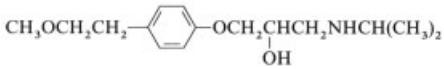

Trade Names. CGP 2175, H.93/26, Beloc, Betaloc, Betaloc SA, Lopressor, Lopresor, Lopresor S.R., Metoros, Prelis, Seloken, Selopral.

Pharmacodynamics. Metoprolol is a competitive antagonist at β_1-receptors without partial agonist activity. It has membrane-stabilizing activity, but only at concentrations higher than those achieved clinically.

Pharmacokinetics. Metoprolol is well absorbed but undergoes extensive, variable, presystemic metabolism in the liver. Systemic bioavailability ranges between $< 5\,\%$ to $> 50\,\%$. The peak concentration is reached in 2 – 3 h, the plasma half-life is ca. 4 h. Most metoprolol is metabolized by the liver, only 10 % is excreted unchanged. About 12 % is bound to plasma protein. Metoprolol has a large volume of distribution (ca. 3 – 6 L/kg).

Uses. Metoprolol has been widely promoted for the treatment of hypertension, angina pectoris, early intervention and long-term use after myocardial infarction, arrhythmias, in the prophylaxis of migraine, and in the management of thyrotoxicosis.

Nadolol [*42200-33-9*], 1-(*tert*-butylamino)-3-[(5,6,7,8-tetrahydro- *cis*-6,7-dihydroxy-1-naphthyl)oxy]-2-propanol, nonselective, SQ 11 725, [246–249]. $C_{17}H_{27}NO_4$, M_r 309.42, *mp* 124 –136 °C, octanol – water partition coefficient 0.066.

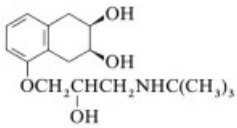

Trade Names. Anabet, Cogard, Solgol.

Pharmacodynamics. Nadolol is a nonselective competitive β-receptor antagonist (i.e., it blocks both the β_1- and β_2-receptors). It does not possess partial agonist or membrane-stabilizing activity.

Phamacokinetics. Nadolol is poorly absorbed (ca. 30 %). There is no presystemic metabolism so that the bioavailability of 30 % is the same as the proportion absorbed. Absorption varies between individuals. The peak concentration occurs 1 – 4 h after oral administration, the plasma half-life is 17 – 24 h. Nadolol is not metabolized, and is excreted in the urine unchanged. It therefore accumulates in patients with renal failure. Nadolol is hydrophilic, about 30 % is bound to plasma proteins. It has a widely varying volume of distribution (ca. 2 L/kg).

Uses. Nadolol has been promoted for the treatment of hypertension, angina pectoris, for migraine prophylaxis, and relief of hyperthyroid symptoms.

Oxprenolol (hydrochloride), [*6452-73-9*], 1-(isopropylamino)-2-hydroxy-3-[*o*-(allyloxy)phenoxy]propane, nonselective, some partial agonist activity, [250–253]. $C_{15}H_{23}NO_3HCl$, M_r 301.8 (free base 265.3), *mp* 107 – 109 °C, octanol – water partition coefficient 1.6.

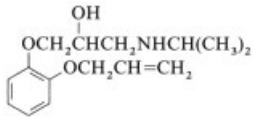

Trade Names. Ba 39 089, Apsolox, Laracor, Paritane, Slow-Pren, Slow-Trasicor, Trasicor, Tracosal, Trasacor.

Pharmacodynamics. Oxprenolol is a nonselective competitive β-receptor antagonist (i.e., it blocks both β_1- and β_2-receptors). It has some partial agonist activity, but less than that of pindolol. Its membrane-stabilizing activity is probably not of clinical significance.

Pharmacokinetics. Oxprenolol is well absorbed but undergoes extensive, variable, presystemic elimination by the liver (ca. 25 – 80 % of the dose). Bioavailability after oral administration is therefore variable and may be as low as 20 %. Absorption is rapid, peak concentrations being reached between 0.5 and 1.5 h, the half-life is ca. 2 h. Oxprenolol is almost entirely metabolized by the liver (ca. 95 %). It is highly bound (70 – 80 %) to plasma. The volume of distribution is 1.6 L/kg.

Uses. Oxprenolol has been promoted for the treatment of hypertension, angina pectoris, certain arrhythmias, hypertrophic obstructive cardiomyopathy, and the control of somatic manifestations of anxiety.

Pindolol [*13523-86-9*], 4-[2-hydroxy-3-(isopropylamino)-propoxy]indole, nonselective, marked partial agonist effect, [254–257].

$C_{14}H_{20}N_2O_2$, M_r 248.32, *mp* 171 – 173 °C, octanol – water partition coefficient 0.41.

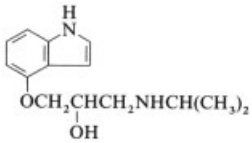

Trade Names. LB 46, Betadren, Betapindolol, Blocklin L, Calvisken, Decreten, Durapindol, Glauco-Viskin, Pectobloc, Pinbetol, Pynastin, Visken.

Pharmacodynamics. Pindolol is a competitive nonselective β-receptor blocking drug (i.e., it blocks both β_1- and β_2-receptors). It has a marked partial agonist activity at the β-receptor.

Pharmacokinetics. Pindolol is completely absorbed after oral administration. There is no presystemic metabolism so that bioavailability approaches 100 %. Peak concentrations are reached 1 – 2 h after oral administration, the plasma half-life is 2.5 – 4 h. Pindolol is 40 % excreted in the urine unchanged, the remainder is metabolized in the liver. About 50 – 70 % of the blood concentration of pindolol is bound to plasma protein. The volume of distribution is 1.2 – 2.0 L/kg.

Uses. Pindolol has been promoted for the treatment of hypertension, angina, and tachyarrhythmias.

Practolol (hydrochloride), [*6673-35-4*], 1-(4-acetamidophenoxy)-3-isopropylamino-2-propanol, β_1-selective, partial agonist activity, [258–260]. $C_{14}H_{22}N_2O_3Cl$, M_r 302.8 (free base 266.34), *mp* 140 – 142 °C, octanol – water partition coefficient 0.06.

Trade Names. I.C.I. 50 172, AY 21 011, Dalzic Eraldin.

Pharmacodynamics. Practolol is a competitive β_1-selective blocking drug. It has strong partial agonist activity at the β_1-receptor. It is no longer used for oral treatment because of severe oculomucocutaneous reactions which may induce irreversible damage of the eyes, fibrotic pleural and peritoneal reactions, and a psoriasiform skin rash.

Pharmacokinetics. Practolol is well absorbed orally, with negligible presystemic metabolism. Bioavailability is >95 %. Peak blood levels are reached 1 – 3 h after oral administration. The half-life is ca. 11.5 h. Practolol is almost entirely excreted unchanged in the urine. Plasma protein binding is negligible. The volume of distribution is 1.6 L/kg.

Uses. Practolol was previously used for hypertension and angina pectoris but this usage was stopped because of the oculomucocutaneous syndrome. The intravenous use for the control of tachyarrhythmia was continued for many years, but this preparation has now also been withdrawn.

Propranolol (hydrochloride), [*318-98-9*], 1-(isopropylamino)-3-(1-naphthyloxy)-2-propanol, nonselective, no partial agonist activity, [261–264]. $C_{16}H_{21}NO_2HCl$, M_r 295.8 (free base 259.3), *mp* 163 – 164 °C, octanol – water partition coefficient 3.65.

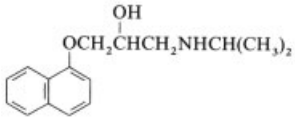

Trade Names. ICI 45 520, AY 64 063, NSC-91 523, Angilol, Beprane, Beta-Tablinen, Caridorol, Dumopranol, Elbrol, Inderex, Kemi, Propahexal, Pylapron, Sloprolol.

Pharmacodynamics. Propranolol is a nonselective β-adrenoceptor blocking drug (i.e., it blocks both β_1- and β_2-receptors). It has no partial agonist activity. It has membrane-stabilizing activity at concentrations which are, however, only rarely achieved with high dosage.

Pharmacokinetics. Propranolol is completely absorbed after oral administration, but because of variable, extensive, first-pass metabolism in the liver bioavailability tends to be low (5 –50 %). There is thus considerable variation in plasma concentration between individuals. Peak concentrations occur 1 – 2 h after oral administration, the plasma half-life is 3.9 h. Propranolol is almost entirely metabolized in the liver, only 1 – 4 % is excreted unchanged in the urine and feces. It is highly bound (80 – 95 %)

to plasma protein. The volume of distribution is high (ca. 3 L/kg).

Uses. Propranolol has been widely evaluated, most of the pioneer work on β-blocking drugs was performed with this drug. It has been promoted for the treatment hypertension, angina pectoris, secondary prevention after myocardial infarction, certain cardiac arrhythmias, catecholamine excess in pheochromocytoma (in combination with α-receptor blockade), somatic manifestations of anxiety, in the management of thyrotoxicosis, tremor of the limbs, migraine prophylaxis, hypertrophic obstructive cardiomyopathy, and to reduce the risk of gastrointestinal hemorrhage in portal hypertension.

Sotalol (hydrochloride), [*959-24-0*], 4′-[1-hydroxy-2-(isopropylamino)ethyl]methanesulfonamide, nonselective, no partial agonist activity, [265,266]. $C_{12}H_{20}N_2O_3SHCl$, M_r 308.8 (free base 272.4), *mp* 207 °C, octanol – water partition coefficient 0.79.

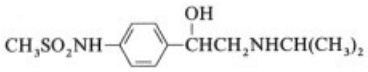

Trade Names. MJ-1999, Beta Cardone, Betapace, Sotacor, Sotalex.

Pharmacodynamics. Sotalol is a nonselective competitive β-adrenergic blocking drug, without partial agonist activity and membrane-stabilizing (local anesthetic) activity. Its β-blocking activity is associated with the L-isomer, but both the D- and L-isomers possess extra antiarrhythmic properties. Sotalol prolongs the duration of the action potential of heart muscle (the so- called Class III effect).

Pharmacokinetics. Sotalol is well absorbed and first-pass metabolism is negligible, bioavailability is therefore 100 %. The peak plasma concentration is reached in 2 – 3 h, the plasma half-life ranges between 7 and 18 h. Sotalol is not metabolized, and is mostly excreted unchanged by the kidney. Plasma protein binding is negligible. The volume of distribution is 1 – 2 L/kg.

Uses. Sotalol is promoted for the treatment of angina pectoris, hypertension, cardiac arrhythmias, the reduction of the recurrence of myocardial infarction, and in the management of thyrotoxicosis.

Timolol (hydrogen maleate), [*26921-17-5*], *S*-(−)-3-(3-*tert*-butylamino-2-hydroxypropoxy)-4-morpholino-1,2,5-thiadiazole, nonselective, no partial agonist activity, [267–270].

$C_{17}H_{28}N_4O_7S$ (hydrogen maleate salt), M_r 432.5 (free base 316.4), *mp* 202 °C, octanol – water partition coefficient 1.16.

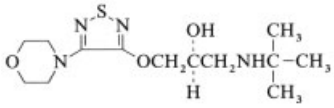

Trade Names. MK 950, Timolol maleate Betim, Blocadren, Proflax, Timacor, Timolate, Timoptic, Timoptol, Timotre, Tenserin.

Pharmacodynamics. Timolol is a competitive antagonist at both β_1- and β_2-receptors. It is without significant partial agonist activity, and has no local anesthetic effect.

Pharmacokinetics. Timolol is almost completely absorbed, but there is considerable first-pass metabolism (50 – 79 %) in the liver, although not as much as with propranolol. Bioavailability is 30 – 50 %. There is a 2 – 7-fold variation in blood levels. The peak concentration occurs 1 – 2 h after oral administration, half is metabolized in the liver but 20 % is excreted unchanged in the urine. Timolol is bound to plasma protein, binding of 10 – 80 % has been reported. The volume of distribution is large (1.74 – 3.64 L/kg).

Uses. Timolol has been promoted for the treatment of hypertension, angina pectoris, migraine prophylaxis, the reduction in the recurrence of myocardial infarction, and as eye drops (Timoptol) in the treatment of glaucoma.

7. References

1. J. W. Black: "Ahlquist and the Development of Beta-Adrenoceptor Antagonists," *Postgrad. Med. J.* **(Suppl. 4), 52** (1976) 11 – 13.
2. R. G. Shanks: "The Discovery of Beta Adrenoceptor Blocking Drugs," in M. J. Parnham, J. Bruinvels (eds): *Discoveries in Pharmacology,* vol. 2. "Haemodynamics, Hormones and Inflammation," Elsevier Sci. Publ. BV, New York 1984, pp. 38 – 72.
3. R. G. Shanks: "The Discovery of Beta-Adrenoceptor Blocking Drugs," *Trends Pharmacol. Sci.* **5** (1984) no. 10, 405 – 409.

4. J. N. Langley: "On the Reaction of Cells and Nerve-Endings to Certain Poisons, Chiefly as Regards the Reaction of Striated Muscles to Nicotine and to Curare," *J. Physiol. London* **33** (1905) 374 – 413.
5. H. H. Dale: "On Some Physiological Actions of Ergot," *J. Physiol. London* **34** (1906) 160 – 206.
6. R. P. Ahlquist: "A Study of Adrenotropic Receptors," *Am. J. Physiol.* **153** (1948) 586 – 600.
7. G. Barger, H. H. Dale: "Chemical Structure and Sympathomimetic Action of Amines," *J. Physiol. London* **41** (1910) 19.
8. W. B. Cannon, A. Rosenblueth: "Studies on Conditions of Activity in Endocrine Organs; Sympathin E and Sympathin I," *Am. J. Physiol.* **104** (1933) 557 – 574.
9. K.-E. Andersson: "Drugs Blocking Adrenoceptors," *Acta. Med. Scand. Suppl.* **665** (1982) 9 – 17.
10. R. J. Lefkowitz: "β-Adrenergic Receptors: Recognition and Regulation," *N. Engl. J. Med.* **295** (1976) 323 – 328.
11. C. E. Powell, I. H. Slater: "Blocking of Inhibitory Adrenergic Receptors by a Dichloro Analogue of Isoproterenol," *J. Pharmacol. Exp. Ther.* **122** (1958) 480 – 488.
12. I. H. Slater, C. E. Powell: "Blockade of Adrenergic Inhibitory Receptor Sites by 1-(3′-4′-Dichloro-Phenol)-2-Isopropylaminoethanol Hydrochloride," *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **16** (1957) 336.
13. N. C. Moran, M. E. Perkins: "Adrenergic Blockade of the Mammalian Heart by a Dichloro Analogue of Isoproterenol," *J. Pharmacol. Exp. Ther.* **124** (1958) 223 – 237.
14. W. E. Glover, A. D. M. Greenfield, R. G. Shanks: "Effect of Dichloroisoprenaline on the Peripheral Vascular Responses to Adrenaline in Man," *Br. J. Pharmacol. Chemother.* **19** (1962) 235 – 244.
15. D. H. Riddell, L. G. Schull, T. F. First, T. D. Baker: "Experience with Pheochromocytoma in 21 Patients: Use of Dichloro Isoproterenol Hydrochloride for Cardiac Arrhythmia," *Ann. Surg.* **157** (1963) 980 –988.
16. J. W. Black, J. S. Stephenson: "Pharmacology of a New Adrenergic Beta-Receptor Blocking Compound (Nethalide)," *Lancet* **2** (1962) no. 7251, 311 – 314.
17. R. M. Fulton, K. G. Green: "Effect of Pronethalol in Angina Pectoris," *Br. Med. J.* **2** (1963) no. 5367, 1228 – 1229.
18. B. Robinson, T. Pilkington: "Effect of Pronethalol in Angina Pectoris," *Br. Med. J.* **2** (1963) no. 5367, 1227 – 1228.
19. B. N. C. Prichard et al.: "Effect of Pronethalol in Angina Pectoris," *Br. Med. J.* **2** (1963) no. 5367, 1226 – 1227.
20. J. P. Stock, N. Dale: "Beta-Adrenergic Receptor Blockade in Cardiac Arrhythmias," *Br. Med. J.* **2** (1963) no. 5367, 1230 – 1233.
21. A. C. Dornhorst, D. R. Laurence: "Use of Pronethalol in Phaecochrome Tumours," *Br. Med. J.* **2** (1963) no. 5367, 1250 – 1251.
22. B. N. C. Prichard: "Hypotensive Action of Pronethalol," *Br. Med. J.* **1** (1964) no. 5392, 1227 –1228.
23. G. E. Paget: "Carcinogenic Action of Pronethalol," *Br. Med. J.* **2** (1963) no. 5367, 1266 – 1267.
24. J. M. Cruickshank, B. N. C. Prichard: *Beta-Blockers in Clinical Practice,* Churchill Livingstone, Edinburgh 1988, pp. 1 – 1003.
25. J. D. Fitzgerald: "Beta Adrenergic Blocking Drugs, Present Position and Future Developments," *Acta Cardiol. Suppl.* **15** (1972) 199 – 216.
26. B. N. C. Prichard: "Beta-Adrenergic Receptor Blockade in Hypertension, Past, Present and Future," *Br. J. Clin. Pharmacol.* **5** (1978) 379 – 399.
27. B. N. C. Prichard, B. Tomlinson, R. J. Walden, P. Bhattacharjee: "The β-Adrenergic Blockade Withdrawal Phenomenon," *J. Cardiovasc. Pharmacol.* **5** (1983) S 56 – S 62.
28. D. A. Richards et al.: "Pharmacological Basis for Antihypertensive Effects of Intravenous Labetalol," *Br. Heart J.* **39** (1977) 99 – 106.
29. B. N. C. Prichard, P. M. S. Gillam: "Assessment of Propranolol in Angina Pectoris, Clinical Dose Response Curve and Effect on Electrocardiogram at Rest and on Exercise," *Br. Heart J.* **33** (1971) 473 –480.
30. D. G. McDevitt: "The Assessment of β-Adrenoceptor Blocking Drugs in Man," *Br. J. Clin. Pharmacol.* **4** (1977) 413.
31. M. Ulrych, E. D. Frohlich, H. P. Dustan, I. H. Page: "Immediate Haemodynamic Effects of Beta-Adrenergic Blockade with Propranolol in Normotensive and Hypertensive Man," *Circulation* **37** (1968) 411 – 416.
32. R. C. Tarazi, H. P. Dustan: "Beta-Adrenergic Blockade in Hypertension," *Am. J. Cardiol.* **29** (1972) 633 – 640.
33. D. Gibson: "Pharmacodynamic Properties of β-Adrenergic Blocking Drugs in Man," *Drugs* **7** (1974) 8 – 38.

34. L. Hansson: "Beta-Adrenergic Blockade in Essential Hypertension, Effects of Propranolol on Haemodynamic Parameters and Plasma Renin Activity," *Acta Med. Scand.* **(Suppl. 1) 55** (1973) 1 – 40.
35. B. Åblad, B. Ljung, R. Sannerstedt: "Haemodynamic Effects of β-Adrenoceptor Blockers in Hypertension," *Drugs* **(Suppl. 1) 11** (1976) 127 – 134.
36. A. J. Man in 't Veld, M. A. D. H. Schalekamp: "How Intrinsic Sympathomimetic Activity Modulates the Haemodynamic Responses to Beta-Adrenoceptor Antagonists. A Clue to the Nature of Their Antihypertensive Mechanism," *Br. J. Clin. Pharmacol.* **(Suppl. 2) 13** (1982) 245 S–257 S.
37. S. H. Taylor, B. Silke, P. S. Lee, A. Hilal: "Haemodynamic Dose Response Effects of Intravenous Beta-Blocking Drugs with Different Ancillary Properties in Patients with Coronary Heart Disease," *Eur. Heart J.* **3** (1982) 564 – 569.
38. B. N. C. Prichard, W. H. Aellig, G. A. Richardson: "The Action of Intravenous Oxprenolol, Practolol, Propranolol and Sotalol on Acute Exercise Tolerance in Angina Pectoris: The Effect on Heart Rate and the Electrocardiogram," *Postgrad. Med. J.* **(Suppl. Nov.) 46** (1970) 77 – 85.
39. T. L. Svendsen, O. Hartling, J. Trap-Jensen: "Immediate Haemodynamic Effects of Propranolol, Practolol, Pindolol, Atenolol and ICI 89 406 in Healthy Volunteers," *Eur. J. Clin. Pharmacol.* **15** (1979) 223 –228.
40. D. G. McDevitt, C. H. Brown, S. G. Carruthers, R. G. Shanks: "Influence of Intrinsic Sympathomimetic Activity and Cardioselectivity on Beta-Adrenoceptor Blockade," *Clin. Pharmacol. Ther.* **21** (1977) no. 5, 556 – 566.
41. A. J. Man in 't Veld, M. A. D. H. Schalekamp: "Effects of 10 Different Beta-Adrenoceptor Antagonists on Haemodynamics, Plasma Renin Activity, and Plasma Norepinephrine in Hypertension: the Key Role of Vascular Resistance Changes in Relation to Partial Agonist Activity," *J. Cardivasc. Pharmacol.* **5** (1983) S 30 – S 45.
42. T. Grandjean, J. L. Rivier: "Cardio-Circulatory Effects of Beta-Adrenergic Blockade in Organic Heart Disease, Comparison Between Propranolol and CIBA 39,089-Ba," *Br. Heart J.* **30** (1968) 50 – 59.
43. F. H. H. Leenen, C. H. M. Coenen, M. Zonderland, A. H. Maas: "Effects of Cardioselective and Non-Cardioselective Beta-Blockade on Dynamic Exercise Performance in Mildly Hypertensive Men," *Clin. Pharmacol. Ther.* **28** (1980) no. 1, 12 – 21.
44. S. C. Morrison et al.: "Selective and Nonselective Beta-Adrenoceptor Blockade in Hypertension: Responses to Changes in Posture, Cold and Exercise," *Circulation* **65** (1982) 1171 – 1177.
45. G. Nyberg: "Effect of Beta-Adrenergic Blockade on Haemodynamic Responses to Dynamic and Isometric Exercise in Angina Pectoris," *Eur. J. Clin. Pharmacol.* **15** (1979) 381.
46. H. J. Waal-Manning: "Atenolol and Three Non-Selective β-Blockers in Hypertension," *Clin. Pharmacol. Ther.* **25** (1979) no. 1, 8 – 18.
47. G. Nyberg: "Blood Pressure and Heart Rate During Sustained Handgrip in Hypertensive Patients Taking Placebo, a Non-Selective β-Blocker (Propranolol) and a Selective β-Blocker (Metoprolol)," *Curr. Ther. Res.* **22** (1977) 828 – 838.
48. R. H. Briant, C. T. Dollery, T. Fenyvesi, C. F. George: "Assessment of Selective Beta-Adrenoceptor Blockade in Man," *Br. J. Pharmacol.* **49** (1973) 106 – 114.
49. B. R. Graham et al.: "Preliminary Observations on the Human Pharmacology of ICI 66 082 in Normal Volunteers," *Br. J. Pharmacol.* **49** (1973) 154 P–155 P.
50. R. J. Walden et al.: "The Effect of Intrinsic Sympathomimetic Activity on Beta Receptor Responsiveness After Beta Adrenoceptor Blockade Withdrawal," *Br. J. Clin. Pharmacol.* **13** (1982) 359 S–364 S.
51. D. McGibney, W. Singleton, B. Silke, S. H. Taylor: "Observations on the Mechanism Underlying the Differences in Exercise and Isoprenaline Tachycardia After Cardioselective and Non-Selective β-Adrenoceptor Antagonists," *Br. J. Clin. Pharmacol.* **15** (1983) 15 – 19.
52. B. Tomlinson et al.: "A Comparison of Penbutolol With Atenolol and Long-Acting Propranolol," *Br. J. Clin. Pharmacol.* **17** (1984) 117 – 182.
53. H.-R. Zerkowski et al.: "Human Myocardial Beta-Adrenoceptors: Demonstration of Both $Beta_1$ and $Beta_2$-Adrenoceptors Mediating Contractile Responses to Beta-Agonists on the Isolated Right Atrium," *Naunyn-Schmiedeberg's Arch. Pharmacol.* **332** (1986) 142 – 147.

54. J. Trap-Jensen, J. E. Carlsen, T. L. Svendsen, N. J Christensen: "Cardiovascular and Adrenergic Effects of Cigarette Smoking During Immediate Non-Selective and Selective Beta Adrenoceptor Blockade in Humans," *Eur. J. Clin. Invest.* **9** (1979) 181 – 183.
55. M. Nc. D. Davidson, R. J. M. Corrall, T. R. D. Shaw, E. B. French: "Observations in Man of Hypoglycaemia During Selective and Non-Selective Beta Blockade," *Scott. Med. J.* **22** (1976) 69 – 72.
56. B. N. C. Prichard, E. J. Ross: "Use of Propranolol in Conjunction With Alpha-Receptor Blocking Drugs in Phaeochromocytoma," *Am. J. Cardiol.* **18** (1966) 394 – 398.
57. S. Freestone, L. E. Ramsay: "Effect of Beta-Blockade on the Pressor Response to Coffee Plus Smoking in Patients With Mild Hypertension," *Drugs* (Suppl. 2) 25 (1983) 141 – 145.
58. A. E. Tattersfield, R. N. Harrison: "Effect of β-Blocker Therapy on Airway Function," *Drugs* **(Suppl. 2) 25** (1983) 227 – 231.
59. P. B. S. Decalmer et al.: "Beta-Blockers and Asthma," *Br. Heart J.* **40** (1978) 184 – 189.
60. D. G. McDevitt: "Pharmacologic Aspects of Cardioselectivity in a Beta-Blocking Drug," *Am. J. Cardiol.* **59** (1987) 10 F–12 F.
61. B. N. C. Prichard, B. Tomlinson: "The Additional Properties of Beta Adrenoceptor Blocking Drugs," *J. Cardiovasc. Pharmacol.* **(Suppl. 4) 8** (1986) S 1 – S 15.
62. M. J. Kendall: "Impact of Beta_1 Selectivity and Intrinsic Sympathomimetic Activity on Potential Unwanted Noncardiovascular Effects of Beta Blockers," *Am. J. Cardiol.* **59** (1987) 44 F–47 F.
63. B. N. C. Prichard: "Pharmacologic Aspects of Intrinsic Sympathomimetic Activity in Beta-Blocking Drugs," *Am. J. Cardiol.* **59** (1987) 13 F–17 F.
64. J. B. Kostis et al.: "Treatment of Angina Pectoris With Pindolol: the Significance of Intrinsic Sympathomimetic Activity of Beta Blockers," *Am. Heart J.* **104** (1982) 495 – 504.
65. S. H. Taylor, B. Silke, P. S. Lee, A. Hilal: "Haemodynamic Dose-Response Effects of Intravenous Beta-Blocking Drugs With Different Ancillary Properties in Patients With Coronary Heart Disease," *Eur. Heart J.* **3** (1982) 564 – 569.
66. A. Svensson, T. Gudbrandsson, R. Sivertsson, L. Hansson: "Haemodynamic Effects of Metoprolol and Pindolol: A Comparison in Hypertensive Patients," *Br. J. Clin. Pharmacol.* **13** (1982) 259 S–267 S.
67. A. J. Man in 't Veld: "Effect of Beta Blockers on Vascular Resistance in Systemic Hypertension," *Am. J. Cardiol.* **50** (1987) 21 F–25 F.
68. P. C. Chang, G. J. Blauw, P. van Brummelen: "Pindolol: a Beta-Blocker With Vasodilating Properties Due to Stimulation of Vascular β_2 Adrenoceptors," *J. Hyperten.* **(Suppl. 2) 1** (1983) 338 – 339.
69. D. G. McDevitt: "β-Adrenoceptor Antagonists and Respiratory Function," *Br. J. Clin. Pharmacol.* **5** (1978) 97 – 99.
70. B. N. C. Prichard: "Bisoprolol: a New Beta-Adrenoceptor Blocking Drug," *Eur. Heart J.* **(Suppl. M) 8** (1987) 121 – 129.
71. W. H. Frishman: "Clinical Significance of Beta_1 Selectivity and Intrinsic Sympathomimetic Activity in a Beta-Adrenergic Blocking Drug," *Am. J. Cardiol.* **59** (1987) 33 F–37 F.
72. R. J. Walden et al.: "Withdrawal Phenomena After Atenolol and Bopindolol: Hormonal Changes in Normal Volunteers," *Br. J. Clin. Pharmacol.* **30** (1990) 547 – 556.
73. T. Lysbo-Svendsen et al: "Effects of Propranolol and Pindolol on Plasma Lignocaine Clearance in Man," *Br. J. Clin. Pharmacol.* **13** (1982) 223 S–226 S.
74. P. Lund-Johansen: "The Pharmacology of Combined α-β-Blockade II," *Drugs* **(Suppl. 2) 28** (1984) 35 – 50.
75. B. N. C. Prichard: "Combined α and β Receptor Inhibition in the Treatment of Hypertension," *Drugs* **(Suppl. 2) 28** (1984) 51 – 68.
76. P. Jaillon et al.: "Beta- and Alpha-Adrenoceptor Antagonism by Medroxalol in Healthy Volunteers: Relationship to Dose and Plasma Concentration," *J. Cardiovasc. Pharmacol.* **4** (1982) 705 – 713.
77. R. C. Dage, H. C. Cheng, J. K. Woodward: "Cardiovascular Properties of Medroxalol, a New Anti-Hypertensive Drug," *J. Cardiovasc. Pharmacol.* **3** (1981) 299 – 315.
78. R. C. Dage, C. P. Hsieth, M. Spedding: "Vasodilation by Medroxalol Mediated by Beta-2 Adrenergic Receptor Stimulation," *J. Cardiovasc. Pharmacol.* **5** (1983) 143 – 150.
79. P. J. Schechter, A. Tanskanen, J. Tuomilehto, J. Koch-Weser: "Treatment of Mild and Moderate Hypertension With Medroxalol, an

Alpha- and Beta-Adrenergic Antagonist," *J. Cardiovasc. Pharmacol.* **4** (1982) 955 – 959.
80. E. Schnurr, L. Widmann, M. Glocke: "Efficacy and Safety of Carvedilol in the Treatment of Hypertension," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 101 – S 107.
81. R. Eggertsen, R. Sivertsson, L. Andrén, L. Hansson: "Acute and Long-Term Haemodynamic Effects of Carvedilol, a Combined Beta-Adrenoceptor Blocking and Precapillary Vasodilating Agent, in Hypertensive Patients," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) 97 – 100.
82. A. G. Dupont et al.: "Effect of Carvedilol on Ambulatory Blood Pressure, Renal Haemodynamics and Cardiac Function in Essential Hypertension," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) 130 – 136.
83. T. Morgan, R. Snowden, L. Butcher: "Effect of Carvedilol and Metoprolol on Blood Pressure, Blood Flow and Vascular Resistance," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 124 – S 129.
84. W. Meyer-Sabellek, K.-L. Schulte, A. Distler, R. Gotzen: "Circadian Antihypertensive Profile of Carvedilol (BM 14 190)," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 119 – S 123.
85. T. Ogihara et al.: "The Effect of Low Dose Carvedilol on Circadian Variation of Blood Pressure in Patients With Essential Hypertension," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 108 –S 112.
86. M. Tanaka, H. Masumura, S. Tanaka, A. Akashi: "Studies on the Antihypertensive Properties of Carvedilol, a Compound with β-Blocking and Vasodilating Effects," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 52 – S 57.
87. U. Abshagen: "A New Molecule With Vasodilating and β-Adrenoceptor Blocking Properties," *J. Cardiovasc. Pharmacol.* **(Suppl. 11) 10** (1987) S 23 – S 32.
88. T. P. Pruss et al.: "Celiprolol: A New Beta Adrenoceptor Antagonist With Novel Ancillary Properties," *J. Cardiovasc. Pharmacol.* **(Suppl. 4) 8** (1986) S 29 – S 32.
89. L. H. Opie: "Qualities of an Ideal Beta-Adrenoreceptor Antagonist and Comparison of Existing Agents with a New Cardioselective Hydrophilic Vasodilator Beta-Adrenoreceptor Antagonist, Celiprolol," *Am. J. Cardiol.* **61** (1988) 8 c – 13 c.
90. P. Lund-Johansen, P. Omvik: "Prizidilol in Essential Hypertension: Long-Term Effects on Plasma Volume, Extracellular Fluid Volume and Central Haemodynamics at Rest and During Exercise," *J. Cardiovasc. Pharmacol.* **4** (1982) 1012 – 1017.
91. D. Deitchman, J. A. LaBudde, R. J. Seidehamel in A. Scriabine (ed.): *Bucindolol, New Drugs Annual: Cardiovascular Dugs,* Raven Press, New York (1983), pp. 1 – 18.
92. P. L. Malini, E. Strocchi, E. Ambrosioni: "Comparison of the Anti-Hypertensive Efficacy of Bucindolol and Propranolol," *J. Hyperten.* **(Suppl. 2) 1** (1983) 348 – 350.
93. D. G. Shand: "Pharmacokinetic Properties of β-Adrenergic Receptor Blocking Drugs," *Drugs* **7** (1974) 73 – 90.
94. P. A. Routledge, D. G. Shand: "Clinical Pharmacokinetics of Propranolol," *Clin. Pharmacokinet.* **4** (1979) 73 – 90.
95. A. Melander, K. Danielson, B. Schersten, E. Wahlin: "Enhancement of the Bioavailability of Propranolol and Metoprolol by Food," *Clin. Pharmacol. Ther.* **22** (1977) no. 1, 108 – 112.
96. B. Åblad et al.: "Pharmacological Effects and Serum Levels of Orally Administered Alprenolol in Man," *Eur. J. Clin. Pharmacol.* **5** (1972) 44 – 52.
97. D. G. Shand, R. E. Rango: "The Disposition of Propranol: I. Elimination During Oral Absorption in Man," *Pharmacology* **7** (1972) 159 – 168.
98. W. Riess et al.: "The Metabolism of β-Blockers in Relation to Their Pharmacokinetic and Pharmacodynamic Behaviour," in: W. Schweizer (ed.): *Beta Blockers – Present Status and Future Prospects,* Ciba, Horsham, UK, pp. 276 – 289.
99. A. S. Nies, D. G. Shand: "Clinical Pharmacology of Propranolol," *Circulation* **52** (1975) 6 – 15.
100. D. G. Shand et al.: "The Disposition of Propranolol VII, The Effects of Saturable Hepatic Tissue Uptake on Drug Clearance by the Perfused Rat Liver," *Drug Metab. Dispos.* **1** (1973) 679 – 686.
101. R. Gugler, W. Herold, H. J. Dengler: "Pharmacokinetics of Pindolol in Man," *Eur. J. Clin. Pharmacol.* **7** (1974) 17 – 24.
102. G. Leopold: "Balanced Pharmacokinetics and Metabolism of Bisoprolol," *J. Cardiovasc. Pharmacol.* (Suppl. 11) 8 (1986) S 16 – S 20.
103. R. Gugler, G. Bodem: "Single and Multiple Dose Pharmacokinetics of Pindolol," *Eur. J. Clin. Pharmacol.* **13** (1978) 13 – 16.

104. J. Sassard et al.: "Pharmacokinetics of Atenolol in Patients With Renal Impairment," *Eur. J. Clin. Pharmacol.* **12** (1977) 175 – 180.
105. S. H. Wan, R. T. Koda, R. F. Maronde: "Pharmacokinetics, Pharmacology of Atenolol and Effect of Renal Disease," *Br. J. Clin. Pharmacol.* **7** (1979) 569 –574.
106. H. C. Brown et al.: "Clinical Pharmacologic Observations on Atenolol, a β-Adrenoceptor Blocker," *Clin. Pharmacol. Ther.* **20** (1976) no. 5, 524 – 534.
107. J. D. Fitzgerald et al.: "Studies on the Pharmacokinetics and Pharmacodynamics of Atenolol in Man," *Eur. J. Clin. Pharmacol.* **13** (1978) 81 – 89.
108. G. Johnsson, C.-G. Regardh: "Clinical Pharmacokinetics of β-Adrenoceptor Blocking Drugs," *Clin. Pharmacokinet.* **1** (1976) 233 – 263.
109. T. Ishizaki et al.: "Pharmacokinetics and Pharmacodynamics in Young Normal and Elderly Hypertensive Subjects: A Study Using Sotalol as a Model Drug," *J. Pharmacol. Exp. Ther.* **212** (1980) 173 – 181.
110. C. T. Dollery, A. F. Junod: "Concentration of (+)-Propranolol in Isolated, Perfused Lungs of Rat," *Br. J. Pharmacol.* **57** (1976) 67 – 71.
111. N.-O. Bodin et al.: "Absorption, Distribution and Excretion of Alprenolol in Man, Dog and Rat," *Acta Pharmacol. Toxicol.* **35** (1974) 261 – 269.
112. C. J. Pepine et al.: "Silent Myocardial Ischaemia: Rationale for Management," *Postgrad. Med.* **(special report) 83** (1988) Feb. 29, 76 – 82.
113. V. Hinstridge, T. M. Speight: "An Overview of Therapeutic Interventions in Myocardial Infarction, Emphasis on Secondary Prevention," *Drugs* **(Suppl. 2) 42** (1991) 8 – 20.
114. S. Yusuf, P. Sleight: "Limitation of Myocardial Infarct Size," *Drugs* **25** (1983) 441 – 450.
115. S. Yusuf et al.: "Beta-Blockade During and After Myocardial Infarction: An Overview of the Randomised Trials," *Prog. Cardiovasc. Dis.* **27** (1985) no. 5, 335 – 371.
116. MIAMI Trial Research Group: "Metoprolol in Acute Myocardial Infarction (MIAMI), A Randomised Placebo-Controlled International Trial," *Eur. Heart J.* **6** (1985) 199 – 226.
117. ISIS-I (First International Study of Infarct Survival) Collaborative Group: "Randomised Trial of Intravenous Atenolol Among 16,027 Cases of Suspected Acute Myocardial Infarction: ISIS-I," *Lancet* **2** (1986) no. 8498 , 57 – 66.
118. J. Herlitz et al.: "The Influence of Early Intervention in Acute Myocardial Infarction on Long-Term Mortality and Morbidity as Assessed in the Goteborg Metroprolol Trial," *Int. J. Cardiol.* **10** (1986) 291 –301.
119. J. Kjekshus: "Comments – Beta-Blockers: Heart Rate Reduction a Mechanism of Benefit," *Eur. Heart J.* (Suppl. A) 6 (1985) 29 – 30.
120. B. N. C. Prichard, C. W. I. Owens: "Clinical Pharmacology of Beta-Adrenoceptor Blocking Drugs," in J. B. Kostis, E. A. De Felice (eds.): *Beta Blockers in the Treatment of Cardiovascular Disease,* Raven Press, New York 1984, pp. 1 – 56.
121. H. J. Waal-Manning: "Comparative Studies on the Hypotensive Effects of Beta-Adrenergic Receptor Blockade," in F. O. Simpson (ed.): *Symposium on Beta-Adrenergic Receptor Blocking Drugs,* Ciba, Auckland 1970, p. 64.
122. B. N. C. Prichard, A. J. Boakes: "The Use of Beta-Adrenergic Blocking Drugs in Hypertension: a Review," *Curr. Med. Res. Opin.* **(Suppl. 5) 4** (1977) 51.
123. F. R. Buhler et al.: "Antihypertensive Beta Blocking Action as Related to Renin and Age: a Pharmacologic Tool to Identify Pathogenic Mechanisms in Essential Hypertension," *Am. J. Cardiol.* **36** (1975) 653 – 669.
124. J. Castenfors, H. Johnsson, L. Oro: "Effect of Alprenolol on Blood Pressure and Plasma Renin Activity in Hypertensive Patients," *Acta Med. Scand.* **193** (1973) 189 – 195.
125. E. L. Bravo, R. C. Tarazi, H. P. Dustan, J. W. Lewis: "Dissociation Between Renin and Arterial Pressure Responses to Beta-Adrenergic Blockade in Human Essential Hypertension," *Circ. Res.* **(6 Supp. I) 36** (1975) 241 – 247.
126. G. Leonetti et al.: "Hypotensive and Renin Suppressing Activities of Propranolol in Hypertensive Patients," *Clin. Sci. Mol. Med.* **48** (1975) 491 – 499.
127. T. O. Morgan et al.: "Beta Adrenergic Receptor Blocking Drugs, Hypertension and Plasma Renin," *Br. J. Clin. Pharmacol.* **2** (1975) 159 – 164.
128. S. N. Anavekar et al.: "The Relationship of Plasma Levels of Pindolol in Hypertensive Patients to Effect on Blood Pressure, Plasma-Renin and Plasma Noradrenaline Levels," *Clin. Exp. Pharmacol. Physiol.* **2** (1975) 203 – 212.

129. A. Verniory, M. Staroukine, M. Telerman, F. Delwiche: "Effect of Sotalol on Arterial Pressure and Renin-Angiotensin System in Hypertensive Patients," in: Snart AG (ed.): *Clinical Advances in Beta-Blocking Therapy,* Sotalol Exerpta Medica, Amsterdam 1974, pp. 36 – 45.
130. G. Simon, W. Kowski, S. Julius: "Anti-Hypertensive and Beta-Adrenoceptor Antagonist Action of Timolol," *J. Clin. Pharmacol. Ther.* **23** (1978) 152 – 157.
131. J. Genest et al.: "β-Adrenergic Blocking Agents and Renin Spectrum," in: A. Scriabine, C. S. Sweet (eds.): *New Anti-Hypertensive Drugs,* Spectrum, New York 1976, pp. 215 – 226.
132. B. G. Hansson, J. F. Dymling, H. Hedeland, U. L. Hulthen: "Long-Term Treatment of Moderate Hypertension With the $Beta_1$-Receptor Blocking Agent Metoprolol," *Eur. J. Clin. Pharmacol.* **11** (1977) 239 –245.
133. J. Wikstrand: "Beta-Blockers and Cardioprotection – Is there any Good News from the Recent Trials," *J. Clin. Pharm. Ther.* **12** (1987) 347 – 350.
134. J. Wikstrand et al.: "Metoprolol Versus Thiazide Diuretics in Hypertension, Morbidity Results from the MAPHY Study," *Hypertension (Dallas)* **17** (1991) 579 – 588.
135. L. Wilhelmsen et al.: "Beta-Blockers Versus Diuretics in Hypertensive Men, Main Results from the HAPPHY Trial," *J. Hypertens.* **5** (1987) 561 – 572.
136. Medical Res. Council Working Party: "MRC Trial of Treatment of Hypertension in Older Adults: Principal Results," *Br. Med. J.* **304** (1992) 405 – 412.
137. A. J. Man in 't Veld, M. A. D. H. Schalekamp: "Pindolol Acts as Beta-Adrenoceptor Agonist in Orthostatic Hypotension: Therapeutic Implications," *Br. Med. J.* **282** (1981) 929 – 931.
138. B. Davis, R. Bannister, C. Mathias, P. Sever: "Pindolol in Postural Hypotension, the Cause for Caution," *Lancet* **2** (1981) no. 8253, 982 – 983.
139. J. Melsen, J. Trap-Jensen: "Xemotolol, a New Selective $Beta_1$-Adrenoceptor Partial Agonist, in the Treatment of Postural Hypotension," *Acta Med. Scand.* **219** (1986) 173 – 177.
140. M. W. Wheat: "Treatment of Dissecting Aneurysms of the Aorta: Current Status," *Prog. Cardiovasc. Dis.* **16** (1973) 87 – 101.
141. R. Doroghazi, E. E. Slater, R. W. DeSanctis: "Medical Therapy for Aortic Dissections," *J. Cardiovasc. Med.* **6** (1981) 187 – 197.
142. J. McFarland et al.: "The Medical Treatment of Dissecting Aortic Aneurysms," *N. Engl. J. Med.* **286** (1972) 115 – 119.
143. A. D. Santos, P. K. Mathew, A. Hilal, W. A. Wallace: "Orthostatic Hypotension: a Commonly Unrecognised Cause of Symptoms in Mitral Valve Prolapse," *Am. J. Med.* **71** (1981) 746 – 750.
144. D. Gibson, E. Sowton: "The Use of Beta-Adrenergic Receptor Blocking Drugs in Dysrhythmias," *Prog. Cardiovasc. Dis.* **12** (1969) 16 – 39.
145. V. Homach et al.: "Electrophysiological Effects of Cardioselective and Non-Cardioselective Beta-Adrenoceptor Blockers With and Without ISA at Rest and During Exercise," *Br. J. Clin. Pharmacol.* **13** (1982) 285 S–293 S.
146. D. J. Coltart, D. G. Gibson, D. G. Shand: "Plasma Propranolol Levels Associated With Suppression of Ventricular Ectopic Beats," *Br. Med. J.* **1** (1971) no. 5747, 490 – 491.
147. D. S. Echt et al.: "Prolongation of the Human Cardiac Monophasic Action Potential by Sotalol," *Am. J. Cardiol.* **50** (1982) 1082 – 1086.
148. J. E. Creamer, A. W. Nathan, A. Shennan, A. J. Camm: "Acute and Chronic Effects of Sotalol and Propranolol on Ventricular Repolarization Using Constant-Rate Pacing," *Am. J. Cardiol.* **57** (1986) 1092 – 1096.
149. B. N. Singh: "Clinical Aspects of the Antiarrhythmic Actions of Beta-Receptor Blocking Drugs, Part 1: Pattern of Response of Common Arrhythmias," *N. Z. Med. J.* **78** (1973) 482 – 486.
150. B. N. Singh, D. E. Jewitt: "Beta-Adrenoreceptor Blocking Drugs in Cardiac Arrhythmias," *Cardiovasc. Drugs* **2** (1977) 119 – 159.
151. J. J. Gallagher et al.: "The Preexcitation Syndromes," *Prog. Cardiovasc. Dis.* **20** (1978) 285 – 327.
152. W. Frishman et al.: "Clinical Pharmacology of the New Beta-Adrenergic Blocking Drugs, Part 5: Pindolol (LB-46) Therapy for Supraventricular Arrhythmia: a Viable Alternative to Propranolol in Patients With Bronchospasm," *Am. Heart J.* **98** (1979) 393 – 398.
153. N. Rehnqvist: "Clinical Experience With Intravenous Metoprolol in Supraventricular

Tachyarrhythmias, A Multicentre Study," *Ann. Clin. Res.* (Suppl. 30) 13 (1981) 68 – 72.
154. C. M. Pratt et al.: "Evaluation of Metoprolol in Suppressing Complex Ventricular Arrhythmias," *Am. J. Cardiol.* **52** (1983) 73 – 78.
155. A. Garson, P. C. Gillette, D. G. McNamara: "Propranolol: the Preferred Palliation for Tetralogy of Fallot," *Am. J. Cardiol.* **47** (1981) 1098 – 1104.
156. M. J. Frank, A. M. Abdulla, M. I. Canedo, R. E. Saylors: "Long-Term Medical Management of Hypertrophic Obstructive Cardiomyopathy," *Am. J. Cardiol.* **42** (1978) 993 – 1001.
157. J. F. Goodwin: "Hypertrophic Diseases of the Myocardium," *Prog. Cardiovasc. Dis.* **16** (1973) 199 –238.
158. J. L. Anderson et al.: "A Randomized Trial of Low-Dose Beta-Blockade Therapy for Idiopathic Dilated Cardiomyopathy," *Am. J. Cardiol.* **55** (1985) 471 – 475.
159. P. Hillon et al.: "Comparison of the Effects of a Cardioselective and a Non-Selective Beta-Blocker on Portal Hypertension in Patients With Cirrhosis," *Hepatology (Baltimore)* **2** (1982) 528 – 531.
160. D. Lebrec et al.: "A Randomized Controlled Study of Propranolol for Prevention of Recurrent Gastrointestinal Bleeding in Patients With Cirrhosis: a Final Report," *Hepatology (Baltimore)* **4** (1984) 355 –358.
161. J. A. Lewis, J. M. Davis, D. Allsop, H. A. Cameron: "Beta-Blockers in Portal Hypertension, an Overview," *Drugs* **(Suppl. 2) 37** (1989) 62 – 69.
162. S. K. Sarin: "Long Term Management of Oesophagal Varices," *Drugs* **(Suppl. 2) 44** (1992) 56 – 59.
163. J. G. Freeman, J. R. Barton, C. O. Record: "Effects of Vasodilators on Portal Pressure in Patients With Portal Hypertension," *Gut* **24** (1983) A 971.
164. E. J. Ross et al.: "Pre-Operative and Operative Management of Patients With Phaeochromocytoma," *Br. Med. J.* **1** (1967) no. 5534, 191 – 198.
165. N. C. Delarue, J. D. Morrow, J. H. Kerr, R. F. Colapinto: "Phaeochromocytoma in the Modern Context," *Can. J. Surg.* **21** (1978) 387 – 395.
166. G. Neil-Dwyer et al.: "Effect of Propranolol and Phentolamine on Myocardial Necrosis After Subarachnoid Haemorrhage," *Br. Med. J.* **2** (1978) no. 6143, 990 – 992.
167. G. Neil-Dwyer, P. Walter, J. M. Cruickshank, C. Stratton: "Beta-Blockade in Subarachnoid Haemorrhage," *Drugs* **(Suppl. 2) 25** (1983) 273 – 277 (Symposium on Beta-Blockers in the 1980s).
168. J. M. Cruickshank et al.: "Reduction of Stress/Catecholamine Induced Cardiac Necrosis by Beta-Selective Blockade," *Lancet* **2** (1987) no. 8559, 585 –589.
169. R. Rimon, R. Kapman, M. Viukari: "Propranolol in the Treatment of Neurocirculatory Asthenia – an Open Pilot Study," *Isr. Ann. Psychiatr. Relat. Discip.* **17** (1979) 144 – 148.
170. T. Hanak: "Efficacy of Alprenolol and Diazepam in the Treatment of Functional Sympathiocotonic Cardiovascular Disorders: a Double-Blind Crossover Study," *Curr. Ther. Res.* **25** (1978) 774 – 785.
171. P. Tyrer: "Current Status of Beta-Blocking Drugs in the Treatment of Anxiety Disorders," *Drugs* **36** (1988) no. 6, 773 – 783.
172. K. E. Andersson, E. Vinge: "Beta-Adrenoceptor Blockers and Calcium Antagonists in the Prophylaxis and Treatment of Migraine," *Drugs* **39** (1990) no. 3, 355 – 373.
173. R. F. Steinert, J. V. Thomas, W. P. Boger: "Long-Term Drift and Continued Efficacy After Multiyear Timolol Therapy," *Arch. Ophthalmol. (Chicago)* **99** (1981) no. 1, 100 – 103.
174. R. P. Wilson, G. L. Spaeth, E. Poryzees: "The Place of Timolol in the Practice of Ophthalmology," *Ophthalmology* (Philadelphia) **87** (1980) 451 – 454.
175. H. M. Beumer: "Adverse Effects of Beta-Adrenergic Receptor Blocking Drugs on Respiratory Function," *Drugs* **7** (1974) 130.
176. H. Astrom: "Comparison of the Effects on Airway Conductance of a New Selective Beta-Adrenergic Blocking Drug, Atenolol, and Propranolol in Asthmatic Subjects," *Scand. J. Respir. Dis.* **56** (1975) 292 –296.
177. H. R. Gribbin, A. D. Mackay, C. J. Baldwin, A. E. Tattersfield: "Bronchial and Cardiac Beta-Adrenoceptor Blockade – a Comparison of Atenolol, Acebutolol and Labetalol," *Br. J. Clin. Pharmacol.* **12** (1981) 61 – 65.
178. E. H. Chester, H. J. Schwartz, G. M. Fleming: "Adverse Effects of Propranolol on Airway Function in Nonasthmatic Chronic Obstructive Lung Disease," *Chest* **79** (1981) 540.
179. A. Ranchod, G. R. Keeton, S. R. Benatar: "The Effect of Beta-Blockers on Ventilatory

Function in Chronic Bronchitis," *S. Afr. Med. J.* **62** (1982) 423.
180. J. M. Patrick, S. B. Pearson: "β-Adrenoceptor Blockade and Ventilation in Man," *Br. J. Clin. Pharmacol.* **10** (1980) 624 – 625.
181. E. A. Amsterdam, R. Gorlin, S. Wolfson: "Evaluation of Long-Term Use of Propranolol in Angina Pectoris," *JAMA J. Am. Med. Assoc.* **210** (1969) 103 – 106.
182. R. M. Norris, N. L. Sammel, E. D. Clarke, P. W. T. Brandt: "Treatment of Acute Myocardial Infarction With Propranolol," *Br. Heart J.* **43** (1980) 617 – 622.
183. S. Yusuf et al.: "Early Intravenous Atenolol in Suspected Acute Myocardial Infarction," *Lancet* **2** (1980) no. 8189, 273 – 276.
184. Beta-Blocker Heart Attack Trial Res. Group: "A Randomised Trial of Propranolol in Patients With Acute Myocardial Infarction," *JAMA J. Am. Med. Assoc.* **247** (1982) 1707 – 1713.
185. Norwegian Multicentre Study Group: "Timolol-Induced Reduction in Mortality and Reinfarction in Patients Surviving Acute Myocardial Infarction," *N. Engl. J. Med.* **304** (1981) 801 – 807.
186. Medical Res. Council Report: "Report of the Medical Research Council Working Party on Mild to Moderate Hypertension: Adverse Reactions to Bendrofluazide and Propranolol for the Treatment of Mild Hypertension," *Lancet* **2** (1981) no. 8246, 539 –542.
187. F. J. Zacharias et al.: "Atenolol in Hypertension: a Study of Long-Term Therapy," *Postgrad. Med. J.* **(Suppl. 3) 53** (1977) 102 – 110.
188. A. J. Marshall, C. J. C. Roberts, D. W. Barritt: "Raynaud's Phenomenon as Side Effect of Beta-Blockers in Hypertension," *Br. Med. J.* **1** (1976) no. 6024, 1498 – 1499.
189. J. M. Cruickshank: "Beta-Blockers, Bradycardia and Adverse Effects," *Acta Ther.* **7** (1981) 309 – 321.
190. J. A. Vale, D. B. Jefferys: "Peripheral Gangrene Complicating Beta Blockade," *Lancet* **1** (1978) no. 8075, 1216.
191. R. Gokal, T. L. Dorman, J. G. G. Ledingham: "Peripheral Skin Necrosis Complicating Beta-Blockade," *Br. Med. J.* **1** (1979) no. 6165, 721 – 722.
192. D. J. Greenblatt, J. Koch-Weser: "Adverse Reaction to Beta Adrenergic Receptor Blocking Drugs, a Report from the Boston Collaborative Drug Surveillance Programme," *Drugs* **7** (1974) 118.
193. T. A. Betts, C. Alford: "Beta-Blocking Drugs and Sleep," *Drugs* **(Suppl. 2) 25** (1983) 268 – 272.
194. J. B. Kostis, R. C. Rosen: "Central Nervous System Effects of Beta-Blockers, a Study With Objective Measures (Abstract I-A)," *Clin. Pharm. Ther.* **39** (1986) 203.
195. A. Westerlung: "Atenolol and Timolol," *N. Engl. J. Med.* **307** (1982) 1343.
196. W. Greil: "Central Nervous System Effects," *Curr. Ther. Res.* **28** (1980) 106 S – 112 S.
197. P. E. Hall, M. J. Kendall, S. R. Smith: "Beta-Blockers and Fatigue," *J. Clin. Hosp. Pharm.* **9** (1984) 283 –291.
198. J. M. Cruickshank: "How Safe are Beta-Blockers?," *Drugs* **(Suppl. 2) 25** (1983) 331 – 340.
199. T. Clausen, J. A. Flatman: "β_2-Adrenoceptors Mediate the Stimulating Effect of Adrenaline on Active Electrogenic Na-K-Transport in Rat Soleus Muscle," *Br. J. Pharmacol.* **68** (1980) 749 – 755.
200. P. Kaiser et al.: "Effect of Beta_1-Selective and Non-Selective Beta-Blockade on Work Capacity and Muscle Metabolism," *Clin. Physiol.* **6** (1986) 197 – 207.
201. R. J. Booth, J. Y. Bullock, J. D. Wilson: "Antinuclear Antibodies During Acebutolol Therapy," *Br. J. Clin. Pharmacol.* **9** (1980) 515 – 517.
202. R. J. Cody et al.: "Development of Antinuclear Antibodies During Acebutolol Therapy," *Clin. Pharmacol. Ther.* **25** (1979) 800 – 805.
203. G. Cocco et al.: "A Review of the Side Effects of β-Adrenoceptor Blocking Drugs on the Skin, Mucosae and Connective Tissue," *Curr. Ther. Res.* **31** (1982) 362 – 378.
204. M. J. Mitchinson: "Some Clinical Aspects of Idiopathic Retroperitoneal Fibrosis," *Br. J. Surg.* **59** (1982) 58 – 60.
205. J. H. Marigold, R. E. Pounder, J. Pemberton, R. P. H. Thompson: "Propranolol, Oxprenolol and Sclerosing Peritonitis," *Br. Med. J.* **284** (1982) 870.
206. B. N. C. Prichard, R. J. Walden: "The Syndrome Associated With the Withdrawal of β-Adrenergic Receptor Blocking Drugs," *Br. J. Clin. Pharmacol.* **(Suppl. 2) 13** (1982) S 337 – S 343.
207. C. Dollery et al. (eds.): *Therapeutic Drugs,* Churchill Livingstone, Edingburgh 1991.
208. S. Budavari (ed.): *Merck Index* 11th ed. (1989).

209. R. G. Main, R. H. Tucker: "Recent Advances in β-Adrenergic Blocking Agents," in G. P. Ellis, G. B. West (eds.): *Progress in Medicinal Chemistry,* vol. 22, Elsevier Sci. Publ., Amsterdam 1985, pp. 121 – 164.
210. May and Baker, ZA 68 08 345, 1969 (Wooldrige and Basil).
211. May and Baker, US 3 857 952, 1974.
212. D. T. Mason (ed.): "Proceedings of the Symposium on Improved β-Adrenergic Blockade Therapy: Results of Acebutolol Trials," *Am. Heart J.* **109** (1985) 1120 – 1223.
213. B. N. Singh, W. R. Thoden, A. Ward: "Acebutolol: a Review of its Pharmacological Properties and Therapeutic Efficacy in Hypertension, Angina Pectoris and Arrhythmia," *Drugs* **29** (1985) 531 – 569.
214. ICI, DE 2 007 751, 1970 (Barrett et al.).
215. ICI, US 3 663 607, 1972; US 3 836 671, 1974.
216. R. C. Heel, R. N. Brogden, T. M. Speight, G. S. Avery: "Atenolol: a Review of its Pharmacological Properties and Therapeutic Efficacy in Angina Pectoris and Hypertension," *Drugs* **17** (1979) 425 – 460.
217. A. N. Wadworth, D. Murdoch, R. N. Brogden: "Atenolol: a Reappraisal of its Pharmacological Properties and Therapeutic Use in Cardiovascular Disorders," *Drugs* **42** (1991) no. 3, 468 – 510.
218. Synthelabo, DE 2 649 605, 1972 (Manoury et al.).
219. Synthelabo, US 4 252 984, 1981.
220. R. Beresford, R. C. Heel: "Betaxolol – a Review of its Pharmacodynamic and Pharmacokinetic Properties and Therapeutic Efficacy in Hypertension," *Drugs* **31** (1986) 6 – 28.
221. M. M. Buckley, K. L. Goa, S. P. Clissold: "Ocular Betaxolol, a Review of its Pharmacological Properties, and Therapeutic Efficacy in Glaucoma and Occular Hypertension," *Drugs* **40** (1990) 75 – 90.
222. E. Merck, BE 859 425, 1978 (Jonas et al.).
223. E. Merck, US 4 171 370, 1979.
224. E. Merck, US 4 258 062, 1981.
225. E. Merck, US 4 522 829, 1985 (Harting, Fuchs).
226. S. G. Lancaster, E. M. Sorkin: "Bisoprolol, a Preliminary Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Hypertension and Angina Pectoris," *Drugs* **36** (1988) 256 – 285.
227. Otsuka, DE 2 302 027, 1973 (Nakagawa et al.).
228. Otsuka, US 3 910 924, 1973.
229. T. D. Giles et al.: "Remission of Mild to Moderate Hypertension After Treatment With Carteolol, a β-Adrenoceptor Blocker With Intrinsic Sympathomimetic Activity," *Arch. Intern. Med.* **148** (1988) 1725 – 1728.
230. P. Crisp, E. M. Sorkin: "Occular Carteolol, a Review of its Pharmacological Properties, and Therapeutic Use in Glaucoma and Ocular Hypertension," *Drug Aging* **2** (1992) 58 – 77.
231. I. P. Merli, J. Levenson, V. Filtti: "Comparative Long Term Vasoactive Effects of Atenolol and Carteolol on the Properties of the Small and Large Arteries of the Upper Extremities in Human Essential Hypertension," *Clin. Pharmacol. Ther.* (St. Louis) 46 (1989) 686 – 692.
232. Boehringer Mannheim, DE 2 815 926, 1979, (Wiedemann et al.).
233. Boehringer Mannheim, US 4 503 067, 1985.
234. R. R. Ruffolo Jr., D. A. Boyle, R. P. Ventui, M. A. Lukas: "Carvedilol (Kredex): a Novel Multiple Action Cardiovascular Agent," *Med. Actual.* **27** (1991) 465 – 495.
235. Chemie Linz, BE 823 411, 1975 (Zölss et al.).
236. Chemie Linz, US 4 034 009, 1977.
237. R. J. Milne, M. M. Buckley: "Celiprolol, an Updated Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy in Cardiovascular Disease," *Drugs* **41** (1991) 941 – 969.
238. Lunts and Collin, Allen and Hamburys, DE 2 032 642, 1971.
239. US 4 012 444, 1977, Allen and Hamburys.
240. K. L. Goa, P. Benfield, E. M. Sorkin: "Labetalol : a Reappraisal of its Pharmacology, Pharmacokinetics and Therapeutic Use in Hypertension and Ischaemic Heart Disease, *Drugs* **37** (1989) 583 – 627.
241. DE 2 106 209, 1971 (Brandstrom et al.) A. B. Hassle.
242. US 3 873 600, 1975, A. B. Hassle.
243. P. Benfield, S. P. Clissold, R. N. Brogden: "Metoprolol, an Updated Review of its Pharmacodynamic and Pharmacokinetic Properties, and Therapeutic Efficacy, in Hypertension, Ischaemic Heart Disease and Related Cardiovascular Disorders," *Drugs* **31** (1986) 376 – 429.
244. R. N. Brogden, R. C. Heel, T. M. Speight, G. S. Avery: "Metoprolol: a Review of its Pharmacological Properties and Therapeutic Efficacy in Hypertension and Angina Pectoris," *Drugs* **14** (1977) 321 –348.
245. J. Koch-Weser: "Metoprolol," *N. Engl. J. Med.* **301** (1979) 698 – 703.

246. Squibb, DE 2 258 995, 1973 (Hauck et al.).
247. Squibb, US 3 935 267, 1976.
248. Squibb, DE 2 421 549, 1974 (Hauck, Sundeen).
249. W. H. Frishman: "Nadolol: a New β-Adrenoceptor Antagonist," *N. Engl. J. Med.* **305** (1981) 678 – 682.
250. Ciba, BE 669 402, 1966.
251. Ciba, GB 1 077 603, 1966.
252. M. J. Kendall, V. A. John: "Oxprenolol: Clinical Pharmacology, Pharmacokinetics, and Pharmacodynamics," *Am. J. Cardiol.* **52** (1983) 27 D–33 D.
253. M. E. Russo, J. O. Covinsky: "Oxprenolol Hydrochloride: Pharmacology, Pharmacokinetics, Adverse Effects and Clinical Efficacy," *Pharmacotherapy (Carlisle, Mass.)* **3** (1983) 68 – 81.
254. Sandoz, CH 469 001, 1969 (Troxler).
255. Sandoz, CH 472 404, 1969.
256. W. H. Frishman: "Pindolol: a New β-Adrenoceptor Antagonist With Partial Agonist Activity," *N. Engl. J. Med.* **308** (1983) 940 – 944.
257. J. G. Riddell, D. W. G. Harron, R. G. Shanks: "Clinical Pharmacokinetics of β-Adrenoceptor Antagonists – an Update," *Clin. Pharmacokinet.* **12** (1987) 305 – 320.
258. ICI, NL 6 512 676, 1966 (Howe, Smith).
259. ICI, US 3 468 387, 1968.
260. A. M. Barrett, J. Carter: "Comparative Chronotropic Activity of β-Adrenoceptive Antagonists," *Br. J. Pharmacol.* **40** (1970) 373 – 381.
261. ICI, BE 640 312, 1964 (Crowther, Smith); BE 640 313, 1964 (Crowther, Smith).
262. ICI, US 3 337 628, 1967; US 3 520 919, 1970.
263. M. E. Conolly, F. Kersting, C. T. Dollery: "The Clinical Pharmacology of Beta-Adrenoceptor Blocking Drugs," *Prog. Cardiovasc. Dis.* **19** (1976) 203 –234.
264. J. D. Fitzgerald: "Propranolol," in A. Scriabine (ed.): *Pharmacology of Antihypertensive Drugs,* Raven Press, New York 1980, pp. 195 – 208.
265. B. N. Singh: "Sotalol: a Review of its Pharmacodynamic and Pharmacokinetic Properties and Therapeutic Use," *Drugs* **34** (1987) 311 – 349.
266. B. N. Singh: "Expanding Clinical Role of Unique Class III Antiarrhythmic Effects of Sotalol," *Am. J. Cardiol.* **65** (1990) 84 A–88 A.
267. DE 1 925 956, 1969 (Weinstock, Frosst).
268. US 3 655 663, 1972 (Frosst).
269. R. N. Brogden, T. M. Speight, G. S. Avery: "Timolol: a Preliminary Report of its Pharmacological Properties and Therapeutic Efficacy in Angina and Hypertension," *Drugs* **9** (1975) 164 – 177.
270. R. C. Heel, R. N. Brogden, T. M. Speight, G. S. Avery: "Timolol: a Review of its Therapeutic Efficacy in the Topical Treatment of Glaucoma," *Drugs* **17** (1979) 38 – 55.

Recording Media → Information Storage Materials
Recycling → Waste
Recycling → Plastics, Recycling

Reduction

Hydrogenation and Dehydrogenation is a separate keyword

Bernhard Kammermeier, Hoechst AG, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Reduction of Individual Classes of Substances 2
3. Electrolysis 2
3.1. Inorganic Processes 2
3.2. Organic Processes 2
4. Individual Reducing Agents 13
4.1. Coal, Petroleum, Natural Gas 13
4.2. Hydrogen 13
4.3. Hydrides 14
4.4. Metal Carbonyls 17
4.5. Metals 17
4.6. Metal Salts 21
4.7. Nonmetallic Inorganic Reducing Agents 22
4.8. Carbon Compounds 24
4.8.1. Carbon Monoxide 24
4.8.2. Formic Acid 24
4.8.3. Carbonyl Compounds 24
4.8.4. Alcohols 25
4.8.5. Alkenes 25
4.8.6. Heterocyclic Compounds 25
5. Special Reduction Methods 25
6. References 27

1. Introduction

The term *reduction* generally denotes a chemical process in which one reactant – the oxidizing agent – accepts electrons. This process can only take place if an electron donor, namely, the reducing agent, is also present; that is, oxidation (donation of electrons) and reduction (acceptance of electrons) always occur in tandem (redox reaction). During reduction a reducing agent is converted into its corresponding oxidizing agent, and vice versa, i.e., the (arbitrary) designation of a redox process as reduction depends exclusively on the synthetic objective.

Historically, reductions – particularly those involved in the extraction of metals from ores – are among the oldest industrial processes, with far-reaching cultural and historical significance (Iron Age, Bronze Age). In addition to the winning of metals from their salts and oxides by means of carbon, carbon monoxide, petroleum, natural gas, electrochemical reduction, inorganic reductants, or hydrogenation, reduction also plays a significant role in the production of low-valence basic chemicals, such as ammonia, silicon and phosphorus compounds, (complex) metal hydrides and carbonyls, hydrogen, and organometallic compounds.

In organic chemistry any change in a functional group is termed reduction if it involves a decrease in oxidation state. The corresponding electron acceptance can proceed by reaction with hydrogen or other electropositive atoms or compounds, as well as elimination of oxygen or other electronegative elements or groups.

If various reducing agents are arranged in the order of their electron-donor properties the so- called electrochemical series is obtained, in which the standard redox potentials provide a measure of the "reductive power" – the electromotive force – of substances relative to a standard hydrogen electrode. The reductive power increases with decreasing standard potential [1–3]. These well-documented numerical values enable predictions to be made of the course and equilibrium position of a reduction, but only in the case of purely thermodynamically controlled reactions (i.e., normally involving inorganic compounds and metals). In organic chemistry, questions of reaction kinetics, selectivity, regiospecificity, and stereospecificity, as well as reaction conditions, are often the subject of detailed experimental investigations.

2. Reduction of Individual Classes of Substances

Among reduction methods, catalytic hydrogenation is extremely important because of its universal applicability [4]; most of the processes

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a22_687

Table 2. Reduction of functional groups with different hydrides [a]

Functional group	Product	$LiBH_4$	$LiBH_4$	$Li(C_2H_5)_3BH$	$NaBH_3CN$	$NaBH_2S_3$	$Li(tert\text{-}C_4H_9O)_3AlH$	$NaH_2Al(OCH_2CH_2OCH_3)_2$	$Na(C_2H_5)_2AlH_2$	$LiAlH_4$	B_2H_6	AlH_3	BH_3NR_3
RCHO	alcohol	+	+	+	+	+	+	+	+	+	+	+	+
R_2CO	alcohol	+	+	+	+	+	+	+	+	+	+	+	+
RCOOH	alcohol	−	−	−	−	+	−	+	+	+	+	−	+
RCOOR	alcohol	−	+	+	−	−	(−)	+	+	+	−	−	+
Lactone	diol	+	+	+	−	+	−	+	+	+	−	−	+
RCOCl	alcohol	+	+	+	−	−	+ [b]	+	+	+	+	+	+
Oxirane	alcohol	−	+	+	−	+	−	+	+	+	−	+	+
$RCONR_2$	amine	−	−	−	−	+	−	+	+	+	+	−	+
RCN	amine	−	−	−	−	+	−	−	+ [b]	+	−	+	+
RNO_2	amine [c]	−	−	−	−	+	−	+	+	+	+	+	+
	azocompound	+	+	+	+	+	+	+	+	+	+	+	+
RHal	hydrocarbon	−	−	−	+	+	−	+	+	−	−	+	−
Olefin	hydrocarbon	−	−	−	−	−	−	−	−	−	+	+	−
Acetylene	hydrocarbon, olefin	−	−	−	−	−	−	−	−	+	+	+	−

[a] + Reduction possible; − reduction not possible.
[b] Reduction to aldehyde.
[c] In some cases reduction takes place only as far as azo compound.

discussed in the following sections can be carried out successfully by catalytic hydrogenation (→ Hydrogenation and Dehydrogenation). Hydrogenation is of further importance due to the rapid development of optically active catalysts, which offer the possibility of carrying out enantioselective syntheses [5–7].

A complete description of all reduction processes and reducing agents is impossible within the scope of this article, given the wealth of material available. The topic of reduction is however discussed in great detail, in both general and selected aspects, in numerous monographs and reviews [8–13]. This article describes the classes of reductive compounds (see Table 1) and deals with various reducing agents.

3. Electrolysis

The importance of electrochemical reduction methods derives from the fact that some reducing agents can be obtained only by means of cathodic reduction. These include, in particular, strongly electropositive metals such as lithium, sodium, potassium, magnesium, calcium, or aluminum [2]. Other reducing agents are often far more expensive or are associated with environmental hazards. Detailed theoretical and experimental information on electrolysis can be found in [3], [70], [188] (see also → Electrochemistry).

3.1. Inorganic Processes

Electrometallurgy occupies a dominant role among the applications of electrolytic reduction processes; either aqueous solutions or melts (electrothermic processes) are used. A distinction can be made between actual winning of a metal (extractive electrolysis) and its subsequent purification (electrolytic refining). In electrolytic refining, impure metals are reoxidized at the anode and migrate as metal ions to the cathode for reductive deposition (electrolytic copper, nickel, silver, and zinc). Electrolytic degreasing is another reductive purification process in which grease is removed by the formation of hydrogen at surface of a metal acting as a cathode and is subsequently saponified in the basic solvent.

Table 1. Commonly used reducing agents for individual compound classes

Compound class	Target product	Reducing agent	Example	Yield, %
1. Hydrocarbons				
Cycloalkane	alkane	Li/NH_3 [14], $Zn/ZnCl_2$ [15]	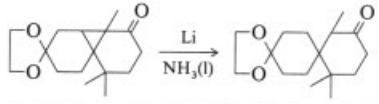	
Alkene	alkane	Na [16], N_2H_2 [17], borane [18]	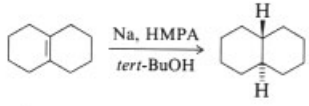	91
	borane	borane [18]	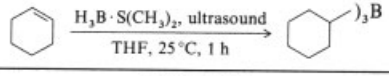	98
Alkyne	alkene	$LiAlH_4$ [19], MgH_2 [20], borane [18]	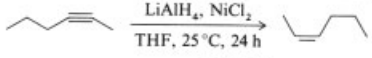	91
	alkane	$LaNi_5H_6$ [21], $KO_2CN{=}NCO_2K$ [22], $LiAlH_4$ [19]	$PhC{\equiv}CC{\equiv}CPh \xrightarrow[THF/MeOH,\ 25\,°C,\ 1\,h]{LaNi_5H_6} Ph(CH_2)_4Ph$	97
Aromatic hydrocarbon	cyclic olefin	Na/NH_3, Li/NH_3 [23], electrolysis [24]	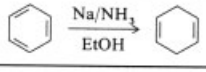	83
2. Halogenated compounds				
Alkyl halide	alkane	$LiEt_3BH$ [25], $LiAlH_4/FeCl_3$ [26]	$H_3C{-}(CH_2)_7{-}Cl \xrightarrow[THF,\ 25\,°C]{LiEt_3BH} H_3C{-}(CH_2)_6{-}CH_3$	91
Dihalide	olefin	Zn [27], Fe–graphite [28]	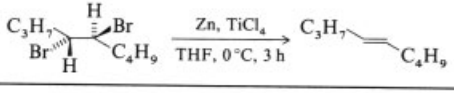	91
Aryl halide	aryl	NaH_2PO_2 [29], i-$C_3H_7OH/h\nu$ [30], $Et_3SiH/Pd/C/h\nu$ [31], $LiAlH_4$ [32]	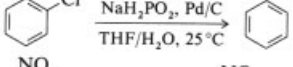	91
	biaryl	Cu [33]	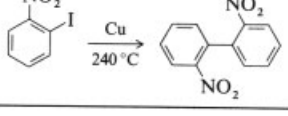	96

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
3. Oxygen-containing compounds				
Alcohol	alkane	HI/HOAc [34], SmI_2/Ac_2O [35]	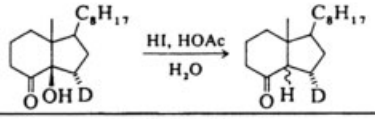	85
Phenol	aromatic hydrocarbon	Li, Na, or K/NH_3 [36], electrolysis [37]	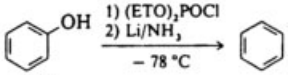	77
	cyclohexanol (cyclohexanone)	electrolysis [38]	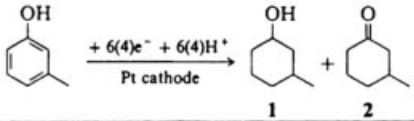	41 (**1**), 50 (**2**)
Peroxide	alcohol	$LiAlH_4$ [39], $NaBH_4$ [40], Na_2SO_3 [41]	$R-O-OH \xrightarrow[H_2O,\ 20\,°C]{NaBH_4} ROH$	
Ozonide	alcohol aldehyde	$BH_3 \cdot S(CH_3)_2$ [42], $LiAlH_4$ [39] $NaBH_4$ [40]	$H_3C-(CH_2)_7-CH=CH_2 \xrightarrow[CH_2Cl_2,\ \text{reflux, 1 h}]{1)\ O_3;\ 2)\ BH_3 \cdot S(CH_3)_2} H_3C-(CH_2)_8-OH$	100
Ether	hydrocarbon	K [43], $LiEt_3BH/Pd(PPh_3)_4$ [44]	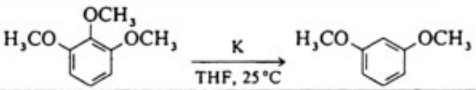	85
Epoxide	alcohol	$LiAlH_4$ [45], NaH [46]	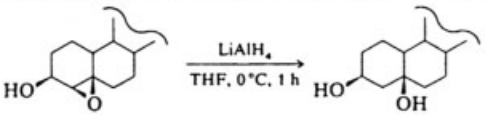	62
	alkene	P_2I_4 [47], $Co_2(CO)_8$ [48]	$C_7H_{15}-\triangle O \xrightarrow[CH_2Cl_2,\ \text{reflux, 3 h}]{P_2I_4} C_7H_{15}-\!\!\!

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
	hydrocarbon	$(PhCO_2)_2BH$ [56], Zn–Hg [57], $BH_3 \cdot THF$ [58]	$H_3C-(CH_2)_8-CHO \xrightarrow[CHCl_3,\ 0\,°C,\ 1\,h]{1)\ TosNHNH_2;\ 2)\ (PhCO_2)_2BH} H_3C-(CH_2)_8-CH_3$	91
	amine	HCO_2H/amine [59], [60]	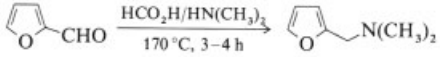	85
	olefin	$TiCl_3$/Zn [61]	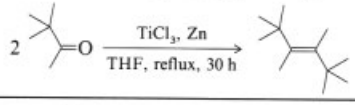	65
Ketone	alcohol	$LiAlH_4$ [62], Li, Na, K, Cs, Ca, Sr, or Ba [63], baker's yeast [51]	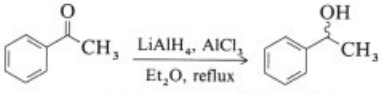	93
	pinacol	electrolysis [64], $TiCl_4$/Zn [53]	$Ph_2CO \xrightarrow[Hg\ cathode,\ CrCl_3]{+\,2e^- + 2H^+} HO-(Ph)_2-(Ph)_2-OH$	70
	amine olefin	HCO_2H/amine [59], [60], [65] $TiCl_4$/Zn [66]	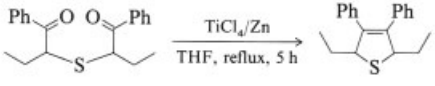	86
	hydrocarbon	N_2H_4 [67], HI/P [68], Zn/Hg [69]	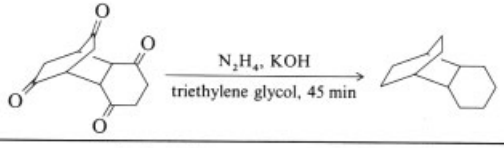	
Sugar	polyol	electrolysis [70], $LiAlH_4$ [71]	$HO-CH_2(CHOH)_4-CHO \xrightarrow[Pb\ cathode,\ 20-30\,°C]{+\,2e^- + 2H^+} HOCH_2-(CHOH)_4-CH_2OH$	
	aldose	$NaBH_4$ [71]	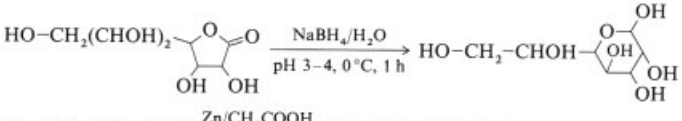	
	ketose	Zn [72]	$HO-CH_2-(CH-OH)_3-CO-CHO \xrightarrow[80-90\,°C,\ 1\,h]{Zn/CH_3COOH} HO-CH_2-(CH-OH)_3-CO-CH_2-OH$	

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
Quinone	hydroquinone	Zn [73], $Na_2S_2O_4$ [74], electrolysis [75], $SnCl_2$ [76]	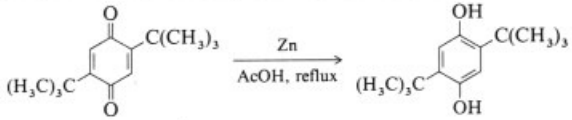	95
	hydrocarbon	HI [77], Zn/pyridine/AcOH [78]	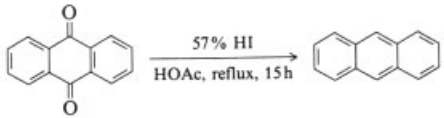	89
Vat dyee	leuco base	$Na_2S_2O_4$, $NaBH_4$ [79], stannite [80]		
4. Carboxylic acids and derivatives				
Carboxylic acid	aldehyde	electrolysis [81], Li [82], $(H_3C)_2NCHO/(CH_3O)_2SO_2$ [83]	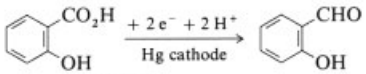	73
	alcohol	$LiAlH_4$ [84], electrolysis [85], AlH_3 [32]	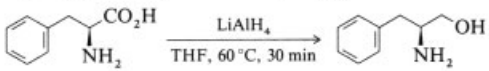	82
Carboxylic ester	aldehyde	$LiAlH_4$, $NaAlH_4$ [86], DIBAH [87]	$C_5H_{11}CO_2CH_3 \xrightarrow[\text{THF, } -60\,^\circ\text{C, 3 h}]{NaAlH_4} C_5H_{11}CHO$	85
	acyloin	$Na/(CH_3)_3SiCl$ [88], [89], Na [90]	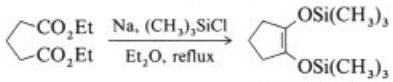	90
	alcohol	$LiBH_4$ [91], $LiAlH_4$ [92], Li, Na [93]	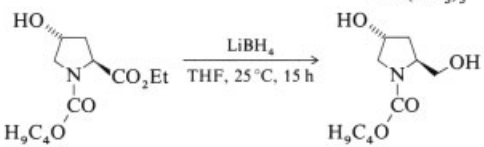	
Carboxylic acid halide	aldehyde	Bu_3SnH [94], $Na_2Fe(CO)_4$ [95]	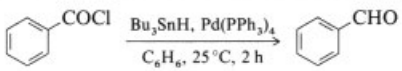	95
	alcohol	$NaBH_4/Al_2O_3$ [96], $NaBH_4$ [40], $LiAlH_4$ [39]	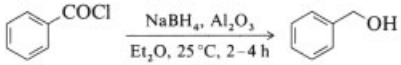	90

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
Carboxylic acid anhydride	lactone	$NaBH_4$ [97], $LiAlH_4$ [81]	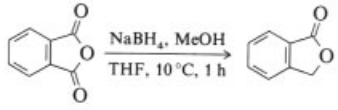	75
	alcohol	$NaBH_4$ [98], catechol borane [97]	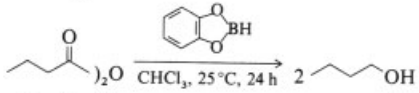	86
	ether	$NaBH_4/BF_3 \cdot OEt_2$ [99], $LiAlH_4$ [39]	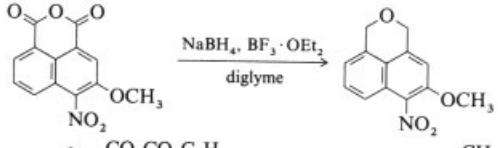	40
	aliphatic	$NaBH_4$ [100]	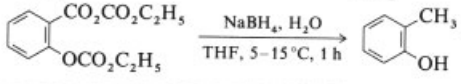	78
5. Nitrogen-containing compounds				
Carboxylic acid amide	aldehyde	$LiAlH_4$ [101], $LiAlH(OEt)_2$ [102], Na/NH_3 [103]	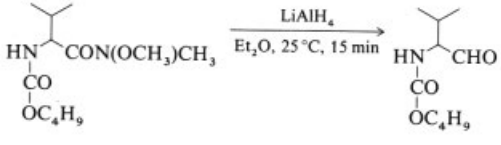	93
	alcohol	$LiEt_3BH$ [104], electrolysis [70]	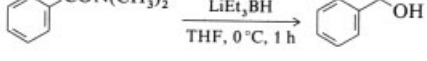	97
	amine	B_2H_6 [105], $Mg(AlH_4)_2$ [106], $NaBH_4$ [107]	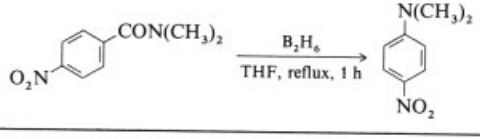	97
Nitrile	aldehyde hydrazone	$LiAlH(OEt)_3$ [108], $SnCl_2$ [109], NaH_2PO_2 [110] N_2H_4 [111]	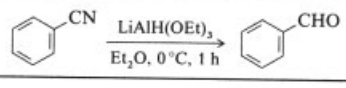	96

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
	amine	$NaBH_4$/Raney Ni [112], $LiAlH_4$ [113]	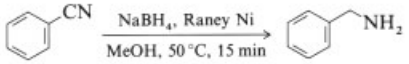	80
Isocyanate	formamide	$LiAlH(tert\text{-}OBu)_3$ [114]	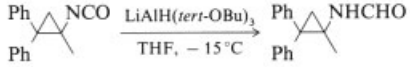	80
	methylamine	$LiAlH_4$ [115], $NaBH_2S_3$ [116]	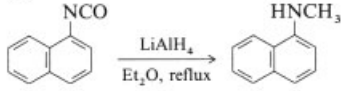	90
Imine	amine	electrolysis [117], $NaBH_4$ [118], Li [119]	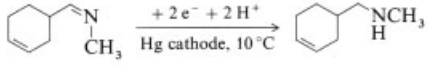	90
Oxime	imine	$TiCl_3$ [120]	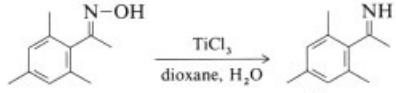	
	hydroxylamine	$NaCNBH_3$, $NaBH_4$ [121], $BH_3 \cdot$ pyridine/HCl [122]	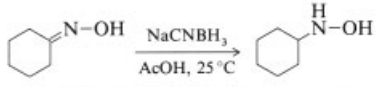	81
	amine	$NaBH_4/ZrCl_4$ [123], $LiAlH_4$ [124]	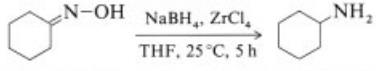	85
***N*-Oxide**	amine, heterocycle	NH_4CO_2H [125], $P(OEt)_3$ [126], electrolysis [127], Ni–Al [128], [129]	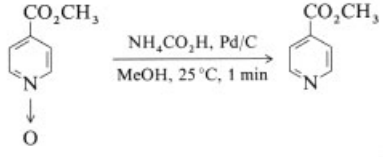	97
Nitrosamine	hydrazine	Zn [130], [131]	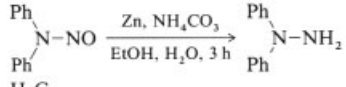	82
	amine	$LiAlH_4$ [132], $SnCl_2$ [133], Ni–Al [129]	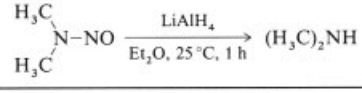	78

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
Nitroalkane	oxime	$Na_2S_2O_3$ [134], $SnCl_2$, $TiCl_3$, $CrCl_2$ [134]	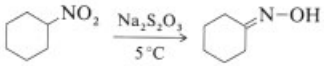	80
	hydroxylamine	Al–Hg [135], electrolysis [136]	$tert\text{-}C_4H_9\text{–}NO_2 \xrightarrow[Et_2O,\ H_2O,\ 1\ h]{Al/H_g,\ reflux} tert\text{-}C_4H_9\text{–}NH(OH)$	75
	amine	N_2H_4/C [137], $LiAlH_4$ [138], electrolysis [139], NH_4CO_2H [140]	$C_2H_5NO_2 \xrightarrow[EtOH,\ reflux,\ 2\ h]{N_2H_4/C} C_2H_5NH_2$	89
Nitro(so)aromatic	azoxy compound	Glucose [141], KOH/EtOH [142], Tl [143], Zn, Mg, Na–Pb [144]	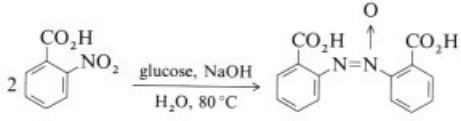	100
	azoxy compound	Zn [145], $LiAlH_4$, Si/NaOH, $NaAlH_2(OCH_2CH_2OCH_3)_2$ [146], $CO/Fe(CO)_5$ [147]	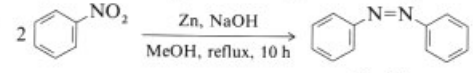	85
	hydrazo compound	Zn [146]	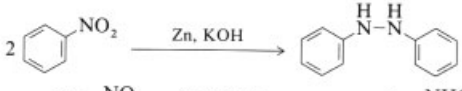	
	hydroxylamine	$NaBH_4$ [148]	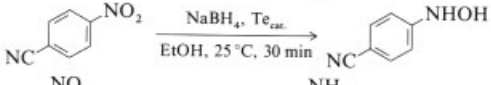	91
	amine	NH_4CO_2H [140], $SnCl_2$ [149], Fe, Zn, H_2S, $Na_2S_2O_4$, H_2SO_3 [150], $CO/Fe(CO)_5$ [147], Ni–Al [129]	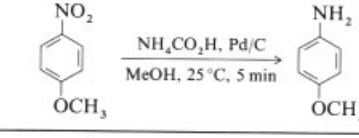	93
Azoxy compound	hydrazo compound	Zn, Na–Hg, H_2S, CH_2O, electrolysis [151], [152], H_2Te [153]	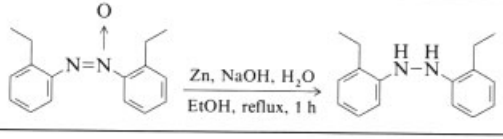	63
Azo compound	hydrazo compound	N_2H_2 [154], Na–Hg, Al–Hg, Zn, H_2S [155], [152], H_2Te [153]	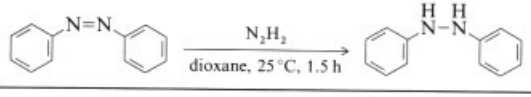	~ 100

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
	amine	cyclohexene/Pd [156], Ni–Al [128], [129], $Na_2S_2O_4$, $(EtO)_2P(S)SH$ [155]	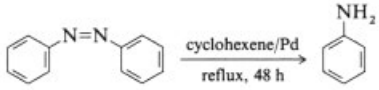	90
Diazonium compound	hydrazine	$Na_2S_2O_4$ [157], $SnCl_2$, Na–Hg, $P(Ph)_3$, PhSeH [158], Zn [159]	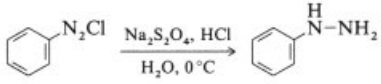	90
	aromatic hydrocarbon	$NaBH_4$ [160], stannite, H_3PO_2, EtOH, hexamethyl phosphoric acid trisamide (HMPA) [159]	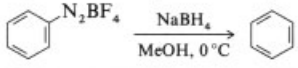	48
Ammonium compound	hydrocarbon	$LiAlH_4$ [161], Na–Hg, electrolysis [162], $NaBH_4$, $NaBH_3CN$ [163]	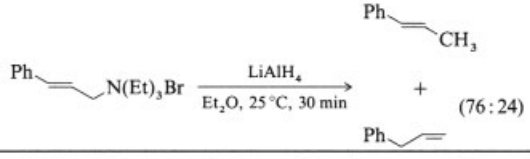	
6. Phosphorus-containing compounds				
Phosphonium compound	hydrocarbon	AlH_3, $LiAlH_4$ [161], NaOH, KOH, electrolysis [164]		90
7. Sulfur-containing compounds				
Sulfonamide	amine	Zn [165], $C_{10}H_8Na$, Na, $SnCl_2$, HI/PH_4I [162]	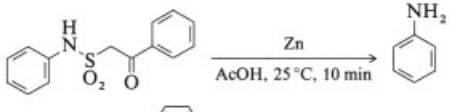	94
	hydrocarbon	$NaBH_4$ [163], Na, $C_{10}H_8Na$ [162]	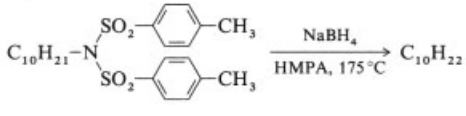	
Sulfonyl chloride	sulfinic acid	$LiAlH_4$, Zn, Na_2S, Na_2SO_3 [162]	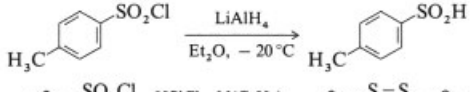	93
	disulfide	$HSiCl_3$ [166], HI [167], N_2H_4 [168]	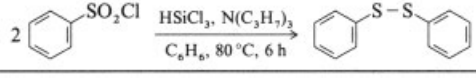	85

Table 1. (Continued)

Compound class	Target product	Reducing agent	Example	Yield, %
	mercaptan	$LiAlH_4$, Zn, P/I_2 [162]	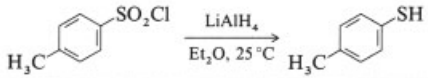	90
Sulfonic acid	aromatic	Na–Hg [169]	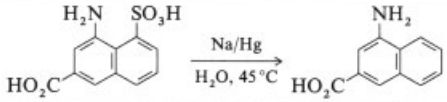	88
Sulfone	sulfoxide	$NaBH_4/Al_2O_3$ [170]	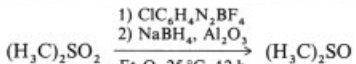	95
	thioether	$LiAlH_4$, Zn [171], DIBAH [172]	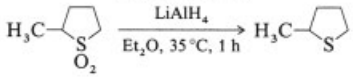	79
	hydrocarbon	$Na_2S_2O_4$ [173], Na–Hg [174], Li [175], Raney Ni [176]	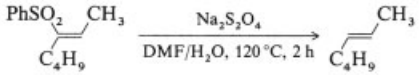	80
Sulfoxide	thioether	$MoOCl_3/Zn$, VCl_2 [177], $SnCl_2$ [178], NaI [179]	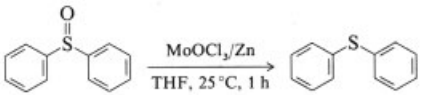	91
	hydrocarbon	Al–Hg [180], Raney Ni [181], Zn [182], Ni_2B [176]	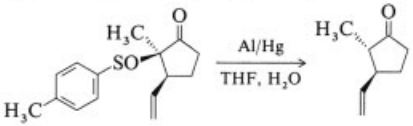	
Disulfide	thiol	N_2H_4 [183], $NaBH_4$ [184], P_2I_4 [185]	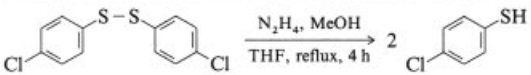	89
	hydrocarbon	Raney Ni, Ni_2B [176]	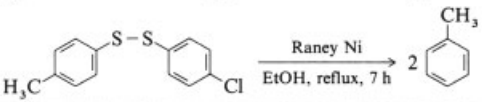	85
Thiol	hydrocarbon	$CO/Co_2(CO)_8$ [186], Raney Ni, Ni_2B [176], $HSnBu_3$ [187]	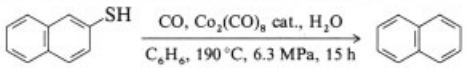	91

Extractive electrolysis is the most important process for producing the strongly electropositive metals of groups 1 – 3 of the periodic table. Metals produced electrothermally include, for example, lithium from LiCl – KCl mixtures, sodium from NaOH or NaCl – $CaCl_2$, magnesium from $MgCl_2$– $CaCl_2$, and aluminum from cryolite melts [2]. In addition, cobalt, cadmium, niobium, and tantalum, among others, are also obtained electrochemically. Hydrogen, which is itself employed as a universal reducing agent, is formed in the industrial production of caustic soda by chlor-alkali electrolysis.

3.2. Organic Processes

Whereas, by definition, oxidative processes such as the dimerization of radical cations or their combination with nucleophilic species, as well as solvolysis, elimination, or rearrangement of carbenium ions, occur at the anode (i.e., the electron acceptor), reactions at the cathode (i.e., the electron donor), involve reductive processes. The great advantage of the cathode as reducing agent is that its potential can be adjusted to any given value, which permits more selective reduction (and also anodic oxidation) than purely chemical methods. Electrical power is normally cheaper than chemicals and greatly reduces pollution. The disadvantages of electrolysis, particularly with organic processes, include possible secondary reactions with the solvent, whose choice is restricted by criteria such as temperature, polarity, and conductivity, as well as the need for special equipment and detailed technical knowledge.

Electrolytic syntheses were first evaluated in detail a hundred years after the discovery of oxidative Kolbe electrolysis [189], [190]. Cathodic reduction was tested for the first time industrially in the hydrodimerizing adiponitrile synthesis [191], [192]. In addition to its use in refunctionalization reactions, such as cathodic hydrogenation, cleavage, and elimination, electrochemical reduction at the cathode is also used for C – C bond formation as a consequence of a reductive umpolung, or for radical pinacol couplings [193]. The removal of protective groups by cathodic cleavage and elimination proceeds smoothly, as does cathodic dehalogenation, reductive carboxylation of aryl halides, or the enantiospecific reduction of appropriate prochiral ketones [193]. The latter reduction proceeds via the two electron mediators methyl viologen and nicotinamide adenine dinucleotide (NADH) from yeast cells, whose reduction equivalents are regenerated continuously at the cathode.

The most important electrochemical process, namely, the hydrodimerizing production of adipic acid dinitrile – with an output of 250 000 t/a – was improved continuously from 1961 to the early 1980s with regard to both space – time yield and economy.

Another industrial process is the reduction of *o*-nitrophenol to the corresponding amine [194]. In addition to these processes, electrochemical reduction plays an important role in the synthesis of valuable, expensive, or otherwise difficultly accessible products in fairly small to moderate amounts.

Electrochemical reactions in the presence of D_2O or T_2O lead to incorporation of deuterium or tritium in appropriate substrates [195], [196].

A detailed review of synthetic aspects of organic electrochemistry can be found in [197]

4. Individual Reducing Agents

The importance of individual reducing agents varies greatly, depending on their respective areas of application. Catalytic processes such as hydrogenation, hydrogenolysis, hydroformylation, reductive amination, or reductive alkylation predominate in large-scale industrial syntheses. The economic, and above all, ecological advantages in the use of hydrogen as an inexpensive reducing agent will become even more important in the future because of increasingly serious environmental problems. The selectivity of catalytic hydrogenation can be extremely fine-tuned by the choice of catalyst metal, carrier, solvent, pressure, and temperature [198]. Developments in the field of asymmetrical catalysis also support this trend [6], [7]. In particular, the latter is consistent with increasingly stringent requirements for the stereochemical uniformity of chemical products,especially pharmaceuticals.

In laboratory chemistry, on the other hand, noncatalytic reduction processes with highly specific and, in some cases, very expensive reactants play an important role. In this case, attention is concentrated on a more product-oriented

approach, and the problem of effluent and waste material is to some extend ignored. Because of the small batch sizes, economic and ecological considerations are not given the fundamental importance that is their due industrially, although here too a gradual change can be foreseen.

Finally, enzymatic processes and thus enzyme-catalyzed reductions have in recent years become much more important in technology as well as in the laboratory. There are many reasons for the major advances in enzymatic reduction. Common objections against the use of enzymes – for example, their restriction to aqueous media, the narrow substrate or product specificity, or lack of thermal stability – have been overcome. Furthermore enzymatic reductions often lead to a substantial simplification of process control, enable synthetic routes to be shortened, or improve regioselectivity or stereoselectivity. The main advantages of biocatalytic process control are the milder (energy- and material-conserving) reaction conditions, extremely high stereoselectivities with a much wider range of substrates, and growing number of commercially available enzymes.

4.1. Coal, Petroleum, Natural Gas

The economic importance of fossil fuels in industrialized countries is associated predominantly with their role as sources of energy. In internal combustion engines and heating systems they represent the reducing agents of choice for the ubiquitous oxygen, although unfortunately only the useful energy of this redox reaction is of interest. Depending on the location and process, various fossil fuels are often interchangeable; however, the area of application of natural gas is expanding for ecological reasons. Coal is an important reducing agent for the production of many metals and nonmetals, such as lead, chromium, nickel, titanium, phosphorus, and hydrogen, as well as carbon monoxide. Generator gas, produced from coal in special furnaces (generators), is a powerful reducing agent:

$$2\,C + \text{air} \longrightarrow 2\,CO + 4\,N_2$$

In the unconsumed part of the heated coke layer, CO and CO_2 are in a strongly temperature- and concentration-dependent redox equilibrium (Boudouard equilibrium), which plays an important part in all industrial processes involving the reduction of oxygen compounds with excess coal. At the temperature prevailing in the blast furnace process a gas mixture termed furnace gas is thus formed.

Carbon monoxide also acts as a reductant in the synthesis of metal carbonyls from metal salts or elemental metals, and opens up a pathway to a large number of valuable organometallic compounds [95], [199]. The carbonyl processes for refining nickel are examples of industrial purification processes involving metal carbonyls (see → Nickel, Chap. 6.3., → Nickel, Chap. 6.4.).

4.2. Hydrogen

About, 77 % of all hydrogen is obtained from petroleum – natural gas, 18 % from coal, and only ca. 4 % electrolytically [200]. Hydrogen is used principally as a chemical raw material, for example, in the synthesis of ammonia and methanol, and in metallurgy. Hydrogen reduces oxides or halides and is used to obtain metals such as tungsten, molybdenum, cobalt, and iron in pure form. Mixtures of CO and H_2 are starting materials for Fischer – Tropsch reactions (C_1 chemistry)

$$x\,CO + y\,H_2 \xrightarrow{\text{catalyst}} \text{CHO products}$$

which can be used together with cheap coal for the synthesis of gasoline and diesel fuel (SASOL Plant, South Africa; see → Coal Liquefaction, Chap. 2.2.6., [201]). Synthesis gas can be obtained by the controlled combustion of petroleum, by reforming natural gas

$$CH_4 + H_2O \rightleftarrows CO + 3\,H_2$$

or by coal gasification

$$C + H_2O \rightleftarrows CO + H_2$$

Finally, hydrogen is required for catalytic hydrogenation, a field that is becoming increasingly important due to the development of asymmetric methods.

4.3. Hydrides

Hydrides (→ Hydrides) can be subdivided into four classes: ionic, metallic, covalent, and complex [1]. Ionic hydrides are formed at moderately high temperature from alkali or alkaline-earth metals and hydrogen. They crystallize in ionic structures of stoichiometric composition and yield hydrogen at the anode on melt electrolysis. The hydride ion that is formed has a standard electrode potential of −2.25 V in the $2H^-/H_2$system and is one of the strongest reduc-ing agents. Transition metals absorb fairly large amounts of hydrogen at elevated temperature and form metallic hydrides of sometimes variable composition (palladium absorbs approximately a thousand times its volume of H_2). Covalent hydrides contain definite bonds between hydrogen and an element of groups 13 – 17. Among the covalent hydrides, boranes, silicon hydrides, tin hydrides, hydrazines, and phosphines are suitable for reduction. Of the complex hydrides, sodium borohydride and lithium aluminum hydride are especially important reducing agents. Like the structurally similar compound lithium gallium hydride, they consist formally of the tetrahedral MH_4^- anion and a cationic counterion. A fine gradation in reactivity can be achieved by replacement of one or several hydride ions in the MH_4^- complex by other ligands. Table 2 illustrates the numerous reduction capabilities of well-known, commercially available complex and covalent hydrides.

Alkali and Alkaline-Earth Hydrides. Metal oxides and metal halides with suitable redox potentials can be reduced to the elementary metals with sodium and calcium hydrides. In the Borol process, NaH serves as reducing agent in the synthesis of sodium borohydride, while lithium aluminum hydride can be prepared from LiH and $AlCl_3$. Lithium hydride reduces carbonyl groups and olefinic double bonds in the presence of transition-metal chlorides [202], whereas NaH in complex mixtures of metal salts and alkoxides is used for the reduction of a number of functional groups [203]. Comparative investigations on LiH, NaH, and KH reveal interesting effects in the reduction of double bonds, depending on experimental conditions [204]. The activity of MgH_2, like that of other saltlike hydrides, depends strongly on the method of preparation [205]. In combination with various alkoxides and dialkylamines, MgH_2 can stereoselectively reduce monocyclic and bicyclic ketones to the corresponding alcohols [206].

Organoboranes, Diborane, and Borane Adducts. Organoboranes and borane adducts, preferably in combination with amines, mercaptans, and ethers, are reducing agents with a wide range of applications in dehalogenation, ether cleavage, deoxygenation, and epoxide ring opening, as well as in the reduction of sulfur and nitro compounds, imines, oximes, aldehydes, ketones, carboxylic acids, acid chlorides, anhydrides, lactones, amides, quinones, alkynes, and olefins [207–213]. (Regio)selective reductions can be achieved with sterically demanding organoboranes [214], such as dicyclohexylborane [92], [215], 9-borabicyclo[3.3.1]nonane [216], diisopinocampheylchloroborane[217], thexylborane [207], borane complexes of sterically hindered tertiary amines [218], [219], or borane pyridine complexes [220]. Catechol borane exhibits lower reactivity [221], [222]. Catechol borane also permits stereoselective reduction of β-hydroxyketones to syn-1,3-diols, whose diastereoselectivities can be increased still further by addition of Wilkinson's catalyst [223]. Very mild deoxygenations can be carried out by bis(benzoyloxy)borane reduction of the corresponding tosylhydrazones [56].

An interesting development is the synthesis of synthetically relevant organoboranes by sonochemical methods [18], [224]. The advantages of ultrasound treatment in the synthesis of various organoboranes from olefins and borohydride adducts include higher yields and substantially reduced reaction times.

Organosilanes. Organosilicon reducing agents [225] have increased greatly in importance, partly due to the fact that they often enable highly stereospecific reactions to be carried out. The use of chirally substituted metal complexes as catalysts permits the enantioselective hydrosilylation of prochiral alkenes, carbonyl compounds, and imines [226], [227]. Tetrabutylammonium fluoride (TBAF) catalyzes the diaste-

reoselective reduction of ketones with diphenylmethylsilane or dimethylphenylsilane [228].

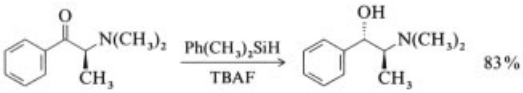

Triethylsilane, sometimes in combination with activators such as protic acids or Lewis acids, reduces lactones, halides, alcohols, nitriles, and diazonium salts [229–231].

Trimethylsilylsilane deoxygenates thiocarbamate derivatives of alcohols in a radical chain reaction[232].

Organotin Hydrides. Among the organostannanes, tributyltin hydride plays a very important role, its principal use being the reduction of organic halides to hydrocarbons via a radical chain mechanism [187], [233–235]. The tin-hydride-reduced intermediates produced by radical generation from halides, alcohols, carbonyl compounds, nitro components, sulfides, and selenides can be used for various consecutive reactions, particularly ring closure [236]. Reduction with tributyltin hydride also makes accessible compounds such as larger ring systems that are otherwise difficult to obtain [237]:

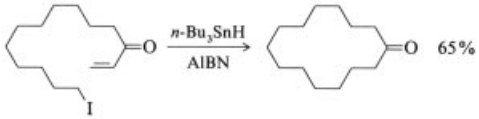

Reductive C – S cleavage of suitably substituted monocyclic azetidinones with tributyltin hydride leads via intramolecular ring closure to carbacepheme bicyclic [238]. Organotin hydrides react, analogously to hydrosilylation, with alkenes and alkynes under hydrostannylation. The mild reaction conditions lead to chemoselective hydrogenation of activated C–C multiple bonds with the preservation of CO and CN double bonds. Recent developments involve the binding of tin hydrides to polymer carriers to produce renewable and easily manipulable reducing agents that can be employed in economical cyclic processes [239].

Complex Borohydrides. Among the borohydrides, sodium borohydride alone or in combination with other metal ions or complex-forming agents is extremely important because of its versatility [40]. With a hydrogen content of 10.2 wt % and a redox potential of −1.24 V, it is an advantageous, specific, and highly selective hydrogenation agent that can also be employed in protic solvents, although with gradual decomposition [240]. In addition to its use as a universal reducing agent, $NaBH_4$ can be used to trap traces of aldehydes, ketones, and peroxides; as a source of pure hydrogen in fuel cells; as an additive in vat dyeing; and for nickel – boron deposition (Nibodur process).

Table 3 illustrates how the reactivity and selectivity of reductions with $NaBH_4$ can be varied by appropriate choice of solvent [40].

In addition to the choice of $Ca(BH_4)_2$, $LiBH_4$, tetraalkylammonium borohydrides, and $Zn(BH_4)_2$ [241–246], combinations of $NaBH_4$ or $LiBH_4$ with $TiCl_4$, $Ti(OR)_4$, $CeCl_4$, Cu_2Cl_2, $CuCl_2$, $CuSO_4$, $ZnCl_2$, ZnI_2, $Ni(CH_3COO)_2$, $ZrCl_4$, $PdCl_2$, $RhCl_3$, Se, I_2, H_2SO_4, CF_3COOH, or $ClSi(CH_3)$ are also suitable for fine adjustment of reactivity [98], [219], [247–259] and for the reduction of nitriles, imines, anhydrides, esters, carboxylic acids, hydrazones, and olefins. A further gradation of the redox potential can be achieved by changing the ligands on boron. The range of reactants extends in this case from stable alkali metal and tetraalkylammonium borohydridoacetates, through the corresponding cyanides, to the highly reactive alkylhydridoborates [25], [44], [260–267].

In enantioselective variants, chiral metal semicorrin or bisoxazoline complexes are added in substoichiometric amounts along with $NaBH_4$ to reduce unsaturated systems with high optical yields [268]. Chiral β-sulfinylenamines are converted diastereoselectively by $NaBH_4$,$Zn(BH_4)_2$, or Bu_4NBH_4 in the presence of excess acetic acid to the corresponding erythroamines [252]. Finally, ketoxime *O*-alkyl ethers can be reduced to amines by $NaBH_4$–$ZrCl_4$ in the presence of chiral 1,2-aminoalcohols with good enantiomeric excess values and high yields [249].

Lithium Aluminum Hydride. Like sodium borohydride, lithium aluminum hydride, which contains 11.5 wt % hydrogen, is one of the most versatile reducing agents. It can reduce virtually all heteroatom – carbon double bonds with the preservation of nonpolarized olefinic double

Table 3. Reduction with $NaBH_4$ in different solvents *

	$NaBH_4$ in various solvents	$NaBH_4$+ LiCl, diglyme/THF	$NaBH_4$+ $AlCl_3$, diglyme	$NaBH_4$+ BF_3, diglyme	$NaBH_4$+ $BF_3 \cdot Et_2O$, THF
Aldehyde	+	+	+	+	+
Ketone	+	+	+	+	+
Acid chloride	+	+	+	+	−
Lactone	+	+	+	+	+
Epoxide	−	+	+	+	+
Ester	−	+	+	±	±
Carboxylic acid	−	−	+	+	+
Nitrile	−	−	+	+	+
Olefin	−	−	+	+	+

* + Reduction possible; − reduction not possible.

bonds [269]. For example, nitro compounds, nitriles, amides, lactams, imines, and oximes can be reduced to amines; carboxylic acids, esters, acid chlorides and anhydrides, ketones, aldehydes, and epoxides can be reduced to alcohols; oxygen – sulfur compounds and disulfides, to sulfides and thioethers; phosphine oxides to phosphines; and hydrazones and hydrazides to hydrazines (see also Table 1). Hydroxylamides of amino acids are converted to labile aminoaldehydes with preservation of the optical activity [101]. On reduction with $LiAlH_4$, various substituted acetylene dicarbinols yield *trans*-configured, conjugated double bonds [270], which permits the synthesis of polyolefinic natural substances [271]. For example, 2,6-dimethylocta-1,7-diene-4-yne affords cosmene (2,6-dimethylocta-1,3,5,7-tetraene).

The reactivity of $LiAlH_4$ can, like $NaBH_4$, be modified considerably by combination with bases or acids. Well-investigated (Lewis) acids include $AlCl_3$ for hydrogenolysis [62], [272]; silver salts [273], [274], and $CeCl_3$ [275] for reductive dehalogenation and reduction of phosphine oxides; LiI for stereoselective reduction of β-alkoxyketones [276]; and hydrogen halides for reductive halogenation of aromatic carbonyl compounds [277]. The addition of amines such as triethanolamine for the reduction of aminoesters and formamidines [278], diethylamine for the synthesis of aldehydes from esters [279], or trialkylamines for the highly selective reduction of acid chlorides, nitriles, and amides in the presence of nitro groups and halides [280] have all proved successful.

In enantioselective variants, 3-ketoesters are reduced to chiral 1,4-diols under the asymmetric induction of a remote group [281]. Early investigations of the mechanism of $LiAlH_4$ reduction of carbonyl groups were carried out by FELKIN et al. [282] and compared with the corresponding Cram models of the relevant transition states, to allow reliable stereochemical predictions to be made. Other mechanistic investigations have dealt with the $LiAlH_4$ reduction of alkenes catalyzed by low-valence titanium and uranium compounds [283], [284].

Other Aluminum Hydrides. In addition to $LiAlH_4$, diisobutylaluminum hydride (DIBAH) is one of the most frequently used complex aluminum hydride reducing agents. It covers nearly the same range of reactions as $LiAlH_4$ but, by virtue of its specific reduction mechanism, can lead to regiospecifically or stereospecifically different products and thus represents an interesting alternative [87]. Other derivatives of $LiAlH_4$ include lithium *n*-butyldiisobutylaluminum hydride for the selective reduction of enones, disulfides, and other functional groups , lithium *tert*-butyldiisobutylaluminum hydride for the stereoselective reduction of cyclic ketones [286], and lithium tris[(3-ethyl-3-pentyl)oxylaluminum hydride for the chemoselective reduction of aldehydes in the presence of ketones [287]. $NaAlH_4$ can be used in combination with $AlCl_3$ and bases [288], whereas $NaAlH_2$–$(OCH_2CH_2OCH_3)_2$ (Red-Al) is used for the release of amines from sulfonamides [289], stereospecific reduction of propargyl alcohols [290], and regioselective epoxide ring opening [291].

Triphenylphosphine Copper(I) Hydride. The stable hexameric compound triphenylphos-

phine copper (I) hydride had been characterized as early as 1971 [292] but was recognized only in 1988 as a valuable stoichiometric reducing agent. Unsaturated ketones, aldehydes, and esters are hydrogenated with good stereoselectivity and excellent chemoselectivity at the activated C – C double bond [293], [294], and alkynes are converted to *cis*-olefins [295]. Catalytic reductions with triphenylphosphine copper chloride under hydrogen pressure interestingly follow a (stereo) chemically nonuniform course [296].

4.4. Metal Carbonyls

The simple and inexpensive methods of preparation of many metal carbonyls from elemental metal and carbon monoxide, by reduction of metal salts in the presence of carbon monoxide, or by oxidation of carbonyl metallates, together with their multiplicity of highly specific reaction capabilities, makes them ideal reducing agents. The formation and decomposition of $Ni(CO)_4$ to obtain high-purity nickel in the Mond nickel process is an industrial example of their use [199]. The inexpensive $Fe(CO)_5$ [297] deoxygenates epoxides to olefins [298] and

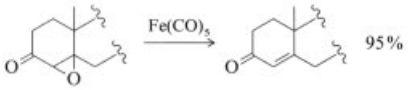

sulfoxides to sulfides [299], hydrogenates α,β-unsaturated carbonyl compounds [300], and decarbonylates aldehydes to hydrocarbons [301]. The carbonyl metallate $Na_2Fe(CO)_4$ (Collman's reagent) obtainable from $Fe(CO)_5$ is also a very selective reducing agent that, in particular, can reduce organic halogenated compounds and perform numerous functional group conversions [95]:

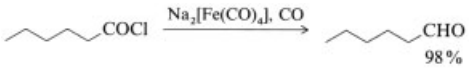

Cobalt carbonyls reduce unsaturated aldehydes [302]; chromium carbonylates hydrogenate α,β-unsaturated carbonyl compounds, imines, and nitriles [303]; molybdenum carbonyl desulfurizes thioethers [304], and dehalogenates α-haloketones, as does the binuclear compound diiron nonacarbonyl [305].

4.5. Metals

Lithium. Lithium can be used as a reducing agent in a number of reactions. In addition to the formation of $LiAlH_4$ from the elements, it is used in the reduction of hydrogen to LiH. The Birch reduction of aromatic ring systems with alkali metals yields partially hydrogenated, nonconjugated olefinic products [306]; the use of lithium instead of sodium suppresses the formation of side products.

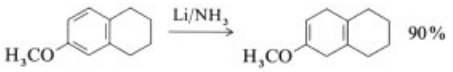

In some cases, nitrile groups can be completely eliminated by lithium [308]. Reductive dehalogenation takes place smoothly [309], as does the formation of organolithium compounds by halogen – metal exchange reactions with halides [310]. Reductive cyclopropane ring opening [14], ether cleavage [310], and ester, alcohol, or thiol reduction [36], [93], [311] also be performed with lithium. The mechanism and stereochemistry of the reduction of cyclic ketones with alkali metals have been investigated in detail [312]. Like other strongly electropositive metals, lithium cleaves the C – S bond in sulfides, sulfoxides, sulfones, and thioketones.

Sodium. Sodium metal is used in the borosilicate process for producing $NaBH_4$ ($\rightarrow$ Hydrides).

$$Na_2B_4O_7 + 7\,SiO_2 + 8\,H_2 + 16\,Na \longrightarrow 4\,NaBH_4 + 7\,Na_2SiO_3$$

Because of its redox potential, sodium can release other metals such as strontium, barium, beryllium, and titanium from their salts. The reduction of SO_2 to sodium dithionite by sodium amalgam is decreasing in importance because of the trend toward mercury-free chlor-alkali electrolysis processes. Sodium amalgam is used for desulfonation [174] and for the synthesis of various carbonyl metallates. Birch reduction by sodium – potassium alloys can be followed by sp^3 C–C bond cleavage [313]. Partial hydrogenation by Birch reduction is one of the

best-known reactions; it involves sodium dissolved in ammonia or amines [306]. In addition to the stepwise reduction of aromatics, especially those with sensitive substituents, conjugated double bonds, triple bonds, or carbonyl groups can also be reduced. Magnetic and spectroscopic investigations indicate that solvated electrons are the actual nucleophilic reagents. An alcohol is often added as a proton source, which is also intended to buffer the solution to suppress undesired isomerization. Carboxylic esters, acid chlorides, and nitriles are reduced in alcoholic solution to the corresponding alcohols and amines. This Bouveault – Blanc reduction is used industrially to synthesize fatty alcohols from glycerides; however, it is increasingly losing its preparative importance due to the use of $LiAlH_4$ [314]. In the absence of an alcohol this reaction leads to an acyloin condensation.

The symmetrical coupling of alkyl halides, known as the Wurtz reaction, to form higher hydrocarbons is now of only minor synthetic importance; in addition to sodium, it can be carried out with lithium, silver, copper, aluminum, zinc, tin, or magnesium [315].

In the synthesis of polyenes, Na(Hg) is used in a modified Whiting reaction [270] for the stereospecific synthesis of *cis*- and *trans*-galbanolenes [(*E, E, Z*)- and (*E, E, E*)-1, 3, 5-undecatriene, respectively) via reductive elimination of dibenzoates [316]. The galbanolenes are important raw materials fo perfumes.

Potassium. Like sodium, potassium is produced by the electrolysis of its molten hydroxide. BERZELIUS first succeeded in obtaining elementary silicon in 1822, by reduction of silicon tetrafluoride with potassium. Because of its extreme reactivity, elemental potassium is rarely used for selective reduction reactions. It can be used as an alloy with sodium or alone for Birch reduction (see above). In the presence of crown ethers, potassium reduces alkyl fluorides and chlorides [317], as well as nitriles [318]. Ethers are reductively cleaved by potassium to give the corresponding hydrocarbons [43]. All higher alkali-metal organyls exhibit this ether cleavage reaction via α-metallation, particularly with cyclic ethers. Finally, nonplanar $4n$ antiaromatics can be converted to planar $(4n+2)$ Hückel aromatics by reduction with potassium [319].

Magnesium. The most important use of magnesium in preparative organic chemistry is the reductive formation of Grignard compounds from organic halides with umpolung of the halogen-bearing carbon atom. Grignard compounds can therefore be used as nucleophiles or as reducing agents.

The Nalco process, which has decreased in importance, combines electrochemical steps with the prior reduction of ethyl chloride by magnesium:

$$2\,C_2H_5MgCl + 2\,C_2H_5Cl + Pb \longrightarrow (C_2H_5)_4Pb + 2\,MgCl_2$$

The Kroll process uses magnesium to prepare titanium from $TiCl_4$; similar reductions yield metals such as zirconium, hafnium, and beryllium. Aldehydes and ketones can be coupled reductively with magnesium to form pinacols. Halides can be converted to the corresponding hydrocarbons via Grignard compounds, which also allow selective access to deuterated compounds [320]. Magnesium also permits the reduction of oximes, nitro compounds, and other functional groups [321] and the chemically selective reduction of conjugated C – C double bonds [322].

In the presence of chlorotrimethylsilane, magnesium leads to reductive cross-coupling of aryl acrylates with aldehydes, giving γ-lactones in moderate to very good yield [323].

Calcium. Calcium is used in metallurgy as a deoxidizing agent for copper, beryllium, and steel (together with silicon) [324], as well as to reduce metals such as zirconium, thorium, vanadium, and uranium. In the manufacture of vacuum tubes, calcium acts as a residual gas getter by forming the corresponding oxides, nitrides, and hydrides. Calcium plays only a minor role in organic synthesis. Like lithium and sodium, it can be used for reduction in liquid ammonia [325]; even difficultly reducible steroid ketones give good results [326]. In the same solvent, calcium also deoxygenates acylated α-hydroxyketones [327].

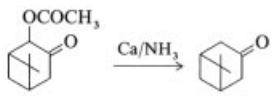

In the presence of ethylenediamine, calcium reduces epoxides to alcohols [328]. Car-

bocyclic and heterocyclic aromatics can be dehalogenated by calcium [329] or converted to Birch-type reduction products [330], [331].

Aluminum is obtained industrially by molten salt electrolysis. It reacts with hydrogen and lithium metal to give $LiAlH_4$. In the Ziegler direct process (→ Aluminum Compounds, Organic, Chap. 1.), aluminum preactivated by titanium is converted with alkenes to give organoaluminum compounds, which in turn serve as valuable reducing agents (e.g., in hydroalumination reactions). In aluminothermic processes (→ Aluminothermic Processes), aluminum liberates elements such as iron, chromium, manganese, titanium, or silicon from their oxides by virtue of its high oxygen affinity. Aluminum amalgam reductively cleaves sulfoxides [180] and leads to pinacol coupling of aldehydes or ketones [332]. In acid solution, aluminum reduces nitro compounds to oximes and amines.

Silicon. In contrast to its organic compounds, silicon is rarely used as a reducing agent, and its principal importance now is as a raw material in the production of silicon chips in microelectronics. Silicon is obtained by reducing silicon dioxide with carbon in a submerged arc furnace and is then purified by zone melting. Like aluminum, it is employed in thermal reduction processes for high-melting metals.

Tin. In organic chemistry, tin has become one of the less common reducing agents. The use of tin in industrial processes is becoming questionable for ecological and economic reasons since it has an unfavorable ratio of reduction capacity per unit mass and the resulting tin salts are toxic.

The conversion of nitro compounds or imines to amines is one of the classic uses of tin reduction [333]. Dinitro compounds can also be converted successfully to diamines [334]. Tin amalgam reduces conjugated enediones smoothly to saturated diketones [335], while mixtures of tin and aluminum assist the regioselective aldol addition of allyl halides to various aldehydes [336–338]. The selectivity of the reductive allylation of aldehydes in the presence of ketones can be enhanced by ultrasound treatment of the tin- containing mixture [339]. Finally, organotin halides are formed by direct reduction of alkyl halides with tin and can be converted directly to organotin hydrides, whose use as specific reducing agents is increasing.

Iron. The Bechamp reduction [340] of aromatic nitro compounds to the corresponding amines is one of the best-known organic reactions involving iron. Systematic investigations of this type of reaction show that it is a useful alternative to catalytic hydrogenation [341] because of the simple reaction control, price, and position of iron in the electrochemical series.

$$Fe \rightleftharpoons Fe^{2+} + 2e^-;\ E_0 = -0.44\ V$$

$$Fe^{2+} \rightleftharpoons Fe^{3+} + e^-;\ E_0 = +0.77\ V$$

The low expenditure on apparatus and equipment is, however, offset by problems arising in the separation and removal of the iron salts formed. The reactivity of iron can be modified by various solvents, impurities in iron, and its active surface. The reduction of nitro compounds can, because of its strong pH dependence, be directed selectively to the azoxy, azo, or hydrazo intermediates.

Iron powder effects the reductive coupling of allyl halides to nonconjugated dienes [342] and plays a major role in the cleavage of S – S bonds and in the reductive dehalogenation of homocyclic and heterocyclic compounds. An interesting reducing agent is iron graphite ($C_{24}Fe$), formed from potassium graphite (C_8K) and $FeCl_3$, and used for the *anti*-debromination of vicinal dibromides [28]. Because it is economical, iron can be used in metallurgy to precipitate more electropositive metals from their salt solutions (cementation). Processes have been developed for precipitating heavy-metal impurities from effluents onto rusty iron.

Zinc is one of the oldest and most widely used reducing agents in organic chemistry, and is the reducing agent in a number of named reactions. In the Clemmensen reduction [343], acid-stable ketones and aldehydes can be converted with zinc amalgam to hydrocarbons.

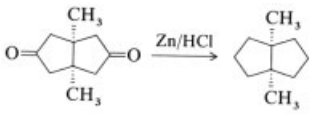

The acid concentration and the choice of solvent are often decisive with regard to reaction course. Other metals give very poor yields in this four-electron transfer.

In the Reformatzky reaction[345],α-halocarboxylic esters undergo reductive condensation with carbonyl compounds in the presence of metallic zinc to give β-hydroxyesters or olefins.

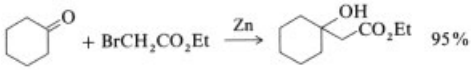

The Simmons – Smith reaction [347] is used to synthesize cyclopropanes by treating olefins with zinc – copper-reduced diiodomethane.

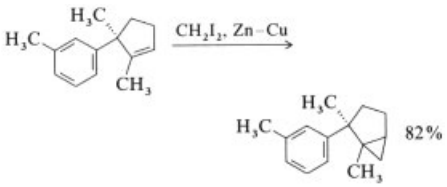

Among a large number of reducing agents, zinc in sodium hydroxide has proved effective in converting nitro compounds to azo components [144]. Furthermore, zinc is suitable for the reduction of nitriles, the reductive cleavage of C – S and C – Hal bonds, the synthesis of *cis*-olefins from alkynes (Rieke zinc [349], [350]), and opening epoxide rings [351]. In the presence of catalytic amounts of aquocobalamin, zinc reduces nitriles smoothly to aldehydes [352]. Metal salts such as $HgCl_2$, CuI, and $NiCl_2$ are often added to activate zinc. A combination of zinc, $NiCl_2$, and ultrasound hydrogenates conjugated unsaturated carbonyl compounds [353–355].

Nickel. A particularly active form of nickel is Raney nickel, a pyrophoric catalyst formed by etching the inactive aluminum component from a nickel – aluminum alloy with alkali. It can be used for catalytic hydrogenation or directly as a reducing agent. The hydrogenative hardening of oils (→ Fats and Fatty Oils, Chap. 8.) in the food industry is the best-known industrial process involving Raney nickel. In organic chemistry, Raney nickel is used mainly as a reductive desulfurizing agent.

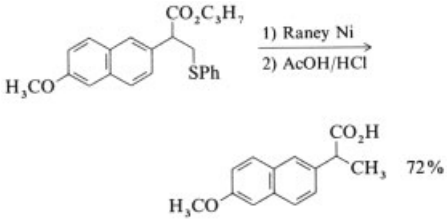

The Wolfrom – Karabinos reduction of carbonyl compounds with Raney nickel via intermediate thioketals or thioacetals is particularly suitable for acid-labile and base-labile compounds [358]. In the synthesis of the non-opiate alkaloid epibatidine, Raney nickel is used to carry out a reductive dehydroxylation at a quaternary carbon center [359]. Nickel – aluminum alloys reduce pyridines, quinolines, isoquinolines [129], [360], and nitroso compounds [361]. Activated nickel dehalogenates *o*-dihaloxylylenes [362]. Nickel – graphite, obtained from $NiBr_2$ and C_8K, reduces alkynes to *cis*-olefins [363].

Copper. In addition to the aforementioned Simmons – Smith reaction, copper is used particularly in reductive dehalogenation. The Ullmann reaction [364] utilizes this property in the synthesis of bisarylenes from corresponding aryl halides [33]. An analogous coupling of halides with phenol derivatives yields bisaryl ethers. The treatment of nitrohalo aromatics with copper powder in molten benzoic acid yields nitro aromatics [365]. Activated copper reacts with epoxyalkyl halides to form the reduced epoxyalkylcopper reagent, which reductively cleaves the epoxy ring with intramolecular cyclization [366].

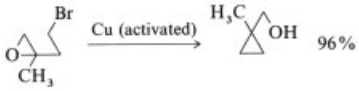

Copper(0) isonitrile complexes allow the synthesis of ring systems from olefins and halides in a Simmons – Smith-type reaction [367], and reduce dibromocyclopropanes to allenes [368].

4.6. Metal Salts

The importance of low-valence metal ions as reducing agents has gradually shifted from stoichiometric reactions to catalytic applications. In this way the selective, fast, and frequently quantitative reductions with low-valence metal ions can be combined with their regeneration by cheaper, less toxic, or more easily manipulable reducing agents in a cyclic process. Often only the combination of redox systems can produce the desired reaction, or particularly reactive species may be formed by in situ reduction. The rapid and quantitative redox reactions of metal salts make them attractive for analytical chemistry (e.g., titanometry) and for inorganic systems [369], as well.

Low-Valence Titanium Ions. The best-known reaction of titanium species in oxidation states 0 to +2 is reductive dimerization of carbonyl compounds to vicinal diols or olefins (McMurry reaction) [53], [370].

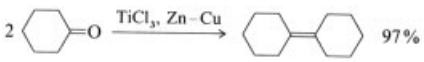

Low-valence titanium is generated in situ from $TiCl_3$ or $TiCl_4$ and K, Mg, Zn–Cu, Zn, or $LiAlH_4$. ESR investigations confirm the occurrence of radical intermediates during the redox reaction. Aldehydes, ketones, esters, and amides are reduced at the carbonyl group to olefins by the Tebbe–Grubbs reagent, prepared from $TiCl_3$–Zn–CH_2Cl_2. Alkynes can be reduced by $TiCl_4$–$LiAlH_4$ to *cis*-alkenes or alkanes [371], while $TiCl_3$–$LiAlH_4$ dehalogenates vicinal dihaldes and bromohydrins [372], [373]. 1,3-Diols are reduced to cyclopropanes, oxiranes are deoxygenated, thioketols are desulfurized, and N–O bonds are reductively cleaved [374].

The reduction of acetals with $TiCl_4$/$LiAlH_4$ leads to pinacol coupling and formation of ethers and alkenes [375].

Vanadium(II) and Niobium(III) Salts. In addition to the reductive removal of hydroxyl groups from acyloin derivatives [376], the transformation of keto groups to quaternary carbon centers [377], and the reductive cyclisation of enals and ynals [378], Vanadium(II) salts have a major potential for the reductive pinacol coupling of aldehydes and ketones [379–381]. This method is also highly suitable for the diastereoselective reductive coupling of peptide alcohols to give C_2 symmetrical HIV protease inhibitors [382]. Niobium(III) salts are also suitable for pinacol coupling, albeit with lower yields [383–386]. They can be used for reductive coupling of imines with aldehydes to give 2-amino alcohols [387]. The related homocoupling of nitriles or silylated imines to afford vicinal diamines is catalyzed by niobium(IV) [388]. Mixtures of $NbCl_5$ or $TaCl_5$ with zinc reduce alkynes to (*Z*)-alkenes [389].

Chromium(II) Salts. Chromium(II) compounds reduce many alkynes stereospecifically to *cis*-olefins, hydrogenate conjugated double bonds, act as dehalogenating agents, convert geminal dibromocyclopropanes and vicinal dihalo alkenes to allenes, and reduce nitro and diazo compounds, as well as epoxides [374].

Iron(II) Salts. The Fe^{2+}/Fe^{3+} redox potential is strongly pH dependent and becomes more positive in acid solution but more negative in alkaline solution. Hence, sensitive nitro compounds can be reduced in basic solution to amines [390].

Mixtures of $FeCl_3$, hydrides, and alkoxides can reduce halides, ketones, alkenes, and alkynes [391]. Organic nitriles can be reductively decyanated with $Fe(acac)_3$/Na [392].

Lanthanide Salts. In the past few years the numerous reduction possibilities of various lanthanide reagents, particularly SmI_2, have been studied in detail [55], [393]. The reactions range from the reduction of unsaturated systems with ytterbium in liquid ammonia to $NaBH_4$ or $LiAlH_4$ reduction of carbonyl-containing substrates, other oxygen-containing compounds, or halides, mediated by $LaCl_3$, $ErCl_3$, SmI_3, and in particular, $CeCl_3$. The increasing area of application of samarium(II) salts is particularly important with regard to a stereochemically uniform reaction [396–398]. In addition to the reduction of nitrogen-containing functional groups, and deoxygenation or dehalogenation reactions, it is used increasingly for Barbier-, Wurtz-, and pinacol-type C–C coupling. Stereocontrolled intramolecular variants of these coupling reactions provide a synthetic route to complex cyclic natural products [399]:

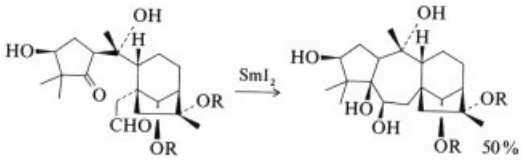

Mixtures of ethylmagnesium bromide and samarium(III) or ytterbium(III) triflate lead to reductive pinacol couplings of aldehydes and ketones that are even more efficient than with SmI_2 [400], [401].

4.7. Nonmetallic Inorganic Reducing Agents

Sulfides, Selenides, and Tellurides [9], [144]. Hydrogen sulfide and the metal sulfides are among the oldest known reducing agents [402], which includes their use in metallurgy and as analytical reagents [403]. They reduce quinones to hydroquinones and azides to amines, as well as reducing sulfur compounds in higher oxidation states, such as disulfides, sulfonyl chlorides, and thiocyanates. Sodium sulfide reduces aromatic or aliphatic nitro groups selectively depending on the amount of water added [404] and can be used for the isomerically pure reduction of aromatic nitro compounds and anilines [405]. Thiols dehalogenate α-haloketones and deoxygenate aromatic ethers [9]. Nitroaromatic compounds are also reduced by ammonium sulfide (Zinin reduction) [406], low-valence selenium compounds [407], H_2Te and its salts, and various organotellurides [9].

Sulfur Dioxide and Sulfites. Sulfur dioxide or sulfites [406] reduce quinones to hydroquinones, ozonides to the bisulfite adducts of the corresponding aldehydes, hydroperoxides to alcohols, diazonium salts to hydrazines, and sulfonyl chlorides to sulfinic acids. In the Piria reduction [408], nitroaromatic compounds are converted to amines and aminosulfonic acids.

Dithionite. Sodium dithionite, $Na_2S_2O_4$, reduces diaryl disulfides to thiophenols, diazonium salts to arylhydrazines, azo components and nitrosoaromatic compounds to amines, quinones to hydroquinones, aliphatic nitro compounds to oximes, and aromatic nitro compounds to amines [406]. Unsaturated ketones are regiospecifically converted by $Na_2S_2O_4$ to saturated carbonyl compounds [12]. Excess rongalite ($HOCH_2SO_2Na \cdot 2\ H_2O$) reduces tellurium in basic solution to the reducing agent Na_2Te [409]. In analysis, dithionite is used as an absorption agent for oxygen. Dithionite and rongalite areused industrially in vat dyeing and as bleaching agents in the paper industry.

Sodium Thiosulfate. $Na_2S_2O_3$, which is prepared from sodium sulfite and sulfur, is used in photography as a complex-forming agent for unreacted silver halide. In bleaching it is used as a reducing antichlorine agent to remove chlorine residues from chlorine-bleached fabric. In iodometry, oxidation equivalents are determined by titration with $Na_2S_2O_3$.

Hydrazine. Hydrazine decomposes in the presence of palladium, platinum, or nickel catalysts into its elements or ammonia, which makes it a convenient source of hydrogen for hydrogenations [410]. The amount of hydrogen released increases with increasing pH of the solvent. Hydrazine can reduce some metal ions to the elements.

Hydrazine reduces ketoximes and nitroaromatic compounds to the corresponding amines, nitriles to hydrazonimines, sulfonyl chlorides to disulfides, and halides to hydrocarbons [8], [9], [410]. Its most important role, however, is in the conversion of carbonyl to methylene groups (Wolff – Kishner reduction).

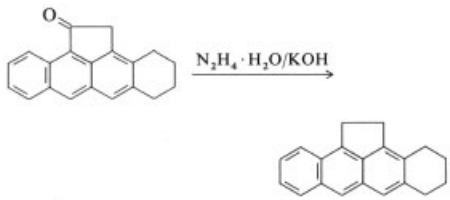

In the Huang – Minlon variant of this reaction the intermediate hydrazone is heated in situ in a high-boiling solvent and converted to the methylene compound. Depending on the stability of the hydrazones, various solvents and bases can be used for their decomposition [9]. In terms of batch scale-up, this variant is often superior to the Clemmenson reduction. The base-catalyzed tosylhydrazone cleavage of carbonyl compounds, known as the Bamford – Stevens

reduction, may also be used for the reductive preparation of olefins from the latter [412].

Diimine. The synthetic potential of N_2H_2 [413] was recognized relatively late, although examples of the reduction of olefinic double bonds in the presence of hydrazine were already known. Diimine, obtained by oxidation of hydrazine with oxygen or H_2O_2, acid hydrolysis of azodi-formate, or decomposition of various sulfonyl hydrazides, reduces C – C double and triple bonds, as well as aldehydes, ketones, and imines. Diimine formed in situ is also suitable for the reduction of highly sensitive endoperoxide olefins [414]:

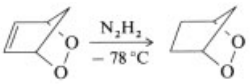

Alkynes are reduced to *cis*-alkenes or alkanes depending on the substituents. Since diimine reduction proceeds under very mild conditions, it preserves a number of functional groups that are attacked during catalytic hydrogenation or by other reducing agents. Allyl halides, esters, and disulfides are thus retained, as are the labile N–O and N–N–O groups [413].

Phosphorus, Phosphines, and Phosphites. Phosphorus is often used in combination with I_2 or HI as a reducing agent. Nitroaromatic compounds are reduced in basic solution by white phosphorus to the azoxy components or amines [415]. Mixtures of red phosphorus and I_2 act as deoxygenating agents and can therefore reductively open lactone rings [416]. Red phosphorus in hydroiodic acid reduces azlactones, dehalogenates aromatic compounds, and deoxygenates hydroxy fatty acids [333]. P_2I_4 reduces dithioacetals to thioethers [417]. Triphenylphosphine [418], because of its high affinity for oxygen and sulfur, is a universal reducing agent. It is a dehalogenating and deoxygenating agent; reduces diaryl sulfides to thiols; and desulfurizes sulfenyl chlorides to the corresponding alkyl chlorides, and sulfur- containing heterocyclics to olefins. Azides are converted by hydrolysis of the intermediate iminophosphines (Staudinger reaction) into amines. In combination with iodine, triphenylphosphine reduces sulfoxides to thioethers and aromatic sulfonic acids to thiols. A combination of triphenylphosphine, iodine, and imidazole reduces vicinal diols to olefins. In the Wittig and Wittig – Horner reactions, triphenylphosphine and other phosphine or phosphite analogues are formally used as reducing agents in the synthesis of olefins from carbonyl compounds and phosphorus ylids.

Trialkylphosphites have a broad range of applications similar to those of triphenylphosphine, and to some extent they complement the properties of the latter. The most commonly used member of this class, triethylphosphite [418], is the classic reagent for desulfurizing cyclic thiocarbonates in the Corey – Winter method for obtaining olefins from vicinal diols. Triethylphosphite is used for reductive photochemical sulfur extrusion in paracyclophane syntheses [419]; it reduces halogenated compounds, hydroperoxides, and oxiranes, as well as nitro and nitroso compounds.

Hypophosphoric Acid. In the presence of Raney nickel, H_3PO_2 reduces nitriles to aldehydes [110], while with Pd – C catalysis, olefinic double bonds are hydrogenated and organic halides are reduced [420]. Aromatic amines can be reduced via their aryldiazonium salts to aromatic hydrocarbons [9].

Hydroiodic Acid is used in combination with phosphorus to reduce a wide range of functional groups (see phosphorus,page 23). It also reduces vicinal diketones and α-hydroxyketones to monoketones, and aromatic sulfonyl chlorides to diarylsulfides [157]; it hydrogenolytically cleaves halogen and silyl groups [333]. Substituted dihydroxyaromatic compounds are reduced to hydrocarbons [421]. Sodium iodide in the presence of various activators reduces α-haloketones to ketones [9].

4.8. Carbon Compounds

4.8.1. Carbon Monoxide

Carbon monoxide owes its outstanding industrial importance as a reducing agent to processes for winning iron and some other metals (see page 19). Furthermore, CO provides access to metal carbonyls, which can be used themselves as reducing agents or as hydrogenation catalysts (see Section 4.4).

Carbon monoxide can be reductively inserted into carbon – halogen bonds [422]. In the presence of various organometallic catalysts, CO reduces nitroaromatic compounds to amines; as a gas mixture with H_2S it represents a cheap alternative to hydrogen for converting dinitroaromatic compounds to diamines [423]. Reductive cyclization of *o*-nitrobenzaldehyde oximes with CO allows straight forward access to 2*H*-indazole derivatives [424]:

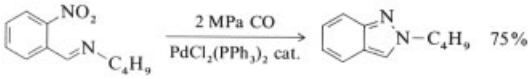

Catalytic reduction of nitroaromatic compounds with CO can also be directed to the isocyanate stage, which permits phosgene-free synthesis of these substances [425].

4.8.2. Formic Acid

In the presence of catalysts, formic acid reduces carboxylic acids and aromatic nitriles to aldehydes [157], [426], [427]. Enamines are reduced in good yields to amines, and activated C – C double bonds are reduced to saturated systems [418]. In the Leuckart – Wallach reaction, formic acid, its amide, or ammonium salt acts as a hydride donor in the reductive amination of carbonyl groups [428]. The special case known as the Eschweiler – Clark reaction uses formaldehyde as the carbonyl component and yields dimethylamines.

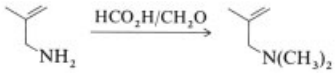

This reaction can also be used to synthesize nitrogen-containing ring systems from acyclic precursors [429]. Sodium formate reductively dehalogenates allyl chlorides to the corresponding olefins in a phase-transfer reaction with a palladium catalyst [430].

4.8.3. Carbonyl Compounds

Aldehydes generally exhibit a marked tendency to oxidize and form the more stable carboxylic acids. In the case of aldehydes that do not contain hydrogen on the α-carbon atom, disproportionation occurs to give the corresponding alcohol and carboxylic acid (Cannizzaro reaction) [52]. This base-catalyzed reaction can occur intermolecularly as well as intramolecularly, and in its crossed version using two different aldehydes leads, at most, to two redox pairs. α-Ketoaldehydes undergo intramolecular disproportionation to give α-hydroxycarboxylic acids.

Formaldehyde used in excess as reactant reduces the other aldehyde to alcohol and is itself converted to formic acid. With regard to industrial application of the Cannizzaro reaction, a wide range of catalysts have been tested so that enolizable aldehydes can also be used under neutral reaction conditions [431]. Formaldehyde is suitable for a number of other reductions, including the preparation of colloidal gold solutions and silver mirrors from the metal salts. It converts nitroaromatic compounds under hydrogenation conditions to alkylated amines [432]. Reductones, which in their tautomeric form also occur as enediols, represent another class of strongly reducing substances [433]. Biologically important members of these acyloins include, for example, the antioxidant vitamin C, which in biological systems is involved in the reductive removal of oxygen radicals.

4.8.4. Alcohols

In the classic Meerwein – Ponndorf – Verley process aluminum isopropoxide acts as a hydride donor in the reduction of ketones [434]. The consumed reducing agent is constantly removed from the reaction as volatile acetone. Acrolein is reduced by 2-propanol to allyl alcohol at elevated temperature in the presence of MgO – ZnO catalysts. Ethanol reductively cleaves aromatic diazonium salts to hydrocarbons, but other functional groups are not attacked. Azoxybenzenes are formed in the reduction of nitroaromatic compounds with benzyl alcohol [435].

Examples of the reductive action of amines analogous to that of alcohols are very rare, although this type of reaction is known as a side reaction in the treatment of simple aldehydes with lithium diisopropylamide [431]. In

the Shell H_2O_2 process (→ Hydrogen Peroxide, Chap. 4.2.), oxygen is reduced to hydrogen peroxide by catalytic H_2 transfer from 2-propanol. Other processes for producing H_2O_2 from oxygen utilize regenerable anthrahydroquinones (→ Hydrogen Peroxide, Chap. 4.1.).

4.8.5. Alkenes

Some alkenes are excellent hydrogen donors for catalytic hydrogenation and hydrogenolysis. Cyclohexene, which is normally used as a reducing agent with a palladium catalyst, is particularly important in this application [418]. Benzyl and benzyloxycarbonyl groups can be removed hydrogenolytically, whereas suitably substituted carbonyl groups can be reduced to methylene groups. Ketones, which are more difficult to reduce, can be reacted with the higher-boiling compound limonene as hydrogen donor. Dinitroaromatic compounds are converted to the corresponding diamines by cyclohexene.

Cycloheptatriene and some of its derivatives act as hydride donors toward many carbenium ions. These reduction mechanisms often proceed under heterogeneous catalysis and are of some industrial interest [431].

4.8.6. Heterocyclic Compounds

The driving force for many heterocyclic systems to transfer hydride ions is formation of the stable aromatic state. 1,4-Dihydropyridines, for example,the commercially available Hantzsch esters, hydrogenate hydride acceptors with the formation of pyridines. 1,4-Dihydronicotinamides and dihydroacridines reduce carbonyl and C – C double bonds. Normally however only electron-deficient double bonds react sufficiently fast under nonenzymatic conditions. In the presence of stoichiometric amounts of $Mg(ClO_4)_2$, optically active *N*-methyldihydropyridines reduce prochiral ketones to alcohols in high optical yields. Dihydropyrimidines and dihydropyrazines are used rarely for reductions because they are often too reactive and difficult to handle [431].

5. Special Reduction Methods

In addition to the aforementioned generally applicable reduction variants, other special procedures also exist, which however are often restricted for experimental reasons to specific applications. Some of these processes, for example, enzymatic reductions, will increasingly complement the general range of methods in the future and nowadays are already the methods of choice in specific problem areas. Pressing ecological and economic questions, as well as the increasing (stereochemical) complexity of chemical structures (e.g., crop protection agents, pharmaceuticals), make biocatalytic process control methods increasingly more advantageous.

Biomimetic Reductions [436]. The best-investigated model structures for simulating biologically relevant reductions are the 1,4-dihydronicotinamides and analogues, which are model compounds for studying NADH-mediated redox reactions. The reductions depend to a great extent on the solvent, the addition of salts, and photosensitizers; they are mainly restricted to allyl acetates.

Photoinduced Reductions. Alcohols, amines, sulfides,and hydrocarbons can transfer electrons to photochemically excited $\pi*$-orbitals of carbonyl compounds or heterocyclic compounds. Similar reductions of nitro, azoxy, and hydrazo compounds are also known [437]. Phenylmethanethiol is desulfurized by UV irradiation in the presence of triethylphosphite [438]. Photo-induced reductions are also used in reprography.

Enzymatic Reductions. Numerous review articles refer to the rapidly growing methods and preparative innovations in biocatalytic reductions [51], [439–445]. Bakers' yeast [446] is one of the most widely used biocatalysts and exhibits an extremely high substrate tolerance in the reduction of ketones and C – C double bonds, and in acyloin condensations [447]. In addition to the multiplicity of possible reactions, an attractive feature is the high enantiomeric and diastereomeric excesses that are obtained, which often can be achieved only with great difficulty by purely chemical methods. In general the most serious problems in enzymatic reduction appear

to be reproducibility, separation of the product from the culture medium, and sufficient purity of the enzymes. The range of applications can be further increased by the use of thermophilic enzymes or by controlled mutagenesis of the amino acid sequence of the active site [448].

Antioxidants. Under the prolonged action of oxygen, a wide range of substances can experience considerable reduction in quality due to undesirable autoxidation processes. To suppress these processes, reducing agents such as electron-rich phenols, hydroquinones, or amines are added to reductively trap peroxide radicals. In biological systems this protective function is performed by enzymes (superoxide dismutases, catalases, peroxidases), vitamins (ascorbic acid, vitamin A, vitamin E), and trace elements [449–454]. The above-mentioned three classes of enzyme catalyze the inactivation of reactive oxygen species such as O_2^- , $HO^\bullet$, H_2O_2, and peroxy radicals. Vitamins C and E act as $^1O_2$ traps, stabilize O_2^- and $HO^\bullet$ radicals, and inhibit the peroxidation of lipid membranes [455–457]. Ascorbic acid can also effectively reduce nitroxide radicals, which are isoelectronic with peroxy radicals [458].

In addition to being effective $^1O_2$ traps, carotinoids protect cells against harmful photooxidation and the action of photosensitizing agents [459]. Plant phenols can also act as antioxidants in other biological systems, e.g., inhibiting autoxidation of unsaturated lipids[460]. A "free-radical theory" explains the process of biological aging in terms of continuous oxidative damage to cellular compartments, caused by reactive oxygen- containing molecules [461, 462,462,463]. A series of biological observations have suggested a connection between aging processes and oxidative stress [462]. The sensitivity of the brain towards radicals is shown in modes of cerebral ischemia in which neuron damage is reduced by antioxidants [455]. An attempt has been made to explain the neurodegenerative Alzheimer's disease in terms of the oxidative, radical, cell-damaging activity of ß-amyloid protein fragments, which could possibly be prevented by prophylatic administration of antioxidants [464–467]. Many of the numerous organosulfur compounds found in garlic and onions have antioxidant activity. The physiological activity profile of Allium species is in part attributable to the reductive cell-protecting activity of certain components [468–470].

Reductive Combustion. A technically interesting method for reducing NO_x emissions from combustion processes is to react the nitrogen oxides with reducing agents such as CO and H_2 [471–479]. A process for the reductive decomposition of NO_x-containing gases with voltages in excess of 20 000 V is being tested. Nitrogen oxides disproportionate into their elements under the influence of high voltage. Technically, this process could be important in reducing emissions from fossil fuel power stations and automobiles [480].

Removal of NO_x by reduction with NH_3 (g) or aqueous NH_3 is known as selective catalytic reduction and is suitable for emission control in power stations, incinerators, glass-melting tanks, and combustion engines [474], [481] [482]. The degree of reduction depends on the nature of the catalyst and the temperature range [483], [484].

6. References

1. H. R. Christen: *Grundlagen d. Allgemeinen und Anorganischen Chemie,* Sauerländer-Salle, Stuttgart 1973, pp. 383 ff.
2. A. F. Hollemann, E. Wiberg: *Lehrbuch d. Anorganischen Chemie,* De Gruyter, Berlin 1971, pp. 188 ff.
3. D. Dobos: *Electrochemical Data,* Elsevier, Amsterdam 1975, p. 247.
4. M. Freifelder: *Catalytic Hydrogenation in Organic Synthesis; Procedures and Commentary,* J. Wiley & Sons, New York 1978.
5. H. B. Kagan, *Bull. Soc. Chim. Fr.* **5** (1988) no. 5, 846.
6. H. Brunner, *Top. Stereochem.* **18** (1988) 129.
7. R. Noyori: "3 rd ed MERCK-Schuchardt Chair in Organic Synthesis," Belgian Organic Synthesis Symposium, Louvain-La-Neuve 1990.
8. B. M. Trost: "Reduction," in I. Fleming (ed.): *Comprehensive Organic Synthesis,* vol. 8, Pergamon Press, Oxford 1991.
9. R. C. Larock: "Reduction," *Comprehensive Organic Transformations,* – 1., VCH Verlagsgesellschaft, Weinheim 1989.

10. M. Hudlicky: *Reductions in Organic Chemistry,* Ellis Horwood Ltd., Chichester 1984.
11. J. March: "Oxidations and Reductions," *Advanced Organic Chemistry,* 3rd ed., J. Wiley & Sons, New York 1985.
12. G. W. J. Fleet, *Org. React. Mech.* 1986, 167.
13. D. C. Liotta, M. Volmer (eds.): *Organic Synthesis Reaction Guide,* J. Wiley & Sons, New York 1991, pp. 235 ff.
14. J. D. White, *Tetrahedron Lett.* 1974, 2879.
15. W. S. Murphy, S. Wattanasin, *Tetrahedron Lett.* 1981, 695.
16. G. M. Whitesides, W. J. Ehmann, *J. Org. Chem.* **35** (1970) 3565.
17. E. E. Van Tamelen, R. J. Timmons, *J. Am. Chem. Soc.* **84** (1962) 1067.
18. R. F. Abdulla, *Aldrichimica Acta* **21** (1988) 31.
19. E. C. Ashby, J. J. Lin, *J. Org. Chem.* **43** (1978) 2567.
20. E. C. Ashby, J. J. Lin, A. B. Goel, *J. Org. Chem.* **43** (1978) 757.
21. T. Imamoto, T. Mita, M. Yokoyama, *J. Org. Chem.* **52** (1987) 5695.
22. S. M. Soll, S. P. Seitz, *Tetrahedron Lett.* 1987, 5457.
23. A. A. Akhrem, I. G. Reshetova, Y. A. Titov: *Birch Reductions of Aromatic Compounds,* IFI/Plenum, New York 1972.
24. H. Nohe, *Chem. Ing. Tech.* **46** (1974) 597.
25. S. Krishnamurthy, H. C. Brown, *J. Org. Chem.* **48** (1983) 3085.
26. E. C. Ashby, J. J. Lin, *J. Org. Chem.* **43** (1978) 1263.
27. F. Sato, T. Akiyama, K. Jida, M. Sato, *Synthesis* 1982, 1025.
28. D. Savoia, E. Tagliavini, C. Trombini, A. Umani-Ronchi, *J. Org. Chem.* **47** (1982) 876.
29. S. K. Boyer, J. Bach, J. McKenna, E. Jagdmann, Jr., *J. Org. Chem.* **47** (1982) 876.
30. J. T. Pinhey, R. D. G. Rigby, *Tetrahedron Lett.* 1969, 1267.
31. J. D. Citron, J. E. Lyons, L. H. Sommer, *J. Org. Chem.* **34** (1969) 638.
32. H. C. Brown, S. Krishnamurthy, *Tetrahedron* **35** (1979) 567.
33. P. H. Gore, G. K. Hughes, *J. Chem. Soc.* 1959, 1615.
34. C. A. Hoeger, A. D. Johnston, W. H. Okamura, *J. Am. Chem. Soc.* **109** (1987) 4690.
35. J. D. White, T. C. Somers, *J. Am. Chem. Soc.* **109** (1987) 4424.
36. R. A. Rossi, J. F. Bunnett, *J. Org. Chem.* **38** (1973) 2314.
37. T. Shono, Y. Matsumura, K. Tsubata, Y. Sugihara, *J. Org. Chem.* **44** (1979) 4508.
38. N. L. Weinberg: "Technique of Electroorganic Synthesis," in A. Weissberger (ed.): *Techniques of Chemistry,* vol. 5/1 + 2, J. Wiley & Sons, New York 1975, p. 224.
39. Metallgesellschaft, Lithiumaluminiumhydrid, Company Brochure, Frankfurt/M. 1974.
40. Bayer, Natriumborhydrid, Leverkusen 1975.
41. R. Hiatt in S. Swern (ed.): *Organic Peroxides,* **vol. 2,** J. Wiley & Sons, New York 1971, p. 50.
42. L. A. Flippin, D. W. Gallagher, K. Jalali-Araghi, *J. Org. Chem.* **54** (1989) 1430.
43. U. Azzena, T. Denurra, G. Melloni, G. Rassu, *J. Chem. Soc. Chem. Comm.* 1987, 1549.
44. D. A. Forsyth, M. R. Estes, P. Lucas, *J. Org. Chem.* **47** (1982) 4380.
45. D. G. Loughhead, *J. Org. Chem.* **50** (1985) 3931.
46. Y. Fort, R. Vanderesse, P. Caubere, *Tetrahedron Lett.* 1985, 3111.
47. H. Suzuki, T. Fuchita, A. Iwasa, T. Mishina, *Synthesis* 1978, 905.
48. P. Dowd, K. Kang, *J. Chem. Soc. Chem. Comm.* 1974, 384.
49. B. C. Ranu, R. Chakraborty, *Tetrahedron Lett.* 1990, 7663.
50. N. Y. M. Fung, P. de Mayo, J. H. Schauble, A. C. Weedon, *J. Org. Chem.* **43** (1978) 3977.
51. R. Csuk, B. I. Glänzer, *Chem. Rev.* **91** (1991) 49.
52. W. Uhl, A. Kyriatsoulis: *Namen und Schlagwortreaktionen in der Organischen Chemie,* Vieweg & Sohn, Braunschweig 1984, pp. 25, 111.
53. J. E. McMurry, *Chem. Rev.* **89** (1989) 1513.
54. J. H. Freudenberger, A. W. Konradi, S. F. Pedersen, *J. Am. Chem. Soc.* **111** (1989) 8014.
55. G. A. Molander, *Chem. Rev.* **92** (1992) 29.
56. G. W. Kabalka, S. T. Summers, *J. Org. Chem.* **46** (1981) 1217.
57. E. Vedejs, *Org. React. (N.Y.)* **22** (1974) 401.
58. C. F. Lane, *Aldrichimica Acta* **10** (1977) 41.
59. *Houben-Weyl,* 4th ed., **11/1,** 648.
60. Ref. [52], p. 102.
61. D. Lenoir, *Chem. Ber.* **111** (1978) 411.
62. R. F. Nystrom, C. R. A. Berger, *J. Am. Chem. Soc.* **80** (1958) 2896.
63. W. S. Murphy, D. F. Sullivan, *J. Chem. Soc. Perkin Trans. 1:* 1972, 999.
64. D. W. Sopher, J. H. P. Utley, *J. Chem. Soc. Chem. Comm.* 1979, 1087.
65. C. Ferri: *Reaktionen d. Organischen Synthese,* Thieme Verlag, Stuttgart 1978, p. 133.
66. J. Nakayama, H. Machida, R. Saito, M. Hoshino, *Tetrahedron Lett.* 1985, 1981.

67. C. H. Foster, D. A. Payne, *J. Am. Chem. Soc.* **100** (1978) 2834.
68. *Houben-Weyl,* 4th ed., **5/1 a,** 276.
69. Ref. [52], p. 32 and references cited therein.
70. F. Beck: *Elektroorganische Chemie,* Verlag Chemie, Weinheim 1974, p. 153.
71. R. L. Whistler, M. L. Wolfrom (eds.): *Methods in Carbohydrate Chemistry,* **vol. 2,** Academic Press, New York 1963, p. 65.
72. E. Fischer, *Ber. Dtsch. Chem. Ges.* **22** (1889) 94.
73. H. M. Crawford, M. Lumpkin, M. McDonald, *J. Am. Chem. Soc.* **74** (1952) 4087.
74. G. A. Reynolds, J. A. Vanallan, *Org. Synth. Coll.* **4** (1963) 15.
75. J. Q. Chambers, S. Patai (eds.): *The Chemistry of the Quinoid Compound,* **vol. 2,** J. Wiley & Sons, New York 1974, p. 737.
76. J. Fatiadi, W. F. Sager, *Org. Synth. Coll.* **5** (1973) 595.
77. M. Konieczny, R. G. Harvey, *J. Org. Chem.* **44** (1979) 4813.
78. J. T. Traxler, *Synth. Commun.* **7** (1977) 161.
79. K. Venkataraman: *The Chemistry of Synthetic Dyes,* **vol. 5,** Academic Press, London 1971, p. 142.
80. Ref. ,[2] p. 527.
81. J. H. Wagenknecht, *J. Org. Chem.* **37** (1972) 1513.
82. A. O. Bedenbaugh, J. H. Bedenbaugh, W. A. Bergin, J. D. Adkins, *J. Am. Chem. Soc.* **92** (1970) 5774.
83. R. Knorr, P. Löw, P. Hassel, *Synthesis* 1983, 785.
84. K. Haaf, C. Rüchardt, *Chem. Ber.* **123** (1990) 635.
85. M. D. Birkett, A. T. Kuhn, *Elektrochim. Acta* **21** (1976) 991.
86. L. I. Zakharin, V. V. Gavrilenko, D. N. Maslin, I. M. Korlina, *Tetrahedron Lett.* 1963, 2087.
87. E. Winterfeldt, *Synthesis* 1975, 617.
88. J. J. Bloomfield, *Tetrahedron Lett.* 1968, 591.
89. K. Rühlmann, *Synthesis* 1971, 236.
90. K. T. Finley, *Chem. Rev.* **64** (1964) 573.
91. G. L. Baker, S. J. Fritschel, J. R. Stille, J. K. Stille, *J. Org. Chem.* **46** (1981) 2954.
92. M. Brudermüller, H. Musso, *Angew. Chem. Int. Ed. Engl.* **27** (1988) 298.
93. Ref. ,[52] p. 20.
94. F. Guibe, P. Four, H. Riviere, *J. Chem. Soc. Chem. Comm.* 1980, 432.
95. C. Elschenbroich, A. Salzer; *Organometallchemie,* Teubner Studienbücher, Stuttgart 1988, p. 285.
96. E. Santaniello, C. Farachi, A. Manzocchi, *Synthesis* 1979, 912.
97. G. W. Kabalka, J. D. Baker Jr., G. W. Neal, *J. Org. Chem.* **42** (1977) 512.
98. K. Soai, S. Yokoyama, K. Mochida, *Synthesis* 1987, 647.
99. A. G. M. Barrett in Ref. [8], p. 241.
100. N. Minami, S. Kijima, *Chem. Pharm. Bull.* **27** (1979) 816.
101. J. A. Fehrentz, B. Castro, *Synthesis* 1983, 676.
102. J. Malek, H. Cerny, *Synthesis* 1972, 217.
103. A. J. Birch, J. Cymerman-Craig, M. Slayton, *Aust. J. Chem.* **8** (1955) 512.
104. H. C. Brown, S. C. Kim, S. Krishnamurthy, *J. Org. Chem.* **45** (1980) 1.
105. H. C. Brown, P. Heim, *J. Am. Chem. Soc.* **86** (1964) 3566.
106. B. D. James, *Chem. Ind. (London)* 1971, 227.
107. N. Umino, T. Iwakuma, N. Itoh, *Tetrahedron Lett.* 1976, 763.
108. H. C. Brown, C. P. Garg, *J. Am. Chem. Soc.* **86** (1964) 1085.
109. H. Stephen, *J. Chem. Soc.* 1925, 1874.
110. O. G. Backeberg, B. Staskun, *J. Chem. Soc.* 1962, 3961.
111. S. Pietra, C. Trinchera, *Gazz. Chim. Ital.* **85** (1955) 1705.
112. R. A. Egli, *Helv. Chim Acta* **53** (1970) 47.
113. R. F. Nystrom, W. G. Brown, *J. Am. Chem. Soc.* **70** (1948) 3738.
114. J. Malek, *Org. React. (N.Y.)* **36** (1988) 249.
115. *Houben-Weyl,* 4th ed., **11/1,** 600.
116. J. R. Brindle, J. L. Liard, N. Berube, *Can. J. Chem.* **54** (1976) 871.
117. T. Pienemann, H. J. Schäfer, *Synthesis* 1987, 1005.
118. J. H. Billmann, A. C. Diesing, *J. Org. Chem.* **22** (1957) 1068.
119. E. M. Kaiser, *Synthesis* 1972, 391.
120. G. H. Timms, E. Wildsmith, *Tetrahedron Lett.* 1971, 195.
121. G. W. Gribble, R. W. Leiby, M. N. Sheehan, *Synthesis* 1977, 856.
122. R. O. Hutchins et al., *Org. Prep. Proced. Int.* **16** (1984) 335.
123. S. Itsuno, Y. Sakurai, K. Ito, *Synthesis* 1988, 995.
124. S. Dev, *Proc. Indian Acad. Sci.* **93** (1984) 1015.
125. R. Balicki, *Synthesis* 1989, 645.
126. J. P. Dirlam, J. W. MacFarland, *J. Org. Chem.* **42** (1977) 1360.
127. L. Horner, H. Röder, *Ber. Dtsch. Chem. Ges.* **101** (1968) 4179.
128. G. Lunn, E. B. Sansone, L. K. Keefer, *Synthesis* 1985, 1104.

129. L. K. Keefer, G. Lunn, *Chem. Rev.* **89** (1989) 459.
130. B. T. Hayes, T. S. Stevens, *J. Chem. Soc. C* 1970, 1088.
131. W. W. Hartmann, L. J. Roll, *Org. Synth. Coll.* **2** (1943) 418.
132. F. W. Schueler, C. Hauna, *J. Am. Chem. Soc.* **73** (1951) 4996.
133. J. S. Buck, C. W. Ferry, *Org. Synth. Coll.* **2** (1943) 290.
134. Ref. [10], p. 182.
135. A. Calder, A. R. Forrester, S. P. Hepburn, *Org. Synth. Coll.* **6** (1988) 803.
136. P. E. Iversen, T. B. Christensen, *Acta Chem. Scand. Ser. B:* **B 31** (1977) 733.
137. B. H. Han, D. H. Shin, S. Y. Cho, *Tetrahedron Lett.* 1985, 6233.
138. A. Hamdan, J. W. F. Wasley, *Synth. Commun.* **15** (1985) 71.
139. H. Ohmori et al., *Chem. Pharm. Bull.* **32** (1984) 3345.
140. S. Ram, R. E. Ehrenkaufer, *Tetrahedron Lett.* 1984, 3415.
141. E. T. McBee, E. Rapkin, *J. Am. Chem. Soc.* **73** (1951) 1325.
142. B. T. Newbold, *J. Chem. Soc.* 1961, 4260.
143. A. McKillop, R. A. Raphael, *J. Org. Chem.* **35** (1970) 1670.
144. G. W. Kabalka, R. S. Varma in Ref. , [8]p. 363.
145. H. E. Bigelow, D. B. Robinson, *Org. Synth. Coll.* **3** (1955) 103.
146. Ref. [10], p. 72.
147. J. E. Kmiecik, *J. Org. Chem.* **30** (1965) 2014.
148. S. Uchida, K. Yanada, H. Yamaguchi, H. Meguri, *Chem. Lett.* 1986, 1069.
149. F. D. Bellamy, K. Ou, *Tetrahedron Lett.* 1984, 839.
150. Ref. [10], p. 73.
151. B. T. Newbold in S. Patai (ed.): *The Chemistry of the Hydrazo-, Azo- and Azoxygroup,* J. Wiley & Son, New York 1975, p. 599.
152. F. G. Thomas, K. G. Boto in Ref. [151], p. 443.
153. N. Kambe, K. Kondo, N. Sonoda, *Angew. Chem. Int. Ed. Engl.* **19** (1980) 1009.
154. J. W. Hamersma, E. I. Snyder, *J. Org. Chem.* **30** (1965) 3985.
155. Ref. [10], p. 95.
156. T. L. Ho, G. A. Olah, *Synthesis* 1985, 1104.
157. Ref. [65], pp. 129 ff.
158. T. L. Gilchrist in Ref. [8], p. 381.
159. Ref. [10], p. 75.
160. E. König, H. Musso, U. I. Zahorsky, *Angew. Chemie* **84** (1972) 33.
161. T. Hirabe, M. Nojima, S. Kusabajashi, *J. Org. Chem.* **49** (1984) 4084.
162. Ref. [10], pp. 90 ff.
163. S. W. McCombie in Ref. [8], p. 826.
164. P. Caubere, P. Coutrot in Ref. [8], p. 860.
165. J. B. Hendrickson, R. Bergeron, *Tetrahedron Lett.* 1970, 345.
166. T. H. Chan, J. P. Montillier, W. F. van Horn, D. N. Harpp, *J. Am. Chem. Soc.* **92** (1970) 7224.
167. W. A. Sheppard, *Org. Synth. Coll.* **5** (1973) 843.
168. A. Furst, R. C. Berlo, S. Hooton, *Chem. Rev.* **65** (1965) 51.
169. *Houben-Weyl,* 4th ed., **9,** 539.
170. I. W. J. Still, S. Szilagyi, *Synth. Commun.* **9** (1979) 923.
171. F. G. Bordwell, W. H. McKellin, *J. Am. Chem. Soc.* **73** (1951) 2251.
172. J. N. Gardner, S. Kaiser, A. Krubiner, H. Lucas, *Can. J. Chem.* **51** (1973) 1419.
173. J. Bremner, M. Julia, M. Launay, J. P. Stacino, *Tetrahedron Lett.* 1982, 3265.
174. T. G. Back, J. R. Proudfoot, C. Djerassi, *Tetrahedron Lett.* 1986, 2187.
175. K. Sato et al., *J. Chem. Soc. Perkin Trans. 1* 1981, 761.
176. W. E. Truce, F. M. Perry, *J. Org. Chem.* **30** (1965) 1316.
177. G. A. Olah, G. K. Prakash, T. L. Ho, *Synthesis* 1976, 810.
178. T. L. Ho, C. M. Wong, *Synthesis* 1973, 206.
179. J. Drabowicz, S. Oae, *Synthesis* 1977, 404.
180. G. H. Posner et al., *J. Org. Chem.* **46** (1981) 5244.
181. L. L. Klein, M. S. Shanklin, *J. Org. Chem.* **53** (1988) 5202.
182. P. Caubere, P. Coutrot in Ref. [8], p. 843.
183. S. N. Matai, P. Spevac, M. P. Singh, R. G. Micetich, *Synth. Commun.* **18** (1988) 575.
184. A. Ookawa, S. Yokoyama, K. Soai, *Synth. Commun.* **16** (1986) 819.
185. H. Suzuki, N. Sato, *Chem. Lett.* 1981, 267.
186. S. C. Shim, S. Antebi, H. Alper, *Tetrahedron Lett.* 1985, 1935.
187. W. P. Neumann, *Synthesis* 1987, 665.
188. H. Gerischer, C. W. Tobias (eds.): *Advances in Electrochemical Science and Engineering,* vol. 1, VCH Verlagsgesellschaft, Weinheim 1990.
189. H. Kolbe, *Justus Liebigs Ann. Chem.* **69** (1849) 257.
190. F. Fichter, *Organische Elektrochemie,* Verlag Steinkopf, Dresden 1942.
191. M. M. Baizer, *J. Electrochem. Soc.* **111** (1964) 215.

192. M. M. Baizer, *Chem. Eng. (N.Y.)* **72** (1965) 238.
193. H. J. Schäfer, *Kontakte (Darmstadt)* **1987,** no. 2, 17;**1987,** no. 3, 37.
194. R. Jansson, *Chem. Eng. News* (1984) Nov. 19, 43.
195. M. F. Semmelhack, R. J. DeFranco, J. Stock, *Tetrahedron Lett.* 1972, 1371.
196. R. N. Renaud, *Can. J. Chem.* 1974, 376.
197. A. J. Fry, *Aldrichimica Acta* **26** (1993) 3.
198. Johnson Matthey Royston, The Catalytic Reaction Guide, Product Information, 1991.
199. H. Werner, *Angew. Chem.* **102** (1990) 1109.
200. D. Wöhrle, *Nachr. Chem. Tech. Lab.* **39** (1991) 1256.
201. Ref. [95], p. 502.
202. E. C. Ashby, S. A. Noding, *J. Org. Chem.* **45** (1980) 1041.
203. R. Vanderesse, J. J. Brunet, P. Caubere, *J. Org. Chem.* **46** (1981) 1270, and references cited therein for various activators and possible reduction reactions.
204. R. Pi et al., *J. Org. Chem.* **52** (1987) 4299.
205. Ref. [206], p. 5.
206. E. C. Ashby, J. J. Lin, A. B. Goel, *J. Org. Chem.* **43** (1978) 1560, 1564.
207. H. C. Brown, J. B. Campbell Jr., *Aldrichimica Acta* **14** (1981) 3.
208. H. C. Brown, U. S. Racherla, *Tetrahedron Lett.* 1985, 2187.
209. G. W. Kabalka, *Aldrichimica Acta* **19** (1986) 11.
210. J. H. Brewster, *Aldrichimica Acta* **20** (1987) 3.
211. M. Srebnik, P. V. Ramachandran, *Aldrichimica Acta* **20** (1987) 9.
212. C. F. Lane, *Chem. Rev.* **76** (1976) 773.
213. C. F. Lane, *Aldrichimica Acta* **10** (1977) 3.
214. J. Mulzer et al. (eds.): "Boron: Reagents for Stereoselectiv Syntheses," in *Organic Synthesis Highlights,* VCH Verlagsgesellschaft, Weinheim 1991, p. 33.
215. H. C. Brown, C. D. Blue, D. J. Nelson, N. G. Bhat, *J. Org. Chem.* **54** (1989) 6064.
216. H. C. Brown, S. Krishnamurthy, N. M. Yoon, *J. Org. Chem.* **41** (1976) 1778.
217. R. K. Dhar, *Aldrichimica Acta* **29** (1994)43.
218. J. V. B. Kanth, M. Periasamy, *J. Chem. Soc. Chem. Commun.* 1990, 1145.
219. O. Kriz, Z. Plzak, J. Plesek, *Collect. Czech. Chem. Commun.* **55** (1990) 2956.
220. M. D. Bormann, I. C. GUCH, M. DiMARE, *J. Org. Chem.* **60** (1995) 5995.
221. G. W. Kabalka, *Org. Prep. Proced. Int.* **9** (1977) 131.
222. U. Wietelmann, *Janssen Chimica Acta* **10** (1992) 16.
223. D. A. Evans, A. H. Hoveyda, *J. Org. Chem.* **55** (1990) 5190.
224. C. Einhorn, J. Einhorn, J. L. Luche, *Chem. Rev.* **89** (1989) 787.
225. D. N. Kursanov, Z. N. Parnes, N. M. Loim, *Synthesis* 1974, 633.
226. "Hydrosilierung" in Ref. [95].
227. A. Kürzinger: "Asymmetrische Hydrosilylierung von Ketonen und Alkaloidvorläufern mit Rhodium/Thiazolidin-Komplexen," Dissertation, Universität Regensburg 1987.
228. M. Fujita, T. Hiyama, *J. Am. Chem. Soc.* **106** (1984) 4629.
229. G. A. Kraus et al., *J. Org. Chem.* **46** (1981) 2417.
230. J. Pataki, R. G. Harvey, *J. Org. Chem.* **52** (1987) 2226.
231. J. Nakayama, M. Yoshida, O. Simamura, *Tetrahedron* **26** (1970) 4609.
232. M. Oba, *Tetrahedron* **50** (1994) 10 193.
233. G. Büchi, H. Wuest, *J. Org. Chem.* **44** (1979) 546.
234. Y. Guindon et al., *Tetrahedron Lett.* 1991, 27.
235. T. Ando, T. Ishihara, E. Ohtani, H. Sawada, *J. Org. Chem.* **46** (1981) 4446.
236. B. Giese: "Radicals in Organic Synthesis," *Organic Chemistry Series,* vol. 5, Pergamon Press, Oxford 1986.
237. N. A. Porter, D. R. Magnin, B. T. Wright, *J. Am. Chem. Soc.* **108** (1986) 2787.
238. T. Kametani, T. Honda, *Heterocycles* **19** (1982) 1861.
239. U. Gerigk et al., *Synthesis* 1990, 448.
240. The selectivity is demonstrated in *G. Carrara et al., J. Am. Chem. Soc.* **76** (1954) 4391. The reduction of an aldehyde to the alcohol takes place in the presence of a nitrogroup, a C – C double bond, a thiophene ring, and a halide-group.
241. J. R. Luly et al., *J. Org. Chem.* **52** (1987) 1487.
242. A. Giannis, K. Sandhoff, *Angew. Chem. Int. Ed. Engl.* **28** (1989) 218, and references cited therein.
243. R. Dharanipragada, A. Alarcon, V. Hruby, *Org. Prep. Proced. Int.* **23** (1991) 396.
244. H. Kotsuki, Y. Ushio, N. Yoshimura, M. Ochi, *J. Org. Chem.* **52** (1987) 2594, and references cited therein.
245. P. Crabbe, G. A. Garcia, C. Ruis, *J. Chem. Soc. Perkin Trans. 1:* 1973, 810.
246. Ref. [259], pp. 1 b, 1 c, 1 u, 1 cc.
247. M. Narisada, I. Horibe, F. Watanabe, K. Takeda, *J. Org. Chem.* **54** (1989) 5308.

248. C. A. Brown, V. K. Ahuja, *J. Chem. Soc. Chem. Commun.* 1973, 553.
249. S. Itsuno, Y. Sakurai, K. Shimizu, K. Ito, *J. Chem. Soc. Perkin Trans 1*: 1990, 1859.
250. C. P. Cartaya-Martin, M. A. Rissolo, *Synth. Commun.* **21** (1991) 951.
251. J. V. B. Kanth, M. Periasamy, *J. Org. Chem.* **56** (1991) 5964.
252. R. Kawecki, L. Kozerski, Z. Urbanczyk-Lipkowska, G. Bocelli, *J. Chem. Soc. Perkin Trans. 1,* 1991, 2255.
253. G. Verardo, A. G. Giumanini, G. Favret, P. Strazzolini, *Synthesis* 1991, 447.
254. Y. Hamada et al., *J. Org. Chem.* **52** (1987) 1252.
255. S. Yoo, S. Lee, *Synlett* 1990, 419.
256. H. C. Brown, C. A. Brown, *J. Am. Chem. Soc.* **85** (1965) 1005.
257. T. E. A. Nieminen, T. A. Hase, *Tetrahedron Lett.* 1987, 4725.
258. J. H. Choi, D. W. Kim, S. C. Shim, *Tetrahedron Lett.* 1986, 1157.
259. D. E. Ward, C. K. Rhee, *Synth. Commun.* **18** (1988) 1927.
260. R. O. Hutchins, N. R. Natale, *Org. Prep. Proced. Int.* **11** (1979) 201.
261. D. A. Evans, K. T. Chapman, *Tetrahedron Lett.* 1986, 5939.
262. R. O. Hutchins, M. Kacher, L. Rua, *J. Org. Chem.* **40** (1975) 923.
263. J. L. Wood, D. R. Jones, R. Hirschmann, A. B. Smith, *Tetrahedron Lett.* 1990, 6329.
264. R. J. Mattson, K. M. Pham, D. J. Leuck, K. A. Cowen, *J. Org. Chem.* **55** (1990) 2552.
265. C. K. Lau et al., *J. Org. Chem.* **51** (1986) 3038.
266. D. A. Evans, K. T. Chapman, E. M. Carreira, *J. Am. Chem. Soc.* **110** (1988) 3560.
267. D. A. Evans, J. A. Gauchet-Prunet, E. M. Carreira, A. B. Charette, *J. Org. Chem.* **56** (1991) 741.
268. P. von Matt, A. Pfaltz, *Tetrahedron Asymmetry* 1991, 691.
269. Merck Index 11, 351.
270. P. Nayler, M. C. Whiting, *J. Chem. Soc.* 1954, 4006.
271. H. R. Pfaendler, F. K. Maier, S. Klar, *J. Am. Chem. Soc.* **108** (1986) 1338.
272. J. H. Brewster, H. O. Bayer, S. F. Osman, *J. Org. Chem.* **29** (1964) 110.
273. N. Shimizu, K. Watanabe, Y. Tsuno, *Chem. Lett.* 1983, 1877.
274. F. Tellier, R. Sauvetre, J. F. Normant, "Depending on the reducing agent ($LiAlH_4$ or $NaBH_4$), fluorinated allyl alcohols give different dehydrogenation products,"*Tetrahedron Lett.* 1987, 3335.
275. T. Imamoto, T. Takeyama, T. Kusomoto, *Chem. Lett.* 1985, 1491.
276. Y. Mori, M. Kuhara, A. Takeuchi, M. Suzuki, *Tetrahedron Lett.* 1988, 5419.
277. C. Bilger, R. Royer, P. Demerseman, *Synthesis* 1988, 902.
278. J. Powell, N. James, S. J. Smith, *Synthesis* 1986, 338.
279. G. Zweifel, G. R. Hahn, T. M. Shoup, *J. Org. Chem.* **52** (1987) 5486.
280. E. M. Marlett, W. S. Park, *J. Org. Chem.* **55** (1990) 2968.
281. Y. Tamai, S. Koike, S. Ogura, S. Miyano, *J. Chem. Soc. Chem. Commun.* 1991, 799.
282. M. Cherest, H. Felkin, N. Prudent, *Tetrahedron Lett.* 1968, 2199.
283. S. Nagahara, T. Uchimoto, W. Yano, Y. Doi, *Nippon Kagaku Kaishi* 1985, 2246; *Chem. Abstr.* **105** (1986) 23798.
284. J. F. Le Marechal, M. Ephritikhine, G. J. Folcher, *J. Organomet. Chem.* **309** (1986) C 1.
285. S. Kim, K. H. Ahn, *J. Org. Chem.* **49** (1984) 1717.
286. S. Kim, K. H. Ahn, Y. W. Chung, *J. Org. Chem.* **47** (1982) 4581.
287. S. Krishnamurthy, *J. Org. Chem.* **46** (1981) 4628.
288. Ethyl Corporation, US 4 474 743, 1984 (E. M. Marlett).
289. E. H. Gold, E. Babad, *J. Org. Chem.* **37** (1972) 2208.
290. T. K. Jones, S. E. Denmark, *Org. Synth. Coll.* **64** (1986) 182.
291. Y. Gao, K. B. Sharpless, *J. Org. Chem.* **53** (1988) 4081.
292. S. A. Bezman, M. R. Churchill, J. A. Osborn, J. Wormald, *J. Am. Chem. Soc.* **93** (1971) 2063.
293. W. S. Mahoney, D. M. Brestensky, J. M. Stryker, *J. Am. Chem. Soc.* **110** (1988) 291.
294. D. M. Brestensky, J. M. Stryker, *Tetrahedron Lett.* 1989, 5677.
295. J. F. Daeuble, C. McGettigan, J. M. Stryker, *Tetrahedron Lett.* 1990, 2387.
296. W. S. Mahoney, J. M. Stryker, *J. Am. Chem. Soc.* **111** (1989) 8818.
297. H. des Abbayes, *New. J. Chem.* **11** (1987) 535.
298. H. Alper, D. des Roches, *Tetrahedron Lett.* 1977, 4155.
299. H. Alper, E. C. H. Keung, *Tetrahedron Lett.* 1970, 53.
300. R. Noyori, I. Umeda, T. Ishigami, *J. Org. Chem.* **37** (1972) 1542.
301. S. J. Nelson, G. Detre, M. Tanabe, *Tetrahedron Lett.* 1973, 447.

302. R. W. Goetz, M. Orchin, *J. Am. Chem. Soc.* **85** (1963) 2782.
303. G. P. Boldrini, A. Umani-Ronchi, *Synthesis* 1976, 596.
304. H. Alper, *Pure Appl. Chem.* **52** (1980) 607.
305. R. Noyori et al., *J. Am. Chem. Soc.* **100** (1978) 1759.
306. Ref. [52], p. 11 and references cited therein.
307. A. L. Wilds, N. A. Nelson, *J. Am. Chem. Soc.* **75** (1953) 5360.
308. P. Q. Huang, S. Arseniyadis, H. P. Husson, *Tetrahedron Lett.* 1987, 547.
309. G. Mehta, S. Padma, *J. Am. Chem. Soc.* **109** (1987) 7230.
310. Ref. [95], p. 31.
311. E. E. van Tamelen, J. G. Carlson, R. K. Russel, S. R. Zawacky, *J. Am. Chem. Soc.* **103** (1981) 4615.
312. S. K. Pradhan, *Tetrahedron* **42** (1986) 6351.
313. R. R. Chambers, C. J. Collins, B. E. Maxwell, *J. Org. Chem.* **50** (1985) 4960.
314. Ref. [52], p. 20.
315. Ref. [52], p. 174.
316. G. Solladie, A. Urbano, G. B. Stone, *Synlett* 1993, 548.
317. T. Oshawa, T. Takagaki, A. Haneda, T. Oishi, *Tetrahedron Lett.* 1981, 2583.
318. T. Ohsawa et al., *Tetrahedron Lett.* 1985, 6103.
319. Ref. [96], p. 49.
320. M. G. Zagorski, R. G. Salomon, *J. Am. Chem. Soc.* **106** (1984) 1750.
321. M. D. Rausch, *Chem. Rev.* **57** (1957) 417.
322. C. Dominguez, A. G. Csaky, J. Plumet, *Tetrahedron Lett.* 1991, 4183.
323. T. Ohno, Y. Ishino, Y. Tsumagari, I. Nishiguchi, *J. Org. Chem.* **60** (1995) 458.
324. *Merck Index* **vol. 11,** p. 1642.
325. J. R. Hwu et al., *J. Org. Chem.* **51** (1986) 4731.
326. J. H. Chapman, J. Elks, G. H. Philipps, L. J. Wyman, *J. Chem. Soc.* 1956, 4344.
327. A. W. Burgstahler, R. E. Sticker, *Tetrahedron* **24** (1968) 2435.
328. R. A. Benkeser, A. Rappa, L. A. Wolsieffer, *J. Org. Chem.* **51** (1986) 3391.
329. H. Neunhöfer, G. Köhler, *Tetrahedron Lett.* 1978, 4879.
330. R. A. Benkeser, J. A. Laugal, A. Rappa, *Tetrahedron Lett.* 1984, 2089.
331. R. A. Benkeser, F. G. Belmonte, J. Kang, *J. Org. Chem.* **48** (1983) 2796.
332. A. A. P. Schreibmann, *Tetrahedron Lett.* 1970, 4271.
333. L. F. Fieser, M. Fieser: *Reagents for Organic Synthesis,* **vol. 1 – 16,** J. Wiley & Sons, New York 1967 –1992.
334. M. F. Asaro, I. Nakayama, R. B. Wilson, Jr., *J. Org. Chem.* **57** (1992) 778.
335. J. P. Schaefer, *J. Org. Chem.* **25** (1960) 2027.
336. J. Nokami, J. Otera, T. Sudo, R. Okawara, *Organometallics* **2** (1983) 191.
337. J. Nokami, S. Wakabayashi, R. Okawara, *Chem. Lett.* 1984, 869.
338. J. M. Coxon, S. J. van Eyk, P. J. Steel, *Tetrahedron* **45** (1989) 1029.
339. C. Petrier, J. Einhorn, J. L. Luche, *Tetrahedron Lett.* 1985, 1449.
340. A. Courtin, *Helv. Chim. Acta* **62** (1980) 2280.
341. D. C. Owsley, J. J. Bloomfield, *Synthesis* 1977, 118.
342. D. W. Hall, E. Hurley, Jr., *Can. J. Chem.* **47** (1969) 1238.
343. E. Clemmenson, *Ber. Dtsch. Chem. Ges.* **46** (1913) 1837.
344. E. Vedejs, *Org. React. (N.Y.)* **22** (1975) 401.
345. M. W. Rathke, *Org. React. (N.Y.)* **22** (1975) 423.
346. R. D. Rieke, S. J. Uhm, *Synthesis* 1975, 452.
347. H. E. Simmons, T. L. Cairns, S. A. Vladuchick, C. M. Hoiness, *Org. React. (N.Y.)* **20** (1973) 1.
348. S. Chandrasekaran, J. V. Turner, *Tetrahedron Lett.* 1982, 3799.
349. W. N. Chou, D. L. Clark, J. B. White, *Tetrahedron Lett.* 1991, 299.
350. For the activation of Zn with acids, KCN, Cu,Cu – Ag or Hg, see Refs. 4 – 8 in [349].
351. S. Murai, T. Murai, S. Kato in Ref. [9], p. 871.
352. A. Fischli, *Helv. Chim. Acta* **61** (1978) 2560.
353. C. Petrier, J. L. Luche, *Tetrahedron Lett.* 1987, 2347.
354. C. Petrier, J. L. Luche, S. Lavaitte, C. Morat, *J. Org. Chem.* **54** (1989) 5313 and references therein.
355. J. Nasielski, A. Standaert, R. Nasielski-Hinkens, *Synth. Commun.* **21** (1991) 901.
356. A. Lubineau, J. Auge, Y. Queneau, *Synthesis* 1994, 741.
357. A. Kumar, R. V. Salunkhe, R. A. Rane, S. Y. Dike, *J. Chem. Soc. Chem. Commun.* 1991, 485.
358. Ref. [52], p. 172 and references cited therein.
359. K. Senokuchi et al., *Synlett* 1994, 343.
360. G. Lunn, E. B. Sansone, *J. Org. Chem.* **51** (1986) 513.
361. H. Hauptmann, W. F. Walter, *Chem. Rev.* **62** (1962) 347.
362. S. Inaba, R. M. Wehmeyer, M. W. Forkner, R. D. Rieke, *J. Org. Chem.* **53** (1988) 339.
363. D. Savoia, E. Tagliavini, C. Trombini, A. Umani-Ronchi, *J. Org. Chem.* **46** (1981) 5340.

364. M. F. Semmelhack et al., *J. Am. Chem. Soc.* **103** (1981) 6460.
365. W. T. Smith, Jr., *J. Am. Chem. Soc.* **71** (1949) 2855.
366. R. D. Rieke, R. M. Wehmeyer, T. C. Wu, G. W. Ebert, *Tetrahedron* **45** (1989) 443.
367. T. Saegusa, Y. Ito, *Synthesis* 1975, 291.
368. A. Ballatore, M. P. Crozet, J. M. Surzur, *Tetrahedron Lett.* 1979, 3073.
369. R. G. Linck, Med. Tech. Publ., Inorg. Chem. Ser.1/9, 340, Butterworth, London 1972.
370. D. Lenoir, *Synthesis* 1989, 883.
371. P. W. Chum, S. E. Wilson, *Tetrahedron Lett.* 1976, 15.
372. J. E. McMurry, T. Hoz, *J. Org. Chem.* **40** (1975) 3797.
373. G. A. Olah, G. K. S. Prakash, *Synthesis* 1976, 607.
374. T. L. Ho, *Synthesis* 1979, 1.
375. H. Ishikawa, T. Mukaiyama, *Bull. Chem. Soc. Jap.* **51** (1978) 2059.
376. T. Inokuchi, H. Kawafuchi, S. Torii, *Chem. Lett.* 1992, 1895.
377. Y. Kataoka, I. Makihira, H. akiyama, K. Tani, *Tetrahedron Lett.* 1995, 6495.
378. T. Inokuchi, H. Kawafuchi, S. Torii, *J. Org. Chem.* **56** (1991) 4983.
379. P. M. Takahara, J. H. Freudenberger, A. W. Konradi, S. F. Pedersen, *Tetrahedron Lett.* 1989, 7177.
380. R. Annunziata et al., *Tetrahedron* 1991, 5737.
381. J. Park, S. F. Pedersen, *Tetrahedron* 1992, 2069.
382. B. Kammermeier et al., *Chem. Eur. J.* 1996, 307 and references cited therein.
383. M. SATO, K. Oshima, *Chem Lett.* 1982, 157.
384. J. Szymoniak, J. Besancon, C. Moise, *Tetrahedron* 1994, 2841.
385. B. Kammermeier et al., *Angew. Chem. Int. Ed. Engl.* **33** (1994) 685.
386. J. Szymoniak, J. Besancon, C. Moise, *Tetrahedron* 1992, 3867.
387. E. J. Roskamp, S. F. Pedersen, *J. Am. Chem. Soc.* **109** (1987) 6551.
388. E. J. Roskamp, S. F. Pedersen, *J. Am. Chem. Soc.* **109** (1987) 3152
389. Y. Kataoka et al. *J. Org. Chem.* **57** (1992) 1615.
390. Ref. [65], p. 119.
391. J. J. Brunet, P. Caubere, *Tetrahedron Lett.* 1977, 3947 and references cited therein.
392. E. E. van Tamelen, H. Rudler, C. Bjorklund, *J. Am. Chem. Soc.* **93** (1971) 7113.
393. G. A. Molander, C. R. Harris, *Chem. Rev.* **96** (1996) 307.
394. T. Hamamoto, Y. Sugimoto, A. Sugino, J. Inanaga, *Synlett* 1994, 377.
395. R. W. Marshman, *Aldrichimica Acta* **28**(1995)77.
396. J. A. Soderquist, *Aldrichimica Acta* **24** (1991) 1.
397. M. Sekiguchi et al., *Heteroat. Chem. Int. conf. 2nd* 1991, 427.
398. J. Collin, C. Bied, H. B. Kagan, *Tetrahedron Lett.* 1991, 629.
399. T. Kan, F. Matsuda, M. Yanagiya, H. Shirahama, *Synlett* 1991, 391.
400. T. HAMAMOTO, Y. SUGIMOTO, A. SUGINO, J. INANAGA, *Synlett* 1994, 377.
401. R. W. MARSHMAN, *Aldrichimica Acta* **28**(1995)77.
402. H. H. Hodgson, H. S. Turner, *J. Chem.* Soc. 1948, 318.
403. *Merck Index* **vol. 11,** p. 4729.
404. D. Huber, G. Andermann, G. Leclerc, *Tetrahedron Lett.* 1988, 635.
405. T. E. Nickson, *J. Org. Chem.* **51** (1986) 3903.
406. Ref. [65], pp. 127 ff.
407. K. Fujimori, H. Yoshimoto, S. Oae, *Tetrahedron Lett.* 1979, 4397.
408. R. Budziarek, *Chem. Ind. (London)* 1978, 583.
409. H. Suzuki, H. Manabe, M. Inouye, *Chem. Lett.* 1985, 1671.
410. Ref. [65], p. 121.
411. R. Sangaiah, A. Golg, *J. Org. Chem.* **52** (1987) 3205.
412. Ref. [52], p. 8.
413. D. J. Pasto in Ref. [8], p. 471.
414. W. Adam, H. J. Eggelte, *J. Org. Chem.* **42** (1977) 3987.
415. N. S. Kozlov, V. A. Soshin, *Chem. Abstr.* **66** (1967) 2290 k.
416. M. G. Sarngadharan, T. R. Seshadri, *Indian J. Chem.* **4** (1966) 503.
417. Y. Shigemasa et al., *Chem. Express* **6** (1991) 849.
418. Lancaster Synthesis Ltd., Lancaster Catalogue, 1988.
419. M. Brink, *Synthesis* 1975, 807.
420. S. K. Boyer, J. Bach, J. McKenna, E. Jagdmann, Jr., *J. Org. Chem.* **50** (1985) 3408.
421. W. E. Parham, Y. A. Sayed, *Synthesis* 1976, 116.
422. G. D. Pandey, K. P. Tiwari, *Synth. Commun.* **9** (1979) 895.
423. C. T. Ratcliffe, G. Pap, *J. Chem. Soc. Chem. Commun.* 1980, 260.
424. M. Akazome, T. Kondo, Y. Watanabe, *J. Chem. Soc. Chem. Commun.* 1991, 1466.
425. *Chem. Week* **121** (1977) Nov. 9, 57.

426. H. W. Gibson, *Chem. Rev.* **69** (1969) 673.
427. T. van Es, B. Staskun, *Org. Synth.* **51** (1971) 20.
428. Ref. [53], p. 102.
429. A. C. Cope, W. D. Burrows, *J. Org. Chem.* **31** (1966) 3099.
430. T. Okano, Y. Moriyama, H. Konishi, J. Kiji, *Chem. Lett.* 1986, 1463.
431. R. M. Kellogg in Ref. [8], p. 79.
432. M. G. Romanelli, E. I. Becker, *Org. Synth.* **47** (1967) 69.
433. K. Schank, *Synthesis* 1972, 176.
434. V. Hach, *J. Org. Chem.* **38** (1973) 293.
435. I. Shimao, *Nippon Kagaku Kaishi* 1974, 515.
436. I. D. Entwistle, W. W. Wood in Ref. [8], p. 977.
437. S. G. Cohen, G. H. Parson, *Chem. Rev.* **73** (1973) 141.
438. F. W. Hoffmann, R. J. Ess, T. C. Simmons, R. S. Hanzel, *J. Am. Chem. Soc.* **78** (1956) 6414.
439. O. P. Ward, C. S. Young, *Enzyme Microb. Technol.* **12** (1990) 482.
440. S. Ghisla, P. Kroneck, *Nachr. Chem. Tech. Lab.* **39** (1991) 834.
441. H. Simon et al., *Angew. Chem.* **97** (1985) 541.
442. H. U. Reißig, *Nachr. Chem. Tech. Lab.* **34** (1986) 782.
443. S. Servi, *Synthesis* 1990, 1.
444. J. B. Jones, *Tetrahedron* **42** (1986) 3351.
445. J. S. Dordick, *Enzyme Microb. Technol.* **11** (1989) 194.
446. N. J. Turner, *Chem. Ind.* 1994, 592.
447. P. Besse, H. Veschambre, M. Dickmann, R. Chenevert, *J. Org. Chem.* **59** (1994) 8288.
448. J. B. Jones, *Aldrichimica Acta* **26** (1993) 105.
449. H. Sies, *Angew. Chem.* **98** (1986) 1061.
450. F. Schnitger, *Selecta* **34** (1992) 18.
451. G. Beck, A. Bergman, K. Keßeler, G. Wess, *Tetrahedron Lett.* 1990, 7293.
452. International Conference on *Antioxidant Vitamins and β-Carotene in Disease Prevention,* London 1989.
453. H. K. Biesalski, *Münch. med. Wschr.* **138** (1996) 315.
454. H. B. Stähelin, *Münch. med. Wschr.* **138** (1996) 321.
455. U. Müller, J. Krieglstein, *Deutsche Apotheker Zeitung* **134** (1994) 1677.
456. A. Zittermann, B. Kling-Steines, *Deutsche Apotheker Zeitung* **134** (1994) 2991.
457. B. Hellwig, *Deutsche Apotheker Zeitung* **136** (1996) 269.
458. Y. C. Liu, Z. L. Liu, Z. X. Han, *Rev. Chem. Intermed.* **10** (1989) 269.
459. R. C. Mordi, *Chem. Ind.* 1993, 79.
460. R. Bitsch, *Naturw. Rdsch* **49** (1996) 47.
461. D. Harman, *J. Gerontol.* **11** (1956) 298.
462. H. D. Osiewacz, *Biol. Unserer Zeit* **25** (1995) 336.
463. R. S. Sohal, R. Weindruch, *Science* **273** (1996) 59.
464. K. Hensley, et al. *Proc. Natl. Acad. Sci. USA* **91** (1994) 3270.
465. S. Borman, *Chem. Eng. News* April 18, 19844.
466. *Chem. Ind.* 1994, 286.
467. T. Thomas, et al., *Nature* **380** (1996) 168.
468. E. Block, *Angew. Chem.* **104** (1992) 1158.
469. H. Isensee, B. Rietz, R. Jacob, *Arzneim-Forsch./Drug Res.* **43(I)** (1993) 94.
470. F. Freeman, B.-G. Huang, R. I-San Lin, *J. Org. Chem.* **59** (1994) 3227.
471. G. G. Börger, W. Hüning, M. Schwetzer, *Chem. Ing. Tech.* **50** (1978) 439.
472. E. Herrmann, W. Weiweiler, *Chemie-Ingenieur-Technik* **64** (1992) 232.
473. P. Necker, *Haus Tech. Vortragsveroeff.* **500** (1986) 50.
474. R. Hamann: *Thermo- Prozess- und Abfalltechnik,* 2nd ed.Vulkan-Verlag, Classen1988p. 289.
475. G. Maletz, C. Baemler, *Forschungsbericht der Sued Chemie,* München 1985.
476. "Katalytische Beseitigung von Stickoxiden aus Abgasen,"*Industrie-Anzeiger* **106** (1984) 26.
477. J. Jacobs, *VGB-Kraftwerkstechnik* **61** (1981) 214.
478. E. W. Haltiner, *Umweltmagazin* **21** (1992) 34.
479. M. J. Kline, Y.C. Lee, P. H. Owen, Environmental Protection Agency, Washington D.C. 1976, EPA 600/7-76 031.
480. "Entstickung durch Hochspannung," *GIT Fachz. Lab.* **36** (1992) IV.
481. G. Lepperhoff, J. Schommers, *Motortechnische Zeitschrift* **49** (1988) 17.
482. U. Kircher, *Verminderung der Schadstoffemission,* Vulkan-Verlag, Classen 1993, p. 5.
483. G. Maletz, *Forschungsbericht der Sued Chemie,* München 1984.
484. V. A. Golodov, *Koord. Khim.* **18** (1992) 750.

Refinery Technology → Oil Refining
Refractories → Refractory Ceramics

Refractory Ceramics

Gerald Routschka, Forschungsinstitut der Feuerfest-Industrie, Bonn, Federal Republic of Germany (Chaps. 1-6, 8-10, Sections 7.1, 7.3)

Karl-Ernst Granitzki, Königswinter, Federal Republic of Germany (Section 7.2)

1. Introduction 2
2. Historical Aspects 2
3. General Classification and Nomenclature 3
4. Raw Materials 3
5. Production 4
5.1. Production of Coarse-Grained Refractory Products 4
5.2. Production of Fine-Grained Refractory Ceramics 9
5.3. Brick Dimensions and Dimensional Tolerances 10
5.4. Packaging, Storage, and Storage Time 10
5.5. Quality Assurance 10
6. General Composition, Properties, and Testing 10
7. Types of Refractory Ceramics . . . 20
7.1. Dense, Shaped Refractory Products 20
7.1.1. Silica Bricks and Fused-Silica Ceramics 20
7.1.2. Fireclay and High-Alumina Bricks . 22
7.1.3. Basic Bricks 25
7.1.4. Zircon- and Zirconia-Containing Bricks 28
7.1.5. Carbon and Graphite Bricks 29
7.1.6. Carbon and Graphite-Containing Refractories 30
7.1.7. Silicon Carbide Bricks 32
7.1.8. Fine-Grained Oxide and Nonoxide Ceramics 32
7.1.9. Fusion-Cast Bricks 33
7.1.10. Refractory Ceramics with Low Thermal Expansion 33
7.2. Unshaped Refractory Materials . . 35
7.3. Heat-Insulating Ceramic Materials 38
7.3.1. Insulating Bricks 39
7.3.2. Insulating Refractory Bricks 39
7.3.3. Refractory Ceramic Fibers 39
8. Selection, Use, and Industry-Specific Consumption of Refractory Ceramics 42
9. Toxicology and Occupational Health 42
10. Economic Aspects 43
11. References 44

1. Introduction

Definitions. According to ISO/R 836, refractory materials are nonmetallic materials or products (including those containing a proportion of metals) having a minimum pyrometric cone equivalent of 1500 °C. Determination of the pyrometric cone equivalent (refractoriness) is described in ISO/R 528. Test pieces of the refractory material in the form of small cones are raised in temperature alongside reference pyrometric cones (blunt-tipped, skewed, triangular pyramids) with a defined temperature of collapse (softening).

Certain products from the refractory industry do not comply with this definition, however, but still fall under refractory classifications and are therefore conventionally termed refractories.

In the common sense, refractories are ceramic construction materials capable of withstanding temperatures above 800 °C. Numerous refractory compositions in a wide variety of shapes and forms are used in a broad range of industrial applications. The three main groups of refractory ceramics are:

1) Shaped refractory products (bricks)
2) Unshaped refractory products (monolithics)
3) Ceramic fiber products

Their three main applications are:

1) The construction of reaction chambers (e.g., furnaces)
2) Thermal insulation
3) Recycling of heat in regenerators or recuperators

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_001

The temperature of use for "hot face" linings ranges from ca. 800 to 1800 °C, and sometimes to 2000 °C or higher.

In addition to high temperature resistance, other essential properties for refractories are resistance to temperature change (thermal shock) and to mechanical and corrosive influences.

2. Historical Aspects

It is believed that the first baked refractory bricks were made from quartz-containing clays, either by the Phoenicians or the Chinese. For several centuries such prerefractory products were sufficient to meet the modest thermal requirements of small glass- or metal-melting furnaces. At the beginning of the nineteenth century the idea of adding prefired clay – the so-called *grog* material – to clay was shown to result in bricks with better dimensional accuracy and stability because of low shrinkage during firing. With this development, "refractory" technology was separated from the heavy clay industry, and special refractory brick works were established.

Historical Dates.

- Ca. 1820: The development of silica bricks for the iron and steel industry began, and fireclay brick was introduced in gas retorts
- Ca. 1860: The application of magnesite as a refractory material was proposed; the magnesite brick was developed about 20 years later in Austria
- Ca. 1870: Use of low-iron bauxite for manufacturing refractory crucibles
- Ca. 1880: Tar – dolomite products for Thomas steel converters and chrome ore bricks for the open-hearth furnace were developed
- 1892: Successful production of SiC in the United States
- 1914: Development of a plastic fireclay mix
- Ca. 1925: First manufacture of a hydraulic (cement) bonded mix (castable)
- 1928: Production of fused cast bricks began in the United States
- Ca. 1945: Earliest development of ceramic fibers and fiber products
- 1969: Initial development of low-cement castables in France
- Early 1980s: Introduction of magnesia–graphite bricks in the steel industry in Japan

The development of the refractory industry was similar to that of other industries. Initially there were only small workshops using manual methods, with process automation and scientific research arising much later. Since the early 1960s, a steadily increasing demand by the consumer industries has led to higher quality, tailor-made refractory products and cost-reducing application techniques.

3. General Classification and Nomenclature

Classifications of refractory products generally make reference only to characteristic limits for specific property values. The nomenclature and forms of designation are derived from features such as raw material base, manufacturing process, nature of bond, and typical application. The main classification criteria are:

1) Chemical and/or mineralogical constitution (see Table 1, ISO 1109-1975)
2) Total porosity
   Dense refractory products: < 45 vol %
   Insulating refractory products: > 45 vol %
3) Form of delivery
   Shaped products (bricks and shapes)
   Unshaped products (monolithics)
   Fibrous products (ceramic fiber products)
4) Type of bonding
   Bonding at < 150 °C; unfired products
   Ceramic bond (clay bond)
   Inorganic chemical bond
   Hydraulic bond
   Organic bond
   Bonding at 150 to 800 °C; thermally treated (tempered) products
   Inorganic chemical bond
   Hydraulic bond
   Bonding at > 800 °C
   Firing (ceramic bond)
   Fusion and solidification
5) Application field

Table 1. Classification of refractory products (ISO 1109 – 1975)

Products	Limiting content of principal constituent	Criteria of subdivision and general observations
High-alumina products Group 1	$Al_2O_3 \geq 56\,\%$	a complete designation of these products shall include an indication of either the raw material actually used or the mineralogical constitution of the final product. In the latter case, the method of ascertaining this constitution shall be stated
High-alumina products Group 2	$45\,\% \leq Al_2O_3 < 56\,\%$	
Fireclay products	$30\,\% \leq Al_2O_3 < 45\,\%$	
Low-alumina fireclay products [a]	$10\,\% \leq Al_2O_3 < 30\,\%$ $SiO_2 < 85\,\%$	
Siliceous products [a]	$85\,\% \leq SiO_2 < 93\,\%$	
Silica products	$SiO_2 \geq 93\,\%$	quality specifications according to use
Basic products		new subdivisions and new criteria of classification may become necessary
Magnesite (magnesia)	$MgO \geq 80\,\%$	products in which the principal constituent is magnesite (magnesia)
Magnesite- chrome (magnesia chrome)	$55\,\% \leq MgO < 80\,\%$	products in which the principal constituents are magnesite and chromite (magnesia)
Chrome-magnesite (chrome magnesia)	$25\,\% \leq MgO < 55\,\%$	products in which the principal constituents are chromite and magnesite (magnesia)
Chromite	$Cr_2O_3 \geq 25\,\%$ $MgO \leq 25\,\%$	products in which the principal constituent is chromite
Forsterite		products in which the principal constituent is forsterite
Dolomite (doloma)		products in which the principal constituent is dolomite (doloma)
Special products		products based on: carbon, graphite, zircon, zirconia, silicon carbide, carbides (other than silicon carbide), nitrides, borides, spinels (other than chromite)
		products based on several oxides (other than those of basic products)
		products based on pure oxides, including alumina, silica, magnesia, zirconia; products of high purity

[a] Products designated in certain countries by the name "semi-silica products" may belong either to the low-alumina fireclay class or to the siliceous products class.

4. Raw Materials [10], [16–19]

The refractories industry consumes a wide variety of raw materials, binders, and special additives.

Most of the raw materials are based on naturally occurring beneficated, upgraded, crushed minerals, usually acquired in captive production by the refractories manufacturer. Natural raw materials subject to large weight losses or volume changes when heated are prefired in rotating, shaft, or tunnel kilns before use (exception: quartz raw materials). The crushed, fired material (grog, grain) normally has a true porosity $< 10\,\text{vol}\,\%$. Natural raw materials that can be used without prefiring are listed in Table 2.

Several high-purity synthetic raw materials have become important in recent years because of the increasing demand for high-performance refractories. Synthetic raw materials are produced by special suppliers, and often supplied with defined particle-size distributions or as fine powders. To produce a grain from fine powders the synthetic materials are sintered in the form of briquettes at high temperature (up to 1800 °C or higher) or fused in electric arc furnaces and crushed. Fused grain materials have very low porosity ($< 3\,\text{vol}\,\%$) and large crystal sizes.

Porous raw materials and *ceramic fibers* are used for the production of thermal insulating refractories.

The grain or aggregate for a refractory is coarse and crystalline. Raw materials for the fines (i.e., for the matrix) are often chosen or prepared in less chemically stable forms (e.g., as "reactive calcines") and finely sized (sometimes $< 1\,\mu$m).

Refractory raw materials are subject to various types of quality control:

clays	**grain**
chemical analysis moisture content mineralogical composition refractoriness grain size distribution	
plasticity	grain density
deflocculant behavior	grain porosity
firing behavior (shrinkage, expansion)	crystallite size sintering behavior (fines)

In most cases, a prepared noncoherent mix of the refractory raw material (or several different

Table 2. Principal raw materials for refractory ceramics

Type	Main mineral phase(s) (minor phases)	Essential chemical components (ignited)	Remarks
Quartzite, silica sand	quartz (SiO_2)	$>97\,\%$ SiO_2, $<1.5\,\%$ Al_2O_3, low alkali content	used unfired; on firing, conversion to cristobalite and tridymite (increase in volume)
Fused silica	glass	$>99.8\,\%$ SiO_2	very low thermal expansion up to 1000 °C, above 1050 °C conversion to cristobalite
Loomy sand	quartz, clay minerals	$>83\,\%$ SiO_2, $4-12\,\%$ Al_2O_3	a clay-bonded sand (used as unshaped refractory material)
Refractory clays	kaolinite [$Al_2SiO_5(OH)_4$] and other clay minerals, quartz	$20-45\,\%$ Al_2O_3, $<3\,\%$ Fe_2O_3, $<3\,\%$ Na_2O+K_2O	special types: kaolin, flint clay, ball clay; on firing, conversion to mullite, cristobalite, and a glassy phase; loss on ignition $6-14\,\%$
Chamotte (burned clay)	mullite ($3\,Al_2O_3 \cdot 2\,SiO_2$), cristobalite ($SiO_2$), quartz, glassy phase	see above	firing temperature ca. 1200 – 1500 °C depending upon Al_2O_3 and alkali content, grain total porosity <10 vol %, special type: porous chamotte, bloated clay ($\sim$5 % Fe_2O_3)
Sillimanite, andalusite, kyanite	$Al_2O_3 \cdot SiO_2$ (different crystal structures)	$60-65\,\%$ Al_2O_3, low alkali content	above 1250 °C beginning of conversion to mullite, SiO_2, and small amount of glassy phase; high volume increase when firing kyanite; sillimanite and andalusite can be used unfired
Mullitized raw materials	mullite	$50-70\,\%$ Al_2O_3	fired mixtures of kaolinitic clay and refractory bauxite (Al-hydrate, $Al_2O_3 \cdot 2\,H_2O$)
Sintered mullite (synthetic)	$3\,Al_2O_3 \cdot 2\,SiO_2$	$72-75\,\%$ Al_2O_3	grain total porosity $\sim 10\,\%$, crystal size $10-150\,\mu m$ speciality: mullite hollow spheres [20]
Fused mullite	$2\,Al_2O_3 \cdot SiO_2$	$\sim 78\,\%$ Al_2O_3	grain total porosity $\sim 2\,\%$, crystal size $100-2000\,\mu m$
Sintered bauxite (refractory bauxite)	corundum (Al_2O_3)	$>85\,\%$ Al_2O_3, $<4\,\%$ TiO_2, $<2\,\%$ Fe_2O_3, low alkali content	producing countries: Guyana, China, Brazil; grain total porosity <10 vol %
Sintered corundum	corundum	$>99.5\,\%$ Al_2O_3, $<0.5\,\%$ Na_2O	grain total porosity $\sim 10\,\%$
Fused white corundum			grain total porosity $\sim 2\,\%$
Fused brown corundum	corundum	$94-97\,\%$ Al_2O_3, $3\,\%$ TiO_2	fused bauxite
Calcined aluminas	γ-Al_2O_3, corundum	$>99.5\,\%$ Al_2O_3, $<0.5\,\%$ Na_2O	fine powders ($<100\,\mu m$)
Corundum hollow spheres (2 – 4 mm diameter)	corundum	$>99.5\,\%$ Al_2O_3	fused Al_2O_3 sprayed by air stream (thermal insulating material)
Sintered magnesia (→ Magnesium Compounds, Chap. 4.)	periclase (MgO) dicalcium silicate, forsterite, calcium magnesium silicates	$85-99\,\%$ MgO, $0.1-3.5\,\%$ CaO, $0.1-3\,\%$ SiO_2, $0.1-8\,\%$ Fe_2O_3, $<0.3\,\%$ B_2O_3	firing of magnesite ($MgCO_3$) or $Mg(OH)_3$ from brines or seawater (→ Magnesium)
Fused magnesia		$97-99.8\,\%$ MgO$<2\,\%$ CaO,$<1\,\%$ SiO_2,$<0.2\,\%$ Fe_2O_3	grain total porosity $<8\,\%$, crystal size $60-200\,\mu m$ [21]grain total porosity $<2\,\%$, crystal size $400-3000\,\mu m$ [22]
Sintered dolomite	periclase (MgO), CaO	$56-62\,\%$ MgO, $36-40\,\%$ CaO, $<2\,\%$ SiO_2	firing of dolomite ($MgCO_3 \cdot CaCO_3$), prone to hydration
Sintered and fused CaO	CaO	$>98\,\%$ CaO	calcined calcite ($CaCO_3$)
Chrome ore (chromite)	$(FeO, MgO) \cdot (Cr_2O_3, Al_2O_3, Fe_2O_3)$ spinel	$10-20\,\%$ MgO, $32-50\,\%$ Cr_2O_3, $12-27\,\%$ Al_2O_3, $<6\,\%$ SiO_2	deposits: Turkey, Iran, Phillipines, South Africa
Magnesia – chrome sinters	periclase, spinel	normally up to $15\,\%$ Cr_2O_3	prereacted granulated mixtures of magnesite (magnesia) and chromite or dolomite
Magnesia – dolomite sinters	periclase, CaO	MgO and CaO in varying content	
Picochromite	spinel ($MgO \cdot Cr_2O_3$)	$19\,\%$ MgO, $79\,\%$ Cr_2O_3	fused mixture of magnesia and Cr_2O_3 powder

Table 2. (Continued)

Sintered and fused spinel	spinel ($MgO \cdot Al_2O_3$)	66 – 78 % Al_2O_3, 21 – 33 % MgO	synthetic material made from fine magnesia and alumina
Olivine, dunite	forsterite ($2\,MgO \cdot SiO_2$) – fayalite ($2\,FeO \cdot SiO_2$) – solid solution	40 – 50 % MgO, 35 – 45 % SiO_2, < 8 % Fe_2O_3	used unfired
Zirconium silicate	zircon ($ZrSiO_4$)	66 % $ZrO_2 + HfO_2$, 33 % SiO_2	extracted from sands (Australia), grain size < 0.3 mm, at temperatures above 1450 °C, decomposition to baddeleyite (ZrO_2) and SiO_2
Zirconia	baddeleyite (ZrO_2) (monocline)	> 99 % $ZrO_2 + HfO_2$	powders, chemically processed from zircon
Stabilized or partly stabilized ZrO_2	cubic/tetragonal phase of ZrO_2	> 94 % ZrO_2, 2 – 6 % CaO, MgO, Y_2O_3	stabilization because of irregular thermal expansion; ground, sintered, or fused material
Chromium oxide	escolaite (Cr_2O_3)	> 98 % Cr_2O_3	fine powder, produced chemically
Silicon carbide [23]	SiC (Si, C)	> 93 % SiC	produced from SiO_2 and C by the Acheson process
Natural graphite [24]	graphite (C), (clay minerals, mica, etc)	75 – 97 % various ash compositions	mainly two types : – flake graphite up to 5 cm crystal diameter – amorphous
7 Coke, anthracite, electrographite		75 – 99 % C	
Diatomite (moler)	opaline hydrous SiO_2, quartz (clay)	> 75 % SiO_2	fine porous siliceous rock (skeletons of aquatic plants)
Perlite	glass	65 – 80 % SiO_2, 12 – 16 % Al_2O_3, high flux content	prefired, hollow, spheric, glassy ingeneous rock
Vermiculite	magnesium, aluminum silicates (similar to mica)	45 % SiO_2, 31 % MgO, 13 % Al_2O_3, 7 % Fe_2O_3	exfoliated by heating, extremely low weight
Pyrophillite	$Al_2Si_4O_{10}(OH)_2$	20 – 30 % Al_2O_3, low alkali content	used unfired, loss on ignition 3 – 7 %; little expansion when heated; formation of mullite and SiO_2
Cordierite [25]	$2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$ (mullite)	31 % Al_2O_3, 53 % SiO_2, 10 % MgO	prefired grog from fine-grained magnesia and talc, low thermal expansion
Recycled refractory ceramics (from used refractory linings)		different compositions	recycled raw materials mainly from used, high-priced refractory products examples: fused mullite – zirconia grain from fused cast bricks, SiC- and zirconia-bearing materials
Ceramic fibers	glassy corundum/mullite (ZrO_2, Cr_2O_3)	system $Al_2O_3 - SiO_2$ 45 – 95 % Al_2O_3	two types available : glassy (amorphous) and polycrystalline
	ZrO_2	> 95 % ZrO_2	(see Section 7.3.2)

raw materials) is bonded by a suitable *binder* or binder combination (Table 3) to ensure sufficient strength at all steps during production or application of the refractory product. Even a small excess of binder may have a detrimental effect on hot strength and corrosion resistance because of its concentration in the matrix (fines).

Special *additives* (Table 4) are used to increase the workability and formability of a refractory mix or to modify refractory structure and the properties of the fired material.

Recycling of used refractory ceramics is complicated by the various contaminants introduced into refractory linings during use. Nevertheless, disposal problems and environmental considerations are expected to encourage efforts toward increased recycling.

5. Production

See also → Ceramics, General Survey.

Various methods are employed to produce refractory ceramics of different types:

1) Refractory products with large grain sizes, normally ≤ 6 mm (Section 5.1), in special cases ≤ 20 mm (pitch- and resin-bonded bricks, Section 7.1.6).

Table 3. Major binders for refractory mixes [6], [26] (often used in combination with water and as binder combinations)

Type of binder	Amount added	Remarks
Clays (plastic clays, ball clays)	$\leq 25\,\%$	strength after drying, ceramic bond at $> 1000\,^{\circ}C$ (formation of a melt, enhancing sintering)
High-alumina cements, calcium aluminate cements [27]	$\leq 25\,\%$	$50-80\,\%$ Al_2O_3, $19-38\,\%$ CaO, hydraulic bond, effectiveness between $20-600\,^{\circ}C$ (dehydration), $> 1000\,^{\circ}C$ ceramic bond
Monoaluminum phosphate (solution)	$\leq 5\,\%$ P_2O_5	hardening at about $200\,^{\circ}C$ by condensation reaction; at ca. $1300\,^{\circ}C$ formation of $AlPO_4$; $> 1500\,^{\circ}C$ onset of vaporization of P_2O_5
Inorganic phosphates, mainly sodium	$\leq 5\,\%$	used mainly in hot-repair mixes, hardening and adhesive strength by melting of the glassy phosphate
Alkali silicates $Na_2O(K_2O) \cdot x\,SiO_2 \cdot y\,H_2O$	$\leq 10\,\%$	normally used as solution, hardening at room temperature by CO_2 reaction or by hardener, thermosetting $> 150\,^{\circ}C$
Ethyl silicate $Si(C_2H_5O)_4$ [28]		used for slurry casting, gelation; curing at ca. $200\,^{\circ}C$
Magnesium sulfate ($MgSO_4 \cdot 7\,H_2O$ solution)	$\sim 5\,\%$	bond for unfired basic magnesitic refractories
Aluminum sulfate	$\sim 5\,\%$	mainly used in plastic mixes to increase drying strength (crystallization of aluminum sulfate)
Colloidal silica (silica sol)		binder mainly for ceramic fibers
Pitch, tar	$< 10\,\%$	organic bond, effectiveness at $20-450\,^{\circ}C$, coking at $600\,^{\circ}C$ (carbon bond)
Various synthetic resins [29]	$< 10\,\%$	modified phenolic resins in powder form or in solution of, e.g., formaldehyde; thermosetting resins (with hardener, hardening at ca. $200\,^{\circ}C$); $> 600\,^{\circ}C$: carbon bond
Variety of organic temporary binders and organometallic binders	$< 5\,\%$	to achieve sufficient strength after drying; burns out; organometallic binders: on heating, formation of sinteractive fine oxides
Boron oxide, other boron compounds	$< 10\,\%$	sintering aid (melt), especially for dry silica mixes
Frits, glass	$< 5\,\%$	mainly for unshaped refractories, low-melting flux- containing powders, to achieve ceramic bond $< 1000\,^{\circ}C$

Table 4. Special additives (normally $< 5\,\%$)

Type of additive	Purpose
Powders of nonoxides (C, Si_3N_4, SiC, BN, AlON, B_4C) and salts (e.g., $BaSO_4$)	to increase corrosion resistance, to reduce wettability by slags and metal melts; carbides also used as antioxidants
Reactive aluminas [10]	improvement of matrix (bond), to increase flowability of castables
Fine silica flour (silica fume) [30]	mainly used in unshaped refractories (castables) to increase flowability and hot strength (mullite formation)
Metal powders (Si, Al, Mg, or alloys)	antioxidants for carbon- containing refractories; to create a special nitridic or carbidic bond
Steel fibers [31]	reinforcement of unshaped refractories
Organic fibers [32]	to eliminate explosion hazard when heating castables
Organic grained materials	to create defined porosity or pore size
Various organic compounds	to increase formability of a mix; lubricants, surface-active substances (addition of $< 0.5\,\%$)
Ethylene glycol	antifreezing material for unshaped refractories

2) Special fine-grained refractory products (fine ceramics, Section 5.2).
3) Products formed via a melt (i.e., fused cast shapes, Section 7.1.9).
4) Ceramic fiber products (Section 7.3.2).

5.1. Production of Coarse-Grained Refractory Products

The production process includes the following steps:

1) Preparation
2) Shaping
3) Drying
4) Thermal treatment up to $800\,^{\circ}C$, or firing
5) Aftertreatment (if necessary)
6) Packing

A flow diagram for the production of dry-pressed fireclay bricks is shown in Figure 1. In order to achieve uniform products most manufacturing steps are now controlled electronically.

The fabrication of unshaped refractory products ends after mixing (Section 7.2), but plastic mixes can also be preformed into blocks using an

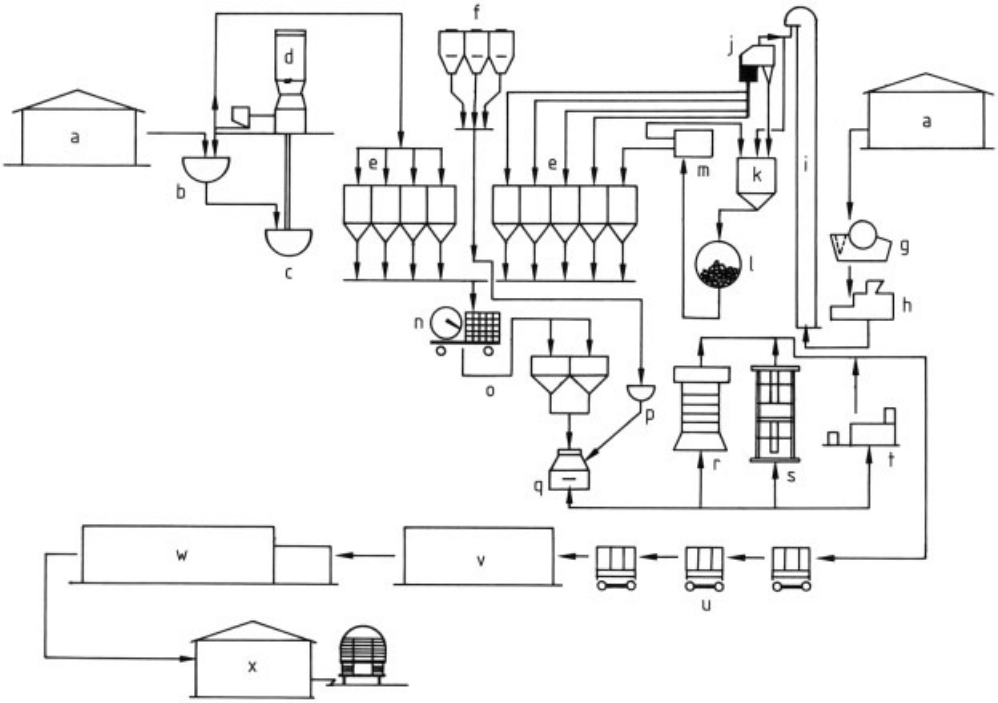

Figure 1. Flow-sheet for the production of dry-pressed fireclay bricks
a) Raw material storage; b) Clay crusher; c) Clay mill; d) Drying tower/classifier; e) Storage silos; f) Mixing of liquid components; g) Coarse crusher; h) Fine crusher; i) Bucket elevator; j) Screen; k) Feed hopper; l) Ball mill; m) Air classifier; n) Weighing scales; o) Feed hopper; p) Dosing of liquid components; q) Mixer; r) Friction press; s) Hydraulic presses; t) Hand shaping/rammer; u) Kiln car; v) Tunnel dryer; w) Tunnel kiln; x) Brick storage/loading

extruder. Other prefabricated shapes are manufactured from appropriate unshaped products.

Preparation. Various devices are used to comminute the raw materials; e.g., jaw crushers, impact crushers, ball mills, and vibratory mills (see also → Size Reduction). The crushed materials are then classified generally on vibrating screens.

These raw material components, a binder, additives, and water (up to 20 %) are fed into a mixer, yielding a moldable mix. Intermittent mixers are usually employed, but muller mixers and pug mills are also applicable. Mixing time, intensity, and sequence have a pronounced influence on the properties of the mix as well as on the product.

The properties of refractory ceramics can also be influenced by the grain-size distribution of the batch and grain sizes of the raw materials. Computational models have been used to establish graded grain sizings for dense packing [33]. Normal practice is to use three or four grain fractions (coarse, medium, fine, and finest). The fine (< 0.1 mm) and finest grains (required, for instance, in "binderless" batches) constitute the matrix of the fired product, which plays an especially important role in corrosion resistance. To increase thermal shock resistance the amount of coarse grain is increased at the expense of medium grain. Table 5 shows typical grain sizings for refractory bricks and castables.

Shaping. The preferred shaping method depends on the properties of the moldable mix (water content, plasticity), the required properties of the product, the number of pieces to be produced, and shape complexity. The following shaping methods are used:

1) Extrusion (pug mill) for plastic mixes (high clay, water content ≤ 8 %), and hydraulic, impact, friction, and rotary table presses (for "dry" mixes with 2 – 6 % water); pressures are normally 40 – 120 N/mm^2

Table 5. Grain sizings for typical refractory ceramics

Refractory	Grain-size fraction, %			Binder/water/additive content
	1 – 5 mm	0.1 – 1 mm	< 0.1 mm	
Silica brick	40	25	35	lime/2 % water/1.5 % sulfite lye
Fireclay brick (dry pressed)	35	25	40 (including clay)	15 % clay/5 % water
Magnesia brick	45	20	35 (10 % < 0.06 mm)	2 % water/sulfite lye
Conventional castable	25	30	35 (including cement)	15 % cement/9 % water (when used)
Low- cement castable	35	25	40 (10 % < 0.01 mm)	5 % cement/5 % fine oxide powder/5 % water (when used)

2) Vibration presses with low surcharge; mainly for prefabricated parts of unshaped refractories
3) Manual or mechanical ramming for complex shapes and small numbers of produced shapes
4) Isostatic pressing; mainly for fine grained high-quality products, with pressures $\leq 300\,\mathrm{N/mm^2}$
5) Slip casting; mainly for fine ceramics and special shapes, water content of slip 10 – 20 %

Drying. After shaping, bricks must generally be dried in drying chambers or tunnel dryers, taking advantage of waste heat from the firing kilns. Only completely dry shapes can be fired without danger of cracking. The drying process can take from several days to several weeks depending upon brick size and the drying sensitivity of the mix. Large shapes are dried under the controlled humidity of the drying air.

Thermal Treatment. Chemical (phosphate), organic, and hydraulic bonded shapes are often heated up to 600 °C to enhance chemical bonding (for higher strength) and to remove water of hydration or volatile organic components (e.g., in a pitch bond; tempering, coking).

Firing. During firing the characteristic ceramic microstructure develops as a result of reactions, recrystallization, and transformation or formation of a liquid phase or melt, respectively (at room temperature this phase consolidates to a glassy phase). With increasing firing temperature the porosity usually decreases, whereas both crystallite size and strength increase. These processes are generally associated with contraction or expansion, so the dimensions of a brick often change during firing, a factor that must be considered when molds are made. Indeed, the firing temperature may be limited by shrinkage or other deformation that would exceed the specified dimensional tolerances.

Various types of kilns (e.g., chamber, hood, and tunnel kilns) are used for firing, fueled by either gas or oil. Firing temperatures and energy requirements for various types of refractory brick types are as follows:

Fireclay bricks	1250 – 1500 °C	ca. 3500 kJ/kg
Silica bricks	1450 – 1500 °C	
High-alumina bricks	1500 – 1800 °C	≤ 1600 °C, ca. 6000 kJ/kg
Magnesia bricks	1500 – 1800 °C	≤ 1800 °C, ca. 10 000 kJ/kg

The firing and cooling curves, and sometimes also the kiln atmosphere (e.g., for magnesia chrome or chromia bricks) are adapted to suit the fired wares. Total firing time, including preheating and cooling, ranges from about three days (tunnel kilns) to about three weeks.

Aftertreatment. For special applications, fired bricks (especially larger shapes or blocks) may be sawn or ground as a way of achieving high dimensional accuracy.

In order to improve slag resistance, basic bricks for the steel industry are impregnated under vacuum with tar, pitch, or resin. Basic bricks can also be provided with steel cladding.

Production of Insulating Refractory Products and Other Porous Refractory Ceramics. To achieve a high porosity (> 45 vol %) four strategies are invoked:

1) Addition of burn-out or volatilizing substances to the mix (e.g., sawdust, coke, granulated plastic)

2) Use of porous raw materials such as diatomite, perlite, vermiculite, bloated clay, or hollow spheres (see Table 1)
3) Foam formation by gas-producing agents (metal powders)
4) Mixing with stabilized foams from colophonium soap or saponin

Most insulating bricks are pressed fireclay products produced with burn-out additives. Unshaped insulating refractory materials contain only porous raw materials.

Other porous refractory ceramics are the so-called porous plugs and ceramic filters (foams) for the steel industry. Unidirectional porosity is achieved by embedding burn-out plastic filaments. In the case of ceramic foam production, the pores of a burn-out sponge are covered with a ceramic suspension (slip).

5.2. Production of Fine-Grained Refractory Ceramics

Fine-grained refractory ceramics are used only to a minor extent, and for special purposes. These are mostly small items, often characterized by specific forms (e.g., plates, tubes, or crucibles), and they usually display little or negligible porosity. Only occasionally there is a demand for large blocks or other objects.

The starting materials for fine-grained ceramics are almost exclusively pure, finely-ground raw materials of the types used in the production of coarse-grained refractories, but other powdered oxides and even non-oxides may also be introduced. Most of the production processes are ones developed for the fine ceramic [34] and powder-metallurgy industries. Forming is normally accomplished by slip casting (→ Ceramics, General Survey) and isostatic pressing (→ Ceramics, General Survey). A freeze-molding process has also been developed [35]. Sintering is carried out at 1450 – 1900 °C. Shrinkage can be considerable (up to 25 % linear shrinkage); strict tolerances with regard to dimensions or surface quality are met by grinding. Dense, fine-grained refractory products are usually very sensitive to thermal shock.

5.3. Brick Dimensions and Dimensional Tolerances

Brick sizes for general use (rectangular and arch bricks, skew backs) are subject to international standards (ISO 5019/1 – 6). The standard rectangular brick sizes are:

1) 230 mm × 114 mm × 64 or 76 mm
2) 250 mm × 124 mm × 64 or 76 mm

Many other nonstandard shapes are employed for specific furnaces and applications (e.g., nozzles, blocks).

Dimensional tolerances are described in the 23rd Fédération Européenne des Fabricants de Produits Refractaires (PRE) Recommendation [36]. Depending on the nature of the refractory material, production method, and intended application, typical specifications are:

1) Dimensional tolerance max: 1 – 2.5 % (min. 2 – 3 mm)
2) Warpage max: 1.5 % (determined across the diagonal)
3) Taper max:
   ± 1 mm for radial height < 250 mm
   ± 2 mm for radial height > 250 mm Rectified (ground) bricks and blocks achieve a dimensional tolerance of 0.1 mm.

5.4. Packaging, Storage, and Storage Time

For purposes of storage and shipment, bricks, shapes, or bags of unshaped refractory materials are normally placed on standard wooden palettes and surrounded by a shrink-wrap film. In the case of heavy paletted shipments steel bands are used. Low-strength bricks (insulating bricks) are often packed in cardboard boxes. Unshaped refractory products are filled into paper bags lined with plastic film or into metallic drums. Delivery is sometimes made in big bags or by silo car.

Refractory ceramics generally must be stored in ventilated, covered, dry surroundings. Basic bricks and unshaped refractories are especially sensitive to moisture (hydration).

Unshaped cement or chemically bonded refractories have a limited storage life, usually 3 – 12 months.

5.5. Quality Assurance

Quality assurance systems are based on ISO 9000 – 9004, supplemented by an analysis of failure-mode effects as well as statistical process control. Quality assurance is integrated directly into every step in the manufacturing process from the delivery of raw materials to shipment of the finished product [37].

Sampling and acceptance testing are described in ISO 5022 (shaped products) and ISO 8656-1 (unshaped products). Inspection of attributes by nondestructive means (visual assessment) produces only a classification like "good" or "poor". More rigorous evaluation of variables normally requires destructive testing directed toward measurable properties. These tests are mainly chemical in nature and normally involve not more than three physical characteristics (open porosity, strength at room temperature, and hot strength or refractoriness under load).

6. General Composition, Properties, and Testing

The properties of refractory products are determined primarily by their chemical compositions, mineral compositions, and structure. Structure here refers to the type, quantity, size, form, and arrangement of phases in the substance involved. Refractory ceramics generally comprise three structural elements: coarse grain, binding matrix, and pores.

The binding matrix is of special importance with respect to refractoriness and corrosion resistance. It is usually associated with a higher degree of porosity than the grains, which have well-defined properties.

Fired bricks possess fixed, measurable properties. By contrast, unfired and unshaped ceramic products acquire their properties through use, in the course of which a temperature gradient is usually established. This temperature gradient gives rise to considerable property gradients between the hot face (sintered at the high-temperature end) and what is often a loose mass at the low-temperature face.

Properties are ascertained by applying various tests, most of which are standardized [36], [38]. Values depend on the test method employed and, in the case of unshaped products, also on the method of preparation of test pieces. Even standardized test methods are of limited value due to problems with reproducibility and comparability, a problem especially prevalent with destructive testing.

Table 6 summarizes important property tests conducted on refractory materials.

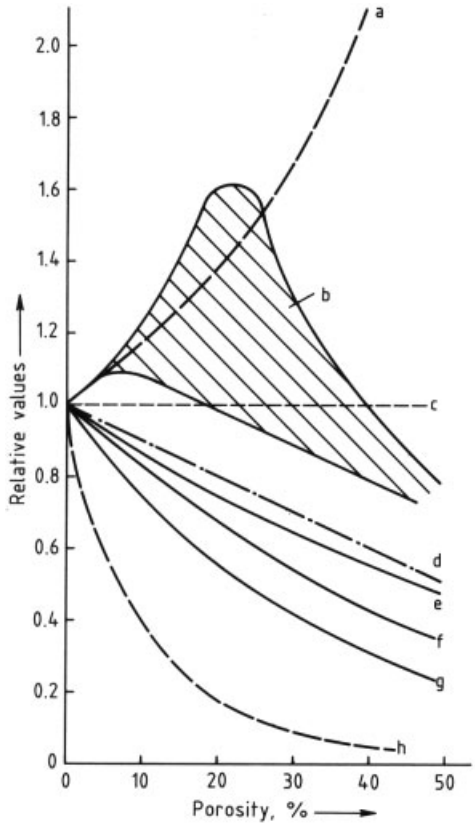

Figure 2. Porosity effects on the properties of refractory ceramics
a) Corrosion; b) Thermal shock resistance; c) Thermal expansion; d) Bulk density; e) Strength (fine ceramics); f) Thermal conductivity; g) Modulus of elasticity; h) Cold compressive strength (coarse refractories)

Chemical and mineralogical composition forms the basis for both classification and service behavior (e.g., with respect to corrosion resistance). In moderately pure materials the chemical composition is of secondary importance with respect to technological properties. However, as quality requirements become more

Table 6. Tested characteristics of refractory ceramics

Chemical, mineralogical	Texture	Cold mechanical	Hot mechanical	Thermophysical	Thermochemical
Chemical analysis[a]	true density	cold compressive strength[a]	hot compressive strength	thermal expansion[b]	refractoriness[b] (pyrometric cone equivalent)
Microprobe analysis	bulk density[a]	cold modulus of rupture[b]	hot modulus of rupture[b]	thermal conductivity[b]	corrosion resistance (slag, melts, vapors)
Microscopy	porosity[a]	abrasion resistance		specific heat	CO resistance
X-ray diffraction	permeability[b] to gas (air)			electrical resistance	oxidation resistance
	pore-size distribution	modulus of elasticity (stress–strain behavior)			hydration resistance acid resistance
			change in dimensions on heating[b]		bubble formation (only for glass melt contacting materials)
	nondestructive testing (mainly ultrasonic)		refractoriness under load[b]		
			creep in compression; thermal shock resistance[b]		
Special and additional tests and test conditions for					
Unshaped refractories	joining and gunning mixtures		carbon-containing refractories		ceramic fibers
Grain-size distribution, consistency or workability, moisture content[b]	bonding or adhesive strength		tests in neutral or reducing atmosphere [c], [d], special chemical analysis methods [c], coking behavior, residual carbon content[b]		compressibility, shot content (nonfibrous particles), tensile strength [e]

[a] For statistical quality control. [b] Additional information for data sheets. [c] [41]. [d] [42]. [e] [43].

stringent even small amounts of foreign oxides can exert a considerable influence.

Porosity has a considerable effect on properties (Fig. 2). It is difficult to optimize a refractory material with respect to porosity (e.g., for increased resistance to corrosion or thermal shock). The different types of porosity and pores are illustrated in Figure 3, together with associated test methods. The proportion of closed pores in bricks with porosity exceeding 15 % is normally small (< ca. 3 %). Relationships between porosity and other properties can be established most accurately by determining pore size distribution [39], [40] and gas permeability.

Bulk density characterizes the weight of a product per unit volume, including all pores. It is needed to calculate porosity, refractory requirement, weight of a refractory lining, stored heat, etc. Bulk density is influenced by the material density (weight per unit volume, excluding pores), which is a constant for each material, and porosity. Densities for refractory ceramics range between 2.2 g/cm^3 (fused silica) and 5.9 g/cm^3 (ZrO_2).

Cold Mechanical Properties. Cold crushing and modulus of rupture (bending) tests provide information about the structure and structural uniformity of a ceramic and assist in statistical quality control. This information is of little value in assessing high-temperature behavior, however. Strength is dependent on the shape of a sample and the conditions of testing. The distribution of strength values within a brick or batch is relatively wide. Resistance to abrasion is determined by abrasive or jet wear, but can be estimated from the above strengths.

Nondestructive Testing [44–47]. There is an increasing tendency to use nondestructive

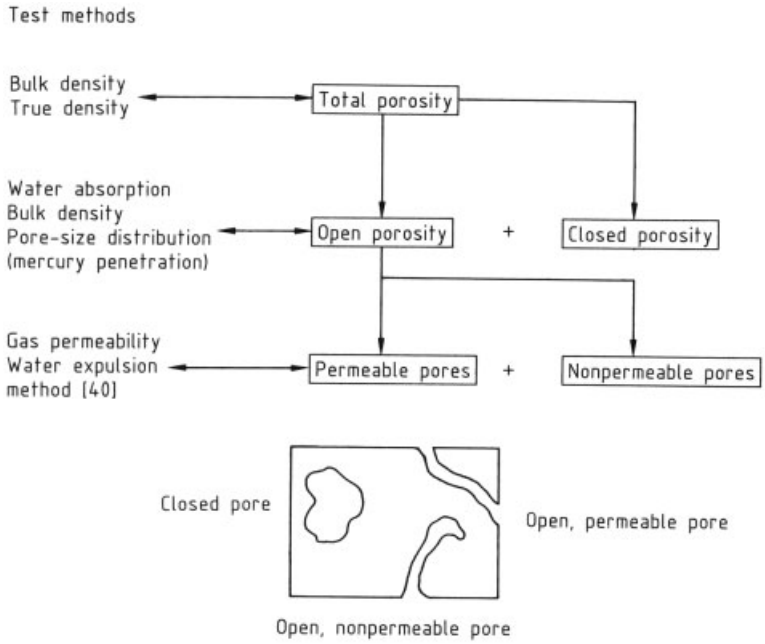

Figure 3. Types of porosities and pores in refractory materials and test methods

testing, particularly for large refractory ceramic objects with a high unit price. Ultrasonic testing is the most common nondestructive procedure for checking the strength, porosity, and integrity of refractory materials. Other methods include resonant frequency, rebound hammer, and, for crack or void locations, X-ray techniques. The interpretation of nondestructive test data is complicated by the fact that results are often not closely related to values obtained from conventional destructive testing.

Refractoriness (see also Chap. 1). For bricks belonging to the silica – firebrick – corundum series, the cone equivalent serves as a measure of the content of fluxing agent. Refractoriness is also determined, however, by ceramic structure, distribution of the fluxing agent, and viscosity of the resulting melt. Cone equivalents are of little value for basic and other highly refractory products.

Most refractory products lack a defined melting point, instead displaying a more or less broad softening interval.

Thermal expansion is a material constant determined primarily by the basic component of the ceramic (e.g., SiO_2, MgO). With the exception of silica and zirconium oxide (Fig. 13) containing products, fired ceramic materials have smooth thermal expansion curves [48] (Fig. 4). During initial heating, unfired bricks and unshaped refractories show overlap between reversible thermal expansion and irreversible expansion/shrinkage effects, which are influenced even by small loads. Reversible and permanent changes in length are of importance for construction of a refractory lining, e.g., when measuring joints. Because thermal expansion cannot be suppressed even with high loads some fraction (usually 0.5 – 0.8) of the laboratory expansion value is used for dimensioning joint thickness; the correct choice depends on furnace construction and operating conditions.

Reheat Change. Heating to high temperatures and subsequent cooling often causes a permanent, time-dependent change in dimensions. If a refractory lining shows substantial shrinkage, joints or cracks may form. The opposite case, after-expansion, is also dangerous, because the resulting pressure may destroy the refractory lining. Dense bricks should not show more than 1 – 1.5 % linear shrinkage at maximum service

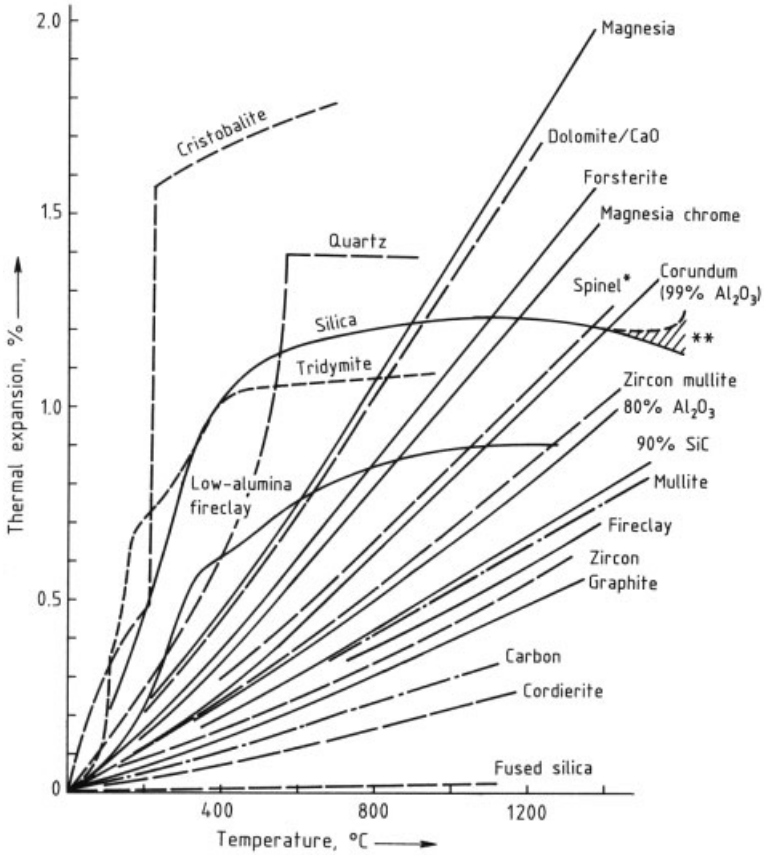

Figure 4. Thermal expansion curves of fired refractory materials
$*$ $MgO \cdot A_2O_3$, $MgO \cdot Cr_2O_3$, chrome ore. $**$ After expansion when residual quartz is present.

temperature. The appropriate tests are described in ISO 2477 (insulating bricks) and ISO 2478 (dense bricks).

The classification-temperature criterion for insulating and unshaped refractory products is limited linear heat shrinkage at a defined temperature and holding time (5 – 24 h). The following shrinkage limits apply:

Insulating fire bricks	max. 2 %	
Ceramic fiber products	max. 4 %	
Castables	max. 2 %	drying and firing shrinkage
Ramming mixes	max. 2 %	drying and firing shrinkage
Moldable mixes	max. 3 %	drying and firing shrinkage

Refractoriness Under Load. Determination of the refractoriness under load (differential, with rising temperature; ISO 1893, DIN 51 053, part 1) is a characteristic test for refractory products. Samples are normally cylinders (height 50 mm, diameter 50 mm) with an internal axial bore (diameter 12.5 mm). These are heated in air under a load of 0.2 N/mm^2 at a rate of 5 °C/min until they become soft. Characteristic temperatures (e.g., T_{05} and T_{10}) are read from a height- change – temperature curve (see Fig. 5). Some products display a broad softening range (e.g., fireclay bricks), while others show a sudden collapse (e.g., silica and magnesia bricks).

The ISO 1893 differential method establishes an actual change in sample length, and was developed from the refractoriness-underload method described in DIN 51 064. The main differences between the two methods are that in DIN 51 064 the measured expansion includes the expansion of the loading columns (carbon), and the test takes place in a carbon resistance furnace (i.e., in a reducing atmosphere). DIN

51 064 characterizes a sample by two parameters, t_a and t_e, where t_a is the temperature at which the sample is compressed by 0.6 %, and t_e that at which it is compressed by 20 %.

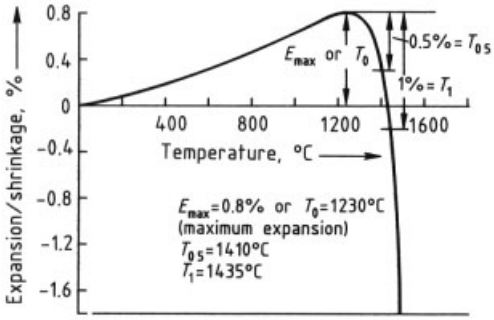

Figure 5. Refractoriness under load curve

The ISO differential method generally produces substantially lower test values than DIN 51 064. Materials that are sensitive to redox conditions (e.g., iron-rich magnesia) show particularly drastic deviations. For example:

	Differential method ISO 1893 T_{05}	DIN 51 064 method t_a
Fireclay brick	1250 °C	1350 °C
High-quality fireclay	1440 °C	1490 °C
Bauxite brick	1450 °C	1550 °C
Iron-rich magnesia brick	1500 °C	1700 °C
Low-iron magnesia brick	1680 °C	1700 °C

Softening behavior under pressure is influenced by several factors: the type of refractory material, its mineralogical and pore structure, and the quantity and viscosity of the melts formed at high temperature. Compression softening behavior improves with increasing firing temperature.

Creep in Compression. Refractory products are often exposed to high temperature and pressure for extended periods of time (e.g., the checkerwork of regenerators). This results in slow shrinkage and creep processes that influence refractory lining lifetimes. Such behavior is characterized by the creep test (ISO 3187), which is performed at constant temperature and pressure. The test is carried out as described in ISO 1893, but once the test temperature is attained it is maintained for 25 h. Loading conditions can be selected to correspond to those expected during service. The result can be read from the "creep curve", for example as percent shrinkage after 25 h (Fig. 6). Creep curves for various refractory materials are shown in Figure 7.

Behavior over longer time periods can be estimated by extrapolation. For rough calculations, the increase in compression creep can be assumed to be proportional to the applied compression load. A temperature increase of ca. 100 °C increases creep by a factor of 5 – 6 [49].

For a brick to show a low degree of long-term creep it should have a high-fired structure, a low porosity, and a low flux content or a low proportion of liquid phase at service temperature.

Hot Mechanical Properties. Thermomechanical behavior is important in the application of a refractory product, and it can be characterized by high-temperature strength testing. Strength is a complex parameter that depends on the structure, quantity, and properties of the glassy phase. The temperature dependence of thermal behavior is a function of the type of loading involved (e.g., bending, compression, tension, or torsion).

One common measure of thermomechanical character is the hot modulus of rupture test (ISO 5013), which is frequently incorporated into the quality control of refractory products. In polycrystalline bricks with low amounts of glassy phase (e.g., corundum or magnesia), strength tends to decrease gradually with increasing temperature. However, if a liquid phase is formed the hot strength can decrease suddenly. In refractory products with a high content of glassy phase (e.g., fireclay brick) a hot-strength maximum occurs at the temperature where the glassy phase starts to soften (high viscosity). Such maxima can also occur if stresses arise in the structure (e.g., in cases with strongly differing expansions of the different refractory components, such as magnesia and chrome ore in magnesia chrome bricks, or cristobalite in fireclay bricks (Fig. 8).

Thermal Conductivity, Specific Heat, and Emission.

Thermal conductivity values are used primarily to calculate the temperature distribution in a refractory wall and the heat loss of a furnace. In some cases products with high conductivity are used to keep the temperature of the hot

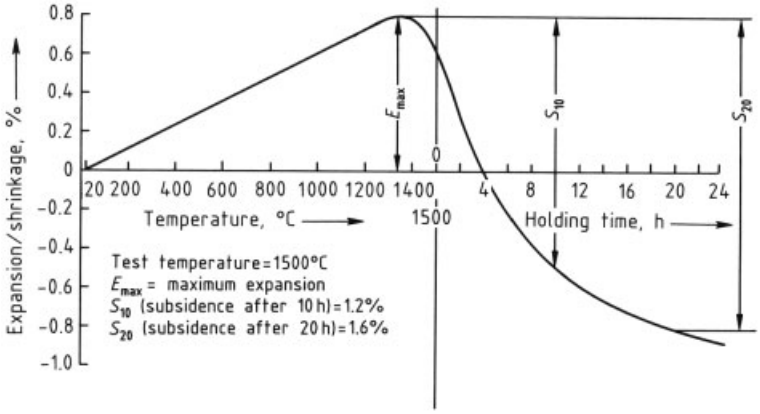

Figure 6. Creep curve

side of the lining low, thereby reducing wear. The thermal conductivity of a refractory material depends on its chemical and mineralogical composition, and especially on its porosity and pore size distribution. The atmosphere is also of great importance, most notably in the case of porous products [50], [51].

The thermal conductivity of refractory products ranges from ca. 0.05 (ceramic fibers) to $100\,\mathrm{W\,m^{-1}\,K^{-1}}$ (graphite) [48] (Fig. 9). Monocrystalline products like corundum brick show a considerable decrease in thermal conductivity with increasing temperature until such time as the heat transfer by radiation (above ca. 800 °C) again increases total conductivity. Products with a complex structure (e.g., silica brick, which has several crystalline phases, and products with a high content of "amorphous" phases) normally have lower thermal conductivities than monocrystalline materials. In this case conductivity increases slowly with temperature.

There are two standard methods for measuring thermal conductivity: the hot-wire test (ISO 8894, ASTM C 1113) and the calorimeter method (ASTM C 201). Values obtained by the two methods may differ by up to 25 %. The method used should therefore be specified whenever thermal conductivity values are given. This is particularly important for unshaped products; in such cases, special attention must also be given to pretreatment of the sample [52].

Specific heat [48] is used to calculate the heat stored by a lining, and for determining the temperature diffusivity that results from thermal conductivity in order to calculate nonstationary temperature fields in, for example, regenerators. The specific heat for conventional refractory products is ca. $1\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$.

The thermal emission of a refractory construction material depends on the type of material and the nature of the surface. Refractory products can be regarded as gray-body radiators, with emissions 50 – 85 % as great as black-body radiators [53], [54].

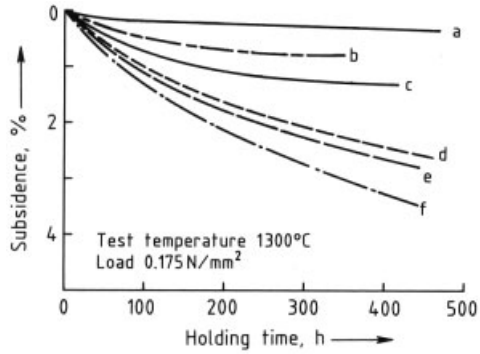

Figure 7. Creep curves for various refractory materials
a) Mullite brick; b) Low-alumina fireclay brick (low-alkali content); c) Corundum brick 70 % Al_2O_3; d) Corundum brick 60 % Al_2O_3; e) Low-alkali fireclay brick; f) Conventional fireclay brick

Electrical Resistance [48]. The electrical resistance of a refractory material is important with respect to the insulation and operation of electrical furnaces. Conventional refractory ceramics display very high electrical resistance at room temperature ($> 10^5\,\Omega\cdot\mathrm{m}$), but this decreases with increasing temperature, reaching values like $1\,\Omega\cdot\mathrm{m}$ at 1400 °C. Even low concen-

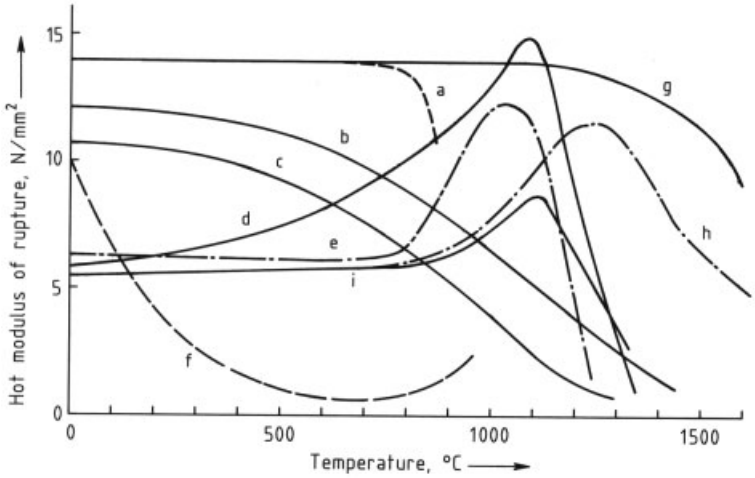

Figure 8. Hot modulus of rupture of refractory bricks
a) Low-iron magnesia with 0.2 % B_2O_3; b) Corundum (99 % Al_2O_3); c) Iron-rich magnesia; d) Low-alkali fireclay; e) Alkali-rich fireclay; f) Chemically bonded magnesia chrome; g) Low-iron magnesia (C_2S bond); h) High-fired magnesia chrome; i) Sillimanite

trations of foreign components can substantially reduce electrical resistance.

Elasticity and Stress – Strain Behavior. A *modulus of elasticity* value is determined in order to estimate the critical stress on a refractory material brought about by mechanical and thermal loading. The *static* modulus of elasticity is established by measuring deformation under compression, bending, and tension. For simplicity, *dynamic* elasticity is determined by means of ultrasonic velocity or resonance frequency. Values obtained by different methods can differ substantially, and they also depend on the type and structure of the refractory material under investigation (Table 7) [55]. The dynamic modulus is higher than the static modulus. Refractory bricks often show strictly elastic behavior only within a narrow deformation range. Increasing stress and temperature leads to irreversible deviations from linearity. This is a consequence of deformations caused by mechanical or creep factors at higher temperatures. Irreversible changes are an indirect result of the porous, inhomogeneous textures of refractory materials. Their structure – a combination of relatively solid, less deformable, coarse and medium particles together with a less solid and more deformable matrix – is particularly important in this respect. The temperature dependence of stress – strain curves for a refractory material provides an indication of the elasticity of the microstructure (relaxation) and the onset of plastic deformation.

Thermal Shock Resistance. Thermal shock resistance is an important property that describes the behavior of a material during a more or less rapid change in temperature. Cracks can lead to considerable wear due to spalling, and they also increase susceptibility to corrosion. The test procedures are quite diverse (water or air quenching, dipping in metal melts, flame impingement tests, cycling between two temperatures, etc.). Thermal shock resistance is highly dependent on sample shape and the test conditions. Densification, uptake of foreign material, and the formation of zones of differing thermal expansion during refractory product use can also lead to substantial changes in thermal shock resistance. Thus, only a test procedure adapted to suit specific service conditions has any chance of success. Another way of circumventing the difficulties with thermal shock tests is to invoke thermal shock parameters. These represent an attempt to classify materials in terms of other properties (e.g., strength, modulus of elasticity, thermal expansion, thermal conductivity, work of fracture, and fracture toughness) [56], [57]. Two important parameters have been defined to

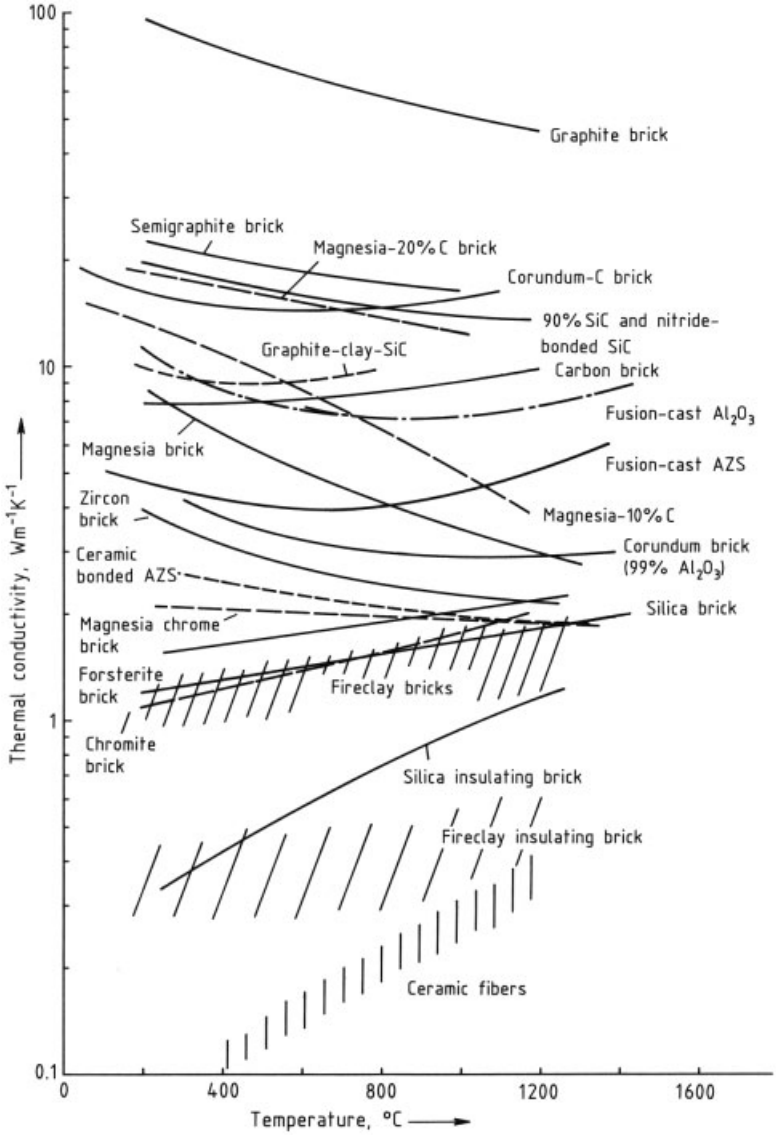

Figure 9. Thermal conductivity of refractory products

assess the thermal shock resistance of refractories:

$$R'''' = \frac{WE}{S^2}$$

(situation involving a fracture of a catastrophic nature)

$$R_{\mathrm{ST}} = \left[\frac{W}{\alpha^2 \cdot E}\right]^{0.5}$$

(situation in which crack propagation is stable or quasi-stable)

where W = work of fracture; E = modulus of elasticity; S = modulus of rupture; and α = thermal expansion coefficient.

Corrosion Resistance. The corrosion resistance of a refractory product is influenced primarily by the chemical compositions of the corrosive agents and the refractory material. Porosity, pore-size distribution, and mineral structure

of the refractory material determine the reaction kinetics. Corrosion is greatly affected by specific operating conditions (furnace atmosphere, temperature, temperature gradient in the lining, crack formation due to additional temperature changes, static or dynamic action of the corrosive media), so conclusions drawn from laboratory slagging experiments are of limited value unless these have been carefully related to real operational conditions.

The diverse character of corrosive attack on refractory materials is reflected in the numerous corrosion tests proposed [58], [59].

Refractory products are subject to three main types of corrosive stress:

1) Effects of the atmosphere and changes in the atmosphere
2) Attack by vapors, flue dusts, and condensates
3) Attack by slag and other fluid media

Effects of the Atmosphere and Changes in the Atmosphere. A change in the valency of components of the refractory material (e.g., iron oxide), as well as absorption of such substances as V_2O_5, can lead to a loosening of brick structure. Even small amounts of metallic iron or iron oxide present in e.g., fireclay or bauxite products can lead to the catalytic deposition of carbon from carbon monoxide (500 – 700 °C) and methane (> 900 °C). The resulting crystallization pressure can destroy the brick structure (CO disintegration). Vaporization processes are accelerated under a reducing atmosphere.

Attack by Vapors, Flue Dusts, and Condensates. Corrosive substances are brought in contact with refractory construction materials by the flow of combustion gases. Components present in certain temperature zones react with the refractory material, especially in their condensed states. A typical example is found in the middle checker layers in glass-furnace regenerators. Corrosion involves wetting, diffusion, impregnation, the formation of dripping melts, and periodic melting and solidification of infiltrated melts. Very great damage is caused by alkali bursting of fireclay and high-alumina bricks, a consequence of the substantial volume increase accompanying formation of nepheline ($Na_2O \cdot Al_2O_3 \cdot 2\,SiO_2$) or kaliophilite ($K_2O \cdot Al_2O_3 \cdot 2\,SiO_2$).

Table 7. Strength and elasticity properties of refractory bricks

Brick type	Microstructure	Cold strength, N/mm^2			Static modulus of elasticity, 10^4 N/mm^2				Dynamic modulus of elasticity, 10^4 N/mm^2		
		Compression	Bending	Tension	Compression	Bending	Tension	Torsion	Young's modulus	Shearing modulus	Poisson's ratio
Magnesia	crystalline, homogeneous	50	6	3.5	4.5	2.0	2.2	0.8	7.0	2.9	0.2
Magnesia–chrome	crystalline, inhomogeneous	65	4	1.5	3.0	1.0	0.8	0.3	2.9	1.2	0.2
Fireclay	crystalline + glassy phase	30	7	4.5	1.0	0.8	0.9	0.4	2.3	1.0	0.1
Fused silica	glass, homogeneous	400	20	10	4.0	3.0	1.6	1.2	4.8	2.1	0.14

Attack by Slag, Melts, and Other Fluid Media. Corrosive processes caused by slags and melts are a consequence of several phenomena (wetting, dissolution, diffusion, free and forced convection) that occur at the same time and place and often interact in a complementary way. Premature wear develops at the three-phase boundary attacking melt – refractory material – atmosphere (*melt line, flux line*) [60].

The intensity of the slagging processes in porous materials exposed to melts is primarily determined by interface properties (i.e., wettability and viscosity) of the attacking molten medium or reaction products. The formation of highly viscous reaction melts, deposits, or protective layers retards the slagging process.

Crucible tests (DIN 51 069) are usually performed in cavities prepared in cubes of the test material. The corrosive material (e.g., glass, slag, or a salt) is placed in the prepared cavity and the crucible is then heated. This test provides only a rough classification of materials.

Button and Drip Tests. The corrosive substance is placed as a pellet or dropped onto the test material. Wettability can be assessed using the sessile drop method (hot stage microscope).

Immersion Tests. The rod-immersion method has achieved considerable importance for testing the attack of melts on refractory materials. Compared to the crucible test this provides a larger quantity of attacking medium relative to the sample surface. In addition, rods or disks of test material can be rotated (dynamic test). Hanging samples can also be used to test corrosion in an atmosphere. This method is employed to test the behavior of refractory bricks on exposure to vapor from alkaline compounds located at the base of the crucible.

Semi-Industrial Corrosion Tests. These tests are somewhat extravagant in terms of time and equipment, and are a response to the demand for a test procedure that provides assessments as equivalent as possible to actual operating conditions. This in turn implies that each procedure must be adapted to a specific application. Special emphasis is placed on the testing of large samples (e.g., a brickwork), which allows one also to test the joining materials. Joints constitute a vulnerable area in a brick lining. For the iron and steel industry, the combined effect of metal and slag is tested. Thermal shock can also be superimposed on the slagging process to obtain a comprehensive picture of wear processes.

The following semi-industrial test procedures are most common:

1) Rotary furnace test
2) Vertical-type rotary furnace tests
3) Chamber slagging furnaces for blowing in or on corroding vapors, etc.
4) Induction and arc furnaces (laboratory scale)

7. Types of Refractory Ceramics

7.1. Dense, Shaped Refractory Products

Characteristics of selected types of conventional refractory bricks are summarized in Tables 8 and 9.

7.1.1. Silica Bricks and Fused-Silica Ceramics

Silica bricks (> 93 % SiO_2) [61–63] consist mainly of silicon dioxide (SiO_2), generally described as silica. Silica occurs in a variety of crystalline modifications (e.g., quartz, cristobalite, tridymite), each with a high- and low-temperature modification, and also as an undercooled melt called *fused silica.* The crystal structures of the SiO_2 modifications differ widely, so distinct density (volume) changes and steps in the thermal expansion are seen during transformation (Fig. 10). This is of great importance during heating and cooling (i.e., during firing and use of silica refractories).

The raw materials for silica bricks are naturally occurring quartzite and other quartz materials with low Al_2O_3 and alkali oxide content. The grained material is normally bonded with about 2 – 3 % slaked lime and some sulfite lye solution as a temporary binder. Dense silica bricks are produced by adding Si- containing powders (e.g., Si_3N_4), which decompose on firing to a dense tridymite-rich SiO_2 matrix [62], [64].

Table 8. Properties of conventional refractory bricks of the system $SiO_2 - Al_2O_3$; for maximum service temperatures see Figure

Type	Al_2O_3, %	SiO_2, %	Fe_2O_3, %	Na_2O^+ K_2O, %	Density, g/cm^3	Bulk density, g/cm^3	Open porosity, %	CCS,∗ N/mm^2	T_{05}, °C
Fused silica, compact		99.5			2.2	2.16	0	200	1640 – 1690 ∗∗
Fused silica, slip cast		> 99			2.2	1.90	13 – 16	50 – 80	1640 – 1690 ∗∗
Silica	< 1.5	> 95 CaO < 3	< 0.5	< 0.2	2.34 – 2.36	1.75 – 2.0	15 – 23	20 – 100	1640 – 1690 ∗∗
Low-alumina fireclay	20 – 30		< 2.5	1 – 3	2.5 – 2.65	1.9 – 2.3	15 – 25	> 35	1250 – 1300
Fireclay	30 – 45		0.6 – 2.5	0.5 – 2.5	2.55 – 2.75	1.8 – 2.4	12 – 25	20 – 85	1300 – 1550
Andalusite, sillimanite	55 – 65		< 1.5	< 1	3.0	2.4 – 2.6	15 – 20	50 – 90	1470 – 1600
Bauxite	80 – 85	8 – 14	1.3	0.5	3.6	2.7 – 2.9	18 – 22	65 – 75	1400 – 1500
Corundum (80 % Al_2O_3)	80 – 85		< 1.3	< 1	3.6	2.8	18	70	1550
Corundum (97 % Al_2O_3)	> 97		0.3	< 0.4	3.96	3.1	< 20	> 50	> 1650

∗ Cold compressive strength.
∗∗ Fused silica, when denitrified to cristobalite. In the glassy state T_{05} ca. 1350 °C.

Table 9. Properties of conventional ceramic-bonded basic bricks

Type	MgO, %	CaO, %	SiO_2, %	Fe_2O_3, %	Al_2O_3, %	Cr_2O_3, %	Density, g/cm^3	Bulk density, g/cm^3	Open porosity, %	CCS,∗ N/mm^2	t_Q, °C
Magnesia (range)	85 – 95	0.8 – 3.9	0.2 – 6	0.2 – 6	0.1 – 5		3.5 – 3.8	2.8 – 3.1	14 – 18	40 – 110	1600 – > 1700
Forsterite bonded	94	1.7	3.8	0.2	0.2		3.5 – 3.8	3.0	15	90	1700
C_2S bonded	96	2	1	0.5	0.2		3.5 – 3.8	3.0	15	60	> 1700
99 % MgO	99	0.1	0.1	0.1				3.0	15	55	> 1740
Magnesia spinel	86		0.3	0.8	13			2.95	15	55	> 1700
Magnesia – chrome	55 – 80	< 3	< 5	5 – 12	< 11	6 – 20	3.7 – 3.85	2.8 – 3.2	14 – 21	> 30	> 1500
Chrome – magnesia	25 – 25	< 2.5	< 5	7 – 15	< 22	15 – 35	3.8 – 4	2.8 – 3.35	16 – 21	> 40	> 1500
Chromite	< 25	< 2	< 7	9 – 18	< 28	> 25	3.8 – 4	3 – 3.3	17 – 21	> 30	> 1500
Forsterite	50 – 60	< 1.5	< 40	< 8	< 8	< 1.5	3.35	2.6 – 2.8	15 – 21	> 25	> 1600
Dolomite	> 36	< 61	< 1.5	< 1	< 1		3.38	2.6 – 2.9	14 – 21	30 – 80	> 1700

∗ Cold compressive strength.

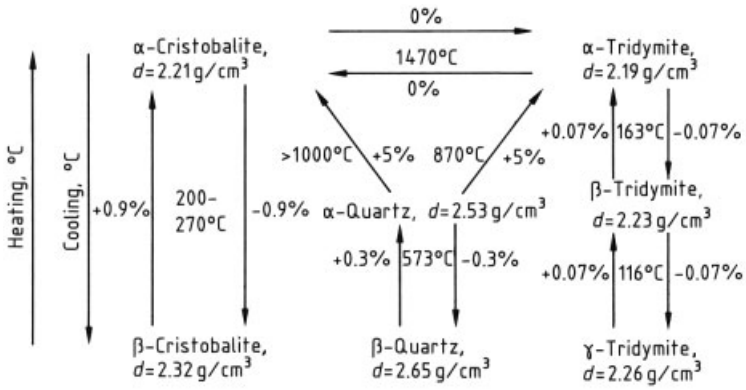

Figure 10. SiO_2 modifications and their transformations

Fired silica bricks contain cristobalite, tridymite, perhaps some residual quartz (usually $< 6\,\%$), some wollastonite ($CaSiO_3$), and about $2-10\,\%$ glassy phase. Silica bricks are characterized by very good thermal shock resistance above 600 °C and high-temperature load-bearing capacity up to about 1600 °C.

Applications: coke ovens, hot blast stoves, glass melting furnaces (crowns), and furnaces for the ceramics industry.

Silica bricks in the applications cited usually show no deformation or critical wear. Acquired impurities migrate with melts from the bricks themselves deep into the cold part of the lining, leading to a zoned texture that is prone to thermal shock. Condensations of alkali (occurring at temperatures $< 1400\,°C$) lead to a melting and dripping of the brick surface.

Fused silica products are fabricated as fused blocks (manufactured with a special sink fuse method) or in special shapes starting from crushed fused grain, formed (often by slip casting [65]) and fired at about 1050 °C. A characteristic feature of fused silica is its very low thermal expansion (i.e., high thermal shock resistance). At about 1150 °C cristobalite crystallizes without any large volume change. This crystallized material is sensitive to cooling below 300 °C (like silica bricks).

Applications: mainly in the glass industry, for hot repairs; submerged nozzles and shrouds for the casting of steel; foundries.

7.1.2. Fireclay and High-Alumina Bricks

These refractory products belong to the system $SiO_2-Al_2O_3$ (Fig. 11) [66]. Only three phases exist in this system: crystalline SiO_2 modifications, corundum (α-Al_2O_3), and mullite. The composition of the mullite ranges from $3\,Al_2O_3 \cdot 2\,SiO_2$ with about 72 % Al_2O_3 (formed on sintering above about 1000 °C) to $2\,Al_2O_3 \cdot SiO_2$ with about 78 % Al_2O_3 (crystallizes from melt). Even small amounts of impurities or flux (always present in industrial refractories) may drastically change the phase relationships and the stability temperatures of the phases. The influence of impurities generally present in the raw materials on the formation temperature for the first melt (eutectic temperature) is as follows:

Pure system	1590 °C	FeO	1250 °C
TiO_2	1480 °C	Na_2O	1050 °C
MgO	1425 °C	K_2O	985 °C
CaO	1345 °C		

The flux content (especially $Na_2O + K_2O$) determines the extent of glassy phase or viscous melt at high temperature, and thus the refractoriness. Fe_2O_3 and TiO_2 are partially incorporated into the mullite lattice and do not completely act as flux. The significance of the Al_2O_3 content is overshadowed by the influence of the flux content.

The morphology of mullite crystals may vary considerably depending on the type and composition of the raw materials, the formation conditions during firing, and subsequent thermal treatment. Mullite shapes range from fineneedled crystals formed at low temperature in the presence of a melt to coarse crystals in materials with low flux content, fired at high temperatures, or in fused mullite. The shape of the mullite crystals and their arrangement in the texture of the brick (e.g., "mullite felt") has a pronounced influence on the properties of fireclay and high-alumina bricks, especially on hot strength.

An overview of refractoriness and maximum application temperatures for fireclay and high-alumina bricks is given in Figure 12. Refractoriness in the field of fireclay refractories falls considerably below the liquidus line of the pure system $SiO_2-Al_2O_3$ because of a high content of alkali oxides and alkaline earths. Maximum application temperatures are distinctly lower due to time-dependent shrinkage or creep, which significantly affects the life of a refractory lining. It is also necessary to distinguish between application under a thermal gradient (as in a refractory wall) and under quasi-isothermal conditions (checkers in regenerators). The broad range of maximum application temperatures for high-alumina bricks shows clearly the influence of the diverse chemical and mineralogical compositions of these bricks.

Fireclay Bricks. According to ISO 1109, fireclay bricks have an Al_2O_3 content of $10-45\,\%$. They are manufactured from chamotte and bonding- clay mixtures. The limit of 45 % Al_2O_3 represents the maximum Al_2O_3

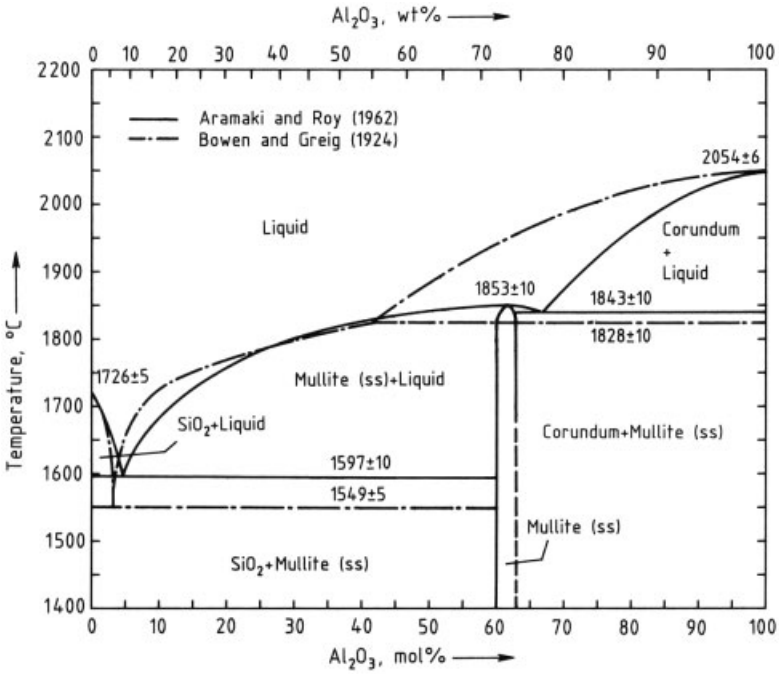

Figure 11. Phase diagram $SiO_2 - Al_2O_3$

content of the clay mineral with the highest alumina content (i.e., ignited kaolinite). The amounts of bonding clay and added water vary, a factor that largely determines both the subsequent forming process (plastic, semidry, dry shaping) and the properties of the resulting fired bricks. Low-alumina or siliceous fireclay bricks may contain as additives granular SiO_2 raw materials. Variation of the clays, grogs, and grain sizes permits the production of numerous brick types (see [67]). Fireclay bricks contain up to ca. 3 % Fe_2O_3 (for low-iron bricks, < 1 %), ≤ 3 % TiO_2, < 1 % CaO + MgO, and alkali oxides, mainly K_2O (≤ 3.5 %) and Na_2O (usually < 0.25 %). High-quality bricks include < 0.5 % $Na_2O + K_2O$.

Mullite, cristobalite, and/or quartz and a glassy phase (composition: ca. 80 % SiO_2, 10 % Al_2O_3, the rest fluxes, mainly alkalies) are the usual components of fireclay bricks. The mullite content is generally proportional to the Al_2O_3 content of the bricks. The amount of cristobalite (0 – 25 %) and of the glassy phase (20 – 60 %) depends largely on the content of alkali oxides, which can be deduced from the phase diagram $Al_2O_3 - SiO_2 - K_2O$ [68]. With rising temperature, cristobalite dissolves in the glassy phase, increasing the amount of melt. On cooling, the cristobalite again crystallizes to a greater or lesser extent from the melt. The amount of residual quartz depends on the firing temperature and the grain size of the quartz.

The resulting highly viscous melt enhances both sintering and densification of the brick texture on firing. A high firing temperature is necessary with a low alkali content to produce a low-porosity brick. Otherwise the glassy phase begins to soften at about 950 °C, imparting a wide softening range under load.

Attack by slag on fireclay bricks normally occurs in relatively small reaction zones; wear progresses by erosion of the resulting highly viscosity reaction zone. In contact with glass melts, alkali migrates into the brick texture and the bricks are slowly dissolved in the glass melt. Other forms of damage include alkali bursting (for bricks with > ca. 35 % Al_2O_3), CO disintegration, and thermal spalling (dense bricks).

Main applications: general furnace construction, blast furnaces and hot-blast stoves, foundries, furnaces for the nonferrous-metal in-

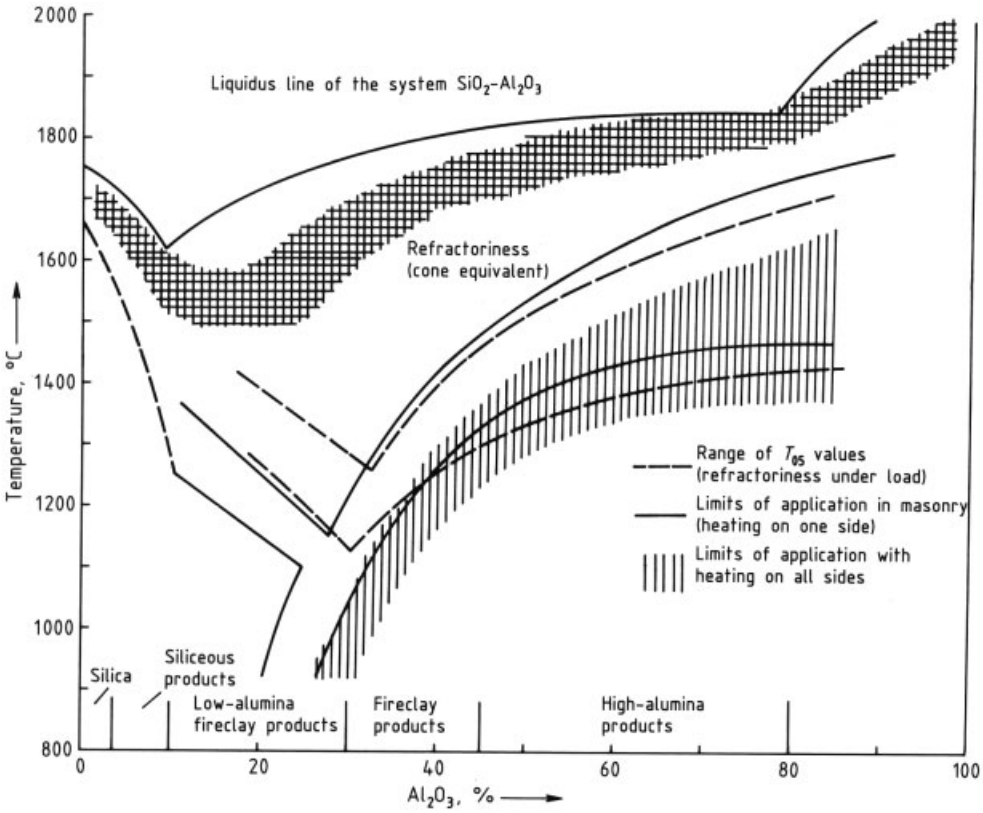

Figure 12. Refractoriness, refractoriness under load, and maximum application temperature of fireclay and high-alumina bricks

dustry, coke ovens, glass industry, cement industry.

High-Alumina Bricks [69]. Bricks with $> 45\,\%$ Al_2O_3 are required for increased thermal and chemical stresses. These are produced from various high-alumina raw materials (sillimanite group, mullitized materials, sintered and fused mullite and corundum, sintered bauxite; see Table 2), alone, in combination, or together with other raw material types, mainly chamotte. The nomenclature applied to the products depends on the primary high-alumina raw material.

A small amount of clay ($< 15\,\%$) is normally used to produce ceramic-bonded fired bricks. Use of a mixture of clay and fine alumina leads to a mullite-bonded corundum brick. A chemical bond on an aluminum phosphate base has also proven successful. This brick type is supplied either fired (1000 – 1500 °C) or heat-treated (150 – 800 °C). The chemical bond gives the brick greater resistance to attacking melts, slag, and alkali oxides. A pitch or resin bond is used for applications in the steel industry. Coking at about 300 °C causes a slag-rejecting carbon bond to develop. In special cases fired bricks are pitch- or resin-impregnated.

Addition to corundum bricks of Cr_2O_3 powder (5 – 15 %), either alone or in combination with zirconium silicate, leads to a highly corrosion-resistant corundum chrome brick [70], [71].

The chemical and mineralogical compositions of high-alumina bricks vary over wide ranges because firing of the bricks normally fails to produce phase equilibrium, and significant changes may occur in the properties of the bricks in service [72]. Bauxite bricks, which constitute the majority of high-alumina bricks, show more or less after-expansion in service caused by reactions between clay bond and corundum and resulting in mullitization [73]. Compared with

fireclay refractories, the alkali oxide content and the extent of glassy phase is low (< 20 %). The main mineralogical components are corundum and mullite. Sillimanite and andalusite bricks normally contain residual amounts of these minerals.

Properties of the different types of high-alumina bricks (for applications in the iron and steel industry) are compiled in [74].

The corrosion resistance of such bricks to glass, slag, and iron and steel melts is good. They are prone to alkali attack: high-alumina bricks with low Al_2O_3 content react to produce low-melting alkali aluminum silicate phases or glasses. With higher alumina content alkali bursting occurs (due to crystallization of, for example, nepheline), as does the formation of β-Al_2O_3.

Main applications: iron and steel industry, nonferrous industry, glass and cement industry, ceramic industry (also as high temperature kiln furniture).

7.1.3. Basic Bricks [75] (→ Magnesium Compounds, Chap. 4.3.1.)

Basic bricks belong mainly to the system MgO – CaO and consist of magnesia (periclase), dolomite, and CaO. Combinations with chrome ore, spinel ($MgO \cdot Al_2O_3$), and carbon (graphite) (see Section. 7.1.6) are also found. Forsterite, chrome ore, and spinel bricks are included as well, since these bricks are often used in place of or in contact with magnesia bricks. An overview is provided in Table 10. The discussion that follows is based on the classification scheme used in [76].

Magnesia Bricks. According to ISO 1109, magnesia bricks contain at least 80 % MgO. A variety of magnesia raw materials has been used to manufacture magnesia and magnesia-containing products (i.e., natural magnesite sinters, seawater magnesia, fused magnesia with various chemical and mineralogical compositions; → Magnesium Compounds, Chap. 4.). Magnesia contains varying amounts of Fe_2O_3 (< 8 %), Al_2O_3 (< 0.3 %), CaO (< 4 %), SiO_2 (< 4 %), and B_2O_3 (< 0.2 %). Beside the main mineralogical component periclase, silicates and ferrites occur as secondary components. The CaO/SiO_2 ratio (→ Magnesium Compounds) and the B_2O_3 content (formation of low melting components) are especially important in determining the mineral compositions and properties (hot strength, corrosion resistance) of magnesia bricks. Monticellite ($CaO \cdot MgO \cdot SiO_2$, CMS) with a CaO/SiO_2 mass ratio of 0.94, has the lowest melting point (1495 °C) of the silicates.

High-performance bricks are highly fired, have a low iron content (< 1.5 % Fe_2O_3), and contain a dicalcium silicate (short formula: C_2S, *mp* 2130 °C) or forsterite (short formula: M_2S, *mp* 1890 °C) bond. These bricks show first subsidence at temperatures ≥ 1600 °C (see Table 9). Magnesia bricks display high thermal expansion and are prone to thermal shock and thermally induced stress in furnace linings. Thermal shock resistance can be increased by a coarse grain distribution or the addition of Al_2O_3 (spinel formation on firing, microcrack formation) or chrome ore (texture elasticity). Special products include magnesia – forsterite (olivine) bricks and zircon- or zirconia- containing magnesia bricks.

High refractoriness, good resistance to attack by basic slags, and alkali resistance characterize magnesia refractories. Slagging causes diffusion and migration of silicate or ferritic melts deep into the brick, producing a zoned structure prone to peeling (structural spalling). A pitch, resin, or carbon bond reduces the infiltration of metal and slags under reducing-atmosphere service conditions. Magnesia bricks are sensitive to hydration (humid atmosphere between 40 – 120 °C), SO_3 (magnesium sulfate formation), and SiO_2 attack (forsterite bursting). Excessive growth of magnesia crystals during service, promoted by acquired impurities and high temperature, can also loosen brick texture.

Applications: steel melting and refining furnaces and vessels, regenerators for glass melting furnaces, cement and lime furnaces.

Dolomite Bricks. These consist of sintered dolomite (doloma), the main components of which are evenly distributed CaO and MgO. The principal impurities are calcium compounds like dicalcium ferrite (C_2F), brownmillerite (C_4AF), tricalcium aluminate (C_3A), tricalcium silicate (C_3S), and dicalcium silicate (C_2S). The SiO_2 content should be < 1 % to avoid excessive

Table 10. Classification and types of basic bricks

Brick type	Groups MgO content, %	Remarks
Magnesia	85 – 98	including pitch- or resin-bonded untempered and tempered bricks (carbon content after coking $< 7\,\%$)
Magnesia – dolomite	50 – 80	
Dolomite	30, 40	
Magnesia – carbon,	30 – 70,	addition of natural graphite
Magnesia – dolomite – carbon,	$C \geq 7\,\%$,	pitch- or resin-bonded untempered and tempered bricks
Dolomite – carbon	(up to 25 %)	with or without antioxidants
Magnesia – chrome	30 – 80	subgroups: direct bonded, rebonded, and chemically bonded bricks
Chrome ore	< 30	
Magnesia – magnesium aluminum spinel	60 – 80	
Spinel		$MgO \cdot Al_2O_3$, $MgO \cdot Cr_2O_3$
Forsterite		fired and unfired bricks
CaO	$CaO > 70$	

C_2S, which tends to cause "dusting" because of changes in the modification. The addition of 1 – 3 % ZrO_2 increases thermal shock resistance [77].

Their free-lime content makes these products susceptible to hydration. For this reason, manufacture, transport, storage (packed in shrink foils), and installation require virtually moisture-free environments and early use or heating. Hydration resistance is increased by pitch or resin bonding and tempering or tar dipping and impregnation of the fired bricks.

Dolomite bricks are sensitive to attack by iron-rich and SiO_2-rich slags. The periclase component is more resistant than the CaO component. Fluid melts are formed with slags rich in FeO. MgO absorbs the FeO in a solid solution.

Applications: mainly converters and ladles for the steel industry (pitch- or resin-bonded, tempered bricks) and the cement and lime industry (fired carbon-free bricks) in competition with magnesia bricks.

Magnesia – dolomite bricks contain mixtures of sintered (sometimes fused) dolomite and magnesia as a way of combining for special purposes (steel converters and ladles) the individual properties of MgO and CaO.

Calcium Oxide Refractories. Despite good corrosion resistance and metallurgical advantages (clean steel), CaO refractories have not yet found appreciable use because of hydration [78], [79].

Basic Bricks Containing Chrome. Most bricks of this type are fired mixtures of magnesia and varying amounts of chrome ore, usually in coarse to medium grain sizes. The refractory chrome ores consist of a chromite, a spinel $(Mg, Fe^{II})O \cdot (Cr, Al, Fe^{III})_2O_3$ of variable composition depending on the ore deposit, and greater or lesser amounts of a magnesium silicate. The extent of the silicate phase should be as low as possible. High-quality magnesia chrome bricks have a SiO_2 content of $< 2.5\,\%$ and a $CaO + SiO_2$ content of $< 3.5\,\%$. In contrast to magnesia bricks, the CaO/SiO_2 ratio and the B_2O_3 content have little influence on hot strength.

The following bricks are fabricated:

1) *Chemically bonded magnesia chrome bricks,* normally metal clad and with a chemical bonding agent (e.g., $MgSO_4 \cdot 7\,H_2O$); this type of product has become less important due to decreasing steel production in open-hearth furnaces
2) *Normally fired bricks,* with a predominant silicate bond between magnesia and chrome ore
3) *High-fired bricks,* with low SiO_2 content and a predominant direct bond between magnesia, residual chrome ore, and so- called secondary spinels (direct-bonded bricks)
4) *Bricks containing prereacted sintered or fused magnesia chrome grain* (rebonded basic bricks); the grain in this case contains a large amount of directly bonded phases

In magnesia chrome bricks with a low SiO_2 content the direct bond and a dense texture develop during firing at temperatures above 1750 °C. The first porous silicate bond that forms between the magnesia and chrome ore grains is subsequently displaced, and a dense direct bond arises between periclase/periclase and periclase/chromite or secondary spinels. At high temperatures spinel is dissolved in the silicate phase, and it precipitates as secondary spinel on cooling [80]. The silicate phase is isolated in pockets; i.e., the bricks show high hot strength and corrosion resistance. An improved dense texture is achieved by adding chromium metal powder [81].

Advantages of adding chrome ore to magnesia mixtures are better thermal shock resistance, increased corrosion resistance to slags with a CaO/SiO_2 ratio < 1.5 (acidic slags) due to the chromite, lower thermal expansion, and lower thermal conductivity.

Chromite bricks consist mainly of chrome ore and a small amount of magnesia. They are quite inert and very resistant to acidic and basic slags.

The wear characteristics of chrome-containing bricks are similar to those of magnesia bricks. Silicate melt penetrates behind the hot phase and spalling occurs. Alternating oxidizing and reducing atmospheres aggravate the attack. Iron oxide infiltration can cause the bricks to swell (bursting). In alkali and calcium environments chromite may decompose to form water-soluble, toxic Cr(VI) compounds. Chrome-bearing refractories are therefore used today only after careful consideration.

Applications: ladles and vessels for vacuum treatment of steel and argon oxygen decarburization (AOD) converters; rotating cement furnaces and glass furnace regenerators (bricks with low Cr_2O_3 content); open hearth furnaces; limited applications for chromite bricks.

Spinel-Containing Bricks. The magnesium aluminum spinel $MgO \cdot Al_2O_3$ is used in the refractory industry [82].

Spinel, $MgO \cdot Al_2O_3$ (71.7 % Al_2O_3, 28.3 % MgO), is the only compound in the system $MgO - Al_2O_3$, and it has a melting point of 2135 °C. First formation of a melt occurs at 2020 °C, so spinel can be classed among the high-temperature materials. With rising temperature the stoichiometric spinel $MgO \cdot Al_2O_3$ can incorporate MgO and Al_2O_3 into its lattice, resulting in spinel compositions ranging from about 66 % Al_2O_3 to 90 % Al_2O_3. Cooling or tempering leads to variable precipitation of MgO or Al_2O_3.

Sintered spinel raw materials are available with 66 – 90 % Al_2O_3, and fused spinels [83].

Magnesia-spinel bricks consist mainly of excess low-iron and low-SiO_2 magnesia together with sintered spinel. This brick type was developed as a chrome-free alternative to the magnesia chrome brick in rotating cement furnaces [84]. In the first step of development fine alumina grain was added to magnesia. On firing in situ, spinel formation occurs.

Ceramic-bonded fused spinel bricks are useful in glass furnace regenerators as an alternative to magnesia bricks in oil-fired furnaces. The spinel shows high resistance to attack by alkali and SO_3 (alkali sulfate).

A special type of spinel brick is the *picrochromite brick*, containing fused chrome spinel ($MgO \cdot Cr_2O_3$) grain, which shows extraordinary resistance to blast furnace slags and acidic coal ashes [85].

Forsterite Bricks [86]. The principal component here is magnesium orthosilicate, forsterite ($2\,MgO \cdot SiO_2$). Olivine is the main raw material, containing < 8 % FeO. During manufacture, magnesia is added to transform the FeO on firing into magnesium ferrite ($MgO \cdot Fe_2O_3$). Chrome ore or spinel additives are used to alter the properties, especially to increase thermal shock resistance. Forsterite bricks display good resistance to iron-containing slags up to 1400 °C, as well as to alkali attack.

In addition to fired forsterite bricks, unfired chemically bonded *olivine bricks* are on the market.

Applications: regenerators for glass melting furnaces, safety linings in the iron and steel industry (restricted use).

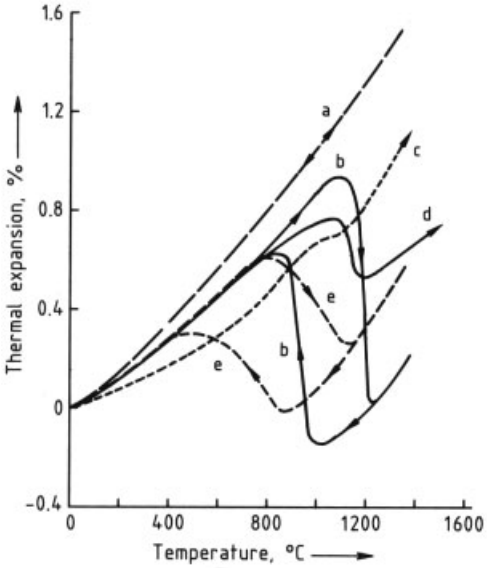

Figure 13. Thermal expansion of ZrO_2 ceramics
a) Fully stabilized; b) Baddeleyite; c) Ceramic-bonded AZS brick; d) Fusion-cast AZS brick; e) Partially stabilized

7.1.4. Zircon- and Zirconia-Containing Bricks [87] (Table 11)

The primary raw material for ZrO_2-containing refractories is zirconium silicate, $ZrSiO_4$ (mineral name: zircon), with a theoretical composition of 67.2 % ZrO_2 and 32.8 % SiO_2. It is stable to ca. 1800 °C, but the temperature for thermal decomposition into ZrO_2 (baddeleyit) and SiO_2 may be far lower depending on the type and amount of accompanying components. Chemical processing of zircon results in powdered zirconium oxide, which is worked up to various types of refractory raw materials (sintered and fused material, stabilized or partially stabilized materials).

Pure ZrO_2 (e.g., the mineral baddeleyite) has a monoclinic lattice, which is transformed reversibly beginning at about 950 °C into a tetragonal modification. This is stable to ca. 2300 °C, at which point it is transformed again into a cubic modification. The monoclinic – tetragonal transformation is accompanied by a volume change of ca. 7 % together with considerable hysteresis (Fig. 13), resulting in stress and cracking in pure ZrO_2 ceramics. Consequently, such ceramics cannot be manufactured. Addition of suitable metal oxides (for refractory purposes, mainly CaO, MgO, and Y_2O_3 in solid solution with ZrO_2) permits the cubic modification to be stabilized down to room temperature. A fully stabilized ZrO_2 can be achieved by the addition of about 5 – 10 % stabilizer. Partially stabilized zirconia raw materials are now used in most cases (3 – 6 % stabilizer added). This leads to smaller volume changes on heating and provides the additional advantage of creating microcracks, which increases the thermal shock resistance of the brick.

The following types of ZrO_2-containing products can be distinguished:

1) Ceramic-bonded zirconium silicate bricks
2) Ceramic-bonded $Al_2O_3 - ZrO_2 - SiO_2$ (AZS) products, manufactured from mixtures of zircon and high-alumina raw material
3) Fusion-cast ZrO_2-containing bricks, such as fusion-cast AZS bricks (see Sec. 7.1.9)
4) Stabilized or partially stabilized sintered ZrO_2 products (zirconia ceramics)

Zirconium silicate bricks are manufactured from zircon sand or prefired zircon grain, with powdered zircon and clay or only very fine zircon powder as a bonding agent. They are fired at ca. 1400 – 1600 °C. Shaping is accomplished by pressing, isostatic pressing, and slip casting. Depending upon the application, shapes with open porosities from 1 – 24 % are produced. The fired bricks contain a considerable amount of undecomposed $ZrSiO_4$ [88–90].

Bricks made from zircon and fused silica are used as hot repair materials because of their high thermal-shock resistance.

Ceramic-Bonded AZS Bricks. The raw materials here are zircon, corundum, calcined Al_2O_3, and mullite (also prefired cyanite). Depending mainly on the firing temperature, two types of products can be distinguished by their mineralogical composition: bricks containing corundum, mullite, and zircon; and bricks containing corundum, mullite, and baddeleyite.

In contrast to fusion-cast AZS bricks (see Section 7.1.9), the ceramic-bonded AZS bricks contain a very small amount of glassy phase and offer far better thermal shock resistance. They also show good resistance to molten glass.

Table 11. Ceramic-bonded zircon- and zirconia- containing refractories

Type	ZrO_2, %	SiO_2, %	Al_2O_3, %	Bulk density, g/cm^3	Open porosity, %	CCS,[a] N/mm^2	Thermal expansion, 1000 °C, %	Thermal conductivity, 1000 °C, W m^{-1} K^{-1}
Zircon silicate								
Pressed or slip cast	62 – 65	33 – 34	0.5 – 1	3.6 – 3.8	17 – 22	20 – 100	0.4	2
Isostatically pressed	62 – 65			3.6 – 4.2	1 – 17	> 70	0.4	
Zircon – mullite (AZS) [b]								
Normally pressed	11 – 23	10 – 15	65 – 73	2.9 – 3.2	15 – 24	60	0.65	1.9 – 2.5
Isostatically pressed	11 – 23			3.2 – 3.5	1 – 14	60		
Zircon – zirconia (nozzle for steel casting)	85			4.25	19			
Stabilized zirconia								
MgO stab.	95			5.0	10	400	0.75	1.5
Dense (CaO stab.)				5.6	0		1.1	

[a] Cold compressive strength.
[b] Mineral composition: zircon 0 – 30 %, baddeleyite 20 – 0 %, mullite 25 – 30 %, glassy phase < 10 %.

Zirconia Ceramics [91], [92]. These are fabricated primarily by fine- ceramic production methods in dense (and therefore thermal-shock-sensitive), small, special shapes. High thermal resistance (melting point, depending on stabilizer, 2500 – 2600 °C) and good corrosion resistance characterize ZrO_2 ceramics. The type of stabilization determines thermal-shock and corrosion resistance. Destabilization may occur in contact with slags.

Applications: zircon- containing refractories are used in glass-melting furnaces (superstructure, bottom, feeders) and in the steel industry as ladle bricks and nozzles. Zircon bricks are resistant to oil residues and especially to V_2O_5. ZrO_2 materials are used as nozzles and sliding nozzle plates for the extremely corrosive conditions associated with the teeming of steel. Other refractory applications of ZrO_2 ceramics are in the ceramic industry as high-temperature kiln linings and kiln furniture.

7.1.5. Carbon and Graphite Bricks
(Table 12)

Carbon is unmeltable and not wetted by melts of metals, slag, or salts under oxygen-free conditions, but it is soluble in pig iron melt. Carbon resists the attack of many acids. High thermal conductivity (especially of graphite) and low thermal expansion impart high thermal-shock resistance. Applications of these bricks are restricted because of susceptibility to attack by oxygen, steam, and CO_2 above 400 °C.

Raw materials with the required grain size (low ash cokes, anthracite, natural and artifical graphite) are mixed together with tar or pitch in heated mixers and, while the mix is still hot, compacted by ramming or vibration. The cold, hardened shapes are subsequently fired at temperatures $\leq$ 1400 °C under the exclusion of air. Pure graphite shapes are produced by a second heat treatment (resistance heating) at ca. 3000 °C.

Depending on the crystallinity of the carbon or graphite and mixtures of the various raw materials, products can be fabricated with widely varying properties (thermal conductivity and oxidation, erosion, and alkali resistance). Addition of, e.g., Al_2O_3 or SiC increases corrosion resistance [93].

Applications: as large carbon blocks in the bottom and bosh of a blast furnace (→ Iron); linings for electric reduction furnaces (electrolytic aluminum cells, FeSi production); acid-resistant linings for the chemical industry. Because graphite- containing bricks display high thermal conductivity they are used as a heat-abducting medium in conjunction with water cooling systems in furnaces.

Table 12. Carbon and graphite refractory bricks

Type	Ash content, %	Bulk density, g/cm^3	Open porosity, %	CCS ∗, N/mm^2	Thermal expansion, 1000 °C, %	Thermal conductivity, 1000 °C, $W\,m^{-1}\,K^{-1}$
Carbon	4.5 – 6.5	1.5 – 1.6	13 – 19	25 – 65	0.4	8 – 14
Semigraphite	1 – 3	1.6 – 1.7	13 – 21	25 – 48	0.38	14 – 24
Graphite	1	1.55	26	15	0.27	51

∗ Cold compressive strength.

Table 13. Graphite- containing refractory products

Type	Application	Al_2O_3, %	SiO_2, %	C, %	ZrO_2, %	SiC, %	Others,[a] %	Bulk density, g/cm^3	Open porosity, %	CCS,[b] N/mm^2
Chamotte and high-alumina graphite (fired)	steel casting nozzles, stoppers	30 – 70		7 – 20				2 – 2.3	17 – 24	
Pitch or resin bonded $Al_2O_3 - SiO_2 - SiC - C$	torpedo ladle bricks	68 – 90	19 – 30	10 – 17		5 – 8	0 – 6	2.7 – 2.8	4 – 10	55 – 80
$Al_2O_3 - C$ base [c]	continuous steel casting (shrouds, nozzles, stopper rods)	42 – 56	14 – 24	24 – 33	0 – 3		0 – 5	2.2 – 2.45	12 – 18	
		60	0.9	24	0 – 9	4		2.7	17	
$ZrO_2 - C$	continuous steel casting (shrouds, nozzles, stopper rods)	1 – 2	5 – 7	7 – 10	75 – 80		0 – 2	3.55 – 4.2	12 – 16	
				20 – 25	67 – 74			3.2 – 3.6	15 – 18	
			0 – 2	5 – 9		MgO 85 – 92	0 – 5	2.45 – 2.5	15 – 18	
MgO – C	continuous steel casting (shrouds, nozzles, stopper rods)		12 – 16	10 – 20		58 – 78		2.25 – 2.5	14 – 17	
	converter, ladle arc furnace bricks [c]			10 – 25		70 – 89	0 – 6		3 – 6	30 – 50

[a] Mainly antioxidants (metal powder, B_2O_3- containing compounds, etc.). [b] Cold compressive strength. [c] Thermal conductivity at 1000 °C; 6 – 13.5 $W\,m^{-1}\,K^{-1}$.

7.1.6. Carbon and Graphite-Containing Refractories (Table 13)

Combinations of different refractory raw materials with carbon (originating from a pitch or resin bond and/or graphite, mainly natural graphite) play a dominant role in the metallurgy of ferrous and nonferrous metals, and above all in the casting of steel.

When manufacturing pitch-bonded, resin-bonded, or impregnated products, preheated refractory grain and binder (usually $< 10\,\%$) are hot mixed (temperatures up to 160 °C depending on the viscosity of the binder) and hot pressed. Liquid resins or solutions containing curing agents (hexamethylenetetramines) allow cold mixing. Complete wetting of the graphite must be achieved. Resin hardening occurs at 200 °C (thermosetting). Impregnation of bricks with pitch or resin is carried out under vacuum and at the appropriate temperature (with resin solutions: room temperature) [29].

Heating causes volatile matter to be released and the "carbon bond" develops. The carbon yield ($\leq 70\,\%$ of binder amount) depends on the type of binder and the coking conditions. The release of volatile organic compounds requires environmental precautions during the production and use of organic-bonded refractories.

Several types of carbon- and graphite- containing refractory ceramics are produced:

1) Fired graphite- containing refractories
   a) *Chamotte – clay – silicon carbide–graphite shapes*, used mainly as crucibles for the melting and casting of nonferrous metals. These contain added SiC and silicon metal to increase oxidation resistance and strength. To minimize oxidation during firing under reducing conditions (≤ 1350 °C) and during use, such shapes

are glazed, or a glaze-forming additive is introduced into the mix. Silicon carbide crucibles are manufactured with a pitch bond; i.e., a carbon bond is formed during firing [94], [95].

b) *Chamotte – graphite and high-alumina – graphite shapes* for steel casting such as stoppers and nozzles.

c) *Pitch- and resin-impregnated fired bricks* (high alumina, magnesia, and dolomite products).

2) Pitch- and resin-bonded refractories

a) *Unfired pitch- or resin-bonded or tempered high-alumina bricks* (after coking, ca. 5 % C).

b) *$Al_2O_3 - SiO_2 - SiC - C$ bricks:* combinations of high-alumina raw materials, SiC (5 –15 %), and graphite (5 – 15 %), developed first for the transport and treatment of pig iron [96].

c) *Pitch- or resin-bonded, unfired or tempered magnesia, magnesia – dolomite* and *dolomite bricks and shapes* [97] (after coking, $\leq$ 6 % C, at which point additional carbon black is added up to 10 %).

d) *Pitch- or resin-bonded or tempered oxide – graphite products* [98–100] from mixtures of mainly pure oxides (sintered and fused grain) together with $\leq$ ca. 30 % graphite, with or without the addition of so- called antioxidants (metal or nonoxide powders), usually < 6 %. These include:

Corundum carbon and *zirconia carbon products,* used mainly as nozzles and shrouds for the continuous casting of steel. Addition of fused silica or ZrO_2- containing aggregates to corundum graphite products increases thermal shock resistance. Special additives like BN reduce the clogging of nozzles.

Magnesia carbon bricks and shapes, used for lining steel converters, for the slag area in ladles and steel refining vessels, and in electrical arc furnaces [101], [102]; also as nozzles for continuous steel casting.

Especially the oxide – graphite products combine the nonwettability of graphite crystals by metal melts and slag (no infiltration into the bricks) with the high corrosion and erosion resistance of dense oxide grain. These products have low open porosity (< 6 %) in the delivered state, but porosity after coking increases to ca. 10 – 12 %. Shapes for the casting of steel normally display a higher porosity (see Table 13) as well as high thermal-shock resistance [100].

The addition of metal powders (e.g., aluminum, magnesium, silicon) or alloys and/or nonoxide compounds like SiC, B_4C, or ZrB_2 inhibits carbon oxidation. Complex reactions occur between the brick components and the atmosphere (CO_2, CO, N_2) during heating in service, leading to such new compounds as Al_4C, which is said to increase strength, and spinel, which increases corrosion resistance. Secondary carbon may also precipitate [103]. The effects of graphite content on essential properties of graphite- containing refractories are shown in Figure 14.

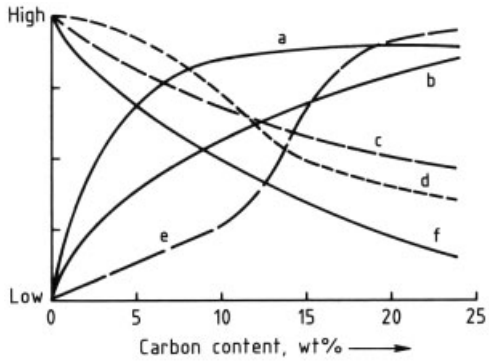

Figure 14. Effect of graphite content on the properties of graphite-containing refractories
a) Slag penetration resistance; b) Thermal shock resistance; c) Oxidation resistance and hot strength with antioxidants; d) Thermal expansion; e) Thermal conductivity; f) Oxidation resistance and hot strength without oxidants

Carbon reduces oxides at high temperatures with the formation of suboxides or metal vapors. Reduction of MgO begins at ca. 1500 °C with the formation of carbon monoxide and magnesium vapors, which may destroy brick texture. On the other hand, the metal vapors can be oxidized in parts of the bricks to form a corrosion-decreasing dense MgO zone.

During service, carbon- containing refractories normally show only a small oxidized working zone, where the effects of corrosion mechanisms known from comparable non- carbon-containing refractories are evident. Carbon still present behind this zone minimizes a deeper infiltration of slag or metal.

Table 14. Silicon carbide refractories

Type	Al_2O_3, %	SiC, %	Si_3N_4/SiAlON, %	Bulk density, g/cm^3	Open porosity, %	CCS,[a] N/mm^2	CMR,[b] N/mm^2	Thermal expansion, 1000 °C, %	Thermal conductivity, 1000 °C, W m^{-1} K^{-1}	Max. service temperature, °C
Silicate-bonded	27	50		2.4	21	70	18		5	
								0.45		1450 – 1550
	8	90		2.5 – 2.7	12 – 20	100	30		10	
Nitride-bonded		70 – 85	≤ 25	2.6	14 – 20	130 – 150	50 – 70	0.45	10 – 15	1500 – 1550
Recrystallized		> 99		2.6	15 – 18	300 – 700	100	0.48	22	1600 ox. 1700 red.
SiSiC		> 80		3.05	0	1250	300	0.45	60	1650

[a] Cold compressive strength.
[b] Cold bending strength.

7.1.7. Silicon Carbide Bricks [104–107] (Table 14)

These can be divided into two classes according to their SiC content: SiC- containing refractories with ca. 40 – 70 % SiC, usually in combination with chamotte and a clay bond; and SiC products with 70 – 100 % SiC and different types of binding (and therefore different properties).

The bonding types encountered are:

1) *Silicate or oxide bond,* where the SiC grains normally are bonded with about 10 – 20 % bonding clay, alone or in combination with fine Al_2O_3, leading to a mullite bond; hot strength is determined largely by the composition of the resulting glass phase
2) *Silicon nitride and sialon* (*SiAlON*) *bond*, produced by firing under a nitrogen- containing atmosphere a green mixture composed of Si – SiC, which may also include Al_2O_3 or other alumina-bearing compounds
3) *Selfbond*, whereby a SiC bond is achieved by adding silicon and carbon and firing in an oxygen-free atmosphere
4) *Recrystallized SiC,* where the bond phase (SiC) forms during high-temperature sintering (at ca. 2000 °C) by vaporization and condensation reactions of fine SiC particles

A special type is the *SiSiC ceramic*, an SiC product with an infiltration of silicon metal. The exceptional properties of SiSiC ceramics include high thermal conductivity, low thermal expansion, very good resistance to thermal shock, high hot strength, and abrasion resistance.

SiC is sensitive to oxygen and such oxygen-containing gases as water vapor. In an oxidizing atmosphere SiC begins to decompose to SiO_2 and CO or CO_2 at ca. 900 °C. Above ca. 1200 °C the SiO_2 layer that forms on the SiC grains becomes glassy and retards further oxidation. During installation it is especially important to take into account the susceptibility of the product to excess air in the temperature range 900 – 1100 °C. Oxidation can be largely prevented by a dense protective coating (e.g., an Al_2O_3 surface layer).

SiC is also sensitive to alkali melts, basic slags, chlorine gas, and metal melts other than lead, zinc, and copper. Alkali resistance can be increased by using a nitride bond.

Applications: blast furnaces, kiln furniture in the ceramic industry (maximum service temperature, 1500 °C), municipal incinerators and special refuse incinerators, retorts for zinc distillation, aluminum industry (submerged burner tubes), tubes for ceramic regenerators.

7.1.8. Fine-Grained Oxide and Nonoxide Ceramics [13], [14]

These normally very pure and highly refractory materials have not yet found appreciable application in the classical markets for refractories (exceptions include ZrO_2, Section 7.1.4, and SiC, Section 7.1.7) because of high price, a restriction to small or thin shapes, and often low thermal-shock resistance. The production

of nonoxide shapes demands special fabrication methods not yet adapted to the refractory industry. Nonoxide refractories (carbides, nitrides, borides) are prone to oxidation at relatively low temperature.

One of the most promising areas of potential development for fine ceramics is their use as engineering or functional refractory materials and in very special applications in industries such as steel.

Uses for special fine ceramics include:

Chromium oxide – used in highly corroded areas of glass tank furnaces (fiber glass production). TiO_2 is added as a sintering aid [89].

Tin oxide – used as an electrode material for electric glass-melting furnaces [89].

Silicon nitride-based materials – skid rails in reheating furnaces for the steel industry, anchors for fixing ceramic-fiber linings, aluminum industry (not wetted by aluminum melt).

Boron nitride-based composites with Si_3N_4 and AlN, tested also in combination with ZrB_2 and TiB_2 or mullite – break rings for horizontal continuous steel castings, reinforcement of the slag zones of submerged nozzles in continuous steel casting [108], [109].

Most special ceramics are very dense, but porous materials based on Al_2O_3, ZrO_2, and SiC with defined pore-size distribution are increasingly being used for filtering steel and nonferrous-metal melts, as well as for high-temperature filtration of gases.

7.1.9. Fusion-Cast Bricks [110]

Most fusion-cast products belong to the system $Al_2O_3 - SiO_2 - ZrO_2(-Cr_2O_3)$. The finely ground raw materials (alumina, zircon, zirconia, chrome ore, chrome oxide, special additives) are usually melted in an electric arc furnace and cast into sand or graphite forms. The shapes produced range from small parts to blocks with lengths up to 1 m or more. During the more or less rapid cooling process (for larger blocks, 6–20 d) the characteristic texture of a fused-cast brick develops: a dense structure (practically free of open pores) with intercrystalline bonding of the mineral phases. The fusion cast $Al_2O_3 - ZrO_2 - SiO_2$ (AZS) bricks contain a high content of a glassy phase (ca. 20 mass %), but it has no bonding function, simply filling the gaps between crystals.

The dense structure imparts high strength and both corrosion and erosion resistance. Disadvantages are low thermal-shock resistance and structural inhomogenities caused by contraction and segregation during cooling.

The various types of fusion-cast bricks are described in Table 15. A special feature of AZS fused-cast bricks is draining out of the glassy phase during first heating, which may lead to bubble formation in contact with glass melts. A newly-developed fused-cast ZrO_2 brick contains a special glassy phase that acts as a buffer against the transformation of baddeleyite [111], [112].

Applications: primarily in glass tank furnaces (→ Glass, Chap. 5.7.), regenerators for glass furnaces, reheating furnaces in the steel industry (skid rails).

7.1.10. Refractory Ceramics with Low Thermal Expansion

As in the case of glassy SiO_2, ceramics with low thermal expansion display extraordinarily high thermal-shock resistance, and they are especially qualified for service conditions involving frequent changes in temperature (e.g., kiln furniture).

These ceramics are products based on cordierite ($2\,MgO \cdot 2\,Al_2O_3 \cdot 5\,SiO_2$), lithium aluminum silicates (eucryptite, spodumene, petalite) [113], and aluminum titanate ($Al_2O_3 \cdot TiO_2$) [114].

Cordierite products have acquired considerable industrial significance. Chamotte or cordierite – chamotte (with or without the addition of corundum or mullite) is bonded with cordierite-forming mixtures of clay, kaolin, talc, or magnesia through firing at ca. 1150 – 1350 °C. The cordierite content of the fired product is ca. 40 – 80 %. The maximum application temperature of normal cordierite shapes is 1250 °C. Addition of mullite and corundum increases both hot strength and service temperature [115] (Table 16).

Applications: kiln furniture.

Table 15. Properties of fusion-cast bricks [a]

Type	Al_2O_3, %	SiO_2,	ZrO_2,	Na_2O,	MgO,	Cr_2O_3,	FeO,	Bulk density, g/cm^3	Thermal expansion, 1000 °C, %	Thermal conductivity, W m^{-1} K^{-1} 300 °C	800 °C	1400 °C	Mineral phases
Corundum (α-Al_2O_3)	99.5			0.2					0.8				corundum
α/β-Al_2O_3	95	1.2		3.5				3.2	0.8	5	4.2	7	corundum, $Na_2O \cdot 11\,Al_2O_3$
β-Al_2O_3	94.5			5.2				2.8	0.8				$Na_2O \cdot 11\,Al_2O_3$
β'''-Al_2O_3 [b]	87.5	0.3		4.5	7.7			2.8	0.8	2.3	2	6.5	$15\,Al_2O_3 \cdot 4\,MgO \cdot Na_2O$
Corundum – mullite	74	20	2	0.8					0.75				corundum, mullite, 15 % glass
Corundum – zircon	50	16	33	1.1					0.75[c]	4.5	4	6.5	corundum, baddeleyite, 16 – 21 % glass
	45.5	12	41	1									
Zirconium oxide	0.8	4.5	94	0.4				5.3	0.7	3.6	2.5	3	baddeleyite
Corundum – chrome	60	2				28	4						ruby, spinel
Corundum – chrome – zircon	32	13	26			26	3	4.0	0.7	5.5	4	6	ruby, baddeleyite, 20 % glass
Magnesia – chrome	7	2.5			57	20	12	3.2 [c]	1	4	4.5	5.5	periclase, spinel

[a] Open porosity < 2 %, cold crushing strength > 200 N/mm^2.

[b] Microporosity ca. 12 %, cold crushing strength 60 N/mm^2;

[c] See also Figure 13; porosity 15 % (block).

Table 16. Cordierite refractory ceramics (kiln furniture) [a]

Type	Al_2O_3, %	SiO_2, %	MgO, %	Bulk density, g/cm^3	Open porosity, %	Strength, N/mm^2	Thermal expansion, 1000 °C, %	Max. service temperature, °C
Fireclay – cordierite	38 – 43	47 – 53	5 – 5.5	2 – 2.05	24 – 30	CCS [b] 50 – 70	0.35	1100 – 1250
Cordierite – mullite	35 – 38	52 – 54	8	1.8 – 2.05	24 – 30	CMR [c] 14 – 24	0.23 – 0.33	1200 – 1400

[a] Thermal conductivity 1 – 1.5 W m^{-1} K^{-1}.
[b] Cold compressive strength.
[c] Cold modulus of rupture.

7.2. Unshaped Refractory Materials [6], [27], [116–118]

Unshaped refractory materials, also called monolithics, are mixtures consisting of an aggregate (grains) and one or more bonds, prepared ready for use either directly or after the addition of one or more suitable liquids (usually water). These mixtures are either dense or insulating.

The CEN Draft Proposal of September, 1992 defines such products according to

1) Intended application:
   Materials for monolithic constructions
   Materials for repairs
   Materials for jointing
2) Product type and placement:
   Castables
   Gunning materials
   Moldable materials (mainly ramming mixes, plastic and tap-hole mixes)
   Refractory cements and mortars
   Others (mainly dry mixes, injection mixes, and coatings, each corresponding to various types of placement)
   Prefabricated shapes made from unshaped materials (mainly castables and moldables)

Classification within the various categories is based on the temperature below which the material does not shrink more than a specified amount (linear shrinkage 2 – 3 %).

Other classification schemes correspond to those for shaped products (see Chap. 3 and Table 1). Additional subdivisions are introduced to reflect application characteristics specified by the principal consumer – the iron and steel industry – in the form of a so- called "code for unshaped products" [119]. This code defines monolithics according to a duodecimal system based on such factors as state as received, type of binder, processing technique, major raw material, composition, and classification temperature.

Four main types of unshaped products can be identified according to chemical composition:

1) Aluminosilicate refractories, including products consisting mainly of aggregates of alumina, silica, and/or silico-aluminates
2) Basic refractories, mainly aggregates of magnesia, dolomite, chromite, and spinel
3) Refractories containing carbon (i.e., graphite or carbon together with other nonwetting agents), including mixes for troughs of blast furnaces [120]
4) Special products containing oxide or nonoxide aggregates different from the above (e.g., silicon carbide, zircon, zirconia)

The raw materials (see Table 2) and compositions of monolithics correspond essentially to those of shaped products. Their properformulation presupposes a thorough knowledge of raw material properties, composition (with special emphasis on bonding and special additives; see Tables 3 and 4), conditioning, and processing. This is due to the fact that a shaped product can be tested prior to application with respect to its suitabilty (e.g., nondestructively), while monolithic materials are transformed into a "shaped" product only after installation in a furnace or heat-up of a lining. This means that production errors have consequences at the time the product is applied, when all costs for the product, including the cost of installation and commencing operation in the facility to be lined with a refractory material, have already been incurred.

The advantages of unshaped refractories (e.g., shorter production and delivery time, quick and mechanized installation, monolithic lining,

flexible construction, and application) and enormous technical advances have led to a considerable increase in the use of monolithics (see Chap. 10). This development became possible because of the introduction of new binder combinations and special additives that result in material properties equal or superior to those of shaped products, as well as improved workability and economical new installation methods.

Unshaped refractory materials used in industrial linings are obviously not prefired. This means that the following factors must be taken into account:

1) A temperature gradient is incurred, one consequence of which is a corresponding property gradient (i.e., a drop in strength across the lining)
2) The maximum temperature at the hot face depends directly on the working temperature of the apparatus
3) Equipment life may be numbered in hours or years.

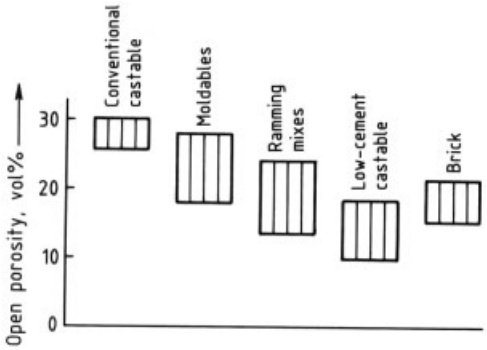

Figure 15. Comparison of open porosities of various fired unshaped refractory materials (60 – 90 % Al_2O_3)

Materials for Monolithic Constructions. An overview is provided in Table 17.

Ramming mixes are noncoherent before use. These are made up of refractory aggregates, bond(s), and (if necessary) liquid(s).

Processing is accomplished by hand or, more frequently, with the help of some mechanical device (pneumatic rammer, slinger, etc.) or vibration. Hardening generally occurs under the influence of heat.

Dry ramming mixes are a speciality product, often vibrated behind steel forms after mixing with a low-melting fritting agent (e.g., boric acid) or a thermosetting resin (hardening at ca. 200 °C).

Plastic mixes are ready-for-use products with a moldable, plastic consistency. The principal bonding agent is plastic clay. These mixes are supplied in a soft or preformed state (extrusion-pressed blocks). They are placed by ramming, vibration, pressing, or "extrusion". Preformed blocks require no molds for formation into a wall, but due to their high water content they shrink to a greater extent. Pitch- or resin-bonded tap hole mixes constitute a special type of moldable material.

Moldable mixes have many different applications, but they are being replaced to some extent by the far more satisfactory modern castables.

Refractory castables are supplied dry, and used after the addition of water or some other liquid followed by mixing. They are placed by casting with or without vibration (self-flowing), or by rodding, pumping, injecting, and sometimes tamping. Bond formation and hardening occur at ambient temperature. Hydraulic-bonded castables are the most common type of monolithics.

Two main families of castables are the *dense* and the *insulating* castables, with a limit of 45 vol % porosity (after total dehydration). Insulating castables are described in detail in [121]. Various types of dense castables are listed in Table 18.

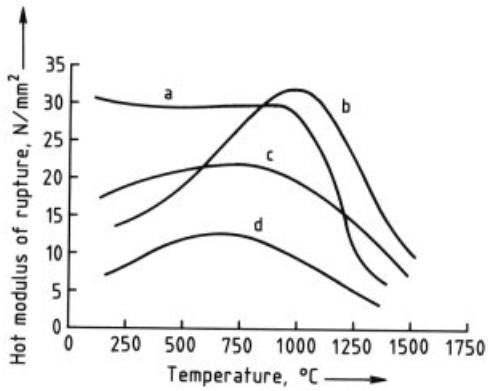

Figure 16. Hot modulus of rupture of low-cement castables, ramming mix, and dense brick (corundum-based materials) [127]
a) LCC; b) ULCC; c) Brick; d) Dry ramming mix

The introduction of low- cement (LCC) and ultra-low- cement (ULCC) castables in the late 1960s led to completely new unshaped materials. Due to a drastic decrease in cement content

Table 17. Types of unshaped refractory materials

Type	Subdivision	Water content upon application, %	Nature of bond
Moldable mixes	dry ramming mixes (fritting mixes)	0 – 1	ceramic
	moist friable ramming mixes	3 – 6	ceramic, chemical, organic
	plastic friable mixes	5 – 8	ceramic, chemical, organic
	preformed plastic mixes	6 – 10	ceramic, chemical, organic
Refractory castables	dense castables	5 – 12	hydraulic, chemical
	insulating castables	10 – 50	hydraulic, chemical
Gunning mixes	moist friable mixes		hydraulic, chemical, ceramic
	different types of castables	5 – 50	hydraulic, chemical, ceramic
	special hot gunning mixes		hydraulic, chemical, ceramic (special binders)

Table 18. Types of dense refractory castables

Type	Classification according to ASTM C 401-89	Mixing water (approx.)	Essential components, bonding (see Table 3)
Conventional castable (CC)	> 2.5 % CaO	9 – 12	main hydraulic bonding: $CA^a + 10\,H^b \rightarrow CAH_{10}{}^c$ (< 20 °C) $2\,CA + 11\,H \rightarrow C_2AH_8{}^d + AH_3{}^e$ (20 – 35 °C)
Low-moisture conventional castable [122]		7 – 9	addition of ultrafine powders (high strength and abrasion resistance)
Low- cement castable (LCC)	1 – 2.5 % CaO	4.5 – 7	deflocculated castables containing aluminous cement, a minimum of 2 % ultrafine particles (SiO_2, Al_2O_3; < 1 micron) and at least one deflocculating agent
Ultralow- cement castable (ULCC)	0.2 – 1 % CaO		thixotropic rheological behavior (vibrational casting)
Clay-bonded castable	see LCC and ULCC	7	contains clay (flocculated) and aluminous cement as a deflocculant
Self-flowing castable [123]		6	consistancy (rheopex rheological behavior), which allows flow and deaeration without vibration
No- cement castable	< 0.2 % CaO		bonding: phosphate, alkali silicate, pitch or resin, Si, Al sol
Mixerless castable [124]			delivered ready-mixed with water for casting (vibrating); contains 0.5 % sodium silicate, thermosetting at 130 – 200 °C by gelation

[a] $CA = CaO \cdot Al_2O_3$. [b] $H = H_2O$. [c] $CAH_{10} = CaO \cdot Al_2O_3 \cdot 10\,H_2O$. [d] $C_2AH_8 = 2\,CaO \cdot Al_2O_3 \cdot 8\,H_2O$. [e] $AH_3 = Al_2O_3 \cdot 3\,H_2O$ (gel)

together with the presence of ultrafine powders (SiO_2, reactive Al_2O_3) and dense packing these castables require little mixing water and display low porosity and high cold and hot strength (due to e.g., mullitization of the matrix).

The composition of the binding-agent mixture of modern low- cement castables is based on a patent from the year 1978 [125]. This patent is the first to mention reduction of the cement content to 1 %. The increase in packing density is achieved by targeted changes in the microstructure through the use of a liquefier and adjusted grain sizes in the super-fine range (< 10 µm).

Figures 15 and 16 describe properties of modern castables, classifying them within the broader context of unshaped materials and comparing them to shaped products.

Conventional fireclay and high-alumina castables develop an anorthitic/gehlenitic ceramic bond when heated above ca. 1100 °C. In conventional corundum castables a high-temperature-resistant calcium aluminate bond develops. The addition of fine SiO_2 powder leads to mullite in low- cement castables, increasing their hot strength.

Compared to conventional castables, modern castables offer better corrosion resistance (due to a lower CaO content), low porosity, and small pore size [126]. Drying and first heat-up steps require careful monitoring to prevent explosive effects caused by excessive vapor pressure inside the lining. For the rheological and setting behavior (hydration), drying, and heat-up of castables see [6], [26]. Prior to heating a castable must be cured for a sufficiently long time (i.e., 24 h) to ensure full hydration and strength. Workability and setting time can be adjusted by the presence of small amounts (e.g., 0.1 %) of special additives. Modern castables require exact dosages of additives and mixing water, thorough mixing, and appropriate forming to achieve high-quality linings or shapes.

Gunning materials are specially prepared for placement by pneumatic or mechanical projection means in cold and also hot application. These may be castables to which water has been added during or prior to gunning, or they may be plastics. Various special bonds have been developed for achieving a dense and adhesive gunned layer. Such materials are used mainly for repairs, especially in the iron and steel industry.

Prefabricated shapes [128] are rammed or cast, and may be pretreated by the manufacturer as a way of providing blocks that are ready for use. Relatively large-volume parts are fabricated and preheated up to 600 °C at the refractory plant. The advantages of prefabricated shapes are:

1) *Optimal Manufacture*. Production of the prefabricated shapes benefits from the special know-how of refractory producers, enabling them to avoid faults during manufacture.
2) *Simple Lining*. Faults in the resulting lining are avoided as a consequence of simplified assembly.
3) *Short Installation Times*. This reduces the cost of refractory lining, although special mechanical equipment is required for handling.
4) *No Drying*. Prefabricated shapes have already been tempered, so on-site drying of the refractory material is not necessary (except for a brief preheating), which reduces energy costs considerably.

Jointing and coating materials are essential for establishing a safe refractory brickwork or lining surface. Such products are fine-grain materials whose characteristics and composition must correspond to those of the associated bricks. Table 19 describes such materials.

Table 19. Joints and coatings [3]

Material	Mixing liquid or water, %	Type of bonding
Refractory mortars	10 – 25	chemical, hydraulic, ceramic
Refractory mastics	12 – 30	chemical
Adhesives	30 – 80	chemical
Trowelling materials	15 – 30	chemical, hydraulic, ceramic
Coatings	30 – 60	chemical

Laying and jointing materials can be classified as follows:

1) *Refractory mortars,* with ceramic bonding starting at 800 °C and chemical or hydraulic bonding starting at 20 °C
2) *Refractory mastics,* with chemical bonding (i.e., high strength) above 110 °C
3) *Adhesives,* with chemical bonding as for refractory mastics, but finer grain size

The lining procedure has been mechanized, and is in most cases very efficient. Various lining methods have been developed. The properties and service behavior of a monolithic lining depend to a great extent on careful and appropriate preparation of the mix, as well as proper lining procedure, curing, drying, and first heat-up.

Applications: principal applications of materials for monolithic construction are in the iron and steel industry, foundries, the petrochemical industry (especially insulating gunning mixes), and incinerators.

7.3. Heat-Insulating Ceramic Materials [129]

Ceramic materials with a high pore volume (i.e., low thermal conductivity) are used extensively for thermal insulation of furnaces and auxiliary equipment. Apart from the effect of decreased thermal wall losses, the low weight of the insulating ceramic material allows lighter furnace construction and low heat-storage losses in discontinuous service.

Thermal conductivity depends not only on total porosity (overall gas volume) but also to a considerable extent on pore size and pore distribution, as well as texture ("thermal bridges") and mineralogical composition. Normally, the finer the pores, the lower the thermal conductivity. A maximum pore size of < 1 mm is the target. Microporous materials have a thermal conductivity lower than air. Above 1000 °C thermal radiation must be taken into account as a heat-transfer mechanism.

The optimal range for porosity or bulk density of a thermal insulating material can be chosen on the basis of the effective conditions for heat transport as a function of temperature [130].

High porosity implies low mechanical strength, high gas permeability, and low corrosion resistance, all of which must be considered in the use of insulating refractory materials. Furthermore, a porous texture leads to lower volume stability at high temperatures than in the case of dense materials (greater sensitivity to over-heating) and a greater susceptibility to reducing atmospheres (reduction of, e.g., Fe_2O_3; increased shrinkage) [131].

The main groups of insulating ceramic products are:

1) Insulating bricks
2) Insulating refractory bricks
3) Ceramic fibers
4) Porous or dense grains and granules of ceramic materials used as poured back-up insulation
5) Insulating castables (see Section 7.2)
6) Special insulating products based on calcium silicate and microporous high-SiO_2 materials,with application temperatures ≤ 1100 °C

Insulating ceramic materials are used mainly for rear lining behind a dense wear lining. Bricks and ceramic fiber products can be placed in direct contact with heat or a flame provided the furnace walls are not under high mechanical stress, there is no corrosion, and application limits are not exceeded.

7.3.1. Insulating Bricks

Insulating bricks are based on diatomite (moler), vermiculite, and perlite, with bulk densities of ca. 0.45 – 0.8 g/cm^3, cold crushing strength of ca. 1 – 7 N/mm^2, and a thermal conductivity at 200 °C of 0.1 – 0.2 W m^{-1} K^{-1}. They are normally used up to 800 °C, but low-flux diatomite bricks may be applicable to 1100 °C. Diatomite bricks have very fine pores, low thermal conductivity, and relatively high strength and abrasion resistance.

7.3.2. Insulating Refractory Bricks

According to official definitions, insulating refractory bricks have a total porosity of at least 45 % and an application temperature of at least 800 °C. Porosity normally lies between 60 and 80 %.

The criteria for classification (ISO 2245) are as follows:

1) The temperature at which the material does not show a permanent linear shrinkage greater than 2 %
2) The bulk density (which for a given product is directly related to porosity, and consequently to thermal conductivity)

Classification temperatures are not to be taken as acceptable service temperature limits, which depend on the particular service conditions.

Insulating refractory bricks are usually fabricated from fireclay (chamotte), sillimanite or mullite, corundum, and silica as raw materials, together with a combustible material. Special types include anorthitic-bonded fireclay bricks (ca. 15 % CaO) and bricks made from bubble alumina.

Types and properties of insulating refractory bricks are summarized in Table 20.

7.3.3. Refractory Ceramic Fibers [132–134]

(→ Fibers, 5. Synthetic Inorganic)

A proposed definition describes ceramic fibers as synthetic mineral fibers suitable for use as heat-insulating materials and containing less than 2 wt % alkali metal or alkaline earth oxides. Most ceramic fibers are based on glassy or crystalline aluminosilicates and crystalline corundum.

Table 20. Properties of insulating refractory bricks

Type	Classification temperature, °C	Al_2O_3, %	SiO_2, %	Fe_2O_3, %	Bulk density, g/cm^3	Porosity, %	CCS ∗, N/mm^2	Thermal expansion, 1000 °C, %	Thermal condutivity, 1000 °C, W m^{-1} K^{-1}
Silica	1500 – 1600	1	92 (3.5 % CaO)	< 0.5	0.8 – 1.2	50 – 65	3 – 4	1.2	0.5 – 0.8
Anorthite fireclay	1100 – 1250	36	45 (15 % CaO)	0.7	0.5	85	1	0.6	0.2
Fireclay	1250 – 1400	35 – 42		< 1.7	0.55 – 1.3	55 – 75	2 – 8	0.5 – 0.6	0.3 – 0.5
High-alumina 50 – 75 % Al_2O_3	1400 – 1700	55 – 75		< 1.5	0.8 – 1.5	45 – 75	2 – 12	0.6 – 0.7	0.35 – 0.55
Corundum hollow sphere	1800	98	0.6	0.1	1.2 – 1.4	60	8 – 18	0.8	0.8 – 1.1
Forsterite	1500	MgO 57	34	6	1.7	45	2.5	1.2	0.8

∗ Cold compressive strength.

Glassy (*amorphous*) *refractory ceramic fibers* (Al_2O_3 content ca. 45 – 60 %) are produced by fusing an aluminosilicate raw material (e.g., calcined kaolin) and/or sand and calcined alumina. Addition of ZrO_2, Cr_2O_3, B_2O_3, and other materials leads to changes in the properties of the fiber. The molten stream is attenuated into fiber form either by impact with a high-velocity jet of air and/or steam or by downward flow onto rotating wheels.

Polycrystalline fibers (62 – 100 % Al_2O_3) are sol – gel derived fibers from, e.g., aluminum salts, colloidal silica, organosilane, and modifiers. Rotary spinning (short staple fibres) is also practiced, as is the production of continuous filaments. After formation the fiber must be dried and calcined. Other types of oxide fibers (e.g., ZrO_2 fiber) are produced in the same way. This type of fiber is relatively expensive. To improve cost-effectiveness, polycrystalline fibers are often blended with low- cost glassy fiber. Properties and application temperatures for fibers and fiber products are shown in Table 21.

The fiber is used in loose form or processed into any of a variety of flexible or essentially rigid forms. Blankets and textiles (ropes, yarns) are examples of products made by an air-laid system. Papers, felts, and boards are produced by wet-laid systems (like the traditional paper-forming process). Boards and special shapes are processed via vacuum casting. The ceramic fiber is slurried in water with organic and/or inorganic binders (e.g., SiO_2 sol) and fillers. The slurry is dewatered onto molds via a vacuum. Fiber is also used in unshaped refractory materials like gunning mixes and castables. Products include ceramic-fiber veneering elements, folded modules, and ready-for-use system blocks with anchoring.

Ceramic-fiber products have very low bulk density (0.05 – 0.3 g/cm^3; i.e., a porosity > 90 %) and unusually low thermal conductivity. In contrast to insulating bricks, thermal conductivity of ceramic-fiber products decreases with increasing bulk density. Thermal conductivity increases more with increasing temperature than in the case of comparable insulating bricks (Fig. 17). Thermal shock resistance is extraordinarily high.

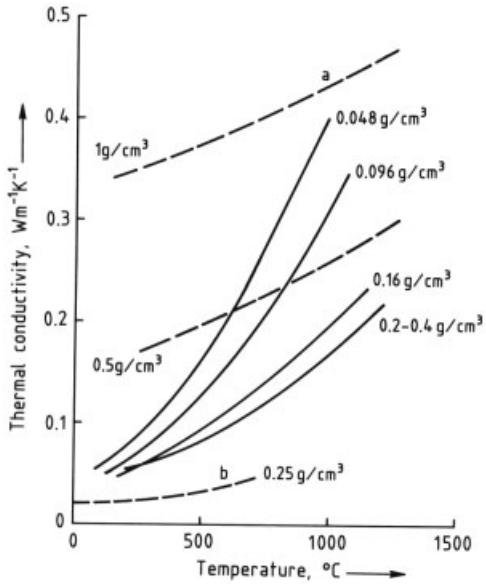

Figure 17. Thermal conductivities of insulating bricks a) Fireclay insulating brick; b) Microporous material

Table 21. Ceramic fibers and fiber products[a]

Type of fiber	Al_2O_3, %	SiO_2, %	Cr_2O_3, %	ZrO_2, %	Liquidus temp., °C	Classification temp., °C	Fiber diameter, µm	Density, g/cm^3	Maximum service temperature (°C)		
									Oxidizing atmos., continuous	Oxidizing atmos., discontinuous	Reducing atmos.
Glassy, amorphous	47	52			1760	1250	2–5	2.58	1100	1250	< 1100
	63	36			1800	1400	3–6	2.63	1200–1400	1400 (1500[b])	ca. 200° lower than in oxidizing atmos.
	41	55	3		1785	1400	2–5	2.61	1200–1400	1400 (1500[b])	ca. 200° lower than in oxidizing atmos.
	35	50		15	1810	1450	≈3	2.60	1200–1400	1400 (1500[b])	ca. 200° lower than in oxidizing atmos.
Polycrystalline	85	15			1950	1600	≈7	3.40	1500	1600 (1700)	ca. 100° lower than in oxidizing atmos.
	95	5			2000	1600	≈3	3.58	1500	1600 (1700)	ca. 100° lower than in oxidizing atmos.

[a] For fiber products in general, the bulk density is 0.05–0.3 g/cm^3 (special products, ≤ 1.3 g/cm^3). Thermal conductivity ranges are 0.08–0.12 W m^{-1} K^{-1} (400 °C) and 0.16–0.40 W m^{-1} K^{-1} (1000 °C). Fiber tensile strength 400–1500 N/mm^2. [b] Mixed fiber from high-alumina glassy fiber and alumina fiber.

Ceramic-fiber products are classified into groups according to the temperature (in 50 °C steps) at which the material does not show more than 4 % shrinkage.

Long-time application of ceramic fibers at maximum service temperature is limited significantly by shrinkage, crystallization and/or recrystallization (crystal growth). Up to 75 % mullite (beginning at 950 °C) and up to 40 % cristobalite (beginning at 1200 °C) can crystallize from amorphous fibers. At high temperatures the fiber becomes brittle, fragile, and subject to crumbling. Compressibility and elasticity are also lost. A crystalline alumina fiber shows crystal transformations (γ-Al_2O_3 to α-Al_2O_3) and crystal growth. Normally, alumina fibers contain about 5 % SiO_2, which is introduced to slow crystal growth; mullite (ca. 20 %) is also formed. In mixed products from glassy and alumina fibers the formation of mullite in service decreases shrinkage [135]. Mullite fibers are more volume-stable at high temperatures [136]. Impurities acquired from a furnace atmosphere (e.g., alkalies) may greatly increase shrinkage and crystal transformation. Ceramic-fiber products are sensitive to mechanical stress as well (e.g., to high-velocity gas streams). Improvement is achieved by brush-like orientation of the fibers perpendicular to abrasion stresses, or by impregnating the surface with SiO_2 sol or the use of rigid ceramically-bonded and prefired shapes.

Applications. Ceramic fibers have found application in all types of low- and high-temperature thermal processes, from use as a filler for joints or seals and insulation over veneering of furnaces walls up to complete fiber-lined furnaces (e.g., in the ceramics industry, and heating furnaces for the steel and glass industries). Most ceramic-fiber products have a service limit in long-time applications of ca. 1500 °C, although in special cases they may be used discontinuously up to 1750 °C.

In the nonferrous industry preshaped parts made from ceramic fibers have found application in direct contact with molten aluminum and copper alloys.

Proper construction and the use of appropriate joining techniques (glueing procedures, anchoring systems) are important factors in the lifetime of fiber furnace linings.

8. Selection, Use, and Industry-Specific Consumption of Refractory Ceramics

The service performance of a refractory material depends in roughly equal measure on:

1) The properties of the refractory material
2) The construction, lining and brick design, installation technique, drying, and initial heating
3) Service conditions

Most data sheets accompanying refractory products specify a service temperature limit or maximum service temperature. This refers to the maximum temperature to which the refractory can be exposed during operation without corrosive influences (heating on one side only). No test yet exists for determining the maximum service temperature. Consequently, the indicated temperatures are based on practical experience supported by suitable hot tests (refractoriness under load, hot strength, reheat change). Classification temperatures for refractory products are not to be considered as service temperature limits, which are usually lower.

In most cases, refractory ceramics are subject to a multitude of thermal, thermomechanical, mechanical, and corrosive stresses. Corrosion is the factor most likely to change the service properties of a refractory material in a significant way. A refractory material must be selected in accordance with those properties that are most important for a given service application. This requires close consultation between the refractory producer, the furnace builder, and the user. Often a refractory ceramic is "tailor made" for a specific application.

The output (tonnage) of refractory products is distributed among various industries in approximately the following way:

Iron and steel industry	55 – 60 %
Cement and lime industry	5 – 8 %
Glass industry	5 – 8 %
Ceramic industry	5 – 8 %
Nonferrous industry	2 – 5 %
Chemical and petrochemical industry	2 – 4 %
Energy-producing industry	2 – 4 %

A high proportion of unshaped refractories (ca. 80 %) is directed toward the iron and steel industry and foundries. Other consumer industries, with the exception of the petrochemical industry, use mainly shaped refractories. Jointing materials, which are included among the unshaped refractories, account for ca. 3 – 5 % of total refractory production. Table 22 provides information on consumption of refractory materials by the iron and steel industry.

The glass industry consumes ca. 0.5 – 1 kg of silica bricks, 1 – 3 kg of basic bricks, and 1 – 3 kg of fusion- cast bricks per tonne of manufactured glass. Consumption in the cement industry is ca. 0.4 – 1.2 kg of refractory per tonne of cement clinker.

Ceramic fibers are used to the extent of ca. 40 % in the process industry (industrial furnaces, kiln construction, and thermal processing) and ca. 25 % in the iron and steel industry [137].

Basic information regarding refractory applications in industrial furnaces, kilns, and ovens, including typical service conditions, is provided in [1]. For a short general overview of the application of refractories in various industries see [138], [139]; for details regarding construction and lining see [14] and [6].

The following list summarizes comprehensive references to the use of refractory ceramics in several industries:

Iron and steel industry	[8], [140–145]
Foundries	[7]
Nonferrous metal industry	[146]
Coke ovens	[62–64]
Cement and lime industry	[147], [148]
Glass industry	[9], [61], [110], [149]
Ceramic industry	[150]
Chemical and petrochemical industry	[151–153]
Energy-producing industry, incinerators	[151], [154], [155]

9. Toxicology and Occupational Health

Industrial hygiene and environmental pollution considerations in the refractory industry are increasingly the subject of national and international regulations and laws. Maximum concentrations have been established for emissions during the fabrication of refractories (gases such as fluorine, chlorine, and sulfur compounds in flue-gas, dust particles) [156] [157] and for concentrations of hazardous substances in the air at

Table 22. Consumption of refractories by the iron and steel industry, expressed in kilograms of refractory material per tonne of crude steel

Country	Bricks		Unshaped refractories	
	1985	1991	1985	1991
Federal Republic of Germany	10.7	8.0	6.4	5.8
Japan	7.3	5.5	6.3	6.1

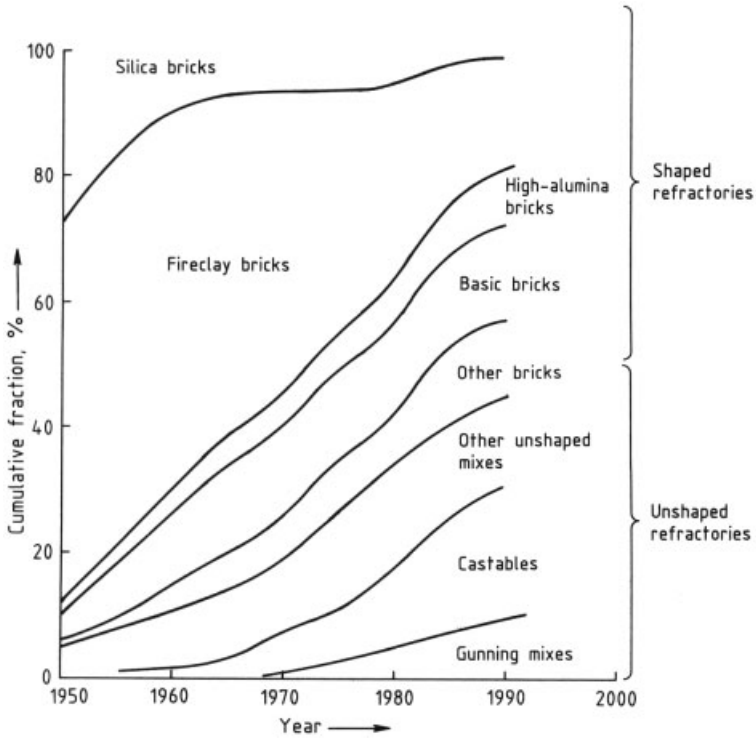

Figure 18. Evaluation of the production of different types of refractories

the workplace [158], [159]. Irritating, harmful, and toxic substances in raw materials (especially binders) and refractory products are indicated in product safety data sheets and marked on packages.

The main groups of hazardous or potentially hazardous substances are: chromium compounds (hexavalent chromium and soluble materials) [160]; SiO_2, i.e., amorphous and all crystalline forms of silica; organic binders (especially pitch) [29], [161]; fibers [162]; boron oxide; and oxides of heavy metals (traces).

10. Economic Aspects [163–166]

The refractories industry represents an independent branch of the ceramic industry. According to [1], about 800 manufacturers worldwide produce refractory ceramics. About 300 manufacturers in the EC produce refractory ceramics. Competition exists between large, increasingly international corporations, medium-sized and small companies, with the latter especially active in the production of unshaped refractory materials. Refractory production data from several regions are presented in Table 23.

The estimated value of refractory production in the EC during 1986 was € 2.8×10^9, ca. 36 % of which could be attributed to the Federal Republic of Germany.

There has been a steady decline in refractory production in all industrialized countries since about 1960, which can be ascribed to a reduction in specific consumption in most user industries together with technological advances resulting

Table 23. Production data for refractory materials

Source	Output, 10^6 t					
	Bricks		Unshaped refractories		Total	
	1988	1991	1988	1991	1988	1991
EC	3.1	2.8	1.8	1.6	4.9	4.4
USA	1.2		0.9		2.1	
Japan	0.95	0.9	0.8	0.8	1.75	1.7
USSR					ca. 10	

in higher quality refractories with a longer life span. Further reduction of overcapacity is necessary.

A shift from mass production of generalized refractories to high-quality and specialized products with diverse properties and applications, and to high-priced "high tech" refractories [167], is quite apparent. Figure 18 provides relative evaluations of different types of refractories used in Japan, which serves as an example of a highly industrialized country. The steady increase in the proportion of unshaped refractories is worth noting.

The following list offers relative price comparisons for several types of refractory supplied as standard-sized bricks:

Fireclay (conventional)	1
Fireclay castable (conventional)	1
High-quality fireclay brick	1.5
Bauxite brick	2
Corundum brick (85 % Al_2O_3)	5
Magnesia brick (conventional)	2.5
High-quality magnesia brick	4.5
Magnesia graphite brick	5
Dolomite brick	0.9
Zircon silicate brick	4
SiC brick (95 % SiC)	7
Fusion- cast brick	10

Refractory Ceramic Fibers. According to [137], total global capacity for ceramic-fiber production increased in 1990 to ca. 190 000 t/a (1.9×10^6 m^3/a) for fibers containing 45 – 60 % Al_2O_3 and about 1500 t/a (15 000 m^3/a) for fibers with 72 – 97 % Al_2O_3. Average selling prices range from ca. € 26/kg for loose glassy fibers to € 307/kg for high-temperature boards and shapes.

Further developments in refractory materials are subject to the following influential factors [163], [165], [168–170]: consumer demand, availability, properties and price of raw materials, automation, and environmental restrictions.

11. References

General References

1. S. C. Carniglia, G. L. Barna: *Handbook of Industrial Refractories Technology. Principles, Types, Properties and Application,* Noyes Publ., Park Ridge, N.J., 1992.
2. W. Schulle: *Feuerfeste Werkstoffe. Feuerfestkeramik – Eigenschaften, prüftechnische Beurteilung, Werkstofftypen,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1990.
3. Didier-Werke AG, "Feuerfeste Werkstoffe und ihre Merkmale," company brochure, Wiesbaden 1990.
4. H. Salmang, H. Scholze: *Keramik,* part 1, "Allgemeine Grundlagen und wichtige Eigenschaften,"Springer Verlag, Berlin 1982; part 2, "Keramische Werkstoffe," Springer Verlag, Berlin 1983.
5. J. H. Chesters: *Refractories. Production and Properties,* The Iron and Steel Institute, London 1973.
6. *Technology of Monolithic Refractories,* Plibrico Japan Comp., Tokyo 1984.
7. K. E. Granitzki, W. Krönert, E. Müller: *Feuerfeste Stoffe im Gießereibetrieb,* Gießerei-Verlag, Düsseldorf 1989.
8. J. H. Chesters: *Refractories for Iron- and Steelmaking,* The Metals Society, London 1974.
9. O. Verworner, K. Berndt: *Feuerfeste Baustoffe für Glasschmelzanlagen,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1977.
10. L. D. Hart (ed.): *Alumina Chemicals Science and Technology Handbook,* American Ceramic Society, Inc., Westerville 1990.
11. *A Handbook of Ceramics. Ceramic Monographs,* Verlag Schmid GmbH, Freiburg, issues beginning 1981.

12. S. Kienow: "Refractory Materials," in [11], part 2.6, 1979, pp. 1 – 18.
13. "Ceramics and Glasses," *Engineering Materials Handbook,* **vol. 4,** The Materials Information Society, 1991, p. 1217.
14. Deutsche Gesellschaft Feuerfest- und Schornsteinbau e.V.: *Feuerfestbau. Stoffe – Konstruktion – Ausführung,* Vulkan-Verlag, Essen 1987 (new edition in preparation).
15. "Global Advances in Refractories," *Unitecr '91 Congress,* 2nd ed., Verlag Stahleisen, Düsseldorf 1992.

Specific References

16. *Am. Ceram. Soc. Bull.* **71** (1992) no. 5, 779 – 824.
17. B. M. Coope, E. M. Dickson (eds.): *Raw Materials for the Refractory Industry,* "IM Refractories Survey 1981," Metal Bulletin Ltd., London 1991.
18. G. Routschka: "Refractory Raw Materials for the Production of Chamotte and High-Alumina Refractory Products – A Survey," *ZI Int.* 1981, no. 12, 690 –704.
19. P. Ney: "Synthetic Ceramic Raw Materials," in [11], pp. 1 – 26.
20. G. K. Sargeant: "Holosphere – A Novel Ceramic Material," *Br. Ceram. Trans. J.* **90** (1991) no. 4, 132 –135.
21. K. Riegel, H. Barthel: "Large-Crystal Magnesia Clinker for Advanced Refractories, an Update and Overview," in [15] pp. 97 – 101.
22. F. Mohri, M. Kobune, H. Tsuri: "The Present State and Future Out-Look of Electrofused Magnesia," *Taikabutsu Overseas* **11** (1991) no. 4, 53 – 57.
23. K.-H. Mehrwald: "Entwicklung und wirtschaftliche Aspekte der technischen SiC-Herstellung," *CFI Ceram. Forum Int.* **69** (1992) no. 3, 72 – 81.
24. N. Yakushige, T. Korenoby, M. Ohe: "Properties of Crystalline Graphite," *Taikabutsu Overseas* **11** (1991) no. 4, 41 – 45.
25. B. Teyton: "Cordierite Chamotte Manufacturing," *Ind. Ceram.*(1988) no. 823, 30 – 33.
26. Recent Progress on Castable Refractories," *Taikabutsu Overseas* **9** (1989) no. 1, 2 – 75.
27. J. E. Kopanda, G. MacZura: "Production Processes, Properties, and Applications for Calcium Aluminate Cements," in [10]pp. 171 – 183.
28. H. G. Emblem: "The Use of Ethyl Silicate as a Binder in Refractory Technology," *Trans. J. Br. Ceram. Soc.* **74** (1975) 223 – 228.
29. V. in der Wiesche et al.: "Synthetic Resins as Carbon Forming Agents for Various Refractories," in [15] pp. 371 – 380.
30. T. Kawakami, M. Yoshimura, J. Yamamoto: "Properties of Silica Flour," *Taikabutsu Overseas* **5** (1985) no. 4, 21 – 25.
31. A. Watanabe, H. Takahashi: "Effect of Stainless Steel Fibers on Thermal Properties of Castable Refractories," *Taikabutsu Overseas* **11** (1991) no. 4, 26 –30.
32. T. R. Kleeb, J. A. Caprio: "Properties and Service Experience of Organic Fiber-Containing Monoliths," *Ind. Heat.* **54** (1987) no. 4, 19 – 21; no. 6, 35 –36.
33. J. E. Funk, D. R. Dinger: "Particle Packing," *Inter- Ceram* **41** (1992) no. 1, 10 – 14; no. 2, 95 – 97; no. 3, 176 – 179.
34. J. Kriegesmann (ed.): *DKG – Technische keramische Werkstoffe,* Lose-Blatt-Sammlung, Wirtschaftsdienst, Köln 1989.
35. D. A. Larsen, T. Weichert: "Superior Refractories by Freeze Molding for Super Alloys and Steel Casting" in [15] pp. 468 – 472.
36. Fédération Européene des Fabricants de Produits Réfractaires (PRE) Brüssel: Refractory Materials. Recommendations 1990 (contains all ISO and PRE classifications and testing standards and recommendations referring to refractory materials).
37. A. Pertl: "Quality Assurance for the Manufacture of Refractories," *InterCeram,* special issue, Aachen Proceedings 1989 (1990) 82 – 84.
38. "Refractories Carbon and Graphite Products, Activated Carbon," *Annu. Book ASTM Stand,* vol. 15. 01. 1992.
39. J. F. Clements, J. Vyse: "A Comparison of Methods for Measuring Pore-Size Distribution in Refractories," *Trans. Br. Ceram. Soc.* **67** (1968) 285 –317.
40. C. Gelines, R. Angers: "Improvement of the Dynamic Water-Expulsion Method for Pore Size Distribution Measurements," *Am. Ceram. Soc. Bull.* **65** (1986) no. 9, 1297 – 1300.
41. H. L. Le Doussal, C. Prieur: "Les nouveaux réfractaires. Aspect, caractérisation et normalisation," *Ind. Ceram.*(1989) no. 837, 255 – 262.
42. L. Hagemann, G. Flohé: "Untersuchungen zur Wärmeleitfähigkeit (WLF) elektrisch leitender, insbesondere kohlenstoffhaltiger

feuerfester Werkstoffe," *Keram. Z.* **41** (1989) no. 9, 639 – 643.
43. P. Dietrichs: "Qualitätskriterien und Prüfmethoden keramischer Fasern und Faserprodukte," *Silikattechnik* **38** (1987) no. 10, 349 – 353.
44. D. A. Bell: "Non-Destructive Testing of Refractories," *Br. Ceram. Trans. J.* **88** (1989) 133 – 137.
45. H. Röpke, H. Hädrich, D. Küstner: "Non-Destructive Testing of Fused Cast Blocks for Glass Tank Furnaces by Ultrasonic Methods," *InterCeram,* special issue 1986, pp. 73 – 74.
46. S. Ulitzka: "Non-Destructive Testing of Bonded Refractories in the Glass Industry," in [150] pp. 85 – 86.
47. E. Hasebe, Y. Hayashi, Y. Aiba: "Application of X-Ray Inspection System to Refractories," *Taikabutsu Overseas* **10** (1990) no. 2, 85 – 98.
48. VDI-Richtlinien 3128, Blatt 4, Induktive Erwärmung. Physikalische Stoffeigenschaften nichtmetallischer Werkstoffe (Ofenbaustoffe), Jan. 1977.
49. K. Konopicky, G. Routschka, L. Hagemann: "Druckfließverhalten von Schamotte- und hochtonerdehaltigen feuerfesten Steinen," *Arch. Eisenhüttenwes.* **42** (1971) no. 6, 433 – 437.
50. R. G. Young, F. J. Hartwig, C. L. Norton: "Effect of Various Atmospheres on Thermal Conductivity of Refractories," *J. Am. Ceram. Soc.* **47** (1964) no. 5, 205 – 210.
51. H. Le Doussal, G. Bisson: "La measure de la conductivité thermique des produits refractaires," *L'ind. Céram.* (1981) no. 4, 283 – 294; no. 5, 349 – 360.
52. G. Routschka: "Thermal Conductivity of Refractory Castables," *InterCeram* **37** (1988) no. 3, 24 – 28, 33.
53. M. S. Glazman, Y. A. Landa, E. Y. Litovshii, N. A. Puchkelevich: "Spectral Emissivity of Refractories (Review article)," *Refractories (Engl. Trans.)* **30** (1989) no. 5, 312 – 319.
54. W. Bauer, R. Steinhardt: "Emissionsgrade feuerfester Baustoffe," *Gas Wärme Int.* **39** (1990) no. 9, 388 – 391.
55. T. K. Das, P. Jeschke: "Spannungen und Verformungen in feuerfestem Mauerwerk," *Ber. Dtsch. Keram. Ges.* **52** (1975) no. 5, 126 – 130.
56. D. P. M. Hasselmann: "Figures-of-Merit for the Thermal Stress Resistance of High-Temperature Brittle Materials: A Review," *Ceramurgia Int.* **4** (1978) no. 4, 147 – 150.
57. S. Uchida et al.: "Fracture of Al_2O_3 – C Submerged Nozzle," *Taikabutsu Overseas* **10** (1990) no. 2, 74 –84.
58. M. Kobayashi, M. Nishi, A. Miyamoto: "Slag Resistance Tests for Refractories," *Taikabutsu Overseas* **2** (1982) no. 2, 5 – 13.
59. K. H. Obst, J. Stradtmann, W. Münchberg: "Verfahren zur Verschlackungsprüfung im Labor- und halbtechnischen Maßstab," *Tonind. Ztg. Keram. Rundsch.* **93** (1969) no. 10, 380 – 385.
60. J. Yoshitomi, M. Harada, T. Eguchi, K. Mukai: "Local Corrosion of Iron Trough Refractories in the Slag – Metal Interface," *InterCeram,* special issue Aachen Proceedings 1988, pp. 85 – 89.
61. K. Konopicky et al.: "Silikasteine für Glasschmelzofen," *Fachausschußbericht no. 66,* Dt. Glastechn. Ges., Frankfurt/M. 1968.
62. P. Jeschke et al.: "Superdense Silica Bricks for Modern Coke Ovens," in *Unitecr '89 Proceedings,* vol. 1, pp. 263 – 283, American Ceramic Society Inc., Westerville 1989.
63. K. E. Lepère, E. Overkott, R. Beckmann: "Requirements with Respect to Silica Bricks as Building Material for the Next Coke Oven Generation," in *Unitecr '91 Congress,* 2nd ed., Verlag Stahleisen, Düsseldorf 1992, pp. 450 – 453.
64. Y. Naruse, Y. Hoshino, T. Tanaka: "Silica Bricks for Coke Ovens," *Taikabutsu Overseas* **2** (1982) no. 1, 110 – 120.
65. P. Schulz: "Schlickergießen von amorpher Kieselsäure von großformatigen und komplizierten Teilen," *Keram. Z.* **38** (1986) no. 7, 442 – 445.
66. J. A. Pask: "Critical Review of Phase Equilibria in the Al2O3 – SiO2 System," in *Ceramic Transactions,* vol. 6, "Mullite and Mullite Matrix Composites," American Ceramic Society, Westerville 1990, pp. 1 –13.
67. Stahl-Eisen-Werkstoffblätter: *SEW 917: Schamottesteine für allgemeine industrielle Einsatzzwecke (Steingruppe A bis 45 % Al_2O_3) and SEW 915: Schamottesteine für Hochöfen und Winderhitzer,* Verlag Stahleisen, Düsseldorf 1984.
68. K. Konopicky: "Allgemeines zum Aufbau der Schamottesteine," *Ber. Dtsch. Keram. Ges.* **36** (1959) no. 11, 367 – 371; "Zur Theorie der Schamotte-Erzeugnisse I," *Ber. Dtsch. Keram. Ges.* **32** (1955) no. 9, 257 – 261.
69. In [10] pp. 427 – 469.

70. E. Thomas, D. Patel: "The Application of Bonded Alumina – Chrome Refractories in the Glass Industry," *InterCeram* **35** (1986) special issue, 50 – 53.
71. H. G. Emblem, T. J. Davies, A. Harabi, V. Tsantzalou: "Alumina – Chrome Refractories," in [15] pp. 415 – 421.
72. K. Konopicky et al.: "Untersuchungen an Steinen mit mehr als 50 % Al_2O_3," *Ber. Dtsch. Keram. Ges.* **42** (1965) no. 5, 213 – 221.
73. H. Schneider, J. Wang, A. Majdic: "Thermal Expansion of Refractory-Grade Bauxites at High-Temperatures," *CFI Ceram. Forum Int.* **63** (1986) no. 9/10, 461 – 470.
74. Stahl-Eisen-Werkstoffblatt SEW 912: *Tonerdereiche Steine,* Verlag Stahleisen, Düsseldorf 1984.
75. A. M. Alper (ed.): *High Temperature Oxides. H 1. Magnesia, Lime, and Chrome Refractories,* Academic Press, New York 1970.
76. Stahl-Eisen-Werkstoffblatt SEW 913: *Basische Steine,* Verlag Stahleisen, Düsseldorf 1984.
77. J. Stradtmann, R. Prange, M. Rähder: Neuentwicklungen auf dem Gebiet dolomitischer Pfannenzustellungen, *Fachber. Hüttenprax. Metallweiterverarb.* **25** (1987) no. 10, 907 – 913.
78. F. Kawano et al.: "CaO Clinker with Improved Anti-Hydration Property," *Taikabutsu Overseas* **11** (1991) no. 3, 29 – 36.
79. E. Marino: "Use of Calcium Oxide as Refractory Material in Steel Making Processes" in *Refractories for the Steel Industry,* Elsevier Applied Science, London 1990, pp. 59 – 68.
80. A. Ikesue et al.: "Formation of Precipitated-Complex Spinel in Magnesia – Chrome Refractories and its Characteristics," *InterCeram* **41** (1992) no. 6, 406 – 412.
81. K. Asano et al.: "Corrosion of Fired Magnesia Chromite Bricks by CaO – CaF2 Slags and Development of Highly Durable Magnesia Chromite Brick," in *Vortragsband XX–XIII. Inter. Feuerfest-Kolloquium,* Aachen 1990, pp. 317 – 327.
82. R. Dal Maschino, B. Fabbi, C. Fiori: "Industrial Applications of Refractories Containing Magnesium Aluminate Spinel," *Ind. Ceram. (Faenza Italy)* **8** (1988) no. 3, 121 – 126.
83. A. Nagasoe, S.-I. Isurumoto, A. Kitamura: "Refractory Characteristics of Spinels With Various MgO Content," *Taikabutsu Overseas* **11** (1991) no. 3, 29 – 36.
84. M. Olbrich, F. Dobrowsky: "Periklas-Spinell-Steine in der Zementindustrie," *Radex Rundsch.* 1990, no. 2/3, 300 – 311.
85. A. P. Starzacher: "Picrochromitsteine," *Radex Rundsch.* 1988, no. 1, 491 – 502.
86. M. Koltermann: "Olivine as a Refractory Material in the Steel Industry," *Taikabutsu Overseas* **8** (1988) no. 1, 3 – 8.
87. E. R. Begley, P. O. Herndon: "Zirconia – Alumina – Silica Refractories," in A. M. Alper (ed.): *High Temperature Oxides,* part 4, Academic Press, New York 1971, pp. 185 – 208.
88. P. Robyn, J. Moreau, G. Soumoy: "Selection of Zircon Bricks for Superstructure of Container Glass Furnaces by Laboratory Simulative Testing," *InterCeram* **35** (1986) special issue, 65 – 67.
89. T. S. Busby, G. B. Shaw: "Special Refractories; Properties and Application," *Glass Technol.* **30** (1989) no. 4, 117 – 119.
90. C. N. McGarry, D. L. Monroe, R. A. Webber: "New Thermal Shock Resistant Dense Zircon and Dense Chromic Oxide Refractories," *Ceram. Eng. Sci. Proc.* **12** (1991) no. 3/4, 473 – 481.
91. H. M. Zender, H. Leistner, H. R. Searle: "ZrO_2 Materials for Application in the Ceramic Industry," *Interceram,* special issue, Aachen Proceedings 1989, Verlag Schmid, 1990, pp. 39 – 42.
92. H. Leistner, D. Ratcliffe, A. Schuler: "Improved Material and Design Devices for Continuous Casting Components," in [15] pp. 316 – 319.
93. K. T. Piel et al.: "Improved Carbon Blocks for Blast Furnaces," *InterCeram,* special issue, Aachen Proceedings 1988, Verlag Schmid, Freiburg 1989, pp. 22 – 25.
94. "Crucible Manufacture in West Germany," *Foundry Trade J. Int.* **8** (1985) no. 28, 183, 185, 187.
95. T. M. Paterson: "How to Select a Crucible and Keep it in Production," *Foundry Trade J.* **160** (1986) no. 3322, 198 – 200.
96. G. Routschka: "Feuerfeste Werkstoffe für den Roheisentransport und die Roheisenvorbehandlung im Spiegel der Literatur," *Sprechsaal* **121** (1988) no. 10, 956 – 966.
97. J. Poirier, F. Etienne, M. Beurotte: "Role of Raw Materials and Physico-Chemical

Properties of Magnesia – Carbon Bricks in the Wear Mechanism of LD Converter Refractory Linings," in *Unitecr '89 Proc.,* vol. 1, American Ceramic Society, Westerville 1989, pp. 326 – 347.
98. W. Zednicek: "Kohlenstoffhaltige basische feuerfeste Steine – eine neue Generation in der Qualitätspalette der Feuerfestindustrie," *Silikattechnik* **39** (1988) no. 11, 385 – 390.
99. G. Zoglmeyer, D. Romai: "Updated Dolomite Brick for Secondary Steel Metallurgy," *InterCeram* **40** (1991) no. 2, 81 – 89.
100. C. F. Cooper: "The Characterisation and Evaluation of Graphitic Continuous Casting Refractories," *InterCeram,* special issue, Aachen Proceedings 1986, Verlag Schmid, Freiburg 1987, pp. 79 – 84.
101. I. Peretz, B. Brezny: "Magnesia – Graphite Refractories for BOF Converters," *Am. Ceram. Soc. Bull.* **71** (1992) no. 9, 1383 – 1390.
102. R. Rathner, J. Knauder, H. Weissensteiner: "Thermomechanical Behaviour of Refractory Linings with Special Emphasis on BOF-Vessels," in [15] pp. 12 –18.
103. A. Yamaguchi, H. Tanaka: "Role and Behaviour of Non-Oxide Compounds Added to Carbon-Containing Refractories," in [15] pp. 32 – 38.
104. E. H. P. Wecht: *Feuerfest-Siliciumcarbid,* Springer Verlag, Wien 1977.
105. A. Fickel, W. Völkel: "Modern Refractory Materials for Blast Furnace Shafts," *InterCeram,* Aachen Proceedings 1988, pp. 38 – 40.
106. W. Völker, A. Fickel: "Suspension Elements Based on Silicon Carbide as Kiln Furniture of the Ceramic Industry," *InterCeram,* Aachen Proceedings 1989, pp. 69 – 71.
107. R. Röttenbacher, P. Pögl: "Silicon Carbide Kiln Furniture," *InterCeram,* Aachen Proceedings 1989, pp. 72 – 74.
108. K. Asano, A. Ishii, Y. Tsutsui: "Reactivity Between BN Composite Ceramics and Molten Stainless Steel," *Taikabutsu Overseas* **11** (1991) no. 3, 3 – 11.
109. "New Refractories for the Glass and Steelmaking Industry Based on Advanced Ceramics," *Ind. Ceram. (Faenza Italy)* **10**(1990) no. 4, 169.
110. A. Krings et al.: "Schmelzgegossene Steine für den Glasofenbau – Eigenschaften und Anwendung," *Fachausschußbericht no. 69,* Dt. Glastechn. Ges., Frankfurt/M. 1973.
111. A. D. Davis, T. M. Wehrenberg: "High Zirconia Glass Refractories: An Overview," *Ceram. Eng. Sci. Proc.* **9** (1988) no. 3/4, 273 – 283.
112. A. Zanoli, G. DuVierre, E. Sertain: "High Zirconia Fused-Cast Refractories," *Ceram. Eng. Sci. Proc.* **12** (1991) no. 3/4, 496 – 517.
113. E. Gugel: "Feuerbeständiger Beton mit niedriger Wärmedehnung," *Sprechsaal* **100** (1967) no. 21, 825 – 830.
114. "Aluminiumtitanat – Keramik Diceron AT," *Sprechsaal* **120** (1987) no. 4, 334 – 335.
115. H. Rasch: "Cordieritkeramik," in [34], chap.4.1.3.0, 1992, pp. 55.
116. O. Thelen: "Ungeformte feuerfeste Baustoffe – von der Stampfmasse bis zum Hochleistungsprodukt," *Keram. Z.* **44** (1992) no. 8, 501 – 507.
117. Shinagawa Technical Report, special issue on Monolithic Refractories, no. 32 (1989).
118. "Ungeformte feuerfeste Erzeugnisse für die Herstellung von Eisen und Stahl." *XXXV. Int. Feuerfest-Kolloquium Aachen,* Stahl und Eisen Special, Verlag Stahleisen, Düsseldorf 1992.
119. Stahl-Eisen-Werkstoffblatt SEW 916: *Kennzeichnung ungeformter feuerfester und feuerbeständiger Erzeugnisse,* Verlag Stahleisen, Düsseldorf 1984.
120. G. Routschka, A. Majdic: "Feuerfeste Erzeugnisse für das Stichloch und das Rinnensystem von Hoch-öfen im Spiegel der Literatur," *Keram. Z.* **34** (1982) no. 9, 468 – 477.
121. R. Krebs: "Feuerleichtbetone im Ofenbau,"*Metall (Berlin)* **33** (1979) no. 11, 1181 – 1185.
122. R. W. Rothrock, L. P. Krietz: "New Development in High Strength, Abrasion Resistant Castables," in P. E. Fisher (ed.): *Ceramic Transactions,* vol. 4, The American Ceramic Society, Inc., Westerville 1990, pp. 272 – 293.
123. F. C. Brachet, R. Avis, B. Clavaud: "Self-Flowing Refractory Castables," in [15], pp. 264 – 266.
124. K. Tabata, T. Yamamura, M. Katayama, K. Enoki: "Development of Mixerless Casting Material for Blast Furnace Troughs," in [15], pp. 46 – 53.
125. J. P. Kiehl, V. Jost, B. Clavaud, FR 76 22 344, 1976.
126. J. P. Moreau: "Development of Low Cement and Ultralow Cement Castables in Combustion Chambers and Boilers of Domestic Waste Incinerators," in [15], pp. 8 – 12.

127. W. Fletcher: "Refractory Castables in the Channel Furnace," *Steel Metals Mag.* **27** (1989) no. 5, 389–390.
128. A. Hey: "Precast Monolithics Refractory Shapes for the Iron and Steel Industry," *InterCeram* **39** (1990) no. 1, 29–32.
129. H. E. Schwiete, K. Konopicky: *Klassifikation und Eigenschaften wärmedämmender Steine,* **4. Beiheft,** Tonindustrie-Zeitung, Hermann Hübner Verlag, Goßlar 1967.
130. W. Schulle, E. Schlegel: "Probleme des inneren Wärmetransports sowie der gezielten Entwicklung von feuerfesten Wärmedämmstoffen," *Keram. Z.* **41** (1989) no. 3, 163–169.
131. J. Lenz: "Feuerfestzustellung für Industrieöfen unter dem Einsatz reagierender Ofenatmosphären," *Fachber. Hüttenprax. Metallweiterverarb.* **21** (1983) no. 5, 312–317.
132. R. D. Smith: "Refractory Ceramic Fiber, in [10], pp. 385–391.
133. T. F. Cooke: "Inorganic Fibers. A Literature Review," *J. Am. Ceram. Soc.* **74** (1991) no. 12, 2959–2978.
134. H. P. Zander: "Keramische Faserstoffe. Eigenschaften, Konstruktionen, Risiken und Grenzen," *Sprechsaal* **122** (1989) no. 2, 126–130;no. 4, 347–350; no. 6, 574–577.
135. P. Dietrichs, W. Krönert: "Eigenschaften, Hochtemperaturverhalten und Einsatzbedingungen keramischer Fasern," *Gas Wärme Int.* **30** (1981) no. 7/8, 338–350.
136. Y. Ohtsuki, Y. Nakahara, N. Sugiyama: "New Mullite Fiber and its Application in the Steel Industry," in [15], pp. 557–566.
137. M. E. Schupp: "A Global View of the Refractory Ceramic Fiber Market," *Am. Ceram. Soc. Bull.* **69** (1990) no. 1, 63–64.
138. "Refractories," in [13], pp. 895–917.
139. L. P. Krietz: "Monolithic and Fibrous Refractories," in [13], pp. 910–917.
140. R. Amavis (ed.): *Refractories for the Steel Industry,* Proc. Elsevier Applied Science, London 1990.
141. "Refractories in the Manufacture and Transport of Pig Iron," *InterCeram,* special issue, Aachen Proceedings, Verlag Schmid, Freiburg 1989.
142. H. Barthel, G. Buchebner, E. Kaltner: "The Magnesia-Based Lining of Oxygen Steel Converters to Conform With Changed Requirements," in *InterCeram,* special issue, Aachen Proceedings '84, pp. 7–10.
143. H. Shikano: "Recent Development and Present Situation of Refining Refractories," *Taikabutsu Overseas* **10** (1990) no. 3, 168–178.
144. "Feuerfeste Werkstoffe in der Sekundärmetallurgie," *33. Int. Feuerfest-Kolloquium,* Aachen 1990.
145. "Refractories for Continuous Casting," *InterCeram,* special issue, Aachen Proceedings '86, Verlag Schmid, Freiburg 1987.
146. G. Routschka, A. Majdic: "Feuerfeste Baustoffe für die Nichteisenmetallindustrie im Spiegel der Literatur," *Radex Rundschau* 1980, no. 4, 263–301.
147. G. Routschka, A. Majdic: "Feuerfeste Erzeugnisse für die Zementindustrie im Spiegel der Literatur," *ZKG Zem. Kalk Gips* **36** (1983) no. 9, 475–485.
148. K. Tabata, Y. Yasuda: "Technical Trends in Refractories for the Cement Industry in Japan," *Shinagawa Giho. Techn. Rep.* **34** (1991) no. 34, 23–42.
149. G. Routschka: "Feuerfeste Erzeugnisse für Glasschmelzwannenöfen im Spiegel der Literatur," *Keram. Z.* **37** (1985) no. 9, 457–466.
150. "Refractories in the Ceramics Industry," Aachen Proceedings 1989, Verlag Schmid, Freiburg 1990.
151. "Refractories in the Chemical Industry and in Waste Incineration," *InterCeram,* special issue, Aachen Proceedings '87, Verlag Schmid, Freiburg 1988.
152. M. S. Crowley, R. E. Fisher: "Petroleum and Petrochemical Applications for Refractories," in [10], pp. 471–488.
153. C. E. Zimmer: "Corrosion-Resistant Linings Used in the Petrochemical Industry," *Am. Ceram. Soc. Bull.* **70** (1991) no. 3, 410–414.
154. P. Dietrichs: "Keramische Auskleidungen in der thermischen Entsorgung," *CFI Ceram. Forum Int.* **68** (1991) no. 3, 92–96; no. 7, 346–350.
155. H. Stein: "Betriebserfahrungen mit feuerfesten Produkten in Müll- und Sonderabfall-Verbrennungsanlagen," *Fortschrittsber. DKG* **7** (1992) no. 2, 259–276.
156. VDI 2585, Entwurf Juli 1991. Emissionsmindung. Keramische Industrie.
157. *Technische Anleitung zur Reinhaltung der Luft,* Heider-Verlag, Bergisch Gladbach 1986.
158. *MAK,* VCH Verlagsgesellschaft, Weinheim 1992.
159. Health and Safety Executive EM 40/92. Occupational Exposure Limits 1992, London HMSO.

160. D. J. Bray: "Toxicity of Chromium Compounds Formed in Refractories," *Am. Ceram. Soc. Bull.* **64** (1985) no. 7, 1012 – 1016.
161. "Environmental Aspects of Basic Refractories," *World Ceramics & Refractories* **2** (1991) no. 3/4, 18, 20.
162. F. W. Löffler, *VDI-Ber.* **853** (1990) 351 – 370.
163. R. S. Redmond: "The Changing Face of Refractories," *CFI, Ceram. Forum Int.* **68** (1991) no. 10/11, 546 – 548.
164. A. J. Owen: "Developments and Trends in the European Refractory Industry," *CFI, Ceram. Forum Int.* **68** (1991) no. 7/8, 351 – 356.
165. K. Harries-Rees: "Refractory Majors. Plagued by Overcapacity," *Ind. Miner.* (London) (1992) Sept., 23 – 43.
166. H. Reh: "Ceramic Production in The Western World," *InterCeram* **37** (1988) no. 6, 20 – 23.
167. P. Bartha: "Betrachtungen zur Feuerfest-Werkstoff-Technik," *Zem. Kalk Gips Ed. B* **42** (1989) no. 10, 526 – 531.
168. G. Routschka: "Refractories History – 1950 to 1990," *InterCeram* **41** (1992) no. 4, 239 – 243.
169. M. Koltermann: "Feuerfeste Baustoffe 1980 bis 1990 – Rückblick und Prognosen," *Stahl Eisen* **111** (1991) no. 8, 41 – 48.
170. C. E. Semler: "Overview of Refractory Problems in Industry," *InterCeram* **40** (1991) no. 7, 534 – 539.

Refractory Coatings → Ceramics, Ceramic-Metal Systems
Refractory Fibers → Fibers, 5.Synthetic Inorganic

Reinforced Plastics

Composite Materials is a separate keyword.

GOTTFRIED W. EHRENSTEIN, Universität Erlangen-Nürnberg, Erlangen-Tennenlohe, Federal Republic of Germany

JOSEF KABELKA, Universität Erlangen-Nürnberg, Erlangen-Tennenlohe, Federal Republic of Germany

1. Introduction, Classification 1
1.1. Fiber-Reinforced Polymers 2
1.2. Particulate Polymer Composites . . 4
1.3. Laminar Polymer Composites 5
2. Reinforced Plastic Components . . . 6
2.1. Reinforcing Fibers 6
2.1.1. Glass Fibers 6
2.1.2. Graphite (Carbon) Fibers 7
2.1.3. Aramid Fibers 7
2.1.4. Other Continuous Fibers 8
2.1.5. Other Discontinuous Fibers 8
2.2. Fillers 9
2.3. Polymeric Matrices 9
3. Fabrication of Polymer Composites 9
3.1. Hand Lay-up Technique 10
3.2. Spray-up 11
3.3. Filament Winding 11
3.4. Pultrusion 12
3.5. Bag Molding 12
3.6. Sheet Molding Compounds (SMC) 12
3.7. Thick Molding Compounds (TMC) 13
3.8. Bulk (or Dough) Molding Compounds (BMC, DMC) 13
3.9. Prepregs 14
3.10. Reinforced Reaction Injection Molding (RRIM) 14
3.11. Resin Transfer Molding (RTM) . . 14
4. Physical and Chemical Properties . 14
5. Quality Specifications 16
6. Economic Aspects 17
7. References 17

1. Introduction, Classification

Reinforced plastics, also called polymer composites, are now an important class of engineering materials. They are distinguished by attractive mechanical properties and corrosion resistance, unique flexibility in design capabilities, and ease of fabrication. The word "composite" means "consisting of two or more distinct parts." Polymer composites consist of one or more discontinuous phases embedded in a continuous-phase polymer matrix.

The discontinuous phase is usually harder and stronger than the continuous phase, and is called the reinforcement. The matrix can be classified as thermoplastic (capable of being separately hardened and softened by increases and decreases, respectively, in temperature) or thermoset (changing into a substantially infusible and insoluble material when cured by the application of heat, or through chemical means).

The properties of composites are strongly influenced by the properties of their constituents and the distribution and interactions among them. The constituents usually interact in a synergistic way, providing properties that are not accounted for by a simple volume-fraction sum of the components. Along with the volume fraction and the distribution of discrete units in the discontinuous phase, the interfacial area plays an important role in determining the extent of interaction between the reinforcement and the matrix and – in this way – the final properties of the composite. The fabrication technology of composites, as well as some of their physical properties, is dominated by the chemistry and rheology of the matrix resin and by the type and physical form of the reinforcement.

The use of reinforcing agents makes it possible for any thermoset – or thermoplastic – matrix property to be improved or changed to meet varying requirements. Most polymer composites developed thus far have been fabricated to improve mechanical properties such as strength, stiffness, or toughness. The strengthening efficiency of the discontinuous phase plays the most important role in these products, and the strengthening mechanism depends strongly on the geometry of the reinforcements. Therefore,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_049

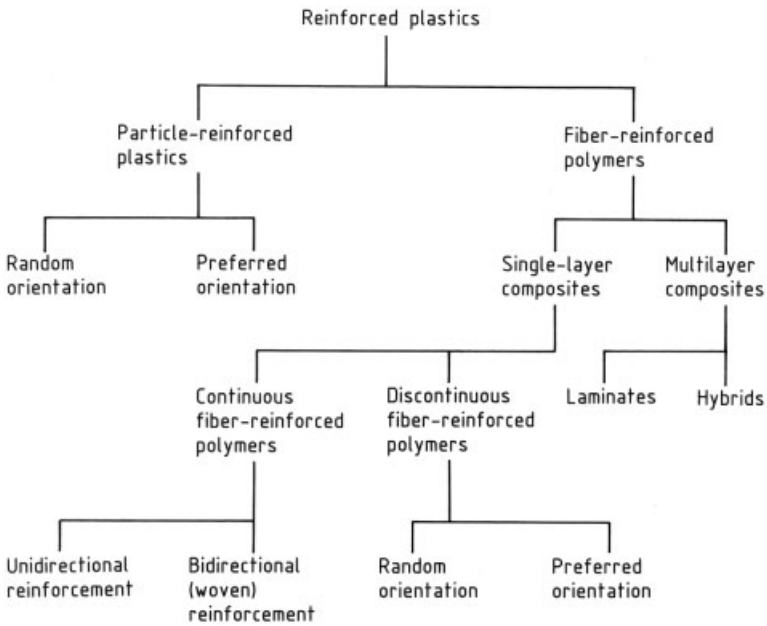

Figure 1. Classification of polymer composites [9]
The single-layer composites include materials composed of identical layers having the same orientation.

polymer composites can be classified according to reinforcement geometry [9]. The three major classes of polymer composites are fibrous, laminar, and particulate. The commonly accepted classification scheme for polymer composites is presented in Figure 1.

1.1. Fiber-Reinforced Polymers

Fibrous composites consist of reinforcing fibers embedded in a matrix. All fibers are characterized by a high aspect (length-to-diameter) ratio. This results in a high surface – volume ratio, permitting good anchorage of fibers within a matrix. Fiber reinforcement can be either continuous or discontinuous. Fibers are generally much stronger and stiffer than the surrounding polymer matrix, but they are small in diameter, and bend easily when pushed axially. Although such composites may have outstanding tensile strength, support is essential to prevent individual fibers from bending and buckling. Thus, the principal purpose of the matrix is to function not as a load-bearing constituent but essentially to bind the fibers together and protect them. Fiber-strengthening efficiency is highest when the direction of loading corresponds to the orientation of the fibers. In the transverse direction, properties of the matrix and the quality of the fiber – matrix interface predominate. The combination of matrix and fibrous elements results in a composite structure that may assume different structural forms depending on the forming and processing methods used.

Because the fibers are oriented in a particular direction, material properties may vary with direction. Such a material is referred to as "anisotropic." Many types of anisotropic materials exist, with special designations such as orthotropic, transversely isotropic, and cubic. The fundamental building block for fiber-reinforced plastics is an element with a unidirectional (UD) but transversely random array of fibers embedded in a surrounding matrix, as illustrated in Figure 2. Thus, a unidirectional composite is transversely isotropic (axes O_T and O_P can be selected arbitrarily). In the $O_L - O_T$ plane, however, the properties of a UD lamina are strongly dependent on direction (Fig. 3). The longitudinal properties of UD composites are controlled by the fiber, whereas transverse properties are matrix-dominated. The transverse properties are unsatisfactory in most engineering applications, a problem which is overcome by bonding together two or more laminae to act as an integral

structural element displaying the desired properties in all directions. Combining layers with different properties or different orientations gives the designer the ability to tailor a laminate (provide the optimal configuration) for a particular set of loading conditions. However, in applications where the stress may not be predictable, or where stresses are likely to be approximately equal in all directions, the composites themselves must have approximately equal strength in all directions.

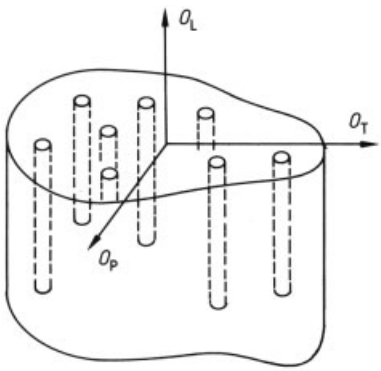

Figure 2. Basic element of fiber-reinforced plastics; coordinate system
O_L = longitudinal; O_T = transverse; O_P = perpendicular

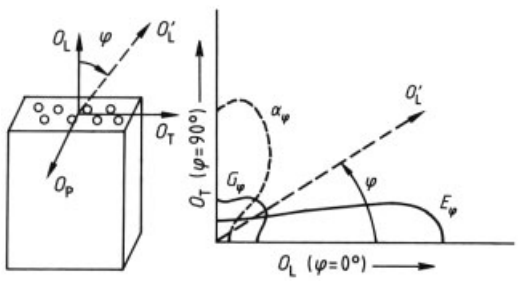

Figure 3. The directional nature of UD lamina properties (E, G, α – Young's modulus, shear modulus, and coefficient of thermal expansion)

An effective way of producing such an isotropic composite material is to use randomly oriented short fibers as reinforcement. Molding compounds consisting of short fibers are usually capable of producing isotropic composites. In short-fiber composites, the so called fiber-end effect plays an important role, and composite properties become a function of fiber length. Loads are not applied directly to the fibers in a composite, but are transferred instead by the matrix to the fiber surface through nonuniform shear stress acting along the fiber – matrix interface (Fig. 4). The tensile stress in the fiber reaches a maximum at the fiber midpoint. The *critical length* of a fiber can be defined as the minimum length at which the ultimate strength of that fiber is reached. Interfacial strength is influenced strongly by the effectiveness of bonding between the fiber and the matrix via coupling agents. The higher the interfacial strength, the shorter is the length that will ensure full exploitation of the fiber's strength.

In the case of short-fiber injection-molded composites, some fibers will have aspect ratios lower than the critical value. These fibers cannot contribute effectively to overall composite strength. Compounding and molding cause fibers to undergo attrition, resulting in a broad distribution in length and a significant fraction of fibers falling below the critical length. In long-fiber injection-molded composites, the mean fiber length is typically 10 times as great as in short-fiber composites. Thus, a higher proportion of the fibers remains effective in load-bearing situations, and the composites themselves are stronger than their short-fiber counterparts. Figure 4 shows that a small region adjoining the fiber end is subject to less than the maximum fiber stress. This in turn affects the elastic modulus. When the fiber length is much greater than the critical length, composite behavior approaches that of a continous-fiber-reinforced polymer. The tensile strength of a composite is more strongly dependent on interfacial bond strength than is its tensile modulus. This is evident in particulate fillers, where the aspect ratio approaches 1. Here the addition of filler contributes to an increase in modulus but generally to a decrease in the strength of the composite.

In general, the incorporation of reinforcing fibers into a polymeric matrix increases short-term stiffness and strength, increases creep resistance and fatigue endurance, and reduces the influence of temperature on all of these properties. Thermal expansion can be reduced as a way of improving dimensional stability.

Impact strength, which is related to the toughness of a material, is reduced in some polymer composites, a circumstance that can be influenced by several factors. The loading of

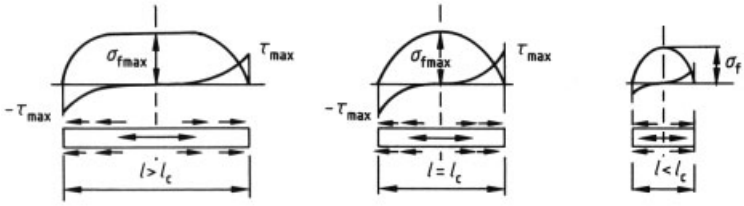

Figure 4. Distribution of tensile stress σ and shear stress τ in short fibers aligned with the load (l = length, l_c = critical fiber length)

fibers may transform the failure mode from ductile to brittle, which can in turn lower impact strength. However, this reduction can be counteracted, at least to some extent, if interfacial adhesion between the polymer matrix and the reinforcement permits the absorption of impact energy through debonding. Nevertheless, in systems where matrix-governed ductile failure occurs, increased interfacial adhesion generally improves impact strength.

1.2. Particulate Polymer Composites

By definition, a particle is nonfibrous; it generally lacks a long dimension, and is not effective in improving fracture resistance when added to a polymer matrix (with the exception of rubberlike particles in brittle polymer matrices, which contribute to an improvement in fracture resistance by promoting and then arresting crazing). Despite this definition, polymers containing very short fibers are generally classified as particulate composites. The introduction of hard particles into a polymer matrix leads to a reduction in strength due to stress concentration in the adjacent matrix material.

The two basic reasons for adding particulate filler or reinforcement to a polymer are (1) to fill the system (i.e., replace some of the more expensive polymer with less expensive filler) and (2) to modify properties of the matrix material (e.g., thermal or electrical conductivity), improve processibility, reduce friction or shrinkage, improve performance at elevated temperature, increase wear or abrasion resistance, improve fire retardancy, provide color, and improve appearance.

Platelet-shaped particles, such as flakes, can also serve as reinforcing agents. They offer two-dimensional geometry and, when placed parallel to each other, can be packed more closely than fibers or spherical particles. However, such possibilities have not been fully exploited due to fabrication difficulties. In terms of reinforcement, the longer the particles, the better is the reinforcement. Most particle-filled polymers are manufactured by injection molding, however, so a balance must be struck in terms of a particle length that gives both optimum properties and good processibility. If relatively long fiberlike particles are used as filling, their distribution and orientation may change within the molding. The flow of fiber-filled plastics into a mold during injection molding is a very complex process, and ensuring flow-induced orientation in the optimum direction is difficult. On the other hand, sometimes the challenge is to minimize undesirable effects of anisotropy (e.g., warping and postmolding shrinkage), which may be accentuated by the flow-orientation of fibers. In these cases particulate fillers may be preferable to fibers.

1.3. Laminar Polymer Composites

Laminar composites consist of at least two different layers (laminae) bonded together into one composite piece. Such layers may differ in material, form, or orientation. Combinations of matrix, fibrous mat, aligned fibers, cloth, paper, and other materials can be laminated together in either continuous or discontinuous processing.

Because of the nearly unlimited variety of possible combinations of layers, material forms, and orientations, generalization regarding laminar design is difficult. This great variety offers the designer considerable freedom in both laminate composition and product form. The unique properties obtained with layered construction encourage the designer to develop composite materials that meet new and innovative design parameters.

The principal design criterion is appropriate fiber orientation within each layer. Polymeric composite structures that incorporate continuously oriented fibers laid in plies can be radically anisotropic in nature. Strength, stiffness, and coefficients of thermal or moisture expansion can all vary more then tenfold in different directions (cf. Fig. 5). Whereas the in-plane strength of such materials may be comparable to the strength of metals, for example, strength in the through-thickness direction can be lower than that of pure matrix. Differing properties of the different layers (e.g., thermal expansion, stiffness) may lead to considerable residual stress when a polymer matrix is cured at elevated temperature. Interlaminar stresses induced through external loading may also be significant, and can cause distortion when the layer-stacking sequence is not symmetric and situated symmetrically to the midplane (Fig. 6).

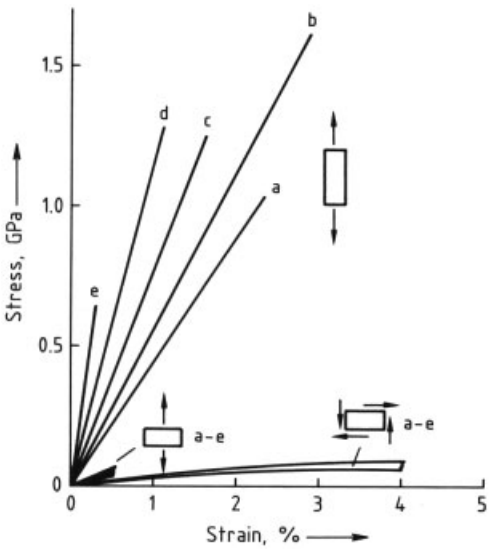

Figure 5. Stress – strain relationships for UD composites under loading in different directions
Matrix = epoxy resin; fiber volume fraction = 0.6
Fiber reinforcement: a) E-glass; b) R-glass; c) Aramid; d) HS graphite; e) UHM graphite

The theory of laminates has been elaborated [9–17], and many computer programs are now available to help determine the optimal configuration of a laminate under specific conditions; [e.g., LAMINATE (Engineering Software, Dallas, Texas), COMPCAL and CYCLAN (American Society for Composites, Dayton, Ohio), MIC-MAC (Think Composites, Dayton, Ohio), and CADFIBER (IKV, Aachen, Germany)]. In addition to these general programs, many special software packages are being developed for special tasks (e.g., computation of fatigue and crack propagation).

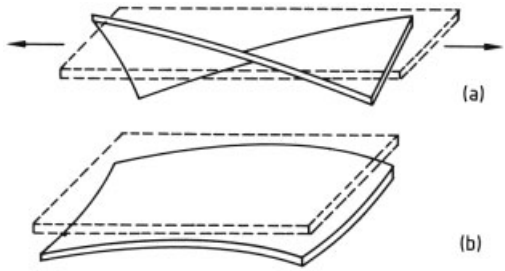

Figure 6. Distortion of an unsymmetrical laminate due to (a) load and (b) temperature change

2. Reinforced Plastic Components

2.1. Reinforcing Fibers (→ Fibers, 4. Synthetic Organic; Fibers, 5. Synthetic Inorganic)

As a rule, fibers used for reinforcement purposes are rather fine, ranging from 2 to 20 μm in diameter. The main exception is boron fiber, available in diameters up to 200 μm. Fibers are typically stronger and stiffer than a corresponding bulk material. Fiber strength generally decreases with increasing diameter and length (Fig. 7). A high fiber aspect ratio permits an effective transfer of load to the fibers via the matrix, which makes such fibers an excellent reinforcing agent.

The major categories of synthetic reinforcing fibers are glass, graphite (carbon), boron, organic, metallic, and ceramic. Natural fibers (asbestos, jute, and sisal), formerly used for economic reasons, have now been generally supplanted by synthetics (in the case of asbestos, because of adverse health effects).

Synthetic fibers are usually prepared in continuous processes and then converted into a form suitable for their intended composite application. The major finished forms are continuous roving, woven roving, fiber mat, chopped strand, and yarn for textile applications.

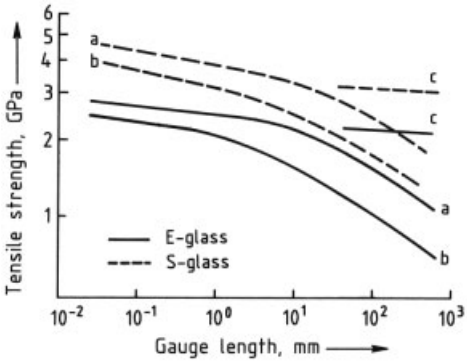

Figure 7. Dependence of fiber strength on gage length
a) Mean filament strength; b) Strength of fiber roving; c) Strength of fiber roving saturated with polyester resin

The form of reinforcement varies with the fabrication process. Three categories can be distinguished: In the first, fibers are placed directly on the component being manufactured. Examples are the filament winding (Section 3.3) and spray-up contact (Section 3.2) molding processes. The second group includes those processes in which a sheet, web, or preform of fibers is prepared and placed in position on a mold where it comes in contact with a resin (e.g., resin-transfer molding, Section 3.11). The third category consists of processes involving the preparation of a compound of fiber and plastic that can flow during molding. Sheet molding compounds (Section 3.6), bulk molding compounds (Section 3.8), and short-fiber-reinforced thermosets and thermoplastics are typical examples within this category.

2.1.1. Glass Fibers (→ Composite Materials; Fibers, 5. Synthetic Inorganic, Chap. 2.)

Glass fibers constitute the most widely used form of synthetic composite reinforcement. Chemically, glass is an amorphous material consisting essentially of a silica network. Other chemical components are added to decrease its viscosity to levels suitable for melting and fiber formation. Both the physical and the chemical properties of the resultant glass are altered to varying degrees by the type and amount of modifiers.

The four main glass compositions used to create continuous fibers for reinforcement are high-alkali glass (A-glass), low-alkali or electrical-grade glass (E-glass), chemically resistant glass (ECR-glass), and high-strength glass (S-glass). Properties of glass fibers are listed in Table 1.

Table 1. Properties of glass fibers

Property	A-glass	ECR-glass	E-glass	S-glass
Density, kg/m^3	2500	2490	2540	2480
Tensile strength, GPa	3.0	3.0	3.5	4.6
Elasticity modulus, GPa		69	72	85
Coefficient of thermal expansion, 10^{-6} K^{-1}	8.6	7.2	5.0	5.6
Tensile elongation, %		2.5	2.4	2.9

Glass fibers are supplied in several basic forms, including rovings, chopped strands, mats, yarns, woven fabrics, woven rovings, nonwoven fabrics, tapes, and other special configurations.

The principal advantages of glass fibers are low cost and high strength. The disadvantages are a low modulus, poor abrasion resistance, and poor adhesion to the polymer matrix. Fiber *sizing* (containing a coupling agent) is used to improve fiber – matrix adhesion; its basic function is to form a chemical link between the glass surface and the matrix. Typically, a coupling agent consists of a silane that is adsorbed onto the glass and is capable of reactions with functional groups in the resin (→ Fibers, 5. Synthetic Inorganic, Chap. 1.5.3.). The precise mechanism of action is not known, but silane sizing performs a useful role in glass-reinforced systems involving unsaturated polyester, epoxy, vinyl ester, or thermoplastic matrices.

Glass fibers generally have good chemical resistance, are noncombustible, and do not absorb water. Substantial reduction of thermal dilatation can be achieved by the addition of glass fibers because their thermal expansion coefficient is an order of magnitude lower than that of most plastics. The useful temperature range is quite large: glass does not soften significantly until a temperature $> 400\,°C$ is reached. Glass fibers are isotropic.

2.1.2. Graphite (Carbon) Fibers (→ Composite Materials, Chap. 4.1.1.; → Fibers, 5. Synthetic Inorganic, Chap. 5.)

Graphite and carbon fibers are generally prepared by the thermal decomposition of three

fibrous organic precursors: polyacrylonitrile (PAN), rayon (cellulose), or pitch. They are available in tows containing typically 2000–10 000 untwisted individual filaments. Fiber cross sections may be circular, dog bone, or irregular, with a diameter of ca. 6–10 μm. The primary advantages of graphite over glass fibers are higher modulus, improved creep-rupture resistance, better fatigue endurance, and lower density. However, they are also characterized by a lower limit of elongation, which results in a relatively low fracture energy. Therefore, impact resistance for graphite-fiber composites is lower than for glass-fiber composites. The high strength and stiffness of graphite fibers result from a high degree of structural orientation. They are strongly anisotropic, with transverse tensile and shear moduli generally an order of magnitude lower than the corresponding axial modulus. Graphite fibers have a negative thermal expansion coefficient in the axial direction, which permits the construction of a multi-ply laminate with zero in-plane expansion. This is achieved, however, at the expense of large deformations in thickness.

Graphite fibers are usually subjected to a surface treatment which results in an increase in hydroxyl and amine groups on the fiber surface, thereby improving the adhesion between fiber and matrix and the laminate shear performance. To improve handling, a light sizing is often applied, consisting of a resin that is compatible with the designated matrix. Graphite fibers are available in various forms: continuous, chopped, woven fabric, or mat. Hybrid fabrics (made of a combination of graphite and glass or aramid fibers) are also available. The graphite fibers applicable as reinforcement are offered in high-strength, high-modulus, and ultrahigh-modulus forms. Representative properties are listed in Table 2. Other grades with intermediate properties have also been produced, and production processes continue to be improved.

2.1.3. Aramid Fibers (→ Composite Materials; → High-Performance Fibers, Chap. 3.2.1.)

Aramid (aromatic polyamide) fibers have been used in advanced composites since the early 1970s. The most widely known are Kevlar fibers. Among commercially available aramid fibers, the two types of interest for reinforcing plastics are high-modulus (Kevlar 49) and, to a lesser extent, low-modulus (Kevlar 29) materials.

Aramid fibers have stiffness and strength characteristics intermediate between glass and graphite. Their bonding to polymeric matrices is rather poor. Plasma amination of the fiber surface seems to be the most promising method for improving this property. In addition to high tensile strength, aramid fibers offer excellent toughness, outstanding heat resistance, fairly good chemical resistance, exceptional wear resistance, good frictional properties at elevated temperature, excellent tensile–tensile fatigue resistance, and a negative axial coefficient of thermal expansion.

Aramid fibers are strongly anisotropic (Table 2). Like graphite, their transverse tensile and shear moduli are about an order of magnitude lower than the axial tensile modulus. The fibers are also weak in compression, a result of their fibril microstructure, leading to use predominantly in tension-loaded construction elements. Their very low density makes them ideal for applications requiring critical weight saving, such as components for aircraft, space vehicles, missiles, and sporting goods. However, an elevated water-absorption capacity results in low values for dielectric properties. Aramid fibers are available in a variety of forms, such as continuous-filament yarns, rovings, woven fabrics, discontinuous staple, and pulp.

2.1.4. Other Continuous Fibers

Almost any organic fiber can be incorporated into a composite. For years, natural fibers such as cotton, jute, hemp, and silk have been studied as reinforcements. They are inferior to synthetics in tensile strength and modulus, but exhibit significantly higher elongation, which results in better composite damage tolerance. Low cost makes natural fibers attractive for applications with low load requirements (e.g., housing construction). A major problem, however, is their susceptibility to fungal and insect attack and to degradation by moisture.

From the category of synthetic fibers, cellulosics, polyamides, polyvinyls, polyacrylonitrile, polyethylene, and polyesters have found

Table 2. Properties of graphite (C) and aramid fibers

Property	C, high strength	C, high modulus	C, ultrahigh modulus	Aramid
Density, kg/m^3	1800	1900	2000	1440
Tensile strength, GPa	2.5	1.8	1.0 – 1.3	2.3
Young's modulus, GPa				
Longitudinal E_L	230	390	520 – 620	124
Transverse E_T	14	8	6	5
Shear, G_{LT}	10	8	7	2.6
Poisson ratio [1]				
Longitudinal ν_{LT}	0.25	0.2	0.2	0.35
Transverse ν_{TT}	0.35	0.3	0.25	0.35
Coefficient of thermal expansion, $10^{-6}\,K^{-1}$				
Longitudinal	− 1.0	− 1.2	− 1.5	− 4.0
Transverse	10	12	15	52
Elongation, %	0.9	0.3	0.2	1.6

application in the preparation of composites because of their favorable contribution to impact strength, electrical properties, and chemical resistance.

2.1.5. Other Discontinuous Fibers

Some fibers, such as single-crystal whiskers and alumina and silicon carbide fibers (→ Composite Materials, Chap. 4.1.3.) lack a corresponding continuous form. They offer very high stiffness and strength, but their full potential for composites has not been realized for two reasons: (1) their original length approaches the critical length, and this low aspect ratio is further reduced during compounding and molding; (2) controlling their alignment in the product is extremely difficult.

2.2. Fillers

Fillers are used to alter certain properties of composites – including price, because they are generally much less expensive than the polymer they replace. Many fillers are available, and in a variety of shapes: fibrous, plate-like, spherical, and irregular. Because of their low aspect ratio, fillers do not improve the tensile strength of a plastic; in fact, strength is usually decreased, especially at higher filler content.

With the addition of most fillers, such as calcium carbonate, kaolin, glass (milled glass fibers, hollow and solid glass spheres, glass flakes), mica, talc, silica, wollastonite, and aluminum oxide, the composite modulus increases, cure shrinkage and thermal expansion decrease, and the heat generated during cure is reduced. Aluminum oxide trihydrate and antimony oxide act both as fillers and as fire retardants; mica and silica improve chemical resistance.

Electrically conductive fillers, such as metallic powders and carbon particles (carbon fibers, carbon black), are used to reduce the accumulation of static charge. Carbon black is used to improve the UV-radiation resistance of polyolefins. Some fillers, such as carbon black, chalk, and titanium dioxide, act as pigments. Addition of extremely fine silica particles affects the thixotrophy of liquid resins. Magnetic properties can be obtained by incorporating such magnetic mineral fillers as barium sulfate.

2.3. Polymeric Matrices (→ Composite Materials, Chap. 4.2.)

Certain physical and chemical properties of the matrix are particularly significant with respect to the properties of a composite. The main advantages conferred by polymer matrices are low cost, easy processibility, good corrosion resistance, good dielectric properties, and low density. Low strength, low stiffness, and (usually) low service temperatures, on the other hand, limit their applications. The primary functions of a matrix are to bind the reinforcing elements together, distribute the loading among them, and protect the reinforcing system; this in turn presupposes good adhesion between matrix and reinforcement. Stress-transfer capability is generally higher in thermosets than thermoplastics. However, thermosets are more difficult to

process, and they require longer molding cycles. They are also high molar mass materials, and therefore inherently relatively brittle. Their creep resistance, particularly at elevated temperature, is generally higher than that of thermoplastic matrices.

Traditionally, thermoplastics were reinforced primarily with discontinuous fibers, but high-temperature resins such as polyether ketone, polysulfone, and polyimide have made thermoplastics competitive with thermosets in the area of high-performance continuous-fiber composites as well.

The most important matrices generally are alkyds, allyl esters, epoxies, furans, melamines, phenolics, polybutadienes, polyesters, polyimides, polyurethanes, silicones, ureas, and vinyl esters. The most widely used, however, are polyesters, epoxies, and vinyl esters. For thermosets, the dominant resins are unsaturated polyesters, followed by acrylic resins, phenolic resins, epoxies, vinyl esters, and unsaturated polyester – urethane hybrids. The complete list of applicable thermoplastic matrices is more extensive, however, and includes acrylonitrile – butadiene – styrenes, acetals, acrylics, fluoropolymers, polyamides, polyamide – imide, polyarylsulfone, polycarbonates, thermoplastic polyesters, polyetherether ketone, polyethersulfones, polyethylenes, polyimides, poly(phenylene sulfides), polypropylenes, polystyrenes, polysulfones, poly(vinyl chloride), and styrene – acrylonitriles.

3. Fabrication of Polymer Composites (→ Composite Materials, Chap. 5.; → Plastics, Processing)

A typical feature of thermoset composites is that material formation and product fabrication occur simultaneously during final molding (e.g., hand lay-up, spray-up, vacuum-bag molding). In some cases, especially with thermoplastic matrix composites, compounding occurs first, and the shape is formed or molded in a second stage. However, in all cases, final composite properties are influenced significantly by processing conditions. Fabrication processes for thermosetting matrices can be classified broadly as wet forming processes and processes involving premixes or prepregs. The wet processes include hand lay-up, filament winding, pultrusion, bag molding, reinforced reaction injection molding (RRIM), and resin transfer molding (RTM). In processes using premixes, compounding is performed separately to yield bulk molding compounds (BMC), sheet molding compounds (SMC), and prepregs. The use of premixes facilitates automation.

Conventional molding and thermoforming methods, (e.g., injection molding, press molding, stamping) are used when processing thermoplastic matrix composites. However, the addition of fibers or filler particles to thermoplastics significantly alters their rheological properties and thermal conductivity, which can greatly affect molding conditions [2].

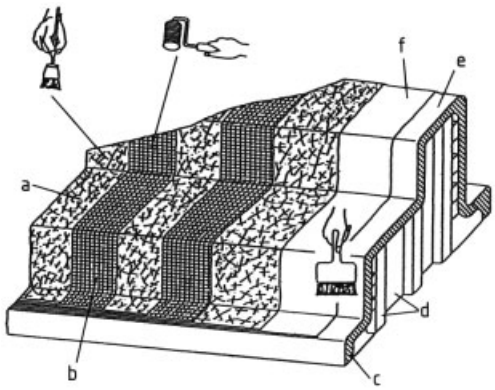

Figure 8. Basic cold- cure contact molding process
a) Strand mat; b) Fabric; c) Mold; d) Mold stiffeners; e) Mold release agent; f) Fine surfacing tissue

3.1. Hand Lay-up Technique

Hand lay-up is an open-mold process based on stacking the plies of reinforcing material in a special orientation and sequence, saturating them with resin, and brushing or rolling them to compact the material and remove entrapped air (Fig. 8). The hand lay-up techniqe is the simplest, most commonly used method for the manufacture of both small and large reinforced products. Typical applications include boats, radomes, pools, and tanks. This process is commonly employed with room-temperature-cure polyester resin and glass reinforcement, but

it can also be used with vinyl ester and suitable epoxy systems as well as with carbon or aramid fiber reinforcement. The reinforcement is usually in the form of chopped strand mat or woven roving. A gel coat, formed from a thixotropic pigmented resin, is widely used to produce a high surface finish.

Performance may be altered greatly by varying the resin-to-reinforcement ratio, the type of resin, the form and direction of reinforcement, and the additive content (Fig. 9).

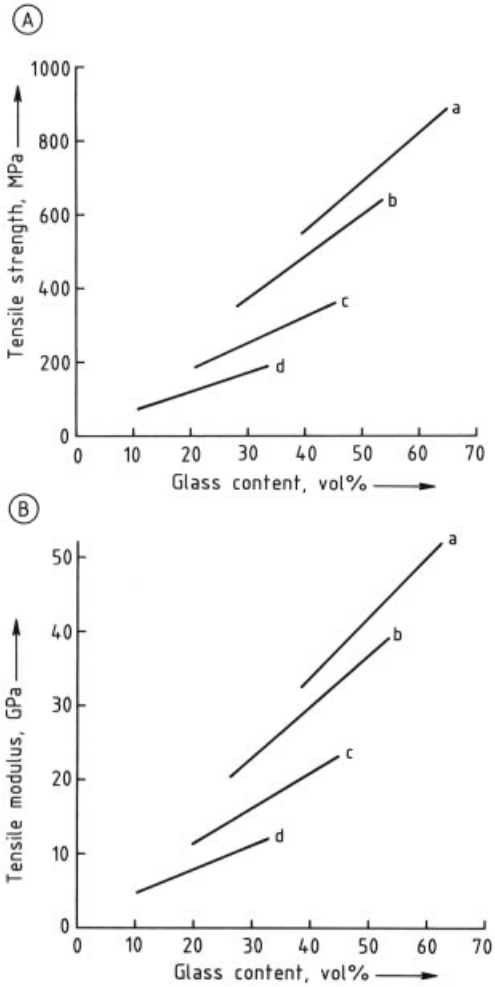

Figure 9. A) Tensile strength and B) stiffness of various glass – polyester laminates as a function of fiber content
a) UD laminate; b) UD fabric; c) Balanced fabric; d) Chopped strand mat

3.2. Spray-up

Spray-up is a partially automated form of hand lay-up. A spray nozzle is used to propel chopped fiber and liquid resin against the mold surface until the desired thickness is reached (Fig. 10). As in hand lay-up, the quality of the product depends heavily on the skill of the operator, but labor and raw material costs are lower than in hand lay-up operations. Spray-up operations can be automated, and are more practical for fabricating large parts (e.g., boat hulls) or creating protective surface coating.

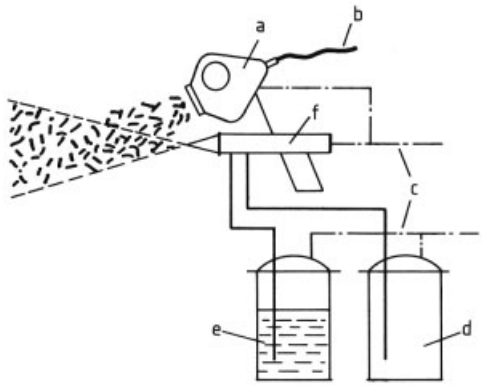

Figure 10. Schematic representation of spray-up process
a) Chopper; b) Fiber strand; c) Pressurized air; d) Resin and accelerator; e) Resin and catalyst; f) Spray nozzle

3.3. Filament Winding

Filament winding is an automated process in which continuous filament (or tape) passes through a resin bath and is then wound on a mandrel. The winding angle and placement of the reinforcement are controlled through specially designed machines in which filament feed is synchronized with mandrel rotation. The device has the capacity to vary the winding angle, filament tension, or resin content in each layer of reinforcement. The winding angle may vary from longitudinal through helical to circumferential or "hoop" (Fig. 11), which permits the design and fabrication of structural parts with a high degree of precision and very high strength. The two basic techniques for filament winding are

1) Wet winding, in which the reinforcement and matrix resin are applied during the winding stage
2) Dry winding, which uses preimpregnated reinforcement

Resin content in the laminate can be controlled more accurately with prepregs.

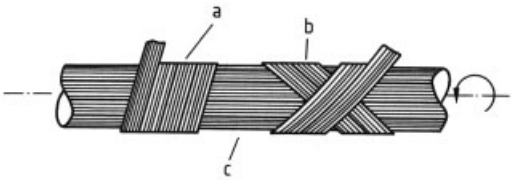

Figure 11. Alternative filament winding patterns
a) Circumferential or hoop; b) Helical; c) Axial

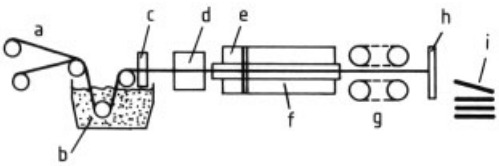

Figure 12. Schematic of the pultrusion process
a) Let-off; b) Impregnation bath; c) Resin control; d) Preformer die; e) Cold die section; f) Heated die section; g) Puller; h) Cut off; i) Product

Reinforcements are available in a number of forms: rovings, fabrics, tapes, mats, foils. Computer-aided design (CAD) methods are widely used to establish optimal laminate composition and fabrication conditions. Filament-wound applications include components subject to very high loads, such as pressure vessels, rocket engine cases, radomes, storage tanks, pipes, helicopter blades, and other aerospace parts.

3.4. Pultrusion

Pultrusion is a continuous processing method for obtaining composites with a constant cross section. The process consists of pulling continuous reinforcement (usually in the form of a roving or mat) through a resin bath (with removal of superfluous) into a shaping die, where the resin is subsequently cured (Fig. 12). A number of profiles, such as rods, tubes, and various structural shapes, with very high longitudinal strength and stiffness, can be produced via appropriate dies. The process is suitable for thermosetting resins (polyester, epoxy).

3.5. Bag Molding

In this process, a sheetlike reinforcement is laid up in a mold, into which the resin is introduced by spreading or coating. A flexible bag or diaphragm is used to apply pressure uniformly over one surface of the laminate. The general process can be subdivided into three basic molding methods: pressure bag, vacuum bag, and autoclave, the last of which is represented schematically in Figure 13. Heat may be applied in the form of steam in the autoclave itself or through the rigid mold. Large parts can be prepared with high-quality surface finishes. Mats, fabrics, preimpregnated forms of reinforcement, and honeycomb materials are used, and excellent adhesion to composite or sandwich core surfaces can be achieved. These features make bag molding attractive in the aerospace and high-technology areas.

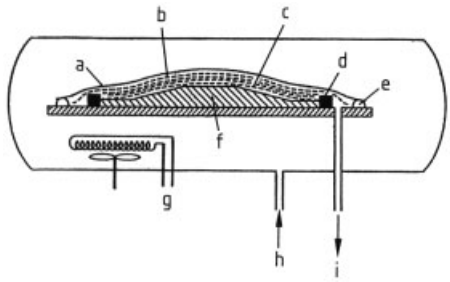

Figure 13. Schematic of an autoclave molding device
a) Bag; b) Bleeder pack; c) Prepreg pack; d) Flexible dam; e) Seal; f) Mold; g) Heater; h) Pressure; i) Vacuum

3.6. Sheet Molding Compounds (SMC)

Sheet molding compounds are produced as continuous flat sheets of ready-to-mold composite material containing fibers, usually together with mineral filler, dispersed in a thermosetting resin. Chopped E-glass fibers 20 – 60 mm long are most often used as reinforcement, although

carbon or aramid fibers are also applicable, either separately or as a hybrid. Both short (discontinuous) and continuous fibers are used in SMC. Short fibers, randomly oriented in the plane of the sheet, are common in applications requiring isotropic material properties. The resin is in an uncured but highly viscous state, achieved by addition of a thickener (MgO). Curing of the resin takes place during the compression-molding operation. With the application of heat and pressure, SMC first flows into and fills the mold cavity, and then cures into a solid. Volume change on cure can be controlled by the addition of various shrinkage compensators, permitting high dimensional stability to be achieved. A schematic of an SMC device is presented in Figure 14.

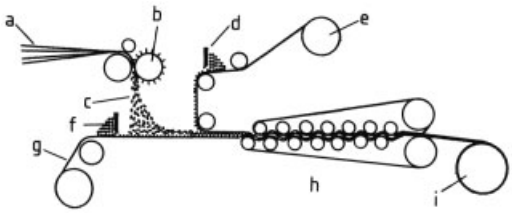

Figure 14. Processing equipment for sheet molding compounds
a) Continuous-strand roving; b) Chopper; c) Chopped roving; d) Doctor blade; e) Carrier film; f) Resin–filler paste; g) Plastic carrier film; h) Compaction section; i) Take-up roll

Several types of SMC are distinguished as a function of the form of reinforcement (Fig. 15):

SMC-R	randomly oriented fibers
SMC-O	oriented SMC
SMC-CR	combination of continuous, parallel, and random fiber orientation
XMC	criss- crossed continuous fibers in a sheet
HMC	SMC with a high fiber content (no filler)

A numeral at the end of the letter designation indicates the fiber content in weight percent. For example, SMC-C (HMC) 20 R 15 contains 20 wt % of continuous parallel and 15 wt % of randomly oriented fibers.

The performance of SMC composites depends primarily on a combined effect of the amount and length of the reinforcement as well as the orientation of the reinforcement in the finished part. Many properties can be influenced favorably by the addition of fillers or other additives. SMC composites can be used to produce structural parts with complex shapes in relatively short molding times. For this reason SMC composites have found broad application in mass production, especially in the transportation industry and appliance manufacture (e.g., washing machines, refrigerators, furniture, computer housings).

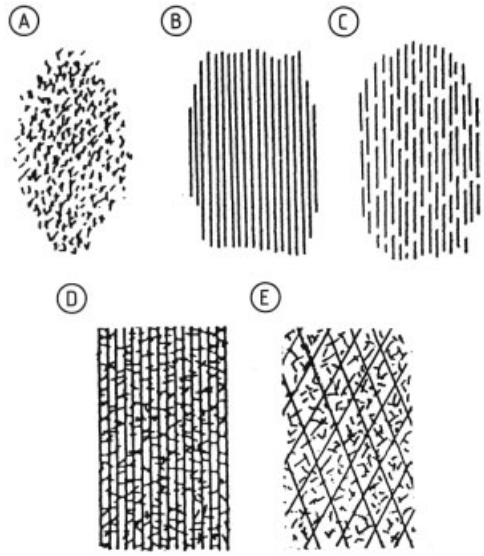

Figure 15. Fiber reinforcement structures of various sheet molding compounds
A) SMC-R; B) SMC-C (HMC); C) SMC-O; D) SMC-CR; E) XMC

3.7. Thick Molding Compounds (TMC)

Thick molding compounds differ from sheet molding compounds only in thickness. This may range up to 50 mm with TMC, whereas a thickness of 3 – 6 mm is achieved with conventional SMC. Greater thickness allows chopped fibers to assume a three-dimensional random arrangement. For this reason, strength and stiffness are both lower with TMC than with SMC.

3.8. Bulk (or Dough) Molding Compounds (BMC, DMC)

BMC is prepared by mixing chopped strands with a filled polyester resin paste. For reinforcement, BMC usually contains glass fibers, with lengths ranging from 3 to 12 mm, but cellulose, cotton, and other fibrous materials can also be

used. Bulk molding compounds are manufactured by compression molding at 130 – 160 °C, depending on the resin formulation. Temperature control is extremely important in the molding process to reduce internal stresses and ensure reproducible physical properties of the product. Major markets for BMC are the transportation and appliance industries. The products are quickly molded, light, and strong, and may possess complex shapes.

3.9. Prepregs

Prepregs are materials that have been impregnated with partially cured polymer matrix. The matrix differs from that in SMC by the absence of fillers, thickening agents, and pigments. Most prepregs are based on epoxy resins. Reinforcements include glass, carbon, and aramid fibers in the form of roving, woven fabric, or mat. Continuous unidirectional fiber sheets give high-performance multilayer composites, used in high-load-bearing construction parts in the aviation and automobile industries. Prepregs are molded by hand lay-up, bag molding, or autoclave molding.

3.10. Reinforced Reaction Injection Molding (RRIM)

By reaction injection molding (RIM), two chemically reactive liquid streams are injected into a mold where polymerization and formation of the part occur. When short fibers or fillers are added, the process is referred to as reinforced reaction injection molding (RRIM). The fibers and particles must be sufficiently small to fit through the mixing head orifice (< 2 mm). Since viscosities of the essentially monomeric materials used in the RRIM process are low, the injection mixture flows into the mold under relatively low pressure and at low temperature, resulting in savings in energy and equipment.

Polyurethane-based polymers have been widely molded by RRIM. Various mineral fillers (e.g., milled glass fibers, glass fibers, glass flakes and beads, mica, wollastonite, and kaolin) are used to improve the modulus and reduce thermal expansion.

3.11. Resin Transfer Molding (RTM)

In this approach a dry reinforcement material is placed in a mold cavity before the mold is closed and resin is injected. Resin flows through the reinforcement preform, expels air in the cavity, and impregnates the reinforcement. After curing, the molded part is removed and the resin reaction may be allowed to continue to completion through postcuring. RTM resins are low-viscosity polyester epoxy and vinyl ester resins. A broad variety of reinforcements is used, such as continuous strand, cloth, woven roving, long fiber, and chopped fiber mat made of glass, aramid, or carbon.

Resin transfer molding has the potential to become an efficient, low- cost manufacturing process for the fabrication of large, integrated, high-performance products.

4. Physical and Chemical Properties

The properties of a resin – matrix composite depend on many factors, among them the properties of the constituents. The form of the reinforcement (particulate, chopped fibers, woven, etc.); its volume fraction; fiber length, distribution, and orientation; interfacial bond strength; and void content all play important roles. Many of these factors depend strongly on the nature of the fabrication process. The constituents of a composite can be combined in various ways. Therefore, determining the properties of all possible combinations is a practical impossibility. Analytical methods have been developed to permit derivation of a set of predicted properties based on the properties of the composite constituents [9–16].

The fundamental building element upon which fiber-reinforced plastics theory is based is a unidirectionally reinforced layer. Many applications involve components in the form of thin plates or shells [17]. For such structures the components are commonly assumed to be in a state of plane stress; that is, through-thickness deformations and stresses are neglected. As a result, material characteristics in the plane of the structure become of primary interest. In such a case, four elastic constants and two thermal expansion coefficients are necessary to describe the

elastic behavior of the UD layer. Like stiffness, strength is a directionally dependent parameter in composites. The strength properties of UD composites are governed by complex internal failure modes and cannot be described in a way similar to that used for elastic characteristics. Although no universally accepted law currently exists for the prediction of composite failure under arbitrary loading conditions, a number of basic strength properties do allow the structural efficiencies of composites to be evaluated. For typical fiber composites, these properties are listed in Table 3. One of the most important factors affecting composite properties is the amount of reinforcing agent present (see Fig. 9). Here the important parameter is percentage by volume, not by weight. Therefore, volume fractions of components must be used to compare composites of different composition.

As can be seen from Table 3, maximum strength and stiffness are reached in the fiber direction. To provide improved transverse properties, composites are commonly arranged in laminates such that individual layers have different orientations. In this way, the transverse properties of the laminate are enhanced at the expense of some of the potential inherent in a UD layer. An isotropic arangement of reinforcement therefore leads to much lower strength and stiffness values (Figs. 9 and 16). Composites reinforced with discontinuous fibers, such as bulk molding compounds, generally have nonuniform properties because local fiber content and orientation are difficult to control. As a result, material properties vary from point to point throughout the material, which contributes to great variation in measured characteristics, especially strength. Relatively high scatter with respect to properties is a typical feature of composites. For generalized comparisons of material weight-savings potential, the strength – density ratio is an adequate guide when loading is uniaxial. This comparison is very favorable for advanced composites (Fig. 16C). Although reinforcing fibers and fillers are elastic over a wide temperature range, the behavior of a polymer matrix is strongly temperature- and time-dependent. Such behavior can be observed in polymer – matrix composites as well, especially for matrix-governed properties [9, 16–19].

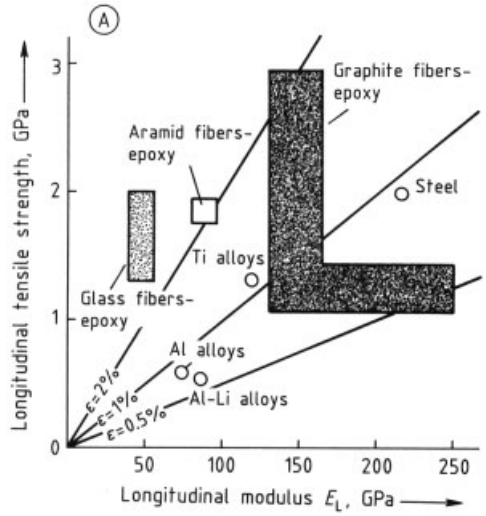

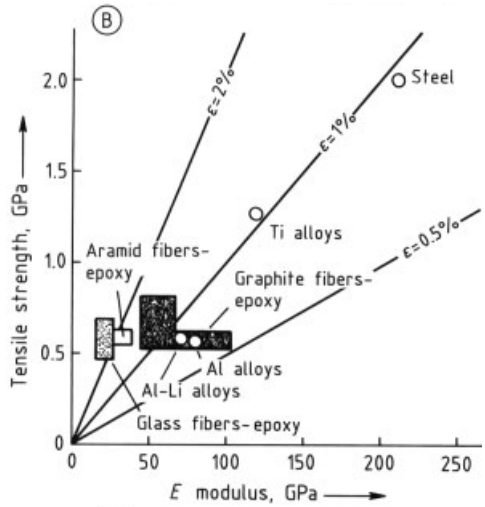

The polymer matrix also plays a dominant role in the chemical resistance of composites. One of the most important tasks of a matrix is to bind the reinforcing elements and protect them against hostile environments. To prevent fiber corrosion, the fiber volume-fraction must be kept low so that absorption and subsequent diffusion of the corrosive medium are minimized. This can be accomplished by protecting the load-carrying layer with a corrosion barrier consisting of thin, resin-rich veil liners. Chemical corrosion of composites can be prevented by careful choice of resin, reinforcement, and additives [20, 21], with the behavior of the resin being most important.

Table 3. Representative mechanical properties of several UD composites (fiber fraction 60 vol %)

Property	E-glass – polyester	E-glass – epoxy	C-HM – epoxy	C-HS – epoxy	Aramid – epoxy
Density, kg/m^3	2000	2000	1650	1600	1400
Young's modulus, GPa					
Longitudinal E_L	43	44	180	140	76
Transverse E_T	12	14	6	8.5	5
Shear G_{LT}	5	6	4	4.5	2.5
Poisson ratio ν_{LT}	0.31	0.30	0.27	0.30	0.36
Tensile strength, MPa					
Longitudinal	800	1000	800	1400	1400
Transverse	25	40	47	50	28
Compression strength, MPa					
Longitudinal	400	450	630	1000	300
Transverse	120	150	190	200	140
Shear strength, MPa					
In-plane	45	80	46	63	60
Interlaminar	40	70	46	70	50
Thermal expansion, 10^{-6} K^{-1}					
Longitudinal	7	7	− 0.7	− 0.3	− 2.7
Transverse	28	26	28	28	54

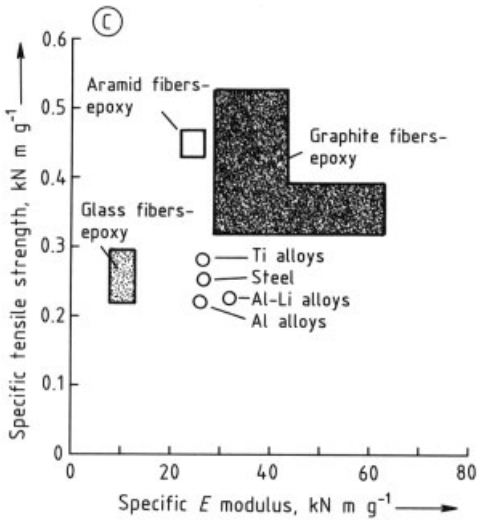

Figure 16. Comparison of strength and stiffness properties of advanced composites with those of high-strength metallic materials
A) UD properties; B, C) Properties of plane-isotropic reinforced composites
ε = typical limit elongation

Of the two categories of polymers used as matrices – thermoplastics and thermosets – thermoplastics show excellent resistance to most acids, bases, and solvents. Among thermoset-based composites, polyester-based resins are generally resistant to acids and solvents, whereas epoxies are more appropriate when resistance to bases is necessary. Special resin systems, such as furans, are also available for specific chemical exposures. Phenolics and epoxies provide superior resistance to various forms of radiation [22].

Good resistance of components to environmental effects can be sustained by good manufacturing practice, appropriate material handling, and the application of consistent control standards. Prolonged storage of resins and fibers, or their exposure to high temperature, UV light, or moisture, may result in degradation and subsequent loss of corrosion resistance [23]. In addition to the effects of temperature, time, and environmental exposure, mechanical loading or fatigue also contributes to the acceleration of corrosion. Stress corrosion can be controlled by a reduction of stress levels, choice of the appropriate laminate structure, and the use of resins with high ultimate elongation [21].

Extensive literature on corrosion and environmental resistance is available from resin and fiber manufacturers, research laboratories, and several government agencies. Corrosion data banks are being developed by the manufacturers and users of composites.

5. Quality Specifications

Quality specification is accomplished by tests to determine the suitability of materials, processes, and designs for the intended applications.

The broad range of applications for reinforced plastics and their diverse properties have necessitated the development of a large number of specialized tests for evaluating mechani-

cal, optical, thermal, electrical, physical, chemical, and durability characteristics. Standard test methods and necessary equipment are described in various sources [24,25], and new testing procedures are constantly being developed [26]. Details about individual methods can be obtained from the literature [27–29]. Specific material features that must be considered when testing polymer composites include:

1) Composition of the constituents. Determination of the fractions of individual components is of prime importance.
2) Anisotropy of properties. The direction of a test (e.g., load application or heat conductivity) should be specified and recorded with the test results.
3) Inhomogeneous structure of the material. With laminates the stacking sequence of layers may play an important role under certain test conditions (e.g., bending, static as well as dynamic fatigue, environmental resistance).
4) Rheological behavior as a function of time. Loading rate and duration, frequency, and loading history must be specified carefully.
5) Relatively great variation in data. A sufficient number of samples must be examined to ensure reasonably reliable test results for the material under evaluation. Statistical analysis of the test data is necessary.
6) Influence of the fabrication process on product properties. Specimens must be prepared and tested under conditions identical to those specified for the product in question.

Tests may be destructive or nondestructive, depending on whether a sample is destroyed or degraded during the test. Nondestructive methods (e.g., ultrasonics, X-ray inspection, acoustic emission, microwave techniques) are becoming increasingly important in ensuring the integrity, reliability, and safety of composite structures [30].

6. Economic Aspects

Based on modern trends in materials technology, the reinforced plastics industry should continue to enjoy rapid growth. The trend can be seen in total shipments of reinforced plastics to the United States, which reached 1.21×10^9 t in 1988 compared to 0.91×10^9 t in 1978, for a mean annual growth rate of nearly 12 %. A similar development can be observed in Europe, where the mean annual growth rate in the 1980s was 9 %. Worldwide growth of composites during 1985 – 1988 averaged 5.5 %, and all predictions for 1990 – 1995 anticipate a growth of ca. 1.5 – 2 times the rate of growth of gross domestic product. Exceptional growth is expected [31] for reinforced plastics with the following characteristics: high performance, high performance – cost ratio, high flame resistance and low smoke index, good adhesion, and corrosion resistance.

Among composites generally, thermosets still dominate thermoplastics as matrix materials. Nevertheless, although thermosets accounted for about 75 % of the demand in the 1980s, this decreased to 70 % in 1990, and the growth in demand for thermoplastics over the first five years of the 1990 s is expected to be double that of composites in general [32].

More than 80 % of the reinforced thermoplastics market belongs to injection-molded resin matrices of nylon, thermoplastic polyesters, olefins, styrene, and polycarbonate. High-performance reinforced thermoplastics in the fastest growing material sector include polyimide, polyetherketone, polybenzimidazole, and polyarylene sulfone. This sector is expected to grow > 10 % annually [33].

The potential exists for significant further advances in composites, in contrast to the relatively minor improvements possible in such wrought sheet-type metals as steel or aluminum. Past and present performance, as well as current areas of research and development, have laid the groundwork for future growth of the composites industry. Effective exploitation of new opportunities is the key to potential large-scale market penetration and consequent profitability for reinforced plastics.

7. References

General References

1. G. Lubin: *Handbook of Composites,* Van Nostrand Reinhold, New York 1982.
2. P. K. Mallick, S. Newmann: *Composite Materials Technology,* Hanser Verlag, München 1990.

3. W. V. Titow, B. J. Lanham: *Reinforced Thermoplastics,* Applied Science Publ., Barking 1975.
4. C. Zweben, H. T. Hahn, Tsu-Wei Chou: *Mechanical Behavior and Properties of Composite Materials, Delaware Composites Design Encyclopedia (DCDE)*, vol. 1, Technomic Publ., Lancaster 1989.
5. M. G. Bader et al.: *Processing and Fabrication Technology, DCDE,* vol. 3, Technomic Publ., Lancaster 1990.
6. H. S. Katz, J. V. Milewski: *Handbook of Fillers and Reinforcements for Plastics,* Van Nostrand Reinhold, New York 1980.
7. D. W. Rosato, C. S. Grove: *Filament Winding,* J. Wiley & Sons, New York 1964.
8. S. M. Lee (ed): *International Encyclopedia of Composites,* **6 vols.,** VCH Publishers, New York 1990 – 1991.

Specific References

9. B. D. Agarwal, L. J. Broutmann: *Analysis and Performance of Fiber Composites,* J. Wiley & Sons, New York 1990.
10. N. N.: *Engineered Materials Handbook,* vol. 1, "Composites," ASM International, Menlo Park, Ohio, 1987.
11. S. W. Tsai, H. T. Hahn: *Introduction to Composite Materials,* Technomic Publ., Westport 1980.
12. R. Christensen: *Mechanics of Composite Materials,* J. Wiley & Sons, New York 1979.
13. S. C. Halpin: *Primer on Composite Materials: Analysis,* Technomic Publ., Lancaster 1984.
14. R. M. Jones: *Mechanics of Composite Materials,* McGraw-Hill, New York 1967.
15. S. W. Tsai: *Composite Design,* Think Composites Publ., Dayton 1987.
16. S. M. Whitney, R. McCullough: *Micromechanical Materials Modelling, DCDE,* vol. 2, Technomic Publ., Lancaster 1990.
17. J. R. Vinson, R. L. Sierakowski: *The Behavior of Structures Composed of Composite Materials,* Martinus Nijhoff Publ., Dordrecht 1987.
18. J. D. Ferry: *Viscoelastic Properties of Polymers,* 3rd ed., J. Wiley & Sons, New York 1980.
19. A. A. Ogale: "Creep Behavior of Thermoplastic Composites" in L. A. Carlsson, R. B. Pipes (eds.): *Thermoplastic Composite Materials,* Elsevier, Amsterdam 1989.
20. P. Schweitzer (ed.): *Corrosion and Corrosion Protection Handbook,* Marcel Dekker, New York 1983.
21. J. H. Mallinson: *Corrosion-Resistant Plastic Composites in Chemical Plant Design,* Marcel Dekker, New York 1988.
22. B. Dolezel: *Die Beständigkeit von Kunststoffen und Gummi,* Hanser Verlag, München 1978.
23. G. S. Springer (ed.): *Environment Effects on Composite Materials,* **vol. 1 – 3,** Technomic Publ. Co., Lancaster 1981 – 1988.
24. ASTM,Standards and Literature References for Composite Materials, ASTM, Philadelphia 1987.
25. ASTM, D 3039, Standard Test Methods for Fiber Resin Composites, Philadelphia 1989.
26. "Composite Materials: Testing and Design," *Proceedings of ASTM Conference,* Philadelphia 1969.
27. G. G. Sih, A. M. Skudra: *Fracture Mechanics and Methods of Testing,* North Holland Publ. Co, Amsterdam 1985.
28. L. A. Carlsson, R. B. Pipes: *Experimental Characterization of Advanced Composite Materials,* Prentice Hall, Englewood Cliffs, N. J., 1987.
29. Yu. M. Tarnopolski, T. Kincis: *Static Test Methods for Composites,* Van Nostrand Reinhold, New York 1985.
30. J. Summerscales (ed.): *Nondestructive Testing of Fiber-Reinforced Plastic Composites,* **2 vols.,** Elsevier, Amsterdam 1987 – 1990.
31. J. S. McDermott: "An Assessment of the Composite Market for the 1990's," *45th Ann. Conf. Comp. Inst.,* SPI, Washington 1990.
32. Frost & Sullivan: "Stetiges Wachstum," *Kunststoff-J.* **4,** Europa-Fachpresse-Verlag, München 1992.
33. G. Sterling: "Developments, Trends and Opportunities for the Plastics Business in the 1990's," *45th Ann. Conf. Comp. Inst.*, SPI, Washington 1990.

Release Agents

HELMUT LAMMERTING, Th. Goldschmidt AG, Essen, Federal Republic of Germany

1. **Introduction** 1
2. **Physical and Chemical Background** . 1
3. **Raw Materials** 2
3.1. **Waxes** 2
3.2. **Metallic Soaps** 2
3.3. **Oils and Fats** 2
3.4. **Silicones** 2
3.5. **Polyoxyalkylenes** 3
3.6. **Fluorocarbons** 3
4. **Methods of Application** 3
5. **Uses** 4
5.1. **Processing Polyurethane Foam** 4
5.2. **Processing Glass Fiber Reinforced Plastics** 4
5.3. **Processing Thermoplastics** 5
5.4. **Processing Rubber** 5
5.5. **Metal Industry** 5
5.6. **Glass Industry** 6
5.7. **Woodworking** 6
5.8. **Food Industry** 6
6. **References** 6

1. Introduction

Release agents are used to reduce or completely eliminate adhesion between two surfaces in contact. They are used in many industrial processes, especially in the plastics processing industry, where the use of effective release agents is indispensable to eliminate friction and ensure economic production.

The many types of release problems that can arise make the choice of a suitable release agent difficult for many users. The material of the mold surface and its physical condition are as important as the chemical constitution and physical properties of the release agent, and process parameters also play a role.

2. Physical and Chemical Background

The relationship between the constitution of release agents and their effectiveness is very complex, so that even today it is only possible to give an incomplete physical and chemical interpretation of release problems. The assessment of release agents is therefore largely subjective. Of course, it is possible to lay down some general requirements for a release agent with good functional properties. The most important of these are:

1) Formation of a smooth, continuous film
2) Good chemical resistance and low reactivity
3) Thermal stability
4) High resistance to oxidation
5) Good stability on the mold surface
6) No effect on the properties of the surface of the molded article, i.e., no blistering, streaking, discoloration, or stress cracking, and low solubility in the plastic
7) Elastic behavior of the release agent film
8) Good physiological properties and skin compatibility

Release agents usually contain several active substances together with additives such as flow improvers and wetting agents, and a carrier. The active substances should be as chemically inert as possible and must be completely free of reactive groups. However, a certain proportion of polar groups is required to give adhesion to the mold surface.

Substances used as release agents should ideally be amorphous, and should have a very low tendency to crystallize, so that they form low energy surfaces (i.e., having low surface tension).

3. Raw Materials

3.1. Waxes

Waxes (→ Waxes) in the form of solvent-based and water-based suspensions or emulsions are widely used in various release agents.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_067

For the product to be effective, it must have a small particle size. Therefore, waxes with a very low crystallizing tendency are used, so that when the suspension dries, dense, elastic films are formed. To achieve the desired properties, different types of waxes must be combined. For use in solvent-based systems, unmodified paraffins are suitable, e.g., Fischer – Tropsch waxes, microwaxes, refined paraffins from mineral oils, and synthetic polyethylene waxes. Fischer – Tropsch waxes and synthetic paraffins have good sealing properties. When used in combination with low-melting paraffins or microwaxes they can give an increase in hardness even at low addition rates. Waxes containing metallic soaps have a lower tendency to crystallize, and therefore form very fine suspensions. They are therefore used with paraffins, which have a higher crystallizing tendency.

For water-based systems, the wax materials described above are less suitable, as they lack hydrophilic groups. Here, acid and ester waxes or ethoxylated waxy alcohols are better, as they can be converted into very finely dispersed emulsions by using an emulsifier. The particle size of the solvent-based suspensions is usually in the range 1 – 10 µm, while for water-based emulsions it can be < 1 µm. Generally, the waxes used are in the solid state under conditions of use. If they are used within their softening range, a deterioration of the release and lubricating properties is usually observed. Above its melting point, the wax loses its lubricating properties and chemical resistance, and has a noticeable effect on the surface of the plastic molding. Consequently, it is unsuitable as a release agent.

Waxes often have the disadvantage that solid residues accumulate on the molds and must be removed mechanically from time to time.

3.2. Metallic Soaps

Metallic soaps (→ Metallic Soaps) have long been known as lubricants and release agents. They have better thermal stability and lubricating properties than waxes. Solutions of metallic soaps usually form very dense films. However, they have the disadvantage that they form rather hard residues on the molds. Removal of these residues is difficult and time consuming.

Metallic soaps are used in combination with paraffin waxes to produce finely dispersed suspensions. Due to their polar character, they can improve the adhesion of the release agent film to the metal surface. This is beneficial, as the film then cannot be dislodged by contact with the molding material.

Metallic soaps are important in the processing of thermoplastics and in reaction injection molding (RIM) of polyurethanes, where they act as internal lubricants and release agents.

3.3. Oils and Fats

Esters of higher fatty acids and gylcerides are used as release agents in processing many types of polymers, especially for molding processes operated below 150 °C. The polar groups present give good adhesion to the metallic molds. Fatty acid esters are also used as internal release agents and lubricants in the plastics industry. Synthetic ester oils, spindle oils, and animal oils are also used to a limited extent in the production of release agents.

3.4. Silicones

Silicones (→ Silicones) have been used in industry as release agents and lubricants for ca. 40 years. They are usually liquids of medium to high viscosity, preferably in the range 100 – 10 000 mPa · s. Very low viscosity siloxanes (< 100 mPa · s) do not possess the necessary mechanical stability. Extremely high viscosity oils ($> 10\,000$ mPa · s) can form adhesive surfaces with reduced lubricating properties. Silicones are chemically inert and have no affinity to other substances. Their most important advantage is their very good thermal stability and a viscosity that varies only slightly over a wide temperature range. Silicones are used both as aqueous emulsions and as solutions in various solvents. Owing to their low surface tension (ca. 20 mN/m) and their incompatibility with most plastics, they are among the most effective release agents. A disadvantage is that the molded articles readily pick up traces of silicone that can interfere with subsequent operations such as painting or bonding with adhesives. The silicone oil release agents can be used either alone or in combination with other substances

having release properties. Medium-viscosity silicone oils (100 – 1000 mPa · s) are especially suitable. The release properties of these linear polydimethylsiloxanes are improved by the addition of small amounts of silicone resin. Such combinations of materials often have improved barrier-layer properties, and a very effective film of release agent is produced.

Organomodified siloxanes are sometimes used as release agents and lubricants. These are mainly linear dimethylpolysiloxanes that contain organic end groups and/or side groups. Substitution by methyl groups causes partial weakening of the silicone character. For example, it is possible to obtain a water-soluble product by the chemical addition of polyoxyalkylenes. Alternatively, alkyl or ester groups can be included, giving products that have some compatibility with other polymers, and can thus be used as internal release agents and lubricants. The introduction of organic groups lowers the tendency of the siloxane to cause paint defects. Water-soluble silicones can very readily be removed from a mold after use by washing.

3.5. Polyoxyalkylenes

Polyoxyalkylenes are used as mold release agents especially in the rubber industry. They have good lubrication properties and impart an attractive finish to the vulcanized products. They can be used neat or in aqueous solution if they are water-soluble. Any residual release agent can easily be removed from the rubber articles by washing with detergent solutions.

3.6. Fluorocarbons

Fluorocarbons are used as permanent release agents in many industries. They are usually applied to the mold surfaces in the form of fluorotelomer dispersions. After the solvent has evaporated, a coherent film is produced by sintering or melting at temperatues above 265 °C.

4. Methods of Application

Release agents are generally used in the form of solutions, dispersions, or emulsions. The most widely used materials are dispersions or emulsions that can be applied to the surfaces of the mold by hand-operated or automatic spraying equipment, or by aerosol cans. The carrier liquids in sprayable release agent dispersions were at one time mainly organic solvents such as aliphatic hydrocarbons, chlorinated hydrocarbons, and fluorochlorohydrocarbons, or mixtures thereof. However, due to the increasing restrictions on emissions, and on the use of chlorofluorocarbons, aqueous release agents will be favored in future.

If the rate of evaporation of the water film is insufficient, release agent concentrates or low-solvent release agent preparations can be used. If these are of low viscosity, they can be applied by air-atomizing spray guns, which give good distribution of the active material.

The use of internal release agents has been increasing for some time. These materials, known as internal lubricants, are mixed with the plastic prior to processing. Internal lubricants have long been used in thermoplastics to reduce internal friction. Materials often used include paraffins, ester waxes, and metallic soaps, which are often soluble in the plastic. Additives are also used which have a definite solubility limit in the polymer, so that the undissolved part of the lubricant migrates to the surface, i.e., to the walls of the mold, where a very thin film is formed, thus giving both internal and external lubrication.

For reaction polymers, especially for polyurethanes, the internal release agents must have special properties. Mixing of the liquid components must lead to good emulsification of the internal release agent or lubricant. On formation of the polymer, this emulsion breaks, and the release agent separates from the polymer matrix. Sufficient release agent then migrates to the outer surface of the molding. For internal lubrication, metallic soaps and some organomodified silicones are mainly used.

5. Uses

There are many industrial processes that require release agents, so that the demand for these materials is high. Manufacturers of release agents include Th. Goldschmidt, E & P Würz, Acmos Chemische Fabriken, Tietgen & Co., Teroson

Werke, Klüber Chemie, and Wacker Chemie (Germany); Air Products & Chemical, Dow Corning, Dexter Corporation, and Chem-Trend (United States); and Rhône-Poulenc (France).

A brief review of the release agents suitable for particular production processes is given below.

5.1. Processing Polyurethane Foam

Polyurethanes are formed by the polyaddition reaction of polyhydroxy compounds with diisocyanates. The initially liquid reaction mixture used to make polyurethane foam develops strong adhesive forces to the mold. Very effective release agents must be used to enable the hardened molding to be released undamaged from the mold. Most release agents used in polyurethanes are mixtures of several waxes. The wax raw materials are dispersed at concentrations of ca. 5 – 10 % in organic solvents, preferably aliphatic hydrocarbons, or in water. A combination of microwaxes with waxes that contain metallic soaps gives very fine dispersions in organic solvents. These are used in a wide variety of polyurethane foams in the molding temperature range 30 – 70 °C. Ester or acid waxes are used in water-based release agents. Ethoxylated alcohol waxes contain water- compatible groups and can be converted to very fine emulsions by using emulsifying agents. The choice of the emulsifier for a water-based product is extremely important. It must by physiologically harmless and must not cause skin irritation; hence, nonionic emulsifiers are preferable. Water-based release agents are preferred for the production of flexible polyurethane foams. The mold temperature must be > 50 °C to evaporate the water film within a reasonable time.

Specialty silicones are also used in polyurethane release agents. Pure silicone oils, i.e., unmodified dimethylpolysiloxanes, are generally unsuitable. They exhibit strong abhesive action, but do not adhere sufficiently to the mold wall and have a strong antifoaming effect on many types of foam. In the production of highly cross-linked foams with densities > 0.3 g/cm^3, special silicone release agents are used to a limited extent. These must have a proportion of polar groups to give adhesion to the mold wall and prevent the release agent from becoming detached on contact with the molding material.

Solvent-based and water-based polyurethane release agents are best applied by air-atomizing spray equipment. This has the advantage that the air stream takes up some of the solvent or water, and the droplet dispersion is improved. Aqueous release agents are mainly used in the production of flexible foams (hot- cured foams) with mold temperatures of 100 – 150 °C. With cold- cured or high-resilience (HR) foam, problems can occur due to the lower molding temperatures (30 – 60 °C). Small amounts of residual water are thus unavoidable, and these react with isocyanate to form hard residues of insoluble, infusible, polymerized urea products. The use of water-based release agents also often leads to problems during later stages of manufacture. Contamination by the release agent gives the foam a greasy or oily surface, and interferes with the action of adhesives used to bond the moldings to other materials.

5.2. Processing Glass Fiber Reinforced Plastics

Glass fiber reinforced plastics (GRP) are usually made with unsaturated polyester or epoxy resins. The polyester resins in particular contain large amounts of reactive solvents such as styrene, and the only release agents that can be used are those that are insoluble or only slightly soluble in the monomers used in the polymerization. Moreover, the molding contracts on polymerization, and can shrink tightly onto the pattern, especially in large molds. Therefore, a good GRP release agent must also have good lubricating properties. Release agents based on hard waxes are most commonly used, usually in the form of liquids or pasty dispersions in aliphatic hydrocarbons. These also contain additives to improve lubrication, such as silicones, fats, oils, or metallic soaps. Silicone release resins are very effective for metallic molds in which epoxy resins are used. They are baked onto the surface of the mold and give extremely durable separating surfaces. Their working life can be still further increased if the mold surfaces are thinly coated with a lubricant, e.g., a silicone grease, between molding operations.

5.3. Processing Thermoplastics

For the injection molding of thermoplastics, fats, oils, and silicone oils are used as release agents, usually in the form of aerosols applied from cans. The decision whether to use a silicone or a fatty oil depends on the operating temperature. As a rough guide, silicone oils are used above 150 °C, and methylphenylsiloxanes above 250 °C. Silicone resins are very effective in the rotational molding process. They are baked on before casting is started, and enable even transparent moldings to be produced free from optical defects caused by scoring or pitting.

5.4. Processing Rubber

Rubber mixtures are formed into a wide range of articles. The release agent must be carefully formulated for each process, taking account of all relevant production conditions. In general, silicones, certain metallic soaps, polyoxyalkylenes, and surfactants (e.g., organic sulfate esters or sulfonates) are used singly or in combination. Sometimes, mineral products such as mica or talc are also used. Silicones are suitable for all vulcanization processes owing to their temperature stability and excellent release and lubricating properties, even when used in very small quantities. Polyethersiloxanes are particularly effective; they are water-soluble and can be removed from the moldings by washing.

Polyoxyalkylenes are very effective release agents for rubber. They give a velvety sheen and a dark appearance to the products, and accentuate their color. Polyoxyalkylenes have good lubricating properties. They can be made water soluble if they contain a proportion of ethylene oxide units, i.e., 30 – 50 % depending on molecular mass and degree of polymerization. Polyoxyalkylenes give almost no residues, and can be removed by a water – detergent mixture. In general, polyoxyalkylenes are used in preference to silicones if the manufacturer or end user requires silicone-free moldings. The release agents used to aid the removal of the rubber articles should, if possible, be biodegradable, and should have no harmful ecological effects. The surfactants used are mainly ethoxylated fatty alcohols, which are usually 80 % biodegradable.

The release agents are usually applied as aqueous emulsions or solutions by spraying with stationary equipment. Another common technique is to pass the preforms through a bath of release agent before molding.

5.5. Metal Industry

The mass production of metal castings requires rapid production of the necessary molds by, e.g., the Croning, hot box, or cold box processes, in which a mixture of molding sand and resin is applied to a pattern and hardened to give a mold or core. In these processes, the release agents must prevent adhesion of the resin binder to the pattern, enabling the sand mold to be slid away from it. In the Croning and hot box processes, the operating temperatures are high, and aqueous emulsions of silicones are therefore used. They have the advantage of forming a very thin coating on the pattern, thereby ensuring dimensional accuracy. In some cases, when temperatures are fairly low (< 300 °C), mineral oil emulsions can be used. Because the cold box process is operated at lower temperatures, release agents based on solvent-based dispersions or solutions are required. Silicones and hard wax dispersions are also suitable. The silicones used must be compatible with any mold coatings used.

For the injection molding of low-melting alloys, e.g., pressure die-casting of aluminum, organomodified silicones are used in combination with waxes.

5.6. Glass Industry

In the molding and shaping of glass, release agents are used both on the mold and on the metallic parts of the conveying system that transports the molten glass from the metering equipment to the mold. Although simple, cheap release agents such as mineral oils and graphite are effective, they produce large amounts of smoke and lead to the build-up of residues. Modern release agents for glass consist essentially of heat-resistant graphite-containing resins, which give hard, smooth, abrasion-resistant films. Phenolic resins, silicone resins, and some silicone – polyester resins are used.

5.7. Woodworking

Release agents are required as auxiliaries for gluing and veneering wood and wood-based materials, and for coating with synthetic resin films. When applying paper impregnated with phenolic and melamine resins, it is necessary to coat the sheet with a release agent to prevent the surface of the press from adhering to the cured surface of the resin. The release agent can be applied by spraying, spreading, or by rollers. For very large surface areas, it can be applied by a transfer process, i.e., by squeezing it out of a sheet of paper, fabric, or nonwoven soaked in the release agent. Suitable release agents include mineral oils, high-melting paraffin waxes, esters of long- chain aliphatic carboxylic acids, and sometimes silicones modified to make them paintable.

5.8. Food Industry

Release agents have been used for centuries in baking or roasting foodstuffs. Vegetable and animal fats are still used in the bakery industry, either pure or as emulsions. In recent decades, however, baking sheets and tins have been coated with silicone release coatings. Household cooking equipment such as saucepans and baking tins are coated with polytetrafluoroethylene or silicone resins to prevent food from adhering during roasting, cooking, and baking.

6. References

1. M. McDonald: *Release Agents,* Noyes Data Corp., New Jersey 1972.
2. H. Lammerting, *Goldschmidt Informiert* **58** (1983) 1.
3. V. M. Folgueras: Formteilfertigung aus PUR-Strukturschaum, Einfluß auf das Entformungsverhalten, Dissertation, TH Aachen, D 82/75.
4. D. B. Cox: "The Role of Release Agents in Urethane Molding", *ACS Symp. Ser.* **172** (1981) 425.

Renewable Energy Resources and Raw Materials → Raw Materials and Energy
Reprography → Imaging Technology

Resins, Natural

KLEMENS FIEBACH, Lackharzwerke Robert Kraemer, Bremen, Federal Republic of Germany (Chaps 1 – 4, Sections 5.1 – 5.4, 5.5.1 – 5.5.6, 5.5.8, 5.5.9, 5.7, 5.8)

DIETER GRIMM, Abieta Chemie GmbH, Gersthofen, Federal Republic of Germany (Sections 5.5.7 and 5.6)

1. Introduction 2
2. Systematics and Classification 2
3. Formation – Collection 3
4. Individual Resins 3
4.1. Acaroid Resin 3
4.2. Amber 3
4.3. Asphaltite 4
4.4. Balsam of Peru 4
4.5. Balsam of Tolu 4
4.6. Benzoin 5
4.7. Canada Balsam 5
4.8. Chinese or Japanese Lacquer 5
4.9. Copal 6
4.10. Damar 6
4.11. Dragon's Blood Resin 7
4.12. Elemi 7
4.13. Frankincense (Olibanum) 7
4.14. Galbanum 8
4.15. Labdanum 8
4.16. Mastic 8
4.17. Myrrh 8
4.18. Sandarac 8
4.19. Shellac 8
4.20. Styrax (Storax) 9
4.21. Utah Resin 9
4.22. Venice Turpentine (Larch Turpentine) 10
5. Rosin (Colophony) 10
5.1. Occurrence 10
5.2. Collection – Extraction 11
5.2.1. Gum Rosin 11
5.2.2. Wood Rosin 11
5.2.3. Tall-Oil Rosin 12
5.3. Composition 12
5.4. Properties 12
5.5. Modified Rosin Types 14
5.5.1. Rosinates 14
5.5.2. Rosin Esters 14
5.5.3. Maleic Acid Adducts (Maleics) 15
5.5.4. Rosin-Modified Phenolic Rosins . . . 16
5.5.5. Polymerized – Dimerized Rosins . . . 16
5.5.6. Hydrogenated Rosins 17
5.5.7. Disproportionated Rosins 17
5.5.8. Rosin Alcohols 17
5.5.9. Rosin Acid Nitrile, Rosin Amine . . . 17
5.6. Economic Aspects 18
5.7. Toxicology and Food Regulations . 18
5.8. Analysis 18
5.8.1. Melting-Point Determination 18
5.8.2. Crystallization Test (Acetone Test) . . 18
5.8.3. Other Methods 19
6. References 19

1. Introduction

Natural resins and balsams were known and used in the earliest human societies. The widely differing uses ranged from surface protection and decoration (e.g., Japanese lacquer, used in the third century B.C.); to preservation and mummification (Egypt), putty and adhesives, medicinal and cosmetic purposes (e.g., disinfectant sealing of wounds); to ritual and ceremonial purposes (incense). Because of industrialization and the ever-deepening knowledge of chemistry and chemical synthesis a large number of the previously important natural resins have been replaced by synthetic ones. Higher standards of living and advanced civilization would presumably make excessive demands on the world capacity of natural resins in every aspect. Nevertheless, some natural resins have retained their industrial importance and are mentioned here.

Current attempts to use nonrenewable resources as sparingly as possible have focused attention more strongly on natural (i.e., renewable) raw materials. This tendency also applies to natural resins (e.g., rosin) as a natural, renewable resource. Rosin is treated in detail in Chapter 5.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_073

2. Systematics and Classification

For a definition of the term resins, see → Resins, Synthetic, Chap. 1.

Natural resins can be classified in several ways:

Botanical Aspects. Resins can be derived from gymnosperm (naked seeders) and from angiosperm (covered seeders); the latter are subdivided into monocotyledons (with one embryonic leaf) and dicotyledons (with two embryonic leaves).

Physical and Chemical Properties. Natural resins are also classified according to their behavior toward typical basic solvents or their melting behavior. This classification takes aspects of application technology into account. However, it does not differentiate finely enough. A classification according to occurrence seems to be more appropriate.

Occurrence. The classification into fossil, recent, and recent – fossil resins gives a basic characteristic that can then be differentiated further (e.g., according to physical and chemical properties). *Fossil resins* are those that come from plants which are extinct; some of these species are not even known for certain. Some fossil resins have undergone considerable change in their chemical structure through aging or ripening, which may have taken thousands of years. The transition from fossil to *recent – fossil* is flexible. The latter can include resins that are both found fossilized and harvested from living plants.

Recent resins are tapped (harvested) from the living plant. The plant – balsam or plant – resin relationship is clearly recognizable. Recent resins are particularly interesting from an ecological point of view because they are the only self-regenerating resins. The terms *balsam* and *turpentine* have practically the same meaning today (i.e., the resinous exudate from living trees and shrubs). Balsam or turpentine is a solution of plant-specific resins in plant-specific solvents (essential oils). Such resins can include resin acids, esters, or alcohols. The exudate is a mobile to highly viscous liquid and often contains resin particles that have already crystallized. In the course of time and as a result of other influences, the exudate has lost the liquidizing components or been chemically converted into a solid material (i.e., by autoxidation).

3. Formation – Collection

The formation of *recent resins* has been most intensively studied in the case of pine balsam. Pine balsam is a plant metabolite (plant exudate) that is formed in extract repositories in the plant. The balsam is collected in so- called resin canals, which are connected to each other lengthwise and crosswise. On opening these canals the so- called *physiological resin* flows out. Severe wounding and additional irritation (irritant resinification) lead to an increase in the number of resin canals and increased balsam production (*pathological resin flow*).

Recent – fossil and *fossil resins* are obtained from the ground by prospecting and mining. These resins often occur in secondary deposits, which are not in the original location of the plant that produced the resin. These secondary deposits may have arrived at their current position through geological faults or through transportation by water (e.g., amber or damar).

4. Individual Resins

Rosin is described in Chapter 5.

4.1. Acaroid Resin

Acaroid resin is a recent resin that is isolated from *Xanthorrhoea* species, which occur in Australia.

Appearance. Yellow and red types of acaroid resin are known, which also differ in their physical properties. Red acaroid resin is the more important.

Properties. Acid value 60 – 110 (red), 125 – 140 (yellow); saponification value 160 – 200 (red), 200 – 240 (yellow). Unlike the red type of resin, the yellow type contains cinnamic and benzoic acids, esters of *p*- coumaric acid, and xanthoresinotannol (ca. 80 %). The red type contains erythroresinotannol (ca. 85 %).

Uses. Acaroid resin is used in cosmetics, paints and coatings, and printing inks, as well as for staining wood (principally the red type) and for fireworks.

4.2. Amber

Amber is a fossil resin that originates from conifers of the tertiary period (*Pinites succinifera*). Amber was presumably formed by secondary resin flow and molded to its present form by pressure, temperature, and other influences in the course of thousands of years.

Occurrence. Amber is found at the Samland coast (Baltic coast) and the Samland peninsula (Palmnicken area). It is obtained either by collecting or fishing on the coast (sea amber) or by isolation from "blue earth," which is mined at Palmnicken (yield: ca. 1 – 2 kg/m^3 excavated material).

Appearance. Amber forms bright-yellow to brownish (amber- colored), angular or rounded pieces of different sizes that can contain inclusions.

Properties. Amber is a very hard material that is virtually insoluble in all solvents and a very good dielectric. Amber acquires a negative charge on rubbing and was therefore called *electron* in the old Greek. It decomposes at ca. 370 °C without melting. In contrast to other resins, amber contains organically bound sulfur. Its composition can vary widely. The following approximate elemental composition can be found in the literature: 78 % carbon, 9.9 % hydrogen, 11.7 % oxygen, 0.4 % sulfur, and 0.2 % ash; acid value 15 – 35, saponification value 90 – 125.

Uses. Even in antiquity amber was a very valuable commercial product and was used in jewelery by the Romans. Amber was also used for special oil varnishes in former times. For this it had to be rendered oil soluble by melting. During this treatment, losses of up to 30 % occur, and amber oil and succinic acid are formed. The term amber varnish no longer signifies that the varnish is produced from amber.

4.3. Asphaltite

Asphaltite or *gilsonite* is a fossil resin (→ Asphalt and Bitumen, Chap. 2.) that is also known as glossy pitch. It is a neutral, thermoplastic hydrocarbon resin. According to DIN 55 946 asphaltite is defined as natural asphalt with a low mineral content and great hardness [2]. This natural product is assumed to be formed from crude petroleum by evaporation of volatile components and chemical reaction (autoxidation, polymerization, etc.). Natural asphalt was known to early civilizations (Babylonians, Sumerians, ca. 3000 B.C.), where it was used as a sealing material.

Occurrence. Hard natural asphalt (asphaltite) is found in Utah/Colorado (gilsonite), Virginia, and Syria. The asphalt deposit in Utah was discovered in ca. 1860 by Samuel H. Gilson and subsequently exploited industrially. *Soft natural asphalts* (pitches) occur in Trinidad (Trinidad asphalt), the Bermudas, Cuba, Mexico, and Palestine.

Extraction. Gilsonite is found in vertical underground seams several meters wide. The depth of these seams can be up to 300 m; they can be several kilometers long. The individual seams run parallel to one another in a northwest – southeast direction.

Appearance. Gilsonite forms black, shiny pieces with conchoidal fracture; it has a dark-brown color in the pulverized state and in solution.

Properties. Gilsonite has a softening point of 145 – 195 °C (ring and ball); an acid value of ca. 2; and a density of 1.03 – 1.15 g/cm^3. It is soluble in carbon disulfide and aromatics, less soluble in aliphatics and mineral oils. Gilsonite solutions readily thicken on storage, but this tendency can be counteracted by addition of polar solvents. Lower alcohols and glycols act as pure dilution agents; gilsonite is only partially soluble in ketones and some glycol ethers. Gilsonite is compatible with many natural and synthetic resins and oils, for example, with modified rosins (e.g., rosinates, maleics, phenol-modified rosins). Gilsonite has very good wetting properties for carbon black.

The approximate elemental composition of *gilsonite* is 85 % carbon, 10 % hydrogen, 2.5 % nitrogen, 1.5 % oxygen, and 0.3 % sulfur. It contains approximately 50 – 65 % asphaltenes, 35 – 45 % resins, 1 – 5 % oils, and 0.2 % ash. With increasing softening point the proportion of asphaltene increases and that of resins and oils decreases. *Trinidad asphalt* contains ca. 82 % carbon, 10 % hydrogen, 1 % nitrogen, 6 % sulfur, and 0.5 % oxygen.

Uses. Gilsonite is an important component of various printing inks (e.g., gravure printing for illustrations and letterpress). Other uses of Gilsonite and Trinidad asphalt are in paints and coatings for buildings, industrial floor coverings, electrical insulating materials, hot-melt materials, and fillers for thermoplastics [3], [4].

4.4. Balsam of Peru

Balsam of Peru is obtained by tapping of living trees of *Myroxylon pereirae* and *M. balsamum* that occur in Central and South America.

Properties. Balsam of Peru is a viscous, yellow-brown liquid, that is nonsticky and nonstringy. It does not dry in air [5]. Balsam of Peru contains large quantities of benzyl cinnamate and benzoate. Its acid value is 60 – 80; the saponification value, 230 – 255. Balsam of Peru is used in perfumes, cosmetics, and medicine (for treating eczema).

4.5. Balsam of Tolu

Balsam of tolu is a recent resin that originates from South America (Columbia, Peru, Venezuela). It is tapped from the living trunks of *Myroxylon toluiferum*.

Properties. Balsam of tolu is a brownish, sticky, semisolid mass, which slowly becomes solid and brittle. Its acid value is 112 – 168; its saponification value 154 – 216; balsam of tolu is soluble in ether, ethanol, and alkali. In addition to resin it contains a high proportion of benzoic and cinnamic acids, their esters, and vanillin.

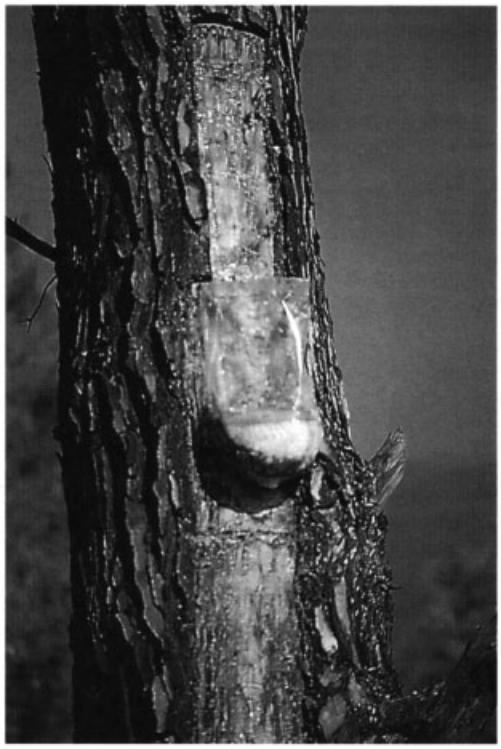

Figure 1. Rosin extraction from a living pine tree

Use. Balsam of tolu is used in perfumes and medicine (expectorants).

4.6. Benzoin

Benzoin resin is a recent resin; the resin forms only after wound irritation (pathological resin). It occurs in diverse Styrax species that are found in Thailand and Indonesia.

Appearance. Two commercial forms of benzoin exist: (1) *beads* or *tears* that are flat, rounded, light brown (their odor is similar to vanilla); and (2) *block benzoin*, which forms irregular brown to reddish pieces (impure).

Properties. Differences exist among Siamese, Sumatran, and Penang benzoin. Unlike Siamese benzoin, Sumatran benzoin contains cinnamic acid in addition to benzoic acid. Sumatran benzoin contains ca. 75 % alcohol-soluble components; Siamese benzoin, up to 90 %.

Uses. Benzoin is used in perfumes and cosmetics, pharmaceuticals (only Siamese benzoin), veterinary medicine, and scented candles.

4.7. Canada Balsam

Canada balsam is a recent resin that originates from *Abies balsamea* (balsam fir, *Abies canadiensis*), which occurs in Canada and the United States. It is tapped from the fir tree by making cuts under the bark. Canada balsam is a colorless to greenish, viscous, slowly solidifying liquid.

Properties. Canada balsam has an acid value of 82 – 87 and a saponification value of 89 – 100; density 0.985 – 0.995 g/cm^3; refractive index 1.5180 – 1.5210 (similar to glass); it is highly soluble in ethanol. Canada balsam belongs to the terpene hydrocarbons.

Uses. Because of the similarity of its refractive index to that of glass, Canada balsam is used mainly for optical purposes (e.g., cementing optical glass elements, embedding microscopy specimens).

4.8. Chinese or Japanese Lacquer

Chinese and Japanese lacquers are recent resins that are isolated from the lacquer tree (*Rhus vernicifera*), which occurs in China, Japan, and India. Chinese – Japanese lacquer is tapped from the living trees and by extracting parts of plants such as the branches and twigs (branch lacquer). The yields are very low.

Appearance. Chinese – Japanese lacquer is a highly viscous, gray to brown milk.

Properties. The latex dries in thin layers in air by enzymatic oxidation, which is favored by moisture, darkness, and low temperature. Daylight and higher temperature (above 30 °C) inhibit the action of the oxidase. An important component of the latex is urushiol, a derivative of catechol with an unsaturated C_{15} side chain, partially soluble in alcohol.

Uses. The well-known Japanese lacquerwork used to be carried out with the latex of the Japanese lacquer tree. It was handcrafted in the thinnest possible layers because of enzymatic oxidation. Japanese lacquer was applied to wood and leather, and could be polished to a high gloss after drying. The current term "Japanese lacquer" is a general quality specification, which has nothing to do with the classical Japanese lacquer from *Rhus vernicifera*.

4.9. Copal

Copal is the collective term for a large number of recent to recent – fossil natural resins with variable structure and very different botanical and geographical origins. Very different qualities thus exist. Copals are subdivided according to geographical criteria: kauri copal, Manila copal, West African copals (e.g., Congo, Angolan, or Cameroon copal), East African copals (e.g., Zanzibar or Madagascar copal), South American copals (e.g., Brazilian or Columbian copal). Kauri, Manila, and Congo copals are still relatively important; the other types are described in [6], [7]. Most types of copal show both an insufficient solubility in common solvents and insufficient compatibility with oils. They must therefore undergo destructive distillation to obtain cleavage products that are more compatible. These pretreated copals can be further processed e.g., by esterification.

Kauri Copal. Various types of kauri copal exist, all of which originate from the kauri spruce (*Agathis australis*) that occurs in New Zealand. They are either mined (fossil), dug from topsoil (bush copal, recent – fossil), or harvested from living trees (tree copal, recent). Kauri copal is whitish yellow, but often dark; the surface is weathered and opaque.

Physical Properties. The physical properties of kauri copal vary with the age of the resin. The acid value is 50 – 100; the saponification value, 75 – 120; and the softening point, 110 – 115 °C (Kraemer Sarnow). Kauri copal is completely soluble in higher alcohols, moderately soluble in lower alcohols, and sparingly soluble in aromatics and aliphatics. Copals used to be important raw materials for copal varnishes, which

had a high gloss, elasticity, and good weathering resistance. They were later replaced by rosin-modified phenolic resins (synthetic copals, see Section 5.5.4). Kauri copal is now used only in determining the kauri – butanol value. A low kauri – butanol value denotes a good solvent power of a solvent (high aromatic content) and is determined by titration to the cloud point of a standard solution of kauri copal in butanol against the solvent to be investigated.

Manila copal is a recent to recent – fossil resin, which is isolated from *Dammara orientalis*, which occurs in the Philippines and Indonesia. Manila copal is harvested from the living tree. An alcohol-soluble soft resin is obtained, which is insoluble in aromatics and aliphatics. A recent –fossil semihard manila copal resin (loba) is insoluble in alcohol. Manila copal needs to be melted, which involves losses of up to 20 %. It is a yellowish, often reddish to dark- colored resin. A main component is agathene dicarboxylic acid ($C_{20}H_{30}O_4$), derived from diterpenes, which is also present in kauri copal.

4.10. Damar

Damar is a recent to recent – fossil resin that is obtained from trees of the Dipterocarpaceae family.

Location – Harvesting. *Recent white damar* is tapped in Batavia, Sumatra, and Thailand, for example. *Recent – fossil brown and black damar* is mined in the East Indies and Sumatra.

White Damar. White damar forms irregular, yellowish- or reddish-white, transparent pieces of different sizes. It has the following composition: 40 % α-damar resin (alcohol soluble); 22.5 % β-damar resin (alcohol insoluble); 23.0 % damarol acid $C_{30}H_{50}O_6$ (a compound with one carboxyl and four hydroxyl groups); 2.5 % water; 3.5 % ash; 8.0 % impurities; and 0.5 % essential oils. White damar has a *mp* of ca. 80 – 90 °C (capillary), an acid value of 20 – 30, and a saponification value of 35 – 60; it is soluble in aromatics, pentanol, pentyl acetate, and carbon disulfide; moderately soluble in other alcohols, ethyl acetate, ethers, ketones, and aliphatics. It is rendered compatible with alkyd resins and oils by cooking. For nitrocellulose combinations, damar is separated from its alcohol-insoluble components (β-damar resin) by dewaxing. Here the β-damar resin components are precipitated from a solution of damar in toluene or toluene – ethyl acetate by the addition of alcohol.

Uses. Damar is still being used in combination with oils for special paints and coatings and for printing inks. The alcohol-soluble part formed during dewaxing process (Liodammar [8]) is used in nitrocellulose combinations and the alcohol-insoluble part (Lio-Resen) [8] in alkali- and acid-resistant paints and coatings.

Black and Brown Damar. Being a recent – fossil resin, black – brown damar is harder than the recent white damar. Black – brown damar has a *mp* of 160 – 180 °C (capillary); an acid value of ca. 25; and a saponification value of ca. 25. If used in combination with oils and alkyd resins, it is often melted to achieve more stable combinations. The resin is then soluble in aromatics and esters, and insoluble in alcohols.

4.11. Dragon's Blood Resin

Dragon's blood resin is a recent resin that originates from *Calamus draco* (climbing palms, also dragon or blood tree), which occurs in the East Indies and Sumatra. The resin is secreted naturally from the fruit of the climbing palm and is removed after solidification (tears). The fruit is subsequently heated in steam or extracted by boiling with water. Dragon's blood resin forms red beads, sticks, or globules. It has an *mp* of ca. 80 – 100 °C; is soluble in alcohols, ethers, and ketones; and insoluble in aliphatics. It gives bright-red films and is used for staining wood and as a colored resin for alcohol coatings and violin varnish.

4.12. Elemi

Elemi (recent) is the collective term for a group of resins derived from plants of the Burseraceae

family (*Canarium indicum*). Each type is described according to its country of origin; the most important type comes from the Philippines (Manila elemi). Elemi is tapped from the living tree; the resin appears only after the plants have been wounded (pathological resin flow) and solidifies with time. Elemi is a yellowish, also greenish, opaque, ointment-like, glutinous, sticky mass that solidifies to a brownish resin which is interspersed with crystals.

Properties. Elemi is soluble in aromatics, alcohols, esters, and carbon disulfide; and less soluble in aliphatics. Elemi has an acid value of 18 – 34, a saponification value of 25 – 60, and a softening point of ca. 80 °C. The exuding balsam contains up to 30 % essential oils [9].

Uses. Elemi was previously used as a (temporary) plasticizer in spirit varnishes and in combination with copals. Elemi oil is obtained from elemi resin by steam distillation, and is used in pharmaceuticals and in cosmetics and perfumes.

4.13. Frankincense (Olibanum)

Frankincense occurs in the United Arab Emirates, Oman, Somalia, Ethiopia, and the East Indies. The recent resin is tapped from the frankincense tree *Boswellia carterii*. The bark is intentionally wounded to obtain a milky extract that is collected after drying.

Properties. Frankincense forms pale-yellow, rounded or globular beads, as well as irregular pieces. It contains ca. 33 % boswellic acid ($C_{32}H_{52}O_4$). Frankincense has an acid value of 30 – 50 (indirect) and is moderately soluble in ethanol and alkali.

Uses. In antiquity, frankincense was known as an important fumigating resin for social and religious purposes (e.g., for funerals). It is also used in cosmetics and medicine. Besides pleasant odor, the disinfectant affect certainly plays an important role in its use. Incense was an important commercial product of the Middle East and the source of its great wealth (Arabia felix) [10]. The incense route of antiquity ran from what is now Yemen through the former kingdom of Sheba to Egypt and via Aqaba – Negev to Gaza. With time, this route by land lost its importance as the regularity of the tides and the monsoon winds were recognized and exploited.

4.14. Galbanum

Galbanum is a recent resin that is obtained from various Iranian Ferula species, e.g., *Ferula galbaniflua*. Galbanum exudes spontaneously from the bark as a milk. It is collected as a resinous gum after drying. It forms brownish to greenish-yellow beads or tears with regular rounded appearance; older resin is darker.

Properties. Galbanum contains ca. 60 – 65 % resin, 35 –40 % vegetable gum, 6 – 10 % essential oils; *mp* ca. 100 °C.

Uses. It was previously used for plasters causing skin irritation and bleeding. It is still being used in perfumery.

4.15. Labdanum

Labdanum is a recent gum resin from the leaves of various Cistacea species (rock rose), a shrub from the Mediterranean area. The resin is a highly viscous, brown mass used in perfumes as a fixative (synthetic ambra).

4.16. Mastic

Mastic is a recent resin that is harvested from the bark of *Pistacia lentiscus*. This shrub is found in the entire Mediterranean area. The resin is greenish to yellow in color and is soluble in alcohols, aromatics, and terpenes. Its acid value is 50 – 70; the saponification value, 60 – 90; *mp*, ca. 105 °C (capillary). Mastic is used in the cosmetics industry as a fixative and in solution in turpentine as a varnish for paintings. It is also used for skin- compatible plasters and adhesives (adhesives for theater make-up).

4.17. Myrrh

Myrrh is a recent tree resin from various *Commiphora* species. The resin comes from Southern Arabia and Somalia and was known in Biblical times. Myrrh is exuded from the plant as

a milk and dries to form beads, which are collected.

Properties. Myrrh forms irregular, rounded, orange-brown colored beads or lumps with lighter areas and a bitter taste [11]. A variety of types can be found, which contain up to 10 % myrrh oil (mixture of terpenes and sesquiterpenes).

Uses. Myrrh was previously used for embalming the dead and as incense. Myrrh oils and myrrh extracts are used in cosmetics as fixatives and in medicine because of their astringent action.

4.18. Sandarac

The recent resin sandarac originates from *Callitris quadrivalvis*, a cypress species that occurs in Australia, North America, and North Africa. It has a pleasant odor, is lemon yellow in color, soluble in ethanol and diethyl ether, and insoluble in water. Sandarac has a *mp* of ca. 135 °C (capillary); an acid value of 115 – 150; and a saponification value of 145 – 155. The main component is sandaracopimaric acid ($C_{20}H_{30}O_2$). Sandarac is used as a fumigant as well as for coatings (paper and leather) and for putty and tooth cement.

4.19. Shellac [11], [12], [13]

Shellac is a recent resin of animal origin that occurs in India and Thailand. It is excreted by the shell louse (*Coccus lacca* or *Laccifer lacca*) that pricks in young shoots of certain trees. The louse forms the resin from the juice of the tree and then excretes it. The word *laksha* means 100 000, because thousands of such shell lice cover the branches of the trees (hence, the term sticklac). The harvest is carried out twice a year (April – May; November – December). Sticklac contains 60 – 80 % pure shellac and 4 – 6 % shellac wax. It is scraped off the branches and twigs and freed from dirt and other foreign matter by intense washing. The material obtained is known as bead lac (85 – 90 % pure shellac, 5 – 8 % shellac wax, 2 – 3 % impurities). By melting and sieving, either sheet shellac or button lac is obtained, depending on the way in which the melt is cooled. Natural shellac contains wax components that produce mat films. It can be dewaxed by dissolving the shellac in aqueous sodium carbonate or by selective dissolving with certain solvents. A particularly light- colored product is obtained by bleaching with hypochlorite or by use of adsorption agents.

Properties. Shellac forms bright-orange to brown, shiny, transparent sheets or beads. The different qualities are Indian (hand-made) button and sheet shellac; sheet shellac machine-processed with or without wax; bleached shellac; and special shellac (water soluble). Special shellac has been pretreated with alkali or amine, and dissolves in water without requiring further additives. Shellac has a *mp* of 65 – 77 °C; an acid value of 67 – 90; and a saponification value of 190 – 260. It is soluble in alcohols, glycol ethers, acetic acid, and aqueous alkali; and insoluble in aliphatics and aromatics. Shellac attacks iron in alcoholic solution. It must therefore be handled in iron-free equipment (e.g., stainless steel, glass, ceramics). Unbleached shellac slowly loses its solubility (over years) as a result of polycondensation. Bleached shellac does so after only ca. six months and the process is accelerated by heat. Special shellac can be stored for an unlimited period. The resin contains a hydrocarboxylic acid with five hydroxyl groups, which lead to polycondensation. Graft polymerization with acrylates and styrene provides new and interesting modifications which can be used for primers, compound foils, and coatings for metals and leather.

Uses. Shellac gives resistant, flexible films with good adhesion. It is used in pharmaceuticals, foods, printing inks, watercolor painting, insulating paints polish, shoe and leather finishing, and hairdressing (hair lacquers, setting agents). Hair preparations alone account for more than 50 % of total shellac consumption. Until about 1960 the shellac gramophone record was the most important area of use for shellac. The polycondensation of shellac, which is initiated by heat (hot hardening) or in the cold by the addition of, for example, 2 % urea, (cold hardening) is widely used in the electric, electronic, and paints and coatings industry.

4.20. Styrax (Storax)

Styrax is a recent resin that is isolated from the bark of *Liquidambar orientalis* (Hamamelidaceae family), which occurs in Asia Minor. Styrax is tapped from the living tree, the resin being formed only after the tree has been wounded. Styrax is often called benzoic resin and is very aromatic, semisolid, and yellowish brown in color. It contains large quantities of cinnamic acid and its derivatives. Its acid value is ca. 105.

Uses. Styrax is used in perfumes and medicine (skin diseases).

4.21. Utah Resin [6]

Utah resin occurs in resin druses in seams of Utah coal from which it is extracted with volatile solvents.

Properties. Utah resin is a practically neutral, natural hydrocarbon resin, which has a blackish appearance as a result of colloidal carbon. It has an acid value of 2 – 3; a saponification value of 2 – 3; a softening point of 160 – 180 °C; and an OH value of 80 – 130. Utah resin is soluble in aliphatics and aromatics, and has a good compatibility with many natural and synthetic binders.

Uses. Utah resin is used in printing inks, pigment preparations, and adhesives.

4.22. Venice Turpentine (Larch Turpentine)

The recent resin venice turpentine is tapped from the tree *Larix europaea* that occurs in Europe.

Properties. Venice turpentine is a pale-green, viscous liquid that does not crystallize. Its acid value is ca. 70; its saponification value, 130 – 145. Venice turpentine contains ca. 38 % resin acid ($C_{20}H_{30}O_2$) and a diterpene alcohol ($C_{20}H_{32}O_2$).

Uses. Venice turpentine is used in perfumes and artists' paints.

5. Rosin (Colophony)

Rosin (colophony) is industrially by far the most important natural resin because, as a recent resin, it is a renewable resource, and as a result of its chemical constitution it can be modified in many different ways to give resins of widely differing compositions. The natural sources have certainly not been exhausted, and producer countries, which have been relatively unimportant until now, will presumably work their way to market in the coming years or decades. In antiquity, rosin was a well-known and commonly used resin. The name colophony is thought to be derived from the Ionic colony Colophon in Asia Minor, where it was traded in the first century A.D. Also the resin wine (retsina) is said to have been known to the Romans.

In English speaking countries the rosins processing industry has been called "naval stores." This term dates back to the 17th century when an independent rosins industry started to develop in North America as an outcome of shipbuilding. The United Kingdom did not have enough indigenous wood for construction of the many ships required. Therefore, the ships were built in the English colonies in North America where enough pine wood was available. Pine balsam, resin, wood tar, pitch etc. were produced and stored in the naval stores. Currently the term naval stores mainly refers to rosin, rosin products, and turpentine oil.

5.1. Occurrence

Gum rosin, as used commercially and industrially nowadays, is the purified and distilled rosin from the balsam of various species of pine (up to 80 different species). Their biosphere is determined by climate, weathering, soil composition, and other botanical and meteorogical factors. The most important species and their habitat are given in what follows:

Pinus	*austriaca* (black pine)	Austria
	caribaea (slash pine)	Central America, United States, Caribbean
	densiflora	Japan
	elliottii	United States [14]
	halepensis (Aleppo pine)	Greece, Portugal, Spain
	langifolia	India
	maritima (seashore pine)	France, Spain, Portugal
	massoniana (Chinese red pine)	China [15]
	mercusii	Indonesia, Burma [16], [17], Philippines
	nigra (black pine)	Austria
	oocarpa	Central America, Honduras [18]
	palustris (swamp pine, longleaf pine)	United States
	pseudostrobus	Central America, Mexico
	sylvestris (Scots pine)	Germany, Poland
	tonkinensis	China [15]
	yunnanensis	China [15]

5.2. Collection – Extraction

Three completely different methods exist for the extraction of rosin. Each of them gives a characteristic composition of the rosin.

5.2.1. Gum Rosin [*8050-09-7*]

Here the *living tree* is systematically wounded so that if the wound is sufficiently long a pathological resin flow ensues. The tree is wounded, e.g., by a number of V-shaped groovelike incisions in a herringbone pattern. Part of the sapwood is also removed so that the resin canals are cut open. Such a wound is known as a "lachte". Irritants, such as sulfuric acid, significantly increase the resin flow from the tree. Without an irritant the wound must be recut practically every other day to achieve a continuous resin flow. Besides groove cutting, wounding of the tree can also be carried out over the surface. The crude balsam, which is exuded, is collected in a tin or earthenware vessel. Small plastic bags are also often used (see Fig. 1). Rosin harvesting begins when the pine trunks have a diameter of ca. 20 cm and can then continue for about 20 – 30 years. The yield of crude balsam depends on many factors, and varies between 2 and 5 kg per tree per year [15]. The species *Pinus elliottii*, introduced into the United States in the 1980s, is said to produce up to 10 kg of crude balsam per year [14].

The crude balsam collected is then stored in a refinery, where it is covered with water to protect it from oxidation. In processing it is melted by careful warming and dissolved in a higher-boiling turpentine oil fraction. Dirt and foreign matter are removed from this free-flowing balsam solution by sedimentation; it is then filtered. Further purification involves washing with hot water, and the dilute balsam is finally subjected to continuous steam distillation, in which water and turpentine oil are removed. The average composition is ca. 70 – 75 % rosin and 20 – 25 % turpentine oil.

5.2.2. Wood Rosin [*8050-09-7*]

The stump of the pine tree is the part that is richest in resin, so developing a process for extraction of this stump resin was considered worthwhile quite early on. In the United States many stumps were left standing in past decades as a result of the clearing of large areas of pine forest for intensive wood production, so extraction of wood rosin became particularly important there. Wood rosin production utilizes stumps that have remained in the soil for 10 years or more so that only the resin-rich heartwood is still available. These types of stump are dug out and then cut up. The chips are extracted with naphtha to obtain a relatively dark resin by extraction (Wood rosin FF). The color can be removed by washing, selective extraction from a petroleum ether solution [19] with furfural, or adsorption onto bleaching clay from a naphtha solution [20]. This dark, oxidized rosin is isolated and used for lacquer applications (Vinsol R) [21]. The naphtha from the wood extraction contains turpentine oil and other substances originating from the wood.

Pine stumps contain ca. 19 % rosin, 4 % turpentine oil, 4 % petroleum ether-insoluble resins, 23 % water, and 50 % cellulose and lignin-type substances. Since pine stumps are only formed in an economically exploitable amount in the clearing of whole forests, the quantity of wood rosin available is limited and has decreased sharply in recent years.

5.2.3. Tall-Oil Rosin [*8052-10-6*] (→ Tall Oil)

The rosin that is also present in felled trees is recovered in wood digestion (sulfate process) from the black liquor. By acidification of the black liquor it is isolated as a dark, very impure resin and is then subjected to multiple distillation steps, including steam distillation. Tall-oil rosin often contains small quantities of higher fatty acids. As a result of the production process, tall-oil rosins are free of oxocarboxylic acid and thus readily soluble in aliphatics.

5.3. Composition

Rosin consists predominantly of rosin acids belonging to the terpene group. The numbering of the carbon atoms in rosin acid molecules is shown below using abietic acid as an example.

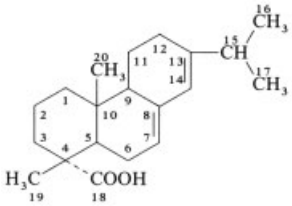

The rosin acids have the molecular formula $C_{20}H_{30}O_2$ and thus belong to the diterpenes (four isoprene units). A large number of isomeric tricyclic rosin acids exist that differ in the position of the two double bonds. The most important isomers are listed in Table 1. The acids are in equilibrium with each other. Depending on the country of origin and plant species, native rosin contains between 30 and 50 % levopimaric acid, which is the most sensitive isomer.

At > 100 to < 200 °C or upon addition of acids, levopimaric acid isomerizes to the more stable abietic acid. A shift in equilibrium in favor of levopimaric acid occurs in the Diels – Alder synthesis because levopimaric acid is removed continuously from the equilibrium by reaction with a dienophile (see Section 5.5.3). Due to the refining method, wood and tall-oil rosins are practically free of levopimaric acid. They have a high content of abietic acid, neoabietic acid, and palustric acid, and contain 15 – 30 % dehydro- and dihydroabietic acid.

5.4. Properties

Rosin has an *mp* of ca. 70 °C and is brownish to bright yellow in color. It exhibits a typical conchoidal fracture. *Gum rosin* has a slightly sweet and aromatic odor, particularly when molten. The color shade is specified by letters.

In the U.S. classification the lightest color shade is X (iodine color index ca. 12.5), followed by WW (water white, iodine color index ca. 18) and WG (window glass, iodine color index ca. 24). The dark types have the color specifications N, M, and K. The French specification recognizes still lighter types with the series AAAAAA (= 6 A) – AA (= 2 A) – AB (iodine color indices 7.5 – 12.5 – 16) and then follows the principle of the U.S. nomenclature WW – WG – N – M – K.

Typical data for rosin and its most important derivatives are given in Table 2. For gum rosin the mandatory specifications are laid down in DIN 55 935 or testing standard LY 204-74 (Peoples' Republic of China), for example. The tendency of rosin to crystallize largely depends on its purity; i.e., the fewer isomers of $C_{20}H_{30}O_2$ present, the greater is the crystallization tendency. A large number of *gum rosins* and especially *tall-oil rosins* tend to crystallize. Tall-oil rosin crystallizes especially easily because of the small number of isomers (ca. 40 % abietic acid, practically no levopimaric acid) contained and because of its content of dehydroabietic acid [22].

North American gum rosin generally has a low tendency to crystallize, which is attributed to the presence of ca. 3 % caribic acid ($C_{20}H_{30}O_3$). Honduras rosin, originating from *Pinus oocarpa*, also has a low tendency to crystallize, attributable to the high proportion of unsaponifiable substances (lower saponification value) [18]. The low crystallization tendency of *wood rosin* is due to posttreatment. The polymerized and hydrogenated types of rosin virtually do not crystallize at all.

The double bonds in the rosin acid molecules can be oxidized easily. The acids of the abietic type are prone to oxidation, but not the dextropimaric types. Oxidation results in an increase in the softening point and a decrease in the solubility in aliphatics (due to formation of oxocarboxylic acids). Modified types (i.e., rosin – maleic or rosin – phenol rosins) are less easily oxidized.

Table 1. Structure and properties of important resin acids and their contents in pine balsam, gum, and wood rosin

	Structural formula	CAS registry no.	$[\alpha_D]$*	*mp*, °C	Content in pine balsam, %	Content in gum rosin, %	Content in wood rosin, %	
Levopimaric acid	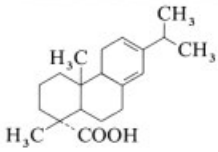		−276°	150–152	30–50	0.1–15	0.1	$C_{20}H_{30}O_2$ Upon heating to > 100°C these acids are in equilibrium with one another; abietic acid disproportionates upon heating to > 200 °C
Neoabietic acid	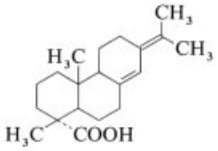		+159°	167–169	15–20	10–60	ca. 10	
Palustric acid	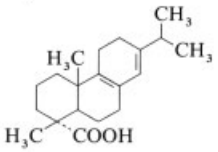		+ 72°	162–167	ca. 10	ca. 20	ca. 12	
Abietic acid	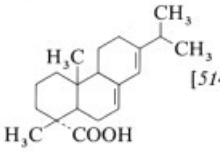	[*514-10-3*]	−106°	173–175	5–10	10–40	ca. 30	
Dehydroabietic acid	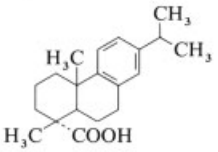		+ 63°	171–173	1	1–10	5–20	$C_{20}H_{28}O_2$
Tetrahydroabietic acid	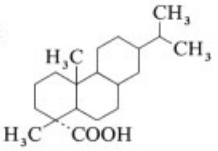				1	1	14	$C_{20}H_{34}O_2$
Dihydroabietic acid	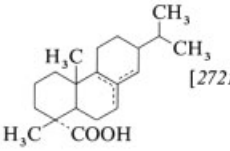	[*27216-04-2*]			4	4	14	$C_{20}H_{32}O_2$ (isomers)
Dextropimaric acid	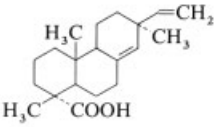		+75°	217–219	8	8	8	$C_{20}H_{30}O_2$ stable
Isodextropimaric acid	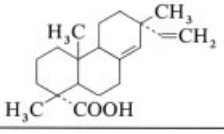		0°	162–164	8	8	8	$C_{20}H_{30}O_2$ stable

* 1% solution in ethanol.

Table 2. Physical properties of the most important rosin types and derivatives

Type	Acid number	*mp* ∗, °C	Saponification number	Unsaponifiable residue
Gum rosin	ca. 170	ca. 85/75 – 65	ca. 170	6 – 8
Tall-oil rosin	160	78/68 – 58	177	ca. 12
Wood rosin	160	73 ∗∗	168	6 – 8
Disproportionated rosin	155	70 ∗∗	160	ca. 10
Polymeric rosin (40 % dimeric acid)	154	100/90 – 80	158	7 – 9
Polymeric rosin (60 % dimeric acid)	154	115/105 – 95	158	7 – 9
Polymeric rosin (80 % dimeric acid)	142	140/130 – 120	155	7 – 9
Hydrogenated rosin	161	68 ∗∗	167	ca. 9

∗ *mp* according to Kofler, unless otherwise noted.
∗∗ Ring and ball (corresponds approximately to the highest value according to Kofler in this temperature range).

Polymerized and hydrogenated rosin types exhibit the best resistance to oxidation.

5.5. Modified Rosin Types

Because of the structure of abietic acid and its isomers, rosin can undergo a large number of reactions to form products with very different physical properties. These products are used in varying applications [23].

5.5.1. Rosinates

The production of complex rosin soaps and other salts of abietic acid is one of the oldest modifications of rosin that is still in use today.

In terms of quantity the *alkali soaps* of abietic acid form the largest proportion of industrial rosinates. They are used in the paper industry for paper sizing. Rosin is digested with a substoichiometric quantity of soda or sodium hydroxide solution. A mixture of alkali soaps and free acid is obtained (free-rosin size). The emulsion is stabilized with protective colloids (casein, animal size). This rosin size is added to the pulp stock in the Hollander.

Rosin pretreated with formaldehyde, rosin – maleic acid adducts, and for special papers, hydrogenated rosin are also used for the production of rosin soaps. Alkali soaps of disproportionated rosin are used in large quantities as emulsifiers in the rubber industry.

Alkaline-earth (Ca, Mg) and are important for the formulation of printing inks, principally for gravure and screen printing. *Calcium rosinate* is the oldest and most important rosin. The formation of calcium rosinates from polymerized rosin leads to a higher-melting lime rosin. If complex rosinates are used, rosins with melting points up to 200 °C are obtained. Even the viscosity (mostly in toluene) can be influenced over wide areas.

Calcium rosinates are currently produced primarily in a melt process. The alkaline-earth component is added as the oxide, hydroxide, or acetate in portions to the molten rosin at > 200 °C, giving rise to vigorous reaction.

Zinc rosinates are produced predominantly from polymerized rosin because, with native rosin, undesirable thickening or even gel formation can occur during production. Areas of use include printing inks, adhesives, and paints and coatings.

Composed rosinates (e.g., Ca – Zn or Zn – Mg) are often used to achieve special properties.

The *heavy-metal rosinates*, such as cobalt, lead, manganese, or copper rosinates can be produced either in a melt process or in solution. The latter generally gives purer products. An aqueous solution of the metal salt is stirred into a warm aqueous solution of sodium rosinate so that the desired metal rosinate is precipitated. Various heavy-metal rosinates are still used today in specialized areas (ceramic dyes, linoleum production, etc.).

5.5.2. Rosin Esters

The esterification of abietic acid is slower than that of fatty acids because of the tertiary acid group in the molecule. Temperatures of > 200 °C (up to 275 °C) and long reaction times are necessary for the esterification. Nevertheless

the reaction goes to practically 100 % completion. Within a reaction time of 16 – 18 h the acid value decreases to < 10, or 10 – 20, depending on the raw material used. The most common and oldest rosin ester is the glycerol ester of native rosin. If reaction conditions are properly maintained, acid values of < 10 and very good solubility in aromatics, aliphatics, esters, and ketones and good compatibility with the common binding agents and elastomers (for adhesives) are achieved. The reaction proceeds, via glyceryl diabietate as intermediate, to diglyceryl abietyl ether, which has very good properties (compatibility, solubility, water resistance). Instead of the abietyl ether, glyceryl triabietate can be formed, which falls markedly short in terms of the above properties. The relatively harsh reaction conditions (temperature – time) also give rise to the formation of dehydro- and dihydroabietic acid (disproportionation reactions), which improve the general properties.

Methanol, glycols, and polyhydric alcohols such as glycerol and pentaerythritol are commonly used as esterification agents. Esterification with methanol either requires pressure or must be carried out with dimethyl sulfate or chloromethane [24]. The methyl and glycyl esters are syrupy liquids. Besides the esters of native rosin (gum rosin, wood rosin, and tall-oil rosin), esters of polymerized – dimerized, hydrogenated, and disproportionated rosin can be found. The hardest rosin ester is the pentaester of dimerized rosin (*mp* ca. 160 °C).

Uses. Rosin esters show a very good compatibility with many elastomers, such as ethylene – vinyl acetate, styrene – isoprene – styrene copolymer (SIS), and styrene – butadiene – styrene copolymer (SBS), natural rubber, chloroprene rubber, etc. The solubility in a large number of solvents (including aliphatics) is very good. Their main area of use is currently as tackifiers in adhesive formulations (solvent-containing adhesives, hot-melt adhesives, dispersion adhesives). Soft rosins are used successfully in dispersion adhesives because they are readily emulsified. Rosin esters are used in formulations of paints and coatings and printing inks because of their good compatibility with nitrocellulose. Certain limitations in nitrocellulose compatibility can be expected with pentaesters, however. Rosin esters with high melting point can also be used as complementary rosins in coating formulations with alkyd rosins, where they improve the body, gloss, and drying time. Deodorized glycerol ester is a component of chewing gum but has to meet definite purity specifications [25].

5.5.3. Maleic Acid Adducts (Maleics)

Because of its conjugated double bonds, levopimaric acid can, as a typical diene, undergo Diels – Alder reactions. Frequently used dienophiles are maleic acid or maleic anhydride, and fumaric acid. Besides these two compounds, acrylic acid and its derivatives are also important industrially. Above 150 °C the isomers which are in equilibrium with levopimaric acid (abietic acid, palustric acid, and neoabietic acid) undergo Diels – Alder reaction. The reaction is exothermic and gives tricarboxylic acids (high acid value, higher melting point). Theoretically, ca. 25 % maleic anhydride can be added to rosin, but in practice much smaller quantities are generally used. The maleic rosins, which are frequently used industrially, are mixed esters from native rosin and Diels – Alder adducts.

All maleic rosins are characterized by their light colors, improved stability toward oxidation, lightfastness, and good pigment wetting properties. They are compatible with many binding agents.

Three different groups of maleic and fumaric acid adducts exist with differing features (see Table 3).

Unesterified adducts are used, for example, in printing inks (moisture set) and in paper size production.

Partially esterified maleic rosins are particularly important because of their high solubility in alcohol and water (after neutralization), their very good compatibility with nitrocellulose, and the broad melting-point range possible. Melting points between ca. 50 and 180 °C are possible, and can be influenced by the degree of adduct formation and the choice of special esterification components. The *areas of use* involve mainly nitrocellulose combinations and aqueous systems (e.g., paints and coatings and printing inks; for the latter, flexographic printing with alcohol and water). Ammonia or a range of amines and ami-

Table 3. Different types of maleic rosins ∗ and their basic features

Type	Acid number	*mp*, °C	Solubility	Compatibility
Unesterified	280 – 310	125 – 160	alcohol, ester	nitrocellulose
Slightly esterified	180 – 205	50 – 70	alcohol, ester, water	nitrocellulose
Slightly esterified	110 – 130	120 – 160	alcohol, ester, water	nitrocellulose
Higher esterified	60 – 100	120 – 160	alcohol, ester	nitrocellulose
Completely esterified	20 – 50	100 – 160	ester	nitrocellulose

∗ Product specifications of Lackharzwerke Robert Kraemer, Bremen, Germany.

no alcohols can be used for neutralization [26], [27].

Even nonneutralized alcoholic rosin solutions show good water tolerance (ca. 30 % based on alcohol in a solution with 30 % solid material). More highly esterified maleic rosins (acid value ≥ 60) lose their water solubility but are completely soluble in alcohol.

Completely esterified rosins are insoluble or only sparingly soluble in alcohols but are readily soluble in esters, ketones, and aromatics. The main *areas of use* are combinations with nitrocellulose (coatings, special gravure printing) but also with oils and alkyd rosins. Rosins with a high content of maleate groups are no longer oil compatible; their nitrocellulose combinations must contain suitable plasticizers (e.g., phthalates, nondrying alkyd rosins, or liquid rosin ester rosins).

5.5.4. Rosin-Modified Phenolic Rosins [28], [29] (→ Phenolic Rosins→ Phenolic Rosins, Chap. 4.3.)

Although the term rosin-modified phenolic rosins has become established, phenol-modified rosins would be more appropriate because rosin predominates in quantity and character. These rosins were also known as synthetic copals because they were able to replace natural copals on account of their technological advantages.

The most common phenols used in the modification of rosin are 4-*tert*-butylphenol and 4,4′-diphenylolpropane (bisphenol A). Through this choice and the wide variation in the ratio of phenol to formaldehyde, a large number of rosin-modified phenolic rosins are accessible. The number is increased still further by the fact that polymerized – dimerized rosins can also undergo this reaction.

The melting point of this group of rosins is in the range of 100 – 180 °C; the acid value varies between 10 and 40. Special types with higher acid values (up to 140) also exhibit good solubility in alcohols. Rosin-modified phenolic rosins can be combined with many alkyd rosins and also show good oil compatibility. They accelerate the drying of coatings, increase the resistance to water and chemicals, and improve the gloss. Rosin-modified phenolic rosins are used in alkyd coatings, printing inks, especially for gravure and offset printing.

5.5.5. Polymerized – Dimerized Rosins [30], [31]

The double bonds present in abietic acid can undergo polymerization. Strictly speaking this is a dimerization, so the expression dimerized rosin would be more appropriate. In practice, however, only rosins with more than 50 % dimeric acid are called dimerized rosins.

The polymerization – dimerization can be carried out in solution at ambient temperature or in the melt in the presence of catalysts such as sulfur, *p*-toluenesulfonic acid, boron trifluoride, zinc chloride, and concentrated sulfuric acid. Disproportionation of the molecule can occur as a side reaction. The common polymerized rosins contain ca. 40, 60, or 80 % dimeric acid. The melting point of this series increases from 70 °C (native rosin) to 100, 120, and 140 °C. All polymerized rosins have a low tendency to crystallize, a high oxidation stability, and an excellent compatibility with many other binding agents and elastomers. The acid value is not affected by polymerization so this group of rosins can undergo typical modification reactions (e.g., rosinate formation, esterification). Since polymerized rosin contains considerable proportions

of native rosin, addition reactions with maleic anhydride or phenol resol can also be carried out. Polymerized rosins and their modifications are used in formulations where enhanced stability or rosins with higher melting point are required.

5.5.6. Hydrogenated Rosins [32], [33]

The double bonds of abietic acid, which are prone to oxidation and are attacked in the presence of light, can be removed by hydrogenation. This route is considerably more expensive than the formation of dimerized rosins, but leads to still more stable rosins. Complete hydrogenation is not necessary; removal of only one of the double bonds per molecule is sufficient. Both the completely and the partially hydrogenated qualities are, however, available industrially. Hydrogenation is performed in the melt at up to 300 °C, up to 35 MPa, and in the presence of noble metal, Raney nickel, or Ni – Al alloy catalysts. The choice of suitable catalysts and the appropriate reaction conditions are particularly important, because otherwise reduction to abietyl alcohol (see Section 5.5.8) or decarboxylation can occur. Partial hydrogenation occurs much more readily and rapidly than complete hydrogenation. If the reaction is carried out properly the carboxyl group remains unchanged and can undergo esterification or rosinate formation. Hydrogenated rosins are used, for example, as particularly oxidation-resistant and lightfast tackifiers in adhesives. The alkali salts are used in the paper industry for special papers and as emulsifiers in emulsion polymerization.

5.5.7. Disproportionated Rosins [34]

Rosin acids with conjugated double bonds such as abietic acid or palustric acid can be disproportionated catalytically to dehydro- and dihydroabietic acids, and tetrahydroabietic acid. Catalysts for this reaction are Pd/charcoal, iodine, sulfur, and organic disulfides. The slightly exothermic reaction starts at ca. 250 °C. With Pd as catalyst 50 – 60 % of dehydroabietic acid and 25 – 35 % of dihydroabietic and dihydropimaric acid together with traces of tetrahydroabietic acid are formed. Decarboxylation leads to ca. 10 % of various neutral compounds.

Disproportionated rosins are highly stable toward oxidation. They are used widely in the form of their alkali soaps as emulsifiers in emulsion polymerization processes for the production of synthetic rubbers and latices. Other application areas are pigments and adhesives.

5.5.8. Rosin Alcohols

In the presence of appropriate catalysts and under harsh conditions the hydrogenation of abietic acid can result in reduction of the carboxyl group to give a rosin alcohol. Since disproportionation occurs simultaneously, a mixture of dehydro-, dihydro-, and tetrahydroabietyl alcohols ("hydroabietyl alcohol") is formed. Heavy metals and copper chromite are also used as catalysts. Industrially, hydrogenation of the methyl ester of rosin is carried out at up to 300 °C and high pressure. Rosin alcohols are light, sticky rosins with good stability toward oxidation. They are used in adhesives as tackifiers and as plasticizers in paints, coatings, adhesives, and plastics. Like many tackifying soft rosins hydroabietyl alcohol also improves the adhesion of a film (coating, printing ink, adhesive) on a base with poor adhesion. The alcohol group can undergo further reactions such as etherification or esterification.

5.5.9. Rosin Acid Nitrile, Rosin Amine

Rosin acid nitrile is formed in the reaction of rosin with ammonia at ca. 300 °C in the presence of catalysts. The main component of rosin acid nitrile is dehydroabietyl nitrile. This is a viscous oil with an acid value of ca. 10 – 15. Industrially, rosin acid nitrile is converted into the amine by hydrogenation under pressure. The resulting dehydroabietylamine is a viscous liquid with an amine content of ca. 92 %. The amino group can undergo many subsequent reactions (e.g., salt formation with inorganic or organic acids, or reaction with polyoxyethylene). Rosin acid amine and its reaction products are used to combat fungi, bacteria, and algae; as corrosion inhibitors in water systems; or as process materials in ore flotation and pharmaceuticals.

5.6. Economic Aspects [35]

Total annual production of rosin is about 1.2×10^6 t/a worldwide in 1995. Of this it is estimated that almost 720 000 t/a, or 60 %, is gum rosin, most of the remainder, about 35 %, is tall oil rosin and the rest is wood rosin.

The naval stores industry is complex and ever-changing. In the early part of the century, gum rosin production was dominant. Wood rosin production, involving the uprooting and extraction of old pine stumps, developed in those countries with large areas of raw materials, such as the United States and the former Soviet Union. Production in the United States peaked during the 1950s and has since declined to a low level. The recovery of tall oil rosin as byproduct of chemical pulping of pine chips also began during the 1950s.

As labor in the more industrialized countries has become more expensive, gum rosin production has declined in European countries. Statistical figures for 1991 to 1993 show that the focus of production of gum rosin today is Southeast Asia. Chinese production accounts for 430 000 t/a (about 60 %) of the total annual production of gum rosin and Indonesia for a further 70 000 t/a. In the last decade the production in Portugal dropped from 100 000 t/a to 20 000 t/a. Production in Russia is estimated to be at a level of some 65 000 t/a. Brazil produces about 45 000 t/a and Mexico and Argentina some 20 000 t/a. All other countries auch as Guatemala, Honduras, Venezuela, India, Vietnam, France, Spain, Greece, or Poland have only a small production mainly for domestic comsumption.

Leading producers of tall oil rosin are the United States and Canada with some 275 000 t/a, followed by the Scandinavian countries with about 50 000 t/a.

5.7. Toxicology and Food Regulations

(see also, → Resins, Synthetic, Chap. 4.)

Native rosin can be regarded as harmless and nontoxic. This also applies to many of its derivatives and modified rosins. Rosin esters are used, for example, in chewing gum, and rosin wine (retsina) is considered to be particularly beneficial. About a hundred workers, who were in daily contact with a large number of modified rosins, have been checked over a period of more than 25 years. No damaging effects on health could be established. Allergic reactions of certain people cannot, however, be ruled out. There is no evidence that allergy initiation can be attributed to any toxicological properties of rosin.

The legal licensing of modified rosins (rosinates, rosin esters, maleics, and modified phenolic rosins) in foods is governed by a large number of national regulations, which are also recognized internationally. The most important regulations are those of the FDA [36], but many regulations of the BGA are observed in other countries [25], [37]. Principally most of the aforementioned modified rosins are permitted in foods. However, certain restrictions or exceptions must be taken into consideration.

5.8. Analysis

Determination of the general and important physical data for rosins is laid down in many nationally and internationally recognized test regulations (e.g., DIN, ASTM). Special test methods for rosins include the following:

5.8.1. Melting-Point Determination

The classical method for melting-point, determination, the capillary method, is often not suitable for the rapid, routine testing of rosins, which can be soft or slightly sticky. In rosin chemistry the *ring and ball* method (DIN 53 784, ASTM D 2398) is used successfully and gives readily reproducible values. Another method in use is the *drop points*method according to Ubbelohde (DIN 51801).

5.8.2. Crystallization Test (Acetone Test)

The crystallization behavior of rosins in acetone gives an important indication of their quality and origin. Tall-oil rosins crystallize sponta-neously; gum rosins, with a few exceptions, within an hour; and polymerized and hydrogenated rosins, not at all. The rosin is coarsely ground in a mortar, and the rosin pieces are freed from dust in a sieve; 10 g of this dust-free substance is weighed

into a test tube (200×30 mm) and covered with 10 mL of acetone. The sample is allowed to stand without stirring or shaking. The time interval after which crystallization begins (on the rosin pieces or on the base of the tube) is determined [27].

5.8.3. Other Methods

All types of rosins (modified and unmodified) can be analyzed easily by IR spectroscopy [38]. Rosin and the different rosin acids can be identified by means of UV spectroscopy and capillary gas chromatography.

A basic identification of rosin is also possible by means of the *Storch–Morawski color reaction*. This and other color reactions of colophony are described in detail in [6]. Various types of rosin can also be differentiated by their optical rotation. Disproportionated and hydrogenated rosins, for example, show a positive rotation, and polymerized–dimerized rosin a negative one.

When a mixture of rosin acid and fatty acids is to be separated the rapid and quantitative formation of the methyl ester of the fatty acid can be advantageously utilized, whereas rosin, with its tertiary carboxyl group, does not react under these conditions [39], [40].

Other methods for the determination of special physical or technical data are described by rosin producers in their leaflets.

6. References

General References

1. *Ullmann,* 4th ed., **12,** 525–538. W. Sandermann: *Naturharze, Terpentinöl und Tallöl-Chemie und Technologie,* Springer Verlag, Berlin 1960. *Römpp: Chemie Lexikon,* Thieme Verlag, Stuttgart. H. Wolff-Berlin: *Die natürlichen Harze,* Wissenschaftliche Verlagsgesell., Stuttgart 1928.

Specific References

2. DIN 95946 Beuth Verlag, Berlin 1983.
3. A. W. Rick, *Farbe Lack* **58** (1952) 105–110.
4. H. Hadert, *Farbe Lack* **57** (1951) 447–450.
5. DAB 7, Deutsches Arzneibuch, 7th ed., Deutscher Apotheker Verlag, Stuttgart, 1968.
6. H. E. Scheiber: *Zerbe, Mineralöle und verwandte Produkte,* 2nd ed., Springer Verlag, Berlin 1969, p. 698.
7. J. Scheiber: *Lacke und ihre Rohstoffe,* Verlag Johann Ambrosius Barth, Leipzig 1926.
8. Liodamar and Lioresen are produced by Synthopol Chemie Dr. Koch & Co, Buxtehude.
9. *Farbe Lack* **72** (1966) 679, 805, 1023.
10. G. Kirchner (ed.): *Terra X, Rätsel alter Weltkulturen,* Umschau Verlag, Frankfurt/Main 1986, p. 85.
11. DAB 9, Deutsches Arzneibuch, 9th ed., Deutscher Apotheker-Verlag, Stuttgart 1986.
12. KPS Kalkhoff GmbH Petersen & Stroever KG, Technical leaflet, Mainz.
13. C. E. Roeper, Technical leaflet, Hamburg.
14. E. Schwenk, *Farbe Lack* **95** (1989) 832–835.
15. G. Weißmann, *Adhäsion* **4** (1985) 10–18.
16. G. Weißmann, H. M. Simatupang, *Farbe Lack* **80** (1974) 932–936.
17. G. Weißmann, *Farbe Lack* **81** (1975) 1012–1014.
18. G. Weißmann, *Farbe Lack* **84** (1978) 486–488.
19. Hercules Powder, US 2 117 572, 1936.
20. Crosby Navels Stores, US 2 181 791, 1938.
21. Product of Hercules Inc., United States.
22. W. Sandermann, G. Weißmann, *Farbe Lack* **71** (1965) 713–716.
23. E. Karsten: *Lackrohstoff-Tabellen,* Curt R. Vincentz Verlag, Hannover, 8th ed., 1987.
24. Hercules Powder US 1 924 934, 1930 (A. C. Johnson) US 1 682 280, 1928 (A. C. Johnson).
25. Verordnung über die Zulassung fremder Stoffe bei der Herstellung von Kaugummi (Kaugummi-Verordnung) BR Deutschland, Bundesgesetzblatt, Teil 1, v. 23. Sept. 1972, 1826–1829.
26. Angus Chemie, Company brochure, Ibbenbüren 1987.
27. Robert Kraemer, Technical leaflet, Bremen 1992.
28. K. Hultzsch: *Chemie der Phenolharze,* Springer Verlag, Berlin.
29. A. Greth, *Kunststoffe* **31** (1941) 345.
30. Hercules Powder, US 2 108 928, 1936 (L. Rummelsburg). US 2 124 675, 1936 (L. Rummelsburg). US 2 136 525, 1936.
31. IG-Farben-Industrie, DT 564 897, 1931 (J. Binapfl).
32. Hercules Powder, US 2 174 651, 1939 (R. J. Byrkit). US 2 155 036, 1937.
33. Du Pont, US 2 086 458, 1932.
34. Hercules Powder, US 2 177 530, 1936 (E. R. Littmann). US 2 407 248, 1946. US 2 503 238, 1950 (F. H. Gayer). US 2 503 268, 1950 (T. Hasselstrom, E. Brennan).

35. J. J. W. Coppen, G. A. Hone: *Gum Naval Stores: Turpentine and Rosin from Pine Resin*, Natural Resources Institute, Food and Agriculture Organisation of the United Nations, Rome 1995
36. Food and Drug Administration (FDA); US-Department of Agriculture, e.g., 175.105 Adhesives, § 175. 125 Pressure-sensitive Adhesives, § 175. 300 Resinous and Polymeric Coatings, § 175. 3870 Rosins and Rosin Derivatives.
37. 47., 109., 135., 162. Mitteilung, Bundesgesundheitsblatt 11 (1968) 192.
38. O. D. Hummel: *Atlas der Polymer- und Kunststoffanalyse,* 3rd ed., Hanser Verlag, München, VCH Verlagsgesellschaft, Weinheim 1991.
39. ASTM D 1585-63; ASTM D 1240-54.
40. E. E. Balder, E. P. Crowell, *J. Am. Soc.* **52** (1975) 14.

Resins, Synthetic

Alkyd Resins, Amino Resins, Epoxy Resins, Phenolic Resins are separate keywords. α- and β-Resins → Tar and Pitch

Gerd Collin, DECHEMA e. V., Frankfurt a. M., Federal Republic of Germany (Chap. 1, Sections 2.1 – 2.8)

Rolf Mildenberg, formerly Rütgerswerke AG, Duisburg, Federal Republic of Germany (Chap. 1, Sections 2.1 – 2.8)

Mechthild Zander, formerly CdF, Forbach, France (Chap. 1, Sections 2.1 – 2.8)

Hartmut Höke, Weinheim, Federal Republic of Germany (Chap. 4)

William McKillip, Advanced Resin Systems, Des Plaines, Illinois 60016, United States (Section 3.1)

Werner Freitag, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Section 3.2)

Wolfgang Imöhl, Schering AG, Berlin, Federal Republic of Germany (Section 3.3, Chap. 4 in part)

1. Definition . . . 2
2. Hydrocarbon Resins . . . 3
2.1. Classification . . . 3
2.2. Raw Materials . . . 3
2.2.1. Indene – Coumarone Feedstocks . . . 4
2.2.2. Petroleum Resin Feedstocks . . . 4
2.3. Production . . . 5
2.4. General Properties . . . 6
2.5. Characterization and Analysis . . . 6
2.6. Resin Types . . . 7
2.6.1. Aliphatic Hydrocarbon Resins . . . 7
2.6.2. Aromatic Hydrocarbon Resins . . . 8
2.6.3. Terpene Resins . . . 8
2.6.4. Resins from Pure Monomers . . . 8
2.6.5. Modified Resins . . . 9
2.7. Uses . . . 9
2.8. Economic Aspects . . . 12
3. Resins by Condensation . . . 12
3.1. Furan Resins . . . 12
3.1.1. Monomers . . . 12
3.1.2. Production – Characteristics . . . 12
3.1.3. Uses . . . 14
3.2. Ketone and Aldehyde Resins . . . 15
3.2.1. Resins from Aliphatic – Aromatic Ketones . . . 16
3.2.1.1. Acetophenone – Formaldehyde Resins . . . 16
3.2.1.2. Modified Acetophenone Resins . . . 18
3.2.2. Resins from Aliphatic Ketones . . . 18
3.2.2.1. Methyl Ethyl Ketone – Formaldehyde Resins . . . 18
3.2.2.2. Acetone – Formaldehyde Resins . . . 19
3.2.2.3. Cyclohexanone Resins . . . 19
3.2.2.4. Cyclohexanone – Formaldehyde Resins . . . 20
3.2.3. Aldehyde Resins . . . 20
3.3. Polyamide Resins . . . 21
3.3.1. Physical and Chemical Properties . . . 21
3.3.2. Raw Materials . . . 22
3.3.3. Production . . . 22
3.3.4. Quality and Analysis . . . 22
3.3.5. Storage . . . 22
3.3.6. Uses . . . 22
3.3.7. Producers, Trade Names . . . 23
4. Natural and Synthetic Resins – Toxicology and Legislation . . . 23
4.1. Toxicology . . . 23
4.1.1. Rosins and Their Derivatives . . . 23
4.1.2. Hydrocarbon Resins . . . 24
4.1.3. Polyamide Resins . . . 24
4.2. Legal Aspects . . . 25
5. References . . . 25

1. Definition

The term "resin" is a versatile expression that is not sharply defined. Originally, the term was applied to low molecular mass natural products, which were usually defined as yellowish to brown- colored, transparent to opaque, soft to brittle, easily fusible, tacky and noncrystallizable materials that are soluble in most common organic solvents and virtually insoluble in water. With the development of the chemical industry, the term resin has also been applied to synthetic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_089

materials that are used as substitutes for natural resins, or to materials with similar physical properties. In the international standards ISO 472 (Plastics – Vocabulary) and ISO 4618/3 (Paints and Varnishes – Vocabulary – Part 3: Terminology of Resins), resins are defined similarly, i.e., as solid, semisolid, or pseudosolid organic materials that have an indefinite and often high relative molecular mass, and generally soften or melt over a range of temperatures on heating. Very often, the term resin is also used to designate any polymer that is a basic material for plastics (e.g., polyethylene resin, → Plastics, General Survey). In the German standard DIN 55 958, resin ("Harz") is defined as a collective term for solid to liquid organic products. Their molecular mass distribution can vary within wide limits. Resins include a large variety of organic substances of different chemical composition, but with many similar physical properties.

Generally, resins are products with medium molecular mass. Their average molecular mass does not exceed 10 000 and, in most cases, is even lower. Resins have an amorphous and often complex structure. They do not exhibit a sharp melting point, but like every amorphous polymeric system, resins have a glass transition temperature and are characterized by their softening point, which denotes the transition temperature from a pseudosolid to a plastic state on heating.

Resins can be subdivided into natural and synthetic resins.

Natural Resins. In ISO 4618/3, natural resins are defined as resins of vegetable or animal origin. The term includes rosins (gum, wood, or tall-oil rosins from tree and plant exudates; wood extracts; or byproducts from paper manufacturing), fossil resins such as amber; mined resins such as asphaltite; shellac as secretion product from an insect; and their main derivatives.

Synthetic Resins. In ISO 4618/3, synthetic resins are defined as resins resulting from controlled chemical reactions such as polyaddition or polycondensation between well-defined reactants that do not themselves have the characteristics of resins. Synthetic resins are also obtained by polymerization (ISO 472) of unsaturated monomers. This term includes two subgroups:

1) Hydrocarbon resins, i.e., synthetic resins from coal tar, petroleum, and turpentine streams, produced by polymerization. These resins are used like natural ones, e.g., in combination with other polymers to impart special properties such as tack, flow, and hardness to a material.
2) Synthetic resins obtained mainly by addition polymerization and polycondensation, which are intermediates in the synthesis of higher molecular mass plastics.

2. Hydrocarbon Resins

2.1. Classification

The term hydrocarbon resin is generally applied to low molecular mass thermoplastic polymers that are composed mainly of carbon and hydrogen. With respect to their origin, hydrocarbon resins can be classified in four main groups:

1) Petroleum hydrocarbon resins, derived from cracked petroleum distillates
2) Indene – coumarone resins, based on coal-tar oils
3) Terpene resins from turpentine fractions
4) Resins from pure monomers

The oldest family of hydrocarbon resins is the indene – coumarone resins, discovered by Krämer and Spilker in 1890 and produced on an industrial scale since about 1910. In the mid-1930s the terpene resins were developed, which are based on byproducts of rosin extraction. Only in the 1940s did the petroleum resins become known, initiated by the rapid development of the petrochemical industry. These resins established their position as the most important group of these products. Resins from pure monomers came onto the market in the mid-1970s.

2.2. Raw Materials

The main sources of raw materials for the chemical industry are pyrolysis processes [e.g., of coal in coke ovens and, to a larger extent, of petroleum fractions in steam crackers (→ Ethylene, Chap. 5.1.)].

Pyrolysis of coal gives high-temperature coal tar as byproduct of coke production. The most important products are:

Coke	ca. 75 wt %
Coke-oven gas	13 wt %
Crude tar	3 wt %
Crude benzole	1 wt %

The first stage of coal-tar distillation gives water and the coal-tar light oil. The latter and the following carbolic oil fraction contain the aromatic unsaturated compounds that are useful in indene – coumarone resin production [4].

The pyrolysis of naphtha fractions can generate the following products and quantities, depending on the origin of the naphtha and the cracking conditions:

Hydrogen	1.0 wt %
Methane-rich gas	15.0 wt %
Ethylene	28.0 wt %
Ethane	3.5 wt %
Propene	14.0 wt %
C_4 fraction	9.0 wt %
C_5 fraction	3.0 wt %
Benzene	7.5 wt %
C_6 fraction	1.0 wt %
Toluene	5.0 wt %
C_8 aromatics	3.0 wt %
C_9 aromatics	3.5 wt %
Residual oil	5.5 wt %

The following crude feeds are used as starting materials in the production of hydrocarbon resins:

1) C_5 streams that contain linear and cyclic olefins and diolefins as resin formers are used for the production of aliphatic hydrocarbon resins
2) $C_8 - C_{10}$ streams containing styrenes, indenes, and dicyclopentadiene (DCPD) are used for the production of aromatic hydrocarbon resins

The composition of the feedstock is very important for the properties of the final resin. It affects color, degree of unsaturation, softening point, molecular mass distribution, solubility, and compatibility of the product. The crude cuts can be processed in various ways to give suitable raw materials for the manufacture of different resin types.

2.2.1. Indene – Coumarone Feedstocks

Coal-tar light oil and carbolic oil, containing aromatic resin formers such as styrene, α-methylstyrene, vinyltoluenes, indene, methylindenes, and coumarone, are treated to remove phenols and pyridine bases by extraction. Further treatment steps to improve resin color and properties are generally applied [5]. A typical composition of an indene – coumarone resin feedstock is as follows:

Styrene	2.0 wt %
α-Methylstyrene	1.0 wt %
Alkylbenzenes	30.0 wt %
Vinyltoluenes	4.0 wt %
Dicyclopentadiene	5.0 wt %
Coumarone	7.0 wt %
Indene	48.0 wt %
Methylcoumarones and methylindenes	3.0 wt %

2.2.2. Petroleum Resin Feedstocks

C_5 Streams. C_5 streams from depentanizer effluents consist of olefins and diolefins such as isoamylene, isoprene, piperylene, and cyclopentadiene, together with C_5 paraffins. A straight depentanizer C_5 stream gives only poor resin quality. The resulting resin is highly susceptible to oxidation. Lowering the cyclopentadiene content by heat soaking and subsequent distillation improves product quality. Modern C_5 resins are made from distilled concentrated piperylene streams that are obtained in isoprene recovery. Special qualities are produced from extractively distilled piperylene concentrates. Some typical compositions are given in Table 1.

The diolefin/monoolefin ratio and the nature of monoolefins present in the feedstock affect the overall course of the polymerization process and the molecular mass distribution of the final resin and are therefore carefully controlled.

C_8/C_9 Streams are obtained from a special separate downstream column. The BTX (benzene, toluene, xylenes) fraction is drawn off as the top product, and a C_8/C_9 material is obtained as a sidestream or residue, depending on the process. This material contains high-boiling unsaturated benzene homo-

Table 1. Typical composition (weight percent) of C_5 resin feeds from steam cracking

Raw materials	Straight	Distilled concentrate	Extractive distillation concentrate
Cyclo- and dicyclopentadiene	19.0	7.5	2.0
Isoprene	30.0	8.0	0.5
Piperylene	14.0	38.0	69.0
Pentenes	15.0	12.5	0.5
Pentanes	11.0	15.0	2.0
Cyclopentene	2.0	8.5	18.0
Other hydrocarbons	9.0	10.5	8.0

logues, such as styrene, α-methylstyrene, vinyltoluenes, indene, and methylindenes, together with some dicyclopentadiene that arises from the depentanizer bottoms by dimerization of cyclopentadiene. Suitable C_8/C_9 streams can be pretreated to improve resin color and properties. A typical composition is given below:

Styrene	2.0 wt %
α-Methylstyrene	4.0 wt %
Vinyltoluenes	20.0 wt %
DCPD and codimers	6.0 wt %
Indene	20.0 wt %
Methylindenes	5.0 wt %
Naphthalene	5.0 wt %
Other nonreactive aromatics	38.0 wt %

Dicyclopentadiene concentrates are generated in the heat-soaking of depentanizer effluents and subsequent distillation of C_5 fractions. Besides dicyclopentadiene, such streams may contain codimers of cyclopentadiene with other dienes, such as butadiene or isoprene, and methyl derivatives of dicyclopentadiene. Dicyclopentadiene concentration may vary from high (90 – 95 %) through medium (70 – 80 %) to relatively low. A typical composition is given below:

Dicyclopentadiene	65 – 98 wt %
Codimers	2 – 35 wt %
Benzene	1 wt %
Sulfur	150 ppm

C_4 streams contain mainly isobutenes. After removal of butadiene, a suitable feedstock for resin production may have the following composition:

1-Butene	ca. 10 wt %
cis- and *trans*-2-butene	10 – 15 wt %
Isobutene	15 – 30 wt % or more
n-butane	ca. 30 wt %
Isobutane	ca. 30 wt %
C_{5+} components	balance

Pure Commercial Monomers. The pure commercial monomers styrene, α-methylstyrene, and vinyltoluenes are used individually, as blends, or together with terpenes or other selected unsaturated aliphatic compounds in resin manufacture. Styrene is produced in large quantities by alkylation of benzene and subsequent dehydrogenation. Its main use is in the manufacture of polystyrene. α-Methylstyrene is a byproduct of the large-scale production of phenol. Vinyltoluenes are produced in smaller quantities by alkylation of toluene, followed by dehydrogenation. Commercial vinyltoluene is a mixture of 64.5 % *m*- and 35.5 % *p*-vinyltoluene.

Terpene Monomers. The raw material for the production of terpene resins is based on natural terpene monomers that are obtained from various sources such as naval stores (gum and wood turpentine) and kraft paper production (tall-oil turpentine). Citrus terpenes, obtained by extraction from orange peel in the production of orange juice, are also used as raw materials. The most important terpenes for resin manufacture are α-pinene, β-pinene, and dipentene (*dl*-limonene). Typical compositions of terpene oils are listed in Table 2.

Table 2. Typical compositions of terpene oils (in wt %)

	Gum turpentine	Wood turpentine	Tall-oil turpentine
α-Pinene	58 – 65	ca. 80	60 – 70
β-Pinene	23 – 35	0 – 2	20 – 25
Camphene		4 – 8	
Other terpenes	5 – 8	15 – 20	6 – 12

2.3. Production

Hydrocarbon resins are produced by polymerization of the unsaturated compounds contained in the fractions described in Section 2.2 by using mainly Friedel – Crafts catalysts such as aluminum chloride and its complexes, as well as boron trifluoride and its complexes. Brønsted acids (e.g., sulfuric acid) may be also used. The term "catalyst" is scientifically incorrect with respect to cationic polymerization and should be replaced by the term initiatior – coinitiatior (initiating system) [6]. The reaction can be performed batchwise or continuously and includes the following steps:

1) Feed preparation (distillation, pretreatment, blending)
2) Polymerization
3) Deactivation of iniating system
4) Resin separation

The most important reaction parameters are feed composition, total and relative concentration of resin formers in the feed, type of catalyst and cocatalyst, catalyst concentration, and temperature. These parameters must be carefully controlled. They affect the reaction widely, in terms of yield, type (polymerization, alkylation), molecular mass, and molecular mass distribution. The particular resin feed is contacted with the initiating system by vigorous stirring. The reaction is exothermic and must be strictly temperature controlled. After reaction, the initiating system is inactivated (e.g., by addition of water, alkali, lime, or ammonia). Then the unreacted feed and the low molecular mass oligomers are removed by distillation, usually in combination with a steam-stripping process. The residual resin is conditioned in flakes, bulk, or other forms.

The proportion of catalyst in the reactant medium influences molecular mass and composition of the resin. High concentrations give resins with lower molecular mass, and lower concentrations give resins with higher molecular mass. A low concentration of catalyst may influence resin composition by selectivity (i.e., only the most reactive components are transformed). Reaction temperature affects product molecular mass, yield, and color. Molecular mass decreases and yield increases with increasing temperature.

Dicyclopentadiene resins are usually manufactured by thermal polymerization of various dicyclopentadiene feedstreams that may also contain styrene or indene.

2.4. General Properties

Depending on raw material and polymerization conditions hydrocarbon resins may range from viscous liquids to very hard solids with light to amber coloration. Special water-white types also exist. Solubility and compatibility characteristics are a function of the special resin type. All unmodified types have in common an apolar structure that gives them good water repellency and high resistance to acid, alkali, and salt solutions. They are soluble in aromatic and most common aliphatic solvents, ethers, esters, ketones, and chlorinated hydrocarbons. Unmodified resins are compatible with a wide range of other resins and polymers. The absence or presence of double bonds in the resin structure determines whether the resin is stable and/or reactive. Thus, both resins with excellent heat and weather stability and highly reactive grades are commercially available. The structure of hydrocarbon resins is very complex. The major principles involved are the following:

1) Mixtures of mono- and diolefin monomers result in paraffinic and naphthenic structures; diolefins (DCPD type) give unsaturated resins

2) Vinyl aromatic monomers (styrenes, indenes) generate chain structures with high aromatic substitution and a low degree of unsaturation
3) Bicyclic terpenes react mainly to produce cyclic structures by rearrangement

2.5. Characterization and Analysis

Hydrocarbon resins are usually characterized by their softening point, color, degree of unsaturation, saponification number, density, and ash content. These properties are determined by standardized methods (see Table 3).

Table 3. Characterization of hydrocarbon resins

Property	Method (unit)	Standard
Softening point	ring and ball RB (°C)	ASTM D 3461-76, ISO 4625
	Krämer – Sarnow KS (°C)	DIN 53 180, NFT 67 001
	Mettler	ASTM D 3461-76
Color	Gardner (50 % solution in toluene)	ASTM D 1544-80 (0-18), ISO 4630
	Barrett (8 % solution in toluene)	
	Saybolt, Hazen, APHA (for very light colors)	
Unsaturation	bromine number (g Br_2/100 g)	ASTM D 1159-84, DIN 51 774, part 1
	Wijs iodine number (g I_2/100 g)	ASTM D 1959-69
Acid number	mg KOH/g	DIN 51 558, DIN 53 402, ASTM D 974-80
Saponification number	mg KOH/g	DIN 51 559, DIN 53 401
Density	at 20 °C (g/cm^3)	DIN 51 757
Flash point	Cleveland open cup (°C)	ISO 2592

Because of their complex structure, hydrocarbon resins do not exhibit a definite melting point but soften on heating within a temperature interval. The *softening point* is defined as the temperature at which a resin, poured in a ring of defined diameter and thickness, flows under the pressure of a given weight. The most widely applied method is the ring-and-ball (RB) method; the older Krämer – Sarnow (KS) method is used more rarely. A more recent method is the Mettler method.

Color is determined optically in solution by comparison with a color standard on the Gardner scale or less commonly on the Barrett scale. For light colors the Saybolt, Hazen, or APHA method is used. A comparison of Barrett and Gardner color numbers is given in what follows:

Barrett	Gardner
B 1/2	1 – 6
B 1	7 – 10
B 11/2	11 – 12
B 2	12 – 13
B 21/2	14
B 3	15
B 4	16 – 17
B 5	18

Unsaturation is indicated by the iodine or bromine number. Neither number gives reliable information on the true degree of unsaturation, because along with addition reactions to the double bonds, substitution reactions may occur.

Other properties determining the end use of resins are solubility, compatibility, melt and solution viscosity, molecular mass, and molecular mass distribution. *Solubility and compatibility* are measured by determining the cloud point. Although generally accepted, this method is not standardized for resins. *Molecular mass and its distribution* are important resin characteristics and are determined by size exclusion (gel permeation) chromatography. Some information on resin structure is obtained by spectroscopic methods such as IR, UV, and NMR. The relation of aromatic to aliphatic protons evaluated by NMR is a useful tool for determining the aromaticity of a resin.

2.6. Resin Types

2.6.1. Aliphatic Hydrocarbon Resins

Aliphatic hydrocarbon resins are based on C_5 feedstreams and can be divided into general-purpose or basic resins, high-quality special piperylene resins, and dicyclopentadiene resins. Typical data for aliphatic petroleum resins are listed in Table 4.

Table 4. Typical data for aliphatic petroleum resins

Property	C_5 Olefin – diolefin resins	Piperylene resins
Softening point RB, °C	70 – 100	liquid – 115
Gardner color	5 – 7	3 – 7
Bromine number, g Br_2/100 g	15	25
Density at 20 °C, g/cm^3	0.93 – 0.97	0.88 – 0.97

C_5 Olefin – Diolefin Resins. Basic resins are produced from more or less narrow C_5 cuts and are available in large quantities as multipurpose resins. They have satisfactory color and color stability and a moderate degree of unsaturation. They are highly soluble in aliphatic solvents and compatible with a wide range of other resins, polymers, and waxes. Their softening point is usually between 70 and 100 °C (RB).

Piperylene resins exhibit an improved quality and are produced from highly concentrated streams that contain up to 70 % piperylene. They have properties similar to those of terpene resins because of their cycloaliphatic structure. Thus they are often called "synthetic terpene resins." They have excellent light stability and light color. A wide range of piperylene resins is produced – from viscous, liquid resins to hard resins.

Dicyclopentadiene Resins. Dicyclopentadiene resins can be divided in two main groups:

1) Those that contain predominantly dicyclopentadiene (DCPD), with high degrees of unsaturation (bromine number ca. 65 – 100 or higher) and which are distinctly heat- and oleoreactive
2) Those containing substantial amounts of aromatic structures (formed by co-reaction with unsaturated aromatics), with moderate unsaturation (bromine number ca. 55 – 57)

Because of the presence of reactive unsaturation these DCPD resins may be modified with other unsaturated components such as maleic anhydride, drying oils, or rosin (see Table 5).

Table 5. Typical data for dicyclopentadiene resins

Property	Straight DCDP resins	Aromatic-modified DCPD resins
Softening point (RB), °C	liquid – 160	10 – 150
Gardner color	6 – 11	5 – 15
Bromine number, g Br_2/100 g	95 – 100	ca. 55
Density at 20 °C, g/cm^3	ca. 1.1	1.05 – 1.12

Polybutenes, manufactured from butadiene-free C_4 fractions, are low molecular mass copolymers of the butene isomers. These polybutenes are normally low-colored, viscous liquids that are chemically stable, oxidation resistant, nondrying, and tacky (→ Polyolefins, Chap. 4.)

2.6.2. Aromatic Hydrocarbon Resins

Aromatic hydrocarbon resins are based on coal-tar-derived indene – coumarone streams or petroleum-derived C_8/C_9 streams. Typical properties of aromatic hydrocarbon resins are listed in Table 6.

Table 6. Typical properties of aromatic resins

Property	Coal/tar based	Petroleum based
Softening point RB, °C	liquid – 170	liquid – 150
Gardner color	6 – 18	5 – 12
Iodine number, g I_2/100 g	35 – 105	25 – 90
Density at 20 °C, g/cm^3	1.08 – 1015	1.06 – 1.11

Indene – Coumarone Resins [4], [5]**.** The oldest family of hydrocarbon resins is now available with an acceptably light color. Indene – coumarone (IC) resins are on the market as viscous liquids and as solids with high softening points, up to 170 °C (RB). The resins are soluble in most common solvents (except low molecular mass alcohols) and compatible with a wide range of polymeric materials. They exhibit high water repellency and resistance to acid and alkali. When not stabilized, they are subject to yellowing by autoxidation under the influence of UV light.

Petroleum Resins. The C_9 aromatic hydrocarbon resins are obtained from streams of cracked petroleum fractions that are purified to a greater or lesser extent. Depending on its origin and pretreatment, the raw material contains a varying percentage of dicyclopentadiene. The degree of unsaturation of the resin can thus be controlled by the choice of raw material. The properties of petroleum resins are generally similar to those of indene – coumarone resins, but petroleum resins are often characterized by a lower degree of unsaturation, higher light and heat stability, and a wider range of solubility and compatibility. They are produced in many different grades from viscous liquids to solids of high softening point.

Table 7. Typical data for terpene resins

Property	Derived from		
	α-Pinene	β-Pinene	*d*-Limonene
Softening point RB, °C	25 – 135	10 – 135	100 – 135
Gardner color	3 – 5	1 – 4	1 – 4
Bromine number, g Br_2/100 g	25 – 30	15 – 30	27 – 28
Density at 20 °C, g/cm^3	0.90 – 0.98	0.93 – 0.99	0.99

Table 8. Typical properties of resins from pure aromatic monomers

Property	Styrene	α-Methylstyrene	α-Methylstyrene – vinyltoluenes
Softening point RB, °C	liquid – 160	70 – 145	75 – 120
Gardner color	1 – 3	<1	<1
Bromine number, g Br_2/100 g	0	2 – 8	2
Density at 20 °C, g/cm^3	0.98 – 1.08	1.06 – 1.07	1.05

2.6.3. Terpene Resins

Terpene resins are based on naturally occurring monomers. The presence of tertiary carbon in their structure makes them very susceptible to oxidation; thus they must be protected by antioxidants.

Terpene resins are usually light colored and soluble in most common solvents. They are widely compatible with other resins and polymers, and have excellent tackifying properties. They are available as viscous liquids and in solid form. Typical properties are presented in Table 7.

2.6.4. Resins from Pure Monomers

Pure monomers such as styrenes or isobutene are excellent resin formers for the manufacture of very stable, light-colored to colorless (water-white) resins. Industrially, resins based on styrene or α-methylstyrene, or blends of mainly α-methylstyrene and vinyltoluenes, are produced. Resins from the copolymerization of terpenes and styrene are also available. Typical data of resins from pure aromatic monomers are given in Table 8.

Table 9. Typical properties of mixed resins

Property	$C_5 - C_9$ Resins	Terpene – aromatic resins
Softening point RB, °C	liquid – 105	105 – 115
Gardner color	2 – 10	2 – 5
Iodine number, g I_2/100 g	<25	not available
Density, g/cm^3	0.86 – 1.07	1.02

2.6.5. Modified Resins

In the previous sections, resins are described that are obtained from steam cracker fractions or monomer mixtures. They contain predominantly carbon and hydrogen, but no polar or functional groups. With the exception of dicyclopentadiene resins, they are chemically nonreactive. This section deals with resins that are either obtained from special, modified feedstreams or are modified after polymerization. Thus, properties of aliphatic- or aromatic-type resins may be adapted by blending initial feedstreams or by addition of selected monomers. Mixed resins of $C_5 - C_9$ types, as well as terpene – aromatic resins, are available. Typical properties are given in Table 9.

Addition of dicyclopentadiene to the initial feedstream increases the degree of unsaturation and thus the reactivity of the final resin. Also coreaction of dienes (e.g., dicyclopentadiene with maleic anhydride) may be applied. Resins made by coreaction of aromatic or terpene resin formers with phenol are well known. Their properties are listed in Table 10.

Table 10. Phenol-modified resins

Property	Based on	
	C_9 and indene – coumarone	Terpene
Softening point RB, °C	liquid – 120	95 – 155
Gardner color	5 – 8	4 – 11
Hydroxyl number, mg OH/g	56 – 112	80 – 260
Density at 20 °C, g/cm^3	1.08 – 1.1	1.03 – 1.09

Table 11. Typical properties of hydrogenated resins

Property	Based on			
	DCPD	C_9 Fraction	Aromatic monomer	Terpene
Softening point (RB), °C	85 – 140	70 – 140	liquid – 125	85 – 115
Gardner color	< 1	> 1	crystal clear	< 1
Bromine number, g Br_2/100 g	2 – 3	not available	not available	10 – 20
Density at 20 °C, g/cm^3	1.10	0.98 – 0.99	0.92 – 1.04	0.94 – 1.01

The high reactivity of terpenes makes them suitable for copolymerization reactions with other monomers, e.g., styrenes or maleic anhydride.

Resins may also be modified by further reaction of the final resin; for example, dicyclopentadiene resins can react with maleic anhydride, unsaturated fatty acids, or rosin upon heat treatment. A very important modification is the catalytic hydrogenation of resins, which gives high-quality water-white resins with a different solubility and compatibility profile than the original resin. Dicyclopentadiene resins are hydrogenated in large quantities. Hydrogenated aromatic resins derived from steam cracker streams or pure monomers and hydrogenated terpene resins are also available. Typical properties are listed in Table 11.

2.7. Uses

Hydrocarbon resins are used in numerous fields of application because of their versatile properties. They are seldom used alone but mainly in conjunction with other materials to adapt their properties to special requirements. Hydrocarbon resins are mostly used as tackifiers, binders, processing aids, film formers, reinforcing agents, and extenders.

With regard to their specific properties, the following classification is often used:

- Hard resins (binders, film-formers) in the paint, varnish, printing and coatings industries
- Tackifiers in the adhesives, sealants, and rubber industry
- Processing aids (plasticizers) in the plastics industry (rubber goods, tires, PVC)
- Fixing agents, e.g., in wood protection
- Water repellents in various applications (e.g., wood protection, coatings)

The most important applications are in adhesives (construction, solvent- or water-based, hot-melt and pressure-sensitive adhesives), coatings (paints, varnishes, roadmaking), the rubber and plastics industry (mainly mechanical rubber goods and tires, also PVC), printing inks, and, to a lesser extent, in various other industrial fields. A generalized view of the main application areas by resin type is given in Table 12. The large group of modified resins is used in applications where special properties are required.

Table 12. Main application areas by resin type

Resin type	Adhesives	Coatings	Paints	Printing inks	Rubber
Aliphatic C_5 *	x	x	x		x
Aromatic C_9 **	x	x	x	x	x
Mixed aliphatic/aromatic	x	x	x		x
Dicyclo-pentadiene		x	x	x	x
Modified DCPD				x	
Butene	x	x			
Pure monomer	x	x	x		
Hydrogenated	x	x			
Terpene	x				

* Petroleum resins; ** Petroleum and indene – coumarone resins.

Aliphatic resins are used extensively in hot-melt adhesives; in pressure-sensitive adhesives based on natural or thermoplastic rubber; in coatings, road-marking formulations, mastics, caulks, and sealants; and in rubber formulations for mechanical goods and tires. *Aromatic C_9* and *indene – coumarone* resins are widely applied in hot-melt formulations based on ethylene – vinyl acetate or thermoplastic rubber; in solvent-based and waterborne adhesive formulations; in the rubber industry (tires and mechanical goods); and in paints, varnishes, coatings, printing inks, sealants, wood protection, and concrete curing. *Terpene resins* are used almost exclusively in

the adhesive industry as excellent tackifiers for pressure-sensitive adhesives and hot-melt formulations. *Dicyclopentadiene resins* are generally applied in paints, varnishes, and coatings where they are introduced in the binder system, and in the printing ink industry. *Water-white resins* are used in adhesives and coating formulations where light color and color stability are important. The liquid polybutylenes are used in sealants, caulks, mastics, and pressure-sensitive formulations. The large group of modified resins is used in applications where special properties are required.

Adhesives [7]**.** Most synthetic adhesives are based on a "backbone" polymer, tackifying resins, and a carrier (solvent, water, or heat) that gives the formulation the necessary consistency for application. Tackifying resins regulate the final rheological properties of the system (tack, specific adhesion, and wetting).

Solvent-borne adhesives, mainly based on SBR and polychloroprene rubber, are frequently combined with hydrocarbon resins (aromatic C_9 and indene – coumarone , aliphatic and mixed C_5/C_9), as well as rosin, rosin esters, terpene – phenolic, and *p-tert*-butylphenolic resins.

Waterborne adhesives are increasingly being used for environmental reasons. Polymer dispersions based on poly(vinyl acetates), latexes, or acrylates may be modified by addition of tackifying resins such as petroleum and indene – coumarone resins, terpene resins, as well as rosin esters, terpene – phenolic resins, and ketone resins.

Hot-melt adhesives are thermoplastic solids which are liquified by heating for application. Backbone polymers for these adhesives are ethylene – vinyl acetate (EVA), thermoplastic rubbers, such as styrene – isoprene – styrene block copolymers, polyethylene, and polypropylene, formulated together with tackifying resins. Polyesters and polyamides, which are also used in hot-melt adhesives, do not need additional tackification. Hot-melt adhesives formulated with the above-mentioned polymers contain paraffins or waxes, tackifying resins, and fillers. Tackifying resins of various types may be used, mainly hydrocarbon resins (aliphatic C_5, aromatic C_9, indene – coumarone, mixed aliphatic – aromatic, terpene and hydrogenated resins), rosin and rosin esters, and terpene – phenolic resins.

Pressure-sensitive adhesives are defined as adhesives which in solvent-free form remain permanently tacky and adhere instantaneously with application of a minimum of pressure. Their main use is in tape and label manufacturing. Formulations are based on backbone polymers such as rubbers (natural rubber, SBR), and block copolymers, together with tackifying resins, or on acrylic resins. Tackifying resins for this use are aliphatic C_5, aromatic C_9, mixed C_5/C_9, pure monomer, terpene, and hydrogenated resins, as well as rosin and rosin esters.

Sealants. Sealants are based on polymers such as butyl rubber, polyisobutene, thermoplastic elastomers, SBR, neoprene, and polysulfides. Addition of tackifying resins is necessary to improve instantaneous adhesion. Practically all hydrocarbon resin types as well as rosin and modified rosin are usable. Other basic raw materials for sealants, such as silicones or urethanes, do not need additional tackification.

Coatings. Coatings [5], [8] range from protective to decorative and can be intended for temporary or permanent use. Hydrocarbon resins are extensively used, as illustrated by the following examples:

- Hot-melt coatings are mainly used in the packaging industry. Formulations and tackifiers used are very similar to those for hot-melt adhesives.
- Paints and varnishes, based on a variety of binder systems, may use rosin derivatives as well as hydrocarbon resins of the aromatic, aliphatic, and DCPD type as hard resins. Aromatic hydrocarbon resins are especially important in the formulation of aluminum paints.
- Road-marking material of the paint or hot-melt type uses mainly aliphatic C_5 resins but also C_9 types, modified by straight or cyclic aliphatics, as well as hydrogenated and special C_9 resins.
- Anticorrosion paints may be based on C_5, C_9 and DCPD resins, mixed C_5/C_9 resins, and phenol-modified aromatic resins.
- Concrete curing membranes, applied to concrete to avoid too rapid drying, contain hy-

drocarbon resins (DCPD, C_5 and C_9) in solution or dispersion.
– Wood protection by impregnation or surface treatment is carried out with, e.g., solutions containing aromatic C_9 resins because of their excellent water-repellent and fixing properties.

Printing Inks. Use of hydrocarbon resins in ink formulation takes advantage of their good compatibility with many ink components, e.g., alkyds, oils, and ink solvents, and gives the ink gloss and good resistance to water and alkali. Hydrocarbon resins are usually formulated together with the classical ink binders such as alkyd resins and modified rosins. High-softening-point hydrocarbon resins of the aromatic C_9 and indene – coumarone type, unmodified and modified, as well as DCPD and functional modified DCPD resins are used. The main use is in offset and gravure inks.

Tires and Mechanical Rubber Goods. Rubber stocks are compounded with a multitude of different additives: processing aids, plasticizers, reinforcing agents, tackifiers, pigments, antioxidants, and curing agents. Resins widely used in this application are generally referred to as processing aids, but in fact they have multiple functions. They promote processing by lowering viscosity, improve the building tack of unvulcanized mixtures, and may improve elongation, tensile strength and flex resistance of the vulcanized materials. Besides alkyl-phenolic resins, the largest volume resins for this application, considerable quantities of hydrocarbon resins, aromatic C_9 and indene – coumarone resins, C_5 aliphatic, mixed aromatic/aliphatic, and DCPD resins, as well as some unmodified rosin are used. The main criteria for choice are availability, consistency, and price.

Other Uses. Hydrocarbon resins are used in many other applications, such as in compounding floor tiles, in the foundry industry as binder and lustrous carbon former, as oil additives, in investment casting, the ceramic industry, in textile treatment and waterproofing, as well as in gypsum waterproofing.

2.8. Economic Aspects [9]

The total capacity for synthetic hydrocarbon resin production is estimated to be about 1×10^6 t/a worldwide in 1991. Main capacities are located in the United States, Canada, and Brazil (ca. 0.45×10^6 t/a); Europe (roughly 0.25×10^6 t/a); and the Asia/Pacific region (about 0.3×10^6 t/a).

Aromatic resins represent a volume of ca. 44 %, C_5 and DCPD resins ca. 37 %, and hydrogenated resins ca. 19 %. Producers and trade names of hydrocarbon resins are listed in Table 13.

3. Resins by Condensation

3.1. Furan Resins

The term "furan resins" describes generic polymers or resins that are derived from the condensation of furfural and/or furfuryl alcohol alone, and with a variety of coreactants. These include formaldehyde or a formaldehyde precursor such as paraformaldehyde, urea – formaldehyde concentrates or syrups, phenol, phenol – formaldehyde resins, aldehydes, ketones, and polyhydroxyl-substituted aromatics. Polymers or resins derived from 2,5-bis(hydroxymethyl)furan (BHMF) have been reported. With 2,5-BHMF, both condensation and addition polymers can be synthesized.

3.1.1. Monomers

Although many chemicals possess a furan nucleus in their molecular structure, furfural and furfuryl alcohol are the only monomers of major industrial significance. *Furfural* [*98-01-1*] is produced commercially from agricultural byproducts such as corncobs, sugarcane bagasse, oat hulls, and wood residues [10] (→ Furan and Derivatives, Chap. 3.). *Furfuryl alcohol* [*98-00-0*] is produced by hydrogenating furfural in a vapor- or liquid-phase process with a variety of catalysts based primarily on copper [11]. *2,5-Bis(hydroxymethyl)furan* [*1883-75-6*] is synthesized by selective hydroxymethylation of furfuryl alcohol [12] or by hydrogenation of the less readily available 5-hydroxymethylfurfural [13].

Table 13. Resin manufacturers and trade names

	Trade names	Resin types
Europe		
Akzo (Netherlands)	Setalin	hydrocarbon-modified phenolic resins
Cray Valley (France)	Norsolene	C_9 and C_5-C_9 resins
DRT (France)	Dercolyte, Dertophene	terpene and modified terpene resins
DSM (Netherlands)	Uratak	C_9 and phenol-modified C_9 resins
Enichem (Italy)	on stream in 1994	C_5 resins
Exxon (France)	Escorez	C_5 and C_5-C_9 resins, DCPD and hydrogenated DCPD resins
Hercules (Netherlands)	Hercures, Regalite Kristalex, Hercoflex	C_5 and C_9 resins, C_5-C_9 resins, resins from pure monomers, hydrogenated resins
Hoechst (Germany)	Karboresin	modified hydrocarbon resins
NEVCIN (Netherlands)	Necires, Nevchem, Nevex, Nevbrite, Nevroz, Nevillac	C_9 and C_5-C_9 resins, DCPD and modified DCPD resins, phenol-modified resins
Rütgers–VFT (Germany)	Novarez	C_9 and modified C_9 resins, IC and phenol-modified IC resins
United States		
Amoco	Amoco resin 18	poly(α-methylstyrene)
Arizona	Zonarez, Zonatac, Newtac, Nirez, Sta-Tac, Betaprene, Beta-Tac	terpene and modified terpene resins, C_5 and DCPD resins
Chemfax	Stygene, Chemprene, Pentaprene, 5 AM	C_9 and IC resins, DCPD resins
Eastman	Eastotac	C_5-C_9 resins, hydrogenated resins
Exxon	Escorez	C_5, C_5-C_9, and C_9 resins, hydrogenated resins
Goodyear	Wingtack	C_5 and C_5-C_9 resins
Hercules	Adtac, Piccodiene, Piccopale, Piccotac, Piccovar, Hercotac, Picco, Regalrez, Piccolastick, Kristalex, Piccotex, Piccolyte, Piccofyn, Endex	C_5, C_9, and DCPD resins, terpene and modified terpene resins, resins from pure monomers, hydrogenated resins
Lawter	Petro-Rez, Alpha-Rez	C_9 and modified C_9 resins, DCPD and modified DCPD resins
Neville	LX, Paradene, Nevillac, Nevchem, Nevoxy, Cumar, Nevex, Nevpene, Nevroz, Nebony, Nevtac	IC resins, C_5, C_9, C_5-C_9, and DCPD resins and modified resins
Polysat	Polysat LC	modified C_9 resins, DCPD and modified DCPD resins
Resinall	Resinall	C_9, modified C_9, and modified DCPD resins
Sartomer	Norsolene	C_9 and modified C_9 resins
Schenectady	SP	terpene phenolic resins
Sunbelt	Sunresin	C_5, C_9, and modified C_9 resins, modified DCPD resins
Japan		
Arakawa	Arkon	hydrogenated resin
Idemitsu	Admaru	styrene resins, DCPD resins, hydrogenated resins
Maruzen	Marukarez	C_5 and DCPD resins, hydrogenated C_5 and DCPD resins
Mitsui	Hi-rez, Petrosin, FTR	C_5 and C_9 resins, resins from pure monomers
Nippon Petrochemical	Nisseki Neopolymer	C_9 resins, liquid resins
Nippon Steel	Nittetsu	IC resins
Nippon Zeon	Quintone	C_5 and modified C_5 resins, DCPD and modified DCPD resins
Toho	Hiresin, Coporex	C_5-C_9 resins
Tonen	Escorez	C_5, C_9, C_5-C_9, and DCPD resins, hydrogenated DCPD resins
Tosoh	Petcoal, Petrotack, Corplex	C_9 and C_5-C_9 resins
Yasuhara	YS, Polystar, Clearon	terpene, modified and hydrogenated terpene resins
Other countries		
Canada St. Lawrence	Salar, Arotac, Coumin	C_9 and IC resins, DCPD resins
Brazil Petroquimica	Unilene	C_9 resins
Korea Kolon Korea	Kolon	C_5, C_9, and IC resins

Table 13. (Continued)

	Trade names	Resin types
Taiwan Yung Tung	Sunrez	C_9 resins
India Indian Petrochemical Corp.	Petrez	C_9 resins

IC = indene – coumarone

3.1.2. Production – Characteristics

Furfural and furfuryl alcohol are converted to resins by *condensation polymerization*.

Furfural-Based Resins. Homopolymerization of furfural, although studied extensively [14], has no industrial importance. Copolymers of furfural with phenol, phenol – formaldehyde resins, and other hydroxyl-substituted aromatics have been available commercially for many years. Since the acid- catalyzed reaction of furfural and phenol is difficult to control, most industrial processes involve the use of alkaline catalysts such as alkali hydroxides or carbonates. Furfural and ketones (usually acetone) condense with alkaline catalysts to form furfurylidene acetone monomer, which polymerizes to a thermoset resin in the presence of strong acid.

Furfuryl Alcohol-Based Resins. The mechanism proposed up to the early 1950s for the polymerization of furfuryl alcohol was reported by Dunlop and Peters in their treatise on furans [10]. As a result of the growing importance of furfuryl alcohol resins since the 1950s, the chemistry of furfuryl alcohol polymerization initiated by heat, acid, and alumina has been studied extensively and has proved to be complex. Studies dealing with the identification of intermediates and byproducts by spectroscopic and chromatographic techniques [15–18], the reaction kinetics involved in polymerization [19–21], the structure of the resulting polymers [22], and the nature of the oxygen- or acid- catalyzed cross-linking of the initial resin have all led to a better understanding of this complex chemistry [23]. A comprehensive review examines the literature up to 1977 and interprets the resinification mechanisms [14]. A more recent review further clarifies the two-stage polymerization of furfuryl alcohol [12]. Based on accumulated data, furfuryl alcohol must be considered a bifunctional monomer in the initial stages of weak acid catalysis, and its "normal" reactions give linear chains or oligomers containing essentially two repeating units (**1**, **2**), with **2** predominating:

1

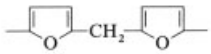

2

The polymerization or resinification of furfuryl alcohol is highly exothermic and, depending on the activity and concentration of the catalyst used, requires careful control of reaction temperature. Generally, weak acids such as carboxylic acids or metallic salts are used to produce oligomeric prepolymers with repeating moieties **1** or **2**. Further conversion of these prepolymers to cross-linked thermoset resins is accomplished by the addition of strong inorganic or organic acids. Aromatic sulfonic acids are the preferred catalyst. Earlier investigators postulated cross-linking through the C-3 and C-4 positions of the furan ring. However, more recent studies [24], suggested, on the basis of analytical data, that the main cause of branching and cross-linking is a condensation reaction between methylene groups within a chain and a hydroxymethyl group at the end of another chain to yield structures such as **3**.

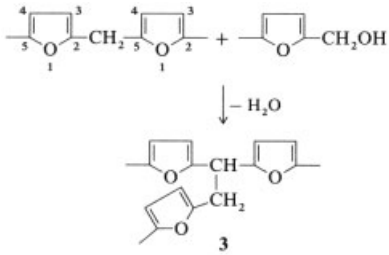

3

2,5-Bis(hydroxymethyl)furan. The 2,5-BHMF monomer behaves as a diol and undergoes esterification (base-catalyzed transesterification), alkoxylation, glycidyl ether formation, cyanoethylation, and urethane formation. In the presence of an acid catalyst, resinification of 2,5-BHMF occurs, similar to that of furfuryl alcohol, although the end groups of the resulting resin are hydroxymethyl. These resins are characterized by enhanced reactivity compared to acid-catalyzed furfuryl alcohol resins [25]. 2,5-Bis(hydromethyl)furan resins convert to highly cross-linked thermosetting polymers that demonstrate unique characteristics in composites, including a desirable high char yield on carbonization, low flame propagation, corrosion resistance, and superior high-temperature stability.

3.1.3. Uses

Furan resins are used in a variety of applications including sand cores and molds for metal casting; corrosion-resistant fiberglass-reinforced plastics; low flammability and low smoke generating composites and foams; carbonaceous products; polymer concretes; wood adhesives; and proppants in oil-well sand consolidation.

The use of *furfuryl alcohol resins* in the worldwide foundry industry remains the only important application. Total world production of furan resins for use as binders to produce sand cores and molds for metal castings exceeds 75 000 t/a.

3.2. Ketone and Aldehyde Resins

Ketone and aldehyde resins are obtained from aliphatic aldehydes and aliphatic or aliphatic–aromatic ketones by self-condensation, or more frequently by cocondensation with formaldehyde [26], [27]. Other monomers such as phenols or urea are also sometimes incorporated. Although these resins have been known for a long time, new types of resin and new areas of application are always being found. In paints and coatings, as in many other areas, these resins are used in combination with other binding agents, plasticizers, pigments, and process materials [28–40].

They are also used in printer and copier technology [41–43], the production of printed circuit boards [44], adhesives [45–48], binding agents for corrugated cardboard [49–51], molding sands [52], [53], and laminates [54–56]. The production of aqueous dispersions has also been described [57].

Reaction Mechanism. The condensation of aliphatic or aliphatic–aromatic ketones with formaldehyde is base catalyzed [58]. Formaldehyde adds first to the ketone, with formation of the aldol:

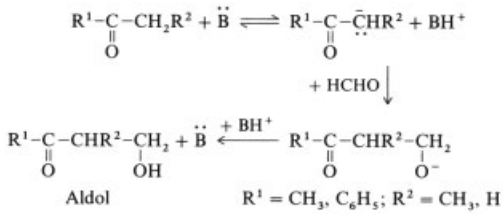

The aldol is then either dehydrated to give vinyl ketones, which can polymerize or reacted with excess formaldehyde to form highly reactive methylol compounds. The reaction pathway followed depends on the molar ratio of ketone to formaldehyde and on reaction conditions (pH, temperature, concentration).

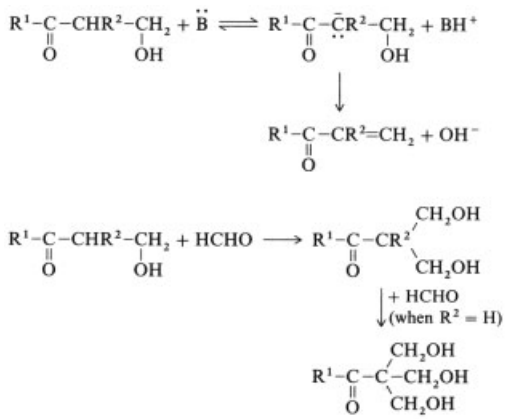

The resins are formed by polymerization of the vinyl ketones to give polymers with an alternating structure (see Fig. 1) or by complex condensation of the methylol compounds with one

another or with other ketone molecules to give partially branched products. In industrial processes, both mechanisms often operate in parallel. An excess of formaldehyde can lead to reduction of carbonyl groups to hydroxyl groups [59]. The tendency of the vinyl ketone intermediate to polymerize is rather independent of the nature of the ketone used [60], [61], provided the vinyl group is not substituted.

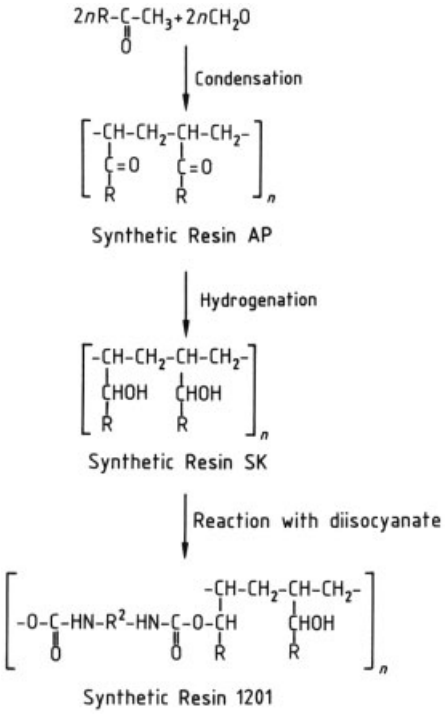

Figure 1. Resins from acetophenone and formaldehyde R = Phenyl. R^2 = Aliphatic hydrocarbon.

3.2.1. Resins from Aliphatic – Aromatic Ketones

Industrially mainly acetophenone is reacted with formaldehyde. Derivatives of acetophenone with alkylated benzene rings, such as methyl-, ethyl-, *tert*-butyl-, or cyclohexylacetophenone can also be employed [62]. Besides the unmodified resins, their hydrogenation products and the resins derived from them are available commercially (see Fig. 1).

3.2.1.1. Acetophenone – Formaldehyde Resins

Acetophenone – formaldehyde resins are light-colored nonhydrolyzable products, with sufficient lightfastness and softening points of 75 – 85 °C. They are readily soluble in aromatic hydrocarbons, esters, ketones, and glycol ethers, but insoluble in alcohols, aliphatic hydrocarbons, and mineral oils. Other physical and technical properties are listed in Table 14. The resins are compatible with nitrocellulose; copolymers of vinyl chloride; chlorinated rubber; maleic, urea, melamine, and ketone resins; plasticizers; and a few alkyd resins. They are incompatible with the majority of alkyd resins, drying oils, polyesters, and polyacrylates. Their relatively low price, gloss-improving properties, and capacity to increase the solid material content of coatings (i.e., the coating can be laden with a higher solids content and the viscosity remains constant when ketone resins are added) should be mentioned in particular.

Production. *Light, soluble, hard resins* are obtained by heating an aqueous mixture of acetophenone, formaldehyde, and potassium or sodium hydroxide to the boiling point. Solvents such as alcohols can also be used. Strong acids, such as sulfuric acid, or dehydrating salts, such as zinc chloride, are also suitable condensing agents but are less common. Quaternary nitrogen or phosphorus bases [63] are used as condensing agent to obtain particularly high-melting resins. Under suitable conditions, sodium sulfite leads to low molecular mass, water-soluble condensates [64]. Formaldehyde is added mostly as a 30 % aqueous solution or as paraformaldehyde.

Larger quantities of alkali hydroxide and higher temperature result in resins that are infusible and not readily soluble, because the methylol intermediates cross-link with one another. Readily soluble, fusible condensates are obtained with ca. 0.2 mol of hydroxide per mole ketone and a sufficiently short reaction time, so that the molecular mass cannot rise above 400 – 800. Generally, increasing the reaction temperature and time results in a higher softening point, while the solubility and compatibility deteriorate. This is probably due to competing reaction mechanisms during condensation. If conditions are such that vinyl phenyl

Table 14. Physical and technical properties of ketone and aldehyde resins

	Synthetic Resin (Hüls)				Laropal (BASF)				
	AP	SK	1201	CA	K 80	A 81	A 101	A LR 8768	A LR 8825
Resin type	ketone aromatic	ketone hydrogenated	ketone modified	ketone aliphatic	ketone aliphatic	aldehyde	aldehyde	aldehyde	aldehyde
Softening point, °C, DIN 53 180	76 – 82				75 – 85	80 – 90	95 – 110	ca. 80	90 – 100
Softening point, °C, DIN 53 181		110 – 120	155 – 170	92 – 108					
Iodine color index (50 % solution) DIN 6162	1 – 2.5	0 – 1	0 – 1	max. 2	max. 2	max. 3	max. 5	max. 5	max. 4
Acid value, mg KOH/g, DIN 53 402	max. 0.1	max. 0.1	max. 0.1	max. 0.3	max. 1	max. 3	max. 3	max. 3	max. 3
OH value, DIN 53 240		ca. 325	ca. 210	ca. 325		40 – 60			
Solubility in									
Alcohols	±	+	+	+	±	+	+	+	+
Esters	+	+	+	+	±	+	+	+	+
Ketones	+	+	+	+	±	+	+	+	+
Aromatics	+	–	±	+	+	+	+	+	+
Aliphatics	–	–	–	–	+	±	–	+	±
Compatibility with nitrocellulose	+	+	+	+	+	+	+	+	+
Chlorinated rubber	+	–	–	+	+	+	+	+	+
Vinyl chloride polymers	+	–	–	+	±	+	+	+	+
Polyacrylates	–	±	±	±	±	±	±	±	±
Urea resins	+	+	+	+	±	+	+	+	+
Melamine resins	+	+	+	+	±	+	+	+	+
Alkyd resins	±	±	±	±	+	+	+	+	+

+ = compatible or soluble; ± = some products compatible or only miscible in certain ratios; – = incompatible or insoluble.

ketone is produced mainly as the intermediate, resins with low softening point and solubility in a broad range of solvents are formed. For example, reaction of 1 mol of acetophenone with 0.05 – 0.5 mol of formaldehyde gives an oil-compatible resin. If methylol intermediates are predominantly formed, the resins have higher melting points and are limited in their solubility and compatibility. Resins with these properties are obtained from acetophenone and formaldehyde in the ratio 1: 1.

Uses. Acetophenone – formaldehyde resins are used as favorably priced binding agents in coatings and printing inks, in combination with other coating raw materials such as nitrocellulose. They improve gloss, intensity, covering capacity, solid material content, adhesion, and drying. In nitrocellulose lacquers the ease of pigmenting and polishing on wood is increased. Acetophenone – formaldehyde resins are also used in vinyl chloride copolymer coatings for metal coating and road-marking paints. Adhesives and hot melts are also formulated with these resins.

Trade Name. Synthetic Resin AP (Hüls).

3.2.1.2. Modified Acetophenone Resins

Hydrogenation of acetophenone – formaldehyde resins, in particular, gives products that differ markedly from the usual ketone and aldehyde resins and therefore open up new areas of application. The reaction of hydroxyl groups, which are formed from the keto groups on hydrogenation, with aliphatic diisocyanates gives high-quality special binding agents (see Fig. 1 and Table 14).

Properties. The hydrogenated resins are very light- colored and particularly stable. They are completely free of formaldehyde, which can no longer be formed by retroaldol-type reactions. Depending on hydrogenation conditions, resins are obtained in which only the keto groups are converted into hydroxyl groups or in which the aromatic ring is also hydrogenated [65]. Only the former are available on the market. Their softening points lie in the range 110 – 120 °C and they are readily soluble in alcohols. The resins can be diluted with aromatics but are insoluble in aliphatics and mineral oil. Their compatibility with alkyd resins is better than that of nonhydrogenated products.

The hydroxyl values of these resins are > 300 mg KOH/g. The OH groups can be reacted with diisocyanates to give extremely high-melting resins (*mp* ca. 160 °C), which are also alcohol soluble. Diisocyanate-modified resins are even more stable and exhibit rapid solvent release.

Production. Under mild hydrogenation conditions only the keto groups are hydrogenated (Fig. 1). The reaction is controlled by determination of the oxygen content and the melting point [66]. Hydrogenation can be carried out catalytically in solution or in the melt [67], [68]. Reactions with diisocyanate take place at slightly elevated temperature with the usual catalysts.

Uses. Modified acetophenone – formaldehyde resins are high melting and stable. They are therefore used as hard resin components together with other binding agents in paints and coatings, printing inks, and adhesives, if high-quality coatings are required. The *hydrogenated resins* are used for lightfast nitrocellulose-based overprinting inks for paper. They are also used in ballpoint pen pastes, inks, and toners; as flow-control agents for powder coatings; as hard resin components in isocyanate reactive systems; and for modification of alkyd resins.

The *isocyanate-modified resins* are used in high-quality printing inks, in paper and wood varnishes, and in the formulation of coatings with good adhesion to plastic surfaces. They are particularly suitable for wetting and stabilizing aluminum pigments in paints and printing inks.

Trade Names. Synthetic Resin SK (hydrogenated resin) and Synthetic Resin 1201 (diisocyanate-modified, Hüls).

3.2.2. Resins from Aliphatic Ketones

Resins are produced from linear aliphatic ketones – principally methyl ethyl ketone and acetone – and formaldehyde. Methyl isobutyl ketone is also used, mainly for the production of adhesives [69], [70]. Cyclohexanone and

methylcyclohexanone are most important in the synthesis of alicyclic resins, although reactions of cyclopentanone, cycloheptanone, and cyclic ketones with longer side chains have also been described [71].

3.2.2.1. Methyl Ethyl Ketone – Formaldehyde Resins

Methyl ethyl ketone is still important for the production of binding agents for paints, coatings, and adhesives. These resins differ from alicyclic and aromatic ketone-based products in certain application-relevant properties, such as solubility in various solvents and compatibility with other lacquer raw materials. These properties may be explained by the high polarity of the ketone and its special behavior in the alkali-catalyzed condensation with formaldehyde [59].

Physical Properties. Methyl ethyl ketone – formaldehyde resins are lightly colored and are soluble in polar organic solvents such as alcohols, esters, ketones, and glycol ethers. The hydroxyl values are in the range 80 – 190 mg KOH/g. The strongly polar resins have a total oxygen content of 21 – 29 wt % and are hygroscopic. Depending on production conditions, the molecular mass lies between 3000 and 5000, and the melting range between 80 and 120 °C. Other physical data and compatibilities are described in [27].

Chemical Properties. Methyl ethyl ketone – formaldehyde resins are unhydrolyzable. The keto and hydroxyl groups and the carbon double bonds give the compounds a certain degree of chemical reactivity. Hydrogenation leads to clear, colorless resins [72]. The hydroxyl groups can be esterified with carboxylic acids or acid anhydrides [73] and can react with isocyanates [74].

Production. Industrially, ketone resins are obtained by alkali-catalyzed condensation of methyl ethyl ketone and formaldehyde (molar ratio 1: 2 – 1: 2.5). Condensation is performed mainly batchwise [59], [75], [76]. Continuous production has also been described [77]. The unpurified ketone is treated with formaldehyde in the presence of water and sodium or potassium hydroxide as catalyst. By increasing the formaldehyde content, the melting range can be raised to 80 – 120 °C.

Special purification processes give light, high-melting products [78]. High softening points are also achieved by using phase-transfer catalysts [72]. The reaction mechanism is not completely understood and involves many concurrent reactions [27].

Uses. Methyl ethyl ketone – formaldehyde resins are used in combination with film-forming agents such as nitrocellulose, acetylcellulose, cellulose ethers, and natural resins. They impart hardness, drying, polishing ability, and good lightfastness. The capability to gelatinize nitrocellulose is exploited in the production of high-gloss nitro lacquers. The free hydroxyl groups are used for cross-linking in isocyanate-reactive coatings, adhesives [79], and molding sands [74].

3.2.2.2. Acetone – Formaldehyde Resins

The alkali-catalyzed condensation of acetone and formaldehyde does not lead to useful, solid synthetic resins. The considerably greater hardening tendency of the unstable methylol intermediates from acetone in comparison to those of other carbonyl compounds [80], [81] leads to a cross-linked, insoluble final product. *Self-hardening precondensates* can, however, be produced, which are used alone or in combination with other hardenable precondensates, such as phenol resols.

Uses. Hard foams [82] are obtained by foaming the methylol compounds in the presence of alkali hydroxide alone or of alkali hydroxide and an elastomer latex [83]. Sands can be hardened with acetone – formaldehyde precondensates for the production of moldings [84–86]. Rapidly hardening moldings made from cement have also been described [87], [88]. Acetone – formaldehyde condensates are used widely in paper [89] and wood adhesives [90]. Phenol resols are often cocondensed to obtain particularly weather-resistant wood materials and chipboard [91–93] or for the waterresistant adhesion of corrugated paperboard [94]. Other uses of acetone

precondensates are in photoreceptors in electrophotography [95], [96] and in production of thickeners and emulsifiers [97], [98].

3.2.2.3. Cyclohexanone Resins

Cyclohexanone and methylcyclohexanone readily undergo self-condensation, in addition to mixed condensation with formaldehyde [99]. The former involves aldol condensation between the carbonyl group of one molecule and the activated methylene group of another. The carbonyl group of the intermediate then reacts with another cyclohexanone molecule, etc. The reaction occurs at elevated temperature and can be catalyzed by basic, acidic, or neutral agents.

Properties. The softening points of the light, neutral resins are between 80 and 120 °C. Under normal conditions the resins are resistant to acids and bases. However, they eliminate water in the presence of acids at elevated temperature (> 80 °C), thus altering their properties significantly. Cyclohexanone resins are lightfast, soluble in many solvents, and compatible with most coating raw materials (see Table 14). They are more expensive than cyclohexanone – formaldehyde resins [99]. An example of the production process (according to [100], [101]) is given in [27].

Uses. Cyclohexanone resins are used in coatings, mainly to improve intensity, gloss, and hardness. They can also improve light and weather resistance, as well as adhesion. In quantities of 5 – 50 % (based on the film-forming agent) these resins are added to coatings based on alkyd resins, vinyl chloride copolymers, chlorinated rubber, nitrocellulose, or oils. Their use as carrier resins for pigment preparations has also been described [102].

Trade Name. Laropal K 80 (BASF).

3.2.2.4. Cyclohexanone – Formaldehyde Resins

Cyclohexanone can react with aldehydes, in particular formaldehyde, to give defined methylol compounds or resinous products. The molar ratio and the reaction conditions determine the final products. A large formaldehyde excess promotes the formation of methylol compounds, whereas basic catalysis leads to resin formation [103], [104]. Higher aldehydes can also be used instead of formaldehyde, but they have not gained any industrial importance. In addition to cyclohexanone, methylcyclohexanone or mixtures of cyclohexanone with aliphatic ketones are also used as the component with the activated methylene group [105]. The resins can be modified with phenols [100], [106], epoxides [107], polyesters [108], and sulfonamides [109]. Gritty products are obtained by the addition of dispersion agents [110]. Continuous production has been described in [111] and [112]. The lightfastness can be increased by hydrogenation [113] and treatment with reducing agents [114].

Properties. Cyclohexanone – formaldehyde resins do not exhibit the broad compatibility and solubility of cyclohexanone resins. However, they are cheaper and nevertheless sufficiently lightfast. They can be combined with a range of binders for coatings, but not with oils (see Table 14). The use of methylcyclohexanone as a raw material generally leads to better solubility and compatibility.

Production. Industrially, cyclohexanone is condensed with formaldehyde in the presence of alkali. An example of the production procedure is given in [27], where the modification with phenol is also described.

Uses. In many cases, cyclohexanone – formaldehyde resins are used to improve drying, hardness, intensity, gloss, and solid material content. In paints and coatings they are always used as additives in addition to other binding agents, for example, in alkyd – acrylate varnishes [115], cement paints [116], epoxy resin systems [117], and marine paints [118]. In addition to their use in conventional printing inks [119], [120], they are increasingly being used in UV-hardening printing inks [121–123]. Another important area of application is in adhesives [124–126] and sealants [127]. The use of these resins in optical recording media has also been described [128], [129].

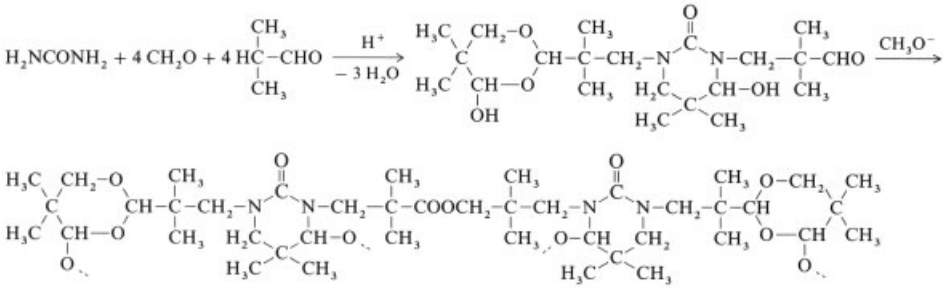

Trade Names. Synthetic Resin AFS (Bayer), Synthetic Resin CA (Hüls), L2 Resin (Leuna), Krumbhaar type resins (Lawter).

3.2.3. Aldehyde Resins

In the 1970s, aldehyde resins synthesized from isobutyraldehyde, formaldehyde, and urea were developed to an industrial level. α-Ureidoalkylation formed the basis of the synthesis and is described in detail in [130]. Aldehyde resins from isobutyraldehyde and phenol are used in melt adhesives [131], [132].

Isobutyraldehyde – Formaldehyde – Urea Resins. Both the reaction of urea with enolizable aldehydes [133] and the reaction of urea, formaldehyde, and enolizable aldehydes [134] in an acidic medium lead, on subsequent treatment with alkali alkoxides, to resinous products. The process can be described as follows (see above).

Properties. The isobutyraldehyde – formaldehyde – urea resins are slightly colored and fairly lightfast. Their melting points lie in the range 80 – 110 °C. The solubility and compatibility with other coating binders can be varied, depending on the production process. Thus products soluble in aliphatics and all other common solvents are on the market. Their compatibility is, however, limited to acrylate resins and some cellulose derivatives (see Table 14). The available functional groups can be seen in the above formula scheme.

Production [134]. Formaldehyde, isobutyraldehyde, and urea are reacted in the presence of sulfuric acid. The condensation is carried out for 3 h at 80 °C. After the addition of xylene the mixture is stirred for another 20 min at 75 °C, and the aqueous phase is separated. The organic phase is treated first with sodium methoxide and then neutralized with 75 % sulfuric acid. The resin resulting after work-up has a pale-yellow color and a softening point of 80 – 82 °C.

Uses. Like ketone resins, aldehyde resins can be added to other coating binders. Improvements in hardness, gloss, intensity, flow, resistance to yellowing, and solid material content are achieved in coatings based on alkyd resins, nitrocellulose, or chlorinated polymers. By the addition of aldehyde resins, the overall formulation can often be made more economical (e.g., in powder coatings, where the flow is improved at the same time). Another area of use is in hot meltings and spray plastics for road markings. Some aldehyde resins are suitable as mixing resins for pigment pastes with broad solubility and compatibility. A high pigment-binding capacity and low solution viscosity have thus far not been achieved with ketone resins. For aqueous dispersions see [135].

Trade Names. Laropal A types (BASF).

3.3. Polyamide Resins

Polyamides are thermoplastic polymers with repeating amide units (−CO−NH−) in the main chain (→ Polyamides).

Polyamide resins (PA resins) also belong to the polyamide group. However, PA resins are based on dimerized fatty acids, and their molecular mass is generally less than 15 000; thus they differ considerably from high molecular mass polyamides in their chemical structure and

properties [136]. Whereas high molecular mass polyamides are important synthetic fibers and engineering plastics, polyamide resins are used as hotmelt adhesives, casting compounds, and binders for printing inks.

3.3.1. Physical and Chemical Properties

The physical and chemical properties of PA resins can be varied within wide limits through the choice and ratio of raw materials used. Brittle, tough-hard, or flexible PA resins can be produced, which have very different tensile strengths and elongation values [137].

The density of PA resins is ca. 0.95 g/cm^3. The *softening point* is in the range of 80 – 250 °C, preferably between 90 and 190 °C. Melt viscosities are between 0.1 and 100 Pa · s at 220 °C. The water absorption of PA resins is between 0.5 and 3 % after 28-d storage in water and is thus much lower than the water absorption of industrially relevant polyamides. All PA resins show good resistance to chemicals. They are generally resistant to water, dilute alkali, weak acid, oils, and fats. Most PA resins are moderately stable to unstable toward esters, ketones, alcohols, and aliphatic hydrocarbons. Some are resistant to chlorinated hydrocarbons.

3.3.2. Raw Materials

The dimerized fatty acids required for the production of PA resins are obtained by heating mono- or polyunsaturated fatty acids, such as oleic, linoleic, or linolenic acids or their esters with or without a catalyst. The dimerized fatty acids formed have an open- chained, cyclic-unsaturated, or aromatic structure [138].

In addition to dimerized fatty acids, dicarboxylic acids are frequently used as comonomer. Examples are adipic acid, azelaic acid, sebacic acid, decamethylenedicarboxylic acid, terephthalic acid or its ester, or isophthalic acid. Ethylenediamine, hexamethylenediamine, 1,12-diamino-4,9-dioxadodecane, piperazine, dipiperidylpropane, and polyoxypropylenediamine are used as the polyamine component [139]. 6-Aminocaproic acid in the form of its lactam and 11-aminoundecanoic acid are the preferred aminocarboxylic acids.

3.3.3. Production

Polycondensation of the aforementioned monomers is carried out in a reactor equipped with a stirrer and condenser at 150 to 250 °C with exclusion of air (Fig. 2). The reaction is performed initially under normal pressure and then under reduced pressure. The water formed in the reaction is distilled off. When the condensation reaction has gone to completion the product is removed as a melt, cooled, and ground or poured into special molds.

3.3.4. Quality and Analysis

For quality testing and characterization of the products the following procedures are used:

The *softening range* is measured by the ring-and-ball method (DIN 52 011 or ASTM E-28). Melting points determined by differential scanning calorimetry or thermomechanical analysis, or microscopic melting points are sometimes given.

Melt viscosities are measured with rotation viscometers. The values obtained can vary considerably, depending on the type of viscometer used. For highly viscous products the *melt index*according to DIN 53 735 or ASTM D-1238 is sometimes determined. The *tensile strength* and *elongation* are determined according to DIN 53 455 or ASTM D-1708. The *water absorption* is determined according to DIN 53 495 or ASTM D-570. The *acid value* is determined according to DIN 53 402 and the *amine value* according to DIN 16 945.

3.3.5. Storage

Polyamide resins are usually supplied as pellets. Because many machines for further processing can only handle rods, cylinders, or rectangular blocks, PA resins are also supplied in these forms.

Since problems can arise in processing if the product is damp, moisture-proof packaging is recommended. With proper storage, the shelf-life of PA resins is virtually unlimited.

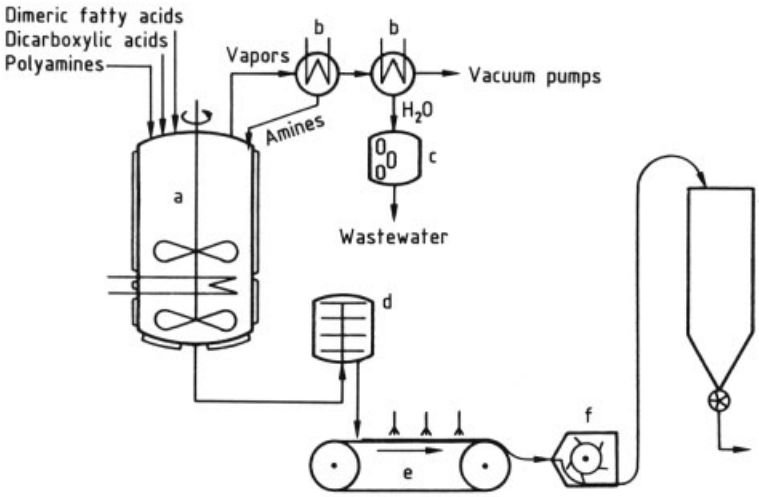

Figure 2. Polyamide resin production
a) Reaction vessel; b) Condenser; c) Separator; d) Filter; e) Cooling band; f) Granulator

3.3.6. Uses

Thermoplastic PA resins are used as high-performance hotmelt adhesives [140], casting compounds [141], binders for printing inks, overprinting lacquers [142], and thixotropic agents for alkyd resin coatings [143].

In the shoe industry, PA hotmelt adhesives have been used for many years as an aid in assembly for bonding the upper leather to the insole (lasting). The furniture industry utilizes PA hotmelt adhesives for heat-resistant edge banding of chip or particle board.

In the automobile industry, PA hotmelt adhesives are used for interior trim. The bonding of air and oil filters by using PA hotmelt adhesives is particularly important. A relatively new area of application is the sealing of wires and cables into car- connectors by potting with PA resins. Connectors are being incorporated more frequently in the car because of the increased use of electronic control systems. In the electrical and electronics industries, PA resins are used because of their relatively high heat resistance and good insulating properties. In addition to bonding radio and television housings and loudspeakers, in particular coil winding, coil ends, and other electronic components are fixed and bonded with PA hotmelt adhesives.

In the telecommunications sector, PA hotmelt adhesives are used to bond and seal cables and heat-shrink components, and to protect telephone cables from moisture and corrosion at the ends or connection points. In the textile industry, PA resins are used for "cleaning-resistant" bonding of interlining to textiles [144].

PA resins are also used as binders for printing inks. Printing inks based on PA resins are particularly suitable for printing on nonporous materials, such as plastic films, because of their good adhesion, flexibility, and gloss. Since the printed films show good resistance to chemicals and abrasion, they are also used as overprint lacquers.

In the area of transparent lacquers, which are applied by gravure or flexographic printing, PA resins are used as cold-seal release lacquers, antislip varnishes, or heat-seal varnishes.

3.3.7. Producers, Trade Names

The PA resins are sold under the following trade names: Euremelt, Eurelon (Schering, Germany); Versamid, Macromelt (Henkel, Germany); Reamide (Chemplast, Italy); Unirez (Union Camp, United States); Versamid, Versalon (Cray Valley Products, United Kingdom); Polymid (Lawter, United States); Evacor (Evode, United Kingdom); Sunkem (Sun Chemical, United States); Terlan (Lambiotte, France); Casamid (Swan, United Kingdom); Emerez

(Emery, United States); Sunmide (Sanwa, Japan); Polymide (Sanyo, Japan).

4. Natural and Synthetic Resins – Toxicology and Legislation

4.1. Toxicology

4.1.1. Rosins and Their Derivatives

For review, see [145].

Crude Rosins (→ Resins, Natural, Chap. 5.2., → Resins, Natural, Chap. 5.3., → Resins, Natural, Chap. 5.4., → Tall Oil). In contrast to refined rosins, crude rosins contain volatile, neutral components (→ Terpenes), which may impose properties on them that resemble those of organic solvents.

Resin Acids. *Abietic acid*, the major component of *refined rosins*, has an acute oral LD_{50} in mice of about 2500 mg/kg. The acute oral LD_{50} of its *methyl ester* is > 5000 mg/kg in rats and rabbits. Resin acids were not mutagenic in the Ames test.

Refined rosins, original or chemically modified, have a negligible acute oral toxicity in rodents and guinea pigs, the LD_{50} values ranging from 4000 to 12 000 mg/kg; no acute dermal toxicity was found in rabbits for pale rosins.

On oral administration, *dark rosins* with a high phenolic fraction and *rosin esters* showed no acute toxic symptoms.

In longterm feeding studies with pale and esterified rosins on rats and dogs (90 days or two years), no toxic symptoms were observed at the 0.2 % to 1 % dietary level. Therefore rosins and their derivatives may be regarded as safe.

Generally, rosin acids, rosins, and their derivatives are not considered skin irritants. However, a small portion of the population shows an allergic response on patch tests. A high sensitization level was reported among printers exposed to gum arabic, a constituent of printing ink [146].

Fumes from rosins may cause ocular and bronchial irritation: ca. 80 % of humans exposed to products of wood rosin at a level > 0.12 mg/m^3 experienced severe irritation of the eye, nose, or throat. The TLV for rosin core solder pyrolysis products is 0.1 mg/m^3 aliphatic aldehydes measured as formaldehyde.

4.1.2. Hydrocarbon Resins

The acute oral toxicity (LD_{50}) of hydrocarbon resins in rats is generally greater than 5 g/kg body weight, and they were found to be nonirritating to the skin and eye under experimental conditions [147]. In the EEC directives, most hydrocarbon resins are not classified as harmful, toxic, or irritant to the skin and eyes. Handling of resins requires the usual precautions for working with chemicals. Inhalation of dust or fumes (the latter generated on heating or combustion) must be avoided. More specific information is given in the manufacturers safety data sheets.

Food legislation concerning hydrocarbon resins is regulated by many national authorities but will likely be harmonized with the requirements of the FDA and the EU. The main attention is directed to migration of residual monomers, oligomers, and additives that might contaminate contiguous media by diffusion, and to the stability of resin formulations under application conditions.

In the United States, numerous resins are CFR approved substances for the following:

175.105: Adhesives
175.300: Resinous and polymeric coatings
176.170: Components of paper and paperboard in contact with aqueous and fatty food
176.180: Components of paper and paperboard in contact with dry food
177.1210: Closures with sealing gaskets for food containers
177.2600: Rubber articles intended for repeated use
178.3800: Preservatives for wood.

In the EU, the use of materials intended – in their final state – to come in contact with products for human consumption is regulated by the framework directive 89/109/EEC, followed by several subordinate directives, one of which concerns the use of plastics materials, e.g., in food packaging (90/128/EEC). Implementation by the EU member states was required by January 1, 1993.

Monomers and additives are not allowed to be used for food packaging unless listed in the annexes of the latter directive (positive-list principle). Not all hydrocarbon resins have yet been covered there. Therefore, an industrial motion to enclose those resins in the additive list, among others of the IC type, was directed to the European Scientific Committee for Food, but has not yet been settled.

4.1.3. Polyamide Resins

Polyamide resins do not need to be labeled [EC guideline 79/831/EEC (classification, packaging, and labeling of hazardous substances), GefStoffV]. Some PA resins are licensed by the FDA for contact with foods.

If PA resins are used in the molten state, vapors can sometimes be formed, which do not impair general health or have any toxic effects, according to the current state of knowledge. Nevertheless, extraction of the vapors above processing equipment and adequate ventilation of the workplace are recommended.

4.2. Legal Aspects

Food legislation concerning resins is regulated by many national authorities, but will likely be harmonized with the requirements of the FDA (United States) and the European Community. The main questions are directed to the migration of residual monomers, oligomers, and additives that might contaminate contiguous media by diffusion, and to the stability of resinous formulations under the conditions applied (→ Foods, 4. Food Packaging, Chap. 8.1., → Foods, 4. Food Packaging, Chap. 8.2.).

In the *European Community,* the application of resins in the food sector will be regulated in connection with materials intended to come in contact with foods [148], [149] and with use of food additives other than colors and sweeteners [150]. Several directives that are especially relevant [151–153] have already been adopted.

In the United States, the use of resinous materials in contact with food is regulated in the Code of Federal Regulations (CFR) under Title 21 (Food and Drugs): Besides the use of resins, other additives such as stabilizers, fillers, and linker reagents are also addressed (e.g., Parts 172, 175, 176).

Natural gums, synthetic as well as natural terpene resins,indene coumarone,and petroleum hydrocarbon resins may be used as chewing gum bases with limitations (121.1059, 121.90). Terpene resins are permitted for coatings in direct use (e.g., coating of fruits) when the precursors originate from wood (172.280). In indirect applications, most resins (synthetically or naturally derived) may be used.

Petroleum hydrocarbons are authorized for indirect use as components of paper and paperboard (Part 176).

5. References

General References

1. W. Vredenburgh, K. F. Foley, A. N. Scarlatti: "Hydrocarbon Resins", in *Encycl. Polym. Sci. Eng.,* vol. 7, John Wiley, New York 1987, pp. 758 – 783.
2. E. W. Flick: *Industrial Synthetic Resins Handbook,* Noyes Publication, Park Ridge, N.J., 1989, pp. 214 –384.
3. R. Mildenberg, M. Zander, G. Collin: *Hydrocarbon Resins – Manufacture, Properties and Applications,* VCH, Weinheim 1997.
4. H.-G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Heidelberg 1968, pp. 92 – 95, 162 – 169.
5. G. Collin, W. Lücke: *Erdöl, Kohle, Erdgas, Petrochem.* **38** (1985) 120 – 126.
6. J. P. Kennedy, E. Marechal: *Carbocationic Polymerization,* John Wiley, New York 1982, pp. 10 –12.
7. E. W. Flick: *Handbook of Adhesive Raw Materials,* 2nd ed., Noyes Publication, Park Ridge, N.J., 1985, pp. 91 – 101.
8. R. Jordan: *Coating* **15** (1982) 335 – 337; **16** (1983) 12 – 14, 143 – 147.
9. DeWitt Comp.Inc., Inter-Tech Ltd., Directory for the Hydrocarbon Resin Industry, Houston 1991.
10. A. P. Dunlop, F. N. Peters: "The Furans," *ACS Monogr. 119,* Rheinhold Publishing Corporation, New York 1953.
11. Süd-Chemie AG: "Furfural Hydrogenation, Catalysts and Chemistry," Technical Bulletin, München.

12. W. J. McKillip: "Chemistry of Furan Polymers," *ACS Symp. Ser.* **385** (1989) chap. 29.
13. Merck & Co., US 3 083 236, 1963 (V. Torleif, J. D. Garber, R. E. Jones).
14. A. Gandini, *Adv. Polym. Sci.* **25** (1977) 47 – 58.
15. J. B. Barr, S. B. Wallon, *J. Appl. Polym. Sci.* **15** (1971) 1079.
16. Y. Hachihama, T. Shono, *Technol. Rep. Osaka Univ.* **7** (1957) 479.
17. E. Wewerka, E. Loughran, K. J. Walters, *J. Appl. Polym. Sci.* **15** (1971) 1437.
18. E. Wewerka, *J. Polym. Sci.* **A19** (1971) 2703.
19. Y. Hachihama, T. Shono, *Technol. Rep. Osaka Univ.* **4** (1954) 413.
20. A. K. A. Rathi, M. J. Chanda, *J. Appl. Polym. Sci.* **18** (1974) 1541.
21. T. A. Krishnan, M. Chanda, *Angew. Makromol. Chem.* **43** (1975) 145.
22. R. T. Conley, I. J. Metil, *J. Appl. Polym. Sci.* **7** (1963) 37.
23. R. T. Conley, I. J. Metil, *J. Appl. Polym. Sci.* **7** (1963) 1083.
24. G. E. Maciel, I. S. Chuang, G. E. Myers, *Macromolecules* **17** (1984) 1087.
25. QO Chemicals Technical Data Bulletin no. 194, 1979.
26. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* **vol. I/1,** Verlag W. A. Colomb, Stuttgart-Berlin 1971, pp. 390 ff.
27. *Ullmann,* 4th ed., **15,** 547 – 552.
28. Chugoku Marine Paint, JP 49 092-137, 1972.
29. Kubo Takashi Paint, JP 60 124-666, 1983.
30. Nippon Seika KK, JP 60 018-551, 1983.
31. Dainippon Paint KK, JP 50 109-234, 1974.
32. Sakai Kagaku Koggo Co, JP 74 001-451, 1967.
33. G. Bassler, DD 140 053, 1978.
34. E. W. Flick: *Printing Ink Formulations,* Noyes Data Corp., Park Ridge, N.J. 1985.
35. W. Ritzerfeld, DE 3 219 893, 1982.
36. Toka Shikiso Chem. KK, JP 52 004-310, 1975.
37. Millmaster Onyx Corp., US 4 070 500, 1976.
38. Canon KK, JP 62 160-273, 1986.
39. Carre SA, FR 2 058 553, 1969.
40. Pentel KK, JP 56 065-062, 1979.
41. Fuji Photo Film KK, DE 2 648 498, 1975.
42. Hitachi KK, JP 61 002-161, 1984.
43. Mitsubishi Paper Mill, JP 53 017-346, 1976.
44. Fuji Photo Film Co, DT 2 232 185, 1971.
45. Annheuser Busch Inc., US 4 033 914, 1976.
46. Hitachi Chem. Co, JP 49 004-731, 1972.
47. H. Kempin, EP 239 658, 1986.
48. Sugiyama Sangyo Kag., JP 5 6024-474, 1979.
49. Aika Kogyo KK, JP 57 056-242, 1980.
50. Hohnen Oil KK, JP 6 0 34-775, 1979.
51. Oji Cornstarch KK, JP 59 152-969, 1983.
52. Aishin Kako KK, JP 55 149-742, 1979.
53. Kao Corp. JP 62 179-847, 1986.
54. Ciba Geigy AG, EP 176 484, 1984.
55. Fa. Munchen, DT 2 438 724, 1974.
56. Nippon Crown Cork KK, JP 57 189-035, 1981.
57. BASF AG, DE 3 406 474, 1984 (K. Fischer, H. Petersen, H. Kasch, E. Wistuba).
58. S. Patai: *The Chemistry of the Carbonyl Group,* Interscience, New York 1966, pp. 589 – 590.
59. F. Josten, *Fette Seifen Anstrichm.* **54** (1952) 673.
60. J. Scheiber: *Chemie und Technologie der künstlichen Harze,* **vol. 1,** Wissenschaftliche Verlagsgesellschaft, Stuttgart 1961, pp. 164 – 165.
61. H. Wagner, H. F. Sarx: *Lackkunstharze,* Hanser Verlag, München 1971, pp. 84 – 85.
62. Chem. Werke Hüls, DT 892 975, 1940.
63. Chem. Werke Hüls, DE 3 324 287, 1983.
64. Farbenfab. Bayer AG, FR 1 595 632, 1967.
65. Chem. Werke Hüls, DE 3 334 631, 1983.
66. Chem. Werke Hüls, DT 870 022, 1944.
67. Farbenfab. Bayer AG, DT 907 348, 1940.
68. Chem. Werke Hüls, DT 826 974, 1949.
69. Arakawa Kagaku Kogyo, JP 01 069-682, 1987.
70. Weyerhauser Co, US 3 947 425, 1972.
71. BASF AG, EP 7 106, 1978.
72. Farbenfab. Bayer AG, DT 907 348, 1940.
73. Rheinpreussen, DE-AS 1 066 020, 1957.
74. Deutsche Texaco AG, DT 2 039 330, 1970.
75. Rheinpreussen, DT 890 866, 1941.
76. Dainichi Seika Co, JP 7 022 218, 1965.
77. Rheinpreussen, DT 1 155 909, 1957.
78. Rheinpreussen, DT 1 300 256, 1960.
79. C. Lüttgen: *Die Technologie der Klebstoffe,* 2nd ed., part 1, W. Panzegrau Verlag, Berlin 1959, p. 212.
80. F. Engelhardt, J. Wöllner, *Brennst. Chem.* **44** (1963) 52.
81. *Houben-Weyl,* 4th ed., **XIV/2,** p. 416.
82. VEB Farbenfab. Wolfen, DD 1 171 156, 1962.
83. Asahi Chem. Ind. Co, JP 7 0 38 425, 1966.
84. Deutsche Texaco AG, DT-AS 1 767 904, 1968.
85. Deutsche Texaco AG, DT 1 922 015, 1969.
86. Deutsche Texaco AG, DT 2 439 828, 1974.
87. Deutsche Texaco AG, DT-OS 2 353 490, 1973.
88. Kaluga Halurgy Res., SU 925 903, 1980.
89. D. C. Sistrunk, CA 2 007 361, 1989.
90. Casco Lab. Inc., EP 66 560, 1981.
91. Deutsche Texaco AG, DT 2 363 797, 1973.
92. Deutsche Texaco AG, DT 1 247 017, 1963.
93. Deutsche Texaco AG, DT 2 264 288, 1972.
94. Owens-Illinois Inc., US 3 591 534, 1968.

95. Canon KK, JP 56 051-748, 1979.
96. Canon KK, JP 57 078-045, 1980.
97. SKW Frostberg AG, DE 3 315 152, 1983.
98. SKW Frostberg AG, DE 3 429 068, 1984.
99. I.G. Farbenind., DT 511 092, 1925.
100. BIOS-Rep. no. 629, no. 743, no. 1243.
101. *Mod. Plast.* **25** (1948) no. 12, 119.
102. BASF, DT-OS 2 400 194, 1974 (K. Heinle, E. Herrmann, J. D. Stetten).
103. M. N. Tilitschenka, *Zh. Obshch. Khim.* **25** (1952) 64.
104. R. Bittner, P. Franke, H. Patzelt, DL 83 412, 1970.
105. H. G. Rosenkranz et al., DT 1 262 600, 1965.
106. Hitachi Chem. Co., JP 7 2 41 648, 1969.
107. Howards of Ilford, GB 864 542, 1957.
108. VEB Leuna, DT 1 042 229, 1955.
109. M. Ikuta, K. Toyoshima, A. Shimizu, JP 7 1 40 875, 1966.
110. VEB Leuna, DL 12 433, 1953.
111. Rheinpreussen, DT 1 155 909, 1957.
112. Hitachi-Kasei, JP 15 098, 1964.
113. Hitachi Chemical KK, JP 50 013-415, 1973.
114. Howards of Ilford, GB 864 541, 1958.
115. Budalakk Festek, HU 029 441, 1982.
116. VEB Chem. Bitterfeld, DD 234 161, 1983.
117. Lechler Chemie GmbH, EP 41 200, 1980.
118. Toyo Soda Mfg. KK, JP 62 236-804, 1986.
119. H. G. Rosenkranz, P. Jodl, M. Schloffel, DL 75 746, 1969.
120. Toyo Ink Mfg. KK, JP 5 2 043-501, 1975.
121. Dainippon Ink Chem. KK, JP 56 093-776, 1979.
122. Suzuka Paint Mfg. KK, JP 50 150-794, 1974.
123. Suzuka Paint Mfg. KK, JP 50 156-594, 1974.
124. Dunlop Ltd., FR 2 221 507, 1973.
125. W. Kunzel, DL 136 149, 1977.
126. Vibac Spa., GB 2 140 439, 1983.
127. Union Carbide Cop., US 3 772 237, 1971.
128. TDK Corp., JP 59 124-896, 1982.
129. TDK Corp., JP 59 124-894, 1982.
130. H. Petersen, *Synthesis* 1973, no. 5, 243 – 292.
131. BASF, DE-OS 2 847 030, 1978 (H. Petersen et al.).
132. BASF, DE-OS 2 941 635, 1979 (H. Petersen, K. Fischer, H. Zaunbrecher).
133. BASF, DT-OS 2 757 176, 1977 (H. Petersen et al.).
134. BASF, DT 2 757 220, 1977 (H. Petersen et al.).
135. BASF, DE 3 406 473, 1984 (K. Fischer et al.).
136. D. E. Floyd: *Polyamide Resins,* Reinhold Publishing Corp., New York 1966.
137. W. Imöhl, *Adhäsion* **18** (1974) no. 1, 7 – 11.
138. W. Link, G. Spiteller, *Fat Sci. Technol.* **92** (1990) 135 – 138.
139. R. Jordan: *Schmelzklebstoffe, Rohstoffe-Herstellung,* Hinterwaldner Verlag, München 1985, p. 93.
140. W. Imöhl, *European Adhesives & Sealants* (1991) June, 17 – 18.
141. W. Imöhl in R. Hinterwaldner (ed.): *Verbund- und Vergußtechnik in der Elektrotechnik, Elektronik und Optik,* Hinterwaldner Verlag, München 1989, pp.115 – 122.
142. Schering AG, *Binders for Printing Inks,* company brochure, Bergkamen.
143. Washburn Co., US 2 663 649, 1952 (W. B. Winkler).
144. Schering AG, DE 2 534 121, 1975 (M. Drawert, E. Griebsch, W. Imöhl).
145. J. J. Domanski in D. F. Zinkel, J. Russell (eds.): *Naval Stores,* Pulp Chem. Assoc., New York 1989, p. 896.
146. W. Gronemeyer, E. Fuchs (eds.): *Karenz und Hyposensibilisierung bei Inhalations- und Insektengift-Allergie,* Dustrie, München 1983.
147. Nevcin Polymers BV: unpublished data, 1986; Reprotox Huntingdon, Germany: unpublished data, 1979.
148. Council Directive 89/109/EEC, 21 Dec. 1988.
149. H. J. Palmen, *Kunststoffe* **81** (1991) 736.
150. EEC Commission: III/3624/91/EN-Rev. 1.
151. Commission Directive 90/128/EEC, 23. Feb. 1990.
152. Council Directive 82/711/EEC, Oct. 18, 1982.
153. Council Directive 85/572/EEC, Dec. 19, 1985.

Resorcinol

KLAUS W. SCHMIEDEL, Hoechst Aktiengesellschaft, Frankfurt am Main, Federal Republic of Germany

DANIEL DECKER, Hoechst Aktiengesellschaft, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties 2
4. Production 3
4.1. Via Sulfonation of Benzene 3
4.2. Via Hydroperoxidation of *m*-Diisopropylbenzene 4
4.3. Other Processes 4
5. Environmental Protection 5
6. Quality Specifications and Analysis . 5
7. Uses 5
8. Transport Classifications and Occupational Health 5
9. References 6

1. Introduction

Resorcinol [*108-46-3*], $C_6H_6O_2$, M_r 110.112, is a white crystalline compound with a weak odor and a bittersweet taste. Other names are 1,3- (or *meta*-) benzenediol, 1,3-dihydroxybenzene or dioxybenzene, and 3-hydroxyphenol. Resorcinol does not occur in nature as such. The first syllable of the name resorcinol is derived from the word resin because HLASIWETZ and BARTH obtained it by the destructive distillation of a natural resin in 1864 [1]. The structure of resorcinol

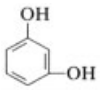

is similar to that of orcinol, 5-methyl-1,3-benzenediol [*6153-39-5*], which accounts for the second part of the name. Resorcinol has been produced industrially for more than 100 years.

2. Physical Properties [2]

The physical properties of resorcinol can be summarized as follows:

mp (101.3 kPa)	109.8 °C
bp (101.3 kPa)	276.5 °C
(1.33 kPa)	151.5 °C
Vapor pressure, liquid	
(110 °C)	0.10003 kPa
(150 °C)	0.99438 kPa
(190 °C)	6.1064 kPa
mp (101.3 kPa)	109.8 °C
bp (101.3 kPa)	276.5 °C
(1.33 kPa)	151.5 °C
Density, solid	
(α-phase, 20 °C)	1.278 g/cm^3
(β-phase, 20 °C)	1.327 g/cm^3
Density, liquid	
(110 °C)	1.1758 g/cm^3
(150 °C)	1.1439 g/cm^3
(190 °C)	1.1103 g/cm^3
Dynamic viscosity, liquid	
(130 °C)	4.994 mPa · s
(150 °C)	3.013 mPa · s
(190 °C)	1.388 mPa · s
Dynamic viscosity, vapor	
(110 °C)	0.00873 mPa · s
(150 °C)	0.00967 mPa · s
(190 °C)	0.01060 mPa · s
Specific heat capacity, solid	
(20 °C)	1191 J kg^{-1} K^{-1}
(50 °C)	1325 J kg^{-1} K^{-1}
(70 °C)	1417 J kg^{-1} K^{-1}
Specific heat capacity, liquid	
(at *mp*)	2185 J kg^{-1} K^{-1}
(140 °C)	2151 J kg^{-1} K^{-1}
(180 °C)	2287 J kg^{-1} K^{-1}
Specific heat capacity, vapor	
(110 °C)	1402 J kg^{-1} K^{-1}
(150 °C)	1512 J kg^{-1} K^{-1}
(190 °C)	1609 J kg^{-1} K^{-1}
Entropy	
(ideal gas, 25 °C)	2888 J kg^{-1} K^{-1}
Heat of formation	
(ideal gas, 25 °C)	− 274.7 kJ/mol
Gibbs free energy of formation	
(ideal gas, 25 °C)	− 181.3 kJ/mol
Heat of fusion (at *mp*)	21.0 kJ/mol
Heat of vaporization	
(110 °C)	78.2 kJ/mol
(150 °C)	75.2 kJ/mol
(190 °C)	72.0 kJ/mol

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_111

mp (101.3 kPa)	109.8 °C
bp (101.3 kPa)	276.5 °C
(1.33 kPa)	151.5 °C
Heat of combustion (25 °C)	− 2719 kJ/mol
Estimated critical data	
t_{crit}	810 °C
p_{crit}	7490 kPa
v_{crit}	0.3 L/mol
Flash point (closed cup)	126.85 °C
Lower explosion limit in air (199 °C)	1.4 vol %
Density of aqueous solution (5.0 mol/L, 20 °C)	1.1101 g/cm^3
Solubility	
in water (0 °C)	66.7 g/100 g solvent
(20 °C)	141.0 g/100 g solvent
(40 °C)	266.3 g/100 g solvent
(60 °C)	506.0 g/100 g solvent
in CCl_4 (40 °C)	0.12 g/100 g solvent
in $CHCl_3$ (40 °C)	0.78 g/100 g solvent
in acetone (40 °C)	243.3 g/100 g solvent
in benzene (40 °C)	0.85 g/100 g solvent
in nitrobenzene (40 °C)	17 g/100 g solvent
pK_1 (20 °C)	9.4
pK_2 (20 °C)	11.4

Resorcinol exists in at least two crystalline modifications (phases) [3]. At normal pressure the α-phase is stable below ca. 71 °C, whereas the β-phase is stable above that temperature up to the melting point. Transition rates from the α- to the β-phase between 72 and 100 °C are given in [4]. Both phases exist as rhombic – pyramidal crystals (Pna 2_1), but the unit cells differ in their dimensions: α-resorcinol: $a = 1.053$ nm, $b = 0.95$ nm, $c = 0.566$ nm; β-resorcinol: $a = 0.791$ nm, $b = 1.257$ nm, $c = 0.55$ nm. The unit cells of both phases contain four resorcinol molecules.

Interatomic distances in the resorcinol molecule are as follows: C – C 0.138 nm, C – O 0.134 nm, C – H 0.108 nm, O – H 0.102 nm.

Literature data on solubilities vary widely because of the presence of the α- and β-phases: according to [5], 198 g of resorcinol is soluble in 100 g of water at 50 °C; according to [6], 371.5 g is soluble in 100 g water at 50.4 °C.

The density of aqueous resorcinol solutions increases approximately proportionally to the resorcinol concentration.

3. Chemical Properties [7]

Resorcinol is a dihydric phenol and exhibits the typical reactivity of a phenol. Its most important reaction is with formaldehyde [*50-00-0*] to form phenolic resins (→ Phenolic Resins→ Phenolic Resins).

The hydrogen atom surrounded by the two *meta*-hydroxyl groups can be substituted much more easily than the other ring hydrogens. Compared to catechol [*120-80-9*] and hydroquinone [*123-31-9*], resorcinol has the highest reactivity toward formaldehyde.

4. Production

Resorcinol is produced commercially worldwide in only a few specialized plants. All of these plants use benzene [*71-43-2*] as the main feedstock (see Fig. 1). In Japan, resorcinol is produced in two plants (Sumitomo Chemical and Mitsui Petrochemical) via 1,3-diisopropylbenzene [*99-62-7*]. The United States (INDSPEC Chemical Corp.) and Germany (Hoechst) each produce it in one plant, using the "classical" route via 1,3-benzenedisulfonic acid [*98-48-6*] (see Fig. 2).

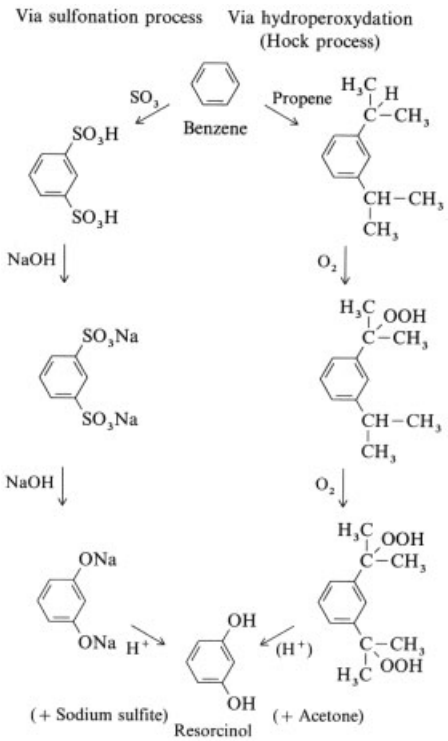

Figure 1. Resorcinol production reactions

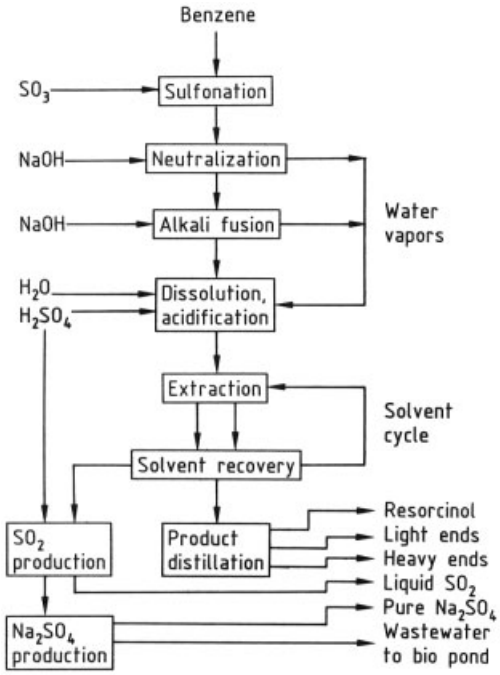

Figure 2. Resorcinol production via sulfonation

4.1. Via Sulfonation of Benzene [8]

The sulfonation of benzene formerly led to the production of considerable amounts of waste because it was carried out with mixtures of sulfuric acid and sulfur trioxide. Excess sulfuric acid was precipitated with lime to form gypsum, which had to be disposed of in landfills.

Process Description. In the process variation used in Germany today, the sulfonation of benzene is carried out continuously with sulfur trioxide [*7446-11-9*] alone so that only small amounts of sulfuric acid are contained in the sulfonation mixture and the addition of lime is not required. After neutralization with sodium sulfite, soda ash, or sodium hydroxide solution, the sulfonation product disodium benzene-1,3-disulfonate [*831-59-4*] is mixed with excess sodium hydroxide and fed to an alkali fusion reactor at 320 – 350 °C. The endothermic solid-state reaction yields a white powder, consisting chiefly of disodium resorcinate [*6025-45-2*], sodium sulfite, and some excess sodium hydroxide.

Depending on the method used, the product of the fusion reaction is treated with either a small quantity of water (in this case, solid sodium sulfite which contains organic impurities, is obtained as byproduct) or a large quantity of water forming an almost saturated solution of the product. In both processes the dissolved disodium resorcinate is then reacted with sulfur dioxide, sulfuric acid, or hydrochloric acid to give resorcinol. The dissolved resorcinol is extracted with an organic solvent. The preferred solvent is diisopropyl ether [*108-20-3*], but benzene, 4-methyl-2-pentanone (methyl isobutyl ketone) [*108-10-1*], or others can also be used. The solvent is then distilled off, and crude resorcinol is purified further by distillation in vacuum. The byproducts of this process are small amounts of light ends, essentially phenol, as well as cresols and 3-mercaptophenol [*40248-84-8*], and small amounts of heavy ends, mostly oligohydroxybiphenylenes.

The salt solution remaining after extraction is processed further. When sulfuric acid is used to neutralize disodium resorcinate, sulfur dioxide and sodium sulfate [*7757-82-6*] can be recovered as byproducts from this stream.

4.2. Via Hydroperoxidation of *m*-Diisopropylbenzene

Benzene, together with a benzene – cumene mixture recycled from the alkylation process, is alkylated with propene [*115-07-1*] in the liquid phase by using an $AlCl_3$ – HCl complex as catalyst (→ Phenol, Chap. 4.1.). After addition of recycled *p*-diisopropylbenzene (*p*-DiPB) and triisopropylbenzene (TriPB), the alkylate is subjected to isomerization/transalkylation. In this step most of *p*-DiPB and TriPB are converted into *m*-diisopropylbenzene (*m*-DiPB). The reaction mixture is then separated by distillation into three fractions (benzene – cumene, *m*-DiPB, and *p*-DiPB – TriPB). The subsequent autoxidation of pure *m*-DiPB proceeds according to a radical chain mechanism. It is accomplished in a cascade of aeration reactors by using compressed air in an aqueous alkaline medium at 90 – 110 °C and 0.5 – 0.7 MPa to yield [1,3-phenylenebis-(1-methylethylidene)]bishydroperoxide (*m*-diisopropyl-

benzene dihydroperoxide, DHP) [*721-26-6*] [9] (see Fig. 1).

After an overall residence time of 6 – 8 h the oxidate contains ca. 20 % DHP and ca. 35 % *m*-diisopropylbenzene monohydroperoxide (MHP). The heterogeneous oxidate is subjected to phase separation. *m*-Diisopropylbenzene dihydroperoxide is then crystallized, centrifuged, dissolved in acetone, and fed to a cleavage reactor (→ Phenol). Higher concentrations of hydroperoxides lead to the above-average formation of byproduct and increased safety hazards. The main byproducts of oxidation are *meta*-substituted acetophenones and dimethylphenyl carbinols. The latter are converted to recyclable *m*-DiPB by acid- catalyzed dehydration and hydrogenation in downstream unit operations.

The cleavage of DHP to resorcinol and acetone [*67-64-1*] is carried out under acid catalysis (preferably with ca. 1 % H_2SO_4) in boiling acetone with a reaction time of 30 min [10, 11]. After neutralization, acetone is distilled off at normal pressure and resorcinol under vacuum. Further purification of resorcinol can be achieved by recrystallization or extraction. The overall process yield is ca. 75 %, based on benzene. Alternatively, cleavage of the oxidate can be carried out in the presence of hydrogen peroxide [*7722-84-1*]. In this case the carbinols formed as byproducts are subsequently oxidized to DHP and thus eventually also converted to resorcinol [12].

4.3. Other Processes

Koppers in the United States and Mitsui Toatsu in Japan have developed a process for manufacturing resorcinol from *1,3-diaminobenzene* [*108-45-2*], which is produced by the hydrogenation of dinitrobenzene [13]. The *dehydrogenation of 1,3- cyclohexanedione* [*504-02-9*] to resorcinol, described by British Oxygen, was further developed by Hoechst in Germany to a four-stage process [14]. The starting materials are acetone and methyl acrylate [*96-33-3*] or acrylonitrile [*107-13-1*]. Neither process has been commercialized. The direct hydroxylation of phenol with hydrogen peroxide, for example, has been investigated extensively, but it yields only catechol and hydroquinone (→ Hydroquinone, Chap. 4.2.) [15].

5. Environmental Protection

Resorcinol must not enter rivers, lakes, or groundwater. Since it is readily biodegradable, small amounts in a biological wastewater treatment plant do not pose any problems. If possible, resorcinol should be collected and returned to the producer. When resorcinol is present in the air in the form of dust, smoke, or vapor, it can be removed by air scrubbing with a small amount of water in a wet filter equipped with a demister (a once-through operation).

6. Quality Specifications and Analysis

Resorcinol can be identified by IR [16] or Raman spectroscopy [17]. The UV and NMR spectra are given in [18].

A violet color is produced when an aqueous solution of resorcinol is treated with an iron(III) chloride ($FeCl_3$) solution.

Freezing-point measurement is the most useful test for determining product quality [19]. Freezing-point depression indicates impurities, most frequently the presence of water. Because of the existence of two solid resorcinol phases, determination of the melting point as a measure of product quality is not recommended. Thus, for instance, the melting range of 110 – 113 °C given in [20], does not indicate the presence of impurities. Impurities are best determined by gas chromatography of a solution of resorcinol in toluene or ethyl acetate, using a fused quartz capillary column with a nonpolar stationary phase, and a flame-ionization detector (FID). Freshly distilled technical-grade resorcinol exhibits approximately the following characteristics:

Appearance	white flakes, crystals, or powder
Freezing point	min. 109.5 °C
Purity	min. 99.5 % (wt %)
Water content	max. 0.1 % (wt %) (Karl Fischer)
Phenol	max. 0.1 % (area/area) (capillary GC)
o-Cresol	max. 0.1 %
m-, *p*-Cresols	max. 0.1 %
Catechol	max. 0.1 %
3-Mercaptophenol	max. 0.2 %
Heavy ends	max. 0.3 %

Resorcinol becomes darker during storage especially when exposed to light, but this does not impair its suitability for most applications.

7. Uses

Resorcinol is used in industry as an intermediate. Worldwide consumption is about 40 000 t/a (1990), the primary consumer (more than 50 %) is the *rubber industry*. In the production of tires and other reinforced rubber products (conveyor belts, driving belts), resorcinol – phenol – formaldehyde condensates are used to enhance adhesion between cord and rubber (dip formulations, dry bonding agents). Furthermore, some rubber mixtures contain resorcinol to improve some properties after curing.

The second largest market for resorcinol (ca. 25 %) is in high-quality wood adhesives, which are made from resorcinol, phenol, and formaldehyde. These resorcinol – phenol – formaldehyde resins are especially suitable for the manufacture of laminated wooden beams – which must be to some extent waterproof – at ambient temperature. The first such use was in the manufacture of wooden aircraft propellers ("propeller glue").

Further uses include

1) A new process for producing the intermediate *m*-aminophenol [*591-27-5*]
2) Production of light stabilizers for plastics
3) Production of sunscreen preparations for the skin
4) Production of dyes (fluorescein, eosin)
5) Manufacture of special pharmaceuticals (e.g., acne preparations)

8. Transport Classifications and Occupational Health [21]

Transport classifications and labeling are as follows:

UN no.	2876
EC no.	604-010-00-1
IMDG Code	6.1/2876/III
RID/ADR	6.1/14 C
DOT/IMO	poison B

Resorcinol must be labeled "harmful" in the EC and *mindergiftig* (less toxic) in Germany, with hazard label Xn.

Occupational Health. No MAK value has been set. Exposure to resorcinol must be kept below 10 ppm or 45 mg/m^3 TWA (OSHA-PEL and ACGIH-TLV).

Toxicology. For oral toxicity, an LD_{50} = 301 mg/kg (rat) has been found. Tests with rats show fast excretion (> 90 % within 24 h), mainly via the urine as the glucuronide [22]. Resorcinol is less toxic than phenol or catechol if absorbed by ingestion or skin penetration. Ingestion of 8 g of resorcinol by a child resulted in hypothermia, hypotension, decrease of respiration, tremors, icterus, and hemoglobinuria.

In a two-year study, no evidence of carcinogenic activity was found when up to 225 mg/kg resorcinol in water was administered orally to male or female rats or mice. However, clinical signs suggestive of a chemical-related effect on the central nervous system, including ataxia, recumbency, and tremors, were observed in the rats and mice during the two-year study [23].

9. References

1. H. Hlasiwetz, L. Barth, *Ann. Chem. Pharm.* **130** (1864) 354 – 359.
2. Database DIPPR, American Institute of Chemical Engineers, host STN, Chem. Abstr. Service, Columbus, Ohio; FIZ, Karlsruhe.
3. A. Kofler, *Arch. Pharm. Ber. Dtsch. Pharm. Ges.* **281** (1943) 8 – 22.
4. B. V. Erofeev, L. T. Mendeleev, *Vestsi Akad. Navuk BSSR, Ser. Fiz. Tekh. Navuk* 1956, no. 4, 99, 101; *Chem. Abstr.* **51** (1957) 13 536.
5. C. L. Speyers, *Am. J. Sci.* **14** (1902) no. 4, 294.
6. W. H. Walker et al., *J. Phys. Chem.* **35** (1931) 3262.
7. *Beilstein,* **6,** no. 554, 796; **E I** 398, **E II** 802; **E III** 4292; **E IV** 5658.
8. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* Verlag Chemie, Weinheim 1976.
9. Mitsui Petrochemical Ind., EP 284 424, 1988 (H. Miki et al.); *Chem. Abstr.* **110** (1989) 78 057 h.
10. Sumitomo Chemical Co., JP 6 1 00 327, 1986 (H. Tomita et al.); *Chem. Abstr.* **105** (1986) 24 062 j.

11. Mitsui Petrochemical Ind., EP 28 931, 1981 (I. Imai et al.); *Chem. Abstr.* **95** (1981) 186 840 p.
12. Mitsui Petrochemical Ind., EP 327 361, 1989 (K. Yorozu, H. Ohno); *Chem. Abstr.* **112** (1990) 76 597 b.
13. Process Evaluation/Research Planning, Third Quarterly Report (1977), Resorcinol from *m*-Phenylenediamine: The Koppers Process, Chem. Systems Inc., Tarrytown, N.Y..
14. Hoechst AG, DE 2 437 983 A 1, 1976 (W. H. Mueller, K. Riedel); *Chem. Abstr.* **84** (1976) 165 541 z.
15. Anic S.p.A., DE-OS 3 309 669, 1983 (A. Esposito et al.); *Chem. Abstr.* **100** (1984) 22 409 a.
16. *FT-IR-Atlas,* VCH Verlagsgesellschaft Weinheim 1988, no. 1007.
17. D. Penot, J.-P. Mathieu, *J. Chim. Phys., Phys. Chim. Biol.* **52** (1955) 829 – 833.
18. Sadtler Research Lab.: *Standard Spectra Collection,* H-NMR no. 6672, C-NMR no. 4428, C-13 no. 4428, UV no. 2572.
19. H. Grimminger, H. Knorsch, A. Vollmer, *Pharm. Ind.* **52** (1990) no. 10, 1280.
20. Aldrich Chemical Co. Inc.: *Catalogue Fine Chemicals,* Milwaukee, WI.
21. Database HSDB of NLM; host DIMDI, Köln.
22. Y. C. Kim, H. B. Matthews, *Fundam. Appl. Toxicol.* **9** (1987) no. 3, 409 – 414.
23. *Chemical Regulation Reporter* **14,** March 15, 1991, no. 49, 1783.

Reverse Osmosis → Membranes and Membrane Separation Processes

Oil and Gas

KLAUS WEGGEN, Hemmingen, Federal Republic of Germany (Chap. 1)

GÜNTER PUSCH, Technische Universität Clausthal, Clausthal-Zellerfeld, Federal Republic of Germany (Chap. 2)

HEINRICH RISCHMÜLLER, Technische Universität Clausthal, Clausthal-Zellerfeld, Federal Republic of Germany (Chap. 3)

1. Origin, Deposition Mechanisms, Exploration, and Occurrence of Crude Oil and Natural Gas . . . 3
1.1. Crude Oil . . . 3
1.1.1. Formation . . . 4
1.1.2. Formation of Deposits . . . 6
1.1.2.1. Migration . . . 6
1.1.2.2. Reservoir Rocks . . . 7
1.1.2.3. Trap Structures . . . 7
1.1.3. Oil Exploration . . . 8
1.1.3.1. Geological Investigations . . . 8
1.1.3.2. Geophysical Investigations . . . 9
1.1.3.2.1. Potential Methods . . . 9
1.1.3.2.2. Seismic Methods . . . 10
1.1.3.3. Geochemical Investigations and Determination of Coalification on Cutting Samples . . . 13
1.1.4. Occurrences, Production, and Reserves . . . 14
1.1.4.1. Production and Reserves . . . 14
1.1.4.2. Regional Distribution of the Reserves . . . 16
1.1.4.2.1. Middle East – Arabian (Persian) Gulf . . . 16
1.1.4.2.2. Europe . . . 18
1.1.4.2.3. Commenwealth of Independent States . . . 19
1.1.4.2.4. Africa . . . 20
1.1.4.2.5. North America . . . 21
1.1.4.2.6. Latin America . . . 22
1.1.4.2.7. Far East . . . 22
1.2. Natural Gas . . . 23
1.2.1. Origin and Deposit Formation . . . 24
1.2.2. Natural Gas Occurrences and Their Reserves . . . 25
1.2.3. Regional Distribution of the Reserves . . . 25
1.2.3.1. Commonwealth of Independent States . . . 25
1.2.3.2. Middle East . . . 26
1.2.3.3. North America . . . 26
1.2.3.4. Africa . . . 27
1.2.3.5. Western Europe . . . 27
2. Reservoir Engineering . . . 28
2.1. Characterization of Oil and Gas Reservoirs . . . 28
2.1.1. Reservoir Rock . . . 28
2.1.1.1. Porosity . . . 28
2.1.1.2. Pore Configuration Models . . . 30
2.1.1.3. Compressibility of the Pore Space 30
2.1.2. Physical Properties of the Pore Saturating Fluids . . . 31
2.1.2.1. Density . . . 32
2.1.2.2. Compressibility . . . 34
2.1.2.3. Viscosity . . . 34
2.1.3. Reserves . . . 35
2.2. Flow Processes in Porous Media 37
2.2.1. Single-Phase Flow . . . 37
2.2.1.1. Steady-State Flow – Darcy's Equation . . . 37
2.2.1.2. Transient Single-Phase Flow . . . 38
2.2.2. Multiphase Flow . . . 40
2.2.2.1. Capillary Pressure and Saturation 40
2.2.2.2. Relative Permeabilities . . . 41
2.2.3. Drive Mechanisms and Recovery Efficiency . . . 42
2.2.3.1. Microscopic Recovery Factor . . . 43
2.2.3.2. Areal Recovery Factor . . . 45
2.2.3.3. Vertical Recovery Factor or Flood Invasion Efficiency . . . 45
2.3. Modeling of Reservoir and Production Performance . . . 46
2.3.1. Material Balances Method . . . 47
2.3.2. Tarner Prediction Method . . . 48
2.3.3. Reservoir Simulation . . . 50
2.3.3.1. Model for Steady-State Flow . . . 50
2.3.3.2. Model for Unsteady-State Flow . 51
2.3.4. Statistical Prediction Methods (Decline Curve Methods) . . . 53
3. Deep Drilling and Production Engineering . . . 55
3.1. Deep Drilling Engineering . . . 55

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_117

3.1.1. Introduction 55
3.1.2. History 55
3.1.3. Rotary Drilling 56
3.1.4. New Developments in Deep Drilling Engineering 60
3.2. Investigation of the Reservoir Rocks 62
3.3. Casing and Cementation 65
3.4. Production Engineering 66
3.4.1. Objectives 66
3.4.2. General Production Engineering . 68
3.4.2.1. Completion 68
3.4.2.2. Perforation 68
3.4.2.3. Well and Reservoir Treatments . 69
3.4.2.4. Workover 71
3.4.2.5. Horizontal Wells 71
3.4.3. Oil Production Engineering 72
3.4.3.1. Collection and Treatment of Crude Oil 76
3.4.3.2. Special Problems in Crude Oil Production 77
3.4.4. Natural Gas Production Engineering 78
3.4.5. Offshore Recovery 79
4. References 83

Chapter 1 is based on the article Erdöl und Erdgas, Bildung, Suche und Vorkommen by Hans-Ulrich Hark † in the 4th edition of Ullmann's.

Abbreviations

A	cross-sectional area, m^2
A_s	swept area, m^2
B	formation volume factor
$B_{o,w,g}$	formation volume factor for oil, water, gas
B_t	two-phase formation volume factor
c	compressibility coefficient, Pa^{-1}
c_f	pore volume compressibility, Pa^{-1}
$c_{o,w,g}$	compressibility for oil, water, gas, Pa^{-1}
c_{pc}	pseudocritical compressibility
c_{pr}	pseudoreduced compressibility
c_r	reduced compressibility
C	critical point
D	production decline rate, s^{-1}
D	hole diameter
E_A	areal oil recovery factor
E_D	microscopic oil recovery factor
E_I	vertical recovery factor
E_R	recovery efficiency
f	fractional flow
$f_{o,w,g}$	fractional flow for oil, water, gas
F	surface area of cutting particles
g	acceleration due to gravity, $m\,s^{-2}$
G_p	gas produced, m^3
h	thickness, elevation, m
k	permeability of rock, m^2
k	drillability coefficient
$k_{o,w,g}$	effective permeability for oil, water, gas, m^2
k_r	relative permeability
$k_{ro,w,g}$	relative permeability for oil, water, gas
l	length, m
m	gradient
M_r	molecular mass
M	mobility ratio
N	initial oil volume, original oil in place, m^3
N_p	oil produced, m^3
N_{pa}	maximum cummulative production rate
OGIP	original gas in place
OOIP	original oil in place
p	pressure, Pa
p_b	bubble point pressure, degassing pressure, Pa
p_c	capillary pressure, Pa
p_e	pressure at reservoir boundary, Pa
p_m	mean pressure
p_{pc}	pseudocritical pressure
p_{pr}	pseudoreduced pressure
p_r	reduced pressure
p_{wf}	flowing pressure in well, Pa
p_{ws}	shut-in pressure, Pa
P	weight on bit
q	flow rate, $m^3\,s^{-1}$
q_a	limiting production rate
$q_{o,w,g}$	flow rate for oil, water, gas, $m^3\,s^{-1}$
r	radius, m
r_e	radius of area drained by the well
$r_{w,c}$	well radius, reservoir boundary radius, m
R_p	produced oil – gas ratio
R_s	gas – oil ratio
R_{si}	initial gas – oil ratio
S	saturation

S	drillability of the rock
$S_{o,w,g}$	saturation oil, water, gas
S_{or}	residual oil saturation
S_{wf}	front saturation
S_{wi}	initial water saturation, irreducible water saturation
t	time, s
t_p	producing time, s
T	temperature, K
T_c	capillary temperature
T_{pc}	pseudocritical temperature
T_r	reduced temperature
v	velocity, $m\,s^{-1}$
$\boldsymbol{v}$	apparent velocity
V	cuttings volume, m^3
V_b	bulk volume, m^3
V_B	total rock volume
V_p	pore volume, m^3
W_i	volume of water, m^3
W_p	water produced, m^3
x	distance
z	real gas factor
γ	Euler constant
θ	contact angle
Θ	static shear stress, dyne/cm^2
μ	dynamic viscosity
μ_g	viscosity of a real gas, $Pa \cdot s$
$\varrho_{o,w,g,f}$	density of oil, water, gas, fluid, kg/m^3
ϱ_m	density of mud
σ	surface tension, interfacial tension
ϕ	porosity
ϕ_d	dead porosity
ϕ_{eff}	effective porosity
$\boldsymbol{\Phi}$	flow potential, Pa
$\lambda_{o,w,g}$	mobility of oil, water, gas

Subscripts

i, 0	initial
r	residual
b	at bubble point
0	standard
j	component j
f	fluid

1. Origin, Deposition Mechanisms, Exploration, and Occurrence of Crude Oil and Natural Gas

1.1. Crude Oil

Crude oil is the name given to all organic compounds which are liquid under reservoir conditions. They can partly solidify at the surface after expansion and cooling. Petroleum is a mixture of hydrocarbons in varying proportions, which are accompanied by compounds of sulfur, oxygen, phosphorus, and nitrogen. Minor constituents are porphyrins, ash-forming metal compounds (usually sulfides of vanadium, nickel, copper, cobalt, molybdenum, lead, chromium, and arsenic), as well as inorganic salts, hydrogen sulfide, and – as a typical constituent of "crude oils" – water in varying amounts. The differing mixing ratios preclude uniform chemical and physical properties. Elementary analyses show average carbon contents of ca. 79.5 – 88.5 wt %, hydrogen contents of ca. 10 – 15.5 wt %, and impurity contents of up to a maximum of 5 wt %. Most crude oils are optically active.

The essential ingredients of the petroleum mixture are alkanes, naphthenes, and aromatics. The oxidation products naphthenic acids, terpenes, and phenols are also present in some cases. According to the predominant constituents, a distinction is made between *paraffin-based crude oils* and *naphthene- or asphalt-based crude oils.* If both hydrocarbon types are contained in considerable amounts in crude oil, it is termed *mixed-based oil.* The composition of crude oils on a worldwide basis includes more than 30 % paraffins, at least 40 % naphthenes, and ca. 25 % aromatics. Naphthene-based crude oils – that contain resinous and asphaltic substances – are frequently found in the upper level of the Earth's crust, mixed-based oils in the middle, and paraffin-based oils in the deeper zones.

Near the surface, the crude oils lose part of their volatile constituents. During this, with chemical processes participating, more or less solid residues are formed. Mainly ozokerite (earth wax) is formed from paraffinic crude oils and natural asphalts (→ Asphalt and Bitumen) from naphthenic crude oils.

For decades, petroleum has been mainly used as raw material for the production of engine fu-

els, lubricants, and fuel oil. Because of the complex compositions of crude oils, the chemical industry has developed many processes through which petroleum has become the basis of numerous plastics. In 1990, more than 10 % of world petroleum production was used in petrochemistry.

1.1.1. Formation

Despite a century's discussion with many contradictory opinions and intensive research, the formation of petroleum can still not be satisfactorily explained. The discussion on the inorganic formation of crude oils continues to this day. Russian experts, in particular Porfirev [56] do not exclude the fact that petroleum is formed from inorganic matter (metal carbides and superheated steam) even though it is generally agreed that the crude oils are predominantly of organic origin. The investigations of potential petroleum source rocks in the past decades have contributed considerably to this opinion. All deposits in the sedimentary basins contain organic residues of terrestrial, limnic, fluvial, and marine species, whose conversion under anaerobic conditions leads – as can be demonstrated chemically – to bitumen (soluble in organic solvents) or kerogen (insoluble). Such rocks are classified as source rocks according to the amount of organic material. In clastic sediments, a petroleum source rock should contain at least 0.5 wt % and in carbonates at least 0.3 wt % of total organic carbon (TOC). A source rock with a TOC value of more than 4.0 wt % is classified as excellent [57].

The organic substance originates from photosynthesis that started on a major scale 2×10^9 years ago. Welte [58] has calculated that in the course of the Earth's history 6.4×10^{15} t organic carbon (C_{org}) has been produced. Only 18 % of the organic carbon, however, has contributed to the genesis of petroleum, 82 % is bound in the sediment as carbonate. The ratio of C_{org} bound to carbonate to C_{org} bound to petroleum is 11 000 : 1. The organic origin of the petroleum can also be deduced from the observation that crude oils are cracked at temperatures above 200 °C to methane and graphite. The simultaneous occurrence of copper, nickel, vanadium and molybdenum (that are also detected in the sapropel) in crude oils is also in favor of its organic origin. Sporadic occurrences of small oil reservoirs in volcanic or in magmatic-metamorphic rocks that have repeatedly been quoted by the "inorganic" protagonists as counter-evidence are always in connection with local elevations of the basement. It can be assumed that the crude oil has migrated from deeper-lying source rocks in intercalated volcanic strata in a higher position between sediments or by lateral migration where there is juxtaposition with magmatic-metamorphic reservoir rocks.

Crude oils are formed principally from phytoplankton and bacteria; up to the Silurian period from lower plants, e.g., blue and green algae, and from the Devonian period from higher forms that occur in increasing amounts. The fats (lipids), proteins, pigments, and amino and nucleic acids formed from them are the starting materials for bitumens.

A prerequisite for the formation of oil is that the organic material lying at the bottom of the sea or lake must be sedimented in the absence of oxygen, so that it cannot be decomposed and lead to the formation of kerogen. Such conditions exist, e.g., in eutrophic calm waters at great depth where the water in the sediment no longer contains dissolved oxygen. A reducing hydrogen sulfide medium prevails. Ideal prerequisites for such conditions are present in sapropels in which the organic matter is decomposed, decayed by anaerobic bacteria. The majority of the organic matter, however, is oxidized and recycled as CO_2.

In the course of the Earth's history, source rocks have been formed in various anoxic zones:

1) *In nonmarine lakes*, a uniform warm climate with little inflow of oxygen- containing water is the precondition for source rock formation. Lake Tanganyika is an example.
2) *In closed inland seas with a positive water balance* (water inflow exceeding rate of evaporation), anoxic conditions arise in warm climatic conditions, at a great depth, and excess water. The Black Sea with its deep zones free from currents and its bays free from breakers is an example.
   With a *negative water balance*, an oxygen excess arises, preventing source rock formation. The Red Sea and the Dead Sea are examples of this.

3) *The ascent of marine currents from greater depths* to the surface leads to the substantial consumption of the scarce oxygen present in these currents as a result of the biological productivity near the surface. From the organic residues that sink in the deep water lacking in oxygen, source rocks have been formed. Examples of this occur off the Southwest African coast (Benguela current) and off the west American coast (Humboldt current off Peru).
4) Source rocks are formed in the *open ocean* as a result of global climatic warming in association with the large transgressions. The many thick source rocks in the Jurassic and Middle Cretaceous periods are attributable to this.

An important factor in source rock formation related to the transgressions of the oceans is the transport of the organic residues from the energy-rich zones into regions of calm water lacking in oxygen. This is supported by the adsorption of organic material onto fine clay particles, which sink to the sea floor in calm regions and compress the organic material content of the sediment. Anaerobic conditions can prevail only a few centimeters below the surface of argillaceous deposits. Since ca. three quarters of all argillaceous and carbonatic sedimentary rocks contain organic carbon, the applied exploration usually assumes that not only typical sapropels but all sediments, with a fairly high content of organic matter, are to be regarded as potential source rocks for hydrocarbons.

A high sedimentation rate with a large proportion of coarsely clastic material, on the other hand, dilutes the content of organic carbon and leads to source rocks of low quality, if to any at all.

Actuo-geological observations on hydrocarbon groups from recent marine deposits off the Californian coasts and comparative infrared spectroscopic studies on oils in the Pliocene Los Angeles Basin indicate that at the time the oil was formed in the Pliocene epoch, water and sedimentation conditions must have prevailed that are similar to those in the offshore regions of the sea nowadays [59].

Even in shallow depths down to 1000 m and temperatures up to 50 °C the organic material embedded in the sediment is converted to kerogen. This *diagenetic process* intensifies with further covering and sinking as well as with temperature rise as a result of heating from the Earth's interior. This process, known as *catagenesis,* proceeds at depths of 1000 – 5000 m and temperatures up to a maximum of ca. 175 °C (Fig. 1). During catagenesis the organic material dehydrogenates, with the formation of methane and hydrogen as well as $C_{13} - C_{30}$ liquid hydrocarbons with increasing temperature. Oxygen and hydrogen escape more rapidly than the carbon and nitrogen from the source rock and so the relative carbon content increases. The H : C ratio decreases with increasing temperature [57].

The conversion of organic material by pressure and temperature occurs over geological ages. The biochemical processes, which are still not known in detail, involve the action of reducing bacteria and the catalytic action of the rocks. The conversion can be described as a function of temperature, time, and pressure by modeling (see [60]). The temperature is the most important factor. According to these ideas, the conversion of the organic matter included in the source rock starts even at low temperature with the intensive contribution of microbes and leads to the formation of proto-bitumens. Methane is formed in minor amounts as a byproduct of the microbial conversion. Through further settling, as a result of increased sediment cover or through heating as a result of subsurface magma intrusions, the temperature of the source rock rises and, by liquefaction of a part of the primary bitumen, leads to true petroleum formation.

The simultaneous compaction of the argillaceous source rocks, and the recrystallization of the clay minerals into more stable forms with lower water content cause large amounts of pore and crystal water to be released. With this squeezed-out dehydration water, the newly formed oil migrates according to the geothermal gradient – at depths ranging between ca. 1000 and 3000 m (the so-called oil window) – possibly in dissolved form or as small droplets or bubbles, out of the source rock and into the overlying or adjacent reservoir rock (see [61]).

This primary migration of oil droplets from the source rocks to the petroleum reservoir rocks can proceed laterally and, in particular, vertically. The low-molecular, possibly gaseous hydrocarbons function as a propellant. While

lateral migration probably occurs by capillary paths, vertical migration through impermeable strata is mainly based on fine fissures that close afterwards [59]. If the thermal stress on the oils in source or reservoir rocks proceeds further, the degradation and break-up of the high-molecular hydrocarbons by chemial maturation leads to ever smaller and more stable compounds. Thus, a natural cracking of the petroleum occurs. Accordingly, paraffin-based oils are found mainly in deeper regions of the Earth's crust. The maturation process concludes with the conversion to methane.

During this maturation of the oils already present, a further part of the organic matter that has remained fixed is possibly converted into petroleum. According to [60], the temperatures necessary for the genesis and maturation of the petroleum are closely related to the geological time available for them. Thus, according to Russian ideas, ca. $40-50\times10^6$ a are required for the start of petroleum formation at an earth temperature of at least 60 °C. The Californian deposits mentioned, though, have formed at temperatures above 100 °C in only $1-2\times10^6$ a [62].

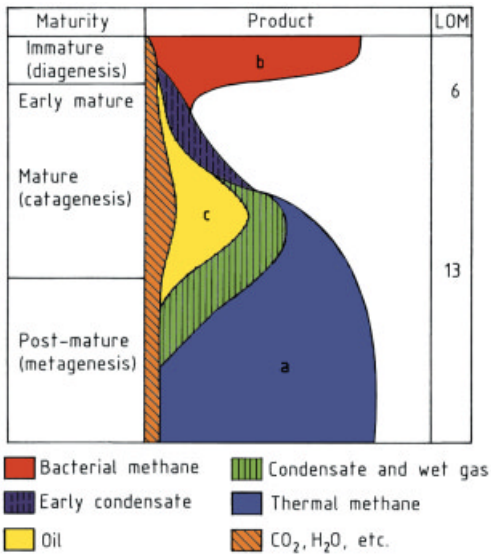

Figure 1. General schemes of hydrocarbon formation (LOM = level of organic metamorphism, according to Grunau, *J. Petrol.,* July 1983)
a) All kerogen types; b) Bacterial methane, certainly kerogen Type III; c) Mainly kerogen types I + II

1.1.2. Formation of Deposits

Deposit formation is the concentration of infinite numbers of oil droplets which form in source rocks and are then squeezed out of them to economically exploitable accumulations. Deposit formation is the result of oil migration. The oil particles move during primary migration from the argillaceous source rocks into the porous reservoir rocks, which usually lie at a higher level. A considerable fractionation of the crude oils can occur even during this movement as a result of selective filtration, adsorption, and condensation [63], [64].

1.1.2.1. Migration

In many cases, however, the true concentration of deposit formation occurs only through migration in the bedding or horizontal secondary migration, i.e., through movement and accumulation within the permeable reservoir horizons. With favorable facial conditions, lateral migration can be very important for deposit formation. It can occur over distances of 10 km to a maximum of 100 km. With changing facies, differential migration can occur, by which the lighter hydrocarbons migrate to the higher reservoir beds. During the course of this, the oils experience a certain buoyancy in relation to the formation water present in the reservoir rocks because of their lower density. The oil is therefore concentrated in the highest zones of the reservoir horizons, which are sealed above by mudstones or salt beds. A gravity differentation of petroleum over edge water within the permeable reservoir rocks then occurs which corresponds to the different densities of the individual components. Provided gravitational separation is not prevented by capillary forces due to small pore diameters, the edge water limits the oil column to the botom of the reservoir (oil – water contact). If the oils are supersaturated with gas, a gas cap forms over the petroleum.

Possible reservoir rocks are permeable and porous sandstones and limestones. They have average porosities of ca. 10 % to a maximum of 40 %. The quality of reservoir rocks is determined, aside from their pore volume, by their permeability, which is of decisive importance for the production capacity of wells.

Petroleum and gas deposits can occur in all geological formations, starting from the Cambrian period, provided reservoir rocks are available that are accompanied or underlain by source rocks of appropriate maturity. Also the reservoir rocks must possess trap structures from the overlaying of sealing horizons, such as clay and salt beds. It is also important that the structural traps were already present at the time of oil formation and survived during the subsequent geological periods.

1.1.2.2. Reservoir Rocks

The quality of deposits is greatly affected by the type of reservoir rock.

Wind-blown sands, for example, form the good gas reservoir beds of the Rotliegendes. They spread from England to Poland over an area of 1000 – 3000 km^2 with thicknesses of up to 200 m and maximum porosity up to 20 %. *Fluvial sands* of meandering rivers can have net thicknesses of 6 – 75 m. The Triassic sands of Prudhoe Bay with reserves of 2.8×10^9 t oil belong to this type. Both *alluvial fans* such as the Cretaceous sands of the Sirte Basin in Libya and the *beach and barrier sands* parallel to the coastlines are good reservoir beds. The latter include the Jurassic Piper field (United Kingdom). The *marine platform sands,* which occur as transgressive or regressive sequences, are equally good reservoir beds. *Deep-water sands,* occurring as so- called deep-sea fans and turbidites, certainly spread over hundreds of kilometers, but are of variable quality. This reservoir type includes the Pliocene Sands in the Los Angeles Basin with a thickness up to 700 m and the Paleocene Sands of the Viking Graben in the North Sea.

Carbonate reservoir beds can be of very differing quality. Thus *reefs* or *reef limestone debris* can have excellent reservoir properties as in West Texas or Alberta/Canada, with net thicknesses of 170 – 200 m, but they are mostly of limited extent and variable quality. The *chalk* forms only average-quality reservoir beds. Mainly composed of coccoliths, the chalk is thin-bedded. The production engineering of the Austin Chalk in West Texas, the Ekofisk field in the Norwegian North Sea, or the reservoir of Reitbrook (Hamburg) is difficult. Owing to their low permeability, these deposits can be extracted profitably only in conjunction with very fine fissuring.

1.1.2.3. Trap Structures

The type of trap structure encountered strongly depends on the worldwide or regional tectonic processes in the Earth's crust. Movements at the plate edges lead to individual stress patterns with a variety of structural phenomena as the result of forces of contraction and expansion. The structures can, therefore, be of various sizes and shapes [65], [66]. The principle trap structures, as listed below, can also occur in combination.

The types of structure possible as traps for petroleum and natural gas can be of very different origin (see Fig. 2). There are anticlines (a), tiltings, faults (d), salt intrusions, and discordant overlappings, and sometimes synsedimentary facies differentiations which arose during the deposition of the beds. Permeability barriers, arising as a result of secondary recrystallization and the formation of new minerals in the pores, laterally seal the reservoir in such a way that a special form of the facies reservoir is formed. This can often be observed in conjunction with faults. The most important tectonic traps are the saddle-shaped anticlines and flat, oval mound structures as well as monoclines originating from antithetic faults. Reservoirs in association with overthrusts are usually arched in the frontal zone and sometimes enrolled. They are known as rollover structures. The first reservoirs discovered were found in association with salt domes (d), where various kinds of structural traps are possible on the flanks and the top of the dome. In the case of discordance reservoirs (c), porous horizons are sealed off after tilting and partial removal of horizontally deposited younger sediments. Other typical facies reservoirs are the reefs (b), which are usually composed of corals, as well as traps which have arisen, e.g., through the argillation of a sandstone with reservoir characteristics. In addition, there is a variety of special forms and combined reservoir types. In all of these cases, it is common that porous reservoir horizons in their top zone are covered and transposed with sealing by clay rocks and salt layers.

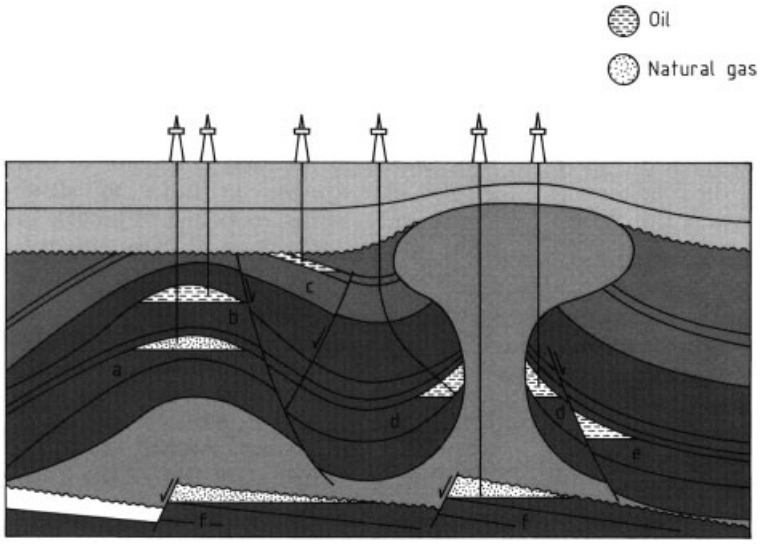

Figure 2. Most important reservoir types and trap structures
a) Under an updoming (anticline); b) In a coral reef; c) Under overlapping strata (discordance); d) On a salt plug tank; e) In a conformable downthrown fault; f) In a reverse downthrown fault

1.1.3. Oil Exploration

The potential for petroleum and natural gas present in a sedimentary basin must be proved by exploratory wells. The total costs of the wells – which rise exponentially as a result of ever increasing depths, extreme environmental conditions, and continually increasing costs of equipment and personnel – require intensive and expensive scientific preliminary exploration. Aside from the geologists, the geophysicist and, with growing importance, the geochemist, have made themselves equally valuable as exploratory experts. No drilling project is commenced unless these experts, after having made their profitability calculations, have given their approval. After calculation of the risk capital for the exploration phase, an analysis must be made of the receipts and expenditure over the whole operating life of the expected oil field: in other words, the driving force for exploration is the return on investment. Despite all care, the risk of sinking dry wells remains high. The ratio of discovery wells to dry wells, from 1 : 8 to 1 : 6, since the 1950s has not greatly improved. The reason is that exploration advances into new basins (or parts of basins) and to greater depths. The ratio is very much more favorable, for instance, in the Norwegian North Sea, where of about 300 exploration wells, of which about half were wildcats (i.e., wells drilled in regions not known to be productive), more than 100 were oilbearing. The success rate was 1 : 3.5. It is therefore necessary to reconstruct the conditions of formation of the hydrocarbons in the chronological and regional framework and to check whether potential reservoir rocks are in trap position for oil or gas. The seismic reflection method is the central element of modern exploration and crucial for almost every well. This method has been operated and continually developed for more than three decades. It is supported by a series of geological, geophysical, and geochemical methods that are used according to the degree of knowledge or development of a region and according to the problems. In an already well-known region – aside from studies on geological facies and thickness of the rock sequences – investigations start with seismic reflection measurements. In contrast, in areas which geologically are still little explored or with very deep parts of a basin, these measurements usually conclude the geophysical preparation.

1.1.3.1. Geological Investigations

In areas which are still little explored, especially on land, geological surface investigations are

used for the first assessment and possible delimitation of prospects and nonprospects. The prerequisite is merely that the stratigraphic sequence outcrops at least partly at the surface and is accessible. In the present state of exploration, the number of less explored areas has greatly diminished. In the 1980s, the number of sedimentary basin prospects is ca. 700. Of these, 350 are at least partly explored. In the remaining sedimentary basins, preliminary geological investigations have generally been carried out which permit conclusions on the nature and thickness of the sediments. From *satellite images,* precise maps are available for all parts of the world in which geologists enter their findings. Geodetic methods are becoming increasingly less important. In regions with poor vegetation, the satellite image also permits conclusions on the geology. Later – when checked and consolidated during the course of geological field surveys by a smaller or greater number of transverse profiles – these conclusions suffice to enable statements on the thickness, facies, and age of the rocks. If the results of the field surveys show that criteria for an oil prospect are present, special *examinations of rock samples* are necessary in addition to *stratigraphic investigations* in order to identify source and reservoir rocks and to provide evidence of clay as well as of evaporites being important as the seal. The evidence of oil and gas as well as asphalt traces in the rock samples alone is not sufficient to predict the presence of oil deposits. Asphalts are residues of originally liquid bitumens that, as a result of the break-up of a reservoir, have moved and oxidized near the surface. Asphalt traces can, therefore, be an indication that oil has formed and migrated. Reservoirs, however, can be left as fairly deep structures below thick, dense sediments, even when younger reservoirs have been destroyed, e.g., by uplift, tilting, and erosion.

1.1.3.2. Geophysical Investigations

1.1.3.2.1. Potential Methods

The nonseismic geophysical methods (potential methods) include electrical, magnetic, and gravimetric methods. They are used both individually and combined, for the first exploration of sedimentary basins. Since gravimetric and magnetic measurements can be made from an aircraft, they are a rapid and inexpensive tool of petroleum exploration.

Magnetic Measurements. These measurements are based on the phenomenon that rocks differ in their magnetic susceptibility. It depends on their content of magnetic minerals such as magnetite or ilmenite. The crystalline basement would cause no marked anomally, owing to the uniform distribution of magnetic minerals within them. Sediments, on the other hand, differ in their susceptibility according to the proportion of such minerals in the strata. The changes of the Earth's magnetic field are recorded with the aid of proton magnetometers and give, for example, indications of the shape and depth of the crystalline basement or of volcanic or magmatic intrusive bodies. *Aeromagnetics* was developed in World War II to detect submerged submarines. The measuring instruments are accommodated in the wingtips of the aircraft or towed behind on cables. In their basic version, the measuring instruments are coils through which current flows. The applied magnetic field is altered by the terrestrial magnetic field and the resulting anomaly is recorded.

Gravimetric Measurements. Gravimetric measurements are used to measure the changes in the Earth's gravitational field. Because of the different densities of rocks, beds of relatively low density, such as salt rocks, with negative anomalies can be demarcated from rocks of higher density, such as intrusives or quartzites, with positive anomalies. As a result it is also possible to detect blocks of identical rocks that are displaced tectonically by faults. The typical measuring instrument is the spring balance, whose spiral spring shows a different extension according to the density of the rocks. The extension is recorded and interpreted (Fig. 3). Aerial gravimetry, combined with aerial magnetics, is a rapid and cheap method of reconnaisance survey for the petroleum exploration of extended areas.

Gravimetric methods are increasingly used again for the solution of special problems. Computer model calculations can be carried out which lead to considerably improved conclusions. Thus gravimetric measurements are an

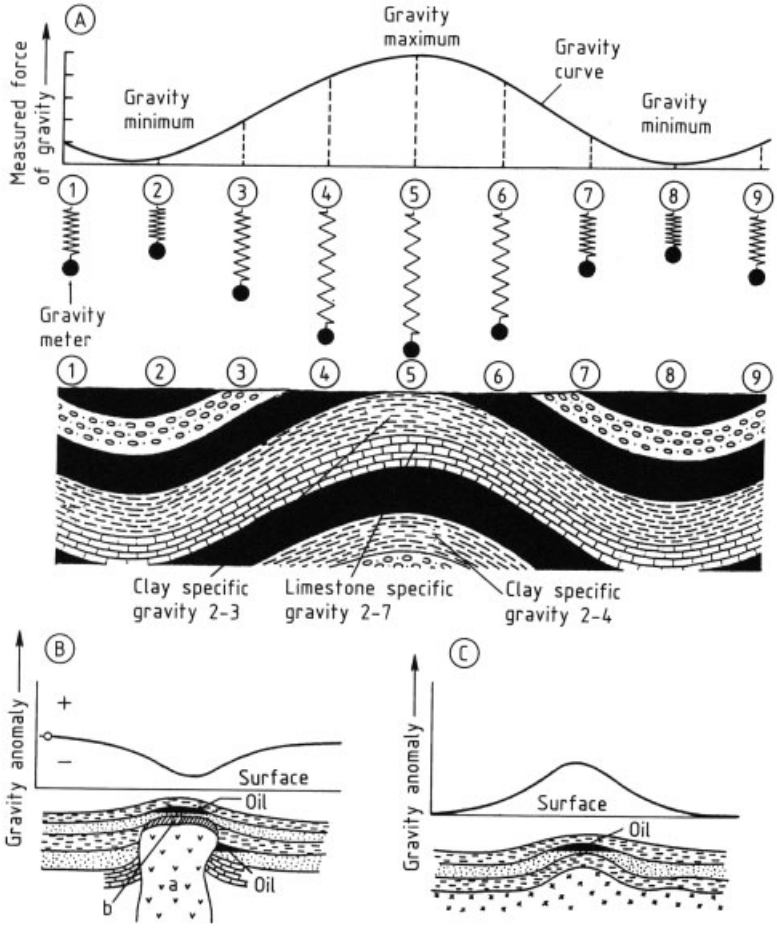

Figure 3. A) Principles of gravitational surveying; Points 1 – 9 represent observation stations where the gravity meter is installed. At these points the attraction of the gravitational force is measured. B) Negative gravity anomaly over a subsurface salt dome with associated oil accumulation; a) Salt dome; b) Cap rock; C) Positive gravity anomaly over an igneous uplift or "buried bill" with an oil accumulation in the overlying sediment

important means for detecting petroleum reservoirs at the flanks of salt domes; for the definition of the drilling and casing program and the allowed deviation for deep gas wells; and also for the interpretation of seismic surveys near salt domes, where the reflections are disturbed by the salt, in order to clearly determine the salt flanks.

Geoelectric Measurements. Geoelectric measurements of the natural Earth currents (tellurics) in the form of magnetotellurics have not fulfilled the hopes placed in them. The numerous interference factors in the industrial countries, such as electrical energy in cross- country transmission lines and on railway lines are as disadvantageous as the large spread of the results. Magnetotelluries are therefore only rarely used.

1.1.3.2.2. Seismic Methods

Since the 1970s seismic methods have become the most accurate and most frequently used exploration methods. This is mainly due to digital recording and the many processing and interpretation programs available. If in the past, seismics was a subsidiary science of geology, today it must be ranked of equal importance alongside it. The seismic methods permit very accurate conclusions on the stratigraphic sequence

and depth of geological beds. These methods measure the transit time of artificially generated elastic waves. The waves have usually been generated by the detonation of explosive charges placed in boreholes (Fig. 4). Apart from this, other energy sources also are used nowadays. The *vibroseis method,* has become established whereby a steel plate mounted below a truck is pressed by the weight of the truck onto the ground. This transmits controlled vibrations of a given frequency that are produced by a vibrator installed in the truck, to the subsoil. Four to six vibrator trucks operate simultaneously. With this method, measurements can be performed on roads and tracks and therefore in populated areas. It is also more rapid than explosion seismics, and cheaper.

The elastic waves spread in all directions and at bed boundaries, e.g., between claystone and sandstone or sandstone and carbonate. They are partly refracted and partly reflected to the surface by the echo sounder principle. The velocity of these seismic waves through the various geological strata is a measure of the type of rock and depends on the rocks' elasticity and density (ca. 1000 – 2000 m/s in loose rocks and up to 6000 m/s in compact rocks). The more accurately the actual velocity in each bed is known, the more accurately can the depth of the reflection be calculated. Computer model calculations provide very useful approximations. There are two different methods: the reflection seismic method and the refraction seismic method, the latter was used almost exclusively until the 1930s. The refraction seismic method is still used nowadays only for wide-area surveying of deep-lying horizons, but is also called in for the determination of sonic velocities in the rocks and to solve special problems.

Seismic Refraction Method. In the seismic refraction method a part of the energy waves is refracted at the refraction horizon, similar to light being refracted at the interface between water and air. Some of the refracted waves then continue to run along the bed boundaries and in the course of this generate other waves that are refracted again to the surface. The measurement is of the transit time from the detonation at the energy center until the refracted wave impinges on the geophones set up on the surface at distances of up to ca. 25 km. The seismic refraction method requires large explosive charges with the geophones at a relatively large distance and enables a relatively large depth of penetration.

Seismic Reflection Method. The seismic reflection method uses, in contrast to the refracted wave, the waves directly thrown back by the reflecting horizons. Depending on the depth of the reflecting horizons, the waves return to the Earth's surface after increasing time intervals and are received by groups of geophones systematically laid out on the ground. They are then converted to electromagnetic impulses and recorded on magnetic tapes in complicated digital recording equipment. Recording the seismic signals on magnetic disks – having already been pretreated in the recording truck by amplification, filtering, and balancing of the vibrations – enables corrections to the raw data and various investigations of the signals with simultaneous suppression of the interference energy (noise) to be carried out at the geophysical playback centers. From the transit time found for the emitted and reflected waves, the depth of the reflecting bed can be determined, provided the velocity of the rock beds passed through, which often shows rapid regional variations, is known. The depth is then half the product of velocity and transit time. Indications of the velocity of the seismic waves come from detailed calculations of the refraction and reflection seismic methods, from seismic well-logging, and directly from a geophone being lowered on a wine line in the bore hole. The geophone is positioned at the boundaries of rock units and receives the waves produced in the usual manner off the well site.

Aside from the compression waves which have up until now been used almost exclusively for scanning the structural conditions in the subsoil, *shear waves* also are gaining increasing importance today. They enable conclusions to be made on the petrographic nature of the rocks and their contents. The combination of compression and shear waves permits at least approximations to the porosity of a reservoir rock and to the pore content.

Seismic Measurements in Offshore Regions. The reflection seismic measurements can also be used in offshore regions. The seismic method based on explosives previously used, has been abandoned and replaced because of the great en-

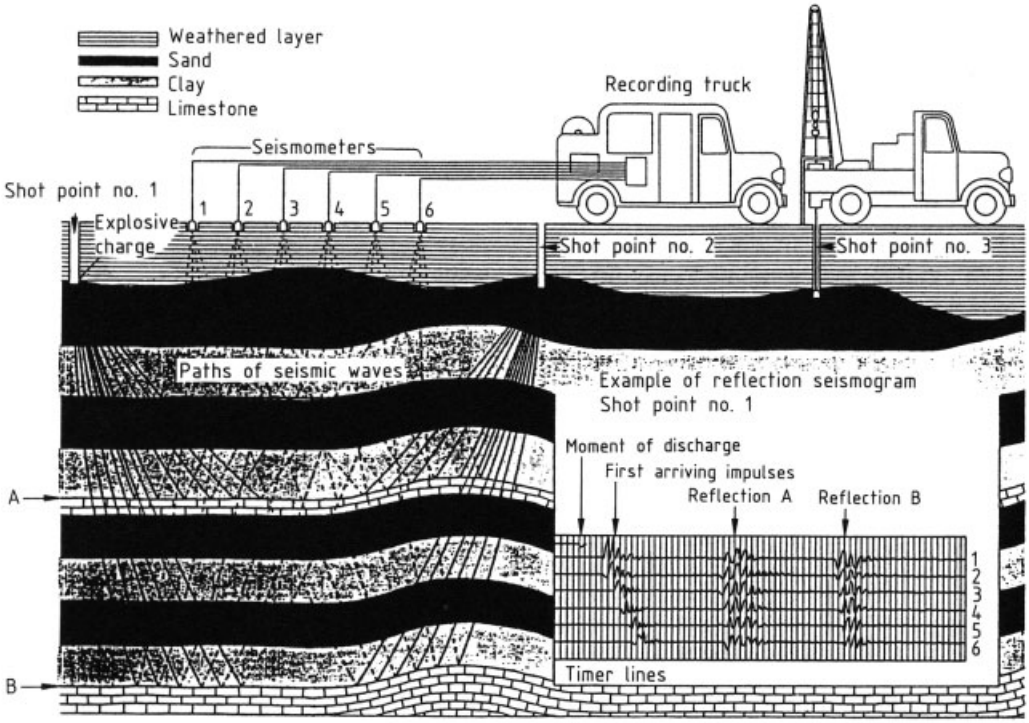

Figure 4. Reflection seismic surveying technique with two reflecting horizons, A and B

vironmental damage, especially where a large fish stock exists; at the same time the airgun method has established itself. In this method highly compressed air is suddenly released. With the large number of available recording vessels, seismic companies not only carry out commissions to order but, in all seas, carry out self-financed measurements which they then offer for sale to the oil companies for reconnaisance purposes.

Three-Dimensional (3 D) Method. In exploration seismics, the *line method* has hitherto been used almost exclusively. According to the exploratory problem definition, individual profiles spaced at a distance of 500 – 2000 m from each other are intersected by crosslines having the same spacing. The 3 D method, which is almost routine in the development of proved deposits, is also being used more and more in exploration. The groups of geophones are placed at profiles with distances of 25 m and the waves are generated by explosives fired from various sides. The digitally recorded results are processed and corrected, as in the line seismic method, and subsequently transmitted to work stations. Because of the small distances between profiles and the possibility of assembling profile lines according to one's own ideas via shot points of neighboring lines (random line), the interpretation of the reflection horizons leads to increased precision. Because of the high cost of equipment – several thousand geophones are required for the geophone installation – and the large number of staff used – up to 350 members – the method is very expensive. It pays, however, especially during development of a productive area, since as a result of the close surveying, false interpretations and consequently false structural descriptions are restricted or even avoided. Characteristics of a 3 D measurement are given below:

Area of survey	300 km^2
Surveying party size	140 employees
	32 vehicles
	1000 tracks
Geophones	14 400
Optical-fiber cable	71 500 m
Bore holes	15 076
Drilling meterage	157 183
Scanning points	480 000
Cost	ca. $\$ 7.5 \times 10^6$
Cost per km^2	\$ 25 000

1.1.3.3. Geochemical Investigations and Determination of Coalification on Cutting Samples

Geochemical investigations and determinations of coalification have become increasingly important. They are used as indicators of maturity and prospectiveness for potential source rocks and for settling questions about migration paths and the origin of crude oils and natural gases. These determinations are carried out partly with rock samples from geological field surveys or partly with rock sequences from boreholes. *Coalification investigations* serve to determine the maturity of a source rock. If it is established that source rocks are immature then any further exploration activity is often abandoned. Coalification studies, preferably reflection photometric measurements, are applied to vitrinite, which is present both in coal seams and in other sedimentary rocks. Depending on the coalification gradients, which show regional variations, the coalification and the vitrinite reflection increase with growing depth while the volatiles content simultaneously falls. From the peat to the graphite stage, the mean reflection values range from ca. 0.2 % – 10 %. According to the carbon ratio theory, the formation and release of oil is restricted to the previously mentioned "oil window" with a coalification range of ca. 0.45 – 1.25 % vitrinite reflection (subbituminous coal to high volatile bituminous A coal stage). Below the oil window light oils and condensates can still be formed and, at greater coalification, natural gases. By *chemical investigations* on source rock samples, fairly reliable conclusions on their degree of maturity are possible. *Optical parameters* are also used. Here the change of color of the dispersed organic material in the sediment on heating is determined. The optical studies have little significance and should be used only together with chemical or pyrolysis data. *Pyrolysis,* performed on a large number of rock samples, is very important for the analysis of kerogen. The Rock – Eval pyrolysis tester provides rapid determination of the type and degree of maturity of the kerogen together with a direct determination of the hydrocarbon potential. The results of the investigation can be used as geochemical logs directly at the well for the determination of the hydrocarbon potential and the maturity. The comparison of several boreholes enables excellent geological/geochemical correlations in a region.

Carbon-isotope analyses of natural gas components are routine in hydrocarbon exploration. These makes use of the close connection between the $^{13}C : {}^{12}C$ ratio in methane and the maturity of the organic matter from which the methane has been formed. It is possible, by isotopic analysis of gases, e.g., gases from cuttings from boreholes, to decide whether these gases have been formed indigenously in the horizons just drilled through or have risen from deeper strata. With isotope investigations, for example, it was possible to demonstrate that the natural gases accumulated in the large fields of Northwest Germany, the Netherlands, and southern North Sea originate from the coal seams of the Pennsylvanian system [67]. The carbon-isotope ratios of petroleum gas also enable reliable estimates of the stratigraphic assignment of the corresponding source rocks [68]. So-called *chemofossils* enable conclusions both on the deposition medium and on the "thermal history" of the sediment. These geochemical fossils, also known as biomarkers, are molecular structures of organisms in the form of lipids (e.g., fats, waxes) and porphyrins, etc., that have undergone only small changes during their transformation to kerogen. The main features of their chemical structure are still present.

Geochemistry has also been concerned intensively with the *primary migration* of crude oils, which for a long time was difficult to interpret. It is assumed that pressure built up in the sediment due to the decomposition of kerogen which led to the formation of microcracks, through which the hydrocarbons could escape from the source rock. They are transported with low-molecular hydrocarbons as propellant in the direction of the pressure gradient to more permeable rocks. An important instrument of exploration in future could be the geochemical modeling of sedimentary basins. The amounts of petroleum which have formed in a basin, migrated, and finally concentrated in a reservoir are calculated using computer methods. This calculation is carried out in close collaboration with the geological development and not only takes into account the sedimentation and facies distribution, but also the settling and the related rise of temperature, as well as the uplift phases with erosion of reservoirs already formed. A three-dimensional ki-

netic model has attracted attention [69]. It is based on a set of differential equations for energy and mass balances that describe the course of various geological processes. The physical parameters, such as porosity and permeability, and the thermal sequences are selected from available lithological data and, as knowledge increases, can be adjusted iteratively.

1.1.4. Occurrences, Production, and Reserves

The petroleum occurrences are associated with the sedimentary basins of the Earth. Their distribution is in no way statistical. Even their tectonic classification is very diverse and shows that deposit formation is not associated with only certain types of basin (Fig. 5). The largest petroleum occurrences lie in basins that have subsided considerably for instance as a result of the continental drift – approximately from the Jurassic to the Middle Tertiary period – and, owing to a lack of contact with oxygen-rich waters of polar origin and to a simultaneously warm climate with high, oxygen-reducing water temperatures, have preserved large amounts of organic material in the anoxic medium. The reserves in older geological formations are relatively small. The reason may be that there were fairly small amounts of organic material in the shallow seas of the epicontinental region during the flooding by the worldwide transgressions. But the smaller reserves in older formations than those of the Mesozoic era may also be the result of thermal evolution (conversion of oil into gas), the destruction of the reservoirs or their sealing horizon by tectonic events and erosion. The sedimentary basins with prospects are nowadays distributed both over the continents and over the regions of the continental shelves covered by the seas (Fig. 6). The methods of exploration and development are in both cases the same, for offshore wells the means of transport and technical installations are adapted to the water medium. The difference lies in the greater costs for offshore wells and production installations.

Apart from the classical offshore fields in Lake Maracaibo (Venezuela) and in Baku (Caspian Sea) that were already developed in the 1930s with modified onshore installations, offshore exploration and production achieved increasing importance only in the 1960s. Technical developments then became necessary that enabled deposits to be developed even at a depth of more than 1000 m water. In 1983, on the west coast of France, a well was drilled to almost 2000 m at a water depth of 1714 m. Such depths of water cause difficulties for production, which then is only possible by means of an underwater completion. This means that the production equipment is totally installed on the ocean floor and must be maintained by divers. The deepest producing field is the Jolliet field in the Gulf of Mexico. It produces from a depth of 535 m of water and is equipped with a tension leg platform, i.e., a floating platform that is fixed to the ground by steel wire ropes. In 1992 the Snorre field in the Norwegian North Sea also went into production by means of a tension leg platform at a depth of up to 350 m of water. Further production installations according to the tension leg system are being prepared up to ca. 1000 m (Auger and Marlin fields). Owing to high investment costs, the systematic recovery of hydrocarbons at great depths of water, in particular on the continental slope, depends on the development of the market price of crude oil.

1.1.4.1. Production and Reserves

The prices of crude oil collapsed in the second half of the 1980s. OPEC saw a possibility of stabilizing the prices only by fixing production quotas. These efforts, however, had no particular success. The control of production by OPEC as well as other economic measures in producing countries outside OPEC, influence production in such a way that the annual production data do not give a clear picture of the productivity of a producing country. Rather, the cumulative production gives only an impression of the geological and technical capacities (Table 1). For better comparison, the annual production for 1989 for the region concerned is indicated. World oil production for the year 1989 was 3.434×10^9 t. In 20 years, it rose only by ca. 20 %. In view of worldwide industrial development this is only a moderate increase that is attributable in part to energy saving efforts which started with the 1973/1974 energy crisis. The ranking of the producing countries for 1989 shows that the Middle East states Iran,

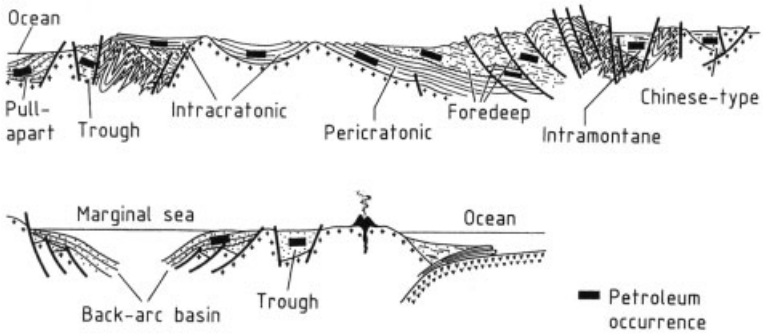

Figure 5. Geotectonic classification of petroleum occurrences

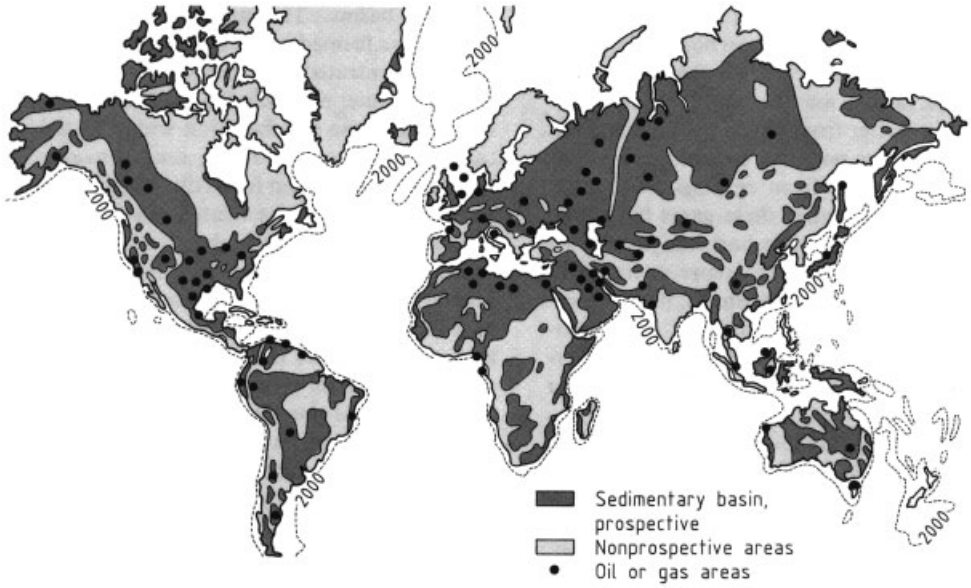

Figure 6. The sedimentary basins of the Earth with the most important production areas

Iraq, Kuwait, Saudi Arabia, and the United Arab Emirates lead the list of producing countries with a production of 922×10^6 t, followed by the Soviet Union with 668×10^6 t, and the United States and Canada together with 525×10^6 t. Following behind are South and Central America (380×10^6 t), Asia/Oceania (365×10^6 t), Africa (350×10^6 t), and Europe (223×10^6 t). The production from offshore fields was about 25 % of world oil production [70]. The proved world oil reserves rose to ca. 164.30×10^9 t through development, in particular, of offshore fields. Compared with reserves published in the 1970s (92.40×10^9 t) this is an increase of >80 %. If the probable reserves determined according to the Gaussian distribution curve are added, then there are currently 240×10^9 t petroleum to be expected. Compared with the production of ca. 3.4×10^9 t in 1990, this gives a ratio of 72 : 1. Thus the ratio of 30 : 1 existing for decades has more than doubled. Even taking only the proved reserves into account, the ratio is about 50 : 1. There may be various reasons for this. One could be the rapid sequence of discovery and development of large deposits – above all offshore. These include new discoveries in East Venezuela and in the Brazilian Campos Basin, as well as in the Russian North Caspian Basin and in Saudi Arabia. Large reserves have recently been developed in the Tarim Basin (West China) and in the Papua Basin of New Guinea, in which up until now no eco-

nomic oil discoveries have been known. Another reason is certainly the nature of the recording or evaluation of the proved reserves. The new assessment is based on the newly proven or re-evaluated reserves in the oil fields, larger extent of the deposits, or improved oil recovery. At the World Petroleum Congress in 1990, a rise of the reserves by 45×10^9 t to more than 164×10^9 t, was reported, i.e., a rise of ca. 33 %. Of this, ca. 32×10^9 t are apportionable to the Middle East and only 3×10^9 t to North and South America. Other authors quote a larger or smaller rise than this. The reason for this may be that, in addition to the newly developed reserves, new evaluations of the fields have been made in which, apart from the proved reserves, probable and possible reserves have been accounted for. The reserves according to MASTERS et al. [71] listed in Table 1 represent, despite all uncertainties, a reliable order of magnitude. The size of the increase in reserves from new discoveries also reflects the number of successful exploration wells. As stated above, the saturation of the petroleum market since 1985 has led to a drastic reduction of exploration activity. If in 1984 in the western countries, 1250 seismic crews and 6500 drilling rigs were in use, their numbers had fallen in 1989 to 447 and ca. 1924, respectively (1990: 2057, 1991: 1896). A speeding up of exploration depends on the general improvement and stability of prices and availability of reserves in the important consumer centers. The further prospects for success in petroleum exploration are limited to where the assumed profitability for foreseeable future ends or when the natural deposit potential of liquid hydrocarbons in the subsoil of the earth is exhausted. The two aspects are closely interlinked, since the estimation of potential reserves and even the definition of a deposit are determined by profitability factors. The ca. 77×10^9 t of conventional crude oil still to be expected according to the most recent estimates must be supplemented by the reserves from very heavy oil deposits and oil (tar) sands (→ Tar Sands), which can be assumed to be at least 100×10^9 t. An accurate demarcation from the "conventional" crude oils listed in Table 1 ist not possible, since MASTERS et al. [71] also take into account the reserves between 10° and 20° API (for a definition of ° API, see → Oil Refining, Chap. 7.1.), which other authors include in the low-quality heavy oils. The heavy oil and bitumen deposits are found to the extent of 87 % in Canada (Athabasca, Cold Lake, Peace River), the CIS (Volga – Urals, East Siberia), and Venezuela (Orinoco). They represent a future reserve that can only be used when, because of the depletion of the conventional fields, so-called tertiary recovery methods are standard, e.g., hot water injection or steam injection, to extract the oil from the deposits. The prices by then must have exceeded the limit of $ 50/bbl (the oil price in 1992 was ca. $ 20/bbl).

1.1.4.2. Regional Distribution of the Reserves

1.1.4.2.1. Middle East – Arabian (Persian) Gulf

The largest reserves are in the countries of the Middle East and in the Arabian Gulf. On 1st January, 1990, proved reserves of ca. 99×10^9 t were established and probable ones of 19×10^9 t calculated (see Table 1). The first economic oil find in recent times in this region was made in 1908 with the discovery of the Iranian Masjid-i-Sulaiman field. Many subsequent discoveries have been added. At present, Saudi Arabia, with more than 41.5×10^9 t of proved reserves, followed by Iran with 11.75×10^9 t, Kuwait with ca. 13.7×10^9 t, and Iraq with ca. 16.85×10^9 t are leading. The largest deposits in the Middle East include the fields of Burgan in Kuwait, Ghawar, and Safaniya in Saudi Arabia, Kirkuk and Rumaila in Iraq, and the deposits of Gach Saran and Agha Jari in Iran.

Geologically, the oilfields of the Arabian Gulf are situated on the shelf area in front of the pre-Cambrian massif outcropping in the west, and south of the Arabian peninsula in the form of intensively block-folded crystalline. Continental and shallow-water sediments are found deposited between the Cambrian and the Cretaceous systems on the slowly sinking shelf. The shelf is divided by numerous North – South trends that are linked with basement highs. North and east of the shelf, including the Taurus – Zagros – Oman mountains there follows a geosynclinal or plate-edge region. This region also received thick tertiary sediments and during alpidic orogenesis was folded by tangential forces and divided into structures with

Table 1. The most important crude oil producing countries; their total production up to December 31, 1989 and their resources (assembled from various international data)

	Cumulative production, 10^6 t	Resources	
		Proven and probable, 10^6 t	Prospective, 10^6 t
Canada	2 353	2 050	7 950
United States	24 722	7 200	7 280
North America	**27 075**	**9 250**	**15 230**
Argentina	795	350	475
Brazil	429	1 700	1 625
Columbia	477	650	930
Ecuador	238	400	415
Mexico	2 639	7 100	3 975
Trinidad	397	250	160
Venezuela	6 693	4 800	4 130
Others	397	400	750
Latin America	**12 065**	**15 565**	**12 460**
Bahrein	127	30	
Iran	5898	11 750	3 020
Iraq	3 323	16 850	5 560
Kuwait	3 800	13 700	320
Neutral Zone	713	2 200	320
Oman	493	1 100	150
Qatar	684	650	
Saudi Arabia	9 157	41 500	7 950
Syria	207	300	
United Arab Emirates	1 860	10 500	800
Others	207	750	
Middle East	**24 469**	**99 330**	**18 120**
Algeria	1 510	1 450	250
Angola	238	450	250
Egypt	747	900	300
Gabon	223	225	200
Libya	2 607	4 908	950
Nigeria	2 067	2 800	1 120
Others	29	650	1 830
Africa	**7 421**	**11 375**	**4 900**
Australia	509	475	325
Burma	79	30	150
China	2 226	5 000	5 100
India	445	675	400
Malaysia	668	1 200	700
Indonesia	2 194	2 100	1 250
Others	32	100	825
Far East	**6 153**	**9 580**	**8 750**
Norway	572	2 500	1065
United Kingdom	1 463	2 800	1 175
Other Western Europe	699	400	300
Eastern Europe	1 097	300	270
Europe	**3 831**	**6 000**	**2 900**

Table 1. (Continued)

	Cumulative production, 10^6 t	Resources	
		Proven and probable, 10^6 t	Prospective, 10^6 t
Soviet Union	16 630	13 250	14 625
World 1989	9 764	164 300	76 985
1990		135 000	
1991		134 630	

a northwest – southeast direction. In accordance with the asymmetric structure of this trough, the large deposits on the Arabian side mainly lie in Jurassic and Cretaceous beds, and the offshore deposits mainly in Cretaceous reservoir rocks. The Persian – Iraqi fields produce in the more deeply settled parts of the trough from Cretaceous and Tertiary horizons. On the Arabian peninsula, anticline-like updomings dominate structurally, whilst in the Persian Gulf salt structures and bending folds, and in the foothills of the Iranian Zagros range anticlinal and overthrust structures can be found. The most recent exploration activities in Saudi Arabia proved deposits with several hundred million tonnes of oil in sandstones of the Paleozoic era. The Permian Khuff, itself a reservoir bed of natural gas deposits, serves here as capping of stratigraphic and structural traps. The oil is of Silurian origin. The Paleozoic era could provide further potential in Saudi Arabia.

1.1.4.2.2. Europe

The petroleum deposits of Europe, apart from the North Sea, are fully developed and, with production falling, are already in the phase of secondary and tertiary recovery methods. If the reserves of all onshore producing countries are combined, proved reserves of ca. 0.7×10^9 t can be established together with exploratory reserves of ca. 0.5×10^9 t. The United Kingdom and Norway possess proved reserves of ca. 5.325×10^9 t from offshore fields, and a further ca. 2.34×10^9 t exploratory reserves are expected. The comparison with the Middle East, however, shows the small importance of the European states as oil-producing countries. Even so, in 1974 the Federal Republic of Germany produced more than 6×10^6 t oil and was able to cover ca. 30 % of indigenous requirements. Of the 823×10^6 t oil reserves originally present, 61×10^6 t proved and probable oil reserves remained after a total production of 225×10^6 t (December 31, 1991). The recovery factor of 35 % of the original reserves means that 537×10^6 t remain unrecoverable. The reserves in Italy and the Netherlands, with developed quantities of ca. 110×10^6 t in each case and prospective reserves of ca. 330×10^6 t in Italy, are – on a worldwide view – of equally low importance as those of Romania, with 190×10^6 t and a prospective reserve of 235×10^6 t. All other countries on the European mainland are globally considered insignificant as producing countries. The oil deposits of northern Germany are bound to sediments of epicontinental floodings which occurred during the Jurassic and Cretaceous periods. The narrow NNW – SSE striking troughs filled with Jurassic deposits are caused by halokinetic movements. The Rhine Valley Graben as a minor oil region dates from the "rifting" (continental drift) related-to young tectonic movements. The European deposits are distributed over the foredeeps of the Pyrenees, the Alps, and the Carpathians, although the deposits in the Aquitaine Basin, the Molasse Basin (Bavaria), the Carpathian Basin, and the Po Basin are not very large. The oil occurrences in Sicily, Spain, and in the Aquitaine Basin of France are linked with Triassic, Jurassic, and Cretaceous carbonates. The intramontane Pannonian Basin and the Paris Basin should be mentioned because of their historic importance for the exploitation of oil in Europe. The Lower-Saxony Basin also is of historic importance, since the deposits of Ölheim (1881) and Wietze (1859) were discovered here almost simultaneously with Colonel Drake's success in the United States.

North Sea. The North Sea has been one of the most attractive exploration regions for the international oil industry since the start of exploration in 1964. All large oil companies, often together with recently founded local partners, have introduced their worldwide know-how into development of this logistically difficult region. Great depths of water, strong currents, and extremely harsh weather conditions require the use of specially developed drilling and production installations. Although the technology used can nowadays be regarded as advanced, the damage to the platforms during construction and operation shows the limits for the prospecting and production of hydrocarbons. The first success in the North Sea was the discovery of the Ekofisk field on the Norwegian Shelf in 1968. Two important successes in the British sector of the North Sea were the discovery of Forties in 1970 and of Brent in 1971. About 17 oil fields are in production in the Norwegian North Sea, and in the British North Sea more than 100 oil and gas fields including numerous small ones. Norway's largest fields are Ekofisk/Eldfisk, Gulfaks, and Statfjord, which have up to now provided ca. 80 % of the Norwegian oil production. In the United Kingdom the Forties, Brent, Magnus, Ninian, Fulmar, and Statfjord (British part) fields provide more than 40 % of the British oil production. The five Danish and seven Dutch offshore fields are negligible compared with those of Norway and the United Kingdom. Geologically, the fields are linked with the large graben that open up from the beginning of the Mesozoic era in conjunction with the rifting of the Atlantic plate. The most important element is the Viking Graben which collapsed between the East Shetland Platform and the Norwegian Vestland Bogen. It runs for a length of ca. 455 km and has a width of ca. 50 km in the south and 180 km in the north approximately on the midline between Norway and the United Kingdom. In addition, there are special basins, some with thick tertiary sediments, to which the oil is bound. The deposits are found in sediments of the Triassic, the Jurassic, the Upper Cretaceous, and the Paleogene periods. The traps are mostly the tilted fault blocks typical of rifting, that are sealed by transgressively overlapping fine-grained Upper Cretaceous sediments. The development of hydrocarbons is limited to the north by the 62nd degree of latitude as a result of Norwegian legislation. Seismic measurements and specifically located wildcats have provided evidence that in the northern North Sea (Tromsø, Haltenbanken) and in the Barents Sea, deposits both of oil and of natural gas can be expected. The time for specific development of these fields will depend on the development of the reserves already developed and also on the oil price. Even today, owing to expensive technology, a price in excess of $ 20/bbl and tax relief by the producing countries is necessary for the continuation of exploration in the North Sea. In 1991, the daily production of all North Sea fields reached 192×10^6 t.

1.1.4.2.3. Commenwealth of Independent States

Industrial oil recovery in the territory of the CIS had already started at Baku in the middle of the last century and reached its first peak around 1900, when North American production was briefly surpassed. After a period of stagnation, a rapid upward development started in the former Soviet Union after World War II with the development of new oil reserves. The proved resources on January 1, 1990, were 13.2×10^9 t, and 14.6×10^9 t of exploratory reserves were expected. Accordingly, the total reserves are more than 200 % higher than those of the United States. Until 1918, almost the whole production came from the Caucasus mountains. In the 1950s the Volga – Urals region and the Timan – Petschora Basin, adjacent to the north, were developed. In addition, the discoveries in the North Caspian Basin (Emba) and in the Mangyschlak trough of Turkmenia developed very rapidly, but have passed their peak. In the 1960s and 1970s, the extended deposits east of the Urals in the West Siberian Basin were discovered. Meanwhile 96 % of the oil developed in West Siberia is produced in the Tjumen province. In the southern provinces of Novosibirsk and Omsk, exploration has given disappointing results. In almost all oil provinces, as a result of technical deficiencies and lack of oil field equipment, production is falling or remaining below expectations. Proved reserves could be mobilized by the use of modern technology, such as the horizontal drilling process in many reservoirs of low permeability. The opening of the CIS to the West leads to the expectation of a

strong involvement of international oil companies which, with their know-how and their modern technology can bring oil exploration and exploitation to a reliable high standard. There are also justified expectations of the development of important new reserves in Turkmenia and East Siberia, where only a small production of oil has been achieved so far. Large oil discoveries in the pre-Caspian Depression promise interesting exploration possibilities. In the recently discovered Thenguise, Astrakhanskoe, and Karatshaganakskoe fields, oil is produced from carbonates and clastics of the Paleozoic era at depths of more than 4000 m below a Permian salt cover. With an area of 500 000 km^2 and a sedimentation thickness of almost 20 000 m, good possibilities also arise here which may increase the number of reservoirs in the CIS substantially. Geologically, the deposits in the Caucasus mountains, the Caspian Sea, Galicia, the Fergana depression, and on Sachalin, etc., are linked above all with Neogen horizons in prealpine depressions and innermontane troughs. The deposits of the Volga – Ural area lie in Paleozoic depressions of the Eastern – European Platform. The main reservoir rocks here are Devonian and Permian reef limestones as well as sandstones of the Carboniferous period. In the large and oil-rich Timan – Pechora Basin, the reservoir systems belong to Upper Devonian sandstones and chalks of the Lower Carboniferous period. Mainly Permian and Mesozoic reservoir horizons are oil-bearing in the deposits of the North Caspian Depression, in Kazakhstan, and Turkmenia. In the deposits of the Western Siberian Platform, the hydrocarbons are assigned to the central part of the Lower or Middle Cretaceous period. In the Molasse Basin of the eastern Ural foreland, oil and gas deposits are linked with Upper Jurassic sandstones and calcarenites.

1.1.4.2.4. Africa

In 1990, the proved and probable oil reserves of Africa were about 11.3×10^9 t, and the exploratory ones ca. 5.0×10^9 t. The resources are concentrated in Libya, Algeria, and Nigeria. There are further deposits of noteworthy size in Egypt, Congo, Angola, Gabon, and Tunisia. The African continent was long considered to have few prospects for hydrocarbons. The breakthrough came in the 1950s with the almost simultaneous discoveries of the first large deposits in Algeria and the Niger delta. The large discoveries in Libya followed in the 1960s. Geologically, the oil deposits of Algeria and the smaller occurrences in West Libya which are not yet producing, are linked to the Paleozoic basin between the Precambrian Massifs of the mid-Sahara and the Atlas range. Fairly large amounts of oil are found in the Neogene basins of the Syrte in Central Libya and in the Niger delta. There are also older oil-producing strata, mostly in the Cretaceous system. At the western edges of the African shield, between several stable cratons, depression fields have been formed, on which marine transgressions deposited thick sediments in the Mesozoic and Tertiary eras. In some cases, between the Permo-Triassic system in Morocco and the Cretaceous system in Gabon, salt deposits occurred that led, possibly in conjunction with the continental drift, to diapirism. The basins not only cover parts of the current mainland but reach far into the Atlantic. The coastal basins range as follows from north to south: Spanish Sahara, Senegal – Mauritania, Ivory Coast, Dahomey, Nigeria, Gabon, Congo, and Angola. In Eastern Africa, of the graben systems which are largely linked to the rifting, only the Red Sea graben with Miocene salt deposits and hydrocarbons below is of importance, as in the Gulf of Suez, which is the only area of East Africa with fairly large reserves.

Libya, with 4.9×10^9 t proved and probable reserves and ca. 1×10^9 t prospective, has the largest oil reserves of Africa. Cumulatively, ca. 2.6×10^9 t oil has been produced in Libya up to now. Geologically, Libya is divided into three large basins: in the west the Murzuk basin, in the central part the Syrte basin, and in the east the Kufra Basin. The oil originates for the greater part from reservoir rocks of the Cretaceous (sandstones) and the Lower Tertiary (reef limestones) systems, but also from the Cambrian and Ordovician systems of the Syrte Basin. Libya's only producing offshore field Bour, with a daily production of ca. 13 000 t (1991), has little importance. The development of hydrocarbons in North Africa began with the discovery of the Zelten (now Nasser) field in 1959. Nasser produces from Lower Eocene limestones of the

Syrte Basin. With original reserves of about 400×10^6 t, it belongs, like Amal and Sarir, with ca. 1.2×10^9 t, to the largest fields of Libya. Libyan oil is low in sulfur and in great demand as a light crude in European refineries.

Nigeria is the second largest producing country in Africa. The proved and probable reserves are indicated as 2.8×10^9 t and the prospective reserves as 1.1×10^9 t oil. The fields are in the Niger Delta Basin that is filled with ca. 12 000 m of clastic rocks of the Tertiary period. In Nigeria, 15 giant oilfields (according to U.S. definition, fields with more than 500×10^6 bbl, ca. 80×10^6 t) have been proved, of which almost half lie offshore.

Algeria, with proved and probable reserves of ca. 1.4×10^9 t, occupies the third position among the producing countries of Africa. The prospective reserves (ca. 240×10^6 t) are quite insignificant. The oilfields lie in the Polignac Basin, near the Libyan border, and in the Ghadames Basin. The fields in Polignac produce from reservoirs in the Carboniferous, Devonian, and Cambro – Ordivician systems. In the Ghadames Basin likewise there are sandstones of the Cambro – Ordivician, that are discordantly overlaid with evaporites.

1.1.4.2.5. North America

Geologically, the North American oil deposits are linked with the large sedimentation region that appeared repeatedly since the Paleozoic era. This basin forms the foredeep to alpine-type mountains in the west (Rocky Mountains) and flattens to the east towards the basement area of the Canadian shield. In the south, this sedimentary basin expands and becomes the thick Tertiary Gulf basin. A lateral arm of the trough zone with Paleozoic rocks follows the southeast edge of the Canadian shield to the mouth of the St. Lawrence river. In accordance with this geological structure, the deposits in the more deeply sunken western sections of the trough appear in Tertiary and Cretaceous horizons, and in the higher-lying eastern marginal areas (midcontinent, eastern states with Appalachians) in sandstones and limestones of the Paleozoic era. The deeply sunken Gulf coast has its main deposits in Cretaceous and Tertiary strata. In the Californian fields, the oil comes from Tertiary and Cretaceous horizons. The new discoveries in Alaska occur mainly in Permo-Triassic sandstones. Other oil discoveries, scarcely developed until now, in the arctic expanse of Canada indicate several very promising basins with possibilities of reservoirs from the Paleozoic era to the Tertiary period.

United States. On January 1, 1990, the United States had available proved reserves of ca. 7.2×10^9 t and prospective reserves of ca. 7.3×10^9 t. Following the producing countries of the Middle East and the CIS, the United States are in the third position. The same ranking applies to the production in the year 1990. The cumulative production of 24.7×10^9 t crude oil was achieved from 597 320 production wells. (The Middle East states produced ca. 26.4×10^9 t from only 7061 wells). The oil resources in the offshore region of Texas, Louisiana, and California were about one quarter of the total reserves. Due to results of exploration in the Gulf of Mexico an increase of the offshore reserves can be anticipated from new discoveries in the deep water region, but this will scarcely lead to compensation for the fall in reserves determined by production. Before Texas, with the coastal strip on the Gulf of Mexico, advanced in 1930 to become the leading petroleum province, the position of the leading producer was occupied by Pennsylvania until the turn of the century, then by California, and later by the midcontinent. On January 1, 1990, the order according to daily production was Texas (ca. 306 179 t), Alaska (281 948 t), Louisiana (171 145 t), California (152 840 t), and Oklahoma (48 900 t).

Canada. Only after World War II did Canada rise to become an important producing country as a result of discoveries in Alberta. On January 1, 1990, the Canadian reserves amounted to ca. 2.1×10^9 t, 7.9×10^9 t of prospective reserves were reported. The daily production of ca. 190 000 t is obtained from 39 159 production wells. The reserves are declining. However, numerous discoveries in arctic regions cannot be developed and therefore cannot be evaluated. More than 80 % of the oil production occurs in Alberta. The hydrocarbons are present in the

foothills of the Rocky Mountains and in the Alberta Plain. The reservoirs for oil, as for natural gas, are present in highly stressed folds, mostly bound to overthrusts, in limestones of the Upper Carboniferous period, more rarely in beds of the Devonian or the Triassic periods. In the Alberta Plain, Paleozoic reef limestones are the reservoir rocks for the oil. Occasionally Cretaceous sandstones are the reservoir rocks, overlaying the Paleozoic discordant.

1.1.4.2.6. Latin America

The most important deposits in Latin America are those of Venezuela and Mexico; Trinidad, Ecuador, Argentina, Peru, Colombia, and Brazil follow some way behind. The main deposits are in the foretroughs of the cordilleras and their branches, and are bound to sediments from the Cretaceous to the Neogene periods. In addition, fairly small coastal basins in Brazil and Argentina are productive. The resource figures have increased due to new discoveries in Venezuela and more recently in Columbia but most of all to the oil accumulations developed in the jungle basins of Ecuador and Peru. Through the development of the Marina – Austral Basin between Tierra del Fuego and the Falkland Islands, the reserves of Argentina have increased considerably. The new reserves are not yet in production because of their geographical position and high gas : oil ratio. The high proportion of prospective reserves in Argentina of ca. 477×10^6 t, as against proved reserves of only ca. 350×10^6 t still points to a good exploration potential. As a result of the large oil deposits in the Bay of Maracaibo and on the Gulf Coast of Mexico, the submarine reserves of Latin America reach 40 % of the total resources. Further offshore deposits are off Trinidad and Brazil.

1.1.4.2.7. Far East

The most important producing regions of the Far East are China, Indonesia, Brunei, and Australia. In addition, oil deposits are known, for example, in Malaysia, Burma, Thailand, Japan, Taiwan, and New Zealand.

China. Only after World War II has China gained importance as a producing country. By the use of modern scientific techniques, the first small fields were discovered in the 1950s. In three regions, groups of fields had been proved by the end of the 1980s that originate from plate-tectonic movements. Up to 1980, 150 fields had been discovered. Of these, merely for reasons of size, only ca. 20 fields are of importance. Even so, the proved reserves have grown to ca. 5.0×10^9 t. The two largest groups of fields, Daqing in East China and Shengli on the Yellow Sea, achieved annual productions of 55.6×10^6 t and 33.5×10^6 t, respectively in 1990; that is ca. 64 % of the Chinese production. In the Daqing field, which was discovered in 1959, between 4000 and 5000 wells distributed over an area of 1200 km^2 are producing from Lower Cretaceous sandstones in very shallow anticlinical deposits. Shengli produces in a group of more than 40 individual structures from Precambrian and Tertiary reservoir beds. Of the numerous sedimentary basins of China, only a few are fully developed. The Tarim Basin (Sinkiang) is expected to be particularly rich in reserves. This extremely large inland basin, of 560 000 km^2, has a sediment thickness of 18 000 m. Since 1983, after intensive geophysical preparation and the sinking of 200 wells – the deepest reaching 6000 m – large quantities of hydrocarbons have been proved in both the North Tarim High and in the center of the Takelmagan Desert. The reserves are estimated at 11×10^9 m^3 oil and almost 8×10^{18} m^3 natural gas. The productive formations are sandstones of the Triassic and Ordovician carbonates. Exploration off the coasts of China is operated as joint ventures with international oil companies. Starting from onshore discoveries, in the Bohai Sea, east of Beijing (Peking), a series of fields has been developed whose potential is estimated to be more than 150×10^6 t crude oil. In the occurrences discovered in the South China Sea, no important reserves are yet known. The crude oil of the Bohai Basin originated from tertiary sediments of various structural types, including stratigraphic traps.

Other Areas. Important new discoveries have been made in recent years, especially in the submarine areas of Indonesia, Brunei, and South Australia (Tasman Strait). Of the devel-

oped reserves of the Far East, offshore deposits could account for almost one third. The most important reservoir horizons are in the Tertiary system; Mesozoic and Paleozoic formations are secondary possibilities. Wide shelf areas off the northwest Australian Coast as well as on the Indonesian and Chinese Shelves are regarded as favorable for future exploration.

1.2. Natural Gas

Natural gas has gained considerable importance in the past decades. Enormous gas reservoirs in Siberia, the Middle East, Algeria, and in Europe, connected by pipeline systems with the consumer countries, contribute substantially to the energy supply. The deposits in North America are, for environmental reasons, an important energy source for households and industry because of the fuel's cleanliness. Natural gas deposits in tropical regions are at present still underprivileged, since this gas can only be used in the industrial countries after liquefaction and transportation by ship. This is economically possible only if the price level offers an appropriate margin for this treatment of natural gas. The composition of natural gas is extraordinarily diverse (see also → Natural Gas, Chap. 1.1.). Aside from the so-called "dry", i.e., pure methane gas, "wet" gases with varying proportions of higher hydrocarbons occur. Inert constituents, such as nitrogen and carbon dioxide, reduce the calorific value. Sulfur and mercury are pollutants when the natural gas is used and must be removed from it (→ Natural Gas, Chap. 2.3.1., → Natural Gas, Chap. 2.3.2.). In connection with this, the elementary sulfur removed from the gas secures its own economically significant market value. The proportions of helium in the nitrogen-rich natural gases are, in general, small and only in a few cases do they allow extraction for economic utilization (→ Noble Gases, → Noble Gases, Chap. 4.2.1.). The natural gases arising in the limestones and dolomites of the Northwest German Zechstein contain up to 20 vol % hydrogen sulfide. The natural gas in the French Lacq field contains 15 vol % H_2S; while 63 % and 88 % are known from reservoirs in Wyoming in the United States and in Alberta, Canada (→ Hydrogen Sulfide, Chap. 4.2.). Since the gases containing hydrogen sulfide usually occur in contact with gypsum or anhydrite beds, formation of hydrogen sulfide is ascribed to bacterial reduction of calcium sulfate. Natural gas deposits with considerable percentages of heavy hydrocarbons are known as condensate deposits. While wet natural gases and condensates mainly lie adjacent to oil deposits or are associated with them, dry natural gases are usually found in separate gas provinces. This already follows from the fact that the majority of dry natural gases can be derived from the kerogen III arising from terrestrial plants. Other dry, natural gases are formed in parts of deep basins where pressure and temperature have exceeded the threshold value at which crude oils form. A distinction is made between *associated gas,* which is closely connected with crude oil, and *nonassociated or natural gas,* which is separately generated, predominantly from kerogen III. The associated gas dissolved in the crude oil is released during production by pressure reduction. If the crude oil is supersaturated, then part of the gas migrates upwards and usually forms a gas cap, which under certain conditions, can be exploited separately.

The *nonassociated gas* stands for deposits of dry natural gases that have been formed from kerogen III. This is methane gas (CH_4) being produced from beds of the Upper Carboniferous, the Rotliegendes, and the Zechstein systems in the British North sea, the Netherlands, Germany, and in Poland. It is beyond doubt that this gas originates from the solid coal seams below the reservoirs or from vegetable residues dispersed in sediments of the lower part of the Upper Carboniferous system. The "petroleum gas", as a byproduct of oil production, plays only a subsidiary role in energy supply – apart from exceptions such as Ekofisk. Large amounts of this gas are used in the production plants or, usually after compression, fed to regional gas supply systems. In low-pressure reservoirs, the petroleum gas is returned to the reservoir to raise the pressure which consequently reduces the production costs by a longer-lasting eruptive phase. In the large producing countries, for example on the Arabian Gulf and in North Africa, enormous quantities of petroleum gas are flared, thus being both wasteful and environmentally polluting.

Since ca. 1980, the theory of the abiogenic origin of the hydrocarbons in the ground has gained interest again. Gold postulates that the

vast quantities of bound (oxidized) carbon, e.g., in carbonates, cannot originate from the atmosphere but escape from the mantel of the Earth. The presence of helium, that originates from the radioactive decay of certain minerals, is also quoted by GOLD as evidence, as are methane inclusions in diamonds, in kimberlite, or the high methane content in Lake Kivu, which lies over a rift fault. The theory is based on the finding of a methane atmosphere for the outer planets and their moons. Evidence for methane of abiological origin was to be provided in the mid-1980s by a deep borehole in the so-called Siljan Ring in Sweden. From this enormous meteorite crater, oil and gas traces were reported from water wells and flammable gas was detected from lakes and pools. Gravimetric measurements showed an anomaly that was interpreted as a zone of good porosity in the granite, created by finely divided hair cracks and clefts due to impact by the meteorite. The borehole sunk to 6080 m from July 1986 to January 1987 and did not establish proof of abiogenic gas. German scientists intend to investigate the problem of the so-called mantle gas in collaboration with the natural gas industry and the federal ministry responsible.

As with abiogenic hydrocarbons from the Earth's mantle, the question of the origin of the nitrogen in natural gases, involving the theory of the inclusion of air in the Earth's crust, is still being discussed. More probable theories suggest an enrichment process of nitrogen as a result of fractional differentiation during gas migration, formation of nitrogen by the action of anaerobic bacteria during oil generation in the source rock, and its release from vegetable materials during coalification. For the origin of the Northwest German nitrogen enrichments, the possibilities under discussion include a relatively late split off from the Upper Carboniferous coal seams and a concentration by fractional differentiation on the migration paths.

1.2.1. Origin and Deposit Formation

Natural gases in the narrower sense, i.e., the combustible hydrocarbons, are often closely related to the crude oils, both genetically and above all with regard to deposit formation (see Sections 1.1.1 and 1.1.2). Petroleum and natural gas frequently occur in the same geological level of a region. On the other hand, coal and petroleum occurrences in the same level of a region are almost entirely mutually exclusive. Dry natural gas as a derivative of coalification – both in regions with thick coal seams as well as in those with dispersely distributed coaly material – is also often found directly adjacent to petroleum regions, e.g., in the Emsland and the Wiehengebirgsvorland of Northwest Germany. The majority of natural gas occurrences in the world is linked with oil provinces. Associated natural gases, often in connection with an economically uninteresting oil seam of low thickness (e.g., the enormous gas deposits in the north of the West Siberian Platform), predominate. As a result of fractional migration effects, accumulations of dry gas deposits can also occur.

Gas formation in petroleum source rocks proceeds in different ways. In the diagenesis phase, occurring in immature source rocks, bacterial methane is generated from kerogen III. At $> 60\,^{\circ}C$, the splitting-off of wet gas from kerogen I and II begins – in addition to oil formation – which continues to the lower end of the oil window at $150\,^{\circ}C$. At higher temperature dry thermal gas is increasingly formed. The natural gas deposits of the Rotliegendes and of the Zechstein in the natural gas belt between the south of the British North Sea and Poland owe their formation, as already mentioned, to the solid coal deposits in the underlying Upper Carboniferous system. The coal-bearing beds were probably subjected, already at the end of the Carboniferous period, to an early coalification which was so low that 30 % of the volatiles remained in the coal. As a result of the further sinking of the coal seams by thick Cretaceous and Tertiary overlying strata, the temperature required for natural gas formation rose again and the coalification process continued with release of further methane. The majority of the gas deposits arising between England and Poland, including the Groningen deposit, owe their formation to such processes.

Theoretically, a further cause for the formation of natural gas is its genesis from oil shales. As in the technical recovery of gas from oil shales, including the Kukersit shales of Estonia (→ Oil Shale, Chap. 7.1.), natural pyrolysis

processes at temperatures far above 200 °C are possible, by which natural gas could have been formed from oil shales.

1.2.2. Natural Gas Occurrences and Their Reserves

The importance of the natural gas producing countries is reflected in their present reserves. The amounts produced annually have increased as a result of the consolidation of the international pipeline network. In addition, interest in natural gas has increased because it is the cleanest of the fossil fuels; it is much less polluting than coal or oil. Large reserves also justify the high investment for the treatment and pipeline transport of the natural gas (→ Natural Gas, Chap. 4.1.). The production quantities reported are not a measure of the natural gas potential available in the producing countries, which follows from the growth of the reserves. This is evident on comparing the United States and the CIS In 1989, production in the United States was 500×10^9 m^3, and that of the CIS 739×10^9 m^3. On January 1, 1990, the proved and probable reserves of the United States were ca. 8.4×10^{12} m^3 and those of the CIS ca. 45×10^{12} m^3. To these there must be added exploratory reserves ca. 11×10^{12} m^3 for the United States and ca. 35×10^{12} m^3 for the CIS The disproportion between production and reserves in the CIS originates in missing pipeline connections, line losses, and also marketing problems. Other large natural gas producers are Canada, the Netherlands, Iran, Venezuela, the United Kingdom, Romania, Mexico, China, and Germany. The large natural gas potentials in the countries of the Arabian Gulf, in Nigeria, and in Indonesia suffer from the distance of the markets, since a basic requirement for transportation from remote countries to the trade centers is the liquefaction of the natural gas. Because of the relatively small capacity of the liquid gas tankers and the large distances to be covered, the economic use of liquid gas is often at or below the break-even point. The natural gas resources of the world – proved and probable – on January 1, 1990, were ca. 127.0×10^{12} m^3, and the prospective another ca. 121×10^{12} m^3. The CIS has developed its leading position through the enormous new discoveries, above all in Western Siberia, and is followed by the Middle East, which also has a very powerful growth to record. The raising of the proved gas reserves in these two large districts originates not only in new discoveries but partly in the extension or higher evaluation of deposits already known and the strengthened inclusion of petroleum gases, particularly in the Middle East and Venezuela. While Africa and North America show a declining trend for the proved reserves, Western Europe has improved its position in the international gas sector by further new discoveries in the North Sea.

1.2.3. Regional Distribution of the Reserves

A breakdown of natural gas production and resources according to region is given in Table 2.

1.2.3.1. Commonwealth of Independent States

The largest reserves in the world, at 45×10^{12} m^3, are those of the CIS The occurrences are widely spread and in some cases combined with petroleum deposits.

The first great gas boom was in the 1950s with the development of the fields of Shebelinka in the Dnieper – Donetz Trough and Gazli in Central Asia, each with primary resources of $450-500\times10^9$ m^3. The discoveries in the second half of the 1960s, some of which were enormous, established the present position of the CIS in the natural gas sector. Important deposits mentioned are those of Krasny – Kholm (Orenburg) in the Volga – Urals district, where ca. 750×10^9 m^3 of gas have been discovered, and the Vuktyl and Layavozh deposits in the Petchora Basin, linked with beds of the same age and each with ca. 500×10^9 m^3 of gas. In Turkmenia, the Malai field with ca. 340×10^9 m^3, is one of the giants. Probably the largest natural gas province in the world is in the north of the Western Siberian platform. The gas occurring here in sandstones of the Cretaceous system (predominantly Cenomanian), in part several hundred meters thick, consists of almost pure (97 – 99 %) methane. The structural traps are extended anticlines and mound-like highs. About 220 fields have been developed in this area up to now.

Table 2. Natural gas production and resources up to December 31, 1989 (assembled from various international data)

	Cumulative production, 10^9 m^3	Resources	
		Proven and probable, 10^9 m^3	Prospective, 10^9 m^3
North America	23 436	11 000	18 700
Latin America	1 734	7 000	10 600
Middle East	1 629	42 000	26 700
Africa	924	7 400	9 300
Far East	2 306	9 800	13 700
Europe	4 970	5 800	7 200
CIS	10 150	44 000	34 800
World	45 149	127 000	121 000

This producing province is the most important for the future export of gas to Europe. Its most extensive deposits include the Urengoy field, the largest deposit proved so far with ca. 6×10^{12} m^3 of initial gas reserves, and also the large deposits of Jubileyny (ca. 2×10^{12} m^3), Taz (ca. 1.15×10^{12} m^3), Medvezhye (ca. 1.05 $\times10^{12}$ m^3), and Yamburg (ca. 850×10^9 m^3) among others. At least a further ten fields of this region have each reported primarily more than 100×10^9 m^3 of recoverable gas reserves. Two new discoveries in the Arctic Ocean deserve attention. In 1987 in the Barents Sea, not far from the Norwegian border, the Shtowmanowskaja field with an estimated reserve of $3-4\times10^{12}$ m^3 was discovered, and in 1988 in the Kara Sea, the Russanowskaja field with 6×10^{12} m^3. Both fields together are credited with 15 % of the future production provided that the production and transportation to Nowaja Semlja is possible under arctic conditions.

1.2.3.2. Middle East

Even if not so dominant as in the case of crude oil, the areas on the Arabian Gulf, with proved natural gas reserves of ca. 68×10^{12} m^3 play an important role in the possibility of future exports. However, a considerable fraction of these quantities is trapped in gas caps or is dissolved petroleum gas linked with the large oilfields of this area and is therefore dependent on their production. The gas cap of the Pazanan oilfield in Iran alone contains recoverable quantities of ca. 1.4×10^{12} m^3 of gas. With the recent discovery of enormous gas accumulations in the Permian Khuff formation, very large dry gas deposits have also been proved for the Middle East. The Kangan field in Iran is said to contain 4×10^{12} m^3 of resources. The deposits available in the Permian formation are sealed above by triadic salt beds. There are further deposits of this type, not hitherto exploited, in Qatar. Iran also is involved in the northern part of the Northwest Dome (North Field) – with $2.8-3.4\ \times10^{12}$ m^3 a "super giant" – which belongs to the Permian Khuff. Notable natural gas reserves are also known in Saudi Arabia and the Emirates.

1.2.3.3. North America

Despite higher quantities produced, the known North American gas reserves have risen again. The reserves of 9×10^{12} m^3 in the mid 1970s have risen to ca. 11×10^{12} m^3 proved and probable. Of these, 8.4×10^{12} m^3 in the mid 1980s were assigned to the United States alone, the traditional land of natural gas. Its large-scale industrial use started here even before the turn of the century. In addition to this, exploratory reserves of almost 20×10^{12} m^3 can be anticipated in North America. The largest deposits in the United States are in Texas, Oklahoma (particularly in the deep Anadarco Basin), Kansas, and New Mexico with the fields of Hogoton (initial reserves ca. 1.1×10^{12} m^3) and Panhandle (primary resources ca. 870×10^9 m^3), which are already in part greatly depleted. A large fraction of the resources are in the Rocky Mountain Province with 1.25×10^{12} m^3. The natural gas here is predominantly bound to Permian carbonate rocks, but is also found in sandstones, arkoses, and weathered granite zones. Other deposits, such as the Blanco Basin field in New

Mexico and Rio Vista in California, occur in Cretaceous sandstones and Lower Paleozoic carbonate rocks (Pucket field, Texas).

The gas accumulations of the American Gulf Coast are mainly associated with oil deposits or are found in separate reservoirs over the oil in Tertiary sandstones. The recoverable gas contents of these latter isolated deposits are only between 100 and 300×10^9 m^3.

The North American gas deposits, especially those on the Gulf Coast, are frequently characterized by a relatively high nitrogen content.

1.2.3.4. Africa

The proximity of Algeria and Libya to the industrial countries of Europe favors the export of natural gas to Europe. Altogether, supply contracts for more than 80×10^9 m^3/a have been closed, of which 12×10^9 m^3 are conveyed by pipeline to Italy and the remainder, after liquefaction, by tanker mainly to Germany and a further part to the United States. In addition to the gas reserves of the oil-producing countries Libya and Algeria, that mostly occur as dissolved gas, there are dry gas deposits, particularly in Algeria. The largest African field, Hassi R'Mel in the Algerian Sahara, discovered in 1956, has been in production for years. The gas that contains 86 % methane occurs in triadic sandstones in the area of an anticline. The original resources were assessed as ca. 1×10^{12} m^3, but are probably more than double that amount.

1.2.3.5. Western Europe

The Western European deposits were of only local significance until the discovery of the large Groningen gas field in 1959/1960 and the subsequent discoveries in the North Sea. The small Western European gas production comes mainly from the French field, Lacq, in the Pyrenean foredeep, with primary reserves of ca. 200 $\times10^9$ m^3, as well as from deposits in Northwest Germany and the Netherlands where they are mostly bound to Zechstein dolomites and from Tertiary reservoir rocks in the German Alpine approaches. There are smaller production areas in the inner-alpine Vienna basin, in the Po plain, and off the Adriatic coast of Italy. The Austrian and Italian reservoirs occur mainly in Tertiary horizons. As a result of the development of the gas deposit of Groningen in the 250 m thick Rotliegendes sands, with initial reserves of ca. 1.9×10^{12} m^3 and a methane content of ca. 82 vol %, the exploration boom for inferred natural gas accumulations of the Rotliegendes in the North Sea and for the neighboring region of Germany started from 1963/1964. In the North Sea a distinction can be made between a dry gas province in the southern North Sea with deposits off the eastern English and Dutch coasts, and the gas deposits of the central and northern North Sea. The latter are mostly associated with oil deposits and linked with the Dan formation but, most of all to sandstones of the Paleogene system.

The deposits in the *southern North Sea* generally occur, as in the Groningen field, in the sandstones of the Rotliegendes. The methane gas, like most of the accumulations in Northwest Germany, is attributed to coalification processes in the underlying seam-bearing Upper Carboniferous system.

The natural gas reservoirs *further to the north* could be of similar origin, in the broadest sense, with the oil deposits proved there. These are partly gas cap gases and partly gas accumulations in separate reservoir horizons above oil deposits, or above economically uninteresting oil seams.

On January 1, 1992, ca. 4400×10^9 m^3 natural gas had been discovered and developed in the North Sea. Of this, ca. 1200×10^9 m^3 was assigned to the United Kingdom, ca. 2700×10^9 m^3 to Norway, ca. 350×10^9 m^3 to the Netherlands, and ca. 110×10^9 m^3 to Denmark. The results of drilling from the German North Sea give no indication of major natural gas deposits comparable with those of the neighboring sectors. Only one minor natural gas deposit in the extreme part of the German North Sea with reserves of ca. 10×10^9 m^3 is under development. Production, in the North Sea, mainly from British gas fields, increased from ca. 50×10^9 m^3 in 1985 to 98×10^9 m^3 in 1991.

The largest known gas deposits in the North Sea include the following fields: Frigg (Paleocene sandstone, ca. 400×10^9 m^3 resources); Leman (Rotliegend sands, ca. 300×10^9 m^3 reserves); Indefatigable (Rotliegend sands, ca. 140×10^9 m^3); Viking (Rotliegend sands, ca.

140×10^9 m^3); L-10/11 (Rotliegend sands, ca. $100-140\times10^9$ m^3); K 13 (Rotliegend sands, ca. $100-140\times10^9$ m^3); Hewett (Triassic sands, ca. 110×10^9 m^3); Troll (Middle/Upper Jurassic, ca. 1600×10^9 m^3).

With continuous exploration on the Norwegian Shelf (Arctic Ocean and Barents Sea) further large natural gas discoveries can be expected.

2. Reservoir Engineering

2.1. Characterization of Oil and Gas Reservoirs

As described in Chapter 1, oil and gas reservoirs are accumulations of hydrocarbons that are stored in porous or fissured strata of the Earth's outer crust and contain a commercially valuable quantity of oil or gas. A precondition for storage in the sedimentary rocks is a continuous cavity, in the form of pores or fissures, that permits transport of the fluids into pressure sinks (wells). Since these sedimentary strata can be found at various depths, it is important that the rock matrix takes on the high rock compression stress without much deformation so that when pore pressure is released, the pore fluids can easily flow to the pressure sink, i.e., the well. If large deformations were to occur in the rock, the flow paths in the neighborhood of the well could close upon pressure release thus resulting in the virtual sealing of the reservoir.

2.1.1. Reservoir Rock

The most important reservoir-forming sedimentary rocks are described in Section 1.1.2.2. Since the reservoir rock rarely has a monomineralic character, its mineral composition and the internal surfaces resulting from it are very important, both for the storage capacity of the rock and for the transmissibility of the fluids stored within it. The crystal structure and specific internal surface of the rock-forming minerals determine the types of porosity of a reservoir rock.

In principle, cavities are found in a mineral structure along the cleavage planes of crystals (*intercrystalline pore space*), within the layer lattice of crystals (*intracrystalline pore space*), and, on a larger scale, along the boundary surfaces of particles (*interparticulate pore space*) or within the particle at mineral boundaries or growth defects (*intraparticulate pore space*). Hydrocarbons are stored mainly in the interparticulate pore space of clastic sediments or the *fissured* pore space of chemical sediments (limestone, dolomite). Fissures form in a rock as a result of deformation when the shear strength is exceeded.

Figure 7 shows the typical composition of a porous clastic sediment (detrital rock). The rock body consists of sediment particles (grain) that are mainly composed of mineral components quartz, feldspar, and mica; of bioclasts; or oolites. In most cases the rock particles are cemented together by calcites, dolomites, silicates, or clay minerals. In the interparticulate pore space so formed, pore fillings of clay minerals, quartz, and mica can lower the available porosity.

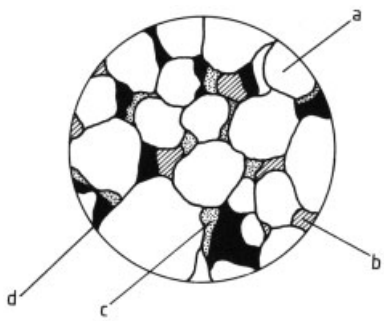

Figure 7. Composition of a detritral rock [72]
a) Grain (texture: sand granules 1/16 – 2 mm diameter; silt 1/256 – 1/16 mm diameter); b) Pore fillings (texture: sand granules 1/16 mm diameter; silt 1/16 – 1/256 mm diameter; clay $<$ 1/256 mm); c) Cement; d) Fluid

2.1.1.1. Porosity

Porosity ϕ defines the ratio of the pore volume V_p of a reservoir rock to the bulk volume V_b of the rock:

$$\phi=\frac{V_\mathrm{p}}{V_\mathrm{b}}$$

According to this definition, the porosity also includes the volume of isolated pores (*dead pores*).

If all interconnected elements of the pore volume are taken into account, the *effective porosity*

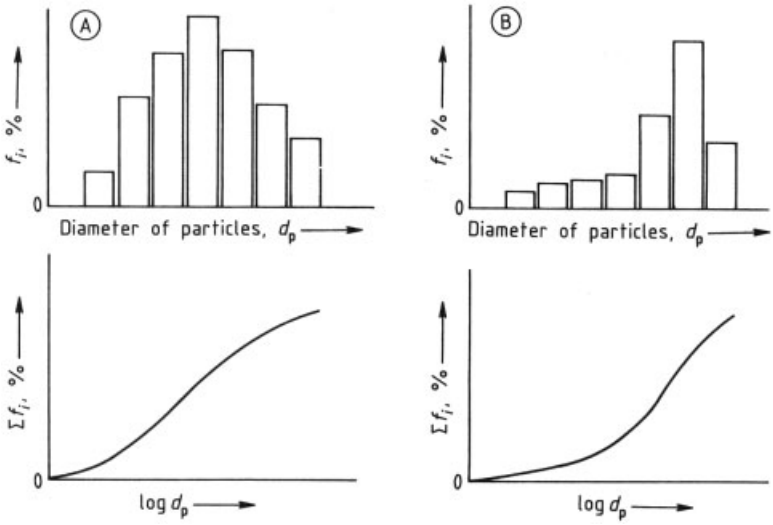

Figure 8. Grain-size distribution and porosity
A) Low porosity; B) High porosity

is obtained. This is an indicator of the effective storage volume of a rock. The sum of effective ϕ_{eff} and dead ϕ_{d} porosities gives the total porosity:

$$\phi_{\mathrm{eff}} + \phi_{\mathrm{d}} = \phi$$

For geometrical or hydrodynamical reasons, not all pores are available for convective mass transport. A fixed portion of the cavity volume can be regarded as *dead-end* pore volume.

The porosity depends essentially on four characteristics of the sedimentary rock:

1) Particle shape
2) Degree of cementation
3) Particle size distribution
4) Particle arrangement

The influence of particle shape and degree of cementation is evident from Figure 7. In general, the more edged the particles, the greater is the interparticulate component of the pore space. The available porosity decreases with increasing cement component. The most important influencing factors, however, are the particle size distribution and the particle arrangement in the rock structure.

Figure 8 A shows a wide distribution of particle diameters which leads to a large reduction in porosity, since the interparticulate pore space formed by the coarse fractions is filled up by the fine components. On the other hand, a narrow particle size distribution (Fig. 8 B) results in a high porosity. In the lower half of the figure, the cumulative particle size distribution is also shown. Usually the steeper the cumulative curve, the greater is the porosity of the rock.

Three special cases of particle arrangements and the corresponding porosities are shown in Figure 9. In the cubic, space-centered sphere packing (Fig. 9 A), in which the particle midpoints form the corners of a cube, the interparticulate pore volume is exactly 47.6 % of the total volume. In this case each particle has six points of contact with the neighboring particles. In the hexagonal arrangement of particles (Fig. 9 B), the spherical particles of the upper lattice plane are displaced by half the particle diameter relative to the lower lattice plane. In this arrangement, the number of contacts rises from six to eight and the porosity decreases to 39.5 %. If in the second lattice plane of the packing, the row of spheres is displaced not only in the x-direction by $d/2$, but also in the y-direction perpendicular to it, by $d/2$, the densest packing of spheres is obtained (C). Here in the rhombohedral arrangement the particles of the upper lattice plane are placed exactly in the holes of the lower plane and accordingly each particle has twelve neighbors and the porosity is only 25.9 %. The porosity of a

natural packing of cemented sedimentary grains usually falls between the limiting values of 39.5 and 25.9 %.

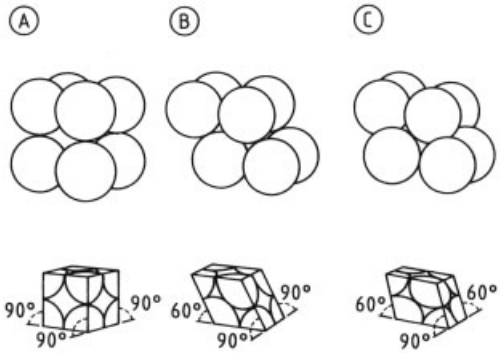

Figure 9. Packing and porosity [73]
A) Space-centered, cubic sphere packing (number of contacts = 6, $\phi = 47.6\,\%$); B) Hexagonal packing (number of contacts = 8, $\phi = 39.5\,\%$); C) Densest packing of spheres (rhombohedral) (number of contacts = 12, $\phi = 25.9\,\%$)

2.1.1.2. Pore Configuration Models

Although, at first glance, the shape of the pores or fissures of a sedimentary rock has very little to do with regular geometric forms, cylindrical capillaries or parallel plates are regarded as geometrically similar to the pores and fissures and are chosen as models for calculating the pore space properties. These very simple geometries are unable, however, to reproduce the irregular variation of the pore or fissure cross section or of the tortuosity of flow path, which are important characteristics of the pore configuration.

In Figure 10, the commonest types of pore configuration models, up to the most advanced representation (fractal network model), are shown. The *capillary bundle model* (Fig. 10 A) is the most frequently applied pore configuration model, mainly used for calculating the pore radius distribution, the saturation with fluids, and the oil displacement process. A particular drawback of this model is that a transverse flow between the capillaries is not possible. A pore configuration model for fissured reservoir rocks derived from the plate model is the *brickstone model* (B), in which the individual blocks symbolize the porous matrix and the gaps between them, the fissures. Here mass transfer between matrix and fissure is possible. The *intersphere network model* (C) is used particularly for determining the phase distribution in the pore space and the capillary forces. With the *cascade network model* (D), dead-end pore effects, turbulence effects, and mixing phenomena can be investigated.

According to the principle of self-similarity of geometric forms in nature, relations are traceable as far as between the crystal forms in the microcosmos and the geological outfits (morphology) of the Earth's surface. In connection with this, the fractal dimension is a scale-up factor between the characteristic dimensions and the image scales. If the spherical shape (circular shape in Fig. 10 E) is chosen as the characteristic elementary shape of the rock structure and of the pore space, a fractal network for matrix and pore configuration can be developed, whose specific internal surface and volume fraction are calculable and correlatable with other physically measurable quantities such as electrical conductivity and permeability.

An example of the use of pore configuration models for calculating pore space properties is given here, based on the determination of the pore radius distribution by the mercury injection method. If the total pore space V_p is assumed to be represented by a bundle of capillaries of various diameters, then on injection of a nonwetting liquid phase (i.e., mercury) the individual sizes of capillary are filled stepwise (ΔV) as the pressure is raised. Since the injection pressure for mercury p_c is inversely proportional to the radius of the capillaries, the proportions by volume $\Delta V/V_p$ of the individual capillary radii can be estimated from the mercury injection curve

$$p_c = \frac{2\sigma_{Hg}\cos\theta_{Hg}}{r}$$

where σ_{Hg} is the surface tension of mercury and θ the wetting angle of contact.

In order to reconcile this measurement with actual pore space configurations, it must be taken into account that each pore has at least two characteristic dimensions – the pore entrance and the pore widening. The characteristic cross-section determined by the mercury injection method corresponds to the pore entrance radius.

2.1.1.3. Compressibility of the Pore Space

The compressibility c of a substance is the change, with external pressure, of its volume

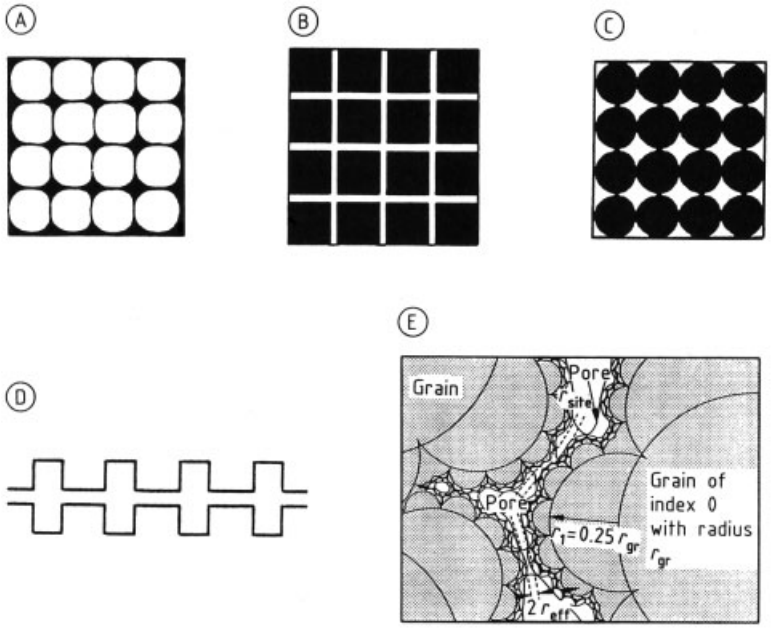

Figure 10. Pore configuration models [74]
A) Capillary bundle model; B) Brickstone model; C) Intersphere network model; D) Cascade network model; E) Fractal network model
White color means pore space.

relative to its initial volume. For a solid the compressibility is negligibly small. A porous medium which is built up of individual particles can, however, change its particle arrangement under pressure and, therefore, is said to have considerable compressibility. The deformability is discussed for the different forms of packing of spherical particles in Section 2.1.1.1. According to the studies of Geertsma [75], the pore volume compressibility c_f depends on the type of pore configuration (porosity of the medium) and the effective pressure which is the difference of external pressure p_e and internal pore pressure p_i.

$$c_f = (-)\,\frac{1}{V_p}\,\frac{\partial V_p}{\partial\,(p_e - p_i)}$$

Figure 11 shows that the pore volume change reacts with increasing sensitivity to an increase in the external pressure with decreasing porosity. The porosity of sedimentary rock under in situ conditions (i.e., under the influence of the overburden pressure of the sediments) can, therefore, be up to 20 % smaller than the porosity determined in the laboratory [77].

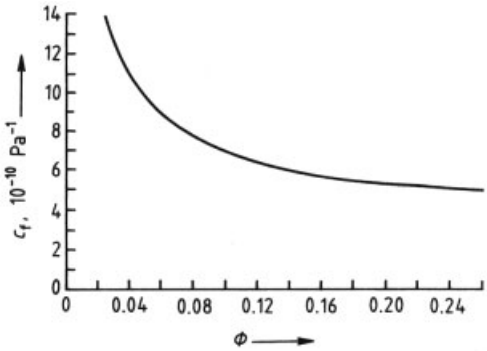

Figure 11. Compressibility of the pore volume [76]

2.1.2. Physical Properties of the Pore Saturating Fluids

Gas, oil, and water are the relevant pore saturating fluids, whose physical properties must be known as a function of the pressure, the reservoir temperature, and the fluid's composition. The material properties of these multicomponent systems can only be predicted for the gas state at low pressure with acceptable accuracy and without an extensive experimental data base.

For liquids and phase mixtures, estimates of the material properties are obtained in accordance with the theorem of corresponding states by normalizing to conditions at the critical point. The state variables, pressure, and temperature, are normalized by introducing the reduced pressure p_r and temperature T_r (→ Estimation of Physical Properties, Chap. 2.2.).

$$p_r = \frac{p}{p_c} \quad T_r = \frac{T}{T_c}$$

Mixtures with complex structure can be described by means of pseudocritical pressure and temperature that are calculated from the molar fractions of the individual components and the corresponding individual pseudocritical quantities. These simple correlations yield satisfactory results only when the individual components have similar critical pressures and critical volumes.

In all other cases three-parameter thermodynamic functions of state are used to improve the accuracy of prediction. These calculation methods are based on the attempt to produce a functional relationship between the critical material properties of a component and the molecular interactions. Nevertheless, no known equation of state is able to determine all properties of hydrocarbon mixtures equally satisfactorily.

For determining the contents of a reservoir and for calculations of recovery and flow processes, the most important physical parameters are the density, compressibility, and viscosity of the fluids. In the following, the methods of calculation applied in practice will be considered without showing the full physical background. The reader is referred to the references [77], [78] given at the end of this article.

2.1.2.1. Density

The density of a *gas* ϱ_g can be calculated directly from the equation of state for real gases

$$\varrho_g = \frac{M_r p}{z R T}$$

where M_r is the molecular mass, p the pressure, z the real gas factor, R the gas constant, and T the temperature.

The reduced temperature and the reduced pressure serve as adjustment parameters for the real gas factor z. For gas mixtures, instead of the molecular mass M_r the mean molecular mass of the mixture is used together with the real gas factor for the mixture.

For direct calculation of *liquid densities,* this simple form of the equation is inadequate. Because of the higher packing density of the molecules in the liquid allowances must be made for the inherent volume of the molecules and for molecular interactions.

If the critical densities or the densities of the individual components under standard conditions are known, the density of the mixture ϱ_f can be determined from the molar fractions according to the rule of additive properties

$$\varrho_f = \frac{\sum_1^n z_j M_{rj}}{\sum_1^n \frac{z_j M_{rj}}{\varrho_{0j}}}$$

For reservoir calculations, the so- called *formation volume factor B* is usually used instead of density for converting gas and liquid volumes to reservoir or standard conditions. The formation volume factor is the ratio of the volume under reservoir conditions to the volume under standard conditions.

$$B = \frac{\text{Volume at pressure } p \text{ and temperature } T}{\text{Volume at standard conditions } (p_0, T_0)}$$

where the standard conditions for gases are: $p_0 = 0.1013$ MPa and T_0 273.15 K. For liquids $p_0 = 0.1013$ MPa and $T_0 = 293.15$ K holds.

In the United States, 60 °F, corresponding to 288.65 K, is used as standard temperature. The *formation volume factor for gas* B_g can be calculated directly from the real gas equation:

$$B_g = \frac{z T p_0}{p T_0}$$

From a double-logarithmic plot of the formation volume factor versus pressure, a linear curve is obtained as long as the influence of the real gas factor at elevated pressure is not noticeable (Fig. 12).

Figure 13 shows the formation volume factor for oil, B_o, as a function of pressure at constant temperature.

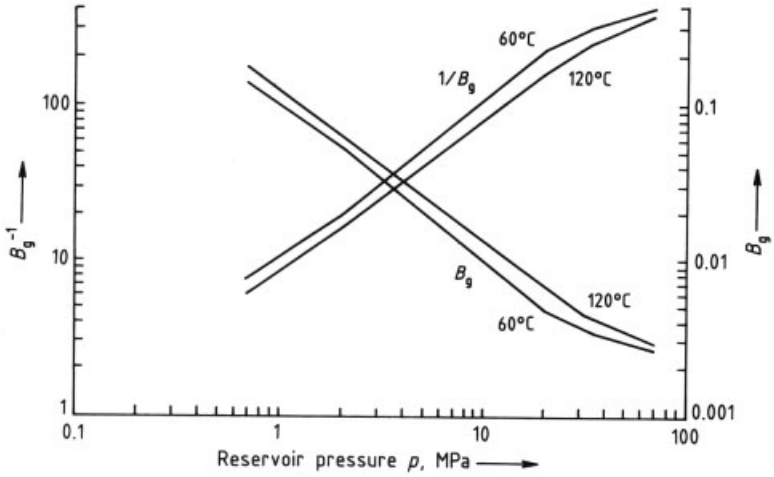

Figure 12. Formation volume factor for gas

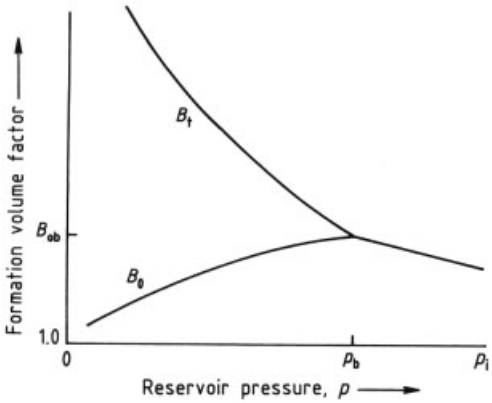

Figure 13. Formation volume factor for oil [78]

For a gas-undersaturated oil, the formation volume factor B_o at first increases with decreasing pressure as a result of the liquid-phase expansion of the oil and reaches its maximum value at the bubble point pressure (p_b). The greater the gradient of the branch of the curve between the initial reservoir pressure (p_i) and the bubble point pressure, the more compressible is the liquid. When the pressure falls below the bubble point pressure, the volume of the oil and with it the formation volume factor decreases rapidly. The final value of 1 is not reached, since the thermal shrinkage of the oil from reservoir-temperature to standard-temperature condition occurs at a final pressure of 0.1 MPa. To describe the volume behavior of a two-phase mixture of gas and oil, the two formation volume factors can be added to obtain the two-phase formation volume factor B_t, taking the gas – oil ratio (R_s) into account.

$$B_t = B_o + B_g\,(R_{si} - R_s)$$

where R_{si} is the initial gas – oil ratio.

In Figure 13 the upper part of the curve is identical to the two-phase formation volume factor B_t. The formation volume factor for water B_w, which is shown in Figure 14, does not show a decrease through shrinkage resulting from degassing below the bubble point pressure (p_b). Instead, it rises further as a result of volume expansion on pressure release. This is due to the relatively low dissolved gas volume in water (ca. 1 m^3 gas/m^3) compared with oil (10 – 50 m^3 gas/m^3).

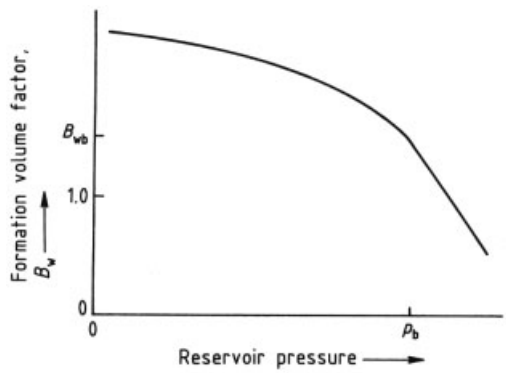

Figure 14. Formation volume factor for water [78]

2.1.2.2. Compressibility

The isothermal compressibility of the pore fluids is given by:

$$c = -\frac{1}{v}\frac{\partial V}{\partial p_T}$$

The pore fluids differ clearly in their compressibility behavior. While gases in the pressure range 0.1 – 100 MPa are extraordinarily compressible ($c = 10^{-5} - 10^{-8}\ \mathrm{Pa}^{-1}$), oils have a moderate to low compressibility ($c = 10^{-8} - 10^{-10}\ \mathrm{Pa}^{-1}$) depending on their dissolved gas content, and water has a compressibility of ca. $10^{-10}\ \mathrm{Pa}^{-1}$. The last figure is of the same order as the compressibility of the rock pore volume. In calculating the compressibility of real substances, the normalized thermodynamic functions of state, pressure, and temperature can be used.

For a *gas mixture* under real conditions, the compressibility equation reduced to normalized-conditions is:

$$c_g = \frac{c_{\mathrm{pr}}}{p_{\mathrm{pc}}} = \frac{1}{p_{\mathrm{pc}}}\left[\frac{1}{p_{\mathrm{pr}}} - \frac{1}{z}\frac{\partial z}{\partial p_{\mathrm{pr}}}\right]$$

Here, instead of the critical data for a single gas, pseudocritical values (p_{pc} and T_{pc}) are used as normalizing quantities.

$$p_{\mathrm{pr}} = \frac{p}{p_{\mathrm{pc}}} \quad p_{\mathrm{pc}} = \sum_1^n y_j p_{c_j}$$

$$T_{\mathrm{pr}} = \frac{T}{T_{\mathrm{pc}}} \quad T_{\mathrm{pc}} = \sum_1^n y_j T_{c_j}$$

where p_{pr} is the pseudoreduced pressure, T_{pr} the pseudoreduced temperature, and y_j the molar fractions of the individual components.

For an ideal gas, whose z factor is 1, the gas compressibility is therefore inversely proportional to its pressure. For liquids, whose critical z factor is 0.27 (most higher molecular mass hydrocarbons satisfy this condition), the compressibility can be determined from normalized pressure – temperature graph [78].

2.1.2.3. Viscosity

According to Newton, the dynamic viscosity μ is the proportionality factor between the shearing stress in the flow direction and the resulting shearing rate change perpendicular to the flow direction. Gas, oil, and water largely behave as Newtonian substances. For the viscosity of *gases* at low pressure and constant temperature, the product of the free path, mean velocity and the density of the gas molecules is a measure of internal friction. At low pressure, only the Brownian motion velocity of the molecules is relevant for this energy transfer and the viscosity of gases therefore increases with the square root of temperature. Only at elevated pressure does a reduction of the internal friction and, therefore, the viscosity, appear with gases and liquids with increasing temperature.

For *gas mixtures* the mixing rule definition for n components applies.

$$\mu_g = \frac{\sum_1^n y_j \mu_{g_j} M_j^{0.5}}{\sum_1^n y_j M_j^{0.5}}$$

To allow for the influence of pressure and temperature on the viscosity of a real gas, the normalized representations of the reduced pressure and temperature can again be employed.

The pressure and temperature dependence of the viscosity of reservoir oils is shown in Figures 15 and 16, respectively. Above the bubble point pressure p_b, the viscosity linearly increases with the reservoir pressure. Below the bubble point pressure, the oil viscosity falls steeply with increasing pressure or gas content.

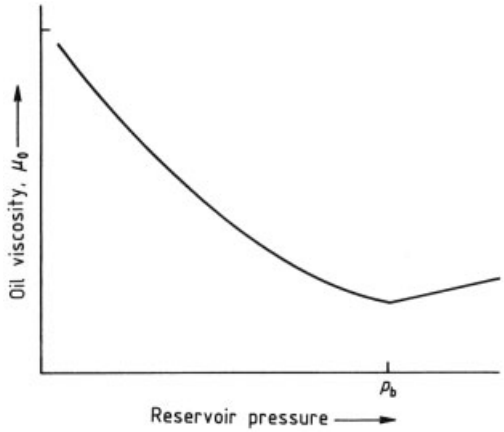

Figure 15. Viscosity of oil as a function of pressure [78]

The temperature dependence of the viscosity of oil is characterized by an exponential function, so that in a semilogarithmic plot versus

the reciprocal of the absolute temperature (see Fig. 16) a straight line is obtained. The slope of this straight line is greater the higher the oil's content of heavy components. For the dependence of the viscosity of water on the temperature the same mathematical relationships apply. In addition, the viscosity of water increases with increasing salt content [79].

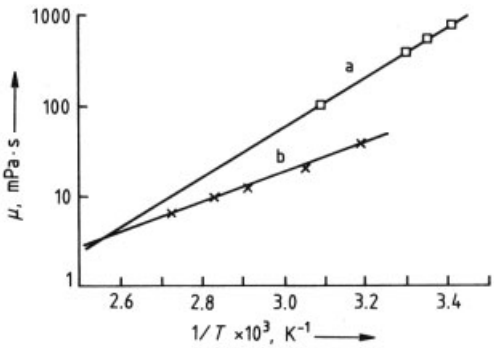

Figure 16. Viscosity of oil as a function of temperature
a) Heavy crude oil; b) Medium crude oil

2.1.3. Reserves

A detailed discussion of worldwide resources of oil and gas is given in Sections 1.1.4 and 1.2.3, respectively.

Resources mean the total volume of hydrocarbon stored in reservoirs. With oil, gas, or condensate reservoirs, the initial volume of the pore contents, the so- called original oil in place (OOIP) or original gas in place (OGIP), is the first quantity used for evaluating a resource. Since for physical, technological, and economic reasons the resources are not completely recoverable, a second concept is necessary for characterizing the profitability of a discovery: the reserves. *Reserves* mean those quantities of hydrocarbons that at a given time are proved on the basis of sufficient geological and reservoir engineering information and can, in future, be produced with currently available methods of recovery and applicable economic conditions. Reduced to a simple formula,

$$\text{Reserves} = \text{Resources} \times \text{Recovery factor}$$

The recovery factors for oil reservoirs, apart from a few exceptions, are between 20 and 45 % of the contents. For gas reservoirs, the recovery factor can be up to 85 %. The degree of certainty in the determination of the reserves increases with increasing development of a reservoir. According to a classification of reserves made by the Committee of the World Petroleum Congress, "proven reserves" are spoken of when the probability of proof exceeds 90 %, and "unproven reserves" when that probability is between 10 and 90 % [80]. The unproven reserves are split by a median value of 50 % into probable and possible reserves. The remainder are the speculative, so far undiscovered, possibly recoverable reserves.

Volumetric Determination of the Reservoir Contents. At the start of development of an oil or gas reservoir, usually only the contours of the reservoir are known; apart from test analyses data, no adequate production and pressure trends are available. The content of a reservoir can then be surveyed only by volumetric methods. The contents related to standard conditions are calculated from the following formulae:

$$\text{OOIP} = \frac{V_\text{b}\,\phi\,(1-S_\text{wi})}{B_\text{oi}}$$

$$\text{OGIP} = \frac{V_\text{b}\,\phi\,(1-S_\text{wi})}{B_\text{gi}}$$

where OOIP denotes original oil in place, OGIP denotes original gas in place, and S_wi is the initial water saturation.

The bulk volume of the reservoir V_b is determined from a geological model of the reservoir. In this procedure the nonproductive regions of the strata must also be taken into account such as clayey interlayers or lenses, water-bearing belts, and the distinction between gas-, oil- and water-bearing zones (gas – oil – water contacts) when delimiting the net rock volume from the gross rock volume. In the evaluation of borehole measurements, so- called cut-off values for porosity, saturation, clay content, and permeability are introduced to differentiate between pay zone and nonproductive thickness and to determine the pay zone thickness (Fig. 17). The black areas correspond to the porous, hydrocarbon-bearing formation intervals that are included in the net layer thickness.

For displaying the geological models, various mapping methods are chosen. For example, in Figure 18, the structural map of a reservoir is

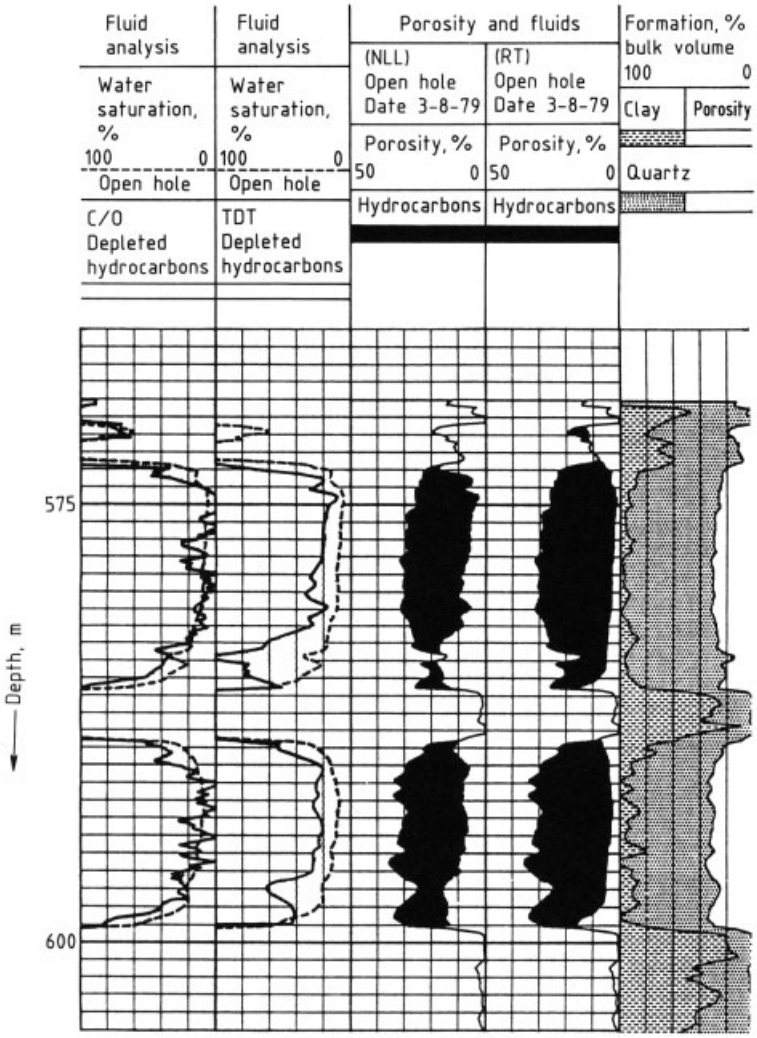

Figure 17. Net pays from well logging [81]

represented by subsurface contour lines (that is in feed). For determining the net rock volume of a reservoir, the reservoir body can be imagined to be divided into a large number of slices of equal thickness. Each of these slices can be represented by a geometrically similar body such as a truncated cone, truncated pyramid, or other, irregular shapes, allowing its volume to be calculated. A more elegant method is the graphical evaluation of the subsurface contour planes. In this method the areas of the contour line map determined by planimeter for the upper and lower boundary of the reservoir are plotted against the relative depth in a system of cartesian coordinates [77].

Furthermore, if the lower and upper boundaries of the reservoir contents – the gas – oil and water – oil contacts, respectively – are taken into account, the rock volume can be determined directly from the area – depth graph by means of millimeter paper or planimetry. These computed or graphically-determined rock volumes are inserted into the formulae for calculating the OOIP and OGIP. In low-permeability reservoirs bearing formation water, the transition zone from the completely water saturated zone to the oil

bearing zone, containing irreducible water, is an uncertainty factor in the oil or gas resource calculation. The transition region is characterized by the capillary pressure curve.

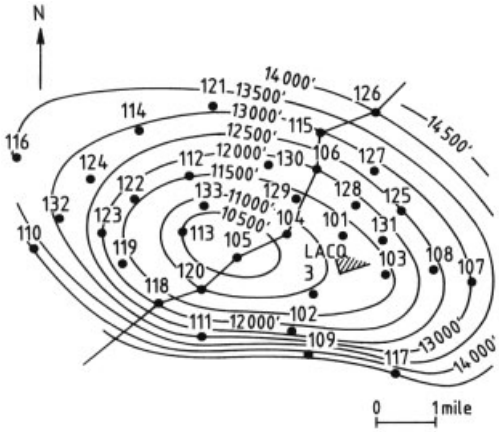

Figure 18. Isomap representation of reservoirs/regular contours [82]

2.2. Flow Processes in Porous Media

The original state of a crude oil or natural gas reservoir is characterized by the thermodynamic equilibrium of the hydrocarbon phases and the capillary equilibrium of the wetting and nonwetting phases, which in turn depends on the prevailing temperature, pore space pressure, and rock properties. Not only the composition but also the properties of the oil, gas, and water phase show a remarkable "memory" for the early history of the formation of the deposit. For example, oils that lack in certain alkanes can be recognized as intermediate products of biological degradation from an original sapropellitic material. Light oils which have been exposed to oxidation as a result of tectonic postsedimentary processes, might have been converted into medium-heavy oils, but they still retain the chemical features (fingerprints) of their origins.

The initial distribution of the phases in the pore space is governed by the equilibrium between interfacial forces and viscous forces during the migration process. If a volatile oil enters a porous rock which is saturated with salt water and flushes the pore space over extended periods then the resulting initial water saturation in the deposit is normally greater than the adsorbed water saturation (i.e., the layer of water retained by physical adhesion to the interior surfaces of the rock). If, in contrast, the migrating oil is composed of heavier hydrocarbon fractions, it displaces the pore water almost down to the level of the absorbed water saturation. After a reservoir has been filled, the materials are in dynamic rather than static equilibrium. Diffusion and convection can influence the material equilibrium and the distribution of the phases in the deposit.

2.2.1. Single-Phase Flow

If a fluid with little or no compressibility (water, oil) is flowing in an infinitely large reservoir system towards a pressure sink (well) then – after a transient phase – a steady-state condition is reached between the quantity removed from the well and the quantity flowing in from the deposit. Under these conditions, the pressure changes only as a function of space, but not of time. However, during the transient state before the steady-state equilibrium is reached the pressure also changes with time.

2.2.1.1. Steady-State Flow – Darcy's Equation

In the laminar flow regime – i.e., in a homogeneously porous medium with a Reynolds number < 1 – Darcy's law holds; a single well in an infinitely large reservoir radial flow geometry can be assumed

$$\frac{q}{2\pi r h} = \boldsymbol{v} = -\frac{k}{\mu}\frac{\partial p}{\partial r} \tag{2.1}$$

where q is the flow rate, r the radius, h the thickness of the porous layer, $\boldsymbol{v}$ the apparent velocity, k the permeability of the rock, μ the dynamic viscosity, and p the pressure. Integration of Equation (2.1) between the well radius r_{w} and the radius r_{e} of the area drained by the well gives the following for steady-state flow:

Boundary conditions: p_{e} at r_{e}; p_{wf} at r_{w} are constant with time

$$q_{\mathrm{w}} = \frac{2\pi h k}{B\mu}\frac{(p_{\mathrm{e}} - p_{\mathrm{wf}})}{\ln r_{\mathrm{e}}/r_{\mathrm{w}}} \tag{2.2}$$

where B is the formation volume factor of the fluid.

However, for a *compressible medium* (gas) the volume flow is not constant along the direction of flow. If Darcy's formula is still to be applied it is necessary to convert the flow rate to some mean pressure p_m between the system's input and output pressures [77]. The average flow rate can be converted to the required flow conditions in the well by using the equation of state for ideal gases.

Boundary conditions for linear geometry: p_e at $x = l$; p_{wf} at $x = 0$, where x denotes the distance from the well.

$$p_m = \frac{p_e + p_{wf}}{2}$$

$$q_m = \frac{Ak}{\mu}\frac{(p_e - p_{wf})}{l}$$

$$q_w\, p_{wf} = q_m\, p_m = -\frac{Ak}{\mu}\frac{(p_e^2 - p_{wf}^2)}{2l}$$

$$q_w = -\frac{Ak}{B\mu}\frac{(p_e^2 - p_{wf}^2)}{2p_{wf}l}$$

2.2.1.2. Transient Single-Phase Flow

Consider a space element in the form of a cylindrical capillary of length Δx and cross-sectional area A (Fig. 19). A mass flow $(\varrho \boldsymbol{v}_x)_x A$ is injected at the inlet end and the mass flow $(\varrho \boldsymbol{v}_x)_{x+\Delta x} A$ emerges at the outlet point at a distance x.

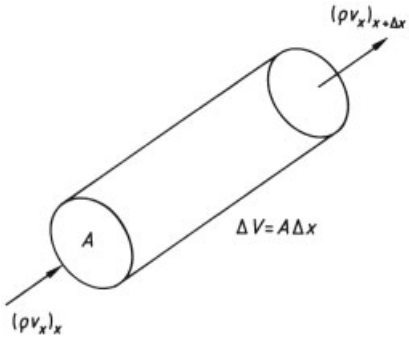

Figure 19. Material balance in a cylindrical capillary of length Δx and cross-sectional area A

Under transient flow conditions, the difference between these two mass flows corresponds to the mass stored in the volume element ΔV. The material balance must be maintained because of the principle of conservation of mass.

$$A\,\Delta t \quad ((\varrho \boldsymbol{v}_x)_x - (\varrho \boldsymbol{v}_x)_{x+\Delta x}) = \text{Mass stored}$$

(Mass in − Mass out)

divided by $A\,\Delta x\,\Delta t$

$$\lim_{\substack{\Delta x \to 0 \\ \Delta t \to 0}} \frac{(\varrho \boldsymbol{v}_x)_{x+\Delta x} - (\varrho \boldsymbol{v}_x)}{\Delta x} = -\frac{\Delta \varrho_{\Delta t}}{\Delta t}$$

$$\frac{\partial\,(\varrho \boldsymbol{v}_x)}{\partial x} + \frac{\partial \varrho}{\partial t} = 0$$

The differential equation for the continuity law is obtained from the difference equation assuming small increments of time and space. The nabla operator and the porosity in the storage volume ΔV are introduced so that this equation can be applied to a porous medium and extended to three-dimensional flow conditions. For horizontal flow of a medium of low compressibility, the extended continuity equation becomes:

$$\nabla\,(\varrho \boldsymbol{v}) + \frac{\partial\,(\phi\,\varrho)}{\partial t} = 0 \tag{2.4}$$

In one dimension:

$$\boldsymbol{v}_x = -\frac{k}{\mu}\frac{\partial p}{\partial x}$$

$$-\frac{\partial}{\partial x}\left(\varrho \cdot \frac{k}{\mu}\frac{\partial p}{\partial x}\right) + \frac{\partial\,(\phi\,\varrho)}{\partial t} = 0 \tag{2.5}$$

$$\frac{k}{\mu}\left[\frac{\partial \varrho}{\partial x}\frac{\partial p}{\partial x} + \varrho\frac{\partial^2 \mathrm{p}}{\partial x^2}\right] = \Phi\,\frac{\partial \varrho}{\partial t}$$

where k/μ is the transmissibility of the pore fluid for a given porous medium.

As an initial condition it was assumed that the transmissibility k/μ and the porosity ϕ are constant throughout the pore space system. The compressibility equation

$$c = -\frac{1}{\varrho}\frac{\partial \varrho}{\partial p}$$

for a fluid of low compressibility is introduced and used to replace the change in density ϱ by the change in pressure p.

$$\frac{k\,\varrho}{\mu}\left[c\left(\frac{\partial p}{\partial x}\right)^2 + \frac{\partial^2 p}{\partial x^2}\right] = \varrho c \Phi\,\frac{\partial p}{\partial t}$$

Both sides are multiplied by the product $\mu/k\varrho$. The flow rate, at some distance from the well, is only in the range of a few centimeters per hour

so the pressure gradient p/x and the square of the pressure gradient are very small in comparison with the other variables in Equation (2.5). This term can therefore be neglected.

This then gives Equation (2.6) which is similar in form to the diffusivity equation, and is referred to as the pressure propagation or hydraulic conductivity equation for the one-dimensional case.

$$\frac{\partial^2 p}{\partial x^2} = \frac{\phi \mu c}{k} \frac{\partial p}{\partial t} \tag{2.6}$$

Equation (2.5) can also be solved for other initial conditions. If the density of the fluid phase is defined by a mass law, e.g., the gas law for ideal gases, then the following form of the hydraulic conductivity equation is obtained

$$\frac{\partial^2 \left(p^2\right)}{\partial x^2} = 2 \frac{\phi \mu}{k} \frac{\partial p}{\partial t} \tag{2.7}$$

These equations can also be represented in three-dimensional form. To solve Equations (2.6) and (2.7) it is necessary to adapt the flow geometry and boundary conditions to match the situation in the deposit. For a cylindrical coordinate system Equation (2.8) is obtained.

$$\frac{\partial^2 p}{\partial r^2} + \frac{1}{r} \frac{\partial p}{\partial r} = \frac{\phi \mu c}{k} \frac{\partial p}{\partial t} \tag{2.8}$$

With a linear pressure sink as the well, and for infinite extension of the reservoir, i.e., $r_e = \infty$, the Boltzmann transformation can be used for solving Equation (2.8).

The following boundary conditions apply:

$t=0$	$r=r_w$	$p=p_i$
	$r=r_e$	$p=p_i$
$t=t_1$	$r=r_w$	$p=p_{wf}$
	$r=r_e$	$p=p_i$

The result of the Boltzmann transformation is the exponential integral equation:

$$p_{wf} = p_i + \frac{q\mu}{4\pi h k} E_i \left(\frac{-\phi \mu c_t r^2}{4kt} \right) \tag{2.9}$$

where the symbol E_i denotes exponential integral and p_{wf} the flowing pressure at the bottom of the well.

In an infinitely large homogeneous reservoir system with a line source as the well, the flowing pressure at any point can be calculated from Equation (2.9) as a function of the radius r and the time t. The natural logarithm can be used in place of the exponential integral provided the argument of the exponential integral is significantly smaller than the limiting value of 0.01 [83].

$$p_{wf} = p_i + \frac{q\mu}{4\pi h k} \ln \frac{\gamma \phi \mu c_t r_w^2}{4kt} \tag{2.10}$$

where γ denotes the Euler constant.

To reconstruct the situation in a deposit, the deposit's hydraulic system must be transformed back to its original state. This can be done by, for example, closing off an operating well. The pressure in the well will then gradually build up again until it reaches the original reservoir pressure, assuming that the deposit is comparable to an infinitely large reservoir system (influence of aquifers). The shut-in of the well is simulated mathematically by superimposing a negative flow rate q on the radial flow system. Combination of the two equations for the influx into the well and the out flow from the well gives an equation for the shut-in pressure p_{ws} of the following form:

$$p_{ws} = p_i - \frac{q\mu}{4\pi h k} \ln \frac{t_p + \Delta t}{\Delta t} \tag{2.11}$$

where $t_f = t_p$ denotes the flow duration and Δt the shut-in time [84]. This equation has been used to extrapolate the initial reservoir pressure p_i from the graphical presentation of the shut-in pressure plotted against the logarithm of the time ratio ($\frac{t_p + \Delta t}{\Delta t} = 1$ (Fig. 20).

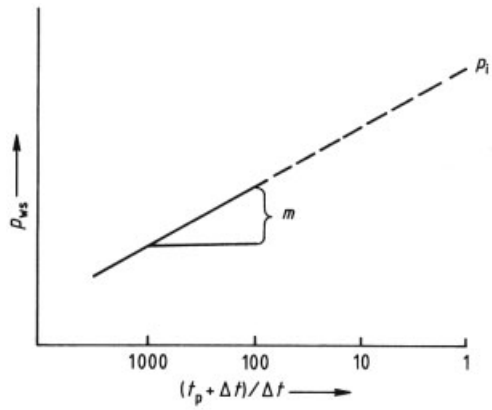

Figure 20. Shut-in pressure p_{WS} versus logarithmic time ratio $\frac{t_p + \Delta t}{\Delta t}$ for extrapolation of initial reservoir pressure p_i [83]

For an infinitely large reservoir the initial pressure in the well will be reconstituted after an infinite shut-in time $\frac{t_p + \Delta t}{\Delta t} = 1$. The gradient of the straight line m corresponds to the coefficient in Equation (2.11) in front of the logarithmic function. As the flow rate is normally determined in a measuring tank above ground, it has to be transferred to the state in the deposit. This is done by incorporating the formation volume factor B in the defining equation for m.

$$m = \frac{q \mu B}{4 \pi h k}$$

From the gradient m the permeability capacity $k \cdot h$ can be calculated for the reservoir system.

$$kh = \frac{q \mu B}{4 \pi m}$$

In addition to this simple radial model with the selected initial and boundary conditions many methods exist for finding solutions for specific reservoir conditions, which can be applied to individual well and reservoir problems. Individual solutions have therefore been found for the differential equations to the following conditions; finite reservoir extension, near wellbore effects, fractured reservoirs, stratified reservoirs, and partial penetration and particular direction of the well in the producing formation.

2.2.2. Multiphase Flow

2.2.2.1. Capillary Pressure and Saturation

If there is more than one phase in the pore space system of a reservoir rock then the distribution of the phases is determined by the interfacial tensions between the fluids and the solid phase. The equilibrium of forces between two liquid phases (oil and water) and a solid phase is shown in Figure 21.

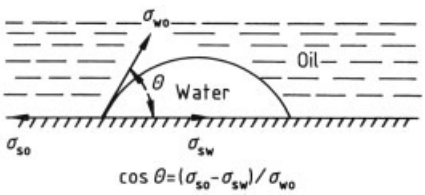

Figure 21. Equilibrium of forces between two liquid phases (water and oil) and the solid phase [77]
σ_{so} = interfacial tension solid – oil; σ_{sw} = interfacial tension solid – water; σ_{wo} interfacial tension water – oil

The wetting angle θ is a measure of the wettability of the solid phase by the oil and water phases. A value of 0° for θ represents an ideally water wet solid, whereas 180° defines a completely oil wet surface. The distribution of the phases for a pore cross section therefore can be idealized as follows:

- Water, as the most strongly wetting phase, is positioned directly against the surface of the rock
- Gas, which is the least wetting, is positioned in the center of the pore duct
- The oil phase is distributed between them (Fig. 22)

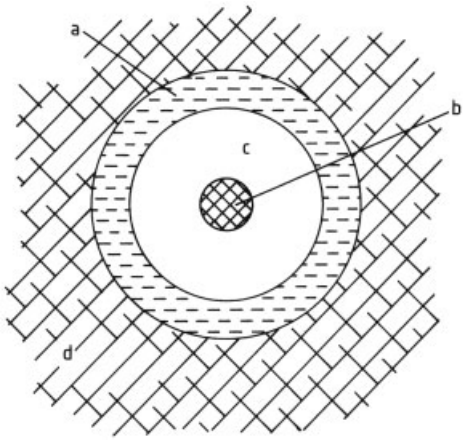

Figure 22. Idealized distribution of oil, water, and gas phase in a pore cross section
a) Water; b) Gas; c) Oil; d) Rock

Due to the adhesive forces the pressure in the wetting phase (water) is lower than in the nonwetting phase (oil). The difference between these two pressures is the capillary pressure p_c where

$$p_c = p_o - p_w$$

Capillary pressure curves (see Fig. 23) are of fundamental importance for the mobility of the phases in multiphase flow. They determine the irreducible, and, therefore, fixed saturation of the wetting phase at infinite capillary pressure and characterize the capillary ascent of the water table at the boundary of the oil – water zones. The capillary pressure can be directly transferred into

elevation by use of the capillary/gravity force equilibrium.

$$\frac{2\sigma_{\mathrm{ow}}\cos\theta}{r} = (\varrho_{\mathrm{w}} - \delta_{\mathrm{o}})_{gh}$$

where σ_{ow} denotes the interfacial tension between oil and water phase, r the equivalent hydraulic radius of the porous system, ϱ_{w} the density of water, δ_{o} the density of oil, g the gravitational constant, and h the elevation above the water table.

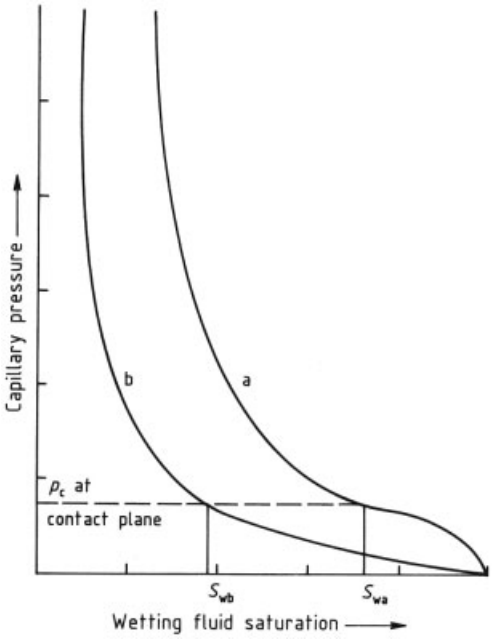

Figure 23. Capillary pressure versus wetting fluid saturation [85]
a) Low permeability; b) High permeability

For the same capillary pressure, a higher water saturation is found in a pore space system of lower permeability (a) than in a highly permeable one (b). If it is assumed that water represents the wetting phase this means a reinforced capillary imbibition of water from zones of high permeability into low permeability and therefore an improved displacement of oil from the pores of lower permeability.

2.2.2.2. Relative Permeabilities

The mobilities of the individual phases in multiphase flow in a porous medium depend on the wettability of the rock, on the interfacial tension between the fluids, and on their viscosities. A new flow coefficient, which is introduced into the phase flow equation as the *relative permeability,* is needed to describe the different mobilities of the phases in the medium as a function of their fractional saturation. Relative permeability k_{r} is defined as the ratio of the effective permeability of a phase (k_{w}, k_{o}, k_{g}) in the presence of one or more adjacent phases to a reference permeability. The absolute rock permeability according to Darcy, or the permeability to water at 100 % water saturation, or the permeability to oil for irreducible water saturation can be used as the reference permeability. The relative permeabilities for a binary oil – water system for a more strongly water-wet and a more strongly oil-wet rock are shown in Figure 24. If the saturation of the wetting phase S_{w} is plotted on the abscissa then the curves are displaced to the right with increasing water wettability. On the other hand, the final value for the relative permeability of the water phase increases if the water wettability decreases.

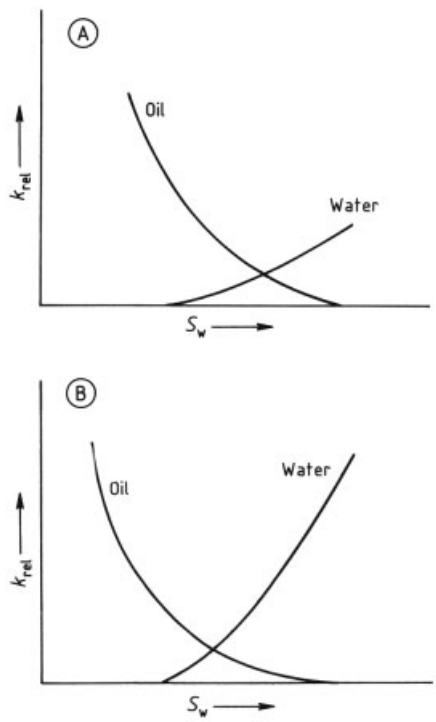

Figure 24. Relative permeabilities for a binary oil – water system [86]
A) Water-wet rock; B) Oil-wet rock

This type of representation can only be used for two-phase permeabilities; for a three-phase system STONE [87] has proposed a paired representation of the water – oil and gas – oil relative permeabilities.

Apart from this, a ternary diagram is the only possible way of representing the relative permeabilities of a three-phase system.

The flow equations for a two-phase system can therefore be represented while taking account of the functions of the relative permeabilities and the capillary pressure.

Oil:

$$\nabla\left[-\frac{k_{\mathrm{ro}}k\,\varrho_{\mathrm{o}}}{\mu_{\mathrm{o}}}\nabla\boldsymbol{\Phi}_{\mathrm{o}}\right]=\boldsymbol{\Phi}\frac{\partial S_{\mathrm{w}}}{\partial p_{\mathrm{c}}}\left(\frac{\partial\boldsymbol{\Phi}_{\mathrm{o}}}{\partial t}-\frac{\partial\boldsymbol{\Phi}_{\mathrm{w}}}{\partial t}\right)$$

Water:

$$\nabla\left[-\frac{k_{\mathrm{rw}}k\,\varrho_{\mathrm{w}}}{\mu_{\mathrm{w}}}\nabla\boldsymbol{\Phi}_{\mathrm{w}}\right]=\phi\frac{\partial S_{\mathrm{w}}}{\partial p_{\mathrm{c}}}\left(\frac{\partial\Phi_{\mathrm{o}}}{\partial t}-\frac{\partial\Phi_{\mathrm{w}}}{\partial t}\right)$$

where $\boldsymbol{\Phi}$ is the flow potential:

$$\boldsymbol{\Phi}\frac{1}{\varrho}\,(p+\varrho g z)$$

and k_{ro} and k_{rw} denote the relative permeabilities of oil and water, respectively.

Depending on the boundary conditions, this system of partial differential equations, can, in general, only be solved numerically.

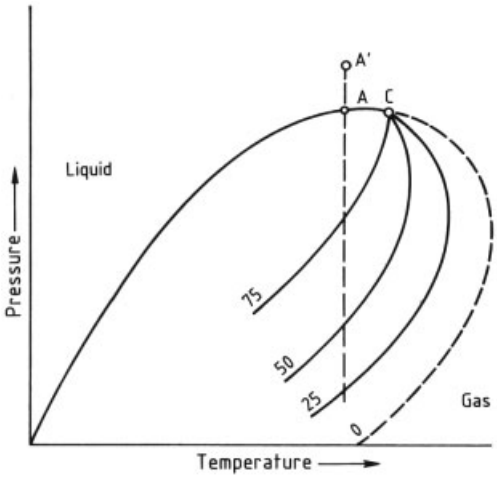

Figure 25. Binary diagram for a gas – oil mixture (each phase consisting of one component) [77]
A = bubble point pressure (pressure depletion at room temperature); A′ = initial reservoir pressure; C = critical point

2.2.3. Drive Mechanisms and Recovery Efficiency

The energy reservoir for the flow of oil, gas, and water in a porous medium consists of the compression energy of the aquifer volume surrounding the deposit; the mechanical energy of the gas liberation process; the influence of gravity; and the capillary pressure effect.

As shown in Section 2.2, the initial pressure in a deposit is mainly hydrostatic. The ratio of the initial pressure to the saturation pressure of the oil – gas mixture characterizes the individual steps of the oil recovery process and the drive mechanism. Figure 25 shows a simplified binary diagram for a mixture of gas and oil, each phase consisting of one component.

In this $p-T$ diagram the oil recovery process is considered isothermal. If the initial reservoir pressure p_{i} lies above the bubble point pressure p_{b} the deposit is termed a gas undersaturated oil deposit. When the pressure is released the volume expansion drive alone is initially effective. When the pressure falls below the bubble point pressure, this is supplemented by the mechanical effect of the gas liberation process. If the bubble point pressure and initial reservoir pressure are equal then the deposit is termed a gas-saturated oil deposit. If the initial reservoir pressure lies below the bubble point pressure then the expansion effect of a gas cap can provide an additional drive mechanism. From Figure 25 it can also be seen that no oil deposits exist to the right of the critical point C and that the gas reservoir or con-densate phase type then begins. The influence of the different drive mechanisms is best calculated from a material balance (see Section 2.3.1). The physical and mathematical treatment of these material balances assumes closed systems in which removal of fluid results in a reduction of pressure. If, however, the removal of fluid is replenished by an aquifer feed, or by water or gas injection then the system pressure remains essentially unchanged. A pressure change in the system also changes the properties of the phases in a way which is disadvantageous to the oil recovery process (see Section 2.1.2); measures which maintain the pressure are, therefore, favorable. *Water drive* is by far the most effective drive mechanism; it can be introduced by artificial flooding measures in locations where

water cannot be fed from natural sources (i.e., inadequate aquifer size or inadequate energy replenishment due to low aquifer permeability). The oil recovery process by natural aquifer expansion and by the other drive mechanisms mentioned rate as *primary extraction processes.* On the other hand, replenishment of the reservoir energy by water or gas injection is one of the *secondary extraction processes*.

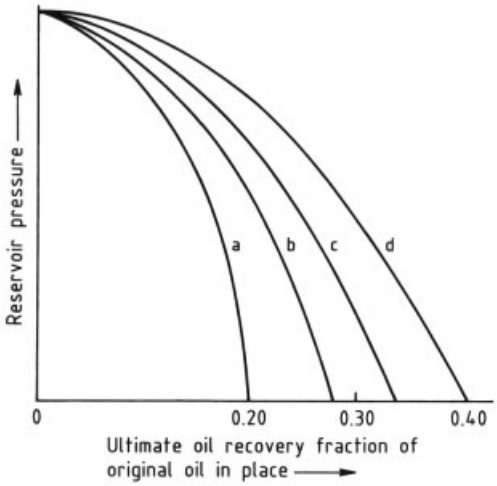

Figure 26. Oil recovery efficiency as a function of pressure for various drive mechanisms [88]
a) Depletion drive; b) External gas drive (small gas cap); c) External gas drive (large gas cap); d) Water drive

The *recovery efficiency* of a deposit is given by the percentage of the cumulative production of the quantity of hydrocarbon originally present. The performance of extraction and the recovery efficiency are determined largely by the geological factors of the deposit and the physics of the oil recovery process. The physical approach to oil recovery processes presupposes a microscopic dimension, and the geological approach a macroscopic one; the recovery factor must, therefore, include microscopic and macroscopic parameters. The total recovery factor or recovery efficiency E_{R} is the product of the microscopic recovery factor E_{D}, the areal recovery factor E_{A}, and the vertical recovery factor (flood invasion efficiency) E_{I}. Figure 26 shows that the recovery efficiencies lie between 20 % and 40 % of the oil present, depending on the characteristic drive mechanisms and oil viscosities. In gas reservoirs the recovery efficiency lies between 80 and 90 % of the gas present. This is attributable to the high mobility of the gas and the favorable form of the volume expansion drive.

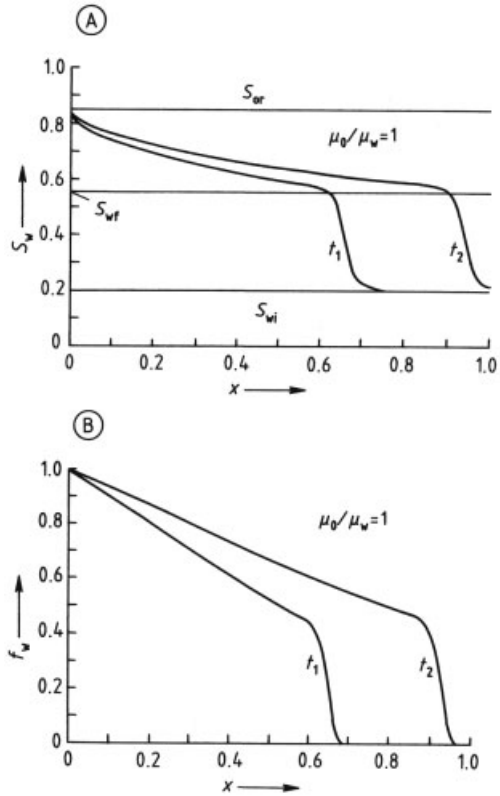

Figure 27. A) Saturation distribution along flooding section (direction x); B) Fractional flow for water along flooding section

2.2.3.1. Microscopic Recovery Factor

BUCKLEY, LEVERETT, and WELGE [89–91] have developed the theory of the displacement of a fluid by a second, immiscible fluid in a homogeneous medium. This microscopic process determines which fluid exhibits the stronger wetting properties towards the porous medium and which has the greater mobility. The displacement process in the pore space is not as perfect as the displacement of a fluid by a piston in a machine; this is because a mobile residual phase saturation of the displaced fluid (oil) remains after the displacement front has passed. Continued flooding can reduce this mobile residual phase saturation down to an irreducible, fixed percentage. Through a linear flooding process it is pos-

sible to determine the saturation distribution S along the flooding section, i.e., in direction x (Fig. 27).

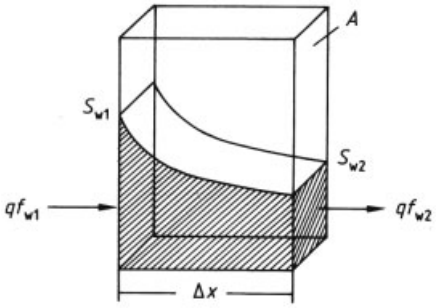

Figure 28. Mass balance for an infinitely small volume unit at the saturation boundary between oil and water

At the inlet cross section the water saturation is only slightly smaller than the difference of 1 minus the physical residual oil saturation S_{or}. Up to the flooding front, the water saturation decreases to a value S_{wf} (front saturation), depending on the mobilities of the phases. The difference between the ultimate water saturation $(1-S_{\mathrm{or}})$ and the actual water saturation at any section is defined by the profile f, the saturation of mobile oil. The front is characterized by a step change in the water saturation down to the *irreducible water saturation* S_{wi}. The proportions of the flowing phases change along this saturation profile and can be expressed by the fractional flow f of one of the two flowing phases, as it is shown for different times t_1 and t_2 in Figure 27 B.

$$f_{\mathrm{w}}=\frac{q_{\mathrm{w}}}{q_{\mathrm{o}}+q_{\mathrm{w}}}=\frac{1}{1+M^{-1}}$$

$$M=\frac{\lambda_{\mathrm{w}}}{\lambda_{\mathrm{o}}}=\frac{k_{\mathrm{rw}}\,\mu_{\mathrm{o}}}{k_{\mathrm{ro}}\,\mu_{\mathrm{w}}}$$

The fractional flow for water is a function of the mobility ratio M, i.e., the mobility of water $k_{\mathrm{w}}/\mu_{\mathrm{w}}$ as the displacing phase to the mobility of the oil $k_{\mathrm{o}}/\mu_{\mathrm{o}}$ as the displaced medium. At the front the fractional flow of water falls to zero, as only the mobile oil phase can flow ahead of the front. By using a continuity approach in the porous medium the velocity has been determined at which the front or the saturation profile moves through a pore continuum [90]. If an infinitely small volume unit is considered at the boundary between oil and water then the difference between input and output water flow volume in the time unit Δt must be equal to the change in saturation in the space element $A\,\Delta x$ (see Fig. 28).

$$\Delta t\,(q\,f_{\mathrm{w}_1}-q\,f_{\mathrm{w}_2})=A\,\Delta x\phi\,(\mathrm{S}_{\mathrm{w}_1}-S_{\mathrm{w}_2})$$

This finite difference equation can then be transformed into a differential equation which gives the velocity of propagation of a point on the saturation profile.

$$\lim_{\substack{\Delta x\to 0\\ \Delta t\to 0}} q\Delta f_{\mathrm{w}}\,u\Delta t=A\Delta x\,\phi\Delta S_{\mathrm{w}}$$

$$\frac{\partial x}{\partial t}=\frac{q}{A\,\phi}\,\frac{\partial f_{\mathrm{w}}}{\partial S_{\mathrm{w}}}$$

Integration between the limits x and 0 and the times t and 0 gives the equation for determining position x which has reached some arbitrary saturation value at the time t after injection of a certain volume of water W_{i}.

$$x=\frac{W_{\mathrm{i}}}{A\,\phi}\,\frac{\partial f_{\mathrm{w}}}{\partial S_{\mathrm{w}}}$$

$$W_{\mathrm{i}}=\int_0^t q\mathrm{d}t$$

If, instead of some arbitrary saturation, the intention is to characterize the saturation at the front, and hence the front velocity v_{f}, then Figure 29 illustrates the graphic method for determining the front saturation.

The saturation at the displacement front S_{wf} is defined by the point where the tangent to the fractional flow curve from its origin intersects the curve itself. An extension of this tangent until it intersects the fractional flow value 1, characterizes the average water saturation $\bar{S}_{\mathrm{wbt}}$ as the front breaks through at the end of the flooding section. The difference $1-S_{\mathrm{or}}-\bar{S}_{\mathrm{wbt}}$ provides a measure for the residual oil which can be mobilized. In practice this mobilizable residual oil can only be extracted after an infinitely long flooding time. The greater the mobility of the displaced phase in comparison with the displacing phase the more rapidly the residual oil saturation reaches a quasi steady-state final condition S_{or}. This physical residual oil saturation

serves as a measure for establishing the microscopic oil recovery factor E_{D}.

$$E_{\mathrm{D}} = \frac{S_{\mathrm{oi}} - S_{\mathrm{or}}}{S_{\mathrm{oi}}} = 1 - \frac{S_{\mathrm{or}}}{S_{\mathrm{oi}}}$$

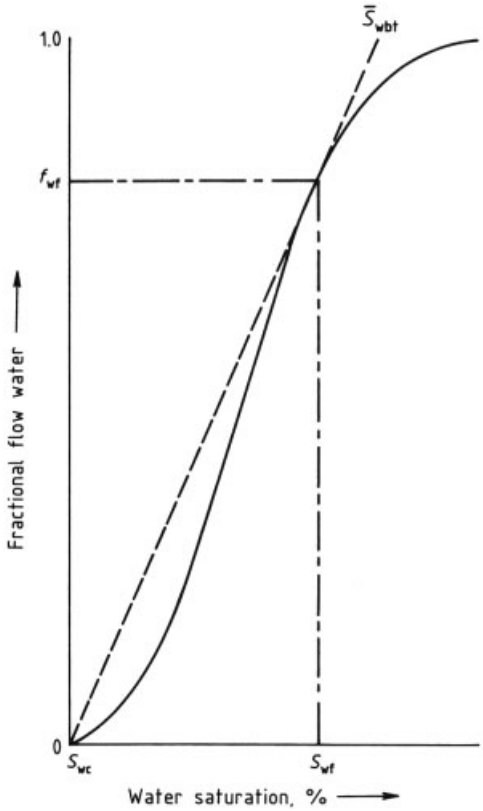

Figure 29. Fractional flow curve for water, definition of the front saturation S_{wf} and break through saturation $\bar{S}_{\mathrm{wbt}}$ according to Welge [92]

2.2.3.2. Areal Recovery Factor

Consider a homogeneous deposit with no vertical stratification and a concentric arrangement of the peripheral production wells at a distance r from an injection well. Here the areal oil recovery factor E_{A}, i.e., the ratio between swept area A_{S} and the total area A, would be equal to 1 by the time the flooding front has reached the peripheral production wells.

For a five-point well pattern as shown in Figure 30 a considerable proportion of the area A remains unswept when the flooding front reaches the four peripheral production wells. The areal oil recovery factor which can be achieved before breakthrough of the flooding front depends on the mobility ratio of the front and the arrangement of the wells.

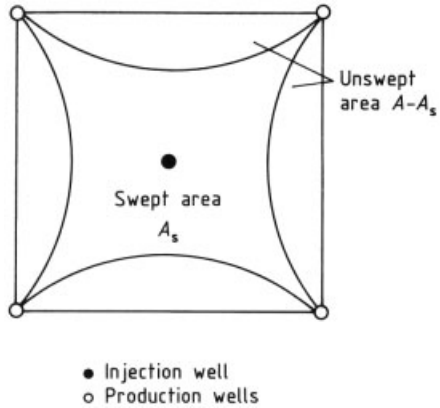

Figure 30. Concentric arrangement of four peripheral production with respect to one injection well, 5-point pattern

In a homogeneous producing formation the well spacing initially plays no part. However, as the capacity of the producing formation to accept the injection medium (water) depends on permeability, the well spacing cannot be set at any arbitrary distance. There is little point in optimizing the well spacing on the basis of a homogeneous deposit model, as in practice more complex deposit formations are often found.

2.2.3.3. Vertical Recovery Factor or Flood Invasion Efficiency

If the oil producing formation were to have no vertical subdivisions then the overall oil recovery factor from this deposit would be the product of the microscopic and areal oil recovery factors. However, the conditions under which the sediments were deposited have normally led to a vertical, facially subdivided stratification, with the result that the vertical recovery factor also has to be taken into account when determining the overall oil recovery.

If water is invading into a stratified producing formation from an aquifer or an injection well the strata fill in the order of their permeabilities. The most permeable layer is flooded first.

Dykstra and Parsons [93] have developed a simple theory for determining the vertical recovery factor in a stratified producing formation. For

this purpose they have further simplified the microscopic displacement process as described by BUCKLEY and LEVERETT and have assumed that no mobile residual oil remains behind the front. However, unlike BUCKLEY and LEVERETT, they have taken into account the fact that the velocity of the displacement front depends on the mobility ratio of the phases and on the position of the displacement front along the flooding section. The higher the mobility of the displacing phase the more rapidly the flooding front extends toward the end of the flooding section. The flooding velocity at the inlet corresponds to the effective Darcy flow velocity of the oil phase (see Fig. 31).

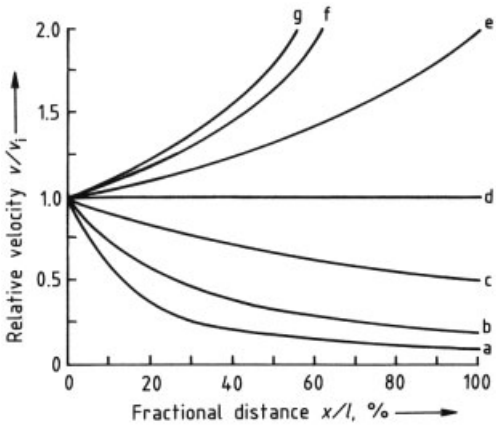

Figure 31. Relative velocity of flooding front as function of fractional distance for different mobilities of displacing phase [94]
a) $M = 0.1$; b) $M = 0.2$; c) $M = 0.5$; d) $M = 1$; e) $M = 2$; f) $M = 5$; g) $M = 10$

The flood invasion efficiency E_{I} is given by integration of the water saturation profile in Figure 32 using the following formula

$$E_{\mathrm{I}} = \frac{\sum_{1}^{i} h_j + \sum_{i+1}^{n} \frac{x_j}{l} h_j}{\sum_{1}^{n} h_j}$$

The position of the flooding front at time t in the incompletely swept strata $j = i + 1$, relative to the layer i in which the water has just broken through, is determined by the equation

$$\frac{x_j}{l} = \frac{\left[M^2 + \frac{k_j}{k_{\mathrm{i}}}\left(1 - M^2\right)\right]^{0.5}}{1 - M} - \frac{M}{1 - M}$$

where x_j/l is the position of the flooding front.

The following formula applies to the time until breakthrough t_i

$$t_i = -\frac{\phi l^2 (M+1)\, \mu_w}{2 k_{wi} \Delta p}$$

With the aid of these three equations the vertical oil recovery factor can be determined at any given time. The product of the three factors gives the overall recovery efficiency E_{R}:

$$E_{\mathrm{D}} = E_{\mathrm{D}}\, E_{\mathrm{A}}\, E_{\mathrm{I}}$$

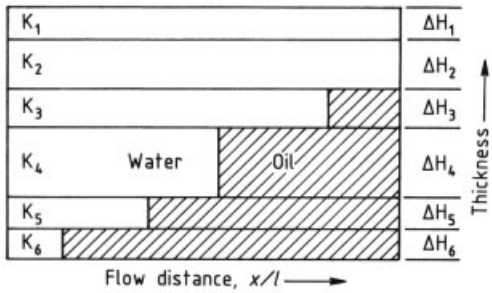

Figure 32. Cross section of a layered reservoir; position of flood front in the individual layers at a certain time level t [94]

2.3. Modeling of Reservoir and Production Performance

An effective reservoir characterization comprises a description of the reservoir's geological boundaries, the petrophysical properties of the rock, and the material properties of the contents. Its aim is to calculate the production behavior and plan measures to optimize production from the reservoir. The calculation methods used are the basis for predictions of future production performance and are a strategic instrument of reservoir management. All methods of prediction are based on the physical fundamentals of the recovery process and the flow properties of the fluids in a geologically complex reservoir body. The credibility and reliability of these prediction methods depend on how closely the calculation techniques model the reality of the recovery process and the reservoir. This accuracy of the model must be checked by a *model validation* before use. Once the model is chosen it must be adjusted to the production history of the reservoir (*model calibration*). The better the model is calibrated, the more reliable the prediction.

The following models for describing the reservoir behavior, with increasing complexity of process description and/or reservoir characterization, can be distinguished:

1) Tank models (material balances)
2) Combined tank – flow model (Tarner prediction method)
3) Multiphase flow models (reservoir simulation)

2.3.1. Material Balances Method

The tank model is the simplest form of reconstruction of the reservoir contents. It is based on the conservation of mass and an equation of state [88]. All forms of material balance use the following approach:

$$\text{Original contents} - \text{Quantity produced}$$

$$= \text{Quantity remaining}$$

In order to set up a material balance the following assumptions must be made:

1) The reservoir is considered as an energetic unit
2) The material properties and the pressure can be described by mean values representative of the reservoir
3) The material laws for the pore contents can be determined with sufficient accuracy

Material Balance for a Gas-Undersaturated Oil Reservoir Without Aquifer Influence This material balance is presented as an example.

Initial State:

1) The initial pressure p_i exceeds the bubble point pressure p_b
2) The reservoir contents consist of the volume of gas-undersaturated oil NB_{oi} and the volume of initial water V_{wi}

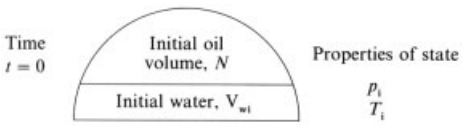

$$V_{pi} = \frac{N B_{oi}}{1 - S_{wi}}$$

$$V_{wi} = V_{pi} S_{wi}$$

(for explanation of terms see the abbreviations list at the beginning of this article).

State Above the Bubble Point Pressure p_b:

1) The actual pressure exceeds the bubble point pressure p_b but is lower than p_i.
2) The oil volume $N_p B_o$ is produced. The volume lost is replaced by the expansion of the rock matrix, the residual oil, and the initial water.

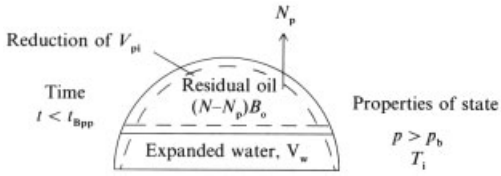

$$V_p = V_{pi} \left[1 - c_f \left(p_i - p\right)\right]$$

$$V_w = V_{wi} \left[1 + c_w \left(p_i - p\right)\right]$$

$$\left(N - N_p\right) B_o = \text{Residual oil}$$

(for explanation of terms see the abbreviations list at the beginning of this article).

State Below the Bubble Point Pressure p_b:

1) The reservoir pressure p is less than the bubble point pressure p_b.
2) Aside from the volume expansion of oil, initial water, and rock matrix, the release of gas from solution occurs as a further drive process. To simplify the calculation of the equilibrium behavior as far as possible, the compressibility equations for the oil, water, and rock phases as well as the real gas equation for the gas phase are used as material laws.

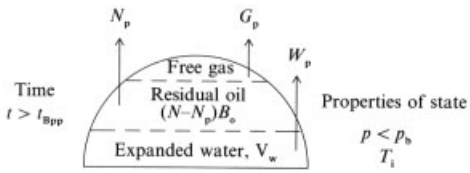

$$V_p = V_{pi} \left(1 - c_f \left(p_i - p\right)\right)$$

$$V_w = V_{wi} \left[1 + c_w \left(p_i - p\right)\right]$$

From the material balance for gas it follows that:

$$\text{Free gas} = (N R_{si} - G_p - (N - N_p) R_s) B_g \tag{2.12}$$

From the displacement process in the pore space it follows that:

$$\text{Free gas} = V_p - (V_w - W_p B_w) - (N - N_p) B_o \tag{2.13}$$

$V_w - W_p B_w$ = net water expansion

$(N - N_p) B_o$ = residual oil volume

Equating (2.12) to (2.13) gives:

$$N R_{si} B_g - G_p B_g - (N - N_p) R_s B_g$$

$$= V_p - V_w + W_p B_w - (N - N_p) B_o$$

Separating reservoir and production terms:

$$N R_{si} B_g - N R_s B_g + N B_o + \frac{N B_{oi}}{1 - S_{wi}}$$

$$\cdot (S_{wi} (1 + c_w \Delta p) - 1 + c_f \Delta p)$$

$$= G_p B_g - N_p R_s B_g + W_p B_w + N_p B_o$$

Combining and rearranging the terms with

$$G_p = N_p R_p$$

gives

$$N \left[(R_{si} - R_s) B_g + B_o + \frac{B_{oi}}{1 - S_{wi}} ((c_f + S_{wi} c_w) \Delta p - (1 - S_{wi})) \right] = N_p [(R_p - R_s) + B_o] + W_p B_w$$

Further rearrangements with

$$\frac{B_o - B_{oi}}{B_{oi} \Delta p} = c_o$$

gives

$$N B_{oi} \left(\frac{B_g}{B_{oi}} (R_{si} - R_s) + \frac{B_o - B_{oi}}{B_{oi}} + \frac{c_f + S_{wi} c_w}{1 - S_{wi}} \Delta p \right)$$

$$N B_{oi} \left(\frac{B_g}{B_{oi}} (R_{si} - R_s) + \frac{c_o S_{oi} + c_w S_{wi} + c_f}{1 - S_{wi}} \Delta p \right)$$

$$= N_p (B_o + B_g (R_p - R_s)) + W_p B_w$$

with

$$c_{eff} = \frac{c_o S_{oi} + c_w S_{wi} + c_f}{1 - S_{wi}}$$

it follows that

$$\frac{N_p}{N} = \frac{B_g (R_{si} - R_s) + c_{eff} \Delta p B_{oi} - \frac{W_p B_w}{N}}{B_o + B_g (R_p - R_s)} \tag{2.14}$$

The ratio N_p/N in Equation (2.14) indicates the recoverable quantity of oil – as a fraction of the oil in place – for the dissolved-gas drive mechanism (first term of the equation) and for the fluid- or rock-expansion drive mechanism (second term of the equation). The third term is normally negligible for reservoirs without an aquifer.

2.3.2. Tarner Prediction Method

A tank model is only capable of describing the energetic aspect of the recovery process. For a complete characterization the adequate inclusion of the flow processes in the porous reservoir rock is absent. Tarner solved this problem by a combination of a material balance and flow equations based on the Darcy law [95].

The calculation is as follows:

For a differential pressure change in the closed reservoir, the expected cumulative oil production is predicted and the cumulative gas – oil ratio calculated by using a gas material balance. To check the accuracy of the assumption, the change in saturation in the porous rock caused by fluid removal and the resulting relative permeabilities for gas and oil are determined. From the flow equations for the gas and oil phases a gas – oil ratio results which is compared with that calculated from the material balance. If the results are identical, the predicted quantity of oil produced during an incremental pressure change is correct. Otherwise, the calculation must be repeated with a new assumption until the two gas – oil ratios are equal.

In the following, the equations for the calculation of the two gas – oil ratios are derived by way of example for a gas-saturated oil reservoir.

Material Balance for Gas. Original gas content = Quantity of gas produced + residual gas

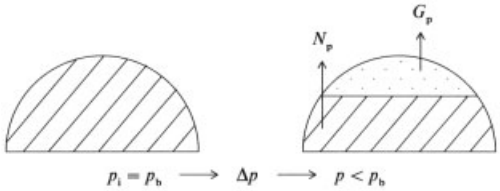

Material Balance for Gas

$$N R_{si} = [N B_{oi} - (N - N_p) B_o] \, 1/B_g$$

Gas cap content

$$+ (N - N_p) R_s + G_p$$

Dissolved gas

$$B_g = \frac{T z p_0}{p T_0}$$

$$G_p = N_p R_p$$

$$R_p = \frac{q_g}{q_0} \quad \text{when } q_g \text{ and } q_0 = \text{constant}$$

Separating the variables:

$$N R_{si} B_g - N B_{oi} + N B_o - N R_s B_g$$

$$= N_p R_p B_g + N_p B_o - N_p R_s B_g$$

$$N [B_o + B_g (R_{si} - R_s) - B_{oi}]$$

$$= N_p (B_o + R_p B_g - R_{si} B_g + R_{si} B_g - R_s B_g)$$

with

$$B_o + (R_{si} - R_s) B_g = B_t$$

$$B_{oi} = B_{ti}$$

it follows that:

$$N [B_t - B_{ti}] = N_p [B_t + B_g (R_p - R_{si})]$$

$$\frac{N_p}{N} = \frac{B_t - B_{ti}}{B_t + B_g (R_p - R_{si})} \tag{2.16}$$

or solved with respect to $N_p R_p$:

$$G_{p1} = N_p R_{p1} = \frac{N (B_t - B_{ti}) - N_p (B_t - R_{si} B_g)}{B_g}$$

The index 1 refers to the gas – oil ratio derived from the material balance.

Change of saturation

$$\text{OOIP} = N B_{oi}$$

$$\text{Residual oil} = (N - N_p) B_o$$

$$V_p = \frac{N B_{oi}}{1 - S_{wi}}$$

$$S_o = \frac{(N - N_p) B_o}{N B_{oi}} (1 - S_{wi})$$

$$S_o = \left(1 - \frac{N_p}{N}\right) \frac{B_o}{B_{oi}} (1 - S_{wi}) \tag{2.17}$$

Dynamic gas – oil ratio equation

$$R_{p2} = \left[\frac{q_g}{q_o}\right]_{T_o, p_o} = \frac{\frac{q_g}{B_g} + \frac{q_o R_s}{B_o}}{\frac{q_o}{B_o}}$$

The index 2 refers to the gas – oil ratio derived from flow equation.

$$R_{p2} = \frac{q_g B_o}{q_o B_g} + R_s$$

Darcy equation (2 immisible phases):

$$q_g = -\frac{k_g A}{\mu_g} \nabla p_g$$

when ∇p_c is rather low it follows $\nabla p_g \simeq \nabla p_o$

$$q_o = -\frac{k_o A}{\mu_o} \nabla p_o$$

$$R_p = \frac{k_g \mu_o B_o}{k_o \mu_g B_g} + R_s \tag{2.18}$$

$$G_{p2} = N_p R_{p2}$$

Each pressure stage can be graphically extrapolated to the convergence of the ΔG_p values or numerically iterated (see Fig. 33).

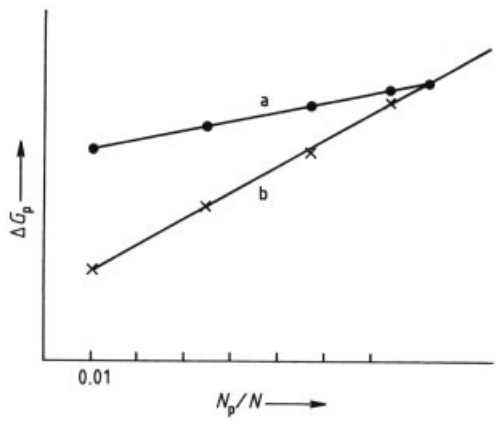

Figure 33. Graphical method of solution in the Tarner method a) Material balance; b) Flow equation

Figure 34 shows the typical development of the produced *gas – oil ratio* R_p plotted against the cumulative oil produced. The gas – oil ratio curve above the bubble point pressure (1 – 2) is constant; when the pressure falls below p_b then initially no free gas reaches the well as long as the required critical gas saturation is not reached (2 – 3) in the surroundings.

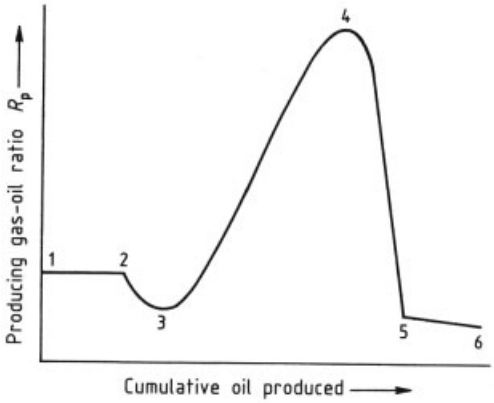

Figure 34. Influence of the critical gas saturation on R_p [88]

Then follows a steep rise of the gas – oil ratio (3 – 4), resulting from the rapid transport of the free gas. The quantity of dissolved gas, which is the source of the gas yield, is limited. Therefore with advancing production time the cumulative gas – oil ratio reaches a maximum, which is then followed by a rapid drop (4 – 5) until the free gas is exhausted. In the last phase of production (5 – 6) it is principally the gas dissolved in the residual oil that is produced.

2.3.3. Reservoir Simulation

The prediction methods discussed in the previous sections are applicable only to problems in which the reservoir can be described by uniform mean properties of the rock and the pore fluids. The individual character of a group of wells or the characteristic drainage volume of their reservoir, and the occurrence of multiphase flow can be described only by reservoir simulation. This entails modeling the whole body of the reservoir and solving, simultaneously or stepwise, the coupled differential equations for pressure and saturation in each of its volume elements. This section can only give a short introduction to the basic principles of the numerical simulation of reservoir behavior. For further details, see [96], [97]. Here the mathematical foundations for an engineering understanding will be treated with reference to two simple models [98].

2.3.3.1. Model for Steady-State Flow

A five-point well pattern with an injection well at the center of the square is taken as the basis. According to symmetry considerations, in a homogeneous formation, only one-eighth of the grid area including the injection well and a production well, need be shown. Figure 35 shows the pattern section dividing the reservoir area into square blocks, the representative portion being bound by two orthogonal sides and the diagonal between injection well and production well.

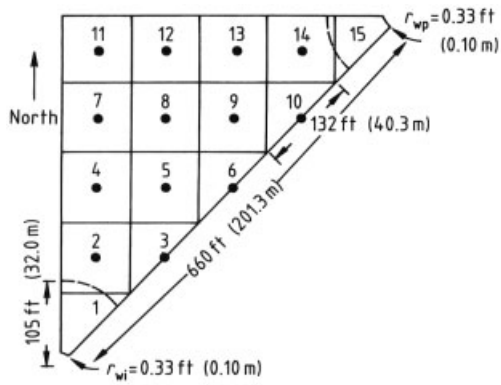

Figure 35. Grid model for steady-state flow [98]

At the center of each block there is a node at which all the block's properties are concentrated. The flow geometry in all blocks is linear, except for the drainage areas of the two wells. Here, radial flow geometry prevails and, instead of a node, there is an external radius r_e and a well radius r_w, representing the equivalent block boundaries. An incompressible medium is assumed to be injected into well 1 and produced from well 15. Because of the prevailing pressure regime, outflow from the nodes is possible only to the north or east and inflow to the nodes only from the south or west. For steady-state flow it is assumed that for each node the sum of inflows and outflows at every instant is zero. This is illustrated here for node 8:

$$q_{7-8} + q_{5-8} = q_{8-12} + q_{8-9} \tag{2.19}$$

If the permeability is equal in all nodes (i.e., in all blocks) and the areas of flow are likewise identical (since the block width is constant), the general form of flow equation for the Darcy law (see Section 2.2.1.1) can be written as follows:

$$q_i = -\frac{Ak}{\mu \Delta x}\Delta p_i \tag{2.20}$$

Since the coefficients before Δp_i are constant in all terms, Equation (2.21) can be written in place of Equation (2.19)

$$(q_7 - q_8) + (q_5 - q_8) = (q_8 - q_{12}) + (q_8 - q_9) \tag{2.21}$$

For node 8 there then follows the pressure balance set up in Equation (2.21) or rearranged to:

$$q_5 + q_7 - 4\,q_8 + q_9 + q_{12} = 0$$

Such a pressure balance can be set up for each of the 13 nodes, in addition radial flow conditions must be included in the pressure balance for nodes 1 and 15. The 15 equations with 15 unknowns are arranged in matrix form and solved for the pressure values at the nodes with, e.g., the Gauss – Jordan or Gauss – Seidel method. The calculation can also be performed with different permeabilities in the individual blocks.

2.3.3.2. Model for Unsteady-State Flow

If the same model is used, but instead of an incompressible fluid a compressible medium is injected into the well, then, instead of a material balance for the individual volume flows in the nodes, a balance for the mass flows must be set up. As already derived in Section 2.2.1.2 from the mass balance in a cylindrical volume element, a variant of the diffusivity equation must be used to describe unsteady-state flows. For linear flow through the volume element of edge length Δx, the following differential equation can be written for the change of flow along the distance Δx:

$$\frac{kA}{\mu}\left(\frac{\Delta p}{\Delta x}\right)_{x+\Delta x}\Delta t - \frac{kA}{\mu}\left(\frac{\Delta p}{\Delta x}\right)_{x}$$

$$\Delta t = A\Delta x\,\phi c\Delta p \tag{2.22}$$

After separating the variables Δx and Δt, the diffusivity equation for unsteady-state flow is obtained (Eq. 2.23):

$$\frac{\Delta p}{\Delta t} = \frac{k}{\phi \mu c}\frac{\Delta\,(\Delta p/\Delta x)}{\Delta x} \tag{2.23}$$

This differential equation is solved by the finite difference method (→ Mathematics in Chemical Engineering, Chap. 7.3.; → Fluid Mechanics, Chap. 5.1.). In this procedure the pressure gradient $\Delta p/\Delta x$ is regarded as the difference of pressure in two successive segments, divided by the distance between the centers of the segments. Figure 36 shows such a linear sequence of segments.

$$\left|\frac{\Delta p}{\Delta x}\right|_{j-1,j} = \frac{p_j - p_{j-1}}{\Delta x} \tag{2.24}$$

$$\left|\frac{\Delta p}{\Delta x}\right|_{j,j+1} = \frac{p_{j+1} - p_j}{\Delta x} \tag{2.25}$$

$$\frac{\Delta\,(\Delta p/\Delta x)}{\Delta x} = \frac{(p_{j+1} - p_j)/\Delta x - (p_j - p_{j-1})/\Delta x}{\Delta x} = \frac{p_{j+1} - 2p_j + p_{j-1}}{\Delta x^2} \tag{2.26}$$

The pressure gradient between segments $j-1$ and j is calculated according to Equation (2.24), and the pressure gradient between segments j and $j+1$ according to Equation (2.25). The change of pressure gradient between segments $j-1$ and $j+1$ is then shown by Equation (2.26). For unsteady-state processes, the change of pressure in the time Δt is superimposed on the change of pressure with position Δx. In Figure 36, the index j designates the position and the index n the time. In the lower part of the Figure 36 the pressure variation at times n and $n+1$ is shown. If Equation (2.23) is applied in general form to define the pressure variation at the beginning and at the end of the time interval, Equations (2.27) and (2.28) respectively result (see below).

The pressure variation at the beginning of the time interval is given by:

$$\frac{p_{j,\mathrm{n}} - p_{j,n+1}}{t_{n+1} - t_n} = \frac{k}{\phi \mu c}\frac{p_{j+1,n} - 2p_{j,n} + p_{j-1,n}}{\Delta x^2} \tag{2.27}$$

and at the end of the time interval by:

$$\frac{p_{j,\mathrm{n}} - p_{j,n+1}}{t_{n+1} - t_n} = \frac{k}{\phi \mu c} \frac{p_{j+1,n+1} - 2p_{j,n+1} + p_{j-1,n+1}}{\Delta x^2} \tag{2.28}$$

Equations (2.27) and (2.28) can give the same result only when the time interval Δt_n approaches zero. If the pressures at time n in all segments are known, the pressure in element j at time $n+1$ can be calculated with Equation (2.27). This method is known as the *explicit solution method.* The advantage of the explicit solution method is that it is very easy to program. However, it is only accurate if very small time steps are chosen.

In contrast to the explicit solution (Eq. 2.27) in which there is only one unknown, i.e., the pressure p_j at time $n+1$, three unknown values are present in Equation (2.28) at the end of the interval (i.e., p_{j-1}, p_j, and p_{j+1}).

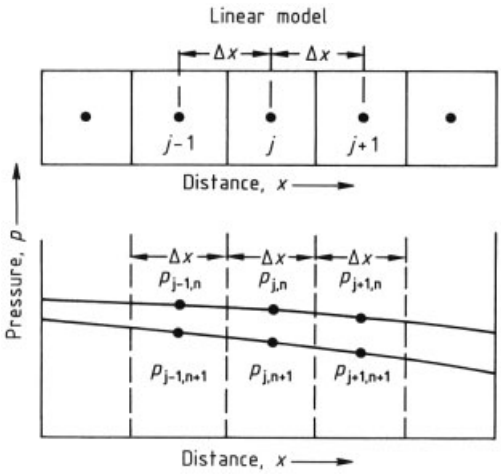

Figure 36. Linear sequence model

If, however, Equation (2.28) is formulated for all three nodes, $j-1$, j and $j+1$, the three unknown pressures can be determined from the three equations. The equations are arranged in matrix form and solved simultaneously either by the Gauss – Seidel or by the Gauss – Jordan method. This procedure is known as the *implicit solution method* and has the advantage of yielding sufficiently accurate results even at quite large time intervals. For modeling problems in multiphase flow, the saturation equation for each block of the reservoir model must be solved in addition to the pressure equation. Hence, the chosen block size of a reservoir model is influenced not only by the availability of geological and petrophysical information but also by the accuracy of the numerical computation method. In regions of the reservoir where relatively rapid changes occur in the transport characteristics – such as pressure gradient, transmissibilities, and saturation changes – fine-mesh grids are necessary for numerical accuracy of the solution method. For complex reservoir studies, simulation models with several ten thousand blocks can be produced which require considerable computer time. There are numerous simulation codes with a dynamicized fine grid size that follows the movement of flood fronts.

Table 3. Types of reservoir simulators

Description	Physical/chemical processes
Multiphase flow simulator	multiphase flow in porous/fissured media solution of gas in oil properties of phases depend on pressure and temperature
Compositional simulator	mixing of phases equilibrium of components between the phases dispersion flow properties depend on interfacial tension properties of phases depend on composition
Thermal simulator	like compositional simulator, in addition : thermal convection, conduction, and radiation flow properties of phase depend on temperature chemical reaction of components of a solid phase (oxidation)
Chemical flood simulator	adsorption/desorption of components retention of phases rheology of aqueous phases properties depend on salt content of the aqueous phase mass transfer between aqueous phase and rock

The powerful computer programs currently available to the natural gas and petroleum industry not only take into account simple processes such as gas dissolution, gravitational force, and capillary pressure effects but also can describe the infuence of compaction in unconsolidated formations and the influence of the diffusion and sorption of components. They can also treat particular rheological properties of the phases, e.g., the structural viscosity of polymer solutions

in polymer flooding the phase behavior during mixing processes, temperature- controlled fractionation of materials (distillation and cracking), thermal conduction, and thermal convection. A survey of common simulator types is given in Table 3.

2.3.4. Statistical Prediction Methods (Decline Curve Methods)

Aside from tank models or flow models to describe and predict reservoir behavior, statistical prediction methods also can be used. These methods are based on observation of the decline of production or pressure in a reservoir; analysis of the mathematical trend; and extrapolation of the decline of production or pressure over the period of prediction. The observed trend curves often show uniform patterns. This is due to the fact that the change in the flow rate is proportional to the pressure change, and the pressure change is linked logarithmically, as a result of the compressibility relation, with the volume removed from a closed system. In many cases, therefore, an exponential decline of the production rate with time can be derived from the trend analysis.

To characterize the decline curve analysis method, some basic definitions are given in advance [98]. The *fractional production decline* is defined as the change of flow rate divided by the initial flow rate

$$-\frac{\mathrm{d}q}{q} = \text{fractional production decline}$$

The change of the fractional production decline with time is called the *production decline rate D*.

$$-\frac{\mathrm{d}q}{q\,\mathrm{d}t} = D$$

In a closed or semiclosed reservoir with no infinitely large, fully active aquifer available for energy replacement, three types of production rate decline can be distinguished:

1) Constant production decline rate:

$$D = \alpha' q^0 \quad \text{exponential production decline}$$

2) Linearly proportional production decline rate:

$$D = \alpha' q^1 \quad \text{harmonic production decline}$$

3) Production decline rate is proportional to a power function of the production rate:

$$D = \alpha' q^n \quad \text{hyperbolic production decline}$$

In order to permit simple prediction methods, the production rate or a function of the production rate is plotted against a function of time or against cumulative production, choosing scales such that straight lines are obtained. The basic method is explained with the aid of exponential and harmonic production declines as examples.

Exponential Production Decline. Figure 37 shows a semilogarithmic plot of the production rate versus time. According to Equation (2.29) the gradient of the straight line is a measure of the production decline rate *D*.

$$D = -\frac{\mathrm{d}q}{q\mathrm{d}t} = \alpha' q^0 = \alpha'$$

after integration

$$\int_{q_0}^{q} \frac{\mathrm{d}q}{q} = -\alpha' \int_0^t \mathrm{d}t$$

$$\ln\frac{q}{q_0} = -\alpha' t = -Dt$$

$$\frac{q}{q_0} = e^{-Dt} \tag{2.29}$$

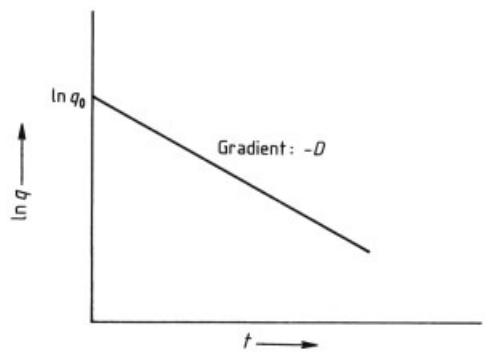

Figure 37. Exponential production decline, production rate versus time

By combining the variables *q* and *t* a new quantity – the cumulative production N_p – is obtained.

$$N_p = \int_0^{t_{N_p}} q\,dt$$

$$\int_{q_0}^{q} \frac{dq}{q} = -\alpha' \int_0^{t_{N_p}} dt$$

$$\int_{q_0}^{q} dq = -\alpha' \int_0^{t_{N_p}} q\,dt$$

$$q = q_0 - D N_p \tag{2.30}$$

A plot of production rate q versus N_p is shown in Figure 38. From this plot the production decline rate D is also obtained from the gradient. As soon as the production rate of a well falls below a marginal value determined by the cost, production must be stopped. The corresponding cumulative amount produced can be determined graphically by extrapolating the production decline curve to the known limiting production rate q_a. The difference between the maximum cumulative production N_{pa} and the present cumulative production N_p gives the reserves of the well or reservoir.

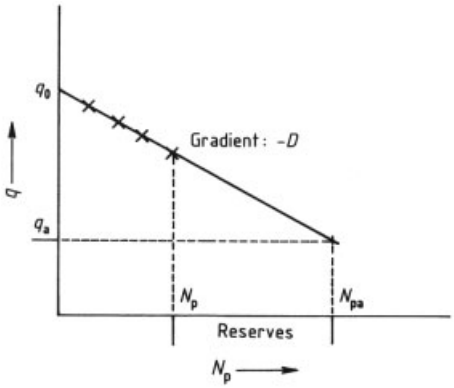

Figure 38. Exponential production decline, production rate versus cumulative production

Harmonic Production Decline. The defining equation for harmonic production decline is:

$$D = -\frac{dq}{q\,dt} = \alpha' q$$

Initial conditions are: when $t = 0$, $D = D_0$ and $q = q_0$,

$$D_0 = \alpha' q_0, \quad \alpha' = \frac{D_0}{q_0}$$

After separating the variables and integrating

$$-\int_{q_0}^{q} \frac{dq}{q^2} = \alpha' \int_0^t dt$$

$$\frac{1}{q} - \frac{1}{q_0} = \frac{D_0}{q_0} t$$

$$q = \frac{q_0}{1 + D_0 t} \tag{2.31}$$

$$\log \frac{q}{q_0} = -\log (1 + D_0 t)$$

For harmonic production decline, the double logarithmic plot of the production rate against the time function $(1 + D_0 t)$ gives a straight line of gradient -1. Again, if a new quantity is introduced by combining q and t a semilogarithmic relation between production rate and cumulative production is obtained after integration (see Fig. 39).

$$N_p = \int_0^{t_{N_p}} q\,dt$$

$$-\int_{q_0}^{q} \frac{dq}{q} = \alpha' \int_0^{t_{N_p}} q\,dt$$

$$\ln \frac{q}{q_0} = -\frac{D_0}{q_0} N_p \tag{2.32}$$

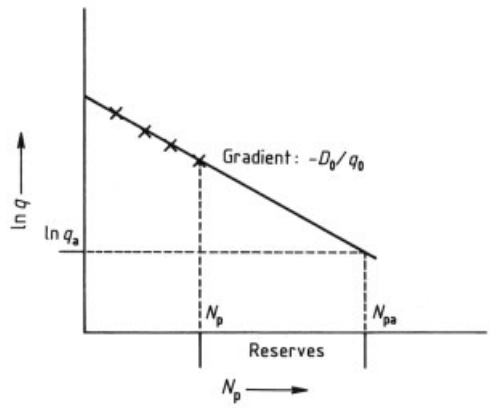

Figure 39. Harmonic production decline, production rate plotted against cumulative production

With the appropriate plot, a straight line of gradient $-D_0/q_0$ is obtained, that again – if the limiting production rate q_a is known – can be extrapolated to give the maximum cumulative production N_{pa}.

Hyperbolic Production Decline. From the equation for the hyperbolic production decline rate, after rearrangement of the variables and integration, the following relations are obtained:

Initial condition: $t = 0$, $q = q_0$, $D = D_0$

$$D = -\frac{\mathrm{d}q}{q\mathrm{d}t} = \alpha' q^n \qquad \alpha' = \frac{D_0}{q_0^n}$$

$$q = \frac{q_0}{(1 + n D_0 t)^{1/n}} \tag{2.33}$$

$$N_\mathrm{p} = \frac{q_0^n}{(1-n)\, D_0} \left(\frac{1}{q_0^{n-1}} - \frac{1}{q^{n-1}} \right) \tag{2.34}$$

A plot of production rate versus $\log(1 + n D_0 t)$ is given in Figure 40.

The analysis of the production decline can also be based on the fraction of oil in the total production rate or on the water cut in the production, instead of on the oil production rate. The same relations then apply. In order to eliminate the influence of individual measurement inaccuracies or of unrepresentative sampling, it is advisable to integrate or smooth the dependent variables q and the water cut f_w respectively over fairly long periods, e.g., over a month's or year's production. These steps enable improved reliability of the prediction methods.

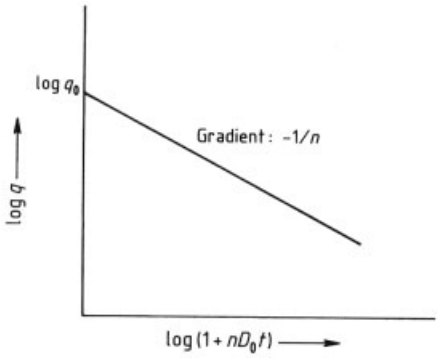

Figure 40. Hyperbolic production decline, production rate plotted against the time function

3. Deep Drilling and Production Engineering

3.1. Deep Drilling Engineering [119–123]

3.1.1. Introduction

To produce raw materials and fuels from the Earth's crust, two basic problems have to be solved: Firstly, the deposit must be discovered, its parameters determined, and then investigated whether it can be developed by considering all the available data (see Chaps. 1 and 2). Secondly, the raw materials present in the deposit have to be produced – i.e., transported to the surface – economically [119]. To solve these problems, particularly for liquid and gaseous materials, deep drilling is indispensable. Boreholes for locating and exploring deposits are known as *exploratory wells*, while boreholes used for producing from the deposits are known as *production wells*. Successful exploratory wells are converted into production wells using suitable equipment.

Exploratory wells must be drilled in such a way that they provide accurate information on the rock formation drilled through and on the deposit to be explored. This is done by sampling the rocks, e.g., in the form of drill cores, sampling the materials present in the rocks (e.g., water, oil, gas), gathering data from a great variety of geophysical borehole measurements, carrying out productivity tests, and then measuring the resulting behavior of the reservoir pressure.

Production wells must be drilled and completed in such a way that the cost of subsequent production is minimal, as much as possible of the reservoir contents can be produced economically, optimum use is made of its productive capacity, a minimum number of wells is required to develop the reservoir and trouble-free production can be achieved, preferably over the reservoir's lifetime [120].

Boreholes compete with mining methods as a means of prospecting for and producing raw material and energy, and are in many cases, superior to them from economic, operational, and safety viewpoints. Deep drilling engineering is therefore becoming increasingly important with th development of deposits and reservoirs that lie ever deeper and in inhospitable regions, under the oceans, or in perpetual ice.

3.1.2. History [124–126]

Deep drilling is far from being a new technology (Fig. 41). As early as 600 B.C. CONFUCIUS reports that wells have been sunk in China to a depth of about 500 m into salt brine with hammer drilling devices. In Europe, both in ancient

times and in the Middle Ages, wells were sunk to obtain water and later for prospecting shallow ore lodes. Development of drilling engineering and well depths that could be achieved stagnated for many centuries. The search for raw materials (potash, iron ore, and coal) in the course of the industrialization in the 19th century forced the development of deep drilling engineering. At that time, drilling depths already reached 2000 m.

In Europe during the second half of the 19th century and the start of the 20th century a extensive search for raw materials took place. It was pioneers like von Oeynhausen, Fauck, Craelius and Anton Raky who encouraged the progression of deep drilling by, for example, developing high-speed rotary percussion drilling with circulating drilling mud. Only then could the frozen boreholes for shaft sinking of mines for coal, potash, and rock salt be produced.

The rotary drilling method, developed in the United States at the end of the 19th century for the exploration and development of oil deposits, was an important step forward. Since then, rotary drilling has been continuously developed and today is still the dominant drilling method by which depths of more than 9000 m have been reached in commercial boreholes. The present depth record has been achieved with the Kola SG 3 well in the Commonwealth of Independent States, which reached a depth of 12 261 m, using turbodrilling.

3.1.3. Rotary Drilling [127–131]

The circulating viscoplastic drilling mud and powerful roller cone drill bits are the most important characteristics of rotary drilling. These, in combination with other typical features, have contributed to the overriding importance of rotary drilling in the exploration and development of deep and high-pressure reservoirs. Spectacular blowouts, which were numerous in the past, are now rare.

The principle of mode of action in rotary drilling is as follows (Fig. 42):

By means of a rotary table (e), a torque is applied to the drill string that consists of individual drill pipes (steel tubes). The drill string is suspended in the derrick by a hook (b) attached to a traveling block. This hook, which allows the drill string to rotate, can be moved vertically with a drawwork on the rig floor. The drill string transfers the torque to the drill bit (n), whose impact on the rock is controllable via the weight of the drill string and drill collars (m) (massive thick-walled tubes arranged above the bit), which is not taken up by the drilling hook.

The rotary table, which is a "surface drive", can be replaced by a hydraulic drilling motor immediately above the bit. This *downhole motor*, which can be a positive-displacement machine on the Moineau principle or a multistage turbine, is installed in the drill pipe and driven by the circulating mud.

Drilling Tool. The drilling tool, which is adapted to the rock through which the hole has to be sunk, performs the drilling work on the hole bottom. This can be done by machine drilling (cutting bit), fragmentation drilling (roller cone bit, Fig. 43 A – E), or abrasive drilling (diamond bit Fig. 43 G).

The key factor with regard to drilling costs is the rate of penetration (v). The rate of penetration depends on the rotary speed of the bit n, the weight on bit P, the drillability of the rock S, and the hole diameter D. In the empirical relation

$$v = \frac{n P^2}{D^2 S^2} k$$

a "drillability coefficient" k combines factors such as rock properties, mud parameters, pressure conditions, bit type, wear of the bit teeth, etc. It is assumed that the hole bottom is clean.

Rotary speed and bit load cannot be increased without limit. However, the "penetration factor", defined as the product of the bit load and bit rotary speed, should be as high as possible.

The *power required* to turn the drill pipe, which depends on the mud lubrication properties, rises with the square of the drill pipe diameter and linearly with the depth. With increasing depth the torque imparted to the bit by a surface drive becomes smaller.

The *bit's mechanical power* is directly proportional to the product of torque and rotary speed. In addition, due to the mud stream emerging from the bit, *hydraulic power* is available at the bit, which is equal to the pump power minus the pressure drops in the mud system.

With limited mud pressure due to conventional borehole and pipe sizes, the optimum hy-

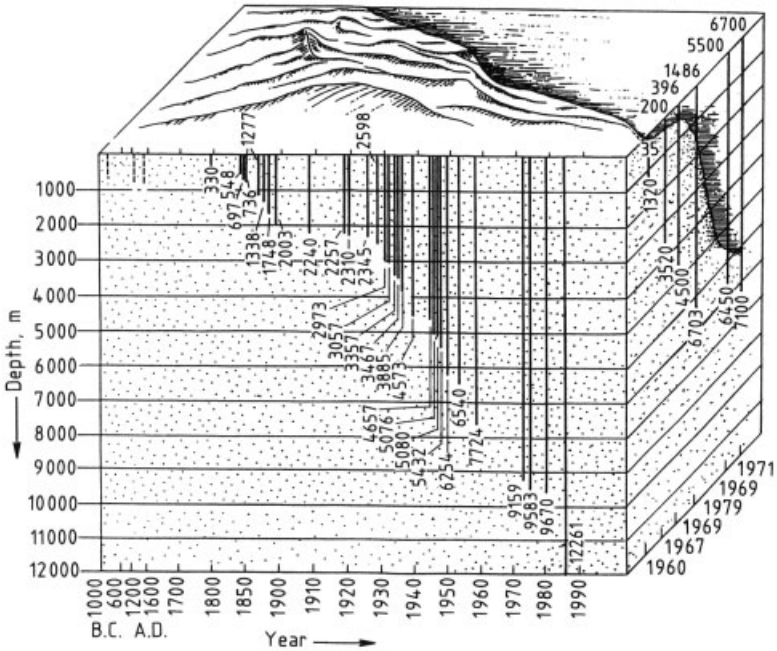

Figure 41. History of drilling depth

draulic power at the bit is reached at ca. two thirds of the power required to pump the mud. The hydraulic power is much greater than the mechanical power, especially with increasing depth. The bit shape, nozzle arrangement, and diameter are chosen so that all requirements for the hydraulic pressure conditions, cleaning of the hole bottom, lubrication, and cooling of the bit, as well as for securing the cuttings' removal capacity are met.

The resistance to the flow of mud on its return through the annulus between drill pipe and borehole wall requires hydraulically optimum compatibility between borehole and drill pipe diameters to achieve minimum resistance to flow. The minimum flow velocity in the annulus should be 0.5 – 0.75 m/s. The pump power must be so designed that sufficiently high pressures and large volumes can be handled.

The drill string is loaded by tension (the drill pipes' own weight), torsion, bending (resulting from borehole deviations), constant or dynamic internal pressure and buoyancy forces, and periodical vibration. In some cases, the superposed stresses in drilling can only be calculated with difficulty.

Drilling Mud. The cuttings are carried in the annulus between drill pipe and borehole wall by the mud. A thixotropic liquid, containing bentonite and cellulose for example, is used as the drilling mud. Its functions include cooling and lubricating the drilling tool and cleaning the borehole bottom, discharging the cuttings, and supporting the borehole wall (hydraulic casing). The mud is pumped with piston pumps through the drill pipe to the bit, from which it emerges at high speed to ensure that the cuttings are removed as completely as possible. The cuttings and finer solids are removed from the mud at the surface and the mud is then recycled by pumping it back again into the hole. Vibrating screens, centrifuges, and hydrocylones are used to seperate the solids (Fig. 44).

The mud parameters (e.g., the rheological properties, the mechanical, and thermal state of the colloid system) are adjusted according to the drilling speed, pump power, bit type, borehole diameter, pore pressure in the rock, temperature, nature of the rock drilled (mud loss zones in the drilled formations), borehole inflows, and solids content.

Mud densities between 1.10 and 1.40 g/cm^3 are usually sufficient for compensating the reser-

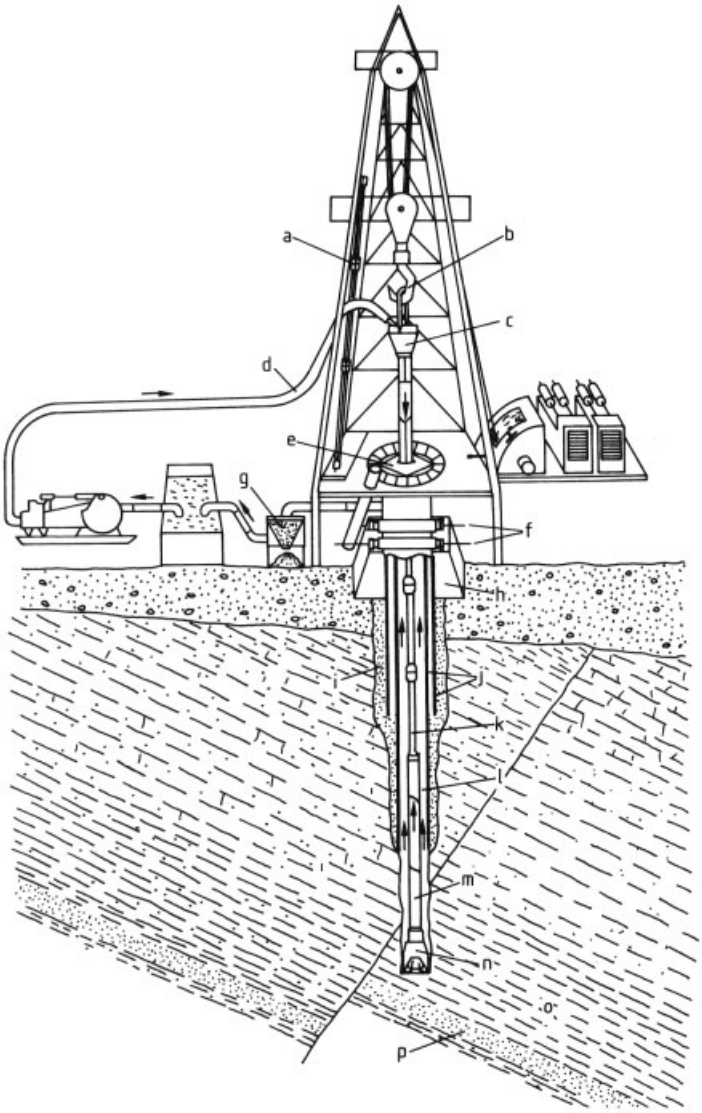

Figure 42. Diagram of a rotary drilling system
a) Parked drill pipe; b) Drill hook; c) Rotary swirrel; d) Flexible hose; e) Rotary table; f) Preventer; g) Vibrating screen (shaker); h) Pit; i) Cement sheath; j) Casing; k) Hollow drill pipe; l) Circulating drilling mud; m) Drill collar as part of the drill string; n) Roller cone bit; o) Rock strata; p) Reservoir rock horizon

voir pressure and the "trip margins", i.e., the piston effect and inertia forces during pulling the drill string. In special cases, however, values below 1 g/cm^3 or up to 2.5 g/cm^3 are also required.

The carrying power of the mud for the cuttings is given by:

$$F\theta = V\,(\varrho_{\mathrm{G}} - \varrho_{\mathrm{m}})\,g$$

where F is the surface area of the cutting particles (cm^2); V the volume of the cuttings (cm^3); ϱ_{G}, ϱ_{m} are the density of rock, and mud, respectively (g/cm^3), g is the acceleration due to gravity (dm/s^2), and θ the static shear stress (dyne/cm^2).

The filtration of water from the mud in porous rock strata, with the formation of a filter cake, proceeds according to Darcy's law of filtration.

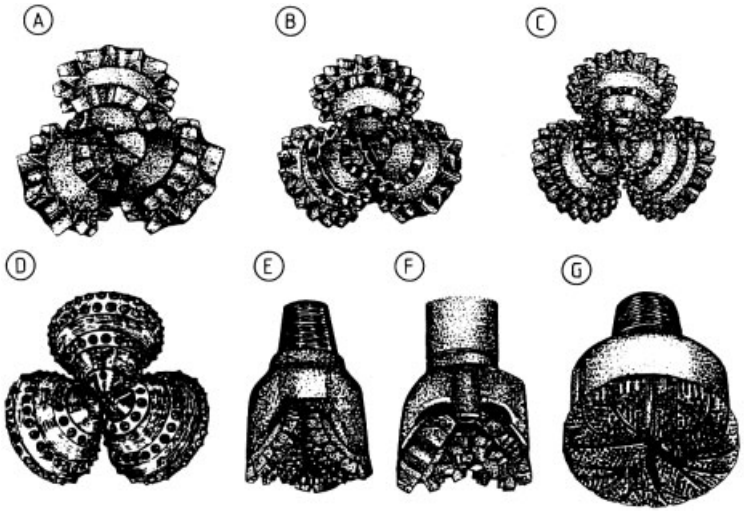

Figure 43. Various types of bit
A) Roller bit with coarse teeth (for soft formation); B) Roller bit (for medium hard rock); C) Roller bit with fine teeth (for tough rock); D) Roller bit with hard metal inserts (for very hard formations); E) Roller cone bit; F) Jet bit; G) Diamond full-hole bit

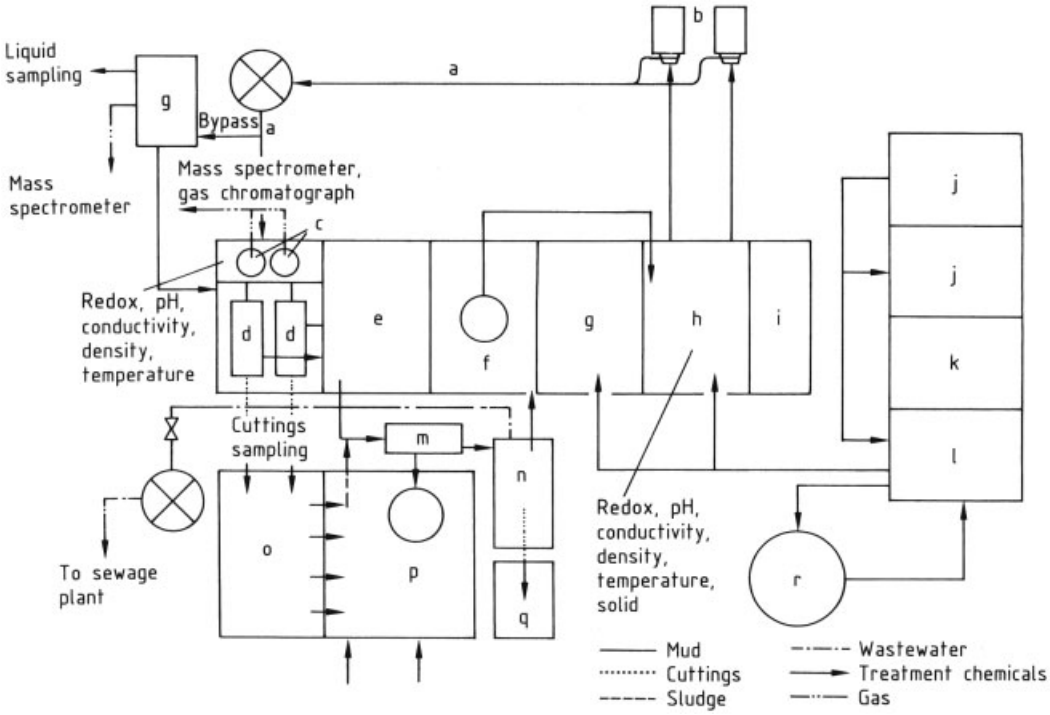

Figure 44. Processing of drilling mud, schematic diagram
a) Flow measurement; b) Mud pumps; c) Gas traps; d) Sieve; e) Centrifuge tank; f) Degasser tank (vacuum degasser); g) Intermediate tank; h) Suction tank; i) Pill tank; j) Water tank; k) Reserve tank; l) Mixing tank; m) Moineau pump; n) Centrifuge; o) Cuttings pit; p) Sludge pit; q) Container, cuttings sampling; r) Mixer; s) Separator sampler

The quantity of filtrate is proportional to the permeability of the filter cake, the applied pressure difference, and the active filter area, and inversely proportional to the filtrate viscosity and the filter cake thickness.

Pressure and temperature changes, particularly in deep holes, make it difficult to assess and calculate the formation of the filter cake. The filter cake can attain a high mechanical strength and cause drilling difficulties. Prolonged filtration can also lead to considerable reservoir damage.

The electrical conductivity of the mud and its pH value and contents of chloride, sulfate, calcium, and hydrogencarbonate ions are continuously monitored. Additives used to adjust the mud system include:

1) Clay materials (e.g., bentonite, attapulgite)
2) Weighting materials (e.g., barite, hematite)
3) Viscosity-lowering substances, dispersants, protective colloids
4) Emulsifiers, lubricants (e.g., oil, to reduce internal friction)
5) Inhibitors

There are water- and oil-based drilling muds. Aside from the technical drilling requirements for the mud, the following requirements should be met: the formations drilled through remain identifiable (mud-logging see below, indication of traces of gas and oil); borehole measurements, tests, core drilling, and other borehole studies can be carried out; the mud causes as little damage as possible to the formations drilled through (obstruction, swelling, mud losses); and corrosion of the drilling equipment is limited.

Drilling muds for use in geoscientific research such as in the German Continental Deep Drilling Program (kontinentales Tiefbohrprogramm, KTB) the main well must satisfy not only the drilling engineering requirements but also those of the geoscientists. For example, such drilling mud must have adequate rheological properties, lubrication, and borehole stabilization without influencing the geoscientific analysis or even disturbing it.

Data on the drilling parameters, mud properties and lithology are collected and stored in a *mud logging system* by modern data acquisition and a computer system for drilling rigs. A network of monitors, work stations, plotters, and printers gives continual access to continuous and discontinuous data and information.

The hole tapers in sections from top to bottom. Each of these sections is secured by means of a string of casing, which is bonded by cementing into the surrounding formation (Fig. 45).

3.1.4. New Developments in Deep Drilling Engineering [131–143]

Superdeep Wells. Commercial wells for hydrocarbons with final depths of at least 5000 m are common nowadays, especially in Europe, where exploration has moved to even deeper strata – at present the Rotliegendes. The technology, especially the control of high pressure and temperature, is essentially available for this depth range. However, considerable development work remains to be done in certain areas, especially to improve the economics of drilling and the operating life of the well.

Superdeep scientific wells such as that of the KTB, which will be more than 10 km deep, represent an intensification of the requirements and hence are of importance for technology development (Fig. 46).

A few degrees' change in the inclination and azimuth of the well bore would cause such large lateral forces, abrasive loads, and wear that the well would have to be abandoned. The 10 000 m drill string is analogous to a wire 1.5 mm thick and 100 m long, at whose lower end a drilling tool has to be controlled with an accuracy in the 0.1 mm range. Active vertical drilling systems have been developed that continuously measure the inclination and direction of the borehole. They pulse the values through the mud to the surface, so that compensatory steering can be applied to the tool by extending hydraulic pistons or ribs (Fig. 47). Straight, superdeep vertical holes are therefore a particularly demanding variant of directional drilling engineering. The size and performance data of the drilling rig for a superdeep hole clearly exceed those for conventional extra-heavy rigs.

Horizontal Drilling. The second main focus of development in deep drilling engineering is directional drilling where the drilling of holes along a defined course over long distances and

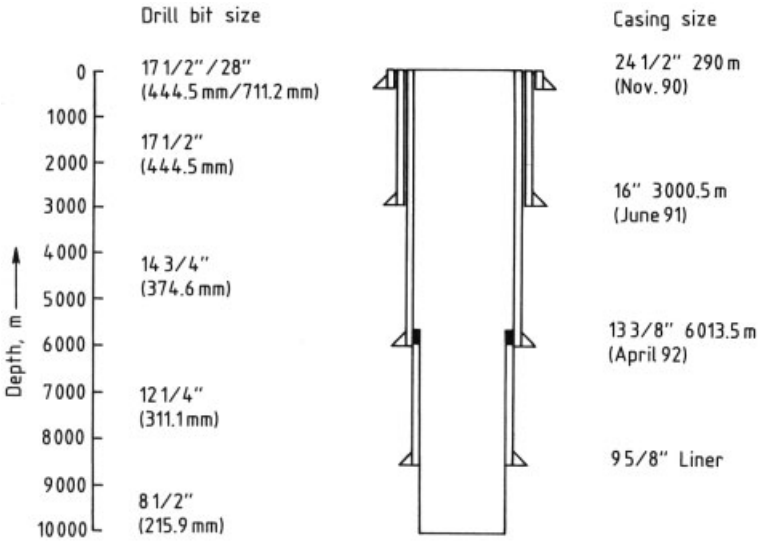

Figure 45. Continental Deep Drilling Program (KTB), drilling and casing diagram

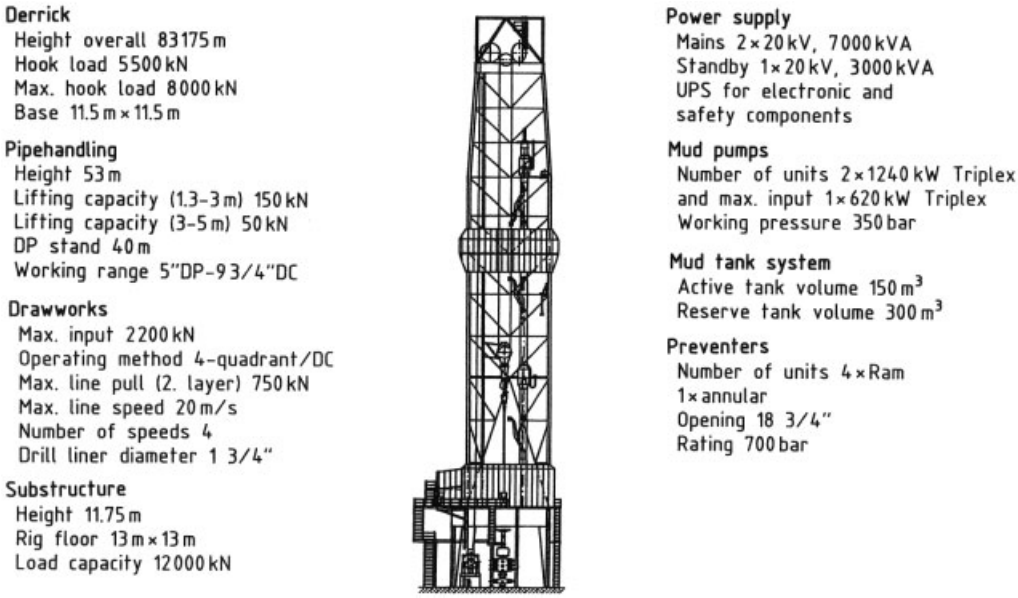

Figure 46. Drilling rig UTB 1 for the 10 000 m main well of the Continental Deep Drilling Program (Germany)

at considerable depth in the underground rock is now possible and economic (Fig. 48).

Offshore, virtually all wells are directional drilled holes, and the oil and gas fields are developed with up to 60 wells per platform. *Onshore*, the classical applications of horizontal deep drilling include recovery under residential areas and traffic routes, borehole deviation in the event of damage, and relief wells in the event of blowouts. A directional drilled hole is produced by using mud-driven drilling motors. The deviation is created with a bent shaft (bent sub) combined with the drilling motor, or the motor itself is tilted. In Western Europe, 85 % of the onshore wells are directional drilled.

Inclination and azimuth (i.e., the course of the borehole) are continually measured and the values transmitted to the surface via pressure

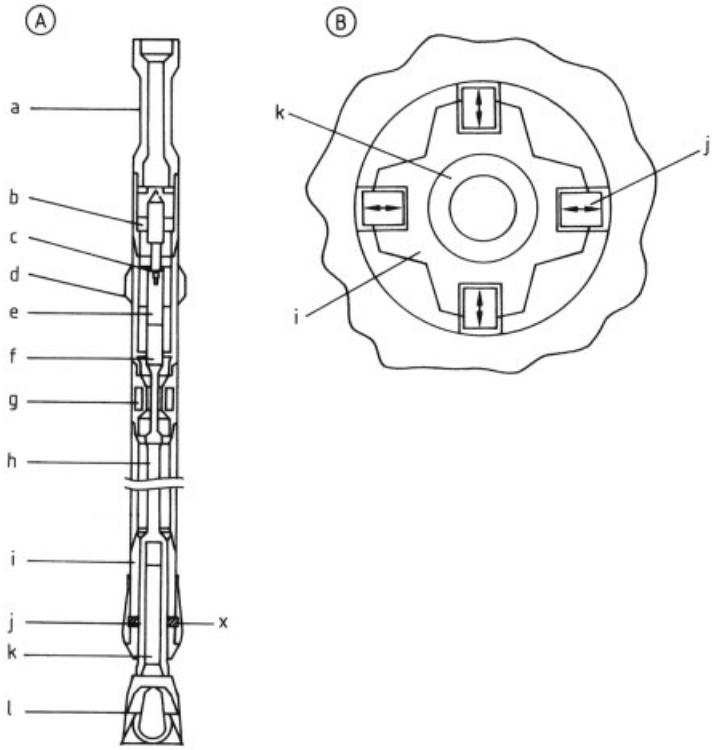

Figure 47. A) Active steering vertical drilling system; B) Cross section of X;
a) Flex sub; b) Data pulser; c) Sensors; d) Top stabilizer; e) Electronic pack; f) Generator; g) Control valves; h) PDM – downhole motor; i) Nonrotating body; j) Hydraulic pistons with movable ribs; k) Rotating shaft; l) Drill bit

pulses through the mud column. From the surface the azimuth can be corrected accordingly. The continual checking, known as *measuring while drilling* (MWD), is crucial for success. The azimuth is measured by magnetometer or, if in the neighborhood of ferromagnetic masses, by gyro compass. The inclination is evaluated by accelerometers (pendulum). In addition, these instruments indicate the tool face, i.e., the angle between the desired direction and magnetic north (notch and bead). The availability of powerful, slow-running drilling motors together with the MWD technique are prerequisites for successful directional drilling engineering.

The inclination build-up of a hole to be deviated into the horizontal can be abrupt or gradual (see Fig. 49). At an inclination change of 10°/10 m, the resulting deviated section has a radius of 6 m, whereas at 1°/10 m, the radius is ca. 600 m. Holes have been deviated to reach destinations of up to 3000 m, and holes exist which have been drilled horizontally for almost 1000 m.

3.2. Investigation of the Reservoir Rocks [144], [145]

With suitable drilling tools, rock samples (mud samples, drill cores) are obtained from the reservoir rock zone and can subsequently be examined comprehensively. Special geophysical measurement (logging) and the first, brief productivity tests (drill-stem tests) are also carried out.

Core Drilling. During core drilling, the drilling bit (full-hole bit) is replaced by a core barrel with crown (ring bit). Single- or double-tube core barrels are used, some with diamond-type cutter heads. Single-tube core barrels have the drawback that the core is directly exposed to

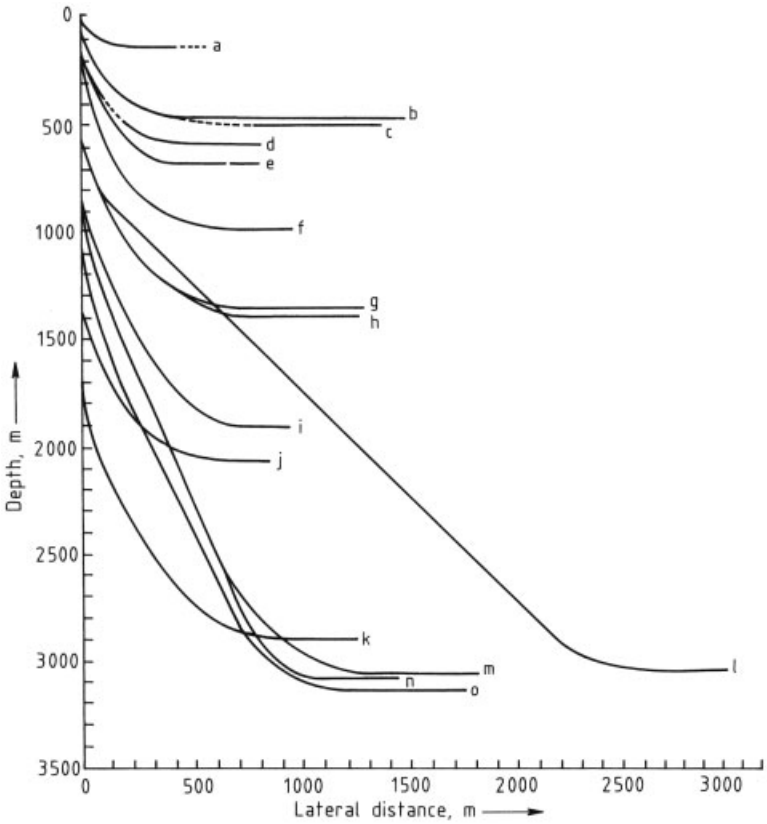

Figure 48. Selected horizontal wells curvature
a) Texaco, 1981 (horizontal section >300 m); b) Esso Resources Canada Ltd., Cold Lake 1978 – 1984 (horizontal section 1000 m); c) Esso Resources Canada Ltd., Norman Wells, 1981 (horizontal section 500 m); d) Elf France, CR 163, 1986 (horizontal section 360 m); e) Elf France, LA 90/LA 91, 1980 – 1981 (horizontal sections 270 and 470 m); f) BP 1985 (horizontal section 324 m); g) Elf Italy, RSM 6, 1982 (horizontal section 600 m); h) Elf Italy, RSM 211, 1986 (horizontal section 600 m); i) Preussag, 1984 (horizontal section 200 m); j) Marsvoskoe, 1968 (horizontal section 320 m); k) Elf France CLU 110, 1983 (horizontal section 330 m); l) Sohio, Y 20, 1986 (horizontal section 305 m); m) Sohio, B 30, 1985 (horizontal section 452 m); n) Sohio, JX 2, 1986 (horizontal sction 415 m); o) Sohio, BW 1, 1985 (horizontal section 579 m)

the mud (core losses); With doubletube core barrels this is prevented by the inner tube of steel or rubber (Fig. 50). Wedgejaw catchers, flap catchers, or combinations of both prevent the core from falling out. Orientation devices and core markings allow an accurate determination of inclination (dip), course (strike), and pattern of the strata drilled through.

Since 1985, remarkable successes have been achieved by a combination of rotary and mining drilling methods. The *mining drilling method* involves rapid rotational drilling with thin-lipped diamond crowns and a double tube core barrel. Externally and internally smooth drill pipe is used which almost fills the whole borehole cross section. After boring out the core, the inner core barrel is left on the cable (wireline coring), so that the bit can remain on the bottom until it is worn out (Fig. 51). The thin-lipped crown and thus the distinctly smaller area of rock which has to be destroyed, makes greater drilling progress possible in comparison with rotary drilling technology. With the combination of both methods the drilling progress can be doubled with complete core recovery; by use of the wireline technique, the number of round trips can be reduced; and by use of a rotary drilling rig, the time for the round trips can be reduced.

Aside from core drilling, cores can also be shot, drilled, or cut (slicer) from the borehole wall.

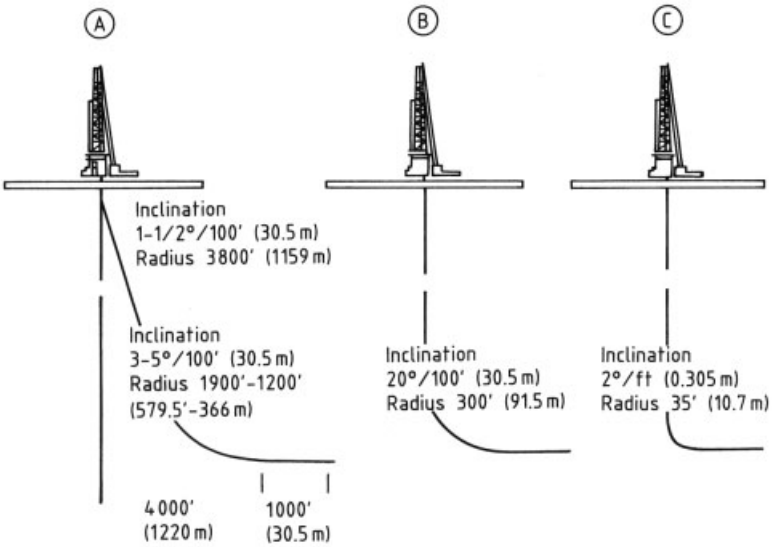

Figure 49. Horizontal drilling strategies
A) Conventional; B) Drainhole, medium radius technology; C) Drainhole, short hole technology

The *core investigations* provide measurements on pore filling, pore volume, liquid saturation, capillary pressure, wettability, grain sizes, chemical composition, petrographic relationships, and strength, etc.

Geophysical Borehole Measurements. Continuous coring is impossible for economic reasons. Geophysical borehole measurements allow the continuous determination of rock parameters, especially the porosity, the original saturation conditions, the nature of the pore fluid, the type of rock, the mineral composition, and the bed thickness.

The laboratory examinations of the core samples taken point-by-point, however, take longer.

A variety of instruments is available for borehole measurements. The methods used can be divided into four main groups

1) Electrical methods: macroresistance, microresistance, conductivity, scanner technique, dip meter
2) Sonic measurements: sonic velocity, travel time, amplitude, reflection, refraction
3) Radioactivity measurements: natural, induced gamma, and induced neutron radioactivity, spectral analysis
4) Determination of geophysical fields: temperature, pressure, stress, magnetics, Earth's gravitation

The temperature limit of application of geophysical methods is ca. 250 °C, corresponding to about 8000 m borehole depth.

The data from borehole measurements are subjected to comprehensive and complex methods of interpretation, for which powerful computers and specialists are required. The results obtained describe the rock, its pore filling, and the geology in situ. Greater reliability is obtained when more drilled rock material from cores and cuttings have been used for calibration and correlation of the measurements.

Tests. In (productivity) tests, qualitative and quantitative conclusions about the inflow process from the formation to the well are obtained in order to decide whether a well allows economic production. At the start of a test, the formation interval to be examined is isolated from the rest of the formation in order to exclude possible effects from other beds and mud pressure. After pressure relief of the interval, the rate of flow to the well, the pressure, and temperature are measured and plotted as a function of time. The inflow into the drill string is then stopped and the pressure build-up which then occurs is recorded. A second flow cycle is initiated by reopening the drill string.

In principle it is also possible to use a test device on a self-supporting electric cable. Here,

however, the inflow rate is strictly limited for design reasons, so that only zones near the boreholes can be tested.

In order to save the casing costs of wells which may later, prove to be uneconomic, tests are carried out in the uncased borehole during or immediately after the end of the drilling operations. For stability reasons, the test duration in the uncased and therefore unbalanced, unstable borehole (drill-stem test) must be kept short, whereas in later capacity tests in the cased hole the test duration is theoretically unlimited.

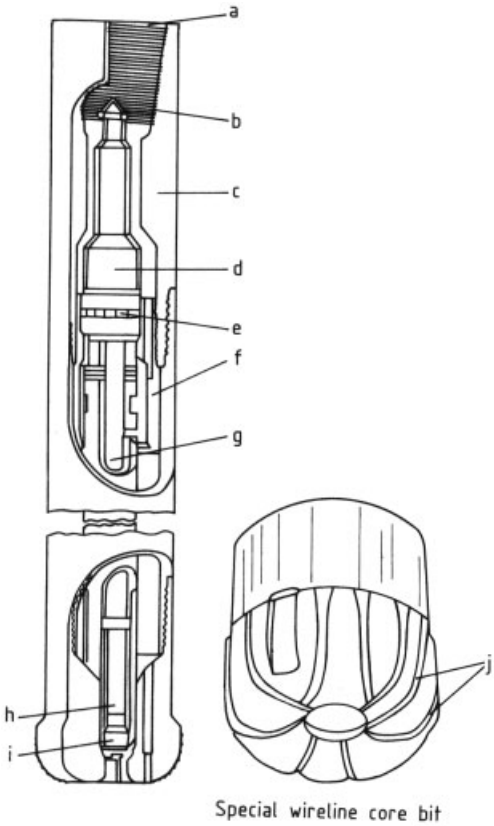

Figure 50. Schematic diagram of double core barrel
a) Coupling to drill pipe; b) Overshot for pulling out with wireline; c) Outer tube, rotates core bit; d) Inner tube, non-rotating; e) Bearing; f) Mud passage; g) Ball valve; h) Core catching spring; i) Core catching cone; j) Mud channels

The interpretation of the measurements must provide information on the following important parameters:

1) Initial reservoir pressure
2) Flow pressure as a function of the amount produced per unit time (productivity index)
3) Shut-in pressure
4) Reservoir and borehole temperature
5) Permeability and permeability capacity
6) Drainage radius of the well
7) Location of formation damage

During these tests, samples of the reservoir contents are also obtained, so as to carry out chemical and physical studies. The results serve as a basis or aid to planning decisions on: the completion of the well, well and reservoir treatments, the production method, the surface installations, the further development of the field, and exploitation strategy.

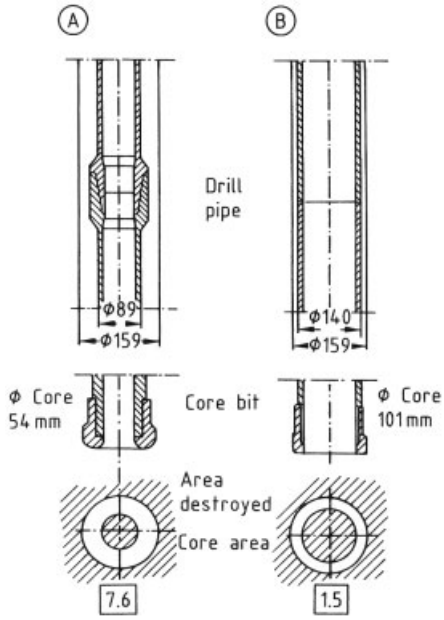

Figure 51. Comparison of oil and mining core drilling
A) Oil core drilling (double tube core bit); B) Mining core drilling (cable core bit)

3.3. Casing and Cementation

Casing [146]. The casing of a borehole usually consists of several concentric strings of casing pipe, which are installed according to the geological – engineering requirements at different depths (Fig. 42). It protects the borehole wall against scouring and blowouts, seals permeable horizons to prevent undesirable inflow or outflow, and prevents sloughing from the formation

into the hole. The casing is in part already inserted during the drilling operation, in order to protect problem sections of borehole. It represents a considerable cost component of the well.

With *exploratory wells*, the planning of the casing program is rather difficult, since knowledge of the formation is still imperfect. For *production wells* a standard program is usually set up, in which tube dimensions, wall thicknesses, and landing depths are optimized.

The casing can either be run to the surface and landed in the wellhead, or hung in the borehole as a so-called liner in a string of casing already installed in the borehole. To ensure a tight bond between casing string and formation, the casing is cemented, either up to the surface or only over a certain length of the borehole.

The casing is subject to a multiplicity of loads, which generally arise in combination:

1) Differential pressure between the outside and inside due to hydrostatic pressure and the pressure of the fluids in the pore space
2) Radial component of the formation stress
3) Tensile stress from the string's own weight, especially during installation
4) Bending stress in holes with changes of direction, especially in horizontal holes
5) Thermal stresses

Since the individual stress contributions superpose one another, precise assessment of them e.g., by landing under tension, is essential. A series of standardized material grades and casing joints is available to meet the requirements.

Tubing String. The tubing string transports the fluid produced to the surface. The tubing is landed in the wellhead at the surface and anchored, if necessary, at the lower end. The annulus between casing and tubing can be sealed with a packer if the conditions in the well require it. *Packers* are usually elastically deformable sealing elements (of elastomers), which are so combined with peripheral sliding wedges that during up and down movements the wedges grip, the seal deforms, and the annulus between tubing and casing is sealed. Tubing, like casing, is manufactured in a range of standardized material grades and dimensions.

Cementing [134]. The cementing of the casings serves the following purposes:

1) Massive bond of casing and formation, in order to divert the forces resulting from the loading of the casing into the formation, i.e., absorption of the axial and radial stresses which act on the casing string
2) Isolation of permeable formations, in order to prevent flow between different rock strata and, in particular, the contamination of the groundwater
3) Corrosion protection of the casing

The cement is mixed with water to form a cement slurry, and a wide variety of additives are added to control the rheological properties, the setting time, the final strength, etc. The slurry is then injected with high pressure pumps through the borehole into the annulus between casing and formation.

The cements used in the oil and gas industry are standardized according to API. Table 4 gives some important characteristics of various API cements. Aside from the cement quality, the technical execution of the cementing is of crucial importance for the quality of the cement sheath. For control, the *cement bond log* is used: this is an acoustic measurement method for determining the bonding of cement and casing or of the cement with the formation.

For cementing at high temperature, Class G cements are used, to which SiO_2 flour is added to avoid a decrease in compressive strength and an increase in the permeability of the cement. The effect of the added SiO_2 depends on the formation of hydration products, for instance according to

$$2\,Ca(OH)_2 + SiO_2 \rightarrow 2\,CaO \cdot SiO_2 \cdot 1.25\,H_2O + 0.75\,H_2O$$

The elevated temperature lowers the hardening time, so retarders and antifriction agents must be added.

3.4. Production Engineering [148], [149]

3.4.1. Objectives [150–153]

The purpose of the exploitation and production planning of hydrocarbon reservoirs is to produce optimum amounts of saleable (transportable) products at minimum cost and with close attention to all aspects of safety and ecology. In initial

Table 4. Characteristics of some API cements [147]

	Class		
	A	D, E, F	G, H
Density, g/cm3	3.14	3.14	3.12
Specific surface, cm^2/g	1500 – 1900	1100 – 1500	1600 – 2000
Optimum water content of the slurry, wt %	46	38	44
Density of the slurry, g/cm^3	1.87	1.97	1.89
Temperature range, °C	75	110 – 160	260 – 370
Sulfate resistance	normal	medium/high	medium

stages of field development, with few wells and few drill-stem and production tests available, the base for planning is not valid enough. However, the data obtained from continuous production and field development lead to continual refinement of production planning, in which reservoir and engineering are regarded as a synergism. Only in this way can an optimum overall technical and economic result be achieved.

Production engineering has to solve the following problems in *oil production*:

1) The time for water injection – when not using it to adjust the pressure above the bubble point (see Section 3.4.3) and/or for adjusting the dynamic level to below a certain material stress, corresponding to the fatigue endurance limit of the pump rods
2) Dependence of the productivity index (IPR) on viscosity of the oil and water cut
3) Gas production and availability in the gas lift method (see Section 3.4.3)
4) Advantages and disadvantages of the artificial lift methods

and for *gas production*:

1) Occurrence of toxic and/or other problematic substances, e.g., hydrogen sulfide, elementary sulfur, carbon dioxide, nitrogen, which requires complex process engineering installations (see also, → Natural Gas)
2) Heterogeneous multilayer and selective water incursion

The safety and ecological aspects have always been important; nowadays it is preeminantly so. In the second half of the 19th century, a blowout was proof of discovery after finding the reservoir by drilling and a reason for celebration. Thanks to the continual improvement of technical equipment and the regular training of technical staff, blowouts have become rare, but they can never be totally avoided due to human error.

Over wide areas, oil and gas production engineering have common principles, which are treated in the following section whereas specific questions of production engineering for oil are dealt with in Section 3.4.3, those for gas in Section 3.4.4.

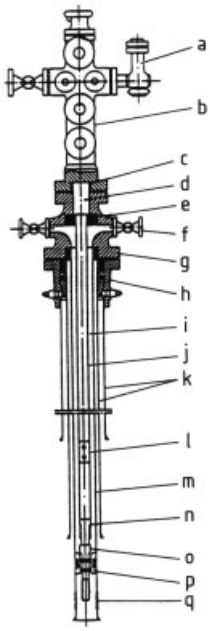

Figure 52. Production equipment of a typical gas well
a) Nozzle; b) Block-type Christmas tree with five shut-off valves; c) Double P-seal; d) Hanger nipple with seat for plug; e) Tubing seal; f) Flange slide valves; g) Tubing double flange with additional casing seal; h) Bottom flange; i) Corrosion-inhibited protective liquid; j) API tubings; C 75 or J 55; k) Intermediate casing; l) Circulation slide valve with seat for plug; m) End casing; n) Tubing receptacle, overshot type, with seat for wireline plug; o) Tubing anchor; p) Production packer; q) Perforation

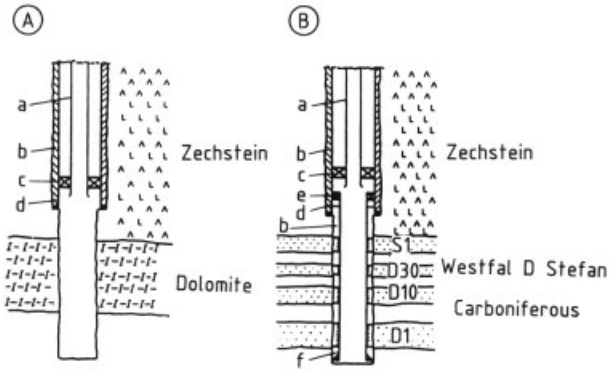

Figure 53. Bottom-hole equipment in the reservoir zone
A) Main dolomite, uncased; B) Carboniferous, casing with liner; a) Tubing, 2 7/8 inch (2.22 cm); b) Cement; c) Production packer; d) Casing, 7 inch (17.8 cm); e) Liner hanger; f) Liner, 5 inch (12.7 cm)

3.4.2. General Production Engineering

General production engineering encompasses completion, setting up production, perforation, well and reservoir treatment, fracturing technology, workover, the inflow – performance relationship, one- and two-phase flow through tubing and lines, safety engineering, environmental protection, and the plugging and abandonment of wells.

3.4.2.1. Completion

If, after borehole measurements and tests, a well is considered worth producing from, it is completed, i.e., it is equipped with all technical installations necessary for production (Fig. 52). Components necessary for the completion are the wellhead (including flanging), casing, cementing, tubing string or strings (single or multiple completion), and the "bottom hole completion". This includes the special equipment used in the reservoir zone, i.e., liners, liner hangers, sand filters, packers, and all the "wireline equipment" which can be operated by cable and wire.

Bottom Hole Completion. Essentially two types of well completion exists in the zone of the producing formation (pay zone), open-hole completion or casing and later perforation (perforated casing) (Fig. 53).

Open-hole completion, in which the producing formation is not cased, is only rarely used. Especially with fairly thick, poorly consolidated formations and at high production rates, difficulties arise with the penetration of sand from the formation into the well.

For this reason, the *casing is set on top of producing formation or below* it and cemented. To allow the production fluid to enter the well, the casing and the cement sheath are then perforated in the reservoir rock zone.

3.4.2.2. Perforation [154]

By perforating the casing and cement sheath, the well is hydraulically connected with the reservoir, so that the production fluid can enter the well with the least possible pressure drop. Currently, shaped charges are exclusively used for perforation. They are introduced into the borehole on a self-supporting cable and ignited. The number of perforations, their arrangement, penetration, and size depend on the rock hardness and borehole conditions.

With the hollow-charge perforators mainly in use, the penetration depth in the rock is 10 to 40 cm, depending on the compressive strength of the rock (Fig. 54).

Normal shaped charge perforators can be used up to ca. 170 °C. For higher temperature, special explosives and techniques have been developed. These explosives are so designed that they do not detonate by self-ignition when overheated but become unusable as a result of decomposition to gaseous phases. Explosives

stable to high temperature such as hexanitrostilbene (HNS), can be used at up to 250 °C and depths down to 8000 m for up to 24 h downhole time. In the high-temperature range which is sensitive to misfiring, safety measures prevent the effects of stray and leakage currents on the charges.

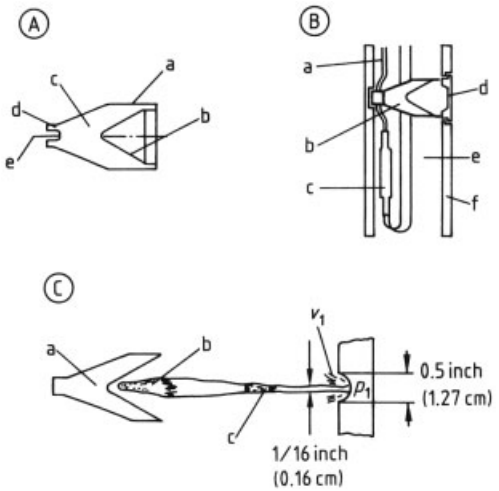

Figure 54. Schematic of a shaped-charge perforation
A); a) Case or container; b) Liner; c) Main explosive charge; d) Booster; e) Detonating cord groove, point of initiation; B); a) Detonating cord; b) Shaped charge; c) Electric detonator; d) Port plug; e) Air; f) Steel carrier; C); a) Shaped charge; b) Slug; c) Jet
Pressure on target $p_1 = 35$ kPA (5×10^6 psi), velocity of foreward jet $v_1 = 6100$ m/s

The method of *tubing-coupled perforating (TCP)* (Fig 55), which has been available since ca. 1980, has captured a significant market, because of the fairly large number of deep gas wells and the rapid development of horizontal drilling. By this method the charge carriers are installed on the tubing string below the packer and landed with the latter. They are then connected (wet) in a separate run with an ignition mechanism on the cable, and then ignited. The advantages of this method are

1) A greater shot density – 39 shots per meter with the TCP instead of 13 shots per meter conventionally
2) Considerably longer perforation sections, since the load limit of the cable used in conventional perforation is no longer limiting
3) No hanging of the charge carriers in the inclined hole (> 70°)
4) Greater safety, because the charges are sensitized underground

3.4.2.3. Well and Reservoir Treatments [155–158]

The treatment of the well and of the reservoir formation is intended to increase the productivity of wells. The decrease of this productivity can stem from the drilling process or arise in the course of production. *Well treatment* is intended to restore the original state or, in other words, to restore an inflow to the well free from restrictions. *Reservoir bed treatment* is aimed at a sustained increase of the flow through the reservoir rock.

Well Treatment. The type of obstruction of the well must be known in order to choose the most suitable treatment from the various possibilities. The obstruction can be caused

1) By solids from the mud (filter cake)
2) By increased water saturation around the well (water block)
3) By swelling of the clay minerals present in the rock's matrix
4) By chemical precipitation
5) By emulsification

In practice the first two causes occur most frequently. The obstructions are usually removed by chemical treatments. The most important chemical treatment agents are: aqueous solutions of hydrochloric acid (5 – 15 wt %), hydrofluoric acid (1 – 6 wt %), acetic acid, citric acid, and surfactants (all in concentrations of 0.1 – 5 wt %).

Corrosion inhibitors are added to the acids. *Hydrochloric acid* is generally used with carbonate-containing reservoir rock. Whose carbonate content should be ca. 20 %. The hydrochloric acid partly dissolves the rock. *Hydrofluoric acid*, usually in conjunction with hydrochloric acid, is used for dissolving silicates: here, the hydrofluoric acid is added to the hydrochloric acid as NH_4HF_2.

Surfactants are added in almost all well treatments in order to promote the dispersion of the solids, to improve the wettability of the rock, and to prevent emulsification.

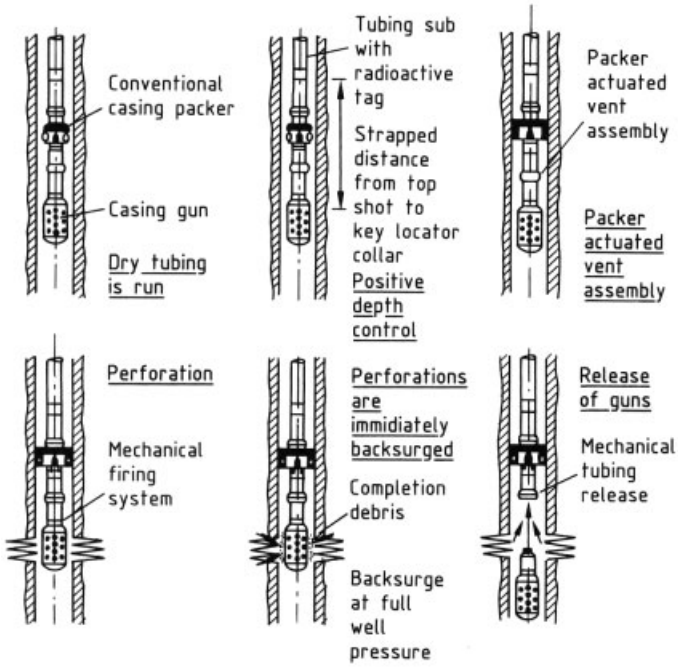

Figure 55. Tubing-conveyed perforation (TCP)

If it is intended to treat only the *borehole wall*, it is sufficient to circulate the treatment liquid into the well zone in question. If the obstruction *extends into the reservoir bed*, e.g., by clay swelling, the treatment liquid must be uniformly injected into the reservoir, avoiding pressure which leads to cracking. In any case, the reaction products should be removed from the reservoir bed or from the well as quickly as possible.

Reservoir Bed Treatments. Two methods basically exist for extensive improvement of the inflow conditions in a reservoir bed of low original permeability:

1) Pressure acidizing (usually in carbonate rocks)
2) Hydraulic fracturing (frac treatment)

The best-known method of reservoir-bed treatments is fracturing technology (Fig. 56), which is mainly used for producing artifical flow paths in the reservoir rock.

Pressure acidizing is also a hydraulic fracturing method by injection of acid. In pressure acidizing, however an increased permeability is achieved, by dissolution of a part of the carbonate matrix along the fracture faces. In *frac treatment* oil, water, or acid are injected together with gelling or viscosity-enhancing agents. In addition, to hold the fractures open and thereby ensure a high-permeability flow path, a proppant (e.g., fluvial, natural quartz sand) is added while pumping in the liquids. This proppant is deposited in the fracture and so produces the high-permeability artificial flow path. The reservoir treatments are very highly developed and so make it possible to solve very specific problems. The fracturing process will be treated briefly in the following:

Fracturing. In fracturing, only a part of the injected liquid is available to widen the fracture and act as a medium for transporting the proppant, or for the chemical reaction along the fracture faces. With a widening fracture, the other part of the liquid increasingly migrates into the pore space of the rock. In order to limit this effect, which is unfavorable for the success of the operation, *fluid loss agents* are added that raise the viscosity and form a filter cake. These agents cause the initially high fluid loss to decrease substantially after the filter cake has built up. Later,

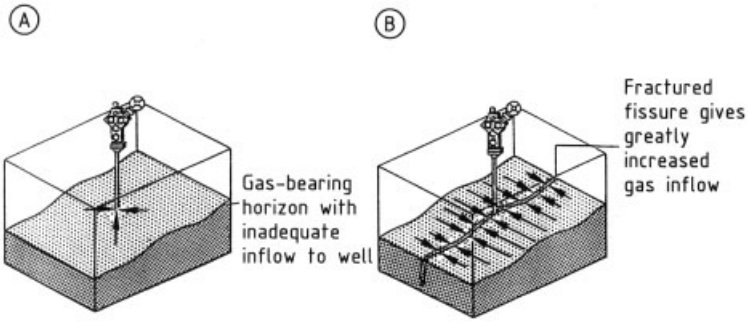

Figure 56. Schematic of fracturing technology
A) Before fracturing; B) After fracturing

these additives become ineffective due to the influence of temperature and pressure, and the production stream reaches the well unhindered.

In order to ensure the success of reservoir bed treatments, a sufficient depth of penetration into the reservoir rock must be achieved. This penetration depth must be greater with less permeable reservoir rock. Penetration depths of 40 – 70 m are usual with vertical or horizontal fractures.

The development of deep (ca. 5000 m), low-permeability gas reservoirs in the Rotliegendes and Carboniferous has led to remarkable developments of this technology. New developments now include a temperature-stable and sufficiently viscous supporting fluid, the availability of pressure-resistant proppants, and powerful equipment. For example, a treatment volume of 1100 m^3 requires a fracture length of 300 m, maximum pressure of 750 bar at a depth of 4200 m, and 15 pumping units with an output of 9000 hp.

3.4.2.4. Workover

An important chapter of constant concern for the production engineer is workover technology. More than 10^6 production wells worldwide, of which three quarters are oil wells, require continual upkeep and maintenance and – repeatedly during the course of their use – repair. Thus the number of workover hoists is three to four times the number of drilling rigs. In addition, there are the modern systems, i.e., the wireline technique (lowering activation of tolds by wireline), the coiled tubing technique, and hydraulic lifting and lowering of the strings with the snubbing unit. Workover should always follow a sequence of diagnosis, therapy (work program), operational execution, and monitoring of success. Reliable measurement techniques and supervision of operation and production are the prerequisites for correct diagnosis. The *workover hoist,* equipped with rotary table or power swivel and pump unit, can carry out the functions of lifting, lowering, rotation, and circulating fluid, likewise the *snubbing unit*. With the *wireline technique*, only running in and out is possible. The *coiled tubing* (Fig. 57) enables lifting, lowering, and circulating. Rotation is possible by using suitable downhole motors. The coiled tubing technology has a high technical standard and a growing market, particularly also with regard to horizontal drilling. Coiled strings with diameters of 1 inch (2.54 cm), $1^1/^4$ inch (3.18 cm), $1\ ^3/^4$ inch (4.44 cm) and 2 inch (5.08 cm) are available, and can be used at a depth of up to 5500 m and in combination depending on the loading situation.

3.4.2.5. Horizontal Wells [132], [135]

Horizontal drilling, which has been developed to a state of maturity since the beginning of the 1980s, makes great demands on production engineering (see page 60). Drilling engineering is able to produce, horizontal drilled sections of a few hundred to more than 1000 m in any rock. In the deviation to the horizontal, the inclination build-up can be abrupt or gradual (Fig. 46). Whether this value is 10°/10 m (final radius 60 m) or 1°/10 m (final radius 600 m) is very important for the subsequent completion.

The essential reason for the rapid growth of horizontal drilling is that a markedly larger

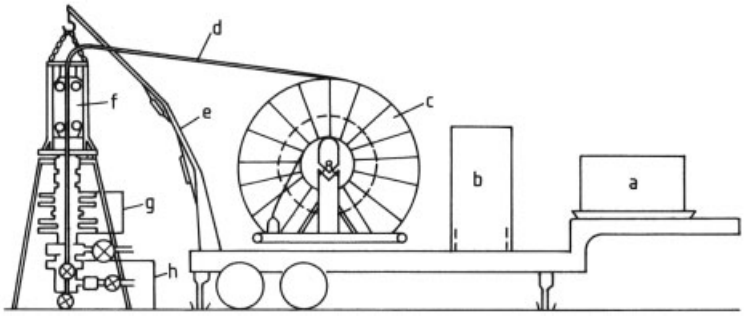

Figure 57. Schematic of a coiled-tube unit
a) Drive unit; b) Operator's cabin; c) Coiled-tube drum; d) Coiled tubing; e) Injector crane; f) Injector head; g) Preventer device; h) Wellhead

zone of the reservoir is developed for production than with vertical wells, so that a considerable (15 – 20-fold) production increase is achieved at lower differential pressure; better sweep efficiency; and a higher recovery. In heterogeneous, fissured rock the effect is even more significant. The potential, especially if specific enhanced oil recovery (EOR) processes are incorporated in this technology, is far from exhausted. In gas reservoirs, turbulence around the borehole is avoided, and in underground storage, a capacity increase of 3- to 8-fold seems possible with horizontal drilling. There is still, however, a considerable need to develop devices and methods specific to the process, cf. Schlumberger coiled-tubing-logging (CTL). Completion in deviated and horizontal wells can be carried out as follows:

1) Open hole (i.e., uncased): only possible with stable formations
2) Open hole with slotted or "prepacked" liner (i.e., plastic-reinforced sand filter)
3) Slotted liner with external casing packers (ECP) in the open hole
4) Cemented and perforated liner

With the first two methods, no isolation of unwanted sections is possible. However, the use of correctly placed ECP's (method 3) and successful cementing (method 4) allows selective production. For liner installation, the clearances and the borehole geometry in the deviation zone and the permissible bending stress must be taken into account. Cementing in the horizontal section is impeded by the fact that the gravitational force favors segregation of liquid and gas at the top and sedimentation of cuttings and solid at the bottom. For the perforation, only the TCP technique (see Section 3.4.2.2) is possible from a 70° inclination; and, as an alternative to the tubing string, coiled tubing or the drill string. The wireline technique in horizontal holes is virtually ruled out. For workover, the coiled tubing technique mainly is applied (Fig. 54). Methods independent of the borehole inclination are suitable as production facilities such as the submersible centrifugal pump, the hydraulic piston pump, and gas lift. The lengths of rigid units must be geared to the maximum possible flexure in the deviation zone, and the necessary adaptation of systems must be carried out.

3.4.3. Oil Production Engineering [159–163]

In oil production engineering, not only the technical instruments and developments determine success. Efficiency and overall profitability crucially depend on qualified supervision, damage evaluation, the logistics of material and distribution, and a sound operating organization. The use of *tertiary methods* (e.g., steam injection and steam flooding) signify additional tasks for production engineering: In *steam injection wells*, and also in the hot production wells of the *steam flooding projects,* extreme thermal stresses (140 bar, 340 °C) and corrosion problems in the underground equipment must be kept under control.

Production engineering has powerful production processes and production facilities available to it. Herewith, the particular conditions of each reservoir can be taken into account. This in turn

enables, firstly, the reservoir fluids to be produced as efficiently as possible with extensive use of the reservoir energy; and secondly, production difficulties, caused for example by accumulated sand or water or by corrosion, to be obviated in good time.

With the opening up and development of a reservoir, its most important qualities are determined: reservoir pressure, permeability of the rocks, viscosity of the production fluid, and the presence and characteristics of a natural water drive. These are the foundations for planning production rates and methods, and equipment.

In principle, the following production processes are available:

1) Flowing production
2) Gas lift
3) Centrifugal pumps
4) Piston pumps, sucker rod, or hydraulic

With these methods, ca. 3.25×10^9 m^3 crude oil per year are produced from depths of up to 6000 m.

Flowing Production [156], [164]. As a result of production, a pressure drop occurs which expands, funnel-shaped, around the hole. Under the influence of this pressure gradient, the reservoir liquid flows to the borehole, enters the casing through the perforation holes, rises up in the tubing string to the surface, and flows at the surface by lines to a collection manifold at a central plant. Here the output from the field is combined, dewatered, treated, and transported to the refinery. The flow pressure at the depth of the reservoir bed is sufficient to overcome the hydrostatic pressure of the production column and the frictional loss in the tubing string and in the surface lines.

The *flowing production rate* is calculated as a function of bottom flow pressure, tubing diameter, and pressure head according to the law of conservation of energy, while taking into account the empirical parameters which determine the two-phase flow.

The relationships become more complicated when the pressure in the tubing string falls below the degassing pressure (bubble point) and free gas evolves (see Section 2.2). Progressive pressure reduction gives rise to larger gas bubbles (see Fig. 58) and the liquid column breaks up (slug flow, Fig. 58 C), or the rising gas bubbles entrain oil drops which, owing to friction at the tube walls, form an upward-moving film (annular flow, Fig. 58 E) which finally changes into mist flow, i.e., the liquid film disappears and the liquid separates into droplets which are mixed with the gas stream (Fig. 58 F).

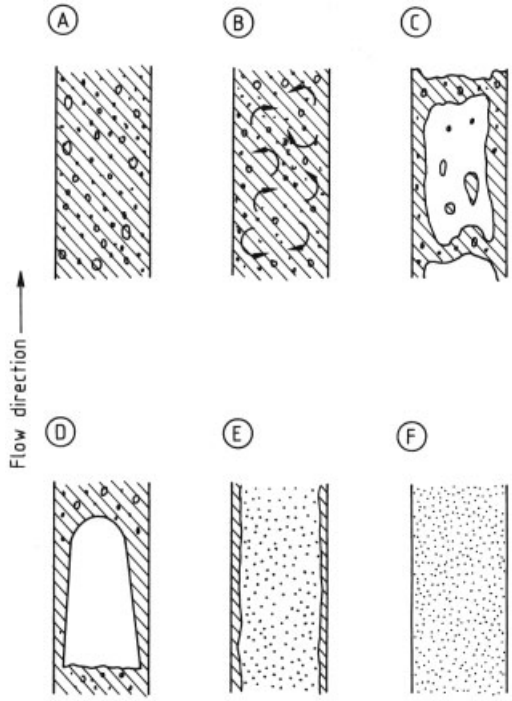

Figure 58. Flow patterns in two-phase flow
A) Bubble flow; B) Bubble flow (turbulent); C) Slug flow; d) Froth Flow; E) Annular flow; F) Spray flow

The duration of this economically advantageous eruptive production phase can be extended by a good matching of the production rate to the tube cross section.

Gas-Lift Production. In the gas-lift method, gas is injected into the tubing string at one point or – to minimize the starting pressure – at several successive points and in this way bubble or slug flow is initiated, as has been described for eruptive production (see Fig. 59). This method is used continuously or intermittently, depending on rate of flow, production rate, and gas availability.

The eruptive production phase is often already supported with gas lift or is switched to gas lift when the pressure gradient between borehole

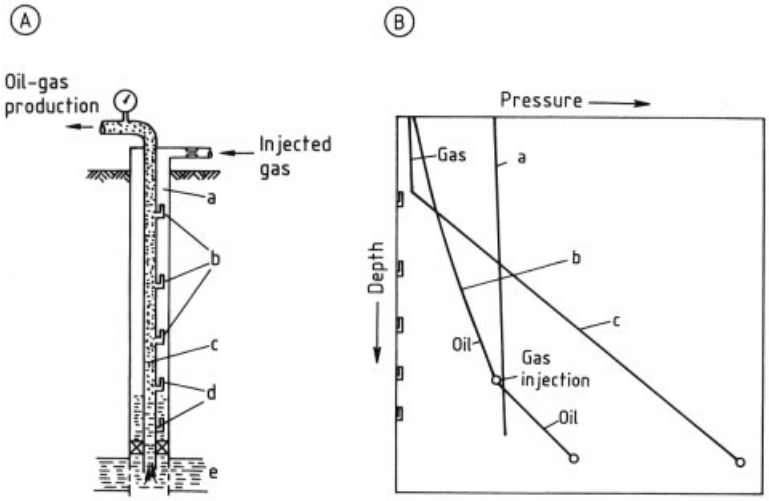

Figure 59. Continuous gas-lift operation
A) Schematic; a) Annulus; b) Valves closed; c) Tubing; d) Valves open; e) Production zone; B) Pressure distribution at different depths; a) Annulus pressure; b) Pressure curve in tubing for production well; c) Pressure curve in tubing for inoperative well

and reservoir decreases. The applicability of this flexible production method is limited when there is a shortage of gas or the gas compression costs are too high. Where sufficient gas is available, the scope for application is wide because of the flexibility of this method with regard to production rate and depth. In strongly deviated boreholes and if abrasive media are coproduced, gas lift is usually superior to other production methods (no moving parts).

Gas lift valves controlled by the injection pressure are fitted in the tubing string, into which they inject the lift gas. The pressure at the valve is controlled via a pressure bellow which operates with or without spring support (according to type). At the surface before installation the bellow is set to a certain opening and closing pressure.

Since on start-up of a well in gas-lift operation, the back-pressure of the whole liquid column requires a high injection pressure, several lift valves are installed and distributed over the tubing string, whose opening pressures are coordinated with each other.

Production with Subsurface Pumps
[165], [166]. Subsurface pumps are the most widespread form for the production of the oil, which is usually contaminated by considerable amounts of water. Types of pump used include electric submersible centrifugal pumps and mechanical or hydraulic subsurface piston pumps. The pumps are installed below the liquid level in the wells and, in difficult underground conditions, must overcome hydraulic heads of several hundred to a few thousand meters (Fig. 60).

Electric Submersible Centrifugal Pumps (Fig. 57 B). These are multistage centrifugal pumps that are submerged in the production liquid. They are installed in the borehole, directly coupled with the motor, and suspended on the tubing string. The pump unit comprises from bottom to top: electric motor (bipolar asynchronous squirrel-cage, oil-immersed, j), centrifugal pump (with numerous stages e, f), protector (sealing adaptor between motor and pump, k), gas separator (for separating free gas from the produced liquid, l), electrical switching device, and power cable (these limit the installation depth, since the electric insulation resistance is temperature dependent).

The hydraulic head (hydrostatic pressure), frictional pressure losses, and wellhead pressure determine the number of pump stages. Because of possible changes in the hydraulic head, pumps with stable characteristics are preferred.

At depths between 1000 and 4000 m, submersible centrifugal pumps cope with minimum delivery rates between 260 and 50 m^3/d

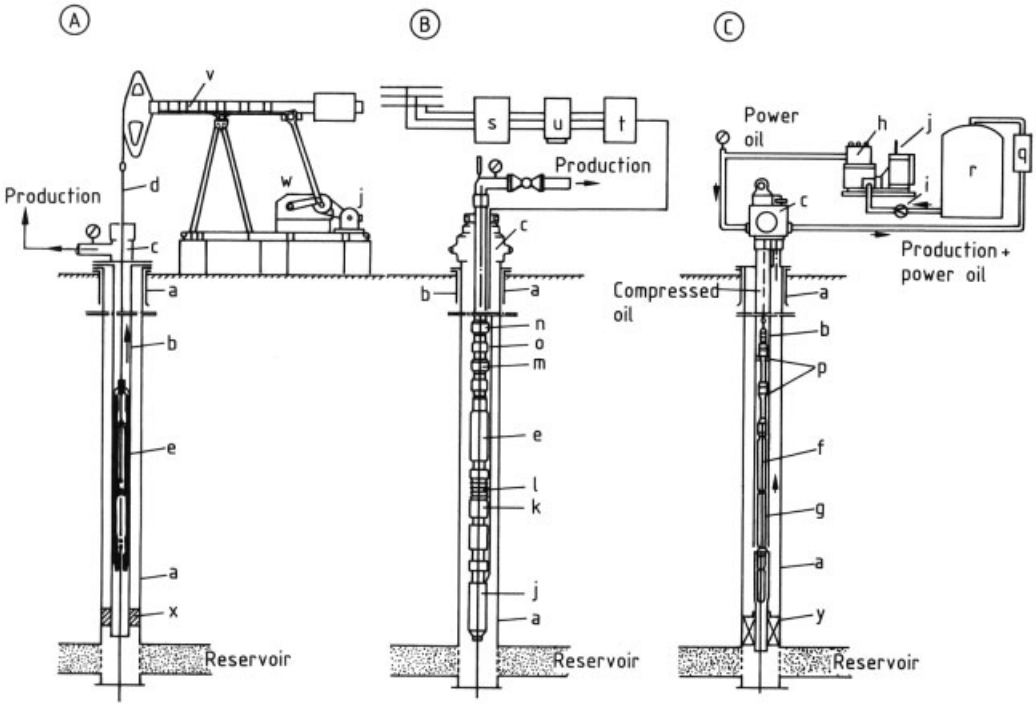

Figure 60. Different types of pump
A) Subsurface rod pump with walking beam drive; B) Electric submersible centrifugal pump; C) Hydraulic subsurface piston pump with pressure oil supply in the open system;
a) Casing; b) Tubing string; c) Wellhead; d) Pump rod; e) Subsurface pump; f) Drive pump, g) Production pump; h) Pressure oil pump with i) Pressure raising pump on the suction side; j) Motor; k) Sealing adaptor; l) Gas separator; m) Nonreturn valve; n) Tubing drain valve; o) Three-core cable; p) Circulation piece; r) Pressure oil tank; s) High voltage transformer; t) Autotransformer; u) Switch cabinet; v) Walking beam; w) Gearing; x) Tubing anchor; y) Production packer

and maximum delivery rates between 340 and 150 m^3/d.

Mechanical Rod Subsurface Pumps (Sucker Rod Pumping [167]) (Fig. 57 A). Oil production was for a long time synonymous with the use of rod pumps, which are installed as tubing pumps with the tubing or are landed at the pumping depth as inserted pumps on the pump rods in the tubing string. These pumps are driven from the surface by beam pumping units with motor, gearing, and a pumping jack, that includes the typical walking beam with "horse's head" on which the pump rod string is suspended. The weak point of the system, the "pump rod string", requires particular attention. Liquid load, rod string weight, vibrations, and mass acceleration, which occurs particularly when the direction of motion reverses, determine the efficiency and the limits of this kind of artificial lift.

The supply of types of subsurface pumps that differ according to material, valve arrangement, piston area, and stroke length also take account of extreme operating conditions. Valve function and pumping efficiency are periodically checked by dynamometer measurements, i.e., by recording the load uptake. Corrosion, sand, or friction in deviated wells reduce the service life of pumps and rod string.

Hydraulic Subsurface Piston Pumps (Fig. 57 C). Like the electric submersible centrifugal pumps, these pumps are run underground, yet by high-pressure fluid. The difficulties with the pump rod string are avoided by hydraulic transmission of the automotive energy with an auxiliary liquid from the surface to a pump installed in the borehole. A high-pressure pumping installation provides the "power oil" that is supplied to the individual producing wells and the pump via the tubing string. The pump combines

two identical sections in one block, the upper of which operates as a motor which is driven by pressure oil (f) and for its part drives the lower pump (g) which is hydraulically in complete equilibrium.

The pressure oil is either taken from the produced oil and carefully cleaned (open system) or is contained in a closed system. Usually the pressure oil is supplied down the tubing and the flow back through the annulus or, where that is not possible, it is fed through additional tubing strings. For repairs, the liquid stream is reversed and the pump pumped to the surface (free pumps in contrast to fixed pumps).

The range of application of hydraulic piston pumps exceeds that of rod pumps by far. As with gas lift, a number of wells are connected to one power oil high pressure pump station, respectively, so that an additional system of lines is required. To account for the particular conditions of each well, a specially devised flow and pressure control system is required.

With the electric submersible centrifugal pumps and rod pumps, neither additional lines above ground nor an expensive control technique are used. These basic differences are significant for economic efficiency [168].

Developments and Limitations. In addition there are *jet pumps* (mainly in the United States); *Moineau pumps* with subsurface drive or driven by pump rods from the surface; and *turbine-driven submersible centrifugal pumps.* However, these pumps are rarely used. The proven methods described above maintain their position, thanks to their high standard of technical maturity and reliability. The correct choice of pump for the pressure and inflow conditions (inflow performance) and the flow resistances in the production path (outflow performance), both determined by the reservoir, and the product composition (gas, oil) ensures a long and optimized service life. An example of this is Shell Oil's Reno field in Wyoming in the United States. There, since 1975, 190 wells at depths ranging from 2000 to 4000 m have been converted from hydraulic piston pumps to rod pumps for economic reasons alone. By using combined pump rod strings of high-strength steel, a further developed heavy-duty beam (Mark II) drive, and subsurface pumps suitable for the great depths, 105 m^3/d wet oil with 30 % water are produced from 4420 m lifting depth. The field is monitored by electronic dynamometers.

3.4.3.1. Collection and Treatment of Crude Oil [146]

Production rates, crude oil properties, and field conditions determine in what way and to what extent the production of individual wells, groups of wells, parts of the overall field or several neighboring fields are combined. Treatment at the field includes gas separation and dewatering (Fig. 61).

Gas Separation. The degassing process starts already in the reservoir and intensifies with the decrease of pressure and temperature through the tubing string and the pipe system during production. Consequently, a two-phase mixture enters the gas separators. To obtain a maximum liquid yield and, in particular, at high gas – oil ratios, a multistage separation should be aimed for. However, apparatus cost and special requirements, such as, e.g., gas delivery at high pressure to the consumer, lead to the widespread use of two- and three-stage separation.

Dewatering and Desalting. Since crude oil can still be produced at 95 % water content, water – together with sand and contaminants – is removed as early as possible to reduce transportation costs and corrosion. The refinery acceptance standard in Germany limits the total contamination (water and sludge) of the crude oil generally to 1 % (salt content below 0.02 %). The water separates only partly as "free water." Water and oil mainly occur as an emulsion, which must be broken by thermal, mechanical, chemical, or electrical methods, or by a combination of these. At the same time, the emulsifiers present in the crude oil (constituents that are concentrated at the interface) must be made ineffective. The free water droplets must coalesce and, with the aid of gravitational or centrifugal forces, separate.

Since the 1980s, a greatly simplified water separation method has evolved from the variety of process combinations used. Particular features of separation plants are:

1) Early feeding of the demulsifier in order to use the pipeline for internal mixing

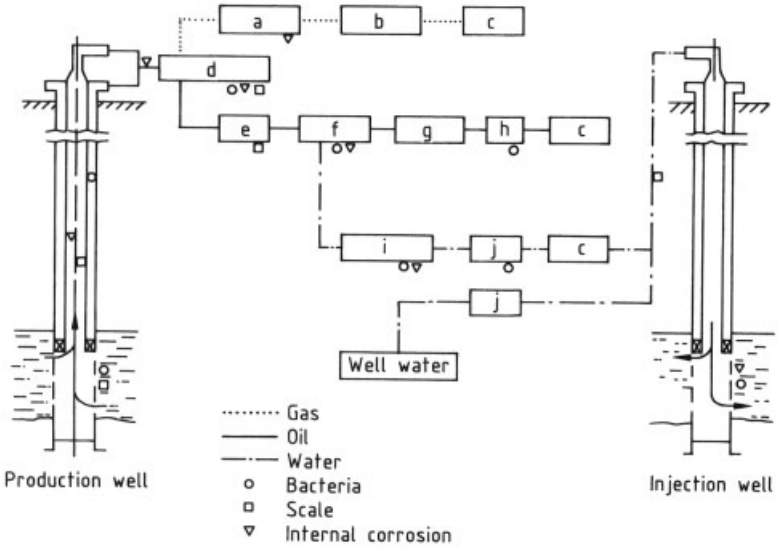

Figure 61. Flow diagram of an oilfield
a) Compressor; b) Drying; c) Line; d) Gas separator; e) Heater; f) Separator; g) Desalter; h) Tank; i) Secondary sedimentation; j) Filter

2) Moderate heating ("cold treatment") to save thermal energy and reduce the vaporization losses
3) Simple apparatus; the water separation can often be carried out in a single tank, whereas previously a downstream fine purification in the gas-heated treater or an electrical fine purification has been required

These improvements have been made possible by the further development of demulsifying chemicals and by continual constructional improvements. These include also the improvement of the gas separation, by which the inconvenient degassing after processing has been eliminated.

3.4.3.2. Special Problems in Crude Oil Production

Paraffin Precipitation. When the crude oil is cooled below the pour point, high-boiling hydrocarbons can precipitate and lead to clogging of the tubing string or the output lines. Countermeasures are internal plastic lining of pipes, hot oil treatment, additions of chemical inhibitors to the oil, and periodic cleaning of tubing and lines with scrapers or go-devils. Incompletely removed paraffin precipitates are often the reason why reservoir bed treatments and measures to improve the inflow are unsuccessful.

Chemical Precipitates. Impurities such as carbonates, sulfates, and sulfur can impair production, treatment, and water injection by obstructions in the reservoir, sealing of the perforation, and narrowing of lines to the point of shutdown.

Sand. Fine sand from the reservoir bed can considerably hinder operation of the pumps and cause damage by blocking and increased wear. Countermeasures are:

1) Borehole measurements for the early detection of danger and the subsequent establishment of a safe production rate which must not be exceeded.
2) Retention of the sand by filters consisting of slotted tubes which can be wound with one or more layers of wire or jacketed with a cylinder of plastics-reinforced sand. When the product liquid flows into the filter, the entrained sand is filtered off, and then it deposits and/or fills the annulus between filter and borehole wall. This effect is also achieved by artificially jetted and injected sand packings (gravel pack, frac pack). The

slot and mesh widths of the filters and the particle size composition of the gravel packings must be carefully matched to the particle size of the entrained sand and the flow rates.

3) Artificial sand consolidation by two-component resins that are injected into the formation. Shrinkage during the consolidation or flushing (flooding after gelling) ensures the necessary permeability. Failures can force the abandonment of a well.

Corrosion. Well production and oilfield installations are exposed to corrosion both externally and internally. External corrosion results from aggressive soils or parasitic electrical currents and internal corrosion from salt water, carbon dioxide, or hydrogen sulfide contained in the produced fluids. Parts of a plant particularly at risk are those in which attack by corrosive media takes place in the presence of mechanical stresses, for example, rod strings and subsurface pumps in high water cut wells. Salt water dramatically lowers the fatigue strength of steels under dynamic conditions. Lines and tanks are protected cathodically and/or by addition of suitable inhibitors to the liquid.

3.4.4. Natural Gas Production Engineering [169–172]

Characteristic features of natural gas production engineering are reservoir depths of more than 5000 m (Fig. 62), the completion problems connected with this, and the technically safe control of the accompanying gases hydrogen sulfide, carbon dioxide, and nitrogen. To avoid stress corrosion cracking induced by hydrogen sulfide, *sour gas wells* are provided with completions that ensure the circulation of solvents and/or corrosion inhibitors. Selection of appropiate materials plays a crucial role for casing and completion. Large treatment plants remove environmentally damaging hydrogen sulfide by gas scrubbing (Grossenkneten) and the nitrogen by liquefaction (Clenze) (see also, → Natural Gas, Chap. 2.3.1., → Natural Gas, Chap. 2.6.).

High pressure and high pressure differentials, extreme temperature differences, and aggressive gas constituents require special attention in the production equipment of gas wells. The gas is kept away from the actual borehole casing by the production equipment (Fig. 49) which, in the simplest case, consists of tubing and packer. The necessity of complete and durable sealing makes extremely high demands on the material and of machining precision. For gas wells, special gas-tight tubing joints have been developed. In the selection of tubing and its joints, the large stress resulting from high pressure and temperature changes, causing large load reversals, must be taken into account.

The evolvement of reservoir pressure with time, its depth, the inflow area of the well, the expected production period, and the possible liquid yield during production determine the selection of tubing size and consequently the production rate.

Packers seal off one or several producing zones from each other and from the annulus. The sealing materials used must be temperature-stable, permanently elastic, and gas- and H_2S-resistant. For the further security of the well, the annulus is usually filled with a protective liquid (inhibited water).

Essential elements of the *tubing string* are devices for well-engineering operations, for the control and monitoring of production, and for remedial measures, such as the injection of inhibitors or agents against hydrate formation. These devices are installed, pulled out, and actuated on the wire line. Multizone completions, borehole repairs, selective test and treatment operations, the killing of wells, and gas-lift production require the selective, manipulable opening of the tubing string. For this purpose sliding side door couplings are installed, which are also actuated with tools on the wire line.

Production is *controlled and monitored* by control valves and measuring sections. It is generally controlled as a function of rate or pressure. The rates are determined by orifice meter measurements. (Fig. 63).

Safety is particularly important. At the *surface*, for example, the valves are normally kept open via a low pressure gas system. If the pressure drops in the sensor system, these valves close the production line. With *underground safety valves*, a safety valve kept open via a hydraulic control line closes the tubing string as soon as the pressure in the control line falls. Other arrangements prevent the freezing-up of control devices and lines at the surface by hy-

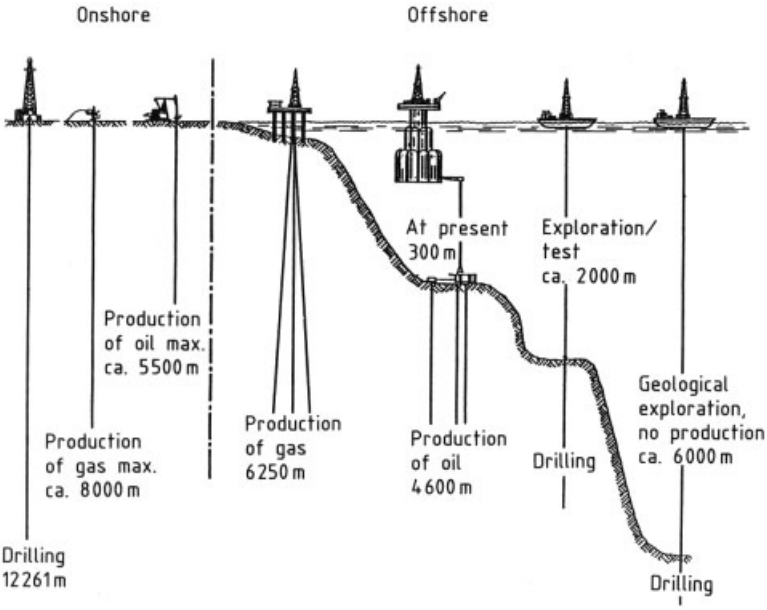

Figure 62. Present development of production depths

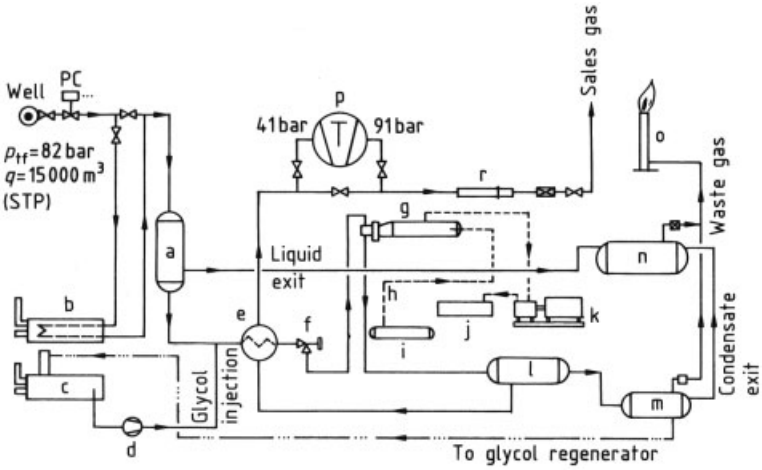

Figure 63. Flow diagram of a natural gas drying plant with external cooling and pressure increase
a) Prefractionator; b) Continuous heater; c) Glycol regeneration; d) Glycol pump; e) Heat exchanger; f) Motor needle valve; g) Chiller (coolant NH_3); h) Refrigeration plant; i) Collector; j) Condenser; k) Piston compressor; l) Low-temperature separator ($-10\,°C$); m) Glycol – oil separator; n) Condensate tank; o) Flare; p) Compressor plant; r) Measuring section

drate formation, and protect the surface installations from excessively high pressure.

Numerous exhausted natural gas fields are converted into natural gas *deep storage reservoirs*, where demand and suitability allow. (→ Natural Gas). Production engineering questions are also very important for storage reservoirs, especially since withdrawal rates at up to ten times the production rate of a gas deposit are aimed at.

3.4.5. Offshore Recovery [168], [173]

Liquid and gaseous hydrocarbons in reservoirs under the sea bottom are recovered and produced by equipment and devices which correspond in

principle to those of onshore installations. A distinction is made between supporting structures designed to incorporate *drilling rigs* used only temporarily, and those to incorporate *production and treatment facilities* which must be installed in the sea over the whole period of exploitation of the deposit, frequently more than 20 years.

Because of the difficult environmental conditions and of the danger to the staff from the processing of highly inflammable hydrocarbons in a very small space (especially gases), the safety requirements for offshore units are particularly high.

The positioning of the *supporting structures* over the defined locations in the sea no longer presents difficulties in principle, since a very wide range of navigation systems, such as satellite, Doppler, and radio navigation can be used.

Offshore wells require, for the distance between sea bottom and platform, a special string of casing known as a *riser*, which must resist the weather conditions, the motion of the sea, and the current. With exploration in ever deeper water, *platforms* are being built and operated in ocean depths down to 600 m (Fig. 64) and development and planning is already geared to ocean depths ca. 1200 m. The riser technology has also been pushed ahead; the safety engineering aspects of continuous well monitoring under extreme weather conditions being a priority. For the Shell Offshore East Coast Project a riser of 4000 feet (2121 meters) is planned in connection with a dynamically positioned drilling vessel. In deeper waters, in areas with shipping lanes, and for smaller fields, *subsea completion* is sometimes used, in which wellhead and manifold are directly on the sea bottom.

Supporting Structures for Offshore Drilling Rigs (Fig. 65).

Barges were the first constructional systems used for installing drilling equipment and are still in use. These consist of flat hulls which are sunk (flooded) in several storeys above each other in shallow water regions. Such installations must not be exposed to currents and waves. The barge platforms are preferably used in coastal swamp and mud flat regions, inland seas, and similar locations. These installations can be used to water depths of up to ca. 10 m. They are dismantled by refloating the barges.

Fixed Platforms. These are working platforms erected over the surface of the sea on piles rammed into the bottom. Sometimes it can be appropriate that only the derrick is erected on such a platform, while the remaining equipment, materials, energy plants, crew accommodation, etc., are accommodated on a ship anchored nearby. The field of use of these platform – ship installations extends to ca. 30 m depth of water. However, this combined construction method can be used only in regions without difficult currents and weather conditions.

Jack-Up Platforms (Fig. 62 B). Jack-up platforms are flat, multi-storey barges that have three, four, six or more legs. As soon as these hulls have been towed to their place of use, the legs are generally lowered vertically into the sea bottom. The platform is raised on the legs so far above the sea floor (which must be stable) that storm waves of maximum height 32 m (measured in centennial storms) cannot beat under the hull of the barge.

These platforms are equipped for a payload of a few 1000 t of material, depending on their operating location. The derrick and all necessary auxiliary equipment, such as a generator (electric drive for the rotary table, hoist equipment, and mud pumps), cementing, and mud installations, measurement services, etc., are permanently installed. Crew rooms and a helicopter landing deck are necessary support facilities.

Material is transferred from the supply ships by permanently installed rotary cranes. The field of application of such rigs extends to about 100 m depth of water. Because of good stability, the dependence of operations on the weather is very small.

Semisubmersibles (Fig. 62 A) are floating offshore platforms that can be self-propelled. The work platform rests on 3 – 8 slender columns, which themselves rest on two long, parallel floats. The latter are usually submerged to 6 – 9 m during transportation to the location, to give more favorable sailing conditions. At the location, sea water ballast is taken on and the draft increased to about 20 – 25 m to lower the center of gravity. This results in small movements with the heaving swell which together with the reduced water plane area prevent large amplitudes of motion when in the drilling condition.

In addition, the semisubmersible is anchored to its location. Despite the use of these measures

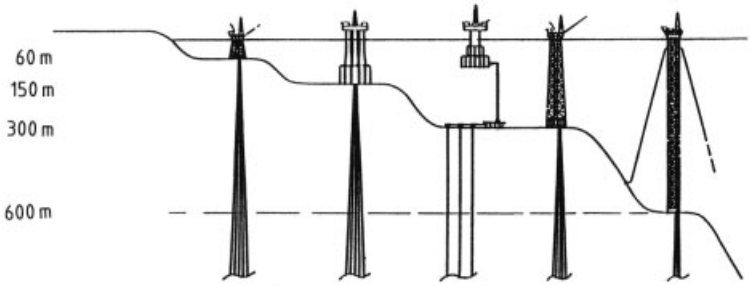

Figure 64. Supporting structures for offshore production plants

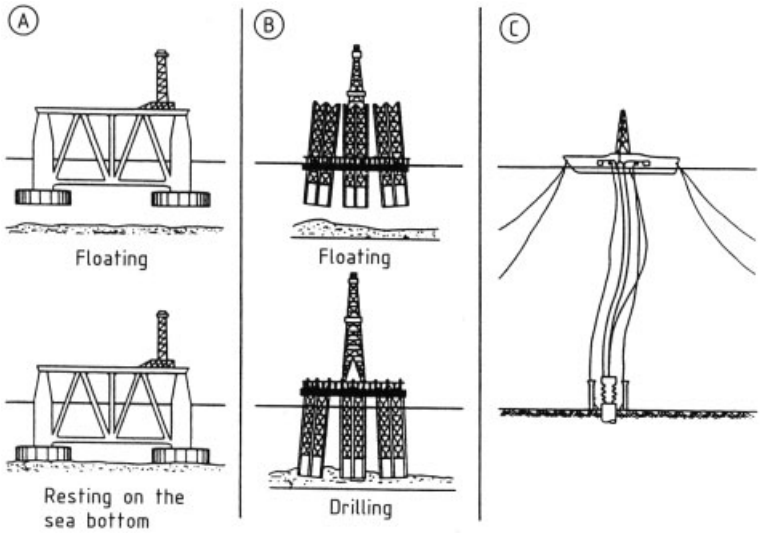

Figure 65. Mobile offshore drilling rigs
A) Submersible, water depth to 300 m; B) Jack-up platform, water depth to 100 m; C) Drill ship, water depth to 500 m

there is an increased effect of waves and storms on these rigs compared to other offshore units so that servicing is more restricted than of the units previously described. The main fields of use are water depths between 50 and 600 m. The variable deck load reaches ca. 5000 t, depending on type. The engineering installations required are the same as those of the jack-up platform.

Drillships (Fig. 62 C). Drillships were used very early in offshore engineering. Conventional ships were converted by erecting the derrick amidships and also accommodating the other engineering installations on the ship. Under the derrick a square recess (moonpool) was provided to make drilling possible. Gradually special drillships with different shapes have been built for very diverse applications. However, drillships are used only in small numbers since, owing to their large motions with the heaving swell and despite dynamic positioning over the borehole, they can be used only to a limited degree on the open seas and oceans.

Demand also plays a part. Because of their greater hydrocarbon expectation, it is mainly the shelf edges which have been explored up to now, and the deep sea hardly at all, so that drillships have largely been dispensed with. But they offer considerable advantages in regions with fairly stable weather, because of their high transit speed and fairly large carrying capacity (and therefore small dependence on external supply).

The deep sea drillships "Joides Resolution," and its predecessor the "Glomar Challenger" have undertaken studies of the deep sea bottom and its underground as part of a broad program of geoscientific research.

Drilling rigs on fixed production platforms are used in fairly deep waters where a field has already been explored by other drilling rigs and a larger number of production wells must now be drilled. These platforms are already prepared for later production. More than 40 holes can be drilled fanwise, directed to defined target points. The production installations for the extraction of the hydrocarbons are installed later as required with the aid of rigging-up ships.

Supporting Structures for Offshore Production Plants (Fig. 61). The production platforms used for offshore oil and gas production are of the same type as those for the drilling rigs of fixed platforms and semisubmersibles. The first platforms were built as wooden structures, but since 1940 they have been constructed only of steel or concrete. All submersed steel structures require corrosion protection in order to ensure a sufficient service life. For this, direct protection paint and/or galvanization is used and in addition cathodic protection devices feeding external current or sacrificial anodes are applied.

In the Norwegian and British sectors of the North Sea, several large concrete structures have been set down on the sea bottom. These consist of three hollow conical bodies of concrete, constructed on the sliding molding principle, that contain, as an upper annex, multistorey production installations and personnel accommodation; the derrick also is arranged there.

From these large structures, a fairly large number of deflected holes are sunk. Two legs contain the production casing of the wells. The third leg receives essentially the supply and line system. The base of the structure is provided with a large number of concrete tanks in which the crude oil, salt water, and other products are stored. The first stages of the structures were built in Norwegian fjords, on a construction deck raised on sheet piles. Then, when the structure had reached a certain stage, the installations were floated off and towed into deeper water, in order to allow immersion during the further weight increase resulting from concreting to the final height. The maximum total height of these installations in the Brent and Statfjord fields is 250 m at a water depth down to 175 m. A small wave pressure and short installation times at the location (no ramming work) were reasons for constructing such structures, which, however, can be set down only on a suitably firm sea bottom. The largest offshore platform in the world is 271 m high and has been installed in the Norwegian Statfjord field.

Offshore Well Equipment. An important technical problem in drilling at sea is the effect of the heaving swell and of the current on the sea bottom. Movements and vibrations of the drilling rig in relation to the pipe strings leading to the sea bed (riser and drill string) must be avoided. However, these undesirable effects cannot completely be removed especially on floating drilling rigs (semisubmersible and drillship). Devices are used to combat these difficulties. They are installed in the hoist installations and in the casing string (riser). For wells in *shallow water areas* there are no longer any special problems. Here completion arrangements can be employed which are also in use on land, with casing held in a "well protector" and led out above the surface of the sea.

However, in *fairly deep water* and in the area of shipping lanes, underwater completion must be used. This was first applied in 1943 in the 11.5 m deep water of Lake Erie. Meanwhile, such systems, usually remote-controlled, have been installed in a great variety of sea areas of the world. Installations that cannot be reached by divers are not yet in use. Two types of completion exist: *wet completion,* in which all the components are in contact with the sea, and *dry completion,* in which the components are in a pressure chamber under atmospheric conditions, so that the personnel can carry out assembly and maintenance of the installation by "submarine passenger ferry." The criteria for the selection of suitable completion include water depth, weather conditions, areas with shipping lanes, icebergs, mechanical properties of the seabottom and the possibility of sediments wandering at sea bottom.

The use of divers, drilling rigs, or workover ships is required for necessary repairs, possible replacement of the Christmas tree and the tubing, and workover operations.

Logistics. As soon as the production of oil and gas in the offshore sector is taken up, provision must be made for the transportation of these products. Underwater pipelines and ships are primarily used. An essential prerequisite for the

smooth and secure operation of offshore drilling and production activities is the existence of a land base. The land base provides the maintenance service, all supplies of materials, as well as the catering, intelligence, and rescue services, and usually the exchange of personnel by helicopter, etc.

4. References

General References

1. A. W. Bally: "Seismic Expression of Structural Styles," *AAPG Stud. Geol.* **15** (1983).
2. T. Bandlova, G. Katzung: "Zum Auftreten von Erdgas in Verbreitungsgebieten kohleführender Ablagerungen," *Z. Angew. Geol.* **21** (1975) 467 – 572.
3. H. Bartenstein, R. Teichmüller: "Inkohlungsuntersuchungen, ein Schlüssel zur Prospektierung von paläozoischen Kohlenwasserstoff-Lagerstätten?" *Fortschr. Geol. Rheinl. Westfalen* **24** (1974).
4. R. Beck, P. Lehner: "Oceans, Now Frontier in Exploration," *Am. Assoc. Pet. Geol. Bull.* **58** (1974) 376 – 395.
5. H. Beckmann: *Geological Prospecting of Petroleum,* Enke Verlag, Stuttgart 1976, pp. 176.
6. F. Bender: "Geologie der Kohlenwasserstoffe," *Angew. Geowissenschaften,* vol. 3, Enke Verlag, Stuttgart 1984.
7. P. Bitterli: "On the Origin of Petroleum," *Bull. Ver. Schweiz. Pet. Geol. Ing.* **30** (1964) no. 79, 74 – 77.
8. H. Boigk: *Erdöl und Erdölgas in der Bundesrepublik Deutschland,* Enke Verlag, Stuttgart 1981, pp. 330.
9. H. Boigk, H. W. Hagemann, W. Stahl, G. Wollanke: "Zur Herkunft und Migration des Stickstoffs nordwestdeutscher Erdgase im Oberkarbon und Rotliegend," *Erdöl Kohle Erdgas Petrochem.* **75/76** (1986) 112 – 139.
10. H. Boigk, H. U. Hark, H. J. Meyer, H. Wehner: "Beziehungen zwischen Geochemie, Migration und Lagerstättengenese im Niedersächsischen Becken (westlich der Ems)," *Erdöl Kohle Erdgas Petrochem.* **74/75** (1986) no. 1, 98 – 114.
11. C. Bois, P. Bouche, R. Pelet: "Global Geologic History and Distribution of Hydrocarbon Reserves," *Am. Assoc. Pet. Geol. Bull.* **66** (1982) no. 9, 1248 – 1270.
12. L. F. Brown, Jr., W. L. Fisher: *Seismic Stratigraphic Interpretation and Petroleum Exploration,* AAP Continuing Education Course Series, 1980, p. 16.
13. M. S. Burštar: *Osnovy teorii formirovanija zeležej nefti i gaza (Grundlagen der Theorie der Bildung von Erdöl- und Erdgaslagerstätten),* Nedra, Moskau 1973, pp. 255.
14. R. E. Chapman: *Petroleum Geology,* Elsevier, Amsterdam 1983.
15. G. Dohr: *Applied Geophysics in Geology of Petroleum,* **vol. 1,** Enke Verlag, Stuttgart 1981.
16. W. Dow: "Petroleum Source Beds on Continental Slopes and Rises," *Am. Assoc. Pet. Geol. Bull.* **59** (1970) no. 6, 1584 – 1606.
17. H. Grunau: "Abundance of Source Rocks for Oil and Gas Worldwide," *J. Pet. Geol.* **6** (1983) no. 1, 39 – 54.
18. H. D. Hedberg: Geologic Aspects of Origin of Petroleum, *Am. Pet. Geol. Bull.* **48** (1964) 1755 – 1803.
19. D. Heling et al.: *Sediments and Environmental Geochemistry, Sedimentological and Petrophysical Aspects of Primary Petroleum Migration Pathways,* Springer Verlag, Berlin 1990, 152 pp.
20. G. Hosson, W. Pohl: *Modern Petroleum Technology,* 4th ed., Inst. of Petroleum, London 1973.
21. J. M. Hunt: *Petroleum Geochemistry and Geology,* W. H. Freeman, San Francisco 1979, pp. 617.
22. V. C. Illings: "Geology Applied to Petroleum," *Proc. Geol. Assoc.* **53** (1942) 156 – 187.
23. A. Jegorow: "Paläogeographische Beziehungen zwischen kohle-, erdöl- und erdgasführenden Gebieten des Karbons," *C. R. Sept. Congrès Internat. de Stratigraphie et de Geologie sur Carbonifère,* **vol. 3**, Krefeld 1974, pp. 249 – 259.
24. H. Jüntgen, J. Klein: "Entstehung von Erdgas aus kohligen Sedimenten," *Erdöl Kohle Erdgas Petrochem.* **74/75** (1986) no. 1, 52 – 69.
25. R. F. K. Kinghorn: *An Introduction to the Physics and Chemistry of Petroleum,* Wiley-Interscience, New York 1983, pp. 432.
26. I. Köster, H. Kulke, H. W. Ladwein: "Neuentdeckte und potentielle Erdölvorkommen," *Erdöl Kohle Erdgas Petrochem.*(1992) no. 3, 106.
27. D. Leythaeuser, H. W. Ladwein: "Wechselbeziehungen zwischen Erdölgenese und Migration," *Erdöl Kohle Erdgas Petrochem.*(1992) no. 3, 108.

28. T. P. Lowell, A. J. D. Harding: "Structural Styles, Their Plate Tectonic Habitat, and Hydrocarbon Traps in Petroleum Provinces," *Am. Assoc. Petrol. Geol. Bull.* **63** (1979) 1016 – 1058.
29. R. McQuillin, M. Bacon, W. Barclay: *An Introduction to Seismic Interpretation–Reflection Seismics in Petroleum Exploration,* Graham & Trotman, London 1986, pp. 287.
30. A. Melton, A. C. Giardini: "A Scientific Explanation for the Origin and Location of Petroleum Accumulations," *J. Pet. Geol.* **6** (1983) 117 – 138.
31. G. B. Moody: *Petroleum Exploration Handbook,* McGraw-Hill, New York 1961.
32. G. J. Moore, G. T. Demaison: "Anoxic Environments and Oil Source Bed Genesis," *Am. Assoc. Pet. Geol. Bull.* **64** (1980) no. 8, 1179 – 1209.
33. R. J. Murris: "Middle East: Stratigraphic Evolution and Oil Habitat," *Am. Assoc. Pet. Geol. Bull.* **64** (1980) no. 5, 597 – 618.
34. N. S. Neidell: "Stratigraphic Modeling and Interpretation: Geophysical Principles and Techniques," *AAPG Education Course Note Series* **13** (1981) 141.
35. F. K. North: *Petroleum Geology,* Allen & Unwin, Boston 1985, pp. 607.
36. M. R. Pegrum, G. Rees, D. Naylor: *Geology of the North-West European Continental Shelf,* **vol. 2, The North Sea.** Graham Trotman Dudlex Publ., London 1975.
37. R. G. Schäfer, D. H. Welte: "Entstehung von Erdöl und Erdgas," *Prax. Naturwiss. Chem.* **41** (1992) no. 2.
38. M. Schoell: "Kohlenstoff- und Wasserstoffisotopenanalysen in der Erdöl- und Erdgasforschung," *3eme Cycle Romand des Sciences de la Terre,* 1980, pp. 1 – 47.
39. M. Schoell: Genetic Characterisation of Natural Gases, *Am. Assoc. Pet. Geol. Bull.* **67** (1983) no. 12, 2225 – 2238.
40. R. C. Selley: *Elements of Petroleum Geology,* Freeman, New York 1985, pp. 449.
41. R. E. Sheriff: "Seismic Detection of Hydrocarbons: The Underlying Physical Principles," *Offshore Technology Conference 6,* Preprint 1, Houston 1974, pp. 637 – 649.
42. V. Sokolov, A. Geodekian, Z. Buniat-Zade: "The General Scheme of Petroleum and Gas Formation, Alteration and Migration in the Earth's Crust," Adv. Org. Geochem. Proc. Int. Meet. 4th (1968) 279 – 288.
43. V. Stroganov: "Principal Phases in Origin of Gaseous and Liquid Hydrocarbons and Conditions of Formation of Zones of Oil – Gas Accumulations," *Int. Geol. Rev.* **16** (1974) 769 – 776.
44. M. Teichmüller: "Anwendung kohlenpetrographischer Methoden bei der Erdöl- und Erdgasprospektion," *Erdöl Kohle Erdgas Petrochem.* **24** (1971) 69 – 76.
45. E. N. Tiratsoo: *Oilfields of the World,* Scientific Press, Beaconsfield 1984, pp. 392.
46. G. D. Tiratsoo, A. E. N. Hobson: *Introduction to Petroleum Geology,* Scientific Press, Beaconsfield 1981, pp. 352.
47. B. P. Tissot: "Recent Advances in Petroleum Geochemistry Applied To Hydrocarbon Exploration," *Am. Assoc. Pet. Geol. Bull.* **68** (1984) no. 5, 545 – 563.
48. A. Treibs: "Entstehung des Erdöls," *Erdöl Kohle* **1** (1948) 137 – 143, 185 – 199.
49. F. Troost, P. J. P. M. van Beek: "The Groningen Gas Field: A Case History of the Development of a Giant Gas Field," *JPT J. Pet. Technol.*(1979) July, 815 – 820.
50. N. B. Wassojewitsch et al.: "Die Hauptphase der Erdölbildung," *Z. Angew. Geol.* **15** (1969) 612 – 615.
51. D. H. Welte: "Zur Entwicklungsgeschichte von Erdölen auf Grund geochemisch-geologischer Untersuchungen," *Erdöl Kohle Erdgas Petrochem.* **20** (1967) 65 – 77.
52. D. H. Welte: "Petroleum Exploration and Organic Geochemistry," *J. Geochem. Explor.* **1** (1972) 117 – 1939.
53. E. Whitehead: "Molecular Evidence for the Biogenesis of Petroleum and Natural Gas," *Proc. Symp. Hydrogeochem. Biogeochem.* **2** (1973) 158 – 211.
54. P. A. Ziegler: "Öl- und Gas-Provinzen der Nordsee," *Erdöl Erdgas Z.* **91** (1975) 207 – 217.
55. P. A. Ziegler: *Geological Atlas of Western and Central Europe,* Shell International Petroleum Maatschappij B.V., 1982.

Specific References

56. V. B. Porfirev: "Inorganic Origin of Petroleum, *Am. Assoc. Petrol. Geol. Bull.* **58** (1974) 3 – 33.
57. B. Tissot, D. Welte: *Petroleum and Occurrence,* Springer Verlag, Heidelberg 1978.
58. D. H. Welte: "Organischer Kohlenstoff und die Entwicklung der Photosynthese auf der Erde," *Naturwissenschaften* **57** (1970) 17.

59. W. Schott: "Aufsuchen von Erdöl- und Erdgaslagerstätten" in A. Bentz, H. J. Martini (eds.) *Lehrbuch der Angewandten Geologie,* vol. 2, part 1, Enke Verlag, Stuttgart 1968, pp. 563 – 671.
60. C. C. Cornelius: "Geothermal Aspects of Hydrocarbon Exploration in the North Sea Area", *NGU-Nor. Geol. Unters.* **1975,** no. 316, 29 – 67.
61. A. Hood, J. R. Castano: Organic Metamorphism: Its Relationship to Petroleum Generation and Application To Studies of Authigenic Minerals, United Nations ESCAP CCOP, Techn. Bull. **8,** 8 – 113.
62. G. Philippi: "The Influence of Marine and Terrestrial Source Material on The Composition of Petroleum," *Geochim. Cosmochim. Acta* **38** (1974) 947 –966.
63. K. Krejci-Graf: "Moderne Anschauungen über die Entstehung des Erdöls," *Erdöl Kohle Erdgas Petrochem.* **13** (1960) 836 – 845.
64. U. Mann: "Revealing Hydrocarbon Migration Pathways," *Geol. Rundschau* **78** (1990) no. 1, 337 –348.
65. H. Klemme: Petroleum Basins – Classification and Characteristics," *J. Pet. Geol.* **3** (1980) no. 187 – 207.
66. P. A. Ziegler: "Faulting and Graben Formation in Western and Central Europe," in: P. Kent, M. H. P. Bott, D. P. McKenzie, C. A. Williams (eds.): "The Evolution of Sedimentary Basins," *Philos. Trans. R. Soc. Ser. London A* **305** (1982) 113 –143.
67. W. Stahl, J. Koch: "13 C/12 C – Verhältnis norddeutscher Erdgase – Reifemerkmal ihrer Muttersubstanzen," *Erdöl Kohle Erdgas Petrochem.* **27** (1974) no. 10.
68. W. Stahl: "Carbon Isotope Ratios of German Natural Gases in Comparison with Isotope Data of Gaseous Hydrocarbons from Other Parts of the World," *Adv. Org. Geochem. Proc. Int. Meet.* **6th** (1973) 453 – 462.
69. D. H. Welte, M. A. Jükler: "Petroleum Origin and Accumulation in Basin Evolution – A Quantitative Model," *Am. Ass. Petrol. Geol. Bull.* **65** (1981) 1387 – 1390.
70. W. Rühl: *Energiefaktor Erdöl,* Edition Interform, Zürich 1989.
71. C. D. Masters, D. H. Root, E. D. Attanasi: "World Resources of Crude Oil and Natural Gas," *Proceedings of World Petrol. Congress,* Publ. J. Wiley and Sons, Chichester 1991.
72. O. Serra: *Fundamentals of Well Log Interpretation,* vol. 2, Acquisition of logging Data; Development in Petroleum Science, Elsevier, Amsterdam 1987.
73. L. C. Graton, H. J. Fraser: "Systematic Pacing of Spheres – with particular Relation to Porosity and Permeability,"*J. Geol.*(1935) Nov.– Dec., 785 – 909.
74. H. Pape, L. Riepe, J. R. Schopper: "Theory of Self-Similar Network Structures in Sedimentary and Igneous Rocks and their Investigations with Microscopical. and Physical Methods," *J. Microsc. (Oxford)* **148** (1987) part 2, 121 – 147.
75. J. Geertsma: "The Effect of Fluid Pressure Decline Oil Volumetric Changes of Porous Rocks," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **210** (1957).
76. N. H. Hall: "Compressibilities of Reservoir Rocks," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **198** (1953) 309.
77. J. W. Amyx, D. M. Bass, R. L. Whiting: *Petroleum Reservoir Engineering, Physical Properties,* McGraw-Hill, New York 1960.
78. W. D. McCain: *The Properties of Petroleum Fluids,* Petroleum Publishing Company, 1973.
79. E. J. Lynch: *Formation Evaluation,* Harper Row, New York 1962.
80. A. R. Martinez et al.: "Classification and Nomenclature Systems for Petroleum and Petroleum Reserves 1987 Report," *Proc. World Pet. Congr.* **12th** (1987).
81. F. Bender: "Methoden der angewandten Geophysik und mathematische Verfahren in den Geowissenschaften," in *Angewandte Geowissenschaften,* vol. 2, Enke Verlag, Stuttgart 1985.
82. E. Winnock, Y. Pontalier: "Lacq Gas Field; Geology of giant Petroleum Fields," *Am. Ass. Pet. Geol. Bull.* **Nov.** (1970) 370 – 387.
83. J. Lee: "Well Testing," *SPE Textbook Series,* **vol. 1,** Texas A & M University, 1982.
84. D. R. Homer: "Pressure Build-Up in Wells," *Proc. Third World Petroleum Congress,* **vol. II,** Leiden 1951, p. 503.
85. R. E. Collins: *Flow of Fluids through Porous Materials,* Reinhold Publishing Corporation, New York 1961.
86. M. Honarpour, L. Koederitz, A. H. Harvey: *Relative Permeability of Petroleum Reservoirs,* CRC Press, Boca Raton 1986.
87. H. L. Stone: "Probability Model for Estimating Three-Phase Relative Permeability," *J. Petr. Techn.*(1970) Feb., 214 – 219.
88. F. W. Cole: *Reservoir Engineering Manual,* Gulf Publ. Corp., Houston 1969.

89. M. C. Leverett: "Capillary Behavior in Porous Solids," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **142** (1941) 152 – 169.
90. S. E. Buckley, M. C. Leverett: "Mechanism of Fluid Disphacements in Sands," *Trans. Am. Inst. Metall. Pet. Eng.* **146** (1942) 107 – 116.
91. H. J. Welge: "A Simplified Method for Computing Oil Recovery by Gas or Water Drive," *Trans. Am. Inst. Metall. Pet. Eng.* **195** (1952) 91 – 98.
92. F. F. Craig: "The Reservoir Engineering Aspects of Waterflooding," *SPE Monograph,* **vol. 3,** New York 1971.
93. H. Dykstra, R. L. Parsons: "The Prediction of Waterflood Performance with Variation in Permeability Profile," *Prod. Monthly* **15** (1950) 9 – 12.
94. B. C. Craft, M. F. Hawkins: *Applied Petroleum Reservoir Engineering,* Prentice Hall, Englerwood Cliffs 1959.
95. J. Tarner: "How Different Size Gas Caps and Pressure Maintenance Programs Affect Amount of Recoverable Oil," *Oil Weekly* **144** (1944) June 12, 32 – 34.
96. D. W. Peaceman: "Fundamentals of Numerical Reservoir Simulation," vol. 6, *Developments in Petroleum Science,* Elsevier, Amsterdam 1977.
97. H. B. Crichlow: *Modern Reservoir Engineering – A Simulation Approach,* Prentice-Hall, Englewood Cliffs, New Jersey 1977.
98. H. C. Slider: *Worldwide Practical Petroleum Reservoir Engineering Methods,* Penn Well Publ. Co., Tulsa 1983.

General References

99. W. Arnold: *Eroberung der Tiefe,* VEB Dtsch. Vlg. f. Grundstoffind., Leipzig 1965.
100. D. Kett, *Gas-Wasserfach* **108** (1967) 221 – 229.
101. *Oil Gas J.* **73** (1975) Jan., 72.
102. *Oil Gas J.* **72** (1974) July, 33.
103. S. J. Pirson, *Oil Reservoir Engineering,* McGraw-Hill, New York 1958.
104. P. L. Randolph, *World Oil* 1974, Aug., 33 – 36;Sept., 65 – 71; Oct., 131 – 134; Nov. ,81 – 84.
105. H. Rischmüller, *Erdöl Erdgas Z.* **19** (1966) 91 – 104.
106. H. Rischmüller, *Erdöl Erdgas Z.* **82** (1966) 90 – 99.
107. H. Rischmüller, H.-M. Tröndle: "Meerestechnik beim Ozeanbergbau," in G. Dietrich (ed.): Erforschung des Meeres, Umschau-Vlg., Frankfurt 1970, 267 – 285.
108. H. Rischmüller, *Erdöl Erdgas Z.* **88** (1972) 110 – 114.
109. K. Trenel, *Erdöl Erdgas Z.* **83** (1967) 188 – 191.
110. W. Tunn, *Erdöl Erdgas Z.* **87** (1971) July, 202 – 208.
111. E. Wichert, K. Aziz, *Can. J. Chem. Eng.* **49** (1971) 267 – 273.
112. R. A. Sumbatov, B. N. Khakhaev, V. N. Gramolin: "Drilling Rigs and Technical Facilities for Superdeep Well Drilling," Paper presented at the IGC, Washington 1989.
113. D. Noran: *Oil Gas J.* **73** (1975) Nov., 77 – 81.
114. *Erdölinformationsdienst* **28** (1975) no. 41.
115. *Erdölinformationsdienst* **29** (1975) no. 5.
116. W. Simmler, *Gas-Wasserfach* **109** (1968) 724 – 730.
117. Mannesmannröhren-Werke: *Rohre für Öl- und Gasfelder,* 6th ed., **vols 1 – 3,** 1962.
118. *International Petroleum Encyclopedia,* The Petroleum Publ. Co., Tulsa 1975.

Specific References

119. R. D. Faniev: *Abbau von Erdöl- und Erdgaslagerstätten,* VEB Dtsch. Vlg. f. Grundstoffind., Leipzig 1963.
120. A. Mayer-Gürr: "Erdöl- und Erdgasgewinnung in der Bundesrepublik Deutschland," *Oel-Zeitschrift für die Mineralölwirtschaft* **1** (1973).
121. K. K. Millheim: "The New Engineering Paradigm and the Emergence of Investigative Engineering," *J. Petrol. Technol.* 1989, Aug.
122. *Ullmann* 4th ed., **11,** pp. 19 – 40.
123. H. Rischmüller: "Tiefbohrtechnik als Schlüsselinstrument in der Energiewirtschaft," *Geowissenschaften* **10** (1990) 317 – 323.
124. Ö. Aliquander: *Das moderne Rotarybohren,* VEB Dtsch. Vlg. f. Grundstoffind., Leipzig 1965.
125. W. Arnold: *Eroberung der Tiefe,* VEB Dtsch. Vlg. f. Grundstoffind., Leipzig 1983.
126. H. F. Spörker: "Entwicklungstendenzen der Tiefbohrtechnik," *Erdöl Erdgas Z.* **88** (1972) 104 – 110.
127. K.-H. Grodde: *Bohrspülungen und Zementschlämme in der Tiefbohrtechnik,* Vlg. d. Erdöl-Z. Otto Vieth, Hamburg 1963.
128. C. Marx, H. Rischmüller: "Drilling and Coring Techniques for Hard Rock," in H. J. Behr, F. G. Stehli, H. Vidal (eds.): *Observation of the Continental Crust through Drilling II,* Springer, Berlin 1987, 149 – 160.

129. Maurer Engineering Inc.: *Deep Drilling Cost Sensitivity Analysis,* Study for Gas Research Institute, Chicago 1989, GRI Contract No. 5088-260-1731.
130. F. P. Springer: "Bohrtechnische Aspekte bei der Ausrüstung von Tiefbohranlagen," *Erdöl Erdgas Z.* **88** (1972) 410 – 417.
131. R. S. Andrews: *A Survey of Recent Technology Development in International Continental Scientific Drilling Programs,* Joint Oceanographic Institutions, Inc., 1990.
132. H. Baker: Horizontal Wells, Reprinted from a series of articles appearing *Oil & Gas J.* Sept. 24 – 31, 1990.
133. H.-J. Behr, R. Emmermann: Scientific Objectives and Site-Selection Studies of the Continental Deep Drilling Programme of the Federal Republic of Germany (KTB), in K.-J. Behr, F. G. Stehli, H. Vidal (eds.): *Observation of the Continental Crust through Drilling II,* Springer, Berlin 1987, 85 – 120.
134. H.-J. Behr, P. Kehrer, H. Rischmüller: "The German Continental Deep Drilling Program, Objectives and State of Work," in A. Boden, K. G. Erikson (eds.): Deep Drilling in Crystalline Bedrock, vol. 2, Springer, Berlin 1988, 64 – 81. W. B. Bleakley, *Oil Gas J.* **72** (1974) March, 69 – 78.
135. T. Burgess, P. Van de Slijke: "Horizontal Drilling Comes of Age," *Oilfield Review* 1980, July.
136. C. Chur, B. Engeser, J. Oppelt: "Das Vertikalbohrkonzept für die KTB-Hauptbohrung," *Erdöl Erdgas Kohle* **106** (1990) 486 – 490.
137. M. Ellins, T. Tran Viet: "Spülungskonzept und Mud Logging-System für die KTB-Hauptbohrung," *Erdöl Erdgas Kohle* **106** (1990) 190, 491 – 495.
138. R. Emmermann: "Das Deutsche Kontinentale Tiefbohrprogramm," *Geowissenschaften in unserer Zeit* **1** (1986) 19 – 33.
139. E. P. Ilaysov et al.: "Ultradeep Drilling for the Investigation of the Earth's Crust," *12th World Petroleum Congress,* Preprint of Review and Forecast, Topic 12, Houston 1987.
140. H. Rischmüller, C. Chur: "Technical Concept and Status of Planning for the Wells of the Continental Deep Drilling Program (KTB)," in H.-J. Behr, F. G. Stehli, H. Vidal (eds.): *Observation of the Continental Crust through Drilling II,* Springer, Berlin 1987, 136 – 149.
141. H. Rischmüller: "Das kontinentale Tiefbohrprogramm der Bundesrepublik Deutschland – ein Synergismus von Geowissenschaften und Technik," in *Bergbau im Wandel,* Akademische Druck- und Verlagsanstalt, Verlag Glückauf, Essen 1988, 143 – 152.
142. A. Sperber et al.: "Das technische Konzept der KTB-Hauptbohrung," *3rd Colloquium DFG-Schwerpunktprogramm,* Gießen 1990.
143. H. F. Spörker: "Thoughts on Problems and Possible Solutions in Ultradeep Drilling," in H.-J. Behr, F. G. Stehli, H. Vidal (eds.): *Observation of the Continental Crust through Drilling II,* Springer, Berlin 1987, 170 – 184.
144. T. C. Frick: *Petroleum Production Handbook,* McGraw-Hill, New York 1962.
145. J. E. Kastrop, *Petroleum Eng.* **47** (1975) Feb., 27 – 31.
146. G. V. Chilingar, C. M. Beeson: *Surface Operations in Petroleum Production,* Elsevier, New York 1969.
147. K. H. Grodde: *Bohrspülungen und Zementschlämme in der Tiefbohrtechnik,* Vlg. d. Erdoel-Z. Otto Vieth, Hamburg 1963.
148. H.-G. Graf, H.-G. Haddenhorst, H. Rischmüller, W. De la Sauce: "Förder- und Speichertechnik," *Erdöl, Erdgas* **100** (1984) no. 4.
149. A. P. Szilas: *Production and Transport of Oil and Gas,* Elsevier, Amsterdam 1975.
150. J. W. Amyx, D. H. Bass, R. L. Whitting: *Petroleum Reservoir Engineering,* McGraw-Hill, New York 1960.
151. N. J. Clark: Elements of Petroleum Reservoirs. *Soc. Petroleum Eng. of Amer. Inst. of Mining, Metall. Petroleum Eng.,* Dallas 1960.
152. B. C. Craft, M. F. Hawkins: *Applied Petroleum Reservoir Engineering,* Prentice-Hall, Englewood Cliffs 1959.
153. C. R. Smith: *Mechanics of Secondary Oil Recovery,* Reinhold Publ. Corp., New York 1966.
154. M. K. Robson: "Introduction to and Benefits of Tubing-Conveyed Perforating," *JPT,* **Feb.** 1990.
155. H.-J. Dietzel, G. v. Hantelmann, H. Rischmüller: "Artifizielle Fließwege im flözführenden Karbon, *Erdöl Kohle-Erdgas,* **37** (1984) no. 12.
156. J. Gulbis et al.: "Taking the Brakes off Proppant-Pack Conductivity," *Oilfield Review* 1991 Jan.
157. G. C. Howard, C. R. Fast: "Hydraulic Fracturing," *H. L. Doherty Series Monograph.* **vol. 2,** Am. Inst. Min. Metall. & Petrol. Eng., 1970.

158. M. L. Slusser, M. Rieckmann: *Erdöl Kohle Erdgas Petrochem. Compendium 75/76,* 149 – 161.
159. J. D. Clegg: "High-Rate Artificial Lift," J. Petrol. Technol. **1988,** March.
160. B. C. Craft, W. R. Holden, E. D. Graves Jr.: *Well Design, Drilling and Production,* Prentice-Hall, Englewood Cliffs 1962.
161. T. E. W. Nind: *Principles of Oil Well Production,* Mc Graw-Hill, New York 1964.
162. H. Rischmüller: "Alternative synergistische Förderhilfsmittelplanung für produzierende Erdölfelder," *Erdöl-Erdgas-Z.* **88** (1972) April.
163. H. Rischmüller: "Kurzer Überblick über den Stand bewährter und die Entwicklung neuer Verfahren in der Erdölfördertechnik," *Erdöl, Erdgas, Kohle* **102** (1986) no. 1.
164. H. Muskat: *The Flow of Homogeneous Fluids through Porous Media,* Edwards Broth. Inc., Ann Arbor 1946.
165. C. J. Gott: "Successful Rod Pumping at 14.500 Feet," *SPE Production Engineering* 1986 Nov.
166. J. Zaba: *Modern Oil-Well Pumping,* Petroleum Publ. Co., Tulsa 1962.
167. H. Rischmüller: *Erdölförderung mit Gestängetiefpumpen,* Verlag Vereinigte Edelstahlwerke AG und Böhler AG, Düsseldorf 1984.
168. K. Horn, R. Putzer, H. Rischmüller: "Komplettierung und Förderung von Offshore-Sonden," *Erdöl Z.* **80** (1964) 379 – 395.
169. D. L. Katz et al.: *Handbook of Natural Gas Engineering,* McGraw-Hill, New York 1959.
170. H. Laurien: *Taschenbuch Erdgas,* 2nd ed., R. Oldenbourg Vlg., München 1970.
171. M. Medici: *The Natural Gas Industry,* Butterworth, London 1974.
172. H. Rischmüller: "Die Gaslagerstätten des Westemslandes," *Erdöl Erdgas Z.* **84** (1968) 322 – 329.
173. K. Hinz, H. Amann, G. Schendl: "Erdölexploration im tieferen Wasser," *Erdöl Kohle* **28** (1975) April, 175 – 182.

Rhenium and Rhenium Compounds

HANS GEORG NADLER, H. C. Starck, Goslar, Federal Republic of Germany

1. Introduction . . . 1
2. Properties . . . 2
3. Occurrence and Raw Materials . . . 2
3.1. Molybdenite . . . 2
3.2. Recovery from Spent Catalysts and Alloys . . . 3
4. Production . . . 3
4.1. Recovery of Rhenium from Flue Gases in Molybdenite Roasting . . . 3
4.2. Production of Pure Rhenium Compounds from Gas Scrubbing Liquors 4
4.3. Production of Rhenium Metal Powder and Pellets . . . 5
5. Rhenium as an Alloying Component . 5
6. Uses . . . 6
7. Rhenium Compounds . . . 6
7.1. Oxides and Chalcogenides . . . 7
7.2. Halides . . . 7
7.3. Carbonyls and Organorhenium Compounds . . . 7
8. Uses of Rhenium Compounds . . . 8
9. Analysis . . . 9
10. Economic Aspects . . . 10
11. Toxicology and Physiological Effects . 10
12. References . . . 10

1. Introduction

Rhenium (Re) is present in the earth's crust at a very low concentration (ca. 7×10^{-8} %, 0.7 ppm). The element was not discovered until 1925, when TACKE, NODDACK, and BERG observed its X-ray spectrum in concentrates and compounds obtained from the beneficiation of columbite [5].

The discoverers isolated the first gram of rhenium from 660 kg of Norwegian molybdenite (MoS_2), which contained 2 – 4 ppm Re [6]. Rhenium is the most recently discovered refractory metal; it is a member of the group of high-melting metals that includes molybdenum, tungsten, niobium, and tantalum. It has the second highest melting point of all metals (3453 K).

Industrial production of rhenium began in the 1930s at Kali Werke Aschersleben in the Harz region [7], and also at H. C. Starck (HCST), formerly Gebr. Borchers AG, in Goslar. The raw material consisted of furnace bottoms, byproducts from the treatment of Mansfeld copper schists. The furnace bottoms were digested by fusion with sodium sulfate [8]. Then, after a complicated chemical separation process, rhenium was isolated as potassium perrhenate, which was reduced with hydrogen to give the impure metal.

An important advance in rhenium technology was the production of the element from flue dusts from the roasting of molybdenite [9]. The amount recovered from the dusts was, however, only 10 – 15 % of the total Re content of the MoS_2 concentrate. The introduction of accurate analytical methods [10] enabled material balances for rhenium extraction to be established. These showed that most of the rhenium was lost up the stack, and only a small amount remained in the flue dust or in the roasted product. The flue gas was therefore scrubbed intensively to ensure that the volatile oxides of rhenium were recovered when the molybdenite concentrate was roasted [11, 12].

2. Properties

Physical Properties. Rhenium [*7440-15-5*] has the electronic configuration [Xe]$4f^{14}5d^5 6s^2$, and consists of the natural isotopes ^{185}Re (37.07 %) and ^{187}Re (62.93 %, β-emitter with a half-life of 10^{11} years). Fused or sintered rhenium has a silvery white luster, and the powdered metal is silver gray. It has a hexagonal close-packed structure (type A 3), with $a_0 = 276$ pm and $c_0 = 446$ pm. Rhenium retains this lattice up to its melting point and hence does not undergo a ductile – brittle transformation, in contrast to other refractory metals.

Other physical data are listed below:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_199

Atomic number	75
Relative atomic mass A_r	186.207
Density (20 °C)	21.0 g/cm^3
Metallic radius	137 pm
Ionic radius (VII)	53 pm
Melting point	3453 K (3180 °C)
Boiling point	6143 K (5870 °C)
Heat of fusion	33 kJ/mol
Heat of sublimation	
ΔH_B (monatomic gas)	+779 (± 8) kJ/mol
Enthalpy of formation of Re_2O_7	−1241 kJ/mol
Electrical conductivity	0.051 $\mu\Omega^{-1}$ cm^{-1}
Specific resistivity (20 °C)	19.3 $\mu\Omega \cdot$ cm
Superconductivity	1.699 K
Tensile strength (20 °C, annealed)	12 MPa
Modulus of elasticity (20 °C)	459.9 GPa
Recrystallization temperature (depending on degree of working)	1573 – 2073 K

Chemical Properties. Rhenium is a very heat-resistant metal, provided it does not come in contact with oxidizing agents. It is practically insoluble in hydrochloric and hydrofluoric acids. In oxidizing acids, it dissolves to form perrhenic acid (E^0 for Re/ReO_2: 0.251 V).

Rhenium forms volatile oxides with oxygen at high temperature. In air at 350 °C, the heptoxide, Re_2O_7, is formed. The rate of oxidation increases with increasing air flow rate. The stability of rhenium metal components is limited due to oxide formation even below 0.133 Pa [13]. When rhenium is heated with fluorine or chlorine, the fluorides or chlorides are formed. Rhenium reacts with silicon, boron, and phosphorus at elevated temperature to form silicides, borides, and phosphides.

3. Occurrence and Raw Materials

3.1. Molybdenite

Molybdenite, MoS_2 [*1317-33-5*], especially from porphyry copper ore deposits, is the preferred host mineral for rhenium, which isomorphically replaces molybdenum in the MoS_2 lattice. The copper – molybdenum porphyry ores contain 0.3 – 1.6 % Cu and 0.01 – 0.05 % Mo. The sulfides of copper and molybdenum are concentrated and separated by flotation, giving a rhenium content in the MoS_2 concentrates of several hundred parts per million.

Rhenium contents (in ppm) of selected molybdenum concentrates of various origins are as follows:

Canada	
Island Copper	700 – 1300
HVC	200 – 400
Endako	< 100
United States	
Pinto Valley	1500 – 2000
Magma San Manuel	800
Bagdad	350
Sierrita	180
Bingham Canyon	250
Climax	30
Henderson	20
Mexico	
La Caridad	570
Peru	
Cuajone	580
Toquepala	600
Chile	
Chuquicamata	300
El Salvador	600
El Teniente	400
Iran	
Sar Chesmeh	800

World rhenium reserves [14] (in t Re) are

Canada	32
Chile	1306
Peru	45
CIS	594
United States	386
Others	91
World total	2453

In copper concentrates, the rhenium content is 10 – 50 ppm, the molybdenum content 200 – 600 ppm. Copper concentrates are therefore a useful source of rhenium.

3.2. Recovery from Spent Catalysts and Alloys

Platinum – rhenium catalysts are used in the petroleum industry in the reforming process. When their activity is too low even after regeneration, the catalysts must be replaced. Valuable metals are recovered by dissolving the carrier (Al_2O_3) in sodium hydroxide solution or sulfuric acid. Platinum remains in the residue, from which it can be recovered. Rhenium, which goes into solution, can be recovered by solid or liquid ion exchangers [15].

In addition to this hydrometallurgical procedure, a pyrometallurgical process (fusion) allows rapid concentration of the valuable metals [16]. The resulting alloys are treated by dissolution and separation processes.

The high economic value of the WRe, MoRe, and rhenium- containing superalloys means that scrap alloys of these compositions must be processed. The following processes are suitable:

1) Oxidation by atmospheric oxygen to form WO_3 and Re_2O_7 (*bp* 362 °C), which are separated by sublimation and distillation [17]
2) Fusion with NaOH and oxidizing agents such as $NaNO_3$ and $NaNO_2$ [18]

These initial operations are followed by isolation of the elements and crystallization of their salts.

4. Production

4.1. Recovery of Rhenium from Flue Gases in Molybdenite Roasting

Porphyry copper ores containing 0.3 – 1.6 % Cu and 0.01 – 0.05 % Mo are ground and concentrated by flotation (→ Molybdenum and Molybdenum Compounds, Chap. 4.1.). Further treatment leads to almost quantitative recovery of molybdenite.

When molybdenite (90 – 95 % MoS_2, remainder: SiO_2, Cu, Fe, etc.) is roasted at 500 – 700 °C to produce technical-grade molybdenum trioxide, MoO_3 [*1313-27-5*], rhenium heptoxide, Re_2O_7 [*1314-68-7*], sublimes due to its high vapor pressure, and escapes with the flue gases. The vapor pressure curves of some oxides and sulfides of rhenium and molybdenum are given in Figure 1 [19].

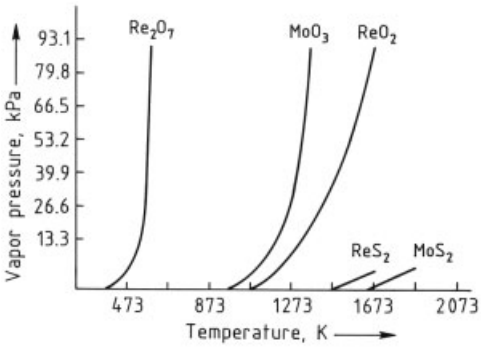

Figure 1. Vapor pressure of some oxides and sulfides of rhenium and molybdenum

The flue gases also contain sulfur dioxide, small amounts of sulfur trioxide, selenium dioxide, and large amounts of dust (MoO_3, MoO_2, and MoS_2). This dust can be recovered (> 99 %) in high-temperature electrostatic filters and recycled to the roasting process. In the final gas scrubbing operation, the remaining dust fraction, rhenium heptoxide, selenium dioxide, sulfur trioxide, and some of the sulfur dioxide are removed from the flue gas. The scrubbing liquor is recirculated to increase the rhenium concentration, allowing the subsequent process stages to be operated economically.

The compound Re_2O_7 is readily soluble in water, forming $HReO_4$, so that the recovery of rhenium from the flue gases is > 90 %. Multistage pressure jet (venturi) scrubbers enable high rhenium recovery efficiencies to be achieved, effectively removing the finest dust fractions and causing rhenium oxides to be absorbed.

The venturi system operates with a pressure difference of 30 hPa in the first stage and 150 hPa in the second stage. The mist formed in the venturi washing process is trapped in a packed tower. This intensive gas purification produces waste gases that contain very pure SO_2, which can be used for sulfuric acid production [11]. The recovery of rhenium from the flue gases in the roasting of molybdenite is shown schematically in Figure 2.

The roasted molybdenum concentrates are used mainly to produce MoO_3 for direct use in metallurgy (production of ferromolybdenum) and in molybdenum chemistry (→ Molybdenum and Molybdenum Compounds). For economic reasons, optimum conditions for rhenium recovery cannot always be maintained (temperature, air excess, etc.). The yield of rhenium is also limited by various impurities in the concentrates, especially alkali-metal and alkaline-earth elements, which form nonvolatile stable perrhenates under roasting conditions.

Other possible processes for the treatment of molybdenum concentrates include electrooxidation [20] and treatment with O_2 or NH_3 at high temperature and pressure [21]. Yields of Mo and Re are ca. 98 %. These processes have not been used thus far on an industrial scale.

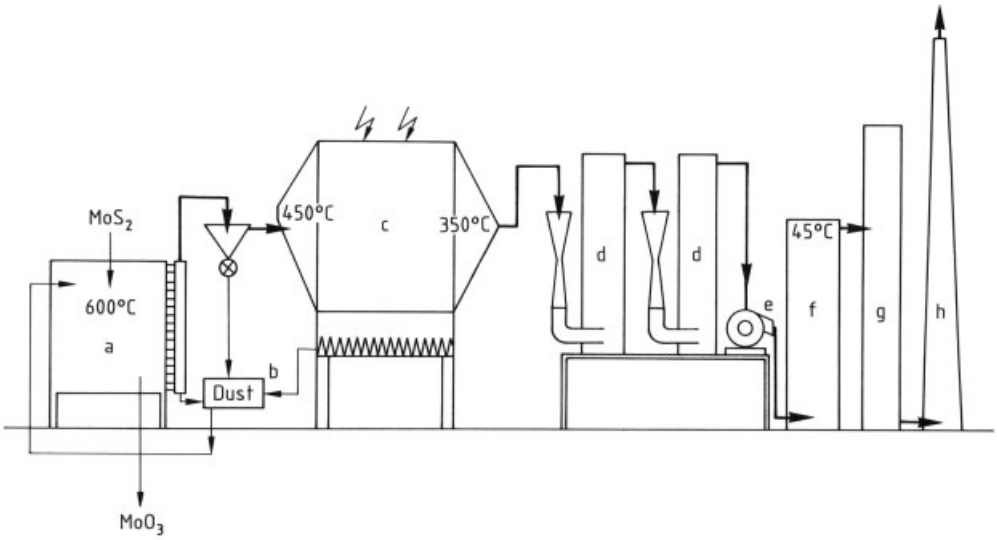

Figure 2. Flue gas purification with rhenium recovery by H. C. Starck
a) Roasting furnace; b) Dust chamber; c) Electrostatic filter; d) Venturi scrubbers; e) Ventilator; f) Secondary washer; g) SO_2 conversion; h) Stack

4.2. Production of Pure Rhenium Compounds from Gas Scrubbing Liquors

Gas scrubbing liquors contain sulfuric acid and, typically, rhenium at 0.2 – 1.5 g/L. After removal of the solid constituents such as molybdenum oxides, sulfides, and selenium, rhenium can be precipitated as its sparingly soluble sulfide Re_2S_7 [*12038-67-4*], or by cementation on iron or zinc. It can then be concentrated further.

However, the use of ion exchangers in solid or liquid form is preferable for recovering and concentrating rhenium. Primary, secondary, and tertiary long- chain amines, dissolved in organic solvents such as kerosene, can be used to extract rhenium and molybdenum at pH 1 – 2. At higher pH, quaternary amines are more effective [22]. Separation of molybdenum from rhenium can be achieved by extraction at an optimum pH. Alternatively, both elements can be extracted from acidic solution and then separated by selective stripping with ammonia at different pH values [23].

Rhenium in neutral, alkaline, or sulfuric acid solution can be fixed by strongly basic solid ion exchangers. The molybdate ion is less strongly bound to the ion exchanger than the perrhenate ion, enabling the molybdenum to be removed by a displacement reaction. Rhenium can be eluted from the ion-exchange resin by strong mineral acids, especially perchloric acid, or ammonium thiocyanate [24].

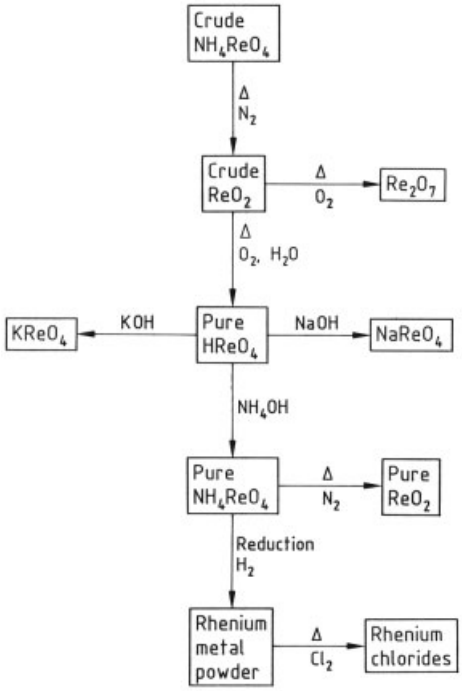

Figure 3. Rhenium production

Technical-grade ammonium perrhenate is produced by crystallization from the eluate solution. Further purification is carried out by recrystallization.

Solvent extraction, combined with the use of solid ion exchangers or crystallization, can be used to produce pure or highly pure ammonium perrhenate, which is the starting material for the production of many rhenium compounds as well as the metal [25]. A schematic of rhenium production is given in Figure 3.

Another method of rhenium production consists of solvent extraction combined with electrowinning of the rhenium from perchlorate solutions [26].

4.3. Production of Rhenium Metal Powder and Pellets

Rhenium powder [*7440-15-5*] can be produced by the reduction of pure ammonium perrhenate NH_4ReO_4 [*13598-65-7*] in a stream of hydrogen. This reduction is usually carried out in two stages:

1) Reduction at 300 – 350 °C to form ReO_2
2) Reduction of the ReO_2 to Re metal powder at 800 °C

The desired properties of the metal powder, such as grain size and surface area,can be achieved by adjustment of the reduction parameters.

In a further stage, the rhenium metal powder with added compaction agents is compressed into pellets and then sintered at ca. 1000 – 1500 °C under an atmosphere of hydrogen. To produce large workpieces, the powdered metal is charged into rubber containers, compressed isostatically at ca. 500 MPa, and then sintered. Rhenium can be produced at almost its theoretical density by hot isostatic pressing (HIP) and by electron-beam or electric-arc melting.

Rhenium has very high strength at high temperature, as well as extremely good ductility and cold working properties. It is therefore a very useful high-temperature material. In some applications, the useful properties of rhenium can best be realized by chemical vapor deposition (CVD) [27]. Plasma coating with spheroidized rhenium metal powder is another possibility, producing dense rhenium coatings when combined with a sintering operation.

5. Rhenium as an Alloying Component

When rhenium is used as an alloying component with the metals tungsten and molybdenum, which are difficult to work, ductility and strength are improved [28]. This is caused by alloy softening, which is defined as the reduction in the yield stress or hardness at low temperature. This effect is observed especially in body-centered cubic alloys. In addition to the improvement in ductility at low temperature, the strength at high temperature increases [29]. Tungsten – rhenium alloys containing 25 – 30 % Re have good cold ductility [30]. As indicated in Figure 4, rhenium is very soluble in tungsten, its solubility reaching 28 wt % at 1600 °C and 37 wt % at 3000 °C. Tungsten is also soluble in rhenium: 11 wt % at 1600 °C and 20 wt % at 2825 °C. Besides a tetragonal σ-phase, a cubic χ-phase exists in a very narrow range in the W – Re system [31].

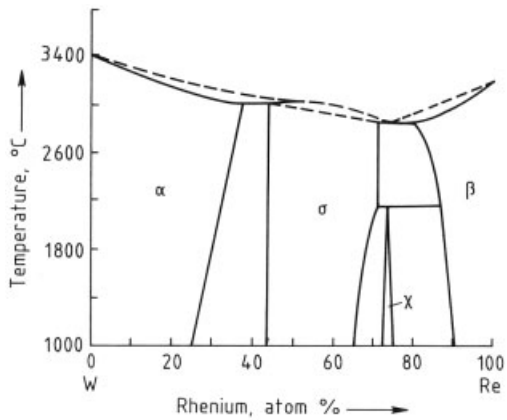

Figure 4. Tungsten – rhenium phase diagram [31]

The poor welding properties of common molybdenum alloys limit their use in high-temperature applications. The addition of rhenium enables these alloys to be welded and used as construction materials. Molybdenum – rhenium alloys containing up to 50 % Re show high tensile strength and good elongation over the temperature range 0 – 1800 °C [32].

Superalloys must have not only high strength, but also good corrosion resistance at elevated

temperature. They may be divided into three groups: nickel-base alloys, cobalt-base alloys, and iron (nickel)-base alloys.

Other typical alloying elements include chromium, molybdenum, tungsten, niobium, tantalum, aluminum, titanium, boron, and zircon. The addition 1 – 3 % Re to a nickel-base alloy improves its toughness at high temperature and prevents fatigue fracture [33].

6. Uses

Rhenium in the form of the powdered metal or pellets is incorporated into alloys by various methods. The W – Re and Mo – Re alloys are used mainly in the manufacture of thermoelements. Other uses include semiconductors, heating elements, electrical and electronic applications, high-temperature welding rods, and metallic coatings [34].

Tungsten – rhenium alloys (W5Re, W10Re) are used in the manufacture of rotating X-ray anodes (Fig. 5). For normal X-ray diagnosis, composite anodes, consisting of a very hard molybdenum base with an annular W – Re coating, are used. Computer tomography requires larger anodes, made by using a hybrid metal – graphite construction combined with a W – Re coating [35]. An alternative technique is to deposit the metals or alloys by CVD or PVD processes [36].

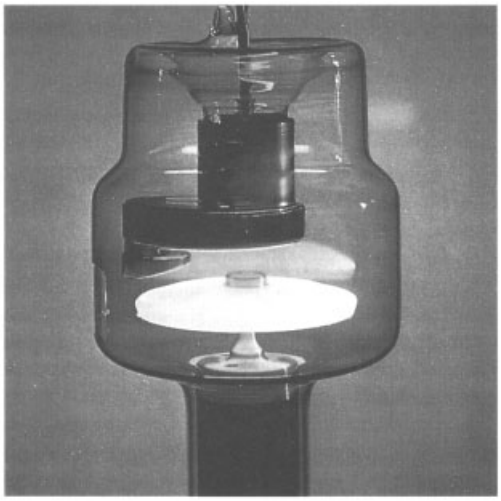

Figure 5. Rotating X-ray anode

Nickel-base alloys containing 1 – 3 wt % Re are used mainly in the production of aircraft turbine blades. They are monocrystalline and have high strength and resistance to oxidation. When these turbine blades are used in the hot zones of an engine, operating temperature can be increased, giving higher efficiency (lower fuel consumption). Similar effects are achieved by coating the gas turbine blades with rheniumcontaining (1 – 20 % Re) MCrAlY alloys [37].

7. Rhenium Compounds

Rhenium closely resembles its neighbors in the sixth row (W, Os) and molybdenum (diagonal relationship) in its physical properties; many chemical parallels also exist. Eleven oxidation states of rhenium have been described [38]. In the most important rhenium compounds such as perrhenic acid, the perrhenates, and dirhenium heptoxide, rhenium has the oxidation state + 7. It exhibits its lowest oxidation states (as low as – 3) in carbonyl complexes.

7.1. Oxides and Chalcogenides

The yellow hygroscopic rhenium(VII) oxide, Re_2O_7 [*1314-68-7*], is the most stable oxide of rhenium. It is formed from rhenium metal powder or other rhenium oxides in dry air or an oxygen atmosphere above 350 °C. The oxide Re_2O_7 is readily soluble in water, forming the colorless perrhenic acid, $HReO_4$ [*13768-11-1*], which is pale yellow at high concentration.

Perrhenic acid forms salts ($MReO_4$) with a tetrahedral ReO_4^- ion. Ammonium perrhenate, NH_4ReO_4, is an important starting material, which can be reduced to Re metal and used for the production of many other rhenium compounds. The preparation of some rhenium compounds is shown in Figure 6. When Re_2O_7 is heated with rhenium powder, the oxides ReO_3 [*1314-28-9*], Re_2O_5 [*12165-05-8*], ReO_2 [*1314-28-9*], and Re_2O_3 [*12060-05-8*] are formed.

In Re_2O_7, ReO_4 tetrahedra and ReO_6 octahedra are linked alternately at their corners. ReO_3 disproportionates on heating to form Re_2O_7 and ReO_2. The bluish black rhenium(IV) oxide, ReO_2, which is insoluble in water, has a distorted

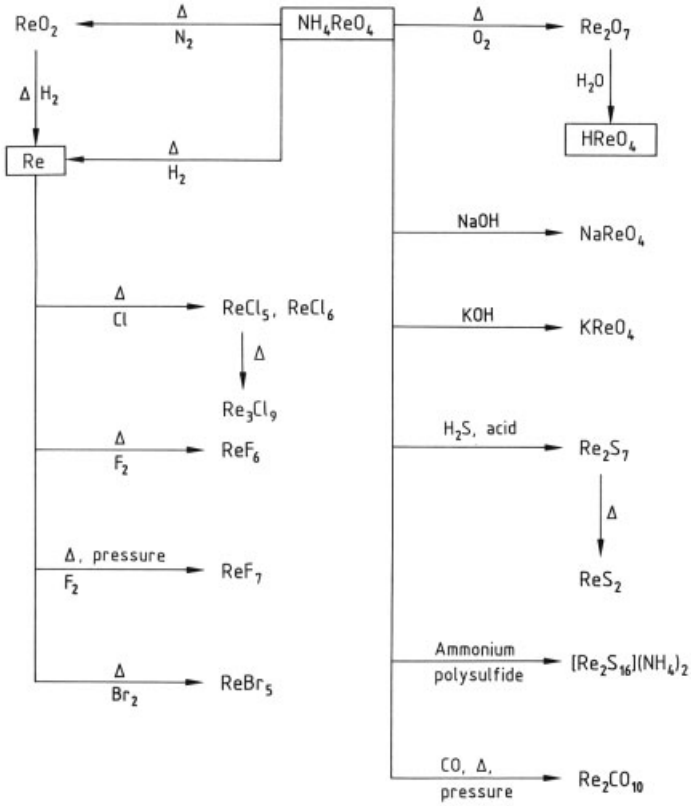

Figure 6. Synthesis of rhenium compounds

rutile structure and disproportionates at 900 °C into Re_2O_7 and metallic Re.

With sulfur, rhenium forms Re_2S_7 [*12038-67-4*], ReS_3 [*12166-08-4*], and ReS_2 [*12038-63-0*]. The black rhenium(VII) sulfide is formed when hydrogen sulfide is passed into acidic solutions of perrhenates. Reduction of Re_2S_7 by hydrogen produces ReS_3. The most stable sulfide ReS_2 is formed by the thermal decomposition of Re_2S_7 or by heating the elements together. The structure of ReS_2 (trigonal prismatic) is isomorphous with the rhombohedral modification of MoS_2 [38].

Table 1 lists the properties of some oxides and sulfides of rhenium.

Ammonium perrhenate reacts with aqueous ammonium polysulfide to form $(NH_4)_2(Re_2S_{16})$, a polynuclear sulfide having a metal – metal bond with marked double-bond character [39].

7.2. Halides

The most important rhenium halides can be prepared from the elements (see Fig. 6). Rhenium(VII) fluoride, ReF_7 [*17029-21-9*], is the highest halide. It is formed at 400 °C under slight pressure. Rhenium(VI) fluoride, ReF_6 [*10049-17-9*], is formed from the elements at 125 °C. Rhenium(V) fluoride, ReF_5 [*30937-52-1*], and rhenium(IV) fluoride, ReF_4 [*15192-42-4*], can be prepared by reduction of ReF_6. Suitable reducing agents include H_2, metallic Re, SO_2, and Zn.

Rhenium pentachloride, Re_2Cl_{10} [*39368-69-9*], is an important starting material for the synthesis of other rhenium compounds. It is formed from the elements at 500 °C. In the crystal, $ReCl_6$ octahedra form edge-sharing dimers. Rhenium (III) chloride, Re_3Cl_9 [*14973-59-2*], can be produced by ther-

Table 1. Properties of some rhenium oxides and sulfides[38]

Compound	*mp/bp*, °C	Structure	Oxidation state
Re_2O_7	300/360	alternating tetrahedra and octahedra	VII
ReO_3	> 300/750	cubic	VI
ReO_2	900 (decomp.) → Re_2O_7 + Re	rutile	IV
Re_2S_7	> 250 (decomp.) → ReS_2 + S		VII
ReS_2	1000 (decomp.) → Re + S	trigonal prismatic	IV

mal decomposition of the pentachloride. All rhenium(III) chlorides, bromides, and iodides contain the Re_3X_9 unit with Re_3 metal clusters.

$ReBr_3$ [*13569-49-8*] or $ReCl_3$ reacts with aqueous solutions of HBr to form $ReBr_3 \cdot \frac{2}{3} H_2O$. Single crystals can be obtained by evaporation of the solution [40]. Rhenium(IV) bromide, $ReBr_4$ [*36753-03-4*], and ReI_4 [*59301-47-2*] are formed by the reaction of $HReO_4$ with HBr or HI.

Some properties of rhenium halides are given in Table 2 [38]. A large number of complex halides and oxyhalides of rhenium are known [41]. In the latter, rhenium is linked to oxygen by both single and double bonds. With nitrogen, triple, double, and single bonds are known [42].

7.3. Carbonyls and Organorhenium Compounds

Rhenium carbonyl, $Re_2(CO)_{10}$ [*14285-68-8*], is formed by the reaction of CO with Re_2O_7 or Re_2S_7 at elevated pressure and temperature (20 MPa, 250 °C). It forms colorless crystals (*mp* 177 °C) and sublimes in vacuum. From the binary rhenium carbonyl, many cluster carbonyls, carbonyl hydrides, and complexes with a wide range of ligands in various oxidation states can be prepared [43].

A large number of organorhenium compounds with a wide range of structures, properties, and reactivities are now known [44]. The syntheses of hexamethylrhenium, $Re(CH_3)_6$ [*56090-02-9*] [45], and of the half-sandwich complex, $[(\eta^5\text{-}C_5Me_5)ReO_3]$ [*90695-83-3*] [46], have enabled new model substances for homogeneous organometallic catalysis to be prepared.

8. Uses of Rhenium Compounds

Rhenium compounds are used in various areas of homogeneous and heterogeneous catalysis in petrochemistry, the pharmaceutical industry, and organic synthesis (e.g., alkylation, dehydroisomerization, dealkylation, isomerization, dehydrochlorination, hydrogenation, dehydrogenation, dehydrocyclization, oxidation, and reforming [47]).

Reforming (→ Benzene, Chap. 5.2.1.; → Oil Refining, Chap. 3.4.) is an important part of petroleum processing. It is used in the production of unleaded antiknock fuels for internal combustion engines and aromatic hydrocarbons for various syntheses [48]. Originally, molybdenum oxide was used for reforming, but it was soon found that platinum on aluminum oxide (Al_2O_3) had better catalytic properties. An increase in catalytic activity, selectivity, and particularly stability was achieved with the development of platinum – rhenium catalysts. The production of Pt – Re bimetallic catalysts by Chevron and UOP began in 1968 (Rheniform process). The rhenium components were solutions of NH_4ReO_4 or $HReO_4$. Multimetallic catalysts such as platinum – rhenium – indium have been developed from the earlier bimetallic catalysts [49]. The reforming process is still the most important catalytic application of rhenium compounds. In 1990, 45 % of rhenium consumption in the United States was for this process [14].

New rhenium compounds such as methyltrioxorhenium, CH_3ReO_3 [*70197-13-6*], which can be produced fairly simply from Re_2O_7 and tetramethyltin, exhibit catalytic activity in the synthesis of alkenes (alkene metathesis), alkene oxidation, and alkenation of aldehydes [50].

Basic research in the field of rhenium compounds is expected to lead to the discovery of further catalytic properties of known rhenium

Table 2. Rhenium halides

Oxidation state	Fluorides	Chlorides	Bromides	Iodides
VII	ReF_7 yellow *mp* 48.3 °C *bp* 73.7 °C			
VI	ReF_6 yellow *mp* 18.5 °C *bp* 33.7 °C	$ReCl_6$ brown – green *mp* 29 °C		
V	ReF_5 yellow – green *mp* 48 °C *bp* 221 °C (extrapolated)	$ReCl_5$ black – brown *mp* 220 °C	$ReBr_5$ dark brown 110 °C (decomp.)	
IV	ReF_4 blue *mp* 124.5 °C sublimes 300 °C	$ReCl_4$ black – violet 300 °C (decomp.)	$ReBr_4$	ReI_4 (decomp. > room temperature)
III		$(ReCl_3)_3$ dark red sublimes 500 °C (decomp.)	$(ReBr_3)_3$ red – brown	$(ReI_3)_3$ black (decomp. on heating)

compounds and of compounds as yet undiscovered.

9. Analysis

The analytical chemistry of rhenium includes the determination of rhenium in raw materials and intermediates, and the determination of trace impurities in high-purity rhenium products.

Rhenium can be determined by various analytical methods after separation by selective precipitation, distillation, extraction, and ion exchange. W. Geilmann and H. Bode carried out important research into the analysis of rhenium in the 1940s and 1950s [51]. The most important practical application is the rapid and convenient determination of small amounts of rhenium in the presence of large amounts of molybdenum, which invariably accompanies it. After treatment of a sample with alkali, rhenium is determined in the Re(VII) state as the perrhenate ion. The accompanying elements are separated and masked, and Re(VII) is then precipitated by addition of the tetraphenylarsonium cation and determined gravimetrically [52]. In acidic solution in the presence of $SnCl_2$, Re(VII) forms colored chelates with oximes and thiocyanates, which are extracted and determined by spectrophotometry [53]. Rhenium can be determined by atomic absorption spectrometry using the resonance line at 346.047 nm. Analysis of solutions by emission spectrometry uses the lines at 346.047 and 346.473 nm [54]. Polarographic methods are also available [55].

The determination of trace elements in high-purity rhenium products, such as NH_4ReO_4 or rhenium metal, is carried out by combustion analysis (C), hot extraction by carrier gas (N, O), and atomic absorption spectroscopy on solutions (metallic impurities) [56].

10. Economic Aspects

The largest world reserves of rhenium are located in Chile which is today the world's largest producer. Other producers include Germany, CIS, Sweden, the United States, and Japan. In Germany, rhenium is obtained from molybdenum concentrates, spent catalysts, and rhenium-containing scrap, and processed to produce high-purity NH_4ReO_4, $HReO_4$, and Re metal powder or pellets [57]. The largest consumer of rhenium is the United States, which imported ca. 15 t of Re in 1990. This could increase to 20 t/a in the next five years. Rhenium is used for aircraft turbine blades (60 %), reforming catalysts (30 %), and other applications (10 %) [14].

11. Toxicology and Physiological Effects

No evidence exists that rhenium is an essential element for living organisms [58]. People who have worked over a long period of time in the extraction and production of rhenium metal and its most important compounds, perrhenic acid and ammonium perrhenate, have not shown any toxic effects. Toxicological investigations of rhenium and rhenium compounds are contradictory and do not enable any OECD guidelines to be issued [59].

Medical studies indicate that rhenium carboxylates have an antitumor effect [60]. Rhenium-188 and rhenium-186 are used in combination with monoclonal antibodies for the diagnosis and treatment of leukemia [61].

12. References

General References

1. K. B. Lebedev: *The Chemistry of Rhenium,* Butterworths, London 1962.
2. R. D. Peacock: *The Chemistry of Technetium and Rhenium,* Elsevier, Amsterdam 1966.
3. B. W. Gonser (ed.): *Rhenium,* Elsevier, Amsterdam 1962.
4. A. Sutulov: *Molybdenum and Rhenium 1778 – 1977,* University of Concepción, Concepción, Chile 1976.

Specific References

5. W. Noddack, I. Tacke, O. Berg, *Naturwissenschaften* **13** (1925) 567 – 574.
6. I. Noddack, W. Noddack, *Z. Anorg. Allg. Chem.* **183** (1929) 353 – 375.
7. W. Feit, *Angew. Chem.* **43** (1930) 459 – 462.
8. V. Tafel: *Lehrbuch der Metallhüttenkunde,* **vol. 3,** S. Hirzel, Leipzig 1954, pp. 224 – 230.
9. A. D. Melaven, J. A. Bacon, US 2 414 969, 1947.
10. S. Tribalat, *Anal. Chim. Acta* **3** (1949) 113 – 125. S. Tribalat, *Anal. Chim. Acta* **4** (1950) 228 – 234.
11. H. G. Nadler, P. Borchers, *Erzmetall* **40** (1987) 293 –298.
12. S. R. Zimmerley, E. E. Malouf, US 2 809 092, 1952.
13. H. Jehn, R. Völker, *Z. Metallkd.* **67** (1976) H. 11, 715 – 719.
14. J. W. Blossom: *Mineral Industry Surveys, Annual Review Rhenium in 1991,* US Bureau of Mines, Washington, DC, 1992.
15. Universal Oil Products Comp., US 3 672 874, 1972 (C. L. Wiley, L. Blanchard). Amer. Cyanamid, US 4 599 153, 1986 (J. H. Bright).
16. Kennecott Corp., EP 004 882 3A1, 1982 (P. Cichy).
17. M. J. Ferrante, F. E. Block, A. D. Fugate, F. A. Skirvin: *Recovery of Rhenium from Tungsten-Rhenium Alloy,* RI7254 U.S. Dept. of the Interior, Bureau of Mines 1969.
18. GTE Products Corp., US 4 557 906, 1985 (A. D. Douglas, K. T. Reilly, J. E. Landmesser).
19. B. Jezowska-Trzebiatowska, S. Kopacz, T. Mikulski: *The Rare Elements,* Elsevier, Amsterdam 1990, p. 361.
20. B. J. Scheiner, R. E. Lindstrom, D. L. Pool: *Extraction and Recovery of Molybdenum and Rhenium from Molybdenite Concentrates by Electrooxidation: Process Demonstration,* RI 8145, US Bureau of Mines 1976.
21. DE 344 380 6A1, 1986 (E. Gock). DE 371 072 5A1, 1988 (E. Gock, J. Friedrich, J. Kähler). American Metal Climax Inc., DE 2 045 308, 1971 (C. J. Hallada, H. F. Barry, R. W. McConnell).
22. G. M. Ritcey, A. W. Ashbrook: *Solvent Extraction,* Part II, Elsevier, Amsterdam 1979, pp. 421 – 425. G. Huifa, Shen Jinglau, M. A. Hughes, *Hydrometallurgy* **25** (1990) 293 – 304.
23. J. Kähler, E. Gock, *Erzmetall* **41** (1988) no. 3, 132 –137.
24. Kennecott Copper Corp., US 3 558 268, 1971 (J. D. Prater, R. N. Platzke).
25. H. G. Nadler, P. Borchers in GDMB Gesellschaft Deutscher Metallhütten- und Bergleute: *Die Anwendung von festen und flüssigen Ionenaustauschern bei der Gewinnung und Raffination von Metallen,* Heft 53, Clausthal-Zellerfeld 1989, pp. 162 – 174.
26. P. E. Churchward, J. B. Rosenbaum: *Sources and Recovery Methods for Rhenium,* **RI6246,** US Bureau of Mines 1963.
27. A. J. Sherman, R. H. Tuffias, R. B. Kaplan, *J. Met.* (1991) July, 20 – 23.
28. G. A. Geach, J. R. Hughes in F. Benesovsky (ed.): *The Alloys of Rhenium and Molybdenum or with Tungsten and Having good High-temperature Properties, Plansee Proceedings,* Pergamon Press, London 1955, pp. 245 –253.

29. E. Pink, R. J. Arsenault, *Prog. Mater. Sci.* **24** (1979) 1 – 52.
30. J. E. Peters, *J. Met.* (1969) April, 27 – 30.
31. J. M. Dickinson, L. S. Richardson, *Trans. Am. Soc. Met.* **51** (1959) 758. S. V. Nagender Naidu, P. Rama Rao, *J. Alloy Phase Diagrams* **6** (1990) no. 3, 129 – 136.
32. R. Eck, *Int. J. Refract. Hard Met.* **5** (1986) no. 1, 43 –48.
33. General Electric Comp., EP 040 368 1A1, 1990 (M. F. Henry).
34. H. Cross Comp.: *Rhenium and Rhenium Alloys,* Weekhawken, New Jersey, USA.
35. H. Bildstein, *BHM, Berg Hüttenmänn. Monatsh.* **136** (1991) no. 9, 359 – 361.
36. Metallwerk Plansee, EP 039 962 1A1, 1990 (P. Rödhammer).
37. Siemens AG, EP 041 239 7A1, 1991 (N. Czech).
38. N. N. Greenwood, A. Earnshaw: *Chemistry of the Elements,* Pergamon Press, Oxford 1984.
39. A. Müller et al, *Angew. Chem.* **103** (1991) 1501 – 1503.
40. B. Jung, H. Ehrhardt, G. Meyer, *Z. Anorg. Allg. Chem.* **603** (1991) 49 – 56.
41. R. D. Peacock in J. C. Bailar, H. J. Emelius, Sir R. Nyholm, A. F. Trotman-Dickenson (eds.): *Comprehensive Inorganic Chemistry,* **vol. 3,** Pergamon Press, Oxford 1973, pp. 905 – 978.
42. K. Dehnicke, J. Strähle, *Angew. Chem.* **93** (1981) 451 –463; *Angew. Chem.* **104** (1992) 978 – 1000.
43. F. A. Cotton, G. Wilkinson: *Anorganische Chemie,* VCH Verlagsgesellschaft, Weinheim 1985.
44. *Gmelin,* **Organorhenium Compounds, part 1, part 2.** A. Segnitz: *Methoden der Org. Chemie, Metallorganische Verbindungen,* **vol. 13/9 a,** Thieme Verlag, Stuttgart 1986, p. 122.
45. K. Mertis, G. Wilkinson, *J. Chem. Soc. Dalton Trans.* 1976, 1488.
46. W. A. Herrmann, *Angew. Chem.* **100** (1988) 1269 –1286.
47. M. H. Corrigan, W. H. Davenport, J. W. Spelman: *A Bibliography on the Catalytic Applications of Rhenium,* Cleveland Refractory Metals, Weekhawken, New Jersey, USA 1968.
48. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie,* VCH Verlagsgesellschaft, Weinheim 1988, 334 – 337.
49. UOP Inc., US 4 629 551, 1986 (G. J. Antos, Li Wang). Exxon Research and Engineering Company, EP 010 653 1B1, 1984 (G. E. Markley, W. E. Winter).
50. W. A. Herrmann et al., *Angew. Chem.* **103** (1991) 183 –185; *Angew. Chem.* **103** (1991) 1704 – 1711.
51. M. E. Peterson, J. S. Mac Duff, M. W. Hovey: *Isolation and Colorimetric Determination of Rhenium,* **RI5889,** US Bureau of Mines 1961.
52. *Gmelin,* System no. 70, Rhenium, Verlag Chemie, Berlin 1941, 69 – 72.
53. O. G. Koch, G. A. Koch-Dedic: *Handbuch der Spurenanalyse,* part 2, Springer-Verlag, Berlin 1974, 976 –986.
54. O. D. Bozhkov, N. Jordanov, L. V. Borissova, Yu. I. Fabelinskii, *Fresenius Z. Anal. Chem.* **321** (1985) 453 –456.
55. H. Y. Chen, R. Neeb, *Fresenius Z. Anal. Chem.* **320** (1985) 247 – 251.
56. N. Jordanov, I. Hazevov, O. Bozhkov, *Fresenius Z. Anal. Chem.* **335** (1989) 910 – 913. A. G. Khatchatrian, *J. Planar Chromatogr.* **2** (1989) April, 142 – 147.
57. Roskill: *The Economics of Rhenium,* 4th ed., Roskill Information Services Ltd., London 1988, pp. 7 – 8.
58. P. Jacquet in H. G. Seiler, H. Sigel, A. Sigel (eds.): *Handbook on Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988, pp. 557 – 560.
59. J. T. Haley, F. D. Cartwright, *J. Pharm. Sci.* **57** (1968) 321 – 323. S. V. Suvorov, *Nov. Dannye Toksikol. Redk. Met. Ikh Soedin.* 1967, 45 – 51; *Chem. Abstr.* **71** (1969) 53274 h.
60. G. W. Estland, G. Yang, T. Thompson, *Methods Find. Exp. Clin. Pharmacol* **5** (1983) no. 7, 435 – 438.
61. Sloan Kettering Inst. Cancer, US 450 918, 1989 (D. A. Scheinberg).

Rodenticides

ALAN BUCKLE, ICI Public Health, Fernhurst, Haslemere, United Kingdom

1.	**Introduction**	1
2.	**Commensal Rodents**	1
3.	**Rodents as Pests of Agriculture**	2
4.	**Control of Rodents**	2
5.	**Rodenticides Used in Poisoned Baits**	2
5.1.	**Acute and Subacute Rodenticides**	2
5.2.	**First-Generation Anticoagulants**	4
5.2.1.	Hydroxycoumarins	5
5.2.2.	Indandiones	6
5.2.3.	Anticoagulant Resistance	7
5.3.	**Second-Generation Anticoagulants**	7
6.	**Other Chemical Methods of Rodent Control**	9
7.	**References**	10

1. Introduction

The order Rodentia is the largest mammalian order with ca. 1500 species, although barely 100 of these are important as pests. Pest rodents belong mainly to the families Sciuridae (tree and ground squirrels), Cricetidae (voles, gerbils, and hamsters), and Muridae (rats and mice). Rodenticides are used in the control of all these animals and, occasionally, for the control of those belonging to the closely related order Lagomorpha (rabbits). Rodents are pests in both commensal and agricultural situations.

2. Commensal Rodents

Evidence of traps made to capture rodents in the early cultures of Mesopotamia and Egypt witness the long relationship between humans and these damaging animals. Three rodent species, the Norway rat (*Rattus norvegicus*), the ship rat (*Rattus rattus*), and the house mouse (*Mus musculus*), have their origins in Asia [1] but have spread to threaten the health and livelihoods of humans across the globe. The same species have even journeyed with humans to hundreds of oceanic islands and have done untold damage to their fragile ecosystems [2].

Diseases are transmitted from rodents to humans and domesticated animals, either directly by bites or exposure to infected feces and urine, or indirectly via a vector, usually an arthropod such as a fly or mosquito. The range of diseases is very large and their causative organisms include viruses, bacteria, protozoa, and helminths. Among the most important diseases carried by rodents and transmitted to humans are plague, leptospirosis, salmonellosis, typhus, leishmaniasis, trichinosis, schistosomiasis, and a number of hemorrhagic and relapsing fevers [3].

Almost all foodstuffs produced for consumption by humans and domestic animals are liable to damage by commensal rodents during processing and storage but financial loss is very difficult to define [4]. Norway rats and house mice prefer dry foods based on cereals, while ship rats prefer moist foods such as fruit.

The incisor teeth of Norway rats have a hardness of 5.5 on Moh's scale and, during gnawing, exert a pressure of up to 500 kg/cm^2 [5]. Armed with these formidable weapons, rodents damage a wide range of natural and man-made materials while gaining access to food and harborage and merely while relieving boredom. Of particular importance is the damage done to pipes, conduits, and cables, both because of the high cost of repair and the fires resulting from electrical short circuits.

An adult rat annually produces some 30 000 droppings and 8 L of urine. These, together with innumerable hairs, are shed indiscriminately and the cost of cleaning soiled commodities frequently far exceeds the value of foodstuffs actually consumed by rodents.

3. Rodents as Pests of Agriculture

Rodents are among the most damaging pests of growing crops. Cereals (e.g., rice, wheat, barley, and maize) and plantation crops (e.g.,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_211

coconuts, oil palm, cocoa, and sugarcane) are highly vulnerable to attack. Some of the species responsible for these depredations are *Sigmodon hispidus* in Central and South America, *Arvicanthis niloticus* and *Praomys natalensis* in sub-Saharan Africa, *Bandicota bengalensis* in the Indian subcontinent, *Rattus tiomanicus* and *R. argentiventer* in south-east Asia, and *Rattus exulans* among the islands of the south Pacific [5]. In Europe, the main rodent pests of agriculture are voles, *Microtus agrestis*, *M. arvalis,* and *Arvicola terrestris* [6]. In the United States, a number of species of ground squirrels and voles are important pests of crops and rangelands.

4. Control of Rodents

Rodenticides are the mainstay of most rodent-control operations. They are principally applied in the form of poisoned baits, often based on cereals, and placed in areas where target rodents are active. In dry locations, such as cereal stores, rodenticides are occasionally presented in the form of liquids. A few compounds are also available as tracking dusts which are applied to surfaces over which the rodents pass. The contaminated rodents then ingest the poisoned dust during grooming.

The safe, effective application of rodenticides requires a degree of skill. Baits must be placed so that all rodents have access to them in sufficient quantity to acquire a lethal dose. Few of the compounds employed possess a substantial degree of specificity to rodents. It is therefore imperative to prevent the consumption of baits by nontargets by the exclusion of these animals from the baited area, by the careful placement of baits, or by the use of bait boxes.

Nonchemical methods of rodent control include the use of traps, the proofing of buildings against the ingress of rodents, modification of the environment to make it less conducive to rodent infestation (removal of food and harborage), and the use of predatory animals, such as cats and dogs. A number of devices are sold which emit ultrasound that is said to repel rodents. These, however, have been shown to be largely ineffective in most practical circumstances [7].

5. Rodenticides Used in Poisoned Baits

5.1. Acute and Subacute Rodenticides

Acute and subacute rodenticides induce rapid effects. The majority of *acute rodenticides* cause death, after the consumption of a lethal dose, within a few minutes to several hours. The fast onset of the symptoms of toxicosis has a drawback however. Rodents that consume only a sublethal amount of the poison associate the painful effects with the new food (i.e., the bait), and refuse to eat it again. These animals are said then to be bait shy. Bait shyness can be prevented, at least in part, by prebaiting. Here, unpoisoned bait of the type later to be used to carry the poison is first presented. The poison is only added when the rodents have overcome their initial suspicion of the novel food and are feeding freely on it. This technique maximizes the quantity of poison taken by rodents at their first feeding, thereby reducing the likelihood of sublethal exposure.

Subacute rodenticides also have relatively rapid effects, one of the most common being anorexia [8]. However, when a potentially lethal dose has been consumed, they cause death after ca. 1 – 3 days.

Zinc phosphide [*1314-84-7*], Zn_3P_2, trizinc diphosphide, M_r 258.06, $mp > 420\,°C$, is one of the most common acute rodenticides. It is generally available as a grey or black crystalline powder, 80 – 95 % purity, with a strong garlic odor. It is applied in baits at concentrations of 1 – 5 %, 2 % is most widely used. The acute oral LD_{50} for Norway rats is ca. 40 – 45 mg/kg [5]. Zinc phosphide is also highly toxic to a wide range of vertebrates (LD_{50} for pig, dog, cat, chicken, and mallard in the range 20 – 40 mg/kg) [9]. The mode of action is by the evolution of phosphine gas in the acid environment of the stomach. The gas enters the bloodstream, causing heart failure and damage to internal organs [10]. There is no specific antidote [9].

Red Squill is extracted from bulbs of the Mediterranean plant, *Urginea maritima.* The active substance, a cardiac and nerve glycoside, *scilliroside* [*507-60-8*], $C_{32}H_{44}O_{12}$, 3β(β-

D-glucopyranosyloxy-17β-(2-oxo-2*H*-pyran-5-yl)-14β-androst-4-ene-6β,8,14-triol 6-acetate, M_r 620.67, *mp* 168 – 170 °C, is slightly soluble in water and soluble in alcohols and glycol:

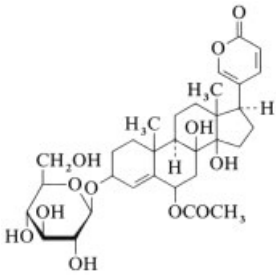

The rodent toxicity of crude extracts is very difficult to determine; an LD_{50} dose to Norway rats of 125 – 1500 mg/kg has been reported [9]. A purified extract is highly toxic to *Rattus norvegicus* (LD_{50} 0.4 – 2.6 mg/kg) and *Mus musculus* ($LD_{50} < 0.2$ mg/kg) [5]. Concentrations used in baits range from 0.015 to 0.05 %, although unpalatability [5] and bait shyness [11] have been reported. Symptoms of poisoning include hind limb paralysis, convulsions, emesis (except in rat), and excessive diuresis and diarrhea. Atropine sulfate is antidotal [9].

Trade Name. Stabilized extract, Silmurin (Sandoz).

Calciferol [*50-14-6*], vitamin D_2, ergocalciferol, $C_{28}H_{44}O$, (3β,5*Z*,7*E*,22*E*)-9,10-secoergosta-5,7,10(19),22-tetraen-3-ol, M_r 396.63, *mp* 115 – 118 °C [12]:

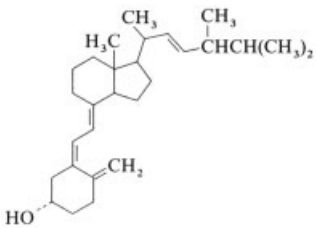

This naturally-occurring compound is used for rodent control both on its own and in combination with 0.025 % warfarin and 0.005 % difenacoum, although synergism is unproven [13]. It is stable in baits at a concentration of 0.1 % for at least 6 months, but after this time its efficiency may diminish [5]. The acute oral LD_{50} for rats and mice is 30 – 100 mg/kg. Smaller doses may be additive over a period of several days. However, there is some indication that sublethal doses cause anorexia and bait shyness [8]. Toxic doses in mammals stimulate the absorption of calcium in the intestines and the mobilization of skeletal calcium. This results in hypercalcemia, osteomalacia, and the calcification of soft tissues, particularly the major arteries and kidneys. Treatment of accidental poisoning is symptomatic with cortisone and sodium sulfate [5].

Cholecalciferol [*67-97-0*], vitamin D_3, has been introduced more recently as a rodenticide and has similar properties [5].

Trade Names. Sorexa-CD (ergocalciferol with difenacoum), Sorexa-CR (ergocalciferol with warfarin), both Sorex. Quintox (cholecalciferol), Bell Laboratories.

Bromethalin [*63333-35-7*], α,α,α-trifluoro-*N*-methyl-4,6-dinitro-*N*-(2,4,6-tribromophenyl)-*o*-toluidine, $C_{14}H_7Br_3F_3N_3O_4$, M_r 577.9, *mp* 150 – 151 °C. A pale yellow crystalline solid, insoluble in water, soluble in acetone, chloroform, and dichloromethane:

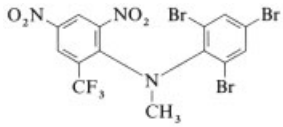

This compound has been known for some time to have rodenticidal properties [14] but has only recently been made available. The acute oral LD_{50} is in the range 2.0 – 8.1 mg/kg for rodents and 1.8 – 13.0 mg/kg for several nontarget species, such as dog, cat, rabbit, and chicken [15]. Bromethalin is used in baits at a concentration of 0.005 or 0.01 % and is effective against rodents resistant to anticoagulants. Anorexia occurs after an effective dose has been consumed. The mode of action is by uncoupling oxidative phosphorylation in cells of the central nervous system. Symptoms of poisoning include tremors, convulsions, prostration, and hind-limb paralysis. No specific antidote is available but a symptomatic treatment has been described [16].

Trade Names. Assault (Purina Mills), Vengeance (Dow-Elanco).

The following acute and subacute compounds are used for rodent control less frequently than those listed above, either because they are not widely available or because restrictions have been placed upon their use due to safety considerations [10, 17].

Sodium monofluoroacetate [*62-74-8*], $CH_2FCOONa$, $C_2H_2FNaO_2$, M_r 100, *mp* 200 °C, commonly known as Compound 1080, is highly toxic to rodents, having an LD_{50} to rats and mice in the range 0.1 – 17.0 mg/kg [9]. Baits contain 0.08 – 0.5 % of the active ingredient, which acts by blocking the tricarboxylic acid cycle, causing the accumulation of citric acid and leading to convulsions and respiratory or circulatory failure. The poison is nonspecific [5]. Because of its high toxicity, the lack of antidote, and its secondary hazard, the use of Compound 1080 is carefully regulated in many countries [9].

Fluoroacetamide [*640-19-7*], CH_2FCONH_2, C_2H_4FNO, M_r 77, *mp* 109 °C, a colorless, odorless, tasteless, white powder that is highly soluble in water [9]. Also known as Compound 1081, it is an analog of Compound 1080 and is similar to it. However, it is used at higher concentration because it is less potent. Used in bait at concentrations of 1 or 2 % it performed better than 0.25 % 1080 and is most often used in the control of rats in sewers and in other closed areas [5].

α-Chloralose [*15879-93-3*], (*R*)-1,2-*O*-(2,2,2-trichloroethylidene)-α-D-glucofuranose, $C_8H_{11}Cl_3O_6$, M_r 309.54, *mp* 187 °C, is a fast-acting narcotizing agent [9]. It slows several metabolic processes, causing hypothermia and eventual death. It is most effective against small rodents, such as mice, which have a high surface area-to-volume ratio, and in cool conditions. It is used in baits at concentrations of 2 – 8 % [5].

The use of thallium sulfate, pyriminyl (Vacor), silatrane, norbormide, and crimidine in rodent control has virtually ceased.

5.2. First-Generation Anticoagulants

The origin of first-generation anticoagulants lies in research conducted in the 1930s in the United States aimed at discovering the causative agent of a hemorrhagic disease of cattle. This was found to be dicoumarin, a contaminant isolated from spoiled sweet clover hay [18]. Further work focused on the potential of this compound and a series of synthetic derivatives in the treatment of human thrombosis. Warfarin, the most active of the series, was soon recognized as a promising therapeutic and rodenticide [19]. In the United Kingdom, dicoumarin itself was used in rodent control [20] until the superior potency of warfarin led to its replacement. The advantages of the anticoagulants over their fast-acting predecessors were quickly recognized and, by the early 1950s, their use dominated rodent control.

The benefits of anticoagulants are related to their chronic mode of action. They interrupt the vitamin K cycle by blocking the enzyme vitamin K epoxide reductase. Vitamin K is required in the production of several blood clotting factors. When recycling of vitamin K is blocked by the anticoagulant, dietary vitamin K is insufficient to maintain adequate clotting factor production. A few days after the ingestion of an effective dose, the factors circulating in the blood are depleted and clotting fails, leading to a fatal hemorrhage [5]. This delayed action means that rodents are unable to relate the symptoms of toxicosis to the anticoagulant that has caused it; therefore no bait shyness occurs. A further important advantage of these compounds is that vitamin K_1 is a specific antidote and, in cases of accidental poisoning, their chronic mode of action allows sufficient time for its administration.

During 1950 – 1970, a number of anticoagulant compounds were commercialized as rodenticides; all were either hydroxycoumarins or members of a related group, the indandiones. The compounds described in Section 5.2 are known as the first-generation anticoagulants. Although acute oral LD_{50} values are given for purposes of comparison, in practice it is unlikely that a lethal dose of any of these rodenticides would be consumed by rodents during a single episode of feeding on poisoned bait.

5.2.1. Hydroxycoumarins

Warfarin [*81-81-2*], 3-(α-acetonylbenzyl)-4-hydroxycoumarin, $C_{19}H_{16}O_4$, M_r 308.22, *mp*

159 – 162 °C, crystalline solid, insoluble in water, soluble in acetone [21]:

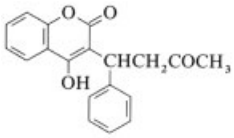

Although values for the acute oral LD_{50} of Norway rats vary between 1.5 and 323 mg/kg [9], the most reliable estimates place the figure between 10 and 20 mg/kg [22]; the strain and sex of the test animals and the carrier used in the administration affect the results. Warfarin is widely used for the control of Norway rats but other pest species, such as *R. rattus* and *M. musculus*, are less susceptible [23] and the compound is not much used in their control. A wide variety of formulations is available, including concentrates containing 0.5 – 1.0 % for bait preparation and use as tracking dusts and baits containing 0.025 – 0.05 % of the active ingredient [17]. Resistance to warfarin was first discovered in the United Kingdom in 1958 and is now established in Norway rat and house mouse populations throughout Europe and North America [23]. Mixtures of warfarin with both sulfaquinoxiline (Proline) and calciferol (Sorexa CR) have been used in proprietary products but the effectiveness of these additives has not been fully demonstrated.

Anticoagulants do not generally produce carcinogenic, teratogenic, or mutagenic effects but there is evidence that warfarin may adversely affect human fetal development [24].

Coumafuryl [*117-52-2*], 3-(α-acetonylfurfuryl)-4-hydroxycoumarin, $C_{17}H_{14}O_5$, M_r 298.28, *mp* 124 °C, is an early anticoagulant now rarely used in rodent control.

Coumachlor [*81-82-3*], 3-(α-acetonyl-4-chlorobenzyl)-4-hydroxycoumarin, $C_{19}H_{15}ClO_4$, M_r 342.79, *mp* 169 – 171 °C, colorless, crystalline solid [25]:

Few reliable data are available on the acute toxicity of coumachlor. The acute oral LD_{50} is given as 187 mg/kg for rats [26] and 900 – 1200 mg/kg for rats and mice [5]. However, the compound is more potent in small daily doses, with a chronic LD_{50} of only 0.1 – 1.0 mg kg^{-1} d^{-1} for 14 – 21 days [27]. The active ingredient is recommended in prepared baits at a concentration of 0.025 – 0.05 %. It is available as concentrates and tracking powders at both 0.5 and 1.0 % [17].

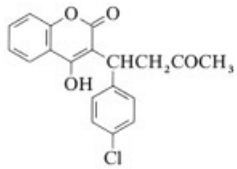

Trade Names. Ratilan, Tomorin (Ciba-Geigy).

Coumatetralyl [*5836-29-3*], 4-hydroxy-3-(1,2,3,4-tetrahydro-1-naphthyl)coumarin, $C_{19}H_{16}O_3$, M_r 292.34, *mp* 172 – 176 °C. Coumatetralyl is produced by the condensation of 4-hydroxycoumarin and 1,2,3,4-tetrahydro-1-naphthol [28]. It is an off-white crystalline powder that is practically odorless, tasteless, and virtually insoluble in water but soluble in many organic solvents:

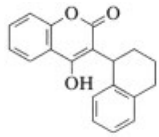

First introduced in 1956, coumatetralyl is one of the most widely used first-generation anticoagulants. It is not, however, available in the United States. The acute oral LD_{50} for Norway rats is given as 16.6 mg/kg [29] but, as with many of these compounds, coumatetralyl is more potent when administered in consecutive daily doses (chronic LD_{50}, 0.3 mg kg^{-1} d^{-1} for 5 days) [26]. No satisfactory data are available on the acute oral toxicity of the compound to mice. However, a complete kill of mice was obtained in the laboratory following 21 days of feeding on coumatetralyl bait [30]. Proprietary baits generally carry 0.375 % coumatetralyl. Powder concentrates are available (strength 0.75 %) which are used for the preparation of poisoned baits and as a tracking dust [17].

Trade Name. Racumin (Bayer).

5.2.2. Indandiones

Diphacinone [*82-66-6*], 2-(diphenylacetyl)indan-1,3-dione, $C_{23}H_{16}O_3$, M_r 340.38, *mp* 145–147 °C, is a yellow powder that is soluble in many organic solvents [31]:

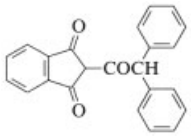

This compound is toxic to Norway rats. The acute oral LD_{50} is given as 2.3 mg/kg [26], 3.0 mg/kg [32], 17 mg/kg [33], and 23 mg/kg for females and 43 mg/kg for males [34]. It is, however, considerably less active against house mice, with acute oral LD_{50} values calculated at 141 mg/kg [29] and 340 mg/kg [32]. The compound is not widely used except for rat control in the United States and, as the sodium salt, for rodent control in China. It is available in a number of formulations, including 0.1–0.5 % powder concentrates; pelleted, meal, and wax block ready-to-use baits containing 0.005–0.05 % of the active ingredient; a 0.1 % water-soluble concentrate based on sugar; and tracking powders containing up to 2 % diphacinone [17]. Generally higher concentrations of the active material are used for mouse control than for rat control.

Trade Names. Diphacin, Ramik (Velsicol).

Chlorophacinone [*3691-35-8*], 2-[2-(4-chlorophenyl)-2-phenylacetyl]indan-1,3-dione, $C_{23}H_{15}ClO_3$, M_r 374.83, *mp* 140 °C. A yellow crystalline solid; sparingly soluble in water; soluble in acetone, ethanol, and ethyl acetate [35]:

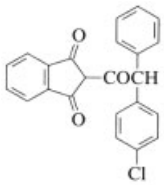

This compound was introduced in 1961 and is now widely used in Europe, the United States, and elsewhere. The acute oral LD_{50} to *R. norvegicus* is 20.5 mg/kg [36]; the compound is used in baits against these animals at a concentration of 0.005–0.01 %. Bait containing 0.025 % chlorophacinone gave a complete kill of house mice after a 7-d feeding period but, in other tests, survivors were recorded following 10 and 21 days of feeding [5]. Chlorophacinone acts as an uncoupler of oxidative phosphorylation as well as an anticoagulant [26], which is unusual among this group of compounds.

Trade Names. Drat (Ciba-Geigy); Rozol (LiphaTech); Caid, Liphadione (Lipha).

Pindone [*83-26-1*], 2-pivaloylindan-1,3-dione, $C_{14}H_{14}O_3$, M_r 230.25, *mp* 108.5–110.5 °C, a yellow-brown crystalline solid, slightly soluble in water, soluble in organic solvents [37]:

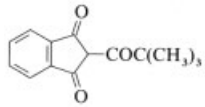

First introduced as an insecticide, and only later commercialized for its rodenticidal properties [38], this compound is used mainly in the United States. The acute oral LD_{50} for Norway rats is given as 50 [29] and 280 mg/kg [39]. Baits containing 0.005–0.05 % have been used for the control of rats and mice. Pindone is similar to warfarin and its efficacy [5].

Trade Names. Pival, Pivalyn [sodium salt] (Motomoco).

5.2.3. Anticoagulant Resistance

The first substantiated case of warfarin resistance was recorded in Scotland in 1958 in Norway rats [40]. The phenomenon has now been widely recorded in Europe and North America in all three common commensal rodent species [5,20,23]. In the tropics, warfarin resistance has been found in *R. rattus* [41] and *R. tiomanicus* [42]. Warfarin-resistant rodent strains often possess cross resistance to many other first-generation compounds. More recently, resistance to difenacoum [43] and bromadiolone [44] has become evident in Europe.

Populations of Norway rats now exist in the United Kingdom and Denmark which are less susceptible to brodifacoum than normal. However, they remain sufficiently susceptible to be controlled with that compound [45]. The mechanism of anticoagulant resistance has been researched [46] but is still not completely understood.

5.3. Second-Generation Anticoagulants

The discovery of resistance to the first-generation anticoagulants threatened the gains made with the introduction of these highly successful compounds. In an attempt to overcome this problem, some chemists investigated alternative modes of action while others continued to work on the hydroxycoumarins. This latter line of research resulted in the discovery of a second generation of anticoagulant rodenticides [47]. These compounds are effective against rodents resistant to first-generation anticoagulants and are also more active against susceptible animals. Some are so active against most target species that, unlike the first-generation compounds, a lethal dose is ingested by rodents during a single episode of feeding on poisoned bait [29]. The second-generation anticoagulants retain all of the advantages of the earlier compounds. In particular, vitamin K_1 is antidotal, although the period required for its effective administration is prolonged [48]. This is because the second-generation compounds have long biological half-lifes [49–51]. The secondary hazard of these materials to predators has been reviewed [52].

Difenacoum [*56073-07-5*], 3-(3-biphenyl-4-yl-1,2,3,4-tetrahydro-1-naphthyl)-4-hydroxycoumarin, $C_{31}H_{24}O_3$, M_r 444.54, *mp* 215 – 217 °C. An off-white, odorless, crystalline solid synthesized by the condensation of 4-hydroxycoumarin and 3-biphenyl-4-yl-1,2,3,4-tetrahydro-1-naphthol; virtually insoluble in water; slightly soluble in benzene and alcohols; soluble in acetone, chloroform, and other chlorinated solvents [53]:

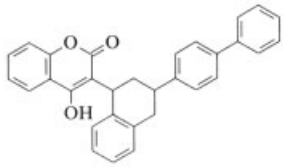

Appearing in 1974, difenacoum was the first of the new generation of anticoagulants to be commercialized for the control of rodents resistant to warfarin and related compounds [47]. LD_{50} values (acute oral) for warfarin-susceptible rodent strains are 1.8 mg/kg for Norway rats and 0.8 mg/kg for house mice. The equivalent figure for warfarin-resistant Norway rats is 3.4 mg/kg. There is also a degree of specificity. The compound is generally less toxic to nontarget animals than to targets (acute oral LD_{50} in mg/kg: pig > 50, dog 50, cat 100, chicken 50). A wide range of bait types containing 0.005 % difenacoum is available, including meals, broken and whole grains, pellets, and wax blocks. A 0.1 % contact dust is no longer widely used. Resistance to difenacoum was detected among a population of Norway rats in the United Kingdom in 1978 [54]. This remains a problem of local importance but relatively low resistance factors indicate that other behavioral factors also play a significant role [55]. Difenacoum resistance has also been recorded in other European countries [44].

Trade Names. Neosorexa (Sorex), Ratak (ICI).

Brodifacoum [*56073-10-0*], 3-[3-(4′-bromobiphenyl-4-yl)-1,2,3,4-tetrahydro-1-naphthyl]-4-hydroxycoumarin, $C_{31}H_{23}BrO_3$, M_r 523.44, *mp* 228 – 232 °C. An off-white/fawn, tasteless, and odorless powder; insoluble in water but soluble in certain organic solvents [53]:

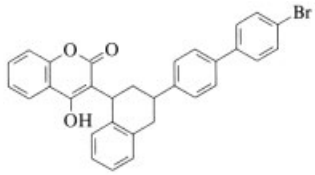

Brodifacoum is the most potent second-generation compound. Rodents succumb after consuming bait as part of their food intake on only one occasion [26]. The acute oral LD_{50} values for brodifacoum against warfarin-susceptible rodents are: Norway rats 0.26 mg/kg, house mice 0.4 mg/kg, roof rats 0.73 mg/kg. The compound is, however, equally effective when employed to control rodent infestations resistant to warfarin and other anticoagulant rodenticides [43]. No difference is discernible between the acute and the cumulative, chronic toxicities of the material. Brodifacoum has been widely evaluated, both in the laboratory and in the field, against commensal and agricultural rodent pests [56]. Commercially available pellet, wax block, and cereal baits generally contain 0.005 % brodifacoum.

Trade Names. Brodifacoum (Sorex); Havoc, Klerat, Matikus, Talon (ICI).

Bromadiolone [*28772-56-7*], 3-[3-(4′-bromobiphenyl-4-yl)-3-hydroxy-1-phenylpropyl]-4-hydroxycoumarin, $C_{30}H_{23}BrO_4$, M_r 527.42, *mp* 200 – 210 °C. A yellowish powder insoluble in water; soluble in ethanol, ethyl acetate, and dimethylformamate:

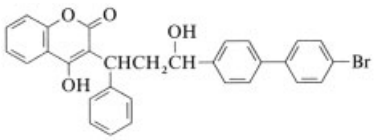

This compound was patented in 1968 [57] and introduced to the market in 1976 [5]. The acute oral LD_{50} for Norway rats is 1.125 mg/kg and for house mice is 1.75 mg/kg [58]. It is generally used in baits against these species at a concentration of 0.005 % [59,60]. It is sufficiently potent to kill warfarin-suceptible Norway rats after one day of feeding but repeated feeding is required to kill resistant rats and house mice [5, 61]. Bromadiolone has a good palatability [5]. It is widely available in a variety of formulations, including cereal-based baits, oil-based concentrates, and powder concentrates, containing 0.1 – 0.5 % of the active ingredient, and tracking dusts at 0.1 – 2.0 % strength [17].

Trade Names. Bromone, Super-Caid (Lipha); Contrac, Maki (LiphaTech).

Flocoumafen [*90035-08-8*], 4-hydroxy-3-[1,2,3,4-tetrahydro-3-[4-(4-trifluoromethylbenzyloxy)phenyl]-1-naphthyl]coumarin, $C_{33}H_{25}F_3O_4$, M_r 542.6, *mp* 181 – 191 °C (*cis*), 163 – 166 °C (*trans*). An off-white powder insoluble in water; soluble in acetone, alcohols, chloroform, and dichloromethane [62]:

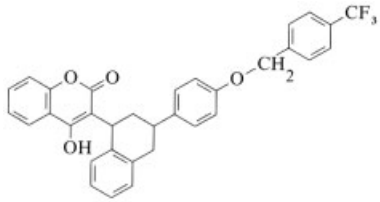

This compound was introduced in 1984 and is one of the most potent second-generation compounds. Acute oral LD_{50} (mg/kg) is 0.25 for Norway rats, 0.8 for house mice, and 0.18 for gerbils. It is less active against nontarget bird species (acute oral LD_{50} values: > 100 mg/kg for chicken, > 300 mg/kg for quail, and ca. 100 mg/kg for Mallard duck) but toxic to dogs (0.075 – 0.25 mg/kg) [63]. The main formulation available is a wax briquette containing 0.005 % flocoumafen; this is effective against rodent strains resistant to other anticoagulants and is used in a wide range of urban, industrial, and agricultural situations [64,65]. Pellet and grain baits are also produced.

Trade Name. Storm (Sorex, Shell).

Difethialone [*104653-34-1*], 3-[(1*RS*,3*RS*; 1*RS*,3*SR*)-3-(4′-bromobiphenyl-4-yl)-1,2,3,4-tetrahydro-1-naphthyl]-4-hydroxy-1-benzothiin-2-one, $C_{31}H_{23}BrO_2S$, M_r 539.2, *mp* 203 – 227 °C [66]:

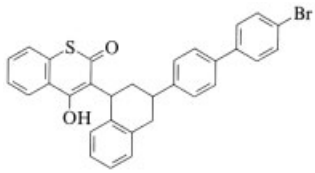

Its structure differs from that of brodifacoum in the substitution of sulfur for the oxygen atom

in the hydroxycoumarin ring. LD_{50} values for warfarin-susceptible animals are 0.51 mg/kg for *R. norvegicus* and 0.48 mg/kg for *M. musculus* [50]. Satisfactory results have been obtained in laboratory tests with 0.0025 % difethialone baits against susceptible and warfarin-resistant rats and mice, both in Denmark [50] and France [67]. Exposure to the compound at this concentration for one day did not, however, give complete control [67]. Wheat baits carrying 0.0025 % of the active ingredient were effective for the control of *Arvicola terrestris* and *Pitymys duodecimcostatus* in the field [68]. Field trials against rats and mice have also been conducted successfully in the United States [69].

Trade Name. Frappe (Lipha).

6. Other Chemical Methods of Rodent Control

Fumigants are employed for rodent control when conventional methods are ineffective or impractical and when the structures and installations to be treated, such as ships, aircraft, and grain silos, can be sealed effectively or enclosed in a gastight membrane. They are also used applied directly to rodent burrows.

One of the most common fumigants is *phosphine* [*7803-51-2*] (PH_3), which is evolved when aluminum phosphide (AlP) formulations (trade names: Phostoxin, Degesch; Delicia Gastoxin, Ernst Freyberg; Celphos, Excel) are exposed to atmospheric or soil moisture. *Hydrogen cyanide* [*74-90-8*] (HCN) is similarly used, particularly for rabbit control in the United Kingdom, and is applied as sodium cyanide (NaCN) powder (trade name: Cymag, ICI). *Methyl bromide* [*74-83-9*] (CH_3Br) and *chloropicrin* [*76-06-2*] (CCl_3NO_2) are used less frequently. Fumigants are available formulated as powders, impregnated cardboard disks, pellets, and tablets, and as gases in steel cylinders. Great care is required in their use and, in many countries, only professionals are permitted to apply them.

Chemosterilants. Alternatives to compounds that cause the death of rodents have been sought for reasons of humaneness and concern for the environment. Chemosterilants cause infertility in animals that ingest an effective dose. Several of these compounds have been studied for their potential as rodent control agents [70–72]. α- *Chlorohydrin* [*96-24-2*] (3-chloro-1,2-propanediol) is one of the few to have been commercialized [73]. It sterilizes male Norway rats at a dose of 90 – 100 mg/kg (acute oral LD_{50} is 150 – 160 mg/kg) and is available as a 1 % ready-to-use bait and as 20 % concentrate (trade name: Epibloc, Gametrics). Chemosterilants are not widely used in pest control because their effects are often transient and the presence of rodents, sterile or not, is considered undesirable.

Repellants. Many substances suppress rodent consumption of foods treated with them. For example, certain rodenticides are unpalatable to rodents when applied at concentrations higher than those normally used. It is surprising, therefore, that a fully reliable chemical rodent repellant has not been discovered. *Thiram*, *tetramethylthiuram disulfide* [*137-26-8*] [74], and *Rotran* [*3304-97-0*], a tertiary butylsulfenyldimethyldithiocarbamate [75], have been used as rodent repellants. A series of naturally-occurring phenolic compounds has promising properties [76].

A useful advance towards the increased safety of rodenticidal preparations is the introduction of the human taste deterrent *denatonium benzoate* [*3734-33-6*], *N*-[2-[(2,6-dimethylphenyl)amino]-2-oxoethyl]-*N,N*-diethylbenzenemethanaminium benzoate, $C_{28}H_{34}N_2O_3$, M_r 446.5, *mp* 163 – 170 °C. This compound is highly repellant to humans when included in rodenticidal baits at a strength of 0.001 % but does not deter their consumption by rodents [77]. Incorporation of this compound is unlikely to decrease the frequency of accidental exposure of humans to baits. It will, however, reduce the quantity of accidentally consumed bait and, thereby, the clinical severity of these incidents.

7. References

1. J. E. Brooks, *CRC Crit. Rev. Environ. Control* **3** (1973) 405 – 453.
2. I. A. E. Atkinson in P. J. Moors (ed.): *Conservation of Island Birds,* ICBP (1985) pp. 35 – 81.

3. N. G. Gratz in I. Prakash (ed.): *Rodent Pest Management,* CRC Press, Boca Raton, Fla., 1988, p. 101.
4. W. B. Jackson, *Bull. OEPP* **7** (1977) 439 – 458.
5. A. P. Meehan: *Rats and Mice, Their Biology and Control,* Rentokil, East Grinstead 1984.
6. M. Lund in I. Prakash (ed.): *Rodent Pest Management,* CRC Press, Boca Raton, Fla. 1988, p. 29.
7. M. Lund in I. Prakash (ed.): *Rodent Pest Management,* CRC Press, Boca Raton, Fla. 1988, p. 407.
8. C. V. Prescott, M. El-Amin, R. H. Smith in R. E. Marsh (ed.): *Proc. Vertebr. Pest Conf.,* in press.
9. J. Hone, H. Mulligan, *Science Bulletin 89,* Department of Agriculture New South Wales, Australia 1982.
10. M. Lund in I. Prakash (ed.): *Rodent Pest Management,* CRC Press, Boca Raton, Fla., 1988, p. 331.
11. K. Ogushi, I. Tokumitsu, *Eisei Dobutsu* **23** (1972) 71 –75.
12. Sorex Ltd., GB 1 371 135, 1973 (M. R. Hadler).
13. J. H. Greaves, R. Redfern, R. E. King, *J. Hyg., Camb.* **73** (1974) 341 – 351.
14. B. A. Dreikorn, G. O. P. O'Doherty, A. J. Clinton, K. E. Kramer, *Proc. Br. Crop Prot. Conf. Pests Dis.* 1979, 491 – 498.
15. S. R. Spaulding, R. B. L. van Lier, M. E. Tarrant, *Acta Zool. Fennica* **173** (1985) 171 – 172.
16. S. R. Spaulding, *BCPC Mono.* **37** (1987) 89 – 95.
17. *FAO Plant Production and Protection Paper 16,* Food and Agriculture Organization, Rome 1979.
18. K. P. Link, *Harvey Lect. Series* **39** (1944) 162 – 216.
19. J. A. O'Connor, *Res. London* **1** (1948) 334 – 336.
20. M. R. Hadler, A. P. Buckle in R. E. Marsh (ed.): *Proc. Vertebr. Pest Conf.* **15** (1992) 149 – 155.
21. Wisconsin Al. Res. Foundation, US 2 427 578, 1947 (M. A. Stahmann, M. Ikawa, K. P. Link).
22. M. Lund, *Annual Report 1982,* Danish Pest Infestation Laboratory, Lyngby 1983, pp. 89 – 100.
23. J. H. Greaves, *Acta Zool. Fennica* **173** (1985) 155 – 157.
24. M. J. Robinson, J. Pash, J. Grimwade, J. Campbell, *Med. J. Aust.* **1** (1978) 157.
25. Geigy AG, DT 884 500, 1950 (W. Stoll, F. Litvan).
26. *Pesticide Manual,* British Crop Protection Council, Brighton 1987, p. 1081.
27. M. von Reiff, R. Weismann, *Acta Trop.* **8** (1951) 97 –130.
28. Bayer, DT 1 079 382, 1960 (E. Enders, A. Muller).
29. A. C. Dubock, D. E. Kaukeinen in W. E. Howard, R. E. Marsh (eds.): *Proc. Vertebr. Pest Conf.* **8** (1978) 127 – 137.
30. F. P. Rowe, R. Redfern, *Ann. Appl. Biol.* **62** (1968) 355 – 361.
31. Upjohn, US 2 672 483, 1954 (D. G. Thomas).
32. J. T. Correll, L. L. Coleman, S. Long, R. S. Willy, *Proc. Soc. Exp. Biol. Med.* **80** (1952) 139 – 143.
33. E. W. Bentley, Y. Larthe, *J. Hyg, Comb.* **57** (1959) 135 – 149.
34. T. Kusano, *Eisei Dobutsu* **24** (1974) 207 – 213.
35. Lipha, US 3 153 612, 1964 (D. Molho, E. Boschetti).
36. M. Lund in I. Prakash (ed.): *Rodent Pest Management,* CRC Press, Boca Raton, Fla., 1988, p. 341.
37. Upjohn, DT 897 103, 1951 (D. G. Thomas).
38. H. Kabat, E. F. Stohlman, M. I. Smith, *J. Pharmacol. Exp. Ther.* **80** (1944) 160 – 170.
39. T. B. Gaines, *Toxicol. Appl. Pharmacol.* **2** (1960) 88 – 99.
40. C. M. Boyle, *Nature (London)* **188** (1960) 517.
41. Y. M. Lam, *MARDI Res. Bull.* **12** (1984) 373 – 379.
42. B. J. Wood, G. F. Chung in L. R. Davis, R. E. Marsh (eds.): *Proc. Vertebr. Pest Conf.* **14** (1990) 129 – 134.
43. J. H. Greaves, P. B. Cullen-Ayres in J. W. Suttie (ed.): *Proc. 17th Steenbock Symposium,* Elsevier, New York 1988, p. 387.
44. M. Lund in D. O. Clark (ed.): *Proc. Vertebr. Pest Conf.* **11** (1984) 89 – 94.
45. M. Lund, *Annual Report 1989,* Danish Pest Infestation Laboratory, Lyngby 1990, p. 77.
46. A. D. MacNicoll, *Pesticide Resistance: Strategies for Management,* National Academy Press, Washington, D. C. 1986, pp. 87 – 99.
47. M. R. Hadler, R. S. Shadbolt, *Nature (London)* **253** (1975) 275 – 277.
48. The Treatment of Anticoagulant Rodenticide Poisoning, Joint publication between: ICI, Lipha, Shell International Chemical Company Ltd., Sorex Ltd., Fernhurst 1988, p. 9.
49. G. Parmar, H. Bratt, R. Moore, P. L. Batten, *Hum. Toxicol.* **6** (1987) 431 – 432.

50. J. C. Lechevin, R. M. Poche in A. C. Crabb, R. E. Marsh (eds.): *Proc. Vertebr. Pest Conf.* **13** (1988) 59–63.
51. K. R. Huckle et al., *Pestic. Sci.* **25** (1989) 297–312.
52. D. E. Kaukeinen in R. E. Marsh (ed.): *Proc. Vertebr. Pest Conf.* **12** (1982) 151–158.
53. Ward Blenkinsop Co., DE-OS 2 424 806, 1973 (M. R. Hadler, R. S. Shadbolt).
54. J. H. Greaves, D. S. Shepherd, J. E. Gill, *Ann. Appl. Biol.* **100** (1982) 581–587.
55. R. J. Quy, D. S. Shepherd, I. R. Inglis, *Crop Prot.* **11** (1992) 14–20.
56. D. E. Kaukeinen, M. Rampaud in T. P. Salmon (ed.): *Proc. Vertebr. Pest Conf.* **12** (1986) 16–50.
57. Lipha, DE-OS 1 959 317, 1968 (O. Molho, E. Boschetti, L. Fontaine).
58. M. Grand, *Phytiatr. Phytopharm.* **25** (1976) 69–88.
59. F. P. Rowe, C. J. Plant, A. Bradfield, *J. Hyg.* **87** (1981) 171–177.
60. C. G. J. Richards, *J. Hyg.,* Camb. 86 (1981) 363–367.
61. R. Redfern, J. E. Gill, *J. Hyg., Camb.* **84** (1980) 263–268.
62. Shell, EP 0 098 629, 1984 (I. D. Entwistle, P. Boehm).
63. D. J. Bowler, I. D. Entwistle, A. J. Porter, *Proc. Br. Crop Prot. Conf. Pests Dis.* 1984, 397–404.
64. A. P. Buckle, *J. Hyg., Camb.* **96** (1985) 467–473.
65. R. A. Johnson, *Bull. OEPP* **18** (1988) 481–488.
66. Lipha, 572 279-B2, 1988 (J. J. Berthelon).
67. K. Nahas, G. Lorgue, M. Mazallon, *Ann. Rech. Vet.* **20** (1989) 159–164.
68. J. C. Lechevin, *Bull. OEPP* **18** (1988) 475–480.
69. E. F. Marshall in R. E. Marsh (ed.): *Proc. Vertebr. Pest Conf.* **15** (1992) 171–174.
70. A. M. Bowerman, J. E. Brooks, *J. Wildl. Manage.* **35** (1971) 618–624.
71. A. B. Lazarus, F. P. Rowe, *Agro-Ecosystems* **8** (1982) 59–67.
72. A. P. Meehan, M. C. Hum, *Int. Pest Control* **21** (1979) 39–41.
73. R. J. Ericsson in R. E. Marsh (ed.): *Proc. Vertebr. Pest Conf.* **10** (1982) 6–9.
74. Du Pont, US 1 972 961, 1931 (W. H. Tisdale, I. Williams).
75. Phillips Petroleum, US 2 621 143, 1952 (D. Goodhue, C. E. Tissol).
76. NRDC, GB 2 229 633, 1990 (M. F. Wilson, P. W. Greig-Smith).
77. D. E. Kaukeinen, A. P. Buckle in J. E. Borrecco, R. E. Marsh (eds.): *Proc. Vertebr. Pest Conf.* **15 (1992)** Newport Beach, CA, March 3–5, 1992, pp. 192–198.

Roll Bonding → Corrosion
Roofing Materials → Asphalt and Bitumen
Roof Tiles → Construction Ceramics
Rosin → Resins, Natural
Rosin Oil → Turpentines
Rotenoids → Insect Control

Rubber, 1. Survey

Heinz-Hermann Greve, Bayer AG, Leverkusen, Federal Republic of Germany(Chap. 1 - 4)

Desmond Threadingham, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 3 in part)

1. Definition 1
2. Rubber Elasticity 1
3. History 2
4. Nomenclature 3
5. References 4

1. Definition

The terms *rubber* and *elastomer* are often used incorrectly. DIN 53 501 and DIN 7 724 give exact definitions, differentiate between these two terms, and also define the term *vulcanization*. Vulcanization is also defined in ISO 1382 no. 1003. In these standards, high molar mass polymeric materials are classified according to the temperature dependence of their mechanical properties.

Rubber (raw rubber) is a non-cross-linked, but cross-linkable (vulcanizable) polymer with rubber-elastic properties at room temperature.

At higher temperature or under the influence of deforming forces, raw rubber exhibits increasing viscous flow, so that it can be molded under suitable conditions. Raw rubber is a starting material for the production of elastomers.

Elastomers are polymeric materials that are cross-linked (vulcanized) up to their decomposition points. They are hard and glassy at low temperature and do not exhibit viscous flow even at high temperature. They have rubber-elastic properties, particularly from room temperature up to their decomposition point.

Rubber-elastic behavior is characterized by a relatively low shear modulus with a rather slight temperature dependence.

Elastomers exhibit a glass transition temperature of $< 0\,°C$ in the torsional vibration test according to DIN 53 520. Their shear moduli lie in the range $0.1 - 1000\,N/mm^2$ and remain almost constant between ca. 20 °C and the decomposition temperature.

Vulcanization is a process in which rubber, through a change in its chemical structure (for example, cross-linking), is converted to a condition in which the elastic properties are conferred or re-established or improved or extended over a greater range of temperatures. In some cases, the process is carried to a point where the substances become rigid (ISO 1382 no. 1003).

2. Rubber Elasticity

The difference between the mechanical properties of elastomers and those of other materials, such as metals and thermoplastics, is the result of different energy storage mechanisms. Metals, for example, alter their energy content on deformation by changing the distances between atoms; thermoplastics do so by deformation of the valence angles of the molecules. Therefore, metals and thermoplastics cool on elongation, and the tension due to deformation is lowered on heating.

Rubber elasticity is caused by entropy changes. An elastomer that is cross-linked by chemical or van der Waals bonds is forced into a more highly ordered conformation under extension. This leads to a decrease in entropy. When the load is removed the polymer molecules return to their original position with an increase in entropy. This is the reason elastomers warm up when elongated and the tension increases on heating. Prerequisites for rubber-elastic properties are

1) Sufficient cross-linking (i.e., polymer chains may not slide away from each other completely or undergo viscous flow)
2) Sufficiently high mobility of the longer chain segments (this is the case only at 20 – 70 °C above the glass transition temperature)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_221

3) A highly amorphous state because crystallization limits chain mobility.

3. History

The old term *caoutchouc* was taken from the word *Caao-Chu* meaning "weeping tree." Many useful objects were made from the milk of the weeping tree by the indigenous people of the Amazon region long before white men discovered South America. When this tree, belonging to the spurge family, became known in Europe, botanists gave it the name *Hevea brasiliensis*. The milk produced is an aqueous emulsion or dispersion of oils, fats, waxes, resins, starch, proteins, and the main component, polyisoprene, which contains a high proportion of double bonds with the *cis* configuration. The milk occurs in more than a thousand species of plants, all of which are dicotyledonous angiosperms.

In 1495, Christopher Columbus, on his second journey to the New World, was the first white man to observe Indians in Haiti playing with an elastic ball. A quarter of a century later, the Spanish adventurer Fernando Cortez brought knowledge of the new, elastic, hitherto unknown material back to Europe.

More than 200 years later, in 1751, the people of Europe were reminded of the existence of natural rubber in an article by the French scientists C. M. de la Condamine and C. F. Fresneau. The first molded articles (rubber tubes and shoes) were produced in 1765, after P. I. Macquer and L. A. M. Hérrissant, succeeded in dissolving natural rubber and achieving the same properties, after evaporation of the solvent (ether and turpentine), that the rubber had before dissolution. In 1770 the Englishman J. Priestley reported the discovery of the erasing effect of natural rubber by the mechanic Edward Nairne. In 1826, M. Faraday discovered that the basic building block of natural rubber molecules was a hydrocarbon with five carbon atoms. This compound was later named isoprene. At this time the world rubber consumption was ca. 150 t. The discovery of hot vulcanization by Charles Goodyear in 1839 led the way to many new applications.

Practical uses were made possible only by the further development of vulcanization by Thomas Hancock in 1843, after which rapid development ensued. In 1845, R. W. Thomson applied for a patent for the first air-filled tires at the London patent office. By 1856, worldwide consumption of rubber had reached 7000 t. Because of the sharply increasing need for rubber arising from the growing number of applications and the Brazilian rubber monopoly, rubber was for a time more expensive than silver.

In 1873 the Englishman Cross succeeded in a first attempt to smuggle 3000 caoutchouc seeds and plants out of Brazil. Despite the most careful treatment, the seeds did not germinate. Three years later, Henry Wickham brought 70 000 seeds to England legally, and these germinated in Kew Gardens [14].

From the original 70 000 seeds, 2600 rubber seedlings were shipped to India, Ceylon, Singapore, and Malaysia. Of these, 1800 grew and ca. 5 – 7 became ready to tap. The latter were the ancestors of all the rubber trees in Southeast Asia.

In 1887 the Frenchman G. Bouchardat was the first to succeed in producing *synthetic rubber* by polymerization of isoprene obtained by dry distillation of rubber. The polymerization was performed over several months. In the years that followed , rapid development in tire production occurred after the manufacture of the first tires by Macintosh, Dunlop, Michelin, and Goodrich.

In 1900 the Russian chemist J. Kondakov succeeded in polymerizing the more readily accessible dimethylbutadiene by lengthy heat treatment to give methyl rubber. This was the first completely synthetic rubber with a composition that differed from natural rubber.

The first usable synthetic rubber (polyisoprene) was produced by the German chemist Fritz Hoffmann in 1909. In 1910 the German researcher C. D. Harries and, independently, the English chemists F. E. Matthews and E. Strange created the basis for large-scale industrial synthetic rubber production through the discovery that the polymerization of dienes can be initiated by alkali, particularly sodium. Because of the increasing use of automobiles at the beginning of the 20th century the price of rubber rose to 14 €/kg. At this time, world rubber consumption was 100 000 t.

The first completely synthetic isoprene rubber tires were produced at the Farbenfabriken

Bayer. With these tires, CARL DUISBERG drove from Leverkusen to Freiburg im Breisgau without a puncture. Impressed by this performance, Kaiser Wilhelm II had all his cars fitted with these tires.

The first industrial production of methyl rubber took place at Bayer, Leverkusen, between 1915 and 1918 because Germany was cut off from natural rubber supplies during World War I. In total, 2500 t was produced. At the beginning of the 1920s a synthetic rubber with the name Buna (butadiene with natrium, i.e., sodium catalysis) was produced. This name is still used today for many types of synthetic rubber.

In 1929, W. BOCK developed the basis for the general economic exploitation of synthetic rubber – the emulsion polymerization of butadiene and styrene. Styrene – butadiene rubber (SBR) is still the most widely used synthetic rubber.

At the beginning of the 1930s the first oil resistant rubber, nitrile rubber (NBR) was discovered by TSCHUNKUR, shortly afterwards, W. H. CAROTHERS developed polychloroprene, which is resistant to ozone and weathering.

Other important dates in the history of rubber are

1937	polyurethane rubber developed by OTTO BAYER
1942	silicone rubber
1943	butyl rubber
1948	fluororubber
1949	acrylate rubber
1953	chlorosulfonated polyethylene
1959	ethylene – propylene rubber developed by G. NATTA and G. CRESBY, based on Ziegler – Natta catalysts
1961	ethylene – vinyl acetate rubber
1962	styrene – butadiene – styrene (SBS) block copolymers, which led to the first thermoplastic elastomers
1965	epichlorohydrin rubber
1970	thermoplastic polyurethane rubber
1973	thermoplastic copolyesters
1975	polynorbornene rubber and hydrogenated nitrile rubber
1978	ethylene – acrylate rubber
1979	thermoplastic elastomers based on cross-linked ethylene – propylene rubbers
1981	polytetrafluoroethylene – polyolefin blends
1985	amorphous thermoplastic chloropolyolefin elastomers and introduction of a very highly resistant fluoroalkoxyphosphazene rubber
1989	ethylene – chlorosulfonated – polyethylene copolymers

In 1996 worldwide consumption of rubber was 15.6×10^6 t, of which 63 % was synthetic rubber. Until 2000 an average annual growth rate of the total rubber market of 2.7 % is expected.

The development of ever more specialized types of rubber, principally since 1965, shows that products require custom tailoring of properties for certain areas of application. Otherwise the ever higher requirements of vulcanized products could no longer be fulfilled.

With the continued development of catalyst systems and process technology new synthetic rubbers have been added so that today there are more than 25 different types of synthetic rubber. Most of these, like natural rubber, are hydrocarbons but with varying degrees of saturation of the carbon polymer chain and differing polarities as a result of the substituent side groups. It is these factors which are largely responsible for resistance to the destructive effects of outside influences such as heat, oxygen, ozone and chemicals and to swelling caused by mineral oils and fuels. Exceptions are epichlorohydrin rubber with carbon and oxygen in the polymer chain, polysulfides based upon carbon and sulfur and inorganic rubbers such as silicone with a silicon – oxygen polymer backbone or polyphosphazene with phosphorus and nitrogen in the main chain. These latter two materials are characterized by exceptional resistance to heat.

Each type of synthetic rubber exists as a family of products with a range of grades to suit different processing and end use requirements. The *Synthetic Rubber Manual* of the International Institute of Synthetic Rubber Producers now lists several thousand products from manufacturers around the globe.

Despite all the progress in the development of new synthetic rubbers, the importance of natural rubber has not diminished. In terms of quantity produced it is still the most important rubber, and in its mechanical properties it has so far not been surpassed by any synthetic rubber.

4. Nomenclature

Only in a very few cases are various types of rubber known by their IUPAC names. The ASTM D 1418-90 names and abbreviations are the most widely used. The most important abbreviations are given in Rubber, 3. Synthetic, Chap. 1.2.

5. References

General References

1. P. W. Allen: *Natural Rubber and the Synthetics,* Crosby Lockwood, London 1972, p. 255.
2. J. Le Bras: *Grundlagen der Wissenschaft und Technologie des Kautschuks,* Berliner Union, Stuttgart 1956.
3. A. S. Craig: *Rubber Technology,* Oliver & Boyd, Edinburgh 1963, p. 222.
4. Synthetischer Kautschuk, Darstellung einer Industrie, Internat. Inst. of Synthetic Rubber Producers, New York 1973, p. 96.
5. W. Hofmann: "Kautschuk Technologie," Habilitationsschrift am Institut für Kunststoffverarbeitung an der Rheinisch-Westf. Techn. Hochschule Aachen, Selbstverlag, Aachen 1975.
6. W. Kleemann: *Einführung in die Rezepturentwicklung der Gummi-Industrie,* 2nd ed., Deutscher Verlag f. Grundstoff-Ind., Leipzig 1966, p. 630.
7. P. Kluckow, F. Zeplichal: *Chemie und Technologie der Elastomere,* 3rd ed., Berliner Union, Stuttgart 1970, p. 593.
8. *Winnacker-Küchler,* 3rd ed., **5,** 142 – 251.
9. G. Kraus: *Reinforcement of Elastomers,* Interscience, New York 1965, p. 611.
10. *Houben-Weyl,* 4th ed., **E 20/1.**
11. M. Morton: *Rubber Technology,* 3rd ed., Van Nostrand Reinhold, New York 1987.
12. P. Schidrowitz, T. R. Dawson: *History of the Rubber Industry,* Heffer & Sons, Cambridge 1952.
13. H. J. Stern: *Rubber-Natural and Synthetic,* 2nd ed., MacLaren, London – Elsevier, Amsterdam 1967, p. 519.
14. J. Schnetger: *Lexikon der Kautschuktechnik,* Hüthig Buch Verlag, Heidelberg 1992.
15. S. Boström (ed.): *Kautschuk-Handbuch,* **vol. 1 – 5,** Berliner Union, Stuttgart 1959 – 1962.
16. Bayer AG (ed.): *Kautschuk-Handbuch,* Berliner Union – Kohlhammer, Stuttgart 1971.
17. G. M. Blow (ed.): *Rubber Technology and Manufacture,* Butterworth, London 1971.
18. WDK Report: *Vom Rohstoff Kautschuk zum Werkstoff Gummi, Zeittafel der Kautschukgeschichte,* Frankfurt am Main, 1990.
19. K. Dinges in H. Batzer (ed.): *Kautschuk und Gummi in: Polymere Werkstoffe,* **vol. 3,** Thieme Verlag,Stuttgart 1984.
20. U. Eisele: "Spezifische Merkmale von Gummi im Vergleich zu anderen Werkstoffen," *Kautschuk, Gummi, Kunstst.,* Sonderdruck Zelle, anläßlich des 60jährigen Bestehens der DKG, Sept., 24 – 26, 1986.
21. U. Eisele: Skript zur Vorlesung *Polymerphysik,* Universität Hannover 1990.
22. H. Seiffert: "Der Weltkautschukmarkt an der Jahreswende," *Gummi, Fasern, Kunstst.* **1** (1992) 6.
23. "Elastomerics 1992 Review and Forecast: An Industry with now here to go but up?" *Elastomerics* **2** (1992) 11.
24. D. Kranz, K. Dinges: Skript zur Vorlesung *Latextechnologie,* Universität Hannover 1990.

Rubber, 2. Natural

HEINZ-HERMANN GREVE, Bayer AG, Leverkusen, Federal Republic of Germany

1. **Introduction** 1
2. **Rubber Extraction** 3
3. **Composition of Natural Rubber Latex** 3
4. **Biosynthesis of Natural Rubber** . . 4
5. **Commercial Extraction of Natural Latex** 4
6. **Production of Natural Rubber** . . . 5
6.1. **Extraction by Evaporation of Water (Evaporation, Spray Drying)** 5
6.2. **Coagulation** 5
7. **Classification** 7
7.1. **Technically Classified (TC) Rubber** 7
7.2. **Standard Malaysian Rubber (SMR)** 7
7.3. **Standardized Indonesian Rubber (SIR)** 7
8. **Physical and Technological Properties of Solid Rubber** 7
9. **Uses** 8
10. **Modification** 8
10.1. **Hydrogenated Natural Rubber** . . . 8
10.2. **Chlorinated Natural Rubber** 9
10.3. **Hydrohalogenated Natural Rubber** 9
10.4. **Cyclized Natural Rubber** 9
10.5. **Resin-Modified Natural Rubber** . . 10
10.6. **Poly(Methyl Methacrylate)-Grafted Natural Rubber** 10
10.7. **Superior-Processing Natural Rubber** 10
10.8. ***N*-Phenylcarbamoylazoformate-Modified Natural Rubber** 11
10.9. **Polystyrene-Grafted Natural Rubber** 11
10.10. **Epoxidized Natural Rubber (ENR)** 11
10.11. **Degraded Natural Rubber** 12
10.12. **Thermoplastic Natural Rubber** . . 12
11. **Compounding** 12
12. **Summary** 12
13. **References** 12

1. Introduction

Many plants contain a milky juice – an aqueous emulsion or dispersion of oils, fats, waxes, resins, starch, and proteins. More than 1000 plant species contain aqueous dispersions of rubber – or, more precisely, of 1,4-polyisoprene with a high content of *cis* double bonds.

Rubber occurs only in seed-bearing plants (spermatophytes),which are also dicotyledonous angiosperms. Within one genus the phylogenetically younger forms always produce more rubber than the older ones. Thus, rubber formation appears to be the result of more recent evolution.

With regard to economic importance, rubber plants can be divided into three groups:

1) Plants from which rubber is currently extracted: *Hevea brasiliensis*, which provides more than 99 % of the rubber from natural sources.
2) Plants that were used in the past or are being investigated for future use: about 50 species, in particular, *Ficus elastica* (rubber plant), *Funtumia castilla*, kok-saghyz (*Taraxacum),* guayule (*Parthenium argentatum*), cassava, and tapioca bush. The guayule shrub is in a special situation. Its economic exploitation is constantly being reinvestigated in the United States and Mexico.
3) Rubber plants whose exploitation has not yet been investigated: ca. 1000 species.

All trees and bushes have a layered structure. From inside out are the heartwood, the sapwood, the cambium, the phloem, and the bark. In the cambium, which consists of only one layer of cells, new cells are continuously formed during the growth period. Those to the inside become wood and those to the outside become phloem. Latex is produced between the cambium and the phloem. Latex can be formed in three types of location: milk vessels, milk tubes, or milk cells.

Milk vessels consist of an associated system of tubes between the bark and the cambium, to which new cells are being joined continuously. However, the whole plant does not have a single

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_225

milk vessel system throughout, but rather many individual ones. Each system extends from the roots via the trunk to the twigs but is not connected to the other systems. Examples of plants with milk vessel systems are *Hevea* and cassava.

Milk tubes consist of a branched tube system permeating the entire plant. They are found, for example, in poinsettias or crown-of-thorns (*Euphorbia spendens*).

Milk cells are individual, isolated, elongated cells. Latex exudes only from cut cells, and tapping is not possible. Such cells are present in lettuce, chamomile, and salsify.

Hevea brasiliensis. The habitat of the rubber tree is the Amazon rain forest. Today, plantations of *Hevea brasiliensis* are also found in Southeast Asia, principally in Thailand, Indonesia, and Malaysia. When growing in the wild these trees live for more than 100 years and grow to more than 40 m high; on plantations their height is kept to less than 20 m. *Hevea brasiliensis* requires temperatures of 25 – 30 °C, at least 2000 mm of rainfall per year, and high atmospheric humidity. For this reason it grows only between 15° North and 15° South.

More than 99 % of natural rubber currently comes from *Hevea brasiliensis*. After an initial growth period of 5 – 7 years, a tree gives latex for ca. 25 – 30 years. *Hevea brasiliensis* is very suitable for rubber extraction for the following reasons:

1) It can be tapped because of its milk vessels.
2) It does not form resins on tapping.
3) Polyisoprene with almost 100 % *cis* double bonds is produced.
4) It tolerates tapping over a period of many years very well.

In the second half of the 19th century the conditions for a rapid expansion of the rubber industry were created by the discovery of vulcanization. The development was impeded only by the difficulties in extracting natural rubber in sufficient quantities. The only source was Brazil, where wild rubber was extracted by inadequate and dangerous methods. In tapping expeditions to the often widely scattered trees, deep cuts were made in the trees with axes, which destroyed the milk tube systems. This careless harvesting resulted in rapid exhaustion of the sources, so the tappers had to penetrate ever further into the jungle, with the increasing danger of disease and attacks by natives. For these reasons the price of rubber was higher than the price of silver for a time.

On the initiative of Sir Clements Markhem the India Office and the London Bontanical Garden decided to cultivate *Hevea brasiliensis*. They commissioned Henry Wickam to procure the seeds of the plant. In 1874 he collected 70 000 seeds and shipped them to England for a fee of £70. Contrary to all claims, the export of *Hevea* seeds was legal. However, a problem arose because Brazilian customs operated very slowly and *Hevea* seeds were capable of germinating only for a relatively short period and therefore had to be brought quickly to Kew Gardens near London. Wickam solved the problem with the aid of the British consul in Para and several £1 notes. Of the 70 000 seeds, about 2600 germinated and about 1800 grew. Most of them were shipped to India and Ceylon, where they perished because of the unsuitable climate. A few plants (50 in 1876 and 22 in 1877) were taken to Malaysia and Singapore, and of those, about 5 – 7 could later be tapped. They formed the basis for the current world natural rubber production.

Today's *Hevea* plantations are predominantly in Southeast Asia: Malaysia (24 % of world production), Indonesia (24 %), and Thailand (25 %); some can also be found in Africa (6 %). In Brazil, rubber trees could not be planted for a long time because of a widespread fungal disease *Dothidella ulei*. In Africa and Asia this disease has thus far not appeared because the climate is slightly different from that in Brazil. An attempt is now being made to solve this problem by cross-breeding with resistant *Hevea* species, so that *Hevea brasiliensis* can again be planted in Brazil.

2. Rubber Extraction

Latex is extracted from *Hevea brasiliensis* by tapping, which involves making a downward-slanting cut in the bark of the tree with a knife around half the circumference of the trunk at an angle of 30 – 35°. The inner cambium must not be damaged in the process because, otherwise, irregular reformation of the bark would lead to difficulties with the next tapping cut. In 2 – 5 h,

ca. 50 cm^3 of latex exudes from the cut (ca. 15 g of rubber). It is collected in a vessel fixed at the lower end of the cut. Later the cut is closed by the drying latex. After two days, this rubber film is removed and tapping is continued.

Although this is the most common tapping method, others also are used (e.g., instead of being cut, the bark is pierced with sharp needles).

Ca. 2 kg of rubber is extracted per tree per year. This corresponds to 1000 kg per hectare per year at 500 trees per hectare. Originally, the annual yield of a plantation was 350 kg per hectare; today, it is 600 – 700 kg per hectare per year, if the entire planted area is considered, including trees that are not ready for tapping, those that are too old, and newly planted areas. Targeted cultivation has enabled average annual yields of more than 1500 kg per hectare to be achieved on large plantations, particularly in Malaysia. On experimental plantations, yields have even been increased to 2500 kg per hectare.

Ca. 80 % of the production originates from smallholdings. The typical smallholder, with his family, works 1.25 hectares with about 625 *Hevea* trees. With 288 harvesting days, he has to tap 90 000 times per year in a 2-day cycle. He can harvest 625 kg of natural rubber per year from the lower-yield trees (500 kg per hectare). At \$ 0.27 – 0.33/kg, the smallholder earns \$ 209 per year. From this he has to feed, on average, 3.2 people, and he is at the minimum subsistence level for a family of three.

In summarizing the problems of natural rubber extraction from *Hevea* (suitable climate; large area required; abundant, cheap work force needed), it becomes clear that natural rubber plantations are uneconomical in highly developed countries. In the United States and Mexico, attempts are being being made to counteract these problems by extraction of natural rubber from guayule. In guayule the rubber latex occurs in individual milk cells and therefore cannot be obtained by tapping. In the autumn harvest, shrubs are pulled up by machine and ground; the rubber is extracted with solvents and separated from resins. The rubber is then processed further in a process that can be automated extensively. Guayule grows in arid deserts that could otherwise not be used, so the large area required does not play an inhibitory role.

3. Composition of Natural Rubber Latex

Natural rubber latex is essentially a dispersion of *cis*-1,4-polyisoprene in water. The average particle size is between 0.15 and 3.0 µm. The particle-size distribution is very broad. The aqueous dispersion contains between 30 and 38 % solid material, depending on the time of the year and the age of the tree. Other components of the latex are 1 – 2 % proteins and phosphoproteins, 2 % resins, 1 % fatty acids, 1 % carbohydrates, and ca. 0.5 % inorganic salts. The rubber particles are surrounded by protein anions and are thus effectively negatively charged, which hinders coagulation of the latex. These proteins are decomposed rapidly by bacteria and enzymes when exposed to air, and the rubber then partially coagulates. In addition, in the presence of air, cross-linking of the rubber occurs within the latex particles, with gel formation and subsequent degradation of the polymer chains.

4. Biosynthesis of Natural Rubber [16–18]

In the plant, natural rubber is not formed by polymerization of unsaturated monomers (e.g., isoprene); instead, a polycondensation occurs, starting from a basic unit of five carbon atoms, isopentenyl pyrophosphate.

If the formation of the basic building block is considered in a highly simplified form (see Fig. 1), a molecule with six carbon atoms (β-hydroxy-β-methylglutaric acid) is built up from three acetyl groups. β-Hydroxy-β-methylglutaric acid is then reduced to mevalonic acid. The latter eliminates CO_2 and water to give isopentenyl alcohol, or more precisely its ester with pyrophosphoric acid. In the second stage of polycondensation, isopentenyl pyrophosphate is condensed with one molecule of dimethylallyl pyrophosphate, which is formed from isopentenyl pyrophosphate by isomerization. In this way, geranyl pyrophosphate is produced, which reacts with another molecule of isopentenyl pyrophosphate to give farnesyl pyrophosphate. The condensation is repeated until the final molecular mass of natural rubber is reached. The entire process is controlled enzymatically.

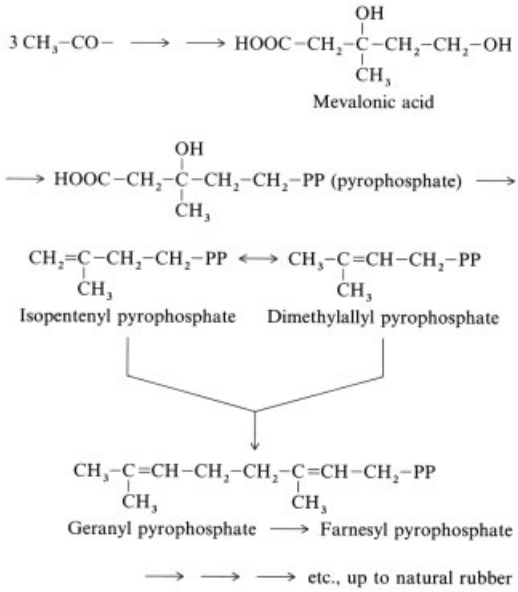

Figure 1. Simplified scheme showing synthesis of natural rubber.

Why the plant forms latex is thus far not clear: the milky juice or the rubber precipitated from it does not appear to have any protective function in the plant. Two theories are currently being discussed:

1) The rubber constitutes an energy reserve that the plant can degrade when energy is needed.
2) The rubber is a product of excretion, with which the plant can render toxic metabolites harmless.

Both theories have been supported by a number of investigations, but refuted by an equal number. Possibly, no single reason exists for latex formation.

5. Commercial Extraction of Natural Latex [19–24]

When latex was first extracted, the problem was to stabilize it toward enzymatic and oxidative attack by bacteria. Today, an aqueous ammoniacal solution (ca. 0.7 %) is generally used to counteract degradation. In some cases, ammonia is used in combination with phenol or alkali phenolates.

Natural latex is concentrated to lower the transportation costs. Three processes are essentially available for this:

1) Centrifugation
2) Creaming
3) Evaporation

Electrodecantation is also sometimes used.

Centrifugation of the fresh latex allows extensive separation of the lighter disperse rubber phase from the aqueous serum. About 80 % of the rubber phase remains in the concentrate, and the solid content is up to 60 wt %. Most of the other latex components and rubber particles with smaller diameters remain in the serum. Thus the concentrated latex has a higher average particle size than the original latex. The process is continuous and therefore very economical, and is the most commonly used method of concentration.

Creaming Method. In the creaming method, 0.25 wt % ammonium alginate and 0.5 wt % ammonium oleate are added to the latex. In the course of several days the latex separates into two layers, of which the upper has a rubber content of ca. 60 %. After separation of the serum and renewed creaming, a concentration of 65 – 68 % solid materials is achieved.

Evaporation. Latices produced by evaporation have a very high solid material content (up to 75 wt %) because the very small particles are still present, unlike centrifuged latex. They also differ from the latex obtained by the other processes in that they contain all the water-soluble solids. Evaporation is the most expensive and least used process.

Uses. When the transportation problem was solved by stabilization, natural latex could be used worldwide and became more and more competitive with synthetically produced latex. The areas of use of natural rubber latex are those in which high toughness, elongation, and resistance to tear propagation in the vulcanizate and the raw material are required. Natural latex is as yet unsurpassed in the sum of these properties. The main areas of use are in diving gear, rubber thread, adhesives, and foam linings for carpets.

6. Production of Natural Rubber

Most of the natural rubber latex is further processed to solid rubber; in this case, no concentration of the latex is necessary. The rubber can be obtained by evaporation, spray drying, or the most important process, coagulation and subsequent drying.

6.1. Extraction by Evaporation of Water (Evaporation, Spray Drying)

The tappers in the Brazilian jungle extract the rubber from the milky juice. The latex is poured onto an oarlike piece of wood covered with clay and dried by rotating the wood over a smoky fire. Addition of urikuri nuts to the firewood gives the smoke a high phenolic content. The process is stopped when a large lump has been produced. This highly smoked wild rubber is called hard-cure fine para. It is qualitatively a very high-value natural rubber, because in this processing method the most valuable parts of the latex solidify while those of low value drop off. The preserving effect of the phenol- containing smoke is so good that the rubber lumps can be stored for years without mold formation.

In *spray drying* the latex is poured onto a rapidly rotating disk atomizer. From this it falls in finely divided form into a drying tower, into which hot air is blown in countercurrent. The dry material is obtained in the form of small flakes that are later pressed together to form blocks. This type of rubber is called sprayed rubber. It contains all the solid components of the natural rubber latex and has very good mechanical properties. Unlike coagulated rubber it is very tough and difficult to process.

6.2. Coagulation

On a plantation the rubber is separated from the latex by coagulation. Acids, such as formic or acetic acid, or salts, such as sodium silicofluoride, are used as precipitating agents. The latex is first diluted to a solid material content of 15 – 20 % and coagulated at pH 5.1 – 4.8 (isoelectric point). The coagulate must be processed further immediately; otherwise, it will be attacked by bacteria. Most of the water-soluble substances in the natural rubber latex remain in the serum.

Two main processes are used to *stabilize* the raw rubber: the smoked-sheet and the crêpe rubber processes.

Smoked sheets still contain a large proportion of other substances present in the latex. In the production of smoked sheets the diluted latex is treated with 1 part of 0.5 % formic acid to 10 parts of latex in long, rectangular vats with constant stirring. After skimming, aluminum plates are placed upright in the vats about 4 cm apart. The next day the coagulated rubber formed between the plates is removed. These soft, completely homogeneous rubber plates are rolled out with a hand roller and processed further on a sheet mangle. The latter consists of five pairs of press rollers arranged in series.

The resulting sheets of 3 – 4-mm thickness are hung in a smokehouse and treated with phenol- containing smoke, produced by burning fresh wood and nutshells, to protect the rubber against oxidation and mold formation. After the smoking process the temperature is raised to 60 °C and the sheets are dried for two to three days.

The standard grades of *ribbed smoked sheets* are rated in terms or purity and quality from 1 (very good) to 5 (poor quality). In addition, an unsmoked grade *air-dried sheets*, is an odorless, top-quality, light-brown product of high purity, whose vulcanizate has very good mechanical properties. Air-dried sheets are produced by grinding the moist coagulate, extruding it, and after renewed grinding, washing it in washing mills. The rubber is then air-dried and molded into plates several millimeters thick.

Crêpe Grades. The crêpe grades are freed extensively from nonrubber components by washing in washing mills. The quantity of coagulating agent is calculated so that, after a few hours, the 15 – 20 % latex yields a soft, cohesive coagulate from which a clear serum gradually separates. To remove the acidic residues and most of the nonrubber components the coagulate is washed under friction in ribbed mills with powerful water scrubbing. The process is repeated with increasingly closely spaced rollers having increasingly fine ribbing, until thin crêpe-like sheets are obtained. These are dried either

for 10 to 12 days in air or with a vacuum dryer at 70 °C for 2 h. With the vacuum dryer the product must be cooled very quickly to avoid surface oxidation. The crêpe grades all have a brownish color and are stabilized with 0.5–0.75 wt % sodium hydrogensulfide before coagulation. Here, the grades also differ according to color, purity, and strength of the sheet.

The *secondary grades* play an additional, economically important role. These generally consist of rubber from small plantations, which has been processed in very different ways and from which a crêpe is produced by mixing and milling them together. This crêpe is sold as *remilled crêpe*. *Flat bark crêpe* is produced from the dried rubber films that form where the tree has been tapped and from waste material.

Special Grades. Two special grades exist: *initial concentration rubber* [25], which is coagulated from undiluted latex, and *Hevea crumb* [26]. The latter is a crumblike rubber that gives a moist coagulate on addition of an incompatible oil, such as 0.7 % castor oil. The coagulate is crumbled by passage through conventional crêpe rollers. Approximately 0.4 % of the castor oil remains in the rubber, but it does not affect the properties significantly. Because of the large surface area of the crumb the rubber can be purified readily by washing with water and is dried in a deep-bed dryer with air circulation at 80–100 °C in 3 to 4 h. The *Hevea* crumb is molded into compact lumps and classified according to the standard malaysian rubber scheme (see Section 7.2).

Superior-Processing Rubber. Another form of natural rubber is superior-processing rubber (SP rubber, see Section 10.7) [27–34]. It is produced by coagulation of a mixture of normal and vulcanized latex. The advantages of SP rubber are better processing properties, particularly in injection molding and calendering.

There is also *premasticated natural rubber*, whose molecular mass is degraded significantly through the addition of chemicals, which is very easy to process.

7. Classification

Even with strict working instructions for coagulation and processing, the standard grades, such as crêpe and smoked-sheet rubber, produced from plantations differ in their processibility and rate of vulcanization. Constant quality control is therefore necessary, and guidelines for the classification of natural rubber were worked out at the suggestion of the French Rubber Institute in Indochina in 1949.

Natural rubber that has been classified according to these guidelines is known as technically classified (TC) rubber. This classification was further improved by Malaysia, and the so-called standard malaysian rubber (SMR) scheme was introduced.

7.1. Technically Classified (TC) Rubber [35]

Natural rubber is classified according to its Mooney viscosity and the tensile modulus of a standard vulcanizate. A blue circle on a lump means, for example, that the natural rubber shows average plasticity during processing and has a high vulcanization rate.

Since 1953 the Mooney viscosity has no longer been quoted by some plantations, because it gives no indication of the processing behavior of the natural rubber. However, the tensile strength of test vulcanizates has proved to be important for calibration of the vulcameter (→ Rubber, 6. Testing).

7.2. Standard Malaysian Rubber (SMR) [36–42]

The quality features of SMR are laid down in a technical specification. They are

1) Dirt content
2) Ash content
3) Copper, manganese, and nitrogen content
4) Content of volatile components, and
5) Plasticity retention index (PRI)

In determination of the *PRI*, rubber is heated to 140 °C for 30 min and the plasticity is measured before and after. The ratio of the former to

the latter number multiplied by 100 is the PRI. The PRI characterizes the degradation behavior of natural rubber. It gives an indication of the expected aging properties and of the mixing viscosity, which is in turn related to tensile strength, dynamic properties such as rebound resilience, and hysteresis. At the end of 1970 the so- called MOD value was introduced, which is also a measure of vulcanization behavior.

There are also constant-viscosity (CV) and low-viscosity (LV) grades. *Constant viscosity* is achieved by saturating the aldehyde groups, which are formed in the oxidation of natural rubber, with hydroxylamine, thus hindering hardening. Naphthenic oil is added to the *low-viscosity* grades to decrease the viscosity.

Another grade is *dynat rubber.* This is an SMR grade processed by air drying.

Since its introduction in 1965, much success has been achieved in the standardization of natural rubber by use of the classification scheme. The proportion of natural rubber classified according to the Malaysian scheme was more than 40 % in 1990.

7.3. Standardized Indonesian Rubber (SIR)

The SIR classification is not widely used and corresponds generally to the SMR classification.

8. Physical and Technological Properties of Solid Rubber

At 20 °C the density of natural rubber is 0.906 – 0.916 g/cm^3, the specific heat is 1.905 kJ kg^{-1} K^{-1}, and the refractive index ca. 1.5191. The heat of combustion is − 45.2 kJ/kg. The UV absorption of thin rubber films begins at 310 nm; complete absorption occurs below 225 nm.

Electrical properties are affected mainly by water-soluble nonrubber components. The electrical conductance is 2×10^{-15} to 1×10^{-13} S/m.

Natural rubber is generally processed after mastication, which gives it excellent processing properties. These include rapid sheet formation in open roll mills, high toughness of the unvulcanized mixture (green strength), and high building tack. The tack and green strength are very important properties for the production of composites, in which various layers must be welded together as in tires or pneumatic springs. Natural rubber also exhibits excellent behavior on extrusion and calendering. Another positive feature is the high vulcanization rate. On *cooling and subsequent tempering* at ca. 10 to −35 °C, natural rubber becomes opaque and inelastic. This effect is caused by partial crystallization.

On *elongation* to more than 100 % the forced alignment of the chains also induces crystallization (elongation crystallization), which can be detected by the appearance of diffraction patterns in X-ray diagrams. The elongation crystallization leads to a strengthening effect and a higher tensile strength in the direction of elongation (anisotropy of mechanical properties). The result is a high green strength of unvulcanized and vulcanized mixtures. Crystallites formed at greater elongation do not melt when the load is removed. Thus they hinder, for example, propagation of an edge tear. In an intermittent stress – strain experiment the elongation crystallization is responsible for the fact that the loading and unloading curves show sharp differences (i.e., there is high hysteresis).

At low deformation (< 10 %), carbon black filled natural rubber exhibits a small loss factor (tan δ) because of the homogeneous carbon black distribution and the uniform network, so that hysteresis losses are low. This means that on small deformations the vulcanizates undergo low heat buildup.

Raw rubber dissolves (or at least swells very strongly) in many organic liquids such as benzene, petroleum ether, crude petroleum, and carbon tetrachloride. In contrast, vulcanized natural rubber can only swell because the chemical cross-linking prevents dissolution.

Compared with synthetic rubber, natural rubber has the following advantages: high structural stability with high elasticity, very good cold flexibility, excellent dynamic properties, and acceptable abrasive wear. Because of this combination of properties, it is superior to synthetic rubber in many applications. Its disadvantages include low oil resistance and unsatisfactory aging properties. In many cases, however, the latter can be improved significantly by the addition of stabilizers.

9. Uses

The main area of use of natural rubber is tires, particularly for trucks, where its property profile makes it much more suitable than synthetic rubber. Because of the low heat conductance, heat buildup is important in large tires. When the elasticity of rubber is too low and the hysteresis too high, too much heat is evolved and heat buildup occurs. This leads to inner burning, which causes destruction of the tire.

Because of its high elasticity associated with low hysteresis, natural rubber is also used in spring elements and buffers.

10. Modification

Since 1801, natural rubber (NR) has been modified in many different ways, and modified forms have been available commercially since 1915. The degree of modification can vary from a few percent to complete modification of the polymer chain.

Modification highly affects physical properties. Even thermoplastic or resinous material can be obtained by modification of rubber.

The most well known types of modification are hydrogenation,chlorination,hydrohalogenation, cyclization, resin modification, methyl methacrylate grafting, superior-processing *N*-phenylcarbonylazoformate modification (see Section 10.8), polystyrene grafting, and epoxidation.

In addition, degraded natural rubber (liquid NR) and thermoplastic natural rubber (TP-NR) exist. Liquid NR is a chemically degraded natural rubber and a true modified rubber, whereas TP – NR is a physical blend of rubber with polypropylene. It is a new form of a thermoplastically processible natural rubber and not a real modification.

10.1. Hydrogenated Natural Rubber

[43–48]

Hydrogenated natural rubber is of purely academic interest. It was first produced by Pummerer and Burkard in 1922 and Harries in 1923, using platinum as hydrogenation catalyst. Further synthetic routes were found by Staudinger and coworkers in 1930. Hydrogenated natural rubber is a colorless, transparent, elastic to waxy solid.

Elastic hydrogenated natural rubber could, in principle, be used in the cable industry because of its good insulating capacity, as well as a raw material for adhesives. However, the relatively high cost of hydrogenation has thus far prevented any commercial applications.

10.2. Chlorinated Natural Rubber

[49–57]

Chlorinated natural rubber was the first modified natural rubber for which industrial applications were developed. Chlorination can be carried out in solution, in the latex, or on solid rubber. The chlorination mechanism was first investigated by Kraus and Reynolds and later by Bloomfield.

The glass transition temperature of chlorinated natural rubber increases with increasing chlorine content. Up to 35 wt % chlorine the material is still a rubber. Chlorinated natural rubber is more thermally stable and has a clearly higher oil resistance than pure natural rubber. Products with the *trade names* Alloprene (ICI), Parlon (Hercules), and Pergut (Bayer) with a chlorine content of ca. 65 % are available commercially. These products are crumbly materials that no longer possess rubbery properties. They are light, cream-colored, thermoplastic powders, which are nonflammable and highly resistant to chemical attack. They are used in paints and coatings that are resistant to chemicals and protect against serious environmental damage. Chlorinated rubber is dissolved in a solvent and mixed with plasticizers and pigments. These types of paint contain 10 – 12 % chlorinated rubber and can be painted or sprayed on. They are used to protect wood, steel, and cement, and for underfloor protection in car bodies and other vehicles.

Chlorinated rubber is also used in adhesives and textile coatings. However, in many areas it has been replaced by polychloroprene.

10.3. Hydrohalogenated Natural Rubber [58–63]

Hydrohalogenation can be carried out with hydrogen chloride, bromide, iodide, or fluoride. Only hydrogen chloride is used industrially. The hydrobrominated product is very unstable, the hydrogen iodide addition has thus far aroused little interest, and HF is highly toxic.

As early as 1900, WEBER found that natural rubber reacted with hydrogen chloride in chloroform. He obtained a very hard, white product with the molecular formula $(C_5H_9Cl)_x$. Hydrochlorination was further investigated in 1942 by BUNN and GARNER. They discovered that the addition of hydrogen chloride to polyisoprene gives an anti-Markovnikov product. Products with varying degrees of chlorination were synthesized by many others. The hardness of the products increases with increasing chlorine content. Completely hydrohalogenated rubber is a highly crystalline material, which assumes highly ordered structures when elongated and has a *mp* of $> 115\,°C$. It is very inert chemically and completely incompatible with natural rubber.

The largest area of use in the past was in a transparent film with the trade name Plio-Film, which was used as a packaging material, particularly for foods. Hydrohalogenated natural rubber was used as a binding agent between rubber and metals. In 1950, ca. 3500 t was produced. Hydrohalogenated natural rubber is, however, no longer available commercially.

10.4. Cyclized Natural Rubber [63–71]

Cyclized natural rubber is probably the oldest known modification of natural rubber. In 1791, LEONHARDI discovered that rubber becomes hard and brittle when treated with sulfuric acid. Similar observations were made at the beginning of the 1920s, but only in 1927 did FISCHER investigate thoroughly the reaction of natural rubber with organic sulfonic acids, sulfonyl chloride, and sulfates. At the same time BRUSON, SEBRELL, and CALVERT found that cyclized rubber is also obtained from reaction of natural rubber with zinc chloride. The mechanism of this internal cyclization is not completely understood. A suggestion based on the results of M. GORDON was worked out in 1956 by BLOOMFIELD and STOKES.

Two types of cyclized rubber were formerly produced commercially:

Products Obtained by Cyclization with Sulfuric Acid or its Derivatives. These cyclized products were subdivided into two further classes, depending on the degree of modification.

1) Brown-black, resinous substances similar to balata materials. They were sold by Goodrich under the names Thermoprene and Fenolac, and were used as adhesives, binding agents, and reinforcing resins.
2) Light, transparent, low molecular mass resins. They were soluble in many solvents and were compatible with many other resins, oils, and plasticizers. The commercial products were Plastoprene from Plastanol and Alpex 450 J from the Societa Italiana Resine. They were used for printing inks and surface varnishes, in particularly for products requiring resistance to chemicals.

Products Obtained by Cyclization with Chlorostannic Acid. The second group of products included thermoplastic, fusible materials, which were sold by Goodyear under the name Plioform. Other commercial products were masterbatches, which were sold by Durham Raw Materials under the name Cyklite and by Hubron Rubber under the name Cyklatex. They were used for shoe soles and nonblack heavy-duty industrial rollers.

10.5. Resin-Modified Natural Rubber [72–74]

Resin-modified natural rubber is not a modified rubber in the classical sense. It is formed by reaction of natural rubber with phenol – formaldehyde resins during vulcanization. Very soft vulcanizates to ebonite-type materials (hard rubber) can be produced. The products have a wide range of applications and are available, for example, under the name Cellobond from BP Chemicals. They have a low compression set, excellent dynamic stability,and good aging properties. Moreover, with reinforcing carbon

black (→ Carbon, Chap. 6.7.) (e.g., N 330) the Mooney viscosity is not increased, which improves processibility.

10.6. Poly(Methyl Methacrylate)-Grafted Natural Rubber [75–85]

Several monomers can be grafted onto natural rubber (e.g., styrene, vinyl acetate, acrylonitrile, and methyl methacrylate). Peroxides or hydroperoxides are used as initiators. The most important graft polymers are those with methyl methacrylate that are marketed under the trade name Hevea Plus MG. Their main areas of use are in self-reinforcing vulcanizates and adhesives. Three qualities of Hevea Plus MG are available commercially: MG 30 with a total content of 30 wt % methyl methacrylate, MG 40 with 40 wt %, and MG 49; the latter consists of 80 % graft polymer, 10 % free methyl methacrylate, and 10 % free polyisoprene. MG 49 is miscible with natural rubber and common carbon blacks in almost all proportions and can be vulcanized with normal sulfur cross-linking systems. About 300 – 500 t of all types is produced per year.

One use of these materials is for poly(vinyl chloride) adhesion. For this application a free-flowing powder form of Hevea Plus MG 49 is available, which dissolves rapidly in toluene or methyl ethyl ketone.

10.7. Superior-Processing Natural Rubber [30–33]

Superior-processing natural rubber is a mixture of vulcanized and unvulcanized natural rubber. The cross-linking occurs in the latex and was investigated in detail by the Rubber Research Institute of Malaysia in 1957. Products are available with 20 – 80 wt % vulcanized phase in the mixture. They are characterized by better processibility, particularly in injection molding. After mixing the vulcanized and unvulcanized latices, they are coagulated and dried together in the normal way. The products are SP 20, SP 40, SP 50, PA 57, and PA 80. The PA products, which contain more than 50 % of the vulcanized phase, are available only as masterbatches. They are used as process materials in rubber processing. The masterbatches do not contain any fillers but are stretched with oil. PA 57 is a variant of PA 80 stretched with 40 wt % oil, so that there is 57 % cross-linked rubber, 15 % unmodified rubber, and 29 % oil in the final product. Unlike other process materials, SP rubber does not alter the properties of the final vulcanizate. It leads only to improvements in calendering, extrusion, and injection molding. In calendering, these improvements include smoother surfaces, less shrinkage, and reduced temperature sensitivity of the mixture. Also, in continuous salt vulcanization the addition of PA 80 helps to prevent problems of porosity.

10.8. *N*-Phenylcarbamoylazoformate-Modified Natural Rubber [86–93]

Like the hydrogenated analogue, natural rubber modified with ethyl *N*-phenylcarbamoylazoformate (ENPCAF) is of scientific interest only. Many scientists have investigated this modification and its mechanism of formation. No commercial applications have yet been found.

10.9. Polystyrene-Grafted Natural Rubber [94–101]

Polystyrene-modified natural rubber was developed for the production of a thermoplastic elastomer. Thermoplastic elastomers frequently have an ABA block structure, where A is styrene (see Rubber, 3. Synthetic). The desired structures cannot be obtained by a common grafting process, as used for grafting with methyl methacrylate. A longer route via special polystyrenes having azodicarboxylate as the terminal group must be used. With simultaneous disintegration of the azo groups a polymer of defined structure is formed, but not a pure ABA block. Since thermoplastic natural rubber is also obtained through blends of natural rubber with polyolefins, the graft products with styrene have not gained any commercial importance.

10.10. Epoxidized Natural Rubber (ENR) [102–115]

The trialkylethylene double bonds of natural rubber readily undergo a reaction with peracids to give epoxides in high yield. This is the most economical method of modifying natural rubber at the latex stage. A mixture of hydrogen peroxide and formic acid, for example, can be used for the epoxidation. The mechanism was extensively investigated by GREENSPAN and GONSOVSKAJA after the reaction had been discovered by PUMMERER and BURKARD in 1922. A 100 % epoxidized natural rubber is a thermoplastic with a glass transition temperature of ca. 100 °C and can be processed by injection molding.

Products epoxidized to 10 – 50 % are commercially available under the names ENR 25 and 50. These products can be cross-linked by using normal vulcanization agents.

The glass transition temperature increases by ca. 1 °C per mole of epoxidation per mole rubber, so that ENR 50 has a T_g near room temperature. ENR 50 is a highly damped rubber at room temperature and at higher temperature it has the same resilience as natural rubber itself. The oil resistance is improved significantly with increasing degree of epoxidation, compared to natural rubber, whereas the gas permeability decreases. Because of this the use of epoxidized products as inner liners in tires has been investigated. However, the degree of modification must be 70 % to achieve the same gas permeability as butyl rubber, and the products are then no longer rubbery and therefore unsuitable for the purpose envisaged. Another application that is being tested is the reinforcement of NR with light fillers, for which epoxidized natural rubber is used as binding agent. These vulcanizates have the same mechanical properties as natural rubber vulcanizates with highly reactive carbon blacks, but they exhibit improved resistance to skidding in the wet and a lower resistance to rolling. Another potential area of use is improvement of the impact resistance of polystyrene. No commercial breakthroughs have occurred in any of these areas, however.

10.11. Degraded Natural Rubber [116–118]

Liquid NR is a dark, viscous material obtained by depolymerization – oxidation of natural rubber with peroxides in a kneader. STEVENS registered the first patent in 1933. The products are sold under the trade names Lorival R 525 and 200, and Hartmann DPR 35, 40, 75, 400, and 01. Their molecular masses vary between 13 000 and 160 000 g/mol.

These products are used as rubber processing materials and are incorporated into the network, so that they do not lead to any significant deterioration of the properties of the vulcanizate.

10.12. Thermoplastic Natural Rubber [119–126]

As already mentioned, thermoplastic natural rubber can be obtained either by grafting with polystyrene or by blending with polyolefins. Products based on blends are available commercially (→ Rubber, 3. Synthetic).

11. Compounding

Like all rubber, for natural rubber the property profile of the vulcanizate depends considerably on compounding. Hardness can be increased as desired by the addition of *inactive fillers* (chalk, kaolin, etc.), with decreases of toughness and elasticity. With *semiactive carbon blacks* a good compromise is achieved among good processibility, low compression set, relatively high hardness, and good toughness. For very high requirements concerning toughness and abrasive wear, *highly active fillers* must be used (highly active carbon black, light fillers with SI 69). To improve the processibility and lower the cost of compounding, *plasticizers* are added. Because of the sensitivity toward oxidation, *aging* and *ozone protection agents* are also necessary.

Vulcanization is carried out by using sulfur and vulcanization accelerators (e.g., dimorphyl disulfide) or with resin systems, depending on the area of use.

Depending on the use, substances can also be added for flameproofing, increasing the building tack, improving tear resistance, etc.

12. Summary [127]

Natural rubber is the most important rubber in terms of quantity produced because of its excellent mechanical properties (and will also remain so, at least in the medium term). The annual consumption is ca. 5×10^6 t, which is 35 % of the total world rubber consumption. The growth between 1970 and 1990 amounted to 65 %, and the growth forecast for the next five years is 1.7 % annually.

Since the retail prices have been relatively low, they should be maintained at this level by the International Rubber Agreement or at least be kept within certain limits.

Countries in which natural rubber is cultivated are attempting to achieve higher yields from their production by increasing the refinement of further processing of NR, in addition to continuous efforts to improve yields (kilograms per hectare). How successful this will be is not known. If the wage level in the producer countries increases in the long term, NR will be replaced increasingly by synthetic rubber because of rising prices. Only in applications where, for example, extremely high resistance to tear propagation is required (medical gloves, condoms, etc.) will NR not be replaced, even at a much higher price.

13. References

General References

1. M. Mortan : *Rubber Technology,* 3rd ed., Van Nostrand Reinhold, New York 1987.
2. W. Hofmann: *Kautschuk Technologie,* Gentner Verlag, Stuttgart 1980.
3. W. Hofmann: *Rubber Technology Handbook,* Hanser Verlag, Munich 1989.
4. C. Barlow: *The Natural Rubber Industry,* Oxford University Press, Kuala Lumpur 1978.
5. A. D. Roberts: *Natural Rubber Science and Technology,* Oxford University Press, Kuala Lumpur 1988.
6. G. W. Drake: "Planting and Production of Natural Rubber and Latex," *Prog. Rubber Technol.* **34** (1970) 25 – 32; **35** (1971) 7 – 122.
7. D. J. Elliott, B. K. Tidd: "Developments in Curing Systems," *Prog. Rubber Technol.* **37** (1973/74) 38 – 126.
8. M. Ulmann: *Wertvolle Kautschukpflanzen des gemäßigten Klimas,* Akademie-Verlag, Berlin 1951.
9. Wirtschaftsverband der deutschen Kautschuk-Industrie: Internationale Normvorschriften über Qualität und Verpackung von Naturkautschuksorten, Frankfurt 1969.
10. D. Kranz, K. Dinges: *Skript zur Vorlesung Latextechnologie,* Universität Hannover 1990.
11. A. K. Bhomwick, H. L. Stephens (eds.): *Handbook of Elastomers,* Marcel Dekker, New York 1988.
12. J. I. Cunneen, F. W. Shipley, *J. Polym. Sci.* **36** (1959) 77.
13. J. I. Cunneen, C. G. Moore, B. R. Shephard, *J. Appl. Polym. Sci.* **3** (1960) no. 7, 11.
14. C. Pinazzi, J. C. Danjard, R. Pautrat, *Proc. Nat. Rubber Res. Conf.* 1960, 551.
15. C. Pinazzi, R. Cheritat, R. Pautrat, *Rev. Gen. Caoutch.* **39** (1962) no. 12, 1961.

Specific References

16. B. L. Archer, B. G. Audley: "Biosynthesis of Rubber," in *Adv. in Enzymol. Relat. Areas Mol. Biol.* **29** (1967) 221.
17. R. G. O. Kekwick: "Biosynthesis of Rubber," in J. d'Auzac, J. Chrestin, J.-L. Jacob (eds.): *Physiology of Rubber Tree Latex,* CRC Press, Boca Raton, Fla., 1987.
18. F. Lynen, U. Henning, *Angew. Chem.* **72** (1960) 820, 826.
19. P. D. Abraham, P. R. Wycherley, S. W. Pakianathan, *Rubber Chem. Technol.* **45** (1972) 883.
20. A. D. T. Gorton: "The Production and Properties of Natural Rubber Latex Concentrate," *Natural Rubber Producers Res. Assoc.,* Welwyn Garden 1973.
21. G. W. Drake: "Planting and Production of Natural Rubber and Latex," *Prog. Rubber Technol.* **34** (1970) 25 – 32.
22. M. E. Cain: "The Future of Rubber from Trees and other Plants," *Royal Agricultural Society of the Commonwealth Conference,* Peterborough 1990.
23. B. C. Sekhar: "Natural Rubber," *Progr. Developm.* **18** (1965) 78.
24. Natural Rubber Technical Information Sheets, No. 9 – 120, Publ. MRPRA, Kuala Lumpur 1979.
25. J. Leveque, *Rev. Gen. Caoutch. Plast.* **43** (1966) 1304.
26. B. C. Sekhar: NR-Techn. Bulletin no. 11, Rubber Research Inst. of Malaysia, 1965.
27. B. C. Sekhar: NR-Techn. Bulletin no. 2, Rubber Research Inst. of Malaysia, 1965.

28. B. C. Sekhar, P. S. Chin, *Kautsch. Gummi Kunstst.* **19** (1966) 80.
29. St. T. Segeman, *Rubber World* **154** (1966) no. 1, 75.
30. F. W. Shipley, *J. Inst. Rubber Ind.* **1** (1967) 149.
31. H. C. Baker, *Trans. Inst. Rubber Ind.* **32** (1956) 77.
32. H. C. Baker, R. M. Foden, *Rubber Chem. Technol.* **33** (1960) 810.
33. H. C. Baker, S. C. Stokes, *Proc. Natural Rubber Res. Conf.* 1960, 587.
34. B. C. Sekhar, J. Drake, *J. Rubber Res. Inst. Malays.* **15** (1958) 216.
35. K. F. Heinisch, *Gummi Asbest* **4** (1951) 40.
36. B. C. Sekhar: Rubber SMR Bulletin no. 11, Research Institut of Malaysia, 1991.
37. B. C. Sekhar: *Malaysian Natural Rubber. New Presentation Process,* Rubber Res. Inst. Malaysia, Kuala Lumpur 1970.
38. I. Bateman, B. C. Sekhar, *Rubber Chem. Technol.* **39** (1966) 1608.
39. M. S. Sambhi: "Degradative Studies Related to the Plasticity Retention Index of the Standard Malaysian Rubber Scheme, I. Kinetics of Degradation," *Rubber Chem. Technol.* **55** (1982).
40. M. G. Smith, *Rubber J.* **149** (1967) 28.
41. B. C. Sekhar, *J. Polym. Sci.* **48** (1968) 133.
42. C. W. Thompson, *Rubber J.* **148** (1966) 24.
43. R. Pummerer, P. A. Burkard, *Ber. Dtsch. Chem. Ges.* **55** (1922) 3458.
44. K. C. Roberts, J. Wilson, GB 577 472, 1946.
45. C. D. Harriers, *Ber. Dtsch. Chem. Ges.* **56** (1923) 1048.
46. H. Staudinger et al., *Helv. Chim. Acta* **13** (1930) 1334, 1339.
47. H. Staudinger, E. O. Leupold, *Ber. Dtsch. Chem. Ges.* **67** (1934) 304.
48. D. R. Burfield, K. L. Lim, P. K. Seow, C. T. Loo: *Proc. Int. Rubber Conf.* **2** (1985) 47.
49. G. F. Bloomfield, *J. Chem. Soc.* 1943, 289.
50. *BIOS,* Final Rep. no. 1626, p. 18.
51. A. Momosee, F. Nakano: EP 356 580, 1990.
52. G. Kraus, W. B. Reynolds, *J. Am. Chem. Soc.* **72** (1950) 5621.
53. G. F. Bloomfield, in W. J. S. Nauton (ed.): *The Applied Science of Rubber,* part 1, chap. II, Edward Arnold, London 1961.
54. D. E. Woods, *J. Soc. Chem. Ind. (London)* **68** (1949) 343.
55. C. D. Harries, *Ber. Dtsch. Chem. Ges.* **46** (1913) 733.
56. V. L. Kofman, V. V. Podmasteren, S. D. Razumovski, L. B. Krentsel, *Vysokomol. Soedin. Ser. A* **29** (1987) no. 5, 1103.
57. G. J. V. Amerongen, *Ind. Eng. Chem.* **43** (1951) 2535.
58. Wingfoot Corp., US 1 989 632, 1933 (W. M. C. Calvert).
59. C. W. Bunn, E. V. Garner, *J. Chem. Soc.* 1942, 654.
60. C. O. Weber, *Ber. Dtsch. Chem. Ges.* **33** (1900) 779.
61. G. J. van Veersen: *Proc. Second Rubber Technol. Conf.,* London 1948, p. 87.
62. F. W. Hinrichsen, H. Quensell, E. Kindscher, *Ber. Dtsch. Chem. Ges.* **46** (1913) 1283.
63. D. H. E. Tom, *J. Polym. Sci.* **20** (1956) 381.
64. J. G. Leonhardi in C. W. Winkelman, H. A. Winkelman (eds.): *Systematic Survey of Rubber Chemistry,* Chem. Catalogue Co., New York 1971, p. 1923.
65. C. D. Harries: *Untersuchungen über die Natürlichen und Künstlichen Kautschukarten,* Springer Verlag, Berlin 1919, p. 6.
66. F. Kirchhof, *Kolloid. Z.* **27** (1929) 311.
67. H. L. Fischer, *Ind. Eng. Chem.* **19** (1927) 1325.
68. H. A. Bruson, L. B. Sebrell, W. C. Calvert, *Ind. Eng. Chem.* **19** (1927) 1035.
69. G. F. Bloomfield, S. G. Stokes, *Trans. Inst. Rubber Ind.* **32** (1956) 172.
70. J. I. Cunneen, E. H. Farmer, H. P. Koch, *J. Chem. Soc.* 1943, 472.
71. J. J. v. Veersen, *Recl. Trav. Chim. Pays Bas* **69** (1950) 11.
72. M. Gordon, *Proc. R. Soc. London Ser. A* **204** (1951) 569.
73. M. Gordon, *Ind. Eng. Chem.* **43** (1951) 386.
74. British Resin Products Ltd., Tech. Manual no. 11, The Westerham Press, London 1964.
75. P. W. Allen, F. M. Merrett, *J. Polym. Sci.* **22** (1956) 193.
76. F. M. Merritt, R. J. Wood, *Inst. Rubber Ind. Trans. Proc.* **32** (1956) 27.
77. P. W. Allen, C. L. M. Bell, E. G. Cockbain, *Rubber Chem. Technol.* **33** (1960) no. 3, 825.
78. P. W. Allen in L. Bateman (ed.): *The Chemistry and Physics of Rubber-like Substances,* chap. 5, Mac Laren, London 1963.
79. J. Scanlan, *Trans Faraday Soc.* **50** (1954) 756.
80. P. W. Allen, F. M. Merrett, J. Scanlon, *Trans. Faraday Soc.* **51** (1955) 95.
81. P. W. Allen, G. P. McSweeney, *Trans. Faraday Soc.* **54** (1958) 715.
82. BRPRA, Tech. Bull. no. 1, "Heveaplus M", London 1954.
83. M. A. Wheelans, *NR Technol.* **8** (1977) 69.

84. A. R. Bevan, G. G. Bloomfield, *Rubber Dev.* **16** (1963) 42.
85. J. le Bras, *Kautsch. Gummi Kunstst.* **18** (1965) 561.
86. M. E. Cain, G. T. Knight, P. M. Lewis, B. Saville, *J. Rubber Res. Inst. Malays.* **22** (1968) no. 3, 289.
87. C. S. L. Baker, D. Barnard, M. R. Porter, *Rubber Chem. Technol.* **43** (1970) 501.
88. P. J. Flory, N. Rapjohn, M. C. Shaffer, *J. Polym. Sci.* **4** (1949) 225.
89. B. M. E. van der Hoff, E. J. Buckler, *J. Macromol. Sci. Chem.* **1** (1967) 747.
90. G. T. Knight, M. J. R. Loadman, B. Saville, J. Wildgoose, *J. Chem. Soc. Chem. Commun.* 1974, 194.
91. D. Barnard, K. Dawes, P. G. Mente, *Proc. Int. Rubber Conf.* 1975, 215.
92. M. Porter, *Plast. Rubber Mater. Appl.* **3** (1978) 32.
93. G. J. van Amerongen, *Rubber Chem. Technol.* **20** (1947) 494.
94. J. C. Falk, R. J. Schlott, D. F. Hoeg, J. F. Pendleton, *Rubber Chem. Technol.* **46** (1973) 1044.
95. P. Sigwalt, A. Polton, M. Miskovic, *J. Polym. Sci. Polym. Symp.* **56** (1976) 13.
96. A. Sundet et al., *Macromolecules* **9** (1976) 371.
97. R. P. Foss, H. W. Jacobson, H. N. Cripps, W. H. Sharkey, *Macromolecules* **9** (1976) 373.
98. D. S. Campbell: *Symposium on Powdered Liquid and Thermoplastic Natural Rubber,* Phuket 1981, p. 57.
99. D. S. Campbell, P. G. Mente, A. J. Tinker, *Kautsch. Gummi Kunstst.* **34** (1981) 636.
100. H. M. R. Hoffmann, *Angew. Chem. Int. Ed. Engl.* **8** (1969) 556.
101. D. S. Campbell, D. E. Loeber, A. J. Tinker, *Polymer* **19** (1978) 1106.
102. F. P. Greenspan in E. M. Fettes (ed.): *Chemical Reactions of Polymers,* Ch. IIE, Interscience, New York 1964.
103. T. B. Gonsovskaya, F. G. Ponomarev, P. T. Poluektor, L. K. Khvatova, *Kauch. Rezina* **30** (1971) 23.
104. G. F. Bloomfield, E. H. Farmer, *J. Soc. Chem. Ind. (London)* **53** (1934) 121 T.
105. S. C. Ng, L. H. Gan, *Eur. Polym. J.* **17** (1981) 1073.
106. T. Colclough, *Trans. Inst. Rubber Ind.* **38** (1962) no. 1, TII.
107. T. Colclough, J. I. Cunneen, C. G. Moore, *Tetrahedron* **15** (1961) 187.
108. A. D. Roberts: *Institute of Physics Conference on New Materials and their Applications,* Warwick University, April 1990, p. 8.
109. D. R. Burfield, S. N. Gan, *J. Polym. Sci.* **13** (1975) 2725.
110. I. R. Gelling, *Rubber Chem. Technol.* **58** (1985) 86.
111. I. R. Gelling, J. Morrison, *Rubber Chem. Technol.* **58** (1985) 243.
112. C. S. L. Baker, I. R. Gelling, R. Newell, *Rubber Chem. Technol.* **58** (1985) 67.
113. C. K. L. Davis, S. V. Wolfe, I. R. Gelling, A. G. Thomas, *Polymer* **24** (1983) 107.
114. K. A. Nordsiek: American Chemical Society, Rubber Division Meeting, Indianapolis May 1984, Paper 48; abstract in *Rubber Chem. Technol.* **88** (1984) 201.
115. G. F. Morton, L. H. Krol: American Chemical Society, Rubber Division Meeting, Chicago 1982; abstract in *Rubber Chem. Technol.* **56** (1982) 508.
116. H. P. Stevens, Clayton & Stevens, GB 309 820, 1933 (H. P. Stevens).
117. BRPRA: Rubber Technical Developments Ltd., Broadsheet, London 1952.
118. H. V. Hardman, K. V. Hardman, US 2 349 549, 1944.
119. D. J. Elliott: *Symposium on Powdered Liquid and Thermoplastic Natural Rubber,* Phuket 1981, p. 23.
120. W. K. Fischer, US 3 806 558, 1974.
121. D. J. Duncan, GB 1 489 108, 1977.
122. A. Y. Coran, R. Patel, *Rubber Chem. Technol.* **53** (1981) 141.
123. D. S. Campbell, D. J. Elliott, M. A. Wheelans, *NR Technol.* **9** (1978) 21.
124. D. J. Elliott: Proceedings of DGF Meeting, Odense 1986.
125. A. J. Tinker, *Polym. Commun.* **25** (1984) 325.
126. D. J. Elliott, A. J. Tinker, *Proc. Int. Rubber Conf.* **2** (1985) 633.
127. H. Seifert: "Internationales Kautschukabkommen," *Gummi, Fasern, Kunstst.* **5** (1990) 250.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Rubber, 3. Synthetic

Standard Article

Desmond Threadingham[1], Werner Obrecht[2], Jean-Pierre Lambert[3], Michael Happ[4], Christiane Oppenheimer-Stix[5], John Dunn[6], Ralf Krüger[7], Heinz-Dieter Brandt[8], Wolfgang Nentwig[9], Nicola Rooney[10], Ronald T. LaFlair[11], Ute U. Wolf[12], John Duffy[13], Judit E. Puskas[14], G. J. Wilson[15], Hermann Meisenheimer[16], Roland Steiger[17], André Marbach[18], Klaus M. Diedrich[19], Jürgen Ackermann[20], Dieter Wrobel[21], Uwe Hoffmann[22], Hans Dieter Thomas[23], Rüdiger Engehausen[24], Stephen D. Pask[25], Hartmuth Buding[26], Andreas Ostrowicki[27], Bernd Stollfuss[28], Gabor Kaszas[10], Mark Drewitt[10], Stephan Glander[10], Wolfgang Wieder[29], Gerhard Wachholz[30]

[1]Bayer AG, Leverkusen, Germany
[2]Bayer AG, Leverkusen, Germany
[3]Bayer Polymères, La Wantzenau, France
[4]Bayer AG, Dormagen, Germany
[5]Bayer AG, Dormagen, Germany
[6]J. R. Consulting, Sarnia, Ontario, Canada
[7]Bayer AG, Leverkusen, Germany
[8]Bayer AG, Dormagen, Germany
[9]Bayer AG, Dormagen, Germany
[10]Bayer Inc, Sarnia, Ontario, Canada
[11]Bayer Inc, Sarnia, Ontario, Canada
[12]Bayer AG, Dormagen, Germany
[13]Bayer Inc, Sarnia, Ontario, Canada
[14]University of Western Ontario, Department of Chemical and Biochemical Engineering, London, Ontario, Canada
[15]Bayer Inc, Sarnia, Ontario, Canada
[16]Bayer AG, Leverkusen, Germany
[17]Bayer AG, Dormagen, Germany
[18]Elf Atochem, Paris La Defense, France
[19]Creannova Spezialchemie GmbH, Marl, Germany
[20]Bayer AG, Leverkusen, Germany
[21]Bayer AG, Leverkusen, Germany
[22]Budenheim, Germany
[23]Bayer AG, Leverkusen, Germany
[24]Bayer AG, Dormagen, Germany
[25]Bayer AG, Dormagen, Germany
[26]Bayer AG, Dormagen, Germany
[27]Polysar Rubber Corporation, Sarnia, Ontario, Canada
[28]Bayer AG, Dormagen, Germany
[29]Bayer AG, Leverkusen, Germany
[30]Bayer AG, Leverkusen, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a23_239.pub4
Article Online Posting Date: April 30, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Market and Areas of Application
1.2. Nomenclature and Classification
1.3. Properties
1.4. Production
1.5. Producers
2. Emulsion Rubbers
2.1. Emulsion Styrene - Butadiene Rubber (E-SBR)
2.1.1. Properties, Grades, and Application
2.1.2. Basic Chemistry and Production Processes
2.1.3. Producers and Production Capacities
2.2. Chloroprene Rubber (CR)
2.2.1. Properties, Grades, and Applications
2.2.2. Basic Chemistry and Production Processes
2.2.3. Producers and Production Capacities
2.3. Nitrile Rubber (NBR)
2.3.1. Properties, Grades, and Applications
2.3.2. Basic Chemistry and Production Processes
2.3.3. Producers and Production Capacities
2.4. Emulsion Polybutadiene (E-BR)
2.4.1. Properties and Applications
2.4.2. Basic Chemistry and Production Processes
2.4.3. Producers and Production Capacities

2.5. Acrylate Rubber (ACM)
2.5.1. Properties, Grades, and Applications
2.5.2. Basic Chemistry and Production Processes
2.5.3. Producers and Production Capacities
2.6. Fluororubbers
2.6.1. Chemical Composition and Properties
2.6.2. Vulcanization
2.6.3. Production Processes
2.6.4. Producers, Production, Capacities, and Markets
3. Solution Rubbers
3.1. Synthesis by Anionic Polymerization
3.1.1. Solution 1,3-Butadiene - Styrene Rubber (S-SBR) and Styrene - Isoprene - Butadiene Rubber (S-SIBR)
3.1.1.1. Properties, Grades, and Applications
3.1.1.2. Basic Chemistry and Production Processes
3.1.1.3. Producers and Production Capacities
3.1.2. Lithium - Butadiene (Li-BR) and Lithium - Isoprene (Li-IR) Rubber
3.1.2.1. Properties, Grades, and Applications
3.1.2.2. Basic Chemistry and Production Processes
3.1.2.3. Producers and Production Capacities
3.2. Synthesis by Ziegler - Natta Polymerization
3.2.1. Polybutadiene and Polyisoprene Rubber
3.2.1.1. Properties, Grades, and Applications
3.2.1.2. Basic Chemistry and Production Processes
3.2.1.3. Producers and Production Capacities
3.2.2. Ethylene - Propene Elastomers (EPM, EPDM)
3.2.2.1. Properties, Grades, and Applications
3.2.2.2. Basic Chemistry and Production Processes
3.2.2.3. Producers and Production Capacities
3.3. Synthesis of Butyl Rubber by Cationic Polymerization
3.3.1. Properties, Grades, and Applications
3.3.1.1. Properties
3.3.1.2. Grades
3.3.1.3. Specialty Rubbers
3.3.1.4. Applications
3.3.2. Basic Chemistry and Production Processes
3.3.2.1. Basic Chemistry
3.3.2.2. Industrial Production
3.3.3. Producers and Production Capacities
3.3.4. Storage and Transportation
3.3.5. Legal Aspects
3.3.6. Toxicology and Occupational Health
4. Synthesis by Radical and Other Mechanisms
4.1. EVM and Ethylene Copolymers
4.1.1. Properties, Grades, and Applications
4.1.2. Basic Chemistry and Production Processes
4.1.3. Producers and Production Capacities
4.1.4. Toxicology
4.2. Epoxide Rubbers (CO, ECO, GECO, GPO)
4.2.1. Properties, Grades, and Applications
4.2.2. Basic Chemistry and Production Processes
4.2.3. Producers and Production Capacities
4.3. Polynorbornene
4.3.1. Production of 2-Norbornene Monomer
4.3.2. Polymerization
4.3.3. Properties
4.3.3.1. Properties of Pure Polymer
4.3.3.2. Properties of Polynorbornene Rubbers (PNR)
4.3.4. Processing of PNR
4.3.5. Uses
4.3.6. Economic Aspects
4.3.7. Toxicology
4.4. Polyoctenamers
4.4.1. Production of Cyclooctene
4.4.2. Polymerization
4.4.3. Properties
4.4.4. Vulcanization, Processing, and Applications
4.4.5. Economic Aspects
4.4.6. Toxicology
5. Synthesis by Polyaddition, Polycondensation, and Other Mechanisms
5.1. Silicone Rubber
5.1.1. Properties, Grades, and Applications
5.1.2. Production Process and Basic Chemistry

5.1.3. Producers and Markets
5.2. Polyurethane Rubber
5.3. Thiokol Rubber
5.3.1. Properties, Grades, and Applications
5.3.1.1. History
5.3.1.2. Grades
5.3.1.3. Mechanical Properties, Cross-Linking, and Applications
5.3.2. Basic Chemistry and Production Processes
5.3.3. Producers and Production Capacities
6. Synthesis by Polymer Modification
6.1. Halobutyl Rubber
6.1.1. Introduction
6.1.2. Properties, Grades, and Applications
6.1.3. Basic Chemistry and Production Process
6.1.3.1. Basic Chemistry
6.1.3.2. Production
6.1.4. Chemical Analysis
6.1.5. Storage and Transportation
6.1.6. Legal Aspects
6.1.7. Producers and Production Capacities
6.1.8. Toxicology and Occupational Health
6.1.9. New Commercial Development
6.2. Chloropolyethylene and Chlorosulfonyl Polyethylene
6.2.1. Chloropolyethylene
6.2.1.1. Properties, Grades, and Applications
6.2.1.2. Basic Chemistry and Production Processes
6.2.2. Chlorosulfonyl Polyethylene
6.2.2.1. Properties, Grades, and Applications
6.2.2.2. Basic Chemistry and Production Processes
6.2.3. Producers and Production Capacities
6.3. Hydrogenated Nitrile Rubber
6.3.1. Properties, Grades, and Applications
6.3.2. Basic Chemistry and Production Processes
6.3.3. Producers and Production Capacities
6.4. Polyphosphazenes
6.4.1. Properties, Grades, and Applications
6.4.1.1. Properties
6.4.1.2. Commercial Grades and Applications
6.4.2. Basic Chemistry and Production Processes
6.4.3. Producers and Markets
7. Thermoplastic Elastomers
7.1. Introduction
7.1.1. Definition
7.1.2. Global Market
7.2. PVC Blends
7.3. Thermoplastic Polyurethanes
7.4. Styrene Block Copolymers
7.5. Thermoplastic Polyolefins
7.6. Copolyesters
7.7. Copolyether- and Copolyetheresteramides
8. Evaluation of the Present Situation and Remarks on Future Trends
8.1. Market, Producers, and Capacities
8.1.1. Market
8.1.2. Producers, Capacities and Capacity Utilization
8.2. Market Requirements and Impact on Development of Synthetic Rubber
8.2.1. Tires and Tire Rubber
8.3. Production Processes
8.4. Ecological and Health Aspects
8.4.1. Introduction
8.4.2. Health Aspects
8.4.2.1. Raw Materials
8.4.2.2. Processing
8.4.2.3. Use
8.4.3. Environment
8.4.3.1. Recycling
8.4.4. Outlook

[Top of Page]

1. Introduction

Desmond Threadingham and Werner Obrecht

1.1. Market and Areas of Application

In 1996 the combined annual consumption of synthetic (SR) and natural rubber (NR) was 15.6×10^6 t worldwide. Consumption grew by 9.5 % in the five years since 1991. The SR and NR consumption in the various regions is shown in Table 1 [1]. For many years, up to 1990, two thirds (ca. 67 %) of the total rubber consumed was synthetic and one third natural rubber. The years from 1991 to 1996 have seen the proportion of SR consumption decline to 63.1 %. This is explained by a large decrease in consumption (– 52.6 %) by CIS and Central Europe (the former Soviet Union and Eastern Europe), which in 1996 still consumed 86.8 % SR. At the same time, a strong increase in consumption (+ 24.4 %) is observed in Asia in Oceania. These include the major NR producing countries of the world, where only 50.6 % SR is used. These influences were partly offset by strong growth (+ 26.1 %) in North America, recovering from recession and a decline in consumption in 1991. In North America 73 % SR is used, which is similar to the European figure (71 % SR).

Table 1. Actual consumption of synthetic and natural rubber in various regions, and forecasts until 2000 (10^3 t) [1].

Region	Synthetic rubber				Natural rubber				Total rubber			
	1994	1995	1996	2000	1994	1995	1996	2000	1994	1995	1996	2000
North America	2 854	3 002	3 026	3 112	1 067	1 116	1 112	1 102	3 921	4 118	4 138	4 214
Latin America	573	576	597	680	253	261	267	303	826	837	864	983
Western Europe	2 039	2 311	2 350	2 543	862	935	944	988	2 901	3 246	3 294	3 531
CIS	1 275	505	602	708	60	48	50	85	1 335	553	652	793
Central Europe	200	290	308	368	60	80	88	130	260	370	396	498
Middle East and Africa	133	164	170	190	174	190	195	210	307	354	365	400
Asia and Oceania	1 843	2 174	2 242	2 561	2 072	2 130	2 183	2 410	3 915	4 304	4 425	4 971
Socialist Countries of Asia	480	519	555	727	795	859	919	1 204	1 275	1 377	1 474	1 931
Total	9 397	9 541	9 849	10 889	5 343	5 618	5 758	6 432	14 740	15 159	15 607	17 321

The areas of SR application and their proportion of total consumption (percent) are as follows.

Tires	59
Automotive mechanical goods	10
Nonautomotive mechanical goods	9
Plastic modification	6
Footwear	4
Construction	3
Wire covering	2
Adhesives	1
Others	6

These figures reflect a strong dependence of SR consumption on tire and automotive production. Table 2 shows the quantity of rubber used (compounded and vulcanized), rubber type, and area of application in a typical medium sized car.

Table 2. Use of rubber in a typical medium size car

Elastomer type	kg per vehicle (gross weight)	Applications
SBR	16.7 (14.9)**	tire, mount, brake hose (tube), coolant hose
BR	6.5 (6.5)**	tire
NR	8.4 (4.0)**	tire, mount, bush, mat, wiper blade
IIR	3.0 (2.9)**	tire, tube, mount
CR	3.1	hose (cover), boot (joint dust cover), fan belt
NBR (HNBR)	2.3	filler neck hose, fuel line hose (tube), power steering hose (tube), oil seal, diaphragm, emission control hose, timing belt
EPDM	5.7	coolant hose, weatherstripping, cable cover, gromment, exhaust hanger
ECO (epichlorohydrin)	0.5	hose (cover), emission control hose, diaphragm
CSM (chlorosulfonyl polyethylene)	0.5	hose (cover), cable cover
ACM (polyacrylate)	0.2	oil seal, oil hose, emission control hose
FPM (fluoroelastomer) *	0.2	diaphragm, fuel line hose (tube), O-ring
Urethane	boot (joint dust cover)	
FVMQ	diaphragm	
Total	47.1	

* For abbreviations of types of rubber, see Section Nomenclature and Classification.
** Figures in brackets indicate rubber used for tires.

1.2. Nomenclature and Classification

In addition to the first SRs which were developed in the 1920s and 1930s and which are still in use nowadays, a variety of more than 20 products now exists, most of them having been developed after World War II [2], [3]. Synthetic rubber can be classified in groups according to the chemical composition of the polymer backbone. These groups can be further subdivided with respect to the substituent groups attached to the polymer chain. The group designation is given in Table 3 [4].

Table 3. Group designation of rubber according to the chemical composition of the polymer chain [4].

Group designation	Chemical structure	Examples
M	rubber having a fully saturated polymer backbone of the polymethylene type	EPM, EPDM, CM, CSM, AEM, ACM, EVM
O	rubber having oxygen atoms in the polymer backbone (polyether rubber)	CO, ECO, GPO
Q	rubber having silicon atoms in the polymer backbone (silicon rubber)	VMQ, MQ, FMVQ
R	rubber having an unsaturated carbon main chain	NR, SBR, BR, NBR, CR, IIR
T	rubber having sulfur in the polymer backbone (polythioether rubber)	OT, EOT
U	rubber of the polyurethane type containing carbon, oxygen, and nitrogen in the polymer backbone	AU, EU
Z	rubber having phosphorus and nitrogen in the polymer backbone	FZ, PZ

This nomenclature allows synthetic rubbers to be identified by a series of letters. The letter indicating the group is the last and is preceded by one or more additional letters representing the name of the monomer or monomers from which the rubber is manufactured. Based on this system the following abbreviations listed below are used for the different types of rubber according to DIN/ISO 1629 [4].

ACM. copolymer of ethyl acrylate or other acrylates and a small amount of a monomer that facilitates vulcanization
[a] *A-CSM.* chlorosulfonated polyethylene with a high content of alkyl groups
[a] *AEM.* copolymers of ethyl acrylate or other acrylates with ethylene
AU. polyester urethane
BIIR. brominated isobutene – isoprene copolymers (bromobutyl rubber)
BR. butadiene rubbers
CIIR. chlorinated isobutene – isoprene copolymers (chlorobutyl rubber)
CM. chlorinated polyethylene
CO. polychloromethyloxirane(epichlorohydrin rubber)
CR. chloroprene rubber
CSM. chlorosulfonated polyethylene
[a] *E-BR.*

emulsion butadiene rubber
ECO. ethylene oxide (oxirane) and epichlorohydrin (chloromethyloxirane) copolymer
[a] *ENM.* copolymer of ethylene and acrylonitrile
[a] *ENR.* epoxidized natural rubber
EPDM. terpolymer of ethylene, propene, and a diene with the residual unsaturated portion of the diene in the side chain
EPM. ethylene – propene copolymer
[a] *E-SBR.* emulsion styrene – butadiene rubber
EU. polyether urethane
[a] *EVM.* ethylene – vinyl acetate copolymer
FMQ. silicone rubber having fluorine and methyl groups on the polymer chain
FMVQ. silicone rubber having fluorine, methyl, and vinyl substituents on the polymer chain
FPM. rubber with fluoro, fluoroalkyl, or fluoroalkoxy substituents on the saturated carbon main chain
FZ. polyphosphazene rubber having fluoroalkyl or fluoroalkoxy substituents on the polymer chain
GPO. copolymer of propylene oxide and allyl glycidyl ether
HIIR. halogenated butyl rubber (halobutyl)
HNBR. hydrogenated acrylonitrile – butadiene rubber (highly saturated nitrile rubber)
IIR. isobutene – isoprene copolymers (butyl rubber)
IR. isoprene rubber, synthetic
MQ. silicone rubber having only methyl substituents on the polymer chain, such as dimethyl polysiloxane
NBR. acrylonitrile – butadiene rubber
PMQ. silicone rubber having phenyl and methyl substituents on the polymer chain
PNR. polynorbornene rubber
PZ. polyphosphazene rubber with phenoxy substituents on the polymer chain
SBR-HS. styrene – butadiene rubber with high styrene content
[a] *S-SBR.* solution styrene – butadiene rubber
[a] *S-SIBR.* solution styrene – isoprene – butadiene rubber
[a] *TPE.* thermoplastic elastomer
VMQ. silicone rubber having both methylvinyl and vinyl substituent groups on the polymer chain
XNBR. carboxylated acrylonitrile – butadiene rubber
XSBR. carboxylated styrene – butadiene rubber
YAU. thermoplastic polyester urethane
YEU. thermoplastic polyether urethane
YSBR. block copolymer of styrene and butadiene
YSIR. block copolymer of styrene and isoprene
[a]. Not officially approved by DIN/ISO 1629 [4].

According to market shares and prices three different categories of SR are usually distinguished [5]:

1. General-purpose or tire rubber
2. Special rubber
3. Speciality rubber or high-performance elastomers

General-purpose rubber. includes various types of rubber with a high market volume that are available at a modest price. This category consists of the R rubbers NR, SBR, BR, and IR that are used mainly in tires. The annual consumption of each of the general-purpose rubbers is > 1×10^6 t (NR: 5×10^6 t; SBR: > 2.5×10^6 t; BR and IR: > 1×10^6 t).

Special rubber includes a series of products that are higher priced and produced in smaller volumes than general-purpose rubber. Representatives of this category are EPDM, NBR, CR, and IIR.

These products are superior to general-purpose rubber in at least one property:

EPDM	good heat and weather resistance
NBR	good oil resistance
CR	medium oil and good ozone resistance, low flammability
IIR	low air permeability

With the exception of IIR, the special rubbers are used for a variety of technical rubber goods such as extruded profiles, seals, gaskets, hoses, or power transmission belts. IIR and its halogenated derivatives — BIIR and CIIR — are used primarily for tire inner liners and tubes because of their low air permeability. The annual consumption of each special rubber is in the range 200 000 – 650 000 t.

The group of *high-performance elastomers* (speciality rubber) consists of small-volume, high-priced products. These products are distinguished by at least two major properties. Examples are

FPM. excellent oil and heat resistance
VMQ. excellent low-temperature flexibility and excellent heat resistance
HNBR. excellent dynamic performance combined with good heat and oil resistance

As for special rubbers, the most important application of high-performance elastomers is in technical rubber goods. Apart from silicon rubber with an annual consumption of more than 100 000 t, the annual consumption of the other high-performance elastomers is generally less than 100 000 t each.

1.3. Properties

SR is superior to natural rubber primarily in two properties; its heat and mineral oil resistance [6-13]. A classification of elastomers according to their heat and mineral oil resistance is shown in Figure 1 (plot according to ASTM D 2000/SAE J 200) [14], [15].

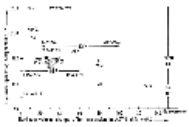

Figure 1. Classification of elastomers according to their heat and oil resistance

To enable a better overall comparison of properties, low-temperature flexibility, mechanical properties, ozone resistance, and price must also be considered (see Table 4). The data and ratings given in Table 4 are based on literature values [7-16]. In spite of various difficulties associated with a quantitative rating of complex properties, marks ranging from 1 – 10 are assigned to each of the technological properties evaluated: 1 reflects poor and 10 reflects outstanding performance. However, the rating marks do not always correlate with the data on specific properties given in Table 4. This is because the data refer to single physical properties that can be quantified rather easily, whereas the ratings refers to technological performance which accounts for many different aspects. It has to be considered that the technological performance of filled vulcanizates is highly influenced by the compounding ingredients and the vulcanization system used. Since the rating marks must account for various properties, they are not really objective from a scientific point of view. On the other hand, Table 4 and Figure 2 provides a summary of relevant technological properties and a guideline for comparing various elastomers.

Table 4. Comparative rating of properties in filled vulcanizates (1 poor and 10 outstanding performance)

Rubber	Low-temperature performance		Maximum service temperature		Mechanical performance		Oil swell (ASTM 2000–90)		Ozone resistance rating	Overall performance index	Price, $/kg	Performance – price ratio
	T_g, °C	Rating	T_{max}, °C	Rating	Tensile strength, MPa	Rating	vol % (T, °C)	Rating				
NR	–72	8	80	1	25	10	>140 (70)	1	1	21	1.3	16.2
SBR	–70/+10	6	95	3	22	7	130 (70)	2	1	19	1.1	17.3
BR	–120/–70	10	85	2	20	6	>140 (70)	1	1	20	1.1	18.2
EPDM	–60/–45	5	145	6	24	8	>140 (70)	1	8	28	1.8	15.6
IIR	–60	6	135	5	15	3	140 (70)	1	6	21	2.3	9.1
NBR	–70/+10	5	125	5	22	7	5 – 60 (100)	7	2	26	1.8	14.4
CR	–39/–35	4	115	4	22	7	80 (100)	3	6	24	2.7	8.8
CM	–25	3	140	6	15	4	80 (150)	4	5	22	2.6	8.5
CSM	–25	3	135	5	16	4	80 (150)	4	9	25	4.0	6.3
EVM	–35/+10	4	170	8	14	3	20 – 100 (150)	6	9	30	3.3	9.1
AEM	–35	4	170	8	15	4	50 (150)	5	9	30	7.3	4.1
ECO	–50	5	130	5	15	4	30 (150)	6	8	28	7.8	3.6
AU	–30	4	80	1	25	10	3 – 25 (100)	7	9	31	6.9	4.5

VMQ	–120	8	250	10	10	1	30 – 50 (150)	6	10	35	9.9	3.5
ACM	–45/–10	4	170	8	14	3	20 – 40 (150)	7	9	31	7.7	4.0
HNBR[a]	–26/ –34	3	150	6	25	10	15 – 40 (150)	8	9	36	25.0	1.4
FPM	–20	2	250	10	14	3	5 (150)	9	10	34	36.0	0.9
FMVQ	–70	8	215	9	10	1	10 (150)	9	10	37	36.0	1.0
FZ	–65	8	180	8	16	4	10 (150)	9	10	39	500.0	0.1

[a] Fully hydrogenated grades.

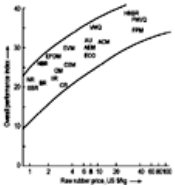

Figure 2. Performance and price of elastomers

The rating marks assigned to *low-temperature performance* (Table 4) account for T_g, rate and degree of crystallization, and compatability with plasticizers. The T_g values given in Table 4 refer to raw rubber [11].

The *maximum service temperature* referred to in Table 4 is the temperature at which a 50 % decrease in elongation is observed for an aging period of 1000 h. The temperature is usually obtained by extrapolation in an Arrhenius-type plot assuming the validity of a linear relation. The data quoted in Table 4 are from [11-15] and are based on static hot-air aging tests.

The data on *tensile strength* (TS) in Table 4 are chosen as one feature characterizing the mechanical performance of a vulcanizate. Again, tensile strength is only one aspect of mechanical performance. Other important properties are elongation at break, moduli at different extension ratios, resistance to cut growth, abrasion, dynamic performance, elasticity, and hysteresis loss (the latter two are determined by static and dynamic experiments over a wide temperature range). Data on tensile strength are primarily based on [11-13]; the ratings given for mechanical performance are based on [16].

Oil-swell data given in Table 4 refer to ASTM D 2000 [14] and SAE J 200 [15]. These data refer to volume swell of the corresponding vulcanizates after 70 h immersion in ASTM oil 3 for the temperature given in brackets. The ratings account for the volume swell and the influence of oil additives commonly used.

For an overall evaluation of elastomer performance, *outdoor performance* must be taken into account. To provide comparative data, ozone resistance seems appropriate [6]. The ratings given in Table 4 are based on [16].

The *overall performance index* given for each elastomer is the sum of five performance ratings. Thus the overall performance index of the “ideal elastomer” is 50. As Table 4 shows, even the best elastomers have overall performance indices well below 50.

Price is often referred to as the most important technological property. Because synthetic raw rubber prices strongly depend on the price of crude oil, the competitive situation, and marketing factors, the data given in Table 4 do not have absolute significance but are comparative figures.

A performance – price evaluation for various rubbers is given in Figure 2; here the overall performance indices are plotted versus the respective raw rubber prices (data from Table 4). As can be seen from this figure, simple economic features are verified. No products are available that provide either outstanding performance at low cost or low performance at a high price.

1.4. Production

The great variety of manufacturing processes for SR can be classified from both a chemical and from a chemical engineering point of view. Chemically, the mechanism governing the polymerization reaction and the chemistry involved in the key reaction are important. However, technologically speaking, the process and the series of process steps involved are major issues. In Table 5 the most important synthetic rubbers are listed according to both chemical and chemical engineering criteria. All important polymerization mechanisms are applied to SR: free-radical, Ziegler – Natta, anionic, and cationic polymerizations, as well as polyaddition and polycondensation. More recently, modification of readily available polymers has become an established route for the production of SR. Technologically important methods for chemical modification of existing polymers are chlorination and bromination of butyl rubber (CIIR and BIIR), chlorination and chlorosulfonation of polyethylene (CM and CSM), hydrogenation of NBR (HNBR), and alcoholysis of polyphosphazenes (FZ).

Table 5. Production of synthetic rubber in terms of chemical mechanisms and process features [*]

Chemical mechanism	Process features				
	Emulsion	Solution	Dispersion	Bulk	Gas phase
Radical polymerization	E-SBR; CR; NBR; E-BR; ACM; FPM; EVM	EVM	EVM	AEM; EVM (ENM)	
Ziegler – Natta polymerization		BR; EPM; EPDR	EPM; EPDM	EPM; EPDM; EPDM (BR)	(EPM); (EPDM)
Anionic polymerization		BR; S-SBR; IR		Q	
Cationic polymerization		ECO; CO	IIR	Q	
Polyaddition and polycondensation		AU; EU	EU	AU; Q	
Polymer modification		CIIR; BIIR; CM; CSM; HNBR; FZ	CM; CSM (HNBR)		

* Processes under development are given in parentheses.

The production technology includes both well-established and more advanced processes. Emulsion, solution, and dispersion processes are well established. More advanced technologies are bulk- and gas-phase processes that are as yet applied only to a few synthetic rubbers when the corresponding manufacturing facilities are available for large-volume products such as thermoplastics. For example, AEM and EVM are polymerized in bulk by using the technology developed for low-density polyethylene production. The bulk copolymerization of ethylene and acrylonitrile (ENM) is not yet applied in commercial-scale production and commerical-scale production of EPDM by the gas-phase process is only now in 1997 being introduced to the market by Union Carbide. The gas-phase process omits several of the steps (i.e., storage and purification of solvents, solution makeup, reactor cooling, solvent flashing and recovery, polymer washing, cleaning and drying) in a typical production sequence. The simplified process should reduce energy requirements and offer the potential for 20 – 30 % cost reduction. At the same time the Du Pont Dow joint venture is building new capacities to produce EPDM using a metallocene catalyst system. This single site catalyst provides a more similiar polymerization reaction rate for ethylene, propene, and ethylidene norbornene. Hence, a more consistent polymer with better distribution of the monomers along the polymer chain is produced while the constrained geometry of the catalyst environment results in a polymer which is more linear. The greater catalyst efficiency means that only minute traces of catalyst remain in the polymer. Therefore, a water washing step to clean the polymer is dispensed with in the new process, which allows cost savings and avoids problems which arise sometimes from residual moisture in the polymer [17] (further references are given in Section Production Processes, Table 64)

Table 64. Capacities, capacity shares, and plant locations of the top 11 producers of synthetic rubber according to 1995 IISRP statistics *

Company	Production capacity for solid rubber, 10^3 t	Share of total SR capacity, %	Location of plants
Bayer	911	11.9	United States, Canada, Germany, Belgium, France
Enichem	668	8.7	Italy, United Kingdom, United States
Goodyear	630	8.2	United States, France
Exxon	622	8.1	United States, France, United Kingdom, Japan
JSR	580	7.6	Japan
Shell	470	6.1	United States, France, Japan, Netherlands, Germany
Nippon Zeon	380	4.9	Japan, United States, United Kingdom
DSM	370	4.8	Netherlands, United States, Japan
Korea Kumho	355	4.6	South Korea
Du Pont Dow	354	4.6	United States, United Kingdom, Japan
Bridgestone / Firestone	260	3.4	United States

* Figures refer to Western World capacities; latex, TPE, and specialities are not included.

.

As outlined in Table 5, different routes exist for the production of various types of synthetic rubber.

The specific features and advantages of various processes and the product properties that are associated with these processes are the focus of the following chapters. Despite the very different synthetic and technological routes applied in the production of SR the same sequence of process steps is generally used:

1. Storage of raw materials and feedstocks
2. Purification of raw materials and feedstocks
3. Making up solutions and mixing of components
4. Polymerization or polymer modification reactions
5. Short stopping
6. Flashing, degassing, and stripping of unreacted monomers, solvents, and other components
7. Stabilization
8. Postmodification reactions (blending with oil, carbon black, polymers, fibers, etc.)
9. Polymer isolation, cleaning, and drying
10. Finishing of polymer, including baling, pelletizing, and packaging

Additional process steps that are not included in the straightforward production scheme are

1. Recycling of monomers, solvents, and catalysts
2. Treatment of waste air and wastewater
3. Treatment of scrap

As a consequence of economic and environmental constraints, these aspects are gaining increasing importance. To avoid additional costs for waste air and wastewater treatment, even well-established processes are currently being analyzed and improved. Economic and environmental aspects will have a great impact on future developments.

1.5. Producers

Due to takeovers, the start-up of new plants, and plant closures the range of products being supplied from various SR manufacturers has changed considerably since the early 1980s. The respective product portfolios of the major SR suppliers are listed in Table 6. Latex, TPE, and E-BR are not included in this list.

Table 6. Major manufacturers of synthetic rubber outside the former Eastern Bloc and their range of solid rubber grades *

SR producer	Range of solid rubber grades
Bayer	AU; BIIR; BR (Co, Li, Nd, Ti); CIIR; CR; EPDM; EPM; E-SBR; EVM; HNBR; IIR; NBR; Q; SBR-HS; S-SBR, XNBR
DSM	EPDM; EPM; E-SBR; NBR
DuPont	AEM; CM; CR; CSM; EPDM; EPM; EVM; EU; FPM

ENI	ACM; BR (Li, Nd, Ti); CR; EPDM; EPM; E-SBR; FPM; NBR; SBR-HS, S-SBR
Exxon	[BR(CO)]; BIIR; CIIR; EPDM; EPM; [E-SBR]; EVM; IIR
Goodyear	BR (Ni); [EPDM]; [EPM]; E-SBR; IR; NBR; SBR-HS; S-SBR; XNBR
JSR	ACM; BR (Ni); EPDM; EPM; E-SBR; [IIR]; IR; NBR; SBR-HS; S-SBR
Korea Kumho	BR (Ni); [EPDM]; E-SBR; NBR; SBR-HS; S-SBR
Michelin	BR (Ti); E-SBR; SBR-HS; S-SBR
Nippon Zeon	ACM; BR (Co); CO; ECO; E-SBR; GPO; HNBR; IR; NBR; PNR; SBR-HS; S-SBR; XNBR

* Latex and TPE are not included; products available from joint ventures are given in square brackets.

[Top of Page]

2. Emulsion Rubbers

Werner Obrecht, Jean-Pierre Lambert, Michael Happ, Christiane Oppenheimer-Stix, John Dunn, Ralf Krüger

2.1. Emulsion Styrene – Butadiene Rubber (E-SBR)

In 1929, the first copolymer of styrene and butadiene was synthesized in the Bayer laboratories by E. TSCHUNKUR and W. BOCK [18] using an emulsion polymerization process. By 1935, the production of the polymer, called Buna S, had reached 25 t per month. In the original process, the polymerization temperature was 50 °C, allowing the use of potassium persulfate as activator. Sodium linoleate was used as emulsifier but acted also as a chain-transfer agent (molar mass regulator). The styrene content of the polymer was ca. 30 wt %. The end product was difficult to process and had to be softened by heat at 130 °C before use. A significant improvement was made in 1942 when the linoleic soap was replaced by a saturated sodium fatty soap as emulsifier and diisopropyl xanthogen disulfide as chain-transfer agent. Softening by heat treatment was also required for this new grade called Buna S3 [19], [20]. The properties and the manufacturing process of Buna S were well known in the United States where research had been conducted in the same area.

At the beginning of World War II the U.S. government initiated a program to produce large amounts of styrene – butadiene copolymer known as “Government Rubber-Styrene” (GR – S). The first plant went on stream in April 1942. In this process dodecyl mercaptan was used as chain-transfer agent, which resulted in a rubber that did not need to be softened by heat. In 1945, the following production capacities were available:

Germany: 175 000 t/a;
United States: 840 000 t/a.

In 1947, the use of the “redox initiation system” allowed a reduction in polymerization temperature from 45 to 5 °C, resulting in a rubber that was much easier to process. Essentially the same “cold polymerization process” is in use in 2003. New forms of E-SBR (oil-extended SBR, black masterbatches, SBR pre-cross-linked with divinylbenzene) were introduced in the late 1940s and early 1950s. Almost all families of E-SBR marketed in 2003 were commercially available by 1954: hot and cold grades including masterbatches with black or oil [21], [22] and cross-linked types. Due to its excellent cost – performance – processing balance, E-SBR in 2003 still has — after more than 50 years — a much larger market share than any other synthetic elastomer, and represents more than 30 % of the synthetic rubber produced worldwide. It is the main synthetic elastomer used for tire manufacturing, which accounts for about 75 % of its usage [23]. However, its usage is slowly decreasing, because polybutadiene and solution SBR are more and more preferred for tire construction [24]. New derivatives of conventional E-SBR are currently being introduced in the market and could counteract that decrease.

In addition, some technological changes took place in E-SBR production to reduce possible health hazards.

2.1.1. *Properties, Grades, and Application*

Compared to other elastomers, E-SBR offers the following advantages:

1. Excellent abrasion resistance,
2. Good physical properties including crack growth resistance, good elasticity, and good stress – strain properties.
3. Excellent water resistance.
4. Easy processing.
5. Adjustable low-temperature flexibility.

Disadvantages are:

1. Medium heat resistance,
2. Poor resistance to organic chemicals,
3. Poor ozone resistance,
4. Low green strength and building tack,
5. Insufficient performances when using silica as reinforcing filler.

E-SBR starts to decompose at ca. 200 °C. Its flash point varies from 210 °C for the oil-extended grades to more than 300 °C for the non-extended grades. All E-SBRs have an autoignition temperature higher than 300 °C (ASTM D 1929). The typical rubber density varies from 0.94 g/mL for a nonextended polymer (23.5 wt % styrene) to 1.13 g/mL for a black oil masterbatch. The glass transition temperature increases with the bound styrene content. It is less than – 30 °C with 40 wt % of styrene and increases to – 20 °C for a styrene content of 50 wt % [26].

Main Families of E-SBR. Many individual grades of E-SBR are produced by using different combinations of variables such as polymerization temperature, presence of a cross-linking agent, ratio of styrene to butadiene, type of emulsifier and antioxidant, use of extender oils, or masterbatching with carbon black. The International Institute of Synthetic Rubber Producers (IISRP) has issued a nomenclature for these grades, assigning a four-digit number to each of them [27]. All the grades are classified in sections. Rubbers that belong to the same sections have the first two digits of the grade number in common.

Section 1: Hot Nonpigmented Rubber (grades 1002 to 1028). These grades offer an improved green strength and a good processibility. They are used in applications such as adhesives and cements [28], light-colored mechanical goods, or chewing gum. The cross-linked grades prepared by adding a third monomer (ethylene glycol dimethylacrylate or divinylbenzene) during polymerization, are also used as process aids when blended with general-purpose rubber. They improve the production rate and aspect in extruded and calendered goods.

Section 2: Cold Nonpigmented Rubber (grades 1500 to 1573). These grades are produced in a large range of viscosities with highly varying contents of bound styrene. They are used in applications such as adhesives, belting, footwear, mechanical goods, tire carcasses, treads, wire, and cable insulation.

Section 4: Cold Oil Masterbatch (grades 1700 to 1779). This form of E-SBR is very popular and represents more than 50 % of the E-SBR sold in the United States. These polymers have a higher molar mass and are extended with aromatic, naphthenic, or paraffinic oils at levels varying from 16 to 50 wt % based on the polymer (16 to 50 parts per hundred rubber, phr). They combine low cost with good physical properties (due to high molar mass of the ESBR used), easy processing, and improved building tack. They are used in a wide number of applications (tires, belting, footwear, hoses, mechanical goods, etc.).

Sections 3 and 5: Cold Oil – Black Masterbatch (grades 1601 to 1622 and 1805 to 1849). The number 16 (section 3) and 18 (section 5) families contain

carbon black and can contain oil. The oil content can vary from 0 to 30 phr for the 16 family and from 17.5 to 62.5 phr for the 18 family. These oil – black masterbatches are made from various elastomers (here E-SBR) carbon black, and oil types. They are used mainly in the tire industry. For the rubber processor, they offer the advantage that carbon black need not be added during processing, and that the number of mixing steps is reduced.

Emulsion Resin–Rubber Masterbatch and High-Styrene Reinforcement Polymers (grades 1900 to 1960). These polymers generally contain more styrene than the conventional E-SBR grades (up to 85 wt %) and are used in a wide range of applications where high hardness is required (e.g., for footwear and packaging).

New Products. In addition to the conventional grades, all known since the 1960s, several new experimental polymers belonging to the E-SBR family are in 2003 said to be available, but have not been reportedly sold in big volumes:

- NSBR: In that grade proposed by Bayer part of the styrene monomer is replaced by acrylonitrile. NBR has been proposed for manufacturing both technical goods and tires [29].
- Functionalized E-SBR and NSBR: Polymers functionalized with mainly hydroxyl and amino groups have been prepared. This modification improves the quality of the reinforcement, mainly when white fillers such as silica are used [30-34].
- E-SBR with a low soap content: The high residual soap level contained in E-SBR is detrimental to reinforcement. E-SBR preparation with a low residual soap has been patented [35].

Uses. E-SBR can be used alone or in blend with other polymers such as butadiene rubber to decrease the rolling resistance of tires, or it can be blended with natural rubber to improve its green strength [36]. It is compounded by using the same basic formulations as applied for other unsaturated hydrocarbon polymers. Sulfur, accelerators, zinc oxide, stearic acid, antioxidants, fillers, and softeners are generally present in the compound.

2.1.2. *Basic Chemistry and Production Processes*

E-SBR was the first commercial polymer produced by emulsion polymerization (⟶ Polymerization Processes). The mechanisms specific to E-SBR synthesis are described here briefly for the case of cold polymerization:

Generation of free radicals

$$\underset{\text{Hydroperoxide}}{ROOH} + \underset{\text{Ferrous ion}}{Fe^{2+}} \longrightarrow \underset{\text{Peroxyl radical}}{RO^{\bullet}} + \underset{\text{Ferric ion}}{Fe^{3+}} + OH^{-}$$

Initiation

$$RO^{\bullet} + \underset{\text{Butadiene}}{CH_2{=}CH{-}CH{=}CH_2} \longrightarrow RO\,(CH_2CH{=}CHCH_2)^{\bullet}$$

Propagation

$$RO(CH_2CH{=}CHCH_2)^{\bullet} + C_6H_5CH{=}CH_2 \rightarrow RO\left(CH_2CH{=}CHCH_2\right)(C_6H_5CH{-}CH)^{\bullet}$$

the growing chain can be designated $P^{\bullet}$

Chain transfer

$$P^{\bullet} + \underset{\text{Mercaptan}}{RSH} \longrightarrow PH + RS^{\bullet}$$

Chain termination by combination

$$P^{\bullet} + P^{\bullet} \rightarrow P{-}P$$

Chain termination by disproportionation

$$P^{\bullet} + P^{\bullet} \rightarrow P^{1} + P^{2}$$

Activation. Several systems have been proposed for initiation of the radical polymerization to form E-SBR. For high-temperature E-SBR (40 – 60 °C), radicals are formed by decomposition of ammonium or potassium persulfate. For the production of E-SBR at lower temperature, hydroperoxide – sugar and peroxide – iron – potassium pyrophosphate systems were first used as activators [37]. In 2003, the peroxide – iron – ethylenediaminetetraacetic acid complex – reducing agent (sodium formaldehyde sulfoxylate) system is largely favored [38].

Chain Transfer. A chain transfer agent (also called molar mass modifier), generally a long-chain tertiary mercaptan, is added to the reactor to regulate the molar mass of the polymer. Because its concentration controls the average molar mass of the polymer produced at each instant, its depletion rate must be adjusted to the polymerization rate. This is often achieved by the use of incremental modifier addition or by using mercaptans with different molar masses (C_9- to C_{16}-mercaptans are commercially available) [39].

Shortstopping. The reaction is stopped between 60 and 80 % conversion, depending on the required rubber properties. Dithiocarbamates, hydroquinones, and hydroxylamines are generally used for this purpose. Dithiocarbamates which were largely used for shortstopping until the late 1990s have generally been substituted by other compounds to avoid the build-up of carcinogenic nitrosamines. The role of the shortstopper is to scavenge radicals and eliminate excess activator still present at the end of the polymerization.

Production Process. (see Fig. 3). E-SBR can be produced batchwise or continuously. That latter process, the most commonly used, is described here: Styrene, butadiene, soap, water, modifier, and the components of the activation system are pumped into a reactor train at a controlled rate. The soaps used for manufacturing E-SBR are “natural” soaps such as tallow fatty acid soap or disproportionated rosin soap (nonmodified rosin soaps act as reaction retarders). Generally the reaction is performed in six to twelve well-agitated reactors. Since the reaction temperature greatly influences polymer structure and thus properties (see Table 8), the heat removal system (jacket, cooling plates, tubes, external exchanger) must be carefully designed if a high content of *trans*-1,4-moieties is to be obtained (reaction temperature 5 °C).

Figure 3. SBR production — simplified flow chart

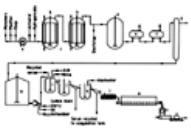

a) Precooler; b) Polymerization reactor; c) Fluid refrigeration unit; d) Blowdown tank; e) Atmospheric flash tank; f) Vacuum flash tank; g) Stripping column; h) Latex blend tank; i) Coagulation tank; j) Soap conversion tank; k) Wash tank; l) Dewater extrudes; m) Dryer; n) Baler; o) Weighscale

Reactor size varies from 10 to 30 m^3. Activator and modifier are sometimes added incrementally to control reaction kinetics and improve the copolymer properties.

When the target conversion is reached, the reaction is stopped and unreacted monomers are recovered. In the standard process, that operation is performed in three steps: In the first and second steps, the latex is flashed in two drums (the first one being at atmospheric pressure and the second one under vacuum at ca. 10 kPa). The latex is then pumped to the top of a stripping column, operating under vacuum, the latex passing downward over perforated plates countercurrent to steam.

The latex is then generally stored before being pumped to *coagulation*. The antioxidant and, in the case of masterbatch production, oil emulsion and carbon black slurry are added before coagulation. The oil used as extender can be aromatic or naphthenic. In the case of aromatic oils, the trend is to reduce the content of carcinogenic polycyclic aromatic hydrocarbons (PAH) to less than 3 wt % by using “mildly refined paraffinic” (MES) or “treated distillate aromatic extracts” (TDAE) oils. Coagulation usually takes place in two well-agitated tanks. During this operation the rubber is precipitated, usually by a mixture of sulfuric acid and brine with the addition of a small amount of animal glue or polyamines to improve the coagulation. Other possibilities are to use multivalent metal ions (calcium, aluminum, magnesium) or “salt-free” systems in which cationic flocculants are used in combination with acid, resulting in rubber with a low ash content. The rubber is then rinsed in a reslurry tank, dewatered in a dewatering extruder or on a vacuum drum filter, and dried by using either a thermal (tunnel, fluidized bed) or an extruder dryer. The rubber crumbs are finally baled and packaged.

Control of the Polymer Structure. Due to the free-radical polymerization process, E-SBR is a random copolymer and its structure can only be statistically controlled. The viscosity of the rubber increases with increasing molar mass. Table 7 illustrates the dependence between the Mooney viscosity ML 1 + 4 at 100 °C and the molar mass *M* measured by gel permeation chromatography. The butadiene moiety, as incorporated in the polymer chain, can form three isomers with *trans*-1,4-, *cis*-1,4-, and 1,2-linkages.

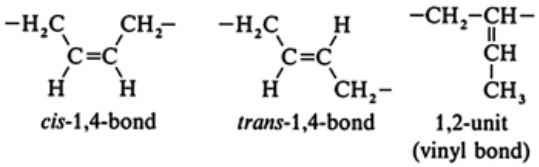

Their respective proportions can be controlled only by varying the reaction temperature. The lower the reaction temperature, the higher the percentage of *trans*-1,4-units (see Table 7) [23].

Table 7. Relationship between Mooney viscosity and molar mass *

	Mooney ML 1 + 4		
	34	**53**	**108**
Molar mass, 10^3 g/mol	350	500	850

* Measurements performed in the Bayer laboratories.

Table 8. Proportion of different structure types as a function of polymerization temperature

Reaction temperature, °C	*cis*-1,4-Units, %	*trans*-1,4-Units, %	1,2-Units, %
5	8	69	23
50	15	58	27

Branching density in E-SBR increases with butadiene content and conversion. Polymer chains start forming gels if more than one branch per chain is formed. Therefore the probability of gel formation increases with butadiene content, conversion, and molar mass [40]. The styrene content of the copolymer depends on the styrene – butadiene ratio in the feedstock. Due to the reactivity ratio r_1 and r_2 of styrene and butadiene, 0.476 and 1.413 at 15 °C [40], [41], the copolymer composition drift during polymerization is limited and there is no need to add any of the monomer incrementally during E-SBR polymerization (see also ⟶ Plastics, General Survey – Copolymerizations).

2.1.3. *Producers and Production Capacities*

There is a large overcapacity for E-SBR production: In 1995, less than 2×10^6 t was manufactured worldwide, although production capacity then was reported to be more than 4×10^6 t. The main suppliers are listed in Table 9 [41].

Table 9. Major suppliers of E-SBR and production capacity (individual capacity over 20 000 t; data from 2003)

Company name	Plant location (s)	Capacity, t
DSM Copolymer Inc	United States	150 000
Goodyear Tire & Rubber Company	United States	267 000
PASA Petroquimica Argentina SA	Argentina	53 500
Petroflex	Brazil	261 000
Industrias Negromex SA	Mexico	74 500
Bayer Elastomères	France	80 000
EniChem SPA	Italy/United Kingdom	210 000
Dow	Germany/Netherlands	176 000

Petkim Petrokimya	Turkey	27 000
Neftochim	Bulgaria	20 000
Chemopetrol	Czech Republic	76 000
Dwory	Poland	104 000
S.C Carom S.A	Romania	100 000
HIP Petrochemija	Yugoslavia	40 000
Omsky Kuchuk JSC	CIS	128 000
Togliattikauchuk	CIS	65 000
Voronezh	CIS	155 000
National Petrochemical Industry Co	Iran	50 000
Karbochem	South Africa	32 000
Quenos	Australia	35 000
Synthetics and Chemicals	India	70 000
Gadjha Tunggal	Indonesia	60 000
JSR Corporation	Japan	215 000
Mitsubishi Kasei	Japan	65 000
Zeon Corporation	Japan	200 000
Sumitomo	Japan	50 000
Hyundai Petrochemical	Korea	60 000
Kumho	Korea	200 000
TSRC	Taiwan	60 000
BST Elastomers	Thailand	60 000
SINOPEC	China	80 000
Petrochina	China	120 000

2.2. Chloroprene Rubber (CR)

2.2.1. *Properties, Grades, and Applications*

Polychloroprene [(poly(2-chloro-1,3-butadiene)] [9010-98-4] was first produced in 1932 by Du Pont under the name Duprene. Besides butyl rubber (IIR), nitrile rubber (NBR), and ethylene – propylene – diene rubber (EPDM), CR is among the most important special rubbers in terms of quantity produced [42-44].

Polychloroprene vulcanizates are characterized by the following combination of the technical properties:

1. Good mechanical strength
2. High ozone and weather resistance
3. Good aging resistance
4. Low flammability
5. Good resistance toward chemicals
6. Moderate oil and fuel resistance
7. Adhesion to many substrates

Polychloroprene can be vulcanized by using various accelerator systems over a wide temperature range. Because of its favorable properties, polychloroprene is used in the following areas:

1. In the rubber and cable industry (ca. 73 %)
2. As a raw material for adhesives (ca. 17 %)
3. As a latex (ca. 10 %)

Industrially, polychloroprene is produced by radical polymerization of chloroprene (2-chloro-1,3-butadiene) (⟶ Chlorinated Hydrocarbons) in aqueous emulsion. The reaction is exothermic (ΔH = – 65 to – 75 kJ/mol). Chloroprene polymerizes 700 times faster than other dienes such as isoprene. The high reactivity of chloroprene favors homopolymerization. Copolymers with the following monomers are important economically:

1. 2,3-Dichloro-1,3-butadiene for rubber grades with a high degree of crystallization [25067-95-2]
2. Esters of acrylic and methacrylic acids with di- and polyhydric alcohols for the production of pre-cross-linked grades
3. Elemental sulfur for the production of sulfur-modified grades
4. Unsaturated carboxylic acids (particularly acrylic and methacrylic acids) for carboxylated latices and solid rubber grades [25053-30-9]

In radical polymerization the individual chloroprene moieties are incorporated into the polymer chain in various structures (see Fig. 4). In addition to the dominating 1,4-*trans*-structure, also 1,4-*cis*-, 1,2-, and 3,4-units can be detected by NMR [45]. In the 1,4-*trans*-moiety the neighboring monomer units can be bonded head-to-head, tail-to-tail, or head-to-tail. The proportions of the individual structures in the polymer strongly depend on the polymerization temperature (see Table 10). Structural irregularities increase with rising polymerization temperature. At the same time, the degree and rate of crystallization decrease. If CR is used as a *raw material for adhesives*, high degrees of crystallinity are desirable [25]. In this case the polymer is produced at 30 °C in the absence of crystallization-distorting monomers. Films produced from solutions of such types of adhesives exhibit high cohesive strength after a short period as a result of crystallization. Hence these grades are particularly suitable for the production of contact adhesives.

Table 10. Proportions of different structure moieties in polychloroprene as a function of the polymerization temperature

Polymerization	1,4-*trans* *			1,4-*cis*	1,2	3,4
temperature, °C	H/T, %	H/H + T/T, %	Total, %	%	%	%
+ 12	83.0	11.5	94.5	3.8	1.0	0.8
+ 30	81.5	12.0	93.5	4.5	1.2	1.0
+ 42	80.5	12.0	92.5	5.2	1.2	1.1
+ 57	80.5	11.0	91.5	5.8	1.4	1.3
+ 70	75.0	13.5	88.5	8.4	1.5	1.4

* H = head; T = tail.

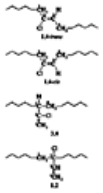

Figure 4. Incorporation of chloroprene units into the polymer chain

In *CR rubber grades*, however, crystallization is undesirable because it leads to an increase in the hardness of unvulcanized rubber mixtures on storage and a decrease in the elasticity of the vulcanizate through embrittlement [46]. At polymerization temperatures between 30 and 65 °C, crystallization can be inhibited. However, polymerization at these temperatures increases the proportion of the 1,2-structure (allylchloro structure) in the polymer chain (see Table 10). These groups are of decisive importance because vulcanization with metal oxides (in particular ZnO and MgO) occurs via the 1,2-structures, but they also promote aging of the rubber. Allylchloro structures are also formed by incorporation of 1-chloroprene, which is present as an impurity in 2-chloroprene. To produce CR rubber grades that crystallize very slowly, polymerization is performed in the presence of crystallization-distorting comonomers (in particular 2,3-dichloro-1,3-butadiene). Two main polychloroprene grades exist: the *solid rubber grades* for elastomers and adhesives, and the *latex grades.* The solid rubber grades are classified according to crystallization tendency and viscosity (solution viscosity for adhesives, Mooney viscosity for rubber articles).

Solid Rubber title.
Elastomers. Elastomers consist of general-purpose grades, pre-cross-linked grades, and sulfur-modified grades. *General-purpose grades* are mostly produced with *n*-dodecylmercaptan as the chain-transfer agent and occasionally with xanthogen disulfides. If xanthogen disulfides are used, the elastomers are more readily processible and give vulcanizates with improved mechanical properties, such as higher tensile strength and resistance to tear propagation. This improved performance allows the incorporation of higher levels of filler in the compounds and thus makes them cheaper [47].

Pre-cross-linked grades consist of a blend of soluble CR (sol) and cross-linked CR (gel). The gel particles are finely dispersed in the rubber matrix. Pre-cross-linked rubber mixtures show less swelling after extrusion (die swell), better calenderability and surface smoothness, and higher dimensional stability under load. Pre-cross-linked grades are particularly suitable for the extrusion of profiled parts.

Sulfur-modified grades are copolymers of chloroprene and sulfur. The viscosity is adjusted mostly after polymerization by cleavage of polysulfide bonds with thiuram disulfides under alkaline conditions (peptization). With a special method, so-called in situ peptization, polymerization and chain cleavage occur in parallel, supported by nucleophiles such as dithiocarbamates or xanthates.

Sulfur-modified grades can be masticated readily (lowering of viscosity) and thus allow the production of soft mixtures with good building tack. Vulcanization can be carried out with MgO and ZnO alone without an accelerator, because the terminal groups introduced by peptization are involved in the curing mechanism. Sulfur-modified grades are used in particular for parts exposed to dynamic stress, such as driving belts, timing belts, or conveyor belts, because of their excellent mechanical properties. The low stability of the raw rubbers to storage and the low hot air resistance of the vulcanizates represent disadvantages compared with general-purpose grades. The most important areas of application for elastomers in Western Europe in 1990 were as follows (in percent):

Molded goods	25 %
Cable	14 %
Transmission belts	8 %
Profiles	7 %
Conveyor belts	3 %
Others	14 %

Adhesives. The CR grades with moderate or high crystallizing tendency are used for adhesives. Special grades contain thiuram compounds, which accelerate production of the adhesive solution by rapid mastication. The most important areas of application for contact adhesives based on polychloroprene are in shoe sole adhesion; gluing of veneers in the furniture industry; adhesion of floor coverings; and do-it-yourself work with wood, textiles, and other materials.

Latex Grades. The latex grades are classified according to the crystallization tendency, type of emulsifier, deformation hardness, solubility of the dry substance, pH of the latex, and coloration tendency. A special grade, a carboxylated latex (e.g., with methacrylic acid as the comonomer), vulcanizes with ZnO even at room temperature. The most important uses of CR latex are:

1. Binding agents for all types of fibers
2. Impregnating, coating, and lining woven tissue
3. Molded foam (e.g., for chair upholstery and mattresses)
4. Improvement of bitumen
5. Dipped articles (e.g., meteorological balloons and rubber gloves)
6. Elastic concrete

2.2.2. *Basic Chemistry and Production Processes*

Polychloroprene is produced exclusively by radical polymerization in aqueous emulsions, mainly between 10 and 45 °C [42]. Polymer properties are determined primarily by the nature and quantity of the chain-transfer agent, the final degree of conversion, the polymerization temperature, and if present, the comonomers.

Other parameters primarily affect *colloidal properties*, but also alter polymer structure indirectly. These are mainly the nature and concentration of emulsifiers and coemulsifiers, the electrolyte concentration, the ratio of liquid monomer phase to aqueous phase, the kind of activation, and the polymerization process. Polymerization is carried out either continuously in a cascade of stirred-tank reactors, semibatchwise (continuous addition process), or batchwise in a stirred-tank reactor. Small quantities of special grades are best produced in a batch process. The continuous addition process is most suitable for production of copolymers from comonomers with unfavorable copolymerization parameters.

The emulsion polymerization is performed at pH 9 – 13; alkali salts of disproportionated rosin acids are mainly used as *emulsifiers*. In addition, sodium sulfonates — such as those derived from naphthalenesulfonic acid (e.g., Baykanol PQ) or alkylbenzenesulfonic acids (e.g., Marlon A 365) — are required for the process. When the alkaline latex is acidified for freeze coagulation (see below) the rosinate is converted into the water-insoluble acid, which has no emulsifying action. The sulfonic acids remain in the form of their salts in both the neutral and the weakly acidic pH regions, and maintain the stability of the latex at room temperature. For cationic or neutral latices, which are also applied in the form of latex, cationic emulsifiers such as long-chain ammonium salts or betaines are used, as well as alkyl sulfates or alkyl sulfonates. Chloroprene polymerizes very readily; in particular in the presence of air, it first reacts to form peroxy compounds, which decompose spontaneously and can produce popcorn polymers (i.e., polymers with a popcorn-like shape that are formed especially from dienes. These undesired polymers are formed by proliferous growth, especially in the gas phase at low levels of chain-transfer

agents or in the presence of cross-linkers.). Chloroprene is therefore protected against oxidation and polymerization by the addition of antioxidants, such as 2,6-di-*tert*-butyl-4-methylphenol, phenothiazine, 4-*tert*-butylcatechol, or combinations of these compounds. Nitrogen oxides and their addition products to alkenes or amines are particularly effective for preventing popcorn formation.

In industrial polymerization the stabilizers are generally not removed from the monomers and they are compensated for by suitable activation. The choice of *activator system* is determined mainly by

1. The nature and quantity of stabilizers in the monomer
2. The polymerization temperature
3. The comonomers (e.g., sulfur)
4. The presence of traces of oxygen

Activation with water-soluble persulfates in the presence of sodium 9,10-anthraquinone-2-sulfonate (silver salt) is widely used. At higher temperature (> 300 °C), sulfinic acids (e.g., formamidinesulfinic acid) can also be employed. Traces of oxygen, inhibitors, or aldehydes (which are formed as impurities in the monomer synthesis) can be eliminated by the addition of sodium sulfite. For low polymerization temperatures (< 30 °C), redox initiators (e.g., peroxodisulfate – dithionite) are generally used, sometimes with addition of iron-(II) sulfate and the disodium salt of ethylenediaminetetraacetic acid (EDTA).

In the absence of *chain-transfer agents*, chloroprene polymerizes to form cross-linked, high molar mass products, which cannot be processed even at low degrees of conversion. To obtain soluble polymers and to adjust the viscosity, chain-transfer agents such as *n*-dodecylmercaptan or xanthogen disulfides, must be added and the conversion must be limited to 70 % (max.) [48]. Polymerization is then terminated by the addition of a shortstopper. To protect the rubber against thermal aging, *stabilizers* are frequently also added at this point. The residual monomer is removed by steam distillation in vacuum (stripping) and recycled to the process. The diameter of the latex particles is generally between 100 and 200 nm. The pH of the degassed latex is adjusted to 5.5 – 7.0 by addition of dilute acetic acid. The latex is then coagulated continuously as a thin film on a slowly rotating cooling drum at ca. –15 °C; thawed by spraying with water; and washed, dried, rolled to produce a rope, cut into chips, and packed (Fig. 5). Direct rotary drying without a separate washing step is an alternative, but it is less widely used than freeze coagulation.

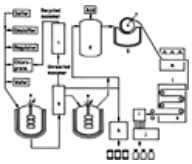

Figure 5. Schematic of the production of polychloroprene

a) Polymerization reactor; b) Stripper; c) Purification; d) Neutralization; e) Peptization; f)Rotating cooling drum; g) Coagulation by freezing; h) Latex concentration; i) Chopping machine; j) Dursting machine; k) Roping machine; l) Dryer; m) Washing

Some *latex grades* are concentrated after degassing.

For *sulfur-modified grades*, elemental sulfur is used in the polymerization instead of a regulator. This gives a cross-linked copolymer with chloroprene and sulfur sequences. The polysulfide bonds are peptized in a subsequent step with thiuram disulfide, for example; the reaction is accelerated by, e.g., dithiocarbamates. For *pre-cross-linked grades* the gel component must be polymerized at a high degree of conversion (> 80 %) to achieve sufficient cross-linking densities in the gel latex particles because of the unfavorable copolymerization parameters of chloroprene and the cross-linking agent. The degassed sol and gel latices are mixed and isolated together as solid rubber. In this way a regular distribution of the gel particles in the rubber is achieved.

2.2.3. *Producers and Production Capacities*

Polychloroprene is produced by seven Western producers at nine production sites. The total nameplate capacity is 325 000 t/a [49]. Production sites also exist in China and the former Soviet Union, which produce for their own consumption. Their production figures are estimates (see Table 11).

Table 11. Polychloroprene producers and production sites

Producer	Product name	Production site	Production capacity, t/a	Production started	Notes
Bayer	Baypren	Dormagen/Germany	60 000	1957	
Denki Kagaku Kogyo	Denka-Chloroprene	Omi/Japan	45 000	1962	
Distugil	Butaclor	Champagnier/France	40 000	1966	joint venture, 50 : 50 BP/ Rhône Poulenc
Du Pont	Neoprene	Louisville/United States	100 000	1941	
		Pontchartrain/United States	36 000	1970	
		Londonderry/Northern Ireland	33 000	1960	
Du Pont – Showa Denko	Neoprene	Kawasaki/Japan	20 000	1963	joint venture, Du Pont/Showa Denko
Miles	Baypren	Houston/United States	27 000	1970	subsidiary of Bayer
Tosoh	Skyprene	Shinnanyo/Japan	24 000	1971	
Raznoimport	Nairit	Eriwan/Armenia	40 000 *	1943	production discontinued at times
People's Republic of China		Changshon	10 000 *	1958	
		Datong	5 000 *	1966	
		Quingdao	5 000 *	1966	

* Estimated.

2.3. Nitrile Rubber (NBR)

2.3.1. *Properties, Grades, and Applications [50-56]*

Nitrile rubber, acrylonitrile – butadiene rubber, or poly(acrylonitrile-*co*-1,3-butadiene) [9003-18-3] is a special rubber, which gives vulcanizates with a balanced combination of properties. The rubber is readily processed and can be cross-linked with most common vulcanization systems. The vulcanizates exhibit excellent resistance to swelling in oils, fats, and fuels; good hot air resistance; good compression set; and abrasion resistance.

The resistance to swelling is determined mainly by the acrylonitrile content in the polymer. With increasing acrylonitrile content (i.e., higher polarity of the

vulcanizate), its resistance to swelling increases (see Fig. 6). This occurs, however, only in nonpolar or low-polarity media. The NBR vulcanizates are strongly swelled by polar solvents such as esters and ketones. Other industrially important properties are also affected by the acrylonitrile content [57]. With increasing acrylonitrile level, the low-temperature flexibility and elasticity, in particular, as well as the gas permeability, decrease, whereas the hardness, compression set, and tensile strength increase (see Fig. 7). Thus, for certain applications the acrylonitrile content of the starting polymer must be chosen in such a way that the best possible compromise between all properties is achieved. Grades of NBR with an acrylonitrile content of 15 – 53 wt % are available commercially. The most important grades have an acrylonitrile content of 28 – 38 wt %. They possess a balanced ratio of low-temperature flexibility and resistance to swelling. The glass transition temperature T_g of NBR grades with 22 – 51 wt % acrylonitrile is between – 40 and – 12 °C, according to differential scanning calorimetry (DSC) measurements [58]. For NBR grades with acrylonitrile contents of < 34 wt %, two different glass transitions are obtained experimentally, depending on whether the polymer was synthesized with or without staggered monomer dosage [59]. This effect is due to the higher chemical heterogeneity of the polymer. If the T_g is determined by using a dynamic method (vibron viscoelastometer), slightly higher T_g values are found than with DSC depending on the stress frequency [60]. Values of T_g between – 48 and – 1.5 °C were obtained for acrylonitrile contents between 19.2 and 49 wt %.

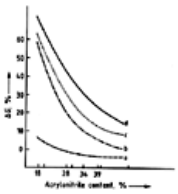

Figure 6. Swelling (change in weight, ΔG) of NBR vulcanizates versus acrylonitrile content. Immersion temperatures: ASTM oils 140 °C, liquid fuels 20 °C, immersion time 14 days

a) ASTM oil no. 1 (predominantly paraffinic mineral oil); b) ASTM oil no. 3 (aromatic – naphthenic mineral oil); c) Fuel B (isooctane – toluene 70:30); d) Fuel C (isooctane – toluene 50:50)

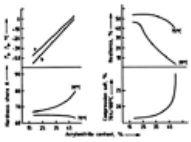

Figure 7. Second-order transition point, brittleness point, resilience, hardness, and compression set versus acrylonitrile content

a) Second-order transition point; b) Brittleness point

The *aging properties* of NBR after aging in hot air are generally better than those of natural rubber, for example. The maximum service temperature for NBR in hot air is ca. 120 °C (according to ASTM-D 2000). After suitable compounding (e.g., by using light fillers such as silica and ether-thioether as plasticizers), short-term heating up to 150 °C can be sustained [61-63]. If air is excluded, NBR vulcanizates are also stable for longer periods at 150 °C [55], [64].

The NBR vulcanizates are used mainly where resistance to oil, fuels, and fats is required (e.g., in the automotive and oil drilling industries). They are chiefly employed for production of technical rubber goods such as seals, O-rings, valves, and bellows. Nitrile rubbers are also used for producing tank and hydraulic hosing, printing and calender rollers, and brake linings.

The molecules of *NBR standard grades* are normally linear or only slightly branched. The molar mass correlates approximately with the melt viscosity, which is given as the Mooney value. Products with low Mooney values (low molar mass) are processed more readily, whereas products with high Mooney values give vulcanizates with better mechanical properties, such as tensile strength, elasticity, and compression set.

The special *cross-linked NBR grades* are produced by copolymerization with a bifunctional monomer, such as divinylbenzene. Compounding of cross-linked grades with standard NBR gives articles with a particularly good dimensional stability, especially after extrusion. The die swell is exceptionally low in these products, and very smooth surfaces are obtained. However, compared with standard grades, these vulcanizates have poorer tensile strength, elongation at break, and resistance to tear propagation.

Another class of special grades are the *carboxylated NBR grades*, *XNBR* [56], [65], [66]. These are acrylonitrile – butadiene copolymers that contain small quantities of a carboxylic acid (e.g., methacrylic acid; acid content up to 9 wt %) [67]. These grades are characterized by improved tensile strength, resistance to tear propagation, and in particular, higher abrasion resistance compared with standard NBR. These improved properties occur because during processing an additional cross-linking occurs via the carboxyl groups in the presence of metal oxides (e.g., ZnO). However, this can also lead to undesirable scorching of the mixture. This problem is counteracted, for example, by use of certain surface-treated types of ZnO or zincperoxide – NBR masterbatches [66]. The addition of sodium stearate to the mixture [68] or cross-linking with epoxy resins [69] has also been suggested. The XNBR rubbers are used mainly in the textile industry (e.g., as spinning cots and aprons) and as pump stators and seals, especially in the oil drilling sector.

An additional class of NBR grades includes those with *bound antioxidants* developed in the 1960s [70-72]. They contain an amine-based antioxidant which is incorporated into the polymer chain by copolymerization. Reportedly, these products have particularly good aging properties on swelling and contact with fuels and oils. They have thus far not been widely used, because they have not brought the expected advantages for all applications. If the vulcanizates are subjected to a so-called sequence aging (i.e., oil-aging followed by hot-air aging), polymers with bound antioxidants show slight advantages. After normal aging marginal improvements are only observed if special compounds are prepared.

To increase the ozone resistance of NBR, *NBR – PVC blends* were developed very early. The proportion of PVC must be at least 30 wt % to achieve a significant effect. A maximum of 50 wt % PVC is used. Either the blends can be produced by mixing the polymers together as solids, or NBR and PVC latices can be co-coagulated before confectioning. Homogeneous mixing of the polymers is essential; this is achieved only if a so-called fluxing occurs between 160 and 170 °C. Only then does a good combination of the resistance to swelling and elasticity of NBR with the good aging and ozone resistance of PVC result [73]. However, the brittleness temperature of the blends increases with increasing amounts of PVC. Such NBR – PVC blends are used mainly in fuel and fire engine hosing, in thermal insulating hosing and in printing rollers.

Further improvement of the aging and ozone resistance of NBR is achieved by *selective hydrogenation* of the double bonds of the butadiene units (see Section Hydrogenated Nitrile Rubber).

Powder grades, a special form of NBR rubber, were developed at the beginning of the 1970s [74], [75]. Baled products are ground to a powder but must then be treated with a dusting agent, such as silica or calcium stearate, to obtain a free-flowing powder. Other processes for the production of NBR powder involve treating NBR latex with a small quantity of a thermoplastic latex (e.g., PVC) followed by coprecipitation. The powder particles are thus coated with a thermoplastic shell, which inhibits their self-adhesion [76]. In another variant the NBR latex is spraydried so that considerable quantities of other latex components remain in the product. This process is used mainly for the production of cross-linked powder grades. Originally a new processing technology (continuous compound production) was to be introduced to the rubber industry with the powder grades. However, thus far this technology has not been able to compete with existing processing procedures. The main applications of the powder grades are in the modification of plastics, particularly of PVC and of phenol – formaldehyde resins.

2.3.2. *Basic Chemistry and Production Processes*

The first synthesis of acrylonitrile – butadiene copolymers was carried out as an emulsion polymerization at IG Farbenindustrie in Leverkusen in 1930. Large-scale industrial production of this new rubber began in Leverkusen in 1934.

The production of NBR essentially involves an emulsion polymerization process consisting of polymerization, degassing, precipitation, washing, drying, and finishing. The production processes of individual NBR producers differ only in the nature of the polymerization process (batch or continuous) or in the

equipment (or equipment combinations) used for subsequent processing steps. Greater differences can be found in the other chemical components used (e.g., emulsifiers, modifiers, initiators, shortstopping agents, and antioxidants). All polymerization processes are performed in the same temperature range.

Industrial polymerization is either a batchwise or a continuous process. *Continuous polymerization* takes place in stirred-tank reactors (see Fig. 8). The operating pressure is between 2 and 7 bar.

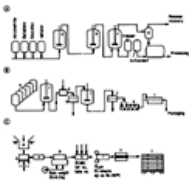

Figure 8. Schematic of NBR production

A) Polymerization; B) Isolation; C) Packaging

a) Stirred tank 1; b) Stirred tank *n*; c) Shortstopping vessel; d) Degassing; e) Storage tanks; f) Operating (lot) tank; g) Precipitation; h) Sieve 1; i) Washing cascade; j) Sieve *n*; k) Water removal; l) Dryer; m) Shaft balance;n) Bale press; o) Visual inspection; p) Control balance; q) Labeling; r) Automatic wrapping; s) Metal detector; t) Automatic paletting

The production of the individual grades differs in

1. The monomer ratio
2. The modifier content (see above)
3. The use of other comonomers (e.g., divinyl-benzene as a cross-linking agent or methacrylic acid in the production of XNBR grades)

The monomer ratio determines the acrylonitrile content in the final product. The copolymerization parameters [77] of acrylonitrile and butadiene are so different that the acrylonitrile becomes incorporated into the polymer more rapidly than butadiene. Therefore, sometimes subsequent dosing of the monomer must be carried out during polymerization to obtain a product that is chemically as uniform as possible. Acrylonitrile and butadiene polymerize azeotropically at a monomer ratio of 38 wt % acrylonitrile to 62 wt % butadiene. Thus, if the starting monomers are applied in this ratio, polymer molecules with the same chemical composition are obtained during the entire reaction time.

A good overview of the production of nitrile rubber can be found in [50]. The monomers are first emulsified in water by using one or several emulsifiers or dispersion agents, and polymerization is started by an initiator (Fig. 8 A). Alkali salts of arylalkylsulfonic acids, rosin acids, long-chain saturated and unsaturated (C_{12} – C_{18}) fatty acids, alkylsulfonic acids, alkyl sulfates, and mixtures of these compounds are usually used as emulsifiers or dispersion agents. Redox systems are predominantly used for initiation, i.e., mixtures of organic peroxides, hydroperoxides, or persulfates with reducing agents such as sodium dithionite or rongalite (sodium formaldehydesulfoxylate). Metal salts, such as iron sulfate, are often used as cocatalysts. If the reaction is performed in an alkaline medium, complexing agents such as EDTA or nitrilotriacetic acid are added to keep the iron ions in solution. Polymerization at relatively low temperature (0 – 30 °C) with good space – time yields can be achieved by using such redox systems.

The molar mass of the products is controlled with a modifier, normally a long-chain mercaptan (e.g., dodecylmercaptan or mixtures of various mercaptans containing C_8– C_{16} chains). Xanthogen disulfides are currently of minor importance.

To obtain gel-free products the reaction is usually terminated at 70 – 80 % conversion by using a shortstopping agent (c). The latter can be, for example, a reducing agent, which rapidly destroys the peroxides used as initiators. The polymer radicals growing in the latex particles must also be scavenged. Therefore shortstopping agents that are soluble in both water and organic media are used (sodium dithionite, diethylhydroxylamine, hydroquinone, hydroxyammonium chloride or sulfate, or sodium dimethyldithiocarbamate).

After shortstopping, the pressure of the polymerization vessel is released and the residual monomers are removed by distillation or steam stripping in a degassing tower (d). A stabilizer is then added to the latex to protect the rubber from oxidation during processing and subsequent storage. Sterically hindered phenols or diphenylamines, for example, are used for this purpose. The latter, however, give rubbers that are prone to discoloration.

The stabilized latex is then precipitated by the addition of electrolytes such as NaCl, $CaCl_2$, or $Al_2(SO_4)_3$, also in combination with an acid, either in a tube [78], in a precipitation vessel, or in a screw machine (i). Auxiliary products such as polysaccharides or polyamines [79], methyl celluloses [80], or similar materials are often added during the precipitation. Nitrile rubber latex can also be coagulated with cationic precipitating agents, such as quaternary condensation products from epichlorohydrin and amines. These form insoluble compounds with the anionic emulsifiers and are included in the rubber. Nevertheless, rubber coagulated this way shows markedly better water swelling, even without an additional washing step, compared to rubber coagulated by using classical precipitation electrolytes [81].

The precipitation temperature can be up to 80 °C to produce a crumb structure, which permits efficient washing out of emulsifier or initiator residues. Normally water is removed from the damp rubber in a screw machine (k). The rubber is then dried using fluidized-bed or band dryers (l), for example. A residual moisture content of < 1 % is achieved. The rubber crumbs formed are pressed into bales and packed.

2.3.3. *Producers and Production Capacities*

Since a restructuring of NBR activities has occurred because of the sale or closure of plants by various producers, at the beginning of 1992 only nine NBR producers were in operation, whose capacities were ≥ 10 000 t/a. These are listed in Table 12 in order of capacity [82]. Nitriflex and President Fine Chemicals are assumed to produce mainly for the home market. The Russian plant in Krasnojarsk with an estimated capacity of ca. 100 000 t/a is not listed. Its production is probably entirely consumed domestically. The largest producer of nitrile rubber is Bayer, followed by Nippon Zeon and Goodyear. After them come Enichem and JSR with comparable capacities.

Table 12. NBR producers, sites, capacities, and trade names

Company	Site	Capacity, t/a	Trade name
Bayer	Leverkusen (Germany)	35 000	Perbunan N
	La Wantzenau (France)	33 000	Krynac
	Sarnia (Canada)	20 000	Krynac
Nippon Zeon/Zeon Chemicals	Kawasaki (Japan)	10 000	Nipol
	Tokuyama (Japan)	10 000	Nipol
	Barry (United Kingdom)	11 000	Nipol
	Louisville (United States)	32 500	Hycar
	Akron (United States)	3 000	Hycar
Goodyear	Sandouville (France)	11 000	Chemigum
	Houston (United States)	28 000	Chemigum
Enichem	Porto Torres (Italy)	30 000	Europrene N
Japan Synthetic Rubber (JSR)	Yokkaichi (Japan)	30 000	JSR N
Uniroyal	Painesville (United States)	20 000	Paracryl
President Fine Chemicals	Kaohsiung (Taiwan)	12 000	Nancar

Nitriflex	Duque de Caxias (Mexico)	11 000	Nitriflex
DSM	Baton Rouge (United States)	10 000	Nysyn

The consumption of NBR solid rubber in 1989 was given as 208 000 t by the IISRP [82] (not including the former USSR and the communist countries of Asia and Eastern Europe). At an overall capacity of 330 000 t (not including the former USSR), large excess capacity is available.

2.4. Emulsion Polybutadiene (E-BR)

2.4.1. *Properties and Applications*

In contrast to BR which is produced in solution (anionic polymerization, see Section Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber; Ziegler – Natta polymerization, see Section Polybutadiene and Polyisoprene Rubber), E-BR [9003-17-2] is produced by an emulsion polymerization process where chain propagation occurs via a free-radical mechanism. Therefore, the microstructure of the polymer and the stereospecificity of the reaction cannot be controlled as in the case of Ziegler – Natta or Li catalysis. A typical E-BR, which is polymerized at 5 °C and shortstopped at 50 – 60 % conversion, has the following microstructure: 14 % *cis*-1,4; 69 % *trans*-1,4; and 17 % vinyl groups [83]. Due to this microstructure, E-BR exhibits a T_g around – 80 °C whereas the T_g of high-*cis* Ziegler – Natta BR is < – 100 °C [84]. Another feature associated with the free-radical emulsion polymerization of BR is gel formation. To effectively avoid gel formation the polymerization must be terminated at 50 – 60 % conversion [84].

Due to its relatively high T_g and its gel content, E-BR is rarely used for standard rubber applications such as tires and technical rubber goods. However, undisclosed quantities of E-BR are apparently used by tire manufacturers with a considerable captive or pseudocaptive supply of SR (Goodyear and Bridgestone). Because these E-BR tire grades are not commercially available, little is known about them outside these two tire companies.

E-BR is used predominantly in *rubber toughening* of brittle thermoplastic polymers, the most important of which are styrene – acrylonitrile (SAN) copolymers [85], [86]. BR containing SAN is referred to as acrylonitrile – butadiene – styrene (ABS) (⟶ Polystyrene and Styrene Copolymers). To a lesser extent, E-BR is used for the toughening of poly(vinyl chloride) (PVC), poly(methyl methacrylate) (PMMA) (⟶ Polymethacrylates), polycarbonate (PC), polyamides (PA), and polyesters (PETP, PBT) (⟶ Polymer Blends – Polyesters). All of these BR-toughened thermoplastic materials exhibit a two- or multi-phase morphology, with the BR phase being dispersed in a continuous polymer matrix. To obtain good properties (see below) of the composite multiphase materials, the BR phase must be grafted with a shell that acts as a compatibilizer between the BR phase and the surrounding polymer matrix (⟶ Polystyrene and Styrene Copolymers – ABS Production by Emulsion Polymerization). Important parameters of the BR phase are particle size, particle-size distribution, and cross-linking density. BR particles with diameters ranging from 300 – 500 nm yield a superior combination of toughness, stiffness, and surface gloss [86]. A good balance of properties can also be achieved with a bimodal particle-size distribution of the BR phase. The optimum particle diameters are 100 and 400 nm, respectively [87]. Little is known about the influence of cross-linking density of the BR phase on the technological properties of the composite materials [85], [88], [89].

2.4.2. *Basic Chemistry and Production Processes*

The auxiliary agents used for the production of E-BR are quite common and known from the production of other emulsion rubbers. The most important *emulsifiers* are the sodium or potassium salts of fully or partially hydrogenated fatty acids and of hydrogenated or disproportionated rosin acids, as well as sulfates and sulfonates of alkyl and arylalkyl derivatives. In order to obtain E-BR latices with a monomodal particle-size distribution, special emulsifiers containing two ionogenic groups per molecule are advantageous [90]. For *activation*, inorganic salts such as peroxodisulfates or peroxodiphosphates and organic activators such as dialkyl- and diaryl-based peroxides and hydroperoxides are used. Activation with redox systems is used at low polymerization temperatures. In contrast to the application requirements which the other emulsion rubbers have to meet, the utilization of gel-free E-BR is not essential in the production of ABS. Therefore E-BR is produced at polymerization temperatures above 50 °C, and the reaction is carried out until conversions are achieved that are well above the critical conversion at which gel is formed [86].

Butadiene has a fairly low propagation rate constant k_p [91-93] and the average concentration of radicals per latex particle n [94-97] is usually lower than predicted by the Smith – Ewart theory (n = 0.5) (⟶ Polymerization Processes). As a consequence of these low k_p and n values the overall polymerization rate V_{Br} depends strongly on the number of latex particles N_L as given by the Smith – Ewart relationship

$$V_{Br} = k_p \cdot n \cdot [M] \frac{N_L}{N_A}$$

where [M] is the concentration of monomer within the latex particles and N_A is Avogadro's number. Because of the strong dependence of V_{Br} on N_L the manufacturing processes differ considerably depending on the final diameter of the latex particles to be achieved. Technologically important diameter ranges are 80 – 120 nm and 300 – 500 nm. These two types of latices differ by a factor of 64 in volume. At the same monomer – water phase ratio the latex designed for larger particle diameters must contain considerably less latex particles (theoretically: 1/64). The lower particle number results in an equivalent reduction of the overall polymerization rate (V_{Br}). In practice this difference in V_{Br} is not as distinct as predicted by theory because it is somewhat compensated by higher n values [98], by particle agglomeration during the course of the reaction, and by higher monomer – water ratios. Nevertheless, the production of latices with final particle diameters between 300 and 500 nm requires a much longer polymerization time than the production of latices with diameters between 80 and 120 nm, if a direct, one-step polymerization is applied.

Consequently E-BR containing *large-sized particles* is not exclusively produced in a time-consuming one-step reaction. Two-step processes that include the agglomeration of small-sized latices are quite common.

The BR latices with *small particles* in the 100-nm-diameter region are produced in a one-step polymerization either continuously or batchwise. Safety risks are reduced and space – time performance of reactors is improved by incremental butadiene addition.

One-Step Production of Latices with Large-Sized Particles. For the one-step production of latices with particles in the 300 – 500 nm diameter range, initially a small number of latex particles is required to allow for the maximum growth of each particle. This can be achieved either by carefully controlling the initial stage of polymerizaton or by applying *seed latex technology* [86]. The surface area of each individual latex particle increases considerably during the course of the reaction. Thus emulsifier has to be added incrementally during polymerization to achieve optimal soap coverage of the latex particles and good colloidal stability of the latex. On the other hand, overdosing of emulsifier must be carefuly avoided. This would lead to the formation of second- and third-generation latex particles, resulting in a reduction of the final average diameter and in a drastic viscosity increase. The particle-size distribution of latices produced by this route is usually distinctively narrow.

Two-Step Production of Latices Containing Large Particles. For the two-step production of latices containing large particles, various agglomeration techniques are used. Physical as well as chemical methods of agglomeration are employed. The agglomeration route involves freeze agglomeration [99]. More advanced technologies apply high-pressure homogenization [100-103] and chemical agglomeration. Chemical agglomeration is achieved by changing the pH [104-106], by the addition of poly(ethylene oxides) of appropriate molar mass [107], by the addition of latices bearing hydrophilic groups [108], or by latices with a core – shell structure exhibiting a hydrophilic shell [109]. The particle-size distribution of agglomerated latices is usually broad and often polymodal.

To obtain good properties of the rubber-toughened thermoplastic materials the BR latex has to be grafted [86]. For the details of this process step, see ⟶ Polystyrene and Styrene Copolymers.

2.4.3. *Producers and Production Capacities*

Only limited quantities of E-BR suitable for thermoplast toughening are readily available on the market. One of the few suppliers of BR latex is Bayer Elastomères France (Lillebonne) (BL 2004 and KA 8318).

Vast amounts of E-BR are captively consumed by the ABS manufacturing industry. Major ABS producers in the United States, Japan, and Western Europe are Asahi, ATO, BASF, Bayer, Chemische Werke Buna, Daicel, Denki Kagaku, DOW Chemical, DSM, ENI, General Electric, Japan Synthetic Rubber, Kaneka, Mitsubishi, Monsanto, Sumitomo Naugatuck, Toray, and Ube. Due to the captive supply situation of most ABS producers, E-BR capacities are not published. By taking world ABS capacities (3.15×10^6 t) (⟶ Polystyrene and Styrene Copolymers) and assuming that the average rubber content of ABS is 15 wt % and that all ABS is based on E-BR, the total E-BR capacity can be estimated to be 473 400 t.

2.5. Acrylate Rubber (ACM)

Acrylate (ACM) rubbers, designated under ISO 1629 nomenclature as copolymers of ethyl acrylate or other acrylates and a small amount of a monomer that facilitates vulcanization (see Section Properties, Grades, and Applications) were first commercialized by B. F. Goodrich in 1948. These are specialty elastomers whose heat, oil, and ozone resistance is superior to those of general-purpose rubbers such as NR and SBR. Their manufacture, properties, and applications have been reviewed [110], [111]. The first review contains an extensive list of references, and the latter cites several earlier reviews. Both reviews refer to [112].

2.5.1. *Properties, Grades, and Applications*

Properties. The saturated backbone of ACM rubbers renders them inherently resistant to ozone. The first commercial grades were poly(ethyl acrylate) containing a small amount of comonomer as a cross-linking site because the poly(ethyl acrylate) homopolymer is difficult to cross-link. Normally, the concentration of such cross-linking sites is 1 – 5 wt %. The ACM rubbers can be used for continuous service up to 175 °C and for short periods at 200 °C, the ultimate mode of failure being embrittlement. Their aging resistance is superior to that of CR or chlorosulfonyl polyethylene (CSM), comparable to that of hydrogenated nitrile (HNBR) but not as good as that of silicone rubbers or fluoroelastomers. In their resistance to oil, they are comparable to NBR with medium acrylonitrile content and are also fairly resistant to lubricants (both lubricants according to the API-SF classification and synthetic esters). However, they are not resistant to gasoline or aromatic hydrocarbons and, generally, are not suitable for use in contact with hot water or steam. Acrylic elastomers are valued for their resistance to extreme-pressure gear lubricants containing sulfur because of their saturated backbone.

The major disadvantage of ACM rubber is its low temperature resistance. An ACM that consists primarily of *poly*(*ethyl acrylate*) has a T_g of – 15 °C [110] as determined by differential thermal analysis. Such elastomers begin to stiffen just below 0 °C. Low-temperature properties can be improved by the addition of 5 – 12 phr of plasticizer, but the most effective low-temperature plasticizers are volatile at service temperatures and are extracted by many of the fluids to which polyacrylates are exposed. *Poly*(*butyl acrylate*) has a $T_g \approx$ – 45 °C, but it has much poorer oil resistance than *poly*(*ethyl acrylate*). Consequently, a compromise is sought by copolymerizing ethyl and butyl acrylates. Homopolymers of *2-methoxyethyl* or *2-ethoxyethyl acrylate* combine a low T_g with good oil resistance, but these monomers are expensive. Consequently, those elastomers that are available on the market are copolymers of ethyl, butyl, and 2-alkoxyethyl acrylate plus a cross-linking monomer [113]. In 1991, Japan Synthetic Rubber announced the development of an ACM that is temperature resistant down to – 40 °C. The composition of the polymer has not been divulged [114].

Cross-Linking Monomers. Since the backbone of ACM elastomers is saturated, cross-linking sites must be incorporated during polymerization. Cross-linking monomers that have been evaluated include the following:

1. Vinyl monomers incorporating a labile chlorine atom that can be easily substituted (e.g., 2-chloroethyl vinyl ether, vinyl chloroacetate)
2. Unsaturated carboxylic acids (e.g., acrylic or methacrylic acid)
3. Vinyl monomers containing an epoxy group (e.g., allyl or vinyl glycidyl ethers and acrylic or methacrylic glycidyl esters)
4. Diene monomers [e.g., 5-ethylidene-2-norbornene (ENB)] [115]

The first chlorine-containing monomer used was 2-chloroethyl vinyl ether, which required the use of diamines as curatives. More reactive chlorine-containing monomers permit curing by an alkali-metal soap with sulfur as an accelerator. The disadvantages of these cure sites are that the resulting vulcanizates are corrosive and potentially toxic. These disadvantages are overcome by the use of polymers containing carboxylic or epoxide groups with *n*-octadecyltrimethylammonium bromide as the curative [113]. Curing epoxy ACM with isocyanuric acid – *n*-octadecyltrimethylammonium bromide – 1,3-diphenylurea results in long scorch time, fast cure, and low compression set [116]. *Ortho*-tolylbiguanide is added to control the reaction rate. The use of diphenylguanidine to modify reactivity has also been described [117]. Polymers containing ENB [115] have the advantage of being curable with conventional sulfur systems, which results in excellent storage stability and no mold corrosion.

Most cure systems for acrylic elastomers require an oven postcure for the development of optimum properties, including satisfactory resistance to compression set. In many applications postcuring may be avoided by the use of a soap – quaternary salt cross-linking system in a polymer containing chlorine or carboxyl cure sites [118]. A proprietary delayed-action curative – accelerator complex for polyacrylates has been described that contains a chloroacetyl group as the cure site [119]. This complex permits fast curing with no postcure, good scorch protection, and long storage stability.

Reinforcement and Additives. Acrylate rubbers have low gum strength, and carbon black reinforcement is required to ensure tensile strengths up to 17 MPa. The carbon blacks most commonly used are N550, N539, and N774. Stearic acid is commonly applied as a release agent. Antioxidants are not essential for aging resistance, but diphenylamines with low volatility marginally improve dry heat resistance.

For compounding ACM rubbers, Banbury mixing is generally preferred, because these elastomers tend to stick to the mill. Early addition of reinforcing pigments is recommended, because polyacrylates are soft and thermoplastic and quickly lose shear resistance.

Uses. The major market for ACM rubbers is in automative seals for transmission, pinions, valve stems, oil pans, camshafts, crankshafts, steering systems, wheel bearings and brakes, and in automotive hoses, which may account for as much as 25 % of the usage. Acrylate rubber containing epoxy groups and vulcanized with an isocyanuric acid system that requires no postcure is resistant to heat, automatic transmission fluid (ATF), and oxidized ATF [120]. It can be used for automatic transmission fluid (ATF) hoses. ACM rubbers are also used in adhesives, as impact improvers for plastics, in caulks and sealants, in roof mastics, and in binders for rocket propellants. The latter application arises because the polymers burn readily and leave no ash.

As self-cross-linking dispersions, polyacrylate latices are used as binders for nonwoven fabrics and as flock binders in paper saturation to render it water and oil resistant, in adhesives, in manufacture of needle-felt carpets and carpet underlays, and in textile and leather finishing.

2.5.2. *Basic Chemistry and Production Processes (see also ⟶ Polyacrylates)*

Acrylate monomers may be derived from acetylene, lactic acid, or propene (⟶ Acrylic Acid and Derivatives – Production).

Polymerization is generally carried out by free-radical reaction in emulsion or, in some cases, in suspension. Bulk polymerization is possible but is difficult to control. Solution polymerization of water-soluble monomers produces a viscous cement that is difficult to handle. Initiation by photochemical means or by radiation has been studied in the laboratory but is not used commercially.

Emulsion polymerization (⟶ Polyacrylates – Emulsion Polymerization) may be initiated by persulfates (e.g., potassium persulfate) or by hydroperoxides (e.g., *p*-menthane hydroperoxide). Because acrylates tend to hydrolyze, polymerization should be performed at pH > 7 by using a long-chain alkyl sulfate as emulsifier. The solid product is recovered by conventional coagulation washing and drying with hot air, in vacuum, or by extrusion. Antioxidant need not be added at the finishing stage since the polymer backbone is saturated and resistant to heat and light. The polymers tend to be hygroscopic, and dry storage is recommended.

2.5.3. *Producers and Production Capacities*

The most important producers of ACM rubber are listed in Table 13.

Table 13. Producers, production capacities, and trade names of ACM rubber (Based on data from International Institute of Synthetic Rubber Producers Inc.)

Producer (country)	Production capacity, t/a	Trade name
PROTEX (France)	2 000	Acrylon
Reichold (United States)	500	Tylac
Rohm & Haas (United States)	2 000	76 Res
Zeon Chemicals (United States)	5 300	Hytemp
JSR, Corp (Japan)	unknown	Arex
Nippon Zeon (Japan)	500	Nipol
NOK (Japan)	unknown	Noxtite
Mitsui Toatsu (Japan)	2 500	Polyac
Takeda (Japan)	2 500	Ultrasol
Toa Paint (Japan)	1 300	Acron
Toa Resin (Taipei)	600	Acron

2.6. Fluororubbers [121-123]

Fluororubbers represent a large class of high-performance elastomers with saturated carbon backbones in which, in comparison to conventional rubbers such as EPM, the hydrogen atoms attached to the backbone are partially or completely replaced by fluorine atoms. Compared with their hydrogenated analogues, the stabilities of both the C – F bonds and the fluorinated carbon backbone increase significantly with an increasing degree of fluorination [124]. This explains the outstanding properties of fluororubbers, such as high resistance toward heat, chemicals, ozone and weathering, and low swelling in oils and fuels. Thus, fluororubbers find application at high temperatures and in aggressive media: low temperature applications are, however, limited by their high glass transition temperature, typical values for the vulcanizates being between – 10 and 0 °C (special grades between – 30 and – 20 °C). The primary applications of the finished fluororubber articles (seals, tubing, etc.) are found in the automotive, aviation and space travel industries, in chemical plants, and in flue gas desulfurization plants (bellows expansion joints in piping), i.e., areas in which physical and chemical stability and durability under hostile conditions are required. Thus, fluororubbers find increasing application in areas completely new for rubbers, and are also replacing less resistance rubbers in conventional applications [125].

2.6.1. *Chemical Composition and Properties*

Fluororubbers (FPM) consist of fluorinated ethylene moieties, e.g., vinylidene fluoride (VDF) and tetrafluoroethylene (TFE), randomly copolymerized with fluorinated vinyl comonomers, e.g., hexafluoropropene (HFP) and perfluoro(alkyl vinyl ether) (PFAVE). The function of the comonomers is to disrupt the crystallinity of the fluorinated polyethylene. However, this is accompanied by an undesired increase in the glass transition temperature of the resulting copolymer compared with that of the homopolymer (e.g., PVDF) [126], [127]. A balanced copolymer composition is therefore required. In Table 14 the main types of fluorinated rubbers are listed according to their monomer composition, fluorine and hydrogen content, and their application temperature.

Table 14. Classification of commercial fluororubbers according to monomers, service temperature, and fluorine and hydrogen content

Fluororubber grade [a]	F content, %	H content, %	Service temperature, °C [b]
VDF – HFP	66	1.9	–18 – 210
VDF – HFP – TFE	67 – 71	1.1 – 1.7	–12 – 230
VDF – TFE – PFMVE	64 – 66.5	1.1 – 1.7	–27 – 230
TFE – PFMVE	73	0	0 – 280
TFE – P	54	4.3	5 – 200
TFE – P – VDF	56 – 60	3.1 – 3.7	– 9 [c] – 230

[a] VDF = vinylidene fluoride, HFP = hexafluoropropene, TFE = tetrafluoroethylene, PFMVE = perfluoro (methyl vinyl ether), P = propene.
[b] Region between the glass transition temperature of the raw rubber (T_g's of filled vulcanizates are 3 – 6 °C higher) and the upper maximum service temperature.
[c] According to [123].

The sequence $–CH_2–CF(CF_3)–$, and in particular its dyad and triad, which are found in fluororubbers containing vinylidene fluoride and hexafluoropropene, are especially prone to attack from bases and nucleophiles. The strong electron-withdrawing effect of the CF_3 group facilitates the elimination of HF in the presence of strong bases (e.g., present as oil additives). This leads to formation of unsaturated units in the polymer backbone [128], with subsequent reduction in the chemical and ozone resistance of the rubber [129]. The proportion of these sequences can be reduced by partial or complete replacement of the hexafluoropropene and vinylidene fluoride by perfluorinated monomers: the ageing and swelling porperties of the fluororubbers, and their stability towards bases increase with increasing degree of fluorination. Thus, perfluorinated elastomers, e.g., copolymers of tetrafluoroethylene and perfluoro (methyl vinyl ether), have the highest thermal stability. Alternatively, the resistance of the rubbers to bases can be enhanced by replacing hexafluoropropene with propenes containing less (or no) fluorine, e.g., co- and terpolymers in which the hexafluoropropene is partially or completely replaced by 2-hydropentafluoropropene [123].

The fluororubbers based on tetrafluoroethylene and propene discovered in the 1960s and subsequently marketed (Asahi Glass, tradename: Aflas) [130-133] offer a good alternative to rubbers with high fluorine contents (> 65 %). Despite their lower fluorine content (ca. 54 %), these alternating copolymers have a very high stability toward bases, are resistant toward polar solvents, experience less swelling in methanol [133-135], but suffer from extremely poor low temperature flexibility. A lowering of the glass transition temperature from – 2 °C for the tetrafluoroethylene – propene copolymer by ca. 10 °C, yet with retention of the advantageous property profile, may be achieved by terpolymerizing tetrafluoroetylene – propene with VDF [136]. Furthermore, reactive sites, such as double bonds, can be introduced into the terpolymers, enabling cross-linking reactions [137].

2.6.2. *Vulcanization*

Because of the extraordinary stability of the bonds in the polymer backbone of fluororubbers, the vulcanization step cannot be carried out by using conventional methods.

For *copolymers containing vinylidene fluoride and hexafluoropropene,* vulcanization is possible after reactive groups have been created in the polymer chain in situ by dehydrofluorination [128], [138], [139].

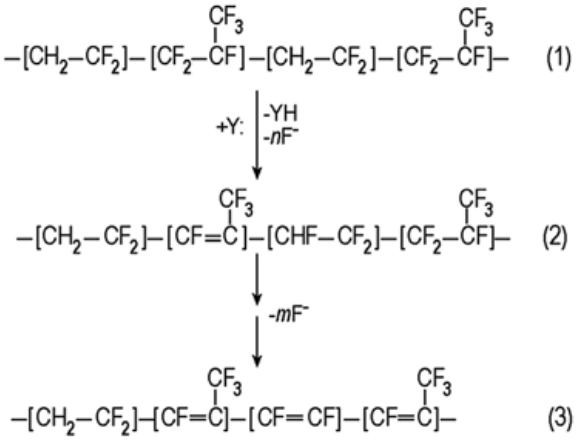

The structural element –CF=C(CF_3)– can be considered as a reactive di- or triene moiety. The fluorine atom attached to the double bond can be replaced by suitable nucleophiles. Cross-linking bridges can be created by the use of bifunctional nucleophiles such as diamines (hexamethylenediamine) or bisphenolates [2,2-bis(4-hydroxyphenyl)hexafluoropropane, i.e., bisphenol AF].

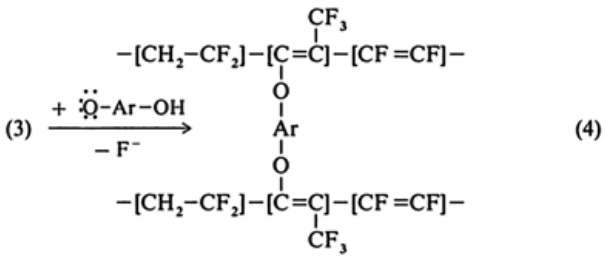

In industrial processing the vulcanization system consisting of bisphenol AF, in combination with magnesium oxide and calcium oxide (acid scavengers) and benzyltriphenylphosphonium chloride (HF elimination catalyst), is most frequently used. Processing does not normally present any problems (good mixture formation, low mold fouling). The resulting vulcanizates are characterized by satisfactory mechanical properties. Typical copolymers of vinylidene fluoride vulcanized with bisphenol have the following tension – extension characteristics: 7 – 17 MPa tensile strength and 150 – 400 % elongation at tear. The compression set values are also low at high temperature: 15 – 30 % after 70 h at 200 °C. Fluororubbers vulcanized in this way have inferior swelling and aging properties owing to dehydrofluorination during cross-linking, which leads primarily to a decrease of the fluorine content. The diene structures thus formed can then undergo undesirable reactions (e.g., they catalyze further HF elimination in the presence of base), which can lead to uncontrolled formation of new cross-linking bridges and oxidative degradation, and finally to loss of the good mechanical properties.

Reactive groups are introduced to circumvent these problems, for example to improve the resistance to automotive lubricating oils and aqueous fluids, and, in addition, to enable the vulcanization of more highly fluorinated and even perfluorinated rubbers. For the commercial fluororubber grades, that are cross-linkable with peroxides (radically), bromine or iodine (or both) can be introduced into the polymer chain. This is performed either through *copolymerization with bromine-containing monomers* [121], [122], [140-142] (in the simplest case with bromotrifluoroethylene) or by using compounds that contain iodine, or iodine and bromine, as chain-transfer agents [143-146]. Cross-linking occurs above 150 °C by using suitable peroxides [e.g., 2,5-bis(*tert*-butylperoxy)-2,5-dimethylhexane]. Di- or trifunctional olefins are added as co-cross-linking agents, of which triallyl isocyanurate forms the most stable cross-linking bridges [140].

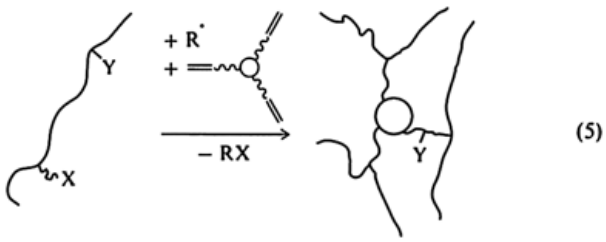

Improved vulcanizate properties (in particular, increased stability toward bases) are counterbalanced by processing problems. Severe mold contamination and the formation of toxic vulcanization cleavage products, such as methyl bromide, are particularly undesirable [125], [129]. More recent developments are aimed at obtaining peroxide-vulcanizable rubbers with optimal processing behavior (low mold fouling) and good vulcanizate characteristics (mechanical and aging properties), by combination of copolymerization with a bromine-containing comonomer and termination with iodine compounds [147, 150, 151].

Other reactive groups, which form cross-linking bridges with correspondingly high termal stability, were developed specially for *perfluoroelastomers*, whose maximum service temperatures reach 280 °C [121], [122]. Perfluoro(vinyl ether) monomers with terminal cyano or perfluorophenyl groups in the saturated side chain are used as reactive groups. They form cross-linking bridges as follows:

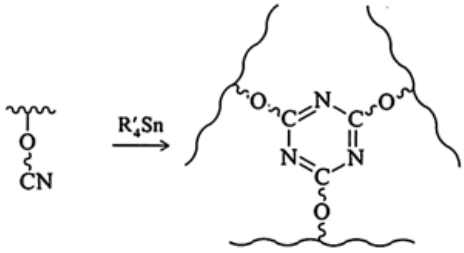

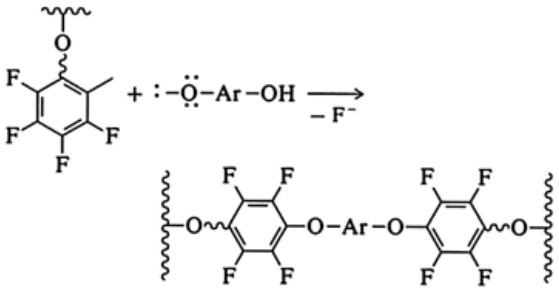

Vulcanized molded articles from perfluoroelastomers are made by Du Pont for special applications (trade name Kalrez). The nature of the reactive group depends on the conditions of use, because differing resistances to heat and chemicals are achieved depending on the type of cross-linking [122].

Production of vulcanizates involves first mixing raw rubbers with low-activity fillers, acid scavengers, and the corresponding co-cross-linking agents and catalysts, depending on the type of cross-linking. Molding and cross-linking are carried out between 160 and 200 °C. For 100 % vulcanization and removal of volatile substances the vulcanizate is then generally tempered for 24 h at > 200 °C.

Recent developments are focused on the improvement of mold flow properties and the optimization of cure characteristics (scorch safety, faster cure rate) in order to produce grades that are better processable in extrusion and injection molding [152], [153].

2.6.3. *Production Processes*

The starting materials for almost all fluoromonomers are hydrochlorofluorocarbons and hydrofluorocarbons, e.g.

$$CHClF_2 \rightarrow TFE, HFP + HCl$$

$$CH_3 - CClF_2 \rightarrow VDF + HCl$$

The most common fluororubbers are produced by radical emulsion polymerization at 5 to > 60 bar and up to 120 °C [121], [148].

Processes for production of vinylidene fluoride copolymers by solution polymerization have also been described [149]. However, they do not appear to be very practicable because of the limited choice of suitable solvents (only alkyl acetates or formates) and the high tendency of almost all organic solvents to transfer hydrogen to the reactive growing fluoropolymer radical. Thus, polymerization is stopped at an early stage, and no high molar mass products are obtained.

For formation and *stabilization* of the emulsions, fluorinated emulsifiers are necessary. Generally these are perfluorinated C_6– C_{10} carboxylates or sulfonates, the ammonium salt of perfluorooctanoic acid being the most frequently used. The use of microemulsions formed by a perfluoropolyoxyalkylene and water is recommended for a smooth introduction of special monomers, such as perfluoroalkyl vinyl ether and/or ethylene, into the copolymer [154]. Water-soluble inorganic peroxides commonly used in emulsion polymerization, such as ammonium or potassium peroxodisulfate, are used as *initiators.* The production of tetrafluoroethylene – propene copolymers is performed at low temperature because the transfer of allyl hydrogen at high temperature leads to retardation and insufficiently high molar masses. Therefore redox systems must be used as initiators only in these cases. *Chain-transfer agents* are used to regulate the molar mass. A wide choice of these is available because of the high radical reactivity already mentioned. Commonly used chain-transfer agents are isopropanol and diethyl malonate. The molar mass is generally not regulated by varying the quantity of initiator, because this leads to an increased content of ionic terminal groups and impairs the processing properties of the rubber.

Since fluororubbers are often composed of monomers with differing reactivities, the monomer concentrations during polymerization must be kept constant to obtain chemically uniform products. For this reason, continuous or batch processes are used exclusively. Continuous polymerization is carried out in a stirred-tank reactor with approximately ideal mixing under steady-state conditions. Water is charged together with initiator, emulsifier, buffer, chain-transfer agent and other additives, and the monomers, and the resulting emulsion is discharged continuously. Vinylidene fluoride copolymers are generally produced by using short residence times (15 min), constant temperatures between 90 and 120 °C, and pressures between 50 and 60 bar. Charge and discharge are adjusted in such a way that conversions of > 90 % are achieved. Further steps, such as monomer recycling, coagulation, washing, and drying, are performed according to the standard techniques used for emulsion rubbers.

2.6.4. *Producers, Production, Capacities, and Markets*

During the last decade fluororubber consumption increased from about 7000 t in 1988 to a 1997 demand of 15 000 – 20 000 t, divided among the following regions: 33 % EU, 40 % NAFTA (United States, Canada, Mexico), and 23 % Japan/South Korea (rest of the world 3 %). In 1993 worldwide production capacities were 22 000 t, essentially divided among five producers. The two major producers have strengthened their position by forming joint ventures: The Du Pont–Dow Elastomers joint venture is the largest producer, selling its rubbers under the trade name Viton. The second largest producer is Dyneon, a joint venture consisting of the 3M fluoroelastomer division and the fluoropolymer division of Hoechst (trade name: Fluorel). Other suppliers are Ausimont–Montefluos (trade name: Tecnoflon), Daikin Kogyo (trade name: Daiel), and Asahi Glass (trade name: Aflas).

[Top of Page]

3. Solution Rubbers

Heinz-Dieter Brandt, Wolfgang Nentwig, Nicola Rooney, Ronald T. LaFlair, Ute U. Wolf, John Duffy, Judit E. Puskas

3.1. Synthesis by Anionic Polymerization

3.1.1. *Solution 1,3-Butadiene – Styrene Rubber (S-SBR) and Styrene – Isoprene – Butadiene Rubber (S-SIBR)*

3.1.1.1. *Properties, Grades, and Applications*

Copolymers of styrene and 1,3-butadiene (SBR) are elastomers of great technical importance that are used predominantly in automobile tires [157]. Copolymerization can be performed either in a free radical emulsion process (E-SBR; see Section Emulsion Styrene – Butadiene Rubber (E-SBR)) or in a solution process by anionic initiation (S-SBR). The emulsion process yields pure random copolymers with a fixed microstructure of the incorporated butadiene moieties. The solution process is more versatile: the microstructure of the butadiene moieties and the styrene sequence distribution (pure

random to true block copolymers) can be adjusted [158], [159]. In addition, a great variety of endgroups can be incorporated by using the solution process [159].

S-SBR with a *random styrene sequence distribution* is predominantly used for standard rubber applications, whereas styrene – butadiene block copolymers are used in plastics applications such as high impact polystyrene (HIPS) and acrylonitrile – butadiene styrene terpolymer (ABS) (see also ⟶ Polystyrene and Styrene Copolymers).

Random Styrene – Butadiene Copolymers (S-SBR). The main application area of S-SBR is in passenger car tire treads, although E-SBR is still predominantly used in this field. However, with investigations focusing on improvements in wet grip (higher safety) as well as rolling resistance (lower fuel consumption) and abrasion resistance, the consumption of S-SBR has been increasing [157]. The proportion of energy loss ratio via the tire tread accounts for 40 % of the energy losses brought about by the whole tire [160], [155]. Therefore, the reduction of the energy loss due to tire tread is the most effective way to lower the rolling resistance. However, any improvement in rolling resistance is normally combined with a deterioration of wet grip performance (wet skid resistance). Elastomers with a low T_g show very low rolling resistance but poor wet grip performance, while elastomers with a high T_g have a good wet grip performance in combination with high rolling resistance [156, 161, 162]. Figure 9 shows the trade-off relationship between rolling and wet skid resistance.

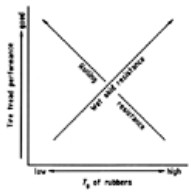

Figure 9. Trade-off relationship between rolling resistance and wet skid resistance for tire tread rubbers [163]

Many investigations to optimize the balance between high wet grip and low rolling resistance in combination with high wear performance are based on S-SBR. Typical polymer and vulcanizate properties of E-SBR and S-SBR are compared in Table 15. Because of its broad molar mass distribution and its high degree of branching E-SBR can be processed more easily than S-SBR. S-SBR cures faster than E-SBR. However, E-SBR has a higher tensile and tear strength than S-SBR. S-SBR shows lower hysteresis and provides a higher modulus. Moreover, the green strength of S-SBR is significantly improved in comparison to E-SBR.

Table 15. Polymer and vulcanizate properties of E-SBR and S-SBR

Property	S-SBR	E-SBR
	SL 705	**EM 1500**
Polymer		
Styrene content, %	25	23.5
Vinyl (1,2) content, %	10	19
cis-1,4-content, %	33	14
trans-1,4-content, %	57	67
Mooney viscosity ML1+4 (100 °C)	48.0	51.5
T_g (by DSC), °C	– 65	– 50
M_n */1000	200	145
M_w */1000	420	651
M_w/M_n	2.1	4.5
Compound (ASTM)		
ML1 + 4 (100 °C)	88.0	86.0
Rheometer tests (160 °C)		
ts1, min	4.7	5.8
tc90, min	14.9	19.4
Tensile strength, MPa	18.0	19.0
Elongation at tear, %	565	635
300 % modulus, MPa	8.9	7.8
Ring tear strength (23 °C), N	240	265
Ring tear strength (70 °C), N	235	255
Rebound resilience (23 °C), %	49	45
Rebound resilience (70 °C), %	49	49
Abrasion DIN S 40, mm^3	135	148
Abrasion DIN S 60, mm^3	78	65
tan δ, 0 °C	0.190	0.216
tan δ, 60 °C	0.172	0.183
T_g (torsional pendulum), °C	– 50	– 39

* M_w = mass-average molar mass; M_n = number-average molar mass.

E-SBR and S-SBR also differ in hydrocarbon content. E-SBR contains only ca. 90 % hydrocarbon due to the high level of soap residue (see Section Basic Chemistry and Production Processes), while the hydrocarbon content of S-SBR is close to 100 %. These soap residues are responsible for the high

hysteresis of E-SBR [164]. However, these residues also improve the processibility and increase tear strength.

Due to the difference in microstructure [different content of 1,2-(vinyl)-configurations], standard S-SBR has a lower T_g than E-SBR at the same level of bound styrene (Table 15). Also the difference in macrostructure is obvious. S-SBR has a narrower molar mass distribution and a more linear backbone.In tire tread applications S-SBR therefore provides lower rolling resistance, improved dry slide traction, and superior wear resistance, while E-SBR provides the better wet slide traction [165], [166].

In 1991 the "green tire" concept was launched with Michelin's patent application [339]. In this concept, the subsitution of silica for carbon black gives rise to lower rolling resistance, improved wet grip, and good wear properties. In terms of rubber, this concept is based on S-SBR or blends of S-SBR with BR or NR: An strong increase in the demand for "green tires" is anticipated [180].

Vinyl-S-SBR (VS-SBR). Traditional S-SBR types with 10 % vinyl content are of minor interest for use in modern high performance tires [157]. S-SBR grades with a vinyl content up to 70 % and with a styrene content up to 40 % are tailor-made for this application and are commercially available. They meet the demand for lower rolling resistance and higher wet grip. By additional variation of the macrostructure (bimodal molar mass distribution), the aforementioned effects are enhanced and, furthermore, the processibility is improved [170-172].

Styrene – Isoprene – Butadiene Terpolymer (S-SIBR). The development of S-SIBR is based on the "wide modal tan δ concept" [173]. S-SIBR rubber is often referred to as *integral rubber*. It has a tailor-made chain consisting of different segments exhibiting different T_gs. The maxima of the damping peaks of the separate segments result in a tan δ curve of the polymer with a wider peak profile. The principle is shown in Figure 10. The absolute values for tan δ given for the three temperature ranges in Figure 10 indicate the generally accepted relationship between tan δ and tire characteristics regarding rolling resistance (50 to 80 °C), wet grip (– 20 to 20 °C), and wear resistance (– 50 to – 100 °C). This concept has been known for several years [174-176, 179]. Commercial production of S-SIBR began in the United States in 1991 [157].

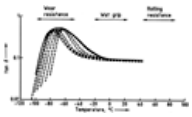

Figure 10. "Integral rubber"—wide modal tan δ concept [173]

Improvements in the manufacture process were reported in the 1990s [180], [181]. The concept of chain-end coupling can be applied also with S-SIBR [182]. Tire tread compounds containing S-SIBR and silica as filler are claimed to have improved properties [183].

Styrene – Butadiene Block Copolymers. These block copolymers exhibit two separate ranges of T_gs: – 80 to – 90 °C for the polybutadiene block and 70 to 80 °C for the polystyrene block [184]. These polymers are not suitable for standard rubber applications such as tires [162].

Butadiene – Styrene Diblock Copolymers (SB). SB diblock copolymers are used for the production of HIPS and mass polymerized ABS, either alone or in blends with polybutadiene. By using diblock copolymers, HIPS properties such as gloss and transparency which are often unsatisfactory can be improved [185], [186]. SB diblock copolymers can also be used to increase impact strength of polystyrene and polyolefins by mechanical blending.

SB diblock copolymers with different styrene content are available: 40 % bound styrene (for HIPS), 30 % bound styrene (for ABS, HIPS), 25 % bound styrene (for carpet pads and bitumen modification), and 10 % bound styrene (for ABS).

Styrene – Butadiene Triblock Copolymers (SBS). Linear [187] and radial SBS types [188], [189] are commercially available. They are typical thermoplastic elastomers (see Section Styrene Block Copolymers; ⟶ Thermoplastic Elastomers).

3.1.1.2. *Basic Chemistry and Production Processes*

The versatility of anionic solution polymerization allows the formation of tailor-made diene – styrene copolymers. The unique features of copolymerization initiated by butyllithium in nonpolar hydrocarbon solvents are:

1. Absence of spontaneous chain terminations (living chains polymerization, ⟶ Plastics, General Survey – Anionic Polymerizations), which allows the sequential addition of monomers and the use of chain-coupling techniques
2. Absence of chain-transfer reactions to give essentially linear and gel-free polymers
3. A wide choice of chain microstructures [159], [190], [194]

The use of butyllithium in nonpolar, hydrocarbon solvents leads to polybutadienes in which the diene units are present mainly in the 1,4-configuration (ca. 90 %), the remainder being in the 1,2-configuration (vinyl structure).

SB Diblock Copolymers. SB copolymers are usually produced in a batch polymerization. Because of the different reaction rates of the two monomers in hydrocarbon solvents (reaction rate of butadiene is higher), the diene dominates the initial copolymerization reaction. This inevitably leads to the formation of tapered diblock copolymers in which the almost pure polybutadiene block is connected to a polystyrene block through a transitionally random polymer segment consisting of increasing proportions of styrene. About 75 % of the total bound styrene content is incorporated as a polystyrene block [158], [195].

SBS Triblock Copolymers (SBS). Different strategies exist for the synthesis of triblock styrene – butadiene copolymers:

Due to the "living character" of the anionic polymerization, SBS can be produced by polymerization of styrene to form *living polystyrene chains*, followed by chain extension with butadiene to form a diblock (SB) which then can be (1) extended with a second polystyrene end-block [187] or (2) coupled with di- or polyfunctional coupling agents (e.g., mono- or diesters, metal halides) to produce linear, 3-, 4- or multibranched copolymers [189].

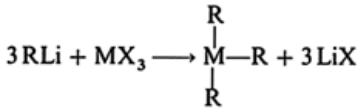

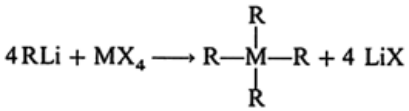

Triblock copolymers can also be produced by dilithium compound-initiated polymerization of butadiene to form *living polybutadiene chains* followed by building the styrene blocks on both ends of the polybutadiene chain by adding styrene [196], [197].

S-SBR. Completely random copolymers can be obtained in nonpolar solvents either by a continuous polymerization process [198] or a semibatch process with programmed addition of butadiene to the reaction mixture [199]. S-SBR made by these routes is characterized by a vinyl content of ca. 10 % in the diene fraction.

Another route utilizes randomizers, in batch or continuous operation. These randomizers or modifiers are polar compounds such as ethers or amines which decrease the difference between the reactivities of butadiene and styrene to such an extent that both monomers are consumed at virtually equal rates [200], [201]. The randomizing action of these polar compounds results from an effect they have on the initiator and coincides with an increased amount of

1,2-addition in butadiene polymerization. The vinyl content in the polybutadiene depends on the modifier type, modifier – catalyst ratio and concentration (modifier and catalyst), and the polymerization temperature [161]. The vinyl content of the diene fraction can be up to 90 %.

This process flexibility is an important and attractive possibility — in addition to changing the bound styrene level — to control the T_g of the derived polymers. The relationship between the vinyl and styrene contents and T_g is illustrated in Figure 11 for a series of random S-SBR. The high linearity together with a narrow molar mass distribution of the S-SBR results in a relatively poor processing performance during mixing and molding operations. This can be improved by the coupling of polymer chains through the "living" chain ends. Various coupling agents (dihalides, tetrahalides, divinylbenzene, etc.) are available [170-172], [202]. Tin tetrachloride shows not only good coupling efficiency but also an improved interaction (e.g., better rolling resistance) with the filler carbon black. A similar effect can be obtained by a chain end modification by terminating the "living" chain end with a modifier such as *N*-methyl-2-pyrrolidone [203], *N*-substituted aziridines [204], aminobenzophenone compounds [177], [178], and substituted imines [205].

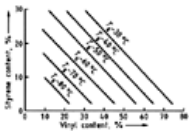

Figure 11. Relationship of T_g to bound styrene and vinyl content [156]

The concept of introducing polymer-bound functionalities can be extended by using functionalized initiators such as lithium cycloalkylamides [191-193], lithium compounds containing imine groups [213], [214], and LiSn alkyls [216], [217].

Novel S-SBR types formed by selective hydrogenation of 1,2-vinyl units have been presented. They are characterized by an extremely small hysteresis loss at high temperature and a large loss at lower temperature [218].

Process Description. S-SBR, VS-SBR, and S-SIBR are commercially produced by batch and continuous anionic solution processes. Typically, these processes consist of four operating stages: polymerization, monomer and solvent recycling, product recovery, and product finishing. The polymerization is carried out at 50 – 100 °C and 1 – 5 bar in a batch reactor or in a reactor train of two to four stirred-tank reactors. Conversion is complete after a residence time of ca. 1 – 4 h depending on the temperature, monomer concentration, catalyst, and modifier levels. Immediately downstream of the reactor, the polymerization is terminated with alcohol or other polar liquids. After the addition of stabilizers and, depending on the polymer type, extender oils, the polymer solution is coagulated and stripped with steam and hot water to remove unreacted monomers and solvents. The rubber crumb is then passed over a screen to remove free water and subsequently dried. Either extrusion or hot air drying may be used. S-SBR, VS-SBR, and SB are sold in the form of bales.

3.1.1.3. *Producers and Production Capacities*

The commercial production of S-SBR started around 1965. The *consumption* increased from 400 000 t in 1980 (16 % of total SBR market) to 440 000 t in 1991 (15 % of total SBR market).

The *production capacity* was ca. 610 000 t in 1996. The S-SBR plant capacity is, in almost all cases, part of a multipurpose anionic solution polymerization facility. Thus, if the demand for S-SBR increases in the future, parts of the multipurpose plants can be used for S-SBR production. Table 16 gives an overview of the S-SBR producers, their S-SBR capacities in 1992, and the capacities of their multipurpose facilities [206].

Table 16. Producers of S-SBR, plant locations, and production capacities (1996)

Company	Location	Capacity *, t	Multipurpose plant	
			Capacity, t	Products
Western Europe				
Bayer	Lillebonne	20 000	75 000	BR, SB
Enichem	Ravenna,	15 000	85 000	SBS, SIS
	Grangemouth	10 000	80 000	BR
Michelin	Bassens	70 000	100 000	BR
Petrochim	Antwerp	20 000	80 000	BR, SBS
Repsol	Santander	40 000	75 000	BR, SBS
Shell	Pernis	5000	30 000	IR
North America				
American Synthetic Rubber	Louisville, KY	60 000	160 000	BR
Firestone	Lake Charles	90 000	150 000	BR. SBS
	Orange	50 000	110 000	BR, SB
Goodyear	Beaumont, TX	20 000	290 000	BR, SIBR, IR
Latin America				
Coperbo	Muncipo d. Cabo (1976)	35 000	110 000	BR, SB, SBS
Negromex	Altamira	45 000	94 000	BR, SBS
Far East				
Asahi Chemical Ind.	Kawasaki	20 000	130 000	BR, SB, SBS
Japan Elastomers	Tsurusaki (Oita)	10 000	48 000	BR, SBS
Japan Synthetic Rub.	Yokkaichi	15 000	25 000	BR
Nippon Zeon	Tokuyama	20 000	78 000	BR, SBS
Sumitomo Chem.	Chiba	10 000		
Hyundai Petrochem.	Daesan	10 000	40 000	BR
Kumho Petrochem.	Yeochon	25 000	150 000	BR, SBS
Taiwan Synthetic	Kaohsiung	16 000	100 000	BR. SBS
Africa				
Carbochem	Newcastle	10 000	36 000	BR, IR

* Estimated.

3.1.2. *Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber*

3.1.2.1. *Properties, Grades, and Applications*

Polybutadiene. produced with lithium initiators is often described as a low- or medium-*cis*-1,4-polymer. When produced in an aliphatic, cycloaliphatic, or aromatic solvent, a polymer of ca. 36 – 38 % *cis*-1,4-, 52 % *trans*-1,4-, and 10 % 1,2-content is produced. The 1,2-content of the polymer can be varied over the range from 10 to 90 % by the addition of polar additives. The higher the 1,2-content, the higher is the T_g of the polymer. Standard lithium polybutadiene of 10 % vinyl (i.e., 1,2-)content has a T_g of – 93 °C, whereas the T_g of a 90 % vinyl polymer can be expected to be ca. – 20 °C [207], [208].

A considerable amount of lithium-initiated polybutadiene is used to produce *high-impact polystyrene* (HIPS) (⟶ Polystyrene and Styrene Copolymers – Polystyrene) [209]. BR is the rubber of choice because it dissolves readily in styrene and gives rubber domains of the appropriate size that inhibit crack propagation. BR is the most efficient impact modifier for polystyrene on the basis of impact improvement relative to rubber level used. Lithium polybutadiene is used in this application because it can be produced with uniformity and high purity. It competes in this application with cobalt polybutadiene produced by Ziegler – Natta catalysis (see Section Polybutadiene and Polyisoprene Rubber). The viscosity of the rubber solution in styrene is important to the HIPS producer. The solution viscosity of the rubber in styrene is a function of the molar mass, molar mass distribution, and branching of the polymer. Various proceses can be used to branch lithium polymers. Small amounts of divinylbenzene (DVB) may be added or the branching process may be carried out thermally. Branching also helps in handling the rubber because linear polymers are very prone to cold flow. Levels of gel in the polymer must be ≤ 200 ppm. A higher 1,2-content of the lithium polymer (10 %) causes faster grafting than that observed with cobalt polymers with a 1,2-content of only 2 %. Because of this the HIPS process must be adjusted to the polymer in use.

Some lithium polybutadiene grades have been used in *tires*, but the production volume of these grades is relatively small. The standard lithium polymer exhibits poor processing behavior and does not give the excellent tread wear that is typical of the high-*cis*-polybutadienes made by Ziegler – Natta polymerization [210]. More recent work to produce lithium polybutadienes for improving tire performance, however, has been relatively successful. Lithium chemistry allows the production of polymers with very specific molecular architecture (branching, end groups, microstructure, copolymerization control, etc.). Specific lithium polybutadiene polymers and solution SBR polymers are now being made to meet the increasing demands of tire manufacturers [211]. Lithium polymers with high vinyl content are used in winter tires and tires with lower rolling resistance.

Specific polar end groups are sometimes added to improve the bonding of the polymer to carbon black. Branched and coupled polymers are made to improve processing and to alter the dynamic properties of the tire. Star-branched lithium polymers are used in very highly carbon black-loaded compounds, such as bead or chafer compounds, because they can easily be compounded with large amounts of carbon black. Copolymers with styrene and, in some cases, isoprene further extend the range of special products available to tire designers. To date, the market volume of such materials is not large (see Section Solution 1,3-Butadiene – Styrene Rubber (S-SBR) and Styrene – Isoprene – Butadiene Rubber (S-SIBR)), but growth of these materials is expected. Functional lithium initiators with an end group that can bond to carbon black are now being studied to further improve rolling resistance.

Lithium polyisoprene. is produced on a small scale for use as a processing aid for natural rubber and in other applications where its high purity is an asset. A tread specialty polyisoprene with a high 3,4-content is also produced, as well as a liquid IR.

3.1.2.2. *Basic Chemistry and Production Processes [212], [226]*

Lithium Polybutadiene. Anionic polymerization offers many possibilities for synthesizing tailor-made polymers. Only the most important aspects of lithium chemistry are covered here. For a broader understanding of the many complex products which can be made using lithium chemistry, see [220]. Polymerization of dienes such as butadiene and isoprene can be initiated by alkali metals or organoalkali compounds. The use of organolithium initiators is advantageous because of their solubility in polar and nonpolar solvents. In the case of *organolithium initiators* (e.g., butyllithium), the initiation step can be regarded as a nucleophilic attack of the initiator on the monomer, generating an allylic organolithium compound, which is able to add the next monomer molecule. Further addition of monomer leads to polybutadiene or polyisoprene chains, which still possess an organolithium unit at the end. Depending on the nature of the solvent and other reaction conditions, the carbon – lithium bond has a more or less ionic character. In sufficiently pure systems at moderate temperature, transfer and termination reactions do not occur to a significant extent. Therefore, the active species retain their reactivity, and the polymerization is often described as living (⟶ Plastics, General Survey – Anionic Polymerizations). In pure systems the polymerizations go to completion. The molar mass of the product can be calculated from the number of moles of monomer divided by the number of moles of butyl lithium. The product has a narrow Poisson distribution, if the initiation is rapid and side reactions are avoided.In the ideal case, the polymer chains are linear. The narrow molar mass distribution and high linearity usually yield polybutadienes that are difficult to process. Thus, introduction of branches or broadening the molar mass distribution can be useful.

Copolymerization of butadiene with small amounts of a branching comonomer such as divinylbenzene yields products that are easier to process and have a reduced cold flow. Partial or quantitative coupling of polymer chain ends [227] can also be used. Common coupling agents are metal halides, such as $SnCl_4$ and $SiCl_4$, or diesters. Moreover, anionic polymerization offers the possibility of using multifunctional initiators and introducing functional groups at the chain end during termination [226], [228]. Of the many functional end groups which can be used to terminate the polymer, two main types have been used to improve tire properties. The use of tin compounds in any form (R_3SnCl, R_2SnCl_2, $RSnCl_3$, or $SnCl_4$) leads to a tin-containing group in the molecule which improves the bonding of the rubber to carbon black [221], [222]. Depending on the number of chlorine atoms in the tin compound, these materials can also introduce branching of the polymer and alter processing characteristics Another class of compounds used commercially to introduce end groups are benzophenone derivatives such as *bis*(4,4'-diethylamino) benzophenone [223]. Many other end groups have been described in the patent literature to be useful for reduction of rolling resistance.

The success at improving rolling resistance by terminal functional groups has prompted considerable research to incorporated similar groups at the initiation end of the molecule [224].

In batch polymerization with a rapid initiator such as *sec*-butyllithium, narrow molar mass distributions can be obtained. In some tire applications bimodel molar mass distributions are desirable. This can be achieved by coupling some of the polymer with, e.g., $SnCl_4$ or $SiCl_4$. To this end, a substoichiometric amount of coupling agents is added, so that some of the polymer is left uncoupled.

In many applications, Li-Br with a narrow molar mass distribution is not needed. Products for these markets are thus often made in a simple continuous process with high polymerization temperature and hexane as the solvent, which results in a broader molar mass distribution.

The *kinetics* of the initiation as well as of the propagation reactions show a first-order rate dependence on monomer concentration that is independent of the solvent used [212], whereas various reaction orders are found with respect to the initiator concentration, depending on solvent type and other parameters. This is a consequence of the different nature of the initiator, the propagating chain end, and the mechanism under different reaction conditions. In nonpolar solvents, the growing polydienyllithium molecules associate in pairs, whereas the addition of polar solvents such as tetrahydrofuran results in partial dissociation of the pairs, and stabilization of the ionic bond by solvation of the lithium ion, leading to charge separation. In addition to the effect of kinetics, the microstructure is strongly affected by the nature of the propagating chain end, by solvation and association, and by the polarity of the alkyl – metal bond; therefore changing the alkali metal in the initiator drastically changes the microstructure. In *nonpolar media*, Li-BR with ca. 90 % 1,4-units is obtained and the *cis* : *trans* ratio increases with decreasing initiator concentration. Temperature does not strongly affect microstructure in nonpolar media. Conventional Li-BR has about 33 – 38 % *cis* content [229] and is therefore called low-*cis*-BR [207]. The content of 1,2-units in the polybutadiene chain is increased by the addition of *polar modifiers*; the level is influenced by the amount and type of polar modifier used and by the polymerization temperature [227]. Considerable attention is needed for the selection of a suitable polar modifier. Some modifiers give rise to side reactions which may prohibit coupling or addition of functional end groups. Amount and cost of the polar modfier may be important, especially if the polymeriization temperature or the desired vinyl level is high. Both conditions necessitate addition of higher levels of modifier. Being polar molecules, modiifiers are of concern as regards hygiene and pollution and may pose problems in solvent reprocessing systems.

Lithium-Polyisoprene. The basic chemistry in anionic polymerization of isoprene via alkyllithium initiation does not differ much from that of butadiene. Since isoprene can be regarded as a butadiene that is methyl substituted in the 2-position, the 1,2- and 3,4-addition products are not identical. Consequently, four different units can exist in the polymer chain. In nonpolar media or with no solvent, very high *cis*-1,4-content is attainable by alkyllithium initiation (e.g., 96 % *cis*-1,4- and 4 % 3,4-units with no solvent and with low concentrations of *sec*-butyllithium). The presence of solvent or an increased initiator concentration can decrease the *cis* and increase the *trans* content [230]. By the addition of polar solvents, polyisoprene with mainly 3,4-units and possibly some 1,2-units can be synthesized [212]. Commercial “low-*cis*” polyisoprene has a *cis* content of ca. 92 % [207] in contrast to the Ziegler – Natta

IR types with higher *cis* content.

Process Description. (see Fig. 12). A typical process for the production of lithium polybutadiene is shown in Figure 12. The polymerization can be carried out in a broad range of aliphatic, cycloaliphatic, or aromatic solvents. The solvent system is chosen to have a solubility parameter appropriate to the polymer. Especially when styrene copolymers are to be made in the same plant, the solubility parameter of the solvent must be sufficiently high to keep the polymer in solution. Simple Li-Br is usually made in mixed hexane isomers, however, in plants where SBR or block copolymers are also produced, it may be necessary to improve the solubility parameters by adding other better solvents such as cyclohexane. Because of the sensitivity of the system to polar impurities, the monomer and solvent streams must be very pure. The feed is typically dried by azeotropic distillations (a) to lower the water content to a few parts per million. The polymerization (b) itself is quite straightforward. The butyllithium initiator is added to the reaction mixture, and polymerization occurs usually going to completion. Different reactor configurations can be used to produce either narrow or broad molar mass distributions as desired, and DVB may be added to increase branching of the polymer.

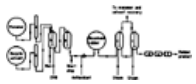

Figure 12. Schematic of a polybutadiene plant

a) Purification towers; b) Polymerization reactors; c) Stripping vessels; d) Dewatering extruder; e) Drying oven; f) Packaging

One of the advantages of lithium chemistry is the ability to polymerize at relatively high temperature where water cooling is adequate and reactions are rapid.

For HIPS application a colorless polymer with a good stability, and a gel level < 200 ppm is required. This means that the polymer must be well protected during the production process. Tradionally, a combination of a hindered phenol as a free radical scavenger and a phosphite as hydroperoxide decomposer has been used as antioxidant. Concerns about the deleterious effects of nonylphenol produced by hydrolysis of trinonylphenyl phosphite has led to the search for new antioxidant systems.

The need for a low gel content requires that the drying and finishing steps be carried out without excessive shear or temperature increase. Effective processes for shortstopping and for extraction of lithium residues are also required, because residues left in the rubber will lead to gel and color formation.

Removal of the solvent and any residual monomer is carried out by contacting the dissolved rubber with steam and water (c). The hydrocarbon is stripped off, and small rubber particles remain suspended in the water. Further contact with steam reduces the content of hydrocarbon in the rubber. Improvement of the stripping process has been necessary because of the need to reduce the pollution of air and water; this has been accomplished with better contactors and additional stripping vessels. The slurry of rubber is then dewatered (d), usually with an extruder that squeezes out most of the water. A final drying stage in a fluidized bed or tunnel dryer (e) completes the process, and the rubber is baled and packaged (f).

Earlier processes for removing solvent directly with drum dryers have now largely been abandoned. A thin layer of polymer cement was dried on a large heated drum but this process is no longer used extensively because of the inherently high solvent losses.

3.1.2.3. *Producers and Production Capacities [229], [231]*

Polybutadiene rubber. is the second largest synthetic rubber product after SBR: In 1995, 1 402 000 t was produced, exclusive of Eastern Europe and the socialist countries of Asia. Worldwide annual BR capacities were about 2 3280 000 t at the end of 1996, as estimated by the International Institute of Synthetic Rubber Producers. This figure is, however, difficult to interpret, because most of the capacity can be used to produce BR, SBR, and block copolymers. Special BR types — e.g., polymers with special end groups, liquid polymers, high-vinyl polymers — are produced only in small amounts.

The technology for the preparation of Li-BR was developed by Firestone Tire & Rubber Co. [208], [232] in the 1950s; Firestone is still proucing Li-BR with the trade name Diene. Among the other producers of low- and medium-*cis*-BR, and in some cases of high-vinyl types as well, are the following producers listed in Table 17 (with trade names).

Table 17. Producers of polybutadiene rubber

Company	Country	Trade name
Asahi Chemical	Japan	Asadene
Bayer	France	Buna
Bayer	United States	Taktene 550
Coperbo	Brazil	Coperflex BR
Enichem	Italy	Intene, Intolene
Firestone	United States	Diene
Japan Elastomer	Japan	Asaprene
Karbochem	South Africa	Afdene
Kumho	Korea	
Negromex	Mexico	Solprene
Petrochim	Belgium	Finaprene
Repsol	Spain	Calpene

Isoprene Rubber. According to IISRP statistics, 116 000 t of synthetic isoprene were produced in 1995 outside the Eastern bloc countries. Most of this is high-*cis* material produced by Ziegler – Natta catalysts. Shell Nederland Chemie produces Li-IR under the trade name Cariflex [233]. A 3,4-polyisoprene previously produced by Hüls (Vestogrip) is now made by Karbochem under the trade name Isogrip. Kurarary produces a small amount of liquid IR.

3.2. Synthesis by Ziegler – Natta Polymerization

3.2.1. *Polybutadiene and Polyisoprene Rubber*

3.2.1.1. *Properties, Grades, and Applications*

High-cis-Polybutadiene. High-*cis*-polybutadiene is the second largest volume synthetic elastomer, second only to emulsion SBR: It is used primarily in the manufacture of tires. Moderate amounts are also used in high-impact polystyrene and ABS. In these markets it competes with lithium polybutadiene (see Section Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber). Small amounts of high-*cis*-polybutadiene (2 – 3 %) are used in other applications, such as golf balls, hose roll covers, and conveyor belts.

The principal characteristic of this polymer is the high content of *cis*-1,4-microstructures, which is generally 92 – 98 % (Table 18). The *cis*-1,4-polymer chain has a very low T_g (– 10 °C). This leads to high resilience, good low-temperature properties, and low heat buildup on repeated deformation [238-240]. These characteristics are needed in tire treads, carcass stocks, and sidewalls [241]. High *cis*-BR provides the extremely rapid chain motions needed for abrasion resistance and resilience of tires. In the tread portion of the tire, *cis*-BR is blended with rubbers with a higher T_g, such as SBR, to improve traction.

The differences in microstructure shown in Table 18 do not decisively affect product performance. The vinyl group has a higher T_g (≈ – 106 °C), but since

the vinyl content is only 0 – 4 %, it has only a small effect on the overall glass transition temperature. The higher level of *cis*-1,4-microstructure seen when neodymium is used as catalyst seems to improve fatigue properties in some applications. The higher linearity of Nd-Br polymers may also contribute to this result.

Table 18. Dependence of microstructure of Ziegler – Natta polybutadiene on catalyst used

	Cis content, %	*Trans* content, %	Vinyl content, %	Molar mass distribution	T_g, °C
Cobalt	97	1	2	moderate	– 106
Titanium	93	3	4	moderate	– 103
Nickel	97	2	1	broad	– 107
Neodymium	98	1	1	very broad	– 109

High *cis*-BR polymers show significant variation in processing characterstics because of differences in molar mass, molar mass distribution, and degree of branching. These factors vary even within polymers produced by the same catalyst system, so overall conclusions about preferred catalyst systems are somewhat difficult to reach. On balance, the *nickel-catalyzed* polymers, with their broad molar mass distribution and slightly higher degree of branching are somewhat easier to process (better mill compound behavior and extrudability), but their tensile properties and fatigue resistance are inferior. *Titanium-catalyzed* polymers, with their lower branching and moderate molar mass distribution, are more difficult to process but better in tensile strength and fatigue resistance. Polymers synthesized with *cobalt catalyst systems* can be produced with a varying degree of branching. The branched grades are easy to process, whereas the more linear polymers have better dynamic mechanical properties. Polymers made with *neodymium systems* are highly linear and have a broad molar mass distribution. This gives them a desirable balance of properties with the exception of poor extrudability and high cold flow. The processing properties of some neodymium-catalyzed polymers are improved by “molar mass jumping” as described below.

Most high-*cis*-BRs do not have the high tack and green strength of natural rubber. When the products were developed, extensive work was carried out to determine the reasons for this and to improve the product, but a commercially attractive route does not seem to have been found. Polymers produced with neodymium systems are somewhat improved in this regard, but still not to the level of natural rubber.

High-*cis*-BR is also prone to cold flow (i.e., the tendency of unvulcanized rubber to flow during storage). In extreme cases, the bales of high-*cis*-butadiene rubber are completely distorted to the point that the film wrap splits. Cold flow is most serious for low molar mass or highly linear materials. Long-chain branching is introduced into some products to reduce cold flow. This can be carried out by postpolymerization reactions, sometimes termed “molar mass jumping”, that involve reacting the polymer in solution with disulfur dichloride.

As noted above, high-*cis*-BR has relatively poor processing properties. In many cases, this is due to stress relaxation, which is too fast. Slower relaxations as a result of branching or molar mass jumping improves processing on mills. Processing problems with high *cis*-polymer are diminished in compounds of higher *cis*-BR where either SBR or natural rubber is also present.

As with all polymers, mechanical properties are a function of molar mass. Tensile strength, abrasion resistance, and fatigue resistance all improve as the molar mass of the polymer increases. Balanced against this, of course, is the greater difficulty of mixing and processing higher molar mass polymers.

Oil extension with naphthenic or aromatic oil is used to make higher molar mass products processible. One producer of high-*cis*-BR also makes a polybutadiene black masterbatch. In this case, both carbon black and oil are added to the polymer while it is still in solution. This facilitates subsequent processing and allows high molar mass polymers to be handled with relative ease.

As mentioned previously, the degree of branching of these polymers can vary. Relatively little is known about the nature of this branching or about its cause. However, the branching differences in the products can be observed fairly easily in dynamic mechanical or stress relaxation tests.

Another important characteristic of some grades of polybutadiene is the gel content. In tire applications this is not generally a concern. However, BR used in high-impact polystyrene must have very low gel levels. It is primarily the cobalt-catalyzed BR that is used to produce HIPS (the other product used is lithium-catalyzed BR; see Section Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber). The neodymium catalyst system may be capable of producing HIPS-quality BR, but unsatisfactory control of solution viscosity has so far kept it out of this application.

In addition to the low gel content, HIPS-BR must also contain low residues of catalyst, have little color, and contain a suitable type and level of antioxidant so that the HIPS polymerization is not retarded. The most common antioxidant system is a combination of a hindered phenol and a phosphite. Concern about the adverse health effect of free nonylphenol formed by hydrolysis of trinonylphenyl phosphite has led to the evaluation of new stabilizers that do not contain trinonylphenyl phosphite.

Although the content of 1,2-units is only 0 – 4 %, it can still influence the HIPS process by affecting the rate of grafting of the polystyrene on the rubber. The solution viscosity of the rubber in the styrene monomer is also of considerable importance because it determines how, and when, the rubber phase breaks up to yield rubber particles. The size and nature of the rubber particles are very important in determining the impact properties of HIPS (see also ⟶ Polystyrene and Styrene Copolymers – Introduction).

High-cis-Polyisoprene. Polyisoprene produced with a Ziegler – Natta catalyst attracted considerable interest when it was developed and commercialized in the 1960s. The polymer has a microstructure very similar to that of natural rubber, but it has inferior green strength and tack [241], [242]. Considerable research has been carried out to determine why high-*cis* polyisoprene does not duplicate all the properties of natural rubber. Although natural rubber contains almost 100 % *cis*-1,4-structures, the microstructure is not the only important difference between the synthetic and the natural products. Functional groups on the natural product and nonrubber components are also important in influencing green strength, tack, and tensile properties. Unfortunately, the availability and cost of isoprene have remained a problem. Isoprene synthesis is more difficult than that of butadiene, and its purification is also more complicated (⟶ Isoprene – Production). Isoprene containing C_5 streams from cracking operations are used to the extent available. As a result, both the monomer and the polymer are relatively expensive. The availability of natural rubber at reasonable cost has made the synthetic polymers unattractive, and a number of plants in the Western World have been closed. Some high-*cis*-polyisoprene is still produced by Goodyear JSR and Nippon Zeon. It is used where its high purity is an advantage and also as a processing aid for natural rubber. Shell also produces a polyisoprene in the Netherlands using a lithium catalyst.

Substantial volumes of Ziegler – Natta polyisoprene were produced in Eastern Europe. However, most of the production capacity cannot be operated economically and has been shut down. The operation of such facilities depends on natural rubber pricing.

Other Commercial Ziegler – Natta-Catalyzed Polydienes.
Trans-1,4-Polyisoprene. A Ziegler – Natta system is used by Kuraray (Japan) to produce *trans-1,4-polyisoprene* that is similar in molecular structure and properties to the natural product balata. Because of its low melting point the crystalline product can be used for splinting or prosthetic applications. Sheets of the product are softened in an oven or with hot water and then molded to the body and allowed to cool. *Trans*-1,4-polyisoprene is also used for golf ball covers, where it competes with ionomers. The process used to make the *trans* polymer gives a lower yield of product, and the relatively high price of the product limits it to low-volume specialty applications.

Syndiotactic 1,2-Polybutadiene. Another low volume polymer produced by Ziegler – Natta catalysis is syndiotactic 1,2-polybutadiene (JSR BR). This partly crystalline product is a thermoplastic elastomer with high purity. It is produced with only 15 – 30 % crystallinity to make it flexible and processible, and is used as a high-permeability film in food packaging. It is also used as a dispersible film wrap for rubber and rubber chemicals, but in this application it faces stiff competition from much less expensive films, such as polyethylene and ethylene – vinyl acetate copolymer films.

Research on tires is devoted to reduction of rolling resistance without detracting from wear or traction properties. This has resulted in a trend toward S-SBR instead of E-SBR (see Section Emulsion Styrene – Butadiene Rubber (E-SBR)). Tread compounds are also increasingly reinforced with silica instead of carbon black. In the endeavor to improve rolling resistance the levels of high-*cis*-BR in tread compounds were also increased. Some research on novel Ziegler – Natta polymers such as butadiene – isoprene copolymers has been carried out. Thes products have had little commercial success.

3.2.1.2. *Basic Chemistry and Production Processes*

High-cis-Polybutadiene. Ziegler – Natta polybutadiene was first produced by Phillips Petroleum in 1956 using titanium catalysts [243]. This was followed very rapidly by development of cobalt and nickel systems [244-246]. The ability of these catalysts to produce high-*cis* polymers with properties similar to those of natural rubber led to rapid commercialization of both high-*cis*-polybutadiene and high-*cis*-polyisoprene. The latter business dwindled due to the lack of acceptably priced monomer and to competition from relatively low-cost natural rubber. In contrast, the production of polybutadiene has benefited from a ready supply of cheap monomer. Technical developments in the field have been fairly modest since the 1980s, but more recent work on rare-earth catalyst systems has revived interest in product and process improvement. Very good early studies have been carried out by GIPPIN [247], [248]. Reviews of the technology have been published [249-251].

Ziegler – Natta polybutadienes are produced by solution processes using a variety of solvents. Some, such as the titanium process, are best carried out in aromatic solvents, whereas others, such as the neodymium-catalyzed process, can be carried out only in aliphatic or cycloaliphatic solvents. As indicated below, each catalyst system has its own unique features. However, the various Ziegler – Natta systems share common elements. All are sensitive to impurities, thus monomer and recycle solvent streams must be well purified. Carbonyl and acetylenic impurities in the monomer must be at the level of a few parts per million, and the polymerization inhibitor (*tert* butylcatechol) is usually removed from the monomer prior to reaction.

The feed streams of monomer and solvent are usually dried by azeotropic distillation and fed to a series of polymerization reactors. Ziegler – Natta catalysts are, commonly, three-component systems, and both the order of addition and the mixing conditions can be important. The catalyst components may be added directly into the first reactor or some components may be added into a premix vessel ahead of the reactor or into the feed line. Various methods are used to control the molar mass of the products. Molar mass modifiers specific to the catalyst system are used where necessary. The final conversion of ca. 70 – 99 % is achieved in a polymerization time of 2 – 3 h.

After polymerization a short stop is added. Water, fatty acids, or alcohols, can be used to terminate the reaction. Antioxidants are added, and steam and water are brought into contact with the polymer solution to strip off the solvent and any residual monomer. This coagulation and stripping operation leaves the rubber in the form of small particles floating in water. The water is drained off through a screen, and the material is dewatered, usually in an extruder. The final drying is carried out in a fluidized-bed dryer, tunnel dryer, or twin-screw drying extruder.

Good operation of the solvent removal step is important to minimize air and water pollution. Solvent loss of 1 or 2 wt %, based on the weight of polymer, occurred in older processes. Improved stripping conditions, additional stripping vessels, and other changes have been implemented to reduce losses of solvent to ≈ 0.3 wt %. Work has been carried out to improve engineering practices throughout the plants in order to reduce losses at valves, flanges, and pumps. In some cases, air from these sources is ducted and used as boiler combustion air to prevent losses to the environment. There has also been some concern regarding exposure to butadiene monomer, and the improved stripping mentioned above is important in minimizing this exposure.

Co-BR. Polymerization to form high-*cis*-BR, based on cobalt catalysts, was disclosed in early patents to B. F. Goodrich [245], and this technology was later licensed to other companies. Various cobalt compounds can be used. Because $CoCl_2$ is insoluble in the aromatic solvent used, work was carried out to use soluble cobalt salts such as the acetylacetonate, the octoate, or the versatate. These materials are readily available and are easy to handle in solution.

The other catalyst components are an alkylaluminum compound and small amounts of water. Various alkylaluminum compounds can be used. The most common material is diethylaluminum chloride [$(C_2H_5)_2AlCl$], but aluminum compounds containing more chlorine, such as the sesquichloride [$(C_2H_5)_3Al_2Cl_3$], can also be employed. Catalysts with lower chlorine content have reduced activity.

The amount of water used in the system is very important. With *diethylaluminum chloride,* very little polymerization occurs in the absence of water. As the water content is increased, the reaction becomes more rapid, then passes through a maximum, and again declines to zero at higher concentration. As the water level is increased, so does the solution viscosity of the polymer at a given Mooney viscosity. This increased solution viscosity results in improved dynamic properties and is commonly interpreted as a lower degree of branching.

Polymerizations carried out with *alkylaluminum* compounds containing more *chlorine* in the molecule than diethylaluminum chloride can be used without adding water, but in practice, this method generally does not offer any advantages.

Polymerization is usually carried out at 5 – 35 °C, depending on the solvent and the heat removal capability of the reactor. Early work was based mostly on benzene as the solvent. However, with increased concern about loss of solvent to air or water, work has been carried out to find other suitable solvents. Cyclohexane can be used in place of benzene, although side reactions leading to gel formation or to fouling of the reactor may be more pronounced. Aromatic solvents with methyl substituents can be added in small amounts. These substituted aromatics stabilize the catalyst complex but also reduce the reaction rate.

The molar mass of the polymer produced can be controlled by monomer concentration or by the addition of modifiers, such as hydrogen, 1,2-butadiene, propadiene, or cyclooctadiene. α-Olefins, such as ethylene, propene, 1-butene, or 1-hexene can also be used to reduce the molar mass of the polymer.

Cobalt polybutadiene is often used in high-impact polystyrene. HIPS used for extrusion applications must contain less than 200 – 300 ppm of gel to prevent visible defects in the HIPS. To meet such stringent conditions, a catalyst system must be selected that does not favor side reactions, and the feed must be very pure. Extra care is required to keep the product free of gel particles. The polymer solutions are filtered after polymerization, and wash steps are incorporated to remove catalyst residues. Extra care imust be taken to ensure that the product is not degraded.

The reaction rates of cobalt-catalyzed polymerization are generally proportional to the concentrations of both cobalt and monomer. With increasing cobalt concentration the molar mass of the polymer decreases and the polymer becomes more branched.

As mentioned above, the rate of polymerization is a complex function of the amount of water present. The water concentration also influences molar mass, branching, and to a small degree, the microstructure of the resulting polymer. The degree of branching decreases with increasing water level. Gel levels in the product also increase with the amount of water charged. It does not appear, therefore, that gel forms because branching has increased to the point at which a network can form.

Cobalt-catalyzed polymerization is usually carried out with very low levels of cobalt. Many chains are formed by each cobalt atom. In contrast, the alkylaluminum is present in large excess. Water levels in the range of 0.1 – 0.5 mol per mole of alkylaluminum compound are generally used.

The mechanism of polymerization is only partially understood [252]. The functional groups attached to the aluminum and cobalt appear to be exchanged, and cobalt is reduced from Co^{2+} to Co^{+}. Coordination of butadiene to the catalyst complex is involved in the stereoregulating mechanism. When aromatic solvents are present, they can form strong complexes with the catalyst and displace the monomer as a ligand. With substituted aromatics, the catalyst complex can become so stable that little polymerization occurs.

Titanium-Catalyzed Polybutadiene. The Ziegler – Natta polybutadiene process developed by Phillips Petroleum uses a catalyst of titanium tetraiodide and triisobutylaluminum [253]. The low solubility of the titanium tetraiodide in the hydrocarbon solvent led to production difficulties, and numerous other systems were developed to improve handling and product uniformity. The common element in the various systems is the presence of iodine with the eventual formation of a titanium – iodine bond [254-257]. Catalysts that contain other halogens do not yield the desired high content of *cis* microstructure in the polymer and are prone to other difficulties such as the formation of gel and low molar mass oligomers.

A precipitate forms during these polymerizations, which further complicates the process. Extra care must be taken to bring the catalyst species together in a desirable and reproducible way.

Polymerization is usually carried out in benzene or toluene. However, mixed aromatic – aliphatic solvents or even totally aliphatic systems can be used with some difficulty. Polymerization can occur over the range 5 – 50 °C.

The efficiency of titanium catalyst systems is influenced by the aluminum – titanium ratio. The *cis* content, molar mass, molar mass distribution of the product, and catalyst activity are all influenced by this ratio. The optimum ratio depends on the system in use.

Nickel-Catalyzed Polybutadiene. Many Ziegler – Natta catalyst systems, based on nickel, have been disclosed in the literature. Those containing fluorine generally give a high content of *cis*-1,4-microstructure in the polybutadiene. These catalysts were developed by Bridgestone [246], [258], [259] and studied in other laboratories [260], [261]. The most common catalyst system consists of a nickel carboxylate with an alkylaluminum and BF_3 etherate or HF. These systems can give *cis*-1,4-contents of up to 98 % in the polymer, along with good reactivity at low levels of nickel and relative freedom from gel. As in most of the Ziegler – Natta catalyst systems used for production of polybutadienes, the molar mass of the resulting polymer is inversely proportional to catalyst concentration. Allenes, acetylenes, and vinylcyclohexane all retard polymerization and reduce molar mass. The molar mass decreases with increasing polymerization temperature. Most polymerizations described in the literature are carried out in either benzene or mixed aromatic – aliphatic solvents.

Neodymium-Catalyzed Polybutadiene. High-*cis*-polybutadiene can be prepared using some of the lanthanide elements as catalysts [262]. A basic patent was filed by Union Carbide in 1963 [263]. Considerable research work on the system has been carried out by Russian and Chinese scientists [264]. The preferred system seems to be based on neodymium. Commercial products are now made by Bayer and Enichem. Rare-earth systems are also in use in South Africa and China. The catalyst system consists of rare-earth carboxylates with alkylaluminum and a source of chloride, such as diethylaluminum chloride [265-267]. Bulky carboxylates are preferred to improve catalyst solubility. The reactivity is a function of the carboxylate chosen and of the chloride: neodymium ratio. The chain-transfer agents used in other Ziegler – Natta catalyst systems (olefins, hydrogen, allene, etc.) are inactive in this system [268]. In the neodymium system the molar mass is varied by using a hydride such as diethylaluminum hydride or diisobutylaluminum hydride.

The neodymium polybutadiene system gives rapid polymerization to achieve very high conversions. Monomer concentrations of 10 – 20 wt % are generally used. The upper concentration limit depends on the ability of the equipment to handle the highly viscous solution and on the efficiency of heat removal. The polymers obtained have a very broad molar mass distribution and thus a high solution viscosity.

The neodymium catalyst system does not tolerate aromatic solvents. Polymerization can be carried out in a variety of aliphatic and cycloaliphatic solvents, generally at 50 – 90 °C. The molar mass of the product is inversely proportional to the reaction temperature, and the *cis* content of the polymer declines slightly with increasing temperature. The system is not sensitive to water at part-per-million levels as are some other Ziegler – Natta systems.

High-*cis*-BR has been improved incrementally, but the basic product has not been altered significantly for many years. Most of the research has been concerned with process improvement to increase yield and reduce cost and environmental hazards. Research is directed toward solvent-free, so-called “gas phase” processes, whereby polymer particles fluidized in monomer vapors are formed. This technology is used for production of polyethylene and polyproplyene (⟶ Polyolefins – Gas-Phase Process, ⟶ Polyolefins – Gas-Phase Processes). Its application to production of high-*cis*-BR has been studied in a number of laboratories.

Syndiotactic 1,2-Polybutadiene. Syndiotactic 1,2-polybutadiene is produced only by Japan Synthetic Rubber in small amounts [269], [270]. This partly crystalline material is produced by using a Ziegler – Natta polymerization process. Although a 99.7 % 1,2-content can be achieved, the product is deliberately produced with only ca. 90 % 1,2-content to reduce the crystallinity to 15 – 30 %.

The polymer is produced with a cobalt catalyst and an alkylaluminum compound. Triphenylphosphine is added to give the syndiotactic 1,2-microstructure [271].

High-cis-Polyisoprene. Titanium catalysts are often used to make high-*cis*-polyisoprene [242]. Titanium tetrachloride together with trialkylaluminum was used in early work, and the catalysts currently in use appear to be very similar [244], [272-274]. This catalyst is not suitable for the production of high-*cis*-polybutadiene, but it can produce a *cis*-1,4-polyisoprene with a *cis* content as high as 97 % [275]. Isoprene is more easily polymerized than butadiene to form polymers with high *cis* content. Even lithium-catalyzed polymers have ca. 93 % *cis*-1,4-microstructure (see Section Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber).

The conditions under which the Ziegler – Natta catalyst components are brought together are quite important. In the preferred process, AlR_3 is slowly added to $TiCl_4$ in an aliphatic solvent at low temperature. Ethers and amines may be used to reduce gel and oligomer formation. The optimum molar ratio of titanium to aluminum is about one. This gives a good reaction rate and minimizes problems of oligomer formation. Molar mass can be influenced by controlling monomer concentration and temperature.

High-*cis*-polyisoprene and butadiene – isoprene copolymers can also be produced with *neodymium catalysts.* This may prove to be of interest because of its potential to produce improved polybutadiene products with low levels of the higher-cost isoprene monomer.

High-trans-1,4-Polyisoprene. Highly crystalline *trans*-1,4-polyisoprene can be produced by using various vanadium, titanium, or mixed vanadium – titanium catalysts [275], [276]. The VCl_3 system has been used commercially [277] because it produces less material with low molar mass and gives a higher overall *trans*-1,4-content. The process is usually performed at 20 – 90 °C. The addition of ethers or $Ti(OR)_4$ is claimed to increase the polymerization rate, which is otherwise slow [276], [278]. High solution viscosity necessitates operation at lower solids levels and this limits the productivity of the system. Aliphatic solvents are preferred because the solution viscosity is reduced in this case.

3.2.1.3. *Producers and Production Capacities*

High-cis-Polybutadiene. Production data on BR including high-*cis*-polybutadiene are given in Section Lithium – Butadiene (Li-BR) and Lithium – Isoprene (Li-IR) Rubber.

Of the high-*cis*-BR technologies, the *cobalt process* is most widely employed. B. F. Goodrich technology was licensed during the 1960s and 1970s in Canada, the United States, the Federal Republic of Germany, Japan, and Australia. Other plants were later built in Turkey, India, the People's Republic of China, and France. This technology has since been refined to produce *cis*-BR with low gel grades for productions of HIPS and to control branching. The more linear *titanium-catalyzed* polymers were first developed by Phillips. These are now produced in France, Germany, Japan, and the United States. *Nickel-catalyzed polymers* are produced in the United States, the People's Republic of China, and Japan. A more recently developed process is based on *neodymium*. High-*cis*-polybutadiene is now produced in the United Kingdom and Germany by this method.

Producers and trade names are listed in Table 19

Table 19. High -*cis*-polybutadiene producers

Company	Country	Trade name
American Synthetic Rubber Co.	United States	Cisdene
Australian Synthetic	Australia	Austrapol CB
Bayer	United States/ Canada/ Germany	Buna CB, Taktene

Bridgestone/ Firestone	United States	Duradene
Enichem	UK/ Italy	Europrene Neocis
Goodyear	United States	Budene
Japan Synthetic Rubber	Japan	JSR BR
Hyundai Petrochemical	Korea	
Indian Petrochemicals CL	India	Cisamer
Karbochem	South Africa	Neodene
Kumho Petrochemical	Korea	Kosyn
Michelin	France	
Nippon Zeon	Japan	NIPOL
Petkim	Turkey	Petcis
Petrochim	Romania	
Taiwan Synthetic Rubber	Taiwan	Taipol
Ube Industries	Japan	Ubepol
Voronezh SR	Russian Republic	SKD
Yefremor	Russian Republic	SKD

High-cis-Polyisoprene. Worldwide production of polyisoprene is presently difficult to estimate and subject to rapid changes. Most of the capacity is in Eastern European but the economics of these plants are marginal now and much of the capacity is shut down. The other major economic factor is the price of natural rubber which can fluctuate widely. Current producers and trade names are given in Table 20.

Table 20. High -*cis*-polyisoprene producers

Company	Country	Trade name
Goodyear	United States	Natsyn
Japan Synthetic Rubber	Japan	JSR
Nippon Zeon	Japan	NIPOL IR
Petkin	Turkey	Petcis
Petrochim	Romania	
Shell Nederland	Netherlands	(Li polymer)
Voronezh SR V/O Raznoimport	Russian Republic	SKI
Yefrenov V/O Raznoimport	Russian Republic	SKI
Shanghai GAO DIAO	People's Rep. China	

Syndiotactic 1,2-polybutadiene. is produced by Japan Synthetic (trade name JSR RB).

3.2.2. *Ethylene – Propene Elastomers (EPM, EPDM) [279-281]*

3.2.2.1. *Properties, Grades, and Applications*

Ethylene – propene elastomers are a family of random elastomeric copolymers with a completely saturated polymer backbone. These elastomers are therefore designated with an “M” according to the ASTM standard DIN/ISO 1629.

There are two basic types of ethylene – propene elastomers:

1. Ethylene – propene copolymers (EPM)
2. Ethylene – propene – diene terpolymers (EPDM)

The *copolymers* are generally amorphous elastomeric polymers without unsaturated bonds, thus necessitating the use of peroxide cure systems for vulcanization. The *terpolymers* contain small quantities (< 12 wt %) of a nonconjugated diene. Only one of the double bonds of each diene molecule takes part in polymerization, resulting in a saturated polymer backbone with a pendant double bond that facilitates the use of conventional sulfur curing.

Properties. The saturated carbon backbone structure accounts for the following features:

1. Outstanding ozone resistance
2. Excellent weathering stability
3. Excellent heat resistance (< 175 °C)
4. Good insulating properties
5. Low-temperature flexibility (glass transition temperature: ca. – 55 °C)
6. Good chemical resistance (toward polar organic and inorganic compounds)
7. Good moisture and steam resistance

EPM and EPDM have the lowest density (0.86 kg/L) of any commercial elastomer. In addition, during compounding, EP(D)M can be heavily loaded with oil and fillers resulting in low compound costs.

The structure of the polymer can generally be described by three features:

1. Ethylene: propene ratio and monomer sequence distribution
2. Molar mass and molar mass distribution
3. Content and type of diene incorporated

These structural features are directly related to the polymerization conditions. By varying the conditions, polymers with different properties can be obtained and a large selection can be provided, suitable for various applications.

As ethylene content increases:

Stress – strain properties improve
Green strength improves
Compression set deteriorates

Low temperature flexibility deteriorates
Hardness increases

As molar mass increases:

Stress – strain properties improve
Green strength improves
Extendability improves
Compression set improves

As diene content increases:

Stress – strain properties improve
Compression set improves
Cure rate increases
Heat aging resistance declines

The diene type (see Section Basic Chemistry and Production Processes) influences compound properties as follows: Ethylidenenorbornene (ENB) gives the fastest cure rate; dicyclopentadiene (DCPD) the best processability; and 1,4-hexadiene (HD) the best heat resistance.

The EP(D)M grades (ca. 250 available worldwide) are constantly optimized to fit specialty applications. Although many manufacturers produce particular grades which are reasonable substitutes for one another, EP(D)M elastomers are not standardized as are general-purpose rubbers. Listings of grades are summarized in [282].

Specifications. The amount of information provided by manufacturers in their specifications differs. Mooney viscosity — the basic qualitative index of molar mass and processability — is always specified. Mooney viscosity is usually expressed at 100 °C, 125 °C, or 150 °C. Higher temperatures are used for accurate Mooney measurements of higher molar mass grades with a Mooney viscosity > 80 – 100 at 100 °C. There are no absolute measurements of molar mass. Each manufacturer quotes his own set of numbers. The conditions under which the number is measured must be clearly specified in all cases. Oil content, if any, is given in phr. Some manufacturers specify ethylene content (40 – 80 wt %), diene content (0 – 12 wt %), iodine number, and maximum residues (e.g., 50 ppm transition metal, 0.3 % ash, 0.5 % water or volatiles content). Performance of a standard compound is usually given, including cure characteristics (delta torque, cure times), tensile strength (4 – 22 MPa). Elongation at break (150 – 800 %), and compression set (5 – 20 % after 22 h at 70 °C). Density varies according to diene content and crystallinity, but is in the range 0.86 – 0.90 kg/L.

Storage and Transportation. The rubber is generally shipped in boxes containing bales of 25 to 34 kg, multiunits of bales, or bags of pellets for certain semicrystalline grades. Rubber processors prefer loosely packed bales, whereas plastic blenders prefer pellets.

Uses. As a result of their versatility, EP(D)M polymers find use in a number of market segments, both in the traditional thermoset areas, and increasingly in thermoplastic blends and alloys. Typical applications for EP(D)M are given in Table 21.

Table 21. Applications of EPM and EPDM (1991)

Market segment	Market breakdown	Typical applications
Automotive	32 %	radiator/heater hoses, weatherstrips, window channels, sponge seals, mounts, car mats
Plastics modification	15 %	impact-modified polypropylene, thermoplastic olefins
Construction	14 %	roofing membrane, window seals
Industrial	11 %	steam and garden hoses, gaskets O-rings, mounts
Oil	10 %	oil viscosity improver
Electrical	6 %	cable jacketing, insulation, cable filler, connectors
Tire	4 %	sidewall construction, inner tubes, tire valves
Miscellaneous	8 %	footwear, carpet underlay

Uses in Tires. In the 1960s, EPDM was expected to become the new tire elastomer. But until recently, its utilization for tire applications has been limited to some inner tubes and to blends for white side-walls. The main deficiency of EPDM is inability to develop adhesion and to covulcanize with unsaturated polymers such as natural rubber, or polybutadiene.

The physical properties of blends with highly unsaturated rubbers are poor, due to a lack of co-cure between the polymers. The conventional explanation is that the unsaturated polymer is overcured before the EPDM phase in the blend can achieve a high state of cure. This is due to the very different cure rates of the two polymers. EPDM is generally much slower than highly unsaturated rubbers [283]. The physical properties of blends can be improved by modifying the EPDM with, for example, maleic anhydride [284], but this approach is too costly for tire applications.

In *blends with NR*, excellent physical properties and adhesion can be achieved when the EPDM has high molar mass and high (> 8 %) ethylidenenorbornene content [285], [286]. This concept is commercially being used in veneers for radial passenger tires. The veneer is resistant to weathering, since the EPDM phase is immune to ozone attack, and the NR phase domains are so small that cracks cannot easily grow. No chemical antidegradants are required in the formulation, so discoloration from staining antioxidants is avoided [287]. Development of this technology for tire sidewalls and other tire components is underway in tire companies worldwide.

3.2.2.2. *Basic Chemistry and Production Processes*

Monomers. The monomers ethylene and propene are normally used as polymerization grade (purity > 99 %) and must be free of impurities such as carbon monoxide, oxygen, water, acetylene, methanol, or methylacetylene (content < 10 ppm). These impurities are potent poisons of Ziegler – Natta catalysts, which are exclusively used for EP(D)M polymerization [281], [288].

Termonomers. The ideal termonomer has to meet several important criteria. It should have two double bonds of widely different reactivity. One double bond should react rapidly in the terpolymerization, while the second one should be relatively inert. However, this second double bond must be reactive toward the vulcanizing agent. The monomer should also be volatile, so that the residual unreacted monomer can be easily removed from the polymer after polymerization. Finally, the termonomer should be readily available and inexpensive. However, no termonomer meets all criteria. Many compounds have been investigated as possible termonomers [279]. Most research has been restricted to nonconjugated dienes as termonomers since they have minimal influence on copolymerization, yield normal molar mass distributions, and only introduce unsaturation in side chains. The different termonomers commercially used for EPDM are shown in Figure 13.

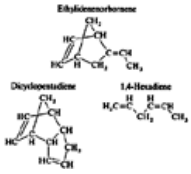

Figure 13. Termonomers used for production of EPDM

Ethylenenorbornene is the most commonly used termonomer because it reacts rapidly in the polymerization and gives high quality, fast curing elastomers. ENB polymers cure faster than polymers containing other termonomers. This is essential for high productivity in continuous curing extruded applications.

Dicyclopentadiene is used as a termonomer due to its low cost and its high reactivity in the polymerization. The main deficiency of DCPD terpolymers is their slow cure rate relative to terpolymers containing ENB. Another disadvantage is that DCPD can be easily cracked to cyclopentadiene (CPD) during stripping and monomer recovery operations. Catalyst efficiency is severely reduced, if CPD levels build-up in the process streams.

There are some similarities between DCP and ENB. Both molecules contain a seven-membered dicyclic norbornene ring in which one double bond is a member of a highly strained five-membered carbon ring. This double bond enters rapidly into the polymerization. The exocyclic ethylenic double bond in ENB is sterically hindered so it does not react during polymerization. In DCPD, the second double bond is not “protected” by steric hinderance. It can, therefore, take part in the polymerization leading to branching, cross-linking, and possibly gel formation. The processability of DCPD terpolymers is improved in comparison with ENB polymers because of the broadening of molar mass distribution as a result of branching.

1,4-Hexadiene (HD) is exclusively used by Du Pont. It is a linear nonconjugated diene. Polymerization favorably occurs via the 1,2 position because of the steric hindrance of the 4,5 double bond. In contrast to ENB and DCPD rings, HD is less reactive in polymerization due to the lack of bond strain. 1,4-Hexadiene has a low boiling point (66 °C) and is not separated from the solvent during the solvent recycle. The cure rate of EPDM using HD as the termonomer is slower than with EPDM using ENB and is comparable with DCPD as termonomer. Due to the relative absence of side reactions, compounds based on HD polymers show the best heat resistance of all EPDM vulcanizates. HD polymers are most suitable for post modification, e.g., reaction with maleic anhydride.

Catalysts. Commercial EP(D)M polymerization processes employ vanadium or titanium catalysts. Metallocene catalysts, although commercially used for polyolefin polymerization (⟶ Polyolefins – Coordination Catalysis) are currently in the pilot plant stage for EP(D)M polymerization [289-291].

Vanadium compounds are used more widely than titanium compounds as catalysts. Vanadium compounds are soluble in hydrocarbons and form an essentially homogeneous system, in addition the elastomeric polymers produced by these catalysts contain no or a minimal amount of crystalline, homopolymer fractions. However, vanadium catalysts are less active than titanium or metallocene catalysts (see Production Processes.).

The most widely used commercial vanadium compounds are vanadium oxytrichloride, vanadium tetrachloride, and vanadium trisacetylacetonate.

There are also many different aluminum alkyls which can be used as *cocatalysts* for EP(D)M polymerization. The most widely used commercial compounds are diethylaluminum chloride and ethylaluminum sesquichloride.

Reactivators. Catalyst efficiency (typically 800 – 8000 g polymer per gram vanadium) can be improved by using reactivators. The active oxidation state of vanadium is + 3; vanadium is reduced from oxidation state + 4 (VCl_4) or + 5 ($VOCl_3$) during alkylation [292]. Vanadium in the inactive oxidation state of + 2 can be reactivated by treatment with compounds containing active halogen atoms which can oxidize the deactivated vanadium species back to the + 3 valence state [293], [294]. Ethyl trichloroacetate and *n*-butyl perchlorocrotonate are commercially used. DSM has promoted the EP(D)M reactivator ethyl dichlorophenylacetate. It is claimed that this compound combines an excellent reactivation effect with lower residual vanadium and chlorine levels in the finished polymer [295], [296].

Molar Mass Control. In most Ziegler – Natta systems, the molar mass of the polymer is influenced by the type of solvent, the concentration of monomers, catalyst and reactivator, the temperature, the aluminum : vanadium ratio and the ethylene : propene ratio [288]. Trace levels of impurities can also have a significant effect on molar mass. As in homopolymerization, transfer agents such as hydrogen and diethyl zinc are widely used for the primary control of molar mass of the polymer during copolymerization.

Production Processes. At present, commercial EPM and EPDM polymers are produced by solution or suspension processes. Most manufacturers use a solution process. The reaction is homogeneous; the process is flexible and capable of producing many different grades of products. Enichem and Miles (Bayer) use a suspension process [281]. Both solution and suspension are continuous processes which typically consist of four operating sections, i.e., polymerization, momomer and solvent recycling, product recovery, and production finishing.

Solution Process. The schematic of a typical solution process is shown in Figure 14. Typcially, polymerization is carried out at 10 – 50 °C and 1 – 20 bar in a single continuous stirred-tank reactor (CSTR) (a). The polymerization is exothermic ($\Delta H \approx -2.4$ MJ per kilogram of polymer). The heat of reaction is removed by precooling the reactor feeds and/or evaporative cooling of the reaction mixture. The residence times in the reactor range from 15 – 60 min, depending on the catalyst employed, temperature, etc. Ethylene conversion is nearly complete in a single pass through the reactor. Unreacted propene is recovered and recycled to the polymerization. Solvent (usually hexane) and unreacted diene are also recovered for reuse after purification. Immediately downstream of the reactor, the polymerization is shortstopped with water, alcohol, or other polar liquids (b). The polymer solution is then mixed with water, aqueous sodium hydroxide, or dilute acid under vigorous agitation to remove catalyst residues (c). After addition of chemicals such as extender oils and stabilizers (d), the polymer solution is coagulated and stripped with steam and hot water to remove unreacted monomers and solvents (e). Antiagglomerants and antifoaming agents may be added to control crumb agglomeration and facilitate the separation of gas and liquid. The rubber crumb is then passed over a screen (f) to remove free water and subsequently dried. Either extrusion drying or hot air drying may be used. EP(D)M is sold in the form of crumb, bales, or pellets.

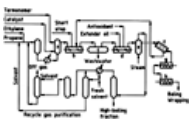

Figure 14. Solution process for production of EPM/EPDM [297], [298]

a) Reactor; b) Shortstopping section; c) Separation of catalyst; d) Addition of chemicals; e) Steam stripping section; f) Screen; g) Expeller; h) Expander

Suspension Process. Suspension polymerization is carried out in an excess of liquid propene. The polymer precipitates out of the reaction medium as soon as it is formed. Consequently, the viscosity increase in the polymerization medium is negligible, compared with that of solution polymerization. The solids content in the reaction medium can therefore be higher than in the solution process, e.g., 20 – 30 wt % instead of 5 – 10 wt %. Monomer recovery, stripping, and finishing are carried out as described for the solution process except that the coagulation step is not required since the polymer already exists as particles. The slurry process is well suited to the production of grades with high Mooney viscosity, which can be oil extended for easier processing.

Dutral (now Enichem) is running a *simplified suspension process* for the production of EPM, using a highly active catalyst based on titanium. The product is recovered directly from the suspension in an extruder-based system, without the prior catalyst residue removal. The unreacted monomers are recycled to the reactor without treatment or distillation [299], [300]. This new slurry process offers the advantages of reduced capital and operating costs. However, presently this process is limited to the production of EPM. Unfortunately, polymers produced with titanium catalysts tend to have a less random monomer

sequence distribution compared with existing vanadium-based commercial copolymers [281], due to multiple active sites with titanium compared with essentially single sites with vanadium.

Gas-Phase Polymerization. Gas-phase polymerization is now commonplace for polyolefins [301], [302] (⟶ Polyolefins – Gas-Phase Process) and some types of modified polypropylenes (reactor-modified thermoplastic olefins) (see Section Thermoplastic Polyolefins) [303]. Extension of this technology to the full range of elastomeric co- and terpolymers is in the advanced pilot plant stage [304-306]. A schematic of the overall Union Carbide gas-phase process is shown in Figure 15.

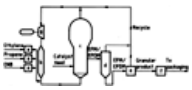

Figure 15. Gas phase EPM/EPDM process of Union Carbide [307]

a) Monomer purification; b) Cycle cooler; c) Cycle compressor; d) Fluidized-bed reactor; e) Product discharge; f) Monomer separation; g) Balers/baggers

Monomers and catalysts are fed to the reactor (c) and the product is removed intermittently. Heat of reaction is removed by circulating the fluidizing gas through heat exchangers outside the reactor. The process uses no solvents and does not require a catalyst residue removal step. Solvent stripping and recovery operations are avoided, although unreacted monomers must be purged before discharging the product. In principle this technology offers the advantages of reduced capital and operating costs when compared to the conventional processes. It has yet to be proven, however, that the fluidized-bed process can produce as wide a range of products, having the same or comparable properties (e.g., high ENB or low Mooney) as those produced by existing processes and that the product grade changeover time is not prohibitive.

3.2.2.3. *Producers and Production Capacities*

The commercial production of EPM and EPDM started in the early 1960s. Today EP(D)M has the largest market volume of all special elastomers: 601 000 t in 1992 [308]. With a growth rate of ca. 4 % per year, the EP(D)M consumption is predicted to increase to 726 000 t by 1996. In 1992, the production capacity of the 12 EP(D)M manufacturers that operate 15 plants was ca. 854 000 t. Table 22 gives an overview of the EPM/EPDM producers, their plant locations, their production capacities (1992), and their trade names.

Table 22. EPM/EPDM producers, locations, production capacities (1992), and trade names [308]

Country, company	Plant location	Start-up	Capacity, t	Trade name
Western Europe				
France				
Socabu/Exxon	Gravenchon	1972	80 000	Vistalon
Germany				
Hüls	Marl	1972	41 000	Buna AP
Italy				
Enichem	Ferrara	1963	85 000	Dutral
Netherlands				
DSM	Beek	1965	60 000	Keltan
North America				
United States				
Copolymer/DSM	Addis, LA	1968	73 000	Epsyn
Du Pont	Beaumont, TX	1963	77 000	Nordel
Exxon Chemical	Baton Rouge, LA	1962	85 000	Vistalon
Miles (Bayer)	Orange, TX	1967	52 000	Polysar EPDM
Uniroyal Chem.	Geismar, LA	1963	91 000	Royalene
South America				
Brazil				
Nitriflex	Triunfo	1988	10 000	
Asia				
Japan				
DSM-IDEMITSU	Chiba	1991	40 000	
JSR	Yokkaichi	1970	45 000	JSR-EP
Mitsui	Chiba	1970	60 000	Mitsui-EPT
Sumitomo Chem.	Chiba	1969	35 000	Esprene
Korea				
Kumho E P.	Yochon	1988	20 000	

3.3. Synthesis of Butyl Rubber by Cationic Polymerization [309], [310]

Butyl rubber is a copolymer of isobutene (2-methylpropene) and a small amount of isoprene (2-methylbuta-1,3-diene, 1 – 3 mol %). It is abbreviated IIR in the ISO nomenclature (ISO 1629, 1987) (butyl rubber see Section Nomenclature and Classification). The isoprene units introduce unsaturated sites into the polymer chain to enable vulcanization. Historically, butyl rubber was the first saturated polymer into which limited olefinic functionality was deliberately introduced [311]; EPDM (see Section Ethylene – Propene Elastomers (EPM, EPDM)) is a more recent example of the same principle.

The principal structure of the polymer chain is as follows:

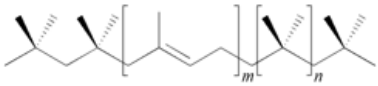

$n >> m$

3.3.1. *Properties, Grades, and Applications*

3.3.1.1. *Properties*

The most significant property of butyl rubber that differentiates it from the other hydrocarbon elastomers, is low permeability to gases; most significantly, in practice, to air and water vapor. Due to its low resilience, butyl rubber is also used when vibration damping is important. For applications, butyl rubber is compounded with fillers, plasticizers, and curing agents (see Section Applications)

The highly saturated IIR chain is resistant to chemical attack. Thus IIR compounds exhibit better heat and ozone resistance than highly unsaturated elastomers such as BR, NR, and SBR. IIR also has excellent resistance to acids, bases, and other chemicals. As are all hydrocarbon polymers, IIR is highly swollen by gasoline and other hydrocarbons. Butyl rubber has good electrical insulating properties and is biocompatible [312]. As a viscoeleastic material, uncured IIR creeps (cold flows) at ambient temperature. This deformation is prevented by vulcanization (i.e., cross-linking via the isoprene units).

Typical properties of the *uncured polymer* are as follows [313]:

Density	917 kg/m^3
Glass transition temperature	– 70 °C
Heat capacity	1.95 kJ kg^{-1} K^{-1}
Thermal conductivity	0.130 W m^{-1} K^{-1}
Refractive index	1.51
Dielectric constant (at 1 kHz)	2.38
Dissipation factor (at 1 kHz)	0.003

The proton NMR spectrum of IIR has singlets at 1.1 and 2.3 ppm from tetramethylsilane (TMS) due to the methyl and methylene groups of the polyisobutene, a triplet at 5.1 ppm that stems from the single hydrogen attached to the double bond [314], as well as a peak around 4.9 ppm that has been identified as a branching point [315]. The isoprene content listed in Table 23 is the sum of the regular and the branched structures.

3.3.1.2. *Grades*

Each producer listed in Table 24 makes several grades of IIR, which differ primarily in the amount of isoprene and the Mooney viscosity (i.e., molar mass and molar mass distribution). Equivalent grades of the two main producers are listed in Table 23 together with illustrative values of specified or important properties. The actual specification ranges differ slightly for different producers, the comparison in Table 23 being based on Mooney Viscosity.

Table 23. Selected grades and properties of IIR

Producer	Butyl grade			
Bayer	100	101–3	301	402
Exxon	065	007	268	365
Mooney viscosity[a], MU	33±3	51±5	51±5	33±4
Isoprene content[b], mol %	0.9±0.2	1.75±0.2	1.85±0.2	2.25±0.2
Cure characteristics [c]				
Maximum torque, dN · m	64±6.0	–	82±6.0	87±6.0
Minimum torque, dN · m	11±4.5	–	16.0±4.5	11.5±4.5
Antioxidant content[b], min wt %	0.03 min	none	0.03	0.03
Ash content[d], wt %	0.3	0.3	0.3	0.3
Volatiles content, max %	0.3	1.0	0.3	0.3

[a] Bayer grade according to ASTM D 1646, ML/1+8 (125 °C).
[b] Bayer standard method.
[c] ExxonMobil grade according to ASTM D-2084, modified (rheometer ODR 2000).
[d] According to ASTM D 5667.

3.3.1.3. *Specialty Rubbers*

Specialty rubbers include Bayer's XL 10000, a cross-linked grade containing DVB [313], and Exxon's Star-branched butyl rubber, containing 5–20 % of a very high molar mass branched fraction [316, 317]. The latter was introduced by Exxon Chemical Co. in 1989. A broad, bimodal molar mass distribution was achieved using branching agents in the polymerization. One mode consists of normal linear chains, as previously described. The second mode consists of branches of normal chains connected through an SBS polymeric branching agent added to the polymerization. The concept was labeled “Star-branched butyls”, and exhibits higher green strength and faster stress relaxation.

3.3.1.4. *Applications*

The majority of the IIR produced is used captively by manufacturers as base polymer for halobutyl rubber production, discussed in Section Halobutyl Rubber. A significant portion of the remainder is used in tire-related applications: inner tubes, tire-curing bladders, or as blends with BIIR—inner liners. A small amount of this is used in other applications, principally sealant tapes and caulks, pharmaceutical closures, membranes for roofing and water retention, and various mechanical seals and mounts.

For most applications, IIR is vulcanized via the isoprene units to cross-link the individual polymer chains into a network. Vulcanization or curing is carried out commercially by three methods, using sulfur, phenolic resins, or quinone dioximes as curing agent. Of these methods, accelerated sulfur cures are used for most applications, phenolic resin cures are used for making tire-curing bladders, and quinone dioxime cures are used infrequently.

The measured properties of the vulcanizate are affected by the method and extent of cross-linking and by the quantity and type of fillers and plasticizers used.

IIR is also used as a component of food and in the manufacture and storage of foods. IIR is an allowed component of chewing gum in most countries [318]. Bayer's XL 10000 may be used in sealing cements, and is also commonly used in condenser caps [315].

3.3.2. *Basic Chemistry and Production Processes*

3.3.2.1. *Basic Chemistry*

Butyl rubber is composed of linear polyisobutene chains with isoprene units distributed randomly along them [319]. The isobutene moieties are linked exclusively head to tail; the isoprene units have mainly the *trans*-1,4-configuration, with about 10 % of the branched structure [315].

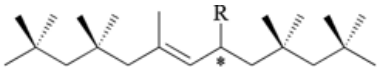

The copolymerization of isobutene and various monomers such as isoprene or isoprene and DVB (divinylbenzene) is a carbocationic process. This is initiated by a species resulting from an activated Lewis acid such as AlX_nR_{3-n} (X = Cl, n = 1 – 3, R = Et) in the slurry or solution process [320], [321]. The Lewis acid is activated by protogenic compounds such as water, hydrochloric acid, or hydrogen sulfide. Specialty rubbers such as XL10000 (Bayer) or Star-branched butyl (ExxonMobil Chemical) can be formed by the addition of a third monomer during the polymerization (see Section Specialty Rubbers). The mechanism of the copolymerization is still not fully understood; the generally accepted scheme is given below:

Initiator formation (speculated)

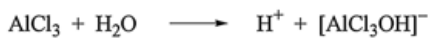

Initiation

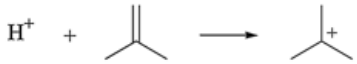

Propagation

Chain transfer

Termination

Alternative mechanisms have been proposed [322].

The incorporation of isoprene is governed by the ratios of homo- and cross propagation rate constants, $r_{1B} = k_{11}/k_{12}$ and $r_{1P} = k_{22}/k_{21}$, where isobutene and isoprene are referred to as monomer 1 and 2, respectively. It was shown that the composition of butyl rubber can be described by r_{1B} alone [315]. The rate constant for isobutene homopolymerization is $k_{11} = (6 \pm 2 \times 10^8)$ L mol^{-1} s^{-1} but this number has always been a source of heated debate [323].

Industrially the copolymerization is carried out at low temperature (ca. – 90 °C) in order to obtain high molar masses. The strong dependence of molar mass on temperature is due to the high activation energy of chain transfer relative to that of the polymerization. In the production process of butyl rubber, isoprene is both the dominating poison and chain transfer agent. In general, most unsaturated and substituted hydrocarbons and some alkyl halides affect the polymerization.- The compounds classified as “poisons” reduce the degree of conversion, while those classified as “chain transfer agents” reduce the molar mass (Mooney viscosity) [324]. The *slurry process* uses chloromethane as a diluent, while the *solution process* uses a hydrocarbon such as isopentane as solvent. Both solvents are inert in terms of polymerization chemistry. Although isobutene is believed to polymerize by a carbocationic mechanism only [312], alternative mechanisms such as a radical – cationic [325] or a cationic – coordination [326] mechanism have been suggested. The claim of the cationic-coordination mechanism has yet to be substantiated but continues to be an area of considerable research due to the cost saving that could be realized by not having to polymerize at ca. –90 °C. Considerable research effort is being directed towards a system that would allow copolymerization at higher temperature [327].

3.3.2.2. *Industrial Production*

Historical Development. During the late 1920s M. OTTO working at the IG Farben laboratories in Germany, investigated synthetic oils made by polymerizing olefins with strong acids. He made the important discovery that the polymerization temperature regulated molar mass. By chilling isobutene with dry ice before addition of boron trifluoride catalyst, OTTO succeeded in making rubberlike products instead of oils. Around 1930 F. A. HOWARD, the President of Standard Oil Development Company, visited Germany and learnt of OTTO's discovery. In 1932 OTTO, invited by HOWARD, came to the USA and worked with R. M. THOMAS to exploit the polymers of isobutene. The process for manufacturing butyl rubber was further developed by R. M. THOMAS and O. C. SLOTTERBECK at Standard Oil Development Company (Esso, now Exxon) (Linden, New Jersey, United States). They firstly used butadiene and then isoprene to introduce sites for vulcanization. During World War II, most of the world was cut off from natural rubber supplies in the Far East and, in a remarkably brief time, large-scale manufacturing of various synthetic rubbers was implemented in North America [328, 329].

The first commercial-scale production unit came into operation in 1943 in Baton Rouge, Louisiana. Units in Baytown, Texas, and Sarnia, Ontario, Canada, followed in 1944.

Today, butyl rubber is produced by two different processes, the slurry and the solution processes. The vast majority of butyl rubber is produced by the slurry process, so this will be discussed in more detail.

Raw Materials. The monomers used for both production processes of butyl rubber, isobutene, and isoprene are extracted from internal petroleum refinery streams (⟶ Butenes – Upgrading of Butenes; ⟶ Isoprene – Recovery from C_5 Cracking Fractions (more details in [330])

Slurry Polymerization Process. In the traditional method of manufacture, practiced since 1943, an aluminum trichloride catalyst is used with a small quantity of water as cocatalyst in chloromethane as diluent. The terms catalyst and cocatalyst reflect established industrial usage; the correct terms are initiator and coinitiator, respectively [312]. The catalyst solution (0.2 wt % $AlCl_3$ in chloromethane with a small amount of water) is prepared by passing pure and dry chloromethane through a bed of solid $AlCl_3$ (Fig. 16, a). Water is added to this concentrated solution, which is further diluted with chloromethane

and chilled to the reaction temperature. The catalyst is quite active; the polymer to catalyst mass ratio is normally on the order of several thousands. The catalyst activity is proportional to the water content of the catalyst solution in the low concentration range. Excessive amount of water destroys the catalyst. The catalyst solution is then fed into the reactor together with the monomer feed. The nominal isobutene cocentration in the feed is ca. 30 wt %. The reaction is carried out continuously in an intensely agitated vessel (Fig. 16, b) at ca. – 90 °C. The polymer is formed instantaneously, and since it is insoluble in the diluent, precipitates as a fine slurry (particle size 10 – 30 µm). Agglomeration is not a problem because the rubber is below its glass transition temperature of – 70 °C. The slurry circulates through concentric rows of heat-exchanger tubes surrounding a central draft tube. The circulation : feed ratio is very high (> 100), but kinetic analysis showed that the reactor flow regime is closer to that of a plug flow reactor than that of a continuous stirred tank reactor because of the extremely high reaction rate [331]. The heat of reaction (– 0.8 MJ/kg) is removed by boiling subatmospheric ethylene in the reactor jacket.

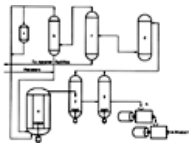

Figure 16. Flow diagram for production of butyl rubber

a) Dissolver; b) Catalyst tower; c) Recycle tower; d) Recycle gas dryer; e) Reactor; f) Flash drum; g) Vacuum stripper; h) Finishing extruders

Reactor Control. Reactors are controlled to produce the desired molar mass of polymer, measured as Mooney viscosity, with the catalyst flow as the principal control variable. An increase in catalyst flow decreases the Mooney viscosity. The isoprene content of the polymer, the second most important control parameter, is determined by the composition of the feed together with the degree of conversion. The reactors are run typically at incomplete conversion for better control. Due to the poisoning effect of isoprene, conversion decreases with increasing isoprene content in the feed. The catalyst flow also has an effect on conversion; thus controlling this process is not easy. The reactors are operated intermittently because of fouling of the heat exchanger surfaces. When the control of the reactor cannot be maintained due to temperature increase, the operations are stopped and the reactors washed. To reduce or prevent fouling, the use of slurry stabilizing agents [332] or addition of 3 – 13 % solvent such as hexane to the feed has been patented [333].

Product Recovery and Finishing. The reactor discharge stream is treated with steam and hot water in a flash drum (Fig. 16, f). Most of the chloromethane and the unreacted monomers are flashed off and the water is separated from the vapors by condensation and subsequent adsorption on activated alumina (d). This step sometimes is proceeded by absorption in glycol. The dried stream is separated by distillation (b, c) into a chloromethane – isobutene stream for feed make-up, a pure chloromethane stream for catalyst make-up, and streams which must be purged from the recycle loop to prevent build-up of unwanted impurities. The polymer is recovered in the form of a wet crumb (6 – 10 wt % rubber in water). To control agglomeration, salts of fatty acids are added in the flash drum. Antioxidants are also added at this stage and sodium hydroxide is added to neutralize catalyst residues. The rubber is converted into the final commercial bale form by further drying. Most of the water is separated by draining, followed by drying extruders and finished by a final drying step in a fluidized bed. The dry rubber is cooled, baled, and packed into boxes or crates for shipment.

When the polymer from the reactor is to be processed further, such as by halogenation (see Section Halobutyl Rubber), the product may be recovered directly as a solution by discharging the reactor content into a hot solvent such as hexane. A further stripping stage is followed by adjusting the rubber concentration to the desired level for further processing, typically 20 – 25 wt % [334]. It is important to remove unsaturated residues before halogenation, otherwise these unsaturated substances would be halogenated, producing undesirable compounds.

Solution Polymerization Process. A process developed in the Soviet Union operates at – 60 to – 90 °C using a hydrocarbon solvent, such as isopentane and an aluminum alkyl halide catalyst [321]. This eliminates chloromethane and makes the water slurry step superfluous when the polymer is to be treated further in solution. However, the polymer concentration in the solution is necessarily lower than that possible in a slurry because the effect of concentration on effective viscosity is greater in solution, which reduces heat-transfer efficiency. Despite the decreased demand for refrigeration, the presence of the additional diluent leads to energy costs higher than those for the slurry process.

3.3.3. *Producers and Production Capacities*

Worldwide, several companies produce butyl elastomers in many different countries. However, the most prominent producers are Bayer AG and ExxonMobil Chemical with collectively two-thirds of the world production capacity. Worldwide production facilities for butyl and halobutyl rubber are listed in Table 24 [335], [336]. The capacities given refer to polymerization, with the exception of Japan Butyl's Kashima plant, which is a halogenation facility only. In general, many of the production facilities are multifunctional, able to produce butyl or halobutyl rubber on demand. The global nameplate capacity was 972 × 10^3 t in 2001.

Table 24. Producers and production capabilities of butyl and halobutyl rubber

	Producer	Location	Plant	Nameplate, 10^3 t/a
North America	Bayer Inc.	Sarnia, Canada	Butyl I + II	140
	Exxon Mobil Chemical	Baytown, USA		121
	Exxon Mobil Chemical	Baton Rouge, USA		132
Western Europe	Bayer	Antwerp, (B)	B.R.N.V.	115
	Exxon Chemical	N. Dame de Gravenchon (F)	Socabu	56
	Exxon Chemical	Fawley (UK)		93
Cental Europe and CIS	Nizhnekamskneftekhim	Nizhnekamsk, Tartarstan		60
	Sintezkauchuk	Togliatti, Russia		35
	Pitesti	Romania		5
Asia	Japan Butyl	Kawasaki		80
	Company	Kashima		35
	Yan Shan Petrochemical	Yan Shan (Beijing), China		30
	Reliance Industries	Hazira, China		25
World total				972

Since the 1970s the relative amount of IIR used has been decreasing and the amount of halobutyl rubber increasing. Differences in the relative amounts of IIR and halobutyl rubber used in different regions reflect the relative number of tubed and tubeless tires.

3.3.4. *Storage and Transportation*

IIR is packaged in 34 kg bales wrapped in synthetic polymer films, and shipped in crates or skids depending on customer needs.

IIR is considered a nonhazardous material for shipping and storage [337]. Stabilized grades should be used within five years of their date of manufacture, after which the molar mass may decrease significantly. The molar mass of grades without antioxidant can decrease after six months of storage.

3.3.5. *Legal Aspects*

The allowable uses and restraints of IIR in foods and storage of foods are covered by regulation of individual countries. These usually follow the United States FDA CFR, although local differences may occur.

The grades of IIR used in chewing gums, Bayer Butyl 101–3 and Exxon Butyl 007, contain no added antioxidant and are made with food-grade fatty acid. The use of IIR in *food manufacture* and *storage* is usually restricted to nonfat or low-fat foods [168]. In these applications, constraints are placed on the types and levels of antioxidants used. If the IIR is vulcanized, the curatives and other additives in the compound are also subject to restrictions.

3.3.6. *Toxicology and Occupational Health*

IIR is a nonhazardous material, and no special precautions are required in its handling [169]. It is generally used in cured compounds, and precautions for handling curatives and other compounding ingredients are frequently more significant.

Butyl rubbers contain only small amounts of the residual monomers isobutene and isoprene, typically < 100 mg/kg in each case. To date, no threshold limit value has been established for either monomer, since they both exhibit very low *acute toxicity*. The 2 h LC_{50} in mice is 415 g/m^3 and 157 g/m^3 for isobutene and isoprene, respectively [338].

A modified Ames test of isobutene showed that it is not mutagenic. Only traces of chloromethane (< 1 mg/kg) are found in butyl rubbers. Thus, the maximum airborne concentration released during processing should be well below the 50 mL/m^3 (8 h weighted average) occupational exposure limit set by the ACGIH.

Butyl rubber will not burn unless its surface temperature reaches 300 – 500 °C and an ignition source is available. Burning butyl rubber gives off dense black smoke and toxic gases such as carbon monoxide and hydrocarbons, resulting from incomplete combustion.

[Top of Page]

4. Synthesis by Radical and Other Mechanisms

John Dunn, Hermann Meisenheimer, Roland Steiger, André Marbach, Klaus M. Diedrich

4.1. EVM and Ethylene Copolymers

4.1.1. *Properties, Grades, and Applications*

Two copolymers of ethylene are industrially important in the solid rubber sector: ethylene – vinyl acetate copolymers, EVM [24937-78-8], and ethylene – methyl acrylate copolymers, AEM [25103-74-6].

EVM grades containing *between 40 and 80 wt % vinyl acetate* have elastomer properties. Their glass transition temperatures T_g are between –36 and 0 °C [219, 235, 236]. Their Mooney viscosities ML 1 + 4 (100 °C) are between 15 and 35 Mooney units, depending on the production process; thus EVM belongs to the readily processible rubbers.

Ethylene – vinyl acetate copolymers with *vinyl acetate content < 40 wt %* have pronounced partial crystallinity and can be processed thermoplastically. Copolymers with > *80 wt % vinyl acetate* are amorphous and have a T_g greater than 0 °C.

Ethylene – methyl acrylate solid rubber grades with methyl acrylate contents between 55 and 70 wt %, and Mooney viscosities ML 1 + 4 (100 °C) between 15 and 30 Mooney units, are available commercially. Their glass transition temperatures lie between –36 and –20 °C [237].

All other ethylene copolymers produced in industry are thermoplastics and are therefore not dealt with here [341], [342].

The following properties can be achieved by proper compounding in vulcanizates based on EVM or AEM:

1. Heat resistance up to 175 °C
2. Halogen-free
3. Good weathering, ozone, and UV resistance
4. Swelling in mineral oil of 0 – 100 vol % (depending on vinyl acetate or methyl acrylate content)
5. Satisfactory low-temperature properties (T_g between – 36 and 0 °C, depending on vinyl acetate or methyl acrylate content)
6. Satisfactory mechanical properties

The basic properties of EVM and AEM vulcanized rubbers are closely related to their chemical structures. The saturated methylene main chain and the associated thermal stability are common to both (see Fig. 17 and Table 4). EVM is only exceeded by silicone and fluorine rubber in its thermal stability [343]. Noticeable decomposition with loss of acetic acid occurs only above 170 – 180 °C [219]. The high resistance to weathering, ozone, and UV irradiation, of both EVM and AEM are attributed to the saturated polymer main chain.

Figure 17. Structure and basic properties of ethylene – vinyl acetate rubber (EVM) [235]

A) Heat, ozone, weather, and UV resistance; B) Low oil swell; C) Compatibility to nonpolar compounds; D) Preferred cure sites for peroxide curing

Both rubbers are halogen free, which is important with regard to their fire protection behavior (formation of halogen-free combustion gases).

The properties of EVM and AEM depend highly on the comonomer content. For a partially crystalline, branched polyethylene, the ethylene sequences capable of crystallization are distorted by the incorporation of vinyl acetate or methyl acrylate. Thus, with increasing co-monomer content the proportion of noncrystalline material also increases [219], [342], [344]. An EVM with 60 wt % vinyl acetate is completely amorphous.

With increasing crystallinity (increasing ethylene content), the polymers become harder, more rigid, and less permeable to gas or steam. The transparency also decreases [219], [342], [345]. The resistance of EVM and AEM toward nonpolar oils and fuels increases with increasing content of polar vinyl acetate or methyl acrylate groups. In contrast, the mobility of the polymer chains and the low-temperature flexibility decrease [236], [346], [347].

In accordance with the Gordon – Taylor relationship [348] the *glass transition temperatures* of the copolymers should vary linearly between the T_g of a purely amorphous low-density polyethylene (– 120 °C) [349] and that of a poly(methyl acrylate) (+ 6 °C) [350] or poly(vinyl acetate) (+ 29 °C/ +42 °C) [350]. For the Tg of poly(vinyl acetate) various data have been reported such as 35 °C [351] or 42 °C [352]. The Tg is dependent on molar mass [353]. Samples prepared in experiments in solution have a Tg between 39 and 42 °C. However, copolymers with low vinyl acetate content have a higher T_g than expected.

For partially crystalline ethylene – vinyl acetate copolymers (< 55 wt % vinyl acetate), T_g values are between –38 and – 30 °C and depend only slightly on the vinyl acetate content (see Fig. 18). Commercially available AEM grades have a T_g of –31 °C (55 wt % methyl acrylate) and – 20 °C (69 wt % methyl acrylate).

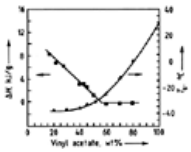

Figure 18. Glass transition temperature T_g and enthalpy of fusion ΔH of ethylene – vinyl acetate copolymers (DSC, second heating, 20 K/min)

Electrical properties also vary with the proportion of polar groups. The specific volume resistance ϱ_D of EVM decreases with increasing vinyl acetate content. From 0 to 150 °C, ϱ_D values are between 10^{17} and 10^{10} Ω · cm.

EVM and AEM rubbers are radically cross-linked, preferably with organic peroxides [343], [354-357]. To increase the rate and degree of cross-linking of EVM, low molar mass, polyfunctional vinyl compounds, such as triallyl cyanurate or triallyl phosphate, are added as well as peroxide [358], [359]. Cross-linking by using high-energy radiation (β radiation) is also possible [360].

Some commercial AEM grades contain ca. 4 wt % of a vinyl carboxylic acid (such as acrylic acid, maleic acid, or monoethyl maleate) as so-called cure site molecules [346], [361], [362]. These terpolymers are cross-linked with hexamethylenediamine and diphenylguanidine. Such nonradical cross-linking can improve the mechanical properties of the vulcanizate. Nonradical cross-linking of EVM by using alkyl aluminum compounds has been reported [363].

Applications.
Ethylene – Vinyl Acetate Rubbers. The main areas of application of EVM rubbers are in the cable sector, adhesives, and the modification of thermoplastics. As a result of the increasing high-temperature requirements in the automotive industry, EVM is being introduced for use in hoses and seals. EVM also has excellent dynamic properties which make it a suitable material for roll covers.

In the cable sector, EVM is used in heat-resistant electrical insulation (e.g., in floor heating, ignition wires, night storage heaters, rear window heating, electric motors, toasters, grills); in halogen-free cable sheathing; and in easy-peel conductive layers over medium-voltage insulation (underground cables) [345].

The so-called flame-retardant, noncorrosive (FRNC) application is especially important. EVM vulcanizates containing flame-retardant fillers, such as aluminum hydroxide, magnesium carbonate, magnesium hydroxide, or zinc borate, meet the most stringent safety requirements in case of fire [364]. They are characterized by

1. Low flammability
2. Low combustion gas density
3. Corrosion-free combustion gases

Wires and cables, as well as floors and noise absorption mats based on EVM, offer increased safety for people and materials in the event of fire. They are used, for example, in subways, shops, offices, hospitals, ships, control stations, computer centers, and radio or television stations. Underground conveyor belts based on EVM also satisfy strict flame resistance regulations [365], [366].

Ethylene – vinyl acetate has a low polymer viscosity for a solid rubber. Therefore, in the preparation of conductive compounds for medium-voltage cables, electrically conductive carbon blacks can be easily dispersed without the addition of plasticizers. The latter would considerably impair the dielectric properties of the insulation.

Another important reason for the use of EVM as a conductive layer in cable construction is the ready removal of this layer during mounting in the cable duct (banana peel effect) [345].

EVM copolymers with up to 45 wt % vinyl acetate are used as solvent-free, hot-melt adhesives [367], [368]. Formulation with waxes permits control of the melt index and the viscosity of the molten adhesive. Adhesion is strengthened by the addition of adhesive resins. EVM copolymers with 50–80 wt % vinyl acetate can also be used as pressure-sensitive adhesives (PSA) [369].

In contrast to most common thermoplastic ethylene-vinyl acetate copolymers with low vinyl acetate content, EVM copolymers can be processed easily for thin-layer applications. The specific molar mass distribution allows the preparation of high quality, transparent and homogeneous composites, either as cast film or as blown film [369], [370].

EVM copolymers are also suitable for *modifying thermoplastics*, such as poly(vinyl chloride) (PVC) [371], [372], styrene – acrylonitrile copolymers (SAN) [373], and cellulose esters [374]. For example, EVM – PVC graft polymers with 5 – 10 % EVM (45 wt % vinyl acetate) give impact-resistant products, which are used in plastic windows [372], [375]. EVM copolymers with 60 – 70 wt % vinyl acetate are completely compatible with PVC and function as polymeric plasticizers. These types of mixtures are processed to give sheets for car fittings.

Low molar mass EVM copolymers with up to 40 wt % vinyl acetate are also used as flow improvers for crude oil and middle distillates [376].

Ethylene – vinyl acetate copolymers can be completely or partially saponified. Here the ester groups are formally converted into vinyl alcohol units, with loss of acetic acid. EVM grades modified in this way are suitable for woven tissue coatings [377] and as hot-melt adhesives [378]. Other applications are in pigment dye printing, paints for paper, and varnishes [379], where they are used as thickeners.

Ethylene – Methyl Acrylate Rubbers. The primary uses of AEM rubbers are in car construction [346], [380]. Because of their good oil-resistance – low-temperature flexibility relationship (see Table 4), they are used for seals, bellows, hosing, and O-rings [346], [361], [362], in addition to cable and wiring insulation [381], [382].

4.1.2. *Basic Chemistry and Production Processes*

Polymerization Mechanisms. The radical copolymerization of ethylene and vinyl acetate has been known since 1938 [383]. At 1000 bar and 90 °C the copolymerization parameters r_1 and r_2 are nearly 1 for both monomers (see also ⟶ Plastics, General Survey – Copolymerizations) [384-386]. Higher r values have been found for ethylene at elevated temperature [387]. Higher rates of ethylene incorporation with increasing pressure have also been found [388]; this was attributed to heterogeneous reaction conditions (phase separation) [389], [390].

NMR investigations confirm the random distribution of monomers in EVM-type copolymers [391], [392].

Polymerization is initiated with peroxides, peresters, hydroperoxides, or azo compounds [341]. The choice of initiators depends on the polymerization temperature and the desired polymerization time. Important crude polymer properties, such as short-chain branching, depend strongly on these parameters [393].

An ethylene – vinyl acetate copolymerization has been described on a laboratory scale that uses organometallic catalysts. The catalysts are based on

rhodium(III) chloride – triethylaluminum complexes [394].

Production of EVM Rubber. EVM solid rubber is produced in industry by three processes (see Fig. 19).

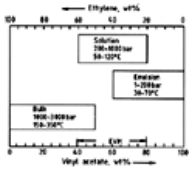

Figure 19. Production processes for ethylene – vinyl acetate copolymers

High-Pressure Process. The high-pressure or bulk process [219] is similar to the production process for low-density polyethylene (LDPE) (⟶ Polyolefins). It is carried out at 1000 – 3000 bar and 150 – 350 °C without the addition of solvents. Stabilized vinyl acetate is first mixed with precooled ethylene, which has been compressed to 100 – 500 bar. The monomer stream is compressed further and charged continuously to a stirred autoclave or tube reactor [341]. The initiator solution and, if required, a chain-transfer agent (e.g., an alkane, ketone, or aldehyde) are injected. The conversion is 10 – 20 %, based on the monomers. The reaction mixture is then depressurized in stages; the residual monomers are separated and recycled to the reaction. The polymer melt is chilled and pelletized. Monomeric vinyl acetate has chain-transfer properties that limit the molar mass of the polymer chain [343]. As a result of the high monomer concentration in the high pressure process, only relatively low average molar masses can be obtained with copolymers with high vinyl acetate content [219]. High-pressure EVM products are therefore produced only up to a vinyl acetate content of 45 wt % and Mooney viscosities of 15 – 20 Mooney units.

Solution Process. The solution process [395-397] is operated in the medium-pressure region (200 – 1000 bar) at 50 – 120 °C by using solvents with low transfer constants. Suitable solvents are *tert*-butanol [395-404], benzene, toluene, methyl acetate [219], methanol [405-407], and dialkyl sulfoxides [408]. The process enables the production of gel-free copolymers [400], [401] that are characterized by high Mooney viscosity (20 – 35 Mooney units) at high vinyl acetate content (40 – 80 wt %).

Polymerization occurs in a cascade of stirred-tank reactors at conversions of 50 to 70 %. During workup the product solution is freed from solvent and residual monomer. Pellets 2 – 4 mm in size are finally produced for formulation, which give a free-flowing material when treated with a suitable coating to lower the self-adhesion of polymer particles.

Emulsion Process. The emulsion process [409] operates between 30 and 70 °C at 1 – 200 bar. The polymers have vinyl acetate contents between 60 and 100 wt % and are produced as latices. The products are not suitable for vulcanization because of their high proportion of gel. Dispersions of EVM copolymers are used as binders for exterior paints and plasters, as components of wood adhesives, and as thickeners in varnishes [379].

Production of AEM Rubber. AEM solid rubbers were first introduced in 1975 as filler-containing masterbatches by Du Pont [346]. No production processes are described in the literature. However, AEM is often produced alternately with EVM in the same high-pressure plants. Although acrylates are clearly more reactive in polymerization than ethylene [342], the process gives random ethylene – methyl acrylate co- and terpolymers [346]. Therefore, either low conversions are used or methyl acrylate is added stepwise [341].

4.1.3. *Producers and Production Capacities*

The capacity for production of EVM or AEM copolymers was 215 000 t in 1996 according to IISRP [410]. It is not clear how many tonnes of elastomer grades are actually produced.

The most important *producers* of EVM and AEM solid rubbers are Bayer [Dormagen, Germany, trade name Levapren (EVM)] [412], Du Pont [Orange, Texas, trade names Elvax (EVM), Vamac (AEM)], and Exxon [Antwerpen, Belgium, trade name Escorene (EVM)].

4.1.4. *Toxicology*

EVM copolymers are physiologically harmless. They are licensed for contact with foods and correspond to recommendations of the BGVV (the former BGA) [411].

4.2. Epoxide Rubbers (CO, ECO, GECO, GPO)

Epoxide rubbers (designations according to ISO 1629) are derived from epichlorohydrin and from propylene oxide. The following epichlorohydrin elastomers are commercially available:

Epichlorohydrin Elastomers
CO. polychloromethyloxiran (epichlorohydrin elastomer) [24969-06-0]
ECO. ethylene oxide (oxiran – chloromethyl oxiran (epichlorohydrin) copolymer [24969-10-6]
GECO. epichlorohydrin – ethylene oxide – allyl glycidyl ether terpolymer [26587-37-1] (ASTM D 1418–90)
GCO. epichlorohydrin – allyl glycidyl ether copolymer [24969-09-3]

Propylene Oxide Elastomers
GPO. propylene oxide – allyl glycidyl ether copolymer [25104-27-2]. Propylene oxide elastomer is no longer commercially available.

Epichlorohydrin elastomers were first introduced in 1965 by B. F. Goodrich under licence from Hercules. GPO was introduced in developmental quantities (as Dyangen XP-139) by General Tire in 1964 and commercialized by Hercules (as Parel) in 1972.

All epoxide rubbers are specialty elastomers; their ozone and aging resistance is superior to general-purpose rubbers such as NR and SBR. *Epichlorohydrin rubbers* also have superior oil and fuel resistance, whereas *GPO* has excellent low-temperature flexibility and some oil resistance (comparable with CR). Its dynamic properties are similar to those of NR.

The chemistry of and markets for epichlorohydrin elastomers were reviewed in [413], [414]. A review of GPO is given in [415].

4.2.1. *Properties, Grades, and Applications*

Epichlorohydrin Rubber. The *epichlorohydrin homopolymer*, CO, has the structure $(CH_2–CH(CH_2–Cl)–O)_n$, i.e., it contains oxygen atoms in the polymer chain and 38 % chlorine as a substituent, which is responsible for its fuel resistance. The chlorine also imparts good flame retardance to the polymer. Epichlorohydrin homopolymer is outstanding in its low gas permeability and high ozone resistance, and is comparable to polyacrylate and NBR with medium acrylonitrile content in terms of oil resistance. The temperature flexibility of epochlorohydrin polymer is not satisfactory since its T_g is ≈ – 20 °C.

The 1 : 1 *epichlorohydrin – ethylene oxide copolymer*, ECO, $(CH_2CH(CH_2Cl)–O–CH_2–CH_2–O)_n$ contains 26 % chlorine. Its T_g is less than – 40 °C; hence its low-temperature properties are excellent. Epichlorohydrin – ethylene oxide copolymer is somewhat less fuel resistant than CO but still comparable in fuel resistance to a high-acrylonitrile NBR. The oil resistance of epichlorohydrin elastomers is intermediate between medium- and high-acrylonitrile NBR. ECO is less ozone resistant than CO and has a higher permeability to gas.

Epichlorohydrin polymers compounded with N 550 carbon black have enough electrical conductivity to dissipate static charges in belts, hoses, etc., since these rubbers are more polar than hydrocarbon-based rubbers.

Epichlorohydrin rubbers are subject to oxidative degradation that is initiated by hydroperoxide formation on the carbon adjacent to the oxygen in the chain. Oxidative degradation is counteracted by addition of protective agents such as nickel dithiocarbamates. With these agents the useful properties of CO are retained after 500 to 1000 h of air aging at 135 °C. Epichlorohydrin homopolymer (CO) retains its tensile strength better than ECO. After 1000 h at 135 °C CO retains its tensile strength while ECO loses 75 % of its strength. Epichlorohydrin rubber tends to soften in oil at elevated temperature (≥ 150 °C), or in gasoline containing hydroperoxides ("sour" gasoline). The softening has been attributed to scission of a carbon – oxygen bond according to the following mechanism [416]:

$$-CH(CH_2Cl)-CH_2-O-CH_2-CH_2- + RO\cdot(\cdot OH)$$

$$\rightarrow ROH(HOH) + -CH(CH_2Cl)-\dot{C}HOCH_2CH_2$$

$$-CH(CH_2Cl)-\dot{C}H-O-CH_2CH_2-$$

$$\rightarrow -CH(CH_2Cl)-CHO + \cdot CH_2-CH_2-$$

Metal dithiocarbamates promote this reaction because they catalyze the decomposition of hydroperoxides.

Epichlorohydrin rubber and ECO are not reactive enough to be vulcanized with sulfur, and require combinations of nucleophile, cross-linker, and acid acceptor, such as ethylene thiourea – lead oxide (1 : 5) as vulcanizing agents. GECO, containing reactive allyl glycidyl ether moieties, is cross-linkable with sulfur or peroxides as well as the cross-linkers used for CO and ECO [413]. When cross-linked by 2,4,6-trimercapto-s-triazine, GECO is lead free, is resistant to metal corrosion and heat softening, and has low compression set [417].

Uses. Epichlorohydrin rubbers are used in automotive radiator hoses, turbocharger hoses, and, if sour gasoline resistance is not a problem, fuel hoses and tubing. They are also used in suspension bushings, seals, gaskets, and coated fabrics, in rubber-covered rolls, and in oil-field specialties.

Propylene Oxide – Allyl Glycidyl Ether Copolymer. GPO is a copolymer of propylene oxide typically with 6 wt % allyl glycidyl ether. It has a T_g of –75 °C and a service temperature range of –60 °C to 135 °C. GPO is inherently ozone resistant.

Because of the reactive allyl glycidyl units the copolymer can be cross-linked by sulfur in the presence of accelerators such as tetramethylthiuram monosulfide (TMTM) and 2-mercaptobenzothiazole (MBT). Peroxides tend to degrade the polymer and are not suitable as cross-linking agents. The properties of GPO vulcanizates are generally similar to those of natural rubber or polychloroprene. The vulcanizates are moderately resistant to oils and aliphatic solvents.

Uses. GPO is principally used in engine and body mounts, because its dynamic characteristics are similar to those of NR over an even broader temperature range [419]. Possible applications include suspension bushings, dust seals, and boots.

4.2.2. *Basic Chemistry and Production Processes*

Epichlorohydrin Elastomers. The development of epichlorohydrin elastomers has been described [418]. Polymerization via a cationic coordination mechanism is carried out in solution in aliphatic or aromatic hydrocarbons, chlorinated hydrocarbons, or ether, with a catalyst formed by reacting R_3Al and water (thought to be $R_2Al–O–AlR_2$). Catalyst activity is improved by the addition of acetylacetone. Some combinations of organozinc and organomagnesium compounds, as well as other additives and chelating agents in combination with alkylaluminum compounds, are also effective catalysts. The polymer is produced in a continuous process using a back-mixed reactor. Polymerization proceeds at reasonable speed at room temperature but can be accelerated by raising the temperature to 50 °C.

Propylene Oxide – Allyl Glycidyl Elastomer. The production of GPO can be carried out in aliphatic, aromatic, or chlorinated hydrocarbon solution by using R_3Al – water – acetylacetone, diethylzinc – water, or complex cyanides as catalysts. Complete conversion and a uniform polymer with the composition of the original monomer feed are obtained.

4.2.3. *Producers and Production Capacities*

Producers of epoxide rubber are listed in Table 26 together with their production capacities.

Table 26. Producers, production capacities, and trade names of epoxide rubbers (based on data from International Institute of Synthetic Rubber Producers Inc.)

Producer (country)	Production capacity, t/a	Trade name
Zeon Chemicals (United States)	9 900	Hydrin H (CO)
		Hydrin C (ECO)
		Hydrin T (GECO)
Nippon Zeon (Japan)	1 500	Gechron 1000 (CO)
		Gechron 1100 (GCO)
		Gechron 2000 (ECO)
		Gechron 3100 (GECO)
Osaka Soda (Japan)	1 200	Epichlomer

4.3. Polynorbornene

Polynorbornene is the common name of polymers obtained by homopolymerization of bicyclo[2.2.1]hept-2-ene [498-66-8] called 2-norbornene. Since the 1940s this highly strained reactive molecule has been subjected to numerous polymerization studies applying many different types of catalysts [420]. 2-Norbornene polymerization can occur via two different modes: ring-opening metathesis polymerization and vinyl-type polymerization. Metathesis is mainly a reaction between two olefins, characterized by cleavage and reformation of carbon – carbon double bonds. With cyclic olefins, it becomes a ring-opening polymerization reaction [421], [422]. Vinyl-type polymerization maintains the cyclic structures. Basically four different polymeric structures can be obtained, depending on the type of catalyst or cocatalyst. Most frequently, a combination of two or three of these structures is observed.

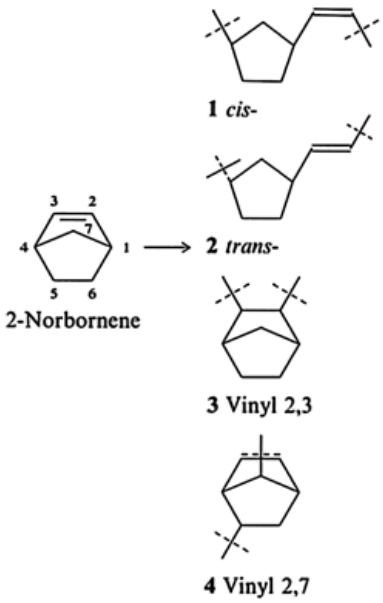

In 1970, the French company CdF Chimie began intensive studies on the polymerization of 2-norbornene, which led to a commercial polynorbornene that has been produced on an industrial scale since 1976. This polymer possesses molecular structure **2** and is characterized by an unusually high molar mass (> 3×10^6 g/mol) and a glass transition temperature of 37 °C; therefore it can be regarded as an intermediate between elastomers and thermoplastics. However, it is highly compatible with plasticizers and acquires elastomeric properties by incorporation of these substances.

Polynorbornene is presently produced in France by Elf Atochem and developed worldwide in its main applications by Nippon Zeon, under the trade name Norsorex.

4.3.1. *Production of 2-Norbornene Monomer*

2-Norbornene is most commonly synthesized by Diels – Alder addition of cyclopentadiene (CPD) and ethylene. Cyclopentadiene is itself produced by cracking of dicyclopentadiene (DCPD) at > 160 °C. Dicyclopentadiene is extracted from the C_5 streams from naphtha steam cracking (⟶ Ethylene – Cracking Conditions, ⟶ Ethylene – Heat Requirements for Hydrocarbon Pyrolysis, ⟶ Ethylene – Commercial Cracking Yields, ⟶ Ethylene – Commercial Cracking Furnaces; ⟶ Resins, Synthetic). The addition reaction occurs without catalyst at high temperature and high pressure.

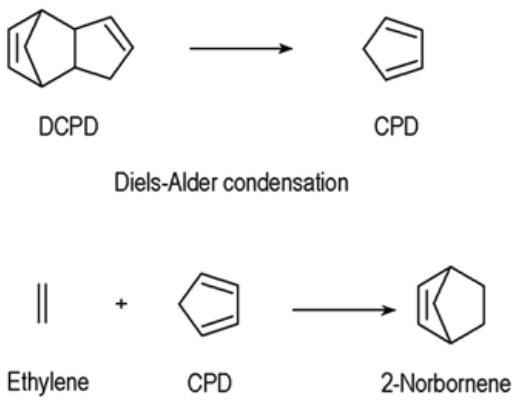

Because of the impurities contained in commercial DCPD (about 5 %) and the byproducts formed by reaction of 2-norbornene with the raw materials, a thorough purification of 2-norbornene is required after the synthesis step, in order to obtain a polymerizable monomer (purity > 99.7 %). Pure 2-norbornene is solid at room temperature (*mp* 47 °C) and must be protected from oxidation.

4.3.2. *Polymerization*

Commercial polynorbornene (Norsorex) is produced by a metathesis type of polymerization, catalyzed by a transition-metal salt [423]. The chemical structure of the polymer results essentially from a ring-opening polymerization, preserving one double bond per repeating unit with mainly *trans* configuration. According to the IUPAC system, its name is poly(1,3-cyclopentylenevinylene) [25038-76-0].

Polymerization is extremely rapid and highly exothermic. By application of a special technology the polynorbornene obtains a convoluted and porous structure. After cooling and drying, the polymer appears as a white, free-flowing, amorphous powder, with a grain size of < 0.8 mm and an apparent density of 0.3 g/cm^3. The porous structure and the presence of double bonds render polynorbornene very sensitive to oxidation. Therefore, a nonstaining antioxidant is added during its manufacture.

4.3.3. *Properties*

4.3.3.1. *Properties of Pure Polymer*

The chemical structure, the uncommonly high molar mass, and the unusual transition temperature are responsible for the unique properties of polynorbornene. Typical properties are given below:

Molar mass	> 3×10^6 g/mol
T_g	37 °C
Density	0.96 g/cm^3
Refractive index	1.534
Solubility parameter	19.8 MPa$^{1/2}$
Thermal conductivity	0.285 W m^{-1} K^{-1}

Dielectric constant (1 MHz)	2.5
Dissipation factor (1 MHz)	10^{-4}
Volume resistivity	$7\times10^{16}\ \Omega\cdot cm$

The thermal behavior of the product classifies it as a thermoplastic with a very low glass transition temperature. Polynorbornene is impossible to melt because of its high molar mass and decomposes at > 200 °C without real fluidization. The polymer exhibits a “shape memory effect” (it recovers its memorized shape when heated to more than 37 °C) [424]. Pure polynorbornene is used in only a few minor applications.

Polynorbornene has a very high affinity for liquid hydrocarbons, promoted by its porous structure. Pure polynorbornene is considered a highly efficient antipollutant because of its ability to absorb and gel usually ten times its weight of common hydrocarbons. Because of the compatibility of polynorbornene with extension oils that act as plasticizers, the transition temperature of the polynorbornene – oil mixture is lowered to – 60 °C, converting the plasticized polynorbornene into an elastomer. The presence of double bonds in the polymeric chain allows further vulcanization with common peroxides or sulfur vulcanization systems.

4.3.3.2. *Properties of Polynorbornene Rubbers (PNR)*

Polynorbornene rubbers are unique because of the high level of plasticizers and fillers that can be incorporated. Plasticizers can be applied at up to 500 parts per 100 parts of polymer; fillers, at up to 2000 parts. Most usual PNR compounds contain on average 20 % pure polynorbornene. This explains the prime importance of an appropriate selection of the compounding ingredients, both qualitatively and quantitatively. The properties of the resulting vulcanizates will thus vary over a wide range and can be “custom-tailored.”

Only some general physical and chemical properties are related strictly to the polymer itself. These can be summarized as follows: good aging and heat resistance (up to 90 °C), excellent mechanical strength, low compression set, and high compatibility with most other rubbers. Polynorbornene rubbers exhibit poor ozone resistance when unprotected; however, efficient protection is obtained by addition of antiozonants or a small amount of EPDM rubber.

Influence of the Plasticizer. The amount of plasticizer included determines the softness of the vulcanizate; unusually low hardness values down to 10 Shore A can be reached in combination with good mechanical properties. The type of plasticizer used affects the cold resistance, energy absorption capability, and compatibility with other rubbers. Figure 20 shows the variation of elasticity as a function of temperature for a given PNR compound in which only the type of oil used as plasticizer varies. This illustrates how the shock absorption ability can be adjusted; in this example a rebound between 5 and 60 % is obtained at room temperature. Table 27 shows the dependence of the cold resistance (brittle point) of the vulcanizates on the type and level of plasticizer.

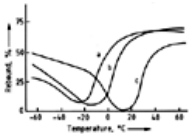

Figure 20. Influence of oil type on rebound

a) Paraffinic oil; b) Naphthenic oil; c) Aromatic oil (high viscosity)

Table 27. Influence of plasticizer type on brittle point

Type	Level, phr				
Aromatic oil	180			160	
Naphthenic oil		180			180
Paraffinic oil			180	20	
Dioctyl adipate					40
Brittle point, °C	–30	–42	–49	–38	–52

Influence of Filler. All types of mineral fillers and carbon blacks can be used with PNR. The importance of their effect depends on their concentration in the rubber compound. The type of filler influences mainly physical properties such as strength, abrasion, and tear resistance; electrical properties; and dynamic damping properties. Table 28 lists the mechanical properties of a given PNR compound containing 200 parts of naphthenic oil and 200 parts of various fillers.

Table 28. Influence of filler type

Characteristic	Type of filler					
	Whiting	MT	SRF	GPF	FEF	HAF
Hardness, Shore A	28	35	43	53	57	61
Tensile strength, MPa	10.5	11.8	14.6	14.8	15.1	16.9
Elongation, %	525	507	369	322	276	268
Rebound at 20 °C, %	55	56	45	35	34	22.5

Figure 21 illustrates the dynamic damping behavior of PNR compounds filled with two different carbon blacks (SAF, MT) in comparison to the elastic natural rubber and the dampening chlorobutyl rubber.

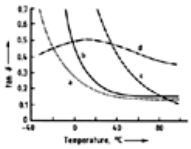

Figure 21. Damping factor of some rubber compounds

a) Natural rubber; b) Norsorex plus MT; c) Chlorobutyl rubber; d) Norsorex plus SAF

Frequency: 125 Hz; Applied force: 10 N

4.3.4. ***Processing of PNR***

Both internal mixers and open mills can be used for PNR compounding. The equipment must be preheated at ca. 80 °C to avoid possible degradation of

the long polymeric chains at lower temperature due to mechanical rupture. The process always involves a preliminary absorption of oil (i.e., plasticizer) by the powdered polymer, so that it becomes an elastomeric material that can be submitted to shearing. For the production of very soft compounds, a masterbatch with a limited amount of oil (150 parts) is generally mixed first; the additional ingredients are introduced in a second mixing step. Subsequent processing is performed with common equipment in the rubber industry (⟶). The compound can also be prepared as a free-flowing dry blend in a high-speed powder blender. The dry blend may be extruded directly or injection molded.

4.3.5. *Uses*

PNR is a specialty elastomer, used mainly in applications where its unique properties are in demand. The most important developments are in the field of low-hardness solid parts and elastomers with tailored damping properties. These characteristics are exploited in very different industries such as the automotive, appliance, office equipment, and leisure industries for the production of molded or extruded goods such as seals, boots, damping pads, bumpers, rolls, and shoe insoles.

PNR is also used in blends with other elastomers to improve some of their characteristics such as green strength, hardness, or damping.

4.3.6. *Economic Aspects*

Worldwide production capacity of polynorbornene is presently 5000 t/a in the French plant of Elf Atochem (Carling). Consumption of pure PNR was about 1000 t in 1991, corresponding to about 5000 t of compounded elastomer.

4.3.7. *Toxicology*

Polynorbornene powder is nonmutagenic (Ames test) and nontoxic: acute oral toxicity for rats is 11 g/kg.

4.4. Polyoctenamers

In the ring-opening metathesis polymerization of cycloolefins, unsaturated polymers are formed — the so-called polyalkenamers [425] or polyalkenylenes [426] (both names are used in the literature and in *Chemical Abstracts)*. Only polynorborene (see Section Polynorbornene), polydicyclopentadiene, and polyoctenamers [28702-45-6] are produced on an industrial scale [427], [428]. At first, research on polyoctenamers concentrated on polymers rich in *cis* structures [429], [430]. However, for economic reasons an industrial breakthrough was achieved only with the *trans*-polyoctenamers [28730-09-8] [431].

4.4.1. *Production of Cyclooctene*

Cyclooctene [931-88-4] is obtained by partial hydrogenation of 1,5-cyclooctadiene [111-78-4]. Originally, the latter was formed as a byproduct in the synthesis of 1,5,9-cyclododecatriene, an intermediate produced on an industrial scale by Hüls AG, now Degussa AG [432]. Because of increasing demand for 1,5-cyclooctadiene the dimerization of butadiene, known since 1953 [433], [434], is also used for its production. Instead of the original Reppe-type catalysts, modified Ziegler – Natta catalysts with zerovalent nickel are now used for dimerization [435], which give 1,5-cyclooctadiene with complete conversion of starting material and high selectivity at elevated temperature. The crude product is separated from byproducts and catalyst residue by flash evaporation. The purity of this intermediate is > 99 %.

The selectivity of the hydrogenation of 1,5-cyclooctadiene to cyclooctene is slightly less than 100 % when commercial heterogeneous catalysts are used. With precise temperature control and hydrogen metering a specially developed hydrogenation catalyst gives complete selectivity. A small degree of overhydrogenation to cyclooctane must be put up with. Otherwise, traces of 1,5-diene can remain in the product that may hinder the sensitive polymerization process, because the 1,5-diene tends toward double-bond isomerization to 1,3-cyclooctadiene. Especially the latter acts as a strong catalyst poison [436-438]. The final product with the purity necessary for polymerization contains 95 – 97 % cyclooctene; the remainder is essentially cyclooctane.

4.4.2. *Polymerization*

In the first ring-opening polymerization of cyclooctene in 1965 by NATTA et al., tungsten or molybdenum salts in combination with organoaluminum compounds were used as catalysts [439]. In 1967, CALDERON et al., succeeded in improving polymerization conditions by using the catalyst system WCl_6 – $EtAlCl_2$ – EtOH. At the same time they reported an extractable oligomer fraction in the polymerization of cyclooctene and other cycloolefins. They concluded that the process was entropy controlled because of the low heat of reaction.

Following that time, many other articles on the metathesis polymerization of cycloolefins, including cyclooctene, have been published. Most of the original work is summarized in three monographs [440-442] and many review articles [429], [430], [443-448].

The mechanism of ring opening during polymerization is still being disputed. The nature of the metal complex formed as a transition state before ring opening, which initiates the actual polymerization, is not certain.

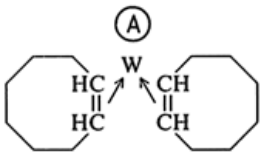

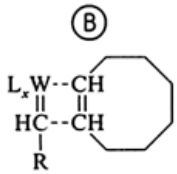

One proposal is that two monomer molecules form a bisolefin – metal complex (A) that undergoes ring expansion to give a cyclic diolefin with double the number of ring members. Since the double bonds of this dimer have the same reactivity as those of the monomer, the ring-expanding metathesis can continue until macrocyclic polymers are obtained.

The alternative theory states that a metal carbene – olefin complex (B) is formed, in which only three carbon atoms participate in the reaction with the metal atom. This leads primarily to linear polymers. The formation of cyclic oligomers is explained by backbiting. This transition state is the preferred model. It is in good agreement with the polymerization pathway, which is assumed to be a chain reaction. The formation of macrocyclic polymers is less easily explained with this model. Since macrocycles with molar masses up to 100 000 g/mol have been detected in polyoctenamers [449], [450], transition state (A) cannot be ruled out completely.

Both theories have advantages and disadvantages. Many questions remain unanswered because of the complex nature of the reaction. All the reactions taking place — for example, stereospecific reactions, oligomerization and degradation of linear and cyclic molecules, and reactions regulating the molar mass — are equilibrium reactions. They are as yet not completely understood theoretically. A detailed discussion of the mechanism of the ring opening and the propagation reaction is given in [442]. However, the polymerization can be controlled reproducibly to a great extent by catalyst composition and reaction conditions.

On an industrial scale, polymerization is carried out in hexane. High-purity monomers and anhydrous conditions are essential. A tungsten-based, modified Calderon catalyst is used. The polymerization is carried out adiabatically, goes to complete conversion, and gives almost 100 % yield. The workup involves flash evaporation of the solvent with subsequent recycling. The final product is filtered in the melt, pelletized after cooling, and packed in bags. Its purity is

> 99.5 %. The content of monomer (mainly cyclooctane) and low molar mass oligomers is very small.

4.4.3. *Properties*

Polyoctenamers have been produced over wide ranges of both molar mass and *cis : trans* ratios of double bonds [429], [451], [442]. All products exhibit bimodal molar mass distributions, which are typical for metathesis polymerization, with the first maximum at low molar mass oligomers. The crystallinity of polyoctenamers depends strongly on their microstructure (i.e., the ratio of *cis : trans* double bonds). The double bonds can be arranged in sequences, and crystallites with defined melting temperatures are then formed. With respect to the crystallinity the effect of the *trans* double bonds is more pronounced than that of the *cis* double bonds. Recrystallization from the melt occurs within seconds. This effect is very important in certain applications involving rubber processing.

Depending on the reaction mechanism, linear, unbranched polymers with an ideal polymethylene structure are formed in the polymerization, in addition to macrocycles. A double bond occurs at every eighth carbon atom of the chain or macrocycle. This explains the high chain mobility and the low glass transition temperature T_g.

Two polyoctenamer grades are available commercially from Degussa AG under the trade name Vestenamer. The low molar mass and crystallinity are chosen specifically so that they give advantages in rubber processing because of their low melt viscosity. In comparison to common baled rubbers the handling of polyoctenamers is simplified because they are delivered as pellets. Data for these two Vestenamer grades are given in Table 29.

Table 29. Typical polymer properties of Vestenamer 8012 and 6213

Property	Vestenamer grade	
	8012	6213
Molar mass, g/mol	70 000	90 000
Glass transition temperature T_g, °C	–65	–75
Crystallinity (at 23 °C), %	30	12
mp, °C	54	< 36
Start of thermal decomposition, °C	275	250
Ratio of *trans : cis* double bonds, %	80 : 20	60 : 40
Mooney viscosity ML (1 + 4) (at 100 °C)	< 10	< 10
Viscosity number *J* (at 23 °C), mL/g	120	130
Density, g/cm^3	0.91	0.89

4.4.4. *Vulcanization, Processing, and Applications*

The molar mass of Vestenamer is kept intentionally low for a rubber. Because of the thermoplasticity the viscosity in the molten state is unusually low compared to other solid rubbers: The Mooney value ML (1 + 4) at 100 °C is only 5 – 8. Because of its high degree of unsaturation, Vestenamer can be vulcanized with all the cross-linking agents commonly used in the rubber industry (e.g., sulfur, sulfur donors and accelerators, peroxides, and vulcanization resins). Therefore, Vestenamer has two functions when used as a blend component in rubber compounds, in which 5 to 20 parts of the basic rubber are generally replaced. Because of its low viscosity, Vestenamer acts as a *plasticizer* or processing agent in compounding and molding. However, after vulcanization it is *cross-linked* like all other rubbers, completely incorporated into the network, and can no longer be extracted. This means lower overall energy consumption for the mixing process, easier filler incorporation, better filler dispersion, and a smaller temperature increase during mixing. In extrusion processes the mass pressure and die swell are decreased, and the surface finish of the extrudates (e.g., profiles or tubing) is improved. The high rate of crystallization of Vestenamer leads to a significantly higher green strength of the compounds. Thus the rigidity and dimensional stability of extrudates, which are stored before vulcanization, are increased markedly. Examples are tire bead and apex compounds, profiles, and braided hoses. Because of the improved flow properties of the mixtures, the mold-filling process can be improved and shortened in the production of molded articles, and complicated mold geometries are cleanly produced. Because of the rapid recrystallization of Vestenamer 8012 the anisotropy of the calender shrinkage — i.e., the different shrinkage in the longitudinal and transverse directions — is significantly reduced. Even very thin sheets can therefore be calendered without defects. The tack of the compounds is reduced due to the completely unbranched polyoctenamer molecules in combination with their crystallinity. (This is a disadvantage in NR compounds for tire production, but an advantage in CR compounds [451-454].)

The *vulcanizate* properties of rubber – polyoctenamer blends are barely affected compared to compounds without Vestenamer. Tensile strength and tear resistance are decreased somewhat because of the low molar mass of the polyoctenamer. The elongation at break decreases and the modulus increases. The hardness is also increased because the residual crystallinity is noticeable after cross-linking. The ozone resistance is not affected. In blends of polar and nonpolar (NBR – EPDM) or emulsion and solution rubbers (NR – BR), addition of small quantities of Vestenamer (5 – 10 phr) clearly increases dispersion of the otherwise mutually not very compatible rubbers, and processing is thus improved.

Over the years new applications have been commercialized, some of them outside traditional rubber processing, e.g., in the modification of plastics. The dynamic properties of natural rubber compounds are improved on blending with Vestenamer, and reversion is suppressed at the same time [455]. These results open up interesting possibilities for tire compounding and production. Interest is also increasing in the use of Vestenamer — blended with other rubbers or as the only rubber component — for the production of extremely hard but nevertheless readily processible tire apex compounds or golf balls [456].

In tire compounding blending with Vestenamer lowers the shrinkage of inner liner compounds and offers the possibility to partly dispense with the expensive silane filler activator in silica-filled tire treads [462].

The pronounced thermoplasticity and low melting point permit the use of Vestenamer as a carrier material for filler batches. The filler incorporation by Vestenamer is high, and the rapid recrystallization from the melt allows granulation of the batches. Vestamer is often used in combination with other carrier materials to inhibit their undesired tackiness. Batches using Vestenamer as carrier material for accelerators, peroxides, very fine powders, or even liquid rubber chemicals, are commercialized.

Vestenamer – oil mixtures are used as coating and binding agents for ground rubber waste powder [457]. The importance of this process for reduction of the large amounts of waste rubber was met with increasing interest during the last years, because recycling of waste rubber is technically difficult.

Based on the experience with rubber recycling, Vestenamer is since 1999 used in rubberized asphalt [458]. Crumb rubber (ground tire rubber, GTR) enhances the durability and performance of asphalt concrete pavements. But the hot GTR-asphalt mix exhibits high tackiness, which causes problems in processing and handling (separation of GTR in the terminal blending tank, sticking to the rolling equipment). Addition of a small amount of Vestenamer (about 0.5 %) successfully overcomes these problems, and increases the PG (Performance Graded) ratings of the asphalt. Vestenamer functions as reactive modifier and forms a uniform, low-tack, rubber-like composite from GTR and asphalt [459, 460]. This patented mixing process [461] reduces the costs of paving. A number of test roads demonstrate the high quality of pavements with rubberized asphalt containing Vestenamer.

4.4.5. *Economic Aspects*

Detailed economic and ecological analyses of the metathesis processes used industrially show clearly their advantages and disadvantages [436]. Analysis makes clear why polyoctenamers can be produced on a commercial scale, unlike polypentenamers, which were intended for very similar applications.

Vestenamer production began in a pilot plant in 1980, and in 1989 a large plant with a capacity of 10 000 – 12 000 t/a was put in operation. Its nominal capacity was reduced to 3000 t/a after Hüls AG had finished their activities in the rubber business by December 31, 1994.

In 2002, 38 % of the Vestenamer produced were sold in Germany, 34 % in the rest of Europe, 23 % in both Americas, and 5 % in Asia. The following quantities of Vestenamer were consumed in different areas of the rubber industry:

Tires	44 %
Extrudates	18 %
Molded articles	7 %
Calendered articles, rubber soles, roll covers	4 %
Recycling	12 %
Custom compounding	5 %
Other	10 %

Vestenamer is used predominantly in blends with the following rubbers: NR, SBR, EPDM, NBR, CR, CSM, and pre-cross-linked butyl rubber.

4.4.6. *Toxicology*

Vestenamer and its precursors have been thoroughly investigated with regard to their toxicology.

1,5-Cyclooctadiene has an LD_{50} of 1900 mg/kg in rats (oral) and of > 10 000 mg/kg in rabbits (dermal). It has a slight irritant effect on the skin (OECD test 404), but shows no sensitizing effect in the corresponding test according to Magnusson and Kligman. It is not irritating to the eyes or mucous membranes (OECD test 405).

Cyclooctene has an LD_{50} of 4550 mg/kg in rats (oral) and of > 10 000 mg/kg in rabbits (dermal) and a slight irritant effect on the skin (OECD test 404). It shows no sensitizing effect in the corresponding test according to Magnusson and Kligman, and is not irritating to the eyes or mucous membranes (OECD test 405).

Trans-Polyoctenamer has an LD_{50} of > 12 500 mg/kg in rats (oral) and has no irritating effect on the skin or on rabbit eyes. In rats, 90-d oral administration at concentrations up to 4000 mg/kg body weight did not result in any toxic effects. Vestenamer showed no mutagenic effects in the Ames test (in vitro) on *Salmonella typhimurium* and the micronucleus test (in vivo) on mice

The Scientific Committee on Food of the European Commission evaluated cyclooctene for food contact and approved it for use in aqueous foodstuffs (restriction: specific migration limit (SML) 0.05 mg/kg). Cyclooctene has been added to Section A of the Positive List for Monomers of the EU-Plastics Directive 90/128/EEC (6th Amendment). Thus Vestenamer can be used safely in contact with aqueous foodstuffs (but not in contact with alcoholic and fatty foods).

[Top of Page]

5. Synthesis by Polyaddition, Polycondensation, and Other Mechanisms

Jürgen Ackermann, Dieter Wrobel, Uwe Hoffmann, Hans Dieter Thomas, Rüdiger Engehausen, Stephen D. Pask

5.1. Silicone Rubber

Silicone elastomers (Q) form the most important class of rubber composed of repeating inorganic Si–O instead of C–C units in the polymer chain. Silicone polymers can bear various organic substituents on silicon.

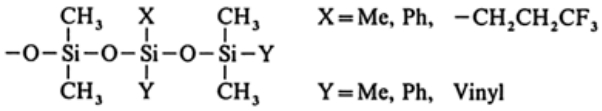

According to ASTM 1418, the initial letters of substituents X and Y (methyl, phenyl, vinyl, and fluoralkyl) are used before the letter Q to differentiate among the individual Q grades (MQ, VMQ, PVMQ, FMQ, FVMQ). ISO 1629 uses the same letters but in a different order (MVQ, PMVQ, FMVQ).

Pure polydimethylsiloxanes with cross-linkable vinyl or SiOH groups as side- or end groups are by far the most important products of the silicone industry. Different substitution of siloxy groups and degrees of polymer chain lengths give rise to an extraordinary number of products and a broad application range [463]. Peroxide -cured silicone rubbers (hot vulcanization, HV) have been marketed since 1944, shortly after the industrial production of silicones began in 1939 [464]. The low strength of unfilled silicone elastomers was improved by the introduction of “white soot” (fumed or pyrogenic silica in 1942) as a reinforcing filler and by incorporation of cross-linkable vinyl groups into the polymer.

Low-viscosity two-component rubbers, which can be cured at room temperature (room temperature vulcanization, RTV), followed in 1954, and in 1958 the first one-component RTV rubber for sealants. In 1976 liquid silicone rubbers (LSR) appeared in the market.

5.1.1. *Properties, Grades, and Applications*

Commercial silicone rubbers are divided into four main groups depending on their applications and viscosities. These can, in turn, be differentiated according to cross-linking mechanism and polymer type [465], [466].

1. Solid rubbers [(HV = hot vulcanization rubber)]: VMQ, FVMQ, PVMQ
2. Liquid silicone rubbers (LSR, two-component): VMQ, FVMQ
3. Molding (casting) compounds (RTV-2 = two-component room temperature vulcanization): MQ, VMQ, FVMQ, PVMQ
4. Sealants (RTV-1 = one-component room temperature vulcanization): MQ, PMQ, FMQ

Solid silicone rubbers are marketed mostly as HV base compounds without peroxide. *Liquid Silicone* Rubbers are sold exclusively as two (A and B) components, which are formulated ready for use. For *casting* or *molding compounds* as well as *sealants* silicone producers also supply raw materials.

HV rubbers are kneadable, noncreeping compounds mostly with a yield point. Their Mooney viscosities at room temperature are comparable to those of other rubbers at 80 – 140 °C. The viscosities of the other silicone rubbers are so low that they can be pumped with commercial metering pumps (piston type) or are free-flowing. Unlike the other rubbers one-component RTV compounds are formulated to exhibit a strong yield point, which hinders undesired flow out of vertical joints [466]. The different viscosity ranges for Q rubbers at a shearing rate $\dot{\gamma} = 1\ s^{-1}$ are as follows:

HV	40 000 – 200 000 Pa · s (Mooney: 20 – 100)
LSR	100 – 5 000 Pa · s
RTV-2	10 – 50 Pa · s

RTV-1	40 – 100 Pa · s

The Q elastomers made from solid HV or liquid silicone rubbers have better mechanical properties than one- or two-component RTV compounds (see Table 30) [467], [468]. HV and LSR are mainly used for production of heat-resistant molded or extruded articles.

Table 30. Mechanical properties of Q rubbers

	HV	LSR	RTV-2	RTV-1
Hardness, Shore A	20 – 90	20 – 80	0 – 90	15 – 25
Tensile strength, MPa	6 – 12	6 – 10	0.5 – 6	1 – 3
Tear ASTM 624B, N/mm	10 – 45	15 – 50	10 – 20	3 – 8

One-component RTV sealants are characterized by good *adhesion* to a wide variety of surfaces. The low tensile strength of these rubbers inhibits their application in the fabrication of molded or extruded articles.

Two-component RTV rubbers used for casting and molding applications have low *viscosities* and undergo vulcanization at room temperature.

The Q elastomers can be characterized by the following properties in comparison with other synthetic rubbers:

1. Higher low-temperature flexibility
2. Higher oxidation and hot air resistance
3. Higher UV and ozone resistance
4. Very high gas permeability
5. High transparency
6. High electrical resistivity
7. Low surface tension
8. Low temperature dependence of mechanical properties
9. Physiological compatibility

The relaxation and elastic properties of silicone rubbers are nearly constant over a broad temperature range of – 45 to 200 °C [470]. Rubbers containing more than 4 mol % diphenyl- or methylphenylsiloxy groups (PVMQ) have an equilibrium melting point T_m of – 70 °C instead of – 45 °C (brittleness point – 105 °C instead of – 55 °C) while the glass transition temperature T_g remains unchanged at – 120 °C [471]. The transparency, the resistance to oxidation and irradiation are also increased by phenyl groups. The introduction of additional siloxy groups with polar substituents, such as CN- or $F_3CC_2H_4$-groups, into the polymer chain reduces swelling in nonpolar solvents, e.g., aliphatic or aromatic fuel [472].

Oxides or complexes of Ti, Zr, Mn, or Ce are used as stabilizers against aging by hot air, preventing oxidation of methyl and alkylene groups under formation of additional cross-links. Exposure to acids (oxidized oils), amines, or steam leads to depolymerization. The action can be suppressed by alkaline earth oxide, carbonate or silicate fillers. VMQ exhibits 50 % residual elongation even after aging under extreme conditions, i.e., after 60 d at 250 °C or 21 d at 300 °C in air, 30 d at 160 °C in steam, 2000 h at 150 °C in ASTM 2 oil. For silicone foams, see ⟶ Foamed Plastics – Silicone (SI) Foams.

Elastomer blends of VMQ with other peroxide curable polymers can be formulated. Blends with other rubbers, tetrafluoroethylene – hexafluoropropylene (FKM) or ethylene – propylene – diene (EPDM) and others are described in the patent literature and are produced commercially [473], [474]. A large number of siloxane block copolymers with thermoplastic hard segments such as polycarbonate, polysulfone, polyimide, or silarlylene are known. They exhibit properties of thermoplastic elastomers [468], [469], [475].

Applications. Uses of different silicone rubbers as a function of their properties are listed in Table 31. Typical articles produced from silicone rubber include: cable insulations, seals and tubing for food and processing equipment, key(board)pads, baby nipples, and pacifiers, shaft seals, anode caps, spark plug boots, membranes, implants, drug release carriers, electronic encapsulation, photocopier rolls, molds for replication of art or denture, window or sanitary seals, adhesives, adhesive controlled release coatings for paper, waterproof and thermal-shielding [465], [476], [477].

Table 31. Properties of Q rubbers that determine application

Grade	Processing	Main property
HV	extrusion	toughness and relaxation –105 to 300 °C
HV/LSR	molded articles	toughness and relaxation – 105 to 300 °C, short cycle times
HV/RTV-2	foam	heat insulation, flame resistance
RTV-1	joints	weathering resistance, adhesion
LSR/RTV-2	coating	surface tension, adhesion
RTV-2	castings, sealing	room temperature vulcanizaton, flow properties
	reproduction	room temperature vulcanization, electrical insulation
	gels	low hardness, damping

5.1.2. *Production Process and Basic Chemistry*

The starting materials for the production of silicone rubbers are linear silicone polymers with reactive SiOH or vinyl end groups. For production of silicone polymers, see ⟶ Silicon Compounds, Organic.

Rubber Compositions. The compositions of different silicone rubbers are given in Table 32. The silicone polymers differ in the lengths of their SiO chain, functionalities, and vinyl or SiOH contents. Fumed or precipitated silicas are used almost exclusively as reinforcing filler in silicone rubber. As exception carbon black is used in electrically conductive rubbers. The performance of silica fillers can be improved by adding hydrophobizing agents acting either before or during the mixing process [478]. These process aids, sometimes incorrectly called plasticizers, are based on either polysiloxanediols, disilazanes or alkoxysilanes [479], [480]. Process aids lower the structural viscosity and prevent the crepe hardening (increase in viscosity during storage) of the compound while simultaneously drying the filler [481]. Quartz powder, diatomaceous earth, micas, talcs, chalks, powdered alkaline earth oxides and of Fe, Ti or Al can be used as inert fillers.

Table 32. Composition of silicone rubbers (in parts by weight per 100 parts by weight of Q polymer)

Component	RTV-1	RTV-2	LSR	HV
Polymer chain length	1000	100 – 1000	500 – 1000	6000
Functionality	SiOH	SiOH/Si vinyl	Si vinyl	Si vinyl
Silica	5 – 10	0 – 30	0 – 40	0 – 75
Vinyl resin		0 – 30	0 – 30	0 – 20
Process aid		0 – 10	0 – 10	0 – 15
Plasticizer	0 – 30	0 – 20		
Inert filler	0 – 100	0 – 100	0 – 50	0 – 300
Cross-linker				
1	$RSi(OR)_3$	$Si(OR)_4$	SiH – siloxane	
2		SiH – siloxane		SiH – siloxane
Catalyst				
1	Sn	Sn	Pt	peroxide
2		Pt		Pt

Moisture-cured one-component RTV compounds are formulated as ready-for-use rubbers. Silanes, such as methyltriacetoxysilane, which undergo hydrolysis upon contact with ambient moisture, act as cross-linker. Fumed silicas (partially hydrophobized) with low water content and chalks coated with stearic acid are common fillers. The cross-linking reaction is catalyzed by a tin compound (e.g., dibutyltin dilaurate). Nonfunctional polydimethylsiloxane fluids are used as plasticizers [482].

Two component RTV compounds are based on polymers with SiOH or vinyl groups and little or no reinforcing filler to achieve lowest viscosities possible. The SiOH containing polymers are cured by condensation, those containing vinyl groups by hydrosilation. Q-resins with vinyl groups enable improved cross-linking density and reinforcement without silica filler. These resins are common in hydrosilation -cured, low viscosity RTV-2 compounds [483]. The formulations of different two compound RTV systems are given in Table 33.

Table 33. Formulation of the two-component systems RTV-2 and LSR *

Component	Condensation cross-linking		Addition cross-linking	
	A component	B component	A component	B component
SiOH polymer	X			
SiVi polymer **			X	X
Filler	(X)		(X)	(X)
SiVi resin **			(X)	(X)
$Si(OR)_4$ cross-linker		X		
Water		X		
SiH cross-linker				X
Sn – Ti catalyst		X		
Pt catalyst			X	
Inhibitor				X

* x = substance is contained in component; (x) = substance that can be optionally incorporated.
** Silicone polymer containing vinyl groups.

Production Process. Solid HV-rubber is produced on roll mills, in kneaders, or in screw mixers [484], [485]. In contrast to other synthetic rubbers silicone rubber compounds are not prone to significant shear heating (heat-build-up) or polymer degradation because of their low viscosity. VMQ rubbers can be heated up to 180 °C to render the filler hydrophobic (by condensation of the process aids with the silanol groups of the filler surface) [480], [485]. Transparent or opaque “basic compounds” are obtained, which are optionally mixed with inert fillers, stabilizers, pigments, and finally with peroxides at 20 – 120 °C on roll mills or in kneaders and extruders.

The *liquid silicone rubbers* RTV-1 and RTV-2 can be formulated in mixers of the coating and paste industry [486]. Planetary mixers and dissolvers are generally used, but kneaders and screw extruders are also suitable. The mixing processes must be carried out under an inert gas because of the moisture sensitivity and at higher temperatures due to danger of ignition of low boiling ingredients.

Vulcanization and Cross-Linking Reactions.
Peroxide Cure. Radical cross-linking of pure methylsilicone rubbers with peroxides leads to a low cross-link density. Vinylsiloxanes are considerable more reactive to give improved cross-link density. Reaction of methyl or vinyl groups with the peroxide initiator generates propene, butene, and a few ethylene bridges [487].

Alkyl- or aroyl peroxides such as dicumyl peroxide, 2,5-dimethyl-2,5-bis(*tert*-butylperoxy) hexane, or bis(2,4-dichlorbenzoyl) peroxide as well as bis(4-methylbenzoyl) peroxide are used as catalysts [488].

Curing by Hydrosilation. Cross-linking by hydrosilation generates exclusively ethylene bridges. In contrast to the peroxide-induced cure no decomposition products are formed. The curing reaction is catalyzed by 5 – 150 ppm platinum as platinum salts, or Pt(0)-vinylsiloxanes, Pt(0)-olefin, or Pt(0)-phosphine complexes.

$$\equiv Si-CH=CH_2 + H-Si\equiv \xrightarrow{Pt} \equiv Si-CH_2-CH_2-Si\equiv$$

SiH – siloxanes of the formula $RMe_2SiO-(Me_2SiO)_n-(MeHSiO)_m-SiMe_2R$ (where R is Me or H) serve as cross-linking agents. Cross-linking agents are applied in concentrations of 0.5 – 15 wt %. To extend the cure time at room temperature up to several days, inhibitors such as 2-methyl-3-butyn-2-ol, 1-ethynylcyclohexanol, vinylsiloxane, amines, or ketones (0.01 – 0.3 wt %) are used in hydrosilation -cured rubbers [489], [490].

Curing by Condensation. Polyfunctional silanes are the most common cross-linkers in rubber systems cured by condensation reactions. *Two-component*

systems contain alkoxysilanes, such as tetraethoxysilane. *One-component systems* also use other silanes such as methyltrimethoxy-, methyl- or ethyltriacetoxy-, methyltris(butanone oxime)-, or methylethoxybis(*N*-methylbenzamide)silane. These silanes react with the SiOH-terminated polymers to form prepolymers which undergo further cross-linking in the presence of moisture. The condensation reaction produces alcohols, oximes, amides, or acids as byproducts. *One-component RTV* compounding involves the formation to a prepolymer endstopped by cross-linker [482], [486]. The condensation reactions are catalyzed by tin dicarboxylates or titanium chelates added at 0.2 – 1 %. The rate of cross-linking of the individual systems can be adjusted within wide ranges.

Rubbers that are cross-linked with *peroxides* have a cure rate determined by the decomposition rate of the peroxide. In *hydrosilation* the cure rate is controlled by the quantity of the inhibitors (alkynols), SiH-cross-linker and the platinum catalyst.

Condensation RTV systems can be modified in viscosity and elongation by the choice of cross-linking agent and the length and functionality of polymer chain. The cure rate of one-component *sealant rubbers* is mainly governed by the diffusion rate of moisture from the surface through the bulk of the material. The vulcanization properties of different Q rubbers are listed in Table 34.

Table 34. Vulcanization properties of Q rubbers

	Pot life, min	Vulcanization temperature, °C	Vulcanization time, min
HV	> 43 200	120 – 200	1 – 2
LSR	> 4320	160 – 200	0.15 – 1
RTV-2 condensation	30 – 120	25 – 40	3 – 1440
RTV-2 hydrosilation	1 – 3	25 – 80	3 – 1440
RTV-1	5 – 25 *	25	1.4 **

* Skin formation.
** µm/min.

Molded articles made of solid or liquid silicone rubbers are mostly subjected to postcuring at 200 °C (2 – 6 h) to remove peroxide decomposition products as well as other volatile components (siloxanes) and to improve (lower) the compression set.

Vulcanization Process. Solid HV-rubbers are vulcanized using common feeding and molding equipment. Machines involved are extruders, transfer molds, and screw conveyors, but also injection molding machines. Press molds can be filled manually either with sheets, strips, or granulates. Curing can be pressureless (hot air) for extrudates or under pressure in heated molds [491], [492].

Extruded articles (profiles, tubing) are cured using bisaroylperoxides [e.g., bis(2,4-dichlorbenzoyl) peroxide] or by hydrosilation. Alkylperoxides (e.g., dicumyl peroxide) which decompose at higher temperatures, are only suitable for press molding.

Liquid silicone rubbers (A and B components) are effectively conveyed by means of metering pumps in a 1 : 1 ratio (+ color paste) first through a static mixer and then into the injection molding machine. Vulcanization occurs in the heated mold. Cold-runner technology is established to reduce loss of material [493].

Casting and molding compounds are stirred with the cross-linking agent in the ratio 1 : 8 to 1 : 10 and are cross-linked at room temperature or slightly above. Dental molding compounds are dosed from double cartridges (dispensing gun) via a static mixer, or the individual components are kneaded together.

RTV-1 rubbers can be extruded out of cartridges with dispensing guns or from drums with appropriate equipment. RTV-1 rubbers cure on contact with atmospheric moisture.

Recycling and Environmental Aspects. Recycling of silicone implies the separate collection of used silicone articles. Silicone elastomers can be reused in limited quantities as rubber powder or reclaims (partially depolymerized by steam hydrolysis). Another way is depolymerization of the Q elastomers to give distillable cyclosiloxanes. This can be carried out by using strong acids (concentrated sulfuric acid), bases (amines), or high temperature. The cyclosiloxanes can be introduced again as raw materials for the production of selected Q polymers (⟶ Silicones) [494]. Toxicological and environmental aspects of siloxanes are described elsewhere (⟶ Silicon Compounds, Organic).

5.1.3. *Producers and Markets*

The producers of VMQ rubber and compounds are listed below in alphabetical order together with the corresponding trade names:

Bayer (Leverkusen, Germany)	Silopren
Chemkomb. Dankow (Lipezk, CIS)	
Chimprom (Cheboksary, Gorki CIS)	
Dow Corning (Midland, MI, United States)	Silastic
General Electric (Waterford, NY, United States)	Blensil, GE-Sil
Sivento (Nünchritz, Germany)	Sivento
Inkos (Retkino, CIS)	
Kasan (Kasan, Tataric Republic CIS)	
Rhône Poulenc (St. Fons, France)	Rhodorsil
Saporoshki Sawod (Saparoshje, Ukraine CIS)	Kremni Polimer
Shinetsu (Isoke, Japan)	KE, Sylon
Toshiba (Ota, Japan)	TSE
Toray (Chiba, Japan)	SH
OSi Specialties Witco (Sisterville, WV, United States)	Sag, Silwet
Usolje (Irkutsk, CIS)	
Wacker (Burghausen, Germany)	Elastosil

The market shares of different VMQ rubbers in the Western world are given in Table 35. A breakdown by application of hot vulcanization rubber according to end-users (% of total) is given in the following:

Electrical industry	30
Automotive	20

Medicine – foods technology	20
Photocopier rolls	5
Construction	10
Others	15

Table 35. VMQ rubbers and their markets in 1996 (t)

	Western Europe	Japan	United States	Rest of world
HV	20 000	18 000	34 000	33 000
LSR	5 400	800	3 400	2 400
RTV 2K	7 000	5 100	7 500	3 400
RTV 1K	100 000	20 000	70 000	50 000

5.2. Polyurethane Rubber (see also ⟶ Polyurethanes)

Properties, Grades, and Applications. Millable polyurethane rubbers that can be processed by the methods of the rubber industry are a special group of polyurethane elastomers. Of greater market importance is the largest group, the thermoplastically processible polyurethane elastomers (see Chap. Thermoplastic Elastomers; see ⟶ Thermoplastic Elastomers), and the polyurethane elastomers produced by casting (⟶ Polyurethanes – Cast Elastomers) [495-498].

Vulcanizates from polyurethane rubber are characterized by their high toughness, high abrasion resistance, good oil resistance, and excellent ozone and UV resistance. Depending on the curing method, they have a high elasticity even at great hardness (diisocyanate cross-linking) or very low gas permeability (peroxide cross-linking) [499], [500]. Their hardness ranges from ca. 50 Shore A to 70 Shore D. The products are available on the market as blocks or pellets. Their Mooney viscosity varies between 20 and 60 (ML 1+4, 100 °C).

The differences in properties of the vulcanizates arise from the choice of *raw materials*. Polyurethane rubber is produced according to the principle of diisocyanate polyaddition, discovered in 1937 by OTTO BAYER. Linear macromolecules can be built up by addition of diisocyanates to dihydroxy compounds [501]. The *dihydroxy monomers* are long-chain compounds, which can be polyester- or polyetherpolyols. Thus polyesterurethanes and polyetherurethanes can be made. To a great extent, the polyols determine the elasticity, low-temperature flexibility, resistance to hydrolysis, and swelling behavior of the rubber. The cross-linking system also influences the properties considerably. Polyurethanes can be cross-linked with diisocyanate, peroxide, or sulfur [496], [499], [502-504].

Diisocyanate cross-linking, which requires the use of a suitable polymer with free hydroxyl groups, gives vulcanizates with a high elasticity, even at hardness levels up to ca. 70 Shore D [499], [502], [504]. *Peroxide cross-linking* generally confers excellent compression set and leads to vulcanizates with improved aging properties [503], [504]. Because of the accelerator systems necessary for *sulfur cross-linking*, which have unfavorable effects on the polyurethane backbone, sulfur cross-linking gives vulcanizates with only moderate aging resistance [497], [501].

Compounding takes place according to the normal guidelines of the rubber industry. Reinforcing carbon blacks are the preferred fillers [502]. The use of a hydrolysis inhibitor is generally recommended for polyesterurethanes [505-508].

The maximum service temperature for peroxidically cross-linked vulcanizates is ca. 125 °C. The material may be used up to ca. 150 °C for a short period [499], [502], [509]. Because of their chemical nature, vulcanizates from polyurethane rubber are attacked under the long-term action of hydrolyzing media (e.g., acid or base, hot water, and steam). This applies particularly to polyester-based products [500], [505], [506], [509]. However, compared to the polyether-based products, which are more resistant to hydrolysis, polyester-based vulcanizates exhibit greater resistance to organic solvents and oils.

The *glass transition temperatures* are in the range – 10 to – 40 °C; the *brittleness temperatures* are significantly lower for certain elastomers (less than – 70 °C) [500].

Uses. Typical uses of polyurethane rubber are wipers, driving belts, ejection springs, diaphragms in suspension systems, bushings, seals, rubber – metal parts, couplings, sleeves, matrices for the ceramics industry, membranes, engine mountings, wheel coverings, grinding discs, and roller coverings [510].

Production Processes and Basic Chemistry. Unlike polymerization and polycondensation, the basis of polyurethane chemistry is polyaddition. Here urethane bonds are formed by bonding diisocyanates to diols. To obtain products that can be milled and are stable to storage, a molar ratio of the diisocyanate component to the diol component of < 1 is used [495-497].

Polyesters or polyethers with a molar mass of ca. 1000 to 2000 g/mol are used as the diols; examples are polyadipates or poly-C_4-ethers. The preferred diisocyanates are toluene diisocyanate (TDI) and 4,4′-methylenediphenyl diisocyanate (MDI). Special grades with good lightfastness contain *m*-tetramethylxylene diisocyanate (TMXDI) [511] or dicyclohexylmethane 4,4′diisocyanate (HMDI, also H12MDI, SMDI, DCHM) [512], [513]. Where a special function is required that enables cross-linking with sulfur, OH-functional compounds with a double bond are used, such as glycerol monoallyl ether or trimethylolpropane monoallyl ether [497], [502].

Producers and Markets. In 2002 the estimated world market (not including the CIS and the Peoples' Republic of China) was ca. 1500 t/a. Suppliers of millable polyurethanes are:

Bayer/ Rheinchemie	Urepan
SBU	Sumipan
TSE Industries	Millathane
Uniroyal	Adiprene, Vibrathane

5.3. Thiokol Rubber

5.3.1. *Properties, Grades, and Applications*

5.3.1.1. *History*

Credited as being the oldest synthetic rubber still in production, thiokol rubbers [63148-67-4] (international abbreviation T) have changed considerably in character since their discovery in the 1920s [514-518]. This family of products was the result of attempts to utilize refinery fractions to produce glycols by chlorinating the olefins contained in these fractions and reacting the products obtained with alkali. The lack of success led PATRICK and MNOOKIN [516], [517] to react the chlorinated products with sodium di- and polysulfides, a polycondensation reaction that is still the basic approach to the production of thiokols.

The initial *type A polymer* was prepared from 1,2-dichloroethane [107-06-2] and sodium polysulfide, to yield a polymer with an average of four sulfur atoms between the alkyl groups in the polymer chain. After initial development by Thiokol Corporation, United States, in 1929, these polymers were prepared in

several countries. Type A polymers all suffer from two disadvantages: the physical properties of their vulcanizates are poor compared to those of natural rubber, and an unpleasant odor is given off during their difficult processing even after they are cured.

In 1934, *type B* tetrasulfide polymers were prepared by using a dichlorinated ether monomer. Later developments included homopolymer disulfide ethers *type D* and copolymers of disulfide ethers and ethylene dichloride (*types EB* and *F*). Today only FA and ST rubbers, a range of liquid rubbers, and the water dispersions MX and WD-6 are on the market.

5.3.1.2. *Grades*

The monomer compositions of the polysulfide rubbers now available are given in Table 36. The use of bis(2-chloroethoxy)methane [111-91-1] — often referred to as "formal" — improved the processibility of polysulfide elastomers, reduced their odor, and improved their low-temperature properties. At the same time the fuel and oil resistance of the formal-based products, although by many standards excellent, is not as good as that of the original polysulfide elastomers.

Table 36. Monomer composition of presently available polysulfide rubbers *

Type	Sulfur content, %	$C_2H_4Cl_2$	Formal	1,2,3-Trichloropropane
FA	47	x	x	
ST	37		x	x
LP	varying		x	x

* x = incorporated in the molecule.

The *FA materials* have terminal hydroxyl groups, although due to their very high molar mass (see Table 37) this generally does not influence curing or properties. The typical properties of FA and ST rubbers are given in Table 37.

Table 37. Typical properties of FA and ST raw rubbers

	FA	ST
Physical form	millable rubber	
Color	light brown	
d (at 23 °C)	1.34	1.27
Mooney viscosity ML1+4		
100 °C		30 – 40
123 °C	60 – 110	
Molar mass, g/mol	200 000	80 000
Odor	slight polysulfide	
Bin stability	limited	

Both *ST and LP grades* have terminal SH groups. The grades of LP rubber available are given in Table 38 together with their typical molar masses and viscosities LP-33 has a content of SH groups of 5.24 % [552].

Table 38. Properties of different grades of LP rubber

Polymer	mol % cross-link	Molar mass, kg/mol	Viscosity, Pa · s
LP 2	2.0	1000	130
LP-12	0.1	4000	4 000
LP-31	0.5	7000	11 000

Polysulfide water dispersions typically contain 50 % (MX grades) and 75 % (WD grades) polymer, stabilized against coagulation by $Mg(OH)_2$ and an added wetting agent. The dispersed rubber has an average particle size of ca. 10 µm.

5.3.1.3. *Mechanical Properties, Cross-Linking, and Applications*

Mechanical Properties. FA and ST polysulfide rubber compounds are flexible down to ca. – 45 °C (FA) or – 51 °C (ST), even over long periods of time in the presence of oils or solvents. Therefore, addition of a plasticizer is not required. The glass transition temperature of poly(ethyl formal disulfide) is ca. – 59 °C [519]. The even better resistance of FA rubbers against swelling in hydrocarbon oils can be attributed to their greater sulfur content. In ASTM no. 3 fluid a typical ST formulation exhibits a volume swell of about 18 % after 7 d [520]. Both types have modest heat resistance up to ca. 70 °C, and the thiol-terminated (i.e., ST) types can be used up to 125 °C where compression set is not critical [520]. A major difference between FA and ST rubbers is their resistance to compression set; FA types yield compounds with poorer compression set. Typical formulations and properties of FA and ST rubbers are shown in Tables 39 and 40.

Table 39. Typical formulations of FA and ST

Composition	FA	ST
Rubber, phr	100	100
N774 Black, phr	60	60
Stearic acid, phr	0.5	1.0
Zinc oxide, phr	10	5.0
Calcium oxide, phr		1.0
MBTS *, phr	0.3	
DPG **, phr	0.1	

* MBTS = dibenzothiazyl disulfide.
** DPG = *N,N′*-diphenylguanidine.

Table 40. Typical properties of FA * and ST ** compounds

Property	FA	ST
Tensile strength, MPa	8.3	9.0
Elongation at break, %	380	390
Hardness Shore A (23 °C)	70	67
Compression set, % (22 h, 70 °C)	20	15

* Cure 40 min at 150 °C.
** Cure 30 min at 155 °C.

Improvements in processing and increased ability to fill FA grades can be achieved by masticating them with, for example, benzothiazyl disulfide or thiuram disulfide in the presence of catalysts such as diphenylguanidine. As an example, addition of ca. 0.3 parts thiazole and 0.1 parts guanidine is quite efficient. An alternative is to blend the thiokol rubber with up to ca. 20 wt % of other polymers such as CR or NBR. The second polymer is chosen according to the solvent resistance required; thus, if resistance to ketones is required, CR should be employed, whereas a better resistance to aromatic or halogenated solvents is achieved with NBR blends [530].

Cross-Linking and Compounding [532, 533, 538]. *Hydroxyl-terminated* types are, after mastication, generally cross-linked by using zinc oxide [521]. For *thiol-terminated* polymers, many cross-linking reactions are possible, but generally zinc peroxide is used for the solid rubber, and lead or magnesium oxide or calcium peroxide is used for the liquid rubber. Although organic peroxides can be used, those that yield acidic byproducts should be avoided, since these can lead to polymer degradation.

Carbon black and *mineral fillers* are added to thiokol compounds and have similar effects as in other rubber systems. However, due to the intractability of the batch, only up to 100 parts of filler are added to the liquid systems. Acidic fillers should, however, be avoided because these too can lead to polymer degradation.

Plasticizers are rarely used because the low-temperature properties of these rubbers are adequate without them. However, plasticizers such as butyl benzyl phthalate are added to liquid thiokols if very soft compounds (down to 20 Shore A) are required. Alternatively the cure reaction can be limited by adding monofunctional mercaptans, such as cetostearyl mercaptan (e.g., 1.5 parts) [522].

Applications. Thiokol rubbers find use where their service temperature range of ca. – 45 to 100 °C (intermittently up to 150 °C) and their excellent resistance against oil and chemicals give them advantages over other elastomers [529]. Thiokols are widely used as sealants for heavy construction, building, or automotive application [528, 547-550, 551-543, 544]. The material is especially used as a sealant material for insulating glass to prevent condensation [534], [531, 535-537, 546]. Other applications are as modifiers for epoxy resins, as adhesives, or printing rolls and blankets [527, 553-555]. The liquid systems find particular use in fabric coatings and as coating materials generally [540, 545, 556].

Although the water dispersions of thiokol rubbers do not, alone, form continuous films on drying, the addition of a “softener” (such as ammonium hydrosulfide, tetramethylthiuram disulfide, benzothiazyl disulfide, or piperidinium pentamethylene dithiocarbamate) leads to chain scission and a transient softening, which allows a continuous film to form. These water dispersions are used extensively as concrete liners and, more generally, for films and fabric coatings having the flexibility and solvent resistance typical of polysulfide rubbers.

5.3.2. *Basic Chemistry and Production Processes [519], [523]*

Thiokols are polycondensates from bis(2-chloroethoxy)methane and sodium polysulfide. The general reaction is:

$$n\,\mathrm{ClRCl} + n\mathrm{Na_2S_x} \rightarrow \left(\mathrm{R-S_x}\right)_n + 2n\,\mathrm{NaCl}$$

where R = $^{\bullet}CH_2CH_2OCH_2OCH_2CH_2{}^{\bullet}$

so that, after isolation, the polymer has the general formula:

$$\mathrm{HO(CH_2CH_2S_xCH_2CH_2OCH_2OCH_2CH_2S_x)_nCH_2CH_2OH}$$

Depending on the thiokol type, either 1,2-dichloroethane or 1,2,3-trichloropropane [96-18-4] is employed as comonomer. The reaction is carried out in an aqueous dispersion in the presence of magnesium hydroxide. Formally, the latter acts as a nucleating agent for the polymer when it is precipitated by adjustment of the pH of the reaction mixture to ca. 5 – 6. The FA type polymer is isolated without further reaction steps, washed to remove inorganic salts and low molar mass material, and finally dried. The last stage involves exposure to air, which completes chain termination. The resulting hydroxyl-terminated polymer has a molar mass of ca. 200 000 g/mol and a very narrow molar mass distribution.

Before isolation of *thiol-terminated polymers* which are polymerized using 1,2,3-trichloropropane as a branching/cross-linking agent, sodium sulfite and sodium sulfhydrate (an aqueous solution of sodium hydrosulfite) are added to the reaction mixture. The former ensures that the number of sulfur atoms between the $CH_2CH_2OCH_2OCH_2CH_2$ groups is reduced to two. The sulfhydrate reacts with disulfide groups to cleave the polymer chain and introduce the thiol end groups. Polymer isolation is similar to that for hydroxyl-terminated polymers.

5.3.3. *Producers and Production Capacities*

The world production capacity for polysulfide rubbers was presumably some 40 000 t/a in 1999 [524]. The major producer of thiokol rubbers is Morton Thiokol, United States, whose capacity is estimated to lie between 9000 and 30 000 t/a [558], [524]. Toray Thiokol (estimated 2000 t in 2000) [525] in Japan, a joint venture between Morton and Toray, and Chemiewerk Greiz-Dolau in Thüringen, Germany — which was taken over by Harrisons & Crosfield (Manchester) in 1992 [539] — also make similar products. The latter has a probable capacity of some 5000 t/a [526].

[Top of Page]

6. Synthesis by Polymer Modification

Michael Happ, John Duffy, G. J. Wilson, Stephen D. Pask, Hartmuth Buding, Andreas Ostrowicki

6.1. Halobutyl Rubber

6.1.1. *Introduction*

The halobutyl rubbers, bromobutyl (BIIR) and chlorobutyl (CIIR) rubbers were first investigated by R. F. MORRISSEY of B. F. Goodrich in the mid-1950s [559-562]. MORRISSEY showed that halogenating the isoprene groups in IIR produced a rubber which could co-cure (vulcanize) with elastomers such as NR, BR, and SBR while preserving the essential properties of IIR. Goodrich marketed a brominated butyl rubber (Hycar 2202) produced by mill-mixing IIR with *N-bromosuccinimide* from 1954 – 1969. Independently, Michelin developed a similar brominated polymer [563] that was used internally in tubeless

radial tires.

Exxon developed the commercial chlorination of IIR with *elemental chlorine*. It showed that the chlorination was a substitution reaction and that HCl was released [564-566]. Polymer degradation, which occurred in chlorination, was overcome by limiting the amount of chlorine used to less than 1 mol per mole of isoprene in the IIR. In the early 1960s, Exxon introduced several grades of CIIR that differed mainly in molar mass.

Bromination of IIR with *elemental bromine* was developed and a commercial product introduced by Polysar in the late 1960s and early 1970s. Both types of bromination are substitution reactions, analogous to chlorination.

Halobutyl rubbers are more easily vulcanized than IIRs because the allylic halide groups make them more reactive.

6.1.2. *Properties, Grades, and Applications*

Properties. BIIR contains approximately 1.8-2.1 wt % bromine, and CIIR contains 1.2 wt % chlorine. Other than the density of BIIR, which is 930 kg/m^3, the physical properties of BIIR and CIIR are similar to those of IIR (see Section Properties, Grades, and Applications).

At temperatures above 150 °C, BIIR under-goes dehydrobromination, cross-linking, isomerization, and chain scission reactions, the latter two reactions being accelerated in the presence of acids. Dehydrobromination of the endo and exo structures (see Section Basic Chemistry) forms conjugated diene units each of which can appear as two isomeric configurations. BIIR is stabilized by calcium stearate and epoxidized soybean oil (ESBO), which react with any released HBr. CIIR is more stable than BIIR, and therefore lower amount of calcium stearate is added as stabilizer. Calcium stearate is also a cure retarder.

For example, when a solution of CIIR is heated with calcium oxide in the presence of a small amount of a zinc carboxylate (usually stearate), the polymer is dehydrochlorinated to give a butyl rubber containing conjugated diene groups [595], which can be cured with peroxides and reacted with styrene and maleic anhydride resins to give interpenetrating networks [585]. BIIR, under the same conditions, rearranges and cross-links to the bromomethyl structure [572].

Both BIIR and CIIR cure when heated with zinc oxide in the presence of stearic acid [559]. For BIIR, a higher fraction of the limited olefinic functionality is converted to cross-links than for CIIR, and the effect is more pronounced for the rearranged form of BIIR [572], [586]. Conventional sulfur curatives achieve a rapid, effective cure, even in the presence of highly unsaturated polymers such as NR, BR, and SBR, which prevent IIR from curing [559]. The faculty of the halobutyl rubbers for curing in the presence of other elastomers has led to their widespread use in tire inner liners.

The halobutyl rubbers cross-link with resins, various amines [586], dienophiles [587], and other curatives [588]. The curing reaction is usually faster than that of IIR, and the vulcanizates are usually formulated to be more resistant to heat. Because of this cure versatility, halobutyl rubbers have displaced IIR in applications such as pharmaceutical closures, tank linings, and mechanical goods.

Applications. During the 1960s the quantity of CIIR used for previous IIR applications increased as CIIR was adopted for the inner liners of tubeless tires. In this application, CIIR is usually blended with NR in ratios of ca. 70 : 30. Since BIIR is more reactive than CIIR it co-cures more effectively with highly unsaturated elastomers. BIIR is the elastomer of choice for inner liners containing no NR, such as for truck tires.

Halobutyl rubbers continue to replace IIR with tubeless tires replacing tubed tires. An estimate of the amounts of IIR and halobutyl rubber used in various applications is given in Section Properties, Grades, and Applications.

Grades. BIIR and CIIR are produced by Bayer and Exxon. Grades differ in Mooney viscosity and, hence, processibility of the polymers. Grades from each producer are listed in Table 41 with typical values of the parameters used for specification purposes. As with IIR, specification ranges differ slightly according to source.

Table 41. Grades and typical properties of halobutyl rubber

Producer and properties	Bromobutyl grades			Chlorobutyl grades	
Bayer	2030	2040	BBX2	1240	1255
Mooney viscosity [a], ML/1+8/125°C	32	39	46	38	50
Halogen content [b], wt %	1.8	1.8	1.8	1.25	1.25
Volatile content [c], wt %	0.7 max.	0.7 max.	0.7 max.	0.5 max.	0.5 max.
Ash content [d], wt %	0.7 max.	0.7 max.	0.7 max.	0.5 max.	0.5 max.
ESBO content, wt %	1.3	1.3	1.3		
Antioxidant content, wt %	0.03 min.	0.03 min.	0.03 min.	0.03 min.	0.03 min.
Cure characteristics [e]					
Maximum torque, dN · m	12.5	13.4	14.3	15.2	16.0
Minimum torque, dN · m	4.3	4.9	5.5	4.5	5.4
Time to 50 % cure, min				3.0	3.0
Time to 90 % cure, min	9.0	9.0	9.0		
ExxonMobil	2222	2235	2255	1066	1068
Mooney viscosity [f], ML/1+8/125°C	34	39	46	38	50
Halogen content [g], wt %	2.0	2.0	2.0	1.26	1.26
Water content [g], wt %	0.3 max.	0.3 max.	0.3 max.	0.3 max.	0.3 max.
ESBO content [g], wt %	1.3	1.3	1.3		
Antioxidant content [g], wt %	0.02 min.	0.02 min.	0.02 min.	0.02 min.	0.02 min.
Cure characteristics [h]					
Maximum torque, dN · m	34.0	37.0	41.0	41	45
Minimum torque, dN · m	12.0	14.0	16.0	14	17
Time to 50 % cure, min	8	7.5	7.5	5	5.0
Time to 90 % cure, min	10.5	10.0	10.0	13	13.0

[a] According to ASTM D 1646 = ISO 289.
[b] Determined by X-ray fluorescence.

[c] According to ASTM D 5668 = ISO 248.
[d] According to ASTM D 5667 = ISO 247.
[e] Test procedure specified in ASTM D-5289 = ISO 6502, Rotorless cure meter, 160°C, 1.7 Hz, 1° arc, 30 min run; determined with the following compound: halobutyl rubber 100 phr, carbon black IRB#7 black 40 phr, NBS zinc oxide 5 phr, NBS stearic acid 1 phr.
[f] According to modified ASTM D 16467.
[g] ExxonMobil test method
[h] According to modified ASTM D 2084 using ODR 2000 rheometer.

6.1.3. *Basic Chemistry and Production Process*

6.1.3.1. *Basic Chemistry*

According to the generally accepted [564, 569, 573-575] mechanism halogenation of butyl rubber proceeds through the formation of a halonium ion. It was also suggested that in the initial stage, a molecular complex forms between the halogen and the C ═C double bond [574], [576]. The presence of the sterically bulky methyl group in the β position prevents the attack of the halide ion at the tertiary carbon atom and therefore the formation of an addition product [569]. As a result, the energetically less favorable process of elimination of hydrogen halide, leading to substitution products, becomes the predominant pathway [569]. Recently it was proposed that the byproduct HCl accelerates chlorination of natural rubber [574] as well as butyl rubber [575] and excess of bromine acts as a catalyst in the bromination of bromobutyl rubber in a nonpolar reaction medium [577].

The main product is a secondary allylic halide structure where the double bond is external relative to the backbone of the polymer. This structure is referred to as “exomethylene” or “exo”. The minor product is another secondary allylic halide in which case the double bond is located in the backbone. This structure is referred to as “endo”. The byproduct of the main reaction, hydrogen halide, can also react with the double bond. The secondary allylic “exo” structure can rearrange to a thermodynamically more stable primary allylic form in the presence of acids, such as the byproduct HX [573], [578]. Dehydrobromination of the already formed allylic structures can also occur leading to the formation of conjugated diene structures. These primary and secondary structures are shown in Figure 22. In [570-572], the structures originally suggested in [573] are designated as shown in Figure 22.

Figure 22. Principal structures present in halobutyl rubber

Commercial BIIR base polymer contains about 1.7 – 1.9 mol % of isoprene monomer, of which 0.5 – 0.6 mol % is not brominated. Of the approximately 1.0 mol % brominated, about 0.9 mol % is converted to the exomethylene form and the remainder to the endo and rearranged structures, with a trace of the endo-conjugated form. CIIR base polymer typically contains 2.1 – 2.3 mol % of isoprene, of which 1.7 – 1.8 mol % is chlorinated. Approximately 1.3 – 1.4 mol % is converted to the exo form and 0.4 – 0.5 mol % to *cis* and *trans* endo isomers. The rearranged form does not occur in CIIR.

6.1.3.2. *Production*

Chemicals Used in Production. The principal chemicals employed in the production of halobutyl rubber from IIR are the elemental halogens, hexane as solvent, sodium hydroxide as neutralizing agent along with calcium stearate and ESBO, which are used to control stability and cure behavior [567].

Solvent Process. Halobutyl rubber is produced by contacting a solution of regular butyl base polymer in an alkane (the solution is known as cement) with halogen in an agitated vessel. Unreacted halogen and hydrogen halide formed as byproduct are extracted from the cement, and additives are incorporated. The solution is steam-stripped to remove the solvent, giving a solid product. The solid material is finished by methods similar to those described in Section Industrial Production for regular butyl rubber. A process flow sheet is shown in Figure 23.

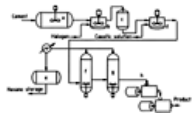

Figure 23. Flow diagram for production of halobutyl rubber

a) Cement storage; b) Halogenator; c) Aging drum; d) Neutralizer; e) Hexane receiver; f) Coagulator; g) Stripper; h) Finishing extruders

Dissolution of Rubber. The halogenation operation begins with a solution of rubber in an alkane, usually hexane. The rubber is dissolved either directly by the solvent replacement process (see Section Producers and Production Capacities), or indirectly by dissolving solid rubber produced by the slurry process (see Section Producers and Production Capacities) [581].

Solutions of butyl rubber are prepared by dissolving rubber crumb directly from the vacuum strippers (see Fig. 16), or from chopped or ground bales of rubber from a conventional finishing process. The time for complete dissolution depends on temperature, particle size, agitation, and molar mass; 1 – 4 h are normally required.

Chlorination and Bromination. Chlorination and bromination are conducted in one or more intensely agitated vessels.

The *chlorination reaction* is carried out by contacting the rubber cement with a solution of chlorine or with chlorine vapor at 25 – 65 °C. Chlorine reacts violently with many hydrocarbons, including butyl rubber solutions, so its concentration must be kept low. As discussed above, the reaction consists of replacement and rearrangement, producing hydrogen chloride as byproduct. The quantity of chlorine added is essentially stoichiometric.

Addition of bromine to the polymer proceeds more slowly than that of chlorine, so the cement can be contacted directly with liquid bromine. Reaction temperatures are comparable to those for chlorination. To provide more time for reaction at low temperature, the reaction mixture passes through a residence time vessel after the mixer. More than a stoichiometric amount of bromine must be used to obtain a sufficient reaction rate without product breakdown. As in CIIR manufacture, hydrogen bromide is a byproduct.

Neutralization. The hydrogen bromide and unreacted bromine are neutralized promptly to avoid undesirable side reactions that reduce product stability. Hydrogen chloride is also neutralized while chlorine reacts to completion. Neutralization is conducted by dispersing the halogenated cement in a sodium hydroxide solution with high-intensity mixers.

Product Recovery. The product is recovered as a 5 – 12 % slurry in water by contacting it with steam and water in a process similar to that described in Section Ethylene – Propene Elastomers (EPM, EPDM) for regular butyl rubbers. The stabilizers calcium stearate and ESBO, as well as an antioxidant, are added to the rubber immediately before recovery.

The rubber is finished with mechanical drying equipment in a process similar to that used for regular butyl rubber (see Section Ethylene – Propene Elastomers (EPM, EPDM)). Because of the greater reactivity of the halogenated product, less severe conditions are employed.

Extruder Halogenation. Butyl rubber has been halogenated in the solid phase in an extruder-based reactor in which the bulk phase is in direct contact with chlorine or bromine [589]. Because the operation is substantially free of water, the hydrohalic acids formed as byproducts do not dissociate and are removed by an extruder devolatilization, thereby avoiding the need for dispersed-phase neutralization.

6.1.4. *Chemical Analysis*

The *IR spectra* of BIIR and CIIR are very similar to those of IIR, except for peaks at 910 and 1637 cm^{-1} due to the $=CH_2$ group [590]. The absorption at 910 cm^{-1} is a shoulder, and the peak at 1637 cm^{-1} is small and broad; however, they can be used to determine the amount of exomethylene structure present in the rubber.

The *proton NMR spectra* of BIIR and CIIR are very similar [591]. Figure 26 shows the 500 MHz ^{1}HNMR spectra of a typical bromobutyl sample with the assignment of the resonances present in the olefinic region.

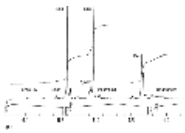

Figure 26. Schematic for HNBR production

a) Solution production; b) Hydrogenation (homogeneous or heterogeneous); c) Catalyst separation; d) Isolation of HNBR

6.1.5. *Storage and Transportation*

The halobutyl rubbers are packaged, shipped, and stored in the same manner as IIR (see Section Producers and Production Capacities). Shelf life is about 60 months if the polymer is not subjected to high temperature during storage.

6.1.6. *Legal Aspects*

BIIR and CIIR comply with U.S. FDA 21 CFR 177.1210 (sealing gaskets for food containers) and are allowed in articles intended for repeated use (CFR 177.2600). Local regulations are applicable.

6.1.7. *Producers and Production Capacities*

BIIR and CIIR are produced by Exxon and Bayer in plants that either are used only for production of halobutyl rubber or can produce both halobutyl and IIR. Estimated production capacities of IIR and halobutyl rubber are given in the section on IIR (Table 25)

Table 25. Uses of IIR and halobutyl rubber by region in 10^3 t/a

	IIR			Halobutyl rubber		
Use	**America**	**Western Europe**	**Asia**	**America**	**Western Europe**	**Asia**
Inner liners				120	50	40
Inner tubes	45	45	85			
Curing bladders	5	5	5			
Nontire applications	20	10	15	10	10	10
Total	70	60	105	130	60	50
Type total		**235**			**240**	

.

6.1.8. *Toxicology and Occupational Health*

Halobutyl rubbers are regarded as nonhazardous and can be processed like other elastomers. Traces of halogenated hydrocarbons are present in both BIIR and CIIR. The levels are too low to be hazardous [593], but must be considered where extraction may lead to unfavorable consequences.

The presence of halogenated polymers must be considered if used tires are incinerated.

6.1.9. *New Commercial Development*

Star-Branched Halobutyl Rubber. Chlorinated or brominated versions of star branched butyl rubber (see Section Industrial Production) are made in an analogous manner to CIIR and BIIR [594].

Brominated Isobutene/p-Methylstyrene Copolymer. Conventional halobutyls are based on the copolymer of isobutene and isoprene. In the early 1990s Exxon developed a new type of bromobutyl rubber, which is made by the free radical bromination of isobutene/*p*-methylstyrene copolymers and commercialized under the EXXPRO trade name. The isobutene/*p*-methylstyrene copolymer is made by conventional cationic polymerization using $AlCl_3$ or alkyl aluminum coinitiator at low temperatures in methyl chloride diluent [581], [579] as it is in the case of IIR. The *p*-methylstyrene content of these polymers is between 2.5 – 7 mol % the rest being isobutene [580]. By the free radical bromination, 0.5 – 1.2 mol % of the incorporated *p*-methylstyrene units are converted to benzylic bromide structure. The fully saturated backbone of this polymer results in an improved ozone resistance compared to conventional halobutyl elastomers.

6.2. Chloropolyethylene and Chlorosulfonyl Polyethylene

Because of its high chain mobility, polyethylene has a very low glass transition temperature. However, because of the regular chain structure it also crystallizes very readily, so that it is a thermoplastic below its melting point. Functional groups, which hinder crystallization, can be introduced by polymer modification reactions. This, however, decreases chain mobility because of steric hindrance, and the glass transition temperature is raised. Within certain limits of chlorine contents, polymers with rubbery properties are obtained. The chlorination (CM) and chlorosulfonation (CSM) of polyethylene are important industrially.

6.2.1. *Chloropolyethylene*

6.2.1.1. *Properties, Grades, and Applications*

The properties of CM are determined principally by the chlorine content [582], [583].

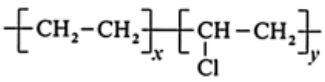

Polymers with chlorine content of < 20 wt % show a noticeable degree of crystallinity because of the presence of polyethylene sequences. At > 50 wt % chlorine the strong interactions of the C – Cl dipoles result in hard, brittle, but noncrystalline products. Polymers with between 25 and 40 wt % chlorine are suitable for rubber applications. At a chlorine content of ca. 35 wt % the crystallinity of the chain disappears completely. These polymers possess the best low-temperature flexibility (brittleness point ≤ – 40 °C) and the lowest softening point (≤ 20 °C). The effect of using different types of polyethylene is observable only at a chlorine content of < 30 wt % [584]. Thus, if the more crystalline and less branched high-density polyethylene (HDPE) is chlorinated to a chlorine content of 15 wt % it exhibits the same degree of crystallinity as unchlorinated low-density polyethylene (LDPE).

The CM grades are available commercially in powder form, and differ in chlorine content and viscosity. Their vulcanizates are characterized by very high

resistance to hot air, oil, and chemicals and by very high flame resistance. At optimal composition they are very resistant to discoloration and ozone cracking. In the range of 36 – 44 wt % chlorine, which is typical for most of the CM grades, oil resistance, tear strength, flame resistance, and gas impermeability improve with increasing chlorine content. However, plasticizer sorption, compression set, and low-temperature flexibility deteriorate [596]. A comparison with vulcanizates of related rubbers is shown in Table 42. The relationships given depend on the composition.

Table 42. Comparison of CM vulcanizates with related rubbers [*][596]

	CSM	CR	NBR	ECO
Price	– –	O	+	– –
Hot-air resistance	O/–	– –	– –	–
Hot-oil resistance	–	– –	– –	–
Oil and fuel swelling	O	O	+	+
Ozone and weather resistance	O	–	—	O
Resistance to discoloration	O	—	–	
Fire resistance	O	+	– –	–
Resistance to chemicals	O	–	– –	– –
Mechanical properties	O	O	O	–
Brittleness point	–	+	O/–	O
Cold flexibility	O	+	O/+	+
Compression set at > 125°C	– –	+	O	– –
Adhesion to metal – fabrics	O	+	+	O

[*] Comparable products are – – much less favorable; – somewhat less favorable; O about the same; + better.

CM can be compounded on conventional rubber-processing machines. Without addition of the common rubber chemicals, CM behaves more like a thermoplastic. It can be compounded particularly rapidly in a kneader by the upside-down method, in which the rubber powder is added last. A thermal softening effect during processing is typical for CM. Like every chlorine-containing polymer, CM must be stabilized against loss of HCl at high processing or application temperature [597]. Metal oxides such as magnesium or lead oxide, basic lead salts, epoxidized soybean oil, and epoxy resins are suitable stabilizers, frequently in combination with one another. Zinc oxide and also other zinc compounds must not be used as stabilizers because they drastically reduce the resistance to aging of the vulcanizate. Because of the absence of double bonds, lower degree of branching (tertiary chlorine), and random incorporation of chlorine atoms in the polymer chain, CM elastomers are thermally more stable than PVC or CR. However, CM cannot be vulcanized with sulfur. Organic peroxides are most frequently used for *vulcanization*, usually in combination with multifunctional vinyl or allyl compounds (e.g., triallyl cyanurate). Cross-linking can also be carried out by using energy-rich radiation or — for articles that do not require high thermal stability — via a thiadiazole – amine system. CM is used mainly in the cable and wire sectors for the production of oil- and flame-resistant sheathing. Other uses include technical rubber goods such as hose (particularly hydraulic and chemical hose), collars, dust caps, and seals in car and plant construction [596].

6.2.1.2. *Basic Chemistry and Production Processes*

The first chloropolyethylenes were produced by chlorination of LDPE in solution (e.g., in CCl_4). This resulted in a random distribution of chlorine atoms along the polymer chain. Currently, HDPE is used as starting material and is chlorinated mainly in aqueous suspension. Polyethylene must be in powder form to provide a large surface area for chlorination. The reaction is performed in several stages to achieve optimal chlorination and avoid agglomeration of the powder particles [598-601]. In the first stage the reaction is carried out below the melting point of polyethylene which leads mainly to chlorination at the amorphous phases. The product is then heated above the melting point of the crystalline phases, and chlorination is continued. Occasionally, chlorination is completed in a third stage at a lower temperature. The CM powder is washed free of HCl, water is removed by centrifugation, and the product is dried and packed in sacks.

6.2.2. ***Chlorosulfonyl Polyethylene***

6.2.2.1. *Properties, Grades, and Applications*

Chlorosulfonyl polyethylene (CSM) contains, besides chlorine atoms, small quantities of sulfonyl chloride groups distributed randomly along the polymer chain.

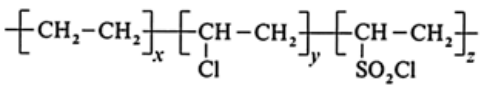

Commercial CSM grades contain between 25 and 43 wt % chlorine and 0.8 – 1.5 wt % sulfur, and have Mooney viscosities between 20 and 110 [602], [603]. In a product containing, for example, 25 wt % chlorine and 1.0 wt % sulfur, ca. every 80th carbon atom carries a sulfonyl chloride group, and every 7th or 8th carbon atom is chlorinated. In many properties, CSM is similar to CM because of the comparable chlorine content. However, the chlorine atoms of sulfonyl chloride groups are significantly more reactive than those bonded to carbon atoms and can undergo cross-linking reactions [603]. The vulcanization of CSM can take place

1. In the presence of moisture and a divalent metal oxide (e.g., MgO or PbO) with the formation of ionic cross-linking bridges. This reaction is slow and is particularly suitable for vulcanizates used in roofing and for linings for reservoirs and basins.
2. With MgO and sulfur-containing accelerators, such as mercaptobenzthiazole or thiuram disulfides. If the vulcanizates are to be used at elevated temperature (up to 150 °C), polyfunctional alcohols (e.g., pentaerythritol) are added. These applications include hydraulic tubes and cable covering.
3. With peroxides, by irradiation, or together with epoxy resins.

In addition CSM is used for fabric coatings, roller coatings, extruded good, seals, and molded articles.

A new development is the production of so-called *alkylated chlorosulfonated polyethylene* (ACSM) [604-606]. Alkylated chlorosulfonated polyethylene has short alkyl side groups on the polymer chain, which replace some of the chlorine atoms (chlorine content: 26 – 30 wt %) and act as internal plasticizers.

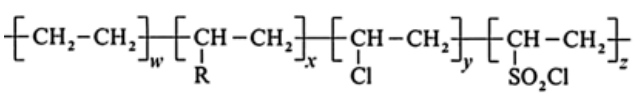

The products have glass transition temperatures about 5 °C lower than CSM, better heat resistance, and lower oil resistance. Unlike standard CSM, ACSM is also suitable for dynamic stressed parts, such as timing belts, axle boots, and poly-V-belts.

6.2.2.2. *Basic Chemistry and Production Processes*

Chlorosulfonyl polyethylene is produced commercially by simultaneous chlorination and chlorosulfonation of *LDPE* or *HDPE* in an inert solvent (e.g., CCl_4) with a mixture of gaseous chlorine and sulfur dioxide under UV irradiation [607]. The product is processed by application of a concentrated solution of the

product to two drying rollers placed close together, which are heated to 100 – 200 °C. The rubber layer is stripped off the rollers with a knife [608]. Spray drying or pouring the polymer solution into hot water above the boiling point of the solvent, with the addition of $MgCO_3$, for example, to avoid lump formation, is also described [609]. CSM can also be produced by chlorosulfonation of *CM powder* with gaseous chlorine and sulfur dioxide in a fluidized-bed reactor at 20 – 100 °C under UV [610]. *Linear low-density polyethylene* (copolymer of ethylene and short-chain α-olefins) is most likely used as the starting material for ACSM.

6.2.3. *Producers and Production Capacities*

In 1991 the worldwide capacity for CSM was 56 600 t, and for CM 51 400 t [611]. The largest proportion of CM is used not in the elastomer sector, but for modifying thermoplastics, mainly PVC. The producers, trade names, production sites, and capacities for CSM and CM are listed in Table 43.

Table 43. CSM and CM producers and production capacities

Producer	Type	Trade name	Production site	Production capacity, t/a
Denki KK	CSM	Denka-CSM	Omi/Japan	2 000
Du Pont	CSM	Hypalon	Beaumont/United States	34 500
Du Pont	CSM	Hypalon	Londonderry/Northern Ireland	18 100
Tosoh	CSM	Toso-CSM	Shinnanyo/Japan	2 000
Dow	CM	Tyrin CM	Plaquemine/United States	40 000 *
Hoechst	CM	Hostapren	Frankfurt/Germany	6 000 **
Osaka Soda	CM	Daisolac	Amagasaki/Japan	2 400
Showa Denko	CM	Elaslen		3 000
Du Pont	CM	CP-R		

* Estimated, no information from the producer.
** Production expansion to 12 000 t/a planned.

6.3. Hydrogenated Nitrile Rubber

6.3.1. *Properties, Grades, and Applications*

Introduction. The selective hydrogenation of the C – C double bonds in NBR leads to a specialty rubber, hydrogenated nitrile rubber (HNBR), which can be processed into vulcanizates with a balanced property profile [612-625]. The vulcanizates are characterized by

1. Excellent mechanical properties
2. Very good abrasion resistance
3. Very good resistance toward industrial oils
4. Very good low-temperature impact and flexural resistance
5. Very good ozone and hot-air resistance
6. Good resistance to alkali
7. Low steam and gas permeability

The reaction scheme of NBR hydrogenation and the chemical formula of HNBR [88254-10-8] are given below:

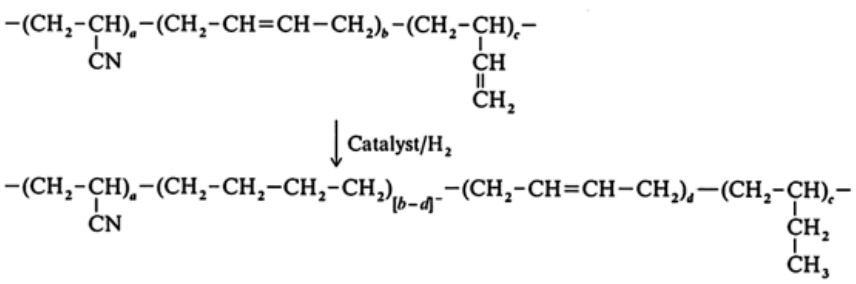

Grades. Completely and partially hydrogenated HNBR grades with acrylonitrile contents between 34 and 49 wt % and Mooney viscosities between 50 and 152 [ML 1+4 (100 °C)] are on the market.

Completely hydrogenated grades are defined here as those in which the degree of hydrogenation is 99 %. *Partially hydrogenated grades* have a degree of hydrogenation < 99 %; those with a degree of hydrogenation between ca. 90 and 98 % are of particular technological interest because they can be vulcanized using traditional sulfur systems.

Raw Rubber. Investigations of the *microstructure* of HNBR with different degrees of hydrogenation using pyrolysis – GC, IR and NMR spectroscopy have been described [626]. The degradation of HNBR under different conditions and various influences has also been investigated [627], [628].

The density of HNBR depends on the acrylonitrile content. For instance, HNBR containing ca. 43 wt % has a density of ca. 0.97 g/cm^3; whereas that containing ca. 34 wt % acrylonitrile, has a density of 0.95 g/cm^3.

Completely and partially hydrogenated NBR exhibits spontaneous and elongation crystallization [629], [630]. The variation of the glass transition temperature of NBR with its acrylonitrile content obeys the Gordon – Taylor relationship for random copolymers. In contrast, completely hydrogenated NBR and ethylene – acrylonitrile copolymers with acrylonitrile content < 55 wt % do not obey this relationship (see Fig. 24). The T_g (mechanical – dynamic measurement at 11 Hz) of products with an acrylonitrile content of 18.9 – 39.0 wt % is between –25 and –21.5 °C. At a content of 48.3 wt % acrylonitrile, T_g is –8 °C and thus markedly higher than expected from the Gordon – Taylor relationship [629]. A corresponding behavior is described for ethylene – vinyl acetate and ethylene – vinyl chloride copolymers of certain compositions, which has been attributed to crystallizing ethylene sequences.

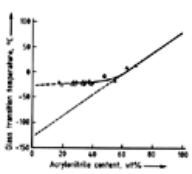

Figure 24. Dependence of glass transition temperature on acrylonitrile content according to [629]

○ = Hydrogenated products; △ = Ethylene – acrylonitrile copolymers

The T_g of HNBR also depends on the degree of hydrogenation [616], [630], [631].

Products that exhibit significantly better low-temperature properties than conventional HNBR (TR 10 test according to ASTM D-1329 and T 10 test according to Gehmann) have been developed. They can be obtained by hydrogenation of ester-group-containing terpolymers based on NBR [632], [633].

Properties of the Vulcanizate. HNBR can be processed easily by the usual methods in the rubber industry. *Completely hydrogenated* NBR is vulcanized by using peroxides or high-energy radiation; *partially hydrogenated* NBR can also be vulcanized with sulfur and sulfur donors. Table 44 shows schematically the properties of vulcanizates derived from completely hydrogenated NBR, as a function of acrylonitrile content. Table 45 indicates the effects of the degree of hydrogenation and the nature of the cross-linking system on some vulcanizate properties.

Table 44. Effect of acrylonitrile content on the properties of HNBR vulcanizates (direction of the arrow indicates an increase in value or an improvement in properties)

Vulcanizates properties	Acrylonitrile content Low	High
Oil resistance	→	
Fuel resistance	→	
Tensile modulus	→	
Shore hardness	→	
Rebound resilience	←	
Gas permeability	←	
Heat buildup	←	

Table 45. Effect of degree of hydrogenation and cross-linking system on the properties of HNBR vulcanizates (direction of the arrow indicates an increase in value or an improvement in properties)

Vulcanizate properties	Completely hydrogenated grades	Partially hydrogenated grades
	Peroxide cross-linking	Sulfur cross-linking
Hot-air resistance	←	
Ozone resistance	←	
Tensile modulus	←	
Elongation at break	→	
Resistance to tear propagation	→	

HNBR can be compounded without a plasticizer in such a way that its vulcanizates have a *brittleness temperature* of <–70 °C according to DIN 53 546. The T_g of completely hydrogenated NBR can be lowered within certain limits by special compounding with plasticizers, without significant loss of heat resistance.

If magnesium and zinc oxides are replaced by zinc peroxide in the compounding of partially hydrogenated NBR, a vulcanizate with improved properties is obtained after cross-linking with sulfur donors in the presence of a small amount of sulfur (efficient vulcanization). These properties are tensile modulus, compression set, ozone and hot-air resistance, and dynamic load-bearing capacity [634]. However, to prevent accidents, only zinc peroxide made into a paste with plasticizer should be used.

Partially hydrogenated NBR that has been cross-linked with peroxides has a better low-temperature compression set than completely hydrogenated grades.

The heat resistance of vulcanizates from completely hydrogenated NBR is sufficient up to 150 °C — and that of NBR vulcanizates up to 120 °C — if residual elongation at break of 50 % after 1000 h aging is used as the criterion [612].

In Table 4 of Section Properties some properties of HNBR vulcanizates are compared with those of other rubber vulcanizates. The balanced property profile of HNBR can be clearly observed.

Applications. Because of its particular performance, HNBR can be used in the *automotive industry* for toothed and driving belts, and for various types of hosing and seals; in the *petroleum industry* for drill pipe protectors, pump stators, packing seals, and blowout preventors; in the *electrical industry* for cable coverings; and in the *machine construction industry* for laminating rollers.

The use of HNBR as a *modifier for thermoplastics* [635-644], as a *blend component for other rubbers* [645-656], and as a basic component for the production of *dynamically cross-linked thermoplastic elastomers* [657] has been claimed in various patent applications. EVM and FPM, for example, have achieved some importance as blend components for HNBR. *HNBR – EVM rubber blends* are suitable for cable coverings in ship construction. They meet the Naval Engineering Standard 518. *HNBR – FPM rubber blends* give vulcanizates with high fuel resistance and good mechanical properties. An *HNBR – PVC blend* has a better fuel resistance than HNBR [658].

6.3.2. *Basic Chemistry and Production Processes*

In the hydrogenation of NBR to HNBR the C – C double bonds must be able to be completely (selectively) hydrogenated if desired, without the triple bonds of the nitrile groups being attacked. Reduction of the nitrile groups leads to an increase in molar mass due to the formation of secondary and tertiary amines and eventually to insoluble products.

Determination of the degree of hydrogenation of HNBR can be carried out by IR and ^{1}H-NMR spectroscopy, and titrimetrically in chloroform by using the iodine number method according to ASTM D-1959 [659-661]. The acrylonitrile content can be determined, for example, by the Kjeldahl method or by elemental analysis according to Dumas.

In the current industrial hydrogenation processes, NBR solutions are reacted with hydrogen batchwise in a stirred vessel in the presence of homogeneous or heterogeneous catalysts. A schematic diagram for both processes is shown in Figure 25. The polymer concentration of the solution to be hydrogenated is ca. 15 wt %.

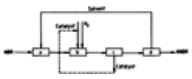

Figure 25. Schematic for HNBR production

a) Solution production; b) Hydrogenation (homogeneous or heterogeneous); c) Catalyst separation; d) Isolation of HNBR

Homogeneous Catalysis. In homogeneous hydrogenation, the catalyst, like the substrate to be hydrogenated, is dissolved in the solvent. Suitable catalysts are, for example, rhodium and ruthenium phosphine complexes [662-677]. With *rhodium catalysts* the process is carried out preferably in chlorinated aromatic hydrocarbons as solvent, in the presence of phosphine ligands if desired.

Depending on the type of *ruthenium catalyst* used, hydrogenation can be carried out in either the presence or the absence of free phosphines. Ketones, chlorobenzene, or mixtures of the two are used as solvent. The use of organometallic hydrogenation catalysts requires careful exclusion of oxygen.

Depending on the catalyst chosen and its concentration, the reaction temperature is usually between ca. 100 and 150 °C. The reaction pressure can vary from a few to ca. 190 bar.

The kinetics of selective hydrogenation of NBR with tris(triphenylphosphine)rhodium(I) chloride in chlorobenzene are known. The C – C double bonds of the 1,2-bonded butadiene units are hydrogenated most rapidly, followed by the 1,4-*cis*-bonded units, and the hydrogenation of 1,4-*trans*-bonded butadiene units is slowest [678], [679].

Heterogeneous Catalysis. In heterogeneous hydrogenation of NBR supported palladium catalysts in particular are used, which are dispersed in the dissolved substrate. The use of other supported noble-metal catalysts is also described. Carbon black, activated carbon, calcium carbonate, silicon dioxide, and titanium or zirconium dioxides can be used as catalyst support [680-685]. The reaction is usually performed in low molar mass ketone solvents. The preferred reaction temperature is ca. 50 °C and the reaction pressure ca. 50 bar. The hydrogenating activity of the catalyst can be improved by doping the palladium with other metals [680-685]. Particularly good hydrogenation results are obtained with silicon dioxide as the carrier material for the active metals if a specific surface area of 600 m^2/g is not exceeded and the average pore diameter is between 80 and 120 nm.

After hydrogenation the catalyst is usually removed from the reaction solution. Homogeneous catalysts can be removed by precipitation [686] or absorption [687], [688]; for heterogeneous catalysts, filtration or centrifugation are suitable [682]. The isolated catalysts are regenerated if necessary and then recycled into the hydrogenation process, or worked up.

HNBR is isolated by blowing steam through the solution, followed by drying or by direct evaporation of the solvent.

Besides the hydrogenation process described, other processes are known, for example,

1. Reaction of NBR with hydrogen in solution with palladium carboxylates [689]
2. Reaction of NBR with hydrogen in solution with a catalyst composed of a palladium and an organoaluminum compound [690]
3. Hydrogenation of olefinic double bonds in unsaturated polymers (including NBR) with hydrogen in solution with rhodium catalysts of the Wilkinson type, which are fixed on a polymer matrix [691]
4. Reaction of NBR latex with hydrogen in the presence of chlorobenzene by using tetrakis(triphenylphosphine)rhodium(I) hydride [665]
5. Hydrogenation of NBR latex with hydrazine as the hydrogen donor in the presence of an oxidizing agent and a metal-ion activator at atmospheric pressure [692]

6.3.3. *Producers and Production Capacities*

The worldwide production capacity of HNBR was ca. 4900 t in 1991. Of this, 1800 was attributed to Nippon Zeon (plant in Takaoka, Japan, trade name Zetpol) [693]; 1500 t to Zeon Chemicals, a subsidiary of Nippon Zeon (plant in Bayport, Texas, United States, trade name Zetpol); and 1600 t to Miles Inc., a 100 % subsidiary of Bayer AG (plant in Orange, Texas, United States, trade names Therban, Tornac) [694].

6.4. Polyphosphazenes (see also ⟶ Inorganic Polymers – Polyphosphazenes)

6.4.1. *Properties, Grades, and Applications*

Introduction. A rapidly growing branch of material science has been the area of inorganic and organometallic polymers. Polymers with main-chain building blocks containing atoms other than carbon offer a great variety in structure and often have greater bond energies than C – C single bonds with ca. 343 kJ/mol. An example is the P – N bond with 615 kJ/mol.

Polyphosphazenes, as one example of a polymer with an inorganic backbone, contain pentavalent phosphorus in a chain of alternating phosphorus and nitrogen atoms joined by alternating double and single bonds. All polyphosphazenes are grouped together under the abbreviation Z, with fluoroalkoxy elastomers being designated FZ and the aryloxyphosphazenes PZ.

The two organic moieties attached to each phosphorus atom can be varied to yield a very broad spectrum of materials with different properties; more than 300 different structures have been documented to date with new application under evaluation [695-719].

History. The ring-opening polymerization of hexachlorocyclotriphosphazene [940-71-6] [16422-79-0] dates back to 1897 and the work of STOKES. Without control, this reaction yields a cross-linked, rubbery material (also known as “inorganic rubber”), which quickly hydrolyzes [720-725].

Synthesis of the cyclic precursors from the reaction of phosphorus pentachloride and ammonia dates back to 1834 to the work of ROSE [726], and LIEBIG and WÖHLER [727]. This reaction usually leads to a mixture of cyclic trimers, tetramers, and linear oligomers with a degree of polymerization up to 20.

$$n + mPCl_5 + n + mNH_4Cl \rightarrow$$
$$\left(PCl_2=N\right)_n + \left(PCl_2-N\right)_m + 4\,HCl$$

n = 3,4; m = 5 – 20

In 1964, ALLCOCK et al. filed a patent for the preparation of a soluble poly(dichlorophosphazene) [26085-02-9] via a ring-opening polymerization of hexachlorocyclotriphosphazene and subsequent conversion of the resulting polymer to poly(organophosphazenes), which are suitable for technical applications [728-731].

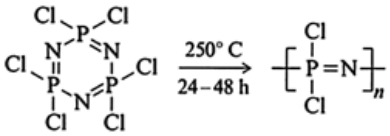

6.4.1.1. *Properties*

The structural variety of polymers with a phosphazene backbone leads to a broad range of physical properties. The common feature of all phosphazene polymers is the inherent flexibility of the main chain and the high dipole moment resulting from the absence of any resonance between the alternating P – N single and double bonds [732-735]. For some polyphosphazenes, rotational barriers for the backbone as low as 420 kJ/mol have been measured. Such low values are in accordance with the low glass transition temperature of many polyphosphazenes [695-697], [736] (see Table 46). Unfortunately, the limited

thermal stability of the simple poly(alkoxyphosphazenes) restricts their technical application despite their low T_g [737], [738].

Table 46. Glass transition temperature (T_g), crystalline melting point (T_m), and decomposition temperature (T_d) for some polyphosphazenes

Substituent	T_g, °C	T_m, °C	T_d, °C	Reference
Cl	–63	30		[739]
OCH_3	–76			[729]
OCH_2CH_3	–84			[729]
OCH_2CF_3	–66	242	360	[729]
OCH_2CF_3, $OCH_2C_3F_7$	–77			[739]
OC_6H_5	–8	390	380	[729]
$NH(CH_3)$	14	140		[740]
$N(CH_3)_2$	–4			[731]
CH_3	–46			[741]

Poly(dichlorophosphazene) with a high molar mass is a transparent elastomer with excellent stress – relaxation behavior. Unfortunately, this polymer degrades on contact with moisture to yield ammonium chloride, ammonium phosphate, and HCl.

In contrast, *poly(organophosphazenes)* are relatively stable to hydrolysis and, depending on the nature of the substituents, behave as thermoplastics or as elastomers. With the exception of the elastomeric polyphosphazenes with C_1 and C_2 alkoxy side groups, most *alkoxy and aryloxy phosphazene homopolymers* are thermoplastics. In these polymers, microcrystallinity prevents the facile conformational changes necessary to impart elasticity. However, reducing the symmetry and regularity by introducing different substituents in a random manner can yield useful elastomeric materials such as the fluoroalkoxides or aryloxides [739-744]. Such “copolymers” also tend to exhibit better thermal stability [745], [746].

Crystalline polyphosphazenes often exhibit a first-order transition at a temperature T_1. Between T_1 and the melting temperature T_m, these polymers are in a mesomorphic or liquid crystalline state, which is neither nematic nor smectic in nature [747-752].

Some polyphosphazenes exhibit outstanding thermal stability as measured by thermogravimetric analysis (TGA), with initial weight losses occurring above 300 °C (Table 46). However, at these temperatures, depolymerization occurs with the formation of small cyclic phosphazenes, and the high decomposition temperatures measured may simply reflect the boiling points of the depolymerization products. Thus, gel permeation chromatography suggests that poly (trifluoroethoxyphosphazene) starts to depolymerize even below 200 °C [753-755]. Studies on the thermal degradation of poly(bis-trifluoroethoxyphosphazenes) and of random and block copolymers bearing trifluoroethoxy and alkoxyalkoxy groups indicate that the thermal stability of the copolymers decreases with increasing content of alkoxyalkoxy groups [756].

The fluoroalkoxy polyphosphazenes exhibit excellent stability toward aqueous acid and base as well as low volume swell in a variety of solvents [757]. Concentrated sulfuric acid does, however, lead to decomposition [730].

Nearly all polyphosphazenes have good flame-retardant properties with oxygen indices up to 70, low smoke emissions and low toxicity of combustion gases [740], [758-764].

6.4.1.2. *Commercial Grades and Applications*

Copolymers with trifluoroethoxide and telomer fluoroalkoxide side groups (trade name Eypel-F, Ethyl Corp., **1**) and with phenoxide and *p*-ethylphenoxide side groups (Eypel-A, Ethyl Corp., Orgaflex, Atochem, **2**) have been commercialized [765], [766].

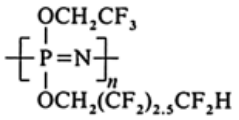

1

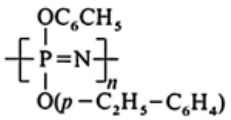

2

Eypel-F is sold as compounds that can be used to manufacture goods having Shore A hardness from 40 to 90. The compounds are priced from about $ 200 to $ 600 per kilogram. Some typical mechanical properties of these compounds are shown in Table 47 [766].

Table 47. Physical properties of typical Eypel-F compounds

Property	Eypel-F		
	6504	7003	7001
Hardness Shore A	65	70	70
Modulus at 100 % elongation, MPa	4.1	9.0	6.7
Tensile strength, MPa	7.6	10.3	9.3
Elongation at break, %	200	120	140
Compression set (70 h at 149°C), %	50	20	30
Density, g/mL	1.74	1.85	1.77
Mooney viscosity (ML4), 100 °C	66	45	45
Crescent tear, die B, kN/m (ASTM D-624)	27	19.3	17.5

To allow a *sulfur or peroxide cure,* ca. 1 % of the side groups introduced carry an unsaturated bond such as, e.g., *o*-allylphenoxy groups. A radiation cure is also possible. Polyphosphazenes can be filled with silica, clay, carbon black, or a combination of these [745], [746]. With a T_g of ca. – 65 °C and a fluorine content of ca. 55 % the fluoroelastomers are suitable for use between – 60 and 175 °C. Their flame retardance is very good as is their fuel resistance. Some typical values of Eypel-F 7003 are given in Table 48 [766].

Table 48. Effects of fuel, lubricant, and hydraulic fluid on Eypel-F 7003

Medium *	Temperature, °C	Swell, vol., %	Tensile strength, %	Hardness, points
ASTM fuel C	100	19	–24	–12
Aviation 100	82	15	–28	–12
ASTM #3	150	2	–16	–2
MIL-H-5606B	135	4	–4	–3
MIL-H-83282A	135	2	–5	–3

* All samples were immersed for 166 h at the temperature given.

Typical *applications* for Eypel-F are as fluid-resistant diaphragms, seals, gaskets, O-rings, and vibration isolators. Although military applications were a major market in the past, they are becoming less important. Automotive and biomedical applications (such as dental liners) are now the main markets [767].

Fluoroelastomers (FKM) and fluorosilicone elastomers (MFQ) are high-performance elastomers with properties similar to polyphosphazenes. Compared to MFQ, commercial fluorophosphazene polymers (Eypel-F) exhibit superior resistance to tear and abrasion, and better flex fatigue.

The properties of FZ at high temperature (up to 140 °C) are comparable to MFQ. The continuous-use temperature for FZ is 175 °C (short term 200 °C), ca. 50 – 70 °C lower than that of FKM. Furthermore, FZ is less stable to hydrolysis than MFQ [755], [768]. The low-temperature flexibility of FZ (to – 65 °C) is comparable to that of MFQ and considerably lower than that of FKM (–18 °C) [768], [769]. At room temperature, FZ shows a greater tendency to swell in a variety of fluids than FKM; but at higher temperature, FZ performs better [743], [769-771]. Table 49 compares some properties of these elastomers.

Table 49. Comparison of some properties of FZ, MFQ, and FKM *

Property	FZ	MFQ	FKM
Abrasion and tear resistance	F	P	G
Heat aging at 140°C	E	E	E
Heat aging at 180 C	G	G	E
T_g, °C	–65	–68	–18
Resistance to			
Aliphatic hydrocarbons	E	E	E
Aromatic hydrocarbons	E	G	E
Alcohols	P	G	G
Aqueous solutions	F	G	G
Oxidation – ozone	E	E	E

* E = excellent, G = good, F = fair, P = poor.

Poly(aryloxyphosphazenes) are sold as foam tubing or foam sheeting with prices ranging from $ 75 – 270 per square meter for sheeting and from $ 33 – 49 per meter for tubing. Foam sheeting finds its major application as a halogen-free, flame-resistant noise insulator. Foam tubing is used mainly for pipe insulation, with military outlets forming the major market. Typical properties of Eypel-A foam are given in Table 50 [772].

Table 50. Typical properties for Eypel A foam

	Typical value
Flammability	
Limiting oxygen index (ASTM D-2863)	44 %
NBS smoke density (ASTM E-662)	
Flaming	70
Nonflaming	40
Surface flammability (ASTM E-162, flame spread index)	11
Acid Gas Generation (MIL-C-24640, 4.8.16)	0 mg HCl/g
Mechanical properties	
Density (ASTM D-1622)	72 kg/m^3
Compressive res. (ASTM D- 1056)	14 – 55 kPa
Tensile strength (ASTM D-412)	170 – 240 kPa
Compression set (ASTM D-1667)	25 %
Noise reduction coefficient (ASTM C-423)	0.3
Thermal conductivity (24°C, ASTM C-518)	0.037 W m^{-1} K^{-1}

6.4.2. *Basic Chemistry and Production Processes*

Both Elf Acquitaine (Atochem) and Ethyl Corp. have pilot plants to produce polyphosphazenes. Although no details have been disclosed by these companies, Ethyl Corp. operates under license from Firestone, whereas Elf Acquitaine apparently has its own technology.

Production of Poly(Dichlorophosphazene). A major stumbling block in the commercialization of polyphosphazenes is the high price of the “monomer” hexachlorocyclotriphosphazene. However, the patent literature documents a process that avoids isolation of the “monomer” [773-775]. In this process,

linear phosphazene oligomers [774] with an average degree of polymerization between 4 and 6 are prepared from phosphorus pentachloride and ammonium chloride in chlorobenzene. This reaction requires a long time at high temperature (e.g., 4 d at 130 °C). The solvent is then distilled, and the cyclic material is extracted with hexane. Ammonium chloride is added to the linear oligomers, and the temperature is increased first to 155 – 165 °C for several hours and then to 200 °C. During the process, ammonium chloride and chlorobenzene are added, while during the nitrogen sweep, cyclic oligomers are removed. Yields up to 60 % are claimed based on the phosphorus pentachloride used.

Elf Acquitaine probably prepares its chlorophosphazene polymers via an alternative method based, formally, on the polymerization of *N*-dichlorothiophosphoryl-*P*-trichloromonophosphazene or *N*-dichlorophosphoryl-*P*-trichloromonophosphazene, the latter being prepared from ammonium sulfate, phosphorus trichloride, and chlorine [776-782]. These are polycondensation reactions and are reported to give high yields of essentially linear poly (dichlorophosphazene).

The mechanism of polymerization, catalysis, and cross-linking reaction of the *cyclic trimer* in bulk and in solution has been summarized by HAGNAUER [783]. Additional development work to improve the efficiency of polymerization and minimize cross-linking has been carried out [784-787]. The simple polymerization of the cyclic trimer is very sensitive to traces of phosphorus pentachloride, which can act as a chain terminator, and also to water, which can act as an accelerator at low concentration but as an inhibitor at higher concentration [788-790]. A wide range of materials have been reported to catalyze polymerization of the cyclic trimer, a reaction that is thought to be a cationic addition polymerization [791].

Polymer Modification. By all these routes the polymer obtained must first be modified to replace all P-Cl groups in order to produce useful polymers. Industrially the introduction of fluoroalkoxy or aryloxy groups is probably carried out in either cyclohexane or a mixture of cyclohexane and C_5 – C_7 alkanes; the alkali-metal alkoxide is added as a solution in tetrahydrofuran [791].

The substitution reaction of linear poly(dichlorophosphazene) with alkali-metal alkoxides occurs readily and completely with sterically nondemanding substituents in refluxing benzene or tetrahydrofuran [788], [791-795]. However, if bulky substituents are introduced, the reaction may be incomplete, and more vigorous reaction conditions are required to achieve complete reaction within a reasonable time [742].

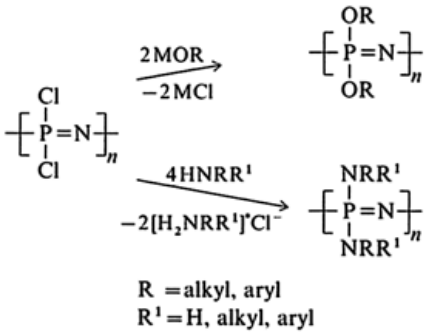

Substitution of the chlorine in poly(dichlorophosphazene) takes place with most nucleophiles, which can also react with the chlorine atoms of simple phosphorus compounds. Thus, a variety of side groups can be attached to the phosphazene backbone. However, organometallic reagents such as organomagnesium or organolithium reagents are exceptions and generally cause simultaneous degradation of the polymer, thus preventing access to the more thermally stable polyalkyl- and polyarylphosphazenes [795].

Two or more different alkoxy or aryloxy groups can be introduced by simultaneous reaction with mixtures of nucleophiles; this method is used commercially [745], [746]. Provided the nucleophiles are similar in shape and reactivity their distribution will be random [739], [742]. An alternative route is the metathetical exchange of ligands by treating an alkoxyphosphazene polymer with an alkali-metal salt of a second nucleophile [796].

Direct Routes to Polyphosphazenes. A direct route to a poly(trifluoroethoxyphosphazene) with moderate molar mass via the thermal polycondensation of tris(trifluoroethoxy)-*N*-trimethylsilylphosphoramine has been described by FLINDT and ROSE [797]. Subsequently, this synthetic route was developed by MATYJAZEWSKI et al., using various anionic initiators, such as alkoxides, fluoride ion, Lewis acids, or strongly nucleophilic amines. The reactions can be carried out at as low as 70 – 120 °C to yield high molar mass linear polymers [798], [799], [800].

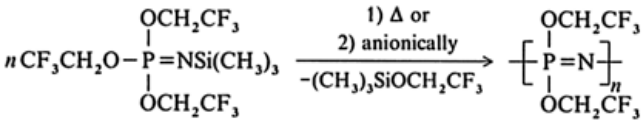

A similar route, leading to poly(organophosphazenes) was introduced by NEILSON and WISIAN – NEILSON, based on the thermal polymerization of *N*-trimethylsilylphosphoranimines having trifluoroethoxy and trimethylsilyl leaving groups [698], [741], [801].

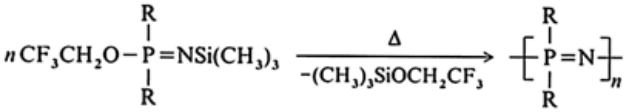

A development to broaden the range of properties is the introduction of other groups or atoms into the phosphazene backbone, allowing, e.g., ring-opening polymerization at lower temperature, as in poly(carbophosphazenes) [802], poly(thiophosphazenes) [803], or poly(thionylphosphazenes) [804]. Another approach is to combine the properties of both the phosphazene backbone and the poly- or oligosiloxane side chains [805], [806], [807].

In the late 1990s a new method for the direct synthesis of organo-substituted polyphosphazenes at ambient temperature has been developed [808], [809]. This route facilitates the control of the molecular mass and provides products with a narrow molecular mass distribution.

6.4.3. *Producers and Markets*

In 2003, only two firms are known to be producing commercial quantities of polyphosphazenes: Ethyl Corp. (Eypel F and A) and Atochem (Elf Acquitain, Orgaflex AMF). Both firms are reported to have pilot plants producing some 25 – 50 t of raw polymer annually [810].

In the past the high price of these elastomers has restricted their use to very specialized applications. New developing markets may be envisioned in which biocompatible, biologically active, or biodegradable polymers are required [811-814], e.g., photo or radiation cross-linkable polyphosphazene compounds for encapsulating biological active materials [815-817, 827, 828]. Novel applications are as materials for flame retardant polymers [818-820], membranes [821-826] and as solid solvents for ionic conduction in rechargeable batteries [829-838]. However, no application is so fast growing at the moment that any expansion of production capacities is announced.

[Top of Page]

7. Thermoplastic Elastomers

Stephen D. Pask, Andreas Ostrowicki, Bernd Stollfuss

7.1. Introduction

Although thermoplastic elastomers are not synthetic rubbers, and are not dealt with in the introduction or in the market analyses of this article, they compete with synthetic rubbers for a variety of applications. The purpose of this chapter is simply to give a brief overview of these materials since thermoplastic elastomers are treated in ⟶ Thermoplastic Elastomers.

7.1.1. *Definition*

Thermoplastic elastomers (TPE) are polymer-based material that can be processed like thermoplastics at elevated temperature, using conventional processes (such as injection molding, extrusion, blow molding), and exhibit elastomeric properties at their use temperature.

Thermoplastic processing behavior is attained by reversible or chemical cross-linking, such as:

Block copolymer association (glassy or crystalline domains)
Random copolymer segment association
Site association
Chemical bonding

The elasticity of the TPE results from the combination of physical cross-links and entanglements between the rubbery chains or is the result of a heterogeneous dispersion of the rubber phase in semicrystalline thermoplastics ⟶ Thermoplastic Elastomers – Morphology.

Materials ranging form soft rubbers to low-modulus thermoplastics are marketed as TPE, which makes a precise definition of this term diffilcult. However, the following groups of materials can be identified in terms of the basic chemistry involved:

1. PVC blends
2. Styrene block copolymers (SBC), e.g., styrene – butadiene – styrene coplymers (SBS), styrene – isoprene – styrene copolymers (SIS), styrene – ethylene/butene – styrene copolymers (SEBS)
3. Non-cross-linked thermoplastic elastomeric alloys (TPEA)
   - Thermoplastic polyolefin rubbers (TPO), mostly blends of PP and EPM
     - Physically blended (compounded) TPO (CTPO)
     - Reactor TPO (RTPO), produced directly in polymerization reactors.
4. Partially or fully cross-linked elastomeric alloys (EA) or dynamically vulcanized alloys (DVA), i.e., thermoplastic vulcanizates (TPV)
   - Olefinic TPV-based EPDM – PP blends (O-TPVs)
   - Natural rubber – polyolefin TPVs (NR-TPV)
   - Nitrile rubber-based TPVs (NBR-TPV)
   - Butyl rubber-based TPVs (IIR-TPV)
5. Thermoplastic polyurethanes (TPU)
6. Copolyesters (COPE)
7. Copolyether- or copolyesteretheramides (COPA, PAE)

Essentially, the mechanical properties at service temperatures of all TPE depend on a physical cross-linking mechanism that, at processing temperature, breaks down and allows the material to flow. The detailed processes are, however, specific to individual materials. An overview of the different classes of materials and some key properties are shown in Figure 27, a comparison of properties of TPE is given in Table 51.

Figure 27. Thermoplastic elastomers — an overview

⟶ = special grades exist that have an even better performance

Table 51. Comparison of properties of TPEs

TPE	Hardness (Shore)	300 % Modulus, MPa	Tensile strength, MPa	Hard block T_m, °C	Soft block T_g, °C	Service temperature, °C	Oil resistance
SBS	60A – 95A	2 – 6	20 – 32	80 (T_g)	– 90	– 70 – 80	poor
TPO	45A – 75D	7 – 14	7 – 25	140 – 160	– 60	– 50 – 120	poor
TPU	70A – 75D	8 – 35	35 – 45	190	– 20 to – 45	– 40 – 120	good
COPE	90A – 70D	8 – 28	7 – 55	180	– 50	– 50 – 150	good
COPA	60A – 75D	15 – 30	21 – 50	120 – 275	– 50	– 40 – 80	good

7.1.2. *Global Market*

A detailed market analysis of TPE is made difficult by the numerous smaller companies compounding the base materials — such as simple polyolefin blends, SBC, and NBR – PVC — to produce customer-specific and captive use compounds. However, the trends and a prognosis for the near future can be seen in Figure 28.

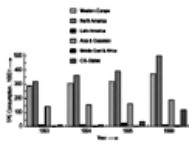

Figure 28. World TPE market [903]

The worldwide TPE growth rate from 1994 to 1999 is predicted to be 5 – 6 % annually according to the IISRP: More detailed analyses suggest that the growth rates for different classes of TPEs could vary considerably. The consumption figures for different types of TPE in Europe for 1991, 1995 and 2000 are listed in Table 52, together with the annual growth rate.

Table 52. Consumption of TPEs in Western Europe by product classes

Type of TPE	Year 1991	1995	2000	Annual growth rate, %
SBC	117	145	165	2.5
TPO	79	90	121	6
TPV (TPEA)	20	24	41	11
TPU	32	52	67	5
COPE	8	12	20	10
COPA	4	5	7	5
Total	260	329	421	5

An estimated breakdown of the TPE market in terms of end use is given in Table 53:

Table 53. Breakdown of the TPE market in terms of end use

Application	Year 1995	2000
Automotive	30	34
Wire, cable	4	5
Shoe industry	13	10
Polymer modification	11	11
Tubes, hoses	3	3
Modified bitumen	12	11
Construction materials	2	2
Adhesives	10	9
Medical goods	1	2
Films, sheets	2	2
Others	3	3

7.2. PVC Blends

Blends of polyvinyl chloride (PVC) and nitrile rubber have been described in the patent literature since the early 1940s and have been the subject of many reviews [839]. PVC is often added to NBR in conventionally cross-linked systems to impart ozone resistance. In a 1 : 1 mixture of PVC and NBR, both with a molar mass of ca. 1×10^5 g/mol, PVC is completely miscible with NBR containing between 35 and 41 % acrylonitrile [840]. More recently, specially developed NBR, more or less precross-linked, has become available for mixing with PVC at levels of 20 – 80 % to produce thermoplastic elastomer compounds.

The addition of NBR to PVC compounds lowers the compression set, increases abrasion resistance, improves low-temperature flexibility, increases fuel and oil resistance, and improves flex- and cut-growth resistance. Comparison with other TPE and with conventional elastomers [841] show, however, that NBR – PVC blends have a considerable compression set and that their properties deteriorate at higher temperature.

The majority of PVC – NBR thermoplastic elastomers are compounded in-house for captive use. An exception to this is Denka LCS, an improved blend sold by Denka (Denki Kagaku Kogyo Kabushiki Kaisha); some of its properties are given in Table 54 [842].

Table 54. Some properties of Denka LCS

Property	Grade Z-6050	Z-6060	Z-6070
Hardness(JIS K 6301), JISA	50	60	70
Tensile strength(JIS K 6723), MPa	11	13	14
Ultimate elongation(JIS K 6723), %	440	420	400
Tear strength(JIS K 6301 (B)), N/mm	25	30	37
Compression set(JIS K 6301), %			
22 h at 70 °C	39	39	38
70 h at 100 °C	52	49	49
Specific gravity	1.12	1.14	1.15

7.3. Thermoplastic Polyurethanes

Polyurethane thermoplastic elastomers (TPU) were one of the first types of TPE to be commercialized, coming on the market in the 1950s. Because of their high strength and abrasion resistance they are used in a variety of applications from ski boots to cable sheathing. An excellent review of these materials is given in [843]. Thermoplastic polyurethanes consist of linear multiblock copolymers with soft segments having low T_g that are separated by hard segments forming semicrystalline domains and providing dimensional stability. In addition to the incompatibility of the hard and soft domains, interurethane hydrogen bonding promotes phase separation and is, at least in part, responsible for typical TPU properties. Phase separation is, however, incomplete [844-848]: The more nonpolar the soft segment, the greater is the degree of phase separation.

Soft Segments. Of the wide range of possible starting materials only a small number are of commercial interest. Some commercially important polyols and the corresponding T_g of the soft domains in TPU are given in the following:

Poly(ethylene adipate)	–25 °C
Poly(butene-1,4 adipate)	–40 °C
Poly(ethylene butene-1,4 adipate)	–30 °C
Poly(hexamethylene 2,2-dimethylpropylene adipate)	–30 °C
Polycaprolactone	–40 °C
Poly(diethylene glycol adipate)	–30 °C
Poly(hexanediol-1,6 carbonate)	–30 °C
Poly(oxytetramethylene)	–80 °C

The molar mass of the glycols used is normally between 600 and 4000 g/mol. Under normal conditions the crystallization of soft segments occurs only at very low levels of hard segment or after tempering, but it is important in terms of stress-induced crystallization, which leads to self-reinforcement of TPU under load [849].

Hard Segments. The hard segments in TPU are formed from diisocyanates and chain extenders. In addition to the most important diisocyanate, 4,4′-methylenediphenyl diisocyanate (MDI), a number of other *isocyanates* are used as raw materials for TPU [843], [850-853].

4,4′-Methylenediphenyl diisocyanate (MDI) [101-68-8]

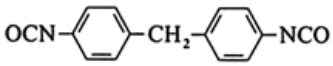

Hexamethylene diisocyanate (HDI) [822-06-0]

$OCN–(CH_2)_6–NCO$

1-Isocyanato-3-isocyanatomethyl-3,5,5-trimethylcyclohexane (isophorone diisocyanate, IPDI) [4098-71-9]

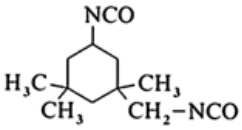

4,4 ′-Diisocyanatodicyclohexylmethane (H_{12}-MDI) [5124-30-1]

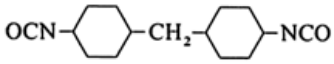

3,3′-Dimethyl-4,4′-biphenyl diisocyanate (TODI)

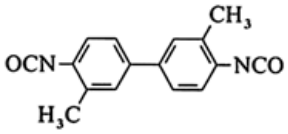

2,4- and 2,6-Toluene diisocyanate (TDI) [584-84-9]

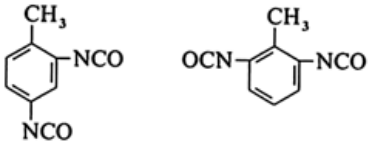

1,5-Naphthalene diisocyanate (NDI) [3173-72-6]

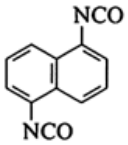

The most important *chain extenders* which, together with the diisocyanate, determine the characteristics of the hard domains in TPU are ethylene glycol, 1,4-butanediol, 1,6-hexanediol, and hydroquinone bis(2-hydroxyethyl) ether.

Thermoplastic polyurethanes are prepared by mixing the ingredients at high temperature using either the so-called one-shot method (⟶ Polyurethanes) or the prepolymer method (⟶ Polyurethanes – Reactive One-Pack Systems), whereby the polyol and the diisocyanate are reacted to give a prepolymer before the chain extender is added. Both processes are carried out commercially by using either reaction extruders or a “belt process.” In the latter, the mixed ingredients are poured onto a belt and allowed to solidify before either being chopped directly into pellets or being blended and pelletized.

The ratio of diisocyanate to the sum of the groups which can react with the isocyanate is critical for the properties of the final product; at less than 0.96 the molar mass of the products falls below ca. 40 000 g/mol and its properties are inadequate [854]; at a ratio greater than 1.1, cross-linking occurs and processing becomes more difficult.

Thermoplastic polyurethanes are available in the Shore hardness range 75 A to 75 D. The *properties* of an ester, an ether, and a special ether – ester type TPU of medium hardness are summarized in Table 55.

Table 55. Some typical properties of ester, ether, and ether – ester TPU

Property	Type of TPU		
	Ester	Ether	Ether – ester
Hardness (DIN 53 505)			
Shore A	92	93	96
Shore D	44	42	56
Tensile strength, MPa	45	40	40
Ultimate elongation (DIN 53 504), %	450	380	400
Tear strength (DIN 53 515), kN/m	90	90	110
Abrasion loss (DIN 53 516), mm^3	30	25	35
Compression set (DIN 53 517), % (24 h at 70 °C)	60	60	50
Density, g/cm^3	1.23	1.14	1.20
Stability against*			
Weathering	+	0	+
Hydrolysis	0	+	+
Microbial attack	+/–	++	++

* ++ excellent, + good, 0 mediocre, – poor, +/– sometimes good, sometimes poor.

Thermoplastic polyurethanes can be *manufactured* by most producers of polyurethane resin. The major producers of TPU with dedicated capacities and trade names are given in Table 20.

Table 56. Major world TPU producers

Producer	Capacity, 10^3 t/a	Trade name
B. F. Goodrich	24	Estane
Bayer (Miles, United States)	23	Desmopan (Texin)
BASF (Elastogran)	24	Elastollan
Others	28	

7.4. Styrene Block Copolymers

Styrene block copolymers represent the largest volume of TPE on the market. Their ability to be extended with both oils and fillers has increased their range of applications so that they are now used both as thermoplastic elastomers (e.g., in shoe soles) and as sealants, coatings, and asphalt modifiers.

These materials are composed essentially of three blocks — styrene – rubber – styrene — whereby the central rubber block is either polybutadiene (B), polyisoprene (I), or hydrogenated polybutadiene. The latter, depending on detailed polymerization conditions, leads to a copolymer block that can be identified formally as a copolymer of ethylene and butene (EB). Thus, the three major types of SBC are abbreviated by the letters SBS, SIS, and SEBS. More complex structures, such as star-shaped $(S\text{-}B)_3$ or $(S\text{-}B)_4$ are also available commercially. In all types, the two phases form separate domains at service temperature. Thus, at room temperature, the rubber domains are physically cross-linked by the polystyrene domains, and the proportion of the two phases determines the hardness (stiffness) of the material. The phase separation is better with SEBS so that the service temperature limit is closer to the T_g of simple polystyrene (ca. 100 °C). These hydrogenated polymers are also less susceptible to aging, which allows them to be used for more demanding applications. In particular, the purity with which these materials can be manufactured is leading to increasing use in medical and food packaging applications [855]. Typical properties for an SBS, an SIS, and an SEBS material are given in Table 57.

Table 57. Typical properties for linear SBS, SIS, and SEBS [856]

Property	SBS	SIS	SEBS
Tensile strength, MPa	32	20	34
Elongation, %	880	1200	500
Set at break, %	10	10	
Hardness, Shore A	71	52	75
Specific gravity	0.94	0.93	0.91
Styrene – rubber ratio	31 : 69	21 : 79	29 : 71

Production of SBS and SIS. Block copolymers based on styrene and butadiene or isoprene are manufactured essentially by three processes:

1. Bifunctional initiator process
2. Three-stage sequential addition
3. Coupling process

The *bifunctional initiator process* uses initiators such as sodium naphthalene or organo-dilithium compounds, with the latter generally being preferred [857]. Although initiation with dilithium compounds leads to materials of very narrow molar mass distribution, the limited solubility of the lithium initiators in nonpolar solvents restricts the process to polar solvents such as ethers. In these solvents, butadiene polymerizes anionically mainly by a 1,2-addition process, and the T_g of the polymer rises. This difficulty can be surmounted by adding small amounts of polar solvent to nonpolar hydrocarbon solvents in which butadiene polymerizes mainly via 1,4-addition [858].

The *three-stage sequential addition process* involves polymerizing first a styrene block, then a butadiene (isoprene) block and finally another styrene block. The process is generally carried out by using hydrocarbon-soluble organolithium initiators (e.g., *n*- or *sec*-butyllithium) in hydrocarbon solvents (e.g., cyclohexane, benzene, toluene, hexane). This process is very effective for producing pure block copolymers but is limited to linear polymers [859].

In the *coupling process* the first two stages are identical to those used for a three-stage sequential addition, but the final stage is a coupling reaction. Here the living polymer chains are reacted with di- or multifunctional reagents, such as silanes [860], haloalkanes [861], esters [862], [863], or divinylbenzene [864]. The advantages of this method lie in the ability to produce not only linear but also radial (star) polymers; the precise control of stoichiometry required in the coupling reaction is the major disadvantage.

For all three methods the reaction system must be controlled carefully to exclude impurities, since the presence of homopolymer and especially diblock polymers leads to a deterioration in the properties of the final products.

Production of SEBS. The hydrogenation of SBS polymers to SEBS is usually carried out on polymers containing a large amount of 1,2-double bonds (≥ 40 %). Selective hydrogenation of the resulting diene segments is achieved by carrying out the hydrogenation at low temperature (ca. 35 – 150 °C) using cobalt or nickel acetyl acetonates as catalyst that are reduced with alkylaluminum compounds in situ. A typical solvent is cyclohexane, and the reaction time is usually <4 h between 1.7 and 10 MPa [865-867].

Economic Aspects. Some 50 % of the world market for SBC is supplied by Shell under the trade name Kraton. Shell's total capacity according to several commercial studies is ca. 200 000 t/a. Shell is the only producer of hydrogenated grades with a capacity for these grades of some 40 000 t/a. According to various sources, this capacity is being expanded, and Enichem Elastomeri is expected to be producing hydrogenated grades commercially by the end of 1992 . Other major producers of SBC include Firestone (32 t/a), Enichem (120 t/a), and DOW – Exxon (32 t/a). However, much of the material, based on SBC, that actually reaches the end user as a TPE has been compounded by smaller companies with fillers, oils, or other polymers to achieve materials having an optimum price – property ratio.

7.5. Thermoplastic Polyolefins

Three groups of materials are included under the term thermoplastic polyolefins: (1) simple (compounded) blends (sometimes abbreviated CTPO); (2) reactor-modified polyolefins (abbreviated RTPO); and (3) partially or fully vulcanized blends (TPV).

Thermoplastic Blends. Thermoplastic polyolefins are polypropylene (PP)-based compounds with sufficient elastomer content to impart partial elastomeric qualities to them. As regards production volume, these PP-based elastomers are among the fastet-growing types of thermoplastic elastomers. Some 70 % of the total TPO volume consists of compounded materials, largely blends of partially crystalline polypropylene (homopolymers or copolymers) and amorphous or,semicrystalline ethylene – propene copolymers (EPM or EPDM). As simple blends, their morphologies depend not only on the viscosities and properties of the base polymers, but also on the manufacturing process. With 20 – 80 vol % rubber the blends usually have a continuous rubber phase and often comprise two cocontinuous phases.

The use of EPM or EPDM with a certain proportion of crystalline ethylene units increases the tensile strength and elasticity at room temperature, but these improvements are lost at higher temperature (> ca. 70 °C). In addition to a large number of proprietary materials [868], many end users are now compounding their own blends for their specific needs since further compounding of the proprietary blends is usually not successful. Work to expand the property range of PP-based elastomers centers on improved performance of the elastomer phase. By upgrading the blendability of these rubbers or using different elastomers, suppliers can produce TPOs with a wider range of properties. For example, ethylene – α-olefin copolymers, produced by Dow's Insite metallocene process, are said to be more miscible with PP than EPM and to provide up to 20 % higher cold-temperature impact strength than equivalent contents of EPM in the blend. These metallocene-catalyzed polymers also provide more rigidity to a TPO than EPM, because of their crystallinity of 10 to 15 % (the crystallinity of EPM is only 0 to 5 %).

The rather poor elasticity of these simple blends can be improved by *partial cross-linking* of the rubber phase during mixing with, e.g., peroxides [869], [870]. The action of peroxides leads to a reduction in the molar mass of the PP and probably also to some grafting.

Both simple blends and partially cross-linked systems can be filled with inorganic fillers and extended with oil.

In the 1990s *blends of natural rubber* (*NR*) *and PP* have become available commercially [871]. In these systems the PP appears to provide some protection for the NR against oxidation, while the double bonds in NR can still be used to prime the article surfaces ready for painting or bonding. These systems also appear to show some strain-induced crystallization typical of NR. The properties of two commercial NR – TPO products are given in Table 58.

Table 58. Properties of two NR – TPO products [871]

Property	Telcar Z 1189	Vitacom 7001
Hardness Shore A	56	70
Tensile strength, MPa	4.0	9.8
Elongation, %	290	320
Compression set, %		
22 h at 70 °C		37
22 h at 100 °C	43	44

Because of the nature of the supply (i.e., a large number of small compounders, several hundred worldwide) and because a considerable quantity of material is prepared for captive use, a breakdown in terms of suppliers of these materials is not only difficult but also of doubtful reliability.

Thermoplastic Vulcanizates. If nonradical cross-linking systems are employed, the rubber phase in a TPO can be essentially completely cross-linked while a thermoplastically processible material is maintained (dynamic vulcanization, see below).

Typical cross-linking systems for TPV are phenol – formaldehyde resins, metal halides, and zinc oxide [872]. Since the original patent of GESSLER [873], CORAN and PATEL have developed the idea of producing materials from most rubbers and a range of thermoplastics [874-880]. In contrast to the simple and partially cross-linked systems, these fully cross-linked systems have a fixed morphology, and their properties are less sensitive to additional processing steps. Indeed, such materials can be reworked several times without the loss of properties [881]. The morphology of TPV comprises small (ca. 1 – 5 µm particles of cross-linked rubber in the thermoplastic matrix, even up to rubber contents of ca. 80 – 90 vol %. The rubber volume can be extended with considerable amounts of mineral oil and the entire blend filled with inorganic filler.

Commercially, the most successful TPV (based on EPDM and PP) is marketed by Advanced Elastomer Systems (capacity probably ca. 60 000 t/a) under the trade name Santoprene. For applications demanding better oil resistance, Advanced Elastomer Systems markets a TPV based on nitrile rubber and PP under the name Geolast. The hot-oil resistance (≥ 120 °C) of this TPV is, however, not as good as that of an NBR vulcanizate because of the lack of resistance of the PP matrix.

A comparison of the properties of simple blends, partially cross-linked TPO, and TPV is given in Table 59.

Table 59. Comparison of the properties of different types of EPDM – PP TPO

Property	Simple blend	Partially cross-linked blend	Fully cross-linked TPO

Hardness, Shore A	78	72	75
Tensile strength, MPa	12	5.5	8.5
Elongation, %	650	350	490
Compression set, %			
22 h at 70 °C	75	60	38

Compounding thermoplastics and rubbers can be carried out either in Banbury mixers, provided they are constructed for the higher temperatures necessary (ca. 200 – 240 °C) compared to those normally encountered in rubber compounding, or in continuous mixing machines [882], [883]. Dynamic vulcanization essentially involves adding the cross-linking chemicals after the polymers have been mixed and continuing the mixing until the cross-linking reaction is complete. Although this reactive processing can be carried out in Banbury mixers, the mixing times are inordinately long (10 – 20 min) so that continuous mixers are employed commercially and running temperatures are increased (for EPDM – PP blends, 220 – 240 °C) to reduce the time required to complete cross-linking [882].

Reactor-modified thermoplastic polyolefins are based on relatively new technology, but due to the cost advantage they represent a very dynamic market. Besides being used alone, they can be used, for example, to modify PP to produce a new generation of TPO. Major areas of application are in films, hoses, and automobile trims.

These materials are produced by gas-phase polymerization using modified Ziegler catalysts and gas-phase, fluidized-bed technology [884]. The most widely known process is probably that licenced by Union Carbide under the trade name Unipol [885], [886] (⟶ Polyolefins). Polymers are sold under the trade name Flexomer. Exxon Chemicals (trade name: Exxtral) is said [887] to have its own catalyst technology, and Montecatini (Himont) has its own process called Spheripol (polymer trade name Spherilene).

7.6. Copolyesters

The copolyesters were commercialized in the early 1970s by Du Pont under the trade name Hytrel [888] and by Toyobo [889] under the trade name Pelprene. They have often been referred to as engineering thermoplastic elastomers because of their high flexural modulus. Because of their excellent processibility, rapid crystallization rates, and broad range of physical properties (see Table 60), these materials are widely accepted for component integration in the machine and automotive industries. Of particular interest is their “perfect spring” behavior at low elongation: up to 7 % for the stiffer materials and 25 % for the softer grades. The service temperature range for COPE (– 40 to + 150 °C) is wider than that of TPU [890], [891].

Table 60. Some properties of COPE [894]

	Hardness (Shore D)		
Property	**32**	**54**	**70**
Tensile strength, MPa	15.4	47.0	55.0
Elongation, %	880	660	535
Yield strength, MPa	4.8	13.0	28.0
Compression set, %			
22 h at 70 °C	90	53	
Melting point, °C	106	189	189

These multiblock thermoplastic elastomers are *synthesized,* for example, from dimethyl terephthalate, poly(tetramethylene ether) glycol, and tetramethylene glycol via a transesterification reaction in the melt using titanium catalysts [892]. During the final stage of polymerization, excess short-chain diol is removed by distillation. The lower the reaction pressure, the faster is the rate of polymerization, as measured by the increase in inherent viscosity. Other monomers are also used [893] to produce a family of materials whose multiblock structure — with hard segments composed of crystalline poly (tetramethylene terephthalate) and soft segments composed of poly(tetramethylene ether terephthalate), for example — is analogous to that of TPU (see Section Thermoplastic Polyurethanes) and polyetheramides (see Section Copolyether- and Copolyetheresteramides).

A more recent, detailed review of this class of materials can be found in [895].

The major *supplier* of COPE worldwide is Du Pont, which also produces material in Japan in a joint venture with Toray. General Electric (Lomod) and Hoechst – Celanese are the major competitors for the world market of some 40 000 t/a. Detailed capacities could not be ascertained.

7.7. Copolyether- and Copolyetheresteramides

The polyether(ester) – polyamide block copolymers, often abreviated generally as PEBA [896], [897], have similar multiblock microstructure to TPU and COPE (see Sections Thermoplastic Polyurethanes and Copolyesters). In copolyetheramides (PEA), the *soft segments* are derived from aliphatic polyesters, whereas in copolyetheresteramides (PEEA) the soft segments are derived from aliphatic polyethers linked to the hard segments via an ester moiety. The *hard segments* can be formed from aliphatic or aromatic amides.

Depending on the amide type the hard segments are either glassy or semicrystalline. These segments also control the high-temperature properties of this class of TPE. The physical properties of these materials are similar to those of COPE, but these amide systems generally exhibit lower hardness (> ca. 60 Shore A), lower moduli, and significantly lower specific gravity (typically ca. 1.01, compared to 1.15 – 1.22). Compared with COPE, PEBA polymers have a smaller service temperature window (– 40 to 80 °C [898]). Those types based on highly crystalline aromatic amides can have good mechanical properties up to ca. 170 °C continuous use [899]. They find application, for example, in gaskets and seals, conveyor belt rollers, tubes and hose, shoes, and medical uses.

Two approaches are used for the *synthesis* of block polyamides: via a *polycondensation* of aromatic diisocyanates with dicarboxylic acids as chain extenders to give the hard amide blocks and a carboxylic acid-terminated polyester or polyether for the soft segments. Typically MDI is the diisocyanate, and adipic or azelaic acid is used as chain extender. The prepolymer soft segments generally have molar masses in the range 500 – 5000 g/mol [900], [901]. Alternatively, the carboxylic acid-terminated polyamide blocks can be synthesized from polyamides using dicarboxylic acid chain limitors at $T \geq 230$ °C and $p \geq 25$ bar. The hydroxyl-terminated polyethers are produced either anionically or cationically from ethylene and propylene oxides or from tetrahydrofuran, respectively. Finally, a *polyesterification* is carried out at 230 – 280 °C under reduced pressure [896], [902].

The major *producers* of polyamide TPE are DOW (Estamid) and Atochem (Pebax). Typical *mechanical properties* of these materials are given in Table 61.

Table 61. Typical properties of polyetheramides

Hardness, Shore D	25	40	63

Tensile strength, MPa	79	36	51
Elongation, %	715	485	380
Compression set, %			
22 h at 70 °C	62	21	5

[Top of Page]

8. Evaluation of the Present Situation and Remarks on Future Trends

Desmond Threadingham, Werner Obrecht, Rüdiger Engehausen, Wolfgang Wieder, Gerhard Wachholz

8.1. Market, Producers, and Capacities

8.1.1. *Market*

Basically two sources of statistical data on rubber production and consumption exist: The London-based IRSG (International Rubber Study Group) and the IISRP (International Institute of Synthetic Rubber Producers), based in Houston, Texas [1238], [1239].

The IRSG is an intergovernmental organization of 17 member states. The strength of the IRSG market data lies in its coverage of NR. The information on SR, however, is less precise, as the data also includes some thermoplastic elastomers such as SBS, SIS, and SEBS and latices [1240]. A differentiation is not always made between carboxylated and noncarboxylated SBR latices. This is crucial insofar as some countries classify carboxylated latices as plastics.

The IISRP is an association of synthetic rubber manufacturers. Approximately 95% of total SR production is now documented by the IISRP, because the Chinese manufacturers also joined this organization. Some of the Russian producers and a few smaller manufacturers in Central Europe and Asia are still absent.

The IISRP publishes data on NR and SR, with the IISRP statistics being characterized by data on individual SR types. In order to have clear distinctions for SR, several years ago the IISRP decided only to publish consumption data on solid rubbers. The only exception is CR which includes a small proportion of latex. For these reasons, the following observations are limited to the solid rubbers, i.e. solid NR and solid SR.[1)]

The trend in rubber consumption in the final decade of the last century was marked by a significant downturn in the early 1990s. (see Fig. 29).

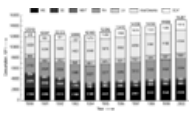

Figure 29. Global consumption (10^3 t) of solid rubber (NR, SR) 1990 – 2000

The collapse of the Communist system in Eastern Europe, and in the Soviet Union in particular, resulted in an economic decline which also significantly affected rubber consumption in this region up until 1994. Superimposed on this trend were the negative after-effects of the Gulf war with a global economic crisis which did not end until 1993.

Development in rubber consumption was correspondingly weak until 1993. From that point onwards, worldwide rubber consumption grew at an annual rate of 3.5% until 2000. Even the Asian crisis in 1998 only stemmed this growth temporarily. The growth in rubber consumption was particularly strong in North America (excluding Mexico) and in the Asian countries, especially China (SCA). China managed to more than double its rubber consumption over this 10-year period.

While worldwide SR consumption only increased by 0.7% p.a. in the 1990s, NR consumption grew by 3.3% per annum (Table 71).

Table 71. Growth in NR and SR consumption, 1990-2000

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	CAGR* (90/2000)
SR Solid, 10^3 t	8.296	7.749	7.614	7.216	7.338	7.787	8.047	8.356	8.281	8.601	8.887	0.7 %
NR Solid, 10^3 t	4.614	4.498	4.759	4.784	5.025	5.259	5.364	5.652	5.777	5.835	6.380	3.3 %
Total, 10^3 t	12.910	12.247	12.373	12.000	12.363	13.046	13.411	14.008	14.058	14.436	15.267	1.7 %

*CAGR: Compound Annual Growth Rate

Worst affected by the economic decline in Eastern Europe (EE) in the early 1990s was SR. The USSR had built up enormous polyisoprene capacities for strategic reasons, in order not to be dependent on NR.

In 1990, NR represented only 9% of total rubber consumption in EE, compared with 30% in North America (NA) and 37% in Western Europe (WE), not to mention the Asian countries, in which the percentages were much higher.

Even in 2000, NR represents only 20% of total rubber consumption in EE (Table 72). The increase compared with 1990 is attributable not so much to a structural adaptation of the rubber-producing industry as in the West, but to SR consumption itself, which is still very low and has not yet reached its previous level. It was not until the end of the 1990s that SR consumption took off as a result of the economic recovery in the Russian Federation.

Table 72. Ratio of solid SR to NR consumption by major region, 2000

	NR Solid	SR Solid
EE	20 %	80 %
NA	33 %	67 %
WE	38 %	62 %
As/Oc	51 %	49 %
SCA	53 %	47 %
World	**42 %**	**58 %**

It remains to be seen whether Russia will make full use of the remaining IR capacities. A short supply of NR in the coming years coupled with increasing prices could create the conditions for further activation of the IR plants.

As expected, NA leads the world with a per capita rubber consumption of 11.7 kg, with WE trailing in second place followed by the other regions.

Although it would clearly be wrong to use NA as a benchmark for other regions because of its transport structure, Table 73. gives an indication of the long-term potential which still exists in other regions in terms of rubber consumption.

Table 73. Solid rubber consumption per capita by region, 2000

Regions	NR and SR 10^3 t	SR 10^3 t	Population $\times 10^6$	NR and SR per capita, kg	SR per capita, kg
NA	3,694	2,464	316	11.7	7.8
WE *	2,920	1,800	458	6.4	3.9
EE	995	794	411	2.4	1.9
As/Oc	4,356	2,126	2,081	2.1	1.0
LA	1,001	599	520	1.9	1.2
SCA	1,916	906	1,383	1.4	0.7
MEA	385	198	902	0.4	0.2
World	15,267	8,887	6,071	2.5	1.5

* incl. Turkey
Source: IISRP, IRSG, United Nations Population Division, estimates by the author.

A glance at the growth in individual SR types shows that consumption of tire rubbers (SBR, BR, IR and IIR) almost stagnated (+0.2 % p.a.) between 1990 and 2000 (see Fig. 30). While the former USSR was still consuming large volumes of SR in 1990, the economic collapse led to a drastic decline in the consumption of SR, and the tire rubbers IR and ESBR in particular. The USSR had a large share of world capacity for these rubbers (81% for IR with 1.1 × 10^6 t and 17% for ESBR with 690 x10^3 t) [1241].

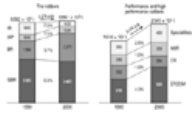

Figure 30. Global consumption (10^3 t) of solid SR by category and type

Unfortunately, the IISRP has not published any data on the solution polymerized rubber SSBR, which is subsumed under SBR in Figure 30. SSBR and BR are tire rubbers which have shown consistently strong growth in the 1990s. In many cases, they have replaced ESBR in the tire treads of “green” passenger car tires. Using silica as a filler, Michelin succeeded at the beginning of the 1990s in developing passenger car tires with improved wet grip and lower rolling resistance without reducing the tire lifetime. These tires rapidly became popular, particularly in WE, because of their lower fuel consumption.

Because of its high air impermeability, butyl rubber (IIR) is virtually irreplaceable as a tire component. The transition from tires with tubes to tubeless tires led to a decline in non-halogenated butyl and increased demand for halobutyl rubber. Butyl rubber is now produced by a total of five manufacturers worldwide: Exxon, Bayer, two manufacturers in Russia, and one in China. Exxon and Bayer together have a capacity share of more than 75% [1242].

In contrast to tire rubbers, the nontire rubber market grew by 2.4% p.a. between 1990 and 2000, with EP(D)M showing the strongest growth at 4.3% p.a.. After ESBR and BR, EP(D)M is now the third most important SR type, with a share of 10%. In the past, its growth has been partly at the expense of CR and NBR.

Specialties meet a growing demand for high-tech articles for use in difficult environments and have shown consistent growth. There are about 16 rubbers in this category with outstanding performance properties. Neither the IISRP nor the IRSG provide detailed information on this segment.

After the highest worldwide rubber consumption so far was reached in 2000, consumption fell (-4.3%) in 2001 as a result of the economic crisis (September 11) with declining markets in cars and tires (Table 74).

Table 74. Solid rubber consumption (10^3 t) forecast by region, 2001-2007

Regions	Rubber groups	2000 Historical	2001 Historical	01/00 (%)	2002 Estimated	02/01 (%)	2003 Forecast	03/02 (%)	2007 Forecast	CAGR (%) 03/07
WE	SR	1,800	1,750	-2.8	1,760	0.6	1,785	1.4	1,895	1.5
	NR	1,120	1,085	-3.1	1,015	-6.5	1,025	1.0	1,080	1.3
	Total	2,920	2,835	-2.9	2,775	-2.1	2,810	1.3	2,975	1.4
EE	SR	794	851	7.2	902	6.0	956	6.0	1,150	4.7
	NR	201	214	6.5	198	-7.5	208	5.1	245	4.2
	Total	995	1,065	7.0	1,100	3.3	1,164	5.8	1,395	4.6
MEA	SR	198	188	-5.1	190	1.1	197	3.7	231	4.1
	NR	187	165	-11.8	167	1.2	174	4.2	206	4.3
	Total	385	353	-8.3	357	1.1	371	3.9	437	4.2
NA	SR	2,464	2,217	-10.0	2,215	-0.1	2,250	1.6	2,440	2.0
	NR	1,230	995	-19.1	1,095	10.1	1,108	1.2	1,220	2.4
	Total	3,694	3,212	-13.0	3,310	3.1	3,358	1.5	3,660	2.2
LA	SR	599	522	-12.9	502	-3.8	515	2.6	606	4.2
	NR	402	378	-6.0	358	-5.3	365	2.0	420	3.6
	Total	1,001	900	-10.1	860	-4.4	880	2.3	1,026	3.9
As/Oc	SR	2,126	2,013	-5.3	2,075	3.1	2,130	2.7	2,350	2.5
	NR	2,230	2,110	-5.4	2,160	2.4	2,210	2.3	2,410	2.2

	Total	4,356	4,123	-5.3	4,235	2.7	4,340	2.5	4,760	2.3
SCA	SR	906	989	9.2	1,090	10.2	1,180	8.3	1,580	7.6
	NR	1,010	1,130	11.9	1,215	7.5	1,300	7.0	1,600	5.3
	Total	1,916	2,119	10.6	2,305	8.8	2,480	7.6	3,180	6.4
World	SR	8,887	8,530	-4.0	8,734	2.4	9,013	3.2	10,252	3.3
	NR	6,380	6,077	-4.7	6,208	2.2	6,390	2.9	7,181	3.0
	Total	15,267	14,607	-4.3	14,942	2.3	15,403	3.1	17,433	3.1

Source: Estimates by the author based on IISRP and IRSG statistics

Only China (SCA) and EE brought a glimpse of cheer to the generally gloomy picture. At the time this article was written, no definitive data were available for 2002, although there should have been some positive growth again in worldwide consumption.

In the medium term, rubber consumption will continue to grow in the wake of a recovering world economy, although differences between the regions may be very great.

In Europe, the main growth will be in EE. In the long-term, NA will be affected by a shift in tire production to more cost-effective regions, as has already occurred in WE. The same applies to countries such as South Korea and Japan. In the coming years, China will continue to be the main growth market.

The IRSG is forecasting 6.4% growth in worldwide rubber consumption for 2003 compared with 2002 and 4.9% for 2004 over 2003 [1243]. Economic growth in the main industrial countries so far does not really support this optimistic prognosis for 2003.

8.1.2. *Producers, Capacities and Capacity Utilization*

As can be seen from Figure 31, worldwide SR capacity was 11.2 × 10^6 t in 2002 [1242]. Consumption, on the other hand, came to 8.7 × 10^6 t (see Table 74). If it is assumed that no major difference exists between consumption and production, this gives a theoretical utilization rate of 78%. It is important to take into account here the fact that the figures mentioned are nominal capacities. Under the assumption that the operating capacity is approximately 10% lower than the nominal capacity, this gives a utilization rate of 86%. In addition to this, some capacity figures appear to be excessive. This applies to the IR capacities in Russia and Romania, for example, with a total capacity of almost 600 × 10^3 t [1242]. According to a different source, the plants in Romania have already been closed for some years [1244]. The breakdown of capacities into BR and SSBR is also problematic, because both types can be manufactured together with certain thermoplastic elastomers (SBS, SEBS, SIS, etc.) in multi-purpose plants, but big manufacturers only give total capacities.

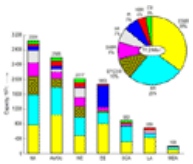

Figure 31. World capacities (10^3 t) of solid SR by region and product group in 2002 (no specialties included)

A comparison of capacity and consumption in the regions shows that excess capacity exists in NA, WE and EE, whereas insufficient cover prevails in China. Because of its expanding demand, China will continue to build up capacity in the coming years.

Whereas capacities continued to be expanded in the second half of the 1990s, many manufacturers suffered from negative or stagnating market growth in rubber consumption in 2001 and 2002, particularly in the large industrialized countries. In the case of commodities, this led to considerable elimination of capacities in the saturated markets. Expansion continued, however, in the case of products with growth potential or in regions with strong accumulated demand.

- *Ameripol Synpol* closed down a 95-kt ESBR plant in the USA in 2001.
- *DSM* wants to part with its 150-kt ESBR plant in the USA. With the sale of its NBR business (tolling agreement) to Zeon in 1999, the biggest EP(D)M producer focused its activities on EP(D)M. A new 80-kt EP(D)M plant is to be started up in The Netherlands at the end of 2002/beginning of 2003 [1245]. DSM is also planning to build a new 40-kt plant in China [1246].
- *Bayer* ceased production of BR (95 × 10^3 t) in Canada at the end of 2002 and concentrated the production of NBR in WE at the La Wantzenau site in France. Bayer had previously increased its BR/SSBR capacities in the USA and WE by 100 × 10^3 t in total.
- *Goodyear* commissioned a new 110-kt BR/SSBR plant in the USA at the end of 2000. It pulled out of NBR production altogether in 2001 and sold the business to Zeon. In the field of solid rubbers, Goodyear now manufactures only tire rubbers.
- *Enichem* put the spin-off company Polimeri Europa up for sale in 2002.
  The following products at seven sites are affected by this: CR, EP(D)M, ESBR, SSBR, NBR, BR, XSBR/ NCSBR latices and thermoplastic elastomers (SBS, SIS) [1247].
- *Exxon* will commission a 90-kt plant for EP(D)M, plastomers and noval polymers in 2003 [1248].
- *DOW* considerably expanded its capacities in the tire rubber market in the second half of the 1990s through acquisitions and takeovers (Karbochem in 1998, Shell in 1999) and became one of the top ten manufacturers. DOW first became involved in tire rubbers in 1997 with the acquisition of BSL Olefinverbund in East Germany. In 2000, DOW expanded its capacity there by another 60 × 10^3 t BR/SSBR. With the takeover of Union Carbide in 2001, DOW acquired a 90 × 10^3 t EP(D)M plant, which operates using gas phase technology. The plant is being used by the 50/50 joint venture DuPont DOW Elastomers, which is also responsible for marketing [1249].
- *DuPont Dow Elastomers* (DDE) was founded by DuPont and DOW as a joint venture in 1996, with both companies pooling their resources in the field of performance and high performance rubber [1250]. DuPont contributed EP(D)M, CR, CM, CSM, EAM and FKM, while DOW contributed the CM business and the “Engage” product class which is manufactured on the basis of “Insite” technology. This technology, which is based on a special metallocene-type catalyst, was also used for the joint venture's new EP(D)M plant in Plaquemine (1997).
- *Shell* pulled out of the rubber business altogether with the sale of ESBR and BR to DOW in early 1999 and of the TPE business (SBS, SIS, SEBS, SEP), including IR, to a private equity investment firm in 2000 [1250].
- Additional capacities came onto the market in the ASEAN countries and in China. China alone increased its capacities by 200 × 10^3 t BR/SSBR between 1996 and 2002. Taiwan Synthetic commissioned a 100-kt ESBR plant in Nantong at the end of 1998. The first Chinese plant for manufacturing butyl rubber (30 × 10^3 t) was started up in Yanshan in 2000.

These few examples show how the focuses are shifting between the regions and how major manufacturers are reacting in markets where growth is less strong.

As can be seen from Table 75, the capacity of the top ten manufacturers worldwide was 5.536 × 10^6 t in 2002, which represented 50% of the total world capacity of 11.2 × 10^6 t. Compared with the major plastic groups such as polyethylene, polypropylene, polystyrene and polyurethane, the SR manufacturing landscape as a whole is relatively fragmented. At most, a high degree of consolidation exists in individual product groups such as IIR and EP(D)M. Apart from that, there is a large number of manufacturers in fierce competition with each other. The Chinese company Sinopec has now moved up to number seven in the world ranking. As in many other areas of industry, China is going for expansion and will try to make increasing headway on the world markets in the coming years. PetroChina, the other big SR manufacturer in China, and currently ranked number 12, is also set to join the top ten in

the next few years.

Table 75. Top 10 producers of solid SR by capacity (10^3 t), 2002*

Companies	Country	ESBR	SSBR	BR	IR	IIR	EP(D)M	CR	NBR	Total	Share(%)
1 Bayer	Germany	100	32	528		255	110	65	95	1,185	10.6
2 Goodyear	USA	267	100	260	90					717	6.4
3 ExxonMobil	USA					432	174			606	5.4
4 Polimeri	Italy	190	40	160			85	40	33	548	4.9
5 JSR	Japan	215	25	72	36		70		40	458	4.1
6 Zeon	Japan	200	20	58	37				110	425	3.8
7 Sinopec	China	80	45	270		30				425	3.8
8 Kumho	S.-Korea	190	27	123			40		20	400	3.6
9 DOW	USA	208	48	139						395	3.5
10 DSM	Netherlands	150					226			376	3.4
Total 10		1,600	337	1,610	163	717	705	105	298	5,536	49.5

* no specialties included
Source: IISRP, Worldwide Rubber Statistics 2002
Remark: Capacities of 50/50 joint ventures were assigned in full to the capacity of the leading producer in that segment

SR manufacturers are having to face increasing concentration and globalization on the part of their customers. This is particularly true of tire manufacturers, although it also applies to manufacturers of technical rubber articles, albeit to a lesser extent. The three top tire manufacturers (Michelin, Bridgestone, Goodyear) now control 60% of the world tire market and can thus exert price pressure on their raw material suppliers. They also operate their own SR production plants and are thus only partially dependent on suppliers. Against a background of lower profit margins, restructuring in the SR industry will continue in the future.

8.2. Market Requirements and Impact on Development of Synthetic Rubber

8.2.1. *Tires and Tire Rubber*

Tire manufacturers have optimized the performance of tires further and further in answer to the constantly increasing requirements regarding safety, economy, and ecology. The key targets for improved tire performance are still lower rolling resistance, better grip performance and higher wear resistance. It is always a challenge to optimize the overall performance in this “magic triangle”, taking into consideration that improvement in one property often can lead to deterioration of another [1252-1257, 1289].

The compromise achieved between these characteristics varies from continent to continent. It is interesting to realize the geographic differences in the requirements for passenger car tires.

Requirement:	Good wet grip	Low rolling resistance	Abrasion resistance
North America	moderate	high	high
Europe	high	high	moderate
Japan	high	high	low

In Europe and Japan, tire grip in wet and dry conditions (safety) and low rolling resistance with the associated reduction in gasoline consumption (environmental protection), have the highest priority. In North America, on the other hand, low abrasion values and low rolling resistance are of particular significance.

Since the 1970s the performance of tires has improved dramatically: The rolling resistance has been reduced by 50%, the life of an average passenger car tire has nearly doubled. In the USA, an average tire lasted about 24 000 miles in the early 1970s; by 2000 that figure had become 43 000 miles. Regional differences are also interesting: in Europe, the average life of a passenger car tire is only between 20 000 and 25 000 miles [1258], [1259].

This development was only possible due to improvements in several areas, such as tire design, computer-aided construction, new mixing concepts, alternative reinforcement materials and filler systems, addition of rubber chemicals for optimized faster curing and better protection of the rubber, and, last but not least, optimized rubber grades.

A summary of the situation in the early 1980s is given in [840]. Since the 1980s, various attempts have been made to provide a better balance of these properties such as: Vinyl-BR and vinyl-IR [841-843], S-SBR and S-BR with optimized micro- and macrostructure [844-851], [1261, 1262], Multiblock S-SBR and S-SIBR (“integral rubber”) [843], [852-867], [1262-1265] CR [840], [868], Epoxidized NR [869], End-group-modified S-SBR and BR [870-884, 1266], Rare-earth-based high-*cis* 1,4-BR [885-898], Blends (rubber–filler, rubber–fiber, rubber–thermoplast, and others) [899-923], Rubber–rubber blends [868], [883], [901], [924-945], Low molecular mass compounding ingredients and modified rubber [910], [918], [946-964], HNBR [965-969], Chemically modified carbon black [970-972, 1267], Resin modification [920], [954], Silicate filler in combination with a filler activator [973].

The energy consumption of the tire tread accounts for nearly two thirds of the rolling resistance of a tire [1268]. In the car tire tread compounds, that are used particularly in North America, carbon black is used as the sole filler. Hysteresis is reduced by using modified S-SBR grades which, by coupling with tin compounds or by end-group functionalization, are able to interact better with the filler than unmodified grades of rubber.

This “linkage” of the polymer to the filler restricts chain mobility, leading to reduced hysteresis and lower rolling resistance [1269]. One disadvantage of this approach lies in the poor wet grip characteristics of the tire tread compounds.

In Europe, an alternative approach to formulation has been established in the late 1990s: by using silica as a filler [1270-1274] it is possible to achieve a simultaneous improvement in rolling resistance and wet grip characteristics.

Introduction of a chemical linkage between the filler and the rubber via a silane group also reduces hysteresis [1275, 1276]. It has proved possible to overcome initial deficiencies in the third component of the “magic triangle”, namely abrasion, by incorporating high-cis-polybutadiene, such as Nd-BR (produced using neodymium catalysts) or Co-BR (cobalt catalyst) into the rubber compound.

A new development in *solution rubbers* involves modification of the entire polymer chain. In conventional polymers, which are used in the aforementioned compounds with carbon black fillers, only the chain ends are modified. Additional functionality may also be incorporated into the middle of the polymer chain by means of a coupling agent (e.g., tin tetrachloride). New research results are indicating the way to innovative new tire rubbers, in which the rubber molecule is modified with functional groups over the entire length of the chain. With this type of functionalization, it is possible to strengthen the interaction between filler and rubber [1277, 1278].

The polymerization processes of E-SBR and S-SBR are different. Coupling and end group modification cannot be carried out in the radical polymerization to form E-SBR. Therefore, other approaches are necessary to introduce functional groups into the E-SBR [1279, 1280]. By using monomers with polar side groups, it is also possible to obtain functionalized polymers in the emulsion process. Using acrylonitrile as a termonomer in addition to butadiene and styrene, a terpolymer, NSBR, can be obtained. In production of NSBR the monomers are polymerized statistically. Usually the ACN content is up to 20% [1281, 1278]. This NSBR shows very good wet grip properties and can be used both in the footwear industry — especially in sports shoe applications — and in tire treads as a wet grip improving agent.

One development in the tire polymer area is the evaluation of the suitability of different nonaromatic oils, such as MES (medium extracted solvate) or TDAE (treated distillate aromatic extract) instead of the standard aromatic DAE (distillate aromatic extract) oil for rubber-like SBR and BR [1282-1285].

Sidewall Performance. Flex cracking and weather and fatigue resistance are important criteria for tire sidewall performance. Compounds consisting of NR – BR blends provide the state-of-the-art solution to these technical requirements, the composition has not changed much since the 1980s. Nd-BR-based compounds exhibit advantages over other BRs (Li, Ti, Ni, Co) [974].

Inner Liners. Increasing performance requirements for inner liners refer primarily to ease of processing, improvement in covulcanization, impermeability, heat and dynamic resistance, and low weight. These requirements are met by a preferred usage of BIIR to CIIR. BIIR-related improvements refer to products with broader molar mass distributions and to brominated rubber grades which are based on isobutene – para-alkylstyrene feedstocks. [1286, 1287-1288], [985-999].

8.3. Production Processes

The present status of SR production and trends in technology are elucidated by review of the patent literature (see Tables 69 and 70). In Table 69 the patent literature is ordered according to chemical and technological features of the key reactions applied to SR production. In Table 70 the literature is categorized according to references that describe improvements of process steps.

Table 69. Literature on synthetic rubber production processes relating to chemical and technological features of key reactions

Chemical mechanism	Type of process for key reaction				
	Emulsion	Solution	Dispersion (slurry)	Bulk	Gas phase
Radical polymerization	[1012], [1013], [1081], [1084], [1117-1125]	[1126-1129]		[1180-1182]	
Ziegler – Natta polymerization	[976]	[957], [958], [967], [969], [970], [1073-1075], [1085],[1134-1141]	[1083]	[959-962], [964], [965], [968]	[1188-1191]
Anionic polymerization		[1080]			
Cationic polymerization		[1029]	[1170], [1171]		

Table 70. Process improvements relating to sequence of process steps for various types of reactions

Process step	Type of reaction			
	Emulsion	Solution	Dispersion (slurry)	Bulk
Makeup of solutions and other components	[1117], [1118]			
Polymerization and polymer modification reaction	[976], [1012], [1013], [1118-1125], [1192]	[966], [967], [969], [970], [1073-1075], [1080], [1126-1129], [1131-1133], [1152-1165], [1170]		[959-965], [968], [1180-1187], [1205-1207]
Short stopping	[1081], [1084]			
Flashing, degassing, and stripping of unreacted monomers, solvents, and other components	[1193]	[1199]	[1175], [1176]	
Postmodification reactions (blending with oil, carbon black, polymers, fibers, etc.)	[972], [993-995], [1009], [1010], [1058], [1072], [1194]		[1177]	
Polymer isolation, cleaning, and drying	[1195], [1196]	[1166-1169], [1200-1202]	[1203], [1204]	
Wastewater treatment	[1197], [1198]			

The patent literature focuses on both well-established and non-state-of-the-art (not yet established) SR technology. *Well-established processes* include radical polymerization in emulsion and solution, as well as polymer modification in solution and dispersion. These processes are reinvestigated for the following reasons:

1. Reduction of environmental impact
2. Improvement of process control and product consistency
3. Reduction of manufacturing costs
4. Production of new products by available technology (in existing plants)

The patent literature dealing with *non-state-of-the-art SR processes* aims at the adoption of readily available technology for the production of close matches for products whose properties are known. Most investigations are in the field of bulk and gas-phase polymerization.

High-pressure bulk polymerizaton using radical initiators is a standard process for production of low density polyethylene. This process has been applied to the copolymerization of ethylene and acrylonitrile [1205], [1206]. It is presently being reinvestigated in order to produce a match for HNBR [1180-1182], [1207] (see Section Hydrogenated Nitrile Rubber).

Recent interest in the bulk polymerization of butadiene by using *Ziegler – Natta catalysts* is a result of the availability of special reactors and captive and market requirements for high-*cis* BR. Ni- [968] and Nd-based [959-966], Ziegler – Natta catalysts are used.

Production of polyethylene, polypropylene, and linear low-density polyethylene by *gas-phase polymerization* is a well-established technology. The obvious problems associated with applying this technology to production of EPDM are the stickiness of the particles formed and the difficult removal of residual ethylidenenorbornene. A U.S.-based firm has apparently tackled these problems and is presently starting up pilot-plant production of EPDM [1208].

A positive impact on SR development can be expected from *Kaminsky – Brintzinger – Sinn-type metallocene alumoxane catalysts* [1130]. Progress in the technological properties of EPDM is reflected in the recent patent literature [1134-1141], [1132], [1133] and papers [1142-1149].

Another breakthrough has supposedly been achieved with the *polymerization of polar monomers* using Ziegler – Natta catalysts. This opens the possibility of SR synthesis by copolymerization of olefins and monomers such as acrylonitrile [1131], methyl acrylate [1150], and vinyl acetate [1151], which have thus far been considered poisons of the active sites of Ziegler – Natta catalysts.

Thus, sufficient know-how exists for a technological “quantum jump” in SR production. Since the SR manufacturing industry faces sluggish market growth and tough competition, investments will be focused on refurbishing existing plants rather than establishing a completely new technology that involves high risks, unless a significant reduction in manufacturing costs can be expected.

8.4. Ecological and Health Aspects

8.4.1. *Introduction*

Environmental and health aspects of rubber extend over the whole life cycle of a rubber part, starting from raw materials and covering all stages of storage, transport, manufacturing, and use. At the end of its life, recycling of the rubber article or its disposal still remains an issue to be solved. Chemical legislation, safety regulations, directives and guidelines apply for each single stage of the article's life, as illustrated by the following list, which reflects the legal situation in Europe, and which is far from being comprehensive. Other regions apply similar legal tools:

Chemicals. Dangerous Substances Directive
Emissions. Integrated Pollution Prevention and Control (IPPC), Ambient Air Quality Management
Water/ Wastewater. Water Framework Directive
Plant safety. Prevention of Industrial Risks Directive (“Seveso Directive”)
Food contact. Framework Directive 89/109/EEC
Waste. Hazardous Waste Directive; Recycling

Special regulations deal with nitrosamines, volatile organic compounds (VOC), transport, etc. Hence, although the risks regarding production and use of rubber parts seem to be manageable, the legal framework necessary to ensure product safety is somewhat complicated and often incoherent.

8.4.2. *Health Aspects*

8.4.2.1. *Raw Materials*

Raw materials for rubber include polymers, rubber additives, processing aids, extenders/plasticizers, fillers, and reinforcing materials, such as steel cord or fibers. Generally, the *polymers* (SR or NR) do not pose health problems within the rubber industry, because trace amounts of residual monomers, e.g., butadiene, styrene, acrylonitrile, or chloroprene are removed during rubber processing. Concerns of sensitive consumers such as food contact, are regulated by laws or directives in some countries. However, some SR *additives or plasticizers* such as aromatic extender oils or phthalates merit special attention. Aromatic extender oils are classified as carcinogens, and some phthalates are teratogenic. No exposure, hence no risk is to be expected during normal handling of extended rubbers, because it has been demonstrated that the extenders/plasticizers are occluded in the polymer matrix and are not bioavailable. In addition, some *rubber chemicals* pose potential health and environmental hazards. Many accelerators and antioxidants are classified and labeled as sensitizers and/or hazardous to the environment. Production, transport, storage, and use of these chemicals are regulated by several directives, hence no risk should be associated with rubber additives when handled under specified conditions. Exposure to chemicals can be minimized by using predispersed or coated products [1290]. One aspect however, which is closely related to the use of accelerators, is the occurrence of nitrosamines. Some of these are proven and potent carcinogens to animals, according to test results. Nitrosamines are generated wherever secondary aliphatic amines are involved (⟶ Rubber, 4. Chemicals and Additives – Toxicology and Ecology of Organic Vulcanization Accelerators). In order to protect workers and consumers, a specific and severe legal framework applies. Guidelines in Germany set airborne limits for carcinogenic nitrosamines [1291], [1291]. Nitrosamine migration from baby teats and soothers is regulated on EU level [1292]. Since for most applications less dangerous accelerators are available, the nitrosamine problem appears to be manageable [1293].

8.4.2.2. *Processing*

Unit operations of rubber processing include raw materials storage, mixing/blending, shaping, vulcanization, and storage of finished articles. Along the whole process, exposure of workers to hazardous substances may occur. Many regulations for industrial hygiene apply, so that in case of occupational exposure countermeasures can be taken. An epidemiological retrospective study on German rubber workers revealed an increased risk of developing cancer [1294], which may have been caused by exposure to aromatic amines, asbestos, solvents, nitrosamines, or curing fumes. Curing fumes have been studied recently [1295]. They consist basically of a great number of organic substances, depending on vulcanization/curing conditions (time, temperature, etc.). Due to the complex nature of the emissions, a risk evaluation is difficult. Different investigations have shown that exposure to curing fumes may slightly increase the health hazard. The findings usually cover a long period of time, thus reflecting the earlier situation when the operating conditions were less developed than today. It can be stated, however, that workplace hygiene has very much improved since the 1980s.

8.4.2.3. *Use*

Consumer contact mainly occurs indirectly, e.g., via food contact. In many countries, food contact regulations exist, which regulate, e.g., compound ingredients or use conditions. Possible migration that might occur between rubber articles and food is strictly limited. Migration of antioxidants into food can be minimized by chemical attachment of the antioxidant to the polymer [1296]. The major use of commercialized rubber parts is in automotive or technical applications where no significant exposure occurs.

Latex protein allergy as a problem for the rubber industry became important with increasing use of gloves and other articles made from NR latex. The allergic reaction is due to a small fraction of residual soluble (extractable) proteins, which may become respirable by powdering the gloves [1297]. It is technically feasible to reduce the content of proteins in latex, so that the protein allergy can be managed. (An alternative is to use synthetic (e.g. XNBR) rubber latex). Skin contact allergy may also be caused by rubber chemicals (accelerators, antioxidants), which are frequently classified as sensitizers.

8.4.3. *Environment*

The production of polymers and rubber additives is subject to emission regulations. Likewise, transport and processing has little impact on the environment. In comparison with other sources, e.g., the traffic, the emission of curing fumes, for instance, contributes only to a very small extent to the pollution in the urban environment [1295]. However, the fate of material resulting from tire abrasion and from waste tires is becoming increasingly important in environmental policy and public interest. In 1998, a comprehensive study of the impact of tire debris was published [1298]. It was shown that rubber chemicals and their metabolites as well as polycyclic aromatic hydrocarbons (PAH) and zinc oxide were detectable in highway run-off water. The concentration of these substances decreased rapidly with increasing distance from the highway. It was concluded that photolytic and oxidative degradation of small abraded particles prevents accumulation of tire additives. Substitution of aromatic oils by safe process oils in tire treads was recommended in order to reduce the environmental impact of tire debris. In 2001 a comprehensive picture of the environmental impact of tires was gained by a life cycle assessment (LCA), where all aspects of a tire were assessed, including energy consumption, raw materials, emissions, and waste disposal [1299]. According to ISO 14040, this was achieved by:

- Compiling an inventory of relevant inputs and outputs of a product system;
- Evaluating the potential environmental impacts associated with those inputs and outputs;

– Interpreting the results of the inventory analysis and the impact assessment in relation to goal and scope;

Based on the study, the following conclusions can be drawn:

The use phase has the highest contribution to the environmental load in the life cycle of a car tire

The most important environmental aspect of car traffic is the fuel consumption that increases with increasing rolling resistance of the tires. Tire debris has a significant contribution to the overall life cycle, but its role is modest in comparison with the fuel consumption

The environmental load in the production phase of a tire is mainly determined by the production of raw materials, and not by the tire manufacturing itself

The average end-of-life scenario has a relatively low contribution to the overall load of the life cycle. Similarly, distribution transport, and end-of-life collection transport has little significance

It is not possible to determine the environmental load of retreaded tires precisely.

8.4.3.1. *Recycling.*

Discarding of rubber products, particularly used tires has caused severe environmental problems in the past. Used tires account for ca. 60% of total rubber waste in Europe. Many attempts to recycle waste tires have been undertaken. Retreading is a proven technology to extend the use phase of tires, but the final problem of end-of-life disposal remains controversial. Several technologies were developed which would allow for recycling of the vulcanized rubber [1300]; e.g., reclaiming [1301], [1302], grinding [1303], and finally incineration in cement kilns which turned out to be the most practical option. In Germany, about 550 000 t of used tires must be managed annually. About 55% are incinerated with energy recovery, retreading accounts for another 30%, and about 10 % are reused after grinding or reclaiming in less demanding applications, e.g., sport ground surfaces. However, toxic eluates may migrate from such surfaces [1304]. In the USA, the major part of approximately 250 × 10^6 scrap tires per year are utilized for tire-derived fuel [1305].

8.4.4. ***Outlook***

In the near future, a more severe regulation of chemical substances along the value chain is expected (“White Paper”). Polymers are less concerned than rubber chemicals and additives, including accelerators, antioxidants, processing aids, etc. The EU White Paper is a starting point for a comprehensive approach to a European substances legislation with the ultimate target to manage all risks to humans and environment regarding the production, use, and discharge of substances, that are produced or imported at volumes > 1 t/a [1306]. The White Paper will result in a regulation, with the REACH process as basis (**R**egistration, **E**valuation, **A**uthorization of **Ch**emicals), in order to assess and manage the risks of a substance through the value chain, which implies that downstream users, e.g. the rubber industry, are involved. The new regulation will replace about 40 different pieces of EC legislation which regulate chemical substances, as well as an amendment to the classification and labeling Directive. Thus, the rubber industry will be affected by some particular features and issues. However, disposal of worn tires remains an issue that requires an economically and ecologically compatible solution. The environmental impact of tires can be improved by substituting aromatic extender oils by safe process oils [1307]. A great progress was made by the development of “green” tires, which significantly reduce fuel consumption without sacrificing safety standard or longevity of the tires [1308].

[Top of Page]

End Notes

1) TPEs not included, only thermoset rubbers

References

1. *Worldwide Rubber Statistics 1996*, International Institute of Synthetic Rubber Producers, Houston 1996.
2. H. Bertram, *Bayer Mitt. f. d. Gummi-Ind.*, (1981) no. 53 3 – 11.
3. M. Morton: “Synthetic Rubber Past, Present and Future,” *Rubber World* **201** (1989) no. 1, 55 – 66.
4. DIN/ISO 1629, Rubber and Latices; Nomenclature, 2nd ed., 1487–12–01.
5. K. Dinges in H. Batzer (ed.): “Kautschuk und Gummi,” *Polymere Werkstoffe, Technologie 2*, vol. **3**, "chap. 2", Georg Thieme Verlag Stuttgart 1984, 330 – 392.
6. R. Casper, E. Rohde: “Zunehmende Bedeutung der Spezialkautschuke in Westeuropa,” *Kautsch. Gummi Kunstst.* **41** (1988) no. 6, 541 – 548.
7. T. Timm: “Die physikalischen Leistungsgrenzen von Hochelastomeren,” *Kautsch. Gummi Kunstst.* **39** (1986) no. 1, 15 – 25.
8. G. P. Streit: “Der Einsatz hochtemperaturbeständiger Elastomere in der Dichtungstechnik,” *Kautsch. Gummi Kunstst.* **41** (1988) no. 9, 908 – 915.
9. H. G. Drössler: “Neuere Materialien – eine Bestandsaufnahme,” *Kautsch. Gummi Kunstst.* **40** (1987) no. 1, 28 – 32.
10. C. M. Blow, C. Hepburn: *Rubber Technology and Manufacture*, 2nd ed., Plastics and Rubber Institute by Butterworth Scientific, London 1982.
11. W. Hofmann: “Technische Elastomere,” *Kunststoffe* **77** (1987) no. 10, 1057 – 1064.
12. J. A. Brydson: *Rubbery Materials and Their Compounds*, Elsevier Applied Science, London 1988, pp. 428 – 432.
13. L. P. Smith, *Plast. Rubber Int.* **10** (1985) no. 3, 31 –34.
14. ASTM Designation D 2000–90 Standard Classification System for Rubber Products in Automotive Applications – Appendix X 1 Type and Class of Polymer Usage, Philadelphia, 1992; pp. 217 – 246.
15. Classification System For Rubber Materials, SAE J 200, Jan. 1990.
16. unpublished results.
17. J. G. Pillow, “*New EPDM Materials as Produced by Constrained Geometry Catalyst Chemistry Compared with Current Commercial EPDM*,” Du Pont Dow Elastomers SA, European Technical Centre, Geneva.
18. IG Farben, DE 570.230, 1933 (E. Tschunkur, W. Bock).
19. R. F. Dunbrook in G. S. Whitby (ed.): *Synthetic Rubber*, Wiley Interscience, New York 1954, p. 32.
20. R. L. Bebb, L. B. Wakefield in G. S. Whitby (ed.): *Synthetic Rubber*, Wiley Interscience, New York 1954, p. 937.
21. E. B. Storey, *Rubber Chem. Technol.* **34** (1961) 1402 – 1484. Links
22. H. J. J. Janssen, K. V. Weinstock, *Rubber Chem. Technol.* **34** (1961) 1485 – 1500. Links
23. J. F. Auchter: “Styrene – Butadiene Elastomers,” *Chemical Economics Handbook*, SRI International, Menlo Park 1991.
24. *Worldwide Rubber Statistics 1996*, International Institute of Synthetic Rubber Producers, Houston 1996.
25. M. Dollhausen, *Adhäsion* **26** (1982) 23 – 25.
26. K. H. Nordsiek, *Kautsch. Gummi Kunstst.* **39** (1986) 559.
27. *The Synthetic Rubber Manual*, 13th ed., International Institute of Synthetic Rubber Producers, Houston 2002.
28. P. T. Stricharczuk, D. E. Wright in I. Skeist (ed.): *Handbook of Adhesives*, 2nd ed., Van Nostrand Reinhold, New York 1977, p. 293.
29. Kumho KR, US 043 964, 2002.
30. Nippon Zeon, EP 0806452; EP 0819731, EP 0870786.
31. Goodyear, EP 0926 192, EP 1 219 676, US 4 574 140.
32. Michelin, WO 01/9090237.
33. JSR, EP 1 184 415.
34. Bayer, WO 02/20632, DE 100 60 222.

35. Goodyear, US 5 583 173.
36. E. T. McDonnel, K. C. Baranwal, J. C. Andries in D. R. Paul, S. Newman (eds.): *Polymer Blends*, vol **2**, Academic Press, New York 1978, p. 263.
37. F. A. Bovey, I. M. Kolthoff, A. I. Medalia, E. J. Meehan: *Emulsion Polymerisation*, Interscience, New York 1955.
38. C. A. Uraneck in J. P. Kennedy, E. G. M. Tornqvist (eds.): *Polymer Chemistry of Synthetic Elastomers*, part 1, chap. 4A, Interscience, New York 1968.
39. C. A. Uraneck, *Rubber Chem. Technol.* **49** (1976) 610 – 649. Links
40. P. J. Flory: *Principles of Polymer Chemistry*, 10th ed., Cornell University Press, London 1978.
41. G. M. Burnett, G. G. Cameron, P. L. Thorat, *J. Polym. Sci. Polym. Chem. Ed.* **8** (1970) 3435 –3442.
42. P. R. Johnson, *Rubber Chem. Technol.* **49** (1976) 650 – 702. Links
43. C. A. Stewart, T. Takeshita, M. L. Coleman, *Encycl. Polym. Sci. Eng.* **3** (1985) 441 – 462.
44. *Houben-Weyl*, 4th ed., **E 20**, 842 – 859.
45. P. Petiaud, Q. T. Pham, *J. Polym. Sci. Polym. Chem. Ed.* **23** (1985) 1333 – 1360.
46. E. Rohde, H. Bechen, M. Mezger, *Kautsch. Gummi Kunstst.* **42** (1989) 1121 – 1129.
47. R. Musch, U. Eisele: "New Polychloroprene Grades with Optimized Structure-Property Relationship," *Meeting of the Rubber Division, American Chemical Society*, Detroit, Oct. 17 – 20, 1989.
48. M. C. Coleman, R. E. Fuller, *J. Macromol. Sci. Phys.* **B 11** (1975) 419 – 439.
49. International Institute of Synthetic Rubber Producers, Inc. (IISPR): *Worldwide Rubber Statistics 1991, Rubber Handbook,* 8th ed., Swedish Institution of Rubber Technology (SFG), 1990.
50. W. Hofmann, *Rubber Chem. Technol.* **36** (1963) 1.
51. J. R. Dunn, D. C. Coulthard, H. A. Pfisterer, *Rubber Chem. Technol.* **51** (1978) 389. Links
52. *Kirk-Othmer*, 3rd ed., **8** (1993).
53. H. H. Bertram, *Rubber Technol.* **2** (1981) 51.
54. J. R. Dunn, *Kautsch. Gummi Kunstst.* **36** (1983) 966.
55. W. Hofmann in S. H. Morell (ed.), *Prog. Rubber Technol.* **46** (1984) 43.
56. P. W. Milner, *Rubber Technol.* **4** (1987) 57.
57. Z. Szentivanyi, H. Magg, F. Leibbrandt, J. Wassen, *Kautsch. Gummi Kunstst.* **41** (1988) 1118.
58. A. H. Jorgensen, L. A. Chandler, E. A. Collins, *Rubber Chem. Technol.* **46** (1973) 1087. Links
59. V. R. Landi, *Rubber Chem. Technol.* **45** (1972) 222. Links
60. W. Obrecht et al. *Angew. Makromol. Chem.* **145/146** (1986) 161.
61. J. T. Byers, N. L. Hewitt, J. P. Tultz, *Gummi Fasern Kunstst.* **42** (1989) 263 – 268, 436 – 446.
62. J. R. Dunn, *Plast. Rubber Process. Appl.* **2** (1982) 161.
63. W. Hofmann, *Plast. Rubber Process. Appl.* **5** (1985) 209.
64. J. R. Dunn, H. A. Pfisterer, *J. Elastomers Plast.* **9** (1977) 193.
65. H. P. Brown, *Rubber Chem. Technol.* **30** (1957) 1347. Links
66. R. J. Weir, R. C. Burkey, *Rubber and Plastics News* (1981) 32.
67. Polysar, EP 156 475, 1988 (K. Sato).
68. M. B. Jerman, Z. Janovic, *J. Macromol. Sci. Chem.* **A 21** (1984) 887.
69. S. K. Chakraborty, S. K. De, *J. Appl. Polym. Sci.* **27** (1982) 4561. Links
70. J. W. Horvath, J. R. Purdon, G. E. Meyer, F. J. Naples, *Appl. Polym. Symp.* **25** (1974) 187.
71. J. A. Kuczkowski, J. G. Gillick, *Rubber Chem. Technol.* **57** (1984) 621. Links
72. D. Miller, M. Gozdiff, F. Shaheen, P. Dean, *Rubber World* **200** (1989) 13.
73. W. J. Abrams, *Rubber Age (N. Y.)* 91 (1962) 255.
74. J. P. Lehnen, *Eur. Rubber J.* **164** (1982) 53.
75. L. P. Smith, *Elastomerics* **110** (1978) 32.
76. B. F. Goodrich, EP 39 065, 1981 (J. C. McCarty).
77. J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook*, 2nd ed., Wiley-Interscience, New York 1975.
78. Polysar, DE-OS 2 952 508, 2 952 526, 1979 (N. Gurak, K. Tebbens).
79. Polysar, DE 3 043 688, 1980 (R. H. Wunder).
80. Bayer, DE-OS 2 425 441, 1974 (H. Schnöring, G. Pampus).
81. B. F. Goodrich, US 4 920 176, 1990 (A. H. Jorgensen).
82. IISRP: *Worldwide Rubber Statistics 1991.*
83. H. de Decker, C. McCall, W. Bahary, *Rubber Plast. Age* **46** (1965) no. 3, 286.
84. *Houben-Weyl*, **E 20**, 798 – 821.
85. K. Dinges, *Kautsch. Gummi Kunstst.* **32** (1979) no. 10, 748 – 756.
86. A. Echte, *Adv. Chem. Ser.* **222** (1989) 15 – 64. Links
87. F. Haaf *et al.*, *J. Sci. Ind. Res.* **40** (1981) 659.
88. L. Morbitzer *et al.*, *J. Appl. Polym. Sci.* **20** (1976) 2691. Links
89. D. Kranz, L. Morbitzer, K. H. Ott, R. Casper, *Angew. Makromol. Chem.* **58/59** (1977) 213 – 226. Links
90. Bayer AG, US 5 071 946, 1991 (A. Schmidt, H. Eichenauer, K. H. Ott, A. Pischtschan).
91. M. Morton, P. P. Salatiello, H. Landfield, *J. Polym. Sci.* **8** (1952) no. 1, 111 – 121.
92. M. Morton, P. P. Salatiello, H. Landfield, *J. Polym. Sci.* **8** (1952) no. 2, 215 – 224.
93. S. Deibert, F. Bandermann, J. Schweer, J. Sarnecki, *Makromol. Chem. Rapid Comm.* **13** (1992) 351 –355.
94. P. A. Weerts, J. L. M. von der Loos, A. L. German, *Makromol. Chem.* **190** (1989) 777 – 788.
95. P. A. Weerts, J. L. M. von der Loos, A. L. German, *Makromol. Chem.* **191** (1990) 2615.
96. P. A. Weerts, A. L. German, R. G. Gilbert, *Macromolecules* **24** (1991) 1622. Links
97. I. A. Maxwell, B. R. Morrison, D. H. Napper, R. G. Gilbert, *Macromol. Chem.* **193** (1992) 303 –313.
98. U. Pallaske, A. Schmidt, K. H. Ott, in K. H. Reichert, W. Geister (eds.): *Polymer Reaction Engineering*, Hüthig und Wepf Verlag, Stuttgart 1986, pp. 19 – 39.
99. L. Talalay, *Rubber Chem. Technol.* **36** (1963) 581. Links
100. H. A. Kurzhals, H. Reuter, *Chem Ing. Tech.* **51** (1979) no. 4, 325.
101. W. J. Lecluse, *Chem. Ing. Tech.* **52** (1980) no. 8, 668 – 669.
102. Doverstrand Ltd., US 3 795 647, 1972 (B. A. Ripley-Duggan).
103. The International Synthetic Rubber Co Ltd., DE-OS 2 101 650, 1970 (D. A. Bennett).

104. Monsanto Co., US 3 988 010, 1976 (M. D. Paster).
105. Monsanto Co., US 4 043 955, 1975 (M. D. Paster).
106. Hüls AG, DE-AS 1 770 934, 1968 (H. Schlüter).
107. Hüls AG, DE 1 213 984, 1963 (H. Schlüter).
108. H. G. Keppler, H. Wesslau, J. Stabenow, *Angew. Makromol. Chem.* **2** (1968) 1. Links
109. The DOW Chemical Co., EP-A 143 858, 1983 (D. E. Henton, T. M. O'Brien).
110. R. W. Strassburg in A. K. Bhowmick, H. L. Stephens (eds.): *Handbook of Elastomers, New Developments and Technology*, "chap. 18", Dekker, New York 1988.
111. D. A. Seil, F. R. Wolf in : M. Morton (ed.): *Rubber Technology*, 3rd ed., "chap. 11", Van Nostrand Reinhold, New York 1987.
112. R. D. DeMarco, *Rubber Chem. Technol.* **52** (1979) 173. Links
113. T. Brewster, G. Cantoni, *Rubber and Plastics News* **9** (1987) March, 18.
114. *Jpn. Chem. Week* **32** (1991) Dec., 5.
115. H. Enyo, S. Zen, Y. Takemura, *Proc. Int. Rubber Conf.* Kyoto, Japan, Oct. 15 – 18, 1985, p. 790.
116. K. Umetsu *et al.*, Paper no. 54, Rubber Division, American Chemical Society, Cleveland, Ohio, Oct. 17 – 20, 1995.
117. T. Ikumoto, M. Sugimoto, T. Kondo, *Nippon Gomu Kyokaishi* **60** (1987) no. 8, 469.
118. R. Montague, *Rubber World* **199** (1988) Dec., 20.
119. E. Chang, E. Mazzone, *Rubber and Plastic News, Technical Yearbook* (1989) 99.
120. K. Hosoya, T. Nakagawa, S. Yagishita, SAE Paper no. 880 028, SAE International, Warrendale.
121. A. L. Logothetis, *Prog. Polym. Sci.* **14** (1989) 251 –296. Links
122. A. L. Logothetis, *Fluorpolymers Conference*, "Paper no. 3", Manchester, Jan. 6 – 8, 1992.
123. D. Cook, M. Lynn, *Rapra Rev. Rep.* **3** (1990) 32–1 –32–112.
124. M. Pianca *et al.*, *Polymer* **28** (1987) 224 – 230. Links
125. W. Hofman, *Kunststoffe* **80** (1990) 1204 – 1209.
126. G. Moggi, P. Bornadelli, C. J. Bart, *Polym. Bull. (Berlin)* **7** (1982) 115 – 122.
127. G. Arjoldi, M. Pianca, M. Fumagalli, G. Moggi, *Polymer* **30** (1989) 2180 – 2187.
128. W. W. Schmiegel, *Angew. Makromol. Chem.* **76/77** (1979) 39 – 65. Links
129. G. Streit, S. Dunse, *Kautsch. Gummi Kunstst.* **38** (1985) 471 – 479.
130. Y. Tabata, K. Ishigure, H. Sobue, *Kogyo Kagaku Zasshi* **65** (1962) 1626 – 1629.
131. Y. Tabata, K. Ishigure, H. Sobue, *J. Polym. Sci. Part A 2*, **2** (1964) 2235 – 2243.
132. G. Kojima, H. Kojima, Y. Tabata, *Rubber Chem. Technol.* **50** (1977) 403 – 412. Links
133. Du Pont, US 3 467 635, 1966 (W. R. Brasen, C. S. Cleaver).
134. D. E. Hull, *Kautsch. Gummi Kunstst.* **38** (1985) 480 – 484.
135. D. E. Hull, *Elastomerics* **122** (1990) 11 – 16.
136. W. M. Grootaert, R. E. Kolb, A. T. Worm, *Rubber Chem. Technol.* **63** (1990) 516 – 522. Links
137. Asahi Glas, US 4 758 618, 1986 (I. Yasumichi, H. Wachi).
138. W. W. Schmiegel, *Kautsch. Gummi Kunstst.* **31** (1978) 137 – 148.
139. P. Venkateswarlu, R. E. Kolb, R. A. Guenther, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **31** (1990) no. 1, 360 – 361.
140. D. Apotheker, B. J. Finlay, P. J. Krusic, A. L. Logothetis, *Rubber Chem. Technol.* **55** (1982) 1004 – 1018. Links
141. Du Pont, US 4 035 565, 1975 (D. Apotheker, P. J. Krusic).
142. Du Pont, US 4 214 060, 1976 (D. Apotheker, P. J. Krusic).
143. Daikin, US 4 158 678, 1977 (M. Tatemoto, T. Nagakawa).
144. Daikin, US 4 243 770, 1978 (M. Tatemoto *et al.*)
145. Daikin, US 4 361 678, 1980 (M. Tatemoto, S. Morita).
146. Nippon Mektron, DE 3 710 818, 1987 (H. Tatsu *et al.*)
147. Ausimont, EP 0 407 937, 1990 (M. Albano, G. Brinati, V. Arcella, E. Gianetti).
148. R. E. Uschold, *Polymer. (Tokyo)* **17** (1985) 253 –263.
149. Montefluos, US 4 739 024, 1987 (G. Moggi, A. Caratto, P. Bonardelli, C. Monti).
150. Du Pont, US 4 948 852, 1989 (A. L. Moore).
151. V. Arcella, R. Ferro, M. Albano, A. Minutillo, *Kautsch. Gummi Kunstst.* **44** (1991) 833 – 837.
152. P. Ferrandez, S. Bowers, *Gummi Asbest Kunstst.* **48** (1995) 626 – 633.
153. D. A. Kotz, R. D. Stevens, *Kautsch. Gummi Kunstst.* **48** (1995) 159 – 163.
154. Ausimont, EP 0 518 073, 1991 (V. Arcella *et al.*).
155. Y. Saito, *Kautsch. Gummi Kunstst.* **39** (1986) 30 –32.
156. K. H. Nordsiek, *Kautsch. Gummi Kunstst.* **39** (1986) 599 – 611.
157. G. W. Marwede, B. Stollfuß, A. J. M. Sumner, *Kautsch. Gummi Kunstst.* **46** (1993) 380 – 388.
158. R. Zelinski, C. W. Childers, *Rubber Chem. Technol.* **41** (1968) 161. Links
159. S. Bywater in H. F. Mark (ed.): *Encyclopedia of Polymer Science and Engineering*, 2nd ed., vol. **2**, Wiley, New York 1985, p. 1.
160. R. Bond, G. F. Morton, L. H. Krol, *Polymer* **25** (1984) 132 – 140. Links
161. A. E. Oberster, T. C. Bouton, J. K. Valaitis, *Angew. Makromol. Chem.* **29/30** (1973) 291. Links
162. G. Day, S. Futamara, *Kautsch. Gummi Kunstst.* **40** (1987) 39 – 43.
163. A. Ueda, H. Watanabe, S. Akita, *Int. Rubber Conf.* 1985, 199.
164. B. Kern, S. Futamara: “Solution SBR as Tread Rubbers,” *ACS Rubber Div. 131st Meeting*, Montreal, Quebec, May 1987.
165. D. C. Moore, G. L. Day, *Elastomerics* **117** (1985) no. 3, 22 – 28.
166. H. L. Brantley, G. L. Day, *Kautsch. Gummi Kunstst.* **40** (1987) 122 – 125.
167. United States FDA 21, CFR 172.615.
168. a) United States FDA 21, CRR 176.180. b) BGA Recommendations XX and XXI.
169. United States 29 CFR 1910–1200.
170. F. Tsutsumi, M. Sakakibara, N. Oshima: “Structure and Dynamic Properties of New Solution SBR Coupled with Tin Compound,” *Rubber Division Meeting*, American Chemical Society, Cincinati, Ohio, Oct. 18 – 21 1988.
171. Japan Synthetic Rubber Co., EP 134 909, 1981 (Y. Takeuchi *et al.*).
172. Bridgestone Tire, EP 254 766, 1981 (M. Ogawa, M. Ikegami).
173. K. H. Nordsiek, K. M. Kiepert, *Kautsch. Gummi Kunstst.* **38** (1985) 178 – 185.
174. Hüls AG, EP 256 261, 1986 (C. Herrmann *et al.*).

175. The Goodyear Tire & Rubber Co., EP 263 054, 1986 (A. F. Halasa, B. B. Gross, T. L. Cox, G. F. Balogh).
176. Hüls AG, EP 362 514, 1988 (W. Hellermann *et al.*).
177. Yokohama Rubber, Nippon Zeon, DE 3 635 366, 1986 (Y. Kikuchi *et al.*).
178. Nippon Zeon DE 3 315 525, 1982 (S. Akita *et al.*).
179. Hüls AG, EP 362 515, 1988 (W. Hellermann *et al.*).
180. H. H. Grever, G. Marwede, *Tyre Tech,* Italy, Oct. 1995.
181. Goodyear, EP 483 046, 1991 (A. F. Halasa *et al.*).
182. Goodyear, EP 698 624, 1995 (A. F. Halasa *et al.*).
183. Goodyear, EP 645 423, 1994 (J. Berg *et al.*).
184. M. Hoffmann, G. Pampus, G. Marwede, *Kautsch. Gummi Kunstst.* **22** (1969) 691.
185. A. Echte, *Angew. Makromol. Chem.* **58/59** (1977) 175 – 198. Links
186. BASF, DE 3 345 377, 1983 (A. Echte, H. Hönl, W. Grohn).
187. Shell, US 3 265 765, 1962 (G. Holden, R. Milkovich).
188. G. Holden, N. R. Legge: *Thermoplastic Elastomers Based on Polystyrene-Polydiene Block Copolymers, Thermoplastic Elastomers*, Hanser, München 1987, pp. 47 – 66.
189. L. Fetters, *J. Polym. Sci. Part C* **26** (1969) 1.
190. M. Morton: *Anionic Polymerization: Principles and Practice*, Academic Press, New York 1983.
191. Bridgestone, EP 742 232, 1993 (D. F. Lawson *et al.*).
192. Bridgestone, EP 725 069, 1996 (D. F. Lawson *et al.*).
193. Bridgestone, EP 725 085, 1996 (D. F. Lawson *et al.*).
194. P. Luijk, *Kautsch. Gummi Kunstst.* **34** (1981) 191 – 195.
195. I. Kuntz, *J. Polym. Sci.* **54** (1961) 569.
196. L. Fetters, *Macromolecules* **12** (1979) 344. Links
197. P. Guyot *et al.*, *Polymer* **22** (1981) 1724. Links
198. Firestone, DE-AS 1 300 239, 1964 (N. Keckler); *Chem. Abstr.* **61** (1964) 808.
199. T. C. Buton, S. Futamara, *Rubber Age (N.Y.)* **106** (1974) 33 – 39.
200. H. L. Hsieh, W. H. Glaze, *Rubber Chem. Technol.* **43** (1970) 22. Links
201. T. A. Antkowiak, A. E. Oberster, H. F. Halasa, D. P. Tate, *J. Polym. Sci. Polym. Chem. Ed.* **10** (1972) 1319.
202. Hüls AG, EP 304 589, 1987 (W. Hellermann *et al.*).
203. T. Kawanaka, A. Yosioka, N. Nagata, H. Watanabe: “Analysis of Chain End Modified Rubber Structure,” *ACS Rubber Div. Meeting*, Detroit, Oct. 17 – 20, 1989.
204. Nippon Zeon, EP 747 425, 1995 (M. Nakamura *et al.*).
205. Bridgestone/Firestone, EP 639 587, 1994 (D. F. Lawson *et al.*).
206. IISRP: *Worldwide Rubber Statistics, 1996,* and own capacity estimations.
207. J. A. Brydson: *Rubbery Materials and their Compounds,* Elsevier Applied Science, London 1988.
208. *Ullmann*, 4th ed., **13**, 595.
209. A. Echti in C. K. Riew, *Adv. Chem. Ser.* **222** (1989) 15.
210. H. Fries, B. Stollfuß, *ACS Rubber Division Meeting*, Cincinnati, Oct. 1988, paper no. 8.
211. W. Kern, S. Futamara, *Polymer* **29** (1988) 1801. Links
212. M. Morton: *Anionic Polymerization: Principles and Practice*, Academic Press, New York 1983.
213. Bridgestone, EP 693 505, 1995 (D. F. Lawson *et al.*).
214. Bridgestone, EP 622 381, 1994 (T. Katamura *et al.*).
215. Bridgestone, Japan Synthetic Rubber, EP 609 010, 1994 (Y. Yamakawa *et al.*).
216. Bridgestone/Firestone, EP 493 839, 1991 W. Hergenrother *et al.*).
217. W. Hergenrother *et al.*, *J. Polym. Sci. Part A: Polym Chem* **33** (1995) 143 – 157. Links
218. Y. Kitagawa:“The Industry looks outwards,”*IISRP 37th Annual General Meeting,* Bath, May 5 – 9, 1996.
219. G. W. Gilby, *Dev. Rubber. Technol.* **3** (1981) 101.
220. F. Tsutsumi, M. Sakakibara, N. Oshima, *Rubber Chem. Technol.* **68** (1990) 8.
221. C. A. Sierra, C. Galan, J. M. Gomeo Falou, V. Ruiz Santa Quitera, *Rubber Chem. Technol.* **68** (1995) 259. Links
222. N. Nagita *et al.*, *Rubber Chem. Technol.* **60** (1987) 837.
223. H. L. Hsieh, R. P. Quirk: *Anionic Polymerization Principles and Practical Applications*, Marcel Dekker, New York 1996.
224. R. P. Quirk, S. H. Jang, J. Kim, *Rubber Chem. Technol.* **69** (1996) 444. Links
225. J. E. Puskas, M. Verhelst, P. Collart, J. B. Schmidt, *Kautsch. Gummi Kunstst.* **48** (1995) no. 12, 866.
226. M. Fontanille, A. H. E. Müller, S. Bywater in G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science,* vol. 3, Pergamon Press, Oxford 1989.
227. I. G. Hargis, R. A. Livigni, S. L. Aggarwal in: A. Whelan, K. S. Lee (eds.): *Developments in Rubber Technology,* vol. 4, Elsevier Applied Science, London 1987.
228. W. Cooper in R. N. Haward (ed.): *Developments in Polymerisation-1*, Applied Science, London 1979.
229. *The Synthetic Rubber Manual*, 11th ed., International Institute of Synthetic Rubber Producers, Houston 1989.
230. M. Morton, J. R. Rupert in F. E. Bailey (ed.): *Initiation of Polymerization*, Americal Chemical Society, Washington, D.C., 1983.
231. *Worldwide Rubber Statistics 1991*, International Institute of Synthetic Rubber Producers, Houston 1991.
232. Firestone, US 3 317 918, 1956 (F. C. Foster).
233. E. W. Duck, J. M. Locke in W. M. Saltman (ed.): *The Stereo Rubbers*, Wiley-Interscience, New York 1977.
234. *Worldwide Rubber Statistics 1996*, International Institute of Synthetic Rubber Producers, Houston 1996.
235. E. Rohde: “Ethylen – Vinylacetat Kautschuk, Einsatz und Einsatzmöglichkeiten für technische Gummiartikel,” *DKG-Bezirksgruppentagung Nordrhein-Westfalen,* Bad Honnef, May 7 – 8, 1992.
236. F. P. Reding, J. A. Faucher, R. D. Whitman, *J. Polym. Sci.* **57** (1962) 483.
237. T. M. Dobel, J. R. Harrel: *New Developments in Ethylene – Acrylic Elastomers*, "paper no. 28", Am. Chem. Soc. Rubber Division Meeting, Detroit, Oct. 8 – 11, 1991.
238. H. Fries, B. Stollfuß: *ACS Rubber Division*, Cincinnati, Ohio, Oct. 1988.
239. W. Hoffmann: *Rubber Technology Handbook*, Hanser Verlag, München 1990.
240. H. Fries, B. Stollfuß: “Structure and Properties of Butadiene Rubber,” *ACS Rubber Division*, Cincinnati, Ohio, Oct. 1988, paper no. 8.
241. S. N.Chakravarty, S. K. Mustafi, A. K. Mukherjee, *Rubber News* **28** (1989) July, 33.
242. E. Schoenburg, H. Marsh, S. Walters, W. Saltman, *Rubber Chem. Technol.* **52** (1977) 527.

243. Phillips Petroleum, GB 848 065, 1956.
244. Goodrich Gulf, US 3 117 743, 1963 (S. E. Horne).
245. Goodrich Gulf, BE 543 292, 1955 (S. E. Horne).
246. Bridgestone, US 3 170 907, 1965 (K. Ueda, T. Yoshimoto, J. Hosono, K. Maeda).
247. M. Gippin, *Ind. Eng. Chem. Prod. Res. Dev.* **4** (1965) 160.
248. M. Gippin, *Ind. Eng. Chem. Prod. Res. Dev.* **1** (1962) 32.
249. S. Horne in R. P. Quirk (ed.): *Transition Metal Catalyzed Polymerizations*, vol. **4**, "part B", Harwood Academic Publishers, New York 1983, p. 527.
250. R. P. Burford, *J. Macromol. Sci. Chem.* **A17** (1982) 123.
251. A. K. Bhowmick, H. L. Stevens, *Handbook of Elastomers, New Developments and Technology*, Dekker, New York 1988.
252. J. Howell, W. Curtis, J. Solomon in R. P. Quirk (ed.): *Transition Metal Catalyzed Polymerizations*, vol. **4**, "part B", Harwood Academic Press, New York 1983, p. 509.
253. Phillips Petroleum, BE 551 851, 1956 (R. P. Zelinski, D. R. Smith).
254. Firestone, US 3 205 213, 1965 (A. Fono, R. S. Stearns).
255. Bayer, GB 974 993, 1964 (N. Schön, G. Pampus, J. Witte).
256. W. M. Saltman, T. Link, *Ind. Eng. Chem. Prod. Res. Dev.* **3** (1964) 199.
257. J. F. Henderson, *J. Polym. Sci. Part C* **4** (1963) 233.
258. T. Yoshimoto *et al.*, *Makromol. Chem.* **139** (1970) 61.
259. T. Yoshimoto *et al.*, *Makromol. Chem.* **139** (1970) 73.
260. M. C. Throckmorton, F. S. Farson, *Rubber Chem. Technol.* **45** (1972) 268. Links
261. W. M. Saltman, L. J. Kuzma, *Rubber Chem. Technol.* **46** (1973) 1055. Links
262. M. C. Throckmorton, *Kautsch. Gummi Kunstst.* **22** (1969) 293.
263. Union Carbide, US 3 297 667, 1967 (W. von Dohlen *et al.*).
264. Q. Jun, W. Fu-Song, H. Bao Tong in R. Quirk (ed.): *Transition Metal Catalyzed Polymerizations*, vol. **4**, "part A", Harwood Academic Press, New York 1983, p. 265.
265. G. Marwede, G. Sylvester, *IISRP, 22nd Annual Meeting Proceedings*, 1981, pp. III-3.
266. J. Witte, *Angew. Makromol. Chem.* **94** (1981) 119. Links
267. Enichem, US 4 444 903 (A. Carbonaro, D. Ferraro, M. Bruzzone).
268. Goodyear – Enichem, US 4 663 405 (M. C. Throckmorton).
269. Y. Takeuchi, A. Sekimoto, M. Abe: "New Industrial Polymers," *ACS Symp. Ser.* (1974) no. 4, 15.
270. Y. Takeuchi, A. Sekimoto, M. Abe: "New Industrial Polymers," *ACS Symp. Ser.* (1974) no. 4, 26.
271. H. Ashitaka, H. Ishikawa, H. Veno, *J. Polym. Sci. Polym. Chem.* **21** (1983) 1853.
272. S. E. Horne: *Transition Metal Catalyzed Polymerizations*, vol. **4**, "part B", Harwood Academic Press, New York 1983, p. 527.
273. W. M. Saltman, *J. Polym. Sci. Part A* **1** (1963) 373.
274. W. Cooper *et al.*: "Elastomer Stereospecific Polymerization," *Adv. Chem. Ser.* **52** (1966) 46.
275. G. J. Van: "Amerongen Elastomer Stereospecific Polymerization," *Adv. Chem. Ser.* **52** (1966) 136.
276. W. Cooper, R. K. Smith, A. Stokes, *J. Polym. Sci. Part B* **4** (1966) 309.
277. E. Lovering, W. B. Wright, *J. Polym. Sci. Polym. Chem. Ed.* **6** (1968) 2221.
278. W. Cooper *et al.*: "Elastomer Stereospecific Polymerization," *Adv. Chem. Ser.* **52** (1966) 46.
279. F. P. Baldwin, G. Ver Strate, *Rubber Chem. Technol.* **45** (1972) 709. Links
280. S. Cesca, *Macromol. Rev.* **10** (1975) 1.
281. G. Ver Strate: "Ethylene-Propylene Elastomers," J. I. Kroschwitz (ed.): *Encycl. Polym. Sci. Eng.,* 2nd ed., vol. 6, Wiley, New York 1986, pp. 522 – 564.
282. *The Synthetic Rubber Manual*, vol. **12** International Institute of Synthetic Rubber Producers, Houston 1992, pp. 47 – 58.
283. P. S. Brown, A. J. Tinker, *J. Natl. Rubb. Res.* **5** (1990) no. 3, 157 – 162.
284. A. Y. Coran, *Rubber Chem. Technol.* **64** (1991) 801 – 812. Links
285. Polysar Ltd., US 4 645 793, 1985 (W. von Hellens, S. A. H. Mohammed, R. Hallman).
286. Bridgestone Corp., US 4 886 850, 1988 (M. Ogawa, Y. Shiomura, T. Takizawa).
287. H. Fries, A. Sumner, *Tyre-Tech. 91*, Berlin, Oct. 1991.
288. J. Boor: *Ziegler Natta Catalysts and Polymerizations*, Academic Press, Inc., New York 1979.
289. Hoechst, EP 384 263, 1989 (D. Volker, M. Antberg, J. Rohrmann, A. Winter).
290. W. Kaminsky, M. Miri, *J. Polym. Sci. Polym. Chem. Ed.* **23** (1985) 2151 – 2164.
291. Exxon Chemical Patents Inc., US 5 001 205, 1988; EP 347 129, 1989 (S. Floyd, E. L. Hoel).
292. A. Gumboldt, J. Hellberg, G. Schleitzer, *Makromol. Chem.* **101** (1966) 229.
293. Hoechst, DE 1 495 698, 1962 (A. Gumboldt, E. Junghans, G. Schleitzer).
294. Chem. Werke Hüls, DE 1 595 442, 1966 (H. Emde, K. Benedikter, P. Hegenberg).
295. Stamicarbon B.V., EP 44 119, 1980 (G. G. Evens, E. M. J. Pijpers, R. H. M. Seevens).
296. Stamicarbon B.V., EP 44 595, 1980 (G. G. Evens, E. M. J. Pijpers, R. H. M. Seevens).
297. B. Schleimer, J. Wolpers, *Werkz. Hüls AG* **161** (1971) 8 – 13.
298. B. Schleimer, J. Wolpers: "Ethylene-Propylene Rubber – Bunawerke Hüls GmbH," *Hydrocarbon Process.* **58** (1979) no. 11, 171.
299. P. Galli, F. Miliani, T. Simonazzi: "New Trends in the Field of Propylene Based Polymers," *International Polymer Symposium*, Kyoto, Japan, Aug. 20 – 25, 1984.
300. *Rubber Plast. News* **15** (1985) no. 4, 138.
301. D. E. James: "Linear Low Density Polyethylene," in H. Mark, C. Overberger (eds.): *Encyclopedia of Polymer Science and Engineering,* vol. 6, Wiley & Sons, New York 1988, pp. 429 – 454.
302. S. P. Sawin, C. J. Baas: "UNIPOL PP: A Gas-Phase Route to Polypropylene," *Chem. Eng.* **92** (1985) May 27, 42.
303. M. R. Rifi et al.; "FLEXOMER TM Polyolefins: Bridging the Gap between Rubber and Plastics," *ACS Rubber Division Meeting,* Las Vegas, June 1990.
304. BP Chemicals Ltd., EP 409 492, 1990 (J. A. Bailly, P. Bres).
305. Union Carbide, US 5 106 927, 1989 (M. R. Riffi, K. H. Lee, M. J. Kriss, H. T. Liu).
306. *Rubber Plast. News* **4** (1991) 5.
307. F. G. Stakem, A. U. Paeglis, J. D. Collins: "Union Carbide Gas Phase EPM and EPDM Rubber," paper presented at *142nd Technical Meeting, Rubber Division,* American Chemical Society, Nashville, Nov. 3 – 6, 1992.
308. IISRP: "Worldwide Rubber Statistics," *Kautsch. Gummi Kunstst.* **45** (1992) 334.
309. *Kirk-Othmer*, 3rd ed., **8**, 470 ff.

310. E. N. Kresge, R. H. Schatz, H-C. Wang: *Encyclopedia of Polymer Science and Engineering*, vol. **8**, Wiley-Interscience, New York, 1987, pp. 423 ff.
311. R. M. Thomas, *Rubber Chem. Technol.* **42** (1969) G9011.
312. J. P. Kennedy: *Cationic Polymerization of Olefins: A Critical Inventory*, Wiley-Interscience, New York 1976, p. 14.
313. L. A. Wood: "Physical Constants of Different Rubbers," in J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook,* 3rd ed., Wiley-Interscience, New York 1989, V/9 – V/10.
314. C. Y. Chu, R. Vukov, *Macromolecules* **18** (1985) 1423. Links
315. J. L. White, T. D. Schaffer, C. J. Ruff, J. P. Cross, *Macromolecules* **28** (1995) 329.
316. Bayer XL 10000 complies with United States FDA Code of Federal Regulations (CFR) Title 21, Part 175.300 (b) (3) (xxxi).
317. D. E. Smith (ed.): *Blue Book 1996*, Lippincott & Peto Inc., Ohio 1996.
318. Bayer Butyl 101-3 complies with United States FDA Code of Federal Regulations (CFR) Title 21, Part 177.1420 Isobutylene polymers, 172.615 Chewing Gum base, 177.2600 Rubber Articles Intended for Repeated Use, and other FDA regulations.
319. I. Kuntz, K. D. Rose, *J. Polym. Sci. Part A: Polym. Chem.* **27** (1989) 107.
320. Standard Oil Development Company, US 2 243 658, 1937 (R. M. Sparks, O. C. Slotterbeck).
321. USSR Ministry, CA 1 019 095, 1973 (N. V. Scherbakova *et al.*). Also issued as DE 104 985.
322. A. Gandini, H. Cheradame, *Adv. Polym. Sci.* **34/35** (1980).
323. M. Roth, H. Mayr, *Macromolecules* **29** (1996) 6104. Links
324. J. P. Kennedy: *Cationic Polymerization of Olefins: A Critical Inventory*, Wiley-Interscience, New York 1976, p. 116.
325. L. Toman, J. Pilar, J. Spevacek, M. Marek, *J. Polym. Sci. Polym. Chem. Ed.* **16** (1978) 2759.
326. F. Barsan, M. C. Baird, *J. Chem. Soc. Chem. Commun* (1995) 1065.
327. K. R. Kumar, A. Penciu, M. J. Drewitt, P. J. McInenly, M. C. Baird, *J. Polym. Sci. Part A Polym. Chem.* **40**, (2002) 3302.
328. IG Farbenindustrie, DE 641 284, 1931 (M. Otto, M. Mueller-Cunradi).
329. A. T. Purcell, J. V. Fusco: *Butyl, The First Fifty Years*, Exxon Corporation, Houston 1987, p. 10.
330. R. Schuster (ed.): "*Handbook of Rubber Technology*", Wiley-VCH Weinheim, Germany, in press.
331. A. A. Berlin, S. A. Wolfson, *Polym. Sci.* **36(4)** (1994) 505.
332. Exxon, US 4 252 710, 1979 (K. W. Powers, R. H. Schatz); US 4 474 924, 1983 (K. W. Powers, H. C. Wang).
333. Bayer AG, US 5,539,065, 1996 (W. Baade, G. Langstein, J. E. Puskas).
334. Esso, US 2 940 960, 1957 (B. R. Tegge *et al.*).
335. International Institute of Synthetic Rubber Producers: World Wide Rubber Statistics, 2001, Houston.
336. *SRI International*, Chemical Economics Handbook, 1999.
337. United States Dept. of Transport, CFR Title 49.
338. B. B. Shugaev, *Arch. Environ. Health* **18** (1969) 879.
339. Michelin, FR 9 102 433, 1991 (R. Rauline). Also issued as EP 501 227.
340. Goodyear, EP 598 193, 1993 (A. F. Halasa *et al.*).
341. N. L. Zutty, J. A. Faucher in *Encycl. Polym. Sci. and Technol.* **6** (1967) 387.
342. D. Mahling, *Kunststoffe* **57** (1967) no. 5, 321.
343. H. Bartl, *Kautsch. Gummi Kunstst.* **25** (1972) no. 10, 452.
344. I. O. Salyer, A. S. Kenyon, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 3083.
345. Bayer AG, Levapren, company brochure KA 30991, Leverkusen, Aug. 1984.
346. J. F. Hagman, J. W. Crary in *Encyclopedia of Polymer Science and Engineering*, Wiley and Sons, New York 1985, pp. 325 – 334.
347. L. E. Nielsen, *J. Polym. Sci.* **42** (1960) 357.
348. B. Vollmert: *Grundriss der Makromolekularen Chemie IV,* E. Vollmert-Verlag, Karlsruhe 1979, pp.153 – 157.
349. R. F. Bayer, *Rubber Chem. Technol.* **36** (1963) 1303.
350. K. Fischer, *Textilveredelung* **19** (1984) no. 5, 146.
351. E. Vidal Russell, N. E. Israeloff, *Los Alamos Natl. Lab., Prepr. Arch., Condens. Matter* (2000), 1 – 14, arXiv: cond-mat/0012245.
352. C. Schick, M. Merzlyakov, A. Hensel, *J. Chem. Phys.* **111** (1999) no. 6, 2695 – 2700. Links
353. A. Nojiri, K. Hattori, S. Okamoto, *Jap. J. Appl. Phys.* **10** (1971) no. 6, 803.
354. B. Krieg, H. Meyer, *Kunststoffe* **75** (1985) no. 10, 751.
355. W. H. Brinkmann, L. W. S. Damen, *Kautsch. Gummi Kunstst.* **19** (1966) no. 5, 267.
356. H. Bartl, J. Peter, *Kautsch. Gummi Kunstst.* **14** (1961) WT 23 – 32.
357. F. Szöcs, O. Rostasowa, *J. Appl. Polym. Sci.* **8** (1974) 2529.
358. Farbenfabriken Bayer AG, DE-AS 1 169 126, 1958 (H. Bartl, J. Peter).
359. Bayer, FR 1 225 704, 1959 (H. Bartl).
360. Farbenfabriken Bayer AG, DE-AS 1 116 394, 1958 (H. Bartl, H. Scheurlen).
361. Du Pont: Polym. Prod. Dep., Elastom. Div., Brochures E-52096–2, E-10770–2, E-08933–2, Genf, 1987.
362. *Winnacker-Küchler*, 4th ed., p. 570.
363. M. Rätzsch, U. Hofmann, *J. Macromol. Sci. Chem.* **A 27** (1990) no. 12, 1489.
364. R. Borngässer, "Neue Lavapren-Typen," *Jugoslawien-Symposium Kautschuk*, Dubrovnik, May 1987.
365. Continental AG, Hannover, DE-OS 3 827 908 (M. Brüggemann).
366. Continental AG, Northeim, Prüfzeugnis LOBA-Nr. 18.43.21–55–110, 20.10.1987.
367. Planatolwerk Willy Hesselmann Chemische und Maschinenfabrik für Klebetechnik GmbH & Co. KG, EP 0 203 251, 1986 (H. Koller).
368. "The Hot Melts 1 and 2," *Mod. Packag.* **38** (1964) no. 2, 113.
369. R. Parg: "*Levamelt^{TM}; A New Class of Adhesive Ethylene-Vinyl Acetate Copolymer for Pressure Sensitive Adhesive (PSA) and coextruded Adhesive Layer Applications*", Dubai Plast, 23-25.04.2001, VI b.
370. R. Parg: "*New Strechable highly Oxygen Impermeable Ultra Thin Films made of EVOH-EVM Levamelt^{TM} Blends*" *Speciality Plastic Films 2001*, Düsseldorf, Germany , 22-24 October 2001.
371. W. Göbel, H. Bartl, D. Hardt, A. Reischl, *Kunststoffe* **55** (1965) no. 5, 329.
372. H. Bartl, *Österr. Chem. Z.* **85** (1984) no. 5, 105.
373. H. Bartl *et al.*, *J. Appl. Polym. Sci. Appl. Symp.* **36** (1981) 165.
374. H. Alberts, W. Fischer, C. Leuschke, H. Bartl, DE-OS 2 951 800, 1979.
375. Farbenfabriken Bayer AG, FR 1 373 236, 1963 (D. Hardt, H. Bartl).
376. Hoechst, EP 0 493 769, 1991 (W. Reimann, M. Feustel, J. Hobes).
377. Farbenfabriken Bayer AG, DT 1 206 848, 1962 (F. Kassack, R. Resz, H. Bartl, G. Berndt).
378. Farbenfabriken Bayer AG, DE-AS 1 222 887, 1964 (R. Resz, H. Bartl, K. Weber, M. Boll).

379. *Ullmanns* 4th ed., **15**, pp. 589 ff.
380. J. F. Hagman: "Ethylene – Acrylic Elastomers," *Conference of the Rubber Division of the Am. Chem. Soc.*, Detroit 1975.
381. J. R. Pailthorp, *Elastomerics* **III** (1979) no. 10, 50.
382. F. V. Bailor, *Rubber World* **180** (1979) no. 1, 42.
383. Imperial Chemical Industries, GB 407 643, 1938 (M. W. Perrin, E. W. Fawcett, J. G. Paton, E. G. Williams).
384. J. Brandrup, E. H. Immergut (eds.): *Polymer Handbook*, 3rd ed., Wiley & Sons, Inc., New York 1989, p. II/181.
385. R. D. Burkhart, N. L. Zutty, *J. Polym. Sci. Part A* **1** (1963) 1137.
386. US 3 168 456 (1965).
387. R. A. Terterjan, A. J. Dinzes, M. W. Rüsakow, *Neftekhimiya* **3** (1965) no. 5, 340.
388. B. Erussalimsky *et al.*, *Makromol. Chem.* **104** (1967) 288.
389. M. Rätzsch, W. Schneider, D. Musche, *J. Polym. Sci. Polym. Chem. Ed.* **9** (1971) 785.
390. F. Leder, M. J. Wisotsky, *J. Polym. Sci. Polym. Chem. Ed.* **10** (1972) 447.
391. T. K. Wu, D. W. Ovenall, G. S. Reddy, *J. Polym. Sci. Polym. Phys. Ed.* **12** (1974) 901.
392. J. Schaefer, *J. Phys. Chem.* **70** (1966) 1975.
393. M. J. Wisotsky, A. E. Kober, I. A. Zlochower, *J. Appl. Polym. Sci.* **15** (1971) 1737. Links
394. Tosoh Corporation, EP 0 453 205, 1990 (K. Soga, Y. Naito).
395. Farbenfabriken Bayer AG, DE-OS 1 495 660, 1957; DT 1 126 613, 1957; DT 1 126 614, 1957 (H. Bartl).
396. Farbenfabriken Bayer AG, DT 1 495 767, 1963 (W. D. Schellenberg, H. Bartl).
397. Farbenfabriken Bayer AG, US 3 325 460, 1967 (W. D. Schellenberg, H. Bartl).
398. Bayer, DE 3 323 531, 1983 (H. Bartl, F. Mietsch, K. Riebel, J. Thörmer).
399. Bayer, DE 38 159 46, 1988 (W. Baade *et al.*).
400. Bayer, EP 0 307 755, 1989 (H. Bartl).
401. Bayer, EP 0 341 499, 1988 (W. Baade).
402. Bayer, EP 0 374 660, 1988 (W. Baade, W. Obrecht, C. Ohm).
403. Bayer, EP 0 374 664, 1988 (A. Kolwert, W. Obrecht, H. Sutter).
404. Bayer, EP 0 374 666, 1989 (W. Baade).
405. Kuraray, US 4 485 225, 1983 (K. Satoh, K. Yonezu).
406. Kuraray, EP 0 136 559, 1983 (K. Matsumoto, J. Tanaka).
407. Du Pont, EP 0 460 936, 1990 (B. Hatch, D. R. Marshall).
408. Kuraray Co., EP 0 495 495, 1992 (Y. Hisamichi *et al.*).
409. B. V. Mitra, M. R. Katti, *Pop. Plast.* **18** (1973) no. 10, 15.
410. International Institute of Synthetic Rubber Producers *Worldwide Rubber Statistics 1996*.
411. BGVV XXI: Bedarfsgegenstände auf Basis von Natur- und Synthesekautschuk.
412. Bayer: "Levapren, Ethylen – Vinylacetat Kautschuk (EVM), *Handbuch für die Gummi-Industrie,* 2nd ed., Leverkusen 1992.
413. D. L. Hertz Jr. in A. K. Bhowmick, H. L. Stephens (eds.): *Handbook of Elastomers, New Developments and Technology*, "chap. 13", Marcel Dekker, New York 1988.
414. E. Scheer: "Epichlorohydrin Elastomers," in R. O. Babbit (ed.): *Vanderbilt Handbook,* R. T. Vanderbilt Co. Inc., Norwalk, 1978, p. 275.
415. D. A. Berta, E. J. Vandenberg in A. K. Bhowmick, H. L. Stephens (eds.): *Handbook of Elastomers, New Developments and Technology*, "chap. 19", Marcel Dekker, New York 1988.
416. K. Mori, Y. Nakamura, *Rubber Chem. Technol.* **57** (1984) 665. Links
417. K. Hashimoto, M. Ingami, A. Maeda, N. Watanabe, Paper no. 8 presented at Rubber Division, American Chemical Society, Atlanta, Oct. 7 – 10,1986.
418. E. J. Vandenberg, *CHEMTECH* 1983, Aug., 474.
419. J. Schaefer, R. Dalton: *Propylene Oxide High Performance Elastomer for Automobile Engine Mounts*, Detroit Rubber Group, March 15, 1990.
420. N. Calderon, *J. Macromol. Sci. Revs. Macromol. Chem., C7* **1** (1972) 105.
421. K. J. Ivin: *Olefin Metathesis*, Academic Press, London 1983.
422. V. Dragutan *et al.*: *Olefin Metathesis and Ring-Opening Polymerization of Cyclo-Olefins*, Wiley Interscience, Chichester 1985.
423. Société Chimique des Charbonnages, FR 74 10368, 1974 (J. Lahouste *et al.*).
424. Nippon Zeon Co. Ltd, JP 16 37751 (H. Nagai *et al.*).
425. M. J. Huggins, *J. Polym. Sci.* **8** (1952) 257.
426. M. J. Huggins, *J. Polym. Sci. Polym. Lett. Ed.* **11** (1973) 389.
427. S. Warwel, *Erdöl Erdgas Kohle* **103** (1987) 238.
428. R. Streck, *CHEMTECH* 1989, 498.
429. G. Dall'Asta, "Polymer Synthesis and Conversions via Metathesis and Investigations on the Structure of Metathesis Catalysts," *Plenary Main Sect. Lect. Int. Congr. Pure Appl. Chem., 24th, 1973*, vol. **1** (1974) 133; G. Dall'Asta, *Rubber Chem. Technol.* **47** (1974) 511. Links
430. G. Gianotti, A. Capizzi, L. Del Guidice, *Rubber Chem. Technol.* **49** (1976) 170. Links
431. A. Dräxler, *Kautsch. Gummi Kunstst.* **42** (1989) 868.
432. K.-D. Hesse, "Vestenamer – a New Rubber Based on Butadiene," *IISRP Annual Meeting*, Edinburgh, June 8 – 12, 1987.
433. ICI, GB 701 106, 1953 (W. H. B. Reed).
434. W. H. B. Reed, *J. Chem. Soc.* 1954, 1931.
435. W. Brenner *et al.*, *Justus Liebigs Ann. Chemie* **727** (1969) 161.
436. R. Streck, *J. Mol. Catal.* **76** (1992) 359.
437. D. Zerpner, W. Holtrup, R. Streck, *J. Mol. Catal.* **36** (1986) 153.
438. W. Holtrup, R. Streck, W. Zaar, D. Zerpner, *J. Mol. Catal.* **36** (1986) 127.
439. G. Natta, G. Dall'Asta, I. W. Bassi, G. Carella, *Makromol. Chem.* **91** (1966) 87. N. Calderon, E. A. Ofstead, W. A. Judy, *J. Polym. Sci. Part A – I* **5** (1967) 2209.
440. K. J. Ivin, J. C. Mol: *Olefin Metathesis and Metathesis Polymerization*, Academic Press, London 1997.
441. V. Dragutan, A. T. Balaban, M. Dimonie: *Olefin Metathesis and Ring-Opening Polymerization of Cyclo-Olefins*, J. Wiley & Sons, Bukarest 1985.
442. V. Dragutan, R. Streck, *Catalytic Polymerization of Cycloolefins*, Elsevier, Amsterdam 2000.
443. K. W. Scott *et al.*, *Adv. Chem. Ser.* **91** (1969) 399.
444. R. Streck, *Chem. Ztg.* **99** (1975) 397.
445. N. Calderon, E. A. Ofstead, W. A. Judy, *Angew. Chem.* **88** (1976) 433.
446. N. Calderon, J. P. Laurence, E. A. Ofstead, *Adv. Organomet. Chem.* **17** (1979) 494.

447. H. Höcker, *Angew. Makromol. Chem.* **100** (1981) 87. Links
448. R. Streck, *J. Mol. Catal.* **15** (1982) 3. N. Calderon *et al.*, *J. Am. Chem. Soc.* **90** (1968) 4133. Links J. L. Hérrisson, Y. Chauvin, *Makromol. Chem.* **141** (1971) 161.
449. T. Usami, Y. Gotoh, S. Takayama, *Eur. Polym. J.* **21** (1985) 885. Links
450. W. Holtrup, R. Hammel, *Chem. Ztg.* **109** (1985) 267.
451. A. Dräxler: "Polyalkenylenes" in A. K. Bhowmick, H. L. Stephens (eds.): *Handbook of Elastomers,* Marcel Dekker, New York 1988, p. 661.
452. A. Dräxler, *Kautsch. Gummi Kunstst.* **34** (1981) 185; *Elastomerics* **115** (1983) 16.
453. A. Dräxler, *Kautsch. Gummi Kunstst.* **36** (1983) 1037.
454. K. M. Diedrich, *Kautsch. Gummi Kunstst.* **42** (1989) 1130.
455. K. zur Nedden, K. M. Diedrich, G. Huhn: *Meeting of ACS Rubber Division*, Oct. 17 – 20, 1989; *Rubber Plastics News* **19** (1990) no. 26, 14.
456. K. P. Richter, K. H. Nordsiek, *Kautsch. Gummi Kunstst.* **43** (1990) 439.
457. Hüls AG, DE-OS 4 111 158 (J. Peter, P. Schmidt, D. Mahlke, K. zur Nedden). D. Mahlke, *Kautsch. Gummi Kunstst.* **46** (1993) 889.
458. K. M. Diedrich, B. J. Burns, *Polymer Recycling* **4** (1999) no. 4 237 – 246.
459. K. M. Diedrich, B. J. Burns, *Elastomery*, **4** (2000) no. 1, 9 – 14.
460. K. M. Diedrich, *Asphalt* **35** (2000) no. 5, 6.
461. Creanova Inc., USP 5 936 015, 1999 (B. J. Burns).
462. K. M. Diedrich, "Vestenamer — New Data on Tire Manufacturing Applications," *International Tire Exhibition and Conference — ITEC '96*, Akron, OH, September 10 – 12, 1996, Paper 24C.
463. E. L. Warrick, O. R. Pierce, K. E. Polmanteer, J. C. Saam, *Rubber Chem. Technol.* **52** (1979) 437 – 525. Links
464. E. G. Rochow, *CHEMTEC* 1980, Sept., 532 – 538.
465. B. R. Trego, H. W. Winnan, *RAPRA Rev. Rep.* **3** (1990) 31, 1 – 91.
466. J. Ackermann, V. Damrath, *Chem. Unserer Zeit* **23** (1989) no. 3, 86 – 97. Links
467. K. E. Polmanteer, *Rubber Chem. Technol.* **61** (1988) 470 – 506. Links
468. D. Wrobel *et al.*: *Silicones Chemistry and Technology*, Vulkan Verlag, Essen 1989, pp. 65 – 79.
469. M. E. Livingstone, P. R. Dvornic, R. W. Lenz, *J. Appl. Polym. Sci.* **27** (1982) 3239 – 3251. Links
470. T. Timm, *GAK Kautsch. Gummi Kunstst.* **39** (1986) no. 1, 15 – 24.
471. M. T. Maxson, C. L. Lee, *GAK Kautsch. Gummi Kunstst.* **39** (1986) no. 10, 532 – 539.
472. R. E. Drake, *ACS-Meeting 1981*, paper no. 6.
473. H. Yoshida, J. Watanabe, Y. Zama, I. Umeda, ACS-Meeting, Oct. 1989 paper no., 122.
474. J. M. Mitchell, T. Wada, K. Itoh, *GAK Kautsch. Gummi Kunstst.* **39** (1986) no. 6, 294 – 301.
475. *Encyclopedia of Polymer Science and Engineering*, 2nd ed., vol. **11**, Wiley and Sons, New York, p. 690.
476. D. Finney, *Rubber World* **202** (1990) June, 33 – 37.
477. B. Arkles, *CHEMTECH* 1983 no. 9, 542 – 555.
478. Wacker, EP 378 785, 1989 (J. Schuster, H. Müller, H. Vorbuchner).
479. Wacker, US 4 008 198, 1975 (H. Krohberger, J. Burkhardt, J. Patzke).
480. Dow Corning, EP 62 427, 1981 (M. C. Murray, J. M. Donatelli).
481. B. B. Boonstra, H. Cochrane, E. M. Dannenberg, *Rubber Chem. Technol.* **48** (1975) 558 – 576. Links
482. Dow Corning, EP 337 791, 1988 (S. Spells).
483. Rhône Poulenc, US 3 264 258, 1961 (L. Ceyzeriat).
484. G. Matthews: *Polymer Mixing Technol.*, Applied Science Publ., London 1982.
485. Rhône Poulenc, EP 258 159, 1990 (M. Alliot-Lugaz, G. Roulett).
486. Degussa, "Basic Characteristics and Applications of Aerosil", *Technical Bull. Pig.,* part 11, no. 63, Frankfurt, Pig 63–0–3–1176k.
487. E. M. Barrell II., R. Hawkins, A. A. Fukushima, J. F. Johnson, *J. Polym. Sci. Polym. Symp.* **71** (1984) 189 – 202.
488. J. D. Van Drumpt, *Rubber World* **197** (1988) March, 33 – 41.
489. Dow Corning, US 4 946 878, 1990 (J. D. Jensen, C. J. Bilgrien).
490. Shinetsu, US 3 671 480, 1972 (T. Wada, K. Itoh).
491. C. M. Blow: *Rubber Technology and Manufacture*, Butterworth, London 1971.
492. Lynch: *Handbook of Silicone Rubber Fabrication*, Van Norstrand Reinhold, New York 1978.
493. Steinbichler, *Kunststoffe* **77** (1987) no. 10, 931 – 938.
494. J. Trejbal, S. Mueller, M. Schaetz, *Plasty Kauc.* **22** (1985) no. 9, 275 T62 – T67.
495. J. H. Saunders, K. C. Frisch: "Polyurethanes: Chemistry and Technology," *High Polymers*, vol. **16**, "part 1 and 2", Interscience Publishers, New York 1962, 1964.
496. P. Wright, A. P. C. Cumming: *Solid Polyurethane Elastomers*, Maclaren and Sons, London 1969.
497. *Kunststoff-Handbuch,* 1st ed., W. Kallert (ed.), vol. 7, 1966, pp. 247 – 268; 2nd ed., W. Kallert (ed.), 1983, pp. 440 – 445; 3rd ed., K. Recker (ed.), in preparation, Hanser Verlag, München.
498. N. R. Legge, G. Holden, H. E. Schröder (eds.): *Thermoplastic Elastomers a Comprehensive Review*, Hanser Verlag, München 1987, pp. 15 – 46.
499. H. Kleimann, *GAK Kautsch. Gummi Kunstst.* **36** (1983) 175 – 179.
500. Bayer, Verschnitte von Urepan 640 mit Urepan 641 und Urepan VP PU 0332, Company Brochure, Leverkusen, 1989.
501. O. Bayer, E. Müller, *Angew. Chem.* **72** (1960) no. 24, 934 – 939.
502. W. Kallert, *GAK Kautsch. Gummi Kunstst.* **19** (1966) 363 – 371.
503. Bayer, Study on the Effects of Varying the Content of Peroxide and Coagent in Urepan 640 G, 1st and 2nd study, Company brochure, Leverkusen 1990.
504. Bayer, Vergleichende Prüfergebnisse niedrig gefüllter Vulkanisate, Company brochure, Leverkusen 1986.
505. Z. T. Ossefort, F. B. Testroet, *Rubber Chem. Technol.* **39** (1966) 1308 – 1327. Links
506. W. Neumann, J. Peter, H. Holtschmidt, W. Kallert, "Improvements in the Ageing and Hydrolysis Behavior of Elastomers Containing Ester Groups," *4th Rubber Technology Conference*, London 1962.
507. Thiokol Chemical Corp., Improving Water Resistance of Peroxide Cured Elastothane, Trenton, Technical Service and Development, E-455–0.14A, 1966.
508. W. Neumann, P. Fischer, *Angew. Chem.* **74** (1962) 801 – 806.
509. T. Timm, *GAK Kautsch. Gummi Kunstst.* **35** (1982) 568 – 584, **36** (1983) 257 – 268.
510. P. H. McKinstry, M. Chase, *Urethanes Technology* (1990), April – May 24 – 28.
511. V. D. Arendt, R. E. Logan, R. Saxon, *J. Cell. Plast.* (1982) 376 – 382.
512. O. Hutzinger (ed.), *The Handbook of Environmental Chemistry*, vol. **3**, "part G", Springer Verlag, Berlin 1991, pp. 28 – 31.
513. *Houben-Weyl*, 4th ed., **E 20**, 1589.

514. R. F. Dunbrook in G. S. Whitby (ed.), *Historical Review, Synthetic Rubber*, Wiley-Interscience, New York 1954, p. 32.
515. C. H. Fisher, G. S. Whitby, E. M. Beavers in G. S. Whitby (ed.), *Miscellaneous Synthetic Elastomers, Synthetic Rubber*, Wiley-Interscience, New York 1954, p. 892.
516. J. C. Patrick, US 1 890 191, 1927.
517. J. C. Patrick, N. M. Mnookin, GB 0 302 270, 1927.
518. J. Baer, CH 127 540, 1926.
519. R. H. Gobran, M. B. Berenbaum: “Polysulfide and Monosulfide Elastomers” in J. P. Kennedy, E. G. M. Törnqvist (eds.): *Polymer Chemistry of Synthetic Elastomers,* part 2, Interscience, New York 1969.
520. T. C. P. Lee: *Progress of Rubber Technology*, Applied Science Publishers, London 1983, p. 145.
521. S. K. Flanders, *J. Am. Chem. Soc. Educational Symp.* **14**, 1985.
522. R. F. Ohm: “Additives that affect mixing” in R. F. Grossman (ed.): *Mixing of Rubber*, Chapman & Hall, London 1997, p. 91 – 102.
523. J. Furukawa, S. Yamashita, S. Ido: “Elastomers by Miscellaneous or Unknown Mechanisms” in J. P. Kennedy, E. G. M. Törnqvist (eds.): *Polymer Chemistry of Synthetic, Elastomers,* part 2, Interscience, New York, 1969.
524. T. C. P. Lee, *RAPRA Review Reports* (1999), 9, 1 – 94.
525. *Worldwide Rubber Statistics* (2001), International Institute of Synthetic Rubber Producers, Houston 2001.
526. D. Shaw, *Eur. Rubber J.* **174** (1992) Feb., 2.
527. E. S. Nefed'ev *et al.*, *Polymer* **33** (1992) 3911 – 3915. Links
528. K. Guempel, *Bauen Kunstst.* (1992) no. 6 19 – 21.
529. H. Lucke, R. Schmidt, *Adhaesion Kleben Dichten* **37** (1993) 34 – 35.
530. A. R. Tripathy, C. K. Das, *Plast. Rubber Compos. Process Appl.* **21** (1994) 263 – 290.
531. FMC Corp., WO 79 743, 2001 (D. D. Janoff, J. C. Vicic).
532. G. B. Lowe, *Int. J. Adhesion & Adhesives* **17** (1997), no. 4, 345 – 348.
533. M. Bowtell, *Adhesives Age* **40** (1997) 62 – 63.
534. *Glass Dig.* **72** (1993) 42.
535. *Glass Dig.* **71** (1992) 41.
536. Morton International Ltd., US 5 567 258-A, 1996 (T. C. P. Lee, G. B. Lowe).
537. H. Lucke, R. Schmidt, *Adhaes. - Kleben Dichten* **40** (1996), p. 33 – 35.
538. G. B. Lowe, M. Futscher, *Proceedings Manchester* (1996) paper 14.
539. *Rubber Plastics News* (1992) Feb. 10, 5.
540. Toray Industries Inc., Chugai Shoko K. K., Osada Giken K. K., JP 070 249, 2002 (Tadayoshi Kamaishi, Maekawa Toshihiro, Kazuo Hayakawa, Hideharu Nagata).
541. K. Oba, *RILEM Proceedings* **36** (1999), 151 – 160.
542. A. Idiyatova, Yu. N. Khakimullin, A. G. Liakumovich, *Kautschuk i Rezina* **5** (1998) 30 – 33.
543. Toray Industries Inc., Toray Thiokol K. K., Osada Giken K. K., JP 097 271, 2002 (Tadayoshi Kamaishi, Takao Sakamoto, Kiyokata Ishii, Hideharu Nagata).
544. Toray Thiokol K. K., JP 064 504, 2001 (Kazuhisa Sakae, Toru Hirose).
545. Springborn Laboratories Inc., US 5 320 871-A, 1994 (S. C. Agro, B. O. Baum, P. J. Courtney).
546. Rohm & Haas Co., Morton International Inc., EP 1 044 936, 2001 (J. R. Gilmore, S. J. Hobbs, K. B. Potts).
547. *Jpn. Chem. Week* **28** (1987) 7.
548. Morton International Inc., US 5 663 219-A, 1997 (K. R. Chokshi, K. B. Potts).
549. J. Comyn, J. Day, S. J. Shaw, *Int. J. Adhesion & Adhesives* **17** (1997), p. 213 – 221.
550. V. V. Markov, S. V. Reznichenko, S. V. Knyazeva, Y. A. Ionov, *Int. Polym. Sci. Technol.* **23** (1996), p. T/30 – 31.
551. J. Murray, *Adhesives Age* **39** (1996), p. 16 – 17.
552. K. Ganesh, K. Kishore, *Macromolecules* **29** (1996) 26 – 29. Links
553. Toray Ind. Inc., WO 94 00638, 1994 (T. Kamaishi, H. Tanisugi).
554. Toray Thiokol Co. Ltd., US 5 393 861-A, 1995 (K. Sakae, K. Echigoya, M. Okajima).
555. Henkel Teroson GmbH, WO 90 222, 2001 (S. Grimm, M. Proebster).
556. Impact Coatings Inc., US 5 494 729-A, 1996 (J. W. Henry, B. C. McCaw).
557. M. Futscher, H. Lucke, *Int. Polym. Sci. Technol.* **22** (1995) T/35 – 38.
558. *Worldwide Rubber Statistics 1995*, International Institute of Synthetic Rubber Producers, Houston 1995.
559. R. T. Morrissey, *Ind. Eng. Chem.* **47** (1955) 1562.
560. B. F. Goodrich, US 2 732 354, 1952 (R. T. Morrissey, M. R. Frederick).
561. B. F. Goodrich, US 2 816 098, 1954 (R. T. Morrissey).
562. B. F. Goodrich, US 2 833 734, 1958 (R. T. Morrissey, H. J. Weiss).
563. Michelin, FR 1 286 688, 1960 (J.-M. Massoubre, G. L. Travers).
564. F. P. Baldwin *et al.*, *Rubber Plast. Age* **42** (1961) 500.
565. Esso, US 2 944 578, 1955 (F. P. Baldwin, R. M. Thomas).
566. Esso, US 2 964 493, 1956 (F. P. Baldwin, R. M. Thomas).
567. D. R. Lide (ed.): *CRC Handbook of Chemistry and Physics*, 72nd ed., CRC Press Inc, Boca Raton, Fla. 1991.
568. G. B. Sergeev, Yu. A. Seguchev, V. V. Smirnov, *Usp. Khim.* **42** (1973) 1545.
569. R. Vukov, *Rubber Chem. Technol.* **57** (1984) 275. Links
570. Exxon, EP-A 161 785, 1985 (I. J. Gardner *et al.*).
571. Exxon, US 4 632 963, 1984 (I. J. Gardner *et al.*).
572. I. J. Gardner, J. V. Fusco: “Halobutyls: Structure and Reactivity,” *ACS Rubber Div. Meeting*, Montreal, May 1987, paper no. 90.
573. A. Van Tongerloo, R. Vukov: “Butyl Rubber – Halogenation Mechanisms,” *Proc. Int. Rubber Conf., Venice* (1979), p. 70.
574. L. B. Krentsel, S. O. Travin, A. D. Litmanovich, K. K. Yutujan, *Eur. Polym. J.* **21** (1985) no. 4, 405. Links
575. L. Kh. Guseinova, L. B. Krentsel, L. V. Semenikhina, A. A. Azizov, V. M. Akhmedov, *Dokl. Akad. Nauk Az. SSR* **44** (1988) no. 7, 44.
576. G. B. Sergeev, Yu. A. Serguchev, A. A. Smirnov, *Usp. Khim.* **42** (1973) 1545.
577. G. Kaszas, *Polymer Preprints* **40** (1999) no. 2, 579.
578. Exxon Research & Engineering Co. US 4 634 741 (I. J. Gardner, J. V. Fusco, F. P. Baldwin).
579. Exxon Chemical Co., US 5 162 445, (K. W. Powers, H-C. Wang).
580. J. V. Fusco in K. C. Baranwal, H. L. Stephens (eds.): *Basic Elastomer Technology 1st ed.*, Rubber Division, ACS, Cleaveland 2001, p. 346.
581. *Kirk-Othmer*, 4th ed., **8**, 934.

582. H. Schwind, F. Keller, *Faserforsch. Textiltech.* **29** (1978) 135 – 141.
583. W. Busch, F. Kloos, J. Brandrup, *Angew. Makromol. Chem.* **105** (1982) 187 – 201. Links
584. P. J. Canterino, G. R. Kahle, *J. Appl. Polym. Sci.* **6** (1962) 20 – 31. Links
585. I. J. Gardner, F. P. Baldwin, *Rubber Chem. Technol.* **49** (1976) 391.
586. D. C. Edwards, *Rubber Chem. Technol.* **60** (1987) 62. Links
587. K. Ho, R. Steevenzs, *Rubber Chem. Technol.* **62** (1989) 42. Links
588. J. Walker, W. Hopkins, R. H. Jones, *Int. Rubber Conf. 1985*, Kyoto, p. 205.
589. R. C. Kowalski, *Chem. Eng. Prog.* **85** (1989) no. 5, 67. Links
590. D. M. Cheng *et al.*, *Rubber Chem. Technol.* **63** (1990) 265. Links
591. C. Y. Chu, K. N. Watson, R. Vukov, *Rubber Chem. Technol.* **60** (1987) 636. Links
592. Z. E. Zolles (ed.): *Bromine and its Compounds*, Ernest Benn, London 1966, pp. 775 – 778.
593. Material Safety Data Sheet, Polysar Rubber Corporation, Pittsburgh 2001, USA.
594. Exxon, WO 13102, 1991 (N. F. Newman, I. J. Gardner).
595. F. P. Baldwin *et al.*, *Rubber Chem. Technol.* **49** (1976) 390.
596. E. Rohde, *GAK Kautsch. Gummi Kunstst.* **35** (1982) 478 – 487.
597. J. Menough, *Rubber World* **166** (1984) May, 14 – 15.
598. Hoechst, DE-OS 1 420 407, 1957 (H. Klug, K.-H. Mittelberger).
599. Dow Chemical, DE-OS 1 495 518, 1964 (W. L. P. Young, B. Posey Jr.
600. Dow Chemical, US 3 563 974, 1968 (C. R. Eckhardt, W. M. Bungo).
601. Hoechst, DE-OS 3 213 687, 1983 (H. J. Hardel, H. W. Keuper, J. Brandrup, D. Fleischer).
602. G. A. Baseden in R. F. Ohm (ed.): *The Vanderbilt Rubber Handbook*, 13th ed., R. T. Vanderbilt Company, Inc., Norwalk 1990, pp. 183 – 189.
603. J. E. A. Williams, *Eur. Rubber J.* **165** (1983) 22 – 25.
604. J. G. Pillow, R. E. Ennis: "ACSM – A New Generation of Chlorinated Elastomer for High Temperature Dynamic Applications," *ACS Rubber Division Meeting*, Detroit Oct. 17 – 20, 1989.
605. T. Nakagawa, K. Sando, N. Morishita, *GAK Kautsch. Gummi Kunstst.* **45** (1992) 21 – 23.
606. J. M. Tixhon, *Autotech Congress Birmingham* **40** (1989) 1 – 26.
607. Du Pont, DE-OS 2 034 592, 1971 (E. J. Breda).
608. Du Pont, GB 804 535, 1957 (J. Calil).
609. Hoechst, DE-AS 1 105 176, 1958 (L. Orthner, F. Landauer, M. Reuter, H. Herzberg).
610. Dow Chemical, EP 177 865, 1986 (R. A. Blanchard).
611. *Worldwide Rubber Statistics 1991*, International Institute of Synthetic Rubber Producers, Inc. (IISRP), Houston 1991, p. 76. A. J. Maldonado in R. F. Ohm (ed.): *The Vanderbilt Rubber Handbook*, 13th ed., R. T. Vanderbilt Company, Inc., Norwalk 1990, p. 190.
612. J. Thörmer, G. Marwede, H. Buding, *GAK Kautsch. Gummi Kunstst.* **36** (1983) 269 – 274.
613. J. Mirza, N. Schön, J. Thörmer, *GAK Kautsch. Gummi Kunstst.* **39** (1986) 615 – 620.
614. J. Thörmer, J. Mirza, N. Schön, *Elastomerics* **118** (1986) 28 – 32.
615. J. Thörmer *et al.*, *Rubber World* **201** (1989) 25 – 30.
616. M. Mezger, U. Eisele, E. Rohde, *GAK Kautsch. Gummi Kunstst.* **44** (1991) 341 – 347.
617. Bayer AG: *Handbuch für die Gummiindustrie*, 2nd ed., Leverkusen 1990, pp. 110 – 131.
618. K. Hashimoto, N. Watanabe, A. Yoshioka, *Rubber World* **190** (1984) May, 32 – 47.
619. K. Hashimoto, N. Watanabe, M. Oyama, Y. Todani, *Gummi Fasern Kunstst.* **37** (1984) 602 – 606.
620. N. Watanabe, K. Hashimoto, M. Oyama, *GAK Kautsch. Gummi Kunstst.* **42** (1989) 912 – 916.
621. A. H. Weinstein, *Rubber Chem. Technol.* **57** (1984) 203 – 215. Links
622. J. R. Dunn, G. C. Blackshaw, J. Timar, *Elastomerics* **118** (1986) 20 – 25.
623. R. J. Weir, G. C. Blackshaw, *Eur. Rubber J.* **171** (1989) 22 – 24.
624. W. A. Wisemann, J. J. Ridland, *SAE Tech. Pap.* 1989, 1 – 11.
625. R. J. Weir, G. C. Blackshaw, *Rubber India* **42** (1990) 15 – 23.
626. A. Kondo *et al.*, *Macromolecules* **21** (1988) 2918 –2924. Links
627. D. J. Carlsson, S. Chmela, D. M. Wiles, *Macromol. Symp.* **27** (1989) 139 – 153.
628. S. Bhattacharjee, A. K. Bhowmick, B. N. Avasthi, *Polym. Degrad. Stab.* **31** (1991) 71 – 87.
629. W. Obrecht *et al.*, *Angew. Makromol. Chem.* **145/146** (1986) 161 – 179. Links
630. U. Eisele, Z. Szentivanyi, W. Obrecht, *J. Appl. Polym. Sci. Appl. Polym. Symp.* **50** (1992) 185 – 197.
631. Y. Kubo, K. Hashimoto, N. Watanabe, *GAK Kautsch. Gummi Kunstst.* **40** (1987) 118 – 121.
632. Nippon Zeon, US 5 013 796, priority March 24, 1988 (JP) (S. Hayashi, Y. Kubo, N. Watanabe, Y. Aimura).
633. Japan Synthetic Rubber, EP-A 0 366 377, priorities Oct. 22, 1988 (JP), Nov. 11, 1988 (JP), Aug. 8, 1989 (JP) (N. Sahabe, T. Ohara, T. Miyabayashi, H. Ise).
634. J. Thörmer, J. Mirza, Z. Szentivanyi, W. Obrecht, *GAK Kautsch. Gummi Kunstst.* **41** (1988) 1208 –1214.
635. Bayer, EP 0 131 196, priority July 9, 1983 (DE) (W. Stix, W. Nouvertné, H. Buding).
636. Bayer, EP-A 0 251 044, priority July 2, 1986 (DE) (H. Eichenauer *et al.*).
637. Bayer, EP-A 0 255 594, priority July 5, 1986 (DE) (C. Lindner *et al.*).
638. Bayer, EP-A 0 285 984, priority April 10, 1987 (DE) (E. Leitz, H.-J. Kress, D. Wittmann, H. Buding).
639. Bayer, EP-A 0 356 729, priority Aug. 13, 1988 (DE) (H. Eichenauer, H. Buding, J. Merten, K.-H. Ott).
640. Bayer, EP-A 0 361 131, priority Sep. 15, 1988 (DE) (H. Eichenauer, H. Buding, T. Eckel, K.-H. Ott).
641. Bayer, DE-OS 3 941 564, priority Dec. 16, 1989 (DE) (K.-H. Köhler, H. Buding, H. Eichenauer, J. Thörmer).
642. Nippon Zeon, EP 0 158 794, priority March 1, 1984 (JP) (M. Oyama, Y. Kubo).
643. Nippon Zeon, DE-OS 3 828 494, priority Aug. 24, 1987 (JP) (M. Aonuma, T. Suzuki, S. Isomura, K. Nishimura).
644. Evode Limited, GB-A 2 210 047, priority Sept. 18, 1987 (GB) (R. S. Whitehouse).
645. Bayer, EP 0 138 140, priority Oct. 14, 1983 (DE) (H. Buding, Z. Szentivanyi, J. Thörmer).
646. Bayer, EP 0 151 691, priority Oct. 29, 1983 (DE) (H. Buding, H. Königshofen, Z. Szentivanyi, J. Thörmer).
647. BICC Public Limited Company, EP 0 272 071, priority Dec. 16, 1986 (GB) (D. S. Ohbi).
648. BICC Public Limited Company, EP-A 0 319 199, priority Nov. 25, 1987 (GB) (P. J. Corish, M. R. Kendle).
649. Otto Haas, EP-A 0 322 521, priority Dec. 24, 1987 (DE) (H. F. Struckmeyer).
650. Karl Joh Gummiwarenfabrik GmbH; DE-OS 3 812 354, priority April 14, 1988 (DE) (W. von Tirpitz).
651. Bayer, EP 0 146 068, priority Dec. 14, 1983 (DE) (H. Buding, H. Königshofen, Z. Szentivanyi, J. Thörmer).

652. Bayer, DE-OS 3 518 273, priority Aug. 18, 1984 (DE), Aug. 23, 1984 (DE), Feb. 27, 1985 (DE) (H. Buding, H. Königshofen, Z. Szentivanyi, J. Thörmer).
653. Bayer, EP-A 0 384 210, priority Feb. 20, 1989 (DE) (Z. Szentivanyi, W. Obrecht, J. Wassen).
654. Bayer, DE-OS 3 394 954, priority Oct. 20, 1989 (DE) (R. Musch, J. Thörmer, H. Buding, F. Leibbrandt).
655. Nippon Zeon, DE 3 232 692, priority Feb. 2, 1981 (JP) (M. Oyama, K. Hashimoto).
656. Nippon Zeon, US 4 614 779, priority May 11, 1984 (JP) (N. Watanabe, H. Shomoda, Y. Kubo).
657. Polysar, EP-A 0 364 859, priority Oct. 13, 1988 (US) (H. P. Plaumann).
658. Nippon Zeon, DE 3 141 573, priority Oct. 20, 1980 (JP) (M. Oyama, K. Hashimoto).
659. D. Brück, *GAK Kautsch. Gummi Kunstst.* **42** (1989) 107 – 110.
660. D. Brück, *GAK Kautsch. Gummi Kunstst.* **42** (1989) 194 – 197.
661. A. J. Marshall, I. R. Jobe, T. Dee, C. Taylor, *Rubber Chem. Technol.* **63** (1990) 244 – 255. Links
662. Shell Oil Company, US 3 700 637, priority Dec. 11, 1967 (US) (A. M. T. Finch Jr.,
663. Bayer, DE 2 539 132, priority Sept. 3, 1975 (DE) (D. Oppelt, H. Schuster, J. Thörmer, R. Braden).
664. Bayer, EP 0 134 023, priority Aug. 19, 1983 (DE) (H. Buding, H. Königshofen, Z. Szentivanyi, J. Thörmer).
665. Polysar, US 4 464 515, priority Dec. 8, 1982 (CA) (G. L. Rempel, H. Azizian).
666. Polysar, US 4 503 196, priority Dec. 8, 1982 (CA) (G. L. Rempel, H. Azizian).
667. Bayer, EP 0 138 141, priority Oct. 13, 1983 (DE) (P. Fiedler, R. Braden, H. Buding).
668. Bayer, EP 0 174 576, priority Sept. 12, 1984 (DE) (H. Buding, P. Fiedler, H. Königshofen, J. Thörmer).
669. Bayer, EP 0 213 422, priority Aug. 8, 1985 (DE) (P. Fiedler, H. Buding, R. Braden, J. Thörmer).
670. Bayer, EP 0 223 151, priority Nov. 19, 1985 (DE) (P. Fielder, H. Buding, R. Braden).
671. Bayer, EP 0 224 139, priority Nov. 11, 1985 (DE) (T. Himmler, P. Fiedler, R. Braden, H. Buding).
672. Bayer, DE-OS 3 921 264, priority June 6, 1989 (DE) (H. Buding *et al.*).
673. Bayer, EP-A 0 405 268, priority June 6, 1989 (DE) (T. Himmler).
674. University of Waterloo, US 4 816 525, priority July 7, 1987 (US) (G. L. Rempel, N. A. Mohammadi, R. Farwaha).
675. University of Waterloo, US 4 812 528, priority June 6, 1988 (US) (G. L. Rempel, N. A. Mohammadi, R. Farwaha).
676. University of Waterloo, EP-A 0 455 154, priority May 5, 1990 (CA) (G. L. Rempel, N. T. McManus, N. Mohammadi).
677. G. L. Rempel, N. T. McManus, US 5 057 581, priority Dec. 13, 1990 (US).
678. D. Brück, D. Oppelt, Z. Szentivanyi, in J. Süderwald, R. Weis (eds.): *Reprints of Short Communications presented at IUPAC Makro Mainz, 26. IUPAC Int. Symp. Macromol. Mainz Sept. 17 – 21*, vol. **I**, 1979, pp. 538 – 541.
679. N. A. Mohammadi, G. L. Rempel, *Macromolecules* **20** (1987) 2362 – 2368. Links
680. Nippon Zeon, DE-AS 2 913 992, priority April 6, 1978 (JP) (A. Maeda *et al.*).
681. Nippon Zeon, DE 3 046 008, priority Dec. 6, 1979 (JP) (Y. Kubo, K. Ohura).
682. Nippon Zeon, DE 3 227 650, priority July 24, 1981 (JP) (Y. Kubo, K. Ohura).
683. Bayer, EP 0 200 029, priority April 20, 1985 (DE) (H. Buding, R. Casper).
684. Nippon Zeon, DE 3 905 432, priority Feb. 22, 1988 (JP) (Y. Kubo, K. Ohura).
685. Nippon Zeon, DE 3 046 251, priority Dec. 8, 1979 (JP) (Y. Kubo, T. Ohishi, K. Ohura).
686. Polysar, EP-A 0 354 413, priority Aug. 1, 1988 (US) (A. Osmann, W. G. Bradford).
687. Polysar, US 4 985 540, priority Nov. 20, 1989 (US) (W. G. Bradford, G. J. Arsenault, A. J. Marshall).
688. Bayer, DE-OS 4 032 597, priority Oct. 13, 1990 (DE) (P. Panster, S. Wieland, H. Buding, W. Obrecht).
689. Nippon Zeon, DE-OS 3 346 888, priority Dec. 23, 1982 (JP) (Y. Kubo, T. Kohtaki, K. Ohura).
690. Shell Internationale Research Maatschappij B.V., EP-A 0 368 381, priority Nov. 9, 1988 (US) (R. J. Hoxmeier, L. H. Slaugh).
691. The Goodyear Tire & Rubber Company, EP-A 0 279 766, priority Feb. 19, 1987 (US) (L. G. Wideman, W.-L. Hsu).
692. The Goodyear Tire & Rubber Company, EP 0 117 221, priority Feb. 22, 1983 (US) (L. G. Wideman).
693. Zeon Chemicals Inc., Corporate Profile Section, *Rubber World* **202** (1990) Aug., 61.
694. *Worldwide Rubber Statistics 1991*, International Institute of Synthetic Rubber Producers, Inc., p. 70.
695. H. R. Allcock, *Angew. Chem.* **89** (1977) 153; *Angew. Chem. Int. Ed. Engl.* **16** (1977) 147. Links
696. P. Potin, R. De Jaeger, *Eur. Polym. J.* **27** (1991) 341. Links
697. H. R. Allcock: *Phosphorous Nitrogen Compounds*, Academic Press, New York 1972.
698. R. H. Neilson, P. Wisian-Neilson, *Chem. Rev.* **88** (1988) 541. Links
699. C. Zhang, P. Wisian-Neilson, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **36** (1995) 481 – 482.
700. M. L. White, K. Matyjaszewski, *J. Macromol. Sci. Part A: Polym. Chem.* **A 34** (1996) 277 – 289.
701. J. A. Gruneich, P. Wisian-Neilson, *Polym. Prepr. (Am. Chem. Soc. Div. Polym. Chem.)* **37** (1996) 611.
702. M. A. Gomez *et al.*, *Eur. Polym. J.* **32** (1996) 717 – 723. Links
703. Y. Ni *et al.*, *Macromolecules* **29** (1996) 3401 – 3408. Links
704. H. R. Allcock, K. B. Visscher, Y.-B. Kim, *Macromolecules* **29** (1996) 2721 – 2728. Links
705. P. Wisian-Neilson, G.-F. Xu, *Macromolecules* **29** (1996) 3457 – 3461. Links
706. G. A. Carriedo *et al.*, *Macromolecules* **29** (1996) 5320 – 5325. Links
707. H. R. Allcock, S. E. Kuharcik, C. J. Nielsen, *Macromolecules* **29** (1996) 3686 – 3693. Links
708. J. Vandorpe, E. Schacht, *Polymer* **37** (1996) 3141 – 3145. Links
709. J. A. Gruneich, P. Wisian-Neilson, *Macromolecules* **29** (1996) 5511 – 5512. Links
710. Elf Atochem, EP 532 396, 1993 (P. Potin, S. Potin, R. Schneider, S. Bordere).
711. M. Gleria, R. de Jaeger, *J. Inorg. Organometl. Polym.* **11** (2001) 1 – 45.
712. M. L. Turner, *Annu. Rep. Prog. Chem. Sect. A: Inorg. Chem.* **97** (2001) 443 – 459.
713. K. Inoue, T. Itaya, *Bull. Chem. Soc. Japan* **74** (2001) 1381 – 1395.
714. C.-O. Turrin, C. Larre, A.-M. Caminade, J.-P. Majoral, *Phosphorus Research Bulletin* **10** (1999) 777 – 781.
715. H. R. Allcock, A. Ambrosio, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **41** (2000) 732.
716. H. R. Allcock, *Polym. Prepr. (Am. Chem. Soc., Div. Polym. Chem.)* **41** (2000) 553.
717. D. P. Gates, I. Manners, *J. Chem. Soc., Dalton Trans.* **15** (1997) 2525 – 2532.
718. K. B. Visscher, I. Manners, H. R. Allcock, *Advances in Chemistry* **239** (1994) 157–178.
719. A. R. McWilliams, H. Dorn, I. Manners, *Topics in Current Chemistry* **220** (2002) 141–167. Links
720. H. N. Stokes, *Am. Chem. J.* **17** (1895) 275.
721. H. N. Stokes, *Ber. Dtsch. Chem. Ges.* **28** (1895) 437.
722. H. N. Stokes, *Am. Chem. J.* **18** (1896) 629, 780.

723. H. N. Stokes, *Am. Chem. J.* **19** (1897) 782.
724. H. N. Stokes, *Am. Chem. J.* **20** (1898) 740.
725. H. N. Stokes, *Z. Anorg. Chem.* **19** (1899) 36.
726. H. Rose, *Justus Liebigs Ann. Chem.* **11** (1834) 129.
727. J. Liebig, *Justus Liebigs Ann. Chem.* **11** (1834) 139.
728. American Cyanamid, US 3 370 020, 1968 (H. R. Allcock, R. L. Kugel).
729. H. R. Allcock, R. L. Kugel, *J. Am. Chem. Soc.* **87** (1965) 4216. Links
730. H. R. Allcock, R. L. Kugel, K. J. Valan, *Inorg. Chem.* **5** (1966) 1709. Links
731. H. R. Allcock, R. L. Kugel, *Inorg. Chem.* **5** (1966) 1716. Links
732. S. Uzaki, K. Adachi, T. Kotaka, *Macromolecules* **21** (1988) 153. Links
733. D. P. Craig, N. C. Paddock: *Nonbenzoid Aromatics*, Academic Press, New York 1971.
734. H. R. Allcock, R. W. Allen, J. J. Meister, *Macromolecules* **9** (1976) 950. Links
735. Y. Chatani, K. Yatsuyanagi, *Macromolecules* **20** (1987) 1042. Links
736. H. R. Allcock, *CHEMTECH* **5** (1975) 552.
737. S. H. Rose, K. A. Reynard, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **13** (1972) 778.
738. W. F. Ferrar, F. V. Di Stefano, H. R. Allcock, *Macromolecules* **13** (1980) 1345. Links
739. S. H. Rose, *J. Polym. Sci. Part B* **6** (1968) 837.
740. H. R. Allcock, W. J. Cook, D. P. Mack, *Inorg. Chem.* **11** (1972) 2584. Links
741. P. Wisian-Neilson, R. H. Neilson, *J. Am. Chem. Soc.* **102** (1980) 2848. Links
742. H. R. Allcock, M. N. Mang, A. A. Dembek, K. J. Wynne, *Macromolecules* **22** (1989) 4179. Links
743. H. R. Penton, *GAK Kautsch. Gummi Kunstst.* **39** (1986) 301.
744. T. N. Bowmer *et al.*, *Macromolecules* **24** (1991) 4827. Links
745. H. R. Penton: “Organic and Organometallic Polymers,” *ACS Symp. Ser.* **360** (1988) 277.
746. H. R. Penton, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **28** (1987) 439.
747. N. S. Schneider, C. R. Desper, J. J. Beres in A. Blumstein (ed.): *Liquid Crystalline Order in Polymers*, Academic Press, New York 1978, p. 299.
748. M. A. Gomez *et al.*, *Polym. Commun.* **31** (1990) 308.
749. M. Kojima, J. H. Magill, *Makromol. Chem.* **186** (1985) 649.
750. R. E. Singler *et al.*, *Macromolecules* **24** (1991) 510. Links
751. M. Kojima, J. H. Magill, M. L. White, K. Matyjaszewski, *Macromol. Chem. Phys.* **196** (1995) 1713 – 1737. Links
752. H. R. Allcock, E. H. Klingenberg, *Macromolecules* **28** (1995) 4351 – 4360. Links
753. H. R. Allcock, W. J. Cook, *Macromolecules* **7** (1974) 284. Links
754. H. R. Allcock, G. Y. Moore, W. J. Cook, *Macromolecules* **7** (1974) 571. Links
755. S. J. Maynard, T. R. Sharp, J. F. Haw, *Macromolecules* **24** (1991) 2794. Links
756. M. L. White, K. Matyjaszewski, *J. Macromol. Sci. Pure Appl. Chem.* **A 32** (1995) 1115 – 1135.
757. J. C. Vicic, K. A. Reynard, *J. Appl. Polym. Sci.* **21** (1977) 3185. Links
758. R. Singler, G. L. Hagnauer in C. Carraher, J. Sheats, C. U. PittmanJr. (eds.): *Organometallic Polymers*, "chap. 26", Academic Press, New York 1978.
759. R. Singler, N. S. Schneider, G. L. Hagnauer, *Polym. Eng. Sci.* **15** (1975) 321. Links
760. E. J. Quinn, B. M. Althouse, P. Potin, *Symp. Chem. Comb. Toxicity*, MARM, Millersville PA, 1988.
761. A. Wilson, *J. Coated Fabr.* **7** (1978) 223.
762. Horizon Research, US 3 994 838, 1976 (J. E. Thompson, R. W. Sicka).
763. R. E. Singler, G. L. Hagnauer, N. S. Schneider, *Polym. News* **5** (1978) 9.
764. J. Green, *Rubber Plastics News* (1993) January 18, 25.
765. M. S. Jones, *Rubber World* (1991) 33.
766. H. Penton, *Eur. Rubber J.* **168** (1986) 20.
767. M. Gleria, *Phosporus Research Bulletin* **10** (1999) 55 – 69.
768. F. Björk, B. Stenberg, *Plast. Rubber Process Appl.* **9** (1988) 227.
769. W. Hoffmann, *Kunststoffe* **78** (1988) 132.
770. J. T. Bocks, *Gummi Fasern Kunstst.* **39** (1986) 374.
771. G. J. Knight, W. W. Wright, *Br. Polym. J.* **21** (1989) 199.
772. Ethyl Corp., Data Sheet, Eypel-A Foam, 9/86.
773. Ethyl Corp., US 4 198 381, 1980 (E. D. Hornbaker, H. M. Li).
774. Ethyl Corp., US 4 374 815, 1983 (H. M. Li).
775. Ethyl Corp., US 4 447 408, 1984 (H. M. Li).
776. Atochem, EP 132 188, 1984 (T. Abou Chakra, R. De Jaeger).
777. Atochem, EP 178 997, 1984 (R. De Jaeger, G. Pegniez, P. Potin).
778. Atochem, EP 299 819, 1987 (N. Pasimourt, G. Pegniez, P. Potin).
779. M. Heloui, R. De Jaeger, E. Puskaric, J. Heubel, *Makromol. Chem.* **183** (1982) 113.
780. G. D. Halluin, R. De Jaeger, J. P. Chambrette, P. Potin, *Macromolecules* **25** (1992) 1254.
781. G. D. Halluin, P. Potin, R. De Jaeger, *Bull. Soc. Chim. Belg.* **98** (1989) 653.
782. Elf Atochem, FR 2 700 170, 1994 (J. Chambrette, P. Potin).
783. G. L. Hagnauer, *J. Macromol. Sci. Chem.* **A16** (1981) 385.
784. M. S. Sennett, G. L. Hagnauer, R. L. Singler, G. Davies, *Macromolecules* **19** (1986) 959; **22** (1989) 4235.
785. A. N. Mujumdar, S. G. Young, R. L. Merker, J. H. Magill, *Macromolecules* **23** (1990) 14. Links
786. Ethyl Corp., US 5 006 324, 1991 (C. H. Kolich, B. R. Meltsner, H. G. Braxton).
787. J. H. Magill, R. L. Merker, US 4 946 938, 1990.
788. H. R. Allcock, J. E. Gardner, K. M. Smeltz, *Macromolecules* **8** (1975) 36. Links
789. E. Fluck, *Z. Anorg. Allg. Chem.* **320** (1963) 64. Links
790. E. F. Moran, *J. Inorg. Nucl. Chem.* **30** (1968) 1405.
791. H. R. Allcock, R. J. Best, *Can. J. Chem.* **42** (1964) 447. Links
792. Firestone Tyre and Rubber Co., US 4 609 724, 1986 (R. A. Elefritz, W. M. Cole, C. R. Johnson).
793. T. M. Conelly, J. K. Gillham, *J. Appl. Polym. Sci.* **20** (1976) 473. Links
794. G. Allen, C. J. Lewis, S. M. Todd, *Polymer* **11** (1970) 44. Links

795. H. R. Allcock, J. L. Desorcie, G. H. Riding, *Polyhedron* **6** (1987) 119. Links
796. G. S. Kyker, T. A. Antowiak, *Rubber Chem. Technol.* **47** (1974) 32. Links
797. E.-P. Flindt, H. Rose, *Z. Anorg. Allg. Chem.* **428** (1977) 204. Links
798. K. Matyjazewski *et al.*, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **32** (1991) 305.
799. K. Matyjazewski, J. Dauth, R. Montague, O. Nuyken, *J. Polym. Sci. Polym. Chem. Ed.* **30** (1992) 813.
800. R A. Montague, J. B. Green, K. Matyjaszewski, *J. Macromol. Sci. Pure Appl. Chem.* **A 32** (1995) 1497 – 1519.
801. R. H. Neilson *et al.*, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **32** (1991) 483.
802. H. R. Allcock *et al.*, *Macromolecules* **24** (1991) 2024. Links
803. J. A. Dodge, I. Manners, H. R. Allcock, *J. Am. Chem. Soc.* **112** (1990) 1268. Links
804. M. Liang, I. Manners, *J. Am. Chem. Soc.* **113** (1991) 4044. Links
805. H. R. Allcock, D. J. Brennan, J. M. Graaskamp, *Macromolecules* **21** (1988) 1. Links
806. H. R. Allcock, W. D. Coggio, *Macromolecules* **21** (1988) 153.
807. G. A. Carriedo, F. J. G. Alonso, P. A. Gonzalez, J. L. Garcia, P. Gomez-Elipe, *Phosphorus, Sulfur Silicon Relat. Elem.* **144–146** (1999) 73 – 76.
808. H. R. Allcock, C. A. Crane, C. T. Morrissey, J. M. Nelson, S. D. Reeves, C. H. Honeyman, I. Manners, *Macromolecules* **29** (1996) 7740–7747. Links
809. H. R. Allcock, J. M. Nelson, S. D. Reeves, C. H. Honeyman, I. Manners, *Macromolecules* **30** (1997) 50–56. Links
810. D. F. Lohr, H. R. Penton: “Polyphosphazene Elastomers”, Plastics Engineering, New York NY 2001, p. 591 – 603.
811. H. R. Allcock, J. Y. Chang, *Macromolecules* **24** (1991) 993. Links
812. H. R. Allcock, S. R. Pucher, *Macromolecules* **24** (1991) 23. Links
813. H. R. Allcock, S. Kwon, S. R. Pucher, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **31** (1990) 180.
814. E. Schacht, J. Vandorpe, Y. Lemmouchi, S. Dejardin, L. Seymour: “Degradable polyphosphazenes for biomedical applications” in R. M. Ottenbrite (ed.): *Frontiers in Biomedical Polymer Applications of Polymers*, Technomic, Lancaster PA 1998, p. 27 – 42.
815. Penn State Res Found, WO 95 28150, 1995 (H. R. Allcock, C. G. Cameron, D. E. Smith).
816. Penn State Res Found, WO 95 32736, 1995 (H. R. Allcock, S. R. Pucher, K. B. Visscher).
817. Ethyl Corp., US 5 104 947, 1992 (J. Crommen, E. Schacht).
818. Mitsubishi Engineering-Plastics Corp., EP 1 160 276, 2001 (H. Urabe, H. Ohyama, M. Suzuki).
819. WO 034 682, 1998 (Y. Ni, M. A. Winnik, I. Manners, Z. Pang, X. Gu).
820. Bayer AG, DE 19 828 539, 1999 (T. Eckel, M. Zobel, D. Wittmann).
821. R. Wycisk, P. N. Pintauro, W. Wang, S. O'Connor, *J. Appl. Polym. Sci.* **59** (1996) 1607 – 1617. Links
822. A. Bac, D. Roizard, P. Lochon, J. Ghanbaja, *Macromol. Symp.* **102** (1996) 225 – 232. Links
823. Penn State Res Found, US 5 457 160, 1995 (H. R. Allcock, K. B. Visscher).
824. Ethyl Petroleum Additives Ltd, EP 511 441, 1992.
825. Air Products and Chemicals Inc., EP 900 588, 1999 (M. Langsam).
826. Q. Guo, H. Tang, P. N. Pintauro, S. O'Connor, ACS Symposium Series **744** (2000) 162 – 173.
827. H. R. Allcock, A. Ambrosio, *ACS Symposium Series* **833** (2003) 82–101.
828. F. F. Stewart, M. K. Harrup, *Polymer News* **26** (2001) 78–85.
829. J. L. Acosta, E. Morales, M. Paleo, J. R. Jurado, *Eur. Polym. J.* **31** (1996) 13 – 18.
830. J. L. Luque, E. M. Bergas, J. R. Egea, *J. Appl. Polym. Sci.* **59** (1996) 1173 – 1180. Links
831. H. R. Allcock, M. E. Napierala, C. G. Cameron, S. J. O'Connor, *Macromolecules* **29** (1996) 1951 – 1956. Links
832. J. L. Acosta, E. Morales, *J. Appl. Polym. Sci.* **60** (1996) 1185 – 1191. Links
833. H. R. Allcock, S. E. Kuharcik, C. S. Reed, M. E. Napierala, *Macromolecules* **29** (1996) 3384 – 3389. Links
834. Penn State Res Found, US 5 414 025, 1995 (H. R. Allcock, W. D. Coggio, C. J. Nelson).
835. Penn State Res Found, WO 96 27630, 1996 (H. R. Allcock *et al.*).
836. H. R. Allcock, M. E. Napierala, D. L. Olmeijer, C. G. Cameron, S. E. Kuharcik, C. S. Reed, S. J. M. O'Connor, *Electrochimica Acta* **43** (1998) 1145 – 1150. Links
837. V. Chandrasekhar, *Advances in Polymer Science* **135** (1998) 139 – 205. Links
838. The Penn State Research Foundation, WO 002 673, 2002 (H. R. Allcock, E. C. Kellam III., V. R. Morford).
839. J. R. Wolfe: *Thermoplastic Elastomers; A Comprehensive Review*, Oxford University Press, New York 1987, p. 107.
840. R. Casper, L. Morbitzer, *Angew. Makromol. Chem.* **58/59** (1977) 1. Links
841. M. K. Stockdale, *J. Vinyl Technol.* **12** (1990) no. 4, 235.
842. Denka LCS, Company brochure, no. 1986–10–0001-PK, Düsseldorf, 1986.
843. W. Meckel, W. Goyert, W. Wieder: *Thermoplastic Elastomers; A Comprehensive Review*, Hanser Verlag, New York 1987, p. 13.
844. G. M. Estes, S. L. Cooper, A. V. Tobolsky, *J. Macromol. Sci. Rev. Macromol. Chem.* **4** (1970) 313.
845. N. S. Schneider, C. R. Desper, J. L. Illinger, A. O. King, *J. Macromol. Sci. Phys.* **B 11** (1975) 527.
846. J. W. C. Van Bogart, P. E. Gibson, S. L. Cooper, *J. Polym. Sci. Polym. Phys. Ed.* **21** (1983) 65.
847. C. S. Paik Sung, N. S. Schneider, *J. Mater. Sci.* **13** (1978) 1684.
848. Z. Ophir, G. L. Wilkes, *J. Polym. Sci. Polym. Phys. Ed.* **18** (1980) 1969.
849. L. Morbitzer, H. Hespe, *J. Appl. Polym. Sci.* **27** (1982) 2891. Links
850. C. G. Seefried, J. V. Koleste, F. E. Critchfield, *J. Appl. Polym. Sci.* **19** (1975) 3185. Links
851. Union Carbide, US 4 055 549, 1976 (R. Roberts).
852. Upjohn, US 3 899 467, 1974 (H. W. Bonk, T. M. Shah).
853. Bayer, DE-OS 3 329 775, 1983 (W. Goyert).
854. C. S. Schollenberger, K. Danbergs, *J. Elastomers Plast.* **7** (1975) 65.
855. L. White, *Eur. Rubber J.* **173** (1991) Oct., 26.
856. Shell, Kraton D and Kraton G Rubbers, Bulletin, no. 09, 1986.
857. G. Holden, N. R. Legge: *Thermoplastic Elastomers; A Comprehensive Review*, Hanser Verlag, New York 1987, p. 47.
858. L. J. Fetter, M. Marton, *Macromolecules* **2** (1969) 453.
859. A. Nostey, J. E. McGrath: *Block Copolymers: Overview and Critical Survey*, Academic Press, New York 1977.
860. P. Zelinski, C. F. Wofford, *J. Polym. Sci. Part A* **3** (1965) 93.
861. Phillips, US 3 560 593, 1971 (H. L. Henry).
862. Shell, US 3 644 588, 1969 (H. L. Howard).
863. Shell, DE 1 905 422, 1968 (H. E. de la Mare, E. F. Bullard).
864. Phillips, US 3 637 554, 1972 (R. D. Lundberg, J. V. Koleske, E. R. Walter).
865. Shell, US 3 700 633, 1972 (M. M. Wald, M. G. Quam).

866. Shell, US 3 595 942, 1971 (M. M. Wald, M. G. Quam).
867. Shell, US 3 810 957, 1965 (H. E. Lunk).
868. C. D. Shedd: *Handbook of Thermoplastic Elastomers*, 2nd ed., Van Nostrand Reinhold, New York 1988, p. 46.
869. Chemische Werke Hüls, DE 1 271 980, 1963 (F. Seifert, J. Bittscheidt, J. Plenikowski).
870. Uniroyal, DE-OS 2 202 706, 1972 (W. K. Fischer).
871. A. J. Tinker, R. D. Icenogle, I. Whittle, *Rubber World* **199** (1989) March, 25.
872. Monsanto, DE 2 848 448, 1977 (M. A. Albert, A.-S. Sabet).
873. Esso Research, US 3 037 954, 1962 (A. M. Gessler).
874. Monsanto, DE 2 632 654, 1975 (A. Y. Coran, B. Das, R. P. Patel).
875. Monsanto, EP 0 004 776, 1978 (A. Y. Coran).
876. Monsanto, DE 2 709 060, 1976 (A. Y. Coran, R. P. Patel).
877. Monsanto, DE 2 726 177, 1976 (A. Y. Coran, R. P. Patel).
878. Monsanto, DE 2 657 109, 1975 (A. Y. Coran, R. P. Patel).
879. Monsanto, EP 0 036 279, 1980 (A. Y. Coran, R. P. Patel).
880. A. Y. Coran, R. Patel, *Rubber Chem. Technol.* **54** (1981) 892. Links
881. D. Perelli, *Technol. Chem.* **11** (1991) no. 6, 80, 94.
882. Monsanto, EP 109 375, 1982 (Y. L. Wang).
883. Monsanto, EP 107 635, 1982 (A.-S. Sabet, K.-S. Shen).
884. Mitsubishi, JP 91 292 311, 1990 (T. Fujita, N. Yamamoto).
885. Union Carbide, EP 417 710, 1989 (M. R. Rifi, K. H. Lee, M. J. Kriss, H. T. Liu).
886. Union Carbide, EP 417 711, 1989 (M. R. Rifi, K. H. Lee, M. J. Kriss, H. T. Liu).
887. R. D. Leaversuch, *Mod. Plast. Int.*, (1992) April, 32.
888. W. O. Murtland, *Rubber World* **166** (1972) May, 37.
889. Toyobo, JP 77 125 597, 1977 (H. Imanaki, T. Sugitoge, Y. Kijima).
890. J. W. Cary, *Gummi Fasern Kunstst.* **39** (1986) 330.
891. S. Fakirov, *Pr. Nauk. Politech. Szezecin* **443** (1991) 23.
892. G. K. Hoeschele, *Chimia* **28** (1974) 544. Links
893. J. R. Wolfe, *MMI Press Symp. Ser.* **3** (1983) 145.
894. J. R. Wolfe, *Adv. Chem. Ser.* **176** (1979) 129. Links
895. R. K. Adams, G. K. Hoeschele: *Thermoplastic Elastomers; A Comprehensive Review*, Hanser Publ., New York 1987, p. 163.
896. G. Deelens: *Thermoplastic Elastomers; A Comprehensive Review*, Hanser Publ., New York 1987, p. 215.
897. W. J. Farrisey, T. M. Shah: *Handbook of Thermoplastic Elastomers*, 2nd ed., Van Nostrand Reinhold, New York 1988, p. 252.
898. *Gummi Fasern Kunstst.* **40** (1987) 440.
899. A. S. Wood, *Mod. Plast. Int.* (1988) April, 48.
900. J. T. Chapin, B. K. Onder, W. J. Farrisey, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **21** (1980) no. 21, 130.
901. Upjohn, US 4 129 715, 1978 (A. T. Chen, W. J. Farrissey, R. G. Nelb).
902. R. G. Nelb, A. T. Chen, K. Onder: *Thermoplastic Elastomers; A Comprehensive Review*, Hanser Publ., New York 1987, p. 197.
903. A. Thorn: *Thermoplastic Elastomers —An Overview*, Rapra Technology Ltd., Shawbury 1996.
904. "USSR Uses Much Less Rubber," *Eur. Rubber J.* **174** (1992) no. 3, 8 – 9.
905. "World Rubber Industry Declines," "Rubber Use to Rise by 3 %," *Eur. Rubber J.* **174** (1992) no. 3, 9.
906. J. D. Carr, P. Jumpasul, H. D. Smit: *The World Rubber Economy Changes and Challenges, A Joint IRSG/ESI Study, May 1988*, IRSG, ESI, London 1988, p. 260.
907. In [906], 202.
908. W. Hofmann, *Kunststoffe* **80** (1990) no. 10, 1204 –1209.
909. B. Davis, *Eur. Rubber J.* **178** (1996) July/Aug.
910. *Rubber and Plastic News*, 15th ed., vol. **21**, 13. *Rubbicana 1992* (Rubber directory and buyer's guide).
911. *Gummi Adreßbuch mit Bezugsquellennachweis*, 12th ed., Curt R. Vincentz Verlag, Hannover 1991.
912. D. J. Schuring (ed.): "Tyre Rolling Resistance," *Rubber Division Symposia*, vol. **1**, American Chemical Society, Chicago, Ill., Oct. 5 – 7, 1982.
913. S. Yamashita, "Progress of Synthetic Rubber Chemistry in Japan," *24th Annual Meeting of the International Institute of Synthetic Rubber Producers*, Kyoto, Japan, May 9 – 13, 1983.
914. A. Ueda, H. Wantanabe, S. Akita, "Structure and Properties of Newly Developed High Vinyl Poly-butadiene Rubbers," *Int. Rubber Conference*, IRC 85, Kyoto, Japan, Oct. 15 – 18, 1985, p. 199 – 204.
915. Japan Synthetic Rubber Co., EP 134 909, priority 17.12. 1980 JP (Y. Takeuchi *et al.*).
916. L. H. Krol, "Synthetic Rubbers: Their Outlook for The Eighties," *Rubbercon 81 [Eighty Ore] Int. Rubber Conference (1981).* vol. **I**, Harrogate, England, June 8 – 12 1981, A 3/1 – A 3/9.
917. K.-H. Nordsiek, K. H. Kiepert, *GAK Kautsch. Gummi Kunstst.* **33** (1980) no. 4, 251 – 255.
918. G. Day, S. Futamura, *Kautsch. Gummi Kunstst.* **40** (1987) no. 1, 39 – 53.
919. Bridgestone Tire Co., EP-A 254 766, priority 27. 08. 1981 JP (M. Ogawa, M. Ikegami).
920. The Goodyear Tire and Rubber Co., EP-A 438 966, priority 28. 12. 1989 US (A. F. Halasa, J. Bergh, F. A. J. Fourgon).
921. Sumitomo Chemical Co., EP 186 985, priority 07. 12. 1984 JP (H. Takao, N. Yoshida, A. Imai).
922. Michelin et Cie., BE 88 25 65, priority 01. 08. 1980 BE.
923. Michelin et Cie., BE 88 25 66, priority 01. 08. 1980 BE.
924. K.-H. Nordsiek: "Model Studies for the Development of an Ideal Tire Tread Rubber," *Meeting of the ACS Rubber Division*, Indianapolis, May 8 – 11, 1984.
925. Bridgestone Corp., EP 476 640, priority 21. 09. 1990 (D. A. Joyner *et al.*).
926. Bridgestone Corp., EP 476 641, priority 21. 09. 1990 (J. W. Kang, T. Hashimoto).
927. Bridgestone Corp., EP 476 662, priority 21. 09. 1990 (J. W. Kang, G. B. Seaver, T. Hashimoto).
928. The Goodyear Tire and Rubber Co., EP-A 263 054, priority 29. 09. 1986 (A. F. Halasa, B. B. Gross, J. L. Cox, G. F. Balogh).
929. The Goodyear Tire and Rubber Co., EP-A 438 967, priority 16. 01. 1990 (W.-L. Hsu, A. F. Halasa).
930. The Goodyear Tire and Rubber Co., EP 483 046, priority 22. 10. 1990 (W.-L. Hsu, A. F. Halasa).
931. Hüls AG, EP-A 256 261, priority 14. 08. 1986 DE (C. Herrmann *et al.*).
932. Hüls AG, EP-A 302 196, priority 28. 07. 1987 DE (C. Herrmann *et al.*).
933. Hüls AG, EP-A 328 774, priority 13. 02. 1988 DE (C. Herrmann *et al.*).

934. Hüls AG, EP-A 344 441, priority 03. 06. 1988 DE (C. Herrmann *et al.*).
935. Hüls AG, EP-A 362 514, priority 05. 10. 1988 DE (W. Hellermann *et al.*).
936. Hüls AG, EP-A 362 515, priority 04. 08. 1989 DE (W. Hellermann *et al.*).
937. Hüls AG, EP-A 384 033, priority 21. 02. 1989 (J. Wolpers, C. Herrmann, W. Hellermann).
938. Shell Internationale Research Maatschappij B.V., EP-A 270 178, priority 02. 12. 1986 GB (A. A. Broekhuis, P. Luijk, J. E. Stamhuis).
939. Sumitomo Chemical Co., EP-A 270 927, priority 29. 11. 1986 JP (Y. Saito *et al.*).
940. Bayer, EP-A 405 216, priority 24. 06. 1989 DE (H.-W. Engels *et al.*).
941. I. R. Gelling, *NR Technol.* **18** (1987) no. 2, 21 – 29.
942. F. Tsutsumi, M. Sakakibara, N. Oshima: "Structure and Dynamic Properties of New Solution SBR Coupled with Tin Compound," *Meeting of the Rubber Division*, American Chemical Society, Cincinnati, Ohio, Oct. 18 – 21, 1988.
943. A. Ueda, T. Ohyama, H. Watanabe, A. Yoshioka: "Vibration Isolation Characteristics of Chemically Modified Solution-Polymerized Rubbers for Engine Mount Use," *Meeting of the Rubber Division*, American Chemical Society, Cincinnati, Ohio, Oct. 18 – 21, 1988.
944. H. Fries, B. Stollfuß: "Structure and Properties of Butadiene Rubber," *Meeting of the Rubber Division*, American Chemical Society, Cincinatti, Ohio, Oct. 18 – 21, 1988.
945. M. Sakakibara, F. Tsutsumi, I. Hattori, Y. Hongu: "Structure and Properties of Diene Rubbers," *Meeting of the Rubber Division*, American Chemical Society, Detroit, Mich. Oct. 8 – 11, 1991.
946. Asahi, EP 252 992, priority 10. 01. 1986 (Y. Kitagawa, H. Yamada, A. Saito).
947. Asahi, EP-A 260 325, priority 13. 03. 1986 (Y. Kitagawa, Y. Hattori, A. Saito).
948. Bayer, EP 218 876, priority 11. 09. 1985 (B. Stollfuß, W. Wieder, W. Nentwig).
949. Bridgestone Corp., EP-A 299 074, priority 14. 01. 1987 (T. Hamada, H. Fukuoka, H. Komatsu, T. Fujmaki).
950. Bridgestone/Firestone, Inc., EP-A 452 603, priority 09. 04. 1990 US (T. A. Antkowiak, D. F. Lawson, R. W. Koch, M. L. Stayer).
951. Bridgestone/Firestone, Inc., EP-A 451 604, priority 09. 04. 1990 US (M. L. Stayer, T. A. Antkowiak, D. F. Lawson, R. W. Koch).
952. The Goodyear Tire and Rubber Co., EP-A 316 255, priority 15. 10. 1987 US (A. F. Halasa, B. A. Matrana, S. E. Robertson-Wilcox).
953. Nippon Zeon Co., EP 180 141, priority 26. 10. 1984 JP (H. Watanabe, K. Noguchi, T. Kase, S. Akita).
954. Nippon Zeon Co., EP 180 853, priority 26. 10. 1984 JP (H. Watanabe, K. Noguchi, T. Kase, A. Shuichi).
955. The Yokohama Rubber Co., Nippon Zeon Co., DE-OS 3 635 366, priority 23. 10. 1985 JP (Y. Kikuchi et al.).
956. Sumitomo Chemical Co., EP-A 341 496, priority 02. 05. 1988 (A. Imai, T. Seki, K. Yamamoto).
957. Bayer, EP-A 7027, priority 08. 07. 1978 DE (G. Sylvester, J. Witte, G. Marwede).
958. Bayer, EP-A 11 184, priority 11. 11. 1978 DE (G. Sylvester, J. Witte, G. Marwede).
959. Enoxy Chimica, EP-A 92 270, priority 15. 04. 1982 IT (A. Carbonaro, D. Ferraro).
960. Enichem. Polimeri, EP-A 127 236, priority 25. 05. 1983 IT (A. Carbonaro, S. Gordini, S. Cucinella).
961. Enichem. Elastomeri, EP 201 962, priority 15. 05. 1985 IT (A. Carbonaro, S. Gordini, S. Cucinella).
962. Enichem. Elastomeri, EP 201 979, priority 15. 05. 1985 IT (A. Carbonaro, S. Gordini, S. Cucinella).
963. Enichem. Elastomeri, EP-A 204 373, priority 05. 06. 1985 IT (A. Carbonaro, D. Ferraro).
964. Enichem. Elastomeri, EP 207 540, priority 05. 06. 1985 IT (S. Gordini).
965. Enichem. Elastomeri, EP 207 558, priority 01. 07. 1985 IT (S. Gordini, A. Carbonaro, S. Spina).
966. Enichem. Elastomers, EP-A 375 421, priority 22. 12. 1988 (D. K. Jenkins, P. J. Ansell).
967. Enichem. Elastomeri, EP-A 386 808, priority 17. 02. 1989 IT (P. Andreussi, R. Bianchi).
968. The Goodyear Tire and Rubber Co., EP 236 253, priority 19. 02. 1986 US (M. C. Throckmorton, R. W. Stachowiak, C. L. Wilson).
969. Japan Synthetic Rubber Co., DE 3 224 288, priority 29. 06. 1981 JP (Y. Takeuchi, M. Sakakibara, T. Shibata).
970. Japan Synthetic Rubber Co., EP-A 406 920, priority 05. 09. 1988 JP (I. Hattori, M. Sakakibara, F. Tsutsumi, M. Yoshizawa).
971. Polysar Ltd., EP 232 592, priority 18. 11. 1975 US (D. C. Edwards, J. A. Crossman).
972. The Firestone Tire and Rubber Co., EP 100 434, priority 14. 07. 1982 US (G. G. A. Böhm).
973. Bridgestone Corp., EP-A 470 693, priority 11. 06. 1990 JP (M. Araki, Y. Morimoto).
974. General Tire Inc., EP 263 216, priority 02. 09. 1986 US (J. L. Palombo, R. L. Miller).
975. The Goodyear Tire and Rubber Co., EP 231 143, priority 23. 01. 1986 US (A. J. Bell).
976. The Goodyear Tire and Rubber Co., EP-A 357 535, priority 24. 08. 1988 US (G. L. Burroway, G. F. Magoun, R. N. Gujarathi).
977. The Goodyear Tire and Rubber Co., EP-A 359 694, priority 14. 09. 1988 US (R. M. Scriver, W.-L. Hsu, A. F. Halasa).
978. The Goodyear Tire and Rubber Co., EP-A 434 596, priority 21. 12. 1989 US (B. R. Hahn, W.-L. Hsu, D. D. Callander, A. F. Halasa).
979. The Goodyear Tire and Rubber Co., EP-A 468 914, priority 26. 07. 1990 US (D. B. Patterson, A. F. Halasa).
980. The Goodyear Tire and Rubber Co., EP 476 305, priority 23. 08. 1990 (W. C. T. Tung *et al.*).
981. The Goodyear Tire and Rubber Co., EP 477 635, priority 17. 09. 1990 (G. Cohn, D. D. Callander, W. C. T. Tung).
982. The Goodyear Tire and Rubber Co., EP-A 470 040, priority 02. 08. 1990 US (G. O. Schulz, D. K. Parker).
983. The Goodyear Tire and Rubber Co., EP 482 478, priority 24. 10. 1990 (W.-L. Hsu, A. F. Halasa).
984. Hüls AG, EP-A 258 554, priority 19. 08. 1986 DE (R. Ollenik, W. Kleinert, M. Müller).
985. Industrie Pirelli S.p.A., GB 2 122 110, priority 24. 06. 1982 IT (P. L. Robicci).
986. Pirelli Coordinamento Pneumatici S.p.A., EP 408 941, priority 21. 07. 1989 IT (M. Mezzanotte).
987. Pirelli Coordinamento Pneumatici S.p.A., EP 454 025, priority 24. 04. 1990 IT (P. Borgonovo, G. Ghilardi).
988. Sumitomo Rubber Industries, EP-A 426 388, priority 26. 10. 1989 JP (S. Takatsu, K. Muraoka).
989. Michelin et Cie., DE-OS 3 149 716, priority 16. 12. 1980 FR (P. Durif).
990. Michelin et Cie., DE-OS 3 209 581, priority 16. 03. 1981 FR (C. Ancel, P. Philibert).
991. Michelin et Cie., EP 220 642, priority 24. 10. 1985 FR (J. Gouttebess, P. Villaine).
992. Höchst AG, EP 149 836, priority 07. 01. 1984.
993. Mitsubishi Rayon Co. Ltd., EP-A 423 759, priority 18. 10. 1989 JP (Y. Fukuda, H. Inada, H. Sato).
994. E. I. Du Pont, EP-A 426 019, priority 31. 10. 1989 US (F. Y. Kafka).
995. Du Pont, EP-A 426 024, priority 31. 10. 1989 US (F. Y. Kafka).
996. Bayer, EP-A 255 632, priority 02. 08. 1986 DE (J. Peter, G. Pampus).
997. Polysar Ltd., EP 265 671, priority 29. 09. 1986 US (J. Odar, W. von Hellens).
998. Bridgestone Tire Co., Japan Synthetic Rubber Co., DE 3 228 547, priority 31. 07. 1981 JP (Y. Takeuchi et al.).
999. The Firestone Tire and Rubber Comp., EP 139 848, priority 06. 09. 1983 US (D. F. Graves).
1000. The Firestone Tire and Rubber Corp., EP 197 285, priority 21. 03. 1985 US (W. W. Schloman, J. A. Davis).
1001. Bridgestone Corp., DE-OS 3 628 844, priority 27. 08. 1985 JP (E. Takiguchi, T. Kikutsugi).
1002. The Firestone Tire and Rubber Corp., EP 234 303, priority 14. 02. 1986 US (S. Futamura).

1003. Continental Gummi-Werke AG, Deutsches Institut Kautschuktechnologie e.V., DE-OS 3 626 240, priority 02. 08. 1986 DE (J. Peter, G. Pampus).
1004. Continental AG, DE 3 720 461, priority 20. 06. 1987 DE (R. Mahling, W. Klapp, W. Holzbach).
1005. Continental AG, DE-OS 3 740 303, priority 27. 11. 1987 DE (G. Kaiser, J. Peter).
1006. Continental AG, EP-A 366 915, priority 31. 10. 1988 DE (H. Stevens, G. Kaiser).
1007. The Goodyear Tire and Rubber Corp., EP-A 448 904, priority 28. 03. 1990 US (J. K. Clark).
1008. Hüls AG, EP 168 524, priority 05. 07. 1984 DE (K.-H. Nordsiek, H. Schwesig).
1009. Hüls AG, EP 173 791, priority 18. 04. 1984 DE (K.-H. Nordsiek, J. Wolpers, J. Tornau).
1010. Hüls AG, EP-A 333 999, priority 23. 03. 1988 DE (R. Tappe).
1011. Japan Synthetic Rubber Co., Bridgestone Corp., DE-OS 3 636 970, priority 30. 10. 1985 JP (F. Tsutsumi et al.).
1012. Japan Synthetic Rubber Co., Bridgestone Corp., EP-A 328 284, priority 01. 02. 1988 JP (M. Sakakibara, F. Tsutsumi, M. Kondo, Y. Yagi).
1013. Japan Synthetic Rubber Co., EP-A 405 904, priority 26. 09. 1989 JP (M. Kondo, I. Sugita, I. Furuta, F. Tsutsumi).
1014. The Goodyear Tire and Rubber Co., EP 477 682, priority 24. 09. 1990 (T. D. Linster).
1015. Pirelli Coordinamento Pneumatici S.p.A., EP-A 448 112, priority 23. 03. 1990 IT (L. Garro, M. Martin).
1016. Sumitomo Chemical Co., EP 263 885, priority 10. 04. 1986 JP (H. Takao *et al.*).
1017. Uniroyal Chemical Co., EP-A 342 288, priority 20. 05. 1988 (J. M. Mitchell, D. E. Wingrove, C. F. Eckert).
1018. Bayer, EP-A 313 917, priority 24. 10. 1987 and 02. 06. 1988 DE (W. Obrecht, Z. Szentivanyi).
1019. The Firestone Tire and Rubber Comp., EP 151 212, priority 09. 01. 1984 US (T. F. Donatelli, P. F. Murray).
1020. Bridgestone Corp., DE 3 540 456, priority 15. 11. 1984 JP (Y. Hirata *et al.*).
1021. The Firestone Tire and Rubber Co., EP 177 695, priority 21. 09. 1984 US (D. P. Tate, D. F. Graves).
1022. The Firestone Tire and Rubber Co., EP-A 287 806, priority 13. 04. 1987 (D. F. Graves, H. W. Engels).
1023. Bridgestone Corp., EP 328 291, priority 02. 02. 1988 (T. Yamada, R. Fujio).
1024. Bridgestone Corp., EP-A 333 317, priority 16. 02. 1988 (I. Nakajima, K. Hatakeyama, Y. Hirata).
1025. Bridgestone Corp., EP-A 405 827, priority 29. 06. 1989 JP (T. Saito, J. Yatsunami, I. Wada, T. Takizawa).
1026. Continental Gummi-Werke AG, DE-OS 3 033 711, priority 08. 09. 1980 DE (R. Sohnemann).
1027. Semperit Reifen Aktiengesellschaft, EP-A 406 214, priority 28. 06. 1989 AT (O. Beckmann).
1028. The Goodyear Tire and Rubber Co., EP 149 956, priority 06. 01. 1984 US (P. H. Sandstrom).
1029. The Goodyear Tire and Rubber Co., EP-A 335 817, priority 29. 03. 1988 US (L. G. Wideman, G. F. Castner).
1030. The Goodyear Tire and Rubber Co., EP-A 363 302, priority 06. 10. 1988 US (S. K. Mowdood, B. K. Kansupada).
1031. Hüls AG, EP-A 385 072, priority, 1989 DE (G. Hörpel, K.-H. Nordsiek).
1032. Hüls AG, EP-A 432 416, priority 12. 12. 1989 DE (G. Hörpel, H. G. Haag, K.-H. Nordsiek).
1033. Sumitomo Rubber Industries, EP-A 386 905, priority 04. 03. 1989 GB (B. D. W. Powell).
1034. Sumitomo Chemical Co., EP-A 345 594, priority 07. 06. 1988 (I. Kurimoto, N. Inui, T. Yamaguchi).
1035. Sumitomo Chemical Co., EP-A 409 565, priority 17. 07. 1989 JP (N. Inui, H. Nagasaki, S. Yachigo, M. Oikawa).
1036. The Goodyear Tire and Rubber Co., EP 476 310, priority 28. 08. 1990 (L. G. Wideman, P. H. Sandstorm, G. F. Balogh, D. J. Keith).
1037. Continental AG, DE-OS 3 924 528, priority 25. 07. 1989 DE (H. Stevens, R. Teves).
1038. Continental AG, DE-OS 3 924 529, priority 25. 07. 1989 DE (H. Stevens, R. Teves).
1039. Continental AG, DE-OS 3 924 530, priority 25. 07. 1989 DE (H. Stevens, R. Teves).
1040. Continental AG, EP-A 410 311, priority 25. 07. 1989 DE (H. Stevens, R. Teves).
1041. Nippon Zeon Co., EP-A 420 449, priority 18. 09. 1989 JP (Y. Saito).
1042. Degussa AG, EP-A 468 140, priority 25. 07. 1990 DE (S. Wolff, U. Göre).
1043. M. Nedev, E. Karadimova, *GAK Kautsch. Gummi Kunstst.* **44** (1991) no. 12, 1133.
1044. M. Nedev, E. Karadimova, *GAK Kautsch. Gummi Kunstst.* **45** (1992) no. 1, 30 – 31.
1045. Compagnie Générale Des Etablissements Michelin – Michelin & Cie, EP-A 501 227, priority 25. 02. 91 (R. Rauline)
1046. G. W. Marwede, B. Stollfuß, A. J. M. Sumner: “Current Status of Tyre Elastomers in Europe,” *IISRP Meeting*, Cancon/Mexico, May 4 – 6, 1992.
1047. Polysar Ltd., US 4 645 793, priority 19. 12. 1985 US (W. von Hellens, Sheikh A. H. Mohammed, R. Hallman).
1048. Bridgestone Corporation, US 4 886 850, priority 17. 06. 1986 JP (M. Ogawa, Y. Shiomura, T. Takizawa).
1049. Bridgestone/Firestone Inc., EP-A 463 205, 1990 (A. E. Oberster, C. M. Roland, G. G. A. Böhm).
1050. The Goodyear Tire and Rubber Co., EP 240 448, priority 05. 03. 1986 US (T. J. Botzman).
1051. The Goodyear Tire and Rubber Co., EP-A 448 903, priority 26. 03. 1990 US (R. J. Gartland, A. F. Finelli, A. J. Bell).
1052. Sumitomo Chemical Co., EP-A 447 191, priority 12. 03. 1990 JP (Y. Keisaku *et al.*).
1053. Uniroyal Chemical Co., US 4 973 627, priority 13. 10. 1989 US (J. M. Mitchell).
1054. Mitsui Petrochemical Industries Ltd., EP-A 446 480, priority 05. 10. 1989 JP (H. Nakahama, T. Mishima).
1055. Mitsui Petrochemical Industries Ltd., EP-A 446 382, priority 05. 10. 1989 JP (H. Nakahama, T. Mishima).
1056. W. von Hellens, D. C. Edwards, Z. J. Lobos, *Rubber and Plastic News*, (1990) Oct. 1, 61 – 64.
1057. Polysar Ltd., US 4 665 963, priority 19. 10. 1984 US (J. Timar, Sheikh A. H. Mohammed, J. Walker).
1058. Polysar Ltd., US 4 614 772, priority 26. 09. 1985 US (D. C. Edwards, D. Padlija).
1059. Polysar Ltd., EP-A 314 416, priority 28. 10. 1987 US (Sheikh A. H. Mohammed, C. W. von Hellens).
1060. Exxon Research and Engineering Co., US 4 384 072, priority 30. 09. 1981 US (N. F. Newman, R. C. Kowalski).
1061. Exxon Research and Engineering Co., US 4 632 963, priority 05. 04. 1984 US (I. J. Gardner *et al.*).
1062. Exxon Chemical Patents Inc., EP-A 320 263, priority 11. 12. 1987 US (K. W. Powers, H.-C. Wang, D. C. Handy, J. V. Fusco).
1063. Exxon Chemical Patents Inc., EP-A 344 021, priority 27. 05. 1988 US (W. K. Powers, H.-C. Wang).
1064. Exxon Chemical Patents Inc., WO 91/18050, priority 14. 05. 1990 (D. F. Kruse, R. C. Keller).
1065. Exxon Chemical Patents Inc., WO 91/19761, priority 15. 06. 1990 (J. V. Fusco, D. F. Kruse).
1066. G. J. Wilson, K. Richie, T. Dolezal: “Compounding Elastomer of Limited Cure Functionality,” *Rubber Division*, American Chemical Society, Montreal, Quebec, May 2 – 5, 1978.
1067. D. Padlija: “Developmental Butyl Rubbers, Part I – A Process for Producing Low Molecular Weight Rubber,” *ACS Rubber Div. Mtg.*, Toronto, May 1991.
1068. A. J. Van Tongerloo, I. Spadola, R. Steevensz, “Developmental Butyl Rubbers, Part II: Preparation of Bromobutyl with Broad Molecular Weight Distribution,” *ACS Rubber Div. Mtg.*, Toronto, May 1991.
1069. K. J. Kumbhani, J. J. Sendorck, H. Tan: “Developmental Butyl Rubbers, Part III: Processing and Performance Characteristics of Bromobutyl Rubbers With Broad Molecular Weight Distributions,” *ACS Rubber Div. Mtg.*, Toronto, May 1991.
1070. J. Machurat, Y. Bomal: “How to Increase the Properties of a Bromobutyl Based Innerliner by the Use of a Precipitated Silica?” *Kautsch. Gummi Kunstst.* **45** (1992) no. 1, 40 – 46.

1071. "Exxon Launches Bromo Rubbers," *Eur. Rubber J.* **174** (1992) March, 18.
1072. Polysar Ltd., EP 180 444, priority 29. 10. 1984 US (D. C. Edwards, D. Padlija).
1073. The Firestone Tire and Rubber Co., DE 2 900 800, priority 13. 01. 1978 US (A. F. Halasa, J. E. Hall).
1074. The Firestone Tire and Rubber Co., EP 166 849, priority 04. 06. 1984 US (I. W. Kang, J. E. Hall).
1075. The Firestone Tire and Rubber Co., EP 167 701, priority 04. 06. 1984 US (I. W. Kang).
1076. Bridgestone/Firestone Inc., EP-A 390 012, priority 30. 03. 1989 US (D. F. Graves, C. M. Roland, G. G. A. Böhm).
1077. Bridgestone Corp., EP-A 476 655, priority 21. 09. 1990 US (J. W. Kang, G. B. Seaver, T. Hashimoto).
1078. The Goodyear Tire and Rubber Co., EP 155 897, priority 27. 02. 1984 US (D. B. Patterson, D. K. Parker, L. D. Jr.
1079. The Goodyear Tire and Rubber Co., EP 269 546, priority 20. 10. 1986 US (R. J. Hopper, C. R. Parks).
1080. The Goodyear Tire and Rubber Co., EP-A 425 425, priority 23. 10. 1989 US (R. W. Stachowiak, A. J. Puccio, T. J. Eberhart).
1081. Hüls AG, EP-A 365 792, priority 28. 10. 1988 DE (G. Hörpel, J. Pangratz).
1082. Hüls AG, EP 248 175, priority 31. 05. 1986 DE (W. Hellermann, K.-H. Nordsiek, J. Wolpers, P. Sunder-Plassmann).
1083. Japan Synthetic Rubber Co., EP 152 175, priority 12. 01. 1984 JP (H. Ono, Y. Matsumura, E. Okuya).
1084. Shell Internationale Research Maatschappij B.V., EP-A 437 293, priority 09. 01. 1990 GB (S. M. de Vries, J. A. Willemsen).
1085. Sumitomo Chemical Co., EP-A 433 943, priority 18. 12. 1989 JP (M. Kakugo *et al.*).
1086. A. Jemain, Michelin, les Secrets du "Pneu Vert," Economiste no. 825 (20. 12. 1991) 44 – 45.
1087. F. Meier, *Wirtschaftswoche, Unternehmen + Märkte*, **4** (1992) Jan. 17, 98 – 99.
1088. Y. Tsuzuki: "Current Trend of Japanese Automotive Requirements for Elastomers", *24th Annual Meeting Proceedings, II-3, IISRP Meeting,* Kyoto, May 1983.
1089. G. Walter: *Kunststoffe und Elastomere in Kraftfahrzeugen*, W. Kohlhammer Verlag, Stuttgart 1985.
1090. W. Behrens: "Elastomere in VW Fahrzeugen, Werkstoffeinsatz und künftige Anforderungen," *DKG-Meeting*, Bad Homburg, Sept. 19, 1990.
1091. E. Okuya: "Present Situation and Future Requirements for Elastomers in the Japanese Automotive Industry," *Assogomma Meeting*, Milan, June 16, 1983.
1092. P. S. Johnson: "The Future of Non-tire Automotive Applications for Rubber," *Automot. Polym. Des.* (1987) June 5 – 6, 25.
1093. R. Casper, E. Rohde, *GAK Kautsch. Gummi Kunstst.* **41** (1988) no. 6, 541 – 548.
1094. J. R. Dunn: "Speciality Elastomers in Underhood Applications," *IISRP Meeting*, Capetown, South Africa, April 25, 1990.
1095. T. Hamanaka, *Int. Polym. Sci. Technol.* **15** (1988) no. 2, T 30 – T 35.
1096. G. P. Streit, *Kautsch. Gummi Kunstst.* **41** (1988) no. 9, 908 – 915.
1097. H. Hill, H.-V. Buchholz, G. Kirchmann: "Elastomer-Dichtsysteme in Kraftfahrzeug-Karosserien," *DKG-Meeting*, Bad Homburg, Sept. 19 – 20, 1990.
1098. S. Lüpfert, *3. Fachtagung des Süddeutschen Kunststoffzentrums über Hochtemperaturbeständige Polymere Werkstoffe*, Würzburg, April 1988.
1099. G. Heinz: "A Powerful Drive in Belts," *Eur. Rubber J.* **173** (1991) June, 31.
1100. The United States of America as represented by the Secretary of the Army, Washington, D.C.,US 4 843 114, priority 29. 01. 1988 (P. Touchet et al.).
1101. M. Aonuima, "Recent Trends in Elastomer Blends," *Int. Polym. Sci. Techn.* **17** (1990) no. 8, T/40 – T/49.
1102. "VW-Golf: Kälteflexibler Thermoplast vervielfacht Dichtungsstulpfen-Lebensdauer," *Kautsch. Gummi Kunstst.* **45** (1992) no. 1, 8.
1103. B. Davis, S. Robinson, "Speciality Elastomers: Niche Headed for Growth Explosion," *Eur. Rubber J.* **171** (1989) March, 20 – 21.
1104. BASF, EP 295 675, priority 19. 06. 1987 DE (K. Bronstert, H.-M. Walter, H.-D. Schwaben).
1105. BASF, EP-A 240 903, priority 08. 04. 1986 DE (K. Bronstert, H. Hoenl, A. Echte, P. Klaerner).
1106. Enichem. Elastomeri S.p.A., EP 161 021, priority 18. 04. 1984 IT (A. Roggero, T. Salvatori).
1107. BASF, EP-A 423 571, priority 14. 10. 1989 DE (D. Bender *et al.*).
1108. BASF, EP-A 288 761, priority 08. 04. 1987 DE (K. Bronstert, S. Bohnet).
1109. The DOW Chemical Comp., EP-A 277 687, priority 28. 01. 1987 NL (N. Alle, H. B. Savenije, T. M. Roovers).
1110. The DOW Chemical Comp., EP-A 418 042, priority 12. 09. 1989 GB (M. Demirörs).
1111. BASF, EP 167 707, priority 16. 03. 1984 DE (H. Hoenl, A. Echte).
1112. BASF, EP 283 731, priority 25. 02. 1987 DE (K. Bronstert, H.-D. Schwaben, A. Echte).
1113. BP Chemicals Ltd., EP-A 329 283, priority 21. 01. 1988 FR (O. Lamblin).
1114. The DOW Chemical Comp., EP-A 412 801, priority 09. 09. 1989 GB (L. M. Aerts, C. T. E. van Nuffel, N. Alle).
1115. The DOW Chemical Comp., EP-A 352 919, priority 26. 07. 1988 US (D. E. Henton).
1116. Hoechst Celanese Corp., EP-A 282 153, priority 12. 07. 1987 US (T. C. Yu, T. J. Dolce).
1117. Du Pont, EP-A 358 388, priority 26. 08. 1988 US (D. E. Swaris).
1118. Enichem. Synthesis S.p.A., EP-A 434 130, priority 21. 12. 1989 IT (A. Villanti, C. Caredda, V. Messori).
1119. Shin-Etsu Chemical Co., Ltd., EP-A 355 575, priority 10. 08. 1988 JP (T. Shimizu, I. Kaneko, M. Watanabe).
1120. Shin-Etsu Chemical Co., Ltd., EP-A 382 102, priority 02. 02. 1989 JP (S. Ueno, I. Kaneko, M. Watanabe).
1121. Shin-Etsu Chemical Co., Ltd., EP-A 397 009, priority 02. 05. 1989 JP (S. Ueno, I. Kaneko, M. Watanabe).
1122. Shin-Etsu Chemical Co., Ltd., EP-A 460 458, priority 25. 05. 1990 JP (T. Shimizu, T. Sato).
1123. Bayer, US 4 814 401, priority 03. 10. 1985 DE (W. Obrecht, P. Wendling, W. Göbel, E. Müller).
1124. Bayer, US 4 920 189, 1987 (W. Obrecht, P. Wendling, W. Göbel, E. Müller).
1125. Du Pont, EP-A 415 337, priority 28. 08. 1989 US (E. W. Slocum, A. C. Sobrero, R. C. Wheland).
1126. Bayer, EP-A 341 499, priority 10. 05. 1988 DE (B. A. Wolf *et al.*).
1127. Bayer, EP-A 374 660, priority 23. 12. 1988 DE (W. Baade, W. Obrecht, C. Ohm).
1128. Du Pont, EP-A 460 936, priority 06. 06. 1990 US (P. B. Hatch, D. R. Marshall).
1129. Hoechst AG, EP-A 426 017, priority 28. 10. 1989 DE (C. D. Eisenbach, U. Bülow, W. Lieberth, F. Fischer).
1130. B. A. Krentsel, L. A. Nekhaeva: "Metalcomplex Catalysis of the Polymerization of 1-Olefins," *Russ. Chem. Rev. (Engl. Transl.)* **59** (1990) no. 12, 1193 – 1207.
1131. Du Pont, US 4 698 403, priority 15. 10. 1985 US (U. Klabunde).
1132. Himont Incorporated, EP-A 384 171, priority 31. 01. 1989 IT (L. Resconi, U. Giannini, E. Albizzati).
1133. Exxon Chemical Patents Inc., EP 347 129, priority 16. 06. 1988 US (S. Floyd, E. L. Hoel).
1134. F. Leibrandt, *Kautsch. Gummi Kunstst.* **48** (1995) 201.
1135. G. Stella, J. R. Schauder, "Advances in EPDM structure design," Rubbercon 1995, Göteborg.
1136. R. T. Sylvest, G. Lancaster, S. Betso, "New Elastomers From Insite Technology: Molecular Architecture Control And Development Process Innovation, " 2nd International Congress on Metallocene Polymers, Metallocenes 1996, Düsseldorf, March 1996.
1137. J. L. Laird, J. A. Riedel, "Evaluation of EPDM Materials as Produced by Constrained Geometry Catalyst Chemistry Against Current Commercial EPDM Products and Performance Rquirements," ACS Rubber Division 148th Technical Meeting, Cleveland, October 1995.
1138. J. C. Stevens, "Insite Catalyst Structure – Activity Relationship for Olefin Polymerisation," METCON 1993.

1139. Y-C Hwang, S. Chum, R. Guerra, K. Sehanobish, “Morphology and Deformation Behaviour of Homogeneous Polyolefin Copolymers made with Insite Technology,” 3rd International Business Forum on Specialty Polyolefins, SPO Conference, September 1993.
1140. M. S. Edmondson, T. P. Karjala, R. M. Patel, D. R. Parikh, “Ethylene–Propylene–Diene Elastomers via Constrained Geometry Catalyst Technology, ” ACS Rubber Division 149th Technical Meeting, Montreal, May 1996.
1141. F. K. Jones, J. L. Laird, B. W. Smith, “Characterisation of EPDMs Produced by Constrained Geometry Catalysts,” ACS Rubber Division 149th Technical Meeting, Montreal, May 1996.
1142. F. Leibbrandt, *Kautschuk Gummi Kuststoffe* **48** (1995) 201.
1143. G. Stella, J. R. Schauder, “Advances in EPDM structure design,” Rubbercon '95, Gothenburg.
1144. R. T. Sylvest, G. Lancaster, S. Besto, “New Elastomers from Insite Technology: Molecular Architecture Control and Development Process Innovation,” 2nd International Congress on Metallocene Polymers, Metallocene '96, Düsseldorf, March 1996.
1145. J. L. Laird, J. A. Riedel, “Evaluation of EPDM Materials as Produced by Constrained Geometry Catalyst Chemistry Against Current Commercial EPDM Products and Performance Requirements,” ACS Rubber Division 148th Technical Meeting, Cleveland, October 1995.
1146. J. C. Stevens, “Insite Catalyst Structure / Activity Relationships for Olefin Polymerisation,” METCON '93.
1147. Y.-C. Hwang, S. Chum, R. Guerra, K. Sehanobish, “Morphology and Deformation Behaviour of Homogeneous Polyolefin Copolymers made with Insite Technology,” 3rd International Business Forum on Speciality Polyolefines, SPO Conference, September 1993.
1148. M. S. Edmondson, T. P. Karjala, R. M. Patel, D. R. Parikh, “Ethylene–Propylene–Diene Elastomers via Constrained Geometry Catalyst Technology,” ACS Rubber Division 149th Technical Meeting, Montreal, May 1996.
1149. F. K. Jones, J. L. Laird, B. W. Smith, “Characterisation of EPDMs Produced by Constrained Geometry Catalysts,” ACS Rubber Division 149th Technical Meeting, Montreal, May 1996.
1150. H. Imamura, A. Ohmura, E. Haku, S. Tsudiya, *J. Catal.* **96** (1985) 139. Links
1151. Tosoh Corp., EP-A 453 205, priority 16. 04. 1990 JP (K. Soga, Y. Naito).
1152. Bayer, EP-A 305 800, priority 29. 08. 1987 DE (R. Fauss, M. Seidler, J. Witte).
1153. BASF, EP-A 381 973, priority 27. 01. 1989 DE (M. Buback, S. Rohde).
1154. Nippon Zeon Co., Ltd., EP-A 324 860, priority 29. 06. 1987 JP (M. Aonuma, H. Sakakida, Y. Aimura).
1155. Shell Internationale Research Maatschappij B.V., EP-A 410 546, priority 28. 07. 1989 US (P. K. Wong).
1156. Polysar Ltd., EP 111 412, priority 08. 12. 1982 CA (G. L. Rempel, H. Azizian).
1157. University of Waterloo, EP-A 298 386, priority 06. 07. 1987 US (G. L. Rempel, N. A. Mohammadi, R. Farwaha).
1158. Kuraray Co., Ltd., EP-A 319 967, priority 07. 12. 1987 JP (M. Ishii, H. Takamatsu, K. Ishiura, S. Minatono).
1159. Japan Synthetic Rubber Co. Ltd., EP-A 366 377, priority 22. 10. 1988 JP (N. Sakabe, T. Ohara, T. Miyabayashi, H. Ise).
1160. Shell Internationale Research Maatschappij B.V., EP-A 360 356, priority 22. 09. 1988 US (C. L. Willis, L. A. Pottick, D. E. Goodwin).
1161. Shell Internationale Research Maatschappij B.V., EP-A 368 381, priority 09. 11. 1988 US (R. J. Hoxmeier, L. H. Slaugh).
1162. Bayer, EP-A 405 266, priority 29. 06. 1989 DE (H. Buding *et al.*).
1163. University of Waterloo, EP-A 455 154, priority 01. 05. 1990 CA (G. L. Rempel, N. T. McManus, N. A. Mohammadi).
1164. Nippon Zeon Co. Ltd., DE-OS 4 110 349, priority 30. 03. 1990 JP (K. Takahashi, Y. Kubo).
1165. Nippon Zeon Co. Ltd., US 5 013 796, priority 24. 03. 1988 JP (S. Hayashi, Y. Kubo, N. Watanabe, Y. Aimura).
1166. Enichem. Elastomers Ltd., EP 312 213, priority 22. 09. 1987 GB (D. G. Timms).
1167. Polysar Ltd., EP-A 325 198, priority 19. 01. 1988 US (A. Osman, W. G. Bradford).
1168. Polysar Ltd., EP-A 346 778, priority 13. 06. 1988 US (D. T. Ahlberg, D. Padliya, J. T. Reed).
1169. Polysar Rubber Corp., EP-A 429 017, priority 20. 11. 1989 US (W. G. Bradford, G. J. Arsenault, A. J. Marshall).
1170. Enichem Elastomeri, EP 209 253, priority 17. 06. 1985 (M. Bruzzone, S. Gordini, K. Wyllie).
1171. Exxon Research and Engineering Company, EP-A 117 635, priority 27. 01. 1983 US (K. W. Powers, H.-C. Wang).
1172. The DOW Chemical Company, US 3 898 208, priority 29. 06. 1973 US (R. L. Krause).
1173. The Goodyear Tire and Rubber Company, EP 117 221, priority 22. 02. 1983 US (L. G. Wideman).
1174. The Goodyear Tire and Rubber Company, EP-A 279 766, priority 19. 02. 1987 US (L. G. Wideman, W.-L. Hsu).
1175. Olin Corporation, WO 90/13578, priority 11. 05. 1989 US (H. W. Schiessl, S. A. Manke).
1176. Exxon Research and Engineering Co., US 415 924, priority 31. 05. 1978 US (B. R. Tegge, R. H. Schatz, D. M. Beeson, L. S. Minckler).
1177. Exxon Research and Engineering Co., US 4 303 766, priority 12. 12. 1977 US (C. P. O'Farrell, F. N. Kresge).
1178. D. K. Parker, H. W. Schiessl, R. F. Roberts: “A New Process for the Preparation of Highly Saturated Nitrile Rubber in Latex Form,” *139th Meeting of the Rubber Division*, Americal Chemical Society, Toronto, Canada, May 21 – 24, 1991.
1179. Hüls AG, DE-OS 3 606 745, priority 01. 03. 1986 DE (B. Stützel, H.-J. Smigerski, M. Müller, G. Stevermüer).
1180. Bayer, EP-A 374 664, priority 23. 12. 1988 DE (H. Sutter, A. Kolwert, W. Obrecht).
1181. Bayer, EP-A 374 662, priority 23. 12. 1988 DE (W. Obrecht *et al.*).
1182. Bayer, EP-A 380 938, priority 24. 01. 1989 DE (M. Buback *et al.*).
1183. Polysar Ltd., EP 111 391, priority 10. 11. 1982 CA (D. C. Edwards).
1184. Bayer, EP-A 157 079, priority 26. 01. 1984 DE (H. Fischer).
1185. Bayer, EP-A 419 952, priority 26. 09. 1989 DE (W. Obrecht, S. Szentivanyi, H. Waniczek).
1186. Bayer, EP-A 420 006, priority 29. 09. 1989 DE (H.-H. Greve, K.-P. Meurer, W. Obrecht, R. Musch).
1187. The Goodyear Tire and Rubber Company, EP-A 464 478, priority 02. 07. 1990 US (R. J. Hopper).
1188. Naphthachimie Société, US 4 343 926, priority 06. 07. 1979 FR (F. Caumartin, J. L. Vidal, P. Mangin).
1189. BP Chemicals Ltd., EP 211 624, priority 06. 08. 1985 FR (J. Collomb, D. C. Durand, L. Havas, F. R. M. M. Morterol).
1190. Union Carbide Chemicals and Plastics Company, Inc., EP-A 417 710, priority 11. 09. 1989 US (M. R. Rifi, K. H. Lee, M. J. Kriss, H.-T. Liu).
1191. Union Carbide Chemicals and Plastics Company, Inc., EP-A 417 711, priority 11. 09. 1989 US (M. R. Rifi, K. H. Lee, M. J. Kriss, J.-T. Liu).
1192. Du Pont, EP-A 451 998, priority 02. 04. 1990 US (L. Ahramjan, C. W. Jr. Halet)
1193. Du Pont, EP-A 426 023, priority 31. 10. 1989 US (F. Y. Kafka, A. R. Bice, D. K. Burchett).
1194. Bayer, US 4 769 407, priority 13. 11. 1985 DE (W. Obrecht, P. Wendling, W. Göbel, E. Müller).
1195. Shell Internationale Research Maatschappij B.V., EP-A 353 802, priority 11. 07. 1988 GB (J. P. M. Peters, G. Van Os, S. M. de Vries).
1196. Atochem North America, Inc., EP-A 460 284, priority 29. 05. 1990 US (N. Kydonieus, J. A. Hall, J. D. Phillips, J. A. Sturm).
1197. J. Dunkel: “Anwendungsmöglichkeiten der Membrantechnik in der chemischen Industrie,” *3. Fachtagung zu Theorie und Praxis der Membrantrennprozesse*, Köthen, Oct. 1988.
1198. J. Wehlmann, “Alternative zur Membrantechnik,” *Chemische Industrie* 1991, no. 12, 20 – 23.
1199. Bunawerke Hüls GmbH, EP-A 462 317, priority 13. 06. 1990 DE (W. Breuker, H. Wagner, E. Möller, B. Schleimer).
1200. List AG, EP-A 267 531, priority 12. 11. 1986 CH (J. List).
1201. Bayer, EP-A 451 602, priority 07. 04. 1990 DE (C. Casper, J. Weinschenk).
1202. Exxon Research and Engineering Co., US 4 104 218, priority 30. 12. 1975 GB (P. Hous).

1203. Toyo Engineering Corp., EP-A 374 879, priority 23. 12. 1988 JP (H. Miyakawa, T. Shinohara, T. Maeda).
1204. Sumitomo Chemical Co. Ltd., EP 141 622, priority 28. 10. 1983 JP (K. Matsuyama, T. Suzuki, T. Chigusa).
1205. Du Pont, US 3 264 275, priority 02. 08. 1966 (D. B. Pattison, H. E. Schroeder).
1206. R. E. Uschold, J. B. Finlay, *J. Appl. Polym. Sci. Appl. Polym. Symp.* **25** (1974) 205.
1207. H. Brackemann, M. Buback, F. Rindfleisch, S. Rohde, *Angew. Chem. Int. Ed. Engl.* **30** (1991) no. 12, 1654 – 1656.
1208. R. Coulter, "Union Carbide Plant Will Test Gas Phase EP," *Rubber Plastic News* (1991) March 4 5.
1209. *Houben-Weyl*, **E 20**, 2330 – 2332.
1210. W. Wenneis, V. Siekermann, "Nitrosamine in der Gummiindustrie," in VDI Gesellschaft Kunststofftechnik (ed.): *Elastomere und Umwelt,* VDI-Verlag, Düsseldorf 1991, pp. 109 – 126.
1211. S. Lüpfert: "Umweltschutz und Rezeptur," *GAK Kautsch. Gummi Kunstst.* **44** (1991) no. 8, 771 – 779.
1212. Deutsche Forschungsgemeinschaft, *Mitteilung XXII der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe vom 10. Juli 1986*, VCH Verlagsgesellschaft, Weinheim, Germany 1986.
1213. Bundesarbeitsblatt 10 (1985); 1 (1987) Kohlhammer Verlag, Köln.
1214. Verordnung über gefährliche Stoffe (Gefahrstoffverordnung – Gef. Stoff. V) Aug. 26, 1986, Bundesgesetzblatt no. 47, 1986, p. 1470.
1215. Gesetz zum Schutz vor gefährlichen Stoffen (Chemikaliengesetz – Chem. G.) Sept. 16, 1980, Bundesgesetzblatt no. 58 (1980) 1718.
1216. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft – TA Luft) Feb. 27, 1986; Gemeinsames Ministerialblatt no. 7, 1986, p. 95.
1217. K. Berendes: *Das neue Abwasserabgabengesetz*, 2nd ed., Beck Verlag, München 1989.
1218. R. Franck, H. Mühlschlegel (eds.): *Kunststoffe im Lebensmittelverkehr, Empfehlungen des Bundesgesundheitsamtes*, Carl Heymanns Verlag, Köln 1985.
1219. R. Henselder-Ludwig: Abfallrecht, Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit, Bundesanzeiger Verlagsgesellschaft mbH, Köln 1991.
1220. H.-J. Schwenke: "Recyclingstrategien für die Reifenindustrie aus ganzheitlicher Sicht," *GAK Kautsch. Gummi Kunstst.* **45** (1992) no. 1, 60 – 67.
1221. H. Puchan: "Das zweite schwarze Leben," *Frankfurter Allgemeine Zeitung*, Mai 05, 1990.
1222. G. von Randow: "Falsche Scheu vor Runderneuerten," *Die Zeit*, Sept. 11, 1990, p. 98.
1223. W. O. Murtland: "Reviews of Recent Rubber Recycling Literature," *Elastomerics* (1983) May 13 –36.
1224. J. Paul: "Rubber Reclaiming," *Encycl. Polym. Sci. Eng.* **14** (1988) 787 – 804.
1225. P. T. Williams, S. Besler, D. T. Taylor, *Fuel* **69** (1990) no. 12, 1474 – 1482. Links
1226. K. Vohwinkel: "Regenerat – eine Alternative zum Thema Recycling und Umweltschutz," *Kautsch. Gummi Kunstst.* **44** (1991) no. 8, 747 – 750.
1227. F. Sommer: "Stoffliches Recycling von Altgummiabfällen," *Plastverarbeiter* **42** (1991) no. 1, 98 – 103.
1228. H. J. Büschleb: "Herstellung von Gummimehl aus Altprotektor, Replastifizierung und Einsatz in Protektormischungen," *Plaste Kautsch.* **37** (1990) no. 6, 205 – 207.
1229. "Altreifen für den Straßenbau?" *Kautsch. Gummi Kunstst.* **38** (1985) no. 11, 986.
1230. P. S. Gentschewa: "Einfluß von vermahlenem Gummi auf die Eigenschaften von Laufflächenmischungen für Runderneuerungen," *Kautsch. Gummi Kunstst.* **39** (1986) no. 4, 351.
1231. W. Kaminski, H. Rößler, H. Sinn: "Pyrolyse von Elastomeren und Gummi zur Werkstoffrückgewinnung," *Kautsch. Gummi Kunstst.* **44** (1991) no. 9, 846.
1232. *Houben-Weyl*, **E 20** (1987) 2332 – 2343.
1233. H. Timmann, A. Dreyer: "Das D.R.P.-Verfahren," *Kautsch. Gummi Kunstst.* **39** (1986) 637 – 640.
1234. I. Mladenow, J. Patov: "Verwertung von Altreifen durch thermische Destruktion," *Plaste Kautschuk* **31** (1984) 149 – 151.
1235. "Die Redaktion sprach am Jahresbeginn 1992 mit Dipl.-Ing. Wilhelm Lenz, dem Geschäftsführer der GAVS, Frankfurt (M) 90", *Kautsch. Gummi Kunstst.* **45** (1992) no. 1, 68 – 69.
1236. J. Bergh, R. O. Simpson: "The Environmental Tyre," *Tyretech '91 Conference*, Berlin, Oct. 24 – 25, 1991.
1237. *Eur Rubber J* **179** (1997) April.
1238. International Rubber Study Group, Heron House, 109/115 Wembley HA9 8DA, UK http://www.rubberstudy.com.
1239. International Institute of Rubber Producers, 2077 South Gessner Road, Suite 133, Houston, Texas 77063, USA http://www.iisrp.com.
1240. *IRSG, Rubber Statistical Bulletin* **57** (2003) no. 6, March, inside front cover.
1241. IISRP, Worldwide Rubber Statistics 1991 and 1992.
1242. IISRP, Worldwide Rubber Statistics 2002.
1243. *IRSG, Rubber Industry Report* **2** (2003) no. 8, Feb., p. 34.
1244. *Chemical Week* 1994, 19 March, 23.
1245. *Kautsch. Gummi Kunstst.* **55** (2002) no. 7-8, 347.
1246. *Chemical Week* 1999, 21 July, 20.
1247. *Financial Times* 2002, Oct. 30, 23.
1248. *European Chemical News* 2002, 29 April-5 May, 28.
1249. *European Chemical News* 2001, 29 Oct-4 Nov, 28.
1250. *Plastic & Kautschuk Zeitung* 1996, Nr. 8, 18. April p. 28.
1251. *Eur. Rubber J.* **182** (2000) no. 10, 3.
1252. E. Bakuniec, ACS Meeting, Dallas, April 2000.
1253. D. Beach, J. Schroeder, *Rubber World* (2000) Sept, 44.
1254. R. Engehausen, G. Marwede, *ITEC 98*, Akron, 1998.
1255. S. Nishina, H. Kobayashi, H. Koga, *IRC 2001*, Birmingham, England, June 2001.
1256. A. W. Niziolek, *ACS Meeting*, Anaheim, 1997.
1257. A. W. Niziolek, J. G. Neilsen, R. H. Jones, *Kautsch. Gummi Kunstst.* **53** (2000) 358.
1258. *Tire Business*, 29.1.2001, p. 10.
1259. *Tire Technology International*, June 2001.
1260. M. Bortolotti, S. Busetti, F. Mistrali, L. Vitalini, *Kautsch. Gummi Kunstst.* **51** (1998) 331.
1261. H. Colvin, S. Henning, Tire Technology Expo 2001, Cannes, 2001.
1262. J. A. Ayala, S. Moctezuma, J. Lopez, *Proc. ACS Div. Polymer Mat. Sci. Eng.* **73**, (1995) 535.
1263. A. F. Halasa, W. L. Hsu, *Polym. Prepr*. **37** (1996) 676.
1264. A. F. Halasa, *Rubber Chem. Tech.* **70** (1998) 295.
1265. M. L. Kerns, Z. G. Xu, S. Christian, G. Thomas, ACS Meeting, Cincinnati, 2000.
1266. H. Akema, IISRP Annual General Meeting, Vancouver, 1997.
1267. M. J. Wang, *Rubber Chem. Tech.* **71** (1998) 520.

1268. R. Engehausen, A. Rawlinson, J. Trimbach, DKT 2000, Nürnberg, 2000.
1269. F. Tsutsumi, M. Sakakibara, N. Oshima, *Rubber Chem. Tech.* **63** (1990) 8.
1270. P. Cochet, ITEC Asia, Korea, 2001.
1271. A. Hunsche, U. Görl, H. G. Koban, T. Lehmann, *Kautsch. Gummi Kunstst.* **51** (1998) 525.
1272. C. Lin, W. L. Hergenrother, E. Alexanian, G. G. A. Böhm, ACS Meeting, Cleveland, 2001.
1273. T. Fujimaki, K. Morita, *Int. Polym. Sci. Tech.* **25** (1998) 26.
1274. A. Saito, H. Yamada, T. Matsuda, N. Kubo, N. Ishimura, ACS Meeting, Savannah, 2002.
1275. H.-D. Luginsland, A. Hasse, ACS Meeting, Dallas, 2000.
1276. B. Schwaiger, IISRP Annual General Meeting, Heidelberg, 2000.
1277. H. Bender, M. Grün, S. Mücke, DKG Fachtagung, Fulda, 2002.
1278. R. H. Schuster, *Rubber World* (2001) Sept., 24.
1279. T. Karato, M. Nakamura, Y. Takagishi, Tire Technology Expo, Cannes, 2001.
1280. V. M. Monroy, S. D. Hofmann, R. Tietz, ACS Meeting, Cleveland, 2001.
1281. R. Engehausen, A. Rawlinson, J. Trimbach, Kautsch. Gummi Kunstst. 54 (2001) 528.
1282. R. Engehausen, ACS Meeting, Savannah, 2002.
1283. C. Arcoleo, I. Bowman, L. Soddu, P. Tortoreto, *Indian Rubber J.* (2000) Sept., 94.
1284. V. Null, *Kautsch. Gummi Kunstst.* **52** (1999).
1285. S. Nielsen, *Tire Tech. International* (2002) 126.
1286. R. Engehausen, A. J. M. Sumner, ACS Meeting, Orlando, 1999.
1287. G. E. Jones, ACS Meeting, Dallas, 2000.
1288. W. H. Waddell, *Rubber Chem. Tech.* **71** (1998) 590.
1289. J. Palombo, ACS Meeting, Savannah, 2002.
1290. Th. Früh, A. Schuch: “Einfluss gebundener Chemikalien auf Arbeitshygiene, Mischungs- und Vulkanisateigenschaften,” *Kautsch. Gummi Kunstst.* **52** (1999) 242.
1291. TRGS 552, BArbBl. 3/1996.
1292. EU Directive 93/11/EEC.
1293. B. G. Willoughby, K. W. Scott, “*Nitrosamines in Rubber*” , RAPRA Technology Ltd. 1997, and literature cited therein.
1294. S. K. Weiland *et al.*: “Workplace risk factors for cancer in the German rubber industry,” *Occup. Environ. Med.* **55** (1998) 317 – 325. Links
1295. J. Asplund: “Curing Fumes - A Large-Scale Study,” *Kautsch. Gummi Kunstst.* **48** (1995) 276.
1296. G. Knobloch: “Ein neuer Weg zu polymergebundenen Alterungsschutzmitteln, *Kautsch. Gummi Kunstst.* **52** (1999) 10.
1297. A. B. Czuppon *et al.*: “The rubber elongation factor of rubber trees (Hevea brasiliensis) is the major allergen in latex,” *J. Allerg. Clin. Immunol.* **92** (1993) 690.
1298. W. Baumann, M. Ismeier: “Emissionen beim bestimmungsgemässen Gebrauch von Reifen”, *Kautsch. Gummi Kunstst.* **51** (1998) 182; W. Baumann, M. Ismeier: “Kautschuk und Gummi,” Springer-Verlag, Berlin-Heidelberg, 1998.
1299. PRé consultants B. V.: “Life Cycle assessment of an average European car tire,” *Third party report*, Amersfoort, The Netherlands 2001.
1300. C. Former, E. Osen: “Stand und Perspektiven des Gummirecyclings,” *Kautsch. Gummi Kunstst.* **56** (2003) 81.
1301. H-J. Masnuel, W. Dierkes, A. Hendriks: “Butyl Reclaim in Innerliner Applications,” *Kautsch. Gummi Kunstst.* **53** (2000) 730.
1302. D. De, S. Maiti, B. Adhikari, *Kautsch. Gummi Kunstst.* **53** (2000) 346.
1303. J. K. Kim, R. P. Burford: “Study on Powder Utilization of Waste Tires as a Filler in Rubber Compounding,” *Rubber Chem. Technol.* **71** (1998) 1028. Links
1304. U. Baumann *et al.*: “Umweltrelevante Untersuchungen an einigen Gummiartikeln,” *Kautsch. Gummi Kunstst.* **51** (1998) 28.
1305. M. Myhre, D. A. MacKillop: “Rubber Recycling,” *Rubber Chem. Technol.* **75** (2002) 429. Links
1306. KOM (2001) 88 (27.02.2001).
1307. V. Null: “Safe Process Oils for Tires with Low Environmental Impact,” *Kautsch. Gummi Kunstst.* **52** (1999) 799.
1308. R. Engehausen, A. Rawlinson, J. Trimbach: “Reifenkautschuke - interessante Neuentwicklungen,” *Kautsch. Gummi Kunstst.* **54** (2001) 528.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Rubber, 4. Chemicals and Additives

HANS-WILHELM ENGELS, Bayer AG, Dormagen, Germany (Chap. 1 and 2)

HERRMANN-JOSEF WEIDENHAUPT, Bayer AG, Dormagen, Germany (Chap. 3 and 4)

MANFRED PIEROTH, Bayer AG, Dormagen, Germany (Chap. 7 and 10)

WERNER HOFMANN, RWTH, Aachen, Germany (Chap. 9)

KARL-HANS MENTING, Schill + Seilacher "Struktol" Aktiengesellschaft, Hamburg, Germany (Chap. 6.1, 6.2, 6.3, 6.4, 6.5 and 6.7)

THOMAS MERGENHAGEN, Schill + Seilacher "Struktol" Aktiengesellschaft, Hamburg, Germany (Chap. 6.6)

RALF SCHMOLL, Degussa AG, Wesseling, Germany (Chap. 4.3 and 4.3.1)

STEFAN UHRLANDT, Degussa Corporation, Piscataway, NJ, USA (Chap. 4.3 and 4.3.1)

1. Introduction 2
2. Vulcanization Chemicals 2
2.1. Basics of Vulcanization 2
2.2. Sulfur-Containing Cross-Linking Agents 4
2.3. Vulcanization Accelerators 5
2.3.1. Classes of Accelerators 8
2.3.2. Toxicology and Ecology of Organic Vulcanization Accelerators 12
2.3.3. Delivery Forms of Vulcanization Accelerators 12
2.4. Accelerator Activators 13
2.5. Sulfur-Free Cross-Linking Agents 13
2.5.1. Peroxides (Organic Peroxides and Peroxycarboxylic Acids) 13
2.5.2. Quinone Dioxime 15
2.5.3. Polymethylolphenol Resins 15
2.5.4. Other Cross-Linking Agents 16
2.5.5. Radiation Cross-Linking 16
2.6. Vulcanization Retarders 16
3. Antidegradants 17
3.1. Appearance of Aging and Fatigue 17
3.2. Classification and Types of Antidegradants 21
3.2.1. Staining Antidegradants that Act as Anti-Flex-Cracking Agents and Antiozonants 22
3.2.2. Staining Antidegradants with Protection against Fatigue, but Not against Ozone 23
3.2.3. Staining Antidegradants with Little or No Protection against Fatigue and None against Ozone . . 24
3.2.4. Nonstaining Antidegradants with Protection against Fatigue or Ozone 24
3.2.5. Nonstaining Antidegradants without Protection against Fatigue or Ozone 24
3.2.6. Nonstaining Antiozonants without Protection against Aging 26
3.2.7. Other Antidegradants 27
3.3. Choice of Antidegradants 27
4. Fillers and Pigments 28
4.1. Typical Properties and Types of Fillers 28
4.1.1. Action of Fillers 28
4.1.2. Properties of Fillers and Their Significance 30
4.2. Carbon Blacks 31
4.3. White Fillers 33
4.3.1. Precipitated Silicas 33
4.3.1.1. Production 34
4.3.1.2. Physicochemical Properties 36
4.3.1.3. Commercial Products 39
4.3.1.4. Essentials of Rubber Silicas 40
4.3.2. Inactive White Fillers 40
4.3.3. Coating of Fillers 41
4.4. Pigments 41
4.4.1. White Pigments 41
4.4.2. Inorganic Colored Pigments 42
4.4.3. Organic Colored Pigments 42
4.5. Organic Fillers 42
5. Plasticizers 42
5.1. General 42
5.2. Mineral Oils 43
5.3. Synthetic Plasticizers 44
5.4. Natural Products 46
6. Processing Additives 46
6.1. Behavior and Classification of Processing Additives 46
6.2. Peptizers 46

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_365.pub2

6.3. Dispersing Agents and Lubricants 48
6.4. Homogenizers 49
6.5. Tackifiers 49
6.6. Release Agents 49
6.6.1. Batch-Off Release Agents 50
6.6.2. Mold Release Agents 50
6.6.3. Release Agents for Special Applications 51
6.6.3.1. Mandrel Release Agents 51
6.6.3.2. Bladder Release Agents 51
6.6.3.3. Inside Tire Paint 51
6.6.3.4. Outside Tire Paints 51
6.7. Other Processing Additives 52
7. Rubber Adhesion 52
7.1. Adhesion of Rubber to Reinforcing Materials 52
7.2. Rubber – Metal Bond 54
8. Chemical Blowing Agents 55
8.1. Cellular Rubber Articles 55
8.2. Product Types and Properties . . 55
9. Latex Chemicals 58
10. Flameresistant Rubber Articles . 60
10.1. Combustion Process in Polymeric Materials 60
10.2. Testing of Flammability 61
10.3. Inhibition of Combustion 61
11. References 61

1. Introduction

Rubber chemicals and additives are essential for the production of rubber-elastic materials. They basically fulfill three functions:

1) They make compounding and processing easier.
2) They determine elastomer properties.
3) They protect the finished product against deterioration in properties under the conditions of use. This broad range of functions explains the large number and variety of rubber chemicals and additives [1 – 4].

2. Vulcanization Chemicals [5 – 9]

2.1. Basics of Vulcanization

Rubber products are characterized by high resilience following mechanical deformation: the original dimension is restored almost immediately after the load has been removed. This rubber elasticity results from a three-dimensional network structure of long-chain rubber molecules that are bonded to one another at cross-linking points (cross-links). Vulcanization is the chemical process of forming these cross-links between polymer chains. The viscous raw rubber, characterized by irreversible Newtonian flow, is converted into the elastomer.

Sulfur and peroxides are used mainly as curing agents; to a small extent, special cross-linking agents may be employed.

The restoring force after deformation is a function of the number of network-forming polymer chains per unit volume. Increasing the number of cross-links distributes tension over more polymer chains and thus increases the restoring force [10].

The vulcanization process can be monitored continuously by using *vulcameters.* The cross-linkable rubber compound is subjected to periodic deformation in a reaction chamber at a given temperature, and at the same time the force or torque necessary to produce deformation is monitored. Basically two types of vulcameters exist that measure linear and torsional shear, depending on the sample deformation [11]. Important industrial parameters can be determined from a plot of the cross-linking or curing isotherm as a function of reaction time (Fig. 1). The *scorch time* t_s (time taken to reach a given increase ΔF above F_{min}) defines the beginning of the cross-linking reaction (onset of cure) and gives information on the flow time and processibility of the compound at a given temperature. The minimum force F_{min} observed correlates with the compound viscosity. Depending on the vulcanization agents chosen, the start of vulcanization can be adjusted to be rapid or delayed. For curing in molds (e.g., in tire production) a longer flow period is necessary for the rubber compound before vulcanization starts, so that the compound can fill all the cavities in the mold perfectly and trapped air can be released. In other vulcanization processes, as in steam curing or ultrahigh-frequency (UHF) vulcanization, a rapid start is desired, so that the onset of cross-linking prevents deformation of articles through the flow processes. Too rapid an onset of cure

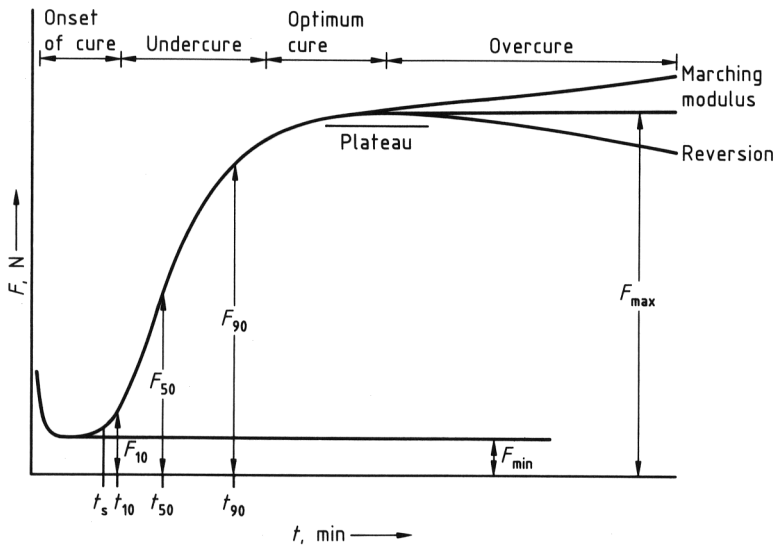

Figure 1. Vulcameter curves (for explanation see text)

(or scorching) is undesirable in most cases because the processing safety in compounding and molding is impaired. Scorched compounds can no longer be perfectly extruded or calendered.

The *steepness of the cross-linking isotherm* is a measure of the rate of the cross-linking reaction. Rate constants can be calculated from the reaction times t_x and the corresponding torque F_t. Generally, a very rapid cure rate is preferred. In the region of optimum cure the maximum cross-link density is achieved. The ideal graph of a cross-linking isotherm is a right-angled step function. The final force achieved $F_{\max}$ is a measure of the cross-link density and often correlates with stress values (force per initial cross section required to produce a specific elongation) obtained from static stress – strain tests [12]. In the overcure region the vulcameter force can be constant, can increase (marching modulus), or for certain rubbers (e.g., natural rubber and polyisoprene) can decrease (reversion).

The course of vulcanization can also be followed discontinuously by plotting typical rubber-elastic properties such as stress, tensile strength, elongation at break, elasticity, or tear resistance, against vulcanization time. Individual physical properties reach their optimum at different vulcanization times. For example, elasticity and compression set values are optimal in the region of slight overcure, whereas tear resistance shows maximum values at slight undercure.

Variation of Vulcanizate Properties with Cross-Link Density. Figure 2 shows schematically the dependence of some vulcanizate properties on cross-link density. The static modulus (curve c) increases with vulcanization more steeply than the dynamic modulus (curve b), which is composed of elastic and viscous parts. Vulcanization leads to a shift in the viscous properties in favor of elastic behavior. The hysteresis losses (i.e., the energy converted irreversibly to heat) decrease (curve f). The tensile strength, tear resistance, fatigue life, and toughness are all a function of energy at break (curve a). These properties increase initially with the buildup of cross-links and then decrease rapidly with increasing cross-link density [13].

Variation of Vulcanizate Properties with Structure of Cross-Links. The mechanical properties of natural rubber (NR) vulcanizates are more strongly affected by the structure of the cross-links than those of synthetic rubber vulcanizates. For example, polysulfidic cross-links increase tear and fatigue resistance, while thermally stable carbon – carbon cross-links or monosulfidic ones lead to aging-resistant vulcanizates with low compression set and higher reversion resistance.

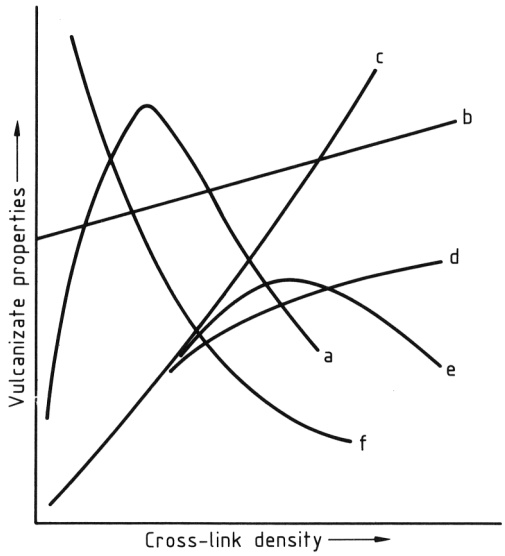

Figure 2. Variation of vulcanizate properties with cross-link density
a) Tear strength, fatigue life, toughness; b) High-speed dynamic modulus; c) Static modulus; d) Hardness; e) Tensile strength; f) Hysteresis, permanent set, friction coefficient

2.2. Sulfur-Containing Cross-Linking Agents

Rubber containing carbon-carbon double bonds, such as NR, styrene – butadiene rubber (SBR), butadiene rubber (BR), and nitrile rubber (NBR) is basically still being cross-linked, as in the past 80 years, by using a combination of sulfur, organic accelerators, and activators (e.g., zinc oxide). The annual worldwide consumption of sulfur for the production of elastomeric materials was estimated to be 225 000 t in 1990, and that of organic vulcanization accelerators and sulfur donors to be more than 200 000 t. Sulfur cross-linking accounted for more than 80 % of all rubber cross-linking reactions [14].

Vulcanization with sulfur and accelerators is preferred for several reasons: (1) economics, (2) excellent controllability of the vulcanization rate with accelerators and retarders, and (3) the possibility of affecting the nature of the cross-links by the type and quantity of accelerators and by the amount of sulfur. Polysulfidic cross-links give good elastic properties and high tensile strength to vulcanizates; mono- or disulfidic cross-links yield vulcanizates with good heat resistance.

Completely saturated rubbers, such as ethylene – propylene elastomers (EPM), ethylene – vinylacetate copolymers (EVM), chloropolyethylene (CM), or acrylate rubber (ACM) cannot be vulcanized with sulfur. Other cross-linking agents must be used.

Elemental Sulfur. Sulfur has been known as a cross-linking agent for more than 150 years. C. Goodyear first succeeded in cross-linking natural rubber in the United States in 1839. The patent claims the production of rubberized cotton tissue at 132 °C using 100 parts NR, 20 parts sulfur, and 28 parts $PbO \cdot PbCO_3$ [15]. Although they were part of the formulation, Goodyear did not recognize the cure-activating effect of metal oxides. At about the same time, T. Hancock in the United Kingdom discovered that NR can be vulcanized by dipping it in molten sulfur [16].

Ca. 0.2 – 5 % sulfur (based on the rubber) is added for the production of *rubbery articles*. For *vulcanite*, the proportion of sulfur is increased to 25 – 50 %. The dosage range of ca. 5 – 25 % sulfur is not used for most applications (apart from floor or certain roller coverings), because the vulcanizates produced lie in the leather-hard region and have unfavorable toughness and elastic properties. To be suitable for vulcanization, sulfur should be 99 % pure (ash content <0.2 %). It must be acid free (acid content, max. 0.2 %), because acids retard vulcanization. Flowers of sulfur, which generally contain a significant quantity of acid, are therefore not used often. Sulfur powder, produced by the Claus process (→ Sulfur) with subsequent grinding, is preferred. This sulfur has particle sizes up to 150 µm, classified according to Chancel grade (old method to determine the particle size of sulfur suspended in ether) [17] and can be distributed readily in rubber compounds. Uniform distribution of sulfur is a prerequisite for a good property profile of rubber. At low sulfur dosages or in rubbers in which sulfur cannot readily be distributed homogeneously (e.g., NBR), sulfur treated with dispersion agents or sulfur pastes are frequently applied.

If so much sulfur is added to the compound that its solubility limit is exceeded, common sulfur tends to bloom on the surface, which makes adhesion or welding during assembling more difficult. Blooming can be avoided by using *insoluble sulfur* (sulfur containing 95 – 97 %

highly polymerized molecules that are insoluble in carbon disulfide). The insoluble sulfur (particle diameter 2.5 – 3.5 µm) is much finer than sulfur powder and is mixed with mineral oil plasticizers, dispersion agents, and fillers to achieve good distribution, or is used as a masterbatch. Its primary uses are for tire compounds and, in particular, for adhesive compounds. The insoluble sulfur form is unstable. It is converted slowly at room temperature and rapidly at elevated temperature into the common soluble form. Insoluble sulfur must therefore be stored in a cool place and kept away from alkaline material. On incorporation, the temperature must not rise too far (maximum 120 °C). The use of insoluble sulfur instead of normal vulcanization sulfur does not influence vulcanization kinetics and the properties of vulcanizates [18].

To achieve uniform distribution in latex compounds and avoid sedimentation, the use of *colloidal sulfur* is necessary. Local overcure, which occurs readily in the environment of coarser sulfur particles and impairs the aging properties of vulcanizates, is avoided as a result of the fineness of the colloidal sulfur.

Rubber can be cross-linked with *selenium* and *tellurium*, but these elements are less active. They have no practical importance for reasons of both economics and, particularly, toxicity.

Sulfur Donors. Apart from elemental sulfur, organic sulfur compounds, which liberate sulfur only at elevated temperature, can be used for vulcanization. Sulfur donors are used mainly in combination with accelerators for the production of heat-resistant articles. Products used in practice are dimorpholyl disulfide (dithioldimorpholine DTDM, **1**), 2-morpholinodithiobenzothiazole (MBSS, **2**), caprolactam disulfide (DTDC, **3**), dipentamethylenethiuram tetrasulfide (DPTT, **4**), and isopropylxanthic tetrasulfide (**5**) [19, 20].

Thioplastics (i.e., thiokol rubbers) can also be considered sulfur donors because they can be used instead of sulfur for the vulcanization of rubber. Vulcanization with thiurams, which achieve cross-linking without actually liberating sulfur, is described in Section 2.5. The combination of sulfur donors with thiurams gives particularly aging-resistant vulcanizates with low compression set.

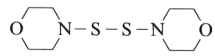

1

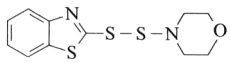

2

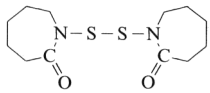

3

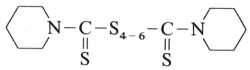

4

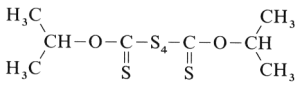

5

2.3. Vulcanization Accelerators

Sulfur is a sluggish cross-linking agent. Without the addition of accelerators and activators a reaction time of 5 h at 140 °C is required for cross-linking NR. In addition, although a sulfur dosage of 8 % is used, unsatisfactory cross-link densities, toughness, and aging properties are achieved. The vulcanization-accelerating effects of litharge, magnesium oxide, calcium hydroxide, and antimony sulfide were recognized soon after the discovery of sulfur vulcanization [21]. Since the introduction of organic accelerators, these inorganic ones have almost completely lost their importance. The history of accelerator development is very instructive for understanding accelerated sulfur cross-linking. In 1906, Oenslager in the United States recognized the activating effect of organic bases such as aniline on sulfur cross-linking [22]. For toxicological reasons, aniline was later replaced by thiocarbanilide (diphenylthiourea), the reaction product of aniline with carbon disulfide.

Since synthetic rubber reacts even more sluggishly than natural rubber, it cannot be vulcanized by using sulfur alone. Within the framework of synthetic rubber development, more active cross-linking systems were necessary. In 1910, Hoffmann and Gottlob discovered that ammonium dithiocarbamates – formed from aliphatic

and aromatic amines with carbon disulfide – aldehyde – amine condensation products, and guanidines were efficient vulcanization accelerators [23]. In experiments on the control of vulcanization kinetics, particularly the onset of reaction, dithiocarbamate derivatives were investigated, and zinc dithiocarbamate [24] and thiurams [25] were discovered (1919 – 1920). At about the same time, zinc oxide and stearic acid were found to be effective activators of accelerated sulfur cross-linking [26, 27]. Both compounds are still included in every sulfur cross-linking system. With organic accelerators and activators a significantly lower sulfur level could be used, which improved the aging properties of the vulcanizate. In 1921, mercaptobenzothiazole was discovered independently by Bedford and Sebrell at Goodyear Tire & Rubber [28] and by Bruni and Romani at Pirelli [29]. The onset of cure could be delayed significantly by derivatives of mercaptobenzothiazole. This finally led to the discovery of sulfenamides based on mercaptobenzothiazole at Bayer in 1931. These sulfenamides are the largest class of accelerators in terms of quantity and value [30].

The importance of organic vulcanization accelerators is based on the following factors: They greatly increase the rate of the cross-linking reaction with sulfur, which allows vulcanization time to be shortened to an economically justifiable level. Use of vulcanization accelerators also permits vulcanization at relatively low temperature and thus avoids heat damage to the rubber and improves aging resistance. Stepwise dosage and a combination of different accelerators enable the scorch and complete cure to be regulated, which are extremely important practical goals. Furthermore, the quantity of sulfur necessary for achieving optimal vulcanization properties is decreased, since the sulfur is used more efficiently for the formation of cross-links, and less is lost in undesirable side reactions. This leads to a considerable improvement in the aging resistance of rubber articles. Another important result of lower sulfur dosage is flattening of the vulcanization isotherm (plateau effect), which reduces the danger of overcure and impairment of vulcanizate properties.

Classification of Organic Accelerators. The list of commercially available accelerators for sulfur cross-linking systems currently consists of about 80 different substances, and the number of compounds that are capable of acceleration is even larger. Table 1 gives a summary of the most important accelerators grouped according to their chemical constitution, together with their international abbreviations.

Important accelerator producers and the trade names of their products are given below:

Akzo	Perkacit
Atochem	Ekaland
Bayer	Vulkacit
Chemetall	Naftocit
Goodrich	Good Rite
Kawaguchi	Accel
Kumho	Kumac
Monsanto	Santocure
Ouchi	Noceler
Robinson Brothers	Robac
Sanshin	Sanceller
Sumitomo Chemicals	Soxinol
Uniroyal	Delac

Figure 3 shows a comparison of the most important accelerators in which the normalized cure time achievable with them is plotted against the normalized scorch time (a measure of the resistance to cure onset) for natural rubber tread compounds. Individual members of each accelerator group naturally differ in their vulcanization behavior and can be classified further within ranges of cure times achievable.

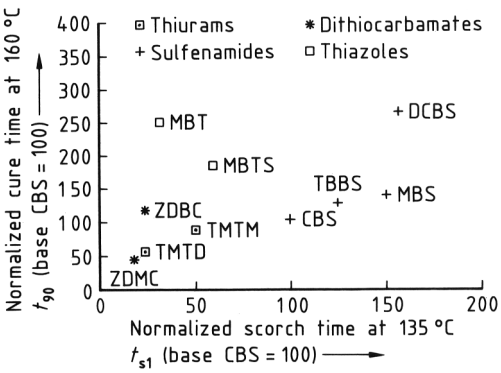

Figure 3. Comparison of accelerators (equimolar amounts) in tread compounds

With skillful choice of an accelerator or a combination of accelerators, almost all desired processing and cross-linking characteristics, as well as vulcanizate properties, can be obtained. Typical sulfur cross-linking systems for carbon-

Table 1. Classification of organic vulcanization accelerators

Group	International abbreviation	Chemical name	CAS registry no.
Sulfenamides	CBS	*N*-cyclohexyl-2-benzothiazylsulfenamide	[*95-33-0*]
	TBBS	*N-tert*-butyl-2-benzothiazylsulfenamide	[*95-31-8*]
	MBS	2-morpholinothiobenzothiazole	[*102-77-2*]
	DCBS	*N,N*-dicyclo-2-benzothiazylsulfenamide	[*4979-32-2*]
	OTOS	*N*-oxydiethylenethiocarbamyl-*N′*-oxydiethylenesulfenamide	[*13752-51-7*]
Thiazoles	MBT	2-mercaptobenzothiazole	[*149-30-4*]
	MBTS	dibenzothiazyl disulfide	[*120-78-5*]
	ZMBT	zinc 2-mercaptobenzothiazole	[*155-04-4*]
Thiurams	TMTD	tetramethylthiuram disulfide	[*137-26-8*]
	TMTM	tetramethylthiuram monosulfide	[*97-74-5*]
	TETD	tetraethylthiuram disulfide	[*97-77-8*]
	DPTT	dipentamethylene thiuram tetrasulfide	[*120-54-7*]
	TBzTD	tetrabenzylthiuram disulfide	
Dithiocarbamates	ZDMC	zinc dimethyldithiocarbamate	[*137-30-4*]
	ZDEC	zinc diethyldithiocarbamate	[*14324-55-1*]
	ZDBC	zinc dibutyldithiocarbamate	[*136-23-2*]
	ZEPC	zinc ethylphenyldithiocarbamate	[*14634-93-6*]
	Z5MC	zinc pentamethylene dithiocarbamate	[*13878-54-1*]
	ZBEC	zinc dibenzyldithiocarbamate	[*14726-36-4*]
Guanidines	DPG	*N,N′*-diphenylguanidine	[*102-06-7*]
	DOTG	*N,N′*-di-*o*-tolylguanidine	[*97-39-2*]
	OTBG	*o*-tolylbiguanide	
Thioureas	ETU	ethylenthiourea	[*96-45-7*]
	DETU	diethylenethiourea	[*105-55-5*]
	DPTU	*N,N′*-diphenylthiourea	[*102-08-9*]
Other accelerators		aldehyde – amine condensation products	
		amines	
		dithiophosphates	
		xanthates	
		triazines	

black-filled rubbers such as NR, SBR, NBR, butyl rubber (IIR), and ethylene – propene – diene copolymers (EPDM) are shown in Table 2. These conventional vulcanization systems (CV systems) differ in their sulfur content and in the nature and dosage of the accelerator, depending on the substrate. When starting from the basic formulations, special requirements can be fulfilled by variation of the sulfur – accelerator ratio and by combination of different accelerators.

Table 2. Typical vulcanization systems for carbon-black-filled elastomers

Component	NR	SBR	NBR	IIR	EPDM
Sulfur, phr	2.5	2.0	1.5	2.0	1.5
Zinc oxide, phr	5	5	5	3	5
Stearic acid, phr	2	2	1	2	1
CBS, phr	0.6	1.0			
MBTS, phr			1.5	0.5	
MBT, phr					1.5
TMTD, phr				1.0	0.5
ZDBC, phr					1.5

Mechanism of Accelerated Sulfur Cross-Linking. The large number of possible factors affecting sulfur cross-linking can be better understood if the molecular course of cross-link formation is considered in more detail. The mechanism of accelerated sulfur cross-linking is, however, extremely complex and consists of many competing reactions. Most studies of the cross-linking of natural rubber by sulfur have been conducted by Malaysian Rubber Producer Research Association (MRPRA) in London, particularly from 1950 – 1980 [31]. Figure 4 summarizes the mechanism of accelerated sulfur cross-linking. The accelerator molecule (here exemplified by sulfenamide accelerators) reacts with the activators zinc oxide and stearic acid to give an amine-coordinated accelerator – zinc complex (A), which then reacts with sulfur to give a sulfur – accelerator complex (B). The exact composition of this complex is unknown; even whether or not it contains zinc is not known for certain. This sulfur – accelerator complex reacts with the rubber polymer, prefer-

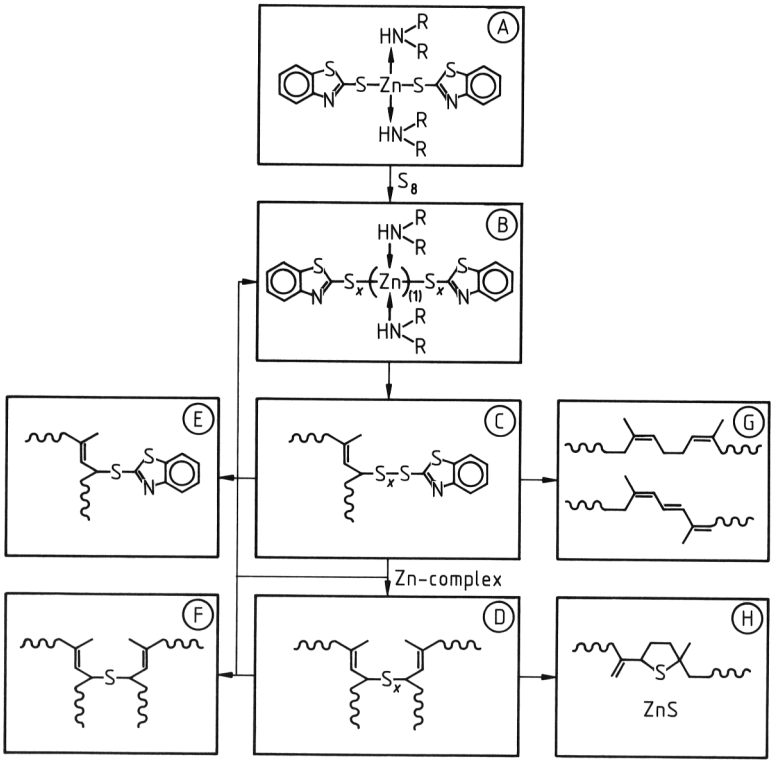

Figure 4. Mechanism of accelerated sulfur cross-linking in natural rubber

ably in the allyl position, and forms a polymer-bound accelerator – polysulfane as the main intermediate (C). The latter reacts with a second rubber chain to form a polysulfidic cross-link only if zinc – accelerator complexes are present (D). However, zinc – accelerator complexes also catalyze the desulfurization of polysulfanes with the reformation of accelerator complexes having high sulfur content, and finally monosulfidic cross-links (F). The sulfur-rich accelerator complexes formed by sulfur extrusion can produce additional cross-links by reaction with the rubber chain. The polymer-bound accelerator – polysulfane can undergo purely thermal elimination reactions as side reactions, which lead to chain modification with the formation of diene and triene structures (G). A second side reaction is the formation of monsulfide bridges (E). The thermal decomposition of polysulfidic cross-links can lead to sulfur ring formation, a process that is associated with a decrease in the cross-link density (modulus reversion, H).

The balance of competing reactions determines the number and structure of cross-links. It depends on the sulfur – accelerator ratio, the structure of the zinc – accelerator complex, the cure temperature, and cure time. Sometimes these sulfur rearrangement reactions continue under service conditions. A high sulfur – accelerator ratio (conventional vulcanization system) favors polysulfidic cross-links, since only a small quantity of the zinc – accelerator complex is available for degradation of the polysulfides. With low sulfur – accelerator ratios (efficient vulcanization, EV system), many monosulfidic cross-links are formed because of a high concentration of the zinc complex.

2.3.1. Classes of Accelerators

Sulfenamides. Sulfenamides are the largest class of organic vulcanization accelerators in terms of quantity and value. They are produced by reaction of 2-mercaptobenzothiazole with

primary and secondary amines under oxidative conditions.

Compared with the accelerators mercaptobenzothiazole (MBT) and dibenzothiazyl disulfide (MBTS, see page 9), the sulfenamides (**6**) result in an extremely delayed onset of cure (longer scorch times) and thus greater processing safety. The thermal stability of sulfenamides varies depending on the nature of the substituents on nitrogen. This, in turn, varies the delay in cure onset. Processing safety increases from *N*-cyclohexyl-2-benzothiazylsulfenamide (CBS) to *N*-*tert*-butyl-2-benzothiazylsulfenamide (TBBS), then 2-morpholinothiobenzothiazole (MBS), and is greatest for *N*,*N*-dicyclo-2-benzothiazylsulfenamide (DCBS).

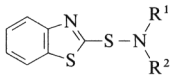

6

CBS: $R^1 = H$, R^2 = Cyclohexyl
TBBS: $R^1 = H$, R^2 = *tert*-Butyl
MBS: $R^1 + R^2 = -CH_2-CH_2-O-CH_2-CH_2-$
DCBS: $R^1 = R^2$ = Cyclohexyl

During vulcanization, sulfenamides are thermally cleaved to give a 2-mercaptobenzothiazyl compound and an amine. The latter acts as a strong secondary accelerator for the former, so that vulcanization proceeds very rapidly after the delayed start. The vulcanization behavior therefore corresponds to that with base-activated thiazole accelerators, except that the sulfenamides have a longer flow period. The vulcanization curves of compounds with sulfenamides thus approach the ideal curves for molded articles (i.e., long flow time and rapid complete vulcanization). The vulcanization rate can be increased by the addition of secondary accelerators, particularly thiurams, and the onset of cure can be delayed by retarders.

Sulfenamides are *used* in the production of tires, conveyor belts, and technical rubber goods that are vulcanized in heated presses. These accelerators are not suitable for hot-air vulcanization. Sulfenamides are decomposed rapidly by hot steam and slowly by moisture; therefore, steam vulcanization starts rapidly with sulfenamides. They are thus also used in combination with secondary accelerators for continuous vulcanization in steam (e.g., cable production).

The thiocarbamylsulfenamides **7** and **8** [32, 33] give vulcanizates with good resistance to reversion.

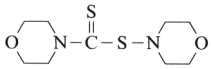

7

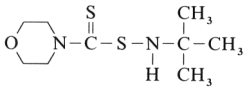

8

The characteristics of vulcanization are broadly similar to those achieved with a combination of MBT-derived sulfenamides and dithiocarbamates. The advantage of thiocarbamylsulfenamides compared to these combinations is the clearly more favorable resistance to cure onset. Thiocarbamylsulfenamides show a strong synergistic effect in combination with other accelerators. Their main uses are in retread compounds for tires and in heat-stable products with low compression set (e.g., engine mountings).

Thiazoles. The thiazoles considered here (**9** – **11**) are derivatives of 2-mercaptobenzothiazole and represent an important class of accelerators.

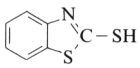

9

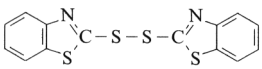

10

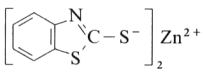

11

2-Mercaptobenzothiazole (MBT),**9** and dibenzothiazyl disulfide (MBTS)**10** are all-purpose accelerators for almost all rubber. With MBT, vulcanization starts rapidly, and with MBTS it is somewhat delayed. The zinc salt of 2-mercaptobenzothiazole (ZMBT)**11** is used mainly in latex compounds. These three "mercapto accelerators" give an average vulcanization rate, a broad plateau, relatively high processing safety, and very good resistance to vulcanizate aging. They are used either alone or, preferably, in combination with other accel-

erators, particularly guanidines, thiurams, and dithiocarbamates. Combinations of mercapto accelerators and basic accelerators, such as *N*,*N*′-diphenylguanidine (DPG), act synergistically. Combinations of, for example, MBTS – DPG or MBT – hexamethylenetetramine are important in industry because of their enhanced vulcanization rate and their high degree of cross-linking. Premixed combinations of this type are available commercially.

2-Mercaptobenzothiazole, dibenzothiazyl disulfide, and their combinations with other accelerators are *used* in the production of technical rubber goods, such as conveyor belts, V-belts, seals, footwear, and rubber threads. The vulcanizates have a bitter taste, so they are not suitable for products that come in contact with foods.

Thiurams. The main thiuram vulcanization accelerators are **12** – **17**.

The most important product in this range is tetramethylthiuram disulfide (TMTD, **12**). Tetrabenzylthiuram disulfide (**17**) [34, 35] an alternative to other thiurams in many applications because the latter can form carcinogenic nitrosamines (see Section 2.3.2) under vulcanization conditions.

The thiuram accelerators produce faster onset of cure and more complete cure than thiazoles. They are therefore employed for more rapid vulcanizations and, in particular, as secondary accelerators for the mercapto and sulfenamide accelerators. Tetramethylthiuram monosulfide (TMTM, **13**) gives a better resistance to cure onset than the disulfide (TMTD).

At higher dosages (2 – 5 %), vulcanization can be carried out with thiuram di- and polysulfides with or without small quantities of sulfur (efficient vulcanization, EV). The vulcanizates formed are characterized by good heat resistance and low compression set, and metals (copper, silver) coming in contact with the vulcanizates are not darkened by sulfide formation. Thiuram vulcanization is used for the production of heat-resistant articles (e.g., seals from NBR). Thiuram accelerators are often used as components of vulcanization systems for ethylene – propylene terpolymers (EPDM rubbers).

Dithiocarbamates. Dithiocarbamates (**18**) have the strongest vulcanization accelerating action (ultraaccelerators).

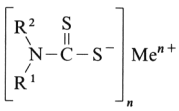

18

The compounds employed most frequently are zinc *N*-diethyldithiocarbamate (ZDEC), zinc *N*-dimethyldithiocarbamate (ZDMC), and zinc *N*-dibutyldithiocarbamate (ZDBC). To a lesser extent, other zinc, selenium, tellurium, lead, copper, or alkaline-earth salts of dithiocarbamic acids are used. Besides these metal salts, dialkyl- and alkylcycloalkylamine dithiocarbamates (e.g., the piperidine salt of pentamethylenedithiocarbamic acid) are used. Zinc dibenzyldithiocarbamate is gaining importance as an alternative to nitrosamine-forming dithiocarbamates (see Section 2.3.2) [36]. Dithiocarb-

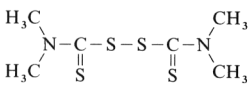

12

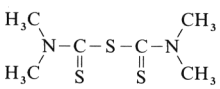

13

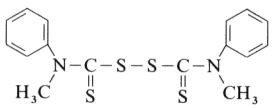

14

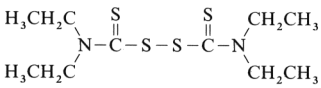

15

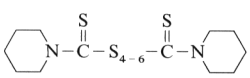

16

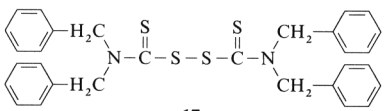

17

amates are added to compounds, that are to be vulcanized at low temperature in the shortest possible time. Examples are dipped articles, rubberized fabrics, self-vulcanizing solutions (sometimes in combination with bases), and latex articles. The rapidly starting vulcanization requires special attention during processing. The vulcanization rate is highest with ammonium salts, while zinc salts exhibit greater processing safety. Besides use in latices, dithiocarbamates are the preferred agents for EPDM cross-linking and are used as activators for thiazole and sulfenamide cross-linking.

Guanidines. Guanidine accelerators, such as *N*,*N*′-diphenylguanidine (DPG), *N*,*N*′-di-*o*-tolylguanidine (DOTG), and *o*-tolylbiguanide (OTBG) are not important as individual accelerators because their accelerating action is low. Without addition of other accelerators, they are used only for the vulcanization of voluminous articles. However, they are important secondary accelerators for thiazoles because such combinations result in faster vulcanization and higher cross-link densities than the individual substances.

Guanidines can be used as activators for sulfenamides to increase reaction rate. Combinations of guanidine and thiazole accelerators are used in compounds with silica fillers, because basic guanidines counteract the delay in cure onset produced by silicic acid. *o*-Tolylbiguanide is an important accelerator for rubber used in contact with foods.

Thioureas. Ethylenethiourea (ETU, 2-mercaptoimidazoline) and diethylenethiourea (DETU) are used mainly as accelerators for the vulcanization of polychloroprene and epichlorohydrin. With zinc oxide they give high cross-link densities. DETU gives a rapid onset of cure; ETU, a somewhat slower one. Thiazoles, such as MBTS, are used as retarders. In EPDM, thioureas are highly activating, particularly in compounds containing calcium oxide. Because of the teratogenic effects of ETU [37], the use of thioureas is decreasing, and they are supplied preferably as polymer-bound granulates,which effectively hinder exposure to and inhalation of thiourea dust.

Other Accelerators for Sulfur Vulcanization.

Aldehyde – Amine Condensation Products. Aldehyde – amine condensation products differ greatly in their accelerating effects. Hexamethylenetetramine (Hexa) is used as an inexpensive, basic secondary accelerator in combination with mercapto accelerators, particularly in light compounds. Its action is similar to, but slower than, that of diphenylguanidine.

Tricrotonylidenetetramine is a special accelerator for the vulcanization of hard rubber compounds, but it vulcanizes relatively slowly. An accelerator that acts more rapidly is the condensation product of butyraldehyde with aniline. This product gives vulcanizates with high elasticity and is used for the production of rubber buffers.

Amines. Amines, such as cyclohexylethylamine, have a strong secondary accelerating effect on dithiocarbamates. For self-vulcanizing adhesive solutions, zinc *N*-ethylphenyldithiocarbamate is used in combination with cyclohexylethylamine. Amines, such as di- or triethanolamines and polyethylene – polyamines, are used as activators for active silica fillers.

Dithiophosphates. The most important dithiophosphate accelerator is zinc dibutyldithiophosphate (**19**).

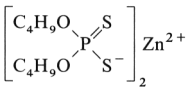

19

Copper diisopropyldithiophosphate and zinc aminedithiophosphates are used to a lesser extent, mainly in cross-linking of EPDM. Here they are used in combination with thiazoles, sulfenamides, and zinc dibenzyldithiocarbamate (ZBEC), which are almost free of nitrosamines (see page 10).

Xanthate Accelerators. Xanthates, such as sodium isopropyl-, zinc isopropyl-, and zinc dibutylxanthate act similar to dithiocarbamates; however, they have not achieved the same importance. They are sometimes used for the production of self-vulcanizing adhesive solutions and with latices for the production of rubber foams.

Triazine Accelerators [38]. Triazine accelerators are similar in structure and activity to thi-

azoles. In addition to its accelerating action the tetrasulfide (**20**) also acts as a sulfur donor.

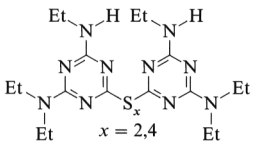

$x = 2,4$

20

Compared with thiazole accelerators, triazines give rise to higher cross-link densities and better stability to reversion of the vulcanizates. Because of their high price and limited availabilty, triazines remain specialized products for use in a few applications.

2.3.2. Toxicology and Ecology of Organic Vulcanization Accelerators

The acute biological (oral and dermal toxicity; irritation of the skin and eyes) and environmentally damaging properties of cross-linking agents have been thoroughly investigated and are well known. They are summarized in the safety data sheets of producers and are shown on packaging labels. Most cross-linking agents are harmless on acute exposure [39]. However, besides the toxicity of the chemicals themselves, the hazards of their decomposition products, which can be present in vapors emitted during vulcanization, must be considered. In the rubber industry, nitrosamines are particularly important. In experiments with animals, most nitrosamines have been shown to be carcinogenic [40]. Stable nitrosamines are formed by nitrosation of secondary amines, which are liberated from accelerators and sulfur donors, and sometimes from short stoppers with the appropriate structure. Fillers (with absorbed NO_x) or production processes (salt baths) can have nitrosating potential. The role of NO_x, ubiquitous in air, is still disputed [41].

In the second amendment of the Chemikaliengesetz (1990), exposure limits with TRK values of 1 and 2.5 µg/m^3 are stipulated for 12 of the nitrosamines frequently occurring in industry. Similar regulations in the EC and the United States are anticipated. Since secondary amines are a necessary prerequisite for the formation of stable nitrosamines, the latter can be prevented by the avoidance of secondary amine sources by product substitution. In the case of sulfenamides this is achieved by the use of compounds based on *primary amines*, such as CBS and TBBS, sometimes in combination with vulcanization retarders. If the use of *secondary amines* is unavoidable, then amines whose nitrosamines have not been found to be carcinogenic in animal experiments (so-called safe amines) should be used. Dicyclohexylamine and dibenzylamine are the most important examples of these. Accelerators based on these amines are the sulfenamide DCBS, the thiuram tetrabenzylthiuram disulfide (TBzTD), and the dithiocarbamate ZBEC. Dicaprolactam disulfide is a sulfur donor that does not form nitrosamines. Since most of the nitrosamines in the rubber matrix are formed during vulcanization, the addition of NO_x scavengers results in a reduction in nitrosamine content. Normal compounding ingredients, such as phenolic antioxidants, amines, and metal oxides, can have scavenging effects. Additives such as α-tocopherol, ascorbic palmitate [42], and urea derivatives [43] are also used as scavengers. Their action depends strongly on the formulation and process used, and requires testing of the targeted nitrosamine reduction in individual cases. Scavengers can contribute to maintaining TRK values and thus represent a short-term solution to nitrosamine minimization. The complete avoidance of nitrosamines is possible only through product substitution.

2.3.3. Delivery Forms of Vulcanization Accelerators [44]

The homogeneous distribution of cross-linking ingredients is a prerequisite for good vulcanizate properties. The level of dispersion depends on factors such as the compound viscosity, shear gradient during compounding, solubility, melting point, and particle size. From the view point of the production engineer, the main requirements of the delivery form are automatic dosability, and suitability for storage in large containers and silos. Industrial hygiene requires that the product be dust free. The wide variety of forms supplied corresponds to this broad catalogue of requirements and includes powders, microgranulates (from spray drying), extrusion granulates with 0 – 5 % oil, pastes, wax granulates (> 5 % oil, wax), and predispersed

polymer-bound granulates. The form supplied must be suited to the requirements in an individual case.

2.4. Accelerator Activators

To develop the full activity of organic vulcanization accelerators, they must always be used in combination with inorganic and organic activators. The addition of *zinc oxide* to rubber compounds containing sulfur and accelerators considerably increases cross-link density. For the general-purpose rubbers NR and SBR, increasing zinc oxide dosage up ca. 5 phr leads to high tensile strength and stress values. Higher dosages have no further effect on these properties.

The effects of *cadmium* and *bismuth oxides* are similar to those of zinc oxide. However, they are not used because of economic and ecological reasons. *Magnesium* and *calcium oxides* are less suitable since they produce much lower cross-link densities. The lead oxides PbO and Pb_3O_4 are used as activators when vulcanizates with particularly low swelling in water are required.

The rubber – sulfur – accelerator – zinc oxide system is further activated by fatty acids. Stearic acid or zinc stearate is generally used. Besides activation, fatty acids and their salts improve processibility and filler distribution, which are important for vulcanizate properties. The use of special zinc salts besides ZnO can lead to an increase in cross-link density and an improvement in reversion behavior [45]. Zinc diamminediisocyanate has been introduced for cross-linking of EPDM [46]. Through activation of cross-linking kinetics and improvement of compression set zinc diamminediisocyanate in combination with MBT, dibutyl dithiophosphate, and ZBEC offers a new EPDM cross-linking system that forms no carcinogenic nitrosamines but nevertheless offers the combination of properties provided by classical accelerator systems.

2.5. Sulfur-Free Cross-Linking Agents

2.5.1. Peroxides (Organic Peroxides and Peroxycarboxylic Acids) [47]

The cross-linking of rubber with peroxides (dibenzoyl peroxide) was first carried out by Ostromislensky in Russia in 1915 [48]. However, industrial interest in this form of cross-linking developed only with the discovery of di-*tert*-butyl peroxide and dicumyl peroxide (in ca. 1950) for the cross-linking of polyethylene, and with the introduction of saturated synthetic rubber grades, such as silicone rubber (Q) and ethylene – propylene – diene copolymers (EPM, EPDM). Peroxides are the most important cross-linking agents for rubber containing few or no double bonds in the chain. Examples are silicone rubber, EPDM, ethylene – vinyl acetate copolymers (EVM), and hydrogenated acrylonitrile – butadiene copolymers (HNBR). Since rubbers containing double bonds, such as NR, SBR, BR, NBR, and chloroprene rubber (CR), generally require the addition of large quantities of antidegradants, cross-linking with peroxide is used only in exceptional cases, for example, to achieve particular heat resistance or low compression set.

For the cross-linking of high molar mass polymers, only peroxides that are sufficiently stable under processing conditions are suitable. The decomposition temperature of peroxides depends on their structure. The half-life of decomposition determines the resistance of cure onset and the cross-linking rate at a given temperature. The addition of activators, such as zinc oxide and stearic acid, is not necessary. The most important cross-linking peroxides are acyl peroxides, perketals, and alkyl or aralkyl peroxides. Acyl peroxides and perketals generally have lower decomposition temperatures than alkyl peroxides.

Acyl peroxide

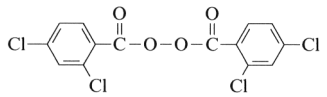

21

Perketal

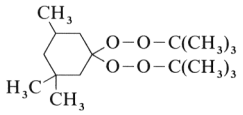

22

Alkyl- and aralkyl peroxides

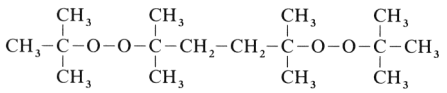

23

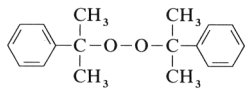

24

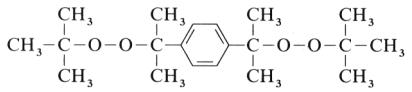

25

Bis(2,4-dichlorobenzoyl) peroxide (**21**) is used mainly for cross-linking of silicone rubber. The perketal *1,1-di(tert-butylperoxy-3,3,5-trimethylcyclohexane* (**22**) results in very rapid cross-linking even at moderate temperature. *2,5-(tert-Butylperoxy)-2,5-dimethylhexane* (**23**) has good processing safety and is used mainly to cross-link EPDM and silicone rubber.

The most important products in terms of quantity are *dicumyl peroxide* (**24**) (for polyethylene cross-linking in cable applications) and *bis(tert-butylperoxyisopropyl)benzene* (**25**) (for cross-linking a wide range of rubber types).

For cross-linking elastomers the preferred peroxides are those that form the following radicals:

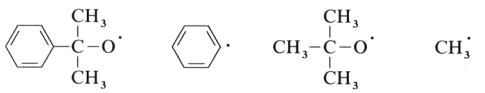

Cumyloxy Phenyl *tert*-Butyloxy Methyl

These radicals are very reactive and can abstract hydrogen atoms from rubber molecules. The polymer radicals thus formed produce a carbon – carbon cross-link upon recombination (Fig. 5).

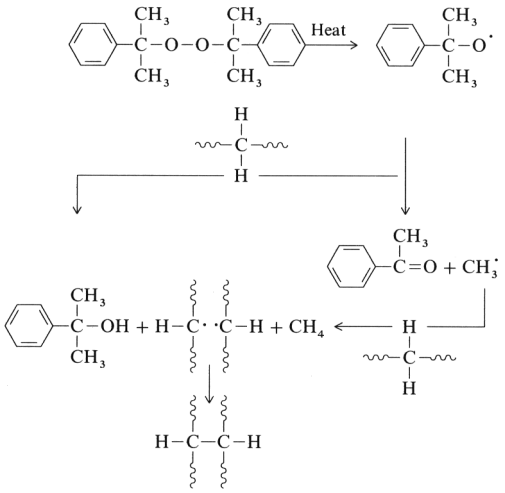

Figure 5. Mechanism of peroxide cross-linking

Besides radical recombination the polymer radical can also be stabilized by chain cleavage. This behavior is especially pronounced with polymers having tertiary carbon atoms, for example, butyl rubber (IIR), epichlorohydrin rubber (ECO), and thermoplastics such as polypropylene. Thus, peroxides are not suitable for cross-linking these polymers.

The kinetics of peroxide cross-linking depend primarily on the structure of the peroxide and the vulcanization temperature, and less on the structure of the polymer. Therefore, blends of polymers that possess widely differing reactivities with regard to sulfur cross-linking can be covulcanized very successfully. Examples are NR – EPDM and NBR – EPDM blends for hosing, NBR – EVM blends for seals with improved light and good oil resistance, and PE – EPDM blends with improved extrusion behavior.

Interactions of peroxide cross-linking agents with various formulation ingredients must be reckoned with. Refined mineral oils (e.g., paraffinic and, under certain conditions, naphthenic oils) are suitable extenders. Highly aromatic oils cannot be used. Synthetic plasticizers generally do not have any negative effects.

Acidic fillers favor the heterolytic decomposition of peroxides, with the formation of ionic

intermediates that do not result in cross-linking. Silica, silicates, kaolin, and talc, which tend to absorb cross-linking agents, should be combined with activators such as diphenylguanidines or polyglycols.

The use of antioxidants is problematic because they are intended to react immediately with free radicals. Small dosages of oligomerized trimethyldihydroquinoline (TMQ), 4,4′-thiobis-6-*tert*-butyl-*m*- cresol, and methylmercaptobenzimidazole (MMBI) allow peroxide vulcanization to proceed without too significant a decrease in cross-link density. Within certain limits the resistance to cure onset [e.g., with butylated hydroxytoluene (BHT)] can be controlled by antioxidants.

Peroxide cross-linking of *polymers that tend to undergo chain cleavage* can often be achieved by using coagents (polyfunctional polymerizable compounds, such as ally and methacrylate compounds) [48]. Coagents significantly increase the cross-link density and improve vulcanizate properties (such as compression set or heat resistance) because they are incorporated into the vulcanizate as cross-links. Common coagents are diallyl terephthalate (DATP), ethylene glycol dimethacrylate (EDMA), trimethylolpropane trimethacrylate (TRIM), triallyl cyanurate (TAC), triallyl isocyanurate (TAIC), *N*,*N*′-*m*-phenylenedimaleinimide, and 1,2- *cis*-polybutadiene.

The choice of the best coagent depends on the compound and the technical requirements. Peroxide vulcanization is carried out mainly in processes such as during mold curing,injection molding, transfer molding, and oxygen-free steam vulcanization, and in salt baths. Hot-air and ultra high-frequency processes are used only in exceptional cases (e.g., silicone cross-linking with 2,4-dichlorodibenzoyl peroxide) since these processes generally produce sticky,adhesive surfaces.

Peroxide- cross-linked vulcanizates generally have lower tear resistance and tensile strength than sulfur- cross-linked elastomers, but excellent heat resistance and low compression set, particularly at high temperature [49].

Forms of Peroxides Supplied. Depending on their thermal stability, peroxides are available commercially in the technically pure form; desensitized on carriers such as chalk, kaolin, or silica; or mixed with polymers such as EPM – EDPM, EVA, or PE (active substance content of the formulation, usually 40 %). The forms of the active substances and mixtures supplied include liquids, pastes, powders, granules, and masterbatch sheets or granules.

Peroxide Producers. Important producers of peroxides are listed below, together with the trade names of the corresponding products:

Akzo	Trigonox, Perkadox, Lucidol, Cadox
Atochem	Luperox, Luperco, Alperox
Herkules	Di-Cup, Vul-Cup
Pennwalt	Luperox, Lupersol
Peroxid Chemie	Interox

Toxicity. Most peroxides used for cross-linking are either slightly toxic or nontoxic. Being oxidizing agents, they cause some skin irritation. Decomposition products of certain peroxides can cause unpleasant odors under vulcanization conditions, for example, the penetrating odor of acetophenone from dicumyl peroxide. Other peroxides are currently available whose decomposition products have considerably less powerful odors. New types of peroxides with groups capable of polymerization permit the incorporation of cleavage products into the polymer (e.g., *tert*-butylperoxy isopropylmethylstyrene initiator D 120) [50].

2.5.2. Quinone Dioxime

p-Benzoquinone dioxime and its dibenzoyl derivative are used mainly for cross-linking butyl rubber in the presence of oxidizing agents, such as lead oxides, because the resulting vulcanizate is especially heat and steam resistant. One area of use is in bladders for tire presses.

2.5.3. Polymethylolphenol Resins

Vulcanization using imcompletely reacted polymethylol resins is also limited mainly to the cross-linking of butyl rubber. In the course of vulcanization the resins are incorporated into the polymer molecules as parts of the cross-links. As in the case of quinone dioxime cross-linking agents, the vulcanizates have good heat

and steam resistance. Compounds such as tin(II) chloride are used to activate the resins. Certain halogenated resins can react without these activators. Since about 1989, resin cross-linking has also been used in the production of thermoplastic polyolefins (TPO) (→ Thermoplastic Elastomers).

2.5.4. Other Cross-Linking Agents

Zinc and magnesium oxides can be used to cross-link rubbers that contain either halogens or carboxyl groups. The mechanisms of reaction of the two groups are quite different.

Rubber Containing Carboxyl Groups. In the case of carboxyl- containing rubbers – of which carboxylated SBR and NBR are economically significant – a pure salt cross-linking occurs. Since salt cross-linking is extremely problematic because of the rapid onset of cure, the processing and cross-linking of carboxyl- containing polymers is generally limited to latex technology. In principle, all di- or polyvalent cations can be used for this type of cross-linking. Zinc oxide is the most commonly used.

Halogen-Containing Rubber. The cross-linking of halogen- containing rubber by metal oxides follows quite a different mechanism. With *unsaturated polymers* such as polychloroprene, and chlorinated or brominated butyl rubber, cross-linking starts with the abstraction of an allylic halogen which is initiated by metal oxides. Both zinc and magnesium oxides can be used for this purpose. With polychloroprene, a combination of both oxides is used to achieve optimum stability on storage and resistance to cure onset. With *sulfur-modified grades*, no accelerators are used, whereas other polychloroprene grades require the addition of organic accelerators for optimal vulcanization. In the cross-linking of halogenated butyl rubber, zinc oxide is used alone or, to achieve particular properties, in combination with accelerators such as MBTS or TMTD. 3-Methylthiazolidine-2-thione is a possible alternative to ETU in polychloroprene cross-linking [51].

If used with *saturated halogen- containing polymers*, zinc oxide catalyzes the formation of hydrochloric acid. For this reason, zinc oxide is not used with chlorinated or chlorosulfonated polyethylene; instead, magnesium oxide is used in combination with special accelerators such as 2,5-dimercaptothiadiazole.

Special rubbers, which require cross-linking agents specifically suited to the type of polymer (e.g., fluoro- or acrylate rubbers), are discussed in → Rubber, 3. Synthetic in the sections dealing with the particular polymers.

2.5.5. Radiation Cross-Linking [52] (see also → Radiation Chemistry, Chap. 1.3.2.2)

Rubber can also be cross-linked without the addition of cross-linking agents, by high-energy radiation such as gamma rays (for example, from cobalt sources) or high-energy electrons (from electron accelerators), with the formation of carbon – carbon bonds. Very compact electron accelerators are now available, which can be incorporated into on-line production processes [53]. An example is the radiation cross-linking of inner liners in tire production, where dimensional stability and building tack are both increased. This allows economy of materials and construction, with a reduced number of building components. The main uses of electron radiation cross-linking are for wire insulation and coatings. A limitation in the use of radiation cross-linking is the relatively short penetration depth. An advantage is the absence of cross-linking agents and their reaction products; therefore, radiation cross-linking is used for the production of disposable medical articles.

2.6. Vulcanization Retarders [54, 55]

The fine adjustment of resistance to cure onset frequently requires the addition of retarders, despite the large number of different accelerators available, particularly when high processing or vulcanization temperatures are used. Acidic substances generally retard the vulcanization of general-purpose diene rubbers. Organic acids and anhydrides, such as benzoic acid, salicylic acid, or phthalic anhydride, are used as vulcanization retarders. Fatty acids and their salts, such as stearic and oleic acids, zinc stearate, and zinc laurate, can also delay the onset of cure. Sulfenic acid derivatives of acidic nitrogen compounds are considerably more effective.

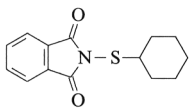

26

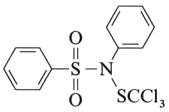

27

The commercially most important and widely used product is cyclohexylthiophthalimide (CTP) (**26**), which exhibits the greatest efficiency with sulfenamide and thiazole accelerators, even at extremely low dosage (0.05 – 0.3 phr). The addition of CTP results in a parallel shift in the vulcanization isotherm along the time axis, while other vulcanization properties are hardly affected. Trichloromethylthiosulfanilide (**27**) is a special retarder from the same class of substances, which is used preferably in combination with thiazole accelerators. Thiurams are barely affected by these types of S – N compounds, but the resistance to onset of cure can be improved by addition of thiazole accelerators and, in particular, sulfenamides.

3. Antidegradants

3.1. Appearance of Aging and Fatigue [4, 56 – 61]

The aging of a material is defined as changes in its properties with time as a result of external influences or use. This does not include changes due to wear and tear (e.g., mechanical damage or abrasion). Aging can manifest itself in vulcanizates in many different ways. The observed changes include:

1) Hardening, embrittlement
2) Softening, becoming sticky
3) Loss of toughness
4) Loss of elastic properties
5) Cracking
6) Changes in the nature of the surface (frosting, crazing, loss of color, gloss, and feel)

In addition, other changes (e.g., in electrical properties, development of odors) can occur.

As seen from this list, the changes can affect either the entire rubber article or only the surface. Aging processes that affect the whole article (bulk processes) are:

1) Aging due to oxygen at low and high temperature (aging in the narrow sense)
2) Aging due to rubber poisons
3) Aging due to heat
4) Aging due to hot water or steam
5) Aging in hot media

Aging processes on the surface include:

1) Dynamic cracking due to oxygen
2) Static and dynamic cracking due to ozone
3) Crazing
4) Frosting

Aging due to Oxygen at Low and High Temperature. The determining factors in aging due to oxygen and heat are the oxygen in the ambient air and the temperature. Even at room temperature, oxygen tends to promote aging processes, which naturally occur more rapidly with increasing temperature. Aging due to oxygen involves continuous absorption of oxygen (initiated by a chain reaction). Atmospheric oxygen is partially bonded to the vulcanizate and partially liberated in the form of carbon monoxide, water, and low molar mass oxidation products. Reactions involving oxygen proceed by a chain mechanism. Basic experiments on autoxidation (i.e., self-accelerating oxidation) involving kinetic investigations of low molar mass saturated hydrocarbons in the liquid phase were carried out by Bolland and Gee [62, 63].

Initiation reaction

$RH \rightarrow R\cdot$ (1)

Chain propagation

$R\cdot + O_2 \rightarrow ROO\cdot$ (2)

$ROO\cdot + RH \rightarrow ROOH + R\cdot$ (3)

Chain branching

$ROOH \rightarrow RO\cdot + HO\cdot$ (4)

$RO\cdot + RH \rightarrow ROH + R\cdot$ (5)

$HO\cdot + RH \rightarrow H_2O + R\cdot$ (6)

Chain termination

$2\,R \rightarrow R{-}R$ (7)

$R\cdot + ROO \rightarrow ROOR$ (8)

$2\,ROO\cdot \rightarrow ROOR + O_2$ (9)

Although the exact mechanism of the initiation reaction (1) is still unclear, the alkyl radical R· is generally assumed to be formed by absorption of energy in the form of heat or UV radiation or by reaction with traces of heavy metals. The actual chain carriers in the chain propagation reaction are alkyl radicals R· and peroxy radicals ROO·. The peroxy radicals are less reactive and

therefore quite selective. They preferably attack very reactive C – H bonds (those next to an aromatic nucleus or in the allyl position, tertiary C – H bonds, or C – H bonds adjacent to oxygen as in aldehydes or ethers). The chain branching step (4) requires high activation energy and is therefore the main reaction at high temperature. The radicals RO· and HO· formed in reaction (4) are capable of reacting with RH – like the peroxy radicals – and thus can also initiate new chain reactions (reactions 5 and 6); the rate of oxidation becomes markedly higher (kinetic chain branching). Chain termination can occur by recombination of radicals, which may however decompose again and initiate new radical chain reactions.

The sensitivity of high molar mass polymers to oxidation depends first on their content of double bonds and second on the groups (atoms) next to the double bonds. Polymers with a high double bond content, such as NR and BR, are therefore more sensitive to autoxidation than EPDM or EVM with their completely saturated main chains. Methyl substituents next to double bonds [NR, isoprene rubber (IR)] favor autoxidation, while the shielding effect of chlorine in polychloroprene gives rise to better resistance to aging. Tertiary carbon-atoms, which are readily converted into hydroperoxides, constitute weak points in the methyl-substituted products.

The oxygen absorbed can initiate the following reactions within the article depending on the type of rubber:

1) Cleavage of the polymer chain, whereby the vulcanizate network is weakened (degradation, softening)
2) Cross-linking, whereby the vulcanizate network density is increased (cyclization, hardening)
3) Chemical bonding of the oxygen (e.g., as hydroxyl or keto groups), without chain cleavage or formation of new cross-links

The relative extent of occurrence of these three competing reactions determines the amount of changes in the mechanical properties that occur during aging. The first two reactions result in large changes, whereas the third is comparatively harmless.

The aging behavior of natural rubber, synthetic isoprene, and butyl rubber can vary depending on the conditions chosen. In contrast, with vulcanizates based on SBR, BR, NBR, EPDM, CR, and many other synthetic rubber types, the second reaction (cross-linking) predominates from the beginning, so that hardening occurs. Because of this, an increase in modulus and a decrease in elongation at break are observed. On hot-air aging at 150 °C, butyl rubber exhibits considerable softening after 13 d (chain cleavage), whereas EPDM hardens. Completely oxidized natural rubber vulcanizates are generally hard, brittle, and thermoplastic. The decrease in the elongation at break is generally used to characterize the aging resistance of polymers. For natural rubber, the decrease in tensile strength is often also evaluated.

Aging resistance depends not only on the type of rubber and the antidegradants, but also the cross-linking system used. In sulfur vulcanization, aging resistance depends on the distribution of mono-, di-, and polysulfidic cross-links in the network. With an increasing proportion of mono- and disulfidic cross-links, aging resistance in the narrow sense (aging due to oxygen and heat) increases. Sulfenamides and, in particular, mercapto accelerators give vulcanizates with better aging resistance than those produced with dithiocarbamates. Vulcanizates produced with sulfur donors are more resistant to aging than those produced with sulfur itself. Thiuram vulcanization also gives good aging resistance. Factice and reclaimed rubber favorably affect resistance to aging as, in general, do high dosages of white fillers. Vulcanizates cross-linked without sulfur (using quinone dioxime, peroxides, or resins) have the greatest aging resistance.

Aging due to Rubber Poisons. The activation energy of the chain reactions in autoxidation is lowered by the catalytic action of heavy-metal ions. Uncontrolled acceleration of the oxidation is initiated at much lower temperature than in the absence of heavy-metal ions. This catalytic action is due to formation of radicals from the alkyl hydroperoxides (hydroperoxide decomposition) and to the formation of radicals in the initiation phase (see below). The most well known rubber poisons are iron [particularly Fe(II)], copper, manganese, cobalt, and nickel.

Initiation phase

$$Me^{2+} + RH \rightarrow Me^{+} + H^{+} + R\cdot$$

$$Me^{2+} + RH \longrightarrow Me^{+} + H^{+} + R^{\bullet}$$
$$Me^{+} \xrightarrow{(O_2)} Me^{2+}$$

Hydroperoxide decomposition

$$Me^{2+} + ROOH \rightarrow Me^{+} + H^{+} + ROO^{\bullet}$$
$$Me^{+} + ROOH \rightarrow Me^{2+} + OH^{-} + RO^{\bullet}$$

Natural rubber and polyisoprene are considerably more sensitive toward rubber poisons, particularly copper and manganese, than synthetic rubbers such as SBR, BR, and NBR. Polychloroprene is in an intermediate position with regard to sensitivity. The chemical form in which the heavy-metal ions are present is very important. The solubility of the metal compound in the rubber seems to play the major role. Thus, the soluble compounds copper stearate and oleate are powerful aging accelerators. After these, in order of decreasing reactivity, are copper sulfate and chloride, copper oxide, and metallic copper. Complexes such as copper phthalocyanine are virtually inactive and harmless, as is, for example, copper dimethyldithiocarbamate. The same applies to iron compounds, so that the iron oxide pigments and complexes (e.g., iron ethylenetetramine) commonly used in the rubber industry can be regarded as harmless. The accelerator system is also important for the sensitivity of vulcanizates toward rubber poisons. Vulcanizates produced with basic accelerators are more sensitive than those produced with acidic ones. In thiuram vulcanizates, copper compounds can even act as antioxidants or at least behave neutrally. Whether the poison enters the rubber during compounding (e.g., with the fillers) or from external contact (e.g., bare copper tubing, textiles containing copper dyes) does not make any difference in principle. As a rule of thumb, the content of rubber poisons should be <0.001 %.

Heat Aging. At temperatures of more than 100 °C even in the presence of oxygen, the heat factor is considered most important, and at this temperature heat aging occurs. The action of heat alone produces post- cross-linking, cyclization, or reversion. With NR and IR, *reversion* predominates after longer warming periods (i.e., thermal degradation occurs). This is sometimes known as overcure.

CR and EPDM are typical examples in which inter- or intramolecular cross-linking (hardening) continues at higher temperature. This *post- cross-linking* is caused by the vulcanization agent, as well as by reactive groups in the rubber itself. BR, SBR, and NBR also undergo hardening, but on strong overheating, softening begins. The term *cyclization* is not always unambiguous. Here, it should be understood as additional cross-linking of the vulcanizate. Cyclization can also be important in hot processing of raw synthetic rubber or its compounds (SBR, NBR) in the absence of cross-linking agents. In this case an increase in viscosity occurs.

Besides these reactions, *rearrangement of network bonds* can occur without changing their effective number. The physical effects observed are small.

Hot Water and Steam Aging. A special case of heat aging – the effect of steam or hot water on the vulcanizate – results in partial or total exclusion of air. In particular, rubber that is not sensitive to hydrolysis can be exposed in water or steam to higher temperature than in hot air. Silicone rubber, ethylene – vinyl acetate polymers, polyurethanes, and some other polymers with ether or ester bonds are, however, degraded quite rapidly. The C – O, C – N, and Si – O bonds are cleaved hydrolytically.

Aging in Hot Media is also a special case of heat aging. In addition to the heat effects already described, the simultaneous action of swelling and aggressive media must be considered. Besides the possible extraction of compounding ingredients such as plasticizers (changes in hardness, elongation at break, etc.) and antidegradants, the resistance of the polymer toward the medium concerned is important.

Dynamic Cracking under the Effect of Oxygen. If rubber is subjected to continuous changes in mechanical stress (e.g., periodically bend backwards and forwards), cracks gradually appear in the surface, which increase in size and can finally lead to complete breakage of the article. These cracks occur perpendicular to the direction of stress. Natural rubber forms these cracks relatively rapidly, but they then develop slowly. Crack formation occurs later in SBR and polybutadiene, but is followed by rapid growth. This is associated with lower tear resistance. Raising the temperature accelerates cracking, as does increasing the degree and frequency of de-

formation. Flex cracking or fatigue occurs particularly rapidly if the deformation cycle is not sinusoidal but shows leaps and bounds.

Even the fillers and the cross-linking system have an effect on the fatigue behavior of vulcanizates. Carbon black is more advantageous (i.e., it counteracts dynamic cracking to a certain extent) than inactive, white fillers. With the sulfur cross-linking system the ratio of mono-, di-, and polysulfidic cross-links plays a decisive role, because the latter are more fatigue resistant than the former.

Static and Dynamic Cracking due to Ozone. If vulcanizates in the statically elongated state or under dynamic load are exposed to weathering, cracks gradually appear perpendicular to the direction of elongation (tension). These cracks gradually increase in size and can also lead to breakage of the vulcanizate. This phenomenon is caused by the small quantity of ozone present in the atmosphere. Investigations have shown that the Criegee mechanism [64] of ozonolysis of olefins is also valid for rubber [65, 66] (see Fig. 6).

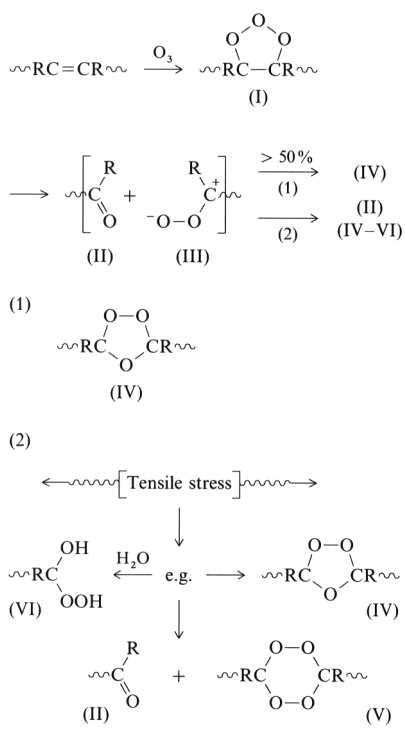

Figure 6. Ozonolysis of polymers

In a 1,3-dipolar cycloaddition the double bond is attacked by ozone. The unstable primary ozonide I decomposes simultaneously into a carbonyl compound II, and a carbonyl oxide III. In the unextended – not stressed state, more than 50 % of the cleaved fragments again undergo a 1,3-dipolar addition, this time to give the actual ozonide IV. In the extended state, these cleaved fragments are removed from the reaction cage and can, for example, react with neighboring molecular chains or other molecules. In molecular chains, which are arranged parallel to the direction of tension, this process leads to microscopically small cracks perpendicular to the direction of stress.

Without extension and stress no cracks are formed in the vulcanizate. A critical extension – which is <10 % for NR – must be exceeded. The number of cracks formed per unit area and time increases with increasing extension, but their length and depth decrease. An increase in temperature has the same effect as an increase in extension: the number of cracks increases, and their depth decreases. In accordance with the reaction mechanism the most highly unsaturated rubbers, such as NR, SBR, BR, IR, and NBR, are the most sensitive toward ozone cracking. CR is much more resistant, because of the shielding effect of the chlorine atom. Completely saturated elastomers naturally have the best ozone resistance.

Although it has been known for some time that light is not to be the major cause of this type of cracking, the expression "sun cracking" is persistently used. This is correct only in so far as high-energy radiation (UV light) produces ozone, whose effect is more pronounced in mountains than in low-lying land and is stronger in spring or summer than in winter. All types of electrical discharge (spark formation on electric motors or electric arcs) cause locally high ozone concentrations, as does gasoline exhaust.

Crazing. On prolonged exposure of light-colored vulcanizate to sun, a network of connected cracks or grooves without a definite orientation is formed on the surface (crazing). The surface resembles a wrinkled orange or an elephant's skin. The process becomes slower as irradiation continues. Destruction of the vulcanizate generally does not occur, but the surface can become completely brittle so that it can no

longer bind the white filler (chalking). Crazing occurs only in light-colored vulcanizates, not in pigmented ones or in those filled with carbon black. It is caused by light- catalyzed oxidation of the vulcanizate surface. High-energy, short-wavelength light is more effective than long-wavelength light. If the high-energy blue and violet rays are filtered off, hardly any damage occurs. Therefore, yellow and red pigments have a favorable effect here. The protective effect of carbon black results from its complete light adsorption.

Frosting. Frosting is a special aging effect. Previously glossy or colorless rubber surfaces become mat or acquire a white coating under certain, not fully understood conditions. The frosting phenomenon occurs in a warm, damp, ozone- containing atmosphere, and only in vulcanizates with white fillers, mainly silica. Acidic accelerators promote the effect, as do thiurams and dithiocarbamates, whereas basic accelerator systems appear not to [67, 68].

3.2. Classification and Types of Antidegradants [69 – 75]

(see also → Antioxidants)

The resistance of a vulcanizate toward the aging effects described is determined primarily by the elastomer forming the basis of the compound. The cross-linking system, the cross-link density, and to a lesser extent, the antidegradant used also affect aging resistance.

The mechanism of action of the antidegradant depends on the presence of active hydrogen-atoms in the molecule. These are present in the form of OH, NH, and SH groups, which act as chain terminators in the radical chain reaction initiated by oxygen, for example. The radical that is formed from the antidegradant after hydrogen splitting is resonance stabilized. Because it is very low in energy and sterically hindered, it cannot initiate a new radical chain reaction. It either reacts with another antidegradant radical (dimerization) or combines with a peroxy or alkyl radical (see Fig. 7). In a single autoxidation reaction chain about 10 – 100 oxygen molecules can be transferred to the polymer substrate. The antidegradant never competes with the substrate with regard to direct oxygen absorption. Much higher dosages would be necessary for this.

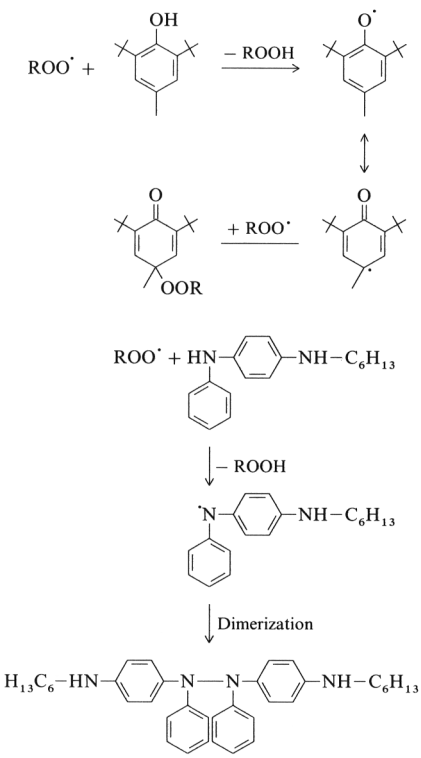

Figure 7. Mechanism of action of antidegradants

For *ozone protection* a scavenger protective film mechanism is assumed, based on the work of Lattimer et al. [76 – 78]. The antiozonant competes with the double bonds of the polymer for ozone. The effectiveness of the antiozonant also depends on its mobility.

With a few exceptions, antidegradants can be subdivided into mono- and oligofunctional aromatic amines, mono- and oligofunctional phenols, and heterocyclic mercapto compounds. For the user, their different mechanism of action is more important. Correspondingly a differentiation is made among *antioxidants*, *antiozonants*, and *anti-flex- cracking agents*. In addition, the action of the antidegradants against discoloration of the vulcanizate itself and against contact staining is also important.

Staining products are more effective than nonstaining ones, but this is not always true.

Aromatic amines are generally staining antidegradants, and phenols nonstaining ones.

The following groups can be distinguished.

3.2.1. Staining Antidegradants that Act as Anti-Flex-Cracking Agents and Antiozonants

p-Phenylenediamines substituted at nitrogen are the most effective antioxidants, anti-flex- cracking agents, and antiozonants under static and dynamic stress. They increase the critical energy necessary for the formation of ozone cracks under static conditions. Cracking therefore occurs only at greater elongation. At the same time the cracking rate under static and dynamic stress is decreased [79].

The range of activity of individual *p*-phenylenediamines (**28**) depends on the nature and size of the substituents on the nitrogen atoms.

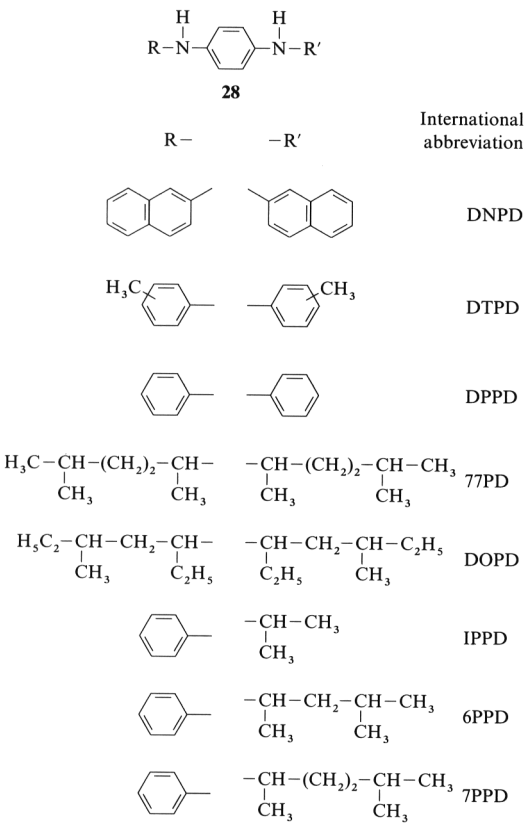

Symmetrical *N*,*N*′-diaryl-*p*-phenylenediamines are very effective antidegradants. For example, *N*,*N*′-di-β-naphthyl-*p*-phenylenediamine (known internationally as DNPD) is the most powerful antioxidant known for rubber. However, it is no longer produced for toxicological reasons. Because of the size of the substituents its mobility in rubber is very low, so that DNPD is, in practice, not transported to the surface to protect against fatigue and ozone. If smaller substituents are used (e.g., *N*,*N*′-ditolyl-*p*-phenylenediamine, DTPD), the mobility is somewhat higher, and thus the protection against fatigue and ozone is better, but the compound is a poorer antioxidant [80, 81]. *N*,*N*′-Diphenyl-*p*-phenylenediamine (DPPD) [*74-31-7*] shows similar effects. Because of its low solubility in rubber, and therefore its greater tendency to bloom, it is seldom used for rubber.

Of the symmetrical dialkyl derivatives, *N*,*N*′-diisopropyl- and *N*,*N*′-di-*sec*-butyl-*p*-phenylenediamine are particularly effective in counteracting ozone cracking. They are, however, not used in rubber because of their high volatility and skin irritating action. *N*,*N*′-Bis(1,4-dimethylpentyl)-*p*-phenylenediamine (77PD) [*3081-14-9*] or *N*,*N*′-bis(1-ethyl-3-methylpentyl)-*p*-phenylenediamine (DOPD) exhibit better properties. Compared with alkylaryl-*p*-phenylenediamines, the symmetrical dialkyl compounds have the advantage of being particularly good antiozonants under static stress. Their disadvantages are their relatively low protective action against oxygen and heat, their unfavorable effect on onset of cure, and their high volatility.

Under static and dynamic stress, *N*,*N*′-alkylaryl compounds exhibit greatest activity when the alkyl group is an isopropyl or *sec*-butyl group and the aryl is a phenyl group. The most active compound is *N*-isopropyl-*N*′-phenyl-*p*-phenylenediamine (IPPD) [*101-72-4*]. *N*-1,3-Dimethylbutyl-*N*′-phenyl-*p*-phenylenediamine (6PPD) [*793-24-8*] is less volatile, causes less water elution from the rubber article (e.g., tires), and thus has a prolonged activity. However, it is a little less active than IPPD, and protection against fatigue is somewhat lower. *N*-1,4-Dimethylpentyl-*N*′-phenyl-*p*-phenylenediamine (7PPD) [*3081-01-4*] is comparable to 6PPD but less important in market terms. Sensitization on skin contact

is possible with IPPD, so labeling as harmful to health is obligatory. This does not apply to 6PPD. IPPD is therefore being increasingly replaced by 6PPD. Longer-chain alkyl groups (e.g., 2-methylheptyl groups) reduce the volatility still further, but the action against ozone and flex cracking (fatigue) is considerably less at the same dosage (in terms of weight) [82].

An addition to *p*-phenylenediamines, 6-ethoxy-2,2,4-trimethyl-1,2-dihydroquinoline (**29**) (ETMQ) [*95-31-8*] also belongs to the group of agents that protect against autoxidation, fatigue, and ozone. Its activity against static and dynamic cracking is comparatively weak.

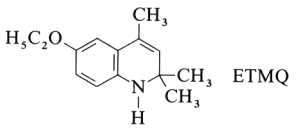

29

Long-term protection is becoming increasingly important technologically. Volatility and water elution play a decisive role, particularly in truck tires that are retreaded many times. Here, combinations of 6PPD and DTPD have proved successful. Antidegradants based on alkylaryl-*p*-phenylenediamines, which are or can be polymer bound, have also been investigated. In agreement with the scavenger protective film model, these compounds are very good antioxidants, average anti-flex-cracking agents, and poor antiozonants [83 – 87]. To better understand the action of *p*-phenylenediamines, the effect of substituents on the protective function was also investigated. The arylalkyl-substituted NH group is primarily responsible for antiozonant properties, and the bisaryl-substituted NH group for protection against fatigue. Antioxidant properties are comparable for the two groups [88].

3.2.2. Staining Antidegradants with Protection against Fatigue, but Not against Ozone

The diarylamines (**30**), such as phenyl-α-naphthylamine (PAN) [*90-30-2*], phenyl-β-naphthylamine (PBN) [*135-88-6*], and substituted diphenylamines, such as octylated diphenylamine (ODPA) or styrenated diphenylamine (SDPA) [*68442-68-2*], are important antidegradants.

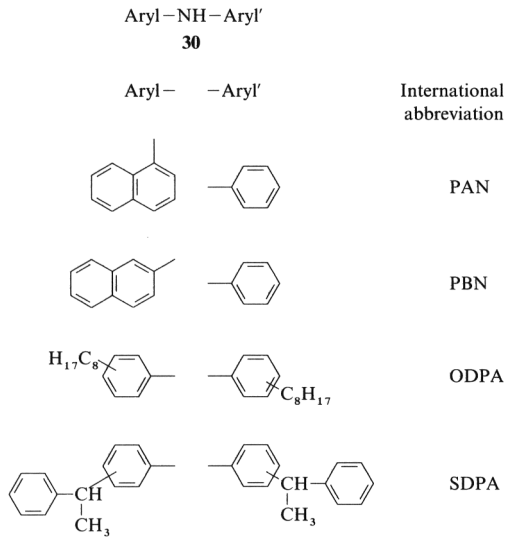

This group protects mainly against oxygen and heat. Because traces of carcinogenic β-naphthylamine are always present in PBN, PBN is seldom used. In addition, PBN has the disadvantage of blooming because of its low solubility in rubber. The level of activity of all antioxidants in this group is about the same in tire rubbers (natural rubber, synthetic polyisoprene, polybutadiene, and styrene – butadiene rubber). ODPA and SDPA are significantly more active in polychloroprene than the other products. In general, ODPA, SDPA, and other diphenylamines offer the best protection against oxygen and heat for chloroprene rubbers.

PAN, PBN, ODPA, and SDPA are effective anti-flex-cracking agents, although not to the same extent as *p*-phenylenediamines. PAN and PBN give about the same protection; that of ODPA and SDPA is a little lower. However, rubbers respond differently to these antidegradants. In NR and IR the activity is high, whereas in SBR and BR these products provide low protection. The tendency of PAN and PBN to stain is lower than that of *p*-phenylenediamines. However, it is still high enough that these antidegredants can only be applied in dark compounds. ODPA and SDPA are weakly staining products (yellow to brownish yellow); only weak contact staining is observed. Thus, ODPA and SDPA can be used

without any problem for brown and red vulcanizates.

3.2.3. Staining Antidegradants with Little or No Protection against Fatigue and None against Ozone

As has already been mentioned, for protection against fatigue or ozone, a certain mobility of the antidegradant molecule in the substrate is necessary. Products in which this mobility is limited are antidegradants only in the narrow sense.

For example, *N*,*N′*-di-*β*-naphthyl-*p*-phenylenediamine (DNPD) protects only against oxidation, not ozone or fatigue. 2,2,4-Trimethyl-1,2-dihydroquinoline (TMQ) [*26780-96-1*] is present in oligomeric form (**31**) and thus not very mobile.

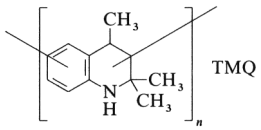

31

It is a good antioxidant, but gives low protection against fatigue. Because of its oligomeric form, TMQ is nonvolatile and is therefore an antidegradant with effective heat protection. Despite its low migrating capacity, vulcanizates containing TMQ can give rise to contact staining. This is, however, lower than in the case of PAN. A special feature of TMQ is its relative inertness toward cross-linking peroxides. It can therefore be used as an antidegradant in peroxidically cross-linked EPDM, EPM, or PE.

Diphenylamine – acetone condensation products (ADPA) [*68412-48-6*] are also effective heat-protection agents, but they have been forced off the market to a great extent by TMQ.

3.2.4. Nonstaining Antidegradants with Protection against Fatigue or Ozone

Nonstaining antidegradants are generally inferior to the staining products under different aging stresses.

Aralkylated phenols (**32**), e.g., styrenated phenols (SPH) [*61788-44-1*], and alkylated, styrenated phenols are nonstaining antioxidants that protect against fatigue but not ozone.

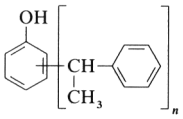

$n = 1,2,3$

32

These monofunctional phenols are relatively low priced and exhibit medium-strong protection toward oxygen and heat. Protection against flex cracking is not comparable with that of *p*-phenylenediamines, but is higher than that of other phenols. These products impart good protection against crazing to light colored vulcanizates.

A range of new products has been developed and presented as nonstaining antioxidants with protection against ozone. None of them are used successfully in the market, however.

3.2.5. Nonstaining Antidegradants without Protection against Fatigue or Ozone

Certain products protect against oxygen and heat, crazing, and to some extent, heavy-metal rubber poisons but show little or no effect against flex cracking, and none against ozone cracking. These antidegradants are mostly phenols, but some are heterocyclic mercapto compounds.

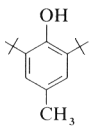

BHT

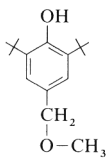

33 **34**

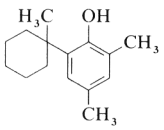

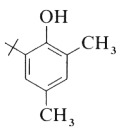

35 **36**

2,4,6−Substituted monophenols

Of the *2,4,6-substituted monophenols* (**33**–**36**), 2,6-di-*tert*-butyl-4-methylphenol [*128-37-0*] (**33**) (BHT) is widely used. However, because of its high volatility, it is an effective antioxidant only at relatively low temperature. Phenols, such as 6-*tert*-butyl-2,4-di-

methylphenol (**36**), 2,4-dimethyl-6-(α-methylcyclohexyl)phenol (**35**), and 2,6-di-*tert*-butyl-4-methoxymethylphenol (**34**) exhibit better thermal stability and thus considerably better protection. Other substituted monophenols, whose structures have not been published, have been marketed by various companies.

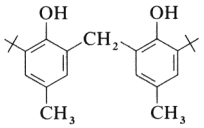

BPH
37

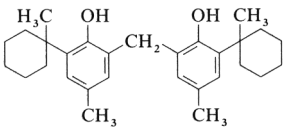

38

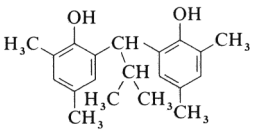

39

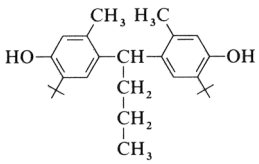

40

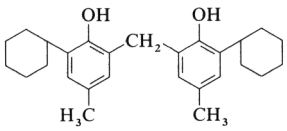

41

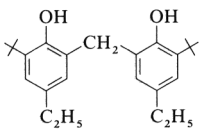

42

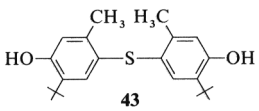

43

Of the *bifunctional phenols*, 2,2′-methylenebis(6-*tert*-butyl-4-methylphenol) (BPH) [*119-47-1*] (**37**), 2,2′-methylenebis(6-α-methylcyclohexylphenol) (**38**), 2,2′-isobutylidenebis(4,6-dimethylphenol) [*33145-10-7*] (**39**), and 4,4′-butylidenebis(2-*tert*-butyl-5-methylphenol) [*85-60-9*] (**40**) offer excellent protection against oxygen. However, with the exception of the two last-named compounds, prolonged illumination of the vulcanizate can produce a pink coloration. 2,2′-Methylenebis(6-cyclohexyl-4-methylphenol) (**41**) or the corresponding 6-*tert*-butyl-4-ethylphenol derivative (**42**) and 4,4′-thiobis-(6-*tert*-butyl-3-methylphenol) [*96-69-5*] (**43**) [89, 90] have a lower staining tendency but are less active.

The reaction product of dicyclopentadiene and 2-*tert*-butyl-4-methylphenol is also a bifunctional product. It has a low staining tendency, particularly in an oxidizing medium (e.g., when latex articles are treated with oxidizing cleaning agents). Its main applications are consequently also in the latex area. The mono- and bifunctional phenols named are added as antioxidants in the compounding of solid rubber, and they are also used as so-called *stabilizers* in polymer production. Polymers can be damaged thermooxidatively in processing (e.g., during drying) and must therefore be protected. The quantity of stabilizer added here is lower than in later compounding. Besides direct protection during production, stabilizers also protect the polymers on storage until use. The stabilizers are generally pure antioxidants, but combinations of nonstaining antioxidants with organic phosphites such as tris(nonylphenyl) phosphite (TNPP) are also used. A disadvantage of phosphites is their low stability to hydrolysis. Liquid stabilizers are preferred for better handling. One example of these is 6-methyl-2,4-bis(octylthiomethyl)-phenol [*110553-27-0*] (**44**).

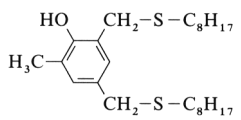

44

Trifunctional or *polyfunctional phenols* are characterized by extremely low volatility and are therefore used for polymers that are processed at high temperature. Examples of trifunctional phenols are tris-1,1,3(5′-*tert*-butyl-

4′-hydroxy-2′-methylphenyl)butane (**45**), and 1,3,5-trimethyl-2,4,6-tris(3′,5′-di-*tert*-butyl-4′-hydroxybenzyl)benzene (**46**). An example of a polyfunctional phenol is pentaerythrityl-tetrakis[3-(3,5-di-tert-butyl-4-hydroxyphenyl) propionate] (**47**).

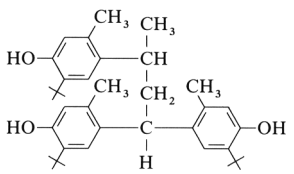

45

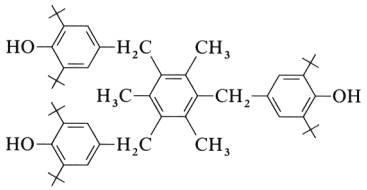

46

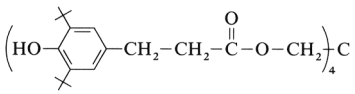

47

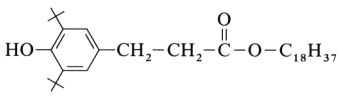

48

These phenols are used less with diene rubbers and mainly with polyolefins such as polyethylene and polypropylene. A monophenol of low volatility, octadecyl-3-(3,5-di-*tert*-butyl-4-hydroxyphenyl) propionate (**48**) is also used. In PE or PP these phenols are used alone or in combination with nonvolatile, hydrolysis-stable, organic phosphites [e.g., tris(2,4-di-*tert*-butylphenyl) phosphite (**49**)].

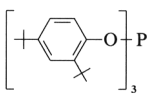

49

The phosphites are also known as processing stabilizers because they protect thermoplastics from polymer degradation during multiple extrusion.

Mercaptobenzimidazole (MBI) (**50**), its zinc salt (ZMBI) (**51**), 4- and 5-methyl-2-mercaptobenzimidazole (MMBI), and their zinc salts (ZMMBI) are antidegradants of average activity.

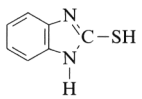

MBI
50

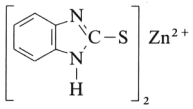

ZMBI
51

However, they are generally not used alone. They are of great importance because they act synergistically with almost all other antidegradants (i.e., the activity of the combination is greater than the sum of the activities of the individual antidegradants).

All products in this group give some protection against copper and manganese rubber poisons. MBI and MMBI or their zinc salts exhibit low activity when applied alone, while a combination of MBI and a bisphenol such as BPH, for example, results in synergistic action. These combinations are therefore always recommended when strong protection against rubber poisons and favorable staining behavior are required. The mercaptobenzimidazoles can be regarded as tautomeric thioureas. This explains why they also act as accelerators, particularly in CR vulcanization. The zinc compounds have less effect on vulcanization kinetics. Like TMQ, MBI and MMBI and their zinc salts interfer in cross-linking with peroxides only to a minor extent.

3.2.6. Nonstaining Antiozonants without Protection against Aging

The production of ozone-resistant, light colored rubber articles is a challenge for the compounder. Within certain limits, light colored rubber articles can obtain some physical protection by the addition of *waxes*. Waxes bloom on the surface of the article above a certain concentration (in natural rubber, ca. 0.8 %) and form a physical barrier against attack by ozone. If the film is damaged, the protective action is lost in that location, and the cracks formed there are often wider and deeper than they would be without the wax. Microcrystalline waxes show improved

action, particularly at high temperature. Waxes alone have no effect in counteracting dynamic crack formation.

The *combination of wax and antidegradants* can lead to considerably improved protection against aging. Antiozonants, antioxidants, and anti-flex- cracking agents are enriched at the vulcanizate surface by the migration of antiozonant waxes. This is most important for good protection against ozone.

Another possibility for the compounder involves *blending diene rubbers with ozone-resistant polymers*. Blends of NBR with PVC are commercially available, but blends of diene rubbers with CR or EPDM are also worth considering. However, attention must be paid to compatibility.

Besides some benzofuran derivatives, special ethers (**52**) and acetals (**53**) with olefinic double bonds have been found to be *chemically active, nonstaining antiozonants*.

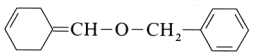

52

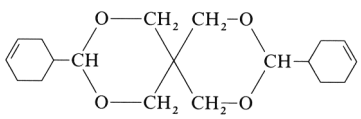

53

Since they contain no active hydrogen, they do not act as chain terminators in radical chain reactions, but the olefinic double bond in the reagent competes for ozone with the double bonds in the rubber. In diene rubbers such as NR and SBR, these special compounds are active only in combination with wax. Without addition of wax they can be used only in CR, halogenated polymers (e.g., HIIR, CM) and its blends.

3.2.7. Other Antidegradants

Polymers containing ester groups, such as ethylene – vinyl acetate copolymers, and also polyurethanes, can be attacked by moisture, acid, base, and sometimes elevated temperature. By the addition of, for example, polycarbodiimides these effects can be inhibited and the durability considerably improved.

3.3. Choice of Antidegradants

An overview of the spectrum of activity of the most important antidegradants is presented in Table 3. The table can be used to estimate which products are suitable for different requirements and for a particular combination.

Nearly all antidegradants can be employed to prevent *autoxidation* of vulcanizates, with the exception of the special antiozonants without protection against aging, or agents giving protection against hydrolysis and light.

The number of suitable *anti-flex- cracking agents* is more limited. An anti-flex- cracking agent is generally active against autoxidation, while the reverse is not true.

There are even less suitable *antiozonants* available. Staining antiozonants are mainly also anti-flex- cracking agents, but not every anti-flex- cracking agent is an antiozonant. Apart from waxes and nonstaining antiozonants, antiozonants are generally also active against heat and oxygen. However, not every antioxidant is active against ozone cracking.

The *level* of antidegradants in the vulcanizates depends on the usage condition of the compound and on the antidegradant itself. Antidegradants show a pronounced plateau effect in their action against autoxidation (i.e., above a certain level the protection against autoxidation is not improved by increasing the dosage). Typical levels are between 0.5 and 2.5 phr. For protection against ozone or fatigue, a plateau effect occurs only beyond 4 – 6 phr; the activity increases continuously with increasing level. The usual levels are therefore higher, between ca. 1.0 and 3.5 phr.

The protective effects achieved are more pronounced in natural rubber than in synthetic, partly because of the generally better resistance of synthetic rubber to aging, and partly because of the stabilizers already added during production, which act as antidegradants. Both staining antidegradants (e.g., *p*-phenylenediamines) and nonstaining ones (e.g., mono-, bi-, and polyfunctional phenols) are used as stabilizers. The dosage depends on the activity of the substance and the desired protective effect, and is between 0.1 and 1.0 %. Staining stabilizers are almost always superior to nonstaining ones, particularly with regard to heat resistance and inhibition of

Table 3. Spectrum of activity of important antidegradants

International abbreviation	Autoxidation [a, b]	Heat [a, b]	Flex cracking [a, b]	(Static) Ozone cracking [a, b]	Metal contamination [a]	Crazing [a]	Staining [a]	Contact staining [c]	Contact with foods [c]		State
									BGA	FDA	
IPPD	2	2 – 3	1	1 – 2	2	6	5 – 6	5	cat. 4	+	solid
6PPD	2	2 – 3	1 – 2	2 [d]	2	6	5 – 6	5	cat. 3 + 4	–	solid
77PD	3 – 4	3 – 4	2	1		6	5	5	cat. 4	–	liquid
DTPD	2	2 – 3	2	3 [d]	2	6	5	4	cat. 4	+	solid
TMQ	2	1 – 2 [e]	4 – 5	6	3 – 4	6	3	2	cat. 4	–	solid
PAN	2	2 – 3	2 – 3	6	2 – 3	6	5	4	cat. 4	+	solid
PBN	2	2 – 3	2 – 3	6	3 – 4	6	5	4	cat. 4	+	solid
SDPA	2 – 3	2 [d]	3 – 4	6	3	6	3	2	cat. 1	+	liquid
ODPA	2 – 3	2 [d]	4	6	3	6	2 – 3	1 – 2	cat. 4	+	solid
BPH	2 – 3	3	6	6	3 [e]	1	1	0	cat. 1 – 4	+	solid
Reaction product of butylated *p*-cresol with dicyclopentadiene	3	3	6	6	4 [f]	1 – 2	0	0	cat. 4	+	solid
BHT	3 – 4	4 – 5	6	6	4 – 5	1	0	0	cat. 1 – 4	+	solid
SPH	4	4	4	6		3	0	0	cat. 1 – 4	+	liquid
MBI	4 [g]	3 [h]	6	6	6 [i]	6	0 [j]	0	cat. 4	–	solid
MMBI	4 [d]	3 [h]	6	6	6 [i]	6	0 [j]	0	cat. 4	–	solid
ZMMBI	4 [g]	3 [h]	6	6	6 [i]	6	0 [j]	0	cat. 4	–	solid

[a] 1 = best, 6 = worst.
[b] NR and IR, etc.
[c] cat. = category; 0 = nonstaining; 6 = very pronounced staining; – not tested or meaningless in this context.
[d] In CR: 1.
[e] In combination with MBI or MMBI: 1.
[f] In combination with MBI or MMBI: 2.
[g] In compounds accelerated by dithiocarbamate.
[h] In combination with IPPD or TMQ: 1 (compound dependent); in steam also without other antidegradants: 1.
[i] In combination with BPH: 1.
[j] Brightening.

gel formation. There are also special stabilizers (e.g., aryl phosphites), which are not antidegradants for vulcanizates.

Producers and Trade Names. The following are producers of antidegradants and stabilizers for rubber: Akzo (Permanax), Bayer (Vulkanox), Bozzetto (Anox), Chemische Werke Lowi (Lowinox), Ciba-Geigy (Irganox, Irgafos), Goodrich (Good-Rite), Goodyear (Wingstay), Kawaguchi (Antage), Monsanto (Santoflex, Santowhite), Raschig (Ralox), Seiko (Nonflex), Sumitomo (Antigene, Sumilizer), and Uniroyal Chemicals (Flexone, Naugard).

4. Fillers and Pigments

4.1. Typical Properties and Types of Fillers [91 – 96]

4.1.1. Action of Fillers

The properties of industrial rubber compounds and their vulcanizates are determined not only by the nature and quantity of the polymers used, the cross-linking system, the plasticizers, and the antidegradants but also by the nature and quantity of the fillers employed. Non-self-reinforcing rubbers such as SBR and BR give unsatisfactory and hardly usable vulcanizates without the addition of reinforcing fillers (white fillers or carbon blacks).

Fillers are classified with regard to their chemistry, first into white fillers and carbon blacks, and second into natural products and synthetic ones. Technologically, a distinction is

made between active or reinforcing fillers and inactive or extender fillers. The former are used to achieve particular vulcanizate properties; the latter, to make the compound cheaper (to extend it). The varying reinforcing effects of fillers *enhance* a range of vulcanizate properties, such as toughness, stress, tear resistance, compound viscosity, and hardness, whereas other properties, such as elongation at break and impact resilience are *impaired.* The activity of a filler is determined by particle size (the smaller the size, the greater is the activity), the geometrical shape of the particles (structure), and the chemical composition. If the dependence of vulcanizate properties on the quantity of filler is investigated, a steady change in properties is observed for inactive fillers, whereas for active fillers, as might be expected, increasing the proportion of filler leads to a maximum, which is characteristic in height and position for the individual filler (see Fig. 8).

At the same level, active fillers give harder vulcanizates, higher tensile strength, better tear resistance, and lower abrasion loss than inactive ones. On addition of inactive to medium-active fillers the *elastic properties* of vulcanizates deteriorate only slightly; they are even improved with finely divided zinc oxide. With active fillers, worsening of elastic properties is often observed. A comparison of activity of various fillers gives conflicting results with regard to modulus. Highly active precipitated silicas only moderately improve the modulus because of their special effect on the vulcanization behavior, and accordingly would have to be classified as semiactive. This would contradict their other behavior in rubber. Besides vulcanizate properties, the mixing temperature during production and the compound viscosity are measures of the activity of a filler. The more sharply the *viscosity* decreases with an increasing proportion of filler, the more active the filler is. For compounding, too high a viscosity is a disadvantage, not only because of the high mixing energy costs, but above all because further processibility (extrusion, calendering, injection molding, etc.) is impaired. A highly active filler can therefore only be added in smaller quantities than an inactive or semiactive one. For classification of a filler, the entire property profile of the green compound produced with the filler and its vulcanizates must be considered.

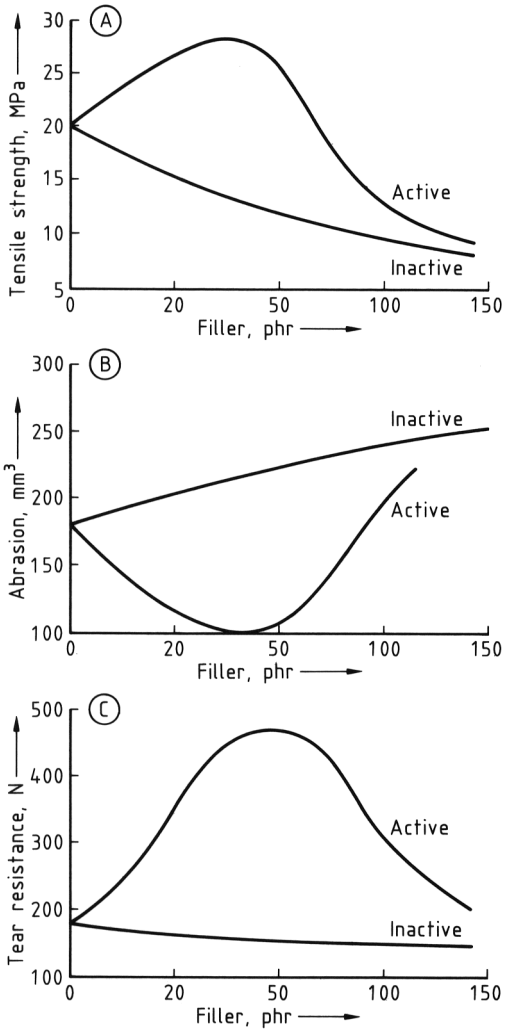

Figure 8. Effect of active and inactive fillers
A) On tensile strength; B) On abrasion; C) On tear resistance of natural rubber

If a filler – rubber mixture is extracted with a suitable solvent (benzene, toluene), and after drying of the the extraction residue, a hard, brittle gel is obtained, which contains all the filler. The rubber content of this gel is known as *bound rubber*. The quantity of bound rubber formed on mixing depends on many factors, such as type of filler, degree of filling, polymer type, mixing technique, and extraction conditions. In the case of carbon blacks, bound rubber is a measure of filler activity, but it is not used for assessment, unlike other physical properties such

as surface area and structure, whose dependence on the compound and vulcanizate properties are known.

4.1.2. Properties of Fillers and Their Significance [97]

The optimum level of fillers to achieve a certain compound property can be found by determining the vulcanizate properties as a function of the degree of filling. Since this method is very laborious and time- consuming, a series of analytical methods are available that give the specialist an indication of how the filler is likely to behave in the rubber. The analytical test methods are found in ASTM, DIN, or ISO standards.

The following important test methods are used to characterize fillers:

The *specific surface area* is determined by nitrogen or iodine adsorption. For inactive fillers, BET values between 0 and 10 m^2/g are found; for semiactive ones, values between 10 and 90 m^2/g; and for active ones, values between 90 and 250 m^2/g. The cetyltrimethylammonium bromide (CTAB) number is used as a measure of the nonporous surface area. Occasionally it is regarded as the more exact way of describing the reinforcing effect.

The primary *particle size* can only be derived inexactly from a determination of the surface area. The smallest physically determinable filler particles, so- called primary particles, can be measured by means of electron microscopy. For active, precipitated silicas the primary particle diameter is in the range 15 – 20 nm. In the case of carbon blacks and active silicas and silicates, the primary particles collect together via the so- called primary bond to form aggregates, which in turn form agglomerates (secondary structure) through van der Waals forces. This secondary structure can be destroyed relatively easily because of the relative weakness of the van der Waals forces. This is the case when carbon blacks or silicas and silicates are mixed into rubber and dispersed. At sufficiently high shearing forces and adequate mixing time, the aggregates form the smallest filler particles present in the rubber. The mixing process is often stopped before this situation is reached. The size of the agglomerates and their distribution in the filler are usually determined by laser beam scattering.

The *adsorption of liquids* by fillers can give indications of the structure of the filler and thus its activity. In the past, mainly oil adsorption was determined, but more recently dibutyl phthalate (DBP) adsorption is preferred. Unlike carbon blacks, the activity of precipitated silicas and silicates in rubber cannot be unambiguously described and classified by using the DBP number. Silicas and silicates with differing reinforcing properties can give the same results on DBP adsorption. For carbon blacks, DBP numbers between 60 and 150 mL/100 g are found; for active precipitated silicas the values are between 150 and 250 mL/100 g.

The *tint strength* is actually a colorimetric property of carbon blacks. It is a comparison of the color strength of carbon blacks with a white pigment (zinc oxide). Since the tint strength is affected by particle size – particularly particle-size distribution – and also by the structure of the carbon black, conclusions about the activity can be drawn. The more finely divided the carbon black is, the greater is the tint strength. Lower structures (i.e., carbon blacks with a lower degree of aggregation, see → Carbon) have higher tint strengths.

The pH of the rubber compound influences vulcanization kinetics. Acidic compounding ingredients delay the onset of cure, while basic ones accelerate it. For this reason the *pH of fillers* in an aqueous dispersion is determined. In all classes of fillers (carbon blacks, natural and synthetic white fillers), both acidic and basic products are found. The gas carbon blacks are acidic because of their oxidized surface, while furnace carbon blacks are normally basic. Precipitated active silicas are neutral to weakly acidic, and synthetic silicates are basic (pH 9 – 12).

The methods of determining *pellet hardness* (usually only for pelletized carbon black types) and *sieve residue* (various possible methods) play an important role in the processing of fillers.

Pellet Hardness. The pellet hardness gives indications of possible problems with dispersion. Too high a hardness can give rise to undispersed carbon black areas in the compound and thus defects in the finished article and its possible rejection. However, pellets that are too soft are also undesirable. During transportation and storage of the filler a high content of fines is formed, which can result in dust exposure at the

place of mixing, on weighing out, or possibly in automatic dosing systems. The fines content and the pellet abrasion are therefore measures of transport stability and conveyability.

Sieve Residue. For the granular active silica types also used in the rubber industry, corresponding considerations apply. Granulate hardness has not yet been determined successfully because of technical measuring difficulties. The fines content in the *Rotation and Tapping* (*ROTAP*) *dry sieve analysis* gives a first indication of transport stability. Certain values may not be exceeded because dust problems can also arise at the workplace in the processing industry.

For powder fillers the sieve residue gives indications of possible processing problems through grit or undispersed filler particles. These cause defects such as holes or unevenness, particularly in thin-walled articles (e.g., profiles, hosing, membranes) which can also lead to rejection of the article.

Bulk and Tapped Density. To determine the space requirement of fillers the bulk and tapped densities are measured. They are important for the volumes necessary for transport and storage, and can also give an indication of dispersion time. A filler with a lower bulk or tapped density requires a longer time for incorporation into the rubber because smaller shearing forces are built up. The specific densities of fillers, as well as those of all other compounding ingredients, are needed for calculating the optimal degree of filling of internal mixers and the density of the resulting compound.

Other filler properties, which are used only occasionally for special types of rubber, are dealt with in that particular section, if important.

Fillers can be classified simply as *carbon blacks* and *white fillers*. The latter include many natural and synthetic products, such as pyrogenic and precipitated silicas, precipitated aluminum and calcium silicates, zinc and magnesium carbonates, calcium carbonate (natural and synthetic), calcium and barium sulfates, soft and hard clays, calcinated clay, silica, talc, barytes, aluminum hydrate, metal oxides, and titanium dioxide. A classification scheme is given in Figure 9.

4.2. Carbon Blacks [98 – 102]

(→ Carbon, Chap. 6)

Chemically, industrial carbon blacks consist almost exclusively of carbon. However, simply using the chemical symbol C would not describe carbon blacks adequately to the user. For characterization a number of physical and chemical properties are therefore determined, which give the specialist an indication of the activity and compounding behavior in polymers. Small proportions of oxygen, hydrogen, nitrogen, and sulfur, and traces of metals in the lower ppm range, can be determined by chemical analysis. Oxygen- containing groups, such as carbonyl, carboxyl, pyrone, phenol, quinone, lactol, and ether groups, can occasionally be determined quantitatively on the surface of carbon blacks. The content of nitrogen and sulfur depends on the raw materials used.

Carbon blacks are produced either by *incomplete combustion* of crude petroleum, natural gas, and tar oil products or by the *thermal cracking* of acetylene and other hydrocarbons. The following are produced by *incomplete combustion:* lampblack, channel black, gas black, and furnace black. Thermal, acetylene, and electric arc blacks are produced by *thermal cracking*.

The annual production capacity for carbon blacks is more than 6×10^6 t worldwide. Furnace blacks are the most important group, accounting for more than 95 %. The furnace black process was developed in the 1920s and differs from the other processes mainly in that the incomplete combustion for carbon black production takes place in a closed reactor. In this process, the small number of basic types could be diversified into a large number of different types by using alkali-metal compounds, which have an effect on the structure of carbon blacks.

In the rubber industry, furnace blacks have forced the others almost completely out of the market. In the past, *thermal blacks* and the *channel blacks* were quite important. For environmental reasons (high emissions), only limited quantities are still produced by these processes. *Acetylene blacks* are important as conducting blacks because of their particular purity. To achieve comparable conductivity of a compound, only small quantities of these types are needed compared to standard types. Generally the conductivity of a filled polymer increases

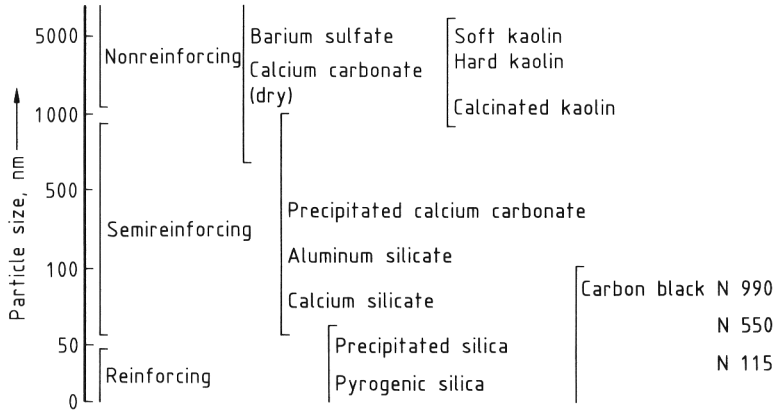

Figure 9. Classification of fillers

with increasing specific surface area and increasing structure of the carbon black.

In the past, carbon blacks were characterized for the rubber industry according to their behavior in the rubber compounds. Carbon blacks, which showed particular resistance to abrasion in compounds, were specified as *high-* or *superabrasion furnace* (HAF or SAF); others with particularly good behavior in extrusion compounds, as *fast-extrusion furnace* (FEF), for example. The ASTM specification D 1765 has now become widely used (Table 4). It consists of a letter that describes the vulcanization rate (N for normal, S for slow) and three numbers, the first being an index of the primary particle size. The physical and chemical characteristics are based on the carbon black currently used as a comparison, International Reference Black 6 (IRB 6). This carbon black was produced in a single production campaign after agreement with other producers. After this comparison carbon black has been consumed a new one will be produced, which will then have the specification IRB 7, etc.

According to the ASTM scheme, carbon blacks are characterized principally by iodine adsorption and nitrogen surface area, which are measures of the particle size and activity of the carbon black, and the DBP adsorption (structure) (→ Carbon). The differences in specific surface area and structure are clearly reflected in the technical rubber properties. Increasing iodine adsorption or nitrogen surface area is associated with increased reinforcing action [103]. *High-structure carbon blacks* are evenly distributed in the compound. They bring about in- creased compound viscosity and low die swell with smooth extrudate surfaces, high modulus, increased hardness, and improved abrasion resistance in the vulcanizate. In the 1960s the high-structure carbon blacks dominated the field of abrasion-resistant tire treads. At the beginning of the 1970s, so- called *improved* or *new-technology carbon blacks* were developed as a result of a new technology in furnace black production [104 – 107]. A characteristic of this new generation compared to standard carbon blacks is improved abrasion resistance of the tire treads at the same iodine adsorption of the carbon blacks. The difference between these and the normal carbon blacks with regard to physicochemical properties lies in the nonporous carbon black surface (CTAB adsorption).

Table 4. Extract for carbon black classification according to ASTM 1765

ASTM specification	Iodine adsorption, g/kg	DBP adsorption, mL/100 g	Nitrogen-surface area, m^2/g
N 115	160	113	145
N 220	121	114	119
N 299	108	124	108
S 315		79	88
N 326	82	72	84
N 339	90	120	96
N 375	90	114	100
N 539	43	111	41
N 550	43	121	42
N 642	36	64	37
N 650	36	122	38
N 762	27	65	28
N 765	31	115	31
N 990		43	9

Carbon blacks with *low structure* are characterized by lower heat buildup under dynamic stress (hysteresis), high tensile strength and tear resistance, good resistance to flex cracking, and low modulus.

Occupational Health. Industrial carbon blacks do not belong to the group of hazardous substances or goods. Years of experience by carbon black producers and the processing industry have shown that no health hazards or damage should be expected from them. Detailed investigations in Europe and the United States have shown that industrial carbon blacks do not cause any structural damage to the lungs. Irritation of the eyes and mucous membranes has also not been observed. Therefore, during production and processing only the general dust limits must be adhered to. These are 3.5 mg/m^3 in the United States (TLV) and 6 mg/m^3 in Germany [108].

Producers and Trade Names. The largest producers of industrial carbon blacks are Cabot (Elftex, Vulcan), Columbian Carbon (Startex, Furnex), and Degussa (Corax, Printex).

4.3. White Fillers

Silicon as a carrier of "inorganic life" occurs in nature almost exclusively in the form of crystalline solids in approx. 800 different siliceous minerals. In the evolutionary history of man, there are numerous indications of the omnipresence of silicates in our natural habitat. Siliceous minerals were either utilized by processing natural silicate deposits, e.g. clay, porcelain, glass or enamel, or obtained by means of chemical conversion (silica, silicones, ceramics). The reason for Degussa dealing with the siliceous chemicals was the company's involvement in the field of carbon black. In addition to their classic application as black pigment, carbon blacks were also increasingly used as active fillers in the rubber industry, in particular in the manufacture of automobile tires. Because the starting material for carbon black at the time (in the 1940s) – natural gas – was not available in sufficient quantities in Germany, a substitute was sought which could be prepared from indigenous starting materials. The concept of 'white carbon black' was born and research into the manufacture of siliceous fillers commenced. The idea to apply the manufacturing method for carbon black to volatile silicon compounds (pyrogenic silica, Aerosil) also originates from that time [215].

4.3.1. Precipitated Silicas

As described above, fine *silicas* can be produced by a pyrogenic reaction (see Section 4.3.2) or by precipitation. *Silicates* are manufactured by a precipitation process. It is an interesting task to precipitate silicas and silicates from aqueous media. The methods are so variable that different products customized for the respective application are attainable. Manufacturers and users of fine silicas and silicates require analytical characteristic values in order to compare the various products [216]. Owing to the great differences in chemical composition of silicas, silicates, and other natural substances such as chalk or siliceous minerals, the characterization usually commences with chemical analysis, which provides important information on the product composition, the main and particularly the secondary constituents. One significant constituent of precipitated silicas and silicates is water, which occurs in various quantities, either chemically bonded or adsorbed. The specific surface area, mean particle size, particle size distribution, pH and absorption of certain oils all play a significant role in the fine products and thus have to be determined. These somewhat "classic" methods of analysis and a number of new methods for characterizing silicas and silicates will be dealt with in greater detail in Section 4.3.1.3. The versatility of fine silicas and silicates is attributable to a variety of fundamental properties. In chemical terms, they are largely inactive and exhibit high thermostability. Targeted modification of the silica surface with organic silicone compounds (see Section 4.3.1.3) converts "water-friendly" or hydrophilic silicas into "water-repelling" or hydrophobic products. These hydrophobic properties have proven to be very useful for certain applications, e.g., in silicone rubber or for defoaming liquids. The oldest application of "white" silicas and silicates, i.e. in shoe soles and technical rubber articles, and the use of silicas in modern automobile tires will be discussed later.

4.3.1.1. Production

Precipitated silicas have only been produced since the 1940s but they have grown to become the most important group of silica products on the basis of production tonnage. The worldwide production capacity of precipitated silica in 1999 was ca. 1 100 000 t, compared to ca. 400 000 t in 1970 [217]. The European production volume for precipitated silica for the year 2000 was 285 500 t [218].

The raw materials used for the industrial production of precipitated silicas are alkali (preferably sodium) silicate solutions, and acid. Silica is precipitated through a reaction of both components. Depending on the intended application of the silica, precipitation parameters such as feed rates, stirring, shearing, temperature, pH, alkali content, etc., can be adjusted allowing for a wide range of silica products with different characteristics, such as specific surface areas ranging from approximately 25–700 m^2/g. Typically, *precipitated silicas* are synthesized under alkaline conditions at pH values >7, while *silica gels* are typically produced under acidic conditions.

Industrially produced synthetic precipitated amorphous silicas are still made today by precipitation from diluted sodium or potassium silicate solutions with the aid of mineral acids such as sulfuric or hydrochloric acid, or by using carbon dioxide. In case of the use of sodium silicate and sulfuric acid the reaction follows Equation 1

$$Na_2O \times 3.3\,SiO_2 + H_2SO_4 \rightarrow 3.3\,SiO_2 + Na_2SO_4 + H_2O \quad (1)$$

In general a module of 3.2-3.5 — the module indicates the ratio between disodium monoxide and SiO_2-content in the sodium silicate — is used in order to reduce formation of byproducts. In all cases a corresponding byproduct is formed which can be sodium sulfate, sodium chloride or sodium carbonate, depending on the acid component employed. If potassium silicate is used, the corresponding potassium salts develop. These byproducts must be removed in the downstream process.

Principally, the polymerization reaction follows the route given in Equation 2

$$\text{Monomer} \rightarrow \text{Dimer} \rightarrow \text{Cyclic molecule} \rightarrow \text{Particle} \rightarrow \text{Precipitated silica} \quad (2)$$

The exact timing of the aggregation in the process and the duration of the particle growth can be controlled by the salt concentration, e.g., by adding sodium sulfate. The underlying mechanism of the process is described in detail in [219].

During the reaction, primary particles are first formed in the reactor; later these particles react with each other, accompanied by dehydration, to form aggregates. Within the aggregates, the primary particles are linked together via siloxane (Si-O-Si) bonds. During this process, the aggregates are deposited to form larger units, or agglomerates. In these agglomerates, the aggregates are held together by hydrogen bonding or van der Waals interactions which are considerably weaker than siloxane bonds. A state of equilibrium, which is dependent on the process conditions and can be easily influenced, is reached between the aggregates and agglomerates (Fig. 10).

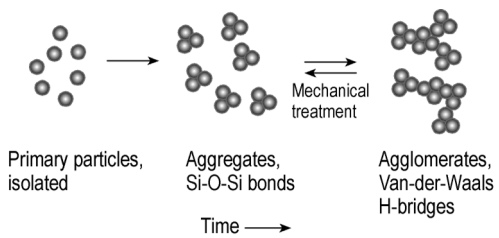

Figure 10. Particle formation during precipitation

The complete industrial production process consists of the following steps: precipitation, filtration, drying, grinding, and, in some cases, compacting and granulation (see Fig. 11):

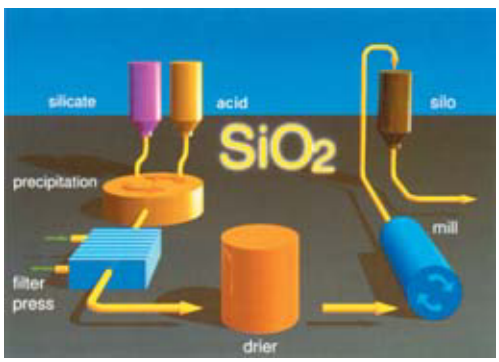

Figure 11. Industrial production process of Precipitated Silica

So far only batch precipitation processes have attained economic importance [220, 221], although continuous precipitations have also been reported [222].

The separation of the silica from the reaction mixture and the removal of the salts contained in the precipitate is carried out in filter aggregates such a rotary filters, belt filters, or chamber-, frame-, and membrane-filter presses.

Since the solid content of the filter cake products is only 15 – 25 %, about 300 to 600 kg of water must be evaporated for each 100 kg of final product. Depending on the desired properties of the end product, drying is carried out in spray, nozzle spray, plate, belt, or rotary driers. Special product properties can be achieved by spray drying the filter cakes after redispersion in water or acid [223, 224]. Figure 12 shows scanning electron micrographs of a nozzle spray-dried, unground silica and a conventionally dried, ground silica.

Various mills and, if required, classifiers [225] are used to control the desired fineness of the particles. Special degrees of fineness can be achieved by air-jet or mechanical grinding [226]. The silica is separated from the air in cyclones or filters.

Processes for the compression and granulation of precipitated silica have been developed [227, 228] to reduce the volume for transportation and for certain uses and also to decrease the formation of dust during handling and processing.

Bags, big bags, containers, or transport silos are used for packaging and shipping. Information about the different methods of packaging and shipping is given in [229]. In Germany the general dust limit of 4 mg inhalable dust per cubic meter air must be adhered to when handling precipitated silicas. Similar regulations apply in other countries.

An easy incorporation of the silica into the rubber mixture and its good dispersion is crucial because it has a considerable effect on the process costs and on the rubber product performance. By selecting appropriate process parameters, it is possible to influence the surface properties and thus the reactivity of the silicas.

At present, a variety of manufacturers supply products that can be divided into the following three groups:

1) Products which can be described as "first-generation", "standard" or "conventional" silica.
2) Products belonging to the second generation of silica. Today, such products are frequently termed "easily dispersible silica" or "semi-highly dispersible (semi-HD) silica".
3) The third and newest generation of silica comprises a product group characterized by excellent dispersion. These products are described as "highly dispersible silica (HD silica)".

With regard to the group of "highly dispersible" silicas, it can be assumed that in addition to special precipitation parameters, only so-called short-retention driers (flash or spray

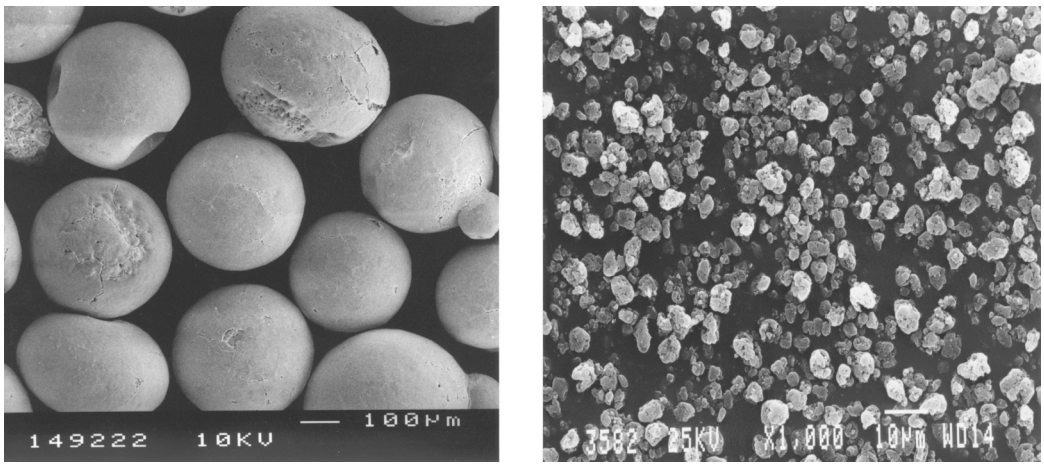

Figure 12. SEM pictures of different Precipitated Silicas

dryers) are used; while the use of long-retention driers (e.g. rotary driers) is typical for silicas of groups (1) and (2).

4.3.1.2. Physicochemical Properties

To characterize a silica means describing its (surface) structure, i.e. its morphology, as accurately as possible. This includes the ratio between the "inner" and "outer" surface, the size and shape of pores, the absorption capacity, the surface roughness, the primary particle size, the formation of aggregates from these primary particles with the development of siloxane bonds, and the combination of aggregates to form agglomerates held together by van der Waals forces. Other methods attempt to characterize the surface chemistry, i.e. they describe the number of silanol groups and their arrangement, surroundings, reactivity and the chemical composition and degradation behavior of the silica surface.

All synthetic *amorphous silicas* are white powders consisting of SiO_2 and between 2–10 wt% of physically adsorbed water. They are amorphous and no crystalline part can be detected. Depending on the production process amorphous silicas have tapped densities (DIN/ISO 787/XI) ranging from 50 g/L to 600 g/L. All amorphous silicas consist of primary particles, aggregates and agglomerates (DIN 53 206). During the production process, primary particles grow together to form chemically bonded aggregates. These aggregates then form agglomerates held together by van der Waals forces and hydrogen bonding. In powder form, only agglomerates exist. Primary particles can be identified as individual particles by means of suitable analytical methods (e.g. transition electron microscopy, TEM). Typically, these spherical primary particles have a diameter in the range of 2 – 20 nm [230]. Aggregates are assemblies of primary particles which are grown together in the form of chains or clusters. The aggregates are formed by the collision of particles, by particle growth and by the further deposition of silica onto these aggregates. In practice, the degree of aggregation defines the structure of the silica, because aggregates can not be broken into primary particles by any dispersion process. On the other hand, when dispersed in a fluid, agglomerates can be broken into aggregates. Typically, the mean particle sizes for agglomerates are between 2 and 250 µm for silicas in powder or microgranular form whereas granulated silicas have a mean particle size of up to 3000 µm.

One important method of characterization is the specific surface area, mainly measured by the Brunauer–Emmett–Teller (BET) method [231] or modernized versions of it [232]. The BET method always provides the sum of the so-called "outer, geometrical" and "inner" surface, i.e. the surface within the porous silica structure. If the value obtained for the surface using the BET method is compared to that from electron micrographs, which only indicate the outer surface, then it is at least possible to estimate the ratio of outer and inner surface. Evaluation of other sections of the BET isotherms provides additional information. The so-called "C value", calculated from the BET isotherms, gives a qualitative indication of the magnitude of the interaction between the surface of the silica and the adsorbed material and hence of the chemical reactivity of the surface [233]. If it is ensured that not only the surface of a sample is covered with nitrogen but also that the pores are filled, then the distribution of mesopores (pores between 2 and 30 nm in size) can also be determined using the Barett–Joyner–Halenda (BJH). The specific surface area of industrial precipitated silica varies widely from 30 m^2/g up to 700 m^2/g.

Another important parameter to describe precipitated silicas and silica gels is the porosity expressed by the pore size distribution. This parameter describes the sorption characteristics of the silica. However, it should be mentioned that the pore size distribution and the surface area are the results of an underlying parameter: the particle sizes of the aggregates.

According to IUPAC [234] the following classifications exist:

Micropores	pore diameter	0 – 2 nm
Mesopores	pore diameter	2 – 50 nm
Macropores	pore diameter	50 – 7500 nm

Precipitated silicas are typically macroporous materials. Here the mercury method is generally most suitable for the reliable determination of pore sizes above 30 nm. WASHBURN et al. [235] introduced the mercury in-

trusion method to measure the pore size of a porous silica.

Typically, precipitated silicas show a broad pore size distribution whereas silica gels exhibit a sharp maximum (Fig. 13). Usually, fumed silicas have only outer surface without any pores. The outer surface is easily accessible in adsorption or desorption processes.

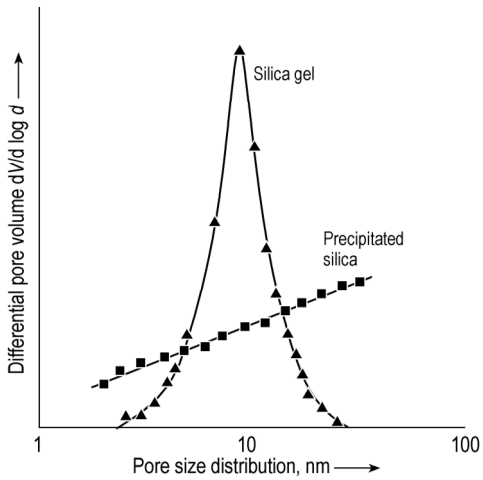

Figure 13. Pore size distribution of a precipitated silica and a silica gel

Hydrophilic precipitated silicas contain between 2 - 10 wt% of physically adsorbed water, hydrophilic fumed silicas only up to 2%. After 2 h of drying in a drying cabinet at 105°C, the physically adsorbed water is removed. The weight difference before and after drying is equivalent to the mass of physically bonded water (method according to ISO 787/2, ASTM D 280) and is called loss on drying. The amount of chemically bonded water is determined by heating a sample for 2 h at 1000°C (loss on ignition, method according to ISO 3262/11, ASTM D 1208). The weight difference before and after heating at 1000°C is equivalent to the weight of physically bonded water (= loss on drying) plus the weight of chemically bonded water.

Generally, the pH value (ISO 787-9) of precipitated silicas is approx. 7, for hydrophilic fumed silica it is around pH 4.

The determination of the sieving residue (ISO 787/18) provides an indication of the difficulty to disperse amount of the fractions in a precipitated silica.

In industrially produced precipitated silicas besides SiO_2 as the main component, traces of other metal oxides, sulfates and chlorides can be determined. The SiO_2 content is determined gravimetrically by fuming off with hydrofluoric acid. The analysis for metal oxides is then performed by means of atomic absorption spectroscopy (AAS) from the residue remaining from the fuming off process. Sulfate and chloride contents can be determined by potentiometric titration.

The *Sears number* is a measure of the number of silanol groups on the surface and therefore a measure of the specific surface of a silica [236, 237]. The Sears number is equivalent to the quantity of 0.1 N NaOH required to titrate a suspension of silica from pH 6 to pH 9. The acidic silanol groups on the silica surface react with NaOH. The Sears number gives an indication of the number of reactive centers on the surface of a silica.

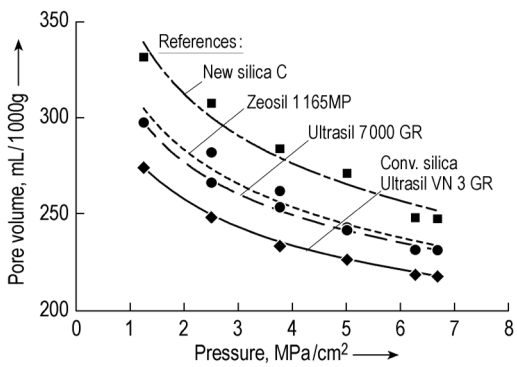

Figure 14. Decrease in pore volume of silica as pressure increases

The void volume [238], i.e. the silica structure as a function of the pressure, is considerably more reliable than the DBP (dibutyl phthalate absorption) number. In this method, the sample is subjected to a defined coarse crushing and then a specified quantity of the sample is transferred to a cylindrical glass chamber with a volume scale. The chamber is sealed by means of a moving piston. For the measurement, a constant pressure is applied to the piston until the volume of the sample in the chamber remains constant. Then the measured value is read off. Subsequently, the pressure exerted on the sample is increased again. Figure 14 shows

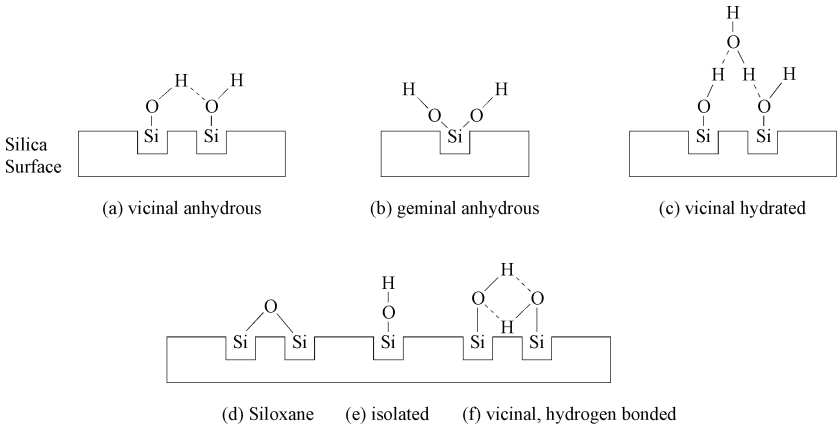

Figure 15. Different types of hydroxyl groups on silica surfaces

the decrease in structure (or the void volume) as the pressure increases. In this figure a new silica, prepared using a modified manufacturing method, is compared with the reference silicas Zeosil 1165 MP, Ultrasil 7000 GR and the conventional silica Ultrasil VN 3 GR with the same cetyltrimethylammonium bromide (CTAB) surface area (see below). It is obvious that the new products have a significantly higher void volume even at the start of the measurement. As the pressure is increased the pore volume decreases, yet always remains well above the level of the respective reference. In other words, the higher porous structure remains intact even when subjected to great stress, e.g. during mixing in a Banbury mixer. The more stable structure can be penetrated more easily by polymers and is indicative of a more advantageous mixing and processing behavior of the sample.

The CTAB method is known from the area of carbon black technology and also used in the case of silica. The method is based on the adsorption of the surface-active CTAB molecules from aqueous solutions [239]. The preferred adsorption site for these large CTAB molecules is the outer, geometrical surface, which correlates quite well with the surface area accessible for the rubber [240, 241]. Comparison with the BET surface, which is the sum of the outer and inner surface of a filler, provides an indication — with a certain margin of error — of the ratio between the inner surface and the total surface. The CTAB surface often coincides very closely with the surface determined from electron-microscopic images (TEM).

Typical physicochemical data of the different silicas are given in Table 5.

Surface Modification. Synthetic amorphous silicas (SAS) have two different functional groups on their surfaces: silanol (Si−OH) groups and siloxane (Si−O−Si) groups (Fig. 15). These two groups substantially influence the surface properties and hence the application properties. The fully hydroxylated surface of precipitated silica has ≈4.6 silanol groups per nm^2 [242 – 244] , for silica gels the value is about 5.5 [245] and for fumed silica ≈2.5 – 3.5 [246] which results in the hydrophilic character of the products.

Whereas the siloxane groups are generally chemically inert, the reactivity of the silanol groups allows chemical surface modification [247, 248]. Thus, the reaction with organosilanes [249 – 253], silicone fluids, or chlorosilanes [254 – 257] leads to hydrophobic silicas.

Hydrophobization is carried out industrially both by wet processes (e.g., addition of organochlorosilanes to the precipitate suspension) and dry processes (e.g. the reaction of fumed silica with dimethyldichlorsilane [258]). A scheme of the reaction and change in surface from hydrophilic to hydrophobic is shown in Figure 16.

In coating processes, in contrast, no chemical reaction takes place; the coating agents are physically adsorbed on the silica surface. Wax coating with wax emulsions has become industrially important [259].

Table 5. Physicochemical data of synthetic amorphous silica

			Precipitated silica	Silica gel	Pyrogenic silica
CAS-Nr.			[*112926-00-8*]	[*112926-00-8*]	[*112945-52-8*]
Specific surface area	ISO 5794-1, Annex D	m^2/g	50–800	20–1000	50–400
Average primary particle size	TEM	nm	2–20	n.a.	7–40
Mean particle size	*	µm	3–3000	0.1–5000	n.a.
Pore volume	IUPAC, App.2, Pt. 1		macroporous	micro- and mesoporous	n.a.
Loss on drying	ISO 787–2	%	3–6	ca. 3	0.5–2
Loss on ignition	ISO 3362–11	%	3–12	5–6	0.5–2.5
pH-value	ISO 787–9		6–8	4–8	3–5
DBP-number	ASTM D 2414	g/100g	50–350	n.a.	100–350
Tapped density	ISO 787–11	g/L	90–450	n.a.	50–150
SiO_2-content	**	%	98–99	>99.5	>99.9
Al_2O_3	**	%		<0.05	<0.05
TiO_2	**	%			<0.03
Fe_2O_3	**	%	<0.03		<0.003
Na_2O		%	<1	<0.1	
HCl	**	%	n.a.		<0.025
SO_3(sulfate)	**	%	<0.8		n.a.

1) various methods
2) based on ignited substance

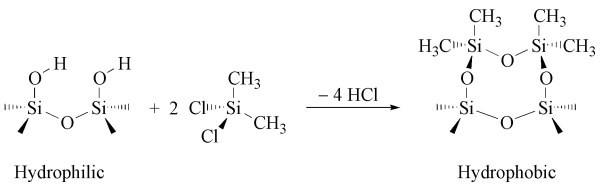

Figure 16. Reaction scheme of dimethydichlorosilane with the silica surface

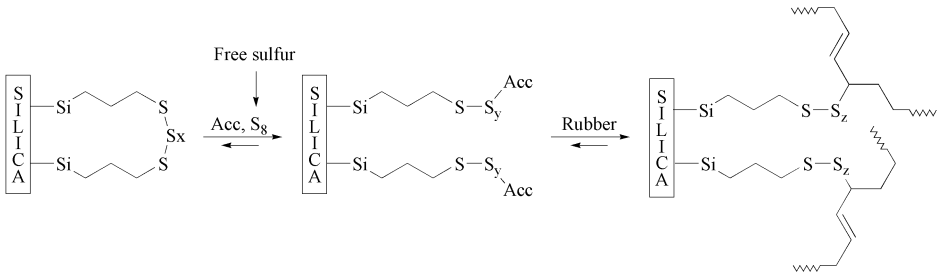

Figure 17. Simplified silica–rubber coupling reaction with Si69® = bis-(triethoxy-silyl-propyl)-tetrasulfane

Since the early 1990s precipitated silica has been used as a high-performance reinforcing fillers in rubber compounds for passenger car tires. To achieve a higher reinforcement silane coupling agents are needed. They provide a chemical link between precipitated silica and rubber polymer [260 – 262] (see Fig. 17). Most used silane for this application is bis(triethoxysilylpropyl)tetrasulfane (e.g. Si69® from Degussa).

4.3.1.3. Commercial Products [263 – 266]

In the previous section, the silicas for the rubber-processing industry were classified according to their dispersion as "conventional", "semi-HD or easily dispersible" and "HD" silicas. This classification has gained acceptance in the literature and will also be adhered to in the following overview in Table 6 (based on the manufacturer's own classification). In Table 6 the products are also classified according to their specific surfaces (in this case CTAB surface). According

to ISO 5794 precipitated silicas can also be classified from A – F according to their BET surface area, with A for silicas with more than 191 m^2/g and F for those with less than 70 m^2/g.

Table 6. Classification of rubber silicas and silicates according to their CTAB surface area and dispersion

Dispersion	CTAB surface area / (m^2/g)		
	100 ± 20	160 ± 20	200 ± 20
Conventional	Ultrasil 360 (GR) Ultrasil AS 7 Ultrasil 880 Hubersil 1613 Hubersil 1633 Hubersil 1635 Hi-Sil 315 Zeosil 125 Gr Zeolex 23 Zeolex 80	Ultrasil VN3 (GR) Hubersil 1714 Hubersil 1715 Hubersil 1743 Hubersil 1745 Hi-Sil 210 Hi-Sil 233 Hi-Sil 255 Hi-Sil 243 LD Zeosil 145 Gr Zeosil 174 G Zeolex 25	Hi-Sil 185/195 Zeosil 195 Gr Hi-Sil 170
Semi-HD	Ultrasil VN2 (GR) Zeosil 115 GR Huberpol 135 Zeosil 1135 MP	Ultrasil 3370 GR Hi-Sil 243 MG Hi-Sil EZ Zeosil 145 MP Zeosil 165 Gr	Hi-Sil 190 G Zeosil 195 MP Zeosil 215 Gr
HD	Zeosil 115 MP Zeopol 8715	Ultrasil 7000 GR Zeosil 1165 MP(S) Zeopol 8745 Zeopol 8755	Ultrasil 7005 Zeosil 1205 MP Hi-Sil 2000

4.3.1.4. Essentials of Rubber Silicas

The oldest use for fine silicas and silicates was in shoe soles. Here, the high demands with regard to wear can be met with ease by such reinforced vulcanized materials. The great advantages of light-colored reinforcements as compared with carbon black is the possibility of satisfying practical, not to mention fashion-related requirements, whether they be that shoe soles be non-marking, transparent or a particular color, or whether it be necessary to stamp cable sheaths with colored lettering.

If the fineness and therefore the high specific surface of a substance is considered to be a main contributing factor to the reinforcing effect in rubber, then fine silicas and silicates must act similarly to rubber black. This is indeed the case, but clear distinctions must be made [267]. In simple terms, the cross-linking of rubber during vulcanization proceeds in a different manner in the presence of carbon black than in the presence of silica or silicates. The reasons for this are: adsorption of cure ingredients and the "immiscibility" of hydrophilic silica with the hydrophobic rubber. Consequently, silicas are mostly used in conjunction with bifunctional organosilanes, such as Si69®. Organosilanes undergo chemical bonding to the silica surface, thereby providing the potential for cross-linking to the rubber.

4.3.2. Inactive White Fillers

Inactive white fillers are preferably used as extenders to make compounds cheaper and to make processing easier. Since these fillers are natural products that are generally untreated, apart from possibly being ground up, the content of rubber poisons is especially important for product quality. The content of rubber poison should generally not exceed 10 ppm, based on the compound. The most important fillers in terms of quantity are calcium carbonate and kaolin.

Calcium carbonate [*471-34-1*] occurs naturally as chalk, limestone, and marble. All three forms can be used as fillers after appropriate processing (grinding, breaking up). The average particle diameter of these products is 2 – 100 µm. Calcium carbonates produced chemically by precipitation are characterized by greater particle fineness, with average particle sizes of 0.1 µm. The specific surface area can be up to 100 m^2/g. Precipitated chalk is therefore a semiactive filler. The differences in color and the effects on processibility, extrudability, and vulcanization of the products used in the rubber industry depend little on the purity and the particle size.

Kaolins [*14808-70-7*], [*14808-60-7*], [*1332-58-7*] are aluminum silicate hydrates. The rubber industry differentiates between soft and hard kaolins, according to the behavior of vulcanizates. Vulcanizates filled with hard kaolins are harder (i.e., hard kaolin has a smaller particle size and is therefore more active). Hard kaolins are suitable for stable injection compounds. The content of crystalline quartz in kaolins used in rubber is generally $<0.5\,\%$. *Calcined kaolins*

have a certain importance because of the good electrically insulating properties in vulcanizates. Besides kaolins, natural calcium or magnesium silicates are used. These products occur naturally as talc [*14807-96-6*], or wollastonite.

Silicon dioxide occurs naturally in both the crystalline and the amorphous forms. Crystalline silicon dioxide is used mostly in the form of ground quartz. Kieselguhr, silica, or diatomaceous earth are among the amorphous forms. *Neuburger Kieselerde* (trade name sillitin) has a particular structure. It consists of corpuscular quartz [*14808-60-7*] and laminar kaolinit [*1318-74-7*]. With a specific surface area of ca. 10 m^2/g, it belongs to the inactive fillers, but combines the special advantages of quartz (good extrusion, little residual deformation) with those of kaolin (facile blending) because of its particular structural properties. The MAK value is 0.15 mg/m^3 because of the crystalline quartz content.

Barium sulfate is occasionally used as a filler because of its high density. It is also thought to give some protection against nuclear radiation.

The zinc oxides found on the market are used mainly as activators for sulfur vulcanization. Only a very small proportion are actually used as fillers for highly elastic, very hard vulcanizates (e.g., for spring elements). The product used here is precipitated zinc oxide with a specific surface area of 30 – 50 m^2/g.

Aluminum hydroxide is used as a flame retarder (release of water) in EVM cable mixtures. Inactive and weakly active types are available, depending on the specific surface area.

4.3.3. Coating of Fillers

A range of processes for surface treatment (coating) has been developed to improve the properties of fillers. Through this coating, an inactive filler is occasionally converted into a semiactive one, without altering the particle size. Calcium carbonate, talc, and zinc oxide can be treated with *stearic acid* to improve distribution, and silicas, silicates, and kaolins with *silanes* to reduce the compound viscosity and to improve mechanical properties of the vulcanizate [121]. Silanes can also be added during the actual mixing process, with results similar to those obtained in filler pretreatment.

4.4. Pigments [122, 123]

(see also → Pigments, Inorganic; → Pigments, Organic)

Both organic and inorganic pigments can be used for coloring rubber compounds. Organic pigments are characterized by very high brightness and particularly attractive brilliance (purity of the tint), but they are inferior to inorganic pigments in their aging properties. The advantages of inorganic pigments are their excellent fastness and their resistance to chemical effects, heat, and weathering.

The preferred pigments for rubber compounds are products that do not contain any rubber poisons; do not impair vulcanization, other processing conditions, or the mechanical properties of the vulcanizate; and are toxicologically harmless. Pigments for the *rubber industry* are available commercially as powders in the pure form or in polymer-bound form. For use in *latex*, aqueous pastes with high pigment concentration or powdered products treated with a dispersion agent to ease dispersion in water are suitable.

4.4.1. White Pigments

(see also → Pigments, Inorganic, Chap. 2)

White pigments can be counted among the inactive fillers because of their particle size. Since high levels of white pigments can lead to a significant decrease in quality of the vulcanizate, they must have good whitening and high hiding power to have a sufficient effect in small quantities.

Lithopone (a mixed pigment produced from zinc sulfide and barium sulfate) is a white pigment with relatively low whitening power. The various types of *titanium dioxide* are better in this respect and are therefore among the most important white pigments in the rubber industry and elsewhere. Both titanium dioxide modifications, rutile and anatase, are suitable as rubber pigments. The rutile types have a ca. 25 % greater whitening and hiding power

than anatase, and are thus preferred on most cases. Rutile types, particularly the inorganically coated rutile pigments, also impart better weather resistance and lightfastness to vulcanizates. In shoe soles, up to 15 phr titanium dioxide is added; in roof sheeting, up to 50 phr. Because of the high price of titanium dioxide, it is partially replaced by zinc oxide or sodium aluminum silicates in low-value compounds.

4.4.2. Inorganic Colored Pigments
(see also → Pigments, Inorganic, Chap. 3)

To achieve muted red, brown, beige, and yellow tints, *iron oxide pigments* are particularly suitable. Since iron oxides often contain rubber poisons such as copper and manganese, they must be analyzed carefully for these impurities. In special pigments for the rubber industry the content of rubber poisons is limited by the producer.

To achieve muted green and yellow-green tints, *chromium oxide* pigments are mainly used. Here too, types that are almost free of rubber poisons are available for the rubber industry.

For blue coloration, *cobalt pigments* are used, as well as *ultramarine blue* with its color nuances of greenish blue to violet. It is favorably priced but has relatively weak color strength.

Cadmium and *chrome yellow* pigments have decreased in importance for ecological reasons. They were used principally because of their bright tints and brilliance.

All desired coloring effects can be obtained by mixing colored pigments. Inorganic colored pigments are inactive fillers because of their low specific surface area. However, they are always used in small quantities.

4.4.3. Organic Colored Pigments
(see also → Pigments, Organic)

Compared with inorganic pigments the organic ones have greater color strength and more brilliant tints, but are less lightfast.

Most often *azo pigments* are used, e.g., the coupling product from diazotized 2-chloroaniline with 3-methyl-1-(4-nitrophenyl)-5-pyrazolone, which is used for orange tones. Other types of compounds are occasionally used (e.g., *alizarin* derivatives). For blue and green tints, *phthalocyanines* are also suitable. On weathering, phthalocyanines lead to crazing on the vulcanizate surface. To achieve brilliant tints with good hiding power, inorganic pigments (hiding power) are used in combination with organic ones (brilliance).

4.5. Organic Fillers

Organic fillers can be divided into naturally occurring and synthetic products. However, determining whether the organic fillers should be regarded as real fillers or should be classified under other categories is often difficult.

Synthetic products, such as phenolic and styrene resins have a strong reinforcing effect, but their advantages with respect to the processing of the compound (plasticizing effect, flow properties) are often more apparent. They are therefore dealt with in more detail in Chapter 6 (processing aids).

Naturally occurring products include cellulose, lignin, and cork. In terms of quantity used, these products play a minor role in the rubber industry. As fibers or tissues, the boundary between real fillers and construction components of composite materials is flexible. Cork is often used in sealants. The very coarse particles can be considered as inactive fillers, or particularly in the case of seals, the rubber can be considered a binding matrix for cork particles.

5. Plasticizers

5.1. General

Plasticizers are the most important chemicals after fillers, in terms of quantity, for the production of rubber compounds. Reasons for their use in rubber compounds and vulcanizates include:

1) Improvement of the processibility by targeted adjustment of compound viscosity
2) Lowering of the elastomer content by high filler and plasticizer level, thus lowering the price of the compound
3) Improvement of filler distribution
4) Changing the vulcanizate properties (e.g., improvement of low-temperature flexibility, elasticity, and extendability)

Plasticizers that bring about an improvement in flow properties and processibility are frequently known as *processing oils*. Products that are used to extend and therefore lower the price of compounds are called *extender oils.*

The plasticizer functions mentioned are generally not fulfilled by any single product.

Apart from these tasks, a plasticizer must meet the following conditions:

1) Compatibility with the particular polymer
2) No interactions with the cross-linking system
3) Little or no volatility at the vulcanization temperature
4) No harmful toxicological properties

Depending on the application of the vulcanizate, the extractability of the plasticizer, staining in light, and contact staining (e.g., with varnish or textiles) must be considered.

Since, in principle, every compounding ingredient that is liquid at the processing temperature can act as a plasticizer, a broad spectrum of substances is sold to the rubber industry as plasticizers. The boundary between plasticizers and processing aids (products that improve processing of the mixture but do not actually exhibit plasticizing action) is flexible.

The theory of plasticizer action [124] differentiates between real and pseudo, or primary and secondary plasticizers. *Primary plasticizers* dissolve or swell the polymer and thus increase the mobility of chain segments. This decreases the viscosity of the unvulcanized compound, and the vulcanizate has lower hardness, improved elasticity, and in general a lower solidification point.

Secondary plasticizers hardly dissolve the polymer, or not at all, and act as a lubricant between the molecular chains of the rubber. They improve the moldability of the compound, without markedly changing its viscosity. Examples are waxes and fatty acids.

No sharp boundary exists between primary and secondary plasticizers. The classification of plasticizers into each group depends on their chemical structure and the nature of the rubber (compatibility).

Plasticizers can be classified as natural products, mineral oils, and synthetic plasticizers. Reasonably priced mineral oils are used mainly as extenders. The choice depends primarily on compatibility and price. The relatively expensive synthetic plasticizers, which are often also PVC plasticizers, are used to improve low-temperature properties. With polar rubbers (e.g., NBR) an improvement of low-temperature properties cannot be achieved with mineral oil because of insufficient compatibility.

5.2. Mineral Oils

Mineral oils constitute by far the most important group of plasticizers in terms of quantity. They are reasonably priced and are known as processing or extender oils for many rubber compounds, depending on their use [124, 125].

Mineral oil plasticizers are divided into various subgroups characterized by the relative percentages of paraffinic, naphthenic, and aromatic carbon atoms. To differentiate among the large number of products of widely differing origin and composition that are available to the rubber industry, differences from mineral oil analysis are used, which permit adequate characterization and classification. Classification by the viscosity – gravity constant (VGC) has now become widely used. This constant is determined according to DIN 51 378 (→ Lubricants and Lubrication). The connection between VGC range and mineral oil types is given below:

Paraffinic	0.790 – 0.819
Relatively naphthenic	0.820 – 0.849
Naphthenic	0.850 – 0.899
Relatively aromatic	0.900 – 0.939
Aromatic	0.940 – 0.999
Highly aromatic	1.000 and higher

Table 7. Properties of different types of mineral oil

Property	Mineral oil		
	Paraffinic	Naphthenic	Aromatic
Density	low	→	high
Color	light	→	dark
Refractive index	low	→	high
VGC	low	→	high
Viscosity	low	→	high
Aniline point	high	→	low

Determination of the carbon distribution in mineral oils was first standardized with ASTM D 2140. The method described in DIN 51 378 follows this standard and can be used to determine carbon distribution via the VGC and the refraction intercept (r_i). The statistical distribution of carbon atoms in paraffinic (C_P), naphthenic

Table 8. General polymer – mineral oil compatibilities

Mineral oil type	Type of rubber according to DIN ISO 1629 *								
	NR	IR	SBR	BR	NBR	CR	CSM	EPDM	IIR
Paraffinic	+	+	+	+	–	–	–	+	+
Relatively naphthenic	+	+	+	+	–	–	–	+	+
Naphthenic	+	+	+	+	–	0	0	+	0
Relatively aromatic	+	+	+	+	0	+	+	0	–
Aromatic	+	+	+	+	+ *	+	+	0	–
High aromatic	+	+	+	+	+ **	+	+	0	–

* + = compatible; 0 = compatible under certain conditions; – = incompatible.
** Low dosage.

(C_N), and aromatic (C_A) bonds is obtained as a percentage.

By knowing the constitution, the general properties (see Table 7) and compatibility of mineral oils with various rubber grades can be derived (Table 8).

The quantities of mineral oil added to rubber compounds are generally between ca. 5 and 50 %; low dosages are added when used as processing aids, and higher dosages for extending. Even higher dosages (>50 %) are used in EPDM compounds. For certain applications, this rubber is cross-linked with peroxide instead of sulfur. In this case, paraffinic mineral oils are preferred in order to exclude interactions with the cross-linking agent.

The largest quantity of mineral oil plasticizers is used in the tire industry to produce compounds based on NR, IR, SBR, and BR. Although these polymers are readily compatible with all types of mineral oils, aromatic oils are used predominantly for tire production.

A considerable quantity of mineral oil is already added to synthetic rubber during production (oil-extended polymers).

Toxicology and Occupational Health. The effects of mineral oils on health are increasingly under discussion [126 – 128] because these oils may contain polycyclic aromatic hydrocarbons, some of which are known to be carcinogenic [129]. Such oils must be labeled according to EC guidelines [130] and OSHA regulations. The OSHA regulations are based on the classification of the International Agency for Research on Cancer (IARC) [131].

On the part of the mineral oil industry, these health considerations have led to the development and production of products which are less harmful [127].

5.3. Synthetic Plasticizers

Because of their much higher prices, the quantity of synthetic plasticizers used in the rubber industry is much lower than that of mineral oils. However, they are indispensable for achieving special vulcanizate properties, in particular for polar rubber grades (NBR, CR, EVM), because these grades are not or only partly compatible with mineral oil plasticizers. Synthetic plasticizers are used because they improve the low-temperature flexibility of vulcanizates. Many different products with greatly differing types of action are commercially available (e.g., ethers, thioethers, esters, and thioesters). Some of these plasticizers are used for thermoplastics, in particular PVC.

The incorporation of synthetic plasticizers can be carried out both by mill mixing and in internal mixers. Plasticizers make the incorporation and dispersion of fillers much easier and simultaneously lower compound viscosity. Plasticized compounds very often exhibit increased building tack. Processing of the compound (extrusion, calendering, injection molding) is generally improved. In the case of vulcanizates, depending on the dosage, a lowering of the Shore hardness and generally a decrease in strength with a simultaneous increase in elongation at break are achieved.

The low-temperature flexibility is considerably improved, in particular by ether and ester plasticizers. However, certain plasticizers have a marked volatility at elevated ambient temperature [132 – 134]. In particular, low molar mass ethers, which give outstanding improvement in the low-temperature flexibility of NBR, are quite volatile at the high temperature of use (in cars). Correspondingly plasticized articles lose their

low-temperature flexibility and harden as a function of time and temperature. Ether – thioether plasticizers are much more satisfactory in this regard [135].

A careful choice must also be made to achieve good or, in individual cases, improved hot-air resistance. Plasticizers with high volatility are definitely unsuitable. However, the use of high molar mass plasticizers with very low volatility (e.g., polyesters or polyadipates) is questionable. They behave favorably with regard to hot-air aging but are less effective with respect to improving low-temperature flexibility.

Polychloroprene constitutes a special case. Synthetic plasticizers improve the low-temperature flexibility of this rubber as expected, but they do not suppress the crystallization tendency [136]. In addition to other compounding measures, aromatic mineral oils should be used in this case [137].

Although synthetic plasticizers show particularly good compatibility with polar rubber grades, only a limited level can be applied depending on the type of polymer, polarity of the plasticizer, and general compounding formulation. Therefore, general dosage limits cannot be given. Dosages of more than 40 phr can be used in only a few cases. The most frequently used dosages of synthetic plasticizers are between 5 and 20 phr.

The different chemical groups of synthetic plasticizers are described below.

Ethers, such as dibenzyl ether or the formaldehyde acetal of diethylene glycol monobutyl ether (BCF), lead to very good low-temperature flexibility in NBR vulcanizates, but have limited use because of their volatility.

Ether – thioethers (e.g., condensation products based on thiodiglycol and alcohols) are very effective plasticizers for improving the low-temperature flexibility of NBR and CR vulcanizates. They are considerably less volatile than monomeric ethers and, in particular, improve the hot-air aging resistance of, for example, NBR vulcanizates [135].

Esters are the most important synthetic plasticizers in terms of quantity. Examples are phthalates [e.g., 2-ethylhexyl phthalate (DOP)], adipates [e.g., dioctyl adipate (DOA)], sebacates [e.g., dioctyl sebacate (DOS)], phosphates [e.g., tricresyl phosphate (TCF)], and trimellitates.

While *DOP* has proved to be a reasonably priced plasticizer with average effectiveness for improving the low-temperature flexibility of NBR and CR vulcanizates, *DOS* has an outstanding effect on CR vulcanizates. Phosphate esters are used to improve flame resistance, and *trimellitates* have been shown to be particularly suitable for HNBR.

Thioesters also have a certain importance, in particular butyl esters of methylenebisthioglycolic acid, which give NBR vulcanizates very good low-temperature flexibility and high elasticity. Also *phenol alkylsulfonates*, which are principally known as PVC plasticizers, are used for NBR and CR because of their low price. Their ability to improve low-temperature flexibility lies between that of polyesters and DOP.

Polyester Plasticizers. Besides monomeric esters, polyester plasticizers (i.e., esters based on adipic and phthalic acids, and propane- and butanediols) are also used. They are relatively nonvolatile and therefore used for high-temperature vulcanization (e.g., UHF processes) and for articles that are exposed to high ambient temperatures. The disadvantage of polyesters is a comparatively small improvement in the low-temperature flexibility of vulcanizates.

Chlorinated paraffins with a chlorine content up to 70 % are used in dosages up to ca. 20 % to decrease the flammability of rubber articles. They are liquid-to-viscous substances and are used mostly in combination with antimony trioxide.

The importance of chloroparaffins for the rubber industry is relatively slight and will decrease even further. In some areas (e.g., the cable industry), aluminum oxide hydrate, magnesium hydroxide, and zinc borate have been established as alternatives to these highly chlorinated compounds.

Outlook. Various possible solutions to the problem of the extraction of plasticizers by fats or oils from finished products (e.g., bellows for universal joints) have been discussed. Polymeric plasticizers (e.g., polyesters) are better than monomeric ones in this respect, but

their improvement of low-temperature flexibility is insufficient. The use of reactive plasticizers has been investigated [136]. These compounds should become bonded to the hydrocarbon chain of the polymer during vulcanization. However, the fixing of these reactive plasticizers (e.g., ethylene glycol dimethacrylate or compounds containing vinyl groups) to the polymer backbone occurs only in unsatisfactory yield or has a considerable effect on the actual cross-linking of the polymer. A practicable solution has not been found.

5.4. Natural Products

A range of natural products are used because of local availability or to achieve particular effects. However, they are of little importance in terms of quantity. These products are oils, waxes, and poorly defined high molar mass products of animal and paraffinic origin, such as lanolin, palm oil, soybean oil, linseed oil, animal gelatin, pine tar, and rosin.

In the usual relatively low dosages, these products act mainly as lubricants (in the case of fatty acids), rather than plasticizers, or are used to improve building tack (rosin).

6. Processing Additives

6.1. Behavior and Classification of Processing Additives

Processing additives are defined as materials used at relatively low dosage levels, that improve processing characteristics of rubbers without significantly affecting their physical properties [204]. The action of processing aids can be physical and/or chemical, depending on their chemical structure, the processing conditions, and the rubber system that they are used in.

Rubber processing is a multiple-step operation, depending on the polymer composition, type and level of filler, the curing system, and shaping and properties of the finished rubber article. A wide range of viscosities, polarities, temperatures, and shear rates has to be handled, often creating specific problems. Inhomogeneous polymer blends, poor filler dispersion, stickiness on machinery, low building tack, bloom, surface failures, and mold fouling are quite common.

Processing additives are often classified by their main function, but some of them are truly multifunctional.

Following the production steps of a rubber compound, the main problems to be solved are:

Viscosity reduction of natural rubber is achieved by *peptizers*, which cleave the long rubber chains and thus enable filler incorporation and blending with synthetic polymers.
Polymer blending is improved by *homogenizers* which control the domain size of the dispersed phase. Some homogenizers also act as filler dispersants and tackifiers.
Dispersing agents improve the incorporation of fillers and reduce the viscosity of the compounds.
Lubricants or *flow improvers* show their full performance mainly in high shear processes such as extrusion and injection molding. They improve productivity and surface quality of the finished article.
Tackifiers enhance the building tack of rubber parts, e.g., in tires or other rubber goods assembled from various compounds.
Some special *resins* are used to increase the hardness of rubber compounds while maintaining good processibility.
Release agents facilitate the demolding of the cured rubber parts. They also improve the filling of the cavity and reduce mold fouling (deposit on the molds).

Table 9 gives an overview on the classification of processing additives according to their effects.

6.2. Peptizers

Although the production of natural rubber with controlled viscosity (CV-grades) is gradually growing, the majority of NR needs viscosity reduction prior to blending with other polymers and/or filler incorporation. This is achieved by chemical peptizers, highly efficient combinations of organic iron complexes and an aromatic sulfur compound.

Table 9. Types and chemical composition of processing additives

Type of processing additive	Chemical composition
Peptizers	2,2′-dibenzamido-diphenyldisulfide
	zinc soaps
Homogenizers	resin blends
Dispersing agents	fatty acid esters
	metallic soaps
	fatty acids
	fatty alcohols
Lubricants	metallic soaps
	fatty acid amides
	fatty acids esters
Tackifiers	phenolic resins
	hydrocarbon resins
	coumarone resins
Reinforcing agents	phenolic resins
	polymer resins
Release agents	cross-linkable resins
	polyesters
	polyethers
	silicone oil emulsions

For many years the preferred peptizer was pentachlorothiophenol (PCTP) or its zinc salt. Meanwhile PCTP is banned in most parts of the world because it forms several teratogenic decomposition products. PTCP is replaced by 2,2′-dibenzamidodiphenyldisulfide (DBD) which is less toxic and reacts similarly.

The viscosity reduction of NR is achieved by two different mechanisms. When mixing on a roll mill or in an internal mixer, mechanical shear force is applied to polymer chains. When a chain is ruptured, free radicals are formed at both chain ends. Those radicals can recombine or react with neighboring NR chains, resulting in long chain branching. The radicals can also react with atmospheric oxygen, which stabilizes the radicals intermediately. The overall effect of this mechanical degradation of the molecular chains, so- called mastication, is low.

DBD reacts with free radicals to form stable compounds that do not react with other polymer chains. Thus the molar mass of the polymer is reduced efficiently, because mainly long chains are attacked by mechanical forces when mixing at low temperatures.

On *open mills* maximum temperatures of 80 – 90 °C are achieved, thus mechanical breakdown and stabilization by DBD is the main mechanism. Beyond 100 °C the viscosity of the NR is reduced so much that mechanical rupture of the chains is minimized.

Most of the NR in industrialized countries is processed in internal mixers. During the mixing cycle much higher temperatures are generated. From 120 °C and above oxidative chain scission is the dominant process, catalyzed by organic iron complexes, so-called boosters. This process is quite efficient, but the purely random attack on both long and short chains results in a strong increase of very short-chain material [205] which tends to stick on machinery and give unfavorable dynamic properties.

Figure 18 shows the molar mass distribution curves (measured by gel permeation chromatography) of untreated NR, the effect of mechanical mastication, and the action of two chemical peptizers. Mechanical mastication reduces the highest molar mass material, and some shorter chains are obtained. Both chemical peptizers act much more aggressively, the maximum peak at a molar mass of about 100 000 is sharply reduced, and much more short chain material is observed. In the case of untreated NR one should keep in mind that the highest molecular material, the so-called gel, is not even detected by gel chromatography. The gel content of the raw NR was 23 %, dropping to 2–5 % when mechanically treated with our without peptizers [205].

Iron complexes in use are mainly (1) iron phthalocyanine (Fe-Pc) and (2) iron hemiporphyrazine (Fe-Hemi). Their activities are comparable. The dosage of iron complexes in the peptizers is rather low, typically 0.1 – 1 wt%. Peptizers are used in dosages of 0.1 – 1 %.

The peptizers described here were originally designed to be used in a separate mastication step to utilize their full activity. For productivity reasons the mastication is now carried out as part of master batch mixing, namely in the tire industry. NR and the peptizer are blended in internal mixers for ca. 30 s, followed by addition of other polymers and fillers. After such a short time the temperature is still rather low. Thus, the iron complex does not work properly but is deactivated immediately by reaction with the carbon black. For this mixing procedure new peptizer concepts have been developed. The radical scavenging action of DBD is combined with the lubricating effect of metallic soaps. This material also acts as a flow improver and reversion inhibitor and thus is a truly multifunctional additive [206].

Producers and Trade Names. Bayer, Germany (Renacit), Kawaguchi Chemical Industry,

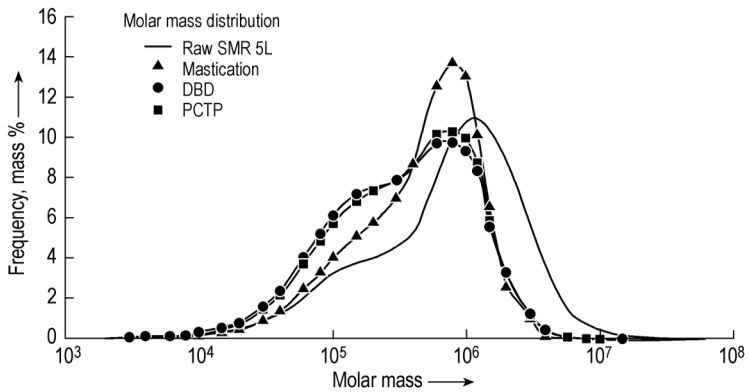

Figure 18. GPC curves of NR compounds with various peptizers

Japan (Peptar), Ouchi Shinko Chemical Industrial, Japan (Noctizer), Schill + Seilacher "Struktol" AG, Germany (Struktol), T. Swan & Co, Ltd., Great Britain (Pepton).

6.3. Dispersing Agents and Lubricants

This group contains the most important processing additives which are widely used to enhance the productivity of rubber mixing and shaping. The relation between chemical structure and main effects is described in Table 9, but most additives have more than one effect. So a clear discrimination is not always possible. Overviews are given in [204, 207, 208].

Dispersing agents and lubricants are mainly based on fatty acid derivatives such as esters, amides, and soaps. Their action depends on their polarity relative to the matrix that they are used in.

Fatty acid esters facilitate "wetting" of the filler surface, thus the filler incorporation is accelerated. The esters are low-melting materials suitable for the use on roll mills and in internal mixers. Stearates of pentaerythritol are commonly used. Because they are fully saturated compounds, they are thermally stable and have little influence on peroxide cure.

Combinations of *fatty acid amides* and *metallic soaps* are also popular. In most polymer systems they result in a good flow improvement. Due to their limited solubility in many polymers they can also act as internal release agents for molding.

Zinc soaps are the classical flow improvers, their effect is most pronounced in high-shear processes such as extrusion and injection molding. The soaps employed normally contain a number of double bonds in their carbon backbone to avoid a too high melting point which could result in dispersing problems. On the other hand the unsaturation limits their use in peroxide cured systems. Zinc soaps of higher molecular fatty acids (<C12) are able to form spherical or lamellar micelles. This behavior is believed to be the reason for their lower apparent viscosity under shear. Layers of soap micelles located between molecule chains can slide easily on each other, because the van der Waals forces between layers of micelles are weaker than the friction between rubber chains [204].

Since 2001 zinc-containing rubber chemicals are under scrutiny, especially in the EU, because of potential environmental problems (aquatic animal toxicity). Meanwhile first alternatives to zinc soaps with equal or better performance have been introduced [209]. Most efficient is a combination of alkali soaps and fatty acid esters. Due to their activating effect on sulfur cure they even open the opportunity to reduce the level of zinc oxide to a minimum.

For high tech applications which require low dosages and a high thermal stability *organosilicones* are available [204]. These condensation products of fatty acid derivatives and silicones provide excellent lubrication and good compatibility with rubber. Dosages as low as 0.5–1 phr can be sufficient, while conventional dispersing agents and lubricants require 2–5 phr, depending on the polymer system and especially the filler level.

Producers and Trade Names. D.O.G. Deutsche Ölgesellschaft, Germany (Deoflow), Kettlitz Chemie, Germany (Dispergator), Rhein Chemie, Germany (Aktiplast), Schill+Seilacher, Germany (Struktol).

6.4. Homogenizers

The most widely used homogenizers, sometimes also called *compatibilizers*, are resin blends. Homogenizers are designed to mix rubbers with different chemical composition, viscosity, and polarity. Homogenizers act by influencing the domain size of the dispersed polymer particles in the matrix of the continuous phase [208]. Thus, some ultimate properties such as tensile strength and elongation at break are improved, but also modulus values and abrasion resistance are increased, because homogenizing resins also enhance the filler dispersion. Often they also improve the tack of an uncured compound.

Homogenizing resin blends, once softened or melted, readily disperse in most elastomers, but they act at the interface of the polymer particles. Due to their adhesive properties they help to compact the batches. Thus higher shear forces can be applied at relatively low temperatures when the viscosities are still high. Normally mixing times are shortened but mixing energies are not necessarily reduced. In homopolymers homogenizers show the well-known and widely used effect of enhanced cured adhesion between bromobutyl innerliners (containing a resinous homogenizer in 3 – 8 phr) and the carcass compound of tubeless radial tires.

The most efficient homogenizers are rather high-molecular materials. As they are dark colored, they cannot be used in light colored compounds, but good homogenizers are available for this purpose, too.

Producers and Trade Names. Flow Polymer, USA (Promix), Rhein Chemie, Germany (Rhenosin), Schill + Seilacher, Germany (Struktol).

6.5. Tackifiers

In rubber technology tackiness means the ability of two uncured rubber compounds to stick together under moderate pressure. Moderate-pressure processes are used in the assembling of tires and tire retreading, of air springs, hoses, multicolored shoe soles, belts and profiles, to name a few. Normally building processes are carried out at ambient temperature, but sometimes also tack at elevated temperatures is required, e.g., in extrusion of reinforced hoses.

Natural rubber normally has sufficient tack because of its content of low-molecular materials, such as fatty acids, resins, and proteins. Synthetic rubbers, such as SBR or BR,however, often show a lack of tackiness.

Depending on the compound and the production conditions, short- or long-term tack might be required. On the other hand, tackifiers should not increase the stickiness to rubber machinery or give a sticky surface of the vulcanized article. Also the influence on the cure characteristics of the compound should be negligible.

The most popular resins used as tackifiers are hydrocarbon resins (C_5 or C_9 based), *p*-substituted phenolic novolaks and a specialty resin made from *p-tert*-butylphenol and acetylene which gives good long term tack [210, 211]. Hydrocarbon resins are relatively inexpensive and give good short-term tack, up to 3–4 days. After this period the tack drops sharply. Phenolic resins have a higher price but can be applied at lower dosages and have a better long-term tack. Their initial tack is similar to that of hydrocarbon resins. If extreme long-term tack is required, like in classical retreading, the butylphenol-acetylene resin is the material of choice. Good tack levels for up to four weeks can be achieved.

Producers and Trade Names. BASF, Germany (Koresin), CECA, France (Ribetak), Exxon, France (Escorez), Kolon, Korea (KC, KNR), Schenectady, USA (SMD), Schill + Seilacher, Germany (Struktol), Sumimoto Bakelite, Japan (Tackirol).

6.6. Release Agents

Release agents are important auxiliaries in the production of rubber parts. Because of the high tackiness of NR and many synthetic rubbers and

the uncured compounds made thereof and because of the tendency of hot cured rubber to adhere to the molding equipment, an economical production of rubber parts is only possible when using release agents, although in principle it is possible to design production processes without use of them.

6.6.1. Batch-Off Release Agents

Rubber compounds are typically mixed in a batch process with at least two (masterbatch and final mix) steps. After each step the compound is given a form (sheets, granules, feeding strips) and cooled, typically in a watery cooling bath. To prevent the formed compound from sticking together to one big lump batch-off release agents are applied, normally simply by addition to the water of the cooling bath. Apart from their release properties batch-off release agents must in some applications also provide anti-slip properties to prevent piles of sheeted rubber compound from slipping during handling and transport.

Blends of metallic soaps of fatty acids, inorganic fillers (silicates, mica), dispersing agents and surface active substances are commonly used in this application.

Producers and Trade Names. Hans W. Barbe, Germany (Promol), Rheinchemie, Germany (Rhenodiv), Kettlitz, Germany, Schill + Seilacher, Germany (Struktol).

6.6.2. Mold Release Agents

Mold release agents are necessary in the rubber curing process to prevent the cured article from sticking to the mold. Rubber, like many other polymers, has quite a low surface energy compared to the metal of a typical mold, leading to quite strong adhesive forces between the rubber and the mold, especially at elevated temperatures. Additionally, many rubbers have low tear strength at high temperatures. Therefore, on mold opening so-called 'hot tear' can occur.

Sacrificial Mold Release Agents. These are simple release agents that are applied to the mold before every curing cycle. In the past all kinds of *oils* and *waxes* and solutions and emulsions made from these have been used for this purpose and are still in use today for some applications. The drawbacks of these materials are insufficient release, uptake by the rubber compound, thus changing its physical properties and a severe mold fouling.

Another class of materials are the *silicone oils* or *silicone oil emulsions* (sometimes also called *silicone fluids*). These silicone oils exhibit superior release properties at very low dosages but unfortunately also show severe drawbacks: If the silicone oil is incorporated into the compound this leads to flow faults and poor knitting. The silicone oils transfer to the rubber part and prevent effective down-line processing like printing and adhesion. Silicone oil emulsions cause severe mold fouling [212].

The automotive industry has widely banned the use of silicone oils as release agents for automotive rubber parts because the silicone oils that are transferred from the mold to the rubber article later migrate to the car body and this creates massive problems in repainting operations.

Producers and trade names for silicone oil emulsions: Bayer, Germany (Levaform), Wacker, Germany, Dow Corning, USA, Schill + Seilacher, Germany (Struksilon).

Semipermanent Mold Release Agents. State of the art in mold release agents are the semipermanent systems, available both as solvent- or water-based systems. These agents consist of mixtures of reactive prepolymers that are applied to hot molds and there cross-link to form a polymeric coating. This coating levels the surface of the, on a microscopic scale, rough metal, thereby reducing the effective surface area, and also reduces the surface energy of the mold. These two effects synergistically lower the adhesion of the cured rubber and give good release properties [213].

The semipermanent coatings withstand multiple molding cycles, typically 30 to 50, but even up to a 100 cycles are possible, greatly depending on the individual rubber compound and the processing parameters.

Semipermanent release agents eliminate the drawbacks of the sacrificial mold release agents and additionally greatly reduce the buildup of mold fouling, giving considerable savings in cleaning costs and thereby greatly enhancing the profitability of molding operations [214].

Producers and Trade Names. Chemtrend (Monocoat), Franklynn, Henkel Loctite, Germany (Aqualine), Schill + Seilacher, Germany (Permalease).

Other Mold Coatings. Many different mold coatings (like chromium plating, PTFE coating, ceramic coatings, etc.) have in the past been tested to eliminate the need for release agents but only with very limited success. Either the release properties are not good enough (chromium plating) or the coatings do not last long enough to be cost effective (PTFE and ceramic coatings).

6.6.3. Release Agents for Special Applications

Apart from the very widely used batch-off and mold release agents there are a number of applications in the rubber industry that have triggered development of special release agents for, in part, very specific requirements.

6.6.3.1. Mandrel Release Agents

Rubber hoses are normally cured in autoclaves on so-called mandrels, typically metal or plastic parts that fill the inside of the hose and have to be removed after vulcanization. The hoses can be divided into long hoses which are straight and are normally extruded directly onto the mandrel, and shaped hoses which are extruded and reinforced without mandrel, cut to lengths, and then pushed onto a shaped mandrel.

Long-hose mandrel release agents consist of blends of surfactants and lubricating materials, sometimes silicone oil emulsions are used.

Producers and trade names : Rheinchemie, Germany (Rhenodiv), Degussa, Germany (Getren).

Mandrel release agents for shaped hoses need to have, in addition to the releasing power, a very high lubricity to reduce the friction forces when the uncured hose is pushed onto the mandrel and especially when removing the cured hose from the shaped mandrel.

Materials mainly used are water-soluble polyesters and blends of polyglycols and surfactants. The polyesters have the added benefit of being readily biodegradable.

Producers and trade names : Degussa, Germany (Getren), Schill + Seilacher, Germany (Struktol MR), Rheinchemie, Germany (Rhenodiv), Marbo, Italy.

6.6.3.2. Bladder Release Agents

Especially tires but also other hollow rubber goods are produced with bladders of butyl rubber that are inserted into the hollow and inflated good to provide the necessary pressure for the vulcanization process. To prevent these bladders from sticking to the inside of the rubber part they are coated with bladder release agents which last for a number of curing cycles. Special blends of reactive silicone resins and lubricants are used in this application.

Producers and Trade Names. Chemtrend, USA, Rhodia, France (Lyndcoat).

6.6.3.3. Inside Tire Paint

Other producers rely on inside tire paints for the release of the bladder from the tire inner liner. These paints are sprayed onto the inside of every tire before mounting it to the press. The inside tire paints consist of blends of surfactants, lubricating agents, inorganic fillers (silicates, mica, carbon blacks) and dispersing agents, typically nowadays based on water as a solvent.

Producers and Trade Names. Chemtrend, USA, (Monocoat), Rheinchemie, Germany, (Rhenodiv), Schill + Seilacher, Germany (Struktol LR).

6.6.3.4. Outside Tire Paints

As with the inside tire paints some producers rely on outside tire paints rather than on mold release agents for their tire molds. These outside paints are also applied to every tire and are of similar makeup as the inside paints, of course this time optimized for release at the tire-mold interface. Producers are the same as for inside tire paints.

6.7. Other Processing Additives

A variety of resinous materials have been in use since the beginning of rubber technology. Natural resins such as *gum rosin* lose importance, mainly because of inconsistent quality and difficult handling.

High styrene-butadiene resins are used to achieve high hardness combined with good processing ability, e.g., in the bead/apex area of radial tires. Styrene levels of 85 % in the copolymer are quite common, while polystyrene itself is incompatible and too high-melting to be incorporated into a rubber matrix.

Factice is a traditional processing additive, made from rapeseed oil or soybean oil by heating with sulfur. It was used much in combination with CR in weather strips, but in this application CR has been replaced by EPDM (without factice). A small amount of *white factice*, made from rapeseed oil cross-linked with disulfur dichloride is used as a softener in erasers. Most traditional factice producers have stopped production.

Sometimes also *mineral oils* are summarized under processing additives, but they only have a softening effect and require relatively high dosages.

Pine tar has also been used for many years, in some applications it improves adhesion but has a strong smell.

Asphalt and *bitumen* (→ Asphalt and Bitumen) are cheap additives for dark rubber compounds. Often an oxidized material called *blown bitumen* is in use, also called *mineral rubber*.

7. Rubber Adhesion

The adhesion of rubber to inelastic materials covers two completely different aspects; the adhesion of rubber to steel cords or textile reinforcing materials, and its adhesion to flat metal surfaces. They differ so sharply in application, technology, and chemistry that they are dealt with separately.

7.1. Adhesion of Rubber to Reinforcing Materials

Rubber articles, such as tires, conveyor belts, driving belts, and hosing, have a common feature; during use they are subjected to periodic extension forces of differing strength, which severely limit the lifetime of the elastic materials. For this reason, these articles are built up as a composite of rubber and reinforcing materials. The elastomer matrix is characterized by high elasticity and low tenacity, while the solid matrix has high tenacity and dimensional stability, but low extendability. The strength of the bonds between the two different materials is decisive for the technical usability of the articles mentioned. Different systems are used to improve the bonding, which depend on the nature of the materials employed.

Rubber Matrix. Sulfur- cross-linked rubbers generally show a considerably higher bonding tendency than polymers with saturated main chains. Within the group of sulfur- cross-linkable rubbers, polar rubbers such as CR and NBR are more readily bonded to the reinforcing materials than nonpolar rubbers such as NR, BR, SBR, and EPDM.

Solid Matrix.

Organic reinforcing materials also differ in their bonding tendency. *Cotton* exhibits a very high tendency to bond to rubber. However, cotton fibers have rarely been used except in the Third World since the 1960s. Until the 1940s, no other reinforcing material was available, but with the advent of rayon, nylon, and polyester, cotton rapidly lost its importance as a technical yarn for rubber applications because of its structure. Being a staple fiber, cotton is clearly weaker than the *man-made fibers* rayon, nylon, and polyester, which are processed as monofilaments in the industrial fiber sector. All commercially important composite materials are built up using these carriers. If particular properties with regard to shrinkage, toughness, modulus, and resistance to chemicals are required, special fibers such as aramid fiber or spun polyethylene are used. Of the man-made fibers, rayon exhibits the greatest adhesion tendency toward rubber, whereas nylon and polyester bond much less strongly because of their chemical structure and morphology.

Inorganic Fibers. Steel cord plays a major role among the inorganic fibers. To improve corrosion behavior and adhesion to rubber, steel cord is usually processed with a brass- or zinc-coated surface. Brass- coated steel cord was incorporated into tire technology through radial tires, which were introduced in the 1960s by Michelin and represent the current state of the art. *Glass fiber* is used in speciality areas such as timing belts.

Various systems can be used to improve *adhesion* between reinforcing carriers and rubber, particularly after aging. These are described below.

RFL-Dip [176 – 179]. Adhesion promoters based on resorcinol – formaldehyde resins in combination with sulfur- cross-linkable latices have been known for impregnating textiles since 1935 [RFL-dip = **r**esorcinol (producers are, e.g., Koppers, Sumitomo Chemical), **f**ormaldehyde, **l**atex (producers are, e.g., BASF, Bayer, Du Pont, Enichem, Hoechst)]. RFL-dip is an aqueous dispersion of resorcinol, formaldehyde solution, and rubber latex. Its solid content depends on the absorption capacity of the textile to be impregnated, and is between 10 and 30 %. The absorption capacity of the textile depends directly on the ratio of amorphous to crystalline parts in the fiber. The higher the crystalline portion, the higher should the solid material content of the RFL-dip be. The textile is passed through this dispersion and predried. In a subsequent step, resorcinol is condensed with formaldehyde at 160 – 220 °C to form a resin. This results in strong mechanical and chemical interactions between the resin and the textile. If the reinforcing material thus impregnated is combined with an unvulcanized rubber compound, sulfur and accelerators from the rubber compound diffuse into the dried latex particles. These are in intimate contact with the resorcinol – formaldehyde resin and are cross-linked with the rubber during subsequent vulcanization. Heat-stable and flexible bonds are built up among the solid carrier, the resin, the latex, and the rubber by this process.

The nature of the latex used depends strongly on the nature of the rubber. For NBR, the best adhesion is achieved by using NBR latex. For NR, SBR, and CR, the use of VP-latex, a terpolymer of styrene (25 – 27 %), butadiene (60 – 63 %), and vinylpyridine (10 – 15 %) is recommended.

Rayon and nylon can be impregnated in this way. *Polyester*, *aramid*, and *glass fiber* must be pretreated first since, because of different reasons, such as high steric hindrance or a high degree of crystallinity, chemical interaction with the resin is not possible or occurs to an insufficient extent.

For *glass fibers*, the pretreatment is carried out by the glass fiber producer (sizing). These pretreated glass fibers can then be treated with an RFL-dip.

To promote adhesion of *polyesters* and *aramid*, a two-bath impregnating process was developed. A preliminary impregnating step is carried out before the RFL-dip, in which a blocked isocyanate is applied in combination with a polyepoxide.

Isocyanates. Polyfunctional isocyanates were used frequently in the rubber industry in the 1940s and 1950s. They have been widely replaced by RFL-dip. The reason for this is the restriction of handling organic solvents (in which the polyfunctional isocyanates are dissolved) and the much lower storage stability of the isocyanate dip prior to treatment with the rubber compound compared to RFL and treated textiles.

Isocyanate impregnating baths are still used occasionally, principally for particular adhesion requirements, since isocyanate-impregnated textiles show very high adhesion to rubber.

The RFS System [180, 181]. As an alternative or supplement to the impregnation of textiles, so- called direct adhesive systems are used. The chemicals required for promoting adhesion are added to the rubber compound during mixing.

The RFS system consists of the components **r**esorcinol, **f**ormaldehyde donors such as hexamethylenetetramine or hexamethylolmelamine (producers of hexamethylenetetramine are, for example, Degussa, Harwick, Housmex, MLPC; producers of hexamethylolmelamine are, for example, ACC, Bayer, Ceca, Koppers, Monsanto, Rousselot), and an active **s**ilica reinforcing filler. The resorcinol – formaldehyde resin, which promotes adhesion, is formed during vulcanization. The silica acts as a catalyst for the resin formation reaction and determines its rate.

In principle, various resin systems can be used. The state of the art involves using resorcinol together with hexamethylenetetramine or etherified hexamethylolmelamines. The importance of the former is rapidly decreasing because of the strongly corrosive action of the decomposition product (ammonia) on polyester and brass- coated steel cord.

Because of the strong tendency of resorcinol to sublime, substitutes for this compound are under investigation. Triazine derivatives or other polar phenols have been described as alternative resin-forming compounds in the patent literature but have not yet been widely used [182].

The RFS system can be used universally, both on textile carriers and on steel cords. It can be employed for the adhesion of rubber to untreated textiles and those impregnated with RFL. The best results are obtained by using RFS- containing rubber compounds with RFL-impregnated textiles.

Cobalt Adhesion Promoters. With the introduction of brass- coated steel cord into tire production by Michelin in the 1960s, organocobalt compounds began to be used as adhesion promoters. They function by stabilizing Cu_xS crystallites, which form a layer between the rubber and the brass. The latter is essential for good adhesion. Cobalt adhesion promoters are used mainly in the tire industry, but also in the conveyor belt sector. Examples are cobalt naphthenate, stearate, decanoate, and abietate.

Producers of cobalt adhesion promoters include Borchers, Manchem, Shepherd Chemical, and Vasset [183].

7.2. Rubber – Metal Bond [184 – 188]

The quality of articles made of rubber – metal composites is determined mainly by the quality of the bond between the two materials. Two processes can, in principle, be considered for the formation fo these articles.

The *adhesion* of vulcanized rubber parts to metal surfaces can be achieved by using polyurethane or polyisocyanate adhesives. This process is employed in linings and coatings. The disadvantages of the system lie in the low thermal, mechanical, and chemical resistance of the bonds.

The *bonding* of unvulcanized rubber compounds to metal surfaces can occur during molding and vulcanization. The bond between the elastomer and the metal is formed during vulcanization at 90 – 180 °C under the usual molding pressures.

The bond is produced by a binder system consisting of a primer and an adhesive. Factors relevant to quality are

1) Surface treatment of the substrate
2) Composition of the primer
3) Composition of the binder

If these three aspects are given proper consideration, the bond strength depends very little on the nature of the elastomer and the substrate to which it is to be bonded.

Surface treatment is naturally very important. The surface must first be cleaned and degreased by using aqueous or solvent- containing degreasing agents. It must then be extended and activated by mechanical or chemical processes (phosphating, see also → Metals, Surface Treatment, Chap. 3). Depending on the nature of the metal, a primer does not need to be used in the subsequent step with appropriate pretreatment. *Primers* are dispersions of chlororubbers in aromatics or ketones containing phenolic resins as the cross-linking agents and titanium dioxide or iron oxide as the pigment. Application to the cleaned and activated metal surface can be carried out by painting, spraying, dipping, or with rollers. The actual *binder* is then applied. Aromatics or ketones and, more recently, aqueous systems are used as solvent for the binder. The binders are complex systems composed of a polymer, resin, and reinforcing filler. Halogenated rubbers with high halogen content provide the best polymer base because of their high polarity. To achieve the necessary hardness, they are mixed with phenolic or epoxy resins. The reinforcement necessary to transfer strength properties is achieved by adding carbon blacks or silicas. Since halogenated polymers eliminate halogen in the form of hydrogen halide at vulcanization and processing temperatures, metal oxides, such as zinc or magnesium oxide, are added to absorb them. A complex cross-linking system must cross-link the polymer chains of the binder with those of the rubber compound and those of the primer, besides cross-linking the polymer chains in the binder. After the binder has dried,

it can be coated with the rubber compound and vulcanized in a mold under pressure.

A range of articles can be produced by using this process. These include shock absorbers, and articles used in sealing, transport, power transmission, and linings.

Producers of rubber – metal adhesive systems include Henkel, Lord Corp., and Chemetall.

8. Chemical Blowing Agents

8.1. Cellular Rubber Articles

Cellular rubber articles are produced using chemical blowing agents. These are almost exclusively organic compounds, which decompose at elevated temperature (see also → Foamed Plastics, Chap. 2.2). At least one of the decomposition products is a gas (e.g., nitrogen), which then produces foam structures in the rubber compound.

The decomposition of the blowing agents currently used is exothermic and irreversible, and takes place within a characteristic temperature range. The properties of an ideal chemical blowing agent were described in 1955 [189].

The decomposition temperature of the blowing agent in the rubber matrix is the most important factor determining the choice of agent. The incorporation of these agents into rubber compounds is carried out either in internal mixers or on open mixing rolls, with the decomposition temperature taken into account. For safety reasons the blowing agent is added separately from other additives as the last compounding ingredient.

Decomposition of the blowing agent occurs either during vulcanization (e.g., in the pressureless production of microcellular rubber profiles in UHF or LCM plants) or predominantly before cross-linking (e.g., in producing cellular rubber sheets by the expansion process).

The decomposition of the blowing agent and the cross-linking of the rubber compound must occur simultaneously for maximum utilization of the gas volume formed. This is particularly important in pressure-free vulcanization (e.g., profile production). If the coordination is not optimal (i.e., if vulcanization begins too early or too late), the blowing agent appears to be inadequate because of insufficient foaming. The decomposition of a blowing agent must not occur too spontaneously – in contrast to an explosive. Nitrogen is preferred as a propellant because of its low diffusion rate, which prevents cells from collapsing during foaming.

Chemical blowing agents are used for several reasons. Often a decrease in weight of the rubber article is desired. Since the latter is sold according to volume, a lowering in density is targeted. In addition, blowing agents are used to produce flexible materials for thermal insulation, which are of very low density. Shoe soles are produced from expanded rubber to increase the comfort of the wearer. The largest consumer of cellular rubber articles is the automotive industry. Here, microcellular rubber profiles are used for door and hood seals.

8.2. Product Types and Properties [190, 191]

Although in the past inorganic blowing agents (ammonium hydrogencarbonate, sodium hydrogencarbonate), which produced carbon dioxide as the propellant gas, were mainly employed, virtually only organic compounds are used today. These are more advantageous, despite their higher price, for the following reasons:

1) Significantly better dispersion in the rubber compounds
2) Higher gas yields
3) Production of gas over a narrower temperature range
4) Formation of nitrogen as the propellant gas, which has a much lower diffusion rate than carbon dioxide
5) Improved storage stability of rubber compounds containing blowing agents

The following organic compounds are sold as blowing agents for the rubber industry: *p*-toluenesulfonhydrazide (TSH), 4,4′-oxybis(benzenesulfonhydrazide) (OBSH), *N*,*N*′-dinitrosopentamethylenetetramine (DNPT), and azodicarbonamide (ADC).

Hydrazides. Hydrazine derivatives of organic sulfonic acids were developed by Bayer in 1949 as a new group of organic blowing agents

that produced nitrogen. Only two compounds from this group are still used for this purpose.

p-Toluenesulfonhydrazide is obtained by reaction of hydrazine with *p*-toluenesulfonyl chloride. The latter is formed in the chlorosulfonation of toluene for saccharin synthesis.

4,4′-Oxybis(benzenesulfonhydrazide) is produced by chlorosulfonation of diphenyl ether and subsequent treatment with hydrazine in an alkaline medium. Both sufonhydrazides are formed as relatively finely divided powders and are sold in this form or as preparations with paraffinic mineral oils.

***N,N′*-Dinitrosopentamethylenetetramine** is obtained by nitrosation of hexamethylenetetramine (urotropine). DNPT is a fine crystalline substance and is frequently produced as an 80 – 90 % mixture with an inert filler (chalk) and oil to desensitize it because of its very high decomposition rate.

Azodicarbonamide is produced from hydrazine and urea. The intermediate hydrazodicarbonamide is then oxidized to azodicarbonamide. Chlorine is used predominantly as the oxidant, but chlorates and hydrogen peroxide can also be used. Depending on particle size, azodicarbonamide is an orange to pale-yellow, crystalline powder (the smaller the particle size the lighter the color). Unlike the blowing agents TSH, OBSH, and DNPT, which are not ground for safety reasons (low decomposition temperature, or high sensitivity to abrasion or impact), ADC is usually ground up (e.g., in an air-swept mill). In this way, partially in combination with air classifying, ADC can be produced with a defined particle-size distribution. For the rubber industry, an average particle size of ca. 5 µm is common.

The blowing agents listed decompose at different temperatures. These temperatures are lowered, sometimes significantly, by the usual additives (zinc oxide, stearic acid, basic accelerators, etc.) present in a rubber compound. Table 10 gives an overview of these effects.

So- called *kickers* are used to lower the relatively high decomposition temperature and increase the rate of decomposition of azodicarbonamide [192]. These additives (zinc oxide, organozinc compounds, urea, and blowing agents such as TSH or DNPT) are added to the rubber compound as such or as a preparation with ADC (modified ADC).

Table 9 also gives an overview of the theoretical gas yields at standard temperature and pressure (STP).

For the sulfonhydrazides the data refer only to nitrogen. The water and steam produced in the decomposition of these blowing agents is not taken into account, because their effect on the blowing capacity can vary widely depending on the vulcanization process.

Reaction Mechanism. The decomposition of *sulfonhydrazides* can be formulated as follows

$$4\,R{-}SO_2{-}NH{-}NH_2 \rightarrow 4\,N_2 + 6\,H_2O + R{-}S{-}S{-}R + R{-}S{-}SO_2{-}R$$

sulfenic acid is first formed from the hydrazide, which disproportionates to form a disulfide and a polymeric thiosulfonate. Nitrogen and water are liberated in the process.

In the decomposition of *DNPT*, nitrogen is formed as the propellant gas. Small quantities of carbon dioxide, oxygen, water, formaldehyde, and nitric oxide have also been found. Traces of amines can be detected from their odor. This is generally regarded as the disadvantage of DNPT, since the fishy odor is transferred to the articles. In addition, a solid white residue is formed that has not yet been identified.

The decomposition of *ADC* is very complex. In the thermal decomposition of the dry substance in air the following quantities of decomposition products are obtained [191 – 193]:

gaseous products	32 %
solid residues	41 %
sublimate	27 %.

The gaseous products consist of

nitrogen	65 vol %
carbon monoxide	32 vol %
carbon dioxide	3 vol %.

The solid residues and the sublimate are composed mainly of urazole (39 %), cyanuric acid (26 %), hydrazodicarbonamide (2 %), and cyamelide (1 %). The figures are based on the quantity of ADC used.

The course of azodicarbonamide decomposition appears to follow two reaction pathways,

Table 10. Properties of organic blowing agents

Blowing agent	Decomposition temperature, °C	Gas yield at STP, mL/g	Heat produced on decomposition *, J/g	Start of decomposition in rubber compounds, °C	Lowering of decomposition temperature by
TSH	ca. 105	120	1531	ca. 85 – 90	basic substances
OBSH	ca. 165	125	1127	ca. 120	basic substances
DNPT	ca. 210	240	4352	ca. 120 – 130	acids, glycols, urea
ADC	ca. 200	220	1842	ca. 150 – 160	zinc oxide, bases, glycols, urea, acids

* Determined in nitrogen at 50 bar.

depending on the compounding formulation and the temperature:

$$H_2N{-}CO{-}N = N{-}CO{-}NH_2 \rightarrow N_2 + CO + H_2N{-}CO{-}NH_2$$

$$H_2N{-}CO{-}NH_2 \rightarrow NH_3 + HNCO$$

$$2\,H_2N{-}CO{-}N{=}N{-}CO{-}NH_2 \rightarrow H_2N{-}CO{-}HN{-}NH{-}CO{-}NH_2 + N_2 + 2\,HNCO$$

The hydrazodicarbonamide formed is the starting material for urazole (**54**).

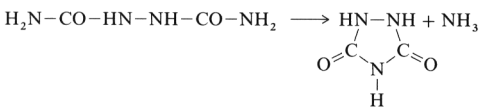

54

Cyanuric acid is formed by trimerization of the intermediate cyanic acid. Both acids are also present in their tautomeric forms, isocyanic and isocyanuric acids. An additional solid polymerization product, cyamelide, is also formed.

The decomposition of the blowing agents is exothermic. Sometimes considerable quantities of heat are produced (Table 9) [194].

Since blowing agents do not decompose spontaneously because of their relatively high dilution in rubber compounds, generally no heat buildup occurs to yield undesirable consequences (e.g., local overcure or "burns" in the interior of the article).

The dosage of blowing agent depends on the desired density of the finished article, the proportion of filler, and the gas yield of the particular agent. The usual dosages are 3 – 10 wt % based on the rubber. In a few cases [e.g., for insulating materials with very low densities (<0.1 g/cm^3)] the dosage can be 30 wt % or more.

Economic Aspects. Individual blowing agents differ considerably in their industrial importance. *Azodicarbonamide* has the greatest market share. Azodicarbonamide and its modified forms are used to produce a wide range of porous rubber articles. The reasons for this are its relatively low price and high gas yield.

Sulfonhydrazides have decreased in importance because of their high price and low gas yields compared to ADC and DNPT. Also, the low decomposition temperature of TSH gives rise to problems in incorporation.

The importance of blowing agents based on *N,N′-dinitrosopentamethylenetetramine* has also decreased despite its high gas yields and low price. The reasons for this are the amine odor, the formation of compounds regarded as toxicologically harmful (formaldehyde, *N*-nitrosamines), and safety considerations concerning the handling of DNPT. The production and use of this blowing agent have been almost completely abandoned in Western Europe. DNPT is used predominantly in Asia for the production of expanded sole plates.

Storage and Transportation. Blowing agents can decompose violently under the effect of heat or on contact with certain chemicals (e.g., bases, acids, or oxidizing agents). On storage, transport, or handling of chemical blowing agents, warming should definitely be avoided. Storage temperature should be kept much less than 50 °C. These agents should therefore not be transported or stored near heaters, and other heat sources such as sunlight should be excluded. The proximity of sparks of flames (e.g., as a result of electrostatic charge buildup or welding work) should also be avoided.

If blowing agents decompose as a result of unforeseen external factors, intensive water cooling must be provided immediately. Only through this measure can the exothermic decomposition reaction be terminated.

Chemical blowing agents must not be transported by air for safety reasons. The various regulations and hazard classes must be adhered to for rail, road, or ship transport.

Producers and Trade Names. Important producers of blowing agents are listed in Table 11, together with trade names.

Table 11. Producers and trade names of chemical blowing agents

Producer	Trade name	Products
Bayer, Germany	Porofor	ADC, TSH
Dong Jin Chemical Ind., Korea	Unicell	ADC, TSH, OBSH, DNPT
Eiwa Chemical Ind., Japan	Cellular	ADC
	Neocellborn	OBSH
	Unifor	TSH
	Vinyfor	ADC
Herbron, Spain	Vinstab	ADC
Kum Yang, Korea	Cellcom	ADC, TSH, OBSH, DNPT
Olin Chemicals, United States	Kempore	ADC
Otsuka Chemical, Japan	Unifoam	ADC
Sankyo Kasei, Japan	Cellmic	ADC, TSH, OBSH, DNPT
Schering Industrial Chemicals, United Kingdom	Genitron	ADC, OBSH
Société Française d'Organo-Synthèse, France		ADC, OBSH
Uniroyal Chemical, United States	Celogen	ADC

9. Latex Chemicals [195], [196, pp. 621 – 652]

Latex is an aqueous, colloidal system. Therefore, water-insoluble substances in powder form must be incorporated into the latex in the form of dispersions because of their poor wettability, and liquid, water-insoluble substances in the form of emulsions. In addition to the usual chemicals in rubber such as sulfur, vulcanization accelerators, or antidegradants, the actual latex chemicals – e.g., dispersion agents, emulsifiers, stabilizers, thickeners, wetting and foaming agents, foam-stabilizers, antifoams, coagulating agents, and preservatives – play an important role in processing.

Dispersion Agents [195]. Sodium methylenebisnaphthalenesulfonate (condensation product of sodium naphthalenesulfonate and formaldehyde), which is used as a 2.5 – 10 % aqueous solution, has proved to be an effective dispersion agent, as has a 1.0 – 2.5 % ammoniacal solution of casein.

Emulsifiers [195, 197]. *Anionic* emulsifiers, such as alkali salts of fatty acids, can be applied in anionic latices. They are particularly important as wetting and foaming agents in latex technology. *Cationic* emulsifiers are of less importance because few cationic latices are used.

The most important emulsifiers in latex technology are the *nonionogenic* ones, such as poly(glycol ethers). They do not affect the charge on the latex particles. While poly(glycol ethers) show good stabilizing action toward the chemical effects of certain compounding ingredients, some poly(ether thioethers) do not. An increase in chemical stability may be undesirable if processing of the latex compound also includes a coagulation step.

Stabilizers [195] should impart stability toward mechanical and chemical effects as well as temperature changes, to the latex compound (for stability toward oxidative degradation, see Chap. 3).

Most emulsifiers (see above) are also stabilizers. Cationic emulsifiers generally give cationic latices particularly high mechanical stability.

Casein in the form of a 5 – 10 % ammoniacal solution is frequently used as a weak stabilizer. A small increase in the mechanical stability of synthetic latices can be achieved by adding sodium methylenebisnaphthalenesulfonate.

Wetting and Foaming Agents [195]. Common latices generally do not wet other materials sufficiently, even when they contain emulsifiers. To reduce the surface tension of the latex so that it is suitable for impregnation, addition of a *wetting agent* – usually an anionic emulsifier – is necessary. Sodium isobutylnaphthalenesulfonate and alkali sulfonates are highly active wetting agents for processing latex compounds. For certain synthetic latices, the addition of a wetting agent is superfluous, depending on the emulsifier system used for their production.

Whereas sodium isobutylnaphthalenesulfonate hardly has any foam-forming effect, which is desirable for the impregnation of smooth materials and coatings, alkali sulfonates are excellent *foaming agents*. They have proved particularly effective in combination with the soaps usually employed (e.g., sodium oleate).

In other impregnation processes, *antifoams* are often added to prevent surface damage caused by foam bubbles. Long-chain alcohols have some effect, but it is often inadequate. Some silicone oils are particularly effective.

Foam Stabilizers (for Production of Foamed Rubber Articles) [195]. In the processing of latex foam – both in the production of molded articles and in the finishing of foam-rubber-coated woven tissue – the foam poured into the mold or painted on the woven tissue often collapses before gelation and coagulation because of an increase in pore size. Poly(ether thioethers) have proved to be effective foam stabilizers, which inhibit the collapse of the foamed-up latex compound and at the same time lower its viscosity, so that it can be poured readily, even into complicated molds; they also prolong the period up to the start of gelation or coagulation. Better processibility, particularly in painting processes, is achieved.

Thickeners [195]. A relatively high viscosity is desirable for latex compounds used to coat textiles; it is also sometimes necessary for compounds for dipping, to achieve a certain wall strength. This applies particularly to certain synthetic latices, which have a relatively low polymer concentration and therefore too low a viscosity for certain processing purposes. However, the viscosity cannot exceed a certain level; otherwise, the very fine air bubbles present in the compound no longer rise to the surface, which leads to defective films.

Thickeners include polysaccharides, albumens, casein, tragacanth gum, pectin, glue, gelatins, and agar-agar. They are applied as 5 – 20 % aqueous solutions. Poly(vinyl alcohol) and polyacrylates are important synthetic thickeners.

Similar to emulsifiers and stabilizers, thickeners increase the swelling in water of articles made from synthetic latices. Therefore, only that level of thickener should be added that is necessary to achieve the targeted increase in viscosity, and excess thickener should be avoided.

Certain thickeners are also creaming agents and bring about creaming of less concentrated latex compounds. This is another reason why no more than the quantity of thickener needed to achieve the desired viscosity is used. The danger of creaming is less pronounced in more concentrated latex compounds.

For the reasons given, sometimes conventional thickeners cannot be used, so the latex compound is sensitized to a certain extent by the addition of substances such as electrolytes, heat-sensitizing agents, or acids. Sensitization is associated mostly with an increase in viscosity.

In certain cases a mineral filler, bentonite, is used, which increases the viscosity of the latex compound even at low dosages. Under some circumstances, however, addition of bentonite can lead to an undesirable thixotropic effect.

The increase in viscosity that can be achieved depends not only on the concentration of the latex and the nature and concentration of the thickener, but also on the chemical composition and specific colloidal properties of the latex, and the molecular structure of the rubber material.

Coagulating Agents [195, 198]. One stage of latex processing is always coagulation. Almost all synthetic latices are considerably more resistant to the sensitizing effects of coagulating agents than natural latex, because of the relatively high content of emulsifiers used in polymerization.

Coagulating Agents with Powerful and Spontaneous Action. Electrolytes, e.g., carboxylic acids (formic acid, acetic acid, and lactic acid), polyvalent metal salts (calcium and aluminum chlorides or nitrates), and organic salts (cyclohexylamine acetate), are also used as coagulating agents. With natural latex, a compact coagulate is obtained; whereas latices of nitrile or styrene – butadiene rubber coagulate as flakes or crumbs.

Like natural latex, most chloroprene rubber latices form a compact coagulate with powerful and spontaneously acting coagulating agents. However, the coagulating agents must be employed with chloroprene rubber in a higher concentration than with natural latex.

Weak Coagulating Agents That Act Only Gradually or on Warming. Ammonium salts,

zinc oxide, sodium and potassium silicofluoride, 2-mercaptobenzimidazole and its zinc salt, zinc 2-mercaptobenzothiazole, poly(vinyl ethers), poly(oxypropylene glycol), and certain functional siloxanes are weak coagulating agents.

The action of weakly coagulating electrolytes is based on a gradual discharge of rubber particles, which leads to an increasing sensitization of the compound. The compound coagulates on warming or with continuing sensitization at room temperature. One theory of heat sensitization states that heat-sensitizing agents are generally compounds with a polyether chain. Such compounds absorb water at room temperature. At increased temperature the water is eliminated. In this state the molecules are capable of forming adducts not only with one another, but also with nonionic and anionic tensides. Through adduct formation the tensides are no longer in a position to exercise their actual function, which is to stabilize the latex compound; they are precipitated as insoluble adducts. This leads to sudden coagulation, whereby a closed seamless gel film is formed on the mold.

Although poly(vinyl methyl ether) and poly(oxypropylene glycol) are very suitable for heat sensitization of natural latex, their action on latices of nitrile and chloroprene rubber is unsatisfactory. The best agents for the latter are certain functional siloxanes.

Preservatives [195]. In many cases, casein or hemoglobin solutions, which tend to putrefy with time, can be formed in latex compounds, thus giving the latex an unpleasant odor. This putrefaction can be hindered by the addition of sodium *o*-phenylphenolate, for example.

Rubber Chemicals for Latices [195]. All rubber chemicals are added in emulsion or dispersion form if they are water insoluble or in aqueous solution if they are water soluble. The water used is soft (i.e., distilled, condensed, or rainwater).

A prerequisite for problem-free vulcanization of latex articles is uniform distribution of sulfur in as fine a form as possible, so that sedimentation in the latex compound is effectively avoided. The use of colloidal sulfur in latex compounds is therefore particularly recommended.

Only the most rapidly acting *vulcanization accelerators* (i.e., dithiocarbamates) are generally used. For *protection against aging*, nonstaining phenolic or weakly staining aminic antidegradants are used.

The addition of *fillers* to latex compounds does not give a reinforcing effect like that observed in solid rubber technology. Therefore, only inactive, white fillers are used, if at all. An exception is the action of certain silica sols in natural latex.

The addition of *plasticizers* is unnecessary. Sometimes paraffinic oil, plant oil, or ether plasticizers are added to achieve particularly low modulus articles (e.g., for toy ballons). Chlorinated paraffins are added to provide flame resistance.

10. Flameresistant Rubber Articles

10.1. Combustion Process in Polymeric Materials

Rubber articles, such as underground conveyor belts or cables, should maintain their function for as long as possible in the event of fire. For this reason, flameresistance is absolutely essential. In the event of fire, thermal decomposition precedes actual ignition of a polymeric material. This decomposition occurs as a result of the effect of energy sources and leads to gaseous decomposition products, which, combined with atmospheric oxygen, form an ignitable mixture. Since further decomposition with exothermic oxidation takes place in the region of the energy source, the temperature of the gas mixture and the concentration of combustible gas increase.

If a concentration threshold and a temperature specific to a particular substance are exceeded, the mixture of oxygen and heated decomposition products ignites. At this point, most of the organic bonds in the polymer molecule are broken down. Carbon black, carbon dioxide, and water are formed in the flame. The energy liberated in this reaction then decomposes more polymer material and increases the quantity of decomposition products.

A complex series of chain reactions involving free radicals has been discussed for this oxidative decomposition of hydrocarbon polymers [199 – 201].

10.2. Testing of Flammability

Testing of the flammability of rubber is carried out mainly on finished articles. In the case of conveyor belts and driving belts, underground behavior is particularly important. Tests are prescribed in appropriate national standards. Similar considerations apply to the cable sector. The oxygen index (LOI), described in ASTM D 2863, gives an initial indication of the fire resistance of rubber compounds. The test determines the concentration of oxygen at room temperature, at which a rubber test article in a glass tube is still extinguished within 3 min after being ignited from above.

10.3. Inhibition of Combustion

As has already been mentioned, the combustion process can be divided into two different stages:

1) Formation of a combustible gas mixture from a polymeric material (pyrolysis)
2) Combustion of the pyrolysis products by a radical mechanism

Effective flameproofing can thus act at either stage. Inhibition of pyrolysis involves suppressing or slowing down the formation of all combustible products. This is achieved by using *fillers*. High dosages of inorganic fillers dilute the combustible polymer per unit volume, so that if ignition occurs, the quantity of pyrolysis products is reduced. This measure has a certain effect if flame-resistant, inactive fillers such as chalk or kaolin are applied, but in most cases is inadequate.

A much better effect is achieved by using inorganic fillers, which also undergo an endothermic reaction as a result of energy supply by an ignition source. The most important example is aluminum hydroxide, $Al(OH)_3$. Energy converts aluminum hydroxide into the inert aluminum oxide. This reaction takes place with elimination of water and is strongly endothermic (1211 kJ/kg aluminum hydroxide). The steam produced dilutes the pyrolysis products in the gas phase, and the energy necessary for its formation is no longer available for pyrolysis of the polymer. If all the aluminum hydroxide is converted into Al_2O_3, the residual oxide acts as a fire retardant only by diluting the polymer. Magnesium hydroxide is also used as a fire retardant but is commercially much less important.

A third effect can be observed when using zinc or calcium borate. In the event of a fire the borate melts and forms a sintered surface so that the rubber layer below is protected for a time against the decomposing effect of the ignition source.

The combustion of pyrolysis products is a radical chain reaction. According to SCHMIDT [202], a determining factor for the combustion rate is the increase in quantity of highly reactive OH radicals. All compounds that can replace these highly reactive species by less reactive radicals lower the combustion rate. The most important groups are chlorine or bromine compounds, which react according to the following mechanism.

$$RX \rightarrow R\cdot + X\cdot$$

$$X\cdot + RH \rightarrow R\cdot + HX$$

$$\cdot OH + HX \rightarrow H_2O + X\cdot$$

where X = Cl, Br

How readily HCl or HBr is liberated is decisive for the degree of flameproofing. In practice, chlorinated or brominated paraffins are used. The use of polychlorinated or polybrominated aromatics is decreasing sharply because of the possible risk of dioxin formation. Antimony oxide has a synergistic effect on the fire-retardant action of halogen compounds. The exact course of the reaction is not yet completely understood. However, the formation of $SbCl_3$, which decomposes in cooler areas near the flame via a series of stripping reactions with the formation of HCl, is undisputed [203].

11. References

General References

1. W. Hofmann: *Rubber Technology Handbook*, Hanser Verlag, München 1989.
2. J. v. Alphen: *Rubber Chemicals*, D. Reidel Publ., Dordrecht 1973.
3. R. O. Babbit: *The Vanderbilt Handbook*, R. T. Vanderbilt Company, Norwalk, Connecticut 1978.

4. Th. Kempermann: *Bayer Handbuch für die Gummiindustrie*, Bayer AG, KA 22 001 Part D, Rubber Chemicals, Leverkusen 1992.

Specific References

5. E. R. Rodger: "Vulcanization Systems" in A. Whelan, K. S. Lee (eds.): *Developments in Rubber Technology,* Applied Science Publ., London 1979.
6. L. Bateman: *The Chemistry and Physics of Rubberlike Substances*, Mac Laren, London – J. Wiley & Sons, New York 1963, p. 449.
7. G. Alliger, I. J. Sjothun: *Vulcanization of Elastomers*, Reinhold Publ. Co., New York 1964.
8. J. A. Brydson: *Rubber Chemistry*, Applied Science Publ., London 1978, p. 194.
9. A. Y. Coran, F. R. Eirich (eds.): *Science and Technology of Rubber*, Academic Press, London 1978, p. 291.
10. P. J. Flory: *Principles of Polymer Chemistry*, Cornell University Press, Ithaca 1953.
11. DIN 53 529 Beuth VdE Verlag, 1988.
12. D. G. Lloyd, *J. Rubber Res. Inst. Malays.* **22** (1969) 399.
13. A. Y. Coran, *CHEMTECH* 1983, 106.
14. W. Hofmann, *Kunststoffe* **80** (1990) 7.
15. Ch. Goodyear, US 3 633, 1839.
16. Th. Hancock, DT 9 952, 1843.
17. P. Kluckow: *Die Praxis des Gummichemikers*, Berliner Union, Stuttgart 1954, p. 349.
18. H. Bartzsch, *Kautsch. Gummi Kunstst.* **41** (1988) 455.
19. Th. Kempermann, U. Eholzer, *Gummi Asbest Kunstst.* **26** (1973) 272.
20. A. Stevenson, R. S. Virdi, *Kautsch. Gummi Kunstst.* **44** (1991) 445.
21. H. L. Fischer, A. R. Davies, *Ind. Eng. Chem.* **40** (1948) 143.
22. G. Oenslager, *Ind. Eng. Chem.* **25** (1933) 232.
23. Farbenfabriken v. F. Bayer & Co., DE 265 221, 1912;
    DE 266 619, 1912;
    DE 269 512, 1913;
    DE 280 198, 1914 (F. Hoffmann, K. Gottlob).
24. PIRELLI & CO., DE 380 774, 1919 (G. Bruni).
25. R. T. Vanderbilt, US 1 413 172, 1920 (B. E. Lorentz).
26. D. F. Cranor, *Ind. Rubber J.* **58** (1919) 1199.
27. A. Maximoff, *Caoutch. Gutta Percha* **18** (1921) 10 944, 10 956.
28. C. W. Bedford, L. B. Sebrell, *Ind. Eng. Chem.* **13** (1921) 1034.
29. G. Bruni, E. Romani, *India Rubber J.* **62** (1921) 18.
30. I.G. Farbenindustrie, DE 573 570, 1931 (E. Zaucker, L. Orthner, M. Bögemann).
31. N. J. Morrison, M. Porter, *Rubber Chem. Technol.* **57** (1983), including cited original references.
32. W. Hofmann, *GAK Gummi Asbest Kunstst.* **32** (1979) 158.
33. D. W. Chasar, *Kautsch. Gummi Kunstst.* **42** (1989) 31.
34. D. B. Seeberger, *Kautsch. Gummi Kunstst.* **42** (1989) 875.
35. Th. Kuhlmann, *Kautsch. Gummi Kunstst.* **42** (1989) 878.
36. Bayer, Technical information, KD 1396 a, Leverkusen 1991.
37. Current intelligence bulletin 22, Ethylene Thiourea (ETU), NIOSH, April 11, 1978.
38. Degussa, EP 0 178 444, 1984 (W. Schwarze, S. Wolf, H. Remmel, H. Lambertz).
39. K. M. Davies, A. Fitch, *Kautsch. Gummi Kunstst.* **44** (1991) 838.
40. M. A. Thomson, *Tyretech 91 Conference*, Berlin, Oct. 24 – 25, 1991, Crain Communication, London 1991, p. 76.
41. D. Henschler (ed.): *Die Nitrosierung flüchtiger Amine am Arbeitsplatz*, VCH Verlagsgesellschaft, Weinheim, Germany 1986/1987.
42. R. H. Schuster, F. Nabholz, M. Gmünder, *Kautsch. Gummi Kunstst.* **43** (1990) 95.
43. H. J. Graf, H. Kleinknecht, K.-H. Menting, *Gummi Asbest Kunstst.* **44** (1991) 679.
44. H. Magg, Th. Kempermann, *Kautsch. Gummi Kunstst.* **35** (1982) 1039.
45. B. G. Crowther, B. Holt, C. Clarke, *Tyretech 91 Conference, Berlin Oct. 24 – 25, 1991*, Crain Communication, London 1991, p. 118.
46. D. G. Bloomfield, *137. ACS Meeting*, Las Vegas, Nevada,
    1990, paper 53.
47. W. C. Enstra, *Kautsch. Gummi Kunstst.* **32** (1979) 756.
48. I. Ostromyslenki, *India Rubber J.* **52** (1916) 420.
49. W. C. Enstra, *Kautsch. Gummi Kunstst.* **43** (1990) 790.
50. W. C. Enstra, D. B. Seeberger, *Kautsch. Gummi Kunstst.* **39** (1986) 929.
51. W. C. Enstra, *Kautsch. Gummi Kunstst.* **42** (1989) 414.
52. U. Eholzer, Th. Kempermann, *Kautsch. Gummi Kunstst.* **33** (1980) 696.
53. PRI (Plastics and Rubber Institute): *Radiation Processing for Plastics and Rubber*, International Conf., Brighton 1981.

54. J. H. Bly: "Radiation Curing of Elastomers," *Rubber Div. ACS Meeting*, Philadelphia, 1982.
55. C. D. Trivette, E. Morita, O. W. Maender, *Rubber Chem. Technol.* **50** (1977) 570.
56. T. Kempermann, *Gummi Asbest Kunstst.* **35** (1982) 238, 402.
57. M. Bravar, T. Kempermann, B. Ljubic, *Kautsch. Gummi Kunstst.* **36** (1983) 95.
58. E. H. Andrews, *J. Appl. Polym. Sci.* **10** (1966) 47.
59. W. Hofmann, *Gummi Asbest Kunstst.* **20** (1967) 602, 714.
60. G. I. Lake et al., *Rubber J.* **146** (1964) 30.
61. O. Lorenz, C. R. Parks, *Rubber Chem. Technol.* **36** (1963) 194, 201.
62. J. L. Bolland, G. Gee, *Trans. Faraday Soc.* **42** (1946) 236, 244.
63. H. S. Dweik, G. Scott, *Rubber Chem. Technol.* **57** (1984) 908.
64. R. Criegee, *Angew. Chem.* **87** (1975) 765.
65. D. Brück, H. Königshofen, L. Ruetz, *Rubber Chem. Technol.* **58** (1985) 728.
66. J. C. Andries, D. B. Bos, H. E. Diem, *Rubber Chem. Technol.* **49** (1975) 41.
67. Beaudry, *Rubber Age (N.Y.)* **69** (1951) 429.
68. E. Kluckow, E. F. Seeberger, *Gummi u. Asbest* **12** (1959) 862.
69. E. W. Bergstrom et al., *Rubber World* **148** (1963) 61.
70. A. Dibbo, *Gummi Asbest Kunstst.* **18** (1961) 120.
71. T. Kempermann, R. Clamroth, *Kautsch. Gummi Kunstst.* **15** (1962) 135.
72. T. Kempermann, *Gummi Asbest Kunstst.* **26** (1973) 90.
73. T. Kempermann, *Gummi Asbest Kunstst.* **27** (1974) 84.
74. D. Brück, *Kautsch. Gummi Kunstst.* **42** (1989) 760.
75. G. P. Langner, T. Kempermann, O. Rosenthal, *GAK Gummi Asbest Kunstst.* **32** (1979) 81.
76. R. P. Lattimer et al., *Rubber Chem. Technol.* **59** (1986) 263.
77. R. P. Lattimer et al., *Rubber Chem. Technol.* **56** (1983) 431.
78. L. Vogel, D. Groß, *Kautsch. Gummi Kunstst.* **44** (1991) 509.
79. M. Braden, A. N. Gent, *J. Appl. Polym. Sci.* **3** (1960) 90; **6** (1962) 449.
80. I. H. Thelin, A. R. Davis, *Rubber Age (N.Y.)* **86** (1959) 81.
81. O. Lorenz, C. R. Parks, *Rubber Chem. Technol.* **34** (1961) 818.
82. T. Kempermann, *Kautsch. Gummi Kunstst.* **15** (1962) 422.
83. H.-W. Engels, H. Hammer, D. Brück, W. Redetzky, *Rubber Chem. Technol.* **62** (1989) 609.
84. H. Fries, *Gummi Fasern Kunststoffe* **40** (1987) 238.
85. T. Kempermann, W. Redetzky, H.-W. Engels, *Kautsch. Gummi Kunstst.* **40** (1987) 1168.
86. R. R. Pandit et al., *Kautsch. Gummi Kunstst.* **36** (1983) 22; **37** (1984) 115.
87. R. R. Pandit et al., *Kautsch. Gummi Kunstst.* **38** (1985) 787.
88. D. Brück, H.-W. Engels, *Kautsch. Gummi Kunstst.* **44** (1991) 1014.
89. P. Schneider, *Angew. Chem.* **67** (1955) 61.
90. I. C. Ambelang et al., *Rubber Chem. Technol.* **36** (1963) 1497.
91. D. Parkinson: *Reinforcement of Rubbers*, Lakeman, London 1957.
92. J. B. Donnet, A. Voet: *Carbon Black-Physics, Chemistry and Elastomer Reinforcement*, Marcel Dekker, New York 1976.
93. G. Kranz: *Reinforcement of Elastomers*, Interscience, New York 1965.
94. B. B. Boonstra: *Reinforcement by Fillers*, Newness-Butterworth, London 1978.
95. E. M. Dannenberg, *Gummi Asbest Kunstst.* **41** (1988) 211.
96. K. Westlinning, *Kautsch. Gummi Kunstst.* **15** (1962) 475.
97. G. Kraus, I. Janzen, *Kautsch. Gummi Kunstst.* **28** (1975) 53.
98. G. Kühner, Degussa AG, *Was ist Ruß*, Frankfurt 1991.
99. S. Wolff, U. Görl, *Kautsch. Gummi Kunstst.* **44** (1990) 941.
100. W. M. Hess et al., *Kautsch. Gummi Kunstst.* **21** (1968) 689.
101. I. Janzen, G. Kraus, *J. Elastomers Plast.* **6** (1974) 142.
102. T. Timm, W. Messerschmidt, *Kautsch. Gummi Kunstst.* **27** (1974) 83.
103. S. Wolff, U.-E. Arnold, *Kautsch. Gummi Kunstst.* **43** (1990) 866.
104. S. Wolff, *Kautsch. Gummi Kunstst.* **27** (1974) 511.
105. B. B. Boonstra et al., *Rubber Chem. Technol.* **47** (1974) 468.
106. E. M. Dannenberg, *Rubber Chem. Technol.* **48** (1975) 410.
107. S. Wolff et al., *Kautsch. Gummi Kunstst.* **28** (1975) 379.
108. European Committee for Biological Effects of Carbon Black, Bulletin no. 4, Carbon Black and Health.

109. J. H. Bachmann et al., *Rubber Chem. Technol.* **32** (1959) 1286.
110. M. P. Wagner, *Rubber Chem. Technol.* **49** (1976) 703.
111. S. Wolff, *Kautsch. Gummi Kunstst.* **41** (1988) 674.
112. F. Bomo, J. C. Morawski, *Gummi Asbest Kunstst.* **41** (1988) 219.
113. I. Shah, E. F. Seeberger, *Gummi Asbest Kunstst.* **27** (1974) 592.
114. R. Bode et al., *Kautsch. Gummi Kunstst.* **20** (1967) 699.
115. H. Ferch, V. Flach, *Wochenbl. Papierfabr.* **10** (1987) 439.
116. G. Cameron et al., *Eur. Rubber J.* **156** (1974) 37.
117. M. W. Ranney et al., *Kautsch. Gummi Kunstst.* **26** (1973) 409.
118. Degussa AG, Verstärkungsadditiv, Si 69, Company brochure, Frankfurt 1988.
119. S. Wolff, *Kautsch. Gummi Kunstst.* **33** (1980) 1000.
120. S. J. Monte, *Gummi Asbest Kunstst.* **44** (1990) 941.
121. F. J. Washabaugh, *Gummi Asbest Kunstst.* **41** (1988) 120.
122. G. W. Ingle, *Mod. Plast.* **31** (1954) 69.
123. D. A. Smith, *Rubber J.* **153** (1971) 19.
124. R. Wommelsdorff, *Kautsch. Gummi Kunstst.* **41** (1988) 248.
125. W.-D. Glenz, R. Wommelsdorff, *Kautsch. Gummi Kunstst.* **39** (1986) 816.
126. W. Hamann, R. Wommelsdorff, *Kautsch. Gummi Kunstst.* **42** (1989) 494.
127. R. Wommelsdorff, W. Hamann: Contribution to the Poster Session of IKT (IRC), Essen, June 24– 27, 1991.
128. W. Lache: "Zur Frage der Toxizität von Schmierstoff-Grundölen", *Schmierstoffe und ihre Wirkung auf Gesundheit, Umwelt und Wirtschaftlichkeit,* Technische Akademie Esslingen, Jan. 19 – 20, 1987.
129. *Med. Res. Counc. (G.B.) Spec. Rep. Ser.* **SRS 306** (1968) 251.
130. EG-Richtlinie 88/490/EEC vom 22. Juli 1988.
131. *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans*, **33** (1984) April.
132. J. R. Griffiths, "The Selection of Plasticizers for Compounding Nitrile Rubber," *Danish Rubber Association (DGF)*, Nov. 30 and Dec. 1, 1990.
133. W. H. Whittington, *Elastomerics* **22** (1990) May, 17.
134. W. H. Whittington, *Rubber World* **183** (1981) March, 33.
135. M. Abele, *Kautsch. Gummi Kunstst.* **44** (1991) 454.
136. T. Inoue, *Int. Polym. Sci. and Technol.* **5** (1978) 99.
137. E. Rohde, H. Bechen, M. Mezger, *Scandinavian Rubber Conference SCR/89*, Tampere/Finland, Jan. 31 – Feb. 2, 1989.
138. Kirchhoff, Trowman, Schill & Seilacher GmbH, *Australian Rubber Conference*, Melbourne 1983.
139. J. A. Brydson: *Rubber Chemistry*, Applied Science Publ., London 1978.
140. U. Eisele, *Kautsch. Gummi Kunstst.* **33** (1980) 165.
141. G. M. Bristow, K. N. G. Fuller, A. G. Thomas, *NR Technol.* **14** (1983) 69.
142. J. L. Leblanc, *Gummi Asbest Kunstst.* **39** (1986) 528.
143. H. Ehrend, *Gummi Asbest Kunstst.* **44** (1991) 438.
144. H. E. Railsback, R. P. Zelinski, *Kautsch. Gummi Kunstst.* **25** (1972) 254.
145. S. R. Moghe, *Rubber Chem. Technol.* **49** (1976) 247.
146. E. S. Dizon, L. A. Papazian, *Rubber Chem. Technol.* **50** (1977) 765.
147. W. Hofmann, *Kautsch Gummi Kunstst.* **38** (1985) 777.
148. J. L. Leblanc, *Rubber Chem. Technol.* **54** (1981) 905.
149. E. A. Collins, J. T. Oetzel, *Rubber Age (N.Y.)* **100** (1969) May, 62.
150. E. A. Collins, J. T. Oetzel, *Rubber Age (N.Y.)* **102** (1970) March, 64.
151. J. R. Hopper, *Rubber Chem. Technol.* **40** (1967) 463.
152. Th. Timm, *Kautsch. Gummi Kunstst.* **31** (1978) 901.
153. B. G. Crowther, *PRI Rubber Conference 84*, Birmingham, March 12 – 15, 1984.
154. A. Haverland, W. Hofmann, *Kautsch. Gummi Kunstst.* **38** (1985) 1012.
155. P. A. Danilowicz, B. C. Howard, *Rubber World* **191** (1985) March, 21.
156. F. M. O. Connor, J. L. Slinger, *Rubber World* **187** (1982) Oct., 19.
157. D. G. Lloyd, *Prog. Rubber and Plast. Technol.* **4** (1988) 21.
158. D. Brichzin, *Kautsch. Gummi Kunstst.* **36** (1983) 451.
159. B. G. Crowther, *Plast. Rubber Int.* **9** (1984) 14.
160. H. Ehrend, L. Steger, *Kautsch. Gummi Kunstst.* **42** (1989) 902.

161. N. B. Fegade, N. M. Deshpande, W. Millns, *Kautsch. Gummi Kunstst.* **37** (1984) 604.
162. F. W. Barlow: *Rubber Compounding: Principles, Materials, and Techniques*, Marcel Dekker, New York 1988.
163. H.-W. Engels, M. Abele, *Rubber World* **204** (1991) 14.
164. Th. Kempermann, *Gummi Asbest Kunstst.* **27** (1974) 566.
165. H. Fries, R. R. Pandit, *Rubber Chem. Technol.* **55** (1982) 309.
166. M. Abele, Th. Kempermann, *Kautsch. Gummi Kunstst.* **42** (1989) 209.
167. E. Esch et al., *Kautsch. Gummi Kunstst.* **30** (1977) 524.
168. R. P. Wool, *Rubber Chem. Technol.* **57** (1984) 307.
169. E. Bister, W. Borchard, G. Rehage, *Kautsch. Gummi Kunstst.* **29** (1976) 527.
170. G. R. Hamed, *Rubber Chem. Technol.* **54** (1981) 576.
171. F. F. Wolny, J. J. Lamp, *Kautsch. Gummi Kunstst.* **37** (1984) 601.
172. J. T. White, *Rubber World* **189** (1984) 41.
173. R. Kretschmer, Th. G. F. Schoon, *Gummi Asbest Kunstst.* **19** (1966) 228.
174. Communication of D.O.G. Deutsche Ölfabrik, Kontakt 27, Hamburg 1982, company brochure.
175. L. Steger, *Kautsch. Gummi Kunstst.* **42** (1989) 124.
176. Du Pont, US 2 128 229, 1935 (W. H. Charch). US 2 561 215, 1945 (C. J. Mighton).
177. W. D. Timmons, *Adhes. Age* **10** (1967) 27.
178. Du Pont, US 3 307 966, 1963 (C. J. Shoaf).
179. D. W. Anderson, *Rubber Age* **103** (1971) 69.
180. Farbenfabriken Bayer, DE 1 301 475, 1965 (J. Dane).
181. A. Peterson, M. I. Dietrick, *Rubber World* **190** (1984) 24.
182. Semperit, EP 0 406 214 A1, 1990 (O. Beckmann).
Yokohama Rubber, DE 3 215 516 C2, 1990 (K. Shrinji).
Goodyear Tire & Rubber, EP 0 476 310 A2, 1991 (L. G. Widemann).
Goodyear Tire & Rubber, EP 0 475 222 A1, 1992 (L. G. Widemann).
183. W. Van Ooij, *Kautsch. Gummi Kunstst.* **44** (1991) 348.
184. Nitsche, *Kautsch. Gummi Kunstst.* **36** (1983) 572.
185. Whittaker Corp., *Kautsch. Gummi Kunstst.* **38** (1985) 105.
186. Sexsmith, *Gummi Fasern Kunstst.* **40** (1987) 558.
187. G. Klement, *Kautsch. Gummi Kunstst.* **5** (1971) 160.
188. S. Buchan: *Rubber to Metall Bonding Crosby*, Lockwood & Son, London 1959.
189. R. A. Reed: "The Chemistry of Modern Blowing Agents," *Plastics Progress 1955*, Iliffe and Sons Ltd., London 1955, p. 51.
190. J. F. Mark, N. G. Gaylord, N. M. Bikales (eds.): "Blowing Agents," *Encyclopedia of Polymer Science and Technology*, vol. **2**, Interscience, New York 1965, p. 532.
191. B. A. Hunter: "Chemical Blowing Agents," *Rubber Age (N.Y.)* **108** (1976) no. 2, 19.
192. D. G. Rowland, *Gummi Asbest Kunstst.* **41** (1988) 405.
193. R. A. Reed, *Br. Plast.* **33** (1960) 468.
194. S. Morisaki, M. Naito, T. Yoshida, *J. Hazar. Mater.* **5** (1981) 49.
195. Latex-Hilfsprodukte, *Bayer-Mitt. Gummi-Ind.* **32** (1963) 56 – 77.
H. Esser, G. Sinn, Bayer AG, Die Verarbeitung von Naturlatex, Company brochure, Leverkusen 1957.
196. Bayer AG: *Handbuch für die Gummi-Industrie*, Berliner Union, Stuttgart 1970, pp. 772 – 801.
197. W. Hofmann: *Nitrilkautschuk*. Berliner Union, Stuttgart 1965.
198. Bayer, Wärmesensible Latexmischungen, Company brochure, Leverkusen 1957.
199. C. Cullis, M. Hirschler: "The Combustion of Organic Polymers" Clarendon Press, Oxford 1981.
200. A. Landrock: "Handbook of Plastics Flammability and Combustion Toxicology", Noyes Publications, Park Ridge 1983.
201. J. Troitzsch: "Brandverhalten von Kunststoffen", Hanser, München 1982.
202. W. Schmidt, *Inst. Plast. Ind. Trans.* **12** (1965) 247.
203. J. W. Hastie, *J. Res. Nat. Bur. Stand.* **A** 77 (1973).
204. *Rubber Handbook*, Schill + Seilacher "Struktol" Aktiengesellschaft, Company Brochure, Hamburg 2000.
205. K. Menting, M. Hensel, H. Umland, Rubber Conference, American Chemical Society, Cleveland 2003.
206. R. Galle-Gutbrecht, M. Hensel, T. Mergenhagen, H. Umland, Int. Rubber Conference, Prague 2002.
207. H. J. Graf, H. M. Issel, A. Kühnberger, *Kautsch. Gummi Kunstst*. **49** (1996) 210.

208. H. Schulz, D. L. Steger, *Kautsch. Gummi-Kunstst*. **51** (1998) 402.
209. R. Galle-Gutbrecht, M. Hensel, K. Menting, T. Mergenhagen, H. Umland, Tire Technology Conference, Hamburg 2002.
210. F. F. Wolny, J. J. Lamp, *Kautsch. Gummi Kunstst*. **37** (1984) 601.
211. J. T. White, *Rubber World* **189** (1984) 41.
212. C. Clarke, T. Mergenhagen, B. Just, Semi-permanent release agents — theory and application, conference proceedings RubberChem 2001, Brussels.
213. R. Utz, M. Hensel, S. Sprenger, *Kautsch. Gummi Kunstst.* **48** (1995) 104.
214. C. Clarke, B. Just, T. Mergenhagen, Maximising Mould Utilisation, conference proceedings IRC 2001, Birmingham.
215. "Immer eine Idee besser", Degussa AG, Frankfurt 1998, 84.
216. Ullmann's, 7th online ed, chapter 7.3
217. Ullmann's, 5th ed. **A 23**, pp. 642 – 647.
218. European Chemical Industry Council(CEFIC): Best Available Techniques Reference (BREF), Document of Synthetic Amourphous Silica (2002) 10.
219. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 174.
220. Degussa, DE 1 467 019, 1963 (A. Becker).
221. PPG, DE 1 283 207, 1961 (G. Lagerstrom).
222. Sifrance, DE 2 224 061, 1972 (J. Donnet).
223. Rhone-Poulenc, EP 0 018 866, 1980 (M. Coudurier).
224. Degussa, DE 2 505 191, 1975 (B. Brandt).
225. Degussa, DE 1 467 437, 1965 (G. Bretschneider).
226. Degussa, DE 1 293 138, 1965 (H. Reinhardt).
227. Degussa, US 4 179 431, 1979 (E. Kilian).
228. Degussa, DE 1 567 440, 1965 (H. Biegler).
229. Degussa: Technical Bulletin Pigments: The Handling of Synthetic Silicas and Silicates, Pig. 28-3-2-598, 3rd ed., May 1998.
230. Kirk-Othmer, 4th ed., **21**, p. 998.
231. S. Brunauer, P. H. Emmet, E. Teller, *J. Am. Chem. Soc.* **59** (1936) 2682.
232. J. Seifert, G. Emig, *Chem. Ing. Tech.* **59** (1978) 475.
233. E. Koberstein, M. Voll: "Charakterisierung von hochdispersem Siliciumdioxid durch physikalische Adsorption von Gasen," *Z. Phys. Chem.* **71** (1970) 275.
234. IUPAC Manual of Symbols and Terminology, Appendix 2, Pt. 1, Colloid and Surface Chemistry, *Pure Appl. Chem.* **31** (1972) 578.
235. E. W. Washburn, *Proc. Nat. Acad. Sci. USA* **7** (1921) 115.
236. G. W. Sears jr., "Determination of Specific Surface Area of Colloidal Silica by Titration with Sodium Hydroxide," *Anal. Chem.* **28** (1956) 1981.
237. W. A. Heston jr., R.K. Iler, G.W. Sears jr., "The Adsorption of Hydroxyl Ions from Aqueous Solution on the Surface of Amorphous Silica," *J. Phys. Chem.* **64** (1960) 147.
238. a) A. Blume, "Analytical Properties of Silica – a Key for Understanding Silica Reinforcement," paper No. 73 presented at the ACS Meeting, April 13-16, 1999, Chicago, Illinois/USA;
b) *Kautsch. Gummi Kunstst.* **53** (2000) 338.
239. ASTM D 3765-04
240. J. Janzen, "Specific Surface Area Measurements on Carbon Black," *Rubber Chem. Technol.* **44** (1971) 1287.
241. A. Voet, J. C. Morawski, J. Donnet, "Reinforcement of Elastomers by Silica," *Rubber Chem. Technol.* **50** (1977) 342.
242. L. T. Zhuralev, *Langmuir* **3** (1987) 316.
243. L. T. Zhuralev, *Colloids Surf., A*, **74** (1993) 71.
244. E. F. Vansant, P. Van der Voort, K. C. Vrancken: Characterization and Chemical Modification of the Silica Surface, *Stud. Surf. Sci. Catal.* **93** (1995) 88.
245. Ullmann's, 5th ed., **A23**, p. 631.
246. J. Mathias, G. Wannemacher, *J. Colloid Interface Sci.* **125** (1988) 61.
247. A. Legrand: *The surface properties of silicas*, J. Wiley & Sons, New York, 1998, p. 286.
248. E. Papirer (ed). *Adsorption on Silica Surfaces*, Marcel Dekker Inc., New York 2000.
249. S. Wolff, *Kautsch. Gummi Kunstst.* **41** (1998) 675.
250. Degussa, DE 1 172 245, 1963 (G. Kallrath).
251. Degussa, DE 2 729 244, 1977 (P. Nauroth).
252. Degussa, EP 0 466 958, 1990 (M. Guenther).
253. Degussa, EP 0 672 731, 1994 (M. Ettlinger).
254. K. Albert, E. Bayer, B. Pfleiderer: *J. Chromatogr.* **506** (1997) 343.
255. Degussa, DE 1 074 559, 1959 (Chem. Fabrik Wesseling).
256. Degussa, DE 2 628 975, 1976 (P. Nauroth).
257. Degussa, DE 19 757 210, 1997 (H. Karbe).
258. Degussa, EP 0 090 125, 1983 (D. Kerner).
259. Degussa, EP 0 341 383, 1989 (D. Kerner).
260. M. P. Wagner: The Consequences of Chemical Bonding in Filler Reinforcement of Elastomers, paper presented at the Colloques Internationaux, Sept. 24-26, 1973, Le Bischenberg-Obernai, France.

261. F. Thurn, S. Wolff: Neue Organosilane für die Reifenindustrie, *Kautsch. Gummi Kunstst.* **28** (1975) 733.
262. S. Wolff, *Kautsch. Gummi Kunstst.* **34** (1975) 280.
263. http://www.degussa.com.
264. http://www.rhodia.com.
265. http://www.ppg.com.
266. http://www.huber.com.
267. a) S. Uhrlandt, A. Blume, "Development of HD-Silicas for Tires - Process, Properties, Performance," paper No 32 presented at the ACS-meeting April 4-6, 2000, Dallas, Texas/USA;
b) *Rubber World* **30** (2002) 226;
c) *Kautsch. Gummi Kunstst.* **54** (2001) 520.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Rubber, 5. Technology

Standard Article

Wolfram von Langenthal[1] and Jochen Schnetger[2]
[1]Bayer AG, Leverkusen, Federal Republic of Germany
[2]Bayer AG, Leverkusen, Federal Republic of Germany

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a23_421
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Latex Processing
1.1. Latex Testing
1.2. Transportation, Storage, and In-House Distribution of Latices
1.3. Compounding and Compounding Formulation
1.3.1. Additives
1.3.1.1. Colloidally Active Substances
1.3.1.2. Rubber Chemicals
1.3.1.3. Fillers and Pigments
1.3.2. Compounding
1.4. Processing Procedure
1.4.1. Processing of Rubber Articles
1.4.1.1. Dipped Goods
1.4.1.2. Rubber-Coated Metal
1.4.1.3. Rubber Threads
1.4.1.4. Cast Articles
1.4.1.5. Molded Foams
1.4.2. Processing of Functional Materials
1.4.2.1. Can Sealants
1.4.2.2. Adhesives
1.4.2.3. Elastified Concrete
1.4.2.4. Modified Mortar and Fillers
1.4.2.5. Paints
1.4.2.6. Bitumen Modifications
1.4.2.7. Chewing Gum
1.4.2.8. Acrylonitrile - Butadiene - Styrene (ABS) Polymers
1.4.3. Application on Substrates
1.4.3.1. Paper and Cardboard
1.4.3.2. Carpeting
1.4.3.3. Shoe Materials
1.4.3.4. Textile Finishing
1.4.3.5. Textile Printing
1.4.3.6. Textile Finishing with Rubber Primer
1.4.3.7. Gaskets
1.4.3.8. Nonwoven Fabrics
2. Solid Rubber Processing
2.1. Internal Mixer (Kneader)
2.1.1. Construction
2.1.2. Mode of Operation

2.1.3. Compounding
2.2. Open Mill
2.2.1. Construction
2.2.2. Mode of Operation
2.2.3. Compounding
2.2.4. Prewarming
2.3. Pellet and Strip Feeding
2.4. Production of Sheeted and Reinforced Rubber
2.4.1. Calenders
2.4.2. Calendering Process
2.4.3. Sheet Production in the Extrusion Process
2.4.4. Vulcanization of Rubber Sheets
2.5. Extruded Articles
2.5.1. Extruders
2.5.2. Vulcanization of Extrudates
2.5.2.1. Batch Processes
2.5.2.2. Pressure-Free Continuous Vulcanization
2.5.2.3. Continuous Vulcanization under Pressure
2.5.3. Profile Geometry Measurement
2.6. Molded Articles
2.6.1. Production Process
2.6.1.1. Molding Process
2.6.1.2. Injection Molding
2.6.2. Mold Cleaning-Contamination
2.6.3. Deflashing
2.7. Solutions of Rubber Compounds

[Top of Page]

1. Latex Processing

Wolfram von Langenthal

Compared to the amount of solid rubber consumed, the quantity of latex that is processed as such is small. In 1992, worldwide natural rubber consumption was 5 125 000 t [of which 560 000 t (10.9 %) was processed directly as latex], and synthetic rubber consumption was 9 192 000 t [of which 1 710 000 t (18.7 %) was processed directly as latex]. These latex quantities are calculated in terms of solid matter so that all figures can be compared directly. The Brussels customs nomenclature differentiates between synthetic rubber latices and other polymer dispersions [i.e., poly(vinyl chloride), ethylene – vinyl acetate copolymer, or polyacrylates]. Irreversible conversion of the polymer from the thermoplastic to the elastomeric state by vulcanization is used as the criterion for identifying a synthetic latex. The conversion is determined by the stress – strain behavior of corresponding polymer films.

Only the smallest quantities of latex are processed by the rubber industry. A much greater demand comes from the textile, paper, and leather industries and the producers of paints and varnishes.

Synthetic latices are virtually always produced by emulsion polymerization. Unlike emulsion rubbers (⟶ Rubber, 3. Synthetic), reaction is carried out to a higher degree of polymerization, which leads to more branching, but also to more intense cross-linking of the polymer chains. The degree of branching and cross-linking is varied, depending on the latex type; for example, it can be held back, as in foam latices, or even deliberately increased, as in reinforcing latices.

Synthetic latices consequently have a higher solid material content than latices that are used for the production of emulsion rubbers. This is desirable, partly for industrial processing but mainly for transportation and storage.

For latices, in addition to the characteristics of the polymer itself, the colloidal properties of the latex are of considerable importance for their application. They affect the processability of the latex and the film-forming characteristics of the polymer.

Synthetic latices can be classified into three categories: those without reactive groups, those with groups capable of reaction, and polymers with groups capable of self- cross-linking. Those without reactive groups are cross-linked by vulcanization provided they are diene polymers. Polymers containing reactive groups, mainly carboxyl, can be cross-linked by the addition of a second component such as a heavy-metal oxide or precondensed amino resins. Reactive polymers contain groups that react at elevated temperature with one another or with other reactive groups introduced by copolymerization, resulting in cross-linking. Examples of these are *N*-methylol groups and their esters or ethers.

The mechanical properties of *vulcanized latex films* are similar to those of corresponding rubber sheets. Their sensitivity to water, i.e., swelling, is however, greater than that of corresponding rubber sheets because of the higher emulsifier content of latex films.

Freshly tapped *natural latex* contains, besides 32 – 40 % rubber, 2 – 3 % proteins and lipoids, 0 – 3 % resin, 1.5 – 4 % glucosides, and 0.2 – 1.0 % inorganic substances. On concentration by centrifugation a large proportion of these other latex components is removed with the serum [1].

The remaining latex components represent the colloidal chemical system, which maintains the metastable latex state. All *synthetic latices* differ from natural latex in this respect, since the former all contain tensides as stabilizers. In some instances, emulsion polymers are free from surfactants. In such cases, however, the polymer contains hydrophilic, preferably sulfate, groups. Because of their soap content the compounding of synthetic rubber latices is different from that of natural rubber (NR) latex. This often results in very different processing techniques. NR latex may be prevulcanized by adding sulfur and ultraaccelerators (⟶ Rubber, 4. Chemicals and Additives); and heating it for a prolonged period of time at 70 – 80 °C under continuous slow agitation. Prevulcanized NR latex is referred to as Revultex. Excess vulcanization agents are removed with the liberated serum in subsequent centrifugation. By drying at low temperature, films, which already show a certain vulcanization effect, can be produced from this Revultex. Similarly pre- cross-linked SR latices can be produced but are not available on the market [2], [3].

Latex polymers are characterized by the same international abbreviations as solid rubber grades. The letter X in front of the abbreviation indicates that the polymer contains reactive groups. Compounding ingredients are always calculated on the basis of 100 parts latex dry substance (phr).

1.1. Latex Testing [4]

Production monitoring by the producer and raw material control by the processor are equally important for maintaining the quality of latices and of the finished products made from them. ISO and respective national standard methods provide a framework for quality assurance. The latices are homogenized before sampling for quality control. Important *latex parameters* are as follows:

The content of coagulum is determined by filtration through variable filtration media, depending on the requirements. Gravimetric evaluation of coagulum content is possible, but visual assessment is usually sufficient [5].

The total solids content is determined by gravimetric analysis of a latex sample dried to constant weight [6].

The dry rubber content is determined by precipitation of the latex with suitable watermiscible solvents, such as acetone or isopropanol, sometimes with the aid of electrolyte solutions. After complete precipitation, the flocculated polymer is carefully freed from the other latex components by washing, and dried to constant weight [7].

The pH value is measured by using glass electrodes [8].

The surface tension is determined with the ring-detachment method.

The density of the latex is determined with a pycnometer [9].

The particle-size distribution can be obtained by various methods, all of which have their advantages and disadvantages. Examples are electron microscopy, ultracentrifugation, and light diffraction measurements. Generally, determination of the average particle size is sufficient (e.g., by turbidity photometry, static light scattering, or laser correlation spectroscopy). The reliability of the method of measurement also depends on the particle-size range, which should be scanned.

The viscosity is measured with a rotation viscometer. In practice, determination of the relative viscosity, which is obtained by measurement of the flow rate from a standard beaker with a defined orifice outlet, is usually sufficient [10].

The mechanical stability of the latex is affected by many factors. Only the measurement of the shear stability by using a high-speed disc stirrer (Klaxon instrument) has been standardized. The stirring time at 14 000 rpm until the first coagulated particles are formed is used as a measure of the mechanical stability [11]. In processing, other factors such as the drying rate also affect the stability of latex compounds. Thus, for control of the mechanical stability in the laboratory, special methods (e.g., squeeze-roller stability, stirrer test) must be worked out.

The chemical stability is a measure of the compatibility of the latex with various additives, such as vulcanization agents, resins, and fillers. Here also, specialized methods are often indispensable. The zinc oxide stability test, which involves addition of a finely dispersed zinc oxide to the latex, subsequent warming, and measurement of the temperature at which coagulation begins, is widely used. It cannot be applied with latices of polymers containing carboxyl groups. The chemical stability of latices is also frequently measured by titration with $CaCl_2$ solution.

The setting point is the temperature at which the latex thickens significantly just before freezing. It is a measure of the minimum storage temperature.

The coagulation point of heat-sensitizable latices is defined as the temperature at which a standard test compound coagualtes quantitatively under defined heating conditions and during slow stirring.

The gelation time is determined at 20 °C for foam compounds, to which a defined quantity of gelling agent (e.g., sodium silicofluoride or ammonium acetate solution) has been added.

The residual monomer content is usually determined by gas chromatography and is specified by latex producers as the upper limit in the safety data sheet or, on special agreement, as a specific value in the test report accompanying the product supplied.

Bacterial contamination is defined by determining the bacterial count.

The following tests are used for characterization of the *polymer* itself.

The white point temperature and the minimum film-formation temperature, as indirect measures of polymer hardness, are determined by drying the latex on a metal block with a defined linear temperature gradient along its length.

The Mooney viscosity and deformation hardness are measured for precipitated polymers in the same way as for solid rubber. Neither determination can be used for polymers with self- cross-linking groups, because the plasticity of the polymer might have been altered uncontrollably by pre- cross-linking and can change further during measurement [12].

For further characterization of the polymer, *films* are cast from the latex and dried.

The film hardness, measured with a Shore instrument, gives information on the polymer hardness. It is also applicable to reactive polymers [13].

The glass transition temperature, as an indirect measure of polymer hardness, is determined on films.

Tensile strength, tensile elongation, and elastic modulus are normally not provided by the latex producers. They are parameters used by the processor to test suitability and can also be used as relative values for determining the cross-linking or cross-linkability of the polymer. The measurements are carried out on films [14], [15].

1.2. Transportation, Storage, and In-House Distribution of Latices

The cost of transportation and storage of latices is very high, as with all liquids. Small quantities are transported in drums or containers, large ones in tankers. Apart from some exceptions, latices are sensitive to frost, which should be taken into account during transportation and storage. To stock up for the winter is therefore not unusual.

In addition to frost protection, containers must be kept closed so that no skin formation occurs as a result of evaporation of water. Storage tanks should be equipped with stirrers so that the latex can be homogenized after long storage periods. Phase separation can be prevented from the beginning by continuous slow stirring.

In the in-house distribution of latices and their compounds, their limited shear stability must be taken into account. Therefore, only feed pumps imposing low shearing force on the substance fed may be used. Feed pipes and pumps must always be kept filled with latex or latex compounds, so that polymer film formation by drying of latex residues is avoided. They should be readily accessible and exchangeable, so that cleaning work can be carried out without halting production.

1.3. Compounding and Compounding Formulation

1.3.1. *Additives*

Additives can be classified as those with colloidal effects, rubber chemicals in the actual sense, or fillers and pigments.

1.3.1.1. *Colloidally Active Substances*

Emulsifiers are added to latices to increase their mechanical and chemical stability. Anionic, cationic, or nonionic surfactants are chosen, depending on processing requirements. Nonionic surfactants show the best emulsifying and stabilizing effect at low temperature, their solubility decreases with increasing temperature. They are therefore preferred as stabilizers in heat-sensitive compounds. Anionic emulsifiers promote foam formation and are particularly suitable for latex compounds that are to be frothed. Cationic emulsifiers are very rarely used, and only in cationic latices, if these might be stabilized to an undesirably large extent by nonionic or amphoteric surfactants.

Protective colloids, such as caseinates, or certain thickeners and complexants, such as sodium hexametaphosphate, are specified as *stabilizers*. These also stabilize the latices against mechanical, chemical, and thermal effects.

Wetting agents can be important in impregnating processes to guarantee perfect penetration of the substrate by the latex compound.

Foaming agents are generally anionic surfactants, such as fatty acid soaps, fatty alcohol sulfates, and alkyl, aryl, or aralkyl sulfonates, succinates, and amidosulfosuccinates (succinamates). Surfactant combinations of uniform ionogenity always froth more intensely than mixtures of anionic and nonionic ones. Thickeners are important as foam stabilizers, in addition to functioning as protective colloids (see above).

Antifoams are used to avoid or destroy foams in various processing procedures. Silicone antifoams are generally more active than others but can easily produce “fish eyes” in smooth latex films (e.g., dipped goods). Their abhesive effect is prohibitive for the use in cord dipping compounds.

Coagulating agents with spontaneous action are used for dipped articles (predipping solution), in precipitation baths for spinning rubber threads, and in latex employed in pulp processes. Acids and salts of polyvalent metals, such as alkaline earths and aluminum, are normally used. Coagulating agents with delayed action, such as sodium silicofluoride or ammonium acetate, are used as gelling agents for foam production.

Heat-sensitizing agents are polyethers that are added to the latex compound and bring about spontaneous quantitative coagulation of the system at an elevated temperature that depends on the compounding formulation. At room temperature these compounds remain stable for long periods. Suitable products are poly(vinyl methyl ether), poly(oxypropylene glycol), poly(siloxane ether), quaternary poly(ether amines), and certain water-soluble polyetherpolyurethanes [16-20].

1.3.1.2. *Rubber Chemicals*

Rubber chemicals that are used in latex must be distributed as finely as possible in aqueous systems. Fine-grained zinc oxides of basic zinc carbonate, and flowers of sulfur or colloidal sulfur, are preferred to other types of zinc and sulfur chemicals.

Only the most active *ultra accelerators* can be used with latices. For NR, nitrile rubber (NBR), and styrene – butadiene rubber (SBR) latices, these are the dithiocarbamate and mercapto accelerators. For polychloroprene (CR) latices, thiourea and guanidine accelerators are used. As with solid rubber, accelerator combinations are almost always applied. The accelerators named are water insoluble. For some processes, water-soluble accelerators are employed (e.g., sodium diethyldithiocarbamate and pentamethyleneammonium-*N*-pentamethylenedithiocarbamate and xanthate).

Polymers containing carboxyl groups can be cross-linked ionomerically with zinc oxide but also with precondensed amino resins by condensation. Self- cross-linking polymers contain thermoreactive groups, such as *N*-methylols. They react with one another and with carboxyl or amide groups by condensation reactions.

With the exception of the paper industry, *antioxidants* are usually added by the producer to latices of polymers having reactive groups. Further addition of antioxidants is then necessary only if very particular stability is required. In the case of all other latices the antioxidant must be added by the processor. Sterically hindered mono-, di-, and polynuclear phenols, benzimidazoles, dihydroquinolines, and arylamines are important antioxidants in latex processing. Synergistic effects are known in latex processing, as with solid rubber. There are no water-soluble antioxidants. Ozone protection is rarely required for latex articles. The effect of ozone protection agents is increased by ozone protection wax.

Water-insoluble, powder rubber chemicals must be dispersed before addition to the latex, and oily substances must be emulsified. Water-insoluble chemicals have the advantage that they remain almost quantitatively in the polymer during processing and are not removed in uncontrollable quantities with the serum, as water-soluble ones are.

1.3.1.3. *Fillers and Pigments*

In contrast to solid rubber, there are no active fillers for latices. Whiting, china clay, slate flour, barytes, and other materials undoubtedly have reinforcing action, but are used primarily to make the product cheaper or to increase its weight. In the case of paper coatings and latex paints the pigment is the most important component; it imparts the effect and the latex acts as a fixative. Aluminum oxide hydrate improves the flame retardancy of articles. Fillers can generally be stirred directly into the latex without predispersion. Colloidal silica is added in some cases to achieve a dry surface of the polymer. It can be added to the latex only as an aqueous preparation.

Inorganic and organic pigments are added for coloration. They must be resistant to vulcanization conditions and free from rubber toxins. Among the carbon blacks used as black pigments, those that are preferentially absorbed by the polymer, and will not be washed out with the aqueous phase, should be selected. In the case of organic pigments, those that have no migration tendency should be employed to ensure constancy of color.

1.3.2. *Compounding*

All *powder additives,* with the exception of most fillers, must first be dispersed in water by using a suitable dispersion agent. Ball, pearl, or for small quantities, laboratory mills with centrifugal action are used for this purpose. The particle size of normal dispersions of rubber chemicals is between 0.1 and 50 µm, and preferably 1 – 5 µm. Perfect distribution of such particles in the latex compound is guaranteed. All solid materials to be added to a latex compound can be dispersed together.

Oily and *waxy compounding ingredients* are emulsified before they are added to the latex. For this purpose, high-speed stirrers are used, which develop strong shear forces. The choice of emulsifier depends on the substance to be emulsified. Two or more emulsifiers are often used to achieve optimum results. Water-in-oil emulsions are rarely used in latex processing, because they generally increase the viscosity of latex compounds.

In production, the order of addition of various components to the latex is often of critical importance for the stability of the compound. No fixed rules exist. Preliminary tests should be carried out in cases of doubt. In any case the latex is stabilized first if necessary. Viscosity and pH are adjusted at the end.

1.4. Processing Procedure

Latex processing is divided into three sections in what follows: Processing of true rubber goods, processing to give compounds, and application to substrates.

1.4.1. *Processing of Rubber Articles*

1.4.1.1. *Dipped Goods [21]*

Dipped articles are seamless hollow articles that are formed from polymer solutions or dispersions by complete coating of a mold surface, followed by drying and cross-linking of the coating. The molds are coated by dipping them briefly into the polymer solution or dispersion so that polymer precipitates on their surface.

In some cases the opposite principle is used. A hollow mold made of separable parts that can be opened is filled with the latex mix, and after forming a layer by precipitation on the inner surface of the mold, the excess is poured out.

The latex compounds precipitated on molds are dried in hot air and vulcanized. Immediately after dipping, the mold is rotated continuously to achieve uniform thickness of the latex layer and prevent drip formation. The mold can be reused more quickly if the article is removed after drying and vulcanized in a tumbler.

Articles with *very thin walls* (e.g., condoms) are produced by dipping the mold briefly into the vulcanizable NR latex compound, drying, further dipping to avoid defects, renewed drying, and vulcanization.

Coagulant or Predipping Process. If thicker walls are required (e.g., for gloves) the coagulant process is used. The heated molds are first dipped into an aqueous, alcoholic, acetonic, or mixed solution, usually of calcium nitrate. The mold is then turned several times so that the solution is distributed uniformly over the mold's surface and is further concentrated by evaporation. The mold thus coated, is then dipped into the latex compound, which is coagulated by the electrolytes and deposited on the mold surface as a gel.

The concentration of the electrolyte solution and the residence time of the mold in the latex compound affect the thickness of the layer of gel formed. With increasing thickness the diffusion rate of the electrolyte solution decreases. The wall thicknesses that can be achieved are therefore limited to ca. 0.8 mm. Before vulcanization the coated mold should be rinsed to remove water-soluble components.

The coagulant process allows two-layer construction of hollow articles. After gel formation on the mold, the latter is dipped into a second latex compound and an additional gel layer is formed by the electrolyte solution diffusing through the first layer. CR – NR laminates are produced in this way. The adhesion to one another is excellent.

Heat Sensitive Process. Very high wall thicknesses (e.g., in nipples for baby bottles and electrician gloves) can be achieved by the heat sensitive process. By addition of heat-sensitizing agents to the latex compound [poly(vinyl methyl ether) to NR latex and organopolysiloxanes to SR latex], compounds are produced that are stable over long periods if stored at ca. 20 °C, but coagulate spontaneously at elevated temperature, generally 36 – 40 °C. Molds that have been warmed to ca. 70 – 90 °C are dipped in the latex compound, which is then spontaneously precipitated on the surface as a gel with relatively high thickness. The temperature and heat capacity of the mold and, to a small extent, the residence time in the latex compound affect the thickness that can be achieved. The limit here is ca. 4.0 mm.

Textile-reinforced gloves can be produced by the coagulation and the heat sensitive processes. A tightly fitting knitted glove is pulled over the mold and the process is carried out as described. For these types of protective gloves, the rubber layer must adhere perfectly to the textile, but the latex compound should not penetrate it. The latex compounds are thickened to inhibit this penetration.

1.4.1.2. *Rubber-Coated Metal*

Corrosion-resistant coatings are produced on metal surfaces by electrophoretic precipitation of latex compounds. Such surface rubberization must often be hard and scratch resistant. For this, ebonite compounds produced from NR latex, which are vulcanized with ca. 45 phr sulfur, are used [22].

1.4.1.3. *Rubber Threads*

A vulcanizable NR latex compound containing zinc 2-mercaptobenzothiazole (ZMBT) as accelerator and secondary sensibilizer is extruded through glass spinnerets into a precipitation bath containing acetic acid. The acetic acid flows in the direction opposite that of extrusion. After complete coagulation the threads are washed to remove acetic acid, which impedes vulcanization, and other water-soluble substances from the gel. The gel is then dried and vulcanized. The latex threads may come into contact with one another only after predrying; before that, they would adhere to one another irreversibly. The rubber threads are produced in diameters of 0.2 – 1.2 mm [23].

1.4.1.4. *Cast Articles*

Articles produced in hollow molds by casting can be considered a special category of dipped articles. Metal molds, which are used with heat-sensitive latex compounds, are comparatively rare. They are employed in the production of meteorological balloons, for example. Porous hollow molds of porcelain or gypsum are most frequently used. Here, the precipitation of the polymer takes place mainly by absorption of the serum through the pores. Very high wall thicknesses are achieved in this way. Hollow articles with small openings, such as hot-water bottles and toy animals, can be produced by this method. The latter are the only latex articles of this type that contain a great proportion of filler — up to 300 phr whiting or china clay. The porous molds gradually become clogged with serum components and thus rendered unusable [24].

1.4.1.5. *Molded Foams [25]*

The well-known molded foams for mattresses, upholstery, shoe parts, etc., are produced from NR or SBR latices. Sulfur, zinc oxide, and accelerator combinations are added for vulcanization, in addition to antioxidants. To achieve foamability, a surfactant is added to the latex (e.g., an aqueous solution of potassium oleate). Surfactant mixtures and, in certain cases, small quantities of thickeners can be used to improve foam formation, stability, and structure (see Section Colloidally Active Substances). A particular characteristic of the foam structure is produced by the addition of diphenylguanidine or triethyltrimethylenetriamine to the vulcanization system. These accelerators contribute less to the vulcanization but more to improving the foam structure and making it more uniform. For perfect foam formation and as small a decrease in volume on drying as possible, compound systems are used that have the highest possible concentration.

In foam production the mechanically produced latex foam is first gelled and then vulcanized. Either the Dunlop or the Talalay process is used.

Dunlop Process. By far the greatest proportion of molded foams is produced by the Dunlop process. Here an aqueous sodium silicofluoride dispersion is added to the readily mixed, vulcanizable latex compound during the mechanical foaming process. The quantity depends on the desired potlife of the foam compound, which is generally adjusted to 8 – 12 min, which is sufficient to fill the foam molds. Contrary to NR latex mixes, SBR latex mixes need not be completely gelled at room temperature before the closed molds are heated in saturated steam (100 °C) to vulcanize the latex foam. After vulcanization the articles are removed from the mold, washed, and dried. Small foam articles for the shoe industry are

produced in open molds. As soon as these are filled and the excess foam has been removed with a doctor blade, they are gelled in an IR field, dried in hot air, and vulcanized. No subsequent washing is necessary.

Talalay Process. In the Talalay process the vulcanizable latex compound is frothed in a mechanical mixer, to give a foam with relatively high density, and then poured into molds. The closed molds are evacuated. The latex compound foams to its final density because of the vacuum applied and thus fills the mold completely. By cooling the mold to ca. – 30 °C the structure of the foam is fixed. Carbon dioxide is led into the mold, thus lowering the pH of the frozen foamed latex compound, which then gels. The foam may now be thawed without changing its structure, and the mold is heated in stages to 110 °C. After vulcanization is complete, the article is removed from the mold, washed, and dried.

Foams from CR latex generally have lower compression resistance than those from NR or SBR latex. The compression deflection properties can be improved by addition of aromatic or aliphatic diisocyanates, e.g., methylenediphenyl-4,4′-diisocyanate (MDI), during foaming. However, in this case the pH of the strongly alkaline latex compound, which can be as high as 12, should be adjusted to 10. Salts of a strong acid and an amine (e.g., diethanolamine hydrochloride) are used for pH adjustment.

The vulcanization agents and antioxidants usually used in processing CR latex are added here also [26].

1.4.2. *Processing of Functional Materials*

1.4.2.1. *Can Sealants*

Filled food cans cannot be sealed by soldering. The joint grooves of the lid are filled with an elastic lining to achieve absolutely tight joints. These aqueous sealing materials must correspond to appropriate food regulations. They consist of polyester resin with a melting point of >100 °C, resinates, filler, high-solid SBR latex, antioxidants, pigments, leveling agents, and thickeners. Very specific rheological requirements are imposed on the finished sealants, because they must be sprayed onto the lid without dripping or be stamped on avoiding stringiness [27], [28].

Because of the high chemical resistance of the polymer, CR latex is used for sealants for *varnish and spray cans*. The compounds also contain components that are resistant to various chemicals and not necessarily the above-mentioned resins.

1.4.2.2. *Adhesives*

Water-based adhesives are continually gaining importance because of safety standards at the workplaces and environmental protection. *Poly(vinyl acetate) dispersions* are probably the oldest and most widely used adhesives of this class. They are mainly used for wood and packaging. *Polyacrylate dispersions* have a substantial market in the building industry (e.g., for tiles and floor coverings). *Polyurethane dispersions* are replacing solvent systems in the automobile and shoe industries. They are also important for the adhesion of formica sheets to chipboard and fiberboard, particularly for kitchen furniture. *SBR latices* are used as packaging, expanded polystyrene, and in the building industry. *CR latices* are playing an increasingly important role as shoe adhesives, but are also used for adhesion of upholstery foam.

In processing to manufacture adhesives, polymer dispersions must often be mixed with thickeners, emulsifiers, and protective colloids for stabilization; resins to increase tackiness; possibly leveling agents; fillers; etc. Cross-linking by using resins or isocyanates that react sluggishly with water is often employed, whereas classical vulcanization is not [29].

1.4.2.3. *Elastified Concrete*

The addition of NR, CR, or SBR latices to cement – sand mixtures gives good adhesion of the concrete to various surfaces (e.g., metal or old concrete). In addition, the concrete is elastified, which renders the layers less prone to cracking. The latex used must be sufficiently stable during compounding to avoid coagulation by the electrolytes present in the cement. In addition, it must not impair the hydraulic binding of the cement. The polymer is not cross-linked. Elastified concrete is used, for example, in ventilation engineering in mines; for coating metal air channels to suppress flow noises and vibration; as a covering for decks of ships; and for vibration-proof floorings. Since it usually hardens rapidly, it is mixed in small quantities and applied by troweling [30].

1.4.2.4. *Modified Mortar and Fillers*

The elastic properties and good adhesion, that are achieved with elastic concrete are used for special mortars and fillers. For the latter, horizontal and lateral adhesion plays an important role, as does the priming effect for the adhesion of the subsequent surface layer (e.g., in PVC floor coverings or carpeting) [30].

1.4.2.5. *Paints*

Latex is important as a binder for latex paints for indoor and outdoor use. Polyacrylate and vinyl acetate copolymer dispersions are the most widely used raw materials. XSBR latex is less important because of its slight tendency of discoloration. Large quantities of filler are incorporated into the latex without any problem, as demonstrated particularly well by structured interior and decorative plasters. Pigment distribution and binding do not pose any problems either. The rheological properties of latex paints are important for surface structure and layer thickness. Therefore, thickeners of varying rheological characteristics are incorporated into the compounds, depending on the desired effect [31].

1.4.2.6. *Bitumen Modifications*

Both hot bitumens and bitumen emulsions are modified with latices. The thermoplastic bitumen is elastified by the incorporation of elastic segments. Optimal effects over long periods are obtained if the rubber particles swell in the bitumen, which achieves efficient blending.

Anionic bitumen emulsions can be mixed readily with latices. Besides elastification, the addition of latex improves the low-temperature flexibility and raises the softening point of the bitumen. The addition of 10 – 20 % latex based on solids

content, is usual. Bitumen emulsions modified in this way are used as moisture barrier in underground structures and in roofing. The bituminous sealing layers can be exposed to building movements without cracking. The lowering of the brittle-point and raising of the softening point contribute to the weather resistance of roofings.

Cationic bitumen emulsions are modified with cationic latices, so that their breaking characteristics are not affected. Modified cationic emulsions are used for surface treatment in road construction. The modified emulsion breaks immediately after application. The bitumen – latex mixture therefore does not flow away from the road even during light rain and remains completely available to bond chips. Modification with latex improves adhesion of the gravel and the weather resistance.

Hot bitumen is modified by spraying the latex on the hot stirred mixture, which is generally at a temperature of more than 180 °C. The water in the latex evaporates immediately on the surface, and only the rubber particles are blended mechanically into the hot bitumen. Modification with latex has the advantage, compared to the use of polymer powders, of much finer distribution because of the considerably smaller size of the latex particles. The quantities of latex used are between 3 and 5 %. This is sufficient to reduce the thermoplasticity to such an extent and to impart so much elasticity to the bitumen that weather-resistant road coverings, which can take up high shearing forces, can be obtained.

1.4.2.7. *Chewing Gum*

Special SBR latices, whose composition and production techniques are specified precisely by FDA, FCC, and BGA regulations, are coagulated by the addition of salts. The coagulum is washed and dried. The gum base is produced from the dried polymer, ester resins, microcrystalline waxes, and whiting by mixing in a kneader. Later, sugars or sugar substitutes and flavorings are introduced into the gum base. The chewing gum material is finally molded and packaged. Apart from different compounding, latices whose polymers have very low Mooney viscosities are used for bubble gum [32].

1.4.2.8. *Acrylonitrile – Butadiene – Styrene (ABS) Polymers (see also ⟶ Polystyrene and Styrene Copolymers)*

Polybutadiene latex is used as grafting base for ABS polymers. Styrene and acrylonitrile are grafted onto the polybutadiene by emulsion polymerization. The demand for latex in this area is very high. The consumers are raw material producers in the plastics industry.

1.4.3. ***Application on Substrates***

The use of latices on widely varying substrates is much more important in terms of quantity than production of rubber articles or functional materials.

1.4.3.1. *Paper and Cardboard*

Paper Coating [33] (⟶ Paper and Pulp, ⟶ Paper and Pulp, ⟶ Paper and Pulp). Demands are continually being increased on the quality of prints on paper and cardboard. Natural paper has an uneven surface. Cellulose fibers and fillers absorb printing inks to different extents, which leads to blurred print.

Paper and cardboard are coated to improve printability. An aqueous pigment dispersion, the paper coating color, is applied on one or both surfaces. A uniform smooth surface is thus obtained and printing can be carried out with perfect contours and as high a color reproduction consistency as possible. To ensure that the pigment layer adheres perfectly to the paper and shows optimum cohesion, a binder dispersion must also be added. These dispersions are now predominantly SBR latices or dispersions of polyacrylates, both types introducing carboxyl groups into the polymer. They are known as synthetic binders, as opposed to natural ones such as starch, casein, or other proteins.

The printability of the coated paper is determined mainly by the nature and quantity of the pigments, as well as by the binder(s). Other additives, such as dispersion aids, wet-strength agents, thickeners, optical brighteners, and dyes, are added as processing aids to control certain properties of the coating, principally printability, as well as processibility of the coating color on the coating machines.

The polymer dispersions must be compatible with one another and with co-binders, e.g., cellulose ethers, starch, poly(vinyl acetate), and poly(vinyl alcohol). These co-binders mainly affect the viscosity and water retention capacity of the paper coating color. This multisubstance system must meet all technical requirements. At the high speeds of modern coating machines the coating color is subject to extremely high shear forces, so all products must also be rheologically compatible. The quantity of latex dry substance based on 100 parts pigment is only ca. 10 – 15 %, to which ca. 0.5 – 6 % co-binder can be added.

Paper Impregnation. Masking tape is impregnated with NBR or SBR latices or polyacrylate dispersions to improve its strength, also when moistened with water or other solvents. The tape can then readily be removed from the masked surface, after painting or varnishing work. When used as packaging materials these tapes have increased strength even under unfavorable weather conditions.

1.4.3.2. *Carpeting (see also ⟶ Floor Coverings) [34]*

The latex consumption of the carpet industry is very high and, in terms of volume, about the same as that of the paper industry. In needle felt floor coverings and tufted carpets, the latex polymer applied is important for the durability of the floor covering. However, woven carpets are treated to give a sizing effect.

Needle Felt Floor Coverings (see also ⟶). Needle felt, whose surface can be structured, is produced from textile fibers with or without woven fabric support. The needle felt is bonded by lick-roll application on the underside or by full-bath impregnation with a latex compound, usually XSBR latex. In exceptional cases, acrylate copolymer dispersions are used. Vulcanization or other cross-linking agents are not added. The use of fillers is possible only in lick-roll application; otherwise, the colors would become hazy. The light and aging resistance of modern XSBR latices is so high that discoloration of the carpeting or premature aging through underfloor heating is not expected. Resistance to water and

shampoo is also achieved, giving an ideal overall range of properties.

Needle felts, which are thermomoldable in three dimensions, are used for floor linings in cars, for carpeting car boots, and for parcel shelves. Here a latex of a thermoplastic polymer is used, which can be molded at < 150 °C in order not to deteriorate the fiber material and still guarantee dimensional stability of the molded article at 90 °C. Special SBR latices with high styrene content in the polymer are employed.

Tufted Carpets (⟶ Floor Coverings – Carpet Properties and Performance). In the tufting process, mechanical bonding of the pile thread to the backing is so low that a latex coating on the backside is essential for perfect anchoring. Highly filled NR, or preferably XSBR latex, is used. Vulcanization or additional cross-linking is unnecessary (see above).

A straight backing, which then also forms the underside of the carpet, is filled with up to 300 % whiting, based on the latex dry matter. Precoatings, which are followed by a foam coating, contain up to 700 % whiting. For bonding the pile and at the same time laminating a fabric onto the back, a maximum of 300 % filler is incorporated into the latex.

In all cases mentioned, the filler is added directly to the latex; it does not have to be predispersed. Depending on the application system and the total solids content of the compound, the viscosity is regulated by the addition of a thickener. For better distribution on the backside of a carpet and in the loops of the pile threads (good filament binding), the latex compound is often slightly foamed. Foam weights of 500 – 1000 g/l are generally used, never lower ones.

Foam Backing. To increase the elasticity and sound insulation, tufted carpets are frequently coated with foam on the reverse side. This foam is compounded similarly to molded foams (see Section Molded Foams). Unlike molded foams, carpet foams are filled. The filler makes the compound cheaper and increases the weight per unit area of the carpet, which is important for perfect laying. High quantities of filler naturally reduce elasticity and increase compression set.

Besides sodium silicofluoride gelling (Dunlop process), ammonium acetate is also used for gelation in foam coating of carpets. The former allows filler quantities of 90 % maximum, the latter, up to 140 %. In addition a nongel foam coating process is also used. Unlike for molded foam or gel foam, sodium sulfosuccin(am)ate is added to the latex as a foaming aid which also prevents the foam from collapsing during drying. The combination and quantity of surfactant added are the only factors determining the foam structure and stability. Filler quantities up to 180 % are possible. A gelling agent is unnecessary. The dried combination of fatty acid soaps and sulfosuccin(am)ates is neither tacky nor greasy and therefore does not impair the quality of the foam layer.

Gelling of the foam or slight predrying of the surface (nongel) takes place under an IR field. Immediately afterward the foam is dried in a tenter frame and vulcanized.

Woven Carpets. Woven carpets and gripper Axminsters are constructed in such a way as to give them suitable end-use properties. The backside finishing, mostly with XSBR latices or vinyl acetate copolymer dispersions, essentially gives a variation in hand. The latices are generally processed with or without small quantities of fillers. The remaining processing corresponds to that of tufted carpets.

Carpet Underlays. Carpet underlays are used mainly under stretch-fitted tufted carpets with double backs or under woven carpets. They increase sound and heat insulation. Needle felt or nonwoven fabrics are sprayed with XSBR latices only at the surface, to improve the cohesion and abrasion resistance of the surface. *Foam underlays* correspond to foam backing of carpets. The foam is applied to thin carrier materials (e.g., spun bonded nonwoven).

1.4.3.3. *Shoe Materials [35]*

The shoe industry is not a direct consumer of latices but uses a large number of materials and adhesives made from them. Shoe parts made with the aid of latex are insoles, toe puffs, long counters, sock and side linings, inserts, and small foam parts. The shaft is fixed onto the insole, and the outsole is later mounted to the bottom. The toe puffs and counters give the tip of the shoe and the quarter box shape and rigidity. Small foam parts are used e.g., as heel lifts, and arch supports.

Insoles.
Cardboard Insoles. Cellulose sheet is impregnated with CR, XSBR, or NR latex, or their mixtures (depending on quality requirements), on a modified size press between the wet end and drying section, when it has a solids content of ca. 50 %. Latex precipitation in the pulp and subsequent sheet formation, instead of wet band impregnation, are also used. The finished cardboard is characterized by good dry and wet abrasion resistance. The rigidity achieved depends on the latex or latex mixture used. Antioxidants are added to the latices if this has not already been done by the latex producer. Zinc oxide is added to CR latex to achieve optimum aging resistance. Aluminum sulfate and sodium aluminate are added to the pulp to give a defined electrolyte content on the surface of the pulp fibers for successive precipitation of the latex particles.

All latices or latex compounds employed must be so stable or appropriately stabilized subsequently that they are not coagulated in the moist web of pulp by the electrolytes contained therein. During drying, which begins after impregnation, coagulation should take place so that polymer migration and thus delamination of the resulting cardboard is avoided. The polymer is not cross-linked or vulcanized, because the insole material must be moldable so that it can be adapted correctly to the contours of the shoe bottom. More than 50 % of the shoes produced have insoles made from this type of cardboard material, which is also used for other leather goods such as briefcases.

Needle Felt. Needle felt, which can also contain reinforcing woven fabrics, is impregnated with XSBR latices with high styrene content processed to form insoles for sport shoes or slippers. Fillers are added, depending on the requirements of the finished product. Vulcanization is not carried out. The stitch tear resistance of these textile articles is greater than that of insoles made from pulp. In addition the wet strength is higher and the water absorption lower than with pulp material. This is particularly important for flexible shoes.

Reclaimed Leather. Leather fibers obtained from vegetable and chrome-tanned leather clippings are formed into a web in the usual wet process on the Fourdrinier sieve. Generally NR latex, sometimes blended with high-styrene content SBR reinforcing latex or poly(vinyl acetate) reinforcing latex, is added to the pulp and precipitated on the leather fibers.

Chrome-tanned fibers contain sufficient electrolytes to bring about precipitation without other additives. For vegetable-tanned fibers the polymer of the latex is precipitated on the fiber by the addition of alum. Generally both types of fiber are used as a mixture. Vegetable tanned fibers increase the rigidity of the finished leather board, and chrome-tanned ones increase the softness. Vulcanization cannot be carried out because leather fibers are irreversibly damaged at temperatures of more than 80 °C. If the content of vegetable tanning agents (e.g., quebracho or mimosa) is high enough, antioxidants need not be added to NR latex, because these substances guarantee adequate protection against aging.

Leather board is used by the shoe industry mainly for making insoles, layered heels, out-soles, and long counters. It is also used as partition walls in leather bags. Particularly thick and impervious material is produced to make special containers (spinning cans), or it can — being flame retardant — be used in protective masks for electrical welders.

Shoe Stiffeners. Leather board is to a limited extent used as stiffening material. More commonly shagged cotton or rayon woven fabrics and thin spun bonded nonwovens are used as substrates for shoe stiffening materials. They are impregnated with highly crystallizing CR, or NR, XSBR, and polystyrene latex, SBR latex with a high styrene content, or their mixtures. The type of latex used depends on the shape of the shoe and the leather, and thus also on fashion. Thin leather and narrow shoe profiles require elastic toe puffs; thick leather uppers and broad shoe profiles require hard ones. Long counters are, in principle, always stiff.

The latex solid matter pick-up of the stiffener substrate is always more than 100 %. The deposited polymer not only hardens the stiffener material, but also acts as an adhesive in the mounting of shoes. Often, however, the impregnated material is additionally coated with an adhesive on one or both sides.

The stiffening material is cut to size; the edges are chamfered, where necessary, and the long counters are often — in the case of leather board, always — premolded in special workshops.

In shoe assembly, the stiffeners can be solvent or heat activated. Under the corresponding effect, these shoe parts soften and become tacky. In this state, they are inserted between the upper material and the lining, and subsequently mounted on the lasting machine. The non- cross-linked polymer applied as a latex here acts as both an adhesive and stiffening element.

Insoles, Inserts, and Small Shoe Parts. All of these are latex foam materials, and their production has already been described (see Sections Molded Foams and Carpeting).

Nonwoven Imitation Leather (see also ⟶). A nonwoven fabric made from fine, intensively needled fibers — preferably polyamide or polyester — with high apparent density, is used as the substrate. It is impregnated with an XNBR latex, which is rendered heat sensitive and vulcanizable. Sometimes a self- cross-linking NBR latex is employed. The latex compound is then coagulated in an IR field. Water-soluble substances can be removed by rinsing. Finally, the fiber – binder composite is dried in hot air and vulcanized.

This material, with its fine fibers intensively intertwined with one another and uniformly deposited binder agglomerates, has a structure very similar to that of natural leather. Like the latter, depending on the thickness of the material it can be split, buffed, sewn, and machine stitched.

NBR latices are used for imitation leather, mainly because of the grease, solvent, and plasticizer resistance of the vulcanized polymer. Elasticity properties are naturally also important. Besides zinc oxide, a small quantity of sulfur and an accelerator are employed as the vulcanization system for XNBR latices.

Nonwoven imitation leather is employed uncoated as heel grips and trimmings for sportsshoes, and coated as sock, quarter box and side lining, and occasionally as the upper material in shoes. It can also be used for leather bags.

Leather Finishing. Leather finishes consist of several layers of different polymer binders. The main components of base coats are dispersions of polyacrylates, polyurethanes, and NBR latices. Good adhesion to the substrate and between individual layers is a requirement of the finish, as are resistance to aging, dry, wet, and low-temperature flexing [36].

1.4.3.4. *Textile Finishing*

In the textile industry, aqueous systems are much easier to be applied and more versatile than those based on organic solvents. Latices are therefore very important.

Textile – textile or textile – foam laminates are produced by using either filled or unfilled latex compounds. Generally, latices of self- cross-linking or carboxyl- containing polymers are employed. The thickened latex compound is applied to the textile web with a doctor blade, and the second web is laminated onto it immediately with pressure rolls. This is followed by drying and thermal cross-linking.

To achieve air permeability of the laminates, the latex compound can be applied with a doctor blade as a thin foam layer, sprayed on, or lickroller applied in such a way that only the raised crossover points of the woven fabric or the peaks of the foam sheet are coated with laminating compound. This method naturally yields weaker laminates.

The *flocking of textiles* is usually carried out in an electrostatic field. The latex compound with a varying degree of filler is applied with a doctor blade to the textile to be flocked. The flocks are then electrostatically shot into this adhesive layer where they are anchored in a vertical position. Again this is followed by drying and thermal cross-linking. The product has

a softer feel if the adhesives applied are mechanically foamed.

Latex coatings are relatively rare. Examples are protective clothing (e.g., butchers' aprons from woven fabric), which are coated with vulcanizable NBR latex compounds, or fabrics for air mattresses, which are coated with vulcanizable SBR latex compounds. The use of NBR and polyurethane latices as tie coats for polyurethane transfer coatings is rather common.

Pile fabrics, such as velvet and plush, are frequently provided with a back sizing to increase pile strength. Self- cross-linking polyacrylate or ethylene – vinyl acetate (EVM) dispersions, or XSBR latices, are used here. The thickened, often foamed, latex is almost always applied by lick rollers, so that the softness of the pile article is minimally impaired. The quantities applied are kept as low as possible to maintain the appealing feel of the article. Examples are pile linings for parkas and plush covering for toy animals.

Base fabrics for emery cloth are finished with latex to achieve rigidity and dimensional stability. Also, through latex treatment a cloth surface is achieved that prevents too high a penetration of the resin applied to bind the abrasive materials. Penetrated resin is not available for binding the grain and impairs the strength of the woven fabric because of its high rigidity. Emery cloths must be water resistant, and often oil resistant, because they are in contact with drilling oil emulsions. Latices of self- cross-linking polymers are used, and NBR co- and terpolymers are the most suitable types.

1.4.3.5. *Textile Printing*

Polymer dispersions — particularly self- cross-linking polyacrylate, NBR, and SBR latices — are used as the pigment binder for textile printing. Good pigment- carrying and binding properties, perfect thickening characteristics, and high shear stability are prerequisites for a pigment binder. The print must be resistant to light, aging, washing, and dry cleaning.

Pigment printing systems are most commonly used in roller and rotary screen printing. In the former, because of the high working speed, enormous shear forces occur, which affect the printing pastes [37].

1.4.3.6. *Textile Finishing with Rubber Primer*

In rubber goods such as tires and conveyer belts, the construction materials rubber and reinforcing textiles have to adhere perfectly to one another. This is maximized by dipping the textiles in a rubber primer. The woven fabric or cord threads are impregnated with an adhesive compound, which in principle consists of **r**esorcinol –**f**ormaldehyde precondensate and **l**atex (RFL-dips, see also ⟶ Rubber, 4. Chemicals and Additives – Adhesion of Rubber to Reinforcing Materials).

Emulsion polymers of butadiene, styrene, and 2-vinylpyridine are used mainly as the latex component. The content of 2-vinylpyridine in the polymer is between 7.5 and 15 %. This content is critical for optimum adhesion. To achieve adhesion of “nonproblematic” substrates, such as rayon cord to rubber, latices with low 2-vinylpyridine content are employed, or the quantity of this polymer component is intentionally reduced by blending with SBR latex. However, polyamide cord requires latex with a high 2-vinylpyridine content. In exceptional cases, instead of pyridine – styrene – butadiene rubber (PSBR) latex, NBR or CR latices are used. For instance, NBR latex is used for the reinforcing fabric of NBR conveyor belts, and CR latex for the reinforcing threads of CR V-belts.

In the impregnation process, only the outer filaments (ca. 2 or 3) are wetted and bound together by the dried resin – latex film in order to maintain the inner flexibility of the cord. After impregnation the textiles are dried and, if this has not already been done, are thermoset in hot air to increase their dimensional stability [38].

1.4.3.7. *Gaskets*

Flat gaskets may or may not be reinforced.

Reinforced gaskets are generally produced ready for fitting according to construction drawings, and nonreinforced ones as sheetings from which individual seals can be die cut. Woven wire or metal band cloth and perforated, so- called spiked metal sheets are used as reinforcing materials. Gaskets made from spiked sheets are always produced ready for fitting. For example, the reinforcing material is coated on both sides with a mixture of mineral fibers, filler, rubber, vulcanizing chemical, and antioxidant, and then vulcanized in a press. The mixture is produced in an internal mixer. NBR latex is used as the rubber component because of its high resistance to heat, oil, and fuel. Slight swelling in the medium which is to be sealed, increases the tightness of the gasket. If this effect alone is predominant, then NR and SBR latices can also be used. The polymer employed is first swollen by addition of toluene to the latex and then mixed with other ingredients in the internal mixer.

Compressed mineral fiber sheets (It plates) are built up in layers on a calender. Two-ply gaskets can be produced in this way. The mixture of mineral fibers, filler, swollen NBR latex, vulcanization agents, and antioxidants is applied to the hot calender rolls. The material is thus formed into sheets and any solvent and water remaining are evaporated. For all gasket materials mentioned, semiultraaccelerators are used — in contrast to the usual latex processing — so that premature vulcanization in processing the mixtures is avoided.

Low-density gaskets are produced on paper machines from both mineral and cellulose fibers. The latex used as a binder is mixed with vulcanization agents (ultraaccelerators, antioxidants, and sometimes emulsifiers for stabilization), diluted, and added to the fibrous pulp in the mixing vat. On mineral fibers the mixture precipitates quantitatively because of the surface charge of the fiber, often without any additives. To achieve optimum properties the surface charge of the cellulose fibers is reversed with the usual retention agents before the latex compound is added to the pulp, so that the latex compound later precipitates perfectly on the fibers. The remaining processing corresponds to that used in paper production. The final step is vulcanization, which takes place in hot air or with contact heat.

NBR latices are generally used because of the oil, gasoline, and heat resistance required. In special cases, CR, NR, or

SBR latices can be used.

1.4.3.8. *Nonwoven Fabrics (see also, ⟶ Nonwoven Fabrics) [39]*

Nonwovens are used for interlinings in the apparel trade. They are also used as coating base, cover stock for hygienic articles, hospital and camping linen, filters, waddings and quiltings, insulating materials, reinforcements, abrasive and polishing disks, draperies, upholstery materials such as rubberized hair, geotextiles, wipers, etc.

These fabrics are, in general, no longer bonded by latex impregnation, but by other techniques. Latex treatment is often applied to achieve special effects (e.g., rigidity and resilience in the case of linings). Because of better decomposition in composting, latex-bound nonwoven fabrics for cover stocks are gaining more importance than thermally bonded ones. For some abrasive nonwovens, spraying with latex not only bonds the fibers to one another but also fixes the abrasive grit on the bonded fiber.

Rubberized hair for car seats is bonded with vulcanizable SBR or NR latex compounds and, if flame resistance is required, with CR latex compounds. The latex mix is sprayed onto the fiber mat and dried but not vulcanized. The web material is later cut to size, placed in molds with varying packing density, and vulcanized in the closed molds.

Wiping cloths made of nonwovens are partially bonded either thermally or by printing with polymer dispersions if only a short lifetime is required. Heavy-duty wipers for floors, windows, and cars are bonded more intensively.

For high-performance *window and car wipers*, the wiping effect is particularly important. This involves the strong adhesion to smooth surfaces (e.g., window panes) shown by natural chamois leather. These cloths function by completely wiping away water, as windshield wipers do, and absorbing it. Also, as with windshield wipers, the effect described is achieved through thin rubber lamellae. A vulcanized latex foam has many such lamellae in the form of cell walls of the foam pores. Relatively thin, needled, nonwoven fabrics, which are generally preimpregnated with a foamed latex compound, are coated on both sides with a thin latex foam layer. Because of the resistance to oil, grease, and to some extent, fuels, which are absolutely essential for these cloths, the use of NBR latices is preferable.

For *floor wipers* such a wiping effect is precluded because they should slide smoothly over the floor and absorb the cleaning water, which must then be wrung out easily.

Needle-punched nonwovens made from cellulose fibers are coated on both sides with foamed latex compounds. The very thin foam layer gives the article good abrasion resistance, without impairing its capacity to absorb or release water. Both vulcanizable and self- cross-linking NBR latices are used for these cloths.

[Top of Page]

2. Solid Rubber Processing

Jochen Schnetger

The production of rubbery materials with industrially useful properties from solid rubber involves the incorporation of a range of additives (⟶ Rubber, 4. Chemicals and Additives). The latter critically affect the properties with regard to process technology and permit cross-linking or vulcanization. To achieve a sufficiently high and uniform property level of the rubber articles produced, homogeneous distribution of all additives in the tough rubber matrix is necessary.

The raw compound is then processed further and molded by different procedures such as calendering, extrusion, pressing, injection molding, or coating processes, and vulcanized through further energy input. Cross-linked structures are thus built up, which convert the rubber blank into a rubber-elastic material. Processing is generally carried out in internal mixers, less frequently in open mills.

2.1. Internal Mixer (Kneader) [40-47]

2.1.1. *Construction*

Processing of rubber compounds in the internal mixer is a batch process. The important components are the ram, mixing chamber, kneading rotors, and the downward-opening drop or hinged door (see Fig. 1). The construction of the kneading rotors is very important for mixing intensity.

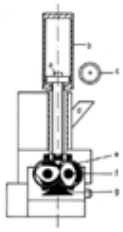

Figure 1. Schematic of an internal mixer with ram

a) Ram; b) Ram cylinder, c) Suction; d) Hopper; e) Mixing chamber; f) Kneading blades; g) Saddle

In principle, two different systems can be found (see Fig. 2):

1. *Tangential rotor systems with friction* (i.e., different peripheral speed of the kneading rotors whose ratio is defined as friction; ca. 1: 1.1). Mixing is performed between the blades, and between the blades and the mixing chamber wall.
2. *Intermeshing rotor systems without friction*, in which the rotor wings intermesh. Because of this intermeshing the rotors cannot run at different speed and friction due to the rotor blades is thus not possible. However, friction results

from different peripheral speed of the wings or screw blocks and the rotor core. This allows a more intensive mixing process compared with the tangential rotor systems.

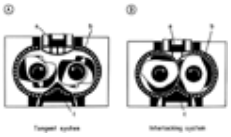

Figure 2. Kneading blade systems

A) Tangent system; B) Interlocking system

a) Ram; b) Cooling; c) Saddle

The advantages of the tangential rotor system are:

1. Good pulling-in action
2. Greater useful volume
3. More rapid emptying

The intermeshing rotor system has the following merits:

1. Rapid mixing
2. Good temperature control
3. Lower dumping temperature
4. High energy input in the shortest time
5. Good filler dispersion [ethylene – propylene – diene (EPDM), etc.]

2.1.2. *Mode of Operation*

After automatic or manual filling of the internal mixer with the compounding ingredients, the hopper is closed by lowering the ram. Ram pressures of 4 to 12 bar are applied.

The ram holds the material being mixed between the blades, and between the blades and the chamber wall, during mixing — thus enhancing mixing efficiency considerably. The ram is normally operated pneumatically. However, models with hydraulic control have been produced.

At the base of the mixing chamber is a *hinged (drop) door*, whose flaps can be closed hydraulically by a motor. The sliding door systems used previously resulted in emptying times that were too long and are no longer used. The kneading rotor systems and the inner wall of the mixing chamber are protected by electroplating against too rapid abrasion. Nevertheless, the chamber volume must be checked at least once a year, so that any changes can be dealt with immediately. Uniformly good dispersion of all compounding ingredients is then guaranteed only if the full volume, once optimized, is not altered.

Driving Mechanisms and Power. Internal mixers are equipped either with electrical driving mechanisms having constant speeds, or one or more fixed speeds, or with variable driving mechanisms and continuously variable speed control of the rotor system. Automatic mixer controls require continuously variable speed control. To deal with very high torques, the unidrive mechanism is generally used (branching of two motor shafts from the gear box). The mechanism is driven by a d.c. motor.

The driving power required to achieve the necessary mixing and shearing action is at least 6 kW, and usually 10 kW per liter.

Heat Removal. Depending on the plasticity of the material being mixed, considerable heat of friction can be produced in the mixing process, which must be removed effectively. This is achieved by cooling the entire working surface, the mixing chamber, the rotor systems, the ram, and the drop door. A reproducible and established temperature course during mixing is extremely important. The temperature of the cooling water and the flow rate must therefore be kept constant. The mixer may also have to be set at a given temperature before the beginning of the mixing cycle (i.e., to thermostat it and avoid water condensation or undesirable temperature jumps on starting up with the first batch).

Metering of Components, Measurement and Control. Modern mixing lines are equipped with *control systems*, which include weighing systems (see Fig. 3). A reproducibly optimized compound quality and an economical production process can be guaranteed only by computerized systems. A basic prerequisite for this is exact metering of all additives. This includes automatic weighing from day tanks that are filled from large silos; computer- controlled, sequenced hand weighing (belt weigher for filling the hopper); and the predominantly non-sequence- controlled weighing of small components. Permitted deviations in, for example, the 100-kg region are ± 0.1 kg.

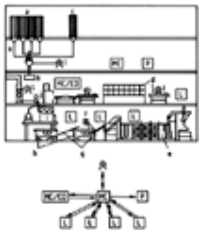

Figure 3. Internal mixer production line

a) Filler; b) Storage container; c) Changing container; d) Weighing small quantities of chemicals; e) Batch-off machine; f) Stock blender; g) Conveyer belts; h) Mill; i) Internal mixer; j) Pump; k) Plasticizer; l) Balance

MC = Master computer; L = Line computer; P = Printer; MC/CD = Master computer and control disk

The control of automatic balances is carried out by additive weighing and control of deviation after each component. The

dosage time and emptying (0-point) are monitored. The belt weigher filling the hopper is connected with a preset weight limit. Total balance results are shown by a monitor with component and tolerance data.

Plasticizers are injected through the side of the chamber wall with the ram closed after the weighing process. Small components, provided they are not weighed by hand under sequence control, are added via the belt weigher. Automatic weighing devices for small components function well.

The charging operation, the movement of the ram (aeration, cleaning), balance readings, speed changes, and emptying are all controlled. Regulation parameters include the cooling water temperature, the temperature of the bulk of the material, and the energy (torque). The last two can be varied by the rotor speed and the ram pressure. Changes in temperature or energy absorption can be controlled by the variable speed of the motor, whereby the given set points for time and temperature or energy input are compared with the actual value, and adjusted by means of the speed within the given tolerance limits. The control and regulation of the mixer, weighing system, other peripheral machines, and alarm indicators; data acquisition from field instruments; production monitoring; etc., are all carried out by the line computer assigned to each production line. Process data are processed by a master computer, which controls formulation, plans production, balances material, carries out documentation of production and process data, and is connected to a central computer (host). Desired process parameters can be stored in the master computer, which then controls the line computer for the mixer and the subsequent machines in the mixing line. These data can also be stored in individual line computers, which then report process parameters back to the master computer.

2.1.3. *Compounding*

Before compounds can be produced in the internal mixer, the *degree of filling* and the batch weight necessary for it must be determined from the filling factor. The filling factor is determined by the supply company in preliminary experiments. Depending on the compound, the degree of filling must, however, be optimized in an individual case for the actual operation. Internal mixers are operated underloaded, with reference to the volume of the chamber. Intermeshing systems (see Section Construction), should be filled even less than tangential systems to achieve good compounding results. Filling factors are 0.7 – 0.9 for tangential and 0.6 – 0.7 for intermeshing systems. The degree of filling can be based on both the mixing chamber volume and the so- called useful volume. On quoting the degree of filling the volume on which it is based must be given.

$$\text{Degree of filling (in\%)(chamber volume)} = \frac{\text{Batch weight} \cdot 100}{\text{Compound density} \cdot \text{chamber volume}}$$

$$\text{Degree of filling (in\%)(useful volume)} = \frac{\text{Batch weight} \cdot 100}{\text{Compound density} \cdot \text{useful volume}}$$

$$\text{Useful volume} = \text{Chamber volume} \cdot \text{filling factor}$$

$$\text{Batch weight} = \text{Chamber volume} \cdot \text{filling factor} \cdot \text{compound density}$$

Important factors affecting the quality of a compound are the degree of filling, the mixing time, the temperature of the bulk material, the speed of the rotor, and the ram pressure.

To be able to incorporate all additives into the rubber matrix, the raw rubber must have a viscosity that is suitable for processing. This is the case for synthetic rubbers, which have a wide range of viscosities, and for specially treated natural rubbers [constant viscosity (CV) grades]. Natural rubber must, however, generally be degraded first (masticated) (⟶ Rubber, 4. Chemicals and Additives – Peptizers). *Mastication* is carried out with the addition of peptizers in the internal mixer or sometimes on the mill. The mastication cycle in the internal mixer lasts 3 – 5 min, depending on the degree of degradation desired.

In a classical mixing operation the rubber, which has a viscosity suitable for processing, is either charged to the open internal mixer hopper via the belt weigher in whole bales (25- and 33.3-kg bales) or cut using a rubber cutter. After closing the hopper and pushing in the ram, the material is plasticized briefly, the ram is then raised, and the filler, plasticizer, antidegradants, stabilizers, or processing aids are added according to a predetermined time cycle. The bulk temperature can reach 180 °C, depending on the viscosity of the raw compound. The mixing time for the production of this *first batch* (basic compound) is ca. 2 – 5 min.

At the end of the mixing time the batch is discharged and shaped in the downstream equipment (see Fig. 3), cooled, and stored. Generally, after 24-h storage, accelerators and cross-linking agents are added in a second mixing stage (*two-stage mixing process*). The temperature of the compound must not exceed the cure onset temperature of the now mixed vulcanization system in this final mixing process, and is generally between 100 and 120 °C. The second, final compounding stage is also sometimes carried out on mills. Since the mixing temperatures for the batch and final compounding are very different, batch and final compounding mixers are used separately for reasons of capacity.

Compounds for highly dynamically stressed articles (e.g., tires) must be mixed very carefully to achieve optimum dynamic load- carrying ability. For this reason, particularly in the tire industry, mixing processes with *up to four stages* are used. For the production of technical rubber goods, two-stage processes, and sometimes one-stage, are usual. In *one-stage processes* accelerators and cross-linking agents are added in the final third of the mixing cycle. In this type of mixing the bulk temperature must not exceed the cure onset threshold of the accelerator system added. In addition, this type of mixing naturally requires the use of accelerator systems that are safe for processing.

The *upside-down process* is a variant for attaining high shearing forces rapidly. Here, fillers, antidegradants, stabilizers, or processing aids are charged first and, after a short premixing period, the raw rubber is added to the internal mixer. This mixing method is frequently used for special rubbers, such as EPDM, or rubber blends with polybutadiene. It is also used when very high quantities of filler are required. If conventional mixing were used in such a case, crumbling of the initially

compact rubber matrix would result, and the crumbly material could only build up sufficient shearing forces to form a compact compound after long rotation times, if at all. Mixing is carried out for a predetermined time and at a predetermined temperature or following a predetermined energy diagram.

Generally, the mixing time for the batch must be long, since too short a cycle time results in insufficient filler dispersion, which is virtually impossible to remedy in further processing steps.

After the mixing stage is finished the hot mixed material is discharged by rapid opening of the drop door. A *mill* situated directly below the hopper is used as a receiver. Here, the unshaped lumps of compound are rolled out to form sheets and are transported on a conveyor belt to a second mill with a stock blender (see Fig. 3) for further homogenizing and intensive cooling. The rubber sheet is then threaded into a batch-off machine (see Fig. 4). This machine coats the still warm sheets with external lubricants, cools, and stacks them. The sheets are passed through the machine in endless loops on cross-bars of a double- chain conveyor. They pass through a curtain of external lubricant spray and a blowing device fitted with ventilators to remove residual moisture from the loops and cool the sheets further. They are deposited in loops (wigwag system) or in strips, which are cut with circular knives before deposition. Coating with release agents lowers the surface tackiness of the raw compound to such an extent that the deposited sheets can be separated easily from one another for further processing.

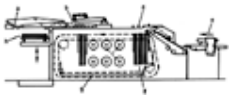

Figure 4. Batch-off machine

a) Knife; b) Gripper; c) Cooling; d) Lubricating agent; e) Input; f) Sheets; g) Lubricating agent bath; h) Wigwag stacking

Instead of the mills, special *extruders* can also be used. Here the compound falls into a large hopper equipped with rams as plugging devices. Depending on the desired form of the raw compound, the extruders can produce sheets or pellets (special extruder heads, pelletizers). Strainers allow the removal of foreign bodies. *Special extruders* are required if large internal mixers are used, as in the tire industry. The very high batch weights used here (up to 500 kg) can no longer be managed by mills in short processing cycles. Outside the tire industry, *strainer-extruders* and *pelletizers* are being increasingly introduced as downstream machines for rubber processing. Granulates or pellets can also be produced in batches on band granulators, which are located downstream of the batch-off machines. Granulates are more expensive to produce but allow further automatic processing steps in extruders or injection molding machines. With technical rubber goods a completely continuous process for the production of compounds — premixing in the internal mixer and processing in subsequent shaping extruders with the formation of strips or pellets — has not proved successful since temperature control, particularly in the case of sensitive vulcanization systems, is extremely difficult. In addition, changing from carbon black-filled to white compounds is extremely difficult because adequate cleaning of the equipment is nearly impossible. Sometimes such continuous operations are used in the production of cable compounds.

2.2. Open Mill [40], [41]

Roll mills are used widely in rubber processing. Their importance as mixers for the complete production of rubber compounds is small compared to the internal mixer. Nevertheless, they are still indispensable as downstream machines to the internal mixer (see Fig. 3), as prewarming roll mills for warm-feed extruders, as feeder roll mills for calenders, for mastication of natural rubber, for breaking stiffened compounds, and for the production of reclaims (refiner mills). They are sometimes also used for the final mixing of raw compounds in the second mixing stage, particularly when small quantities of compound or limited amounts of compounds with light fillers are produced. In the laboratory, they are employed for the production of compounds for quality control using rapid tests and for general compounding development (laboratory mills) as well.

2.2.1. *Construction*

An open mill consists of two parallel, horizontal, chilled casting rolls, which are fixed into gliding or rolling bearings, and rotate in contrary directions at different speeds (see Fig. 5). The front roll turns more slowly than the back roll.

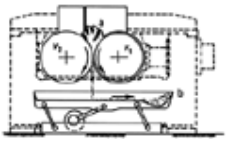

Figure 5. Construction of rubber mill

a) Rubber compound; b) Vibrating trough

v_1 = speed of front roll; v_2 = speed of back roll; $v_1 < v_2$

Roll mills can have constant speeds and fixed friction or controllable speeds and continuously variable friction. The machine is predominantly driven electrically (d.c. motor), or sometimes hydraulically. In modern mills, the rolls run on rolling bearings. The necessary driving power is 2 kW per liter of compound. Industrial roll mills have a friction of 1: 1.1 to 1: 1.2. Frictions of 1: 1.3 are also possible.

The dimensions vary from 100×225 (diameter × roll length in mm) to 200×450 for laboratory roll mills, and from 300×700 to 850×3000 for industrial machines. The rolls can be heated and cooled. While mixing, the rolls are generally cooled intensively by cooling channels below the roll surface (rolls with holes on their periphery, see Fig. 6).

Figure 6. Roll with peripheral holes

a) Cooling agent; b) Cooling channel; c) Roll

2.2.2. *Mode of Operation*

Mills operate with independent cooling and heating systems. The roll nip is regulated by moving the front roll. Two material guides are fixed on the sides and prevent penetration of the rubber compound into the bearings. Additional homogenization and cooling of the compound are achieved by setting up a stock blender. A *stock blender* consists of a pair of draw rolls, fixed above the front roll, to which the sheet is fed by the backward and forward motion of the contact roll (see Fig. 7). The sheet then passes back into the roll mill over a guide plate.

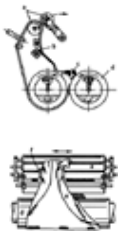

Figure 7. Mill with stock blender

a) Stock blender; b) Guide plate; c) Compound; d) Mill; e) Compound sheet; f) Guide roll

If the pressures built up in the roll nip are too high, the nip is automatically raised either by pressure plates with a preset breaking point or by pressure sensors that produce a corresponding signal. Mills can be operated almost automatically by memory-programmable sequence control. This clearly improves production safety, which is limited by the need for manual servicing.

Cracking and Refining Mills. In the rubber processing industry, so- called cracking and refining mills are used as well. *Cracking mills* with corrugated surfaces and high friction (1: 1.78, 1: 3.24) are used to break up waste, old rubber, and hard compounds. *Refiner mills* with high friction (1: 2.0, 1: 2.50) and a narrowly fixed roll nip are used mainly in processing reclaims.

2.2.3. *Compounding*

In the mixing process the material to be mixed is drawn into the roll nip. This requires sufficient static friction between the raw rubber and the surface of the two rolls, which rotate with friction. Because of this friction, different forces act on the compound being drawn through and continuously shift the particles toward one another. Compacting then occurs in the roll nip. If the rolls were to run without friction, the material would be compacted inefficiently. The different rates at which the material is drawn through also make the rubber head on the roll nip turn, and thus contribute to homogenization of the compound. The compound is finally homogenized by manual cutting and reintroduction of the sheet formed into the roll nip. Use of the stock blender aids this process (see Fig. 7). The excellent mixing action of the mill is related to the fact that very intensive cooling is possible due to the large surface area; thus, high shearing forces can always be introduced.

The *mixing procedure* for the production of a rubber compound on the mill is analogous to that in the conventional internal mixer. Mastication of natural rubber is effected either purely mechanically or by using peptizers. Rubber bales must be broken up in the cutter before being introduced onto the mill. The roll nip is made wider with increasing use of additives so that the necessary rolling material head is maintained until all compounding ingredients are mixed. Otherwise, an ineffective, “dead” head is formed, which does not move and cannot contribute to the mixing operation. Only when a closed rolled sheet passes around without tearing, is the roll nip contracted to a predetermined width. The compound is homogenized by manually cutting the sheet, rolling it out, and passing it through again; it is then cooled. Finally, the sheets are removed from the mill, dipped in release agent solution, and hung to dry.

2.2.4. *Prewarming*

If compounds are stored, they stiffen with time. For charge calenders or warm-feed extruders, and in certain cases cold-feed extruders also, the sheets must be plasticized again on the mill and warmed. A strip is then cut from the sheet turning on the front roll with a pair of circular cutters. This strip passes continuously from the mill into the extruder or calender. Compounds without accelerators are also stored and are then mixed with the cross-linking agent and vulcanization system during warming.

2.3. Pellet and Strip Feeding [40], [48]

Processing of ready-mixed *pellets* (rectangular with ca. 8-mm sides) allows automatic feeding of extrusion and injection molding machines (vibration channel, spiral vibrating conveyor, suction conveyor system, dosage via a balance). Well-prepared pellets pass uniformly into the feed screw of the extruder, so that the danger of pulsation is decreased significantly. In principle, several batches of pellets can be homogenized well in drum mixers. The correct dosage of the powdery release agent (talc) is important for processing. On overdosage the pellets are drawn into the screw inefficiently and smear, which can lead to pulsation of the extrudate. Too little powder promotes the tendency to bridge in the feed hopper, bringing about irregular feeding and pulsation of the extrudate. The talc should be dosed automatically, and excess shaken off or removed by air classifiers. The addition of external lubricants in liquid form can lead to trapped moisture, which causes porosity in the extrudate. Irregular feeding into the screw is also caused by differing pellet sizes. If pellets are produced by band granulators, the attachment of a sieving device is recommended. For very soft compounds the granulate has a strong tendency to bridge, while strips tend to tear. Granulate production requires additional investment, which makes the process more expensive than strip production.

On *strip feeding*, the dimension of the strip must be matched to the width of the drawing-in nip of the particular machine. Splicing leads to faults in strip feeding. Constant surveillance is necessary (e.g., danger of tearing, see above). Processing very hard compounds (>80 Shore A, ML4>100) as strips is virtually impossible. Strips of soft compounds also tend to stretch, leading to irregular extruder feeding.

The possibility of automation and the uniform operating process are advantages of pellet production.

2.4. Production of Sheeted and Reinforced Rubber [40], [49], [50]

2.4.1. *Calenders (see also* ⟶ Plastics, Processing – Rolling and Calendering*)*

Calenders are used to produce raw rubber sheets and for frictioning or coat reinforcing materials (fabric or steel cord). In special cases they are used for the production of profiles or profiled sheets (profile calender). The calender consists of two to four rolls, which are made either from chilled cast iron or from heat-treated steel. In modern machines the rolls are fitted with roller bearings in the calender frame (connecting links), bored on the periphery, and equipped with heating and cooling devices and with independent feedback control systems. The usual lengths and diameters of the rolls vary from 1250 to 2500 mm and 500 to 900 mm, respectively. The rolls are driven individually by d.c. motors, and continuous adjustment for different friction is possible. Since the roll nip must be adjusted precisely to maintain the exact thickness of sheeting or the raw compound covering of the fabric or steel cord, very accurate testing of the nip width must be carried out. Roll feelers located on the roll pins and connected to each other by a resistance bridge are used for this. Any change in distance results in a change in resistance, which is then recorded as an electrical signal at the control desk and visualized digitally there. Modern calender machines — including peripheral devices such as feed-roll-mill-extruders and doubling machines — are run by programmable control systems. Temperature, roll nip adjustment, friction, roll velocities, skew mounting (slope adjustment), and roll bending (see below) are controlled. Measuring devices continually test the thickness of the calendered sheet (examples are devices for measuring weight per unit area by transmission reflection, electronic path-measuring systems, laser probes, or pneumatic sensors). The measuring systems can be equipped with controlling devices that, on deviation from set points, send adjustment signals to the skew mounting or roll bending instruments. Calenders differ and are named according to their construction (see Fig. 8).

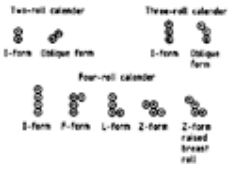

Figure 8. Types of calender construction

In processing rubber compounds, high pressures can arise, which result in bending of the rolls. This in turn gives inexact and varying thickness of the sheet or coating. To give a uniform thickness of the sheet or coating, the calender rolls can be cambered or crowned (i.e., shaped like a double cone, Fig. 9). The result is that the nip is straight, and the sheet thickness therefore uniform, when the rolls are under pressure. However, the roll geometry is then suitable only for particular operating parameters such as calender speed, sheet thickness, viscosity of the compound, and roll temperature. Fixed crowning severely limits the range of use of a calender and is therefore rarely used. The state of the art is represented by variable processes for *skew mounting* (slope adjustment) and *roll bending* (see Figs. 9 and 10) to compensate for the bending of the calender rolls under load. In the roll bending process a hydraulic pressure is exerted on the extended roll pins to compensate for the pressure caused by the compound in the roll nip. Crowning can thus be varied by these measures. The skew mounting can be ± 25 mm and the roll bending ± 0.15 mm.

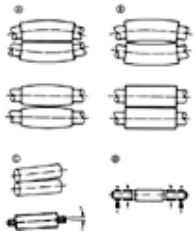

Figure 9. A) Crowned rolls (nip straight under pressure); B) Cylindrical rolls (nip broadening under pressure); C) Skew mounting; D) Roll bending

a) Supplementary bearing; b) Bearing; c) Hydraulic cylinder

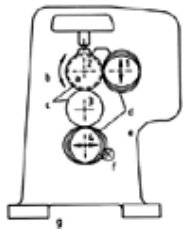

Figure 10. F calender with skew mounting and roll bending

a) Roll 2 with roll bending; b) Rolls driven via universal-joint propeller shafts (d.c. motors); c) Rolls 2 and 3 coupled by gear bair for uniform speed during coating on both sides; d) Rolls 1 and 4 can be skew mounted; e) Roll 4 (or 2) guiding roll for doubling machine; f) Pressure roll; g) Calender speed 1: 10 adjustable, friction (max.) in one nip 1: 2

Two-roll calenders are used predominantly in combination with extruders (extruder-roll-head, extruder-roll-die machine, see Section Sheet Production in the Extrusion Process) or as underrubber calenders (coating tread strips with rubber cushioning) and in the low-temperature coating of steel cord. In the rubber processing industry, *three-* and *four-roll calenders* are used most frequently; the preferred four-roll calenders are of F- and L-type construction.

To draw sheets, F calenders are frequently used, in which the upper roll nip is charged with material. The L calender, used for drawing sheets and also as a laboratory calender, has the advantage of favorable feeding possibilities because of the rolls situated underneath. Calenders can be charged with prewarmed rubber compound by using roll mills with strip drawing or extruders. Z calenders are often used in the tire industry. These calenders are relatively small; the roll bending affects only one roll nip, and all roll nips are readily accessible.

2.4.2. *Calendering Process*

Drawing Sheets — Doubling (Fig. 11). Calendered sheets should have smooth, shiny surfaces and be drawn without blisters. Correct roll thermostatting is very important for surface quality. If the temperature is too low, cold spots are formed on the calender surface; if it is too high, blisters can form. In addition, the roll head of the compounding material must be kept to a low diameter in the roll nip and must rotate in one direction to prevent blister formation. Choosing the correct roll friction is necessary here. During drawing of sheets a certain forced orientation of the rubber molecules (calender effect) occurs, which, after release of the drawing force, leads to shrinking. Shrinking can be controlled by fixing the velocity of the take-off conveyor, the cooling drum, and the take-up roll. Shrinkage can also be minimized by adjusting the compounding formulation (medium-active fillers; for natural rubber, sufficient mastication; use of pre- cross-linked rubbers). Production of thicker sheets by passing several rubber sheets over one another in the calendering operation is called doubling. Sheets are doubled up to 6 or 8 mm (see Fig. 11). Thicker sheets can be obtained by warm – warm doubling (combining two warm components), warm – cold doubling (doubling a compound sheet under a finished cold sheet) and cold – cold doubling (doubling an already finished compound sheet from a roll).

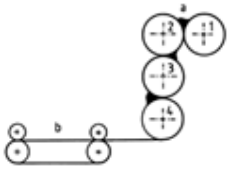

Figure 11. Drawing sheets and doubling

a) Rubber compound; b) Doubling band

Frictioning (Fig. 12). In frictioning, a stored fabric is run in the calender and rendered adhesive by mechanical rubbing with a soft rubber compound having good surface adhesion. The fabric is then ready for use as a reinforcement in rubber articles. The rubber compound is fed into the middle roll as a closed band. Roll 2 turns faster than roll 3 (see Fig. 12). The friction is 1: 1.2 – 1: 2.0. Synthetic fabric must be predried and impregnated. Fabrics can be frictioned either on one side or, via a second operation, on both sides. The load on the fabric is very high in the frictioning process. More frequently, therefore, the fabrics are dipped, particularly when the fabric structure is very dense. Frictioned or dipped fabrics are used as reinforcements in conveyor belt, V-belt, and tire technology.

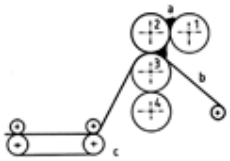

Figure 12. Frictioning of fabrics

a) Rubber compound; b) Fabric; c) Frictioning of fabric: $v2 > v3$; v = roll speed

Coating Fabrics and Steel Cord. Frictioned or dipped fabric or steel cord must be coated with a compound layer of defined thickness for further processing (e.g., in tire or conveyor belt production), so that further processing steps (i.e., construction with other compound sheets to give blanks) can be carried out accurately and rapidly. The reinforcement can be coated with rubber compound on one or both sides.

Warm Coating (Fig. 13). The warm compound is pressed onto the fabric or steel cord as it runs in. When coating endless cord, the threads of bobbin creels are drawn into the calender. Fabrics are fed out from a fabric storer. For coating on both sides, a four-roll calender must be used. Here, a second rubber sheet is drawn in the third roll nip (see Fig. 13 B, C).

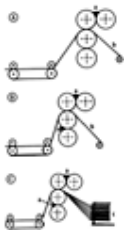

Figure 13. Warm coating — one and both sides (fabric – steel cord)

A) Warm coating on one side; B) Warm coating on both sides; C) Warm coating of steel cord on both sides

a) Rubber compound; b) Fabric; c) Steel cord

The thickness of the layer of the drawn sheets can be adjusted to exactly ± 0.1 mm by skew mounting and roll bending. This is extremely important, particularly for the production of high-value radial tires. Unlike the low-temperature coating process the machine can be operated at high velocities (with parts for passenger car tires, ca. 30 m/min; with parts for truck tires ca. 15 – 20 m/min). In addition, in the warm- coating process, good penetration (i.e., good flow of the coating compound around the threads of the cord of the reinforcement) is achieved. Good penetration contributes significantly to rubber – cord adhesion, in particular to steel cord adhesion in dynamically stressed articles such as tires.

Low-Temperature Coating (Fig. 14). The low-temperature coating process is cheaper than the warm- coating process with respect to the machinery involved (two-roll calender), but it is not as productive and has disadvantages in terms of penetration. Servicing the machine is simpler, since the tire cord is kept in the middle of the calender rolls by the preprocessed, uniform-thickness calendered rubber sheets. Low-temperature coating machines must be operated slowly (speed ca. 50 % less) to achieve sufficient coating of the cord.

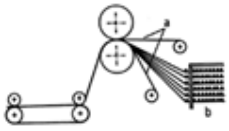

Figure 14. Low-temperature coating of steel cord on both sides

a) Rubber sheet; b) Steel cord

Skim Coating (Fig. 15). If fabric is to be coated or frictioned with relatively hard, dry compounds, working with uniform roll velocities has proved successful [v (roll 2) = v (roll 3)]. Skim coating can take place on one side or both by using a second operation.

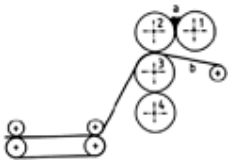

Figure 15. Skim coating with uniform roll speeds $v_2 = v_3$

a) Rubber compound; b) Fabric

Profiling (Fig. 16). For profiling three- or four-roll calenders are used, which have a roll with an exchangeable profiled cylinder. The applications include tire parts that are difficult to extrude, profiled sheets, bicycle tire segments, etc.

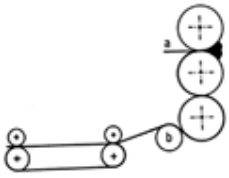

Figure 16. Profile roll calender

a) Rubber compound; b) Profile roll

2.4.3. *Sheet Production in the Extrusion Process*

Extruder-roll-head machines (Fig. 17) consist of combinations of extruders with a broad roll extrusion head and two-roll calenders. The rubber extruder can be designed as a warm- or cold-feed extruder. Modern extruder-roll-head machines are equipped with pin cylinder extruders. Rubber sheets can be extruded to a width of 2400 mm and a depth of 1 – 25 mm. Here, the rolls operate at the same velocity.

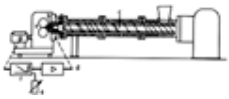

Figure 17. Extruder-roll-head machine

a) Pair of rolls; b) Broad extruder head; c) Extruder; d) Actual pressure; e) Pressure set point; f) Pressure control system

Extruder-roll-die machines are constructed very similarly to extruder-roll-head machines. These units, set up downstream from the internal mixer, with a differently constructed extrusion head, are used as shaping calenders for the rubber sheet.

The one-roll-head extruder is another type of construction for the production of rubber sheets. Here an extruder is combined with a roll that is flange-mounted at the head. Sheet widths up to 500 mm at 300 mm roll diameter can be drawn.

2.4.4. *Vulcanization of Rubber Sheets*

The vulcanization of the sheets, which are taken up from the calender in the form of rolls, can occur batchwise in heating chambers or autoclaves.

Continuous Vulcanization. For continuous vulcanization a rotation process is available. The rotation vulcanization machine (see Fig. 18), also known as rotocure (in German-speaking countries the name AUMA = Automatische Matten Maschine is used), consists of a drum that can be rotated and heated. The rubber sheet is pressed onto the drum with a steel band or a rubber- coated sliding band. A subsequent heating channel mounted on the machine stand can also be used. The discharge capacity is clearly increased by this measure.

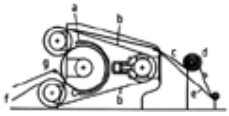

Figure 18. Rotation vulcanization machine

a) Additional heating; b) Steel band; c) Vulcanized rubber; d) Take-up reel; e) Fabric; f) Back cloth; g) Raw rubber sheet

The rotocure can be connected in series to an extruder-roll-head machine (completely continuous operation) or used behind a calender. Preprocessed articles, such as conveyor belts, are also vulcanized on the AUMA.

Batchwise Vulcanization. Large containers coated with rubber sheets that cannot be placed in autoclaves are subjected to

warm-water vulcanization. The coated containers are filled with water, which is brought to the boiling point by passing steam through it.

However, more often so- called self-vulcanizing compounds are used, particularly for vulcanization in large containers, such as ship containers or wastewater treatment towers. The vulcanization systems of these compounds are adjusted in such a way that they undergo complete vulcanization after processing on the walls of the container at ambient temperature or, after filling, at the temperature of the filling medium. For this process, compounds based on polychloroprene rubber with self-vulcanizing accelerator systems are widely used.

2.5. Extruded Articles [40], [51-55]

2.5.1. *Extruders (see also* ⟶ Plastics, Processing – Extrusion*)*

In the rubber processing industry, extrudates are produced by using *single-screw extruders*. In principle, extruders allow the production of storable intermediate products (pellets, strips), unvulcanized finished products, and further processing to give semifinished products. Special designs are used for the dehydration and drying of rubber after polymerization. An extruder essentially consists of an extrusion die, a screw, a cylinder, machine housing, motor, and feeding device including feed roll, plus heating, cooling, and lubricating systems. Technologically, the *screw and cylinder* form one unit. The screw draws in the compound to be extruded in association with the cylinder, compresses it, conveys it, and shapes it through the die. The thermostatted *feed roll* supports the uniform drawing-in of feed strips or bands. Modern extruders are equipped with measuring devices for bulk pressure and temperature, and with programmable control. This control allows set points to be determined for machine parameters such as the temperature profile of the heating and cooling system or the rate of the take-off conveyor, and possibly the extruder screw as well, and allows them to be compared with the actual value and regulated automatically in some cases. The weight, pressure, and temperature can also be determined by using sensors and displayed graphically.

The size of a rubber extruder is given by the screw diameter *D* (usual diameter 60 – 300 mm standardized) and the screw length *L* (which is identical to the length of the cylinder) as a multiple of the diameter. For example, the usual size of a cold-feed extruder is $L = 16\ D$ (i.e., L: $D = 16$). The extruder is charged either with blank strips or with pellets (see Section Profile Geometry Measurement). There are warm-feed extruders, cold-feed extruders, and cold-feed extruders with homogenizing action.

Warm-Feed Extruders (Fig. 19). The warm-feed extruder is much shorter than the cold-feed extruder. The screw lengths are 4 – 6 *D*. The compression ratio between the volume of the first screw flight in the filling zone and the last screw flight before the extrusion die is 1: 1.1 to 1: 1.4. Single- or double-helix, core- or flight-progressive screws are used. The driving power of such extruders is considerably lower than that of cold-feed extruders with the same screw diameter.

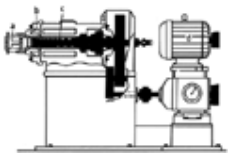

Figure 19. Construction of the warm-feed extruder

a) Extruder head; b) Heating and cooling channels; c) Screw; d) Motor

An open mill is set up before the warm-feed extruder to feed and warm the compound. The compound prewarmed to 85 – 100 °C is fed in strips from the mill into the charge hopper of the extruder. Since the prewarmed compound should be conveyed and compacted uniformly, extruders with short lengths can be used. Warm-feed extruders are used widely in the tire industry and are also used when a high discharge rate is required, e.g., as downstream machines to internal mixers or in combination with cold-feed pin cylinder extruders (duplex extruders) used in the tire industry. In the technical rubber goods industry they have been widely replaced by cold-feed extruders. In the tire industry they are increasingly being replaced by pin cylinder extruders.

Cold-Feed Extruders (Fig. 20). In cold-feed extruders the compound is not preplasticized in feed or prewarming mills. The extruders are considerably longer (screw length 16 – 24 *D*). The necessary plasticization of the rubber compound in the extruder is supported by exact temperature control in the cylinder wall and the screw during extrusion over the entire extruder length. The screws must be self- cleaning.

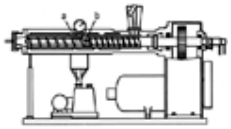

Figure 20. Cold-feed extruder

a) Vacuum zone; b) Dam

For production of industrial extrudates the cold-feed extruder is equipped with a vacuum pump connected to a vacuum zone in the extruder. The production of pore-free profiles is possible only with this device (vacuum of 0.2 – 0.4 kPA). This additional vacuum zone requires a total screw length of at least 16 *D*. In the vacuum zone the screw is cut deeper or wider, and the flight volumes are increased. While being transported through this zone the compound is decompressed, thus favoring degassing. The so- called dam is situated between the drawing-in and vacuum zones. The compound enters the vacuum zone as a thin film via the dam, so that a sealing effect is achieved at the same time. In the vacuum section the compound head rolls continually between the screw and the cylinder so that new surfaces are constantly being formed, which can be degassed more efficiently. As large a temperature difference as possible (ca. 10 °C) between the screw and the cylinder is also beneficial. Compounds with a high proportion of elastomer, which can be nervy, are more difficult to degas. To bind moisture in the compound and thus make degassing easier, calcium oxide is added. When feeding with raw compound pellets instead of strips, core-progressive screws are used (compression 1: 1.4). For extrusion with strip

feeding, flight-progressive screws are usual.

Cold-Feed Extruder with Homogenizing Action.
Pin Cylinder Extruder (Fig. 21). The cold-feed extruder with homogenizing action has largely replaced warm-feed extruders in the tire industry. The pin cylinder or pin extruder, also known as a cross- current mixing extruder, works on the principle of dividing and diverting the bulk flow.

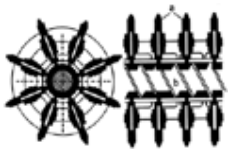

Figure 21. Principle of construction of pin cylinder extruder

a) Pins; b) Screw

In this type of extruder, pins, which project into the screw flight, are arranged in the cylinder. Through these pins the laminarly rotating rubber compound is divided into partial streams, which are recombined behind the pins. New surfaces are continually created, and intensive heat and oxygen exchange is achieved. The pins act at right angles to the bulk flow — hence, the name of the extruder. By dividing the flow, the formation of a cold core of material that occurs in a feed screw of usual construction is prevented. Mechanical shearing does not occur. Nevertheless, further mixing is achieved by the pin system.

The advantages lie in the considerably higher discharge compared to conventional cold-feed extruders, the smaller increase in bulk temperature, and the lower specific energy consumption. The reason for the higher discharge is the length and width of the flight of the screws used.

Extruders with Shearing Action. Some extruders work on the principle of shearing with flow division and recombination systems [Einwelliger Verteiler Kneter (EVK), Werner & Pfleiderer] or shearing with diversion of the bulk flow (Mailleffer) (⟶ Plastics, Processing). In these special constructions the cold-fed raw compound is plasticized at constant temperature. However, temperature control is extremely difficult, so premature cure onset of the rubber compound often occurs.

Transfer Mix (⟶ Mixing of Highly Viscous Media – Single Screw Mixers). The transfer mix extruder is also equipped with feed flights in the cylinder. The raw compound fed in is sheared and transferred continually from the screw to the cylinder and back to the screw.

Extruders with shearing action are still sometimes used as shaping extruders for internal mixers in mixing lines and as prewarming feed extruders for calenders. They are not important for the production of technical rubber goods.

Shearing Head Extruder (Fig. 22). The shearing head extruder is a machine part equipped with a rotatable core, situated between the extruder and the die. In principle the rubber compound fed in is sheared in a circular gap formed between the rotating core and the cylinder. Through shearing in the circular gap the compound is supplied with a large part of the energy necessary for vulcanization. Vulcanization is then completed in a downstream hot-air channel. The low die swell that can be achieved is an advantage. It results from the cross-linking that begins in the extruder head, the associated dimensional stability of the extruded blanks, and the shorter time and lower energy required for complete vulcanization, compared to hot-air and ultrahigh-frequency (UHF) vulcanization. The downstream hot-air section can be kept relatively short. The system is used for producing extrudates with small dimensions.

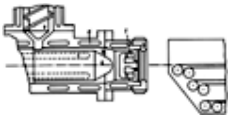

Figure 22. Principle of a shearing extruder heat

a) Extruder; b) Shearing gap; c) Core thermostatting; d) Jacket thermostatting; e) Rotating core head; f) Die; g) Union nut

2.5.2. *Vulcanization of Extrudates [56-64]*

To start the chemical cross-linking and vulcanization reaction in the blanks after the shaping process, and bring vulcanization to completion in the shortest possible time, an increased amount of energy must be supplied in the form of heat or, in special cases, radiation. Batch or continuous processes can be used. The heat required can be supplied by heat transfer from outside or through dissipative processes within the blank (UHF vulcanization, see Ultrahigh-Frequency Preheating — Hot-Air Vulcanization (UHF Vulcanization)).

The *batch processes* consist of free heating in autoclaves, vulcanization under lead in autoclaves, and autoclave vulcanization of wrapped articles (e.g., rollers). *Continuous processes* are divided into those with or without pressure. *Pressureless processes* include salt bath, hot-air, and fluidized-bed vulcanization, ultrahigh-frequency prewarming with hot-air vulcanization, and irradiation cross-linking processes. *Pressurized processes* include pressurized salt bath, steam tube, and hot-gas vulcanization.

2.5.2.1. *Batch Processes*

Free Heating in Autoclaves. The vulcanization of extruded blanks is carried out in longitudinal vulcanization vessels or in pots. The term *free heating* is derived from the introduction of free-lying blanks, which are stored in pans. Pots are sunk into the floor to make charging easier.

Pressurized vessels can be heated directly by introduction of steam or indirectly by using hot air. However, vulcanization in saturated steam is generally preferred for economic reasons. For hot-air vulcanization the vessel must be equipped with a

circulation ventilator to maintain a uniform temperature over its entire volume. Heating is carried out indirectly via a steam radiator, seldom electrically or with gas. The air in the vessel is pressurized; the pressure can be chosen independently of the vulcanization temperature. To avoid changes in the shape of the blanks before vulcanization, they can be placed in supporting frames of profiles. If they are rolled up in pans, the extrudates are embedded in talc to support them and prevent them from sticking together. Tubing is pulled onto mandrels to avoid deformation. Water-bath vulcanization can also take place in the vessel (production of hard rubber articles). In saturated steam vulcanization, spots of condensed water are formed, which give the vulcanizate an unattractive appearance and can lead to partial undercure on the surface. The condensing of water can be counteracted to some extent by treating the blanks with lubricants before introduction into the vessel. The vulcanization temperature for synthetic rubber is generally 140 – 200 °C; for natural rubber it is 120 – 170 °C because of the danger of reversion.

Vulcanization under Lead. Vulcanization under lead is used for tubing and cables. They are subjected to continuous extrusion coating with lead or batch coated with lead in a lead press. The lead- coated articles are cooled in a water bath, rolled onto large drums, placed in autoclaves, and vulcanized. After cooling, the lead jacket is split and the vulcanized cable or tubing is freed.

To give a uniform thickness of the sheet or coating, the calender rolls can be cambered or crowned (i.e., shaped like a double cone, Fig. 9). The result is that the nip is straight, and the sheet thickness therefore uniform, when the rolls are under pressure. However, the roll geometry is then suitable only for particular operating parameters such as calender speed, sheet thickness, viscosity of the compound, and roll temperature. Fixed crowning severely limits the range of use of a calender and is therefore rarely used.

2.5.2.2. *Pressure-Free Continuous Vulcanization*

The vulcanization section is located after the extruder in the production line. To produce pore-free profiles, vacuum extruders and moisture absorbers in the rubber compound (calcium oxide) must be used.

Hot-Air Vulcanization. Vulcanization of the extrudate takes place in hot-air channels or casing (in the case of coated fabric), whereby the blanks are transported on bands or conveyors. The channels are placed either in a line or more compactly in a three-story construction (length per story ca. 9 – 18 m). Hot air gives poor heat transfer, so relatively high temperatures are necessary, usually 200 – 300 °C. For silicone rubber, temperatures up to 800 °C can be used. Hot-air vulcanization is the oldest continuous process and an important constituent of the production of silicone profiles. Peroxidic cross-linking cannot be carried out by hot-air vulcanization. Such vulcanizates can only be produced continuously in salt baths or steam, as in cable technology. A special type of peroxide (bis-2,4-dichlorobenzoyl peroxide), used in silicone rubber technology, is an exception. This peroxide cross-links at relatively low temperature and can be applied only to the very soft silicone rubber extrudates (there is no heat buildup during extrusion and thus no premature cure onset). The poor heat transfer of hot air is improved considerably by high flow rates. Nitrogen is still used sometimes as a medium, so that peroxidic vulcanization is possible. By using shearing head extruder technology (see Shearing Head Extruder), rapid subsequent complete cure is possible even in relatively short hot-air channels through the homogeneous and intense heating of the entire bulk of the raw compound. The hot-air vulcanization channels can be heated electrically, by IR radiation, with gas, or by a steam-driven heating radiator.

Ultrahigh-Frequency Preheating — Hot-Air Vulcanization (UHF Vulcanization) (Fig. 23). UHF vulcanization is a continuous process in a hot-air channel with dielectric preheating of a polar blank in an electromagnetic alternating field. Machines in Western Europe work at 2450 MHz, in Eastern Europe at 2350 MHz, and in the United Kingdom and the United States at 915 MHz. Adequate preheating of the blank can be achieved only with a high loss factor or loss capacity. The more polar a polymer or compound is, the higher are the energy uptake and the associated loss factor (i.e., the heating effect in the UHF machine). Preheating can occur in hollow conductors or in resonator chamber systems. In the *resonator chamber* the energy fed in is distributed uniformly by reflection off the chamber walls. As a result of the broad energy distribution, large profile cross sections can be vulcanized (e.g., expansion joint bands up to 34 cm long). The *hollow conductor system* gives a very narrow energy distribution. The profile is passed through a stationary electromagnetic wave in the hollow conductor (high energy density, high penetration capacity; profile cross sections up to ca. 5×5 cm can be vulcanized). The hollow conductor system has been almost completely replaced by the resonator system. The resonator chamber (see Fig. 23) allows the use of blades with multiple bores and is more economical. The quantity of compound Q (in kg/h) that can be heated per unit time can be calculated as follows:

$$Q = \frac{P_m \cdot 860\eta}{\Delta T \cdot c}$$

where P_m is the magnetron power (kW); ΔT the temperature rise (°C; temperature of the profile behind the UHF section relative to the extrudate temperature); c the specific heat capacity of the compound (kcal kg^{-1} $°C^{-1}$); η the efficiency of the heating system (dependent on the machine and the polarity of the compound — can be 90 % or more); and 860 is the conversion factor (kcal/h) to kilogram per hour.

Figure 23. Continuous UHF vulcanization

a) Warm air feed; b) Magnetron power: 2.5 kW; c) HF power source; d) Resonator chamber system; e) Extruder

The length of the prewarming section is generally between 4 and 5 m. Hot-air channels, which are completely integrated, have lengths of 6 – 8 m. For UHF preheating, two or four magnetrons are used, for example, usually with a power of 2.5 or 6 kW per magnetron for resonator chambers. The magnetrons can be connected in series, thus adding to the power available. The feeding of electromagnetic energy into the vulcanization chamber is at right angles, either from above or from the side. This arrangement gives an optimally uniform energy field in the cross section. In addition, warm air is passed through the UHF channel in countercurrent, to achieve good and uniform heating throughout the entire chamber

and prevent contamination by condensed plasticizer vapors. This also ensures that the heat content of the compound is not lowered by loss through radiation emission (hot-air temperature ca. 250 °C).

A high efficiency is achieved if as little electromagnetic energy as possible is reflected from the profile thread. The modulation of the microwave position to achieve such a low-reflection adaptation is effected by the introduction of modulation pins into the energy input. In the production of metal-laminated profiles, a radiative energy loss due to the metal band is prevented by the rubber compound extruded onto it.

Fluidized-Bed Vulcanization. In fluidized-bed vulcanization, the extruded compound is passed through a fluidized bed of glass beads heated with an air – steam mixture. The beads adhering to the profile cross section are removed by a stream of air or by brushing. Compared to the salt bath, this process is more suitable for the vulcanization of complicated profile cross sections, but is has been successful to some extent only in the United Kingdom and the United States.

Liquid Curing Medium (LCM) Vulcanization (Fig. 24). Vulcanization takes place in a salt bath directly following the extruder. The extrudate is fed into a molten eutectic salt mixture consisting of 53 % potassium nitrate, 40 % sodium nitrite, and 7 % sodium nitrate. After vulcanization and before unrolling, the salt partially adhering to the vulcanizate surface is removed with brushes in a water bath and reprocessed in special recovery plants to avoid wastewater contamination. Eutectic mixtures of potassium nitrate and lithium nitrate have also been used, which do not contain nitrite and thus do not contribute to the formation of nitrosamines under vulcanization conditions. Compounds with peroxidic cross-linking systems can be vulcanized in the salt melt, in contrast to UHF and hot-air vulcanization. Heat penetration through thick-walled articles is not as good as in UHF vulcanization. Another disadvantage of LCM vulcanization is that the raw profile must be submerged in the salt melt using a steel band to counteract buoyancy. Articles with complex cross sections can be severely distorted. To avoid this, constructional aids are used, e.g., salt spray machines or specially constructed, long run-in sections before the steel band with salt melt sprinkling. Nevertheless, if possible, another vulcanization process should be used for complex articles. The maximum vulcanization temperatures of various types of rubbers in LCM vulcanization are listed in what follows:

NR	up to 210 °C, above reversion
SBR	up to 220 °C
Oil-extended SBR	up to 220 °C, above pore formation
NBR	up to 230 °C
CR	220 – 230 °C
CSM	up to 220 °C, above extensive pore formation
EPDM – sulfur vulcanization	230 – 240 °C
EPDM – peroxide vulcanization	250 °C
CM	≈220 °C
Epichlorohydrin – ethylene oxide terpolymer (ETER)	200 °C

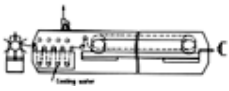

Figure 24. Liquid curing medium (LCM) machine

The use of eutectic metal alloys or organic liquids, such as glycerol, silicone oil, or poly(ethylene glycol), instead of the eutectic salt mixture, has not proved successful.

Radiation Cross-Linking (⟶ Radiation Chemistry). Polymers can be cross-linked through the absorption of radiation. γ-Radiation or β-radiation (electrons) can be used. The energy supplied by such radiation can also lead to depolymerization, as well as cross-linking. Generally, a dose of 50 – 300 Gy (5 – 30 mrd) is applied. Vulcanization systems and cross-linking agents are not necessary; neither is special heat treatment. The elastomer compound can be made up with fillers, plasticizers, and pigments. Activators are sometimes added. Cross-linking follows a radical mechanism. Complex radiation protection measures are required. Another disadvantage is the short penetration depth of β-rays. Also, uniform irradiation of complicated profile geometries is very difficult. On irradiation, ozone, nitrogen pentoxide, and nitrogen monoxide are formed in the airspace. The machines are shielded by concrete walls, whose strength depends on the accelerator voltage. Radionuclides (^{60}Co) can be used as a source of γ- and β-rays. β-Rays are produced mainly by electron accelerators (see also ⟶ Radiation Chemistry; ⟶).

Irradiation cross-linking has been introduced to the cable industry, to some extent into the cross-linking of sheeting (roof sheeting, coated fabric), and to the tire industry to increase the raw toughness of tire parts for further processing by pre-cross-linking with electron rays.

The penetration depth of the electron rays depends on the accelerator voltage and the density of the material. In cable technology, the coating layer thickness to be vulcanized is generally ca. 1 mm (500-kV accelerator).

2.5.2.3. *Continuous Vulcanization under Pressure*

Pressurized Liquid Continuous Vulcanization (PLCV). The PLCV process was developed in the United Kingdom for the vulcanization of cables and reinforced tubing. The molten eutectic salt mixture acting as a heat-transfer agent is passed through a closed system to which the vulcanization trough is connected. The salt melt is pressurized with air or nitrogen. The salt drawn out with the vulcanized cable is fed back into the circulation. A water cooling bath is located downstream of the vulcanization section. Despite some advantages of the system (e.g., production of cross-linked polyethylene with perfect surfaces, higher vulcanization temperatures, higher vulcanization rates compared to steam tube vulcanization, horizontal vulcanization of longer cables), the system has not proved widely successful. The reasons for this are the complex plant construction and the fundamentally difficult handling of molten salt mixtures.

Steam Tube Vulcanization. The continuous process of steam tube vulcanization is used in cable technology. The cable is extrusion coated with rubber compound in an angular extruder head and fed into a steam-filled tube. The tube is flanged directly to the extruder. The steam pressure is ca. 9 – 16 bar, at 180 – 200 °C. The length of the section is generally 50 – 80 m, followed by a cooling water section of 4 – 10 m. The steam chamber is bordered by the angular extruder head on one side and the cooling water section on the other. Cooling water is fed continuously by a high-powered pump. Steam tubes are set up in a rigid horizontal position or freely suspended horizontally (for vulcanization of up to 15-mm cable diameter). Vertically constructed machines are used for vulcanization of heavy duty cable (problem-free guide in high tower plants). Steam tube vulcanization is sometimes also used for the production of armored tubing.

Hot-Gas Vulcanization. Hot-gas vulcanization under pressure in a well-sealed tube is used for the production of hosing and cables. Nitrogen is used as the gas, so that both peroxidic and sulfur cross-linking are possible. An advantage of hot-gas compared to steam tube vulcanization is the fact that pressure and temperature are independent of each other, and the process can run at low pressure but high temperature. Products with a dense, very smooth surface result. In addition, so- called water trees, which can be formed in saturated steam vulcanization of cables by the diffusion of steam into microhollows in the insulation, causing voltage punctures, are not formed here. Since steam also has adverse effects on the steel cord adhesion in hydraulic tubing, the hot-gas technique can be employed to advantage there also.

2.5.3. *Profile Geometry Measurement [54], [65], [66]*

In the continuous production of extrudates the required dimensional tolerance must be adhered to. For complicated profile constructions, monitoring is predominantly carried out destructively in batches, but for simpler profile geometries, continuous, nondestructive testing can be carried out.

The *batch measurement* is still the safest method for establishing the dimensional tolerances of complicated profile cross sections exactly. A thin profile cross section is cut, enlarged, and projected onto a standard drawing by using transmitted light. However, the disadvantage of this process is that faults occurring in continuous production are recognized very late, and articles produced in the meantime must be rejected.

The trend is now increasingly in the direction of *continuous methods of measurement*, which immediately adjust the profile dimension to the given standards by feedback, for example, to the band removal rate. Continuous measuring processes can operate mechanically by scanning rolls, determination of pressure variation by passage through a measuring cell, and determination of weight variations by an inserted band weigher, or optically, by laser technology. Measurement and control by *scanner rolls* are used successfully on rotationally symmetrical profiles, such as cables and hosing. The measurement of *air pressure variations* takes place by passage through a measuring cell, which is heated to a given temperature and has an exactly adjusted internal pressure. A variation in profile thickness is determined from the resulting variation in air pressure. This system is sometimes used for measuring small profile cross sections. Continuous measurement of *weight* to detect variations in profile dimensions is used in the tire industry during tread extrusion. Continuous measurement of the profile geometry by low-powered, *monochromatic lasers* is being increasingly used. Measurement takes place very rapidly without contact. For example, the diameter of round profiles is scanned by high-frequency laser scanners, and a square wave signal is produced at a receiver with the start at t_1 and the end at t_2 (see Fig. 25). The time between t_1 and t_2 corresponds to the diameter of the scanned profile. With the laser scanner method, the profile sag can be measured and controlled immediately on exit from the die before entry into the vulcanization section. To determine complicated profile cross sections the laser head can be rotated around the profile, or two axial measuring heads can be used. The thickness of sheeted or band-shaped profiles can be determined by *laser reflection sensors*. A luminous spot is produced on the profile surface and reflected to the receiving optical instrument. The position of the luminous spot changes when the thickness of the measured object varies. Laser measurement allows the construction of a control system whereby the sensor is set up at the extruder head and usually regulates the take-off conveyor speed. In addition, a second laser measuring head can be connected at the exit of the vulcanization machine.

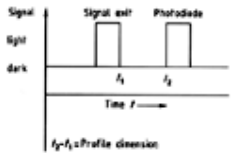

Figure 25. Determination of profile dimension from the measured signal

2.6. Molded Articles [40], [61], [67-72], [78]

Molded articles are produced by compression, transfer, and injection molding, and in special cases, by embossed molding. The raw compound placed or injected into the mold is vulcanized after the mold has been filled.

2.6.1. *Production Process [67], [73]*

2.6.1.1. *Molding Process*

Compression Molding (Fig. 26). The preformed compound blank is placed in the mold and vulcanized after the press is lowered.

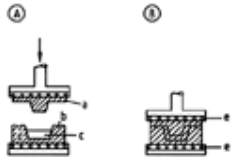

Figure 26. Compression molding process

A) Mold open; B) Mold closed

a) Upper part of mold; b) Flash groove; c) Blank; d) Lower part of mold; e) Heating

Generally, blanks are preformed by a special extruding press so that the same constant blank measurement always results. The blanks can also be preheated (high-frequency preheating 20 – 27 MHz). This preheating guarantees uniform vulcanization and shortened complete cure time. Excess compound (flash) is pressed into the parting lines and flash grooves of the mold. After vulcanization the articles must be deflashed (see Section Deflashing).

Transfer Molding (Fig. 27). The rubber compound is placed in the upper part of the mold. When the mold is closed the compound is pressed through channels into the cavities. Compared to compression molding, higher injection pressures must be used. In addition, more vulcanized waste is produced in this process through the formation of a flat cake in the loading chamber with runners into the injection channels. The blanks must be preformed before being placed in the mold. In transfer molding the rubber compound is plasticized to a certain extent while being injected into the mold cavities. The transfer molding process is useful for the production of complicated small articles. Usually the mold consists of three parts. The middle part can be removed. The upper part is constructed as a piston.

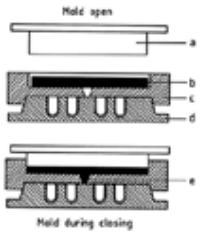

Figure 27. Transfer molding process

a) Piston; b) Compound; c) Middle mold section; d) Lower mold section and caliber; e) Injection nozzle

The *middle part* contains the cylinder, which takes up the rubber compound to be molded, and the injection nozzles. The *lower part* is the actual mold. When the compound is forced through the narrow flow channels, a high heat of friction is produced. This must be considered in compounding since the additional heat can lead to scorching of the compound. Overall, however, a more rapid and homogeneous warming of the total bulk of the blank is achieved compared to compression molding, and vulcanization times can be shortened considerably. Trapped air is pumped out between the parting lines of the mold. Less flash occurs in the parting lines than during compression molding, which also makes degassing considerably easier.

Flashless Process. A modification of the transfer molding process allows the production of molded articles requiring minimum flash. The material is inserted into a cylinder equipped with sealing lips, and pressed through runners to overflow into mold cavities that are mounted floating. One or more runners are assigned and run conically into each cavity. The quantity of raw material must be such that after all cavities have been filled, the formation of a rubber cushion is still possible. This acts as a hydraulic cushion and provides a uniform closing pressure for all cavities. The vulcanizing cushion must subsequently be discarded as waste. When the mold is opened and the cushion is removed, excess material is torn from the molded article at the conical entries to the cavities. The flashless process requires very stable platen presses with well- controlled, plane-parallel heating plates.

Presses. Compression and transfer molding processes are carried out in presses into which the molds are built or inserted. They include all platen presses, the boiler press, and occasionally the articulated lever press.

Platen Press (Fig. 28). Platen presses can be divided into hydraulic column and frame presses, the latter being more recent. Column processes, however, have the advantage of plane-parallel control of the press plates. Platen presses have several heating plates between which the vulcanization molds are inserted. Heating is effected mainly by steam, but also electrically or with oil.

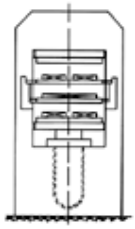

Figure 28. Hydraulic platen press

For the vulcanization of industrial molded articles, the frame press has largely replaced the column press. The number of platens is limited by the manipulation time and the height of the press. To avoid possible heat loss, the vulcanization presses must be jacketed. If the mold is inserted into the platen, the latter is closed with reduced pressure. A surface pressure, normally 35 – 100 bar, is exerted on the mold (high pressure). Such pressures are necessary to achieve uniform pressing of the rubber compound. Articles of large volume must be returned to low pressure (15 – 40 bar) after 10 – 20 min to avoid backrinding (on mold parting lines).

Boiler Press. A vulcanization press with column construction is surrounded by a boiler that can be heated directly. The machine is used if very tall molds need to be heated (e.g., in the production of tall buffers). A range of processes can be found for special applications, such as shoe, conveyor belt, V-belt, and tire presses, as well as jaw-type presses that open on three sides for the production of sheets, mats, or floor coverings.

2.6.1.2. *Injection Molding (see also* ⟶ Plastics, Processing – Injection Molding*)*

The injection molding process is one of the most important processes in rubber technology for producing molded articles. It is suitable both for precision articles and for articles with complicated shapes. The two types of injection molding are screw injection molding and ram injection molding with screw preplasticization. Construction can be horizontal or vertical. With vertical machines, injection can take place from above, below, or the side. With horizontal machines, injection occurs in the direction of the mold. Horizontal systems are predominantly designed as screw injection molding machines, while systems with screw preplasticization and ram injection can be constructed both horizontally and vertically.

Ram Injection Molding Machine with Screw Preplasticization (Fig. 29). The compound has to be redirected on its way from the screw preplasticization into the mold via the injection ram, which is a problem. It contradicts the first-in – first-out principle (i.e., the material that is plasticized first should, for temperature reasons, also be injected first). Specific injection pressures are, for example, between 800 and 1600 bar. Injection volumes of 6400 cm^3 can be realized easily. Construction with *injection from below* (see Fig. 30) is readily accessible. The compound is introduced into the drawing-in shaft via the shortest route. With soft compounds, the fed-in strips are in no danger of tearing.

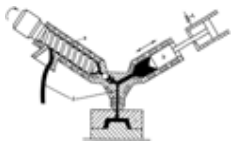

Figure 29. Ram injection molding with screw preplasticization

a) Plasticizing screw; b) Ram; c) Mold; d) Rubber compound

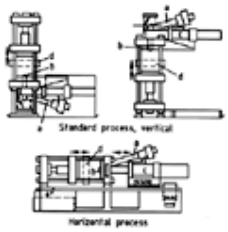

Figure 30. Types of injection molding machine construction

a) Plasticizing screw; b) Injection; c) Ram; d) Mold

Injection from above is advantageous for fixing metal inserts in the production of rubber – metal articles. The metal inserts are placed at the bottom of the mold and are fixed by the vertical pressure during injection of the rubber compound. *Horizontal construction* is characterized by the short height of the entire machine. In addition, the automatic release in fully automated operations is made considerably easier, as are removal and cleaning of the mold. For the production of rubber – metal articles, it is, however, less suitable. A different machine construction allows injection into different parting lines by using a height-adjustable injection unit (injection from the middle, above, or below). This allows a more flexible construction of the mold since the position of the parting line is not fixed, but can be adjusted to suit the molding (e.g., to give better demolding). Thus, various types of mold can be used in a single molding machine.

The plasticized raw compound charged to the injection cylinder is kept at ca. 80 – 110 °C (max.) to avoid too high a temperature of the reactive compound during its residence time in the injection cylinder. Under all circumstances, the cure onset may begin only when all cavities are definitely filled after the injection process has finished.

Screw Injection Molding Machine (Forcing Screw) (Fig. 31). An axially movable screw acts as the ram and the plasticizing unit. A backflow blocker is built into the head of the screw to hinder backflow during injection. To achieve uniform plasticization over the entire shot volume, the axial shift of the screw should not exceed 4 *D* screw length. The temperature of the compound in front of the screw can be 120 °C; after passage through the nozzle, an increase to 140 °C is possible. This type of machine is preferably of horizontal construction and allows shot volumes of ca. 400 cm^3. The shot volume is limited by the cylinder diameter of the screw. The forced screw process works on the first-in – first-out principle and is also used to operate multistation units arranged on circular tables.

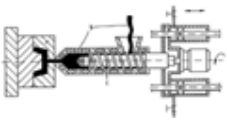

Figure 31. Screw injection molding machine (forcing screw system)

a) Mold; b) Rubber compound; c) Plasticizing screw

A variation is the *screw injection molding machine with storage*. Here, the screw cylinder, not the screw, acts as the ram for injection. For plasticization, the full length of the screw is available over the entire plasticization time. The screw cylinder moves backward or downward with the screw, depending on construction. It is used relatively rarely. These machines are available in horizontal and vertical construction.

In principle, short working cycles can be achieved by using the injection molding process with short vulcanization times, which are very clearly determined by the entry temperature of the raw compound into the mold cavities. The vulcanization systems used must have long flow times and short complete cure times, so that after filling has finished, vulcanization

starts immediately and proceeds completely. The rubber compound is charged as strips or pellets.

The advantages and disadvantages of various injection molding, machines are summarized in Table 1.

Table 1. Advantages and disadvantages of various injection molding machines

Characteristic	Ram injection machine with screw preplasticization (vertical, horizontal)	Screw injection molding machine (forcing screw)	Screw injection molding machine with storage (vertical, horizontal)
Bulk temperature distribution, plasticization	advantage: uniform	disadvantage: nonuniform at axial displacement of the screw >4 *D*	advantage: uniform
Production of articles of large volume	advantage: short screw, short cycle times for large volumes	disadvantage: limited by cylinder diameter	disadvantage: accuracy of dosage difficult to adjust
Fixing of metal inserted part	advantage vertical version: fixing not complicated	disadvantage: fixing difficult	disadvantage: fixing difficult
Rotary bulk article release from mold	disadvantage: only horizontal version can be used, also with automatic release from mold	advantage: automation easily possible, good article release from mold	advantage: automation easily possible, good article release from mold
Corrosion of plasticization screw head	advantage: low danger of corrosion	disadvantage: high danger of corrosion because of additional shear stress (screw injection ram)	advantage: low danger of corrosion because screw cylinder serves as injection ram
Bulk redirection	disadvantage: first in last out	advantage: no bulk redirection, first in first out	advantage: no bulk redirection, first in first out
Compound changing	disadvantage: complex cleaning	advantage: simple cleaning	disadvantage: cleaning relatively difficult, sealing of storage against backflow of material difficult

Machine Control — Process Data Determination [71], [72]. With the aid of sensors built into various operational points of injection molding machines, data can be collected, transmitted, and processed in a computer. Through computerized data collection the evaluation of measured data has become much simpler. Cycles can be compared, and graphical representations displayed immediately making the interpretation of measured data considerably easier. Also, statistical methods of evaluation and documentation of measured values, as demanded by car producers, are possible. In the injection molding machine, important process data such as cylinder temperature, hydraulic pressure, ram distance, bulk temperature, nozzle temperature, mold temperature, inner mold pressure, injection time, subsequent pressure time, dosage time, heating time, and mold opening time can be measured and evaluated by the corresponding built-in sensors. Injection time, subsequent pressure time, dosage time, heating time, and mold opening time are set directly through the machine control. The measured values can be displayed on-line on a screen. Deviations from given tolerances are communicated to the operator by an alarm signal. Upon deviation of actual values from the set point, the mold temperature and heating time are automatically adjusted, whereby the heating time is varied by changing the mold and the bulk temperatures. The injection rate can be controlled by hydraulic pressure.

The heating time determines the progress of the vulcanization, which is given by the stored vulcameter values of the compound. The subsequent pressure, which is an important parameter for the production of articles with minimized flash formation, can also be regulated by adjusting the hydraulic pressure. If strongly undercut articles are produced, trapped air, or burned points as a result of local overheating due to trapped atmospheric oxygen, are a danger. Here, deaeration programs or, preferably, the application of a vacuum (vacuum pressure ca. 5 kPa) can be used to remedy the situation. Evacuation of molds leads to a clear improvement in product quality.

Special Processes. The *injection stamping process* is used for the production of precision articles with extremely low plane-parallel measurement tolerances. In addition, these articles can exhibit only very low flash formation and no formation of knitlines. To carry out the process, a rubber compound, which is predosed exactly, is injected into an injection mold with an opened gap, whose diameter is defined exactly. At the end of the injection process the mold is closed to an equally precisely defined final gap, and the very fine flash formed is cut off (injection – compression process). Exact regulation of the residual gap is very important to maintain the dimensions of the molded articles in the direction of the injection axis. Plane-parallelisms up to ± 0.01 mm can be controlled on the article. To support the plane-parallelism, particularly in multicaliber molds, cold channel molds, which are in themselves more rigid with regard to deformation, are used.

Cold Channel Mold (Fig. 32). In injection molding the vulcanized rubber compound located in the distributor spinner and the feed orifice usually have to be discarded as waste. Particularly in the use of very expensive rubber, this can become extremely uneconomical. The cold channel mold has therefore been developed. The flow channel area is thermostatted with warm water (ca. 90 °C) at a low temperature and well insulated, so that onset of cure of the compound contained in the flow channels is prevented. This part of the compound can thus be reused. Sufficient thermal insulation must be applied particularly between the channel region and the highly heated mold. The injection nozzle cannot have direct contact with the hot mold. The construction requires a great deal of experience and precision.

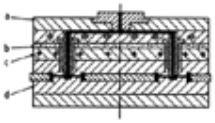

Figure 32. Cold channel mold

a) Cooled cold channel plate; b) Insulating plate; c) Heating plate; d) Mold

2.6.2. *Mold Cleaning — Contamination [74], [75]*

Contamination of molds represents a serious problem, since it considerably shortens their service life. Contamination can, for example, be caused by high vulcanization temperatures or by the escape of volatile components from the compound, so- called rubber plate-out. Compound residues, which gradually decompose, or residues from mold-release agents with organic components can also be left in the mold. Molds can be cleaned mechanically, physically, chemically, electrochemically, or thermally. Processes are also combined. The different cleaning processes are listed below:

Mechanical	brushing, rubbing, abrasive blasting (lapping), cleaning compounds
Physical	ultrasound
Thermal	fluidized bed, flame treatment, inductive prewarming
Chemical	acid or base treatment
Electrochemical	oxidizing agents, electrophoresis

In the *abrasive blasting process* the mold is cleaned by spraying an aqueous dispersion of abrasive materials (very small glass beads, nutshell chips, sawdust – quartz powder). Abrasive blasting can be combined with a precleaning process in a chromic – sulfuric acid mixture. The latter can also be used alone. Cleaning can be carried out in alkaline baths with and without oil spray support; in solvent baths, or in an alkali bath with ultrasound support (frequency 20 kHz) as well. Molds must be dried sufficiently and protected with washable corrosion protection oil when stored to avoid corrosion or rust.

2.6.3. *Deflashing [75-77]*

A process that allows the production of absolutely flash-free articles is technically possible in principle, but generally not practicable from an economic point of view. The removal of flash, even when the quantity is very small, is therefore necessary and frequently leads to increased rejects because articles can be damaged in the different trimming processes. The flash and burrs can be removed manually by brushes or by cutters or simply by separation at tearing edges that have been provided beforehand. A very important technique is *breaking off the flash in the frozen state* in rotating drums. Freezing is carried out by using dry ice or liquid nitrogen. Temperatures of – 79 °C can be achieved with dry ice, and – 196 °C with liquid nitrogen. Drum trimming is not suitable for articles with internal burrs, but is used frequently for rubber articles with strong walls (trimming by freezing, with steel balls, chains, or beating arms). The investment costs for drum trimming are relatively low. In another technique the article is hit with kinetically accelerated steel or plastic particles (*abrasive blasting*). Abrasive blasting (see Fig. 33) is technically more flexible; the shape of the machine depends on the size of the molded article.

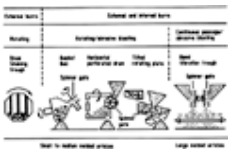

Figure 33. Abrasive blasting process

2.7. Solutions of Rubber Compounds [40], [78]

Solutions of rubber compounds are usually produced in two-shaft trough mixers with a trough that can be heated and cooled. The solid rubber compound previously produced in the internal mixer is charged to the trough mixer together with the solvent and dissolved with the aid of the kneading shaft. The solvent can be gasoline, esters, aromatic hydrocarbons, or a combination of different solvents. A dissolving process in a 70-L trough mixer takes 24 h. The solutions produced are further processed in coating and dough processes or used as adhesive solutions (with a natural rubber base or, for hard rubber modifications, a polychloroprene rubber base). So- called solution dipping for the production of thin-walled articles has been almost completely replaced by corresponding latex technology processes (see Section Dipped Goods).

Coating Processes — Doughs. In the *coating processes,* rubber solution is spread over the impregnated fabric by using special spreading machines with doctor blades, which also regulate the strength of the coating (see Fig. 34).

Figure 34. Spreading machine

a) Machine stand; b) Take-off roll for back cloth; c) Take-up for coated fabric; d) Take-off for uncoated fabric;

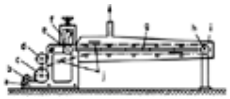

e) Spreading roll; f) Spreading knife; g) Hot plate; h) Reversing roll; i) Casing; j) Fabric sheeting

Drying takes place over hot tables or in drying towers. Vulcanization is carried out after the last coating either in hot air, on the rotocure (see Section Vulcanization of Rubber Sheets), or by radiation cross-linking. The solvent can be recovered by a condensation process or by activated charcoal adsorption with subsequent distillation. Awning cloths, cloak and poncho materials, etc., can be rubberized by this process.

In the *dough process*, highly viscous solutions of rubber compounds are applied to very dense textile fabrics. Unlike the coating process, the compounds are applied by a pair of rolls that work at the same velocity, but in opposite direction (coating calender). After application is finished, the fabric is dried. The pretreated fabric is then coated with rubber compound on the calender (e.g., finishing of conveyor belts) (see Fig. 13).

[Top of Page]

References

1. R. J. Noble: *Latex in Industry*, Rubber Age, New York 1953, pp. 59 – 60.
2. K. O. Calvert: *Polymer Latices and their Applications*, Applied Science Publ., London 1982, p. 19.
3. D. C. Blackley: *High Polymer Latices*, Maclaren & Sons, London/Palmerton Publishing Co., New York 1966, pp. 383 – 401.
4. D. C. Blackley in [3] pp. 437 – 522.
5. ISO 706; DIN 53 594.
6. ISO 124; DIN 53 563.
7. ISO 126; DIN 53 564.
8. ISO 976; DIN 53 606.
9. ISO 705; DIN 53 597.
10. ISO 1652; DIN 53 211.
11. ISO 35; DIN 53 567.
12. ISO 289; DIN 53 523.
13. ISO 868; DIN 53 505.
14. ISO 37; DIN 53 504.
15. ISO 34; DIN 53 515.
16. G. Sinn: “Heat Sensitizing Agents for Natural and Synthetic Latices,” *Polym. Paint. Colour J.*, Proceedings-Hastings Symposium, May 14/15 (1974).
17. Bayer, DE 2 756 770, 1977 (H. Perrey, M. Matner).
18. Bayer, DE 2 756 774, 1977 (H. Perrey, M. Matner).
19. Sandoz, DE 2 837 314, 1978 (H. Gerber).
20. Bayer, DE 3 330 197, 1984 (H. Perrey, M. Matner).
21. in [2], pp. 173 – 205.
22. in [1] , pp. 538 – 549.
23. in [2] , pp. 229 – 236.
24. in [2], pp. 239 – 242.
25. E. W. Madge: *Latex Foam Rubber*, Maclaren & Sons, London, Interscience Publishers, New York 1962.
26. DuPont, DE 2 029 990, 1970 (J. W. Crary).
27. in [1] , pp. 681 – 684.
28. W. R. Grace, US 4 160 752, 1976 (R. H. Akiyama, L. B. Towle).
29. in [2] , pp. 119 – 143.
30. H. Esser: “Baypren-Latices und ihre industrielle Anwendung,” *Gummi Asbest Kunstst.* **14** (1973) no. 7, 574 – 582.
31. in [2], pp. 145 – 171.
32. in [1] , pp. 693, 717.
33. G. Wallpott: “Aufbau und Auftragsverfahren von Papierstreichmassen. Einführung und überblick,” *I Bayer Farben Rev.* **22** (1972) 70 – 85; *II Bayer Farben Rev.* **23** (1973) 66 – 91.
34. in [2], pp. 71 – 91.
35. W. von Langenthal: “Latices – Ihr Einsatz in der Schuhindustrie,” *Kautsch. Gummi Kunstst.* **40** (1987) no. 9, 847 – 858.
36. *Ullmann*, 4th ed., **16**, 159 – 166.
37. *Ullmann*, 4th ed., **22**, 625 – 630.
38. M. Renner, W. von Langenthal: “Ausrüstung von Textilfilamenten für verbesserte Gummi-Gewebe-Haftung,” *Gummi Fasern Kunstst.* **44** (1991) no. 12, 664 – 670.
39. J. Lünenschloß, W. Albrecht: *Vliesstoffe*, Thieme Verlag, Stuttgart 1982.

General Reference

40. J. Schnetger: *Lexikon der Kautschuktechnik*, Hüthig Verlag, Heidelberg 1992.

Specific Reference

41. J. P. Lehnen: *Kautschukverarbeitung*, Vogel, Würzburg 1983, pp. 14 – 24.
42. VDI-Gesellschaft Kunststofftechnik (ed.): *Der Mischbetrieb in der Gummiindustrie*, VDI-Verlag, Düsseldorf 1984.
43. P. K. Freakley: *Rubber Processing and Production Organization*, Plenum Press, New York 1985.
44. F. Sommer, H. Hokner in VDI-Gesellschaft Kunststofftechnik (ed.): *Der rechnergestützte Mischbetrieb*, VDI-Verlag, Düsseldorf 1989, pp. 21 – 47.
45. H. M. Schmid: "Produktionssteigerung und Wirtschaftlichkeitserhöhung von Mischerlinien," *Kautsch. Gummi Kunstst.* **42** (1989) no. 1, pp. 49 – 53.
46. A. Limper, P. Barth, F. Grajewski: *Technologie der Kautschukverarbeitung*, Hanser Verlag, München 1989, p. 21 ff.
47. Werner & Pfleiderer, EP 0 272 338 B 1, 1988 (K. Ruthenberg, F. Müller).
48. P. Kristukat: "Beschicken von Extrudern mit Streifen oder Granulat," in Verein Deutscher Ingenieure (eds.): *Extrudieren von Elastomeren,* VDI-Verlag, Düsseldorf 1986, pp. 43 – 51.
49. H. Kopsch: *Kalandertechnik*, Hanser Verlag, München 1978.
50. A. W. M. Coaker: "Calendering," in H. F. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 2, John Wiley & Sons, New York 1985, pp. 606 – 672.
51. E. Harms: *Kautschukextruder, Aufbau und Einsatz aus verfahrenstechnischer Sicht,* vol. 2: "Kunststoff Technologie," Krauskopf, Mainz 1974.
52. S. Meißner *et al.*, *Technologische Verfahren zur Elastverarbeitung*, Deutscher Verlag für Grundstoffindustrie, Leipzig 1983, p. 252.
53. J. P. Lehnen: "Gegenwärtiger Stand und Zukunftsperspektiven der Kautschukextrusion bedeutender Gummiartikel," *Kautsch. Gummi Kunstst.* **37** (1984) no. 1, pp. 25 – 31.
54. Verein Deutscher Ingenieure (eds.): *Extrudieren von Elastomeren*, VDI-Verlag, Düsseldorf 1986, p. 186 ff.
55. G. Capelle: "Extrusion und Vulkanisation an Profilen," *Kautsch. Gummi Kunstst.* **40** (1987) no. 11, pp. 1058 – 1066.
56. J. Ippen: "Formulation for Continuous Vulcanization on Microwave Systems," *Rubber Chem. Technol.* **44** (1971) no. 1, pp. 294 – 306.
57. D. Pollard: "Cables," in C. M. Blow, C. Hepburn (eds.): *Rubber Technology and Manufacture,* 2nd ed., Butterworth, London 1982, pp. 501 – 502.
58. C. W. Evans: *Developments in Rubber and Rubber Composites 2*, Appl. Sci. Publ., Essex, England 1983, p. 16 ff.
59. U. Schmidt, J. Stimpel: "Abluftreinigung bei der kontinuierlichen Vulkanisation," in Verein Deut-scher Ingenieure (eds.): Extrudieren von Elastomeren, VDI-Verlag, Düsseldorf 1986, pp. 245 – 267.
60. T. L. Wilson: "Dielectric Heating," in H. F. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 5, John Wiley & Sons, New York 1986, pp. 1 – 23.
61. V. D. McGinnes: "Crosslinking with Radiation," in H. F. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 4, John Wiley & Sons, New York 1986, pp. 418 – 449.
62. Ingenieurbüro Hager und Ladwein: "Abgasreinigung von UHF- und Heißluft-Vulkanisationsanlagen," *Kautsch. Gummi Kunstst.* **41** (1988) no. 8, pp. 823 – 824.
63. R. Franho: "Continuous Vulcanization: New Technology Improves Endproducts," *Elastomerics* **121** (1989) no. 1, pp. 30 – 32.
64. H.-H. Spendlin: "Methoden der Abluft- und Abwasserreinigung bei Vulkanisationsanlagen," *Kautsch. Gummi Kunstst.* **45** (1992) no. 1, pp. 32 – 35.
65. M. von Gunten: "Berührungslos abtasten, Lasersysteme überwachen, Produktdimensionen beim Extrudieren von elastomeren Kunststoffen," *Maschinenmarkt/MM-Industriejournal* **93** (1987) no. 13, pp. 58 – 62.
66. O. Marchon: "Automatisches Messen auf Profilprojektoren," *Qual. Zuverlässigkeit* **33** (1988) no. 10, pp. 541 – 543.
67. W. Seyderheim, J. Frey: "Zusammenfassende Darstellung der Rohlingsvorbereitung in der Gummi-Industrie," *Kautsch. Gummi Kunstst.* **28** (1975) no. 6, pp. 335 – 340.
68. Verein Deutscher Ingenieure (eds.): *Spritzgießen technischer Gummiformteile*, VDI-Verlag, Düsseldorf 1981.
69. Verein Deutscher Ingenieure (eds.): *Spritzgießen*, VDI-Verlag, Düsseldorf 1983.
70. Verein Deutscher Ingenieure (eds.): *Von der Kautschukmischung zum Formteil*, VDI-Verlag, Düsseldorf 1987.
71. T. Krehwinkel, C. Schneider: "Verarbeitungsfenster für den Elastomerspritzgießprozeß," *Kautsch. Gummi Kunstst.* **41** (1988) no. 2, pp. 164 – 168.
72. P. Barth, G. Fischbach, L. Setiawan: "Entwicklungsschritte zum automatisierten Spritzgießen von vernetzten Werkstoffen," *Kautsch. Gummi Kunstst.* **41** (1988) no. 8, pp. 800 – 815.
73. Einbrodt, Jörg: "Formpressverfahren," in: *Technologie der Plast- und Elastverarbeitung,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1989, pp. 137 – 146.
74. T. Barth: "Werkzeugverschmutzung und Möglichkeiten zu deren Behebung," *Kautsch. Gummi Kunstst.* **41** (1988) no. 10, pp. 1003 – 1008.
75. Verein Deutscher Ingenieure (eds.): *Das Elastomerwerkzeug*, VDI-Verlag, Düsseldorf 1990.
76. Verein Deutscher Ingenieure (eds.): *Spritzgießen von Elastomeren*, VDI-Verlag, Düsseldorf 1978.
77. G. Köhler: "Automatisationseinsatz bei der Vulkanisation von Gummiformteilen," in Verein Deutscher Ingenieure (eds.): *Kostensenkungspotentiale in der Gummiindustrie,* VDI-Verlag, Düsseldorf 1993, pp. 155 – 192.
78. F. N. Teumac: "Fabrics, Coated," in H. F. Mark et al. (eds.): *Encyclopedia of Polymer Science and Engineering,* 2nd ed., vol. 6, John Wiley & Sons, New York 1986, pp. 632 – 645.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Rubber, 6. Testing

RÜDIGER KOOPMANN, Bayer AG, Leverkusen, Federal Republic of Germany

JOSEF NEU, Bayer AG, Leverkusen, Federal Republic of Germany

DIETER BRÜCK, Bayer AG, Dormagen, Federal Republic of Germany

1. Purpose, Scope, and Limitations of Testing . . . 1
2. Physical and Mechanical Tests . . . 2
2.1. Test of the Viscoelasticity of Vulcanized and Unvulcanized Materials . . 3
2.2. Tests on Unvulcanized Materials . . 4
2.2.1. Rheometric Tests . . . 4
2.2.2. Curemetric Tests . . . 5
2.2.3. Special Tests . . . 5
2.3. Tests on Vulcanized Materials . . . 5
2.3.1. Properties of the Rubber . . . 6
2.3.2. Endurance and Stability . . . 6
2.4. Physical Material Constants . . . 7
3. Electrical Tests . . . 7
4. Rubber Analysis . . . 7
4.1. Compound Formulation . . . 8
4.2. Polymer Analysis . . . 8
4.3. Filler Analysis . . . 8
4.4. Extract Analysis . . . 8
4.5. Special Analyses . . . 8
5. References . . . 8

1. Purpose, Scope, and Limitations of Testing

Cross-linked rubber is an important chemical material. Its ease of deformation to a very high extent (e.g., five to ten times the original length), which is almost completely reversible, is its major characteristic. On oscillatory deformation, part of the mechanical energy applied is converted into heat as a result of internal friction. Depending on the composition of the cross-linked rubber, it can act as an effective barrier to gases and liquids, and in extreme cases, it has very low electrical conductivity. Because of these properties it is used in many technical goods, as well as tires.

Testing is performed to determine characteristic properties of the rubber. With the aid of these data the applicability of a rubber for a particular use and its endurance and stability can be evaluated. Properties of the raw rubber and the unvulcanized rubber compound must also be determined to assess processibility (mixing, production of preforms, vulcanization), including changes on long-term storage. The wide variety of uses of rubber means that many properties must be tested. Rubber testing is described in more detail in [1–6].

To some extent, *vulcanized rubber* is difficult to describe in physical terms. Its deformation behavior deviates considerably from the usual Hooke's law for solid materials (deformation proportional to the force, with the modulus of elasticity as the constant of proportionality). Besides elastic deformation, cross-linked rubber also exhibits viscous flow. *Raw rubber and unvulcanized compounds* are also difficult to describe in physical terms. In principle, these are liquids with extraordinarily high viscosity, but the deviations from the usual Newtonian law for liquids (deformation rate proportional to the force with viscosity as the constant of proportionality) are extreme. In addition, these materials exhibit elastic deformation, as well as viscous flow which influences shaping processes. This particular behavior of rubbers depends strongly on time, intensity of stress, and temperature. The result of a test is therefore strongly influenced by these parameters.

Test Standards. Since complete description of the properties of a material by comprehensive testing is normally not possible for economic reasons, test standards have been developed that allow typical properties to be determined reproducibly when certain test parameters are adhered to. Test standards are also the basis of specification agreements in business dealings.

The wide variety of uses of rubber and the particular physical characteristics of raw rub-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_463

ber, unvulcanized rubber compounds, and vulcanized rubber have led to a large number of test standards. In its "Catalogue 1992" the ISO lists about 320 test standards of the Technical Committee 45 for Rubber and Rubber Products, most of which are specific for certain rubber goods [7]. In the general test standards for solid rubber and its compounds, the ASTM Test Standards Yearbook 1991 includes ca. 85 physical test standards [8]. The German DIN – VDE handbook 47 has about the same number of physical test procedures [9], and the most recent DIN handbook 131 has about 35 standards for chemical test procedures [10].

Each test result can only be assessed if the test standard and all parameters are specified. If there is no national or international test standard, the test must be carried out according to an in-house standard. Only in this way does a comparison with earlier tests become feasible. This applies particularly to tests in the quality control and quality assurance sectors in the production of rubber articles, compounds, polymers, and chemicals. For this area only a limited number of tests are used for permanent control for economic reasons. Their choice is based on experience so that critical deviations can be recognized with certainty. A prerequisite for successful choice is the existence of precise data: long-term repeatability of the test in one laboratory [over weeks to months (some typical values are given in [11])], as well as reproducibility in different testing laboratories. These precision data of a test procedure have been determined systematically for several years and are now included in the test standards. The ASTM standards are particularly far advanced in this regard [8]. A second prerequisite is a knowledge of the sensitivity of the tests, to allow technically relevant material deviations – i.e., differences in test results that are greater than the test precision by a factor of 3 or more [11] – to be recognized with certainty. Further requirements in the important field of quality assurance testing are discussed in [12] and [13].

Tests in the *developmental sector* are generally much more extensive. For development work, sufficiently comprehensive material characterization is necessary to allow prediction of actual technological behavior. The problems that arise can be great (see also Section 2.3.2). This difficulty is clearly indicated with regard to the most important ASTM test standards [14]. For any particular case, sometimes the necessary experience involving the possible use and limitations of selected laboratory test procedures can be obtained only by costly correlation investigations with tests on articles or tires (see [15] for an example concerning abrasion).

However, in this field test precision is not so critical because testing is performed in direct comparison with a standard material. An improvement in a material property can be more easily detected because the improved short-term repeatability of a test can be used to its full advantage.

With laboratory processibility tests of rubber polymers and compounds, analogous problems concerning the accurate prediction of practical behavior can occur [16].

2. Physical and Mechanical Tests

Some review articles supplement the general references cited in Chapter 1 and contain some detailed references [17–20]. Some areas are dealt with comprehensively (e.g., processibility tests [21], performance limits of vulcanized rubber [22], fatigue and fracture [23], viscoelasticity [24], friction [25], and abrasion [26]).

2.1. Test of the Viscoelasticity of Vulcanized and Unvulcanized Materials

(see also → Plastics, Properties and Testing, Chap. 2.1.2., → Plastics, Properties and Testing, Chap. 2.1.3.)

The unique stress – strain behavior of rubber (see Chap. 1) is determined by its molecular structure. Long polymer chains with thermally excited chain segment mobility form loose coils for thermodynamic reasons (entropy maximization). These coils are matted together. If a force is applied, the coil is first slightly deformed, with relatively weak entropy-elastic restoring forces (rubber elasticity); and second, the polymer chains begin to uncoil to achieve maximum entropy again (onset of flow).

The flow processes are severely limited by cross-linking but are still present.

This viscoelastic behavior can be determined by the following tests:

Creep Test. A test specimen is loaded with *constant stress,* and its deformation is determined over a given period of time. The specimen becomes elastically deformed immediately and subsequently flows.

Relaxation Test. A test specimen is subjected to a *constant strain,* and the stress is determined over a period of time. First a stress peak occurs caused by elastic restoring forces. This peak is followed by a decrease in stress (relaxation) because of the onset of flow processes. With unvulcanized material the force ultimately drops to zero. A spectrum of relaxation times (specific to the particular material) characterizes the rate of this relaxation; the spectrum can be determined from the decrease in stress [27]. A reliable test instrument is available [11], [28].

Test Using Periodic, Cyclic Stresses Through Forced Vibrations. A test specimen is subjected to an oscillating sinusoidal deformation and the consequent force variation, which is also sinusoidal, is analyzed [29]. The quotient of force amplitude per cross section of the test specimen and the deformation amplitude is a measure of the dynamic modulus. For rubber this modulus is lower than for metals by a factor of ca. 10 000. In practice, for cross-linked materials the dynamic modulus corresponds to the modulus of elasticity and therefore characterizes the elastic behavior of the material. Because of the flow processes that also occur, the force precedes the deformation [in the region of the deformation maximum, the stress has time to relax to a certain extent (see above)] so that it reaches its maximum earlier to a degree given by the phase angle δ. The phase angle increases with increasing flow; the loss factor $d = \tan\delta$ is a relative measure of the energy loss occurring per cycle, which is converted into heat as a result of internal friction. For vulcanized materials, $\tan\delta$ is typically 0.01 – 0.3; for unvulcanized materials, $\tan\delta$ is > 0.4.

Tests Using Periodic, Cyclic Stresses Through Free Oscillations. Tests can be performed with a torsion pendulum [30]. A rubber band as the test specimen, together with a suspended weight, forms a system that carries out free damped torsional oscillations after impulse. The dynamic modulus of the rubber can be calculated from the frequency of the oscillations. Because of the flow processes occurring in parallel to elastic deformation, mechanical energy is converted into heat and the amplitude of the oscillations continually decreases. The logarithmic decrement $\Lambda = \ln(A_1/A_2)$ (where A_1 and A_2 are the amplitudes of two consecutive oscillations) is a measure of this energy loss and is approximately proportional to the loss factor d.

The last three tests are used both for unvulcanized and vulcanized rubbers. The results obtained in one test can in principle be converted to those obtainable from another test, since they can all be derived from the relaxation time spectrum and a limiting modulus [5]. The spectrum must, however, be determined over a sufficiently long period of time. In practice, because of the experimentally limited range of measuring frequency or observation time, the test temperature is varied greatly, whereby relaxation times can be moved into the measuring range, as given by the Williams – Landel – Ferry (WLF) equation [5].

The derivations mentioned above are, however, only strictly valid in the region of small deformations, as long as there is a linear stress – strain relationship. Stress softening of rubber (Mullins effect [31]) can sometimes be observed in these types of test. Materials containing reinforcing fillers exhibit extreme nonlinearity even at relatively small deformations [32–35].

Low-Temperature Limit of Rubbers. If the test temperature is greatly lowered, the relaxation times of the chain segment mobility are ultimately so long that the polymer chain appears to be immobilized under the stress conditions of the test. The rubber is hard and glassy, and can be deformed only by using correspondingly large forces; it has lost its typical rubber elasticity.

The latter two dynamic tests mentioned, are also used for determination of the low-temperature limit of the rubber [36]. The torsion pendulum [30] has been introduced especially for the determination of the glass transition temperature T_g, i.e., the temperature of the maximum value of the logarithmic decrement. The forced vibration [29] is often used for determination of the technologically relevant low temperature limit T_L for entropy-elastic behavior (e.g., the

temperature at which the dynamic modulus has risen by a factor five times its value at room temperature).

The relaxation time spectrum of raw rubber and rubber compounds is important for processing. Short relaxation times promote filler incorporation and the adhesion of compounds in the molding (building-up) of preforms of rubber articles. Long relaxation times are responsible for the elastic behavior of the compound (e.g., die swell on extrusion and collapse resistance of unvulcanized profiles).

In vulcanized rubber (e.g., in tires), short relaxation times are important for improved adhesion to wet or icy roads, and long relaxation times increase rolling resistance.

Only elongation and shearing of rubber can be achieved by relatively small forces, not compression. This means that in the linear-viscoelastic region and with a Poisson's ratio of $\mu \approx 0.5$, the simplification $E = 3\,G$ (E = modulus of elasticity, G = shear modulus) is valid (i.e., only one modulus value needs to be determined).

2.2. Tests on Unvulcanized Materials

Raw rubber and unvulcanized rubber compounds are characterized principally with regard to their processibility in the fresh state, after storage, and after thermomechanical load.

Besides the tests for linear viscoelastic behavior described in Section 2.1, other tests involve greater loads that are more similar to those experienced in practice. Here, possible changes in the material (e.g., polymer chain scission and possible subsequent reactions) must be determined by rheometric, curemetric, and special tests. Generally, the previous history of the material and the test specimen greatly affect test results and therefore should be defined precisely.

2.2.1. Rheometric Tests

The aim is to characterize the rheological behavior (flow properties) of the material under conditions that are relevant to actual use. Many test procedures are available to shear, compress, or extrude the material [16], [21].

Mooney Shearing Disk Viscometer. The Mooney shearing disk viscometer is the most widely used method [37]. The resistance of the material to prolonged shearing is measured. This resistance is a measure of the viscosity under standardized test conditions. The Mooney viscosity is a very important specification parameter for raw rubber. Problems relating to the precision of the test are discussed in [38–41]. Improvements have been proposed or have already been introduced in more recent standards [41–43], including the pretreatment procedure of the material before the actual test [44].

Extensions of the test using the Mooney viscometer have been described. They involve variations in the shear rate (see, for example, [45]) and measurement of the relaxation (decrease in the torque of the shearing disk after shearing has stopped at the end of the test [46–48]). The former allow determination of the non-Newtonian, i.e., shear-rate dependent viscosity; the latter characterizes the elasticity of simple polymer systems.

Williams and Defo Test and Other Compression Plastometers. By compressing a cylindrical test specimen between two parallel plates, the viscous behavior and subsequent elastic recovery can be determined. The Williams test is standardized [49]. The defo test, preferred in Germany, was standardized [50]; it has gained new importance through modernization and further development to give a very sensitive and meaningful test method [46], [51]. In a special rapid test, only a viscosity-related plasticity number is determined [52] which is used to test the stability of natural rubber to oxidation [53].

Capillary Viscometer. In a high-pressure capillary viscometer the material is extruded through a capillary, whereby extremely high shear rates occur if the material adheres to the capillary wall. This is intended to simulate extrusion and injection molding processes as far as possible. A measure of the viscosity is determined from the pressure and extrusion rate, and characterization of the elasticity from the die swell [54], [55]. A widely used special instrument is described in [56].

Other Tests and Applications. Other rheometric tests are discussed in [21]. Basically all the tests mentioned can be used to assess the storage stability of rubber polymers or compounds by testing before and after storage. For the sake of speed, storage is usually carried out at high temperature (hot air storage). The Mooney viscosity test is used mainly for polymers, and the Mooney scorch test (see below) for compounds.

Material changes resulting from mechanical or thermomechanical load can be determined analogously by testing before and after application of the load [57], [58], or they can be integrated in a test involving strong mechanical loading at elevated temperature [59]. A prolonged Mooney viscosity test can also be used for this purpose.

2.2.2. Curemetric Tests

Curemetric tests are used to characterize the cross-linking reaction in which the rubber compound is vulcanized. The material is sheared weakly through periodic sinusoidal oscillation in a heated chamber, and the resulting stress is measured. The stress increases with increasing cross-linking of the material. The method using a rotor in the chamber [60] leads to delayed heating of the material; the rotor-free test method is more suitable [61], [62].

Analysis of the course of the stress signal with time gives detailed information on material properties relevant to processing, such as the scorch and cure time, the viscosity of the compound, the cross-link density, and sometimes the reversion behavior (decrease in cross-link density after optimum cure). With modern instruments the decrease in the loss factor $d = \tan \delta$ (see Section 2.1) with increased cross-linking can also be determined.

2.2.3. Special Tests

The onset of cross-linking of a compound can still be tested in the Mooney shearing disk viscometer [63]. This Mooney scorch test is also carried out on stored compounds to characterize the scorch behavior and the stability to storage.

A special capillary rheometer, with a ramification flow mold built on, allows simulation of the injection molding process in the laboratory [64], [65]. The mixing process can be simulated and analyzed on a very small scale [66] or on a normal laboratory scale [11], [67]. A special relaxation test is described in [68] and [69].

Tests have been described to assess the tack and green strength of compounds, which are important in compound processing, especially in the building-up of complicated articles such as tires [70–72]; some of these tests have been standardized [73].

For rapid monitoring of the mixing process, the compound density, for example, is determined. For the special monitoring of the filler dispersion achieved in a compound, which is important for rubber properties, a range of tests can be used [74].

2.3. Tests on Vulcanized Materials

Vulcanized rubber needs to be characterized in relation to its performance during a particular use. The properties of the freshly produced rubber and the changes in properties after application of a load are determined, thus simulating a practical situation.

2.3.1. Properties of the Rubber

The linear viscoelastic behavior is tested according to Section 2.1. In addition, many technological tests exist, which are either particularly simple and economical or are based on intended applications.

Hardness. The simplest characterization of elastic deformation behavior is given by the measurement of hardness. The resistance of the rubber to indentation by an object is determined. The Shore A and Shore D hardness are standardized by using a truncated cone and a cone, respectively, as the indenter [75], and the international degree of rubber hardness is determined by using a spherical indenter [76]. Hardness is an important parameter in the specifications for rubber articles. The relation between hardness and modulus is described in [77] and [78]; problems of precision are dealt with in [79].

Rebound Resilience. The standardized test of rebound resilience is the simplest characterization of viscous hysteresis behavior. A metal object rebounds after a free fall from a defined starting position onto rubber [80]. From the height to which it bounces, the percentage of impact energy that is recovered can be calculated. The remainder of the energy is converted irreversibly into heat by rubber hysteresis. Since the hysteresis of rubber depends strongly on frequency, this test, to be regarded as a single-point test in frequency and time, represents only a first rough material characterization. The temperature dependence should also be taken into account.

Tensile Test. The tensile test has great practical importance. A rubber test specimen (dumbbell-shaped rod or ring) is extended in a tensile test machine up to breaking. Typical modulus values (force per cross section) are determined for certain elongations (e.g., 100 %, 300 %) from the forces measured, extensions, and test specimen cross sections, and the limiting values of tensile strength and elongation at break are also determined [81]. Stress softening (Mullins effect [31]) and test specimen preparation (see, for example, [82], [83]) must be taken into account.

Tear Resistance. Mechanically damaged rubber can tear further under relatively low loads. The standardized tear resistance is determined in a tension test by using notched test specimens with different geometries [84]. A critical comparison is given in [85] and a general critical review in [14], which indicates that sufficient experience is still not available to simulate practical conditions with enough certainty in the tests mentioned so far.

Lower Temperature Limit. The lower temperature limit for rubber, which is defined as the transition to the glassy state is important. This limit varies widely depending on the type of rubber and can be affected to a certain extent by compound formulation. Besides the tests described in Section 2.1, tests that are more strongly orientated toward technical applications have been described (brittleness point [86], Gehman test [87]). Other tests are given in [14] and [22].

Other Properties. Tests of the adhesion between rubber and fabric [88] or cord [89], and between rubber and metal [90], and the test for skid resistance [91], are technically relevant. For special applications, the gas or fuel permeability, standardized according to [92] or [93], must be tested.

2.3.2. Endurance and Stability

Changes in properties under load with time are determined. The standardized loads simulate the practical situation. They include static (seals) or dynamic (tires) deformations, contact with aggressive gases or liquids, irradiation and normal aging phenomena, and contact with air or water. Generally in the test standards, only one load at a time is effective. However, this can lead to problems when the results are transferred to practical conditions where the rubber is subjected to several interacting loads in an article (thus shortening its lifetime). The necessary overloading in the laboratory test, in order to obtain any change at all in the relatively short laboratory test times (accelerated process), represents an additional problem. A successful example of testing under accelerating conditions is given in [94]. However, the molecular processes that lead to property changes do not always remain constant on increasing the load (e.g., raising the temperature in an aging test [95]). Carrying out endurance and stability tests and interpreting the results require a great deal of experience (see, for example, [96], [97]). The most important tests are described briefly.

Stress Relaxation. The stress relaxation in constant compression [98] simulates the load on a rubber seal. A less expensive, and therefore often used, test determines the compression set after longer compression of a cylindrical test specimen [99].

Continuous dynamic loads are simulated in flexometer tests [100].

Aging Resistance. The resistance to aging is determined by storage in hot air, air, or oxygen under pressure at elevated temperature and by comparing properties of stored and unstored

test specimens [101]. Generally hardness is determined and the tensile test is carried out; the change in the elongation at break is generally evaluated [15].

The resistance to liquids is characterized by storing test specimens in them and determining changes in weight and volume (swelling) [102]. Comparison tests are possible and generally performed.

Ozone Resistance. Ozone in the air leads to cracks in the surface of deformed rubber. The ozone resistance test simulates this by storing deformed test specimens in air with increased ozone concentration. The test specimen surface is assessed visually for crack formation at defined time intervals [103]. According to ISO, a cyclic deformation of the test specimen is also standardized [104].

Other Tests. In a standardized continuous flexure test, the resistance of the rubber to *crack formation* [105] or *crack growth* [106] is determined. A test of resistance to *cyclic extension* is given in [107].

A rapid laboratory test of *abrasion,* more than 1000 times faster than tire abrasion, should at least make screening tests possible [108]. Laboratory weathering tests are described in [109], and a daylight test is given in [110].

2.4. Physical Material Constants

The physical constants of materials can be determined by the generally valid experimental methods. Thermal expansion, specific heat, thermal conductivity and thermal diffusivity, and density are important.

3. Electrical Tests

Natural rubber and many synthetic rubbers possess very low electrical conductivity and are therefore suitable as electrically insulating materials. By incorporating carbon black, however, the conductivity can be increased by several powers of ten until the antistatic region is reached so that even conducting materials are formed. Rubber compounds can therefore be applied in many different fields.

The test of the *insulating properties* is described in [111], that of *anstistatic* and *electrically conducting rubber* in [112]. Other tests for evaluating electrically insulating materials are standardized [113]. In assessing the results of these tests, it should be taken into account that they are carried out in a standard climate (23 °C, 50 % humidity). The conditions of use of the article can differ considerably from this, with corresponding effects on the electrical properties.

Conductivity measurements are also used to assess the *dispersion of carbon black,* as a simple, rapid monitoring of the mixing process; see [114] for an overview, and [115–120] and [58] for examples. Special conductivity measuring probes in the internal mixer allow, according to [121], incorporation of the carbon black to be monitored during the mixing process [122–127], [11].

With dielectric measurements the course of the cross-linking reaction can be monitored in situ [128], [129].

4. Rubber Analysis [130]

The qualitative and quantitative characterization of polymers, additives, and fillers in a compound generally requires separation into these components and the use of different instrumental techniques. In the past, standardized chemical analysis methods were preferred for determination of the components of rubber compounds and vulcanizates. In [131] the use of the ASTM standard D 297 [132] is described for the reconstruction of vulcanizates. In the following, physical analytical techniques are listed briefly [133], [134] because they require little time for the analysis and give more information on the products investigated. This applies particularly to spectroscopic methods (e.g., solid-state ^{13}C NMR [135], Fourier-transform IR (FT-IR) [136], and pyrolysis mass spectrometry (PY-MS) [137]), which allow a direct analysis of the polymers and other components in carbon-black-filled vulcanizates.

4.1. Compound Formulation

For a quantitative determination of the compounding ingredients in vulcanizates, thermogravimetry is particularly suitable [138]. If the

decomposition regions of the components overlap, as in the case of aromatic oils and polymers [139] or carbonates and carbon blacks [140], the exact proportions of the compounding ingredients cannot be determined. The same applies to polar rubbers, which form carbon black on decomposition under nitrogen [141].

4.2. Polymer Analysis

Identification of the type of polymer in the vulcanizate can be carried out by several methods. In one method, the previously extracted vulcanizate is pyrolyzed [142], e.g., in a high-frequency pyrolyzer, and the gaseous and/or liquid decomposition products are investigated by GC [143], MS [144], or IR [133]. Alternatively, the sample is devulcanized in boiling 1,2-dichlorobenzene [145], and the soluble polymer components are analyzed, e.g., by IR [136]. In the determination of microstructure and the quantitative analysis of polymer blends, solid-state ^{13}C NMR [135] is superior to other analytical methods.

4.3. Filler Analysis

Because of their high IR absorption, inorganic fillers, such as quartz, silicates, or carbonates, can be identified in vulcanizates by FT-IR spectrometry if the technique of attenuated total reflection (ATR) [146] or, in the case of carbonblack-filled vulcanizates, photoacoustic spectroscopy (PAS) [147] is used. Further information, (e.g., on individual filler particles and their distribution) can be obtained by combination of scanning electron microscopy of thin sections of vulcanizates and X-ray microanalysis [138], [148]. The types of carbon black used in the vulcanizate can be determined by using a transmission electron microscope (TEM). Down to the size of HAF blacks, this is also possible with the optical microscope, because of the different coloration of the blacks in the thin section.

4.4. Extract Analysis

Additives such as oils, plasticizers, antidegradants, and accelerators can be extracted from rubber compounds and vulcanizates with methanol or acetone [142]. After evaporation of the solvent, the extract residue is analyzed.

Whereas with spectroscopic methods – apart from mass spectrometry [149] – analysis is often possible only after the components have been separated, with chromatographic methods [150], both qualitative and quantitative determinations can be carried out. HPLC is particularly suitable for the detection of antioxidants, free sulfur, aromatic oils, and accelerators [136] since several samples can be analyzed simultaneously and the individual components can be identified by specific color reactions [151].

4.5. Special Analyses

The identification of exudates on the vulcanizate surface and the cross-link structure in natural rubber vulcanizates cross-linked with sulfur are described in [152] and [153].

5. References

General References

1. R. P. Brown: *Physical Testing of Rubber,* 2nd ed., Elsevier Applied Science Publishers, London 1986.
2. H. Schmiedel et al.: *Prüfung hochpolymerer Werkstoffe,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1977.
3. F. S. Conant in M. Morton (ed.): *Rubber Technology,* 3rd ed., Van Nostrand Reinhold Company, New York 1987, p. 134 – 178.
4. R. Clamroth in W. Gohl et al. (eds.): *Elastomere Dicht- und Konstruktionswerkstoffe,* Lexika-Verlag, Grafenau 1975, p. 49 – 69.
5. J. D. Ferry: *Viscoelastic Properties of Polymers,* 3rd ed., J. Wiley & Sons, New York 1980.
6. R. S. Lenk: *Polymer Rheology,* Applied Science Publishers, London 1978.
7. *ISO Catalogue 1992,* Genève, Switzerland 1992.
8. "Rubber, Natural and Synthetic – General Test Methods; Carbon Black" *1991 Annual Book of ASTM Standards,* **vol. 09.01,** ASTM, Philadelphia 1991.
9. "Kautschuk und Elastomere 1 – Physikalische Prüfverfahren," *DIN-VDE-Taschenbuch 47,* 5th ed., Beuth Verlag, Berlin 1988.

10. "Kautschuk und Elastomere 2 – Normen für chemische Prüfverfahren, Bodenbeläge, Latex, Ruße und Schaumstoffe," *DIN-Taschenbuch 131,* 3rd ed., Beuth Verlag, Berlin 1991.
11. R. Koopmann, G. Marwede, *Kautsch. Gummi Kunstst.* **45** (1992) 133 – 140.
12. S. Lüpfert, *Kautsch. Gummi Kunstst.* **43** (1990) 35 –40.
13. N. M. Levin, *Polym. Test.* **9** (1990) 315 – 327.
14. R. D. Stiehler in H. Long (ed.): *Basic Compounding and Processing of Rubber,* Rubber Division ACS, Akron 1985, pp. 170 – 186.
15. J. R. Beatty, M. L. Studebaker, *Elastomerics* **109** (1977) no. 8, 33 – 40.
16. P. S. Johnson, *Rubber World* **201** (February 1990) 16 – 24.
17. A. R. Payne, *Prog. Rubber Technol.* **34** (1970) 51 – 66.
18. R. P. Brown, J. R. Scott, *Prog. Rubber Technol.* **36** (1972) 67 – 90.
19. R. Ecker et al., *Kautsch. Gummi Kunstst.* **25** (1972) 5 – 10.
20. P. Kainradl, *Kautsch. Gummi Kunstst.* **31** (1978) 341 – 355; *Rubber Chem. Technol.* **52** (1979) 232 – 254.
21. R. H. Norman, P. S. Johnson, *Rubber Chem. Technol.* **54** (1981) 493 – 534.
22. Th. Timm, *Kautsch. Gummi Kunstst.* **39** (1986) 15 –25.
23. G. J. Lake, *Prog. Rubber Technol.* **45** (1983) 89 –143.
24. Th. Timm, *Kautsch. Gummi Kunstst.* **31** (1978) 901 –917.
25. A. D. Roberts, *Prog. Rubber Technol.* **41** (1978) 121 –158.
26. A. Schallamach, *Prog. Rubber Technol.* **46** (1984) 107 – 142.

Specific References

27. T. D. Ferry: *Viscoelastic Properties of Polym.,* 3rd ed., J. Wiley & Sons, New York 1980, pp. 81 – 82.
28. B. Devis, *Eur. Rubber J.* **167** (1985) no. 11, 21 – 25.
29. ISO 4664, 1987, ASTM D 2231 – 1987; DIN 53 513/ 01.1983.
30. ISO 4663, 1986; DIN 53 445/08. 1986.
31. L. Mullins, *Rubber Chem. Technol.* **21** (1948) 281 –300.
32. A. R. Payne, *Rubber Plast. Age* **42** (1961) 963 – 967.
33. A. R. Payne, *J. Appl. Poly. Sci.* **6** (1962) 57 – 63.
34. A. R. Payne, *J. Appl. Poly. Sci.* **7** (1963) 873 – 885.
35. L. Mullins in C. Hepburn, R. J. W. Reynolds (eds.): *Elastomers – Criteria for Engineering Design,* Appl. Sci Publishers, London 1979, pp. 1 – 18.
36. DIN 53 545, 12.1990.
37. ISO 289: 1985, ASTM D 1646 – 1990; DIN 53 523, part 2/05.1991; part 3, 11.1976.
38. J. Markert, *Gummi Asbest Kunstst.* **29** (1976) 568 –578.
39. S. Niemiec, *Polym. Test.* **1** (1980) 201 – 209.
40. H. Kramer, *Kautsch. Gummi Kunstst.* **33** (1980) 20 –25; *Polym. Test.* **2** (1981) 107 – 123.
41. R. Koopmann, *Kautsch. Gummi Kunstst.* **38** (1985) 281 – 287.
42. H. Kramer, *Kautsch. Gummi Kunstst.* **43** (1990) 912 –916; *Rubber World* **204** (1991) April 35 – 39.
43. H. Kramer, *Kautsch. Gummi Kunstst.* **45** (1992) 721 –725; *Rubber Plast. News* **13** (1992) March, 16 – 19.
44. DIN 53 523, part 1/03. 1992.
45. N. Nakajima, E. R. Harrell, *Rubber Chem. Technol.* **52** (1979) 9 – 19.
46. R. K. Koopmann, H. Kramer, *J. Test. Eval.* **12** (1984) 407 – 414.
47. H. Burkin, W. Spreutels, J. Sezna: "MV 2000 Mooney Viscometer – Mooney Relaxation Measurements on Raw and Compounded Rubber Stocks," *ACS Rubber Div. Meeting,* Detroit, Oct. 1989, paper no. 74.
48. N. Vennemann, S. Lüpfert, *Kautsch. Gummi Kunstst.* **44** (1991) 270 – 282.
49. ISO 7323: 1985; ASTM D 926 – 1989.
50. DIN 53 514/12. 1958, withdrawn 1973.
51. R. Koopmann, *Kautsch. Gummi Kunstst.* **36** (1983) 108 – 119; *Polym. Test.* **5** (1985) 341 – 373.
52. ISO 2007: 1991.
53. ISO 2930: 1981; ASTM D 3194 – 1984.
54. H. Wesche, *Kautsch. Gummi Kunstst.* **31** (1978) 495 –507.
55. J. L. Leblanc, *Polymer* **17** (1976) 235 – 240.
56. J. A. Sezna, M. T. Saake, *Rubber Wold* **195** (1987) Feb., 26 – 33.
57. H. Kramer, R. Koopmann, *Kautsch. Gummi Kunstst.* **37** (1984) 874 – 880.
58. R. Koopmann, J. Schnetger, *Kautsch. Gummi Kunstst.* **39** (1986) 131 – 139.
59. J. Markert, *Kautsch. Gummi Kunstst.* **38** (1985) 122 – 127; *Eur. Rubber J.* **166** (1984) no. 7, 28 – 30.
60. ISO 3417: 1991; ASTM D 2084 – 1991.

61. ISO 6502: 1991; DIN 53 529 part 1, 03.1983; part 2, 03.1983; part 3, 06.1983.
62. J. A. Sezna, H. A. Pawlowski: "New Test Results from Rotorless Curemeters,"*ACS Rubber Div. Meeting,* Detroit, Oct. 1989, paper no. 78; J. A. Sezna, *Rubber World* **205** (1992) Jan. 21 – 27.
63. ISO 667: 1981; ASTM D 1646 – 1990; DIN 53 523, part 4, 11.1976.
64. O. Göttfert, *Kautsch. Gummi Kunstst.* **35** (1982) 849 –857.
65. K.-H. Moos, *Int. Polym. Process.* **3** (1988) no. 2, 86 – 90.
66. E. Diani et al., *Kautsch. Gummi Kunstst.* **38** (1985) 912 – 916.
67. K. Gark, S. Sharman, *Kautsch. Gummi Kunstst.* **44** (1991) 256 – 263.
68. S. R. Moghe, *Rubber Chem. Technol.* **49** (1976) 247 –257.
69. D. F. Hasman, D. W. Askea: "Polymer Processability through Fundamental Viscoelastic Properties," ACS Rubber Div. Meeting, Detroit, Oct. 1991, paper no. 65.
70. O. K. F. Bussemaker, *Rubber Chem. Technol.* **37** (1964) 1178 – 1189.
71. G. Koszterszitz, *Colloid Polym. Sci.* **258** (1980) 685 –701.
72. G. R. Hamed, *Rubber Chem. Technol.* **61** (1988) 548 –553.
73. ISO 9026: 1991.
74. W. M. Hess, *Rubber Chem. Technol.* **64** (1991) 386 –449.
75. ISO 868: 1985; ISO 7619: 1986; ASTM D 2240 –1986; DIN 53 505, 06.1987.
76. ISO 48: 1979; ISO 1400: 1975; ISO 1818: 1975; ISO 7619: 1986; ASTM D 1415 – 1988; DIN 53 519, Bl. 1, 05.1972, Bl. 2, 05.1972.
77. K. Tobisch, *Kautsch. Gummi Kunstst.* **34** (1981) 105 –109.
78. St. Sponagel, Th. Lutz, *Kautsch. Gummi Kunstst.* **43** (1990) 861 – 865.
79. R. P. Brown, A. Soekarnein, *Polym. Test.* **10** (1991) 117 – 137.
80. ISO 4662: 1986; ASTM D 2632 – 1988; ASTM D 1054 – 1991; DIN 53 512, 03.1981.
81. ISO 37:1977; ASTM D 412 – 1987; DIN 53 504, 03.1985.
82. R. P. Brown, *Polym. Test.* **2** (1981) 29 – 37.
83. W. V. Chang et al., *Rubber Chem. Technol.* **54** (1981) 449 – 459.
84. ISO 34: 1979; ASTM D 624 – 1986; DIN 53 515, 01.1990; DIN 53 507, 03.1983.
85. R. Clamroth, Th. Kempermann, *Kautsch. Gummi Kunstst.* **38** (1985) 206 – 217.
86. ISO 812: 1991; ASTM D 2137 – 1989; DIN 53 546, 06.1980.
87. ISO 1432: 1988; ASTM D 1053 – 1989 a; DIN 53 548, 01.1983.
88. ISO 36: 1985; ASTM D 413 – 1982; DIN 53 530, 02.1981.
89. ISO 4647: 1982; ISO 5603: 1986.
90. ISO 813: 1986; ISO 814: 1986; ASTM D 429 – 1981; DIN 53 531, part 1, 12.1972; part 2, 06.1981.
91. ASTM E 303 – 1974.
92. ISO 1399: 1982; ISO 2782: 1977; DIN 53 536, 02.1985.
93. ISO 6179: 1989; ASTM D 814 – 1986; DIN 53 532, 09.1975.
94. R. A. Youmans, G. C. Maassen, *Rubber Chem. Technol.* **28** (1955) 1133 – 1140.
95. J. A. Kuczkowski, D. E. Miller in E. Meinecke, F. Stuchal (eds.): "Elastomer Properties – Predictive Testing," *136th Educational Symposium,* no. 23, ACS Rubber Division Meeting, Detroit 1989, paper no. F, p. 3.
96. G. Walter, *Rubber Chem. Technol.* **49** (1976) 775 – 822.
97. P. M. Lewis, *Polym. Test.* **1** (1980) 167 – 189.
98. ISO 3384: 1991; ASTM D 1390 – 1976 (discont. July 1989); DIN 53 537, 09.1985.
99. ISO 815: 1991; ASTM D 395 – 1989; DIN 53 517, 04.1987.
100. ISO 4666-1: 1982; ISO 4666-2: 1982; ISO 4666-3: 1982; ASTM D 623 – 1978; DIN 53 533, part 1 – 3, 07.1988.
101. ISO 188: 1982; ASTM D 573 – 1988; ASTM D 454 –1988; ASTM D 572 – 1988; DIN 53 508, 07.1977.
102. ISO 1817: 1985; ASTM D 471 – 1979; DIN 53 521, 11.1987.
103. ISO 1431-1: 1989; ASTM D 1149 – 1986; ASTM D 1171 – 1986; DIN 53 509 part 1, 05.1990.
104. ISO 1431-2: 1982.
105. ISO 132: 1983; ASTM D 430 – 1973; DIN 53 522, part 1 – 2, 01.1979.
106. ISO 133: 1983; ASTM D 813 – 1987; DIN 53 522, part 3, 01.1979.
107. ISO 6943: 1984.
108. ISO 4649: 1985; DIN 53 516/06.1987.
109. ISO 4892: 1981; DIN 53 387/06.1982.
110. ISO 877: 1976; DIN 53 388/08.1982.
111. ISO 2951: 1974; ASTM D 257 – 1991; DIN 53 482/VDE 0303, part 3/05.1983.
112. ISO 1853: 1975; ISO 2878: 1987; ASTM D991 –1989; DIN 53 482/VDE 0303 part 3/05.1983.

113. IEC 112, 1979; DIN 53 481: 1974; DIN 53 482: 1983; DIN 53 483 part 1/1969; DIN 53 483 Part 2/1970; DIN 53 483 part 3/1969; ASTM D 149 – 1990; ASTM D 150 – 1987.
114. W. M. Hess, *Rubber Chem. Technol.* **64** (1991) 386 –449.
115. B. B. Boonstra, *Rubber Chem. Technol.* **50** (1977) 194 – 210.
116. R. J. Cembrola, *Rubber Chem. Technol.* **56** (1983) 233 – 243.
117. R. J. Cembrola, *Polym. Eng. Sci.* **22** (1982) 601 – 609.
118. J. L. den Otter, G. A. Gerritse, *Plast. Rubber Process. Appl.* **4** (1984) 63 – 67.
119. G. E. Wardell: "Electrical Properties as Related to Rubber – Carbon Black Interaction" *124th Meeting Rubber Div.,* ACS, Houston 1983, paper no. 71.
120. G. Menges et al., *Gummi Fasern Kunstst.* **3** (1986) 126 – 129.
121. A. I. Medalia, *Rubber Chem. Technol.* **59** (1986) 432 –454.
122. S. V. Usachor et al., *Kauch. Resina* 1976 no. 12, 17 –18.
123. J. Spaček, Z. Vitek: "Bewertung der Homogenität von Kautschukmischungen beim Mischen im Kneter," *6th Internat. Kautschuksymposium,* Gottwaldov 1977.
124. J. Jentzsch, J. Spaček et al., *Plaste Kautsch* **30** (1983) 700 – 702.
125. J. Spaček et al., "The Continuous Monitoring of the Mixing Process," *IRC '84,* Moskau 1984, paper C 38.
126. T. A. Henneka, C. A. Rotz: "Monitoring the Carbon Black Loading and Dispersion in Rubber Compounds by Measuring Mixture Dielectric Properties," *39th Antec Proceedings,* Boston 1981, 223 –226.
127. J. Jentzsch et al., *Plaste Kautsch.* **38** (1991) 87 – 90.
128. S. Persson, *Polym. Test.* **6** (1986) 47 – 78.
129. S. Persson, *Plast. Rubber Process. Appl.* **7** (1987) 111 – 125.
130. W. C. Wake et al.: *Analysis of Rubber and Rubber-Like Polymers,* 3rd ed., Applied Science Publishers, London 1983. H. Ostromow: *Analyse von Kautschuk und Elastomeren,* Springer Verlag, Berlin 1981.
131. *1991 Annual Book of ASTM Standards,* **vol. 09.01,** Method D 297-1990.
132. J. J. Leyden and J. M. Rabb, *Elastomerics* **112** (1980) 34.
133. F. Mersch, R. Zimmer, *Kautsch. Gummi Kunstst.* **39** (1986) 427; **43** (1990) 281.
134. D. Brück, *Kautsch. Gummi Kunstst.* **39** (1986) 1165.
135. D. Groß, J. Kelm, *Kautsch. Gummi Kunstst.* **38** (1985) 1089; **40** (1987) 13; **42** (1989) 388. R. A. Komorowski, *Rubber Chem. Technol.* **56** (1983) 959.
136. D. Brück, *Kautsch. Gummi Kunstst.* **41** (1988) 875.
137. W. A. McClemen et al., *Rubber Chem. Technol.* **59** (1986) 178.
138. R. L. Zeyen III, *Kautsch. Gummi Kunstst.* **41** (1988) 974.
139. J. J. Maurer, D. W. Brazier, *Kautsch. Gummi Kunstst.* **36** (1983) 37.
140. J. Sickfeld, D. Neubert, D. Groß, *Kautsch. Gummi Kunstst.* **36** (1983) 760.
141. J. Harris, *Synthesis* **8** (1977) 20.
142. DIN 53 553/ 1979.
143. ISO 7270: 1987; ASTM D 3452 – 1978.
144. J. B. Pausch et al., *Rubber Chem. Technol.* **56** (1983) 1031.
145. ISO 4650, 1984; ASTM D 3677-1990.
146. J. L. König, *Adv. Polym. Sci.* **54** (1983) 87.
147. T. Nguyen, *Prog. Org. Coat.* **13** (1985) 1.
148. *Ullmann,* 4th ed., **5,** 620 – 621.
149. R. P. Lattimer et al., *Rubber Chem. Technol.* **61** (1988) 639.
150. R. H. Schuster, *Kautsch. Gummi Kunstst.* **40** (1987) 642.
151. ASTM D 3156- 1990.
152. D. Holtkamp et al., *Kautsch. Gummi Kunstst.* **44** (1991) 1007.
153. R. C. Hirst: "Detection of Sulfur Crosslinks in Cured Natural Rubber Using Solid State C-13 NMR Spectroscopy," *ACS Rubber Div. Meeting,* Detroit, Oct. 8. 1991, paper no. 69. W. Gronski et al., *Kautsch. Gummi Kunstst.* **44** (1991) 119. Y. Tanaka, *Rubber Chem. Technol.* **64** (1991) 325. D. Brück et al., *Kautsch. Gummi Kunstst.* **43** (1990) 477.

Rubidium and Rubidium Compounds

WINFRIED LENK, Chemetall GmbH, Frankfurt/Main, Federal Republic of Germany(Chap. 1-6, 8)

HORST PRINZ, Chemetall GmbH, Frankfurt/Main, Federal Republic of Germany (Chap. 7)

1. Introduction 1
2. Properties 1
3. Raw Materials and Production 1
4. Rubidium Compounds 3
5. Quality Specifications and Analysis . . . 3
6. Storage and Transportation 3
7. Uses and Economic Aspects 3
8. Toxicology and Occupational Health . . 4
9. References 4

1. Introduction

Rubidium [*7440-17-7*] was discovered in 1861 by BUNSEN, who found hitherto unknown spectral lines during experiments on cesium salts obtained from spa water from Bad Dürkheim, Germany. Soon after this, he succeeded in preparing rubidium metal from its tartrate.

2. Properties

Rubidium, the penultimate alkali metal element that is used in industry, is named after two characteristically colored spectral lines (Latin: *rubidus* = dark red). It is a very reactive soft metal with a silvery luster and has the typical chemical properties of an alkali metal. It ignites spontaneously in air and burns with a violet flame. It reacts explosively with water to form colorless rubidium hydroxide [*1310-82-3*], the second strongest base after cesium hydroxide. Rubidium forms four oxides, all of which are colored [1]. Rubidium is monovalent in all its compounds. The isotopic composition of natural rubidium is 72.15 % ^{85}Rb [*13982-12-2*], a stable isotope, and 27.85 % ^{87}Rb [*13982-13-3*], a β^- emitter. Eighteen artificial isotopes are known. Between mass numbers 79 and 84, these are mainly β^+emitters, and between mass numbers 85 and 95 they are exclusively β^-emitters. Some important physical properties of natural elementary rubidium are:

Atomic number	37
Relative atomic mass	85.468
Melting point	39 °C
Boiling point	696 °C
Ionization potential	4.16 eV
Work function	2.19 eV
Normal electrode potential	−2.99 V
Electrical conductivity (at 0 °C)	$8.86 \times 10^4\ \Omega^{-1}\ \mathrm{cm}^{-1}$
Thermal conductivity (liquid)	$29.30\ \mathrm{W\,m^{-1}\,K^{-1}}$
Specific heat capacity	
Solid	$0.332\ \mathrm{J\,g^{-1}\,K^{-1}}$
Liquid (from *mp* to 1093 °C)	$0.368\ \mathrm{J\,g^{-1}\,K^{-1}}$
Gas	$0.242\ \mathrm{J\,g^{-1}\,K^{-1}}$
Heat of fusion	25.71 J/g
Heat of sublimation (at 25 °C)	1004.3 J/g
Heat of vaporization	887.6 J/g
Vapor pressure (427 – 1093 °C)	log (p/mbar) = 7.0493 – 3891.8 K/*T*
Mohs hardness	0.3
Atomic radius	0.243 nm
Covalent radius	0.211 nm
Ionic radius	0.149 nm
Density at 18 °C	1.522 g/cm^3
at *mp*, liquid	1.472 g/cm^3
Crystal structure	Body- centered cubic

3. Raw Materials and Production

[2–6]

Rubidium is very widely distributed and always occurs in combined form owing to its high chemical reactivity. Unlike its homologues lithium and cesium, it is not a major component of any mineral. Nevertheless, it is more abundant in the Earth's crust than lithium and cesium (ca. 120 g/t). Its ionic radius [3] is similar to that of potassium, and it is found in association mainly with this element, but also with the other alkali

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_473

metals. The lithium-containing mica lepidolite [*1317-64-2*] has a high rubidium content (ca. 3 % Rb_2O). Also, pollucite [*1308-53-8*], a cesium aluminum silicate from Bernic Lake in Canada, can contain ca. 1 % Rb_2O.

Due to the similarity in their ionic radii, rubidium is always found in potash salt deposits, e.g., carnallite, sylvite, and langbeinite, which contain 10 – 300 g/t rubidium oxide, Rb_2O [*12509-27-2*].

Rubidium is also widely distributed in very small concentrations in spring water, rivers, lakes, and seas.

Although rubidium is more abundant than lithium and cesium, it is much more difficult to extract than the other alkali metals. Owing to its very widespread distribution and consequent low concentration in ores, rubidium production is nearly always linked to the production of lithium, potassium, and cesium [7]. The process based on the fractional crystallization of carnallite is now only of historical interest. The more selective methods of recovering rubidium from potash salts (in which the quantities of rubidium are large but the concentrations are very low) are based on extraction processes using substituted benzenes as extractants [8–10]. Other techniques include the use of liquid crystals in a d.c. electric field [11], selective cation exchangers based on nitrated and sulfonated polystyrene [12–14], and inorganic cation exchangers consisting of complex salts based on hexacyanoferrates [15], [16]. Acidic MnO_2 can be used in an adsorption – elution cycle to concentrate rubidium [17]. The liquid waste from the production of cesium from pollucite can also be treated to give rubidium (see → Cesium and Cesium Compounds, Chap. 3.).

Rubidium is obtained from the mixed alkali metal carbonates that are a byproduct of the abandoned production of lithium from lepidolite in the United States. After precipitating the less soluble dicesium hexachlorostannate, the rubidium is precipitated as dirubidium hexachlorostannate [*17362-92-4*]. It is converted to rubidium chloride [*7791-11-9*] by pyrolysis. Rubidium carbonate [*584-09-8*] is obtained by precipitating dirubidium zinc hexacyanoferrate(II) [*128804-74-0*] followed by thermal oxidation. In all processes, the main difficulty is separating the rubidium from the potassium and cesium.

In more recently published research work, the following processes are described: the extraction from weathered rocks by sodium chloride solution in a high-frequency electrical field or with dipicrylamine and ultrasound [18], and the thermal decomposition of rubidium alum (rubidium aluminum sulfate [*7488-54-2*]) in the presence of calcium oxide followed by dissolution to form rubidium sulfate [*7488-54-2*] [19]. Another proposal is the recovery of rubidium from intermediates that arise in the production of aluminum oxide from bauxites and nephelines by means of solvent extraction with long-chain *p*-alkylphenols [20].

4. Rubidium Compounds

Important compounds of rubidium are:

- Rubidium hydroxide, RbOH [*1310-82-3*], M_r: 102.48, *mp*: 301 °C, solubility in water: 180 g/100 mL at 15 °C
- Rubidium carbonate, Rb_2CO_3 [*584-09-8*], M_r: 230.95, *mp*: 837 °C, solubility in water: 450 g/100 mL at 20 °C
- Rubidium hydrogencarbonate, $RbHCO_3$ [*19088-74-5*], M_r: 146.49, *mp*: 175 °C (decomp.), solubility in water: 54 g/100 mL at 20 °C
- Rubidium fluoride, RbF [*13446-74-7*], M_r: 104.47, *mp*: 775 °C, *bp*: 1410 °C, solubility in water: 130.6 g/100 mL at 20 °C
- Rubidium chloride, RbCl [*7791-11-9*], M_r: 120.92, *mp*: 715 °C, *bp*: 1390 °C, solubility in water: 77 g/100 mL at 0 °C
- Rubidium bromide, RbBr [*7789-39-1*], M_r: 165.38, *mp*: 682 °C, *bp*: 1340 °C, solubility in water: 98 g/100 mL at 5 °C
- Rubidium iodide, RbI [*7790-29-6*], M_r: 212.37, *mp*: 642 °C, *bp*: 1300 °C, solubility in water: 152 g/100 mL at 17 °C
- Rubidium nitrate, $RbNO_3$ [*13126-12-0*], M_r: 147.47, *mp*: 310 °C, *bp*: ca. 1700 °C, solubility in water: 44 g/100 mL at 16 °C, solubility in water: 42 g/100 mL at 10 °C

These inorganic compounds are colorless or white, and have the properties typical of salts of alkali metals. They are mostly prepared by wet chemical reactions.

5. Quality Specifications and Analysis

Rubidium salts are marketed with purities up to > 99.5 %. Rubidium metal is obtainable with purities of 99.5 % and 99.8 % [2].

Metallic impurities are usually determined by atomic absorption spectroscopy, and more recently with ICP (inductively coupled plasma) excitation. The elements are then detected by conventional optical detectors or by mass spectroscopy (ICP-MS). The homologous elements sodium, potassium, and cesium can be analyzed by emission spectroscopy. The analysis of rubidium metal must be preceded by reaction to give an aqueous salt solution. Anions can be determined photometrically (e.g., sulfate, phosphate) or argentometrically (chloride).

6. Storage and Transportation

Because rubidium metal is very reactive toward water, steam, ice, air, and many other gases and substances, special precautions are necessary. The metal is shipped in sealed evacuated glass ampoules, or is hermetically sealed in metal containers packed in vermiculite and enclosed in a second outer metal container. Due to its low melting point (*mp* 39 °C), melting cannot be prevented during transportation to or through hot climatic zones. Rubidium salts are hygroscopic and must be protected from moisture during storage and transport. Radioactive rubidium preparations are subject to special storage and transport regulations.

7. Uses and Economic Aspects

Uses. Rubidium compounds are used in psychiatry and medicine [21], in magneto-optic modulators [22], solid-state lasers [23], phosphors (rubidium aluminate) [24], Rb – C molecular sieves for hydrogen absorption [25], paper pigments [26], glass hardening by ion exchange [27], components of electrolytes for fuel cells [28], in Na – Rb zeolites for separating mixtures of hydrocarbons [29], and for density gradient centrifugation. In the early investigations of the Fischer – Tropsch synthesis, RbOH was used to activate the iron catalyst [30]. Owing to its high ionic conductivity silver rubidium iodide, $RbAg_4I_5$, is of interest as a solid electrolyte for primary and secondary cells [31] and as an electrographic picture recording material [32].

The use of rubidium compounds to activate catalysts of various types may increase in terms of quantity. However, potassium and cesium often compete with rubidium in this application. Three types of catalyst may be distinguished. The most important group includes rubidium-activated transition metal oxide catalysts for the production of organic chemicals, e.g., the oxidation of acrolein to acrylic acid [33] or of C_4 hydrocarbons to maleic anhydride [34], the ammoxidation of isobutene to methacrylonitrile [35], or the oxidation of *o*-xylene or naphthalene to phthalic anhydride [36]. The second category includes the activation of metallic catalysts, e.g., silver contact catalysts for ethylene oxide manufacture [37]. The third category comprises rubidium compounds that themselves have catalytic properties, e.g., Rb_2O used in polyester production by polycondensation [38] and in NH_3 synthesis [39], and RbOH for the dehydrocyclization of 1-hexene to form benzene [40].

Economic Aspects. Rubidium manufacturers do not publish production or sales figures. World consumption in 1979 was estimated to be ca. 2 t [2], and has increased slightly since then. The demand for the metal is small compared to the demand for its compounds. In 1996, the price of 99.5 % pure metal was 0.70 – 1.00 $/g, and for 99.8 % pure metal, 0.80 – 6.00 $/g [2]. Prices for rubidium compounds (99.5 %) are in the range 250 – 550 $/kg, and for > 99.5 % pure compounds, 550 – 750 $/kg [2].

8. Toxicology and Occupational Health

It is not possible to give a precise assessment from the small amount of available data on the toxicity of rubidium. In plasma and blood, the properties of rubidium are not dissimilar to those of potassium, as far as can be determined from physiological investigations. The hazard potential is therefore regarded as low [41].

Rubidium metal is designated as a“flammable solid” and “dangerous when wet”, and

should be handled with great care under a protective gas or inert solvent.

In its compounds, the precautions and protective measures to be taken are determined by the anion. Thus, RbOH is very caustic, RbF shows the toxicity typical of the alkali metal fluorides, and $RbClO_4$ [*13510-42-4*] and $RbNO_3$ are oxidizing agents.

9. References

1. *Kirk-Othmer,* 3rd ed., **17,** pp. 492 – 499.
2. *Gmelin,* 8th ed., **System-no. Rubidium, 24.**
3. *Ullmann,* 4th ed., **20,** pp. 295 – 297.
4. C. E. Berthold: "Rubidium and Cesium – Where are they going?" *J. Met* **14** (1962) 355 – 358.
5. E. G. Kuznetsova, L. A. Kulikova, A. K. Ru, *Sovrem. Khim.-Metallurg. Metody Pererab. Kompleks. Rud. Syrya, M.* 1985, 102 – 6; *Chem. Abstr.* **107,** 25 536.
6. E. Schroll, *BHM Berg Hüttenmänn. Monatsh.* **131** (1986) 110 – 115.
7. Li, Minqian, CN 85 101 989, 1986; *Chem. Abstr.* **107,** 179 516.
8. E. M. Kuznetsova, N. A. Kresova, G. M. Panchenkov, *Zh. Fiz. Khim.* **41** (1967) 1041 – 1046; *Chem. Abstr.* **67,** 57 665.
9. V. E. Plyushchev et al., SU 323 459, 1970; *Chem. Abstr.* **79,** 22 004.
10. L. I. Pokrovskaja et al., *Zh. Prikl. Khim.* **50** (1977) 546 – 549; *Chem. Abstr.* **87,** 9148.
11. V. N. Golubev, A. G. Petrov, V. N. Verkhotvrov, SU 1 279 650, 1986; *Chem. Abstr.* **106,** 140 437.
12. R. Ratner, L. Kogag, D. Kohn, *J. Chromatogr.* **148** (1978) 539 – 544.
13. F. Kh. Ismailov, R. G. Osichkina, *Vzb. Khim. Zh.* (1988), no. 5, 56 – 58; *Chem. Abstr.* **110,** 138 080.
14. R. Ratner, J. Mittelman, D. H. Kohn, *J. Chem. Technol. Biotechnol. Chem. Technol.* **35A** (1985) no. 5, 209 –216; *Chem. Abstr.* **104,** 112 171.
15. S. A. Kolesova, V. V. Volkhin, V. V. Alikin, SU 552 105, 1975; *Chem. Abstr.* **87,** 137 936.
16. R. M. Shklovskaya, S. M. Arkhipov, SU 1 245 555, 1986; *Chem. Abstr.* **105,** 136 478.
17. V. V. Volkhin, G. V. Leonteva, SU 295 399, 1968; *Chem. Abstr.* **81,** 51 787.
18. Jiucheng Li, Xiaohong Li, Xian-Gan Li, CN 1 037 127, 1989; *Chem. Abstr.* **113,** 194 434.
19. S. M. Arkhipov, N. I. Kashina, SU 1 546 425, 1990; *Chem. Abstr.* **112,** 237 478.
20. A. M. Reznik, V. I. Vukin, L. G. Romanov, A. O. Baikonvrova, *Trav. Com. Int. Étude Bauxites Alumine Alum.* **17** (1982) 63 – 66; *Chem. Abstr.* **99,** 91 593.
21. H. L. Meltzer, R. R. Fieve, *Curr. Dev. Psychopharmacol.* **1** (1975) 203 – 242.
22. IBM, US 3 527 577, 1968.
23. Optical Materials Inc., US 3 506 584, 1966.
24. Toshiba, JA 74 90 689, 1973.
25. Ichiro Kato, JA 7 4 01 487, 1972.
26. Bayer, DE-OS 2 148 599, 1971.
27. Daini Seikosha Co., JA 72 23 885, 1970.
28. Leesona Corp., US 34 883 035, 1965.
29. Esso Research and Engineering Co., US 3 674 425, 1966.
30. F. Fischer, H. Tropsch, *Brennst. Chem.* **5** (1924) 201 –216.
31. B. Scrosati, *J. Appl. Chem. Biotechnol.* **21** (1971) 223 –228.
32. Nippon Electric, JA 74 16 440, 1972.
33. Standard Oil of Ohio, DE-OS 2 634 791, 1976.
34. Standard Oil of Ohio, US 4 087 444, 1976.
35. Montedison, DE-OS 2 403 716, 1973.
36. BASF, DE-AS 2 546 268, 1975.
37. Hoechst AG, DE-AS 2 740 480, 1977.
38. Mitsubishi Rayon, JA 74 13 635, 1968.
39. Chevron Research Co., US 3 653 831, 1968.
40. Phillips Petroleum, DE-OS 2 832 654, 1978.
41. U.S. Department of Health, Education and Welfare, NIOSH, Health Hazard Evaluation, Report 71 – 72, July 1973.

Rum → Spirits
Ruthenium → Platinum Group Metals and Compounds
Rutherfordium → Transuranium Elements
Saccharin → Sweeteners
Salicylaldehyde → Benzaldehyde

Salicylic Acid

Olivier Boullard, Rhône-Poulenc Chimie S.A., Paris, France

Henri Leblanc, Rhône-Poulenc Chimie S.A., Paris, France

Bernard Besson, Rhône-Poulenc Chimie S.A., Paris, France

1. Introduction 1
2. Physical and Chemical Properties 1
3. Production 2
4. Quality Specifications, Storage and Transportation, and Environmental Protection 3
5. Uses and Economic Aspects 4
6. Salicylic Acid Derivatives 5
7. Toxicology 6
8. References 7

Based on the corresponding article in Ullmann's 4th ed., written by Klaus-Peter Heise and Werner Bunge

1. Introduction

Salicylic acid [*69-72-7*], also called 2-hydroxybenzoic acid or *o*-hydroxybenzoic acid, is widely distributed in the plant kingdom in the form of esters.

Some 2400 years ago Hippocrates prescribed decoctions of willow leaves as a treatment for fever and pain. The active principle in willow leaves is salicylic acid, whose biosynthesis is based on the deamination of phenylalanine to *trans*- cinnamic acid which, when hydrolyzed and oxidized at the β- carbon atom, gives salicylic acid.

Salicylate esters are found in several plant genera, the principal ones being *Salix*, *Spiraea*, and *Gaultheria*. For example, methyl salicylate is present in large quantity in the form of a glucoside in birch bark, various spiraeas (meadowsweet or *Spiraea ulmaria*), and wintergreen leaves.

Salicylic acid was first obtained in 1838 by R. Piria, an Italian chemist at the Institute of Pisa, by melting salicylaldehyde, obtained from meadowsweet, with caustic potash. In 1844 the French chemist Cahours obtained salicylic acid by hydrolysis of methyl salicylate. The final step, namely preparation from a natural source, was accomplished in 1874 by the German chemist H. Kolbe, who synthesized the acid by carboxylation of sodium phenoxide, a process still in use.

Salicylic acid and its derivatives are used mainly to synthesize pharmaceutical products and as intermediates in the production of dyes and agrochemical and perfumery products.

2. Physical and Chemical Properties

Physical Properties. Salicylic acid, $C_7H_6O_3$, M_r 138.12, crystallizes in the form of colorless needles (water) or monoclinic prisms (ethanol), *mp* 159 °C; salicylic acid begins to sublime at 76 °C; flash point (closed cup) 157 °C; heat of sublimation 81.8 kJ/mol; density d_4^{20} 1.443; dissociation constants K_1 1.05×10^{-3}, K_2 4.0×10^{-14} (19 °C); vapor pressure 1.66 mbar (110 °C) and 19.3 mbar (150 °C); solubility (per 100 g of solution) in methanol 38.46 g (21 °C), ethanol 34.87 g (21 °C), chloroform 1.55 g (30 °C), benzene 1.00 g (30 °C); solubility (per 100 mL of solution) in diethyl ether 23.4 g (17 °C), acetone 31.3 g (23 °C); extremely soluble in ammonia, insoluble in liquid sulfur dioxide. Solubility in water:

°C	0	10	20	40	60	80
g/L	0.8	1.2	1.8	3.7	8.2	20.5

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_477

Salicylic acid is an aromatic *o*-hydroxy carboxylic acid, and in contrast to its *meta* and *para* isomers (→ Hydroxycarboxylic Acids, Aromatic, Chap. 4.1.) it is subject to intramolecular hydrogen bonding and steam volatility. It also sublimes more readily than its isomers (→ Sublimation, Chap. 4.2.1.), and is substantially more acidic: dissociation constants for the *meta* and *para* isomers are 8.7×10^{-5} and 3.3×10^{-5}, respectively.

Chemical Properties. The difunctional salicylic acid molecule combines the properties of phenols with those of aromatic carboxylic acids. One equivalent of alkali hydroxide neutralizes only the carboxyl group. An excess of hydroxide is required to form the dialkali salt, from which the free OH group reforms in the presence of carbon dioxide. Chelation occurs with some metal ions such as Fe(III), leading to a violet coloration. Salicylic acid can be used as an indicator in EDTA determinations of Fe(III) [6].

Salicylic acid is esterified by alcohols in the presence of strong acids without significant etherification. Combined ether – esters can be prepared from dialkali salicylate in the presence of alkyl halide; these are converted by alkaline hydrolysis into the corresponding alkoxybenzoic acids. The phenolic OH group is etherified by an alkaline aqueous solution of dialkyl sulfate, and esterified by the action of acyl halides or acid anhydrides. Reduction with sodium and amyl alcohol affords tetrahydrosalicylic acid, whose oxidation product is pimelic acid [7]. Catalytic hydrogenation of salicylic acid esters over Raney nickel produces esters of *cis-trans*-2-hydroxycyclohexane carboxylic acid [8].

Since electrophilic reactants attack the less sterically hindered 5-position in preference to the 3-position, it is possible to obtain either 5-substituted or 3,5-disubstituted derivatives directly. For example, the monosubstituted products obtained on nitration, sulfonation, halogenation, alkylation, acylation, or coupling with diazonium salts are generally salicylic acids substituted at the 5-position. Derivatives substituted exclusively at the 3-position are obtained by indirect means, such as substitution of sulfosalicylic acid at the 5-position, followed by elimination of the sulfonic group. More severe conditions may lead to decarboxylation as well. Thus, 2,4,6-trinitrophenol (picric acid) is obtained upon treatment of salicylic acid with fuming nitric acid, and tribromophenol results from treatment with bromine in the presence of water.

When heated at or above its melting point salicylic acid decomposes into phenol and carbon dioxide. Under a carbon dioxide atmosphere at 230 °C the main product is phenyl salicylate. At 250 °C, xanthone is formed in parallel with phenol. For the behavior of salicylic acid as a function of temperature see [9].

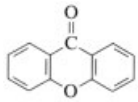

Xanthone

Analysis. Quantitative determination of salicylic acid is generally accomplished by titration in an alkaline medium. Kolthoff's method permits salicylic acid also to be determined with a bromide solution, which converts it to tribromophenol [10]. For trace analysis it is convenient to use colorimetric methods with the reagent iron(III) chloride [11]. Thin-layer chromatography (TLC) provides a convenient test of purity, since each of the isomeric acids can be detected by alkaline coupling (caustic soda) with diazotized *p*-nitroaniline, spraying with Fe(III) chloride, or fluorescence under UV light. The TLC approach is often replaced today by reversed-phase high-performance liquid chromatography (HPLC).

3. Production

Kolbe – Schmitt Synthesis. Salicylic acid is prepared on an industrial scale by the Kolbe – Schmitt synthesis from dry sodium phenoxide in a stream of carbon dioxide at 150 – 160 °C and a pressure of 5 bar.

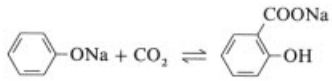

The use of pressure (Schmitt, 1884) results in a yield of about 90 %, whereas without pressure (Kolbe, 1874) the yield does not exceed

50 % because disodium salicylate and phenol are formed in equivalent amounts.

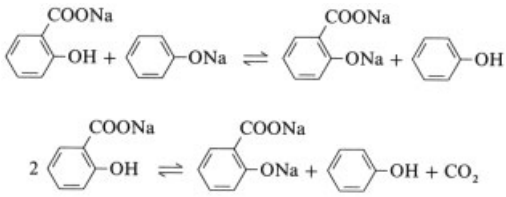

Alternative reaction mechanisms for the carboxylation of phenoxide salts have been described in the literature [12]. 4-Hydroxyisophthalic acid, which yields salicylate by direct release of CO_2 (or by carboxylation of phenolate), has been discussed as a possible intermediate in the reaction process [13]. An electrophilic substitution mechanism via a complex formed between phenol, one molecule of CO_2, and an alkali metal, has also been suggested [14]. Temperature, the nature of the alkali metal, and the CO_2 pressure are all of decisive importance with respect to the reactivity and selectivity of the phenoxide [12, 15].

The various stages in the industrial synthesis of salicylic acid [16] are outlined in Figure 1. Carboxylation may be carried out if in autoclaves heated with steam or heat-exchange fluids and equipped with counterblades, or in powerful mill autoclaves. The process is still conducted mainly in a batchwise manner. To ensure that reaction proceeds satisfactorily, the reaction mass must not only be in a finely ground state during the carboxylation stage, but water must also be rigorously excluded. The presence of water reduces the yield by protonating the phenoxide and releasing alkali-metal hydroxide, which then converts CO_2 into carbonate with the regeneration of water. Water may also be formed in situ via a secondary etherification reaction [15]. Sodium phenoxide is prepared with a 1 – 2 % molar excess of caustic soda; larger amounts of alkali lead to the formation of water, as described above. Anhydrous sodium phenoxide may be prepared either in the autoclave mixer itself by evaporation of an aqueous solution of phenoxide, starting at normal pressure and then gradually introducing vacuum, or in special drying equipment. In order to prevent discoloration and tar formation it is important that the carbon dioxide contains as little oxygen as possible (< 0.1 %).

The carboxylation step is exothermic: $\Delta H = -90.1$ kJ/mol. Using a 6 m^3 reactor, about 3 t of sodium phenoxide is converted into salicylic acid in 25 h. Phenol derived from the formation of disodium salicylate is recovered by distillation. The crude sodium salicylate is dissolved in a mixture of water and a decolorizing agent (e.g., activated charcoal, aluminum, or zinc powder [17]). Salicylic acid is then precipitated with sulfuric acid. This synthesis can be accomplished in a continuous manner by working with a solution of the anhydrous phenoxide in a suitable medium. Recommended solvents are phenol itself [18], higher alcohols [19], dialkyl ketones [20], nitrobenzene [21] and, as a dispersant [22], gasoline.

Other Processes. Consideration has been given to producing salicylic acid by air oxidation of *o*- cresolate at 230 °C in the presence of a copper-based catalyst [23] or copper benzoate (175 – 215 °C) [24]. Alkaline copper benzoate can also be converted by heat treatment into salicylate directly [25]. Benzoylsalicylic acid is an intermediate in the synthesis of phenol from toluene by the Dow process (→ Phenol, Chap. 4.2.).

Salicylic acid can also be obtained by fermentation of such polycyclic aromatic compounds as naphthalene with the aid of microorganisms [26] (→ Biotechnology).

4. Quality Specifications, Storage and Transportation, and Environmental Protection

Quality Control. Technical-grade salicylic acid obtained from the Kolbe – Schmitt process is already extremely pure: salicylic acid content 99.5 %; phenol, *p*-hydroxybenzoic acid, or 4-hydroxyisophthalic acid 0.05 – 0.1 % (as impurity); ash < 0.1 %; water 0.2 %.

An even higher quality acid (pharmaceutical grade) can be obtained by crystallizing the sodium salicylate from water at a temperature not exceeding 20 °C [27], or by sublimation of the acid at 20 mbar and a temperature of 154 °C

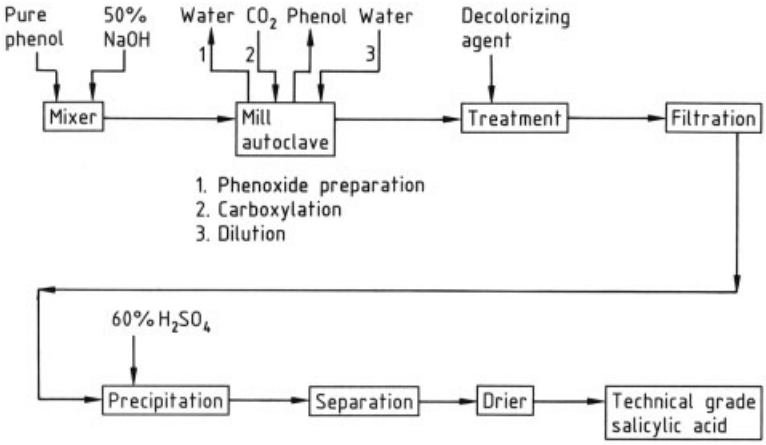

Figure 1. Simplified representation of salicylic acid production by the Kolbe – Schmitt method

[28] or with the aid of a carrier gas. A more modern process achieves sublimation directly by utilizing the heat of neutralization from the reaction of sodium salicylate with hydrogen chloride [29].

The quality specifications of the European Pharmacopoeia are limted to the following items: identity check; color (ethanolic solution); melting point (158 – 161 °C); assay (99.0 – 100.5 %); if sulfated ash (< 0.1 %); heavy metals (< 20 ppm); chloride (< 100 ppm); if sulfates (< 200 ppm); and loss on drying < 0.5 %.

Storage and Handling. Salicylic acid dust is combustible in air. The low combustion energy for the process indicates a high level of combustion sensitivity, so appropriate measures must be adopted to avoid sources of ignition and to protect against potentially severe explosive effects. Explosive conditions can be avoided by maintaining the oxygen level at < 8 %.

The explosive characteristics of salicylic acid dust (particle size < 100 µm) are as follows: minimum combustion temperature, 490 °C; minimum combustion concentration, 30 g/m^3; minimum combustion energy, < 5 mJ; maximum pressure produced, 7.2 bar; maximum rate of pressure increase, 216 bar/s [unpublished work carried out by Rhône-Poulenc in accordance with ISO 6184/1].

Transportation. Salicylic acid is not subject to any transport restrictions. The acid is transported in bags (25 kg), drums (50 kg), big bags (500 – 1000 kg), or bulk containers (15 – 20 t).

Environmental Protection. Effluent and gaseous emission problems are comparable to those posed by phenol.

Freshwater ecotoxicity of salicylic acid on *Daphnia magna*: (ED_{50} 24 h; immobilization): 180 mg/L (AFNOR T 90 301 Standard, French). Salicylic acid is readily biodegradable and very slightly bioaccumulable.

5. Uses and Economic Aspects

Salicylic acid is used mainly in the synthesis of acetylsalicylic acid, the most commonly dispensed pharmaceutical product. In the form of esters, amides, and salicylic acid salts it serves as a starting material for other pharmaceutical products. Technical-grade salicylic acid is used primarily as an intermediate in the production of agrochemical products, dyes, and colorants, as well as in the rubber industry and in the manufacture of phenolic resins.

Salicylic acid itself offers therapeutic benefits in the treatment of rheumatic disorders, for which purpose it is usually administered in the form of the readily soluble sodium salt. On account of its keratolytic action the acid is also widely used for cleaning the skin and removing scales.

By virtue of its bacteriostatic properties it is used as a disinfectant or preserative; however, its presence is not permitted in foods.

Economic Importance. The distribution of salicylic acid in terms of application can be estimated as follows: acetylsalicylic acid 55 %, esters and salts 18 %, resins 10 %, dyes and colorants 10 %. The development and introduction of new analgesics that compete aggressively with acetylsalicylic acid has had a direct effect on market evolution, and a steady decline in the consumption of acetylsalicylic acid has been observed. It is expected that the discovery of new uses will stabilize the situation.

6. Salicylic Acid Derivatives

Salts. The carboxyl group in salicylic acid is easily converted into salts by the action of metal carbonates. In order to prevent discoloration, aqueous salt solutions should be kept slightly acidic. Salts are obtained in solid form by concentrating their aqueous solutions.

Sodium salicylate [*54-21-7*], $C_7H_5NaO_3$, M_r 160.11, forms white, odorless, shiny crystalline flakes; solubility (in a solution of 100 mL): water 125 g (25 °C), ethanol 17 g (15 °C). The technical-grade product (99.5 %) is obtained by evaporating a solution of sodium salicylate; pharmaceutical-quality material [30] is prepared by two successive crystallizations of sodium salicylate hexahydrate from a 45 % aqueous solution at 10 °C. Sodium salicylate is used as an analgesic, antipyretic, and antineuralgic.

Other Salts. Besides sodium salicylate, a number of other common salts are known (e.g., ammonium, magnesium, calcium, aluminum), as is morpholine salicylate. Several of the salts are marketed under various trade names.

Esters. Several important salicylate esters are described below (→ Flavors and Fragrances, Chap. 2.6.5.). These are formed by reaction with the appropriate alcohols. All are soluble in both ether and alcohol, but only sparingly soluble in water.

Methyl salicylate [*119-36-8*], oil of wintergreen, $C_8H_8O_3$, M_r 152.15, is a colorless, oily liquid with a characteristic odor, *mp* − 9 °C, *bp* 222 °C, d_4^{20} 1.184. Methyl salicylate is synthesized by heating a mixture of salicylic acid and methyl alcohol in the presence of sulfuric acid. It is used to treat neuralgia and rheumatism, as well as to stimulate capillary blood circulation. It is also used as an insecticide, sunscreen, fragrance, and synthetic intermediate.

Benzyl salicylate [*118-58-1*], $C_{14}H_{12}O_3$, $M_r = 228.25$, is a clear liquid or colorless to opaque crystalline mass with a characteristic odor, *mp* 24 °C, *bp* 318 °C, d_4^{15} 1.180. It is obtained from a mixture of sodium salicylate and benzyl chloride, or by transesterification of methyl salicylate in the presence of benzyl alcohol. It is present in ylang-ylang and carnation oils, and is widely employed as an additive in soaps, in detergents, and in perfumery products (→ Benzyl Alcohol, Chap. 8.4.).

Isoamyl salicylate [*87-20-7*], $C_{12}H_{16}O_3$, M_r 208.26, is a colorless liquid with the fragrance of orchids, *bp* 270 °C, d_4^{20} 1.050. Isoamyl salicylate is used as a fragrance and stabilizer in perfumery and as an antirheumatic agent (topical application).

Phenyl salicylate [*118-55-8*], salol, $C_{13}H_{10}O_3$, M_r 241.22, is a colorless, crystalline powder, *mp* 43 °C, *bp* 172 °C at 16 mbar. It is obtained from a mixture of salicylic acid and phenol in the presence of sulfuric acid, or by transesterification of methyl salicylate in the presence of sodium phenoxide. Phenyl salicylate is used as an antiseptic, preservative, as a sunscreen and a general photoprotective agent for synthetic products, and as an emollient.

Acetylsalicylic Acid (→ Analgesics and Antipyretics, Chap. 2.1.,; → Anti-inflammatory – Antirheumatic Drugs, Chap. 3.1.; → Cardiovascular Drugs, Chap. 5.3.1.1.). Acetylsalicylic acid [*50-78-2*], $C_9H_8O_4$, M_r 180.15, is isolated as monoclinic colorless needles or crystalline powder (from water), or as flat platelets (from isoamyl alcohol), *mp* 143 – 144 °C depending on the heating rate and crystalline form; dissociation constant $K = 2.8 \times 10^{-4}$ (25 °C); solubility (in 100 mL of solvent): water 0.25 g (15 °C), ethanol 20 g (25 °C), ether 5 g (18 °C). The compound hydrolyzes to some extent during recrystallization from water.

Acetylsalicylic acid is prepared by reacting acetic anhydride with salicylic acid at a temperature of < 90 °C either in a solvent (e.g., acetic acid or aromatic, acyclic, or chlorinated hydrocarbons) or by the addition of catalysts such as

acids or tertiary amines [31]; for quality specifications see [30]. Acetylsalicylic acid is used as an antipyretic, analgesic, and anti-inflammatory agent, and it has an antirheumatic effect. Acetylsalicylic acid also has antithrombosis and anticoagulant properties. It is marketed by Bayer as a consumer product under the trade name Aspirin and as a bulk product by Rhône-Poulenc (Rhodine).

Other Derivatives.

Salicylamide [*65-45-2*], $C_7H_7NO_2$, M_r 137.13, is a white crystalline powder, *mp* 140 °C, sparingly soluble in water and soluble in alcohol. It is obtained by the ammonolysis of methyl salicylate [32]. Salicylamide is used as an analgesic, antipyretic, antirheumatic, sedative, and fungicide.

Salicylanilide [*87-17-2*], $C_{13}H_{11}NO_2$, M_r 213.33, forms colorless and odorless crystals, *mp* 136 – 137 °C. It is sparingly soluble in water and soluble in alcohol and chloroform. The substance is prepared from a mixture of salicylic acid and aniline in the presence of PCl_3 [33]. Salicylanilide and particularly its derivatives are powerful fungicides, and are used in the synthesis of dyes, colorants, lacquers, and textiles, and also as a disinfectant. It is distributed commercially under the trade name Shirlan (ICI).

3-Chlorosalicylic acid [*1829-32-9*], 3-chloro-2-hydroxybenzoic acid, $C_7H_5ClO_3$, M_r 172.57, *mp* 180 °C, is obtained either by the carboxylation of sodium 2-chlorophenoxide (Kolbe – Schmitt synthesis) or by chlorination of 5-sulfosalicylic acid, followed by displacement of the sulfonic group by means of superheated steam [34].

4-Chlorosalicylic acid [*5106-98-9*], 4-chloro-2-hydroxybenzoic acid, $C_7H_5ClO_3$, M_r 172.57, *mp* 211 °C, is prepared from 4-aminosalicylic acid by the Sandmeyer process [35] and used as a synthetic intermediate. The *N*-butylamide derivative of 4-chlorosalicylic acid (Jadit, Hoechst) is an antimycotic.

5-Chlorosalicylic acid [*321-14-2*], 5-chloro-2-hydroxybenzoic acid, $C_7H_5ClO_3$, M_r 172.57, *mp* 176 °C, is synthesized either by the chlorination of salicylic acid or by the Kolbe – Schmitt method starting from sodium 4-chlorophenoxide and CO_2 at 150 °C (higher temperature should be avoided since the decomposition reaction may be explosive) [34, 36]. It is used as an intermediate in the synthesis of agrochemical products and disinfectants. The ethanolamine salt of 5-chlorosalicylic acid 4-nitro-2-chloroanilide (Bayluscid, Bayer) is used as an antiparasitic agent to combat schistosomiasis (→ Molluscicides).

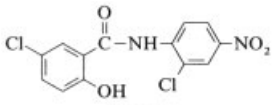

Bayluscid

5-(2,4-Difluorophenyl)salicylic acid [*22494-42-4*] (Diflunisal, Merck), $C_{13}H_8F_2O_3$, M_r 250.20, *mp* 212 – 213 °C.

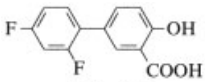

Diflunisal

This compound can be prepared by acetylation of 2,4-difluorobiphenyl, followed by oxidation to 4-acetyloxy-2′,4′-difluorobiphenyl which, when heated under pressure in the presence of potassium carbonate and carbon dioxide, undergoes hydrolysis and carboxylation [37]. Diflunisal is a newly marketed analgesic [38] (→ Analgesics and Antipyretics, Chap. 2.1.; → Antiinflammatory – Antirheumatic Drugs, Chap. 3.1.).

7. Toxicology

Salicylic acid and its derivatives are resorbed by the skin and mucous membranes. Salicylic acid esters dissociate hydrolytically under the influence of esterases. Depending on the pH of the urine, salicylic acid is either oxidized to gentisic acid, or it is eliminated by the kidneys in the form of salicyluretic acid or glucuronide salicylate. Since the elimination rate of salicylic acid is lower than the rate of resorption, in certain cases there is a danger that it may accumulate in the body.

Salicylic acid has keratolytic activity and is a tissue irritant. In the stomach this irritant action affects mainly the mucous-producing cells and striated cells. Long-term treatment with

salicylates slows the rate of blood coagulation by reducing platelet aggregation. Prostaglandin biosynthesis is also inhibited by salicylic acid and its derivatives, which partly explains their anti-inflammatory action. The acid and its derivatives display antipyretic and analgesic effects, and are fungicides and bacteriostatic agents [39].

Symptoms of cutaneous and pulmonary hypersensitivity to salicylates are well known, but it appears that these reactions are not always the result of true allergy. A decrease in the synthesis of prostaglandins and, consequently, other biologically active mediators is evidently involved here as well. Cross-sensitization has been established, for example, between methyl salicylate and acetylsalicylic acid [40–42].

One-time doses in excess of 10 g may prove fatal. The main cause of such *acute* intoxication is the disturbance of the acid – base equilibrium. Severe cases of intoxication may produce delirium, tremor, respiratory insufficiency, sweating, exsiccosis, hyperthermia, or coma. Symptoms of less severe intoxication include hyperventilation, tinnitus, nausea, vomiting, impaired vision and hearing, vertigo, and nervous disorders.

In the case of *chronic* intoxication, symptoms include digestive disorders and gastric and intestinal pain, sometimes with serious hemorrhaging, which nevertheless often remains hidden. Salicylate-induced anaemia, which may be observed in cases of chronic administriation, is an iron deficiency due to hidden bleeding. In the elderly the symptoms of chronic intoxication are often of a neuropsychiatric nature, including confusion and agitation. Despite the very harmful effects of salicylates on tissues, no hepatic or renal problems have been reported to date following chronic administration of pure forms of the compound [34, 43–45].

Experiments with various species of animals have shown that administration of relatively large doses of salicylic acid and its derivatives may have a teratogenic effect. The causative agent is probably salicylic acid. These findings have never been confirmed in humans, however, and there are no grounds for concern assuming the maintenance of appropriate hygiene and proper working conditions [46–49]. In addition to possible local symptoms of irritation due to salicylic acid and its derivatives, the major safety consideration is a risk of allergy. All renewed contact with these products should be avoided by those with a previous history of hyperallergic reactions.

8. References

General References

1. *Beilstein,* **10 H,** 43 ff.; **10 I,** 20 ff.; **10 II,** 25 ff.; **10 III,** 87 ff.; **10 IV.**
2. *Rodd's Chemistry of Carbon Compounds,* **vol. III B,** Elsevier, Amsterdam 1956, pp. 756 – 768.
3. A. S. Lindsey, H. Jeskey: "The Kolbe – Schmitt Reaction," *Chem. Rev.* **57** (1957) 583 – 620.
4. M. J. H. Smith, P. K. Smith: *The Salicylates,* 2nd ed., Wiley, New York 1967.
5. S. Negwer: *Organisch- chemische Arzneimittel und ihre Synonyma,* 5th ed., Akademie-Verlag, Berlin 1978.

Specific References

6. F. T. Welcher: *The Analytical Uses of EDTA,* D. van Nostrand, New York, p. 58.
7. A. Müller, *Org. Synth.* (collect. vol. 2) (1943) 535.
8. J. Pascual et al., *J. Chem. Soc.* 1943, 1949.
9. M. Wesolowski, *Thermochim. Acta* **31** (1979) 133 –146.
10. I. M. Kolthoff, *Pharm. Weekbl.* **58** (1921) 699.
11. J. E. Heestermann, *Chem. Weekbl.* **32** (1935) 463.
12. A. S. Lindsey, H. Jeskey, *Chem. Rev.* **57** (1957) 592 –600.
13. K. Ota, *Mem. Kyushu Inst. Technol. Eng.* **6** (1976) 165 – 175.
14. A. J. Rostron, A. M. Spivey, *J. Chem. Soc.* 1964, 39.
15. Ueno, EP 254 596, 1986.
16. BIOS, no. 116, 664, 1246; CIOS, XXIII-25; FIAT, no. 744. W. L. Hardy, *Ind. Eng. Chem.* **49** (1957) no. 6, 55 A.
17. Bayer, DE-OS 2 644 318, 1976.
18. Wacker, DE 624 318, 1931. Bayer, DE 955 598, 1954. Dow, DE-AS 1 215 173, 1961.
19. Monsanto, US 2 824 892, 1956.
20. Ilford, GB 638 196, 1948.
21. J. Hirao, Y. Hara, *Yuki Gosei Kagaku Kyokaishi* **25** (1967) 577 – 581.
22. Ueno, JP 76 23 494, 1968.
23. Asahi, US 3 360 553, 1967.
24. Dow, BE 665 631, 1965.
25. W. W. Kaeding, A. T. Shulgin, *J. Org. Chem.* **27** (1962) 3551.

26. Sun Oil, US 3 183, 1963.
27. Monsanto, GB 353 921, 1930.
28. BIOS, no. 1141.
29. Dow, US 4 137 258, 1978.
30. Pharmacopée Européenne, Monography 366, 1985.
31. McKetta: *Encyclopedia of Chemical Processing and Design,* **vol. 4,** Dekker, New York 1977, p. 28.
32. E. R. Kline, *J. Chem. Educ.* **19** (1942) 332.
33. C. F. H. Allen, J. van Allan, *Org. Synth.* **(collect. vol. 3)** (1955) 765.
34. N. W. Hirwe et al., *Proc. Indian Acad. Sci. Sect. A* **8** (1938) 208.
35. J. T. Sheehan, *J. Am. Chem. Soc.* **70** (1948) 1665.
36. E. Plazek, *Rocz. Chem.* **10** (1930) 761.
37. Merck, US 4 131 618, 1977; DE-OS 2 532 559, 1975.
38. *Chem. Eng. News* **56** (1978) April 10, 7; **58** (1980) April 14, 24.
39. W. Forth, D. Henschler, W. Rummel: *Pharmakologie und Toxikologie,* Bibliographisches Institut, Mannheim 1975, p. 425.
40. A. Szczeklik et al. *J. Allergy Clin. Immunol.* **60** (1977) 276.
41. C. Hindson, *Contact Dermatitis* **3** (1977) 348.
42. P. Kallos, H. D. Schlumberger, *J. Pharm. Pharmacol.* **30** (1978) 67.
43. L. Lagplante, C. Beaudry, *Union Med. Can.* **108** (1979) no. 3, 280.
44. W. Siegenthaler et al., *Med. Trib. (German ed.)* **13** (1978) no. 18, 66.
45. R. Ludewig, K. Lohs: *Akute Vergiftungen,* G. Fischer Verlag, Stuttgart 1979, p. 381.
46. K. T. Szabo et al., *Toxicol. Appl. Pharmacol.* **19** (1971) 371 – 372.
47. C. A. Kimmel et al., *Teratology* **4** (1971) 15 – 24.
48. A. B. G. Lansdown, P. Grasso, *Experienta* **25** (1969) 885 – 887.
49. I. D. G. Richards, *Br. J. Prev. Soc. Med.* **23** (1969) no. 4, 218 – 225.

Salpeter → Nitrates and Nitrites
Saluretics → Diuretics
Samarium → Rare Earth Elements

Saponins

Michael W. Schwarz, E. Merck OHG, Darmstadt, Federal Republic of Germany

1. **Introduction** 1
1.1. **General Properties** 2
1.2. **Distribution** 2
1.3. **Isolation** 2
1.4. **Structure Elucidation** 3
1.5. **Analysis** 3
2. **Pharmacology** 3
3. **Plant Saponins** 4
3.1. **Steroid Saponins** 4
3.2. **Glycoalkaloids** 7
3.3. **Triterpene Saponins** 9
4. **Animal Saponins** 12
4.1. **Asterosaponins** 12
4.2. **Holothurins** 13
5. **References** 13

Abbreviations

api	β-D-apiose
ara	α-L-arabinose
drib	β-D-2′-deoxyribose
f	furanoside
fuc	β-D-fucose
fru	β-D-fructose
gal	β-D-galactose
galA	β-D-galacturonic acid
glc	β-D-glucose
glcA	β-D-glucuronic acid
qui	β-D-quinovose
rha	α-L-rhamnose
xyl	β-D-xylose

1. Introduction

Saponins are glycosides (see → Carbohydrates) occurring primarily in plants but also in starfish (*Asteroidea*) and sea cucumbers (*Holothuridea*). Properties generally considered to be shared by this group of natural products are surfactant activity, hemolytic action, steroid- complexing ability, and biocidal capability.

The characteristic soapy lather formed when saponin- containing plant extracts are agitated in water provides this group of secondary metabolites with its common name (Latin *sapo* = soap), although this property is shared with structurally related compounds (→ Cardiac Glycosides). Other properties are characteristic of particular types of saponins rather than all members of the class.

Saponins are categorized according to the structure of the aglycone moiety (*sapogenin*) and the number of linked sugar chains. Aglycones can be divided into triterpenoid and steroid sapogenins. Steroid saponins with basic properties are termed glycoalkaloids.

Hexoses common in all saponins are β-D-glucose (glc), β-D-galactose (gal), and α-L-rhamnose (rha); pentoses are α-L-arabinose (ara) and β-D-xylose (xyl). Pentoses occur primarily as pyranosides, less frequently as furanosides (f).

Saponins linked to one or two sugar chains are called *monodesmosides* and *bisdesmosides*, respectively (Greek *desmos* = chain). Trisdesmosidic saponins are rare [1,2]. Mild hydrolysis of saponins yields *prosapogenins*, neutral glycosides with less than three sugar units, or acidic or basic glycosides with one sugar unit. Prosapogenins are less soluble than saponins and do not have typical saponin properties.

1.1. General Properties

Surfactant Activity. The well-known ability of saponins to cause frothing has long been used to detect these materials in plants or plant extracts. The combination of a hydrophobic aglycone and hydrophilic substituents accounts for the amphiphilic nature of saponins. Increased length and branching of the sugar chain correlate with enhanced surfactant power. Hence, triterpenoidal bisdesmosides are more active in this sense than monodesmosides [3].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_485

Hemolytic Activity. Saponins vary considerably in their ability to lyse erythrocytes. Different glycosides of the same sapogenin differ in activity as a function of the type and structure of substituents [4,5]. Bisdesmosides are generally less potent than monodesmosides. Introducing polar hydroxyl groups into the aglycone decreases activity, whereas esterification restores it. The linkage position of the sugar chain in triterpenoids also influences reactivity; thus, C-3 glycosides are more reactive than saponins with a C-28 glycosidic substituent.

Steroid-Complexing Ability. Saponins form insoluble complexes with cholesterol [6] and other sterols. This affinity is more pronounced with steroid saponins and glycoalkaloids than with triterpenoids.

Biocidal Activity. Monodesmosidic saponins exhibit fungitoxic or fungistatic and weak antimicrobial activity [7]. Examples include the specific action of α-tomatine [8] and the oat (*Avena sativa*) root saponin avenacin [9]. Steroid saponins and glycoalkaloids generally demonstrate more pronounced action, whereas triterpenes exhibit a broader spectrum of effectiveness.

The permeabilization of crucial cell membranes probably explains the toxicity of saponins to mollusks [10] and other invertebrates [11]. Ecdysterone-like activity [12] and the repellent or antifeedant action of saponins against termites [13], leaf- cutting ants [14], and the potato beetle [15] have been reported. Aquatic vertebrates such as fish and tadpoles are affected through permeabilization of their gills [16], resulting in a rapid loss of physiological function.

1.2. Distribution

Saponins occur mainly in plants of the subdivision *Angiospermae* of the division *Spermatophytae*. Steroid saponins are restricted mostly to the class *Monocotyledoneae*, a known exception being the saponins of foxglove (*Digitalis* sp.). Triterpene and glycoalkaloid saponins are found primarily in plants of the class *Dicotyledoneae*, a more recently discovered exception being saponins isolated from montbretia (*Crocosmia crocosmiiflora*) [17]. Certain families of the class *Dicotyledoneae* (e.g., *Solanaceae, Hippocastanaceae, Primulaceae, Rosaceae,* and *Caryophyllaceae*) are especially rich in saponin- containing genera.

Monodesmosides are concentrated in the outer tissues of seeds, roots, and bark, consistent with their function as a barrier against microorganisms, whereas bisdesmosides are more abundant in leaves and stems. Secretion of proteolytic enzymes by fungal hyphae triggers conversion of the more soluble (and, therefore, transportable) bisdesmosides by specific glycosidases [18] into the more active monodesmosidic saponins.

Animal saponins are restricted to representatives of the phylum *Echinodermata*. Species of the class Holothuridea (sea cucumbers) contain triterpene saponins, whereas saponins from species of the class *Asteroidea* (starfish) contain steroids. These saponins induce avoidance responses in susceptible predator species [19].

1.3. Isolation

Methods for saponin isolation and purification are covered in several reviews [20–22]. Ground dried or fresh specimens are extracted after optional defatting with solvents such as methanol, ethanol, their aqueous mixtures, or, in the case of glycoalkaloids, slightly acidified water. Formation of artifacts due to enzymatic or solvolytic cleavage of glycosidic bonds or acid- catalyzed elimination and rearrangement of the aglycone must be prevented by suitable choice of conditions.

Distribution of a crude extract between water and butanol separates the saponins from such water-soluble compounds as oligosaccharides. Precipitation with organic solvents (e.g., ether or acetone) or complexing agents (e.g., cholesterol) permits further concentration.

Saponins with free carboxylic acid groups can be purified by ion-exchange chromatography. For cases in which further enrichment by crystallization is not possible, chromatographic techniques are used for the separation of crude saponin mixtures. Methods such as high-performance thin-layer chromatography (HPTLC), high-performance liquid chromatography (HPLC), and droplet countercurrent

chromatography (DCCC) [23] have gained wide acceptance. Methods have been reported for isolating saponins from roots of ginseng (*Panax schinseng*) [24] and *Bupleurum falcatum* [25] by preparative HPLC of crude extracts.

1.4. Structure Elucidation

Historically, saponin structures were deduced after methylation and acid hydrolysis to aglycones and partially methylated sugars. However, this method leads to artifacts, so milder hydrolysis methods have been developed. Relatively large amounts of purified material are still needed for classical structural analysis.

The advent of modern spectroscopic methods has simplified the task. Fast-atom bombardment mass spectrometry (FAB – MS) facilitates the sequencing of sugar chains, while ^{1}H and ^{13}C NMR [27] techniques are used to establish specific linkages and resolve ambiguities. Structures of saponins from the starfish *Asterias forbesii* [28], *Astragalus ernestii* [29], and *Allium giganteum* [30] were elucidated by modern two-dimensional NMR methods.

1.5. Analysis

Analytical procedures must be adapted to the particular application in question. Crude estimates of saponin content are obtained via gravimetric methods [31]. Most technical-grade saponins are mixtures rather than single substances. Quantitative estimation of their makeup is accomplished by measuring the surface tension of solutions, their foaming power, and their hemolytic activity [32].

Saponins or saponin mixtures for pharmaceutical applications require analytical methods more specific to particular active compounds, such as spectrophotometric quantitation of a saponin color reaction [33, 34]. Combinations of methods including thin-layer chromatography – densitometry (TLC – DM) [35], gas chromatography – mass spectrometry (GC – MS) [36], LC – MS [37], HPLC [38], and radioimmunoassay (RIA) [39], have been reported for ginseng saponins and other compounds.

2. Pharmacology

Toxicology. In contrast to poikilothermic animals, saponin toxicity to homeothermic species by oral intake is quite low (50 – 100 mg/kg). The significance of saponins as components of the human diet or of animal feeds has been reviewed extensively [20]. Saponins exhibit different spectra of toxicity on parenteral or intravenous application (LD_{50} 0.7 – 50 mg/kg). The hemolytic action of many saponins prevents their intravenous administration, although detoxification by complex formation with serum cholesterol, albumin, or other plasma constituents reduces this effect.

Membranolytic saponins combine irreversibly with cell membrane systems and produce lesions [40] with a pore diameter of about 8 nm. Various structural models for the complex have been proposed [40, 41]. The role of cholesterol as the primary binding site for saponins in cell membranes has been challenged [42], and various saponins have been shown to interact in different ways with erythrocytes and liposomal membranes with respect to cholesterol, phosphatidylcholine, and distearoyllecithin [43]. The theory that β-glucosidase-induced hydrolysis of the glycosidic bond is an essential step in membranolysis [4] cannot explain changes induced by saponins in the osmotic behavior of liposomes devoid of β-glucosidase [44]. A unifying theory of saponin action on membranes has still to be developed.

Pharmacology. Claims of therapeutic activity for saponins are numerous, with the result that saponins have been denounced by some as panaceas [45].

Resorption of saponins by the small intestine is generally low [46], but enzymatic or bacterial decomposition in the large intestine [47] has been demonstrated. Permeabilization of cells of the small intestinal mucosa [48] reduces their capacity to transport nutrients and increases secretion. A therapeutic benefit is increased secretion in the nasal pharyngeal cavity after administration of cough syrup containing extracts of roots from *Primula officinalis*, *Glycyrrhiza glabra,* and *Gala senega*.

Anti-inflammatory activity has been reported for saponins from the seed of the horse chestnut (*Aesculus hippocastanum*) [49] and *Panax*

inseng (Greek *pan* = all, *axos* = cure) [50,51]. In the former case, pharmacological action was blocked by adrenalectomy, hypophysectomy, or sympatholytic drugs. Antiexudative and antigranulomatous action was explained in terms of activation of corticosterone secretion.

The saponin from *Glycyrrhiza glabra* and the hemisuccinate of the corresponding aglycone (carbenoxolone) are used to treat gastric ulcers [52]. Chemical modifications of the aglycone have been reported to reduce side effects [53]. The ammonium salt of the aglycone is used as an antiallergic agent, and the same effect is reported for saponins from *Ilex crenata* [54].

Hypocholesterolemic effects of saponins might be due to a reduction in the uptake of dietary cholesterol by complex formation, increasing bile salt excretion, or direct interaction with sterols in the membranes of mucosal cells [20]. Active saponins have been found in soybean (*Glycine max*) [55] and *Quillaja saponaria* [56].

Psychotropic effects are reported for ginseng saponins [57], and patents have been filed for analgesic, sedative, and cardiovascular-activating effects [58] as well as antitumor activity [59].

Antiviral activity of saponins has been reported against the herpes- and poliovirus [60], Epstein – Barr virus [61], and human immunodeficiency virus (HIV-1) [62].

Bisdesmosidic saponins solubilize monodesmosides [63], an adjuvant effect that seems to increase immune response to vaccines [64] and the uptake of β-lactam antibiotics [65] after oral administration.

3. Plant Saponins

3.1. Steroid Saponins

Aglycones of monodesmosidic steroid saponins from plants consist of a hexacyclic spirostanol system, whereas bisdesmosides are derived from the pentacyclic furostanol system or a hexacyclic structure such as that in nuatigenin. Hydrolysis of C-26 glycosidic linkages of furostanol sapogenins results in cyclization to spirostanols. Therefore, furostanol saponins carry the designation "proto-" prefixed to the name of the corresponding spirostanol saponin. Especially in the older literature, the suffix "-oside" provides another indication of a furostanol saponin. Nuatigenin saponins rearrange to the spirostanol derivative isonuatigenin on acid hydrolysis [66].

Aglycones are named after the plant sources from which they were first isolated. C-25 (*S*)-epimers of known C-25 (*R*)-sapogenins are assigned the prefix "neo-", whereas C-25 (*R*)-epimers of known C-25 (*S*)-sapogenins are given the prefix "iso-". Epimers with a C-3 α-OH instead of a C-3 β-OH are designated by the prefix "epi-".

Monodesmosides have linear or branched sugar chains linked in most cases to the C-3 β-OH group, whereas furostanol bisdesmosides display an additional glycosidic linkage at C-26. Besides the common sugars listed previously, less prevalent sugars include D-fucose (fuc), β-D-2′-deoxyribose (drib), and β-D-apiose [api(f)]. For a review of steroid saponins, see [67].

Spirostanol Saponins. Figure 1 provides the structures for a number of spirostanol aglycones. Most spirostanol saponins are monodesmosides with a single glycosidic bond at the C-3 β-OH. This is also true of hydroxylated sapogenins such as digitogenin (**15**) and paniculogenin (**14**). Monodesmosidic steroid saponins exhibit typical saponin properties, such as strong hemolytic activity.

Important saponins derived from diosgenin (**1**; $\rightarrow$ Hormones) and its C-25 (*S*)-epimer yamogenin (**2**) are dioscin [rha1 $\rightarrow$ 2(rha1 $\rightarrow$ 4glc)-**1**] (see Fig. 2 for an example of the signifance of the abbreviated notation used here and elsewhere) and gracillin [rha1 $\rightarrow$ 2(glc1 $\rightarrow$3glc)-**1**], two glycosides first isolated from wild yam (*Dioscorea* sp.). Dioscin has been reported in a number of plant sources, including palms [68]. The aglycones are important sources for steroid synthesis, as are hecogenin (**12**), smilagenin (**16**), and its epimer sarsasapogenin (**17**). The latter were first isolated from *Agave* sp. and *Smilax crenata*, respectively. Sarsasapogenin (**17**) yields the monodesmoside parillin, glc1 $\rightarrow$ 2[rha1 $\rightarrow$ 4 (glc1 $\rightarrow$ 6glc)]-**17**, and the bisdesmoside sarsaparilloside (**18**).

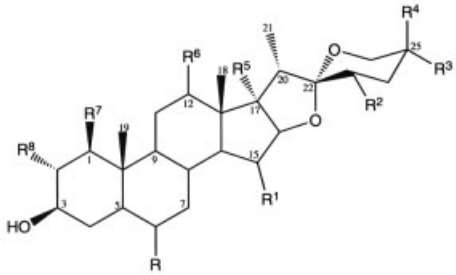

Sapogenin	CAS registry no.	Δ	C-5 H	R	R^1	R^2	R^3	R^4	R^5	R^6	R^7	R^8
Diosgenin (**1**)	[*512-04-9*]	5		H	H	H	CH_3	H	H	H	H	H
Yamogenin (**2**)	[*512-06-1*]	5		H	H	H	H	CH_3	H	H	H	H
Pennogenin (**3**)	[*507-89-1*]	5		H	H	H	CH_3	H	OH	H	H	H
Ruscogenin (**4**)	[*472-11-7*]	5		H	H	H	CH_3	H	H	H	OH	H
Yuccagenin (**5**)	[*511-97-7*]	5		H	H	H	CH_3	H	H	H	H	OH
Kammagenin (**6**)	[*564-44-3*]	5		H	H	H	CH_3	H	H	O=	H	OH
Isonuatigenin (**7**)	[*7050-41-1*]	5		H	H	H	OH	CH_3	H	H	H	H
Tigogenin (**8**)	[*77-60-1*]		α	H	H	H	CH_3	H	H	H	H	H
Neotigogenin (**9**)	[*470-01-9*]		α	H	H	H	H	CH_3	H	H	H	H
Chlorogenin (**10**)	[*562-34-5*]		α	α-OH	H	H	CH_3	H	H	H	H	H
Neochlorogenin (**11**)	[*511-91-1*]		α	α-OH	H	H	H	CH_3	H	H	H	H
Hecogenin (**12**)	[*467-55-0*]		α	H	H	H	CH_3	H	H	O=	H	H
Gitogenin (**13**)	[*511-96-6*]		α	H	H	H	CH_3	H	H	H	H	OH
Paniculogenin (**14**)	[*16750-37-1*]		α	α-OH	H	OH	H	CH_3	H	H	H	H
Digitogenin (**15**)	[*511-34-2*]		α	H	β-OH	H	CH_3	H	H	H	H	OH
Smilagenin (**16**)	[*126-18-1*]		β	H	H	H	CH_3	H	H	H	H	H
Sarsasapogenin (**17**)	[*126-19-2*]		β	H	H	H	H	CH_3	H	H	H	H

Figure 1. Spirostanol-type sapogenins

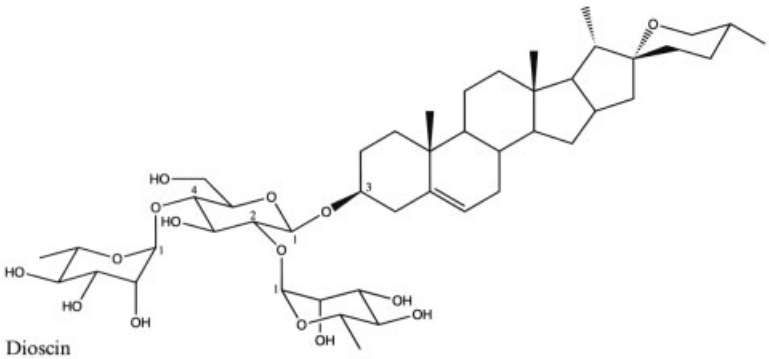

Dioscin

rha1 → 2(rha1 → 4glc)−3−O−diosgenin

rha1 → 2(rha1 → 4glc)−**1**

rha1

$^{4}_{2}$glc → diosgenin

rha1

Figure 2. Alternative notational schemes for saponins, using dioscin as an example

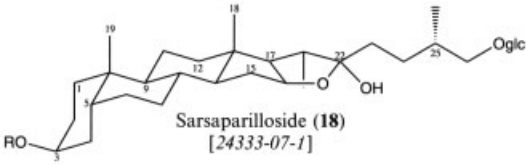

A number of steroid saponins have been isolated from foxglove (*Digitalis purpurea, D. lanata*), of which digitonin, glc1 → 3gal1 → 2 (xyl1 → 3glc)1 → 4gal-**15**, and its aglycone digitogenin (**15**) are the most important. Digitonin forms very insoluble complexes with cholesterol, and has been used as a reagent for the determination of this sterol. Minor sapogenins in *Digitalis* sp. are gitogenin (**13**), tigogenin (**8**), and its epimer neotigogenin (**9**). The saponin glc1 → 3glc-**13** was isolated from Agave cantala [69].

Together with diosgenin (**1**) derivatives, the pennogenin (**3**) saponin dioscinin, rha1 → 4rha1 → 4(rha1 → 2glc)-**3** was isolated from the palm *Trachycarpus wagnerianus* [70]. Another saponin was isolated from *Dracaena mannii* [71] and characterized as rha1 → 2rha1 → 3glc-**3**, showing antimicrobial activity.

Saponins of ruscogenin (**4**) were isolated from *Ophiopogon* sp. [72] and identified as ophiopogonin B (rha1 → 2fuc-1-O-**4**) and D [rha1 → 2(xyl1 → 3fuc)-1-*O*-**4**]. In both cases the sugar chain is attached to the C-1-hydroxyl group.

Chlorogenin (**10**) was first isolated from *Chlorogalum pomeridianum*. Its saponins were used historically as a poison to stun fish, while at the same time leaving them edible by humans.

Furostanol Saponins. As already mentioned, furostanol saponins are readily converted to spirostanol saponins by dilute acid or β-glucosidases.

Leaves and stems often yield the protosaponins of spirostanols isolated from other parts of the plant.

The hydroxyl group at C-22 can be displaced by alkyl alcohols to yield alkylprotosaponins [73], or eliminated with acetic acid. The resulting $\Delta^{20,22}$-furostene saponins are termed pseudoprotosaponins. The isolation of furostene bisdesmosides [69] represents the first report of a genuine pseudoprotosaponin in nature.

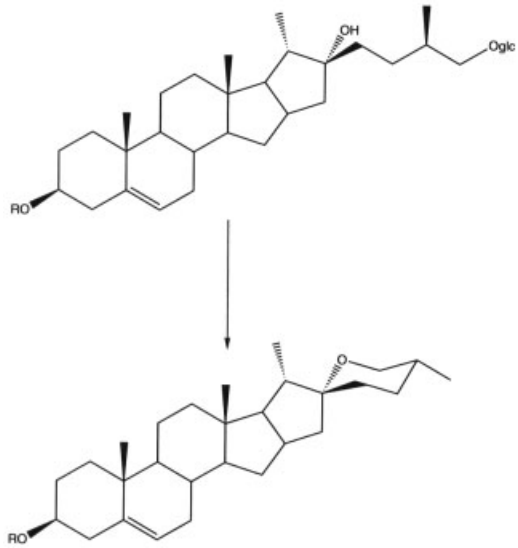

Furostanols are easily distinguished from spirostanol saponins with Ehrlich's reagent [74], and also by IR spectroscopy, since furostanols lack the characteristic absorption bands of a C-22 spiroketal moiety.

The best known example of a bisdesmosidic steroid saponin is sarsaparilloside (**18**) (R = glc1 → 2[rha1 → 4(glc1 → 6glc)] –) from sarsaparilla, the dried root of *Smilax aristolochiaefolia*, used as a flavoring in beverages.

Furostanol glycosides isolated from garlic (*Allium sativum*) [75] were shown to be derived from chlorogenin (**10**). Mild hydrolysis yields antifungal spirostanol saponins.

Furostanols of tigogenin (**8**) were isolated from *Nicotiana tabacum* [76], of sarsasapogenin (**17**) from *Asparagus sp.* [77], of diosgenin (**1**) from *Balanites aegyptiaca* [78], and of gitogenin (**13**) from *Trigonella foenum-graecum* [79]. Trisdesmosidic furostanol glycosides have been reported for ruscogenin (**4**) saponins [72].

Avenacosides A and B were isolated from oat seeds [80] and identified as glc1 → 2(rha1 → 4glc)- and glc1 → 3glc1 → 2(rha1 → 4glc)-glycosides of nuatigenin-26-glucopyranoside (**19**), respectively. A highly specific β-glucosidase (avenacosidase) [81] yields the antifungal 26-desglucoavenacosides. Nuatigenin saponins from *Solanum aculeatissimum* have recently been proposed as a starting material for steroid synthesis [82].

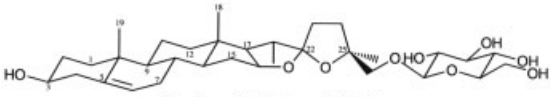

Nuatigenin-26-glucoside (**19**)

Other Steroid Systems. Saponins from *Polypodium vulgare* are derived from a rare type of cholestanone system with a hydroxytetrahydropyran ring at C-17 [83].

Uses. Spirostanol and furostanol saponins of diosgenin (**1**) and its epimer yamogenin (**2**) were the first natural starting materials for the industrial synthesis of pregnenolone and progesterone (→ Hormones, Chap. 3.2.). Hecogenin (**12**), smilagenin (**16**), and sarsasapogenin (**17**) are also used as educts for steroid hormones. However, the growth in demand could not be satisfied by saponins alone, so the microbiological conversion of sterols such as stigmasterol to steroid precursors and total synthesis have both become important alternatives.

3.2. Glycoalkaloids (→ Alkaloids)

Glycoalkaloids are limited principally to the genera *Solanum, Lycopersicum, Cestrum*, and *Veratrum* of the family Solanaceae. They share a common C_{27} cholestane structure but differ in the structures of their side chains. Because of the presence of these materials in cultivated plants [84, 85], the corresponding structures have been known for some time.

Nitrogen analogues of the spirostanol steroid saponins, with nitrogen replacing an oxygen in the spiroketal ring, are called spirosolanes. Depending on the stereochemistry at C-22 they belong to either the solasodane (22*R*) or the tomatidane (22*S*) group. Analogues in which the 3-β-hydroxyl group has been replaced by an amino group are called aminospirostanes.

Solanidanes and *solanocapsines* have contiguous six-ring systems, whereas epinitrilocholestanes have a tetrahydropyridine ring bound to C-20 of a cholestane structure. Glycosides of the spirosolane and solanidane types are bound exclusively through the C-3 β-OH group.

The spirosolane-type aglycone solasodine (**20**) and its saponin α-solamargine (rha1 → 2 (rha1 → 4glc)-**20**) are analogues of diosgenin (**1**) and dioscin, respectively. Soladulcidine (**21**) and 15-hydroxysoladulcidine (**22**) correspond to tigogenin (**8**) and chlorogenin (**10**), respectively (cf. Fig. 3).

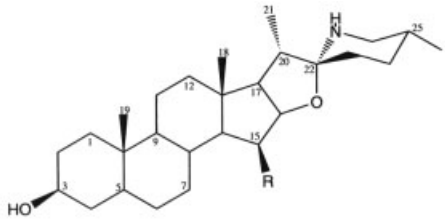

Sapogenin	CAS registry no.	Δ	C-5 H	R
Solasodine (**20**)	[*126-17-0*]	5		H
Soladulcidine (**21**)	[*511-98-8*]		α	H
Hydroxysoladulcidine (**22**)	[*4912-43-0*]		α	OH

Figure 3. Solasodane-type glycoalkaloids

Solasodine (**20**) can be converted by *N*-acetylation and subsequent rearrangement to a $\Delta^{20,22}$-furostene, useful as a precursor in steroid synthesis [86]. This reaction is also feasible with the C-22 epimers tomatidine (**23**) and tomatidenol (**24**). Saponins such as α-tomatine (gal1 → 2(xyl1 → 3glc)1 → 4gal-**23**) can be used as natural sources for steroid precursors. The distribution of α-tomatine and its significance have been reviewed [87].

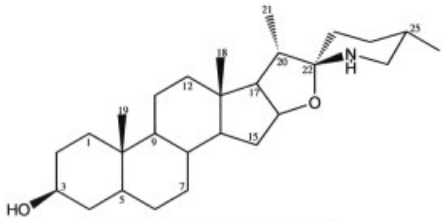

Tomatidine (**23**) [*77-59-8*]
Δ^5, tomatidenol (**24**) [*546-40-7*]

The glycoalkaloid solanine was found in the cultivated potato (*Solanum tuberosum*) and identified as consisting of saponins derived from solanidine (**25**), namely, α-chaconine (rha1 → 2(rha1 → 4glc)-**25**) and α-solanine (rha1 → 2(glc1 → 3gal)-**25**) [88]. Other sapogenins found in *Solanum* sp. (see

Fig. 4) are leptinidine (**26**) [89] and demissidine (**27**), with the saponins demissine (glc1 → 2(xyl1 → 3glc)1 → 4gal-**27**) and commersonine [glc1 → 2(glc1 → 3glc)1 → 4gal-**27**].Chemical degradation of solanidanes to steroids has not been achieved [90].

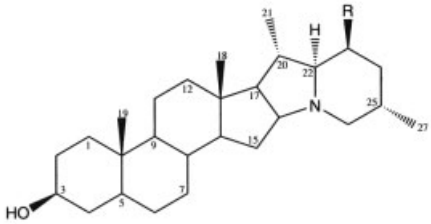

Sapogenin	CAS registry no.	Δ	C-5 H	R
Solanidine (**25**)	[*80-78-4*]	5		H
Leptinidine (**26**)	[*2448-17-1*]	5		OH
Demissidine (**27**)	[*474-08-8*]		α	OH

Figure 4. Solanidane-type glycoalkaloids

Solanocapsine (**28**) has been isolated as a sapogenin from *S. pseudicapsicum* [91]; no glycoside has yet been reported.

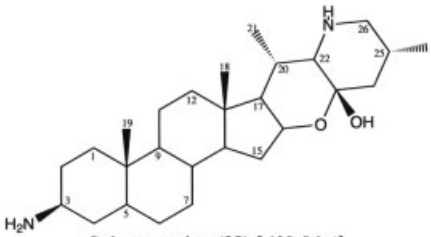

Solanocapsine (**28**) [*639-86-1*]

Natural epinitrilocholestanes (Fig. 5) include verazine (**29**) and tomatillidine (**30**). Intermediates in the conversion of spirosolanes to steroids [92] and in the transformation of spirosolanes to solanidanes [93] have similar structures.

The aminospirostanes jurubidine (**31**) and paniculidine (**32**) from *S. paniculatum* [94] are sapogenins of furostane glycosides such as jurubine (**33**).

Uses. The spirosolanes solasodine (**20**), tomatidine (**23**), and tomatidenol (**24**) serve as natural sources for steroid synthesis. The use of enzymes [95] for glycolysis of solasodine saponins produced by plant cell cultures has been proposed [96].

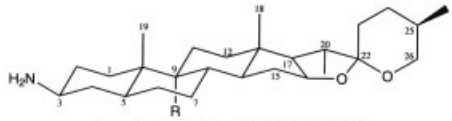

R = H, jurubidine (**31**) [*6084-44-2*]
R = OH, paniculidine (**32**) [*138464-37-6*]

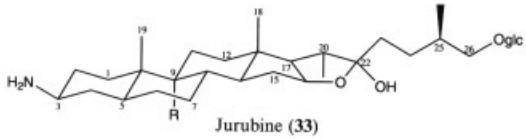

Jurubine (**33**)

Other glycoalkaloids or sapogenins have gained little commercial interest.

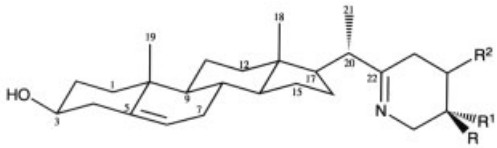

Sapogenin	CAS registry no.	Δ	R	R^1	R^2
Verazine (**29**)	[*14320-81-1*]	5	H	CH_3	H
Tomatillidine (**30**)	[*986-45-8*]	5	CH_3	H	O=

Figure 5. Epinitrilocholestane-type glycoalkaloids

3.3. Triterpene Saponins

Triterpene glycosides are the most common saponins in nature, and they have been reviewed extensively [20–22]. A derivative of triterpenoids was found in tertiary sediments [97].

Monodesmosides have linear or branched sugar chains, usually linked to the C-3 β-OH group. The common sugars listed previously are supplemented by such less frequently encountered sugars as β-D-fucose (fuc), β-D-fructose [fru(f)], β-D-quinovose (qui), β-D-glucuronic acid (glcA), and β-D-galacturonic acid (galA).

Oleananes. Saponins derived from β-amyrin (**34**) are the most common triterpene glycosides (cf. Fig. 6).

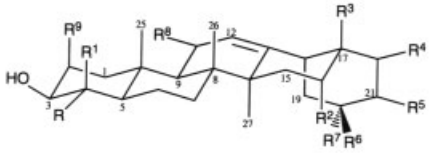

Sapogenin	CAS registry no.	R	R^1	R^2	R^3	R^4	R^5	R^6	R^7	R^8	R^9
β-Amyrin (**34**)	[*559-70-6*]	CH_3	CH_3	H	CH_3	H	H	CH_3	CH_3	H	H
Sophoradiol (**35**)	[*6822-47-5*]	CH_3	CH_3	H	CH_3	α-OH	H	CH_3	CH_3	H	H
Canoniensistriol (**36**)	[*83718-68-7*]	CH_3	CH_3	H	CH_3	α-OH	α-OH	CH_3	CH_3	H	H
Glycyrrhetol (**37**)	[*14226-18-7*]	CH_3	CH_3	H	CH_3	H	H	CH_3	CH_2OH	O=	H
Glycyrrhetic acid (**38**)	[*471-53-4*]	CH_3	CH_3	H	CH_3	H	H	CH_3	COOH	O=	H
Liquiritic acid (**39**)	[*10379-72-3*]	CH_3	CH_3	H	CH_3	H	H	COOH	CH_3	O=	H
Epikatonic acid (**40**)	[*76035-62-6*]	CH_3	CH_3	H	CH_3	H	H	COOH	CH_3	H	H
Liquiridiolic acid (**41**)	[*20528-70-5*]	CH_3	CH_2OH	H	CH_3	H	β-OH	COOH	CH_3	H	H
Melilotigenin (**42**)	[*114702-59-9*]	CH_3	CH_2OH	H	CH_3	O=	H	COOH	CH_3	H	H
Soyasapogenol B (**43**)	[*595-15-3*]	CH_3	CH_2OH	H	CH_3	α-OH	H	CH_3	CH_3	H	H
Soyasapogenol E (**44**)	[*6750-59-0*]	CH_3	CH_2OH	H	CH_3	O=	H	CH_3	CH_3	H	H
Soyasapogenol A (**45**)	[*508-01-0*]	CH_3	CH_2OH	H	CH_3	α-OH	α-OH	CH_3	CH_3	H	H
Azukisapogenol (**46**)	[*86425-21-0*]	CH_3	CH_2OH	H	CH_3	H	H	COOH	CH_3	H	H
Protoescigenin (**47**)	[*20853-07-0*]	CH_3	CH_2OH	OH	CH_2OH	β-OH	α-OH	CH_3	CH_3	H	H
Primulagenin A (**48**)	[*465-95-2*]	CH_3	CH_3	OH	CH_2OH	H	H	CH_3	CH_3	H	H
Theasapogenol D (**49**)	[*53227-91-1*]	CH_3	CH_3	OH	CH_2OH	β-OH	H	CH_3	CH_3	H	H
Theasapogenol B (**50**)	[*13844-01-4*]	CH_3	CH_3	OH	CH_2OH	β-OH	α-OH	CH_3	CH_3	H	H
Theasapogenol C (**51**)	[*14440-27.8*]	CH_2OH	CH_3	OH	CH_2OH	β-OH	H	CH_3	CH_3	H	H
Theasapogenol A (**52**)	[*13844-22-9*]	CH_2OH	CH_3	OH	CH_2OH	β-OH	α-OH	CH_3	CH_3	H	H
Theasapogenol E (**53**)	[*15399-41-4*]	CHO	CH_3	OH	CH_2OH	β-OH	α-OH	CH_3	CH_3	H	H
Oleanolic acid (**54**)	[*508-02-1*]	CH_3	CH_3	H	COOH	H	α-H	CH_3	CH_3	H	H
Echinocystic acid (**55**)	[*510-30-5*]	CH_3	CH_3	OH	COOH	H	H	CH_3	CH_3	H	H
Entagenic acid (**56**)	[*5951-41-7*]	CH_3	CH_3	H	COOH	β-OH	β-OH	CH_3	CH_3	H	H
Serratagenic acid (**57**)	[*6488-64-8*]	CH_3	CH_3	H	COOH	H	H	COOH	CH_3	H	H
Sperculagenic acid (**58**)	[*18671-48-2*]	CH_3	CH_3	H	COOH	H	H	CH_3	COOH	H	H
Hederagenin (**59**)	[*465-99-6*]	CH_2OH	CH_3	H	COOH	H	H	CH_3	CH_3	H	H
Bayogenin (**60**)	[*6989-24-8*]	CH_2OH	CH_3	H	COOH	H	H	CH_3	CH_3	H	OH
Polygalacic acid (**61**)	[*22338-71-2*]	CH_2OH	CH_3	OH	COOH	H	H	CH_3	CH_3	H	OH
Platycodigenin (**62**)	[*22327-82-8*]	CH_2OH	CH_2OH	OH	COOH	H	H	CH_3	CH_3	H	OH
Esculentic acid (**63**)	[*56283-68-2*]	CH_2OH	CH_3	H	COOH	H	H	CH_3	COOH	H	H
Jaligonic acid (**64**)	[*51776-39-7*]	CH_2OH	CH_3	H	COOH	H	H	CH_3	COOH	H	OH
Gypsogenin (**65**)	[*639-14-5*]	CHO	CH_3	H	COOH	H	H	CH_3	CH_3	H	H
Quillaic acid (**66**)	[*631-01-6*]	CHO	CH_3	OH	COOH	H	H	CH_3	CH_3	H	H
Dianic acid (**67**)	[*91652-29-8*]	CHO	CH_3	H	COOH	H	H	CH_2OH	CH_3	H	H
Gypsogenic acid (**68**)	[*5143-05-5*]	COOH	CH_3	H	COOH	H	H	CH_3	CH_3	H	H
Medicagenic acid (**69**)	[*599-07-5*]	COOH	CH_3	H	COOH	H	H	CH_3	CH_3	H	OH

Figure 6. Oleanane-type sapogenins

The variability of substituents is immense. The most frequently hydroxylated carbon centers of the ring system are C-2, C-16, C-21, and C-22. Saponins of hydroxylated aglycones are esterified with organic acids (e.g., acetic, butyric, 2-methylbutyric, tiglic, angelic, and benzoic acids). One example is the principal saponin glc1 → 2(glc1 → 4glcA)-**47** of horse chestnut, which is esterified at C-22 with acetic acid and at C-21 with tiglic or angelic acid. The complete mixture of saponins is called escin.

Of the methyl substituents, C-23, C-24, C-28, C-29, and C-30 are most frequently oxidized to hydroxymethyl, aldehyde, and carboxylic acid groups. Sapogenins with a C-28 carboxylic acid function are present in plant leaves and stems in the form of bisdesmosides with an acyl – glycosidic linkage to a second sugar chain. Enzymatic cleavage of the acyl – glycosidic bond yields the biologically more active monodesmosides. One example is found in a report on the isolation of two oleanolic acid (**54**) saponins from *Momordica cochinchinensis* [98], momordins Id [xyl1 → 2 (xyl1 → 3glcA)-3-*O*-**54**] and IId [xyl1 → 2(xyl1 → 3glcA)-3-*O*-**54**-28-*O*-glc]. Quinoside A from *Chenopodium quinoa* has been reported [2] to be a trisdesmoside of hederagenin (**59**), 3,23-bis (glc)-**59**-28-*O*-ara3 ← 1glc.

Saponins from licorice, the extract of *Glycyrrhiza glabra*, have a keto group at C-11. The resulting enone structure facilitates UV detection during HPLC separation. The composition, uses, and analysis of licorice have been reviewed [99]. The main sapogenin from this source is glycyrrhetic acid (**38**), with the principal saponin being glycyrrhizin (glcA1 → 2glcA-38).

A subgroup of the oleanane-type saponins consists of glycosides with an ether bridge between C-13 and C-18 referred to as epoxyoleanane saponins (Fig. 7).

Examples are protoprimulagenin A (**70**), found in *Primula* sp., cyclamiretin A (**72**) and its saponin cyclamin [xyl1 → 2(glc1 → 3glc)-1 → 4(glc1 → 2-ara)-**72**] from *Cyclamen europeaum*, and the saikogenins (**75**) – (**77**) from *Bupleurum falcatum*, used in traditional oriental medicine [100].

Epoxyoleanane-type saponins are very sensitive to extraction conditions, and they yield artifacts upon acidic cleavage of the ether linkage. Protoprimulagenin A (**70**) yields primulagenin A (**48**), whereas saikogenins (**75**) – (**77**) rearrange to $\Delta^{11,13(18)}$-dienes not found in natural sources.

Ursanes. Aglycones of the ursane type (cf. Fig. 8) are derived from α-amyrin (**78**). Reports describe the isolation of saponins of ursolic acid (**79**) from both maté (*Ilex paraguayensis*) [101] and *Cynara cardunculus* [102]. Ilexgenin A (**82**) was first isolated from *I. pubenscens* [103]. An example of acyl – glycosidic linkages in ursane-type saponins is the asiaticoside rha1 → 4-glc1 → 6glc-28-O-**84**.

Dammaranes. In addition to saponins derived from oleanolic acid (**54**), the roots of *Panax schinseng* contain saponins originating from the tetracyclic dammarane-type aglycones (see Fig. 9). Elucidation of their true structures was complicated by rapid epimerization at C-20 during isolation. Progress in the identification of saponins from *Panax* sp. has been reviewed [104].

Ginsenosides of protopanaxadiol (**87**) are bisdesmosides with glycosidic linkages at C-3 and C-20, whereas protopanaxatriol (**88**) has glycosides linked to C-6 and C-20. Major saponins are ginsenosides Rb1 (glc1 → 2glc-3-O-87-20-O-glc6 ← 1glc) and Rg1 (glc-3-*O*-**88**-20-*O*-glc).

Lupanes and Hopanes. Sapogenins of the pentacyclic lupane and hopane systems are not very common compared to other triterpenes. Although unsubstituted ring systems are present in fruit waxes and other nonpolar plant materials, saponins are scarce. A saponin from *Asparagus gonocladus* has been isolated and identified as the glc1 → 2rha-glycoside of betulinic acid (90) [105]. Known hopane-type sapogenins include **91** and **92**.

Uses. Commercial applications of triterpene saponins are based largely on their surfactant properties and their ability to form oil-in-water emulsions. Formerly used as emulsifiers in X-ray film manufacture (→ Disperse Systems and Dispersions, Chap. 5.3.7.) and as components of detergents or cosmetics, saponins have been largely replaced in these applications by synthetic compounds.

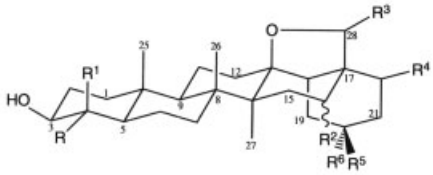

Sapogenin	CAS registry no.	Δ	R	R^1	R^2	R^3	R^4	R^5	R^6
Protoprimulagenin A (**70**)	[*2611-08-7*]		CH_3	CH_3	a-OH	H	H	CH_3	CH_3
Priverogenin B (**71**)	[*20054-97-1*]		CH_3	CH_3	a-OH	H	β-OH	CH_3	CH_3
Cyclamiretin A (**72**	[*5172-34-9*]		CH_3	CH_3	a-OH	H	H	CH_3	CHO
Anagalligenin A (**73**)			CH_3	CH_3	a-OH	OH	β-OH	CH_3	CH_3
Anagalligenin B (**74**)	[*33722-92-8*]		CH_2OH	CH_3	a-OH	H	H	CH_3	CH_3
Saikogenin E (**75**)	[*13715-23-6*]	11	CH_3	CH_3	e-OH	H	H	CH_3	CH_3
Saikogenin F (**76**)	[*14356-59-3*]	11	CH_2OH	CH_3	e-OH	H	H	CH_3	CH_3
Saikogenin G (**77**)	[*18175-79-6*]	11	CH_2OH	CH_3	a-OH	H	H	CH_3	CH_3

Figure 7. Epoxyoleanane-type sapogenins

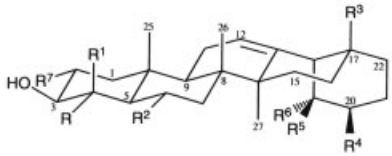

Sapogenin	CAS registry no.	R	R^1	R^2	R^3	R^4	R^5	R^6	R^7
Amyrin (**78**)	[*638-95-9*]	CH_3	CH_3	H	CH_3	CH_3	H	CH_3	H
Ursolic acid (**79**)	[*77-52-1*]	CH_3	CH_3	H	COOH	CH_3	H	CH_3	H
Pomolic acid (**80**)	[*13849-91-7*]	CH_3	CH_3	H	COOH	CH_3	OH	CH_3	H
Cincholic acid (**81**)	[*5948-32-3*]	CH_3	CH_3	H	COOH	CH_3	OH	CH_3	OH
Ilexgenin A (**82**)	[*108524-94-3*]	CH_3	COOH	H	COOH	CH_3	OH	CH_3	H
Rotundic acid (**83**)	[*20137-37-5*]	CH_2OH	CH_3	H	COOH	CH_3	OH	CH_3	H
Asiatic acid (**84**)	[*464-92-6*]	CH_2OH	CH_3	H	COOH	CH_3	H	CH_3	OH
Madecassic acid (**85**)	[*18449-41-7*]	CH_2OH	CH_3	OH	COOH	CH_3	H	CH_3	OH

Figure 8. Ursane-type sapogenins

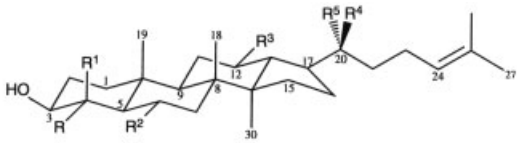

Sapogenin	CAS registry no.	R	R^1	R^2	R^3	R^4	R^5
Dammarenediol (**86**)	[*14351-29-2*]	CH_3	CH_3	H	H	OH	CH_3
Protopanaxadiol (**87**)	[*7755-01-3*]	CH_3	CH_3	H	OH	OH	CH_3
Protopanaxatriol (**88**)	[*1453-93-6*]	CH_3	CH_3	OH	OH	OH	CH_3

Figure 9. Dammarane-type sapogenins

The extract of *Glycyrrhiza glabra* containing glycyrrhetic acid (**38**) and glycyrrhetin is used by the tobacco and food industries for flavoring purposes [99]. As already mentioned, this sapogenins and saponins also have pharmaceutical value for the treatment of ulcers and gastritis [52]. Saponins used as expectorants are isolated from *G. glabra*, *Gala senega, Primula officinalis,* and *Hedera helix* [106].

A vast number of folk medicines contain saponins, and they are used as tonics and stimulants. The best-known examples are ex-

tracts of *Panax schinseng,* containing ginsenosid saponins, and *Bupleurum falcatum*, containing saikosaponins, which are important in traditional oriental medicine.

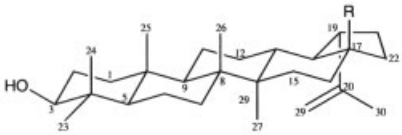

R = CH_3, lupeol (**89**) [*545-47-1*]
R = COOH, betulinic acid (**90**) [*472-15-1*]

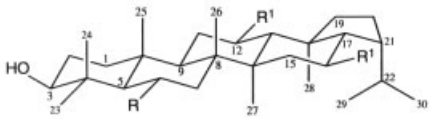

$\Delta^{15,17(21)}$, R = OH, R^1 = H, mollugogenol B (**91**) [*22554-64-9*]
$\Delta^{21(22)}$, R = H, R^1 = OH, spergulatriol (**92**) [*61127-09-1*]

Molluscicidal saponins have been proposed as a means of controlling the spread of schistosomiasis in tropical countries through elimination of the snail vector *Biomphalaria glabrata* [10].

4. Animal Saponins

Marine invertebrates of the phylum Echinodermata are categorized into five classes [107]: *Holothuridea* (sea cucumbers), *Asteroidea* (starfish), *Echinoidea* (sea urchins, sand dollars), *Ophiuroidea* (brittle stars, basket stars), and *Crinoidea* (sea lilies, feather stars). Only members of the classes *Holothuridea* and *Asteroidea* produce saponins. Asterosaponins are based on a steroid cholestane system, whereas holothurins have a triterpenoid lanostane framework. Besides the common sugars listed previously, less common sugars encountered are the 6-deoxy compounds β-D-quinovose (qui) and β-D-fucose (fuc), both in their pyranose forms.

4.1. Asterosaponins

Hydrolysis of crude saponin isolated from starfish yields the steroid asterone (**93**) [107] together with asterogenol, the corresponding 20*S*-hydroxyl reduction product [108]. A noteworthy feature is the $\Delta^{9(11)}$ double bond. Asterone (**93**) is derived from the genuine sapogenins thornasterol A (**94**) and B (**95**) [109] via retroaldol cleavage.

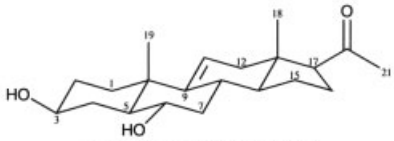

Asterone (**93**) [*37717-02-5*]

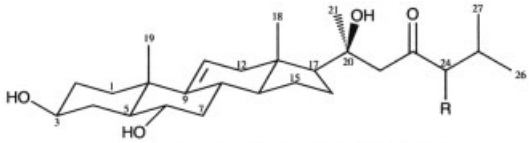

R = H, thornasterol A (**94**) [*55897-77-3*]
R = CH_3, thornasterol B (**95**) [*55897-78-4*]

Most saponins from starfish are sulfated at the 3-β-OH group and glycosylated at the 6-α-OH moiety. Their sugar chains consist of five to six monomeric units, with branching at the second sugar. Examples are forbesides A, B, and C, isolated from *Asterias forbesi* and characterized as gal1 $\rightarrow$ 3fuc1 $\rightarrow$ 2gal1 $\rightarrow$ 4(qui1 $\rightarrow$ 2xyl)1 $\rightarrow$ 3qui-6-O-**94**-3-O-$SO_3^-Na^+$, qui1 $\rightarrow$ 2gal1 $\rightarrow$ 4 (qui1 $\rightarrow$ 2xyl)1 $\rightarrow$ 3qui-6-O-**94**-3-O-$SO_3^-Na^+$, and fuc1 $\rightarrow$ 2gal1 $\rightarrow$ 4(qui1 $\rightarrow$ 2qui)1 $\rightarrow$ 3(6-deoxy-β-D-xylo-4-hexosulopyranosyl)-6-*O*-**94**-3-*O*-$SO_3^-Na^+$, respectively [28].

Sapogenins with an epoxy function in place of the keto group in the thornasterol side chain [110] or penta- and hexahydroxylated cholestane ring systems [111] have been reported.

Uses. Biological activities reported for asterosaponins [112] include cytotoxicity to tumor cells and antiviral activity. However, the high toxicity of these compounds prevents their pharmaceutical use. Use as precursors for steroid synthesis [113] is not technically feasible.

4.2. Holothurins

Holothurins are similar to the asterosaponin glycosides isolated from sea cucumbers. They have a $\Delta^{9(11)}$ double bond, and are prone to yield artifacts. The genuine aglycone holothurigenol

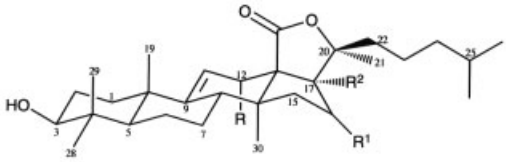

Sapogenin	CAS registry no.	Δ	R	R^1	R^2
Deoxybivittogenin (**97**)	[*77394-02-6*]		H	H	H
Bivittogenin (**98**)	[*67797-17-5*]		OH	H	H
Echinogenol (**99**)			OH	H	OH
Holotoxigenol (**100**)		25	H	O=	H

Figure 10. Holothurinogen-type sapogenins

(**96**) yields a $\Delta^{7,9(11)}$-diene elimination product on acidic hydrolysis.

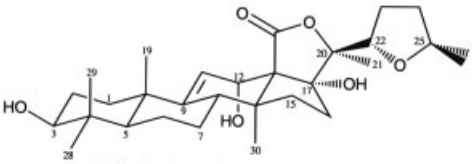

Holothurigenol (**96**) [*72244-90-7*]

Aglycones are glycosylated at the 3-β-OH group with sugar chains containing two to six monomer units. In contrast to asterosaponins, hydroxyl groups on the aglycone system are not sulfated. Instead, hydroxyl moieties in the glycoside chain are methylated and sulfated.

Holothurin B, isolated from *Holothuria leucospilata*, was identified as qui1 $\rightarrow$ 2(Na^+ ^-O_3S--*O*-4′-xyl)-3-*O*-**96** [114], whereas holothurin A was shown to be CH_3-*O*-3′-glc1 $\rightarrow$ 3glc1 $\rightarrow$ 4-holothurin B [115].

Other examples are saponins derived from aglycones with linear side chains at C-20 (cf. Fig. 10). Bivittosides A through D are derived from **97** and **98** [116], echinosides from **99**, and holotoxins A and B from holotoxigenol (**100**) [117].

Sapogenins with a Δ^7 – instead of $\Delta^{9(11)}$ – double bond have been reported in *Cucumaria frondosa* [118] and *C. echinata* [119].

Uses. Although interesting biological activities have been reported (e.g., cytotoxicity to tumor cells [119] and positive inotropic and chronotropic action on atrial muscle [120]), the pharmaceutical use of holothurins is prevented by their severe toxic effects.

5. References

1. I. Kitagawa, H. K. Wang, M. Yoshikawa, *Chem. Pharm. Bull.* **31** (1983) 716 – 722.
2. B. N. Meyer et al., *J. Agric. Food Chem.* **38** (1990) 205 – 208.
3. C. Adler, K. Hiller, *Pharmazie* **40** (1985) 676 – 693.
4. R. Segal, P. Shatkowsky, I. Milo-Goldzweig, *Biochem. Pharmacol.* **23** (1974) 973 – 981.
5. H. Abe et al., *Planta Med.* **34** (1978) 160 – 166.
6. R. Schönheimer, H. Dam, *Z. Physiol. Chem.* **215** (1933) 59.
7. B. von Wolters, *Planta Med.* **14** (1966) 392.
8. P. A. Arneson, R. D. Durbin, *Plant. Physiol.* **43** (1968) 683.
9. W. M. L. Crombie, L. Crombie, *Phytochemistry* **25** (1986) 2069 – 2073.
10. A. Marston, K. Hostettmann, *Phytochemistry* **24** (1985) 639 – 652.
11. Y. Birk in I. E. Liener (ed.): *Toxic Constituents of Plant Foodstuffs,* Academic Press, New York 1969, p. 169.
12. C. Arnault, K. Sláma, *J. Chem. Ecol.* **12** (1986) 1979 –1986.
13. W. Sanderman, H. Funke, *Naturwissenschaften* **57** (1970) 407.
14. G. Febvay, P. Bourgeois, A. Kermarrec, *Agronomie (Paris)* **5** (1985) 439 – 444.
15. R. Kuhn, I. Löw, *Chem. Ber.* **94** (1961) 1088.
16. I. Ishaya, Y. Birk, A. Bondi, Y. Tencer, *J. Sci. Food Agric.* **20** (1969) 433.
17. Y. Asada, T. Ueoko, T. Furuya, *Chem. Pharm. Bull.* **37** (1989) 2139 – 2146.
18. R. Segal, E. Schlösser, *Arch. Microbiol.* **104** (1975) 147.

19. H. M. Feder, A. M. Christensen in R. A. Boolootian (ed.): *Physiology of Echinodermata,* J. Wiley & Sons, New York 1966, p. 87.
20. K. R. Price, I. T. Johnson, G. R. Fenwick, *CRC Crit. Rev. Food Sci. Nutr.* **26** (1987) 27 – 135.
21. S. B. Mahato, S. P. Sarkar, G. Poddar, *Phytochemistry* **27** (1988) 3037 – 3067.
22. T. Schöpke, K. Hiller, *Pharmazie* **45** (1990) 313 – 342.
23. K. Hostettmann, *Planta Med.* **39** (1980) 1 – 18.
24. H. Kanazawa et al., *Chem. Pharm. Bull.* **38** (1990) 1630 – 1632.
25. S. Sakuma, H. Motomura, *J. Chromatogr.* **400** (1987) 293 – 295.
26. D. Fraisse, J. C. Tabet, M. Mecchi, J. Raynaud, *Biomed. Environ. Mass Spectr.* **14** (1986) 1 – 14.
27. P. K. Agrawal, D. C. Jain, R. K. Gupta, R. S. Thakur, *Phytochemistry* **24** (1985) 2479 – 2496.
28. J. A. Findlay, M. Jaseja, D. J. Burnell, J.-R. Brisson, *Can. J. Chem.* **65** (1987) 1384 – 1391, 2605 – 2611.
29. H. K. Wang, *Chem. Pharm. Bull.* **37** (1989) 2041 –2046.
30. Y. Sashida, K. Kawashima, Y. Mimaki, *Chem. Pharm. Bull.* **39** (1991) 698 – 703.
31. M. E. Wall, R. C. Eddy, M. C. McClennan, M. E. Klump, *Anal. Chem.* **24** (1952) 1337.
32. R. Wasicky, M. Wasicky, *Qual. Plant. Plant Foods Hum. Nutr.* **8** (1961) 65.
33. W. Winkler, *Arzneim. Forsch.* **18** (1968) 1031.
34. V. H. Honerlagen, H.-R. Trelter, *Dtsch. Apoth. Ztg.* **119** (1979) 1483.
35. L. E. Liberti, A. D. Marderosian, *J. Pharm. Sci.* **67** (1978) 1487 – 1489.
36. E. Bombardelli, A. Bonati, B. Gabetta, E. M. Martinelli, *J. Chromatogr.* **196** (1980) 121 – 132.
37. H. Hattori et al., *Chem. Pharm. Bull.* **36** (1988) 4467 – 4473.
38. H. Yamaguchi et al., *Chem. Pharm. Bull.* **36** (1988) 4177 – 4188.
39. U. Sankawa et al., *Chem. Pharm. Bull.* **30** (1982) 1907 – 1910.
40. P. Seeman, *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **33** (1974) 2116 – 2124.
41. S. Takagi, H. Ohtsuka, T. Akiyama, U. Sankawa, *Chem. Pharm. Bull.* **30** (1982) 3485 – 3492.
42. R. Segal, I. Milo-Goldzweig, *Biochim. Biophys. Acta* **512** (1978) 223.
43. T. Nakamura et al., *J. Pharmacobio. Dyn.* **2** (1979) 374 – 382.
44. B. S. Yu et al., *Chem.-Biol. Interact.* **56** (1985) 303 –319.
45. G. Reznicek, J. Jurenitsch, *Pharm. uns. Zeit* **20** (1991) 278 – 281.
46. T. Odani, H. Tanizawa, Y. Takino, *Chem. Pharm. Bull.* **31** (1983) 1059 – 1066.
47. M. Karikura et al., *Chem. Pharm. Bull.* **38** (1990) 2859 – 2861.
48. I. T. Johnson et al., *J. Nutr.* **116** (1986) 2270.
49. S. Hiai, H. Yokoyama, H. Oura, *Chem. Pharm. Bull.* **29** (1981) 490 – 494.
50. US 4 755 504, 1988 (Y. Liu).
51. H. Matsuda, K. Samakawa, M. Kubo, *Planta Med.* **56** (1990) 19 – 23.
52. M. H. Khan, F. M. Sullivan in J. Robson, F. Sullivan (eds.): *Symposium on Carbenoxolone Sodium,* Butterworths, London 1968, p. 5.
53. S. Shibata et al., *Chem. Pharm. Bull.* **35** (1987) 1910 –1918.
54. T. Kakuno, K. Yoshikawa, S. Arihara, *Tetrahedron Lett.* **32** (1991) 3535 – 3538.
55. C. R. Sirtori et al., *Lancet* 1977, 275.
56. D. G. Oakenful, D. L. Topping, R. J. Illman, D. E. Fenwick, *Nutr. Rep. Int.* **29** (1984) 1039.
57. H. Yoshimura, K. Watanabe, N. Ogawa, *Eur. J. Pharmacol.* **146** (1988) 291 – 297; *Eur. J. Pharmacol.* **150** (1988) 319 – 324.
58. Takeda Chem. Ind., GB 2 179 042, 1986 (H. Oshio, M. Kuwahara, T. Komiya).
59. S. Arichi, T. Hayashi, M. Kubo, DE-OS 2 828 851, 1980; T. Takamoto, S. Arichi, Y. Uchida, DE-OS 3 042 117, 1981.
60. M. Amoros, R. L. Girre, *Phytochemistry* **26** (1987) 787 – 791.
61. H. Tokuda, T. Konoshima, M. Kozuka, T. Kimura, *Cancer Lett. (Shannon, Irel.)* **40** (1988) 309 – 317.
62. K. Hirabayashi et al., *Chem. Pharm. Bull.* **39** (1991) 112 – 115.
63. X. H. Zhou et al., *Chem. Pharm. Bull.* **39** (1991) 1250 – 1252.
64. I. Maharaj, K. J. Froh, J. B. Campbell, *Can. J. Microbiol.* **32** (1986) 414.
65. Wakunaga Yakuhin K.K., DE-OS 3 207 841, 1982 (O. Tanaka, S. Hiroshima, N. Yata).
66. R. Tschesche, K. H. Richert, *Tetrahedron* **20** (1964) 387.
67. S. B. Mahato, A. N. Ganguly, N. P. Sahu, *Phytochemistry* **21** (1982) 959 – 978.
68. Y. Hirai, S. Sanada, Y. Ida, J. Shoji, *Chem. Pharm. Bull.* **32** (1984) 295 – 301.
69. D. C. Jain, *Phytochemistry* **26** (1987) 1789 – 1790.
70. Y. Hirai, S. Sanada, Y. Ida, J. Shoji, *Chem. Pharm. Bull.* **34** (1986) 82 – 87.

71. C. O. Okunji, C. N. Okeke, H. C. Gugnani, M. M. Iwu, *Int. J. Crude Drug Res.* **28** (1990) 193 – 199.
72. Y. Watanabe, S. Sanada, Y. Ida, J. Shoji, *Chem. Pharm. Bull.* **32** (1984) 3994 – 4002.
73. G. Wulff, T. Tschesche, *Chem. Ber.* **102** (1969) 1253.
74. S. Kiyosawa et al., *Chem. Pharm. Bull.* **16** (1968) 1162.
75. H. Matsura et al., *Chem. Pharm. Bull.* **36** (1988) 3659 – 3663.
76. S. Grünweller, E. Schröder, J. Kesselmeier, *Phytochemistry* **29** (1990) 2485 – 2490.
77. S. C. Sharma, R. Chand, B. S. Bhatti, O. P. Sati, *Planta Med.* **46** (1982) 48 – 51.
78. M. S. Kamel et al., *Chem. Pharm. Bull.* **39** (1991) 1229 – 1233.
79. R. K. Gupta, D. C. Jain, R. S. Thakur, *Phytochemistry* **25** (1986) 2205 – 2207.
80. R. Tschesche, P. Lauven, *Chem. Ber.* **104** (1971) 3549.
81. H. U. Lüning, E. Schlösser, *Z. Pflanzenkrankh. Pflanzenschutz* **82** (1975) 699.
82. T. Ikenaga et al., *Planta Med.* **54** (1988) 140 – 142.
83. I. Jizba et al., *Chem. Ber.* **104** (1971) 837.
84. J. A. Maga, *CRC Crit. Rev. Food Sci. Nutr.* **15** (1980) 371.
85. S. F. Osman in T. Swain, R. Klieman (eds.): *Recent Advances in Phytochemistry,* **vol. 14,** Plenum Press, New York 1979, p. 75.
86. Y. Sato, N. Ikegawa, E. Mosettig, *J. Org. Chem.* **24** (1959) 893; *J. Org. Chem.* **25** (1960) 783.
87. J. G. Roddick, *Phytochemistry* **13** (1974) 9 – 25.
88. R. Kuhn, I. Löw, H. Trischmann, *Chem. Ber.* **88** (1955) 1492, 1690.
89. R. Kuhn, I. Löw, *Chem. Ber.* **94** (1961) 1088, 1096.
90. L. H. Briggs et al., *J. Chem. Soc.* 1950, 3013.
91. K. Schreiber, H. Ripperger, *Liebigs Ann. Chem.* **655** (1962) 114.
92. G. Adam, K. Schreiber, *Chem. Ber.* **99** (1966) 3173.
93. Y. Sato, H. B. Latham, *J. Am. Chem. Soc.* **78** (1956) 3146.
94. H. Ripperger, K. Schreiber, H. Budzikiewicz, *Chem. Ber.* **100** (1967) 1741.
95. A. Ehmke, *Planta Med.* **52** (1987) 507.
96. A. Uddin, H. C. Chaturvedi, *Planta Med.* **37** (1979) 90 – 92.
97. P. Adam, J. M. Trendel, P. Albrecht, *Tetrahedron Lett.* **32** (1991) 4179 – 4182.
98. N. Kawamura, H. Watanabe, H. Oshio, *Phytochemistry* **27** (1988) 3585 – 3591.
99. G. R. Fenwick, J. Lutomski, C. Nieman, *Food Chem.* **38** (1990) 119 – 143.
100. H. Ishii et al., *Chem. Pharm. Bull.* **28** (1980) 2367 – 2383.
101. G. Gosmann, E. P. Schenkel, *J. Nat. Prod.* **52** (1989) 1367 – 1370.
102. S. Shimizu et al., *Chem. Pharm. Bull.* **36** (1988) 2466 –2474.
103. K. Hidaka et al., *Chem. Pharm. Bull.* **35** (1987) 524 –529.
104. O. Tanaka, R. Kasai, *Prog. Chem. Org. Nat. Prod.* **46** (1984) 1.
105. D. Mandloi, P. G. Sant, *Phytochemistry* **20** (1981) 1687 – 1688.
106. H. Wagner, H. Reger, *Dtsch. Apoth. Ztg.* **126** (1986) 1489 – 1493, 2613 – 2617.
107. J. W. ApSimon, J. A. Buccini, S. Badripersaud, *Can. J. Chem.* **51** (1973) 850 – 855.
108. J. W. ApSimon et al., *Can. J. Chem.* **58** (1980) 2703 –2708.
109. I. Kitagawa, M. Kobayashi, T. Sugawara, *Chem. Pharm. Bull.* **26** (1978) 1852 – 1863.
110. R. Riccio et al., *J. Nat. Prod.* **48** (1985) 756 – 765.
111. M. Iorizzi et al., *J. Nat. Prod.* **49** (1986) 67 – 78.
112. N. Fusetani et al., *J. Nat. Prod.* **47** (1984) 997 –1002.
113. J. E. Gurst, Y. M. Sheikh, C. Djerassi, *J. Am. Chem. Soc.* **95** (1973) 628. J. W. ApSimon, J. Burnell, J. Eenkhoorn, *Synth. Commun.* **9** (1979) 215 – 217.
114. I. Kitagawa et al., *Chem. Pharm. Bull.* **29** (1981) 1942 – 1950.
115. I. Kitagawa, T. Nishino, M. Kobayashi, Y. Kyogoku, *Chem. Pharm. Bull.* **29** (1981) 1951 – 1956.
116. I. Kitagawa, M. Kobayashi, M. Hori, Y. Kyogoku, *Chem. Pharm. Bull.* **37** (1989) 61 – 67.
117. I. Kitagawa et al., *Chem. Pharm. Bull.* **26** (1978) 3722 – 3731.
118. M. Girard et al., *Can. J. Chem.* **68** (1990) 11 – 18.
119. T. Miyamoto et al., *Liebigs Ann. Chem.* 1990, 39 – 42.
120. Y. Enomoto et al., *Br. J. Pharmacol.* **88** (1986) 259 –267.

Sealing Materials

Richard A. Palmer, Dow Corning Corporation, Auburn, Michigan 48611, United States

Jerome M. Klosowski, Dow Corning Corporation, Auburn, Michigan 48611, United States

1. Introduction . . . 1
1.1. Definition of Sealing Materials . . . 1
1.2. Performance Characteristics . . . 2
1.3. Specifications and Physical Properties . . . 2
1.3.1. General Considerations . . . 2
1.3.2. Modulus of Elasticity . . . 2
1.3.3. Movement . . . 3
1.3.4. Adhesion . . . 3
1.3.5. UV and Weathering Stability . . . 4
1.3.6. High-Temperature Stability . . . 4
1.3.7. Low-Temperature Flexibility . . . 4
1.3.8. Reversion . . . 4
1.3.9. Staining . . . 5
2. Medium-Performance Sealants . . . 5
2.1. Butyl Sealants . . . 5
2.2. Latex Acrylic Sealants . . . 5
2.3. Solvent-Release Acrylic Sealants . . . 6
3. Polysulfides . . . 6
3.1. Composition . . . 6
3.2. Properties . . . 7
3.3. Advantages, Markets, and Limitations . . . 8
3.4. Major Suppliers . . . 9
4. Urethanes . . . 9
4.1. Composition . . . 9
4.2. Properties . . . 10
4.3. Advantages, Markets, and Limitations . . . 11
4.4. Major Suppliers . . . 11
5. Silicones . . . 12
5.1. Composition . . . 12
5.2. Properties . . . 13
5.3. Advantages, Markets, and Limitations . . . 15
5.4. Major Suppliers . . . 16
6. Environmental Protection . . . 16
7. References . . . 16

1. Introduction

1.1. Definition of Sealing Materials

Sealing materials, or sealants, (sometimes referred to as caulking materials) constitute a broad class of materials used to prevent the intrusion of water, air, and dirt into joints, holes, or gaps. Many of the gaps in question are "working" joints that open and close with changes in temperature. For this reason, most sealants are required to display adhesive properties so that they will adhere to the joint and maintain a complete seal. In fact, some types of sealant serve not only as a seal, but also to bind two materials together structurally. Sealants of the latter type are often called "adhesive sealants."

1.2. Performance Characteristics

There are many ways to classify sealants, but one of the most common addresses the sealant's ability to withstand joint movement. Among the first sealants developed were oil-based putties. These putties are often formulated with soybean or linseed oil, fibrous filler, calcium carbonate, gelling agents, and a catalyst. Such "low-performance" materials are very inexpensive, but generally effective only in joints with movements limited to $< 5\,\%$ of the total joint width. Some of these materials are still in use today, but generally only in applications entailing no movement, or as crack fillers. Because of their low cost they are in fact often misused.

"Medium-performance" sealants tolerate movements of $5 - 12.5\,\%$ of the joint width. Typical sealants applicable in this movement range are butyls, latex acrylics, and solvent acrylics, materials discussed briefly in Chapter 2.

"High-performance" sealants are designed for joints with movements $> 12.5\,\%$. These chemically curing elastomers include polysulfides, the best acrylics, urethanes, and silicones. Many such sealants can accept movements of $\pm 25\,\%$, and the best continue to perform satisfactorily with movement as great

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_499

as ±50 % or even +100 %/−50 %. High-performance sealants are the focus of this article. Note that the movement levels cited are as defined by ASTM C 719, which incorporates hot/cold cycles but tests a sealant only in its optimum condition (after 21 days of cure), not after the sample has been aged by weathering. ASTM C 719 is the only sealant-movement test used consistently throughout the United States, and thus provides a good test for comparison of performance (see also Section 1.3.3).

1.3. Specifications and Physical Properties

1.3.1. General Considerations

As the development of sealants has progressed, many types of specifications have appeared. The International Organization for Standardization is working toward a system of global specification. This has been adopted in Europe, but not yet in other parts of the world. In the United States, ASTM specification C 920 is commonly employed, whereas the Japanese Industrial Standard JIS A 5757 defines elastomeric sealants for industrial high-performance construction applications in Japan [1]. The standards themselves will not be discussed here, but their use is described in the reference cited.

Many physical properties of a sealant are subject to direct measurement. Some of the more important properties are described in the sections that follow for use later in this article.

1.3.2. Modulus of Elasticity

The term "modulus of elasticity" as used by sealant manufacturers refers to the secant modulus, a measure of the stress experienced by a sealant at a given strain. The modulus is very important because it reflects the force exerted by a sealant on an adhesive bond during movement. A high-modulus sealant transmits a high force to the bond, which may prove undesirable in the case of a weak substrate such as concrete. Although concrete is very strong in compression, it becomes quite weak if exposed to the tensile forces of a high-modulus joint sealant. For this reason, the modulus of a sealant must be closely matched to the strength of the substrate. Sealant failures involving high-modulus sealants on weak substrates (like concrete) actually reflect cohesive failures in the substrate itself at its surface, a direct result of the high stress transmitted to the substrate from the sealant as the joint opens.

Low-modulus sealants are gaining in popularity for construction joints (with the important exception of structural glazing) because with such materials less stress is imposed on the sealant, the substrate, and the bond line, maximizing the probability of a successful seal. Medium-to-high-modulus sealants are often used in applications in which the sealant must also serve as an adhesive. Structural glazing, which involves the relatively strong substrates glass and metal, is one such application.

When reviewing a manufacturer's modulus claims it is important to check actual force values generated by a given movement rather than relying only on a nondimensional specification like low, medium, or high. No such qualitative evaluations are presented in this paper because various companies use the terms in different ways. Modulus values from different manufacturers must in any event be compared with caution, since the actual values can vary significantly as a function of specimen size, shape, and the rate of strain.

1.3.3. Movement

The movement capability of a sealant describes the amount of displacement the sealant can withstand during extension and compression when placed in a joint modeling the final application. Since this test is specific to a particular substrate and sealant, it measures both the adhesive properties of the sealant to the substrate and the sealant's ability to stretch and recover. Movement is commonly expressed as "$\pm x\%$," where the positive value refers to extension and the negative value indicates compression. For example, a sealant with a +50/−25 % rating can be extended by 50 % or compressed to 25 % of the original joint width. Like modulus values, movement capabilities of sealants must be interpreted carefully. Movement capability is significantly affected by the nature of the substrate, cure time, curing conditions, temperature, and

rate of cyclic movement. One of the most severe movement tests specified in the United States is ASTM C 719, "Standard Test Method for Adhesion and Cohesion of Elastomeric Joint Sealants Under Cyclic Movement". This test exposes a sealant to several cycles of movement at 3.2 mm/h during a series of exposures to water and both high and low temperatures. In general, sealants with lower modulus values provide higher movement capabilities. Movement tests are among the most useful tests for sealants because they come close to simulating conditions to which the sealant will actually be exposed. However, a sealant manufacturer's movement claims must always be interpreted with caution. For example, a reported "50 % movement ability" may in fact be an alternative way of expressing "± 25 % movement." Movement claims presented in the absence of a test temperature may mean that the test was performed only at room temperature, and the sealant in question might not perform well under low-temperature conditions. Movement tests carried out according to the complete ASTM C 719 procedure indicate that a sealant has been tested under both hot and cold movement conditions.

1.3.4. Adhesion

One of the primary functions of a sealant is to adhere or stick to the substrate being sealed. Without adhesion, the sealant would fail, causing the joint to leak. Although there are many so- called general-purpose sealants, no single sealant provides good adhesion to all substrates. Typically, a given sealant is designed for a specific market and a few target substrates. Adhesion to other substrates can sometimes be achieved through the use of a primer, but sealants are generally designed to be self-priming. Adhesion can be tested in a variety of ways, but the best procedure is to measure adhesion in a test joint that models as closely as possible the proposed application and the anticipated weather conditions. Accelerated weathering with water, heat, and ultraviolet (UV) light can give a useful preliminary indication of the adhesion life-expectancy of a sealant. Rapid data collection is possible with accelerated tests, but devising the proper accelerated model, one that will give an accurate prediction of sealant performance, can unfortunately be quite difficult. Sealant suppliers often prefer simply to cite previously successful sealing applications, leaving the customer to decide whether the proposed application is analogous. Certain adhesion tests are also specific to individual sealant manufacturers. Typically, such tests specify a particular substrate, sealant, and type and length of weathering exposure, as well as what is considered to be acceptable performance. In order to model true sealing applications, tests of this type should always include some indication of movement value.

1.3.5. UV and Weathering Stability

Most sealants are subject to years of exposure to sunlight, heat, cold, rain, wind, and pollutants. Sealants can deteriorate on the basis of several different mechanisms, but ultraviolet radiation from sunlight is one of the more common destructive agents. Upon exposure to sunlight some urethanes and polysulfides develop surface cracks that may penetrate the seal during joint movement and with continued exposure. Polysulfides and silicones generally harden with age, although the extent of hardening varies widely. Some may harden within 5 or 10 years, while others may still perform at greater than 50 % of their original capacity after 30 years.

Sealants are usually exposed to accelerated weathering tests involving heat, water, and UV radiation, but there is significant disagreement over the best way to extrapolate accelerated weathering data to actual life expectancy. Nevertheless, these data may provide the only way to predict longevity for a particular product. Weathering data can often be obtained either from the sealant manufacturer or from architects knowledgable in construction sealants. In general, such accelerated data should be based on thousands of hours of testing.

1.3.6. High-Temperature Stability

In construction applications, elastomeric sealants may need to withstand temperatures as high as 90 °C on the dark- colored, south-facing walls of a building. Other industrial applications also subject sealants to various extremes of heat. Depending upon the application, the

heat stability of a particular sealant may be a critical factor. Long exposure to heat causes some sealants to harden, while others revert to a putty-like paste. Most manufacturers provide upper limits for the service lives of their sealants, and individual sealants often have upper usable temperature limits below the highest temperatures predictable in certain applications. It is therefore important to take the manufacturer's recommended temperature limit carefully into account when a sealant is selected. One should also consider the probable duration of anticipated temperature extremes. For example, some sealants can be exposed successfully to high temperature over the course of one week, but not for a month. Knowledge of the duration of a likely stress is as fundamental in choosing usable sealants as knowing the severity of that stress.

1.3.7. Low-Temperature Flexibility

Some sealants are also subject to harsh winter temperatures and winds. While most sealants remain flexible to $-30\,^{\circ}C$, the extent of this flexibility can vary. For applications involving joint movement at low temperature, tests should be conducted to verify the ability of a particular sealant to perform well at the specified temperature. It is crucial that low-temperature testing be carried out on thoroughly cured sealants with a test such as ASTM C 719. Partially cured sealants may successfully pass a test such as this in the laboratory, but fail on the job once they have fully cured and reached their ultimate modulus. If low-temperature durability is important and time permits, sealants should be cured for six months prior to low-temperature testing.

1.3.8. Reversion

High-performance elastomeric sealants "cure" after extrusion to form an elastic rubber. The modulus of this rubber is not constant, but changes as a function of the conditions to which the sealant is exposed. High-performance sealants generally change more slowly than low- or medium-performance sealants. Under certain conditions, a cured elastomer may "revert," or return to a soft, pasty consistency resembling the uncured sealant; other sealants may become stiff and hard. Some sealants survive for 2 years, some 5 years, some 10 years, some 20 years, and some much longer before modulus changes prevent them from performing satisfactorily. Vast differences exist even within a single generic class of high-performance sealants. For instance, some urethanes become hard within a year or two as a result of weathering in sunlight (although their behavior in the shade may be quite different), while others may retain a high fraction of their joint movement capacity for several years. Other urethanes cure initially to an elastomer, but then become soft and mushy after one or two years of exposure to direct sunlight [2]. This is the phenomenon referred to as reversion. On the other hand, certain two-part, mix-on-site sealants become soft and mushy simply because they were not mixed in the correct ratios. The difference between these two rather similar manifestations – one from a poorly cured material, the other due to reversion – is that the reverted sealant was once a good rubber that later deteriorated, while the incorrectly mixed sealant never fully achieved the rubbery state. The former represents a product problem, but the latter is an application problem.

1.3.9. Staining

Staining is a concern with all sealants. Noncuring butyl materials, some reverting urethanes, certain polysulfides, and some silicones are among the sealants known to discolor glass and other surfaces. Coated glass, porous ceramics, and white marble are especially prone to surface staining. Testing is often carried out prior to large sealing jobs to ensure that the proposed sealant and substrate are compatible. ASTM C 510 and D 2203 are screening tests useful in establishing whether a particular sealant will stain a substrate [3]. As with other performance criteria for sealants, manufacturers are prepared to provide the user with staining data upon request.

A related concern is surface contamination caused by a sealant. Paint shops are very wary of sealants, because surface contamination may interfere with the adhesion of paint to a surface undergoing treatment. Silicones for use in the automobile industry are often specially formulated to be free of traces of volatile components

that might otherwise affect paint adhesion to a substrate.

2. Medium-Performance Sealants

2.1. Butyl Sealants

Butyl sealants constitute a class of medium-performance sealants that includes moldable, preformed butyl tapes as well as hot-applied butyl, solvent-release butyl, and polyisobutylene sealants. Butyl rubber is a copolymer of isobutylene with 15 – 45 wt % isoprene. Polyisobutylene (PIB) is produced by the polymerization of isobutylene, $H_2C{=}C{-}(CH_3)_2$ (→ Butenes; → Polyolefins, Chap. 3.;→ Polyolefins, Chap. 4.; → Rubber, 3. Synthetic). These two materials have been used as sealants since the 1930s. Both contain fillers, typically carbon black, zinc oxide, clays, or silicas. Tackifiers such as the pentaerythritol ester of rosin are also incorporated. Solvent-release butyls commonly contain cyclohexane to plasticize the sealant and facilitate hardening as the solvent evaporates [4].

Preformed butyl tapes re completely solid sealants used in the glazing industry to bond glass to aluminum. They do not shrink, but have a short service life: up to ten years depending upon the environment. *Hot-applied butyls* find particular application in the glazing industry. They are applied at 150 – 204 °C, and form somewhat elastomeric sealants as they cool to room temperature. *Solvent-release butyls* are low in cost and generally provide good adhesion. However, they are limited to about 10 % movement, display moderate shrinkage, and can stain certain substrates. *Polyisobutylene*, a solid sealing material, is used as the primary moisture – vapor seal in the construction of dual-sealed insulating glass units. Low moisture – vapor transmission is what makes polyisobutylene the sealant of choice for this purpose. Secondary seals in insulating glass are typically fabricated from silicone, urethane, or polysulfide. Solvent-release forms of polyisobutylene are also available, but these have the same limitations as solvent-release butyl sealants. Because of their low cost, butyls have become the sealants of choice in the mobile home industry and many other cost-sensitive markets where longevity is not demanded.

2.2. Latex Acrylic Sealants

Latex acrylic sealants were introduced in the late 1960s. Very high molecular mass, low viscosity polyacrylates are prepared by a batch emulsion polymerization process in which monomer, water, surfactants, and an initiator are added to a reactor and polymerized until the acrylic monomer is depleted (→ Polyacrylates, Chap. 3.1.). The maximum level of solids is 60 – 65 %. Above this point viscosity increases significantly and latex stability is poor. Compounding into sealants is accomplished by adding fillers, a freeze – thaw stabilizer, a thickener, and an adhesion promoter [5]. Both clear and pigmented forms are available.

Latex acrylics are easy to apply, and they can be cleaned up with water. Most are low in volatile organics, paintable, and inexpensive, and some provide good adhesion. On the other hand, they are subject to considerable shrinkage, and their modulus changes significantly with temperature [6]. Movement capability is generally near ± 12.5 %, although the literature claims that some formulations achieve ± 25 % movement capability. Latex acrylic sealants are very popular in residential construction and in the consumer market as a result of relatively low cost and ease of application and clean-up.

Many "siliconized" latex acrylics are also available, but they are similar or even identical to conventional latex acrylics with no distinctive silicone characteristics.

2.3. Solvent-Release Acrylic Sealants

Solvent-based acrylic polymers, like latex acrylics, are also produced by batch polymerization, but a solvent is used in place of water and surfactants. The molecular masses of solvent-based acrylic polymers are lower than those of latex acrylic polymers due to a high inherent viscosity. However, the solids content is higher, typically in the range 80 – 83 %. Because these materials contain a considerable amount of solvent and (typically) at least some low molecular mass reactable species, they are able to penetrate dirt and grease and still develop adhesion,

unlike other sealants, which require clean surfaces for optimal adhesion. The UV (sunlight) durability of acrylics is quite good, and their cost is moderate. The high level of solvent and the presence of relatively volatile acrylic components does lead to considerable odor, however, so these sealants are generally limited to outdoor applications. Sealants in this category tend to become permanently deformed as a result of both extension and compression, and they become stiff in cold weather. This restricts their use in high-movement joints. Joint movement capability for solvent acrylics is low, typically 12.5 % maximum.

While all medium-performance sealants are subject to a variety of modes of failure, the most common cause of failure for this group of sealants is hardening. Hardening is usually caused by a gradual loss of plasticizer, together with some oxidative degradation.

3. Polysulfides

3.1. Composition

The basic components of a polysulfide sealant are polymer, plasticizer, filler, curing agent, accelerators/retarders, thixotropes, and adhesion promoters (→ Rubber, 3. Synthetic, Chap. 5.3.1.2.). The polymer backbone may have any of several structures, but most cure through reactions at the terminal mercaptan ends. Several types of polysulfide polymers are illustrated in Figure 1.

Conventional polysulfide polymers are available in molecular masses ranging from 1000 to ca. 8000, with cross-link densities of 0.05 – 2.0 mol %. Modulus and hardness values for a polysulfide sealant are controlled mainly by the extent of cross-linking.

Although condensation or addition reactions can also be used to cure polysulfide sealants, the most common approach is oxidation. Several different oxidative curing agents are employed. Among the most common are lead dioxide, calcium peroxide, zinc peroxide, manganese dioxide, and cumene hydroperoxide. The choice of a specific curing agent is often dictated by the intended application, as shown in Table 1. Some curing agents lead to unwanted byproducts that can affect the mechanical properties of the cured polymer, which requires control by a second additive. For example, lead dioxide curing can result in the formation of an undesirable lead mercaptide species, but this can be eliminated by heating or introducing a small amount of elemental sulfur.

Table 1. Polysulfide curing agents [7]

Curing agent	Application
Lead peroxide	building/construction sealants
Manganese dioxide	improved heat and UV resistance – aircraft and insulating glass
Calcium peroxide	one-component sealants (activated by exposure to moisture)
Inorganic chromates	improved heat resistance – aircraft sealants

The curing rate of a polysulfide sealant can be influenced through the use of accelerators or retarders. Such *cure-rate modifiers* usually work by altering the pH of the sealant, and may change the cure time for a two- component sealant by a matter of minutes or even days. In general, acids are used to reduce the cure rate, while bases accelerate the process. Table 2 lists several cure-rate modifiers for common curing agents.

Table 2. Polysulfide sealant cure-rate modifiers [8]

Curing agent	Retarder	Accelerator
PbO_2	stearic acid metallic stearate	sulfur
CaO_2		water
ZnO_2	sulfur	amines
MnO_2	stearic acid	bases
Cumene hydroperoxide		copper ions, bases

Fillers are added to polysulfide sealants to reinforce, extend, and control the rheology of the system. Common fillers include ground calcium carbonate, precipitated calcium carbonate, furnace or thermal carbon black, and calcined clay. The particle size of a filler plays an important role in determining the extent of reinforcement and the resulting viscosity. The pH of the filler is also an important factor, because it can have an impact on the cure rate and the final stability of the cured sealant. Some fillers, such as the calcium carbonates, are often coated with treating agents as a way of reducing their effect on the pH of the system.

Plasticizers are added to polysulfide sealants to extend the system and lower the modulus. These must be compatible with polysulfide polymers and low in volatility to minimize shrinkage

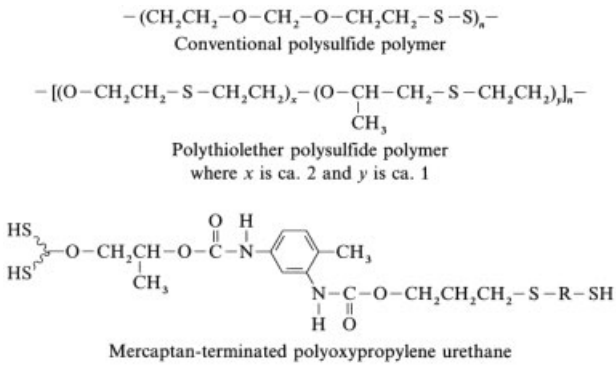

Figure 1. Examples of various types of polysulfide polymers

of the cured sealant. Commonly used plasticizers include butyl benzyl phthalate, octyl benzyl phthalate, dibutyl phthalate, hydrogenated terphenyl, chlorinated hydrocarbons, and modified styrene.

Although polysulfides display some inherent adhesion to common construction substrates, adhesion-promoting additives are typically introduced to generate the high levels of adhesion required in a modern sealant.

3.2. Properties

Modulus of Elasticity. Polysulfide sealants are subject to a high degree of stress relaxation. As a result, measured modulus of elasticity values are dependent upon test-specimen size and testing rate. Stress relaxation can be beneficial, because it allows a polysulfide sealant to stretch without transmitting excessive force to the sealant – interface bond line. However, it also reduces a polysulfide's ability to recover after the stress is removed. This slow recovery characteristic can lead to problems in rapidly cycling joints. Stress relaxation also limits the use of polysulfides in compressive gaskets [8].

Movement. With proper formulation, many polysulfides can accommodate joint movement of $\pm 25\,\%$, which is adequate for most construction joints. Joints subject to rapid cyclic movements can pose a problem for polysulfides if the cycling rate exceeds the rate at which the sealant can recover. Rapid movement in this sense can result in wrinkling or cohesive failure of the sealant [9].

Adhesion. As in the case of other common sealants, additives are typically introduced into polysulfides to ensure adhesion to specific substrates. Phenolic resins, epoxy- and mercapto-functionalized silanes, and epoxy resins are typical additives that improve adhesion to glass, aluminum, and steel. Each of these promoters has the ability to bond chemically to mercaptan groups in the polymer and to reactive sites at the substrate interface [10]. Other substrates, such as masonry or wood, often require primers.

UV and Weathering Stability. Performance in this respect can vary widely among different formulations and brands, making it difficult to characterize the stability of a "typical" polysulfide. Low-quality polysulfides show surface crazing or "alligatoring" after weathering, whereas higher quality polysulfides have successfully sealed certain structures for periods of 10 – 20 years. While the weatherability of polysulfides is not as good as that of silicone, it is adequate for many types of outdoor exposure. All polysulfides are subject to some UV damage; although they can be used for glazing applications in which the bond line is protected from sunlight by a mullion or corner, they are never used in structural glazing situations where the sun shines through the glass onto the bond line.

High-Temperature Stability. In general, polysulfides are intended for use below 100 °C (typically 70 °C), and their hardness can in-

crease with time or heat aging. If exposure to high temperature is contemplated, tests should be conducted to establish the heat stability of a particular sealant at the anticipated service temperature.

Low-Temperature Flexibility. Polysulfides become stiffer at low temperature, but hardening is not significant above − 34 °C.

Chemical Stability. Polysulfides show high resistance to chemicals, heat, oils, and solvents, leading to reliable long-term adhesion in applications involving exposure to these agents. The solvent resistance of polysulfides is the key to their wide use as fuel-tank sealants in the aircraft industry and in many other fuel- containment applications.

Staining. Staining is generally not a problem with polysulfides. However, certain phenolics that are added as adhesion promoters have caused staining problems on masonry.

Modes of Failure. Typical modes of failure for polysulfide sealants vary with the application. In glass and glazing applications, a common cause of failure is destruction of the bond to glass by direct or reflected sunlight. A second failure mechanism affecting many polysulfide sealants is hardening due to oxidation and loss of plasticizer. In joints subject to considerable movement, compression set is a serious problem, especially with the very slow- curing polysulfides. In construction applications, installation defects, such as failure to use a recommended primer, or poor surface cleaning, can result in adhesive bond failure.

3.3. Advantages, Markets, and Limitations

A major advantage of polysulfides is their excellent chemical resistance. This property has led to applications in aircraft fuel-tank sealing and in special construction joints where chemical resistance is important. Polysulfides are among the materials used to seal spill- containment areas for airports.

The insulating glass industry, especially in Europe, is another large consumer of polysulfides. Mercaptan-terminated polyoxypropylene sealants, a special category of polysulfides, also find use as insulating glass sealants. They show relatively good weatherability characteristics and generally low inert gas transmission rates in comparison with conventional polysulfides. This is important, because inert gases such as argon are finding increasing use to improve insulating values in insulating glass units.

3.4. Major Suppliers

Major U.S. suppliers of polysulfides are Chem-Rex, Morton International, Pecora, Products Research and Chemical, and Tremco. In Europe, major suppliers include Chemetal, Evode, Expandite, Kommerling, Le Joint Français, Products Research and Chemical, Servicized, and Teroson [11].

4. Urethanes

4.1. Composition

The three major components in a polyurethane (or urethane) sealant are polymer, filler, and plasticizer (→ Polyurethanes). Isocyanates are the basic starting materials for forming urethane polymers. These react with compounds containing active hydrogen, including alcohols, amines, acids, and water, as illustrated by the example below:

$$\underset{\text{Isocyanate}}{R-N=C=O} + \underset{\text{Alcohol}}{R'-OH} \longrightarrow \underset{\text{Urethane}}{R'-O-\overset{\overset{\displaystyle O}{\|}}{C}-NH-R}$$

Isocyanates are made by the phosgenation of amines, and are available in both aromatic and aliphatic forms. To reduce costs, most urethane sealants are prepared with aromatic isocyanates despite the fact that they are subject to yellowing and UV (sunlight) instability. Aliphatic isocyanates are used in special sealants in which yellowing might prove a problem, and where increased stability is necessary. Aliphatic isocyanates normally react more slowly and are usu-

ally more expensive, but they show improved resistance to hydrolysis and thermal degradation [12].

The starting isocyanates are not present in free form in urethane sealants but rather are incorporated into oligomers, commonly called "prepolymers", which result from the reaction of a difunctional isocyanate with a hydroxy-containing compound. Polyesters were common components of the early urethanes, but the most widely used hydroxy compounds today are hydroxy-terminated polyethers, made by the base-catalyzed addition of propylene oxide (PO) to di- or polyfunctional alcohols or amines. Typical polyfunctional alcohols used for this purpose are propylene glycol, glycerol, and trimethylolpropane. The structure of a typical polyether used in polyurethane sealant formulation is shown below:

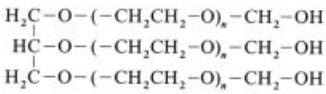

The molecular mass of the polyether component can vary over a wide range, from ca. 440 to > 6000. Both diols and triols are available with this spectrum of molecular masses.

The first step in preparing a urethane sealant is reaction of the diisocyanate with the polyether. Normally one mole of diisocyanate is used for each mole of hydroxyl groups, to produce a polymer with a terminal isocyanate group, as illustrated below:

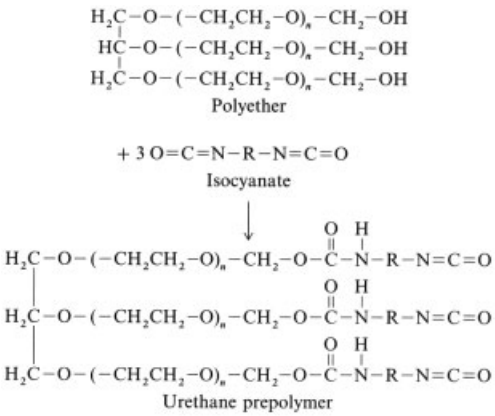

The isocyanate ends of the resulting prepolymer are still reactive with respect to active-hydrogen containing materials. A *two-part urethane* sealant can be cured to an elastomer by mixing it with, for example, 1,4-butandiol, leading to a cross-linked network in the cured elastomer. *One-part urethanes* are more complicated. Reaction of the terminal isocyanate groups with water vapor from the air produces carbon dioxide, which causes bubbling in the sealant. One method of solving this problem involves reacting the prepolymer with phenol and then mixing the resulting "blocked" prepolymer with diketimine. When exposed to moisture, the diketimine is transformed into a diamine with the release of 2-butanone. The diamine then cleaves the bond between phenol and the urethane, causing a chain-extension reaction to occur [13].

As implied above, isocyanates are very reactive with water, and the urethane formulator must be very careful to exclude water and to predry all the components of the formulation. Too much water can reduce shelf-life or cause a formulation to gel. One approach to eliminating moisture during formulation is the addition of quicklime (calcium oxide), which reacts with water to form calcium hydroxide. This reaction is quite slow, and can require hours to occur. It is also reversible, and small amounts of water may reform at a later time. Too much lime can also be detrimental, because it can reduce shelf-life and cause poor weatherability.

Plasticizers are often added to urethane sealants to reduce the viscosity of the formulation. Lower viscosity simplifies handling of the sealant. It also leads to a lowering of the modulus, and in some cases the elongation potential increases as well. Plasticizers are normally introduced in the 5 – 20 % range; typical examples include phthalates, phosphates, chlorinated hydrocarbons, and aromatic oils.

Some urethane sealants contain coal tar or asphalt as extenders, both to reduce cost and to add moisture-barrier properties. Coal tar contains active hydrogens, another factor that must be taken into account by the formulator. Also, tar-modified urethanes display rather poor UV resistance.

Common fillers in urethane sealants include calcium carbonate, talcs, clays, and fumed or precipitated silica. The choice of filler depends

on the desired final properties and anticipated uses of the sealant. Electrical properties can be significantly affected by a filler, as can modulus and elongation, since fillers can be either reinforcing or nonreinforcing.

4.2. Properties

Modulus of Elasticity. The modulus of elasticity of a urethane sealant is influenced by the prepolymer type, isocyanate functionality, plasticizer level, and filler type and surface area. Unless specially plasticized, most urethanes have a rather high modulus of elasticity and are very tough and tear resistant. This is one reason for their use in high sealant-abuse areas, such as sidewalks and parking garages. Unfortunately, a high modulus can also limit a sealant's joint movement capability. Unlike polysulfides, the modulus of a urethane sealant is not affected by the rate of testing, a consequence of the high recovery ability of urethanes. Good urethane sealants yield recoveries of 70 – 90 %, which is better than polysulfides but in general not as good as silicones.

Movement. One- and two- component urethanes display a movement capability of as much as ± 25 %, with some manufacturer's literature claiming even higher values.

Adhesion. Urethane sealants have inherent adhesion capabilities with respect to most common substrates. However, this adhesion is often destroyed by exposure to water. Adhesion is improved by the use of adhesion promoters or primers. The most common adhesion promoters for urethane sealants are silane coupling agents. Epoxy-, amino-, and mercaptosilanes have been successfully used due to their dual reactive nature. The alkoxy ends typically react with surface hydroxy groups of the substrate, while the epoxy, amino, or mercapto ends react with the urethane polymer. Most urethanes show an increase in modulus with age. For this reason the adhesion must be tenacious and durable, since a high degree of stress is placed on the sealant/substrate bond, and the stress becomes greater as the sealant ages and hardens.

UV and Weathering Stability. While not as UV-resistant as silicones or acrylics, urethane sealants can be formulated to show reasonable UV resistance by the addition of UV stabilizers and antioxidants. Life expectancies for most urethane sealants range from 5 – 20 years depending on the application and exposure. If protected from direct sunlight, urethanes have long, useful lives in many areas of application, such as the automotive industry and construction. If not protected from sunlight, urethanes may develop surface cracks that can propagate through the sealant at an actively moving joint, leaving an open, unsealed joint. As with all sunlight-sensitive sealants, it is very important with urethanes to provide a sufficiently deep (thick) sealant bead so that considerable time will elapse before a crack traverses the entire depth.

High-Temperature Stability. Most urethane sealants are used in applications with high-temperature limits $< 100\,^{\circ}$C, and temperatures in excess of 80 °C can often be quite detrimental.

Low-Temperature Flexibility. In general, urethanes remain flexible to $-40\,^{\circ}$C. As with any sealant, testing is important to ensure that a particular sealant will satisfy low-temperature requirements.

Reversion. No single urethane can be regarded as typical. All are unique. With exposure to weather some tend to harden, while others become quite soft under the same conditions. Heat and moisture may exert some influence, but the primary influence in aging is ultraviolet (sunlight) radiation. Therefore, UV absorbers and UV-blocking pigments are very important. Differences in the type and content of UV stabilizers are one reason for the observed wide variation in urethane performance. Under certain conditions of temperature, light, and humidity, some urethanes revert to very soft or mushy sealants; others remain intact, and certain types even become quite hard.

Staining. Staining is not normally a problem with urethanes, apart from those that undergo reversion. Even with reversion any urethanes or plasticizers that reach the surface tend to be degraded and eliminated from the sealant. Thus,

except in cases where the sealant becomes very soft, urethanes are considered to be nonstaining.

Modes of Failure. Failure with urethane sealants is very much product- and application-dependent, although certain failure modes are more prevalent than others. One common cause of failure in direct sunlight is the formation of surface cracks, which can propagate completely through the seal in an active, moving joint. Another general mode of failure is sealant hardening. In this case, the sealant rarely tears, but joint movement places so much stress on the bond that adhesive bond failure occurs, or there may be cohesive substrate failure (in concrete, for example).

As mentioned previously, urethanes react readily with moisture to produce carbon dioxide. Bubbling that occurs during sealing with a urethane is often a result of moisture present on the substrate [14].

4.3. Advantages, Markets, and Limitations

Urethanes can be formulated to display a wide variety of properties, including modulus values ranging from very low to very high. In general they show good chemical resistance. Urethanes are tough and abrasion resistant, making them very useful in traffic-bearing joints such as those found in parking decks. Much perimeter building sealing is accomplished with urethanes. Another major market for urethane sealants is the automotive sector. Thus, urethane sealants are commonly used to bond windshields to metal frames (using primers) and to seal taillights. Curtain-wall glazing is generally carried out with silicone rather than urethanes due to some loss of urethane adhesion from UV exposure. Specially formulated urethanes can be used in insulating glass, but not in places where sunlight is likely to shine on the bond line. Urethanes are also not recommended for use where the service temperature will exceed 80 °C, and dry substrates are required to prevent bubbling.

4.4. Major Suppliers

Major suppliers of urethane sealants in the United States include ChemRex, Mameco International, Pecora Chemical, Products Research and Chemical, Sika, and Tremco [15].

5. Silicones (→ Silicones)

5.1. Composition

Silicone sealing materials are classified according to the three principal mechanisms by which they cure: moisture-cure sealants, silicone latex sealants (which cure by releasing moisture), and addition- curing sealants, which do not release any volatiles. The silicone latex referred to here is a true silicone based on silicone polymers, and should not be confused with siliconized latex acrylic sealants. Siliconized latex acrylic sealants contain less than 2 % silicone, and have all the characteristics of ordinary acrylic sealants.

Moisture- cure silicones are by far the most common silicone sealing materials. They consist of a silicone polymer, a filler, and a moisture-reactive silicone cross-linker. Depending upon the cross-linker used, a catalyst may also be needed. Most polymers used in silicone sealants are polydimethylsiloxanes, with repeating dimethylsilicon and oxygen units. The general structure of these polymers is shown below:

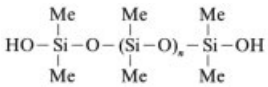

Although many different terminal groups can be attached to such a polymer, most sealant polymers invoke hydroxy groups, through which curing of the silicone sealant takes place. The viscosity of the polymers can range from < 100 cp to > 100 000 cp.

Specialty polymers can be produced to meet specific performance requirements. For example, the brittle-point of a silicone sealant can be lowered from − 65 °C to − 110 °C for aerospace applications by the introduction of diphenylsiloxy groups. Sealants that will be exposed to jet aviation kerosene or gasoline are made low-swell by attaching 1,1,1-trifluoropropyl groups to each silicon atom [16].

Moisture-curing silicones generally require moisture-reactive cross-linkers to convert the paste-like uncured sealant into an elastomer. Many different silicone-based cross-linkers are used. However, most have three (trifunctional) or four (tetrafunctional) reactive ends that react with both moisture in the air and the hydroxyl ends of the silicone polymer to cross-link the sealant. This cross-linking reaction in turn releases a hydrolyzable group, which diffuses out of the sealant. Often an organometallic condensation catalyst is added to accelerate the cross-linking reaction, although some cross-linkers react rapidly without a catalyst. Common catalysts are stannous octoate (where "octoic acid" is an industrial product consisting of eight- carbon acids), tin carboxylates, and organotitanates.

Several important cross-linkers and the corresponding leaving groups are listed in Table 3.

The cure of these moisture-reactive cross-linkers is dependent upon moisture diffusing into the sealant and the byproduct diffusing out. This diffusion- controlled process causes a cured elastomer skin to form on the exposed sealant surface, which then progresses inward. Due to their moisture-reactive nature, silicone sealants must be compounded in the absence of atmospheric moisture.

The siloxane polymer network is quite weak if cured alone, so it is reinforced with various inorganic fillers. Fumed silicas and carbon blacks are often used to prepare high-strength sealants. Sealants that do not require high strength are made with low- or nonreinforcing fillers such as ground calcium carbonate, ground quartz, diatomaceous earth, and various clays. A wide range of rheological properties can be obtained, from nonslump to self-leveling. Moisture- curing silicones are the most widely used and best known of the silicone sealants.

The choice of a particular sealant is often dictated by the nature of the leaving group. The most common leaving group is acetic acid, leading to a vinegar odor, but acetic acid is corrosive to some surfaces. For this reason other, more neutral leaving groups are gaining in popularity. Nonacid cures should always be specified for use on reactive surfaces such as concrete, marble, zinc, or copper.

A second group of silicone sealants consists of the *silicone latex* or *water-based silicones.* These are relative newcomers compared to moisture- curing silicones. A sealant of this type is basically a silicone-in-water emulsion that cures through evaporation of the emulsifying water. It is prepared by first emulsifying a low molecular mass siloxane polymer and then polymerizing it to the desired degree. Emulsion polymerization permits the preparation of very high molecular mass polymers with low viscosity. A cross-linker is added next, along with a condensation catalyst. In contrast to moisture-curing silicones, cross-linking in silicone latex occurs during manufacture. The cross-linker is typically of the $R'\text{-}Si(OR)_3$ configuration, and must diffuse through the aqueous phase into the siloxane in order to react with the polymer. The pH of the aqueous phase must be carefully controlled to prevent the cross-linker from reacting with water before contacting the siloxane phase. Fillers and freeze – thaw stabilizers are also added to complete the compounding of a silicone latex sealant [17].

The advantages of a true silicone latex are similar to those of an acrylic latex. It can be applied to damp surfaces, contains relatively little volatile material, and is subject to very easy clean-up and application.

Addition- cure silicones are generally based on the platinum- catalyzed reaction of a polymeric silicon hydride with a vinyl group attached to a silicone polymer:

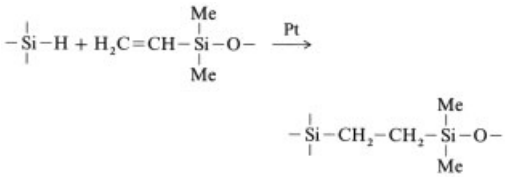

This vulcanization can be accelerated by heat, causing cure to occur within minutes in deep section. No byproducts are produced by this particular cross-linking reaction. Addition- cure silicones are used mainly for molding, encapsulation, and sealing applications. Buildings with penetrations through floors and walls that must be sealed with fire-rated materials are often sealed with addition- cure silicones. The primary advantages of addition- cure silicones are a rapid deep-section cure rate, cure in the absence of air, negligible amounts of volatile species, and no

Table 3. Common silicone cross-linkers and leaving groups

Cross-linker	Formula	CAS registry no.	Leaving group
Methyltrimethoxysilane	$CH_3Si(O-CH_3)_3$	[*1185-55-3*]	methanol
Methyltriacetoxysilane	$CH_3Si(O-\overset{\overset{\displaystyle CH_3}{\vert}}{C}=O)_3$	[*4253-34-3*]	acetic acid
Methyltri(methylethylketoxime)silane	$CH_3Si(O-N=C\begin{matrix}CH_3\\C_2H_5\end{matrix})_3$	[*22984-54-9*]	2-butoxime $(CH_3)(C_2H_5)C=N-OH$
Vinyltris(isopropenyloxy)silane	$CH_2=CHSi(O-\underset{\underset{\displaystyle CH_3}{\vert}}{C}=CH_2)_3$	[*15332-99-7*]	acetone

shrink with thick sections. The disadvantages of this type of silicone are sensitivity to contamination and problems with adhesion to some surfaces.

5.2. Properties

Modulus of Elasticity. Of the high-performance elastomeric sealants, silicones offer the widest modulus range available. Low-modulus silicones can be formulated with high molecular mass polymers containing a high level of plasticizer. However, this approach to low modulus can lead to staining with certain substrates. A second approach is to use a chain extender in combination with a cross-linker [18]. Reaction rates must be carefully matched to ensure the successful cure of a low-modulus sealant. In either approach, nonreinforcing fillers are added to keep the modulus low. Low-modulus silicones are commonly used in construction applications involving concrete. As noted previously, concrete has a low tensile strength, and high-modulus sealants can be a cause of cohesive failure. Very-low-modulus silicones are fast becoming the sealants of choice in substrate-sensitive environments and in almost all applications with very large joint movements.

High-modulus silicone sealants are prepared with relatively large amounts of highly reinforcing fillers, such as fumed silica. Tetrafunctional and trifunctional cross-linkers also contribute to an increased modulus. These sealants are used for applications in which adhesive seals are important, e.g., auto windshield applications, form-in-place gaskets, auto headlights, and appliance sealants.

Movement. Silicone sealants can be formulated to survive movements of +100/−50 %,thereby surpassing the qualities of other elastomeric sealants. Low-modulus sealants typically show the highest movement capability. Indeed, low-modulus silicones set the industry standard to which all other sealants are compared for movement ability in joints. The high movement ability of low-modulus silicones is a primary reason for their selection in remedial applications to seal joints with greater movement than can be tolerated by conventional sealants. Not all silicones are low modulus/high movement materials, however, and an appropriate sealant must be carefully selected for each application. Some applications, such as insulating glass secondary sealants, require high-strength, low-movement sealants, and thus a quite different formulation.

Adhesion. The adhesion characteristics of silicone sealants vary widely. Some general-use silicone sealants provide unprimed adhesion to all the usual construction and industrial surfaces. Such sealants generally contain one or several silane coupling agents to enhance adhesion to various surfaces. Formulators have access to many silane coupling agents, and a proper choice is essential for optimum performance. With some silicones a primer is required to ensure adequate adhesion on certain surfaces.

Addition- cure silicones can be formulated to provide very little inherent adhesion, and for this reason they are often used in molding or encap-

sulation applications in which release properties are important. An example is the encapsulation of electrical parts, where easy removal of the sealant is desirable in case the part must be repaired or modified. Adhesion is improved by the use of primers or proprietary adhesion promoters. Silicone latex sealants offer good adhesion, although it is generally inferior to that of moisture- curing silicones.

Silicones, like most other sealants, are usually inappropriate for applications in which the sealant is continuously submerged in water. A few exceptions exist, but the sealant supplier should be consulted before any attempt is made to seal a wet environment.

UV and Weathering Stability. Because silicones do not absorb UV radiation, they are not subject to degradation in sunlight. Accelerated UV-weatherometer exposures of 20 000 hours are common with no surface cracking. For this reason, silicones are frequently the sealants of choice for construction applications, and almost always for glazing applications in which sunlight will directly or indirectly reach the seal. This UV stability is a result of the absence of unsaturation in the silicon – oxygen backbone of the polymer. Extreme stability of this type is essential in structural glazing applications (i.e., in buildings with all-glass surfaces), where the sealant is the adhesive that actually holds the glass to the building. Indeed, silicones are the only materials permitted for structural glazing in which the sealant will be exposed to direct or reflected sunlight.

High-Temperature Stability. Silicones provide the greatest heat stability of any available class of elastomeric sealants. Industrial grades of silicone sealants can withstand temperatures of 160 –205 °C in air for a year without significant change in their properties [19]. A major contributor to this thermal stability is the polymer's silicon – oxygen backbone, but stability is also influenced by thermal stabilizers, fillers, and vulcanization catalysts. Heat stability is important in oven gaskets, appliances, and many industrial applications. Dark- colored, south-facing buildings also experience temperatures exceeding the 80 °C limit of most other sealants, and the heat stability of silicones makes them suitable for this application as well. While most silicones tolerate 80 – 100 °C very well, only a few survive for extreme lengths of time at temperatures > 200 °C.

Low-Temperature Flexibility. Siloxane polymers provide several important low-temperature advantages. Because the viscosity of the polymer changes only slightly at low temperature, extrusion of uncured sealant from cold cartridges is only slightly affected. The low-temperature flexibility of the polymer also permits the cured sealant to remain flexible to − 54 °C, and specialty polymers extend this low-temperature flexibility even further. More important, silicone modulus values change only slightly at low temperature, in contrast to other sealants. The brittle point of a silicone sealant can be lowered from − 65 °C to − 110 °C by the addition of diphenylsiloxy groups, a common expedient in aerospace applications.

Reversion. Silicones are extremely stable under normal weathering conditions, and show little change or drift in their properties, although some silicones prepared by tin carboxylate catalysis may revert at temperatures > 80 °C and if confined. Reversion due to cleavage of the silicon – oxygen bond can occur if a silicone is exposed continuously to strong acids or bases. Strong acids can also attack silicones that contain marble fillers.

Staining. Staining is the most widely cited flaw with respect to silicone sealants, but the e tent of staining varies considerably as a function of product and surface. Staining is most likely to be a problem with certain porous substrates, especially if the silicone contains low molecular mass nonreactive silicone plasticizers. Testing is often undertaken prior to sealing a large construction job to ensure that staining of the substrate will not occur.

Modes of Failure. Silicones have exceptional weatherability, so many of the failures actually encountered with these products are caused by loss of adhesion, usually because the sealants have such long lifetimes that they outlive the adhesive bond. Bond life is dependent on both the environment and the nature of the silicone – substrate combination. Many applications continue to be successful after 30 years,

whereas some 20-year-old applications have lost most of their original bond strength. As with other high-performance sealants, adhesion problems with silicone sealing materials are often caused by inadequate cleaning and joint preparation.

Most of the early silicones (and some of the present ones) show poor tear resistance, and for this reason they are not generally recommended for high-abuse areas. If a poor-tear-resistance silicone suffers a cut, the cut may propagate along the sealant bead and lead ultimately to failure. It should be noted, however, that some modern silicones have excellent tear resistance, and are used even in sidewalks. A large market for highly tear-resistant, low-modulus silicones is as highway and airport runway sealants.

5.3. Advantages, Markets, and Limitations

Silicone sealants offer the advantages of superior weatherability, great longevity, wide temperature service range, fast cure, and wide modulus range. In the construction industry, common applications are associated with insulating glass, structural glazing, pavement sealing, and perimeter weathersealing. In the home, silicones are most often used as bathtub caulks, grouts, and sink sealants because of their stability with respect to common cleansers and disinfectants as well as their longevity. In the automotive market, cure-in-place gaskets are commonly made of silicone. Many high-temperature industrial applications also specify silicones, and electronic components are often sealed with silicones. The fact that so many different types of silicones exist makes it difficult to generalize. This is particularly true because of the wide diversity of byproducts emitted when various types of silicones cure (acetic acid, amines, ketoximes, and alcohols). Each type of silicone sealant is limited by certain restrictions, and close scrutiny is required to ensure a proper choice.

In general, most contemporary silicones have the following limitations:

1) Silicone sealants are not abrasion resistant, and they must be recessed below the substrate surface to avoid abrasive contact (e.g., with vehicles in highway applications).
2) Except for specialty silicones (aquarium sealants and similar products), silicone sealants are generally not recommended for applications involving continuous water immersion because of the potential for adhesive failure.
3) Most moisture-cure silicones are not paintable. Certain silicones show improved properties in this respect, but they are still not as paintable as other elastomeric sealants.
4) While silicone sealants adhere well to most other sealants, other sealants usually do not adhere to silicones. Thus, it is generally acceptable to repair nonsilicone sealant failures with a silicone product, but usually not to repair a silicone with a nonsilicone.

5.4. Major Suppliers

The major manufacturers of silicone sealants are Dow Corning, General Electric, Wacker Silicones, Rhône-Poulenc, Bayer, and Shin-Etsu.

6. Environmental Protection

Virtually all sealants in all the generic categories release some material to the environment. However, the amount of volatiles released can vary significantly among different generic types and as a function of formulation within each type. In general, there is a trend toward replacing high-volatile formulations, such as solvent-based sealants, with more environmentally friendly sealant systems such as those that are $> 90\,\%$ solid, water-borne systems, and systems with special chemistries like the addition-curing silicones, which release little or no volatiles.

7. References

1. L. Carbary: "Types of Sealants," in *Engineered Materials Handbook,* vol. 3, ASM International, Materials Park, Ohio 1990, p. 188.
2. T. Bridgewater, L. Carbary: "Accelerated Weathering and Heat Stability of Various Perimeter Sealants," *2nd Symposium on*

Science and Technology of Building Seals, Sealants, Glazing, and Waterproofing, Fort Lauderdale, Fla., February 5, 6, 1992; ASTM Committee C-24 on Building Seals and Sealants.
3. J. Klosowski: *Sealants in Construction,* Marcel Dekker, New York 1989, p. 106.
4. A. Usmani: "Environmental Considerations Unique to Sealants," in [1], p. 675.
5. K. Potts: "Replacement of Solvent Based Glazing Compounds with High Performance Water Based Acrylics," *2nd Symposium on Science and Technology of Building Seals, Sealants, Glazing, and Waterproofing,* Fort Lauderdale, Fla., February 5, 6, 1992; ASTM Committee C-24 on Building Seals and Sealants.
6. In [1], p. 189.
7. J. Cook, J. Panek: *Construction Sealants and Adhesives,* 2nd ed., J. Wiley & Sons, New York 1984, p. 109.
8. L. Boller: "Polysulfide Sealants," in J. Panek (ed.): *Building Seals and Sealants,* American Society for Testing and Materials, Baltimore 1976, p. 45.
9. J. Cook: *Construction Sealants and Adhesives,* J. Wiley & Sons, New York 1970, p. 81.
10. E. Peterson: "Formulating Sealants with Mercaptan Terminated Polymers," *Sealant Short Course,* Adhesives and Sealants Council, Rosemont, Ill., May 4 – 6, 1992.
11. R. Giangiordano, C. Sorensson: "Polysulfides," in [1], p. 196.
12. J. Regan: "Urethane Formulations," *Sealant Short Course,* Adhesives and Sealants Council, Rosemont, Ill., May 4 – 6, 1992.
13. R. Evans, R. Greene: "Urethane Sealants," in J. Panek (ed.): *Building Seals and Sealants,* American Society for Testing and Materials, Baltimore 1976, p. 122.
14. J. Regan: "Urethanes," in [1], p. 205.
15. In Ref. [14], p. 204.
16. J. Klosowski, D. Beers: "Silicones," in [1], p. 215.
17. D. Liles, N. Shephard: "Silicone Rubber Latex Sealants," *2nd Symposium on Science and Technology of Building Seals, Sealants, Glazing, and Waterproofing,* Fort Lauderdale, Fla.,February 5, 6, 1992, ASTM Committee C-24 on Building Seals and Sealants.
18. In [3], p. 293.
19. In [16], p. 217.

Seasonings

Flavors and Fragrances is a separate keyword.

Leonard G. Fischer, McCormick & Company, Incorporated, Hunt Valley, Maryland 21031, United States

1. **Introduction** 1
2. **Uses** 1
3. **Resources and Raw Materials** 2
4. **Production** 2
5. **Storage and Transportation** 3
6. **Quality Specifications** 3
7. **Economic Aspects** 3
8. **References** 4

1. Introduction

A seasoning is a mixture of ingredients added to a food in order to improve its palatability and appearance.

"Seasoning" is the broadest term commonly applied in describing modifications to the flavor, appearance, or texture of foods. Its definition includes the use of spices, herbs, extracts, flavors, sweeteners, and salts, in addition to other food ingredients. By popular convention seasonings are generally understood to be dry blends, whereas wet products used for the same purposes are commonly referred to as condiments. The term seasonings is often thought of as covering two general classes of tastes: (1) sweet, typically applicable to fruit preparations and bakery goods, and characterized by the presence of sugars in addition to spices or spice extractives with sweet characteristics (e.g., anise or cinnamon), and (2) savory, which generally implies salty, and utilizes seasoning ingredients that are pungent, tangy, woody, etc. (e.g., pepper, dill, sage).

The use of seasonings predates recorded history, and their invention is a subject for archaeological speculation. Many seasonings or specific seasoning blends have regional or ethnic origins, and the cuisine of many countries and peoples is characterized by a particular type and level of seasoning, as in curries from India and chilies in Mexican cooking.

It is commonly believed that spices first came into use in Europe after the Crusades, probably to provide some preservative effect, or to disguise the flavor of spoiled or rancid meat. Scientific evidence now exists for the antibacterial effects of certain spices, and for the antioxidant effect of some herbs.

From a commercial standpoint, the market for seasonings is generally divided into two parts: (1) the retail or home-use market, in which pure spices and simple seasoning blends are sold directly to consumers for home cooking, and (2) the industrial market, where seasoning blends are carefully formulated for use in prepared foods, snacks, soups, sauces, gravy mixes, salad dressings, etc.

In the industrial market especially, seasoning blends consist of very complex mixtures of spices, herbs, and other ingredients, which might include natural and artificial flavors, dry dairy ingredients, salts, flavor enhancers, and sugars. Such seasoning blends frequently provide a flavor and appearance "signature" as typical of a specific brand as the manufacturer's logo.

2. Uses

Foods commonly seasoned in the retail marketplace are subject to the many subtleties of region and custom. However, a list of foods commonly incorporating seasonings would include:

- Meats – sausages and other cased meats, precooked and preserved (pickled, corned, or salted) products, marinades
- Salad dressings
- Snack foods – potato and corn chips, extruded snacks made from wheat and other grains, popcorn, etc.
- Soups, sauces, and gravies – wet or dry mixes
- Stuffings (bread-based preparations)
- Baked goods – pastry, cookies, frosting, biscuits, crackers

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_515

– Dips and spreads – salsa, cheese and other dairy spreads
– Pickles

3. Resources and Raw Materials

Because the term seasonings encompasses such a broad range of materials, it is difficult to enumerate all the items that might be used. Typically, seasonings would be drawn from the following groups of materials (→ Flavors and Fragrances, Chap. 3.):

– Spices – whole or ground
– Spice extracts – oleoresins or essential oils
– Salts – sodium chloride, sodium phosphates, sodium aluminum phosphates, sodium diacetate, etc.
– Sugars – sucrose, dextrose, lactose, maltose
– Flavor enhancers – monosodium glutamate (MSG), ribonucleotides
– Alliums – onion, garlic, shallots, chives
– Dairy products – cheeses, dry milk, dry whey
– Acids – acetic, lactic, citric, malic
– Roots – ginger, licorice, horseradish
– Hydrolysates – hydrolyzed vegetable protein (HVP), autolyzed yeast
– Natural and artificial flavors – smoke, meat, dairy, etc.
– Vegetable products – tomatoes, peppers, celery, etc.

Many of the ingredients used in seasoning blends are agricultural products, requiring specific environmental conditions, so the supply of these products in consistent volume and quality is a major concern. Many of the traditional businesses or "spice houses" are global in nature, with a network of joint ventures or other business arrangements designed to protect quality and supply in the source country. The source countries are commonly third-world countries, and spice houses often supply both capital and technology in an effort to improve their raw ingredient position. Until quite recently, seasoning ingredients were collected within the source country and then transported to the country of use before cleaning and processing. The current trend is to provide capital and technology for cleaning, processing, and packaging in the country of origin. Pricing and the cost of spice ingredients tend to be volatile, because supplies have historically been unpredictable due to fluctuations in weather, political stability, or the availability of labor.

4. Production

As noted previously, seasoning blends are normally dry materials, and several basic processes are important in their production.

Milling. Hard spices and herbs are commonly milled or ground to reduce the particle size as required for a specific application. For example, salad dressings might require herbs and garlic in the 2 – 4 mm size range, whereas a taco-chip seasoning would require particles sizes of < 0.1 mm. In addition to reducing the particle size, milling and grinding operations release flavoring components, and great care is exercised with regard to temperature and packaging during and after milling or grinding in order to retain volatile flavors to the greatest extent possible.

Spray Drying. Products that are liquids, pastes, or semi-solids often must be converted into dry, free-flowing powders for utilization in seasoning blends. Dairy products are typically spray dried and then introduced in the form of cheese powder, buttermilk powder, whey powder, etc. Spray drying is also useful for protecting essential oils or oleoresins in a dry form. Examples include spray-dried oil of dill, oleoresin black pepper, etc.

Dry Blending. Vertical and horizontal mixers and blenders of all types are utilized to produce seasoning blends. The principal effect is to produce a uniform blend of the various raw materials. Many seasoning components generate considerable dust, so vegetable oils are sometimes incorporated to reduce the dust level. In the case of seasoning blends that contain significant amounts of fine granular materials such as salt or sugar, liquid ingredients (e.g., extracts, oleoresins, and essential oils) can be added directly to the blend and "plated" onto the crystalline compounds. Coloring agents, such as oleoresin paprika, are frequently added to a dry blend as well, but thorough mixing and considerable care

must be invoked to produce uniform and predictable color levels.

Quality control of seasoning blend production requires special care in the following areas:

Raw Material Selection. Since seasoning blends are used to enhance flavor and appearance, the principal quality-assurance techniques involve sensory testing together with analysis of such measurable parameters as particle size, essential oil content, and microbiological levels. Trained panels and controlled standard samples are typically required to assure consistent flavor and appearance.

Formulation. Many of the basic ingredients in seasonings are raw agricultural products, with variable levels of flavor, taste, or color, so some latitude in formulation must be allowed in order to produce a consistent finished product. Where consistency of flavor and appearance is of ultimate concern, the use of isolated essential oils, oleoresins, extracts, and flavors is preferred, as these products can be standardized more easily than raw agricultural products.

Process Control. Processing usually results in elevated temperatures due to friction or heat application, and great care must be taken to protect the heat-sensitive flavoring compounds. Milling is commonly carried out under cryogenic conditions. Blending is often accompanied by cooling or other temperature-minimizing actions, and spray drying is adjusted to protect flavor yield rather than to maximize flavor impact. Relative to other food ingredients, seasonings are very expensive. Protection against flavor loss during processing and shipping is a significant economic consideration in the production process.

5. Storage and Transportation

Seasonings are dry products whose value is defined by their ability to provide flavor enhancement to food. Because flavor compounds are typically volatile and subject to change or loss in storage and transportation, temperature control and the selection of appropriate packaging is a vital concern. Seasoning blends are usually packed in multi-wall paper bags with moisture- and vapor-barrier layers. Fiber drums with plastic liners and vapor-proof lid seals are also common. A storage temperature $< 25\,^{\circ}\mathrm{C}$ is recommended, as is a relative humidity $< 70\,\%$. In warm climates transportation via refrigerated conveyance has been found to be economically justifiable as a means of preserving flavor quality.

6. Quality Specifications

Quality specifications rely heavily on sensory testing to monitor flavor quality and level. It is critical that standard samples be selected jointly by the customer and the supplier as references for use in sensory testing. Such samples are typically packaged in glass containers and held below $-10\,^{\circ}\mathrm{C}$ to minimize flavor change over time. Standard samples must be renewed from time to time by mutual agreement.

Special sensory panels must be selected and trained for evaluating various control parameters. Color control is maintained by the use of tristimulus colorimeters (ASTM E 1347-90, D 2244, E 308; DIN-5033). Typical analytical determinations usually involve salt and sugar. Microbiological analysis is directed toward total plate count, yeast and mold, *Escherichia coli*, and *Salmonella*. Particle-size determinations are also important, with the most common methods based on standard sieve analysis.

7. Economic Aspects

Because the definition of seasonings is so broad, no reliable production or consumption data exist. However, it has been estimated that the average person consumes on the order of $1-2$ kg of seasonings per year, excluding salt. Published figures in the 1990s suggest a total annual consumption in the United States of 3.5×10^5 t; a corresponding figure for Japan is 1×10^5 t.

The list of producers of industrial seasoning blends is long: the United States alone has over 45 major producers, including Borden, Durkee-French, General Spice, Givaudan, Heller Seasonings & Ingredients, Kikkoman, Kraft Food Ingredients, McCormick, Quest International, Spice Islands, and Universal Foods, to name

but a few. Important producers elsewhere in the world include Fuchs, Haarmann & Reimer, and Maizena (Germany); Unilever (Netherlands); Buitoni Perugina and Findus (Italy); Ajinomoto, Itoham Food, Kikkoman Shoyu, and Takeda (Japan); Dragoco, Firmenich, and Nestlé (Switzerland); and Reckitt & Colman, H. E. Daniel, Soutra Foods, and Spicehampton (United Kingdom). McCormick is also represented by a number of international production divisions.

8. References

General References

1. H. D. Belitz, W. Grosch: *Lehrbuch der Lebensmittelchemie,* 4th ed., Springer, Heidelberg 1992.
2. G. Vollmer et al.: *Lebensmittelführer,* vol. 2, Thieme, Stuttgart 1990.
3. K. T. Farrell: *Spices, Condiments, and Seasonings,* AVI Publishing Co., Westport, Conn., 1985.
4. C. Dence: *Season to Taste,* Food Trade Press, Orpington, United Kingdom 1985.
5. N. D. Pintauro: *Food Flavoring Processes, Food Technology Review No.* **32,** Noyes Data Corporation, Park Ridge 1976.
6. A. Askar, H. J. Bielig, H. Treptow: "Gewürze: Eigenschaften, Verunreinigungen, und Qualitätsveränderungen," *Fortschr. Lebensmittelwiss.* 1975, no. 3.
7. H. Melchior, H. Kastner: *Gewürze: Botanische und Chemische Untersuchung,* Paul Parey, Berlin 1974.
8. D. R. Tainter, A. T. Grenis: *Spices & Seasonings: A Food Technology Handbook,* VCH, New York 1993.

Sebacic Acid → Dicarboxylic Acids, Aliphatic
Secondary Cells → Batteries

Sedatives

Barry Dubinsky, The R. W. Johnson Pharmaceutical Research Institute, Spring House, PA 19477-0776, United States

1. **Introduction** 1
2. **Rationale for the Selection of Individual Agents** 2
2.1. **Therapeutic Spectrum of Action and Requirements** 2
2.2. **Mechanism of Action** 2
3. **Individual Agents** 2
3.1. **Benzodiazepines** 2
3.2. **Barbituric Acids** 4
3.3. **Antihistamines** 5
4. **References** 6

1. Introduction

Fabled that the halcyon broods upon her nest floating on the calm sea in the halcyon days.

Halcyon (αλκυων = kingfisher), a bird of which the ancients fabled to breed about the time of the winter solstice in a nest floating on the sea, and that it charmed the wind and waves so that the sea was specially calm during the period [6].

Sedative drugs are used in clinical situations where the induction of a sleepy, calm, tranquil state is the objective of therapy. For example, the sedative and anxiolytic properties of these drugs may be used to moderate excitement and calm the frightened patient who is anticipating surgery. Preanesthetic sedation may even permit a reduction in the administered dose of the general anesthetic and may shorten the time required for recovery from general anesthesia. The ideal agent for conscious sedation should allay anxiety, produce antegrade amnesia, have a rapid onset of action, and a duration of action short enough to permit a rapid recovery. Its dose – response relationship should be well defined and predictable with a broad therapeutic ratio regarding respiratory and cardiovascular safety. Drugs for intravenous or intramuscular administration should be water soluble and should not produce venous irritation or pain at the injection site [1–5].

There are several reasons for administering sedative drugs before the induction of general anesthesia. A primary goal is to reduce anxiety without producing excessive sedation. Sedated patients with cardiovascular disease are less likely to suffer postoperative complications related to myocardial ischemia and large fluctuations in blood pressure. Also, reduction of anxiety may make the patient more cooperative. Probably because so many factors influence postoperative recovery, there may not be a consistent relationship between relief of anxiety and benefit to postoperative recovery; nevertheless, relief of anxiety is humane and should be attempted for every patient [7].

Another goal of preanesthetic sedation is the induction of amnesia for the surgical procedure. Antegrade amnesia for brief procedures, such as endoscopy, is desirable and prepares the patient to accept repeated examinations [8], [9]. However, the induction of amnesia by drugs used for preanesthetic sedation can be a problem, especially in ambulatory patients [7]. The frequency of occurrence of amnesia in postsurgical patients is directly related to the administered dose of drugs, such as the benzodiazepines. Amnesia may be observed occasionally when sedative doses of benzodiazepines are used alone. It may, however, also occur in the majority of patients who receive benzodiazepines in combination with other drugs that can be used as preanesthetic adjuncts such as the anticholinergic agent, scopolamine, or meperidine (pethidine), an analgesic agent with anticholinergic properties.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_519

2. Rationale for the Selection of Individual Agents

2.1. Therapeutic Spectrum of Action and Requirements

Since sedation may be achieved by drugs belonging to diverse chemical classes that work by different pharmacological mechanisms, the sedatives are not a sharply defined group of drugs. The choice of the individual drug depends upon its therapeutic spectrum and the clinical situation in which it is to be used. This review is restricted to drugs used for preanesthetic sedation and does not include agents that are used to relieve preoperative pain with or without the induction of accompanying sedation. Various drugs, including the barbiturates, phenothiazines, butyrophenones [10], and the narcotic analgesics, have been used for the relief of anxiety and for their sedative effects prior to surgery but, they have been supplanted largely by the benzodiazepines [1–5], [7]. The preanesthetic use of narcotic analgesics is limited principally to relief of preoperative pain, to ameliorate pain associated with regional anesthesia, and for attenuation of the cardiovascular response to endotracheal intubation [7]. Certain neuroleptics such as the phenothiazine, chlorpromazine; the butyrophenones, haloperidol or droperidol; and the analgesics, fentanyl, alfentanil, or sufentanil are used often as general anesthetic adjuncts. For example, potent analgesics are used in combination with small doses of neuroleptics to induce neuroleptanalgesia. Preanesthetic use of individual neuroleptic or analgesic drugs is not discussed herein; this subject has been adequately covered elsewhere (→ Anesthetics, General).

2.2. Mechanism of Action

Benzodiazepines and barbiturates modulate the affinity of the oligomeric receptor for the inhibitory neurotransmitter, γ-aminobutyric acid (GABA). Enhanced GABA binding results in an increase in intraneuronal chloride ion concentration with reduced neuronal excitability. Sedative doses of the benzodiazepines inhibit neuronal transmission within forebrain limbic regions but higher doses also depress brainstem loci that modulate skeletal muscle tone. The barbiturates directly depress the sensory and motor cerebral cortex and reduce the excitability of the reticular activating system (RAS). See also → Neuropharmacology, Chap. 3.2.3., → Neuropharmacology, Chap. 3.2.4. The phenothiazines and butyrophenones block receptors for dopamine. This results in a reduction of afferent transmission to the brainstem RAS by inhibition of a parallel system. The sites of the sedative action of the antihistamines and the opioids are unclear and are probably diffuse. The euphoriant actions of the opioids probably involve hypothalamo – cortico – limbic structures while their analgesic effects are localized within brainstem nuclei which inhibit nociceptive afferent input.

3. Individual Agents

3.1. Benzodiazepines

Benzodiazepines possess a wider therapeutic margin of safety than other agents and are used worldwide for preanesthetic sedation. The range of doses for their sedative and hypnotic effects is broad. The range of doses in the following paragraphs have been recommended for adults, doses for children should be lower. Compared to the barbiturates, the benzodiazepines are safer with regard to their effects on cardiovascular and respiratory function. However, caution should be exercised in the elderly, in patients with chronic liver disease, or when they are used in combination with other CNS depressant agents. In addition, the benzodiazepines have been developed for use as psychopharmacological agents (→ Psychopharmacological Agents, Chap. 4.1.), anticonvulsant agents (→ Antiepileptics, Chap. 2.6.), hypnotics (→ Hypnotics, Chap. 5.4.), and skeletal muscle relaxants (→ Skeletal Muscle Relaxants).

Diazepam [*439-14-5*] has been in use as an intravenous sedative since the early 1960s.

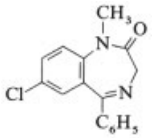

It is very effective for the relief of anxiety before surgery. Its primary use is for oral premedication in doses of 5 – 15 mg. Although it has a very rapid onset of action when administered parenterally (intravenously or intramuscularly), its anxiolytic and sedative effects do not onset until 1 h after oral administration. It is effective for several hours and, as with other benzodiazepines, its elimination half-life is increased in the elderly. Diazepam is metabolized to desmethyldiazepam which is pharmacologically active. Formerly, diazepam was administered extensively in intravenous doses of 0.1 – 0.3 mg/kg. Its use by this route has declined because the injectable formulation contains propylene glycol which may produce venous thrombosis, local irritation, swelling, and vascular impairment. Because of the possibility of precipitation in the infusion fluid, diazepam injection should not be diluted and administered as an intravenous infusion. Also, it is unstable in poly(vinyl chloride) infusion containers and tubing. Diazepam is effective when doses of 0.1 – 0.3 mg/kg are administered intramuscularly but its effects are not as reliable by this route. It is commercially available as tablets, sustained-release capsules, oral solutions, oral drops, and an injectable solution.

Trade Names. Diazepam (various manufacturers); Diazepam Oral Solution, Diazepam Intensol (Roxane); Valium, Valrelease (Roche); Zetran (Hauck).

Lorazepam [*846-49-1*] is a benzodiazepine with greater sedative and amnestic activity and potency than diazepam.

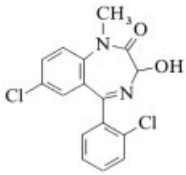

Sedation can be obtained in most adults after administration of 0.044 mg/kg intravenously or 0.05 mg/kg intramuscularly. The injectable product is formulated with a mixture of poly(ethylene glycol) 400 and propylene glycol and is tolerated better than diazepam when administered intravenously. Even when injected, the onset of action of lorazepam may be delayed for up to 40 min. The duration of its sedative and amnestic action may, however, be three to four times that of diazepam. Lorazepam is metabolized to a biologically inactive product by conjugation with glucuronic acid.

Trade Name. Atavan (Wyeth-Ayerst).

Triazolam [*28911-01-5*] is a benzodiazepine that is primarily used as a sleep-inducing hypnotic agent rather than as an anxiolytic.

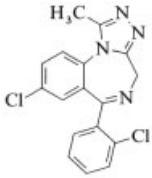

Triazolam has sedative activity and is typically administered orally (0.125 – 0.25 mg), however, up to 0.5 mg may be administered for sleep induction. Anesthetic requirements may be considerably reduced because of its high degree of efficacy as a CNS depressant. Since triazolam has a relatively short half-life (1.5 – 3 h), it may be administered for preoperative sedation when a short-acting agent is desired. Peak plasma levels are obtained within 2 h. It is metabolized to a biologically active α-hydroxylated product. Benzodiazepines with short half-lifes generally do not have residual effects; nevertheless, the patient should be held in the recovery room until locomotor ataxia or amnestic effects have abated.

Trade Name. Halcion (Upjohn).

Midazolam [*59467-70-8*]. The hydrochloride is a water-soluble, benzodiazepine CNS depressant that is formulated for intravenous or intramuscular administration.

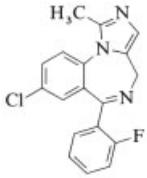

It does not produce venous irritation when injected. It has a more rapid onset of action, shorter duration of action, higher potency, and more

rapid elimination half-life than other benzodiazepines. When used for conscious sedation, its CNS effects are dependent upon the dose and route of administration. Midazolam has anxiolytic activity and is capable of inducing amnesia for the surgical procedure in most patients. It is metabolized to 1-hydroxymethylmidazolam, which is then conjugated with glucuronic acid; both steps proceed rapidly. Its onset of action occurs approximately 15 min after intramuscular administration and recovery may be complete within 2 h. Since intravenous administration has been associated with respiratory depression and arrest, dosage should be individualized, especially for the elderly. One way to accomplish this is by administering the drug slowly over at least 2 min or until the patient's speech becomes slurred. This effect may be accomplished with as little as 1 mg. The maximum intravenous or intramuscular dose per patient is 5 mg.

Trade Name. Versed (Roche).

3.2. Barbituric Acids

The barbiturates such as pentobarbital, secobarbital, and phenobarbital produce sedation at doses that are lower than their hypnotic doses. See also →Hypnotics, Chap. 5.1. However, these agents are used only occasionally as sedatives because they may have a narrow therapeutic window, i.e., they can produce respiratory and cardiovascular depression when administered at relatively small multiples of their sedative or hypnotic doses. Tolerance to the sedative effects of these agents develops rapidly when they are administered repeatedly. Therefore, tolerance may be observed in patients who have been taking barbiturates, alcohol, or other CNS depressant drugs. Barbiturates are metabolized primarily by the hepatic microsomal enzyme system and induction of this system contributes to the development of tolerance. Barbiturates should not be given to patients who are experiencing pain because they may cause disorientation rather than sedation. Their use for preanesthetic sedation has been largely supplanted by the benzodiazepines.

Pentobarbital [*76-74-4*] is a short-acting barbiturate for preanesthetic sedation that may be administered orally or by deep intramuscular injection.

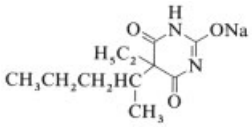

No more than 5 mL should be injected into a single intramuscular site because of its tissueirritating properties. Doses are calculated on the basis of the patient's age, weight, and physical condition. Cardiovascular or respiratory depression is minimal at sedative doses. Pentobarbital solution is strongly basic so its intravenous use should be restricted to conditions in which other routes of administration are not feasible as media with high pH values injure the lining of the vein. When intravenous administration is imperative because prompt sedation is required, the rate of administration should not exceed 50 mg/min.

Trade Names. Nembutal, Nembutal Sodium (Abbott).

Secobarbital (sodium salt) [*309-43-3*] is a short-acting barbiturate for preanesthetic sedation that may be administered orally or by deep intramuscular injection.

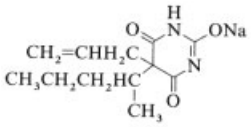

No more than 5 mL should be injected into a single intramuscular site because of its tissue-irritating properties. Doses are calculated on the basis of the patient's age, weight, and physical condition. Cardiovascular or respiratory depression is minimal at sedative doses. Since secobarbital sodium solution is strongly basic, its intravenous use should be restricted to conditions in which other routes of administration are not feasible. When intravenous administration is imperative because prompt sedation is required, the rate of administration should not exceed 50 mg/15 s. Secobarbital sodium injectable solution is formulated with 50 % poly(ethylene glycol) and 0.25 % phenol.

Trade Name. Seconal Sodium (Lilly).

Phenobarbital [*50-06-6*] is a long-acting barbiturate for preanesthetic sedation that may be administered orally or by deep intramuscular injection into a large muscle.

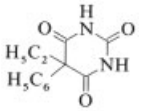

Care must be taken to avoid injecting solutions into or near peripheral nerves or arteries because of the possibility of permanent tissue damage. Doses are calculated on the basis of the patient's age, weight, and physical condition. Luminal sodium injectable solution is formulated with 10 % alcohol and 67.8 % propylene glycol.

Trade Names. Barbita (Vortech), Luminal Sodium (Winthrop-Breon), Solfoton (Poythress).

3.3. Antihistamines

Histamine-blocking activity was first detected in1937in one of aseries of amines with aphenolic ether function synthesized by FOURNEAU. This compound was found to be useful in treating anaphylactic shock and some derivatives were found to possess highly potent and clinically useful antihistaminic activity. Since many antihistaminic drugs (→ Antiallergic Agents, Chap. 2.) also have anticholinergic and sedative effects, they have been found to be effective general anesthetic adjuncts. When used for preanesthetic sedation, the range of doses mentioned in the following paragraphs are recommended for adults. Doses for children should be lower.

Hydroxyzine [*68-88-2*] is a sedative piperazine antihistamine with anxiolytic, antiemetic, antisecretory, and bronchodilator effects.

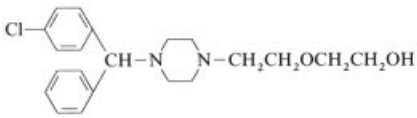

Hydroxyzine hydrochloride is administered orally in a dose range of 50 – 100 mg for preanesthetic sedation; however, it is more commonly administered intramuscularly in a dose range of 25 – 100 mg. While this drug potentiates other CNS depressants and lowers the required amount of the general anesthetic, it does not prolong anesthesia. Sedative doses have minimal effects on respiration or cardiovascular function.

Trade Names. Atarax (Roerig), Vistaril (Roerig-Pfizer).

Diphenhydramine [*58-73-1*] is a sedative antihistamine belonging to the ethanolamine class.

It also has antiemetic and antiparkinson effects. It may be administered intravenously or intramuscularly in a dose range of 10 – 50 mg for preanesthetic sedation.

Trade Name. Benadryl (Parke-Davis).

Promethazine [*60-87-7*] is a phenothiazine and has a long history of use.

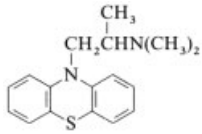

It was introduced in 1946 for use in allergic conditions and has been used as a preanesthetic medication for its sedative and antiemetic effects. Promethazine hydrochloride [*58-33-3*] may be administered orally in a dose range of 25 – 50 mg. The preferred parenteral route is by deep intramuscular injection at a similar dose. The duration of action of promethazine is 4 – 6 h.

Trade Name. Phenergan (Wyeth-Ayerst).

Propiomazine [*362-29-8*] is a phenothiazine that is a short-acting sedative, antiemetic, antihistamine.

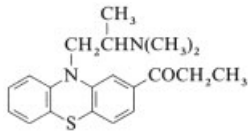

Propiomazine hydrochloride may be administered by the intravenous or intramuscular route in a dose of 20 – 40 mg.

Trade Name. Largon (Wyeth-Ayerst).

4. References

General References

1. B. E. Marshall, D. E. Longnecker in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 285.
2. T. W. Rall in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 345.
3. R. J. Baldessarini in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 383.
4. B. R. Olin et al. (eds.): *Drug Facts and Comparisons,* J. B. Lippincott, St. Louis 1991.
5. E. R. Barnhart: *Physician's Desk Reference,* 45th ed., Medical Economics Company, Oradell 1991.

Specific References

6. *The Compact Edition of the Oxford English Dictionary,* Oxford University Press, Oxford 1971.
7. J. L. Lichtor in R. D. Miller (ed.): *Anesthesia,* 3rd ed., Churchill Livingstone, New York 1990, p. 895.
8. P. M. Lauven, *Scand. J. Gastroenterol.* **25** (1990) suppl. 179, 1 – 6.
9. R. F. McCloy, R. C. Pearson, *Scand. J. Gastroenterol.* **25** (1990) suppl. 179, 7 – 11.
10. M. M. Clark, *Anesthesia* **24** (1969) 36 – 37.

Selenium and Selenium Compounds

BERND E. LANGNER, Norddeutsche Affinerie AG, Hamburg, Federal Republic of Germany

1. **Introduction** 1
2. **Properties** 1
3. **Occurrence and Raw Materials** 4
4. **Production** 4
4.1. **Primary Production** 4
4.2. **Secondary Production** 7
4.3. **Refining** 7
5. **Environmental Protection** 8
6. **Uses** 8
7. **Quality Specifications and Analysis** . 10
8. **Selenium Compounds** 10
9. **Economic Aspects** 12
10. **Biological Activity and Toxicology** . . 12
11. **References** 13

1. Introduction

Selenium [*7782-49-2*], atomic number 34, A_r 78.96, is a member of group 16 of the periodic table and thus belongs to the chalcogens. Consonant with its position between the nonmetal sulfur and the metalloid tellurium, it has mainly nonmetallic properties.

Naturally occurring selenium consists of six stable isotopes: ^{74}Se (0.9 %), ^{76}Se (9.0 %), ^{77}Se (7.6 %), ^{78}Se (23.5 %), ^{80}Se (49.8 %), and ^{82}Se (9.2 %). Selenium is 66th in order of natural abundance in the earth's crust, with an average concentration of 0.05 ppm.

Selenium was discovered in 1817 by BERZELIUS in deposits in lead chambers used for the production of sulfuric acid from copper pyrites. The name selenium is derived from the Greek word for moon and reflects the close resemblance between selenium and tellurium (= earth), which was discovered about 20 years earlier.

Selenium is a byproduct of the extraction of metals such as copper, zinc, and lead, and has many industrial applications based on its chemical and physical properties. Its semiconducting properties and the light-induced variation of its electrical conductivity are of particular importance.

2. Properties [1–4]

Like sulfur, selenium exists in several allotropic forms. The structures of these and the mechanism of transformation of one into another are not completely understood. Red selenium is a thermodynamically unstable, high-energy modification produced by precipitation at low temperature, crystallization of selenium from solution, or quenching of selenium vapor or molten selenium. The three modifications of monoclinic selenium (α-, β-, and γ-selenium) consist of Se_8 rings and differ only in the packing of these in the crystal (Fig. 1). The red, amorphous form consists of chains in an irregular arrangement. It is soluble in carbon disulfide and to some extent in benzene, giving red solutions from which it can be crystallized. Red selenium is an electrical nonconductor. In the vapor phase, selenium exists mainly as Se_8 rings. At higher temperature these decompose into smaller units, eventually forming Se_2 molecules.

Thermodynamically unstable vitreous or amorphous forms are obtained by rapid cooling of molten selenium. The structure of these modifications is not yet fully understood. Presumably, high molecular mass selenium rings of various sizes are present, which explains the poorly defined melting point of industrial-grade vitreous selenium. This product softens at ca. 50 °C to a rubberlike state and is transformed at > 100 °C into the gray, crystalline form, with release of the heat of crystallization. The black, nonconducting form is slightly soluble in carbon disulfide. Structures similar to those in vitreous selenium are present in molten selenium.

On heating to > 100 °C, both the red and the black forms are transformed into the stable, gray, hexagonal ("metallic") form in an exothermic reaction. This modification is also produced by reducing selenous acid with sulfur dioxide at high temperature. Unlike the ring structure of the unstable forms, it consists of helical chains

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_525

in which the structure is repeated at every third atom (Fig. 1). Gray selenium is a semiconductor whose electrical conductivity increases by a factor > 1000 under the influence of light (photoconductivity).

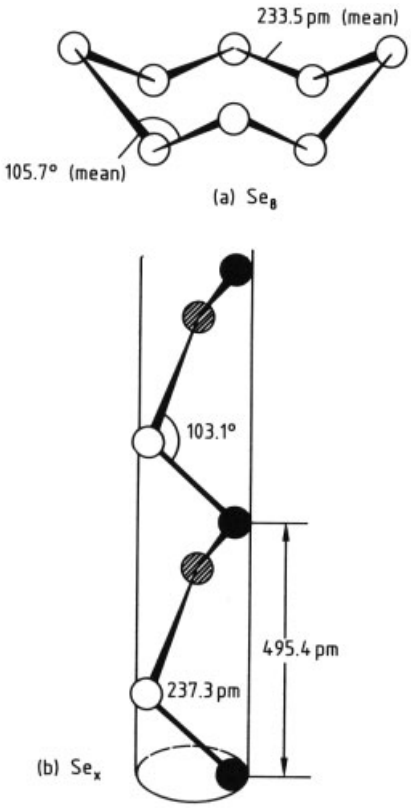

Figure 1. Structures of selenium
a) Se_8 molecule in red α-, β-, and γ-selenium; b) Helical chains in hexagonal gray selenium [4]

Impurities have a great influence on crystallizing properties, so results given in earlier research work, where impurity content is not quoted, are often erroneous. Also, the previous history of the material (e.g., high-temperature treatment) plays an important role in the transformations between various allotropic forms. The thermal properties of selenium can be conveniently investigated by differential thermal analysis [8]. Figure 2 shows the various transformations of distilled selenium on melting. The brittle vitreous form is transformed at 50 °C into the ductile vitreous form (glass transition). Above ca. 120 °C, vitreous selenium changes into the stable hexagonal crystalline form, which melts at 220 °C.

Important physical properties of selenium are given below [1,4]:

Atomic number	34
Electronic structure	$[A_r]3d^{10}4s^24p^4$
Relative atomic mass	78.96
Oxidation states	−II,0,+IV,+VI
Atomic radius, nm	0.140
Ionic radii, nm	
Se^{2-}	0.191
Se^{4+}	0.050
Se^{6+}	0.042
Ionization energy, kJ/mol	940
Ionization potential, eV	
I	9.75
II	21.5
Electronegativity (Pauling)	2.4
Density at 25 °C, g/cm^3	
Hexagonal	4.189
α-Monoclinic	4.389
Vitreous	4.285
Liquid (490 K)	3.975
Melting point, °C	217
Boiling point, °C (101.3 kPa; 760 mm Hg)	684.7 (±0.03)
Vapor pressure, kPa	
344.6 °C	0.133 (1 mm Hg)
431.5 °C	1.333 (10 mm Hg)
547.4 °C	13.33 (100 mm Hg)
636.5 °C	53.32 (400 mm Hg)
Electrical resistivity, Ω/cm∗	
25 °C	10^{10}
400 °C	1.3×10^5
Glass transition temperature, °C (vitreous modification)	30.2∗∗

∗ Strongly dependent on purity and light intensity.
∗∗ Strongly dependent on purity.

The electrical conductivity is strongly dependent on purity. Thus, the conductivity of selenium, including the poorly conducting amorphous form, can be increased considerably by traces (a few parts per million) of halide ions or by alloying elements such as tellurium or arsenic. Because the electrical properties are sensitive to the presence of crystal defects, the purity of selenium, which can be deliberately controlled by doping or alloying with other elements, plays an important role in its use in electrical and electronic applications.

The chemical properties of selenium are intermediate between those of sulfur and those of the metalloid tellurium. In its compounds, it mainly exhibits the oxidation states −II, +IV, and +VI. The oxidation state +VI is considerably less stable than for sulfur, and in acid solutions is attained only by the action of strong

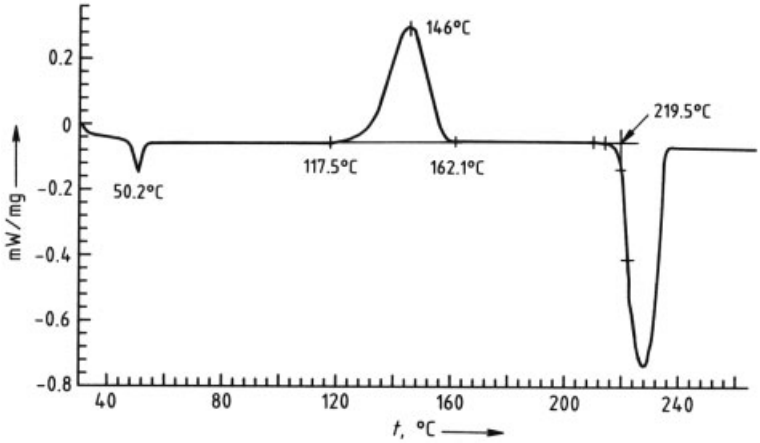

Figure 2. Differential thermal analysis of distilled high-purity selenium after 3 h at 450 °C to give a material of uniform structure [8]

oxidizing agents such as chlorine or hydrogen peroxide. Selenate(VI) is more easily obtained by oxidation in alkaline solution. The redox potentials of selenium in acid and alkaline solution are as follows [4, 7]:

Oxidation State	−II		0		
Acid	H_2Se	$\xrightarrow{-0.40\ V}$	Se	$\xrightarrow{+0.74\ V}$	
	+IV		+VI		
	H_2Se	$\xrightarrow{+1.15\ V}$	SeO_4^{2-}		

	−II		0		
Alkaline	Se^{2-}	$\xrightarrow{-0.92\ V}$	Se	$\xrightarrow{-0.37\ V}$	
	+IV		+VI		
	SeO_3^{2-}	$\xrightarrow{+0.05\ V}$	Se^{2-}		

Selenium burns in air to form selenium dioxide, which has the smell of rotten radishes. Selenium is oxidized by nitric acid to selenous acid. No reaction occurs with nonoxidizing acids such as hydrochloric acid. Hot, concentrated sulfuric acid dissolves selenium, giving a green color and forming polymeric selenium cations, for example:

$$Se_8 + 3\,H_2SO_4 \longrightarrow Se_8^{2+} + 2\,HSO_4^- + SO_2 + 2\,H_2O$$

With chlorine, vigorous oxidation to selenium tetrachloride occurs. Selenium reacts with electropositive elements (e.g., many metals) to form selenides and is therefore strongly corrosive, especially at high temperature. With hydrogen, the toxic gas hydrogen selenide is formed. Selenium combines with sulfur to form nonstoichiometric alloys.

Selenium dissolves in solutions of strong alkali, disproportionating into selenide and selenite. It dissolves in alkali-metal sulfides, sulfites, and cyanides, with formation of addition compounds.

3. Occurrence and Raw Materials

Pure selenium minerals are very rare. Thus, berzelianite (Cu_2Se), tiemannite (HgS), and naumannite (Ag_2Se) are never processed simply to obtain selenium. However, selenium occurs in various concentrations as an accompanying element in many sulfidic ores, although isomorphic substitution in copper pyrites, iron pyrites, or zinc blende cannot exceed a theoretical maximum of 0.5 % because of the larger ionic radius of selenium (Se^{2-}: 0.191 nm, S^{2-}: 0.174 nm). The selenium content of pyrites, zinc blende, galena, and molybdenite is too low for economic recovery. However, most copper concentrates from ore flotation contain ca. 100 – 400 g/t selenium and are hence the most important sources of selenium. Selenium production is therefore closely linked to copper production.

Selenium occurs in low concentrations (ca. 1 – 5 g/t) in some types of coal and mineral oil.

In addition to extraction of selenium from copper ore, the recycling of photocopying drums and rectifiers is becoming an important source.

4. Production

4.1. Primary Production

Selenium is an accompanying element obtained during copper production (→ Copper). The sulfidic copper concentrate (e.g., chalcopyrite) is smelted in an oxidizing atmosphere to give a copper matte containing 40 – 70 % Cu. This is then blown in a converter to yield crude copper. After reduction of excess copper oxide, the crude copper is cast into anodes. Along with other impurities, these copper anodes contain 25 – 70 % of the selenium, depending on the smelting process. The rest of the selenium vaporizes during smelting, mainly as selenium dioxide. This is reduced to the element by the sulfur dioxide produced during smelting and collects in the slime formed in the gas purification equipment. However, recovery is generally uneconomical due to the complex composition of the slime. Likewise, in the roasting of pyrites and zinc or lead ore, most of the selenium volatilizes and concentrates in the slime in the gas purification equipment.

The selenium remaining in the copper anodes is mainly in the form of Cu_2Se, or sometimes CuAgSe and Ag_2Se, depending on the silver content. Because they are insoluble in the sulfuric acid electrolyte, these selenides remain in the anode slime during electrolytic refining of the copper, as do the noble metals. As a result of the high stability of noble-metal selenides, the copper selenides initially present in the anodes can be converted to silver selenide or gold selenide under the influence of the electrolyte, depending on the noble-metal content. The selenium in the anode slime is therefore in the form of various compounds, depending on the silver content. The selenium content of the anode slime varies with the raw material and can reach > 20 %.

Selenium contents (%) of anode slimes (from which residual copper has not been removed) from various copper tankhouses are as follows [9]:

CCR (Noranda), Canada	10.0
Chuquicamata, Chile	4,0
Copper Refineries Pty, Australia	0.7
IMI Refiners, United Kingdom	2.0
Inco, Canada	8.4
Kidd Creek, Canada	19.5
MHO, Belgium	5.1
Norddeutsche Affinerie, Germany	7.5
Palabora, South Africa	3.6
Phelps Dodge, United States	8.8
Southwire, United States	0.6

The precious metals can be recovered only by decomposing the very stable selenides with very strong acid or alkali, or by roasting. Selenium must be removed from the anode slime before the noble metals can be recovered. The important processes are [10]:

1) Fusion with sodium carbonate
2) Roasting with sodium carbonate
3) Direct oxidation (roasting)
4) Oxidative leaching with sodium hydroxide solution under pressure
5) Roasting with concentrated sulfuric acid
6) A hydrometallurgical process of chlorination in hydrochloric acid

In roasting with sodium carbonate, the anode slime is first treated with sulfuric acid solution to remove the copper and then mixed with sodium carbonate and water to give a stiff paste. The paste is extruded or pelletized, and dried. It is then roasted at 530 – 650 °C in a stream of oxygen – air. Sodium carbonate prevents vaporization of the selenium as selenium dioxide by converting it completely into sodium selenate. The roasted pellets are ground and leached with water to dissolve sodium selenate. The precious metals remain in the solid residue, which also contains tellurium in the form of sodium tellurate if higher roasting temperatures are used. The selenate is crystallized from solution and reduced with active carbon to sodium selenide, which is redissolved in water and oxidized with air to elemental selenium. However, this process leads to the formation of large amounts of wastewater.

Elemental selenium can also be precipitated from selenate solution by strongly acidifying with hydrochloric acid and reducing with iron(II) salts or sulfur dioxide. Alternatively, the sodium carbonate solution can be decomposed by adding acid to precipitate impurities such as tellurium. After filtration and further acidification,selenium is precipitated by reduction (Fig. 3).

In the direct oxidation process, copper is removed first from the anode slime by sulfuric acid. The slime is then dried and mixed with inert materials that prevent sintering, and the sele-

nium is volatilized as selenium dioxide vapor at 600 – 900 °C. Silver selenite is formed initially at < 400 °C

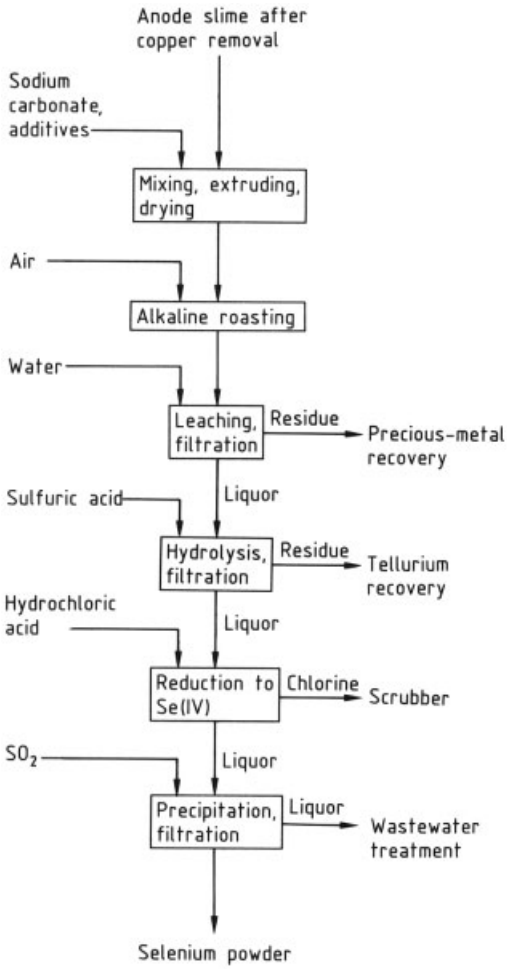

Figure 3. Recovery of selenium from copper anode slime by roasting with sodium carbonate

$$Ag_2Se + 1.5\,O_2 \longrightarrow Ag_2SeO_3$$

and decomposes at higher temperature

$$Ag_2SeO_3 \longrightarrow 2\,Ag + SeO_2 + 0.5\,O_2$$

Selenium dioxide and some of the sulfur dioxide formed from the sulfate fraction of the anode slime are removed from the gas phase by scrubbing with water or sodium hydroxide solution. After acidifying the wash liquor, the selenium is reduced with sulfur dioxide (Fig. 4). Very high volatilization rates are achieved at < 600 °C if the oxidative roasting is carried out in the presence of sulfur dioxide. The reaction is believed to occur via the intermediate formation of sulfur trioxide, which reacts with silver selenite:

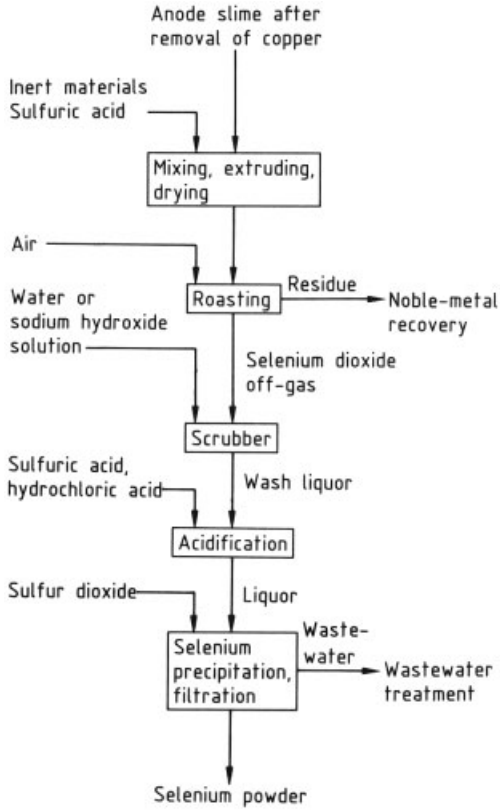

Figure 4. Recovery of selenium from copper anode slime by acidic roasting

$$Ag_2SeO_3 + SO_3 \longrightarrow Ag_2SO_4 + SeO_2$$

This method is used by Outokumpu [11]. The selenium dioxide produced is absorbed in dilute sulfuric acid and immediately precipitated as elemental selenium by the excess sulfur dioxide in the scrubbing equipment.

Whereas roasting is carried out below the melting point of the anode slime, in the classical Doré process (Fig. 5) the anode slime is melted after addition of fluxes. A matte consisting mainly of silver selenide and a high-lead slag containing precious metals are formed. The slag can be processed in a lead smelting works, or the noble metals can be recovered by flotation.

The selenium-containing matte is oxidized in the molten state in a second stage, and selenium is vaporized as its dioxide and absorbed in a scrubbing stage. In a process operated by the Canadian copper producer Noranda, these two stages are carried out in succession in a top blown rotary converter (TBRC) [12].

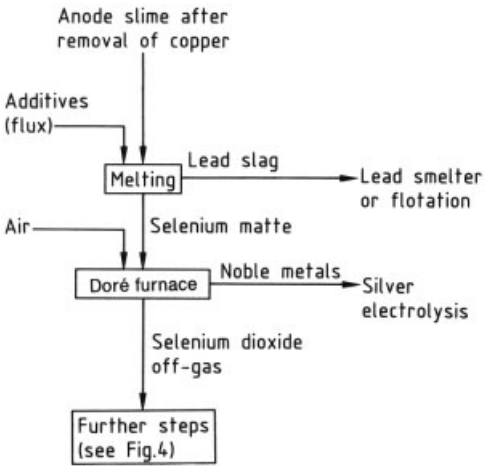

Figure 5. Recovery of selenium from copper anode slime by the Doré process

Alternatively, roasting with sulfation is carried out in the presence of concentrated sulfuric acid at 500 – 600 °C after removal of tellurium [9], although the large quantities of sulfur dioxide and sulfur trioxide aerosols formed can present environmental problems, and the severely corrosive properties of sulfuric acid at high temperature are difficult to deal with.

Apart from the widely used high-temperature processes, hydrometallurgical extraction of selenium from copper anode slimes is of some importance in small plants [13–15]. For example, in the copper refinery Austria-Metall [14], the anode slime is leached with hydrochloric acid and chlorine, whereby the gold, platinum metals, and selenium (as selenous acid) dissolve, and silver remains in the residue as silver chloride. Selenium is then precipitated from the hydrochloric acid solution by reduction with SO_2.

4.2. Secondary Production

Although most selenium comes from primary production, recycling of the selenium in used rectifiers and photocopying drums is becoming increasingly important for environmental reasons. An older recovery process involves fusion with sodium hydroxide to dissolve the selenium, forming polyselenides:

$$6\,NaOH + (2x+1)\,Se \longrightarrow 2\,Na_2Se_x + Na_2SeO_3 + 3\,H_2O$$

Selenium is obtained by acidifying the solution, but toxic hydrogen selenide may be formed. This can be prevented if an alkaline solution of sodium sulfite is used in the treatment process. This dissolves selenium selectively, forming sodium selenosulfate:

$$Na_2SO_3 + Se \longrightarrow Na_2SeSO_3$$

Selenium can be recovered by acidification, which produces SO_2.

In another process, selenium is treated with sodium hydroxide solution and hydrogen peroxide to give a mixture of sodium selenite and sodium selenate. The solution is acidified, and selenium is precipitated by sulfur dioxide [16].

Selenium can also be removed mechanically from photocopying drums by using a water jet. The aluminum drum can be reused whether selenium is removed chemically or mechanically; the selenium on photocopying drums is thus 100 % recyclable without any deterioration of its quality. Recycling of photocopying drums is performed on a large scale (e.g., by the company Retorte).

Selenium can be dissolved selectively from residues by leaching with ammonia or amines in an aprotic solvent in the presence of carbon monoxide, to form amine salts of selenocarbamic acid. It can then be precipitated in a pure state by heating, with recovery of carbon monoxide and the amine [17].

4.3. Refining [3]

The roasting processes, followed by precipitation with sulfur dioxide, gives selenium with a purity usually exceeding 99 %, which is adequate for many chemical applications. However, for electronic applications or xerography, selenium must be purified further, because of the sensitivity of its physical properties to the presence of impurities.

The simplest method of purification is distillation at atmospheric pressure and ca. 650 °C, which gives a purity up to 99.99 %. All the nonvolatile elements, such as copper, nickel, iron, and tellurium (if present as the dioxide), remain in the distillation vessel. However, sulfur and mercury (as HgSe) also distill over and are thus not removed. Therefore, the crude selenium should not contain mercury. A sufficiently pure product is obtained if some of the selenium is not distilled, ca. 50 % being left behind in the distillation vessel to be returned later to the recovery process. The corrosive properties of selenium vapor cause problems during distillation, and purified selenium that is collected in cast iron vessels sometimes becomes contaminated with iron. Improved separation is obtained by vacuum distillation at 0.7 – 1.3 kPa and 360 – 400 °C, or at 0.1 Pa and 260 °C, which also reduces corrosivity [18]. A further improvement in quality is obtained if the melt is selectively oxidized by atmospheric oxygen before distillation. This gives better separation of the more readily oxidized impurities.

An alternative to distillation for producing high-quality selenium is vaporization with oxidation, the selenium dioxide vapor formed being absorbed in pure water. Mercury can be removed from the solution by adsorption, and the other metallic ions by ion exchangers. Elemental selenium is then precipitated with high-purity sulfur dioxide. Purities > 99.99 % can be achieved in this way [19]. After precipitation, the selenium powder is washed and can be remelted if desired (e.g., to produce granules or to homogenize the structure).

High-purity selenium is also obtained by the hydride process. Impure selenium is treated with hydrogen at ca. 550 – 685 °C to form hydrogen selenide, which is then decomposed at 1000 °C. Alternatively, it can be cooled to remove gaseous impurities before thermal decomposition is carried out [20].

5. Environmental Protection

Selenium is toxic at high concentration and must therefore be effectively removed from off-gases and wastewater. Since oxidizing processes are used, selenium is present in the off-gas as selenium dioxide. After cooling, it can be removed either by dust filters or, because of its high solubility, by absorption in water or sodium hydroxide solution in gas-scrubbing equipment. Because selenium dioxide is readily reduced by organic compounds or sulfur dioxide, the off-gas always contains a small amount of elemental selenium. The selenium-containing dust filtered out thus generally has a reddish color.

The wastewater treatment process and the minimum attainable selenium content both depend on the form in which selenium is present and on the other components in the wastewater. Although selenites in acid solution are reduced readily to insoluble selenium, complete removal is often difficult, because the selenium sometimes precipitates as a colloid and the reduction is very slow, especially in dilute solution. Therefore, precipitation aids such as hydroxides of aluminum or iron must be added usually. Adsorptive precipitation then enables a final content of <10 mg/L Se to be achieved in the wastewater.

Selenium-containing wastewater, as produced in the sour waters from petroleum refining, can be purified by ion exchange [21].

6. Uses [5,6]

The uses of selenium and selenium compounds can be divided into the following categories [22]:

Electronic applications	ca. 25 %
Pigments	ca. 10 %
Glass	ca. 40 %
Metallurgy	ca. 10 %
Agricultural and biological applications	5 %
Other uses (e.g., additives for rubber vulcanization, oxidation catalysts)	ca. 10 %

Rectifiers. Selenium conducts electricity preferentially in one direction, and this property has been utilized since 1920 for converting alternating to direct current. A selenium rectifier consists of a steel or aluminum plate coated with nickel, with an additional coating of selenium, 50 – 60 µm thick. A sprayed cadmium alloy is used as a counterelectrode. Several of these rectifiers can be stacked in series. The efficiency of rectification is ca. 85 %. Although selenium rectifiers have been increasingly replaced by germanium and silicon diode rectifiers, selenium is still used where the performance of the rectifier

must be maintained even in the presence of high countervoltage peaks.

Xerography (→ Imaging Technology , Chap. 2.1.1.) [23]. The use of selenium in photocopying technology is based on its sensitivity to light. Selenium, which may be highly pure, doped, or alloyed, is deposited by a vaporization process to form a film, ca. 50 µm thick, on an aluminum drum. In the copying process, the selenium surface is first charged by a high-voltage corona discharge. The drum is then exposed to the projected light from the picture. Where the light falls on the drum, the charge on the selenium flows away, and a latent picture is thereby formed. After the illumination stage, the thermoplastic toner powder is applied to the drum, attracted electrostatically to the charged parts (not illuminated), and then fixed by heat. A photograph of a used photocopier drum after 60 000 photocopying runs is shown in Figure 6. The pure, amorphous selenium originally used to coat the drums is very fragile and crystallizes at high temperature. Therefore, selenium stabilized with tellurium or arsenic is now often used (selenium – tellurium alloys or arsenic triselenide), generally doped with small amounts of chlorine. The alloys containing high-purity tellurium enable copying rates to be increased and light of a wider spectral range to be used. A still-higher copying rate and wider spectral range are possible with arsenic triselenide, which has been used increasingly despite its toxicity and problems in manufacture. The drums also have a considerably longer life. The selenium-based system is now being replaced by organic photoconductors and amorphous silicon in low-cost applications. However, selenium is maintaining its position in industrial applications because of its long life and high photocopying capacity. Whereas a drum with organic photoconductors has a lifetime of ca. 10 000 – 30 000 copies, drums coated with selenium can withstand more than 100 000 copying operations.

A potential application of xerography with selenium as the photoconductor is in medicine. Here, X-rays are used instead of visible light to form the latent image. The advantage of this process lies in the possibility for direct digitization without producing a copy or a film, so that the X-ray image is generated directly by a computer on a visual display unit, and can be stored and processed. Although several patents exist and comprehensive research has been carried out, the process is not yet in general use [24, 25].

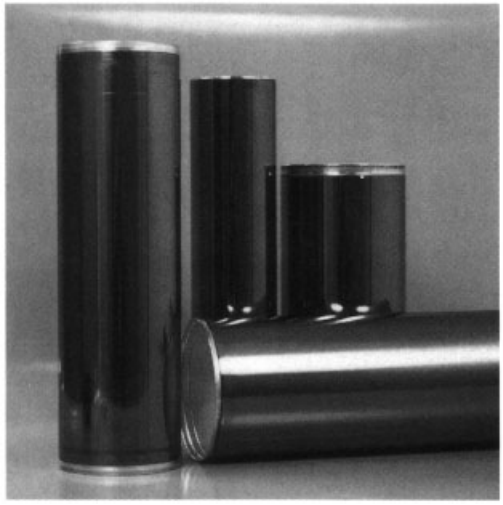

Figure 6. Selenium-coated photocopier drums after 60 000 copies

Glass Manufacture. Selenium and selenium compounds are used in silicate glass both in decolorization and as a colorant for producing intense yellow to red glass for special applications (selenium pink glass).

The main application in glass manufacture is in the decolorization of glass, which is always colored green to yellow green by traces of iron(II) and iron(III). The addition of 30 – 150 g/t selenium compensates for this color by adding the pink of the selenium, with no apparent reduction in transparency [26]. Formerly, selenium was added mainly as the element. However, elemental selenium volatilizes below the melting temperature of glass, so the yields were usually $< 50\,\%$. Therefore, selenium compounds such as sodium selenite, sodium selenate, barium selenite, or zinc selenite are now increasingly used. Even so, some volatilization losses are unavoidable. If an excess of selenium is used, pink glasses are obtained that used to be important for artistic purposes.

The addition of cadmium sulfoselenide to glass gives intense colors such as those required for signal technology (e.g., traffic lights). The

colors vary from yellow to red, depending on the CdSe content and melting conditions. Because of the volatility of selenium compounds, reproducible color control is difficult.

Pigments (→ Pigments Inorganic, Chap. 3.2.). Selenium is used as a pigment in the form of cadmium sulfoselenide. Replacement of part of the sulfur in cadmium sulfide by selenium changes the yellow color to an intense red. The pigment has good brightness and high stability. It is an important colorant for plastics, because the addition of up to 1.5 % cadmium sulfoselenide does not significantly affect their physical properties; it is also stable at the temperatures used for injection molding and extrusion. However, the importance of these pigments has declined considerably because of the toxicity of cadmium compounds. In some areas, the use of cadmium sulfoselenide is prohibited.

Metallurgy. The addition of selenium to steel improves its machinability, resulting in higher cutting speeds, lower energy consumption, and a longer lifetime for the machine tool. Although this is often achieved by adding sulfur, selenium is always used where good hot and cold formability, corrosion resistance, and surface finish are also required. For stainless steel, the addition rate is ca. 0.1 – 0.3 % Se. The machinability of copper can also be improved by addition of selenium, but tellurium is used more often. Addition of up to 0.1 % selenium to magnetic silicon steels improves their magnetic properties for use in transformer cores. Selenium is also used in lead accumulators to improve the casting and mechanical properties of lead – antimony alloys with low antimony content.

Ferroselenium, copper selenide, and nickel selenide are used in metallurgy.

Agriculture. Although the toxicity of soil with high selenium content has long been known, only much later was the need for selenium as an essential trace element in low concentrations discovered. A number of diseases formerly thought to be due to vitamin E deficiency are in fact the result of a deficiency of selenium. For example, the addition of 0.1 – 0.3 ppm selenium to the feed of chickens and pigs considerably increases their growth rate and prevents certain diseases, although the mechanism of this is not fully understood. Selenium may act as an antioxidant or free-radical trap.

The high rate of heart disease in Finland is thought to be due to a selenium deficiency, the soil being low in selenium, and the law now requires that 6 – 16 g/t selenium be added to fertilizers used there.

Selenium is added to fertilizers or animal feed in the form of sodium selenite or sodium selenate, which are absorbed more readily by plants.

Other Uses. Selenium and selenium compounds (e.g., Selenac, dialkylselenium carbamates) are added with sulfur in rubber manufacture. This improves the thermal stability of the product and controls the vulcanization rate.

Selenium compounds are used to a small extent as additives to oxidation catalysts for hydrocarbons. Selenium dioxide is also used as a selective oxidizing agent in organic chemistry.

Selenium acts as an absorbent for mercury, forming stable mercury selenide, in gas purification [27].

Addition of up to 5 % selenium has been proposed in the preparation of dental amalgams [28].

7. Quality Specifications and Analysis

Standard-grade selenium containing 99.5 % Se is adequate for chemical and metallurgical use, although many producers supply a 99.9 % Se material. A special pigment grade containing 99.8 % Se is also marketed. The standard grades are supplied as powders of various bulk densities, depending on precipitation conditions. The standard grade is used mainly as starting material for the production of high-purity selenium for photocopiers and electronic applications, which require a product containing > 99.999 % Se. Here, the heavy-metal content must also be as low as possible, and the oxygen content should be much less than 5 ppm. Electronic-grade selenium is supplied in the form of granules, which can be produced by pouring molten selenium into deionized water. In addition to chemical analysis, the quality assessment of high-purity selenium includes mea-

surement of electrical and especially the photoelectrical properties [29]. The crystallization temperature, determined by thermal analysis, can give a useful indication of the impurities in selenium (Fig. 2), although interpretation of the results is not always unambiguous.

The methods of analysis for selenium depend on the matrix material. Direct spectroscopic determination in raw materials and intermediate products is usually impossible due to lack of standards. Hence, selenium-containing materials are first treated with nitric or sulfuric acid, for example, to dissolve the selenium as selenous acid. Most methods of determination involve reduction of selenous acid with various reducing agents, usually in solution in strong hydrochloric acid. Selenium can be determined gravimetrically after reduction with iron(II), for example, if the hydrochloric acid concentration is such that precipitation is complete and no loss of $SeCl_4$ vapor occurs [30]. The precipitated selenium is weighed, or redissolved and determined by atomic absorption spectroscopy, inductively coupled plasma, or direct coupled plasma spectroscopy [31]. In one of the oldest methods still in use, selenous acid is titrated with thiosulfate according to the equation:

$$Na_2SeO_3 + 4\,Na_2S_2O_3 + 6\,HCl \longrightarrow Na_2SeS_4O_6 + Na_2S_4O_6 + 6\,NaCl + 3\,H_2O$$

Sodium monoselenopentathionate is formed, and the excess of thiosulfate is back-titrated iodometrically [32].

8. Selenium Compounds

Selenium forms a large number of chemical compounds, some of which are of major industrial importance.

Selenium Dioxide, Selenous Acid, Selenites. Volatile selenium dioxide (sublimes at 315 °C) is formed by the combustion of selenium-containing materials. Selenium dioxide [*7446-08-4*] dissolves readily in water (800 g/L at 85 °C) to form selenous acid [*7783-07-5*]. Unlike sulfur dioxide, selenium dioxide and selenous acid are easily reduced to selenium by reducing agents such as sulfur dioxide. Even traces of organic material reduce part of the selenium dioxide to selenium, resulting in a slightly reddish color of the SeO_2. The sensitivity to reduction is increased by the influence of light. Selenium dioxide is used to oxidize alkyl groups to carbonyl groups in organic chemistry, as a lubricant additive, and as an additive in the electrolysis of manganese. Selenous acid is more stable than sulfurous acid and can be isolated in the pure state in the form of colorless prisms. Selenous acid decomposes above 70 °C, forming selenium dioxide and water. Sodium selenite [*10102-18-8*], zinc selenite [*13597-46-1*], and barium selenite [*13718-59-7*] are important additives in glass manufacture. These salts are prepared by the reaction of selenous acid with the corresponding hydroxides or oxides, followed by crystallization.

Selenium Trioxide, Selenic Acid, Selenates. Selenous acid is oxidized to selenic acid [*7783-08-6*] by strong oxidizing agents such as chlorine, chloric acid, ozone, or anodic oxidation. Selenic acid is a hygroscopic solid that melts at 59.9 °C. It is a much stronger oxidizing agent than sulfuric acid; for example, a mixture of concentrated selenic acid with concentrated hydrochloric acid evolves chlorine. Selenic acid can be dehydrated with phosphorus pentoxide or by applying vacuum (200 Pa) at 160 °C to give the strongly oxidizing selenium trioxide [*13768-86-0*] (*mp* 118 °C). Above 185 °C, selenium trioxide decomposes into selenium dioxide and oxygen. Selenates are formed by the reaction of selenic acid with metal oxides or hydroxides, and resemble the isomorphous sulfates in both structure and solubility. Sodium selenate [*13410-01-0*] has become an important fertilizer additive.

Hydrogen Selenide, Selenides. Hydrogen selenide [*7783-07-5*] can be obtained either directly from the elements at ca. 400 °C or by decomposition of the selenides of iron, magnesium, or aluminum with hydrochloric acid. It is a colorless gas (*bp* −41.3 °C) smelling like rotten radishes, is even more toxic than hydrogen sulfide, and has a very irritating effect on the mucous membranes, causing bronchial and nasal irritation. Hydrogen selenide is less stable than hydrogen sulfide and can be oxidized to selenium by air, especially in the presence of light. Heavy-metal selenides such as copper selenide

form the basis of selenium extraction from ores. Cadmium selenide [*1306-24-7*], produced from cadmium and selenium, is used in the pigment industry together with cadmium sulfide.

High-purity zinc selenide [*1315-09-9*] is an important infrared optical material. Single crystals can be produced by reacting zinc vapor with hydrogen selenide (chemical vapor deposition, CVD). Copper indium selenide could be the basis for solar cells of the future; light efficiencies of 14 % at an energy flux density of $> 100\,W/m^2$ can be achieved. However, the manufacturing process (CVD) is expensive [23, 33].

Selenium disulfide [*7488-56-4*] is formed by reaction of hydrogen sulfide with selenous acid or by direct combination of the elements. The red powder (*mp* ca. 100 °C) is used in concentrations of < 1 % in shampoos and hair preparations for the treatment of dandruff.

Selenium Halides. The action of halogens or halides on elemental selenium yields various selenium halides, depending on the temperature and stoichiometric ratio. Selenium monochloride [*14457-70-6*], Se_2Cl_2 (ϱ 2.774 g/cm^3), and selenium monobromide, Se_2Br_2 (ϱ 3.604 g/cm^3), are intensely colored, heavy liquids. $SeCl_4$ [*10026-03-6*] and $SeBr_4$ [*7789-65-3*] are crystalline substances that readily form complexes with alkali-metal halides. Selenium exhibits the oxidation state +VI only in SeF_6 [*7783-79-1*], the corresponding chloride and bromide being unstable. Of the oxyhalides, seleninyl chloride [*7791-23-3*], $SeOCl_2$, is liquid at room temperature and is remarkable for its extreme reactivity, combining with almost all elements.

Organic Selenium Compounds. Like sulfur, selenium forms a large number of organic compounds and can be a component of aliphatic, aromatic, and heterocyclic compounds. From the simple selenols (RSeH), which correspond to the alcohols, to the amino acids, enzymes, and proteins, a large number of selenium-containing organic compounds are known, several of which are of biological importance. The starting compounds for the introduction of selenium into organic compounds are often dialkyl selenides (R_2Se) or dialkyl diselenides (R_2Se_2), which are formed from the corresponding alkyl halide and sodium selenide:

$$2\,RBr + Na_2Se \longrightarrow R_2Se + 2\,NaBr$$
$$2\,RBr + Na_2Se_2 \longrightarrow R_2Se_2 + 2\,NaBr$$

9. Economic Aspects

The total consumption of selenium in 1989–1991 was ca. 2100–2300 t/a (Table 1) [22, 34]. Average consumption in previous years was ca. 1700 t/a, although the published figures are unreliable due to the lack of information from Eastern European countries. Of the total, about one-third is produced as high-purity selenium for electronic applications. Since selenium is a byproduct of copper production, its price is subject to very wide variations due to over- or underproduction from time to time. The price of selenium has decreased considerably (see Table 1). However, most dealing takes place not on the open market, but directly between the producer and consumer, so the true prices differ from the published prices. The price for standard-grade selenium in 1990 was ca. \$ 10/kg, and for high-purity selenium ca. \$ 30–50/kg [22]. Manufacturers of elemental selenium in the Western world include the large copper producers Noranda Mines (Canada); Norddeutsche Affinerie AG (Germany), with its subsidiary Retorte producing selenium compounds and high-purity selenium; Metallurgie Hoboken-Overpelt (Belgium); Mitsubishi Metals, Nippon Mining, Sumitomo (Japan); Outokumpu (Finland); ASARCO, Kennecott, and Phelps Dodge (United States) [35]. Selenium compounds are also produced by Degussa (Germany).

Although the consumption of selenium compounds in the glass industry and as a fertilizer additive is increasing, the use of cadmium selenide in pigments is decreasing, and selenium is being replaced by organic photoconductors in some types of xerography. Therefore the total consumption of selenium is not expected to increase in the near future. However, the supply of selenium could decrease in the future if copper is produced partly by hydrometallurgical processes, which do not yield a selenium byproduct.

10. Biological Activity and Toxicology

Although selenium is an important trace element in low concentrations, it is toxic at higher concentrations. It is an essential trace element in the selenoenzyme glutathione peroxidase. This enzyme, together with vitamin E, is important for the protection of cells from oxidative attack. The activity of glutathione peroxidase in red blood corpuscles correlates well with the selenium concentration up to 100 µg/L [37, 38]. Low selenium levels in the blood cause heart disease. For instance, in China, Keshan disease, a cardiomyopathy, caused the deaths of a large number of people, especially children, until the daily consumption of selenium was found to be very low. The transition from essential trace quantities to toxic quantities of selenium in terms of intake in food is summarized below [38]:

Low-selenium area with Keshan disease	11 µg/d
Selenium adequate area	116 µg/d
National Academy of Sciences, Food and Nutrition Board "safe and adequate" range	50 – 200 µg/d
Lowest observed effect level (LOEL)	3200 µg/d
High-selenium area with chronic selenosis	4990 µg/d

The concentration of selenium in the blood should be 60 – 100 µg/L. Although small amounts of selenium may reduce the risk of cancer, as indicated by some investigative studies, this has not been proved conclusively.

A chronic high selenium intake of > 5 mg/d gives rise to nonspecific symptoms such as hair loss, changes to the fingernails, diarrhea, effects on the central nervous system, and nonspecific changes to the liver. Also, a loss of appetite and a tendency to walk in a circle can occur. High concentrations of selenium in the workplace can cause headaches and "selenium catarrh." Also, the breath can have a garliclike smell due to the formation of methyl selenide, although this effect is much less pronounced than with tellurium.

A few selenium compounds exhibit acute toxicity. Irritation of the mucous membranes of the digestive system, liver damage, and toxic edema of the lungs have been observed. The most toxic compound is hydrogen selenide, for which the MAK is 0.05 ppm, the MAK value for other selenium compounds being 0.1 ppm. In the United States, the TLV is 0.2 mg/m^3. In a study carried out in Canada on people working in selenium production, if the TLV was exceeded slightly (up to 0.8 mg/m^3) for limited periods, no symptoms resulted [39]. In Germany, the limit in drinking water is 0.008 mg/L [40], and in the United States, the EPA recommends a maximum contamination level (MCL) of 0.01 mg/L.

Table 1. Estimated production (t), consumption (t), and price ($/kg) of selenium [22, 35]

	1989	1990	1991	1995*
Production				
Belgium	275	260	260	250
Canada	270	340	240	300
Chile	40	40	40	45
Finland	10	25	25	31
Germany	110	110	100	120
Japan	534	550	631	595
South Korea	20	25	25	
Mexico	25	25	25	
Peru	20	20	20	
Philippines	120	120	120	40
Sweden	60	30	30	50
United States	253	287	240	360
Former Yugoslavia	50	60	40	
Zambia	50	60	40	
Others	100	100	100	
Total	1937	2052	1918	1791
Consumption				
United States	560	530	500	
Japan	283	289	263	
Europe	830	800	770	
China	200	300	350	
Others	300	300	300	
Total	2173	2119	2283	
Average price	13.81	11.26	10.56	

*[36]

11. References

General References

1. R. A. Zingaro, W. C. Cooper: *Selenium,* Van Nostrand Reinhold, New York 1975.
2. K. W. Bagnall: *The Chemistry of Selenium, Tellurium and Polonium,* Elsevier, Amsterdam 1966.
3. *Gmelin,* System no. 10, Teil A 1953, Teil B 1949, Ergänzungsbände 1979, 1980, 1981, 1984.
4. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* VCH Verlagsgesellschaft Weinheim 1988.
5. S. C. Carparella (ed), *Proc. Int. Symp. Ind. Uses Selenium Tellurium* **3rd** (1984) .
6. S. C. Caparella (ed.), *Proc. Int. Symp. Uses Selenium Tellurium* **4th** (1989) .

7. A. J. Bard (ed.): *Encyclopedia of Electrochemistry of the Elements,* Dekker, New York 1975.

Specific References

8. M. Nachtrab, Retorte Company, personal communication.
9. W. C. Cooper, *J. Met.* **42** (1990) Aug., 45.
10. J. E. Hoffmann, *J. Met.* **41** (1989) July, 33.
11. O. Hyvarinen, L. Lindroos, E. Yllö, *J. Met.* **41** (1989), July, 42.
12. Noranda Minerals, US 4 581 064, 1986 (B. H. Morrison, J. G. Lenz, J. Pageau, J. G. Bard).
13. H. Bußmann, R. F. Dobner, *Erzmetall* **43** (1990) 362.
14. Austria Metall, EP 0 263 910, 1986 (J. Bertha, J. Wallner, H. Wörz).
15. J. E. Hoffmann, *J. Met.* **42** (1990) Aug., 50.
16. L. A. Teixeira, R. E. Reber, L. Y. Tavares: "The Mineral, Metals & Materials Society," *EPD Congress 90,* Warrendale 1990, p. 209.
17. N. Sonoda, N. Hosoda, K. Hori, US 4 663 141, 1985.
18. F. Eckardt, H. Berg, *Naturwissenschaften* **45** (1958) 335 – 336.
19. Nippon Mining, JP 31 08, 1956 (N. Imai, M. Endo); *Chem. Abstr.* (1957) 10 858.
20. Bayer, DE-OS 1 170 912, 1962 (H. Giesekus, H. Schöning).
21. P. J. Marcantonio, US 4 915 928, 1990.
22. Gardener, *Met. Miner. Rev.* (1991) 91.
23. M. Caffaray, *Trans. Inst. Min. Metall. Sect. C* **97** (1988) C 87.
24. Siemens, DE-OS 3 236 137, 1984 (K. Kempter).
25. U. Schiebel, T. Buchkremer, G. Frings, P. Quadflieg, *J. Non Cryst. Solids* **115** (1989) 216.
26. B. Simmingsköld in [5] , p. 274.
27. K. Itagaki, A. Yazawa, *Erzmetall* **35** (1982) 358.
28. A. Sato, US 4 528 034, 1982.
29. M. Benson, H. J. Davis in [6] , p. 241.
30. A. M. Scherfhaufer, *Fresenius Z. Anal. Chem.* **164** (1958) 327.
31. M. Verlinden, H. Delestra, E. Adriaenssens, *Talanta* **28** (1981) 637.
32. J. F. Norris, H. Fay, *Am. Chem. J.* **20** (1898) 278.
33. K. Mitchell, C. Eberspacher, K. Pauls in [6] , p. 572.
34. Metall Bulletin (1990) 15th March, p. 15.
35. D. Edelstein in [6] , p. 7.
36. *Kirk-Othmer,* **vol. 21,** Wiley, New York, 1997, p. 704.
37. M. Piscator in [6] , p. 88.
38. R. A. Goyer in [6] , p. 97.
39. J. P. Robin in [6] , p. 103.
40. Bundesgesetzblatt, "Trinkwasserverordnung," Part I, no. 16, 1975, p. 453.

Semiconductor Lasers → Electroluminescent Materials and Devices

Semiconductors

Silicon is a separate keyword

Masamichi Mizutori, Dia Research Institute, Inc., Tokyo 105, Japan

Rokuro Yamada, Dia Research Institute, Inc., Tokyo 105, Japan

1. Definition of Semiconductors . 3
2. History 3
3. Theory 3
3.1. Structure of Semiconductor Crystals 3
3.2. Conduction Mechanism of Semiconductor Materials 4
3.2.1. Band Model 4
3.2.2. Doping 5
3.2.3. The $p-n$ Junction 5
3.3. Lifetime and Recombination of Charged Particles 7
3.4. Effect of Various Factors on Semiconductors 7
3.4.1. Effect of Radiation 7
3.4.2. Behavior in a Magnetic Field . . . 7
3.4.3. Behavior under High Electric Field 7
3.4.4. Semiconductor – Metal Contacts . 7
4. Materials, Properties, and Manufacture 9
4.1. Outline 9
4.1.1. Requirements for Commercial Use 9
4.1.2. Properties and Uses 9
4.1.2.1. Group IV Elements 9
4.1.2.1.1. Germanium 9
4.1.2.1.2. Diamond 10
4.1.2.2. Others 10
4.1.2.2.1. Group VI Elements 10
4.1.2.2.2. Multicomponent Chalcogenides . 10
4.1.2.2.3. Oxide Semiconductors 10
4.1.2.2.4. Organic Semiconductors 11
4.2. Compound Semiconductors . . 11
4.2.1. Properties 11
4.2.2. Single Crystals 12
4.2.2.1. Boat-Grown (BG) Method 12
4.2.2.2. Liquid-Encapsulated Czochralski (LEC) Method 13
4.2.2.3. Comparison of the BG and LEC Methods 14
4.2.3. Wafer Processing 15
4.2.3.1. Slicing 15
4.2.3.2. Lapping and Polishing 15
4.2.3.3. Surface Defects of the Polished Wafer 16
5. Epitaxy 16
5.1. Liquid-Phase Epitaxy (LPE) . . 17
5.1.1. Principles and Apparatus 17
5.1.2. LPE of AlGaAs 18
5.2. Vapor-Phase Epitaxy (VPE) . . 18
5.3. Organometallic Vapor-Phase Epitaxy (OMVPE) 19
5.3.1. Principles and Apparatus 20
5.3.2. Operating Conditions 21
5.4. Molecular Beam Epitaxy (MBE) 22
5.4.1. Principles and Apparatus 22
5.4.2. Gas-Source MBE 22
6. Characterization 23
6.1. Specific Resistivity 23
6.2. Lifetime of Charge Carriers . . 24
6.3. Crystal Defects 24
6.3.1. X-Ray Method 24
6.3.2. Electron-Beam Diffraction 24
6.3.3. Electron Microscopy 25
6.3.4. Etching 25
6.4. Contamination of the Wafer Surface 25
7. Semiconductor Devices 25
7.1. Brief Description of Various Devices 26
7.1.1. Diodes 26
7.1.2. Transistors 27
7.1.3. Integrated Circuits 28
7.1.3.1. Memory 28
7.1.3.2. Logic Circuits 28
7.1.3.3. Analog Integrated Circuits 29
7.1.4. Hybrid Integrated Circuits 29
7.1.5. Application-Specific Integrated Circuits (ASICs) 29
7.2. Wafer Fabrication 30
7.2.1. Formation of Insulation Layer . . 31
7.2.2. Lithography 31
7.2.3. Doping 32
7.2.3.1. Thermal Diffusion 32
7.2.3.2. Ion Implantation 32
7.2.4. Etching 32
7.2.5. Washing 33
7.2.6. Wiring Technology 33

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_537

8. Supporting Industries 33
8.1. Building 33
8.2. Clean Room 33
8.3. Ultraclean Water 35
8.4. Gases 35
8.5. Other Chemicals 36
8.5.1. Quartz 36
8.5.2. Pyrolytic Boron Nitride 36
8.5.3. Graphite 37
8.5.4. Others 37
9. Environmental Aspects 37
9.1. Gases 38
9.2. Liquid Chemicals 38
9.3. Metals 38
9.4. Chlorofluorocarbons (CFCs) . . 38
10. Economics 39
11. References 40

1. Definition of Semiconductors

Semiconductors are crystalline materials whose electric resistivity is between that of insulators and conductors (i.e., in the approximate range 10^9 to $10^{-3}\,\Omega\cdot\mathrm{cm}$). In general, insulators include glass, rubber, plastics, etc., and conductors are metals.

Besides electric resistivity, semiconductors have three major characteristics:

1) The resistivity decreases with increasing temperature, which is substantially different from metals.
2) Not only is the current – voltage relation inconsistent with Ohm's law, but the curve is also nonsymmetrical with respect to the origin. This results in the rectifying action of semiconductors.
3) The conductivity of a semiconductor is extremely sensitive to impurities even at the part-per-million (ppm) level or lower. This specific feature is the most important factor for the commercial use of semiconductors in the electronics industry.

2. History

In the mid-1800s, M. Faraday, F. Braun, and others discovered crystals that have the specific features mentioned in Chapter 1. The rectifying action of selenium was discovered in 1876 by Adams. Although classical physics could not interpret semiconductor phenomena theoretically, the commercial application of selenium and copper(I) oxide in rectifiers and crystal detectors began in Europe in the 1920s without knowledge of the theory. The crystal detector played a very important role in the 1930s, because its rectifying action helped to amplify the weak current of radio broadcasting.

The development of quantum mechanical theory in the 1920s helped to interpret the behavior of semiconductors. In 1931, A. H. Wilson applied quantum theory to semiconductors and established a basic theoretical model (Wilson model). This theory was developed further by A. Schottky and others.

However, the theory was not confirmed experimentally at this time because purification technologies for the experimental materials required were inadequate and the materials did not exhibit their inherent properties. During World War II, the development of radar became an urgent problem, and countries such as the United States, United Kingdom, Germany, and Japan attempted to develop low-noise crystal detectors, for which ultrahigh-purity materials were essential. They selected silicon and germanium, and succeeded in obtaining very pure germanium crystals before the end of the war. This was followed by the discovery of the transistor effect by J. Bardeen and W. H. Brattain in 1948 and the principle of the $p-n$ junction by W. Shockley in 1949 at Bell Laboratories in the United States. Due to its higher melting point and chemical reactivity, silicon is more difficult to process than germanium, and development was therefore delayed. However, with the progress of metal oxide technology for making the insulating layer, silicon became the main semiconductor material. Since the 1950s, silicon semiconductor devices such as diodes, transistors, and integrated circuits (ICs) have been developed, and the density of integration has increased astronomically. The group III – V compounds developed by H. Welker in 1952 have become commercially established primarily in optoelectronics (→ Electroluminescent Materials and Devices).

3. Theory

3.1. Structure of Semiconductor Crystals

A typical semiconductor material is silicon. Silicon is an element of group IV of the periodic table, and in the crystal the four electrons in the outermost orbital of an atom form covalent bonds with adjacent atoms, resulting in a stable crystal. The silicon crystal consists of a diamond-type lattice in which four adjacent atoms form the cubic-system tetrahedron shown in Figure 1. Perfect silicon crystal is an insulator because no free mobile electrons exist. Conductivity is imparted to the silicon crystal by crystal defects, residual impurities, or free electrons generated by doping with impurities. However, the conductivity is much lower than that of metals.

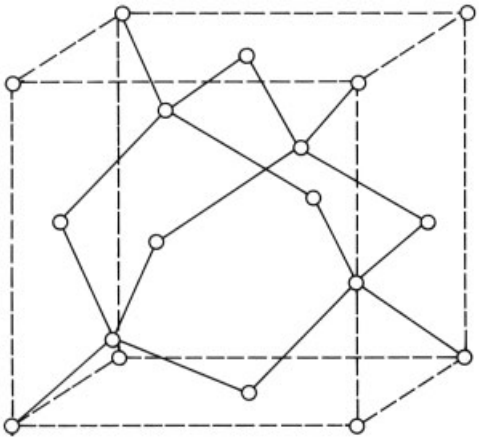

Figure 1. Tetrahedral structure of diamond

Metal atoms have free mobile electrons within the crystal. Because a metal crystal tends to form a close-packed structure, electrons of the outermost orbital of each atom are isolated and are present as free electrons, which are responsible for conductivity. Consequently, a metal atom does not have specific electrons, and the atom exists as the cation.

3.2. Conduction Mechanism of Semiconductor Materials

3.2.1. Band Model

The conduction mechanism of semiconductor materials can be interpreted by band model theory, which is a statistical – mechanical extension of the atomic model proposed by Bohr in 1913. Bohr proposed that electrons surrounding an atomic nucleus can exist only in specific orbitals. When this theory is extended to a crystal – a congregation of atoms – the specific orbitals in which electrons exist can be considered as a band that has a certain width.

When silicon atoms are arranged in an ordered crystal, the band in which electrons exist is called the valence band (see Fig. 2). The arrangement depends on the crystal and the lattice constant, which is related to the energy level of the valence band. If energy such as heat or light, which is higher than the covalent bond energy of the crystal, is supplied, the valence electrons are activated and jump to a higher band. This higher band level, in which the electrons are freely mobile and can participate in the conduction of electricity, is known as the conduction band. The zone between the two bands, in which electron occupancy is forbidden by virtue of quantum theory, is called the forbidden band or band gap. The width of the band gap corresponds to the energy required to break a covalent bond and depends on the atoms and their arrangement. The band gaps of insulators are much wider than those of semiconductors, whereas metals can be defined as materials whose band gap is zero.

When an valence electron jumps to a higher level, a hole is formed in the valence band. This hole is freely mobile within the valence band, behaving as a positively charged particle and participating in the electrical conductivity of a semiconductor together with the free electron in the conduction band.

Based on the above theory, the conductivity of semiconductors can be described by

$$a = nq\mu_{\mathrm{e}} + pq\mu_{\mathrm{p}} \tag{1}$$

where

a = conductivity
n = concentration of electrons
p = concentration of positive holes

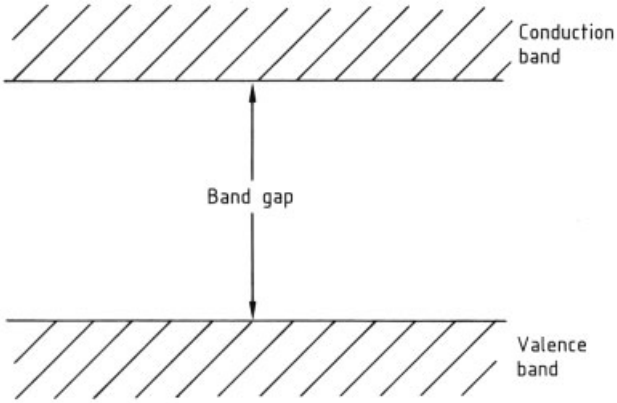

Figure 2. Band model

q = charge of an electron
μ = mobility of an electron (μ_e) or positive hole (μ_p) in a unit electric field

At thermal equilibrium, $n = p$.

The mobility depends on the bonding state of the crystal lattice and the material itself, and is less at higher temperature because the electrons tend to collide with atoms due to the increased atomic vibration resulting from the higher thermal energy of the crystal. In general, a positive hole is less mobile than an electron.

On the other hand, n and p increase exponentially with increasing temperature in accordance with the law of mass action.

$$np \sim e^{-E/kT} \tag{2}$$

where E is the band gap energy, k the Boltzmann constant, and T the absolute temperature.

Because the increase of n and p with temperature is much larger than the decrease in μ, the conductivity of semiconductors increases with increasing temperature. In the case of metals, since the electrons in the outermost orbitals always act as free electrons, n is independent of temperature and μ decreases with temperature. Hence the conductivity of metals decreases with increasing temperature.

Mobility is also affected by impurities in the crystal. Positively charged impurities attract and deflect electrons, and thus decrease the mobility. Therefore, historically, the measured value of mobility of various semiconductors increased from year to year until the late 1950s because of the improvement in crystal quality. Even now, the crystals of compound semiconductors still have fairly high contents of crystal defects and impurities. Therfore, the measured mobility is not necessarily the inherent mobility of each material.

3.2.2. Doping

As discussed above, the electric conductivity of semiconductors is the result of electrons jumping up to the conduction band and the vacant positive holes that remain in the valence band. In contrast to the vacuum tube, which utilizes thermal excitation, impurity implantation is used to generate free electrons and positive holes in semiconductors, and this process is known as doping.

If a group V element such as arsenic, phosphorus, or antimony is implanted in silicon (group IV) crystal, four out of five electrons in the outermost orbital of a group V element remain in the valence band by bonding to adjacent silicon atoms, but the remaining electron is freely mobile in the crystal and jumps up to the conduction band on thermal excitation. A semiconductor that contains many negatively charged electrons is known as an n-type semiconductor.

On the other hand, when a group III element such as boron, aluminum, or gallium is implanted in silicon, a positive hole is formed because a group III element has only three electrons in the outermost orbital. A semiconductor of this type is called a p-type semiconductor. Because of this mechanism, a group V dopant

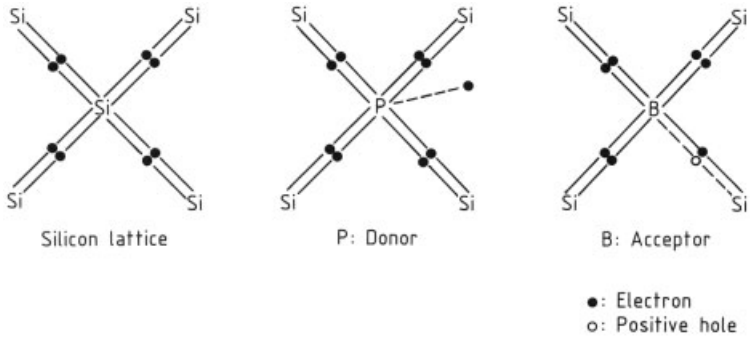

Figure 3. Donor and acceptor

is called a donor (of electrons) and a group III dopant is known as an acceptor (Fig. 3).

Impurities and defects in a crystal generate free electrons and positive holes. Therefore, to achieve very fine control, a defect-free single crystal of extremely high purity must be made, and the dopant implanted quantitatively.

3.2.3. The *p – n* Junction

The $p-n$ junction is the basic structure of semiconductor devices. As shown in Figure 4, the $p-n$ junction is an interface of a p-type and an n-type semiconductor crystal. Each domain contains many free electrons or positive holes. If the $p-n$ junction is forward biased, positive holes move to the n-domain and electrons to the p-domain, where they remain as minority carriers and contribute to conduction until they recombine with majority carriers and reach thermal equilibrium.

The concentration of minority carriers entering the other side, Δp and Δn, depends on the voltage as follows:

$$\Delta p, \Delta n \sim \mathrm{e}^{aV/kT} \tag{3}$$

a = conductivity
V = voltage
k = Boltzmann constant
T = absolute temperature

Therefore, Δp and Δn increase exponentially with voltage.

If the $p-n$ junction is reverse biased, positive holes move to the cathode and electrons to the anode, and a so-called depletion layer is formed in the vicinity of the junction. Therefore, very few minority carriers cross the junction, and the backward electric current is negligible compared to the forward current. Because of this feature, the $p-n$ junction exhibits a strong rectifying characteristic, as shown in Figure 5. When the reverse voltage reaches a very high level (e.g., several tens of volts), a high current is generated by the electron avalanche phenomenon.

3.3. Lifetime and Recombination of Charged Particles

Implanted minority carriers recombine with existing majority carriers and restore the original thermal equilibrium condition. This interval is called the lifetime of the charged particles (numerically, the lifetime is defined as the time constant of the exponential damping curve). The lifetime depends on the type of crystal and is generally $10^{-6}-10^{-3}$ s. However, recombination does not occur only between electrons and positive holes; various defects and impurities in the crystal often act as recombination centers. They decrease the apparent lifetime of a crystal. Therefore, as in the case of mobility and doping, the production of high-purity, defect-free single crystals is the basic requirement of the semiconductor industry.

3.4. Effect of Various Factors on Semiconductors

3.4.1. Effect of Radiation

If a semiconductor is irradiated by light whose photoquantum energy is higher than the band

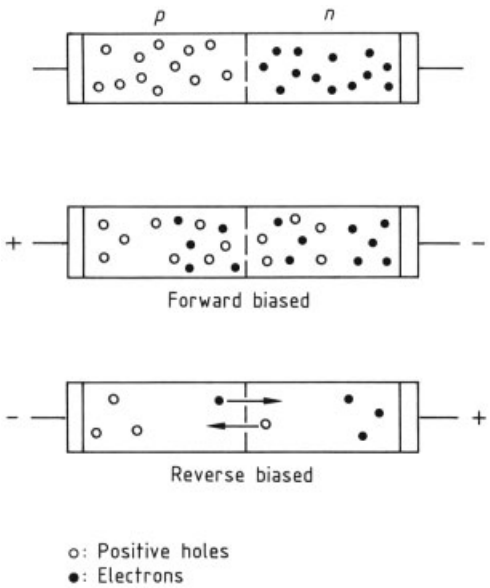

Figure 4. $p-n$ Junction

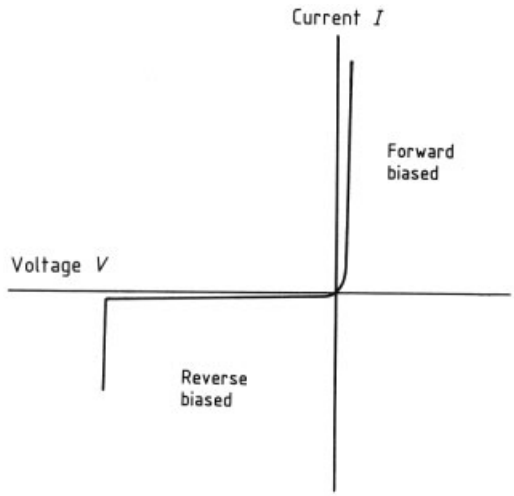

Figure 5. Rectifying action of a $p-n$ junction

gap energy of the semiconductor, an electron – positive hole pair is formed by excitation and the conductivity increases (photoconductivity). If the pair is formed near a $p-n$ junction, photoelectromotive force is generated. This is the principle of the photocell. Ionizing radiation (X-ray, β, or γ) also releases electrons in a crystal and forms electron – positive hole pairs. If a semiconductor is irradiated by higher-energy particles such as protons, α-particles, or neutrons, crystal defects are formed and performance deteriorates. Group III – V materials such as GaAs have better radiation hardness than silicon, which is an advantage in their use as solar cells for satellites that are affected by cosmic rays.

3.4.2. Behavior in a Magnetic Field

If a semiconductor crystal is placed in an electric field and a magnetic field is applied perpendicular to the direction of the electric current, an electric field is generated in a direction perpendicular to both the electric current and the magnetic field. This is called the Hall effect. The Hall voltage depends on the concentration of charged particles and the electron mobility. Therefore, the concentration of charged particles and the type of conduction (n- or p-type) can be determined by measuring the Hall voltage. A device utilizing the Hall effect is used for magnetically operated switches in rotary machines and made mainly from GaAs or InSb.

3.4.3. Behavior under High Electric Field

Under high electric fields, semiconductor materials often exhibit anomalous behavior. This includes the tunnel effect, Zener breakdown, and the Gunn effect. When the dopant concentrations in the n- and p-domain of a $p-n$ junction are both very high (e.g., $>10^{18}$ cm^{-3}, p^+-n^+ junction), the depletion layer formed by the very low reverse bias becomes extremely narrow and a high voltage arises due to the generation of negative resistance. Under these conditions, electrons in the p-domain can pass through the depletion layer to the n-domain as a wave and move as particles in the n-domain. This phenomenon – rapid conversion of particle to wave and then wave to particle – is called the tunnel effect, and the resulting abrupt increase of current is known as Zener breakdown. The tunnel diode is used for high-speed switching, and the Zener diode is used for voltage control.

Semiconductors whose band structure is of the direct-transition type (→ Electroluminescent Materials and Devices, Chap. 2.), such as GaAs and InAs, tend to generate high-resistance layers when the electric field increases; electrons then move from one conduction band to the other because the mobility of electrons differs greatly between the two conduction bands. For this reason,an oscillation phenomenon called the Gunn effect is observed in the microwave region, and is used for amplification and oscillation of microwaves and for shaping of pulse waves.

3.4.4. Semiconductor – Metal Contacts

Semiconductor – metal contacts are a basic structure of semiconductor devices, and the properties of the contact play an important role. Among the many historical hypotheses regarding the contact, the Schottky barrier model explains its properties fairly reasonably. As shown in Figure 6, SCHOTTKY proposed that an electrical dual layer is formed at the contact, with negative charges distributed on the surface of the semiconductor and positive charges distributed within the semiconductor. Since a metal possesses many free electrons, the electric potential inside the metal is uniform and, therefore, an ordered arrangement of charges cannot be maintained. When electrons pass the dual layer, they must overcome the barrier at the contact, and the height of this barrier was regarded as the difference in work function generated by the contact:

$$\begin{aligned}&\text{Height of Schottky barrier}\\&= \text{Work function of semiconductor}\\&\quad - \text{Work function of metal}\end{aligned}$$

However, when the electrical properties of various metal – semiconductor contacts were measured, these properties were found not to depend on the metal but to be strongly affected by the treatment of the semiconductor surface.

To solve this problem, J. BARDEEN proposed, as shown in Figure 7, that at the surface of a semiconductor an inherent quantum mechanical layer exists that traps electrons. Consequently, positive charges remain inside the semiconductor as an electrical dual layer. This hypothesis was experimentally verified by BRATTAIN and led to the invention of the point-contact transistor.

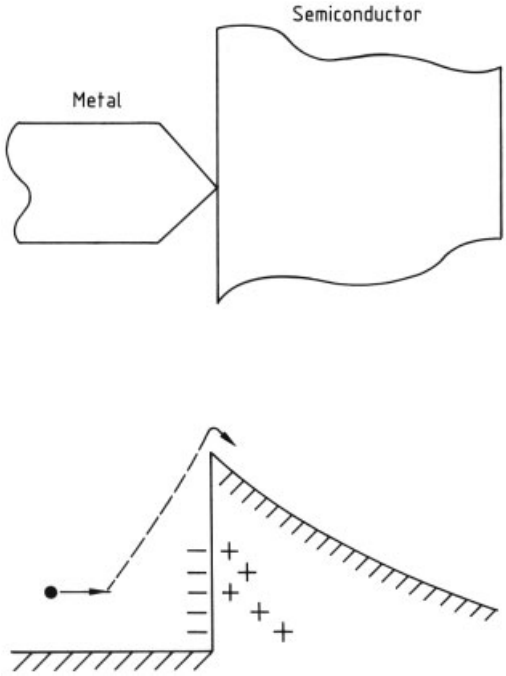

Figure 6. Schematic diagram of the electric dual layer of a metal – semiconductor contact

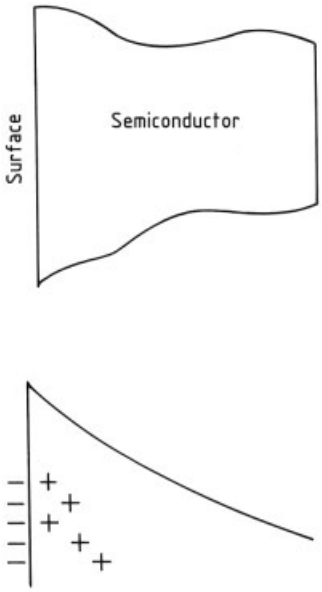

Figure 7. Surface states of a semiconductor (Bardeen's theory)

The behavior of a semiconductor surface is very complicated. For making devices, only the thin surface layer of a wafer is actually used. Therefore, polishing and treatment of the wafer surface are very important.

Although the Schottky barrier theory was disproved, diodes with metal – semiconductor contact structure are called Schottky barrier diodes and are important high-speed devices for use in the microwave and ultrahigh-frequency range.

4. Materials, Properties, and Manufacture

4.1. Outline

4.1.1. Requirements for Commercial Use

The following conditions are important for the commercial use of semiconductor materials as electronic devices:

1) The $p-n$ junctions can be made practically
2) Semiconductor properties should be retained up to at least 150 °C
3) The mobility of electrons and positive holes must be high
4) The diffusion length of the carrier must be high ($\geqq$ 0.1 mm)

Silicon is used overwhelmingly for most of the high-volume devices (→ Silicon), followed by group III – V compounds (GaAs, GaP, etc.), which are used mainly for optoelectronic devices such as light-emitting diodes (LEDs), laser diodes (LDs), and some special microelectronic devices. Germanium used to be a very important material and was the first to be used commercially. However, it was replaced by silicon, which offers much better performance and lower cost; germanium is now used only for limited purposes. Other materials such as group VI amorphous metals, multicomponent chalcogenides, and oxides have specific uses.

Academic attention has been paid to organic semiconductor materials because of their unique properties; tetracyanoquinodimethane (TCNQ) has been used commercially for organic condensers.

Important properties of major semiconductor materials are listed in Table 1.

4.1.2. Properties and Uses

4.1.2.1. Group IV Elements

4.1.2.1.1. Germanium

Germanium was used for crystal detectors before World War II, and in 1948 the first transistor was made from germanium by J. Bardeen, W. H. Brattain, and W. Shockley. Therefore, germanium was a key material for diodes and transistors. Nowadays, small amounts are used for specific diodes.

The band gap of germanium is 0.66 eV, which is lower than that of silicon (1.08 eV) or GaAs (1.4 eV). This means that the upper temperature limit to which germanium retains its semiconductor characteristics is ca. 100 °C, compared to 200 °C for silicon. Since the reverse breakdown voltage of the germanium $p-n$ junction is also lower than that of silicon, it cannot compete with the latter in high-voltage power uses.

However, a low energy gap results in a low starting voltage in the forward voltage – current property, which is advantageous for small signal diodes. The electron and positive hole mobility of germanium is three and four times higher than that of silicon and GaAs, respectively. This gives better performance in high-frequency devices. However, because it is compensated for by improvement of the device structure, germanium has been replaced by silicon even in this market.

The transparency of germanium to light extends up to a wavelength of 1.8 μm, compared with 1.1 μm for silicon. Therefore, germanium avalanche photodiodes are used as detectors in optical fiber communication. However, with the development of quartz optical fiber and InGaAs avalanche photodiodes, germanium is gradually being replaced by these products.

Germanium polycrystal is made by reduction of high-purity GeO_2, obtained from zinc smelting sludge, and purification by natural solidification and zone melting. For production of a single crystal, a horizontal zone leveling process is normally used. If especially high purity is required (e.g., for use in avalanche photodiodes and γ-ray detectors), the Czochralski pulling process is employed as in the case of silicon (→ Silicon).

Table 1. Physical properties of semiconductor materials

Material	Lattice constant, nm	Crystal∗ structure	*mp*, °C	Band gap energy, eV	Type of transition ∗∗	Mobility, $cm^2 V^{-1} s^{-1}$	
						Electron	Positive hole
Diamond		D		5.5		2000	2100
Si	0.5431	D	1420	1.11	I	1500	450
Ge	0.5646	D	936	0.66	I	3900	1600
GaAs	0.5654	Z	1238	1.43	D	8800	400
AlAs	0.5662	Z	≥ 1700	2.16	I	1000	≤ 100
GaP	0.5451	Z	1465	2.26	I	300	
InP	0.5869	Z	1062	1.35	D	4600	150
ZnS	$a = 0.381$, $c = 1.244$	W	1830	3.66	D	160	10
ZnSe	0.5668	Z	1520	2.67	D	600	40
CdTe	0.6481	Z	1090	1.50	D	1050	80
SiC	$a = 0.309$, $c = 1.512$	W	2800	2.996	I	400	50

∗ D = diamond structure; Z = sphalerite structure; W = wurtzite structure.
∗∗ D = direct transition; I = indirect transition.

4.1.2.1.2. Diamond

Diamond has a very high band gap energy (5.5 eV), and the mobility of both electrons and positive holes is also high compared to other semiconductor materials (see Table 1). In addition, since its other properties such as low dielectric constant and high thermal conductivity are advantageous for use in high-speed devices, the use of diamond in heat-resistant devices, short-wavelength LEDs, and high heat-generating power devices is desirable. However, the formation of a conductive $p-n$ junction and the formation of a single-crystal layer, especially on a silicon substrate, have not been achieved. The application of buckminsterfullerene as a semiconductor is being studied, but is still in the exploratory stage.

4.1.2.2. Others

4.1.2.2.1. Group VI Elements

Crystalline and amorphous selenium and tellurium are members of group VI. Their crystal structures belong to the trigonal system. Single crystals of selenium in the form of needles are normally made by sublimation as needle crystals, and tellurium single crystals are made by the Bridgman process or the rotational pulling process. Both selenium and tellurium are electrically *p*-type, and bismuth and antimony are used as dopants. A selenium layer made by vacuum deposition is used for making rectifiers (→ Selenium and Selenium Compounds), in which a heterojunction between an *n*-type CdSe and a *p*-type selenium layer is utilized as the rectification junction. Amorphous semiconductor selenium is made either by high-speed quenching of the melt or by vacuum deposition. Amorphous selenium has lower optical absorption and reflectivity, and is photoconductive. High-purity amorphous selenium made by vacuum deposition is used for xerography.

4.1.2.2.2. Multicomponent Chalcogenides

The main components of chalcogenides are group IV chalcogen atoms; examples include Ge – As – Te and Se – Ge. Thin layers are made by vacuum deposition or sputtering. Applications of these materials are:

1) Switches and memories utilizing a large change of electrical resistance on phase transition
2) Optical memories utilizing changes in reflection factor, transmittance, and refractive index or photodoping
3) Resist materials utilizing a combination of photodoping and photodarkening

4.1.2.2.3. Oxide Semiconductors

In general, oxide compounds exhibit semiconductor characteristics if their composition is made nonstoichiometric by reduction or oxidation. Examples of the former are ZnO, TiO_2,

CdS, and SnO_2 (*n*-type) and, of the latter, NiO, FeO, and Cr_2O_3 (*p*-type). Applications of oxide semiconductors are gas sensors (SnO_2), ballistors (ZnO), and ceramic condensers ($BaTiO_3$).

4.1.2.2.4. Organic Semiconductors

An organic semiconductor is an organic or high molar mass compound that exhibits conductivity and other semiconductor characteristics. The π-electrons or unpaired electrons contribute to conductivity. Polycyclic aromatic compounds, such as anthracene, and compounds containing conjugated double bonds, such as phthalocyanine and polyacetylene, are π-electron carrier types, whereas charge-transfer complexes and organic compounds that contain free radicals are unpaired-electron carrier types. Examples of charge-transfer complexes are complexes of perylene with iodine or sodium. In 1962, tetracyanoquinodimethane (TCNQ) was announced as a strong acceptor whose conductivity is competitive with inorganic materials (several kS/cm) [7]. TCNQ has good high-frequency performance and is used commercially as a raw material in condensers for notebook or laptop personal computers and handheld video cameras in Japan. The chemical structures of perylene and TCNQ are shown in Figure 8.

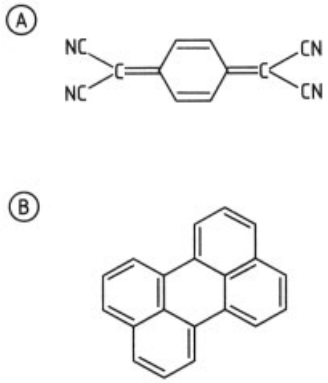

Figure 8. A) Tetracyanoquinodimethane (TCNQ); B) Perylene

Organic semiconductors are now being evaluated for use as photoconductive and photochromic materials, high-energy converters, heat – electricity converters, and in piezoelectricity. Because they exhibit different functions when minor changes are made in their chemical structure, various possible commercial application exist. However, their electron mobility is much lower than that of inorganic semiconductors, and most of them are not used commercially.

4.2. Compound Semiconductors

4.2.1. Properties

Compound semiconductors are intermetallic compounds between groups III and V, II and VI, or IV and IV elements. Examples are GaAs, GaP, InP, InSb, AlGaAs (III – V); ZnS, ZnSe, CdS, CdTe (II – VI); and SiC (IV – IV). Among these, GaAs, GaP, and InP have been developed because of the expanding market for opto- and microelectronic devices such as light-emitting diodes, laser diodes, and field-effect transistors (FETs). With the progress of epitaxial growth processes and device processing technology, mixed crystals such as AlGaAs and InGaAsP have been developed as specialty products.

Specific features of compound semiconductors are as follows:

1) The charge carrier mobility of group III – V compounds is generally several times higher than that of silicon. Hence, GaAs is being developed for devices with high speed, high frequency, low noise, and low electricity consumption.
2) The energy band structure of GaAs, InP (III – V), CdTe, ZnSe (II – VI), etc., is of the direct-transition type; therefore, their efficiency of light emission is higher than that of silicon or germanium, whose energy band structure is of the indirect-transition type. Thus, these compound semiconductors are used widely as optoelectronic devices for information processing and optical communication. GaP became a key material for visible LED with the development of nitrogen-doping technology [8].
3) The emission of electrons by GaAs when exposed to electromagnetic radiation (photoelectric effect) is utilized in photodiodes and solar cells.

4) The broad band gap width of GaAs helps to improve high-temperature performance and radiation hardness against cosmic rays.
5) By preparing mixed crystals such as AlGaAs and InGaAsP, the band gap width and lattice constant can be changed over a fairly broad range. Various mixed crystals are used as raw materials for high-performance devices with specific properties.
6) SiC (IV – IV) is used commercially as the raw material for blue LEDs (→Electroluminescent Materials and Devices, Chap. 4.2.).

4.2.2. Single Crystals

The single crystals of compound semiconductors are generally manufactured by either the boat-grown or the liquid-encapsulated Czochralski method. Here, the manufacturing methods of GaAs are described as representative of group III – V semiconductors.

4.2.2.1. Boat-Grown (BG) Method

There are two boat-grown (BG) methods: the gradient freeze (GF) and the horizontal Bridgman (HB), shown in Figures 9 and 10. Both methods use a quartz boat in which either gallium and arsenic or GaAs polycrystals are charged. Molten GaAs is cooled gradually from one end, and a single crystal grows from the seed crystal. In the GF process, the temperature of the entire system is lowered slowly while the temperature gradient is kept constant. In the HB process the boat moves through an electric furnace in which the temperatures T_1, T_2, and T_3 are kept constant.

Generally, the HB process uses GaAs polycrystal, which is made separately, whereas in the GF process, polycrystalline GaAs is synthesized by subliming arsenic on to gallium in the quartz boat. The growth direction of a single crystal is normally (111), but (511) is also used in some cases.

The vertical Bridgman (VB) method (Fig. 11) is used for the manufacture of GaAs and CdTe single crystals. In the process, a temperature gradient is formed in the vertical-type furnace, and the ampule containing polycrystal moves vertically to grow the single crystal. Neither the HB nor the BG process can make round wafers. However, the VB process can be used to make round wafers by means of a cylindrical ampule because the growth direction is (100). Although VB is still in the developmental stage, it is important as a method of making integrated circuit (IC) grade, large, round, low-defect wafers.

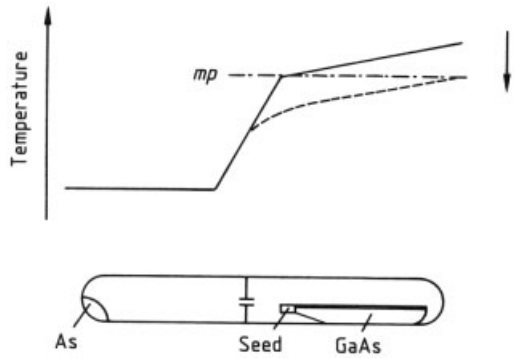

Figure 9. Gradient freeze method

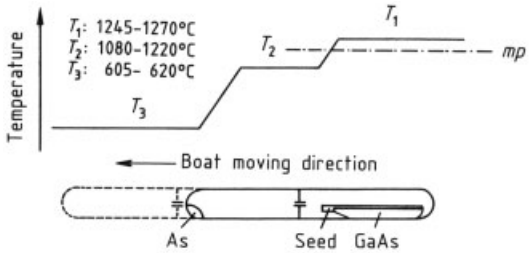

Figure 10. Horizontal Bridgman method

High purity and low defect concentration are two main technological targets of single-crystal manufacture. Silicon contamination is a problem in the BG method, because a quartz boat is generally used. Using 6 N or 7 N gallium and arsenic results in GaAs containing $(5-10)\times10^{16}$ cm^{-3} silicon. For use in optoelectronic devices, where Si-doped *n*-type or Zn-doped *p*-type GaAs is used, silicon contamination can be compensated by adjusting the amount of dopant. However, for use in microelectronic devices such as FETs and ICs, silicon must be compensated by chromium, which forms a deep acceptor level.

To make low-defect crystals, a mild temperature gradient, lower growth temperature, and precise stoichiometry control must be consid-

ered. Of the BG methods, GF is superior to HB as far as defect content is concerned. The typical etch pit density (EPD) of BG wafers is 3000 – 5000 (Si doped and Zn doped) or 5000 – 6000 (Cr doped), and they are generally accepted on the market. Ingots made by the BG method are generally 300 – 500 mm long, and the wafer diameter is 2 – 3 inches.

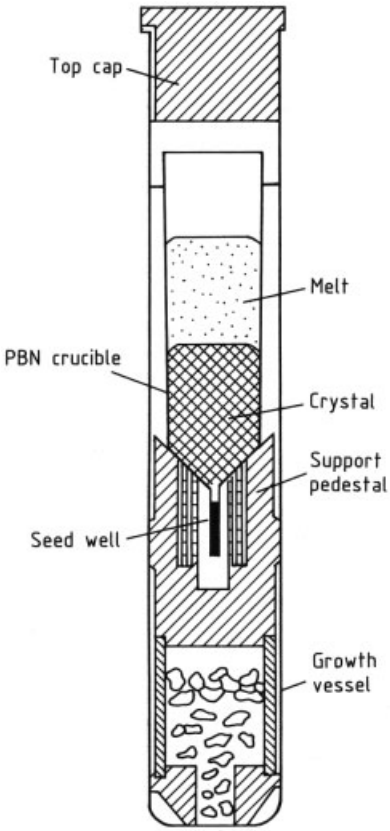

Figure 11. Vertical Bridgman method
PBN = pyrolytic boron nitride

4.2.2.2. Liquid-Encapsulated Czochralski (LEC) Method

In the liquid-encapsulated Czochralski (LEC) method, gallium and arsenic or GaAs polycrystal is charged to a crucible, which is then placed in a pressure vessel. A seed crystal is dipped into the molten GaAs and seed is slowly pulled up with rotation to form a cylindrical single crystal, as shown in Figure 12. The molten raw materials are covered by molten B_2O_3, and the vessel is filled with a high-pressure inert gas (N_2 or He) to prevent vaporization. The crystal is pulled up through the molten B_2O_3 layer. If GaAs polycrystal is used as the raw material, the pressure of the inert gas can be as low as 10 bar (low-pressure LEC), but if a mixture of gallium and arsenic is used, the pressure should be > 70 bar to prevent vaporization of arsenic during the synthesis of GaAs (direct-synthesis LEC). The apparatus for low-pressure LEC is less expensive than that for direct-synthesis LEC, but the GaAs polycrystal must be made separately. Therefore the total manufacturing cost is higher in low-pressure LEC, and chances of increased contamination are greater. For these reasons, the high-pressure, direct-synthesis process is used predominantly today.

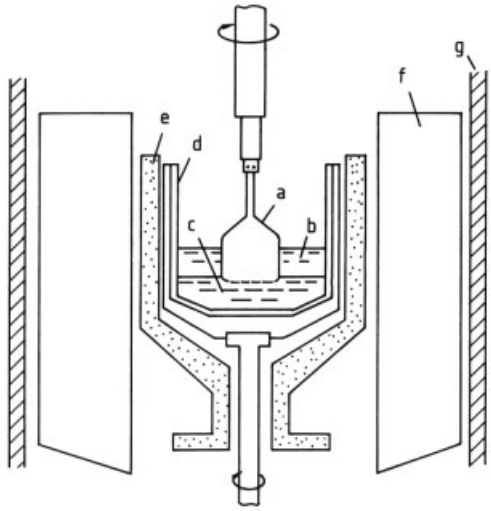

Figure 12. Liquid-encapsulated Czochralski method
a) GaAs crystal; b) Encapsulant, B_2O_3; c) GaAs melt; d) PBN crucible; e) Resistance heater; f) Heat shield; g) Chamber

As in the case of BG, various ideas have been tested to reduce the defect content of LEC wafers. In particular, EPDs of several hundreds per square centimeter can be attained by adding indium, though this has not been commercialized because of the high cost. Precise control of the stoichiometry, homogenization of the temperature distribution in the furnace, and thorough removal of impurities are the major countermeasures used to keep the EPD low.

At present, the standard level of EPD is (2–5)$\times 10^4$ cm^{-2}. Although this level is not fully satisfied, customers are more concerned about homogeneous EPD distribution in a wafer and throughout a crystal ingot.

Diameter control and diameter enlargement are also very important. However, apart from the case of silicon, automatic diameter control by optical methods is impossible in LEC because of the encapsulation by B_2O_3 and the slow heat response. This problem was solved by total computerization of the LEC process. Now most of the ingots are either3or4inches in diameter,and 5-inch wafers are in the developmental stage. Ingots of 3-inch diameter, 300 mm long (6 – 7 kg), are produced commercially. In addition, the successful development of 4-inch-diameter, 320-mm-long ingots (17 kg) has been reported [9].

For uses in microelectronics such as FETs or ICs, the single-crystal ingot must have semi-insulating properties, but since conventional processes use quartz crucibles, the ingot is always contaminated by silicon and becomes conductive. Therefore, the addition of small amounts of chromium (ca. 0.1 %) is required to compensate for silicon. However, because of the very low segregation coefficient of chromium (6×10^{-4}), its distribution in the crystal is inhomogeneous, and crystal properties vary within an ingot. The introduction of crucibles made from pyrolytic boron nitride (PBN) allowed semi-insulating crystals to be produced without adding chromium, which contributed greatly to the development of GaAs ICs. This method gives silicon concentrations of $< 5\times 10^{14}$ cm^{-3}, which are compensated by EL 2, an inherent defect in the crystal, which forms a deep-level donor. Semi-insulating undoped crystal can be produced without difficulty by using PBN crucibles.

To meet the requirement of GaAs IC devices, various modifications were tested to obtain large-diameter, round wafers with low defect content and high purity. Among these modifications, which included the use of an arsenic atmosphere, arsenic pressure control, and arsenic injection, magnet-application LEC (MLEC) is now used commercially. Application of a magnetic field reduces thermal convection and temperature variation in the furnace, and increases the viscosity of the melt, as in the case of silicon (→ Silicon). The surface of the crystal is smoother and has much less striation compared to crystals made by conventional LEC. A major reduction in the mean EPD level was not found, but the defects are distributed much more homogeneously within a wafer, which gives very uniform electrical properties. The MLEC apparatus is shown in Figure 13, and Table 2 lists the properties of semi-insulating wafers made by the MLEC process.

Table 2. Properties of semi-insulating GaAs polished wafers

Property	
Diameter, mm	100 ± 0.4
Thickness, µm	625 ± 10
Growth method	MLEC
Resistivity, $\Omega \cdot$ cm∗	$\geq 1\times 10^7$
Mobility, cm^2 V^{-1} s^{-1}	6000
Etch pit density, cm^{-2}	$\leq 7\times 10^4$

∗ after annealing at 850 °C for 20 min.

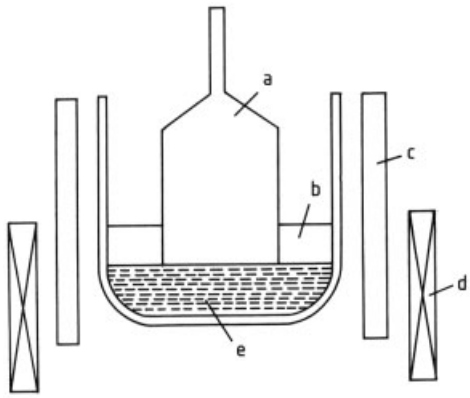

Figure 13. Magnetic LEC method
a) GaAs crystal; b) B_2O_3; c) Heater; d) Solenoid; e) GaAs melt

4.2.2.3. Comparison of the BG and LEC Methods

Both the BG and the LEC processes are used commercially, and each has its own characteristics. In general, BG wafers are used for optoelectronic devices such as LEDs and LDs, and LEC wafers are used for microwave and microelectronic devices, because the most important requirements are low defect concentration for the former, and large size and round shape for the latter.

Major characteristics of the BG method are as follows:

1) Automatic operation is easy, but unit production capacity is small and high man-power is required to prepare the equipment for startup
2) Dopant control is easy because no liquid encapsulation is applied
3) The control of temperature gradient is more moderate and precise than LEC (1 – 5 K/cm in BG, 30 K/cm in LEC); this gives low-EPD crystals
4) Round wafers cannot be made

The LEC method has the following features:

1) Large, round wafers can be made with minor cutting loss
2) Growth velocity is higher than in the BG method
3) By using a PBN crucible, the production of semi-insulating undoped wafers is possible

As mentioned above, each method has its own advantages and disadvantages, and both will coexist in future.

Single crystals of GaP have been used widely as the substrate for epitaxial production of visible-light LEDs. Because the vapor pressure of phosphorus is higher than that of arsenic, and the melting point of GaP is higher than that of GaAs, only the LEC method can be used, but the operating conditions are more severe than with GaAs. Wafers, 2 – 2.5-inches in diameter are made, and the EPD is ca. 2×10^5 cm^{-2}, which is higher than that of GaAs. A reduced EPD level is especially required of substrates for high-brightness green LEDs.

4.2.3. Wafer Processing

Single-crystal ingots are processed into wafers with the crystal orientation, shape, size, and surface properties required to meet specific uses. Basically, the methods of wafer processing are the same for BG wafers and LEC wafers except for the slicing direction due to the different crystal growth directions. Because device fabrication for high-speed ICs requires very fine, accurate pattern formation compared to discrete devices, the specifications for wafer processing, such as flatness and cleanliness, are much stricter. Here, the method for IC-grade wafers is described.

Figure 14 shows the process from an LEC ingot to mirror-surface wafers.

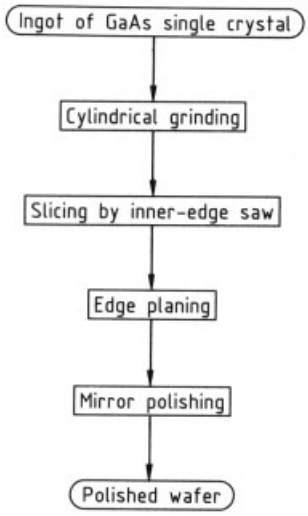

Figure 14. Block diagram of wafer processing

The ingot is ground to a fixed diameter, sliced by using an inner-edge saw, beveled to prevent chipping of the outer edge, and lapped to remove saw marks formed during slicing. The wafers are then etched by strong acid to remove the degraded layer and are processed to a mirror surface by polishing.

4.2.3.1. Slicing

Inner-edge saws are used exclusively for slicing, although wire saws have been tested. The flatness of the mirror wafer depends strongly on slicing operation parameters such as blade mounting, extent of blade bending, blade sharpness, and slicing speed. The orientation of the sliced wafer is inspected by X-ray diffraction or the light-figure method.

4.2.3.2. Lapping and Polishing

Lapping is the preliminary polishing process using fine grinding powder to remove saw marks and adjust wafer thickness. Group III – V materials tend to cleave easily, which causes scratches or cracks. Therefore, more careful attention must

be paid in selecting equipment and processing conditions than for silicon. Like final polishing, both single-side and double-side lapping are used, but the latter generally gives better accuracy and productivity.

Final polishing to give a mirror surface is conducted either by mechanical polishing with Al_2O_3 or SiO_2, by mechanochemical polishing with aqueous NaOCl, or by a combination of both. The mechanochemical method is used more widely and the composition of the polishing solution and the operating conditions are part of the important expertise of each manufacturer. In the case of single-side polishing in which wafers are mounted on a plate, the accuracy of the thickness depends strongly on the type of plate and the accuracy of wafer mounting, and washing with organic solvent is necessary to remove the mounting wax. The polished surface is easily washed, and its appearance is good. In the case of double-side polishing, in which no wax mounting is required, the accuracy of thickness is better and wafer washing is easy. During the collecting process, however, wafers tend to come in contact with each other, creating haze and scratches in the polished surface. Final polishing usually operates in a clean room of Class 1000 or better.

4.2.3.3. Surface Defects of the Polished Wafer

Surface defects include scratches, haze, dust, and pits, which are related to the inherent crystal defects. In addition, because a thin oxide layer is formed during wafer processing, polished wafers must usually be etched before device fabrication. So-called etching-free, fab-ready wafers, whose oxide layer is extremely thin, have been developed and commercialized in Japan.

In the case of silicon, wafer processing is fully automated, but for GaAs it is still in the partial automation stage because of the low production volume and high brittleness. To further improve the cleanliness of wafers, however, automation is very important.

Typical values of the surface properties of IC-grade GaAs polished wafers are listed in Table 3.

Table 3. Surface properties of GaAs wafers

Diameter, mm	100 ± 0.4	76.2 ± 0.4
Surface		
Polished side	damage-free finish	
Back	lapped and etched	
Native oxide layer	very thin	
Flatness		
Total thickness variation, µm	≤ 5	≤ 3
Nonlinear thickness variation, µm	≤ 5	≤ 3
Bow, µm	≤ 8	≤ 5

5. Epitaxy

Epitaxy is the growth of a crystal on another crystal along essentially the same crystal axis; this process is called epitaxial growth. The four epitaxial processes are

1) Liquid-phase epitaxy (LPE), in which the crystal grows on the substrate from a melt solution
2) Vapor-phase epitaxy (VPE), in which the elements or their volatile compounds are transported to the substrate, where they react to form a crystal
3) Organometallic VPE (OMVPE), which utilizes the thermal decomposition of organometallic compounds (OMCs)
4) Molecular beam epitaxy (MBE), in which VPE is conducted with a molecular beam generated under ultrahigh vacuum

Table 4. Comparison of expitaxial methods

Method	Advantages	Disadvantages
LPE	simple apparatus	inhomogeneous in thickness and composition
	high purity and high quality	rough surface
	constant stoichiometry	limitation in thickness
VPE	high productivity	limitation in thickness
	multilayer structure	toxicity of hydrides
	flat surface	
OMVPE	thin multilayer structure	purity of raw materials
	homogeneity and controllability of thickness	dependability of apparatus
	stoichiometry control	toxicity of OMCs and hydrides
		more expensive than LPE and VPE
MBE	best in thickness control	very expensive
	low-temperature operation	low productivity
	ultrathin multilayer structure	toxicity (in case of gas-source MBE)
	in situ observation	

The advantages and disadvantages of these four processes are summarized in Table 4.

The purpose of epitaxy is to make high-purity or mixed thin-layer crystals that cannot be made by bulk crystal growth. OMVPE and MBE are especially suitable for making ultra-thin multilayer structures that are used for high-performance devices.

Although epitaxy is used for silicon (→ Silicon, Chap. 4.2.5.), it is used much more widely for compound semiconductors.

5.1. Liquid-Phase Epitaxy (LPE)

LPE has long been used for compound semiconductors and magnetic garnets. NELSON first attemped to grow GaAs from gallium solution in 1963 [10]. Since then, many improvements have been made, and LPE is now used to make visible and infrared LEDs and LDs.

5.1.1. Principles and Apparatus

In LPE, a low-melting metal is used as the solvent and the supersaturated melt solution of the components is brought in contact with the substrate. The Ga – As phase diagram (Fig. 15) explains the principle of LPE [11]. The melt solution is saturated with arsenic at temperature T_1 and composition C_1. If this solution is cooled to temperture T_2, an amount of arsenic equivalent to $C_1 - C_2$ is supersaturated and the corresponding amount of GaAs is ready to deposit. GaAs LPE is carried out at ca. 800 °C, and the solution is supercooled by less than 10 °C ($T_1 - T_2 < 10$ °C), to avoid the formation of nuclei, and is brought in contact with the substrate.

The slide boat and dipping methods are both used commercially, the former more widely than the latter. The slide boat method is further divided into horizontal and vertical rotational types. Figure 16 shows the horizontal slide boat type. Raw materials and substrates are separately filled in a graphite boat, which is set in a quartz tube under a reducing atmosphere. The melt solution is then brought in contact with the substrate. In the case of GaAs, the GaAs solution is in the upper holder of the boat and the substrate is in the lower holder. The upper part of the boat moves, keeping contact with the lower holder, and when the melt solution moves over the substrate and comes in contact with it, supersaturated GaAs deposits and a crystal grows on the substrate. The movement of the boat is controlled automatically by a quartz rod. Vertical rotation LPE uses a cylindrical boat, and the melt solution and substrate are brought in contact by rotating either the upper or the lower holder.

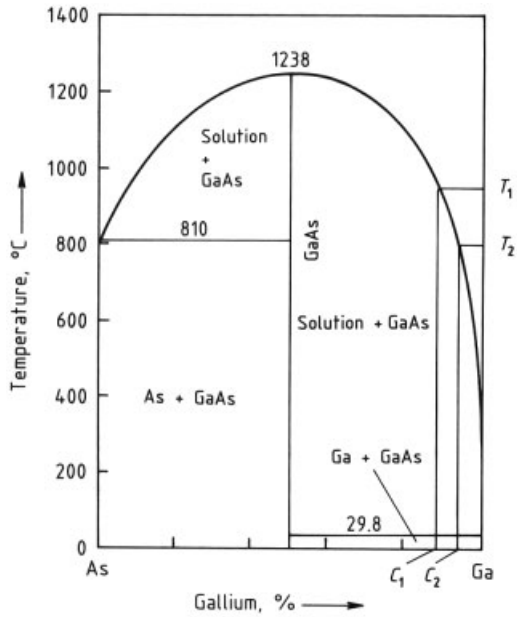

Figure 15. Ga – As phase diagram

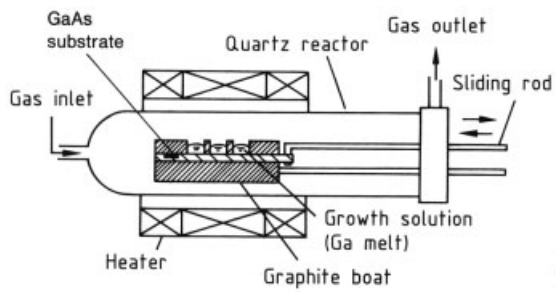

Figure 16. Slide boat LPE

The dipping process is shown in Figure 17. Melt solution is poured in the lower holder into which the substrate is dipped and an epitaxial layer grows. In this process, thermal deterioration of the substrate surface can be avoided by keeping the substrate at low temperature be-

fore starting growth. However, this process is not suitable for multilayer growth. It is used commercially for the LPE of GaP for green and red LEDs.

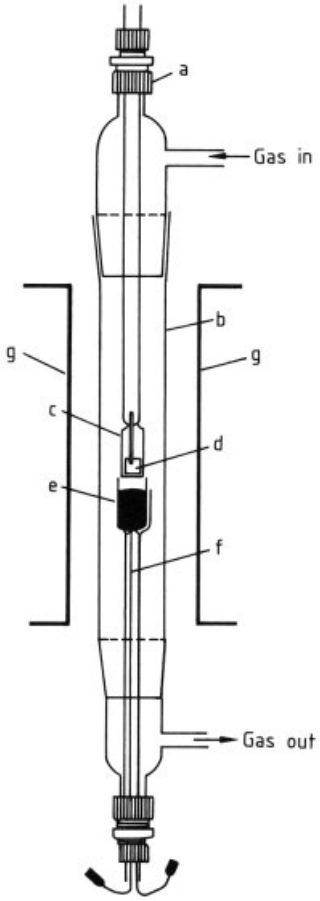

Figure 17. Dipping LPE
a) Polytetrafluoroethylene or O-ring connector; b) Silica furnace tube; c) Substrate holder; d) Substrate; e) Silica crucible containing solution in silica holder; f) Thermocouples to controller and monitor; g) Furnace

5.1.2. LPE of AlGaAs

Multilayer AlGaAs LPE on GaAs, with varying composition and dopant, is very important commercially for manufacturing high-brightness red LEDs. In LPE, melt solution and substrate are brought in contact at a fixed temperature, whereby supersaturated components deposit and crystal growth starts. Therefore, the composition of the melt solution that results in phase equilibrium at the fixed temperature must be found, so that ternary or quaternary compounds $Al_xGa_{1-x}As$, $Ga_xIn_{1-x}As_yP_{1-y}$, etc., can be grown with a required band gap energy, while the lattice match with the substrate is maintained.

Generally, simultaneous control of the lattice constant and the band gap energy is impossible, because only one parameter determines the composition. However, in the $Al_xGa_{1-x}As$ – GaAs system, since the lattice constants of GaAs and AlAs are very close (5.65325 versus 5.6605 at 27 °C), the band gap energy can be changed merely by changing the composition x and keeping the lattice matching.

LPE can make a very pure crystal layer, compared to other epitaxial processes, because high-purity gallium and GaAs are available, and only LPE can make a flat epitaxial layer over rough surfaces. However, it has inherent drawbacks such as uneven thickness caused by intermittent movement of the melt solution and a rough epitaxial surface due to residual melt on the surface. Therefore, LPE must coexist with other epitaxy process, depending on the device structure, productivity, and requirements for surface quality.

5.2. Vapor-Phase Epitaxy (VPE)

In a broad sense, OMVPE and MBE are also VPE processes. However, in this section only halogen-transportation VPE, which uses halides of group III elements as the transportation media, is discussed. The two halogen transportation VPE processes are (1) the chloride process, which uses $AsCl_3$, and (2) the hydride process which uses AsH_3 and PH_3.

The chloride process is simple, and the raw materials $AsCl_3$ and $GaCl_3$ are readily available in very high purity. It was used to make epitaxial wafers for FETs and ICs but has been largely replaced by the use of semi-insulating GaAs substrate made by the LEC method. In the hydride process, because commercially available HCl gas is not pure enough, making a very pure epitaxial layer is difficult. However, in the hydride process the amount of group III and V elements supplied can be controlled independently, and it is useful for growing mixed crystals. Commercially, the hydride process is used only to make $GaAs_xP_{1-x}$ mixed crystals for visible LEDs, for

which the content of arsenic and phosphorus must be controlled to change the wavelength of light emittance. The mechanism of crystal growth and the type of the reactor are similar for the two processes; here, the hydride process is described.

Principles and Apparatus. The apparatus for hydride VPE is shown in Figure 18. For the growth of $GaAs_xP_{1-x}$ epitaxial crystals, the vertical type is used almost exclusively because of its much better productivity.

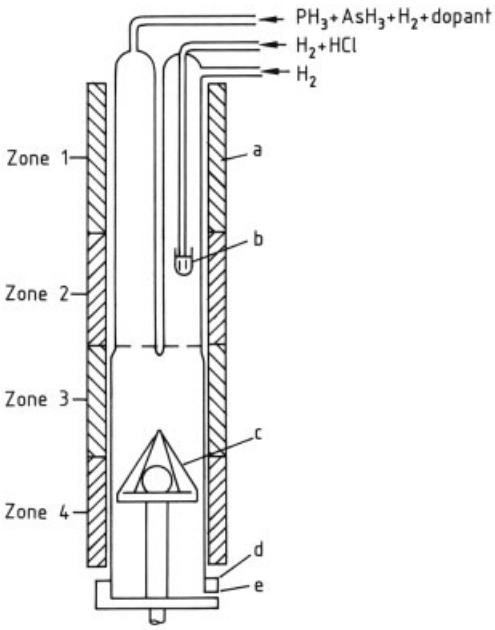

Figure 18. Multiwafer VPE of $GaAs_{1-x}P_x$
a) Furnace; b) Ga reservoir; c) Rotating wafer holder; d) Reactor seal; e) Vent

Heated dry HCl gas is introduced to the gallium reservoir, the GaCl formed disproportionates to Ga and $GaCl_3$, and gallium is supplied to the substrate surface.

$$Ga + HCl \longrightarrow GaCl + 1/2\,H_2$$
$$3\,GaCl \longrightarrow GaCl_3 + Ga$$

The group V elements arsenic and phosphorus are supplied to the substrate surface by the thermal decomposition of AsH_3 and PH_3:

$$n\,AsH_3 \longrightarrow As_n + 3/2\,n\,H_2$$
$$m\,PH_3 \longrightarrow P_m + 3/2\,m\,H_2$$

Mixed crystal $GaAs_xP_{1-x}$ grows on the substrate surface, and the composition x can be changed by controlling the amount of arsenic and phosphorus. The higher x is, the longer is the wavelength of light emittance, as shown in Table 5.

Table 5. Composition of $GaAs_xP_{1-x}$ mixed crystal and wavelength of light emittance

Composition of epitaxial layer, $GaAs_{1-x}P_x$	Substrate	Typical wave-length, nm	Color
$GaAs_{0.90}P_{0.10}$	GaAs	780	IR
$GaAs_{0.80}P_{0.20}$	GaAs	740	IR
$GaAs_{0.70}P_{0.30}$	GaAs	700	IR
$GaAs_{0.61}P_{0.39}$	GaAs	660	red
$GaAs_{0.45}P_{0.55}$	GaP	650	red
$GaAs_{0.35}P_{0.65}$	GaP	630	orange
$GaAs_{0.25}P_{0.75}$	GaP	610	soft orange
$GaAs_{0.15}P_{0.85}$	GaP	590	yellow

In hydride VPE, excess amounts of gallium, arsenic, and phosphorus are supplied to the substrate, the lower the reaction temperature, the higher is the growth velocity. However, a reaction temperature of 750 – 800 °C is generally used because surface quality decreases with decreasing temperature. By using vertical type reactors, more than thirty 2 – 2.5-inch wafers can be charged in a reactor in $GaAs_xP_{1-x}$ epitaxy. Either sulfur or tellurium is used as n-type dopant, and the $p-n$ junction is made by diffusing zinc over the epitaxial layer. The dopant source is H_2S or diethyltellurium.

The brightness of LEDs made from $GaAs_xP_{1-x}$ epitaxy is very sensitive to small amounts of impurities. Therefore, care must be taken with raw material handling and operating conditions, including the material of the gas cylinders and the cylinder filling operation for AsH_3 and PH_3.

5.3. Organometallic Vapor-Phase Epitaxy (OMVPE)

Organometallic vapor-phase epitaxy is also known as MOCVD (metal organic chemical vapor deposition) or MOVPE (metal organic VPE).

LPE and VPE are suitable for making LEDs and FETs that have a simple layer structure and require high purity. With the development of various electronic and optoelectronic devices, such as high-speed and high-frequency HEMTs (high

electron mobility transistors), and quantum-well LDs, which have very complicated device structures, OMVPE has been of special interest, followed by MBE.

OMVPE has the following characteristics:

1) Formation of ultrathin multilayer structure with abrupt interfaces of the layers
2) Formation of multilayers including mixed crystals of precisely controlled composition
3) High uniformity and high productivity compared with MBE

In 1959, Reuhrwein referred to OMVPE, and the first research report was written in 1968 by Manasevit who succeeded in making a GaAs expitaxial layer on sapphire, germanium, and GaAs [12, 13]. In the 1970s, GaAs epitaxy was actively studied; with improvements in the purity of starting materials, the crystal properties and purity of the epitaxial layer have been greatly improved, and Gunn diode, Hall element, and microwave FETs have been reported. In 1977 the successful development of double hetero LD, multiquantum-well LD, and superlattice LD, quickly attracted attention to OMVPE.

5.3.1. Principles and Apparatus

Organometallic compounds and metal hydrides are used as raw materials that thermally decompose and supply metals. In the case of GaAs, trimethyl- or triethylgallium (TMG or TEG) is used as the gallium source and AsH_3 as the arsenic source. In group III – V or II – VI compound semiconductors, OMCs of group III or II elements and hydrides of group V or VI elements are thermally decomposed, and the resulting elements react on the substrate and deposit group III – V or II – VI compounds. The organometallic compounds should have the following properites:

1) Their vapor pressure should be ca. 1 torr (130 Pa) at room temperature
2) They decompose on heating
3) High-purity materials are available commercially
4) Thermal decomposition products neither prevent crystal growth nor contaminate the epitaxial layer

As shown in Figure 19, various steps occur in parallel and in series, and each step must be analyzed in detail to improve the quality of the epitaxial layer. However, this has not yet been achieved, because the mechanism of the thermal decomposition of OMCs is not clear, and no way exists to observe crystal growth directly. Even in the case of the TMG – AsH_3 system, which has been most widely studied, various model mechanisms have been proposed, such as (1) gas-phase reaction between TMG and AsH_3, (2) reaction between gallium and arsenic deposited on the substrate, and (3) formation of GaAs from TMG and AsH_3 adsorbed on the substrate.

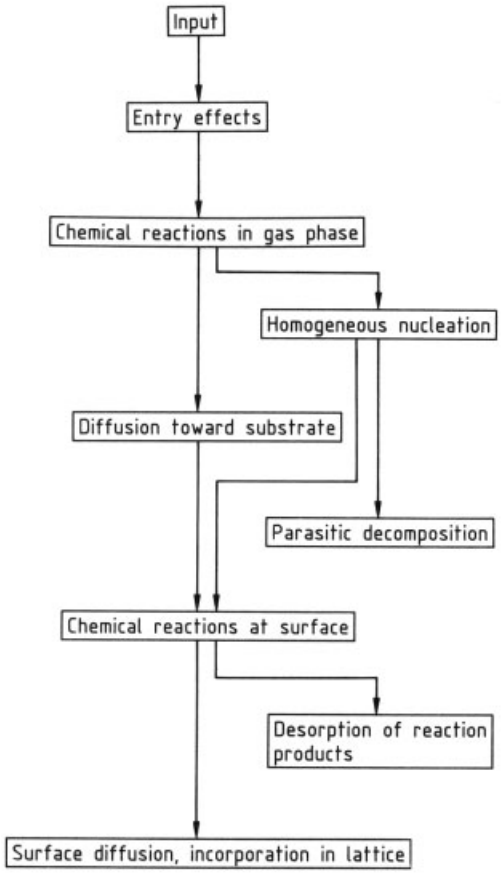

Figure 19. Parallel and series steps of OMVPE

The OMVPE apparatus for $Ga_{1-x}Al_xAs$ is shown in Figure 20. This is a vertical reactor that operates at normal pressure. Raw materials are TMG, trimethylaluminum (TMA), and

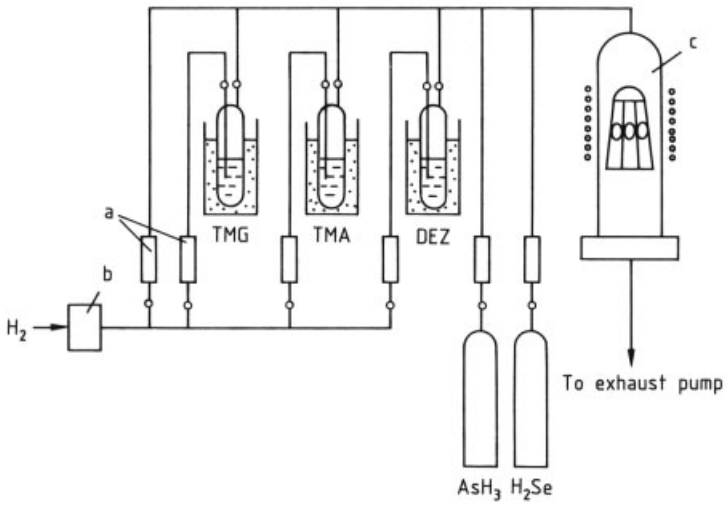

Figure 20. Organometallic vapor-phase epitaxy (OMVPE)
TMG = trimethylgallium; TMA = trimethylaluminum; DEZ = diethylzinc
a) Mass flow controller; b) Purifier; c) Reactor

AsH_3, and the dopants are diethylzinc (DEZ; *p*-type) and H_2Se (*n*-type). AsH_3 and H_2Se are diluted by high-purity H_2 to 5 – 10 % and to several hundred parts per million, respectively, and then put in high-pressure cylinders.

The apparatus consists of three parts: gas inlet, reactor, and waste-gas treatment. In the gas-inlet system, stainless bubblers containing TMG, TMA, and DEZ are set in very precisely controlled thermostats. High-purity H_2, used as carrier gas, is purified further to remove H_2O and O_2, and is introduced to the bubblers to form a saturated vapor that is transported to the reactor through stainless steel piping. The flow rate of the raw materials must be controlled precisely, and to prevent contamination the piping should be leaktight and easily bakable for cleaning. The reactor part is made from quartz tube, and substrates held on a susceptor are heated by high-frequency radiation or by resistance heating. The saturated vapor supplied to the reactor furnace decomposes on the substrate and grows as *n*- or *p*-type $Ga_{1-x}Al_xAs$. In this case, the gas flow rate should be kept high enough to prevent a gas-phase reaction that curtails epitaxial growth. Gas is then sent to the waste-gas treatment to remove or burn residual gas or reaction products.

5.3.2. Operating Conditions

The reaction temperature is usually 650 – 700 °C, and in this temperature range, growth rate is proportional to the concentration and amount of TMG supplied. In OMVPE, growth rate is determined by the amount of group III elements supplied, because the amount of group V element is much higher. Therefore, in contrast to LPE and VPE, in which growth rate varies with time, in OMVPE the thickness of the layer can be controlled precisely by varying the concentration of TMG and the reaction time; high-quality, uniform epitaxial crystals can be obtained.

The growth of various mixed crystals such as InGaAs, InGaP, GaAsP, and GaAsSb has been attempted by using OMVPE. In these cases, composition control is very important, and to achieve it the concentration of group III atoms in the crystal should be proportional to that in the gas supplied.

The purity of the crystal layer has been increased greatly by improving raw material purity and by optimizing reaction conditions. The purity is now comparable to LPE, and GaAlAs – GaAs double hetero LD and HEMT devices are made by OMVPE as well as by MBE. For the growth of phosphorus-containing compounds such as InP and InGaAsP, OMVPE is superior to MBE, because the high vapor pressure of phos-

phorus makes vapor pressure control and waste-gas exhaust difficult in MBE.

5.4. Molecular Beam Epitaxy (MBE)

In MBE, the elements of the epitaxial layer are heated and vaporized under ultrahigh vacuum (1.3×10^{-8} Pa), and the resulting molecular beam is introduced to the substrate to grow the crystal. This is a highly specialized vacuum deposition technology. MBE was first reported by GUENTHER and was developed by CHO [14, 15]. Since then, this technology has been actively developed for device applications, mainly for group III – V compound semiconductors. The quality of the crystal has been improved with progress in raw material purification and ultrahigh-vacuum technology, and has reached a level comparable to LPE or VPE. MBE is now used to make thin layers not only of group III – V but also of group II – VI and single-element semiconductors, metals, and magnetic materials. Especially for group III – V compounds, as the control of doping and thickness at the atomic layer level became possible, the formation of superlattice structures by an abrupt heterojunction, control of dopant concentration at the atomic-layer level, and gradient doping have been used to make various high-performance devices.

5.4.1. Principles and Apparatus

The elements of the epitaxial layer are held in separate molecular beam cells and heated under ultrahigh vacuum. Molecules leaving the molecular beam cell go straight to the substrate, because under ultrahigh vacuum the mean free path of the molecule is much longer than the distance between the cell and the substrate holder. Therefore, the molecular beam has essentially no chance of being deflected by collision with other molecules. The standard MBE apparatus for group III – V compounds is shown in Figure 21.

During MBE growth, because of the high temperature of the cell and substrate, gases vaporize from these hot parts of the apparatus, making it difficult to maintain ultrahigh vacuum. To avoid the problem, a large liquid-nitrogen shroud and heat cutoff are used. A load – lock mechanism separates growth and substrate exchange areas with a gate valve, so that the ultrahigh vacuum of the growth area can be maintained when the substrate is exchanged.

By using ultrahigh-vacuum analytical methods such as Auger or electron-beam diffraction, both the growth speed and the appearance of the growing crystal surface can be observed. This is a unique advantage of MBE, which makes it useful for studying both growth mechanisms and manufacturing apparatus.

Since the cell must withstand $>1000\,^{\circ}$C, pyrolytic boron nitride is often used, and high-purity, high-melting metals such as tantalum are used as heater materials.

The growth rate of group III – V compounds is determined by the molecular beam strength of the group III elements, which depends on cell temperature. Beam strength is controlled precisely by measuring the intensity of the molecular beam and feeding it back to the temperature control system of the cell. The beam intensity is measured by using a quadrupole mass spectrometer or ion gauge. Generally, the growth rate is around one molecular layer per second.

In GaAs MBE, the substrate temperature is $500-600\,^{\circ}$C, and with excess arsenic, gallium is first adsorbed on the substrate, and then an equivalent amount of As reacts with gallium to form GaAs stoichiometrically. Excess arsenic desorbs from the substrate and is removed by the pump. In GaAlAs MBE, the cell temperatures of gallium and aluminum should be controlled separately for composition control.

5.4.2. Gas-Source MBE

In conventional solid-source MBE, the growth chamber must be opened to the atmosphere for refilling, and several days are required to reevacuate to 1.3×10^{-8} Pa. Because this greatly lowers productivity, processes that use a gas source such as organometallic compounds or metal hydrides were developed. The gases form molecular beams at $1.3\times10^{-2}-1.3\times\ 10^{-6}$ Pa, and material refilling can be obtained just by gas cylinder exchange without exposing the growth chamber to the atmosphere. In addition, gas-source MBE has the following advantages:

1) In solid-source MBE, controlling the composition of mixed crystals that contain more

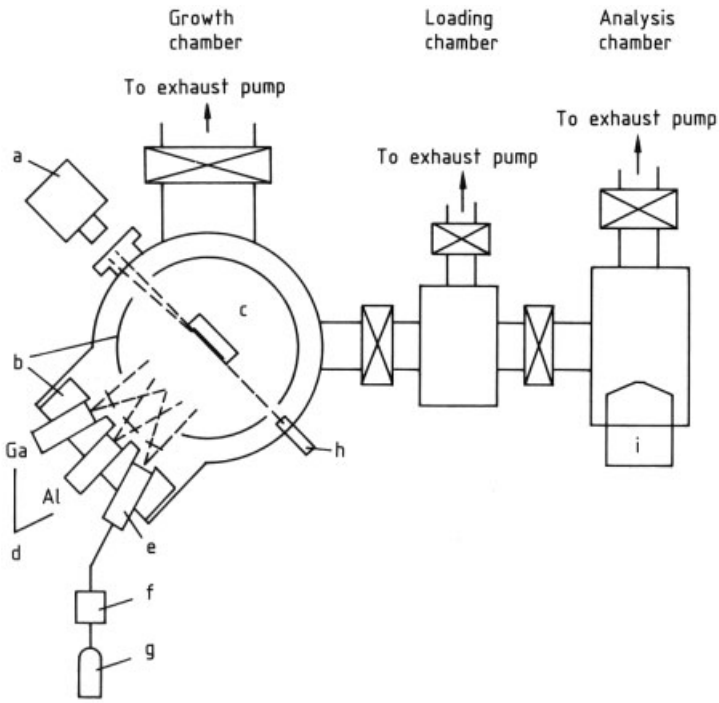

Figure 21. Molecular beam epitaxy (metal and gas source)
a) Reflection high-energy electron diffraction (RHEED) pattern analysis system; b) Liquid-N_2 shroud; c) Substrate holder; d) Effusion cells; e) Gas cell (cracking); f) Pressure controller; g) Gas cylinder; h) RHEED (E-gun); i) Surface analysis system (AES, XPS)

than two group V elements is extremely difficult. In gas-source MBE, control of pressure and flow rate is easier.

2) Solid gallium sources tend to form oval defects on the crystal surface. Gas-source MBE can overcome this problem.
3) Gas-source allows crystal growth on the insulating layer to be avoided.

MBE is used widely for ultrathin multilayer epitaxy, especially for the multilayer GaAlAs structures of HEMTs and LDs. Among various epitaxial processes, MBE can make the thinnest layers, but the apparatus is very expensive and the growth rate is slow.

6. Characterization

The main factors that determine the quality of semiconductor crystals are the specific resitivity, the lifetime of charge carriers, and various kinds of crystal defects. Although cleanliness of the polished wafer surface is not an inherent property of the crystal, it is important in device manufacturing.

6.1. Specific Resistivity

Specific resistivity varies with the difference in the amount of dopants, which is related to the solid – liquid interface of semiconductor crystals and varies in the axial and radial directions of growth.

Specific resistivity is a function of the mobility and concentration of electrons and positive holes in the crystal as follows:

$$\varrho = 1/q\,(\mu_n\, n + \mu_p\, p)$$

where

ϱ = resistivity
q = charge of an electron

μ_n, μ_p = mobilities of electron and hole
n, p = concentrations of electron and hole

Resistivity is measured by two- and four-point probe measurements and by spreading resistance measurements (contact methods). The eddy current and microwave techniques are noncontact methods.

Four-point probe measurement is most widely used. Four probes make contact with the wafer in a straight line; a direct current is applied to the two outer probes; and the potential between the two inner probes is measured. The measured value is the average specific resistivity of the volume between the two probes.

In spreading resistance measurements, the diameter of the probe is very small (2 – 7 μm), and this method can measure the resistivity of extremely small parts of a wafer.

6.2. Lifetime of Charge Carriers

When a semiconductor crystal is excited by light or by application of a voltage, the concentration of electrons and positive holes generated decreases exponentially to restore thermal equilibrium. The lifetime is defined as the time constant of the damping curve. As discussed in Chapter 3, the lifetime of charge carriers is directly related to the properties and efficiency of semiconductor devices; it is therefore a very important parameter for the design and manufacture of devices.

The lifetime is measured by noncontact techniques such as the reflection waveguide method, which utilizes excitation by infrared for the generation of charge carriers and microwaves for the detection of output. In principle, the change in electrical or optical properties caused by excess charge carriers is measured as a function of time, and the lifetime is calculated from the time constant.

6.3. Crystal Defects

Three types of crystal defect exist: point, line, and two dimensional.

Point defects are empty lattice positions or interstitial atoms, and line defects are dislocations. When a surface in a crystal slips, the boundary between that surface and the standing surface is called dislocation. Dislocations cause strain in the crystal lattice and act as electron acceptors, which prevent carrier movement. Dislocation-free silicon crystals can be manufactured commercially, but GaAs crystals still have fairly large amounts of dislocations. Typical two-dimensional defects are multilayered defects and twin-crystal boundaries, which are often formed during crystal growth. The former is an irregularity of crystal lattice arrangement, and the latter is the boundary between crystals that have different growth directions.

Dislocation-free crystals exhibit microscopic defects known as swirls, which are observed as eddies on etching or as shallow pits by microscopy. Swirls are considered to be either aggregates of empty lattice sites or complexes of interstitial silicon atoms with oxygen.

During crystal growth, temperature distribution and growth velocity vary with crystal rotation, and the amount of dopant also varies periodically. These changes are observed as continuous stripes on selective etching or by X-ray topography, and are known as striations.

To evaluate the crystal structure and defects, X-ray topography, electron-beam diffraction, electron microscopy, and etching are used.

6.3.1. X-Ray Method

X-ray topography is most widely used for the characterization of crystal structure and lattice defects. If X-rays are applied to a crystal, they are scattered by each atom and interfere with each other, and the strongly scattered ray is found only in specific directions. In imperfect crystals, diffraction occurs not only at the lattice sites but also at the lattice defects. The picture of lattice defects is obtained by X-ray topography.

6.3.2. Electron-Beam Diffraction

To observe the surface structure and lattice defects of crystals, low- (up to several hundred volts) and high-energy (several thousand volts) electron-beam diffraction is used. The former is known as low-energy electron diffraction (LEED) and the latter as reflection high-energy electron diffraction (RHEED).

Both can be used to analyze the surface structure of crystals qualitatively and quantitatively. Because RHEED utilizes a higher accelerating voltage, a much cleaner image can be obtained, and it allows in situ monitoring of crystal growth in MBE.

6.3.3. Electron Microscopy

To observe crystal defects and crystal microstructure, transmission electron microscopy (TEM) is used. In TEM, an electron beam of uniform current density is applied to a thin-layer probe, and the transmitted or diffracted electron beam is analyzed. The accelerated voltage of the electron beam is usually 100 – 400 kV or 1 – 3 MV in the case of ultrahigh-voltage TEM.

6.3.4. Etching

The observation of lattice defects by etching is the most popular method of crystal evaluation, because etching speed depends on growth direction, type of conduction, type of dopant, lattice defects, and surface structure of the crystal. Various etching solutions are used (Table 6). Etching processes involve an oxidation reaction, followed by decomposition and removal of the oxide.

6.4. Contamination of the Wafer Surface

About 70 % of the problems in the manufacturing of large-scale integrated circuits (LSIs) are a result of particles adhering to the wafer surface, which leads to defects in the shape and appearance of the pattern. Adhering particles cause pattern defects during dry etching and ion implantation because either the particles act as masking materials, or atoms of the particles diffuse into the substrate. Therefore, particle formation must be checked and reduced in each manufacturing stage by measuring and controlling the condition of the wafer surface.

To inspect the particles on the surface, laser-beam diffraction is widely used. Laser beams are applied to the wafer surface, and the resulting diffraction or scattering of the beam caused by particles is received by a collecting lens. Scanning electron microscopy (SEM) is used to observe the shape of particles.

To analyze elements on the surface, electron probe microanalysis and auger electron spectroscopy are used. The former measures specific X-rays generated when a very fine electron beam scans the probe, and the latter measures the electron energy generated by the auger state when an electron beam of 3 – 10 keV is applied.

These methods can be used to analyze inorganic materials. For organic materials, other methods such as infrared microspectroscopy, fluorescence analysis, and Raman spectroscopy are used.

7. Semiconductor Devices

As shown in Figure 22, two main types of semiconductor device can be found: discrete devices and integrated circuits. Discrete devices are unit components that have specific electronic functions such as rectification, amplification, oscillation, and photoelectric conversion. ICs are small structural units in which discrete devices, such as resistors or condensers, are mounted on a silicon semiconductor substrate and inseparably bonded. Although discrete devices have their own important applications for broad uses, with the development of IC manufacturing technology, major advances in miniaturization and cost reduction have been achieved.

ICs have many advantages over discrete devices, including small size, low weight, lower consumption of electricity, high speed, and high reliability. Because the IC manufacturing process is capital intensive and well suited to mass production, the cost per unit function becomes much cheaper with increasing degree of integration.

However, with the diversification of usage, multigrade small-volume production for specific uses became more important. To meet this requirement, application specific ICs (ASICs), such as gate arrays, were developed. ASICs consist of standardized unit circuits whose specific function can be varied by changing the bonding between them.

Table 6. Characteristics of etching solutions

Solution	Composition		Etching velocity, µm/min	Remarks
Dash	HF	$10\,cm^3$	0.1	slow
	HNO_3	$30\,cm^3$		suitable for microdefects
	CH_3COOH	$120\,cm^3$		
Sirtl	HF	$100\,cm^3$	1	depends on growth direction
	Cr_2O_3	50 g		
	H_2O	$100\,cm^3$		
Secco	HF	$100\,cm^3$	1.5	high speed
	$K_2Cr_2O_7$ (0.15 mol % solution)	$50\,cm^3$		egg-shaped etch pits
Wright	HF	$60\,cm^3$	1.0	applicable to all growth directions
	HNO_3	$30\,cm^3$		sharp etch pit shape
	CrO_3 (5 mol % solution)	$30\,cm^3$		
	$Cu(NO_3)_2$	2 g		
	H_2O	$60\,cm^3$		
	CH_3COOH	$60\,cm^3$		
Schimmel	HF	$100\,cm^3$	2	high speed
	CrO_3 (1 mol % solution)	$50\,cm^3$		egg-shaped etch pit

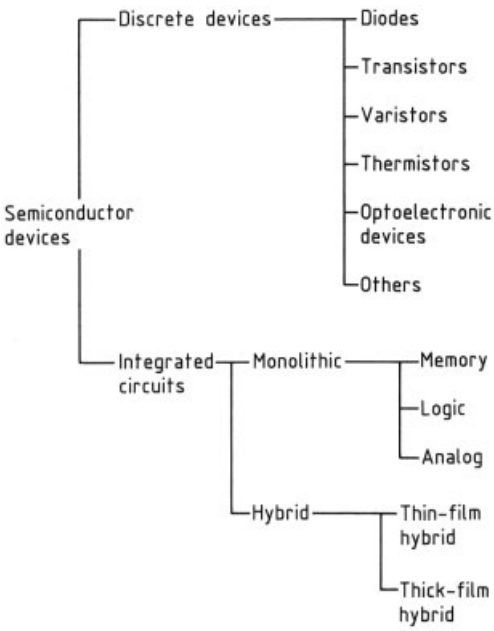

Figure 22. Classification of semiconductor devices

7.1. Brief Description of Various Devices

7.1.1. Diodes

Diode is the general term for devices that have two electrodes. A wide variety of diodes exist including the rectifier diode, voltage reference diode, varactor diode, Gunn diode, impatt diode, and tunnel diode. The three types of rectifier diode are the $p-n$ junction, the Schottky junction, and the pin diode. The pin diode has a high-resistance depletion layer between a $p-n$ junction. The pin diode is used as a high-voltage rectifier because a high reverse voltage can be obtained.

The voltage-reference diode keeps the voltage of both terminals constant by utilizing the breakdown phenomenon; that is, the electric current abruptly increases at a specific reverse voltage.

The varactor is a variable-capacitance diode used for frequency tuning. It utilizes the fact that the depletion-layer capacitance of a $p-n$ or Schottky junction varies with changing voltage.

Impact avalanche transit time (IMPATT) diodes, which are used for the oscillation and amplification of microwaves, utilize negative resistance. This diode has an n^+-n-p junction, and when biased at breakdown, this structure forms negative resistance due to the avalanche effect and transit time effect of carriers (Gunn diode, Schottky diode, tunnel diode, see Chap. 3). The IMPATT diode has a depletion layer adjacent to the junction which makes the transit time of the carriers between the electrodes very short, enabling them to pass through the depletion layer.

7.1.2. Transistors

Transistors consists of two $p-n$ junctions (i.e., three terminals). Various electric actions – amplification, oscillation, rectification, etc. – occur among these three terminals by the action of electrons or positive holes. The three terminals are the emitter (E), collector (C), and base (B). In the $n-p-n$ transistor, as shown in Figure 23, the emitter and collector are n-type and the base is p-type. In the base only a few charge carriers exist as minority carriers, and when forward biased, charge carriers are injected into the base from the emitter and then absorbed in the collector by reverse biasing. The electron flow from emitter to collector is controlled by the base potential with no power expenditure because the base is much thinner than the emitter and collector.

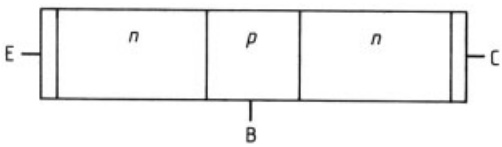

Figure 23. Junction transistor

The transistor is used as an amplifier: the current injected into the base at low voltage flows through a high reverse bias, resulting in a higher voltage (voltage or power amplification).

This mechanism is very similar to the vacuum tube. In a vacuum tube, electrons are emitted from the cathode when it is heated to $> 1000\,°C$ and these electrons reach a positively biased plate. This electron flow is controlled by a grid potential. Since the grid is negatively biased and is a grating, electrons can pass through it easily without loss of power.

Both the transistor and the vacuum tube have an electron supply source and collector, and electron flow is controlled by an electrode that absorbs practically no electrons. The major difference is the electron source. In the vacuum tube, because electrons are emitted from metals such as tungsten and must cross a potential barrier $>1\,eV$, the cathode must be heated to $>1000\,°C$. On the other hand, in the transistor, release of electrons or positive holes by donor or acceptor requires only 30 meV of activation energy, which can be achieved at room temperature. This is the reason the transistor has totally replaced the vacuum tube.

There are two types of junction transistor: bipolar transistors and field-effect transistors (FETs). Whereas bipolar transistors utilize both electrons and positive holes, FETs use either electrons or holes. The bipolar transistor was first developed by Brattain and Bardeen as a point-contact transistor in 1947; then Schockley developed the junction type based on his theory of the $p-n$ junction. The bipolar transistor is suitable for obtaining high current and high speed, but is structurally unsuited to high integration.

FETs are used for amplification, switching, and other applications. In the FET, the current from the source (emitter) to drain (collector) is controlled by the voltage impressed to a gate terminal (base) as shown in Figure 24. The gate is a metal electrode separated from the semiconductor surface by an insulating layer on the substrate. SiO_2 is most widely used as the insulating layer, and this type of FET is called a metal oxide semiconductor FET (MOSFET). MOSFETs are basic units of large-scale ICs, especially memories. GaAs FETs that use a semi-insulating GaAs substrate instead of an SiO_2 insulating layer show good performance in microwave amplification and very high-speed digital ICs. The HEMT is a heterojunction FET, which has a two-dimensional electron – gas interlayer that gives higher electron mobility and lower noise at microwave frequencies than conventional GaAs FETs.

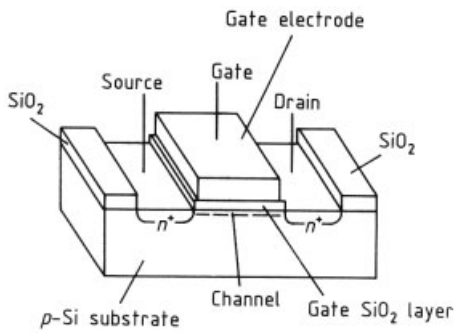

Figure 24. Structure of a MOSFET

7.1.3. Integrated Circuits

Integrated circuits are miniature structures in which many circuit elements are inseparably bonded on or within a substrate.

The original purpose of the development of ICs was simply miniaturization and weight reduction. However, during development, many other advantages were observed and pursued.

The first of these is cost reduction, achieved mainly by the introduction of photoetching, which utilizes photographic technology. When finely patterned negatives are applied to millions of transistors, their arrangement and wiring can be made by repetition of the treatment. The higher the degree of integration, the more effective is this method. Furthermore, because IC packaging is performed as a totality, rather than for each transistor, the number of bonding points and the consumption of packaging materials can be reduced; this also contributes greatly to cost reduction.

Second, the reliability of performance has increased considerably. Most failures originate at bonding points, but in ICs, the number of bonding points has decreased because many components are integrated in a silicon wafer. In addition, reduction in the number of manufacturing steps is also advantageous for performance reliability.

Third, performance itself has also improved. The smaller the transistor, the shorter is the wiring, and the higher the speed. Higher integration becomes possible.

The concept of ICs was proposed shortly after the invention of the transistor, and the invention of the metal oxide semiconductor (MOS) planar transistor by FAIRCHILD provided the basis for LSI, because its structure was suitable for high degrees of integration.

7.1.3.1. Memory

Since the introduction of the 1 kbit MOS memory by Intel in 1971, semiconductor memories have been used widely in many electronic appliances. The three types of MOS memories, pMOS, nMOS, and cMOS, use positive holes, electrons, and a combination of these as carriers, respectively. The pMOS was commercialized first because the separation of the individual elements is easy. Then, with increasing integration, nMOS replaced pMOS because of its higher speed. When the integration reached 1 Mbit, cMOS became the main technology because of its low electric consumption.

Memory is classified as random-access memory (RAM) and read-only memory (ROM). RAM can both write and read, whereas ROM can only be read. RAM is divided into DRAM (dynamic) and SRAM (static) memory. DRAM needs refreshing to maintain memory, but SRAM does not.

The memory action of DRAM is maintained by the electric charge stored in the capacitor, but the charge is lost by leakage current. Therefore, refreshing is needed. Because DRAM consists of only one transistor and one capacitor, it is the simplest semiconductor device. Memory capactiy has been increasing by a factor of four every three years. Capacity has now reached 4 Mbit and will reach 1 Gbit by the end of the 20th century. The basis of this progress is the development of fine processing technology. The progress in device fabrication is shown in Table 7. The wiring line width must be decreased by a factor of 0.6, when integration increase by a factor of four.

SRAM consists of six circuit elements to maintain memory stability, and the integration is about one generation behind DRAM.

DRAM is most widely used for all types of computer and control apparatus, whereas SRAM is used for the main memory of supercomputers and the cache memory of office computers and minicomputers, which especially require high speed.

The two types of read-only memory are mask ROM and programmable ROM. Mask ROM cannot eliminate the memory once written, whereas programmable ROM can rewrite. Both types of ROM are widely used for program storage of word processors, personal computers, workstations, televisions, etc.

7.1.3.2. Logic Circuits

Logic circuits are basic circuits for computer and digital control. Logic circuits judge an input signal or a combination of plural input signals and decide the level of output signals. In the early stages, bipolar logic circuits were standard. However, with the development of small

Table 7. Progress of device fabrication

Integration (DRAM)	256 kbit	1 Mbit	4 Mbit	16 Mbit	64 Mbit
Minimum line width, μm	1.5	1.0	0.8	0.5	0.3
Source of lithography	g-line	g-line	g- or i-line	i-line	excimer laser
CVD method for insulation layer	atmospheric or vacuum	vacuum	plasma	plasma	photo
Formation of metal layer	sputtering	sputtering	sputtering	metal CVD	metal CVD
Element area, mm^2	30 – 40	45 – 75	75 – 95	95 – 105	175 – 230
Tolerable diameter of particles, μm	0.2	0.1	0.05	≤ 0.05	≤ 0.05
Wafer diameter used, inch	4	5	6	8	8 – 10

portable appliances, MOS logic circuits are also being used. Bipolar logic circuits can treat signals about ten times faster than MOS logic. However, the power consumption of MOS logic circuits is lower than that of bipolar circuits.

Bipolar logic circuits consist of a transistor, a diode, and a resistor. They are used mainly in computer components, data transmission, and various industrial controls such as instrument control, numerical control, and process control, which require very rapid speed.

Because MOS logic circuits consists of only MOS FETs, a high degree of integration can be achieved fairly easily at low cost and small size. MOS logic circuits were first used for pocket calculators and then for microcomputers and video game machines, etc. They have made a major contribution to the miniaturization of these devices.

Applications for supercomputers, instrumentation, and military uses require ultrahigh speed and low power consumption. In these cases, GaAs logic ICs are gradually replacing bipolar logic circuits, especially in the United States.

7.1.3.3. Analog Integrated Circuits

Analog ICs, also known as linear ICs, process analog signals. A wide variety of analog ICs are used for both industrial and consumer purposes. Industrial uses include general-purpose devices such as amplifiers and voltage regulators, and specific uses such as motor regulators, temperature regulators, and smoke detectors. Specific consumer uses are in TVs, radios, tape recorders, cameras, air conditioners, and refrigerators. The integration of analog ICs is generally much lower than digital ICs, but has now reached 400 elements per square millimeter, and more than 10 000 elements are integrated in a single chip. Some of these process both analog and digital signals. For instance, analog – digital systems are used for digital audio tape (DAT) and digital TV, and their applications are broadening. In industrial uses, higher speed and higher frequency are being pursued for image processing.

7.1.4. Hybrid Integrated Circuits

In hybrid ICs, conductors, condensers, and resistors are made as films on the insulator, and semiconductor devices are mounted on it. The major features of hybrid ICs are

1) High precision and high reliability
2) Suitable for high-power, high-voltage, and high-frequency circuits
3) Small size but sufficiently high function
4) Easy to apply to custom uses
5) Low development expenses and short delivery time

To manufacture hybrid ICs, the circuit patterns of the resistor, conductor, and condensers are printed on the ceramic insulator with paste and then heat treated (thick-film hybrid ICs). If thin film is required, the circuit pattern is made by a combination of lithography with deposition or sputtering (thin-film hybrid ICs).

7.1.5. Application-Specific Integrated Circuits (ASICs)

Nowadays, as customer requirements diversify, manufacturers must respond with small-volume, multigrade production. In ASICs, standardized circuit blocks for signal treatment are combined with the specific circuit blocks of each customer by using computer-aided design (CAD). This system enables the requirements of each specific customer to be met in a short delivery time. With the increasing variety of equipment functions, the trend toward ASICs is becoming stronger.

Gate arrays and standard cells are the two most important ASIC products. The main part of the gate array is a standard logic circuit, but the wiring is different for each user. A standard cell is a block that consists of a logic gate memory or a central processing unit, etc., and each user can assemble custom LSIs by combining these cells.

Since gate arrays entered the market in 1977, their speed, integration, and functions have been greatly improved. Now many kinds of gate array are used in the industrial and consumer markets. Bipolar gate arrays are used mainly for computers, communications, instruments, and image processing which require very rapid information processing. The cMOS gate arrays consume less power than bipolar arrays, and high integration can be achieved easily. Thus cMOS gate arrays with 3000 – 5000 gates are used widely for both consumer and industrial applications, and they contribute to miniaturization of equipment, cost reduction, and improvement of reliability. The development of very fine processing technology has allowed an array with $>10^5$ gates to be produced. Bi- cMOS gate arrays, which are a combination of bipolar and cMOS gate arrays, have been introduced for special uses.

Because standard cells require trial production for each user, longer and more expensive development is required compared to gate arrays. However, standard cells offer greater design freedom, and more highly functional products can be made. Now standard cells with several hundred thousand gates are manufactured commercially.

7.2. Wafer Fabrication

In the manufacturing of ICs, major advances have been made toward higher integration, finer patterning, and more precise processing. On the other hand, the huge investment costs for sophisticated equipment and the R & D expenses required to meet the above requirements have resulted in increased manufacturing costs. Therefore, manufacturers are making efforts aimed at process rationalization, yield improvement, automation, manpower reduction, and standardization.

The basic flow sheet of wafer processing is shown in Figure 25:

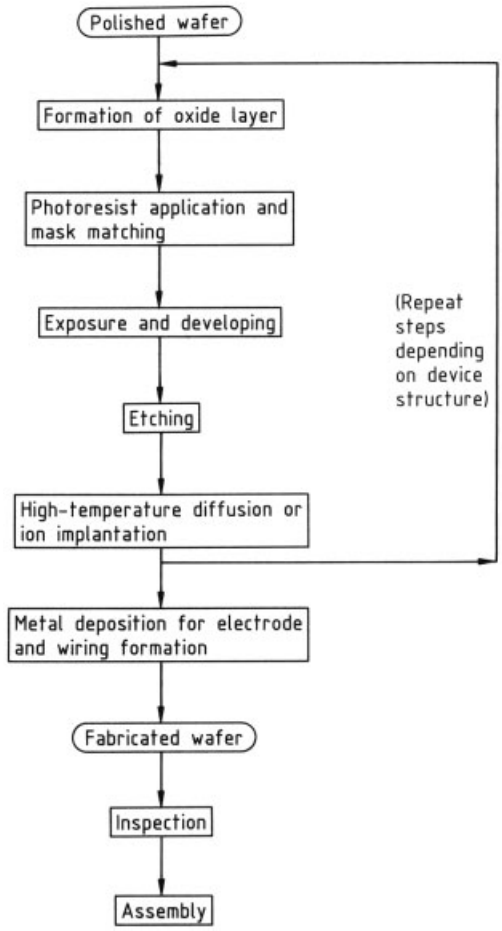

Figure 25. Wafer fabrication process

1) The polished wafer is heated to form an SiO_2 insulating layer on the surface.
2) The SiO_2 layer is coated with photoresist (a photosensitive resin) and prebaked. The circuit pattern is formed by using a photomask on the prebaked surface for mask matching.
3) The wafer is then exposed to UV light, for example, and developed to remove exposed or unexposed photoresist, depending on the type of resist.
4) The wafer is washed and dried and after postbaking, the SiO_2 layer is removed by wet or dry etching.
5) Dopants are then implanted by high-temperature diffusion or ion implantation in areas of the wafer surface where SiO_2 has been removed. (Steps 1 to 5 are repeated, depending on the device structure.)

6) Metals are deposited on the wafer for the formation of wiring and electrodes.
7) The electrical properties of the pattern are inspected by probing after the underside of the wafer has been polished.

The IC elements thus fabricated are divided into individual pieces by scribing and then sent to the assembly process, which includes die bonding, wire bonding, packaging, sorting, and screening.

7.2.1. Formation of Insulation Layer

The establishment of insulation layer technology contributed greatly to the development of the MOS transistor because the quality of the wafer surface on the outside can be maintained as well as the inside of the wafer. Various kinds of insulation layers [SiO_2, Si_3N_4, borophosphosilicate glass (BPSG), etc.] are formed by high-temperature oxidation or chemical vapor deposition (CVD) for gate insulation, capacitor insulation, interlayer wiring insulation, element separation, passivating layers, etc.

SiO_2 layers made by high-temperature oxidation have been widely used as insulation layers for gates and capacitors. SiO_2 layers are formed by reaction with atmospheric or high-pressure dry or wet oxygen at 700 – 1200 °C. Batch diffusion furnaces or lamp furnaces are used.

To form interlayer insulation for multilayer wiring, CVD is used. CVD is classified into low-pressure, atmospheric, and plasma CVD. For SiO_2 and Si_3N_4, the insulating layer can be made at temperatures as low as 300 °C by using the plasma CVD method. In this case, silane is generally used, but tetraethoxysilane gives layers with better flatness.

To form a smoother coated layer, phosphosilicate glass (PSG) or BPSG is applied by using atmospheric CVD. The latter can be deposited at lower temperature.

The method of local oxidation of silicon (LOCOS) has been used for the separation of electrical elements, and this technology is the basis of the development of MOSLSI. LOCOS consists of the following steps:

1) An SiO_2 layer is formed by high-temperature oxidation
2) An Si_3N_4 masking is applied to allow selective oxidation for element separation
3) Dopant is introduced into the separated areas
4) A thick SiO_2 layer is formed by oxidizing the separate areas

To improve and stabilize the electrical properties of the semiconductor surface, a passivation layer is formed to prevent contamination of and damage to the surface. Two methods are used for this purpose. One is to make a protective layer of Si_3N_4; the other is so-called gettering, in which contaminants such as sodium ions are absorbed or deactivated by the PSG layer. Sometimes both processes are combined.

Other purposes of passivation are (1) protection of the SiO_2 – Si interface, (2) protection of aluminum wiring, and (3) protection of electrical contact pads after wire bonding.

7.2.2. Lithography

Lithography (see also → Imaging Technology, Chap. 4.3.) is a technology for pattern formation derived from photography. To form the pattern of a device or circuit, the semiconductor substrate is coated with photoresist and is exposed, developed, and etched through a photomask. This technology is an important basis for achieving high integration of semiconductor devices, and finer patterning must be pursued for higher integration. Table 7 shows the relationship among the degree of integration, minimum line width, chip area of a DRAM, and the exposure technology.

Depending on the source of light, the technology of lithography is called photolithography, electron-beam lithography, ion-beam lithography, etc. Photolithography is most widely used. The equipment used in lithography is called a stepper, which employs g-line light (wavelength 436 nm) from a high-pressure mercury lamp. The stepper exposes the photomask negatives through a lens by a step-and-repeat process because the field is too small to image the entire wafer.

A further increase in integration requires improvements in resolution, depth of focus, field size, alignment, etc. This means using shorter-wavelength light. The first step is to use the ultraviolet i-line (365 nm) for 16-Mbit DRAMs; then excimer lasers of KrF (248 nm) and ArF

(193 nm) should be used for higher integration. For the next generation, IC exposure by electron beam or soft X-rays is being studied for >1-Gbit integration. The wavelengths of these sources are sufficiently short, but the equipment itself is still in the developmental stage.

7.2.3. Doping

Impurity doping is a basic technology of device fabrication. Thermal diffusion, ion implantation, plasma doping, and laser doping are all used for impurity doping. Among these, thermal diffusion and ion implantation are the most popular and are described here.

7.2.3.1. Thermal Diffusion

In thermal diffusion, the diffusion source – either gas (PH_3, AsH_3, B_2H_6), liquid ($POCl_3$, BBr_3), or solid ($Al_2(PO_3)_3$, BN, B_2O_3) – is heated to ca. 1000 °C. The resulting dopant oxides adhere to the surface of the substrate and are reduced to their elements by the substrate. Dopant atoms then diffuse into the substrate. This process is known as two-step diffusion.

The other process is solid-state diffusion. In this case, a dopant-containing layer is formed on the substrate surface by spin application or CVD and then heated.

The productivity of thermal diffusion is very high, and the controllability in the high-concentration diffusion range is excellent because the maximum surface concentration is determined by thermal equilibrium. However, in the low-concentration diffusion range the controllability is unsatisfactory. Therefore, the thermal diffusion process is generally used when the requirements for concentration control are not very strict.

7.2.3.2. Ion Implantation

The principle of ion implantation was invented in the 1950s by Western Electric or Nishizawa and was applied in the 1970s to the threshold control of the MOS transistor at low dopant concentrations. This technology has now replaced thermal diffusion in many cases.

In ion implantation, a gas or vapor that contains the dopant element is converted to the plasma state; the ions in the plasma are accelerated by a high-energy field and implanted in the semiconductor substrate. This technology has various advantages over thermal diffusion:

1) Since the amount of dopant implanted is controlled by the integrated value of the beam current, the accuracy is very high
2) Ions are implanted very uniformly, and the reproducibility is good
3) Because the implantation mechanism can be modeled theoretically, the distribution of implanted ions can be estimated
4) Dopants can be implanted through a surface layer

Nowadays, ion implantation is applicable to almost all LSI device fabrication. Because the collision of ions with atoms in the substrate results in many defects on crystal lattices, however, the damaged crystallinity must be restored by heat treatment (annealing).

As shown in Table 8, the equipment is classified into four types depending on the strength of the electric current applied. With the development of submicron devices, the trend is toward high current and high energy.

Table 8. Methods of ion implantation

Beam	Acceleration and voltage		
	10–60 keV	60–200 keV	400–1000 keV
Low (10–100 μA)			High-energy machine
Middle (100 μA–2 mA)	Middle-current machine	Middle-current machine	High-energy machine
High (2–30 mA)	High-current machine	High-current machine	
Ultrahigh (ca. 100 mA)			Oxygen implantation (developmental)

7.2.4. Etching

Etching is applied for either surface treatment or selective etching during lithography. The two types of etching are wet and dry. In wet etching, liquid chemicals such as hydrofluoric acid are

used, whereas dry etching utilizes the chemical or physical action of highly reactive gases.

Wet etching has been used for many years because it causes no damage to the substrate, and the throughput and selectivity are high. However, because the etchant solution penetrates into the boundary between photoresist and substrate, the accuracy of device fabrication deteriorates. Therefore, wet processing is not applicable to the fine patterning of highly integrated circuits.

The processing accuracy of dry etching is much better than that of wet etching, and automation is easy. For these reasons, dry etching is replacing wet etching despite low throughput and damage to the substrate surface.

Plasma etching, reactive etching, reactive-ion etching, and photoetching processes are used, depending on the material, shape, and structure of the devices. Dry etching processes are listed in Table 9.

Table 9. Dry etching processes

Mode	Mechanism	Source of reaction	Reactive gas
Chemical	plasma	radical	CF_4, $CF_4 + O_2$
	photo (UV, laser)	active ion	Cl_2, $Cl_2 + SF_6 + NF_3$, SF_6
Physicochemical	reactive ion	active ion	CCl_4, CF_4, $SiCl_4$
	reactive ion beam	active ion	SF_6, CHF_3
Physical	sputtering	inactive ion	Ar
	ion beam	inactive ion	Ar

7.2.5. Washing

Semiconductor surfaces are very sensitive, and device characteristics are affected not only by impurities in the bulk of the semiconductor but also by impurities adhering to the surface. These surface effects increase with increasing fineness of pattern. Therefore, washing technology is very important to obtain expected device characteristics with high yield.

Surface impurities include particles, metals, organic substances, oxide layer, etc., which are removed by wet or dry washing. With increasing integration and finer patterning of devices, the requirements for surface cleanliness are becoming more stringent. The wet washing process is used mainly because it can be applied to most impurities, including fine particles.

7.2.6. Wiring Technology

To increase the integration of ICs, the area used for wiring on the chip must be decreased. However, reduction of wiring-pattern width is limited by the current-carrying capacity and the wiring resistance. To solve this problem, multilayer wiring technology was developed for bipolar LSIs and MOSLSIs in the early 1970s. As wiring materials, aluminum and polysilicides, such as $MoSi_x$ or WSi_x, are used.

8. Supporting Industries

8.1. Building

The most important issues in building specifications for semiconductor materials and device manufacturing are cleanliness and absence of vibration. The former is discussed in the next section, but basically the building should be designed to avoid contamination by recycling air from the utility or maintenance zone where various chemicals and gases are treated. As for vibration, the base and surrounding conditions such as roads and railways should be studied carefully. Also, the building structure should prevent the propagation of vibration from equipment installed in the building. For equipment that is especially sensitive to vibration a local antivibration system can be installed.

8.2. Clean Room

With the increasing density of integration, the effect of contamination on quality, manufacturing yield, and product reliability has become extremely critical because of the finer patterning. Therefore, the production of ultraclean wafers is a core technology of semiconductor device fabrication. Of course, the same considerations apply to wafer processing (polishing, washing, inspection, and packaging).

The primary factors in yield drop are pattern defects caused by particles adhering to the wafer or photomask. Determining the minimum particle size that causes pattern defects is very complicated, but empirically a particle size one-tenth

of the minimum line width (design rule) must be considered for removal from the standpoint of yield and reliability. In the case of memories, when the integration is increased by a factor of 4, the chip area increases by 40 – 50 %. This increase in area requires a further reduction of the remaining finer particles per unit area. Particles come from the environment, equipment, and materials (including gases and chemicals), but the major source of particles is the operators. Therefore, a strong trend exists toward automated operation.

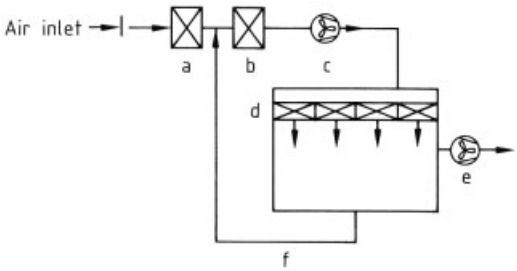

Figure 26. Schematic of a laminar-flow clean room
a) Primary filter; b) Middle filter; c) Intake fan; d) Final filter; e) Exhaust fan; f) Recycled air

Table 10. U.S. Federal Standard of Cleanliness∗

Class	Particle size, µm∗∗				
	0.1	0.2	0.3	0.5	5.0
1	35	7.5	3	1	NS
10	350	75	30	10	NS
100	NS	750	300	100	NS
1 000	NS	NS	NS	1 000	7
10 000	NS	NS	NS	10 000	70
100 000	NS	NS	NS	100 000	700

∗ Number shown means tolerable upper limit of particles of each size per cubic foot (1 $Ft^3 \approx 28$ L).
∗∗ NS = not specified.

The basis of clean room specification is the U.S. Federal Standard 209 D (Table 10). However, more recent LSI processes necessitate a more stringent specification of 1 – 10 particles of > 0.1-µm diameter per cubic foot (super clean room level). Clean rooms use either the total laminar flow or partial laminar flow system; the former is generally used in advanced clean room technology. As shown in Figure 26, outer air is supplied to the clean room through primary, secondary, and final filters. The final filter, installed in the ceiling of the room, is specified by 0.1-µm particle size and is known as a high-efficiency particular air filter (HEPA filter) or ultralow-penetration air filter (ULPA filter). The volume of air supplied is equivalent to that which replaces the room air at a rate of 70 – 500 times per hour. Dust particles generated in the room are taken out with the recycled air and removed by filters. About two-thirds of the total air is recycled.

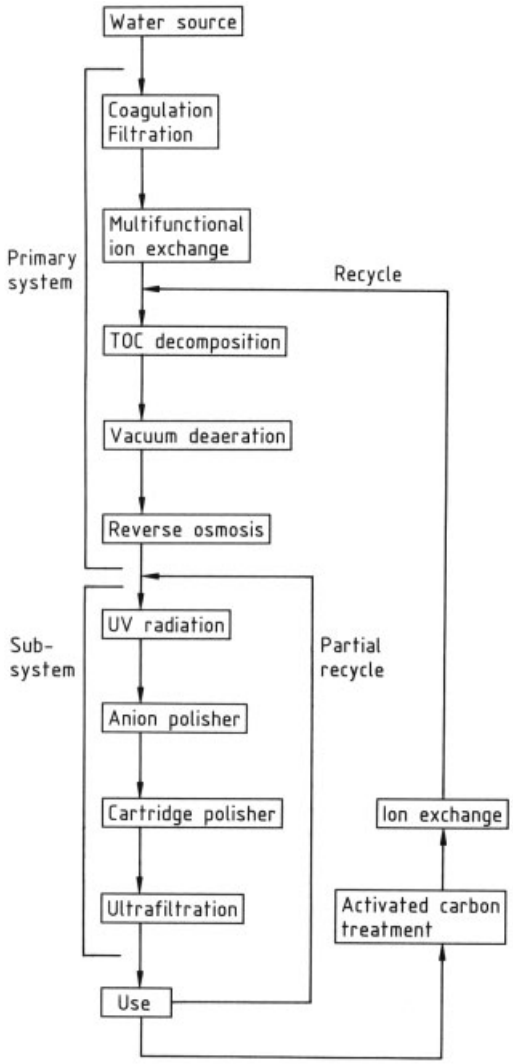

Figure 27. Modern ultraclean water system

8.3. Ultraclean Water

Because ultraclean water is generally used at the final stage of processing, its quality strongly affects the cleanliness of wafers. As shown in Table 11, specifications are becoming increasingly rigorous. Major items to be checked are number and size of particles, ions, and live bacteria. In addition, removal of total organic carbon (TOC) and dissolved oxygen is required to prevent formation of an oxide layer.

Particle specifiction is based on past experience. Metal ions such as chromium and iron must be limited to the 0.01-ppb level. Live bacteria cause shape defects, and the elements phosphorus, sodium, and potassium can diffuse from them to the inside of wafer on heat treatmet to form abnormal conducting areas or to cause performance to deteriorate. The TOC is a nutrient for live bacteria and causes defects in the oxide layer of the gate. Dissolved oxygen helps increase live bacteria, and the thickness of the natural oxide layer formed by water rinsing depends on the dissolved oxygen concentration.

An ultraclean water system consists of a primary clean-water system, subsystems, and a wastewater recovery system. In the primary system, industrial or well water is treated by filtration or multifunctional ion exchange to remove particles, ions, organic compounds, live bacteria, and dissolved oxygen. The purified water is sent to subsystems where it is treated by UV light, anion polishers, and an ultrafilter to remove TOC and other microimpurities. To lower manufacturing costs and volume consumption, high-efficiency wastewater recovery systems must be used.

An ultraclean water system for 16-Mbit DRAMs is shown in Figure 27. This system utilizes the following technologies and achieves the target levels listed in Table 11:

1) Removal of fine particles by a double-skin ultrafilter module
2) Removal of TOC, dissolved salt, and SiO_2 by two-stage low-pressure reverse osmosis
3) Use of low-elution ion-exchange resins to remove TOC
4) Sterilization by high-energy UV light
5) Sterilization by hot water (80 – 90 °C) using heat-resistant poly(vinylidene fluoride) piping
6) Lowering of dissolved oxygen concentration by nitrogen purging and membrane degassing

Stains or water marks caused by water droplets remaining on the wafer and electrostatic breakdown can be avoided by using high-pressure water jets. On-line continuous measurements of fine particles and TOC are available. With increasing wafer diameter, the consumption of ultraclean water is increasing, and the improvement of space efficiency, the development of recycling systems or of fully closed systems, and cost reduction are important issues for ultraclean water.

8.4. Gases

Various gases are used in LSI fabrication:

1) Conventional industrial gases such as H_2, N_2, O_2: These are used widely in many steps, and the consumption volume is high. LSI plants usually have on-site production or storage facilities, and these gases are supplied by pipeline.
2) Highly toxic and explosive gases such as SiH_4, AsH_3, PH_3, and B_2H_6 for epitaxy and dopant diffusion.
3) Halogenated gases or chlorofluorocarbon gases for etching (HCl, Cl_2, CCl_4, $SiCl_4$, CF_4, etc.).

These gases are classified in Table 12 from the standpoint of application.

Table 12. Major gases used in semiconductor process

Process	Gases
Epitaxy and CVD	SiH_4, AsH_3, PH_3, $(C_2H_5)_2Te$, $(C_2H_5)_2Zn$, O_2, N_2O, NH_3
Surface oxidation	O_2, N_2, H_2, H_2O
Diffusion and ion implantation	AsH_3, PH_3, B_2H_6

With the development of finer patterning and higher-performance devices, raw materials of very high purity, including gases, are always requested. In many cases, however, the analytical methods have not yet been established, and the relationship between gas purity and device performance is not clear. To obtain high-purity gases at the point of use, attention must be paid not only to the manufacturing of the gases but

Table 11. Quality requirement for ultraclean water

	Integration			
	1 Mbit	4 Mbit	16 Mbit	Currently achieved
Minimum line, width, µm	1.0 – 1.2	0.8	0.5	
Specific resistivity, MΩ · cm	18 (min.)	18 (min.)	18.1 (min.)	18.2 (min.)
Particles				
Size, µm	0.1	0.08 – 0.1	0.05	0.07
Number, L^{-1}	20 000 (max.)	10 000	10 000	1
Number of live bacteria, L^{-1}	50 (max.)	50	10	1
Total organic carbon, ppb	50 (max.)	30	10	1
Dissolved oxygen, ppb	100 (max.)	50	50	10

also to transportation, storage, piping, etc. In particular, the cylinder materials, including valves, O-rings, and packing, and the method of filling are very important.

Environmental issues relating to gases are discussed in Chapter 9.

8.5. Other Chemicals

8.5.1. Quartz

Quartz glass is widely used in semiconductor manufacturing processes as boats, crucibles, piping, bell jars, etc., for the following reasons:

1) Very high heat resistance and low coefficient of thermal expansion
2) Chemically stable except toward HF and H_3PO_4 (>300 °C) and stable to molten metals
3) Good electrical insulation and low dielectric loss factor regardless of frequency

However, when quartz glass is used as a boat or crucible in contact with molten materials, possible contamination must be carefully checked in advance. In the case of silicon single-crystal pulling, oxygen and metal impurities seriously affect the quality of the silicon wafer and the electrical properties of devices. Silicon permeation from quartz affects the conductivity of GaAs.

Major impurities in quartz are aluminum (5 – 30 ppm), boron (0.1 – 0.3 ppm), iron, sodium, potassium (0.1 – 2 ppm); copper (0.05 – 0.1 ppm); and 1 – 200 ppm of water. Different impurities are critical, depending on the processes in which quartz is used.

For the process tube used to supply gases for doping, CVD, etc. in LSI fabrication, heavy and alkali metals must be removed. For this purpose, natural quartz powder is melted by an oxyhydrogen flame, and copper and alkali metals are removed by evaporation.

For the diffusion process, quartz tubes are made by electrical melting in vacuum furnaces. This quartz has a higher content of metal impurities, but it is resistant to high-temperature flexure because it contains almost no water.

Synthetic quartz tube made from $SiCl_4$ by oxidative melting in an oxyhydrogen flame contains very small amounts of metallic impurities, but the water content is high. This is used as crucibles for making high-purity single crystals.

8.5.2. Pyrolytic Boron Nitride

Pyrolytic boron nitride (PBN) is made from BCl_3 and NH_3 by vacuum chemical vapor deposition. The reaction product BN deposits on a graphite mold, which is subsequently removed. PBN crucibles are used for making semi-insulating GaAs by LEC (see Chap. 4). The characteristics of PBN are as follows:

1) Very high purity because the CVD process is used
2) Very inert to chemicals and molten metals
3) Highly heat resistant (in an inert atmosphere it is resistant up to 3000 °C and has better thermal shock resistance than quartz)
4) Very expensive but can be used repeatedly

PBN is also used for MBE crucibles. For the production of silicon single crystals, PBN is not used because the electrical properties are very

sensitive to boron contamination. Even in GaAs LEC, boron contamination remains a problem.

8.5.3. Graphite

High-purity graphite is an indispensable material for the semiconductor industry and is used as a susceptor, heater, and thermal insulator for single-crystal and epitaxial growth of silicon, GaAs, and others.

Manufacturing high-purity graphite is a time-consuming process, and quality control is very difficult. Coke and tar pitch are mixed and molded by extrusion or by a rubber press, preheated at ca. 900 °C, and then graphitized at ca. 2800 °C. Preheating and graphitization are repeated for purification. This process takes several months. At this stage, the product still contains about 0.5 % impurities (Al, V, Si, B, SiC, etc.). For further purification, graphite is heated with chlorine or fluorine, which do not react with it, at 2500 – 2800 °C. A typical analysis of purified graphite is shown in Table 13.

Table 13. Typical analysis of purified graphite

Constituents	Content, ppm (max.)
Ash	10
B	0.1
Fe	0.1
Cu	0.01
K	0.01
Na	0.01
Al	0.02
Si	2

With the increase of single-crystal diameter, strength improvement of graphite became very important, and extrusion molding was replaced by the use of rubber presses, which can make more uniform and stronger products.

For the group III – V materials GaAs and GaP, the thermal condition of LEC is more severe than with silicon and high pressure is also applied. Therefore, the requirements for graphite quality are more rigorous.

8.5.4. Others

As shown in Table 14, many kinds of chemicals are used in lithography, etching, washing, and other steps. Particle content and metal impurities strongly affect the quality of devices. Currently the quality of electronic-grade (EL) chemicals basically satisfies most users.

Table 14. Chemicals used in device fabrication

Washing	
Particles	$NH_4OH - H_2O_2 - H_2O$, Choline $- H_2O_2 - H_2O$
Metals	$H_2SO_4 - H_2O_2$, $HCl - H_2O_2 - H_2O$, $HNO_3 - HF - H_2O$
Organics	$H_2SO_4 - H_2O_2$, $NH_4OH - H_2O_2 - H_2O$
Natural oxide layer	$HF - H_2O$, $HF - NH_4F - H_2O$
Drying	isopropanol
Lithography	tetramethylammonium hydroxide for development butyl acetate for rinsing $H_2SO_4 - H_2O$, phenol for peeling off
Etching	
of SiO_2	HF, etc.
of Si	$HF - HNO_3 - CH_3COOH$
of Si_3N_4	H_3PO_4
of Al	$H_3PO_4 - HNO_3 - CH_3COOH$

Electronic-grade chemicals require further purification (e.g., by filtration, distillation, and ion exchange).

The particle content should be less than $20\,mL^{-1}$ of 0.1-µm particles for the production of 16-Mbit DRAM. Typical particle contents are, however, higher than this, and the particle content increases during transport. Chemical producers and LSI manufacturers are cooperating to solve this problem.

With regard to metal impurities, the requirements are also becoming much more rigorous as integration increases. In the case of 16-Mbit DRAM, $<0.1 - 1$ ppm is specified depending on the metal. Most commercially available chemicals (H_2O_2, NH_4OH, H_2SO_4, HCl, and HF) meet this requirement.

9. Environmental Aspects

In the electronics industry, many toxic and flammable substances are used from crystal growth to device fabrication. In 1985, NIOSH registered 182 substances in this area, and their number is increasing and changing with technological progress. In addition, most of these substances are discharged in various states (gas mixtures, liquids, fine particles, sludge, etc.), which makes treatment before they are released to the environment very complicated.

9.1. Gases

Many kinds of gases are used in epitaxial growth, CVD, formation of SiO_2 layers, doping, and dry etching, as shown in Table 12.

Among them, the hydrides (SiH_4, B_2H_6, PH_3, and AsH_3) are very toxic, highly explosive, and inflammable. When they leak, they tend to combust spontaneously. Hence special safety measures such as installation of a gas-leak detector, thorough ventilation, waste-gas treatment, emergency counterplans, and periodic inspections should be followed. The same attention should be paid to H_2 and NH_3.

When mixed with flammable gases, O_2, N_2O, and NF_3 accelerate explosion. The use of NF_3 as a dry etchant is increasing. With inert gases such as N_2 and argon the danger of an unbreathable atmosphere must also be considered.

9.2. Liquid Chemicals

Acid, alkali, and salts are used mainly in wet etching and washing processes. The consumption of HF and mixed acid (mixture of nitric and sulfuric acid) is very high. Both result in serious burns, and HF readily penetrates the skin and destroys tissues. Toxic gases are sometimes evolved during acid washing. Organic solvents are used in deoiling and photoresist processes.

These chemicals must be used in the specified closed station under a ventilation system to prevent the exposure of operators and the contamination of clean rooms.

9.3. Metals

Metals such as aluminum, antimony, arsenic, lead, zinc, titanium, and nickel are used in vacuum deposition or ion implantation. They are all extremely pure, and consumption is low. For arsenic and lead, special attention must be paid to the exposure of workers, for example, in the cleaning of GaAs crystal pullers.

9.4. Chlorofluorocarbons (CFCs)

In June 1990, to avoid destruction of the ozone layer in the stratosphere, the decision was made internationally to ban the use of certain CFCs (CFC 11, 12, 113, 114, 115) and 1,1,1-trichlorethane, for example, by the year 2000. In addition, a ban on hydrochlorofluorocarbons (HCFCs) was also discussed. Since then, most electronic companies are aiming to essentially stop the use of these CFCs by 1995.

In the IC manufacturing process, CFC 113 and 1,1,1-trichloroethane are used in washing, and other CFCs are used in dry etching. Consumption of the first two is especially high. Besides destruction of the ozone layer, another issue is that CFCs are carcinogenic, leak into the soil, and contaminate ground water. In response to the CFC regulations, both short- and long-range actions are being taken. The following are the short-range actions:

1) Use of closed systems to avoid the evaporation of CFCs
2) Reduction of consumption by improving equipment and operation
3) Recovery and recycle of used CFCs

Medium- or long-range counter measures include the development of more environmentally friendly CFCs such as HCFC 134 a and HCFC 225. World CFC producers have established an international program for alternative fluorochlorocarbon toxicity (PAFT) to confirm the safety of these products before commercialization.

Other developments are the replacement of CFCs by alcohol or water for washing, or even conversion to a washingless process. In alcohol washing, ethanol, isopropanol, and 5-fluoropropanol, for example, are being studied. Water washing includes the use of emulsifiers and terpenes besides pure water. Replacement of CFCs by other substances is still in a developmental stage and requires substantial changes in the entire manufacturing process to meet customers' quality requirements.

Attempts are being made to replace other dangerous chemicals by safer ones; the following are examples of such trials:

1) Use of tetraethoxysilane instead of silane in plasma CVD for formation of the SiO_2 layer
2) Formation of the SiO_2 layer at lower temperature by using liquid chemicals such as tetraalkoxysilanes

3) Use of arsenic metal or organoarsenic and phosphorous compounds to replace AsH_3 and PH_3 in OMVPE and ion implantation
4) Replacement of cellosolves by safer chemicals in the photoresist process

10. Economics

Silicon. Currently (1991), five Japanese (Shin-etsu, Mitsubishi Materials, Osaka Titanium, Komatsu Electronic Metals, and Toshiba Ceramics) and two German (Hüls and Wacker Chemitronic) companies are manufacturing silicon wafers; Texas Instruments, Motorola, and AT & T in the United States are minor producers who use their wafers captively. The production share of Japanese companies is ca. 70 %. Most of these seven companies have global production systems in the United States, Europe, and Asia, or in two of these three areas.

Worldwide production statistics for single-crystal silicon ingot are not available, but according to the Japan Society of Newer Metals, estimated Japanese production volumes in 1991 and 1992 were 2210 and 2300 t, respectively. Based on these figures, the estimate of worldwide production volume in 1991 and 1992 is 3200 and 3300 – 3400 t. On the other hand, Dataquest estimated worldwide consumption of silicon wafers in 1991 and 1992 at 2133 and 2237 million square inches, which corresponds to <200 t in weight. This means that the yield of polished silicon wafers from single-crystal ingots is very low. Production of silicon wafers in 1991 amounted to $\$ 2.27 \times 10^9$.

Wafers are becoming larger with increasing chip size. Now, 6-inch wafers are the most widely used, and 4- or 5-inch wafers are limited to specialty uses. Eight-inch wafers were expected to be used for the commercial production of 4-Mbit DRAM. However, their introduction has been delayed because of the slow market for 4-Mbit DRAM.

Compound Semiconductors. Worldwide statistics for compound semiconductors are not available. However, except for minor captive production by Hewlett – Packard (United States) and Siemens (Germany), about 90 – 95 % are produced in Japan, including epitaxial wafers. According to the Japan Association of Compound Semiconductor Manufacturers, Japanese wafer production amounted to $\$ 159 \times 10^6$ in 1989 and $\$ 186 \times 10^6$ in 1990. Therefore, worldwide production in 1990 was ca. $\$ 200 \times 10^6$, which is less than 10 % of silicon wafer production. The unit price of compound semiconductor wafers is several times higher than silicon, and the percentage of the more expensive epitaxial wafer is high. Based on these, the production volume of compound semiconductors is only 1 – 2 % that of silicon.

Both visible and infrared LEDs are made exclusively from compound semiconductors, GaAs, GaP, and InP. GaAs is becoming established in amplifiers in high-frequency areas such as satellite broadcasting and communications, and devices that require low consumption of electricity such as computer cache memories, although large-volume consumption is not expected.

In LED production, 2 – 2.5-inch epitaxial wafers are used; 3- and 4-inch substrates are used in other applications. Device manufacturers are interested in using 5-inch wafers or larger, but this would necessitate a large investment by both wafer and device producers, and is not economically justifiable at this stage.

Major Japanese wafer makers are as follows:

1) Epitaxial wafers for LEDs: Mitsubishi Kasei, Showa Denko, Shin-etsu
2) Substrates for microelectronic uses: Mitsubishi Kasei, Sumitomo Electric, Hitachi Cable

Semiconductor Devices. Semiconductor devices are now used not only in computers but also in almost all industries relating to everyday life such as communications, displays, home appliances, automotives, aeroplanes, instrumentation, and automation.

According to the forecast of the World Semiconductor Trade Statistics (WSTS), worldwide consumption is estimated as shown in Table 15. Production and consumption of semiconductor devices began in the United States and was dominant there until the middle of the 1980s. Since then, Japan has become the largest producer. In the 1990s market shares were Japan 40 %, United States 30 %, Europe 17 %, and remainder 13 %. In the 1990s, the share of Asian countries

except Japan increased because of the rapid expansion of Korean companies and the transferral of their plants to the Asian area by device fabricators in Japan, the United States, and Europe.

Table 15. Forecast of worldwide semiconductor consumption by WSTS ($ 10^9)

Country	1991	1992	1993	1994
United States	15.6	18.0	20.1	21.7
Japan	21.5	24.5	27.1	29.2
Europe	10.2	11.1	12.4	13.4
Others	8.3	10.0	11.6	12.8
Total	55.5	63.6	71.2	77.2

The top ten device manufacturers in 1991 are as follows, according to Dataquest (J = Japan, U = United States, E = Europe): (1) NEC (J); (2) Toshiba (J); (3) Hitachi (J); (4) Intel (U); (5) Motorola (U); (6) Fujitsu (J); (7) Texas Instruments (U); (8) Mitsubishi Electric (J); (9) Matsushita Electronic (J); (10) Philips (E).

11. References

General References

1. M. Kikuchi, *Story of Semiconductors, Physical Properties and Applications,* **NHK Books No. 52,** Japan Broadcasting Publishing Association, Tokyo 1967.
2. Y. Tarui, *Story of IC, From Transistor to VLSI,* **NHK Books No. 411,** Japan Broadcasting Publishing Association, Tokyo 1982.
3. *Semiconductors Handbook,* Revised Edition, Ohm, Inc., Tokyo 1977.
4. *Semiconductor Materials Handbook for Electronic Devices,* Fuji Techno Systems, Tokyo 1986.
5. *Encyclopedia of Advanced Electronic Materials,* CMC, Inc., Tokyo 1991.
6. *Electronic Components Handbook,* Dempa Shimbun, Inc., Tokyo 1980.

Specific References

7. L. R. Melby, et al., *J. Am. Chem. Soc.* **84** (1962) 3374.
8. Monsanto, JP 1 337 385, 1 227 669.
9. T. Inada, et al., 52nd Autumn Meeting of the Japan Soc. of Appl. Phys., (Oct. 9, 1991).
10. H. Nelson, *RCA Rev.* **24** (1963) 603.
11. V. W. Koester, B. Thoma: *Z. Metallkde* **46** (1955) 291.
12. R. A. Reuhrwein, JP Toku-kou-shou 38 7709 (1963).
13. H. M. Manasevit, *Appl. Phys. Letter* **12** (1968) 156.
14. G. C. Guenther, *Z. Naturforsch. A* **13 a** (1958) 1081.
15. A. Y. Cho, et al., *Prog. Solid State Chem.* **10** (1975) 157.

Sensitizer Dyes → Photography
Sensors → Chemical and Biochemical Sensors

Shoe Polishes

PETER A. BURKE, Kiwi Brands Inc. (Division of Sara Lee Corp.), Douglassville, Pennsylvania 19518, United States

1. **Introduction** 1
2. **Footwear Materials and Processes** . . 1
3. **Solvent-Based Shoe Polish** 2
3.1. **Composition** 2
3.2. **Production** 4
3.3. **Quality Specifications** 4
4. **Emulsion Shoe Creams** 4
4.1. **Composition** 5
4.2. **Production** 5
4.3. **Quality Specifications** 5
5. **Liquid Polishes** 6
5.1. **Composition** 6
5.2. **Production** 7
5.3. **Quality Specifications** 7
6. **Other Shoe Dressings** 7
7. **Environmental Protection** 7
8. **Economic Aspects** 7
9. **References** 7

1. Introduction

Ever since the Romans began wearing sandals, shoe care has been practiced as a way of softening, protecting, and extending the duration of footwear use. Animal fats were applied to leather footwear to replenish natural fatty oils lost as a consequence of exposure to the elements. As the chemistry of leather treatment progressed, providing corrected-grain, chrome-tanned, finished footwear, the sophistication of shoe dressings advanced concomitantly. Modern polishes consist of a complex mixture of waxes, silicones, and polymers dissolved in appropriate solvents. This delicate balance of materials provides a protective coating that is easily applied, adheres strongly, and is easily buffed to the desired gloss. The selection of solvents is driven by chemical properties of the gloss components, and ranges from water to volatile organic solvents.

It is interesting to note that approximately the same shoe-care formulae are used in different countries irrespective of location, with the exception that higher melting waxes are employed in tropical climates. However, the *type* of shoe dressing preferred does reflect certain cultural differences, as exemplified by the fact that liquids are most common in Mexico, southern Europe, and the Far East, whereas paste polishes have a greater share of the market in North America, Africa, and Australia, and in northern Europe there is heavier use of shoe cream.

2. Footwear Materials and Processes

Most dress shoes utilize fully corrected-grain leather for the upper section, the part that requires polishing. The first stage in the conversion of animal skin into leather is tannage, which is the impregnation of chromium or tannic acid salts into the hide (→ Leather). Tanning the leather prevents resident microorganisms from putrefying the hide upon rewetting, a process that causes irreversible damage. A typical shoe upper utilizes rough tanned ox hide, such as zebu or brahma. The hide is first *chrome tanned*, whereby chromium sulfate is introduced to fix the hide; that is, chromium ion complexes with carboxyl groups in the hide proteins. Basic conditions favor the formation of large chromium complexes, permitting cross-linkage between adjacent protein molecules. This cross-linking reaction in turn provides a stabilizing effect against excessive drying and the otherwise detrimental action of hot water. The hide is then dyed and softened using a fat liquour (an oil-in-water emulsion used to introduce oil into the hide).

Finally, texture is imparted to the leather by buffing the grain with fine sandpaper (e.g., #220 paper) to make it smoother and more uniform. Generally, the grain of the leather requires additional finishing to ensure an acceptable "break", manifested as creases in the shoe as it flexes during walking. The finishing of a fine, fully corrected-grain leather requires the application of several finishing coats prior to a high-gloss

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_575

top coat such as nitrocellulose. Typical materials involved in achieving a conventional corrected-grain finish by a four-step process are listed in Table 1.

Table 1. Compositions of the individual coats applied in a four-step leather finishing process ([9], p. 417)

Component	Content, wt %			
	Step 1	Step 2	Step 3	Step 4
Impregnating polyurethane polymer dispersion	20.0			
Water	70.0	52.5	50.0	50.0
Solvent penetrator (2-butanone)	10.0			
Pigment paste		20.0	20.0	
Wax emulsion		2.5	2.5	
Casein emulsion			7.5	
Acrylate polymer dispersion		25.0	20.0	
Nitrocellulose emulsion				50.0

In the scheme outlined in Table 1, a solvent penetrator (2-butanone) is used to promote the adherance of a polyurethane polymer, which serves as a bottom base coat. This is usually followed by an acrylate polymer dispersion to provide gloss and bind pigment to the substrate. In the third step, more polymer solution is applied with casein to improve adhesion and/or a wax base for additional gloss. A nitrocellulose top coat is applied last to ensure a brilliant protective coating.

Leather for the upper parts of casual shoes is treated less extensively, providing a more natural "leather" look. Such shoe uppers are typically manufactured from silicone- or oil-tanned leathers, which are subjected to fewer finishing steps and often remain undyed and relatively porous. Special shoe- care products are required to maintain the resulting natural-looking finish.

Man-made plastic-type shoe materials were introduced a number of years ago, but they still play a only minor role in shoe construction. Thus, the descriptions of formula attributes that follow are predicated on leather as the upper shoe construction material.

3. Solvent-Based Shoe Polish

Cream and paste shoe polishes have been utilized since the early 1800s to protect, repair scuffs, and provide shine. Three fundamental groups of materials contribute to the formula properties: solvents, wax blends, and dye dispersions. A balanced formula is required to produce a shoe polish that possesses such mandatory attributes as high solvent-binding capacity, good heat stability, ease of application, rapid buffing, scuff coverage, high spectral reflectance, and satisfactory paste-surface physical characteristics.

3.1. Composition

Wax Blend. The most critical decisions during formulary development center around the selection of wax components, wax ratios, and total wax concentration. After volatilization of the solvent, the coating properties that remain are a direct reflection of interactions between the wax components and the dye system.

A typical solvent paste polish must contain 20 – 40 % wax components to produce a paste with satisfactory hardness as measured by a precision penetrometer. Care must be taken to balance the relative percentages of high-melting waxes versus low-melting paraffin-type materials in order to achieve the desired surface characteristics and paste firmness. For instance, if low-melting waxes were to be used exclusively in a formulation with only 20 % wax solids, the resulting paste would be unacceptably soft, with poor solvent-binding properties. ("Solvent-binding" is the phenomenon that prevents a paste from rapidly becoming a dry solid.) Most shoe-polish formulae contain a mixture of microcrystalline natural ester waxes, oxidized montan waxes, and a paraffin base (→ Waxes). Selection of the wax components requires a consideration of solvent-binding capacity, gloss value, and hardness index. Synthetic waxes like the non-emulsifiable polyethylenes and Fisher – Tropsch wax can be substituted to some extent to produce equivalent results.

To maintain satisfactory paste attributes, a good balance must be achieved between solvent-binding capacity and gloss value. Paraffin is a major wax ingredient, but it displays both poor gloss and poor solvent-binding properties. Carnauba, candelilla, refined montan, and certain oxidized montan ester waxes from BASF and Hoechst improve the gloss properties markedly. Montan, ozokerite, and Hoechst Wax O, OM,

or OP also improve the solvent binding capacities of paraffins. Table 2 summarizes the general solvent-binding and gloss properties of various waxes based on their contributions to paste-polish characteristics.

Table 2. Characteristic properties of various waxes

Wax	Solvent binding	Gloss
Paraffin	poor	poor
Microcrystalline	moderate	poor
Carnauba	moderate	excellent
Hoechst Wax O	good	fair
Hoechst Wax OM	poor	good
Hoechst Wax OP	good	good
Hoechst Wax S	good	fair
Hoechst Wax F	poor	fair
Montan (refined)	good	good
Ozokerite	good	poor
Candelilla	poor to moderate	very good

Solvent. Essentially, a solvent acts to solubilize wax components in a form that permits the development of a relatively uniform, thin film coating on the leather substrate. Upon evaporation, a thin wax film can be buffed to a desirable gloss. Many different solvents have been utilized in shoe polish [e.g., turpentine, mineral spirits, V.M. & P. ("Varnish Makers and Painters") naphtha, and Stoddard solvent (ASTM D-484-52)], both singly and in combination. Generally, turpentine is considered the solvent of choice due to its relatively high kauri-butanol value, its volatility, and a characteristic "shoe-polish" odor. However, the high cost of turpentine has led to its being largely replaced by mineral spirits in most water-free pastes. The physicochemical properties of mineral spirits (boiling range, kauri-butanol value, evaporation rate, and density) compare favorably to those of turpentine. Nevertheless, the replacement of turpentine with mineral spirits must be undertaken carefully to ensure retention of sufficient aromatic content, since this affects how wax recrystallizes during manufacture. An excellent compromise is a solvent blend of turpentine and mineral spirits with a 10 – 50 wt % turpentine concentration.

Dye Dispersions. To provide appropriate depth of gloss, a basic solvent dye (e.g., Nigrosine Solvent Black #2) must be predispersed in a carrier prior to its incorporation into a wax blend or liquid molten paste. Basic nigrosine dye (→ Azine Dyes, Chap. 2.2.4.) must be neutralized with a fatty acid to make it soluble in the solvent/wax mixture. Normally, a nigrosine base is reacted with either oleic or stearic acid. Nigrosine stearate has the advantage of not lowering the melting point of the wax blend. Being a solid, however, it has the disadvantage of imposing extra processing steps. Rhodamine (→ Fluorescent Dyes, Chap. 4.) and chrysoidine (→ Azo Dyes, Chap. 5.1.1.) dye materials are treated in a similar fashion, leading to cordovan, oxblood, and brown polishes. Dyes supplied originally in an oil-soluble form can be incorporated directly into the solvent, molten wax, or paste.

Table 3. Representative shoe-polish formulations

Chemical component	Relative content, wt %					
Crude carnauba wax		2.0		1.5	5.5	2.0
Montan wax (refined)	2.5			2.0	3.0	2.5
Wax OP ∗	3.5		5.0	1.5		
Wax O ∗		6.5		6.0		3.0
Wax OM ∗	3.0		4.0			
Wax E ∗		2.0				
Wax F ∗	1.0		1.5		2.0	
Microcrystalline wax	1.0	0.5	0.5	1.0	1.5	1.0
Paraffin wax, 52 – 54 °C	12.0	14.0	16.0	12.5	12.0	7.5
Paraffin wax, 58 – 60 °C				3.5	4.0	12.0
Turpentine	74.0		35.0	49.0	69.0	39.0
Mineral spirits		72.0	35.0	20.0		30.0
Dye mixture	3.0	3.0	3.0	3.0	3.0	3.0
Properties ∗∗						
Gloss	G	VG	G	VG	VG	VG
Paste hardness	M	M	M	H	H	H

∗ American Hoechst Corp., a division of Hoechst Cellanese.
∗∗ Abbreviations: G, good; VG, very good; M, moderate; H, hard.

Formulae. The number of potential permutations of wax blends is large. Table 3 provides several representative formulae.

3.2. Production

Manufacture of a water-free shoe paste requires a minimum of two vessels to produce molten paste prior to filling. A steam- or electrically-jacketed vessel with good temperature control is essential for melting the wax and ensuring homogeneity. Typically, the paraffin wax is

melted first to provide a heat sink for the subsequent processing of higher-melting waxes. The stearin – dye mixture is added last due to its thermal lability and susceptibility to charring. The temperature required to obtain a completely soluble wax blend is typically 100 – 130 °C depending upon the nature of the wax materials.

The wax blend is then added slowly over 5 – 10 min to the solvent (turpentine or mineral spirits). The solvent temperature must be maintained at ca. 50 – 60 °C during this paste-formation step. Depending on the size of the vessel, agitation can be provided by a propeller mixer for most polishes. The mixture should be agitated for 15 – 30 min and then filled at a pouring temperature of 40 – 45 °C.

For more efficient manufacture, development of the proper consistency for a polish can be facilitated by passing the filled containers through a cooling tunnel maintained at 5 – 10 °C. Some formulations need not be cooled on an accelerated basis due to the nature of their wax constituents and the paste characteristics that develop naturally. Suitable precautions are required during solvent handling due to the potential for combustion of the mixture.

3.3. Quality Specifications

The resultant paste should present a hard, smooth, glossy surface appearance in the can. Unbalanced formulation or poor chemical processing may lead to dye incrustation, solvent exudation, bloom, or cake contraction, each of which constitutes a critical defect from the standpoint of consumer response.

Dye incrustation and bloom both result from poor binding of the dye – stearin component, causing preferential migration and recrystallization of the stearin at the surface.

Solvent exudation occurs with a wax blend lacking adequate solvent-binding capacity. Pastes rich in paraffinic wax often suffer from this phenomenon. *Cake contraction* is caused by shrinkage of the paste mass in the can, resulting in contraction from the side walls. This is caused by a formulation with an excessive hard-wax content, and leads to what is known in the trade as a "rattler". Such contraction may render the product totally unsalable within a few weeks or months.

4. Emulsion Shoe Creams

Shoe creams are generally much softer in consistency than paste polishes. They are offered in a wide range of colors for application on fashion shoes, and provide excellent staining of scuff marks with perfect matching of the original shoe color. Shoe cream formulae may or may not contain solvent, although most do include some solvent to improve wax solubilization and ease of application.

4.1. Composition

Wax/Solvent Emulsification. The formulator of a shoe cream must select waxes that are both emulsifiable and capable of displaying high gloss properties and high solvent-binding capacity. Examples include carnauba, beeswax, montan, and various oxidized montan waxes (e.g., Hoechst waxes E, S, KSL, and LP). The simplest emulsions are formed in situ by adding ammonia or morpholine together with a fatty acid like oleic or stearic acid. Fugitive emulsifiers like ammonia are excellent for use in coatings, since they evaporate after application and thus do not affect the coating properties. Other surface-active agents can be utilized as well to decrease the interfacial surface tension and thereby facilitate emulsion formation. Hydrophilic – lipophilic balance (HLB) values provide the simplest guide to preparing an oil-in-water or water-in-oil emulsion, since these reflect the hydrophobicity of the surfactant molecules (→ Emulsions, Chap. 2.8., → Emulsions). Extensive literature is available from ICI Americas on the use in this context of the polysorbate esters known commercially as Spans and Tweens. The more hydrophilic the system, the more likely it is to form water-in-oil emulsions (e.g., HLB 4 – 6). In a complex emulsion, stabilization requires a combination system such as that formed in situ between morpholine stearate and a Tween/Span mixture.

Dye Systems. Shoe creams must provide excellent scuff coverage, since this is a primary consideration from the consumer standpoint. This effect is achieved with a combination of

oil-soluble and acid dyes. Acid dyes display particularly effective staining characteristics. In some systems, pigment dispersions are introduced concomitantly due to acid-dye instability with respect to the alkaline materials required to form many emulsions.

Formulae. Table 4 provides examples of emulsion shoe-cream formulations. Such formulations result in an oil-in-water emulsion in which the emulsifier, morpholine stearate, is formed in situ. The use of turpentine as solvent permits incorporation of an excellent wax blend, providing a cream with moderate gloss and excellent staining characteristics. The actual concentration of colorant depends upon the desired color. Carnauba, Wax E, and Wax KSL form especially stable emulsions. Emulsions can also be formed with a Tween/Span combination or a single-component surfactant if so desired, although residual surfactant slightly reduces water-resistance properties.

Table 4. Representative shoe-cream formulations

Component	Relative content, wt %			
Carnauba wax	4.0		7.5	3.0
Montan wax (refined)	2.0		2.5	1.5
Wax E ∗		4.5	3.0	4.0
Wax KSL and/or LP ∗		3.5		2.5
Paraffin wax	8.0	6.0	5.0	7.0
Morpholine	1.5	1.5	1.5	1.5
Stearin	0.5	0.5	0.5	0.5
Turpentine	32.0	32.0	30.0	30.0
Water	50.0	50.0	48.0	48.0
Color mixture	2.0	2.0	2.0	2.0

∗ American Hoechst Corp.

4.2. Production

One steam- and water-jacketed vessel is required to heat the wax phase and form an emulsion at ca. 85 °C. After the wax phase has reached this temperature the solvent should be added, followed by the water phase while mixing. Then the pigment – dye dispersions are added to the emulsion kettle and the mixture is agitated for 15 min to ensure homogeneity. Finally, morpholine is incorporated to produce a stable emulsion. Mixing with an appropriate propeller mixer should result in an emulsion within 5 – 10 min. Cold or ambient water is then introduced into the jacket to cool the shoe cream. Alternatively, it may be allowed to cool passively. The shoe cream can be poured into jars at 45 – 52 °C; alternatively, it can be cooled to ambient temperature and then dispensed as a cream into tubes or glass jars.

4.3. Quality Specifications

An acceptable shoe cream should have a smooth, creamy appearance in the jar or tube, matching the color represented on the label. Emulsion stability is critical, as evidenced by the absence of free solvent or water in the container. Emulsion stability can easily be tested either by accelerated high-temperature storage (100 – 125 °C) or successive freeze – thaw cycles to establish emulsion integrity. Poorly formulated creams will fail such a test, leading to distinct phase separations. Moreover, dye systems require careful material selection to prevent dye interactions that would lead to color drift.

5. Liquid Polishes

Liquid (or "dry-bright") shoe polishes have been used for many years as a way of providing a quick, convenient shine with high pigment concentrations capable of hiding scuffs (e.g., on white nurses' shoes and children's footwear). A thin, flexible, polymeric coating is laid down on the shoe surface in such a way that it requires no buffing. Generally, these formulations contain three principal functional components: a polymeric binder, a wax emulsion, and a pigment dispersion. Drying may entail a combination of solvent evaporation and polymer cross-linking.

5.1. Composition

Polymeric Binder. Selection of the correct polymer is critical to formulating an acceptable dry-bright liquid shoe dressing. The polymer must be capable of binding the heavy pigment load, and it must also provide gloss, adhere to the substrate, and form a flexible film at room temperature. To produce an acceptable film coating the polymer must have a glass-transition temperature well below ambient temperature.

Styrene – acrylic copolymers, acrylate copolymers, and polyurethane polymers have all been used successfully in liquid shoe polishes. Styrene – acrylic copolymers are especially common in dry-bright emulsions because the styrene ensures good gloss properties while the acrylate contributes flexibility. Polyurethanes, while providing excellent gloss and flexibility, are not favored because they are difficult to remove from the leather substrate.

Pigment Dispersion. Most dry-brights are formulated with a high pigment concentration (5 – 30 wt %) for the purpose of covering the leather surface with a film of dispersed pigments. White shoe polishes require an extremely high pigment concentration to provide thorough coverage. Common pigments include yellows and oranges based on lead chromate, browns from iron oxides, and blues of synthetic organic origin (e.g., Acid Blue #9). White shoe polish always contains titanium dioxide (TiO_2) (→ Pigments, Inorganic, Chap. 2.1.). Anatase-grade TiO_2 was formerly used exclusively due to its favorable particle size and its slightly bluish tint, but the more modern chloride manufacturing process leads to an acceptable rutile-grade TiO_2 with excellent coverage.

Formulae. The core formula for a dry-bright shoe polish consists of a polyethylene wax emulsion, polymer, and colorant. Adhesion is provided by a *resin solution* (e.g., an aqueous ammoniacal solution of rosin-modified maleic resin), sometimes supplemented by a casein – borax solution, which significantly improves film adhesion. The resin solution plays an important role as a leveling agent. In its absence the polymer may pool on the surface. The resin helps fill the holes between wax – polymer molecules, thereby contributing to a smooth film. Crazing or cracking results from the brittleness inherent in most hard, high-gloss polymers. To prevent crazing, *plasticizers* may also be incorporated [e.g., tributoxyethyl phosphate (KP-140) or dibutyl phthalate]. Finally, a *glycol* is usually added to provide freeze – thaw stability by depressing the freezing point.

The following are simple examples of formulae for a dry-bright shoe polish and a corresponding wax emulsion (figures denote weight percentage):

Dry-bright shoe-polish:	
Wax emulsion	15
Polymer dispersion (40 % solids)	25
Resin solution (15 % solids)	18
Tributoxyethyl phosphate	2
Ethylene glycol	2
Casein solution (15 % solids)	10
Pigment dispersion	10
Morpholine	2
Water	16
Wax emulsion:	
Polyethylene	15
Oleic acid	2
Morpholine	2
Water	81

5.2. Production

Mild heating (45 °C) facilitates manufacture of a dry-bright emulsion using simple propeller mixing assuming the pigments are supplied in predispersed form. In the case of titanium dioxide, high-speed dispersion equipment is required. Generally, the resins and casein solution, as well as the polymer, are added first, and these are emulsified by the addition of morpholine or some other base. Polyethylene wax emulsion, plasticizer, etc., are then added, followed by the pigment dispersion. After all components have been combined the batch is mixed for 15 – 30 min prior to filling into bottles.

5.3. Quality Specifications

Typically, a dry-bright emulsion is tested for pH, percent solids, emulsion stability, and pigment reincorporation. Liquid shoe-polish efficacy is tested with respect to spectral reflectance (gloss meter), leveling, and pigment dispersion. The two latter analyses are conducted simply by pouring the product onto glass or an appropriate substrate and examining the results visually for "fish eyes" indicative of poor leveling or unincorporated pigment particles.

6. Other Shoe Dressings

Weatherproofing Liquids and Aerosols. Fluoropolymers such as FC-905 (Minnesota Mining and Manufacturing) are utilized to provide fine dress shoes with water and oil repellency. Titanate- curing polysiloxanes such as

Dow-Corning 477 are utilized for heavy-duty water repellency on work boots. These compounds provide an impervious coating upon drying; moisture from the air subsequently serves to cross-link the silicone polymer. Liquid products of this type are available, but aerosols are most popular because of their ease of application.

Leather Dyes. Solvent-based leather dyes are manufactured for professional and military applications. These products consist of black and brown solvent dyes in penetrating solvents such as toluene or xylene.

Silicone-Impregnated Sponges. Silicone polymers, both as emulsions or directly in the form of dimethyl silicone fluid of the appropriate viscosity, have been incorporated into sponges for the purpose of creating a quick temporary shine without any accompanying scuff restoration.

7. Environmental Protection

The heavy utilization of solvents in pastes, shoe creams, and weatherproofers is an industrial concern that offers the technical challenge of reducing the solvent levels without compromising the quality or efficacy of the products. Most major manufacturers are phasing out the use of ozone-depleting chemicals such as 1,1,1-trichloroethane and methylene chloride. There remains a heavy dependence on solvents for efficacious products, but on balance the extent of volatile organic compounds released into the atmosphere from this source is negligible in comparison with automobile emissions and other products and industries.

From a regulatory perspective, the toxicity profile of shoe-care products reflects the content of aromatic solvents. However, the infrequent use of such products, coupled with small package sizes (e.g., 30 – 60 g net content), mitigates any health risk. Dyes and pigments with severe carcinogenic profiles have been eliminated from most products.

Shortenings → Margarine and Shortenings
Silanes → Silicon Compounds, Inorganic

8. Economic Aspects

Major worldwide manufacturers of shoe dressings include Kiwi Brands, Tana, S. C. Johnson, and Reckitt & Colman. Worldwide production in each major product category is estimated as follows:

Paste polish	1.75×10^8 units
Shoe creams	1.90×10^8 units
Liquid polishes	2.25×10^8 units

9. References

General References

1. E. N. Abrahart: *Dyes and Their Intermediates,* Chemical Publishing Company, Inc., New York 1977.
2. H. Bennett: *Industrial Waxes,* **Natural and Synthetic Waxes, vol. 1,** Chemical Publishing Company, Inc., New York 1963.
3. H. Bennett: *Industrial Waxes,* **Compounded Waxes and Technology, vol. 2,** Chemical Publishing Company, Inc., New York 1975.
4. H. Bennett: *The Chemical Formulary,* Chemical Publishing Company, Inc., New York 1971.
5. J. Brandup, E. H. Immergut: *Polymer Handbook,* 3rd ed., J. Wiley & Sons, New York 1989.
6. A. Davidson, B. M. Milwidsky: *Polishes,* C.R.C. Press, Cleveland, Ohio, 1968.
7. F. O'Flaherty, W. T. Roddy, R. M. Lollar: "Preparation for Tannage," *The Chemistry and Technology of Leather,* **vol. 1,** Robert E. Krieger Publishing Company, Malabar, Fla., 1977.
8. F. O'Flaherty, W. T. Roddy, R. M. Lollar: "Types of Tannage," *The Chemistry and Technology of Leather,* **vol. 2,** Robert E. Krieger Publishing Company, Malabar, Fla., 1977.
9. J. H. Sharphouse, *Leather Technician's Handbook,* Leather Producers' Association, London 1971.
10. *Hoechst Waxes – Production, Properties, Application, Guidelines for Wax Users (American Hoechst Corp. trade literature).*

Silica

Otto W. Flörke, Institut für Geologie, Mineralogie und Geophysik, Ruhr-Universität, Bochum, Federal Republic of Germany (Chap. 1 and 2)

Heribert A. Graetsch, Institut für Geologie, Mineralogie und Geophysik, Ruhr-Universität, Bochum, Germany (Chap. 1 and 2)

Fred Brunk, Dr. C. Otto Feuerfest Gmbh, Bochum, Federal Republic of Germany (Chap. 1.7.2)

Leopold Benda, Niedersächsisches Landesamt für Bodenforschung, Hannover, Federal Republic of Germany (Chap. 3)

Siegfried Paschen, Kieselgur- und Kalksandsteinindustrie, Henrich Meyer-Werke Breloh GmbH & Co. KG, Munster, Federal Republic of Germany (Chap. 3)

Horacio E. Bergna, E. I. du Pont de Nemours & Company Inc., Wilmington, Delaware 19 880–0262, United States (Chap. 4)

William O. Roberts, E. I. du Pont de Nemours & Company Inc., Wilmington, Delaware 19 880–0262, United States (Chap. 4)

William A. Welsh, Grace Davison, W. R. Grace & Co.-Conn., Columbia, Maryland, United States (Chap. 5)

David M. Chapman, Grace Davison, W. R. Grace & Co.-Conn., Columbia, Maryland, United States (Chap. 5)

Manfred Ettlinger, Degussa-Hüls AG, Standort Wolfgang, Hanau, Federal Republic of Germany (Chap. 6)

Dieter Kerner, Degussa-Hüls AG, Standort Wolfgang, Hanau, Federal Republic of Germany (Chap. 7)

Monika Maier, Degussa-Hüls AG, Standort Wolfgang, Hanau, Federal Republic of Germany (Chap. 7)

Walter Meon, Degussa-Hüls AG, Standort Wolfgang, Hanau, Federal Republic of Germany (Chap. 7)

Ralf Schmoll, Degussa-Hüls AG, Werk Weeseling, Wesseling, Federal Republic of Germany (Chap. 7)

Hermann Gies, Ruhr-Universität, Bochum, Federal Republic of Germany (Chap. 8)

Dietmar Schiffmann, Institut für Toxikologie, Universität Würzburg, Würzburg, Federal Republic of Germany (Chap. 9)

1. Silica Modifications and Products . 2
1.1. Occurrence and Geochemistry . . . 2
1.2. Si–O Bond and Crystal Structures 4
1.3. Crystalline Silica Phases 5
1.3.1. Crystalline Silica Minerals 5
1.3.2. Crystalline Nonmineral Silica Phases 12
1.4. Noncrystalline Silica Minerals . . . 13
1.5. Colored Silica Minerals 14
1.6. Silica Rocks 15
1.7. Crystalline Silica Products 15
1.7.1. Cultured Quartz Single Crystals 16
1.7.2. Polycrystalline Silica Products 20
1.8. Noncrystalline Silica Products . . . 21
2. Quartz Raw Materials 22
2.1. Physical Forms and Occurrence . . 22
2.2. Processing 23
2.3. Uses 25
3. Diatomites 28
3.1. Introduction 28
3.2. Formation, Composition, and Quality Criteria 29
3.3. Occurrence and Mining 30
3.4. Processing 31
3.5. Analysis 33
3.6. Storage and Transport 33
3.7. Environmental and Health Protection 33
3.8. Uses 34
3.9. Recycling 35
4. Colloidal Silica 35
4.1. Introduction 35
4.2. Structure of Colloidal Silica Particles 37
4.3. Physical and Chemical Properties . 42
4.4. Stability 44
4.5. Production 46

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_583.pub2

4.6. Analysis and Characterization . . . 48
4.7. Uses 49
4.8. Storage, Handling, and Transportation 50
4.9. Economic Aspects 50
5. Silica Gel 51
5.1. Introduction 51
5.2. Structure, Properties, and Characterization 52
5.3. Production 55
5.4. Uses 56
5.5. Economic Aspects 58
5.6. Legal Aspects 58
6. Pyrogenic Silica 58
6.1. Flame Hydrolysis 58
6.1.1. Production Process 59
6.1.2. Morphology 59
6.1.3. Solid-State Properties 61
6.1.4. Surface Chemistry 62
6.1.5. Surface Modification with Silicon Compounds 63
6.1.6. Characterization 63
6.1.7. Uses 64
6.1.8. Industrial Hygiene and Safety 65
6.2. Electric-Arc Process 65
6.3. Plasma Process 65
7. Precipitated Silicas 66
7.1. Introduction 66
7.2. Production 66
7.3. Properties 67
7.3.1. Physicochemical Properties 67
7.3.2. Surface Chemistry and Surface Modification 68
7.3.3. Chemical Composition and Analysis . 69
7.4. Uses 69
7.5. Industrial Hygiene and Safety 71
8. Porosils 71
8.1. Introduction 71
8.2. Physical and Chemical Properties . 73
8.2.1. Zeosils 74
8.2.2. Clathrasils 75
8.3. Manufacture of Porosils 75
8.3.1. Synthesis of Porosils 75
8.3.2. Dealumination of Aluminosilicate Zeolites 75
8.3.3. Formation of Melanophlogite 76
9. Toxicology 76
9.1. Experiences with Humans 76
9.2. Animal Experiments 77
10. References 79

Silica (SiO_2) is a commonly occurring compound in nature. Various crystalline and noncrystalline silica minerals of inorganic and biogenic origin are known. Depending on temperature, pressure, composition of the precursor phases, etc., several different crystalline modifications are formed. Quartz is the most abundant and well-known polymorph. Silica occurs as meter-sized quartz crystals or as amorphous masses made up of submicroscopic crystallites having different microstructures or growth fabrics. In addition, synthetic polymorphs have been produced for which no natural counterparts exist. In industry, quartz sand and silica rocks are widely used as raw materials for production of glass, ceramics, and silicon. Single crystals of quartz are grown artificially for application as resonator crystals in clocks, communication technology, and computer electronics. Both natural and synthetic silica powders are used as fillers to improve the mechanical properties of plastics.

1. Silica Modifications and Products

1.1. Occurrence and Geochemistry

Silica [*7631-86-9*], silicon dioxide, SiO_2, is the major constituent of rock-forming minerals in magmatic and metamorphic rocks. It is also an important component of sediments and soils [1 – 3]. Bound as silicates it accounts for ca. 75 wt % of the Earth's crust [4]. Free silica predominantly occurs as quartz, which makes up 12–14 wt % of the lithosphere.

Quartz is the thermodynamically stable modification at ambient conditions (Fig. 1) occurring as one of the main products of slowly cooled silica-rich magmas such as granites, granodiorites, and related rocks. It is also a primary mineral of pegmatites and hydrothermal veins formed during late stages of magmatic processes. In metamorphic rocks like schists, migmatites, and quartzites which suffered high temperature and pressure due to deep burial in the Earth's crust, quartz is also a major constituent. Owing to its hardness and low solubility, quartz is able to resist erosion and occurs abundantly in sedimentary rocks such as sand

and sandstones. Quartz is also the final product of the diagenesis process of noncrystalline biogenic silica deposited in sea floor sediments. In the form of chalcedony, microcrystalline quartz occurs as secondary filling of veins and cavities in rocks and in petrified wood and other organic matter. Flint nodules, frequently occurring in limestone formations, also consist of chalcedony.

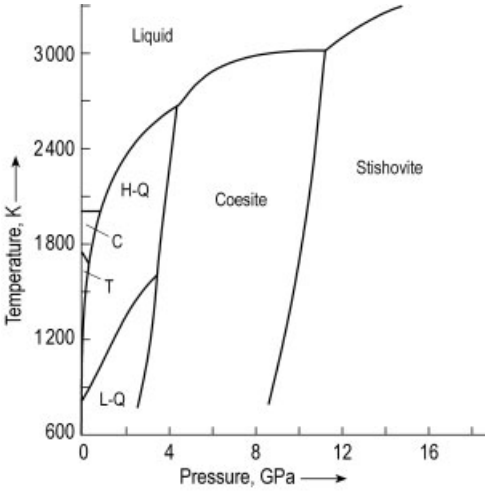

Figure 1. Calculated phase diagram of the SiO_2 system [6]
H-Q: high quartz, L-Q: low quartz, C: high cristobalite, T: high tridymite (for the stability field of tridymite, see text).

Quartz owes its abundance to the action of water in rock-forming processes. In rocks that were formed at high temperatures under dry conditions – volcanic eruptiva, silicate meteorites, moon rocks and silica slags, for example – quartz is rare. Here the high-temperature silica modifications cristobalite and tridymite crystallize (Fig. 1). Under the influence of high pressure induced by burial in great depth or by meteoritic impact, the high-pressure modifications coesite and stishovite are stable.

Under more ambient conditions waterborne unstable silica varieties such as micro- and noncrystalline opals and chalcedony are formed. Deep-sea sediments consisting of large masses of biogenic silica, for example, are altered in a diagenetic process with increasing burial depth from noncrystalline opals via microcrystalline opals to chalcedony and chert. Micro- and noncrystalline silica minerals are described in more detail in [5].

Keatite, silica W, and porosils (except for melanophlogite) are artificial crystalline silica products that do not occur in nature. They do not have a stability field of their own and can be synthesized under hydrothermal conditions only in the presence of sodium (keatite) or with the aid of organic template molecules (porosils). The latter are described in more detail in Chapter 8.

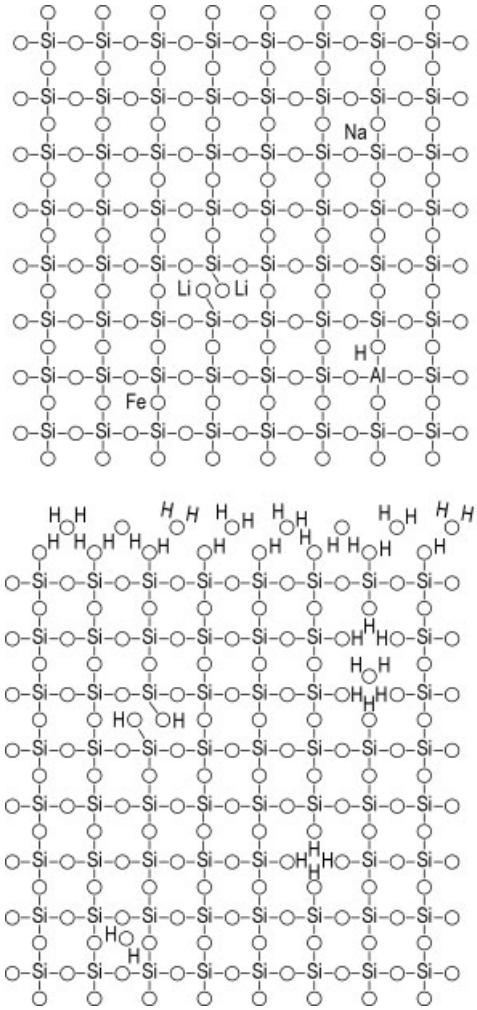

Figure 2. Two-dimensional scheme of incorporation of impurities in silica
Top: chemical point defects introduced by occupation of silicon lattice sites and of interstitial sites by foreign atoms.
Bottom: water incorporated in the silica framework by forming hydrolytic nonbridging oxygen (German: *Trennstellen*) and by occupation of silicon vacancies or interstitial sites, and water located at the surface as silanol groups and adsorbed water molecules.

Silica crystals are almost pure SiO_2 which do not contain large quantities of impurities. Water, however, can be incorporated in concentrations from hundreds to several thousands of parts per million in quartz (Fig. 2). The transition from structural incorporation to microstructural inclusion is fluent [7]. At elevated temperatures and pressures water migrates rapidly through the structural framework by splitting Si−O−Si bonds and forming silanol groups. This mechanism of hydrolytic weakening activates plastic deformation, solid-state flow [8], reconstructive transformations, and recrystallization. Traces of alkali metal ions act as efficient transformation activators (Fig. 2) [9]. Alkali metal ions as well as Al and Fe are the most common nonvolatile impurities in silica. Quartz usually contains about 100 ppm of trace elements or less [5, 10, 11]. The concentration of impurities can be higher in tridymite and cristobalite than in quartz, ranging up to 5000 ppm Al, 3500 ppm Na and 1500 ppm K [12, 13]. The trace elements can be incorporated in at least two different ways (Fig. 2) [14]. The first is the substitution of Si^{4+} by Al^{3+}, or Fe^{3+} on lattice sites and simultaneous addition of Li^+, Na^+, K^+, Mg^{2+}, or Ca^{2+} on interstitial sites. The second is the incorporation of additional alkali metal ions to form alkaline nonbridging oxygen sites (German: *Trennstellen*).

1.2. Si−O Bond and Crystal Structures

Despite its chemical simplicity, SiO_2 displays a remarkable diversity of crystal structures. Apart from the ultrahigh-pressure modification stishovite with octahedrally coordinated silicon (Fig. 3) and artificial silica W with chains of edge-sharing tetrahedra, all crystal structures of silica are made up of a three-dimensional framework of corner-sharing tetrahedra (Fig. 3). The noncrystalline silica phases consist of a continuous random network of corner-sharing tetrahedra. All four sp^3 orbitals of silicon overlap with 2p orbitals of the four coordinating oxygen atoms to form strong σ bonds [15]. Additional minor contributions from π bonding may arise from overlap of Si 3d orbitals with other O 2p orbitals, which partly explains the observed shortening of Si−O bond lengths with increasing Si-O-Si bond angles [16]. The mixed bond character of about 50 % ionic and 50 % covalent results in bent intertetrahedral Si-O-Si angles in the range of 120 to 180° with a mean of about 147° [17 – 19]. Band theory treatment is given in [20]. The effective ionic radius of Si^{4+} is 0.026 nm for tetrahedral coordination by oxygen (O^{2-}: r = 0.135 nm), and 0.040 nm for octahedral coordination [21]. The mean Si−O bond length is 0.162 nm in the tetrahedra [22] and the mean O−O distance is 0.264 nm (Fig. 3). The short Si−O bond length is responsible for the high bond strength.

The strong bonds and the three-dimensional connectivity of the tetrahedra are the reasons for the high hardness (quartz: Mohs hardness 7), the lack of good cleavage, high elasticity, high melting point (cristobalite ca. 2000 K), high activation temperature of the quartz–cristobalite transformation (1300 K) and high transformation temperature of silica glass (1300 K). Contrasting with these properties, the resistivity of silica to electron irradiation damage is relatively low [23].

Silica is an insulator, a property that is utilized in the fabrication of silicon-based microelectronic devices by precipitation of ultrathin insulating films of noncrystalline silica on the semiconductor substrate crystals.

The tetrahedra can be considered as rigid units which remain almost unchanged upon thermal expansion or compression by high pressure. They are rotated or tilted instead. The variability of the intertetrahedral Si-O-Si angles and the unrestricted torsion angle of connected tetrahedra (Fig. 3) also account for the topological diversity of the crystal structures of silica and the tendency for glass formation. At elevated temperatures quartz, cristobalite, tridymite, and melanophlogite displacively transform into dynamically disordered high-temperature phases. Softening of rigid unit modes plays an important role in these thermally induced ferroic transitions [24]. In the variable framework structures, planar defects such as stacking faults in cristobalite and tridymite, and twin planes in quartz and moganite frequently occur. Growth twinning and transformation twinning are also common. Normally, neighboring tetrahedra adopt the electrostatically favorable trans configuration (Fig. 3). Rare exceptions with cis and trans configurations are tridymite [9], melanophlogite, and coesite [26].

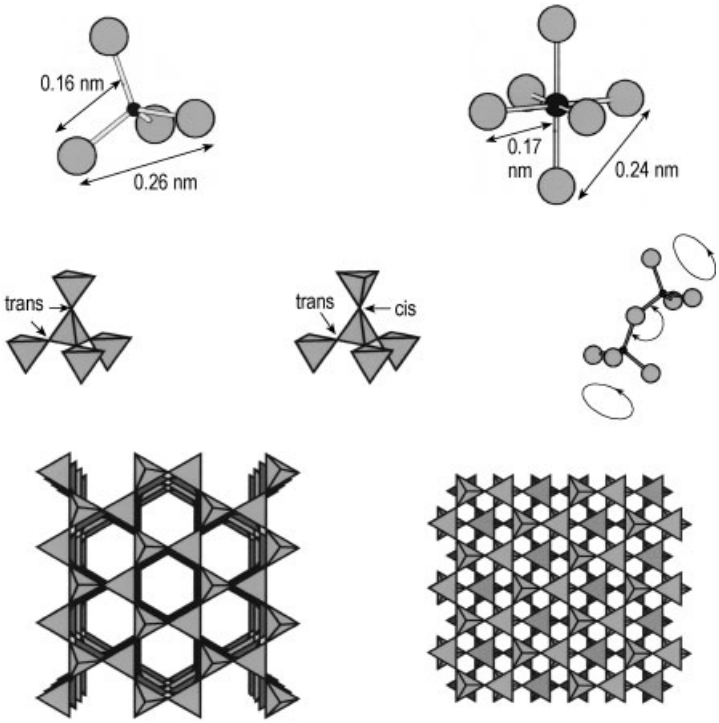

Figure 3. Top: $SiO_{4/2}$ tetrahedron and $SiO_{6/3}$ octahedron.
Middle: tetrahedra configurations in the silica framework structures.
Bottom: stacking principles of idealized six-membered rings of tetrahedra: two-layer sequence in tridymite (left), three-layer sequence in cristobalite (right).

1.3. Crystalline Silica Phases

The crystalline silica minerals are listed in Tables 1, 2, and 3 together with their crystallographic data. Table 1 also lists the nonmineral crystalline phases keatite and silica W. The porosils are treated in Chapter 8.

1.3.1. Crystalline Silica Minerals

The *p–T* diagram of silica, with stishovite, coesite, quartz, and cristobalite, is shown in Figure 8 [6]. Tridymite forms only in the presence of impurities which act as mineralizers. The phase transformations between the stable phases are reconstructive, requiring breaking and formation of new Si−O bonds (except for the high quartz–low quartz transition). Apart from high quartz, all high-temperature and high-pressure phases can persist metastably under ambient conditions. Quartz and most of the phases that are metastable under ambient conditions undergo displacive transitions at elevated temperatures without breaking of bonds. Melanophlogite, moganite, and the microcrystalline mineral species have no thermodynamic stability field.

Quartz is the thermodynamically stable form under ambient conditions. The tetrahedra are arranged in the open structure of low or α-quartz as helices along the trigonal *c*-axis. Two types can be distinguished: a small helix formed by three tetrahedra and a larger double helix consisting of six tetrahedra. All threefold screw axes are either left- or right-handed, and this results in the optical activity of quartz with a rotatory power of 27.71°/nm for Na light [10]. The crystals are enantiomorphous with the symmetry of the acentric space groups $P3_121$ or $P3_221$. The right- and left-handed crystals are mirror images of each other (Fig. 4). The twofold symmetry axes perpendicular to the screw axis are polar, and mechanical stress along the axis produces a direct piezoelectric effect. The reverse effect is

Table 1. Crystalline silica phases: space groups and unit cell parameters

Phase	Space group	Unit cell dimensions in nm					
		a	b	c	α	β	γ
Mineral phases							
Stishovite	$P4_2/mnm$	0.418	0.418	0.267	90 °	90 °	90 °
Coesite	$C2/c$	0.713	1.240	0.717	90 °	120.3 °	90 °
Low quartz	$P3_221/P3_121$ [a]	0.491	0.491	0.540	90 °	90 °	120 °
High-quartz [b]	$P6_222/P6_422$ [a]	0.500	0.500	0.546	90 °	90 °	120 °
Moganite	$I12/a1$	0.876	0.488	1.072	90 °	90.1 °	90 °
Low cristobalite	$P4_12_12/P4_32_12$ [a]	0.497	0.497	0.693	90 °	90 °	90 °
High cristobalite [c]	$Fd\bar{3}m$ [d]	0.713	0.713	0.713	90 °	90 °	90 °
Low tridymite [e]	Cc [f]	1.849	0.499	2.583	90 °	117.8 °	90 °
L1–T_0 (MC)	Cc [g]	1.852	0.500	2.381	90 °	105.8 °	90 °
Low tridymite [h]	$F1$	0.993	1.722	8.186	90 °	90 °	90 °
L2–T_D ($PO_{5/10}$)							
Low tridymite	Cc [f,i]	0.501	0.860	0.822	90 °	91.5 °	90 °
L3–T_0 (MX-1)							
High tridymite [j]	$P2_12_12_1$	0.499	2.617	0.820	90 °	90 °	90 °
H4–T_0 (OP)							
High tridymite [k]	$C112_1$ ($\alpha\beta0$) [i,l]	0.501	0.875	0.821	90 °	90 °	90.3 °
H3–T_0 (OS)							
High tridymite [m]	$C222_1$	0.502	0.876	0.821	90 °	90 °	90 °
H2–T_0 (OC)							
High tridymite [n]	$P6_3/mmc$ [f]	0.505	0.505	0.826	90 °	90 °	120 °
H1–T_0 (HP)							
Melanophlogite	$Pm3n$	1.344	1.344	1.344	90 °	90 °	90 °
Nonmineral phases							
Keatite	$P4_12_12/P4_32_12$ [a]	0.746	0.746	0.858	90 °	90 °	90 °
Silica W	$Icma$	0.836	0.472	0.516	90 °	90 °	90 °

[a] Enantiomorphic (right- and left-handed);
[b] At 863 K;
[c] At 573 K;
[d] Average structure;
[e] For the nomenclature of the tridymite phases, see Fig. 8;
[f] Setting according to [53, 62];
[g] Setting according to [54, 63];
[h] Stacking disordered;
[i] Incommensurately modulated;
[j] At 428 K;
[k] At 443 K;
[l] Four-dimensional super space group;
[m] At 493 K;
[n] At 733 K.

utilized in oscillatory devices. Standard values of important properties are compiled in [27], and the relationships between crystal structure and properties are given in [28]. The low specific free surface energy of the rhombohedral faces $\{10\bar{1}1\}$ results in a poor but distinct cleavage.

The thermal expansion of α-quartz is high ($\alpha_{11} = 13.3 \times 10^{-6}$ K^{-1}, $\alpha_{33} = 7.1 \times 10^{-6}$ K^{-1}) and drops to even slightly negative values at the displacive phase transition to high or β-quartz at 573 °C (Fig. 5). With increasing temperature the tetrahedra are rotated clockwise or anticlockwise around the dyad axes towards their higher symmetry positions in β-quartz. Additional dyad axes coinciding with the threefold screw axes in the double helices are generated at the transition, turning them into sixfold screw axes. The handedness of the helices is preserved, and the space groups for left- and right-handed β-quartz are $P6_222$ and $P6_422$ with additional twofold rotation axes perpendicular to the hexagonal c-axis (Fig. 4). The crystal structure of β-quartz is disordered by intense thermal vibrations of the tetrahedra, and the structure shown in Figure 4 is only an average. An intermediate phase occurs in a narrow temperature range of 1.7 K between α- and β-quartz. It is incommensurately modulated and consists of a regular array of microdomains with the orientation of twins according to the

Table 2. Crystalline silica phases: crystal dimensions, density, and displacive phase transitions

Phase	Crystal dimension [a]	Density g/cm^3	Displacive phase transitions in K
Mineral phases			
Stishovite	2	4.29	–
Coesite	2	2.92	–
Low quartz	1,2,3	2.65	846
High-quartz	4	2.53	846
Moganite	3	2.62	–
Low cristobalite	1,2	2.33	540 [b]
High cristobalite	4	2.21	540 [b]
Low tridymite L1–T_0 (MC)	1,2	2.27	398, 423, 493, 723
Low tridymite L2–T_D ($PO_{5/10}$)	1,2	2.28	398, 423, 493, 723
Low tridymite L3–T_0 (MX-1)	1,2	2.26	338, 398, 423, 493, 723
High tridymite H4–T_0 (OP)	4	2.24	398, 423, 493, 723
High tridymite H3–T_0 (OS)	4	2.22	398, 423, 493, 723
High tridymite H2–T_0 (OC)	4	2.21	398, 423, 493, 723
High tridymite H1–T_0 (HP)	4	2.19	398, 423, 493, 723
Melanophlogite	1,2	ca. 2.0	ca. 333
Nonmineral phases			
Kealtte	2	ca. 2.5	–
Silica W	2 (fibers)	ca. 1.97	–

[a] 1: macroscopic, 2: microscopic, 3: submicroscopic, 4: does not exist at ambient conditions (not quenchable);
[b] Hysteresis of more than 20 K.

Dauphiné law. The size of the domains rapidly decreases with increasing temperature [29]. Heating untwinned single crystals of quartz to temperatures above the transition point usually results in the formation of Dauphiné twins [30].

Dauphiné twins are intergrowths of two right- or two left-handed quartz individuals which are rotated by 180° with respect to each other around their common *c*-axis. This additional dyad axis is identical with the twofold axis which is lost during the high–low transition. Dauphiné twinning is also referred to as electrical twinning. Twinning according to the Brazil law also occurs abundantly in natural quartz, i.e., in amethyst and chalcedony. In this case right- and left-handed individuals are intergrown via the $\{10\bar{1}1\}$ plane. The new symmetry element introduced by the twinning is a mirror plane parallel to $\{11\bar{2}0\}$. Brazil twinning is also called optical twinning. Both kinds of twinning neutralize or reduce the piezoelectric effect of quartz, and crystals suitable for oscillatory and optical purposes must be free from twinning. Twins according to the Japanese law are less abundant. The twin plane is $\{11\bar{2}2\}$ and the *c*-axes of two twinned crystals form an angle of 84°33′.

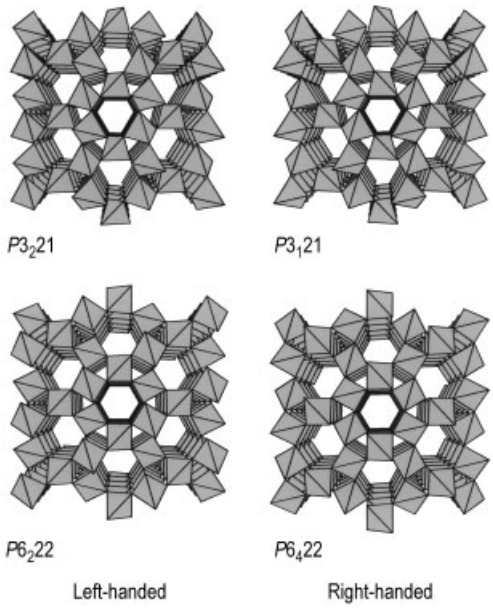

Figure 4. Tetrahedral model of the crystal structures of left- and right-handed quartz. Top: low-temperature quartz; bottom: high-temperature quartz. Perspective view on (001) slices along the *c*-axis; [100] direction vertical.

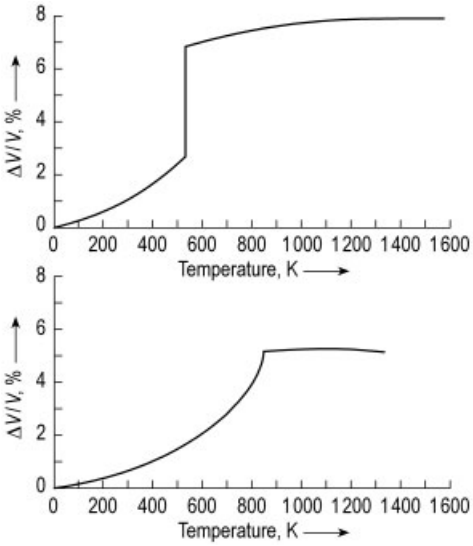

Figure 5. Thermal expansion of the unit cells of cristobalite (top) and quartz (bottom).

Other compounds crystallizing with the quartz structure type are phosphates MPO_4 and arsenates $MAsO_4$ (M = Al, Ga, Fe), BeF_2, the high-temperature form of GeO_2, and the high-pressure modifications of BPO_4, $BAsO_4$, and phosphorus oxynitride (PON) [31 – 33]. β-Eucryptite $LiAlSiO_4$ and γ-spodumene $LiAlSi_2O_6$-III are stuffed derivatives of β-quartz in which the silicon atoms are partly replaced by Al and Li incorporated on interstitial tetrahedral sites (I_4 position) within the large helices [34].

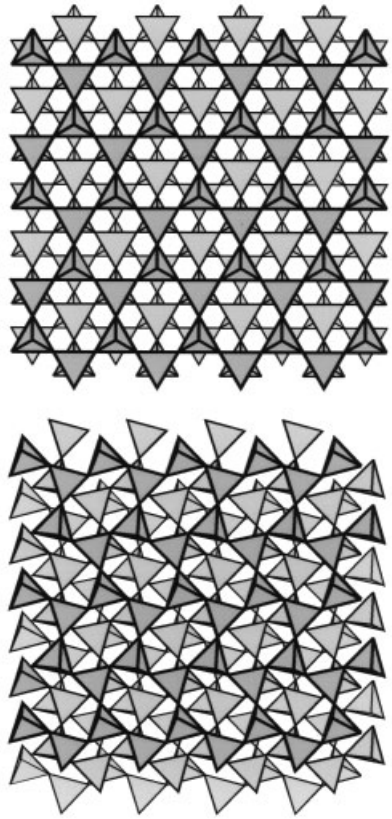

Figure 6. Tetrahedral model of the crystal structure of cristobalite.
Top: (111) slice of the cubic high-temperature form ([$\bar{1}$10] direction vertical).
Bottom: (101) slice of the tetragonal low-temperature modification ([010] direction vertical).

Cristobalite is the low-pressure, high-temperature modification of silica (Fig. 1). It persists as metastable phase at low temperatures. A displacive phase transition from the tetragonal low-temperature α-form to the cubic high-temperature β-form takes place near 540 K. The transition is of first order with an hysteresis of more than 20 K and a volume discontinuity of ca. 5 %. The crystal structure of β-cristobalite is a derivative of the diamond structure type in which carbon is replaced by silicon and oxygen is located midway between neighboring Si atoms. The framework can be described as a stacking of parallel layers of six-membered rings of tetrahedra. Alternate tetrahedra point upwards and downwards (Fig. 6). The stacking is a three-layer sequence in cristobalite.In tridymite, similar layers are stacked in a two-layer sequence with antiparallel mutual orientation. Stacking faults occur frequently in cristobalite and tridymite (cf. microcrystalline opals). The space group of the average structure of β-cristobalite is $Fd\bar{3}m$ [35]. The rigid tetrahedra perform rotational vibrations by precessing around the straight Si-O-Si lines [36 – 39]. At the transition temperature, these motions become unstable, and the structure collapses to form oval rings instead of the hexagonal six-membered rings of tetrahedra in the average high-temperature structure. In α-cristobalite the tetrahedra are arranged in helices containing either left- or right-handed fourfold screw axes parallel to the tetragonal c-axis [36]. Two enantiomorphous forms exist with the space groups

$P4_12_12$ or $P4_32_12$. Loss of the threefold rotation axis, the inversion center, and the centering of the unit cell during the phase transition result in the formation of 12 twins and antiphase domains [40]. The volume discontinuity at the transition leads to characteristic fish-scale microcracking. These scales are rectangular if the cristobalite is structurally ordered and are rounded in the presence of stacking disorder [41]. The displacive cubic-to-tetragonal transition cannot be suppressed by quenching, but stacking faults lower the transformation temperature [138]. The structure of β-cristobalite can be stabilized by way of chemical stuffing. In $Ca_{0.02}Al_{0.09}Si_{0.91}O_2$ some of the silicon atoms are replaced by aluminum, and calcium ions are incorporated on interstitial sites in the large cages between the layers of tetrahedra [34].

Other compounds with the α-cristobalite-type of structure are the high-temperature forms of $AlPO_4$, $GaPO_4$, $MnPO_4$, and $AlAsO_4$ [31, 32]. $BAsO_4$, BPO_4, $BeSO_4$ and phosphorus oxynitride (PON) crystallize with the β-cristobalite structure. Stuffed derivatives are carnegieite $NaAlSiO_4$ [42] and a variety of compounds with the formula Na_2MSiO_4 (M = Ca, Be, Zn, Mg) [34].

Tridymite forms between 1200 and 1800 K at ambient pressure, but traces of foreign ions, preferably alkali metals, or hydrothermal conditions [43] are necessary. The growth of tabular crystals always starts with cristobalite nuclei [44].

The crystal structure is a derivative of the hexagonal wurtzite type with Si occupying the Zn and S positions and oxygen placed between the silicon atoms. Layers with six-membered rings of tetrahedra are stacked in antiparallel orientation to give a two-layer sequence. Viewed along the hexagonal *c*-axis the layers are in the totally eclipsed position in high-temperature tridymite, leaving open channels in the structure (Figs. 3 and 7).

With respect to bonding requirements, tridymite is much less balanced than cristobalite. As a consequence tridymite undergoes a cascade of displacive phase transitions [9, 45] (Fig. 8). Similar to β-cristobalite and β-quartz, the hexagonal high-temperature phase of tridymite is dynamically disordered [46]. Upon cooling, different rigid unit modes of the tetrahedra successively freeze in by forming phases with col-

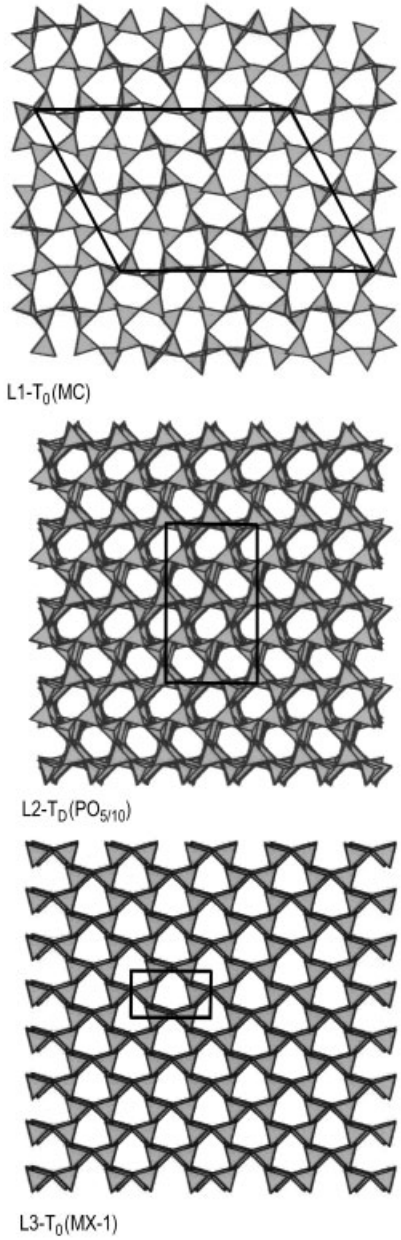

Figure 7. Tetrahedral model of the crystal structures of the three room temperature forms of tridymite.
Top: (010) slice of L1-T_O(MC) with vertical [305] direction.
Middle and bottom: (001) slices of L2-$T_D(PO_{5/10})$ and L3-T_O(MX-1) with [010] and [100] direction vertical, respectively.

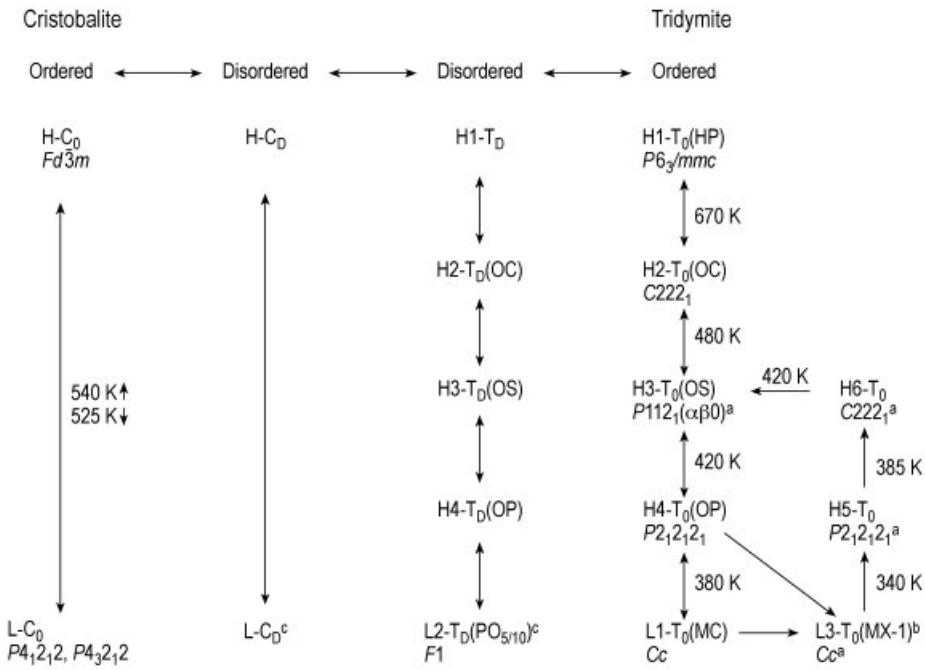

Figure 8. Displacive polymorphism and reconstructive stacking polytypism of cristobalite and tridymite.
[a] phases with incommensurate displacive modulations; [b] phase formed by thermal or mechanical stress (quenching or grinding); [c] transition temperatures depend on the degree of stacking disorder.
Nomenclature of the tridymite phases: L = low-temperature phase, H = high-temperature phase, T_O = ordered tridymite, T_D = stacking-disordered tridymite; in parentheses: M = monoclinic, O = orthorhombic, H = hexagonal, PO = pseudo-orthorhombic, C = C-centered, P = primitive Bravais type, S = superstructure, X = unknown.

lapsed rings of tetrahedra [47]. In a first step alternate layers of tetrahedra are laterally shifted with respect to each other near 400 °C, reducing the symmetry to orthorhombic [48]. In a second step an incommensurately modulated phase with temperature-dependent lattice parameter $\sqrt{3}a_{\mathrm{hex}}$ is formed at 220 °C [49, 50] which discontinuously locks in at ca. 150 °C resulting in a superstructure with a $3\sqrt{3}a_{\mathrm{hex}}$ lattice parameter [51, 52]. Finally, a monoclinic phase with a complex superstructure is formed at 116 °C in a first-order transformation [53, 54]. This modification is usually obtained by flux syntheses and has been found in moon rocks and meteorites. The transformation path, however, can be modified by thermal or mechanical stress or by lattice imperfections like stacking faults. Most tridymites occurring in terrestrial rocks contain stacking faults and have a triclinic (pseudo-orthorhombic) $2\times2\sqrt{3}\times10$ superlattice [55]. The number of high-temperature phases of pseudo-orthorhombic tridymite and the respective transition temperatures depend on the degree of stacking disorder [45]. A third room-temperature modification can be produced by grinding monoclinic tridymite or by quenching from high temperatures [56]. The phase is monoclinic and displays an incommensurate structural modulation [57 – 59]. Further modulated high-temperature phases can be obtained by reheating incommensurate low tridymite [60].

All three low tridymite modifications and cristobalite can be intergrown in the same crystal [60, 61]. Impurities can be incorporated by coupled substitution on lattice sites and on interstices in the open structural channels parallel to the *c*-axis [13]. In refractory-grade crystals, trace element concentration is 1 wt % [13].

A compound with isotypic phases is $AlPO_4$ [31]. Nepheline ($Na_3KAl_4Si_4O_{16}$) and kalsilite and kaliophilite (both $KAlSiO_4$) are stuffed derivatives of tridymite. Another stuffed derivative is $CaAl_2O_4$ [34].

Coesite is the high-pressure modification of silica with the densest framework of $SiO_{4/2}$ tetrahedra. It is composed of four membered rings of tetrahedra which are linked in the form of chains of rings. The chains are interconnected by symmetrically different four-membered rings of tetrahedra (Fig. 9). In this group the tetrahedra are in *cis* configuration with an exceptionally large Si-O-Si angle, restricted by symmetry to 180° [25, 26].

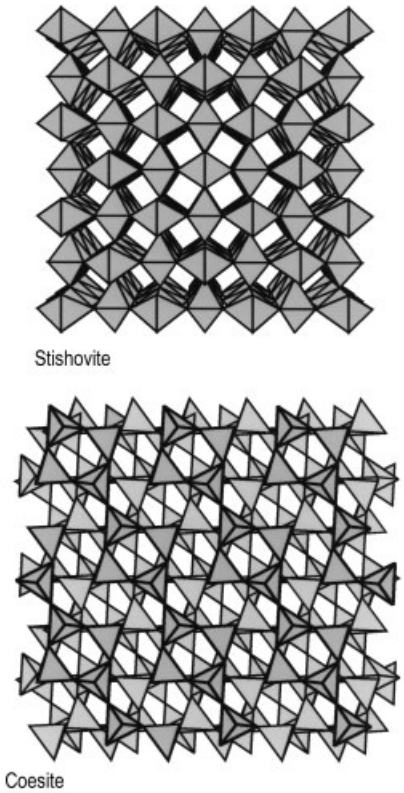

Figure 9. Polyhedral model of the high-pressure modifications coesite and stishovite.
Top: perspective view along the *c*-axis of stishovite ([110] direction vertical).
Bottom: (010) slice of coesite with vertical [101] direction.

Coesite metastably persists up to about 1300 K at ambient pressure. It is much less soluble in hydrofluoric acid than quartz but has a similar solubility to quartz in water under ambient conditions. Coesite and stishovite are rare and were synthesized prior to their detection in nature.

Stishovite crystallizes in the rutile structure type with silicon in sixfold coordination (Fig. 9). The SiO_6 octahedra form chains by sharing two opposite edges. Different chains are connected by linking of octahedra corners [64, 65]. The arrangement results in a very close packing of oxygen, even though the Si−O bonds are longer than in SiO_4 tetrahedra. Stishovite is about 43 % denser than coesite [66]. It is the stable form of silica under the conditions of the Earth's mantle. In the lower mantle stishovite probably transforms from the tetragonal rutile structure to a distorted structure with orthorhombic symmetry (stishovite II). A ferroelastic transformation to the denser $CaCl_2$ structure type is expected at higher pressures which involves tilting of the octahedra [66].

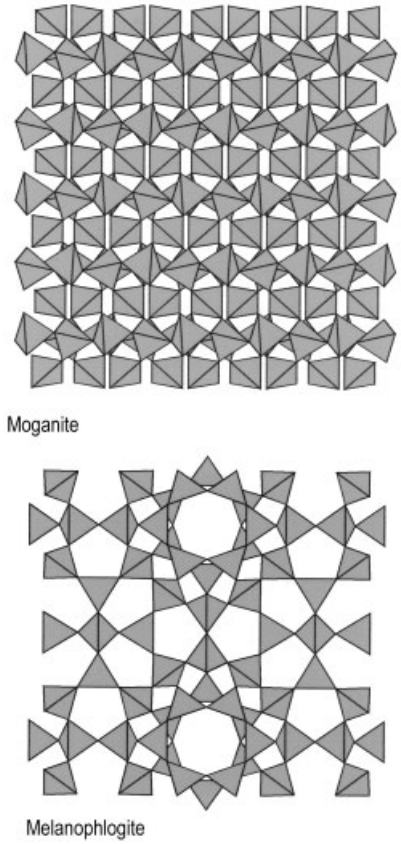

Figure 10. Tetrahedral models of the crystal structures of moganite and melanophlogite. (001) slices with vertical [010] direction.

Table 3. Microcrystalline silica minerals

Mineral or phase	Species *		Microstructure	Refractive index	Water content, wt %
Quartz	microquartz		granular grains $< 20\,\mu m$	1.54–1.55	< 0.3
	chalcedony		length fast fibers, fiber axis <11.0>		
		Ch-w_{LF}	parallel fibrous	1.53–1.54	0.8–2.0
		Ch-h_{LF}	spherulites of radiating fibers	ca. 1.54	0.3–0.8
	quartzine	Ch_{LS}	length slow fibers, fiber axis <00.1>	ca. 1.54	0.5–1.0
Moganite			crystal blades		1.5–3.0
Opal	opal-CT	O-CT_{LS}	length slow fibers (lussatite)	1.45–1.46	
		O-CT_M	lepidospheres of platelets	1.43–1.45	3.0–10.0
	opal-C	O-C_{LF}	parallel platy $\{101\}$	ca. 1.46	1.0–3.0
		O-C_M	massy, tangled platy		1.0–8.0

* Ch: chalcedony, w: wall lining, h: horizontally banded, LF: length fast (negative optical character of the fiber elongation), LS: length slow (positive optical character of the fiber elongation), C: disordered cristobalite, CT: disordered cristobalite/tridymite, M: massy tangled

Moganite is closely related to quartz but represents a new structure type with a lower symmetry. The structure can be described as a composition of alternate $\{10\bar{1}1\}$ layers of right- and left-handed quartz (Fig. 10), that is, a periodic twinning according to the Brazil law on the cell-dimension scale [67]. The only known occurrence of pure moganite is on Gran Canaria, Spain, where it forms microcrystalline fillings of fissures and cooling cracks in the ignimbrite streams [68, 69]. It has been reported that moganite also occurs as a minor constituent of microcrystalline quartz minerals [70].

The synthesis of a high-pressure, high-temperature modification of phosphorus oxynitride (PON) with moganite-type crystal structure has been reported [33].

Melanophlogite is a very rare mineral with the composition $46\,SiO_2 \cdot 6\,(CO_2,N_2) \cdot 2\,(CH_4,N_2)$. Agrigento and Livorno in Italy and Mt. Hamilton in California are the only known noteworthy occurrences [71, 72]. The guest molecules CH_4, CO_2, and N_2 are incorporated in two types of large cagelike voids of the silica framework (Fig. 10) [73]. A displacive phase transition from tetragonal to cubic symmetry takes place between 40 and 65 °C. Melanophlogite is the only mineral species among the large group of artificial microporous silica phases which are treated in Chapter 8.

Chalcedony and quartzine are microcrystalline quartz minerals (Table 3) [5] consisting of submicroscopic fibers. The fiber elongation occurs perpendicular to the trigonal *c*-axis in chalcedony and parallel to it in quartzine. The microcrystals are highly defective and contain a large number of twin planes (left- and right-handed quartz twinned according to the Brazil law) [74]. Chalcedony is the main constituent of banded agate geodes, flint nodules, and chert. Quartzine sporadically occurs as thin bands in agates. The purity is usually high with less than 1 wt % nonvolatile impurities and less than 2 wt % water. The latter is incorporated at defects of the crystal structure and in the microstructure between the fibers [75].

Microcrystalline opals consist of highly stacking disordered cristobalite and tridymite. A distinction is drawn between opal-CT and opal-C [76, 77]. In opal-CT the stacking is random tridymitic (two-layer sequence) and cristobalitic (three-layer sequence) [78], whereas cristobalite dominates in opal-C [77, 79]. The microstructure consists of fibers (lussatite) or of lepidospheric aggregates of thin crystal blades (Table 3). The water content can reach 10 wt % with a large portion of water molecules dissolved in the defective crystal lattice [80, 81].

1.3.2. Crystalline Nonmineral Silica Phases

Keatite, silica W, zeosils, and most of the clathrasils are synthetic crystalline forms of silica for which no counterpart has been found in nature. The porosils are described in Chapter 8.

Keatite has a tetragonal framework of corner-sharing tetrahedra which contains five-, six-, and seven-membered rings of tetrahedra

[82]. Structural channels run parallel to the *a*-axes, and helices with fourfold screw axes parallel to the *c*-axis (Fig. 11) [83]. Under mild hydrothermal conditions microcrystalline fibers of keatite are obtained by using silica glass as starting material and dilute NaOH solutions [84]. It is a metastable intermediate crystallization product. Keatite can also be prepared under dry conditions by heating lithic silica gels with chromium oxide as additive [85].

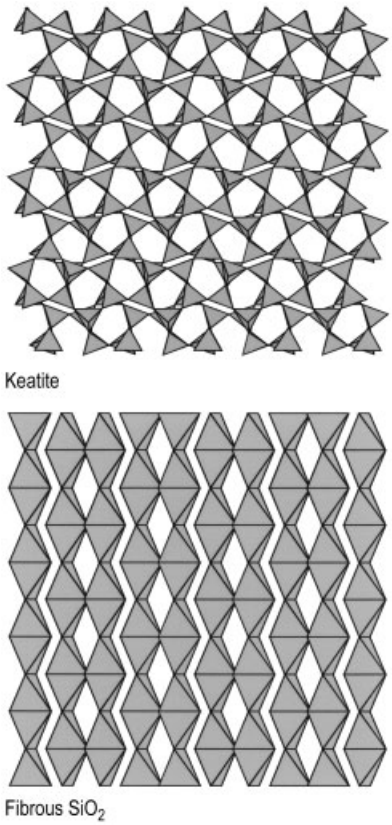

Figure 11. Tetrahedral model of the crystal structures of keatite and silica W.
Top: (100) slice of keatite with vertical [001] direction.
Bottom: (110) slice of silica W with vertical [001] direction.

Stuffed derivatives of keatite are β-spodumene $LiAlSi_2O_6$ and $HAlSi_2O_6$, which show ionic conductivity.

Silica W [86] forms above ca. 1500 K in the vapor transport reaction:

$$H_2\,(g) + SiO_2\,(s) \rightarrow SiO\,(g) + H_2O\,(g)$$

The crystal structure is isotypic with that of SiS_2. Infinite chains of edge-sharing tetrahedra run parallel to one another and are held together by weak interatomic forces (Fig. 11). At ambient temperature, this very unstable configuration collapses immediately on coming into contact with traces of water. Tempering under the formation conditions produces cristobalite [86].

SiO_2-X is an alkali metal silicate hydrate with the approximate formula $MSi_{11}O_{20.5} \cdot 3\,H_2O$ (M = K, Rb, Cs) [87].

1.4. Noncrystalline Silica Minerals

Noncrystalline (antiquated: amorphous) silica has a short-range order of corner-sharing SiO_4 tetrahedra like crystalline silica but no long-range order and consists of a continuous random network instead of an ordered framework structure. The intertetrahedral Si−O−Si angles cover the full range from 120° to 180° with a maximum near 144° in silica glass [88]. The density of the network and other properties are variable. In nature, noncrystalline silica minerals are formed by melting and subsequent quenching (lechâtelierites), by pressure densification (diaplectic glasses), by rapid quenching of silica from hydrothermal solutions or by precipitation from a solution, sol, or gel near ambient conditions (opals). The range of Si−O−Si angles is narrower in opals than in silica glass [89]. The mean Si−O−Si angle is reduced in densified glass [90]. The mineral species are listed in Table 4.

Due to the lack of structural order and stoichiometry, the boundary between silica glass and silicate glass is fuzzy. For natural silica glass it can be arbitrarily drawn at about 3 mol % of nonsilica components. Opals, however, can contain considerable amounts of water and different varieties (opal-A_G and -A_N) can be distinguished by their microstructure and water content.

Lechâtelierites originate from molten silica formed by lightning strikes (fulgurites) or by meteoritic impact (meteoritic glass) on quartz sand or silica rocks [5, 10]. The glass body is featureless like synthetic fused silica but may contain some pores, inhomogeneities, and relicts of

Table 4. Noncrystalline silica minerals

Mineral or phase *	Species *	Microstructure	Refractive index	Water content, wt %
Opal-A (water containing noncrystalline silica particles)	opal-A_G precious opal	close packing of homometric $SiO_2 \cdot nH_2O$ spheres	1.45–1.46	4–6
	potch opal	irregular packing of heterometric $SiO_2 \cdot nH_2O$ spheres	1.45–1.46	4–6
(water containing natural silica glass)	opal-A_N or hyalite	botryoidal with striations and bubbles from degassing	1.45–1.46	2–4
Lechâtelierite (almost water free natural silica glass)	silica fulgurite	relics of quartz grains and pores		
	meteoritic silica glass **	schlieren, meteoritic detritus and reaction products	ca. 1.458	< 0.1

* A: noncrystalline; G: gel structure; N: network structure;
** e.g., Libyan desert glass.

quartz grains. The water content is usually less than 0.1 wt %.

Opal-A_G consists of a close packing of noncrystalline silica spheres with diameters ranging from ca. 100 to 500 nm. Water fills the interstices of the packing and is incorporated within the spheres [80]. Homodisperse spheres form a regular cubic, hexagonal or mixed closest packing [91], which gives rise to Bragg diffraction of visible light and produces the characteristic play of color of precious opals [92]. Heterodisperse spheres form irregular packings that result in diffuse scattering rather than in diffraction of light (potch opal).

Geyserite is precipitated in hot springs as sinter from silica-rich aqueous solutions. *Biogenic opal-AG* stems from plants (grass can contain up to 5 wt % silica, rice husks even up 19 wt %) or from skeletal remains of silica plankton. Silica deposits on the sea floor locally formed massive sediment layers with a total thickness of several hundreds of meters (e.g., Monterey, California). The water content usually ranges from 5 to 8 wt % (mostly molecular water) [5, 80] but can reach up to 16 wt % in biogenic opals [93].

Opal-A_N or hyalite is a water-containing silica glass that was formed in volcanically active areas by quenching of a silica-rich fluid or steam. It can be found as botryoidal crusts and fillings of vesicles in lava rocks [94]. The microstructure is characterized by striations parallel to the surface. Opal-A_N contains up to ca. 3.5 wt % water, almost 40 % of which is bound as silanol groups within the glass network [80, 95].

1.5. Colored Silica Minerals

Pure silica (e.g., rock crystal) is colorless. Except for precious opal, the color of silica originates from selective absorption by incorporated impurities forming so-called color centers. In several cases X-ray or γ radiation is necessary for the activation of the color centers. Heat treatment has an influence on the color, too. Several colored varieties of silica are traded as gem stones. Review articles on the origin of the color are given in [10, 96, 99].

Amethyst has a violet color due to the incorporation of Fe^{3+} on lattice sites and at I_4 interstitial sites in the helices of the quartz structure. Irradiation removes an electron from iron on lattice sites, turning it into Fe^{4+}. The electron is trapped by interstitial iron, which is reduced to Fe^{2+}. The color is not stable above ca. 300 °C but can be restored by renewed irradiation [97, 98].

Citrine is yellow to yellow brown with a smoky cast. It contains iron only at I_4 interstitial sites. If amethyst is heated in air at 300–560°C, iron oxide particles of submicroscopic size are exsolved, producing a yellow to yellow-brown color with reddish tints.

Rose quartz occurs as massy aggregates of quartz rather than as single crystals. The pale pink to deep rose color originates from Fe – Ti intervalence charge transfer in fibrous nanoinclusions of dumontierite or a closely related phase [100].

Smoky quartz or *morion* contains substitutional Al^{3+} on Si^{4+} lattice sites [101]. The black or grayish brown color is caused by irradiation, which produces an electron hole at oxygen bonded to aluminum. The color center is destroyed by heating above 200 °C but can be renewed by irradiation.

Other colored single-crystal quartz varieties such as pink or blue quartz have been found in nature or produced artificially but have gained only minor commercial importance as gemstones [102]. The color centers are often nonuniformly distributed, but for example in amethyst they are concentrated at Brazil twin boundaries parallel to the $\{10\bar{1}1\}$ faces [103].

Microcrystalline quartz minerals can be colored by inclusion of foreign pigment particles in the microstructure rather than by color centers in the crystal lattice. *Chrysoprase* is green due to submicroscopic nickel compounds, and the red or brown color of *jasper* is caused by Fe_2O_3 or $Fe(OH)_3$. *Agate* mainly consists of layers of translucent gray or white chalcedony which can be naturally or artificially colored by infiltration of iron or other compounds between the fibers. *Fire opal* is probably also colored red by Fe^{3+} pigmentation.

The colors of *precious opals*, however, are caused by Bragg diffraction of light and not by absorption. Therefore, the color is of higher spectral purity than in other gems. The diffraction originates from the ordered arrangement of homometric silica spheres and water with a lower refractive index in the interstices of the sphere packing. The maximum wavelength of the color depends on the size of the spheres. Opals containing large spheres produce all colors up to red whereas opals with smaller spheres show only a blue or violet color. Different orientations of grains with coherent sphere packing yield different colors. The play of color observed on turning a stone is enhanced by abundant stacking faults in the packing sequence of the silica spheres, producing diffraction streaks rather than sharp diffraction spots [91, 92].

1.6. Silica Rocks

Many special, trade, and trivial names exist for silica rocks, often with various meanings and imprecise definitions. A classification based on constituent silica phases and microstructure is given in Table 5 [2, 104, 105]. Silica rocks contain more than 90 wt % SiO_2. An ordinary sandstone, for instance, contains only about 65 wt % silica mainly as quartz grains. The rest is matrix and cement (feldspar, $CaCO_3$, mica, clay minerals, etc.). Transitions occur between silica rocks with increasing burial depth in the earth crust which not only result in a reduction of porosity and water content. Quartz sand is solidified to quartz sandstone (arenite or ganister) by cementation of the grains and grades by recrystallization into quartzite under pressure at elevated temperatures. Similarly, noncrystalline biogenetic silica remains of diatoms, radiolaria, etc. are transformed under compression into porcellanite and finally to chert. When the origin is known these rocks are designated as diatomite, radiolarite, or spiculite. Lydite and novaculite are radiolarites of paleozoic age. Tripoli is a friable porous porcellanite but the term is often used in a broader sense for abrasives and polishing media (Section 2.1). Ganister is a quartz sandstone cemented by chalcedony or opal-CT. Sandstone with less than 10 % matrix and cement is called arenite and with more than 10 % matrix wacke (quartz arenite and quartz wacke, respectively, when the total SiO_2 content is higher than 90 wt %). Chert nodules concentrated in layers of cretaceous chalk or limestone are called flint, silex (French), or Feuerstein (German). Agate geodes are fillings of vesicles in volcanic rocks, consisting mainly of chalcedony bands, interstratified by granular microquartz, quartzine, opal-C, opal-CT and coarse quartz crystals [107]. Further silica rocks and stones such as pegmatite quartz, vein quartz, sand, and gravel that are used as raw materials are treated in Chapter 2. A survey of industrial silica rocks is given in [108 – 110].

1.7. Crystalline Silica Products

Here only silica products made by crystallization from mobile mother phases or by transformation in the solid state are discussed. In contrast to the many noncrystalline silica products, crystalline products are limited to hydrothermally manmade quartz crystals and thermochemically produced polycrystalline silica.

Table 5. Silica rocks (> 90 wt % SiO_2)

Classification	Name	Main constituent silica minerals, microstructure	Special, trade, and trivial names (selection)
Sedimentary, nondetrital	diatomite (Chap. 3	opal-A_G; porous coherent diatom accumulation	kieselguhr [3]
	porcellanite	opal-CT; microporous rather strong rock	tripoli [106], opoka
	chert	chalcedony, granular microquartz; nonporous tough rock with conchoidal fracture	hornstein, novaculite, lydite, radiolarite, chert
Sedimentary, detrital	quartz sandstone	quartz grains cemented by direct contact, authigenic outgrowth, chalcedony or opal-CT	quartz arenite, quartz wacke, ganister, zementquartzite
Metamorphic	quartzite	recrystallized quartz grains and cement	crystalline quartzite, felsquartzite
Volcanic	geyserite	opal-A_G, opal-CT; friable, porous	silica sinter

Man-made or cultured quartz is of major economic importance and its use is growing. In contrast, the production of silica refractories for use in high-temperature processes is stagnating.

1.7.1. Cultured Quartz Single Crystals

Manufacture. The growth of quartz single crystals is carried out hydrothermally in sealed steel autoclaves with constant temperature difference and density-driven convection [111] (Fig. 12). Hydrothermal syntheses are defined as reactions or crystal growth at high pressures and temperatures in aqueous solutions, whether in the sub- or supercritical state [112].

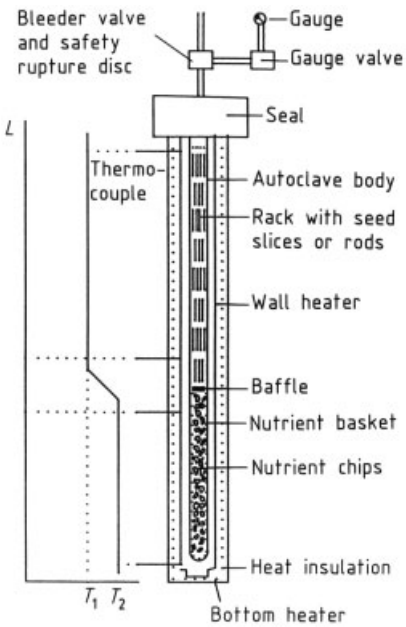

Figure 12. Autoclave for quartz single crystal bar growth

Natural quartz [113] is fractionated into chips (lascas) of controlled size distribution (2.5–4 cm) and surface area. They serve as nutrient and are filled into an iron wire basket which is placed in the lower part of the autoclave. The lower part is separated from the upper part by a baffle of defined opening. Seed slices or rods, cut with defined crystallographic orientation from natural or man-made untwinned quartz crystals with a low content of structural defects, are hung in a rack and placed in the upper part of the autoclave. Preferred seeds are Z plates, cut parallel to $\{0001\}$, and Y rods, cut parallel to $\{11\bar{2}0\}$ and $\{0001\}$ and perpendicular to one of the a-axes.

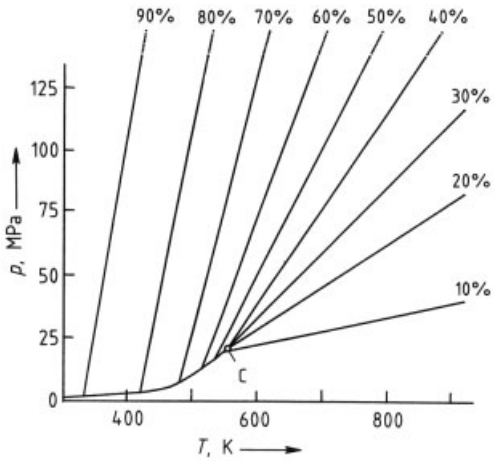

Figure 13. p–T diagram of H_2O, isochores for various percentages of autoclave filling [115, 116]
C: critical point at 22.04 MPa, 547.4 K

Aqueous growth solutions with different mineralizers (mainly 0.5–2.0 M NaOH or Na_2CO_3 and additives such as Li_2O) fill ca. 70–80% of the inner autoclave volume. Special

syntheses, e.g., the growth of amethyst, which needs structural incorporation of traces of iron, require K_2CO_3 or NH_4F solutions [97, 114]. After sealing, the autoclave is electrically heated to a working temperature of ca. 675 K at the outer wall of the nutrient part (T_2) and ca. 20–40 K less at the seed part (T_1) (Fig. 12). The actual temperature difference ΔT in the reactor is much smaller (5–10 K). With increasing temperature, the pressure increases according to the percentage filling (Fig. 13). Under working conditions, it is ca. 100–150 MPa. A rupture disk in the pressure tubing at the seal prevents hazardous pressure increase.

During heating and cooling of the autoclave, at pressures below ca. 80 MPa, the hydrothermal system passes through a region of retrograde SiO_2 solubility. Above 80 MPa the solubility increases with temperature and pressure (Fig. 14). The mineralizers increase the solubility strongly, for example, at 673 K and 100 MPa from ca. 0.15 wt % SiO_2 for pure water to about 2–3 wt %. The solubility is approximately proportional to the mineralizer concentration [117]. Concentrations of NaOH or Na_2CO_3 above 2 mol/L, however, can lead to deposition of sodium silicate glass at the bottom of the autoclave, which obstructs the growth process.

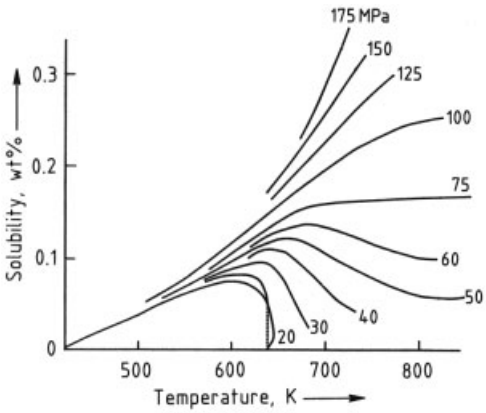

Figure 14. Solubility of SiO_2 in H_2O [117]

In the nutrient section SiO_2 goes into solution. The solution at the bottom part of the autoclave has a higher temperature than the solution in the seed section and rises in the gravitational field due to its lower density. Above the baffle, it becomes cooler and supersaturated with SiO_2, and growth starts at the seeds. The cooled solution has a higher density and sinks into the nutrient section where it is warmed and becomes subsaturated. It dissolves silica from the nutrient and ascends, and so on in a permanent circulation driven by density difference until the desired quantity of nutrient has been transported to the growing quartz crystals.

The effective nonequilibrium distribution coefficient is defined as

$$k_{\mathrm{eff}} \approx C_{\mathrm{sact}} / C_{\mathrm{lact}}$$

where $C_{\mathrm{s\ act}}$ is the actual trace element concentration in the solid, and $C_{\mathrm{l\ act}}$ the actual trace element concentration in the liquid. The trace impurities accumulate in the solution [118] and the grown quartz crystals are of higher purity than the nutrient quartz. Doping with traces of special components changes the physical properties. For example, doping with LiF improves the electrical quality.

Raising *p*, *T*, and ΔT increases the rate of SiO_2 transport and of quartz growth. However, the high-temperature strength of the autoclave alloy sets limits on *p* and *T*. A low-pressure and a high-pressure process are usually applied. A pressure of 70–100 MPa, a temperature of 620 K, a temperature difference of 10 K, and a 0.6 to 0.8 M sodium carbonate solution are typical of the former and yield a growth rate of about 0.4 mm/d. The high-pressure process (100–150 MPa, 650 K, ΔT = 25 K, 1.0 M sodium hydroxide solution) yields a growth rate of up to 1.0 mm/d, but at the expense of higher costs for equipment [100]. The transition of convection from laminar to turbulent or even cellular flow with increasing ΔT limits these factors as well. Increasing face instability [118] and defect concentrations with increasing rates of transport set further restrictions.

The corrosion of the autoclave steel alloy is inhibited by formation of a dense coating of microcrystalline acmite $NaFe^{III}[Si_2O_6]$ and traces of emeleusite $LiNa_2Fe^{III}[Si_6O_{15}]$ by reaction with the growth solution.

The duration of growth runs varies between one and four months. The grown quartz crystals are formed by grinding to ca. 15–20 cm long bars. The orientation of the bars is determined by X-ray diffraction before cutting them

Table 6. Dimensions and yield capacities of autoclaves for man-made quartz

	Inner diameter, cm	Inner length, m	Inner volume, L	Yield per run, kg
Japan	65	13	4500	2000
United States	40	4.5	560	250

into wafers. An autoclave with 250 mm inner diameter and an inner length-to-diameter ratio of 12 gives a yield of ca. 100 kg of quartz bars per month. The loss by removal of seeds from the bars is ca. 20 wt %. Part of the cutting waste is used in the jewelry industry. The growth parameters which determine yield and quality of man-made quartz, are complexly coupled [120]. Dimensions and yield capacities of autoclaves are given in Table 6.

The largest production capacities exist in Japan, the United States, Russia, and China; smaller capacities are found in Europe, South Korea and South Africa. The total production of cultured quartz was estimated at about 200 t/a in the USA at an average value of ca. $ 186 per kilogram in 2005.

Small amounts of colored quartz [96] for jewelry are grown in Russia. While blue (Co-containing) and light green and yellow (Fe) quartz are already colored in the growth process, the Al and Fe color centers for smoky quartz and for amethyst must be activated after growth by γ radiation. Seeds for the first-mentioned color variations are Z plates, whereas for smoky quartz and amethyst X plates cut parallel to $\{10\bar{1}1\}$ are used because of face-specific incorporation requirements of the coloring elements.

Filter and oscillator quartz must be free of twinning and need low defect concentrations [121]. Structurally incorporated impurities are determined by chemical trace element analysis. Impurities which produce paramagnetic defects (e.g., Fe) can be detected by electron spin resonance spectroscopy [122]. Traces of water are determined by IR spectroscopy [123, 124]. Inclusions of foreign elements, depending on type, size, and position in the microstructure of the crystal are recognizable by X-ray diffraction, under the polarizing light microscope (PLM), or with the scanning electron microscope (SEM) and can be chemically analyzed by energy dispersive X-ray spectroscopy (EDS) in the SEM or with the electron microprobe [125]. Dislocations produce diffraction contrast in X-ray topography [126, 127]. Brazil twinning can be detected with PLM, but Dauphiné twinning can only be made visible by etching. Both types of twinning can be differentiated by X-ray topography [128, 129].

The quality of resonators is described by the electromechanical Q value. This dimensionless quality factor is defined as:

$$Q = \frac{2\pi f L}{R}$$

where f is resonance frequency, L inductance, and R resistance.

The Q value increases with decreasing energy loss due to relaxation and internal friction. The slower the decay of the oscillation amplitude after cutting off the excitation, the higher is Q. The quality of the quartz crystals contributes to the resonator Q value. The latter is not measured directly but a material Q is determined instead. A correlation between the IR absorption of structural OH groups (hydrolytic nonbridging oxygen, Fig. 2) at 3580 and 3410 cm^{-1} and the material Q value is usually utilized for quality control [130]. An empirical infrared a value is defined:

$$a = 1/t \log\ (T_{3800}/T_{3500})$$

where a is the extinction coefficient at 3500 cm^{-1}, t the thickness of the sample in centimeters, and T_n the transmittance at wave number n.

The material Q can be evaluated from

$$10^6/Q = 0.144 + 7.74a + 0.44a^2$$

The three highest grades have a material Q of 1.8, 2.2, or 3.0×10^6

Radiation in space, from uranium or thorium impurities, or from other sources affects the oscillator frequencies by acting on alkali metal impurities. Therefore the quartz must be grown to extreme purity Alkali metal ions on interstitial sites can be removed (swept) electrolytically after growth. This is done at elevated temperature (below the transition temperature of 573 °C in order to avoid electrical twinning) by applying a field of 1 kV/cm [119].

For further specification, the quartz bars are examined for inclusions which might contain

acmite impurities that stem from reaction of the hydrothermal solution with the steel walls of the autoclaves. The density of dislocations is determined microscopically after etching with ammonium bifluoride by counting the etch pits or channels. The best grades have dislocation densities of 10, 30, or 100 per cm^2.

Most of the cultured quartz is optically right-handed, as right-handed seed crystals are normally used. Electronic applications are treated in [131, 132]. Natural single-crystal quartz is now used only for a few special applications.

Despite the increasing demand, growth capacities are not markedly increasing because of improved cutting technologies, photolithographic etching, and miniaturization.

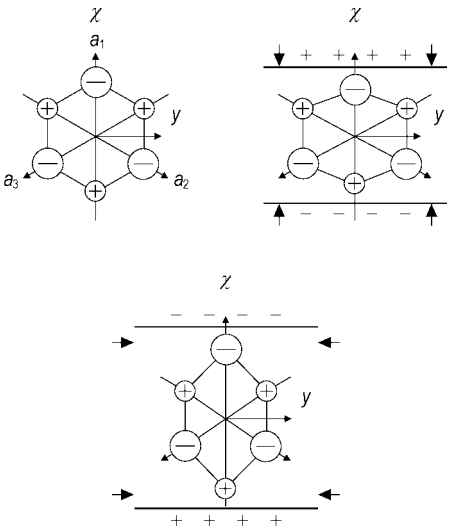

Figure 15. Illustration of the piezoelectric effect in quartz. Left: Simplified structural model of quartz viewed along the trigonal *c*-axis; small circles represent silicon and large circles two oxygen atoms. The trigonal a_i-axes and the orthogonal *x*-and *y*-axes are shown. Middle: Deformation of the structure due to stress applied along the *x*-axis (one of the three polar twofold rotation axes) produces charges on surfaces perpendicular to the *x*-axis (longitudinal direct piezoelectric effect). Right side: Stress applied along the *y*-axis also produces charges on the faces perpendicular to the *x*-axis (transverse direct piezoelectric effect). Change from compression to tension reverses the signs of the electric charges.

The Piezoelectric Effect of Quartz. If stress is applied along one of the three polar twofold rotation axes of quartz an electric polarization is produced (direct piezoelectric effect). The direction of the polarization vector is reversed upon changing from compressive to tensile stress and vice versa. Stress perpendicular to a polar axis and perpendicular to the trigonal *c*-axis (along the *y*-axis) also yields an electric polarization along the polar axis.

$$P_i = d_{ijk}\sigma_{jk}$$

i, j, k = 1, 2, 3
P_i = electric polarization
d_{ijk} = piezoelectric moduli
σ_{jk} = stress tensor

The anisotropy of the piezoelectric effect is described by a third-rank tensor with 27 components. Due to the symmetry of the stress tensor and the point-group symmetry of quartz (32) the number of independent components is reduced to two for quartz

$$d_{111} = -2.31\times10^{-12}\ \mathrm{C/N}$$

$$d_{123} = -0.727\times10^{-12}\ \mathrm{C/N}$$

for right-handed quartz (the signs of d_{ijk} are reversed for left-handed quartz [133]).

If an electric field is applied in appropriate direction the shape of a quartz crystal changes (<0.1 %). This is called the converse piezoelectric effect. An alternating electric field causes the crystal to vibrate.

$$\varepsilon_{jk} = d_{ijk}E_i$$

i, j, k = 1, 2, 3
E_i = electric field
d_{ijk} = piezoelectric moduli
ε_{jk} = strain tensor

Uses. The piezoelectric effect was detected by Pierre and Jacques Curie in 1880/1881 using quartz crystals. In 1917, the first sonar was constructed by Langevin to detect submarines. Since the 1920s quartz oscillators have been used for frequency control in radio broadcasting and communication electronics. The first oscillator for frequency stabilization, cut from natural quartz, was described 1921 by Cady [131]. A quartz crystal is stimulated to vibrate mechanically at its resonance frequency. The piezoelectric response in turn is then utilized for stabilization of the frequency of the circuit. The

use of quartz resonators increased as a consequence of the mass production of color TVs since the 1960s, quartz clocks and watches since the 1970s, and the developments in digital microelectronics. Today, virtually no appliance or device is working in communication electronics or computer technology without a quartz oscillator.

The resonance frequency depends on the cut and the shape of the quartz crystal. Initially, only cuts perpendicular to the *x*-axis were used for sonar transducers. Today, quartz tuning forks for wristwatches are produced from *x*-cuts. For other purposes, different cuts are preferred. So-called AT cuts are most commonly used. An AT-slice cuts the *y*-axis and forms an angle of 35.3° with the trigonal *c*-axis. As the d_{111} piezoelectric constant has a negative and the d_{123} constant a positive dependence on temperature, an AT cut provides independence of the resonance frequency from temperature [133].

In the megahertz range bulk waves are utilized. For the gigahertz range surface acoustic-wave devices are now produced by photolithographic techniques similar to those used in the semiconductor industry [119].

Due to its good elastic properties and very high stability, quartz is still the most important piezoelectric material. However, in the fields of ultrasonics and piezoelectric sensors and actuators, quartz is widely replaced by materials like lead zirconate titanate (PZT) ceramic which show a much stronger piezoelectric effect [134].

Other uses of quartz single crystals utilize optical properties like optical activity, birefringence, and dispersion.

1.7.2. Polycrystalline Silica Products

Refractory Silica. Dense shaped silica products consist per definitionem of greater than 93 wt % SiO_2 [135]. Usually, the silica content varies between 95 and 97 wt %. Refractory silica is available in three grades: superduty, regular duty, and coke-oven quality [136, 137]. Silica bricks are used up to 1925 K. The major applications of dense silica bricks are coke ovens, superstructures of glass-melting tanks, and hot-blast furnaces (checker bricks). Unshaped silica mixes are used for monolithic structures and repairs.

Silica Bricks. Raw materials for industrial silica brick production are high-grade quartzites or silica cemented quartz sandstones with >97 wt % SiO_2. The raw material is crushed, washed, and sieved. The sieve fractions are mixed to the desired grain size distribution. Small amounts of quartz sand are also used, and 1–4 wt % $Ca(OH)_2$ is added as mineralizer (transformation activator) and bonding agent. Sulfite lye acts as plastifier of the batch and provides sufficient strength for manual or mechanical shaping and drying. The bricks are fired at about 1700–1760 K in tunnel or shuttle kilns for one to three weeks (depending on their shape). The quartz raw material transforms into stacking-disordered cristobalite and tridymite under these conditions. The volume of the bricks increases by 4–5 % due to the higher specific volume of the high-temperature silica modifications. The amounts of cristobalite and tridymite depend on the raw material blend as well as on production parameters, mainly temperature and time of firing [138 – 140]. Tridymite forms only in the presence of suitable impurities and mineralizers; the longer the duration of firing the more is formed (see Section 1.3.1) [141 – 144]. The rate of the quartz transformation varies for the different types of raw materials. It depends mainly on the crystallite size distribution and on the amount and distribution of impurities (Al_2O_3, TiO_2, alkali metal oxides, Fe_2O_3). Quartzite transforms more sluggishly than quartz sandstone with a high content of microcrystalline quartz in the cement.

The burnt silica bricks contain small amounts of residual quartz (<6 wt %) and wollastonite $CaSiO_3$. The latter is a crystallization product of an intergranular glass phase ($SiO_2/CaO/Al_2O_3$). An excess of residual quartz must be avoided because of an irreversible after-expansion (Fig. 15) in service above ca. 1500 K, which is due to the transformation of the residual quartz into cristobalite/tridymite.

The content of residual quartz is easily determined by X-ray powder diffraction [145]. The quantitative determination of cristobalite and tridymite, however, is difficult due to their structural defects [138, 139], and large errors can result if these are not taken into account [146].

Silica bricks expand on heating and reach a maximum between 1075 and 1275 K of ca.

Table 7. Material properties of dense silica bricks [149]

Application	Bulk density, g/cm^3	Apparent porosity, vol %	Cold crushing strength, MPa	Refractoriness under load (differential) T_{05}, °C	Thermal expansion 1000 °C, %	Thermal conductivity 1000 °C, W m^{-1} K^{-1}
Coke oven	1.78–1.90	18–23	30–60	1610–1650	1.3	1.8–2.2
Glass furnace	1.81–1.85	19.5–21.5	30–40	1630–1670	1.4–1.5	1.8–1.9

1.3–1.5 vol % and have a small negative expansion at higher temperatures (normally 0.1–0.2 vol %) which is attributed to a structural contraction of the tridymite and cristobalite components [147, 148]. Due to the comparatively low thermal expansion, refractory silica has a very good thermal shock resistance above 875 K. Below this temperature, however, extreme care must be taken on heating and cooling to avoid cracking. A further valuable feature is the extraordinarily good creep behavior of refractory silica under compression at high temperatures. This is attributed to the very small content of liquid phase which is due to the low content of impurities (in particular, < 1.5 wt % Al_2O_3 and < 0.4 wt % alkali). Silica bricks are used up to 1925 K. The corrosion resistance against acid slags is good. However, alkali-containing vapor may cause corrosion by melt formation below 1750 K.

Table 7 gives a survey of the properties of silica bricks. For the production of very dense silica bricks with high thermal conductivity, Si_3N_4 or SiC is mixed into the raw batch. During firing these compounds are oxidized, and the resulting crystalline silica fills the intergranular pores [150, 151].

During operation in a temperature gradient, silica bricks develop zones with different microstructures and silica phases without significant change of shape. Above ca. 1730 K cristobalite predominates, and between 1580 and 1680 K tridymite [147, 152, 153].

The major applications of dense silica bricks are coke ovens, superstructures of glass-melting tanks, and hot-blast furnaces (checker bricks). Unshaped silica mixes are used for monolithic structures and repairs.

Cristobalite sand, powder, and flour are produced from quartz sand by calcination in rotary furnaces at 1800 K or from flint in shaft furnaces with subsequent grinding in ball mills. Another route works with processed sand and addition of mineralizers at 1500 to 1700 K, leading to products with > 99 wt % SiO_2. The former products are used as refractory mortars and ramming mixes in ceramic kilns and for blending with quartz or other mineral sands for various uses (e.g., in foundries) preferably as a modifier of thermal expansion. Heat-insulating lightweight (up to 60 % porosity) silica refractories are produced by addition of material that burns off during firing (e.g., sawdust, cork). Diatomaceous earth (see Chap. 3) can also be used as raw material. Burned diatomite consists of about 80 wt % crystalline silica, mainly cristobalite with some quartz. It serves as heat-insulating material for use up to 1200 K.

High-purity cristobalite from processed sand serves as weathering- and exhaust-resistant white filler which has no yellow shade for thermoplastic road marking (flour as pigment, sand as antiskid), for production of plastics, adhesives, wall paint, and texture coatings. Surface treated (preferably with siloxanes) flour is incorporated into silicone rubber and cable coverings and into EVA or PE foils as IR absorber. It is used for glaze frits, engobes, and other ceramics inter alia as modifier of thermal expansion. In silicate chemistry it serves as raw material.

Rice husk ash is prepared by burning the husks at about 800–900 K for ca. 2 h. The husks contain ca. 9–19 wt % noncrystalline silica, the ashes ca. 85 wt %. They are used as additives to Portland cement for reduction of the alkali–silica reaction and the detrimental cracking that is caused by this reaction [154].

Refractory silica cloth is replacing asbestos. It consists of 8–10 mm thick glass fibers with a silica content of more than 96 %.

1.8. Noncrystalline Silica Products

Most of the noncrystalline silica products are treated in Chapters 3, 4, 5, 6, 7. Only silica glass and silica fume are briefly described here.

Table 8. Physical forms of quartz raw materials

Name	Formation and properties	Remarks, synonyms, and trivial names
Idiomorphic quartz	freely grown, euhedral faced, monocrystalline, possibly twinned, transparent colorless or smoky to dark brown or black	rock crystal, smoky quartz
Lump quartz	fragmented, anhedral, monocrystalline, may be twinned, transparent, colorless, smoky, milky, or white	chunk quartz, lascas, brazil pebbles, lump silica
Quartz pebbles	weathering product from quartz veins, worn by action of water, smoothly rounded, mono- or polycrystalline, transparent to white or yellowish white or gray	if transparent: rhinestone or rheinkiesel, grain size 2–60 mm
Granular quartz	xenomorphic product of natural (weathering) or industrial processing from granitic–pegmatitic rocks, monocrystalline, may be twinned, transparent to translucent, yellowish, gray, or smoky	grain size 0.5–1 mm [159]
Quartz sand	unconsolidated, wind- or water-eroded and transported weathering product of fragmented rocks, grumbled into mono- or polycrystalline grains, rounded or angular anhedral, the majority of grains are discernible with the naked eye, white to yellowish in bulk, in the individual grain transparent to translucent	silica sand, grain size 0.06–2.0 mm
Quartz powder	unconsolidated product of grinding or milling of coarser quartz, individual grains not discernible with the naked eye, white opaque in bulk	quartz flour, silica powder, silica flour, grain size 0.002–0.06 mm
Quartz rock	intergrowth of quartz grains or consolidated by cementation of grains	quartzite, quartz sandstones, see text and Table 5

Silica glass is pure vitreous silica with diverse industrial applications. It is produced either by melting natural quartz crystals above 2050 K (fused quartz) or by vapor-phase hydrolysis of $SiCl_4$ in a methane oxygen flame (fused silica). The use of vitreous silica produced by sol–gel processes is restricted to thin-film technologies. The main benefits of silica glass are broad transparency to ultraviolet to near infrared, high resistance to thermal shock due to the low thermal expansion of 5.5×10^{-7} K^{-1} and high chemical corrosion resistance [155]. These properties led to uses as optical material, optical waveguides and chemical apparatus. Silicate glass is treated in detail in → Glass.

Silica fume is a very fine powder consisting of noncrystalline silica spheres with an average diameter of ca. 0.1 µm. It forms at temperatures of about 2050 K as a byproduct of the reduction of quartz in electric arc furnaces for the production of silicon and ferrosilicon. The silica vapor is condensed and collected to prevent its emission. Silica fume is mainly used as an admixture in cement, concrete, and mortars to improve strength and durability. Due to the small particle size and the large surface area (up to 30 m^2/g), it acts both as a microfiller that reduces the porosity and as a pozzolan which reacts with the calcium hydroxide from the cement [156].

2. Quartz Raw Materials

2.1. Physical Forms and Occurrence

The range of quartz in nature is vast. Table 9 describes the physical forms and Table 9 lists the geological occurrences of raw materials [157 – 166].

Table 9. Geological occurrence of quartz raw materials

Name	Occurrence	Remarks
Idiomorphic quartz	grown in veins, vugs, vesicles, and shear fissures of igneous, metamorphic, or sedimentary rocks	chemical purity and transparency increases from base to top of crystal
Vein quartz	xenomorphic and dense quartz filling of veins, cracks, faults, dikes, bedding planes, fractures, or joints in zones of structural weakness of igneous, metamorphic, or sedimentary rocks	
Pegmatite quartz	polymineralic veins, most frequently of granitic composition (quartz, feldspar, mica), however, it may be of any other composition, or coarse crystalline areas in granites or other igneous plutonic rocks but also in metamorphic rocks with all transitions to the rock itself	
Quartz from disseminated hydrothermal deposits	polymineralic fillings of dispersed vein systems in large volumes of rocks, dissemination along cracks and fractures or in zones of high permeability for hydrothermal solutions	quartz frequently contains fluid and solid inclusions [174, 205] (Fig. 16)
Quartz gravel and sand	occurrences are widespread and diverse, mainly in cretaceous, tertiary, and alluvial beddings; detrital product obtained from rocks by various weathering, transport, selection, and bedding conditions [204 – 207]	

The main characteristics of quartz raw materials are grain size and shape, their distribution functions, surface area and roughness, sorption capacity [163], chemistry, physical properties, and cost. Increasing demand for extremely pure quartz for silica glass production and in the quartz oscillator industry, and the growing use of refined grades of sand makes consistency and long-term availability important [161].

Impurities and their accessibility to refining processes are of major importance. In most cases, trace element analysis gives bulk composition and no information about the location of impurities in the microstructure. This, however, is essential for choice of quartz processing methods. Figure 16 shows the principles of impurity distribution in a crystal.

Quartz gravel and sand are the most important raw materials on a volume basis; the construction industry consumes about 95 % of production (ca. 65 % structural and ca. 30 % civil engineering). In 1995 Germany produced 18×10^6 t of sand (56 % river sand, 44 % high-purity sand with > 98 wt % SiO_2). The overwhelming part of the sand supply for Europe comes from Belgium, Germany, and the Netherlands. In 1995 the EU produced ca 16×10^6 t of high-purity sand. The main consumers were the float- and container-glass industry with rising customer specifications [167]. Recycling of glass reduced the demand for sand in container glassworks.

Tripoli contains ca. 98–99 wt % SiO_2. It occurs in sedimentary strata as a leaching product of calcareous cherts or siliceous limestones. White tripoli is low in iron impurities (< 0.1 wt % Fe_2O_3); rose or cream tripoli contains up to 1 wt % Fe_2O_3. It consists of submicroscopic quartz crystallites. Particles with sizes 5 to 100 µm are crystallite aggregates. White Southern Illinois Tripoli has the composition SiO_2 99.5, Al_2O_3 0.009, Fe_2O_3 0.025, CaO 0.15, TiO_2 0.005 wt %.

Kieselerde is a natural mixture of quartz and kaolinite [168]. The mineralogical and chemical composition of Kieselerde from Neuburg (Donau) follows [168]: 62–84 wt % quartz, 10–30 wt % kaolinite (content increases with decreasing grain size), 6–8 wt % accessorial minerals; 78–87 wt % SiO_2, 8–14 wt % Al_2O_3, $<$ 1 wt % Fe_2O_3, 60 ppm P, and 35 ppm S (both relative to Si).

Chalcedonies (Section 1.3.1) are classified according to their deposition as wall-banded Ch-w or horizontally-banded Ch-h (Table 3). According to the optical character of their fiber elongation, two varieties can be distinguished: chalcedony with length-fast elongation: Ch-w_{LF} and Ch-h_{LF} and quartzine Ch_{LS} with length-slow elongation. Chalcedonies occur in various geological environments, filling veins, vugs, and vesicles in volcanic rocks, sediments, ore deposits, or covering and impregnating plants in regions with volcanic activity [169, 170].

Ch-w is bluish gray in reflected and brownish in transmitted light due to Rayleigh scattering. It has a brown color if it was impregnated with ferrous humic acid waters (then called sarder). It possesses an open submicroporosity which makes it artificially impregnable with coloring solutions and allows penetration of liquid impurities. The microstructure is monomineralic, of high toughness, and high purity, which makes it a good material for milling devices, mortars, and bearing stones.

Chalcedony is slightly weaker than quartzine. If quartzine is interstratified into chalcedony layers, it forms a raised abrasion relief [171]. Since the resistance to HF etching is reverse, etch leveling is possible. Ch-h has almost no open porosity.

2.2. Processing

Impurities may be fixed on the surfaces of the individual quartz crystals or grains. Then they are rather easily accessible for removal. If they are included as minerals or fluids in the microstructure (Fig. 16) [163, 172, 173], the crystals must be crushed for removal. If the impurities are incorporated into the crystal structure, dissolution of the crystal is required for removal. Detrimental accessory minerals in quartz sands are mainly rutile, zircon, tourmalines, magnetite, hematite, ilmenite, chromite, Mn oxides, feldspars, and clay minerals.

Purity grades of processed natural quartz raw materials are [159]:

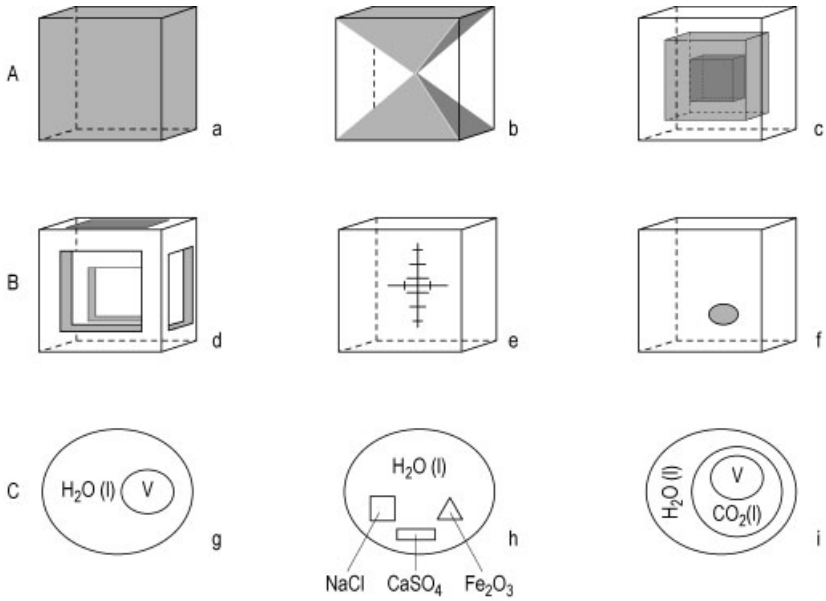

Figure 16. Principles of impurity distribution in a single crystal.
A) Structural incorporation
a) Random; b) Growth sector specific; c) Zonal
B) Microstructural inclusion
d) Skeletal growth and face specific adsorption, possibly with subsequent overgrowth; e) Dendritic growth or precipitation; f) Accumulated and accidentally ingrown, e.g., in fissures or dendritic cavities.
C) Types of fluid inclusions in quartz [174 – 176].
v = vapor, l = liquid; squares, rectangles, and triangles represent crystallites.

Grade	**Impurity content (ppm Si)**
Low	> 500
Medium high	300–500
High	2–50
Ultrahigh	1–8

Processing of coarse monocrystalline quartz to lumps or lascas [159]:

- Size reduction (weight specification for lascas ca. 15–50 g)
- Under water: sorting out of pieces with visible inclusions; under UV light: fluorescent pieces
- Under visible light: sorting of different grades of transparency.

Product: various grades of lascas.

A special process is γ-irradiation-aided selection of lascas [159]:

- Irradiation with doses ≤ 1 Mrad
- Color sorting [177]; the depth of color is mainly a function of the Li/Al trace element concentration ratio (Fig. 17)

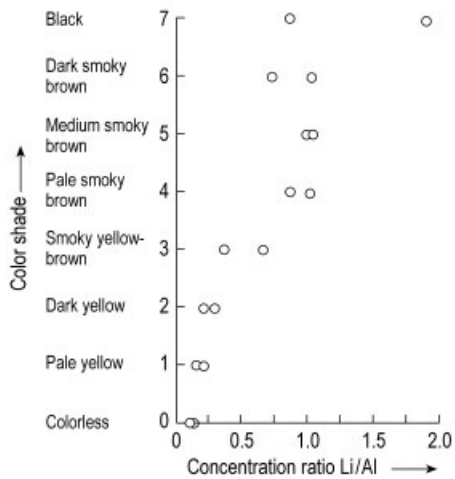

Figure 17. Irradiation color (γ radiation, 1 Mrad) and Li/Al concentration relationship for initially colorless first-grade lascas (Minas Gerais, Brazil) [159].

Processing of quartz rocks, vein quartz, and quartz from disseminated hydrothermal deposits:

– Breaking and sieving.

Product: quartz of low chemical purity for production of high purity silicon compounds.

Processing of granular quartz from deeply weathered feldspathic rocks [159]:

– Crushing and grinding
– Multistage flotation
– High-field magnetic separation
– Chemical treatment with dilute sulfuric or oxalic acid or chlorination at ca. 600 K
– Sieving

Product: medium- to high-purity granular quartz (type QUINTUS, Unimin Corp. New Canaan, Connecticut, USA).

– Acid (HF) leaching
– Elutriation with deionized water to pH ≥ 6
– Vacuum drying
– High-field magnetic separation
– Chlorination at ca. 1400 K

Product: high-purity granular quartz (type IOTA, Unimin Corp., New Canaan, Connecticut, USA; Table 10) preferably for production of silica glass implements: halogen and high-intensity discharge lamp tubings and semiconductor applications such as Czochralski crucibles and diffusion tubes.

Processing of quartz gravel:

– Separation of sand
– Washing
– Sizing and drying

Product: filter pebbles, beach pebbles for milling stones, gravel for concrete.

Processing of quartz sand [178, 179] varies depending on the required grade. Process steps are:

– Washing in water to remove loose impurities
– Friction washing (attrition)
– Crushing of agglomerated grains in water
– Rubbing off strongly adhering impurities (iron oxides and oxide hydrates, clay minerals)
– Screening
– Multistage flotation to remove accessory minerals
– Chemical treatment with aqueous acids (sulfuric, oxalic, fluoric)
– Chlorination at ca. 600 K
– Physical treatment by magnetic and electrostatic separation of iron oxide and feldspar contaminants
– Surface treatment with amino-, epoxy-, methyl-, methacryl-, or vinylsilanes [180, 181].

Sand for use in cement-based products should be free of chlorides and clay minerals. Sand for special mortars must be freed from organic residues by calcination [182].

Quartz powders are produced by dry grinding in encapsulated mills with chert or corundum lining and balls. Refining of powder from high-purity vein quartz gives grades with 8–50 ppm Fe/Si.

Processing of quartz rocks includes breaking, milling, sieving, washing, and sizing [178, 179].

Biological removal of iron impurities with cultures of acid-producing bacteria or fungi (e.g., *Aspergillus niger*) on a nutrient of molasses is discussed in [183].

2.3. Uses

Important uses of quartz raw materials or semifinished materials are listed below [161]:

Use	Raw or semifinished material
Abrasives and grinding media	sand, powder, tripoli, novaculite
Ceramics	quartzite, quartz-cemented sandstone, sand, powder
Chemicals	sand, powder
Construction of buildings and roads	gravel, sand, powder, quartzite, sandstone
Fillers, extenders	powder, tripoli
Filters	gravel, sand
Foundry	sand
Gems	monocrystalline quartz, chalcedony, agate, chert
Glasses, enamels, glazes	sand, powder
Hydrothermal growth of quartz single crystals	lump quartz, granular quartz

Table 10. Trace element analyses (ppm/Si) of UNIMIN-JOTA-quartz grades [208]

Grade	Li		Na		K		Ca		B		Al		Fe	
	min.	max.	min.	max.	min.	max.	min.	max.	min.	max.	min.	max.	min.	max.
Standard ST	0.7	1.5	0.9	1.5	0.7	1.5	0.4	1.5	0.08	0.10	15.2	22.0	0.3	1.5
JOTA-4	0.2	1.0	1.0	1.3	0.4	1.0	0.6	1.0	0.04	0.05	7.9	10.0	0.6	1.0
JOTA-6	0.2	0.3	0.1	0.2	0.1	0.2	0.5	0.7	0.03	0.04	7.9	9.5	0.2	0.3

Metallurgy	quartzite, sandstone, sand, powder
Milling	silex, chalcedony, flint, quartzite
Optical and oscillator devices	monocrystalline man-made and natural quartz
Silicon metallurgical grade, silicon alloys	lump quartz, quartzite, quartz-cemented sandstone
Vitreous silica	lump quartz, granular quartz, sand
Well fracturing	sand

Abrasives and Grinding Quartz. Quartz sand and powder are used for grinding, scouring, and polishing. The conchoidal fracture, hardness, and brittleness provide and maintain particles with sharp cutting edges during use. The use of quartz in sandblasting is decreasing due to health hazards (Chap. 9). Quartz sand is also of declining importance for sandpaper. Quartz powder is utilized in abrasive soaps and in scouring powders. Sandstones are used for grinding. Tripoli is friable and its grains have no sharp edges. Tripoli grades with 98–99 wt % SiO_2 are used for polishing [184]. Novaculite (Table 5) from Arkansas and Missouri, USA, has a microstructure of minute angular quartz grains with open porosity. It is used as oil-soaking whet- and honing stone. Whetstone grades range in porosity from 0.1 to 15 % [184].

Ceramics and Refractories. Batches for stoneware, porcelain, fireclays, and aluminum silicate refractories contain quartz as a nonplastic filler and glass former. White-burning, low-iron quartz powders produced from sand, flint, or quartz rocks are used. Flint is calcined before grinding. In pottery and earthenware, quartz sand and powder act as nonplastic fillers [185]. For silica refractories, see Section 1.7.2.

Chemicals. A wide range of chemicals is produced from quartz. Here, only one important product that is representative of different reaction processes and reaction temperatures, is briefly described.

Sodium or potassium silicate (waterglass; → Silicates) are melted from quartz sand and the alkali metal carbonates. Usually, the sand is of container-glass grade with < 0.1 wt % each of Al_2O_3 and CaO and a very low iron content ($<$ 0.02 wt % Fe_2O_3). Very high purity products require sand with < 0.02 wt % Al_2O_3 and CaO each. Melting is performed in tank furnaces at ca. 1700 K (furnace route) [186].

Waterglass from the hydrothermal route is produced in autoclaves by reaction of quartz sand with aqueous NaOH solution at ca. 430–500 K.

Construction of Buildings and Roads. Quartz sand and gravel are an integral part of asphalt, mortars, and concrete. For good rheological behavior sand or gravel should consist of well-rounded grains or pebbles. Optimized grain size distribution minimizes water demand and shrinkage in the setting reaction and contributes to the development of high binder strength. Quartz sandstones with various cementing material and quartzites have long be used as building stones [187]. Bricks and tiles are fired from mixtures of quartz sand and clay.

Since 1894 when the first calcium silicate (CS) units were industrially produced [188], their use for masonry in central Europe has rapidly developed [189].

Quartz sand (fluviatile) from a broad coastal strip of the Netherlands and North Germany (about 18×10^6 t/a) is processed according to the following (simplified) process scheme:

- Mixing and pretreatment of sand, lime, and water
- Densification and shaping by pressing
- Hardening in the autoclave with steam or water at about 16 bar and 475 K for ca. 6–8 h

In the hardening reaction, the calcium silicate hydrate (CSH) phases, preferably tobermorite and xonotlite, form a cement between the quartz grains. This gives the units a high com-

pression strength. Essential process parameters for the quartz sand are: grain size distribution (sieve line) and the shape and surface of the grains. The accessory minerals [mostly (in wt %) ≤ 5 feldspars and micas, ≤ 3 clay minerals, ≤ 3 calcium carbonates, ≤ 1 iron manganese oxides and oxide hydrates] [190] take part in the hardening reaction. Fe and Mn enter the crystal structures of the CSH phases as doubly charged ions and produce almost no hue. Humic matter spoils the process. About 1.5 t of sand is needed for the production of 1 m^3 CS units. In central Europe CS masonry is appreciated mainly due to its high raw density and sound absorption, its heat capacity and resistance against fire, and the good compression strength which allows the construction of space-saving walls.

Oil Well Cementing. Heat in the depth of oil wells converts the calcium silicate hydrate (CSH) gel of tobermorite character in the settling portland cement with its good cementing properties into α-dicalcium silicate hydrate (α-C_2SH). This causes the development of a porous and weak microstructure with open pathways for corrosive fluids. Addition of quartz sand and powder prevents this deteriorating transformation. Portland cements for use at very high temperatures contain 35–50 wt % of SiO_2 additives [191].

Fillers and Extenders [192 – 194]**.** Quartz and cristobalite fillers are hard and strong with high resistance against chemicals and weathering. They give the final product a characteristic dependence on thermal expansion. In paints, quartz increases the acid resistance and improves flowability, durability, and wear resistance. Silane-treated fillers [195] are used for corrosion prevention in primers instead of heavy metal pigments [196]. In engineering plastics and resins, compressive and flexural strength and thermal shock resistance are improved. The insulator character of quartz and its resistivity to surface leakage currents make it a valuable filler for epoxy resin encapsulations, high-voltage insulation, bushings, switch cases, cast resin transformers, and for filling and fixing of coils. For example, reaction hardening resins may contain up to 80 wt % SiO_2. Silane-treated quartz powder in epoxy castings reduces shrinkage and electrical loss, increases E moduli, compression strength, and thermal conductivity [197, 198]. Quartz is used as a filler in rubbers instead of carbon black when colors other than black are required. In tire linings, it gives improved heat aging, adhesion, and tear strength.

Filters. Quartz pebbles, preferably monocrystalline, and sand are the main filter materials for drinking water purification and softening.

Foundry Technology [160, 178, 179, 199]. Quartz sand is used in molds and cores. Refractoriness (i.e., the temperature at which sintering begins; ca. 1500–1700 K, depending on the content of feldspar), thermal shock resistance, cohesiveness in the moist state, and high-temperature strength are the most important properties. For cast iron, the mold must contain > 85 wt % SiO_2 and for steel castings 95 wt %. Granulometric factors such as particle size distribution, grain shape, state of grain surfaces, and surface area are also important. Coarse sand with low surface area requires less binding agent (usually montmorillonitic clay). Monocrystalline well-rounded grains ensure good permeability to gases evolved during casting. Mold surfaces made of fine sand give smooth molding surfaces that do not require machine finishing. The sands may naturally contain enough clay minerals for good mold strength, or require additional binders, which increases costs.

Specifications for foundry sand grades of different refractoriness:

Beginning of sintering, K	> 1775	1575–1775	< 1575
Quartz, wt % (rest: feldspar and clay minerals)	>99	95–99	< 95

Glasses, Enamels, and Glazes [161, 167, 186, 200 – 204]. Different types of glasses and enamels require various grades of quartz raw materials. The silica content of glazes is mainly supplied by the silicate ingredients (feldspar, clay). If additional SiO_2 is required, quartz or flint powder is added.

Hydrothermal Growth of Quartz [159] (Section 1.7.1). Brazilian lascas were used worldwide as nutrient chips until the Brazilian quartz embargo in 1974. Since then, many other sources for high-purity low-cost lump quartz and granular nutrient were made accessible.

Metallurgical-Grade Silicon, Silicon Alloys and Silicon Carbide (→ Silicon; → Silicon Carbide). For chemical- or metallurgical-grade silicon, quartz lumps or pebbles (25–150 mm) are reduced with coke in electric-arc furnaces. The lumps and pebbles must not decrepitate during heating and therefore should preferably be monocrystalline and free of fluid inclusions. The grade of quartz raw material depends on the required grade of the products. Metallurgical-grade silicon with 99.0 wt % Si requires (in wt %): $\geq$ 99.0 SiO_2, $\leq$ 0.2 Al_2O_3, $\leq$ 0.1 Fe_2O_3, $\leq$ 0.1 MgO and CaO each, $\leq$ 0.02 TiO_2.

Low-grade silicon (98 wt % Si) for steel and aluminum production needs quartz with (in wt %): 98 SiO_2, $\leq$ 0.4 Al_2O_3, 0.2 Fe_2O_3, $\leq$ 0.2 MgO and CaO each, $\leq$ 0.1 P_2O_5, $\leq$ 0.002 TiO_2.

Production of 1 t metallurgical-grade silicon requires ca. 2.5 t of quartz. The worldwide silicon production is ca. 650 000 t/a.

Metallurgy. Quartz in the form of lumps or granules of vein quartz, quartzite, or sandstone is used as a flux in melting metal ores, for slagging iron oxides and basic oxides, and for balancing the silica/lime ratio of blast furnace burdens. In both cases, the SiO_2 content must be >90 wt % with lumps and granules in the size range of 0.8–2.5 cm.

For welding fluxes, quartz powder is used with a controlled grain size distribution of < 0.25 mm with an average of 0.06 mm for coated standard electrodes. The size of quartz particles in agglomerated fluxes must be < 0.06 mm.

Milling. Nodular chert (flint, silex) and very finely crystalline quartzite with > 96 wt % SiO_2 and homogeneous microstructure serve as linings for mills. Rounded flint nodules are preferably used as grinding pebbles.

Belgian silex consists of chalcedonic silica with evenly distributed calcite as secondary mineral phase and represents a material with high overall toughness. Yugoslavian silex has a coarser microstructure and sometimes contains coarse quartzite but is nevertheless of very good durability. Quartzite can only be used if it is free of accessory minerals. Translucent iron-free chalcedony from agates is used in ball mills and mortars for size reduction of high-purity quartz and silica glass.

Quartz for Optical and Oscillator Devices. Natural monocrystalline idiomorphic quartz is preferably used for optical devices as wafer plates, Brewster windows and prisms, birefringent filters, and tuning elements for laser optics. Only small amounts of man-made quartz are used in these applications. Oscillator quartz devices, however, are almost exclusively made from man-made quartz, for reasons of quality (low damping) and cost advantages.

Vitreous Silica. Transparent silica glass is melted from lascas or granular quartz with an impurity content < 30 ppm, the latter being by far the most important feed. Translucent to white opaque fused silica is produced from quartz sand with an impurity content of < 300 ppm. In both cases low contents of aluminum and iron impurities are essential.

A titanium content indicates the presence of submicroscopic rutile crystallites, which cause bubble problems in transparent silica glass during tube drawing. Silica glass for semiconductor uses, if not produced from high-purity silicon, must be very low in heavy metal impurities.

Well Fracturing. For hydraulic fracturing of oil, gas, or geothermal wells, a fluid with suspended quartz sand is pumped at high pressure into the well. Existing openings are enlarged and new voids are created. When the fluid is withdrawn, the remaining sand holds the fractures open. Frac sand must have >98 wt % SiO_2 and well-rounded grains to make placement easier and to provide good permeability. It must consist of single-crystalline grains which are clean and, in particular, free of adhering feldspar, clay, or carbonate minerals. The solubility in hydrochloric acid (a measure of the carbonate content) must be < 0.3 wt %. The content of soft particles must be < 1 wt %.

3. Diatomites

3.1. Introduction

The term diatomite (kieselguhr) refers to sedimentary rocks that are mainly composed of the skeletons of single- celled diatoms.

The low density, high porosity – and thus high absorption capacity for liquids – low thermal conductivity, and outstanding filtration properties make diatomite a versatile raw material [209 – 211].

Deposits with similar petrographic character or analogous physical properties, named according to their origin or area of application, include siliceous earth, "Bergmehl", "Saugschiefer", "Polierschiefer", and tripolite, not all of which are true diatomites.

The designation "tripolite" is derived from the ancient name *terra tripolitana* (after the city of Tripolis). The name referred mainly to minerals used as grinding and polishing agents, but also as lightweight building blocks. Today, the term tripolite refers to extremely fine-grained quartz pelites of inorganic origin [212].

3.2. Formation, Composition, and Quality Criteria [229 – 231]

Diatoms can be described as being almost "ubiquitous", occurring in marine, brackish, and freshwater environments.

The plant organisms possess a silicified membrane and vary in size between several micrometers and ca. 2 mm (generally 10–150 μm).

The skeletons are composed of opal-like, amorphous silica ($SiO_2 \cdot x\ H_2O$) and exhibit a wide range of porous fine structures and shapes. Including fossil forms, at least 15 000–20 000 different forms have been distinguished.

Rock-forming processes involving diatoms occurred worldwide from the Upper Cretaceous period onwards. Commercially exploitable deposits have been formed since the tertiary period.

The deposits are associated with sedimentation areas in which nutrients were abundant and the supply of silicates or dissolved silica was sufficient that a rich diatom flora developed and the frustules of the dead organisms accumulated to an appreciable thickness on the bed of the water body.

Such conditions are found worldwide, mainly in areas with volcanic activity, but also to a lesser extent in lakes with a largely decalcified environment and intensive chemical weathering of quartz sands or silicates (e.g., the German diatomite deposits) [212, 213].

Formation of diatomites of considerable thickness also occurred in marine coastal areas with continual subsidence, a periodical barrier to the open sea, and a regular supply of nutrients [214]. In addition, diatomites of brackish origin were formed in former saltwater lakes.

Important quality criteria for crude diatomite include the particle-size distribution, shape, and fine structure of the diatoms (Fig. 18). However, these can be altered with regard to their structural framework (and porosity) by calcination (see Section 3.4) [215, 216], which is important for the various applications as filtration agents. Furthermore, the chemical and mineralogical composition plays an important role, since higher contents of certain components can limit economically viable processing and refinement of the crude product.

Apart from water and organic substances the most frequent additional components of diatomites are quartz, calcium carbonate, and clay minerals.

The content of water and organic substances is only of minor significance for quality assessment, since they can be removed by drying or calcination.

Quartz sands or other rock fragments of similar particle size can be removed by air classification. In this respect, the conventional specification of the SiO_2 content of a crude diatomite is not very informative. For the assessment of a diatomite it is not the total SiO_2 content which is decisive, but the content of opal (diatom skeletons).

In contrast, the presence of fine-grained carbonates [$CaCO_3$, $CaMg(CO_3)_2$] in amounts greater than 5–10 mol % is problematic because they form complexes during processing which adversely affect the filtration properties and lower the chemical inertness.

Clay minerals restrict the range of applications, but are advantageous for the production of insulating or refractory bricks or for use as powders.

Higher iron contents, particularly as pyrite or markasite, require refinement by calcination (see Section 3.4). The color of the diatomites becomes reddish-gray or pink due to conversion into Fe_2O_3. The filtration properties are hardly influenced, whereas the use as fillers can be affected considerably by this coloration. In the lat-

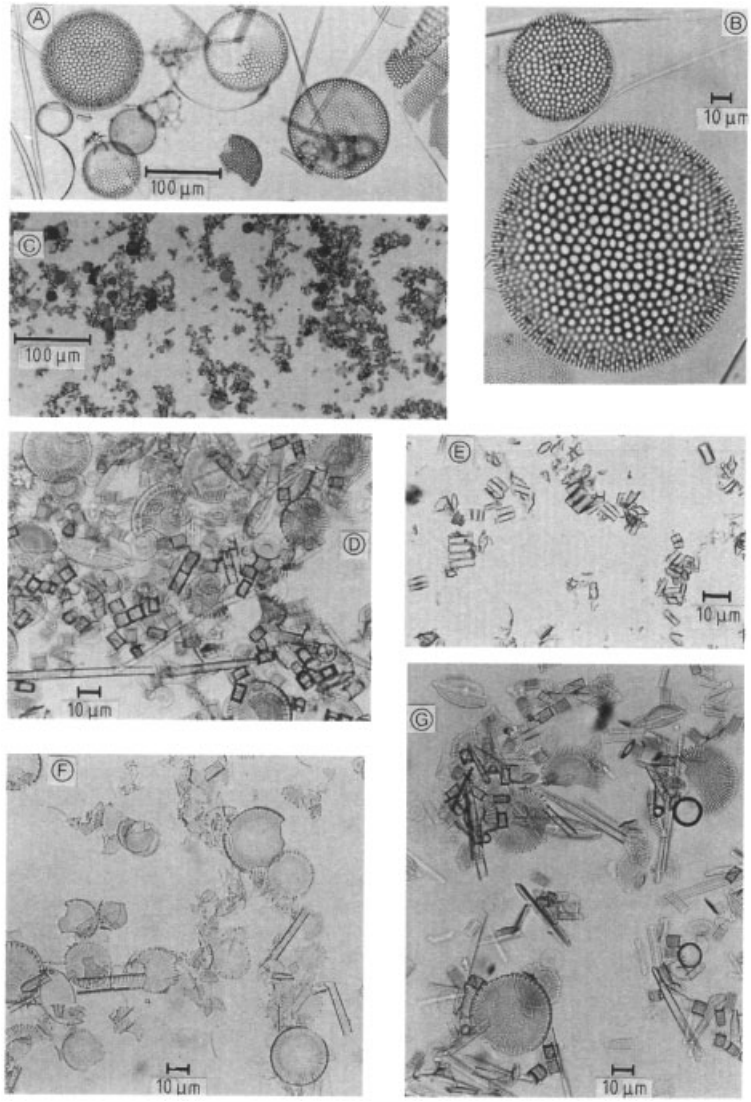

Figure 18. Light micrographs of various diatomites (diatom fraction)
A), B) Lompoc, California, United States; C) Myvatn, Iceland; D), E) Munster-Breloh, Germany; F) Hetendorf/Hermannsburg, Germany; G) Neuohe/Unterlüß, Germany

ter case, a Fe_2O_3 content of > 5–6 % can already be a limiting factor.

3.3. Occurrence and Mining

Diatomite deposits are distributed worldwide. Their economic exploitability and market potential depend on numerous factors such as mining conditions, and processing and transport costs. The size and quality of the deposit, exploitation costs, transport conditions as well as the question whether the diatomite is for a domestic market or whether it is to be exported play a significant role. In 1995, the world demand for diatomite was ca. 1.2×10^6 t/a [217].

The largest diatomite deposits in the world are found in the United States in the district of Lompoc, California. These are upper Tertiary (Miocene) marine deposits with an economically exploitable diatomite thickness of ca. 300 m. In addition, large quantities of marine

and freshwater diatomites are mined in Georgia, Mississippi, Nevada, Oregon, and Washington. In 1986, the United States accounted for about one-third of world production (575 000 t), followed by Rumania, the CIS, and France, with outputs of $(246–272)\times10^3$ t (1986). The production in Rumania and the CIS is almost entirely used in the building construction industry. Further, noteworthy deposits and production sites are located in Algeria, Argentina, Australia, Brazil, Germany, Iceland, Italy, Kenya, Korea, Mexico, Peru, and Spain, whereby in several of these countries the export of diatomite is of no significance [218]. Danish "Moler" is a clay-rich diatomite (diatomaceous earth), which is produced and exported in considerable quantities (63 500–72 600 t/a). It is used predominantly for the production of insulation materials (e.g., fire-resistant insulating bricks).

The only large, economically workable diatomite deposits in Germany are freshwater formations of Pleistocene age. They are situated on the edge of the Lüneburger Heide in the area of Hetendorf/Hermannsburg, Neuohe/Unterlüß, and Munster-Breloh as well as in the Luhetal (Bispingen/Hützel/Schwindebeck). Here industrial usage began with the establishment of the first open- cast mine in 1863. Thus, German diatomite mining represents the earliest modern exploitation of this raw material. Up to World War I almost the total world requirement could be met, with an export quota of $> 30\,\%$.

Diatomite is generally mined in open- cast mines, occasionally also by subsurface or underwater mining (Myvatn, Iceland). Mining by hand is now rare (Third World countries). Almost everywhere, modern, cost-effective mining methods are used, with excavators and wheel/front loaders, and trucks or conveyor belts for transport.

3.4. Processing

Worldwide, the processing methods for crude diatomite differ only slightly from one another. The most important principle is minimization of mechanical damage to the structure of the diatom skeletons. Hollow forms should not be destroyed and pores not blocked (e.g., by sintering during calcination). The high porosity, the most important property of diatomite, must be retained. The processing steps are shown in Figure 19.

Preliminary size reduction of crude material is carried out with spiked rollers, charging boxes, or sieve kneaders to give a uniform grain size that ensures even heat transfer during further treatment. This is particularly important because of the low thermal conductivity of diatomite.

The following types of processed diatomite are produced.

Dried Diatomite [*61790-53-2*]. Depending on the climatic conditions at the production site, the crude diatomite from the deposits is either stored in the open for predrying or transported immediately to storage hoppers. Crude diatomite contains 30–65 % water, which is removed in countercurrent driers. Simultaneous grinding and drying is also practiced, whereby preheated air is blown into the grinders. However, this is only possible for crude products with a low moisture content. Drying is followed by gentle grinding and screening. Granulation gives products that are used as absorbents (e.g., for oil) or as pet litter.

If the starting material is of high purity and has a low content of organic substances, the dried diatomite can also be used as a filtration agent or catalyst support. These diatomites vary in color from white to yellowish or gray-brown, depending on their origin, and consist predominantly of amorphous silica with a low crystalline content (generally $< 3\,\%$).

Calcined Diatomite [*91053-39-3*]. To remove organic substances and alter the particle size, dried diatomite is calcined in gas- or oil-fired rotary kilns at 800–1000 °C. During calcination, the diatom frustules and their fragments are hardened and partially agglomerated by sintering. The degree of sintering can be controlled by altering the temperature and duration of the process. The diatomite is then ground and classified by cyclones and sorters into grades with different particle-size distributions (fine and medium calcined diatomites). Iron impurities, present as Fe_2O_3 after calcination, color the products pink or yellowish to dark brown.

The old-fashioned furnaces with several decks, similar to pyrite-roasting kilns, are still used to a limited extent. They have the advantage of a slow calcination process at relatively low

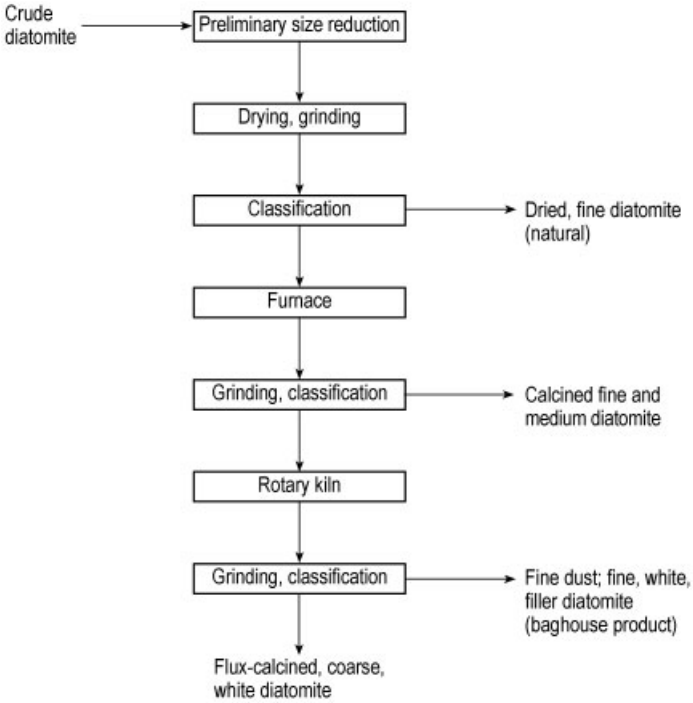

Figure 19. Diatomite processing

temperature which causes little damage to the structure. The kilns are loaded with moist crude diatomite, which is dried on the upper decks, calcined in the middle, and cooled on the lower decks.

Fluidized-bed drying and calcination processes were tested as early as the 1920s, but were unsuccessful due to technical problems. Nowadays modern fluidized-bed technology is also used for the processing of diatomite. Advantages include energy savings, short production times, mild treatment of the diatomite, and more compact design of the plants.

Flux-Calcined Diatomite [*68855-54-9*]. In this process, dried or calcined diatomite powder is calcined with the addition of alkaline fluxes. Sodium carbonate is generally used as flux in amounts of 1–6 %, either mixed with the diatomite before calcination or continuously added during calcination. The calcination temperature is 1000–1200 °C. The addition of sodium carbonate leads to the formation of a sodium silicate glass melt which binds the diatom frustules and their fragments into agglomerates. Since the iron contained in the diatomite is also bound, a white product is formed. The degree of agglomeration can be controlled by varying the temperature, the type and amount of flux, and the reaction time. This process is also known as activation. The high calcination temperatures in the presence of alkali lead to a partial transformation (up to 65 %) of the amorphous silica into cristobalite, a crystalline silica modification. The porosity and the specific surface area are strongly decreased. The annealing and cooling processes are followed by grinding and classification, to give white filter diatomite (the main product) and a dust-like fine fraction which is used as white filler and auxiliary.

Commercially Available Forms, Products, and Producers. Powder diatomite is sold in paper sacks containing 15–25 kg as well as loose in big-bags and silo trucks. Granulated material is available in paper or plastic sacks containing 30 kg, sometimes also in big-bags or silo trucks.

Examples of commercial products are:

Celatom (Eagle-Picher Minerals Inc., Reno, United States)

Celite (Celite Corp., Newark, United States)
Clarcel (Ceca SA, Paris, France)
Dicalite (Grefco Inc., Torrance, United States)
Fina/Optima (Meyer-Breloh, Munster, Germany)
Skamol (Skarrehage Molervaerk a/S Nykøbing, Mors, Denmark)

Table 11 lists the chemical and physical properties of several commercially available diatomites.

3.5. Analysis

Chemical and physical testing begins with the crude diatomite in the mine, continues during the individual production stages, and finishes with the test certificate of the end product. The results of the crude diatomite tests determine the working plan for the different regions of an open- cast mine.

The following analyses and tests are undertaken:

Chemical Analysis. The contents of silica, iron, aluminum, calcium, magnesium, manganese, titanium, sodium, potassium, and, depending on requirements, also other elements are determined after alkaline dissolution or acidic extraction. Filter diatomites for breweries are also tested for beer-soluble iron, calcium, and sometimes other elements [219].

Physical Tests. Water content, loss on ignition (at 1050 °C), bulk density, tapped density, pH value (10 % in water), water absorption capacity, and specific surface area are determined by conventional methods [219]. The particle-size distribution is determined by laser scattering or air-jet screening. For the fine fraction, the sedimentation method is used. The wet density is the density in g/L of a filter cake flushed with water. Since the majority of diatomite is used for filtration, the determination of the permeability is of major importance since it determines the performance and the retaining power of a filter diatomite. In general, the methods developed by the individual producers in which the rate of permeation of water is measured, are sufficient. The company CECA (France) possesses a patent for a permeameter which allows fully automatic measurement. The EBC filter is of simpler construction and allows Darcy values to be precisely measured [220]. Apart from the relative units used by the diverse producers, the Darcy has become generally accepted as the unit for the permeability of a filter agent. For filter materials with the permeability of 1 Darcy the flow rate of a liquid of viscosity of 1 mPa · s is 1 mL · s^{-1} · cm^{-2} filter area for a filter- cake thickness of 1 cm and a pressure difference of 0.1 MPa. Since the permeability can be determined rapidly, this is a suitable method for differentiating between fine, medium, and coarse diatomites [219, 221 – 223].

The amorphous silica content of a diatomite is determined by the alkali solubility. This property is of significance for catalyst diatomites, or for diatomites which serve as a silica source in cement, or as chemical reagents in other silica-containing products. The specific surface area is used to assess the porosity of a diatomite. It is generally determined by nitrogen adsorption (BET method). The degree of whiteness is measured with a leucometer.

3.6. Storage and Transport

Processed diatomite is stored in steel or stainless steel silos, mostly with a circular cross section and equipped with various discharge aids, since diatomite tends to form bridges. Material packed in sacks is stored and transported on wooden palettes in amounts of 600–1000 kg enveloped by a shrink film, and also in big-bags made from plastic- coated webbing with a capacity of 600–1500 kg. Loose material is transported in silo road vehicles or rail containers with a capacity of 30–75 m^3. The diatomite is conveyed pneumatically or with screw and vibrational conveyors.

3.7. Environmental and Health Protection

Diatomite rarely contains toxic compounds. Since, however, it can contain crystalline modifications of SiO_2 in addition to amorphous silica, dust safety measures are of special significance. This is especially true of white activated

Table 11. Chemical and physical properties of commercially available diatomites*

Property	1	2	3	4	5	6	7	8	9
Color	white-gray	yellow-brown	pink	yellow-brown	brown	white	white	white	yellow-brown
SiO_2, %	89.0	72.5	90.7	87.5	86.0	89.5	90.7	91.5	90.2
Al_2O_3, %	3.5	7.1	3.9	4.3	2.8	4.1	3.9	1.6	2.8
Fe_2O_3, %	0.9	5.0	1.4	2.9	4.7	1.6	2.1	0.7	2.5
CaO, %	1.1	1.2	0.5	1.9	0.6	0.5	1.0	4.4	0.7
Na_2O, K_2O, %	0.8	1.4	0.9	0.8	0.7	3.6	3.5	1.9	0.9
Ignition loss, %	2.0	4.7	0.5	0.7	0.3	0.2	0.1	0.1	0.4
Bulk density, g/L	107	290	120	140	125	229	200	195	209
pH value	7.0	5.2	7.5	6.9	7.0	10.0	9.7	9.5	6.7
Water uptake, %	255	200	250	205	201	156	160	200	196
Specific surface area, m^2/g	19.2	25.4	15.2	13.0	16.1	1.9	1.6	3.0	10.6
Average particle size, µm	14.2	19.3	15.9	14.1	13.9	22.5	30.1	6.5	14.7
Wet density, g/L	228	280	271	255	209	297	290	350	357
Permeability, Darcy	0.06	0.09	0.28	0.09	0.08	1.20	1.60		0.08
Crystalline content, %	2.0	2.2	7.6	9.2	9.8	58.1	59.7	62.7	10.3

* 1) American filter diatomite, dried; 2) Danish filler diatomite, calcined; 3) American filter diatomite, calcined; 4) French filter diatomite, calcined; 5) German filter diatomite, calcined; 6) American filter diatomite, flux calcined; 7) French filter diatomite, flux calcined; 8) Spanish filler diatomite, flux calcined, very fine; 9) German regeneration filter diatomite, calcined.

diatomites, which can contain up to 65 % crystalline cristobalite. The MAK value for uncalcined diatomite is 4 mg/m^3 and for calcined diatomite, 0.3 mg/m^3. The TLV for uncalcined diatomite is 10 mg/m^3. The International Agency for Research on Cancer (IARC) published a monograph concerning crystalline silica in 1987 [224]. This summarizes the research results concerning the influence of crystalline silica on humans and concluded that apart from causing silicosis it may also cause cancer (see also Chap. 8). Hence, crystalline silica is classified in group 2 A. As a result of this, the diatomite producers founded the International Diatomite Producers Association (IDPA), whose aims include informing personnel working in the diatomite industry and the consumer about these hazards, and to ensure adherence to the work protection regulations by increasing the information available [225, 226]. The regulations of individual countries concerning the handling of dusts containing crystalline silica differ widely:

United States: 29th Federal Statute book 1910.1000 from 19. 01. 1989, which classifies crystalline silica as a hazardous substance and specifies tolerance levels.

EC: Directive 67/548/EEC from 18. 09. 1979 concerning the classification, packaging, and labeling of hazardous substances.

United Kingdom: Control of substances which damage health No. 1857 from 1988.

Germany: UBG 119–4.86 Quartz: protection against dust dangerous to health; UBG 100–4.85, rule G.1.1.: concerning the health precautions; GefStoffV 8.86: labeling regulations.

France: No. 50.1289 from 16. 10. 1950 changed into No. 63 576 from 11. 06. 1963 concerning medical precautions against occupational silicosis. No. 11 453 from 19. 07. 1982 determines the dust concentration in the air of the work place. No. 87–200 from 25. 03. 1987 regulates the content of safety data sheets for dangerous substances. L 231–6 from 10. 10. 1983 lists dangerous substances and determines the packaging and labeling regulations.

Spain: from 27. 11. 1985 classification and labeling of dangerous substances.

Italy: No. 256 from 29. 05. 1974, No. 927 from 24. 11. 1981, No. 141 from 20. 02. 1988 concerning the classification, packaging, and information for hazardous materials.

3.8. Uses

The main uses of diatomite are the following: 60 % of world production is used as a filtration agent; ca. 25 % as filler diatomites, including carrier materials; and ca. 15 % for other uses.

Filtration Agents. Diatomite is used in both pressure and vacuum filters of various designs. Fine, medium, and coarse diatomites, as well

as mixtures thereof, are employed, depending on the desired clarity and flow rate [227]. Coarse diatomites are used for precoat filtration in pressure filters in amounts of ca. 1 kg/m^2 of filter area. Fine and medium diatomites are added to the slurry to be filtered, usually in amounts of 100–200 g/m^2 filter area (→ Filtration, Chap. 11). For use as filter aids diatomites must fulfil the following requirements:

1) Good preservation of the diatom structure
2) Optimal particle-size distribution
3) Low content of soluble components
4) High volume and therefore high capacity for materials causing turbidity
5) Formation of filter cakes with high permeability, low compressibility, and low tendency to form cracks

The most important uses of diatomite are for the filtration of beer, wine, fruit juice, sugar cane juice, swimming pool water, solvents for chemical cleaning, wastewater, varnishes, and paints etc. Large quantities of diatomite are used for the production of cellulose-based filter beds, filter candles, and special filter papers.

Fillers. Diatomite is used as filler and auxiliary, for example, to modify the rheological properties of polymers. It increases the thermal stability of bitumen, the Shore A hardness of silicone rubber, and is a reinforcing filler in plastics, rubber, and adhesives. Dust-like, white, activated diatomite is used as a delustering agent, for adjusting the viscosity of paints, and as an antiblocking agent for plastic films.

Insulators. Loose diatomite has long been used for insulating double-walled kilns made from fireclay bricks, as well as for producing insulating sheets and bricks. Danish Moler is especially well suited for producing insulating bricks, since its clay content functions as a ceramic binder. In the VSI process (vacuum super insulation), a special form of diatomite acts as a supporting and insulating layer in double-walled elements that contain a high vacuum. Thus, extremely high insulation values are achieved. The VSI process is being tested for long-distance heat transport.

Absorption Agents. Due to their high capacity for liquids, diatomites are used to produce gas purification agents as well as absorption agents such as cat litter and drying agents. Diatomite is also used to ensure the flowability and to prevent clotting of foodstuffs, fire extinguisher powders, and seeds.

Other Uses. Diatomite serves as a fine scourer in polishes and cleaners such as car polishes, toothpastes, and silver polishes. Diatomite is employed as a catalyst support, for the production of pyrotechnics and matches, as a packing material for the transportation of hazardous liquids, and for the filling of acetylene bottles. In some countries it is used as the silica source for the production of cement and calcium silicate.

3.9. Recycling

Attempts to regenerate diatomite have been undertaken for many years [228]. This is only economically viable for diatomites loaded with organic residues, which can be purified by leaching or calcination. In Germany, the company Tremonis operates a thermal process for recycling filter residues from breweries. The flow diagram of the plant is shown in Figure 20. The regenerated diatomite is classified as fine diatomite. It is added to the fresh diatomite for beer filtration in quantities of up to 50 %.

4. Colloidal Silica

4.1. Introduction

Silica sols are stable disperse systems in which the dispersion medium (or continuous phase) is a liquid and the disperse or discontinuous phase is silicon dioxide in the colloidal state of subdivision. This state comprises particles with a size sufficiently small (< 1 µm) not to appear affected by gravitational forces but sufficiently large (> 1 nm) to show marked deviations from the properties of typical solutions (see → Colloids, Chap. 4).

The limits given for the colloidal size range are not rigid since they depend to some extent on the properties under consideration [232].

Silica sols are commonly known as *colloidal silica.* Although the ultimate particles that constitute silica gels, xerogels, cryogels, aerosols,

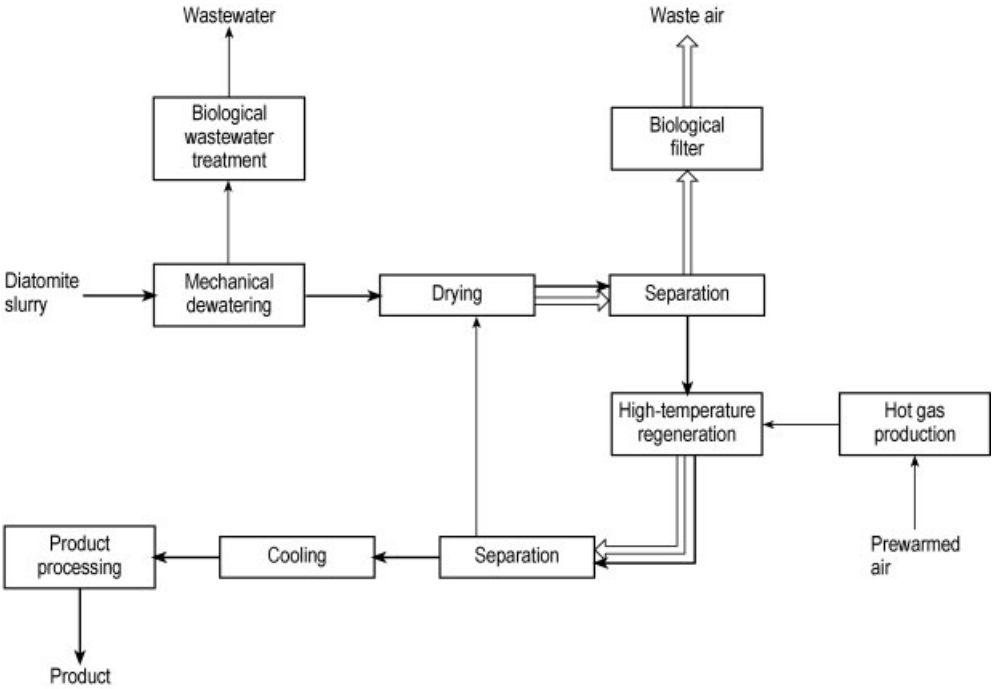

Figure 20. Flow sheet of thermal diatomite regeneration (Tremonis GmbH, Dortmund, Germany)

and pyrogenic (fumed), precipitated, and coacervated silica have lost their mobility by aggregation. They are also in the colloidal range of particle size and are therefore known as colloidal silicas. However, in this chapter the terms silica sol and colloidal silica are used as synonyms. Figure 21 illustrates the relationship between silica sols, gels, and powders [233].

In commercial silica sols the disperse silica is amorphous and the dispersion medium in most cases is water (aquasols or hydrosols). Dispersions in organic solvents are also commercially available (organosols). Silica sols are fluid and stable toward gelation and settling. Most commercial sols are close to monodisperse and consist of dense discrete spheres with a diameter range between ca. 4 or 5 nm and 100 nm. Also commercially available in the form of sols or powders are grades with particles between 0.1 and ca. 1.5 µm. Particle size, particle-size distribution, and concentration of solids determine the appearance of sols. Silica sols look milky if the particle size is large and the concentration is high, opalescent if the size is intermediate, or clear and almost colorless if the particles are of the smallest size range.

Although there are earlier references to what are now known as silica sols, it was Graham [234] who coined the term *colloidal* in 1862 to refer to products such as the one he obtained by reacting acid with silicate and removing the electrolytes by dialysis. Stable and relatively concentrated silica sols were not available until the 1930s when I. G. Farbenindustrie first made 10 wt % ammonia-stabilized silica sols [235]. However, the real breakthrough in colloidal silica technology came in 1941 when Bird [236] patented a process for removing the alkali from a dilute solution of sodium silicate by ion exchange. The next landmark in the development of concentrated silica sols was the first process for making colloidal silica particles of uniform and controlled size reported in 1951 by Bechtold and Snyder [237].

By 1990 colloidal silica constituted a growing market valued at an estimated \$ 50×10^6 in North America [238] and € 15×10^6 in Europe [239]. Applications of silica sols are based on characteristics such as particle size, high specific surface area that gives them good binding ability, stability towards gelation and settling, and surface properties. These characteristics enable

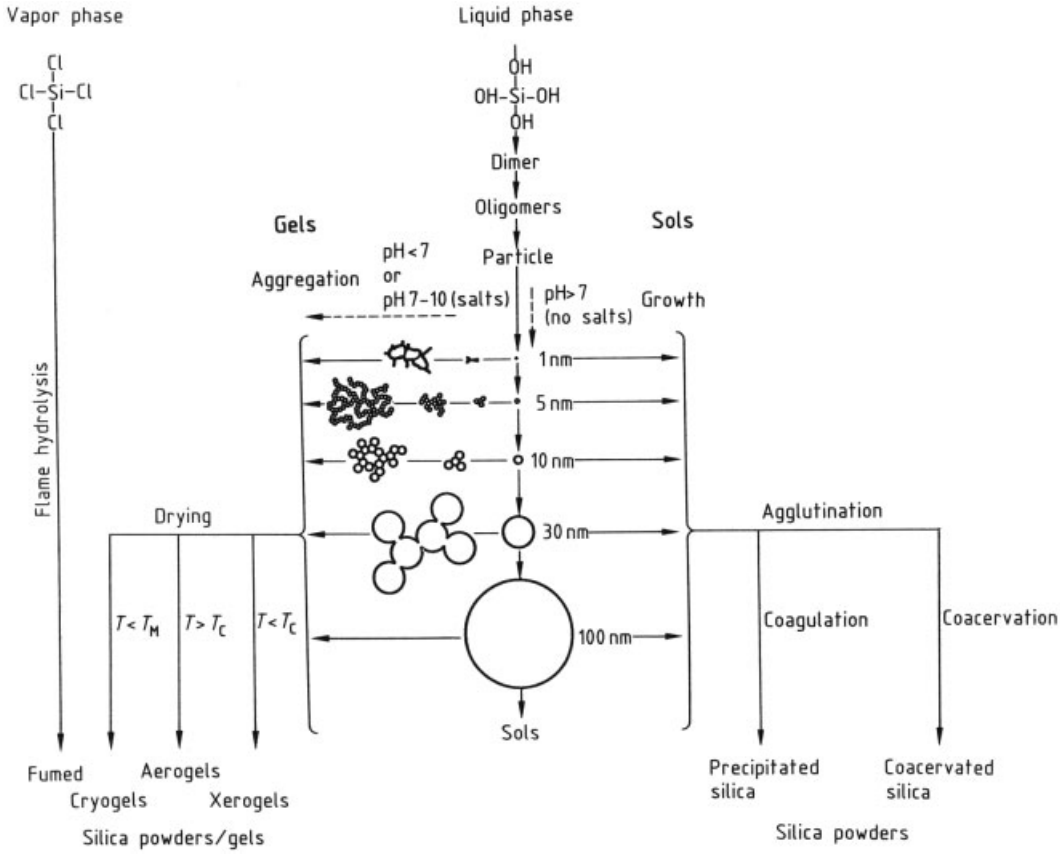

Figure 21. Formation of silica sols, gels, and powders by silica monomer condensation–polymerization followed by aggregation or agglutination and drying [233]
Growth of nascent colloidal particles with decrease in numbers occurs in basic solutions in the absence of salts. In acid solutions or in the presence of flocculating sols the colloidal silica particles form gels by aggregation into three-dimensional networks. Aerogels are made by drying wet gels under supercritical conditions, that is, above the critical temperature T_C and critical pressure of the liquid. Cryogels are made by drying wet gels below the melting temperature T_M of the liquid. Fumed or pyrogenic silicas are formed at high temperature by flame oxidation–hydrolysis of silicon halides

colloidal silica to be used in a wide variety of applications. Major uses are in investment casting, silicon-wafer polishing, and fibrous ceramics.

4.2. Structure of Colloidal Silica Particles

The building block of silica is the SiO_4 tetrahedron, four oxygen atoms at the corners of a regular tetrahedron with a silicon ion at the central cavity or centroid [240] (see Chap. 1). The oxygen ion is so much larger than the Si^{4+} ion that the four oxygens of a SiO_4 unit are in mutual contact and the silicon ion is said to occupy a tetrahedral hole [240]. In amorphous silica the bulk structure is determined, as opposed to the crystalline silicas, by a random packing of $[SiO_4]^{4-}$ units, which results in a nonperiodic structure (see Fig. 22). As a result of this structural difference amorphous silica has a lower density than the crystalline silicas: 2.2 g/cm^3 versus 3.01, 2.65, 2.26, and 2.21 g/cm^3 for coesite, α-quartz, β-tridymite and β- cristobalite, respectively.

Figure 23 represents a two-dimensional random network of a dehydrated but fully hydroxylated amorphous silica particle.

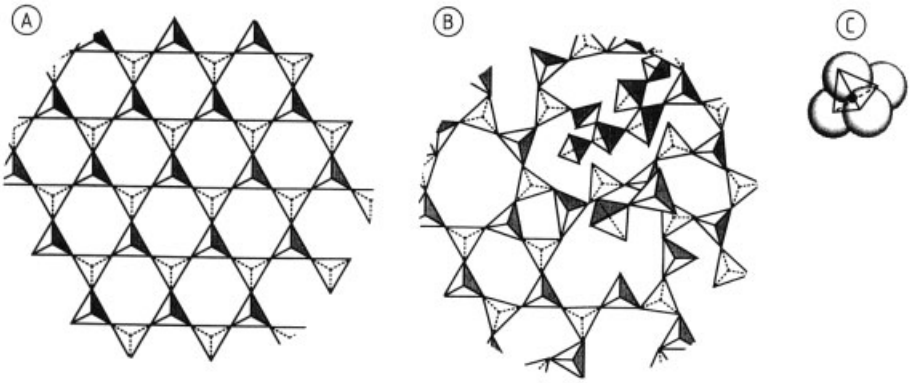

Figure 22. Tetrahedral configuration of crystalline (A) and amorphous (B) silica; (C) represents a silicon atom (filled circle) coordinated by four oxygen atoms (large atom spheres) [241]

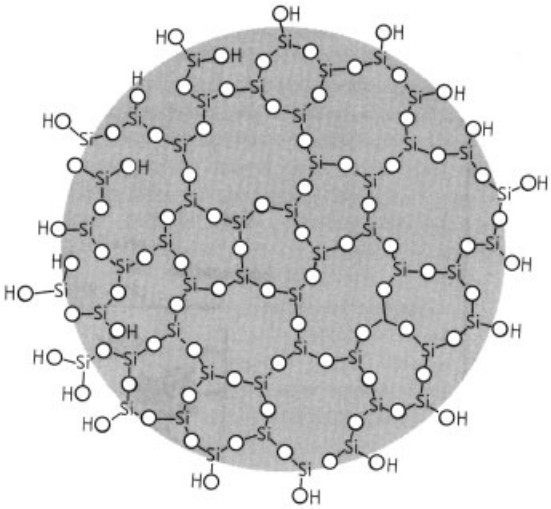

Figure 23. Two-dimensional random network of a dehydrated but hydroxylated colloidal silica particle
Four rules apply [242]: each oxygen ion is linked to no more than two silicon ions. The coordination number of oxygen ions about the central Si cation is 4 (the fourth oxygen not represented in the figure is directly above or below the Si cation). Oxygen tetrahedra share corners, not edges or faces. At least two corners of each tetrahedron are shared. Since the particle is fully hydroxylated each surface Si ion is bonded to one or two hydroxyl ions: the silanol number is 4.6 OH/nm^2. Depending on the method of formation some internal Si ions may be linked to hydroxyl ions. Surface of occluded alkali-metal ion impurities may replace surface or internal protons. Aluminum ion impurities or added modifiers may replace internal or surface tetrehedral Si. The particle surface may be esterified, silanized (silylated), or ion exchanged. The concentration of OH groups on the surface decreases monotonically with increasing temperature when silica is calcined

Impurities such as Na, K, or Al, picked up during synthesis of silica aquasols in alkaline media, may be occluded inside the colloidal particles, replacing internal silanol protons (Na, K) or forming isomorphic tetrahedra (Al) with an additional negative charge on the surface or inside the particles (see Fig. 24). Sols obtained by hydrolysis of alkoxysilanes or by dispersion of pyrogenic silica in water or an organic solvent are pure and substantially free of alkali metals or aluminum.

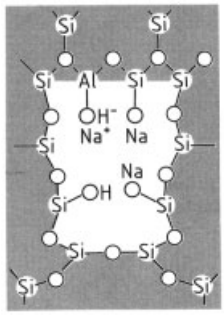

Figure 24. Internal structure of an amorphous silica particle showing an internal silanol group and occluded Na and Al ions

Many of the adsorption, adhesion, chemical, and catalytic properties of colloidal silica depend on the chemistry and geometry of its surface. In 1934 Hofman [243] postulated the existence of silanol (Si–OH) groups on the silica surface. It is now generally accepted that surface silicon atoms tend to have a complete tetrahedral

configuration and in an aqueous medium their free valence becomes saturated with hydroxyl groups forming silanol groups. Under appropriate conditions, silanol groups in turn may condense to form siloxane bridges: ≡Si−O−Si≡.

In the meantime most of the following postulated groups have been identified on the surface or in the internal structure of amorphous silica (see Fig. 25).

1) Single silanol groups, also called isolated or free silanol groups
2) Silanediol groups, also called geminal silanols
3) Silanetriols, postulated but existence not yet generally accepted [244]
4) Hydrogen-bonded vicinal silanols (single or geminal), including terminal groups
5) Internal silanol groups involving OH groups, sometimes classified as structurally bound water (see Fig. 25 B)
6) Strained and stable siloxane bridges
7) Physically adsorbed H_2O, hydrogen-bonded to all types of surface silanol groups (see Fig. 25 B)

Silanol groups are formed on the silica surface in the course of its synthesis during the condensation polymerization of $Si(OH)_4$ or by rehydroxylation of thermally dehydroxylated silica with water or aqueous solutions.

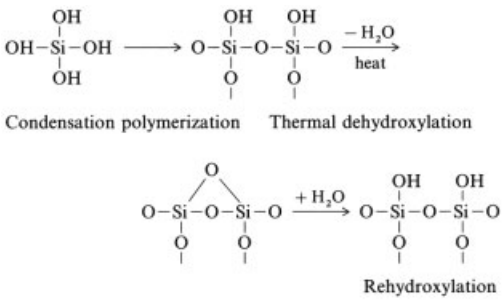

The silanol groups on the silica surface may be classified according to their nature, multiplicity of sites, and type of association [245].

An isolated silanol has an OH group sufficiently remote from neighboring hydroxyl groups that hydrogen bonding cannot occur (≥ 0.33 nm). A silicon site of this kind is designated as Q^3 in NMR Q^n terminology, where n is the number of bridging oxygens bonded to the central silicon. The isolated silanol shows a sharp band at around 3750 cm^{-1} in the IR spectrum [246, 247].

Geminal silanols are silanediol groups located on Q^2 silicon sites. Their existence was postulated by Peri [246 – 248] but only confirmed experimentally with the advent of solid-state ^{29}Si cross-polarization magic angle spinning nuclear magnetic resonance (CP MAS NMR) [249].

Vicinal or hydrogen-bonded or associated silanols are SiOH groups located on neighboring Si atoms such that the OH - - - O distance is sufficiently short for hydrogen bonding to occur. Hydrogen bonding causes a reduction in the O−H stretching frequency, the magnitude of which depends on the strength of the hydrogen bond and thus on the O−H−O distance [250]. The characteristic IR band of vicinal groups occurs at about 3660 cm^{-1} [251].

Geminal Q^2 silanol sites bonded to a neighboring Q^3 silicon through a single siloxane bridge also result in a hydrogen-bonded pair. The remaining OH group experiences very weak hydrogen bonding.

Silanol groups are also found within the structure of the colloidal particles. These groups are designated as internal silanols and in some cases are erroneously referred to as structurally bound water. The concentration of internal silanol groups depends on the synthesis temperature and other variables.

Surface and internal silanol groups may condense to form siloxane bridges. Strained siloxane bridges are formed on the hydroxylated silica surface by thermally induced condensation of hydroxyl groups up to about 500 °C. At higher temperature, the strained siloxane bridges are converted into stable siloxane bridges [250].

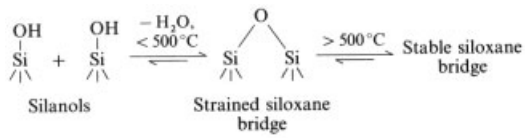

Strained siloxane groups undergo complete rehydroxylation on exposure to water in a matter of hours or a few days, depending on type of silica powder and conditions of exposure [251]. Stable siloxane groups are rehydroxylated at a

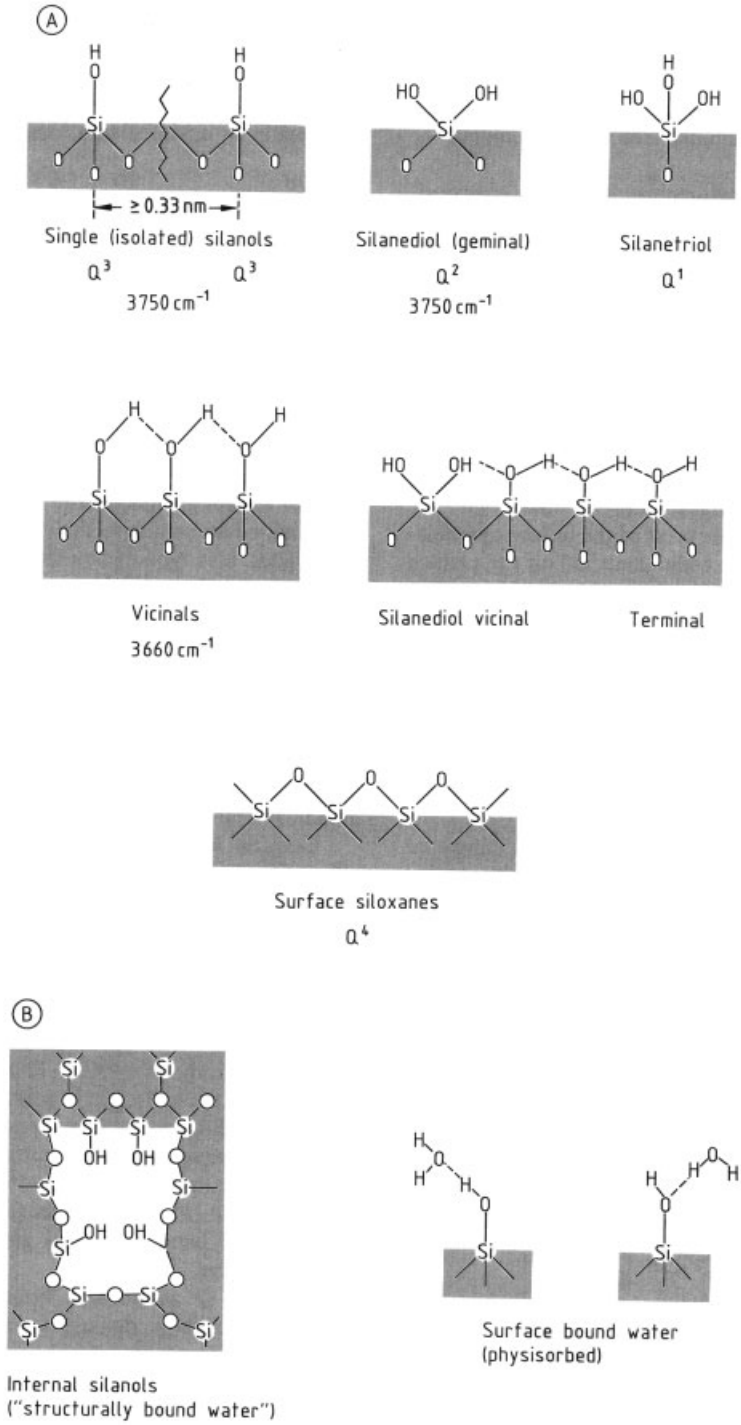

Figure 25. A) Possible types of silanol groups and siloxane bridges occurring on the surface of colloidal silica particles Characteristic bands in the infrared spectrum and Q^n site designation is included (n is the number of bridging oxygens bonded to the central silicon atom); B) Surface bound water and internal silanols

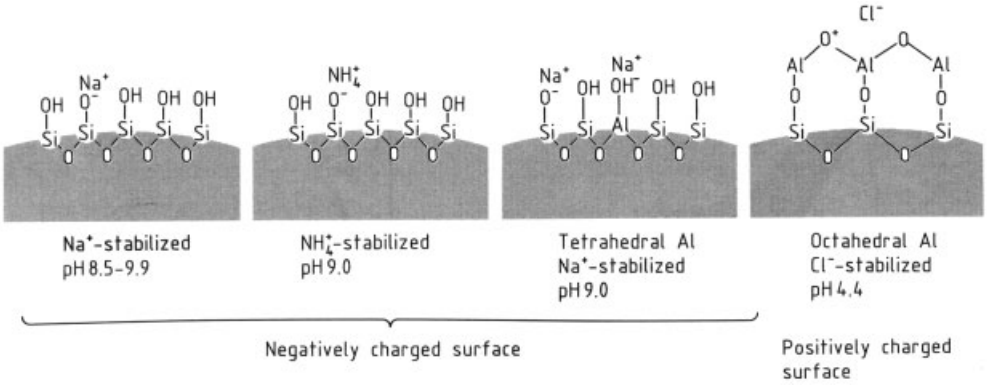

Figure 26. Modified colloidal silica surface

slower rate. For example, a wide-pore sample of 340 m^2/g calcined in air at 900 °C required five years of contact with water at room temperature for complete rehydroxylation [252].

Surface silanol groups are the main centers of adsorption of water molecules [252]. Water can associate by hydrogen bonding with all types of surface silanols, and in some cases with internal silanol groups.

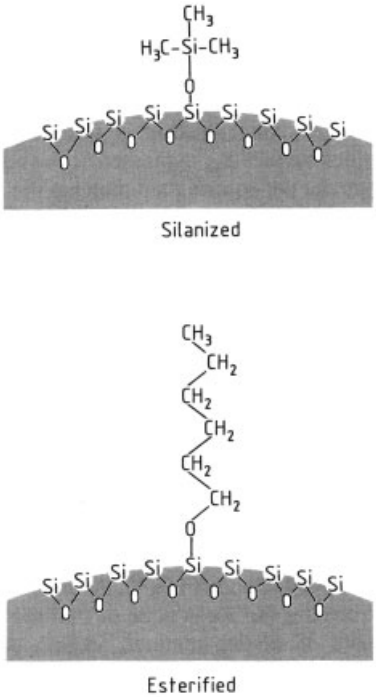

Figure 27. Silylated and esterified silica surface

Surface silanol groups of silica aquasols stabilized in an alkaline medium exchange protons for ions such as Na^+, K^+, or NH_4^+ (Fig. 26). The surface silanol groups can be esterified, as in the case of silica organosols, or silanized (silylated) (Fig. 27). Derivatization, that is modification of the silica surface, is the basis for the use of silica in analytical and process chromatography.

The surface charge of pure silica aquasols is negative at a pH higher than ca. 3. The isoelectric point (IEP) is ca. 2. The IEP can be shifted by partial substitution of surface silicon groups by tetrahedral aluminum (see Fig. 26). In this way the pH range of stability of aquasols may be varied. The surface charge can be reversed by adsorption of octahedral aluminum ions such as those present in basic aluminum chloride [253] (Fig. 26). Aquasols obtained in this manner are commercially available in various particle sizes and concentrations as *positive sols.*

The concentration of silanol groups on the silica surface α_{OH} expressed as the number of OH groups per square nanometer is the silanol number [252]. For a dehydrated but fully hydroxylated amorphous colloidal silica powder α_{OH} is 4.6 [252].

The threshold temperature corresponding to the completion of dehydration, that is removal of physisorbed water from the silica surface, and the beginning of dehydroxylation by condensation of surface OH groups, is estimated to be 190 $\pm$ 10 °C [252].

The concentration of OH groups on the surface decreases monotonically with increasing temperature when silicas are heated under vacuum (Fig. 28). Most of the physisorbed water is

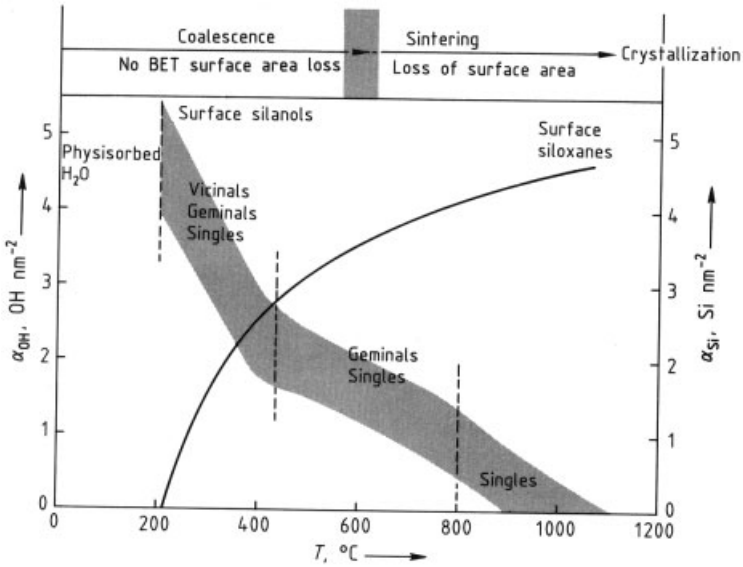

Figure 28. Effect of temperature on silanol groups on the surface of colloidal silica

removed at about 150 °C. At 200 °C all the water from the surface is gone so that the surface consists of single, geminal, vicinal, and terminal silanol groups and siloxane bridges. At ca. 450–500 °C all the vicinal groups condense yielding water vapor and only single, geminal, and terminal silanol groups and strained siloxane bridges remain. The estimated ratio of single to geminal silanol groups on the surface is 85/15 [252] and is believed not to change with temperature, at least up to ca. 800 °C. Internal silanols begin to condense at ca. 600–800 °C and in some cases at lower temperatures. Complete evolution of internal silanols occurs at higher temperatures. It appears that at temperatures from ca. 800 °C to ca. 1000–1100 °C, only isolated (single) silanol groups remain on the silica surface.

At a sufficient surface concentration, the OH groups make the silica surface hydrophilic, whereas a predominance of siloxane groups on the silica surface makes the surface hydrophobic.

4.3. Physical and Chemical Properties

Most commercial silica sols consist of dense, discrete, spheroidal particles typically 4–60 nm in diameter and amorphous to X-ray or electron diffraction. Because they are amorphous the silica particles are only slightly soluble in water (100–150 ppm at 25 °C and pH 2–8) and insoluble in alcohol or in inorganic acids except HF. They are soluble in organic bases such as tetramethylammonium hydroxide, forming water-soluble quaternary ammonium silicates, and, as opposed to crystalline silicas, are very soluble in hot sodium hydroxide solution. Specific surface areas vary with particle size, and for the common grades of silica sols are between 50 and 750 m^2/g. Less typical grades have particle diameters between 60 nm and 1.5 µm with specific surface areas much lower than 50 m^2/g, depending on their particle diameter. The specific surface area A in m^2/g of a system made of solid monodisperse amorphous silica spheres with a smooth surface and bulk density 2.2 g/cm^3 is:

$$A = 6000/2.2\,D$$

where D is the diameter of the spheres in nanometers.

Table 12 lists typical properties of representative commercial silica aquasols. The table also shows the dependence of relative density and viscosity on the silica content. Viscosity also depends on particle size and degree of cross-linking of the particles. In 30 % silica aquasols it is generally < 10 mPa · s.

Silica sols obtained by dispersion of pyrogenic silica powders generally differ from those

Table 12. Typical properties of commercially available negative and positive silica aquasols* [254]

	Particle diameter, nm	Specific surface area, m^2/g	Silica, wt %	pH	Na_2O, wt %	Viscosity at 25 °C, mPa · s	Relative density at 25 °C
Negatively charged surface	4	750	15	10.4	0.80	18	1.10
sodium stabilized	7	360	30	9.9	0.56	6	1.22
	12	230	40	9.7	0.41	20	1.31
	21	130	50	9.0	0.21	35	1.40
	60	50	50	8.5	0.25	10	1.40
	50–80	NA	40	9–10	NA	NA	NA
	70–100	NA	40	9–10	NA	NA	NA
	100	NA	20	8–10	NA	NA	NA
	300	NA	20	8–10	NA	NA	NA
	500	NA	20	8–10	NA	NA	NA
alumina modified surface	12	230	30	8.9	0.24	11	1.21
ammonium stabilized	21	130	40	9.0	0.08	9	1.30
Positively charged surface	13–15	210	30	4.4	NA	NA	NA
chloride stabilized							
alumina surface							

* NA = not available.

synthesized in solution. The dispersion of pyrogenic silica to a sol of separate, discrete ultimate units is difficult, even using intense mechanical shearing forces, due to the extensive coalescence of the particles. Thus, the disperse product consists mainly of short chain-like aggregates made up of small silica particles (see Section 4.5). Typical properties of commercial sols made in this way are listed in Table 13.

Table 14 lists particle size, silica concentration, and viscosity of commercially available organosols.

Silica sols are said to be stable because they do not settle or aggregate for long periods of time. Aggregation and rate of settling, as well as color, appearance, viscosity, density, growth, and solubility in water are functions of particle size. At optimum pH, electrolyte and SiO_2 concentration, aquasols of particle size 4 to ca. 40 nm are extremely stable to settling, whereas aquasols of particle size larger than ca. 60 nm tend to show some settling in a period of months. Particles larger than 100 nm settle on standing, leaving a clear upper layer after a few weeks or days. When the concentration of the silica aquasol is $>$ 10–15 wt % the particle size can be judged visually by the turbidity. If the particles are smaller than ca. 7 nm in diameter the sol is almost as clear as water; from 10–30 nm there is a characteristic opalescence or translucency when seen in bulk; above 40 or 50 nm the appearance is white and milky [235].

Commercial silica aquasols are stabilized near the optimum pH and are concentrated to the maximum concentration permitted by the particle size [257]: ca. 15 wt % for 4–5 nm, ca. 30 wt % for 8–9 nm, ca. 40 wt % for 14 nm, ca. 50 wt % for 22 nm (Fig. 29).

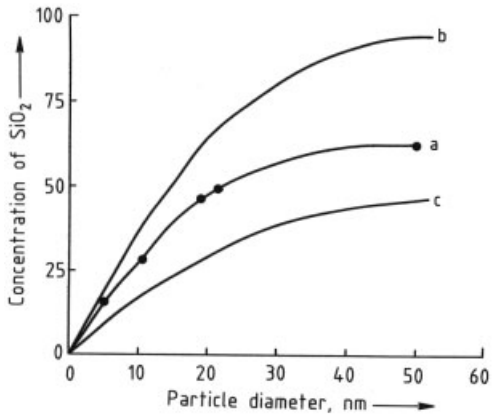

Figure 29. Maximum concentration versus particle size in stable aqueous silica sols at about pH 9.5
a) Concentration in wt %; b) Concentration in grams SiO_2 per 100 mL; c) Volume fraction of SiO_2 ($\times$100) [257]

When silica sols are gelled, dried, or frozen the original degree of dispersion cannot be restored without dissolution of the interparticle bonds that develop. Special dried grades are made of particles with the surface esterified with primary or secondary C_2–C_{18} alkoxy groups. These are organophilic and hydrophobic and can be dispersed in organic solvents or in organic products such as elastomers or plastics. Silica

Table 13. Typical properties of silica aquasols made of short chain-like aggregates of colloidal particles * [255]

Stabilizer	Ultimate particle size, nm	Solids, wt %	pH	Relative density	Viscosity, mPa · s
Acidic medium	**11–14**	**20**	**2.0–4.0**	**NA**	**5–25**
NH_4^+	7	12	7.5–7.8	1.07	<100
	14	17	9.5–10.0	1.10	<250
	14	12	5.0–5.5	1.07	<100
	17	17	9.5–10.0	1.10	<100
	30	20	5.0–5.5	1.12	<100
K^+	17	18	8.6–9.0	1.11	<150
	30	28	7.5–7.8	1.18	<200
	30	30	10.0–10.3	1.19	<150
Na^+	7	14	10.0–10.5	1.08	<100
	8	15	9.5–10.0	1.09	<100
	14	17	9.5–10.0	1.10	<100

* NA = not available.

Table 14. Properties of commercially available silica organosols * [256]

Liquid phase	Particle diameter, μm	SiO_2, wt %	Viscosity at 25 °C, cP	Relative density
Methanol	0.01	30	1–5	NA
2-Propanol	0.01	30	3–20	NA
Water/2-propanol	0.02	30	10	NA
Ethylene glycol mono-*n*-propylether	0.01	20	5–20	NA
Dimethylacetamide	0.01	20	1–10	NA
Ethylene glycol	0.01	20	10–20	1.23
Ethylene glycol	0.18 ± 0.03	20		1.23
DMF	0.02	35	5	NA
Oil	0.02	50	80	NA
Ethylene glycol	0.28 ± 0.03	20		1.23
Ethylene glycol	0.43 ± 0.03	20		1.23
Ethylene glycol	0.50	20	10–20	1.23
Ethylene glycol	0.53 ± 0.03	20		1.23
Ethylene glycol	0.73 ± 0.03	20		1.23
Ethylene glycol	0.80 ± 0.04	20		1.23
Ethylene glycol	0.90 ± 0.05	20		1.23
Ethylene glycol	1.50 ± 0.10	20		1.23

* NA = not available.

sols surface-modified with organic base cations such as $(CH_3)_4N^+$ can be evaporated to a dry powder that disperses spontaneously in water [258].

4.4. Stability

Three types of stability may be distinguished in colloidal systems [259]:

1) Phase stability, analogous to the phase stability of ordinary solutions.
2) Stability to change in *dispersity* [259], that is particle size or particle-size distribution. For example, commercial concentrated silica sols, normally said to be 3 or 4 nm in particle size usually grow within hours or a few days to 5 nm or more on standing at room temperature. Concentrated sols normally labeled 7 nm, if not substantially homodisperse, may grow within months to 8–9 nm. This ripening effect also occurs for larger particle sizes, although at a much slower rate. A concentrated silica sol of 14 nm particle size was found to grow to 17 nm in twenty years at room temperature [260].
3) *Aggregative stability* [259], the central issue in colloidal silicas, and for that matter in colloidal systems in general. In this case *colloidally stable* means that the colloidal particles do not aggregate at a significant rate [232]. The term aggregate is used to describe the structure formed by the cohesion of colloidal particles [232].

Two mechanisms of sol stabilization are generally believed to exist: electrostatic stabilization and steric stabilization (see → Colloids,

Chap. 5). Electrostatic stabilization is based on an interplay of electrostatic repulsion between electrically charged colloidal particles and van der Waals forces of attraction between particles. The DLVO theory (Derjaguin, Landau, Verwey, and Overbeek) constitutes an attempt to describe quantitatively this interplay (see → Colloids, → Colloids). Steric stabilization is generally caused by long-chain molecules or macromolecules adhering on the colloidal particle surface (e.g., by grafting or by physical adsorption), thus preventing the particles from aggregating [261].

Colloidal silica particles aggregate by linking together and forming three-dimensional networks as is the case in gelation, coagulation or flocculation, or by coacervation [262]. In coacervation the silica particles are surrounded by an adsorbed layer of material which makes the particles less hydrophilic but does not form bridges between particles [263].

When a sol is gelled it first becomes viscous and then develops rigidity, filling the volume originally occupied by the sol. On the other hand when a sol is coagulated or flocculated, a precipitate is formed that settles out. A simple way to differentiate between a precipitate and a gel is that a precipitate encloses only part of the liquid in which it is formed [262].

Stability of silica aquasols against irreversible gelation decreases with increasing silica concentration, increasing electrolyte concentration of the aqueous medium, decreasing particle size, and increasing temperature. Water-miscible organic liquids have a similar destabilizing effect on silica aquasols as added electrolyte. The variation of stability as a function of pH and salt concentration is shown in Fig. 30 [264].

According to Iler [265] the basic step in gel formation is the collision of two silica particles with sufficiently low charge on the surface that they come into contact so that siloxane bonds are formed, holding the particles irreversibly together. Formation of this linkage requires the catalytic action of hydroxyl ions (or, as interpreted by some, the dehydration of the surface of particles at higher pH). This is evidenced by the fact that the rate of gel formation in the pH range 3–5 increases with pH and is proportional to the hydroxyl concentration.

Above pH 6, scarcity of hydroxyl ions is no longer the limiting factor on the rate of gelling. Instead, the rate of aggregation decreases because of fewer collisions between particles owing to the increasing charge on the particles and thus decreases with higher pH. Lines in Figure 30 schematically represent the increase in the catalytic effects of hydroxyl ions with increasing pH, and the decrease in the number of effective collisions between particles with increasing pH and particle charge. The net result of these two effects is a maximum in rate of gelling at around pH 5. In the pH range 8–10, sols are generally stable in the absence of salts.

There is also a region of temporary stability at about pH 1.5. Below pH 1.5, traces of HF catalyze aggregation and gelling [266]. In essentially all silicas, traces of fluoride ions, even less than 1 ppm, are present so that the concentration of HF increases with increasing acidity. The fluoride effect is influenced by the aluminum impurities present, since these inactivate some of the fluoride by forming complex ions such as AlF_6^{3-} and other species [267]. However, the gelling rate increases as the pH falls below 3 even when fluorine is absent.

Once the siloxane bonds have formed between particles, further deposition of silica occurs at the point of contact owing to the negative radius of curvature. This occurs rapidly above pH 5, and is slow at pH 1.5.

Since the curves shown in Fig. 30 are constructed on the basis of irrefutable experimental evidence it is quite obvious that silica sols do not conform to the DLVO theory as originally formulated. For example, the DLVO theory predicts minimum stability at the point of zero charge (pH 2–3), whereas the experimental curve shows metastability. Also, the plot shows minimum stability in the pH range 4–7, whereas the DLVO theory predicts a continuous increase in stability in this pH range. Research is being conducted to study the possibility of modifying or amending the DLVO theory or developing a new theory of stability applicable to silica sols.

In addition to common electrolytes such as NaCl, NH_4Cl, and KF, silica aquasols are destabilized and gelled by positively charged sols. When frozen at ca. 0 °C or lower, silica sols gel irreversibly. Long-chain nitrogen bases are effective flocculating agents for silica aquasols that form planar rather than spheroidal aggre-

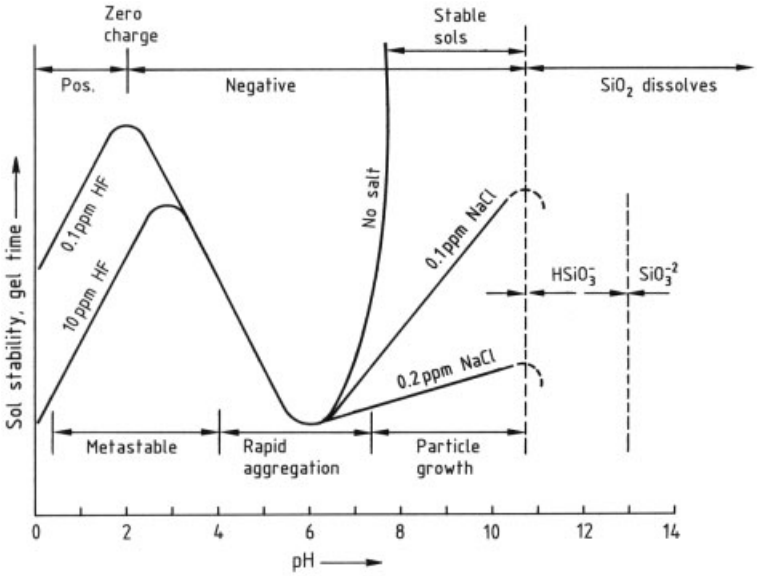

Figure 30. Effect of pH in the colloidal silica–water system [258]

gates. Flocculation may also result from the addition of water-miscible organic solvents to alkali-stabilized silica sols.

Commercial silica aquasols are stabilized at pH 8.5–10.5 by alkalies, usually sodium hydroxide. Ammonia or potassium hydroxide is used when the presence of sodium ions is undesirable.

Surface-modified silica aquasols made by coating the silica particles with tetrahedral aluminum (e.g., sodium aluminate) are much more stable towards gelling in the pH range 4–6 where unmodified sols gel most rapidly. Silica sols modified in this manner are also less sensitive to salts [268].

Coating the negative silica surface with oxides of polyvalent metals such as Al, Cr, Ga, Ti, and Zr reverses the charge of the surface to produce positive aquasols stable at acid pH values [269]. An important characteristic of these sols is that they can be dried and repeptized. Only polymeric hydrolyzed species such as $Al_2(OH)_5Cl$ and not the single Al^{3+} ion can cause charge reversal of silica sols [270, 271].

4.5. Production

The classic silica aquasols of particle size 5–100 nm are prepared by nucleation, polymerization, and growth in aqueous systems. The particle-size range can be extended to at least 300 nm by autoclaving. Silica organosols can be obtained by transferring aquasols to an organic solvent or by hydrolysis of a silane precursor in a mixture of alcohol/ammonia and sufficient water [272] followed by transfer to the solvent.

The most important processes to make silica sols are based on neutralization of soluble silicates with acids, ion exchange, hydrolysis of silicon compounds, dispersion of pyrogenic silica, electrodialysis, dissolution of elemental silicon, and peptization of gels [273].

Most commercial silica sols are prepared by ion exchange of dilute solutions of sodium silicate. Several methods have been proposed since the pioneering work of Bird [236], Bechtold and Snyder [237], Alexander [274], Atkins [275], Wolter and Iler [276], Mindick and Reven [277], and Dirnberger [278]. Sodium silicate can be deionized in a batch operation by adding simultaneously a dilute solution of sodium silicate and a cationic ion-exchange resin in the hydrogen form to a vigorously stirred weakly alkaline aqueous reaction medium in the pH range around 9, at 60–100 °C [276] (Fig. 31). Under these conditions the system is stabilized against aggregation and the original silica nuclei grow while the sol concentration in-

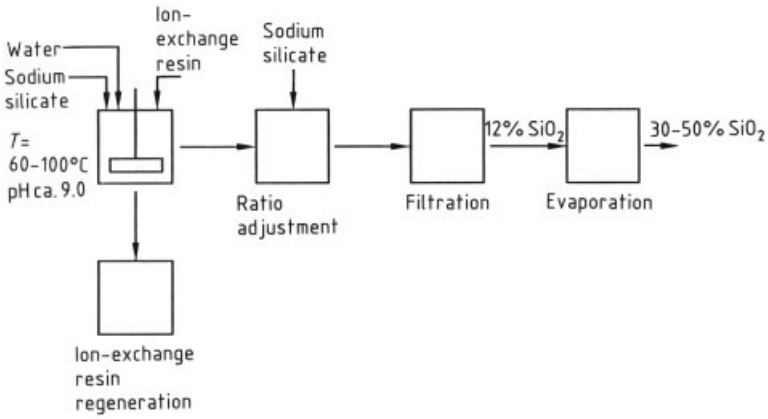

Figure 31. Ion-exchange process for silica aquasols

creases to about 10–15 % silica. Rate of addition, pH, and temperature determine the particle size and quality of the sol.

After separation of the resin for regeneration, more sodium silicate is added to adjust the SiO_2/Na_2O ratio as needed for further stabilization of the sol. The product is filtered and concentrated to the desired level. Sols of small particle size, such as 7 nm, may be used as "heels" to build up the particle size, for example, to 14 or 22 nm.

An alternative method of deionizing sodium silicate is to pass a relatively dilute solution through a bed of ion-exchange resin to produce a sol which is then stabilized with alkali and concentrated. The particles grow during evaporation. The addition of further deionized sol to the evaporating liquid allows silica to build up on previously nucleated particles resulting in larger particle size.

Silica sols can also be made by dispersion of pyrogenic silica in water or in an organic solvent. Commercial pyrogenic silica, also known as fumed silica or aerosil, is made from silicon tetrachloride at high temperatures by a flame hydrolysis–oxidation process (see Chap. 6) [279]. The product is a highly aggregated silica powder. Pyrogenic silicas can be partially disagregated and dispersed to obtain aquasols or organosols of relatively high silica concentration. The silica units in this case consist mainly of short chain-like aggregates composed of silica particles ca. 7–25 nm in diameter.

Monodisperse silica sols of exceptionally large particle size can be prepared by hydrolysis of tetraethoxysilane (TEOS) in a basic solution of water and alcohol [272]. It is claimed that the largest mean particle size that can be produced with TEOS is about 0.8 µm [280]. However, particles up to ca. 2 µm can be made by using tetrapentoxysilane [279], performing the reaction at low temperature [281], or growing the particles by adding more alkoxide after the particles have formed [281]. Wolf [282] has adapted the TEOS method to produce commercial quantities of silica sols.

Viable processes for making commercial silica sols are based on the electrodialysis of sodium silicate solutions to continuously remove sodium ions until a sol is obtained. In the Iler [283] and the Bergna [284] processes there are three compartments separated by two parallel, closely spaced cation-exchange membranes between which the process or "heel" solutions of silica sol in dilute Na_2SO_4 as a conducting electrolyte are rapidly circulated at a specified temperature. Sulfuric acid is circulated in the anode compartment and an alkali in the cathode compartment where sodium hydroxide is generated. Dilute sodium silicate is fed to the third compartment to form SiO_2 by electrolysis. The silica or polysilicic acid that forms is deposited on the silica particles of the "heel" solution, and circulation of the solution is continued until the desired particle size is achieved. Sols of 25 % silica, 15 nm particle size can be prepared directly. Concentrated sols of ca.8 nm particle size with very narrow particlesize distribution can be obtained by seeding in a first step at a lower tem-

perature and raising the temperature in a second step to accelerate silica deposition [284].

Among the advantages of the electrodialysis process are that alkali, oxygen, and hydrogen can be recovered, and that there is much less salt-containing wastewater to be disposed of. The disadvantages of the process are higher electrical and maintenance costs.

4.6. Analysis and Characterization

Characterization of silica sols is aimed at identifying:

1) Physical and chemical properties of the sol
2) Compatibility and stability properties of the sol
3) Purity of the sol
4) Use-related characteristics

Selection of silica sols for commercial uses is commonly based on the nature of the liquid phase (water or organic solvent), particle size, particle-size distribution, degree of aggregation, pH, silica and counterion (stabilizing agent) concentration, viscosity, relative density, and specific surface area.

However, to fully understand and predict the behavior of silica sols and derived powders, it is necessary to determine other characteristics such as particle and particle surface structure; chemical composition including total carbon, organic carbon, soluble salts of alkali metals, total solids, nonsiliceous ash, and trace metals (especially Al and Fe); and physical properties such as turbidity, percent S (the percent by weight of silica in the dispersed phase of a silica sol [285]), refractive index, light scattering, sedimentation rate by ultracentrifugation, porosity, rate of dissolution of the particles, and coalescence factor of derived powders (the coalescence factor is the percent silica that must be dissolved to restore the light transmission under standard conditions of a silica powder redispersed in water [286]). In addition, most manufacturers of colloidal silicas use special tests for specific uses.

A new concept in the characterization of colloidal silica is the application of the fractal approach to sols and gels. The concept of fractal geometry, developed by Mandelbrot [287] in the early eighties, provides a way of quantitatively describing the average structure of certain random objects. For the application of the fractal concept to colloidal silica, see [241, 250].

Most of the methods utilized to characterize silica sols for commercial uses and for research purposes are listed below:

Chemical analysis
% total solids
% and nature of liquid phase
% SiO_2
% NH_4
SiO_2/Na_2O weight ratio
Sulfates as % Na_2SO_4
Chlorides as % NaCl
Carbon including CO_2
Organic carbon
Metals including Al and Fe
Nonsiliceous ash

Physical characteristics
pH
Relative density
Viscosity
Turbidity
Refractive index
Light scattering
Specific conductivity

Particle characteristics
Size and size distribution
Specific surface area (Sears titration of the sol [260] and BET measurement of the derived dried powder)
Porosity
Degree of aggregation
% S (silica in the dispersed phase)
Turbidity
Viscosity
Surface charge
Electrokinetic potential (isoelectric point)
Silanol number

Stability tests
Gelation tests
Flocculation tests

The following methods are used to measure the particle size and particle-size distribution of silica sols [288 – 291]:

Chemical methods
Gas adsorption (e.g., BET) [288]
Titrations (e.g., Sears titration) [289]
Rate of dissolution of particles [290]

Instrumental methods [291, 292]
Transmission electron microscopy (TEM)
Dynamic light scattering: photon correlation spectroscopy (PCS): Malvern Correlator, Coulter N4 Analyzer
Sedimentation field flow fractionation (SF3)
Size-exclusion chromatography (SEC)
Hydrodynamic chromatography
Capillary zone electrophoresis (CZE)

Methods for characterizing the surface of colloidal silica particles are [250, 292]:

Spectroscopy
FTIR (Fourier transform infrared spectroscopy)
DRIFT (diffuse reflectance)
Probe molecule adsorption by DRIFT
^{29}Si CP MAS NMR spectroscopy (^{29}Si cross-polarization magic angle spinning nuclear magnetic resonance spectroscopy)
^{1}H MAS NMR spectroscopy
^{1}H CRAMPS (combined rotation and multiple pulse spectroscopy)
SIMS (secondary ion mass spectrometry)
SNMS (sputtered neutrals mass spectrometry)
EXAFS (extended X-ray absorption fine structure spectroscopy)
XPS (X-ray photo emission spectroscopy) (ESCA)

Thermal and calorimetric methods
Microcalorimetry
TGA (thermogravimetric analysis)
DTA (differential thermogravimetric analysis)
DSC (differential scanning calorimetry)
Thermoporometry

Heterogeneous isotopic exchange
Use of deuterated and tritiated substances combined with mass spectrometry, IR, and ^{1}H NMR

Adsorption and wetting
Adsorption of gases and vapors combined with IR, microcalorimetry
Fractality (fractal dimension)
TPD (temperature programmed desorption) of adsorbed substances (NH_3, pyridine, etc.) coupled with spectrometry and IR
Adsorption of bases (*n*-butylamine) using Hammett or arylmethanol indicators
Wetting with liquids

Microscopy
TEM (transmission electron microscopy)
SEM (scanning electron microscopy)
FESEM (field emission scanning electron microscopy)

Scattering
SAXS, SANS (small angle X-ray and neutron scattering)

Chemical methods
Use of chemically reactive substances: chlorine, metal halides, Grignard compounds, reactive chloro- and alkoxysilanes

4.7. Uses

This section gives a brief description of the major industrial uses of silica sols. The many minor uses are too numerous to be discussed here.

Colloidal silica is widely used as a binder in the modern version of the ancient lost wax process for casting metal [293]. In this process, known as *shell investment casting,* a wax original or a cluster of originals is dipped in a slurry of colloidal silica and refractory powder. Excess slurry is drained from the wax parts and a dry refractory sand is applied to the wet surface. The coating is then allowed to dry. These coating steps are repeated until a ceramic shell of sufficient thickness, usually about 5–10 mm, is built up around the wax.

The wax is then melted out, and the ceramic shell fired to increase its strength and to remove the last traces of wax. Molten metal is poured into the hollow left by the wax. When the metal has cooled, the shell is broken away and the metal, now an exact replica of the original wax shape, is recovered.

This process is widely used to produce jet engine components as well as a large number of other metal parts. The tolerances of the finished casting can be held very close to the final requirements, thus minimizing the need for additional finishing operations.

Refractory fibers can be bonded with colloidal silica to give *insulators* that have excellent high-temperature resistance. The process is similar to papermaking in that the fibers are suspended in water and the mixture passed through

a screen which retains the fibers and allows the water to pass through and be recycled. The thickness of the fiber mat can vary from 0.1–10 cm depending on the insulating properties required. The screen is often contoured to give a shaped insulator. Usually, vacuum is used to assist the flow of water through the screen.

Colloidal silica is used two ways in this process. It is often used in small quantities together with starch in the original fiber suspension to help flocculate the fibers and the starch for better drainage and retention on the screen. The starch acts as the binder in the unfired or green state. Additional colloidal silica is often added to the final shape to stiffen and strengthen it and provide additional strength when the part is heated and the starch is burned out.

More recently, colloidal silica has been applied to the *papermaking* process itself. By adding small amounts of high surface area colloidal silica and a high molecular mass starch to the paper pulp, drainage rates and fiber and filler retention are improved. This allows for higher filler loadings, use of more recycled pulp, and in some cases, higher production rates [294 – 298].

Small amounts of colloidal silica increase the coefficient of friction of surfaces. One of the earliest uses of colloidal silica was to diminish the slipperiness of floor waxes while not affecting their gloss [299, 300]. It is widely used to improve the frictional character of paper and boxboard, which facilitates handling and reduces breakage resulting from falling boxes [301].

Carpets and other surfaces coated with colloidal silica resist soiling because the colloidal silica occupies the sites which would most likely retain visible soil.

The strength and adhesion of latex-based adhesives and paints can be enhanced by the addition of colloidal silica.

Silicon wafers cut from silicon single crystals must be polished to an almost atomically smooth surface before being used as substrates for electronic chips [302, 303]. Colloidal silica is the main component of the final polishing compounds for these materials. They act both as a fine, uniform abrasive and as a scavenger for reaction products of polishing additives that chemically attack the silicon.

Colloidal silicas have been used to bond and improve the attrition resistance of catalyst powders used in streams of reacting gas or liquid. The silica provides sufficient strength and hardness to prevent the catalyst pellets from being broken down and swept away by the stream. Acrylonitrile is made from propylene and ammonia by such a fluidized bed reaction [304, 305]. Ammonia-stabilized colloidal silicas are generally used in these applications because of the poisoning effect of sodium.

Photographic films and papers often incorporate colloidal silicas as grain growth regulators or dye receptors.

Beverages, such as wine, beer, or fruit juices can be clarified using colloidal silica as an aid to the flocculation of the proteins which cause the materials to be hazy.

4.8. Storage, Handling, and Transportation

Colloidal silicas generally undergo irreversible precipitation of the silica if frozen. Therefore, they are generally stored in heated buildings. If outdoor bulk storage is required, tanks should be heated and insulated in climates where freezing might occur. Heated trucks are typically used in cold climates for shipping colloidal silicas.

Colloidal silica is sometimes freeze-stabilized by addition of organic substances such as glycols. The amount added is insufficient to prevent freezing, but does prevent irreversible precipitation.

Storage in plastic, fiberglass-reinforced plastic, stainless steel or lined steel tanks is usually recommended.

Typically, the main hazard associated with colloidal silicas is their alkalinity. However, since the pH of most commercially available materials is < 10.5 they may irritate the skin or eyes, but do not cause irreversible burns. In applications where a respirable mist can be formed, operators should be protected by engineering design or suitable respirators. Since colloidal silica is amorphous, it is less toxic than crystalline silica.

4.9. Economic Aspects

Major manufacturers of colloidal silica and their trade names are as follows:

Bayer	Baykisol	(Germany)
Du Pont	Ludox	(United States)
Eka Nobel	Bindzil, Nyacol	(Sweden)
Hispano Química	Hispacil	(Spain)
Monsanto	Syton	(United Kingdom)
Nalco Chemical	Nalcoag	(United States)
Nippon Shokubai	Seahostar	(Japan)
Nissan	Snowtex	(Japan)

Estimates of the colloidal silica market discussed in this section are based on [238, 239].

The U.S. market for colloidal silica is the largest in the world. The total market for North America is estimated at > 14 000 t (100 % silica basis) and valued at $ 50×10^6 [236]. These figures include ca. 1500 t consumed in Canada. Kline estimates that the total West European market for colloidal silica was 5500 t (100 % silica basis) in 1992, valued at over € 15×10^6 [239]. In Japan colloidal silica production in 1988 was estimated at 4500 t (100 % silica basis) and was expected to grow significantly [238].

The total annual capacity for the manufacture of colloidal silica in North America is estimated at 16 000–21 000 t (100 % silica basis). The total capacity in Europe is estimated at 6000 t.

Nalco is the largest producer for the North American market followed by Du Pont and PQ Corporation. Eka Nobel of Sweden has a relatively small plant in the United States (Pro-Comp Inc.). Alchem, a joint venture of C-I-L Inc. and Nalco in Ontario, is the sole Canadian producer.

In the United States, Monsanto imports colloidal silica from its plant in Wales and small amounts are also imported from Japan. Exports from the United States are believed to be relatively small at about 1800 t/a. Nalco and Du Pont are the main exporters, and Japan, Taiwan, and Canada are the principal destinations. Some colloidal silica is also exported to Europe.

The largest producers of colloidal silica in Western Europe are Bayer in Germany, Eka Nobel in Sweden, and Monsanto in the United Kingdom, followed by Akzo in the Netherlands and Hispano Química in Spain. There are a few smaller producers in Italy and France.

Western Europe imports some colloidal silica from the United States and exports relatively small amounts to the United States, Eastern Europe, the Far East, and the CIS.

In 1987–1988 the North American market for colloidal silica was expected to grow at an average rate of 5 %/a, and the Western European market at 3.5 %/a in volume. The author estimates that the actual growth for both markets has been lower due to the world economic situation.

Colloidal silica prices vary with grade, silica concentration, volume purchased, and country in which it is sold. In the United States the price is about $ 1.80–1.95 per pound for electronic-grade material, and $ 1.15–1. 40 per pound for other grades (100 % silica basis). In Western Europe the prices vary in the range 5.5–8.3 DM/kg (100 % silica basis).

5. Silica Gel

5.1. Introduction

Silica gel [*63231-67-4*] (revised 1990 SiO_2: [*112926-00-8*]) has the nominal chemical formula $SiO_2 \cdot x\,H_2O$ and is a solid, amorphous form of hydrous silicon dioxide distinguished by its microporosity and hydroxylated surface. The structure of silica gel is an interconnected random array of ultimate polymerized silicate particles, called micelles, which are spheroidal and 2–10 nm in diameter (resulting in high surface areas of ca. 300–1000 $m^2/g\ SiO_2$). The properties of silica gel are a result of the silica micelles, their state of aggregation, and the chemistry of the micelle surface.

Control of the surface area, porosity, and surface chemistry of silica gel has led to numerous and diverse uses. Examples include adsorbents for water or other species, abrasives and thickeners in dentifrice, efficient matting agents in coatings, chromatographic media, and catalyst supports. The utility is reflected in an estimated world consumption of > 110 000 t/a, not including China, the CIS, and Eastern Europe.

The development of silica-gel science and technology began with the pioneering work of Graham reported in 1861 [306], although observations of gelation had been reported earlier [307, 308]. Commercial production began at the Silica Gel Corporation (now part of W. R. Grace & Co.-Conn.) with the process invented by Patrick in 1919 [309]. Commercial development and applications expanded worldwide [307], and the colloid chemistry of silica was explored further [310]. With the application of

NMR spectroscopy, small-angle X-ray scattering, and fractal geometry [311], further fundamentals of silica gel chemistry are being elucidated.

5.2. Structure, Properties, and Characterization

Structure. Silica gel is prepared by the neutralization of aqueous alkali metal silicate with acid [307, 309, 310]; for example:

$$Na_2O{\cdot}3.3\,SiO_2 + H_2SO_4 \rightarrow Na_2SO_4 + 3.3\,SiO_2\downarrow + H_2O$$

A complete synthesis of silica gel by this typical commercial route is outlined in Figure 32. Alternatively, stable silica sols may also be gelled [310]. Another method involves the hydrolysis of silicon alkoxides with water, catalyzed by acid or base [311].

The neutralization of sodium silicate initiates a polymerization of silicate tetrahedra in a random, amorphous manner to form small spheroids called micelles. The solution containing the micelles while still liquid is known as a hydrosol.

Gel formation occurs when the interaction of separate micelles through hydrogen bonding and eventual interparticle condensation becomes significant. The rate of gelation depends on many variables such as SiO_2 concentration, pH, temperature, and mixing; this is discussed in detail in [310, 311]. Polymerization and cross-linking continue after the hydrosol has solidified into a hydrogel. The random, amorphous structure is reflected in a lower skeletal density for silica gel of 2.1–2.2 g/cm^3 compared to quartz (2.65 g/cm^3) and the lack of X-ray crystallinity. The micelle, which is the ultimate particle, consists of SiO_2 in its interior and Si−OH on its surface. Solid-state ^{29}Si NMR indicates tetrahedral geometry about the silicon atoms and confirms the presence of bulk SiO_2 and surface Si−OH [312]. The micelle's size determines the specific surface area of the silica gel; a typical micelle size of ca. 2.5 nm in diameter corresponds to ca. 1000 m^2/g.

Once the gel network forms, several processing steps are performed to give the finished gel [313]. First, washing is normally performed to

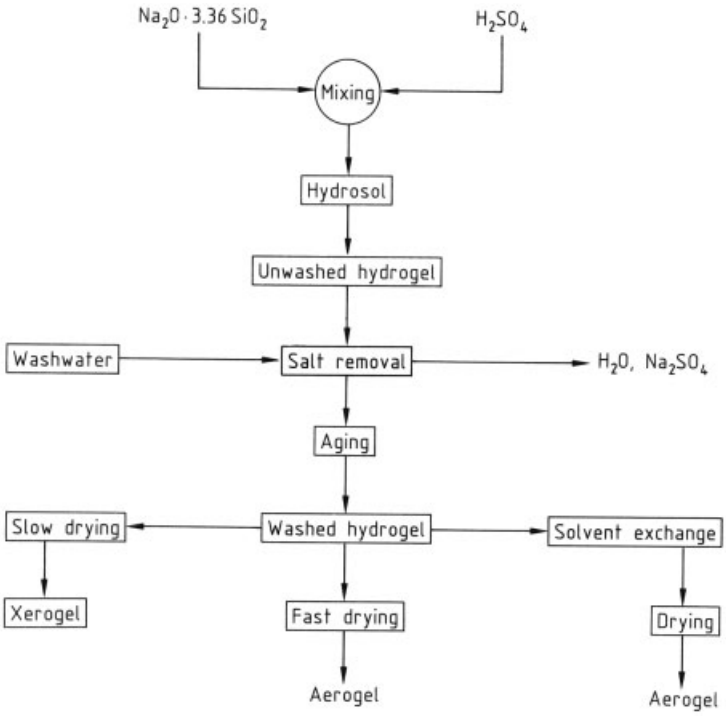

Figure 32. Schematic of silica gel manufacture

remove dissolved salts. Because silica gel has a very low ion-exchange capacity for cations or anions at moderate pH, salt removal is a diffusion-controlled dilution process. Next, if desired, the interaction between micelles and micelle growth can be accelerated by aging in aqueous media under conditions where silica is slightly soluble. A process in which silica dissolves from regions of positive curvature and redeposits at regions of negative curvature reinforces the hydrogel network. Thus, a continuous gel structure is formed.

During the drying process, the surface tension of the solvent in the pores can act to shrink the hydrogel volume. In slow drying, as water is evaporated from a silica hydrogel, the structure collapses gradually due to the surface tension of water. Eventually a point is reached where even though water is still evaporating the gel structure no longer shrinks. At this point the gel is called a xerogel. Fast drying can minimize the shrinkage, and removal of water by solvent exchange followed by drying has the same effect. Materials that are dried with negligible loss of pore volume are known as aerogels. Solvent exchange with a water-miscible liquid such as ethanol to remove water from the hydrogel preserves pore volume after drying. Single or multiple solvent exchanges to lower surface tension results in less pore collapse during drying.

Drying of silica gels can also be carried out under conditions where the solvent in the pores is above its critical point and is vented while maintaining temperature and pressure. KISTLER [314] dried gels supercritically after replacing the water of the hydrogel with alcohol. Alternatively the liquid in the pores can be exchanged for liquid CO_2, which has more convenient critical-point properties [315, 316]. Considerable interest has been focused on such techniques due to the high porosity and mechanical properties achievable [311, 315, 317].

The resulting products from a scheme as in Figure 32 are high purity silica gels with controlled porosity. Transmission electron micrographs showing the pore structure of a silica gel are shown in Figure 33. Typical properties are as follows:

Chemical analysis (dry basis)	
SiO_2	99.71 wt %
Al_2O_3	0.10 wt %
TiO_2	0.09 wt %
Fe_2O_3	0.03 wt %
Trace oxides	0.07 t %
Physical analysis	
Total volatiles	5–6.5 wt %
Surface area	750–800 m^2/g
Pore volume	0.43 cm^3/g
Average pore diameter	2.2 nm
Apparent bulk density	0.72 g/cm^3
Skeletal density	2.19 g/cm^3
Specific heat	920 J kg^{-1} K^{-1}
Thermal conductivity	522 J m^{-1} h^{-1} K^{-1}
Refractive index	1.45

Surface Chemistry and Stability. The physical microporosity of silica gel is an important property, as is the chemistry of its bulk and surface species (Fig. 34). Pure silica gel has a hydroxylated surface covered with silanol groups. These hydroxyl groups are neither very acidic

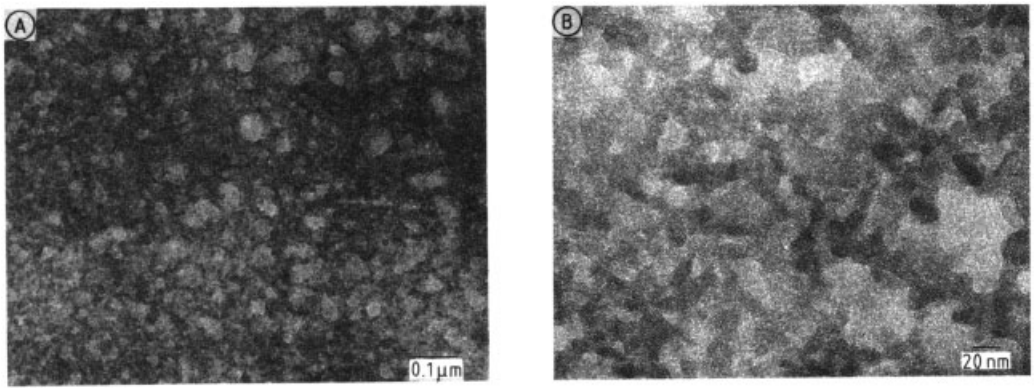

Figure 33. Transmission electron micrographs of a silica gel similar to the fast-dried, high pore volume silica gel in Table 15, showing the micelle structures (ca. 20 nm) and apparent pores also ca. 20 nm in diameter
A) magnified ×235 000; B) magnified ×688 000

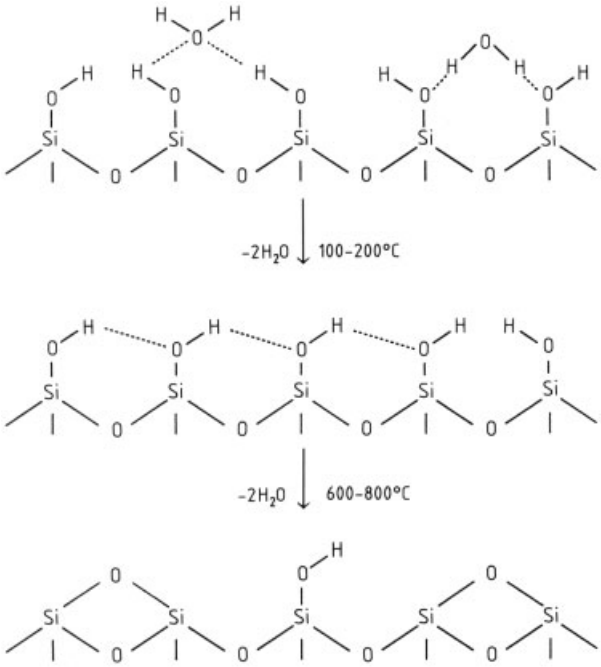

Figure 34. Schematic representation of the dehydration of a silica gel surface

nor basic, with p$K_a \approx 6$ and an isoelectric point at ca. pH 2. The hydroxylated surface is hydrophilic and adsorbs moisture readily. This adsorbed moisture can be desorbed thermally at 100–200 °C leaving behind ≈5.5 OH/nm^2 (or ca. 5 wt % silanol groups on a 300 m^2/g silica). These silanols are more difficult to remove but this can be done thermally by condensation to form siloxanes and water. Temperatures near 600–800 °C are required to dehydroxylate to ca. 1 OH/nm^2. At this silanol concentration the surface is hydrophobic.

Characterization. Numerous methods have been developed to characterize the porosity, structure, and chemistry of silica gels [310] and new methods continue to be applied [311].

With regard to porosity, differently prepared silica gels can be described on the basis of pore diameter d as microporous ($d < 2$ nm), mesoporous ($d \approx 2$–50 nm), or macroporous ($d > 50$ nm) [318]. The most generally useful methods for characterizing these structures with respect to surface area, pore volume, and pore-size distribution are nitrogen adsorption/desorption techniques. The primary method for surface area is the BET procedure [319] (ASTM D 3663–84 or DIN 66 131). Methods have been proposed to improve on the BET technique in small pores, such as the t-plot method [320], and the α_s method [321]. The related N_2 pore volume and pore-size distribution (pore volume as a function of pore diameter) methods are ASTM procedures D 4222–83 and D 4642–87, respectively. The porosity of a silica gel, determined by N_2 porosimetry, is shown in Figure 35. Alternative porosity measurement techniques include mercury porosimetry, water pore volume, and oil absorption.

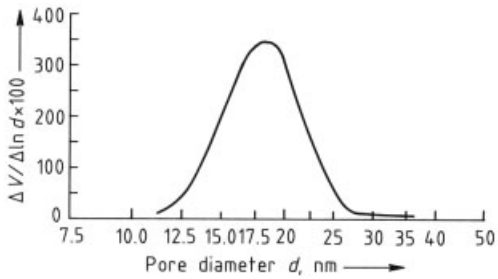

Figure 35. Nitrogen pore-size distribution of a silica gel similar to the fast dried sample of Table 15
V = pore volume

Elemental analysis of silica gel focuses on trace elements in the range of < 1 wt %. Inductively coupled plasma atomic emission spectroscopy (ICP-ES) is becoming the preferred technique. pH measurement is performed on slurries and reflects the balance of residual trace elements (ASTM D 1208, DIN ISO 787 1X, JIS K 5101/24).

The moisture content of silica gel is important to provide an anhydrous basis for chemical analysis and because the hydration of the surface and pores (Fig. 34) is important to performance in applications. Most standards require two measurements: moisture loss on drying for 2 h at 105 °C (ASTM D 280, DIN ISO 787/II, or JIS K 5101/21) and loss on ignition after this drying by heating for 2 h at 1000 °C (ASTM D 1208, DIN 55 921, or JIS 5101/23). Another method often used in the United States prior to analysis to provide a solids basis is to measure total volatiles as a weight per cent loss on heating for 1 h at 954 °C.

The surface hydroxyl groups of silica gel can be characterized by thermogravimetric analysis [310], vibrational spectroscopy [311, 322], and nuclear magnetic resonance [311, 312, 323]. The information obtained concerns hydroxyl concentration, hydrogen bond interaction between hydroxyl groups, and distribution of $\equiv$Si$-$OH and =Si$-(OH)_2$ species on the surface. The adsorptive properties of variously hydroxylated silica materials are an indirect means of characterizing the surface.

Particle-size measurement is also important in silica-gel characterization (DIN 53 206). The method for granular gel is standardized sieve screening (ASTM D 4513). For particle sizes of roughly 1 to 1000 µm, static light scattering (ASTM) and conductivity methods are preferred for particle size analysis. For particle sizes of roughly 10 to 1000 nm, dynamic light scattering [352], electron microscopy, and small-angle X-ray scattering are used.

5.3. Production

Processes. The original production process of Patrick [309] as practiced by the Silica Gel Corporation starting in 1919 contains the main elements of modern processes using aqueous silicate raw materials. The initial step is the batch neutralization of sodium silicate (1.185 g SiO_2/cm^3) at 35–80 °C with 10 % HCl to form a hydrosol. A hydrogel then forms after 3–5 h at ambient temperature and is sized, washed with water to remove salts, and dried first at 75 –120 °C then 300–400 °C. A typical small pore gel made by this process is described in Table 15.

The Patrick gelation conditions, with relatively minor variation, have been the basis of most silica gel manufacture since. The next major change in silica gel products was the development of gels of higher pore volume and pore diameter by washing and aging hydrogels under basic conditions. Early references [324, 325] refer to this technology but large scale commercial introduction occurred around 1945. Base washing was a common industrial practice before 1950. In one example [326] a 17 % SiO_2 hydrosol gelled in about one hour. The hydrogel was broken up and washed with an aqueous ammonia solution of pH 9 for ca. 45 h then tray dried. Similar conditions are described for intermediate-density gel [313]. Alternative washing schemes are numerous and affect rates and equipment design or cost [353, 354]. In one such method that provides for faster throughput of silica hydrogel through the wash stages [354], the hydrogel particles containing reaction product salts are small (50–200 µm) and washing occurs via a hydrocyclone. In this case, the dried product typically has a surface area of 500 m^2/g and a nitrogen pore volume of 0.25 cm^3/g. A further variation involves fast drying [327] which increases the pore volume and diameter of such gels. Typical properties of acid-washed regular density, ammonia-washed intermediate density, and a flash-dried product are shown in Table 15.

Several methods of forming hydrogel, typically into spherical shapes, have been developed. Some methods employ a rapid gelation after mixing in a nozzle followed by spray setting [328 – 331], while others form beads in an oil phase immiscible with the hydrosol [332]. In some bead processes, recycled, dried, and powdered gel is mixed with the fresh sol to improve the physical integrity of the dried gels toward disintegration after contact with liquid water [333]. The direct formation of 5–10 µm gel particles by emulsion polymerization or particle-forming techniques is also possible. A process for the direct formation of 1–5 µm particles with

Table 15. Typical properties of various silica gels

Property	Regular density	Intermediate density	
	Slow dried	Slow dried	Fast dried
Total volatiles [a]	5	6	6
Na_2O dry basis, wt %	<0.1	<0.1	<0.1
SO_4 dry basis, wt %	<0.1	<0.1	<0.1
Al_2O_3 dry basis, wt %	0.05	0.05	0.05
pH [b]	4.0	7.5	4.0
Specific surface area, m^2/g	750	325	390
Pore volume, cm^3/g	0.40	1.1	1.8
Average pore diameter, nm	2.1	13.5	18.5
Water adsorption, wt %			
10 % R.H.	7.5	2.0	
80 % R.H.	35	17.0 [c]	

[a] Weight loss after heating in air at 954 °C.
[b] pH of an aqueous slurry of the gel at 5 wt % solids.
[c] >90 wt % water adsorption at 100 % R.H.

"double structure", i.e., with a core of dense silica and a shell of bulky amorphous silica, has been disclosed [334].

The supercritical preparation of aerogels has undergone a renaissance since earlier commercial production [335] because of renewed interest in the insulation and light-transmittance properties of these materials. The shift in applications toward large panels has created the need for what are described as aerogel monoliths [311]. One such process [315] forms a silica alcogel then exchanges the alcohol for liquid CO_2 before supercritical drying. The resulting monoliths show good light-transmittance properties. The monoliths, typically 500 × 500 × 2.5 cm, can be used to fill the space in the middle of double-pane windows. Another process involves reacting silanol groups on the internal pore surface with a reagent such as trimethylchlorosilane, which then minimizes shrinkage of the gel under drying conditions [355]. Dried gels of very high porosity and low density are thus obtained.

Environmental Protection. In the sodium silicate/H_2SO_4 processes the main environmental concerns are the fate of byproduct process chemicals, normally aqueous Na_2SO_4 and small amounts of aqueous NH_3, and pH control of this stream.

Quality Specifications. Typical quality criteria for silica gel are chemical purity, moisture content, and porosity. These properties depend on manufacturing conditions such as concentrations, gelling and washing times, pH, and drying conditions. Typical properties of silica gels are reported in Table 15. Application-specific properties are often measured for certain end uses (see Section 5.4).

Silica gel has been described by numerous organizations for the purpose of defining the quality necessary for use in food, cosmetic, and pharmaceutical applications. In the United States, monographs for silica gel are included in the Food Chemicals Codex (1981) and The United States Pharmacopeia–The National Formulary (1990). In Japan it is described in the Japanese Pharmacopeia, Standard for Cosmetic Materials, and JIS Z 0701. It is also on the "Positive" list for use in food-grade polyolefins. The EC is compiling permitted silica gel uses into a complete listing for member states. Until that listing is complete member states will maintain individual regulations.

5.4. Uses

Silica gel is used extensively in a wide variety of applications; here, some major uses are described.

Desiccants. The earliest well identified use for silica was as an adsorbent for water. The affinity of silica gel for water is affected by its state of activation (or moisture loading) and the degree of saturation of the surrounding fluid by water vapor. A plot of the affinity of silica gel for water vapor as a function of equilibrium relative humidity at ambient temperature is shown in Figure 36. The properties of two silica desiccants have been described in Table 15.

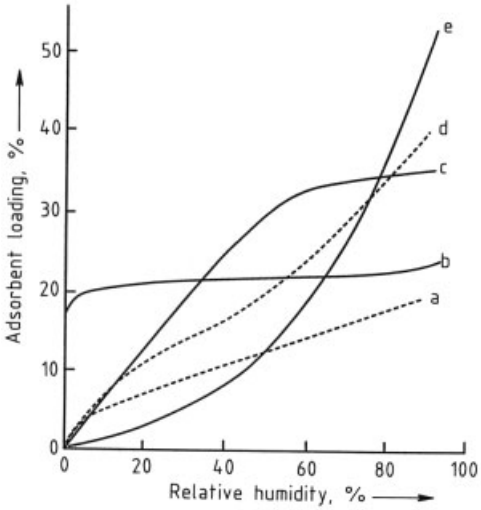

Figure 36. Water adsorption isotherms at room temperature as a function of relative humidity for several materials
a) Granular Al_2O_3; b) Zeolite A (Na^+ form); c) Regular density silica gel; d) Spherical Al_2O_3; e) Intermediate density silica gel

Silica gel is widely used in the narrow pore form to keep enclosed spaces dry, such as in cable junctions, pharmaceutical containers, and consumer goods. Silica gel is also used for drying natural gas. The wider pore form of silica is used to prevent local condensation in high-moisture environments and to absorb condensed mists of moisture. Another application is to use silica gel to maintain a constant humidity environment around art objects [336].

Desiccants have been described in several countries by standards Mil-D-3716B (United States), DEF STAN 80–22 (United Kingdom), and JIS Z 0701–77. These descriptions include standard properties and methods of analysis.

Adsorbents. Silica gel can adsorb many species other than water, particularly polar substances [337]. In addition to chemical processing, silicas have many uses in food processing. Treatment of beer with hydrogels, hydrous gels, or xerogels to remove haze-forming proteins is one example [325, 338]. Another food use is the adsorptive purification of glyceride oils [339]. Silica gels also are used extensively in chromatography [340].

Dentifrice. Silica gels have long been components of dentifrice formulations [341] but more recently were employed more for their suitable abrasiveness for cleaning than simply as carriers or thickeners. Xerogels [342] and hydrous gels [343] are used, but must compete with precipitated silicas [344] in this application. The abrasiveness of a dentifrice can be measured [345] and generally correlates with how effectively it cleans the teeth.

Coatings. Fine sized silicas (2–15 μm) are used extensively to control the reflectance of coatings [326, 327, 335]. By interrupting the surface of the coating the gloss of the surface (as measured by ASTM D 523) is reduced but particles too large can harm the appearance of the surface. The fineness of grind (ASTM D 1210–79) can be measured and specified. In clear coatings, the refractive index of the silica is important to maintain clarity. A typical matting agent has a particle size of ca. 10 μm, a pore volume of ca. 1.8 cm^3/g, and a fineness of grind on the Hegman scale of > 4 (ASTM test). A relatively new application for silica gels is in paper coatings for digital imaging, particularly by ink-jet printers [356]. In this application, the silica-rich coating serves to absorb ink-liquids and to control image appearance.

Microelectronics. The application of silica gels as insulators with low dielectric constants in microelectronics is developing. This use takes advantage of the air-containing pore structure to reduce the dielectric constant [357].

Catalysts and Catalyst Carriers. The high surface area of silica gel makes it an attractive solid on which to effect catalysis. Silica itself does not catalyze many reactions of commercial significance. Silica aluminas have acidic catalytic properties best exemplified in catalytic cracking [307, 346]. Silica gel is also used as a support in olefin polymerization [347 – 349].

Suppliers. Silica gel suppliers (with trade names following in parentheses) are summarized: Akzo, Asahi, BASF (Silica-Perlen, Silica-Pulver), Crosfield Chemicals (Gasil, Lucilite), Eagle, Fuji Silysia (Sylopute, Art Sorb), W. R. Grace & Co.-Conn. (Syloid, Sylobloc, TriSyl), Kali Chemie (KC-Trockenperlen), Millenium (Silcron), Mizusawa, PQ (Britesorb), Qingdao Hyiang, and Uetikon (Diamantgel).

5.5. Economic Aspects

Between 30–50 % of silica gel is estimated to be used in desiccant applications. Estimated production of silica gel in 1994 was > 110 000 t.

5.6. Legal Aspects

Toxicology and Occupational Health (see also Chap. 9). Silica gel is listed under TSCA in the United States, the Domestic Substances List (DSL) in Canada, the European Inventory of Existing Commercial Substances (EIECS), the Australian Inventory of Chemical Substances, and the Japanese Core Inventory (MITI) [358].

Silica gel is generally classified as synthetic amorphous silica. Epidemiological studies have indicated low potential for adverse health effects in humans [350]. Silica gel is not listed on IARC, National Toxicology Program (NTP), or OSHA carcinogen lists.

The IARC evaluation of amorphous silica is that there is inadequate evidence for carcinogenicity to humans or experimental animals. The phrase "inadequate evidence" indicated that the studies cannot be interpreted as showing either the presence or absence of a carcinogenic effect. For synthetic amorphous silica, there is no evidence of carcinogenicity to animals or humans [359].

A study found that amorphous silica particles are regarded as rather innocuous, not causing chronic adverse pulmonary effects [359]. In contrast, exposure to crystalline silica particles can cause both acute and chronic pulmonary inflammatory responses. The objective of this study was to evaluate mechanisms of pulmonary responses to amorphous and crystalline SiO_2 after subchronic inhalation. The study compared the pulmonary response of rats inhaling amorphous silica for three months at concentrations which resulted in a similar acute inflammatory response for both compounds at the end of exposure. The conclusion is that the persistence of crystalline SiO_2 appears to be most important for its long-term effects and that amorphous silica shows significant in vivo solubility. A further conclusion is that acute inflammatory parameters cannot be used as predictors for long-term outcome and that type II cell activation appears to be associated with fibrotic lesions. The study suggests that persistence of high level of cell replication and inflammation may be mechanistically involved in the development of neoplasia.

Cellular changes identified as neoplasms in this study were seen for rats exposed to crystalline silica and were not seen in the rats exposed to amorphous silica. Also, any pathology seen in animals exposed to high doses of amorphous silica was reversible, whereas the pathology for animals exposed to crystalline silica continued to the development of neoplasia [360].

Tests conducted for United States FDA approval for use in foods (see 21 CFR 160.105, 160.815, and 172.480) showed:

LD_{50} (mouse) 8000 mg/kg (limit of test)
LD_{50} (rat) 4500 mg/kg (limit of test)
6-month feeding tests (rat) at levels up to 10 % of diet produced no effects.

A long-term bioassay of chronic toxicity on mice and rats concluded that "proper dietary administration of micronized silica has proven to be generally safe with no long-term toxic effects" [351].

Storage and Transportation. Silica gel should be stored in sealed containers to protect product quality, particularly to avoid moisture adsorption or desorption. Silica gel is not considered a hazardous material by the International Air Transportation Association (IATA Resolution 618, Attachment "A", 1992), the U.S. DOT (49 CFR), or in EEC Council Directive 67/548/EEC.

6. Pyrogenic Silica

The term pyrogenic silica refers to highly dispersed silicas formed from the gas phase at high temperature. Nowadays, the most important production process is flame hydrolysis. The electric-arc process is of little and the plasma process of no economic significance. The designation pyrogenic used in this chapter therefore always refers to silica produced by flame hydrolysis.

6.1. Flame Hydrolysis

The flame hydrolysis process was developed in the late 1930s and patented in 1942 [361]. The

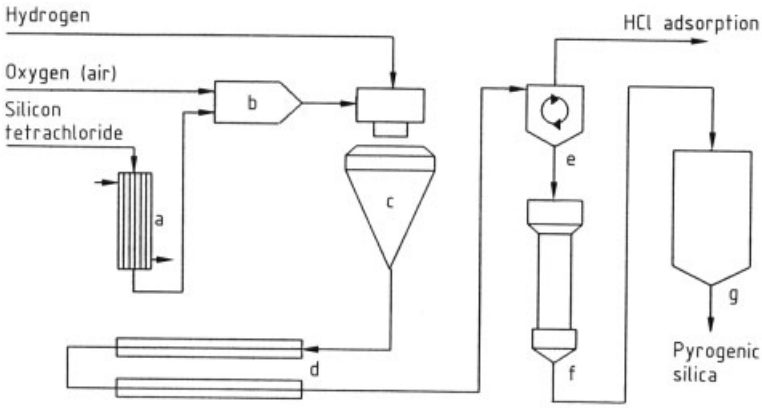

Figure 37. General process scheme of flame hydrolysis
a) Vaporizer; b) Mixing chamber; c) Burner; d) Cooling section; e) Separation; f) Deacidification; g) Hopper

original aim was to develop a "white carbon" as a reinforcing filler for rubber. However, for technical and economic reasons precipitated silicas (see Chapter 7) were adopted for this use.

In the meantime pyrogenic silica [*112945-52-5*] has found applications in numerous branches of industry.

Capacities for pyrogenic silicas amounted to ca. 100 000 t worldwide in 1991. Producers are Degussa (Aerosil), Cabot (Cab-o-sil), Wacker (HDK), and Tokuyama Soda (Reolosil).

6.1.1. Production Process

Silicon tetrachloride is the usual raw material for flame hydrolysis. It is continually vaporized, mixed with dry air and then with hydrogen, fed to a burner, and hydrolyzed:

$$2\,H_2 + O_2 \rightarrow 2\,H_2O$$

$$SiCl_4 + 2\,H_2O \rightarrow SiO_2 + 4\,HCl$$

The gases leaving the reaction chamber contain all of the silica in the form of an aerosol.

The silica is separated almost quantitatively from the hydrochloric acid- containing off-gas by cyclones (centrifugal separators) or filters. Treatment with steam and air in a fluidized-bed reactor is then carried out to remove residual hydrochloric acid adsorbed on the large surface of the silica. The hydrogen chloride is washed from the off-gases in adsorption columns to give hydrochloric acid in commercial concentrations. The hydrochloric acid formed can be reused by reacting it with silicon to produce silicon tetrachloride and hydrogen.

The properties of pyrogenic silica can be controlled by varying reaction parameters such as flame composition and flame temperature. Thus, desired specific surface areas in the range 50 to 400 m^2/g can be produced selectively.

The product, with a bulk density < 20 g/L, is pneumatically transported to packing machines, where the tapped density is increased to 50 –120 g/L by compacting rollers or vacuum packers.

A schematic of the flame hydrolysis process is shown in Figure 37.

It is also possible to use methyltrichlorosilane alone or mixed with silicon tetrachloride as raw material. In this case, however, simultaneous hydrolysis and oxidation is required, which requires modification of the production process.

6.1.2. Morphology

Flame hydrolysis produces extremely fine, mostly spherical particles with diameters of ca. 10 nm. The size of the average primary particles, which can be measured by transmission electron microscopy (TEM), ranges from 7 to 40 nm. The particle-size distribution becomes narrower with decreasing primary particle size (see Fig. 38).

The small particles give rise to high specific surface areas of ca. 50–400 m^2/g. This consists almost entirely of outer surface, which is not formed by pores, and is readily accessible to diffusion processes. The specific surface areas determined by adsorption (e.g., by the BET method

[362]), are in agreement with those determined by TEM within certain limits. Figure 39 shows the BET surface area as a function of the mean particle size measured by TEM.

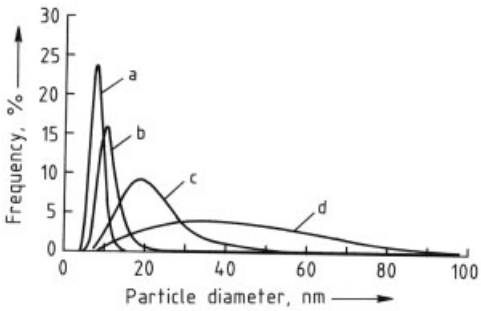

Figure 38. Primary particle-size distribution of pyrogenic silicas with various specific surface areas
a) 300 m^2/g; b) 200 m^2/g; c) 90 m^2/g; d) 50 m^2/g

The primary particles do not occur in isolation, but form aggregates by intergrowth, and agglomerates through cohesion forces. In contrast to precipitated silicas, pyrogenic silicas cannot be produced in the form of defined aggregates or agglomerates. The primary particles can be identified by TEM, but it is not possible to distinguish aggregates from agglomerates [363].

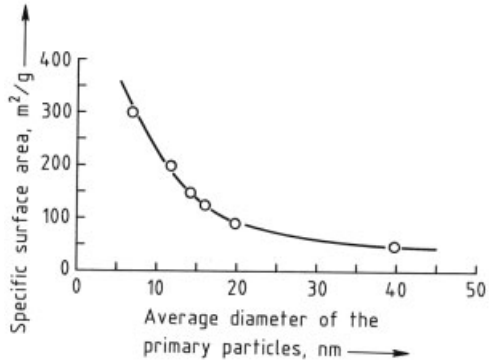

Figure 39. Correlation of the average primary particle size with the specific surface area of different pyrogenic silicas

The size of the agglomerates actually present in a liquid or powder mixture depends mainly on the dispersion and mixing intensity during preparation.

Figure 40 shows a TEM photograph of a pyrogenic silica with an average primary particle size of ca. 40 nm and a specific surface area of

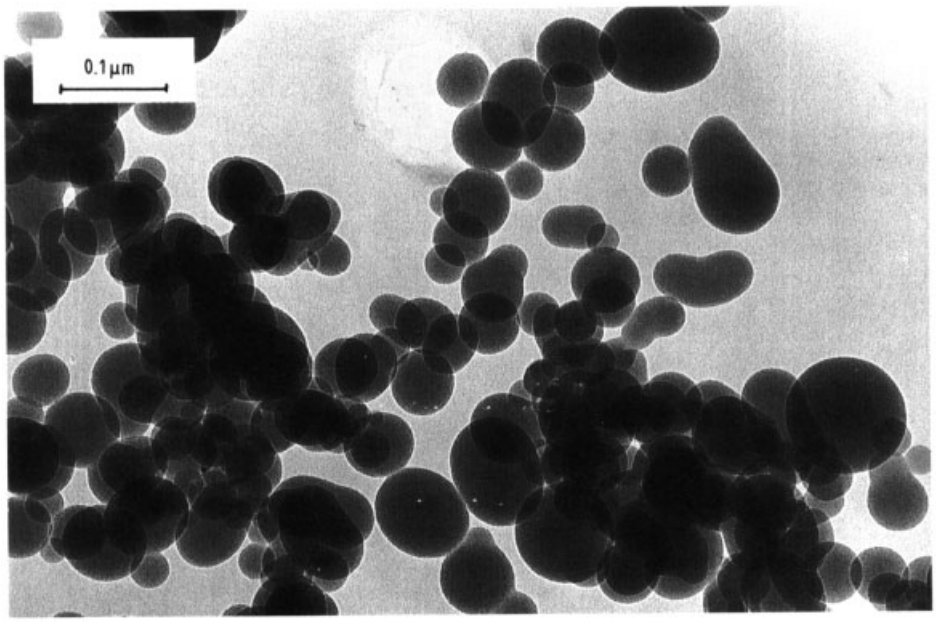

Figure 40. TEM photograph of a pyrogenic silica with an average primary particle size of ca. 40 nm and a specific surface area of ca. 50 m^2/g

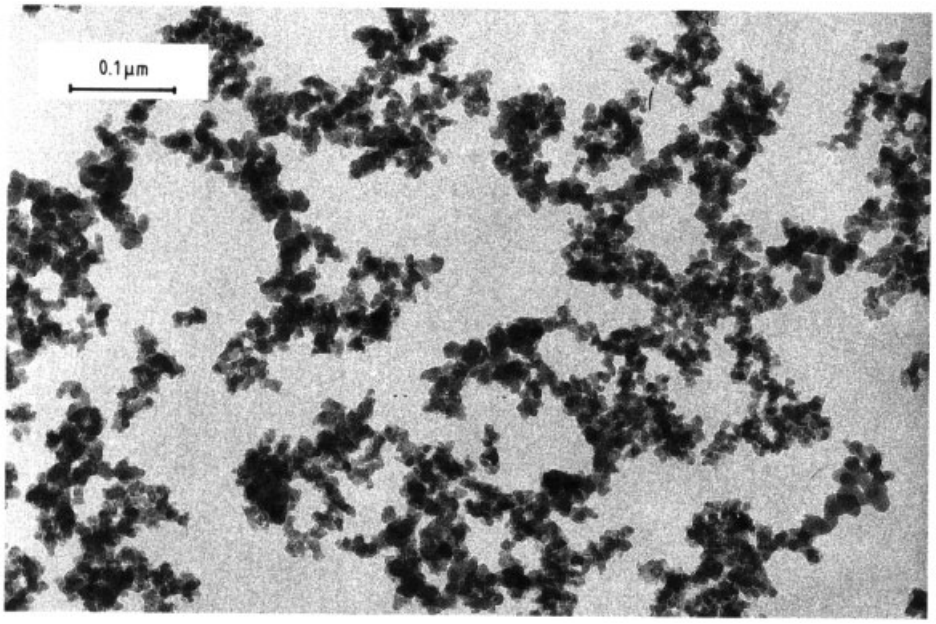

Figure 41. TEM photograph of a pyrogenic silica with an average primary particle size of ca. 7 nm and a specific surface area of ca. 300 m^2/g

ca. 50 m^2/g, while Figure 41 shows one with an average primary particle size of ca. 7 nm and a specific surface area of ca. 300 m^2/g.

6.1.3. Solid-State Properties

Pyrogenic silicas are X-ray amorphous. For example, X-ray diffraction photographs with a detection limit of 0.2 % for Aerosil 200, a hydrophilic pyrogenic silica with a specific surface area of ca. 200 m^2/g, show no crystallinity [364]. This fact is of considerable significance for industrial hygiene, since current experiences indicate that the development of silicosis on inhalation of silica dust is associated with the crystallinity.

The refractive index of 1.45 is similar to that of silica glass. The particle size and the surface chemistry have a small but measurable influence on the refractive index. Transparent mixtures can be prepared from pyrogenic silicas and most organic polymers.

Pure pyrogenic silicas are thermally quite stable. Thus, heating for 7 d at 1000 °C results in no change of the morphology and no crystallization. The thermal stability is, however, significantly lower if other substances are present. Traces of alkali or alkaline-earth metal ions in particular act as mineralizers [365].

Since pyrogenic silicas are produced from readily vaporizable starting materials that can be easily purified by distillation, impurity concentrations resulting from the raw materials are very low. Generally, the silicon dioxide content is > 99.8 % after ignition for 2 h at 1000 °C to remove chemically and physically adsorbed water. The content of hydrochloric acid byproduct can be reduced to less than 250 ppm by suitable measures, which is adequate for most applications.

The purity of pyrogenic silicas is specified particularly for their use in pharmacy. The most important monographs concerning this are:

1) European Pharmacopeia 2 (Silica Colloidalis Anhydrica)
2) DAB 10 (Hochdisperses Siliciumdioxid)
3) U.S.P./National Formulary XVII (Colloidal Silicon Dioxide)

Pyrogenic silicas are largely inert chemically. They dissolve in strong alkali solutions with silicate formation and in hydrofluoric acid with the formation of silicon tetrafluoride. The solubility in pure water is similar to that of quartz (ca. 150

mg/L). Figure 42 shows the solubility behavior of a pyrogenic silica with a specific surface area of 200 m^2/g in the alkaline pH range.

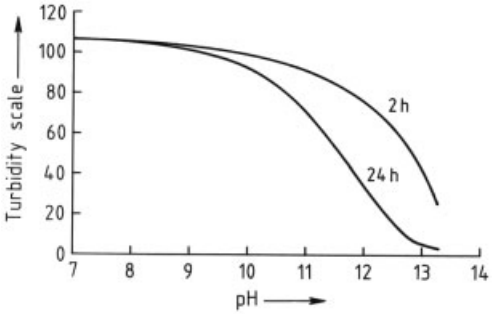

Figure 42. Time dependence of the solubility of a pyrogenic silica in caustic soda

6.1.4. Surface Chemistry

Siloxane and silanol groups occur on the surface of the silica particles. The former are hydrophobic while the latter are hydrophilic and make the pyrogenic silica wettable. The silanol groups can occur isolated, bridged, or geminal (Fig. 43).

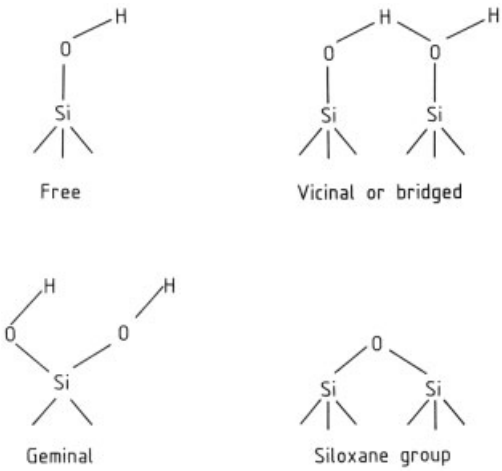

Figure 43. Siloxane and silanol groups on the silica surface

The silanol groups can be determined quantitatively by reaction with lithium aluminum hydride and measurement of the amount of hydrogen formed. The densities lie between 2.5 and 3.5 SiOH/nm^2 [366, 367]. The silanol groups are weakly acidic, which causes a zeta potential for the pyrogenic silicas of zero at a pH value of approximately two [368]. Pyrogenic silicas become preferably negatively charged by friction [369].

IR spectroscopy is important for the analysis of the surface chemistry of pyrogenic silicas [367, 370 – 373], and methods such as solid-state NMR spectroscopy promise further results. The following IR bands are of significance:

– O–D	2760 cm^{-1}
– C–H	2900–3000 cm^{-1}
–SiOH isolated	3750 cm^{-1}
–SiOH bridged	3000–3800 cm^{-1}
–SiOH combination band	4550 cm^{-1}
H_2O adsorbed	5200 cm^{-1}

Immediately after production pyrogenic silicas show mainly isolated silanol groups. In the course of time, adsorbed water reacts with strained siloxane groups and forms bridged silanol groups. This aging can easily be followed by IR spectroscopy (Fig. 44) [367].

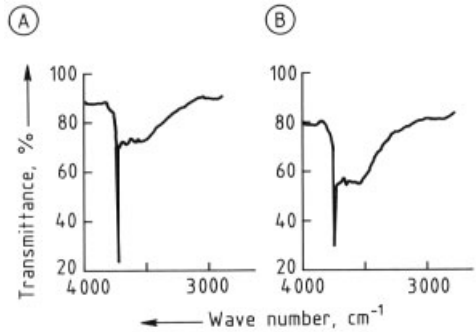

Figure 44. IR spectra of pyrogenic silica (200 m^2/g specific surface area) shortly after production (A) and after storage for one year (B)

The desorption energies of both the chemically and physically bound water can be determined by gravimetric adsorption and temperature-programmed desorption [374]. Thus, the following energies were determined: 50 kJ/mol, for water hydrogen bonded to a silanol group, and 84 kJ/mol for water bonded to two neighboring silanol groups. The elimination of water from neighboring silanol groups and isolated silanol groups requires 122 kJ/mol and 130 kJ/mol, respectively.

The silanol groups undergo acid-specific reactions; examples include:

1) With hydride ions (e.g., $LiAlH_4$) [366]

$$=Si-OH+H^- \rightarrow =Si-O^- + H_2$$

2) Deuterium exchange [367]

$$=Si-OH+D_2O \rightarrow =Si-OD+HDO$$

3) Neutralization [375]

$$=Si-OH+NaOH \rightarrow =Si-ONa+H_2O$$

4) Chlorination [376]

$$=Si-OH+SOCl_2 \rightarrow =Si-Cl+SO_2+HCl$$

5) Esterification with diazomethane [376]

$$=Si-OH+CH_2N_2 \rightarrow =Si-O-CH_3+N_2$$

Summaries of the surface chemistry of finely divided silicas are given in [377, 378].

6.1.5. Surface Modification with Silicon Compounds

Pyrogenic silicas are hydrophilic; however, a hydrophobic character is more favorable for some applications. This requirement can be met by reaction with dimethyldichlorosilane [379, 380]. The resulting silica thus is no longer wettable by water and was introduced on the market in 1962 under the name Aerosil R 972. Apart from chloroalkylsilanes, mainly hexamethyldisilazane, alkoxyalkylsilanes, and polydimethylsiloxanes are used.

Apart from methyl and longer chain alkyl groups, organofunctional groups can also be anchored on the surface of pyrogenic silicas by using silanes. This form of modification was first used on a large scale for applications in dental materials. A pyrogenic silica with a low specific surface area is treated with γ-methylacryloxypropyltrimethoxysilane to give an active filler that can be chemically bonded to the polymethylmethacrylate matrix and provides the dental material with very good mechanical properties [381 – 383].

Another use of pyrogenic silicas modified with organofunctional silanes is in toners. Until recently photocopiers generally fixed the latent electrostatic image on a positively charged selenium layer. An electrostatically negatively chargeable toner is required for development, which contains a hydrophobic pyrogenic silica to improve the flow behavior and chargeability. In more modern copier systems the latent image is formed on an organic semiconductor layer, which is negatively charged and requires a positively chargeable toner. However, pyrogenic silicas tend to become negatively charged due to the presence of silanol groups even in hydrophobized products. Hence, a surface modification is necessary to allow positive charging. This can be achieved by using aminofunctional silanes, which undergo hydrolysis-stable anchoring to the silica surface [384 – 386].

6.1.6. Characterization

Pyrogenic silicas can be characterized well on the basis of chemical and physicochemical data; however, these do not allow a definite assessment of their suitability for a particular use, which requires application testing. The following test methods have proved useful to test the suitability for many uses and also for the quality control of hydrophilic, pyrogenic silicas:

1) BET surface area, m^2/g (DIN 66 131)
2) Drying loss, % (2 h at 105 °C; DIN ISO 787/II)
3) pH value (4 % aqueous suspension; DIN ISO 787/IX)
4) Silicon dioxide content, % (fuming with hydrofluoric acid, relative to the substance heated at 1000 °C for 2 h)

For hydrophobic or surface-modified pyrogenic silicas it is advisable to determine the carbon content instead of the silicon content, since the former provides information on the extent of surface loading.

6.1.7. Uses

Reinforcing Fillers. Pyrogenic silica is used in large quantities as an active filler in silicone rubber (→ Silicones). In RTV silicone sealants, both hydrophilic and hydrophobic pyrogenic silica are used (ca. 10 %) to modify mechanical properties such as Shore hardness, tensile strength, and tear strength. High temperature vulcanizing (HTV) silicone rubber requires ca. 30 % pyrogenic silica to improve the mechanical properties. Hydrophobic pyrogenic silicas with strongly reduced thickening action are used in liquid silicone rubbers. Theories of the reinforcing action of pyrogenic silica in silicone rubber are discussed in [387 – 389].

The tensile strength of vulcanized silicones with various silica fillers is shown in Figure 45.

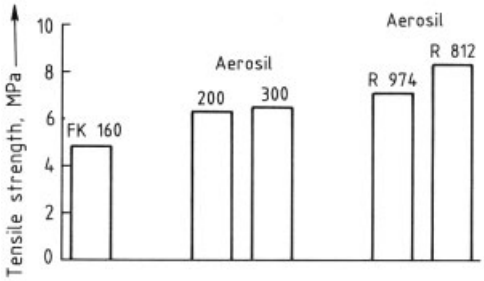

Figure 45. Tensile strength of silicone vulcanizates containing a precipitated silica (FK 160), two hydrophilic (Aerosil 200/300), and two hydrophobic (Aerosil R 974/R 812) pyrogenic silicas

Pyrogenic silicas are also used as active fillers in natural and synthetic rubber, particularly if extremely high mechanical properties are required, as in cable sheathing, seals, or conveyor belts.

In fluoroelastomers, both hydrophilic and hydrophobic silicas improve the mechanical properties [390, 391], even under the action of aggressive gases.

Thickening and Thixotropization. An important application for pyrogenic silicas is the adjustment of the rheological properties of liquid systems such as paints, thermosetting resins, and printing inks. The increase in viscosity generally plays a minor role; thixotropization and the development of a yield value are decisive in practice. These effects are generally attributed to the formation of a three-dimensional network of silica agglomerates, which fixes the liquid in "cells" and thus increases the viscosity. The silica particles interact via silanol groups. The spatial network is destroyed by mechanical stress (e.g., stirring or shaking) to an extent which depends on the intensity and duration of energy input, whereby the viscosity decreases. The silica network regenerates when the stress is removed, and the original viscosity returns. This theory has proved suitable for simple, nonpolar liquids, but is not applicable to complex and/or polar systems, in which adsorption and solvation effects determine the rheological activity [392], especially in the case of hydrophobic pyrogenic silicas.

Hydrophilic pyrogenic silicas show the best thickening action in nonpolar liquids (Fig. 46).

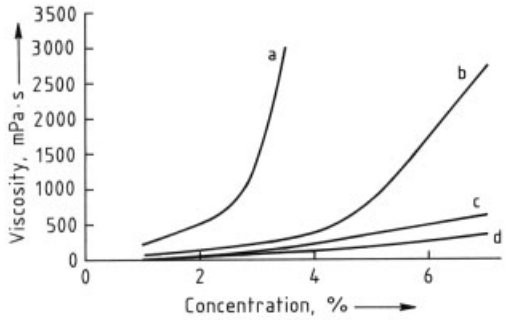

Figure 46. Thickening effect of a pyrogenic silica with 200 m^2/g specific surface area (Aerosil 200) in various solvents a) Xylene; b) Butylacetate; c) Diacetone alcohol; d) Butanol

Antisettling Agents. Closely linked with the control of the rheological properties of liquid systems is the action of pyrogenic silicas as antisettling agents. The establishment of a flow limit hinders settling of suspended particles, depending on their density, size, and shape. If settling does occur, pyrogenic silica generally prevents caking of the sediment and ensures easy dispersibility.

This effect is mainly utilized in filler- or pigment-containing paints and plastics. Not only suspensions of lighter solids can be stabilized, such as silica-based matting agents, but also heavy anticorrosion pigments such as zinc dust or micaceous iron oxide. Both hydrophilic

and hydrophobic pyrogenic silicas can be used as antisettling agents. The most suitable type must be determined for each individual case, whereby additional effects, such as the water repellency of hydrophobic products in corrosion protection systems, must also be taken into consideration [393, 394].

Dispersants. In solid-containing liquid systems, pyrogenic silicas reduce the reagglomeration of the solid particles during the dispersion process, whereby a more favorable state of distribution is achieved. This effect is particularly important in pigment-containing systems, where both the gloss and also the tinting strength, and in the case of carbon black pigments also the jetness, can be improved by adding small quantities of pyrogenic silica [393].

Free Flow Agents. Some powders exhibit poor fluidity or they agglomerate on storage, particularly under the influence of pressure or moisture. The causes of agglomeration include solid bridges, which form by drying or pressure sintering processes, liquid bridges, electrostatic charging, and van der Waals forces [395, 396]. This can make use of the material difficult or even impossible. In many cases this can be remedied by the addition of hydrophilic or hydrophobic silicas. The reasons for this are complex. Thus, the small silica particles can envelop the powder particles and act as "ball bearings", hydrophilic pyrogenic silica can adsorb moisture and "render it harmless", or hydrophobic pyrogenic silica can slow down the absorption of moisture. The significantly smaller silica agglomerates can decrease the van der Waals forces between two powder particles by acting as spacers [396, 397]. Pyrogenic silica is used in fire extinguisher powders, tablet mixtures, toners, powder coatings, and table salt.

Other uses of pyrogenic silica include:

1) Adsorbents
2) Antiblocking agents for plastic films
3) Coatings
4) Catalyst supports
5) Matting agents
6) Grinding agents
7) Polishing agents
8) Raw materials for silica glasses
9) Thermal insulation
10) Additive carriers

6.1.8. Industrial Hygiene and Safety

Pyrogenic silica does not cause silicosis on inhalation (see also Chap. 9). So far no indication for this illness has been found for those employed in production who undergo regular medical tests [398]. However, since it can lead to irritation of the respiratory tract due to the fine particles, a MAK of 4 mg/m^3 total dust in inhaled air was applicable in 1991. This MAK value is not exceeded if suitable precautions are taken, e.g., lowest possible dust development upon handling or extraction of air (ventilation). Otherwise dust masks should be used. According to the findings so far pyrogenic silica is basically harmless by oral intake [399, 400]. Pyrogenic silica can cause a feeling of dryness on the skin, which can be easily remedied by washing and creaming.

Pyrogenic silica is susceptible to electrostatic charging on handling. If it is likely to come into contact with combustible gases or vapors, it is necessary to take the corresponding safety measures, such as careful earthing and utilization of explosion-proof plants.

6.2. Electric-Arc Process

The reduction of quartz sand with coal in an electric-arc furnace gives gaseous silicon monoxide, which is oxidized to finely divided amorphous silica by atmospheric oxygen [401]:

$$SiO_2 + C \rightarrow SiO + CO$$

$$SiO + CO + O_2 \rightarrow SiO_2 + CO_2$$

In a variant of this process the thickening effect of the electric-arc silica is increased by adding steam during oxidation followed by thermal treatment [402].

The matting agent TK 900 [*112926-00-8*] used in paints is a silica produced by Degussa using this process.

6.3. Plasma Process

Finely divided amorphous silicas can be produced in plasma burners. In one case quartz sand

is reduced to silicon monoxide with methanol at ca. 2000 °C. Subsequently, reduction with steam is carried out [403]. At present such silicas are not available commercially. A schematic of the plasma process can be found in [404].

7. Precipitated Silicas

7.1. Introduction

Precipitated silicas, like pyrogenic silicas and silica gels, are a synthetic, finely divided, white, amorphous form of silicon dioxide. Precipitated silicas have only been produced commercially since the 1940s. In the meantime, they have become the most important group of silica products on the basis of the amounts produced. Worldwide production capacity in 1999 was ca. 1 100 000 t, compared to ca. 400 000 t in 1970. Capacities in Europe, America, and Asia amount to 400 000, 340 000, and 360 000 t, respectively.

The largest producers worldwide are Degussa-Hüls, Rhodia, Akzo, and Crosfield in Europe, as well as PPG and Huber in the United States. The Asian market is additionally supplied by numerous local producers. Among these, the companies with the largest capacities are Nippon Silica and Tokuyama Soda in Japan, and Sam Yung and Kofran in Korea.

7.2. Production

Raw materials for the production of precipitated silicas are aqueous alkali metal silicate solutions (e.g., water glass) and acids, generally sulfuric acid [405 – 420]. Precipitation with hydrochloric acid [421, 422], organohalosilanes [423, 424], carbon dioxide [425, 426], or a combination of the latter with mineral acids [427] are of minor economic importance.

In the reaction of alkali metal silicate solution with mineral acid, the silica precipitates according to the following equation:

$$Na_2O{\cdot}3.3\,SiO_2 + H_2SO_4 \rightarrow 3.3\,SiO_2 + Na_2SO_4 + H_2O$$

In contrast to silica gels, which are produced under acidic conditions, in this case precipitation is carried out in neutral or alkaline media. The properties of the precipitated silicas can be influenced by the design of the plant equipment and by varying the process parameters.

The production process consists of the following steps: precipitation, filtration, drying, grinding, and, in some cases, compacting and granulation (Fig. 47).

Numerous possibilities exist for carrying out the precipitation [405 – 427]. So far only batch precipitation processes have attained economic importance, although continuous precipitations have also been reported [420]. Generally, acids and alkali metal silicate solution are fed simultaneously into water in a stirred vessel with the formation of silica seeds. In the course of the precipitation, three-dimensional silica networks are formed, accompanied by an increase in viscosity. The networks are reinforced by further precipitation of oligomeric silica and grow further into discrete particles with a decrease in viscosity. The formation of a coherent system and thus a gel state is avoided by stirring and increasing the temperature.

The properties of the silica can be influenced by varying important precipitation parameters such as pH, temperature, duration, and solid concentration. Additionally, the rate of addition of the acid and the intensity of stirring, which can be supplemented by a shearing action, also influence the properties of the silica.

The separation of the silica from the reaction mixture and the washing out of the salts contained in the precipitate is carried out in filter aggregates such as rotary filters, belt filters, or filter presses (chamber, frame, and membrane filter presses).

The solids content of the washed filter cake is 15–25 % depending on the type of silica and filter; therefore, drying requires the evaporation of large quantities of water. Depending on the desired properties of the end product, drying is carried out in spray, nozzle spray, plate, belt, or rotary dryers. Special product properties can be achieved by spray drying the filter cake after redispersion in water or acid [410, 413, 428].

To control the desired fineness of the particles, various mills and, if required, classifiers [429] are used. Special degrees of fineness can be achieved by air-jet or mechanical grinding [410, 430]. The silica is separated from the air in cyclones or filters.

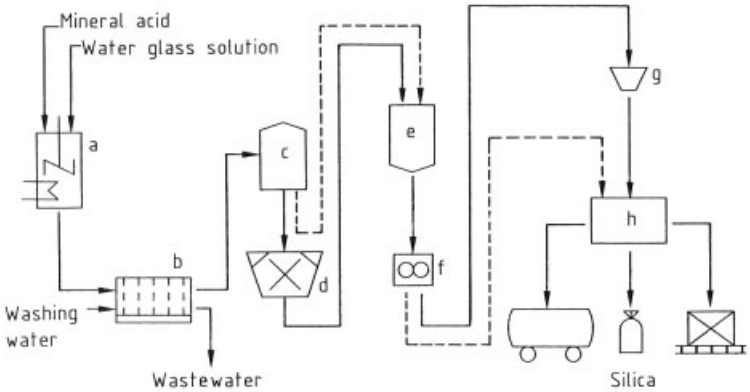

Figure 47. Process scheme for the production of precipitated silicas
a) Precipitation; b) Filtration; c) Drying; d) Grinding; e) Storage; f) Compacting; g) Granulation; h) Packaging

Figure 48 shows scanning electron micrographs of a spray-dried, unground silica and a conventionally dried, ground silica.

a)

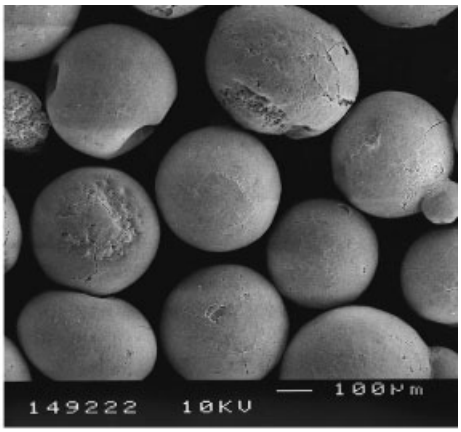

b)

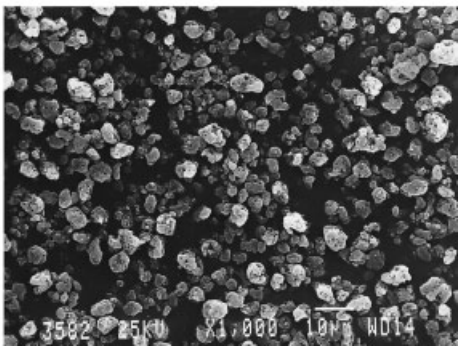

Figure 48. REM photographs of precipitated silicas
Top: spray-dried unground silica; bottom: conventionally dried, ground silica

Processes for the compression and granulation of precipitated silica have been developed [431, 432] to reduce the volume for transport and for certain uses and also to decrease the formation of dust on handling and processing.

Bags, big bags, containers, or transport silos are used for packaging and shipping. Information about the different methods of packaging and shipping is given in [433]; dust-free handling of precipitated silicas is described in [434]. In Germany the general dust limit of 4 mg total dust/m^3 air must be adhered to when handling precipitated silicas. Similar regulations apply in other countries.

7.3. Properties

7.3.1. Physicochemical Properties

Precipitated silicas consist mainly of spherical primary particles, generally intergrown in three-dimensional aggregates. These aggregates can pack together to form larger agglomerates. The terms primary particle, agglomerate, and aggregate are defined in DIN 53 206.

Some important physicochemical properties of several precipitated silicas are listed in Table 16.

The specific surface area is generally determined by the BET method (DIN 66 131; ISO 5794–1, Annex D), which measures both the outer surface and the surface of accessible pores. The specific surface area is of importance especially in reinforcement and adsorption applica-

Table 16. Physicochemical data and chemical composition of several precipitated silicas [CAS registry nos.: [*112926-00-8*] except for Sipernat D17 [*60842-32-2*]; appearance: flutty white powder except for Utrasil 7000 GR (granulate); behaviour towards water: hydrophilic except for Sipernat D17 (hydrophobic)]

	Ultrasil 7000 GR	**Sident 9**	**Sipernat 22**	**Sipernat 50 S**	**Sipernat 570**	**Acematt OK 412[n]**	**Sipernat D 17[o]**
BET surface area (N_2)[a], m^2/g	170	45	190	450	750	130	100
Mean particle size, µm	3000[l]	10.5[j]	100[l]	7.5[j]	6.7[j]	3[m]	7[j]
Tamped density[b], g/L	280	430	280	90	150	120	150
Loss on drying[c], %	5	3	6	6	6	5	4
Loss on ignition[d, g], %	4	3	5	5	5	11.5	7
pH value, 5 % in water [e]	6.5	6.9	6.5	6	6	6	8.0[p]
DBP absorption[f, g], g/100g	210[q]	120	260	325	250	265	225
SiO_2 content[h], %	98	99	98	98.5	99	99	98
Na content as Na_2O[h], %	1	0.5	1	0.6	0.5	0.3	1
Fe content as Fe_2O_3[h], %	0.03	0.03	0.03	0.03	0.02	0.03	0.03
Sulfate content as SO_3 [q], %	0.8	0.3	0.8	0.7	0.3	0.3	0.8
Sieve residue 45 µm[i], %		0.1	0.5	0.1	0.01	<0.01	0.1

[a] ISO 5794–1, Annex D.
[b] ISO 787–11.
[c] ISO 787–2.
[d] ISO 3262–11.
[e] ISO 787–9.
[f] ASTM D 2414.
[g] Based on dried substance.
[h] Based on ignited substance.
[i] ISO 787–18.
[j] Coulter Multisizer, 100 µm capillary.
[k] Coulter Multisizer, 140 µm capillary.
[l] Alpine sieving, ISO 8130–1.
[m] Determined from SEM photos.
[n] Matting agent.
[o] Contains ca. 2 % chemically bound carbon.
[p] 5 % in water : methanol (1 : 1).
[q] Based on original substance.

tions. Precipitated silicas can be produced with specific surface areas of 25–800 m^2/g.

The mean particle size is important in the use of precipitated silicas as matting or antiblocking agents for paint systems and plastic films. The particle size can be determined by various methods either on the powder or after incorporation in a medium (e.g., water, paint, silicone rubber) [435].

The tamped density (ISO 787–11) is a measure of the weight of the powder and thus gives indications of the spatial requirement for transportation and a general guide for handling. The tamped densities of precipitated silicas lie in the range of 50–500 kg/m^3.

The loss on drying (ISO 787–2) includes the majority of the water physically bound to the silica and for precipitated silicas generally lies between 2.5 and 7 % (on leaving the supplier).

The loss on ignition (ISO 3262–11) includes the additionally chemically bound water (a siloxane group is formed from two neighboring silanol groups with the loss of a water molecule). Ignition losses for precipitated silicas generally lie between 3 and 7 %.

Generally the pH value (ISO 787–9) of precipitated silicas is ca. 7.

The dibutyl phthalate (DBP) absorption (ASTM D2414) is a measurement of the absorptive capacity and for precipitated silicas lies in the range of 175–320 g/100 g. The absorptive capacity is important for the conversion of liquids and pastes into powders.

The determination of the sieving residue (ISO 787–18) provides an indication of the amount of difficultly dispersible fractions in a precipitated silica.

7.3.2. Surface Chemistry and Surface Modification

Precipitated silicas have two different functional groups on their surfaces: silanol (Si−OH) groups and siloxane (Si−O−Si) groups.

These two groups substantially influence the surface properties and hence the application

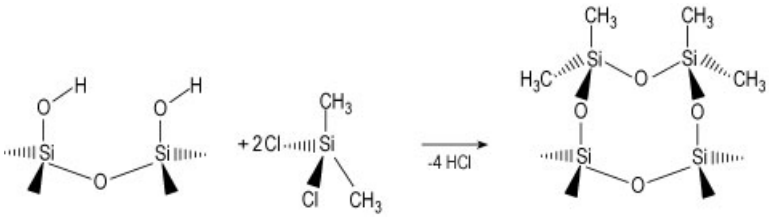

Figure 49. Schematic of the reaction of the surface silanol groups of a precipitated silica with dimethyldichlorosilane

properties of precipitated silicas. The surface has 5–6 silanol groups per nm^2, which results in the hydrophilic character of precipitated silicas. Whereas the siloxane groups are generally chemically inert, the reactivity of the silanol groups allows chemical surface modification [436] of precipitated silicas. Thus, the reaction with organosilanes [437 – 439] or silicone fluids [440, 441] leads to hydrophobic precipitated silicas.

Hydrophobization is carried out industrially both by wet processes (e.g., addition of organochlorosilane to the precipitate suspension; see Fig. 49) and dry processes (e.g., the reaction of precipitated silicas with silicone fluids).

In coating processes, in contrast, no chemical reaction takes place; the coating agents are adsorbed on the silica surface. Wax coating with wax emulsions has become industrially important [410, 442, 443].

Surface modification is advantageous for certain applications (e.g., free flow agents, defoamers, and matting agents; see Section 7.4).

7.3.3. Chemical Composition and Analysis

The chemical composition of precipitated silicas (see Table 16) is generally less important than the physicochemical data with regard to the application properties. Furthermore, commercially available precipitated silicas differ only slightly in their composition.

The SiO_2 content is determined gravimetrically by fuming off with hydrofluoric acid (ISO 3262–17); the analysis of the metal-containing impurities is performed by atomic absorption spectrometry (AAS). The sulfate content is determined by complexometric titration (DIN 38 405/5), colorimetric titration (DIN 2462/8), or potentiometric titration.

The carbon content of modified silicas is determined by converting the carbon-containing component into CO_2 (ISO DIS 3262–19: 1994).

A survey of the chemical analysis of precipitated silicas is given in [444].

7.4. Uses

Precipitated silicas can be tailor-made to meet the requirements of various uses. Only the most important of the numerous applications (Fig. 50) are mentioned here.

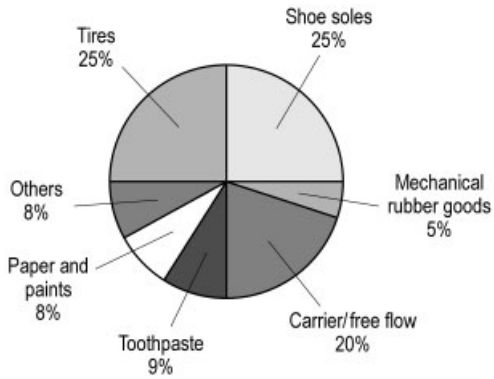

Figure 50. Worldwide consumption and use of precipitated silica in 1999

The oldest and most important use of precipitated silicas is the reinforcement of elastomer products such as shoe soles, technical rubber articles, cable sheating, and tire compounds [445, 446].

The addition of 20–100 parts by weight of precipitated silica to 100 parts by weight of natural or synthetic *rubber* improves the tensile strength, hardness, tear strength, and abrasion resistance of the vulcanized material [447]. Furthermore, special precipitated silicas enable the

production of transparent vulcanized materials. The reinforcing effect of precipitated silicas is superior to that of natural or mineral fillers and, in contrast to carbon blacks, they allow the production of white and colored rubber articles.

In *cables*, they are used mainly in sheathing compounds for cables that are not fixed and must be differentiated from the background by a colored exterior (e.g., mining or excavator-hauling cables). The high resistance of the cable sheath to friction and tearing protects the inside of the cable from abrasion and damage.

Since the early 1990s precipitated silica has been used as a high-performance reinforcing filler in rubber compounds for *tires*. To take full benefit of silica in terms of tire performance, it is essential to use highly dispersible silica grades such as Ultrasil 7000 GR (Degussa-Hüls) or Zeosil 1165 MP (Rhodia) with a silane coupling agent. The use of highly dispersible silica in tread compounds leads to significant improvements in rolling resistance and wet traction of tires. Due to the possibility of achieving the best compromise of the three main tire-performance criteria ("magic triangle") – rolling resistance, wet traction, and tread wear – silica is now widely used in Europe, and demand in North America and Asia is growing significantly. The use of silica in the bodies of car and truck tires is becoming increasingly interesting owing to the need to further reduce rolling resistance to cut fuel consumption and CO_2 emissions [448 – 450].

Precipitated silicas are also used as reinforcing agents in *silicone rubber* [451, 452], where they replace the more expensive pyrogenic silicas in certain formulations.

In *thermoplastics*, precipitated silicas are used to improve specific properties. They act as antiblocking agents and are used to prevent plate-out effects in films and film production. They are also used to improve the mechanical properties of PVC flooring.

A further major application is the use as *carrier silicas* for liquids and semiliquids and as free flow agents for powder formulations, particularly of hygroscopic and caking substances. Of particular importance is high adsorptivity and good flowability of the silica. Such precipitated silicas absorb liquids or solutions to give powder concentrates that contain up to 70 % of the liquid [453]. In this way liquid materials are converted into a dry, free-flowing form, which can be mixed in any ratio with other dry substances. This is of particular importance in the animal-feed industry for the mixing of feed additives [454]. Materials converted into powder form include: choline chloride, ethoxyquin, molasses, and vitamin E acetate [455]. Fats can be converted into free-flowing powders for use as milk-substitute feeds [453, 455], mainly for the fattening of calves (Fig. 51). The same effects are utilized in the food industry, where products such as coffee whiteners (nondairy creamers) and mixed spices are produced and kept intact with the aid of the specific properties of synthetic silicas [462, 463]. Precipitated silicas are used as carriers for crop protection agents and insecticides [456, 464, 465].

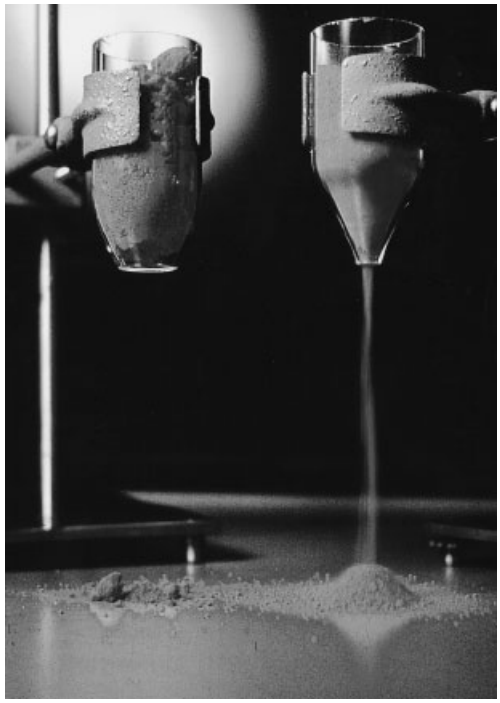

Figure 51. Milk substitutes
Left: without addition; right: after addition of 1.5 % Sipernat 22 S

Precipitated silicas are being used increasingly in *toothpastes* as cleaning agents, since they clean effectively with a minimum of scratching [457]. Additionally, they control the rheological properties of toothpastes and also enable the production of transparent gels [458].

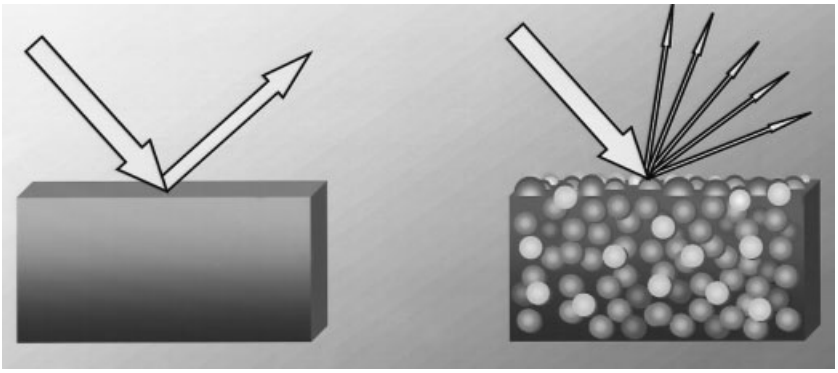

Figure 52. Reflection of light from shiny (left) and matted surfaces (right)

The surfaces of objects are matted for reasons of fashion but also to increase safety by avoiding dazzle. Special silica *matting agents* have been developed for matting of paints and varnishes [410, 429, 443]. The matting effect is the result of roughening of the surface on the microscopic scale, whereby the light is diffused and no longer directionally reflected (Fig. 52).

In the *paper industry* precipitated silicas are used as advanced fillers in, for example, newsprint to improve adsorptivity for printing ink. Strike through is reduced, and the coefficient of friction is adjusted. Silicas are used in special coated papers for ink-jet and direct thermal printing to enhance ink absorption [466].

Hydrophobic precipitated silicas are used increasingly in mineral-oil and silicone-oil *antifoam agents* to increase the antifoaming effect. They are used as foam regulators in laundry detergents [459] and in the production of paints and paper.

Further uses for precipitated silicas include the purification and stabilization of beer [460], and the analysis of blood [461].

Trade Names. Trade names for precipitated silicas include Sipernat, Ultrasil, Sident, Acematt (Degussa-Hüls); Hi-Sil, Flo-Gard, Lo-Vel (PPG); Tixosil, Zeosil (Rhodia); Zeodent, Hübersil (Hüber); Ketjensil, Perkasil (AKZO); Nipsil (Nippon Silica); and Neosyl (Crosfield).

7.5. Industrial Hygiene and Safety

Precipitated amorphous silica is nontoxic in rats after oral intake [462 – 464]. Neither eye irritation nor skin irritation is observed [465, 466]. Dryness of the skin after skin contact can be alleviated by washing and creaming. Inhalation of precipitated amorphous silica does not result in silicosis [467]. In addition, several epidemiological studies show no indication of silicosis in exposed employees [468 – 471].

When working with precipitated amorphous silica, the MAK value (4 mg/m^3 total dust) must be observed [472]. When appropriate precautions such as lowest possible dust development during handling and ventilation are taken, this value is not exceeded. Otherwise, dust masks should be worn.

8. Porosils

8.1. Introduction

Porosils are crystalline porous silicas of the general composition $SiO_2 \cdot M$, where M is an organic or inorganic guest molecule residing within the pores of the silica host framework. R. M. Barrer coined the name porosils to differentiate between porous aluminosilicate zeolites and a new class of compounds of distinct composition [473]. Due to their porous crystal structure porosils have properties similar to the zeolites; however, differences in chemical composition are responsible for their specific properties in catalysis, sorption, and molecular sieving.

Work in the field was initiated in 1978 by Flanigan et al. [474] who reported on the synthesis and crystal structure of the siliceous analogue of the commercially most important zeolite ZSM-5.

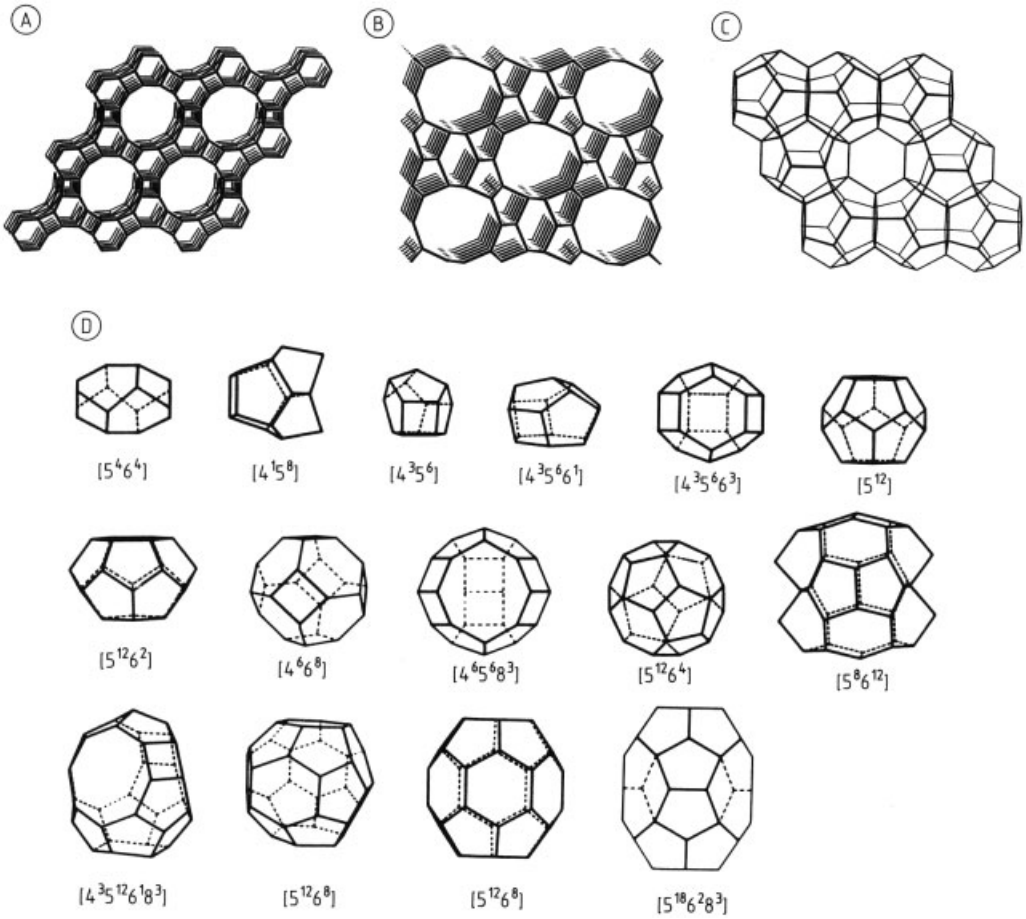

Figure 53. Skeletal projections of porosil framework structures
Si atoms are located in the nodes of the framework O atoms, which are close to the midpoint between 2 Si atoms, but are omitted for clarity
A) Projection of the AFI structure along the channel axis. The channel pore opening is confined by 12 Si atoms (12 MR); B) Projection of the FER, TON, MTT, MEL, and MFI structures along the straight channel axis. The channel pore opening is confined by 10 Si atoms (10 MR); C) Projection of the dodecasil layer. Pentagonal dodecahedra are arranged in layers which are the basic structural feature of DOH, MTN, DDR, and decadodecasil 3 H; D) Compilation of types of cage as found in clathrasil structure types ($[5^{12}]$ represents pentagonal dodecahedron, i.e., 12 five-membered rings)

The guest species M can be expelled by thermal treatment to give the pure silica form of the porosil. Therefore, porosils are considered to be the porous polymorphs of the structure family SiO_2. Up to now, 34 porosil structure types have been made, either by template-directed synthesis or post-synthesis treatment of aluminosilicate zeolites (Table 17) [475, 476]. In the synthesis of porosils the guest species M act as templates for the pores, defining their size and geometry [477]. During the post-synthesis treatment framework aluminum is substituted by silicon, and M is atmospheric gases or sorbate molecules [478].

Porosils are subdivided into zeosils and clathrasils according to the pore structure of the silica framework. Zeosils have channel-type pores with pore openings larger than 0.4 nm (Fig. 53). The channels may intersect to give a two- or three-dimensional pore system (see Table 17). Clathrasils have cage-type voids with pore openings much smaller than the free diameter of the cage and less then 0.45 nm (Fig. 53).

With the exception of melanophlogite, a rare mineral formed by low-temperature, low-pres-

Table 17. Porosil structure types [475, 476]

Porosil	Framework density (Si atoms per 1000 Å^3)	IUPAC code	Pore system**
Zeosils			
Silica-faujasite*	12.7	FAU	3 D 12 MR
Silica-ECM*	12.7	ECM	3 D 12 MR
Silica-A*	12.9	LTA	3 D 8 MR
ITQ-7	15.0	ISV	3 D 12 MR
Silica-beta	15.1	BEA	3 D 12 MR
Silica-levyne	15.2	LEV	2 D 8 MR
Silica-offretite*	15.5	OFF	1 D 12 MR
ITQ-3	15.7	ITE	2 D 8 MR
ITQ-1	15.9	MWW	2 D 10 MR
Silica-mazzite*	16.1	MAZ	1 D 12 MR
CIT-5	16.8	CFI	1 D 14 MR
UTD-1F	17.1	DON	1 D 14 MR
ITQ-4	17.2	IFR	1 D 12 MR
Silica-mordenite	17.2	MOR	1 D 12 MR
Silica-SSZ-24	17.5	AFI	1 D 12 MR
Decadodecasil 3R	17.6	DDR	2 D 8 MR
Decadodecasil 3H	17.6		2 D 8 MR
Silica-ZSM-5	17.9	MFI	3 D 10 MR
Silica-ZSM-11	17.9	MEL	3 D 10 MR
Silica-ferrierite	19.3	FER	1 D 10 MR
Silica-ZSM-12	19.4	MTW	1 D 12 MR
Silica-theta-1	19.7	TON	1 D 10 MR
Silica-ZSM-23	19.7	MTT	1 D 10 MR
VPI-8	20.2	VET	1 D 12 MR
Silica-ZSM-48	19.9		1 D 10 MR
MCM-35	20.7	MTF	1 D 8 MR
Clathrasils			
Silica-A*	12.9	LTA	0 D 8 MR
Silica-levyne	15.2	LEV	0 D 8 MR
ITQ-3	15.7	ITE	0 D 8 MR
Silica-sodalite	17.4	SOD	0 D 6 MR
Decadodecasil 3R	17.6	DDR	0 D 8 MR
Decadodecasil 3H	17.6		0 D 8 MR
Octadecasil	17.6	AST	0 D 6 MR
Sigma-2	17.8	SGT	0 D 6 MR
Dodecasil 1H	18.4	DOH	0 D 6 MR
Dodecasil 3C	18.6	MTN	0 D 6 MR
Silica-RUB-10	18.7	RUT	0 D 6 MR
Melanophlogite	19.0	MEP	0 D 6 MR
Nonasil	19.3	NON	0 D 6 MR
MCM-35	20.7	MTF	0 D 8 MR

* Dealuminated zeolites.
** Pore dimensionality and pore size (n-membered ring of SiO_4 units).

sure metamorphosis in, e.g., serpentine veins [479], all porosil structure types are synthetic materials. The ever increasing number of new porosil structure types reflects the efforts in the systematic investigation of the mechanism of formation of zeolites, for which porosils represent a simplified system. Besides porosils, water (clathrate hydrates), $AlPO_4$ (AlPO's), $GaPO_4$, and other III–V oxides (e.g., arsenates) also form neutral inorganic porous framework structures when synthesized in the presence of structure-directing guest molecules [475].

8.2. Physical and Chemical Properties

The physical and chemical properties of porosils are closely related to their crystal structure and composition. Porosils are colorless materials of hardness close to that of quartz. Since they are framework silicas, they exhibit no pronounced cleavage. Refractive indices vary systematically with the density of the silica framework (Fig. 54) [480]. This confirms that the nature of the Si−O bond accounts for most of the physical properties and is closely related to the natural

polymorphs of the SiO_2 structure family. The porosils are thermodynamically metastable and transform reconstructively to cristobalite at high temperature (e.g., melanophlogite: T > 950 °C [481]). Below 200 °C structural phase transitions to lower symmetry space groups have been reported for melanophlogite [482], dodecasil 3 C [483], decadodecasil 3 R [484], silica-ZSM-5 [485], and silica ZSM-11 [486].

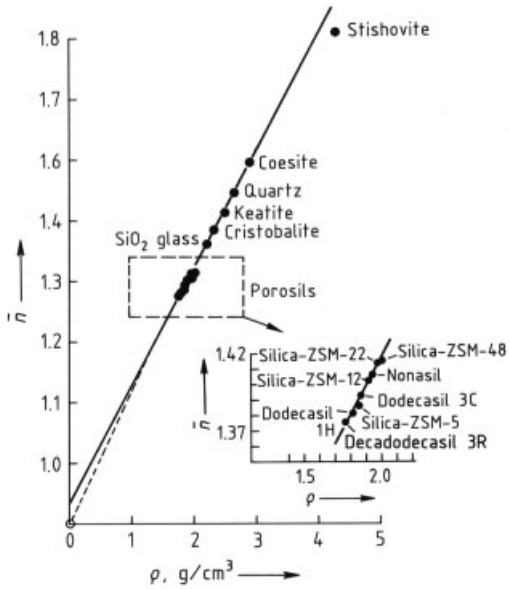

Figure 54. Density of the silica framework of porosils and dense silicas versus refractive indices

For dodecasil 3 C, three consecutive phase transitions have been reported, with transformation from cubic to tetragonal, orthorhombic, and finally monoclinic symmetry [487]. This reflects the flexibility of the silica framework. Several phase transitions are ferroic (e.g., ferroelastic) indicating interesting physical properties for the material.

A number of porosils in the high-symmetry form exhibit negative volume expansion coefficients β. For example, silica-ZSM-5 contracts in the temperature range 393–975 K with $\beta = -15.1 \times 10^{-6}\ K^{-1}$ [488].

Since the composition of the host framework is SiO_2, porosils are electroneutral. The pores, therefore, contain only electroneutral guest species (i.e., ion pairs, molecules, and/or atoms). In contrast to aluminosilicate zeolites, porosils are hydrophobic [474, 489], preferentially sorbing aprotic organic molecules. The weak Lewis acidity of the electroneutral SiO_2 framework of the porosils reduces considerably their potential as materials for heterogeneous acid catalysis, for which the isostructural aluminosilicate zeolites are widely used [490]. The pore geometry of the porosil structure types varies from zero-dimensional pores (i.e., cavities) to three-dimensional pore systems (i.e., intersecting channels) (see Table 17). The pore openings are confined by $[SiO_4]$ units arranged in a ring (Fig. 53) (e.g., ten-membered ring, 10 MR: 10 SiO_4 units per ring). In general the sorption properties of the materials are related to the largest pore opening of the structure, which provides for the release of guest molecules and the access of sorbates to the internal surface. In clathrasils the maximum pore opening (6–8 MR) is too small for the exchange of the occluded species. The template-free material can be obtained only by calcination at > 600 °C. For 6 MR as maximum pore opening only very small atoms or molecules such as H_2 or He could be sorbed onto the silica framework (Table 18). Zeosils with pore openings of 8–12 MR are calcined at lower temperatures and could be used for the selective sorption of, e.g., *n*-alkanes (8 MR materials) or *p*-substituted aromatic compounds (10 MR materials) (Table 18). Because of the uniform character of the pore opening and the possibility of tailoring the size of the pores, porosil- containing membranes, such as those already available for zeolites, could be useful for separation processes [491, 492].

Table 18. Idealized pore diameters and potential sorbates for various pore openings

Pore opening	Planar ring diameter, nm	Potential sorbates (diameter in nm)
6 MR	0.27	He (ca. 0.23) H_2 (ca. 0.23)
8 MR	0.44	alkane (ca. 0.43)
10 MR	0.60	aliphatic and aromatic six-membered rings (ca. 0.6)
12 MR	0.77	substituted ring systems

8.2.1. Zeosils

Physical and chemical properties specific to zeosils are related to their crystal structure. The channel-like pores allow access to the internal

surface of the material for sorbates that meet the geometrical requirement of the pore window.

The free diameter of the individual materials, however, varies considerably due to the flexibility of the silica framework. Idealized values for the pore openings are given in Table 18. Because of the hydrophobic properties of the silica framework and the geometry of the internal pore system, zeosils are of particular interest for various applications. In the presence of water zeosils sorb selectively hydrophobic organic molecules in the vapor and liquid phase [493]. Sorption of monomers into the silica framework followed by polymerization leads to one-, two- or three-dimensional polymeric filaments inside the zeosil framework, as already demonstrated for zeolites [494]. The deposition of semiconducting materials [495] and metal atom clusters [496] inside, e.g., the zeosil cavity of silica-A or silica-faujasite creates new materials with interesting properties due to the limited size of crystallites built within the pores. The sorption of polar organic molecules such as *p*-nitroaniline also leads to changes in symmetry and polarization of the host–guest composites, leading, e.g., to nonlinear optical properties [497]. For a review on new applications of zeosils, see [498].

8.2.2. Clathrasils

The clathrasil structure types have cagelike pores. Species inside the cavities are too large to pass through the 4 MR, 5 MR, 6 MR, or 8 MR windows without their decomposition. In general, clathrasils in the as-synthesized form have no exchange capacity or sorption properties. After calcination at high temperature (> 600 °C) the decomposition of the guest species may lead to a guest-free porosil with very narrow pores that is suitable for sieving and sorption processes (cf. Table 18). The properties of clathrasils are, therefore, linked to properties introduced by the guest molecules clathrated in the formation process. Because of the high thermal and chemical stability of the silica host framework, clathrasils have been proposed as containers for radioactive gases (e.g., krypton) [499]. The synthesis of the clathrasil dodecasil 3 C with a number of nonglobular guest molecules leads to a distortion of the cubic symmetry of the silica framework. At room temperature the materials have acentric symmetry and, therefore, nonlinear optical properties [500]. The temperature of the phase transition to the acentric space group depends on the nature and the concentration of the guest molecule in the silica framework and is in the range between slightly below room temperature and ca. 200 °C.

8.3. Manufacture of Porosils

With the exception of melanophlogite, all porosil structure types are synthetic materials made directly from an active form of SiO_2 or by dealumination of high-silica zeolites.

8.3.1. Synthesis of Porosils

Structure-directing templates play a key role in the synthesis of porous silicas [501, 502]. Since the templates are occluded within the pores during synthesis, it is their size and shape which determine the size and shape of the pores of the silica framework. Organic amines have proved to be the most efficient templates. Up to now ca. 100 different molecules have successfully been used for the synthesis of porosils [501, 503, 504]. Globular species such as xenon, quinuclidine, and 1-aminoadamantane preferentially stabilize the cage-like voids characteristic of clathrasils. Chain-like amines such as the α,ω-diamines of propane, butane, pentane, etc., stabilize channel-like voids. The size of the pore depends on the size of the template. Whereas pyrrolidine leads to preferential formation of dodecasil 3 C by occupying a 16-hedral cage, 1-aminoadamantane is occluded within a 20-hedral cage in dodecasil 1 H (Fig. 55). Zeosil structure types with 10 MR channels are formed in the presence of aliphatic chain-like templates (α, ω-diamino-*n*-alkanes); 12 MR channels are formed with molecules containing sequences of *p*-substituted heterocycles (e.g., 4,4′-trimethylenedipiperidine).

8.3.2. Dealumination of Aluminosilicate Zeolites

Porosils marked with an asterisk in Table 17 have been obtained exclusively by post-synthesis treatment of aluminosilicate zeolites. Aluminum

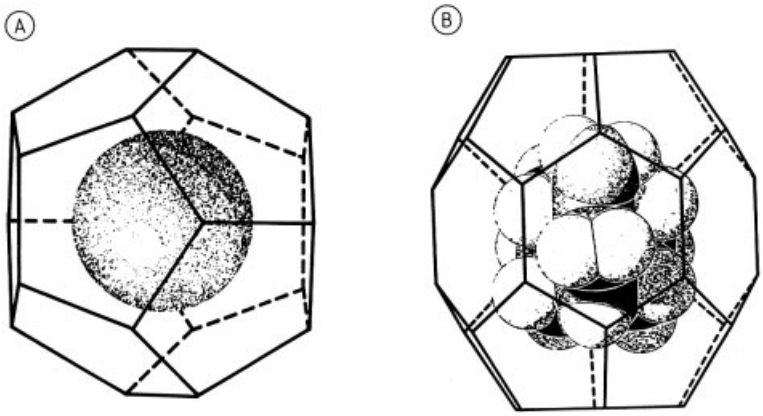

Figure 55. A) Van der Waals model of Xe as guest atom within the pentagonal dodecahedral cage; B) Van der Waals model of 1-aminoadamantane as guest molecule within the icosahedral cage

is selectively extracted from the silicate framework and replaced by silicon [505]. The dealumination procedure includes successive chemical and hydrothermal treatment of the material in order to extract and remove the framework aluminum atom. A typical procedure is the calcination of the synthesis product at 300–500 °C to expel the organic template, exchange of the cations with ammonium, and steaming at ca. 600 °C to remove the aluminum.

8.3.3. Formation of Melanophlogite

So far melanophlogite is the only known natural porosil. In synthesis experiments the conditions of formation have been simulated and are of low temperature (ca. 100–200 °C) and low (authogenic) pressure. The material has been described as a trace occurrence from five localities in very different geological settings. However, the nature of the porosil structure types and the wide range of structural features suggest that further natural variants may exist, especially in hydrothermally active areas (e.g., geysers).

9. Toxicology

The effect of the dust that arises when working with silica is of primary importance from a toxicological point of view, especially in the case of long-term exposure of several hours per day. This applies to naturally occurring and synthetic as well as to crystalline and amorphous. As far as industrial medicine is concerned, local contact with the skin and mucous membranes is important. In the case of inhalation of the dust, the effect on the respiratory tract is significant.

The effect of silica-containing dust on the respiratory tract depends on:

1) The physical and chemical properties (e.g., the particle size and the specific surface area)
2) The concentration administered
3) The daily and total duration of exposure

9.1. Experiences with Humans

After long-term exposure at work, the effects of crystalline or amorphous SiO_2 dust on the respiratory tract of humans are well known and have been widely reported (for reviews, see [506 – 509]).

Crystalline Silica Dust. The classification into common, accelerated, and acute forms of silicosis [506] is based on the time interval between the beginning of exposure to silica and the first appearance of silicotic changes that are detectable in a X-ray picture. A rapid manifestation and an accelerated course of the disease are caused primarily by long-term inhalation of higher concentrations of SiO_2 dust.

The common form of silicosis (simple silicosis) occurs roentgenologically after about twenty or more years of exposure. The content of

crystalline SiO_2 in the dust that causes silicosis in workers is often less than 30 % [506].

The accelerated form of silicosis (5–15 years) develops after a shorter exposure to higher concentrations of SiO_2 dust. Progressive massive fibrosis is often observed on the X-ray picture [506].

The acute form, also known as silicoproteinosis, develops in 1–3 years. The clinical symptoms include rapidly deteriorating function of the lungs, which is always fatal.

The progression of silicosis even after the termination of exposure is especially characteristic, and there is still no effective treatment.

For this reason, it is essential that the legal regulations for working with crystalline SiO_2 are obeyed [510 – 512].

Amorphous Silica Dust.

Amorphous Synthetic Silica. From experimental long-term inhalation tests on rats, it is known that high concentrations of some synthetic amorphous silicas can cause nodular (fibrotic) changes in the lungs of the animals during the exposure phase [513]. However, the regression of the nodules (fibrosis) in the lungs after the termination of the exposure phase is characteristic of the biological action of these amorphous silica dusts. In a treatment-free, follow-up observation phase of several months, a clear time-dependent regression of the changes in the lungs of the experimental animals could be detected after three, eight, and up to twelve months (see below) [513, 514]. In the case of the neighboring lymph nodes, the regression of the morphological changes observed within the inhalation phase was established histologically [513, 514].

In-depth investigations carried out by company doctors over periods of up to 12 years [515] and from 1 to 32 years [516] have confirmed that until now long-term contact with amorphous synthetic silicas has not resulted in a single case of silicosis.

Established MAK [511] and TLV regulations [512] must be complied with when working with amorphous synthetic silicas. If these fixed limiting values are exceeded – especially for a longer period of time – an overtaxing of the respiratory passages and a corresponding biological defense reaction in the bronchial tubes and, especially, in the lung tissue must be expected.

Vitums et al. observed that fibrosis of the lungs was caused by exposure to flue dust containing silica with a wide particle-size distribution, such that as produced in the manufacture of ferrosilicon or quartz glass [517].

Amorphous Natural Silicas. A distinction is made between exposure to amorphous synthetic silica dust and exposure to natural amorphous silica dust (diatomite). The latter can contain small amounts of crystalline SiO_2 and, according to some studies, can cause silicosis in humans on long-term exposure [507].

Carcinogenicity of Silica Dusts in Humans. In many cases, industrial workers exposed to crystalline silica dust exhibit a higher incidence of lung cancer. However, exposure to silica dust is often complicated by other hazards such as polycyclic aromatic hydrocarbons, asbestos, and smoking, which makes a causal analysis extremely difficult. According to Goldsmith et al. [518], at least two studies show a strong association between silicosis and lung cancer: the analysis of the Swedish pneumoconiosis register by Westerholm [519] and a similar study by Finkelstein [520] in Ontario, Canada. However, the same authors [518] also stated that a completely convincing causal relationship cannot be concluded. A possible association between silicosis and lung cancer is also discussed in [521], based on data from a clinical cancer monitoring register. A relationship between tumors of the upper intestinal tract and the ingestion of quartz particles in food in North China has been suggested [522]. A correlation was made between the incidence of cancer of the esophagus and the consumption of plant products that contain silica fragments and fibers for a region in North China, in northeast Iran, and in Transkei.

9.2. Animal Experiments

Long-Term Inhalation of Silica Dust. In specifically designed, practice-oriented long-term inhalation studies, various types of crystalline and amorphous silica dust have been thoroughly tested on rodents. Rats [513, 514, 523, 524] were mainly used, but rabbits [514, 525, 526] and guinea pigs [514] were also employed.

The elimination rate in rats has been determined in an inhalation experiment with highly

disperse *amorphous silica*. In the case of Aerosil 200, for example, after the termination of inhalation, 65–75 % of the dust was eliminated in one month, 80–90 % in two months, and 95 % in three to five months. It is possible that the rate of solubilization of these hydrophilic silicas also plays a role [513].

It was found in retention studies on rats [513] that in the case of longer term inhalation (28 times four hours in 36 days) of highly disperse *amorphous silica* dust, the continuous elimination during the exposure phase is very high. This means that the formation of large SiO_2 deposits in the lungs is improbable [513]. In contrast, in long-term inhalation experiments with *quartz* dust the tendency towards formation of deposits was higher [513].

After the termination of a long-term inhalation study on rats with Aerosil 200 (12 months, 4 per day; concentration: 45 mg/m^3), some of the animals were subjected to treatment-free observation for another 3 to 8 months and then killed [513]. Subsequently, a histological investigation and chemical analysis of SiO_2 in the lungs and lymph nodes were performed. In the follow-up observation period, it was found that the nodules in the lungs were reduced to small residual foci. The lymph nodes, which were still enlarged after three months, had greatly regressed after eight months. Similar observations have also been made in previous inhalation experiments with Aerosil.

In a long-term inhalation experiment on rats with *quartz* [524] (12 months, 5 per day, 5 per week; concentration:10 mg/m^3; standard quartz No. 12, Dörentrup quartz), some of the animals were subjected to treatment-free observation for another 4 weeks to 6 months after the termination of exposure and then killed. In this follow-up observation period, it was possible to detect silicotic reactions in the lungs and reticulohistiocytic cells typical of quartz in the lymph nodes. There was an increase in the silicotic nodules in the lungs after six months and fibrosis of the mediastinal lymph nodes had clearly increased.

Examination of lungs and accompanying lymph nodes of the individual groups of animals at various time intervals showed that in the case of *quartz*, a long residence time in the lungs must always be expected, but that deposits of finely divided *amorphous silica* are relatively quickly and extensively eliminated [513, 524]. This is true of amorphous silicas with a specific surface area of > 100 m^2/g [513].

Injection of Silica Dust. The tissue reaction typical of a particular silica dust can be determined in animal experiments by means of a single intratracheal or intraperitoneal injection.

The intratracheal injection of a high dose of *quartz* resulted in a massive local accumulation of dust [524]. The tissue reaction caused by an injection of this type is similar to the picture of acute silicosis. Fast reactions of this kind in the lung tissue – comparable with the exposure of workers to high concentrations of quartz [515] – can also be produced in animal experiments with very high inhalation concentrations (400 mg/m^3) [524].

The general effect of a particular dust (acute toxicity) and the local harmful (cytotoxic) effects are deduced from the tissue reaction observed in rats after intraperitoneal SiO_2 injections. Cytotoxicity can occur in the form of nodular fibrosis (dust nodules) or surface fibrosis that leads to extensive coalescence of the abdominal organs [513].

The reaction of lung tissue on longer term inhalation of SiO_2 dust cannot be directly inferred from the results of the intratracheal and intraperitoneal injections. Thus, the physiologically important reactions of the organism in the course of SiO_2 retention and SiO_2 elimination can be determined only in long-term inhalation tests with a sufficiently long, treatment-free, follow-up observation period.

The morphological changes in the lungs after long-term inhalation of *natural crystalline* SiO_2 progress after the termination of exposure, whereas with *synthetic amorphous silica* dust the morphological changes regress after the termination of exposure (reversibility).

Carcinogenicity of Silica Dusts in Animal Tests. Lymphomas have been found in rats after intrapleural, i.p., i.v., and direct application in the thymus gland of various silica modifications [527, 528]. The tumor incidence differed for various species of rat, and no tumors were observed in mice and hamsters. A synergistic effect has been reported for silica dust and benzo[*a*]pyrene [529], as has been found for other dusts (e.g., coal, TiO_2). This is presumably the result of chemicophysical surface ef-

fects, which may also be involved in tumor promotion. Tests have shown that Min-U-Sil causes tumors in rats in the intratracheal test [530]. Tissue scarring has been discussed as the cause. Min-U-Silis strongly fibrogenic and extremely finely divided.

10. References

1. R. Siever in Ref. [4]
2. H. Füchtbauer (ed.): *Sedimente und Sedimentgesteine*, 4th ed., Schweizerbarth, Stuttgart 1988.
3. *Ind. Miner. (London)* **236** (1987) 22–39.
4. K. C. Condie in K. H. Wedepohl (ed.): *Handbook of Geochemistry*, vol. **II-2**, Springer, Berlin 1970, Section 14.
5. O. W. Flörke et al., *Neues Jahrb. Mineral. Abh.* **163** (1991) 19–42.
6. V. Swamy, S. K. Saxena, B. Sundman, J. Zhang, *J. Geophys. Res.* **99** (1994) 11787–11794.
7. R. D. Aines, G. R. Rossman, *J. Geophys. Res.* **89** (1984) 4059–4071.
8. A. Nicolas, J. P. Poirier: *Crystalline Plasticity and Solid State Flow in Metamorphic Rocks*, Chap. 5.9. Wiley, London 1976.
9. O. W. Flörke, A. Nukui, *Neues Jahrb. Mineral. Abh.* **158** (1988) 175–182.
10. C. Frondel: "Silica Minerals," *The System of Mineralogy*, 7th ed., vol. **III**, Wiley, New York 1962.
11. M. J. Mombourquette, J. A. Weil, *Can. J. Phys.* **63** (1985) 1282–1293.
12. H. Schneider, O. W. Flörke, *Neues Jahrb. Mineral. Abh.* **145** (1982) 280–290.
13. H. Schneider, *Neues Jahrb. Mineral. Monatsh.* 1986, 433–444.
14. M. J. Buerger, *Amer. Mineral.* **39** (1954) 600–614.
15. L. Pauling, *Amer. Mineral.* **65** (1980) 321–323.
16. R. J. Hill, G. V. Gibbs, *Acta Crystallogr.* **B 35** (1979) 25–30.
17. M. O'Keefe, B. G. Hyde, *Acta Crystallogr.* **B 34** (1978) 27–32.
18. W. H. Baur, *Acta Crystallogr.* **B 36** (1980) 2198–2202.
19. F. Liebau: *Structural Chemistry of Silicates*, Springer-Verlag, Berlin–Heidelberg 1985.
20. W. A. Harrison: *Electronic Structure and the Properties of Solids*, Freeman, San Francisco 1980, pp. 257–288.
21. R. D. Shannon, *Acta Crystallogr.* **A32** (1976) 751–767.
22. J. R. G. Da Silva, D. G. Pinatti, C. E. Anderson, M. L. Rudee, *Philos. Mag.* **31** (1975) 731–737.
23. B. Martin, O. W. Flörke, E. Kainka, R. Wirth, *Phys. Chem. Minerals* **23** (1996) 409–417.
24. M. T. Dove, *Amer. Mineral.* **82** (1997) 213–244.
25. G. V. Gibbs, C. T. Previtt, K. J. Baldwin, *Z. Kristallogr.* **145** (1977) 108–123.
26. L. Levien, C. T. Prewitt, *Amer. Mineral.* **66** (1981) 324–333.
27. S. Haussühl: *Kristallphysik*, Chap. 11.4, Verlag Chemie, Weinheim, Germany 1983.
28. R. E. Newnham: *Structure-Property Relations*, Springer, Berlin 1975.
29. G. Dolino in R. Blinc, A. P. Levanyuk (eds): *Incommensurate Phases in Dielectrics*, vol. **2**, Elsevier, Amsterdam 1986, 205–232.
30. W. A. Deer, R. A. Howie, W. S. Wise, J. Zusmann: "Framework Silicates: Silica Minerals, Feldspathoids and the Zeolites" in *Rockforming Minerals* Vol. 4B, 2nd ed., The Geological Society Publishing House, Bath, UK, 2004, pp. 15ff.
31. K. Kosten, H. Arnold, *Z. Kristallogr.* **152** (1980) 119–133.
32. O. Muller, R. Roy: *The Major Ternary Structural Families*, Springer-Verlag, Berlin 1974.
33. C. Chateau, J. Haines, J.-M. Léger, A. LeSauze, R. Marchand, *Amer. Mineral.* **84** (1999) 207–210.
34. D. C. Palmer: "Stuffed Derivatives of the Silica Polymorphs," *Rev. Miner.* **29** (1994) 83–118.
35. R. W. G. Wyckoff, *Am. J. Sci.* **209** (1925) 448–459.
36. D. J. Peacor, *Z. Kristallogr.* **138** (1973) 274–298.
37. A. F. Wright, A. J. Leadbetter, *Philos. Mag.* **31** (1975) 1391–1401.
38. R. L. Withers et al., *Phase Transitions* **16/17** (1989) 41–45.
39. W. W. Schmahl, I. P. Swainson, M. T. Dove, A. Graeme-Barber, *Z. Kristallogr.* **201** (1992) 125–145.
40. D. M. Hatch, S. Ghose, *Phys. Chem. Minerals* **17** (1991) 554–562.
41. O. W. Flörke, H. Schneider, *Ber. Dtsch. Keram. Ges.* **63** (1986) 368–372.
42. H. Schneider, O. W. Flörke, R. Stoeck, *Z. Kristallogr.* **209** (1994) 113–117.
43. O. W. Flörke, K. Langer, *Contrib. Mineral. Petrol.* **36** (1992) 221–230.
44. O. W. Flörke, *Glastechn. Ber.* **32** (1959) 1–9.

45. H. Graetsch, O. W. Flörke, *Z. Kristallogr.* **195** (1991) 31–48.
46. K. Kihara, *Z. Kristallogr.* **152** (1980) 95–101.
47. A. K. A. Pryde, M. T. Dove, *Phys. Chem. Minerals* **26** (1998) 171–179.
48. W. A. Dollase, *Acta Crystallogr.* **23** (1967) 617–623.
49. A. Nukui, H. Nakazawa, M. Akao, *Amer. Mineral.* **63** (1978) 1252–1259.
50. H. A. Graetsch, M. Brunelli, Z. Kristallogr. **220** (2005) 606–613.
51. K. Kihara, *Z. Kristallogr.* **146** (1977) 185–203.
52. H. Graetsch, *Phys. Chem. Minerals* **28** (2001) 313–321.
53. K. Kato, A. Nukui, *Acta Crystallogr.* **B32** (1976) 2486–2491.
54. W. H. Baur, *Acta Crystallogr.* **B33** (1977) 2615–2619.
55. J. H. Konnert, D. E. Appleman, *Acta Crystallogr.* **B34** (1978) 391–403.
56. W. Hoffmann, M. Kockmeyer, J. Löns, C. Vach, *Fortschr. Mineral.* **61** (1983) 96–98.
57. J. Löns, W. Hoffman, *Z. Kristallogr.* **178** (1987) 141–143.
58. H. Graetsch, I. Topalović-Dierdorf, *Eur. J. Mineral.* **8** (1996) 103–113.
59. H. A. Graetsch, Z. Kristallogr. **218** (2003) 531–535.
60. H. Graetsch, *Amer. Mineral.* **83** (1998) 872–880.
61. A. Nukui, O. W. Flörke, *Amer. Mineral.* **72** (1987) 167–176.
62. W. Hoffmann, *Naturwiss.* **54** (1967) 114.
63. W. A. Dollase, W. H. Baur, *Amer. Mineral.* **61** (1976) 971–978.
64. C. B. Sclar, A. P. Young, L. C. Carrison, C. M. Schwartz, *J. Geophys. Res.* **67** (1962) 4049–4054.
65. M. A. Spackmann, R. J. Hill, G. V. Gibbs, *Phys. Chem. Miner.* **14** (1987) 139–150.
66. R. J. Hemley, C. T. Prewitt, K. J. Kingma: "High-Pressure Behavior of Silica," *Rev. Miner.* **29** (1994) 41–74.
67. G. Miehe, H. Graetsch, *Eur. J. Mineral.* **4** (1992) 693–706.
68. O. W. Flörke, J. B. Jones, H. U. Schmincke, *Z. Kristallogr.* **143** (1976) 156–165.
69. O. W. Flörke, U. Flörke, U. Giese, *Neues Jahrb. Mineral. Abh.* **149** (1984) 325–336.
70. P. J. Heaney, J. E. Post, *Science* **255** (1992) 441–443.
71. M. Grasselini Troysi, P. Orlandi, *Atti. Soc. Tosc. Sci. Nat. Mem.* **A79** (1972) 245–250.
72. H. Gies, H. Gerke, F. Liebau, *Neues Jahrb. Mineral. Monatsh.* (1982) 119–124.
73. H. Gies, *Z. Kristallogr.* **164** (1983) 247–257.
74. G. Miehe, H. Graetsch, O. W. Flörke, *Phys. Chem. Minerals* **10** (1984) 197–199.
75. H. Graetsch, O. W. Flörke, G. Miehe, *Phys. Chem. Minerals* **12** (1985) 300–305.
76. J. B. Jones, E. R. Segnit, *J. Geol. Soc. Australia* **18** (1971) 57–68.
77. O. W. Flörke, *Fortschr. Mineral.* **44** (1967) 181–230.
78. O. W. Flörke, *Neues Jahrb. Mineral. Monatsh.* (1955) 217–223.
79. H. Graetsch: "Structural Characteristics of Opaline and Microcrystalline Silica Minerals," *Rev. Miner.* **29** (1994) 209–232.
80. K. Langer, O. W. Flörke, *Fortschr. Mineral.* **52** (1974) 17–51.
81. H. Graetsch, H. Gies, I. Topalović, *Phys. Chem. Minerals* **21** (1994) 166–175.
82. J. Shropshire, P. P. Keat, P. A. Vaughan, *Z. Kristallogr.* **112** (1959) 400–413.
83. F. Liebau, *Amer. Mineral.* **63** (1978) 918–923.
84. B. Martin, K. Röller, *Neues Jahrb. Mineral. Monatsh.* (1990) 462–466.
85. M. G. Fereira da Silva, J. M. F. Navarro, *J. Non-Cryst. Solids* **109** (1989) 191–197.
86. Al. Weiss, Ar. Weiss, *Z. Anorg. Allg. Chem.* **276** (1954) 95–112.
87. O. W. Flörke, J. B. Jones, U. Köster, E. Robarick, *Acta Crystallogr.* **A 46** (1990) Suppl. C-247.
88. R. L. Mozzi, B. E. Warren, *J. Appl. Cryst.* **2** (1969) 164–173.
89. H. Graetsch, A. Mosset, H. Gies, *J. Non-Cryst. Solids* **119** (1990) 173–180.
90. R. J. Hemley, H. K. Mao, P. M. Bell, B. O. Mysen, *Phys. Rev. Lett.* **57** (1986) 747–750.
91. H. Graetsch, K. Ibel, *Phys. Chem. Minerals* **24** (1997) 102–108.
92. J. V. Sanders, *Acta Crystallogr.* **A24** (1968) 427–434.
93. D. C. Hurd in S. R. Aston: *Silicon Geochemistry and Biogeochemistry*, Academic Press, New York 1983, pp. 179–186.
94. O. W. Flörke, J. B. Jones, E. R. Segnit, *Neues Jahrb. Mineral. Monatsh.* (1973) 82–89.
95. H. Graetsch, I. Topalović-Dierdorf, *Chem. Erde* **56** (1996) 387–391.
96. G. Lehmann, H. U. Bambauer, *Angew. Chemie* **85** (1973) 281–289.
97. V. S. Balitsky, *J. Crystal Growth* **41** (1977) 100–102.
98. O. W. Flörke, K. Röller, F. Siebers, *Gemmologie* **49** (2000) 85–104

99. G. R. Rossmann: "Colored Varieties of the Silica Minerals," *Rev. Miner.* **29** (1994) 433–467.
100. J. S. Goreva, Chi Ma, G. R. Rossman, *Am. Mineral.* **86** (2001) 466–472.
101. R. Pankrath, O. W. Flörke, *Eur. J. Mineral.* **6** (1994) 435–457.
102. O. W. Flörke, *Mitt. Österr. Miner. Ges.* **140** (1995) 9–34.
103. A. C. McLaren, D. R. Pitkethly, *Phys. Chem. Minerals* **8** (1982) 128–135.
104. F. J. Pettijohn: *Sedimentary Rocks*, 3rd ed., Harper & Row, New York 1975, pp. 195 ff, 300 ff, 392 ff.
105. H. Williams, F. J. Turner, C. M. Gilbert: *Petrography*, 2nd ed., Freeman, San Francisco 1982.
106. P. Harben, *Ind. Miner. (London)* **184** (1983) 28–29, 32.
107. O. W. Flörke, B. Köhler-Herbertz, K. Langer, I. Tönges, *Contrib. Mineral. Petrol.* **80** (1982) 324–333.
108. A. Peschel: *Natursteine*, 2nd ed., Dtsch. Verlag Grundstoffind., Leipzig 1983, pp. 77–80, 243–251, 274–275.
109. S. J. Lefond (ed.): *Industrial Minerals and Rocks, (Nonmetallics other than Fuels)*, 4th ed., Amer. Inst. Mining Metallurg. Petrol. Eng. (AIME), New York 1975.
110. M. Kuzvart: *Industrial Minerals and Rocks*, Elsevier, Amsterdam 1984.
111. M. Smith, *Ind. Miner. (London)* **203** (1984) 19–25.
112. K.-T. Wilke: *Kristallzüchtung*, H. Deutsch, Thun 1988, pp. 1027–1058.
113. E. D. Kolb, K. Nassau, R. A. Laudise, E. E. Simpson, K. M. Kroupa, *J. Crystal Growth* **36** (1976) 93–100.
114. C. G. Kennedy, W. L. Knight, W. T. Holser, *Amer. J. Sci.* **256** (1958) 590–595.
115. W. T. Holser, C. G. Kennedy, *Amer. J. Sci.* **257** (1959) 71–77.
116. G. C. Kennedy, *Econ. Geol.* **45** (1950) 629–653.
117. R. A. Laudise, E. D. Kolb, *Endeavour* **28** (1969) 114–117.
118. N. B. Hannay (ed.): "Change of State," *Treatise on Solid State Chemistry*, vol. **5**, Plenum Press, New York 1975.
119. 113a G. Johnson, J. Foise, *Encyclopedia of Applied Physics* **15** (1996) 365–375.
120. V. A. Kuznetsov, A. N. Lobachev, *Sov. Phys. Crystallogr.* **17** (1973) 775–803.
121. A. Armington, J. F. Balascio, *38th Ann. Frequency Control Symp.* (1984) 3–8.
122. J. A. Weil, *Phys. Chem. Minerals* **10** (1984) 149–165.
123. M. S. Paterson, *Bull. Mineral.* **105** (1982) 20–29.
124. P. Cordier, J. C. Doukhan, *Amer. Mineral.* **76** (1991) 361–366.
125. J. Flicstein, M. Schieber, *J. Cryst. Growth* **24/25** (1974) 603–609.
126. A. C. McLaren, C. F. Osborne, L. A. Saunders, *Phys. Stat. Solidi* **4** (1971) 235–247.
127. A. C. McLaren, R. F. Cook, S. T. Hyde, R. C. Tobin, *Phys. Chem. Minerals* **9** (1983) 79–94.
128. P. P. Phakey, *Phys. Stat. Solidi* **34** (1969) 105–119.
129. A. C. McLaren, P. P. Phakey, *Phys. Stat. Solidi* **31** (1969) 723–737.
130. D. M. Dodd, D. B. Frazer, *J. Phys. Chem. Solids* **26** (1965) 673–686.
131. W. G. Cady: *Piezoelectricity*, Dover Publ., New York 1964.
132. P. Profos (ed.): *Handbuch der industriellen Meßtechnik*, 4th ed., Vulkan-Verlag, Essen 1987.
133. J. C. Brice, *Rev. Modern Phys.* **57** (1985) 105–146.
134. S. Fujishima, *IEEE Trans. Ultrasonics Frequency Control* **47** (2000) 1–7.
135. prEN 12475, part 11: Classification of dense shaped refractory products. Alumina–silica products, Oct. 1996.
136. DIN 1089, part 1: Refractories for use in coke ovens. Silica bricks, requirements and testing, Feb. 1995.
137. ASTM C 416–97: Standard classification of silica refractory brick.
138. O. W. Flörke, *Ber. Dtsch. Keram. Ges.* **34** (1957) 343–353.
139. O. W. Flörke, *Refractories J.* **11** (1958) 493–516.
140. F. Brunk, D. Schmidt, IREFCON' 96 Proc. Vol. I, Indian Refractory Association, Calcutta 1996, 226–230.
141. H. Schneider, K. Wohlleben, A. Majdic, *Mineral. Mag.* **43** (1980) 879–883.
142. H. Schneider, A. Majdic, *Neues Jahrb. Mineral. Monatsh.* (1984) 559–883.
143. V. Lach, *Sprechsaal* **123** (1990) 811–813.
144. S. J. Stevens, R. J. Hand, J. H. Sharp, *J. Mater. Sci.* **32** (1997) 2929–2935.
145. A. Majdic, H. Schneider, K. Wohlleben, *cfi/DKG* **63** (1986) 176–189.
146. I. Patzak, *Tonind. Ztg.* **96** (1972) 291–297.
147. H. Schneider, O. W. Flörke, A. Majdic, *Proc. Brit. Ceram. Soc.* **28** (1979) 267–279.

148. D. Taylor, *Trans. Brit. Ceram. Soc.* **83** (1984) 129–134.
149. F. Brunk in G. Routschka (ed.): *Refractory Materials*, Vulkan-Verlag, Essen 1997, p. 43.
150. Y. Naruse, Y. Hoshino, T. Tanaka, *Taikabutsu Overseas* **2** (1982) 110–120.
151. H. J. Koschlig: *UNITECR Proc.,* vol. 1, Amer. Ceram. Soc., Columbus, Ohio 1989, pp. 244–283.
152. F. Brunk, W. Weßling: *XXXVI. Int. Col. on Refractories*, Verlag Stahleisen, Düsseldorf 1993, pp. 20–26.
153. J. D. Panda, G. Goswami, *Iron & Steel Review* **39** (1996) 26–28.
154. A. Kumar in S. N. Gosh, S. L. Sarker, S. Harsch (eds.): *Progress in Cement and Concrete*, vol. **4**, Abi Books, New Dehli 1993, pp. 342–367.
155. G. H. Beall: "Industrial Applications of Silica," *Rev. Miner.* **29** (1994) 469–506.
156. K. H. Khayat, P.-C. Aitcin in S. N. Gosh, S. L. Sarker, S. Harsch (eds.): *Progress in Cement and Concrete*, vol. **4**, Abi Books, New Dehli 1993, pp. 226–265.
157. F. Press, R. Siever: *Earth*, 4th ed., Freeman, San Francisco 1986.
158. B. J. Skinner, S. C. Parker: *Physical Geology*, Wiley-Interscience, New York 1987.
159. L. Jung: *High Purity Natural Quartz,* part I: "High Purity Natural Quartz for Industrial Uses"; Part II: "High Purity Natural Quartz Markets for Suppliers and Users," Quartz Technology Inc., Liberty Corner, New Jersey 1992.
160. R. Weiss: *Formgrundstoffe*, Giesserei-Verlag, Düsseldorf 1984.
161. J. Griffith, *Ind. Miner. (London)* **235** (1987) 25–43.
162. H. Füchtbauer (ed.): *Sedimente und Sedimentgesteine*, 4th ed., Schweizerbarth, Stuttgart 1988.
163. H. J. Blankenburg, J. Götze, H. Schulz: *Quartzrohstoffe*, 2nd ed., Dtsch. Verl. Grundstoffind., Leipzig 1994.
164. W. Koensler: *Sand und Kies*, F. Enke, Stuttgart 1989.
165. W. Lorenz, W. Gwosdz: "Bewertungskriterien Industrieminerale, Steine, Erden", *Quartzrohstoffe,* Geol. Jahrb., part 3, Series H, H. 6, Bundesanst. Geowiss. u. Rohstoffe, Hannover 1999.
166. O. Stier, H. Ortleb: *Gefahrstoffe – Reinhaltung Luft* **59** (1999) 27–31.
167. M. Saller, *Ind. Minerals (London)* **202** (1999) 6, 25–33.
168. Hoffmann Mineral (ed.): *Neuburger Kieselerde*, Neuburg/Donau 1999.
169. O. W. Flörke, B. Köhler-Herbertz, K. Langer, I. Tönges, *Contrib. Mineral. Petrol.* **80** (1982) 324–333.
170. C. Frondel: "Silica Minerals," *The System of Mineralogy*, 7th ed., vol. **3**, J. Wiley & Sons, New York 1962.
171. H. Graetsch, O. W. Flörke, G. Miehe, *Phys. Chem. Minerals* **14** (1987) 249–257.
172. C. Barker, W. D. Underwood, *Analyst* **117** (1992) 1407–1410.
173. R. J. Bodner, *Pure Appl. Chem.* **67** (1995) 873–880.
174. J. Mullis in M. Frey (ed.): *Low Temperature Metamorphism*, Blackie, Glasgow 1987, pp. 162–199.
175. J. T. Nash, T. Theodore, *Econ. Geol.* **66** (1971) 385–399.
176. B. Perney, P. Eberhard, K. Ramseyer, J. Mullis, R. Pankrath, *Amer. Mineral.* **77** (1992) 534–544.
177. Sawyer Research Products, US 3 837 826, 1974 (B. Sawyer).
178. *Ullmann*, 4th ed., **21**, 444–450.
179. R. Weiss, *Erzmetall* **27** (1974) 169–177; **31** (1978) 450–457.
180. D. H. Solomon, D. G. Hawthorne: *Chemistry of Pigments and Fillers*, John Wiley & Sons, New York 1983.
181. *Oberflächenbehandelte Füllstoffe,* Data Sheets, Quartzwerke, Frechen, March 8, 1999.
182. R. Oberste-Pattberg, ARDEX GmbH, Witten, *personal communication.*
183. S. N. Groudev, V. J. Groudeva, *Ind. Minerals (London)* **222** (1986) 81–84.
184. P. Harpen, *Ind. Minerals (London)* **184** (1983) 28–29, 32.
185. H. Scholze (ed.): "Keramische Werkstoffe," *Keramik*, 6th ed., **part 2**, Springer-Verlag, Berlin 1983.
186. B. M. Coope, *Ind. Minerals (London)* **298** (1989) 43–55.
187. A. Peschel: *Natursteine*, 2nd ed., Dtsch. Verlag Grundstoffindustrie, Leipzig 1983.
188. H. Gundlach: *Dampfgehärtete Baustoffe*, Bauverlag, Wiesbaden 1973.
189. W. Eden: *Handbuch zur Herstellung von Kalksandstein*, 2nd ed., Bundesverband Kalksandsteinindustrie, Hannover 1998.
190. W. Eden, E. Eberhard: *Forschg. Ber. Nr. 84,* Forschungsvereinigung Kalk-Sand, Hannover 1996.
191. D. Schlumberger: *Cementing Technology*, Nova Communications, London 1984, p. 218.

192. D. Skudelny, R. Weiss: *GAK – Gummi Asbest Kunststoffe* **36** (1983) 3–11.
193. U. Fuhrmann, D. Skudelny in M. Beyer et al. (eds.): *Epoxidharze i. d. Elektrotechnik*, Expert-Verlag, Grafenau 1983.
194. J. Griffith, *Ind. Minerals (London)* **241** (1987) 23–45.
195. F. A. Hofmann, H. D. Merzbach: *Silanbehandelte Füllstoffe und Verwendung in der Farbenindustrie*, Quartzwerke Frechen, 1990.
196. W. Grass, D. H. Merzbach, D. Skudelny, *Farbe + Lack* **92** (1986) 1030–1033.
197. D. Skudelny, *Kunststoffe* **77** (1987) 1153–1156.
198. F. A. Hofmann, D. Skudelny: *Surface treated mineral fillers – growing markets for speciality products*, Quartzwerke Frechen, 1990.
199. H. Insley, V. D. Fréchette: *Microscopy of Ceramics and Cements*, Academic Press, New York 1955.
200. R. Weiss, *Glastech. Ber.* **49** (1976) 12–15.
201. J. Lange: *Rohstoffe der Glasindustrie*, Dtsch. Verlag Grundstoffind., Leipzig 1993.
202. A. G. Pincus, D. H. Davies (eds.): "Major Ingredients", *Raw Materials in the Glass Industry*, part 1, Ashlee Publ., New York 1983.
203. H. Petzold, H. Pöhlmann: *Email und Emailtechnik*, Springer-Verlag, Berlin 1987.
204. *Ind. Minerals (London)* **197** (1984) 19–37, 39–45.
205. E. Roedder in H. L. Barnes (ed.): *Geochemistry of Hydrothermal Ore Deposits*, 2nd ed., John Wiley & Sons, New York 1979.
206. C. Müller (ed.): *Databook für Keramik-Glas-Baustoff-Produktion*, 2nd ed., Sprechsaal-Verlag, Coburg 1975.
207. R. Weiss, *Ber. Dtsch. Keram. Ges.* **55** (1978) 82–91.
208. UNIMIN Corp., New Canaan, Connecticut, USA, *data sheet,* 1998.
209. N. SeveringhausJr. in S. J. Lefond (ed.): *Ind. Miner. Rocks* 1975, 235–249.
210. Roskill Inform. Serv. Ltd.: *The Economics of Diatomite*, 5th ed., London 1988, p. 73.
211. L. Pettifer, *Ind. Miner. (London)* **175** (1982) 47–69.
212. L. Benda, H. Brandes, *Geol. Jahrb. Reihe A* **21** (1974) 3–85.
213. L. Benda, *Schriftenr. GDMB* **38** (1981) 134–140.
214. W. B. Clark, *Calif. Geol.* **31** (1978) 3–9.
215. L. Benda, B. Mattiat, *Geol. Jahrb. Reihe D* **22** (1977) 3–107.
216. F. L. KadeyJr. in S. J. Lefond (ed.): *Ind. Miner. Rocks* 1975, 605–635.
217. *Ind. Miner. (London)* 1987 May, 22–39.
218. *Ullmann*, 4th ed., **21**, 452–456.
219. Mitteleuropäische Brautechnische Analysenkommission (MEBAK): Methodensammlung, Freising 1982, III, pp. 651–659.
220. Supplier: VEL Geldenaaksebaan 464, B-3030 Leuven (Heverlee), Belgium.
221. S. Ward: "Liquid Filtration Theory," in M. J. Matteson, C. Orr (eds.): *Filtration, Principles and Practices,* New York, Basel 1987.
222. J. Kiefer: "Kieselgurfiltration," *Brauwelt* **40** (1990) 1730–1749.
223. E. Krüger, N. Wagner, B. Lindemann: "Einflußgrößen bei der Bierfiltration," *Brauwelt* **50** (1989) 2434–2444.
224. International Agency for Research on Cancer (IARC): "Silica and Some Silicates," *Carcinogenic Risk of Chemicals to Humans*, **no. 42**, Lyon 1987.
225. International Diatomite Producers Association (IDPA): *Silica Facts Questions and Answers; Health and Safety Aspects of Diatomaceous Earth; The Nature and Safe Handling of Diatomaceous Earth*, San Francisco 1989.
226. *Br. J. Ind. Med.* **46** (1989) 289–291.
227. O. Walton: "Filter Media," *The Brewer* **64** (1978) 261–265.
228. F. Kainer, *Kieselgur*, 2nd ed., Enke, Stuttgart 1951, pp. 39–43.
229. L. Benda, S. Paschen: "Kieselguhr (Diatomit) – ein vielseitig genutzter Rohstoff," *Geol. Jahrb.* **A142** (1993) 383–398.
230. R. O. Y. Breese: "Diatomite," in D. D. Carr (ed.): *Industrial Minerals and Rocks*, 6th ed., AIME, Littleton, CO 1994, pp. 397–442.
231. W. Lorenz, W. Gwosdz: "Bewertungskriterien für Industriemineralien, Steine und Erden, Teil 3: Quartzrohstoffe," *Geol. Jahrb.* (1999) H6, 3–119.
232. D. H. Everett: "Definition Terminology and Symbols in Colloid and Surface Chemistry," *Symbols and Terminology for Physiocochemical Quantities and Units*, **Appendix II, part 1**, International Union of Pure and Applied Chemistry, Butterworths, London 1971.
233. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 174.
234. T. Graham, *Justus Liebigs Ann. Chem.* **121** (1862) 36.
235. R. Griessbach, *Chem. Ztg.* **57** (1933) 253.
236. P. G. Bird, US 2 244 325, 1945.

237. M. F. Bechtold, O. E. Snyder, US 2 574 902, 1951.
238. "Silicates and Silicas Report," *Chemical Economics Handbook*, SRI International, Menlo Park 1990.
239. *Opportunities in Specialty Silicas*, Kline & Co., Fairfield, N. J., 1992.
240. E. S. Gould: *Inorganic Reactions & Structures*, Henry Holt & Company, New York 1957.
241. G. J. Fine, *J. Chem. Ed.* **68** (1991) 765.
242. W. D. Kingery, H. K. Bowen, D. R. Uhlmann: *Introduction to Ceramics*, 2nd ed., J. Wiley & Sons, New York 1976.
243. V. Hofman, K. Endell, D. Wilm, *Angew. Chem.* **30** (1934) 539–558.
244. V. Khavryutchenko et al.: "Abstracts of Papers," in R. K. Iler (ed.): *Memorial Symposium on the Colloid Chemistry of Silica,* Washington, D.C., 1990, Paper no. 100.
245. A. P. Legrand et al., *Adv. Colloid Interface Sci.* **33** (1990) 91–330.
246. R. S. McDonald, *J. Phys. Chem.* **62** (1958) 1168.
247. M. L. Hair: *Infrared Spectroscopy in Surface Chemistry*, Marcel Dekker, New York 1967, p. 82.
248. J. B. Peri, A. L. HensleyJr., *J. Phys. Chem.* **72** (1968) no. 8, 2926–2933.
249. G. E. Maciel, D. W. Sindorf, *J. Am. Chem. Soc.* **102** (1980) 7606–7607.
250. K. D. Unger: "Porous Silica," *J. Chromatogr. Libr.* **16** (1979) .
251. M. L. Hair in A. T. Bell, M. L. Hair (eds.): "Vibrational Spectroscopies for Adsorbed Species," *ACS Symp. Ser.* **137** (1980) .
252. L. T. Zhuravlev: "Abstracts of Papers," in R. K. Iler (ed.) *Memorial Symposium on the Colloid Chemistry of Silica,* Washington, D.C., 1990, Paper no. 93.
253. R. K. Iller: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 410.
254. Du Pont Company, Nalco Inc., and Nissan Chemicals, based on product information sheets (not specifications).
255. Cabot Corp., and Nissan Inc. and Nissan Chemicals, based on product information sheets (not specifications).
256. Nalco Inc., Nissan Chemicals, and Nippon Seahostar, based on product information sheets (not specifications).
257. R. K. Iller: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 325.
258. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979.
259. B. V. Derjaguin: *Theory of Stability of Colloids and Thin Films*, Consultants Bureau, New York 1989.
260. G. W. Sears, R. K. Iler, private communication.
261. D. H. Napper: *Polymer Stabilization of Colloidal Dispersions*, Academic Press, London 1983.
262. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 364.
263. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 396.
264. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 367.
265. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 366.
266. R. K. Iler, *J. Phys. Chem.* **56** (1952) 680.
267. E. Matijevic et al., *J. Phys. Chem.* **73** (1969) 564.
268. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 409.
269. G. B. Alexander, A. H. Bolt, US 3 007 878, 1961.
270. J. T. Overbeek, private communiction, 1955, unpublished.
271. E. Matijevic, *J. Colloid Sci.* **19** (1964) 333.
272. W. Stober, A. Fink, *J. Colloid Interface Sci.* **26** (1968) 62.
273. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 331.
274. A. B. Alexander, A. H. Bolt, US 3 007 878.
275. R. C. Atkins, US 3 012 973, 1961.
276. R. K. Iler, F. J. Wolter, US 2 631 136, 1953.
277. M. Mindick, L. E. Reven, US 3 468 813, 1969.
278. L. A. Dirnberger, US 2 703 314, 1955.
279. J. Kloepfner, DT 762 723, 830 786, 1942.
280. C. J. Brinker, G. W. Scherer: *Sol-Gel Science*, Academic Press, San Diego 1990.
281. C. G. Tan et al., *J. Colloid Interface Sci.* **118** (1987) 290–293.
282. J. Wolf, private communication.
283. R. K. Iler, US 3 668 088, 1972.
284. H. E. Bergna, US 4 410 405, 1983.
285. G. B. Alexander, US 2 750 345, 1956.
286. H. E. Bergna, F. A. Simko Jr., US 3 301 635, 1967.
287. H. H. Mandelbrot in D. Avnir (ed.): *A Fractional Approach to Heterogeneous Chemistry*, J. Wiley & Sons, New York 1989.
288. S. Brumauer, P. H. Emmett, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309.
289. G. W. Sears, *Anal. Chem.* **28** (1956) 1981.
290. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979, p. 354.
291. T. Allen, R. Davies in R. K. Iler (ed.): "Abstracts of Papers," *Memorial Symposium on the Colloid Chemistry of Silica*, , Washington, D.C. 1990, Paper no. 37.

292. J. Kirkland in H. E. Bergna (ed.): "Colloid Chemistry of Silica," *Adv. Chem. Ser.* **234** 1994.
293. P. F. Collins, US 2 380 945, 1945.
294. P. G. Batelson, CA 1 154 564, 1983.
295. K. Anderson et al., *Nord. Pulp Pap. Res. J.* (1986).
296. K. A. Johnson, US 4 643 801, 1987.
297. S. C. Sofia et al., US 4 795 531, 1989.
298. J. D. Rushmere, US 4 798 653, 1989.
299. R. K. Iler, US 2 597 871, 1947.
300. R. K. Iler, US 2 726 961, 1955.
301. I. V. Wilson, US 2 643 048, 1953.
302. R. J. Walsh, A. H. Herzog, US 3 170 273, 1965.
303. G. W. Sears, US 3 922 393, 1975.
304. J. L. Callahan, US 2 974 110, 1961.
305. J. L. Callahan et al., US 3 322 847, 1967.
306. T. Graham, *Philos. Trans. R. Soc. London* **252** (1861) 204.
307. J. G. Vail: "Soluble Silicates," *ACS Monograph Series*, vol. **2**, Reinhold, New York 1952, p. 549.
308. *Ullmann*, 4th ed., **21**, pp. 458–462.
309. W. A. Patrick, US 1 297 724, 1918.
310. R. K. Iler: *The Chemistry of Silica*, J. Wiley & Sons, New York 1979.
311. C. J. Brinker, G. W. Scherer: *Solg-Gel Science*, Academic Press, San Diego 1990.
312. Wijnen et al., *J. Colloid Interface Sci.* **145** (1991) 17.
313. M. E. Winyall in B. E. Leach (ed.): *Applied Industrial Catalysis*, vol. **3.** Academic Press, Orlando 1984, p. 43.
314. S. S. Kistler, *J. Phys. Chem.* **36** (1932) 52.
315. P. H. Tervari, A. J. Hunt, US 4 610 863, 1986.
316. BASF, US 4 667 417, 1987 (F. Graser, A. Stange).
317. L. C. Klein (ed.): *Sol-Gel Technology for Thin Films, Fibers, Preforms, Electronics, and Specialty Shapes*, Nayes, Park Ridge, New Jersey, 1988.
318. M. B. Kenny, K. S. W. Sing in H. Bergna (ed.): "The Colloid Chemistry of Silica," *Adv. Chem. Ser.* **234** (1994) .
319. S. Brunauer, P. H. Emmett, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309.
320. R. W. Cranston, F. A. Inkley, *Adv. Catal.* **9** (1957) 143.
321. K. Sing in G. Parfitt, K. Sing (eds.): *Characterization of Powder Surfaces*, Academic Press, New York 1976, p. 34.
322. R. L. White, A. Nair, *Appl. Spectrosc.* **44** (1990) 69.
323. P. Wijnen, T. Beelen, R. van Santen in H. Bergna (ed.): "The Colloid Chemistry of Silica," *Adv. Chem. Ser.* **234**, 1994.
324. V. K. Markov, N. A. Nagornaya, *J. Applied Chem. (USSR)* **10** (1937) 853.
325. E. Heimann, US 2 316 241, 1938.
326. Davison Chemical, US 2 625 492, 1949 (L. O. Young).
327. W. R. Grace & Co., US 2 856 268, 1954 (L. O. Young).
328. Houdry, US 2 232 727 (A. G. Peterkin, H. A. Shabaker).
329. Mizusawa Ind. Chem., DE 1 442 778, 1963 (T. Kuwata, Y. Sugahara, T. Nakazawa).
330. Asahi Chem. Ind., JP 48 013 834, 1966 (K. Takizawa, Y. Ohba).
331. BASF, DE 2 103 243, 1979 (G. Merz, H. Gehrig, W. Chorbacher).
332. Socony-Vacuum Oil Comp., US 2 385 217, 1945 (M. M. Marisic).
333. Kali-Chemie, DE 1 567 617, 1970 (J. Dultz).
334. Mizusawa, EP 884 277, 1998 (T. Masanori, W. Yuuji).
335. J. F. White, I. V. Wilson, *Ind. Eng. Chem.* **33** (1941) 1169.
336. Fuji-Davison, Art-Sorb-Technical Information (in English), 4.11.27 Meieki, Nagoya-shi, Japan 450.
337. R. S. McDonald, *J. Am. Chem. Soc.* **79** (1975) 850.
338. Crosfield, US 3 617 301, 1968 (D. Barby, J. P. Quinn).
339. W. R. Grace & Co., US 4 629 588, 1984 (W. A. Welsh, Y. O. Parent).
340. A. Bertod, *J. Chromtogr.* **549** (1991) 1.
341. J. Freng, GB 203 248, 1922.
342. Lever Bros., US 3 538 230, 1966 (M. Pader, W. Wiesner).
343. W. R. Grace & Co., US 4 303 641, 1978 (R. B. DeWolf, R. Glemza).
344. J. M. Huber, US 4 340 583, 1977 (S. K. Wason).
345. M. Pader: *Oral Hygiene Products and Practice*, Marcel Dekker, New York 1988.
346. J. S. Magee, J. J. Blazek in J. A. Rabo (ed.): "Zeolite Chemistry and Catalysis," *ACS Monogr*. **no. 171**, 1976, 615.
347. Phillips Petroleum, US 3 378 540, 1964 (D. R. Witt).
348. F. Karol et al., *J. Polym. Sci., Polym. Chem. Ed.* **10** (1972) 2621.
349. Exxon Chemical, US 4 701 432, 1986 (H. C. Welborn).
350. D. D. Dunnom (ed.): *Health Effects of Synthetic Silica Particulates*, **no. 732**, ASTM Special Technical Publication, Philadelphia 1981.

351. Y. Takizawa et al., *Acta Med. Biol. (Niigata)* **36** (1988) 27.
352. T. Provder: *Particle Size Distribution: Assessment and Characterization,* ACS Symposium Series 332, American Chemical Society, 1987.
353. W. R. Grace & Co., US 3 800 031, 1974 (H. S. Sale, A. Delgrado, C. F. Doyle).
354. Unilever Patent Holdings B. V., US 5 215 733, 1993 (J. K. Potter).
355. University of New Mexico, US 5 565 142, 1996 (R. Deshpande, D. M. Smith, C. J. Brinker).
356. D. M. Chapman, *Proceedings of the 1997 TAPPI Coating Conferenz* (1997) 73–93.
357. Texas Instruments Inc., US 5 847 443, 1998 (C. C. Cho et al.).
358. "National Chemical Inventories" (TM) CD-ROM, Copyright 1999, *American Chemical Society.*
359. International Association for Research on Cancer (IARC): *Silica, Some Silicates, Coat Dust and Para-aramid Fibrils*, vol. **68**, 1997, pp.
360. C. J. Johnston et al., *Toxicol. Sci.* **56** (2000) 405–413.
361. Degussa, DE 762 723, 1942 (H. Klöpfer).
362. S. Brunauer, P. H. Emmett, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309.
363. H. Ferch, K. Seibold, *Farbe Lack* **90** (1984) 88.
364. Degussa, *Schriftenreihe Pigmente*, Paper no. 11,
Frankfurt 1992.
365. M. Schmücker, "Untersuchungen an nichtkristallinen technischen Kieselsäuren unter besonderer Berücksichtigung des thermischen Transformationsverhaltens," Diplomarbeit, Ruhr-Universität, Bochum 1988.
366. H. Wistuba, "Die Silanolgruppen der Siliciumdioxidoberfläche und ihre chemischen Reaktionen," Dissertation, Universität Heidelberg 1967.
367. J. Mathias, G. Wannemacher, *J. Colloid Interface Sci.* **125** (1988) 61.
368. D. Kerner, W. Leiner, *Colloid Polym. Sci.* **253** (1975) 960.
369. H. Ferch, *Seifen Öle Fette Wachse* **101** (1975) 17, 51.
370. B. A. Morrow, A. J. McFarlan, *J. Non Cryst. Solids* **120** (1990) 61.
371. B. A. Morrow, A. J. McFarlan, *Langmuir* **7** (1991) 1695.
372. K. Yoshinaga et al., *Chem. Lett.* 1991, 1129.
373. C. P. Tripp, M. L. Hair, *Langmuir* **7** (1991) 923.
374. G. Hellwig, *Farbe Lack* **81** (1975) no. 8, 705.
375. A. Meffert, A. Langenfeld, *Fresenius Z. Anal. Chem.* **249** (1979) 231.
376. H.-P. Boehm, M. Schneider, *Z. Anorg. Allg. Chem.* **301** (1959) 326.
377. H.-P. Boehm, *Kolloid Z. Z. Polym.* **227** (1968) 17.
378. H.-P. Boehm, *Angew. Chem.* **78** (1966) no. 12, 617.
379. Degussa, DE 1 163 784, 1962 (H. Brünner, J. Diether, D. Schutte).
380. Degussa, DE 3 211 431, 1982 (H. Klebe et al.).
381. Ivoclar, DE 3 632 215, 1986 (V. M. Rheinberger, T. Büchl).
382. Blendax, WO 8 503 220, 1985 (W. Kuhlmann).
383. Kulzer, DE 2 405 578, 1974 (A. Groß, R. Schaefer).
384. Canon, EP 369 443, 1989 (Y. Sato, T. Kukimoto).
385. Nippon Aerosil Corp., JP 58 185 405, 1983 (N. Furuya, T. Morii).
386. Canon, EP 216 295, 1986 (K. Tanaka et al.).
387. V. M. Litvinov et al., *Acta Polym.* **39** (1988) no. 5, 244.
388. G. Berrod, A. Vidal, E. Papirer, J. B. Donnet, *J. Appl. Polym. Sci.* **23** (1979) 2679; **26** (1981) 833, 1015.
389. P. Vondracek, M. Schätz, *J. Appl. Polym. Sci.* **21** (1977) 3211.
390. L. A. Peters, J. C. Vicic, D. E. Cain, *GAK Gummi Asbest Kunstst.* **44** (1991) no. 2, 69.
391. H. Ferch, A. Reisert, R. Bode, *Kautsch. Gummi Kunstst.* **39** (1986) 1084.
392. Degussa, *Schriftenreihe Pigmente*, Paper no. 23, Frankfurt 1986.
393. Degussa, *Schriftenreihe Pigmente*, Paper no. 68, Frankfurt 1985.
394. Degussa, *Schriftenreihe Pigmente*, Paper no. 18, Frankfurt 1989.
395. K. Borho et al., *Chem. Ing. Tech.* **63** (1991) no. 8, 792.
396. H. Rumpf, *Chem. Ing. Tech.* **46** (1974) no. 1, 1.
397. R. Tawashi, *Pharm. Ind.* **25** (1963) 64.
398. H. Ferch, H. Gerofke, H. Itzel, H. Klebe, *Arbeitsmedizin Sozialmedizin Präventivmedizin* **22** (1987) 6; **22** (1987) 23.
399. F. Leuschner, unpublished results.
400. J. J. Sarre, unpublished results.
401. Degussa, DE 1 180 723, 1963 (H. Biegler, W. Neugebauer, H. Kempers).
402. Degussa, DE 1 933 291, 1969 (A. Illigen, W. Neugebauer).

403. Lonza, DE 2 337 495, 1972 (C. Schnell, M. Wissler, K. Hengartner).
404. E. R. Schnell, S. M. L. Hamblyn, K. Hengartner, M. Wissler, *Powder Technol.* **20** (1978) 15.
405. Degussa, DE-AS 1 467 019, 1963 (A. Becker, P. Nauroth).
406. Degussa, DE-AS 1 168 874, 1962 (G. Roderburg, P. Nauroth).
407. Degussa, DE-AS 1 299 617, 1965 (H. Reinhardt, K. Achenbach, P. Nauroth).
408. Degussa, EP 0 078 909, 1982 (P. Nauroth, H. Esch, G. Türk).
409. Degussa, DE 1 767 332, 1968 (G. Türk, J. Welsch).
410. Degussa, EP 0 341 383, 1989 (D. Kerner, A. Wagner, F. Schmidt, D. Bauer).
411. Degussa, DE-OS 3 114 493, 1981 (P. Nauroth, R. Kuhlmann, G. Türk, A. Becker).
412. Akzo, DE-AS 2 020 887, 1970 (G. Steenken, E. Seeberger).
413. Rhône-Poulenc, EP 0 018 866, 1980 (J-L. Ray, M. Condurier).
414. Rhône-Poulenc, EP 0 170 578, 1985 (Y. Chevallier).
415. Rhône-Poulenc, EP 0 396 450, 1990 (Y. Chevallier).
416. Degussa, DE-OS 3 639 845, 1986 (H. Reinhardt, A. Becker, R. Kuhlmann, P. Nauroth).
417. Tokuyama Soda, JP-Kokai 59 141 416, 1983.
418. Crosfield, US 4 127 641, 1977 (D. Aldcroft, D. Barby, A. L. Lovell, J. P. Quinn).
419. Huber, US 4 260 454, 1978 (S. K. Wason, R. K. Mays).
420. Sifrance, DE-AS 2 224 061, 1972 (J.-B. Donnet, B. Baudru, M. Condurier, G. Vrisakis).
421. Degussa, DE 966 985, 1950 (K. Andrich).
422. Nippon Silica, JP 03 75 215, 1989 (Y. Mikamoto, K. Sakata).
423. Degussa, DE 1 229 504, 1962 (G. Kallrath).
424. Bayer, DE-OS 3 525 802, 1985 (L. Puppe, O. Schlak, J. Ackermann, T. Naumann).
425. Degussa, DE-AS 1 192 162, 1962 (L. Hüter, G. Steenken).
426. PPG, DE-AS 1 131 196, 1955 (F. S. Thornhill, E. M. Allen).
427. PPG, DE 1 283 207, 1961 (C. B. Lagerstrom).
428. Degussa, DE-AS 2 505 191, 1975 (B. Brandt, P. Nauroth, A. Peters, H. Reinhardt).
429. Degussa, DE-OS 1 467 437, 1965 (G. Bretschneider, K. Pfeiffer, R. Schwarz, K. Laun).
430. Degussa, DE-AS 1 293 138, 1965 (H. Reinhardt, P. Nauroth).
431. Degussa, US 4 179 431, 1979 (E. Kilian, A. Kreher, P. Nauroth, G. Türk).
432. Degussa, DE-AS 1 567 440, 1965 (H. Biegler, K. Trebinger).
433. Degussa, *Schriftenreihe Pigmente*, Paper no. 70, 2nd ed., Frankfurt/Main 1984.
434. Degussa, *Schriftenreihe Pigmente*, Paper no. 28/1, 5th ed., Frankfurt/Main 1998.
435. Degussa, *Schriftenreihe Pigmente*, Paper no. 32, 4th ed., Frankfurt/Main 1989.
436. H. Ferch, *Chem. Ing. Tech.* **48** (1976) no. 11, 922.
437. Degussa, DE 1 172 245, 1963 (G. Kallrath).
438. Degussa, DE 2 513 608, 1975 (H. Reinhardt, K. Trebinger, G. Kallrath).
439. Degussa, DE 2 729 244, 1977 (P. Nauroth et al.).
440. Degussa, DE 1 074 559, 1959.
441. Degussa, DE 2 628 975, 1976 (P. Nauroth et al.).
442. Degussa, DE 1 667 465, 1967 (P. Nauroth, R. Kuhlmann).
443. Degussa, DE 1 592 865, 1967 (O. Kuehnert, G. Türk, E. Eisenmenger).
444. Degussa, *Schriftenreihe Pigmente*, Paper no. 16/1, 4th ed., Frankfurt/Main 1992.
445. H. J. Bachmann, J. W. Sellers, M. P. Wagner, R. F. Wolf, *Rubber Chem. Technol.* **32** (1959) 1286.
446. G. Kraus: *Reinforcement of Elastomers*, Interscience, New York 1965.
447. S. Wolff, *Kautsch. Gummi Kunstst.* **41** (1988) no. 7, 674.
448. S. Wolff, *Kautsch. Gummi Kunstst.* **34** (1981) no. 4, 280.
449. S. Wolff, *Kautsch. Gummi Kunstst.* **36** (1983) no. 11, 969.
450. Degussa, DE-OS 3 305 373, 1983 (S. Wolff, P. Golombeck).
451. R. Bode, A. Reisert, *Kautsch. Gummi Kunstst.* **32** (1979) no. 2, 89.
452. P. Vondracek, M. Schätz, *Kautsch. Gummi Kunstst.* **33** (1980) no. 9, 699.
453. K. H. Müller, *Mühle Mischfuttertech*. **114** (1977) 28.
454. K. H. Müller, *Kraftfutter* **53** (1970) 436.
455. Degussa, *Schriftenreihe Pigmente*, Paper no. 31/1, 3rd ed., Frankfurt/Main 1999.
456. Degussa, *Schriftenreihe Pigmente*, Paper no. 1, 5th ed., Frankfurt/Main 1989.
457. O. Pfrengle, *Fette Seifen Anstrichm.* **63** (1961) no. 5, 445.
458. Degussa, Technical Information, Paper no. TI 1159, Frankfurt/Main, 1994.

459. W. Leonhardt, *Speciality Chemicals* 1989, 441.
460. Degussa, DE 1 517 900, 1965 (H. Reinhardt, P. Nauroth, K. Achenbach).
461. General Electric, US 4 018 564, 1975 (J. H. Wright).
462. Degussa, unpublished report, 77–0016-DKT, 1977.
463. Degussa, unpublished report, 79–0004-DKT, 1979.
464. Degussa, unpublished report, 81–0016-DGT, 1981.
465. Degussa, unpublished report, 91–0131-DGT, 1991.
466. Degussa, unpublished report, 91–0132-DGT, 1991.
467. Degussa, unpublished report, 87–0004-DGT, 1987.
468. E. R. Plunkett, B. J. DeWitt, *Arch. Environ. Health* **5** (1962) 469–472.
469. Degussa, unpublished report, 75–0018-DKT, 1975.
470. Degussa, unpublished report, 81–0108-DKT, 1981.
471. D. Choudat, C. Frisch, G. Barrat, A. El Kholti, F. Couso, *Br. J. Med.* **47** (1990) 763–766.
472. Deutsche Forschungsgemeinschaft (ed.): *MAK und BAT-Werte-Liste*, WILEY-VCH, Weinheim, Germany 1999. pp. 157–165.
473. R. M. Barrer: "Porous Crystals: A Perspective" in Y. Murakani, A. Jijima, J. W. Ward (eds.): *New Developments in Zeolites Science and Technology,* Kodansha, Tokyo, and Elsevier, Amsterdam, Tokyo 1986, pp. 3–11.
474. E. M. Flannigen et al., *Nature* **271** (1978) 512–516.
475. W. M. Meier, D. H. Olson, C. Baerlocher: *Atlas of Zeolite Structure Types*, Butterworth, London 1987.
476. http://www.iza-online.org/
477. H. Gies, B. Marler, *Zeolites* **12** (1992) 42–49.
478. R. Szostak: "Modified Zeolites" in H. van Bekkum, E. M. Flannigan, J. C. Jansen (eds.): *Introduction to Zeolite Science and Practice,* Elsevier, Amsterdam 1991, pp. 153–199.
479. H. Gies et al., *Neues Jahrb. Mineral. Monatsh.* **3** (1982) 119–124.
480. B. Marler, *Phys. Chem. Miner.* **16** (1988) 286–290.
481. J. F. CooperJr., G. E. Dunning, *Am. Mineral.* **57** (1972) 1494–1499.
482. C. A. Fyfe, H. Gies, *J. Inclusion Phenom.* **8** (1990) 235–239.
483. H. Gies, *Z. Kristallogr.* **167** (1984) 13–82.
484. C. A. Fyfe et al., *Zeolites* **10** (1990) 278–282.
485. C. A. Fyfe et al., *J. Am. Chem. Soc.* **110** (1988) 3373–3380.
486. C. A. Fyfe et al., *J. Am Chem. Soc.* **111** (1989) 2470–2474.
487. J. A. Ripmeester et al., *J. Chem. Soc., Chem. Commun.* 1988, 608–611.
488. S. H. Park et al.: "The Thermal Expansion of the Zeolites MFI, AFI, DOH, DDR, and MTN in their Calcined and As Synthesized Forms" in H. Chon, S.-K. Ihm, Y. S. Uh (eds.): *Progress in Zeolite and Microporous Materials, Studies in Surface Science and Catalysis* vol. **5**, Elsevier, Amsterdam 1997, pp. 1989–1994.
489. a) N. Y. Chen, *J. Phys. Chem.* **80** (1976) 60–64. b) D. H. Olson et al., *J. Catal.* **61** (1980) 390–396.
490. R. M. Lago et al.: "The Nature of the Catalytic Sites in HZSM-5-Activity Enhancement" in Y. Murakami, A. Jijima, J. W. Ward (eds.): *New Developments in Zeolite Science and Technology,* Elsevier, Amsterdam 1986, pp. 677–684.
491. T. Bein et al.: "Molecular Sieve Films from Zeolite-Silica Microcomposites" in P. A. Jacobs, R. A. van Santen (eds.): *Zeolites: Facts, Figures, Future,* Elsevier, Amsterdam 1989, pp. 887–896.
492. J. C. Jansen et al.: "Zeolite Coatings – Potential Use in Catalysis and Separation" in M. M. J. Treacy, B. K. Marcus, M. E. Bisher, J. B. Higgins (eds.): *Proceedings of the 12th International Zeolite Conference*, MRS, Warrendale 1999, pp. 603–611.
493. Y. S. Lin, Y. H. Ma: "A Comparative Study of Adsorption and Diffusion of Vapor Alcohols and Alcohols from Aqueous Solutions in Silicalite" in P. A. Jacobs, R. A. van Santen (eds.): *Zeolites: Facts, Figures, Future,* Elsevier, Amsterdam 1989, pp. 877–886.
494. P. Enzel, T. Bein, *J. Phys. Chem.* **93** (1989) 6270 –6272.
S. D. Cox, G. D. Stucky, *J. Phys. Chem.* **95** (1991) 710–720.
495. N. Herron, Y. Wang, *J. Phys. Chem.* **91** (1987) 2757.
S. Miyazaki, H. Yoneuyama, *Denki Kagaku* **58** (1990) 37–40.
496. S. T. Homeyer, W. M. H. Sachtler: "Design of Metal Clusters in NaY Zeolite" in P. A. Jacobs, R. A. van Santen (eds.): Zeolites: Facts, Figures, Future, Elsevier, Amsterdam 1989, pp. 975–984.
K. Möller et al., *J. Phys. Chem.* **93** (1989) 6116 –6120.

497. S. D. Cox et al., *Chem. Mater.* **2** (1990) 609–619.
498. G. Ozin et al., *Angew. Chem.* **101** (1989) 373–390.
499. H. Gies et al., *Angew. Chem.* **94** (1982) 214–215.
500. H. K. Chae et al.: "Clathrasils: New Materials for Nonlinear Optics" in G. Stucky (ed.): "New Materials for Nonlinear Optics," *Am. Chem. Soc. Symp. Ser.* 1991.
501. H. Gies: "Clathrasils and Zeosils: Inclusion Compounds with Silica Host Framework" in J. L. Atwood, J. E. D. Davies, D. D. McNicol (eds.): *Inclusion Compounds,* vol. 5, Inorganic and Physical Aspects of Inclusion, Oxford Unversity Press, Oxford 1991, pp. 1–36.
502. H. Gies et al.: "Synthesis of Porosils: Crystalline Nanoporous Silicas with Cage- and Channel-Like Void Structure" in H. G. Karge, J. Weitkamp (eds.): *Molecular Sieves 1. Science and Technology. Synthesis*, Springer, Berlin 1998, pp. 35–65.
503. K. R. Franklin, B. M. Lowe, *Zeolites* **8** (1988) 491–516.
504. E. W. Valyocsic, L. D. Rollman, *Zeolites* **5** (1985) 123–125.
505. G. T. Kerr, *J. Phys. Chem.* **71** (1967) 4165. C. A. Fyfe et al., *J. Chem. Soc., Chem. Commun.* 1984, 1093–1094.
506. "Silica Flour: Silicosis," *NIOSH, Current Intelligence Bulletin* no. 36, 1981. DHHS (NIOSH) Publication no. 81–137, 1981.
507. Information Profiles on Potential Occupational Hazards, vol. II, Chemical Classes: Amorphous Silicas, NTIS, no. PB 81–147 886, 1979.
508. *Patty*: 3rd ed., **2 B**, p. 3012.
509. D. D. Dunnom (ed.): "Health Effects of Synthetic Silica Particulates," *ASTM Spec. Tech. Publ.* **732**, (1981) no. 04–732000–17.
510. E. Quellmalz: "Silikogener Staub," *Das neue Chemikaliengesetz; Arbeitsstoffverordnung*, Anhang II, 12/2.8, WEKA-Verlag, Kissing 1982.
511. Deutsche Forschungsgemeinschaft (ed.): *MAKund BAT-Wert-Liste*, VCH, Weinheim, Germany 1992, pp.96–102.
512. *Documentation of the Threshold Limit Values,* 4th ed. 1980, Suppl. Docum. 1981, Silica (amorphous), p. 362, Silica (fused), p. 363, Silica (quartz), p. 364. *Am. Conf. Gov. Ind. Hyg.,* Cincinnati, Ohio, 1981.
513. W. Klosterkötter, *Arch. Hyg. Bakteriol.* **149** (1965) 577–598.
514. W. H. Schepers et al.: *AMA Arch. Ind. Health* **16** (1957) 125–146.
515. H. Volk, *Arch. Environ. Health.* **1** (1960) 125–128.
516. H. Gerofke: *Beobachtungen an Beschäftigten des Aerosil-Betriebes*. Rheinfelden 1981, unpublished.
517. V. C. Vitums et al.: *Arch. Environ. Health* **32** (1977) no. 2, 62–68.
518. F. D. Goldsmith, T. L. Guidotti, D. R. Johnston: *Am. J. Ind. Med.* **3** (1982) 423.
519. P. Westerholm: *Scand. J. Work Environm. Health (Suppl).* **2/6** (1980) 1.
520. M. Finkelstein, R. Kusiak, G. Suranyi: *J. Occup. Med.* **24** (1982) 663.
521. H. Gulbergsson, K. Kurppa, H. Koskinen, M. Vasama: "Association between Silicosis and Lung Cancer: A Register Approach", *VI. International Pneumoconiosis Conference,* Bochum, Sept. 20–23, 1983.
522. C. G. O'Neill et al.: *Lancet* **202** (1982) 1202.
523. W. Klosterkötter, H. J. Einbrodt: *Arch. Hyg. Bakteriol.* **149** (1965) 367–384.
524. W. Klosterkötter: *Retention, Penetration und Elimination von Quarz bei einem Langzeit-Inhalationsversuch (Konzentration 10 mg/m^3)* vol. **9**, Verlag Glückauf, Essen 1973, p. 177.
525. H. Gärtner: *Arch. Hyg. Bakt.* **136** (1952) 451–467.
526. K. W. Jötten: *Gutachtliche Untersuchung und Beurteilung der Wirkung von Aerosilstaub auf das Lungengewebe von Versuchstieren*, Deutsche Gold- und Silber-Scheideanstalt, Frankfurt am Main 1952.
527. M. M. E. Wagner: *J. Natl. Cancer Inst.* **57** (1976) 509.
528. M. M. E. Wagner, J. C. Wagner, R. Davies, D. M. Griffiths.: *Br. J. Cancer* **41** (1980) 908.
529. F. Stenbäck, J. Rowland.: *Oncology* **36** (1979) 63.
530. L. M. Holland, M. Gonzales, J. S. Wilson, M. I. Tillery: "Pulmonary Effects of Shale Dusts in Experimental Animals," *Environmental Health Conference,* Park City, April 6–9, 1982.

Silicates

Asbestos, Cement and Concrete, Clays, Glass, Glass Ceramics, Mica, Silica, Talc, and Zeolites are separate keywords.

GERHARD LAGALY, Institute of Inorganic Chemistry, University of Kiel, Kiel, Federal Republic of Germany (Chap. 1)

WERNER TUFAR, Department of Geosciences, Philipps University Marburg, Marburg, Federal Republic of Germany (Chaps. 1, 2)

A. MINIHAN, Unilever Research Port Sunlight Laboratory, Bebington, Wirral, United Kingdom (Sections 3.1 – 3.4 and 3.10)

A. LOVELL, Crosfield Chemicals, Warrington, Cheshire, United Kingdom (Sections 3.5 – 3.9)

1. Structural Chemistry of Silicates . . 2
1.1. Structural Classification 3
1.2. Oligo- and Cyclosilicates 4
1.3. Polysilicates 4
1.4. Phyllosilicates 6
1.4.1. Monophyllosilicates 6
1.4.2. Diphyllosilicates 8
1.4.3. Alkali Silicates 8
1.4.4. Crystalline Silicic Acids 8
1.4.5. Clay Minerals 10
1.4.6. Intracrystalline Reactions 13
1.5. Tectosilicates 14
2. Natural Silicates 15
2.1. Feldspar 15
2.1.1. Structure and Composition 15
2.1.2. Characterization of Individual Feldspars 17
2.1.3. Production 18
2.1.4. Properties 19
2.1.5. Mineral Deposits and their Extraction 20
2.1.6. Processing and Quality Requirements 21
2.1.7. Uses and Economic Aspects 21
2.2. Nepheline and Related Compounds 24
2.3. Leucite 27
2.4. Olivine 28
2.5. Andalusite 31
2.6. Kyanite 34
2.7. Sillimanite 36
2.8. Mullite 37
2.9. Vermiculite 40
2.10. Perlite 44
2.11. Pumice 46
2.12. Basalt 47
2.13. Wollastonite 47
2.14. Toxicology 52
3. Alkali Silicates 52
3.1. Introduction 52
3.2. Raw Materials 53
3.3. Amorphous Anhydrous Alkali Silicates (Solid or Lump Glasses) 53
3.4. Silicate Solutions 55
3.5. Hydrated Water-Soluble Silicates . 58
3.6. Crystalline Solids 58
3.7. Uses and Applications 59
3.8. Economic Aspects 61
3.9. Storage, Safety, Labelling and Transportation 61
3.10. Analysis 62
4. References 63

1. Structural Chemistry of Silicates

More than 95 vol % of the earth's crust is composed of quartz and a few rock-forming silicates (plagioclase 42 vol %, potash feldspar 22 vol %, quartz 18 vol %, amphibole 5 vol %, other silicates 12 vol %). Silicon compounds are also present in the hydrosphere, mainly as dissolved silica. A large number of silicates have been identified in extraterrestrial material [1].

The widespread abundance of silicate minerals is accompanied by a diversity of practical uses (Table 1). Silicates are components of technical materials such as cement (→ Cement and Concrete) and glass (→ Glass, Glass Ceramics). Granite and sandstone are used as building materials. Quartz (→ Silica) is the raw material for producing waterglass, silicon (→ Silicon), and silicon compounds (→ Silicon Compounds, Inorganic, → Silicon Compounds,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_661

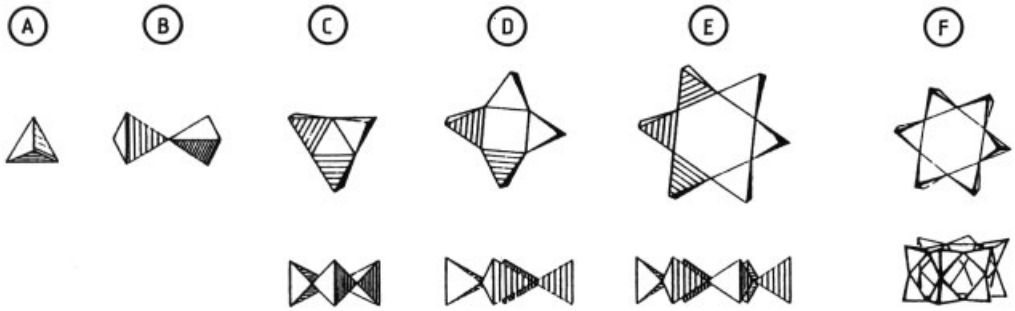

Figure 1. Oligosilicates and cyclosilicates
A) Monosilicates (forsterite, olivine, phenakite, garnets, zircon; formulae see text); B) Disilicates (thortveitite, $Sc_2[Si_2O_7]$; C) Monocyclosilicate: dreier single ring (benitoite, $BaTi[Si_3O_9]$); D) Monocyclosilicate: vierer single ring (taramellite, $Ba_4(Fe,Ti)_4B_2[Si_4O_{12}]_2O_5Cl_x$); E) Monocyclosilicate: sechser single ring (beryl, $Be_3Al_2[Si_6O_{18}]$; tourmalines $XY_3Z_6B_3[Si_6O_{18}]O_9(O, OH, F)_4$ ($X = Na^+, Ca^{2+}$; $Z = Al^{3+}, Mg^{2+}$; $Y = Li^+, Mg^{2+}, Fe^{2+}, Mn^{2+}, Fe^{3+}, Al^{3+}$); dioptase, $Cu_6[Si_6O_{18}] \cdot 6\,H_2O$); F) Dicyclosilicate: sechser double ring (milarite, $KCa_2(Be_2Al)\,[Si_{12}O_{30}] \cdot 0.75\,H_2O$)

Organic, → Silicones) and modern nonclay ceramics (SiC, Si_3N_4) (→ Silicon Carbide). Modern life is not only determined by the use of synthetic polymers but also by the widespread application of silicates and products derived from them. For example, silicon forms the basis of microelectronics.

Table 1. Practical applications of natural silicates

Mineral (rock)	Uses	Production, 10^6 t/a
Kaolinite (kaolin)	paper industry, ceramics, chemical and pharmaceutical industries	25 (kaolin)
Smectite (bentonite)	foundry molding sands, iron ore pelletizing, oil well drilling, building industry, chemical and pharmaceutical industries	7 (bentonite)
Talc, pyrophyllite	filler, electrical insulation, ceramics	6
Feldspars	glass and ceramic industry, filler	3.5 ∗
Micas	filler, electrical industry	2.7
Olivine, forsterite	refractory masses, fluxes, foundry molding sands	2.3
Perlite	lightweight material for thermal and sound insulation	2.5
Vermiculite	thermal insulation, filler, lightweight material	0.65
Sillimanite	refractory masses	0.15
Andalusite	refractory masses	
Kyanite	refractory masses	
Mullite	refractory masses	0.005
Zircon	refractory material, ceramics, zirconium compounds	0.1
Spodumene, petalite	production of lithium	
Beryl	production of beryllium	
Chrysotil asbestos	(high toxicity)	4 (1978)

∗ Including nepheline, leucite: 4.5×10^6 t/a.

1.1. Structural Classification

Silicates comprise the largest, most extensive, and manifold class of minerals. One reason is that the bond energy of the Si – O bond is higher than that of the Si – Si and Si – H bonds (bond energies, kJ/mol: Si – O, 452; Si – Si, 222; Si – H, 318). Comparison with the values for carbon (C – O, 358; C – C, 346; C – H, 413) reveals why the skeletons in silicates are Si – O – Si – O chains instead of the C – C – C chains in carbon chemistry. The second reason is the ease with which silicon – oxygen tetrahedra can be linked to form rings, chains, layers, or three-dimensional frameworks. Diversity is increased by isomorphous substitution of silicon by small cations (mostly Al^{3+}, but also other tri- and divalent cations $B^{3+}, Fe^{3+}, Ga^{3+}, Be^{2+}, Zn^{2+}, Mg^{2+}$). In the 2 : 1 clay mineral sauconite $Zn_3[Si_{4-x}Zn_xO_{10}(OH)_2]$, Si^{4+} is substituted by Zn^{2+}. Synthetic mica phases in which Co^{3+} and Ni^{3+} replace Si^{4+} have been reported [1].

The common coordination number of silicon bonded to oxygen is four, but a few silicates together with the SiO_2 modification stishovite contain sixfold coordinated silicon, e.g., $Mg[SiO_3]$ (high-temperature modification), $(NH_4)_2[SiP_4O_{13}]$, $Ca_3[Si(OH)_6][SO_4][CO_3] \cdot 12\,H_2O$ (thaumasite), modifications of $Si[P_2O_7]$.

The most comprehensive classification of silicates is presented by Liebau [1–3] (Table 2). The term *multiplicity* (M) denotes the number of single polyhedra, rings, chains, or layers that are linked to the complex anion. *Dimensional-*

ity refers to the infinite extension of the anion: $D = 0$, groups; $D = 1$, chains; $D = 2$, layers; $D = 3$, framework. A silicate that contains double tetrahedra, as in $Sc_2[Si_2O_7]$ (thortveitite, Fig. 1B), is called a disilicate ($M = 2$, $D = 0$). Orthoenstatite $Mg_2[Si_2O_6]$ (Fig. 2A) is a monopolysilicate that contains single chains of silicon – oxygen tetrahedra.

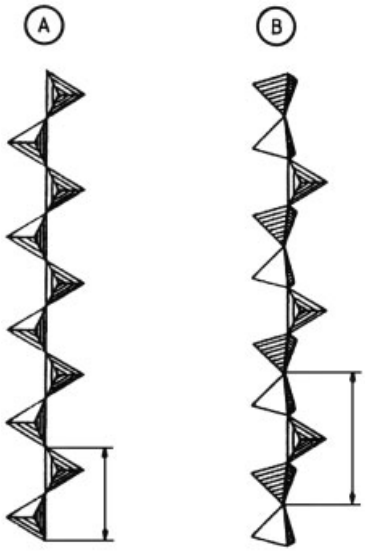

Figure 2. Monopolysilicates
A) Zweier single chains (pyroxenes); B) Dreier single chains (wollastonite, $Ca_3[Si_3O_9]$

Sometimes, particularly in the technical literature, silicates of composition $MO \cdot 2\,SiO_2$, $MO \cdot 3\,SiO_2$ (M = metal) are called di- and trisilicates, respectively. This practice is in contradiction to IUPAC rules and must be abandoned. Also, the term "metasilicates" for silicates M_2SiO_3, $MSiO_3$, $M_2(SiO_3)_3$ should be abandoned because these silicates can be cyclo- or polysilicates.

Connectedness of $[SiO_4]$ tetrahedra is a further parameter that should be mentioned in classifying silicate structures. A $[SiO_4]$ tetrahedron can share common oxygen ions with 0, 1, 2, 3, or 4 $[SiO_4]$ tetrahedra; this is indicated by Q^0, Q^1, Q^2, Q^3, or Q^4, respectively. The connectedness of silicon – oxygen tetrahedra can be obtained from magic angle spinning – NMR (MAS – NMR) measurements. This is particularly advantageous in the study of amorphous materials or silicates with high levels of defects or of unknown structure.

Mineralogists classify the silicates into neso-, soro-, cyclo-, ino-, phyllo-, and tectosilicates (Table 3).

In classifying silicates on the basis of the way in which the silicon – oxygen tetrahedra are linked to each other, it should be remembered that the real structure is determined by the interaction between the $[SiO_4]$ tetrahedra and the nontetrahedral cations (Section 1.3).

1.2. Oligo- and Cyclosilicates

When not specified otherwise, references concerning structure, crystallographic data, or synthesis of the silicates mentioned are found in [1].

Single $[SiO_4]$ tetrahedra are found in *monosilicates* (Fig. 1A) such as forsterite $Mg_2[SiO_4]$, olivine $(Mg, Fe)_2[SiO_4]$, phenakite $Be_2[SiO_4]$, garnets $M^{12}_3M^6_2[SiO_4]_3$ ($M^{12} = Mg^{2+}$, Fe^{2+}, Mn^{2+}; $M^6 = Al^{3+}$, Fe^{3+}, Cr^{3+}), and zircon $Zr[SiO_4]$. Because of the high charge of the $[SiO_4]^{4-}$ anion, the stability of monosilicates increases in the order:

$$K_4[SiO_4] < Ca_2[SiO_4] < Ca_3Al_2[SiO_4]_3 < Zr[SiO_4]$$

Examples of *disilicates* (Fig. 1 B) are thortveitite and barysilite $MnPb_8[Si_2O_7]_3$. Synthetic $Ag_{10}[Si_4O_{13}]$ is a *tetrasilicate*. *Oligosilicates* with unbranched chains $M = 8$, 9, 10 have been synthesized.

A few *monocyclosilicates* (with single rings, $M = 1$) and a *dicyclosilicate* are shown in Figure 1 C – F.

Several mono- and cyclosilicates are of practical importance. Magnesium silicate, $Mg_2[SiO_4]$, is used as a refractory material (forsterite products, Section 2.4). Olivine (ca. 2.3×10^6 t/a) is mainly used as flux for slags in blast furnaces but small amounts are needed as foundry molding sands (Section 2.4). Garnets are utilized as bearings, polishing powder, abrasive products, and precious stones. (The garnets used as ferrimagnetic materials are yttrium and rare-earth iron garnets $M^{3+}_3Fe_5O_{12}$ which consist of $[FeO_4]$ tetrahedra in place of $[SiO_4]$ tetrahedra.) Thortveitite (with scandium ions partially substituted by yttrium ions) is an important scandium mineral.

Table 2. Classification of silicates [1]

Dimensionality		Multiplicity			
		1	2	3	4
0	Oligosilicates	monosilicates	disilicates	trisilicates	tetrasilicates
0	Cyclosilicates	monocyclosilicates	dicyclosilicates	tricyclosilicates	tetracyclosilicates
1	Polysilicates	monopolysilicates	dipolysilicates	tripolysilicates	tetrapolysilicates
2	Phyllosilicates	monophyllosilicates	diphyllosilicates	triphyllosilicates	tetraphyllosilicates
3	Tectosilicates	tectosilicates			

Table 3. Nomenclature of silicates used by mineralogists

Dimensionality	Multiplicity				
	1		2	3	4
0	nesosubsilicates	nesosilicates	sorosilicates		
0			cyclosilicates		
1			inosilicates		
2			phyllosilicates		
3	tectosilicates				

Several oligo- and cyclosilicates are valued as semiprecious or precious stones: garnets, olivine (chrysolite), zircon (hyacinth, jargon), beryl (emerald, aquamarine, heliodor, dioptas), turmaline (rubellite, schorl, indigolit).

Zircon is an abundant accessory mineral, for example in glass ceramics (→ Glass Ceramics). The radioactivity of rocks is often caused by the thorium and uranium content in zircon.

1.3. Polysilicates

Monopolysilicates. Polysilicates are abundant minerals, several are of practical importance. Single- chain polysilicates (dimensionality 1, multiplicity 1) are classified according to the number of silicon – oxygen tetrahedra contained in the repeat unit of the chain (*periodicity,* $P = 2$, 3, etc.). Figure 2 shows polysilicate single chains with periodicities of two and three. Monopolysilicates are known with periodicities in the range 3 – 24 ($P = 24$ in synthetic $Na_{24}Y_8[Si_{24}O_{72}]$ [1]).

The large group of pyroxenes (Fig. 2 A) consists of monopolysilicates with zweier single chains. Examples include enstatite ($Mg_2[Si_2O_6]$), diopside ($CaMg[Si_2O_6]$), and spodumene ($LiAl[Si_2O_6]$). Spodumene is a source for lithium (→ Lithium and Lithium Compounds, Chap. 3.1.1.) and glass ceramic production (→ Glass Ceramics). Enstatite and diopside are sometimes found in ceramic masses (→ Glass Ceramics). Several calcium silicates contain dreier single chains, for example wollastonite ($Ca_3[Si_3O_9]$, modifications 1T, 2M, 7T, Fig. 2 B). Wollastonite is used as a ceramic material, as a filler in plastics and paints, and as a substitute for asbestos.

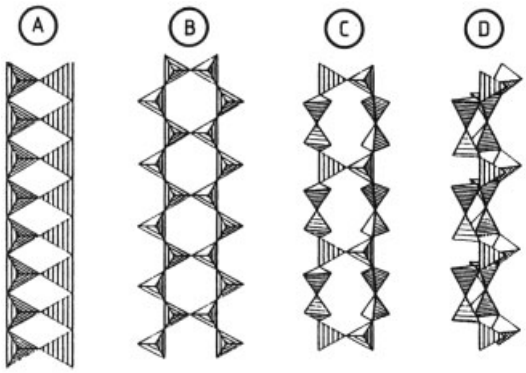

Figure 3. Dipolysilicates
A) Einer double chain (high-temperature sillimanite, $Al^6[AlSiO_5]$ (hT)); B) Zweier double chain (amphiboles, e.g., tremolite, $Mg_5Ca_2[Si_4O_{11}]_2(OH)_2$); C) Dreier double chain (xonotlite, $Ca_6[Si_6O_{17}](OH)_2$; D) Dreier double chain (epididymite, $Na_2Be_2[Si_6O_{15}] \cdot H_2O$)

Dipolysilicates. The arrangement of the tetrahedra in sillimanite (high-temperature modification) is shown in Figure 3 A. However, the tetrahedral Al^{3+} and Si^{4+} ions are almost completely ordered so that sillimanite contains zweier double chains (Fig. 3 B) [1]. Zweier double chains are also the skeletons in am-

phiboles (Fig. 3 B). Xonotlite and epididymite are dipolysilicates with dreier double chains (Fig. 3 C, D).

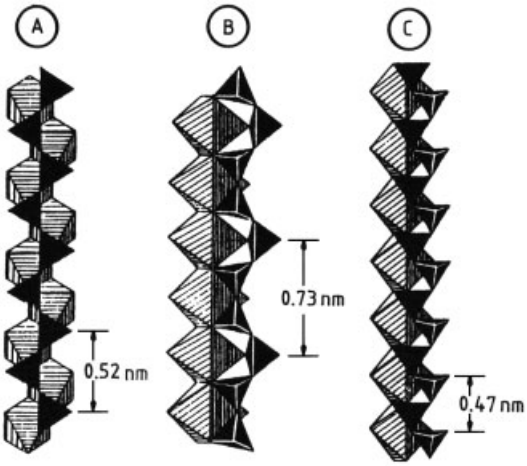

Figure 4. Mutual adjustment of polysilicate anions and columns of $[MO_6]$ octahedra
A) Enstatite, $Mg_2[Si_2O_6]$; B) Wollastonite, $Ca_3[Si_3O_9]$ (low-temperature modification); C) Synthetic $Ba_2[Si_2O_6]$ (high-temperature modification) $Na_2Be_2[Si_6O_{15}] \cdot H_2O$)

Oligopolysilicates. Synthetic oligopolysilicates are $Ba_4[Si_6O_{16}]$ (zweier triple chain), $Ba_5[Si_8O_{21}]$ (zweier fourfold chain), and $Ba_6[Si_{10}O_{26}]$ (zweier fivefold chain).

Influence of Nontetrahedral Cations on the Structure of Polysilicates. The nontetrahedral cations exert a significant influence on the structure of polysilicates, in particular on their periodicity [4]. The mutual adaption of the conformation of the silicate chains to columns of $[MO_n]$ polyhedra is shown for enstatite, wollastonite, and synthetic $Ba_2[Si_2O_6]$ (high-temperature modification) in Figure 4. The dimensions of two $[MgO_6]$ octahedra (ionic radius of Mg^{2+}: 0.072 nm) in enstatite (Fig. 4 A) correspond to the repeat unit (identity period; two tetrahedra) of the stretched zweier single chain. In wollastonite (Fig. 4 B) the dreier single chain fits the column of the considerably larger $[CaO_6]$ octahedra (two octahedra and three tetrahedra; ionic radius of Ca^{2+}: 0.10 nm). A further increase of the ionic radius of the octahedral cation (Ba^{2+}: 0.135 nm) allows the adjustment of two $[SiO_4]$ tetrahedra to one $[BaO_6]$ octahedron: the polyanion assumes the form of a zweier single chain with the identity period distinctly smaller than in enstatite. Further examples of the adaptation of the silicate chain to edge-sharing $[MO_6]$ octahedra are shown for pyroxenes and pyroxenoids in Figure 5 [4].

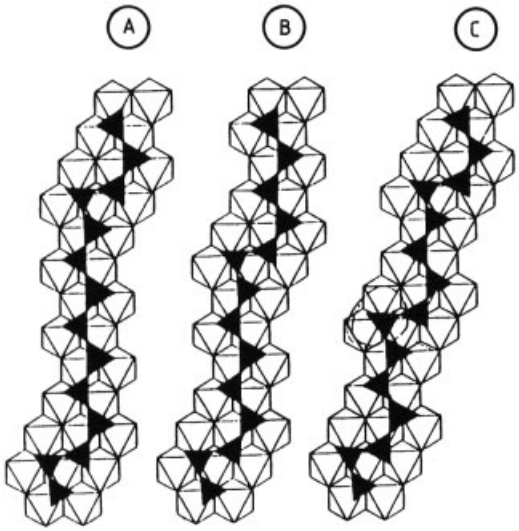

Figure 5. Adjustment of polysilicate anions to "slabs" of $[MoO_6]$ octahedra
A) Ferrosilite-III, $FeSiO_3$ (high-temperature, medium-pressure modification) neuner single chain and $[FeO_6]$ octahedra; B) Pyroxmangite-type $MnSiO_3$ (medium-pressure modification) (pyroxmangite = $(Fe,Mn)_7[Si_7O_{21}]$), siebener single chain and $[MnO_6]$ octahedra; C) Rhodonite-type $MnSiO_3$ (low-pressure modification) (rhodonite = $(Mn, Ca_5)[Si_5O_{15}]$), single chain and $[MnO_6]$ octahedra

Other factors besides the size of the nontetrahedral cations (e.g., electronegativity and valency) also control the conformation of the polysilicate anion [2, 4].

1.4. Phyllosilicates

1.4.1. Monophyllosilicates

Layers of silicon – oxygen tetrahedra are formed when silicate single chains are linked with each other. A simple possibility is to build up the zweier single layer (Fig. 6) by connecting zweier single chains. The term *directedness* indicates the way in which the apical oxygen ions of the $[SiO_4]$ tetrahedra are oriented. In Figure 6 A all apical oxygen ions point in the same direction.

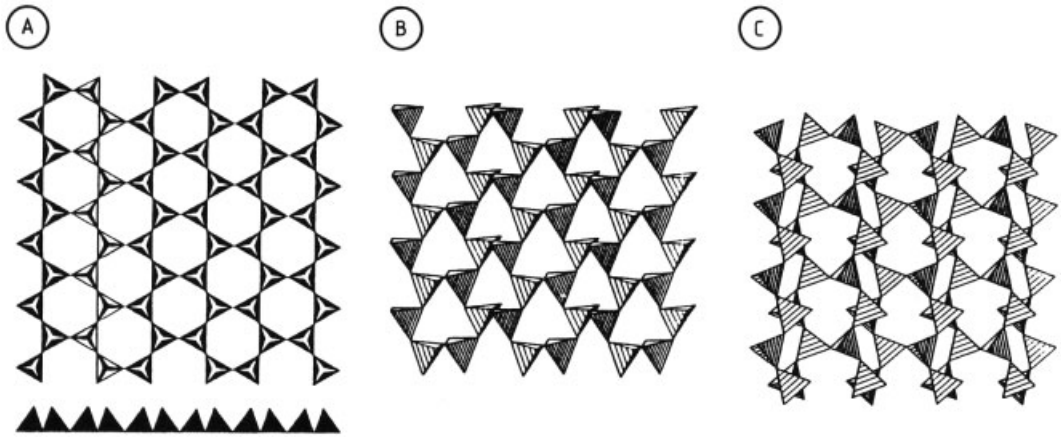

Figure 6. Formation of a zweier single-layer phyllosilicate by linking zweier single chains
A) Unfolded layer with terminal oxygen ions pointing in the same direction; B) Folded layer with terminal oxygen ions pointing up and down: $Li_2[Si_2O_5]$; C) Folded layer in petalite, $Li^{[4]}Al^{[4]}[Si_2O_5]_2$

In many other cases, however, they point in opposite directions (Fig. 6 B, C; Fig. 7).

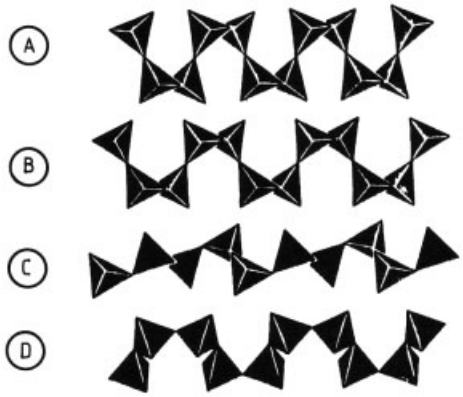

Figure 7. Folding of tetrahedral layers in anhydrous monophyllosilicates
A) $Li_2[Si_2O]_5$; B) α-$Na_2[Si_2O_5]$; C) β-$Na_2[Si_2O_5]$; D) Sanbornite, $Ba[Si_2O_5]$ (low-temperature modification)

The layers shown in Figure 6 are highly flexible, even when the terminal oxygen ions alternate between both sides. The layers can be stabilized in three ways:

1) The layers are folded
2) The layers are linked to form diphyllosilicate anions or structures with higher multiplicities
3) The layers are linked to octahedral layers of nontetrahedral cations (Section 1.4.5)

Folded zweier single layers are found in several alkali and barium phyllosilicates (Fig. 7 A–D). The degree of folding depends on the cation. The layers are highly folded in $Li_2[Si_2O_5]$ (Fig. 7 A). The occurrence of different modifications of $Na_2[Si_2O_5]$ (Fig. 7 B, C) reveals that the convolution (degree of folding) is not only determined by the nontetrahedral cation.

A divalent cation must balance the charge of two $[SiO_4]$ tetrahedra. This increases the folding of the layer. The silicates $Ca[Si_2O_5]$ and $Sr[Si_2O_5]$ do not exist because the folding of the layers, which in turn is a consequence of the electrical field of the cations, provides coordination numbers too small for calcium and strontium ions. In the presence of barium ions (Fig. 7 D), which have a larger ionic radius, the degree of folding is not as large and provides sites for barium ions with sufficiently high coordination numbers (9 in sanbornite $Ba[Si_2O_5]$, low-temperature modification).

Anhydrous, single-layer silicates containing solely trivalent cations are not known because even strong convolution cannot produce a well-balanced charge distribution. However, Li^+ and Al^{3+} form a very stable, anhydrous, layer silicate with highly folded layers (petalite, $Li^{[4]}Al^{[4]}[Si_2O_5]_2$, Fig. 6 C). The stability is related to the tetrahedral coordination of Li^+ and Al^{3+} ions which is a consequence of the high degree of layer folding induced by the trivalent aluminum ion.

Examples of monophyllosilicates with layers composed of silicate chains with $P > 2$ are apophyllite (Fig. 8) and gillespite $BaFe[Si_4O_{10}]$ with vierer single layers. Synthetic $CaCu[Si_4O_{10}]$, isostructural to gillespite, was used as a blue pigment in wall paintings ("Egyptian Blue"). The mineral name is cuprorivaite.

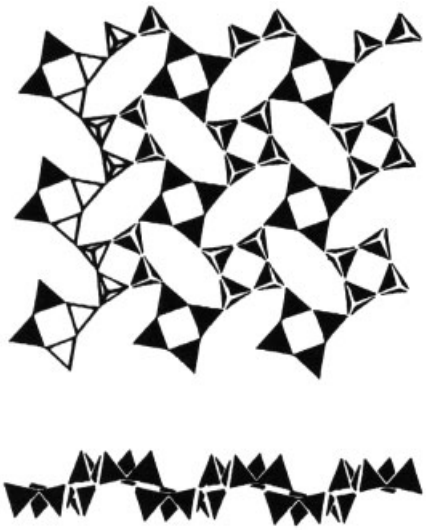

Figure 8. Vierer single layer in apophyllite, $KCa_4[Si_4O_{10}]_2(F, OH) \cdot 8H_2O$

1.4.2. Diphyllosilicates

Stabilization of zweier single layers by linkage to diphyllosilicate anions is illustrated in hexacelsian, $Ba(AlSiO_4)_2$ (high-temperature modification, Fig. 9). Further examples are vierer double layers in naujakasite and reyerite and even neuner double layers in stilpnomelane.

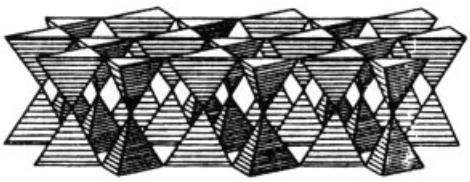

Figure 9. Diphyllosilicates: zweier double layer in hexacelsian, $Ba[(Al, Si)O_4]_2$ high-temperature modification

1.4.3. Alkali Silicates

See also Chapter 3. Among the anhydrous alkali silicates only natrosilite, $Na_2[Si_2O_5]$, occurs as a mineral [5]; all others are synthetic.

$Na_4[Si_2O_6]$ is a monopolysilicate containing zweier single chains. Most of the alkali silicates are monophyllosilicates with zweier single layers: $Li_2[Si_2O_5]$ (Fig. 7 A); α-, β-$Na_2[Si_2O_5]$ (Fig. 7 B, C); $Na_4[Si_6O_{14}]$; $KH[Si_2O_5]$ [6]; and $K_2[Si_4O_9]$. Silicates with Si : O ratios $< 2 : 5$ contain Q^4 tetrahedra besides the Q^3 tetrahedra. Potassium analogues of $Na_2[Si_2O_5]$ are unknown [1].

Anhydrous alkali silicates are formed when quartz is melted with sodium or potassium carbonate or hydroxide (alkali silicates, waterglass). The reaction product is dissolved in water to form waterglass solutions. δ-$Na_2[Si_2O_5]$ is considered as a builder and possible substitute for zeolites in washing powders [7].

Several hydrated alkali silicates occur as minerals:

Kanemite	$NaH[Si_2O_5] \cdot 3H_2O$	[8]
Makatite	$Na_2[Si_4O_8](OH)_2 \cdot 4H_2O$	[9]
Magadiite	$Na_2[Si_{14}O_{29}] \cdot 11H_2O$ *	[10]
Kenyaite	$Na_2[Si_{20}O_{41}] \cdot xH_2O$ *	[10]
Revdite	$Na_2[Si_2O_5] \cdot 5H_2O$	[11]
Grumantite	$NaH[Si_2O_5] \cdot 0.9H_2O$	[12]
Silinaite	$NaLi[Si_2O_5] \cdot 2H_2O$	[13]

* MAS – NMR measurements reveal that the formula should be written as $Na_2H_2Si_{14}O_{30} \cdot xH_2O$, $Na_2H_2Si_{20}O_{42} \cdot xH_2O$ [14]. EUGSTER [10] reported the composition $Na_2Si_{22}O_{45} \cdot 10H_2O$ for kenyaite.

The structures of sodium monosilicates $Na_2H_2[SiO_4] \cdot xH_2O$ ($x = 1 - 8$) [15], makatite [16], and silinaite have been determined [13]. Makatite contains vierer single layers. The zweier single layer in silinaite shows a directedness not observed previously: the slightly kinked zweier single chains alternate such that in one chain the apices point up and in the adjoining chain the apices point down. The different dimensions of the $[LiO_4]$ tetrahedra and $[NaO_2(OH_2)_4]$ octahedra which hold the layers together, impose a slight wave in the silicate layer.

The natural alkali silicates and several other alkali silicates ($Na_2H_2Si_8O_{18} \cdot xH_2O$ [17]; $K_2H_2Si_{14}O_{30} \cdot xH_2O$; $K_2H_2Si_{20}O_{42} \cdot xH_2O$) can be easily synthesized from dispersions of SiO_2 in aqueous NaOH or KOH solutions [17–20]. A characteristic property of these sili-

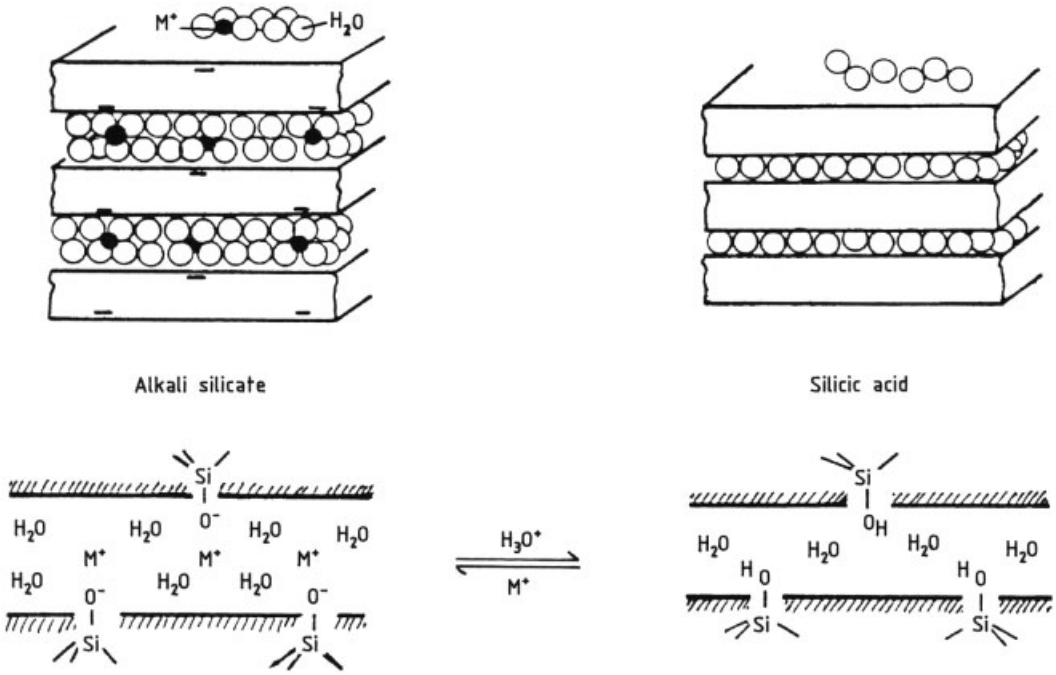

Figure 10. Preparation of crystalline silicic acids from alkali phyllosilicates by exchanging protons for the interlayer cations

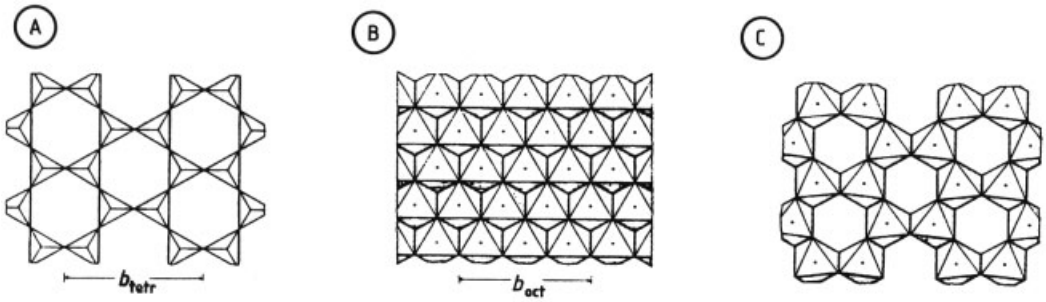

Figure 11. Clay mineral structure
A) Zweier single layer ("tetrahedral sheets"); B), C) Layers of edge-sharing $[MO_6]$ octahedra ("octahedral sheets")
Clay minerals with octahedral sheets as in (B) are called "trioctahedral", those with sheets (C) are called "dioctahedral".

cates is their cation exchange capacity (see Section 1.4.4). Possible practical uses include ion exchangers, adsorbents, or builders in detergents [21].

1.4.4. Crystalline Silicic Acids

Exchanging protons for the cations sandwiched between the silicate layers transforms the alkali phyllosilicates into crystalline acids (Fig. 10). In spite of the fact that the first crystalline silicic acid was prepared from α-$Na_2Si_2O_5$ in 1924 [22], most textbooks of inorganic chemistry refer to the orthosilicic acid $Si(OH)_4$ as the only oxo acid of silicon. The crystal structures of two $H_2Si_2O_5$ acids prepared from α-$Na_2Si_2O_5$ were determined by Liebau [23].

Presently, more than 15 crystalline acids are known and comprise at least six modifications of $H_2Si_2O_5$. The acids are derived from the natural and synthetic alkali silicates mentioned above, from several synthetic and natural cuprous silicates, and from a few other silicates such as apophyllite, carletonite, and gillespite [24–26]. The acids derived from

$Na_2H_2Si_{14}O_{30} \cdot xH_2O$, $K_2H_2Si_{14}O_{30} \cdot xH_2O$, and $K_2H_2Si_{20}O_{42} \cdot xH_2O$ have been considered for practical uses. These acids are interesting alternatives to various forms of silica. Practical applications may be related to adsorption properties [27] and the capability to intercalate neutral organic molecules (Section 1.4.6). A few acids even adsorb anionic surfactants (together with the gegen ions) in the interlayer space [28].

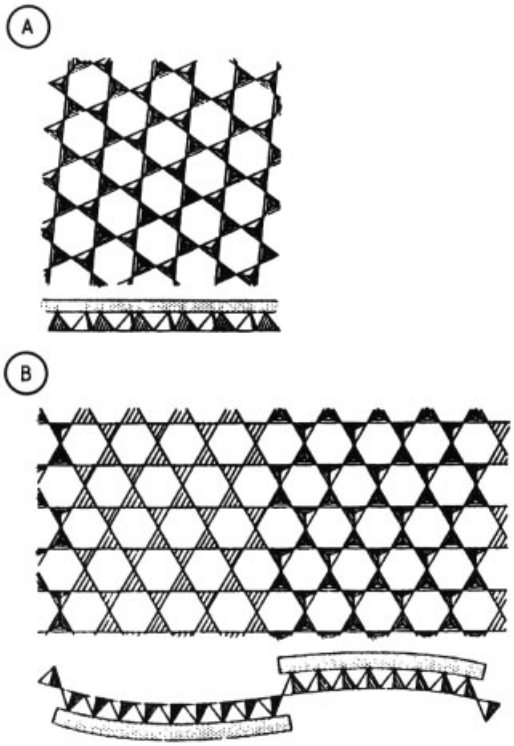

Figure 12. Linking of zweier single layers to octahedral sheets (dotted areas) in clay minerals
A) Kaolinite, $Al_2[Si_2O_5(OH)_4]$; B) Antigorite, idealized $Mg_3[Si_2O_5(OH)_4]$

1.4.5. Clay Minerals

See also Chapter 2.

In clay minerals (Table 4) zweier single layers (Fig. 11 A) are linked to octahedral sheets (Fig. 11 B, C). (In clay mineralogy the single tetrahedral and octahedral layers are called tetrahedral and octahedral sheets; the assembly of these sheets is called a layer [29].) The structures provide a good example of different types of directedness. The terms "clays, kaolins and bentonites" are used somewhat differently by geologists and in practical applications (see Chaps. 1, 5, 7 in [31]). A clear distinction must be made between "clay" (the rock) and "clay mineral" (the mineral) [32].

1:1 Clay Minerals. In kaolinite ($Al_2[Si_2O_5(OH)_4]$) all terminal oxygen ions of the zweier single layer (tetrahedral sheet) point to one side and simultaneously belong to the octahedral sheet (Figs. 11 A, 12 A,13 A). Thus, the octahedra around the octahedral cations consist of these oxygen ions and additional OH^- ions.

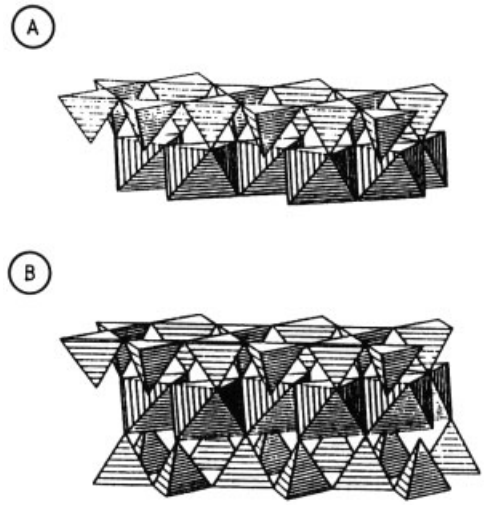

Figure 13. Linking of tetrahedral and octahedral sheets in clay minerals
A) 1 : 1 clay minerals (kaolinite, serpentines); B) 2 : 1 clay minerals (micas, vermiculites, smectites)

As the dimensions of the octahedral and the tetrahedral sheets do not match exactly ($b_{oct} \neq b_{tetr}$, Fig. 11), mechanical stress arises between both sheets (for details see [1, 29]). This causes curling of the crystals when their thickness decreases below a certain value [33]. Halloysite crystals are composed of kaolinite-type layers separated by a water layer, and often assume a tubular morphology.

The misfit between both types of sheets in the serpentine minerals (chrysotile, antigorite, lizardite), which are the trioctahedral counterparts of kaolinite (idealized formula $Mg_3[Si_2O_5(OH)_4]$), also leads to the curling of

Table 4. Classification of 1 : 1 and 2 : 1 clay minerals [30]

ξ [a]	Group	Octahedral character [b]	Predominant octahedral cation [c]	Species
Layer type 1: 1				
	kaolin – serpentine	tri	Mg^{2+}	lizardite, chrysotile [d], antigorite [d]
		di	Al^{3+}	kaolinite, halloysite [d]
Layer type 2: 1				
0	talc – pyrophyllite	tri	Mg^{2+}	talc, willemseite
		di	Al^{3+}	pyrophyllite
0.2 – 0.6	smectite	tri	Mg^{2+}	stevensite, saponite
			Zn^{2+}	sauconite
			Mg^{2+}, Li^{+}	hectorite
		di	Al^{3+}	montmorillonite, beidellite
			Fe^{3+}	nontronite
			Cr^{3+}	volkonskite
0.6 – 0.9	vermiculite	tri	Mg^{2+}	(trioctahedral) vermiculite
		di	Al^{3+}	dioctahedral vermiculite
0.6 – 0.9	illites	di	Al^{3+}	
		tri(?)		
0.6 – 1	mica	tri	Mg^{2+}	phlogopite
			Mg^{2+}, Fe^{2+}, Fe^{3+}	biotite
			Li^{+}, Al^{3+}	lepidolite
		di	Al^{3+}	muscovite, illite
			Fe^{3+}, Al^{3+}	glauconite
			Fe^{3+}, Al^{3+}, Mg^{2+}, Fe^{2+}	celadonite
ca. 1.8 – 2	brittle mica	tri		clintonite
		di		margarite
Variable	chlorite	tri	Mg^{2+}, Al^{3+}	clinochlore
		di	Al^{3+}	donbasseite
		di – tri [e]	Al^{3+}, Mg^{2+}	sudoite

[a] Layer charge (equivalents/unit), see general formula in page 11;
[b] Tri- or dioctahedral;
[c] Exact compositions see [31, Chap. 2];
[d] Nonplanar phyllosilicates;
[e] Interlayer: dioctahedral, 2: 1 layer: trioctahedral.

the chrysotile crystals. This results in the formation of tubes (tubular morphology of asbestos) or in the alternation of the terminal oxygen ions in a bandlike manner as in antigorite (Fig. 12 B).

The major mineral of this group is kaolinite, the most important mineral in kaolins. Total world production was ca. 25×10^6 t in 1992 (US Bureau of Mines, 1992) (→ Clays). It is used in the ceramic industry (ca. 20 %) and as a filler and coating pigment in paper production (ca. 50 %) [31, 34]. Kaolin is also needed for the production of zeolites (→ Zeolites) and ultramarines (→ Pigments, Inorganic, Chap. 3.5.).

2: 1 Clay Minerals. 2 : 1 clay minerals consist of two zweier single layers ("tetrahedral sheets") which enclose a central octahedral sheet (Fig. 13 B). The diversity of this group of phyllosilicates (Table 4) results from:

1) The occurrence of di- and trioctahedral layers (Fig. 11 B, C)
2) Isomorphous substitutions in the octahedral sheet
3) Isomorphous substitutions in the tetrahedral sheet, mainly Al^{3+} for Si^{4+}

In addition, cationic sites can be vacant, and protons may dissociate from OH^- ions or associate to them or to O^{2-} ions. Depending on the type and concentration of the substitutions and defects, the layers can assume negative charges. The number of charges per formula unit ($[(Si,Al)_4O_{10}]$ unit) is the most essential parameter used to characterize the 2 : 1 clay minerals.

The chemical composition of the 2 : 1 clay minerals is represented by:

$$\underbrace{M^{\nu+}{}_{\xi/\gamma}\,(H_2O)_n}_{\text{Interlayer space}}$$

$$\underbrace{\{(M^{3+}, M^{2+}, M^{+})^{(6-x)+}_{2-3}[Si_{4-y}Al_yO_{10}(OH)_2]\}^{\xi-}}_{\text{Layer}}$$

where ν is the valency of the interlayer cation and ξ is the layer charge.

The general formula indicates that the number of octahedral cations (M^{3+}, M^{2+}, M^{+}) varies between 2 and 3. Mostly, the occupancy is near 2 (dioctahedral minerals) or 3 (trioctahedral minerals). An extensive survey of chemical compositions is given in [31].

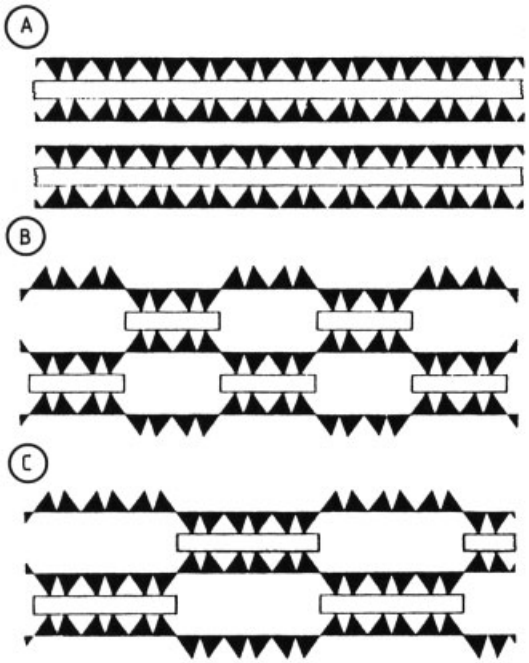

Figure 14. Structural relations between some silicates
A) Talc, $Mg_3[Si_2O_5]_2(OH)_2$; B) Palygorskite, idealized $Mg_5[Si_2O_5]_4(OH)_2(OH_2)_4 \cdot 4\,H_2O$, for exact formula see [29]; C) Sepiolite, idealized $Mg_4[Si_2O_5]_3(OH)_2 \cdot 4\,H_2O$, for exact formula see [29]

In Table 4, the clay minerals are classified according to their layer charge ξ and their di- or trioctahedral character. The silicates with uncharged layers are pyrophyllite (idealized $Al_2[Si_4O_{10}(OH)_2]$) and talc ($Mg_3[Si_4O_{10}(OH)_2]$). *Talc* (Fig. 14 A) is a raw material in the ceramic industry (→ Talc). It is used in the production of insulating materials (→ Insulation, Electric, Chap. 3.1.2.2.), as a filler (paper, paints, cosmetics, elastomers, thermoplastics), and as a carrier material for pesticides. *Pyrophyllite* finds similar uses [31].

Montmorillonite is the most important smectitic mineral (Table 4) of the bentonite rocks (→ Clays). The various applications of bentonites (production ca. 8×10^6 t in 1992; US Bureau of Mines, 1992) are based on the special properties of montmorillonite, in particular its fine particle size ($< 2\,\mu m$), intracrystalline reactivity (swelling in water and organic liquids, ion-exchange properties, see Section 1.4.6), binding of organic materials and colloidal behavior [31]. Only a few of the many hundreds of applications (→ Clays, Chap. 5.2.) [31, 35] are mentioned here. The special rheological properties of bentonites are exploited in the application as a foundry sand binder, for oil well drilling, iron ore pelletizing, in the building industry (in particular for subterranean curtains), in geotechnics, and as sealing materials; for pelletizing animal feed, as fillers and adsorbents (e.g., in clarifying and decolorizing mineral, animal and vegetable oils, beer, and fruit juices) and as a carrier material for pesticides. The profitable use as pet litter should not be overlooked.

Synthetic smectites are hectorite-type materials which are produced more easily than montmorillonites under mild hydrothermal conditions. Hectorite is a very low-charged smectite ($\xi \approx 0.20 - 0.25$ eq/unit) that contains Li^+ ions in the octahedral sheet. It has similar uses to bentonite [36]. The synthetic materials are used in the pharmaceutical industry (e.g., as thickeners, thixotropic agents, gel-forming materials) and seem to provide suitable properties as components in muds for drilling very deep wells [31, 35].

The largest amounts of *vermiculite* (production about 0.65×10^6 t/a) are used in the form of light, expanded aggregates consisting of exfoliated crystals (see Section 2.9).

Illites are very fine, mica-type minerals, and essential components of soils.

Mica is used in the form of sheets, ground powders, or micronized and is described in detail elsewhere (→ Mica). Most synthetic products are fluorinated micas. Formation of fluorophlogopite in glass ceramics gives this mate-

rial a precision machinability with conventional metal-working tools (→ Glass Ceramics) [37].

Chlorites. Chlorites are composed of silicate layers (similar to the 2 : 1 clay minerals mentioned above) that are separated by metal hydroxide layers $M(OH)_2$ or $M(OH)_3$ [29]. Chlorites are not used on an industrial scale but are abundant in soils and are found in clays.

Mixed-Layer Minerals. Mixed-layer minerals are very abundant in common clays and soils. The crystals are composed of layers which either differ in the composition and layer charge or are of different structural types. In most cases, the sequence of the layers is random. Typical combinations are [31] illite – smectite, mica – vermiculite, mica – chlorite, smectite – chlorite, chlorite – vermiculite, and kaolinite – smectite.

Some mixed-layer minerals also have a regular sequence of layers [31]:

Rectorite = muscovite (or paragonite) – montmorillonite
Tosudite = dioctahedral chlorite – smectite
Corrensite = chlorite – vermiculite (or trioctahedral smectite)

Palygorskite and Sepiolite. Palygorskite (Fig. 14 B) and sepiolite (Fig. 14 C) provide further examples of phyllosilicates with single layers of different directedness. The two-dimensional octahedral sheet is split into one-dimensional bands comprising eight or five octahedra. In the tetrahedral sheet some of the tetrahedra are inverted. Bands of three zweier single chains with different directedness alternate in sepiolite whereas the bands in palygorskite are composed of two zweier single chains. Channels are therefore formed in the structure which contain water molecules and exchangeable cations. The cations are bound as a consequence of isomorphous substitutions in the tetrahedral and octahedral sheets. The minerals have zeolitic properties and show a certain cation-exchange capacity. Sepiolite crystallizes as fibers with diameters of 5 – 30 nm and a length of 10 nm – 4 µm. Fibers of palygorskite are 15 – 30 nm thick and 0.5 – 4.5 µm long.

Sepiolite and palygorskite are being increasingly used as thickeners, thixotropic agents, adsorbents, supports of pesticides and catalysts, and pet litter [31].

1.4.6. Intracrystalline Reactions

Kaolinite, 2 : 1 clay minerals, and alkali phyllosilicates are intracrystalline-reactive, layered materials [38]. The interlayer space is accessible to guest molecules or ions (Fig. 15).

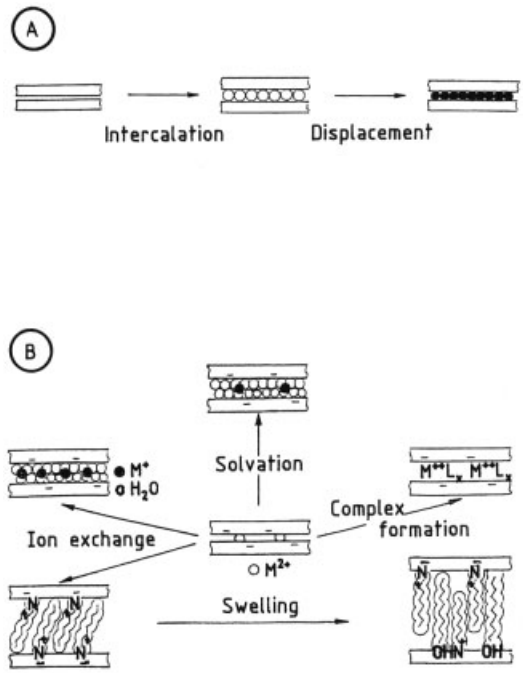

Figure 15. Intracrystalline reactions of phyllosilicates
A) Phyllosilicates with uncharged layers (kaolinite, crystalline silicic acids); B) Phyllosilicates with charged layers (2 : 1 clay minerals, hydrated alkali silicates)

Kaolinite takes up various organic compounds (urea, hydrazine, dimethyl sulfoxide, short- chain acid amides, potassium and ammonium acetate) between the layers so that they are separated by a mono- or bilayer of neutral guest molecules (Fig. 15 A). In most cases, these molecules can be displaced by other molecules [31]. Intercalation of organic compounds promotes delamination of kaolinite particles and can improve the mechanical stability of porcelain. Kaolin modified by intercalation of urea was used in China (mainly in the Sung period, 960 – 1280) in the production of very thin porcelain (egg-shell porcelain) [39].

The most striking difference between the behavior of di- and trioctahedral minerals is that the

trioctahedral 1 : 1 clay minerals such as antigorite do not intercalate guest molecules. Pronounced intercalation capability is observed for the crystalline silicic acids [28,38].

Intracrystalline-reactive phyllosilicates with charged layers can undergo various reactions (Fig. 15 B) [31]. Most important for technical applications is the intracrystalline swelling of smectites by interlamellar adsorption of water. The layer separation of calcium smectites does not normally exceed 1 nm. In the presence of sodium ions, however, the layers separate completely and form colloidal dispersions [31]. This process and the related changes in flow behavior are of great importance in many practical applications (building industry, subterranean curtains, foundry sand binder, oil well drilling fluids etc. [31]).

A further essential reaction is the exchange of the interlayer cations [31]. Soda activation is an important technical process in which interlayer calcium ions are displaced by sodium ions. Fixation of potassium ions by more highly charged 2 : 1 clay minerals plays an important role in soils. Clay minerals are considered as useful adsorbents in environmental technology because of their ability to bind heavy metal ions by cation exchange [31,40,41]. Radioactive nuclides may be removed by clay adsorbents, for instance cesium ions by vermiculites and shales [42]. Even radioactive iodide ions can be adsorbed by modified clay minerals [43].

The third reaction of technical importance is the binding of organic cations. Organo-modified bentonites are used in many applications, mainly as adsorbents, thickeners, or thixotropic agents in organic systems (lubricants, greases, paints, or asphalt). In environmental technology, organic derivatives of clay minerals adsorb toxic substances [31,40,41,44].

1.5. Tectosilicates

Almost all tectosilicates (framework silicates) contain quaternary tetrahedra such that the ratio of tetrahedral cations to oxygen is 1 : 2. The fundamental chain periodicities are $P = 2$, 3, 4, 6: cristobalite, tridymite $P = 2$ (Fig. 16 A); quartz, keatite $P = 3$; zeolites: laumontite, gismondine, harmotone $P = 4$; chabazite, silicalite-1 $P = 6$. Nephelines, $(Na_{1-x}K_x)[AlSiO_4]$, have a tridymite structure (Section 2.2). The feldspars (Section 2.1) have loop-branched dreier frameworks (Fig. 16 B) based on a branched dreier single chain. The black tetrahedra in Figure 16 B form a dreier single chain and additional tetrahedra (stippled) bridge two tetrahedra of this chain [3]. Cordierite, $Mg_2[(Al_4Si_5)O_{18}]$, has an open-branched vierer framework.

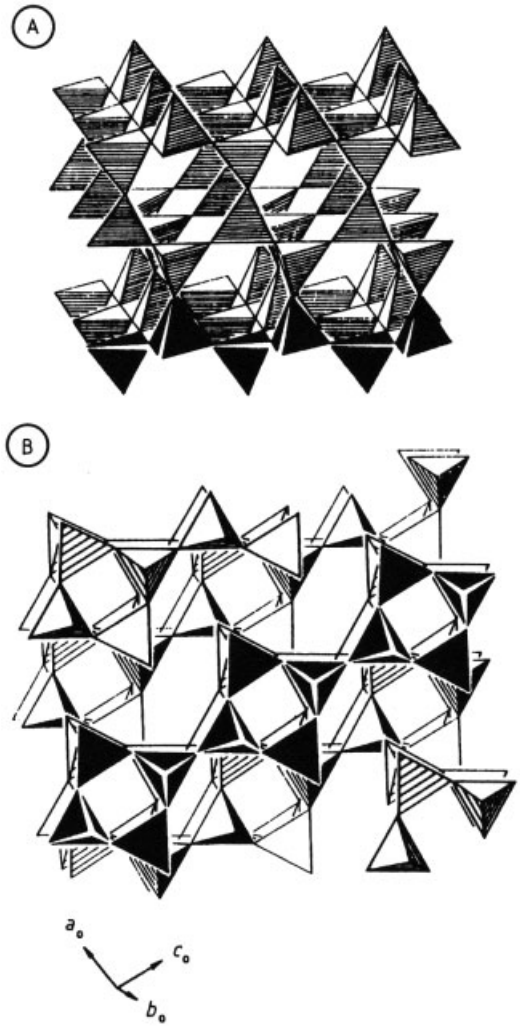

Figure 16. Framework silicates
A) Zweier framework of tridymite $[SiO_2]$; B) Loop-branched dreier framework of feldspars, e.g., orthoclase $K[AlSi_3O_8]$

In Figure 16 the tetrahedral framework is thought to be composed of chains of a certain multiplicity. Another more common view is to

look at the polyhedra which form the framework, and the voids or polyhedral cavities generated between the framework oxygen ions. The notation $[5^{12}]$, for example, indicates a polyhedron composed of 12 pentagonal faces (a pentagon-dodecahedron); $[5^{12}6^2]$ is a polyhedron with 12 pentagons and two hexagons (see Fig. 17). According to the nature of the voids and polyhedral cavities tectosilicates can be divided into three groups [45–47]:

1) *Pyknosiles.* Tectosilicates which have no polyhedral cavities but small interstices between the framework oxygen ions; these voids can be occupied by cations and sometimes water: quartz, cristobalite, tridymite, feldspars.
2) *Clathrasils.* Tectosilicates with polyhedral cavities. The windows of the cages are small and the encaged ions, molecules, or their decomposition products cannot diffuse through them: melanophlogite, dodecasils, sodalite (→ Silica, Chap. 8.2.2.).
3) *Zeolites.* The polyhedral cavities are interconnected by large windows or tunnels, and the enclosed ions or molecules can diffuse through the crystal: analcime, mordenite, faujasite, zeolite A, silicalite (ZSM-5) (→ Zeolites).

Under hydrothermal conditions silica crystallizes from aqueous solutions as quartz, cristobalite, keatite, or coesite (→ Silica) [47,48]. In the presence of inorganic or organic guest compounds, clathrasils are formed. The neutral molecules become enclosed in the polyhedral cavities during synthesis. The guest compounds act as templates for the formation of the polyhedra, thus determining the type of clathrasil [46, 48]. Figure 17 shows polyhedra of melanophlogite and dodecasils occupied by guest molecules. The orientation of the molecules in the cages was derived from single-crystal X-ray diffraction [48].

Minerals of technical importance are the feldspars orthoclase $K[AlSi_3O_8]$, albite $Na[AlSi_3O_8]$, and anorthite $Ca[Al_2Si_2O_8]$ (Section 2.1). The largest amounts of feldspars (and nepheline) are used in the glass and ceramic industries, mainly as fluxes. Cordierite glass ceramics have good thermal stability and shock resistance (→ Glass Ceramics).

The most important synthetic tectosilicates are the zeolites (→ Zeolites) [49,50]. Their Si/Al ratio varies considerably. The most startling development was the synthesis of the almost Al^{3+}-free end members of silica-rich zeolites (ZSM-5 = silicalite-1, ZSM-11 = silicalite-2).

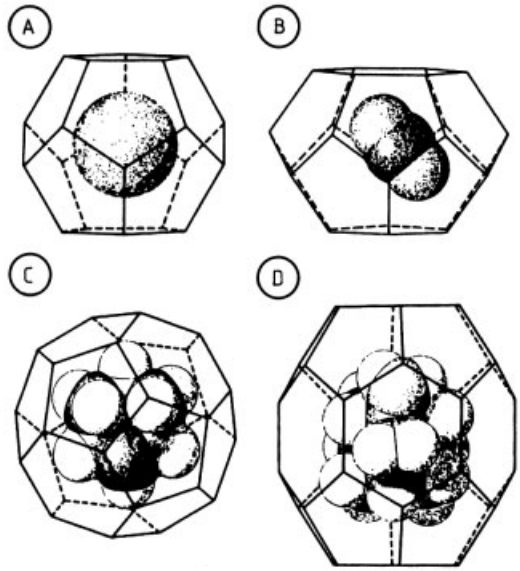

Figure 17. Guest molecules included in cages of the host framework of clathrasils
A) Xenon in the $[5^{12}]$ cage of melanophlogite; B) CO_2 in $[5^{12}6^2]$ of melanophlogite; C) Piperidine in $[5^{12}6^4]$ of dodecasil 3C; D) Adamantylamine in $[5^{12}6^8]$ of dodecasil 1H

2. Natural Silicates

2.1. Feldspar

Feldspars [*68476-25-5*] are anhydrous alkali/alkaline-earth aluminosilicates that closely resemble each other in structure and properties. They are among the most common and important mineral groups, constituting ca. 60 – 65 wt % of the Earth's crust. Approximately two-thirds of them occur as plagioclases.

2.1.1. Structure and Composition

The feldspars are tectosilicates (framework silicates, see Chap. 1). The $[(Si, Al)O_4]$ tetrahedra are linked at all four vertices yielding a

framework $[Si_{4-x}Al_xO_8]^{x-}$. The voids within the tetrahedral framework contain alkali metal or alkaline-earth ions for charge compensation. The structure is composed of four-membered rings, $[(Si, Al)_4O_{12}]$, containing $(Al\,Si_3)$ or (Al_2Si_2) in each ring. The rings are linked by common oxygen atoms (two on each side) in the *a* direction to two neighboring rings to form zigzag bands (or "double crank-shaft chains") making use of three of the four common vertices. These bands are bonded in both the *b* and *c* direction by the fourth common oxygen atom of each (Si, Al) atom (see Fig. 18: in the *b* direction in $y = 0$ or 1/2, in the *c* direction in $y = 0.14$ or 0.36) to form three-dimensional tetrahedral frameworks. As a result, the (010) and (001) planes are rather weakly bonded and readily cleaved. This property is characteristic of all feldspars. The cations K^+, Na^+, and Ca^{2+}, and more rarely Sr^{2+} and Ba^{2+}, occupy the large spaces within the framework of tetrahedra, and are coordinated to oxygen in a fairly irregular manner.

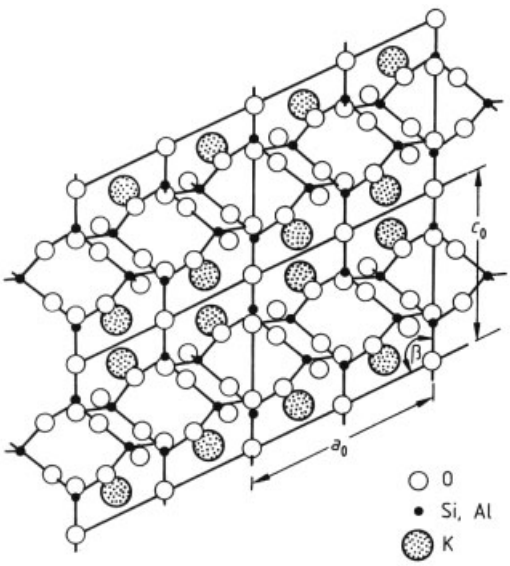

Figure 18. Structure of orthoclase, projection parallel to *b* on (010) showing two crank-shaft bands along a_0

Three main types of feldspar ("molecules") can be distinguished:

1) $K[AlSi_3O_8]$ = potassium feldspar = orthoclase (Or)
2) $Na[AlSi_3O_8]$ = sodium feldspar = albite (Ab)
3) $Ca[Al_2Si_2O_8]$ = calcium feldspar = anorthite (An)

In nature, these three substances seldom occur in a pure state. At high formation temperature (e.g., in some volcanic rocks), a solid solution series between Or and Ab exists: the *alkali feldspars*. On slow cooling [i.e., in most plutonic (igneous) rocks] an immiscibility gap below 650 °C leads to partial unmixing (perthitization). Albite and anorthite form an almost infinite solid solution series: the *plagioclases.* Between orthoclase and anorthite a wide immiscibility gap exists (see Figs. 19, 20, 21).

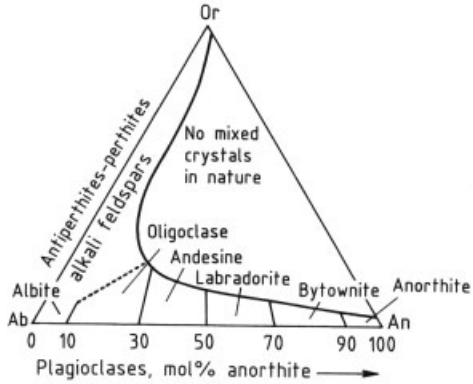

Figure 19. Triangular phase diagram of orthoclase – albite – anorthite showing mixed crystal formation found in nature (at ca. 700 °C)

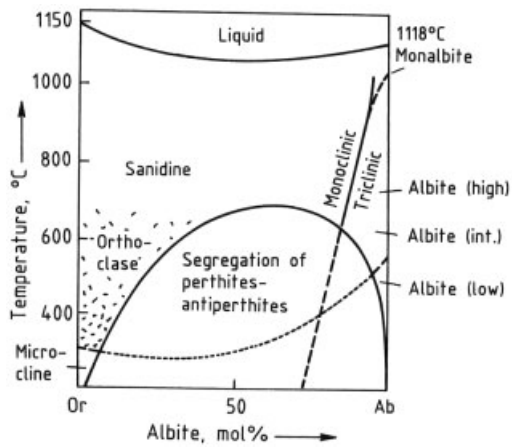

Figure 20. Phase diagram of the alkali feldspars

Pure potassium feldspar has the composition 64.8 wt % SiO_2, 18.3 wt % Al_2O_3 and 16.9 wt % K_2O. Potassium can be replaced by rubidium to a small extent, with the rubidium content in orthoclases from pegmatites reaching ca. 3 wt %. In the plagioclases, the Al_2O_3 content increases on going from the acidic to the basic types: from >19 wt % in albite to almost 37 wt % in anorthite with a concomitant decrease in the SiO_2 content from almost 69 wt % in albite to 43 wt % in anorthite.

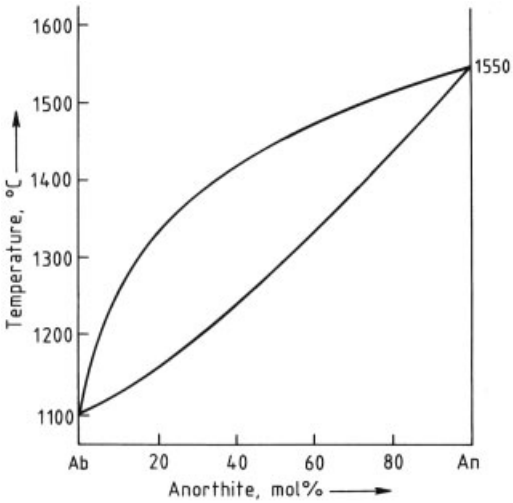

Figure 21. Phase diagram of plagioclases (dry *mp* can be lowered by 500 °C in the presence of water)

All feldspars have a high- and a low-temperature form. In the high form, the distribution of $[SiO_4]$ and $[AlO_4]$ tetrahedra is at random, while in the low form there is a high degree of order. There are also intermediate types with varying degrees of Al/Si order. Ordering changes the monoclinic orthoclase to triclinic microcline. The high forms of albite and anorthite and of albite and orthoclase are completely miscible, whereas the low forms are not.

On slow cooling of the (K, Na) feldspars, segregation (perthite formation) takes place due to selective migration of the small and large cations (Fig. 20). This leads to a regular intergrowth of orthoclase (or microcline), and albite, with characteristic formation of string- or worm-shaped bodies, which can be macroscopic (macroperthite), microscopic (microperthite) or submicroscopic (cryptoperthite). Microcline – perthite consists of microcline containing albite strings, and antiperthite consists of albite with orthoclase strings.

The feldspars are very widely distributed, mainly in igneous rocks (volcanic and plutonic rocks) and metamorphic rocks (contact and regional metamorphites). They also occur in pegmatitic and hydrothermal formations, sedimentary rocks and sediments. This means that the petrographic classification of rocks is largely based on the mass ratios and compositions of the individual feldspars and quartz. If there is a deficiency of SiO_2, leucite $K[AlSi_2O_6]$ (see Section 2.3) or nepheline $Na[AlSiO_4]$ (see Section 2.2) is formed instead of the alkali feldspars.

2.1.2. Characterization of Individual Feldspars

The alkali feldspars can have monoclinic and triclinic crystal forms, but the plagioclases only triclinic forms. The most important feldspars follow.

Potassium Feldspars. The term *sanidine* denotes the metastable, glass-clear monoclinic high form of $K[AlSi_3O_8]$ with a random distribution of (Al, Si) produced by rapid cooling in volcanic rocks (volcanites). During the very slow cooling of plutonic rocks (plutonites) an Al/Si ordering takes place to a greater (triclinic microcline) or lesser (orthoclase) extent. Orthoclase is not visibly triclinized despite prolonged cooling in plutonites. Sanidine melts incongruently with separation of *leucite.*

Sanidine (the high-temperature form of $K[AlSi_3O_8]$) forms monoclinic – prismatic crystals, with space group $C_{2h}^3 - C\,2/m$; lattice constants: $a_0 = 0.856$ nm, $b_0 = 1.303$ nm, $c_0 = 0.718$ nm (a_0, b_0, and c_0 are somewhat smaller when sodium is present); the axial ratio is $a_0 : b_0 : c_0 = 0.657 : 1 : 0.551$; $\beta = 115°59'$; $Z = 4$ (number of chemical formula units per unit cell).

Orthoclase (the intermediate form of $K[AlSi_3O_8]$) forms monoclinic – prismatic crystals, with space group $C_{2h}^3 - C\,2/m$; lattice constants: $a_0 = 0.856$ nm, $b_0 = 1.300$ nm, $c_0 = 0.719$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.659 : 1 : 0.553$; $\beta = 116°01'$; $Z = 4$.

Orthoclase often consists of thick platelets parallel to [010], extended in the [001] direction or having columnar development in the [100] direction. Crystals may be several centimeters or even tens of centimeters in size. Twinning is very characteristic.

Microcline (the low-temperature form of $K[AlSi_3O_8]$) forms triclinic – pinacoidal crystals, with space group $C_i^1 - C\bar{1}$; lattice constants: $a_0 = 0.857$ nm, $b_0 = 1.298$ nm, $c_0 = 0.722$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.660 : 1 : 0.566$; $\alpha = 90°41'$, $\beta = 115°59'$, $\gamma = 87°30'$; $Z = 4$.

Barium Feldspars.

Hyalophane, (K, Ba) $[Al(Al, Si)Si_2O_8]$, forms monoclinic (or pseudomonoclinic?) crystals, with space group $C_{2h}^3 - C\,2/m$; lattice constants: $a_0 = 0.854$ nm, $b_0 = 1.298$ nm, $c_0 = 0.715$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.658 : 1 : 0.551$; $\beta = 115°35'$; $Z = 4$.

Celsian [*001302-50-7*], $Ba[Al_2Si_2O_8]$, forms monoclinic crystals, with space group I 2/c; lattice constants: $a_0 = 0.865$ nm, $b_0 = 1.313$ nm, $c_0 = 2 \times 0.730$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.659 : 1 : 2 \times 0.556$; $\beta = 115°02'$; $Z = 8$.

Plagioclases. The plagioclases form ca. 40 % of the solid crust of the Earth and are widely distributed. At high temperature (see Fig. 21) there is complete miscibility between the end members anorthite, $Ca[Al_2Si_2O_8]$ (basic) and albite, $Na[AlSi_3O_8]$ (acidic). At lower temperature, miscibility is incomplete. The plagioclases are usually twinned along [010], leading to lamellar replication multiple twinning, and typical striation.

The plagioclases have a structure similar to that of sanidine, the replacement of K^+ by the smaller Na^+ and Ca^{2+} ions leading to some deformation with triclinic symmetry. The low-calcium plagioclases [albite (An_{0-10}), oligoclase (An_{10-30}), and andesine (An_{30-50})] have a cell size similar to that of sanidine, with $C\bar{1}$ symmetry. The high-calcium plagioclases (labradorite (An_{50-70}), bytownite (An_{70-90}), and anorthite (An_{90-100})) show doubling of the c_0 lattice constant, so that bytownite has $I\bar{1}$ symmetry and anorthite $P\bar{1}$ symmetry. The plagioclases are usually characterized by the content (in mol %) of anorthite (An), albite (Ab), and orthoclase, whereby the content of orthoclase is usually negligibly small. An important characteristic of plagioclases is the degree of order, i.e., not only of (Al, Si), but also of (Na, Ca) and of (Na, Ca, K) in high-temperature mixed crystals containing potassium. The high- and low-temperature plagioclases can be distinguished optically. In the low-temperature plagioclases, there is an ordered distribution of (Al, Si). For example, in pure anorthite, which has no high-temperature form, the $[AlO_4]$ and $[SiO_4]$ tetrahedra alternate, whereby each oxygen is neighbored by an aluminum and a silicon ion. Despite the Al/Si disorder in high-temperature plagioclases, Al – O – Al groupings do not occur. Also, because the (Na, Ca) ions are smaller, the triclinic form is maintained at higher temperature, unlike sanidine. The triclinic high-temperature form of albite is also known as *analbite*. *Monalbite* is a monoclinic potassium-containing high-temperature phase of albite.

Albite, $Na[AlSi_3O_8]$, An_{0-10} or $Ab_{100}An_{00} - Ab_{90}An_{10}$; $C\bar{1}$ lattice constants: $a_0 = 0.814$ nm, $b_0 = 1.2789$ nm, $c_0 = 0.716$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.636 : 1 : 0.559$; $\alpha = 94°19'$, $\beta = 116°34'$, $\gamma = 87°39'$; $Z = 4$.

Oligoclase, An_{10-30} or $Ab_{90}An_{10} - Ab_{70}An_{30}$; $C\bar{1}$ lattice constants: $a_0 = 0.817$ nm, $b_0 = 1.284$ nm, $c_0 = 0.713$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.636 : 1 : 0.556$, $\alpha = 93°49'$, $\beta = 116°27'$, $\gamma = 88°59'$; $Z = 4$.

Andesine, An_{30-50} or $Ab_{70}An_{30} - Ab_{50}An_{50}$; $C\bar{1}$ lattice constants: $a_0 = 0.818$ nm, $b_0 = 1.288$ nm, $c_0 = 0.711$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.635 : 1 : 0.552$; $\alpha = 93°24'$, $\beta = 116°10'$, $\gamma = 90°24'$; $Z = 4$.

Labradorite, An_{50-70} or $Ab_{50}An_{50} - Ab_{30}An_{70}$; $C\bar{1}$ lattice constants; $a_0 = 0.816$ nm, $b_0 = 1.286$ nm, $c_0 = 2 \times 0.710$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.635 : 1 : 2 \times 0.552$; $\alpha = 93°34'$, $\beta = 116°06'$, $\gamma = 89°47'$; $Z = 8$.

Bytownite, An_{70-90} or $Ab_{30}An_{70} - Ab_{10}An_{90}$; $I\bar{1}$ lattice constants: $a_0 = 0.817$ nm, $b_0 = 1.287$ nm, $c_0 = 1.418$ nm $= 2 \times 0.709$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.635 : 1 : 1.102$; $\alpha = 93°22'$, $\beta = 115°58'$, $\gamma = 90°31'$; $Z = 8$.

Anorthite, $Ca[Al_2Si_2O_8]$, An_{90-100} or $Ab_{10}An_{90} - Ab_0An_{100}$; $P\bar{1}$ lattice constants: $a_0 = 0.818$ nm, $b_0 = 1.288$ nm, $c_0 = 1.417$ nm $= 2 \times 0.785$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.635 : 1 : 2 \times 0.550$; $\alpha = 93°10'$, $\beta = 115°51'$, $\gamma = 91°13'$; $Z = 8$.

2.1.3. Production

Feldspar can be synthesized by solid-state or liquid-phase reactions:

1) *Albite* is formed by heating a mixture of 1 mol Al_2O_3, 6 mol SiO_2, an excess of Na_2O, and sodium tungstate as a mineralizer at 900–1000 °C with a long reaction time (days).
2) *Orthoclase* is formed by heating a mixture of 1 mol Al_2O_3, 6 mol SiO_2, and a large excess of K_2O, with tungstic acid for mineralization at 900 – 1000 °C, or by heating a mixture of Al_2O_3, SiO_2 and acidic potassium tungstate at 900 °C. Crystals are obtained on treatment with boiling water.
3) *Anorthite* is obtained by fusing a mixture of Al_2O_3, SiO_2 and $CaCO_3$. The temperature is maintained slightly below the melting point ($\leq$ 1500 °C), and the mixture is slowly cooled. Four modifications (α, β, γ, and δ) are known.

Preparative methods using hydrothermal reactions in autoclaves or high-pressure reactors are also known:

1) Albite is synthesized by heating precipitated aluminum silicate and excess sodium silicate at ca. 500 °C
2) Orthoclase is synthesized by heating a mixture of Al_2O_3, SiO_2 and KOH with water in a ratio corresponding to $K_2O \cdot Al_2O_3 \cdot 6\,SiO_2$ at 400 °C, or from an aqueous slurry of potassium aluminate or $Al(OH)_3$ with potassium silicate in the presence of CO_2 at 360–400 °C
3) Anorthite is synthesized by heating a mixture of aqueous slurries of SiO_2, Al_2O_3, and CaO in the presence of $CaCl_2$ as a mineralizer at 470 °C

All synthetic feldspars are Al/Si disordered. Ordering requires geologic periods of time. Feldspars also occur in industrial products (ceramics). In synthetic feldspars, Si^{4+} can be completely replaced by Ge^{4+}, and Al^{3+} by Ga^{3+} or Fe^{3+}. This gives new end members that are only found as synthetic compounds. For example, δ-anorthite, δ-$Ca[Al_2Si_2O_8]$, which corresponds to natural anorthite, occurs in fly ash and sometimes in blast-furnace slags. It is formed when CaO- containing slags attack the chamotte (fire bricks) in blast furnaces and lime kilns.

The high-temperature form of albite, α-$Na[AlSi_3O_8]$, occurs in chamottes that have been exposed to Na_2O, e.g., in tank linings and checker bricks used in furnaces for glass-making.

α-Celsian, α-$Ba[Al_2Si_2O_8]$, is the principal component of special bricks used to line the arched roofs of the firing zones in electric tunnel furnaces for glazing stoneware and sanitary ware. It is very resistant to vapors from lead and borate glazes. Other refractory materials react with these vapors to form liquid phases.

2.1.4. Properties

Physical Properties. Feldspars can be glass clear, transparent, translucent, cloudy, matt, colorless, white, gray, or colored (e.g., greenish, reddish, red, or brown) due to inclusions. Labradorite and orthoclase can be iridescent (labradorescent) and are used for decorative purposes. The feldspars can have a glassy or sometimes nacreous luster. Fused feldspars, even if originally dark in color, usually solidify as a transparent, colorless glass.

Cleavage in the (001) plane is very good, moderate in the (010) plane, and just observable in the (110) plane [plagioclases also cleave in the (110) plane]. Orthoclase is named from the cleavages in the mutually perpendicular (001) and (010) planes. The plagioclases form triclinic crystals and owe their name to the fact that the angle between the (001) and (010) planes deviates by a few degrees from 90°. The fracture is conchoidal, uneven, and splintered or brittle.

The *Mohs hardness* of albite is 6 – 6 1/2, and of anorthite 6 – 7. *Densities* (g/cm^3) are: orthoclase 2.53 – 2.56, albite ca. 2.62, and anorthite ca. 2.76.

In thin sections, the feldspars show low refraction of light and birefringence (double refraction).

The average *specific heat* (J g^{-1} K^{-1}) of orthoclase between 0 and 100 °C is ca. 0.79, of albite 0.82, and of anorthite 0.80. The corresponding values for the range 0 – 1100 °C are 1.05, 1.1, and 1.12.

In the crystalline state the alkali feldspars become *electrically conductive* at moderate temperatures. *Ionization* is observed above 900 °C.

In the liquid phase, albite is one of the most strongly dissociated silicates.

The rate of *crystallization* in alkali feldspar melts is extremely small due to their high melt viscosity. The fusion range of natural orthoclase is 1140 – 1300 °C (incongruent melting, see also Section 2.3). Fused anorthite has a comparatively low viscosity, however, and crystallizes readily and rapidly on cooling to form large, well-formed crystals. The maximum crystallization rate occurs 20 – 25 °C below the temperature at which melting starts. A glass is formed only if the cooling rate is very rapid.

Chemical Properties. Chemical attack, including weathering, readily decomposes or changes the feldspars. The prolonged action of water on pulverized feldspar leads to preferential dissolution of the alkali ions, but also dissolves CaO and Al_2O_3, eventually giving a residue of polysilicic acids. Water containing dissolved CO_2 is more aggressive; alkali metals and Al_2O_3 are dissolved preferentially, and the reaction proceeds with formation of kaolinite ($Al_2O_3 \cdot 2\,SiO_2 \cdot 2\,H_2O$).

Strong mineral acids decompose alkali feldspars slowly, producing a residue of polysilicic acids. However, powdered anorthite dissolves completely in hydrochloric acid at 100 °C. Under high pressure (i.e., in autoclaves at ca. 200 bar) potassium feldspars are converted to kaolinite by 5 % hydrofluoric acid at 225 °C or by 0.5 – 1 mol/L HCl at 320 – 330 °C. This first causes complete hydrolysis of the starting material, followed by the formation of kaolinite and silicic acid. At lower temperatures, the reaction only produces a residue with a high silicic acid content. Anorthite is almost completely converted into kaolinite by treatment with 0.5 mol/L HCl at 340 °C.

Concentrated solutions of alkali hydroxide decompose the alkali feldspars completely above 110 °C. Reaction with 1 – 2 % solutions of alkali hydroxide leads to the formation of soluble products.

2.1.5. Mineral Deposits and their Extraction

The most important raw materials are the potassium feldspars (orthoclase, microcline), followed by sodium feldspar (albite). The other members of the plagioclase group are of very little value, although rocks with a high plagioclase content has been tested as a raw material for aluminum production. Nepheline syenite is industrially important (see Section 2.2), as are other rocks and sediments rich in feldspar (e.g., granites, syenites, and porphyries). Feldspar also occurs in some apatite deposits (e.g., in the Kola Peninsula, Russia). Furthermore, feldspar is extracted from sedimentary rocks (e.g., arkoses).

Economically useful deposits of high-grade feldspars are of limited occurrence. They are mostly plutonic rocks, pegmatites, and volcanic rocks with feldspar inclusions. When exposed, feldspathic rocks are often partially or completely kaolinized. Feldspar often arises as a byproduct during the processing of kaolin or mica. World production of feldspar is ca. 6.7×10^6 t/a (all weights specified are in metric tons).

The demands of European industry can be met from European deposits, the most important being in Italy, Turkey, Germany, France, Spain, the Czech Republic, Portugal, Russia, Norway, Sweden, Poland, Hungary, Finland and Greece.

The most important producers are Germany: Amberger Kaolinwerke, Hirschau; Birkenfelder Feldspatwerke Schweyer & Vollmer, Ellweiler; Gebrüder Dorfner, Hirschau; Kali-Chemie, Hannover; Keramische Rohstoffwerke Bauscher-Mandt, Weiden; Villeroy & Boch, Keramische Werke, Mettlach; Italy: Gruppo Maffei SpA., Milano; Minerali Industriali SpA.; France: Denain Anzin Minéraux (Feldspaths du Midi, Feldspaths du Morvan, Feldspaths Baux); Norway: Norfloat A/S; Spain: Unimin; Finland: Partek; Sweden: Svenska Forshammar AB, Gothenburg; Spain: Industrias del Cuarzo and Lansa; Turkey: Esan; Elginkam; former Soviet Union; United States: The Feldspar Corp., Spruce Pine/NC.; IMC Chemical Group Inc., Industrial Minerals Div., Boston/Mass.; Foote Minerals Co., Exton/Pa., Kings Mountain; Mica Co. Inc., Kings Mountain/NC.; Pacer Corp., Custer/SD.; San Antonio Mica Mine, Ajo/Ariz.; Mexico: Materias Primas Monterrey SA, Monterrey/N.L.; Argentina: Piedra Grande SA, San Luis; Japan: Kinsei Kogyo Co Ltd., Osaka; Morimura Bros. Inc., Tokyo; Nissho-Iwai Co. Ltd., Tokyo; Sumitomo Corp., Tokyo; India: Golecha Palawat & Co., Rajastan; Shree Modi Levigated Kaolin Pvt. Ltd., Rajastan; South Africa: Ceramic Minerals (Pty.) Ltd.,

Dunswart/Transvaal; General Overseas Traders S.A. (Pty.) Ltd., Johannesburg; Martin & Robson (Pty.) Ltd., Germiston; Otavi Mining Co. (Pty.) Ltd., Johannesburg; Rand London Corp. Ltd., Johannesburg; Rondebult Clayworks (Pty.) Ltd., Boksburg.

Opencast mining is used to extract sedimentary and pegmatitic deposits, although pegmatite is also obtained by underground mining. The usual methods such as drilling, explosives or cutting are used for massive rock deposits. Scrapers and front loaders are used for sedimentary deposits (loose rock), working in stepped faces.

2.1.6. Processing and Quality Requirements

Processing. Usually, processing of feldspar consists of crushing and/or grinding with subsequent classification, but some feldspar is also processed by flotation (e.g., feldspar in granitic rocks). The methods used to process the massive or loose rock depend on the end product, the mineral composition and the available equipment. In the simplest case, the feldspathic rock is progressively broken down by jaw crushers, ball mills, roller mills or pin mills; classified by vibratory sieves; washed if necessary; and dried. It is then ground using impact mills or ball mills to produce a coarse powder and classified by air elutriation. The final product contains not only feldspar, but also accompanying minerals (e.g., quartz or mica). Marketable products should have a favorable mineral composition with a high feldspar content. A quartz content of the product is not detrimental, since quartz is needed in any case for many applications (e.g., production of porcelain, glass and enamel). Any mica can be removed by air classification after breakdown of the rock. Accompanying minerals with a high magnetic susceptibility can be removed by magnetic separation. The content of heavy metal ions in the end product can be reduced to a safe level by this method.

The flotation process was introduced for processing feldspar in 1960 by the Finnish company Lohja Oy. The rock is first crushed, sometimes washed and dried and then ground. Mica, quartz, other accompanying minerals and feldspars are obtained in the froth or as a sediment in a series of flotation stages. Selectivity depends on the choice and dosage of reagents (activators, collectors, and foaming agents) and adjustment of the pH. Albite and potassium feldspar can also be separated from each other if necessary.

Feldspar also arises as a byproduct in the wet processing of kaolin. Quartz is first removed from an aqueous suspension of the crude ore. Kaolinite and feldspar are then separated in hydrocyclones, dewatered, and dried with rotary or tray dryers (→ Clays, Chap. 4.1.2.). The degree of separation (feldspar : quartz ratio) is ca. 9 %.

In another process, the kaolinite and accompanying clay minerals are slurried, and the feldspar – quartz concentrate (particle size 0.1 – 0.5 mm) is separated into feldspar and quartz by an electric field of 40 kV. The degree of separation (feldspar : quartz ratio) is ca. 96 %. High-intensity magnetic separators are used to remove paramagnetic impurities. Grinding is performed in nonferrous equipment and is followed by air elutriation, giving products with a maximum grain size of 40 – 150 μm.

Quality requirements vary considerably and depend on the end product consumer. High-quality products generally require high-purity raw materials of defined composition. For white glasses, light-colored ceramics, glazes and enamels, the content of heavy metal ions must be extremely small (≤ 0.08 wt %). Contents of Na_2O and/or K_2O and Al_2O_3 must be specified to ensure satisfactory physical and chemical properties, especially strengths of glasses. When the products are used as fluxes, the alkali oxide content must be increased to bring about the desired reduction of the melting point. Important quality criteria for some applications are the particle size, the particle size distribution, and the tolerance range for the fractions of over- and undersized particles.

Chemical and physical data of some feldspar grades are given in Table 5; some particle size distributions for feldspar sands and powders are given in Tables 6.

2.1.7. Uses and Economic Aspects

The glass and ceramic industries are the main consumers of sodium and potassium feldspars and also of related minerals of lower quality

Table 5. Chemical composition and physical data for selected feldspar grades

	Theoretical		Feldspar sand, AKW [a]			Glass	Norfloat [c]		Forshammar [d]	Maffei [ed]	
			FS 800 S	FS 900 L	FS 960 L	FFF [b]	K	Na		F 5	F 7 SE
Total feldspar content, wt %	100	100	81	94	92	90	96.5	86.5	67	80	90
K feldspar, wt %	100		76	87	43	45	65.5	16.5	41		
Na feldspar, wt %		100	5	7	43	42	28.5	61	25	80	90
Ca feldspar, wt %						3	2.5	9	1		
Quartz, wt %			12	4	6	5	3	12	29	12 - 15	5
Other silicates			7	2	2	2	0.5	1	4	5 - 8	5
SiO_2, wt %	64.8	68.7	67.5	65.8	67.7	67.5	65.4	69.2	75.5	70.9	70.10
Al_2O_3, wt %	18.2	19.4	17.0	17.8	18.2	18.5	18.7	18.7	14.7	17.52	18.20
Na_2O		11.8	0.60	0.80	5.00	5.00	3.36	7.20	4.85	9.30	10.20
K_2O, wt %	16.9		12.70	14.65	7.20	8.12	11.10	2.80	4.25	0.65	0.30
CaO, wt %			0.02	0.04	1.2	0.56	0.51	1.82	0.20	0.60	0.60
MgO, wt %			0.04	0.02	0.03				0.10	0.10	0.10
Fe_2O_3, wt %			0.12	0.042	0.08	0.10	0.06	0.11	0.13	0.30	0.10
TiO_2, wt %			0.50	0.044	0.03					0.03	0.05
Loss on ignition, wt %			1.0	0.25	0.20		0.29	0.19	0.45	0.75	0.50
Hardness (Mohs)			6	6	6			6		$6\frac{1}{2}$	$6\frac{1}{2}$
Density, g/cm^3			2.6	2.6	2.6			2.6		2.65	2.65
Begin of sintering, °C			1130	1175	1150	1140		1140			1135
End of sintering, °C			1340	1365	1250			1270			1270
Sintering interval, K			210	190	100			130			135
Begin of fluxing, °C			1430	1440	1330			1290			1300
Softening interval, K			90	75	80			20			30
Half-sphere point, °C			1560	1570	1420	1460		1330			1365
End of fluxing, °C			1630	1600	1580			1410			1500
Fluxing interval, K			200	160	250			120			200

[a] AKW Amberger Kaolinwerke, Hirschau, Germany. [b] Glass-grade feldspar, floated, Lohja Oy, Virkkala, Finland. [c] H. Björum, K/S Norfloat, Oslo, Norway. [d] Svenska Forshammars AB, Gothenburg, Sweden. [e] Gruppo Maffei S.p.A., Monte Orno, Milan, Italy.

Table 6. Typical particle size distributions (in %) for selected feldspar sand grades

Screen analysis	Grading				
DIN 66 165, mm	FS 800 S	FS 900 SF	FS 900S	FS 900L	0305
> 1.0				0.10	0.20
0.5 – 1.0	0.05	0.05	0.20	30	65
0.1 – 0.5	18	21	85	70	35
0.063 – 0.1	25	29	11	0.10	
0.040 – 0.063	22	23	3		
< 0.04	35	27	1		

Dry screen residue	Grind number			
DIN 53 734, µm	3	6	8	10
> 125	6			
> 100	12			
> 90	18	5		
> 71	25	10	3	
> 63	34	18	8	3
> 40	48	35	20	10

with respect to heavy metal ions and accompanying minerals (i.e., feldspathic pegmatites, aplites, phenolites and Cornish stone). See also → Glass, Chap. 5.1., → Ceramics, General Survey, Chap. 2.4. The alkali feldspars have a fluxing action due to their alkali-metal oxide content. They reduce the temperature at which melting begins (i.e., the sintering point of the raw material mixture) and thus shorten the time required for melting or firing. This reduces energy consumption and increases production capacity and the life of the refractory linings. On cooling the melt, the oxidic components of the feldspar form a glassy structure, which provides the glassy matrix in ceramic products.

Al_2O_3 increases melt viscosity, facilitating mechanical forming of the glass. It also increases hardness and impact strength (toughness) and prevents crystallization (devitrification). In ceramic masses, the Al_2O_3 in the glassy matrix reduces the tendency to fire cracking and increases impact strength.

Sodium feldspar is usually added at a level of 0.2 – 17 wt % as a raw material for glass. Potassium feldspar is generally employed for ceramic products, although sometimes calcium or calcium/sodium feldspar is also used.

Feldspar (admixtures of 4 – 20 wt %) is used in the coating for arc-welding electrodes to improve the flow of the molten slag. In blast furnaces, the addition of feldspar also improves slag flow during tapping.

Powdered feldspar is used as a mild abrasive due to its hardness, conchoidal fracture and platelet-shaped particles. Light- colored alkali feldspars, anorthite and plagioclase are added to coatings in the form of fine sands or coarse powders to give antiskid properties to floor coverings and surfaces intended for vehicles and pedestrians.

Finely ground feldspar grades are good fillers for exterior paints because of their chemical inertness, lack of photoreactivity and good dry brightness. As a filler feldpsar competes with other, often cheaper, raw materials such as calcite and dolomite. They are also used as fillers for pigments, adhesives, rubber (especially foam rubber), thermoplasts and thermosets. They can be used as fillers in large amounts approaching the critical pigment volume concentration (CPVC), giving transparent or nearly transparent paints and polymer products, because the refractive indices of many paint resins, plasticizers and polymers are almost the same as those of the feldspars ($n \approx 1.55$).

Feldspathic rocks that can be dressed are also used as building stone, as hard core in roads, and as bright, reflective chippings in bitumen-bonded road surfaces.

In Western Europe, the approximated total consumption of feldspar is ca. 3.7×10^6 t/a and is divided as follows:

Ceramics	57 %
Glazes and enamels	4.5 %
Glass	34 %
Abrasives	1.5 %
Fillers	0.4 %
Others	2.6 %

Alkali feldspars can be replaced completely or partially by leucite (Section 2.3), nepheline or nepheline syenite (Section 2.2), thereby increasing the alkali oxide and Al_2O_3 content and reducing the SiO_2 content. If low levels of accompanying minerals and heavy metal ions are unnecessary, other substitutes can be used: feldspathic pegmatites, aplites, phonolites and porphyries; and to a limited extent plagioclase, anorthosite (plutonic rocks containing ca. 90 % plagioclases), Cornish stone (though this can liberate fluorine on thermal decomposition of the melts) and the blast furnace slags "Carolina stone" and "Calumite" (both consisting of kaolinite, feldspar, quartz and some fluorspar).

Worldwide extraction of feldspars is estimated at $> 6.7 \times 10^6$ t/a. Together with nepheline, nepheline syenite, and leucite, total production exceeds 7.5×10^6 t. In 1990, 247 000 t feldspar and 135 000 t pegmatite were produced in the Federal Republic of Germany.

The production figures of pegmatite in 1988 – 1993 in the United States are ca. 670 000 t/a, in the CIS 310 000 t/a, in Italy ca. 295 000 t/a, in France ca. 195 000 t/a, in Mexico ca. 127 000 t/a, in Brazil ca. 110 000 t/a, in Spain ca. 90 000 t/a, in Turkey ca. 73 000 t/a, in Norway ca. 71 000 t/a, in Venezuela ca. 70 000 t/a, in Finland ca. 68 000 t/a, in Rumania ca. 60 000 t/a, in Sweden ca. 52 000 t/a, in India ca. 50 000 t/a, in Argentina and in the Republic of South Africa ca. 47 000 t/a, respectively.

The total feldspar mining production in 1994 amounted to 1 806 935 t in Italy, 765 202 t in the United States, 528 027 t in Thailand, 500 000 in Turkey, 385 000 t in France, 379 427 t in Ger-

many, 319 658 t in South Korea, 274 789 t in Venezuela, 250 000 t in Spain, 170 000 t in the Czech Republic, 140 000 t in Brazil, 133 400 t in Mexico, 92 440 t in Portugal, 79 000 in Iran, 76 188 t in Colombia, 70 000 t in Russia, 70 000 t in India, 62 905 t in Norway, 55 000 t in Sweden, 53 503 t in Japan, 50 000 in Argentina, 46 000 t in Poland, 43 805 t in the Philippines, 43 600 t in Hungary, 41 389 t in Finland, 40 000 t in Cuba, 39 745 t in Egypt, 37 156 t in the Republic of South Africa, and 35 000 t in Greece. Furthermore, the feldspar mining production amounted in 1994 e.g., to 23 500 t in Australia, to 8000 t in the United Kingdom, and 4883 t in Austria.

2.2. Nepheline and Related Compounds

The alkali aluminosilicate nepheline, $KNa_3[AlSiO_4]_4$, a feldspathoid, belongs to the nepheline group of tectosilicates without nontetrahedral anions. The aluminum : silicon ratio is 1 : 1.

Structure and Mineralogy. As in all tectosilicates, the oxygen ions at the vertices of the $[AlO_4]$ and $[SiO_4]$ tetrahedra in nepheline are linked to the four neighboring tetrahedra. This produces a three-dimensional open framework in which the relatively large Na^+ and K^+ cations are located in the spaces between the tetrahedra. In tectosilicates, these spaces can be occupied by alkaline-earth ions, nontetrahedral anions (e.g., in the feldspathoids sodalite and scapolite), or water (in zeolites). In minerals of the nepheline group the tetrahedra exhibit a hexagonal or pseudohexagonal arrangement. In nepheline (Fig. 22), alternating $[AlO_4]$ and $[SiO_4]$ tetrahedra are linked together at common vertices to form an easily distorted high-tridymite structure with six-membered rings. The apices of the $[AlO_4]$ tetrahedra point parallel to the *c* axis, and those of the $[SiO_4]$ tetrahedra point in the opposite direction. Unlike the high-tridymite structure, Si^{4+} ions in nepheline are replaced by Al^{3+} ions in half of the tetrahedral positions. In order to maintain charge neutrality, 3 Na^+ ions and 1 K^+ ion are found per formula unit in the centers of the channels parallel to the *c* axis.

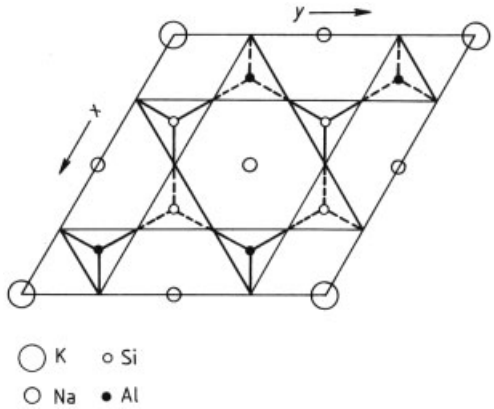

Figure 22. Idealized structure of nepheline, $KNa_3[AlSiO_4]_4$, with $[(Si, Al)O_4]$ tetrahedra linked to form six-membered rings
Projection parallel to *c* on (001)

In natural nepheline, the ratio of K^+ to Na^+ can vary considerably, and Ca^{2+} can also be incorporated in the lattice. There is often an excess of SiO_2, i.e., less than half the Si^{4+} is replaced by Al^{3+}, so that less Na^+ is required to maintain charge neutrality.

Nepheline [*012251-27-3*], $KNa_3[AlSiO_4]_4$, crystallizes in the hexagonal – pyramidal system with space group $C_6^6 - P\,6_3$ and has the following lattice constants: $a_0 = 1.001$ nm, $c_0 = 0.841$ nm; the axial ratio is $c_0/a_0 = 0.840$; $Z = 2$. $\varrho = 2.65$ g/cm^3 (also potassium-free with $a_0 = 0.996$ nm, $c_0 = 0.836$ nm).

The following minerals of the nepheline group have properties similar to those of nepheline:

Kalsilite, $K[AlSiO_4]$ (Fig. 23), crystallizes in the hexagonal – pyramidal system with space group $D_6^6 - P\,6_3$; lattice constants: $a_0 = 0.518$ nm, $c_0 = 0.869$ nm; the axial ratio is $c_0/a_0 = 1.678$; $Z = 2$.

Trikalsilite, $(Na, K)[AlSiO_4]$, crystallizes in the hexagonal system, probably with space group $C_6^6 - P\,6_3$; lattice constants: $a_0 = 1.54$ nm, $c_0 = 0.86$ nm; the axial ratio is $c_0/a_0 = 0.559$; $Z = 18$.

Kaliophilite, $K[AlSiO_4]$, crystallizes in the hexagonal – trapezohedral system with space group $D_6^6 - P\,6_3 22$; lattice constants: $a_0 = 2.706$ nm, $c_0 = 0.861$ nm; the axial ratio is $c_0/a_0 = 0.318$; $Z = 54$; pseudocell: $a_0' = 1.562$ nm $= a_0/3$, $c_0 = 0.861$ nm, $c_0/a_0 = 0.551$; $Z = 18$.

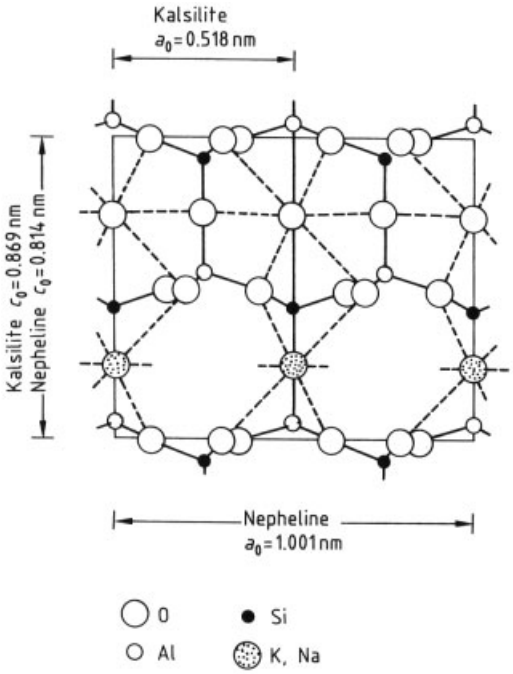

Figure 23. Structure of kalsilite, $K[AlSiO_4]$. The structure of nepheline, $KNa_3[AlSiO_4]_4$, is obtained by replacement of K_4 by KNa_3

High eucryptite, $Li[AlSiO_4]$, only exists as a synthetic product, and is obtained by heating natural eucryptite (low eucryptite), $LiAl[SiO_4]$; it is stable above 962 °C. It crystallizes in the hexagonal – trapezohedral system with a superlattice structure, with space group $D_6^4 - P\,6_222$ or $D_6^5 - P\,6_422$; lattice constants: $a_0 = 0.528$ nm, $c_0 = 1.127$ nm; the axial ratio is $c_0/a_0 = 2.134$; $Z = 3$.

Nepheline is a feldspathoid and occurs only in rocks with a low SiO_2 content (< 50 wt % SiO_2); it has a lower SiO_2 content than the alkali feldspars. It is formed in alkaline rocks in place of albite if the magma is sufficiently unsaturated with SiO_2. It is often associated with other feldspathoids such as sodalite, cancrinite, or leucite, or less often with kalsilite. Nepheline is a typical main constituent of the Atlantic-type igneous rocks, plutonic rocks (e.g., foyaites, theralites and essexites), hypabyssal and volcanic rocks (e.g., phonolites, tephrites and nephelinites). In the Mediterranean suite, minor amounts of nepheline are usually found with the common leucite (see Section 2.3). Nepheline crystallizes immediately after mafic minerals in magmatic rocks, and at about the same time as alkali feldspars in plutonic rocks. In hypabyssal and effusive rocks, nepheline occurs as phenocryst and in the matrix.

The ratio potassium : (potassium + sodium) in nepheline varies between 7 % and 37 % in effusive rocks and also in plutonic rocks, in which nepheline was formed in the presence of anorthoclase or sanidine. In plutonic rocks containing orthoclase and albite, nepheline contains 14 – 23 atom % potassium. Igneous nepheline (slow cooling) has a cloudy or greasy luster and is known as elaeolite. On rapid cooling (effusive rock), the crystals remain clear. Albite can form isomorphous mixtures with nepheline in sodium-rich solid solutions with a maximum albite content of ca. 20 mol % at 700 °C; the albite is often irregularly distributed, i.e., in zones. As with feldspar (see Section 2.1.2), the order/disorder behavior of Al/Si in the nepheline is temperature dependent.

Occurrence in Industrial Products. In refractory technology nepheline occurs in four modifications that do not contain potassium, $NaAlSiO_4$:

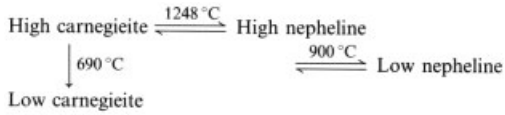

High carnegieite [*001302-34-7*], α-$Na[AlSiO_4]$, crystallizes in the cubic system, with space group $T^4 - P\,2_13$; lattice constants: a_0 = ca. 0.74 nm, $Z = 4$. It has the cubic high cristobalite-type lattice with additional sodium, and can form mixed crystals with cristobalite to a limited extent. High carnegieite occurs only as a synthetic product; in mineralogy it is often termed *β-carnegieite.*

High nepheline [*012251-27-3*], β-$Na[AlSiO_4]$, is orthorhombic. *Low nepheline* [*012251-27-3*], γ-$Na[AlSiO_4]$, is hexagonal – pyramidal. It occurs, for example, in chamotte and SiO_2- containing corundum bricks. It is converted into mullite in the walls of blast furnaces, glass-melting furnaces and checker bricks by the attack of Na_2SO_4 and Na_2O. *Low carnegieite,* β-$Na[AlSiO_4]$, is triclinic.

In refractory technology, five modifications of the polymorphic substance $KAlSiO_4$ are known: hexagonal – trapezohedric kalsilite, orthorhombic (high) kaliophilite, hexagonal kaliophilite (kalinepheline), natural hexagonal kaliophilite and anomalous kaliophilite. A characteristic feature of kalsilite is that it "polymerizes": trikalsilite and tetrakalsilite are known.

Kalsilite occurs in incrustations on the chamotte walls of blast furnaces and in the chamotte checker bricks of Siemens – Martin furnaces. *Potassium – nepheline* is formed at $< 1540\,°C$ in chamotte bricks of industrial furnaces that are exposed to K_2O vapors (e.g., in the brick linings of blast furnaces and tuyeres). *High kaliophilite* is stable above 1540 °C. It also occurs in the chamotte linings of blast furnaces and tuyeres, as well as in the chamotte checker bricks in Siemens – Martin furnaces and glass furnaces used to melt optical K_2O glasses.

Properties and Composition of Industrial Products. The following physical properties are important for industrial applications.

Density:
Nepheline: 2.56 – 2.665 g/cm^3
Kalsilite: 2.59 – 2.625 g/cm^3
β-Carnegieite: 2.513 g/cm^3
α-Carnegieite: 2.343 g/cm^3

Mohs hardness:
Nepheline: 5 1/2 – 6
Kalsilite: 6

Specific heat at 20 – 100 °C:
Crystalline: 0.77 J g^{-1} K^{-1}
Amorphous glass: 0.804 J g^{-1} K^{-1}

Melting point:
Sodium nepheline + albite, 100 – 70 % $NaAlSiO_4$: 1526 – 1410 °C
Kalsilite (0 – 20 % $NaAlSiO_4$): 1300 – 1350 °C
β-Carnegieite (ca. 100 % $NaAlSiO_4$): 1526 °C
Eutectic in the system nepheline – albite: 1068 °C

Transformation temperature:
Nepheline $\rightarrow \alpha$- carnegieite: 1248 ± 5 °C
α-Carnegieite $\rightarrow \beta$- carnegieite: 687 – 692 °C
Nepheline + 30 % kalsilite $\rightarrow \alpha$- carnegieite: 1400 °C
Nepheline + 35 % anorthite $\rightarrow \alpha$- carnegieite: 900 °C

Nepheline shows piezoelectric properties in a high-frequency field.

Compositions of commercially mined nephelines are given in Table 7.

Nepheline has a markedly alkaline reaction in water. Dilute acids, especially HCl, decompose nepheline, producing gelatinous silicic acid.

Deposits. Only a small number of workable nepheline syenite deposits are known, the most important of which are in Canada, Norway and Russia. The rock in a certain Canadian deposit consists of ca. 54 % albite, 18 % microcline, 18 % nepheline, < 4 % biotite, 6 % muscovite and magnetite. In the Western world, only four companies extract and process nepheline-containing rocks (only nepheline syenite), the combined output being ca. 1.2×10^6 t/a.

Extraction. The rock is usually extracted by open- cast mining using explosives. Adhering impurities are removed by sieve grates (grizzlies) prior to size reduction. Ferromagnetic impurities (e.g., biotite) are removed by means of rotary magnets or high-intensity magnetic separation.

Quality Specifications. The quality required depends on the use, e.g., Grade A (Table 8) for the glass industry, Grade B for glass wool manufacture, ceramic grade for the ceramic industry and a lower grade for use as a filler. The grades differ in iron content (Grade A ≤ 0.08 % Fe_2O_3, Grade B: ≤ 0.4 % Fe_2O_3) and in color, which varies with heavy metal content and mean particle size (in some applications).

Uses. The most important consumer is the *glass industry,* which accounts for ca. 70 % of the nepheline marketed. Here, the silicate acts not only as a source of SiO_2 and Al_2O_3, but also as a flux. Mixtures of raw materials for bottle manufacture can contain 5 – 15 % nepheline. Nepheline can replace feldspars in most ceramic and glass-making applications (see Section 2.1.7) with only small adjustments to batch chemistry being necessary. The choice between feldspar and nepheline (nepheline syenite) is often a question of cost.

The *ceramic industry* consumes ca. 15 % of the total nepheline output. The nepheline content in sanitary and whiteware ceramics can be

Table 7. Analyses of natural nephelines and kalsilites (in wt %)

	A	B	C	D	E	F	G	H
SiO_2	44.65	41.88	40.20	38.47	38.48	63.36	60.20	60.8
TiO_2	0.00	0.03	0.05	0.00	0.05			
Al_2O_3	32.03	32.99	32.51	30.81	31.01	22.05	23.30	23.0
Fe_2O_3	0.59	0.74	1.82	1.63	1.12	0.09	0.07	0.08
MgO	0.00	0.00	0.10	0.63	0.00			0.03
CaO	0.71	0.78	1.44	0.20	0.03	5.80	0.30	0.05
Na_2O	17.25	16.11	10.86	2.09	0.30	5.77	10.60	10.4
K_2O	3.66	6.82	12.22	25.65	28.33	2.43	5.10	4.6
H_2O (105 °C)	0.21	0.03	0.00	0.00				0.7
H_2O (105 – 1000 °C)	0.96	0.71	0.20	0.20	0.67			
Total	100.06	100.16	99.77	99.94	100.00	99.50	99.57	99.66
n_E	1.535		1.543	1.543	1.539			$\simeq 1.53$
n_0	1.531		1.539	1.537	1.533			$\simeq 1.53$
Nepheline	81.9	76.1	54.7	9.9	1.4			
Kalsilite	12.9	23.6	45.3	88.9	96.8			
Others	5.2	0.3		1.2	1.8			

A) Nepheline phonolite, New Zealand. B) Nepheline foyaite Transvaal. C) Nepheline potassium ankaratrite, Angola. D) Kalsilite, venanzite, Italy. E) Kalsilite, complex granular intergrowth of kalsilite and nepheline, Angola. F) Nepheline – aplite (whole rock), Feldspar Corp., USA. G) Nepheline – syenite (whole rock), Sobin Chem. Inc., Canada. H) Nepheline Senite (Treminex 958), Quarzwerke GmbH, Frechen, Germany.

Table 8. Composition (wt %) of Canadian nepheline syenite products (Grade A)

Company	SiO_2	Al_2O_3	K_2O	Na_2O	CaO	Fe_2O_3	LOI∗
Indusmin	59.8	23.6	5.0	10.2	0.6	0.07	0.6
IMC	60.2	23.3	5.1	10.6	0.3	0.08	0.4

∗ Loss on ignition.

25 – 30 %, in industrial chemical porcelain and translucent porcelain china 15 – 30 %, and in floor and wall tiles 15 – 55 %.

Finely ground nepheline and feldspar (see Section 2.1.7) are becoming increasingly important as *fillers* for plastics [especially unsaturated polyester resins, poly(methyl methacrylate), and thermoplastic polyurethanes], rubber, adhesives, paints and pigments (Table 9). Since the refractive indices of nepheline and the matrix are almost the same, e.g., in nepheline/poly(vinyl chloride) compositions, transparency is maintained, even at high pigment volume concentrations. The high surface and volume resistance of nepheline mean that it can also be used as a filler in insulating paints, casting resins for electrical applications and other products.

In Russia, nepheline syenite is an important raw material for *aluminum production.* In a multistage process, the Al_2O_3 is extracted and then treated by melt electrolysis. High activity *silica gels* can be produced from nepheline by decomposition with hot 22 % sulfuric acid.

2.3. Leucite

The potassium aluminum silicate leucite, a feldspathoid, belongs to the analcime – leucite group of tectosilicates without nontetrahedral anions. The aluminum : silicon ratio is 1 : 2.

Structure and Mineralogy. As with all the tectosilicates, the oxygen ions at the vertices of the $[AlO_4]$ and $[SiO_4]$ tetrahedra are linked to four neighboring tetrahedra. This produces a three-dimensional open framework in which K^+ ions are located in the spaces within the framework. Whereas in the nepheline group the arrangement of tetrahedra is hexagonal or pseudohexagonal, the minerals of the analcime – leucite group have a cubic or pseudocubic arrangement.

Leucite [*001302-34-7*], $K[AlSi_2O_6]$, is dimorphous. Below 605 °C, it exists as low leucite (α-leucite) and above 605 °C as high leucite (β-leucite). *Low leucite* crystallizes in the tetragonal – pseudocubic system, with the space group C^6_{4h} – $I4_1/a$; lattice constants: $a_0 = 1.304$ nm,

Table 9. Physical data for nepheline fillers

Product∗	Particle diameter, µm		Brightness	Density,
	D_{50} value	Upper limit	(555 nm)	g/cm^3
Minex 4	10.8	30	90	2.6
Apex 400	9.1	32	90.5	2.6
Minex 7	3.9	20	90	2.6
Apex 700	5.9	25	92	2.6
Minex 10	2.7	15	90	2.6

∗ Trade names: Minex is produced by Indusmin and Apex by IMC.

$c_0 = 1.385$ nm; the axial ratio is $c_0/a_0 = 1.062$; $Z = 16$. *High leucite* crystallizes in the cubic – hexakis-octrahedral system, with the space group O_h^{10} – Ia 3 d; $a_0 = 1.343$ nm; $Z = 16$.

Leucite is isotypical with analcime, $Na[AlSi_2O_6] \cdot H_2O$; pollucite, $(Cs, Na)[AlSi_2O_6] \cdot H_2O$, and synthetic $Cs[FeSi_2O_6]$; it is homeotypic with wairakite, $Ca[AlSi_2O_6]_2 \cdot H_2O$. In leucite, K^+ can be replaced to a small extent by Na^+, rarely by Rb^+, and very rarely by Ca^{2+}; these minor substituents sometimes appear as enriched isomorphous layers. The pseudocubic, pale green iron(III) leucite, $K[FeSi_2O_6]$, also known as a synthetic material, can occur in very small quantities with the rock-forming leucite.

The apparently cubic crystals of leucite are anisotropic and consist of numerous twin lamellae, which are arranged parallel to the surfaces of the apparent rhombic dodecahedron. In thin sections, leucite shows a complex twinning on {110} under crossed polarizers. On raising the temperature to 605 °C, leucite becomes isotropic and is then genuinely cubic (high leucite). The transformation is reversible.

Leucite is the most characteristic component of SiO_2-unsaturated magmatites with a preponderance of potassium (Mediterranean rocks) and is typical of magmatic regions (e.g., the Rome region in Central Italy). It is found in many foyaitic and theralitic volcanic rocks.

Leucite occurs in refractory products, e.g., in chamotte bricks from blast furnaces and tuyeres.

High leucite melts congruently at 1686 °C. When potassium feldspar is heated, it melts incongruently above 1150 °C forming a mixture of leucite crystals in a SiO_2-rich melt. At 1150 °C, the melt contains 57.8 wt % leucite crystals, and at 1530 °C all the leucite becomes molten. On cooling, leucite crystallizes first, and the melt therefore becomes richer in SiO_2. The leucite crystallization is completed at 1150 °C and potassium feldspar is then formed.

Properties. Leucite crystals are usually white or gray, glassy, translucent, but very seldom clear. Cleavage is poor parallel to (110), the fracture being conchoidal and brittle. The following physical properties are important for the use of leucite as a raw material for glass:

Density: leucite 2.47 – 2.50 g/cm^3, molten glass produced from leucite 2.41 g/cm^3.

Mohs hardness: 5 1/2 – 6.

Melting point: pure leucite 1686 °C, impure leucite 1298 – 1430 °C.

Heat of crystallization: 47.65 kJ/mol. The tendency to crystallize is very low, and the melt therefore usually solidifies as a glass. Volume increase on solidification of the melt: 2.9 %.

Specific heat (10 – 100 °C): natural mineral 0.8 J g^{-1} K^{-1}, in the amorphous state 0.733 J g^{-1} K^{-1}.

Leucite can contain up to 1.5 wt % Na_2O and 0.8 wt % CaO. It is decomposed by concentrated hydrochloric acid with precipitation of SiO_2.

Deposits and Processing. Leucite deposits occur in the United States (Montana and Wyoming) and in Italy (Mount Vesuvius). Processing and classification of the leucite grades are very similar to that of feldspar and nepheline (see Sections 2.1.6 and 2.2).

Uses. Leucite, like feldspar, is mainly used as a raw material in glass manufacture. Other applications include the production of potassium fertilizers, alum and potassium nitrate.

2.4. Olivine

Olivine (peridot, chrysolite) [*1317-71-1*] is a rock-forming nesosilicate (an orthosilicate, i.e.,

a silicate with isolated $[SiO_4]^{4-}$ tetrahedra held together by metallic ions) belonging to the olivine series of minerals $(Mg, Fe)_2[SiO_4]$. There is an infinite series of solid solutions of the end members forsterite (Fo) [*015118-03-3*], $Mg_2[SiO_4]$, and fayalite (Fa), $Fe_2[SiO_4]$. The intermediate members are olivine $(Mg_{90-70}, Fe_{10-30})_2[SiO_4]$, hyalosiderite $(Mg_{70-50}, Fe_{30-50})_2[SiO_4]$, hortonolite $(Mg_{50-30}, Fe_{50-70})_2[SiO_4]$, and ferrohortonolite $(Mg_{30-10}, Fe_{70-90})_2[SiO_4]$. Other members of the olivine series include tephroite $Mn_2[SiO_4]$, knebelite, $(Mn, Fe)_2[SiO_4]$, and iron knebelite $(Fe, Mn)_2[SiO_4]$.

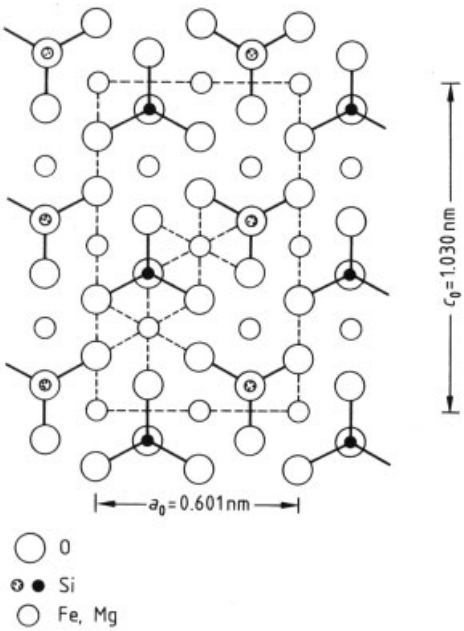

Figure 24. Structure of olivine, projection parallel to *b* on (010)
The $[SiO_4]^{4-}$ tetrahedra point alternately upward and downward

Structure and Mineralogy. The minerals of the olivine series form orthorhombic crystals, space group D_{2h}^{16} – Pmcn, $Z = 4$, with the following lattice constants (in nm):

	Forsterite	Olivine	Hortonolite	Fayalite
a_0	0.600	0.601	0.607	0.617
b_0	0.478	0.478	0.479	0.481
c_0	1.028	1.030	1.034	1.061

Olivine (Fig. 24) can include small amounts of other substituent elements (e.g., nickel). Weathering removes MgO and SiO_2, forming the important nickel serpentine ore mineral garnierite (nickel deposits of the New Caledonia type).

Olivine occurs as intergrown, loose, prismatic or thick tabular crystals and as granular aggregates. Cleavage in the (001) plane is good, and also takes place imperfectly in the (010) plane, the fracture being conchoidal.

Olivine crystallizes at an early stage from a magma. It is found typically as the main mineral constituent in ultrabasic igneous rocks (peridotites and dunites), and as a residual material in the serpentinites formed from them. In the ultrabasics, olivine is accompanied by pyroxene, etc. Olivine is often found in basic magmatites both in plutonic and volcanic rocks (e.g., gabbros, diabases, basalts and melaphyres). It also occurs in meteorites. Olivine is a main constituent of the upper earth mantle.

Olivine of jewelry quality is known as chrysolite, and forms clear, pale green stones that can be polished.

In the presence of water, olivine readily forms other silicates, e.g., serpentine (serpentinization – antigorite, chrysotile). Olivine often then becomes pseudomorphosed to form serpentine, $Mg_6[(OH)_8/Si_4O_{10}]$, and talc, $Mg_3[(OH)_2/Si_4O_{10}]$. The magnesium removed during serpentinization occurs as brucite, $Mg(OH)_2$, magnesite, $MgCO_3$, and cryptocrystalline magnesite of the Kraubath type (Styria, Austria). The iron present in olivine forms secondary magnetite Fe_3O_4 $[Fe^{3+}(Fe^{2+}Fe^{3+})O_4]$ during serpentinization.

Synthesis. Synthetic olivine is produced by calcination of chrysotile tailings from the asbestos mines in the Thetford Mines area of Canada. The tailings are calcined at 850 °C in a rotary kiln to give a mixture of forsterite and silica. A purer forsterite can be synthesized by adding magnesite during calcination. The pelletization of magnesite and the tailings is followed by calcination in a rotary kiln at 800 °C and further heating to a temperature of 1350 °C in a static furnace, whereby magnesite combines with the silica to form forsterite.

Forsterite in Industrial Products. Forsterite is a common component of refractory magnesia products (especially chrome–magnesite bricks), and is the principal component of refractory forsterite bricks (→ Glass, Chap. 5.7.). It is used in high-frequency insulators and also occurs in special metallurgical slags and upon reaction of refractory magnesia products with acidic slags.

Properties. Olivine has a glassy luster and is transparent or translucent, but seldom colorless (i.e., iron-free). The yellowish to green color (which gives it its name) varies from light to dark and may be opaque or even black with high iron content. The double refraction (birefringence) indices are usually high and depend on the magnesium : iron ratio.

The following physical properties are important for industrial applications:

Density: forsterite (Fo) 3.21 g/cm^3, Fo_{90}/Fa_{10} (90 mol Fo/10 mol Fa) 3.25 g/cm^3, fayalite (Fa) 4.34 g/cm^3.

Mohs hardness: forsterite 6 1/2 – 7, fayalite 6 1/2.

Melting point: forsterite (no decomposition) 1885 °C (volume change 3 %), fayalite 1205 °C.

Temperature stability limit for magnesium-rich olivine in an oxidizing atmosphere 1750 °C; softening temperature under a load of 2 kg/cm^2 1600 – 1700 °C; thermal shock resistance moderate.

Heat of fusion of forsterite 3.444 kJ/mol; specific heat (21 – 51 °C) 0.791 kJ/g; thermal conductivity (kJ m^{-1} h^{-1} K^{-1}) at 100 °C 19.3, at 500 °C 11.3, and at 1000 °C 8.8; mean coefficient of thermal expansion over the range 0 – 500 °C 8.3×10^{-6} K^{-1}, 0 – 1000 °C 9.5×10^{-6} K^{-1}, 0 – 1500 °C 11.0×10^{-6} K^{-1}.

Free energy of formation: forsterite − 1922 kJ/mol, fayalite − 1337.5 kJ/mol.

Electrical resistivity: at 800 °C 2×10^6 Ω · cm, at 1200 °C 10^4 Ω · cm, at 1400 °C 10^3 Ω · cm.

Ideal chemical compositions: forsterite ($2\,MgO \cdot SiO_2$): 57.3 wt % MgO, 42.7 wt % SiO_2; fayalite ($2\,FeO \cdot SiO_2$): 70.5 wt % FeO, 29.5 wt % SiO_2.

Olivine is insoluble in water, even at elevated temperature and pressure (620 °C, 30.4 MPa). It is stable towards cold acid, but complete decomposition takes place on heating in dilute acid (0.2 mol/L HCl), forming silica gel. Olivine reacts similarly with HNO_3 and H_2SO_4, but dissolves only slowly in 20 % H_2F_2. Solutions of the alkali metal sulfates readily decompose it. Olivine is resistant to alkaline solutions.

The chemical resistance of olivine in refractory materials is good towards reducing atmospheres, moderate toward carbon, poor toward acidic flux, good toward basic flux, and good toward molten metals.

The following data are characteristic of commercial olivines:

Density: 3.1 – 3.3 g/cm^3
Bulk density (dry sand): 1.4 – 1.7 g/cm^3
Tamped density (dry sand): 1.7 – 1.9 g/cm^3
Bulk density (molding sand): 1.7 – 2.1 g/cm^3
Moisture content (max.): 0.5 %
Softening point: > 1700 °C
Specific heat: 0.84 – 1.38 J g^{-1} K^{-1}
Coefficient of thermal expansion (20 – 1200 °C): 11.6×10^{-6} K^{-1}
FeO content (max.): 8 wt %
pH: 9

The most useful properties of forsterite-rich olivines are their high softening point, low and nearly linear thermal expansion, low thermal conductivity, high density, low abrasivity and (unlike silica) lack of toxicity.

Deposits and Commercial Products. An important economic factor in the extraction of olivine is the fact that olivine-rich ultrabasics or olivine rocks and the serpentinites derived from them are associated with chromite, platinum, and nickel deposits. The principal olivine deposits extracted today are in Norway, Japan, Austria, the United States, Italy, Spain and Pakistan. Smaller producers are Mexico, Sweden, India and Brazil. Japan has a large production of olivine and a production of serpentine which is several times higer than that of olivine. [6]t. The first deposit was extracted in 1948 in Aaheim (Norway).

World production of forsterite-rich olivine and dunite is ca. 6.8×10^6 t/a, of which ca. 2.1×10^6 t/a are in Europe.

Commercial olivines usually contain at least 85 % forsterite. Their chemical composition is MgO ca. 45 – 50 %, SiO_2 ca. 40 – 43 %, FeO ca. 5 – 7 %. Loss on ignition is ca. 1 – 2 %.

Rocks that contain higher proportions of derived minerals such as serpentine and talc are known as dunite. The chemical composition is MgO ca. 36 – 42 %, SiO_2 ca. 36 – 39 %, with a correspondingly higher loss on ignition.

Extraction and Processing. Olivine is extracted by both open- cast and underground mining. The rock is usually broken up by explosives.

The mined rock is processed by multistage size reduction (jaw crushers, cone crushers, and hammer mills), and classified by multiple-deck vibratory screens, with intermediate washing to remove any soft accompanying minerals (e.g., pyroxene, serpentine, chlorite or talc) in the form of a slurry. After washing, the material is mechanically dewatered by continuous centrifuges and dried by rotary or belt dryers (residual moisture content < 0.5 %).

Powdered olivine is produced by impact or hammer mills, giving products with narrow particle size ranges and low dust contents.

Uses. Important uses for magnesium-rich olivines are:

1) *Ore additive* in steel blast furnaces to condition the slag and improve its flow. Here, olivine replaces dolomite and silica and enables the ratio $(CaO + MgO) : (SiO_2 + Al_2O_3)$ to be adjusted to ca. 0.8 – 1.2 %. Also, less heat is required to melt the slag components.
2) Olivine is well suited for the *slag conditioning* of furnace slag and for basicity control in blast and electric arc furnaces. Olivine is added in the form of lumps (10 – 40 mm), sinter (3 – 6 mm) and pellets.
3) Olivine sand is used in *foundry sands* for casting manganese steel, high-alloy steels, brass, bronze and aluminum and in the production of gray iron and alloy steel, which usually require a fine finish. Olivine sands have several advantages over silica sands: the casting surfaces are smoother, fettling and cleaning are easier. Olivine sands can be reused many times, and there is no danger of silicosis. However, the basicity of olivine prevents its use in many applications, for example, with acid-catalyzed furan and phenolic no-bake binders.
4) *Highly refractory forsterite bricks* containing magnesia binders and possibly chrome ore (→ Glass, Chap. 5.7.). The added MgO combines with excess SiO_2 to form forsterite. The addition of chrome ore improves thermal shock resistance and reduces the modulus of distortion (V modulus). The most important advantage compared with magnesia bricks is that scale and slags with a high iron oxide content do not adhere to the wall surfaces below 1400 °C because the MgO is bonded. Also, forsterite bricks are less liable to sulfate attack in the temperature range 800 – 1100 °C. However, olivine can be attacked by Al_2O_3, CaO, SiO_2 and alkali due to the formation of lower melting eutectics. For refractory-grade olivine an Mg-rich composition closer to that of forsterite is required. Refractory olivine is used in the steel industry and in the manufacture of precast refractory linings for incinerators, where its heat-storage properties can result in more complete combustion, and its low expansion rate allows heat cycling without spalling of the lining.
5) *Highly refractory cements* (particle size range 0 – 0.2 or 0 – 1 mm). These are formulated as for forsterite bricks, but with a higher content of fluxing materials.
6) *Highly refractory ramming mixtures* with similar compositions to those of bricks. These are incorporated into furnace linings by manual or compressed-air ramming.
7) *Sand-blasting material* which does not involve a silicosis hazard (unlike silica). Olivine is used to some extent as a blasting abrasive and as a grit in water-jet cutting. Together with other minerals, it is used in cleaning buildings and occasionally for etching structural steel.
8) *Olivine bricks as heat reservoirs* in electrical night storage heaters.

2.5. Andalusite

Andalusite [*12183-86-1*] belongs to the nesosubsilicates (orthosilicates that contain isolated $[SiO_4]$ tetrahedra and additional nontetrahedral anions), and, like kyanite (Section 2.6) and sillimanite (Section 2.7), is a mineral of the Al_2SiO_5 group.

Structure and Mineralogy. Al_2SiO_5 can form three different crystal structures, i.e., Al_2SiO_5 exhibits polymorphism in the form of andalusite, sillimanite and kyanite. In all three phases one aluminum atom (Al_I, Fig. 25) always has the coordination number six, i.e., it is surrounded octahedrally by six oxygen ions. The $[Al_IO_6]$ octahedra are linked by common edges to form chains parallel to the *c* axis. The c_0 dimension in all three minerals is therefore approximately the same (ca. 0.55 nm). The other aluminum atom (Al_{II}, Fig. 25) in andalusite has the coordination number five, in sillimanite four, and in kyanite six. The chains of octahedra in andalusite are held together by $[AlO_5]$ and $[SiO_4]$ groups. In andalusite none of the bonds break preferentially, so that andalusite has a higher Mohs hardness (7 1/2) than sillimanite or kyanite. Libethenite $Cu_2[OH/PO_4]$, adamine $Zn_2[OH/AsO_4]$ and eveite $Mn_2[OH/AsO_4]$ are isotypical with andalusite.

The formula of andalusite is $Al^{[6]}Al^{[5]}[O/SiO_4]$. It forms orthorhombic – dipyramidal crystals with space group D_{2h}^{12} – Pnnm; lattice constants: $a_0 = 0.778$ nm, $b_0 = 0.792$ nm; $c_0 = 0.557$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.982 : 1 : 0.703$; $Z = 4$.

Some Mn^{3+} and Fe^{3+} and a small amount of Cr^{3+} and Ti^{4+} always replace the aluminum in andalusite (usually 2 – 3 atom % in total). Viridinite is a variety that contains several percent of Mn_2O_3 and Fe_2O_3 ($a_0 = 0.7808$ nm, $b_0 = 0.5929$ nm, $c_0 = 0.5567$ nm). Manganese andalusite is also known. Andalusite has a greater potential for ionic substitution (diadochy) than kyanite and sillimanite.

Andalusite occurs rather widely as intergrown, fairly large, thick columnar crystals. It can also occur in aggregates that may be granular, in the form of bunches or tufts, as radially arranged columns, and as coarsely tangled masses and lenses. Andalusite also occurs as chiastolite in argillaceous and spotted slates (Knotenschiefer), forming yellowish-white crystals containing a gray to black carbonaceous impurities in a regular manner.

Andalusite usually has a matt, cloudy, translucent, or opaque appearance and can be red, green or reddish-gray with a vitreous luster. It shows a variable degree of cleavage parallel to {110}. Transparent crystals of gem quality are rare (California, Brazil).

Andalusite represents a typical example of a mineral formed by contact metamorphism (hornfels, etc.). Typically, it often occurs in argillaceous rocks with a low lime content in contact aureoles around igneous intrusions. In places, andalusite is found in granite when the magma resorbed aluminum-rich sedimentary rocks. This mineral is also found in pegmatites and in accidental blocks (volcanic ejecta). Furthermore, andalusite is observed in mesozonal, aluminum-rich mica schists along with sillimanite, muscovite, biotite, almandine, and in some places also with cordierite and staurolite, although never with microcline.

Properties. The following physical properties are important for industrial applications:

Density: 3.09 – 3.16 g/cm^3
Mohs hardness: 7 1/2
Phase changes:
> 1350 °C: decomposition to mullite and vitreous SiO_2; on cooling, volume increases by 3 – 6 % with respect to the starting material
1840 °C: incongruent melting point, conversion to corundum and vitreous SiO_2
1920 °C: complete fusion
Specific heat capacity (0 – 100 °C): 0.705 J g^{-1} K^{-1}
Thermal conductivity: at 0 °C 2.26 kJ m^{-1} h^{-1} K^{-1}, at 189 °C 4.08 kJ m^{-1} h^{-1} K^{-1}

The ideal chemical composition is the same as that of kyanite and sillimanite: 62.93 wt % Al_2O_3, 37.07 wt % SiO_2.

Andalusite is not attacked by mineral acids apart from hydrofluoric acid, which attacks it very slowly in the cold, but readily when hot. Andalusite is stable towards dry HCl and Cl_2. Fused KOH strongly attacks andalusite, as does a solution of CO_2 in water at 160 °C and 0.6 – 0.7 MPa.

Deposits and Extraction. Andalusite deposits are mined in France (Côtes du Nord), Portugal (Porto), the Republic of South Africa (in three areas of the Transvaal: Groot Marico, Lydenburg, Thabazimbi), the People's Republic of China and the United States (White Mountains; Hawthorne, Nevada).

The overlying rocks above the andalusite-bearing schists and sands are not usually very

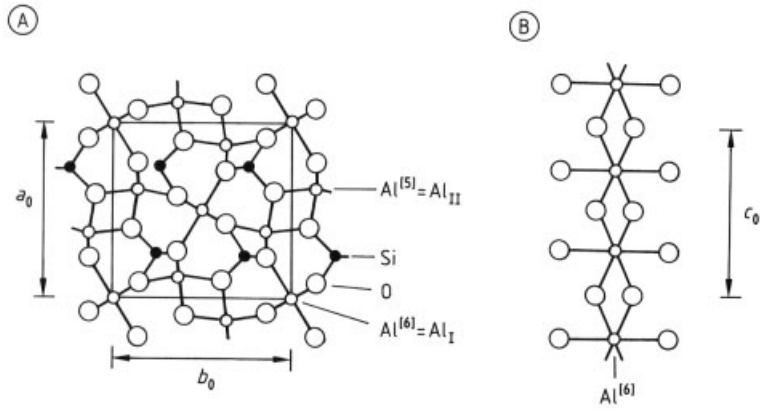

Figure 25. Structure of andalusite, $Al^{[6]}Al^{[5]}[O/SiO_4]$
A) Projection parallel to *c* on (001); B) $[AlO_6]$ octahedra along the *c* axis

thick. Therefore, extraction is mainly by opencast mining using explosives, compressed air hammers and front loaders.

The French deposits contain schists in which the andalusite content, which is coarsely crystalline, is ca. 15 %. The andalusite crystals are up to several centimeters long in the direction of the *c* axis and 1 – 4 mm wide in the direction of the *a* and *b* axes. Accompanying minerals are biotite, iron-rich chlorite, and quartz. The material is processed to give two grades of andalusite concentrate: 59 % Al_2O_3 with 1 % Fe_2O_3, and 53 % Al_2O_3 with 1 % Fe_2O_3. Processed andalusite sand has a particle size of 0.3 – 1.6 mm and is slightly colored.

In the Republic of South Africa, the andalusite is interbedded with schists. The primary deposits consist of andalusite-bearing schists in which the average content of extractable andalusite is 7 – 10 %. The secondary deposits are alluvial andalusite sands (placer deposits), in which the content of extractable andalusite is 10 – 50 %, reaching 50 – 60 % in the Burgersfort mine (Zimro).

In the White Mountains area of the United States 3 – 20 m thick deposits containing 70 – 80 % andalusite occur. In Hawthorne, Nevada, the vein-like deposits are 0.5 – 1.5 m thick and 1000 m long. They are extracted to a depth of 30 m.

Recently, a promising deposit was found near Tomduff in the Republic of Ireland. It consists of a series of andalusite schists and quartz-biotite schists, is 200 m thick, and continues for 3000 m along the strike.

Ore Treatment and Commercial Products. The broken schists and sands are crushed to a grain size of 0.3 – 1.6 mm (or to 0.6 – 3.35 mm by Weedons Minerals). Biotite, chlorite and other ferromagnetic minerals are removed by high-intensity magnetic separators. Quartz and other low-density minerals are removed by a wet process using settling tanks. These processes are described in detail elsewhere (→ Magnetic Separation; → Gravity Concentration). The dried andalusite concentrate is the commercial product. Commercial grades, with the amounts of major impurities, are listed in Table 10.

The refractory properties of the mullite – vitreous SiO_2 matrix (that is formed from the andalusite concentrate by reaction at high temperature) improve with increasing Al_2O_3 content (ideally 62.93 %). They also improve with decreasing TiO_2 content (max. 2 %), decreasing Fe_2O_3 content (< 1 % if possible) and decreasing alkali-metal content (fluxing effect).

Uses. Andalusite, like sillimanite, kyanite and synthetic mullite, is an important raw material for the production of highly refractory materials and is marketed with sillimanite and kyanite as "sillimanite minerals". The increase in volume when andalusite is fired above 1350 °C is small (3 – 6 %), whereas the corresponding increase for kyanite is considerably greater.

Andalusite of the proper fineness is suitable for use as a filler in coatings and plastics subjected to great mechanical or chemical stress, due to its crystal chemical properties and mor-

Table 10. Grades of andalusite concentrates

Company ∗	Grade	Al_2O_3, wt %	Fe_2O_3, wt %
Denain Anzin	keraphalite	59	≃ 1.531
Mineraux	KB	53	1
Cullinan Andalusite	standard	52.5	2
	premium	54.3	1.6
Export minerals		56	< 1
Marico Minerals	normal	58.8	1.3
	upgraded	61.6	0.6
Zimro		59	1.2
Lager Mining		57	1.2 – 2
Hoogenoog Andalusite		59.9 – 60	< 1

∗ Denain Anzin Mineraux is located in Paris, all other companies are in the Republic of South Africa.

phology. The length : width ratio of the crystals is 10 : 1 – 50 : 1. However, the energy required for grinding is considerable owing to the hardness of the mineral.

Economic Aspects. The annual consumption of crude rock by the Federal Republic of Germany, including material fired to > 1350 °C, is ca. 51 000 t. In France the annual consumption of "sillimanite minerals" is ca. 30 000 t and in the Republic of South Africa ca. 135 000 t. The Western countries have a consumption of ca. 170 000 t and the worldwide consumption of "sillimanite minerals" is ca. 300 000 – 350 000 t/a in the last few years.

2.6. Kyanite

Kyanite [*1302-76-7*] (cyanite, disthene) belongs to the nesosubsilicates (orthosilicates that contain isolated $[SiO_4]$ tetrahedra and additional nontetrahedral anions) and, like andalusite (Section 2.5) and sillimanite (Section 2.7), is a mineral of the Al_2SiO_5 group.

Structure and Mineralogy. One aluminum atom (Al_I, Fig. 26) always has the coordination number six, the $[Al_IO_6]$ octahedra being linked by common edges to form chains parallel to the *c* axis. The other aluminum atom (Al_{II}, Fig. 26) in kyanite also has the coordination number six. The chains of $[Al_IO_6]$ octahedra are joined to other $[Al_{II}O_6]$ octahedra, which are attached alternately to the right and left and are also linked via isolated $[SiO_4]$ tetrahedra to form stable planar structures parallel to [100]. Vacant octahedra form open channels parallel to (001). This accounts for the extension of the crystals in the *c* direction [001], giving them a ruler-like appearance and, on the other hand, complete cleavage along (100) and the anisotropic hardness, i.e., 4 – 4 1/2 Mohs in the *c* [001] direction, and 6 – 7 Mohs perpendicular to this in the *b* [010] direction.

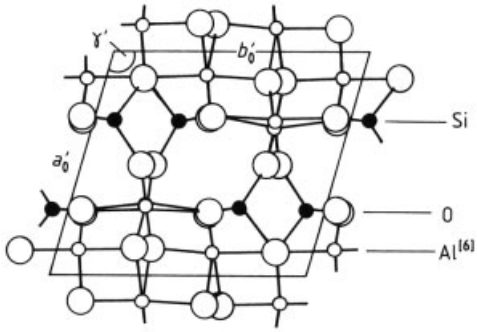

Figure 26. Structure of kyanite, $Al^{[6]}Al^{[6]}[O/SiO_4]$, projection parallel to *c* on (001)
Due to the low degree of symmetry, there are four different types of aluminum atoms, all with six- coordination. Al_I and Al_{II} in kyanite correspond to Al_I in sillimanite (see Fig. 27) and andalusite (see Fig. 25) and form chains of octahedra.

The mineralogical formula of kyanite is $Al^{[6]}Al^{[6]}[O/SiO_4]$. It forms triclinic – pinacoidal crystals with space group $C_i^1 - P\bar{1}$; lattice constants: $a_0 = 0.710$ nm, $b_0 = 0.774$ nm, $c_0 = 0.557$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.917 : 1 : 0.720$; $\alpha = 90°05^1/^2{}'$, $\beta = 101°02'$, $\gamma = 105°44^1/^2{}'$; $Z = 4$. The oxygen ions in kyanite are approximately in cubic close packing, yielding the highest density (3.53 – 3.65) of the Al_2SiO_5 polymorphs.

Ionic substitution is possible to a somewhat larger extent with kyanite than with sillimanite (Section 2.7), but is appreciably less than with andalusite (Section 2.5). Fe^{3+} can replace alu-

minum in the lattice to a small extent, as can Cr^{3+} in rare cases, which leads to an increase in the refractive indices. Low levels of alkali metals and hydroxide have also been found in completely transparent kyanite crystals.

Kyanite occurs mainly as radially arranged lamellar aggregates that may be tangled together. The crystals are intergrown, shaped like broad stalks or rulers that extend in the direction of the *c* (001) axis, and are frequently bent into a slightly waved shape and transversely striated due to translational movement. Kyanite is translucent to transparent and is usually blue, but is often flecked and irregularly colored. It can also be white or yellowish-white (rhatizite), pink or pinkish, with a nacreous or vitreous luster. Cleavage parallel to {100} is fully developed, is well-marked parallel to {010} and is poor or fibrous parallel to {001}. Cleavage parallel to {001} has the character of a fibrous fracture, since {100} also represents a slip plane with [001] as the slip direction.

Kyanite is a typical mineral of (aluminum-rich) metamorphic rocks, especially of mesozonal crystalline rocks, in which it is mainly associated with rutile, staurolite, almandine, hornblende, micas and corundum. It is occasionally found in gneisses, micaceous schists, granulites and eclogites. In catazonal rocks, kyanite only occurs in high-pressure formations because sillimanite is stable at lower pressures. Kyanite is the high-pressure form of Al_2SiO_5; andalusite and sillimanite both crystallize at lower pressures, with sillimanite being the high-temperature form.

On heating at atmospheric pressure, both kyanite and andalusite are converted monotropically to sillimanite. The conversion of kyanite to mullite and vitreous SiO_2 is of industrial importance, because these reactions take place above ca. 1200 °C.

Properties. The following physical properties of kyanite are important for industrial applications:

Density: 3.53 – 3.65 g/cm^3
Mohs hardness: $4\frac{1}{2} - 7$
Phase changes:
< 1300 °C: unstable with respect to sillimanite 1335 °C: decomposition to mullite and vitreous SiO_2, on cooling, volume increases by 14 – 16 % with respect to the original mineral
1830 °C: incongruent melting, forming corundum and vitreous SiO_2
1920 °C: complete fusion

The ideal chemical composition is 62.93 wt % Al_2O_3, 37.07 wt % SiO_2.

Kyanite is stable towards strong mineral acids. Attack by chlorine and dry HCl is only slight at red heat. Hydrofluoric acid decomposes ca. 84 % of the material. Molten KOH strongly attacks kyanite. However, there is little attack by aqueous solutions of alkali hydroxides.

Deposits and Extraction. Kyanite is a common constituent of crystalline schists. Few deposits can be extracted economically, and these consist mainly of schists that contain kyanite in crystal aggregates or as small coarse masses. Kyanite can also occur in lenses in pegmatites and in similar formations or nest-like masses in quartz veins.

The world's largest reserves are in the United States and stretch from northeast Georgia to South and North Carolina and Virginia in the Appalachians. Other large deposits are found in Canada and the Republic of South Africa. Deposits in the United States, Brazil, the Republic of South Africa, Zimbabwe, Spain and India are exploited commercially.

In the United States, kyanite-containing quartzite is extracted using explosives and excavation, followed by grinding and flotation. The flotation concentrate contains 91 % kyanite and 2 – 5 % iron oxides, with silica making up the balance. The Fe_2O_3 content of the dried concentrate is reduced to 0.16 – 0.94 % by high-intensity magnetic separation. The commercial product contains 59.5 – 61.8 % Al_2O_3.

A kyanite – sillimanite concentrate is produced as a byproduct of the treatment of zirconium sands in Australia.

In Spain, alluvial deposits of kyanite-containing gravel are dug by front-loading equipment. After washing and wet tumbling, the kyanite is picked out by hand, calcined at 1480 °C, crushed and classified into grain sizes of 0 – 1, 1 – 3 and 3 – 5 mm. Its composition is Al_2O_3 58 – 63 %, Fe_2O_3 0.5 – 1.5 %, TiO_2 0.15 – 0.5 % and alkali metals ca. 0.7 %.

Uses. Kyanite is an important raw material for refractory products (ceramics) and is marketed with sillimanite and andalusite as an industrial mineral.

Unlike sillimanite, andalusite and mullite, kyanite undergoes a large volume increase (16–18 %) when heated above 1325 °C. Powdered kyanite is therefore used in mixtures for ceramic and refractory products to compensate for shrinkage on firing. The refractory properties of the matrix of mullite and vitreous SiO_2 formed from kyanite concentrates at high temperature are a function of its chemical composition.

2.7. Sillimanite

Sillimanite (fibrolite) belongs to the nesosubsilicates (orthosilicates that contain isolated $[SiO_4]$ tetrahedra and additional nontetrahedral anions) and, like andalusite (Section 2.5), kyanite (Section 2.6) and mullite, is a mineral of the Al_2SiO_5 group.

Structure and Mineralogy. One aluminum atom (Al_I, Fig. 27) always has the coordination number six. The $[Al_IO_6]$ octahedra are linked by common edges to form chains parallel to the *c* axis. The other aluminum atom (Al_{II}, Fig. 27) in sillimanite has the coordination number four. The chains of octahedra in sillimanite are laterally bonded by $[Al_{II}O_4]$ and isolated $[SiO_4]$ tetrahedra which form bands of $[Al_2Si_2O_{10}]$ tetrahedra from two linked chains of $[AlSiO_6]$ (see Fig. 27). The four- coordinated aluminum in sillimanite cannot be substituted by silicon. Parallel to *b* (010), very strongly bonded layers are formed from the $[AlO_6]$ chains and $[Al_2Si_2O_{10}]$ bands. These can easily be cleaved from the neighboring layers by separation at the apices of the $[AlO_6]$ octahedra. This explains the perfect cleavage of sillimanite parallel to $\{010\}$. The chain structure also accounts for the generally fibrous structure of sillimanite crystals (fibrolite). As sillimanite has chains of tetrahedra, it is classified as an inosilicate (chain silicates, i.e., silicates with infinite chains of $[SiO_4]^{4-}$ tetrahedra).

The mineralogical formula of sillimanite is $Al^{[6]}Al^{[4]}[O/SiO_4]$. It forms orthorhombic–dipyramidal crystals with space group D_{2h}^{16}–Pbnm; lattice constants: $a_0 = 0.744$ nm, $b_0 = 0.760$ nm, $c_0 = 5.75$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.979 : 1 : 0.757$; $Z = 4$.

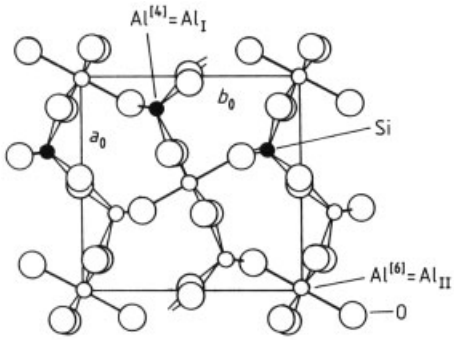

Figure 27. Structure of sillimanite, $Al^{[6]}Al^{[4]}[O/SiO_4]$, projection parallel to *c* on (001)

Sillimanite has low tolerance to ionic substitution, the aluminum atoms in the lattice being replaceable only by some Fe^{3+} and a small amount of titanium. Although alkali metals and hydroxide are often reported, they either occupy the spaces in the lattice or are impurities in the mineral.

Sillimanite usually forms columnar, fibrous or matted aggregates. Single crystals are rare, but large crystals (about the size of a fist) are produced synthetically. Dense masses intergrown with quartz are known as fibrous quartz. Sillimanite is yellowish gray, gray, brownish or grayish-green in color, is transparent to translucent and has a vitreous luster or a silky luster when in fibrous form.

Sillimanite is usually a common minor constituent of ortho- and paragneisses (e.g., feldspar–quartz–cordierite gneisses), micaceous schists, granulites, eclogites, and of pegmatites and quartz veins that occur in these rocks. It also occurs in contact-metamorphic rocks, sometimes in the form of very fine fibers. At high temperature (> 650 °C theoretical, ca. 1480 °C in practice), kyanite and andalusite are converted to sillimanite. At even higher temperature (1545 °C), sillimanite is converted to mullite, which has a lower SiO_2 content, and vitreous SiO_2. Accompanying minerals include feldspars, quartz, cordierite, muscovite, biotite, almandine, staurolite, corundum, andalusite, kyanite, rutile and ore minerals such as magnetite and ilmenite.

Properties. Sillimanite is the modification of $Al_2[O/SiO_4]$ that is stable at high temperature and low pressure. Cleavage on (010) is fully developed, but the fracture is uneven.

The following physical properties are important for industrial applications:

Density: 3.23 – 3.25 g/cm^3 (2.95 g/cm^3 after heating at 1545 °C)
Mohs hardness: 6 – 7
Phase changes:
1545 °C: decomposition to mullite and a vitreous matrix containing ca. 95 % SiO_2
1810 °C: liquid phase reaction to form corundum and a vitreous phase
1860 °C: complete fusion
> 1860 °C: boiling of the melt, producing a vapor phase with a high SiO_2 content and a liquid phase with a high Al_2O_3 content
Temperature stability limit in an oxidizing atmosphere: 1600 °C
Softening temperature under a load of 2 kg/cm^2: 1550 – 1650 °C
Thermal coefficient of expansion: at 20 – 100 °C 2.8×10^{-6} K^{-1}, at 300 °C 4.1×10^{-6} K^{-1}, at 800 °C 4.3×10^{-6} K^{-1}, and at 1200 °C 4.6×10^{-6} K^{-1}.
Specific heat capacity: at 300 °C 0.92 kJ kg^{-1} K^{-1}, at 1200 °C 1.0 kJ kg^{-1} K^{-1}
Thermal conductivity: at 300 °C 5.4 – 5.9 kJ m^{-1} h^{-1} K^{-1}, at 1200 °C 4.6 – 5.0 kJ m^{-1} h^{-1} K^{-1}

Sillimanite ($Al_2O_3 \cdot SiO_2$) has the ideal chemical composition 62.93 wt % Al_2O_3, 37.07 wt % SiO_2. It is insoluble in water, is decomposed by ammonium fluoride on heating and is resistant to weathering.

Like the other $Al_2[O/SiO_4]$ modifications, sillimanite is very difficult to melt and highly resistant to acids. Its chemical resistance when used as a refractory material is as follows:

Toward reducing atmospheres	moderate
Toward carbon	moderate
Toward acidic fluxes	good
Toward basic fluxes	moderate
Toward molten metals	good

Deposits and Extraction. Sillimanite deposits occur in Australia, India, Namibia, the Republic of South Africa, Brazil and Sri Lanka. The sillimanite is usually associated with quartz and kaolin (Australia) or with corundum (India, Republic of South Africa). World production is ca. 30 000 t/a.

The deposits are mainly extracted by opencast mining. Due to the hardness of the rock, drilling and blasting are necessary. The rock is crushed and classified with screening equipment.

Quality Specifications (see Table 11). Sillimanite used in the production of clay- containing refractory products of group 1 (→ Refractory Ceramics and ISO R 1109) must contain at least 56 % Al_2O_3 (ISO R 1109).

Uses. The most important use for sillimanite is in refractory materials. It is normally precalcined to ca. 1550 °C to convert the sillimanite to sinter-mullite (see Section 2.8), which has a high thermal shock resistance and a very low coefficient of thermal expansion (5.6×10^{-6} K^{-1} at 20 – 1000 °C). Sillimanite with a high corundum content is more resistant to slags with a high iron oxide content. Mechanically cut natural rock and bonded bricks are only used for lining blast furnaces, glass melting furnaces and roasting furnaces.

Coarsely ground material is used for highly refractory mortars and ramming compounds. Finely ground material is used in the production of architecturally applied ceramics (e.g., sanitary ware or wall and floor tiles).

Sillimanite is used as a filler in paints, coatings, adhesives and plastics, especially in systems cross-linked by polyaddition due to the nature of its surface (functional SiOH groups for polyaddition and polycondensation), high hardness, and ability to reinforce the matrix due to its acicular (i.e., needle-like) or fibrous structure.

2.8. Mullite

Mullite [*1302-93-8*] belongs to the nesosubsilicates (orthosilicates that contain isolated $[SiO_4]$ tetrahedra and additional nontetrahedral anions) and is very similar to sillimanite (Section 2.7) in its structure and properties.

Structure and Mineralogy. Mullite has a sillimanite structure, but differs from sillimanite in having a deficiency of oxygen. Since some of

Table 11. Sillimanite specifications (wt %)

	Al_2O_3	SiO_2	Fe_2O_3	TiO_2	Na_2O/K_2O
Highly enriched	59 – 65	39 – 33	< 1	max. 2	max. 0.1
With a high corundum content	65 – 78	33 – 20	< 1	max. 2	max. 0.1

the silicon atoms in the $[SiO_4]$ tetrahedra are replaced by Al^{3+}, some of the positions normally occupied by oxygen atoms are left unoccupied to maintain charge balance. Therefore, mullite has a higher aluminum content than sillimanite and has a chemical composition between $3\,Al_2O_3 \cdot 2\,SiO_2$ and $2\,Al_2O_3 \cdot SiO_2$.

The mineralogical formula of mullite is $Al_4^{[6]}Al_4^{[4]}\,[O_3(O_{0.5}, OH, F)/Si_3AlO_{16}]$ or $Al_4^{[6]}Al_4^{[4]}\,[O_{4-x/2}/Si_{4-x}Al_xO_{16}]$, where $x = 1.0 - 1.6$. It forms orthorhombic – dipyramidal crystals, with space group D_{2h}^9 – Pbam; lattice constants: $a_0 = 0.7566$ nm, $b_0 = 0.7682$ nm, $c_0 = 0.2884$ nm (i.e., half of c_0 for sillimanite); the axial ratio is $a_0 : b_0 : c_0 = 0.985 : 1 : 0.375$; $Z = 1$. Synthetic mullite, $3\,Al_2O_3 \cdot 2\,SiO_2$, has lattice constants $a_0 = 0.7576$ nm, $b_0 = 0.7688$ nm, $c_0 = 0.2884$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.981 : 1 : 0.375$. Synthetic mullite, with the composition $2\,Al_2O_3 \cdot SiO_2$, has lattice constants $a_0 = 0.7576$ nm, $b_0 = 0.7687$ nm, $c_0 = 0.2883$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.986 : 1 : 0.375$.

Another modification of Al_2SiO_5 appears to be a newly discovered phase in the system $Al_2O_3 - SiO_2 - H_2O$, crystallizing in the orthorhombic – dipyramidal system, with space group D_{2h}^9 – Pbam; lattice constants: $a_0 = 0.755$ nm, $b_0 = 0.827$ nm, $c_0 = 0.283$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.913 : 1 : 0.342$.

These cells of mullite are subcells (i.e., the ordering effects of the oxygen defects and of the silicon and aluminum ions are neglected). There are always six (silicon + aluminum) ions per subcell in mullite and twelve (silicon + aluminum) ions with doubled c lattice constants in sillimanite. In a few cases, ordering is found, and additional superlattice reflexes are observed corresponding to a larger unit cell, e.g., $a_0 = 3 \times a_0'$, $b_0 = 8 \times b_0'$, and $c_0 = 6 \times c_0'$, and $a_0' = 0.74$ nm, $b_0' = 0.76$ nm and $c_0' = 0.29$ nm. Sillimanite is ordered, with $c_0 = 2 \times c_0'$, and the same must also be true for "prägite", $2\,Al_2O_3 \cdot SiO_2$, which is only known as a synthetic product.

In naturally-occurring mullite, Fe^{3+} or Ti^{4+} can replace Al^{3+} to a small extent (up to 1 wt % Fe_2O_3 and up to 2 wt % TiO_2).

Natural mullite forms small pseudotetragonal colorless, white, yellowish or reddish crystals with columnar or needle-like development in the c [001] direction. Cleavage is well developed on (010), but is difficult to observe due to the normally small size of the crystals. Synthetic mullite is colorless, but green when Al^{3+} is substituted by Cr^{3+}. In synthetic products (e.g., chamottes and porcelain bodies), mullite has extremely fine crystals (< 1 µm). When these crystals come into contact with slags, larger prismatic crystals often form at the reaction interface. Melt-mullite products contain even larger crystals.

Occurrence and Formation. Mullite is not very often found in natural rocks. It greatly resembles sillimanite, with which it can easily be confused. It occurs almost entirely in contact-metamorphosed formations in primary argillaceous sandstones and phyllites contained as inclusions in basalts and also in a narrow exogenous contact zone (buchite) with anatexis at the edge of basalt pipes. The name mullite stems from its occurrence in sintered lumps of sedimented clay in the basalt of the island of Mull in Scotland.

The formation of mullite in synthetic products, such as chamotte products, porcelain and high-aluminum slags (e.g., from coal and coke) is of great industrial importance. It is formed when kaolinite is strongly heated or calcined to 1200 – 1600 °C and is also produced when kyanite, andalusite, sillimanite or other aluminosilicates are strongly heated, decomposing them into mullite and vitreous SiO_2. At 1810 °C, mullite dissociates into an equilibrium mixture of corundum and molten SiO_2. However, some investigations have shown that mullite can have a congruent melting point. In silicate slags, mullite only crystallizes if minor amounts of basic oxide components are present. Very small amounts of Na_2O, etc., significantly reduce the dissociation temperature of mullite.

Properties. The following physical properties are important for industrial applications:

Density: 3.10 – 3.16 g/cm^3
Mohs hardness: 6 – 7
Phase changes:
1810 °C: incongruent melting with decomposition to corundum and a vitreous phase
1920 °C: complete melting
Crystallization:
Occurs readily from the melt, well-developed crystals (usually finely matted together) are formed, even on very rapid cooling
Temperature stability limit in an oxidizing atmosphere: 1850 °C
Thermal expansion coefficient:
at 500 °C 4.0×10^{-6} K^{-1},
at 1000 °C 4.5×10^{-6} K^{-1},
at 1500 °C 5.3×10^{-6} K^{-1}
Specific heat capacity:
at 25 °C: 0.633 Jg^{-1} K^{-1}
Thermal conductivity:
at 100 °C 22.2 kJ m^{-1} h^{-1} K^{-1}
at 500 °C 15.5 kJ m^{-1} h^{-1} K^{-1}
at 1000 °C 14.2 kJ m^{-1} h^{-1} K^{-1}
Electrical resistivity:
at 500 °C 3×10^{-5} Ω · cm,
at 1000 °C 7×10^{-3} Ω · cm,
at 1500 °C 5×10^{-2} Ω · cm

The ideal composition of mullite is 71.8 wt % Al_2O_3, 28.2 wt % SiO_2. Both natural and synthetic mullite often deviate from the composition 3 Al_2O_3 · 2 SiO_2 (60 mol % Al_2O_3), and an Al_2O_3 content of 63 mol % is not uncommon. Mullite is obtained as a solid crystalline mass by sintering a mixture of Al_2O_3 and SiO_2 in the molar ratio 3 : 2 at 1775 °C.

Water and strong mineral acids do not attack mullite even when heated. Mullite also resists concentrated hydrofluoric acid at room temperature for long periods, although rapid decomposition takes place on heating to 70 – 100 °C. Resistance to molten metals and basic compounds is moderate. To achieve complete breakdown, mullite must be fused with sodium peroxide or a mixture of sodium carbonate and borax glass.

The chemical resistance of mullite in refractory materials is:

Toward reducing atmospheres	moderate
Toward carbon	moderate
Toward acidic fluxes	good
Toward basic fluxes	moderate
Toward molten metals	moderate

Deposits. Natural mullite is mined from only one significant deposit in northern Transvaal in the Republic of South Africa (Otavi Mining Company). The production rate is ca. 5000 t/a.

Synthesis. Synthetic mullite is of considerable economic importance. Outputs are as follows: United States ca. 27 000 t/a , United Kingdom: > 12 000 t/a. Producers in *Germany* Vereinigte Aluminiumwerke and Hülst AG; in the *United Kingdom* Keith Ceramic Materials Ltd.; in *Hungary* Hungary "Motim" Alumina and Electrocorundum Factory; in the *United States* Washington Mills Electro Minerals Corp.; in *Brazil* Elfus Ltda.; in *Japan* Showa Denko KK. For the very great demands placed on refractory materials, mullite is superior to the sillimanite minerals (sillimanite, andalusite, kyanite), which can be converted into crystalline mullite and a vitreous SiO_2 matrix only after calcination.

Two main grades of synthetic mullite are produced: fused mullite (electromullite) and sinter mullite. Fused zirconia mullite is also produced. In general, sintered mullite is manufactured in a rotary kiln, whereas fused mullite is produced either in a Higgins furnace or using a tilt-pour electric arc furnace.

Electromullite. A mixture of kaolinite and calcined aluminum oxide in a precisely controlled ratio is melted in an electric arc furnace. The product is crushed and classified into various grain sizes. In the selection of the raw materials, careful attention must be paid to levels of impurities, especially iron, other heavy metal ions and alkali metals. Metallic contaminants generated by abrasion of the crushing and classifying equipment are removed by magnetic separators.

The reaction product contains up to 95 % mullite and 76.4 – 75.2 % Al_2O_3. The following composition is given for an English grade: mullite 95 %, vitreous matrix and corundum 5 %, SiO_2 23 %, Al_2O_3 76.2 %, Fe_2O_3 0.1 %, CaO 0.2 %, MgO 0.1 %, $K_2O + Na_2O$ 0.4 %. The melting point (Seger cones) is ca. 1900 °C, and the material withstands long-term loading at 1750 °C.

Sinter mullite. The purity of this material is somewhat lower than that of electromullite. Briquettes of kaolinite, aluminum hydroxide and some water for bonding are sintered in tunnel furnaces at $\leq 1750\,°C$ and are then crushed and classified into size fractions. The product contains ca. 86 % mullite (i.e., 72.9 % Al_2O_3), ca. 0.7 % Fe_2O_3 and small amounts of other impurities. It withstands long-term loading at 1700 °C.

The sinter mullite produced by the Vereinigte Aluminiumwerke (Germany) has the following specification: Al_2O_3 72 %, SiO_2 16 %, density 3.1 g/cm^3, mean particle diameter 30 µm, maximum particle size 100 µm, color white.

Uses. The main area of use for mullite is in highly refractory materials for lining metallurgical melting furnaces and glass melting tanks, along with other applications where strongly acidic attack takes place at high temperatures. Mullite is also used as a raw material for porcelain for electrical insulation, as an opacifier for high-quality enamels, as a polishing medium (due to its great hardness), as a filler in plastics and paints, and as an abrasion-resistant reinforcing filler in coatings for surfaces and roadways that are subject to severe mechanical stress.

Some of the synthetic mullite's distinguishing attributes (e.g., excellent thermal shock resistance, good hot strength under load, volume stability at high temperatures, low shrinkage, high resistance to aggressive melts and slags, good abrasion resistance, and additionally high creep resistance of fused mullite) also permit its wide use. Refractories for the steel and glass industries are the most important market for sintered and fused mullite and fused zirconia mullite. Synthetic mullite is also used in the ceramic industry.

2.9. Vermiculite

Structure and Mineralogy. Vermiculite belongs to the phyllosilicates (sheet or layer silicates) and has a mica-like foliated structure. It displays the typical properties of the montmorillonite – saponite group to a pronounced degree: the layers have a greater excess charge and a greater capacity for cation exchange than talc or pyrophyllite layers. The structure contains layers similar to those found in talc, pyrophyllite (→ Talc), biotite, and muscovite (→ Mica, Chap. 2.). It comprises infinite, two-dimensional double silicate layers ${}^{2}_{\infty}[Si_4O_{10}]^{4-}$ (Fig. 28). Part of the Si^{4+} is substituted by Al^{3+}, and the terminal oxygen ions of the $[(Si, Al)O_4]$ tetrahedra are always on the same side (Fig. 28 A). The hexagons formed by these oxygen ions have hydroxyl groups at their centers. In this double layer, the oxygen layers formed from O (bound to one Si) and OH groups face each other directly, forming octahedral voids. These voids are occupied mainly by Mg^{2+}, but also by Fe^{3+} and Al^{3+} (Fig. 28 B), which hold together the two Si_2O_5/OH sheets. The octahedral voids are formed by four oxygen atoms and two hydroxyl groups. Due to the partial replacement of Si^{4+} by Al^{3+}, the double layer pocket $Mg_3[(SiAl)_4O_{10}/(OH)_2]$ has a slight excess negative charge, which is compensated by additional cations, e.g., Mg. These cations, together with water molecules, are interlayered between the layer pockets, usually Mg^{2+} or, less frequently, Ca^{2+}. This interlayer (intermediate layer) has a positive excess charge and consists of a double layer of $H_2O - Mg^{2+} - H_2O$, in which the H_2O locations may be only partially occupied. Each H_2O in this interlayer is linked to an oxygen atom in the neighboring silicate double layer by a hydrogen bond. Thus, vermiculite structurally resembles a kind of talc expanded by H_2O that has been formed from (hydrous) mica by potassium depletion.

The swelling capacity (ability to be hydrated) of vermiculite is due to the fact that the silicate layers have a slight excess negative charge and thus repel each other; this facilitates the penetration of water into the interlayer with its small positive excess charge. Unbonded water can be removed from the vermiculite by heating to $< 110\,°C$, but the water fixed by cations in the interlayer is only released at $> 300\,°C$. On heating, the water in the interlayer is released either gradually or, on sudden heating, with expansion and formation of worm-shaped structures. Exfoliation causes unidirectional expansion perpendicular to the cleavage plane (001) and the volume increases greatly (from 8- to 50-fold). If the trapped water is driven off slowly, even at 250 °C vermiculite shows no exfoliation. Exfoliation of vermiculite can also be caused by treatment with hydrogen peroxide, which leads

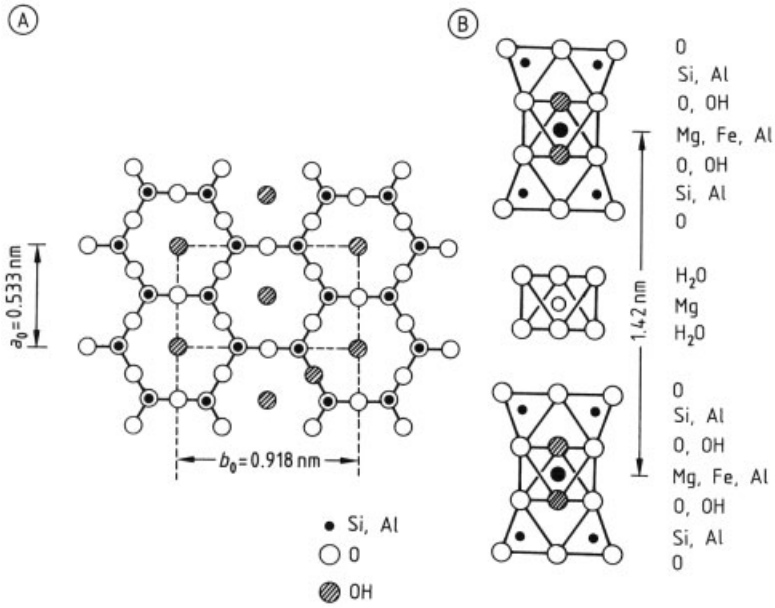

Figure 28. Structure of vermicullite
A) The two-dimensional infinite silicate layer $[(Si, Al)_4O_{10}]^{x-}$, projection parallel to *c* on (001); B) Intermediate $H_2O-Mg^{2+}-H_2O$ layer between two silicate double layers, projection perpendicular to the *a* axis

to release of oxygen by reaction with interlayer magnesium ions.

The mineralogical formula of vermiculite is:

$$(Mg, Fe^{3+}, Al)_3[(OH)_2/Al_{1.25}Si_{2.75}O_{10}] \cdot Mg_{0.33}(H_2O)_4$$

or

$$Mg_{2.36}Fe^{3+}_{0.48}Al_{0.16}[(OH)_2/Al_{1.28}Si_{2.72}O_{10}]^{0.64-} \cdot Mg_{0.32}(H_2O)_4$$

Vermiculite forms monoclinic crystals, with the following lattice constants: $a_0 = 0.533$ nm, $b_0 = 0.918$ nm, $c_0 = 2 \times 1.436$ nm; $\beta = 97°$; the distance between the layers, $d_{002} = 1/2c \cdot \sin\beta$, is ca. 1.425 nm.

Batavite and *protovermiculite* are related to vermiculite and have similar compositions and lattice constants. In nickel – vermiculite, a high proportion (10 – 25 atom %) of the Mg^{2+} ions is substituted by Ni^{2+}. The alkali metal content of vermiculite is always very small, and high contents are due to replacement of layers of vermiculite by layers of biotite. An increase in the number of these biotite layers can eventually lead to the formation of *hydrobiotite.* Another "mixed-layer" mineral is formed by interlayering of the vermiculite with chlorite; when there is a high iron content, this is known as *jefferisite*, and when the interlayering is regular, with a 1 : 1 ratio, it is known as *corrensite*.

Crystals of vermiculite can reach a size of several centimeters and are then easily confused with biotite. However, they are usually much smaller (2 µm) and can be of the size of clay particles when they occur in sediments. Vermiculite forms monoclinic – pseudohexagonal, short, columnar platelets and mainly occurs in aggregates, which can be bent into worm-like shapes. Fibrous formations can also occur, but are very rare. Vermiculite is brown, bronze-colored, yellow, green, or colorless. Cleavage on (001) is perfect and mica-like, but the platelets are not elastic as in micas.

Deposits of vermiculite often occur in ultrabasic and basic rocks (e.g., pyroxenites, hornblendites, peridotites and their metamorphic equivalents) that have undergone hydrothermal potassium metasomatism and alteration to actinolite – biotite formations. These in turn can be altered in the final stages potassium of this metasomatism or by anhydrous processes to formations of vermiculite and hydrobiotite with relicts of phlogopite, biotite, tremolite and actinolite. The alteration of biotite to vermiculite takes

place by leaching of K^+ from between the silicate layers by hydrothermal activity, percolating ground water or supergene weathering. Vermiculite found as an alteration product of biotite in other metamorphic rocks, such as schists and gneisses, is of similar origin. Vermiculite is only rarely found as an alteration product of mica in marine sediments. In places, vermiculite is found in these sediments as an alteration product of, e.g., chlorite and amphibole. Vermiculite occurs occasionally in marble and in carbonatites. Of some economic importance are vermiculite deposits in the contact zones between ultrabasic and basic rocks and acid intrusives.

Vermiculite is transformed into enstatite and talc on heating to 650 °C. Under increasingly severe hydrothermal conditions ($<$ 200 °C, 70 MPa) it is converted into pseudochlorite, which is itself converted to chlorite and montmorillonite at ca. 550 °C. Vermiculite is easily altered into biotite by potassium metasomatism, although this process is only rarely observed in nature.

Properties. The following physical properties are of interest for industrial applications:

Density: 2.3 – 2.75 g/cm
Mohs hardness: 1 1/2 – 2
Sintering temperature: ca. 1260 °C
Fusion point: 2000 – 2400 °C
mp: 1315 °C
Specific heat capacity: 0.84 kJ kg^{-1} K^{-1}
Thermal conductivity:
0.062 – 0.065 W m^{-1} h^{-1} K^{-1}

Deposits and Extraction. Deposits of economic importance are found in northern Transvaal (Republic of South Africa), at Libby in Montana, at Louise in Virginia, and at Encoree in Laurens County in South Carolina (USA). Further deposits occur in Brazil, Zimbabwe, Kenya, Australia, India, China, Russia and Japan. The vermiculite in these deposits has been formed from phlogopite and biotite by weathering (leaching of K) and hydrothermal alteration. The vermiculite at Encoree in South Carolina occurs in metamorphic rocks in contact with acidic intrusions interspersed with pegmatites. Accompanying minerals include corundum, apatite, serpentine, chlorite, talc and mica.

At Palabora in the Transvaal (Republic of South Africa), the Archean contains a Proterozoic intrusive alkali rock complex that consists of, e.g., pyroxenite, syenite, ultrabasic pegmatoids and carbonatite. This intrusive complex is not only well known for its large copper deposit, but also has an outstanding vermiculite deposit. The vermiculite is an alteration product of phlogopite and is free of asbestos.

The vermiculite bearing zones are selectively extracted by open-cast mining, The ore is loosened by blasting and removed by front loaders. The vermiculite ore is beneficiated by screening, jigging and classification.

Processing has three aims:

1) To separate the vermiculite from the accompanying rock, including rudimentary biotite and phlogopite
2) To delaminate the "books" formed from the vermiculite layers and to split the "leaves" (thickness $>$ ca. 0.8 mm) into "flakes"
3) To classify the flakes into commercial particle size fractions

In the first stage, the rock is usually broken into $<$ 20 mm or $<$ 16 mm granules, and the adhering fine material is removed by washing and screening. Further size reduction separates the vermiculite from the accompanying rock.

In a process operated by Palabora Mining (RSA), the vermiculite flakes are separated from the accompanying rock, including mica, by multistage screening in a horizontal stream of air using various suspension techniques.

The products from both the United States and South Africa are separated into five or six size grades in the millimeter range. The chemical compositions are listed in Table 12. Compared with the North American material, South African vermiculite has undergone a higher degree of physical transformation, gives a higher yield on thermal expansion, and therefore has a higher market value. The difference can also be seen from the chemical composition.

Expansion. Vermiculite, similar to perlite (exhibiting three-dimensinal expansion), expands when heated., but only in one direction (one-dimensional) showing the unique phenomenon of exfoliation by sudden heating to ca. 300 °C or higher. Responsible for this exfoliation is a double layer of water (i.e., an

Table 12. Chemical composition of raw vermiculite (average values, wt %)

Component	Libby, Montana	Encoree, South Carolina	Palabora (RSA)
SiO_2	38.64	38.66	39.37
MgO	22.68	20.04	23.37
Al_2O_3	14.94	17.36	12.08
Fe_2O_3	9.29	8.45	5.45
K_2O	7.84	5.24	2.46
Na_2O			0.80
CaO	1.23	0.75	1.46
Cr_2O_3	0.29	0.50	
MnO	0.10	0.11	0.30
P_2O_5	Trace	Trace	0.15
S	Trace	Trace	0.18
Cl	0.28	0.52	0.02
H_2O	5.29	8.71	14.24
Total	100.58	100.34	99.88

H_2O-Mg^{2+} – interlayer) trapped between the silicate double layers and liberated upon heating, whereby the rapidly generated vapor cannot be set free without buckling and separating the structural layers. The low thermal conductivity and low density of expanded vermiculite lead to its widespread use in thermal and acoustic insulation applications, as well as in refrigeration engineering. Expansion occurs prerpendicular to the cleavage plane (001).

The suitability of a particular vermiculite for conversion to expanded aggregates can be assessed by differential thermal analysis. However, the temperature – time profile of the furnace used for expansion can only be established with certainty by a laboratory-scale expansion test.

For the expansion and exfoliation of vermiculite various furnaces (e.g., rotary, cascade, vibrating tray, injection tray) can be used. The vermiculite concentrate is heated to between 430 and 1100 °C. The water in the vermiculite is liberated in three stages in 0.25 – 8 s, beginning with a reversible reaction at ca. 150 °C, followed by the onset of exfoliation above 500 °C and vaporization of the residual water at 870 °C. The bulk density is reduced from about 640 – 1040 to 64 – 176 kg/m^3. In some industrial processes, the vermiculite flakes are expanded in 5 – 8 s at 1250 – 1500 °C and then cooled immediately to 400 °C. Since exfoliated vermiculite shows a huge increase in bulk, only shipping of the vermiculite ore or concentrate is economic. Plants for expansion of the vermiculite are therefore generally located close to the consumers.

Uses. In its most important applications, vermiculite is used in its expanded state to exploit its low density, good thermal and acoustic insulating properties and high absorption capacity for liquids.

Vermiculite is utilized in the construction industry for thermal and acoustic insulation. Thermal insulation is the largest market.

Loose filling is widely used in the construction industry and in industrial applications for insulating hollow spaces (e.g., cavity walls), ceilings, walls and roofs. Loose-fill vermiculite is commonly a coarse grade with a particle size up to 16 mm and a bulk density of 56 – 72 kg/m^3. At cryogenic temperatures perlite is a better insulator, but vermiculite is also used in limited quantities for such purposes.

Vermiculite is also utilized in the manufacture of castable insulation products. For producing high-alumina vermiculite cement, a fast-hardening cement, alumina and lime are heated in a furnace to a high temperature. The product is ground to a powder, and vermiculite is added. Vermiculite cement can also be manufactured by combining it with Portland cement. Vermiculite cement is used in the production of ultra-lightweight aggregates for the construction industry (Table 13) and as lightweight insulation.

Vermiculite is utilized as a constituent in the manufacture of refractory blocks and fire bricks for use in furnaces up to 1200 °C.

In packaging technology, vermiculite is used to cushion impacts, for thermal insulation and to absorb liquids in case of breakage of the container.

Table 13. Physical properties of vermiculite lightweight concrete

Vermiculite/ cement ratio	Density, kg/m^3	Compression strength, N/mm^2	Thermal conductivity, $Wm^{-1}\,h^{-1}\,K^{-1}$	Shrinkage on drying, %
8 : 1	400	0.70	0.094	0.35
6 : 1	480	0.95	to	to
4 : 1	560	1.23	0.158	0.45

In metallurgy, it is used to cover molten metal for thermal insulation during intermediate storage prior to casting.

Expanded vermiculite is also used with garden and house plants to store water and plant nutrients (hydroculture), as a filter for swimming pools and for absorbing spilled liquids.

Economic Aspects. Total world production of crude vermiculite is ca. 512 500 t/a. Because of the short resources the production declined in the United States to ca. 177 000 t/a. Vermiculite is invariably expanded at a location close to the consumer to facilitate transport. The largest consumer country is the United States. European consumption is modest, most being imported from the Republic of South Africa. Although the total consumption of silicates used for thermal and acoustic insulation (including expanded perlite, hydrated and expanded broken glass, glass wool and rock wool) continues to increase, that of vermiculite is decreasing due to its comparatively high price.

2.10. Perlite

Perlite of rhyolitic composition is a natural, volcanic glass that is usually black or gray, but sometimes brownish red. It has curved shrinkage cracks and therefore breaks up into spheroidal granules. It is usually produced by rapid cooling of rhyolitic melts (rhyolite being the volcanic equivalent of granite). Being a volcanic glass, perlite contains few crystals and only a small amount of water (2 – 6 % combined water). Naturally-occurring glasses containing 3 – 8 % water are known as pitchstone (hydrated glass).

Perlite was first mined and expanded in 1946 in the United States. It has since been put to a wide range of uses in Europe. Thermally expanded perlite is increasing in importance as a lightweight filler in constructional units for thermal and noise insulation.

Composition varies according to the type of deposit:

SiO_2	72.0 – 76.0 wt %
Al_2O_3	11.0 – 17.0 wt %
K_2O	4.0 – 5.0 wt %
Na_2O	2.9 – 4.0 wt %
CaO	0.5 – 2.0 wt %
MgO	0.1 – 0.5 wt %
Fe_2O_3	0.5 – 1.5 wt %
TiO_2	0.03 – 0.2 wt %
MnO_2	0.03 – 0.1 wt %
SO_3	0 – 0.2 wt %
H_2O, chemically bound	2 – 7 wt %
Density	2.2 – 2.4 g/cm^3
Softening temperature	810 – 1090 °C
Melting temperature	1260 – 1340 °C
Specific heat	831 $J\,g^{-1}\,K^{-1}$
Bulk density, broken, classified	1.45 kg/L

The most significant variations are in water content, which has an important influence on the physical properties after expansion.

Deposits and Processing. The most important European deposits are found in Greece (Milos, Kos), Italy (Sardinia), Slovakia, Hungary, Bulgaria and Iceland. Further important deposits are located in Turkey (Cumaovasi, Izmir), Canada, Australia, Japan and the United States (New Mexico). Perlite is commonly extracted by open-cast mining. It is then crushed (usually by cone crushers) and classified using vibratory screens into several grain sizes with closely specified tolerances.

The industrial value of the classified ore depends on the temperature and time necessary for expansion, relative volume increase, ratio of closed to open bubbles (cells), nonexpandable fraction and compression strength of the product.

Expansion. Like vermiculite, perlite expands when heated, but in three dimensions. Expanded perlite has excellent insulation properties, comparable to those of vermiculite, which it can substitute to some extent. Expansion is performed in two types of furnace: (1) tilted

horizontal rotary furnaces through which the feed material passes in countercurrent flow to the combustion gas and (2) vertical rotary or blast furnaces through which the expanding grains and hot gases pass in cocurrent flow. Vertical expanders are more common. The expansion results from shock heating the grains to a temperature in the softening range (727 – 1127 °C), which results in bursting. Then the water bound in the perlite vaporizes and the heat-softened grains expand into light, cellular particles with up to 15 – 30 times the original volume. The cellular structure consists of numerous tiny, closed, thin-walled, glass-sealed air bubbles, which account for the excellent insulating properties of the expanded perlite. The optimum temperature – time profile must be closely followed. Longer heating times lead to shrinkage of the heated aggregates, and higher temperatures cause shrinkage and/or melting. The expanded material is classified in cyclones. In some cases, the expanded perlite is ground by pendulum or roller mills. If the product is to be used as a filler or additive, it is treated to make it hydrophobic; it may be silanized with dilute aqueous silane suspensions, or coated with dilute, aqueous silicone oil emulsions.

Quality. Important quality criteria for the expanded material include bulk density, ratio of closed to open air bubbles (cells), unexpanded fraction, and, if used as a filler, compression strength of the bulk material and binder requirement or water absorption properties. The bulk density of expanded perlite depends on the amount of chemically bonded water in the unexpanded material and on the temperature – time profile used for expansion and can be 0.03 – 0.2 kg/L. The fraction of the material with open cells and the fraction of nonexpanded grains should not exceed 5 %. These parameters are both determined by a water-immersion test using material that has not been rendered hydrophobic.

Properties. The most important properties for the diverse applications of expanded perlite are its extremely low density, low thermal conductivity, low thermal expansion, good sound absorption, adequate compression strength at a bulk density of > 0.1 kg/L, and resistance to acids, weathering, and microbial attack. The average properties of industrial perlite (from Milos) after expanding and grinding to 98 % < 40 µm are given below:

Refractive index	1.49
Brightness index Ry (according to DIN 53 163)	82 – 85
Bulk density	0.080 kg/L
Tamped density	0.240 kg/L
Oil absorption number (standard grade)	500 – 600 g/100 g
Water absorption number (standard grade)	400 – 500 g/100 g
Water absorption number (hydrophobic grade)	20 – 25 g/100 g
pH	8 – 9
Thermal expansion coefficient (0 – 100 °C)	0.007 K^{-1}
Thermal conductivity (25 °C)	
Bulk density	
0.04 kg/L	0.04 $W\,m^{-1}\,h^{-1}\,K^{-1}$
0.09 kg/L	0.05 $W\,m^{-1}\,h^{-1}\,K^{-1}$
0.13 kg/L	0.06 $W\,m^{-1}\,h^{-1}\,K^{-1}$
Thermal conductivity (bulk density 0.09 kg/L)	
20 °C	0.05 $W\,m^{-1}\,h^{-1}\,K^{-1}$
400 °C	0.11 $W\,m^{-1}\,h^{-1}\,K^{-1}$
800 °C	0.22 $W\,m^{-1}\,h^{-1}\,K^{-1}$

When perlite is used as a filler and as an additive to lightweight concrete, very high compression strength and resistance to abrasion are required during mixing. These grades have higher densities (0.12 – 0.14 g/L) and are produced from rhyolitic rock.

Uses. Expanded perlite is used as an ultralightweight aggregate in the construction industry and as lightweight insulation. The most important uses of expanded perlite (normally treated to render it hydrophobic) include loose-fill insulation, filled plaster board and wall elements, insulating silicate-bonded boards for thermal and flame protection, insulating bitumen-bonded boards for back facing, insulating boards and coating materials with polymer binders, mortar (bonded with cement, plaster, or silicate), polymer dispersion mortars and plasters for thermal insulation and prevention of condensation on walls and ceilings (e.g., in swimming pools and high humidity rooms).

Coarse-grade expanded perlite (2 – 5 mm) is used in metallurgy to limit the loss of heat from molten metal surfaces. A fine grade (0.15 – 1.8 mm) is used as a thermal insulator in cryogenic applications. In cryogenic applications, expanded perlite is normally uses unfer vaccum conditions.

Expanded perlite, with a bulk density of 0.04 – 0.06 kg/L, is used as an adsorptive filtering aid in precoated filters (e.g., for drinking water or medicinal products).

Economic Aspects. Next to the United States, Western Europe is now the largest consumer of perlite. World annual production of raw perlite is estimated to be ca. 1.8×10^6 t. In Germany ca. 100 000 t/a perlite is expanded, of which ca. 70 000 t (corresponding to ca. 580 000 m^3) is used as a raw material for the building industry. The amount of perlite (from Italy, Greece and Iceland) expanded in the United Kingdom is > 120 000 t/a. The total for France is ca. 90 000 t/a.

2.11. Pumice

Pumice [*001332-09-8*] is not a crystalline silicate mineral, but a light-colored, highly vesiculated, foamed volcanic glass of mainly acid character (> 66 % SiO_2, rhyolite), with a high melt viscosity, high pore volume (> 50 %) and variable water content. Rhyolite itself is an acid volcanic rock consisting of feldspar, quartz, biotite, etc. and glassy groundmass. By rapid undercooling (quenching) of rhyolitic magma, volcanic rocks are produced which consist completely or almost completely of volcanic glass, named obsidian (containing less than 1 wt % H_2O and composed of more than 80 wt % of volcanic glass including mostly very negligible amounts of minute inclusions of quartz, tridymite, feldspar, fayalite, etc.) or pumice, the latter highly vesiculated. Volcanic lava of rhyolitic composition releases enclosed gases (water vapor, carbon dioxide, nitrogen, sulfur dioxide, chlorine, etc.) at ca. 1000 °C explosively as a result of an abrupt pressure drop forming bubbly pumice. Due to its high porosity, pumice is very light and floats on water.

The specific heat capacity of pumice is 1.00 J g^{-1} K^{-1} and its thermal conductivity at 50 °C is 0.274 W m^{-1} h^{-1} K^{-1}.

Crushed and classified pumice is employed as an additive in light insulating concrete used to make constructional units. Due to its limited compression strength, this building material can only be used for one- or two-storey buildings or for cladding load-bearing frameworks. The specific heat capacity of pumice concrete is 1.005 kJ kg^{-1} K^{-1} and its thermal conductivity is 0.457 W m^{-1} K^{-1} for a density of ca. 1.1 kg/L.

Building blocks (e.g., Bisotherm, Bisothermwerke Carl Riffer, Urmitz) consist of pure, natural pumice granules treated to remove organic and inorganic impurities. Denser inorganic impurities such as fragments of obsidian and lavastone are removed by air separation and organic impurities by calcining at ca. 400 °C in rotary kiln (DIN 18 151, 18 152). They are mainly used to build houses.

Pumice in different industrial grades has a wide range of applications: stonewashing of denim and cotton, in agriculture (soil substitute, additive), paint manufacture (non-skid coatings, acoustic insulating ceiling paints, fillers for textured paint), chemical industry (filtration media, chemical carriers, pesticide carriers), cleaning and polishing of metals and plastics, compounders (powdered hand soaps, glass cleaners), dental and cosmetic (polishing natural teeth and dentures, smoothing rough skin), rubber (erasers, mold-release agents), glasses and mirrors (TV tube processing, glass buffing and polishing), furniture (hand-rubbed satin finishing, piano keys, picture frame gold leafing), leather (buffing), electronics (cleaning circuit boards) and pottery (filler).

Powdered pumice is used as an abrasive, as a polishing medium, as an antiskid material and as an additive to pumice soap. It is also used as an absorption medium due to its large surface area (e.g., in packages carrying breakable liquid containers).

As an insulator, pumice competes with expanded perlite (see Section 2.10) and expanded, exfoliated vermiculite (see Section 2.9).

The output of pumice in Germany is ca. 2.4×10^6 t/a, worldwide ca. $6.5 - 7 \times 10^6$ t/a.

Important commercially exploited pumice deposits occur in Germany (Koblenz-Neuwied district), in Italy (isle of Lipari – northeast to Sicily, and around Naples), on volcanic Greek islands (Dodecanese group: islands of Yali and Nisyros, Cyclades group: Thera Island), in Spain (Province of Ciudad Real, Canadas area on isle of Tenerife), Portugal (on the volcanic island of Sao Miguel in the Azores), the United States (in central New Mexico in the area of Santa Fe, at Gate in Oklahoma, and in Idaho and California),

on Caribbean islands (Dominicana, Martinique), on volcanic islands of Japan (southern parts of Hokkaido and Kyushu), and in New Zealand (North Island).

2.12. Basalt

Basalt is the most common volcanic rock, occupying large areas of the Earth's surface (e.g., plateau basalt). It contains $< 52\,\%$ SiO_2 and is the volcanic equivalent of the gabbro rock family. Basalt has a dense, compact, sometimes granular or porphyric structure and is dark to black in color. The main constituents of the matrix are plagioclase (anorthite 50 – 75 %) and pyroxene (mainly augite). Accessory minerals include olivine, magnetite, ilmenite, alkali feldspar, alkali pyroxene, alkali amphibole, feldspathoids and volcanic glass. Basalts are worldwide by far the most extensive exposed crustal igneous rocks, and their formation can be observed even now (e.g., numerous volcanoes in Asia, the United States, Europe and at the mid-ocean ridges).

The most important types are the tholeiitic basalts (often associated with some olivine and, when supersaturated with SiO_2, even quartz) and the alkali basalts (feldspathoids, alkali pyroxene, alkali amphibole and olivine). Medium- to coarse-grained basalts are known as *dolerites,* while olivine basalts (which contain feldspathoids as well as plagioclases) are known as *basanites,* and anchimetamorphic basalts of the Permian/Carboniferous era are known as *melaphyres, diabases* and *greenstones.*

The largest single basalt mass in Europe is in the Vogelsberg region of Germany. Other large deposits are in Hegau, the Eifel, the Westerwald and the Rhön mountains. Another deposit extends from the Oberpfalz (Bavaria), through the Erzgebirge and as far as the Lausitz region.

Some basalts have high strengths and are therefore used in the construction industry, e.g., as hard core, paving stones and building materials. In another important application, molten basalt is converted into refractory fibers (mineral wool for insulation and fire protection) (→ Fibers, 5. Synthetic Inorganic, Chap. 2.2.).

The physical properties required depend on the use intended for the basalt. These include high compression strength (2500 – 4000 kg/cm^2, with a maximum of 5800 kg/cm^2) and transverse strength (150 – 250 kg/cm^2, DIN 52 100). It should be as homogeneous as possible and should be resistant to frost, acids and waste gases (e.g., flue gas). The shapes of the naturally-occurring blocks of material also influence possible uses, e.g., columnar masses with polygonal cross sections, having 4 – 7 sides (feldspar basalts).

The density of basalt varies between 3.00 and 3.15 g/cm^3, the bulk density between 2.95 and 3.00 g/cm^3 and the bulk density of basalt lava between 2.20 and 2.35 g/cm^3. The porosity of basalt varies between 0.2 and 0.9 vol %, and of basalt lava between 20 and 25 vol %. Thus, the water absorption of basalt varies between 0.1 and 0.3 wt %, and of basalt lava between 4 and 10 wt %. The specific heat capacity (0 – 770 °C) is ca. 1.09 J g^{-1} K^{-1}, and the thermal conductivity (0 – 100 °C) is ca. 0.028 W cm^{-1} K^{-1}. The linear expansion coefficient is ca. 9×10^{-6} K^{-1}. Anhydrous basalt begins to melt at ca. 1280 °C, but at only ca. 1100 °C if it contains 1 % water. Basalt wool is stable up to 900 °C.

Sunburnt basalt is unsuitable for many applications. It is affected by weathering very soon after it is mined, especially by the heat of strong sunlight in the presence of moisture. Characteristic white spots are formed, and the rock crumbles. This phenomenon is not limited to basalts alone, but can also occur with other effusive rocks (e.g., phonolite). The "sunburning disease" of basalt is probably mainly due to the conversion of nepheline into analcime ($Na[AlSi_2O_6] \cdot H_2O$), which is accompanied by a volume increase of ca. 5.5 %.

2.13. Wollastonite

Wollastonite [*013983-17-0*], $Ca_3[Si_3O_9]$, is polymorphic. Three modifications occur naturally, the commonest being low wollastonite which has two polytypical structural modifications: triclinic wollastonite (-1T) and monoclinic wollastonite (-2M, parawollastonite).

Structure and Mineralogy. In monoclinic wollastonite, the $[SiO_4]$ tetrahedra are linked to form infinite one-dimensional dreier single chains of $[Si_3O_9]^{6-}$ units (Fig. 29). These are bonded together by Ca^{2+} ions (which balance

the electrical charge), and the chains lie parallel to b [010]. This explains why the crystals are always extended in the direction of b [010]. The monoclinic structure is derived from the triclinic by "inner twinning" on the (100) plane. Wollastonite (-1T) and wollastonite (-2M) are therefore inosilicates (silicates containing $[SiO_4]$ tetrahedra in infinite chains).

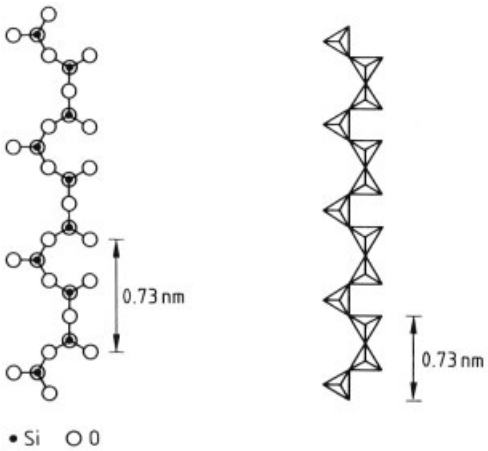

Figure 29. Basic configuration of $[SiO_4]$ tetrahedra in the infinite one-dimensional dreier single chains $[Si_3O_9]^{6-}_{\infty}$ of wollastonite, $Ca_3[Si_3O_9]$

Cyclowollastonite (high wollastonite or pseudowollastonite) is a cyclosilicate and is stable above 1126 °C. The $[SiO_4]$ tetrahedra are arranged in three-membered rings $[Si_3O_9]^{6-}$, which are linked together and neutralized by Ca^{2+} ions (Fig. 30). High-pressure $Ca_3[Si_3O_9]$ also contains rings of tetrahedra.

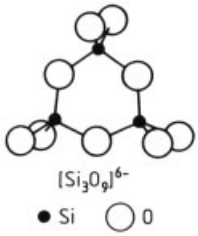

Figure 30. Basic configuration of $[SiO_4]$ tetrahedra in the dreier single rings of $[Si_3O_9]^{6-}$ of cyclowollastonite, $Ca_3[Si_3O_9]$

Wollastonite (-1T), $Ca_3[Si_3O_9]$, forms triclinic – pinacoidal crystals, with space group C_i^1 – P 1, and has the following lattice constants: $a_0 = 0.794$ nm, $b_0 = 0.732$ nm, $c_0 = 0.707$ nm; the axial ratio is $a_0 : b_0 : c_0 = 1.084 : 1 : 0.966$; $\alpha = 90°02'$, $\beta = 95°22'$, $\gamma = 103°26'$; $Z = 2$.

Wollastonite (-2M), (parawollastonite), $Ca_3[Si_3O_9]$, forms monoclinic – prismatic crystals, with space group C_{2h}^5 – $P\,2_1/a$; lattice constants: $a_0 = 1.536$ nm, $b_0 = 0.729$ nm, $c_0 = 0.708$ nm; the axial ratio is $a_0 : b_0 : c_0 = 2.107 : 1 : 0.971$; $\beta = 95°24^1/^2{}'$; $Z = 4$.

Cyclowollastonite (high wollastonite, pseudowollastonite), $Ca[Si_3O_9]$, forms triclinic – pseudohexagonal crystals ($b{:}a = 1.707 \approx \sqrt{3}$); lattice constants: $a_0 = 0.690$ nm, $b_0 = 1.178$ nm, $c_0 = 1.965$ nm;
the axial ratio is $a_0 : b_0 : c_0 = 0.586 : 1 : 1.668$; $\alpha = \gamma = 90°$, $\beta = 90°48'$; $Z = 8$.

High-pressure $Ca_3\,[Si_3O_9]$ forms triclinic crystals with lattice constants $a_0 = 0.6695$ nm, $b_0 = 0.9257$ nm, $c_0 = 0.6666$ nm; the axial ratio is $a_0 : b_0 : c_0 = 0.723 : 1 : 0.720$, $\alpha = 86°38'$, $\beta = 76°08'$, $\gamma = 70°23'$.

In natural wollastonite, Ca^{2+} can only be substituted to a very small extent (ca. 1 – 2 atom %) by Mg^{2+}, Fe^{2+} and Mn^{2+}. Alternatively, it can be substituted by 2 wt % Al_2O_3.

In the wollastonite (-1T) found in industrial products, (β-$Ca_3[Si_3O_9]$), Ca^{2+} can be substituted by Fe^{2+} up to 70 wt % $FeSiO_3$. In industrial cyclowollastonite, (α-$Ca_3[Si_3O_9]$), Ca^{2+} cannot be substituted by Fe^{2+}, but can be substituted by Mn^{2+} up to 1 – 2 atom % and by Sr^{2+} over the entire concentration range.

Wollastonite is usually intergrown and well-formed crystals are very rare. It occurs mostly in the form of thick plates parallel to {100} and {001} and is always extended on the b axis [010]. It is often found in coarse masses, in flat, thin, divergently radiating masses, or in stalked, spiky, tabular, lamellar or finely fibrous aggregates and masses.

Magmatic formation of wollastonite is possible where the magma resorbed calcium while in contact with limestones. This mineral is found in some alkaline igneous rocks and in phonolites, andesites, etc. Wollastonite is commonly a product of contact metamorphism, and sometimes also of regional metamorphism. Under the conditions of contact metamorphism, wollastonite is formed at high temperatures from quartz and calcite in siliceous limestones and by SiO_2 metasomatism in pure limestones. Cyclowollas-

tonite is only rarely found in metamorphic rocks (e.g., south-west Iran, Kropfmühl near Passau in Bavaria).

Cyclowollastonite is formed when wollastonite is heated to $>1126\,°C$. It melts congruently at 1544 °C. In industrial products, cyclowollastonite is the modification that is most often formed, e.g., in silicate bricks, in blast furnace slags and their reaction products with silicate bricks, in slightly basic open-hearth slags, or as a devitrification product in high-calcium bottle glasses.

Properties. The physical properties of interest for industrial applications are given in Table 14. Wollastonite is usually white, translucent and sometimes pale greenish, bluish or brownish. It has a vitreous luster, a nacreous luster on fractured surfaces and coarse masses, and a silky luster in fibrous aggregates. Cleavage is perfect on (100) and (001) and well developed but variable on $(\bar{1}01)$ and $(\bar{2}01)$. The Mohs hardness is 4 1/2 – 5 and the density 2.8 – 2.9 g/cm^3.

Cyclowollastonite usually forms irregular grains and is colorless, with cleavage on (001). The Mohs hardness is 5 and the density 2.91 g/cm^3.

The lower limit of transparency to UV light is at $\lambda = 290$ nm. Wollastonite shows a yellowish thermoluminescence at moderate light intensities. The phosphorescence is yellowish-white and is intensified by doping with 0.5 wt % Nd_2O_3. Cathodic luminescence is influenced by vacuum, temperature and additives.

The electrical conductivity at 1300 °C is $0.18\,\Omega^{-1} \cdot cm^{-1}$ and the dielectric constant at 20 °C is 6.17.

The ideal chemical composition for $CaO \cdot SiO_2$ is 48.29 wt % CaO, 51.71 wt % SiO_2. Aqueous mineral acids decompose wollastonite and pseudowollastonite, producing gelatinous SiO_2. Alkaline solutions do not attack it. It is resistant to weathering and industrial atmospheres.

Production. Pseudowollastonite (cyclowollastonite) is produced industrially as a sintered phase by a diffusion reaction:

$$CaO + SiO_2 \xrightarrow{1400\,°C} Ca[SiO_3] \text{ or}$$
$$CaCO_3 + SiO_2 \xrightarrow{1550\,°C} Ca[SiO_3] + CO_2$$

The reaction is accelerated by the addition of mineralizers (e.g., 1 wt % NaCl, Na_2WO_4 or CaF_2). The solid-state reaction between CaO and SiO_2 is also accelerated in the presence of water vapor.

Finely ground CaO and SiO_2 are melted together in the molar ratio 1 : 1, the melt is then rapidly cooled to form a glass, and the vitreous product is then heated to 800 – 1000 °C.

Deposits and Processing. Although wollastonite (-1T) and wollastonite (-2M), which is somewhat rarer, are widely distributed, in the past only a few deposits were extracted economically up to the 1980s, when the demand for wollastonite and its production increased and doubled in less than 10 years. The most important deposits are in the United States (Willsboro and Harrisville in the State of New York, Riverside in California, Tonopah in Nevada), China (Meiyaoshan and Tiengongshan in Jilin province, Fulapu and Shijianfang, near Faku in Liaoning province, Xinyu in Jiangxi province, Daye in Hubei province, Changxing in Zhejiang province, Qinghai province), India (Belka in the Sirohi District, Kheratarala in Udaipur District), Finland (Lappeenranta), along with Mexico (La Blanca and Panfillo Natera in Zacatecas State), Chile, Namibia (Usakos), Kenya (Kajiado), Turkey, Pakistan, North Korea, Russia, Canada, Spain, Greece, Morocco, New Zealand and Cuba.

In the Willsboro mines, the rock contains ca. 60 % wollastonite (-1T) and (-2M), ca. 40 % almandine, and some diopside. It is crushed in three stages to < 16 mesh (1.2 mm). The almandine is then removed by magnetic separators. The wollastonite concentrate is finally ground and classified by air elutriation.

The rock from Lappeenranta contains an average of 20 % wollastonite (-1T) and (-2M) and accompanying minerals: calcite, dolomite, and quartz. Treatment consists of wet grinding (rod mills), classification (hydrocyclones), flotation (the wollastonite becomes concentrated at the bottom of the flotation cell), magnetic separation of the wollastonite concentrate, filtration and drying. The finished product consists of 90 % wollastonite, 4 % quartz, 4 % calcite and 2 % other silicates.

Table 15 gives the chemical composition of commercial wollastonite products. Table 14

Table 14. Physical properties of wollastonite

Property	Wollastonite∗	Pseudowollastonite	Vitreous $CaO \cdot SiO_2$
Density, g/cm^3	1.8 – 2.9	2.91	2.9
Mohs hardness	4 1/2 – 5	5	
Temperature of phase change, °C	1190		
Enthalpy of phase change, kJ/mol	7.42		
mp, °C		1512	
Heat capacity at 1400 °C, kJ kg^{-1}		1434	
Melt viscosity, kg m^{-1} s^{-1}			
at 1550 °C		2.73×10^{-3}	
at 1600 °C		2.40×10^{-3}	
at 1650 °C		2.38×10^{-3}	
Specific heat capacity, J g^{-1} K^{-1}			
at 1100 °C	1.09	1.08	
at 0 – 700 °C			0.99

∗ (-1T) and (-2M).

Table 15. Chemical composition of commercial grades of wollastonite (-1T) and (-2M)

Component	Finland	United States	India	Theoretical
SiO_2	51.80	50.90	52.00	51.79
CaO	44.50	46.90	47.00	48.29
Al_2O_3	0.44	0.25	0.14	
Fe_2O_3	0.22 – 0.3	0.55	0.31	
MnO	max. 0.01	0.10		
TiO_2	max. 0.05	0.05		
MgO	0.56	0.10		
Na_2O	0.10			
K_2O	0.01			
Loss on ignition	2.20	0.90	0.55	

gives the physical properties of Finnish wollastonite (-1T) and (-2M).

Uses. Wollastonite is widely used in ceramics, plastics, asbestos substitution, metallurgy, paints and glass manufacturing. Wollastonite is classifed into high aspect ratio wollastonite and milled or powdered wollastonite. High aspect ratio wollastonite has an aspect ratio of 1:10 to 1:20 and is well suited as a reinforcing functional filler in plastics, rubber, paints, and coatings and as an asbestos substitute. In these applications, wollastonite is characterized by enhanced hardness, flexural strength, and impact resistance, and also improves the electrical properties and the thermal and dimensional stability of plastics. Milled or powdered wollastonite is used as a filler in ceramics and in metallurgy. The milled wollastonite grades are cheaper than the high aspect ratio grades.

The highest volume use of high-grade wollastonite, i.e., modifications (-1T) and (-2M), is as a raw material for *ceramic products* (see Tables 15 and 16). Teh considerable increase in use o wollastonite can be attributed to the spread of fast-firing techniques, significantly reducing the consumption of energy and improving productivity. Its important advantages include its lack of chemically bonded water (which could break down the structure and even cause explosions during firing), the ease of water removal on drying via the capillaries associated with its acicular (i.e., needle-like) crystal structure, the low shrinkage on drying and firing, the very low coefficient of expansion, the fairly low melting point, the mechanical strengthening of the "green" molded articles due to the acicular structure present until firing and the high breaking strength and thermal shock resistance of the finished product. The high lightness value (see Table 16), which is maintained after firing, enables wollastonite (-1T) and (-2M) to be used in the production of neutral, white ceramic products (e.g., white tableware). The raw material mixture should contain 20 – 50 % wollastonite. Other uses include wall and floor tile bodies and glazes, sanitary ware, earthenware, frits, frit

glazes, and porcelain enamels and specialized applications.

Table 16. Physical properties of Finnish wollastonite (-1T) and (-2M)

Property	FW 200	FW 325
Density, g/cm^3	2.9	
Mohs hardness	4.5 – 5	
Refractive index	1.63	
mp, °C	1394 – 1410	
Coefficient of thermal expansion, K^{-1}	1.5×10^{-6}	
pH of 10 % aqueous suspension	9.0	
Particle size (<74 µm), %	96.5	98.9
Specific surface area, cm^2/g	2980	3040
Oil number, g oil/100 g substance	27	35
Brightness value FMY/C∗, %	86	86
Solubility in water (5 min, 100 °C), mg/100 mL	0.0071	0.0076
Particle shape	Columnar to needle-shaped	

∗ Tristimulus value *Y* measured with standard illuminant according to DIN 5033, color temperature 2856 K, ca. 546.1 nm using ELREPHO equipment.

The addition of wollastonite to frits for ceramic glazes and enamels reduces the firing temperature and the formation of hairline cracks and helps stop cracking, crazing, breaking and glaze defects. It also improves surface hardness and luster. Blister formation due to outgassing of volatile constituents is also reduced. Wollastonite added in amounts of 1 – 3 % to flint and/or flux reinforces semivitreous products and decreases shrinkage and moisture expansion. In some other ceramic applications, feldspar and quartz are substituted by 2 – 5 % wollastonite in vitrified bodies to lower vitrification temperatures and to decrease shrinkage.

The (-1T) and (-2M) modifications are preferred for use as fillers in industrial chemical products and plastics. The chain structure of the lattice leads to the formation of acicular crystals and short fibers, which impart structural viscosity both to molten thermoplastics and to coating materials (paints, plastisols, etc.) and improve the mechanical properties of the solid matrix. In polymer emulsion paints, texture paints, plasters and putties, addition of wollastonite improves abrasion resistance and weathering resistance and reduces shrinkage cracking.

The *plastics* industry is a major consumer of wollastonite. In plastics most grades of wollastonite can be utilized, but most significant are those with high aspect ratio, fine and surface-coated grades. Wollastonite is used in both thermoplastics (polyamides, polyesters, liquid crystal polymers and engineering resins) and thermosets (phenolic molding compounds, epoxies, polyurethanes, polyurea and unsaturated polyesters). Typical loadings of wollastonite in plastics are: nylon 50 %, low-density polyethylene 40 %, polypropylene 23 – 28 % and polystyrene 30 %. A further important market for wollastonite is automotive plastics, in which it acts chiefly as a reinforcement, e.g., in reinforced reaction injection molded components.

In *paints and coatings* high aspect ratio wollastonite is used as a reinforcing filler. In paint films, wollastonite improves the mechanical strength, durability and resistance to weathering, cracking and other ageing effects. Wollastonite has a pH of 9.9 in water and hence functions as an effective alkaline buffer in latex paints, preventing destabilization due to pH drift, particularly when acidic pigments and extenders such as kaolin and aluminum silicate are utilized.

Acicular wollastonite can be mixed with mineral filler materials exhibiting platelet-shaped, lamellar particle shapes (kaolinite, mica, talc, and pyrophyllite) to give filler mixtures with desirable rheological working properties (e.g., adjustable thixotropy).

Wollastonite is an important substitute for short-fiber asbestos and is used widely in reinforcement applications, in fire-resistant wall board and cement products and friction products (brake pistons, brake linings, clutches etc.).

In metallurgical applications the low temperature fluxing of wollastonite imparts better surface finish in continuously cast steel and inhibits sparking in welding.

Due to its luminescent properties, wollastonite, especially the synthetic variety, is the base material of many phosphors, whose characteristics can be varied by doping with heavy metal and rare earth ions.

Synthetic wollastonite is used in synthetic bone implants, to replace bone loss and as a vertebral prosthesis in medical cases requiring load-bearing capacity. Synthetic wollastonite quickly forms strong, biologically safe bonds with osseous tissue.

The favorable electrical properties of wollastonite (high volume resistance and surface resistance, low dielectric loss angle) lead to its use

as an additive to electrical insulating paints and casting materials.

Acicular wollastonite particles can be mixed with filler materials with platelet-shaped particles (kaolinite, mica, talc and pyrophyllite) to give filler mixtures with desirable rheological working properties (e.g., thixotropy that can be adjusted).

The most important producers and suppliers of natural and synthetic wollastonite are the NYCO Minerals Corporation (Willsboro, New York, United States) and the Partek Industrial Minerals Oy (Lappeenranta, Finland). World production of natural wollastonite is estimated at 365 000 t/a. Of the 120 000 t/a of wollastonite that is expected to be produced in the United States, some 90 000 t is consumed by the domestic market and the surplus is exported mainly to Europe. The production of natural wollastonite by Mexican producers is estimated to be 30 000 t/a, by Indian ca. 60 000 t/a, by China 120 000 t/a, and by Finnish producers ca. 30 000 t/a. The European consumption is estimated to be well over 70 000 t/a, of which ca. 65 % is used in the ceramics industry.

2.14. Toxicology

Most of the silicates described in Chapter 2 are not known to be toxic or carcinogenic. However, the inhalation of fine particles ($< 5\,\mu m$, fibrogenic dusts) of silicates and hydrated silicates which have an acicular or fibrous morphology and are insoluble in the environment inside the lungs (pH < 7) should be filtered out by means of effective industrial dust-removal equipment or respiratory protection masks.

The MAK values for inert granular and platelet-shaped silicate dusts is set at 6.0 mg/m^3 in view of possible effects on the respiratory organs. This figure applies to exposure to the dust for one year, or, if each individual exposure to dust is known and documented, for a total exposure time of one year spread over a period of five years.

In the case of exposure to fine wollastonite dust (respirable dust $< 5\,\mu m$), respiratory protection is recommended. It is not yet known whether there is a health hazard from the inhalation of cyclowollastonite dust. However, respirable wollastonite (-1T) and (-2M) dusts are suspected to be carcinogenic.

3. Alkali Silicates

3.1. Introduction

Historical Aspects. The solubility in water of a fused mixture of flint pebbles with potash was first described by the Brussels physician Jean Baptiste van Helmont (1577 – 1644) in his posthumous work "Ortus medicinae", Amsterdam 1648.

Independently, Johann Rudolph Glauber (1604 – 1670) studied the properties of a "thick" solution that he obtained from a fused mixture of sand, flint pebbles, or crystalline quartz with potash, and named "Liquor Silicum". He recommended its use as a curative agent, for the production of liquid fluxes in metal smelting, and for glazing earthenware vessels (Furni Novi Philosophici, Amsterdam 1648 – 1650).

In 1768, during the course of his alchemical studies, Goethe became interested in potassium silicate [116].

Although there were many later reports, the water-soluble silicates were seldom put to practical use until about 1825 when Johann Nepomuk von Fuchs in Munich investigated the industrial production of water-soluble potassium and sodium silicates, which he named "waterglasses" (i.e., water-soluble glasses), and proposed that their solutions might be used as adhesives, for sealing porous stone, and as binders for fresco painting [117]. He exchanged ideas with Justus von Liebig and C. F. Kuhlmann. The latter first produced waterglass in France in 1841 [118, 119]. Soon afterwards, similar work was undertaken in England by W. Gossage, and in various parts of the United States [120].

Waterglass was more widely used during the course of the American Civil War (1861 – 1865), when the northern states utilized it in soap manufacture as a replacement for rosin formerly obtained from the southern states.

Definitions. Commercial alkali silicates are generally specified according to:

1) *Source of alkali* (soda, potash, lithia).

2) *Ratio of silica to alkali-metal oxide.* The usual ratio given is the weight (wt %) ratio or modulus, R_W. The equivalent molar ratio, R_M, can be more useful, particularly when comparing ratios of potassium and sodium silicates. For sodium silicates, the conversion factor is $R_M = 1.032\,R_W$, while for potassium silicates, $R_M = 1.568\,R_W$.
3) *Water content of the silicate.* For solid, crystalline silicates, this corresponds to the water of crystallization. For example,sodium metasilicate, $Na_2O \cdot SiO_2$, can be prepared in anhydrous form, or as the pentahydrate or nonahydrate [121]. Amorphous alkali silicate glasses are essentially anhydrous. The water content of silicate solutions is customarily defined indirectly by giving the density of the solution, which together with the silica to alkali-metal oxide ratio, defines a unique composition for the silicate solution.

Solid alkali-metal silicates used in industry are divided into two groups:

1) Water-soluble silicates solidified as a glass from a melt, normally containing ca. 1.5–4 mol SiO_2/mol M_2O (Section 3.3). These silicates and their aqueous solutions are commonly referred to as waterglass (solid glass or liquid glass). The most common commercial products are concentrated waterglass solutions, which are usually produced by dissolution of silicate glasses in water or by direct dissolution of sand in sodium hydroxide (see Section 3.4).
2) Solid, crystalline sodium silicates, often with water of crystallization. Commercial products of this type are known for $Na_2O \cdot n\,SiO_2$, where $n = 0.5$ (orthosilicate), 1.0 (metasilicate), and 2.0 (disilicate) (see Section 3.6).

3.2. Raw Materials

The dominant silica component used in industry is quartz sand (grain size 0.1–0.5 mm) which has been washed to remove clays and other impurities. The most problematic impurity in such sands is iron, normally present to the extent of ca. 300 ppm. To obtain extremely pure alkali silicates, mainly for scientific use, very finely divided, very pure silica (e.g., fume silica such as Aerosil) is recommended. The alkali-metal oxides are provided in the form of their carbonates or hydroxides. The purity of the starting materials is dictated by the purity requirements for the end products.

An appropriate sulfate may also be used as a cheap source of alkali in silicate production, but requires the presence of a reducing agent (normally coal dust). Very little sulfate-derived glass is produced because the melting process is longer and this, together with the presence of sulfur oxide gases, increases corrosion of the furnace walls. In addition, residual alkali sulfate and ash from the coal result in a relatively impure product.

3.3. Amorphous Anhydrous Alkali Silicates (Solid or Lump Glasses)

Properties. Sodium and potassium silicate glasses are prepared commercially as glass lumps or powders. They are normally colored owing to the presence of impurities, the most usual colors being blue to green or yellow-brown caused by the presence of di- or trivalent iron, respectively.

Alkali silicate glasses with compositions close to those of stoichiometric compounds (e.g., $Na_2O \cdot 2\,SiO_2$, $K_2O \cdot 2\,SiO_2$, $K_2O \cdot 4\,SiO_2$) can be made to crystallize by holding them just below their melting points for long periods. In general, however, commercial alkali silicates are not stoichiometric compounds. Instead, their molar ratio, SiO_2/M_2O, usually fluctuates around average values at which stable solutions can be obtained (Table 17). The vast majority of commercially produced alkali silicate glass is sodium silicate. Furthermore, sodium silicate liquors with a wide range of R_W values may be produced by blending two different liquors with each other or with alkali hydroxide. Virtually all commercial glass production is of sodium silicate of $R_W = 2.0$ (alkaline glass) or $R_W = 3.3$ (neutral glass) [122]. A smaller amount of potassium silicate glass is manufactured and is normally restricted to the ratios $R_W = 2.5$ and $R_W = 2.0–2.1$. Lithium silicate glasses are not readily watersoluble and hence are not produced on a commercial scale [123].

Table 17. Typical, commercially available alkali silicate glasses

Product name	Na_2O, wt %	K_2O, wt %	SiO_2, wt %	Weight ratio	Mol ratio
Pyramid AL	33.2		66.3	2.00	2.06
Pyramid NL	22.9		76.7	3.35	3.46
HK40		32	68	2.1	3.3
HK28/30		28	72	2.5	3.9

Table 18. Physical properties of commercial silicate solutions

Commercial name	Density at 20°C, kg/m^3	SiO_2, wt %	Li_2O, wt %	Na_2O, wt %	K_2O, wt %	Total solids, wt %	Weight ratio	Mol ratio	Viscosity at 20°C, cP	pH at 20°C
Crystal L29	1150	17.50	1.55			19.0	11.3	5.6	< 10	10.7
Crystal 52	1260	22.2		5.8		28.0	3.85	3.97	20	10.5
Crystal 79	1400	29.2		8.9		38.1	3.30	3.41	250–500	10.8
Crystal 100S	1500	31.2		12.5		43.6	2.50	2.58	400	11.2
Crystal 100A	1500	28.1		14.0		42.1	2.00	2.06	200	11.4
Pyramid K53	1260	21.3			8.6	29.9	2.48	3.89	180	10.9
Pyramid K66	1330	23.3			11.4	34.8	2.05	3.21	30	11.8
Pyramid K120	1600	30.8			21.6	52.4	1.43	2.24	200	12.6

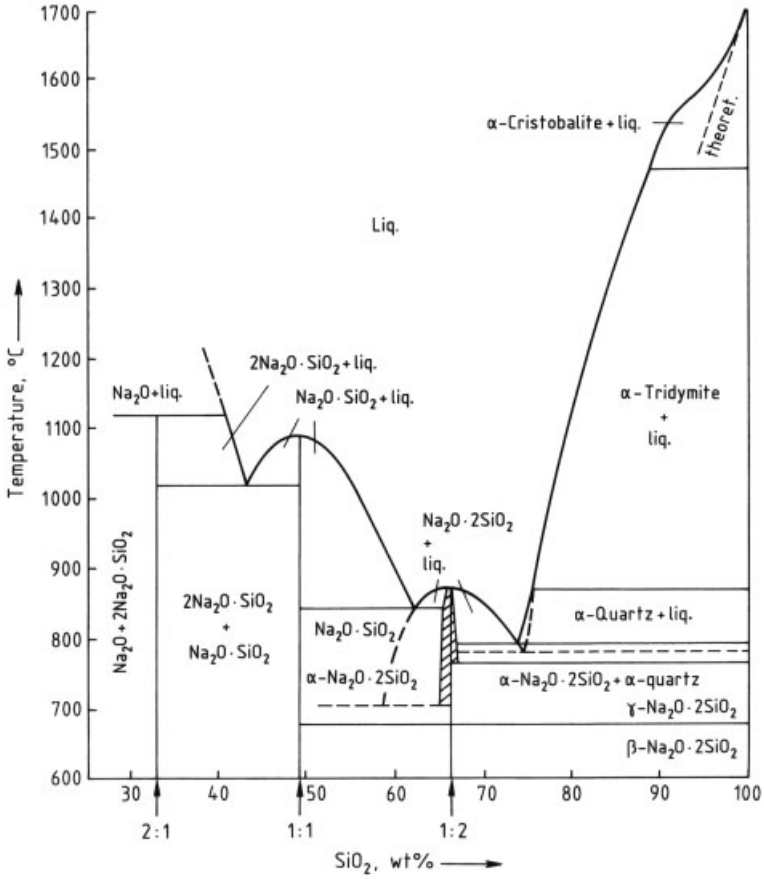

Figure 31. $SiO_2 - Na_2O$ system

The structure of silicate glasses is based on tetrahedral SiO_4 groups that are associated as polysilicate anions by sharing corners of the tetrahedra. The environment of a single Si atom is referred to as Q^n, where n is the number of shared corners for the tetrahedron [124]. The structures of the glasses are discontinuous, with SiO_2-rich domains separated by clusters of M^+ ions.

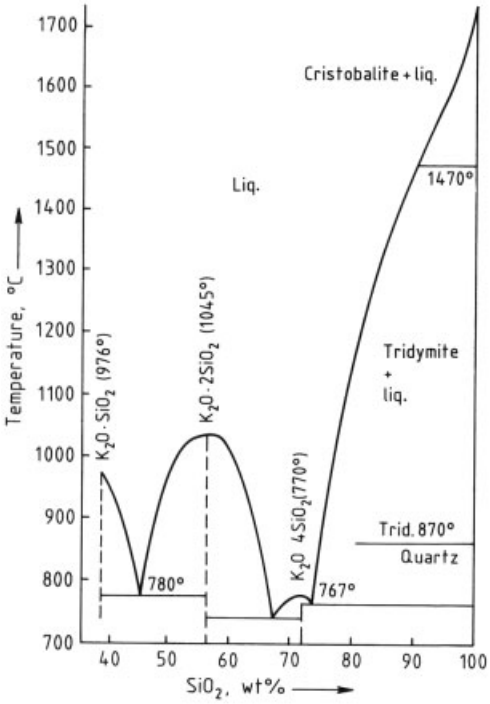

Figure 32. $SiO_2 - K_2O$ system

The melting points of the solid silicates may be obtained from the liquidus diagrams of the systems $SiO_2 - Na_2O$ and $SiO_2 - K_2O$ (Figs. 31 and 32).

Sodium silicates with $R_W \geq 3$ are only slightly attacked by water at room temperature. Sodium silicates with $R_W = 2$ are hygroscopic and cake on storage in moist air. At equivalent molar ratios, potassium silicate glasses are more hygroscopic than their sodium counterparts – potassium silicate of $R_W = 2.0$ ($R_M = 3.1$) also cakes in contact with moist air. Addition of a large excess of hot water to silicate glass causes hydrolysis and leaches out some of the alkali [125]. This alkali then attacks the residual silica, causing dissolution. Solid silicates can be dissolved at temperatures in the region of the boiling point in concentrated circulating solution, but silicate solutions are almost invariably produced by operating at pressure.

Production. Mixtures of pure silica sand (glass sand) and alkali carbonate in the required ratio are continuously fed by a cooled screw conveyor into furnaces that are regeneratively heated by oil or gas. The raw material ratio must allow for losses of alkali carbonate due to vaporization and attack on the material of the furnace. The temperature of the gas space is ca. 1600 °C, the final melt temperature being 1300 – 1500 °C. The carbon dioxide is driven off by reaction of the metal carbonate with the silica:

$$M_2CO_3 + nSiO_2 \longrightarrow M_2O \cdot nSiO_2 + CO_2$$

The furnaces are all of similar construction and size (capacity 150 t) to those used in the glass industry (→ Glass, Chap. 5.2.). They must be lined with high-quality refractory bricks (e.g., sillimanite) since the alkaline melts are very aggressive. The melt flows through the furnace and leaves continuously via an overflow into small molds which are moved along in an endless chain. The cast pieces (diameter ca. 8 cm, thickness 1 – 2 cm) cool as they are transported and often shatter.

Silicates are also produced in a batch process in small furnaces or continuously in rotary furnaces.

The cast blocks of silicate glass are dissolved in water while still warm or after intermediate storage (see Section 3.4). For some applications, they are ground and sieved.

3.4. Silicate Solutions

Properties. Solutions of sodium and potassium silicates with $SiO_2/M_2O \geq 1.5$ are colorless, water white, and viscous. The viscosity increases with concentration and, for a given constant silica content, with the SiO_2/M_2O ratio. When a critical concentration is reached, it rises very rapidly with further concentration (see Fig. 33). For a given ratio there is a limiting concentration above which the solution becomes either too viscous or too unstable for industrial use.

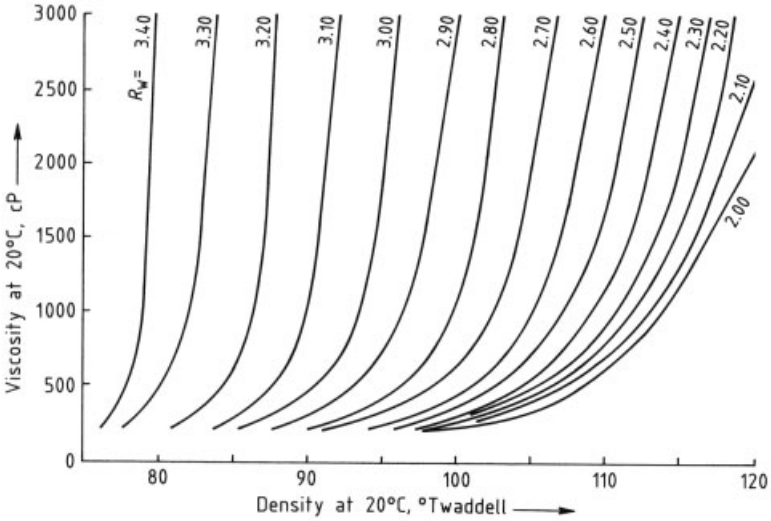

Figure 33. Change of viscosity with density at constant $SiO_2 : Na_2O$ ratios

The structures of the silicate solutions have been probed by a wide variety of techniques such as IR/Raman spectroscopy, light scattering, and chromatography [126–130]. The most valuable techniques are, however, trimethylsililation and ^{29}Si-NMR [129, 131–133]. Alkali silicate solutions contain, in addition to Me^+, OH^-, and orthosilicate ($H_xSiO_4^{(4-x)-}$) ions, a wide variety of linear, cyclic, and highly cross-linked polysilicate ions. The degree of polymerization of the silicate ions increases with increasing concentration and increasing ratio of the solution. Relatively small polysilicates (≤ 8 Si atoms) can be identified by comparing their silylated derivatives with derivatives of known structures from silicate minerals [134]. In addition, ^{29}Si-NMRalso allows identification of oligomeric silicate ions in dilute solutions [131–133]. In more concentrated solutions, however, and particularly at high ratios, NMR signals of individual species cannot be observed, peaks in the NMR spectrum then correspond to the Q^n species (see Section 3.3) [135]. This is illustrated in Figure 34, which shows ^{29}Si-NMR spectra for sodium silicates of $R_W = 2.0$ and 3.3 at typical commercial concentrations. The spectrum of the material with $R_W = 2.0$ is dominated by silicons of the Q^2 type, indicating a preponderance of small cyclic or linear silicate ions; in the material with $R_W = 3.3$, on the other hand, more than 50 % of the silicon atoms are present as Q^3 or Q^4 species, indicating a higher level of three-dimensional, branched polysilicates. These large polysilicate anions are mainly responsible for the high viscosity of silicate solutions. The silicate anions are in dynamic equilibrium and addition of alkali hydroxide (tantamount to reducing the ratio) causes hydrolysis of the polysilicates, thus reducing the viscosity, although even highly alkaline silicate solutions still contain polysilicates. Extensive dilution with CO_2-free water also depolymerizes the polysilicate anions. A wide variety of silicate solutions are available commercially. Sodium silicates are marketed with R_W ranging from 1.5 to 3.85, while potassium silicates are available with R_W from 1.43 to 2.48. Lithium silicates are also available commercially. Table 18 gives physical data for a few typical commercial silicate solutions, while Table 19 lists typical composition ranges for a commercial sodium silicate ($R_W = 3.3$).

All silicate solutions are highly alkaline. Dilution reduces the pH, but less than might be anticipated owing to the buffering action of the silicate. For example, dilution of Crystal 79 ($R_W = 3.3$, 25 wt % silica) to a 1 wt % solution reduces the pH from 11.6 to 10.8.

Addition of acids and acid salts (including carbon dioxide and bicarbonates) to silicates liberates silica. This is the basis of the industrial preparation of silica sols, silica gels, and precipitated silicas, the type of silica being formed depending on the concentration, temperature and

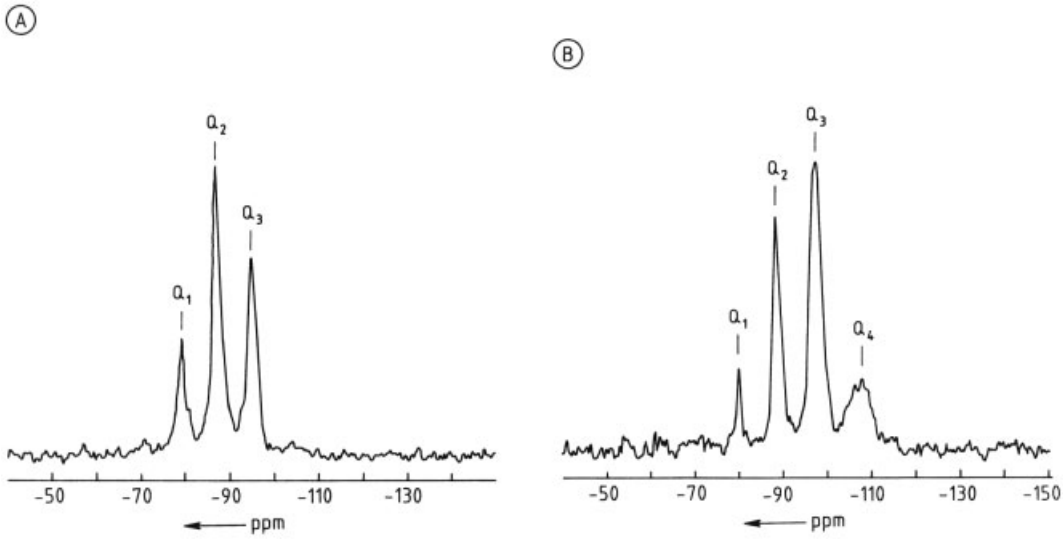

Figure 34. ^{29}Si-NMR spectra for sodium silicates
A) Crystal 100 A; B) Crystal 79

salt content of the solution (→ Silica) [122, 136]. Salts of polyvalent cations yield mixtures of silica gel with metal hydroxide [123]. Well-defined silicates are usually formed only after prolonged heating of the solution, sometimes under pressure. Under these conditions, industrially valuable crystalline silicates such as zeolites (→ Zeolites) or clays (e.g., kaolin) may be formed [136].

The addition of saturated sodium chloride or sodium nitrate solutions to silicate solutions precipitates alkali-containing silica gels which redissolve on dilution [121]. Saturated sodium sulfate and sodium carbonate solutions can be mixed with silicate solutions at room temperature without causing precipitation, as can strongly alkaline salts such as trisodium phosphate and sodium metaborate.

Most organic compounds are incompatible with soluble silicate solutions. Some polyhydric alcohols (e.g., glycerol) are compatible with sodium silicates and yield stable single-phase solutions. Other organics, such as polyol esters, produce a homogeneous system with silicate solutions but slowly react to give a solid mass [137]. Such systems have found use as setting systems (e.g., in the foundry industry and soil consolidation). Heating of sodium and potassium silicates in the absence of carbon dioxide initially produces very viscous solutions; as water loss continues, solid foams are formed that still contain significant amounts of water. Lithium silicates are stable at and around room temperature, but precipitate on heating to > ca. 60 °C. The effect is reversible upon cooling [138].

Table 19. Composition range of typical commercial sodium silicate solution ($R_W = 3.3$)

Assay	Content ∗
SiO_2	29.2 wt %
Na_2O	8.9 wt %
Li	0.2 – 0.5
K	20 – 50
Mg	5 – 20
Ca	1 – 80
Sr	1 – 5
Ba	< 1 – 5
Al	50 – 200
P	< 1 – 10
S	10 – 30
Ti	30 – 80
V	0.1 – 0.8
Cr	< 1
Mn	< 0.5 – 1
Fe	25 – 100
Co	< 1
Ni	< 0.5
Cu	< 0.1 – 0.2
Zn	< 0.2 – 1.0
La	0.2 – 1.0
Ce	< 0.3 – 2.0
Zr	5 – 20
W	< 1 – 25

∗ ppm unless otherwise stated.

Production. Cold or warm lump glass (the latter to improve heat conservation) is usually dissolved at ≥ 5 bar and ca. 150 °C. Stationary systems may be used, but rotating vessels or circulation of the liquor accelerate dissolution rates. Direct steam heating is usually employed in large ($\geq 15\,m^3$) vessels. High glass : water ratios are used to give a satisfactory dissolution rate. Since the alkali silicates do not have a definite solubility in water, prolonged heating can lead to setting of the glass – liquid mix into an intractable elastic mass [122]. An excess of solid glass is heated with water until the required concentration is reached. The silicate solution, containing solid impurities (silicates of, for example, iron, titanium, alkaline earths) is removed from the residual excess glass. The solutions are then clarified by settling at elevated temperatures or by filtration. Some commercial silicate solutions are too viscous to be handled in this way, and are produced in a more dilute form, followed by evaporation of water under reduced pressure.

A variety of molar ratios and concentrations can be produced by appropriate mixing of a small number of silicate solutions produced directly from glass, either with each other or with the appropriate alkali hydroxide.

Sodium silicate solutions of molar ratio ≤ 2.5 can be prepared by direct dissolution of quartz sand in sodium hydroxide solution under pressure at $\leq$ ca. 150 °C. Even higher ratios can be obtained by this method if a more soluble form of silica is substituted for the quartz sand. Sodium silicates with R_W up to 3.3 may be obtained by this method using silica sources such as amorphous silica, diatomite, or cristobalite [123, 139].

Processes of this kind are likely to increase in importance because they require less energy than glass formation followed by dissolution.

As lithium silicate glasses are insoluble, lithium silicate solutions are prepared commercially by dissolution of amorphous silica in lithium hydroxide. Owing to the temperature instability of lithium silicate solutions, this process is carried out at or near room temperature.

3.5. Hydrated Water-Soluble Silicates

Hydrated water-soluble silicates are normally manufactured in the sodium form and are available as fine powders or as granules. In general they are prepared by evaporating silicate solutions to the point where solids are formed which are sufficiently stable to allow commercial handling but which retain adequate solubility. Typically this occurs at water contents of ca. 20 %. In principle a wide range of products with different ratios can be made, although in practice the ratio tends to be restricted to the range $R_W = 2.0$ – 3.3 where the products are amorphous. Unlike finely divided silicate glasses, these materials dissolve in water under normal conditions to give silicate solutions; the rate of dissolution is a function of ratio, water content, and particle size. Thus at low ratios ($R_W = 2.0$), materials in the size range 100 – 200 µm dissolve completely at room temperature in a few minutes, whereas a similar product with a higher ratio ($R_W = 3.3$) requires a higher temperature (≥ 50 °C) to achieve the same dissolution rate. These are, however, both considerably faster than the anhydrous glass analogues, for which dissolution takes many hours and may not go fully to completion.

In dissolving hydrated water soluble silicates, too large an excess of water should be avoided since rapid leaching of the alkali hydroxide can give a solution pH that is too low to attack the residual silicic acid, thus leaving it as an insoluble residue.

Examples of commercial materials are listed in Table 20. The manufacturing processes employed include spray or drum drying of silicate solutions followed by sieving to the required size. Granular materials are formed by mechanical compaction of the spray dried powders to give products which are more easily handled and are less dusty.

3.6. Crystalline Solids

Numerous crystalline alkali silicates have been reported both in anhydrous and hydrated forms [121]. The ratios vary typically from $R_W = 0.5$ to $R_W = 2.0$ and they are often referred to by the more traditional names:

Common name	R_W
orthosilicate ($SiO_2 \cdot 2Na_2O$)	0.5
sesquisilicate ($2SiO_2 \cdot 3Na_2O$)	0.67
metasilicate ($SiO_2 \cdot Na_2O$)	1.0
disilicate ($2SiO_2 \cdot Na_2O$)	2.0

Table 20. Hydrated, water soluble sodium silicates

Commercial name	Na_2O, wt %	SiO_2, wt %	Weight ratio	Molar ratio	Water, wt %	Bulk density, g/L
Pyramid P 40	26.7	53.3	2.00	2.06	20.0	400 – 500
Pyramid P 60	18.5	61.5	3.30	3.41	20.0	550 – 650
Pyramid P 10	28.0	56.0	2.00	2.06	16.0	80 – 120
Pyramid G 80	26.0	52.0	2.00	2.06	22.0	750 – 850
Pyramid G 90	26.0	52.0	2.00	2.06	22.0	800 – 950

In addition, several higher ratio silicates have been discovered and synthesized with ratios such as 4 : 1, 8 : 1, 9.4 : 1, 14 : 1, and 22 : 1 [140, 141]. Many of the latter are complex layered structures, often occurring naturally and having some cation exchange properties. The structured principles of all of these crystalline silicate species have been reviewed [142].

The only crystalline silicates which are presently of any commercial significance are the metasilicate family represented by the general formula $Na_2O \cdot SiO_2 \cdot nH_2O$. Products are available with $n = 0$, 5, 8 and 9, and within this range the anhydrous silicate ($n = 0$) and the pentahydrate ($n = 5$) are the most important. Metasilicate and its hydrates differ in structure. Whereas the anhydrous material contains infinite chains of general formula Na_2SiO_3, the hydrates, more properly represented by the formula $Na_2[SiO_2(OH)_2](n-1)H_2O$, contain the $[SiO_2(OH)_2]^{2-}$ ion as a discrete entity.

The metasilicates crystallize well and are readily soluble in water. Anhydrous sodium metasilicate melts congruently above 1018 °C, and above 70 °C it is the solid phase in equilibrium with aqueous solutions. Below this temperature, the hydrolyzed ion $[SiO_2(OH)_2]^{2-}$ is formed and the various hydrates can be crystallized. Although potassium metasilicates behave similarly, they are not commercial products.

Production. Anhydrous metasilicate is produced by two general methods:

1) It can be prepared in a furnace by direct fusion of sand and sodium carbonate at temperatures above the melting point of metasilicate (1088 °C), followed by solidification, grinding, and sieving. In some cases a solid-state reaction has been used above the melting point of sodium carbonate (850 °C) but below that of the metasilicate in rotary or tunnel kilns. The corrosive nature of metasilicate melts on furnace refractories makes this a less attractive route.
2) Alternatively, it can be obtained by drying solutions at the metasilicate composition in a fluidized bed [143] or a drum granulator [144]. In the latter case the hot liquor is sprayed on to the powder bed whose particles grow and dry uniformly in a current of hot air. At the end of the drum the beaded product is screened and the fines recycled to provide powder for further growth.

Metasilicate hydrates can be obtained by direct hydration of anhydrous metasilicate or by crystallization of a solution of metasilicate with the appropriate composition. In the case of metasilicate pentahydrate for example the heated solution of composition $Na_2O \cdot SiO_2 \cdot 5H_2O$ is kept above 72.2 °C (its melting point) and is crystallized by spraying on to a bed of particles cooled to below this temperature in a rotary drum granulator using a similar procedure to that described above for the anhydrous metasilicate.

Typical products and properties of metasilicates are shown in Table 21.

Of particular interest lately has been the preparation of synthetic layered silicate structure [145]. These materials of general composition $Na_2O \cdot 2\,SiO_2$ have ion-exchange properties.

3.7. Uses and Applications

The uses of alkali metal silicates fall into two broad categories:

1) As raw materials from which further important industrial products can be derived. This application utilizes the silicate anions as building blocks for the formation of a range of silica- containing materials and represents the single largest use of soluble silicates. These derived products include silica

Table 21. Typical properties of metasilicate products

Commercial name	Chemical formula	Water content, wt %	Bulk density, g/L	Average particle size, mm
Metso A 100	$Na_2O \cdot SiO_2$		1200	0.8
Metso 510	$Na_2O \cdot SiO_2 \cdot 5\,H_2O$	43	1000	0.8
Metso 520	$Na_2O \cdot SiO_2 \cdot 5\,H_2O$	43	1000	1.0
Metso 950	$Na_2O \cdot SiO_2 \cdot 9\,H_2O$	57	800	0.6

sols, silica gels, precipitated silicas, zeolites, aluminosilicates, magnesium silicates, synthetic clays, ceramics, and catalysts.

2) As functional additives in a wide range of applications exploiting the varied properties of the silicates.

These properties include: a source of buffered alkalinity; an inorganic polymeric system with adhesive, binder, and film-forming capabilities; nontoxic, nonflammable and environmentally benign; wide range of available compositions; capable of forming hybrid systems with organics (e.g., polymers) and other inorganics (phosphates, borates); complexation or precipitation of ions in solution; intumescent properties on heating; low cost.

Detergents represent the largest functional use of silicates. In broad terms use is made of the alkalinity and buffering capacity to aid in soil removal, assist in the suspension of soil particles, enhance the effectiveness of surfactants by sequestering calcium and magnesium ions, and inhibit corrosion of metal surfaces. In fabric washing powders [146] liquid silicates have traditionally been incorporated into the spray dried powders at levels of 5 – 15 %, whereby the silicate also provides crisp and easily handled detergent granules. The levels of incorporation of liquid silicates have been decreasing with the change to zeolite-based powders and more use is now being made of powdered soluble silicates added after spray drying. In industrial cleansing agents and dishwasher detergents the silicate product of choice has been metasilicate, either anhydrous or pentahydrate. There are now some concerns about the high alkalinity of such materials in domestic products and a move to granular soluble powders with higher ratio ($R_W \geq 2.0$) is likely to occur.

Foundries. Large quantities of liquid silicate (typically with $R_W = 2$) are used in the foundry industry to bind together sand molds and cores prior to pouring the molten metal. The silicate is set by reaction with CO_2 [147] blown through the silicate – sand mixture, or by incorporation of glycerol esters which hydrolyze and cause the setting reaction by release of organic acids [148].

Adhesives. Silicate liquors are used widely as adhesives. The largest consumer is the paper and board industry for such duties as spiral tube winding, corrugated boxboard, and fiber drums. Silicates are also used to bind insulating materials such as vermiculite in building panels, coal dust briquettes, roofing tiles, bricks and ceramics, refractory cements, and in the manufacture of welding rods. In the latter case, lithium silicate has proven to be particularly effective since it not only provides the necessary strength and degree of water resistance, it also reduces significantly the amount of Cr^{VI} evolved during the welding, thus providing safety benefits [149, 150].

Surface Coatings. Silicate liquors are used in a variety of surface coating applications. Examples include the sealing of porous surfaces such as concrete; as a vehicle for the preparation of paints for masonry and glass surfaces and for general fire proofing duties; and as a component in spray coating systems based on concrete and glass fiber. Silicates are also used to coat TiO_2 pigment particles to improve stability.

Paper Deinking and Bleaching. Sodium silicate liquors are used in the paper industry to promote deinking of recycled newsprint and also to act as a promoter to achieve enhanced bleaching with hydrogen peroxide. Both of these are environmentally driven applications, and demand is expected to grow as pressure increases to recycle more paper and to switch from chlorine to peroxide- containing bleaches.

Water Treatment. Silicates are added to water, including potable water, in concentrations

of a few ppm to prevent discoloration due to dissolved iron. The precise mechanism is not fully understood, but it appears that the silicate forms a "soluble" iron – hydroxide – silicate complex which prevents the formation of visible iron flocs. An additional benefit of such water treatment with silicate is the corrosion protection afforded to metal pipes, presumably by similar complex formation at the metal surface. Other water treatment uses have included a coagulation aid in alum flocculation processes and for boiler water treatment.

Civil Engineering. Silicates are used to stabilize soils in many civil engineering projects involving drilling, tunneling, and mining. Silicate solutions mixed with appropriate setting agents are pumped into the ground where they harden the soil [151]. An interesting application of this technology is the Sanipor process for repairing sewers without the need to dig up city roads and streets [152–154]. Other similar uses include sealing around landfill sites, waste fixation, and coastline stabilization.

Enhanced Oil Recovery. Use of silicates with considerable growth potential is in enhanced or tertiary oil recovery, particularly in offshore locations. Flooding the oil field with silicate solutions promotes the formation in situ of soaps and surfactants which improve the oil flow from the porous rock [155, 156].

3.8. Economic Aspects

Estimated 1990 world silicate consumption (expressed as dry glass) is ca. 4×10^6 t (including captive use). The approximate geographic split is United States, Japan and Western Europe 20 % each with the remaining 40 % for the rest of the world. The latter figure is speculative since it does not include Eastern Europe and China. The sales value of the noncaptive volume is ca. $\$1.5\times10^9$. Consumption is estimated at 65 – 75 % of installed capacity.

The volume of silicate supplied to the merchant market, expressed as 100 % solids, is estimated to be:

United States	470 000 t
Western Europe	500 000 t
Japan	500 000 t

Table 22 shows the breakdown of the total volume by application for Western Europe and the United States.

Table 22. Silicate usage by industry application (in percent)

Application	Western Europe	United States
Detergents	28.1	32.6
Paper	7.3	7.5
TiO_2	4.6	4.6
Water treatment	2.9	3.5
Civil engineering	7.7	
Miscellaneous	6.4	5.8
Total	57	54
Derivatives (silica, zeolites etc)	43	46
Total	100	100

3.9. Storage, Safety, Labelling and Transportation

The CAS und EINECS numbers are shown in Table 23 for various soluble silicates.

A comprehensive review of the health and safety aspects of silicates has been published [157, 158]. The primary hazard of all soluble silicates is their alkalinity. Contact effects range from irritation to corrosion. Ingested or inhaled silicates are rapidly eliminated in the urine. Quoted LD_{50} (oral, rat) values are 1280 mg/kg for sodium metasilicate, 1300 mg/kg for sodium silicates with $R_W = 2.0$, and 1600 mg/kg for $R_W = 3.0$ according to the Registry of Toxic Effects of Chemical Substances 1981 – 1982. There are no exposure limits quoted for sodium or potassium silicates by any regulatory body. A limit of 2 mg/m^3 is given for caustic soda and caustic potash by both ACGIH and the HSE (under the EH40/92 exposure regulations) and exposure to silicate should therefore be limited, according to its composition, to remain in line with the caustic soda and potash figures. For lithium silicates a similar situation prevails, but the exposure limit for lithium hydroxide is 1 mg/m^3.

Alkali silicates are not flammable or explosive. Due to their alkalinity, solutions attack aluminum, tin, and zinc, with evolution of hydrogen, at concentrations above ca. 10 %. Powdered silicates tend to be hygroscopic and packaging must therefore provide adequate protection against water vapor ingress to avoid caking. Avoidance of prolonged contact with CO_2

Table 23. CAS and EINECS numbers for various soluble silicates

Product	Formula	CAS registry no.	EINECS no.
Sodium silicate solutions, glasses, or solids	$Na_2O(SiO_2)_x$ $x > 1.6$	[*1344-09-8*]	2 156 874
Sodium metasilicate and hydrate	$Na_2O \cdot SiO_2(H_2O)_y$ $y = 0, 5, 9$	[*6834-92-0*]	2 299 129
Sodium orthosilicate	$2Na_2O \cdot SiO_2$	[*13472-30-5*]	2 367 413
Sodium sesquisilicate	$3Na_2O \cdot 2SiO_2$	[*15593-82-5*]	2 396 711
Potassium silicates	$K_2O(SiO_2)_x$ $x > 1.6$	[*1312-76-1*]	2 151 991
Lithium silicate solutions	$Li_2O(SiO_2)_x$ $6 \leq x \leq 12$	[*12627-14-4*]	2 357 300

is also necessary since both solids and solutions will readily absorb this from the air to form alkali metal carbonates.

Under the rules of the EEC Dangerous Substances Directive 67/548/EEC as amended by 92/32/EEC all silicates with $R_W \geq 1.6$ are classified as irritant. However, the labeling classification for sodium silicates is under review in Europe and a new set of safety phrases will be agreed taking into account both R_W and concentration.

In the environment, commercial silicates rapidly depolymerize on dilution to give molecular species indistinguishable from natural dissolved silica in ground waters. The pH of most silicate solutions is however above the acceptable limit for direct discharge to rivers and sewers, and dilution or neutralization must be carried out first. For disposal to landfill, the classification of silicates varies according to composition and also by countries. In general the more alkaline the products the more likely it is to be classified as hazardous or special waste although concentration is also important. Thus in the United Kingdom metasilicates above 0.1 % and all silicate solutions above 10 % qualify as special waste [159].

3.10. Analysis

The alkali-metal content of soluble silicates is determined by titration in dilute solution with hydrochloric or sulfuric acid using methyl orange as indicator. The silica is determined gravimetrically. Alkali metals are removed as the chlorides by fuming the diluted silicate solution several times with hydrochloric acid in a flat porcelain dish. The insoluble silica is then filtered, washed, calcined, and then weighed. If very great accuracy is required, the silica is volatilized by fuming with hydrofluoric and sulfuric acids. The sesquioxides remaining after this are accounted for in the calcination.

Potassium silicate glasses are dissolved and fumed with hydrochloric acid to give silica-free filtrates whose sodium content is determined by flame photometry. The potassium content of sodium silicate glasses is determined analogously.

The silica content of aqueous silicate solutions can be determined more conveniently by titration. The solution is neutralized using methyl red as indicator, an excess of pure sodium fluoride is then added, and the alkali that is liberated according to the equation

$$SiO_2 + 6\,F^- + 2\,H_2O \longrightarrow SiF_6^{2-} + 4\,OH^-$$

is backtitrated with hydrochloric acid.

Chloride is titrated by the argentometric method in a dilute aqueous solution of the silicate. Sulfate is determined by precipitation as barium sulfate from the filtrate from which the silica has been removed.

Other possible impurities (CO_2, Al_2O_3, Fe_2O_3, TiO_2, CaO, MgO) are determined by the methods used for analyzing ordinary glass. Spectrophotometry is recommended, particularly for iron and titanium.

Alkali silicates that are normally insoluble in water are dissolved by pressure dissolution and investigated as for soluble silicates. Remaining insoluble matter is weighed and may be analyzed as for water-insoluble silicates.

Since 1972, the following ISO specifications for alkali silicate analysis were published, 1686: general, 1687: density of silicate solutions, 1688: density of the dry substance, 1689:

determination of ratio, 1690: gravimetric determination of SiO_2 content, 1691: gas volumetric determination of CO_2 content, 1692: titration of alkalinity, 2122: preparation of solutions of solid alkali silicates and determination of insoluble components, 2123: determination of dynamic viscosity of silicate solutions, 2124: titrimetric determination of SiO_2 content, 3200: gravimetric determination of SO_4^{-2} content of alkali silicates, 3201: photometric determination of iron content of alkali silicates.

4. References

1. F. Liebau: *Structural chemistry of silicates (structure, bonding, classification),* Springer Verlag, Berlin 1985.
2. F. Liebau: "Die Systematik der Silicate," *Naturwissenschaften* **49** (1962) 481 – 491; "Silicon" in K. H. Wedepohl (ed.): *Handbook of Geochemistry,* vol. II/3, chap. 14-A, Springer Verlag, Berlin 1972, pp. 1 – 32; "Classification of Silicates," in P. H. Ribbe (ed.): *Orthosilicates, reviews in mineralogy,* vol. 5, Min. Soc. Am., 1980, pp. 1 – 24.
3. F. Liebau: "Silicates with Branched Anions: a Crystallochemically Distinct Class," *Am. Mineral.* **63** (1978) 918 – 92.
4. F. Liebau: "The Influence of Cation Properties on the Shape of Silicate Anions," *Z. Kristallogr.* **155** (1981) 139 – 153.
5. M. Fleischer: "Natrosilite," *Am. Mineral.* **61** (1976) 339 – 340.
6. M. Th. Le Bihan, A. Kalt, R. Wey: "Étude structurale de $KHSi_2O_5$ et $H_2Si_2O_5$," *Bull. Soc. Fr. Mineral. Cristallogr.* **94** (1971) 15 – 23.
7. F.-J. Dany et al.: "Kristallines Schichtsilikat – ein neuer Builder," *Seifen öle Fette Wachse* **116** (1990) 805 – 808.
8. Z. Johan, G. F. Maglione: "La kanemite, nouveau silicate de sodium hydraté de néoformation," *Bull. Soc. Fr. Mineral. Cristallogr.* **95** (1972) 371 – 382.
9. R. A. Sheppard, A. J. Gude, R. L. Hay: "Makatite, a new Hydrous Silicate Mineral from Lake Magadi, Kenya," *Am. Mineral.* **55** (1970) 358 – 366.
10. H. P. Eugster: "Hydrous Sodium Silicates from Lake Magadi, Kenya: Precursors of Bedded Chert," *Science (Washington D.C.)* **157** (1967) 1177 – 1180.
11. M. Fleischer: "Revdite," *Am. Mineral.* **67** (1982) 1076.
12. J. Puziewicz: "Grumantite", *Am. Mineral.* **73** (1988) 440.
13. G. Y. Chao, J. D. Grice, R. A. Gault: "Silinaite, a new Sodium Lithium Silicate Hydrate Mineral from Mont Saint-Hilaire, Quebec," *Can. Mineral.* **29** (1991) 359 – 362; J. D. Grice: "The Crystal Structure of Silinaite, $NaLiSi_2O_5 \cdot 2H_2O$: a Monophyllosilicate," *Can. Mineral.* **29** (1991) 363 – 367; K. Benneke, P. Thiessen, G. Lagaly: "Synthesis and properties of sodium lithium silicate slinaite," *Inorg. Chem.* **34** (1995) 900 – 907.
14. G. Scholzen, K. Beneke, G. Lagaly: "Diversity of Magadiite," *Z. Anorg. Allg. Chem. 597* (1991) 183 – 196.
15. L. S. Dent Glasser, P. B. Jamieson: "Sodium Silicate Hydrates. The Structure of $Na_2O\ SiO_2 \cdot 8H_2O$," *Acta Crystallogr. Sect.* **B 32** (1976) 705 – 710.
16. H. Annehed, L. Fälth, F. J. Lincoln: "Crystal Structure of Synthetic Makatite $Na_2Si_4O_8(OH)_2 \cdot 4H_2O$," *Z. Kristallogr.* **159** (1982) 203 – 210.
17. G. Bergély et al.: "Chemical Characterization, Structural Features, and Thermal Behavior of Sodium Hydrogen Octosilicate," *Clays Clay Miner.* **39** (1991) 490 – 497.
18. K. Beneke, G. Lagaly: "Kanemite – Intracrystalline Reactivity and Relations to Other Sodium Silicates," *Am. Mineral.* **62** (1977) 763 – 771; "Kenyaite – Synthesis and Properties," *Am. Mineral.* **68** (1983) 818 – 826; "A Hydrated Potassium Layer Silicate and its Crystalline Silicic Acid," *Am. Mineral.* **74** (1989) 224 –229.
19. W. Schwieger, W. Heyer, F. Wolf, K.-H. Bergk: "Zur Synthese von kristallinen Metallsilicathydraten mit Schichtstruktur," *Z. Anorg. Allg. Chem.* **548** (1987) 204 – 216.
20. K. H. Bergk, P. Grabner, W. Schwieger: "Die hydrothermale Kristallisation des Magadiits," *Z. Anorg. Allg. Chem.* **600** (1991) 139 – 144.
21. K.-H. Bergk, W. Schwieger, M. Porsch: "Aluminiumfreie Schichtsilicathydrate – Synthese und Eigenschafts-Anwendungs-Beziehungen," *Chem. Tech. (Leipzig)* **39** (1987) 459 – 466, 508 – 514; K.-H. Bergk, D. Kaufmann, M. Porsch, W. Schwieger: "Herstellung und Verwendung von Magadiit als Phosphatsubstitut in Waschmitteln," *Seifen öle Fette Wachse* **113** (1987) 555 – 561.

22. R. Schwarz, E. Menner: "Zur Kenntnis der Kieselsäuren," *Ber. Dtsch. Chem. Ges.* **57** (1924) 1477–1481.
23. F. Liebau: "über Kristallstrukturen zweier Phyllokieselsäuren, $H_2Si_2O_5$," *Z. Kristallogr.* **120** (1964) 427–449.
24. J.-L. Guth et al.: "Un nouveau type de silice hydratée cristallisée," *C.R. Hebd. Seances Acad. Sci. Ser. D* **286** (1978) 5–8.
25. C. Frondel: "Crystalline Silica Hydrates from Leached Silicates," *Am. Mineral.* **64** (1979) 799–804.
26. G. Lagaly, R. Matouschek: "Die kristallinen Kieselsäuren aus Apophyllit, Carletonit und Gillespit," *Neues Jahrbuch Miner. Abh.* **138** (1980) 81–93.
27. H.-H. Kruse, K. Beneke, G. Lagaly: "Gas Adsorption by a Crystalline Silicic Acid," *Colloid Polymer Sci.* **267** (1989) 844–852.
28. J. Döring, K. Beneke, G. Lagaly: "Adsorption Properties of Crystalline Silicas: II. Adsorption of Anionic Surfactants and Delamination," *Colloid Polym. Sci.* **270** (1992) 609–616.
29. G. W. Brindley, G. Brown: *Crystal Structures of Clay Minerals and their X-ray Identification,* Mineralogical Society, London 1980.
30. Report of the Clay Minerals Soc. Nomencalture Commitee, *Clays Clay Miner.* **39** (1991) 333–335; **45** (1997) 298–300.
31. G. Lagaly, K. Jasmund (eds.): *Tonminerale und Tone. Struktur, Eigenschaften, Anwendungen und Einsatz in Industrie und Umwelt,* Steinkopff Verlag, Darmstadt 1993.
32. "Definition of Clay and Clay Minerals," *Clays Clay Miner.* **43** (1995) 255–256.
33. A. Weiss, J. Russow: "Über das Einrollen von Kaolinkristallen," in I. Th. Rosenqvist, P. Graff-Petersen (eds.): *Proc. Int. Clay Conf.,* Stockholm 1963, v. 1, 69–74.
34. W. B. Jepson: "Kaolins: their Properties and Uses," *Philos. Trans. R. Soc. London Ser. A* **311** (1984) 411–432.
35. I. E. Odom: "Smectite Clay Minerals: Properties and Uses," *Philos. Trans. R. Soc. London Ser. A* **311** (1984) 391–409.
36. K.-H. Bergk, D. Wolf, W. Schwieger, M. Porsch: "Hectorit – ein kristallines Magnesiumsilicat mit Zukunft," *Chem. Tech. (Leipzig)* **41** (1989) 245–251.
37. W. Vogel, W. Höland: "Zur Entwicklung von Bioglaskeramiken für die Medizin," *Angew. Chem.* **99** (1987) 541–558.
38. G. Lagaly, K. Beneke: "Intercalation and Exchange Reactions of Clay Minerals and Nonclay Layer Compounds," *Colloid Polym. Sci.* **269** (1991) 1198–1211.
39. A. Weiss: "Ein Geheimnis des chinesischen Porzellans," *Angew. Chem.* **75** (1963) 755–762.
40. M. Stockmeyer: "Adsorption organischer Substanzen an organophilen Bentoniten," *Z. Dtsch. Geol. Ges.* **141** (1990) 445–451.
41. M. Stockmeyer, K. Kruse: "Adsorption of Zn and Ni Ions and Phenol and Diethylketone by Bentonites of Different Organophilicities," *Clay Miner.* **26** (1991) 431–434.
42. S. Komarnini, D. M. Roy: "Shale as a Radioactive Waste Repository: the Importance of Vermiculite," *J. Inorg. Nucl. Chem.* **41** (1979) 1793–1796.
43. J. Bors: "Sorption of Radioiodine in Organo- clays and -Soils," *Radiochim. Acta* **51** (1990) 139–143.
44. L. J. Michor, T. J. Pinnavaia: "Adsorption of Chlorinated Phenols from Aqueous Solution by Surfactant-modified Pillared Clays," *Clays Clay Miner.* **39** (1991) 634–641.
45. F. Liebau: "Zeolites and Clathrasils – two Distinct Classes of Framework Silicates," *Zeolites* **3** (1983) 191–193.
46. F. Liebau: "Synthesis of Porous Tectosilicates: Parameters Controlling the Pore Geometry" in E. R. Corey, J. Y. Corey, P. P. Gasper (eds.): *Silicon chemistry,* chap. 29, Ellis Horwood, Chichester 1988, pp. 307–323.
47. F. Liebau: "Structural Similarities and Dissimilarities between SiO_2 and H_2O," in: R. A. B. Devine (ed.): *The Physics and Technology of Amorphous SiO_2*, Plenum Press, New York 1988, pp. 15–35.
48. H. Gies, B. Marler: "The Structure Controlling Role of Organic Templates for the Synthesis of Porosils in the System SiO_2/Template/H_2O," *Zeolites* **12** (1992) 42–49; H. Gies: "Clathrasils and Zeosiles: Inclusion Compounds with Silica Host Frameworks" in J. L. Atwood, J. E. D. Davies, D. D. MacNicol (eds): *Inclusion Compounds,* vol. 5, Univ. Press, Oxford 1991, pp. 1–36.
49. R. M. Barrer: *Hydrothermal Chemistry of Zeolites,* Academic Press, London 1982.
50. O. Szostak: *Molecular Sieves. Principles of Synthesis and Identification,* van Nostrand Reinhold, New York 1989.

General References

51. H. U. Bambauer, F. Taborszky, H. D. Trochim in W. E. Tröger (ed.): "Bestimmungstabellen,"

Optische Bestimmung der gesteinsbildenden Minerale, Part 1, 4th ed., E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1971.
52. H.-R. Bosse et al.: "Untersuchungen über Angebot und Nachfrage mineralischer Rohstoffe," *XIX Industrieminerale,* Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover; Prognos AG Europäisches Zentrum für Angewandte Wirtschaftsforschung Basel 1986.
53. W. E. Tröger, H. U. Bambauer, F. Taborszky, H. D. Trochim in O. Braitsch (ed.): *Optische Bestimmung der gesteinsbildenden Minerale,* Part 2, E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1967.
54. B. Coope et al.: *Industr. Minerals Directory,* Metal Bulletin Books Ltd., Surrey 1977, p. 601.
55. D'Ans-Lax, Springer-Verlag, Heidelberg 1967, pp. 1 - 217, 1 - 662, 1-681, 1-757, 3-493.
56. D'Ans-Lax, Springer-Verlag, Heidelberg 1970, pp. 1 - 216, 1 - 341, 1 - 420, 1 - 754, 1 - 1289, 3 - 492.
57. W. A. Deer, R. A. Howie, J. Zussman: "Ortho- and Ring Silicates," *Rock-Forming Minerals,* **vol. 1,** Longmans, Green and Co., London 1962.
58. W. A. Deer, R. A. Howie, J. Zussman: "Sheet Silicates," *Rock-Forming Minerals,* **vol. 3,** Longmans, Green and Co., London 1962.
59. W. A. Deer, R. A. Howie, J. Zussman: "Framework Silicates," *Rock-Forming Minerals,* **vol. 4,** Longmans, Green and Co., London 1963.
60. W. A. Deer, R. A. Howie, J. Zussman: "Disilicates," *Rock-Forming Minerals,* **vol. 1A,** Longman Scientific and Technical, London 1997.
61. W. A. Deer, R. A. Howie, J. Zussman: "Orthosilicates," *Rock-Forming Minerals,* **vol. 1A,** Longman Scientific and Technical, London 1997.
62. W. A. Deer, R. A. Howie, J. Zussman: "Disilicates," *Rock-Forming Minerals,* **vol. 1B,** Longman Scientific and Technical, London 1997.
63. W. A. Deer, R. A. Howie, J. Zussman: "Single-chain Silicates," *Rock-Forming Minerals,* **vol. 2A,** Longman Scientific and Technical, London 1997.
64. W. A. Deer, R. A. Howie, J. Zussman: "Double-chain Silicates," *Rock-Forming Minerals,* **vol. 2B,** Longman Scientific and Technical, London 1997.
65. W. A. Deer, R. A. Howie, J. Zussman: *An Introduction to the Rock-Forming Minerals,* 2nd ed., Longman Scientific and Technical Group, London 1992.
66. H. Freund: *Handbuch der Mikroskopie in der Technik* **vol. 4,** Umschau Verlag, Frankfurt/Main 1974.
67. *Gmelin,* **System-Nr. 35, Al, [A],** 35, 41, 47, 57, 69.
68. *Gmelin,* **System-Nr. 35, Al, [B],** 312, 390, 492.
69. *Gmelin,* **System-Nr. 27, Mg, [A],** Lief. 1, 92.
70. *Gmelin,* **System-Nr. 35, Al, [A],** 41.
71. *Gmelin,* **System-Nr. 35, Al, [B],** 313.
72. *Gmelin,* **System-Nr. 27, Mg, [B],** 351.
73. *Gmelin,* **System-Nr. 1 5, Si, [8],** 482.
74. *Ind. Miner. (London)* 1970, 2, 11.
75. *Ind. Miner. (London)* 1975, 7, 15.
76. *Ind. Miner. (London)* 1975, 11, 9.
77. *Ind. Miner. (London)* 1975, 12, 47.
78. *Ind. Miner. (London)* 1976, 1, 22.
79. *Ind. Miner. (London)* 1976, 9, 71.
80. *Ind. Miner. (London)* 1977, 4, 17.
81. *Ind. Miner. (London)* 1977, 5, 39.
82. *Ind. Miner. (London)* 1977, 6, 41.
83. *Industrial Minerals Directory,* 1st ed., Metal Bulletin Books, Surrey 1977, pp. 374, 565, 591, 620.
84. T. Hahn, M. J. Buerger: "The Detailed Structure of Nepheline, $KNa_3Al_4Si_4O_{16}$," *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **106** (1955) 308.
85. H. Kittel in W. A. Colomb (ed.): *Lehrbuch der Lacke und Beschichtungen,* **vol. 2,** Heenemann, Berlin 1974, p. 379.
86. O. Lückert: *Pigment + Füllstoff Tabellen,* M. u. O. Lückert, Laatzen 1980.
87. *Landolt-Börnstein,* 5th ed., p. 1259.
88. F. Machatschki: *Spezielle Mineralogie auf geochemischer Grundlage,* Springer Verlag, Wien 1953.
89. St. v. Naray-Szabo: "Die Struktur des Pollucits $CsAlSi_2O_6 \cdot x\,H_2O$," *Z. Kristallogr. Kristallgeom. Kristallphys. Kristallchem.* **99** (1938) no. 4, 277.
90. W. P. Pearson: *Structure Reports for 1969,* **vol. 34 A,** Oosthoek, Scheltma & Holkema, Utrecht 1975, p. 366.
91. P. Ramdohr, H. Strunz: *Klockmann's Lehrbuch der Mineralogie,* 16th ed., Enke Verlag, Stuttgart 1978.
92. Rheinisch Westfälisches Elektrizitätswerk Aktiengesellschaft (RWE): *Bau-Handbuch Technischer Ausbau,* 10th ed., Essen 1990.

93. H. J. Rösler: *Lehrbuch der Mineralogie,* 3rd ed., VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1984.
94. H. Strunz: *Mineralogische Tabellen,* 7th ed., Akademische Verlagsgesellschaft Geest & Portig, Leipzig 1977.
95. F. Trojer: *Die oxydischen Kristallphasen der anorganischen Industrieprodukte,* E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1963.
96. F. Trojer, K.-H. Obst, M. Münchberg: *Mineralogie basischer Feuerfest-Produkte*, **vol. 12** in V. D. Ferchette et al. (eds.): *Applied Mineralogy/Technische Mineralogie*, Springer Verlag, Wien – New York 1981.
97. L. Weber, G. Zsak: Welt-Bergbau Daten – World Mining Data, Series A, vol. 13, Bundesministerium für wirtschaftliche Angelegenheiten, Vienna 1998.
98. H. G. F. Winkler: *Petrogenesis of Metamorphic Rocks,* 4th ed., Springer Verlag, New York 1976.
99. *Winnacker-Küchler,* 2nd ed. **2,** 403.
100. J. K. Winter, S. Ghose: *American Mineralogist,* **vol. 64,** Richmond, Virginia, 1979, p. 573.

Specific References

101. "Feldspar and Nepheline Syenite – Rivals in a Flux," *Ind. Miner.* 1976 Jan., 15 – 26.
102. "Feldspar Fluxes – the Rivalry Reviewed," *Ind. Miner.* 1981 April, 21 – 45.
103. W. A. Deer, R. A. Howie, J. Zussman: *An Introduction to the Rock-Forming Minerals,* Longmans Group, London 1952, p. 361. T. G. Sahama, *Am. J. Sci.* (1952) 460.
104. *Ind. Miner.* 1977 June, 55.
105. "Vermiculite: A Market in Transition," *Ind. Miner.* 1970 Nov., 11.
106. "Vermiculite, Increased Capacity Encourages Optimism – but Construction Still the Key to Demand," *Ind. Miner.* 1977 April, 19 – 23.
107. *Ind. Miner.* 1982 no. 175, 117.
108. "Feldspar and Nepheline Syenite - At the Mercy of Glass Markets," *Ind. Miner.* 1990 Aug., 21 – 33.
109. "Olivine – Long Live the Evolution," *Ind. Miner.* 1995 Feb., 23 – 31.
110. "Sillimanite Minerals," *Ind. Miner.* 1996 Nov., 63 – 67.
111. "Synthetic Mullite – Refractories Shock Absorber," *Ind. Miner.* 1991 Feb., 39 – 45.
112. "Minerals in Lightweight Insulation – Filling the Market," *Ind. Miner.* 1991 Oct., 21 – 35.
113. "Pumice Markets – Volcanic Rise of Stonewashing," *Ind. Miner.* 1990 May, 22 – 37.
114. "Synthetic Mineral Fibres – From Rocks to Riches," *Ind. Miner.* 1986 Sept., 20 – 43.
115. "European Industrial Minerals – World Class but Under Threat," *Ind. Miner.* 1997 Aug., 55 – 59.
116. W. Goethe: *Dichtung und Wahrheit,* 1768, p. 8.
117. J. von Fuchs, *Dinglers Polytech. J.* **17** (1825) 465.
118. C. F. Kuhlmann, *Ann. Chim.* **41** (1842) 220.
119. C. F. Kuhlmann, *Ann. Chim.* **64** (1848) 289.
120. UK 762, 1854, W. Gossage.
121. J. G. Vail: *Soluble Silicates,* Reinhold, New York 1952.
122. D. Barby, T. Griffiths, A. R. Jacques, D. Pawson: *The Modern Inorganic Chemicals Industry,* Chemical Society, London 1977, pp. 320 – 352.
123. *Kirk-Othmer,* **20,** 855 – 880.
124. G. Englehardt et al., *Z. Anorg. Allg. Chem.* **418** (1975) 17.
125. T. M. El-Shamy, J. Lewins, R. W. Douglas, *Glass Technol.* **13** (1972) 81.
126. A. Marinangelli, M. A. Morelli, R. Simoni, A. Bertoluzza, *Can. J. Spec.* **23** (1978) 173.
127. A. J. Walker, N. Whitehead, *J. Appl. Chem.* **16** (1966) 230.
128. P. K. Dutta, D. C. Shieh, *Appl. Spectrosc.* **39** (1985) 343.
129. A. P. Brady, A. G. Brown, H. Huffer, *J. Colloid Sci.* **8** (1953) 252.
130. J. J. Fitzgerald, J. A. Hoffman, C. E. Nebo, *J. Chrom. Sci.* **27** (1989) 186.
131. G. Englehardt, H. Janck, D. Hoebbel, W. Wieker, *Z. Chem.* **14** (1974) 109.
132. R. K. Harris, C. T. G. Knight, *J. Chem. Soc. (Faraday II)* **79** (1983) 1525.
133. C. T. G. Knight, *J. Chem. Soc. (Dalton)* 1988, 1457.
134. C. W. Lentz, *Inorg. Chem.* **3** (1964) 574.
135. I. L. Svensson, S. Sjoberg, L. O. Ohlmann, *J. Chem. Soc. (Faraday I)* **82** (1986) 3635.
136. R. K. Iler, *The Chemistry of Silica,* John Wiley, New York 1979.
137. M. Roberts, *Foundry Trade J.* **133** (1972) no. 2925, 783.
138. J. M. Ordway, *Am. J. Sci.* **24** (1907) (ser. 4), 473.
139. R. Novotny, A. Hoff, J. Schurtz, DE 3 902 754, 1990.
140. J. H. Wills, *J. Am. Chem. Soc.* **72** (1952) 5369.
141. L. McCulloch, *J. Am. Chem. Soc.* **74** (1950) 2453.
142. D. Barby et al.: *The Modern Inorganic Chemicals Industry,* Royal Soc. Chemistry, London 1977, p. 334.

143. S. L. Bean, A. W. Mouton, US 3 748 103.
144. C. L. Baker, P. W. Holloway, US 3 208 822.
145. H. P. Rieck, US 4 585 642.
146. R. S. Schreiber, *ACS Symp. Ser.* **194** (1982) 271.
147. K. E. L. Nicholas: *The CO_2 Silicate Process in Foundries,* BCIRA, 1972.
148. M. Roberts, *Foundry Trade Journal* **133** (1972) 783, 2925.
149. T. Griffith, *Welding Review* 1989 .
150. T. Griffith, EP 342 036.
151. R. H. Karol: *Chemical Grouting,* Marcel Dekker, New York 1990.
152. T. Szekely et al., US 4 988 238.
153. I. Clarke, *NoDig International* 1993, Feb., 8.
154. D. Bradshaw, *Financial Times,* 1990, Oct. 12., p. 35.
155. L. Holm, S. D. Robertson, *J. Pet. Tech.* **30** (1978) no. 1, 161.
156. W. L. Schleyer, J. G. Blumberg, *ACS Symp. Ser.* **194** (1982) 187 – 227.
157. W. L. Schleyer, J. G. Blumberg: "Health and Environmental Aspects of Soluble Silicates," *ACS Symp. Ser.* **194** (1982) 49 – 69.
158. W. L. Schleyer, J. G. Blumberg: "Current Regulatory States of Soluble Silicates," *Am. Chem. Soc.* 1982, 31 – 47.
159. Waste Management Paper no. 23, Department of Environment, London 1981.

Silicon

Werner Zulehner, Wacker-Chemitronic GmbH, Burghausen, Federal Republic of Germany (Chaps. 1 – 4)

Bernd Neuer, SKW Trostberg AG, Federal Republic of Germany (Chaps. 5 – 7)

Gerhard Rau, SKW Trostberg AG, Federal Republic of Germany (Chaps. 5 – 7)

1. Introduction 1
2. History 2
3. Physical and Chemical Properties . 2
4. Production 3
4.1. Metallurgical Silicon 3
4.2. Semiconductor Silicon 5
4.2.1. Purification 5
4.2.2. Chemical Vapor Deposition 5
4.2.3. Single-Crystal Growth 8
4.2.4. Wafer Manufacture 15
4.2.5. Epitaxy 17
4.2.6. Characterization 19
4.3. Solar Silicon 19
4.3.1. Crystalline Silicon 20
4.3.2. Amorphous Silicon 22
5. Silicides 22
6. Ferrosilicon 23
7. Calcium Silicon 27
8. References 29

1. Introduction

Silicon, Si, is a gray, metallically lustrous element with atomic number 14 and atomic mass 28.086. It is a member of group 14 of the periodic table, situated between carbon and germanium, and has three natural isotopes: ^{28}Si (92.2 %), ^{29}Si (4.7 %), and ^{30}Si (3.1 %). Corresponding to its position in the periodic table, silicon has four valence electrons, two of them in the 3 *s* and two in the 3 *p* levels. It crystallizes in the diamond structure like diamond, germanium, and α-tin. Silicon is a nonmetallic, semiconducting element with a resistivity of ca. 400 kΩ · cm at room temperature in its purest state. Its electrical resistivity decreases with increasing temperature and increasing concentration of electrically active elements such as B, Al, Ga, In, Tl, P, As, and Sb. By doping silicon with boron, for example, its room-temperature resistivity can be reduced by nine orders of magnitude to ca. 0.4 mΩ · cm. Except for very thin foil, silicon is impermeable to visible light, but it is highly permeable to infrared light. The IR permeability decreases with decreasing electrical resistivity.

After oxygen (50.5 %), silicon is the second most abundant element (27.5 %) in the lithospere. It does not occur in elemental form, but only in oxides and silicates.

Modern electronics is almost exclusively (> 95 %) based on silicon devices, both in low-power and in high-power electronics, a situation that will continue for at least the next few decades. Because of the eminent and ever increasing importance of electronics in technology and everyday life, silicon is one of the most important technical materials, although the quantity required for this application is relatively small (ca. 10 000 t of electronic-grade silicon in 1992). This small quantity results from the very small dimensions of electronic devices (e.g., 0.5×4×4 mm).

Of secondary importance are the uses of silicon for metallurgy and chemistry (silicon compounds), although they consume the major portion of the silicon produced.

2. History

The name silicon derives from the Latin word *silex* (= flint). Even in ancient times, silicon compounds were used for the production of glass. The first attempts to obtain elemental silicon from silica were made in the 1700s, but only in 1810 did Berzelius succeed in isolating silicon from ferrosilicon by reduction with carbon and iron at the melting point of iron. In 1823, he obtained iron-free silicon by reducing SiF_4 with potassium metal.

In 1854, Sainte-Claire Deville obtained coarse-grained silicon by fusion electrolysis

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_721

of silicon-containing $NaAlCl_4$. Further important developments were achieved by WÖHLER (1855), WINKLER (1864), SCHEID (1899), and KÜHNE (1902). Commercial production of ferrosilicon (iron-rich silicon), which had proved to be an excellent deoxidant in steel production, began in the early 1900s as an offshoot of calcium carbide production. Ferrosilicon is produced in the same type of furnace as calcium carbide and is a byproduct of calcium carbide production.

3. Physical and Chemical Properties

Physical Properties. Some physical properties of silicon are as follows (values without references are taken from [1]):

Atomic number	14
A_r	28.086
Lattice constant	0.5431 nm (= edge length of the cubic unit cell)
Interatomic distance in ⟨111⟩ direction	0.2352 nm
Atomic density	5.00×10^{22} atoms/cm^3
Density at 300 K	2.329 g/cm^3
Density at melting point	
Solid	2.30 g/cm^3
Liquid	2.51 g/cm^3
Volume increase at transition from liquid to solid	+9.1 %
Specific heat (300 K)	0.713 J g^{-1} K^{-1}
Thermal expansion (300 K)	2.6×10^{-6}/K
Thermal conductivity (300 K)	1.5 W cm^{-1} K^{-1}
mp	1687 K
bp	3504 K [2]
Critical temperature	5193 K
Critical pressure	145 MPa
Latent heat of fusion	50.66 kJ/mol
Heat of evaporation	385 kJ/mol
Combustion heat, H_0 (Si/SiO_2)	850 kJ/mol
Bulk modulus (300 K)	97.84 Pa
Vickers hardness	10.2 GPa
Mohs hardness	7
Surface tension at melting point	885 mJ/m^2
Band gap (300 K)	1.126 eV
Electron mobility	1440 cm^2 V^{-1} s^{-1}
Hole mobility	484 cm^2 V^{-1} s^{-1}

Chemical Properties. As an element of group 14 of the periodic table, silicon represents the transition from the base metals of groups 1, 2, and 13 to the metalloids and nonmetals of groups 15–17. With nonmetals of higher electronegativity, silicon forms predominantly nonpolar compounds such as Si_3N_4 or SiC, as well as more polar compounds such as SiO_2 and SiF_4 with oxidation state 4^+. If Si–Si bonds are present, the mean oxidation state is lower (e.g., 3^+ in Si_2I_6).

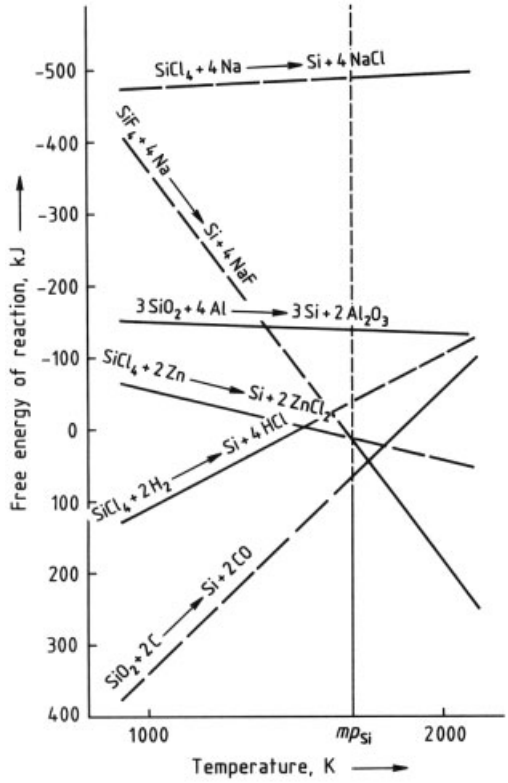

Figure 1. Free energy of selected reactions for production of silicon from SiO_2 and silicon tetrahalides [5]

Silicon binds oxygen dissolved or present as compounds in metal melts to give SiO_2 or SiO, and it may partially evaporate in the form of the volatile SiO. With metals, silicon forms alloys, intermetallic compounds, or silicides.

Bulk silicon is very resistant to acid, including HF or HNO_3, but not to a mixture of HF and HNO_3. In this mixture, HF first etches away the SiO_2 layer from the silicon surface, then HNO_3 oxidizes silicon again, the new SiO_2 layer is removed by HF, and so on. Silicon reacts violently with even dilute alkaline solutions to generate hydrogen. Heating in an oxygen-containing atmosphere forms an SiO_2 layer that inhibits further oxidation. Above 520 K, silicon reacts with HCl to form $SiHCl_3$ and $SiCl_4$ (→ Silicon Compounds, Inorganic), and with CH_3Cl in the presence of copper to give alkylchlorosilanes

(→ Silicon Compounds, Organic). At ca. 620 K, silicon and HBr give $SiBr_4$ and $SiHBr_3$. At room temperature, silicon and fluorine produce SiF_4.

4. Production

4.1. Metallurgical Silicon

Since the turn of the century, silicon has been produced almost exclusively by carbothermal reduction of silicon dioxide:

$$SiO_2 + 2\,C \longrightarrow Si + 2\,CO \quad \Delta H_{298} = 695\,kJ$$

Another way to produce silicon from SiO_2 is aluminothermal reduction:

$$3\,SiO_2 + 4\,Al \longrightarrow 3\,Si + 2\,Al_2O_3 \quad \Delta H_{298} = -619\,kJ$$

Figure 1 shows the free energy of reaction ΔG_0 of the two reactions for the range 900 to 2100 K. Whereas carbothermic reaction of SiO_2 is unattractive below 2000 K, reduction with aluminum is strongly exothermic over the entire temperature range. In carbothermic reaction, the side reactions

$$SiO_2 + C \longrightarrow SiO + CO$$

and

$$SiO_2 + 3\,C \longrightarrow SiC + 2\,CO$$

must be avoided.

For the carbothermal production of silicon and ferrosilicon, electric-arc furnaces (Fig. 2) are used. For silicon production, graphite electrodes are used, whereas for the production of ferrosilicon, Söderberg electrodes (steel tubes filled with a mixture of coal powder, tar, and pitch) are employed. The carbothermal process yields silicon with a purity of ca. 98 % (metallurgical-grade silicon, MG-Si), for which a typical analysis is 0.5 % Fe; 0.4 % Al; 100 – 400 ppm each of Ca, Cr, Mg, Mn, Ni, Ti, and V; and 20 – 40 ppm each of B, Cu, P, and Zr. The usual starting materials are chunks of quartzite, coke, as well as charcoal and wood chips for good ventilation of the charge. To produce 1 t of silicon, 2.9 – 3.1 t of quartz or quartzite, 1.2 – 1.4 t of coke (gas coke and petroleum coke), and 1.7 – 2.5 t of charcoal and wood are required. The consumption of electrode graphite is 120 – 140 kg/t Si, which must be added to the amount of reducing agent, and of electrical energy is 12.5 – 14 MW · h/t Si. The purity of the silicon produced depends primarily on the purity of these materials and secondarily on the purity of the graphite electrodes and the furnace lining.

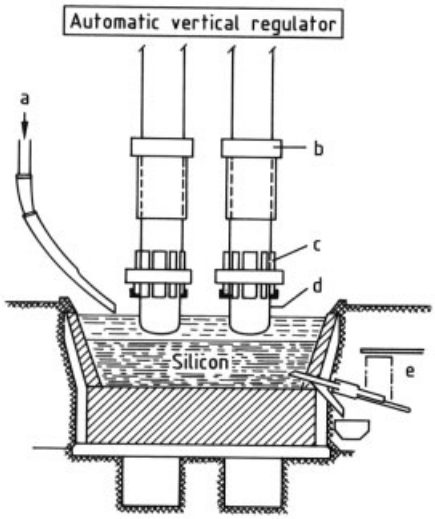

Figure 2. Arc furnace for silicon production [5]
a) Input of quartz and carbon; b) Electrode holder; c) Electric contact; d) Carbon electrode; e) Tapping electrode

The purity of silicon can be increased by using purer starting materials, by improving the furnace construction, and by optimizing the production process. Dow Corning [3], [4] has introduced an improved arc furnace, the so-called direct arc reactor (DAR), which yields purer silicon. With adequately pure starting and furnace materials and with optimized operation, the DAR process gives silicon with a purity of 99.99 %. Residual impurities are Al, Fe (50 – 100 ppm each), and other metals, as well as B and P (< 10 ppm each).

Another method of producing silicon starts with silicon tetrafluoride, which is a byproduct of the production of superphosphate fertilizer from phosphate rock. The ore contains CaF_2, which is converted to HF when treated with H_2SO_4; HF then reacts with SiO_2 to form SiF_4, with an estimated 75 000 t/a being recovered in the United States as fluosilicic acid, H_2SiF_6. Sodium fluoride is added to the H_2SiF_6 to form sodium fluosilicate:

$$H_2SiF_6(aq.) + 2\,NaF\,(s) \longrightarrow Na_2SiF_6(s) + 2\,HF\,(aq.)$$

After filtration and drying, Na_2SiF_6 is decomposed at 920 K:

$$Na_2SiF_6(s) \xrightarrow{\Delta} SiF_4(g) + 2\,NaF(s)$$

Then, SiF_4 is purified by passing it over iron at 1070 K to remove air and SO_2, and by subsequent fractional distillation. Silicon tetrafluoride can be reduced readily to silicon by reaction with sodium:

$$SiF_4(g) + 4\,Na(l) \longrightarrow Si\,(powder) + 4\,NaF(s)$$

The exothermic character of this reaction provides sufficient heat to maintain the process at 770 K. To separate silicon from NaF, the Si –NaF mixture is melted to form two immiscible phases for which liquid – liquid extraction (migration of impurities from silicon to NaF) provides additional purification. This reduction process is known as the SRI (Stanford Research Institute International) process [6], [7]. Because the reduction reaction yields four times as much NaF as silicon, low- cost silicon is attainable only if the byproduct can be sold.

An alternative way to transform H_2SiF_6(aq.) into elemental silicon is used by Ethyl Corporation [8]:

$$H_2SiF_6(aq.) \xrightarrow{\Delta} SiF_4 + 2\,HF$$
$$SiF_4(g) + NaAlH_4(s) \longrightarrow SiH_4(g) + NaAlF_4(s) \qquad (1)$$

Also, $KAlH_4$ or a mixture of $NaAlH_4$ and $KAlH_4$ can be used. $NaAlF_4$ is used in aluminum production, and SiH_4 is decomposed at ca. 1000 K to generate elemental silicon:

$$SiH_4(g) \longrightarrow Si(s) + 2\,H_2(g) \qquad (2)$$

About 150 other reactions yield elemental silicon, but most of them are not used in industrial production.

Ferrosilicon production, ca. 2×10^6 t in 1992, constitutes the major part of industrial silicon production. Ferrosilicon is used in steelmaking, mainly as a deoxidizer, but also as an alloying element (e.g., for the production of transformer sheet metal).

World production of metallurgical-grade silicon in 1991 was ca. 650 000 t, with a market price of ca. $ 1.5/kg. About 60 % was used as an alloying element for nonferrous metals, particularly aluminum alloys; ca. 37 % for the production of silicon compounds; and ca. 3 % for electronic applications (electronic devices, solar cells, sensors, micromechanics). The main producer countries for MG silicon are Norway, the United States, France, Canada, the Republic of South Africa, the CIS, Italy, Spain, and Yugoslavia.

4.2. Semiconductor Silicon

4.2.1. Purification

For applications in electronics, photovoltaics, sensors, and micromechanics, metallurgical-grade silicon must be refined to a much higher purity. Purity requirements for the fabrication of high-efficiency solar cells, for example, are nearly as high as those for advanced semiconductor devices.

At present, the main refining steps are the adsorption and distillation of a silicon compound. For this purpose, silicon is first converted into a compound such as $SiHCl_3$, $SiCl_4$, SiH_2Cl_2, or SiH_4. After purification, the silicon compound is transformed back into elemental, high-purity silicon by chemical vapor deposition (CVD). Alternative refining methods [5] have not been used in industrial production up to the present.

The starting material for the production of semiconductor-grade silicon (SG-Si) is about 90 % carbothermal silicon and about 10 % silicon produced by the Ethyl process [8].

Carbothermal silicon, milled to a sand or powder, is fed into a fluidized-bed reactor where it is fluidized by a stream of hydrogen chloride flowing through small orifices in the bottom of the reactor (see Fig. 3). The reactor is held at ca. 650 °C and the exothermic reaction yields predominantly trichlorosilane and hydrogen,

$$Si + 3\,HCl \rightleftharpoons SiHCl_3 + H_2 \qquad (3)$$

as well as silicon tetrachloride:

$$Si + 4\,HCl \rightleftharpoons SiCl_4 + 2\,H_2$$

In addition to $SiHCl_3$ and $SiCl_4$, other chlorosilanes and compounds of the impurities with chlorine or hydrogen are formed. After removal of silicon powder, impure trichlorosilane is condensed and stored in an intermediate tank, and

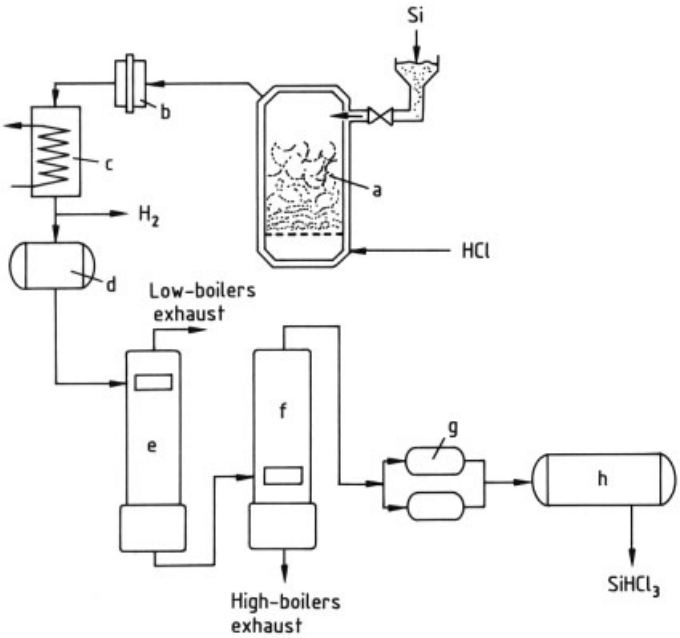

Figure 3. Flow chart of the preparation and refining of trichlorosilane
a) Fluidized-bed reactor; b) Dust filter; c) Condenser; d) Tanks, e) Distillation of low-boiling impurities; f) Distillation of high boilers; g) Tanks; h) Storage tanks

subsequently purified by adsorption and distillation (see Fig. 3). Because of its favorable boiling point of 305 K, trichlorosilane can be purified effectively by fractional distillation. To improve the removal of boron and phosphorus, the two most important electrically active elements in silicon, boron(III) and phosphorus(III) chloride are transformed into less volatile compounds before distillation by use of complexing agents. In this way, the production of trichlorosilane with boron and phosphorus contents of < 1 ppba (parts per billion atomic) is possible.

From the point of view of chemical process engineering, trichlorosilane is easiest to handle and most economical. About 90 % of the electronic-grade silicon is purified by distillation of trichlorosilane, and the rest by distillation of silane, SiH_4. The distillation of SiH_4 is applied only after the Ethyl process (Eq. 1). Silane is much more dangerous to handle than trichlorosilane.

4.2.2. Chemical Vapor Deposition

All semiconductor-grade silicon is produced by converting silicon compounds to elemental silicon by chemical vapor deposition (CVD). For ca. 75 % of worldwide CVD production, trichlorosilane is used as the silicon compound (e.g., by Wacker, Hemlock, Tokuyama Soda, Kyundo, OTC, MEMC); for ca. 25 %, silane (ASiMI, Ethyl), partly produced from trichlorosilane, is used.

The method of converting trichlorosilane into high-purity silicon was first developed by Siemens & Halske in 1952 and patented in 1956 [9]. It employs the reverse of Equation (3), which is strongly endothermic and therefore requires a high reaction temperature (ca. 1400 K). This reaction also yields much tetrachlorosilane as byproduct:

$$4\,SiHCl_3 + 2\,H_2 \rightleftharpoons 3\,Si + SiCl_4 + 8\,HCl \quad \Delta H = 964\,kJ$$

The byproduct $SiCl_4$ is partly converted back to trichlorosilane in the reactions

$$SiCl_4 + H_2 \rightleftharpoons SiHCl_3 + HCl$$

$$Si + 3\,SiCl_4 + 2\,H_2 \rightleftharpoons 4\,SiHCl_3$$

The remaining $SiCl_4$ is used for the production of highly dispersed SiO_2 (e.g., HDK from

Wacker) and synthetic silica glass and glass fibers, and for silicon epitaxy in the production of electronic devices.

To avoid the deposition of silicon on unwanted and contaminating surfaces, the CVD arrangement shown schematically in Figure 4 is used. In a bell-jar reactor, three silicon slim-rods (ca. 8 mm in diameter, semiconductor grade, float zone pulled in a specially designed apparatus) are mounted in the form of an inverted U and are resistance heated electrically to ca. 1400 K, while the bell jar is kept cold. In this way, the heat required for the reaction is available only on the surface of the silicon rods; silicon forms there exclusively and adheres to the rods. Elemental silicon not adhering to the rods is transformed back to a gaseous silicon compound at lower temperature (e.g., see Eq. 3) in the reactive atmosphere around the rods.

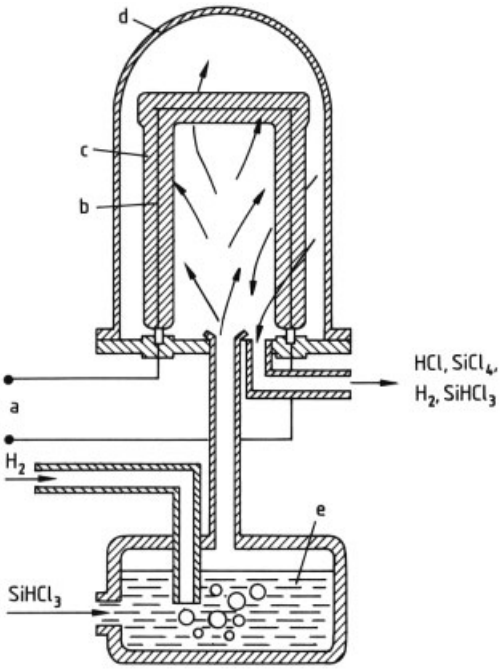

Figure 4. Chemical vapor deposition of silicon
a) Electrical current; b) Starting silicon slim-rod; c) CVD polycrystalline silicon rod (1400 K); d) Reactor (silica, metal); e) Saturator

In the CVD process, the concentration of impurities is again reduced [10], in the case of boron and phosphorus by a factor of up to 1000. The trichlorosilane CVD process yields silicon with < 0.02 ppba boron and < 0.01 pbba phosphorus, so that further refining of the CVD silicon by zone melting for electronic applications is no longer necessary in most cases, in contrast to the situation at the beginning of semiconductor-grade silicon production.

Silicon single crystals can also be grown by CVD if monocrystalline slim-rods are used. However, this process is very expensive because of the very low growth rate that must be used for single- crystal growth, and it yields low-quality single crystals. Therefore, the CVD silicon produced today is a fine-grained, polycrystalline material used as starting material for the melt growth of single crystals.

In the CVD process, much energy is lost by thermal radiation from the silicon rods. The radiant energy is absorbed largely by the cold reactor walls. Energy is therefore the primary cost factor in this process; thus endeavors are made to maximize the deposition rate. Theoretically, the highest growth rate is ca. 2 mm/h; however, growth rates used in industry are lower – ca. 1 mm/h or less.

The upper limit for the deposition rate is determined by the required silicon quality, especially in the production of silicon rods for float-zone crystal growth (see page 11), which requires long cylindrical silicon feed rods, free of cracks and voids, and with a smooth surface, as starting material. Too rapid a growth rate results in dendritic growth, which gives a rough surface, voids, and cracks in the material. Such a feed rod would either crack or emit silicon particles on melting during float-zone growth.

Whereas the diameter of the feed rods for float-zone growth is determined by the requirements of the crystal growth process (the diameter of the feed rods should be a little larger than that of the crystals to be grown), CVD silicon rods for the growth of crucible-pulled crystals may be of any diameter. Therefore, CVD rods for crucible pulling are always grown to the largest possible diameter for economic reasons: the CVD yield (kg/W · h) and the CVD output (kg/h) increase linearly with the radius of the rod at a constant deposition rate. However, the maximum diameter of CVD rods is ca. 25 cm, since at larger diameters, the center of the rod begins to melt and the rod system breaks down, causing damage to the reactor and contamination of the silicon.

Therefore, the deposition process is stopped at a rod diameter known to exhibit no melting.

At present, CVD processes run automatically. Process parameters such as current, temperature, pressure, gas flow, and concentration are measured continuously and are used by a computer to calculate new control parameters. Rod lengths > 2 m and rod numbers of 8 – 16 per reactor are usual today.

A different method of producing CVD silicon is used by ASiMI (formerly UCC), which transforms $SiHCl_3$ into SiH_4 and performs the CVD process with SiH_4 at ca. 1000 K. The CVD reactor and arrangement are similar to those used for deposition with $SiHCl_3$ (deposition on silicon rods, see Fig. 4). The transformation of trichlorosilane into silane occurs according to the following reactions:

$$6\,SiHCl_3 \overset{cat.}{\rightleftharpoons} 3\,SiCl_4 + 3\,SiH_2Cl_2$$

$$4\,SiH_2Cl_2 \rightleftharpoons 2\,SiHCl_3 + 2\,SiH_3Cl$$

$$2\,SiH_3Cl \rightleftharpoons SiH_4 + SiH_2Cl_2$$

$$4\,SiHCl_3 \rightleftharpoons 3\,SiCl_4 + SiH_4$$

The CVD reaction is the decomposition of silane:

$$SiH_4 \longrightarrow Si + 2\,H_2 \qquad (4)$$

A very different way of transforming silane into elemental silicon is used by Ethyl, United States [11], and KRICT, Korea [12], based on a Texas Instruments patent [13].

A mixture of ca. 20 mol % silane and 80 mol % hydrogen is fed through a gas-dispersing frit into the bottom of a reactor; a fine silicon powder is let in through an opening at the top of the reactor and fluidized by the upflowing gas. The silicon powder is heated to ca. 1000 K, and silane decomposes according to Equation (4) into silicon and hydrogen on the hot silicon particles. The silicon particles grow, and the larger ones fall to the bottom of the reactor where they are removed through an opening.

In the Ethyl process, the silicon particles are heated from outside the reactor by resistance heating (see Fig. 5). This leads to problems with unwanted deposition of silicon on the hot walls of the reactor.

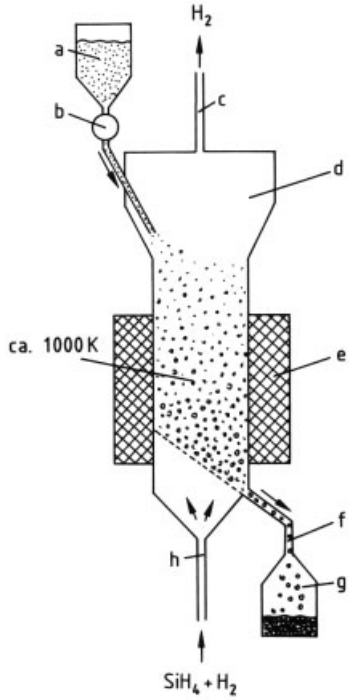

Figure 5. Chemical vapor deposition of granular silicon
a) Silicon powder; b) Powder-feeding device; c) Exhaust; d) Fluidized-bed reactor; e) Heating (resistance or microwave); f) Removal of coarse silicon granules; g) Container for silicon granules; h) Gas injection

This problem is overcome in the newer KRICT process. Silicon particles are heated with microwaves, and the reactor wall remains cold if it is made from an appropriate material (e.g., silica).

Granular CVD silicon has major advantages over rod-shaped CVD silicon for feeding, particularly for continuous feeding of the melt with silicon during crystal growth, because the round silicon granules flow much better than the sharp-cornered grains of broken CVD silicon rods. Moreover, breaking the rods contaminates the silicon severely.

In 1992, the world production of electronic-grade CVD silicon was ca. 10 000 t with a market value of ca. $\$\,500 \times 10^6$ (\$ 50/kg).

4.2.3. Single-Crystal Growth

More than 95 % of all electronic devices are made from single-crystal silicon. In addition, single-crystal silicon is used at present for ca.

40 % of all photovoltaic solar cells and for all micromechanical devices. Two crystal growth methods are used in industry for the production of silicon single crystals: the Czochralski (CZ) method, which yields about 85 % of all silicon crystals, and the float-zone (FZ) method, which yields 15 % of the silicon crystals. The CZ crystals are used mainly for manufacturing highly integrated low-power devices, typically for microprocessors, RAMs, DRAMs, ASICs, etc., whereas the FZ crystals are used mainly for making discrete or low-integrated high-power devices such as diodes, transistors, and thyristors.

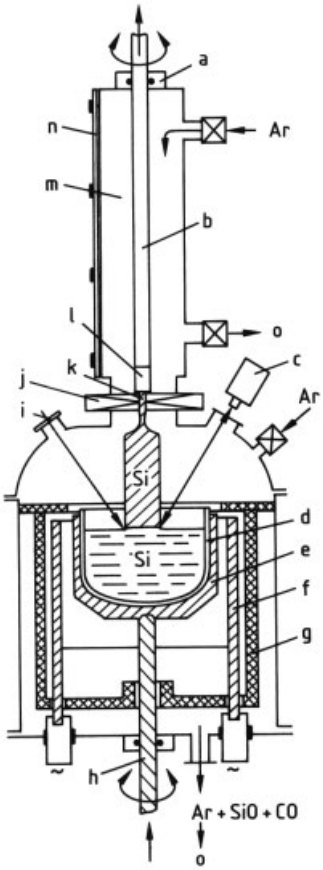

Figure 6. Czochralski silicon crystal puller
a) Seal; b) Seed shaft; c) Optical system; d) Silica crucible; e) Graphite crucible; f) Graphite heater; g) Thermal insulation; h) Crucible shaft; i) Viewing port; j) Separation valve; k) Seed crystal; l) Seed holder; m) Front opening chamber; n) Front opening door; o) Valve

Czochralski Crystal Growth. The Czochralski method [14] uses a crucible for crystal pulling and is therefore also known as crucible pulling (see Fig. 6; for more detailed information, see [15–17]). A silicon melt is held in a crucible, and a single- crystal seed of the proper orientation is dipped into the melt and slowly withdrawn vertically. With suitable pulling speed and crystal and crucible rotation rate, a single crystal of the desired shape can be grown.

To eliminate dislocations, which always occur on immersing the seed crystal, a thin crystal neck (ca. 3 – 5 mm in diameter and several centimeters long) must be grown with a growth rate of several millimeters per minute. Because of the small diameter, the thermal strains in the neck are too low and the growth rate is too high for propagation of dislocations downward into the growing crystal, and the crystal neck becomes dislocation free after a few centimeters. The same procedure is used in float-zone pulling.

The dislocation-free state is stable, and large diameters can be grown despite the high cooling strains in large crystals. After neck growth, the diameter is increased to the nominal crystal diameter and then kept constant by adjusting the pulling rate and the heating power. Dislocation-free crystal growth can be completed by reducing the crystal diameter to nearly zero at the end of the crystal.

Table 1. Equilibrium distribution coefficients k_0 and typical concentrations of impurities in CVD-Si and CZ-Si

Element	k_0	Impurity concentration, ppba	
		CVD-Si	CZ-Si
C	0.07	≤ 200	≤ 200
N	7×10^{-4}	$< 4*$	$< 4*$
O	ca. 1	≤ 100	ca. 15 000
B	0.8	≤ 0.02	ca. 0.2
Al	2.8×10^{-3}	≤ 0.01	≤ 0.4
Ga	8×10^{-3}	$< 0.002*$	$< 0.002*$
In	4×10^{-4}	$< 0.002*$	$< 0.002*$
P	0.35	≤ 0.01	ca. 0.1
As	0.30	ca. 0.01	≤ 0.01
Sb	0.023	ca. 0.0008	ca. 0.01
Li	0.01	$< 0.3*$	≤ 0.1
Na	ca. 1×10^{-3}	$< 0.3*$	≤ 0.1
K		$< 0.2*$	$< 0.1*$
Ti	2.0×10^{-6}	$< 0.5*$	$< 0.5*$
Cr	1.1×10^{-5}	$< 0.02*$	$< 0.004*$
Fe	6.4×10^{-6}	$< 0.5*$	≤ 0.01
Ni	ca. 3×10^{-5}	ca. 0.4	≤ 0.01
Co	1.0×10^{-5}	ca. 0.002	$< 0.001*$
Cu	8×10^{-4}	ca. 0.05	≤ 0.01
Ag	ca. 1×10^{-6}	0.01	$< 0.0008*$
Au	2.5×10^{-5}	0.00001	$< 0.00001*$
Zn	ca. 1×10^{-5}	ca. 0.1	$< 0.05*$
Zr	$< 1.5\times10^{-7}$	$< 0.5*$	$< 0.5*$

$*$ Concentration below the detection limit listed here.

Usually, a small residual melt remains in the crucible, the amount of which depends on quality demands and economic factors. Owing to the segregation effect, the majority of the impurities remain in the melt since most of them have very low segregation coefficients (see Table 1).

The silicon melt reacts with all materials to some extent. Only silica can be used as crucible material, because the product of its reaction with the silicon melt, silicon monoxide (SiO), evaporates readily from the melt. Thus supersaturation of the melt with oxygen or SiO is prevented. Supersaturation of oxygen in the melt would cause supersaturation of oxygen in the crystal, oxygen precipitation, and the formation of dislocations in the crystal near the freezing interface. The growth of a dislocation-free crystal would be impossible. All other crucible materials either contaminate the silicon crystal with impurities (e.g., Al_2O_3 or ZrO_2) or cause dislocated crystal growth by supersaturation (e.g., SiC and Si_3N_4).

Both to promote the evaporation and removal of SiO and to prevent the backflow of CO and other reaction and evaporation products, the furnace is purged with a strong argon stream and the pressure is kept low (generally 10 – 100 mbar). Table 2 gives an overview of the basic data for industrial silicon crystal growth.

About 1.5 g of silica is dissolved per 1000 g of CZ-Si crystal during growth, but $< 1\,\%$ of the oxygen in the dissolved silica is found in the crystal. This indicates that the evaporation rate of SiO is very high and the oxygen content of the melt is exchanged every 10 min. Accordingly, oxygen transport from the crucible to the crystal and thus the oxygen content of the crystal can be manipulated over a wide range by influencing convective flows in the melt and the evaporation of SiO at the melt surface.

Impurities in the dissolved silica are absorbed by the silicon melt. However, only a small portion of impurities are incorporated into the crystal owing to the segregation effect, which is very strong for most elements (see Table 1). For example, from 100 000 iron atoms in the melt, only one atom is incorporated into the crystal. As a result of this strong segregation, the impurity content of the CZ-Si crystal is lower than that of the starting material for most elements.

Figure 6 shows the basic arrangement of a CZ-Si crystal puller, and Figure 7 shows a photograph of such a crystal puller during removal of a grown crystal. The best and most economical heating method is resistance heating with a graphite heating element. The soft silica crucible must be supported by a graphite crucible. The diameter of the crystal is monitored by an optical system (e.g., a CCD camera) and, in most cases, the melt and furnace temperatures are measured by pyrometers. The crystal and the crucible can be moved and rotated very precisely. A valve between the front-opening chamber and the furnace allows the removal of the crystal while the furnace is hot and refilling of the hot crucible(containing residual melt) with polycrystalline silicon and dopant. The growth control of modern crystal pullers allows completely automatic crystal growth. However, with monitoring and correction by an operator, the yield of the crystal grower can be improved. In 1992, the standard diameters of commercial CZ-Si crystals ranged from 50 to 200 mm, averaging 150 mm, and CZ-Si crystals 300 mm in diameter are being developed.

Figure 7. Removal of a large CZ silicon crystal (diameter 204 mm, weight 50 kg) that is still hanging from the seed crystal

Float-Zone Crystal Growth. Figure 8 shows the basic arrangement for large diameter float-zone silicon crystal growth in a modern FZ puller. The FZ-Si method differs from the CZ-Si method mainly in the absence of a crucible and in the small melt volume. The absence of

Table 2. Data for industrial silicon single-crystal growth

		Industrial standard production, 1992	
Property	Maximum value∗	CZ crystals	FZ crystals
Diameter, mm	$\geq$ 300	$\leq$ 200	$\leq$ 150
Length, mm	$\geq$ 2500	$\leq$ 2000	$\leq$ 1800
Weight, kg	$\geq$ 100	$\leq$ 100	$\leq$ 50
Orientation	any	⟨100⟩, ⟨111⟩, ⟨110⟩, ⟨115⟩	⟨111⟩, ⟨100⟩, ⟨110⟩, ⟨115⟩
[B], cm^{-3} ∗∗	$\leq 3\times10^{20}$	1×10^{14}–1×10^{20}	5×10^{11}–1×10^{17}
[P], cm^{-3}	$\leq 2\times10^{20}$	1×10^{14}–1×10^{18}	1×10^{12}–5×10^{18}
[As], cm^{-3}	$\leq 3\times10^{20}$	5×10^{18}–8×10^{19}	
[Sb], cm^{-3}	$\leq 2\times10^{19}$	2×10^{17}–1×10^{19}	
[Al], cm^{-3}	$\leq 5\times10^{17}$		
[Ga], cm^{-3}	$\leq 2\times10^{18}$		
[In], cm^{-3}	$\leq 1\times10^{17}$		
[C], cm^{-3}	$\leq 1\times10^{18}$	1×10^{15}–3×10^{16}	1×10^{15}–3×10^{16}
[N], cm^{-3}	$\leq 5\times10^{15}$	$< 1\times10^{14}$	$< 1\times10^{14}$–1×10^{15}
[O], cm^{-3}	$\leq 2\times10^{18}$	1×10^{17}–1×10^{18}	1×10^{15}–2×10^{16}
Crucible material		silica	no crucible
Crucible diameter, mm		$\leq$ 600	
Rotation rate, rpm		$\leq$ 50 crystal, $\leq$ 30 crucible	$\leq$ 50 crystal, $\leq$ 2 feed rod
Growth rate, mm/min		$\leq$ 2.5	$\leq$ 4.5
Pressure, mbar		5 – 1000	5×10^{-4}–5000
Argon flow, L/h		400 – 5000	

∗ The maximum concentration for dislocation-free growth is generally lower than the maximum solubility in silicon (at ca. 1550 K) due to the effect of retrograde solubility for most of the elements (for most elements, solubility increases up to ca. 1550 K, then decreases continually up to the melting point of 1687 K). The maximum growth and rotation rates decrease with increasing crystal and crucible diameters.

∗∗ Concentration in number of impurity atoms per cm^3 of silicon; 1 cm^3 Si contains 5.00×10^{22} atoms.

a crucible results in a very low oxygen content and in lower concentrations of impurities compared to CZ-Si crystals. The small melt volume leads to an axially constant boron or phosphorus concentration in contrast to CZ-Si (see Fig. 9). Boron and phosphorus are the only dopants used in industrial FZ-Si production.

The starting material for the growth of an FZ-Si single crystal is a polycrystalline silicon feed rod clamped to a movable shaft. The feed rod can be unmelted CVD silicon or FZ-premelted silicon (CVD rods that have been float zone grown to give cylindrical, bubble- and crack-free polycrystalline rods). An induction coil melts the feed rod at the bottom, generating a large melt drop. A slim seed crystal is moved through the narrow opening in the center of the induction coil and brought into contact with the melt drop. Upon movement of the seed crystal downward the melt crystallizes on it and the seed crystal grows.

As in CZ-Si pulling, a dislocation-free state must be generated by growing a neck prior to growth of the large-diameter crystal. After the crystal diameter is increased, melt and crystal are heated from above by the induction coil, not laterally as in the case of small crystal diameters. The melt lies on top of the crystal and is held together mainly by its surface tension and partly by electromagnetic forces. It is connected to the feed rod by a slim melt neck through which the melt flows from the feed rod to the crystal. At the end of the crystal growth process, the FZ melt freezes on top of the crystal and generates dislocations. The dislocations multiply and propagate into the crystal over a distance of about one-half the crystal diameter.

The height of the melt at the edge of the crystal is only several millimeters and does not depend on crystal diameter. Therefore, the growth of large crystal diameters is not limited by heating problems, as was assumed earlier. The FZ-Si crystals can be grown to similar dimensions as CZ-Si crystals. The main problem in the growth of large-diameter FZ-Si crystals is the high thermal stress. The temperature gradients are larger

in growing FZ-Si crystals than in growing CZ-Si crystals of the same diameter because FZ-Si rods cool faster. The CZ-Si crystals are reheated by thermal radiation from the melt, the crucible, and the heater. For the growth of large diameters, cooling of the FZ-Si crystals must be reduced.

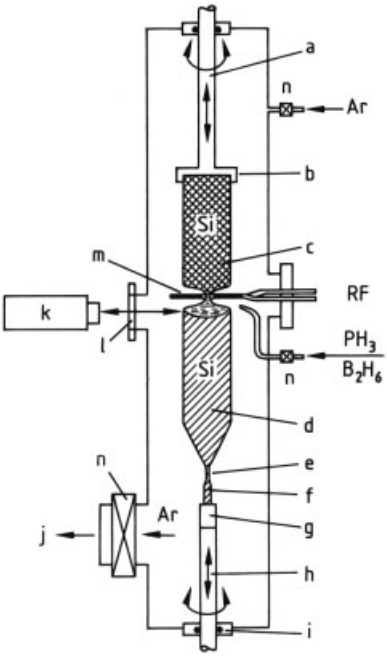

Figure 8. Float-zone silicon crystal puller
a) Feed rod shaft; b) Feed rod holder; c) Feed rod, polycrystalline; d) Single crystal; e) Neck; f) Seed; g) Seed holder; h) Seed shaft; i) Seal; j) To the vacuum pump; k) Camera; l) Viewing port; m) RF heating coil; n) Valve

Growing large-diameter crystals requires large amounts of energy and thus high RF voltages. To prevent sparkover at the induction coils, FZ-Si growth is carried out at higher gas pressures (> 1 bar). In contrast to CZ-Si growth, no argon flow takes place during FZ-Si growth; the FZ pulling chamber is first evacuated to high vacuum and then filled with argon to a certain pressure. The doping of silicon is realized during FZ crystal growth by the controlled addition of a doping gas (B_2H_6 or PH_3) to the argon during growth.

The monitoring and control of crystal diameter and the degree of automation of FZ-Si growth are similar to those of CZ-Si growth. Standard commercial production in 1992 (see Table 2) had crystal diameters up to 150 mm (usually ca. 100 mm). Larger diameters could be developed if they were in industrial demand.

Float-zone crystals can be pulled with higher growth rates than CZ crystals, because the solidification energy of FZ crystals is removed more quickly. With FZ growth the entire environment of the cooling crystal is cold (ca. 300 K, see Fig. 8), whereas the environment of the CZ crystal is very hot in places (melt, crucible, and furnace near 1700 K), and the growing CZ crystal receives much energy from thermal radiation.

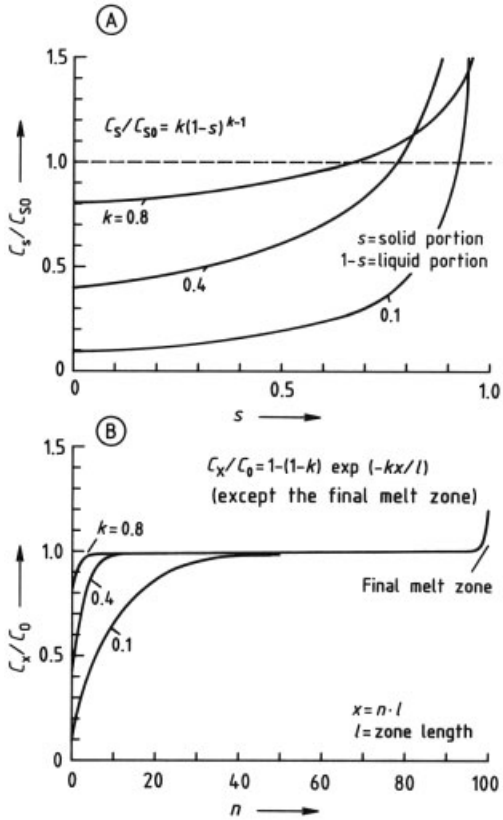

Figure 9. Relative axial concentration variation C_x/C_0 in silicon crystals produced from starting material (melt or feed rod) in which the impurities are distributed homogeneously
A) CZ crystals (normal freezing); B) FZ crystals (final melt zone freezes according to normal freezing); in this example, total length of feed rod is 100 l, and a constant cylindrical shape and diameter is assumed for the system feed rod – FZ melt – FZ crystal
k = Distribution coefficient

The maximum growth rate depends strongly on the crystal diameter in both FZ and CZ growth. With increasing diameter the ratio between the heat-emitting surface (predominantly

the parts of the surface hotter than 1200 K) and the heat-absorbing and generating growth interface becomes smaller, and additionally, the diffusion length of the heat increases owing to increasing distance between the crystal axis and the crystal surface. The maximum growth rate decreases from about 10 mm/min at 5-mm crystal diameter to about 1 mm/min at 150-mm crystal diameter for CZ-Si growth; the corresponding values for FZ-Si growth are 12 mm/min and 2 mm/min. In industrial production, the growth rates are lower than the maximum growth rates for reasons of yield and quality.

Segregation of Impurities. The axial variation of impurity concentration in CZ- and FZ-Si crystals due to segregation is shown in Figure 9. The segregation of impurities, in particular of dopants, has important qualitative and economical effects. Silicon device manufacturers demand silicon materials with tight resistivity tolerances, but CZ-Si crystal rods meet these tolerances only in small sections of their length.

In this respect, FZ-Si growth has major economic advantages. The resistivity of a boron- or phosphorus-doped FZ-Si crystal is nearly constant over its entire length. Strong variations occur only in the seed cone and in the finally freezing melt because of the small FZ melt volume. The height and depth of the melt are only a few centimeters (near the crystal axis) and depend on the crystal diameter and the curvature of the concave solid – liquid interface. The melt volume in FZ growth is much smaller than in CZ growth and remains nearly constant during growth of the cylindrical crystal.

In CZ growth the silicon charge is melted completely at the beginning. The melt volume decreases and the dopant concentration increases continuously during crystal growth. This is valid for nonvolatile or moderately volatile dopants such as B, P, Al, and Ga. In the case of volatile dopants such as As, Sb, or In, the enrichment of the melt with dopant can be compensated or overcompensated by evaporation of the dopant. In FZ growth the enrichment of boron or phosphorus is continuously compensated by the supply of unenriched silicon.

Many attempts and developments have been made to eliminate the axial resistivity decrease during CZ-Si growth. A partial solution of this problem is to consume only a part (e.g., one-half) of the melt and then to refill the crucible by adding polycrystalline silicon chips or by immersion and melting of a polycrystalline silicon rod.

Conditions similar to those during FZ growth can be realized with small CZ melts that are fed quasi- continuously with silicon and dopant (Fig. 10). The crucible is held in a fixed position and filled with melt to a certain level. A silica baffle is suspended from the cover of the furnace and dips several centimeters into the melt. The baffle has openings above the melt level to allow purging with argon. Additionally,a conical shield covers a large part of the melt surface to prevent the backflow and condensation of SiO on the feeding tube and baffle and to reduce heat losses by radiation. The silica feeding tube reaches to a short distance above the melt. Silicon granules and dopants are fed continuously through this tube at the same rate melt is removed by the growing crystal.

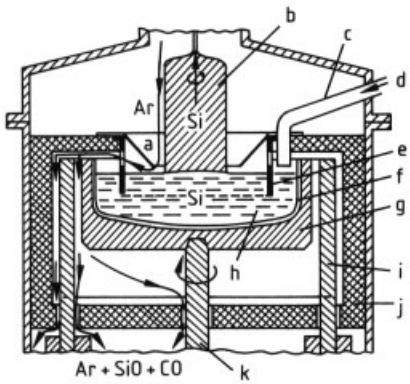

Figure 10. Quasi- continuous feeding of a CZ melt with silicon granules
a) Radiation shield, graphite; b) Crystal; c) Feeding tube, silica; d) Silicon granules; e) Baffle, silica; f) Crucible, silica; g) Susceptor, graphite; h) Melt; i) Heater, graphite; j) Thermal insulation, carbon; k) Crucible shaft, graphite

The problem with this technique is that fusion energy for the granules must be supplied during crystal growth, resulting in higher temperature of the crucible walls and problems with corrosion of the silica, with oxygen content, and with dislocation-free growth. Additional problems result from silicon crystallites that "swim" underneath the baffle from outside to the growing crystal where they generate dislocations.

The fusion energy for the silicon need not be supplied to the melt from which the crystal is grown if the melt is fed with molten silicon.

Melt Flows. Constant-volume melts have the advantage that their convective flows are constant over time. On a macroscopic scale they give a constant oxygen content and a constant radial impurity distribution. Small melts have the advantage of more controllable convective flows because of their weaker thermal convection. The thermal convection in the CZ-Si melt depends on the Grashof number *Gr:*

$$Gr = \frac{g\alpha \, \mathrm{d}T h^3}{\nu^2}$$

where g is the acceleration due to gravity, α is the volume thermal expansion coefficient, dT is the temperature difference in the melt, h is the height of the melt, and ν is the kinematic viscosity.

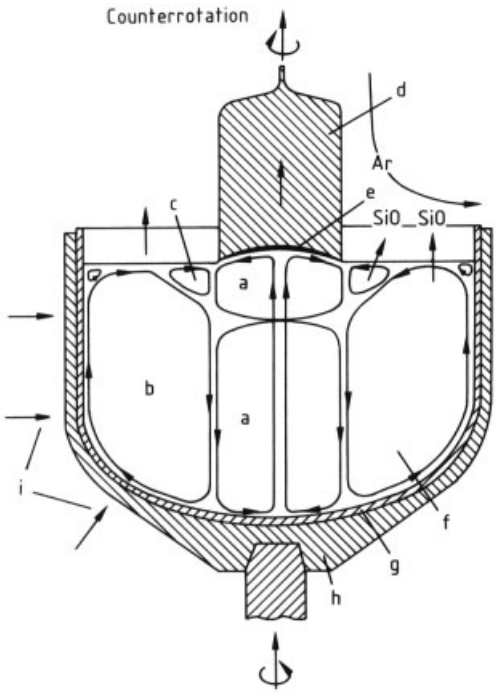

Figure 11. Schematic of the convective flows in a CZ silicon melt; flows are both temporally and locally inhomogeneous, creating temperature and concentration fluctuations
a) Mechanically forced convection; b) Thermal convection; c) Thermocapillary (Marangoni) convection; d) Crystal; e) Diffusion layer; f) Melt; g) Crucible; h) Susceptor; i) Heat flow direction

Owing to the dependence on h^3 and dT (the temperature differences in large melts are greater than in small melts), the convective flows in large CZ-Si melts become very turbulent, unsteady, and nearly uncontrollable. The unsteady flows result in large temperature fluctuations at the solid – liquid interface and in large fluctuations in oxygen transport to the crystal.

Figure 11 shows schematically the various convective flows in a CZ-Si melt. In the classical "open" growth furnace, thermal convection is strongest. A thermocapillary convection – also called Marangoni convection – always occurs below the melt surface at the edge of the crystal and at the crucible wall. It is driven by the differences in surface tension due to differences in the melt temperature. The effects of thermocapillary convection are not well understood.

Mechanically forced convection is generated by rotation of the crystal and crucible. Crystal rotation must be used for two reasons. First, without rotation, the crystal would grow with a noncylindrical shape and dislocations would be generated; second, crystal rotation reduces the strong impurity enrichment toward the center of the crystal caused by thermocapillary convection and segregation.

The incorporation of impurities depends strongly on the thickness of the diffusion layer in the melt at the solid – liquid interface: the thicker the diffusion layer, the higher the incorporation of impurities into the crystal. The thickness of the diffusion layer is strongly influenced by melt flows.

Impurity Striations. Temperature fluctuations at the solid – liquid interface lead to unsteady growth and backmelting. These result in microscopic impurity variations in the crystal, also called impurity striations. Unsteady growth can be reduced by using a high crystal rotational velocity because, first, mechanically forced convection decreases the influence of unsteady thermal convection at the growth interface, and second, when the crystal is moved quickly over the temperature irregularities, temperature variations in the crystal are decreased owing to its thermal response time.

In general, the oxygen distribution does not correspond to the distribution of other elements; this is true on a microscopic scale as well. The effective distribution coefficient of oxygen is

nearly unity and depends only slightly on the growth rate. However, the incorporation of oxygen into the growing crystal depends strongly on the pattern of convective flows in the melt. Therefore, unsteady melt flows result in an unsteady supply of oxygen and in oxygen striations in the crystal.

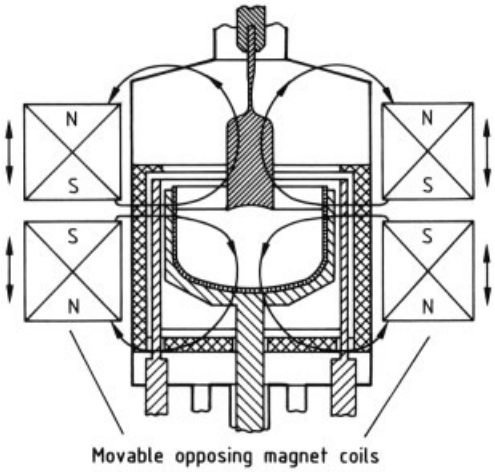

Figure 12. Growth of CZ silicon with two opposing vertical (also called cusp or configured) magnetic fields

Modifying Melt Flows. Melt flows in a CZ-Si melt can be modified in several ways. In most cases, this is achieved by altering growth parameters such as crystal rotation rate, crucible rotation rate, melt dimensions, furnace dimensions, furnace construction, and so on.

A more expensive way to influence the flows in a silicon melt is by the application of magnetic fields, which interact with the free electrons of the silicon melt and retard melt flows.

The easiest way to produce a magnetic field in a CZ-Si melt is to wind a magnetic coil around the furnace vessel and generate a vertical magnetic field in the melt. However, purely vertical fields were shown to result in poor radial impurity distributions. Better results were obtained with transverse magnetic fields, which damp the melt particularly in the vertical direction. This gives a more homogeneous radial impurity distribution of both oxygen and dopant and a low oxygen content. However, the isotherms in the melt are elliptically distorted, which causes backmelting at each revolution of the crystal and rotational striations of oxygen and dopants. Transverse magnetic fields up to 4000 G with diameters up to 100 cm are applied at present.

The third and newest method is the use of two opposing magnetic fields (see Fig. 12); this is probably the best method of the three, but it is not yet fully developed. Two magnetic coils surround the furnace vessel and generate opposing vertical magnetic fields. The coils can be moved vertically to adjust the field to the requirements of the melt; this can also be done during crystal growth. The coils and the field strength are arranged in such a way that a strong field is present near the crucible wall and a weak or zero field in the melt below the crystal. Thereby, melt flows are damped near the crucible wall and the transport of oxygen from the crucible to the crystal is reduced, but they are not damped below the crystal, where mechanically forced convection may result in a homogeneous radial impurity distribution. Good radial oxygen and dopant distributions and low oxygen contents are possible with this method.

Neutron Transmutation Doping. In the case of phosphorus doping, the problem of macro- and microscopical resistivity variation in silicon crystals can be overcome by neutron transmutation doping (NTD). The isotope ^{30}Si, which is present in natural silicon at 3.1 atom %, can be transformed to ^{31}P by absorption of a thermal neutron and emission of an electron:

$$^{30}\mathrm{Si} + \mathrm{n} \longrightarrow {}^{31}\mathrm{Si} + \gamma$$
$$^{31}\mathrm{Si} \xrightarrow{2.6\,\mathrm{h}} {}^{31}\mathrm{P} + \mathrm{e}$$

In the nuclear reaction, the ^{31}Si/^{31}P atoms are displaced from their lattice sites and travel a short distance through the lattice, causing lattice defects. Annealing the crystal at ca. 1100 K restores the silicon lattice, and the ^{31}P atoms return to lattice sites.

The transformation takes place very homogeneously in silicon, because most of the neutrons pass through the silicon crystal without being captured, and each ^{30}Si atom in the crystal has nearly the same probability of capturing a neutron as of being transformed into phosphorus. Consequently, the ^{31}P atoms generated by transmutation are distributed very homogeneously in the crystal.

4.2.4. Wafer Manufacture

After they are grown, silicon crystals are not perfectly cylindrical and do not have the precise nominal diameter despite highly automated and sophisticated crystal pulling. There are always lines of small nodes (at ⟨100⟩ crystals) or facets (at ⟨111⟩ crystals) on the crystal surface parallel to the crystal axis and small variations in crystal diameter. Therefore, the crystals are grown slightly larger in diameter and then ground precisely to cylinders of the stated diameter. To allow crystal-oriented positioning of the wafers in subsequent process steps, the wafers must have flats or notches at special crystallographic sites. For this purpose, planes or notches with defined crystal orientations (e.g., ⟨100⟩, ⟨110⟩, ⟨112⟩) are ground into the cylinders. After grinding, the ingots are polish-etched to remove grinding damage and to reduce wafer breakage on slicing.

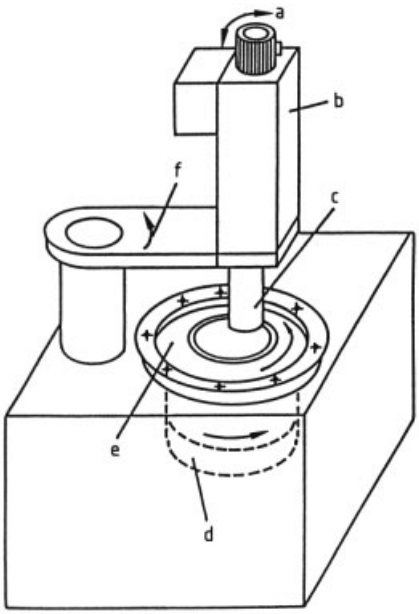

Figure 13. Inner-diameter (ID) sawing machine for slicing silicon
a) Crystal tilting; b) Crystal support and vertical feeding; c) Crystal; d) Saw-blade drive unit; e) ID saw blade; f) Horizontal crystal feeding

Depending on the application of the wafers, the ingot is cut perpendicular to the crystal axis or with a well-defined misorientation of several degrees. These off-axis wafers, after polishing, are normally used as substrates for epitaxial processes. Slicing is performed with inner-diameter saws (see Fig. 13), the saw blade of which is a 0.1-mm-thick high-grade steel foil clamped under high tension between two steel rings. The steel foil has a large circular opening in its center, the edge of which is covered with small diamonds attached to the foil by a galvanic nickel layer. During sawing, the steel foil rotates at ca. 2000 rpm and consumes 0.3 mm of the ingot length (0.3-mm cutting loss). This method generates the flattest and most parallel cuts, and allows the cutting of silicon wafers down to ca. 150-µm thickness with little wafer breakage. For the manufacture of electronic devices, wafers with thicknesses between 150 and 1000 µm are cut.

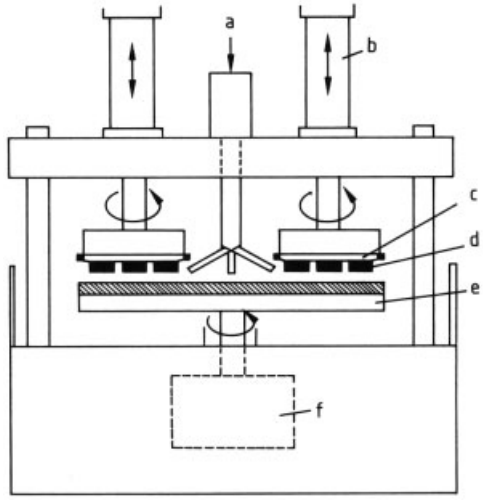

Figure 14. Polishing machine for silicon wafers
a) Polishing slurry feeding; b) Polishing pressure control; c) Wafer carrier; d) Wafers; e) Polishing plate; f) Drive unit

After cutting, the wafers are edge-rounded in special machines, which enable the grinding of different round forms. Edge rounding substantially reduces chipping of the edges in subsequent process steps, imparts a higher mechanical strength to the wafers, and prevents the so-called epi- crown (irregular outgrowth of silicon at the wafer edge) in epitaxial deposition processes and the piling up of photoresist at the edges.

Especially for the mask printing techniques of modern VLSI (very large scale integrated) device manufacturing, the flatness and parallelism of the wafers must be improved by a lapping step. The silicon wafers are lapped between two

large counterrotating steel disks, whereby the wafers are held and led by a lapping carrier. A suspension of an abrasive (a powder of silicon carbide, aluminum oxide, or a mixture of oxides with tightly specified grain size) is added to the process. After lapping, the wafers are etched to remove mechanical damage and then polished.

Figure 14 shows a schematic and Figure 15 a photograph of a polishing machine for silicon wafers. Nowadays, polishing methods basically involve the combined action of chemical and mechanical forces. The polished wafer surface must be free of any residual damage, because such damage may cause stacking faults during subsequent oxidation treatments. Another very stringent demand on surface quality stems from modern photolithographic processing techniques. The polished surface must lie within the depth-of-focus range of the optical systems used to project a masking pattern on the photoresist cover-ing the surface of the wafer.

With the above technology, 200-mm-diameter wafers with both total flatness and total thickness variation of $< 1\,\mu m$ are obtained. After polishing, a careful multistep cleaning process follows that yields a wafer free of particles ($> 0.2\,\mu m$) and residues.

4.2.5. Epitaxy

In epitaxy a crystallographically oriented film is grown as an extension onto an oriented crystal substrate. The growth of an epitaxial film or layer may be realized either by chemical vapor deposition, by physical vapor deposition (e.g., molecular beam epitaxy, MBE), or by liquid-phase epitaxy (LPE). In industrial silicon epitaxy, chemical vapor deposition is employed [18], [19]. With this method, homoepitaxial silicon layers with thicknesses of $0.5 - 120\,\mu m$ are deposited on perfectly polished single- crystal silicon wafers or substrates. The epitaxial or epi layers extend the crystal structure of the substrate exactly and differ only in resistivity and dopant. For device manufacturing, resistivities of ca. $0.03 - 50\,\Omega \cdot cm$ are generated in the layers by doping, mainly with boron, phosphorus, and arsenic.

Epitaxy is performed both on the whole wafer surface (after polishing and prior to device manufacturing) and on small areas of structured wafers (during device manufacturing). In most cases, only one layer is deposited, but in some cases two or three additional layers of different thickness, resistivity, and dopant are deposited. The reasons for using epitaxial layers in device technology are numerous and various [18], [19]. The use of vapor-phase epitaxy in the semiconductor industry began in the early 1960s.

Epitaxial deposition of silicon is in principle the same process as the chemical vapor deposition used for the production of high-purity polycrystalline silicon (see Section 4.2.2). It is carried out at about the same temperature (usually ca. 1400 K) but with a lower growth rate and more precise control. Epitaxial deposition is possible down to ca. 1100 K. At lower temperature the silicon layer becomes polycrystalline or amorphous. The chlorosilanes $SiCl_4$, $SiHCl_3$, and SiH_2Cl_2 are used most frequently as silicon sources, with hydrogen as a reducing gas. The overall simplified reactions are

$$SiCl_4(g) + 2\,H_2(g) \rightleftharpoons Si\,(s) + 4\,HCl\,(g)$$

$$SiHCl_3(g) + H_2(g) \rightleftharpoons Si\,(s) + 3\,HCl\,(g)$$

$$SiH_2Cl_2(g) \rightleftharpoons Si\,(s) + 2\,HCl\,(g)$$

By using trichlorosilane, $SiHCl_3$, and hydrogen at 1400 K, an epitaxial silicon layer grows at about $5\,\mu m/min$. Table 3 gives epitaxial growth rates for various silanes and temperatures [19].

Table 3. Chemical vapor epitaxial growth of silicon films

Silane	Temperature range, K	Attainable growth rate, µm/min
$SiCl_4$	1420 – 1530	0.4 – 15.0
$SiHCl_3$	1370 – 1480	0.4 – 2.0
SiH_2Cl_2	1320 – 1430	0.4 – 3.0
SiH_4	1220 – 1330	0.2 – 0.3

Prior to deposition, the native oxide on the silicon wafer is removed by baking in hydrogen:

$$SiO_2(s) + Si\,(s) \longrightarrow 2\,SiO\,(g)$$

After that, a brief hydrogen chloride etch is generally applied to remove impurities from the wafer surface. During epitaxy, the wafer is held by a susceptor made of high-purity graphite coated with a layer of silicon carbide. The graphite susceptor is also hot during deposition

Figure 15. View of a polishing machine for silicon wafers

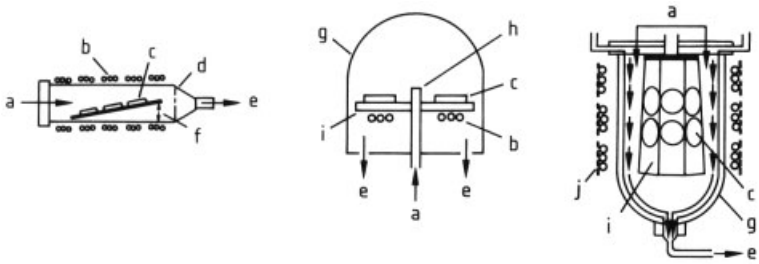

Figure 16. Three types of epitaxy reactors for silicon wafers
a) Gas inlet; b) Induction coil; c) Silicon wafers; d) Silica tube; e) Exhaust; f) Tilt angle; g) Silica bell jar; h) Inlet nozzle; i) Wafer susceptor; j) Radiant heaters

and therefore is also covered with a silicon layer after epitaxy. This unwanted silicon layer is removed by etching with hydrogen chloride. In this way, deposition conditions are identical for each wafer.

To generate electrical doping in the epi layer, gases such as B_2H_6, PH_3, and AsH_3 are added to the chlorosilane – hydrogen mixture. The doping gases decompose and the doping elements are absorbed or dissolved by the growing silicon layer. Because of their low diffusivity, the dopants diffuse only a short distance into the silicon substrate. Epitaxial layers with net charge carrier concentrations in the range of $10^{14} - 10^{20}/cm^3$ can be obtained readily. Figures 16 and 17 show schematically some epitaxial reactors.

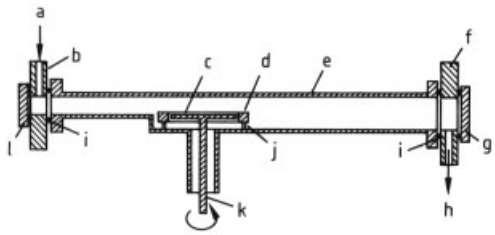

Figure 17. Single-wafer epitaxy reactor for large wafers (according to ASM)
a) Gas injection; b) Gas injector flange; c) SiC- coated, rotating wafer susceptor; d) SiC- coated saturn ring; e) Silica process tube; f) Back flange; g) Rear access port; h) Exhaust; i) O-ring; j) Silica saturn ring support; k) rotating susceptor support; l) Gate valve

4.2.6. Characterization

After crystal pulling, the properties of the silicon crystals, particularly the CZ-Si crystals for megabit devices, must be measured and tested in manifold respects. A great and constantly growing number of measuring and testing methods are employed on silicon. The extent of their application depends on the type and use of the respective material. The most common measuring and testing methods are described briefly below.

Bulk properties:

1) Macroscopic axial and radial distribution of resistivity by two- and four-point probe measurements
2) Microscopic resistivity distribution by spreading resistance measurements
3) Oxygen and carbon concentrations by Fourier transform infrared absorption (FTIR) measurements
4) Oxygen content in samples with high dopant concentration by gas fusion analysis (GFA)
5) Oxygen precipitation behavior in standardized annealings by FTIR measurements and microscopic visual or instrumental determination of the precipitate density and distribution
6) Determination of the content of stacking fault nuclei by oxidation tests and microscopic analysis
7) Testing of the gate oxide integrity (GOI) by a gate oxidation and a subsequent electrical measurement
8) Trap analysis by measuring the minority carrier diffusion length [e.g., with the surface photovoltage (SPV) method]
9) Concentrations of N, B, Al, Ga, P, As, Sb, C−O complexes, thermal donors, etc., by low-temperature (ca. 8 K) FTIR measurements
10) Concentration of electrically active impurities (e.g., metals) and compounds (e.g., FeB) by deep-level transient spectroscopy (DLTS)
11) Concentration of impurities by instrumental neutron activation analysis (INAA; detection limit = 0.00001 – 10 ppba depending on the element)

Geometrical and structural wafer properties:

1) Thickness, parallelism, and flatness by diverse mechanical, optical, or electrical measurements ($\geq 0.1\ \mu m$)
2) Chipping off, scratches, spots
3) Number and size of particles $> 0.15\ \mu m$ on the wafer surface
4) Roughness and haze on the surface by visual or instrumental optical inspection

Purity of the wafer surface: determination of diverse elements and compounds by

1) Inductively coupled plasma mass spectrometry (ICPMS)
2) Total reflection X-ray Fourier analysis (TRXFA)
3) Inductively coupled plasma Auger electron spectroscopy (ICPAES)
4) Vapor-phase decomposition Auger electron spectroscopy (VPDAES)
5) Ion chromatography (IC) and other methods.

4.3. Solar Silicon

Detailed information on solar silicon and its production is given in [5].

A distinction can be made between solar cells for terrestrial applications and solar cells for extraterrestrial applications. In the latter case, the efficiency (output per square meter) is the most important factor; therefore, the best CZ silicon single- crystal qualities are used. Float-zone silicon is not used for extraterrestrial applications because it degrades too strongly on exposure to cosmic radiation, eventually reaching a lower efficiency than CZ silicon. The CZ silicon is more stable owing to its high oxygen content (ca. 15 ppma).

In the case of terrestrial photovoltaic solar cells, the prevailing opinion was initially that solar silicon should be as cheap as possible, and this led to the development of low-priced silicon of low efficiency for solar applications. However, the majority of the costs of a solar module come not from the silicon wafers but from costs that are proportional to the area of the module and are difficult to reduce. At present, about 30 % of the module costs result from the silicon wafers, 30 % from the manufacturing of the solar cell structures, and 40 % from module costs

(metals, glass, machining, etc.). Therefore, developments concentrate increasingly on the production of high-efficiency silicon of moderate or low price. High-efficiency silicon means that the purity and crystal quality of the material must be similar to those of silicon for electronic applications (i.e., monocrystalline and free of dislocations and microdefects). Consequently, solar-grade silicon and electronic-grade silicon will be produced by similar methods in the future.

This trend can already be seen in the development of the solar energy market. The market shares of amorphous silicon and nonsilicon solar materials such as $CuInSe_2$ and CdS are decreasing, while the market shares of multicrystalline silicon and, in particular, single- crystal silicon are increasing.

Table 4 shows the situation of the photovoltaic market in 1990 [20–22]. Photovoltaic (PV) modules for 48 MW_p (p = peak) were produced in total. The value of these PV moduls (ca. $\$400\times10^6$) was about 0.5 % of the value of the electronic devices made from silicon (ca. $\$90\times10^9$).

Table 4. Solar silicon market in 1990

Type of material	Efficiency, %	Module price∗, $/$W_p$	Production MW_p	Share, %
Mono-Si	ca. 15	ca. 4.5	21	44
Multi-Si	ca. 13	ca. 4.5	12	25
a-Si : H	ca. 5	ca. 4.5	15	30
Others		in development		ca. 1

∗ The selling prices are coupled to the output of the modules and not to the manufacturing cost.

4.3.1. Crystalline Silicon

At present, the industry uses the same starting material (high-purity CVD polycrystalline silicon) for the production of crystalline solar silicon as is used for the growth of single- crystal silicon for electronic devices. The production of CVD polycrystalline silicon is described in Sections 4.2.1 and 4.2.2. In 1990, ca. 8000 t of CVD polycrystalline silicon was produced, of which ca. 7000 t was consumed for the production of electronic devices, ca. 700 t for the production of solar cells, and the rest for other applications.

Monocrystalline Silicon. For terrestrial solar cells, CZ and FZ silicon single crystals of $0.5-5\,\Omega\cdot$cm, mostly p-type and sometimes n-type, are used. Float-zone silicon has a market share of about 3 % and is used only for manufacturing very high-efficiency solar cells for light concentrators.

For extraterrestrial applications, CZ crystals of higher resistivity ($7-13\,\Omega\cdot$cm) are used to reduce the degradation of the material by cosmic radiation. The resistivity chosen results from the optimization of the internal resistivity of the solar cell and the diffusion length of the minority charge carriers that are generated by photons and form the electrical current in the solar cell. Increasing the internal resistivity and decreasing the diffusion length reduce the efficiency of a solar cell. The diffusion length of the minority charge carriers is reduced by "traps" that capture the carriers and by too high a dopant concentration. Traps are formed by impurities (e.g., transition metals such as Fe, Mn, Cr, V, and Ti), by small impurity precipitates, and by crystal defects having dangling bonds.

Single- crystal silicon for solar cells is produced in the same way as that for electronic applications (see Section 4.2). The cylindrical silicon crystals are cut to square ingots and then sawed into slices. Square slices give a more complete area covering than round slices.

At present, the efficiency of PV modules made from monocrystalline silicon reaches 18 % in industrial production [23] and 24 % in laboratory production [24].

Multicrystalline Silicon. Multicrystalline silicon is produced by ingot or sheet crystallization.

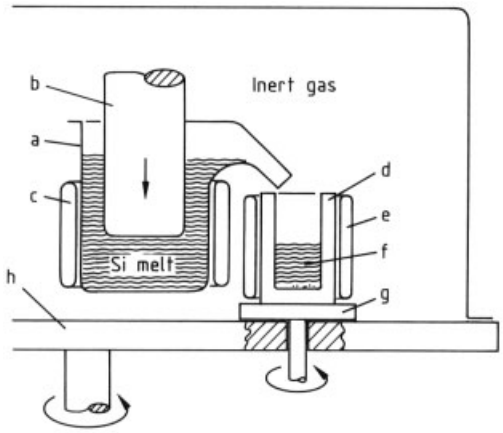

Figure 18. Equipment for semicontinuous production of multicrystalline silicon ingots by mold casting
a) Crucible; b) Piston; c) Heater; d) Mold; e) Heating; f) Cast ingot; g) Rotating cooling plate; h) Base plate

Ingot Crystallization. Several methods are used to produce multicrystalline silicon ingots:

1) Melting in a crucible and crystallization onto a cold plug, which is immersed in the melt and then withdrawn
2) Melting and solidification in the same crucible
3) Melting in a crucible, and casting and solidification in a mold (see Fig. 18)
4) Quasi-continuous melting in and pulling downward through a so-called cold crucible, which is open at the bottom (see Fig. 19)
5) Quasi-continuous casting through a cold crucible or graphite shaper analogous to (4)

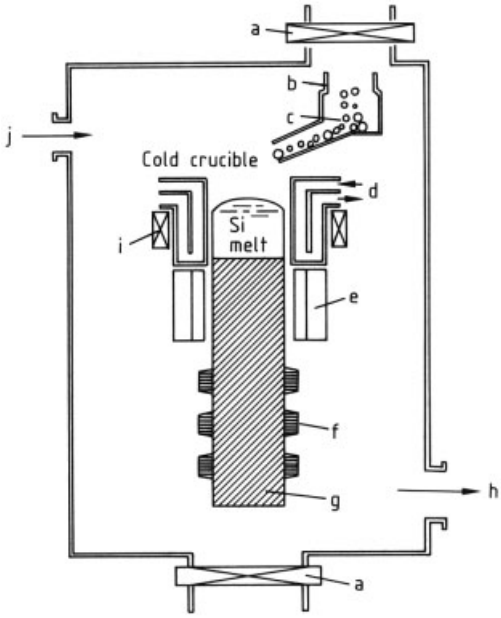

Figure 19. Cold crucible induction casting for multicrystalline silicon ingots
a) Valve; b) Granule feeder; c) Silicon granules; d) Cooling water; e) Auxiliary heater; f) Ingot puller; g) Ingot; h) Vacuum port; i) Induction coil; j) Inert gas inlet

Method (1) is not used for several reasons. Method (2) yields a good crystalline structure owing to its very slow crystallization but includes high crucible costs (crucible is not reusable). Method (3) is the most economical (reusable mold) and also yields good crystal quality. Methods (2) and (3) both yield square ingots, which is a major economic advantage (100 % area coverage).

The ingots must be cut into thin slices (200 – 500 µm thick) for manufacturing solar cells. At present, inner-diameter saws (see Fig. 13) are generally used. However, multiple-wire sawing (Fig. 20) is gaining in importance. Multiple-wire sawing will be the slicing technique of the future for solar silicon because it results in lower material losses and can cut thinner slices.

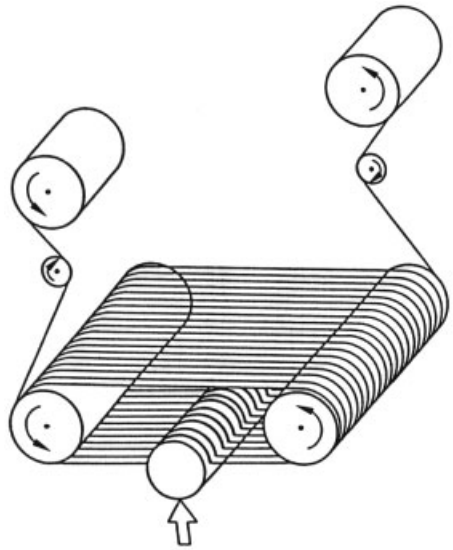

Figure 20. Multiple-wire slurry slicing technique
A single cutting wire is wound onto a delivery spool. The wire is arranged in multiple loops around grooved drive and idle rollers and fixed to the take-up spool

The current efficiencies of multicrystalline solar cells reach 15 %.

Sheet crystallization. The advantage of sheet crystallization is that slicing, which causes high machining costs and material loss, is unnecessary. For economic reasons, sheet crystallization must be carried out at a very high growth rate. However, with such a growth rate, the segregation coefficients of the impurities become unity, and the impurities from the bulk and the surface of the starting polysilicon are incorporated completely into the silicon sheet. In contrast, at the low growth rates in single-crystal pulling or ingot crystallization the impurity concentration of the silicon is reduced drastically (e.g., by a factor of 10^5 for iron). Therefore, the starting silicon for sheet growth must be of a much higher purity than that for crystal or ingot growth.

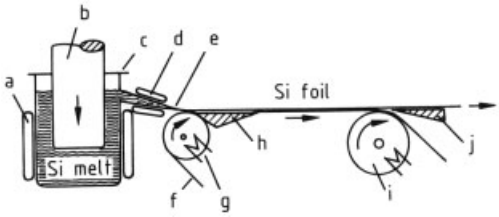

Figure 21. Continuous production of silicon ribbons
a) Heater; b) Piston; c) Crucible; d) Heater; e) Die; f) Foil carrier; g) Heated cylinder; h) Ramp; i) Cooled cylinder; j) Separator

A second disadvantage is that the very high growth rate leads to poor crystal quality (high density of dislocations and grain boundaries). The higher content of both impurities and crystal defects results in lower solar cell efficiency. These two major drawbacks may not be overcome in the future.

Figures 21 and 22 give examples of multicrystalline silicon sheet growth. Many other methods exist for producing such materials [5], [25].

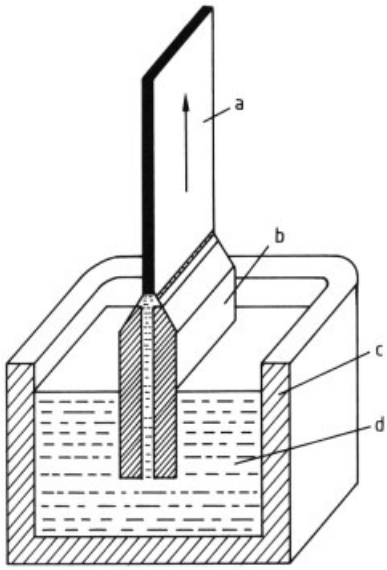

Figure 22. Edge-defined film-fed growth (EFG) or capillary action shaping technique (CAST) for the production of silicon ribbons
a) Silicon ribbon; b) Die; c) Crucible; d) Melt

4.3.2. Amorphous Silicon

Amorphous silicon is often abbreviated to a-Si : H in contrast to c-Si for crystalline silicon. The H in a-Si : H stands for hydrogen, which is needed to neutralize the detrimental dangling bonds. Most of the a-Si : H is used for photovoltaics and is prominent in consumer and low-power applications. In addition to photovoltaics, a-Si : H is employed for electrophotography, in photodetectors, and in accumulation-model field-effect transistors and other electronic devices. In particular, it can be used in integrated functions.

The a-Si : H solar cell is a typical thin-film device. The film is ca. 1 µm thick and is manufactured in large areas in integrated thin-film deposition lines (CVD at 470 – 530 K; see Section 4.2.2). The stabilized efficiency of modules in the field is ca. 5 %. Multijunction devices incorporating a-Si : H/a-Si : H tandems or a-Si : H/a-Si : H/a-Si, Ge : H triple junctions are under development for a stabilized efficiency of 8 % [22].

5. Silicides

Silicides are compounds of silicon with a metal. Figure 23 shows the silicide phases, arranged according to the position of the metal in the periodic table [26], [27].

Many metals of groups 1 – 3 and most transition metals form silicides. Physical properties and structures of silicides are summarized in [26], [27]. The corresponding binary phase diagrams often show the existence of several compounds, having mainly nonstoichiometric compositions and intermetallic character.

Production processes for silicides are based on the following reactions:

1) Fusing or sintering of the metal with silicon:

$$M + Si \longrightarrow MSi$$

2) Reduction of metal oxides by silicon:

$$MO + Si \longrightarrow MSi + SiO_2$$

3) Reaction of metal oxides with SiO_2 and carbon:

$$MO + SiO_2 + C \longrightarrow MSi + CO$$

4) Aluminothermic reaction:

$$MO + SiO_2 + Al \longrightarrow MSi + Al_2O_3 + slag$$

1	2	3	4	5	6	7	8	9	10
H_4Si									
$Li_{15}Si_4$ Li_2Si									
NaSi	Mg_2Si		Ti_5Si_3 TiSi $TiSi_2$	V_3Si V_5Si_5 VSi_2	Cr_3Si Cr_5Si_2 CrSi $CrSi_2$	Mn_3Si Mn_5Si_3 MnSi $MnSi_2$	Fe_3Si Fe_5Si_3 FeSi $FeSi_2$	Co_3Si Co_2Si CoSi $CoSi_2$	Ni_3Si Ni_2Si Ni_5Si_2 Ni_3Si_2 NiSi $NiSi_2$
KSi KSi_6	Ca_2Si CaSi $CaSi_2$	Sc_5Si_3 $ScSi_2$	Zr_4Si Zr_2Si Zr_3Si_2 Zr_4Si_5 ZrSi $ZrSi_2$	Nb_4Si Nb_5Si_3 $NbSi_2$	Mo_3Si Mo_3Si_3 $MoSi_2$		Ru_2Si RuSi $RuSi_2$	Rh_2Si Rh_5Si_3 Rh_3Si_2 RhSi Rh_2Si_3	Pd_3Si Pd_2Si PdSi
RbSi $RbSi_6$	SrSi $SrSi_2$	Y_5Si_4 Y_5Si_3 YSi YSi_2							
CsSi $CsSi_8$	BaSi $BaSi_2$	$LaSi_2$	Hf_2Si Hf_5Si_3 Hf_3Si_2 HfSi $HfSi_2$	$Ta_{4.5}Si$ Ta_2Si Ta_5Si_3 $TaSi_2$	W_5Si_3 WSi_2	Re_3Si Re_5Si_3 $ReSi_2$	OsSi $OsSi_2$ $OsSi_3$	Ir_3Si Ir_2Si Ir_3Si_2 IrSi $IrSi_2$ $IrSi_3$	Pt_3Si Pt_2Si PtSi

Figure 23. Silicides in the periodic table [26]

Silicides have an iron-gray, metallic luster and are resistant to oxidation. Most of the compounds shown in Figure 23 are currently of only academic interest. The silicides of molybdenum, iron, and calcium have achieved commercial significance.

Molybdenum disilicide, a brittle alloy that is very resistant to oxidation, is used as an alloy for heating conductors. It is converted to cermets, from which resistors are then made. Ferrosilicon and calcium silicon are produced in large quantities for use as alloying and refining reagents in steelmaking and casting.

6. Ferrosilicon [28], [29]

The term ferrosilicon [*8049-17-0*] refers to iron – silicon alloys with Si contents of 8 – 95 %.

Table 5 lists the composition of commercial ferrosilicons.

FeSi 75 is commercially the most important, although FeSi 45 is still produced in large quantities for the North American market. Alloys containing > 95 % silicon are classified as silicon metal.

Physical Properties. The Fe – Si phase diagram (Fig. 24) shows the existence of four compounds: Fe_2Si, Fe_5Si_3, FeSi, and $FeSi_2$ [30]. Commercial alloys differ from the compounds shown in the phase diagram with regard to their stoichiometry.

During solidification of the standard grade, FeSi 75, silicon crystallizes initially, followed by solidification of the ξ-phase – the high-temperature phase of $FeSi_2$ – at 1207 °C. FeSi 45 is composed of initially separated FeSi and $FeSi_2$, which is stable below 955 °C [31].

Transformation of the ξ-phase to stable $FeSi_2$ is accompanied by an increase in volume, which can lead to disintegration of the alloy in the range of 45 – 65 % silicon. The transformation and thus the disintegration can be suppressed by supercooling (i.e., rapid solidification in shallow molds).

Structural data for the iron silicides are summarized in Table 6. Table 7 lists the densities and melting ranges of commercial alloys.

Table 5. Composition of ferrosilicons (ISO 5445)

Designation	Chemical composition									
	Si		Al		P	S	C	Mn	Cr	Ti
	Over	Up to and including	Over	Up to and including	(max.)	(max.)	(max.)	(max.)	(max.)	(max.)
FeSi 10	8.0	13.0		0.2	0.15	0.06	2.0	3.0	0.8	0.30
FeSi 15	14.0	20.0		1.0	0.15	0.06	1.5	1.5	0.8	0.30
FeSi 25	20.0	30.0		1.5	0.15	0.06	1.0	1.0	0.8	0.30
FeSi 45	41.0	47.0		2.0	0.05	0.05	0.20	1.0	0.5	0.30
FeSi 50	47.0	51 .0		1.5	0.05	0.05	0.20	0.8	0.5	0.30
FeSi 65	63.0	68.0		2.0	0.05	0.04	0.20	0.4	0.4	0.30
FeSi 75 Al 1	72.0	80.0		1.0	0.05	0.04	0.15	0.5	0.3	0.20
FeSi 75 Al 1.5	72.0	80.0	1.0	1.5	0.05	0.04	0.15	0.5	0.3	0.20
FeSi 75 Al 2	72.0	80.0	1.5	2.0	0.05	0.04	0.20	0.5	0.3	0.30
FeSi 75 Al 3	72.0	80.0	2.0	3.0	0.05	0.04	0.20	0.5	0.5	0.30
FeSi 90 Al 1	87.0	95.0		1.5	0.04	0.04	0.15	0.5	0.2	0.30
FeSi 90 Al 2	87.0	95.0	1.5	3.0	0.04	0.04	0.15	0.5	0.2	0.30

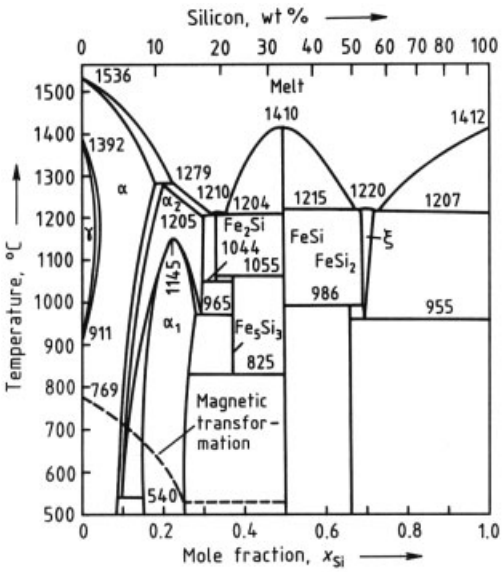

Figure 24. Iron – silicon phase diagram [30]

Table 6. Crystal structure of iron silcides

Compound	Crystal structure	Lattice constants, pm
Fe_5Si_3	hexagonal	$a = 675.5$ $c = 471.7$
FeSi	cubic	$a = 448.9$
$FeSi_2$	tetragonal	$a = 269.2$ $c = 513.7$

Table 7. Density and melting range of commercial FeSi alloys

Alloy	ϱ, g/cm^3	Melting range, °C
FeSi 10	7.3	1280 – 1350
FeSi 25	6.5	1270 – 1350
FeSi 45	5.1	1250 – 1350
FeSi 75	3.2	1250 – 1350
FeSi 90	2.5	1300 – 1400

Raw Materials. Pure quartzes with SiO_2 content $> 98\,\%$ are used for the production of FeSi 45 to FeSi 90. Such quartzes occur naturally as pebbles or rock (vein quartz, quartzite). Quartz sands are processed into briquettes, in which the reducing coal required for a quasi-self-fluxing burden can be incorporated. For good gas flow through the furnace, the starting materials should be free of fine particles, which also requires that they remain dimensionally stable on heating and do not decrepitate too early.

Depending on the location, high-quality iron ore or scrap iron is used as the iron source. In industrialized countries, washed and dried unalloyed steel turnings are generally used.

Flaming coals with low ash content, metallurgical coke, high-temperature brown coal coke, petroleum coke, brown coal briquettes, or wood chips are employed as reducing agent, depending on availability. In calculating the amount of reducing agent required, the carbon resulting from consumption of the Söderberg electrodes must be taken into account; the electrodes supply 5 – 10 % of the total requirement.

The trend is toward the use of finer-grained quartz and coal. For quartz (previously 15 – 100 mm) the particle size now used is 8 – 40 mm; for coal, 2 – 20 mm.

Production. Ferrosilicon is produced in three-phase, submerged arc furnaces with power consumptions of 10 – 70 MW, corresponding to an annual capacity between 9×10^3 and 60×10^3 t per furnace.

Figure 25 shows a schematic of such a furnace. With few exceptions, production is carried out in open furnaces, which allow the burden surface to be poked with stoker machines to prevent crust formation and thus maintain uniform gas flow through the furnace. In modern furnaces the furnace casing can be rotated to reduce encrustation in the lower areas.

Closed furnaces must be built with a rotating furnace vessel to avoid sintering of the burden, since stoker machines can no longer be used. Closed furnaces operate more efficiently (cheaper) because the carbon monoxide gas formed can be used. However, this requires a more elaborate construction of the furnace: with a rotatable hearth and extensive dedusting of the hot off-gases [32].

Generally, large furnaces with capacities of > 40 MW present additional construction and operational problems. The thermal stress of the refractory materials and the off-gas losses also increase because these furnaces have higher operating temperatures [33], [34].

Reactions in the furnace occur according to the simplified scheme:

$$SiO_2 + 2\,C \longrightarrow Si + 2\,CO$$

$$Fe + Si \longrightarrow FeSi$$

Side reactions also occur that result in a lower yield of the desired product, especially when insufficient carbon is used:

$$SiO_2 + C \longrightarrow SiO + CO$$

$$SiO_2 + Si \longrightarrow 2\,SiO$$

Gaseous SiO is oxidized by atmospheric oxygen at the burden surface to give SiO_2 dust, which is carried out of the furnace with the off-gas.

An excess of carbon leads to the formation of SiC, which also lowers the yield.

Specific material and energy requirements for the production of FeSi 45 and FeSi 75 are summarized in Table 8.

Table 8. Specific material and energy consumption for FeSi production

	FeSi 45	FeSi 75
Quartz, kg/t	1200 – 1300	1800 – 2000
Iron turnings, kg/t	550 – 650	230 – 260
Carbon, kg/t	450 – 520	700 – 900
Söderberg paste, kg/t	40 – 50	55 – 70
Energy consumption, kW · h/t	5800 – 6500	8500 – 10 000

Heating the batch to the reaction temperature of up to 1800 °C is achieved mainly by electrical energy (ca. 80 %). The energy uptake of the batch depends on the so-called hearth resistance, whereby the conductivity of the FeSi burden is provided mainly by the coal. Thus, changing the type of coal influences the resistance. The grain size also plays a role: the smaller the grain, the higher is the resistance. The hearth resistance automatically regulates the depth of insertion of the Söderberg electrodes and thus the power consumption of the furnace. A uniformly high hearth resistance and low electrodes are desirable.

The liquid alloy, that accumulates in the hearth is tapped off to pouring ladles at regular intervals (about once an hour) and poured into shallow molds.

Ferrosilicon production is a slag-free process, which means that all the impurities present in the raw materials are transferred to the product. To obtain high purities the alloy must be purified by further treatment outside the furnace.

Aluminum and calcium impurities are removed by oxidation:

1) By injection of gaseous oxygen through immersed lances or through nozzles or sparging blocks in the base of the pouring ladle
2) By treatment with oxidizing siliceous slag, which can be stirred or blown in

Treatments used to obtain the lowest aluminum and calcium concentrations (in the extreme case, 0.02 % max.) also result in silicon losses. The silicon content for these grades drops from 75 to ca. 65 %.

Due to the low affinity of titanium for oxygen, the content of titanium impurities cannot be lowered metallurgically. Low Ti contents must

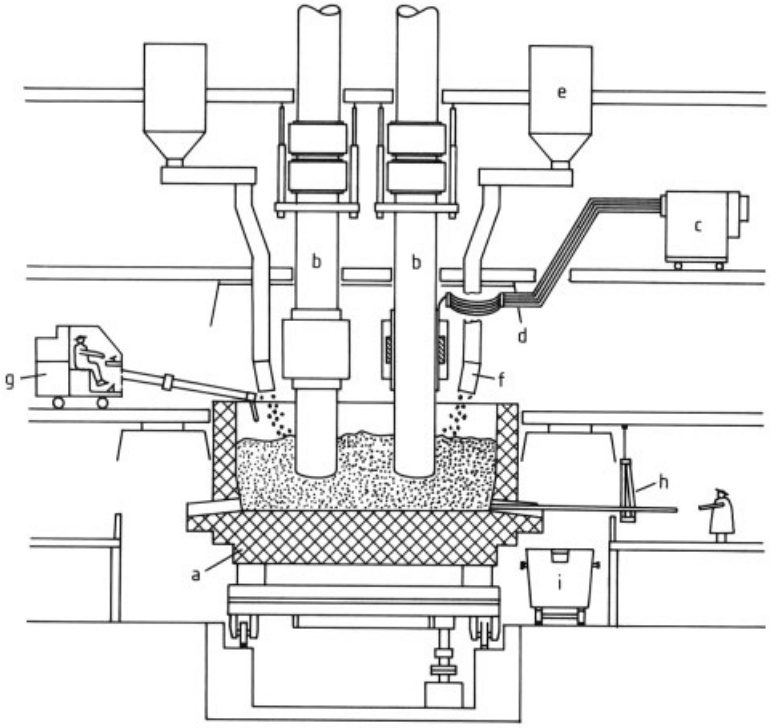

Figure 25. Submerged arc furnace
a) Furnace casing with lining (rotatable); b) Electrodes; c) Transformers; d) Secondary power supply; e) Raw material bunker; f) Charging pipes; g) Stoker machine; h) Burning-out unit; i) Tapping-off ladle

be obtained by using raw materials that contain as little titanium as possible.

Environmental Protection. Silicon monoxide vapor is formed in the reduction process. The SiO is oxidized in air to produce extremely fine SiO_2 dust (particle size 0.1 – 1 µm), which is carried out of the furnace with the off-gas. Furthermore, small amounts of fine particles from the burden are also entrained. A total of 250 – 350 kg of dust is produced per tonne of FeSi 75 [35].

The off-gas from the furnace can be cleaned in filter chambers (bag filters) to the maximum values allowed by TA Luft (20 mg/m^3). The collected dust is sometimes dumped but is now mainly sold as a raw material. Pure SiO_2 dusts are used as cement additives, for example [36], [37].

Since the FeSi process is slag free, it has the advantage of having no slag that must be disposed of.

Quality Specifications. Quality requirements and terms of delivery are summarized in ISO 5445 (see Table 5) and DIN 17 560. In addition, numerous customer specifications exist, some of which require drastically reduced impurity contents. The sieve analysis of ferroalloys is set out in ISO 4551; sampling is described in ISO 4552; and chemical analysis, in ISO 4139 and ISO 4158. Recognized analysis methods are also described in [38–40].

Storage, Transport, and Toxicology. Ferrosilicons produce hydrogen on contact with water. Alloys containing 30 – 70 % Si are classified as dangerous substances, Class 4.3, for transport by road and rail according to GGVS/GGVE and RID/ADR. Alloys with < 30 % and > 70 % Si are excluded from this regulation for land transportation. All commercial FeSi alloys are classified as dangerous goods, Class 4.3, for transportation by ship (IMDG Code) and by air (UN no. 1408, IATA-DGR: Class 4.3).

Ferrosilicon itself is nontoxic. For handling the dust the general MAK values for fine dust (6 mg/m^3) are applicable [41], as well as the usual safety measures such as goggles and face masks.

On contact with water, traces of toxic gases such as phosphine can be formed by reaction with slag adhering to the ferrosilicon. However, if stored in ventilated rooms in accordance with regulations, the critical MAK concentrations are rarely attained (PH_3: MAK 0.1 ppm) [41].

Uses. Approximately 75 % of the FeSi produced is used in the steel industry, where a requirement of 3 – 3.5 kg of FeSi 75 per tonne of steel can be considered normal. For melting almost all grades of steel, silicon is added as a deoxidizer and alloying agent. Silicon binds oxygen dissolved in steel melts, leading to noncritical concentrations. To increase this effect, FeSi is generally added together with other deoxidizers such as aluminum, calcium, and manganese.

Silicon that is not consumed in deoxidation dissolves as an alloying element in the steel and increases its strength and yield point [42]. Conventional construction steels contain 0.2 – 0.4 % silicon. More highly alloyed grades include tool steels, in which silicon improves the hardenability and wear resistance; hot-work steels that contain silicon for better tempering properties; spring steels (up to 2 % Si); and transformer and electro steels (up to 4.5 % Si). Low hysteresis losses are characteristic of the latter materials. Since the magnetic losses depend on the purity of the steel, high-purity FeSi, with drastically reduced content of aluminum, carbon, and titanium, is used.

Foundries consume almost 25 % of the ferrosilicon produced worldwide. Normal cast iron contains 2 – 3 % silicon for improved precipitation of graphite and increased strength.

Ferrosilicon alloys are also used as carrier alloys for metals such as barium, strontium, calcium, and titanium, usually in amounts of 2 – 4 %. These treatment alloys for gray cast iron are added to the ladle or mold to improve the precipitation of graphite. Ferrosilicon is also the prealloy for FeSiMg alloys; magnesium causes the formation of spherical graphite in spherulitic graphite iron.

Economic Aspcects. In 1990, world consumption of ferrosilicon, calculated as FeSi 75, was 3.4×10^6 t, which compared to a capacity of 5.5×10^6 t. Thus capacities were utilized only to ca. 62 %. Usually, the rate of utilization varies between 65 and 75 %. Thus, the economic situation at the beginning of the 1990s was characterized by marked pressure on prices and by the closure of unprofitable operations. However, new capacities of ca. 250×10^3 t/a are now planned, with almost 200×10^3 t/a in Venezuela alone.

The choice of plant location depends less on the raw materials (quartz, coal) than on the availability of cheap electricity. Production is spread worldwide; the CIS and China are important producers (each with a capacity of ca. 10^6 t/a), followed by Norway, the United States, and Brazil (each ca. 0.5×10^6 t/a).

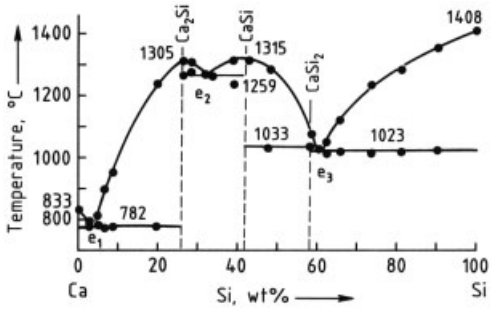

Figure 26. Calcium – silicon phase diagram [44]

7. Calcium Silicon [43]

Like FeSi, calcium silicon [*8049-17-0*] is produced carbothermally in submerged arc furnaces and is consumed in the iron and steel industry.

Physical Properties. The Ca – Si phase diagram (Fig. 26) shows the existence of three compounds [44]: Ca_2Si (cubic, $a = 474$ pm), CaSi (orthorhombic, $a = 391$ pm, $b = 459$ pm, $c = 1080$ pm), and $CaSi_2$ (hexagonal rhombohedral, $a = 1040$ pm, $\alpha = 21.5°$). Commercial CaSi contains ca. 30 % calcium and 60 % silicon; the remainder is iron and impurities. The alloy solidifies mainly as $CaSi_2$ with small amounts of primary and secondary eutectically precipitated silicon. The density of the commercial alloy is 2.2 g/cm^3, and its melting range is 1050 – 1150 °C.

Raw Materials and Production. The raw materials for the production of CaSi are quartz, lime, and coal. Whereas calcium carbide was formerly used as the source of calcium, it has now been replaced by lime with the following approximate specifications: CaO 94 – 97 %, CO_2 1.5 – 3 %, H_2O 0.5 – 1.5 %.

The requirements for quartz and coal, as well as the particle size of the burden, correspond to those of ferrosilicon production.

The three-phase submerged arc furnaces generally used for the production of CaSi are similar to the furnaces used for the production of FeSi (see Fig. 2). In practice, CaSi and FeSi are also produced alternately in the same furnace. Calcium silicon is produced only in open furnaces whose power consumption is 8 – 30 MW, corresponding to a capacity of $(6-20)\times 10^3$ t/a. The basic reaction of the commercial process is as follows:

$$CaO + 2\,SiO_2 + 5\,C \longrightarrow CaSi_2 + 5\,CO$$

In practice the processes are very complex. The process is sensitive to variations in the amount of reducing agent, which may also occur locally in the burden. An excess of carbon in the burden results in the formation of SiC and CaC_2, which in the extreme case leads to the clogging of the furnace. The use of too little carbon results in the formation of lime – silicate slags that are difficult to reduce and are tapped with the metal, which lowers the capacity of the furnace.

Even with correct operation, up to 30 % of slag is carried out with the CaSi. The slag, which is only slightly denser than CaSi, is separated after being tapped off into the pouring ladle. The slag sinks to the bottom, and the metal is then decanted by pouring into shallow molds. Some companies use a cascade of four to five tapping-off molds, which results in refinement because the slag preferably remains in the upper molds and the lower ones contain very pure metal. Careful separation from slag is the only feasible method for refining CaSi. The oxidative treatment used for FeSi is not possible because calcium has a high affinity for oxygen.

The slag, which also contains CaC_2, is processed and recycled to the furnace.

Typical values for the specific material and energy requirements of the CaSi process are as follows:

Quartz	1500 – 1 800 kg/t
Lime	550 – 700 kg/t
Coal	680 – 800 kg/t
Recycled slag	250 – 300 kg/t
Söderberg paste	80 – 100 kg/t
Power consumption	9800 – 13 000 kW · h/t

Environmental Protection. The large amount of dust generated in the CaSi process (up to 300 kg per tonne of CaSi) is a result of vapor losses of SiO and calcium. Both components are oxidized in air and carried out of the furnace with the off-gas, from which they are separated in bag filters. The dust, which consists of ca. 60 % SiO_2 and 30 % CaO (remainder: MgO, carbon, Fe_2O_3), is dumped.

Since all of the slag is reintroduced to the process, none remains to be disposed of.

Quality Specifications. The specifications for standard-grade commercial CaSi are summarized in Table 9. A number of multicomponent CaSi-based alloys contain up to 30 % aluminum, manganese, barium, titanium, magnesium, mischmetal, or iron. These compositions are often specified by the customer and are not defined in standards.

DIN 17 580 also defines procedures for analysis, testing, and arbitration [39], [45], [46].

Storage, Transport, and Toxicology. On contact with water, acid, or base, CaSi can produce hydrogen. The reaction proceeds extremely slowly with pure water and the alloy is therefore not classed as a hazardous substance in most countries (GGVS/GGVE) for transport by road.

For sea and air transport, CaSi is classed as a hazardous substance, Class 4.3 (IMDG-Code; UN no. 1405). Since fine CaSi dusts are explosive, inert gases (nitrogen or argon) must be used for pneumatic transport. The critical oxygen content for carrier or protective gases is 8 %.

Calcium silicon should be stored in dry, ventilated rooms. On contact with moisture, slag impurities can release small amounts of phosphine, although with correct storage the concentrations are not critical.

Uses. Calcium silicon is used in steelmaking as a deoxidizer and desulfurizer, as well as for

Table 9. Composition of CaSi (DIN 17 580)

Name	Symbol	Material no.	Composition, wt %					
			Ca	Ca + Si (min.)	C (max.)	Al (max.)	P (max.)	S (max.)
Calcium – silicon	CaSi	0.3650	29 – 33	90	1.20	1.80	0.070	0.060
Calcium – silicon, low C	CaSiC 50	0.3655	29 – 33	90	0.50	1.80	0.070	0.050

the modification of nonmetallic inclusions [47], [48].

Calcium is the active element in the alloy due to its high affinity for oxygen and sulfur; silicon serves as a carrier element and is alloyed in the steel melt, analogous to FeSi (see page 27).

The efficiency of calcium is very low if the alloy is added in lumps because of the high vapor pressure of calcium at the utilization temperature (p_{Ca} at 1600 °C: 0.2 MPa) and the low solubility of calcium in molten iron (max. 0.03 %).

The efficiency has been improved by the development of injection techniques. Calcium silicon powder is blown deep into the steel melt through immersion lances, or CaSi-filled hollow wires are coiled rapidly into the pouring ladle. These techniques produce very good metallurgical results and are rapidly replacing conventional addition of the alloy as lumps.

The modification of inclusions by calcium improves the fluidity of steel in automated continuous casting plants. The calcium liquefies the suspended solid inclusions of aluminum oxide that are always present in aluminum-killed steels. Furthermore, the material properties of the steel are improved: the calcium aluminates formed are largely undeformable in the solidified steel and allow high isotropy of the material properties to be maintained after working of the steel.

Economic Aspects. The amount of CaSi required by the steel industry has decreased during the 1980s due to changes in process technology: the very efficient injection techniques, especially the cored wire process, greatly reduce the amount of CaSi required per tonne of steel.

In spite of the wide range of uses, world consumption fell from ca. 100×10^3 t/a in 1981 to ca. 80×10^3 t/a in 1991. In the Western steelproducing countries, ca. 50 % is used in the form of filled wire, and this fraction will increase further.

Production capacity is estimated to be 120×10^3 t/a, distributed evenly among the former Eastern-bloc countries, South America, and Western Europe.

8. References

1. *Properties of Silicon,* EMIS Datareviews Series No. 4, INSPEC, London 1988.
2. *J. Phys. Chem. Reference Data* **14** (1985) 1.
3. L. P. Hunt, V. D. Dosaj, J. R. McCormick, *Quart. Rep. DOE/JPL-95 4559-76/2* 1977, Jan.
4. Dow Corning, DE-OS 3 013 319, 1979 (V. D. Dosaj, L. P. Hunt).
5. J. Dietl, D. Helmreich, E. Sirtl: *Crystals,* **vol. 5,** Springer-Verlag, Berlin 1981, p. 43.
6. A. Sanjurjo et al., *J. Electrochem. Soc.* **128** (1981) 179.
7. L. Nanis et al.: *Final Rep. DOE/JPL-95 4471* 1980, March.
8. Ethyl Corp., EP-A 0 112 151 A 1, 1984 (E. M. Marlett).
9. Siemens & Halske, DT 1 102 117, 1954; DT 1 140 549, 1954. Siemens-Schuckertwerke, US 3 042 494, 1958 and 1962 (H. Gutsche).
10. E. Sirtl: *Festkörperprobleme VI,* Pergamon-Vieweg, Braunschweig 1967, p. 1.
11. Ethyl Corp., US 4 784 840, 1988 (M. F. Gautreaux, R. H. Allen); US 4 820 587 and 4 883 687, 1989 (M. F. Gautreaux, R. H. Allen).
12. KRICT, US 4 900 411, 1990 (Y. Poong, S. Yongmok).
13. Texas Instruments, US 3 963 838, 1976 (H. S. N. Setty, C. L. Yaws, B. R. Martin, D. J. Wangler).
14. J. Czochralski, *Z. Phys. Chem.* **92** (1917) 219.
15. W. Zulehner, D. Huber: *Crystals,* **vol. 8,** Springer-Verlag, Berlin 1982, p. 1.
16. W. Zulehner in *Materials Science Engineering,* **vol. B 4,** Elsevier Sequoia, Lausanne 1989, p. 1.
17. W. Zulehner in K. Sumino (ed.): *Defect Control in Semiconductors,* Elsevier Science, North Holland 1990, p. 143.

18. S. B. Kulkarni in N. G. Einspruch, G. B. Larrabee (eds.): *VLSI Electronics: Microstructure Science,* **vol. 6,** Academic Press, New York 1983, p. 73.
19. T. L. Chu in N. G. Einspruch (ed.): *VLSI Handbook,* Academic Press, New York 1985, p. 285.
20. H. Herzer, *Proceedings of the 10th European Photovoltaic Solar Energy Conference,* Lisbon, Portugal, April 1991, p. 501.
21. New Energy Development Organization (NEDO, Japan), Report no. 1, 1990; *Photovoltaic Insider Report* (*PVIR*) **2** (1991).
22. S. Wagner, D. E. Carlson in [20], p. 1179.
23. N. B. Mason, D. Jordan in [20], p. 280.
24. M. A. Green in [20], p. 250.
25. A. Eyer, A. Räuber, A. Goetzberger, *Optoelectronics – Devices and Technologies* **5** (1990) no. 2, 239.
26. R. Kiefer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963.
27. R. Kiefer: *Sondermetalle,* Springer Verlag, Wien 1971.
28. K. A. Feldmann, K. D. Frank in R. Durrer, G. Volkert (eds.): *Metallurgie der Ferrolegierungen,* Springer Verlag, Berlin 1972.
29. W. P. Eljutin, I. A. Palow, B. E. Levin: *Ferrolegierungen,* VEB Verlag Technik, Berlin 1953.
30. E. Schürmann, N. Hensgen, *Arch. Eisenhüttenwes.* **51** (1980) no. 1, 1 – 4.
31. J. A. De Huff, V. D. Coppolecchia, A. Lesnewich, *Electr. Furn. Proc.* **27** (1969) 167 – 174.
32. H. Krogrund, *ISLAFA Conference,* Acapulco 1978.
33. DEMAG company brochure, Submerged – Arc Furnaces, Duisburg 1979.
34. T. W. Lopuszynski, J. P. Trunco, W. L. Wilbern, *Electr. Furn. Proc.* **30** (1972) 89 – 93.
35. U.S. Environmental Protection Agency Office of Air and Waste Management Research Triangle Park, N. C., 27711 EPA – 450/2-74-008, 1974.
36. G. Rau, *Erzmetall* **44** (1991) no. 11, 557 – 559.
37. P. C. Aitcin, *Electr. Furn. Proc.* **42** (1984) 301 – 310.
38. ASTM Designation: *Annual Book of ASTM Standards, E 360 – 70 T, ASTM, Philadelphia,* 1970.
39. *Handbuch für Eisenhüttenlaboratorium,* **vol. 2,** Verlag Stahleisen mbH, Düsseldorf.
40. *Analyse der Metalle,* **vol. 1**, Springer Verlag, Berlin 1980.
41. DFG: *Maximale Arbeitsplatzkonzentration und Biologische Arbeitsstofftoleranzwerte,* VCH Verlagsgesellschaft, Weinheim 1991.
42. E. Houdremont: *Handbuch der Sonderstahlkunde,* Springer Verlag, Berlin 1956.
43. H. Walter in R. Durrer, G. Volkert (eds.): *Metallurgie der Ferrolegierungen,* Springer Verlag, Berlin 1972.
44. E. Schürmann, H. Litterscheid, P. Fünders, *Arch. Eisenhüttenwes.* **45** (1974) no. 6, 367 – 371.
45. "Probenahme", *Handbuch für das Eisenhüttenlaboratorium,* **vol. 3**, Verlag Stahleisen, Düsseldorf.
46. "Schiedsverfahren", *Handbuch für das Eisenhüttenlaboratorium,* **vol. 4**, Verlag Stahleisen, Düsseldorf.
47. B. Neuer, *Erzmetall* **42** (1989) no. 6, 239 – 243.
48. *Calcium Treatment Symposium Proceedings,* The Institute of Metals, London, 1988, Book 459.

Silicon Carbide

KLAUS LIETHSCHMIDT, Elektroschmelzwerk Kempten GmbH, Werk Grefrath, Frechen, Federal Republic of Germany

1. Introduction 1
2. Properties 2
3. Production 4
3.1. Reaction Mechanism 4
3.2. Raw Materials 5
3.3. The Acheson Process 5
3.4. The ESK Process 6
3.5. Processing of the Crude Product . . . 7
3.6. Other Production Methods 7
4. Quality Control 8
5. Toxicology and Occupational Health . 9
6. Uses 9
6.1. Abrasives 9
6.2. Refractories 9
6.3. Metallurgy 10
6.4. Electrical Technology 10
6.5. Electronics 11
6.6. Structural Ceramics 11
7. Economic Aspects 11
8. References 11

1. Introduction

Although silicon carbide has been an industrial product for over 100 years, it is still as important as ever, and developments have led to further new applications. This success is due to silicon carbide's special chemical and physical properties, especially its high chemical resistance, its great hardness, and its semiconducting properties. It is consequently used in the production of abrasives, refractory materials, heating elements, voltage-dependent electrical resistors, light-emitting diodes, and structural ceramic components. Silicon carbide also has the ability to dissolve in molten iron, and it is therefore used in ferrous metallurgy as an alloying and deoxidizing agent.

History [1,2]. BERZELIUS first reported the formation of silicon carbide in 1810 and 1821. It was later rediscovered during various electrothermal experiments (DESPRETZ, 1849; MARSDEN, 1881). However, it was ACHESON who first recognized the potential industrial importance of silicon carbide and exploited it. In the summer of 1891, ACHESON passed a high electric current through a mixture of powdered coke and clay. He noticed some very hard, shiny crystals on the carbon electrode that served as electrical supply. He assumed that these were a compound of carbon with corundum, and named the material carborundum. However, the chemical analysis carried out by ACHESON's coworker MÜHLHÄUSER showed that these crystals consisted of carbon and silicon in the atomic ratio 1: 1, i.e., it was silicon carbide, with the chemical formula SiC. ACHESON soon discovered that these crystals had much better abrasive properties than the emery powder then in general use, and he founded the Carborundum Company in September 1891 for the production of silicon carbide. ACHESON protected his production process, which used a resistance furnace, by several patents (1892 – 1894). The Acheson process for the production of SiC is still operated today with some improvements in detail (see Section 3.3).

An important development in the production of SiC in resistance furnaces took place in Germany in 1972 – 1974. Elektroschmelzwerk Kempten (ESK) developed a furnace in which the gases produced in the reaction are collected and used for energy production. The ESK process also made an important contribution to environmental protection (see Section 3.4).

Natural Occurrence [3]. MOISSAN found some hexagonal SiC crystals in a sample of a meteorite from the Diablo Canyon in Arizona (1905). The mineral was named moissanite after its discoverer. Later it was found that the Diablo Canyon sample was contaminated with synthetic SiC. Not until 1957 was the natural occurrence of SiC confirmed when it was discovered in Siberian kimberlite. Since then, moissanite has been found in various parts of the world, in plutonites, vulcanites, sediments, and metamor-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a23_749

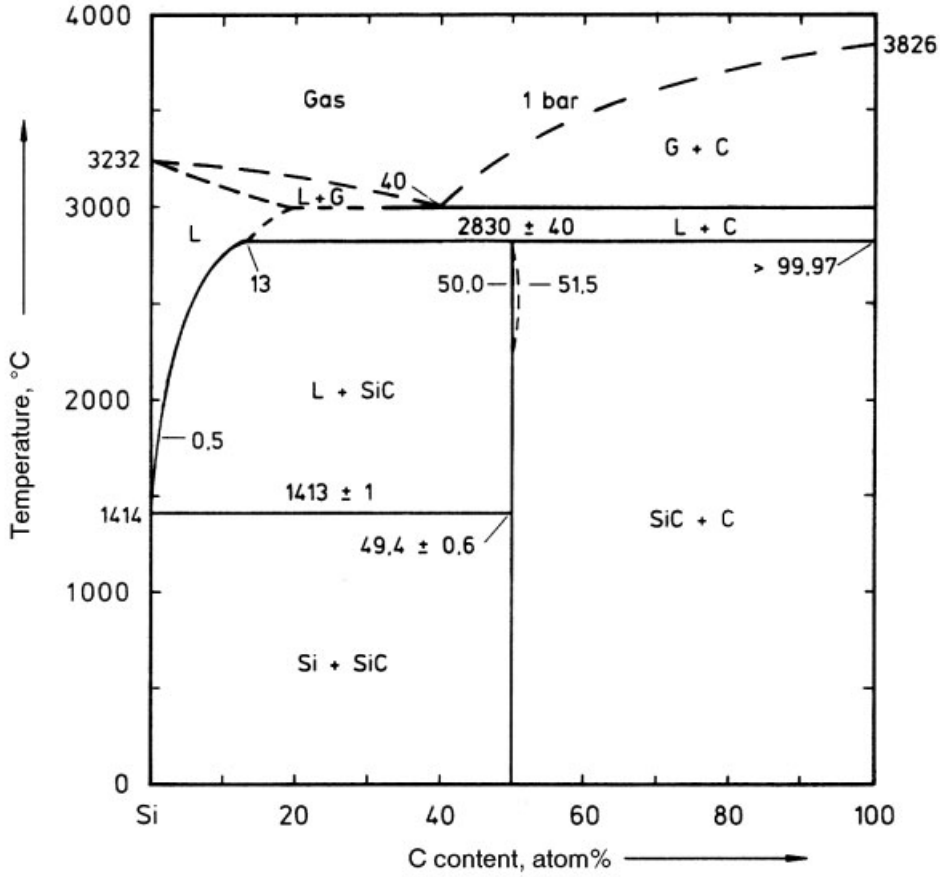

Figure 1. Phase diagram of a silicon – carbon system [4]

phites. It always exists as single crystals with a cross section of ca. 1 mm. Extraterrestrial occurrences of SiC are even rarer. Its existence in interstellar space has been shown by spectral measurements, and it has also been found in samples of moon rock.

2. Properties

The two-component system silicon – carbon contains the single compound silicon carbide [*409-21-2*], SiC, M_r 40.097, $\varrho = 3.21$ g/cm^3, with the composition 70.05 wt % Si, 29.95 wt % C. The Si – C bond is homopolar. Silicon carbide does not exhibit a congruent melting point. In a closed system, i.e., under equilibrium conditions at a total pressure of 0.1 MPa, SiC decomposes at 2830 ± 40 °C into carbon and a silicon-rich melt (Fig. 1) [4]. In an open system (e.g., in the Acheson furnace) decomposition begins at ca. 2300 °C whereby graphite remains behind as a pseudomorphic residue (i.e., having the shape of SiC).

Crystal Structure. Silicon carbide exists as both polymorphs and polytypes. Each C atom is surrounded tetrahedrally by four Si atoms, and each Si atom by four C atoms. These tetrahedra are arranged in planes perpendicular to the *c*-axis. Relative to one other, they can occupy the A, B, or C position of the close-packed structure. Depending on the stacking sequence, the unit cell therefore has a variable number of SiC layers in the direction of the *c*-axis. Consequently silicon carbide occurs in two modifications: cubic β-SiC and α-SiC, which includes a large

number of hexagonal and rhombohedral crystal types. The latter have been termed polytypes by several authors. RAMSDELL [5] classifies β-SiC and the various α-types by the number of SiC layers per unit cell and the crystal classification (C: cubic, H: hexagonal, R: rhombohedral). Industrial SiC contains almost exclusively β-SiC 3 C and the α-types 4 H, 6 H, and 15 R [6].

β-SiC is metastable and, in accordance with Ostwald's rule, is formed initially in SiC production. It changes into α-SiC above 1900 °C. In a nitrogen atmosphere at high pressure, β-SiC is the stable form, so that above 2000 °C α-SiC transforms to β-SiC [7]. LUNDQUIST [8] observed that the proportion of crystal types in α-SiC depends significantly on the aluminum content (Fig. 2). Pure α-SiC mainly crystallizes as 6 H. With small amounts of Al, 15 R also occurs. Above 0.1 % Al, 4 H predominates.

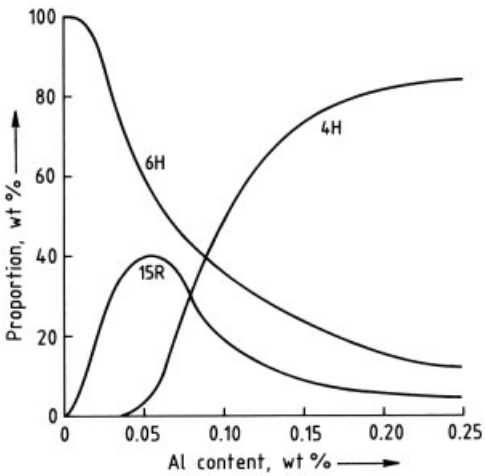

Figure 2. Proportions of polytypes in technical-grade SiC as a function of Al content

Color. Pure α-SiC is colorless, while the cubic β-modification is yellow. The only other elements that can be included in the SiC crystal lattice in amounts > 1 ppm are N, Al, and B. Nitrogen gives a green color to 3 C and 6 H, and a yellow color to 4 H and 15 R. The presence of the trivalent elements boron and aluminum gives all the modifications and polytypes a blue-black color [9].

Hardness. The industrial importance of silicone carbide is mainly due to its extreme hardness of 9.5 – 9.75 on the Mohs scale. Only diamond, cubic boron nitride, and boron carbide are harder. The Knoop microhardness depends on the crystal orientation, and lies in the range of $H_{K,0.1}$ 2100 – 2900 [10]. The effects of crystal type and doping (color) are small and insignificant compared with this hardness range.

Silicon carbide is very brittle, and can therefore be crushed comparatively easily in spite of its great hardness.

Thermal and Calorific Properties. For a ceramic material, silicon carbide has an unusually high thermal conductivity: 150 kJ m^{-1} h^{-1} K^{-1} at 20 °C, and 54 kJ m^{-1} h^{-1} K^{-1} at 1400 °C [11]. α-SiC has a very low thermal coefficient of expansion of 4.7×10^{-6} K^{-1} (20 – 1400 °C), which is the reason why silicon carbide refractory products are so outstandingly resistant to thermal shock.

The specific heat capacity of SiC is 0.67 J g^{-1} K^{-1} at room temperature, and 1.27 J g^{-1} K^{-1} at 1000 °C. The standard enthalpy of formation ΔH^0_{298K} is -71.6 ± 6.3 kJ/mol, and the entropy S^0_{298K} is 16.50 ± 0.13 J mol^{-1} K^{-1} [12].

Optical Properties. α-SiC is birefringent due to its crystal structure: $n_0 = 2.649 - 2.649$, and $n_E = 2.688 - 2.693$ (Na 589 nm, 20 °C) [13]. For β-SiC, a refractive index of ca. 2.63 (Li 671 nm) has been reported [14].

Electrical Properties. Silicon carbide is a semiconductor. Doping with the trivalent elements aluminum and boron gives the SiC *p* character, while the pentavalent element nitrogen produces *n* character when incorporated into the SiC crystal lattice. The resistivity lies between 0.1 $\Omega \cdot$ cm and 10^{12} $\Omega \cdot$ cm, depending on the concentration of lattice defects [15].

Whereas compact, homogeneous silicon carbide obeys Ohm's Law, aggregates of SiC grains show nonlinear current – voltage behavior. At low applied voltages they behave as insulators, but when the applied voltage is increased above a certain value, the current increases exponentially. Thus, the points of contact between the grains cause the electrical resistance to be voltage dependent [15, 16].

Chemical Properties. In addition to its hardness, an important property of silicon carbide is

its chemical resistance. It is resistant to organic solvents, alkalies, acids, salt solutions, and even aqua regia and fuming nitric acid.

The behavior of silicon carbide towards oxygen and oxygen- containing gases is of industrial importance. The oxidation of pure SiC begins at ca. 600 °C, forming a coating of SiO_2 on the surface of the SiC that prevents further oxidation [17].

$$SiC + 2\,O_2 \longrightarrow SiO_2 + CO_2 \quad (1)$$

The reaction rate varies with time according to a parabolic law [18]. The kinetics are determined by the diffusion of oxygen through the SiO_2 layer. The effect of temperature on the reaction obeys the Arrhenius Law.

The "active" oxidation of SiC is distinguished from the "passive" oxidation reaction described above.

$$SiC + O_2 \longrightarrow SiO + CO \quad (2)$$

Active oxidation takes place under conditions of oxygen deficiency above 1000 °C, and leads to decomposition of the SiC and formation of SiO [17, 19].

Thus, silicon carbide is attacked and decomposed by an oxidizing agent if the protective layer of SiO_2 is removed, thereby enabling the reaction to proceed unhindered. SiC is decomposed in this way by fused alkalies such as $Na_2CO_3 + Na_2O_2$ or $Na_2CO_3 + KNO_3$. Fused lead borate in the presence of oxygen causes complete decomposition of SiC. Pb_3O_4 or a mixture of $K_2Cr_2O_7$ with $PbCrO_4$ reacts very vigorously on heating with silicon carbide.

A mixture of hydrofluoric acid, nitric acid, and sulfuric acid slowly attacks silicon carbide, β-SiC being somewhat more reactive than α-SiC [20]. However, complete dissolution only takes place if the SiC is very finely divided, under pressure, and at a high temperature (e.g., 250 °C for 16 h).

Chlorine reacts with SiC above 800 °C, forming silicon tetrachloride and carbon.

Silicon carbide behaves towards molten metals in various ways. It is not attacked by molten zinc or zinc vapor [21]. Molten aluminum attacks SiC slowly, forming Al_4C_3 and Si, but as the Si content increases, the reaction eventually ceases because an equilibrium is established [22]. Molten iron dissolves SiC, forming iron carbide and iron silicide.

3. Production

3.1. Reaction Mechanism

Silicon carbide is produced industrially from silicon dioxide and carbon, which react according to the overall equation:

$$SiO_2 + 3\,C \longrightarrow SiC + 2\,CO \quad (3)$$

The reaction is strongly endothermic with $\Delta H_{298K} = +618.5$ kJ/mol [12], corresponding to 4.28 kWh/kg α-SiC.

The reaction in fact takes place in several stages. The initiation reaction

$$SiO_2 + C \longrightarrow SiO + CO \quad (4)$$

is thermodynamically possible above ca. 1700 °C and in fact begins when the SiO_2 melts. The next stages are

$$SiO + C \longrightarrow Si + CO \quad (5)$$
$$Si + C \longrightarrow SiC \quad (6)$$

The SiC is formed as an incrustation on the carbon grains, preventing further formation of SiC, and ultimately causing the reaction to cease. The SiC then reacts further

$$SiC + 2\,SiO_2 \longrightarrow 3\,SiO + CO \quad (7)$$
$$SiC + SiO \rightleftharpoons 2\,Si(g) + CO \quad (8)$$

The equilibrium reaction (Eq. 8) is mainly responsible for the formation of the large SiC crystals [23].

3.2. Raw Materials

The yield and quality of silicon carbide are seriously affected by any impurities present in the raw materials. Therefore high-purity raw materials must be used.

The source of silicon is very pure, natural silica sand or ground quartz. The SiO_2 content is 99.0 – 99.9 %, preferably > 99.7 %, depending on the quality of SiC required. The usual impurities are Fe_2O_3 (0.01 – 0.05 %) and Al_2O_3 (0.02 – 0.50 %) [9].

The preferred carbon source is low-ash petroleum coke, with a typical composition: 93.7 % C_{fix}, 6.2 % volatiles, and 0.1 % ash [24]. Charcoal and anthracite are also used to produce silicon carbide, but these materials are not sufficiently pure to produce the highest quality SiC grades such as green SiC.

The raw materials should have a maximum grain size of 10 mm. The mixing ratio is precisely determined from the analysis of the raw materials, and corresponds to the stoichiometry given by Equation 3. To produce 1 t SiC, 1.5 – 1.6 t silica sand and 1 – 2 t coke or coal are required [25].

Some producers add 1 – 2 % sodium chloride to the mixture of raw materials, which is intended to remove metallic impurities by converting them to volatile chlorides. However, the effectiveness of this technique is disputed [26].

Sawdust is mixed with the raw materials to give a loosening effect. This promotes release of the carbon monoxide formed at the same time as the SiC.

3.3. The Acheson Process

The process developed by Acheson is still operated today essentially unchanged [1,27, 28]. Electrical resistance furnaces of rectangular cross section are used (Fig. 3). The electric current is fed into the furnace via graphite or carbon electrodes inserted in the end walls. The side walls are removable to facilitate discharging of the product.

The furnaces are filled with the raw material mixture by a crane. When about half of the material has been charged, the heating resistance, made of graphite, calcined pitch coke, or petroleum coke, is inserted, joining together the two electrodes in the end walls of the furnace. This so- called core is then covered with the remainder of the reaction mixture.

The electrodes are then connected to the transformer, and the current is switched on. The silicon carbide is formed around the hot core in the form of a polycrystalline, compact cylinder (the so- called roll, see Fig. 4). The surrounding reaction mixture acts as a thermal insulator. The carbon monoxide byproduct escapes through the surface of the reaction mixture and through slits in the side walls, where it burns. Because the reaction mixture is consumed during SiC formation, more reaction mixture is charged periodically to the process.

After the current is switched off, the furnace is allowed to cool for several days. Then the side walls are removed, and the excess reaction mixture is carried off. The roll of silicon carbide contains not only the original resistance core, but also graphite formed by decomposition of silicone carbide. The boundary between the graphite and the SiC is very sharply defined. The innermost SiC zone is of the highest quality, as the purest and largest crystals can form in this region. Moving outwards, i.e., with decreasing temperature, the crystal growth rate decreases, and the SiC roll becomes more and more finely crystalline. The outer zone is composed of very fine crystals of β-SiC. The SiC roll is covered with a whitish crust (siloxicon or firesand), which consists of coke grains partially converted to SiC, and compounds of silicon with oxygen and nitrogen. This is the region of the initiation reaction (Eq. 4), with which SiC formation begins.

When the SiC roll has been isolated and cooled, it is broken down piece by piece and taken for further treatment (see Section 3.5). The graphite from the core is recovered and reused.

Without exeption, several furnaces are operated in a group and associated with a single transformer. In this way, one furnace is always in operation while the others are cooling, being broken down, or being recharged.

Acheson silicon carbide furnaces can be up to 25 m long, 4 m wide, and 4 m high, consuming up to 5000 kVA, and running for ca. 130 h. The largest furnaces operate at up to 25 000 A.

The electrical energy is usually supplied as three phase a.c. Therefore, a silicon carbide production plant usually employs three groups of furnaces, so that a symmetrical load is applied to the supply network.

For optimum operation of the installed power supply, the wattless current should be as low as possible. This is achieved by suitable design and construction of the secondary circuit, and by capacitative compensation for the wattless current. An alternative to this is to operate the furnaces with d.c. [29].

Acheson silicon carbide furnaces are installed in large buildings with open roof hoods. The CO_2 and SO_2 formed by the combustion

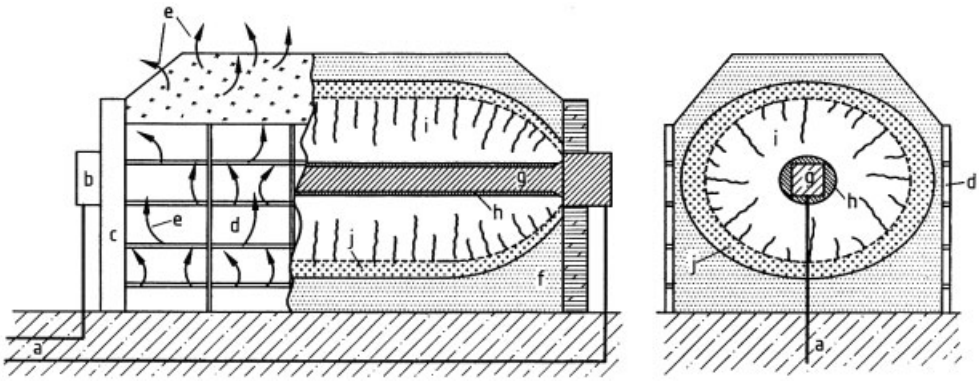

Figure 3. Acheson furnace
a) Electrical supply; b) Electrode; c) Stationary end wall; d) Movable side wall with openings for furnace gases; e) Furnace gases; f) Reaction mixture; g) Heating resistor (core); h) Graphite from decomposition of SiC; i) Silicon carbide: coarsely crystalline inner zone; j) Silicone carbide: finely crystalline outer zone ("amorphous")

of the CO and sulfur at the top of the furnace, and other odorous components, escape through the openings in the roof. The walls of the building are provided with openings near ground level through which fresh air enters. Generally, the natural convection caused by the furnaces adequately ventilates the building, so that additional ventilation equipment is unnecessary. Most Acheson plants are not provided with waste-gas purification equipment.

A comparative review of some Acheson plants is given in [30].

3.4. The ESK Process

A major disadvantage of the Acheson process is that the byproduct CO cannot be recovered and used. Moreover, the SO_2 and dust in the waste gases cause pollution.

These disadvantages led to the development of the ESK process [31–36], which also uses a resistance-heated furnace, although the only design feature it has in common with the Acheson furnace is the horizontal heating core. The ESK furnace (Fig. 4) has no walls, and the electrodes are located in the floor. A graphite column is placed upon each electrode, and the coke–sand mixture is charged by wheel loaders until it reaches the top of the graphite columns. The heating core, which can be linear or U-shaped, is then put in position and connected to the graphite columns. The heating core is then covered with the rest of the mixture, forming a mound ca. 6 m in height. The mound is covered with a plastic sheet which traps the reaction gases evolved during operation of the furnace, causing the sheet to inflate and protecting it from the heat. The rate of gas removal is adjusted so that the pressure under the sheet remains constant.

This technique enables the byproduct CO to be collected, purified, and used. The most modern ESK furnace plant to date is in Delfzijl, in the Netherlands. It includes a desulfurization plant from which elemental sulfur can be obtained, and also includes a power station which recovers some of the energy used in the SiC production.

After the power to the furnace is switched off and the sheet pulled off, the excess reaction mixture is removed by heavy equipment. Dust formation is reduced by damping down with water.

ESK furnaces are considerably larger than Acheson furnaces, the transformers having an installed power of up to 15 MVA. The heating core can be 60 m in length when the U-shaped configuration is used.

3.5. Processing of the Crude Product

The lumps of silicon carbide obtained on dismantling the furnace are taken to a sorting area and broken down with hydraulic equipment and compressed air hammers. The finely crystalline outer zone (the so-called amorphous zone) is further size reduced (e.g., in a jaw crusher) and is

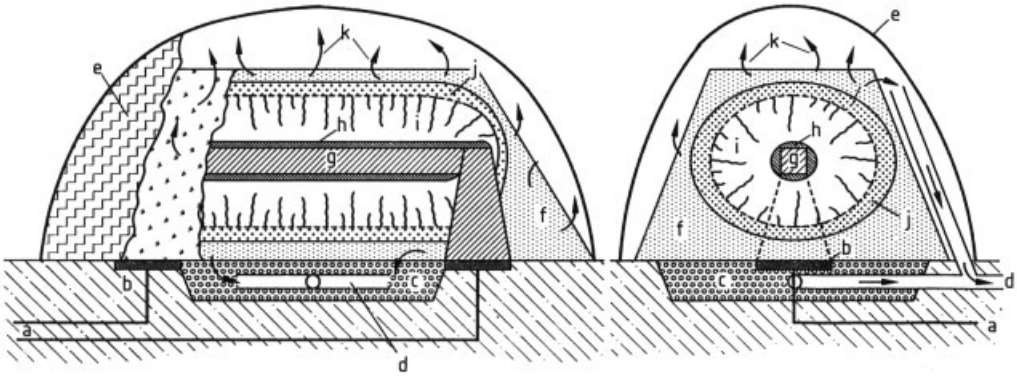

Figure 4. ESK furnace
a) Electrical supply; b) Floor electrode; c) Gas-permeable furnace bed; d) Gas collection; e) Plastic sheet; f) Reaction mixture; g) Heating resistor (core); h) Graphite from decomposition of SiC; i) Silicon carbide: coarsely crystalline inner zone; j) Silicone carbide: finely crystalline outer zone ("amorphous"); k) Furnace gases

usually returned to the furnace process by adding it to the coke – sand mixture. Alternatively, this 90 – 95 % SiC material is a valuable raw material for the production of grain refiners for the ferrous metal industry (see Section 6.3).

The coarsely crystalline, compact pieces from the inner zone of the SiC roll are sorted according to criteria such as color, structure, etc. This top-grade material forms the largest fraction of the SiC yield from a furnace charge.

The large pieces of silicon carbide are size reduced in stages using conventional size-reduction equipment such as jaw crushers, gyratory crushers, hammer mills, roll mills, ball mills, and jet mills. The choice of equipment depends on the desired size and shape of the SiC grains.

Due to the hardness of silicon carbide, some contamination of the product by abrasion of the grinding equipment cannot be avoided, whatever material of construction is used. It is therefore best to use steel, so that the abrasion products can easily be removed by magnetic separators.

For applications that require a particularly high grade of silicon carbide, such as industrial ceramics, the refractory industry, and ceramic-bound abrasives, the granular product must be washed with alkali or acid to remove adhering traces of elemental silicon, metals, metal compounds, graphite, dust, and silica.

Sieving, classification, and sedimentation processes are used to prepare various fractions with very narrow and precise grain size distributions to meet international standards for abrasive grits.

The same technique is used to obtain products for ceramics and refractory materials. These generally have a broader particle size distribution than that required for abrasives, which, however, is specified for each particular application.

Particle sizes range from ca. 5 mm down to the submicron range.

3.6. Other Production Methods

Attempts have been made to produce silicon carbide continuously from coke and sand. However, due to the relatively short reaction times that can be achieved, only agglomerates of very fine β-SiC crystals can be produced, and these are unsuitable for most applications of silicon carbide.

The manufacture of sintered SiC articles requires ultrafine SiC powder, which is generally produced by grinding commercial α-SiC in ball mills, colloid mills, or attritors to give the desired particle size of < 1 µm. An important new process is the direct manufacture of sinterable β-SiC powders from the gas phase by decomposing gaseous silicon halides in the presence of hydrocarbons in a plasma jet reactor [37]. In another CVD process, the starting materials are organosilicon compounds [38]. Nucleation must take place in the gas phase for powders to be formed.

Another method of producing a sinterable β-SiC powder is the carbothermic reduction of finely divided SiO_2 [39].

The thermal decomposition of rice husks gives a very finely divided form of β-SiC [40]. The yield is ca. 10 % based on the total raw material used. Some of the product has the form of whiskers.

The formation of silicon carbide from the elements is used in the manufacture of ceramic components (see Section 6.6). In one process, carbon or carbon-bonded silicon carbide moldings are infiltrated with molten silicon and converted into SiC (SiSiC). In another process, the batch contains elemental carbon and silicon, which are converted to SiC by subsequent heat treatment. The product is known as reaction-bonded silicon carbide (RBSiC).

Carbosilane resins have been proposed as precursors which yield SiC on pyrolysis. The high cost of this process, however, prevents its application on a wider scale.

Fibrous SiC (whiskers) is obtained in various CVD processes in which SiC is precipitated in the presence of catalysts. The preferential one-dimensional growth of these crystals proceeds by a vapor – liquid – solid (VLS) mechanism [3].

4. Quality Control

The SiC content of silicon carbide products is now usually determined by measuring the carbon content. The total carbon content is determined by combustion of the sample in a stream of oxygen at 1050 °C in the presence of lead borate. The CO_2 produced is absorbed in $Ba(ClO_4)_2$ solution and determined by coulometry [41,42]. An alternative technique is to oxidize the SiC with oxygen in a high-frequency induction furnace containing a flux metal, and to detect the CO_2 produced by IR absorption.

Free carbon can be directly determined coulometrically by combustion at 850 °C without any additives, provided that the SiC itself is not appreciably oxidized. If the SiC samples are very finely divided, i.e., have large specific surfaces, and if long heating times are required due to high free carbon contents, this method cannot be used. In such cases, the weight change on combustion is measured, and the decrease in total carbon content is determined. This enables the effect of the unavoidable oxidation of the SiC to be eliminated, and the free carbon content to be calculated [41,42].

Analyzers with IR detection are also suitable for the direct determination of free carbon provided that they allow precise temperature control during combustion of the free carbon.

The SiC content is calculated from the difference between total and free carbon.

The free silicon content is determined by measuring the volume of hydrogen produced on treatment with sodium hydroxide solution. Alternatively, metallic silver is precipitated from a silver fluoride solution, dissolved in nitric acid, and determined by a standard method [41,42]. In the chemical analysis of SiC abrasive grits, emphasis is placed on the determination of accompanying materials such as carbon, silicon, silicon dioxide, and metallic oxides. The SiC content (plus the free carbon) is found by weighing the residue after treatment with a mixture of hydrofluoric, nitric, and sulfuric acids [42]. For special grades of SiC for refractory and metallurgical applications, volatile components can interfere with the SiC determination. In this case, the sample must first be annealed under argon [43].

Besides chemical analysis, physical properties such as particle size, particle size distribution, and bulk density are also important.

The particle size distribution of coarse abrasive materials (5 mm – 50 µm) is determined by sieving. For particle sizes less than 50 µm, sedimentation and laser diffraction methods are used.

Both the particle size distributions and the methods for their determination are standardized. The standards for bonded abrasive applications [44] differ from those for coated abrasives [45]. The bulk density depends both on the particle size distribution and on the particle shape. It is measured by weighing a known volume of SiC grains [46]. The packed density is often determined instead of the bulk density [47].

5. Toxicology and Occupational Health

Silicon carbide is nontoxic, and is therefore a nonhazardous material as defined by GefStoffV.

For fiber-free, finely divided SiC, the MAK value is 4 mg/m^3 [48]. If the MAK is exceeded, a dust mask with a P 1 filter (for inert dusts) must be worn. However, since these SiC dusts are not fibrogenic [49], so that there is no MAK value.

Conclusive experimental studies on the effects of respirable fibrous SiC dusts (whiskers) are not yet available [50].

Silicon carbide can be disposed of together with ordinary domestic refuse provided that local regulations are observed. There is no legal requirement for documentation.

6. Uses

6.1. Abrasives (see also → Abrasives)

The extreme hardness of silicon carbide led to its use as an abrasive immediately after its discovery. Silicon carbide is used directly as a loose-grain abrasive for cutting with wiresaws and for lapping. It is also used in bonded form. The latter use can be divided into two groups:

1) Bonded abrasives include grinding wheels, whetstones, hones, abrasive cutting-off wheels, and monofiles. The SiC grains can be bonded with plastics or ceramics.
2) Coated abrasives include abrasive paper and cloth in sheet or band form. They are produced by strewing the SiC grains onto a substrate coated with glue or bonding resin and then covering with a second layer of bonding agent [51].

6.2. Refractories

High resistance to oxidation, high thermal conductivity, and low coefficient of thermal expanison make silicon carbide ideal for use in refractory raw materials and finished products. Its hardness and abrasion resistance are additional advantages, so that the range of applications in the field of refractory materials is wide [13]. The following examples are therefore by no means complete.

The commonest products are clay-bonded moldings and monolithics with SiC contents of 40 – 90 %. These are used for blast furnace linings and runners; muffle plates and chambers in industrial furnaces; column trays for zinc distillation; boiler furnace linings in thermal power stations; incinerators for domestic refuse; and for kiln furniture used in the manufacture of porcelain and earthenware products.

Quality standards continually increase (especially for kiln furniture), with regard to durability, strength, and weight saving, so that classical clay bonding is being increasingly replaced by the use of synthetic oxides or Si_2ON_2, Si_3N_4, and SiC.

In the case of nitride or oxynitride bonding, the starting material is a mixture of SiC with a few percent elemental silicon. This is shaped by pressing or casting, and then calcined in the presence of precisely controlled amounts of nitrogen or air.

Kiln furniture made of recrystallized SiC (RSiC) is manufactured from a mixture of 70 % silicon carbide with a grain size of ca. 100 µm (the structural particles) and 30 % of a very fine grade of SiC having a particle size distribution ranging from 10 µm to the submicron range (the bonding particles). The mixture is formed by slip casting. The green castings are fired at 2100 – 2500 °C under a protective gas (e.g., argon). No shrinkage takes place. The coarse particles of SiC grow at the expense of the small particles. This takes place mainly at the contact angles between the structural particles, resulting in high strengths, which are maintained even at working temperature (ca. 1500 °C in air) [52]. The finished RSiC articles, which are used as supporting plates, saggars, and transporting rolls for fast-firing kilns, and as supporting beams for firing suspended articles, have a wall thickness ca. 50 % lower than those made of oxide-bonded silicon carbide [53].

A similar development has taken place in the field of silicon carbide blast furnace linings. Here, silicon carbide bricks are bonded with Si_3N_4 and SiC, and sometimes other nitrogen-containing bonding agents such as Si_3N_4/sialon (→ High-Temperature Materials, Chap. 9.3.8.) and Si_2ON_2. Pure SiC bricks are also used. Bricks bonded with α-SiC (firing temperature > 2000 °C) have improved durability compared with reaction-bonded SiC (RBSiC), in which the bonding phase is β-SiC due to the lower firing temperature of 1400 – 1500 °C. When selecting a construction material, the best compromise

must be found with respect to its resistance to alkali attack, oxidation, thermal shock, and abrasion [54, 55].

Silicon carbide is also used in combination with carbon or graphite for refractory applications, e.g., in the graphite- clay bricks used in nonferrous metallurgy. They contain up to 35 % SiC to improve the refractoriness under load.

The use of SiC in the manufacture of filters for casting metals has increased. These filters are fitted to the mold ingate and give a considerable improvement in casting quality. Filters made of SiC are used in gray iron foundries [56]. To produce the filter body, a molding made of open-cell polyurethane foam is saturated with a slurry of SiC, Al_2O_3, and a high-temperature binder such as aluminum dihydrogenphosphate. After drying, the plastic is burnt off, and the ceramic is fired [57].

Structural ceramic materials that are also refractory components include heat exchangers, burner nozzles, etc., made of SiC. The most commonly used varieties of this material are recrystallized silicon carbide (RSiC) and silicon-infiltrated silicon carbide (SiSiC).

6.3. Metallurgy

The iron and steel industries consume large quantities of silicon carbide for deoxidizing high-alloy steels and for adding silicon and carbon to cast iron [58]. The SiC used is of the quality found in the outer zone of the SiC roll (see Section 3.3), as this material dissolves more readily in the molten ferrous metal owing to its fine crystalline structure. In cast iron production, the SiC is added to the feed of the cupola in the form of cement-bonded briquettes. Silicon carbide is used in granular form for deoxidizing steel in the ladle [59].

6.4. Electrical Technology

The semiconducting properties of silicon carbide have long been utilized in resistance heating elements. Metallic heating elements cannot be used in laboratory and industrial furnaces that operate at up to 1500 °C. Instead, elements made of pure SiC in the form of rods, tubes, or helical shapes are used. These are produced from SiC grains bonded by silicon infiltration or by the reaction bonding method (see Section 3.6).

Voltage-dependent resistors (varistors) consist of SiC particles bonded with ceramic or polymer binders (see Section 3.6). They are produced in the form of discs or blocks, or in cuboid or ring shapes, and are the active elements in overvoltage protection equipment such as lightning protection devices and surge arresters. Not only the electrical properties of the silicon carbide are exploited, but also its resistance to chemical attack, which gives the SiC conductors outstanding stability. Silicon carbide allows arresters to be produced that combine low protection levels with good stability to prolonged overvoltage, such as can occur in earth faults in electrical supply systems with an isolated or resonant earthed neutral [60].

High-voltage cables and generators are protected from local discharges by using SiC powder as a filler in the insulating coating (corona shielding).

When used in floor coverings, silicon carbide gives antislip and antiabrasion properties, and eliminates static electricity.

6.5. Electronics

Silicon carbide is an outstanding material for the construction of electronic equipment such as rectifying diodes for high temperatures and/or high energy densities. Blue light-emitting diodes can be produced from SiC single crystals (→ Electroluminescent Materials and Devices, Chap. 4.2.) [61]. However, it is not yet possible to mass produce these economically.

Hydrogenated amorphous SiC compounds of the type $Si_{1-x}C_x$: H, which can be produced as thin films by glow discharge deposition onto silicon or other substrates from mixtures of SiH_4, C_2H_4, and H_2, seem to have a range of potential applications, e.g., in rectifiers [62], light-dependent resistors [63], electrophotographic receptors [64], and light-emitting diodes [65].

6.6. Structural Ceramics

Pumps used in the chemical industry for feeding and metering strongly corrosive and abrasive materials are provided with slide-ring seals

made of sintered SiC. For similar requirements involving radial loads, sliding bearings and shaft housings are manufactured from sintered silicon carbide. The same material is used to produce ball valves and other parts for metering pumps and fittings [66]. The starting material for the manufacture of these structural ceramic components is submicron SiC powder (see Section 3.6).

For the production of sintered silicon carbide moldings, a wide range of sophisticated production steps are available, depending on the properties required of the end product. The SiC powder is first mixed with sintering aids and temporary binders, and is conditioned preparatory to the forming stage. The latter is carried out by slip casting, spray casting, or compression molding. A complex heat treatment process is often included at this stage to remove the intermediate binder before the workpiece is sintered. This can be pressureless (SSiC), under unilateral pressure (hot pressing, HPSiC), or under isostatic pressure (hot isostatic pressing, HIPSiC) [67].

7. Economic Aspects

In 1995, the world capacity for silicon carbide manufacture was ca. 750 000 t/a. Of this, 165 000 t/a was from Western Europe, 145 000 t/a from Eastern Europe, 178 000 t/a from North and South America, and 266 000 t/a from the Far East.

8. References

1. A. L. Lea, *Trans. Br. Ceram. Soc.* **40** (1941) 93 – 118.
2. P. Schützenberger, *Comp. Rend.* **92** (1881) 508; **114** (1892) 1089 – 1093.
3. *Gmelin,* **Si Suppl. vol. B 3,** Silicon Carbide part 2.
4. H. Kleykamp, G. Schumacher, *Ber. Bunsenges. Phys. Chem.* **97** (1993) no. 6, 799.
5. L. S. Ramdsdell, *Am. Mineral.* **32** (1947) 64 – 82.
6. I. Patzak, K. Wohlleben, K. Konopicky, *Glas Email Keramo Tech.* **22** (1971) 2 – 9.
7. R. Kiefer, E. Gugel, P. Ettmayer, A. Schmidt, *Ber. Dtsch. Keram. Ges.* **43** (1966) 621 – 623.
8. P. Lundquist, *Acta Chem. Scand.* **2** (1948) 177 – 191.
9. *Ullmann,* 4th ed., **21,** 431 – 438.
10. P. T. B. Shaffer, *J. Am. Chem. Soc.* **47** (1964) 466.
11. E. Gugel, P. Schuster, G. Senftleben, *Stahl Eisen* **92** (1972) 144 – 149.
12. JANAF Thermochemical Tables, 2nd ed., NSRDS-NBS37, Washington D.C., June 1971.
13. P. Wecht: "Feuerfest-Siliciumcarbid", *Applied Mineralogy,* vol. 11, Springer Verlag, Wien, New York 1977.
14. *Gmelin,* 8th ed., **15** Silicium, part B.
15. K. Zückler in W. Schottky (ed.): *Halbleiterprobleme* III (1956), pp. 207 – 229.
16. S. H. Hagen, *Ber. Dtsch. Keram. Ges.* **47** (1970) 630 –634.
17. B. Frisch, W. R. Thiele, R. Drumm, B. Münnich, *Ber. Dtsch. Keram. Ges.* **65** (1988) 277 – 284.
18. G. Wiebke, *Ber. Dtsch. Keram. Ges.* **37** (1960) 219 –224.
19. J. Schlichting, *Ber. Dtsch. Keram. Ges.* **56** (1979) 196 –200.
20. K. Konopicky, I. Patzak, H. Dohr, *Glas Email Keramo. Tech.* **23** (1972) 81 – 87.
21. P. T. B. Shaffer, *Ceram. Age* **82** (1966) 42 – 44.
22. T. Iseki, T. Kameda, T. Maruyama, *J. Mater. Sci.* **19** (1984) 1692 – 1698.
23. H. von Zeppelin, *Schweiz. Arch.* **34** (1968) 19 – 29.
24. H. Fuchs, *Chem. Ing. Tech.* **46** (1974) 139 – 142.
25. *Winnacker-Küchler,* 4th ed., **2,** 627.
26. K. H. Mehrwald, *Ber. Dtsch. Keram. Ges.* **44** (1967) 148.
27. E. G. Acheson, DE 76 629, 1892.
28. E. G. Acheson, DE 85 197, 1894.
29. Lonza AG, CH 595 283, 1978 (G. Ferrazzi, G. Nizzola, S. Lanini, O. Troendle).
30. K. H. Mehrwald, *Ber. Dtsch. Keram. Ges.* **69** (1992) 72 – 81.
31. Elektroschmelzwerk Kempten GmbH, DE 2 364 106, 1976 (G. Wiebke, L. Korndörfer, E. Korndörfer).
32. Elektroschmelzwerk Kempten GmbH, DE 2 364 107. 1976 (A. Korsten et al.).
33. Elektroschmelzwerk Kempten GmbH, DE 2 364 108, 1976 (G. Wiebke, A. Korsten, T. Benecke, F. Petersen).
34. Elektroschmelzwerk Kempten GmbH, DE 2 364 109, 1975 (G. Wiebke, A. Korsten, T. Benecke, F. Petersen).
35. Elektroschmelzwerk Kempten GmbH, DE 2 421 818, 1977 (G. Wiebke, A. Korsten, T. Benecke, F. Petersen).
36. Elektroschmelzwerk Kempten GmbH, DE 2 630 198, 1978 (F. Petersen, A. Korsten).

37. General Electric Co, DE 2 518 950, 1975 (S. Prochazka).
38. W. Böcker, H. Hausner, *Ber. Dtsch. Keram. Ges.* **55** (1978) 233 – 237.
39. K. A. Schwetz, A. Lipp, *Radex Rundsch.* 1978, 489 –498.
40. J. G. Lee, J. B. Cutler, *Am. Ceram. Soc. Bull.* **54** (1975) 195 – 198.
41. DIN 51 075 part 1 – 5.
42. ISO 9286.
43. DIN 51 076,part 1.
44. DIN 69 101,part 1 – 3.
45. DIN 69 176,part 1 – 3.
46. FEPA 44-X-1986.
47. ISO 787,part 11.
48. D. Henschler (ed.), *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* VCH Verlagsgesellschaft, Weinheim 1987.
49. J. Bruch, Institut für Hygiene und Arbeitsmedizin, Universitätsklinikum Essen, Publication in print.
50. I. D. Birchall et al., *J. Mater. Sci. Lett.* **7** (1988) 350 –352.
51. L. Coes, Jr.: “Abrasives”, *Applied Mineralogy,* vol. 1, Springer Verlag, Wien 1971.
52. J. Kriegesmann, *Keram. Z.* **42** (1990) 481 – 484.
53. W. Völker, A. Fickel, *Aachen Proceedings* (1989) 69 –71 (special issue of Inter Ceram).
54. A. Fickel, W. Völker, *Aachen Proceedings* (1988) 38 –40 (special issue of Inter Ceram).
55. R. W. Brown et al., *Aachen Proceedings* (1988) 41 – 46 (special issue of Inter Ceram).
56. Georg Fischer AG, EP 126 847, 1984 (R. Rietzscher, J. Otto, G. Trinkl, W. Kaettlitz).
57. Georg Fischer AG, WO 82/03 339, 1982 (F. Hofmann et al.).
58. Elektroschmelzwerk Kempten GmbH, Th. Benecke, Der Einsatz von metallurgischem Siliciumcarbid, company brochure, München 1978.
59. E. A. Loria, A. P. Thompson, *Trans. Am. Soc. Met.* **39** (1947) 651 – 669.
60. Elektroschmelzwerk Kempten GmbH, Siliciumcarbid für spannungsabhängige Widerstände, company brochure, München 1990.
61. W. von Münch, W. Kürzinger, I. Pfaffeneder in *Solid State Electronics,* **vol. 19,** Pergamon Press, Oxford 1976, pp. 871 – 874.
62. M. M. Rahman, S. Furukawa, *Jpn. J. Appl. Phys.* **23** (1984) 524 – 525.
63. Matsushita Electric Industrial Co, JP 80 127 080, 1980; *Chem. Abstr.* **94** (1981) 56964.
64. Matsushita Electric Industrial Co, JP 81 14 241, 1981; *Chem. Abstr.* **95** (1981) 52673.
65. H. Kukimoto, *Semicond. Semimetals* **D 21** (1984) 239 – 48; *Chem. Abstr.* **102** (1985) 175 100.
66. H. Knoch, J. Kracker, *Ber. Dtsch. Keram. Ges.* **64** (1987) 159 – 163.
67. K. A. Schwetz, *Radex Rundsch.* 1989, 26 – 39.

Silicon Compounds, Inorganic

Walter Simmler, Bayer AG , Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Silicon Hydrides 2
2.1. Monosilane 2
2.2. Disilane 3
2.3. Purification 4
2.4. Uses 4
2.5. Safety Measures 4
2.6. Photochemistry 4
3. Silicon Halides 5
3.1. Trichlorosilane 5
3.1.1. Preparation from Elemental Silicon . 5
3.1.2. Preparation from Tetrachlorosilane . . 6
3.1.3. Purification 7
3.1.4. Waste Disposal 7
3.1.5. Uses 8
3.2. Tetrachlorosilane 8
3.2.1. Production 8
3.2.2. Purification 10
3.2.3. Uses 10
3.3. Hexachlorodisilane 10
3.3.1. Production 10
3.3.2. Uses 11
3.4. Other Silicon Chlorides 11
3.5. Silicon Fluorides 11
3.6. Silicon Bromides 12
3.7. Silicon Iodides 13
4. Silicon Oxides 13
4.1. Silicon Monoxide 13
4.1.1. Production 13
4.1.2. Properties 14
4.1.3. Uses 14
4.2. Silicon Sesquioxides 15
4.3. Other Silicon – Oxygen Compounds 15
5. Silicon Sulfides 16
6. Silicon Nitrogen Compounds 17
7. Silicon Borides 17
8. References 17

1. Introduction

A number of inorganic compounds of silicon have gone through a remarkable development in the past years, prompted chiefly by the electronics industry and efforts of environmental protection. A tremendous growth of knowledge in physics and chemistry, as well as technical applications, has taken place around certain hydrides, halides, and oxides of silicon. Progress in solid state physics, the material sciences, and an increase in unwanted byproduct formation in the course of production of large-scale silicon compounds have necessitated that one look anew at process engineering. Energy consideration is of direct bearing in this technology since most processes in technical silicon chemistry require high temperature.

Environmental protection and elimination of safety hazards during production and handling of the above silicon products has been the concern of all ever since these compounds became large-scale technical products. The moisture sensitivity of the halides, the oxygen sensitivity of the hydrides, and pyrophoric properties of some of the distillation residues from reaction steps involving elementary silicon or silicides have caused waste problems and particular emphasis on process safety. Greater flexibility of chemical transformation within at least partly recycling procedures have both decreased waste and enhanced safety. Making use of byproducts has been extended to comprise substances of the system C – Si – O – Hal, its phases and transformations. Alongside product making, byproduct formation, its investigation, suppression, or transformation into useful silicon derivatives has been to some degree due to efforts of environment protection.

With respect to use, many new applications were found for special inorganic silicon compounds, optical fibers and solar cells being example for both surface as well as bulk properties.

Surface treatments, not necessarily confined to protective coatings but including migration beneath the surface, are carried out to increase hardness, thermal stability, to influence cohesion and adhesion, to resist oxidation, erosion, corrosion, and to enhance reflexion, etc. Interfaces and multilayer cermets and plastics need silicon in various bonding states as intermediates during production or permanently in a final product.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_001

2. Silicon Hydrides

Silicon hydrides are sensitive to hydrolysis and oxidation; volatile higher silanes are spontaneously flammable [1–3]. The first five members of the homologous series Si_nH_{2n+2} are known in the pure state (see Table 1).

All silanes are decomposed in reactions of the type:

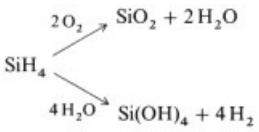

The thermal decomposition of monosilane takes place as follows:

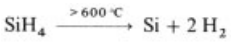

Higher silanes decompose at lower temperature.

High-purity silicon can be produced in this way from pure silanes. Organosilicon compounds can be obtained by addition of silanes to unsaturated organic compounds.

2.1. Monosilane

Monosilane, SiH_4, is formed by the hydrolysis of magnesium silicide (Mg_2Si) with 10 % hydrochloric acid at 50 °C. It is, however, accompanied by higher homologues which must be removed by high-vacuum fractionation [4,5]. However, if magnesium silicide is hydrolyzed with ammonium chloride in anhydrous ammonia under its own vapor pressure at room temperature, SiH_4 is obtained exclusively [6]. Ammonium bromide has also been used [7]. The ammonia is condensed out as far as possible, and any remaining traces are removed by washing the silane with dilute nitric acid.

Silane can readily be prepared from silicon halides and metal hydrides. The most convenient laboratory method is the reaction of $SiCl_4$ with lithium aluminum hydride in diethyl ether [8] or THF [9].

$$SiCl_4 + LiAlH_4 \longrightarrow SiH_4 + LiCl + AlCl_3$$

The hydrogenation can also be performed with alkali metal hydride/aluminum alkoxide or alkali metal hydride/boron triethyl in paraffin oil [10], or by alkali metal aluminum hydride/calcium aluminum hydride in other organic solvents [11,12].

The alkali or alkaline-earth metal hydrides can be formed in situ by passing silicon tetrachloride and hydrogen into an electrolyzed melt of the metal halide [13].

In a two-step process, in which the rate of silane formation is adjusted by means of the temperature difference between the two stages, silicon halides other than the tetrafluoride are first reacted with excess alkali metal hydride or alkali metal aluminum hydride, followed by degassing and then contacting the solid reaction product with excess silicon halide [14]. Monosilane is obtained in low yield but with sufficient purity to be used in the manufacture of semiconductors with toluene as a solvent instead of ethers [15].

Benzyltriethylammonium chloride or other quaternary ammonium salts are used as catalysts in concentrations up to 25 mol %. High yields of monosilane are obtained from the reaction of silicon halides with a mixture of aluminum products containing dialkyl aluminum hydrides, trialkylaluminum, ethylaluminum chloride, and aluminum powder in liquid paraffin at 40 °C.

Catalytic disproportionation of chlorosilane mixtures originating from various sources in the silicon industry provides a route to monosilane for the manufacture of polycrystalline solar silicon. For a mixture of trichloro- and dichlorosilane, copper [16], a mixture of silica and alumina at 300 °C [17], and quaternary ammonium groups bonded to carbon in an ion-exchange resin have been reported as catalysts for the main reactionleading to monosilane [18].

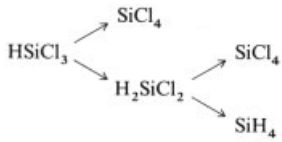

The catalytic conversion of intermediate dichlorosilane can also be accomplished by alkyl or aryl nitriles on carbon carriers at 50 °C [19]. Boron catalysts, however, enhance the formation of dichlorosilane in chlorosilane mixtures obtained by high-pressure plasma hydrogenation of tetrachlorosilane [20]:

Table 1. Silanes Si_nH_{2n+2}

	M_r	*mp*, °C	*bp*, °C	d_4^{20}
SiH_4	32.12	−185	−112	0.68
Si_2H_6	62.23	−130	− 14.5	0.69
Si_3H_8	92.33	−117	− 53	0.72
Si_4H_{10}	122.44	− 91	−108	0.79

$$SiCl_4 + H_2 \longrightarrow HSiCl_3 + H_2SiCl_2$$

In-process purification is achieved by contacting part of the monosilane produced with the starting mixture of chlorosilanes, which contains the impurities as chlorides, chiefly those of arsenic, phosphorus, and boron. This procedure converts the contaminants into their more volatile hydrides, which are removed with the off-gas [18].

Properties and physical data of monosilane, particularly with regard to obtaining high-purity silicon by chemical vapor deposition, are compiled in [21].

2.2. Disilane

Increasing amounts of hexachlorodisilane are being produced as a byproduct of the chlorosilane industry. Since it is of no immediate use, it is subjected to hydrogenation or disproportionation to convert it to useful derivatives, mainly tetra- or trichlorosilane. When the production of amorphous silicon from disilane became commercially viable, the hydrogenation of hexachlorosilane became a feasible alternative to disproportionation. The production of disilane resembles that of monosilane, the reducing agents being metal hydrides, mainly lithium aluminum hydride. The use of di-*n*-butyl ether as a solvent is reported to give a smooth reaction at 35 °C [22]. Addition of lithium hydride increases the yield molar ratios $LiH : LiAlH_4$ of 0.8, 2.0, 12.0, 20.0, and 40.0 resulted in yields of 83, 88, 90, 89, and 70 %, respectively. Lithium hydride can be used as the sole reducing agent if aluminum bromide or chloride is added to the ether solvent [23]. Lithium aluminum hydride in refluxing diethyl or diisopropyl ether gave disilane in 88 % yield [24].

2.3. Purification

Removal of boron compounds from silanes that are to be used for the production of high-purity silicon is very important. When SiH_4 is produced from alkali metal hydrides [12], it is recommended that water should be added to hydrolyze boron hydrides under the alkaline reaction conditions. Water vapor is added during distillation for the same reason [25]. Boron compounds can be removed from SiH_4 by adsorption from the gas phase onto activated carbon or silica carriers loaded with compounds that form complexes with boron compounds ($AlCl_3$, $FeCl_3$, C_2H_5CN, NH_2CN, PCl_3) [26].

If alkali metal hydrides MH are reacted in organic solvents, it is often unnecessary to include a special stage for removing boron compounds from the silicon compounds, as the former react to form boranate complexes under the following conditions:

$$SiCl_4 + 2\,BCl_3 + 10\,MH \longrightarrow 10\,MCl + SiH_4 + B_2H_6$$

$$B_2H_6 + 2\,MH \longrightarrow 2\,MBH_4$$

A systematic study has been carried out on the liquid – vapor distribution coefficients of volatile trace impurities in the hydrides of boron, silicon, phosphorus, arsenic, sulfur, and selenium. Experimental and calculated data are reported for the temperature range between the boiling point and critical temperature. Linear regression coefficients for the temperature dependence of the impurity distribution coefficients are included [27].

2.4. Uses

Direct use of silicon hydrides involves their decomposition on surfaces, leading to deposition of elemental silicon or silicon compounds.

Polycrystalline films of elemental silicon of controllable grain size are deposited on glass or silica-coated surfaces by heating them to 450–800 °C in an atmosphere consisting of monosilane and tetrabromo- or tetraiodosilane [28]. Chemical vapor deposition is generally followed by annealing and recrystallization with an energy beam. This is used to deposit silicon on insulator surfaces [29]. Amorphous silicon photoconducting layers on electrophotographic plates are made by using a plasma for silane decomposition and silicon deposition [30]. Liquid crystal displays, based on thin film transistors, for screens in monitoring devices, are made from amorphous silicon which is produced on alkali-free glass surfaces in situ by an electric discharge in an atmosphere of monosilane at reduced pressure. Residual bonds are saturated by hydrogen atoms to avoid diminishing the semiconductor effect [31]. On the surfaces of transition metals, cleaned by sputtering in an argon atmosphere, monosilane and the higher silicon hydrides can be thermally decomposed at ca. 300 °C to form metal silicides [32]. Exposure to air leads to the formation of silicon oxides.

2.5. Safety Measures

Silicon hydrides are combustible; the lower members are spontaneously flammable. Particular attention has been given to the system monosilane – oxygen and its explosibility [33]. Oxygen-rich mixtures (>70 % O_2) give water and silicon dioxide as combustion products. As the mixtures become richer in monosilane, hydrogen replaces water as a combustion product. For monosilane-rich mixtures (>70 % SiH_4) the products are hydrogen, silicon, and silicon oxides. The explosion of monosilane-rich mixtures is primarily a thermal explosion of monosilane itself.

Hazardous waste gases containing silicon hydrides and halides are passed through columns containing pelletized oxidants to remove the silicon compounds. The oxidants are made by treating a mixture of copper and zinc nitrate with sodium carbonate [34].

2.6. Photochemistry

Irradiation of silanes and silyl compounds leads to fragments, namely the silylene (: SiH_2) intermediate. Silane mixtures with boron trichloride [35] and hydrogen chloride [36] have been investigated with infrared laser techniques.

Monosilane and boron trichloride react along two pathways, one being hydrogenation of boron compounds, leading to boron dichloride, diborane, and trichlorosilane. In a side reaction, decomposition of silane takes place, forming hydrogen and amorphous silicon. Carrying out this procedure with methane instead of silane leads to a pyrolysis of methane, the released hydrogen partly converting the boron trichloride to hydrides.

In the presence of hydrogen chloride, the primary reaction seems to be the formation of intermediate silylene which in turn reacts with silane and hydrogen chloride. Simultaneous formation of chlorosilanes also occurs. The activation energy for the insertion of silylene into hydrogen chloride was determined to be <1.3 kcal/mol. In general, silylene insertions are considered in the context of a triplet state, and symmetry allows insertions to proceed without excitation energy when the antibonding orbital, relating to the bond into which the insertion occurs, is symmetrical with respect to the reaction coordinate [37].

3. Silicon Halides

Silicon halides (see Table 2) are mostly colorless liquids that fume in air, are readily hydrolyzed, and consequently have an irritating effect on the mucous membranes. Silicon hydridohalides with a high hydrogen content are spontaneously flammable. All 18 mixed silicon halides with the composition $SiHal^1{}_nHal^2{}_m$ $(n+m=4)$ are known.

3.1. Trichlorosilane

Trichlorosilane, $HSiCl_3$, M_r 135.4, *mp* −134 °C, *bp* 36.5 °C (Table 2) is the most important silicon hydridohalide. The vapor is highly flammable; mixtures with oxygen or air explode violently on ignition or on contact with a hot surface. It is decomposed by water with the evolution of hydrogen.

Table 2. Melting and boiling points of halogenosilanes

	M_r	*mp*, °C	*bp*, °C
SiH_3F	69.12		−98.6
SiH_2F_2	106.11	−119	−77.5
$SiHF_3$	143.10	−131.5	−97.5
SiF_4	180.09	− 90.2 [a]	−98.6
SiH_3Cl	66.57	−118	−30.5
SiH_2Cl_2	101.01	−122	− 8.5
$SiHCl_3$	135.45	−134	36.5
$SiCl_4$	169.89	− 70	57.6
SiH_3Br	111.04	− 94	2.0
SiH_2Br_2	189.95	− 70	66
$SiHBr_3$	268.86	− 73.5	112
$SiBr_4$	347.77	5	153
SiH_3I	158.02	− 56.5	46
SiH_2I_2	283.91	− 1	149.5
$SiHI_3$	409.80	8	220
SiI_4	535.69	121	290

[a] At 1753 mbar triple point.

3.1.1. Preparation from Elemental Silicon

Trichlorosilane was first prepared by F. WÖHLER, who used the same method as that now used for laboratory and industrial scale production [38]:

$$Si + n\,HCl \longrightarrow HSiCl_3 + SiCl_4$$

A mixture of chlorinated silanes is produced when a mixture of hydrogen chloride and hydrogen at ca. 250 – 400 °C is passed over high-silicon (ca. 97 % Si) ferrosilicon or silicon mixed with aluminum, nickel, or copper(I) chloride [39]. The process is sometimes carried out in a turbulent [40–42] or nonturbulent fluidized bed [43, 44].

Due to the growing demand for pure precursors for high-purity silicon, process improvements have been made continuously to raise the yield of trichlorosilane and to lower energy consumption as the chief cost cutting measures in production. The fluidized-bed reactor [45–54] has become the predominant process base for the manufacture of trichlorosilane from metallurgical-grade silicon or other silicon solids. Additional feed gases, both inert (e.g., nitrogen, argon, and helium) or reactive (e.g., chlorine) have been used [45, 46]; feeding hydrogen chloride in pulses is reported to give a higher yield [47]. Tetrachlorosilane added to the feed gas serves as a carrier and fluidizing agent when silicon residues are utilized as a source of silicon [48]. These residues can originate from the direct synthesis of organochlorosilanes (→Silicon Compounds, Organic, Chap. 2.1.). Higher temperature in the fluidized bed followed by rapid quenching are reported to increase yield 1.8-fold. Quenching is effected by spraying the top of the bed with tetrachlorosilane at 20 °C such that the reaction gases leaving the reactor are cooled to 400 °C in less than one second [50].

Extremely fine particles of silicon (50 – 800 µm), obtained by atomizing molten elemental silicon in a nitrogen atmosphere, are claimed to give a high yield of trichlorosilane [51]. A screw conveyor has also been used for producing trichlorosilane from ferrosilicon [55]. Depending on the cooling rate the tetrachlorosilane to trichlorosilane ratio can be adjusted between 4 : 1 to 0.2 : 5.

In a two-step [56] process silicon is first reacted with tetrachlorosilane at 1100 – 1300 °C and the resulting product is treated with hydrogen chloride. The conversion of tetrachlorosilane to trichlorosilane is 50 – 60 %.

Adding tetrachlorosilane to the hydrogen chloride prior to contact with powdered silicon is claimed to enable complete conversion to trichlorosilane [57].

Silicon powder with 2 % copper [58] or 6 % copper [59], has been used at 350 – 600 °C. The yield of trichlorosilane was 7 – 27 mol % for 2 % Cu, and 80 – 90 mol % for 6 % Cu. Much lower temperatures (down to 260 °C) were possible when antimony pentachloride was used as a co-

catalyst with copper; the trichlorosilane yield was 21 % [60].

Silicon residues from direct synthesis of organochlorosilanes can also be reacted with hydrogen chloride to give trichlorosilane [61].

3.1.2. Preparation from Tetrachlorosilane

Tetrachlorosilane, up to now used as a starting material for silica fillers, quartz, and special glasses, has become an important precursor for trichlorosilane. The basic reaction is

$$1.5\,SiCl_4 + H_2 + 0.5\,Si \rightarrow 2\,HSiCl_3 \quad (1)$$

Although use of a graphite heating element at 1100 – 1400 °C has been reported [62], most processes are operated between 800 and 1400 °C. The residence time is critical and is adjusted by quenching or by using a two-step process. The residence time of the feed should be <2.5 s [63], and quenching carried out below 600 °C within <1 s [64]. A feed stream of tetrachlorosilane – hydrogen mixture with a molar ratio of 1 : 2 in the absence of silicon at 1100, 1200, and 1500 °C, gave trichlorosilane in 27, 31, and 35 %, respectively.

In two-stage processes the reaction with silicon is carried out at lower temperature in the second-stage reactor.

Typical temperatures are 500 – 700 °C for the first stage, and 300 – 350 °C for the second [65]. The yield of trichlorosilane can be doubled by adding hydrogen chloride to the gas from the first reactor prior to entering the second [66]. A reaction mixture obtained by passing tetrachlorosilane and hydrogen through a reactor at 1050 – 1250 °C is then cooled to 250 – 350 °C and reacted with silicon in a second reactor [67], giving a product containing 38 % trichlorosilane and 61 % tetrachlorosilane. It is also possible to react the intermediate dichlorosilylene with hydrogen in a separate reaction chamber [68].

Single-step processes are mostly carried out in fluidized-bed reactors [69, 70] with quenching of the gaseous reaction products. For furnaces operating at ambient pressure, preheating the reactant gases [71], and removing surface oxide coatings from the silicon powder with aqueous acid or alkaline hydrogen fluoride solution raises the trichlorosilane yield to 70 % [72]. Pretreatment of silicon can also be carried out with gaseous hydrogen chloride [73]. While particle size of the silicon and its distribution seems to have little influence on the reaction [74], it is strongly affected by catalysts. Copper, the main catalyst, is added as such to the silicon [75–81]. The reaction kinetics have been studied in the presence of copper chloride [82]. The hydrogenation of tetrachlorosilane has an activation energy of 20 – 25 kcal/mol which is lowered to 10 – 15 kcal/mol by the catalyst; conversion and throughput are improved.

Copper has also been used with cocatalysts such as metal oxides [83] and metal halides. Aluminum halides give trichlorosilane yields of 20 – 30 mol % [84–86]; iron and vanadium halides give comparable yields [87–89].

Using nickel salts as cocatalysts [90, 91] between 500 and 600 °C gave trichlorosilane yields of 36 % whereas powdered nickel alone gave 27 % [92]. Antimony chloride in a pressure reactor gave 21 % [93].

Platinum compounds have also been used [94, 95]. Silicon with addition of 10^{-3} mol % platinum black [96] catalyzes tetrachlorosilane conversion at 550 °C to give an 80 % yield of trichlorosilane. At the same temperature, platinum on carbon gives a mere 5 % yield [97]. Carbon as the sole catalyst has also been reported in a two-step process [98], in a fluidized-bed reactor [99], and in a packed-bed reaction chamber [100], the latter two giving trichlorosilane yields of 20 vol % and 11 mol %, respectively.

The conversion of tetrachlorosilane to trichlorosilane has also been effected electrolytically [101] with 1,2-dimethoxyethane as a solvent, hydroquinone as hydrogen donor, and tetrabutylammonium perchlorate as supporting electrolyte.

3.1.3. Purification

Extreme purification of trichlorosilane is necessary if it is to be used for production of high-purity silicon for semiconductors or photoconductors. This is mainly to remove impurities such as chlorides of calcium, aluminum, titanium, copper, magnesium, iron, boron, and phosphorus, which would remain in the silicon formed by hydrogen reduction. The trichlorosilane is treated with complexing agents

such as thioglycolic acid, β-naphthylamine, and salts of ethylenediaminetetraacetic acid [102]. An extremely pure product is obtained by extraction with CH_3CN [103]. Other methods include adsorption of the impurities on columns of activated silica [104], activated carbon, ion exchangers [105, 106], titanium sponge [107], or by treatment with acetals [108] or salt hydrates, which cause partial hydrolysis of the impurities [109–111]. The purification of trichlorosilane [112, 113], the development of purity criteria, and industrial control of purity [114] are well developed subjects.

The ultimate aim in trichlorosilane purification is the elimination of trace amounts of boron and phosphorus. Both elements are retained in chemical processing in sufficient amounts to impede or even prohibit certain uses in electronics. Boron impurities are first converted from volatile halides or hydrides to nonvolatile acids or oxides, for example, by passing moist nitrogen into the boilers of a multistage distillation [115]. The addition of 0.1 – 1.0 g/t of trichlorosilane decreased the boron content from 1.6 ppm to 0.2 ppb. A similar procedure [116] gave 12 ppb residual boron in a one-step distillation. The addition of silica with at least 0.25 wt % of total hydroxyl groups is also claimed to effect partial hydrolysis [117].

Phosphorus, the other chief contaminant of silicon compounds, is present in chlorosilanes as phosphorous trichloride. During their distillation, it is trapped in the high-boiling residue by complex formation with molybdenum oxychloride and covalent nickel compounds. The phosphorus-metal complexes are thermally stable. Residual phosphorus contents of chlorosilanes distilled were below the analytical detection limit [118, 119]. Manganese dioxide is also reported as a purification agent. It forms phosphorus oxychloride which has a much higher boiling point than phosphorus trichloride and hence remains in the distillation residues [120]. Boron and phosphorus impurities in trichloro- or dichlorosilane can be oxidized with oxygen to give less volatile compounds [121]. Polycrystalline silicon prepared from purified chlorosilanes had residual impurities of 0.1 ppb boron and 0.18 ppb phosphorus.

3.1.4. Waste Disposal

The in-plant, partly on-line conversion of waste from chlorosilane manufacture and use has two chief sources: distillation residues and off-gases which are both hazardous materials. To provide safety, environment protection, and an economic conversion of material that can be used elsewhere or safely deposited, several procedures have been adopted. Distillation sludge is evaporated in the presence of metal chloride slurries giving chlorosilanes [122] and then hydrolyzed with steam in azeotropic hydrogen chloride which continuously gives additional hydrogen chloride [123].

3.1.5. Uses

The purity of trichlorosilane greatly influences its end use. High purity silicon is produced by the pyrolysis of trichlorosilane in the presence of hydrogen [124, 125], for which a large number of techniques have been used, e.g., pyrolysis in a fluidized bed [126], pyrolysis in or over molten silicon [127–129], and epitaxial growth on silicon [130–132] or graphite [133]. Other processes reported include the production of polycrystalline silicon by the pyrolysis of trichlorosilane [134–137], the production of amorphous silicon coatings by CVD [138], and the production of solar silicon [139, 140]. A reactor made of silicon has been described for the pyrolytic process [141].

In fluidized-bed reactors caking of silicon on the inner wall is prevented by starting with a bed of polycrystalline silicon of grain size 700 – 3000 µm, fluidized by a mixture of trichlorosilane and hydrogen [142]. The polycrystalline silicon granules are continuously removed from the reactor.

Other important uses for trichlorosilane include the manufacture of very finely divided ("micronized") silica for use as a filler [143], the manufacture of silicon nitride ceramics, sometimes together with metal nitrides [144], the surface treatment of boron and borides [145], the production of free-flowing granulated fertilizers [146], and in screen printing [147].

3.2. Tetrachlorosilane

Silicon tetrachloride, $SiCl_4$, M_r 169.9, *mp* −69.4 °C, *bp* 57.3 °C, T_c 234.0, p_c 37.5 bar, V_c 326.3 cm g^{-1} mol^{-1}, is a clear, colorless liquid that fumes strongly in air and hydrolyzes rapidly in water, producing SiO_2 gel. When dry, it does not attack steel, and can therefore be transported in steel tanks. It is soluble in benzene, ether, chloroform, and petroleum ether. It reacts with alcohols to give esters of silicic acid. On partial hydrolysis (e.g., with ether – water mixtures), silicon oxychlorides, $Si_nO_{n-1}Cl_{2n+2}$, are formed as colorless, viscous liquids. These oxychlorides, together with the corresponding metallic chlorides, are formed when silicon tetrachloride reacts with metal oxides (e.g., of Mn, Cu, Ca, Zn, Mg, Ag, or Hg) in organic solvents [148].

3.2.1. Production

Silicon tetrachloride is prepared in the laboratory by the reaction of silicon or high-silicon ferrosilicon with dry chlorine at >400 °C. At lower temperature, Si_2Cl_6 and other subchlorides are also formed [15].

In industry, $SiCl_4$ is mainly produced by reacting ferrosilicon, pure silicon, or silicon carbide with chlorine. Other possible processes include the reaction of SiO_2 or a SiO_2/SiC mixture with carbon and chlorine (or with $COCl_2$ or CO/Cl_2 mixture). This older process, which uses silica sand as the cheap raw material, has attracted renewed interest since the introduction of the fluidized-bed reactors.

From Ferrosilicon (>90 % Si) or Pure Silicon and Chlorine. The reaction between silicon and chlorine is exothermic. Therefore, after the reaction has been initiated, cooling must be provided. Ferrosilicon, with a high silicon content, is cheaper than pure silicon, and due to the greater reactivity of the iron, is more readily chlorinated. However, the temperature must be maintained at >400 °C so that the $FeCl_3$ formed is removed, and blockages in the pipework are prevented.

In the Dynamit Nobel process, large pieces of ferrosilicon containing >90 % silicon are continuously fed into a furnace at >500 °C fitted with a cooling jacket. Silicon tetrachloride is formed together with iron chloride and the chlorides of the other elements present in the ferrosilicon, mainly aluminum chloride, calcium chloride, and titanium tetrachloride. The nonvolatile chlorides (e.g., $CaCl_2$) remain in the furnace, and are removed from time to time together with unchlorinated components (e.g., SiO_2).

The volatile chlorides pass out of the furnace and are then partially condensed by cooling. The collecting vessel is maintained at such a temperature that the $SiCl_4$ distils over while the $FeCl_3$ and $AlCl_3$ remain behind. Most of the $SiCl_4$ is then condensed out, and the residual gases are cooled to −35 °C, causing any remaining tetrachloride to separate.

The crude silicon tetrachloride still contains small quantities of iron and aluminum chlorides, and also some chlorine, titanium tetrachloride, and hexachlorodisiloxane formed by partial hydrolysis of $SiCl_4$ by the traces of water that enter the furnace during charging. This is purified by simple evaporation and condensation, followed by fractional distillation. In the evaporation process, which is also applied to the residues from the fractional distillation, chlorides of iron and aluminum remain behind along with some of the medium boiling point compounds, i.e., hexachlorodisiloxane and titanium tetrachloride. The crude distillate obtained is fractionated in a double distillation column.

It has been proposed that a molten salt reaction medium should be used [149], and that inert materials such as carbon, aluminum oxide, and silica should be added in both the batch [150] and the continuous process [151].

From Silicon Carbide and Chlorine.

$$SiC + 2\,Cl_2 \longrightarrow SiCl_4 + C$$

The advantage of using the readily available industrial-grade silicon carbide as the starting material must be balanced against the disadvantages that the reaction with chlorine starts at a higher temperature, and the heat of reaction is not sufficient to maintain the material at the reaction temperature without heating. Also, the carbon formed remains behind as a residue.

Silicon is added to the silicon carbide to facilitate the reaction. The chlorination is carried out in a vertical shaft furnace made of steel or cast iron lined with carbon plates. The downstream condensation system includes a spray of

liquid silicon tetrachloride to prevent blockage by the iron and aluminum chlorides formed from impurities in the raw material [152].

In a modified process, the chlorine is mixed with oxygen, which oxidizes some of the carbon formed in the reaction to carbon monoxide [153, 154]:

$$SiC + 2\,Cl_2 + 1/2\,O_2 \longrightarrow SiCl_4 + CO$$

From Silica, Carbon, and Chlorine. Information on the temperature required for the reaction of various silica modifications with chlorine and carbon is given in [155].

Chlorides of alkali metals or alkaline earths can be mixed with the finely powdered coke and kieselguhr or other silica-containing materials to lower the reaction temperature. The mixture is treated with chlorine at red heat [156]. In a variation of the process, the energy is provided by an electric arc, which is applied to the reaction mixture in a fixed [157] or fluidized bed [158].

As the reaction of SiO_2 with carbon and chlorine is endothermic, it is necessary to couple it with an energy-producing process, e.g., silicon and chlorine, ferrosilicon and chlorine, or silicon carbide and chlorine. In the Stauffer process [159], silicon carbide, silica sand, and coke are treated with chlorine gas in a fluidized-bed reactor. A separator removes any solid material from the gaseous reaction product, which is then cooled in heat exchangers. Silicon tetrachloride condenses out and is purified by the usual distillation process.

Other Processes. These include: treatment of silicon or a Si – Cu alloy with CCl_4 vapor at 260 – 300 °C [160], reaction of alkali metal silicofluorides (e.g., Na_2SiF_6) with $AlCl_3$ at 280 °C [161] or with $MgCl_2$ at 500 – 1000 °C [162], and recovery of byproduct silicon tetrachloride from the production of methylchlorosilanes (→ Silicon Compounds, Organic).

The formation of incrustations in a fluidized-bed reactor can be prevented if a mixture of metallurgical silicon and silicon carbide with silicates of iron, titanium, vanadium, or chromium is used. Increases in yield of up to 12 % have been observed [163].

Lower reaction temperatures, down to 200 °C, and the use of silicides, preferably calcium silicide, yields substantial amounts of tetrachlorosilane together with an appreciable amount of hexachlorodisilane [164, 165]. The conversion of higher chlorosilanes into tetrachlorosilane is effected by tertiary amines in catalytic amounts below 160 °C. Refluxing of hexachlorodisilane with 0.8 vol % of pyridine gave tetrachlorosilane [166].

3.2.2. Purification

The main aim of purification is to remove halides of boron and phosphorus by preferential hydrolysis, complexing agents, or the incorporation of adsorption stages in the distillation process. Examples include preferential hydrolysis of BCl_3 by adding water or ice [167, 168], adding substances to the distillation process that combine with BCl_3 such as CH_3CN and $(C_6H_5)_3CF$ [169, 170], substances that combine with PCl_3 such as $AlCl_3$ [171], and passing the silicon tetrachloride through adsorption columns charged with substances containing tertiary amine groups, quaternary ammonium groups [172], or water-containing oxides or silicates [173]. A countercurrent zone-melting process has also been developed for the purification of $SiCl_4$ [174].

Removing boron impurities is also accomplished by high-temperature hydrogenation, which converts boron trichloride, the chief impurity to volatile diborane [175]. The fact that considerable amounts of silicon hydrides, chiefly trichlorosilane and dichlorosilane, are formed as byproducts, is of no apparent disadvantage since both can now be used in high-purity silicon manufacture.

Low-loss optical fibers require precursors with extremely low contents of hydroxyl-containing impurities. These are probably chloro – hydroxy complexes of metals, mainly iron. They are removed by partial hydrolysis of the tetrachlorosilane with adsorption of the impurities on the in situ precipitated silica gel. The residual contaminant content is 12 ppm Fe, 2 ppm OH, and 18 ppm HCl [176].

3.2.3. Uses

High-purity silicon is produced by the thermal decomposition of $SiCl_4$ in the presence of

hydrogen, and highly dispersed silica is produced by flame hydrolysis of $SiCl_4$ (→ Silica, Chap. 6.1.1.).

Corrosion-resistant silicon coatings are produced on metal surfaces, e.g., steel and iron [177, 178], molybdenum, tungsten, tantalum [179], molybdenum, or tungsten [180] by heating the metal component in an atmosphere of $SiCl_4 - H_2$ or $SiCl_4$ at 1000 – 1400 °C.

3.3. Hexachlorodisilane

For years, hexachlorodisilane was a typical byproduct; its formation was suppressed as far as possible or a conversion to more versatile products carried out. It is now a compound of its own standing leading directly to the production of several special silicon products. The manufacturing processes, however, are still dominated by the basic reaction of silicon or silicides with chlorine giving a number of chlorosilanes, hexachlorodisilane being just one.

3.3.1. Production

To obtain hexachlorodisilane as a high share in a crude chlorosilane product mixture, conversion or degradation processes have been designed. Calcium silicide is the preferred raw material and a low reaction temperature of 200 °C is used [181]. Continuous production in a multicolumn reactor [182], and batch production [183] with a comproportionation step have been described:

$$Si_3Cl_8 + SiCl_4 \longrightarrow 2\,Si_2Cl_6$$

Reaction of ferrosilicon, with a silicon content of 50 %, and chlorine at 160 °C in a vibrating reactor gave 55 wt % hexachlorodisilane, and 44 wt % tetrachlorosilane with a 70 % ferrosilicon conversion [184, 185].

High-boiling residues, predominantly containing polychloropolysilanes, are cleaved by chlorine in a fluidized bed [186] at 250 – 450 °C.

$$2\,Si_3Cl_8 + Cl_2 \longrightarrow 3\,Si_2Cl_6$$

At 500 °C on α-alumina beads [187] comproportion of tetrachlorosilane is readily achieved, with a conversion of 70 %. The reverse reaction is carried out at a temperature of 700 °C aided by lithium – magnesium double oxides as catalysts and alumina as the carrier.

$$2\,SiCl_4 \longrightarrow Si_2Cl_6 + Cl_2$$

The reaction product consists of 50 % unchanged tetrachlorosilane, 40 % hexachlorodisilane, and 10 % higher homologues [188].

Liquid-phase chlorination allows a lower reaction temperature to be used. Using elemental silicon and a catalyst consisting of the chlorides of sodium, potassium, and zinc gives a product mixture that contains tetrachlorosilane, hexachlorodisilane, and tetrachlorodisilane in a weight ratio of 12 : 17 : 11 [189]. Suspension of silicides and antimony pentachloride can be chlorinated to hexachlorodisilane with 34 % conversion [190]. In a solvent mixture in which some components exhibit catalytic properties [e.g., perfluoroalkanes, tris(perfluoroalkyl)amines, and cyclic perfluoroalkyl ethers] elemental silicon reacts with chlorine at 150 °C [191] to give hexachlorodisilane and octachlorotrisilane.

3.3.2. Uses

In most of its applications, hexachlorodilane can be replaced by other special silicon compounds, but in some cases, the reactions take place at much lower temperature than with tetrachlorosilane or trichlorosilane. The deposition of elemental silicon by purely thermal decomposition has been carried out on silicon wire at 750 °C [192], on a quartz plate at 500 °C [193], or on quartz crucibles [194].

Hydrogen as a reducing agent has been used to make bulk polycrystalline silicon above 250 °C [195] or to produce atomic layers of silicon by homoepitaxy [196]. A viscous oil obtained by reacting hexachlorodisilane with tetrabutylphosphonium chloride at 200 °C can be pyrolyzed to give finely divided amorphous silicon [197]. The reaction of hexachlorodisilane with propane on graphite surfaces at 700 °C [198] gives β-silicon carbide. On a cobalt-containing substrate, silicon carbide whiskers are obtained [199]. Mixed carbides of silicon and titanium are made by hydrogenation of a mixture of hexachlorodisilane, titanium tetrachloride, and alkyl halide [200].

Optical fibers can be made by direct surface coating of preforms by flame hydrolysis of hexachlorodisilane [201, 202]. Ultrafine silica particles are formed by thermal decomposition of hexachlorodisilane in an oxygen atmosphere [203]. The method is suitable for the manufacture of quartz glass as a substrate for wave guides [204].

3.4. Other Silicon Chlorides

Hexachlorodisiloxane, intermediate or byproduct in oxidation or hydrolysis reaction, can be obtained as the chief product by reacting tetrachlorosilane with sulfur trioxide 500 °C.

$$2\,SiCl_4 + SO_3 \longrightarrow (Cl_3Si)_2O + SO_2 + Cl_2$$

By comparison, other monosilanes react differently. The bromide and hydride at room temperature give silicon dioxide, the fluoride, not at all even at 600 °C [205]. Formation and degradation of hexachlorodisiloxane can be shown by high-temperature reactions. It is formed from tetrachlorosilane and oxygen [206] together with other oligomeric siloxanes but with mono- and polycyclic siloxanes also. Thermodynamic studies indicate a certain metastability of halosiloxanes in general. No direct use has been made of hexachlorodisiloxane, conversion to chlorosilanes, i.e., recycling, is practised [207]:

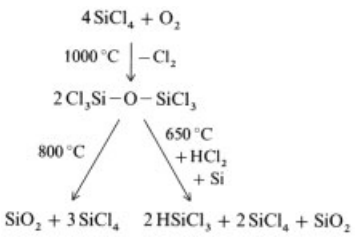

An important intermediate is dichlorosilylene, the first reaction product of the attack of chlorine on silicon. Dichlorosilylene can be obtained directly in high-vacuum apparatus from the elements at 1250 °C [208, 209], and 1100 °C [210]; it is a transparent, amber-like solid.

3.5. Silicon Fluorides (Silicon Tetrafluoride, see → Fluorine Compounds, Inorganic, Chap. 5.1.)

Difluorosilylene Silicon tetrafluoride and silicon react at 1200 °C in a high vacuum, yielding a condensate with the composition $(SiF_2)_x$. This turns yellow on cooling with liquid air and ignites on contact with atmospheric oxygen. It loses its color on heating to −78 °C and becomes highly polymeric [211].

A plasma flame in an argon atmosphere [212, 213] can also be used as a reaction medium. A molar ratio of tetrafluorosilane to silicon of 0.8 : 1 gives a SiF_2 yield of 70 %.

Difluorosilylene can be used to produce silicon difluoride filaments [214]. The production of amorphous silicon films from SiF_2 by chemical vapor deposition leads to material with residual Si – F groups [215], which is claimed to afford a higher doping efficiency [216] for boron and phosphorus.

Reacting SiF_2 with methane and hydrogen in a glow discharge [217] gives a photoconductive amorphous silicon carbide. Silicides can be prepared by reacting hexafluorides of tungsten or molybdenum with difluorosilylene in a hydrogen atmosphere at 300 – 600 °C [218]. Oxidation of SiF_2 with oxygen, nitrous oxide, or nitrogen dioxide [219] gives silicon dioxide with residual Si – F groups, suitable for the manufacture of optical fibers.

3.6. Silicon Bromides

Silicon tetrabromide, tetrabromosilane, $SiBr_4$, M_r 347.7, is produced from silicon and bromine vapor at >600 °C. It is a colorless liquid, with *mp* 5 °C and *bp* 153 °C.

With ferrosilicon powder a 70 % yield of $SiBr_4$ is obtained. Crude tetrabromosilane is heated with zinc to remove residual bromine [220].

The main use for silicon tetrabromide is as a raw material for the production of silicon and its compounds. The reaction used is the reverse of the preparation method, i.e., it is pyrolytically decomposed [221, 222], usually together with silanes, using fluidized-bed processes [223], or

crystal growth techniques [224]. These techniques are also used in the manufacture of optical fibers [225–228] and photoconductive silicon [229].

In a plasma, tetrabromosilane is readily cleaved to the elements, and the deposition of the silicon layer is much faster than with tetrachlorosilane [230]. Tribromosilane can be produced by reacting metallurgical-grade silicon with tetrabromosilane at 600 – 800 °C in hydrogen atmosphere [231].

Other processes include the electrolytic deposition of silicon from solutions of silicon tetrabromide in aprotic organic solvents [232–234].

Silicon tetrabromide is also used to produce silicon nitride itself [235–238], as a coating, or to seal porous solids by exposing them to $SiBr_4$ vapor followed by reaction with ammonia [239]. Boron – silicon phases have been obtained by the copyrolysis of halogenated silanes and boranes [240].

Silicon tetrabromide is used as an auxiliary in the production of SnO_2 whiskers [241], the etching of aluminum and its alloys [242], and the copolymerization of dienes and vinyl compounds [243].

The entire series of bromosilicon hydrides are obtained by disproportionation [244]. With Amberlite A 21 as catalyst at 135 °C the major product is dibromosilane. Fluorobromosilanes are made by reacting tetrabromosilane with antimony trifluoride [245].

Fluorobromosilicon hydrides are made by reaction with alkyltin hydrides [246] at room temperature under reduced pressure:

$$SiBr_4 \xrightarrow{SbF_3} SiF_2Br_2 \xrightarrow{R_3SnH} HSiF_2Br$$

Comproportionation can be extended to the disilanes:

$$Si_2Br_6 \xrightarrow{Si_2Cl_6} Si_2Br_xCl_y$$

Separation of the products by distillation is difficult due to continuing halogen redistribution [247]. Since aryl groups on silicon can be cleaved by hydrogen halides in the presence of aluminum halide catalysts, this method can be used to prepare higher silicon bromides:

$$Ph_3Si-SiCl_3 \xrightarrow{HBr/AlBr_3} Br_3Si-SiCl_3$$

A well-defined, very reactive subhalide with the composition $(SiBr_2)_x$ is obtained by the reaction of $SiBr_4$ and silicon at 1150 – 1200 °C in a high vacuum. It is a transparent, almost colorless, amber-like solid, soluble in benzene, xylene, and CCl_4 [15,248,249].

3.7. Silicon Iodides

Silicon iodide, SiI_4, M_r 535.7, *mp* 121 °C, is a crystalline substance that is very sensitive to hydrolysis, decomposes into its elements on heating or on exposure to light, and reacts with oxygen at elevated temperature with liberation of iodine. SiI_4 is produced in 70 % yield by passing a carrier gas saturated with iodine vapor over silicon containing 4 % copper at 600 – 700 °C [15]. Silicon tetraiodide can be obtained on the kilogram scale in quantitative yield by passing nitrogen saturated with iodine vapor at ca. 200 °C over granulated silicon at 1150 – 1200 °C. Approximately 1.5 kg SiI_4 can be produced in 5 h by this method [250].

Other methods include the reaction of mixtures of iodine with bromine or chlorine with silicon at 600 – 800 °C [251], and heating alkali metal fluorosilicates with alkaline-earth iodides at 500 – 1000 °C [252] or with aluminum iodide at 300 °C [253].

Like the tetrachloride, tetraiodide can be used for the production of high-purity silicon by thermal decomposition (sometimes in the presence of hydrogen). The silicon tetraiodide can be purified beforehand by repeated sublimation and zone melting [254,255].

Silicon tetraiodide has some value as an additional reactant in the pyrolysis of silicon hydrides for the production of high-purity silicon [135], in the reduction of other silicon halides in a fluidized bed [256], and for the purification of silicon by zone melting [222].

Silicon tetraiodide is used as a starting material for the preparation of high-purity silicon films for solar cells by electrolyzing a solution of SiI_4 in aprotic organic solvents [232,233].

Special applications of silicon tetraiodide as a raw material include the manufacture of cermets [230], and of optical fibers with a germanium core and an outer coating of SiO_2 [257].

A subhalide having the composition of diiodosilylene has been reported [3].

4. Silicon Oxides (Silicon Dioxide, → Silica)

With the oxides, more so than with the halides, the growth of interest in use and production, i.e., formation, is brought about by the same driving forces as for the halides, i.e., electronics manufacture and environment protection.

Sources of silicon oxides are chiefly high-temperature processes in heavy industry like blast furnaces, carbide manufacture, coke making, and silicate processing. Main uses are found in high-tech precision preparation of chips, wafers, integrated circuits, layers, and coatings.

4.1. Silicon Monoxide

Silicon monoxide, SiO, M_r 44.06, is formed initially as a gas when SiO_2 and silicon, or SiO_2 and an amount of carbon insufficient for complete reduction, are heated to >1250 °C. It therefore plays a part in the industrial production of silicon and silicon carbide from SiO_2 and carbon, and in the production of ferrosilicon [258, 259].

At lower temperature (ca. 600 – 1000 °C), SiO disproportionates into silicon and silicon dioxide:

$$2\,SiO \rightleftharpoons SiO_2 + Si$$

There has, therefore, been prolonged controversy over the question whether SiO should be considered a true compound. Nevertheless, rapid chilling of the gas to room temperature enables the monoxide to be obtained in a solid metastable state. It has a wide range of uses as an additive in applied solid-state physics, and as a raw material in the production of other silicon compounds.

4.1.1. Production

A stoichiometric mixture of finely powdered SiO_2 and silicon in a reaction tube closed at one end is heated at its lowest point in a high vacuum for ca. 4 h at 1250 °C. A brown mixture of SiO_2 and Si, formed by disproportionation, collects in a short transition zone in the part of the tube that just protrudes from the furnace, and the SiO collects as black, compact, shellac-like masses in the parts further away from the furnace.

If an intimate mixture of SiO_2 and silicon is heated for 9 h to 1300 °C and rapidly chilled, cubic crystals of SiO are formed [15]. On an industrial scale, the monoxide can be produced by heating SiO_2 with silica, carbon, or silicon carbide to 1500 – 1800 °C in a resistance furnace [260, 261].

Heating the vapors of the reaction

$$SiO_2 + Si \xrightarrow{1500\,^\circ C} 2\,SiO$$

to 1800 °C in an ammonia atmosphere gives amorphous silicon monoxide powders [262].

Silicon monoxide can be made directly in the form of fibers by reducing sand flour with methane at 2500 °C in a fluid-wall reactor [263].

The formation of silicon monoxide by side reactions or as a byproduct of the main process takes place in the course of a number of industrial production streams. In terms of chemical categories these production steps are based on reductions as well as oxidations.

$$SiO_2 \begin{matrix} + Si \\ + SiC \\ + C \end{matrix} \longrightarrow SiO \longleftarrow \begin{matrix} Si \\ SiC \\ SiMe \end{matrix} + O_2$$

4.1.2. Properties

At room temperature, silicon monoxide is a light brown, finely divided powder, or, if it has been produced by slow condensation, it is a black or very dark brown, shellac-like, glassy mass. The lower limit of stability for gaseous SiO is not below 1025 °C, and is very probably >1175 °C [264]. The best temperature range for its production from SiO_2 and Si is 1200 – 1400 °C [265–267].

At 600 °C, SiO decomposes in several hours, at 700 – 900 °C in ca. two hours, and at 1000 – 1200 °C almost instantaneously into SiO_2 and silicon [266, 268]. Oxygen causes surface oxidation to SiO_2 even at room temperature, and complete oxidation takes place at ca. 500 °C.

Finely divided SiO is pyrophoric and burns with flame formation if there is rapid access of air.

With the increase in the number of possible uses for SiO, the silicon – oxygen system has been the subject of extensive investigations [269], particularly with regard to suppressing the formation of SiO during the production of high-purity silicon [270], and of Si – O – N mixed phases in atmospheres containing O_2 and N_2 [271].

4.1.3. Uses

Silicon monoxide is used in small quantities to produce protective films on semiconductors [272–274], vidicon targets [275], and in antireflective coatings on optical glass and plastics [276–279]. Silicon monoxide prevents the formation of anatase in rutile coatings on glass [280]. In entertainments electronics and data processing, silicon monoxide is used in recording tape, other types of data storage media, data transfer heads [281, 282], and in the alignment of liquid crystals for digital displays [283, 284]. Other applications include the coloring of glass surfaces by the simultaneous or sequential deposition of metals and silicon monoxide [285].

Another application of silicon monoxide electronics is the manufacture of heat-resistant films and devices in which corrosion resistance is required. It can be used as a heat-resistant protective layer in laser recording media [286], and in multilayer thermal heads for thermographic printers [287]. Multilayer devices consisting of aluminum, germanium, and silicon monoxide are selective solar absorbers for thermal conversion [288]. Ceramics consisting of silicon carbide whiskers in a graphite matrix can be reinforced by decreasing the porosity by filling the voids with silicon prepared in situ [289].

Silicon monoxide is used as a raw material, a construction material, and as an auxiliary. The best known application is the oxidation of gaseous silicon monoxide to amorphous finely divided silica [290], (see → Silica, Chap. 6.2.). Other uses include the production of silicon carbide [291, 292], silicon nitride [293–295], Si – O – N mixed phases [296], and cermets [297, 298]. Silica can be removed from silica-containing materials (e.g., kaolin) by adding silicon and heating the mixture at 1450 °C in a vacuum [299–301]:

$$Al_2O_3 \cdot 2\,SiO_2 + 2\,Si \longrightarrow Al_2O_3 + 4\,SiO_{Gas}$$

Silicon monoxide can be used to remove phosphorus from blast furnace slag, allowing it to be recycled [302].

Hard coatings of SiO_2 on metals, especially on metallic mirrors, can be produced by deposition of SiO from the vapor state (→Thin Films). The SiO is then oxidized by atmospheric oxygen to form SiO_2.

In the production of silicon nitride ceramic powders, silicon monoxide is reacted with a reducing atmosphere consisting of a mixture of ammonia and hydrocarbons at 1000 °C. When methane is used as sole reductant, the product is cubic β-silicon carbide [303].

A convenient method to prepare members of the halosilane family is the reaction of silicon monoxide with hydrogen halides [304].

Low-silicon metals are produced by effecting silicon depletion of the melt by limited oxidation. In the production of high-purity vanadium, vanadium pentoxide is added to the melt and the silicon monoxide formed is evaporated [305]. Silicon in molten indium is oxidized with water in a hydrogen atmosphere at 800 °C; ca. 90 % of the initial silicon content is converted to SiO [306].

4.2. Silicon Sesquioxides

Depending on temperature and pressure, the following disproportionation steps can take place:

$$4\,SiO \rightleftarrows Si_2O + Si_2O_3$$
$$Si_2O_3 \rightleftarrows SiO + SiO_2$$

When the monoxide is evaporated at reduced pressure the vapor invariably condenses as the sesquioxide, Si_2O_3 [307].

The reaction of sulfur trioxide with hexabromodisilane at room temperature affords an almost quantitative yield of the sesquioxide; no cleavage of the Si – Si bond takes place [308]:

$$Br_3Si{-}SiBr_3 + 3\,SO_3 \longrightarrow Si_2O_3 + 3\,SO_2 + 3\,Br_2$$

The sesquioxide can be prepared from hexachlorosidilane by hydrolysis via a sol – gel route [309].

Polymeric silicon sesquioxide can be prepared by thermal dehydrogenation of silicon oxyhydride in high vacuum [310].

$$2\,HSiO_{1.5} \xrightarrow{1000\,^\circ C} Si_2O_3 + H_2$$

Silicon sesquioxide can be prepared in situ on surfaces. Plasma – vapor deposition from a mixture of monosilane and nitrous oxide [311] gives a moisture-resistant film of silicon sesquioxide, used as a protective layer for *p*-Si substrates. On polycarbonate sheets, Si_2O_3 layers are prepared by electric discharge [312] or vacuum deposition [313]. On polyethylene [314] a Si_2O_3 layer has been applied by sputtering with an oxygen plasma, using silicon monoxide as the silicon source.

The suboxides, the sesquioxide being the predominant species, are present in semiinsulating films of polycrystalline silicon [315]. Annealing at 1100 °C converts the other suboxides to the sesquioxide only. The suboxides are also formed on etched surfaces of silicon [316] and in layers obtained by silicon implantation oxidation [317]. Thermal annealing leads to a composition corresponding to that of the sesquioxide when carried out at 600 °C.

Less for effects in high-tech physics than essentially in basic chemical processes, the sesquioxide has found quite different applications. Cleaning and purification spans from that of arsenic vapors [318] to the treatment of waste fluid from sweet potato processing [319]. Highly permeable membrane separation elements are made from a ceramic material based on alumina and silicon sesquioxide [320]. The material is stable in aqueous environment having a water penetration rate of 4.2 m^3/h. In ceramic material the sesquioxide has been used as an oxidizing agent during manufacture [321] and as bits for welldrilling [322]. Wear resistance is much lower though, than for tungsten carbide.

4.3. Other Silicon – Oxygen Compounds

Oxyhydrides of the composition $(HSiO_{1.5})_x$ are obtained by the hydrolysis of trichlorosilane [323–325]. In organic solvents the product is formed as an amorphous powder or as transparent mica-like flakes [310], depending on the type of solvent. Well-defined, low molecular oligomers – the smallest [326, 327] being a cage of 8 silicon atoms arranged at the corners of a cube with 12 oxygen atoms forming the bridges along the edges – can be prepared by hydrolysis in a two-phase system [328–330]. Oligomers with *x* an even number ranging from 8 to 16 are obtained.

The Si_8-cube is also the smallest cage in zeolite structures; it is the 4 – 4 secondary building unit in zeolite A. Since the silver zeolite reacts with water evolving oxygen, ligand-to-metal charge transfers in the Si_8-cube are of interest in the context of processes involving photolysis of water by visible light [331]. By photochemical oxidation of tetrachlorosilane the trichlorosilanol is produced, a compound existing only in equilibria with others [332]. Utilizing structural possibilities new silicon – oxygen compounds were prepared, the oxygen taking part as the donor atom in penta-coordinate silicon alkoxide complexes. Alkoxysilanes react with metal hydrides to trialkoxy anions which then rearrange to the tetraalkoxy anion [333]

$$4\,HSi(OR)_3 + 4\,KH \longrightarrow 4\,\&H_2Si(OR)_3]^-K^+$$
$$\longrightarrow 3\,\&HSi(OR)_4]^-K^+ + SiH_4 + KH$$

Likewise the reaction of silica with glycol is made possible through penta-coordinate complex formation around the silicon atom [334].

5. Silicon Sulfides

Silicon sulfide, SiS_2, M_r 92.18, *mp* 1090 °C, sublimation point 1130 °C, forms white, asbestos-like needles, which readily hydrolyze. It is formed from silicon and sulfur at 1100 – 1300 °C [335], but preferably at >1400 °C [336, 337]. Silicides react with sulfur at >700 °C, hydrogen sulfide at 1100 – 1200 °C, or sulfur chlorides (e.g., S_2Cl_2, at its boiling point of 135 °C). The SiS_2 formed must be recovered from the other reaction products (e.g., MgS) by sublimation.

The simplest laboratory method consists of a combination of these two processes. The reaction of Mg_2Si with S_2Cl_2 is used to start the reaction of silicon with sulfur. A mixture of 3 parts by weight of Si, 0.5 parts by weight of Mg_2Si, and 8 parts by weight of sulfur is impregnated

with 1 part by weight of S_2Cl_2 in a glass tube and heated at one point with a small flame. The reaction starts with a glow. The SiS_2 formed is sublimed off in a vacuum [336].

On a large scale, SiS_2 can be produced by the thermal decomposition of esters of thiosilicic acid. The ethyl ester decomposes at 300 °C:

$$Si(SC_2H_5)_4 \longrightarrow SiS_2 + 2\,C_2H_5SC_2H_5$$

The reaction temperature can be lowered to 200 °C by adding sulfur [336, 338].

The formation of silicon – sulfur compounds, including SiS_2, has also been observed during the drilling of holes in crystalline silicon plates by laser beams in an atmosphere of sulfur hexafluoride [339].

Direct synthesis from the elements is still being carried out and improved. Continuous feeding into reactors [340] allows a lower reaction temperature of 800 °C to be used. Under a pressure of 2.5 – 6.0 GPa and 1000 – 1200 °C the conversion to the sulfide is almost complete, yielding four different polymorphic modifications of silicon sulfide [341].

Silicon disulfide can be conveniently obtained from silica, carbon, and sulfur vapors in an atmosphere of hydrogen sulfide above 1130 °C [342]. The product consists of glassy solid and needle-like crystals.

Silicon disulfide reacts with ammonia to give silicon nitride, Si_3N_4. Between 800 and 1400 °C the nitride is obtained as an amorphous powder. If a lower temperature and silicon disulfide in the form of whiskers are used, the product retains the morphological characteristics and consists of silicon nitride whiskers [343].

Because of their volatility, the formation of silicon sulfides at high temperature is utilized for desiliconization in purification processes. For example, pure zirconium oxide is obtained by conversion of silica impurities followed by vaporization at 1300 °C [344]:

$$SiO_2 + C + 2\,S \longrightarrow SiS_2 + CO_2$$

Glassy phases containing silicon disulfide can be used as solid electrolytes in conducting glasses for batteries, especially microbatteries. A system consisting of a $Li_2S - SiS_2(P_2S_5)$ phase between a $Li/V_2O_5 - TeO_2$ couple has an open circuit voltage of 2.8 – 3.1 V [345, 346].

From a melt of equimolar amounts of silicon disulfide and lithium sulfide, lithium silicon sulfide is obtained in three phases: equilibrium, metastable crystalline, and glassy Li_2SiS_3 [347]. The conductivity is greatly enhanced by the addition of lithium bromide [348, 349], iodide [350, 351], or chloride [349]. The highest conductivity reported [333] is 1.8 s/cm at 25 °C, measured in a glass consisting of the phase $0.24\,SiS_2 - 0.36\,Li_2S - 0.4\,LiI$.

Glassy phases containing silicon disulfide are also obtained by the action of elemental silicon on arsenic sulfide [352]

$$5\,Si + 2\,As_2S_3 \longrightarrow 2\,SiS_2 + 2\,As_2Si$$

6. Silicon Nitrogen Compounds

The prevailing chemical of this class is the silicon nitride, Si_3N_4, a solid of regular valence composition that has found commercial use in high-temperature engineering ceramics and in the form of whiskers as reinforcing material in cermets. A number of processes are now available to produce the nitride as a powder, a film, or as needles.

For nitride production, choice of process is largely influenced by the material, of having it at disposal or having to dispose of it. Byproducts and wastes from other production lines of silicon chemistry manufacturing decide on availability and necessity.

The direct synthesis from the elements has been carried out only to study the effects on silicon surfaces. Between 1350 and 1450 °C the nitride is formed [353]. Nitridation is enhanced by the presence of iron in silicon and by hydrogen in the nitrogen atmosphere, the latter also increasing the ratio of α-/β-silicon nitride [354]. Gas-phase reactions of monosilane and ammonia [355] can be used for deposition of silicon nitride.

Of the silicon halides, the tetrachlorosilane can be reacted with a nitrogen – hydrogen mixture giving amorphous nitride [356]

$$3\,SiCl_4 + 2\,N_2 + 12\,H \longrightarrow Si_3N_4 + 12\,HCl$$

Hexachlorodisilane is useful to obtain silicon nitride whiskers [357] by chemical vapor deposition in the course of the reaction with ammonia. Between 950 and 1700 °C the resulting

whiskers are coiled in shape; they exhibit good dispersibility [358]. The reaction can be extended to include other halide derivatives of disilane [357]. Finely divided nitride is obtained using silicon fluorides as the silicon source [359]. Making silicon nitride from the disulfide can result in needles [343,360]. By reacting evaporated silicon disulfide with ammonia above 1130 °C [361] or in a two-step process first at 800 °C, then at 1400 °C [362] the product is a powder of α-Si_3N_4. During the synthesis of silicon carbide the nitride is formed through contact of the carbide with nitrogen at temperatures below 2050 K [363]. For a variety of silicon or nitrogen sources the reaction temperature of 900 to 1100 °C seems practical, the energy furnished by a plasma induced by radio frequency discharge [364]. A critical review of thermal plasma reactions leading to silicon nitride from different silicon sources presents data on the thermodynamics and nucleation kinetics [365]; i.e., a high supersaturation of silicon should always be avoided.

7. Silicon Borides

The long-known binary silicon borides, SiB_3 and SiB_6, are produced simultaneously when the constituent elements are heated between carbon electrodes in an electric furnace [366]. Phases with the composition SiB_4 have since been produced and evaluated for a large number of application areas.

The compound SiB_4 is produced from its elements at 1090 – 1370 °C by resistance heating and is purified by treatment with a mixture of nitric and hydrofluoric acids. It can also be produced from boron trioxide and silicon [367]. It is used as a refractory material in blast furnaces. Its use as a catalyst for the liquid-phase oxidation of alkenes with organic hydroperoxides to form oxiranes has also been reported [368]. Mixed phases of silicon borides and molybdenum silicides resistant to oxygen up to 1600 °C are prepared by sintering a pasty mixture of these materials with colloidal silica at 1300 – 1500 °C. They can be used as abrasives and as highly refractory materials [369].

8. References

1. A. Stock: *Hydrides of Boron and Silicon,* Ithaca, New York 1933 ; Z. Elektrochem. 32 (1926) 341.
2. A. G. McDiarmid, *Quart. Rev.* **10** (1956) 208.
3. *Gmelin,* **System. no. 15, part B,** 227 ff.
4. A. Stock, C. Somieski, *Ber. Dtsch. Chem. Ges.* **49** (1916) 125, 155.
5. H. J. Emeleus, A. F. Maddock, *J. Am. Chem. Soc.* (1946) 1131.
6. H. Clasen, *Angew. Chem.* **70** (1958) 179.
7. W. C. Johnson, S. Isenberg, *J. Am. Chem. Soc.* **57** (1935) 1349.
8. A. E. Finholt, A. C. Brown, K. E. Wilzbach, H. I. Schlesinger, *J. Am. Chem. Soc.* **69** (1947) 2692.
9. Internat. Standard Electric Corp., US 2 888 328, 1956.
10. Kali-Chemie, DT 1 034 159, 1956; DE-AS 1 049 835, 1956.
11. Allied Chem. Corp., DT 1 080 531, 1959.
12. Standard Telephone & Cables, GB 832 333 and 832 334, 1956.
13. W. Sundermeyer, O. Glemser, *Angew. Chem.* **70** (1958) 625.
14. Ethyl Corp., EP 300 320, 1989 (J. A. Bossier, D. M. Richards, L. T. Crasto).
15. Ethyl Corp., US 4 778 668, 1988 (E. M. Marlett, R. N. DePriest).
16. Union Carbide, CA 1 151 305, 1983 (W. C. Breneman).
17. Tokuyama Soda, JP 59 174 516, 1984.
18. Union Carbide, CA 1 184 017, 1985 (W. C. Breneman, L. M. Coleman).
19. Tokuyama Soda, JP 60 060 917, 1985.
20. Solavolt International, US 4 542 004, 1985 (K. R. Sarma, C. S. Chanley).
21. J. Moriyama: "Production of High-Purity Silicon. A Review," *Suiyokaishi* **20** (1988) no. 10, 671 – 679.
22. Mitsui Toatsu Chemicals, JP 58 156 522, 1983.
23. Toa Gosei Chemical, JP 01 234 316, 1989 (S. Watanabe, T. Hattori).
24. Toa Gosei Chemical, JP 61 191 512, 1986 (O. Hirano et al.).
25. Bell Téléphone, BE 566 348, 1960.
26. Licentia Patentverwertung, DT 1 073 460, 1958.
27. V. V. Balabanov et al., *Vysokochist. Veshchestva* **6** (1990) 60 – 66.
28. Toshiba Corp., JP 58 048 417, 1983.
29. Seiko Instruments and Electronics, JP 63 132 420, 1988 (N. Shimizu, M. Shinho).
30. Canon, DE 3 208 494, 1982 (I. Shimizu, K. Ogawa, E. Inoue).

31. S. W. Depp, W. E. Howard, *Spektrum der Wissenschaft* 1993, 547.
32. Bell Laboratories, US 4 579 752, 1986 (L. H. Dubois, R. G. Nuzzo).
33. J. R. Hartmann, J. Famil-Ghiriha, M. A. Ring, H. E. O'Neal, *Combust. Flame* **68** (1987) no. 1, 43 – 56.
34. Pioneer Electric Corp., JP 61 090 726, 1986 (K. Kitahara, T. Shimada).
35. A. Y. Adamova et al., *Kim. Vys. Energ.* **15** (1981) no. 4, 365 – 369.
36. C. B. Moore, J. Biedrzycki, F. W. Lampe, *J. Am. Chem. Soc.* **106** (1984) no. 25, 7761 – 7765.
37. T. N. Bell, K. A. Perkins, P. G. Perkins, *J. Chem. Soc. Chem. Commun.* 1980, no. 22, 1046 – 1047.
38. H. Buff, F. Wöhler, *Liebigs Ann. Chem.* **104** (1857) 94.
39. G. Braver: *Handbuch der präparativen anorganischen Chemie,* vol. 1, F. Enke, Stuttgart 1960.
40. Wacker, DT 1 067 010, 1951; BASF, DT 1 105 389, 1960.
41. Linde, US 2 499 009, 1947.
42. Union Carbide, US 2 595 620, 1948.
43. Dynamit Nobel, BE 856 480, 1977.
44. Wacker, DT 623 290, 1977.
45. Telecomunicações Brasileiras, BR 8 006 892, 1980 (J. W. C. Carvalho et al.).
46. A. C. F. de Arruda, L. C. Barbosda, *Congr. Ann. ABM, 35th,* **vol. 3,** Assoc. Brasileira Metais, Sao Paulo 1980, pp. 375 – 390.
47. V. F. Popenko et al., SU 833 494, 1981.
48. General Electric, DE 3 303 903, 1983 (A. Ritzer, N. B. Shah, E. Daniel).
49. Toa Nenryo, JP 63 170 210, 1988 (H. Yamada, H. Ogawa, T. Hosokawa, M. Tachikawa).
50. Hüls, DE 3 809 784, 1989 (K. Ruff).
51. Elkem, DE 3 938 897, 1990 (K. Forwald, G. Schüssler, O. Soerli).
52. High Pure Silicon Co., JP 02 145 413, 1990 (M. Takaguchi, N. Moriika).
53. T. C. Hwang, H. H. Hsieh, *Energy Res. Abstr.* **24** (1984) no. 24.
54. K. X. Li, S. H. Peng, T. C. Ho, *AIChE Symp. Ser.* **84** (1988) 114 – 125.
55. Nippon Aerosil Co., DE 3 230 590, 1983 (T. Ito, H. Hori).
56. Mitsubishi Kakoki Kaisha, JP 02 172 811, 1990 (K. Akaike).
57. C. Zhang, CN 87 100 535, 1987.
58. Shin-Etsu Chemical Ind., JP 58 161 915, 1983.
59. Osaka Titanium Co., JP 56 073 617, 1981.
60. Mitsui Toatsu Chemicals, JP 63 100 014, 1988 (K. Inoe et al.).
61. General Electric, US 4 390 510, 1983 (A. Ritzer, B. Shah, D. E. Silva).
62. Mitsubishi Metal Corp., JP 57 003 711, 1982.
63. Council of Sci. and Ind. Research, IN 166 734, 1990 (V. G. Neurgaonkar et al.).
64. Denki Kagaku Kogyo, JP 60 081 010, 1985.
65. Motorola, EP 133 209, 1985 (W. M. Ingle, S. M. Peffley, H. S. N. Setty).
66. High Pure Silicon Co., JP 51 118 017, 1982.
67. Wacker, DE 3 024 319, 1982 (A. Goeppinger et al.).
68. Nippon Steel Corp., JP 63 112 410, 1988 (M. Onozawa).
69. Shin-Etsu Chemical Ind., JP 57 140 309, 1982.
70. Mitsubishi Kakoki Kaisha, JP 63 008 207, 1988 (K. Akaike).
71. Nippon Steel Corp., JP 62 021 707, 1987 (M. Onozawa).
72. Mitsubishi Metal Corp., JP 62 288 109, 1987 (E. Kimura, K. Ogi, T. Kurashige).
73. Wacker, DE 3 341 340, 1985 (A. Schnegg, R. Rurlaender).
74. J. Y. P. Mui, D. Seyferth, *Sci. Tech. Aerosp. Rep.* **19** (1981) no. 18.
75. J. Y. P. Mui, D. Seyferth, *Energy Res. Abstr.* **6** (1981) no. 6.
76. Rhône-Poulenc, EP 100 266, 1984 (J. L. Lepage, G. Simon).
77. Osaka Titanium, JP 57 140 312, 1982.
78. Osaka Titanium, JP 57 140 311, 1982.
79. Osaka Titanium, JP 57 129 817, 1982.
80. Denki Kagaku Kogyo, JP 59 035 017, 1984.
81. Denki Kagaku Kogyo, JP 55 045 919, 1984.
82. W. M. Ingle, M. S. Peffley, *J. Electrochem. Soc.* **132** (1985) no. 5, 1236 – 1240.
83. Mitsui Toatsu, JP 01 313 317, 1989 (K. Inoue et al.).
84. Mitsui Toatsu, JP 63 095 107, 1988 (K. Inoue et al.).
85. Mitsui Toatsu, JP 63 095 108, 1988 (K. Inoue et al.).
86. Mitsui Toatsu, JP 63 100 015, 1988 (K. Inoue et al.).
87. Mitsui Toatsu, JP 63 095 109, 1988 (K. Inoue et al.).
88. Mitsui Toatsu, JP 63 100 016, 1988 (K. Inoue et al.).
89. Mitsui Toatsu, JP 63 095 110, 1988 (K. Inoue et al.).
90. Osaka Titanium, JP 57 156 319, 1982.
91. Osaka Titanium, JP 58 011 042, 1983.
92. Z. Shen, J. Wang, CN 85 107 465, 1987.
93. Mitsui Toatsu, JP 63 095 111, 1988 (K. Inoue et al.).

94. Nippon Kokan, EP 255 877, 1988 (N. Yoneda et al.).
95. Nippon Kokan, JP 01 100 011, 1989 (K. Ozaki et al.).
96. Mitsubishi Metal Corp., JP 62 256 713, 1987 (E. Kimura, K. Ogi, T. Kurashige).
97. Idemitsu Kosan, JP 62 270 413, 1987 (N. Tanaka).
98. High Pure Silicon Co., JP 57 156 318, 1982.
99. V. F. Kochubei, *Otkrytiya, Izobret., Prom. Obraztoy, Tovarnye, Znaki* 1981, no. 45, 122.
100. Osaka Titanium, JP 62 235 205, 1987 (S. Morimoto).
101. Mitsui Toatsu, JP 63 250 484, 1988 (M. Murakami et al.).
102. Wacker, GB 906 617, 1960.
103. Siemens & Halske, DT 955 415, 1955.
104. Wacker, DT 1 134 973, 1958.
105. Union Carbide, DT 2 507 864, 1975.
106. Smiel, DT 2 852 598, 1979.
107. Bochkarev, SU 504 699, 1976.
108. Dynamit Nobel, DT 1 767 667, 1976.
109. Wacker, DT 2 546 957, 1977.
110. Akademie der Wissenschaft, DL 135 613, 1979.
111. Akademie der Wissenschaft, DL 136 493, 1979.
112. A. A. Efremov, Y. D. Zel'venskii, *Tr. Mosk. Khim. Tsuetn. Inst. im. D. J. Mendeleeva* **96** (1977) 4 – 10.
113. Z. A. Elunika, V. J. Lobov, J. P. Polokhina, *Tovetn. Met.* 1980, no. 3, 72 – 73.
114. Wacker, DT 2 558 183, 1977.
115. B. Koehler et al., DD 158 322, 1983.
116. L. Y. Shvartsman et al., *Vysokochist, Veshchestva* 1989, no. 5, 136 – 140.
117. Dow Corning, EP 107 784, 1984 (R. S. Doornbos).
118. General Electric, DE 3 423 611, 1986 (W. D. Kray).
119. General Electric, JP 59 097 519, 1984.
120. General Electric, JP 59 097 518, 1984.
121. Motorola, US 4 409 19, 1983 (R. D. Darnell, W. M. Ingle).
122. Union Carbide, DE 3 709 577, 1987 (W. C. Breneman, C. C. Yang, G. Henningsen).
123. Hüls Troisdorf, DE 3 642 285, 1988 (B. Falk, W. Graetz, K. Ruff).
124. Osaka Titanium, JA 73 047 500, 1973.
125. Mitsubishi Metal, JA 78 097 996, 1978.
126. Texas Instruments, US 3 963 838, 1976.
127. Pennish, US 4 176 166, 1979.
128. Mitsubishi Paper Mill, JA 77 144 959, 1978.
129. Wacker, BE 857 689, 1978.
130. Harris-Intertype, US 3 935 040, 1976.
131. Siemens, DT 2 535 813, 1976.
132. Hitachi, JA 77 044 166, 1977.
133. Dow Corning, US 3 961 003, 1976.
134. Kuratomi, JA 74 076 721, 1974.
135. Kuratomi, JA 74 059 729, 1974.
136. Union Carbide, US 3 900 660, 1975.
137. Motorola, NL 75 008 684, 1976.
138. Siemens, DT 2 508 802, 1976.
139. Hitachi, JA 76 027 077, 1976.
140. Nippon Electric, JA 76 086 080, 1976.
141. Siemens, BE 841 241, 1976.
142. Osaka Titanium, JP 02 279 512, 1990 (M. Kuramoto, T. Nakai).
143. Cabot, BE 850 891, 1977.
144. IBM, US 3 974 003, 1976.
145. Siemens, DT 2 519 572, 1976.
146. UKR Agric Acad. ZAP, SU 660 968, 1980.
147. Hiratsuka Resin, JA 76 033 149, 1976.
148. Dow Corning, US 2 571 884, 1950.
149. BASF, FR 1 009 837, 1948.
150. Monsanto, GB 815 276, 1950.
151. Th. Goldschmidt, DT 685 728, 1936. Union Carbide Belge, BE 562 868, 1956.
152. Carborundum, US 1 271 713, 1918.
153. Stauffer Chem. Co., US 2 739 041, 1953.
154. Cabot, US 2 843 458, 1955.
155. E. Grunder, J. Elöd, *Z. Anorg. Allg. Chem.* **195** (1931) 269.
156. Konsort. für Elektrochem. Ind., GB 176 811, 1922.
157. Degussa, DT 1 079 015, 1960.
158. Stauffer Chem. Co., GB 865 939, 1959.
159. Stauffer Chem. Co., *Chem. Eng.* **69** (1962) 71.
160. L. V. Laine, SU 108 316, 1957.
161. Kali-Chemie, DT 901 412, 1954.
162. Horizons, GB 735 617, 1953.
163. Union Carbide, EP 26 860, 1979.
164. Mitsui Toatsu Chemicals, JP 59 195 519, 1984.
165. Mitsui Toatsu Chemicals, JP 59 184 720, 1984.
166. Wacker, DE 3 503 262, 1986 (R. Standigl, R. Griesshammer).
167. Int. Standard Electric, DAS 1 074 560, 1958.
168. Du Pont, US 2 820 698, 1954.
169. Bell Telephone Lab., US 2 812 235, 1955.
170. Westinghouse Electric, DT 1 088 473, 1957.
171. Western Electric, DT 1 029 811, 1956.
172. G. A. Wolff, US 2 877 097, 1958.
173. H. C. Theuerer, *J. Electrochem. Soc.* **107** (1960) 29.
174. G. G. Deryatykh, N. Kh. Agliulov, A. E. Wikolaev, E. M. Shcheplyagin, *Zh. Prikl. Khim. (Leningrad)* **53** (1980) no. 8, 1845 – 1848.
175. Denki Kagaku Kokyo, JP 02 196 014, 1990 (M. Fujimoto, M. Shinoyama, H. Matsumura).

176. Bell Telephone Lab., US 4 282 196, 1981 (T. Kometani, D. L. Wood).
177. General Electric, GB 309 393, 1928.
178. Union Chimique Belge, GB 1 042 076, 1951.
179. Faustel Metallurgie, US 2 665 997, 2 665 998, 1950.
180. E. Fitzer, DE 178 779, 1952.
181. Mitsui Toatsu Chemicals, JP 59 162 121, 1984.
182. Mitsui Toatsu Chemicals, JP 59 232 910, 1984.
183. Mitsui Toatsu Chemicals, JP 59 207 829, 1984.
184. Toa Gosei Chemical, DE 3 623 493, 1987 (M. Ito, T. Hattori, Y. Miwa).
185. Toa Gosei Chemical Industry, JP 63 021 211, 1988 (M. Ito, T. Hatsutori, Y. Miwa).
186. Mitsui Toatsu Chemicals, JP 61 205 614, 1986 (A. Hiai, K. Wakimura, T. Use).
187. Mitsui Toatsu Chemicals, JP 59 207 830, 1984.
188. Mitsubishi Metal Corp., JP 62 070 425, 1987 (H. Ikeda).
189. Mitsubishi Metal Corp., JP 63 089 414, 1988 (H. Ikeda, M. Tsunashima).
190. Mitsubishi Metal Corp., JP 62 036 014, 1987 (Y. Kuroda, K. Yamakawa).
191. Mitsubishi Metal Corp., JP 63 091 138, 1988 (H. Ikeda, M. Tsunashima, M. Sato).
192. Mitsubishi Metal Corp., EP 282 037, 1988 (H. Ikeda, M. Tsunashima).
193. Mitsubishi Metal Corp., EP 264 722, 1988 (H. Ikeda, M. Tsunashima, M. Sato).
194. Mitsubishi Metal Corp., JP 63 210 015, 1988 (H. Ikeda, M. Tsunashima, M. Sato).
195. Mitsubishi Metal Corp., JP 01 192 716, 1989 (K. Ogi, T. Kurashige, E. Kimura).
196. NEC Corp., JP 62 296 511, 1987 (A. Ishitani).
197. Dow Corning, FR 2 486 057, 1982 (J. H. Gaul, D. R. Weyenberg).
198. Toa Gosei Chemical, JP 61 295 372, 1986 (S. Motojima, T. Hattori, N. Ishikawa).
199. Toa Gosei Chemical, JP 63 277 599, 1988 (S. Motojima, T. Hattori, N. Ishikawa).
200. Mitsubishi Metal Corp., JP 01 301 507, 1989 (T. Hirai, H. Ikeda).
201. Mitsui Toatsu Chemicals, JP 62 143 839, 1987 (T. Iwao, K. Morii, Y. Tajima).
202. Mitsui Toatsu Chemicals, JP 62 143 840, 1987 (T. Iwao, K. Morii, Y. Tajima).
203. Mitsui Toatsu Chemicals, JP 61 227 914, 1986 (T. Iwao, K. Morii, Y. Tajima).
204. Wacker, DE 3 518 620, 1986 (R. Griesshammer, H. Herrmann, R. Standigl).
205. B. S. Suresh, D. K. Padma, *Polyhedron* **5** (1986) no. 10, 1579 – 1579.
206. A. Kornick, M. Binnewies, *Z. Anorg. Anal. Chem.* **587** (1990) 157 – 166.
207. Dynamit Nobel, DE 3 615 509, 1987 (B. Falk, K. Ruff, K. Schrage).
208. P. Voss, Ph.D. thesis, Aachen 1961.
209. P. Voss, Dissertation, Aachen 1962.
210. H. Bloching, Dissertation, Freie Universität Berlin 1961.
211. Du Pont, US 2 840 588, 1956.
212. Kemira Oy, FI 822 32, 1990 (V. P. Judin, A. Hayha, P. Koukkari).
213. Kemira Oy, US 5 077 027, 1991 (V. P. Judin, A. Hayha, P. Koukkari).
214. Semiconductor Energy Laboratory Co., JP 60 054 913, 1985.
215. M. Janai et al., *J. Appl. Phys.* **52** (1981) no. 5, 3622 –3624.
216. H. Matsumura, H. Ihara, T. Uesugi, *J. Appl. Phys.* **57** (1985) no. 12, 5483 – 5485.
217. H. Matsumura, T. Uesugi, H. Ihara, *Mater. Res. Soc. Symp. Proc.,* **49** (1985) 175 – 180.
218. État Français, FR 2 620 736, 1989 (Y. Pauleau, P. Lami, J. Pelletier).
219. Semiconductor Energy Laboratory Co., JP 60 144 940, 1985 (S. Yamazaki).
220. K. Zhang, Q. Yuan, *Huaxue Shijie* **23** (1982) no. 12, 353 – 354.
221. Schumacher, GB 2 028 289, 1980.
222. Ceres Corp., WP 8 001 489, 1980.
223. Schumacher, US 4 084 024, 1978.
224. Siemens, JA 77 035 605, 1977.
225. Western Electric, US 3 932 160, 1976.
226. Nippon Teleg. & Tel., JA 77 088 348, 1977.
227. Nippon Teleg. & Tel. and Furukawa Electric, JA 77 104 936, 1977.
228. Nippon Teleg. & Tel., JA 79 127 914, 1979.
229. Fuji Photo Film, JA 80 167 119, 1981.
230. Micron Technology Inc., US 5 082 524, 1992 (D. A. Cathey).
231. Schumacher, US 4 318 942, 1982 (L. M. Woerner, E. B. Moore).
232. Battelle Dev., US 3 990 953, 1976.
233. Exxon Res. & Eng., US 4 192 720, 1980.
234. Licentia Patentverwertung, DT 2 929 669, 1981.
235. Tokyo Shibaura Electric, JA 78 140 311, 1979.
236. UBE Industries, GB 2 020 264, 1979.
237. Asahi Glass, JA 79 134 098, 1979.
238. Asahi Glass, JA 79 160 600, 1980.
239. Rosenthal, DT 2 635 167, 1977.
240. B. Armas, C. Combescure, G. Male, M. Morales, *Less-Common Met.* **67** (1979) no. 2, 449 – 453.
241. Toyota Cent. Res. & Dev., JA 75 108 183, 1977.
242. Hitachi, GB 19 915, 1980.
243. Phillips Petroleum, BE 858 462, 1978.

244. Asahi Glass, JP 62 070 219, 1987 (K. Sato et al.).
245. K. Zhang et al., *Huaxue Shiji* **5** (1983) no. 4, 251 –253.
246. Dow Corning, EP 301 678, 1989 (K. G. Sharp, J. J. D'Errico).
247. H. Schmaelzer, E. Hengge, *J. Organomet. Chem.* **225** (1982) no. 1, 171 – 176.
248. M. Schmeisser, M. Schwarzmann, *Z. Naturforsch.* **B 116** (1956) 278.
249. Kali-Chemie, DT 955 414, 1954.
250. K. Friedrich, Thesis, RWTH-Aachen 1964.
251. Licentia Patentverwertung, DT 1 045 381, 1957.
252. Morrizons, GB 735 617, 1953.
253. Kali-Chemie, DT 901 412, 1954.
254. B. Rubin, G. H. Moates, J. R. Wenier, *J. Electrochem. Soc.* **104** (1957) 656.
255. Japan Telegram & Telephone, JA 13 669, 1988.
256. Texas Instruments, US 4 154 870, 1979.
257. Int. Telephone & Telegraphe, US 3 961 926, 1976.
258. H. Schäfer, *Chem.-Ztg.* **75** (1951) 48.
259. E. Zintl et al., *Z. Anorg. Chem.* **245** (1940) 1.
260. Goodrich, DT 1 079 608, 1956.
261. O. K. Botoinkin, SU 139 657, 1960.
262. Fumio Hori, JP 59 008 613, 1984.
263. J. M. Huber, EP 176 770, 1986 (D. E. Schramm).
264. L. Brewer, *Chem. Rev.* **52** (1953) 39.
265. W. Blitz, P. Ehrlich, *Naturwissenschaften* **26** (1938) 188.
266. G. Grube, H. Speidel, *Z. Elektrochem.* **53** (1949) 339.
267. H. v. Wartenberg, *Z. Elektrochem.* **53** (1949) 343.
268. H. Schäfer, R. Hörnle, *Z. Anorg. Allg. Chem.* **263** (1959) 261.
269. J. Berak, D. Grzeskowiak, *Zesz. Nauk. Politech. Slask. Vhem.* **88** (1979) 154 – 155.
270. Kemanord, RD 199 058, 1980.
271. A. Marcks, H. Hausner, *Therm. Anal. Proc. Int. Conf. 6th* **2** (1980) 223 – 228.
272. Hammatsu, NL 74 006 827, 1975.
273. Hitachi, DT 2 365 056, 1975.
274. Fujitsu, JA 79 018 689, 1979.
275. Stepanov, SU 423 203, 1975.
276. Hoya Glass Works, DT 2 443 718, 1975.
277. Unitika, JA 74 041 469, 1975.
278. Richter, DL 120 190, 1976.
279. Teijin Lens, JA 80 110 127, 1980.
280. BFG Glassgroup, BE 859 428, 1978.
281. AS Phys. Metall. Inst., SU 609 121, 1979.
282. Sony Corp., JA 79 031 721, 1979.
283. ITT Industries, FR 2 362 089, 1978.
284. Matsushita Elec. Ind., JA 79 024 866, 1979.
285. M. Schmitt, FR 2 369 103, 1978.
286. Matsushita Elec. Ind., JP 63 276 724, 1988 (K. Kimura et al.).
287. NHK Spring Co., JP 62 071 665, 1987 (T. Ebihara, T. Yokoo).
288. S. Yoshida, S. Misawa, *Denshi Gijutsu Sogo Kenkyusho Iho.,* **44** (1980) no. 1/2, 14 – 22.
289. United States Dept. of Energy, US 389 802, 1983 (J. V. Milewski).
290. Lonza, DT 2 531 481, 1976.
291. Osaka Titanium, JA 74 098 807, 1975.
292. Suzuki, DT 2 909 104, 1979.
293. Toyota Chuv Kenkyusho, JA 76 048 799, 1976.
294. Lucas J. Industries, GB 1 470 171, 1977.
295. N. Azuma, T. Yamada, S. Kuwabare, *Yogyo Kyokaishi* **86** (1978) no. 9, 418 – 424.
296. Ceraver, BE 858 704, 1978.
297. Bludov, SU 438 050, 1975.
298. Victor Co. of Japan, JA 75 038 200, 1976.
299. E. Zintl et al., *Z. Anorg. Allg. Chem.* **245** (1940) 1.
300. I.G. Farbenind., CH 215 137, 1939.
301. Duisberg, US 2 242 497, 1939.
302. Kawasaki Steel, JA 78 134 794, 1979.
303. Toyota Motor Co., DE 3 602 647, 1986 (S. Abe, M. Ogawa).
304. Dow Corning, US 5 051 248, 1991 (G. N. Bokerman, J. P. Canady, C. S. Kuivida).
305. O. N. Carlson, F. A. Schmidt, *Proc. Vac. Metall. Conf. Spec. Met. Melting Process* 1985, 129 – 136.
306. L. M Ericsson, SE 450 391, 1987 (T. S. Ganev, M. Jansson).
307. K. L. Day, B. Donn, *Science* **4365** (1978) 307 – 308.
308. B. S. Suresh, D. Krishnamurth Padma, *J. Chem. Soc. Dalton Trans.,* 1984, no. 8, 1779 – 1780.
309. V. Belot et al., *J. Non-Cryst. Solids* **127** (1991) no. 2, 207 – 214.
310. E. Wilberg, W. Simmler, *Z. Anorg. Allg. Chem.* **283** (1956) 401.
311. Hitachi, JP 59 074 637, 1984.
312. Asahi Glass, JP 55 058 230, 1980.
313. Asahi Glass, JP 55 145 739, 1980.
314. Toppan Printing Co., JP 02 122 924, 1990 (M. Sekiguchi et al.).
315. Y. Wang, C. Jiang, *Bandaoti Xvebao,* **7** (1986) no. 6, 596 – 601.
316. N. A. Takasaki, E. Ikawa, Y. Kurogi, *J. Vac. Sci. Technol. B* **4** (1986) no. 4, 806 – 811.
317. R. E. Stahlbush, W. E. Carlos, S. M. Prokes, *IEEE Trans. Nucl. Sci., NS-34* **6** (1987) part I, 1680 – 1685.

318. E. P. Bochkarev et al., *Vysokochist. Veshchestva* **1** (1989) 73 – 77.
319. Ei Shokusan K. K., JP 62 183 899, 1987 (F. Baba, K. Kozono, T. Hashiguchi, K. Maemura).
320. Kubota Ltd., JP 62 250 910, 1987 (Y. Okumura, Y. Okamoto, K. Yanai).
321. Unisearch Ltd., WO 8 906 222, 1989 (J. A. Bourdillon et al.).
322. I. N. Buyanovskii et al., *Trenie Iznos* **6** (1985) no. 1, 119 – 124.
323. A. Stock, F. Zeidler, *Ber. dtsch. chem. Ges.* **56** (1923) 986.
324. R. Müller, *Chem. Techn.* **2** (1950) 1.
325. G. H. Wagner, A. N. Pines, *Ind. Eng. Chem.* **44** (1952) 321.
326. R. Müller, R. Kohne, S. Sliwinski, *J. prakt. Chem.* **9** (1959) 71.
327. K. Larssow, *Ak. Kemi* **16** (1960) 215.
328. C. L. Frye, W. T. Collins, *J. Am. Chem. Soc.* **92** (1970) 5586.
329. H. Bürgi, G. Calzaferri, *J. Chromatography* **507** (1990) 481.
330. P. A. Agaskar, *Inorg. Chem.* **30** (1991) 2707.
331. G. Calzaferri, *Nachr. Chem. Tech. Lab.* **40** (1992) 1106.
332. R. Gooden, J. W. Mitchell, *J. Electrochem. Soc.* **129** no. 7, 1619.
333. R. J. P. Corrir, *Organometallics* **10** (1991) 3574.
334. M. Laine, *Nature* **353** (1991) 642.
335. W. Humpel, V. Haasy, *Z. Anorg. Allg. Chem.* **23** (1900) 32.
336. M. Schmeisser, H. Müller, W. Burgermeister, *Angew. Chem.* **69** (1957) 781.
337. L. E. Gorsh, G. N. Dolenko, ZZV, *Akad. Nauk. SSSR Neorg. Mater.* **15** (1979) no. 9, 1521 – 1523.
338. Kali-Chemie, DT 1 008 265, 1955, 1 029 810, 1957.
339. Mitsubishi Electric, JA 77 070 498, 1977.
340. Toyota Motor Corp., JP 63 079 716, 1988 (S. Abe, O. Kazuaki, M. Ogawa).
341. N. V. Tolstoguzov, *Izv. Vyssh. Uchebn. Zaved., Chern. Metall.* 1991, no. 4, 22 – 25.
342. Asahi Chemical Industry, JP 62 252 310, 1987 (Y. Yamada, N. Tamura, K. Makita).
343. Rockwell International Corp., US 4 552 740, 1985 (P. E. D. Morgan, E. A. Pugar).
344. Asahi Chemical Industry, JP 62 252 321, 1987 (Y. Yamada, N. Tamura, K. Makita).
345. R. Creus et al., *Mater. Sci. Eng. B* **B 3** (1989) no. 1/2, 109 – 112.
346. J. H. Kennedy, Z. Zhang, *J. Electrochem. Soc.* **136** (1989) no. 9, 2441 – 2443.
347. B. T. Ahn, R. A. Huggins, *Mater. Res. Bull.* **25** (1990) no. 3, 381 – 389.
348. S. Sahami, S. W. Shea, J. H. Kennedy, *J. Electrochem. Soc.* **132** (1985) no. 4, 985 – 986.
349. J. H. Kennedy et al., *Solid State Ionics* **18 – 19** (1986) no. 1, 368 – 371.
350. Matsushita Electric Industrial Co., JP 04 133 209, 1992 (K. Yamamura, K. Takada, S. Kondo).
351. J. H. Kennedy, Y. Yang, *J. Electrochem. Soc.* **133** (1986) no. 11, 2437 – 2438.
352. A. A. Godovikov, R. G. Kuryaeva, *Fiz. Khim. Skekla* **9** (1983) no. 4, 502 – 509.
353. P. Barlier, J. P. Torre, *Ann. Chim. (Paris)* **5** (1989) no. 8, 701 – 709.
354. N. J. Shaw, *Sci. Tech. Aerosp. Rep.* **20** (1982) no. 6, 113.
355. J. F. Lartigue, M. Ducarroir, B. Armas, *J. Mater. Sci.* **19** (1984) no. 9, 3079 – 3089.
356. P. M. Jeffers, S. H. Bauer, *J. Non-Cryst. Solids* **57** (1983) no. 1, 189 – 193.
357. Toa Gosei Chemical, JP 02 034 598, 1990 (S. Motojima, T. Hattori, N. Ishikawa).
358. Toa Gosei Chemical, JP 02 034 599, 1990 (S. Motojima, T. Hattori, N. Ishikawa).
359. Kemira Oy, DE 3 939 048, 1990 (V. P. Judin, A. Hayha, P. Koukkari).
360. P. E. D. Morgan, E. A. Pugar, *J. Am. Chem. Soc.* **68** (1985) no. 12, 699 – 703.
361. Asahi Chemical Industry, JP 62 252 304, 1987 (Y. Yamada, N. Tamura, K. Makita).
362. Asahi Chemical Industry, JP 62 252 303, 1987 (Y. Yamada, N. Tamura, K. Makita).
363. V. D. Pershin et al., *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **24** (1981) no. 11, 1323 – 1326.
364. Unión Explosives Río Tinto, ES 2 006 119, 1989 (F. Gómez Aleixandre et al.).
365. Y. Chang, R. M. Young, E. Pfender, *Plasma Chem. Plasma Processes* **7** (1987) no. 3, 299 – 316.
366. H. Moissan, A. Stock, *Ber. Dtsch. Chem. Ges.* **33** (1990) 2125.
367. Allis-Chalmers, US 3 427 131, 1969.
368. Texaco Dev., US 4 038 292, 1977.
369. Aktiebolag Kauthal, SW 201 850, 1966.

Silicon Compounds, Organic

LUTZ RÖSCH, Wacker Chemie GmbH, Burghausen, Federal Republic of Germany

PETER JOHN, Wacker Chemie GmbH, Burghausen, Federal Republic of Germany

RUDOLF REITMEIER, Wacker Chemie GmbH, Burghausen, Federal Republic of Germany

1. Introduction 2
2. Fundamental Synthetic Routes for Organosilicon Compounds 4
2.1. Direct Synthesis of Organohalosilanes 4
2.2. Grignard Synthesis 7
2.3. Syntheses with Alkali Metals 8
2.4. Addition Reactions (Hydrosilylation) 8
2.5. Substitution Reactions of Si–H Bonds 10
2.6. Coproportionation and Disproportionation 10
3. Silicon-Functional Organosilicon Compounds 11
3.1. Halo- and Pseudohalosilanes 12
3.2. Alkoxy- and Aryloxysilanes 12
3.3. Acyloxysilanes 16
3.4. Oximino- and Aminoxysilanes . . . 16
3.5. Aminosilanes, Amidosilanes, and Silazanes 16
3.6. Hydrogensilanes 17
4. Organofunctional Organosilicon Compounds 18
4.1. Vinyl and Other Alkenyl Compounds 18
4.2. Organohalogenated Compounds . . 19
4.3. Nitrogen-Containing Compounds . 21
4.3.1. Cyanoalkyl Compounds 21
4.3.2. Organic Amino Compounds 21
4.3.3. Other Nitrogen Compounds 22
4.4. Organosulfur Compounds 23
4.4.1. Mercapto and Sulfidic Organofunctions 23
4.4.2. Compounds Containing Sulfur–Oxygen Groups 23
4.5. Oxygen-Containing Compounds . . 24
4.5.1. Epoxy and Other Oxy Compounds . . 24
4.5.2. Acrylates and Other Ester Functions . 24
4.5.3. Acid Anhydrides and Other Carboxy Groups 25
4.6. Other Organofunctions 26
5. Other Organosilanes 26
5.1. Tetraorganosilanes 26
5.2. Disilanes 27
5.3. Polysilanes 27
6. Uses 27
6.1. Silylating Agents 27
6.2. Physiologically Active Organosilicon Compounds 28
6.3. Silanes for Silicone Modification . . 29
6.4. Silanes for Modification of Organic Polymers 29
6.5. Silanes in Catalysis 31
6.6. Silanes for Use in Liquid Media . . 32
6.7. Silane Coupling Agents 32
6.8. Silanes for Hydrophobic Surface Treatment 34
6.9. Analysis 34
7. Toxicology and Environmental Aspects 34
8. Economic Aspects 35
9. References 35

1. Introduction

Organosilicon compounds are substances that contain at least one organic group directly bonded to silicon via a silicon–carbon bond. They can be subdivided into two main groups: monomeric compounds and polymers. Of the latter, the polymeric silicones, in which organosilicon units are linked by Si–O–Si bonds, are economically the most important. They are obtained from monomeric organohalosilanes by hydrolysis or methanolysis followed by polycondensation. Silicone production is thus the largest consumer of monomeric silanes. Polysilanes – compounds with Si–Si bonds – are increasing in importance. The monomeric organosilicon compounds are important components in a wide range of other applications apart from silicone chemistry. Their uses as auxiliary materials and as protecting groups in organic syntheses, as

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_021

catalyst components in olefin polymerization, as protective agents for buildings, and as precursors for silicon carbide and silicon nitride coatings by chemical vapor deposition are important. An area with great potential is that of biologically active silicon compounds. Some organosilicon-based pharmaceuticals and crop protection agents are already on the market. Further developments in this area include silylated fragrance materials.

History. The first organosilicon compound, tetraethylsilane, was obtained in 1863 by FRIEDEL and CRAFTS by treatment of diethylzinc with tetrachlorosilane (Eq. 1) [15]. Two years later they prepared tetramethylsilane from dimethylzinc by an analogous method [16]. Subsequently, a series of new organosilicon compounds was prepared by treating tetrachlorosilane or tetraethoxysilane with organometallic reagents (Eq. 2). The work of PAPE, LADENBURG, and KIPPING was associated with this [17, 18]. The introduction of Grignard reagents into organosilicon chemistry by DILTHEY and KIPPING in the early 1900s was an important advance (Eq. 3) [19].

$$2\,Et_2Zn + SiCl_4 \longrightarrow Et_4Si + 2\,ZnCl_2 \quad (1)$$

$$Si(OEt)_4 + Et_2Zn \xrightarrow{Na} Et_xSi(OEt)_{4-x} \quad (2)$$

$$SiCl_4 + EtMgI \longrightarrow Et_xSiCl_{4-x} \quad (3)$$

Despite much research, organosilicon compounds remained a laboratory curiosity for a long time, and in 1936, KIPPING – in a summary of his investigations of nearly 40 years – still maintained that for this class of compound, important applications were neither known nor to be expected [18]. At the same time, however, Corning Glass and General Electric in the United States had already recognized that insulating materials resistant to high temperature could be produced from these compounds [20]. With the development of the direct synthesis – the production of methylchlorosilanes by the reaction of silicon with chloromethane in the presence of a catalyst – by ROCHOW [21] in the United States and MÜLLER in Germany [22], the basis for economical silicone production was finally established. Another milestone was the discovery in 1947 of the hydrosilylation reaction, in particular the oxidative addition of Si–H groups to carbon – carbon multiple bonds [23]. The discovery that this reaction can be controlled by using transition-metal catalysts [24] allowed economical access to a large number of new organosilicon compounds and their use in silicone chemistry and other areas.

Nomenclature. In nearly all organosilicon compounds, particularly the industrially important ones, the silicon atom is tetravalent. Most organosilicon compounds therefore contain tetracoordinate silicon. Compounds with higher coordination numbers occur as reactive intermediates in many reactions, while compounds with multiple bonds to silicon exhibit lower coordination numbers. The group of biologically active silatranes [25] with pentacoordinate silicon and the numerous investigations into hypervalent organosilicon compounds [26], whose chemistry may achieve industrial importance in the future, are also worth mentioning. In principle, organosilicon compounds can be regarded as derivatives of silane (SiH_4) in which the hydrogens have been replaced completely or partially by hydrocarbon groups:

$$SiH_4 \xrightarrow[-nH]{+nR} R_nSiH_{4-n} \quad (4)$$

The remaining hydrogens can also be replaced by functional groups X. The resulting compounds are known as mono-, di-, or trifunctional organosilicon compounds, depending on the number of groups present in the molecule.

$$R_nSiH_{4-n} \xrightarrow[-mH]{+mX} R_nSiH_{4-n-m}X_m \quad (5)$$

Almost all monomeric organosilicon compounds that contain tetracoordinate silicon can be derived from this general formula.

For a more precise classification the R groups are divided into "organofunctional" and "simple" organic groups. The latter are generally alkyl or aryl groups to which, however, a certain degree of functionality must be assigned, since it is possible to exchange alkyl groups for other alkyl or aryl groups, hydrogen, or other substituents in the presence of catalytically active Lewis acids such as $AlCl_3$. Alkyl groups can also be converted into organofunctional groups: e.g., chloromethylated silanes are formed in the

chlorination of methylsilanes under UV irradiation. Examples:

Simple organic groups:	$-CH_3$, $-C_2H_5$, $-C_6H_5$
Organofunctional groups:	$-CH=CH_2$, $-C\equiv CH$, $-CH_2-CH=CH_2$, $-CH_2Cl$, $-\underset{\underset{O}{\|}}{C}-R$, $-CH_2CH_2NR_2$, $-CH_2-\underset{\underset{O}{\diagup\diagdown}}{CH-CH_2}$
Silicon-functional groups:	halogens, $-NH_2$, $-NR_2$, $-S-C\equiv N$, $-OR$, $-OH$ (silanols), $-OM$ (metal silanolates)

The special IUPAC nomenclature rules for organosilicon compounds have not yet been issued. According to general IUPAC rules, tetravalent organosilicon compounds are called silanes and the ligands should be listed in alphabetical order, but hydrogen should not be named. The following order, which deviates from this, has become established in the technical literature: organofunctional ligand, simple organic ligand, alkoxy groups, halogen.

Compounds in which silicon is incorporated into a ring are known as silacyclo compounds (e.g., silacyclobutane). Silenes contain a Si–C double bond. Silicon chains and rings are named according to the number of silicon atoms: disilane, trisilane, cyclohexasilane. Certain functional groups have their own names: ≡Si–OH = silanol, ≡Si–SH = silanethiol, ≡Si–O–Si≡ = siloxane. Some examples of nomenclature are:

Chemical formula	Name
Me_4Si	tetramethylsilane
$Me(H)SiCl_2$	dichloromethylsilane
Ph_2SiH_2	diphenylsilane
$H_2NCH_2CH_2(Me)Si(OMe)_2$	2-aminoethyldimethoxymethylsilane
Me_3SiOH	trimethylsilanol
$(Me_2SiO)_3$	hexamethylcyclotrisiloxane
$Me_2(H)SiSiH_2Me$	1,1,2-trimethyldisilane
$(Me_2Si)_6$	dodecamethylcyclohexasilane

Structure and Reactivity. Like carbon, silicon preferentially assumes a tetrahedral arrangement of its four valences in forming compounds. The larger atomic radius of silicon compared to carbon (C, 77 pm; Si, 177 pm) leads, for example, to lower barriers to rotation at silicon–element bonds than in the corresponding carbon compounds. Also, the completely different bond energies of silicon and carbon (for examples, see Table 1) and the different electronegativities of the two elements (Pauling electronegativities: Si = 1.8, C = 2.5) result in the organic chemistry of silicon being fundamentally different from that of carbon. The high affinity of silicon for oxygen is the reason why only compounds with Si–O bonds (SiO_2, silicates) occur naturally. Although addition reactions to unsaturated systems are important in synthetic organic chemistry, in silicon chemistry this type of reaction is primarily of academic interest because, thus far, species with multiple bonds to silicon have been difficult to prepare. If the problem of synthesizing compounds with Si–C or Si–Si double bonds economically is solved, the door could be opened to a new sphere of industrial organosilicon chemistry. In organosilicon chemistry, S_N2 nucleophilic substitution at silicon plays a much more important role than in the case of carbon. Depending on the nucleophile, an Si–C bond can be formed or broken. It is particularly likely to be broken if a functional group is present in the β-position (β-effect). Nucleophilic substitution always occurs particularly readily if the bond strength of the new silicon–substituent bond is significantly higher and no steric hindrance occurs. Because of this, alkylhalosilanes react with water, for example, with the loss of hydrogen halide to form silanols in the first step (Eq. 6). The latter can then react further under alkaline (nucleophilic attack of OH^- at the silicon atom) or acid (electrophilic attack of H_3O^+ at the oxygen atom) conditions to give oligo- or polysiloxanes (see → Silicones, Chap. 2.1.1.).

Table 1. Examples of bond energies for silicon and carbon [27]

Compound	Bond energy, kJ/mol	
	X = Si	X = C
H_3X-H	378	439
H_3X-CH_3	343	378
H_3X-OH	452	386
H_3X-F	565	460
H_3X-Cl	381	354
H_3X-SiH_3	310	343

$$R_{4-n}\,SiX_n + n\,H_2O \longrightarrow R_{4-n}\,Si(OH)_n + n\,HX \quad (6)$$

The conversion of Si–X into Si–O is facilitated by the ability of the silicon atom to form pentacoordinate transition states.

The high reactivity of Si–X bonds (X = halogen, N, S), together with the high chemical and thermal stability of bonds between silicon and simple organic groups, is characteristic of organosilicon compounds. These properties have led to the broad range of applications of this class of compounds.

Because of the very large number of known organosilicon compounds, the most important fundamental synthesis routes for organosilanes are discussed first. Following this the classes of organosilicon compounds that are particularly important industrially and economically are discussed.

2. Fundamental Synthetic Routes for Organosilicon Compounds

2.1. Direct Synthesis of Organohalosilanes

The process described independently by Rochow [21] and Müller [22], involving the direct reaction of haloalkanes with elemental silicon in the presence of a copper catalyst, is currently the standard synthesis used by all major silicone producers.

$$\begin{array}{l} Si + RX \xrightarrow{250-320\,^{\circ}C} R_3SiX + R_2SiX_2 + RSiX_3 + \ldots \\ R = CH_3: Cat. = Cu \\ R = C_6H_5: Cat. = Ag \end{array} \qquad X = F, Cl, etc. \quad (7)$$

Although the process is, in principle, applicable to any organic halide, it has been successful commercially only in the production of methylchloro- and phenylchlorosilanes. Longer- chain haloalkanes and unsaturated compounds such as vinyl chloride or allyl chloride generally give low yields of organosilanes, consisting of mixtures that are difficult to separate. Similarly poor results are obtained when polychlorinated compounds such as α,ω-dichloroalkanes or dichloromethane are used [28–30]. For this reason, only the reaction of silicon with chloromethane is described below.

Mechanism. Although the mechanism of the direct process has been investigated intensively since its discovery, it has not yet been fully elucidated. A detailed review of older work on this subject can be found in [31], while more recent studies are described in [32]. The reaction of chloromethane and silicon in the presence of copper is a heterogeneously catalyzed gas–solid reaction. In the first step the copper catalyst and silicon react to form Cu_3Si. Whether this η-phase is the actual reactive species has been disputed, but it does guarantee that silicon and copper are in close proximity. This permits the formation of active Si–Cl groups on the surface via the chlorination of silicon by CuCl. The latter is formed by the decomposition of a copper–chloromethane adduct, which also yields hydrogen, methane (and other hydrocarbons), and carbon as side products. Since the Si–Cl species are formed next to the copper–chloromethane adducts, the transfer of a methyl group with formation of a surface-bound Cl–Si–Me transition state is probable. Further chloromethane transfer finally leads to Me_2SiCl_2, which is liberated. CuCl remains behind at the catalyst surface and can initiate a new cycle. The chain reaction can be terminated by combination of two surface-bound Si–Cl groups or by reaction of Si–Cl with CuCl. The resulting $SiCl_2$ groups give $MeSiCl_3$ by further reaction with a Cu–MeCl adduct. The formation of Me_3SiCl and other side products can be explained similarly. Methyl radicals or silylenes in the gas phase [33] are not important for propagation of the reaction.

Promoters. If copper alone is used as catalyst in the direct synthesis a slow reaction occurs with low degrees of silicon conversion and unsatisfactory selectivity for the target product dichlorodimethylsilane. By using metallic promoters, both the activity of the catalyst mixture and the selectivity of the synthesis can be improved. The addition of antimony, cadmium, aluminum, zinc, tin, or combinations thereof has been found to be effective. For example, a catalyst mixture with 94.49 wt % Si, 5 wt % Cu, 0.5 wt % Zn, and 0.01 wt % Sn gave a dichlorodimethylsilane selectivity of 80 mol % [34]. A selectivity of 90 % has been reported with a similar catalyst system [35]. The promoter concentration is critical; for example, too high a tin content inhibits reaction. The action of promoters can be explained as follows: they facilitate the transfer of chlorine and methyl groups to silicon and the diffusion of silicon into copper. Further

information is given in [34] and references cited therein.

Nonmetals such as phosphorus can also be used as promoters [36].

Product Distribution. The direct process never follows exactly the "ideal equation," according to which only dichlorodimethylsilane should be formed. Many other products are formed in varying quantities in temperature-, catalyst-, or raw material-dependent side reactions. Besides methyltrichlorosilane and chlorotrimethylsilane, incompletely substituted silanes, such as dichloromethylsilane, are also formed. Furthermore, the crude silane mixture contains saturated and unsaturated hydrocarbons and 3–8 % disilanes of general formula $Me_nSi_2Cl_{6-n}$. Table 2 lists the distribution of the main monomeric silicon-containing products and their boiling points. The compounds Me_4Si [*75-76-3*], $HSiCl_3$ [*10025-78-2*], $SiCl_4$ [*10026-04-7*], and others are also formed in small amounts. All the organosilicon compounds formed can be used in silicone chemistry or other fields, the disilanes generally after catalytic or thermal cleavage to monosilanes. The product ratio causes problems in certain cases. The target is to obtain a crude silane mixture in which the individual silanes are obtained as closely as possible to the ratio in which they will be used in the production of functional silanes and silicones, by means of suitable reaction conditions. Fortunately, dichlorodimethylsilane, which is particularly important, is formed in ca. 80 % yield. However, with chlorotrimethylsilane, chlorodimethylsilane, and dichloromethylsilane, shortages sometimes occur. These can be counteracted by altering the product ratio in the direct synthesis via addition of gases or metals.

Table 2. Main monomeric silicon-containing products of direct chloromethylsilane synthesis

Compound	*bp*, °C	Proportion in the crude silane, wt %
Me_2SiCl_2 [*75-78-5*]	70	70–90
$MeSiCl_3$ [*75-79-6*]	66	5–15
Me_3SiCl [*75-77-4*]	57	2–4
$MeHSiCl_2$ [*75-54-7*]	41	1–4
Me_2HSiCl [*1066-35-9*]	35	0.1–0.5

Zinc, magnesium, or aluminum in quantities up to 10 % or more has a strong alkylating effect via methylmetal halides formed as intermediates, so that the content of chlorotrimethylsilane in the crude silane can increase to 5–20 %. The metals ultimately act as halogen scavengers, forming the halides $ZnCl_2$, $MgCl_2$, or $AlCl_3$ [37, 38].

Calcium silicide (10–80 % Ca) reacts with chloromethane in the presence of zinc and copper as catalysts to give a mixture with a high proportion of more highly alkylated silanes (ca. 61 % Me_3SiCl and ca. 8 % Me_4Si) [39].

The proportion of hydrogen-containing organochlorosilanes can be increased considerably by adding H_2 [40] or HCl [41]. A $MeHSiCl_2 + Me_2HSiCl$ selectivity of > 80 mol % has been reported [42]. If silicon is reacted with HCl alone, $HSiCl_3$ and $SiCl_4$ are obtained.

Raw Materials. Formerly, ca. 97 % pure silicon was used. After the significant effect of even small quantities of added metals on the selectivity of the reaction and the catalyst activity was recognized, a purity of > 99 % is now aimed for. The impurities are generally aluminum, calcium, iron, and titanium. The presence of certain elements (e.g., Pb), which act as inhibitors, is particularly critical. Even 50 ppm of lead in the catalyst mixture halves the silicon conversion, and the Me_2SiCl_2 selectivity decreases considerably. The reactivity of silicon depends on its morphology, which can be affected by the conditions under which liquid silicon is cooled in the production process.

The purity of other raw materials should also be as high as possible. Copper, produced by cementation or electrolysis, must therefore be subjected to further purification. Partially oxidized copper is used, whose average composition is approximately Cu_2O. The use of a Cu–Si alloy or other copper compounds (e.g., CuCl or copper oxalate) is also possible [43]. Promoter elements such as zinc are used either as the metal, the oxide, or the chloride. Most producers obtain chloromethane by treating methanol with hydrogen chloride. In this way the hydrogen chloride formed in the conversion of organochlorosilanes to silicones can be recycled. If methanolysis is carried out instead of hydrolysis (Eq. 8), chloromethane is formed directly and can be recycled into the reaction after purification.

$$2 \equiv SiCl + 2\,CH_3OH \longrightarrow \equiv Si-O-Si\equiv + 2\,CH_3Cl + H_2O \quad (8)$$

Production and Composition of Catalyst Mixture. Silicon and the copper catalyst are either finely ground separately and then mixed with other catalyst components (promoters) or used in the form of a ground copper – silicon alloy. A particularly active catalyst mixture is obtained by atomizing silicon or Cu–Si alloys, which may already contain promoters [44]. During grinding the particle size must be adjusted so that homogeneous fluidization is obtained in the fluidized-bed reactor. For silicon the particle sizes are 30 – 350 µm, and for the copper catalyst, 5 – 30 µm. The catalyst mixture can be activated before use by tempering at 1000 °C in a stream of N_2 or preferably H_2. The direct synthesis, however, also activates the catalyst mixture. Usual catalyst mixtures contain 2 – 6 % copper and 0.05 – 0.5 % zinc. Other promoters such as antimony or tin can also be added in small amounts [35, 45].

Industrial Process. The catalyst mixture is introduced into a fluidized-bed reactor and fluidized with a strong, tangentially introduced stream of chloromethane at ca. 300 °C and 2 – 5 bar (Fig. 1). The gas phase is partially freed from solid components in a cyclone at the head of the reactor, and separated solids are returned to the reactor. The fine portion passing through the cyclone enters the crude silane after condensation of the gas stream and is isolated during workup. Part of it can then be reprocessed. Condensation gives gaseous chloromethane and a crude liquid silane mixture. After purification, chloromethane is recycled to the reactor, and the crude silane mixture is distilled. Since silicon is consumed in the reaction and some of the catalyst is entrained in the gas stream, new catalyst mixture must be added continually. Its composition in the reactor must be kept as constant as possible. Silicon conversions of 90 – 98 % are achieved, and chloromethane conversions of 30 – 90 %, depending on process conditions (pressure, reactivity of the catalyst mixture, etc.). The strongly exothermic direct synthesis requires precise and effective temperature control because high time – space yields and high selectivity for dichlorodimethylsilane are possible only at ca. 300 °C. This is achieved by several measures, which generally supplement each other. Heat is removed by a reactor cooling jacket and sometimes by a heat exchanger in the reactor. The reaction can be slowed by pumping in an inert gas so that less heat is liberated. Thorough mixing of the catalyst mixture in the CH_3Cl stream is particularly important.

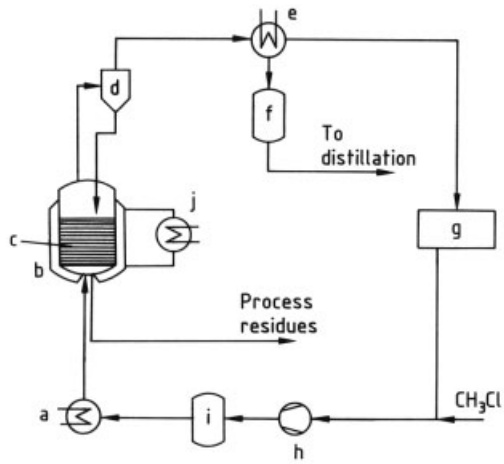

Figure 1. Production of chloromethylsilanes
a) Preheater; b) Reactor; c) Catalyst mixture; d) Cyclone; e) Heat exchanger; f) Crude silane container; g) MeCl purification; h) Compressor; i) MeCl container; j) Heating/cooling system

Distillation of Crude Silane. The crude silane mixture (for composition, see page 5) is separated into individual silanes in columns with very high separating capacity, connected in series. Since the organic chlorosilanes are building blocks for silicones, they must fulfill high purity requirements. Even a small amount of tri- or monofunctional silanes in the dichlorodimethylsilane can bring about chain branching ($MeSiCl_3$) or chain termination (Me_3SiCl). The high purity required for Me_2SiCl_2 is achieved by distillation of the Me_2SiCl_2 obtained in the fractional distillation of the crude silane mixture.

Corrosion Problems. Steel is the preferred construction material for the direct process plant. Because methylchlorosilanes, particularly CH_3Cl, react with zinc, tin, magnesium, and aluminum at elevated temperature, these metals

must not be used as alloy components. In some parts of the plant, enamel or Teflon must also be used.

Direct Synthesis with Halogen-Free Compounds. Many experiments have been described on the application of direct synthesis to halogen-free compounds. None of these routes have become important economically. A summary can be found in [46]. Treatment of silicon with dimethyl ether would be particularly interesting because it should give halogen-free methoxymethylsilanes. Such compounds could also serve as starting materials for the production of silicones. This would avoid the corrosion problems caused by HCl in the methods used up till now. Although a reaction between silicon and dimethyl ether can be achieved [47], the drastic reaction conditions required and the low yields of useful products make this route unlikely to become an economical alternative to conventional direct synthesis.

2.2. Grignard Synthesis

The Grignard synthesis of organosilanes, in particular organohalosilanes, was introduced by DILTHEY and KIPPING [19]. It still attracts much interest because of its universal preparative possibilities. For the synthesis of special silanes by the introduction of certain saturated or unsaturated hydrocarbon groups, this is frequently the best synthetic route.

Silanes with halogen or alkoxy groups can be used as starting materials. For example, $SiCl_4$ and $HSiCl_3$ can be treated with methyl Grignard reagents to give the silanes known from the direct synthesis (e.g., Eq. 9):

$$HSiCl_3 + RMgX \longrightarrow MgXCl + RSiHCl_2 + R_2SiHCl + R_3SiH \quad (9)$$

Mono-, di-, and trichlorosilanes can act as starting materials for mixed alkyl–arylchlorosilanes:

$$RSiCl_3 + ArMgX \longrightarrow MgXCl + RSiArCl_2 + RSiAr_2Cl + RSiAr_3 \quad (10)$$

Ar = Aryl

Ethers are the most frequently used solvents in Grignard synthesis (diethyl ether, dibutyl ether, THF, ethylene glycol ethers). The reactivity is particularly high in THF. In this solvent, vinyl groups have been introduced for the first time by using vinylmagnesium bromide [48], and Si–H groups can be substituted with alkyl groups [49].

The first step in Grignard synthesis is preparation of the Grignard solution. In some cases it can be prepared in situ in the presence of the halo- or alkoxysilane. This route also offers the possibility of a continuous Grignard synthesis [50]. The nature of the halogen or alkoxy group bonded to silicon, the number and nature of the organic groups already present in the starting silicon compound, and the size of the group to be introduced are some important factors for the result of the Grignard synthesis. Steric hindrance can make reactions proceed sluggishly or not at all, but it can also allow the formation of a desired, partially substituted product in high yield. Without steric hindrance the reaction is generally complete only after all halogens have been replaced. If fluorosilanes are used, particularly high degrees of substitution with bulky organic groups can be achieved because of the small size of the fluorine atom.

In some cases the reaction can be controlled by addition of catalysts. The reaction of dichlorodiphenylsilane [*80-10-4*] with *tert*-butyl Grignard in THF in the presence of CuCN or CuSCN gives *tert*-butylchlorodiphenylsilane [*58479-61-1*] in 75 % yield (Eq. 11); without catalyst, no reaction occurs [51].

$$tert\text{-}BuMgX + Ph_2SiCl_2 \xrightarrow{\text{Cat.}} MgXCl + tert\text{-}BuPh_2SiCl \quad (11)$$

Cat.: e.g., CuCN, CuSCN, Me_3SiCN

Even in the absence of steric hindrance, the yield of partially substituted intermediates can be increased by using an appropriate stoichiometry of starting materials. Generally, however, separation of the resulting mixtures is uneconomical.

Silicon-bonded alkoxy groups react with Grignard reagents more slowly than halogens; the reactions are thus easier to control. For example, diisobutyldimethoxysilane can be synthesized from tetramethoxysilane and isobutyl Grignard in 95 % yield and > 97 % purity [52].

Disadvantages of Grignard syntheses are the high consumption of anhydrous, mostly volatile,

and readily flammable solvents and the formation of large amounts of salts. They are also often accompanied by side reactions. Because of this they are used industrially only in exceptional cases (e.g., for the production of organosilicon compounds of particularly high value).

2.3. Syntheses with Alkali Metals

The reaction of halosilanes with organic compounds of alkali metals or in situ with the reaction products of haloalkanes and alkali metals is an alternative or a supplement to the Grignard synthesis and is also usually carried out only on a laboratory scale:

$$RLi + ClSi{\equiv} \longrightarrow RSi{\equiv} + LiCl \quad (12)$$

$$RCl + ClSi{\equiv} + 2\,Na \longrightarrow RSi{\equiv} + 2\,NaCl \quad (13)$$

Because of the high reactivity of organometallic compounds, they can also attack solvents. However, their advantage is that highly sterically demanding organic groups can often be introduced when this is not possible with a Grignard reagent [53].

With the smaller methyl or ethyl groups, all halogens are generally replaced. Even Si−OR and Si−H groups can react with organolithium compounds [54, 55].

Organozinc and organoaluminum compounds can also be used for alkylations [56].

2.4. Addition Reactions (Hydrosilylation)

Silanes and siloxanes that contain Si-bonded hydrogens can add to compounds with C−C multiple bonds (Eqs. 14 and 15). On addition to alkynes, alkene derivatives are formed initially, to which a second molecule of the Si−H compound can be added, provided no steric hindrance occurs.

$$\equiv SiH + R_2C{=}CR_2 \longrightarrow \equiv SiCR_2CR_2H \quad (14)$$

$$\equiv SiH + RC{\equiv}CR \longrightarrow \equiv SiCR{=}CRH \quad (15)$$

Hydrosilylation is used industrially for synthesizing alkylsilanes and functional silanes, cross-linking silicone polymers, binding silicone polymers to organic polymers, and removing impurities that contain Si−H groups from silane fractions of the direct synthesis.

Two different hydrosilylation routes are possible. Both make use of the high enthalpy of formation, estimated to be 160 kJ/mol for the addition of an Si−H group to a simple alkene, and the fact that the Si−H bond is readily activated. The first method developed is based on a chain mechanism involving silyl radicals [23]:

$$HSiCl_3 \longrightarrow H^{\cdot} + {}^{\cdot}SiCl_3 \quad (16)$$

$${}^{\cdot}SiCl_3 + CH_2{=}CHCH_3 \longrightarrow Cl_3SiCH_2\dot{C}HCH_3 \quad (17)$$

$$Cl_3SiCH_2\dot{C}HCH_3 + HSiCl_3 \longrightarrow Cl_3SiCH_2CH_2CH_3 + {}^{\cdot}SiCl_3 \quad (18)$$

Silyl radicals can be generated thermally, by decomposition of radical initiators (e.g., acyl peroxides, azonitriles), by UV light, or by γ-radiation. With this method a large number of side products can be formed by decomposition of the initiators and polymerization of the alkene components. It is used industrially only in exceptional cases.

The route used almost exclusively in industry is the transition-metal-catalyzed addition of Si−H groups to multiple-bond systems.

In 1953, some supported transition metals were found to be very effective hydrosilylation catalysts. In particular, Pt – active charcoal is an highly active catalyst for the reaction of $HSiCl_3$ with acetylene, ethylene, butadiene, allyl chloride and $H_2C{=}CF_2$ at ca. 130 °C [57]. In 1957, Speier reported that chloroplatinic acid, $H_2PtCl_6 \cdot 6\,H_2O$, and its salts are catalysts that allow very rapid addition of a wide range of silanes to various alkenes, sometimes even below room temperature. This provided the possibility of hydrosilylating complex organic molecules. A summary of this work can be found in [58].

Mechanism. In the transition-metal-catalyzed addition reaction the ability of the transition metal to change its coordination number with a simultaneous change in oxidation state is exploited. This is shown in the example of platinum, where the reaction takes place at particularly active atoms on the metal surface or in the coordination sphere of the dissolved

complex (Eq. 19). The Si—H bond adds oxidatively to platinum, and a π-bond to the olefin is formed.

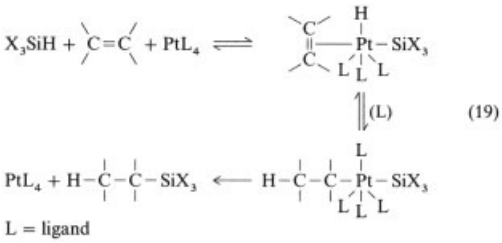

(19)

Both steps are reversible. The reaction is completed by addition of hydrogen to the π-bonded molecule, which is also reversible. This is followed by reductive elimination of the silicon group with the σ-bonded organic ligand. When the final product is liberated the platinum catalyst is reformed and can reenter the cycle [59]. This mechanism explains most phenomena observed in the reaction. It is supported by the fact that in some cases, the intermediates can be characterized unambiguously. A more recent example is described in [60]. Colloidal solutions of platinum and other metals are also particularly active catalysts (see [61, 62], and references cited therein).

In general the β-addition product is preferentially formed in hydrosilylation:

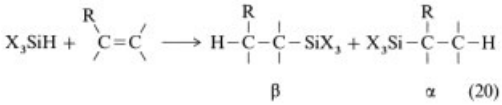

(20)

However, by using special catalysts (e.g., Rh or Pd compounds), the α-addition product can be isolated as the main component [63].

Olefins with terminal double bonds generally react faster than inner olefins. With the latter the terminal addition product is also formed because isomerization of the olefin occurs in the reversible first reaction step [64]. Silanes vary in reactivity depending on the nature of the substituents. Chlorosilanes are the most reactive, followed by alkoxysilanes; purely organic silanes and siloxanes frequently react sluggishly.

Catalysts. In industry, platinum catalysts such as Pt – activated charcoal, Pt – silica gel, $H_2PtCl_6 \cdot 6\,H_2O$ – vinylsiloxane (Karstedt solution), $H_2PtCl_6 \cdot 6\,H_2O$ (Speier catalyst), and Pt olefin complexes are the most widely used. Palladium and rhodium compounds are also frequently used. For special applications a large number of catalysts based on other metals (main-group, transition, and rare-earth metals) have been developed [65]. In homogeneous catalysis the usual platinum concentrations are in the milligram-per-kilogram range, based on the silane used.

Hydrosilylation of Other Multiple-Bond Systems. Other multiple-bond systems such as

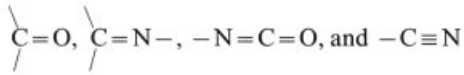

can be hydrosilylated catalytically. These reactions and hydrosilylation in the presence of carbon monoxide are not described here; details can be found in [11, 58, 65].

2.5. Substitution Reactions of Si—H Bonds

Organic groups can also be bonded to silicon by substitution of Si—H bonds. The syntheses of $PhSiCl_3$ [*98-13-5*] (Eq. 21, an alternative to the direct synthesis) and $MePhSiCl_2$ [*149-74-6*] (Eq. 22) are important industrially.

$$HSiCl_3 + C_6H_5Cl \xrightarrow{\text{ca. } 500\,^\circ\text{C}} C_6H_5SiCl_3 + HCl \qquad (21)$$

$$MeHSiCl_2 + C_6H_5Cl \xrightarrow{\text{ca. } 500\,^\circ\text{C}} C_6H_5MeSiCl_2 + HCl \qquad (22)$$

The use of Friedel – Crafts catalysts, proposed earlier, is unnecessary because they favor side reactions and decomposition. The yields are 60 – 70 %; $SiCl_4$ or $MeSiCl_3$ and benzene are formed as side products [66].

Trichlorovinylsilane [*75-94-5*] can be obtained similarly from vinyl chloride and trichlorosilane [67].

In the presence of BCl_3 or $AlCl_3$, hydrocarbons can also undergo this type of reaction with the elimination of H_2 (Eq. 23).

$$\equiv SiH + RH \longrightarrow \equiv SiR + H_2 \qquad (23)$$

A variant is shown in Equation (24):

$$\equiv\mathrm{SiH} + \mathrm{ArH} + \mathrm{Cl_2} \longrightarrow \equiv\mathrm{SiAr} + 2\,\mathrm{HCl} \qquad (24)$$

For the mechanism of this reaction, see [68].

Other substitution reactions of Si−H bonds and other Si−X bonds are discussed in Chapter 3.

2.6. Coproportionation and Disproportionation

Co- and disproportionations are used for the conversion of silanes with undesired degrees of substitution into industrially utilizable products (Eq. 25):

$$\mathrm{SiR_4} + \mathrm{SiR_4^1} \underset{\text{Disproportionation}}{\overset{\text{Coproportionation}}{\rightleftharpoons}} 2\,\mathrm{Si(R, R^1)_4} \qquad (25)$$

The reaction is accelerated by Friedel – Crafts catalysts ($AlCl_3$, Na_3AlCl_6, CuCl), organic amines, and sodium alkoxides in the case of alkoxysilanes.

There are few limitations as to the nature of the substituents that can be exchanged because alkyl and aryl groups, alkoxy groups, halogen, and hydrogen atoms are all exchangeable [69, 70]. A simple example is the coproportionation shown in Equation (26), which proceeds at moderate temperature.

$$\mathrm{MeSiCl_3} + \mathrm{Me_3SiCl} \xrightarrow{\text{Cat.}} 2\,\mathrm{Me_2SiCl_2} \qquad (26)$$

In practice, to increase the yield of Me_2SiCl_2 high- and low-methyl fractions from the direct synthesis are coproportionated catalytically in a similar way [71].

3. Silicon-Functional Organosilicon Compounds

This chapter deals mainly with organosilicon compounds of the general formula R_nSiX_{4-n} having at least one Si−X group that can be cleaved hydrolytically (X = halogen, OR^1, NR_2^1; R, R^1 = H, alkyl, aryl). Fundamental aspects of these silanes are described in [8, 10]. An overview of the data for important individual compounds is given in [72].

Reactivity. Equation (27) describes the protolysis of silanes with cleavable groups X [73].

$$\mathrm{R}_n\mathrm{SiX}_{4-n} + (4-n)\,\mathrm{HY} \longrightarrow \mathrm{R}_n\mathrm{SiY}_{4-n} + (4-n)\,\mathrm{HX} \qquad (27)$$

Industrially this is carried out with many reactants HY and has led to various industrial processes, depending on the nature and number of R groups, and the nature of the leaving group X and the nucleophile Y (Scheme 1, with X = Cl). Some syntheses also exploit the interconversion of the silicon-functional derivatives [e.g., shifting the equilibrium by removing volatile components through distillation (Eqs. 28 and 29)].

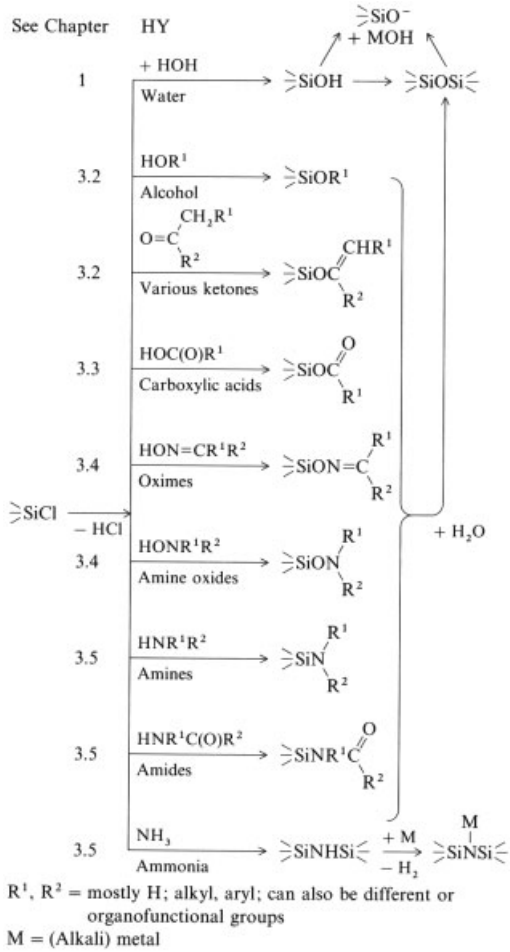

Scheme 1. Industrially important reactions of organochlorosilanes R_nSiCl_{4-n} with nucleophiles Y, according to Eq. (27) (protolysis)

$$\equiv SiNHR + HON{=}CR^1R^2 \xrightarrow{Cat.} \equiv SiON{=}CR^1R^2 + H_2NR \quad (28)$$

$$\equiv SiOR^1 + (MeCO)_2O \xrightarrow{Cat.} \equiv SiOCOMe + MeCOOR^1 \quad (29)$$

If silane compounds are reacted with one another, the thermodynamic equilibrium (as opposed to a purely statistical distribution) can be shifted in the direction of a specific component (Eqs. 30 and 31).

$$2\,RSi(NHR^1)_3 + RSiCl_3 \xrightarrow{Cat.} 3\,RSiCl(NHR^1)_2 \quad (30)$$

$$RR^1Si(OR^2)_2 + RR^1SiCl_2 \xrightarrow{Cat.} 2\,RR^1Si(OR^2)Cl \quad (31)$$

This can be attributed to differences in the free enthalpy of reaction, which has been calculated for many combinations [74]. Other metal atoms can also be incorporated in such reactions:

$$TiCl_4 + n\,Me_3SiOR^1 \longrightarrow Cl_{4-n}\,Ti(OR^1)_n + n\,Me_3SiCl \quad (32)$$

$n = 1-3$

3.1. Halo- and Pseudohalosilanes

Chlorosilanes are produced by the processes described in Chapter 2. They are generally liquids and are thermally quite stable. Their overall high reactivity in protolytic processes is intermediate in the series SiF << SiCl < SiBr < SiI and decreases with increasing substitution at silicon in the order $SiCl_4 > RSiCl_3 > R_2SiCl_2 > R_3SiCl$ and with increasing steric hindrance in the group R in the series H, CH_3, C_2H_5, $CH(CH_3)_2$, $C(CH_3)_3$.

The interest in organic synthesis has concentrated for some time on trialkylsilyl derivatives with sterically demanding R groups and particular leaving groups X. The highly reactive [75], light-sensitive iodotrimethylsilane is being used increasingly in industry. It is produced by in situ cleavage of hexamethyldisilane (see Section 5.2) in solution with elemental iodine. The crystalline, volatile *tert*-butylchlorodimethylsilane (*mp* 89 °C, *bp* 125 °C) also has specific applications. Besides its synthesis from *tert*-butyllithium and dichlorodimethylsilane, new routes have also been proposed, analogous to Equation (11). The sterically demanding chlorodimethylhexylsilane [*67373-56-2*] is formed from chlorodimethylsilane and tetramethylethylene over aluminum chloride [76] according to Equation (14). For the uses of these compounds and information on pseudohalosilanes Me_3SiX (e.g., X = CN/NC, NCO, NCS, N_3) the literature cited in Section 6.1 should be consulted.

3.2. Alkoxy- and Aryloxysilanes

The alkoxylation of chlorosilanes to give alkoxysilanes (Eq. 33) is of major industrial importance.

$$R_nSiCl_{4-n} + (4-n)\,HOR^1 \longrightarrow R_nSi(OR^1)_{4-n} + (4-n)\,HCl \quad (33)$$

R = H, alkyl, aryl; also organofunctional; R^1 = Me, Et, etc.

Alkoxysilanes are preferred because of their advantages in further use. They do not produce any hydrogen chloride (TLV 7.5 mg/m^3; in comparison, methanol = 262 mg/m^3, ethanol = 1880 mg/m^3). Acid-sensitive substances (e.g., acetone) or basic catalysts can be used in processing. Higher boiling points and flash points compared to chloro compounds (Table 3) permit the use of open apparatus under certain circumstances, and corrosion-resistant materials need not be employed. Alkoxysilanes hydrolyze more slowly and in a more controlled manner than chlorosilanes and allow access to silanols [77], defined silicone resins, or sol – gel precursors (→ Inorganic Polymers, Chap. 9.). Mixed condensations of alkoxysilanes give better yields of the target product [78] (Eq. 34), compared to chlorosilanes.

$$RSiCl_3 + Me_3SiCl \xrightarrow[\text{2) Water}]{\text{1) Isopropanol}} \underset{94\,\%}{RSi(OSiMe_3)_3} \quad (34)$$

$R = CH_2CH_2CH_2Cl$

Alkoxysilanes also offer advantages in the Grignard synthesis (Section 2.2). Coupling with

Table 3. Important chloro- and ethoxymethylsilanes Me_nSiX_{4-n}

Compound	CAS no.	M_r	*bp*, °C	Flash point∗, °C
$SiCl_4$∗∗	[*10026-04-7*]	169.9	57	
$MeSiCl_3$	[*75-79-6*]	149.5	66	11
Me_2SiCl_2	[*75-78-5*]	129.1	70	−5
Me_3SiCl∗∗	[*75-77-4*]	108.6	57	−20
$Si(OEt)_4$	[*78-10-4*]	208.3	168	50
$MeSi(OEt)_3$	[*2031-67-6*]	178.3	143	33
$Me_2Si(OEt)_2$	[*78-62-6*]	148.3	113	13
Me_3SiOEt	[*1825-62-3*]	118.3	76	−14
Me_4Si	[*75-76-3*]	88.2	26	< −20

∗ Measured on ≥ 98 % pure silanes (DIN 51 755).
∗∗ $SiCl_4$ and Me_3SiCl form an azeotrope.

sodium and organic halogen compounds proceeds similarly [79].

Significant uses of silanes [e.g., as coupling agents or for polyethylene cross-linking (Section 6)] were able to be developed to their current major importance only with the introduction of neutral alkoxy compounds.

Synthesis. The formation of alkoxysilanes according to Equation (33) is often associated with undesired side reactions, such as elimination of haloalkanes (Eq. 8) or olefins, resulting in the simultaneous formation of siloxanes. The reactions proceed more cleanly if the hydrogen chloride liberated is removed by (tertiary) amines. The additional costs of these bases and the necessary solvents and the moderate yields led to the use of processes that do not require neutralization and allow recycling of the large quantities of HCl liberated. The best industrial process is continuous, solvent-free production in a corrosion-resistant distillation column [80]. Chlorosilane, prereacted with alcohol, is fed into the upper third of the column countercurrent to the alcohol vapor coming up from the circulation evaporator at the bottom. HCl gas with a low silane content escapes via the condenser. If pure silanes are used, the quality of the alkoxysilane removed from the base of the column depends on the boiling point and water content of the alcohol. It does not generally need to be distilled. Usually alcohol-sensitive groups such as SiH, $SiCH_2CH_2Cl$, or $SiCH_2CH_2CN$ remain unreacted, since the alcohol can be used in strictly stoichiometric quantities. The action of excess alcohol on SiH groups, especially in the presence of bases, leads to loss of hydrogen according to Equation (35), often with a dangerous pressure buildup.

$$\equiv SiH + HOR^1 \xrightarrow{Cat.} \equiv SiOR^1 + H_2 \qquad (35)$$
$$R^1 = \text{Alkyl (or also H)}$$

In the absence of free alcohol, $HSi(OEt)_3$ [*998-30-1*] and potassium ethoxide form the synthetically interesting pentacoordinate compound $[HSi(OEt)_4]K$ [81]. Reactions of the type shown in Equation (35) can also be used for the synthesis of silanols or alkoxysilanes (e.g., by using Pd catalysts; see also Section 3.6).

The reaction of simple alcohols with a hydrolyzable group other than SiCl is used only in exceptional cases for economic reasons. Silyl ethers of higher-boiling alcohols or phenols are formed from methoxy- or ethoxysilanes, the best results being obtained with alkaline catalysts (e.g., alkoxides) and by distillation of the volatile alcohol. The quite selective formation of diethoxymethylsilane [*2031-62-1*], which is difficult to obtain by the above route (see Eq. 33), from oligomeric methylhydrogensiloxane [82] (Eq. 36) is noteworthy:

$$\left[-\overset{H}{\underset{Me}{Si}}O-\right]_n + n\,MeSi(OEt)_3 \xrightarrow{AlX_3} \left[-\overset{OEt}{\underset{Me}{Si}}O-\right]_n + n\,MeSiH(OEt)_2 \qquad (36)$$

Alkoxytrimethylsilanes are readily formed from the highly reactive nitrogen derivatives (see Section 3.5). Ethers of *tert*-butanol are obtained from the chlorosilane with amines or alcoholates, and phenol ethers are obtained without the addition of HCl-binding reagents. The phenoxysilanes have particularly high thermal sta-

bility: $Si(OC_6H_5)_4$ [*1174-72-7*] boils without decomposition at ca. 405 °C.

The most important reaction of the alkoxysilanes is with water. While earlier work was concerned with preparative aspects, more recent investigations on hydrolytic behavior have focussed on the kinetics in dilute, homogeneous, predominantly solvent-free solutions [73, 83–86]. Alkoxysilanes are generally insoluble or sparingly soluble in water. They undergo loss of alcohol (Eq. 37) to give reactive silanols. The first step is rate limiting.

$$R_nSi(OR^1)_{4-n} + H_2O \xrightarrow{\text{Cat.}} R_nSi(OR^1)_{3-n}(OH) + R^1OH$$
$$\downarrow + H_2O \qquad (37)$$
$$R_nSi(OR^1)_{2-n}(OH)_2 + R^1OH$$
$$\downarrow \text{etc.}$$

The silanols are solvated by water because of their higher polarity and their capacity to form hydrogen bonds, sometimes resulting in complete miscibility. Condensation to give siloxanes, which are generally insoluble, occurs parallel to hydrolysis (Eq. 38).

$$2 \equiv SiOH \longrightarrow \equiv SiOSi \equiv + H_2O \quad (38)$$

Compared with halogens the alkoxy group is a poorer leaving group. Its reactivity toward nucleophiles is limited and decreases roughly in the order $OR^1 = OCH_3 > OC_2H_5 > OCH(CH_3)_2 > OC_6H_5 > OC(CH_3)_3$. The relationships among pH, rate of hydrolysis, and stability shown for 3-glycidoxypropyltrimethoxysilane [*2530-83-8*] (Fig. 2) hold approximately for the formation and condensation of all silanols. In general, precise pH control is important for the reproducibility of the reaction and the quality of the hydrolysis product. Even traces of chlorinated hydrocarbons, formed during the alkoxylation of industrial methylchlorosilanes from olefinic side products of Rochow silanes (Section 2.1), can disturb the reaction by uncontrolled liberation of HCl [87]. The stability of the silanols $R_nSi(OH)_{4-n}$ toward condensation increases with increasing n. Unlike the quite insensitive phenylsilanols, such as $Ph_2Si(OH)_2$ [*947-42-2*], compounds with the general formula R_3SiOH condense in the presence of traces of acid with the elimination of water to give disiloxanes. Under neutral conditions they can be stored almost indefinitely. Steric hindrance in the R group also leads to stabilization. Because of hydrogen bonding, Me_3SiOH [*1066-40-6*] (M_r 90) boils at ca. 100 °C, as does its "dimer" $Me_3SiOSiMe_3$ (M_r 162). The unstable dimethylsilanediol [*1066-42-8*] [77] may be an intermediate in the hydrolysis of dichlorodimethylsilane to form silicone fluids. In addition to linear polydimethylsiloxanes (PDMSs) the primary hydrolysate contains cyclic compounds $(Me_2SiO)_n$ ($n \geq 3$; for CAS numbers and physical properties, see Table 4). The high ring strain of $(Me_2SiO)_3$ not only allows facile ring opening to give siloxane polymers (Eq. 39) but also reaction with silanols to give ambivalent siloxanes [88] (Eq. 40).

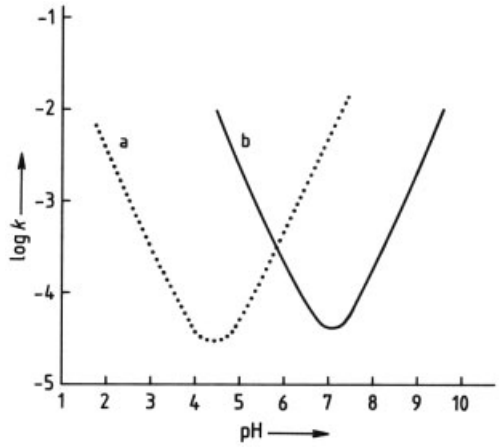

Figure 2. pH-dependent behavior of 3-glycidoxypropyltrimethoxysilane in aqueous solution (ca. 0.25 mol/L) [83]
a) Silanetriol condensation; b) Silane hydrolysis; k = rate constant

$$n\,[SiMe_2O]_3 \xrightarrow{\text{Cat. or heat}} [SiMe_2O]_{3n} \qquad (39)$$
$$n\,[SiMe_2O]_3 + RMe_2SiOH \xrightarrow{\text{Cat.}} R[Me_2SiO]_{3n+1}H \qquad (40)$$

R = (Functional) alkyl group

Cyclic siloxanes react particularly readily with chlorosilanes such as $MeRSiCl_2$ (R = H, Me, $CH{=}CH_2$) to form linear, short-chain, α,ω-halogenated siloxanes [89], which can be worked up hydrolytically to yield polymers.

Table 4. Industrially important methyloligosiloxanes

Structure *	CAS no.	*mp*, °C	*bp*, °C
$HMe_2SiOSiMe_2H$	[*3277-26-7*]	<−100	71
$Me_3SiOSiMe_3$	[*107-46-0*]	ca. −65	100
$VMe_2SiOSiMe_2V$	[*2627-95-4*]	ca. −103	140
$(Me_2SiO)_3$	[*541-05-9*]	62	135
$(Me_2SiO)_4$	[*556-67-2*]	17	170
$(Me_2SiO)_5$	[*541-02-6*]	−44	210
$(HMeSiO)_4$ **	[*2370-88-9*]	ca. −69	135
$(VMeSiO)_4$ **	[*2554-06-5*]	ca. −40	229

* $V = CH_2{=}CH$.
** Mixture of positional isomers.

Siloxane cleavage also occurs with highly reactive silanols formed in situ from chloro- or alkoxysilanes, water, and a catalyst [90], often selectively (silanolysis, Eq. 41; X = Cl, OR^1).

$$2\ RSiX_3 + 3\ HMe_2SiOSiMe_2H \xrightarrow[H^+]{3\ H_2O} 2\ RSi(OSiMe_2H)_3 + 6\ HX \qquad (41)$$

The reactivity of well-defined oligosiloxanes is reported in [91].

Water-soluble alkylsilanetriols can be stabilized only for a limited period in dilute solution under weakly acidic conditions. On warming, aeration of the solvent, or increasing the pH, they condense to resinous, generally water-insoluble polymers.

In strongly alkaline media, alkali-metal siliconates **1** are formed from $RSi(OR^1)_3$ (Eq. 42).

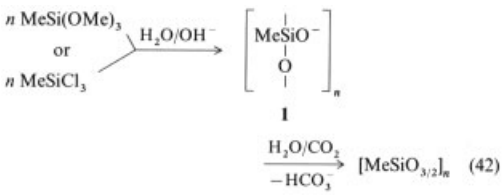

These are widely used to render mineral substrates hydrophobic. They react with atmospheric carbon dioxide to form insoluble, water-repellent methylsilicone resin, which is permeable to water vapor. It is used for impregnating mineral construction materials such as roofing tiles, or for protection against rising damp, particularly in old walls with low alkali content. More recently, silicone microemulsions have been used successfully (see Section 6.8).

A special category of alkoxysilanes is the silenol ethers, which can be obtained according to Equation (43) [92].

$$CH_3SiCl_3 + 3\ O{=}\underset{\underset{R^2}{|}}{C}{-}CH_2R^1 + 3\ B \longrightarrow CH_3Si(O{-}\underset{\underset{R^2}{|}}{C}{=}CHR^1)_3 + 3\ B\cdot HCl \qquad (43)$$

Moisture converts the enolized leaving group back into the starting neutral ketone [e.g., acetone (R^1 = H, R^2 = Me) or 5,5-dimethyl-1,3- cyclohexanedione (see Section 6.3)].

Esters of orthosilicic acid, $Si(OR^1)_4$, and their lower condensation stages are not organosilanes in the strictest sense. However, because of their industrial importance they are covered here. Unlike organo(organoxy)silanes, tetraalkoxysilanes are synthesized directly from silicon [93] or suitable natural silicates [94] and alcohols. Many side products are formed in this reaction, so that the overall energetic advantages of this direct synthetic route are offset by an uneconomical workup. Therefore products from the reaction of $SiCl_4$ with ethanol and water, which have a calculated SiO_2 content of 40 wt %, are used in practice [95]. These so-called ethyl silicates 40 [*26352-16-9*] are mixtures of linear, branched, and cyclic ethoxysiloxanes, which are stable on storage; their viscosity is ca. 4 mm^2/s. Unlike analogous condensation products of the toxic tetramethyl silicate, the cheap ethoxy compounds are used in large quantities. Water or atmospheric moisture causes loss of ethanol to give a strongly adhering silicic acid binding agent. Hydrolysis and cross-linking can be accelerated by catalysts. Examples of uses are:

1) Cross-linking of silicone rubber (see Table 6).
2) Inhibition of fiber pilling on polyester fabrics [as a mixture with alkoxyphosphorus compounds; 2-(diethylphosphono)-ethyltriethoxysilane has been used as an alternative, see Section 4.6].
3) Consolidation of weathered sandstone for renovation of buildings (catalyst alkylacyltin compounds) [96], with retention of the open-pored character of the stone.
4) Protection of metallic construction components (bridges, containers, ships, pipelines): finely divided zinc pigments and other additives form impermeable coatings (cold galvanization) with silicic acid on steel, which protect against water and chloride ions and are resistant up to 420 °C. Even if the layer is damaged, the protective action is retained, since a local element is formed with zinc as the anode [97].
5) Making molds for precision casting: In many work cycles thin layers of ethyl silicate and fireproof pigments are deposited on top of each other on wax patterns and hardened in air or an ammonia atmosphere. After melting of the wax and firing at ca. 1000 °C the metal is cast into the porous mold.

3.3. Acyloxysilanes

Acyloxysilanes, $R_nSi(OCOR^1)_{4-n}$, are less readily formed than the alkoxy compounds. Acetoxysilanes ($R^1 = Me$) are very important industrially as cross-linking agents in one-component silicone sealants (RTV 1 systems; see → Silicones, Chap. 4.4.3.; and Section 6.3). They can only be produced economically from the reactive chlorosilanes:

$$RSiCl_3 + 3\,Ac_2O \longrightarrow RSi(OAc)_3 + 3\,AcCl \quad (44)$$

$$RSiCl_3 + 3\,HOAc \longrightarrow RSi(OAc)_3 + 3\,HCl \quad (45)$$

$$R = CH_3, C_2H_5, HC{=}CH_2;\ Ac = COCH_3$$

The first, older variant (Eq. 44) requires removal of acetyl chloride by distillation. Heating in the presence of traces of catalytically active metals can lead to reformation of acetic anhydride and oligomerization (Eq. 46).

$$2 \equiv SiOAc \longrightarrow \equiv SiOSi\equiv + Ac_2O \quad (46)$$

Continuous countercurrent reaction with acetic acid is more suitable (Eq. 45) [98]. The products are practically free of chloride, acid, and anhydride, and yields are close to 100 %. The compounds, known as acetic acid hardeners, are spontaneously hydrolyzable, corrosive, high-boiling liquids with melting points between 5 and 45 °C.

3.4. Oximino- and Aminoxysilanes

The trend toward RTV 1 systems, whose cross-linking agents eliminate neutral products, has made the long-known oxime and aminoxy cross-linking agents even more important industrially. Intensive work has been carried out on the production of alkyloximinosilanes; the oximino group (Ox) is generaly $-ON{=}C(CH_3)C_2H_5$ [99]. The best process is the reaction shown in Equation (47) [100].

$$MeSiCl_3 + 6\,HOx \xrightarrow[<60\,^\circ C]{\text{Solvent}} MeSi(Ox)_3 + 3\,HOx \cdot HCl \quad (47)$$

The amphoteric 2-butanone oxime acts as both a reactant and a "base". With hydrogen chloride it forms a liquid, readily separable phase that can be reacted with alkali to give 2-butanone oxime again. Careful dosage and monitoring lower the danger of vigorous decomposition of the acidic oxime hydrochloride.

Aminoxysilanes, $R_nSi(ONR^1R^2)_{4-n}$, are formed by reaction of hydroxylamines with alkylchlorosilanes, with removal of the liberated hydrogen chloride by bases [101]; another possible route is shown in Equation 53.

3.5. Aminosilanes, Amidosilanes, and Silazanes

(Alkylamino)alkylsilanes. Because of the high affinity of amines for HCl the industrial synthesis according to Equation (48) requires an excess of base. Triethylamine (A) can be used as co-base and solvent [102]. The granular, crystalline, readily separable hydrochloride formed can be worked up for reuse.

$$\equiv SiCl + H_2NR^1 + A \longrightarrow \underset{\text{Silylamine or (mono-)silazane}}{\equiv SiNHR^1} + A \cdot HCl \qquad (48)$$

When carbon dioxide is passed through, carbamatosilanes $\equiv SiOCONHR^1$ are formed from silylamines [103].

Amidosilanes. The amidosilanes $R_nSi-[NR^1C(O)R^2]_{4-n}$ ($R, R^1 = Me$; $R^2 = Me$ or $C_6H_5CH_2$), which are useful neutral cross-linking agents, are formed from chlorosilanes and organic amides by using strongly alkaline HCl scavengers (e.g., sodium metal or methoxide) or from Si–H compounds [104].

The silicon-monofunctional derivatives of urea and acetamide (see Table 5), used in large quantities for the synthesis of pharmaceuticals, also belong to this group of substances. *N,N′*-Bis(trimethylsilyl)urea (BSU) is produced industrially in quantitative yield from hexamethyldisilazane (HMDS) and urea, with, for example, a catalytic quantity of an ammonium salt or an organic sulfonamide [105] (Eq. 49).

$$(Me_3Si)_2NH + H_2NCONH_2 \xrightarrow[120\,^\circ C]{Cat.} (Me_3SiNH)_2CO + NH_3 \qquad (49)$$

The *N,O*-bis(trimethylsilyl)acetamides are among the most powerful trimethylsilyl-transfer agents. Production involves the conversion of acetamide into the liquid BSA by means of Me_3SiCl – trialkylamine via the solid monosilyl derivative (MSA) [106] (Eq. 50).

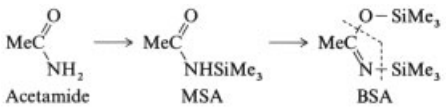

(50)

On heating in the presence of Fe ions, for example, bis(trimethylsilyl)amides decompose to hexamethyldisiloxane and the corresponding nitrile, as indicated by the dashed lines in Equation (50).

Disilazanes. Many chlorosilanes react with ammonia to give not stable, simple silylamines $\equiv SiNH_2$, but disubstitution products with the SiNHSi building block. Hexamethyldisilazane has achieved major industrial importance. The exothermic reaction according to Equation (51) can be carried out in solvents such as hexane, toluene, or isododecane, but is best performed in warm HMDS itself [107].

$$2\,Me_3SiCl + 3\,NH_3 \longrightarrow (Me_3Si)_2NH + 2\,NH_4Cl \qquad (51)$$

The latter variant leads to the formation of coarse-grained ammonium chloride, which can be readily stirred and filtered and gives better turnover and yield. The suggestion has been made that the salt be separated by treatment with water. HMDS forms strong hydrogen bonds and is only weakly basic. Alkali and alkaline-earth metals form salts with it, e.g., $LiN(SiMe_3)_2$ [*4039-32-1*].

3.6. Hydrogensilanes

Organohydrogensilanes (H-silanes) are obtained by the general synthetic methods described in Chapter 2. The cheapest raw material for Si–H compounds, trichlorosilane [*10025-78-2*], is produced from silicon and hydrogen chloride (→ Silicon Compounds, Inorganic, Chap. 2.1.1.). An Si–H synthesis from halosilane and metal hydride can only be considered for special, high-value sil(ox)anes [108]. Cleavage of Si–H bonds (and Si–Ph) is effected by halogens, giving halogen derivatives. H-Substitution also takes place on protolysis (e.g., Eq. 35). Other reactions of Si–H silanes involving loss of hydrogen are summarized in [109].

Many of the reactions shown in Scheme 1 can be carried out with retention of the Si–H group. Insertion into Si–H bonds occurs on hydrosilylation (Section 2.4). In this way, the compounds $R_nSiH(OR^1)_{3-n}$, which can be produced from the corresponding chlorosilanes and alcohols, can be converted into functional alkoxyalkylsilanes (see Chapter 4). With amine or metal catalysts the alkoxy H-silanes sometimes undergo H-redistribution. In the absence of SiC-bonded groups, monosilane SiH_4 [*7803-62-5*], which is spontaneously flammable in air, can be formed [110]. The methylhydrogensiloxanes, which are used widely in industry, can react spontaneously with traces of alkali [111] according to Equation (52), to give gaseous methylsilane, $MeSiH_3$ [*992-94-9*].

$$3\left[\text{MeSi(H)O}\right] \xrightarrow{\text{Cat.}} \text{MeSiH}_3\text{(g)} + 2\left[\text{MeSiO}_{3/2}\right] \quad (52)$$

Siloxanes with the structural elements **2**, **3**, or **4**, formed according to Equation (53), are effective condensation cross-linking agents.

$$\left[\text{MeSi(H)O}\right] + \text{HY} \xrightarrow{\text{Cat.}} \left[\begin{array}{c}\text{MeSiO}\\ |\\ \text{Y}\end{array}\right] + \text{H}_2 \quad (53)$$

2 – 4

2: $Y = NR^1C(O)R^2$ [104];

3: $Y = OCH_2CH_2NR^1_2$ [112];

4: $Y = ONR^2_2$ [113]; R^1 = e. g., Me; R^2 = e. g., Me , Et

4. Organofunctional Organosilicon Compounds

Compared with the compounds described in Chapter 3, which react exclusively at the labile substituent X on the silicon, in this chapter derivatives with functionalized organic groups R^1 are considered. These silanes can be described by the general formula $X_nR_{3-n}Si{-}R^1{-}Y$, except for the alkenyl-substituted compounds. The alkyl substituents R and the leaving groups X correspond to those in Chapter 3. The Y groups, which are generally attached by stable bonds, are decisive for the activity of this class of compounds. These organic functional groups are used to bond quite different organic polymers to one another and to inorganic surfaces via siloxane bridges. The Y group generally participates itself in chemical bonding (e.g., in coupling agents, cross-linking agents) or at least increases the compatibility of different materials and surfaces.

The industrial development of this class of substances is closely associated with that of hydrosilylation. Vinylsilanes were first used, followed by aminopropyl- and mercaptosilanes [114]. In 1963 the methacrylic function, and later the epoxy and cationic styrene derivatives, came on the market.

4.1. Vinyl and Other Alkenyl Compounds

Industrial use of organofunctional silanes began in the early 1950s with the development of vinylsilanes and their reinforcing effect in glass-fiber – polyester plastics. Alkenyl compounds still dominate as strengthening additives in polymer applications and, in particular, as reactive components in many silicone formulations.

Vinylsilanes are produced on a large scale mainly by continuous processes. The preferred route is catalytic addition of hydrogen-bearing chlorosilanes to acetylene.

$$\text{HC}{\equiv}\text{CH} + \text{H}\overset{\overset{\text{X}}{|}}{\text{Si}}\text{RR}^1 \xrightarrow{\text{Pt}}$$

X = Cl

$$\text{H}_2\text{C}{=}\text{CH}{-}\overset{\overset{\text{X}}{|}}{\text{Si}}\text{RR}^1 + \text{R}^1\overset{\overset{\text{X}}{|}}{\text{R}}\text{Si}\text{CH}_2\text{CH}_2\overset{\overset{\text{X}}{|}}{\text{Si}}\text{RR}^1 \quad (54)$$

In 1953 the use of platinum on solid carriers for this reaction was patented [57]. For example, trichlorosilane and acetylene react at ca. 130 °C in a tubular reactor to give a product mixture with 70 % $CH_2{=}CHSiCl_3$ and 19 % of the diadduct after a few minutes. Formation of $Cl_3SiCH_2CH_2SiCl_3$, derivatives of which are used to a limited extent as "bridging hardeners," can be suppressed by using a large acetylene excess and short residence time. Since heterogeneous catalysts such as Pt – active charcoal are often rapidly deactivated, soluble complexes such as hexachloroplatinic acid (Speier catalyst) [58] are now used. These are added in high-boiling solvents such as dichlorobenzene and lead to formation of the desired vinylsilanes in >90 % yield without overpressure [115]. Through control of the temperature and reagent stoichiometry, the target product can also be removed continuously from the reactor [116]. In this way the subsequent reaction to give the diadduct, which acts simultaneously as a solvent, is suppressed still further. Addition reactions of alkylsilanes ($RHSiCl_2$) or H-substituted siloxanes proceed analogously. In particular, reactions with dichloromethyl- and chlorodimethylsilane are of industrial interest; those with the less reactive H_2SiCl_2 are somewhat less so.

Vinyl compounds can also be obtained from vinyl chloride by substitution reactions of Si–H bonds (see Section 2.5) [67].

The Grignard synthesis is worthwhile only for vinyl-rich intermediates that are virtually inaccessible by hydrosilylation.

Allyl compounds are also produced on a large scale. Tri- and dichloroallylsilanes can be prepared by replacing the halogen (e.g., in allyl chloride) with H-silanes. The resulting condensation with loss of HCl can be carried out at elevated temperature or with tertiary amines and CuCl as an effective catalyst [117, 118]. The allyl group can also be generated by subsequent HCl elimination from chloropropylsilanes, which are available on an industrial scale. Like the allyldimethylsilyl chain stoppers, these compounds also play an important role as reactive building blocks for polysiloxanes. The monochlorosilane precursor is produced better by cleavage of methyl-rich disilanes with allyl chloride, as is allyltrimethylsilane, used in pharmaceutical synthesis [64, 119]. The Grignard coupling of $CH_2{=}CHCH_2X$ with Si–X compounds (Section 2.2) is the most universal method, but because of the low time–space yield, it is limited to laboratory syntheses and the production of intermediates that cannot be produced directly by the above methods.

The lower stability of the Si–C bond toward strong acids and bases, compared with vinylsilanes, is also a sign of the special reactivity of the allyl group.

Other industrially interesting alkenyl compounds include adducts of H-sil(ox)anes with vinylcyclohexene, linear diolefins, or substituted alkynes.

The addition of trichlorosilane or dichloromethylsilane to vinylcyclohexene gives the desired cyclohexenylethyl compound with excellent selectivity because of the considerably higher reactivity of the terminal double bond. The cyclohexene structure is, moreover, suitable for the production of epoxysilanes and epoxysiloxanes (see Section 4.5.1) after replacement of the chlorine substituent. The process is more difficult to control when several double bonds of equal reactivity are present in the starting material. Diolefins such as hexadiene or octadiene must be present in at least two- to fourfold excess if good yields of the monoadduct are desired. Isomers derived from internal olefin structures may also be formed.

The resulting alkenylsilicon compounds are used, like the allyl derivatives, for particularly reactive silicones [e.g., LTC (low-temperature curing) systems for coating plastic films or paper] [120]. Such systems can also be produced by incorporating industrially relevant allyl ether precursors. Some of these newly developed siloxanes can even be cross-linked cationically after rearrangement of the allyl group to a propenyloxy group (cf. vinyl ethers) [121].

Butadiene gives mono adducts with high selectivity, but the resulting allylsilanes ($\equiv SiCH_2CH{=}CHCH_3$) are not suitable for addition reactions because of the terminal methyl group.

Sil(ox)anes with a butadienyl substituent are, like the conjugated diolefin itself, very reactive. Such compounds can easily be produced by the hydrosilylation of butynols, which are available on an industrial scale [122]. Monoaddition to the triple bond (Eq. 55) is highly selective and gives two isomeric adducts, both of which readily eliminate water.

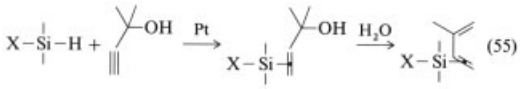

This elimination is generally carried out in situ, in the presence of the next reaction partner, which then becomes bonded to the isoprenylsilicon intermediate, for example, in a [2 + 4] cycloaddition (see Section 4.5.3).

Silylacetylenes can be produced relatively easily (e.g., from chlorosilanes and acetylene derivatives), but because of their low stability, particularly toward protic compounds, they have not achieved industrial importance thus far.

4.2. Organohalogenated Compounds

Silanes with halogen-containing side groups play a key role in silicone chemistry and particularly as reactive intermediates for the introduction of various Y functions. The downstream products of allyl chloride are the most important. 3- or γ-chloropropylsilanes are produced on a large scale. The production process starts with a catalytic addition, which is

often carried out continuously (see Scheme 2). In the case of trichlorosilane it gives the desired chloropropylsilane with a selectivity up to ca. 83 % [123]. The formation and addition of propene, occurring as a side reaction, can be interpreted as Cl−H exchange. It increases in the order $HSiCl_3 < MeHSiCl_2 < Me_2HSiCl < Me_2HSiPh$ [124]. The resulting losses can, however, also be explained by the intermediate decomposition of a β-adduct, which is perhaps therefore not isolable. Accordingly, methallyl chloride gives fewer side products.

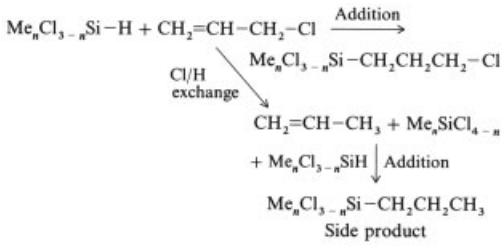

Scheme 2. Addition of silanes to allyl chloride

The (chloropropyl)chlorosilanes are alkoxylated on a large scale by a reaction analogous to that shown in Equation (33) (Section 3.2). The methoxy and ethoxy derivatives are almost the only industrially important ones. They are additives for various formulations and, in particular, reactive intermediates for the production of other organofunctional silanes. They are also accessible directly by a new process starting from allyl chloride and H-alkoxysilanes with iridium catalysts [125].

The corresponding bromo- and iodopropyl compounds are of only minor importance, despite their higher reactivity in nucleophilic reactions, since the addition of iodine or phase-transfer catalysts is generally sufficient to activate the organochlorine group.

Chloromethyl compounds have achieved a certain importance (e.g., in the synthesis of biologically active substances) because of the many derivatives that can be formed from them (see Section 6.2). They are generally produced by radical chlorination of methylsilanes from the direct synthesis. Such processes generally exhibit modest selectivities and in the silicone industry are used only for methyltrichloroand chlorotrimethylsilane. The corresponding monochlorinated derivatives, $ClCH_2SiCl_3$ and $ClCH_2(CH_3)_2SiCl$, can be used in the same way as the chloropropylsilanes described above. Compounds with additional organic functional groups can be obtained from chloromethyllithium as the reactive intermediate [126].

Chlorophenylsilanes are produced by direct chlorination with Friedel – Crafts catalysts. Substitution in the aromatic ring leads to stabilization, making the compounds suitable for high-temperature applications.

β-Chloroethylsilanes show exceptionally high reactivity, which is further enhanced in the presence of bases. These compounds are readily produced by the reaction of gaseous HCl with chlorovinylsilanes, which undergo smooth anti-Markovnikov addition in the presence of soluble aluminum catalysts [127]. The terminal carbon atom is rendered positively charged by the so-called β-effect, and the loss of the anionic leaving group is activated. In the presence of moisture the functional group decomposes according to Equation (56). The rate of ethylene formation depends on the pH and the nature of the groups bonded to silicon [128]. This has been used to control ripening processes.

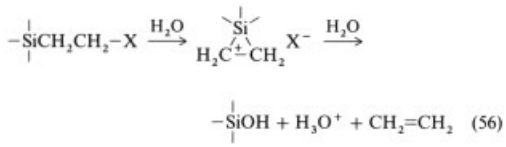

Industrial fluoroorganosilanes differ fundamentally from the organohalogen compounds considered thus far. The systems that are of interest in industrial applications are strictly speaking not organofunctional since they are actually very inert. Because α- and β-C-fluorinated compounds are of comparatively low stability, fluoro substituents are incorporated only from the γ-position onward. The most important monomers are derived from 3,3,3-trifluoropropene. Trifluoropropylsilanes are produced by the catalytic addition of H-silanes.

Longer- chain highly fluorinated side groups lead to considerably lower surface tension. A monolayer of $CF_3(CF_2)_7CH_2CH_3SiO\equiv$ reaches values of < 8 mN/m [129]. In practical applications, such substituents, which can also be introduced by hydrosilylation, give partic-

ularly favorable effects with regard to solvent resistance, wettability, and adhesion/release. In addition, syntheses with SiH – fluoroaromatics (condensation, H_2 cleavage), Cl–F exchange, or Grignard precursors are known but have not achieved industrial importance.

4.3. Nitrogen-Containing Compounds

Parallel to the development of vinyl systems, nitrogen derivatives – in particular aminoorganosilanes – gained increasing industrial importance. These very reactive compounds play an important role in many silicone formulations, as well as in a large number of other applications involving compatibility, adhesion, or bonding of different substrates.

4.3.1. Cyanoalkyl Compounds

Cyanoethylsilanes were synthesized very early on by Si–H additions of chlorosilanes to acrylonitrile. Such processes have been carried out on a large scale since reduction of the alkoxylated downstream products to give the industrially interesting aminopropyl group became possible [130]. The acrylonitrile process is, however, carried out not with the usual transition-metal catalysts (which sometimes give the unstable α-adducts exclusively), but in the presence of tertiary amines or phosphines. Presumably, anionic intermediates are involved [131].

$$Cl_3SiH \xrightarrow{NR_3} \{Cl_3S\bar{i} - H - \overset{+}{N}R_3\} \xrightarrow{CH_2=CHCN} Cl_3SiCH_2CH_2CN \qquad (57)$$

Addition of CuCl leads to more effective systems [132]. The corresponding trialkoxy compounds are produced by an alkoxylation process analogous to that described in Section 3.2 or by the addition of HCN to alkoxyvinylsilanes [133].

Longer- chain cyano compounds can be produced by noble-metal-catalyzed hydrosilylation of alkenyl nitriles such as allyl cyanide. The resulting cyanopropylsilanes are used, for example, in chromatography.

Organically bonded nitrile functions can be hydrolyzed to give free carboxylic acids. Hydrogenation is described in Section 4.3.2.

4.3.2. Organic Amino Compounds

Aminopropyl systems are the most important organofunctional sil(ox)anes. "Primary" aminosilanes are preferred because of their high reactivity.

Aminopropyltriethoxysilane [*919-30-2*], the most important individual compound, can be produced by three methods.

The acrylonitrile process (Section 4.3.1) is still regarded as very economical despite the necessary expenditures for safety. Cyanoethylsilane originating from the addition step (Eq. 57) is ethoxylated and then hydrogenated over Raney nickel at high hydrogen pressure [130].

$$Cl_3SiCH_2CH_2-CN \xrightarrow[2)\ H_2,\ Ni]{1)\ EtOH} (EtO)_3SiCH_2CH_2CH_2-NH_2 \qquad (58)$$

The hydrosilylation of allylamine appears as a direct route to be actually simpler than the other, multistage processes, which in addition require pressurized reactors. The addition of triethoxysilane is, however, somewhat inhibited with the normal platinum catalysts and gives too much of the β-adduct (i.e., internal addition, 10 – 25 %), which causes problems in many applications. Rhodium phosphine complexes increase the overall yield of the desired γ-product to > 80 % so that the process now appears to be competitive [134].

$$(EtO)_3SiH + CH_2=CHCH_2NH_2 \xrightarrow{Rh} (EtO)_3SiCH_2CH_2CH_2NH_2 + (EtO)_3Si\overset{Me}{\overset{|}{C}}HCH_2NH_2,\ Si(OEt)_4,\ H_2 \qquad (59)$$

This reaction is an example of the advantageous use of expensive rhodium compounds, which generally exhibit higher selectivity and lower sensitivity to inhibiting substances. The reaction mode and yield are now comparable to those of the more reactive (and with platinum compounds sufficiently selective) *N*-trimethylsilylallylamine [65].

The ammonia substitution process starts from chloropropyltriethoxysilane [*5089-70-3*], and requires a large excess and high pressure of NH_3.

With a 75-fold molar quantity of NH_3 with respect to the chloro component, selectivities for the primary amine of up to 81 % have been reported [135]. The correspondingly low space–time yields can only be improved if increased proportions of di- and even trisubstituted downstream products of the initially formed aminopropylsilane are acceptable.

$$(RO)_3SiCH_2CH_2CH_2Cl \xrightarrow{NH_3} [(RO)_3SiCH_2CH_2CH_2]_nNH_{(3-n)} \quad (60)$$

Whether the NH_3 process is economical depends strongly on the possibilities for using the side products, particularly the bis(3-trialkoxysilylpropyl)amine (e.g., as coupling agents).

The substitution process with primary or even secondary amines or with ethylene-bridged polyamines as the nucleophiles is considered more advantageous. γ-(Ethylenediamino)propylalkoxysilanes are produced industrially by a reaction analogous to Equation (60). A three- to fivefold excess of diamine is generally sufficient to achieve yields of monomer $> 80\,\%$. One equivalent of the strong base acts as an HCl acceptor and forms a readily separable liquid phase. The very reactive methoxy compounds are of particular industrial importance.

In many applications *N*-(2-aminoethyl)-3-aminopropyltrimethoxysilane [*1760-24-3*] competes with or supplements aminopropyltriethoxysilane, the most widespread organofunctional silane. The analogous dimethoxymethylaminosilane [*3069-29-2*] is the active building block of many amine fluids (for treating textiles, etc.).

Other important amino compounds are the *N*-methyl-, *N,N*-dimethyl-, and *N*- cyclohexylaminopropylsilanes. These are also produced by the substitution process. Here the removal of the amine hydrochloride, which is produced simultaneously, is problematic. Chloride-free pure products in good yields are obtained only after subsequent treatment with alkoxides or other strong bases [124]. Such secondary or tertiary amino groups impart advantages in certain applications (e.g., less thermal yellowing). They can therefore be considered alternatives to the standard silanes discussed previously. They are also used in the preparation of quaternary ammonium and betaine structures.

Organic amino functions at the end of siloxane chains are becoming increasingly important for the synthesis of block structures with organic polymer segments. Primary amino groups are preferred.

Bis(aminopropyl)tetramethyldisiloxane (aminopropyl chain stopper [*2469-55-8*]) is not sufficiently accessible either by substitution with NH_3 or by a route analogous to Equation (57). Since the direct addition of tetramethyldisiloxane (cf. Eq. 59) is also unsatisfactory, cyclic hydrosilylation of the silazane from chlorodimethylsilane and allylamine has been proposed [136].

$$CH_2{=}CH_2CH_2{-}NH{-}Si(Me)_2H \xrightarrow{Pt} Me_2\overline{SiCH_2CH_2CH_2{-}NH} \quad (61)$$

The resulting Si–N ring is very reactive and can be cleaved hydrolytically to give the free aminopropyl group. The route involving phthalimide (Gabriel synthesis) or the treatment of the corresponding chloropropyldisiloxane with sodium diformylamide appears to be more straightforward [137].

As an alternative for certain applications the ethylenediaminopropyl stopper function may be considered, which is readily produced by direct substitution (cf. Eq. 60).

Aromatic amino compounds, which are used for the modification of particularly thermostable composites (e.g., in polyimides) can be produced by hydrogenation of nitrophenyl precursors (accessible according to Eq. 66 [138]).

4.3.3. Other Nitrogen Compounds

Other interesting nitrogen compounds can be produced by nucleophilic substitution of chloropropylsilanes, including isocyanate, urethane, ureido [123], and imidazole derivatives. Some of these compounds can also be produced by hydrosilylation of the corresponding allyl precursors and further derivatization. They are versatile intermediates in silicone chemistry.

Quaternary ammonium compounds (quats) are produced mainly by substitution reactions

of chloropropyl precursors. The reaction with long-chain amines such as octadecyldimethylamine [139] is of industrial importance because it leads to surface-fixable biocides (Section 6.2). Methyl-substituted quats, which are also accessible from alkylaminosilanes, such as the side products shown in Equation (60) ($n = 2, 3$) act as ion exchangers and, if the chains are longer, also as phase-transfer catalysts [123].

The many compounds produced from aminosilanes include the cationic styrene compound $CH_2{=}CHC_6H_4CH_2N^+H_2CH_2CH_2NH{-}CH_2CH_2CH_2Si(OMe)_3$ [*34937-00-3*], which is produced by substitution of *p*-vinylbenzyl chloride and was introduced in 1972 as a copolymerizable coupling agent [114].

4.4. Organosulfur Compounds

4.4.1. Mercapto and Sulfidic Organofunctions

Mercaptofunctional silanes with spacers of more than one carbon between sulfur and silicon were reported as stable compounds in the early 1960s. These monomers, which are important in the rubber industry, are produced mainly from the corresponding chloropropyl precursors. Since alkali-metal hydrogensulfides give too much bis(silylpropyl)sulfide in this reaction, the mercaptopropyl compound can, in practice, be obtained directly only with H_2S – amine combinations [140]. Pure monomers are better produced from protected *S*-nucleophiles such as thiourea. The isothiuronium salts formed initially are then cleaved with base (e.g., ammonia).

$$(RO)_{3-n}Me_nSi(CH_2)_3Cl \xrightarrow{+\,S{=}C(NH_2)_2} (RO)_{3-n}Me_n\,Si(CH_2)_3S^+C(NH_2)_2Cl^- \xrightarrow[-(H_2N)_2C{=}NH_2{}^+Cl^-]{+\,NH_3} (RO)_{3-n}Me_nSi(CH_2)_3{-}SH \quad (62)$$

Since complete separation of the guanidinium salts is often difficult, addition of solvents or cleavage with alkylamines, which form liquid phases, has been proposed [123, 141]. The same mercapto compounds can also be produced by addition of H_2S to the corresponding cyanoethylsilanes or by their reductive cleavage with $H_2{-}CoS_x$ [142]. β-Mercaptoethyl compounds can be produced from vinyl compounds and H_2S [143].

The thiol group is very versatile, bonding, for example, by radical addition to alkenyl systems under mild conditions. This predominantly anti-Markovnikov addition is of considerable industrial importance in the cross-linking of various polymers.

Silanes with sulfidic sulfur bridges are of considerable industrial importance. They are produced mostly by direct substitution with sodium polysulfides Na_2S_x [123]. The average chain length of S_x is decisive for the reactivity of the compounds (e.g., in rubber vulcanization). The chain length can be regulated by addition of sulfur or by adjusting the pH during substitution with disodium tetrasulfide. This process is used industrially on a large scale. The most important compound, the tetrasulfane $[(EtO)_3SiCH_2CH_2CH_2]_2S_4$ [*40372-72-3*], is used as a strengthening additive for the treatment of fillers and for the cross-linking (vulcanization) of polyolefin elastomers.

Disulfidic compounds, whose sulfur bridges can be readily cleaved by oxidation (see Section 4.4.2) or reversible reduction to thiol groups, can be obtained by the same route.

4.4.2. Compounds Containing Sulfur – Oxygen Groups

The ready oxidizability of sulfidic sulfur compounds (e.g., with H_2O_2) is used for the synthesis of sulfonate-functional siloxanes, which are used as ion exchangers, for example [144].

Such groups are preferably produced directly from the corresponding chloropropyl precursors. Substitution with Na_2SO_3 gives sulfopropylsilane – silanol intermediates, which are suitable for the production of hydrophilic to tensidic silicone oils [145]. Sulfoethylsilanols can be synthesized from vinylsilanes and bisulfite [146].

Thiosulfate groups are introduced by direct substitution with $S_2O_3^{2-}$ (Eq. 63) [147]. The organically bonded thiosulfate group (Bunte salt) undergoes hydrolytic cleavage to give mercaptopropylsiloxanes.

$$(RO)_{3-n}Me_nSi(CH_2)_3Cl \xrightarrow[-NaCl]{+\,Na_2S_2O_3,\ H_2O} (RO)_{3-n}Me_nSi(CH_2)_3S_2O_3^-\ Na^+ \quad (63)$$

Compounds in which sulfur is bonded via oxygen are generally less stable to hydrolysis and are not dealt with here.

4.5. Oxygen-Containing Compounds

4.5.1. Epoxy and Other Oxy Compounds

In epoxy groups, oxygen is part of a highly activated ring structure. Since oxiranes can be opened even by traces of HCl, chlorosilanes are generally not suitable as starting materials. Industrial production of such silicon compounds starts mostly from alkoxy H-silanes or siloxanes and epoxidized alkenes such as allyl glycidyl ether or vinylcyclohexene oxide [124, 148]. Hydrosilylation with chlorine-free platinum catalysts or, for less pure reactants, with rhodium catalysts leads directly to monomers such as 3-glycidoxypropyltrimethoxysilane [*2530-83-8*] or 3,4-epoxycyclohexylethyltrimethoxysilane [*3388-04-3*], which are used as, e.g., coupling agents (Section 6.7).

The epoxidation of alkenylsil(ox)anes (Eq. 64) avoids toxic, potentially mutagenic, oxirane precursors [149, 150].

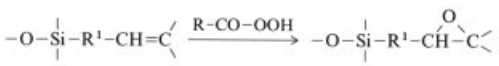

(64)

This relatively mild oxidation is particularly suitable for polysiloxanes with epoxy groups in the chain. The degree of functionality can be controlled most readily by equilibration with the alkenylsilane precursor.

Whereas vinylic oxiranes, such as the α- or β- chloro or hydroxy compounds, rearrange relatively easily or eliminate olefins (Peterson reaction [151]), higher epoxy compounds behave like their organic analogues. Their high reactivity with acids and, in particular, with nucleophiles has led to a large number of individual applications.

Hydroxyalkyl groups are stable only if the OH group is in the γ-position or further from the silicon; they then react like normal alcohols. Direct hydrosilylation of allyl alcohol is now rarely used industrially because of considerable hydrogen formation in alcoholysis side reactions [124] and because of the toxicity of the alcohol. Since these problems are only partially solved by using silyl protecting groups, less critical precursors are used, such as allyl acetals or allyl esters, which can be hydrolyzed readily to hydroxy compounds after Si–H addition. Comparable intermediates are also obtained by the esterification of chloropropyl precursors with carboxylic acid salts. Similar hydroxy compounds are accessible by hydrosilylation of higher alkenols or glycol monoallyl ethers. Losses due to alcoholysis decrease with increasing chain length and with decreasing polarity of the components. This also holds for the addition to allylphenol.

Longer-chain poly(ethylene glycol) or poly (propylene glycol), which lead to systems with hydrophilic to tensidic properties, are added in the form of their monoallyl ethers, mostly directly to the oligomeric sil(ox)anes (siloxanes–oxyalkylenes, etc.).

Hydroxyl groups can also be formed by reaction of *O*-nucleophiles, in the simplest case water, with epoxyalkylsilanes.

4.5.2. Acrylates and Other Ester Functions

As early as 1963, methacrylic-functional silanes were introduced as particularly effective modifiers for organic polymers containing carboxyl groups. These are produced almost exclusively by hydrosilylation of allyl methacrylate (Eq. 65) [24].

$$X_3Si{-}H + H_2C{=}CHCH_2OCO\overset{Me}{\overset{|}{C}}{=}CH_2 \xrightarrow{Pt} X_3Si(CH_2)_3OCO\overset{Me}{\overset{|}{C}}{=}CH_2 \quad (65)$$

This reaction leads to the desired 3-methacryloxypropyl group with good selectivity because of the considerably higher reactivity of the allyl group. The most important of these compounds, the trimethoxy compound [*2530-85-0*] (X = OMe), is produced industrially either by direct addition of trimethoxysilane or preferably via the corresponding trichlorosilyl

precursor [*7351-61-3*] (X = Cl) in the presence of radical scavengers [152]. The second method provides access to other alkoxy derivatives.

The reaction of γ-functional precursors with acrylates is an alternative. Chloropropylsil(ox)anes can, for example, be esterified directly with salts of acrylic or methacrylic acid. Such processes, which are carried out in solution or heterogeneously with phase-transfer catalysts, are particularly suitable for products richer in siliconbound methyl groups, which are less readily accessible from allyl acrylates. The formation of esters from hydroxypropyl precursors and acrylic acid or the comparable transesterification with methyl methacrylate in the presence of tin catalysts [153] is preferably carried out on linear siloxanes. According to patents, similar systems can be produced by reaction of aminofunctional sil(ox)anes with acrylic acid chloride [154].

Other oxygen-bonded esters include the acyloxypropylsil(ox)anes discussed in Section 4.5.1.

Alkyl-bonded esters, such as the addition products of acrylates and vinylacetic acid, are dealt with in Section 4.5.3.

4.5.3. Acid Anhydrides and Other Carboxy Groups

Carboxy compounds play an important role in industrial polymer chemistry because of their dipolar nature. The modification of such systems can take place by radical grafting with acrylate groups or by nucleophilic coupling with activated carboxyl groups.

Anhydride-functional silicon compounds are particularly suitable for this. The introduction of such side groups is generally carried out by hydrosilylation of appropriate alkenyl precursors. Whereas the addition of sil(ox)anes to 5-norbornene-2,3-dicarboxylic acid anhydride [*826-62-0*] leads only to mixtures of isomers that are difficult to purify [155], the analogous reaction with allylsuccinic acid anhydride [*7539-12-0*] gives readily distillable monomers. The easy-to-handle 3-triethoxysilylpropylsuccinic acid anhydride [*93042-68-3*] is of particular interest [156].

Carboxyorgano groups can also be introduced into silanes by palladium-catalyzed reaction of disilanes with aromatic acid chlorides (e.g., from 1,2,4-benzenetricarboxylic anhydride [*1204-28-0*]) [138].

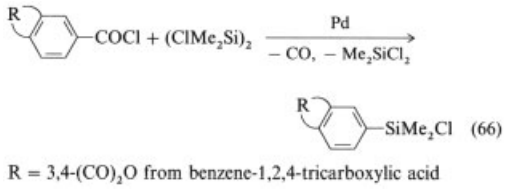

4-(Chlorodimethylsilyl)phthalic anhydride [*116088-82-5*], produced by this route, can be used as a stopper function for bonding longer-chain siloxanes to OH- or NH-functional organic blocks. Comparable structures are synthesized better by the [2 + 4] cycloaddition of maleic anhydride to the isoprenyl (butadienyl) compounds described in Section 4.1 [157].

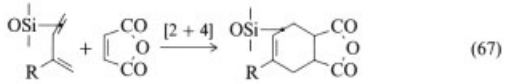

The hydrosilylation of acrylates has attracted renewed interest as a result of the use of silylketene acetals in group-transfer polymerization (GTP, see Section 6.5). The resulting reactive intermediates can be regarded as neutral silyl-transfer agents like the α-silyl esters, which contain a labile Si−C bond. Such compounds, which are relatively unstable compared to silicones, are formed along with others in the hydrosilylation of acrylic esters [158].

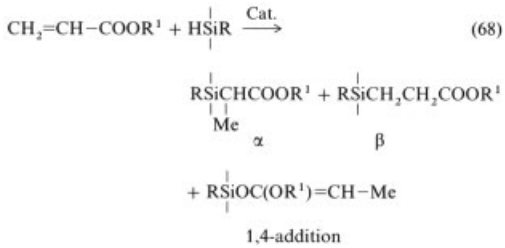

The product ratio depends on the ester group, the catalyst used, and especially on the starting silane and the stoichiometry. The stable β-adducts are mainly of interest in silicone chemistry.

Methacrylic esters give terminal products with considerably higher selectivity, particularly

when dichloromethylsilane is the H- component [24]. The resulting propionates can be equilibrated into organosiloxanes and subsequently saponified to the free acid, as is possible for cyanoalkyl groups under more forcing conditions.

Such carboxy-functional siloxanes have been produced from the corresponding silalactones [159]. The ester bond of the β-adducts is activated by the neighboring Si−Cl group, so that even catalytic quantities of tertiary or quaternary bases B are sufficient to liberate R^1Cl (usually chloromethane) according to Equation (69).

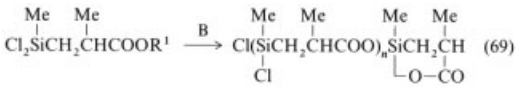

(69)

Above 130 °C the lactones gradually rearrange into oligomeric acid chlorides [160], which are also suitable for the production of propionate-functional organosiloxanes.

Longer- chain carboxy functions are accessible by hydrosilylation of the trimethylsilyl esters of vinylacetic acid, undeceneoic acid, etc. [161].

The organometallic introduction of acid groups via haloalkyl precursors and carbonyl compounds with activated CH groups, or by using magnesium and CO_2, appears too costly for industrial applications.

Direct nucleophilic substitution of haloalkylsilanes with carboxylates plays a significant role. Besides the monoesters or amides of dicarboxylic acids, derivatives of chloroacetic acid should be mentioned in particular here.

Treatment of chloroacetic acid derivatives with dialkylaminopropylsil(ox)anes gives betaine-modified products, which are intended for cosmetic and tenside applications [162]. Similar carboxy-functional products are also accessible via mercaptoacetic acid [163].

4.6. Other Organofunctions

Carbon-bonded phosphorus groups are particularly stable in the β-position or further from silicon. The phosphonate esters $\equiv SiCH_2CH_2-P(O)(OR)_2$ can, for example, be saponified with strong acids without cleaving the ethylene bridge between the heteroatoms [164]. These compounds, which are of major industrial interest (see Section 6.4), are produced by radical addition of phosphite esters to vinylalkoxysilanes (hydrophosphorylation). The reaction with diethyl phosphite in the presence of peroxide catalyst to give 2-(diethylphosphono)ethyltriethoxysilane [*757-44-8*] is carried out on a large scale [165]. Similar phosphonate esters can be obtained by the Arbuzov reaction of trialkyl phosphites with haloalkylsilanes.

Phosphine groups can also be introduced by radical addition (e.g., of $HPPh_2$ to vinylsilanes) or by substitution with alkali-metal phosphides [166]. Such ligands are suitable for fixing noble metals, particularly rhodium catalysts.

Organometallic groups bound to organosilyl groups directly or via carbon bridges generally occur only as reactive intermediates in the synthesis of the compounds discussed here. Examples are ferrocene, titanium, or tin complexes, numerous Grignard reagents, and the π/σ-bonded active catalysts of various hydrosilylation reactions. They are not described in detail here.

5. Other Organosilanes

5.1. Tetraorganosilanes

In tetraorganosilanes, four organic groups are bonded to silicon via carbon. Symmetrical compounds have four identical organic substituents, and unsymmetrical ones up to four different organic groups on silicon. Carbosilanes, compounds in which silyl groups are bonded, for example, via CH_2 groups also belong to this class [30].

Production. Symmetrical tetraorganosilanes are generally produced by treatment of SiX_4 (X = halogen, OR, H) with organometallic reagents [167]. Tetramethylsilane is formed in small quantities in the direct synthesis. It is also produced as the target product by treatment of Me_3SiCl with methyl Grignard. For the synthesis of unsymmetrical products, numerous processes have been proposed in which $SiCl_4$ or R_nSiX_{4-n} ($n = 1-3$, X = mostly halogen) is treated with mixtures of organometallic reagents; for examples, see [168].

Properties and Uses. Tetraorganosilanes are relatively inert chemically except for compounds bearing organofunctional groups. Pyrolytic decomposition of tetramethylsilane occurs only above 600 °C. This compound is attacked only slowly by concentrated sulfuric acid (Eq. 70).

$$Me_4Si \xrightarrow{H_2SO_4} Me_3SiOSiMe_3 + CH_4 \quad (70)$$

Tetraorganosilanes can, however, be used as alkylating agents [169] (Eq. 7).

$$R-\overset{\overset{O}{\|}}{C}-Cl + Me_3Si-C\equiv C-R^1 \xrightarrow{AlCl_3} R-\overset{\overset{O}{\|}}{C}-C\equiv C-R^1 + Me_3SiCl \quad (71)$$

Tetramethylsilane is used in coproportionation reactions [170] (see Section 2.6). It is also used as a reference in ^1H, ^{13}C, and ^{29}Si NMR spectroscopy and to fill ionization chambers in particle accelerators [171].

5.2. Disilanes

Chloromethyldisilanes $Me_nSi_2Cl_{6-n}$ ($n = 1-6$) are formed as side products in the direct synthesis (3 – 8 % in the crude silane). Only hexamethyldisilane [*1450-14-2*] is produced specifically and on a large scale from Me_3SiCl by the Wurtz reaction (Eq. 72).

$$2\,Me_3SiCl + 2\,Na \longrightarrow Me_6Si_2 + 2\,NaCl \quad (72)$$

To achieve complete conversion of sodium, many process variants have been proposed; see, for example, [172]. The electrochemical synthesis of Me_6Si_2 from Me_3SiCl is still not used in industry [173]. Most of the disilane fraction from the direct synthesis ($Me_nSi_2Cl_{6-n}$, $n = 1-6$, main products $n = 2-4$) is cleaved to give monosilanes. On acidic cleavage with hydrogen chloride, H- containing silanes are also obtained (Eq. 73).

$$Me_2ClSi-SiMeCl_2 + HCl \xrightarrow{Cat.} Me_2SiCl_2 + MeHSiCl_2 \quad (73)$$

Cleavage with amine- or phosphorus- containing catalysts gives chlorine- containing polysilanes as well as monosilanes [174]. Chlorine-free polysilanes can be produced as the target product by disproportionation of methoxymethyldisilanes in the presence of NaOMe [175].

The synthetic potential of disilanes is noticeably gaining interest in industry. Besides the reactions described above, a range of methods has been developed in which new Si–C bonds can be formed by catalytic cleavage of the Si–Si bond. An example is shown in Equation (66). For further examples, see [176].

5.3. Polysilanes

Organopolysilanes are compounds in which organosilyl groups are linked by Si–Si bonds. They are increasingly being used as starting materials for silicon carbide ceramics, fibers [177] (see also → Fibers, 5. Synthetic Organic, Chap. 3.2.2.), and composites derived therefrom [178]; as photoresists in microelectronics; and as photoinitiators. They also possess nonlinear optical properties. The production, properties, and uses of polysilanes are described in → Inorganic Polymers, Chap. 2. A more recent review can be found in [179].

6. Uses

The rapid development of industrial silane syntheses created economical access to completely new substances, reaction principles, and surprising applications. The possible uses of mono- and oligomeric organosil(ox)anes, apart from the production of polymeric silicones (see → Silicones), are given below (see also Chap. 4). Only some important uses are described here. Further information can be found in [123, 180].

6.1. Silylating Agents

Silanes have long been used in pharmaceutical research and production. One group of processes is characterized by the fact that silicon compounds enter only temporarily into the production cycle of the target product.

Silanes as Protecting Groups. In organic synthesis the presence of sensitive protic groups (OH, NH, COOH) can hinder reactions or make purification more difficult. These problems can be circumvented by temporarily substituting the protons using suitable silylating agents R_3SiX (X = halogen, amine, amide; see Table 5) [181]. The choice of the X group must take into consideration the possible pH sensitivity of the substrate molecule. The protecting group R_3Si can generally be introduced with high reactivity and selectivity. It can activate desired reactions or hinder undesired attack in another position. The derivatized molecule exhibits better solubility in organic solvents and increased volatility due to the absence of hydrogen bonds. After the desired reactions have been carried out, the protecting group is carefully removed by alcoholysis or hydrolysis, to finally form $Me_3SiOSiMe_3$ (HMDSO). To increase the overall economy and decrease environmental pollution, recycling of Me_3Si groups has been established since 1972 (Fig. 3). About 70 % of the HMDSO wastes returned to the silicone producers can be recovered. The cleavage of HMDSO to give chlorotrimethylsilane can, for example, take place by continuous reaction with dichlorodimethylsilane [89] or hydrogen chloride [182]. The more selective, nitrogen-containing, silylating agents that can be produced from it (Section 3.5) are quite inert in the pure state toward the groups to be protected (and also to moisture). By addition of acids or metal catalysts their activity is drastically increased under certain conditions.

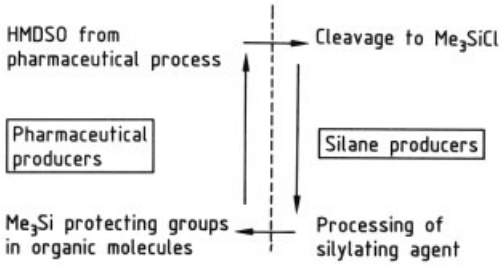

Figure 3. Recycling of Me_3Si protecting groups

One of the most important uses of trimethylsilanes is in the production of completely or semisynthetic β-lactam antibiotics; for this and the use of more recent silylating agents, see [75, 181, 183].

The use of silanes as reagents in organic synthesis has also been investigated intensively [8, 184]. The stereochemical behavior of organosilanes is particularly interesting [187]. Relatively little is known of the extent to which this has been exploited industrially.

Silanes for Enzyme Immobilization. Many biotechnological processes [186] involve the covalent anchoring of enzymatically active macromolecules on inert carriers (e.g., porous glass) over which the reaction medium is (continuously) passed. One variant of this is covalent bonding by using silane coupling agents [123] (see Section 6.7).

6.2. Physiologically Active Organosilicon Compounds

In physiologically active organosilicon compounds at least one silicon atom is a component of the active molecule itself. Because of the differences in reactivity between carbon and silicon atoms, in some cases the synthesis can be carried out more specifically and in higher yield than with the carbon analogue [187].

In many cases the silicon compound has modified activity. Provided hydrolysis of the SiOC bonds can occur, they are degraded more rapidly in the organism and thus permit a higher dosage with fewer side effects under certain circumstances.

Organosilicon Pharmaceuticals. The above-mentioned organosilanes exhibit specific physiological activity only in exceptional cases (Chapter 7). Following the relatively late discovery and investigation of silatranes **5** (from ca. 1960 onward) [25], which have a unique, broad spectrum of activity depending on the structure of the SiC-bonded group, a range of potential organosilicon pharmaceuticals has been systematically researched [188].

Table 5. Important silanes for introduction of protecting groups

Chemical name	Structure	CAS no.	Cleavage product
For standard applications			
Dichlorodimethylsilane	Me_2SiCl_2	[*75-78-5*]	acidic
Chlorotrimethylsilane ∗	Me_3SiCl	[*75-77-4*]	acidic
Hexamethyldisilazane ∗	$(Me_3Si)_2NH$	[*999-97-3*]	alkaline
N,N′-Bis(trimethylsilyl)urea ∗	$(Me_3SiNH)_2CO$	[*18297-63-7*]	neutral
For special applications			
N,O-Bis(trimethylsilyl)acetamide ∗	$CH_3C[{=}NSi(CH_3)_3]OSi(CH_3)_3$	[*10416-59-8*]	neutral
Iodotrimethylsilane ∗	Me_3SiI	[*16029-98-4*]	acidic
Hexamethyldisilane ∗	$Me_3SiSiMe_3$	[*1450-14-2*]	
tert-Butylchlorodimethylsilane	Me_3CSiMe_2Cl	[*18162-48-6*]	acidic

∗ Incorporated into industrial Me_3Si recycling.
∗∗ The highly reactive 1 : 1 molar mixture gives NH_4Cl.

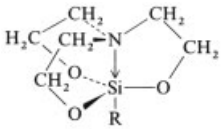

5

Only a few of these agents are used in the treatment of humans: Cisobitan (against prostate cancer) in Northern Europe; Migugen and Mival (hair growth and wound-healing agents) in the CIS.

Organosilicon Biocides. Tetraorganosilanes with specific substitution in (hetero)aromatic ring systems have achieved importance as biocides. The first important market product was the broad-spectrum fungicide Flusilazol (**6**) [189], which is soon to be followed by the insecticide Silafluofen (**7**) [190].

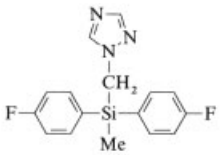

6 [*85509-19-9*]

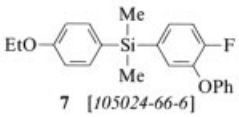

7 [*105024-66-6*]

The antimicrobial activity typical of ammonium compounds with longer-chain alkyl substituents is also exhibited by silanes such as $(RO)_3Si(CH_2)_3N^+MeR_1R_2Cl^-$. Here the Si function acts only as an anchor toward mineral substrates or organic fibers. The primary uses are in the treatment of textiles (carpets, sportswear) and the purification of water [191].

6.3. Silanes for Silicone Modification

A relatively small range of moisture-sensitive organosilanes is used as cross-linking agents in cold-cross-linking silicone sealants (RTV systems; see → Silicones, Chap. 4.4.3.). The most important compounds are listed in Table 6. The quantities used are several thousand tonnes per year. Silane coupling agents can also be used for cross-linking (see Section 6.7) [192].

Organofunctional silicones are obtained by introducing the corresponding silanes (see Table 7) into polysiloxanes by condensation or equilibration reactions. The generally fluid products are used in reprography, in textile finishing, and in polish additives.

In coating technology, specific oligosiloxanes that undergo cross-linking by cationic ring-opening polymerization are now used [193]. In the production of siliconized release paper, the reaction of dimethylpolysiloxanes or methylcyclosiloxanes containing 2-(3-epoxycyclohexyl)-ethyl groups is initiated by strong Brönsted acids. The latter are formed, for example, from diaryliodonium salts on UV irradiation [194]. The reaction is not inhibited by oxygen.

6.4. Silanes for Modification of Organic Polymers

Organosil(ox)anes are being used increasingly for improving the production and properties of polymers.

Cross-Linking of Polyethylene. In some applications the SiOH condensation mechanism,

Table 6. Important cross-linking agents for RTV silicone rubber *

Silane type	CAS no.	Cleavage product
Two-component systems		
$Si(OEt)_4$	[*78-10-4*]	Ethanol
Ethyl silicate 40	[*26352-16-9*]	Ethanol
One-component systems		
$MeSi(OMe)_3$	[*1185-55-3*]	Methanol
$H_2C{=}HCSi[OC(Me){=}CH_2]_3$	[*15332-99-7*]	Acetone
$Me/EtSi(OCOMe)_3$	[*4253-34-3*]/[*17689-77-9*]	Acetic acid
$MeSi[ONC(Me)Et]_3$	[*22984-54-9*]	Butanone-2-oxime
Mixture of oligomers		Diethylhydroxyl amine + *N*-methylacetamide
$MeSi(OEt)[N(CH_3)C(O)C_6H_5]_2$	[*16230-35-6*]	*N*-methylbenzamide
$MeSi(NHC_6H_{11})_3$	[*15901-40-3*]	Cyclohexylamine

* Difunctional analogues are chain extenders.

which is so successful for silicones, is used for cross-linking and better adhesion. Two industrial processes were introduced in the 1970s under the names Sioplas and Monosil, in which trimethoxyvinylsilane [*2768-02-7*] is grafted onto polyethylene (PE, sometimes modified) by means of dicumyl peroxide (DICUP) [195]. Although the industrial processes differ (Fig. 4), the reaction principle is identical. After molding, the silyl function bonded to the polymer via an ethylene bridge is hydrolyzed by external moisture (hot water bath, steam). The resulting silanol groups then undergo cross-linking, which is catalyzed by premixed dibutyltin dilaurate (DBTL) [*77-58-7*] (Eq. 74).

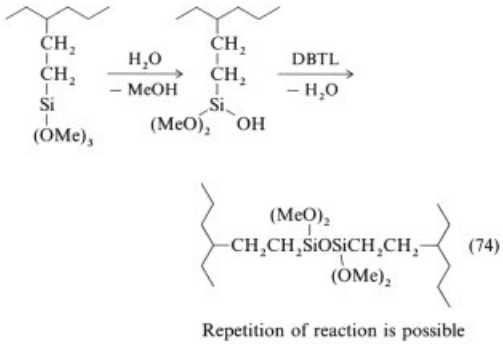

The Sioplas process separately produces two granulated polyethylene compounds in classical hot mixers. These are stored temporarily and can later be extruded as ca. 19 : 1 mixture with standard equipment. In a new process variant [196], vinylsilane is copolymerized with ethylene under high pressure. The resulting peroxide-free compound 1 (in Fig. 4) is very uniform and more stable to storage than grafted types. The cross-linked products are used in quantities of several hundred thousand tonnes per year as insulating materials, predominantly in low-voltage cables. They remain relatively stable dimensionally on thermal overloading. Their use in underfloor heating pipes or in self- cross-linking EPM or EDPM rubber articles is also common. The flameproofing of XL-PE with metal hydroxides is described in [197].

Cross-linking via silanol groups is also used in other polymers (e.g., vinyl acetate – acrylate dispersions [198], polyacrylates, or polyoxyalkylenes [199]) that are used as coatings, adhesives, or casting compounds.

Special Polymers. Unsaturated sil(ox)anes, in particular, have been proposed for copolymeric incorporation into standard polyolefins or polyacrylates. An important use is in raw materials for hard contact lenses, where a proportion of defined methacrylic-functional siloxanes [200] improves oxygen permeability. To enhance resistance to scratching, the surfaces of plastics are treated with silica-reinforced hydrolysates from, for example, $MeSi(OMe)_3$ (acrylic glass panes and consumer articles; polycarbonate spectacle lenses [201]). Polymers from short-chain or cyclic methylhydrosiloxanes, such as $[HMeSi(O)]_4$ [*2370-88-9*], with dicyclopentadiene and sometimes other diolefins, which are formed by platinum- catalyzed addition, are relatively new. They are recommended as nonstick, prehardened, impregnating resins for glass-fiber prepregs. The latter are laminated to

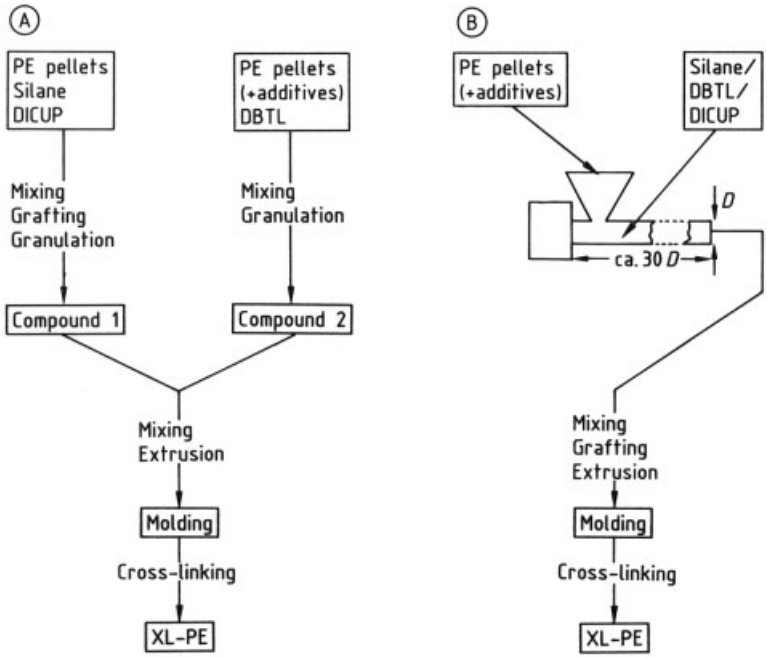

Figure 4. Polyethylene cross-linking (XL-PE)
A) Sioplas process (two stages); B) Monosil (one stage)

form carrier plates for printed electrical circuits [202]. The field of sil(ox)ane-modified organic polymers is described in detail in [203].

The rubber additive polysulfidesilane (Section 4.4.1 and Table 7) is both a coupling agent in mixtures with silicate fillers and a source of sulfur for rubber cross-linking. This unique, low-toxicity substance lowers the tendency to scorch (to undergo cross-linking too early), so that, for example, thick-walled tires for heavy-goods vehicles can be vulcanized without difficulty. In addition, it gives rubber a high reversion stability. In terms of quantity used, this is one of the most industrially important silanes [123].

Stabilizers. Silane polymer additives have a wide range of structures and effects. Incorporation of $(CH_3O)_3SiCH_2CH_2CH_2Cl$ [*2530-87-2*] or polysulfidesilane (see above) increases the thermal stability of filled EPDM rubber and of EPDM covulcanized with silicone rubber [204]. Phenylsil(ox)anols are suitable for flameproofing high-temperature polymers [205]. Sterically hindered phenolic or amine-containing alkoxysilane antioxidants bind migration-free to polymers and fillers owing to hydrolysis and condensation and thus ensure long-term stability [206]. Addition of (chlorine-free) organosilanes can delay the aging of PE insulating materials in a damp environment (treeing) [207]. Cyanacrylates adhere to deactivating surfaces (e.g., wood) more reliably after the addition of sila crown ethers [208].

6.5. Silanes in Catalysis

Silanes themselves are generally not catalytically active. However, their ability to form complexes with active metal compounds, whereby the catalyst system displays specific action, and the fixing of catalytically active centers to sil(ox)ane structures are of interest.

Donor Silanes. In the production of polypropylene with classical Ziegler – Natta catalysts the desired isotactic material is formed in unsatisfactory proportions. By using alkoxysilanes as external donors [209] an increase to $> 96\,\%$ isotactic material (heptane extraction) is achieved. The total catalyst requirement can be decreased to less than 1 : 20 000, so that separation from the polymer is unnecessary. Silanes $(CH_3O)_2SiRR^1$

with at least one sterically demanding group (e.g., R = cyclo-C_6H_{11}, R^1 = Me) [*17865-32-6*] have been shown to be the most effective donors. Bulk polymerization (Spheripol, Addipol, and Catalloy processes) and gas-phase processes (Unipol) are widely carried out by using these cocatalysts.

Siloxanes in Other Catalyst Systems. In high-temperature vulcanizing silicone rubbers, the Karstedt catalyst [60] is the most important platinum complex. It contains vinyl siloxane(s) as the complex-forming components, which give rise to the extremely high activity.

Mechanically tough, spherically uniform micro- and mesoporous particles can be produced by hydrolysis of suitable organofunctional silanes. They adsorb traces of metal or organic materials from (aqueous) solutions [210]. If they are coated with (noble) metal ions, highly active catalyst systems are formed [123]. Catalytically active groups can also be bonded to standard carriers via ambivalent silanes. For example, silane quats fixed on aluminosilicate (see Section 6.2) are excellent for the diproportionation of $MeHSiCl_2$ to MeH_2SiCl [*993-00-0*] or $MeSiH_3$ [*992-94-9*] [211].

In group-transfer polymerization [212] the reaction of α,β-unsaturated carbonyl monomers [e.g.,(meth)acrylates]is initiated by *O*-silylketene acetals and traces of certain anions at room temperature. This relatively new process works quasi-anionically and is a living polymerization, whose end products have a very narrow molecular mass distribution. They are used as paint additives.

6.6. Silanes for Use in Liquid Media

As a result of their high thermal stability, low pour points, and relatively flat temperature–viscosity curves, silicic acid esters of higher alcohols and peralkylated silanes have been proposed as high-performance lubricants for air- and spacecraft [213]. Phenylmethyltrisiloxanes are used as diffusion pump oils, and methylsilane glycol esters have been developed as brake fluids for auto racing [214]. For durability, modern car engines (specifically those made from light alloys) require corrosion inhibition on the interior metal walls. To prevent gel formation, various organofunctional silanes are added to the water–glycol cooling medium, whose pH is adjusted to slightly alkaline [215]. The addition of water-soluble organofunctional silanes to detergents [216] improves dirt removal from hard surfaces (rinsing agents) and lowers surface corrosion in enameled dishwashers, washing machines, and dryers. Effective wetting additives are of interest in agriculture, e.g., the glycol-modified trisiloxane $Me_3SiOSi(Me)[(CH_2)_3(OCH_2CH_2)_nOCH_3]OSiMe_3$ (n = ca. 8) [*87244-72-2* *]. It temporarily lowers the surface tension of water (at 0.1 % addition) to < 21 mN/m [217].

6.7. Silane Coupling Agents

Of all the organosilanes the silane coupling agents have resulted in the most important technological progress. From the early 1950s they were intensively developed, starting with the vinyl- and aminoalkoxysilanes, and they have influenced nearly all areas of technology, including the medical and analytical branches [114]. The field of silane coupling agents is covered by extensive literature [83, 84, 218, 219]. Characterization, synthesis, and some product-specific remarks on the use of coupling agents can be found in Chapter 4. Here, only the mode of action and the commercial products (Table 7) are mentioned.

Silane coupling agents are usually applied as prehydrolyzed, dilute (ideally 0.1–0.5 %, aqueous) solution, but also from the gas phase, by spraying, dipping, or direct mixing and (less often) in the form of primers (silane partial hydrolysates) on surfaces, fibers, or filler particles. Typical positively reacting inorganic [83] substrates are glass (fibers, hollow spheres), quartz (powder), (synthetic) silicas, wollastonite (calcium metasilicate), mica, kaolin, hydrated aluminum oxide, and among the metals, aluminum, steel, zinc, and copper (alloys) [220]. Additional complex-forming interactions of NH, SH, or COOH groups are believed to occur on metal surfaces.

The function Y must generally be adapted to the polymer type. Epoxy and aminoorganosilanes can be used quite universally. Reinforced or filled thermosets, thermoplastics, and elastomers become industrially and

Table 7. Organosilicon coupling agents in general use

Functionality	Chemical structure	CAS no.
Vinyl	$(MeO)_3SiCH{=}CH_2$	[2768-02-7]
	$(MeOC_2H_4O)_3SiCH{=}CH_2$	[1067-53-4]
Chloro	$(MeO)_3SiC_3H_6Cl$	[2530-87-2]
Monoamino	$(EtO)_3SiC_3H_6NH_2$	[919-30-2]
Diamino	$(MeO)_3SiC_3H_6NHC_2H_4NH_2$	[1760-24-3]
Mercapto	$(MeO)_3SiC_3H_6SH$	[4420-74-0]
Polysulfide	$[(EtO)_3SiC_3H_6]_2S_x$ (x = ca. 4)	[40372-72-3]
Epoxy	$(MeO)_3SiC_3H_6OCH_2CH{-}CH_2$ (O)	[2530-83-8]
	$(MeO)_3SiC_2H_4$	[3388-04-3]
Methacrylic	$(MeO)_3SiC_3H_6OCOC(CH_3){=}CH_2$	[2530-85-0]
Styrylcationic	$(MeO)_3SiC_3H_6NHC_2H_4NHCH_2$ $CH{=}CH_2 \cdot HCl$	[34937-00-3]

economically more competitive by using silane coupling agents. Producers' leaflets give more detailed information on the areas of use and handling. The following is valid as a chemical basis for the coupling-promoting effect of organofunctional silanes with the structure $(RO)_3SiCH{=}CH_2$ or $(RO)_3SiR^1Y$ (Table 7) at interfaces between mineral or metallic substances and organic polymers. Silanols formed by optimal hydrolysis (see Section 3.1) of functional alkoxysilanes condense with the elimination of water, and strong MOSi bonds are formed with the hydroxyl containing substrate surface (M = Si, Al, Fe, etc.) and by the silane molecules with one another. A flexible siloxane intermediate phase, occasionally of varying thickness, must be formed, whose functional groups on the side facing the polymer can become bonded to the organic phase (covalent chemical bonding and interpenetration). The basic mechanisms and their effect on the solution of coupling problems are currently being investigated physicochemically [221]. Nevertheless, relationships between theory and activity are not always clear, so a test of practical application must be the deciding factor. Occasionally, silanes also act between different polymers. Compounds with fewer than three OR leaving groups can also be active but are rarely used in practice (to some extent for cost reasons, cf. Section 4.2), although their more hydrophobic intermediate phase increases their resistance to water.

Mode of Action. Reaction of the silanol groups of the intermediate phase with the OH groups of the inorganic substrate strongly alters the polarity and wettability of the latter, so that the polymeric binding agents, adhesives, or coatings, which are generally of low polarity, can spread and adhere better. As a result of the breakup of hydrogen bonds, penetration of the boundary layer by the diffusion of surrounding moisture is strongly hindered. Thus, consumer articles made of such fiber- or particle-reinforced composites (boats, tanks for liquids, sports articles) have a considerably increased lifetime. Often the mechanical properties can be clearly improved, even in the dry state, and the filler content increased (more economical at the same performance level), or areas of use and temperature ranges can be extended. The electrical values (dielectric strength, loss factor tan δ) are higher and less affected by the entry of water.

At the performance level attained today, improvements in activity can be achieved with new silane structures only in individual cases. Furthermore, restrictive laws in industrial nations are forcing deeper study of interaction phenomena in the combined use of known basic materials [219, 222]. Problems associated with the binding of mineral fibers and fillers in bulk plastics such as polyethylene or polypropylene [223], the incorporation of nonoxide fillers such as calcium carbonate (organotitanates, for example, are used here), and the reinforcement of high-temperature polymers, such as polycarbonates, polyimides, or polysulfones, have still not been solved satisfactorily. The standard silanes (Table 7) tend to yellow above 130 – 180 °C and bring about structural degradation. More recent developments include temperature-stable aromatic silanes containing carboxyl groups, for

example [138], multifunctional silanes [224], stable aqueous silane solutions free of alcohol groups, or "polymeric silanes" [225]. If economic considerations are not taken into account, cationic, silanized, finely divided filler particles could become important as additives for washing powders, particularly for hard water [226].

6.8. Silanes for Hydrophobic Surface Treatment

The modification of (mineral) surfaces with nonfunctionalized silanes has been studied to a similar extent as the use of silane coupling agents [227]. The aim of treatment (generally with inexpensive alkylchloro- or alkylalkoxysilanes) is the elimination of polar HOM groups, the removal of adsorbed moisture, or improved wetting by organic media. Hydrophilic surfaces become hydrophobic and lipophilic. This is particularly effective for particle fillers, increasing the maximum degree of filling or improving mechanical, electrical, and chemical resistance. Silanized silicas attract much interest for the thixotropic modification of certain epoxide resins and especially for improving the mechanical properties of silicone rubber. A compound produced by cocondensation of $Si(OEt)_4$ and $Me_3SiOSiMe_3$ is claimed to exhibit particularly good properties [228]. Another example of this use of silicon compounds is in the protection of building materials of all types (masonry, plaster, natural stone, concrete) against weathering. Alkylalkoxysilanes [229] or silicone microemulsions [230] are used for this purpose, for example. Water-repellent silanization is applied to glass instruments for chemical and medical use (e.g., ampules) or in the production of light bulbs. The solvent-, dirt-, and fat-repellent (soil release effect) action of alkylsilyl groups is surpassed by certain fluoroalkylsilanes (see Section 4.2).

Small quantities of volatile sil(ox)anes (e.g., $Me_3SiOSiMe_3$) are used in plasma chemical vapor deposition. The extremely thin polymeric silicon-containing layers formed, protect, for example, vapor-deposited aluminum headlight reflectors [231] and are used for the production of microstructures on electronic chips [232].

6.9. Analysis

The analysis of silanes and silicones is described in [233]. In liquid chromatography (reversed phase HPLC) of mixtures of substances [234], stationary phases of defined polarity that are attached by stable bonds to the silica gel carrier are produced by using silanes (frequently octyl or octadecyl, but also organofunctional types). These phases are suitable for the separation of pesticides, protein components, or optical isomers and for ion chromatography. Information is provided by equipment suppliers. The methods for protective silylation described in Section 6.1 enable the analysis of sensitive organic (natural) products as well as short-lived silanols by GC or GC/MS (Lentz technique). Even mineral analyses are possible by using Me_3Si derivatization [235]. Traces of silicon in the environment can be characterized by $Me_3SiOSiMe_3$ – catalyst [233], and surfaces contaminated with silicones can be cleaned by using the same mixture.

7. Toxicology and Environmental Aspects

The physiological action of silanes is strongly structure dependent. An overall description is not possible, so in this chapter only some important aspects are discussed.

Tetraorganosilanes are nontoxic, provided they do not contain any toxic organofunctional substituents. However, for silicon-functional compounds, toxicity must always be a consideration, particularly when they contain readily hydrolyzable groups. This leads to attack on the mucous membranes and the eyes, particularly at high volatility. Organohalosilanes can also cause severe skin irritation or corrosion, due to the hydrogen halides liberated on exothermic hydrolysis and the associated localized dehydration. In the case of contamination, washing with copious amounts of water is indispensable as a first-aid measure.

Organoalkoxysilanes do not generally exhibit pronounced toxicity; however, the formation of alcohol during hydrolysis must be considered. The widely used tetraethoxysilane has a TLV of 10 ppm. In comparison, the volatile compounds

$Si(OMe)_4$ (TLV = 1 ppm) and $HSi(OMe)_3$ [85, 236] are highly toxic. On inhalation, severe headache and visual disorders occur, which can lead to permanent blindness. Liver and kidney damage has also been reported [237].

Organofunctional silanes are generally considered of low toxicity. Depending on the substituents, special effects can also occur [188]. In the case of physiologically active silicon compounds (Section 6.2) these properties are specifically exploited.

Widely varying toxicity can be found in series of silanes that contain both organofunctional and silicon-functional groups. The combination of amino and alkoxy groups in particular, present in silatranes, for example [25, 238], leads to quite pronounced toxicity in some cases.

A considerable proportion of physiologically active organosilicon compounds is processed in-house to polymers that are of low activity or harmless. During production and transport, attention must be given to safe handling of the monomers. Depending on the vapor pressure, flammability, corrosivity, etc., suitable precautions must be taken, which are set out in safety data sheets. In industry, halosilanes should be stored in double-walled steel tanks under an inert gas. Laboratory quantities are kept in glass vessels; plastic containers are generally unsuitable.

8. Economic Aspects

The main source of monomeric organosilicon compounds is the crude silane mixture produced by the direct synthesis. Its production has increased greatly during the last decade. In 1977 the quantity of crude silane was 380 000 t; in 1985, 600 000 t; and in 1991, ca. 10^6 t. About 95 % of this is used for the production of silicones. This explains why the largest silicone producers are also the most important producers of monomeric silanes. In the United States, these companies are Dow Corning, General Electric, and Union Carbide; in Europe, Bayer (Germany), Hüls (Germany), Rhône-Poulenc (France), and Wacker-Chemie (Germany); in Japan, Shin-Etsu, Toray (joint enterprise with Dow Corning), and Toshiba (joint enterprise with General Electric). Most of these companies operate worldwide; the assignment by country is made by head office. There are also production units in the CIS and in China.

According to estimates, ca. 40 % of the world market is accounted for by the United States, 30 % by Western Europe, 15 % by Japan, and 15 % by the rest of the world. The total turnover in silicones was ca. 7.5×10^9 DM in 1991 [239]. Besides the crude silane mixture, essentially only trichlorosilane is used as a starting material for other silanes, especially organofunctional silanes; ca. 30 000 t of $HSiCl_3$ is estimated to have been used for this purpose in 1991. The high-value silanes obtained by this route (price per kilogram between 5 and 30 €; for special silanes, even higher) are used, among other things, as coupling agents, cross-linking agents, and for surface treatment. The world market for these applications was estimated to be 5×10^8 € in 1991 (United States 46 %, Europe 46 %, Japan 8 %) [240]. These types of silanes are produced also by other companies besides silicone producers, such as PCR Chemicals (United States), Degussa (Germany), and Chisso (Japan).

Other applications of monomeric organosilanes are of comparatively minor economic importance. Silylating agents, which are mainly used as reagents in the pharmaceutical industry, had a market volume of 38×10^6 € (ca. 7000 t) in 1991. Organosilanes used as catalyst components in polypropylene production had a turnover of ca. 7.7×10^6 € (400 t) in 1991.

9. References

General References

1. A. R. Bassindale, P. P. Gaspar (eds.): *Front. Organosilicon Chemistry,* R. Soc. Chem., Cambridge, U.K., 1991.
2. S. Patai, Z. Rappoport (eds.): *The Silicon-Heteroatom Bond,* Wiley-Interscience, Chichester 1991.
3. S. Patai, Z. Rappaport (eds.): *The Chemistry of Organic Silicon Compounds,* **2 vols.,** Wiley-Interscience, Chichester 1989.
4. J. M. Zeigler, F. W. G. Fearon (eds): "Silicon-Based Polymer Science," *Adv. Chem. Ser.* **224** (1990) .
5. E. Lukevics, O. Pudova, R. Sturkovich: *Molecular Structure of Organosilicon Compounds,* Ellis Horwood, Chichester 1989.

6. J. Y. Corey, E. R. Corey, P. P. Gaspar (eds.): *Silicon Chemistry,* Ellis Horwood, Sussex 1988.
7. E. G. Rochow: *Silicon and Silicones,* Springer Verlag, Berlin 1987.
8. S. Pawlenko: *Organosilicon Chemistry,* W. de Gruyter, Berlin 1986.
9. H. Sakurai (ed.): *Organosilicon and Bioorganosilicon Compounds,* Ellis Horwood, Chichester 1985.
10. D. A. Armitage: "Organosilanes," in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* vol. 2, Pergamon Press, Oxford, 1982, pp. 3 – 204.
11. F. O. Stark, J. R. Falender, A. P. Wright: "Silicones," in G. Wilkinson, F. G. A. Stone, E. W. Abel (eds.): *Comprehensive Organometallic Chemistry,* vol. 2, Pergamon Press, Oxford 1982, pp. 306 – 363.
12. *Winnacker-Küchler,* 4th ed., **6,** 816 – 852.
13. *Houben-Weyl,* 4th ed., **13/5,** 1 – 502.
14. R. J. Voorhoeve: *Organosilanes, Precursors to Silicones,* Elsevier, Amsterdam 1967.

Specific References

15. C. Friedel, J. M. Crafts, *Justus Liebigs Ann. Chem.* **127** (1863) 28 – 32.
16. C. Friedel, J. M. Crafts, *Justus Liebigs Ann. Chem.* **136** (1865) 203 – 211.
17. A. Ladenburg, *Justus Liebigs Ann. Chem.* **173** (1874) 143 – 166; C. Pape, *Justus Liebigs Ann. Chem.* **222** (1884) 354 –374.
18. F. S. Kipping, *Proc. R. Soc. London Ser. A* **159** (1937) 139 – 147.
19. W. Dilthey, F. Eduardoff, *Ber. Dtsch. Chem. Ges.* **37** (1904) 1139 – 1142; F. S. Kipping, *Proc. Chem. Soc. London* **20** (1904) 15 – 16.
20. In [7] pp. 64 – 80.
21. General Electric, US 2 380 995, 1941 (E. G. Rochow).
22. VEB Silikonchemie, DD 5448, 1942 (R. Müller).
23. L. H. Sommer, E. W. Pietrusza, F. C. Whitmore, *J. Am. Chem. Soc.* **69** (1947) 1881; C. A. Burkhard, R. H. Krieble, *J. Am. Chem. Soc.* **69** (1947) 2687 – 2689.
24. J. L. Speier, J. A. Webster, G. H. Barnes, *J. Am. Chem. Soc.* **79** (1957) 974 – 979.
25. M. G. Voronkov, V. M. Dyakov, S. V. Kirpichenko, *J. Organomet. Chem.* **233** (1982) 1 – 147.
26. R. J. P. Corriu, J. C. Young in [2] pp. 1 – 66.
27. J. Y. Corey in [3] pp. 1 – 56.
28. In [14] pp. 186 – 219.
29. R. Müller, *Z. Chem.* **25** (1985) 309 – 318.
30. G. Fritz, *Angew. Chem.* **99** (1987) 1150 – 1171; *Angew. Chem. Int. Ed. Engl.* **26** (1987) 1111 – 1132.
31. In [14] pp. 244 – 282.
32. M. P. Clarke, *J. Organomet. Chem.* **376** (1989) 165 –222.
33. M. P. Clarke, I. M. T. Davidson, *J. Organomet. Chem.* **408** (1991) 149 – 156.
34. L. D. Gasper-Galvin, D. M. Sevenich, H. B. Friedrich, D. G. Rethwisch, *J. Catal.* **128** (1991) 468 – 478.
35. W. J. Ward, A. Ritzer, K. M. Carroll, J. W. Flock, *J. Catal.* **100** (1986) 240 – 249.
36. Dow Corning, US 4 898 960, 1986 (V. D. Dosaj, R. L. Halm, O. K. Wilding).
37. General Electric, US 2 464 033, 1949 (W. F. Gilliam).
38. Wacker Chemie, US 2 877 254, 1959 (E. Enk, S. Nitzsche).
39. General Electric, US 2 887 501, 1959 (B. A. Bluestein).
40. Union Carbide, US 4 973 725, 1989 (K. M. Lewis, R. A. Cameron, J. M. Larnerd).
41. Dow Corning, US 4 966 986, 1989 (R. L. Halm, R. H. Zapp).
42. M. G. R. T. de Cooker, J. H. N. de Bruyn, P. J. van den Berg, *J. Organometal. Chem.* **99** (1975) 371 – 377.
43. VEB Nünchritz, DD 250 536, 1986 (W. Walkow et al.).
44. Bayer, DE 3 841 417, 1988 (K. Feldner, B. Degen, G. Wagner, M. Schulze).
45. R. Schliebs, J. Ackermann, *Chem. Unserer Zeit* **21** (1987) 121 – 127.
46. J. Y. Corey in [3] p. 24.
47. Union Carbide, US 4 593 114, 1985 (M. L. Kenrick, B. Kanner).
48. H. Normant, C. R. Hebd. Seances Acad. Sci. **239** (1954) 1510 – 1512.
49. H. Gilman, E. A. Zuech, *J. Am. Chem. Soc.* **79** (1957) 4560 – 4561.
50. V. I. Zhun, M. K. Ten, *Khim. Promst. (Moscow)* 1989, 15 – 18; *Chem. Abstr.* **111** (1989) 153 903 g.
51. Dow Corning Toray, EP-A 405 560, 1990 (A. Shirahata). A. Shirahata, *Tetrahedr. Lett.* **30** (1989) 6393 – 6394.
52. Schering, EP-A 348 693, 1989 (J. Graefe, W. Uzick, U. Weinberg).
53. M. Weidenbruch, K. Kramer, *J. Organomet. Chem.* **291** (1985) 159 – 163.
54. H. Gilman, S. P. Massie, *J. Am. Chem. Soc.* **68** (1946) 1128.

55. M. G. Voronkov, N. G. Romanova, L. G. Smirnova, *Collect. Czech. Chem. Commun.* **23** (1959) 1013.
56. Dow Corning, US 4 888 435, 1989 (K. M. Chadwick, R. L. Halm, B. R. Keyes).
57. Union Carbide, US 2 637 738, 1953 (G. H. Wagner).
58. J. L. Speier, *Adv. Organometal. Chem.* **17** (1979) 407 – 447. B. Marciniec (ed.): *Comprehensive Handbook on Hydrosilylation,* Pergamon Press, New York 1992.
59. A. J. Chalk, J. F. Harrod, *J. Am. Chem. Soc.* **87** (1965) 16 – 21.
60. P. B. Hitchcock, M. F. Lappert, N. J. W. Warhurst, *Angew. Chem.* **103** (1991) 439 – 441; *Angew. Chem. Int. Ed. Engl.* **30** (1991) 438 – 440.
61. L. N. Lewis, R. J. Uriarte, N. Lewis, *J. Catal.* **127** (1991) 67 – 74.
62. L. N. Lewis, K. G. Sy, G. L. Bryant, P. E. Donahue, *Organometallics* **10** (1991) 3750 – 3759.
63. B. Marciniec, E. Máckowska, J. Guliúski, W. Urbaniak, *Z. Anorg. Allg. Chem.* **529** (1985) 222 – 228.
64. In [11] pp. 310 – 313.
65. I. Ojima in [3] pp. 1479 – 1526.
66. In [12] p. 827.
67. Hüls, EP-A 438 666, 1990 (W. Hange, H. Dietsche, C.-D. Seiler).
68. In [11] pp. 317, 318.
69. In [11] pp. 307 – 309.
70. V. D. Sheludrakov, V. I. Zhun', M. K. Ten, *Zh. Obshch. Khim.* **57** (1987) 567 – 571; *J. General Chem. USSR (Engl. Transl.)* **57** (1987) 495 – 499.
71. Bayer, DE-OS 3 410 644, 1984 (K. Feldner, W. Grape).
72. O. L. Flaningham in A. L. Smith (ed.) : *The Analytical Chemistry of Silicones,* Wiley & Sons, New York 1991, pp. 523 – 541.
73. A. L. Bassindale, P. G. Taylor in [3] pp. 839 – 892.
74. D. R. Weyenberg, L. G. Mahone, W. H. Atwell, *Ann. N. Y. Acad. Sci.* **159** (1969) 38 – 55.
75. G. A. Olah, G. K. Prakash, R. Krishnamurti in G. L. Larson (ed.): *Advances in Silicon Chemistry,* **vol. 1,** JAI Press, Greenwich 1991, pp. 1 – 64.
76. Ciba-Geigy, EP-A 177 454, 1985 (K. Oertle, H. Wetter).
77. General Electric, US 4 395 563, 1981 (S. E. Hayes).
78. Dow Corning, BP 914 460, 1961 (J. L. Speier).
79. J. L. Speier, D. L. Kleyer, *Organometallics* **10** (1991) 3046 – 3049.
80. Wacker Chemie, US 4 298 753, 1980 (A. Schinabeck et al.).
81. R. J. P. Corriu, C. Guerin, B. J. L. Henner, Q. Wang, *Organometallics* **10** (1991) 3200 – 3205.
82. Bayer, EP-A 4 310, 1979 (O. Schlak, H.-H. Moretto).
83. E. P. Plueddemann: *Silane Coupling Agents,* 2nd ed., Plenum Press, New York 1991 (and references cited therein).
84. F. D. Blum, W. Meesiri, H.-J. Kang, J. E. Gambogi, *J. Adhes. Sci. Technol.* **5** (1991) 479 – 496. F. D. Osterholtz, E. R. Pohl, *J. Adhes. Sci. Technol.* **6** (1992) 127 – 149.
85. G. J. Kallos, J. C. Tou, R. M. Malczewski, W. F. Boley, *Am. Ind. Hyg. Assoc. J.* **52** (1991) 259 – 262.
86. D. E. Leyden, J. B. Atwater, *J. Adhes. Sci. Technol.* **5** (1991) 815 – 829.
87. R. H. Chung, S. E. Hayes, *J. Organomet. Chem.* **265** (1984) 135 – 139; Dow Corning, US 4 732 996, 1987 (K. W. Moorhead, K. L. Reading, D. J. Rengering, A. P. Wright).
88. Toray, US 5 045 621, 1989 (T. Suzuki).
89. Wacker Chemie, US 4 113 760, 1977 (V. Frey et al.).
90. Wacker Chemie, DE-OS 3 716 372, 1987 (C. Trieschmann, J. Müller, K. H. Wegehaupt); Shin-Etsu, EP-A 435 654, 1990 (Y. Yamamoto, T. Matsuda).
91. P. V. Wright in K. J. Ivin, T. Saegusa (eds.): *Ring-Opening Polymerisation,* **vol. 2,** Elsevier, London/New York 1984, pp. 1055 – 1133. C. J. C. Edwards, R. F. T. Stepto in J. A. Semlyen (ed.): *Cyclic Polymers,* Elsevier, London/New York 1986, pp. 85 – 166.
92. Shin-Etsu, US 3 819 563, 1974 (T. Takago, T. Sato, H. Aoki). C. Rochin, O. Babot, R. Duboudin, *J. Organomet. Chem.* **281** (1985) C 24 – C 28; General Electric, US 4 210 596, 1978 (J. A. Cella).
93. R. J. Ayen, J. H. Burk, *Mater. Res. Soc. Symp. Proc.* **73** (1986) 801 – 808.
94. G. B. Goodwin, M. E. Kenney, *Inorg. Chem.* **29** (1990) 1216 – 1222.
95. Wacker Chemie, US 4 209 454, 1978 (W. Graf, V. Frey, P. John, N. Zeller).
96. M. Roth, *Bautenschutz + Bausanierung* **2** (1979) 12 –15.
97. E. le Coz, *DEFAZET Dtsch. Farben Z.* **32** (1978) no. 12, 56 – 60.
98. Wacker Chemie, US 4 176 130, 1978 (P. John, W. Feichtner, W. Graf, V. Frey). General Electric, US 4 329 484, 1981 (L. P. Petersen).

99. Rhone-Poulenc, US 4 918 209, 1988 (P. Baule, F. Chizat). Hüls, EP-A 381 840, 1989 (G. Zoche). Wacker Chemie, DE 4 104 725, 1991 (P. John, C. Braunsperger).
100. Allied, US 4 400 527, 1981 (C. T. Mathew, H. E. Ulmer).
101. K. C. Pande, R. E. Ridenour, *Chem. Ind. (London)* **1970,** 56. Shin-Etsu, JP 51 (76)–19 728, 1974 (M. Takamizawa, Y. Inone, H. Yoshioka); *Chem. Abstr.* **85** (1976) 33181 t.
102. D. A. Armitage in [2] pp. 367 – 394 and 448 – 466.
103. Union Carbide, US 4 400 526, 1982 (B. Kanner, C. L. Schilling, S. P. Hopper). General Electric, US 4 631 346, 1985 (J. L. Webb, C. E. Olsen).
104. General Electric, US 4 602 094, 1985 (T. D. Mitchell), and references cited therein. Bayer, US 4 739 088, 1986 (R. Endres, A. de Montigny).
105. Wacker Chemie, US 3 992 428, 1976 (H. Müller, I. Bauer, E. Schmidt, R. Riedle). Gist-Brocades, EP 43 630, 1981 (C. A. Bruynes, T. K. Jurriens).
106. Dynamit Nobel, DE-OS 3 443 960; DE 3 443 961, 1984 (H.-J. Kötzsch, H.-J. Vahlensieck).
107. Th. Goldschmidt, DE-AS 2 645 703, 1976 (G. Körner, H.-J. Patzke).
108. Th. Goldschmidt, EP-A 266 633, 1987 (G. Körner, G. Weitemeier, D. Wewers). Dow Corning, US 5 015 624, 1990 (W. J. Schulz).
109. J. Y. Corey in G. L. Larson (ed.): *Advances in Silicon Chemistry,* **vol. 1,** JAI Press, Greenwich 1991, pp. 327 – 387.
110. Dynamit Nobel, US 4 016 188, 1975 (H.-J. Kötzsch, H.-J. Vahlensieck).
111. E. L. Zichy, *J. Organomet. Chem.* **4** (1965) 411 – 412.
112. Imperial Chemical Ind. GB 1 152 251, 1966 (R. M. Gibbon, E. K. Pierpoint).
113. General Electric, US 3 441 583, 1965 (R. A. Murphy).
114. E. P. Plueddemann, *J. Adhes. Sci. Techn.* **5** (1991) 261 – 277.
115. Rhone-Poulenc, US 3 404 169, 1964 (M. Gaignon, M. Lefort).
116. Wacker Chemie, US 3 793 358, 1972 (S. Nitzsche, I. Bauer, W. Graf, N. Zeller).
117. T. K. Sarkar, *Synthesis* 1990, 969 – 983.
118. N. Furuya, T. Sukawa, *J. Organomet. Chem.* **96** (1975) C 1 – C 3.
119. H. Matsumoto et al., *J. Organomet. Chem.* **148** (1978) 97 – 106.
120. Dow Corning, US 4 609 574, 1985 (J. R. Keryk, P. Y. K. Lo, L. E. Thayer).
121. Wacker Chemie, EP-A 396 130, 1990 (C. Herzig, D. Gilch).
122. Wacker Chemie, US 5 041 594, 1991 (C. Herzig).
123. U. Deschler, P. Kleinschmit, P. Panster, *Angew. Chem.* **98** (1986) 237 – 253; *Angew. Chem. Int. Ed. Engl.* **25** (1986) 236 – 252.
124. In [11] pp. 314 – 316.
125. Union Carbide, US 4 658 050, 1986 (J. M. Quirk, B. Kanner).
126. T. Kobayashi, K. H. Pannell, *Organometallics* **10** (1991) 1960 – 1964.
127. Wacker Chemie, DE 3 144 020, 1981 (R. Artes, V. Frey, P. John, M. Scherer).
128. M. J. Gregory, *J. Chem. Soc., Perkin Trans. II* 1973, 1699 – 1702.
129. M. J. Owen, D. E. Williams, *J. Adhes. Sci. Techn.* **5** (1991) 307 – 320.
130. Union Carbide, US 2 930 809, 1956 (V. B. Jex, D. L. Bailey).
131. R. A. Benkeser, K. M. Foley, J. B. Grutzner, W. E. Smith, *J. Am. Chem. Soc.* **92** (1970) 697 – 698.
132. A. B. Rajkumar, P. Boudjouk, *Organometallics* **8** (1989) 549 – 550.
133. Union Carbide, US 3 595 897, 1968 (E. S. Brown, E. A. Rick, F. D. Mendicino).
134. Union Carbide, US 4 556 722, 1985 (J. M. Quirk, S. Turner).
135. Dynamit Nobel, DE-AS 2 749 316, 1977 (F.-R. Kappler, C.-D. Seiler, H.-J. Vahlensieck).
136. General Electric, DE-OS 3 546 376, 1985 (J. L. Webb, E. Cathryn).
137. H. Yinglin, H. Hongwen, *Synthesis* 1990, 122 –124.
138. J. D. Rich, *J. Am. Chem. Soc.* **111** (1989) 5886 –5893.
139. Dow Corning, US 3 560 385, 1968 (C. A. Roth).
140. J. E. Bittell, J. L. Speier, *J. Org. Chem.* **43** (1978) 1687 – 1688.
141. Dynamit Nobel, DE 3 346 910, 1983 (C.-D. Seiler, H.-J. Vahlensieck).
142. Union Carbide, DE-OS 2 300 912, 1973 (J. Y. Pui Mui).
143. Phillips Petroleum, US 3 890 213, 1973 (R. P. Louthan).
144. Degussa, DE-OS 3 226 093, 1982 (P. Panster, H. Grethe, P. Kleinschmit).
145. Wacker Chemie, DE-OS 4 135 170, 1991 (R. Hager, B. Deubzer, J. Wolferseder).
146. Minnesota Mining, US 4 267 213, 1979 (B. R. Beck, F. T. Sher, G. V. D. Tiers).

147. Wacker Chemie, DE-OS 4 135 142, 1991 (R. Hager, B. Deubzer).
148. Union Carbide, EP-A 262 642, 1987 (J. M. Quirk, B. Kanner).
149. Degussa, DE-OS 3 528 006, 1985 (U. Deschler, A. Grund, G. Prescher).
150. Wacker Chemie, DE 4 128 894, 1991 (C. Herzig, D. Gilch, J. Bindl).
151. P. R. Hudrlik, E. L. O. Agwaramgbo in [6] pp. 95 – 104.
152. Dow Corning, US 4 780 555, 1988 (H. M. Bank).
153. Rhône-Poulenc, US 4 940 766, 1987 (M. Gay, E. Canivenc).
154. Dow Corning, US 4 861 907, 1989 (A. P. Wright, D. J. Bunge).
155. S. A. Swint, M. A. Buese, *J. Organomet. Chem.* **402** (1991) 145 – 153.
156. Wacker Chemie, DE-OS 3 301 807, 1983 (T. Lindner, P. John, N. Zeller, R. Riedle).
157. Wacker Chemie, US 5 015 700, 1990 (C. Herzig, J. Esterbauer).
158. K. Takeshita et al., *J. Org. Chem.* **52** (1987) 4864 – 4868.
159. Dow Corning, US 4 788 313, 1983 (G. Chandra, D. R. Juen).
160. Dow Corning, US 4 329 483, 1981 (J. Speier).
161. Dow Corning, EP-A 196 169, 1986 (K. J. Woodward, R. M. Edmund).
162. Dow Corning, US 4 847 397, 1988 (F. Sawaragi, H. Taniguchi).
163. Dow Corning, US 4 599 438, 1983 (J. W. White, S. Westall, B. J. Griffiths).
164. G. H. Barnes, M. P. David, *J. Org. Chem.* **25** (1960) 1191 – 1194.
165. Wacker Chemie, DE-AS 2 219 983, 1972 (G. Künstle, H. Liberda, H. Spes).
166. M. Czakova, M. Capka, *J. Mol. Catal.* **11** (1981) 313 – 322.
167. In [13] pp. 31 – 78.
168. Ethyl Corp., US 4 711 965; 4 711 966, 1987 (G. E. Nelson).
169. In [8] pp. 124 – 127. J. Ipaktschi, A. Heydari, *Angew. Chem.* **104** (1992) 335 – 336; *Angew. Chem. Int. Ed. Engl.* **31** (1992) 331.
170. M. Bordeau, S. M. Djamei, R. Calas, J. Dunogues, *J. Organomet. Chem.* **288** (1985) 131 – 138.
171. S. Ochsenbein, D. Schinzel, A. Gonidec, W. F. Schmidt, *Nucl. Instrum. Methods Phys. Res., Sect. A* **273** (1988) , 654 – 656; *Chem. Abstr.* **110** (1989) 65467 s.
172. Rhône-Poulenc, EP-A 255 453, 1988 (J. S. Ferlet).
173. A. Kunai, T. Kawakami, E. Toyoda, M. Ishikawa, *Organometallics* **10** (1991) 2001 – 2003.
174. Dow Corning, US 4 534 948, 1981 (R. H. Baney).
175. Wacker Chemie, EP 214 664, 1986 (V. Frey, B. Pachaly, N. Zeller).
176. Y. Ito in [1] pp. 391 – 398.
177. J. Lipowitz, *Am. Ceram. Soc. Bull.* **70** (1991) 1888 – 1894.
178. K. S. Mazdiyasni (ed.): *Fiber Reinforced Ceramic Composites-Materials,* Processing and Technology, Park Ridge (USA), 1990.
179. R. West in [3] pp. 1207 – 1240.
180. A. Tomanek: *Silicone und Technik: Ein Kompendium für Praxis, Lehre und Selbststudium,* Hanser Verlag, München 1990. A. Tomanek: *Silicones and Industry: A Compendium for Practical Use, Instruction and Reference,* Hanser Verlag, München 1992.
181. H. Menzel: *Pharmaceutical Manufacturing International,* Sterling Publication, London 1989, pp. 129 – 134. P. Kochs, *Chem. Ztg.* **113** (1989) 225 – 238.
182. Dynamit Nobel, DE-OS 3 151 677, 1981 (H.-J. Kötzsch, H.-J. Vahlensieck).
183. M. Lalonde, T. H. Chan, *Synthesis* 1985, 817 – 845; H. Wetter, K. Oertle, *Tetrahedron Lett.* **26** (1985) 5515 – 5518.
184. W. P. Weber: *Silicon Reagents for Organic Synthesis,* Springer Verlag, Berlin 1983. R. Anderson, *Synthesis* 1985, 717 – 734. E. W. Colvin: *Silicon in Organic Synthesis,* R. E. Krieger Publishing, Malabar 1985. E. W. Colvin: *Silicon Reagents in Organic Synthesis,* Academic Press, London 1988. G. L. Larson in [2] pp. 763 – 808. J. K. Rasmussen in G. L. Larson (ed.): *Advances in Silicon Chemistry,* **vol. 1,** JAI Press, Greenwich (1991) pp. 65 – 187.
185. I. Fleming in [9] pp. 197 – 211. R. J. P. Corriu, C. Guerin, J. J. E. Moreau in [2] pp. 305 – 370. I. Fleming, *Pure Appl. Chem.* **62** (1990) 1879 – 1886.
186. A. Rosevear, J. F. Kennedy, J. M. S. Cabral (eds.): *Immobilized Enzymes and Cells,* Adam Hilger, Bristol 1987.
187. S. Barcza in [6] pp. 135 – 144.
188. R. Tacke, H. Linoh in [3] pp. 1143 – 1206. R. Tacke et al., *J. Organomet. Chem.* **417** (1991) 339 – 353.
189. Du Pont, US 4 510 136, 1983 (W. K. Moberg). W. K. Moberg, D. R. Baker, J. G. Fenyes, *ACS Symp. Ser.* **443** (1991) 1 – 14.

190. Hoechst, EP-A 224 024, 1986 (H. H. Schubert et al.). Dainippon, JP-Kokai 03 (91)–99 003, 1989 (K. Sugamoto, Y. Namite); *Chem. Abstr.* **115** (1991) 201146 u.
191. W. C. White, R. L. Gettings in D. E. Leyden (ed.): *Chemically Modified Surfaces,* **vol. 1,** Silanes, Surfaces, and Interfaces, Gordon and Breach, London 1986, pp. 107 – 140. K. J. Hüttinger, M. F. Jung, M. C. Schnell, *Chem.-Ztg.* **114** (1990) 161 – 165.
192. Wacker Chemie, US 4 801 673, 1987 (E. Bosch, F. Neuhauser, A. Schiller, O. Sommer).
193. J. Stein, R. P. Eckberg, *J. Coated Fabr.* **20** (1990) 24 –42.
194. J. V. Crivello, J. L. Lee, *J. Polym. Sci., Polym. Chem. Ed.* **28** (1990) 479 – 503. General Electric, EP-A 412 430, 1990 (J. V. Crivello, J. L. Lee).
195. H.-G. Fritz, S. Ultsch in Verein Dt. Ingenieure (ed.): *Polymerreaktionen und reaktives Aufbereiten in kontinuierlichen Maschinen,* VDI-Verlag, Düsseldorf 1988, pp. 243 – 295. S. Ultsch, H.-G. Fritz, *Plast. Rubber Process Appl.* **13** (1990) 81 – 91.
196. Mitsubishi, EP 193 317, 1986 (I. Ishino, A. Ohno, T. Isaka). Nippon Oil, US 4 412 042, 1982 (K. Matsuura, N. Noboru, M. Miyoshi).
197. Nippon Unicar, EP-A 370 518, 1979 (K. Horita, S. Hayashi, T. Koshijama).
198. Wacker Chemie, DE-AS 2 148 456 – 2 148 458, 1971 (E. Bergmeister, P.-G. Kirst, H. Wiest).
199. Kanegafuchi, US 4 837 401, 1985 (R. Hirose, S. Yukimoto, K. Isayama). Kanegafuchi, US 5 011 900, 1989 (S. Yukimoto, T. Hirose, H. Wakabayashi, K. Isayama).
200. B. Arkles, *CHEMTECH* **13** (1983) 542 – 555.
201. Seiko Epson, US 5 015 523, 1987 (H. Kawashima, M. Nakashima, T. Mogami).
202. Hercules, US 5 008 360, 1989 (J. K. Bard, J. S. Burnier). Hercules, US 5 013 809, 1989 (R. T. Leibfried).
203. I. Yilgör, J. E. McGrath, *Adv. Polym. Sci.* **86** (1988) 1 – 86. W. Gardiner, J. W. White : "High Value Polym.," *Spec. Publ. R. Soc. Chem.* **87** (1991) 98 – 108.
204. Shin-Etsu, US 4 201 698, 1978 (K. Itoh, T. Oshima). Mitsui, EP-A 314 396, 1988 (T. Tojo, K. Okamoto, A. Matsuda, E. Louis).
205. General Electric, EP-A 415 072 and A 415 073, 1990 (L. N. Lewis et al.).
206. Enichem Sintesi, EP-A 162 523 and A 162 524, 1985 (A. Greco, L. Cassar, C. Neri et al.).
207. Licentia, DE-OS 3 628 554, 1986 (J. Wartusch). Licentia, DE-OS 3 702 209, 1987 (J. Wartusch, H. Andreß, W. Gölz).
208. Loctite, US 4 906 317, 1985 (J.-C. Lin).
209. M. Härkönen, J. V. Seppälä, T. Väänänen, *Makromol. Chem.* **192** (1991) 721 – 734.
210. Degussa, EP-A 416 271 and A 416 272, 1990 (P. Panster et al.).
211. Wacker Chemie, US 4 870 200, 1988 (R. Ottlinger, A. Rengstl, R. Jira).
212. D. Y. Sogah, W. B. Farnham in [9] pp. 219 – 230.
213. V. K. Gupta et al., *Lubr. Eng.* **46** (1990) 706 –711. K. J. L. Paciorek et al., *Ind. Eng. Chem. Prod. Res. Dev.* **30** (1991) 2191 – 2194.
214. Castrol, US 3 994 948, 1974 and US 4 141 851, 1978 (H. F. Askew, C. J. Harrington, G. J. J. Jayne).
215. Dow Chemical, EP-A 111 013, 1982 (R. T. Jernigan). Korea advanced Institute (KAIST), US 4 873 011, 1988 (I. N. Jung, S. Y. Hwang, C. S. Lee).
216. Procter and Gamble, US 4 005 024/25/28/30, 1975 (D. C. Heckert, D. M. Watt et al.); EP 75 986-90, 1982 (C. R. Barrat, J. R. Walker, J. Wevers, H. Ernst).
217. M. Knoche, H. Tamura, M. J. Bukovac, *J. Agric. Food Chem.* **39** (1991) 202 – 206.
218. V. Chvalovsky in V. Chvalovsky, J. M. Belama (eds.): *Organosilicon Compounds,* Plenum Press, New York 1984, pp. 1 – 33.
219. G. Tesoro, Y. Wu, *J. Adhes. Sci. Technol.* **5** (1991) 771 – 784.
220. P. Walker, *J. Adhes. Sci. Technol.* **5** (1991) 279 –305.
221. K. P. Hoh, H. Ishida, J. L. Koenig, *Polym. Compos.* **9** (1988) 151 – 157; D. J. Ondrus, F. J. Boerio, K. J. Grannen, *J. Adhes.* **29** (1989) 27 – 42. W. J. van Ooij, A. Sabata, *J. Adhes. Sci. Technol.* **5** (1991) 843 – 863. E. Nishio, N. Ikuta, H. Okabayashi, *J. Anal. Appl. Pyrolysis* **18** (1991) 261 – 268.
222. P. G. Pape, E. P. Plueddemann, *J. Adhes. Sci. Technol.* **5** (1991) 831 – 842.
223. PCR, US 4 975 509, 1988 (W. G. Joslyn, A. D. Ulrich, M. E. Wilson).
224. Toshiba, DE 4 010 128, 1990 (H. Motegi, T. Sunaga, M. Zanbayashi). Toshiba, JP-Kokai 03 (91)–188 085, 1989 (H. Mogi, T. Sunaga, M. Zenbayashi); *Chem. Abstr.* **116** (1992) 83922 q.
225. B. Arkles, J. Steinmetz, J. Zazyczny, P. Mehta, *J. Adhes. Sci. Technol.* **6** (1992) 193 – 206.
226. E. P. Plueddemann in D. E. Leyden, W. T. Collins (eds.): *Chemically Modified Surfaces,* **vol. 3,** "Chemically Modified Oxide Surfaces," Gordon and Breach, London 1990, pp. 281 – 294.

227. D. E. Leyden, W. T. Collins (eds.): *Chemically Modified Surfaces,* **vol. 2, "Chemically Modified Surfaces in Science and Industry,"** Gordon and Breach, London 1988; H. A. Mottola, J. R. Steinmetz (eds.), *Chemically Modified Surfaces,* **vol. 4,** Elsevier, Amsterdam 1992, supplements the series of monographs on silanes on surfaces cited in [196, 226]
228. K. E. Polmanteer, H. L. Chapman, M. A. Lutz, *Rubber Chem. Technol.* **58** (1985) 939 – 974.
229. K. M. Roedder., *Bautenschutz + Bausanierung* **10** (1987) 143 – 148.
230. H. Mayer, M. Roth, *Bautenschutz + Bausanierung* **13** (1990) pp. 1 – 4.
231. G. Benz, *Bosch Tech. Ber.* **8** (1986/87) 219 – 226.
232. J. N. Helbert, N. Saha, *J. Adhes. Sci. Technol.* **5** (1991) 905 – 925. E. Babich et al., *Microelectron. Eng.* **13** (1991) 47 –50.
233. A. L. Smith (ed.): *The Analytical Chemistry of Silicones,* J. Wiley & Sons, New York 1991.
234. W. Cheng, M. McCown, *J. Chromatogr.* **318** (1985) 173 – 185. K. K. Unger (ed.): *Handbuch der HPLC,* GIT Verlag, Darmstadt 1989. K. Kimata, N. Tanaka, T. Araki, *J. Chromatogr.* **594** (1992) 87 – 96.
235. A. M. Dunster, J. R. Parsonage, E. A. Vidgeon, *Mat. Sci. Technol.* **5** (1989) 708 – 713.
236. G. B. Kolesar et al., *Fundam. Appl. Toxicol.* **13** (1989) 285 – 295.
237. In [8] pp. 130 – 134 and references cited therein.
238. M. A. Horsham, C. J. Palmer, L. M. Cole, J. E. Casida, *J. Agric. Food Chem.* **38** (1990) 1734 – 1738.
239. D. Hunter, *Chem. Week,* Feb. 19 (1992) 24 – 25.
240. *Compatibilizers,*Hewin International Inc., Amsterdam 1991.

Silicones

Hans-Heinrich Moretto, Bayer AG, Leverkusen, Federal Republic of Germany

Manfred Schulze, Bayer AG, Leverkusen, Federal Republic of Germany

Gebhard Wagner, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 2
2. Linear and Cyclic Polyorganosiloxanes 3
2.1. Production 3
2.1.1. Hydrolysis 3
2.1.2. Methanolysis 4
2.1.3. Cyclization 4
2.1.4. Polymerization 5
2.1.5. Polycondensation 7
2.1.6. Industrial Production of Linear Polysiloxanes 8
2.2. Polydimethylsiloxanes 8
2.3. Siloxane-Based Copolymers 8
3. Silicone Fluids 10
3.1. Methylsilicone Fluids 11
3.2. Methylphenylsilicone Fluids 11
3.3. Other Types of Silicone Fluids . . . 11
3.4. Properties 11
3.5. Formulation 14
3.6. Applications 14
4. Silicone Rubbers and Elastomers . 15
4.1. General Properties 15
4.2. Rubber Compounds 15
4.3. Rheology 17
4.4. Curing Systems 17
4.4.1. Radical Curing with Peroxides 17
4.4.2. Hydrosilylation Curing 17
4.4.3. Condensation Curing 18
4.4.4. Radiation Curing 19
4.4.5. Oxidative Coupling 19
4.5. Peroxide-Cured High-Temperature Vulcanizing Silicone Rubbers 19
4.6. Liquid Silicone Rubbers 20
4.7. Room Temperature Curing Silicone Rubbers 21
4.7.1. Two-Component RTV Systems 21
4.7.2. One-Component RTV Systems 22
4.8. Paper and Textile Coatings 23
4.8.1. Paper Coating 23
4.8.2. Textile Coating 24
4.9. Properties of Silicone Elastomers . 24
4.10. Applications 26
5. Silicone Resins 27
5.1. Structure and General Properties . 27
5.2. Production 28
5.3. Curing 28
5.4. Properties 29
5.5. Applications 29
6. Block and Graft Copolymers 29
6.1. Polysiloxane – Polyether Copolymers 30
6.2. Other Block Copolymers 30
6.3. Graft Copolymers 31
6.4. Applications of Block and Graft Copolymers 32
7. Analysis 32
8. Toxicology 34
9. Environmental Aspects 35
10. Economic Aspects 35
11. References 36

1. Introduction

Nomenclature and Structure. The term silicones is used for compounds in which silicon atoms are linked via oxygen atoms, each silicon atom bearing one or several organic groups. In industrially important silicones, these groups are usually methyl or phenyl. The silicones are known as polyorganosiloxanes according to IUPAC rules.

The structure of the industrially important "methylsilicones" can involve the units listed in Table 1.

Polymer structures can be described by using the letters M, D, T, and Q to designate the monomer units. Linear silicone fluids are composed mainly of D units. The base polymers for silicone elastomers or silicone rubbers consist of D units that bear cross-linkable functional groups. The main structural feature of the highly branched silicone resins are T units, often combined with D units to make the resins more flex-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_057

ible. Silicone resins can also contain Q and M units.

Table 1. Origin, functionality, and fields of application of silicone structural units

Starting compound	Structural unit		Application
	Symbol	Functionality	
$(CH_3)_3SiCl$ Monochlorotrimethylsilane	M	$H_3C-Si(CH_3)_2-O_{0.5}$ monofunctional	chain end in silicone fluids, trimethylsilyl protecting groups
$(CH_3)_2SiCl_2$ Dichlorodimethylsilane	D	$H_3C-Si(CH_3)-O_{0.5}$ $O_{0.5}$ difunctional	linear siloxane polymers, silicone fluids, rubbers, elastomers
CH_3SiCl_3 Trichloromethylsilane	T	$H_3C-Si(O_{0.5})-O_{0.5}$ $O_{0.5}$ trifunctional	silicone resins for paints, impregnating agents, masonry protection
$SiCl_4$ Tetrachlorosilane	Q	$O_{0.5}-Si(O_{0.5})-O_{0.5}$ $O_{0.5}$ tetrafunctional	silicone resins

General Properties. Silicones have many outstanding properties, which are described in detail for the individual product groups. In general, methylsilicones exhibit greater stability to high temperature, UV radiation, and weathering than organic polymers; marked surface-active behavior (low surface tension, high spreading power); good dielectric properties, as well as low temperature dependence of their physical properties.

2. Linear and Cyclic Polyorganosiloxanes

2.1. Production

Linear and cyclic polyorganosiloxanes are generally produced by reacting organodichlorosilanes with water. The mixture of oligomeric siloxanes arising from hydrolysis can be converted either entirely to cyclic siloxanes (e.g., octamethylcyclotetrasiloxane) or directly polymerized to linear polysiloxanes. A special case of the production of cyclic or linear oligomeric dimethylsiloxanes that has gained in importance is the methanolysis of dimethyldichlorosilane.

The most important chlorosilanes used industrially (> 90 % of the total) are the methylchlorosilanes. Other chlorosilanes that contain silicon-bound H, C_6H_5, CH_2=CH, or CF_3CH_2-CH_2 groups, either exclusively or in combination with methyl groups, are produced in smaller quantities. Chlorosilanes and siloxanes containing other organic ligands, such as C_2H_5 or $HOCH_2$, are not discussed in detail here because of their minor importance.

In the following sections, the processes of hydrolysis, methanolysis, polymerization, and polycondensation are described in more detail; the production of oligomeric polydimethylsiloxanes from dimethyldichlorosilane is used as an example throughout.

2.1.1. Hydrolysis

The complete hydrolysis of dimethyldichlorosilane leads to an oligomer mixture consisting of cyclic dimethylsiloxanes and hydroxyl-terminated dimethylsiloxanes:

$$n\,(CH_3)_2SiCl_2 + n\,H_2O \longrightarrow [(CH_3)_2SiO]_n + 2n\,HCl \quad n = 3, 4, 5, \text{etc.}$$

and simultaneously:

$$m\,(CH_3)_2SiCl_2 + (m+1)\,H_2O \longrightarrow HO[(CH_3)_2SiO]_m\,H + 2m\,HCl \quad m = 4 -> 100$$

Hydrolysis with a deficiency of water gives linear dimethylsiloxanes with terminal SiCl groups:

$$(n+2)\,(CH_3)_2SiCl_2 + (n+1)\,H_2O \longrightarrow Cl(CH_3)_2SiO[(CH_3)_2SiO]_n Si(CH_3)_2Cl + 2(n+1)\,HCl$$

Complete hydrolysis with excess water is carried out continuously in the liquid phase with ca. 25 % hydrochloric acid or in the gas phase at ca. 100 °C. Liquid-phase hydrolysis gives cyclic and linear oligomeric dimethylsiloxanes in the approximate ratio 1: 1 to 1: 2, depending on reaction conditions, together with ca. 30 % hydrochloric acid. The hydrochloric acid can be reacted with methanol to produce chloromethane,

which is used for producing methylchlorosilanes in the Rochow synthesis (→ Silicon Compounds, Organic, Chap. 2.1.) [1–4] (chlorine recycling).

The ratio of cyclic to linear dimethylsiloxanes and the chain length of the linear oligomers can be varied over a relatively wide range by means of hydrolysis conditions. For example, rapid removal of hydrochloric acid from the reaction mixture by neutralization leads almost exclusively to short-chain siloxanediols. Cyclic siloxanes represent up to two-thirds of the reaction product if prolonged contact with HCl occurs. The tetramer octamethylcyclotetrasiloxane is the predominant cyclic siloxane.

Control of the hydrolysis reaction to give predominantly cyclic or linear oligomeric dimethylsiloxanes is important because high-molecular mass polydimethylsiloxanes are produced both by equilibrating polymerization and by polycondensation. The preferred starting materials for equilibrating polymerization are cyclic siloxanes, while polycondensation is only possible with hydroxyl-terminated oligomers.

The hydrolysis process can also be used to produce organosiloxanes modified with functional groups. For example, the reaction of a mixture of dimethyldichlorosilane, methylhydrogendichlorosilane, and trimethylchlorosilane with excess water affords a trimethylsilyl end-stopped linear polymethylsiloxane with a random distribution of dimethylsiloxy and methylhydrogensiloxy units. The ratio of the two difunctional methylsiloxane units in the mixture can be varied at will. The average chain length of the siloxanes formed decreases with increasing trimethylchlorosilane content in the silane mixture.

$$n\,(CH_3)_2SiCl_2 + m\,CH_3(H)SiCl_2 + 2(CH_3)_3SiCl + (n+m+1)\,H_2O \longrightarrow (CH_3)_3SiO[(CH_3)_2SiO]_n[CH_3(H)SiO]_m Si(CH_3)_3 + 2(n+m+1)\,HCl$$

The hydrolysis product mixture is influenced to a certain degree by the addition of organic solvents. This method is especially useful for obtaining siloxanes that contain silanol groups and silicone resins (see Section 5.2).

Figure 1 shows a schematic of the hydrolysis process.

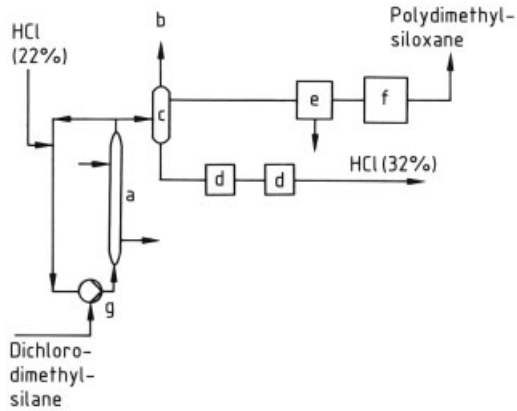

Figure 1. Continuous hydrolysis of dichlorodimethylsilane in a circulation apparatus
a) Cooler; b) Exhaust; c) Phase separation; d) Settling vessel; e) Water separator; f) Neutralization, g) Pump

2.1.2. Methanolysis

The methanolysis process for the production of siloxanes from dimethyldichlorosilane allows direct recovery of the chlorine from methylchlorosilanes as chloromethane. The silane reacts with methanol to give oligomeric dimethylsiloxanes and chloromethane. Two variants of this process exist, one of which leads to linear hydroxyl end-stopped oligomeric siloxanes, whereas the other affords mainly cyclic siloxanes. In the former, siloxanes are removed from the bottom of the reaction column, and the lower-boiling cyclic compounds in the overhead are returned to the process; in the latter, volatile cyclic oligomers are removed continuously from the reaction mixture by distillation.

In both cases the methanolysis plants are operated in association with methylchlorosilane production. Figures 2 and 3 show how methanolysis or hydrolysis of dimethyldichlorosilane is integrated with the Rochow synthesis and the plants for further processing.

Methanolysis occurs according to the following overall equations:

$$n(CH_3)_2SiCl_2 + 2n\,CH_3OH \longrightarrow HO\,[(CH_3)_2SiO]_nH + 2n\,CH_3Cl + (n-1)H_2O$$

$$n(CH_3)_2SiCl_2 + 2n\,CH_3OH \longrightarrow [(CH_3)_2SiO]_n + 2n\,CH_3Cl + n\,H_2O$$

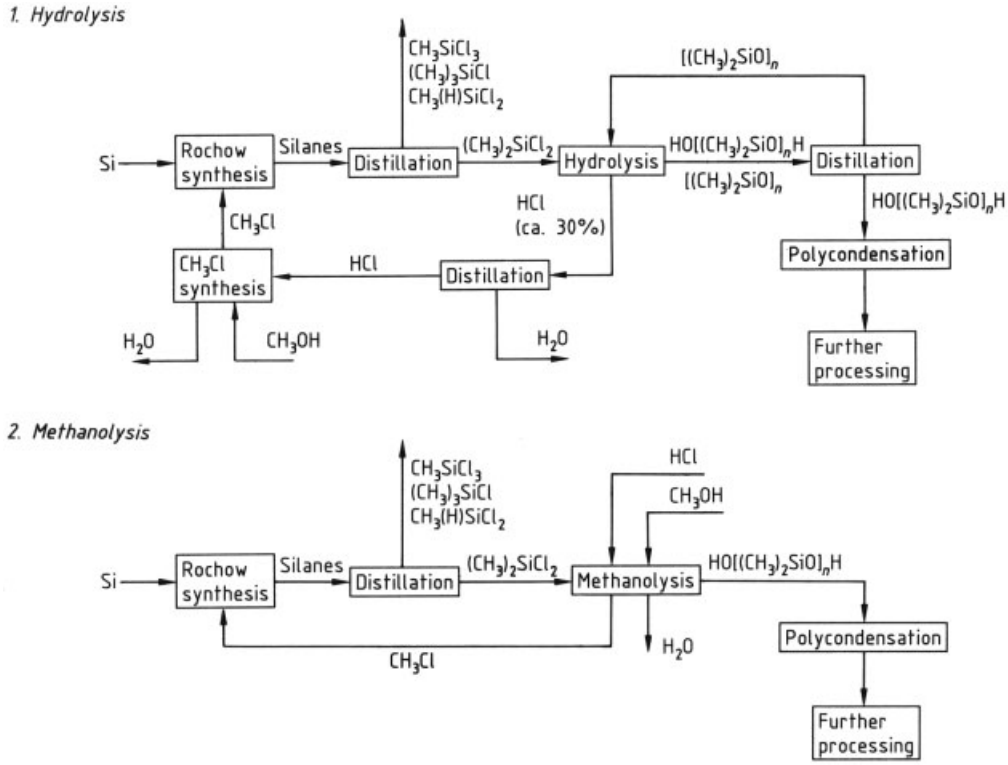

Figure 2. Integration of the hydrolysis or methanolysis of dimethyldichlorosilane via intermediate linear dimethylsiloxanes in silicone production

Dimethyl ether formed as a byproduct is purged from the process or allowed to react with hydrochloric acid to form chloromethane, depending on the process variant.

If the synthesis and methanolysis of dimethyldichlorosilane are regarded as a single process unit, then overall, dimethylsiloxane is formally produced from silicon and methanol, with water as the only byproduct. Recycling of chlorine as chloromethane avoids HCl waste, making silicone production cleaner and more efficient.

Rochow synthesis:

$$n\,\mathrm{Si} + 2n\,\mathrm{CH_3Cl} \longrightarrow n\,(\mathrm{CH_3})_2\mathrm{SiCl_2}$$

Methanolysis:

$$n\,(\mathrm{CH_3})_2\mathrm{SiCl_2} + 2n\,\mathrm{CH_3OH} \longrightarrow [(\mathrm{CH_3})_2\mathrm{SiO}]_n + 2n\,\mathrm{CH_3Cl} + n\,\mathrm{H_2O}$$

Overall:

$$n\,\mathrm{Si} + 2n\,\mathrm{CH_3OH} \longrightarrow [(\mathrm{CH_3})_2\mathrm{SiO}]_n + n\,\mathrm{H_2O}$$

2.1.3. Cyclization

Pure cyclic siloxanes are produced by cyclization. Octamethylcyclotetrasiloxane and decamethylcyclopentasiloxane are major industrial products, which are either marketed as such or used for the production of polydimethylsiloxanes. Cyclization is performed by heating the hydrolysis or methanolysis mixture with potassium hydroxide. Process aids are used to prevent polymerization of the siloxanes to high-viscosity liquids. Potassium hydroxide catalyzes an equilibration reaction in which the Si–O–Si bonds are cleaved and reformed. During the reaction, the lower-boiling octamethylcyclotetrasiloxane and decamethylcyclopentasiloxane are distilled continuously from the reaction mixture. The siloxanes are constantly reformed to maintain the equilibrium until the siloxane mixture is completely converted to the desired cyclic siloxanes.

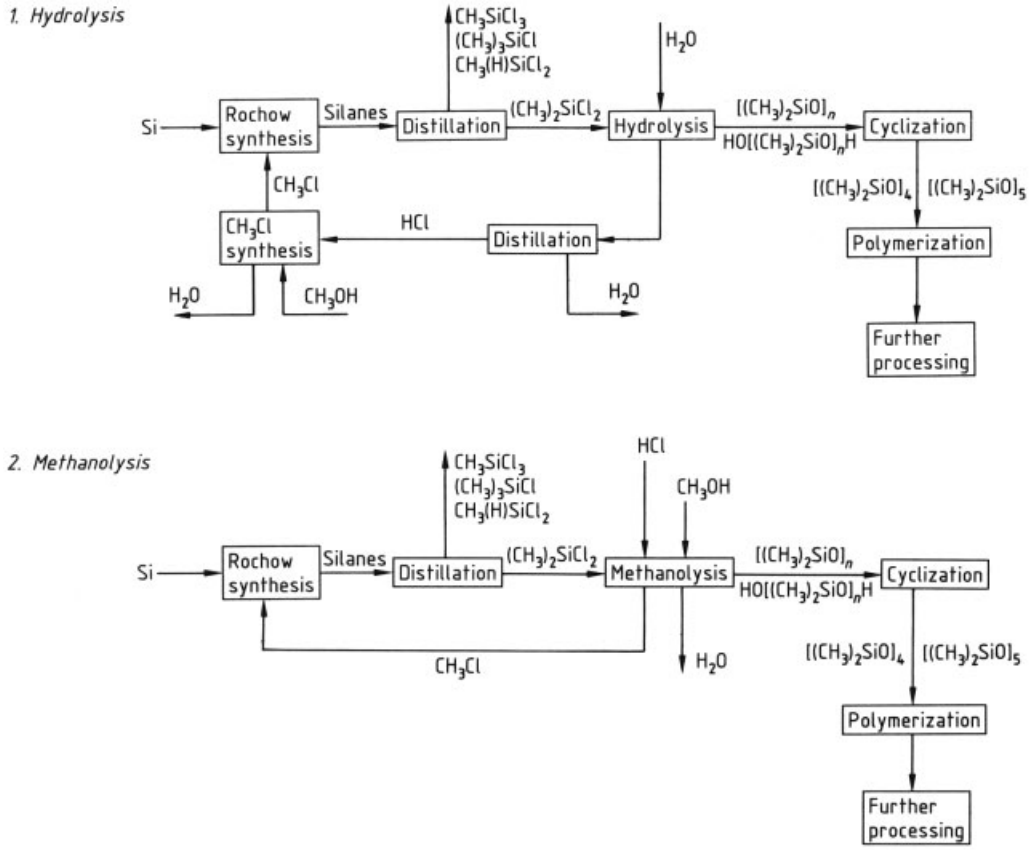

Figure 3. Integration of the hydrolysis or methanolysis of dimethyldichlorosilane via intermediate cyclic dimethylsiloxanes in silicone production

2.1.4. Polymerization

Linear polyorganosiloxanes can be prepared from cyclic organosiloxanes by equilibrating ring-opening polymerization, which is promoted by both anionic and cationic catalysts [5,6]. Octamethylcyclotetrasiloxane is the preferred starting material. Other siloxanes are used for chain end-stopping or for producing copolymers.

Anionic polymerization with alkali-metal hydroxide catalysts is of industrial importance. The catalytic activity decreases in the series Cs > Rb > K > Na > Li [7]. KOH is the most common catalyst. Rapid polymerization is observed above 140 °C on adding as little as a few ppm KOH, e.g., as a suspension in octamethylcyclotetrasiloxane. The polymerization mechanism involves initial formation of potassium siloxanolate, which then catalyzes chain growth and equilibration via cleavage of Si – O – Si bonds.

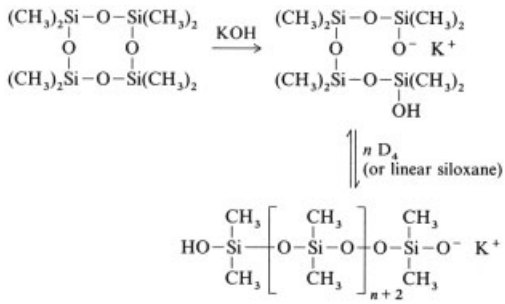

The polymerization reaction leads to an equilibrium mixture of linear polysiloxanes and ca. 15 – 18 wt % cyclic siloxanes. The average chain length of the linear polysiloxanes is controlled by the ratio of end groups to D units in the polymerization mixture. Various substances (called

regulators) added during polymerization determine the end groups. A Poisson distribution of molecular mass is obtained.

For example, hydroxyl-terminated polydimethylsiloxanes are formed in the presence of water. The average molecular mass depends on the amount of water added. If trimethylsiloxy-containing siloxanes (e.g., MD_2M) are added to the polymerization mixture then trimethylsilyl-terminated polymethylsiloxanes are formed.

Cation- complexing additives such as poly(ethylene glycol), crown ethers, methyl ethyl ketone, dimethylformamide, and dimethyl sulfoxide accelerate the reaction [5, 8] by promoting the dissociation of inactive silanolates:

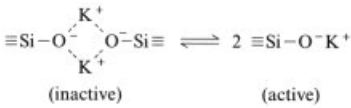

After the molecular mass distribution has reached equilibrium, the catalyst must be deactivated. Numerous deactivation methods have been described. KOH is generally neutralized with phosphoric acid or chlorosilanes. Careful neutralization is essential for the stability of the polymer, since both alkaline and acid residues lead to degradation of the siloxane chain by depolymerization. After neutralization, the volatile low molecular mass constituents (mostly cyclic compounds) are removed by distillation and recycled to the polymerization process.

Cationic polymerization of cyclosiloxanes is carried out with strong protic or Lewis acids. Industrially important catalysts are perfluoroalkanesulfonic acids or sulfuric acid [5, 6]. The reaction mechanism has not yet been fully elucidated.

Another method of cationic polymerization employs acidic solids such as ion-exchange resins and acid-activated silicates as catalysts. Cationic polymerization also leads to an equilibrium mixture of linear polysiloxanes.

2.1.5. Polycondensation

Linear oligomeric dimethylsiloxanes from the hydrolysis or methanolysis process can be polymerized by polycondensation. This process requires the use of high-purity dimethyldichlorosilane if strictly linear siloxane polymers are to be produced, since purification of the siloxanediols by distillation is not possible.

Polycondensation is carried out as a batch or continuous process in the presence of acid catalysts, preferably polychlorophosphazenes $(PNCl_2)_x$ [9–11]. The water produced in the reaction must be removed. When the desired chain length is reached, the catalyst is deactivated with ammonia or an amine [12, 13].

Polycondensation of siloxanediols proceeds very rapidly at elevated temperature under vacuum (Fig. 4, curve a). In the presence of a short- chain, $R(CH_3)_2SiO$-terminated dimethylsiloxane ($R = CH_3$, $CH{=}CH_2$, H) that regulates the chain length via equilibration, a combination of polycondensation and slower equilibrating polymerization occurs, and the system therefore passes through a marked viscosity maximum (Fig. 4, curve b). With a siloxanol as chain-length regulator, the final viscosity is reached more quickly (Fig. 4, curve c) because only condensation steps occur.

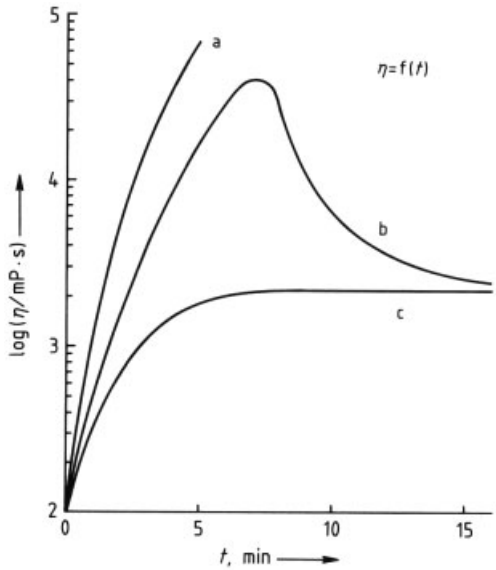

Figure 4. Polycondensation – equilibration of siloxane diols
a) $HO[(CH_3)_2SiO]_nH$; b) $HO[(CH_3)_2SiO]_nH + R(CH_3)_2SiO[Si(CH)_2O]_xSi(CH_3)_2R$;
c) $HO[(CH_3)_2SiO]_nH + R(CH_3)_2SiO[Si(CH_3)_2O]_xH$
$R = CH_3$, $CH = CH_2$

$R(CH_3)_2SiO$-terminated polydimethylsiloxanes can be obtained by polycondensation

Table 2. Polymerization processes

Starting material	D_4	D_4	$HO[(CH_3)_2SiO]_nH$
Catalyst	KOH	H_2SO_4	$(PNCl_2)_x$
Amount of catalyst, ppm	5 – 20	100 – 1000	5 – 200
Polymerization time, min	10 – 90	15 – 30	10 – 20
Reaction temperature, °C	140 – 180	20 – 160	40 – 160
Volatile oligomers in the product mixture, %	13	13	2
Regulator	MD_xM	MD_xM	$[R(CH_3)_2Si]_2NH$
Neutralization	H_3PO_4	ZnO Na_2CO_3	amines

and equilibration of the siloxanediols with $R(CH_3)_2$ SiO-terminated dimethylsiloxanes or by polycondensation of the siloxanediols followed by silylation of the silanol end groups with hexaorganodisilazanes, $R(CH_3)_2SiNHSi(CH_3)_2R$, for example.

2.1.6. Industrial Production of Linear Polysiloxanes

For small-volume products, the polymerization of oligomeric siloxanes is carried out in stirred vessels; batches up to 15 t are readily controllable. Larger quantities are produced in continuous plants (Fig. 5).

Depending on the type of polymerization reaction, the following process steps:

1) Purifying and drying the starting materials
2) Catalyst and regulator metering
3) Establishing equilibrium and condensation
4) Neutralization
5) Oligomer removal and distillate recycling

pose different technical requirements, which are summarized in Table 2.

When the variants are considered, five process designs for performing the polymerization have essentially prevailed:

1) Single-stage polymerization vessel
2) Stirred-tank cascade
3) Screw extruder reactor
4) Cell reactor (tubular reactor with spiral stirrers that generate approximately plug flow)
5) Solid (catalyst) reactor

2.2. Polydimethylsiloxanes

Linear polydimethylsiloxanes are the most important industrial polysiloxanes. The polymers are classified according to their viscosity (i.e., average chain length) and the nature of the end groups. The most important polymer types characterized by their end groups are listed in Table 3.

The endgroups determine the use. For example, trimethylsilyl-terminated polydimethylsiloxanes are typical silicone fluids. Hydroxy- and vinyl-terminated polymers find major application in silicone rubbers.

Table 3. Types of polymer

End group	Structure
Methyl	$(CH_3)_3Si-O-[(CH_3)_2Si-O-]_x-Si(CH_3)_3$
OH	$HO-(CH_3)_2Si-O-[(CH_3)_2Si-O-]_x-Si(CH_3)_2-OH$
Vinyl	$CH_2{=}CH-(CH_3)_2Si-O-[(CH_3)_2Si-O-]_x-Si(CH_3)_2-CH{=}CH_2$
H	$H-(CH_3)_2Si-O-[(CH_3)_2Si-O-]_x-Si(CH_3)_2-H$

2.3. Siloxane-Based Copolymers

Siloxane copolymers containing other siloxy groups in addition to dimethylsiloxy groups are used to obtain special properties. With a few exceptions, these copolymers have the general structure

$$R(CH_3)SiO[(CH_3)_2SiO]_x[(R^1)(R^2)SiO]_ySi(CH_3)_2R$$

The substituents R, R^1, and R^2 can be identical or different. The most important D units are:

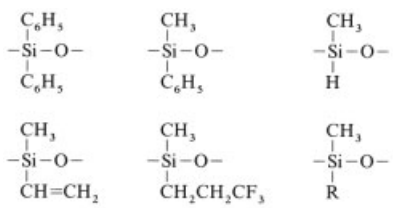

R = C_2–C_{14} alkyl

Table 4. Important types of silicone fluid

Structural formula	Brief designation
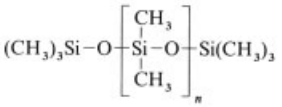	methylsilicone fluids $n = 2-4000$
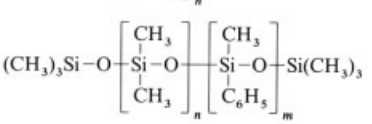	methylphenylsilicone fluids $n = 0-1000$, $m = 1-1000$
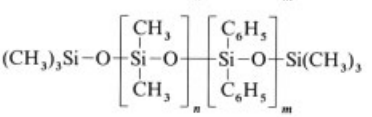	methylphenylsilicone fluids $n = 0-1000$, $m = 1-1000$
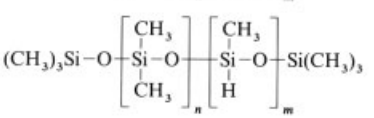	methylhydrogensilicone fluids $n = 0-100$, $m = 1-100$
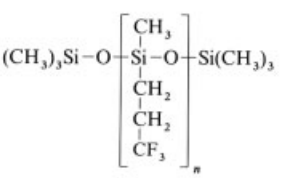	fluorosilicone fluids $n = 10-100$
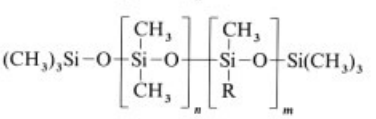	methylalkylsilicone fluids $n = 0-100$, $m = 1-100$

Table 5. Physical properties of polydimethylsiloxanes ∗

Property	Average molecular mass					
	600	1800	5800	26 000	62 000	160 000
Viscosity (25 °C), mPa · s	3	10	100	1000	12 500	500 000
Viscosity – temperature coefficient	0.55	0.57	0.60	0.61	0.61	0.61
Density (25 °C), g/cm^3	0.90	0.94	0.97	0.97	0.97	0.97
Flash point (DIN 51 376), °C	62	170	300	320	> 320	> 320
Pour point, °C	−100	−90	−50	−50	−50	−40
Refractive index, n_D^{25}	1.390	1.398	1.402	1.403	1.404	1.404
Thermal conductivity (150 °C), W m^{-1} K^{-1}	0.10	0.14	0.16	0.17	0.17	0.17
Specific heat (20 °C), J g^{-1} K^{-1}	1.50	1.50	1.50	1.50	1.50	1.50
Thermal expansion coefficient (0 – 150 °C), cm^3 cm^{-3} K^{-1}	11.4×10^{-4}	10.3×10^{-4}	9.9×10^{-4}	9.9×10^{-4}	9.9×10^{-4}	9.9×10^{-4}
Surface tension, mN/m	19.3	19.9	20.9	21.2	21.4	21.5
Dielectric constant (25 °C, 50 Hz)	2.5	2.6	2.6	2.7	2.8	2.8
Dielectric strength, kV/mm	13	14	14	14	15	15
Resistivity (25 °C), Ω · cm	2×10^{14}	2×10^{14}	2×10^{14}	2×10^{14}	2×10^{14}	2×10^{14}
Loss factor tan δ (25 °C, 50 Hz)	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}	5×10^{-5}

∗ Some of these are approximate values that can vary depending on the producer.

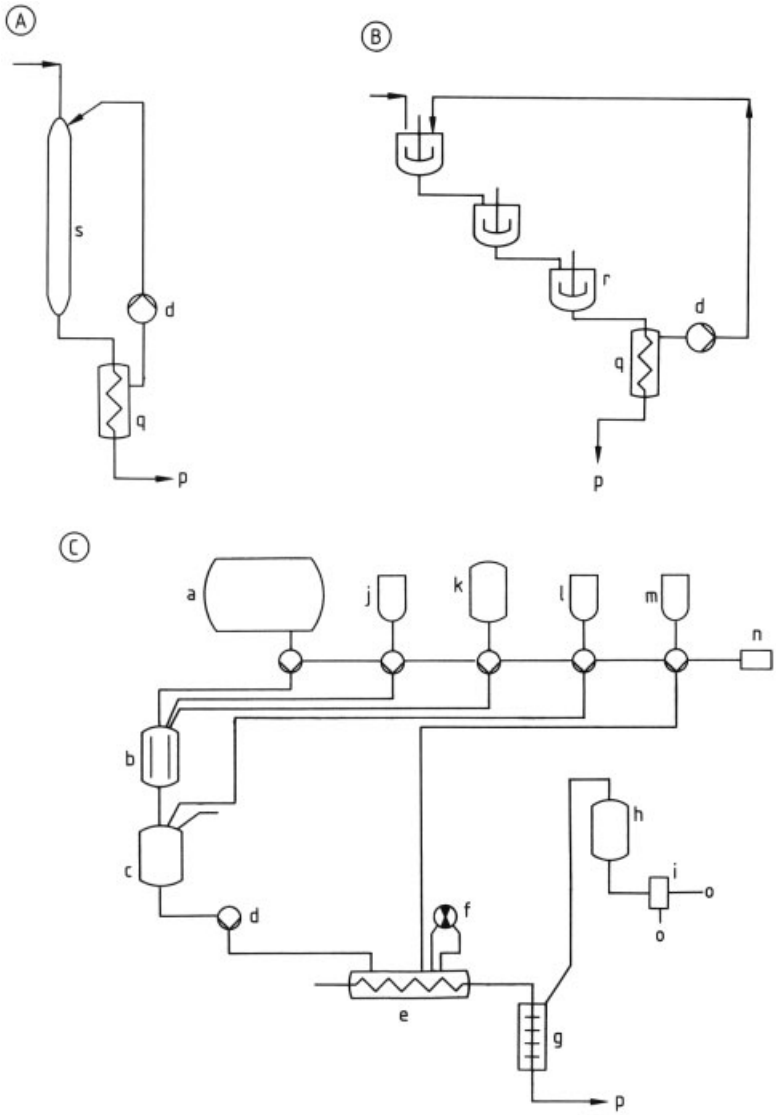

Figure 5. Industrial processes for production of silicone polymers
A) Fixed-bed catalysis; B) Stirred-vessel cascade; C) Continuous process in a screw mixer
a) D_4 cyclosiloxane vessel; b) Dryer; c) Static premixer; d) Metering pump; e) Screw mixer; f) Process viscometer; g) Degassing extruder; h) Condenser; i) Vacuum pump; j) Vessel for regulator; k) Vessel for vinylmethylsiloxane; l) Vessel for potassium siloxanolate; m) Vessel for phosphoric acid; n) Control of metering pumps; o) Recycle to c; p) Finished polymer; q) Falling-film evaporator; r) Vessel; s) Ion-exchange column

The copolymers are usually produced by copolymerization of the appropriately substituted cyclosiloxanes and D_4 under equilibrating conditions. In this way, the various D groups are distributed randomly along the polymer chain. Exceptions are block copolymers that are produced by anionic polymerization under nonequilibrating conditions.

Polymers with trifluoro groups are also prepared by the kinetically controlled polymerization of $[CH_3(CF_3CH_2CH_2)SiO]_3$, since 96 wt % cyclics are formed at equilibrium [14].

For example, polydimethyl – polydiphenyl block copolymers are produced by stepwise polymerization of D_3 and $(DPh_2)_3$ with Li- containing bases in tetrahydrofuran. The resulting

block copolymers have a twophase morphology and elastomeric properties.

3. Silicone Fluids

The structure of linear silicone fluids can generally be described by the composition MD_xM ($x = 2-4000$).

Silicone fluids are distinguished from common organic fluids by a number of unique properties:

1) Good thermal stability (150 – 250 °C)
2) Good low-temperature performance (<−70 °C)
3) Strong hydrophobicity
4) Excellent release properties
5) Antifriction and lubricating properties
6) Pronounced surface activity
7) Good dielectric properties
8) Very good damping behavior
9) Good radiation resistance
10) High solubility of gases
11) Physiological inertness
12) Low temperature dependence of physical properties

These properties can be modified over a wide range by varying the organic substituents. The most important silicone fluids are listed in Table 4.

3.1. Methylsilicone Fluids

The most important silicone fluids are the methylsilicone fluids (polydimethylsiloxanes, PDMS). Silicone fluids exhibit a chain-length distribution. The average chain length largely determines the viscosity. Low molecular mass volatile constituents are removed during production, which increases the flash point of the final product. Fluids with viscosities ranging between 1 and 10^6 mPa · s are commercially available.

The most important physical properties of polydimethylsiloxanes are listed in Table 5, which shows that the physical properties of silicone fluids depend on the molecular mass only up to a certain degree of polymerization. With increasing molecular mass, they reach a limiting value.

3.2. Methylphenylsilicone Fluids

Methylphenylsilicone fluids exhibit higher thermal stability, better low-temperature properties, more powerful solvent action for organic substances, and better lubricating properties than PDMS. Their radiation resistance is also particularly high. The physical properties of methylphenylsilicone fluids of high and low phenyl content are listed in Table 6.

Table 6. Physical properties of methylphenylsilicone fluids

	Baysilone fluid PN 200 ∗	Baysilone fluid PH 1000 ∗
Ratio of phenyl to methyl groups	4.5	1.5
Viscosity (25 °C), mm^2/s	200	1000
Viscosity temperature coefficient	0.74	0.89
Density (25 °C), g/cm^3	1.03	1.08
Flash point, °C	ca. 300	ca. 300
Pour point, °C	ca. − 65	ca. − 30
Refractive index, n_D^{25}	1.471	1.512
Thermal conductivity (50 °C), $W\,m^{-1}\,K^{-1}$	0.13	0.13
Specific heat, $J\,g^{-1}\,K^{-1}$	1.56	1.60
Thermal expansion coefficient (25 – 180 °C), K^{-1}	8.7×10^{-4}	7.8×10^{-4}
Surface tension, mN/m	23	25
Dielectric constant (25 °C, 50 Hz)	2.9	2.8
Dielectric strength, kV/mm	15	15
Loss factor tan δ (25 °C, 50 Hz)	1×10^{-4}	3×10^{-4}

∗ Trademarks of Bayer.

3.3. Other Types of Silicone Fluids

Methylhydrogensilicone fluids have silicon –hydrogen groups that react with protic compounds with the formation of hydrogen. Inorganic and organic surfaces having reactive groups, OH or NH_2 groups (such as glass), for example, can be modified. As a result these surfaces acquire completely new properties (see Section 3.6).

Fluorosilicone fluids have gained importance due to their outstanding low-temperature lubricating properties, low solubility in mineral oils, and the unique properties of fluorosilicone-treated surfaces.

Methylalkylsilicone fluids containing C_2 – C_{14} alkyl groups have better lubricating properties than pure PDMS. Like the methylphenylsiloxanes, they have better compatibility with organic substances (e.g., solvents or organic paint systems).

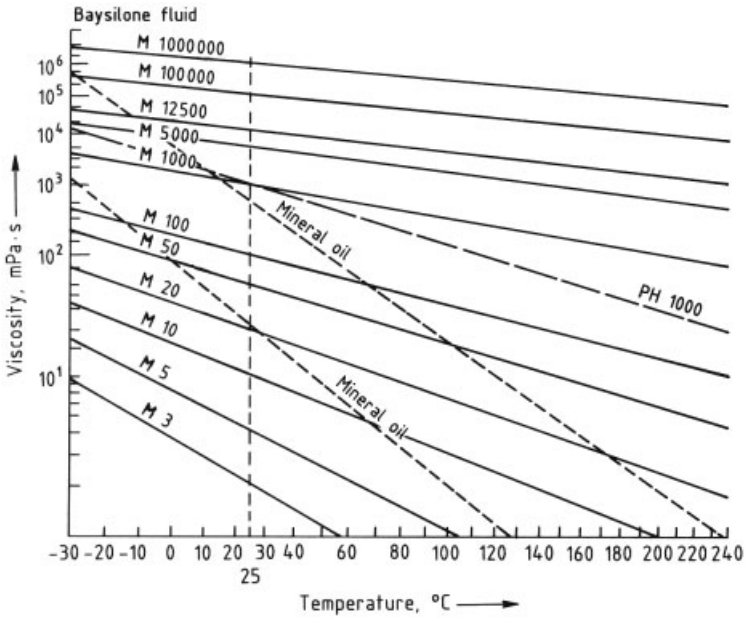

Figure 6. Temperature dependence of the viscosity of silicone fluids compared with mineral oils

3.4. Properties

Viscosity and Molecular Mass. The relation between the viscosity and the molecular mass of PDMS can be expressed approximately by the equation

$$M = \frac{464 \cdot (\eta_{25})^{0.825}}{2 + 0.0905 \cdot (\eta_{25})^{0.555}}$$

Other equations that are applicable to specific molecular mass ranges are given in [15, 16].

Determination of the molecular mass of linear and cyclic polymers from the intrinsic viscosity $[\eta]$ (in mL/g) by means of the formula $[\eta] = K \cdot M^{\alpha}$ is described in [17].

Temperature Dependence of Viscosity. An important property of silicone fluids is the small dependence of the viscosity on temperature, compared to other fluids. The viscosity of PDMS as a function of temperature is shown in Figure 6, which includes mineral oils for comparison.

The relatively low viscosity of silicone fluids at low temperature and the low pour point are important in many applications. The viscosity–temperature coefficient (VTC)

$$\mathrm{VTC} = 1 - \frac{\nu^{99\,^\circ\mathrm{C}}}{\nu^{38\,^\circ\mathrm{C}}}$$

which is frequently reported in the technical literature, is ca. 0.6 for silicone fluids.

As also shown in Figure 6, increasing the phenyl content (phenylsilicone fluids PH 1000) increases the VTC, and the viscosity–temperature behavior approaches that of mineral oils.

Mixture Viscosity. Formulas are available for determining the viscosity of a mixture of two silicone fluids [18]. Silicone fluid manufacturers provide this information in mixture diagrams (Fig. 7).

The required mixing ratio can be determined with the aid of the mixture diagram. For example, if a fluid of viscosity 6000 mPa · s is to be blended at 25 °C from a silicone fluid with viscosity of 1000 mPa · s and a silicone fluid with a viscosity of 12 500 mPa · s, the value 1000 mPa · s is marked on the left-hand side and the value of 12 500 on the right-hand side, and the two values are joined by a straight line. The intersection of this straight line with a line drawn parallel to the x-axis through the value of 6000 mPa · s indicates the mixing ratio on the x-axis, in this case 70 % silicone fluid of viscosity 12 500 mPa · s and 30 % silicone fluid of viscosity 1000 mPa · s.

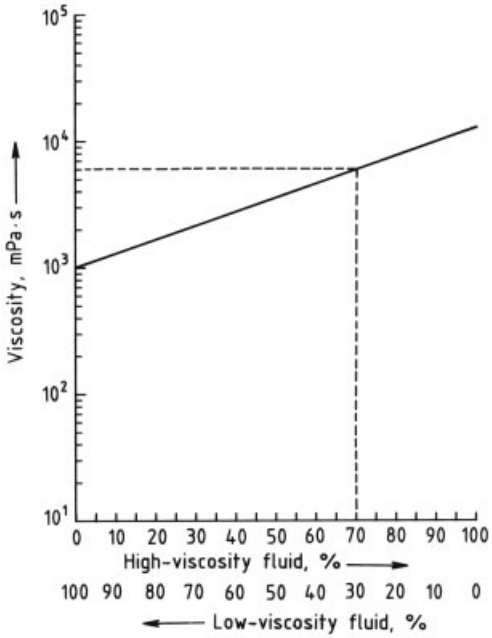

Figure 7. Viscosity adjustment by blending

Newtonian Behavior and Pseudoplasticity. Silicone fluids of low or medium viscosity exhibit Newtonian behavior up to high shear rates (see Fig. 8). However, high-viscosity silicone fluids have lower Newtonian plateaus and show pseudoplastic behavior at lower shear rates. If the viscosity is decreased by raising the temperature, the pseudoplastic behavior begins at a higher shear rate.

Compressibility. The very high compressibility of polydimethylsiloxanes is apparent in the compressibility coefficient of 100×10^{-11} m^2/N. Methylphenylsilicone fluids have lower values of ca. 60×10^{-11} m^2/N, whereas mineral oils exhibit coefficients of ca. 50×10^{-11} m^2/N.

Thermal Properties. Thermal conductivities and specific heats of various silicon fluids are listed in Tables 5 and 6. Both values are nearly independent of temperature.

The thermal stability of dimethylsilicone fluids is very high. Little change is observed in the physical properties even after prolonged exposure to temperatures of 150 – 200 °C in air. Methylphenylsilicone fluids are stable at temperatures as much as 50 °C higher. In inert atmosphere or under vacuum the thermal stability is even higher.

At higher temperature, atmospheric oxygen leads to cleavage of the organic groups and consequently to gelling by cross-linking. By contrast, prolonged heating in a closed system in the absence of oxygen causes chain scission and a decrease in viscosity. These processes are very dependent on the nature of the contacting surfaces and the presence of catalyst traces or impurities.

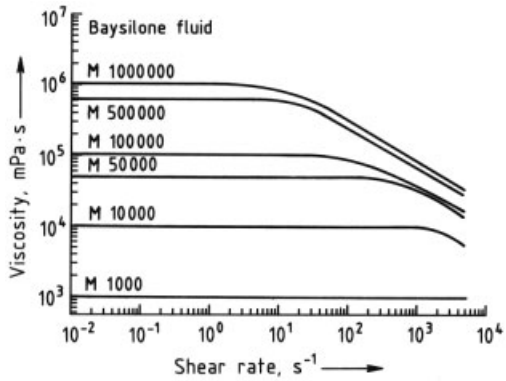

Figure 8. Flow behavior of silicone fluids at 25 °C

Solubility. The PDMS and methylphenylsilicone fluids are highly soluble in aliphatic, aromatic, and chlorinated hydrocarbons, as well as in most ethers, esters, and higher alcohols. They are insoluble in water, methanol, and ethylene glycol. The solubility in organic solvents decreases with increasing viscosity of the polymer. PDMS with viscosities >20 mPa·s exhibit only limited solubility in acetone, ethanol, butanol, and isopropanol. PDMS of different viscosities can be mixed, but PDMS and methylphenylsilicone fluids are immiscible.

The solubility of gases in polydimethylsiloxanes is considerable. Thus, 1 g of silicone fluid dissolves 0.19 mL of air, 0.17 mL of N_2, or 1 mL of CO_2 at room temperature.

Interfacial Properties. Silicone polymers exhibit low surface tension: ca. 20 mN/m for

polydimethylsiloxanes, and ca. 25 mN/m for methylphenylsiloxanes (see Tables 5 and 6).

Since most solids have surface tensions > 20 mN/m, silicone fluids readily form surface films. Exceptions are polyolefins and polytetrafluoroethylene, whose surface tensions are also low. The interfacial tension of polydimethylsiloxane against air and water is listed in Table 7. A water drop on siliconized glass has a contact angle of 80 – 110°, which decreases on prolonged contact due to changes in the state of orientation of the silicone film [19]. Four different states are observed. The spreading of silicone fluids on water leads to ordered systems [20].

Table 7. Interfacial tension σ of polydimethylsiloxane (140 mm^2/s) against air and water at different temperatures [18]

Temperature, °C	σ, mN/m	
	Against air	Against H_2O
10	22.9	28.5
20	22.2	29.5
30	21.5	30.5
40	20.9	32.5
50	20.2	33.0
90	17.5	

3.5. Formulation

Some properties of polyorganosiloxanes can be particularly advantageous in certain formulations. Silicone fluid-in-water emulsions are preferred forms for applying thin films or for metering very small amounts. Silicone fluid emulsions are used, for example, as release agents, for impregnating surfaces, for water repellency treatment, in antifriction applications, and as antifoam agents. Emulsions are generally sold as 10 – 35 % concentrates, which are diluted to usually less than 1 %, sometimes only a few parts per million, before use. Silicone microemulsions, some of which are transparent, are aqueous systems with silicone particle sizes of 10 – 80 nm [21]. For certain applications, silicone fluids are formulated with silica or other consistency-increasing additives such as metal soaps, polytetrafluoroethylene (PTFE), boron nitride, and ureas to make pastes (see Section 3.6).

3.6. Applications

The characteristic properties of silicone fluids are exploited in numerous applications [22]. Thus, because of their high thermal stability and good low-temperature performance, they are used as *heat-transfer media* in heating circuits in the chemical, petrochemical, pharmaceutical, and food industries and in solar power plants, and as *refrigerants* in cryostats, freeze dryers, and climate simulation plants. Their low surface tension leads to use as *release agents* in the processing of plastics and rubber articles.

Silicone fluids, especially functional copolymers with Si – H or Si – OH groups, spread on surfaces to form oriented films with their hydrophobic organic groups aligned opposite to the phase boundary of the substrate. These functional copolymer fluids have many uses as additives in water-repellent *polishes*, as *waterproofing agents* for textiles, and as *protective coatings* for building materials. Their high water vapor permeability is a particular advantage, permitting good ventilation of water vapor while largely preventing the entry of liquid water.

Because of their surface activity, polydimethylsiloxanes are used as *antifoams* in aqueous systems, in petroleum processing, and in laundry detergents. Special formulations with highly dispersed silica are effective antifoams even in the part-per-million range. In human and veterinary medicine, siloxane antifoams are used as *antiflatulent* agents. The antifriction properties of polydimethylsiloxanes make them useful as *lubricants* (e.g., for films, yarn, medical articles, wine corks, and fillers). Special silicone fluids that dissolve spermicides of the Nonoxinol-9 type can be used as prophylactic coatings for condoms. Silicone fluids are increasingly important as *dielectric coolants* for transformers and rectifiers because of their flame resistance [23], resistance to ageing, material compatibility, and physiological inertness. These properties are also important for use as *power transmission fluids* in viscous and fan couplings [24] and as *hydraulic* and *damping fluids* in shock absorbers, railway buffers, and vibration insulation systems, at both high and low temperature. The more advantageous lubricating properties of siloxanes compared to pure organic products in the temperature range from less than

$-20\,°C$ to $>150\,°C$ have led to their use as *lubricants*. The greases or pastes obtained by incorporating consistency improvers exhibit a particularly good lubricating behavior when the base fluids are modified by phenyl, long-chain alkyl, or fluoro groups. They are used chiefly for lubricating electric motors, motor bearings of furnace blowers and ventilators, pump bearings for liquid gases, and machine bearings for low-temperature operation. Flexure-stable silicone fluid pastes are used as *embedding compounds* for glass-fiber cables in communication engineering. Since silicone fluids have a good dissolving power for organic vapors, they can be used as *absorbents* for organic vapors of low water solubility in off-gas purification. Modified silicone fluids are used as *paint additives* to influence the leveling or to obtain special finishes (e.g., hammer effect). Their tolerance by the skin and their physiological inertness [25] make silicone fluids suitable as additives for *ointments* and *cosmetic preparations*.

4. Silicone Rubbers and Elastomers

4.1. General Properties

Silicone polymers exhibit low glass transition and equilibrium melting temperatures, weak intermolecular interactions, and high chain mobility. These properties make them highly suitable for use in rubbers.

Silicone polymers can incorporate a variety of functional groups as potential cross-linking points. The position (exclusively chain ends, or along the polymer chain) and the content of these functional units can be readily varied. Together with the wide range of polymers and their compound viscosities these different cross-linking systems lead to many applications.

The development of silicone elastomers is reviewed in [26–28]. Specialized rubbers with unique properties such as low-temperature flexibility down to $-70\,°C$ according to DIN 53548 (phenyl, ethyl) [29], oleophobicity and solvent resistance (fluoroalkyl, cyanoalkyl), or higher surface tension (polyether) are obtained by replacing the methyl groups with other organic groups.

Curing (vulcanization) converts unvulcanized silicone compounds into silicone rubbers/elastomers. Unlike other rubber polymers, unfilled silicone rubbers achieve only low mechanical strengths when cured. Adequate strengths are only obtained by incorporating reinforcing fillers. High surface area silicas are used almost exclusively for this purpose.

The stress – strain curves of filler-free siloxane networks at small deformations can be well described by the semi-empirical Mooney – Rivlin equation [30, 31]. The ultimate strength of silicon elastomers is determined by the attainable average chain length between cross-links [32]. The Einstein – Guth – Gold equation is also used to describe filled silicone elastomers [33].

4.2. Rubber Compounds

Reinforcing Fillers. Whereas carbon black is most commonly used to reinforce standard organic elastomers, the best reinforcing fillers for silicones are finely-divided silicas. Both pyrogenic and precipitated silicas with BET surface areas of $150-400\,m^2/g$ are used. These silicas have average primary particles of only 7 – 30 nm, which, however, are present in silicone compounds as larger aggregates [34]. Transparent silicone elastomers can be produced with silica fillers because of the small primary particle size.

The tensile strength of silicone elastomers reinforced with high surface area silicas can be as much as 50 times that of unfilled systems, reaching values of up to 12 MPa. Common organic elastomers are significantly stronger and can attain tensile strengths of 20 – 25 MPa.

The silica filler is normally added to the silicone polymer during formulation. However, precipitated silicas can also be generated in situ in solvents [35] or even directly in the rubber polymer from orthosilicate esters [36, 37]. Silica fillers cause a large increase in the viscosity of the formulated rubber [38]. Fillers also influence the low-temperature crystallization rate of the elastomer [26].

Reinforced silicone elastomers usually contain 5 – 38 wt % silica filler. An increase in the viscosity is observed above 2 wt % filler [34]. A higher filler content, like a higher cross-link density, leads to an increase in both the modulus and the hardness of cured rubbers. Increasing the filler content or filler surface area also

raises the tensile strength, while at the same time causing a deterioration of the relaxation behavior of the polymer network. The tensile strength reaches a maximum at filler concentrations of 25 – 30 wt % [34]. The relaxation behavior is influenced by the inflexible filler – filler network, which hinders reversible deformation of the polymer. The stress – strain curves of filled elastomers show a large reduction in stress and hardness after the first deformation (Mullin effect: breakdown of the solid – solid interactions) [39]. A significant increase in the storage modulus is observed above 20 wt % filler. At this filler concentration, known as the percolation point, the silica aggregates meet and interpenetrate the entire polymer network [40].

Inert Fillers. Lower surface area, "inert" fillers are used for cost reduction and for adjusting certain properties. Common inert fillers are quartz powders, diatomaceaous earth, siliceous and other chalks, talcs, micas, calcium or zirconium silicates, and alumina trihydrate. Inert fillers can be used to modify the stress – strain behavior. Substitution of silica by inert fillers leads to a lower hardness and a decrease in the modulus. Treatment of inert fillers with special silane coupling agents can partially compensate for this effect [41].

Inert fillers increase the thermal conductivity of the elastomers [42,43]. Electrically conducting elastomers can be produced by addition of special furnace blacks or other conductive materials (carbon fibers, metal powders) [44,45].

Process Aids. Pyrogenic silicas used in reinforcing silicone elastomers have a pronounced thickening action due to hydrogen bonds between the filler aggregates. Three-dimensionally cross-linked structures, formed in the nonpolar silicone polymer matrix, hinder flow even at low concentration (> 2 wt %). At high filler concentration (ca. 25 wt %), such as are necessary for optimal reinforcement, the viscosity is so high that the compounds can no longer be remilled. Process aids are used in silicone rubber manufacturing to reduce the viscosity. These process aids react to form a hydrophobic filler surface and thereby restrict filler – filler interactions. Silylating agents such as silylamines and silylacetamides [46] are particularly effective. Hexaalkyldisilazanes are preferred.

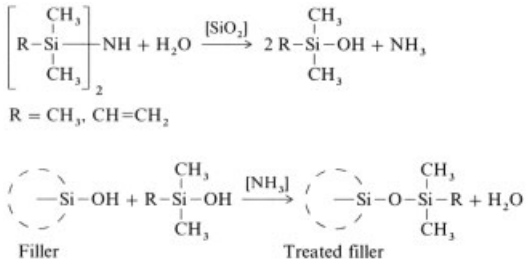

Silylation decreases the proportion of surface-bound rubber (gel content or bound rubber) and prevents hardening and viscosity increase on storage (crepe hardening). However, hydrophobic fillers are less strongly reinforcing [47]. Siloxanediols and alkoxysilanes are used as process aids when a higher pseudoplasticity is required. This applies to solid rubbers, for which stability under load is necessary for processing (e.g., extrusion).

As an alternative to treatment of fillers during rubber manufacture, pretreated fillers can also be used.

Stabilizers. Specific performance properties of silicone elastomers, such as resistance to hot air, chemicals, or fire, can be improved by using additives. Generally, very small quantities (0.001 – 10 wt %) are required. Metal oxides; salts of iron, titanium, zirconium, cesium, nickel, copper, cobalt, or manganese; or carbon blacks are suitable for inhibiting the cross-linking by hydroxyl radicals produced in hot air at > 220 °C [48,49].

The action of water, acids, or bases can lead to depolymerization of the siloxane polymer network [50]. Stabilizers against these agents are alkaline-earth silicates, certain amphoteric hydroxides, and specific organic polymers [51,52].

Silicone elastomers can be protected from continued burning following ignition by addition of conventional flame retardants such as haloaromatics with Sb_2O_3 [53] or $Al(OH)_3$ with zinc borate [54,55]. A method of flame retardation specific to silicones is the use of platinum (10 – 60 ppm), TiO_2 [56], Fe_2O_3, or carbon black, and certain nitrogen compounds [57], which under the action of fire result in ceramization of the surface and reduced depolymerization [58,59].

4.3. Rheology

Both the viscosity and its dependence on shear rate are important in the processing of silicone rubber compounds. The requirements of various rubber processing methods can be readily met because of the large range of silicone polymer viscosities and the different methods of hydrophobic filler treatment [60].

Maximum strength is achieved with high-viscosity polymers and small primary particle size silica fillers. In the case of pyrogenic silicas, the best fillers are those with the highest surface area [34].

Deviations from Newtonian flow behavior can be adjusted for the various silicone rubbers by means of the filler surface area, filler dispersion, and the filler surface treatment. Thus rubbers with weak (two- component RTV) or strong (HTV) shear thinning are available, as well as pumpable liquid silicone rubbers (LSR) and rubbers with thixotropic behavior (one- component RTV).

4.4. Curing Systems

4.4.1. Radical Curing with Peroxides

Silicone rubbers are cross-linked at high temperatures with peroxides but not with sulfur. Both peroxides common in the rubber industry and specific to silicone processing are used. Cross-linkable silicone polymers must contain unsaturated groups in sufficient amounts. Preferred polymers contain 0.03–2 wt % methylvinylsiloxy groups. The concentration of methylvinylsiloxy groups determines the cross-link density and thus important elastomeric properties such as elongation. The vinyl group distribution in the polymer can be determined by means of ^{29}Si NMR spectroscopy [61].

Four groups of peroxides are commonly used industrially: dialkyl, peroxyketal, diaroyl, and alkyl aroyl. The peroxides used preferably are:

Dicumyl peroxide
tert-Butyl peroxybenzoate
tert-Butyl cumyl peroxide
2,5-Dimethylbis(2,5-*tert*-butylperoxy)-hexane
Bis(2,4-dichlorobenzoyl) peroxide
Bis(4-methylbenzoyl) peroxide.

Dialkyl peroxides cross-link mainly via the vinyl groups [62, 63]. The more reactive diaroyl peroxides are less specific and give the higher cure rates required for pressureless curing.

The structure of the polymer cross-links has been elucidated by ^{29}Si NMR spectroscopy and by depolymerization to low-molecular weight units in several cases [62]. Nonspecific peroxides cross-link via both methyl and vinyl groups, forming mostly propylene and butylene bridges, but also a few ethylene bridges, between polymer chains [63, 64]. Diaroyl peroxides can react with the vinyl silyl groups to form ester and ether groups [65]. The major side products of peroxide decomposition are acids.

The cross-link yield with low-vinyl content polymers can be increased by using co- cross-linkers such as triallyl isocyanurates or acrylates [66].

4.4.2. Hydrosilylation Curing

The highly selective hydrosilylation curing process is becoming increasingly important for the vulcanization of silicone rubbers. The cross-linking reaction is based on the addition of Si–H groups to C=C double bonds (see → Silicon Compounds, Organic, Chap. 2.4.). Usually, long- chain polydimethylsiloxanes containing two or more vinyl groups are reacted with short-chain methylhydrogensiloxanes (cross-linkers) in the presence of metal catalysts according to the equation:

$$\equiv Si-H + CH_2=CH-Si\equiv \xrightarrow{Pt} \equiv Si-CH_2-CH_2-Si\equiv$$

Preferred catalysts are platinum compounds. Concentrations as low as a few ppm platinum lead to adequate curing rates.

Hydrosilylation curing, unlike cross-linking by peroxides, produces no decomposition products. The reaction rate can be varied over a wide range by means of the catalyst concentration and by using inhibitors (e.g., 2-methyl-3-butyn-2-ol) [67]. Inhibitors also extend the processing time (pot life) at room temperature. The catalysts are

poisoned by various substances such as compounds of heavy metals (e.g., tin), sulfur compounds (thiols, sulfides, etc.), and nitrogen compounds (amines, isocyanates, etc.).

The molar ratio of the reactants is especially important for optimum curing. A 1.5- to 2-fold molar excess of Si – H groups is generally used to achieve the optimum cross-link density (i.e., by complete reaction of the vinyl groups). It is also possible to produce very soft or gel-like products by incomplete cross-linking.

4.4.3. Condensation Curing

Condensation- curing compounds cure at room temperature and are known as RTV (room temperature vulcanizing) compounds. The elastomeric network is formed by reaction of hydroxy functional polysiloxanes and tri- or tetrafunctional silanes containing hydrolyzable Si – O or Si – N bonds (Table 8). The hydrolyzable groups react with SiOH groups of the polymer or with water to form cross-links.

Table 8. Typical silane cross-linkers

Silane	Cleavage product
$Si(OC_2H_5)_4$, $[SiO(OC_2H_5)_2]_n$, $Si(OC_3H_7)_4$, $CH_3Si(OCH_3)_3$, $CH_3Si(OC_2H_5)_3$	alcohols
$CH_3Si(OCOCH_3)_3$, $C_2H_5Si(OCOCH_3)_3$	acetic acid
$CH_3Si[NH(s{-}C_4H_9)]_3$, $CH_3Si(NHC_6H_{11})_3$	amines
$CH_3Si[O{-}N{=}C(CH_3)C_2H_5]_3$, $Si[O{-}N{=}C(CH_3)C_2H_5]_4$, $CH_2{=}CHSi[O{-}N{=}C(CH_3)C_2H_5]_3$	butanonoxime
$CH_3SiOC_2H_5[N(CH_3)COC_6H_5]_2$	*N*-methylbenzamide, ethanol

Most silane cross-linking agents react spontaneously with SiOH groups or with water. Metal catalysts (Sn or Ti) are generally added to these systems to give complete curing and improve the properties. Alkoxysilanes are an example of cross-linkers that do not cure in the absence of appropriate catalysts.

Condensation curing is used in one- and two-component products, which differ in composition and cure rate.

One- component RTV compounds contain excess cross-linking agent. During compounding the hydroxyl functional polysiloxanes and the cross-linker react to produce polymers with reactive end groups, for example:

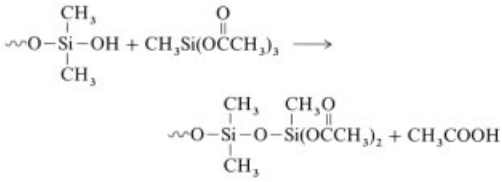

In this way, cross-linking of the polymers can be repressed. One- component RTV products are sold in containers sealed against humidity. Curing starts when the compounds are exposed to atmospheric moisture during application. As the remaining free cross-linking agent and the reactive polymer endgroups are hydrolyzed, the resulting SiOH groups react with unhydrolyzed groups to form an elastomer network. The following simplified reaction scheme applies:

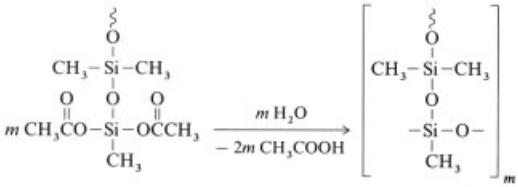

One- component RTV compounds are distinguished according to the cure byproducts as acidic, basic or, neutral cure systems (Table 8). Acidic systems contain methyl- or ethyltriacetoxysilane as curing agent and produce acetic acid as byproduct [68]. Basic systems produce amines [69]. Several neutral systems are commercially available. Alkoxy systems use mainly methyltrimethoxysilane [70, 71]. Oxime systems contain methyl(tributanone oximo)silane and give butanone oxime as byproduct [72]. Benzamide systems produce *N*-methylbenzamide as a typical hydrolysis byproduct. The latter remains as a finely divided solid in the cured elastomer [73].

Tetrafunctional alkoxysilanes are generally used as cross-linkers in two- component RTV compounds. They must be used in combination with tin catalysts, whose specific reaction mechanism ensures cross-linking of the OH-functional polysiloxanes, even in the presence of excess cross-linker [74, 75]. The presence of

water is also necessary for curing. Unlike one-component RTV compounds, the process does not involve an intermediate endstopped polysiloxane. The compounds cure uniformly, in contrast to one-component RTV systems, which cure from the surface inward.

4.4.4. Radiation Curing

Radiation-cured products have been unimportant hitherto. This kind of cross-linking is applied principally in processes for which high curing temperatures are not possible [76].

Curing by Ultraviolet Light. Ultraviolet curing has been described for coating temperature-sensitive substrates (paper, films) [77].

Cross-Linking by γ Rays and Electron Beams. Cross-linking under the influence of γ rays or electron beams proceeds via the formation of free radicals [78].

Polydimethylsiloxane rubbers cured with 1 – 5 Mrad of γ rays have similar mechanical properties to peroxide-cured rubbers.

4.4.5. Oxidative Coupling

Polysiloxanes containing mercaptoalkyl groups react rapidly at room temperature in the presence of metal catalysts to form elastomers cross-linked by disulfide bridges [79]. These rubbers lack some of the advantages typical of silicones, such as odorlessness, physiological inertness, and weather resistance.

4.5. Peroxide-Cured High-Temperature Vulcanizing Silicone Rubbers

High-temperature vulcanizing (HTV) silicon rubbers are designed for the processing methods and equipment of the rubber industry.

The viscosity of HTV compounds is in the range of 20 – 100 Mooney units at 25 °C (concentric-disc viscometer DIN 53 523), typical of other rubbers but only at temperatures of 80 – 120 °C.

The base polymers are high-viscosity, vinyl-containing polysiloxanes with only methyl (VMQ), or a mixture of methyl and phenyl (PVMQ) or trifluoropropyl (FVMQ) groups. The vinyl substituents can be present as end groups and along the polymer chain. The vinyl group concentration is commonly between 0.03 and 2.0 mol %.

Production. HTV silicone rubbers can be produced batchwise or continuously from the siloxane polymers, silica fillers, and process aids in conventional rubber mixing equipment.

Suitable mixing equipment includes e.g., twinscrew compounders (2 – 6000 L), internal mixers (Banbury, 1 – 300 L), roll mills (1 – 100 L), twinscrew extruders or Buss co-kneaders (2 – 500 kg/h). The individual machines have different mixing times, but comparable space – time yields [80, 81].

Since hydrophobic silica fillers are not usually employed for mixing, dispersion at temperatures up to ca. 160 °C is often necessary. Unlike the production of temperature-sensitive, high-viscosity organic rubbers, heat – not cooling – must be applied during silicone rubber production.

The rubbers are usually manufactured as pigment- and cross-linker-free compounds. The addition of cross-linking agents (e.g., peroxides), pigments, stabilizers, and optional processing aids is usually carried out by the user on a smaller scale in easily cleaned roll mills. Ready-to-use compounds are marketed on a limited scale by silicone producers.

Processing. HTV silicone rubbers are pasty, translucent or colored materials. In the uncured state at 25 °C they have a lower green strength than other rubbers. To improve the processibility on roll mills (roll workability) and for better demolding, release agents [82, 83], or viscosifying additives such as PTFE powder [84] are sometimes used.

Molding. Peroxide-containing rubbers are prepared for molding as sheets, strips, or granules. Compression-molded articles are produced in steel molds filled by means of a piston or in automatic transfer molding presses and casting machines [60]. Tubes and insulated wires are produced continuously by extrusion. Coatings or coverings are produced on calenders.

Vulcanization. Vulcanization of HTV moldings is carried out in hot compression molds, in steam chambers under pressure, or in unpressurized hot air ovens. Diaroyl peroxides are used for pressureless vulcanization because of their low decomposition temperature (60 – 90 °C) and high cure rate (5 – 20 min at 100 °C).

Curing with diaryl peroxides occurs in a few minutes at 160 – 180 °C. The cure kinetics can be determined with conventional curemeters, which measure an increase in torque as an index of the degree of cross-linking. The time required for a 50 % increase in torque roughly corresponds to the half-life of the peroxide [85]. After curing, volatile siloxanes (0.5 – 1.5 %) and residual peroxide decomposition products must be removed. This is normally accomplished by heat treatment (post-curing) for 2 – 6 h at 200 °C.

High-viscosity silicone rubbers can also be vulcanized by hydrosilylation curing. A precondition is a sufficient concentration of cross-linkable vinyl groups and their appropriate distribution in the polymer. These systems produce high-strength, transparent elastomers free of discoloration. Both compression molding and unpressurized curing can be employed [86, 87].

Cross-linking by condensation reactions is not common for these rubbers due to the long cure times.

Other Forms of Application. Dispersions of HTV rubbers in organic solvents are used to produce coatings such as coated glass fibers and braided cables.

Silicone foam rubber can be produced by adding blowing agents or by hydrosilylation curing in the presence of SiOH groups or water. The latter method involves the evolution of hydrogen (see → Foamed Plastics, Chap. 4.7.).

4.6. Liquid Silicone Rubbers

Liquid silicone rubber (LSR) is a new class of rubber that has grown rapidly in importance since the early 1980s. This rubber was developed specifically for processing on injection molding equipment. LSR compounds exhibit a number of favorable processing properties such as low viscosity and high curing rate and represent an economical alternative to conventional rubber processing due to the lower cost of the finished article.

Liquid silicone rubbers are medium-viscosity materials that can be pumped from the storage vessel to the injection molding machine. Hydrosilylation curing is the exclusive mechanism of vulcanization for this class of rubbers (see Section 4.4.2). Peroxide-cured liquid silicone rubbers are of no commercial importance.

Formulation. Liquid silicone rubber is based on vinyl-containing polydimethylsiloxanes with viscosities about 1000 times lower than for HTV rubber stock. However, the chain length, which is decisive for the development of the elastomer network is only six times lower. To compensate for the relatively short chain length and still maintain an adequate network density, the majority of the polymer must have vinyl end groups. Acceptable reinforcing fillers are pyrogenic or precipitated silicas made hydrophobic by treatment with silylating agents. Filler treatment reduces the mutual interaction of the filler agglomerates to produce compounds that are free flowing despite the high filler content.

Liquid silicone rubbers are formulated as two-component systems to ensure a long shelf life. Only after the two components (A and B) are mixed can the rubber be cured. The A component usually contains the platinum catalyst in addition to the base polymer, while the B component contains the cross-linker (polymethylhydrogensiloxane) and an optional inhibitor in addition to the base polymer.

Liquid silicone rubbers with other base polymers [e.g., methylphenylsiloxane (PVMQ) or methyltrifluoropropylsiloxane (FVMQ)] [88, 89] or other fillers (e.g., carbon black for electrically conductive types) are also available [45].

Processing. Liquid silicone rubber is supplied ready for processing. The two components are conveyed with a metering pump to a mixing head in which an optional dye paste can be separately charged. The material then flows through a static mixer into the automatic injection molding machine (Fig. 9). In a predetermined cycle, the material is injected into the hot mold, held under pressure to cure, then automatically demolded. The molds are mainly of the so-called

cold-runner type, in which the sprue channels are cooled so that no loss of material occurs.

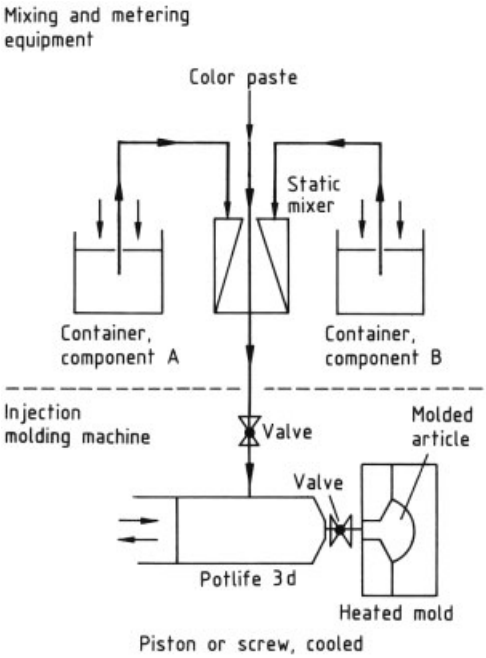

Figure 9. LSR processing

Liquid silicone rubber can be inexpensively processed like thermoplastic elastomers, but has the advantage of being chemically cross-linked. Thus, use temperatures of well above 100 °C are possible.

The cure temperature of LSR is generally between 170 and 230 °C. Cycle times of 15 – 60 s can be achieved, depending upon the wall thickness of the moldings. Many moldings are used as obtained without further treatment. With correct mold design, flash removal is unnecessary. For certain applications, e.g., in the foods industry, post-curing the articles (usually 4 h at 200 °C) is required to remove residual volatiles. Post-curing also removes excess Si–H groups by hydrolysis and partial condensation.

$$\equiv Si-H + H_2O \xrightarrow{Pt} \equiv Si-OH + H_2$$
$$\equiv Si-H + HO-Si\equiv \xrightarrow{Pt} \equiv Si-O-Si\equiv + H_2$$

Post-curing results in new disiloxane bridges and an increase in the network density. After post-curing, further reactions of the Si-H groups are no longer possible. As a consequence, the compression set, an important property for many applications, is reduced to a desirably low level.

Liquid silicone rubber is suitable for producing coatings due to the relatively low viscosity. Special grades with low filler content and, therefore, lower viscosity are available for this application. Unlike many other elastic coating materials, LSR formulations contain no solvent.

4.7. Room Temperature Curing Silicone Rubbers

Silicone RTV (room temperature vulcanizing) systems are classified according to their cure mechanism (hydrosilylation or condensation cure). One- or two-component systems have been developed for different applications. Two-component products, which must be mixed before application, are available with either curing mechanism. One-component RTV compounds are based almost exclusively on a condensation mechanism. These systems cure upon exposure to atmospheric moisture.

Typical RTV silicone products contain polymers with a polydimethylsiloxane backbone [90]. One-component systems, composed of organic polymers with silane end groups [91], are not discussed here.

Water-based silicone sealants are unique one-component systems that consist of an aqueous emulsion of a silicone polymer, fillers, an organotin condensation catalyst, and an anionic surfactant. On application, the water evaporates to give a material that is cross-linked mainly by the filler [92, 93].

4.7.1. Two-Component RTV Systems

Two-component RTV systems are used as pourable potting or casting compounds.

Condensation-Cured Systems.

Composition. The base component contains a hydroxyl-end-capped polydimethylsiloxane (M ca. $10^4 - 10^5$), which may also contain non-reinforcing fillers, unreactive silicone fluid and pigments, as well as small amounts of water and other additives.

The second component (cross-linker), which is sensitive to hydrolysis, contains the cross-linking agent (generally an alkoxysilane) and a catalyst (often the condensation product of an alkoxysilane with a dialkyltin compound).

Processing. The two components are generally used in ratios between 100 : 1 and 10 : 1 (base component to cross-linker). Automatic mixing and dispensing equipment requires a ratio of 10: 1 for reliable operation.

Curing begins shortly after the two components are mixed and is usually complete within several hours to one day at room temperature. Alcohol is generally eliminated during curing and must completely evaporate from the cured elastomer to prevent reversion. The rate of cross-linking and evaporation of the alcohol can be accelerated by raising the temperature.

Hydrosilylation-Cured Systems.

Composition. Hydrosilylation- cured systems consist of two components which are often mixed in the ratio 1: 1. One component contains a polydimethylsiloxane with vinyl end groups, optional fillers, pigments, and the platinum catalyst. The second component contains the cross-linking agent, which is either a polymethylhydrogensiloxane or a copolymer thereof with a polydimethylsiloxane. This component can also contain fillers and a vinyl- containing siloxane.

Processing. Curing begins immediately after mixing the two components. The reaction can be retarded, and hence the pot life increased, by adding low molecular mass cross-linkable components or inhibitors.

4.7.2. One-Component RTV Systems

Uses. One- component RTV silicone compounds are mainly used as sealants in building construction owing to their good stability to weathering [94, 95]. The pastes are formulated to be nonsagging and easily pumpable, so that they can be readily applied. The pastes are packaged in plastic cartridges or, increasingly, tubular plastic bags, which prevent the ingress of atmospheric moisture prior to use.

Numerous industrial applications also exist in which one- component RTV compounds are used as sealants, adhesives, and coatings. The properties of the product are adjusted to suit the application, ranging from easily pourable to highly viscous or rigid compounds, and from low-modulus rubbers to adhesives with high tear strength. Examples include automobile sealants [96], textile coatings, production of insulating glass, contruction adhesives (e.g., for cladding panels), sealing of furnaces, and applications in electrical and electronic engineering [89].

Composition. The base polymers are polydimethylsiloxanes with terminal hydroxyl groups and a viscosity of 1 – 500 Pa · s, preferably 10 – 150 Pa · s. For building construction, where soft vulcanizates are mostly required, nonfunctional polydimethylsiloxane fluids are added as plasticizers. The rigidity is adjusted by adding 7 – 15 wt % pyrogenic silica. Silica of low surface area is preferred (130 – 150 m^2/g). Pigmented compounds often contain inert fillers in addition to silica to lower the cost; ground natural chalk, treated with stearic acid, is used preferentially.

The silane cross-linking agents are used in quantities of 3 – 6 wt %. The amount of cross-linking agent must exceed the quantity required to react with the hydroxyl groups of the raw materials. Neutral systems generally contain an organofunctional silane as coupling agent (e.g., aminopropyltrialkoxysilanes). Dialkyltin carboxylates or chelate complexes of Ti(IV), preferably with acetylacetone or esters of acetylacetic acid, are used as catalysts. Tin compounds are generally employed in amounts of < 0.5 wt %, titanium compounds are used in higher quantities [97].

Building sealants contain inorganic pigments as additional auxiliaries for coloring and biocides to prevent the growth of mold in damp rooms.

Processing. The material is applied by direct injection into the joint from the cartridge or tube by using manual or compressed-air pistols. After a few minutes, an elastic skin forms on the surface under the action of atmospheric moisture. Curing proceeds from the surface to the inside and, with usual joint dimensions, is complete after at most a few days. The rate of curing depends not only on the ambient temperature and humidity, but also on the type of sealant. Silicone sealants generally exhibit high flexibility and good adhesion to a wide range of substrates.

Use of a primer may be necessary for porous substrates and in the case of insufficient adhesion.

Whereas the classical acetate systems dominate in the sanitary sector, neutral systems are being used in increasing amounts, for example, as window sealants. Low-odor, noncorrosive systems can be expected to further increase in importance.

General guidelines for processing of silicones for industrial fabrication cannot be given because of the wide range of applications.

4.8. Paper and Textile Coatings

Silicone elastomers are used in a variety of coating applications. The most important substrates are paper and textiles. The choice of polymer and curing system, the formulation and the coating method depend on the substrate and the desired properties.

4.8.1. Paper Coating

Paper and films are often coated with very thin layers of silicone to make them repellent to adhesive substances.

Paper coating plants operate at high speed (up to 500 m/min), and thus require silicone coating systems that cure rapidly while maintaining a sufficient pot life. Silicone coating materials are two- or multicomponent systems that cure at substrate temperatures of 100 – 130 °C in 5 – 15 s.

The coating mixture must exhibit good wetability, both for the rolls and for the substrate surface, to give pore-free films. Completely cured coatings contain no migrating constituents that can impair the adhesive properties of a subsequently applied adhesive layer. The adhesive is applied directly (on-line process) or in a later manufacturing step. The adhesive forces between the adhesive layer and the silicone coating must remain constant over fairly long periods of storage. This is especially important when controlled-release additives are used to increase adhesion between the silicone and adhesive layers.

At present, the following silicone coating systems are used:

Solvent-Containing Systems.

Condensation systems (I) consist of hydroxyl-end- capped polydimethylsiloxanes and polymethylhydrogensiloxanes as cross-linkers. The cross-linking reaction is catalyzed by organotin compounds:

$$\equiv SiOH + HSi\equiv \xrightarrow{Sn} \equiv SiOSi\equiv + H_2$$

The system is used in dilute form (3 – 5 % in organic solvents). In emulsion form it is diluted to about 10 % with water before use.

Condensation systems (II) contain reactive silanes instead of polymethylhydrogensiloxane. Their curing mechanism is the same as for the moisture- cured RTV compounds (see Section 4.4.3). Systems I and II have been virtually displaced from the market by hydrosilylation-cured systems because of disadvantages such as sensitivity to variations in atmospheric humidity and slow cure even at high curing temperature (20 – 40 s at 120 – 180 °C).

Hydrosilylation- cured systems consist of higher viscosity polydimethylsiloxanes with reactive vinyl groups and polymethylhydrogensiloxanes as cross-linkers. They cure in the presence of platinum complexes (30 – 120 ppm Pt) in 15 – 20 s at 120 – 140 °C (for curing mechanism see Section 4.4.2). They are applied as solutions in organic solvents or as emulsions. With sufficient dilution it is possible to prepare very thin films. The solvent is removed by evaporation during cross-linking and is recovered or burned.

Solvent-Free Systems.

Heat-cured systems are also based on the principle of hydrosilylation curing and differ from the corresponding solvent- containing systems by having shorter- chain vinyl- containing polymethylsiloxanes with a viscosity of 200 – 500 mPa · s. In contrast to solvent- containing systems, the thickness of the solvent-free coating does not change on curing. Therefore, stringent requirements must be met by both the application equipment and the substrates to obtain pore-free films with reproducible coating thicknesses of ca. 1 µm (ca. 1 g/m^2) silicone. Solvent-free systems cure in 5 – 15 s at substrate temperatures of 100 – 120 °C. About 50 % of all sub-

strates are coated with solvent-free systems, and the proportion is increasing.

Photochemically cured systems are not well established in paper coating, in spite of the advantage provided by minimal thermal treatment of the substrates. This method is, however, used for special coating applications [98]. The cross-linking mechanisms are varied, but most systems use UV-absorbing sensitizers to increase the quantum yield.

Radiation-cured systems use high-energy radiation (e.g., electron beams) and do not require sensitizers, but they do require a nitrogen atmosphere to avoid formation of interfering oxygen radicals.

Silicone- coated papers and films are assessed by the following criteria:

Curing of the coating
Abrasion resistance (no rub-off)
Adhesion to the substrate
Porosity of the coating
Magnitude of the release force values
Constancy of the release force values
Residual adhesive forces
Coating strength
Film smoothness

Fields of application for the abhesive coated papers are [100]:

Self-adhesive labels (74 %)
Adhesive tapes (10 %)
Hygienic articles (6 %)
Construction, packaging, decoration, etc. (10 %)

4.8.2. Textile Coating

Textiles can be rendered water-repellent with special silicone fluids that simultaneously provide the fabric with a soft feel. Elastomer systems that are similar in composition to those described for paper coating or for two- component and LSR silicone rubbers are used for waterproof textile coatings that are permeable to water vapor.

4.9. Properties of Silicone Elastomers

The typical properties of silicone polymers are also exhibited by cured silica-reinforced products. They account for the many applications of silicone elastomers. Different combinations of properties can be obtained by choice of polymer, cross-linking system, and compound composition.

Temperature Limits for Silicone Elastomers. Silicone elastomers have the widest operating temperature range of commercially important rubbers. VMQ rubbers exhibit better relaxation properties than other synthetic rubbers, i.e., the compression set has a smaller temperature dependence in the range −40 to 170 °C (Fig. 10) [99].

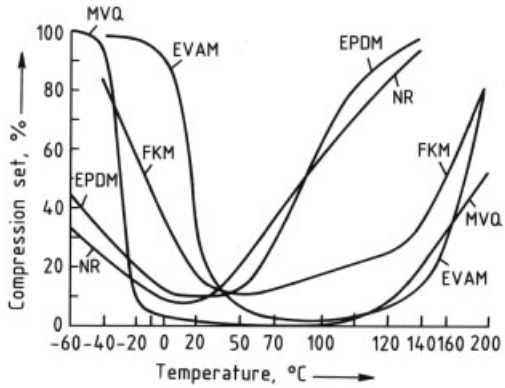

Figure 10. Compression set versus temperature for some elastomers

Mechanical properties such as tensile strength and tear strength also show only a small temperature dependence between − 40 and 150 °C. The tensile strength of VMQ is strongly enhanced at − 40 °C.

This general behavior must be distinguished from heat stability or changes in the room temperature properties that occur at elevated temperature. The heat stability of elastomers is evaluated by measuring the retained properties after high-temperature ageing (see hot-air resistance).

A phenyl content of 5 – 12 mol % extends the low-temperature flexibility (measured as torsional stiffness, DIN 53 548) from − 45 to − 70 °C and the brittleness point from − 55 to − 105 °C (ASTM D 380) [88].

Phenyl groups improve the resistance to γ rays by a factor of 2 – 5.

Hot-Air and Oxidation Resistance. The high-temperature and oxidation resistance of pure cross-linked polydimethylsiloxane polymers is adequate only up to 200 °C. Above this temperature, oxidation reactions at the alkylene cross-links and methyl groups cause embrittlement. Oxidation can be inhibited by using hot-air stabilizers (see Section 4.2). Further improvement is achieved by limiting the content of acidic Si – OH groups in the silica filler.

Since even low contents of acidic and alkaline impurities lead to depolymerization (reversion) of the network, the polymers must be thoroughly neutralized. Stabilizers that neutralize acidic decomposition products from cross-linking agents, for example, are added optionally. Elastomers stabilized in this way withstand exposure to hot air at 300 °C for up to 21 d. After an induction period, embrittlement increases logarithmically with time.

Steam, Chemicals and Oil. Despite the sensitivity of the Si – O bond to acids, bases, and to water at elevated temperature, the stability of silicone elastomers can be significantly improved by increasing the cross-linking density and by using additives [52, 101].

Patents describe the use of combinations of CaO and silanes containing methacrylic groups to improve the oil resistance [102]. Additives such as mica or diatomaceous earth have a similar effect [103, 104]. The swelling rate can be appreciably reduced by using fairly large amounts of vinyl resins [105]. Stabilized VMQ rubbers withstand immersion in ASTM 2 oil for up to 2000 h and in ASTM 3 oil for up to 500 h.

Swelling rate depends on the solubility parameters χ of the siloxane and the contact medium. Similar χ values result in higher swelling rate [106]. Additional degradation by acids or bases occurs in technical applications. Silicone elastomers with enhanced oil resistance usually have a high filler content and thus a reduced siloxane and stabilizer content.

A large decrease in the swelling rate can be achieved by changing the solubility parameter of the polymer (e.g., by using fluoroalkyl-modified siloxanes) [89].

Another method to improve the oil resistance is to blend silicone elastomers with solvent-resistant rubbers such as acrylonitrile or fluorinated rubbers [107–110]. These blends can exhibit improved low-temperature and swelling behavior. Interest in blends has increased with the introduction of new peroxide- curable fluorinated rubbers [107, 111].

Reversion – Compression Set. Reversion and increased compression set are often induced by the action of water vapor and acid or base. The resulting breakdown of the polymer network can be counteracted by polymer neutralization and by using the maximum possible cross-link density. Compression set can also be reduced by using neutralizing additives or by the addition of methylhydrogensiloxanes [52, 101, 112].

High-Strength Formulations. The room-temperature tensile strength of silicone elastomers of 6 – 12 MPa is lower than that of other common elastomers (10 – 30 MPa). The tear strength is also only moderate (10 – 30 N/mm, compared to 20 – 80 N/mm), according to ASTM 624 D die B. Much effort has been targeted at improving the tear strength of silicone elastomers [113–119]. The tear propagation strength can be increased to 40 – 55 N/mm by optimizing the polymer structure.

Adhesion. Silicone elastomers are often used in composite structures with other materials. Here, adhesion of the elastomer to the other material is particularly important.

Owing to their cross-linking mechanism, condensation cross-linking systems show good adhesion to surfaces with hydroxyl groups. The hydroxyl groups on the surface react with the SiOH or SiOR groups of the polymer. Good adhesion is observed to glass and to metals with oxide or hydrated surfaces, such as aluminum and iron. Condensation systems with and without solvent are also used as coupling agents or adhesives for other silicone elastomers.

For good adhesion peroxide- and hydrosilylation- cured rubbers frequently require primers that are applied as coupling agents to the surface of the material. With the proper choice of primer, good adhesion can even be obtained on many plastic surfaces [120–123].

Electrical Properties. Siloxanes and their mixtures with pyrogenic silica have a very good electrical insulating capacity. The electrical properties are temperature dependent and are strongly affected by exposure to water. Accordingly, the resistivity of $10^{16}\,\Omega\cdot$cm under dry conditions at 20 °C decreases to $10^{12}\,\Omega\cdot$cm at 160 °C and 50 % R.H. [32].

The use of precipitated instead of pyrogenic silica leads to a decrease in resistivity and, because of the higher water content of the filler, to an increase in dielectric constant. The addition of carbon blacks [124], sometimes in combination with conducting powders or fibers [125, 126], can lower the resistivity to $3\,\Omega\cdot$cm (ISO 1853).

Thermal Properties. The thermal conductivity of silicone elastomers can be increased by large amounts of inert fillers such as quartz, AlN, Al_2O_3, Si_3N_4, or MgO [41, 43, 127, 128].

Burning Properties. The heat evolution from siloxanes on burning is lower and accordingly more favorable than that of pure organic elastomers [129, 130]. A residual electrical insulating capacity is maintained, even after burning, as a result of the remaining silica framework.

Flame retardants can effectively suppress the further combustion of silicones after ignition. In this way, transparent or colored elastomers can be produced that pass flammability tests such as UL 94 V1 or V0 [59]. The combustion gases formed are less toxic than those from halogen- or nitrogen-containing elastomers [129–132].

Gas Permeability. Silicone elastomers dissolve many gases well and are highly permeable. The high gas permeability [133] is somewhat selective and depends on the substituents [134–137]. Silicone elastomers are therefore suitable for the production of gas-separation membranes [138] and for oxygen-permeable contact lenses or dressings for wounds [139–141].

Transparency. The extremely small primary particles of well-dispersed high surface area silica fillers in silicone elastomers cause only minor light scattering. With a UV transmission (DIN 5033) of up to 90 %, they are almost transparent [46] and are therefore used as contact and other lenses [140]. This high transparency is achieved by special hydrophobic treatment of the filler with strong silylating agents [142]. A further enhancement of the transparency is possible by incorporating phenyl groups into the polymer, which shifts the refractive index n_D^{20} from 1.39 to 1.43 – 1.47 [143–145]. Recently, more rigid silicone – acrylate copolymers have been preferred for contact lenses [146].

Physiological Behavior. Because silicone elastomers exhibit low surface tension and are abhesive, interactions with surfaces of living tissues, cells, or blood platelets are very weak [146–148]. Therefore, filler-free elastomers, in particular, are suitable for implant applications. Numerous surgical and medical applications exist. Another advantage of silicone elastomers is that they can be sterilized in steam above 100 °C.

4.10. Applications

The typical properties of silicone elastomers lead to their application in areas involving:

1) Electrical insulation or conduction
2) High and/or low temperatures
3) Weather and UV exposure
4) Oil and hot-air contact
5) Dynamic stress
6) Flame resistance, arcing resistance
7) Gas permeability
8) Contact with foods
9) Contact with living tissue
10) Abhesive surfaces
11) Transparent articles of high optical quality

Examples of uses of silicone elastomers are listed below:

Electrical
wire insulation
cable sheaths
power cable endcaps

Electronic
anode caps
coatings
encapsulation
fiber-optic coatings

Household
pot seals
coffeemaker tubes and seals

baby bottle nipples
oven gaskets
anti-stick papers

Automobile
gaskets
cable guides and connectors
headlight seals
shaft seals
O-rings
cooling and turbo hoses
oil pan gaskets
spark plug boots
air bags
ignition cables
fuel line diaphragmas and seals
exhaust pipe hangers
axle boots

Airplane
headlight seals
window seals
interior joints and floors
vibrational damping

Office machines
keyboard pads
copier rolls

Medicine and dentistry
tooth impression compounds
implants
pumps
heart valve seals
catheters
membranes
lenses
controlled drug release

Paper and Textiles
coatings
conveyor belts

General
casting forms
membranes
impression molding compounds
anti-stick bladders
divers' masks
protective masks

5. Silicone Resins

5.1. Structure and General Properties

Alongside fluids and elastomers, silicone resins are the third important class of materials produced by hydrolysis or alcoholysis of organochlorosilanes [149, 150].

Whereas silicone fluids and elastomers are based on linear polymers, silicone resins are highly branched, containing significant quantities of trifunctional (T) or tetrafunctional (Q) units.

By combination of T and Q units with difunctional (D) and monofunctional (M) units, a large number of different resin structures can be formed, in which the proportion of the various silicone units can vary over a wide range.

The properties of silicone resins are strongly influenced by the organic groups bonded to silicon. The most frequently used monomers are methyltrichlorosilane, phenyltrichlorosilane, dimethyldichlorosilane, and diphenyldichlorosilane. Pure T resins are relatively brittle and exhibit little thermoplasticity. The hardness can be decreased by adding D and M units. An increased dimethylsiloxane content, in particular, leads to products of increased elasticity and adhesion, but also increased tendency to yellow at high temperature. With increasing temperature, the hardness of the resins decreases. This effect is stronger when diphenyldichlorosilane is used. The introduction of phenyl groups has a plasticizing effect. Simultaneously, the resistance of the polyorganosiloxanes to heat and weathering is increased, and their compatibility with organic polymers, their pigmentability, and their yellowing resistance are improved. Products with increased alkali resistance required for masonry protection are obtained by using special silicone resins.

In addition to pure silicone resins, combination resins also play an important role [151, 152].

Combination resins are produced by cocondensation of silicone resins having hydroxyl or alkoxy groups with functionalized organic resins such as polyester, alkyd, acrylic, or epoxy resins. Compatibility of the resins is improved by using methylphenylsilicones.

5.2. Production

Silicone resins and their precursors are generally produced from mixtures of the corresponding organochlorosilanes by hydrolysis or alcoholysis. From a given mixture, completely different cross-linked structures with markedly different properties can be obtained, depending on process conditions. The simplest case – addition of the silane mixture to the hydrolysis medium – is known as direct hydrolysis. Since the chlorosilanes are immiscible with water and the individual silanes hydrolyze at different rates, this can lead to nonuniform products and, in the worst case, gelation.

In reverse hydrolysis, in which the hydrolyzing medium (usually an alcohol – water mixture) is added to the silane mixture, a more uniform cohydrolysis is achieved and gelling is largely avoided. However, low-boiling starting materials may be lost, together with escaping hydrogen chloride.

Both processes can be improved by adding solvents. Solvents immiscible with water, such as toluene, xylene, or cyclohexane, lower the silane concentration and the concentration of reactive hydrolysis condensate at the interface, and repress the influence of hydrochloric acid, which promotes condensation. The process can be further improved by using solvents miscible with both water and silane (two-phase solvents), such as acetone. A short- chain alcohol such as butanol is often used to lower the reactivity of the system by forming alkoxysilanes. A drawback of the two-phase solvent is the relatively high concentration of organic substances in the wastewater.

Silicone resins are commonly bodied following hydrolysis by heating in the presence of mild condensation catalysts. This step is used to control the viscosity and cure properties of the resin. In this reaction the functional groups condense according to

$$\equiv\mathrm{Si}-\mathrm{OH}+\mathrm{HO}-\mathrm{Si}\equiv \longrightarrow \equiv\mathrm{Si}-\mathrm{O}-\mathrm{Si}\equiv+\mathrm{H_2O}$$

or

$$\equiv\mathrm{Si}-\mathrm{OH}+\mathrm{RO}-\mathrm{Si}\equiv \longrightarrow \equiv\mathrm{Si}-\mathrm{O}-\mathrm{Si}\equiv+\mathrm{ROH}$$

The condensation process is frequently accompanied by viscosity monitoring and terminated by cooling, neutralization, or dilution when the desired properties are attained.

The nature and concentration of the remaining reactive groups and the molecular mass distribution determine many of the further processing and product properties.

Combination resins are formed by reaction of siloxane precursors with hydroxy-functional organic esters (e.g., polyesters), alkyd resins, acrylates, or epoxy resins

$$\equiv\mathrm{Si}-\mathrm{OH}+\mathrm{HO}-\mathrm{C}\equiv \longrightarrow \equiv\mathrm{Si}-\mathrm{O}-\mathrm{C}\equiv+\mathrm{H_2O}$$

and

$$\equiv\mathrm{Si}-\mathrm{OH}+\mathrm{RO}-\mathrm{C}\equiv \longrightarrow \equiv\mathrm{Si}-\mathrm{O}-\mathrm{C}\equiv+\mathrm{ROH}$$

The hydrolysis of chlorosilanes with subsequent condensation is irreversible. Alcoholysis products can be improved by an optional acid- or base- catalyzed equilibration stage. Although silicone resins can be produced continuously [145, 153–155], many products with relatively low demand are produced in batch processes.

5.3. Curing

Silicone resins can be cured by the same reaction mechanisms as silicone elastomers. The most common method of curing silicone resins is metal- catalyzed condensation, preferably with complexes of tin or titanium. Resins containing vinyl groups can be cross-linked by peroxides. The simultaneous presence of silicon-bonded hydrogen and of vinyl groups enables curing by noble-metal catalyzed hydrosilylation (see Section 4.2). Radiation-induced curing reactions are also possible. Above 150 °C in air, oxidative degradation occurs and is accompanied by curing, for example,

$$\equiv\mathrm{Si}-\mathrm{CH_3}+\mathrm{CH_3}-\mathrm{Si}\equiv \xrightarrow{\Delta,\ \mathrm{O_2},\ \mathrm{cat.}} \equiv\mathrm{Si}-\mathrm{O}-\mathrm{Si}\equiv$$

This curing mechanism is exploited in the use of silicone resins as binders in the paint industry, for example, in high-temperature anticorrosion coatings, release coatings, or decorative coatings for kitchenware [156].

5.4. Properties

The characteristic properties of silicone resins are their high resistance to temperature and weathering. The weather resistance includes UV and oxidative stability, as well as stability to aqueous acids, fats, and oils. The behavior toward fats and oils also enables the use of silicone resins as release coatings. Silicone resins are water repellent and, like all silicone polymers, permeable to gases. They also show good wetting properties on many inorganic and organic substrates.

After curing, silicone resins exhibit good adhesion to most surfaces due to chemical or physical bonding to the surface. They are electrically nonconductive, and their electrical and mechanical properties show little dependence on temperature. The susceptibility to oxidation, especially of the pure methylsilicone resins, is low. Decorative and gloss-retaining effects, as well as elasticity and pigment compatibility, increase with increasing phenyl content, but the hardness decreases simultaneously. By varying the organic substituents of the resin, desirable resin properties can be adjusted selectively. Provided certain preconditions are met, silicone resins are physiologically inert [157, 158].

5.5. Applications

Silicone resins have a wide range of applications, the most important being paint binders and masonry protection. In the electrical industry, silicone resins are used as binders for compression molding compounds and laminates (glass fabric, mica), as well as for impregnating resins and insulating varnishes. Electronic chips and other components are protected from moisture, dust, and chemicals by silicone-resin-based coatings. Unlike conventional coating materials such as phenolic or epoxy resins, silicones exhibit both water repellency and good heat resistance. For high-temperature applications [155], methylsilicone resins are generally used. Silicone-resin-bonded mica laminates are widely used as supporting elements for resistance heaters. In anticorrosion paints, the resins are pigmented with aluminum, zinc dust, or micaceous iron oxide. In decorative coatings, where paint properties such as pigmentability, elasticity, adhesion, and gloss are the overriding concern, methylphenylsilicone resins are generally used. The required temperature resistance of between 250 °C (household appliances) and 400 °C (ovens) can be met. Silicone resins are also used as release coatings (e.g., for baking pans). Silicone resins can be used to protect buildings against graffiti.

Heat-resistant stoving paints, such as those used for the decorative coating of household appliances or for coating metal cladding panels by the coil-coating process [159], frequently consist of silicone–polyester combination resins. The silicone resin gives increased resistance to weathering and chalking, especially in the coil-coating field.

Special silicone resins are used as paint additives to improve specific paint properties [160]. Silicone resins are used in masonry protection. Porous inorganic building materials are rendered water repellent by impregnation with silicone resin. This also increases the resistance to chemicals while maintaining gas and water vapor permeability.

The effectiveness of silicone resins in the protection of building materials depends on the depth of penetration into the substrate. To meet the demand for solvent-free products, silicone emulsions and microemulsions are available [161, 162]. Microemulsions contain up to 20 wt % solids and have particle sizes of 10–80 nm; they have good penetrating power for most building materials. Silicone resin microemulsions are diluted with water before use.

Other applications of silicone resins are as scratchproof coatings for glass and plastics and in polishes. QM resins are commonly used for coatings and in elastomer reinforcement.

6. Block and Graft Copolymers

Polysiloxanes (A) and organic polymers (B) can be chemically linked to give block or graft copolymers. Block copolymers with linear AB, ABA, or $(AB)_n$ structures can be prepared selectively. Grafting reactions lead to statistically distributed branches on the base polymer.

6.1. Polysiloxane – Polyether Copolymers

The most important siloxane-containing copolymers are the polysiloxane – polyether copolymers [163–170]. Both linear and branched block copolymer structures are possible. Branched copolymers consisting of a polysiloxane core and linear polyether side chains are of particular importance. The polyether segments can be polyoxyethylene, polyoxypropylene, or poly(oxyethylene – oxypropylene) units (see → Polyoxyalkylenes). The copolymers can have Si–O–C as well as Si–C linkages between the polysiloxane and polyether segments. Copolymers with Si – O – C linkages are produced by reaction of siloxanes having reactive end groups with polyether alcohols, for example,

$$\equiv\mathrm{Si{-}X} + \mathrm{HO(CH_2CH_2O)}_n\mathrm{R} \longrightarrow \equiv\mathrm{Si{-}O{-}(CH_2CH_2O)}_n\mathrm{R} + \mathrm{HX}$$

X = Cl, OR, OAc, NR_2, H

The Si – C linked copolymers are generally produced by platinum-catalyzed hydrosilylation of polyethers having unsaturated end groups with SiH-containing siloxanes

$$\equiv\mathrm{SiH} + \mathrm{CH_2{=}CHCH_2O(CH_2CH_2O)}_n\mathrm{R} \longrightarrow \equiv\mathrm{SiCH_2CH_2CH_2O(CH_2CH_2O)}_n\mathrm{R}$$

R = H, alkyl

Polysiloxane – polyether copolymers have surfactant properties. The surface tension of aqueous solutions of these copolymers can be as low as 20 mN/m (see Table 9) [169], which is 5 – 8 mN/m lower than comparable organic surfactants. The chain length and structure (linear or branched) of the siloxane component have a strong influence on surface tension, whereas the chain length of the polyether segment has only a minor effect.

Unlike most organic surfactants, polysiloxane – polyether copolymers can also lower the surface tension of organic liquids [169, 170].

Polysiloxane – polyether copolymers are used in large quantities as foam stabilizers in the production of polyurethane foams [164–170]. In this application, the role of the copolymer includes not only lowering the surface energy to promote nucleation, buildup, and stabilization of the foam cells, but also emulsification of the polymer and blowing agent. Branched copolymers with poly(oxyethylene – oxypropylene) segments have a particularly good combination of these properties. The foam-stabilizing action of these copolymers can also be applied to phenolic resin foams [171].

Table 9. Surface tension of aqueous solutions of polysiloxane – polyether copolymers

Copolymer*	Surface tension, mN/m
$[MD]_2D'(EO)_{11.8}Me$	20.0
$MD_3M'(EO)_{11.8}Me$	20.2
$D_3D'(EO)_{8.0}Me$	20.5
$MM'(EO)_{4.8}Me$	21.4
$D_4[M'(EO)_{8.5}Me]_2$	23.4
$[D'(EO)_4Me]_4$	26.3
Water alone	72.0

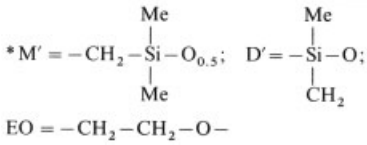

6.2. Other Block Copolymers

Siloxane-containing block copolymers are known with a variety of organic building blocks, such as polycarbonate [172–175], polyurea [176], polyimide [177], polyester [178], polyurethane [179, 180], polyarylether [181, 182], polylactam [183], polystyrene [184–191], polybutadiene [192], and carbodiimide [193].

Suitable synthetic routes involve terminally reactive polysiloxanes, which either are present during polymerization of the organic monomer [172]:

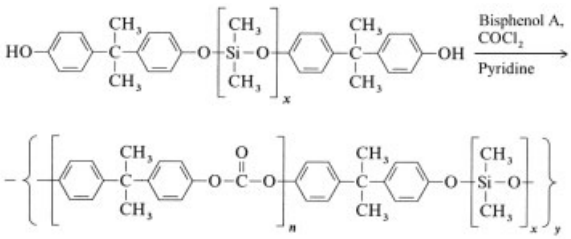

or react with stoichiometric amounts of functional organic polymer blocks [178]:

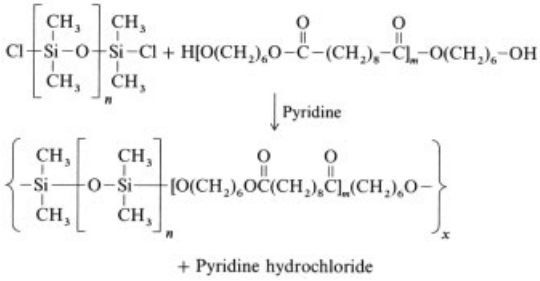

Block copolymers can be prepared from some organic monomers by means of living anionic polymerization (e.g., styrene, α-methylstyrene, and butadiene). In this process, the organic monomer is polymerized anionically; then the still-active ends of the polymer are used to polymerize hexamethylcyclotrisiloxane under nonequilibrating conditions [163, 184–189]. By varying this process, AB, ABA, or $(AB)_n$ block copolymers can be produced selectively. Polysiloxanes with bis(silylpinacolate) groups initiate radical polymerization of certain organic monomers to give block copolymers [190, 191].

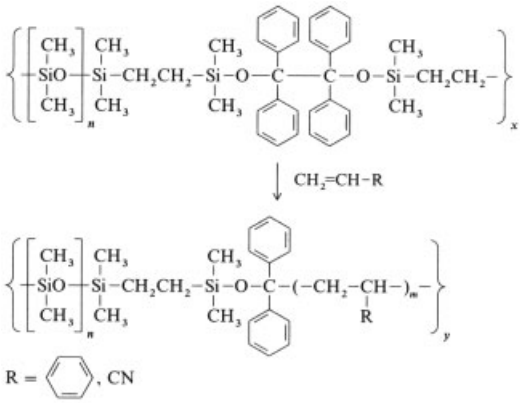

Silarylenesiloxane – siloxane block copolymers with different silylarylene groups and various organic substituents on the siloxane segments are known:

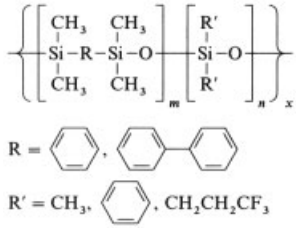

The silarylene segments readily form crystalline phases, which give the copolymers elastomeric properties. Copolymers have been synthesized with good mechanical properties, stability to oxidation at elevated temperature, and good chemical resistance [194–197].

The polysiloxane and polyorgano segments of block copolymers are usually incompatible. Most block copolymers therefore have a detectable two-phase morphology. For example, in dynamic – mechanical studies on polysiloxane – polystyrene $(AB)_n$ copolymers, two glass transition temperatures were detected, one at $-110\,°C$ for the polysiloxane phase and a second at $+90\,°C$ for the polystyrene phase [189].

Block copolymers with thermoplastic segments can behave as thermoplastic elastomers (TPE). The thermoplastic phase causes physical cross-linking of the elastic polysiloxane segments [198, 199]. The morphology and mechanical properties of the TPEs depend critically on the content of thermoplastic and the chain length of the siloxane segments. The mechanical properties obtained are in some cases superior to those of conventional filled silicone elastomers.

6.3. Graft Copolymers

The radical copolymerization of polysiloxanes with various vinyl monomers (e.g., styrene, acrylates, or vinyl acetate) leads to multiphase graft copolymers [200–206]. Polysiloxanes with reactive functional groups on the siloxane chain, such as mercaptopropyl and vinyl groups, can also be grafted onto organic polymers with reactive groups.

Graft copolymers can stabilize dispersions of organic thermoplastics in polysiloxanes, which can then behave as reinforced elastomers. If hydroxyl-end- capped polysiloxanes are grafted

with thermoplastics, copolymers are obtained that can be formulated with condensation cross-linking agents to give one- or two-component RTV systems. These elastomers can have good mechanical properties and have various applications as embedding materials and coating compounds [207, 208].

6.4. Applications of Block and Graft Copolymers

Polysiloxane – polyether copolymers are important surfactants that are used not only as foam stabilizers but also as leveling agents in paints, release aids, textile auxiliaries, and additives for cosmetics and polishing agents. Siloxane-containing copolymers have various applications as modifiers for thermoplastics. They improve the impact resistance [209] and surface properties [210], and act as flame retardants [211]. The stability of polysiloxanes toward oxidation at elevated temperature and the good low-temperature flexibility are advantageous in these applications. Because of the high oxygen permeability of polysiloxane-containing copolymers, they can be used for making contact lenses. The high gas permeability of polysiloxanes leads to applications as gas-separation membranes. Their use in numerous biomedical applications such as dialysis membranes is described in [141].

7. Analysis

The wide range of chemical and physical methods available for analysis of silicones is outlined below. A comprehensive updated description of these chemical analyses, spectroscopic, chromatographic, X-ray, and microscopic methods, together with ca. 1400 references to the original literature, can be found in a standard work on silicone analysis [212].

Silicon Content. In many cases, determination of the silicon content is sufficient to characterize the silicone content of the substance to be analyzed. Gravimetric, spectrophotometric, and atomic spectroscopic methods are available. To an increasing extent, however, species-specific analysis is required. This is performed by gas chromatography or by combining separation methods with spectroscopic determination.

Determination of Total Silicon Content by Gravimetric and Spectrophotometric Methods. The standard gravimetric method for determining total silicon involves fusing the sample with Na_2O_2 in a nickel bomb. The fusion cake is treated with water, and silicon is separated as SiO_2 by repeated fuming with hydrochloric acid. The final determination is performed by differential weighing before and after removal of the silicon as SiF_4 by fuming with $HF-H_2SO_4$ [213].

Another established gravimetric method is to convert the silicon in the fusion solution to molybdosilicate, to precipitate the latter with organic nitrogen compounds (e.g., pyridine hydrochloride), and to heat the precipitate to give the stoichiometrically defined mixed oxide $SiO_2 \cdot 12MoO_3$. Because of the favorable weight increase, this method is particularly suitable for determining low silicon contents [214].

Very low silicon contents are determined spectrophotometrically after conversion to molybdosilicate. Both the yellow coloration at 390 nm and the blue coloration at 815 nm obtained after reduction with ascorbic acid or 1-amino-2-naphthol-4-sulfonic acid are suitable for this purpose [215].

Determination of Total Silicon Content by Atomic Spectroscopic Methods. Organosilicon compounds can be determined in organic solvents by atomic absorption spectrometry in a nitrous oxide – acetylene flame or by atomic emission spectrometry in an argon plasma. The most suitable solvents are methyl isobutyl ketone, toluene, and xylene. These methods are readily applicable only to nonvolatile organosilicon compounds. Volatile compounds give variable signals, which causes problems in calibration.

Determination of Functional Groups.

Si – Cl. The Si – Cl groups are hydrolyzed to Cl^-, which is titrated argentometrically. Potentiometric end-point detection is recommended, with a chloride-selective or silver electrode as indicator electrode [216]. With low chloride concentrations, the titration should be performed in acetic acid – water mixtures (ca. 4: 1).

Si – H. The Si – H bond is reacted with excess oxidizing agent (e.g., $HgCl_2$ [217], bromine

Table 10. Important IR bands of silicones

Group	Wave number, cm^{-1}, and intensity ∗ of band		
≡SiOH	3685	(m)	free silanol groups
	3200 – 3500	(s)	associated silanol groups
≡SiH	2100 – 2300	(s)	
$\equiv SiCH_3$	1250 – 1280	(s)	
	860 – 750	(s)	
$\equiv SiC_6H_5$	1120	(s)	
	1590	(m)	
≡Si−O−Si≡	1130 – 1000	(s)	broad, possibly several maxima
$\equiv Si-CH{=}CH_2$	1400	(s)	

∗ s = strong; m = medium.

[218], or bromosuccinimide [219]), and the excess oxidizing agent is determined iodometrically. The solvents used are di-*sec*-butyl ether, carbon tetrachloride, acetic acid, or mixtures thereof. The Si – H bond can also be determined gas volumetrically after decomposition by alkali. *n*-Butanol is suitable as solvent and sodium butoxide as decomposition reagent.

Si – OH is determined gas volumetrically, the method being based on the formation of methane in the Zerewitinoff reaction with methylmagnesium chloride or on the formation of hydrogen in the reaction with lithium aluminum hydride [216].

Si – vinyl is determined gas chromatographically as ethylene after alkaline cleavage of the Si – C bond [220]. Iodometric titration of the Si – vinyl bond is possible after reaction with excess bromine solution in a 1: 1 acetic acid – carbon tetrachloride mixture. In the case of polymers, $HgCl_2$ may be required as catalyst. The method is not applicable in the presence of compounds containing other unsaturated groups or SiH groups [221].

Infrared Spectroscopy. The effectiveness of IR spectroscopy in qualitative and quantitative analysis has been increased markedly by the use of computer-interfaced FT-IR spectrometers. It is a rapid and easily used method for qualitative characterization, even of mixtures. The very specific IR bands of silicones allow their detection in a wide variety of preparations. Infrared spectroscopy thus complements other chemical and physical methods. Aside from the usual absorption techniques, ATR (attenuated total reflection) and FMIR (frustrated multiple internal reflection) are important for the investigation of cross-linked, insoluble silicones (rubbers, coatings).

Quantitative analysis can also be conveniently performed by using the specific IR bands of silicones. The intense $Si-CH_3$ band at 1258 cm^{-1} is suitable for determination of the silicone content in the principal and trace ranges. The Si – OH content is determined from the intensity of the free SiOH band at 3685 cm^{-1} in dilute CCl_4 solutions by differential spectroscopy against an SiOH-free siloxane [222]. Another method uses the Si – OD bands after deuteration [223].

Table 10 lists some particularly prominent IR bands of silicones. Further details can be found in [212].

Nuclear Magnetic Resonance Spectroscopy. Of the three possible methods, ^{1}H, ^{13}C, and ^{29}Si NMR spectroscopy, ^{1}H and ^{29}Si are preferred for the analysis of silicone polymers. The recording of a ^{1}H NMR spectrum requires 3 – 10 min, and that of a ^{29}Si NMR spectrum 3 – 10 h. The ^{13}C spectrum usually yields no information beyond that of ^{1}H NMR, while requiring a longer acquisition time.

^{1}H NMR. The typical chemical shifts of substituents bonded to silicon permit both qualitative identification and quantitative determination of the structural groups in silicone polymers. Table 11 lists typical shift values.

Use of *^{29}Si NMR* spectroscopy provides detailed information on the silicone backbone (see also Table 12). It enables:

1) Quantification of the proportions of M, D, T, and Q units in the polymer

2) Qualitative and quantitative analysis of neighboring groups
3) Determination of the proportion of functional groups and end groups in the polymer
4) Conclusions about asymmetry, tacticity, ring conformations, etc.

Table 11. Typical ^{1}H NMR chemical shifts of silicone polymers (relative to $SiMe_4$, solvent $CDCl_3$)

Group	Chemical shift, ppm
$\equiv Si-CH_3$	0.0
$\equiv Si-CH_2-$	0.5
$\equiv Si-OCH_3$	3.5
$\equiv Si-H$	4.7
$\equiv Si-CH{=}CH_2$	5.8–6.2
$\equiv Si-C_6H_5$	7.3 (intensity 3); 7.6 (intensity 2)

Table 12. Typical ^{29}Si NMR shifts in silicone polymers (relative to $SiMe_4$, solvent $CDCl_3$)

Group		Chemical shift ∗, ppm
$-Si(-CH_3)_3$	(M)	+ 8
$-Si(-CH_3)_2(-H)$		− 5
$-Si(-CH_3)_2(-CH{=}CH_2)$		− 5
$-Si(-CH_3)_2(-OH)$		− 10
$-Si(-CH_3)_2(-OCH_3)$		− 12
$=Si(-CH_3)_2$	(D)	− 20
$=Si(-CH_3)(-H)$		− 35
$=Si(-C_6H_5)_2$		− 45
$=Si(-CH_3)(-OCH_3)$		− 56
$\equiv Si(-CH_3)$	(T)	− 65
$\equiv Si(-C_6H_5)$		− 78
$=Si=$	(Q)	−100

∗ Typical values when all other substituents of Si are attached via $-O-Si$.

8. Toxicology

The toxicology of silicones, especially polydimethylsiloxanes, has been thoroughly studied because they are used in medicine and medical technology, as well as in cosmetics.

The inertness of silicones toward warm-blooded animals has been demonstrated in a number of tests [224]. A single large oral dose of PDMS has only a laxative effect in test animals. The LD_{50} in rats is > 50 g/kg. After single dermal or inhalative administration, no effects on animals were observed. Skin contact with PDMS does not cause irritation. After contact with the eyes, temporary irritation of the conjunctiva and possible lachrymation can occur.

Studies of chronic oral administration of PDMS to various animals did not reveal harmful effects. Oral intake of PDMS by humans (e.g., in the treatment of flatulence) causes no undesirable side effects. Regular dermal application in humans (e.g., as a skin cream) is also tolerated without problems, and sensitization has not been observed. The inhalation of PDMS-containing aerosol for 90 d by experimental animals had no effect. However, injection into the trachea led to pathological changes in the lungs. In animals, subcutaneously injected PDMS remains at the place of injection in subcutaneous cysts. Acute and chronic studies show that polymeric siloxanes cannot penetrate cells, and that cleavage of the Si–C or Si–O bonds does not occur in the body.

Various tests have given no indication of mutagenicity.

Studies on the carcinogenic effect of PDMS after oral administration in rats and mice showed no positive findings.

Repeated subcutaneous injection of PDMS in animals led to increased tumor formation at the site of injection. The tumors were not attributed to the chemical activity of PDMS but to the physical properties of the substance.

Implantation of weakly cross-linked silicone gels under the skin of human beings is currently a subject of controversy. Cases are reported in which health problems have arisen.

A panel of experts of the FAO and WHO considers the daily intake by human beings of 1.5 mg of PDMS per kilogram of body weight in the form of food additives to be unobjectionable. The molecular mass of the PDMS used must exceed 200.

In Germany, silicones are permitted in the production of consumer articles for food contact, provided Recommendation XV of the BGA is followed. Similar EC regulations are currently being prepared on the use of silicones in the production of consumer articles that come in contact with foods.

In the case of organofunctional polysiloxanes, toxic effects cannot be excluded, because of the reactivity of the organofunctional groups. Polysiloxanes containing trifluoropropyl groups form toxic decomposition products when heated above 280 °C [225].

9. Environmental Aspects

Introduction into the Enviroment, Behavior in the Environment. Owing to their diverse applications, silicones, especially silicone fluids, enter the environment from diffuse sources (e.g., in wastewater).

Polydimethylsiloxanes are nonbiodegradable [226]. However, PDMS in wastewater is largely eliminated in treatment plants by absorption on sewage sludge [227]. Sewage sludge is burned or applied, either as compost or directly, to areas under agricultural use. Traces of siloxanes have been detected in soil [228]. As a result of dumping sewage sludge at sea, siloxanes enter marine sediments [229].

Siloxanes in natural bodies of water are adsorbed on suspended particles and deposited in the sediment [230]. Traces of siloxanes have been found in sediment [228, 229, 231].

Abiotic degradation paths of siloxanes have been demonstrated in soil and aqueous media. Siloxanes adsorbed on soil are degraded by the catalytic effect of certain clay minerals [232]. Siloxanes dissolved in water (e.g., siloxanols) can be degraded to silicates by indirect photochemical reaction [233, 234].

Oligomeric siloxanes can enter the atmosphere due to their volatility. In the atmosphere these compounds are degraded with a half-life of a few days by photochemically induced oxidation [235–237].

After use, solid silicones (e.g., silicone rubber, resins) are disposed of as waste or incinerated. On complete combustion, SiO_2, CO_2, and H_2O are formed. Solid SiO_2 is removed from the flue gas by dedusting. It is not known whether silicones are degraded (biotic or abiotic) in landfill sites.

Behavior toward Organisms in the Environment. The solubility of PDMS in water is very low. Polydimethylsiloxanes dissolved in water have no harmful effects on aquatic organisms (plankton, crustacea, mussels, fish) [238]. Silicone fluid administered with food exhibits no effect on fish [239]. The symbiosis of microorganisms (bacteria, algae, protozoa) is also not disturbed by siloxanes dissolved in water [230].

Polydimethylsiloxanes in the form of an emulsion enter surface water at higher concentrations than the solubility of pure PDMS. In tests, emulsion concentrations up to 10 000 ppm PDMS in water exhibited toxic effects on fish, but these are attributed to the emulsifier [226, 240, 241].

Sediment enriched in PDMS (up to 1000 mg/kg) has no effect on organisms living in the sediment [242].

Polymeric siloxanes of high molecular mass do not accumulate in aquatic organisms, even when they are fed PDMS- containing nutrients [241, 243].

Oligomeric siloxanes of molecular mass up to ca. 1000 are absorbed by tissue and accumulated. After transfer of the fish to uncontaminated water, absorbed siloxanes were rapidly eliminated from the tissue [243–246].

Summary. Liquid and volatile siloxanes enter the environment as a result of use. Soil (sewage sludge) and sediment are sinks, and siloxanes are found in trace amounts in these environmental compartments. Siloxanes have no marked harmful effects on organisms in the environment. Therefore they are assumed not to represent an environmental hazard.

10. Economic Aspects

Silicones have been produced industrially since the early 1940s. With the industrial-scale realization of methylchlorosilane synthesis (Rochow direct synthesis), the basis was created for a rapid increase in sales worldwide, beginning in the United States. Given the considerable innovation potential of this product group, this trend is likely to continue.

One reason for this development is the diverse molecular structure of siloxanes, and the intimate interrelationship of structure and properties not obtainable for any other class of polymers. Thousands of industrial products have been developed, ranging from fluids to rubbers and resins. Silicones as such or in various formulations, are used in most industries and spheres of life.

Worldwide silicone production can be estimated from the production of organochlorosilane precursors. From 1 kg of dimethyldichlorosilane, about 0.5 kg of dimethylsiloxane is obtained. Thus, an organochlorosilane production of ca. 800 000 t/a in 1991 corresponds to a

siloxane output of ca. 400 000 t/a. Actual sales are higher due to formulation of silicones with fillers, solvents, and auxiliary chemicals.

The sales of > 600 000 t/a of silicone products in 1992 represented a business volume of ca. $ 5.5×10^9. The United States is the largest market (ca. 40 %), followed by Europe and Japan.

All economically important regions of the world have integrated silane – silicone production facilities. The individual producers are as follows:

United States:	Dow Corning
	General Electric
	Union Carbide
	Wacker (no silane production)
Germany:	Bayer
	Goldschmidt (no silane production)
	Hüls
	Wacker
France:	Rhône-Poulenc
United Kingdom:	Dow Corning
Japan:	Shin Etsu
	Toray Silicones
	Toshiba Silicones

Brazil, Mexico, Italy, Spain, China, the CIS, and India also have production facilities for silicones and in some cases organochlorosilanes, which however are far less important than those in the United States, Europe, and Japan.

11. References

1. Rhône-Poulenc, US 2 556 897, 1947 (A. F. Bidaud).
2. General Electric Comp., US 2 741 630, 1955 (C. E. Reed, J. M. Tome).
3. Wacker, US 4 032 557, 1975 (H. Spork, R. Strasser, R. Riedle).
4. Bayer, US 4 060 537, 1975 (G. Maass, H. Luecking, W. Buechner, B. Degen).
5. T. C. Kendrick, B. Parbhoo, J. W. White in *The Chemistry of Organic Silicon Compounds,* Wiley, New York 1989, pp. 1289 – 1361.
6. P. V. Wright in K. I. Ivin, T. Saegusa (eds.): *Ring-Opening Polymerisation,* **vol. 2,** chap. 14, Elsevier, New York 1984.
7. D. I. Hurd, R. C. Osthoff, M. L. Corrin, *J. Am. Chem. Soc.* **76** (1954) 249.
8. I. Chojnowski, M. Mazurek, *Makromol. Chem.* **176** (1975) 2999 – 3023.
9. Wacker, EP 0 258 640, 1986 (J. Burkhardt, W. Strekkel, A. Boeck).
10. Wacker, EP 0 208 285, 1985 (J. Burkhardt).
11. Wacker, US 3 839 388, 1972 (E. Wohlfahrt, S. Nitzsche, W. Hechtl).
12. Wacker, US 4 203 913, 1975 (J. Burkhardt, K. Wegehaupt).
13. Wacker, US 3 398 176, 1964 (S. Nitzsche, M. Wick, K. Wegehaupt).
14. E. D. Brown, J. B. Carmichael, *J. Polym. Sci. Part B* **3** (1965) 473.
15. A. J. Barry, *J. Appl. Phys.* **17** (1946) 1020 – 1024.
16. E. L. Warrick, W. A. Piccoli, F. O. Starr, *J. Am. Chem. Soc.* **77** (1955) 5017 – 5018.
17. Y. Ito, S. Shishido, *J. Polym. Sci Polym. Phys. Ed.* **11** (1973) 2283 – 2289.
18. T. Kataoka, S. Ueda, *J. Polym. Sci. Polym. Chem. Ed.* **5** (1967) 3071 – 3089.
19. H. Steinbach, C. Sucker, *Colloid Polym. Sci.* **255** (1977) 452 – 459.
20. H. Steinbach, C. Sucker, *Colloid Polym. Sci.* **252** (1974) 306 – 316.
21. H. Meyer, *Farbe + Lack* **97** (1991) 301 – 305.
22. P. Preiss, *Seifen Oele Fette Wachse* **116** (1990) 175 –180.
23. J. Dirnböck, P. Preiss, H.-W. Schiffer, R. Schmitz, *ETZ Elektrotech. Z.* **105** (1984) 832 – 835.
24. Bayer, DE 3 915 066, 1989 (G. Marquardt, P. Preis).
25. J. Roidl, *Seifen öle Fette Wachse* **109** (1983) 91 – 94.
26. E. L. Warrick, O. R. Pierce, K. E. Polmanteer, J. C. Saam, *Rubber Chem. Technol.* **52** (1979) 437 – 525.
27. B. R. Trego, H. W. Winnan, *RAPRA Rev. Rep.* **3** (1990) no. 31, 1 – 99.
28. K. E. Polmanteer, *Rubber Chem. Technol.* **61** (1988) 470 – 506.
29. L. Markova, *Intern. Polym. Sci. Tech.* **12** (1985) no. 2, 565 – 571.
30. W. Oppermann, *Prog. Colloid & Polym. Sci.* **75** (1987) 49 – 54.
31. S. Wang, J. E. Mark, *ACS-Meeting,* Cincinnati, Oct. 1988, paper no. 18.
32. U. Eisele, *Gummi Asbest Kunstst.* **33** (1980) no. 3, 165 – 174.
33. S. Wolff, J.-B. Donner, *Rubber Chem. Technol.* **63** (1990) 32 – 45.
34. R. Bode, H. Ferch, H. Fratzscher, *Gummi Asbest Kunstst.* **20** (1967) no. 12, 699 – 706.
35. H. L. Chapman, M. A. Lutz, K. E. Polmanteer, *Rubber Chem. Technol.* **58** (1985) 939 – 963.
36. S. Wang, P. Xu, J. E. Mark, *Rubber Chem. Technol.* **64** (1991) 746 – 759.
37. Y. P. Ning, J. E. Mark, *ACS-Meeting,* Los Angeles, 1985, paper no. 59.

38. Y. Todani, A. Ueda, *Nippon Gomu Kyokaishi* **50** (1977) no. 6, 379 T27 – T33.
39. C. M. Roland, *Rubber Chem. Technol.* **62** (1989) 880 – 895.
40. A. Pouchelon, P. Vondracek, *Rubber Chem. Technol.* **62** (1989) 788 – 799.
41. Dow Corning, DE 29 34 202, 1978 (M. C. Murray).
42. Stauffer Wacker Silicones, EP 184 649, 1985 (M. J. Streusand).
43. Dow Corning, EP 382 188, 1990 (A. L. Peterson).
44. M. R. Toub, *ACS-Meeting,* Los Angeles 1985, paper no. 99.
45. T. M. Aminabhavi, P. E. Cassidy, *Rubber Chem. Technol.* **63** (1990) 451 – 471.
46. Wacker, US 40 08 198, 1977 (J. H. Burkhardt, Krohberger, J. Patzke).
47. M. Aranguren, C. Makosko, B. Thakhar, M. Tirell, *Mater. Res. Soc. Symp. Proc.* **170** (1990) 303 – 308.
48. Toray, DE 25 54 498, 1975 (K. Kishimoto, Y. Koda, S. Sasaki, M. Suzuki).
49. Dow Corning, US 45 28 313, 1984 (T. J. Swihart, J. E. Jones).
50. N. Grassie, I. G. Mac Farlane, *Eur. Polym. J.* **14** (1978) 875 – 884.
51. Wacker, US 4 701 490, 1985 (J. Burkhardt, W. Rauchberger).
52. Bayer, DE 28 47 481, 1978 (H. Steinberger, W. Michel, W. Kniege).
53. Perrenatorwerk, DE 29 09 462, 1980 (P. Hagen, R. Jonas).
54. General Electric, US 48 33 190, 1989 (J. A. Cella, E. A. O'Neil, D. A. Williams).
55. Dow Corning, EP 329 332, 1988 (R. R. Buch, M. A. Cabey, C. M. Monroe).
56. Toray, US 4 156 674, 1979 (S. Sumimura).
57. Bayer, EP 51 212, 1981 (J. Ackermann, W. Rauer, W. Kniege).
58. K. Lagarde, L. Lahaye, *Eur. Polym. J.* **13** (1977) 769 – 774.
59. M. R. McLaury, *J. Fire & Flammability* **10** (1979) 175 – 198.
60. C. M. Blow, *Rubber Tech. a. Manuf.,* Butterworth, London 1971.
61. M.-J. Ziemelis, J. C. Saam, *Macromolecules* **22** (1989) 2111 – 2116.
62. E. M. Barrell II., R. Hawkins, A. A. Fukushima, J. F. Johnson, *J. Polym. Sci. Polym. Symp.* **71** (1984) 189 – 202.
63. M. L. Dunham, D. L. Bailey, R. Y. Mixer, *Ind. Eng. Chem.* **49** (1957) no. 9, 1373 – 1376.
64. K. Beshah, J. E. Mark, A. Himstedt, J. L. Ackerman, *Polym. Prepr. Am. Chem. Soc. Dir. Polym. Chem.* **23** (1982) 52 – 53.
65. J. P. Bop, B. Maillard, C. Filiatre, J. J. Villenave, *Thermochim. Acta* **58** (1982) 155 – 167.
66. General Electric, DE 36 19 206, 1986 (E. R. Evans).
67. Dow Corning, US 3 445 420, 1966 (G. Kookootsedes, E. Plueddemann).
68. Rhône-Poulenc, FR 1 198 749, 1958 (L. Ceyzeriat).
69. Wacker, DE-AS 1 255 924, 1966 (M. Wick, P. Hittmair, E. Wohlfahrt, S. Nitzsche).
70. Dow Corning, US 3 334 067, 1966 (D. R. Weyenberg).
71. General Electric, US 4 417 042, 1982 (J. J. Dziark).
72. Dow Corning, US 3 189 576, 1962 (E. Sweet).
73. Bayer, US 4 434 283, 1984 (H. Sattlegger, K. Schnurrbusch, B. Degen, T. Achtenberg).
74. V. V. Severnyi, R. M. Minas'yan, J. A. Makarenko, N. M. Bizynkova, *Vysokomol. Soedin Ser. A.* **18** (1976) 1464 – 1471 (engl.).
75. F. W. van der Weij, *Macromol. Chem. (Oxford)* **181** (1980) 2541 – 2548.
76. R. Bradley: *Radiation Tech. Handbook,* chap 7, Marcel Dekker, New York, pp. 139 – 176.
77. U. Müller, H.-J. Timpe, *Gummi Fasern Kunstst.* **41** (1988) 1131 – 1138.
78. S. A. H. Mohammed, J. Walker, *Rubber Chem. Technol.* **59** (1986) 482 – 496.
79. Dow Corning, EP 031 640, 1979 (G. R. Homan, C. Lee).
80. Rhône-Poulenc, EP 258 159, 1990 (M. Alliot-Lugaz, G. Roullet).
81. G. Matthews: *Polymer Mixing Tech.,* Appl. Sci. Publ., London 1982.
82. Shin Etsu, US 4 888 374, 1988 (M. Takahashi, T. Yoshida).
83. Shin Etsu, EP 423 717, 1990 (Y. Inoe, M. Takahasi, K.-I. Takita, T. Yoshida).
84. Dow Corning, US 41 04 351, 1978 (J. D. Blizzard, C. M. Monroe).
85. J. D. Van Drumpt, *Rubber World* **197** (1988) no. 6, 33 – 41.
86. Amer. Optical Corp., US 39 96 189, 1976 (E. A. Travnicek).
87. Dow Corning, US 47 85 047, 1987 (J. D. Jensen).
88. M. T. Maxson, C. L. Lee, *Gummi Fasern Kunstst.* **39** (1986) 532 – 539.
89. D. J. Cornelius, C. M. Monroe, *Polym. Eng. Sci.* **25** (1985) no. 8, 467 – 473.

90. W. Buechner, R. Schliebs, G. Winter, K. H. Buechel: *Industrial Inorganic Chemistry,* VCH Verlagsgesellschaft, Weinheim 1989, pp. 299 – 302.
91. Kanegafuchi, EP 0 333 222, 1988 (H. Wakabayashi, H. Iwakiri, K. Tamai, K. Isayama).
92. J. C. Saam, D. Graiver, M. Baile, *Rubber Chem. Technol.* **54** (1981) 976 – 987.
93. Bayer, EP 143 877, 1985 (W. Grape, F. Saykowski, O. Schlak, T. Würminghausen).
94. C. G. Cash, *Appl. Polym. Symp.* **14** (1970) 47 – 50.
95. M. J. Owen, J. M. Klosowski, *Polym. Sci. Technol. (Plenum)* **37** (1988) 281 – 291.
96. Bayer, US 4 458 055, 1984 (H. Sattlegger, K. Schnurrbusch, H.-G. Metzinger).
97. Bayer, DE 1 258 087, 1966 (H. Sattlegger, W. Noll, K. Damm, H.-G. Gölitz).
98. *Release Liner Markets and Technology,* Conference Proceedings, 14 – 15 May 1991, Amsterdam, Pira International.
99. H.-J. Jahn, *Gummi Asbest Kunstst.* **21** (1968) no. 9, 469 – 477.
100. A. Watson, *Release Liner Markets and Technology,* Conference Proceedings 14 – 15 May 1991, Amsterdam, Pira International.
101. Wacker, US 43 01 056, 1980 (J. Patzke, K.-H. Wegehaupt).
102. Rhône Poulenc, EP 235 048, 1987 (R. Lagarde).
103. Toray, US 41 16 920, 1978 (H. Honma, H. Kakuno).
104. Toray, DE-OS 30 34 232, 1980 (S.-I. Sumimura, S. Miyakoshi).
105. General Electric, EP 415 180, 1990 (E. M. Jeram, B. J. Ward, D. A. Martin).
106. R. E. Drake, *ACS-Meeting,* Minneapolis 1981, paper no. 6.
107. H. Yoshida, J. Watanabe, Y. Zama, *ACS-Meeting,* Detroit Oct. 89, paper no. 122.
108. Dow Corning, EP 380 104, 1990 (K. Kunamatsu, A. Komatsu).
109. Bendix, US 35 38 028, 1970 (C. P. Morgan).
110. J. R. Falender, C. M. Monroe, *ACS-Meeting,* Denver Oct. 1973.
111. J. Umeda, Y. Takenu, J. Watanabe, Y. Funahashi, *ACS-meeting,* Cincinnati Oct. 1988, paper no. 79.
112. Dow Corning, US 47 74 281, 1988 (R. G. Chaffee, C. M. Monroe).
113. K. E. Polmanteer, *Rubber Chem. Technol.* **54** (1981) 1050 – 1080.
114. Dow Corning, DE-AS 10 13 070, 1955 (N. G. Dickmann).
115. Dow Corning, EP 240 162, 1987 (B. I. Gutek).
116. Bridgestone/Toshiba, US 47 14 734, 1985 (Y. Funahashi, T. Hashimoto, A. Maehara).
117. Shin Etsu, US 36 71 480, 1972 (T. Wada, K. Itoh).
118. Shin Etsu, US 39 50 299, 1975 (K. Itoh, N. Kuga, T. Fukuda).
119. J. R. Halladay, R. L. Warley, *134th ACS-Meeting,* Cincinnati Oct. 1988, paper no. 7.
120. T. Suzuki, A. Kasuya, *J. Adhes. Sci. Tech.* **6** (1989) 463 – 473.
121. Toray, EP 244 952, 1987 (M. Saito, K. Shimizu, M. Hamada).
122. Shin Etsu, EP 431 881, 1990 (T. Fukuda, S. Ide, M. Fukushima).
123. Toray, EP 349 897, 1989 (S. Sasaki).
124. General Electric, US 4 020 014, 1977 (A. L. Service, G. Christi).
125. Dow Corning, DE 3 012 772, 1979 (G. Kehrer, W. Smith).
126. Dow Corning, EP 367 562, 1988 (R. L. Cole, M. A. Lutz).
127. H. J. Ott, H.-A. Büscher, D. Skudelny, *Kunststoffe* **70** (1980) no. 3, 156 – 161.
128. General Electric, DE 24 58 507, 1973 (J. H. Wright).
129. W. Taylor, K. A. C. Scott, *Notes Doc.* **95** (1979) no. 2, 259 – 266.
130. C. J. Hilado, C. J. Casey, *J. Fire and Flamm.* **10** (1979) 140 – 168, 227 – 239.
131. J. Lipowitz, M. J. Ziemelis, *J. Fire and Flammability* **7** (1976) 504 – 529.
132. R. R. Buch, *Fire Saf. J.* **17** (1991) 1 – 12.
133. F. R. Eirich: *Rheology,* Acad. Press, New York 1968, p. 229.
134. R. J. Wilcock, J. L. McHale, R. Battino, E. Wilhelm, *Fluid Phase Equilib.* **2** (1978) 225 – 230.
135. Teijin Ltd., EP 94 050, 1983 (F. Ueda, E. Hashimoto, T. Yamada).
136. H. Tajima, T. Masuda, T. Hijashimura, *J. Polym. Sci. Part A,* Polym. Chem. 25 (1987) 2033 – 2042.
137. H. Schuck, *Gummi Asbest Kunstst.* **33** (1980) no. 9, 705 – 715.
138. R. Rautenbach, R. Albrecht, *Chem. Ing. Tech.* **57** (1985) no. 2, 119 – 130.
139. Titmus Eurocon, EP 033 754, 1981 (G. Kossmehl, D. Quast, H. Schafer).
140. Amer. Optical Corp., US 39 96 187, 1976 (E. A. Travnicek).
141. B. Arkles, *CHEMTECH* 1983, no. 9, 542 – 555.

142. Dow Corning, US 36 24 023, 1971 (J. V. Hartlage).
143. Dow Corning, US 44 18 165, 1981 (K. E. Polmanteer, H. L. Chapman).
144. Amer. Optical Corp., DE 26 16 147, 1976 (W. Lambert).
145. General Electric, DE 25 56 252, 1975 (J. H. Wright).
146. H. M. Leeper, R. M. Wright, *Rubber Chem. Technol.* **56** (1983) 523 – 555.
147. A. S. Chawla, *J. Biomed. Mater. Res.* **16** (1982) 501 – 508.
148. M. S. Lucas, D. J. Moore, *J. Prosthet. Dent.* **42** (1979) no. 4, 447 – 451.
149. B. Deubzer in: *Silicone – Chemie und Technologie,* Vulkan-Verlag, Essen 1989, pp. 99 – 116.
150. E. Schamberg, G. Koerner, *Goldschmidt informiert* **63** (1984) no. 4, 49 – 56.
151. K. A. Earhardt, *Paint Varn. Prod.* **62** (1972) no. 1, 35 – 43; **62** (1972) no. 2, 37 – 42.
152. W. Brushwell, *Farbe + Lack* **82** (1976) no. 3, 219 –221.
153. Wacker, EP 0 003 610, 1978 (W. Graf, V. Frey, P. John, N. Zeller).
154. Wacker, EP 0 032 376, 1980 (T. Lindner, N. Zeller, A. Schinabeck, G. Engelsberger, R. Riedle).
155. Goldschmidt, EP 0 167 924, 1984 (H. Giesing et al.).
156. W. A. Finzel: "Properties of High Temperature Silicone Coatings," *J. Prot. Coat. Lin.* **4** (1987) no. 8, 38 – 43.
157. BGesundh. Bl. **23** (1980) 122.
158. BGesundh. Bl. **27** (1984) 190.
159. E. V. Schmid, *Farbe + Lack* **85** (1979) 744 – 748.
160. S. F. Thames, *ACS Symp. Ser. (Appl. Polym. Sci. 2nd Ed.)* **285** (1985) 1117 – 1140.
161. H. Mayer, M. Roth, *Bautenschutz + Bausanierung* **13** (1990) 1 – 4.
162. H. Mayer, I. König-Lumer, G. Kolleritsch, C. Wochinger, *Bautenschutz + Bausanierung* **14** (1991) 27 –33.
163. H. B. Plumb, J. H. Atherton in D. C. Allport, W. H. Jones (eds.): *Block Copolymers,* Allied Sci. Publ., London 1973, chapt. 6.
164. M. J. Owen, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 97 – 103.
165. T. C. Kendrick, B. M. Kingston, N. C. Lloyd, M. S. Owen, *J. Colloid Interface Sci.* **24** (1967) 135 – 140.
166. M. S. Owen, T. C. Kendrick, B. M. Kingston, N. C. Lloyd, *J. Colloid Interface Sci.* **24** (1967) 141 – 150.
167. M. S. Owen, T. C. Kendrick, *J. Colloid Interface Sci.* **27** (1968) 46 – 52.
168. R. J. Bondreau, *Mod. Plast.* **44** (1967) 133.
169. B. Kanner, W. D. Reid, I. H. Petersen, *Ind. Eng. Chem. Prod. Res. Dev.* **6** (1967) no. 2, 88 – 92.
170. B. Kanner, B. Prokai, C. S. Eschbach, G. J. Murphy, *J. Cell. Plast.* **15** (1979) 315.
171. M. J. Owen, C. Denis, *J. Cell. Plast.* **13** (1977) 264.
172. H. A. Vaughn, *J. Polym. Sci. Part B* **7** (1969) 569 –572.
173. R. P. Kambour, J. Polym. Sci. Part B **7** (1969) 573 –577.
174. A. Noshay, M. Matzner, T. C. Williams, *Ind. Eng. Chem. Prod. Res. Dev.* **12** (1973) no. 4, 268.
175. J. S. Riffle, R. G. Freelin, A. K. Banthia, J. E. McGrath, *J. Macromol. Sci. Chem.* **A15** (1981) 967 –998.
176. I. Yilgör, J. S. Riffle, G. L. Wilkes, J. E. McGrath, *Polym. Bull. (Berlin)* **8** (1982) 535 – 542.
177. I. Yilgör et al., *Polym. Prepr. (Am. Chem. Soc. Dir. Polym. Chem.)* **24** (1983) no. 1, 170 – 173.
178. J. J. O'Malley, T. G. Pacansky, W. J. Stauffer, *Macromolecules* **10** (1977) no. 6, 1197 – 1199.
179. Y. Yue-hai et al., *J. Polym. Sci. Polym. Phys. Ed.* **23** (1985) 2319 – 2338.
180. Bayer, DE 2 730 744, 1977 (H.-H. Moretto, A. de Montigny, H. Steinbach, H. Sattlegger).
181. M. Matzner et al., *Appl. Polym. Symp.* **22** (1973) 143 – 156.
182. B. C. Auman, V. Percec, H. A. Schneider, H.-J. Cantow, *Polymer* **28** (1987) 1407 – 1417.
183. P. P. Policasto, P. K. Hernandes, *Polym. Bull. (Berlin)* **16** (1986) 43 – 45.
184. J. C. Saam, D. J. Gordon, S. E. Lindsey, *Macromolecules* **3** (1970) 1.
185. J. W. Dean, *J. Polym. Sci. Part B* **8** (1970) 677 – 679.
186. I. Jansen, G. Lohmann, K. Rühlmann, *Plaste Kautsch.* **31** (1984) 441 – 447.
187. M. Morton, A. A. Rembaum, E. E. Bostick, *J. Appl. Polym. Sci.* **8** (1964) 2707 – 2716.
188. P. Banjaj, S. K. Varshney, A. Misra, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 295 – 309.
189. S. K. Varshney, C. L. Beatty, P. Banjaj, *Polym. Prepr. (Am. Chem. Soc. Dir. Polym. Chem.)* **22** (1981) 321.
190. J. V. Crivello, D. A. Conlon, J. L. Lee, *J. Polym. Sci. Polym.* **24** (1986) 1197 – 1215.
191. J. V. Crivello, J. L. Lee, D. A. Conlon, *J. Polym. Sci. Polym.* **24** (1986) 1251 – 1279.
192. G. Lohmann, K. Rühlmann, *Plaste Kautsch.* **32** (1985) no. 6, 206 – 210.

193. Bayer, DE 2 730 743, 1977 (H.-H. Moretto, H. Steinbach, I. Larking, H. Sattlegger).
194. W. A. Dunnavant, *Inorg. Macromol. Rev.* **1** (1971) 165 – 189.
195. C. U. Pittmann, W. J. Patterson, S. P. McManus, *J. Polym. Sci. Polym. Chem. Ed.* **14** (1976) 1715 – 1734.
196. H. Rosenberg, E. W. Choe in C. E. Carraher, J. E. Sheats, C. U. Pittmann (eds.): *Organometallic Polymers,* Academic Press, New York 1978.
197. Y. Nagase, J. Ochiai, K. Matsui, M. Uchikura, *Polymer* **29** (1988) 740 – 745.
198. I. Yilgör, J. E. McGrath, *Adv. Polym. Sci.* **86** (1988) 1 – 86.
199. J. D. Summers et al., *ACS Symp. Ser.* **360** (1988) 180 – 198.
200. G. Greber, E. Reese, *Makromol. Chem.* **55** (1962) 96.
201. J. R. Falender, C. M. Monroe, *Rubber Chem. Technol.* **47** (1974) 57.
202. J. C. Saam, C. H. Tsai, *J. Appl. Polym. Sci.* **18** (1974) 2279 – 2285.
203. S. D. Smith, J. E. McGrath, *Polym. Prepr. (Am. Chem. Soc. Dir. Polym. Chem.)* **27** (1986) no. 2, 31.
204. Y. Kawakami, R. A. N. Murthy, Y. Yamashita, *Makromol. Chem.* **185** (1984) 9 – 18.
205. G. G. Cameron, M. S. Chisholm, *Polymer* **27** (1986) 437, 1420.
206. T. R. Williams, *J. Appl. Polym. Sci.* **31** (1986) 1293 – 1308.
207. K. Marquardt, F.-H. Kreuzer, M. Wick, *Angew. Makromol. Chem.* **58/59** (1977) 243 – 257.
208. W. Kiel, M. Marquardt, *Gummi Asbest Kunstst.* **30** (1977) no. 2, 76 – 84.
209. J. C. Saam, C. M. Mettler, J. R. Falender, T. J. Dill, *J. Appl. Polym. Sci.* **24** (1979) 187 – 199.
210. M. J. Owen, T. C. Kendrick, *Macromolecules* **3** (1970) 458.
211. R. P. Kambour, H. J. Klopfer, H. J. Smith, *J. Appl. Polym. Sci.* **26** (1981) 847.
212. A. L. Smith (ed.): *The Analytical Chemistry of Silicones,* J. Wiley & Sons, New York 1991.
213. H. R. Shell in I. M. Kolthoff, P. J. Elvin (eds.): *Treatise on Analytical Chemistry,* part II, vol. 2, Intersci. Publ., New York 1962, p. 139.
214. C. Harzdorf, *Fresenius' Z. Anal. Chem.* **227** (1966) 96.
215. H. R. Shell in I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* part II, vol. 2, Intersci. Publ. New York 1962, p. 169.
216. R. C. Smith, N. C. Angelotti, C. L. Hanson, in A. L. Smith (ed.): *Analysis of Silicones,* J. Wiley & Sons, New York 1974.
217. G. Fritz, *Z. Anorg. Allg. Chem.* **280** (1955) 134.
218. G. Fritz, H. Burdt, *Z. Anorg. Allg. Chem.* **317** (1962) 35.
219. C. Harzdorf, *Fresenius' Z. Anal. Chem.* **256** (1971) 192.
220. C. L. Hanson, R. C. Smith, *Anal. Chem.* **44** (1972) 1571.
221. C. Harzdorf, *Fresenius' Z. Anal. Chem.* **276** (1975) 279.
222. G. W. Griffith, *Ind. Eng. Chem. Prod. Res. Dev.* **23** (1984) 590 – 593.
223. E. D. Lipp, *Appl. Spectrosc.* **45** (1991) 477 – 483.
224. BIBRA (The British Industrial Biological Assoc.): *Toxicity Profile: Polydimethylsiloxane,* Carshalton/Surrey 1987.
225. *Chem. Regul. Reporter,* Bureau of National Affairs, Nov. 28, (1986) 1147 – 1148.
226. E. J. Hobbs, M. L. Keplinger, J. C. Calandra, *Environ. Res.* **10** (1975) 397 – 406.
227. H. Friege et al., *Korresp. Abwasser* **36** (1989) 601 – 608.
228. F. Siebert, Thesis, Universität Heidelberg 1988.
229. R. E. Pellenberg, *Mar. Pollut. Bull.* **13** (1982) 427 –429.
230. R. L. Gettings, T. H. Lane, *Dow Corning Int. Rep.* 10005 – 10957 (1982).
231. Environmental Agency Japan, Office of Health Studies, *Chemical Assessment Annual Report 1981.*
232. R. R. Buch, D. N. Ingebrigtson, *Environ. Sci. Technol.* **13** (1979) 676 – 679.
233. C. Anderson, K. Hochgeschwender, H. Weidemann, R. Wilmes, *Chemosphere* **16** (1987) 2567 – 2577.
234. R. R. Buch, T. H. Lane, R. B. Annelin, C. L. Frye, *Environ. Toxicol. Chem.* **3** (1984) 215.
235. Y. Abe, G. B. Butler, T. E. Hagen-Esch, *J. Macromol. Sci. Chem.* **A 16** (1981) 461 – 417.
236. R. Atkinson, *Environ. Sci. Technol.* **25** (1991) 863 – 866.
237. R. Sommerlade, H. Parlar, D. Wrobel, P. Kochs, *Environ. Sci. Technol.* **27** (1993) (in press).
238. P. Maggi, C. Alzien, *Sci. Peche* **269** (1977) 1 – 3.
239. H. Mann, B. Ollenschläger, H. H. Reichenbach-Klinke, *Fisch Umwelt* **3** (1977) 19 – 22.
240. A. W. Hill, J. E. Caunter, G. J. Eales, *Report ICI Brixham Laboratories,* no. BLIB/2392 A (1984).

241. M. Aubert, J. Aubert, H. Augier, C. Gillemaut, *Chemosphere* **14** (1985) 127 – 138.
242. N. C. D. Craig, J. E. Caunter, *Chemosphere* **21** (1990) 751 – 759.
243. A. Opperhuizen, H. W. J. Damen, G. M. Asyee, J. M. D. Van der Steen, *Toxicol. Environ. Chem.* **13** (1987) 265 – 285.
244. W. A. Bruggeman et al., *Toxicol. Environ. Chem.* **7** (1984) 287 – 296.
245. R. B. Annelin, C. L. Frye, *Sci. Total Environ.* **83** (1989) 1 – 11.
246. Springborn Laboratories, SLI Report no. 91-6-3809, Report to Silicone Health Council/USA, Wareham, Mass., 1991.

silicon halides → silicon compounds, inorganic
silicon tetrafluoride → silicon compounds, inorganic

Silk

HELMUT ZAHN, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Germany

ANITA KRASOWSKI, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Germany

1. Introduction 1
1.1. Definitions 1
1.2. History 2
1.3. Life Cycle of the Mulberry Silkworm 2
2. Structure and Properties 2
2.1. Hierarchy and Crystal Structure . . . 3
2.2. Chemical Composition 3
2.3. Properties 5
2.4. Analysis 6
3. Production 6
3.1. Sericulture 6
3.2. Production of Silk Thread 6
3.3. World Production 7
3.4. Silk Yarns – Classification and Testing 7
4. Processing 8
4.1. Degumming 8
4.2. Weighting 9
4.3. Bleaching 10
4.4. Dyeing 10
4.5. Printing 11
4.6. Special Finishing 11
5. Uses and Other Aspects 11
6. References 13

1. Introduction

The importance of silk as a renewable raw material for textiles and other goods cannot be appreciated from its minute weight-based share in the world fiber production, but rather from its value in trade and processing. Despite the triumphal advance of synthetic fibers, silk maintains its place in the raw material market, in the textile and clothing industries, and in retail trade because of its unique properties.

Silk has played an important role not only in political, cultural, and economic history (Silk Road, national costumes), but also in the history of technology (textile machines, Jacquard) and in science (protein chemistry, genetics, and genetic engineering).

1.1. Definitions [13]

The term silk generally refers to the thread spun by certain caterpillars, spiders, and mussels. Only silk spun by the larvae of the silk moth in the form of threads that form a cocoon is of practical importance.

According to DIN 60 001 [14], silk is a natural fiber, classified under the term animal fibers. A distinction is made between fibers from the cocoon of the mulberry silk moth *Bombyx mori* L. (mulberry silk) and those of the wild tussah (tussah silk): Chinese (*Antheraea pernyi* guerin); Japanese (*Antheraea yamamai* guerin); and Indian (*Antheraea mylitta* drury, as well as *Actias selene* HBN.). Other wild silks are Indian eri silk (*Samia ricini* Bsd.) and Indian muga silk (*Antheraea assamensis* Ww.).

1.2. History [15, 16]

According to WIEDEMANN [15], sericulture and the processing of silk fibers have occupied an important place in the cultural history of China for about 5000 years.

There is no doubt that the mulberry silkworm was first domesticated in China. All historical records confirm that this country was the cradle of sericulture. A broken cocoon shell and remnants of silk, velvet, ribbons and other fabrics were found in neolithic sites in the Shanxi and Zhejiang Provinces. These remnants were traced back to 2750 – 2850 B.C. by carbon-14 dating. Further research proved that sericulture was thriving at that early stage along the Huang He and Chang Jiang (Yellow and Yangtze Rivers). During the Shang Dynasty (16th Century B.C.), sericulture and weaving techniques attained high levels.

During the Western Han Dynasty (206 B.C. to 8 A.D.), Chinese sericulture entered its golden

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_095

age, and weaving and printing were developed to a high degree of perfection.

During the reign of Han Wudi (140 – 87 B.C.), large amounts of silk were carried over the "Silk Road" from Xi'an, Shaanxi to the Phoenician parts of the eastern Mediterranean, from where they were shipped to all major cities of the west.

Nevertheless, it took some centuries for sericulture to become known in Europe because the Chinese dynasties tried to keep it as a secret. Silk cultivation was established for the first time in 552 A.D. in Byzantium under Emperor Justinian. Byzantium remained the largest silk market in Europe until the 11th century. The spreading of silk is also associated with the advance of the Arabs into the western world (e.g., into Spain). It was not until the end of the 16th century, however, that sericulture became established in Europe, especially in Italy and Southern France.

1.3. Life Cycle of the Mulberry Silkworm [17]

As with other moths, the metamorphosis of the silk moth [7] involves the following four stages: egg, larva (silkworm), pupa (chrysalis), moth. After mating, the female lays flat eggs (up to 500) that are as large as a poppy seed. Larvae ca. 3 mm in size hatch from the eggs in spring after wintering for several months. Normally, they live on the fresh, tender leaves of the mulberry tree (*Morus alba*), although nowadays there are also artificial diets available for raising silkworms. In the following four to six weeks, the larva increases its weight many times by consuming ca. 30 g of mulberry leaves. After molting four times, it reaches a length of ca. 8 – 9 cm and begins to spin a protective cocoon. The silk-producing glands are two long tubes of variable thickness. Each gland consists of an anterior, middle, and posterior section. Fibroin is expressed in the posterior section via the usual DNA – RNA protein pathway and is stored in the middle section. Sericin is synthesized in the middle section and accumulates in a layer around the fibroin core. Fibroin and sericin move into the anterior section, which tapers and combines with the second gland to form a single duct leading into the spinneret. On extrusion in air, the aqueous solution of proteins hardens to form a solid silk thread made of two fibroin filaments surrounded by a cementing layer of sericin. The conversion of liquid fibroin to partly crystalline, fibrillar fibroin (crystallite content ca. 60 %) results from the rearrangement and orientation of the folded fibroin molecules caused by extrusion stress and by stretching exerted on the thread by the head of the larva [8].

During the spinning process, the larva moves its head in a figure-of-eight form, producing ca. 60 cm of thread per minute. The finished, spun cocoon is spherical, oval, or constricted in the middle and hangs in the previously spun pupal bed. From the pupa formed after the last molting, a moth emerges after about one week and leaves the cocoon, which has already been softened by the secretion of a proteolytic liquid and made permeable at one spot. The moths, which have a life span of only a few days, mate shortly afterward, and the females lay their eggs, completing the life cycle of the silkworm.

2. Structure and Properties [1–13]

As a protein fiber readily available in pure form, silk has long been a popular subject of research in microscopy, in analysis of the crystal structure of multiphase systems, and in elucidation of the chemical formulas of proteins.

2.1. Hierarchy and Crystal Structure [18]

The raw silk thread (bave) has a slight lengthwise striation and consists of two round to trilobal, on average 10 – 14 μm fibroin filaments (brins) embedded in sericin or silk gum (Fig. 1). Thus, raw silk has a "sheath-around-two- cores" composite structure.

The hierarchy of a fibroin filament, consisting of fibrils, microfibrils, and polymer molecules, is shown in Figure 1 [19].

As with cultivated silk, the tussah larva spins two filaments simultaneously in a sericin matrix. However, the 25 – 30 μm filaments have a much flatter cross section than those of cultivated silk, and the fibroin-to-sericin ratio is much higher. The bound filament pair or bave

is approximately 1.2 – 3.3 denier, and the filaments exhibit lengthwise striations because of the gum channels in the interior [20].

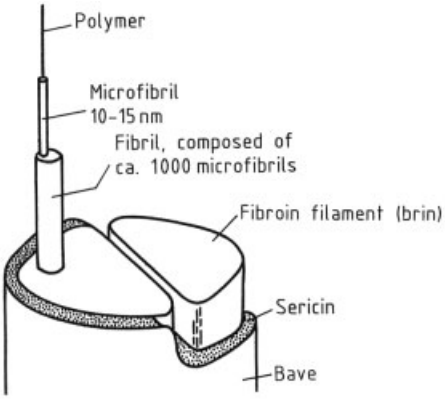

Figure 1. Texture of silk thread [19]
Reprinted from [19] by permission of the author (slightly modified)

Silk is a highly oriented crystalline fiber [21]. MARSH et al. have postulated a structure for the crystalline regions of silk fibroin [22] in which the polypeptide chains have an antiparallel pleated-sheet conformation and are oriented such that in any one sheet, all the glycyl residues project on one side, and the alanyl and seryl residues on the other. The sheets are then packed so that the glycyl surfaces are in contact, leading to an alternation in intersheet spacing between 0.35 and 0.57 nm. This structure with pleated-sheet molecules arranged in an antiparallel manner and polar projections of the side chains has been widely accepted. However, in 1991, TAKAHASHI et al. [23] proposed a new model in which the pleated sheet molecules are packed in an antipolar manner (Fig. 2).

A more recent analysis of the crystal structure of silk was based on 35 newly observed X-ray reflection intensities. The amino acid sequence in the crystalline region of silk fibroin is $-(\text{GAGAGS})-_n$ (G = glycine; A = alanine; S = serine), which is first approximated as $-(\text{GA})-_n$. The fiber period was estimated to be 0.698 nm. The observed reflections can be indexed by a rectangular unit cell with the parameters $a = 0.938$ nm, $b = 0.949$ nm, and c (fiber axis) = 0.698 nm and the space group $P2_1$-C_2^2.

The crystalline region is composed of a rather irregular stacking of sheet structures formed by hydrogen bonds in which the pleated sheet molecules are arranged in an antiparallel manner (antipolar), different from the polar arrangement proposed by MARSH et al. [22].

2.2. Chemical Composition [5,7,8]

Silk fibroin accounts for 75 – 83 wt % of raw silk. The content of sericin, which surrounds the two fibroin filaments, varies from 17 to 25 %, depending on the species, origin, and culture conditions of the raw silk. Fibroin and sericin are proteins. Waxes and fats (ca. 1.5 %), as well as colorants and mineral components (ca. 1 %), occur exclusively in the gum layer. The colorants include xanthophyll, carotene, and flavones. Small amounts of amino sugars have also been detected [24].

Fibroin. Of the 20 amino acids involved in the chemical composition of fibroin from *Bombyx mori* (see Table 1), four of them – glycine, alanine, serine, and tyrosine – make up > 90 mol %. Fibroin contains very little cystine. Alanine is the most common amino acid in tussah fibroin.

Table 1. Amino acid composition of fibroin and sericin (in mol %) [5]

Amino acid	Fibroin	Sericin
Glycine	44.5	14.7
Alanine	29.3	4.3
Serine	12.1	37.3
Tyrosine	5.2	2.6
Valine	2.2	3.6
Aspartic acid and asparagine	1.6	14.8
Glutamic acid and glutamine	1.0	3.4
Threonine	0.9	8.7
Isoleucine	0.7	0.7
Phenylalanine	0.6	0.3
Leucine	0.5	1.4
Arginine	0.5	3.6
Lysine	0.3	2.4
Proline	0.3	0.7
Histidine	0.2	1.2
Methionine	0.1	
Tryptophan	0.2	
Half cystine	0.04	0.1

To determine the molecular mass of silk fibroin (370 000), aqueous fibroin dispersions of the secretion from the posterior glands of the silkworm are analyzed.

According to KIKUCHI et al. [25], the majority of silk fibroin molecules produced by *Bom-*

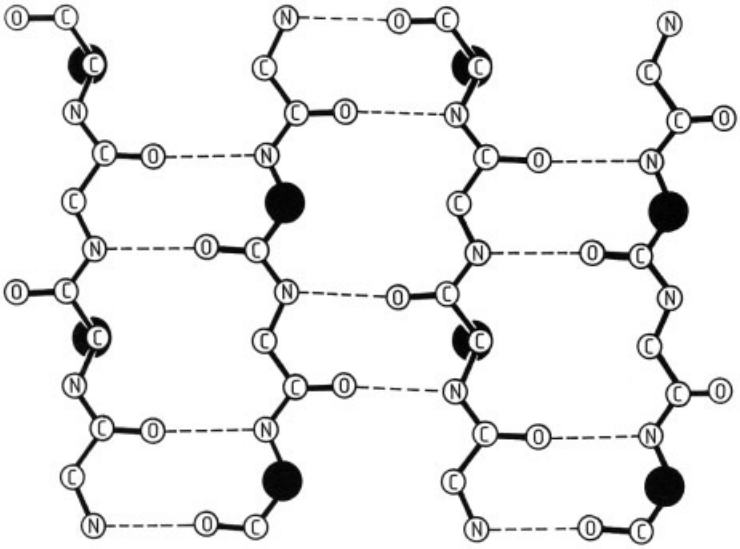

Figure 2. Antipolar – antiparallel (AA) fibroin model for the sheet structure formed by hydrogen bonds [23]
The pleated sheet is seen perpendicular to the sheet. In the antiparallel sheet, the up-and-down molecules alternate, and the methyl groups or hydroxymethyl groups (black circles) of the alanine or serine residues, respectively, alternately point to both sides of the sheet along the hydrogen bond direction (hydrogen atoms not shown).
Reprinted from [23] by permission of J. Wiley & Sons, Inc.

byx mori consist of one heavy (H) chain and one light (L) chain, linked by a single disulfide bond.

Sericin. The secretion with which the larva bonds the two fibroin filaments that come out of the glands is a protein or protein mixture with rubber-elastic properties. It can be dissolved in hot soap solution, alkali solution, or by enzymes. A comparison of the amino acid data of sericin and fibroin (see Table 1) shows that these two components have different compositions; the most frequently occurring amino acid in sericin is serine.

Several authors subdivide sericin in different types based on their solubility in hot water or buffer solutions. Sericin was either isolated from ready spun raw silk filaments [24, 26–28] or from the silk glands of *Bombyx mori* [28, 29]. Dependent on the experimental conditions two [24, 26], three [27], four [28], or five [29] different sericin fractions were obtained. Whether these fractions are really different types of sericin proteins or in part hydrolysis products of sericin, is not definitely known yet [51, 30]. Furthermore it is not known, if all types of sericin isolated from the silk glands also occur in the ready spun raw silk filament [29, 30]. Molecular masses between 1.6×10^4 and ca. 3×10^5 were determined for sericin proteins [28].

2.3. Properties

In the case of fibers, the geometry, the texture, and the mechanical, physical, and chemical properties largely govern mechanical and chemical processing (spinning, weaving, and knitting – preparation, dyeing, printing, and finishing). Much attention has been paid in the literature to the description of silk fiber properties, partly because of the exceptional variety of outward forms. Review articles can be found in [1–13].

The single filaments of degummed silk (i.e., after removal of the sericin, see Section 4.1) have a fineness that varies between 1 and 3.5 dtex depending on its origin. At 3.5 to 4.5 cN/dtex raw silk exhibits, under standard conditions (20 °C, 65 % R.H.) a tensile strength (based on fineness) that almost reaches the values of synthetic fibers. The wet tenacity is 75 – 95 % of the tenacity under standard conditions. The elongation under standard conditions is 18 – 25 %, and in the wet state 15 – 60 %. The corresponding values for degummed silk are a little lower and they de-

pend on the degumming method and the testing conditions (see Sections 2.4 and 4.1) [32]. In addition to its fine and characteristic handle (feel), degummed mulberry silk has good elastic properties.

Silk is hygroscopic. Under normal atmospheric conditions (20 °C, 60 % R.H.), degummed silk absorbs about 10 % water based on the weight of the water-free fiber. The silk trade calculates with 11 % water (regain).

At 1.6 % (longitudinal direction) and 18.7 % (crosswise), fiber swelling is strongly anisotropic. Silk has a relatively high electrical resistance. At low air humidity, it becomes electrostatically charged when subjected to friction.

Silk is virtually unaffected by temperatures up to 140 °C; above 150 °C, thermal decomposition occurs.

Along with wool and the synthetic textile fibers polyester and polyamide, silk is a poorly inflammable fibrous material [33].

As a protein, silk fibroin is a polyampholyte with cationic and anionic side chains and, therefore, absorbs strong alkalis and mineral acids with the formation of salts. The isoionic point is at pH 5.0. Moderately strong and weak acids undergo primarily molecular bonding to silk via hydrogen bonds.

In the isoionic range (pH 4 – 5), swelling and reactivity are at their lowest.

The degree of decomposition of silk after chemical textile operations is measured mechanically (tenacity and elongation) and chemically (viscometric determination of the molecular mass of silk). The weight loss in blank dyeing is determined indirectly from the amount of silk fibroin that has dissolved (nitrogen analysis).

In comparison with wool, silk is more sensitive to acid because of the virtual absence of acid-stable cystine bridges. However, the absence of alkali-sensitive cystine bridges gives silk a higher resistance to alkali. For this reason, silk can be treated with 16 – 18 % sodium hydroxide solution (mercerizing bath) at low temperature to produce crepe effects in mixed fabrics containing cotton yarn in the warp and the weft.

Silk can absorb considerable amounts of salts, a property that is exploited in weighting (see Section 4.2). Salts can degrade silk under certain conditions, especially in the presence of light, air, moisture, and perspiration. *Bombyx mori* fibroin is soluble in certain concentrated aqueous salt solutions, e.g., lithium bromide, lithium iodide, lithium thiocyanate, and copper ethylenediamine solutions. The best of these solvents is probably concentrated (49 %) aqueous lithium thiocyanate, because the solution is approximately pH neutral, and the solvent is unlikely to cause peptide bond hydrolysis [34].

Sunlight degrades silk rapidly. Oxidizing agents degrade silk proteins, causing discoloration, the formation of acidic groups, and deterioration of the mechanical properties.

Although chemically similar to cultivated silk, tussah is more resistant to corrosive chemicals (see Table 2). This increased resistance, particularly to alkali, can be an advantage in dyeing. It permits the application of reactive dyes to tussah by means of an alkaline fixation procedure [20].

Table 2. Dissolution time of silk in chemicals [20]

Treatment	Cultivated silk	Tussah
Boiling 1 % NaOH	rapid	30 – 60 min
Cold concentrated HCl (*d* 1.16)	rapid	very slow
$ZnCl_2$ solution (*d* 1.67)	rapid	slow

2.4. Analysis

The sericin content in incompletely degummed silk is determined gravimetrically after high-temperature alkaline extraction of sericin. It can also be determined by amino acid analysis or from the sum of the primary amino groups [35].

When silk is heated in acids or alkalies, losses in tenacity occur. The greater the losses in tenacity are, the lower is the molecular mass of the silk fibroin. This can be determined by measuring the viscosity of a solution of silk in lithium bromide solution [36].

Blends of polyester with silk or wool are commonly analyzed by using a boiling solution of 5 % sodium or potassium hydroxide to dissolve the protein fiber. A faster room-temperature technique using a trichloroacetic acid – dichloromethane reagent to eliminate the polyester in 5 min has been proposed as a reliable method for determining the protein fiber content [37].

The comfort and handle of silk fabrics are tested by the methods of Kawabata [38, 39].

Silk that has been graft polymerized with methacrylamide (see Section 4.2) can be identified by solubility tests or IR spectroscopy [40].

3. Production

3.1. Sericulture [6,11]

The silk industry is closely tied to cultivation of the mulberry tree. Although it is also possible to rear silkworms on an industrial scale by using completely or partly artificial diets [11,41], the cultivated silkworm, *Bombyx mori*, usually is fed by mulberry leaves. The life cycle of the silk moth is linked to the cycle of the mulberry tree. Since young silk-worms require young leaves to eat, the hatching of the eggs is timed to coincide with the appearance of new leaves.

The lives of the silk farmer and his family and employees are dominated by the silkworm. From the time of hatching until the spinning of the cocoon, the larva eats prodigiously. Since the larvae from 30 g of eggs (40 000 – 60 000) consume about one tonne of leaves during their lifetime of 35 d, much time is spent in picking and preparing leaves. Young silkworms require chopped leaves, which adds to the work. In addition, the trays on which silkworms are raised must be cleaned daily. After they have been laid, the eggs must pass through a period of low temperature before they will hatch. This tendency is a holdover from the time when the silk moth was a wild insect that laid its eggs in the fall for hatching in the following spring. Cultured silk eggs are exposed to artificial low temperature for a period of time.

The reproduction of eggs is distinctly separate from the farmer's activity of rearing cocoons. These two phases are kept separate because of the many precautions that must be taken not only to prevent the introduction of diseased eggs and keep the silk produced of high quality, but mainly to avoid degeneration by inbreeding.

The silk farmer receives his eggs on cards. About a week before the mulberry trees bring forth their first leaves, the cards of eggs are placed in an incubator at 27 °C. The eggs hatch in about ten days.

A large number of breedings of *Bombyx mor*i exist. They differ in thickness and length of the thread spun by the silkworms, in their demands and adaptation towards climate and diet.

Tussah silkworm growing is quite different from the cultivation of mulberry silkworms. Whereas mulberry silkworms are raised indoors, tussahs grow in toothed-oak forests. After baby tussah worms have hatched in early spring, the cultivators bring them to oak forests in hilly areas and place them around tree trunks. The wild silkworms then begin to climb the trees and feed on the leaves. To prevent the silkworms from being eaten or harmed, their cultivators fire a shotgun into the air at intervals to scare birds away. The cocoons are harvested from the trees in autumn [42].

3.2. Production of Silk Thread

To permit the storage of cocoons over a longer period of time, the pupae contained in the cocoons are first killed with hot air or superheated steam.

In the sorting department, the well-preserved cocoons are separated from the faulty ones (damaged, pierced). After the loose floss silk has been removed, the cocoons are sorted according to size, color, and evenness. Cocoons of the same breed season, class, and size are mixed to obtain an even batch.

In the filature (silk-reeling factory), 40 – 60 cocoons are placed in a small basin filled with boiling water. The cocoons are lightly beaten by means of an alternately rotating besom and kept below the surface of the water. The silk gum softens, and the outer layers of the cocoon that cannot be reeled (floss silk) is removed. The cocoons prepared in this manner, which have exposed, reelable thread ends, are now placed in the reeling basin (water temperature ca. 40 °C). The threads from several cocoons (the number depends on the titer desired) are drawn off together and wound into a hank (grège or reeled silk). The hanks are then dried.

Depending on the origin and species, the total length of thread in a mulberry silk cocoon is 1000 – 3000 m, only 500 – 1500 m of which can be reeled. To obtain 1 kg of grège, 5 – 10 kg of cocoons are required, depending on the technological standard in the particular country.

Together with the waste from silkworm raising – watt silk, defective cocoons, pierced co-

coons – the outer and inner layers of the cocoon that cannot be reeled off are processed into high-quality schappe and bourette silk. Oils and fish feed are produced from the pupae.

Unlike cultivated silk, tussah is very difficult to reel off without preliminary bleaching of the cocoons with a weak solution of hydrogen peroxide and sodium silicate [20].

3.3. World Production

In 1998, the worldwide production of raw silk amounted to ca. 70 727 t. The most important silk-producing countries are China, India, Japan, Brazil, and Uzbekistan (see Table 3).

World production of wild silk is estimated to be 10 – 15 % of the production of mulberry silk. The largest exporter of raw silk is China. In China, the income of about 100×10^6 people is based on silk production or finishing [43].

Silk has less than 0.2 % of the total world fiber market. At present, silk is produced in about 50 countries and the number is continuously increasing. New silk producing countries are for instance Ivory coast, Uganda, and Colombia [44].

The data in Table 3 show that silk production rapidly increased from the beginning of the 1990s until 1995, and then a strong decrease occurred. This was due to the diminished demands of the consumers for silk and to the overabundance of silk products on the world markets. The image of silk has been downgraded for several years as a result of the marketing of very cheap made-up goods from China and India, i. e., since the launching of poor quality ″sand-washed silk″ [45]. The official Chinese price for raw silk (grade 3A, 20/22 denier) is today (1998) about $ 20 per kilogram (ten years ago $ 51).

3.4. Silk Yarns – Classification and Testing

Fine silk yarns for weaving are divided into two groups [46]:

1) Filament yarns made from raw grège: grège, tram, organzine, crepe, and grenadine
2) Spun yarns such as schappe silk made from the waste from cocoon processing and bourette silk made from the waste from schappe spinning

The fineness of grège threads is expressed in denier (grams per 9000 m) or dtex (grams per 10 000 m) and that of spinning fiber yarns in Nm (metric number = length in meters of 1 g).

According to the International Silk Association (Lyon), the standard classification for grège threads involves the following 12 quality classes: 6A, 5A, 4A, 3A, 2A, A, B, C, D, E, F, and G. Class 6A represents the best quality, which is rarely attained; G is the lowest grade. The quality normally found on the market is 20/22 den, 3A and 4A.

Silk threads are tested objectively and subjectively.

Objective testing includes tests for titer, size, tenacity, elongation, cohesion, and winding breaks, as well as determination of moisture and sericin content [11,47,48].

Subjective tests include the seriplane test, in which grège threads are wound on black board (seriplane) and the evenness, cleanness, and neatness are evaluated visually. However, this is being increasingly performed by electronic testing [11,48].

Many silk producing countries have their own systems for classification of raw silk quality. The Chinese classification system is the most important one, because the lion's share of silk is produced in China. The countries of CIS and Brazil use quite similar systems. Classification is mostly carried out by subjective-visual assessment and therefore the reproducibility is not good. The Chinese classification system for grège threads involves the following 12 quality classes: 6A, 5A, 4A, ..., 2A, A, B, ..., F and substandard. Class 6A represents the best quality; substandard is the lowest grade.

The International Silk Association (ISA, respectively AIS - Association Internationale de la Soie) competes for the introduction of a new international raw silk classification system, based on the tex-system and reducing the 12 (Chinese) grades to 8 grades. Classification can be carried out either by the traditional "Seriplane test" (subjective, visual assessment of evenness, cleanness, neatness; with the grades S6, S5, ..., S0, Ssub) or by electronic testing methods (automated testing of titer, size, evenness, thick and thin defects, impurities, tenacity, elongation, co-

Table 3. Estimated world production, imports, and exports of raw mulberry silk (in tonnes) (Source: ISA member-countries)

Production	1992	1993	1994	1995	1996	1997	1998
China	54 480	69 300	72 000	77 900	59 000	55 117	49 430
India	12 600	13 200	13 200	12 884	12 384	12 954	13 400
Japan	5100	4200	3900	3240	2579	1902	1108
Brazil	2280	2340	2520	2468	2270	2120	1821
Vietnam	N.A.	N.A.	N.A.	2100	1500	834	862
Uzbekistan	2160	1800	1800	2320	2500	2000	1500
Thailand	1589	1229	1377	1313	1144	1039	900
N. Korea	1200	1200	1200	600	360	200	150
Iran	423	427	396	750	600	500	400
S. Korea	910	840	491	346	146	72	60
Turkey	140	120	80	65	30	27	20
Others *	1537	1303	1424	1217	1165	1666	1076
Total	82 419	96 006	100 388	104 203	83 678	78 431	70 727
Main Importers of Raw Silk							
European Union	3337	3935	5313	4767	4727	5370	4012
of which Italy	2711	2824	3813	3252	2927	3191	2477
Japan	1559	1483	1543	1992	2076	2100	1740
India	2076	2768	4892	5403	4149	2911	2760
Korea (Republic)	2433	2914	2908	2763	3233	2470	1228
Main Exporters of Raw Silk							
China	10 825	6203	13 049	9674	11 200	10 120	8481
Brazil	2026	1920	2068	2320	2174	1892	1678

N.A.: Data not available
* Bangladesh, Bulgaria, Cambodia, Colombia, Egypt, Indonesia, Malaysia, Paraguay, Philippines, Romania, Sri Lanka, CIS

hesion and winding properties; with the grades E6, E5, ..., E0, Esub) [48].

4. Processing

4.1. Degumming [7,35,49–51]

The term degumming (boiling-off) refers to partial or complete removal of the sericin that covers the two fibroin filaments. The sericin sheath is responsible for the brittleness, harsh handle, mat appearance, and yellowish to yellowish green color of raw silk. Only after degumming does silk attain its full gloss and its soft, smooth feel, which can be improved further in a finishing process following dyeing. Degumming results in a weight loss that can amount to 2 – 30 %, depending on the desired degree of degumming.

Most fabrics are degummed in the producing countries. The degumming agent used is still mainly Marseilles soap, an alkali-free olive oil soap. In this classical degumming method, silk is treated in a soap bath at 90 – 98 °C for 2 – 4 h, depending on the quality and type of fabric. It is then boiled off again and subjected to weakly ammoniacal rinsing. In the case of tussah silk, alkaline boiling off with soap and soda removes the gum adequately because of the low sericin content. Other degumming methods use enzyme solutions (e.g., papain, trypsin, or various bacterial enzymes), boiling water under pressure, weakly alkaline solution ($NaHCO_3$, ammonia), or acidic solution [52] (e.g., tartaric acid or citric acid). The degumming process can be accelerated by treatment with ultrasound [53,54].

New degumming methods have been developed to save time, water, and energy in the production process. The latest powder products consist mainly of sodium salts, such as phosphates and carbonates, sometimes with a soap additive. In addition to sodium salts, liquid scouring agents also contain anionic surfactants and other organic substances such as sequestering and fiber-protective agents. In contrast to soap, these newer products can be used in continuous processes [55,56].

The degree and evenness of degumming of silk can be evaluated qualitatively by means of a dyeing test with C.I. Direct Red 80 [35]. In contrast to fibroin, the sericin present is dyed red. Even degumming must be achieved because

otherwise levelness problems could arise during dyeing.

Raw silk is degummed either as yarn in the form of hanks in an open beck, in a foam degumming machine or on perforated tubes, or as fabric in a star machine, winch beck, jigger, or continuous machines (e.g., open-width washing machine). In principle, any dyeing machine can be used. The time required to degum silk fabrics depends on the weight of the fabric. New machines for degumming, washing, and dyeing of all spool yarns and star-hung fabrics which operate at ambient or high temperatures are now commercially available.

4.2. Weighting [7,49,57]

Weighting of silk refers to the incorporation of metal salts, plant materials, or synthetic substances into fibroin to compensate for the loss in weight during degumming. Moreover, increased fiber volume and improved luster and handle are achieved in this process. Hanks are usually weighted before dyeing, in vats, on sticks, or in special centrifuges, hank-washing machines, or continuous open-width washing machines.

The weighting agent is fixed so that it is washfast. Weighting with water-soluble compounds such as dextrin increases the weight without producing a permanent improvement in fiber properties.

Today, tin phosphate silicate weighting (mineral weighting), a method known for decades, plays only a minor role. Silk is treated in succession with aqueous solutions of tin(IV)-chloride and disodium hydrogenphosphate. Basic tin phosphate formed in the fiber is subsequently converted to washfast tin sodium silicate in a concentrated silicate bath. The degree of weighting depends on the number of treatments with tin salt and phosphate; increases in weight of more than 100 % are possible. The degree of weighting is based on the weight of raw silk. If weighting compensates exactly for the loss in weight incurred during degumming, the silk is referred to as par. For example, 50 % above par means that the weight of 100 kg of raw silk has increased by 50 kg.

Mineral/plant or pure plant weighting with tanning agents or logwood extract is used principally for material that is to be dyed black. This type of weighting is often combined with black dyeing.

More recent methods use synthetic resins (e.g., methacrylamide) for weighting silk yarns and fabrics. The material is generally dyed first and then weighted by graft polymerization with methacrylamide. This procedure gives improved fastness but can be used only with dyes that are insensitive to graft polymerization conditions [57]. When silk is grafted with methacrylamide, increases in weight of 20 to 35 % are usual.

In general, the degree of weighting depends on the intended use of the silk. However, very high mineral weightings reduce the quality, and the silk becomes brittle and damaged; exposure to sunlight and moisture result in premature wear. Furthermore, the affinity for dyes and the washfastness of the dyed material are reduced. For this reason, silk is weighted today only for certain articles (e.g., ties, borders, embroidery) to improve luster, handle, drape, and fullness.

In comparison with cultivated silk, much less weight is lost when tussah is degummed. Therefore, weighting is of less importance in processing tussah [20].

4.3. Bleaching [7,49]

In most cases, the colored pigments present in sericin are removed during degumming. To achieve clear white tones and brilliant light shades, oxidative or reductive bleaching can be used. Bleaching is usually carried out after degumming.

Reductive bleaching agents that can be used for silk are sulfur dioxide, dithionite, or hydrogen sulfite. Sulfur dioxide is preferred to other reductive bleaching agents because of easy handling, lower cost, and fiber preservation.

However, oxidative bleaching with hydrogen peroxide or H_2O_2-releasing salts is of much greater importance. The resulting whiteness is, in contrast to sulfur dioxide bleaching, persistent and the material is free from odor. The concentration and temperature of the bleaching liquor depend on the amount of colored pigment remaining on the fiber.

Optical brighteners are added to the bleach to achieve full whiteness. This step is usually performed in a reductive bleaching liquor after oxidative prebleaching.

For silk piece goods, degumming and bleaching are often carried out in the same bath. Since bleaching damages the fibers less than mineral weighting, bleaching is usually performed after weighting. Because of its intense inherent color, which cannot be removed even by complete degumming, tussah silk requires considerably stronger bleaching conditions than mulberry silk. Depending on the quality of tussah, full whiteness can be achieved only with great difficulty, and fiber damage is unavoidable [58].

4.4. Dyeing [7,59]

As a protein fiber, silk can be dyed with almost all types of dyes used for wool dyeing; acid, basic, metal-complex, reactive, and vat dyes, as well as pigments (see → Textile Dyeing). Also older natural dyes such as logwood and tanning agents are still used. The fastness required often determines the choice of dye. Acid dyes are the most frequently used dyes for both hank silk and piece goods because they give good to very good lightfastness, and an extensive range of colors is available. Although basic dyes give glossy, clear colors, the lightfastness is lower. However, the levelness can be improved by adding acetic acid, and the wetfastness can be improved by aftertreatment with tannic acid solution. Preformed chromium complex dyes with good leveling ability produce high wetfastness and lightfastness. The 1:2 metal complex dyes are also used in silk dyeing.

Although vat dyes produce washfast colors, they cause considerable fiber damage when exposed to strong light (e.g., decorative materials and drapes). Weighted silk cannot be dyed with vat dyes because the weighting of the thread would be removed by the alkaline dye liquor. Reactive dyes give good results, especially when used under alkaline conditions. In many cases, dyeing with reactive dyes leads to additional fixation of the dye to the fiber through covalent bonding with the numerous functional groups present.

For blackening, logwood black or tanning agents are still being used because they give a special depth of color and filling effect. The 1:2 metal-complex, afterchrome, and Naphtol AS dyes can also be used to produce deep black tones.

In comparison with wool, silk exhibits a markedly lower affinity for wool dyes, especially for reactive dyes. This is due to the high orientation of fibroin, impeding the insertion of dyes into the filaments [60] and the lower acid-binding capacity of silk compared to wool [49, 61]. Due to its fine titer, about twice the amount or, in the case of tussah, almost three times the amount of dye is required to achieve the same depth of color as obtained with wool [13,62, 63]). As a rule, degummed silk is dyed, even degumming being a requirement for level dyeing. The degumming liquor can be used in silk dyeing as an excellent protective colloid and leveling agent.

The dyeing temperature should not exceed 90 °C because the luster and handle of silk can be negatively influenced on boiling.

The dyeing processes used are the exhaust, cold pad-batch, and the pad-steam methods. The exhaust dyeing process is the most important because it can be used for dyeing piece goods and all types of yarn.

In principle, all dyeing machines employed for other textile fibers can be used for silk. The machines must be modified because of the sensitivity of silk to tensile force and mechanical stress. Silk is dyed in hank (yarn) form on perforated tubes or in piece form. The dyeing aggregates used for yarn dyeing are the hank-dyeing machine, package dyeing apparatus, and spraying arm dyeing machine, and those used for piece dyeing are the star frame, winch vat jig, beam, and jet dyeing machine.

4.5. Printing [7,64]

Textile printing is particularly important in silk finishing. Based on total silk production, the percentage of printed articles is higher than for all other natural and man-made fibers.

Articles produced by silk printing are predominantly dress materials, scarves, shawls, ties, and increasingly, drapery and upholstery fabrics.

Like dyeing, printing also requires the even degumming of silk. If necessary, unevenly degummed fabric must be improved in a rewashing step. For yarn-dyed articles, a prewash is helpful to prevent bleeding during the rewashing process.

Silk fabric made from nonweighted silk is generally used for printing. In the printing of mineral-weighted silk, the weighting should not exceed 15 % above par because otherwise the material becomes brittle on steaming, and color changes occur during discharge printing due to the formation of tin sulfide. Both direct printing and discharge printing are suitable. As with other fibers, special effects can be achieved by the application of pigment, matte white, lacquer, and bronze printing to silk, sometimes in combination with direct and discharge printing.

Generally, silk prints are produced with the same types of dyes and printing methods as wool prints (see → Textile Printing). For reasons of fastness, acid, metal- complex, and reactive dyes are preferred; in addition to these, vat dyes can also be used for discharge printing.

Silk printing is carried out mechanically as hand screen printing or machine printing (automatic flat screen and rotary screen printing machines).

4.6. Special Finishing [7,65]

Objectives of special finishing processes are to improve some rather unfavorable properties of silk (creasing properties, launderability, UV and light resistance), if possible, without damaging the characteristic luster and handle of silk.

Because of the high price of silk and low percentage of worldwide annual fiber production, relatively little attention has been paid to the development of special finishing processes. For instance, both wet tenacity and water-drop fastness can be improved by treatment with hexamethylenetetramine (as a source of formaldehyde for cross-linking). Fabrics finished with polyurethanes exhibit good wet crease recovery and are dimensionally stable on washing. Cross-links resulting from the reaction with dihydroxymethyl urea or thiourea confer a higher crease resistance on silk but do not improve its tenacity. Silk that has been treated with hydroxymethylacrylamide and tin chloride shows increased wet and dry crease recovery. Finishing with mono- and multifunctional epoxides can improve moisture uptake and pleat retention. An effective flameproof finish results from reaction with phosphoric acid derivatives and quaternary ammonium salts.

At the moment, sand-washed silk fabrics are highly fashionable. The sand-washing process is a special washing procedure for silk fabrics to produce a certain degree of fibrillation with an acceptable degree of loss of tenacity. Sand-washed fabrics have a special appearance and handle and a higher surface roughness than "normal" silk. In 1997, 70 to 80 % of the silk fabrics produced in China were finished with sand-washing processes [66].

Other finishing processes serve for scrooping, softening, and building up a light, firm handle [49].

Schappe articles are treated in order to improve the slippage behavior. Tussah articles are made water repellent by special finishes.

5. Uses and Other Aspects [13,15]

The uses of silk cover a wide range of applications from ready-to-wear articles and home textiles to technical fabrics and cosmetic articles. Ready-to-wear articles include mainly mens' and womens' wear such as shirts, blouses, scarves, shawls, ties, and high-quality evening clothes, as well as lining materials, underwear, night clothes, national dress (kimonos and sarongs). In home textiles, silk is most often used for furnishing fabrics, upholstery, wallpaper, velvet, plush, carpets, and covers. In technical fabrics, silk is employed for typewriter ribbon, insulating material for cable covering, and surgical articles. Silk powder is added to lipstick, skin cream, and soap; fibroin or sericin is added to shampoos and lotions.

Silk is also a very suitable material for sewing and embroidery thread because it can be dyed to give any shade.

According to the German textile labeling law (Textilkennzeichnungsgesetz, TKG), the name silk is permitted only for fibers that are isolated from the cocoons of silk-spinning insects (mulberry silk and wild silk). The trademark for fabric made of 100 % fine or wild silk is awarded by the International Silk Organization in Lyon (see Fig. 3). In this context it has to be mentioned that according to the TKG and the corresponding European guideline no special labeling is necessary for weighted silk. The fabrics may be traded as "pure silk".

Figure 3. International silk symbol

The silk atlas describes 150 different silk fabrics. Of these, nine are mentioned here with their main applications:

Crêpe Georgette (for evening dresses, blouses), crêpe de Chine façonné (for ladies' dresses, blouses), satin crepe (for blouses, evening dresses, negligees), faille (for dresses, coats), taffeta (for robes, frills, cocktail dresses, blouses, lining), honan silk (for sports clothes, two-piece outfits), organdy (also organza; for bridal dresses, blouses, ladies' hats), toile de soie (for blouses, shirts), and doupion fabrics (blouses, dresses) [15].

According to Mecheels [67], the thermophysiological wearing comfort of silk textiles is very high because silk material is very agreeable to the skin, and the water balance seems to be optimal.

Clothes made of silk function excellently under everyday conditions, even when they come in very close contact with the skin. The same applies to their appearance when worn and the drape of silk clothes. However, some drawbacks must be accepted. Silk materials are highly sensitive to abrasion owing to the fibrillation of the fiber surface, which is exploited in sand-washed schappe fabrics. The wearing time of silk clothes is limited by abrasion and perspiration damage. The low resistance of silk to light and the insufficient lightfastness of some colorants can lead to the yellowing of natural- color areas and color changes in silk outer clothing.

As a rule, the drycleaning of silk clothes poses no problems if specific conditions are maintained. Depending on the type of fabric and degree of fastness of the color or design, silk articles can also be washed. However, the manufacturer's instructions for care should be followed.

No health risks are involved in wearing silk clothes, even if they come in direct contact with the skin. Allergies are observed very rarely and are caused not by the silk itself, but by substances added during finishing (see Chap. 4). On June 6, 1991, a seminar was held in Hof (Germany) on the medical compatibility of chemically finished textiles and the following statement was made [68]:

"Properly finished and treated textiles represent no health risk for humans.

Long-term studies at the Dermatological University Clinic in Erlangen have clearly shown that real allergies to finished textiles for clothing are extremely rare, even when formaldehyde and individual disperse dyes are included.

The possible allergenic potential for sensitive persons can be eliminated by following the recommendation to wash textiles before use.

Properly finished and dyed textiles exert no carcinogenic effect."

More recent studies also draw the conclusion that allergies and other health risks caused by wearing textiles are extremely low [69, 70].

Today, silk garments are often priced in the same range as garments made of cotton or synthetic fibers. As already mentioned in Section 3.3, the image of silk has suffered a lot in the last ten years due to the overabundance of low-price, low-quality made-up goods on the markets. As a result, silk is not unreservedly associated with luxury and exclusiveness any more. Nevertheless, the upgrowing standard of living in Southeast Asia and increasing demands for comfortable clothes made of natural, ecologically friendly processed fibers permit a confident prognosis for the consumption of silk in the future. In order to refine the image of silk, efforts are made to assure certain standards of quality during all stages of silk processing.

Acknowledgments. The authors would like to thank Ursula Behn (Deutsches Wollforschungsinstitut) for revising the manuscript and K.-H. Phan (Deutsches Wollforschungsin-

stitut), ADOLF FAES (President of the International Silk Association, Zürich), B. TRUDEL (Trudel Limited, Zürich), R. CURRIE (General Secretary of the International Silk Association, Lyon), R. CREMER (Verband der Deutschen Seiden- und Samtindustrie, Krefeld), H. FREI (Siber Hegner Textil AG, Zürich), L. ZOSSO (DESCO von Schulthess AG, Zürich), Mr. SCHWARZENBACH (Schubiger & Schwarzenbach, Uznach/Switzerland), E. IIZUKA (Shinshu University, Ueda, Nagano/Japan), S. MIZUNO (Tohoku University, Sendai/Japan) and Y. Takahashi (Osaka University, Osaka/Japan) for valuable contributions and information.

6. References

General References

1. F. Bock, L. Pigorini: *Die Seidenspinner – Ihre Zoologie, Biologie und Zucht,* Springer-Verlag, Berlin 1938.
2. E. R. Kaswell: *Textile Fibers, Yarns and Fabrics,* Reinhold Publ., New York 1953.
3. M. Harris (ed.): *Handbook of Textile Fibers,* Harris Research Laboratories, Washington 1954.
4. P. A. Koch, G. Satlow (eds.): *Großes Textil-Lexikon,* Deutsche Verlags-Anstalt, Stuttgart 1965/1966.
5. F. Lucas, K. M. Rudall in M. Florkin, E. H. Stotz (eds.): *Comprehensive Biochemistry,* **vol. 26 B,** Elsevier, Amsterdam 1968, p. 475.
6. W. G. Wolfgang: "Silk" in *Encyclopedia of Polymer Science and Technology,* vol. 12, Interscience, New York 1970, p. 578.
7. *Ullmann,* 4th ed., **21,** 201.
8. R. M. Robson: "Silk. Composition, Structure and Properties," in M. Lewin, E. M. Pearce (eds.): *Handbook of Fiber Science and Technology,* vol. 4, "Fiber Chemistry," Marcel Dekker, New York 1985, 647 – 700.
9. *Seide – Zur Geschichte eines edlen Gewebes,* 3rd ed. Verlag Neue Zürcher Zeitung, 1986.
10. C. D. Livengood: "Silk," *Encycl. Polym. Sci. Eng.* **15** (1989) 309.
11. L. Rheinberg, *Text. Prog.* **21** (1991) 1 – 43.
12. P. R. Brady, *Rev. Prog. Color. Relat. Top.* **22** (1992) 58 – 78.
13. H. Zahn, B. Wulfhorst, M. Steffens "Faserstoff-Tabellen nach P.-A. Koch: Seide," *Chemiefasern/Textilind.* **44** (1994) 40 – 59.

Specific References

14. ISO 6938-1984: Textiles – natural fibers – generic names and definitions. DIN 60 001, part 1, 10.90: Textile Faserstoffe – Naturfasern, Deutsches Institut für Normung, Berlin 1990.
15. H. G. Wiedemann: *Chinesische Seide. Geschichte eines Naturproduktes,* Verlag Dr. Wiedemann, Greifensee, Switzerland, 1984.
16. China: Sericulture, FAO Agricultural Service Bulletin 42, Rome 1980, 7 – 8.
17. J. Magoshi, Y. Magoshi, M. A. Becker, S. Nakamura: "Biospinning" in J. C. Salamone (ed.): *Polymeric Materials Encyclopedia,* CRC Press, 667 – 679.
18. Y. Shen, M. A. Johnson, D. C. Martin, *Macromolecules* **31** (1998) 8857 – 8864.
19. E. Iizuka, *Proc. 7th Int. Wool Text. Res. Conf.,* **vol. 1,** Tokyo 1985, 319 – 328.
20. A. G. Liddiard, *J. Soc. Dyers Colour.* **98** (1982) 394 –397.
21. R. D. B. Fraser, T. P. MacRae, F. H. C. Stewart, E. Suzuki, *J. Mol. Biol.* **11** (1965) 706 – 712.
22. R. E. Marsh, R. B. Corey, L. Pauling, *Biochim. Biophys. Acta* **16** (1955) 1 – 34.
23. Y. Takahashi, M. Gehoh, K. Yuzuriha, *J. Polym. Sci. Polym. Phys. Ed.* **29** (1991) 889 – 891.
24. H. Sinohara, *Comp. Biochem. Physiol.* **63B** (1979) 87 – 91.
25. Y. Kikuchi et al., *Gene* **110** (1992) 151 – 158.
26. E. M. Sheldon, T. B. Johnson, *J. Amer. Chem. Soc.* **47** (1925) 412 – 418.
27. H. H. Mosher, *Am. Dyestuff Rep.* **21** (1932) 341 – 345.
28. K. Komatsu, *Proc. 7th Internat. Wool Text. Conf.,* Vol. 1, Tokyo (1985), 373 – 382.
29. T. Gamo, T. Inokuchi, H. Laufer, *Insect Biochem.* **7** (1977) 285 – 295.
30. R. Voegeli, J. Meier, R. Blust, R. Hofsteter, *Cosmet. Toiletries* **108** (1993) 101 – 108.
31. S. Mizuno: "Symposium on Silk Protein Structure and Silk Gene Expression," *19th International Congress of Entomology,* Beijing, China, June 30 – July 2, 1992.
32. A. Krasowski: "Optimierung von Prozessen in der Seidenveredlung", Thesis RWTH Aachen (1998).
33. D. Wolber, K. Bläsi, *Textilveredlung* **25** (1990) 269 –275.
34. F. Lucas, J. T. B. Shaw, S. G. Smith: "The Silk Fibroins", *Advances in Protein Chemistry* **13** (1958) 107 – 242.
35. J. Knott, G. Freddi, M. Belly, *Melliand Textilber.* **64** (1983) 481 – 483.

36. Schweizer Normvorschrift SNV 19 5595, 1969.
37. S. Sivakumaran, S. Rajendran, *Text. Chem. Color.* **20** (1988) 25 – 26.
38. M. Matsudaira, S. Kawabata, *J. Text. Inst.* **79** (1988) 458 – 503.
39. M. Jarrigeon, F. Payot, S. Rotival, M. Sotton, *L'Industrie Textile* 1990, no. 1206, 41 – 44.
40. W. Schindler, U. Wiesel, *Melliand Textilber.* **80** (1999) 546 – 548.
41. H. Ito, Y. Muraoka, T. Yamazaki, T. Imamura, H. Mori, M. Ichida, M. Simuda, F. Matsubara, *Text. Res. J.* **65** (1995) 755 – 759.
42. M. Xianmin, *Wool Record* **151** (1992) no. 3570, 41.
43. P. Giger, *Wool Record* **156** (1997) no. 3636, 17.
44. R. Currie, *Notiziario Tecnico Tessile* **47** (1997) no. 1, 9 – 14.
45. Anonym, *Textile Asia* **30** (1999) no. 1, 47.
46. DIN 60 550, 10.84: Webgarne aus Seide. Deutsches Institut für Normung e.V., Berlin 1984.
47. R. Freitag, *Melliand Textilber.* **66** (1985) 566 – 569.
48. Study Group of the Silk Standards Committee (eds.): *Raw Silk 1995,* International Silk Association, Congress Brighton, May 1995.
49. H. Flensberg, I. Hammers, *Text. Prax. Int.* **43** (1988) 739, 742 – 747.
50. A. Svilokos Bianchi, G. M. Colonna, *Melliand Textilber.* **73** (1992) 68 – 75.
51. M. L. Gulrajani, *Rev. Prog. Color. Relat. Top.* **22** (1992) 79 – 89.
52. G. Freddi, G. Allara, G. Candiani, *JSDC* **112** (1996) 191 – 195.
53. A. Krasowski, B. Müller, J. Föhles, H. Höcker, *Melliand Textilber.* **80** (1999) 543 – 545.
54. J. Jeyakodi-Moses, *The Indian Textile Journal* **107** (1997) no. 11, 26 – 28.
55. K. Mahall, I. Goebel, *Textilveredlung* **23** (1988) no. 1, 8 – 16.
56. A. S. Bianchi, G. M. Colonna, *Melliand Textilber.* **73** (1992) 68 – 75.
57. W. Schindler, C. Gluth, A.-R. Czerny, *Melliand Textilber.* **73** (1992) 65 – 67.
58. S. Das, *J. Soc. Dyers Colour* **108** (1992) 481 – 486.
59. R. Rohrer, *Textilveredlung* **20** (1985) no. 3, 85 – 87.
60. "Seide und Farbe", Broschüre no. 9307.00.88, Sandoz AG, Basel 1988.
61. H. Rath, in *Lehrbuch der Textilchemie,* 3. Aufl., Springer-Verlag, Berlin 1972, 506 – 508.
62. R. Hofstetter, *Melliand Textilber.* **72** (1991) 366 – 373.
63. A. Kremer, H. Kremer, *Melliand Textilber.* **73** (1992) 580 – 582.
64. A.-R. Czerny, *Melliand Textilber.* **69** (1988) 438 –439.
65. B. Tsang, R. C. Dhingra, *Textile Asia* **27** (1996) no. 5, 56 – 63.
66. V. V. Hadimani, V. G. Halliyal, P. P. Raichurkar, R. Bhandiwad, *The Indian Textile Journal* **108** (1997) no. 4, 96 – 103.
67. J. Mecheels, *TPI Text. Prax. Int.* **43** (1988) 735 – 738.
68. W. Loy, *TPI Text. Prax. Int.* **46** (1991) 1241 – 1244.
69. T. Platzek, *Melliand Textilber.* **77** (1996) 774 – 778.
70. F. Klatschka, *Melliand Textilber.* **75** (1994) 193 – 202.

Sillimanite → Silicates
Siloxanes → Silicon Compounds, Organic

Silver, Silver Compounds, and Silver Alloys

Hermann Renner, Degussa AG, Hanau, Federal Republic of Germany (Chaps. 1 – 4, Sections 10.3, 13.2 and 13.3)

Günther Schlamp, Demetron GmbH, Hanau, Federal Republic of Germany (Sections 2.2, 10.6.1, 10.7, Chap. 12)

Klaus Zimmermann, Degussa-Hüls AG, Hanau, Federal Republic of Germany (Chaps. 5 – 8, Sections 10.10, 10.11)

Wolfgang Weise, Fachhochschule Esslingen, Esslingen, Federal Republic of Germany (Chap. 9)

Peter Tews, Allgem. Gold und Silberscheideanstalt (AGOSI), Pforzheim, Federal Republic of Germany (Sections 10.1, 10.2)

Klaus Dermann, Ducera Dental GmbH und Co. KG, Hanau, Federal Republic of Germany (Section 10.4)

Alfons Knödler, Forschungsinstitut für Edelmetalle und Metallchemie, Schwäbisch Gmünd, Federal Republic of Germany (Section 10.5)

Karl-Heinz Schröder, Linsengericht-Walrode, Federal Republic of Germany (Section 10.6.2)

Bernd Kempf, Degussa-Hüls AG, Hanau, Federal Republic of Germany (Sections 10.8, 10.9)

Hans Martin Lüschow, Degussa-Hüls AG, Hanau, Federal Republic of Germany (Chap. 11)

Ralf Drieselmann, dmc2 Degussa Metals, Catalysts AG, Frankfurt, Federal Republic of Germany (Section 12.3)

Catrin Peter, Klinikum der Friedrich-Schiller-Universität, Jena, Federal Republic of Germany (Section 13.1)

Rainer Schiele, Klinikum der Friedrich-Schiller-Universität, Jena, Federal Republic of Germany (Section 13.1)

1. History 2
1.1. Centers of Silver Production 2
1.2. Extent of Production 4
1.3. Development of Production Processes 4
1.4. Monetary Significance and Price Structure 6
2. Properties 6
2.1. Atomic Properties 6
2.2. Physical Properties 7
2.3. Chemical Properties 7
3. Occurrence and Raw Materials . . 8
3.1. Formation, Abundance, and Distribution of Ores 8
3.2. Silver Minerals 8
3.3. Deposits 9
3.4. Secondary Silver 10
3.5. Resources and Reserves 11
4. Extraction from Ores 11
4.1. Extraction from Silver Ores 11
4.1.1. Smelting 11
4.1.2. Amalgamation 11
4.1.3. Cyanidation 12
4.1.4. Thiosulfate Leaching (Patera Process) 13
4.1.5. Metallurgical Processes 13
4.2. Extraction from Lead and Lead – Zinc Ores 13
4.2.1. Production of Lead Bullion 13
4.2.2. Cupellation without Prior Silver Enrichment 13
4.2.3. Silver Enrichment by the Pattinson Process 14
4.2.4. Silver Enrichment by the Parkes Process 14
4.2.5. Cupellation of Enriched Lead 14
4.2.6. Silver Extraction from Electrolytic Lead Refining 15
4.3. Extraction from Copper and Copper – Nickel Ores 15
4.3.1. Formation of Silver-Containing Copper Anode Slimes 15
4.3.2. Pretreatment of Copper Anode Slimes 15
4.3.3. Processing of Copper Anode Slimes 16
4.3.4. Silver Extraction from Copper Matte 18
4.4. Extraction from Gold Ores 19
4.5. Extraction from Tin Ores 19
5. Recovery from Secondary Silver . 19
5.1. Via Copper Smelters 19

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_107

5.2. Via Lead Smelters 19
5.3. Via the Lead – Silver Smelting Process 20
5.4. Via Scrap Metal Leaching 22
5.5. Via Scrap Metal Electrolysis 22
5.6. Processing of Flue Dust 23
5.7. Processing of Copper Matte 24
5.8. Processing of Photographic Materials 24
5.9. Surface Desilvering 26
5.10. Processing of Special Scrap 26
6. Silver Refining 26
6.1. Fine Smelting 27
6.2. Refining with Nitric Acid (Inquartation) 27
6.3. Refining with Sulfuric Acid (Affination) 27
6.4. Möbius Electrolysis 27
6.5. Balbach – Thum Electrolysis 28
7. Silver Compounds 29
7.1. Silver Nitrate 29
7.2. Silver Halides 31
7.3. Silver Oxides 33
7.4. Other Soluble Silver Compounds . 33
7.5. Other Insoluble Silver Compounds 34
7.6. Silver Complexes 34
7.7. Explosive Silver Compounds 35
8. Disperse Silver 36
8.1. Silver Powder 36
8.2. Colloidal Silver 37
9. Silver Alloys 38
9.1. Binary Silver Alloys 38
9.2. Ternary Silver Alloys 44
9.3. Quaternary Silver Alloys 46
9.4. Manufacturing 46
10. Uses 50
10.1. Coins 50
10.2. Jewelry 51
10.3. Medicine 52
10.4. Dentistry 54
10.5. Coatings 54
10.5.1. Silver Electroplating 54
10.5.2. Silver Plating by Chemical Reactions 55
10.5.3. Mechanical and Thermomechanical Plating 56
10.5.4. Physical Vapor Deposition. 56
10.5.5. Firing Processes 56
10.6. Electronics and Electrical Technology 57
10.6.1. Electronics 57
10.6.2. Electrical Technology 61
10.7. Brazing Alloys 64
10.8. Chemical Equipment 65
10.9. Catalysts 67
10.10. Photography 68
10.11. Other Uses 68
11. Specifications and Analysis 69
11.1. Qualities and Commercial Grades 69
11.2. Sampling 69
11.3. Qualitative Analysis 69
11.4. Quantitative Analysis 70
11.5. Purity Analysis 71
11.6. Trace Analysis 71
11.7. Argentometry 72
12. Economic Aspects 72
12.1. Production 72
12.2. Demand 73
12.3. Silver Market and Trading 73
13. Toxicology, Occupational Health, and Safety 73
13.1. Toxicology 73
13.2. Ecotoxicology 74
13.3. Explosion Hazards 75
14. References 75

1. History [19]

1.1. Centers of Silver Production [1–6]

Ancient Times. Like gold, silver occurs in native form and therefore has been known to humans since prehistoric times. The oldest finds of silver whose age can be determined date from around 4000 B.C. in Egypt (see Fig. 1). In the following years, silver appeared in the entire cultural region of the eastern Mediterranean. In the second millennium B.C., the Phoenicians obtained large quantities of silver from deposits on the Iberian Peninsula, which were distributed by their companies of traveling merchants.

The first upsurge in the production of silver took place when the Greeks exploited the mines of Laurion, which reached their peak productivity in ca. 500 B.C. The Greeks also worked silver mines in Thrace, Asia Minor, and the entire Aegean region. The Athenian empire and its claim to power were based to a large extent on the role of silver for coinage. Modern methods of trace analysis have elucidated the relationship between the extraction and processing of the Greek silver deposits and their effects on the political developments at that time [7–11].

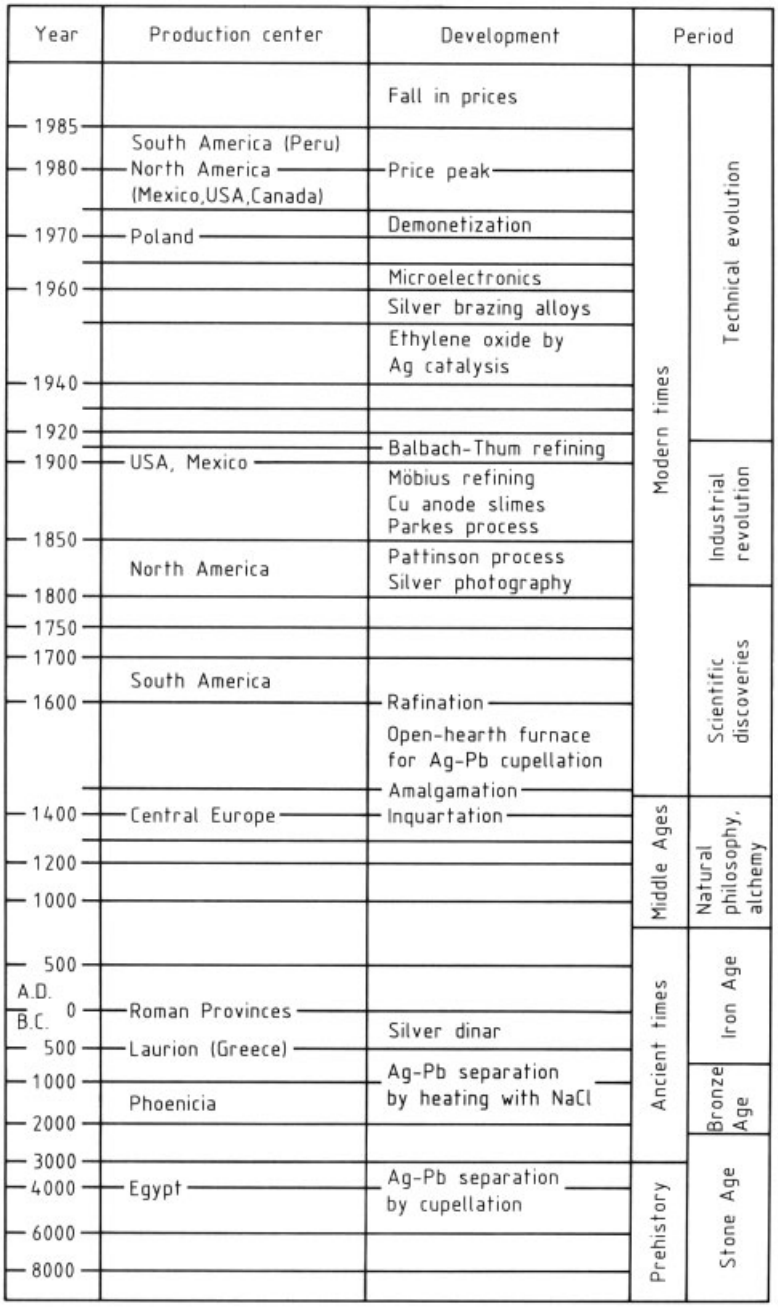

Figure 1. Historical survey of important silver production centers and silver processing

Rome was initially poor in silver. However, considerable quantities were soon obtained in the provinces that fell to the Roman Empire (i.e., southern Spain, the Balkans, the Carpathians, etc.). In many Roman expeditions of conquest (e.g., the Punic wars and those in the period of the civil wars), the seizing of silver mines was often a cause of war-like disputes. The introduction of the silver denarius as the common cur-

rency in ca. 300 B.C. elevated the political importance of the national metal.

Undoubtedly, silver was also extracted in pre-Christian times in middle and northern Europe, although this only became known through the Romans' written records. Likewise in India, China, and Japan the production of silver probably dates back to ancient times.

When the migration of the Germanic people began and the political dissolution of the Roman Empire occurred, silver production practically ceased. This situation changed with the reorganization of the western lands by the Carolingians.

Middle Ages. During the Middle Ages, the extraction of silver became widespread in central Europe. Some of the production locations are listed here, with their starting dates: Bohemia (Pribram: 753; Miss: 1137; Kuttenberg: 1240), Sweden (Sala: 8th century), Saxony (Freiberg, which produced a total of 5500 t silver: 1103; Mittweida and Frankenberg: 922), Erzgebirge (Joachimsthal: 1510; Annaberg, Schneeberg), Alsace (Markirch: 7th century), the Lahn region, Siegerland (ca. 860), Silesia (Beuthen, Tarnowitz), Hungary (Schemnitz), Norway (Kongsberg), Tyrol, Steiermark, Salzburg, the southern Black Forest (Freiburg im Breisgau), Harz Rammelsberg, which provided much of Europe's silver in the 11th and 12th centuries: 968; St. Andreasberg; Mansfeld, which was the source of most of the silver produced in Germany at that time: 1450.

Many of these deposits were exhausted by the end of the Middle Ages and often long before this. However, the mines in the Harz, Siebenbürgen, Freiberg, Norway, and Alsace retained some of their importance up to the industrial revolution.

The deposits in India, Japan, China, and pre-Columbian America, were also mined during the Middle Ages.

Modern Times. After the discovery of America, silver was plundered by the Spanish and brought to Europe in large quantities. Also, new deposits were developed. Soon, America was producing more silver than Europe, and prices collapsed. Up to ca. 1700, world production was dominated by Central and South America (Peru, Bolivia, Chile, and Argentina).

In the 1800s, the main center of production moved to North America, i.e., Mexico, Canada, and the United States (Nevada). Also, significant quantities of silver were produced in Russia and Australia, and the output from Europe increased, although by then this was almost entirely associated with lead and zinc ores.

After World War II, Poland became an important producer following the exploitation of newly discovered silver-containing copper deposits.

1.2. Extent of Production

By the beginning of the Middle Ages (ca. 800 A.D.), a total of ca. 50 000 – 100 000 t of silver is estimated to have been extracted. By then the silver deposits of the European Mediterranean region, almost the only regions to be mined in ancient times were effectively exhausted.

At the end of the Middle Ages, world production had almost dropped to 50 t/a. During the whole of the Middle Ages, ca. 200 000 – 300 000 t silver was extracted, most of which came from Europe, mainly the German-speaking area.

By 1975, average world primary silver production ca. $> 10\,000$ t/a, and by 1991, it reached a peak of 11 500 t/a. The countries of North and South America today account for more than two-thirds of world production. In 1975, Poland was by far the largest producer in Europe. For primary production figures, see Table 1.

In total, $(1 - 1.2) \times 10^6$ t silver has been extracted worldwide. This is about 10 times the figure for gold production, 100 times the total world production of platinum, and 50 times that of the platinum-group metals.

1.3. Development of Production Processes [1–13]

Very little has been written about the techniques of silver extraction in ancient times. The Egyptians are thought to have separated gold from silver by heating with salt. They must therefore have known how to convert the silver chloride produced into metallic silver.

Table 1. Silver production excluding CIS and China (t/a)

Region	1989	1991	1992	1994	1996	1997
Mining production	11 481	11 560	11 480	10 650	11 700	12 160
Mexico	2306	2250	2400	2500	2535	2580
United States	2007	1850	1820	1770	1550	1800
Peru	1840	1770	1500	1520	1970	2070
Canada	1306	1290	1150	680	1240	1170
Australia	1168	1180	1080	1100	1150	1160
Other countries	2854	3220	3530	3020	3250	3380
Secondary silver	3344	3380	3450	3300	4030	4170
Total	14 825	14 940	14 930	13 950	15 730	16 330

The main source of silver in ancient times was silver-containing lead ores. In the Middle Ages, deposits of native silver were also important, especially in Central Europe. The rich ores could be smelted simply after hand separation (i.e., after crushing and separating from the worthless rock). Large pieces weighing more than 100 kg, or large blocks weighing many tons and containing native metal, were found and processed in this way.

The Greeks, the Romans, and the miners of the Middle Ages knew how to process silver-containing lead ores by smelting to produce lead. In the Middle Ages, the lead was removed by oxidation in cupellation furnaces to produce silver with a fineness of 650/1000 – 875/1000. This could then be oxidized further to give a fineness of 990/1000.

When the silver was accompanied by copper in the deposit, the liquation process [14] was used. The silver-containing copper was melted with lead, and the silver-rich liquid phase recovered by liquation was blown with air.

Introduction of the amalgamation process in 1566 improved silver productivity and production rates in Mexican and South American mines. This process later became dominant and was used in the United States and, from 1800, in Europe in a modified form suitable for a wide range of raw materials.

After the mid-1800s several new techniques were developed that replaced the old processes or improved economics. For example, the lead cupellation process was made more efficient by introduction of the Pattinson process (1833; see Section 4.2) which for a time was very widely used. This was based on the segregation of solid lead from the lead – silver melt by cooling. This increased the silver content to 2.5 % and thereby decreased the load on the cupellation process. The Parkes process (1850; see Section 4.2) soon replaced the Pattinson process because of its superior economics. In this process, silver was transferred from the lead to zinc floating on the surface of the lead; this formed a crust that could be removed. The Parkes process is still virtually the only process used today for extracting silver from lead ores. It is also used to extract silver from lead – zinc and zinc ores.

The chloridizing roasting of copper ores followed by leaching with brine to dissolve the silver chloride formed was introduced in 1843 by Augustin (see Section 4.3.4); it was soon replaced by sulfation roasting (see Section 4.3.4), at temperatures below the decomposition temperature of silver sulfate and above that of copper sulfate, followed by leaching with water. This was known as the Ziervogel process, and for almost 100 years most of the silver in German ores was extracted in this way. The sulfuric acid extraction of dead-burnt copper ore in a stream of air (see page 16), to form a silver-containing slime, was used for only a short period. After 1856, thiosulfate leaching (see Section 4.1.4) idized roasted ore was used primarily in Mexico and California. In the Patera process, sodium thiosulfate was used; in the Kiss process, calcium thiosulfate; and in the Russel process, sodium copper thiosulfate. Since ca. 1900, cyanide leaching (see Section 4.1.3) is still used for silver ores that are not smelted, has been the most important aqueous treatment process.

Since ca. 1876, almost all the silver associated with copper has been obtained from the electrolytic refining of copper. The silver appears in the copper anode slimes, which are generally treated by pyrometallurgical processes (see Sections 4.3.3, 4.3.4).

To separate silver from the gold in auriferous silver (the end product of nearly all smelting and

hydrometallurgical extraction processes), the nitric acid process (developed in 1433–see Section 6.2) and the sulfuric acid process (dating from 1802–see Section 6.3) have been completely replaced by electrolytic refining.

1.4. Monetary Significance and Price Structure [1,15,16]

From ancient times to the Middle Ages and later, even into the present century, silver was the classic metal for currency. In comparison, gold played an almost insignificant role.

In the 6th century B.C., King Croesus ordered the minting of silver coins. In ancient Greece, silver currency was the basis of the economic system. In 296 B.C. the Romans introduced the silver denarius as the general method of payment, and gold ceased being used in coins from 217 B.C. Silver was the basis of the coinage reform by the Byzantine Empire in A.D. 310, and was recognized throughout the Mediterranean region, so that silver maintained its importance as a coinage metal for many centuries. Charlemagne made silver the basis of the currency system of his empire.

During the last two centuries, gold began to displace silver, mainly due to the reduction in the price of silver caused by the surplus from new production, as well as the demonetization of silver.

Between the two world wars, strong moves, triggered by inflation and the world economic crisis, were made to remonetize silver, but these largely came to naught.

After World War II, because of the increasing demand and the shortage of supply, the price of silver rose so much that by 1970 the value of the silver in the most important silver coins still in circulation (U.S. dollars and Swiss francs) exceeded the face value of the coins. Silver was therefore virtually completely demonetized.

After the 17th century, the ratio of the prices of silver and gold remained fairly constant at ca. 1 : 15 for a long period. However, free price movements have continually tended to shift this ratio, and it has often been kept constant by state intervention. A fairly gradual increase in the price of gold began ca. 120 years ago. At the beginning of World War I, the ratio reached 1 : 40. During the World War II it rose to 1 : 100, and during the following three decades it fluctuated around 1 : 40. In 1981 it was 1 : 50, and in 1992 it was 1 : 95.

Silver prices have been freely determined by supply and demand from 1970, after which the price rose steadily from ca. 90 €/kg to 150–260 €/kg. In late 1979 and early 1980, speculative dealings of a hitherto unimagined volume occurred, and the price of silver briefly reached nearly 1406 €/kg [17]. In the following decade, and up to the present, a rapid and fairly continuous decline in price has taken place, reaching a low of 8 €/kg in 1992. The price of silver now reflects its status as an industrial metal and depends on the current economic situation. In parallel with this, silver has lost its importance as an investment metal. The high stock levels (see Chap. 12) and sales of these stocks from time to time also affect price reductions.

2. Properties [18–24]

Silver [*7440-22-4*], like gold and the platinum-group metals, is a noble metal. Silver, gold, and copper are known as the coinage metals.

2.1. Atomic Properties [25]

Silver has an atomic number of 47 and a relative atomic mass of 107.8682. Natural silver consists of the stable isotopes ^{107}Ag (51.8 %) and ^{109}Ag (48.2 %). Many artificial isotopes with relative atomic masses between 102 and 117 are known.

The electronic configuration of silver is [Kr]$4d^{10}5s^1$. The $5s^1$ shell is the principal valence orbital, which accounts for the preference for oxidation state 1. The $4d$ shell can also act as a valence orbital, resulting in oxidation states 2 and 3. Thus, silver behaves as a transition metal due to the contribution from the inner valence orbital. Since the differences between the energies of the valence electrons are small, silver has a tendency to form complexes with covalent bonds and to exhibit other properties such as catalytic activity typical of transition metals.

The atomic radius of silver (with 12-fold metallic coordination) is 0.144 nm; the ionic radius of Ag^+ (six-coordinate) is 0.137 nm, and of Ag^{3+} (six-coordinate) 0.075 nm.

2.2. Physical Properties

The physical properties of silver are generally intermediate between those of copper and gold. These three elements have strong similarities, although silver has some anomalous properties, which are directly related to the $d^{10}s^1$ electronic configuration and the irregularity of the atomic radii. Some physical data are given below:

Lattice constant at 20 °C	0.4086 nm
Atomic radius (in metal)	0.1445 nm
Melting point	961.9 °C
Boiling point	2212 °C
Specific heat capacity at 25 °C	0.23 J kg^{-1} K^{-1}
Thermal conductivity	418 W m^{-1} K^{-1}
Vapor pressure at	
1030 °C	1.33 Pa
1190 °C	13.3 Pa
1360 °C	133 Pa
1580 °C	1330 Pa
1870 °C	13 300 Pa
Density at 20 °C	10.49 g/cm^3
Density of liquid at *mp*	9.35 g/cm^3
Density of liquid at 1250 °C	9.05 g/cm^3
Brinell hardness	26
Modulus of elasticity	82 000 MPa
Tensile strength	140 MPa
Resistivity at 0 °C	1.465 Ωcm

Like copper and gold, silver crystallizes in a face-centered cubic structure in which each metal atom is surrounded by 12 neighbors. This high degree of symmetry results in a structure with many slip planes and is the reason for its good mechanical formability. Silver is only slightly less ductile than gold. It can be given a high polish and has a warm, white luster. Very thin films are blue or green by transmitted light, thicker films being yellow to brown.

Silver has the highest electrical conductivity, the highest thermal conductivity, and the lowest electrical contact resistance of all metals.

2.3. Chemical Properties

Molten silver absorbs oxygen and releases it during solidification, with bubbling of the metal surface. At 1000 °C, 1 cm^3 of molten silver dissolves a maximum of ca. 20 cm^3 of oxygen, corresponding to a stoichiometric composition of $AgO_{0.01}$. The solubility of oxygen in silver just below its melting point is only ca. 1/40 of that at 1000 °C. Hydrogen is very slightly soluble in molten silver, whereas nitrogen, carbon monoxide, carbon dioxide, and the noble gases are insoluble.

Halogens react violently with silver at red heat. Moist chlorine gas corrodes silver even at low temperature. Ozone blackens the surface of silver due to oxide formation.

Hydrogen sulfide (gaseous or in solution) and aqueous solutions of sulfides immediately form a black coating of silver sulfide on the surface of the metal. At high temperature, the reaction with hydrogen sulfide gas or sodium polysulfide solution can lead to complete conversion of silver to silver sulfide.

The solubility of metallic silver in water is very low, but it is sufficient to exhibit an oligodynamic (bactericidal) effect (see Section 13.2).

The preferred medium for dissolution of silver is hot, concentrated nitric acid (see Fig. 2):

$$3\,Ag + 4\,HNO_3 \longrightarrow 3\,AgNO_3 + 2\,H_2O + NO_3$$

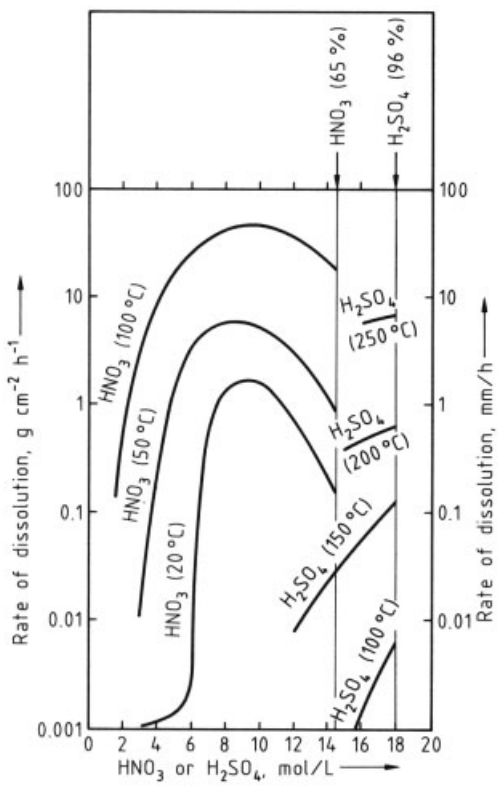

Figure 2. Rate of dissolution of silver in nitric acid and sulfuric acid as a function of concentration and temperature

Silver also dissolves rapidly in hot concentrated sulfuric acid (Fig. 2):

$$2\,Ag + 3\,H_2SO_4 \longrightarrow 2\,AgHSO_4 + SO_2 + 2\,H_2O$$

Rates of dissolution in dilute sulfuric acid are appreciable if oxygen or hydrogen peroxide are present. The initially vigorous reaction with aqua regia

$$2\,Ag + 3\,HCl + HNO_3 \longrightarrow 2\,AgCl + 2\,H_2O + NOCl$$

is rapidly slowed by the formation of a coating of silver chloride. Most common oxidizing aqueous media attack silver, including chromic acid, permanganate solutions, persulfuric acid, selenic acid, and aqueous solutions of free halogens. The following also attack silver: hydrochloric acid, phosphoric acid, bromine water, solutions of alkali-metal chlorides, copper chloride, and iron chloride. The reactions are often strongly dependent on temperature and in some cases on the formation of protective coatings. Silver is resistant to aqueous solutions of organic acids and aqueous alkali.

Silver is fairly resistant toward fused sodium hydroxide at ca. 550 °C in the absence of atmospheric oxygen and moisture. Potassium hydroxide is more aggressive. Silver is attacked by fused salts such as sodium peroxide, potassium nitrate, sodium carbonate, potassium hydrogensulfate, and potassium cyanide.

3. Occurrence and Raw Materials

[1–3, 19, 26–29]

3.1. Formation, Abundance, and Distribution of Ores

Because of its chalcophilic character, silver became concentrated in oxidic sulfidic materials during the formation of the Earth's mantle, which has a silver content of ca. 10 ppm. The metallic core of the Earth should have a much lower silver content because silver does not alloy with iron or nickel. This is confirmed by investigation of iron meteorites. In the region accessible to mining (i.e., to a depth of 5 km), the Earth's crust has an average silver content of 0.05 – 0.1 ppm. Thus, silver is 100 to 1000 times more abundant than gold or platinum. The average silver content of the oceans is ca. 0.001 ppm [30], but it often reaches much higher values at the mouths of rivers. This enormous reserve of silver (10^9 t) cannot, however, be extracted economically.

Silver became concentrated at particular sites in the Earth's crust by the recrystallization of silicate magmas caused by volcanic action in impregnation zones. Most sulfidic, arsenidic, and antimonidic formations were originated in this way. Hydrothermal processes caused conversion to metallic deposits. Also, the action of the atmosphere and hydrosphere was involved (e.g., in the formation of caps over magmatic deposits). Unlike gold and the platinum metals, silver has not formed sedimentary deposits to any extent, since it is dissolved much more readily than gold or platinum by the action of the hydrosphere. However, silver-containing sedimentary deposits of other elements are known (e.g., Mansfeld copper slate).

3.2. Silver Minerals [3]

Nearly all silver minerals are compounds of silver with sulfur, its homologues selenium and tellurium, and their neighbors in the periodic table, arsenic, antimony, and bismuth. These ores generally have a metallic luster and are semiconductors. Other silver minerals include the halides and the free metal. The major mineralogical groups – oxides, silicates, spinels, phosphates, gypsum, and carbonates – are not represented.

Native silver occurs in the form of nuggets, dendrites, sand-like material, or sometimes as large lumps or blocks. It often forms mixed crystals with gold, and can also contain copper and other metals and metalloids.

Argentite, Ag_2S, is the most common silver compound found in the sulfidic ores of other metals, especially galena, PbS, in which it is present with other sulfidic minerals in the form of very fine inclusions. In the Middle Ages, argentite was often found in high concentrations and was a mineral in its own right with a high silver content.

Pyrargyrite, antimony silver blende, Ag_3SbS_3, is an important mineral occurring in true silver ores, which are now found and mined mainly in Central America.

Proustite, arsenic silver ore, Ag_3AsS_3, is of similar occurrence to the corresponding antimony-containing sulfur compound Ag_3SbS_3.

Stephanite, Ag_5SbS_4, is very characteristic of the silver ores familiar to miners from early times.

Cerargyrite, horn silver, AgCl, does not occur in large quantities. It has extremely low hardness and crystallizes with the NaCl lattice type.

Apart from the compounds mentioned above, many other well-defined silver minerals exist. Some of them, particularly the sulfur compounds, still make a significant contribution to silver production.

Some examples are given below:

Miargyrite	$AgSbS_2$
Polybasite (Eugenite)	Ag_9SbS_6
Freibergite	$(Cu, Ag)_{12}Sb_4S_{13}$
Freieslebenite	$Ag_3Pb_2Sb_3S_8$
Andorite	$AgPbSb_3S_6$
Stromeyerite	$Ag_2Cu_2S_2$
Schapbachite	$AgBiS_2$
Argyrodite	Ag_8GeS_6
Sternbergite	$AgFe_2S_3$
Argentopyrite	$AgFe_3S_5$
Argyropyrite	$Ag_2Fe_7S_{11}$
Sundite	$AgFeSbS_9$
Hessite	Ag_2Te
Petzite	Ag_3AuTe_2
Silvanite	$AgAuTe_4$
Bromyrite (Bromargyrite)	AgBr
Embolite	Ag(Cl, Br)
Iodite (Iodargyrite)	AgI

Silver is a component of most sulfidic minerals, at least at the parts-per-million level. The silver compounds are present as mixed crystals or as microscopic inclusions with a crystal structure that deviates from that of the matrix.

3.3. Deposits [3, 31–36]

The known silver deposits are associated mainly with vulcanism in the Tertiary period running, for example, in a continuous line along the entire west coast of the American continent (Fig. 3).

True silver ores, in which silver is practically the only valuable metal, are mined mainly in Mexico (Pachuca, Frenillo, and San Fernando de l'Oro) and to a small extent in the United States (Utah). In Europe, such deposits are almost completely exhausted. Worldwide, they still account for more than one-tenth of the silver mined. Depending on the price of silver, it can be economic to process materials containing 0.01 % (100 ppm) silver, or even less. The deposits in Ontario, which contain native silver associated with cobalt sulfide and arsenide, can also be regarded as true silver ores, since only the silver is extracted.

Argentiferous lead and lead – zinc ores, which often occur in association, are distributed widely. In almost all silver-producing countries, they constitute an important fraction of the silver produced by mining and account for ca. 50 % of world silver production. The separation and recovery of silver during lead production is profitable even for a silver content of a few parts per million. The important reserves in Europe include those in Spain (Lineares, Sierra Morena), Germany (Rammelsberg/Harz), Yugoslavia (Trepca), and to a lesser extent Italy (Montevecchio, Monteponi), Greece (Laurion), and France (Savoy). Argentiferous lead and lead – zinc ores are mined in large quantities in the United States (Helen, Montana; Denver, Colorado; Coeur d'Alene, Idaho), Mexico, Peru, Bolivia, and Argentina. In Australia and Tasmania, lead and lead – zinc ores are the primary source of silver. In North Africa (Djebel Hallouf, Tunisia; Beni-Sat, Algeria), Nigeria, and Namibia (Otavi), these are of some importance. In the Republic of South Africa, other silver ores are more important. In India, the extraction of silver from lead and lead – zinc ores is extremely important, and this appears to also be true of the Commonwealth of Independent States (CIS).

Argentiferous copper ores, which sometimes also contain nickel and cobalt, satisfy ca. 30 % of the world silver demand. Even very low contents of silver can be recovered profitably from copper ores, since almost all the copper is electrorefined, and the silver remains in the anode slimes. In Europe, the largest reserves of argentiferous copper ore are in Poland (Katowice), Yugoslavia (Bor), and Sweden. Smaller reserves can be found in Finland. The CIS, from which little information is available, is also believed to have reserves of this type. The largest argentiferous copper reserves in the world are in Canada (Sudbury, Ontario, Rouyn-Noranda, Quebec). Silver is also recovered during copper smelting in Chile, Bolivia, Peru, Katanga, Namibia, Zimbabwe, and the Republic of South Africa.

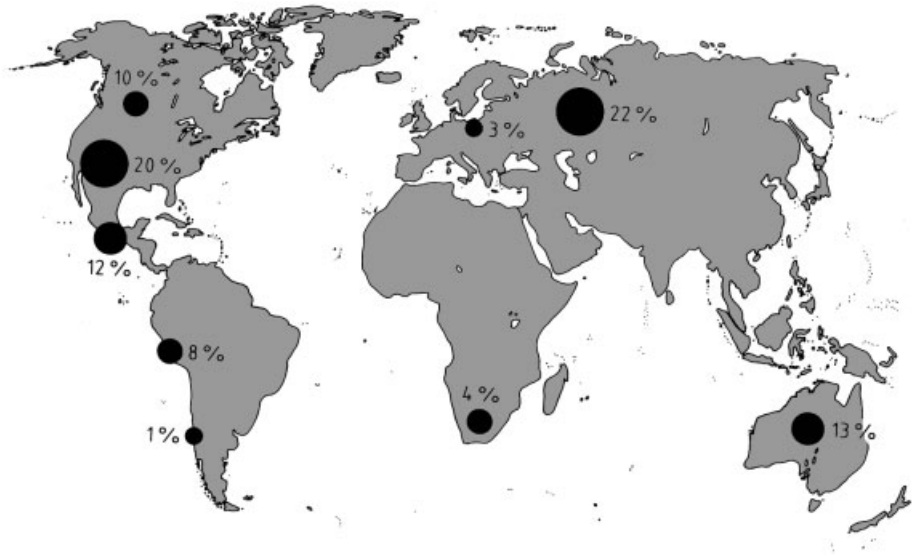

Figure 3. Reserves of silver by region

Extraction of silver from gold ores is carried out mainly in the Republic of South Africa, and also in Canada, Alaska, and Sweden (Boliden). The quantity obtained in this way is <10 % of world silver production. Silver-containing tin deposits are concentrated in Bolivia (Cerro de Pasco, Potosi) but also occur in Burma and Argentina. Silver produced from tin ores amounts to only 2 % of world production.

3.4. Secondary Silver [37–42]

The demand for silver that cannot be met by mining must be made up by recycling and by release from stocks (see Section 1.4). However, stocks of silver are often in a condition (e.g., type of alloy or low purity) that is not accepted directly by the market. Hence, these materials must generally be refined.

Silver coins have been almost entirely withdrawn from circulation. Some were refined immediately, but a smaller number were saved. From time to time, large stocks of these coins are released unpredictably, usually in association with speculative dealings on the silver market.

Even when the coins are of uniform composition, the alloy generally cannot be reused directly because, for example, the quantity is too large or the alloy is too complex. The silver content of coinage alloys is 40 – 90 %. Copper is the other main alloy component, and nickel, zinc, or traces of other metals are generally present.

The old Indian silver reserves come onto the market in the form of fittings and similar articles; the silver content is 90 – 95 %.

Photographic slimes, produced during the manufacture of films and photographic paper, have a silver content of 5 – 30 %. The silver is present as a halide, the remainder being mainly gelatin and water.

Approximately 50 % of the silver used in the photographic industry ends up in fixing baths. The silver-containing fixing bath solutions are rarely supplied directly to refineries. Instead, silver is usually extracted from them where they were produced, either by electrolysis to yield compact crude silver of fairly high purity (see Section 5.5) or by chemical reduction to give slime concentrates (10 – 60 % Ag).

The archives of X-ray films and motion pictures, which are stored for a limited period only, are a continuing source of recycled silver. In developed X-ray film the silver content is 1 – 2 %; in developed black-and-white film it is 0.2 %, but in developed color film the amount is negli-

gible. The films are converted to ashes or slimes with silver contents up to 90 %.

Offcuts of X-ray and graphic film from the photographic industry and film material stored too long (e.g., from military stocks) have a silver content up to 3 %.

Photographic printing paper offcuts from the photographic industry have a silver content of ca. 0.4 %. These papers used to contain up to 50 % barium sulfate filler, but modern papers are mainly polymer coated and have a low filler content.

Scrap silver from tableware and ornaments usually with silver content of 80 or 92.5 %, is important during periods of high silver prices. The treatment of silver-plated utensils (e.g., cutlery with ca. 4 – 5 % Ag) is usually profitable only in large quantities.

The electrical industry and its suppliers generate waste materials during the manufacture of electrical contacts at various stages of the production processes. Waste also arises from the production of small batteries for domestic electronic use and larger units for the military and air transport areas [43]. These cells often contain mercury. Silver-plated switching elements are produced in the electronics industry.

The chemical industry provides spent silver-containing catalysts [38]. These include silver wire mesh or silver crystals used in the oxidation of methanol to formaldehyde and Ag – α-Al_2O_3 catalysts for the oxidation of ethylene to ethylene oxide.

A regular, though small, quantity of silver comes from equipment and silver-plated vessels or pipework used in the chemical industry. Finally, the amount of silver waste from the production of dental amalgams is not inconsiderable. It consists of granules of Ag – Sn – Hg alloys and is usually contaminated with other materials, including organic substances.

3.5. Resources and Reserves [15, 24, 26, 36]

Mineral "resources" include all the principal known deposits, including those that have not yet been proved, but are only assumed. They also include deposits that are known for certain to exist, but cannot be extracted for economic or ecological reasons [44].

Mineral "reserves" include only those deposits that are already economically exploitable, or whose existence, based on prospecting work, is certain or probable [44]. These mineral reserves amount to ca. 250 000 t of silver at the present time. With a demand of ca. 10 000 t/a, silver is one of those metals whose known raw material reserves can supply the demand only for a limited period. By far the largest proportion of the silver is contained in lead, zinc, and copper ores, so silver production is linked primarily to the demand for these metals. This is probably not a good situation from viewpoint of future supplies. However, the advantage is that silver reserves are much more widely distributed than those of gold and the platinum metals, so the supply of silver to industrialized nations would be much less threatened by possible political crises than the supplies of other important raw materials.

4. Extraction from Ores [1, 3, 12, 19, 20, 44–46]

4.1. Extraction from Silver Ores

4.1.1. Smelting

In the early days, considerable quantities of silver were extracted simply by smelting high-percentage silver ores under oxidizing or reducing conditions with conversion of accompanying materials into slag. The starting materials were almost exclusively ores containing elemental silver, silver sulfide, or silver halides.

4.1.2. Amalgamation

For many centuries, amalgamation was the most important process for the treatment of true silver ores. Today, it is rarely used because of high costs, poor yields, and the toxicity of mercury. It is now carried on in privately owned small mines. The amalgamation process can be used directly only if the silver is native or is present as the chloride. In this form it will amalgamate on contact with mercury without any other treatment.

Amalgam formation does not take place as rapidly as with gold, but the reaction velocity

and rate of dissolution of silver in mercury are nevertheless fairly high. The copper-plate amalgamation process used with gold cannot be used with silver because reaction times are too short to give a useful yield. To ensure problem-free amalgam formation, the silver or silver halide particles should first be liberated, preferably by wet grinding. However, an extremely fine particle size is detrimental to the process.

Sulfidic, arsenidic, and antimonidic silver minerals do not react directly to form an amalgam and therefore require pretreatment. This can consist of a simple oxidative roasting process or roasting accompanied by chloridization or sulfation.

The reaction is accelerated and the yield improved by the addition of base metals (e.g., iron and copper), sodium chloride, copper (I) or copper (II) chloride, or copper sulfate. The following are some of the reactions that occur:

$$CuSO_4 + 2\,NaCl \longrightarrow Na_2SO_4 + CuCl_2$$

$$Ag_2S + CuCl_2 \longrightarrow 2\,AgCl + CuS$$

$$2\,Ag_3AsS_3 + 3\,CuCl_2 \longrightarrow 6\,AgCl + 3\,CuS + As_2S_3$$

$$CuCl_2 + Cu \longrightarrow 2\,CuCl$$

$$2\,Ag_3AsS_3 + 6\,CuCl \longrightarrow 6\,AgCl + 3\,CuS + 3\,Cu + As_2S_3$$

$$AgCl + Cu + Hg \longrightarrow CuCl + AgHg$$

The operating techniques and the choice of reaction conditions have undergone many changes. In the Patio process, a paste of finely ground ore, sodium chloride, mercury, and water was kneaded. Other processes employ copper or iron pans, stirring apparatus, or rotating vessels, either with or without heating. The resulting metallic phase is washed to remove accompanying materials, and the residual liquid mercury is separated from the solid amalgam. The latter is heated in retorts to distill off the mercury and yield impure silver. The overall yields of silver obtained in the amalgamation process reach 95 % under favorable circumstances. The loss of mercury amounts to 0.2 – 2 kg for each kilogram of silver.

4.1.3. Cyanidation

Because of its favorable economics, the aqueous cyanide process is the dominant process for extraction of silver from true silver ores. The conditions are essentially the same as those for the gold cyanide process (see → Gold, Gold Alloys, and Gold Compounds, Chap. 4.3.1.).

The ore is crushed, wet milled, and treated with an aqueous solution of sodium cyanide (5 – 10 g NaCN/L) for several days in a Pachuca tank [3]. Solids are removed by a thickener and filter, and silver is precipitated from solution by cementation on zinc dust. Other possible methods of recovering silver are ion exchange or electrolytic reduction [4]. Another process involves adsorption of the silver cyanide complex onto active carbon [47, 48], analogous to the carbon-in-pulp process for gold (→ Gold, Gold Alloys, and Gold Compounds, Chap. 4.3.2.). This process has been fully developed, but has not achieved the same importance or productivity as in the case of gold.

Silver chloride, either naturally occurring or produced by a chloridizing roast, is dissolved according to the equation

$$AgCl + 2\,NaCN \longrightarrow Na[Ag(CN)_2] + NaCl$$

Elemental silver is dissolved by passing air into the solution:

$$4\,Ag + 8\,NaCN + O_2 + 2\,H_2O \longrightarrow 4\,Na[Ag(CN)_2] + 4\,NaOH$$

Silver sulfide is also attacked, but more slowly, according to the following equations:

$$2\,Ag_2S + 8\,NaCN + 2\,O_2 + H_2O \longrightarrow 4\,Na[Ag(CN)_2] + Na_2S_2O_3 + 2\,NaOH$$

and

$$2\,Ag_2S + 10\,NaCN + O_2 + 2\,H_2O \longrightarrow 4\,Na[Ag(CN)_2] + 2\,NaSCN + 4\,NaOH$$

If the ores contain cyanide-insoluble silver compounds such as proustite (Ag_3AsS_3) and pyrargite (Ag_3SbS_3), cyanide extraction must be preceded by chloridization roasting. The dissolution of selenidic and telluridic compounds of silver by cyanide is also difficult, but is improved by fine grinding and roasting, as well as by the use of cyanogen bromide, which also increases the solvent action of the sodium cyanide on metallic silver:

$$2\,Ag + BrCN + 3\,NaCN \longrightarrow 2\,Na[Ag(CN)_2] + NaBr$$

Coarse metallic silver particles and manganese-containing silver ores can also lead to complications.

In general, the leaching of silver ores proceeds more slowly than with gold ores. Also, the yield with gold is higher than that with silver. If only metallic silver and silver chloride are present the yield can reach 98 %, but with other ores it is sometimes only 80 %. The final silver content of the powdered ore after the extraction of silver can be as low as 0.5 ppm in favorable cases.

4.1.4. Thiosulfate Leaching (Patera Process)

The process of leaching silver ores with thiosulfate solution was developed originally in Europe under the name of the Patera process. In the 1800s its importance increased in Central and South America, and later in North America as well. It replaced the amalgamation process, mainly because of its higher yield and lower cost. However, in a few decades it was replaced by the more economical cyanide leaching process and is now seldom used. Halide-containing silver ores or chloridized roasted ores undergo the same reactions that form the basis of the photographic fixing:

$$AgCl + 2\,Na_2S_2O_3 \longrightarrow Na_3[Ag(S_2O_3)_2] + NaCl$$

and

$$AgCl + 3\,Na_2S_2O_3 \longrightarrow Na_5[Ag(S_2O_3)_3] + NaCl$$

An excess of sodium thiosulfate should be present to prevent formation of the sparingly soluble complex $Na[Ag(S_2O_3)]$. The sodium thiosulfate concentration is usually 2.5 – 25 g/L. In variations of the process, sodium copper thiosulfate or calcium thiosulfate is used. Silver is precipitated from solutions by the addition of sodium sulfide. Yields of silver in the thiosulfate process are lower than in the cyanide process. Processing techniques are similar. A disadvantage of the thiosulfate process is that gold is insoluble and therefore not recovered.

4.1.5. Metallurgical Processes

True silver ores are often added directly to copper or lead smelting operations. This is especially true if the accompanying gangue minerals are of use in the smelting process (e.g., quartz, which benefits the copper shaft furnace process).

4.2. Extraction from Lead and Lead – Zinc Ores [34]

4.2.1. Production of Lead Bullion

In all process stages in the extraction of lead from lead and lead – zinc ores, silver follows lead. This applies to ore flotation; ore sintering; reductive smelting in the blast furnace (see → Lead, Chap. 4.); removal of copper by sulfur (see → Lead, Chap. 5.1.1.); and removal of arsenic, antimony,and tin by selective oxidation (see → Lead, Chap. 5.1.2.).

The next stage is the extraction of silver and other noble metals. Historically, several processes have been used in succession, each one replacing its predecessor, namely, cupellation (oxidation), cupellation after the silver has been concentrated by the Pattinson process, and concentration by the Parkes process followed by zinc removal and cupellation.

4.2.2. Cupellation without Prior Silver Enrichment

Before processes for concentrating the silver in lead bullion were developed (i.e., up to the 1800s), all the lead had to be converted to litharge (PbO) by oxidation (cupellation), and then the litharge removed from the silver-enriched metallic phase. This yielded an impure product that was converted to contemporary commercial-grade silver by a refining operation, which consisted essentially of further oxidation in a small furnace. Here, the elements copper, selenium, and tellurium were removed by repeated addition of pure lead, and subsequent addition of silver sulfate removed bismuth and lead. Oxidizing agents such as sodium nitrate and sodium carbonate were also added, and removal of the last traces of litharge formed was assisted by the addition of fluxes (e.g., glass).

4.2.3. Silver Enrichment by the Pattinson Process

A significant technical advance in the process of silver concentration was achieved by PATTINSON (1833). After this discovery, silver-containing lead no longer had to be treated by the costly cupellation process because the crude lead could be separated into two fractions, one with a low silver content and the other containing 2 % Ag. The process was based on the fact that when molten silver-containing lead is cooled (e.g., 0.2 % silver), pure lead separates until the silver content in the liquid phase reaches 2.5 % (see phase diagram for Ag – Pb in Fig. 27). However, because of the difficulty of separating lead crystals from the silver-containing molten metal, complete separation cannot be achieved in a single operation. If the process is repeated, a grade of lead with a low silver and bismuth content ($\leq$10 ppm silver), and a grade of lead with a high silver content ($\leq$2 % silver) that is suitable for cupellation, can be produced, at an acceptable cost.

The fraction that contains silver and bismuth also contains gold. The theoretical silver content of 2.5 % is not achieved in practice, since the small temperature difference between the melting point of pure lead (327 °C) and that of the eutectic mixture (304 °C) means that separation becomes more difficult as the eutectic point is approached. The presence of copper, nickel, arsenic, antimony, and bismuth can also interfere with the removal of silver. Larger amounts of these elements therefore have to be removed first. Some copper goes into the Pb – Ag alloy, and some into the lead crystals, so that purified lead from the Pattinson process always contains some copper. Arsenic is found mainly in the lead crystals, whereas nickel and antimony follow bismuth into the silver-containing melt. In the usual operating procedure, either the lead crystals were removed by hand sieving, or the silver-containing melt was siphoned or tapped off. The fractions obtained were then treated again by crystallization. A series of cast iron vessels of 10 – 15 t capacity was employed, using countercurrent flow from the center to both sides. About 15 units were needed. Two-thirds of the contents of the vessel were recovered as crystals, and one-third remained as melt.

4.2.4. Silver Enrichment by the Parkes Process

The silver in bullion lead, together with gold and platinum metals, is concentrated by the Parkes process (→ Lead, Chap. 5.1.3.). The process involves the production of a solid alloy of lead, zinc, and silver by stirring zinc into molten impure lead. This so-called zinc crust floats on the surface of the molten metal (450 °C). It is skimmed off, the adhering liquid lead is removed by hot pressing, and the zinc is distilled in a retort or vacuum furnace. The residue consists of a lead – noble-metal alloy with up to 50 % silver, ca. 1 % zinc, and some copper, arsenic, antimony, and bismuth. It is oxidized by cupellation to obtain crude silver or, if appreciable amounts of gold are present,auriferous silver.

4.2.5. Cupellation of Enriched Lead

The cupellation process removes lead and small amounts of other base elements from the silver and accompanying noble metals by selective oxidation.

In the cupellation process, the lead melt is blown with air to produce molten lead oxide (litharge, PbO), while the silver, gold, and platinum-group metals do not react. The liquid layer of lead oxide produced covers the molten metal completely, but the oxidation reaction is not seriously impeded since the hot litharge takes up oxygen and then releases it to the molten metal. Copper, which is usually present, also assists the transport of oxygen. During cupellation, the temperature must be sufficiently high for the litharge (*mp* 886 °C) to form a low-viscosity melt that does not incorporate metallic and noble-metal components. At the end of the process, when the melt has a higher silver content and therefore a higher melting point, the operating temperature must be significantly greater than the melting point of silver (963 °C). In practice, the process is operated at 900 – 1100 °C. If the temperature is too high, volatilization can occur, which leads to losses of silver. The molten litharge formed must be removed continuously to prevent the layer from becoming too thick. The end of the smelting process is recognized when the litharge layer becomes thin and finally

breaks up. The bright surface of the silver can then be seen. The last traces of litharge are mixed with marl (calcium carbonate) to facilitate its complete removal. Molten silver is usually cast into anodes for electrorefining. The copper and lead contents of the crude silver should preferably not exceed 1.0 and 0.1 %, respectively. Of the normal impurities, zinc, arsenic, antimony, and copper go mainly into the litharge that is first produced, although some go into the slag that can be removed after the initial melting operation. Bismuth becomes concentrated in the end fraction of the litharge.

The so-called English cupellation furnace (cupel) is now nearly always used for the cupellation process in lead smelting plants. The hearth, which is replaceable, is lined with magnesite bricks, and can be tilted to pour off the litharge and molten silver. It is fired under a flat hood. Air is blown through nozzles over the surface of the melt, freeing part of the metal surface of molten litharge. The litharge layer is removed when it reaches a thickness of ca. 1 cm. About 100 kg of lead is oxidized per hour per square meter of furnace area; the capacity is ca. 1000–2000 kg silver. The so-called German cupel is larger and has a fixed hearth with a launder at the front that can be lowered continuously, so that as the volume of molten metal decreases, the litharge is continually removed. Molten silver must be skimmed off. The greatest advantage of the smaller English smelting furnace is that less silver is required for its operation. The lining of the hearth of the cupel picks up a certain amount of silver, so that after dismantling, the broken lining with litharge attached must be treated in a lead shaft furnace.

4.2.6. Silver Extraction from Electrolytic Lead Refining

The aqueous electrorefining of lead bullion (see → Lead, Chap. 5.2.) usually produces anode slimes that are difficult to process, whose composition is similar to that of the anode slimes from the electrorefining of copper [Cox] (Table 3). They contain all the noble metals and other impurities associated with crude lead. They are treated by pyrometallurgical processes to obtain auriferous silver and other valuable metals. Copper anode slimes and lead anode slimes are sometimes mixed and treated together.

4.3. Extraction from Copper and Copper–Nickel Ores [33,47–53]

4.3.1. Formation of Silver-Containing Copper Anode Slimes

In silver-bearing ores of copper, copper–nickel, and copper–nickel–cobalt, the silver stays with the copper-rich fractions during all concentration stages, including mechanical processing and smelting (see → Copper, → Nickel). It is transferred from the copper matte into the crude copper. Gold also follows the copper and silver, but only a small fraction of the platinum metals go to the crude copper, the majority accompanying the nickel. Today, all crude copper is electrorefined. The copper anode slimes formed contain almost all the silver and noble metals, and large amounts of copper, lead, nickel, selenium, tellurium, arsenic, antimony, and sulfur. Because of their composition, the economic and ecologically acceptable treatment of copper anode slimes is one of the most difficult tasks in the metallurgy of noble metals. Before the introduction of electrorefining, silver had to be extracted from copper matte produced by the copper shaft furnace (see Section 4.3.4).

4.3.2. Pretreatment of Copper Anode Slimes

Coarse copper particles that become detached from the anodes or appear as growths on the cathodes are removed first by sieving which eliminates ca. 30 % of the weight of the slime. In the following pretreatment process, elemental copper (formed by disproportionation of Cu^+ in the electrolyte), copper telluride, and copper selenide are dissolved by treatment with 4–5 mol/L sulfuric acid, with addition of air or oxygen at 60–95 °C. The effectiveness of this technique is limited, especially if the selenium content is high.

Examples of the composition of untreated and pretreated copper anode slimes are given in Table 2. The wide variation in the compositions

Table 2. Typical compositions [%] of untreated and pretreated copper anode slimes and untreated lead anode slime

Source	Ag	Au	Se	Te	As	Sb	Bi	Cu	Pb	Fe	Ni	SiO_2	Al_2O_3	S
Untreated copper anode slime														
Asarko, USA	18	1	9	1	4	4		21	10					5
Boliden, Sweden	11	2	21	1	1	1	1	40	10		1			3
Nordd. Affinerie, Hamburg	10	0.3	5			7		17	15					
Outokumpu, Finland	10	0.5	4		0.7	0.1		11	3	0.6	45	2		2
Mt. Lyell, Australia	1	0.2	3		0.7	0.1	0.01	67	1		0.1	1		10
El. Ref. Smelt. Co., Australia	8	2	3	3	4	8	0.1	14	24	0.4	0.5	10		8
Canadian Copper Ref., Montreal	19	1	20	3	0.6	0.5		37		0.6	0.2			
Noranda, Quebec	13	2	28	4	0.3	0.3		46		0.4	0.2			
Rio Tinto, Spain	8	0.7	8	0.5	2	3	0.5	25	10	0.3	0.1	20		
Palabora, South Africa	15	0.6	5	0.5				15	5		4	30	10	
Cerro, Oroya, Peru	28	0.1	2	2	2	11	1	2	24					
Pretreated copper anode slime														
Impex	49	4	3	6				3	1		0.1	12	20	
Codelco, Chile	24	1	16	0.3	0.3	5		2	1					
Nordd. Affinerie, Hamburg	13	0.3	5	1	3	8	0.6	3	15	0.2	3	7		6
Outokumpu, Finland	37	2	13		0.3	0.2		0.5	12	0.3	1	13		
El. Ref. Smelt. Co., Australia	12	3	0.7	4	1	10	0.1	2	28		0.4			5
Can. Copper Ref., Montreal	41	4	7	3				4						
Boliden, Sweden	29	5	20			4	2	2	30		0.7	2		
Enami, Chile	50	0.9	6	2	1	10	0.4	0.4	7			6		
Untreated lead anode slime														
Cerro, Oroya, Peru	10	0.01	0.1	0.7	5	33	21	2	16					

of the slimes both before and after treatment results in many process steps being carried out under very specific conditions that the processing industries often do not disclose.

4.3.3. Processing of Copper Anode Slimes [54,55]

Processing of Copper Anode Slimes by Roasting.

Oxidizing Roasting at 300–400 °C. Roasting is performed on a thin layer of material or in an air-blown rotary furnace. Metallic copper is oxidized fairly easily. The selenides and tellurides of copper and silver are converted to selenites and tellurites. Only a small fraction of selenium dioxide is volatilized. The product is treated with dilute sodium hydroxide solution, which dissolves mainly the selenium. Treatment with dilute sulfuric acid dissolves tellurium preferentially, and some silver as well. Most of the selenium and tellurium (75–90 %) can be brought into solution.

Oxidizing Roasting at 600–800 °C. The material is mixed with bentonite as a binder and with copper(II) oxide and iron(III) oxide as reaction accelerators, pelletized, and reacted with air blowing. Up to 98 % of the selenium can be recovered as SeO_2 sublimate. The product is then leached with dilute sulfuric acid, which dissolves copper, tellurium, and some silver.

Roasting with Sodium Carbonate. A mixture of anode slime with sodium carbonate (sometimes with the addition of sodium hydroxide or sodium nitrate) is generally pelletized and heated at ca. 400 °C in air, so that selenium and tellurium are converted into water-soluble compounds. Copper, which is also oxidized, and most of the tellurium can be dissolved with dilute sulfuric acid. The resulting solution also contains some silver in the form of its sulfate.

Processing of Copper Anode Slimes by Sulfation. Dried anode slime is mixed with concentrated sulfuric acid and heated, first to 300 °C, and then to 400–500 °C, causing sulfation and oxidation. Most of the selenium sublimes as selenium dioxide. Most of the copper, silver, and nickel can be dissolved with hot water, whereas the tellurium oxide formed remains undissolved in the residue.

Hydrometallurgical Treatment of Copper Anode Slimes [56]. Copper anode slimes can be dissolved in oxidizing acids. Treatment with $HCl-Cl_2$ [57] gives a residue that contains silver chloride, silica, and lead chloride. Gold, plat-

inum, selenium, and tellurium must be recovered from solution, and other accompanying elements are eliminated by wastewater treatment.

Copper anode slime is reportedly digested in a high-pressure process, either after the removal of copper or after one of the roasting processes [58]. Temperatures between 160 and 180 °C are recommended. If 10 – 40 % sodium hydroxide solution is used, selenium and tellurium go into solution, while silver and most of the copper remain undissolved. If dilute sulfuric acid is used in the presence of air, selenium, tellurium, silver, and copper dissolve. Problems arise if selenium and tellurium exceed an oxidation state of 4. Solutions of selenate(VI) and tellurate(VI) cannot be readily reduced to the elements. The favorable economics of the pressure process are mainly a consequence of the improved sharpness of separation compared to other processes and of the short reaction times.

Processing of Copper Anode Slimes in a Cupellation Furnace. If pretreated anode slimes contain only moderate amounts of copper, selenium, tellurium, antimony, and arsenic, the noble metals can be recovered in a cupellation process. However, a two-stage process in separate furnaces is necessary for treating the low-grade and high-grade materials.

Processing of Copper Anode Slimes in the Doré Furnace. In the Doré process, pretreated anode slimes (i.e., with copper removed) are treated with slag-forming and oxidizing materials for a long period in the molten state. In addition to the crude silver fraction, three types of slag are usually produced and removed separately. Flue dust is also obtained.

Sodium carbonate, lime, cullet (broken glass), and sand are added to form a silicate slag that takes up primarily iron, arsenic, antimony, lead, nickel, and tin. Selenium matte, which consists mainly of selenides and tellurides of copper and silver, forms an intermediate layer between the silicate slag and the molten metal if the starting material has a high selenium and tellurium content. After removal of the silicate layer, selenium matte is blown with air and treated with sodium carbonate and sodium nitrate. This produces a slag that contains sodium selenite, sodium selenate, sodium tellurite, sodium tellurate, and only ca. 1 % silver. This slag is tapped off, and the molten metal is reacted with additional sodium nitrate to remove the remaining copper, lead, selenium, and tellurium. The doré silver produced has a noble-metal content of 99.0 –99.5 % and is sent to the silver electrorefining process. Its main impurity is copper, but it also contains some lead, bismuth, selenium, and (depending on the starting material) tellurium. Doré furnaces are usually oil fired along their length, and are similar to German fixed hearth or English tilting hearth cupels. For hearth sizes between 3 and 6 m^2, a charge of 8 – 15 t anode slime is treated in 100 – 200-kg lots. The reaction time is 3 – 6 d or longer. Waste gases are passed through dedusting chambers, coolers, scrubbers, and electrostatic gas purifying equipment.

Slags produced by oxidation of selenium matte, and flue dust and slimes from the scrubbers are major sources of selenium (→ Selenium and Selenium Compounds) and tellurium (→ Tellurium and Tellurium Compounds). The most valuable material in the wash liquor is selenium. All other drosses, including residues from selenium and tellurium production, which contain noble metals, are recycled to the copper shaft furnace or to the doré process.

Typical compositions of intermediate products are given in Table 3. These can, however, vary over a wide range.

Treatment of Copper Anode Slimes with Molten Salts. Powdered anode slimes can be digested below their melting point, without forming molten metal by fusing with Na_2CO_3 and $NaNO_3$ or similar mixtures. The reaction is also possible at sintering temperature. After leaching with water, a separation process must be used similar to that used after acidic hydrometallurgical digestion of copper anode slimes.

Other Methods of Processing Copper Anode Slimes. Dried, decopperized anode slimes can be treated with chlorine at ca. 700 °C. Silver chloride remains in the residue, and the chlorides of selenium, tellurium, copper, tin, arsenic, etc., are sublimed or distilled.

If the decopperized anode slimes are heated to 700 °C in a vacuum, silver selenide decomposes and selenium sublimes.

Table 3. Percentage compositions of intermediate products from copper anode slimes and copper matte

	Ag	Au	Se	Te	As	Sb	Bi	Cu	Pb	Ni	SiO_2
Cupellation											
After melting	1		1		3	10		4	40	5	10
Tapped off selenium matte	20	0.1	20						25		
Tapped off litharge	1		0.5			2	3	10	60		
Argentiferous lead	40	0.5	1					5			
Auriferous silver	97	2									
Doré Process											
Selenium matte	55	0.1	25	4				13	1		
Selenite – tellurite slag	1		18	6				2			
Potassium nitrate slag	4		6	1				16			
Flue dust	4	0.3	30	3		10		5			
Gas scrubber sediment	10	0.2	35	2				0.5			
Crude metal (before treatment with $NaNO_3$)	79	4	4	1				4	1		
Doré metal	95	4	<0.1				<0.1	1	0.1		
Ziervogel Process											
Mansfeld copper matte (before Ziervogel process)	0.5							75			
Mansfeld copper matte (after Ziervogel process)	0.02										

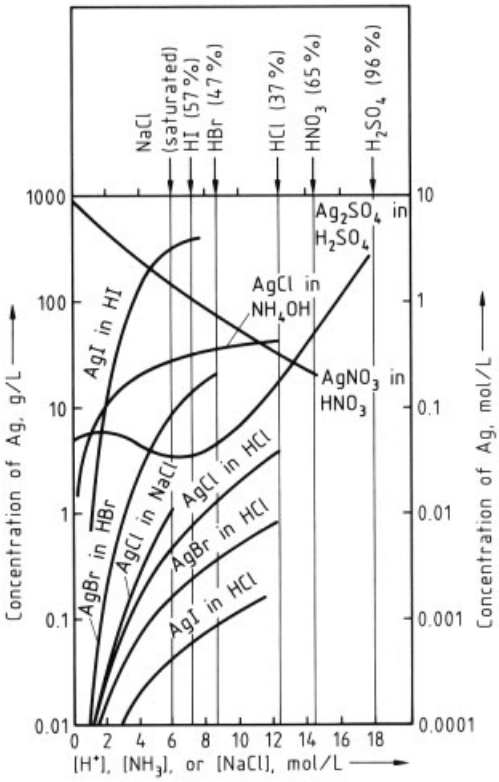

Figure 4. Solubilities of silver compounds in aqueous electrolytes at 20 °C

4.3.4. Silver Extraction from Copper Matte

Before electrolytic technology became available for refining copper, copper – silver separation had to be performed on copper matte from the copper shaft furnace. Two main processes were used.

In the Augustin process, copper matte was roasted under chloridizing conditions, and the silver chloride formed was leached out with hot brine, leaving behind the unchanged copper sulfide (Fig. 4).

In the mid-1800s, the Augustin process was replaced by the Ziervogel process, which involved sulfation roasting of powdered copper matte with a precisely controlled temperature program. This first decomposes silver sulfide to silver and sulfur. The silver reacts with sulfur trioxide liberated from iron sulfate and copper sulfate, forming the more stable silver sulfate, which is water soluble. The copper is then in the form of copper(II) oxide. Treatment with water acidified with a little sulfuric acid dissolves silver sulfate, leaving copper(II) oxide in the residue. More than 90 % of the silver goes into solution, 5 % is in the flue dust, and the rest stays in the copper oxide (see Table 3). Gold is undissolved and remains in the copper. Thus, the gold content of copper matte cannot be recovered by the Ziervogel process.

4.4. Extraction from Gold Ores

The silver present in most gold ores follows the gold in all pretreatment processes, as well as in the aqueous cyanide process. Crude gold as mined usually contains 10 – 12 % silver. In the commonly used Miller process, silver and base metals are separated from gold by conversion to their chlorides (AgCl, CuCl, etc.). The reaction is carried out by passing chlorine through the molten metal (see → Gold, Gold Alloys, and Gold Compounds, Chap. 5.2.). Silver chloride and the base-metal chlorides form a slay on the molten gold to which borax is added as a flux. Solidified slag is leached with a solution of sodium chlorate in hydrochloric acid, which dissolves the base-metal chlorides and leaves silver chloride as residue. The silver chloride is reduced with zinc dust in suspension to precipitate crude silver, which is then electrorefined (see Section 6.4).

Small amounts of silver and gold are also recovered from sodium carbonate-containing slag and broken crucible fragments from the melting of precipitated gold in the cyanidation process. These materials are treated by the lead – silver blast furnace process and the cupellation process (see Section 5.3).

4.5. Extraction from Tin Ores

Silver-containing concentrates from the treatment of Bolivian tin ores usually contain a few hundred parts per million of silver, present mainly as sulfidic compounds. The ore is roasted under oxidizing conditions in the presence of sodium chloride, to convert silver to silver chloride, which can then be leached out. Alternatively, the concentrates can be treated by an amalgamation process, with or without a preliminary roasting operation (see Section 4.1.2). A more economical solution is often to add these materials to the feed for the copper smelting process. Silver then ends up in anode slimes from the electrorefining of copper.

Leaching with Hydrochloric Acid or Aqueous Sodium Chloride. Leaching with HCl is carried out at ca. 100 °C. Tin dioxide is not dissolved, but most of the other materials are, including silver. Silver can be cemented with scrap iron. The cementate produced is treated in copper or lead smelting processes.

Leaching with Aqueous Thiosulfate. When the silver content is high, aqueous thiosulfate is often used because of the high and selective solubility of silver chloride in thiosulfate. The process was formerly used widely to treat true silver ores (see Section 4.1).

5. Recovery from Secondary Silver

[3, 20, 37–42, 56, 57]

The collection and treatment of silver-containing scrap and residues is of major economic importance. These materials can be separated into metallic products that can be melted to form a homogeneous material and nonmetallic concentrates (drosses, see Section 3.4). The amount of silver in this field corresponds to about 20 – 40 % of primary silver production. As in the production of silver from ores, the aim is to obtain a crude silver that can be treated in the refining process. Only seldom is good-delivery silver obtained that does not need refining.

5.1. Via Copper Smelters

In principle, all silver-containing recycled materials can be treated in the usual copper production processes (→ Copper ; see also Section 4.3). Since the silver is tied up in the process for a long period, only a few types of silver-containing scrap are treated in this way. Copper must be the principal base metal present. This category of scrap includes fabrication waste and recycled material from the electronics industry. Usually, they would be fed to the copper shaft furnace or the cupel. Flue dust and copper matte from the lead – silver blast furnace are also treated in copper smelters.

5.2. Via Lead Smelters

The only use of lead smelters in silver recycling is in treating the silver-containing flue dusts from lead – silver blast furnaces. These smelters are very suitable for dealing with the high lead and zinc content of flue dusts.

5.3. Via the Lead – Silver Smelting Process [23, 24, 58]

Some companies that recover noble metals from scrap use a special lead-based smelting process. The pyrometallurgical reactions are the same as those used for noble-metal extraction in ancient times and correspond in principle to those of the modern lead smelting process. There are two main stages of the process: the blast furnace stage and the oxidation stage. In the blast furnace, silver and the other noble metals are dissolved in the lead formed in the process. In the oxidation furnace (converter), the lead is then oxidized to lead oxide (PbO, litharge), which can be easily separated from the metal melt. The main product is a metal with a high percentage of silver (auriferous silver). Lead oxide is recycled by the blast furnace process. The accompanying elements that enter the system in the raw material are eliminated via three byproducts of the blast furnace: copper matte, slag, and flue dust. Slag, which contains the oxide and silicate components, is worthless and can be discarded. Copper matte, which contains the chalcophilic elements copper and nickel, and flue dust, which contains zinc and cadmium, must be treated separately.

Nonmetallic concentrates (drosses) containing noble metals are generally smelted in the lead – silver blast furnace. Metallic scrap materials are often treated directly in the converter. In this case, base metal oxides go into the litharge, though some are in the flue dust. The choice of treatment process is often determined by the cost of tying up the value of the noble metal, thereby incurring loss of interest, rather than by actual operating costs.

Typical compositions of materials involved in the lead – silver smelting process are given in Table 4.

The Blast Furnace. In the blast furnace (Fig. 5), the following reactions take place:

$$C + 1/2\,O_2 \longrightarrow CO$$

$$PbO + CO \longrightarrow Pb + CO_2$$

Here, lead is formed from the litharge produced in the converter and acts as a collector for noble metals. The chemical reactions, which also form the copper matte and slag phases, take place at 1100 – 1300 °C.

In the operation of the lead – silver blast furnace a large number of parameters must be controlled and various boundary conditions observed. The widely varying composition of the raw materials and of the recycled litharge, must be compensated by varying the amount of quartz, limestone, sulfur, iron oxide, pyrites, and coke added. This ensures the formation of a low-viscosity slag that does not incorporate droplets of metal. The amount of sulfur added is calculated from the amount of nonferrous metal. The ability of the slag to take up aluminum oxide is limited. If the Al_2O_3 content reaches 20 %, the slag viscosity increases. This must be taken into account when catalysts supported on γ- or α-aluminum oxide are processed.

The ability of lead and auriferous silver to alloy with gold, platinum, and palladium is unlimited. However, rhodium, iridium, ruthenium, and osmium, together with other refractory metals such as tungsten and molybdenum, collect at the bottom of the blast furnace or at the bottom of the converter. For this reason, as well as the volatility of ruthenium and osmium compounds in the blast furnace, the presence of these metals is undesirable. Halides in the raw materials end up almost completely in the flue dust in the form of lead and silver halides. Zinc and cadmium take the same route.

The elements arsenic, selenium, and tellurium are kept out of noble-metal smelters as much as possible because they form compounds with noble metals (speisses) that are difficult to process.

Instead of using a blast furnace, the lead smelting stage can be operated in a flame-fired furnace. This, too, is followed by an oxidation stage.

Oxidation in the Converter. The following reaction takes place in a converter:

$$Pb + 1/2\,O_2 \longrightarrow PbO$$

The molten lead oxide (litharge) produced is removed continuously from the surface of the liquid metal and recycled to the blast furnace. Silver and other noble metals remain in the elemental state and are concentrated from ca. 20 % in

Table 4. Typical compositions of materials produced during noble-metal smelting

Source	Material			Nonnoble impurities, %												Relative quantities, t
		Ag	Au	Cu	Pb	S	Halogen	Fe	Ni	Zn	Cd	Sn	SiO_2	Al_2O_3	CaO	
Blast furnace process	slag	100 ppm	2 ppm	0.5	1			15		2			35	20	15	30
	copper matte	3 %	100 ppm	40	10	25		15	1							3
	lead bullion	20 %	1 %	15	60	1		2	2	4		3				6
Cupellation process	litharge	4 %	100 ppm	15	60			2	2			3				6
	auriferous silver	97 %	1 %	1	0.1											1
Gas purification	flue dust	1 %	100 ppm	3	30	2	25	2	1	8	3					1

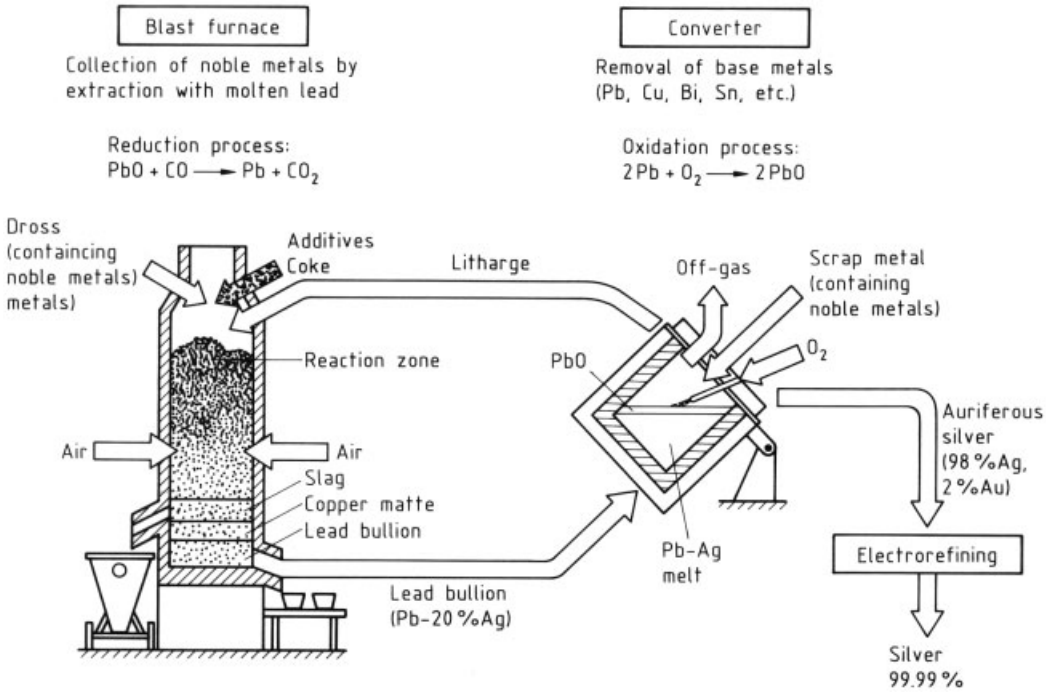

Figure 5. Silver recovery in a blast furnace and converter

the starting material (lead bullion) to ca. 99 % in auriferous silver (Table 4). The lead oxide contains silver particle inclusions, considerable quantities of copper, and small amounts of other accompanying base metals as their oxides (Table 4).

The oxidation of argentiferous lead was carried out for a long time exclusively in shallow flame-fired furnaces (cupels) like those used in the lead smelting process. Lead was oxidized by applying a flame and a stream of air to the surface of the melt. The reaction temperatures during cupellation were ca. 900 °C at the start and ca. 1100 °C at the end of the process. The lower limit is determined by the melting temperature of PbO (890 °C), silver (961 °C), and auriferous silver (ca. 1050 °C).

Since ca. 1975, rotary noble-metal converters have been used instead of cupels. The oxidizing agent is gaseous oxygen [59]. Such converters contain several tonnes of metal, and the litharge is poured off through a spout. A natural gas – oxygen burner produces very intense heating. The great economic advantages compared to

conventional cupellation hearths result from the much higher rate of reaction due to convection in the melt, low energy requirements, improved working conditions, longer lifetime of the furnace lining and the potential for automation.

If scrap alloys are processed directly in the converter and not in the shaft furnace, to speed up noble-metal throughput, the litharge picks up more base metal, mainly copper. The copper can be removed in copper matte only after the litharge has been transferred to the blast furnace. Therefore the noble metal is sometimes concentrated by a preliminary treatment of the raw material (see Sections 5.4 and 5.5).

Both the oxidation process with direct feeding of scrap metal and the shaft furnace process can be carried out with copper instead of silver. This is appropriate if the raw material has a high content of platinum-group metals that do not readily form alloys with silver (e.g., rhodium). The copper oxide formed is then usually processed by leaching with sulfuric acid.

5.4. Via Scrap Metal Leaching [60]

Alloys based on silver – copper, which can also contain nickel, zinc, other base metals, gold, and platinum-group metals, can be treated hydrometallurgically to recover the noble and base metals.

The basis for this is the very wide miscibility gap between silver and copper in the alloy system (see Fig. 15). Copper can be selectively dissolved by dilute sulfuric acid, with addition of air, gaseous oxygen, or hydrogen peroxide:

$$Cu + CuSO_4 \longrightarrow Cu_2SO_4$$

$$Cu_2SO_4 + H_2SO_4 + 1/2\,O_2 \longrightarrow 2\,CuSO_4 + H_2O$$

Copper can also be oxidized in a prior roasting stage. Dissolution by sulfuric acid is then possible without addition of oxidizing agents. The preferred reaction temperature is ca. 90 °C. In the concentration range Ag 30 Cu 70 to Ag 70 Cu 30, the silver in the alloy retains its structure during the process. The reaction rate decreases nonlinearly with increasing depth of attack due to the restriction of diffusion through the silver skeleton. For acceptable times of solution, the thickness of the alloy pieces should not exceed 0.1 – 0.2 mm. This can be achieved by machining the alloy to form chippings or by spraying the molten metal to form a powder. In both cases, a material is obtained that can easily be kept in suspension by stirring. If the solidified metal is slowly cooled to ca. 200 °C, formation of mixed crystals on the silver-rich side of the system can be inhibited (see Fig. 15), which decreases the rate of copper dissolution. The silver that remains is generally electrorefined. Copper sulfate or electrolytic copper can be recovered from solution.

Silver-containing alloys can be dissolved in nitric acid or hydrochloric acid – chlorine, and silver can be precipitated as silver chloride. However, this is not economical on an industrial scale. Copper nitrate can be converted to copper oxide or copper(I) chloride. Copper cementation is also useful if it is connected with a converter process.

5.5. Via Scrap Metal Electrolysis [61]

Dietzel Electrolysis (see Fig. 6) [62]. A recycled alloy, usually of silver, gold, and copper, is anodically dissolved in an electrolyte of slightly acidified copper nitrate. The anode and cathode spaces are separated by a textile diaphragm. The anolyte is removed continuously via an overflow, and silver is precipitated from it by cementation with copper. The desilvered anolyte is fed continuously into the cathode space, where copper is deposited on rotating roller-shaped cathodes and is removed in sheet form from time to time. The flow of desilvered solution into the cathode space, and the outflow of silver-containing solution from the anode space, which is caused hydrostatically, produce a flow of solution through the diaphragm in the direction of the anode. Very little dissolved silver reaches the cathode space. Copper with a low silver content is recovered at the cathode.

The process is still used industrially. One disadvantage is the high cost of equipment. Several variations of the process have been described [63].

Demag Electrolysis [63]. Copper can be dissolved from silver – copper anodes by a dilute sulfuric acid electrolyte and deposited on the cathode, while silver remains undissolved. If the

metal has an average silver content, as is necessary for the sulfuric acid leaching of scrap silver (see Section 5.4), a felt-like mass of silver is produced that retains the shape of the original anode. With increasing depth of penetration into the anode, the dissolution reaction is inhibited by the silver structure, which presents diffusion. Therefore, the thickness of the anode sheet must not exceed ca. 5 mm. If anode baskets are used, avoiding the production of zones from which the copper has not been completely removed is even more difficult.

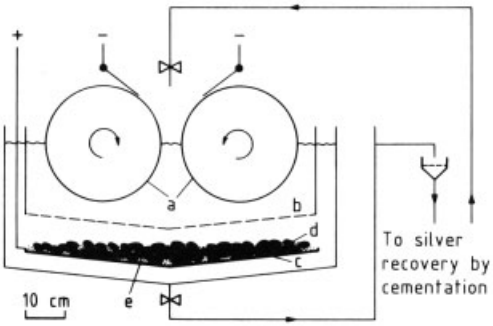

Figure 6. Dietzel electrolysis (front view)
a) Roller-shaped cathodes; b) Diaphragm; c) Anode; d) Pieces of scrap silver; e) Anode slime

As dissolution proceeds, the current – voltage characteristic of the electrolytic cell changes. Potentiostatic control of electrolysis is therefore essential to prevent excessive voltage, which would cause silver to dissolve and be deposited with copper on the cathode. As in the sulfuric acid dissolution process, preliminary thermal treatment must be carried out to prevent excessive amounts of copper from appearing in the silver crystals due to the formation of silver – copper mixed crystals. Also, the copper-containing electrolyte must be carefully washed out of the silver felt, which contains very fine pores.

Electrolysis in sulfuric acid also removes the other metals that usually accompany copper and silver, but nickel accumulates in the electrolyte where it causes problems. If zinc, cadmium, lead, and tin are present in the recycled alloy, it should be oxidized first in a converter. Tin, especially, can interfere with electrolysis by precipitation of stannic acid.

5.6. Processing of Flue Dust

Flue dust from the waste gases of the lead – silver smelting process nearly always contains silver. Flue dust can be recovered quantitatively by means of settling chambers, electrostatic processes, wet and dry filtration, and scrubbing towers.

Dust from the lead – silver shaft furnace and the cupel generally contains the following: lead(II) oxide, lead sulfide, silver sulfide, elementary silver, copper oxide, copper sulfide, slag components, etc. It can normally be recycled to the shaft furnace.

If halides, mainly from photographic materials, are introduced into the smelting process, they are essentially bound as halides of lead and silver, which because of their low boiling points (*bp* of $PbCl_2$ = 916 °C), are found almost entirely in the flue dusts. If the flue dust is recycled to the shaft furnace the halides revolatilize, and only the other components are reduced. The halide content of flue dust increases due to the recirculating system, and when it exceeds a certain figure, the dust is treated with sodium carbonate solution or milk of lime. Soluble halides are disposed with wastewater. All heavy metals, nonferrous metals, and noble metals remain as oxides, carbonates, or metals in the residue, which is fed back to the shaft furnace process.

Like the halides, the elements zinc and cadmium become concentrated in the flue dust, passing unchanged through the reducing shaft furnace due to their vapor pressure if they have not already been removed in the slag during cupellation or by distillation from the feed material. After sufficient concentration, they are dissolved as sulfates by treatment with sulfuric acid. The main component of the residue, which is fed back to the shaft furnace, is lead sulfate. The distribution of silver between the solution and the residue must be monitored.

The arsenic, antimony, selenium, and tellurium in flue dusts are especially troublesome. Noble-metal smelters attempt to remove these elements, which are readily sublimable in the elemental state, before the flue dust is introduced into the recycling system. Recycling of dusts of this type is usually confined to copper and lead smelters, where the elements can be picked up in the slag, either in the shaft furnace or the doré furnace.

5.7. Processing of Copper Matte [64]

Because of its high silver content, copper mate (see also Section 4.3.4) from the lead – silver shaft furnace used in noble-metal smelting must be treated without delay to minimize interest losses. Silver can be recovered quickly by the lead shaft furnace process, followed by isolation of silver by the Parkes process (see Section 4.2.4). Treatment in the shaft furnace of a copper refinery is possible, but silver is bound for a long time in the copper electrorefining stage and then in the copper anode slimes. In principle, copper matte can be ground, roasted in air, and leached with dilute sulfuric acid to dissolve the copper and leave a silver-containing residue. The same result is obtained by directly leaching ground copper matte with dilute sulfuric acid at ca. 160 °C in autoclaves, with introduction of air. Direct electrolysis of copper matte anodes has also been proposed. Copper is deposited on the cathode, whereas silver remains in the anode slime. However, mechanically stable anodes are difficult to produce.

5.8. Processing of Photographic Materials [40,42,64]

Films. Film material (1 – 3 % Ag) in which the film base is nonuniform [usually poly(ethylene terephthalate), PET, or cellulose triacetate, CTA], and therefore difficult to reuse, is best treated by incineration, generally after size reduction in a cutting mill. The combustion temperature must be controlled carefully to minimize volatilization of the silver, which can lead to high losses in the combustion gases, and consequent high recycling costs. The durability of the furnace lining is also strongly dependent on temperature. Usual incineration temperatures are in the range of 1000 – 1200 °C. The temperature can be lowered by water spraying [65].

If pyrolysis [66] is used rather than incineration, flue dust problems are largely avoided. The end products of incineration and pyrolysis can be treated by smelting or electrorefining, depending on the silver content. In pyrolysis, temperatures of 600 – 700 °C are used. The silver content in the remaining ashes ranges from 15 to 35 %, depending on the process. Typical concentrations of silver in the flue dust are in the region of 20 %.

If the film substrate is to be reused, a washing process is employed that destroys the gelatin coating and liberates silver or silver halide. Most important for the washing process is the kind of film carrier material. Triacetate film, mainly applied in amateur photography, is relatively easy to process. PET material from graphic and X-ray applications is more critical because of the need for complete removal of the adhesion promoter. Suitable washing solutions include hot dilute sulfuric acid, hot dilute sodium hydroxide solution, sodium hypochlorite solution, or enzyme solutions [67]. If developed films are to be treated, the grains of silver can be oxidized with a solution of iron (III) chloride, into which air is injected continously for reoxidation of iron (II), and the silver halide formed can be washed off with a solution of sodium thiosulfate. The silver concentrates, which must be reduced to silver if silver halides are present, generally have a sufficiently high silver content to be treated by electrorefining.

Older films, such as those from archives, must be tested carefully to determine whether they are made of extremely flammable nitrocellulose film material. The storage of such materials requires elaborate safety precautions.

Photographic Paper. Washing processes are uneconomic because the silver content of photographic papers is low and the substrate materials cannot be reused after washing. Instead, silver is usually recovered by incineration or pyrolysis.

Photographic Emulsions. The photographic industry produces unusable, usually pasty or lumpy residues of photographic emulsions, which are supplied to recovery plants for treatment. Apart from silver halides, they contain water and considerable amounts of gelatin. Treatment can be by combustion, pyrolysis, or chemical breakdown of the gelatin in solution or suspension. The silver halides can be reduced at the same time or in a subsequent operation. The isolated silver halides can be converted into metallic silver by fusion with sodium carbonate or by hydrometallurgical reduction.

Fixing Baths [68–75]. Silver-containing fixing baths from photographic plants and laboratories, hospital X-ray departments, and doctors' surgeries are rarely sent directly to recovery

plants. They are generally treated chemically or electrochemically in situ to precipitate the silver. Alternatively, they may be purchased by specialist companies that convert then to concentrates and detoxify the highly contaminated wastewater.

In times of high silver prices, ca. 75 % of all fixing baths are treated to recover the silver in spite of their wide dispersal. In addition to the economic value of the silver recovered from fixing baths, ecological considerations are important.

In concentrations of $\geq$0.04 ppm, silver ions have harmful effects on biological systems. However, these effects are much reduced if the silver in photographic wastewater is in the form of a complex. The potentially harmful effects on biological wastewater treatment systems are prevented by rapid reaction of dissolved silver with the sulfide that is always present in such systems forming insoluble and, therefore nontoxic, silver sulfide. Although no harmful effects on biological sewage treatment plants have been observed in practice, the silver concentration in wastewater passing into the main drainage system is usually limited to 1 – 2 ppm.

Many processes have been used to treat silver-containing fixing baths. Chemical precipitation processes have largely been displaced by electrolytic processes for ecological reasons. The original method of precipitating silver sulfide from the fixing bath by adding sodium sulfide is obsolete and unacceptable for reasons of industrial hygiene. Precipitated silver sulfide was often reacted with iron powder above the melting point of silver to form metallic silver and iron sulfide. Later, silver was precipitated by adding trimercapto-*s*-triazine [73]. This reaction takes place without liberation of hydrogen sulfide and can be followed potentiometrically. The method is also suitable for treating the much more dilute fixing and bleaching baths from color photography, for rinsing solutions, and for wastewater that contains copper, cadmium, mercury, nickel, and lead. Final concentrations of ca. 1 mg/L silver are obtained. The cementation of silver with zinc powder is dangerous due to the hydrogen evolved and has led to severe accidents. It also produces zinc-containing wastewater. The use of sodium borohydride as reducing agent has been suggested, although this also results in evolution of hydrogen. The use of sodium dithionite as a reducing agent is recommended in the photographic industry:

$$2\,Na_3[Ag(S_2O_3)_2] + Na_2S_2O_4 + 4\,NaOH \longrightarrow 2\,Ag + 2\,Na_2SO_3 + 4\,Na_2S_2O_3 + 2\,H_2O$$

Most fixing bath treatment companies now use a process in which silver is recovered quantitatively by electrolysis. Fairly pure silver (90 – 99.5 %) is removed from the stainless steel cathodes in the form of sheets and budlike growths. In large laboratories and hospital X-ray departments, part of the silver is recovered continuously in electrolytic cells, so that the fixing bath can be used for a longer period and no longer need be discarded so frequently and replaced with new material. For smaller laboratories and doctors' surgeries, the recovery of silver from fixing baths is carried out by using containers packed with steel wool [74]. The spent fixing bath solution is run through slowly, whereby silver is reduced and precipitates on the steel wool. The suppliers take back the containers with the silver and send them to recovery plants.

Electrolysis equipment [75, 76] of various sizes is also rented out by many business establishments who undertake the refining and purchase of the silver produced. This system has become important mainly in European countries except Germany [77].

The complex-forming thiosulfate in the fixing bath can be decomposed by sodium hypochlorite or chlorine and neutralized with sodium hydroxide solution:

$$Na_2S_2O_3 + 4\,Cl_2 + 10\,NaOH \longrightarrow 2\,Na_2SO_4 + 8\,NaCl + 5\,H_2O$$

The silver chloride formed can be filtered off and reduced. In practice, oxidation with hydrogen peroxide is often preferred [78]. The advantages are ease and safety in handling, and the avoidance of contamination with toxic chlorinated organic substances. The reactions are

$$2\,Na_2S_2O_3 + H_2O_2 \longrightarrow Na_2S_4O_6 + 2\,NaOH$$

and

$$Na_2S_4O_6 + 7\,H_2O_2 + 6\,NaOH \longrightarrow 4\,Na_2SO_4 + 10\,H_2O$$

which can be summarized as

$$Na_2S_2O_3 + 4\,H_2O_2 + 2\,NaOH \longrightarrow 2\,Na_2SO_4 + 5\,H_2O$$

This can be taken to either the tetrathionate or the sulfate stage, depending on the desired BOD value of the wastewater produced. The silver precipitates as a mixture of Ag_2O, AgBr, and Ag_2S.

In low-silver fixing baths (e.g., washing baths or fixing baths from which silver has been removed electrolytically) the small amounts of residual silver can be removed by ion exchange [79–81].

5.9. Surface Desilvering

In silver-plated material for jewelry, the base metal is usually German silver, a copper–nickel–zinc alloy. The base materials of silver-plated electronic components are generally copper–iron, iron–nickel–cobalt alloys, or brass. In addition they contain tin as a brazing alloy component and gold as a contact material. In electrical low-current contacts, carrier materials for the silver coating can be copper, copper–tin, copper–beryllium, brass, German silver, or copper–nickel–tin.

The processing of these materials in the lead–silver smelting process (see Section 4.2) is uneconomical. The large amounts of copper result in formation of additional copper matte, which can often incorporate more silver than is introduced in the scrap material. Nickel and iron lead to the formation of undesired furnace bottoms. Silver-plated materials are better treated by the copper-smelting route (see Section 4.3), which allows recovery of copper, nickel, silver, and gold in high yield. A disadvantage is that the noble metal is tied up for a long period.

In the dissolution of silver coatings from metallic substrates by acid treatment, both silver and the base metal are attacked, especially in the reaction of nitric acid with German silver. A mixture of concentrated sulfuric acid and concentrated nitric acid gives better results. Silver-plated items may also be reacted with hot sodium polysulfide solution, which rapidly and selectively converts the silver to silver sulfide. The sulfide coating can then be removed by mechanical or thermal shock, or by reduction with aluminum or reducing agents, e.g.:

$$Ag_2S + Na_2S_2O_4 + 4\,NaOH \longrightarrow 2\,Ag + 2\,Na_2SO_3 + Na_2S + 2\,H_2O$$

The dissolution of silver coatings in sodium cyanide solution in the presence of oxygen is slow and therefore suitable only for very thin films (a few micrometers), such as those on electronic products. As with gold, higher rates of dissolution are achieved if water-soluble aromatic nitro compounds (e.g., nitrobenzenesulfonic acid or nitrobenzoic acid) are used as oxidizing agent instead of air [82]. When the base material consists of copper or copper alloys, unwanted side reactions of the type

$$Cu + 4\,NaCN + H_2O \longrightarrow Na_3[Cu(CN)_4] + NaOH + 1/2\,H_2$$

are difficult to avoid. Silver can be obtained from the resulting solution by electrolysis or by decomposition of the cyanide with chlorine.

5.10. Processing of Special Scrap

Silver residues for recovery often arise in combination with other valuable materials. In these cases, conventional processes may fail, either because the accompanying substances cause problems, or because they cannot be recovered. In many cases (e.g., Ag–α-Al_2O_3, Ag–W, Ag–Ta), silver can be dissolved by nitric acid without attacking the accompanying material. More reactive metals such as nickel, tin, cadmium, or indium can generally be dissolved – even if they are alloyed with silver – by using acid, acid mixtures, or alkali,or by anodic dissolution, leaving silver as the residue.

Before dental amalgam wastes are treated chemically to recover silver, they are usually heated to $>500\,°C$ in closed equipment with strict control of the surrounding atmosphere to distill the mercury.

6. Silver Refining [2,3,20,56,61,83]

Crude silver, from ores in primary production and from recycling operations generally has a purity of 98–99.5 %, which may include gold and platinum-group metals because these processes only separate noble metals from base metals.

The main task of silver refining (world output: 10 000 t/a) is the production of silver with

a purity of >99.9 % and >99.97 %. A purity of >99.99 % is increasingly being specified for industrial applications. The first and most important task is the recovery of gold and platinum metals from crude silver. The value of the gold in doré silver and auriferous silver often exceeds that of the silver itself. The refining of silver therefore generally includes a gold – silver separation process. Before refining, the noble metals are in very concentrated form, so the process must involve the shortest possible tying-up of the noble metal.

6.1. Fine Smelting

If the crude silver contains hardly any gold or platinum-group metals, silver of fineness 999/1000 can be obtained by an oxidation process, generally carried out in a separate refining furnace. The purification process is improved by adding alkali-metal nitrates, alkali-metal carbonates, or silver sulfate, and by blowing with oxygen.

6.2. Refining with Nitric Acid (Inquartation) [3]

The long-known process of dissolving a silver – gold alloy in nitric acid is still used today, although to a minor extent. Silver dissolves according to the equation

$$3\,Ag + 4\,HNO_3 \longrightarrow 3\,AgNO_3 + 2\,H_2O + NO$$

(see Fig. 2), and the gold remains undissolved. The separation becomes increasingly efficient as the composition of the alloy approaches Ag 75 Au25. Ag – Cu – Au alloys can also be treated by this process. In addition to silver, the base-metal components and considerable amounts of the platinum metals are dissolved. Silver nitrate can be isolated from the solution (see Section 7.1); metallic silver can be obtained by cementation and chemical reduction or electrolytic reduction; or silver chloride can be precipitated. Historically, the process was temporarily replaced by the sulfuric acid separation process, because of the high consumption of nitric acid and the large amount of recycled noble metal. Its advantage is rapid recovery of the gold. Hence alloys with high gold content (e.g., 10 – 20 %) are treated by this process.

6.3. Refining with Sulfuric Acid (Affination) [3]

Silver – gold alloys are treated with ca. 86 % sulfuric acid at 240 °C, whereby the silver is oxidized according to the equation

$$2\,Ag + 2\,H_2SO_4 \longrightarrow Ag_2SO_4 + 2\,H_2O + SO_2$$

and remains in solution as the more soluble silver hydrogensulfate. The solution is diluted with water to precipitate silver sulfate, which is filtered off and treated with iron in a cementation process. The silver sulfate still in solution is treated by cementation with copper to produce copper sulfate. This refining process can deal with a wider range of alloy compositions than the nitric acid process (Section 6.2). The alloys treated may contain appreciable amounts of copper. Also, the consumption of acid is much lower than in the nitric acid process. On the other hand, this refining process requires more costly equipment and results in considerable off-gas problems. The sulfuric acid process has not been used in large-scale applications since World War II.

6.4. Möbius Electrolysis [23, 24]

The Möbius electrolysis process for refining silver is used mainly in Europe (Fig. 7). A silver nitrate – sodium nitrate – nitric acid electrolyte is employed. Silver and the accompanying base metals are dissolved by anodic oxidation, and the concentration and current – voltage ratios are controlled so that only silver is deposited at the cathode. Cast crude silver anodes are suspended and surrounded by anode bags in which anode slime collects. This contains all the gold in the form of metal, selenide, or telluride, as well as platinum-group metals, silver particles that become detached, lead dioxide, and oxidic copper compounds. Dendritic silver crystals are deposited on the stainless steel cathode sheet and removed continuously by automatic scrapers to prevent short circuits. Copper accumulates in the electrolyte. If the concentration of copper becomes too high, it can deposit with the silver, and the purity specification may not be met. The

same is true of selenium. Palladium goes partly into solution but precipitates on the cathode to a very minor extent.

The silver concentration in the electrolyte is ca. 50 g/L, with 10 g/L of free HNO_3, and the anodic current density is 400 – 500 A/m^2. The cell voltage is 2.0 – 2.5 V. The anodes and cathodes are usually arranged in parallel in the cell, and the energy consumption is ca. 0.6 kW · h/kg of pure silver. To compensate for the anodically dissolved base metals that are not removed at the cathode (mainly copper), silver nitrate solution is fed into the electrolyte to prevent depletion of silver.

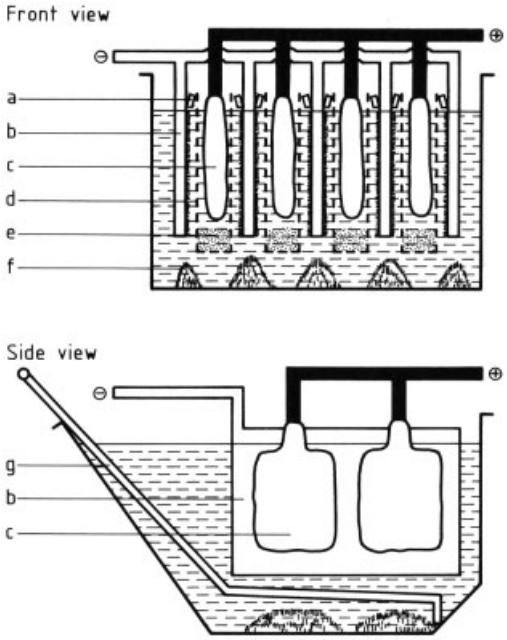

Figure 7. Silver electrorefining by the Möbius process
a) Scraper; b) Cathode (stainless steel); c) Anode (crude silver); d) Anode bag; e) Anode slime (gold, platinum-group metals); f) Silver crystals; g) Scraper

At one time, the Möbius process had to be interrupted to remove crystals of fine silver. The anode and cathode carrying device was lifted from the bath, and the perforated false bottom was raised to enable the silver crystals to be collected. Today, cells are usually constructed so that the silver crystals can be scraped manually over the edge of the vessel without interrupting electrolysis (Fig. 7). The spent anodes are replaced continually by new anodes, so that interruption of electrolysis is necessary only if the electrolyte must be replaced due to excessive copper concentration, or if the anode bags have to be emptied. The electrolyte is usually treated with copper to recover the silver by cementation, and the copper is then recovered by cementation with iron. The electrolyte can also be treated by chemical precipitation methods or by the "black melt" process in which a mixture of silver nitrate and copper nitrate is obtained by evaporating to dryness (see Section 7.1). The main accompanying metal, copper, is reported to be removed by liquid – liquid extraction, in which case the remaining silver-containing solution could then be returned to the system [84, 86, 85].

The Möbius electrolysis process is most suitable for crude silver containing >90 % Ag. With high copper content, the electrolyte must be regenerated too often, and with high gold content, the anode bags require frequent emptying. Both are costly operations. A single electrolysis gives a purity of 99.95 – 99.99 % Ag and a double electrolysis gives a purity of 99.995 – 99.999 %.

Current developments in silver electrolysis are aimed at fully automated high-speed cells which are able to process higher gold and PGM contents. Possibilities are the application of high Au contents (up to 25 – 30 %) so that a Au/PGM scaffold remains from the anode, or the use of alloy grains. In the latter case the process works at higher currents (same current density). Feeding of silver grains and removal of anode-slime can be automated by special anode-cell constructions.

6.5. Balbach – Thum Electrolysis [23, 24]

The Balbach – Thum electrolysis process uses virtually the same electrochemical conditions as the Möbius process. Only the geometrical arrangement is different (see Table 5 and Figure 8). The anodes and cathodes are arranged horizontally. Each cell is provided with only one anode and one cathode. The cathode is a stainless steel plate covering the entire floor area. Above this, at a distance of 10 cm, is a slightly smaller framework of nonconducting material, covered with fabric and containing the anodes. The anode and cathode are supplied with electric cur-

rent via silver contacts. The crude silver anodes are completely covered by electrolyte, and therefore completely dissolved, unlike the anodes of the Möbius process, which leave behind some unconsumed metal. Another advantage of Balbach – Thum electrolysis is that the anode slime is recovered more easily, which is especially important when the crude silver has a high gold content. Disadvantages include the large space requirement and high energy consumption. The Balbach – Thum process is used mainly in the United States.

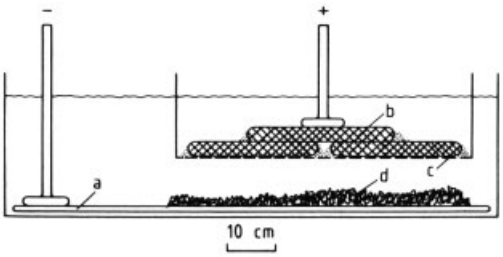

Figure 8. Balbach – Thum electrorefining
a) Cathode; b) Silver anode; c) Diaphragm; d) Silver crystals

7. Silver Compounds [20, 87–89]

7.1. Silver Nitrate [2]

Properties. Silver nitrate [*7761-88-8*], M_r 169.87, theoretical silver content 63.50 %, colorless rhombic crystals, *mp* 209 °C, temperature of transformation to the trigonal form 160 °C, decomposition temperature 444 °C, density 4.352 g/cm^3, is very soluble in water, the solubility being strongly dependent on temperature (see Table 5 and Fig. 9), with complete miscibility in the $AgNO_3 - H_2O$ system at >159 °C. Its solubility in water is decreased considerably by the presence of nitric acid (see Fig. 2). The solubility in ethanol is 20.8 g/L, in methanol 35 g/L, and in benzene 2.2 g/L.

Silver nitrate has an oxidizing action on organic materials, which can lead to ignition and combustion, sometimes with explosion. Silver nitrate is decomposed even by traces of dust, with formation of finely divided, black silver. The reaction is accelerated by light. Silver nitrate and its solutions are unstable toward light over long periods. Complete drying after production leeds to instable material. Traces of acid on the surface can prevent decomposition for longer terms.

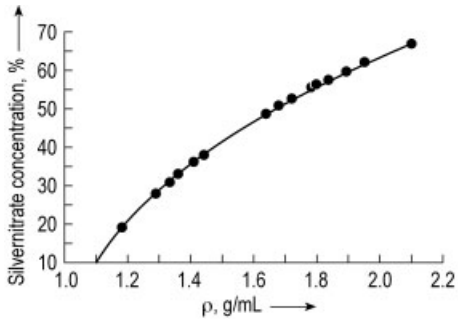

Figure 9. Correlation between concentration and density in aqueous silver nitrate solution at 20 °C

Production. Silver nitrate is generally produced by dissolving pure silver in hot 32 % nitric acid:

$$4\,Ag + 6\,HNO_3 \longrightarrow 4\,AgNO_3 + NO + NO_2 + 3\,H_2O$$

If all the nitrogen dioxide is recycled by reaction with the water in the system to form nitric acid (see below), the overall reaction is

$$3\,Ag + 4\,HNO_3 \longrightarrow 3\,AgNO_3 + NO + 2\,H_2O$$

In practice, the mass balance lies somewhere between these two equations. A slight excess of nitric acid is required, which can be recovered when the product is crystallized by evaporation. Many producers pass oxygen or air through the reaction mixture, giving improved recovery of nitric acid:

$$4\,Ag + 4\,HNO_3 + O_2 \longrightarrow 4\,AgNO_3 + 2\,H_2O$$

However, the formation of nitrogen oxides can be largely suppressed by using a closed vessel, so that off-gas treatment can be avoided. The nitrogen oxides produced were formerly thermally decomposed at >600 °C in a flame, to produce nitrogen and oxygen, but energy costs are high. An alternative method is to use the reactions

$$2\,NO + O_2 \longrightarrow 2\,NO_2$$

and

$$3\,NO_2 + H_2O \longrightarrow 2\,HNO_3 + NO$$

Table 5. Saturated solutions of silver nitrate in water

Temperature, °C	$AgNO_3$, g per 100 ml H_2O	$AgNO_3$, g/L of solution	$AgNO_3$ molarity	$AgNO_3$, wt %	Density, g/cm^3
−7	94	773	4.6	48.5	1.60
0	115	909	5.4	53.5	1.70
10	160	1170	6.9	61.5	1.90
20	215	1438	8.5	66.6	2.11
25	241	1551	9.1	70.7	2.19
30	282	1710	10.0	73.8	2.32
40	335	1892	11.1	77.0	2.46
50	400	2083	12.3	80.0	2.60
60	471	2261	13.3	82.5	2.74
80	652	2608	15.4	86.7	3.01
90	762	2768	16.3	88.4	3.13
100	1024	3052	18.0	91.1	3.35
110	1105	3120	18.4	91.7	3.40
125	1624	3430	20.2	94.2	3.64
133	1941	3552	20.9	95.1	3.74
> 159	∞				

or the direct reaction

$$2\,NO_2 + H_2O_2 \longrightarrow 2\,HNO_3$$

These processes are best carried out in countercurrent gas scrubbing equipment. The nitric acid formed is recycled to the process.

Since silver nitrate for the photographic industry is of a purity higher than that of refined silver, purification is required. Typical silver nitrate quality for photographic industry (in ppm) is given below:

Pb	≤ 0.2
Cu	≤ 0.2
Fe	≤ 0.2
Cd	≤ 0.2
Hg	≤ 0.01
Cr	≤ 0.2
Mn	≤ 0.2
Ni	≤ 0.2
Al	≤ 0.5
Sb	≤ 0.1
Bi	≤ 0.1
Se	≤ 0.1
Sn	≤ 0.1
Zn	≤ 0.1
K	≤ 5
Na	≤ 5
Ca	≤ 1
Pt	≤ 0.001
Pd	≤ 0.005
Au	≤ 0.005
Rh	≤ 0.001
Ir	≤ 0.001
Ru	≤ 0.001
Sulfate	≤ 20
Nitrite	≤ 10
Chloride	≤ 5

The classical crystallization process is in principle very effective, but due to the high solubility of silver nitrate even in the cold (see Table 5) and in the presence of HNO_3 (see Fig. 2), the amount recycled is high. Nevertheless, consequtive crystallization steps are the main purification process in the photographic industry. Precipitation and adsorption of copper, iron, lead, tin, etc., as their oxide-hydrates by the addition of silver(I) oxide are recommended. This is added to the weakly acid aqueous solution until the pH reaches 6. Chromatographic purification methods based on γ-Al_2O_3 have also been described [90, 91]. Small traces of compounds of the platinum-group metals can lead to undesirable photochemical sensitization and can be converted into the inactive elemental form by irradiation with UV light [92]. Final purification can be carried out by heating molten anhydrous silver nitrate to ca. 350 °C, which mainly removes copper [20, 36]. At this temperature, silver nitrate does not decompose to the nitrite, but the decomposition temperature of most other heavy-metal nitrates to their oxides is exceeded. The copper oxide formed gives a dark-colored melt (black melt process). Precipitated impurities are removed by dissolving the product in water and filtering. The purified solution is then concentrated, and the silver nitrate is crystallized, washed carefully, and dried.

In the production of silver nitrate on an industrial scale, the equipment must be closed so that no loss of silver nitrate solution, aerosols, or dusts occurs. All operations can be carried out in stainless steel equipment.

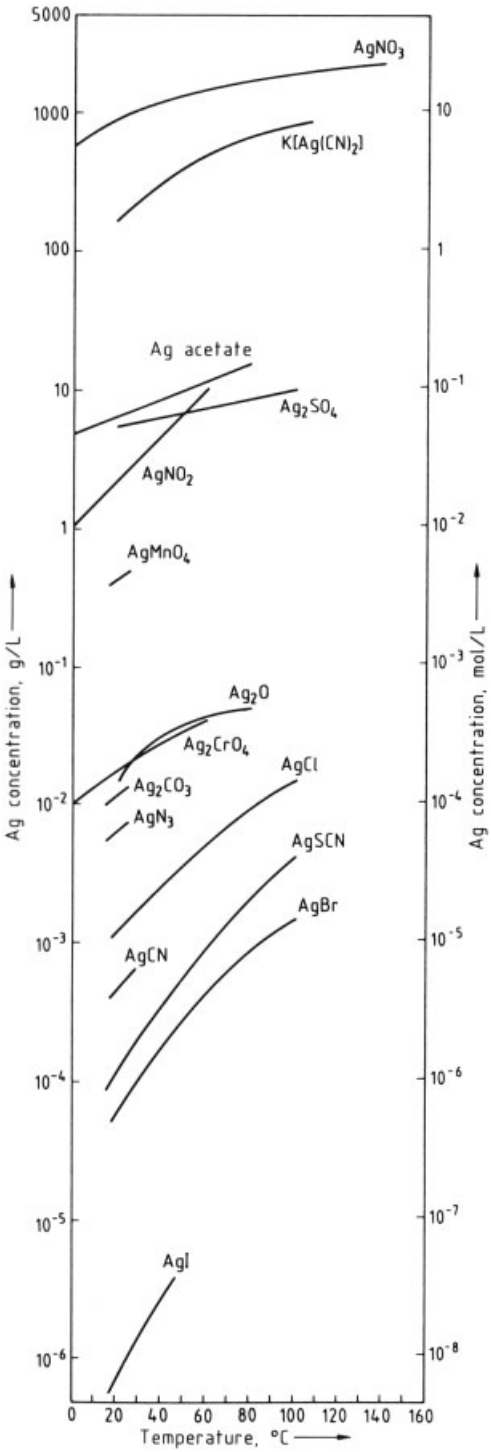

Figure 10. Solubilities of silver compounds in water as a function of temperature

The silver may be recovered from the mother liquor by cementation on copper, or the mixed nitrates can be treated by the black melt process to separate silver and copper. Alternatively, silver may be precipitated as silver chloride.

Uses. Most of the silver nitrate produced commercially is used in the photographic industry (see Section 10.10, and → Photography). Therefore, the quality specification for this application is the one in general use. Silver nitrate is also the starting material for the production of most other silver compounds and preparations, e.g., potassium dicyanoargentate(I) (see page 34), silver oxide (see Section 7.3), and supported catalysts. The process of electroless silver plating is also based on silver nitrate (see Section 10.5). Approximately 33 % of the worldwide annual consumption of industrial silver (ca. 22 000 t) is in the form of silver nitrate.

7.2. Silver Halides

Silver chloride [*7783-90-6*], AgCl; silver bromide [*7785-23-1*], AgBr; and silver iodide [*7783-96-2*], AgI (see Table 6 and Fig. 10) resemble one another in properties. Silver astatide (AgAt) is a member of this series, although it is likely to be sensitive to oxidation, even by atmospheric oxygen [93].

Silver halides become increasingly insoluble in the series AgCl, AgBr, AgI. Their solubility can be reduced further by addition of the respective common ion in low concentration. Higher concentrations of the common ion cause a marked increase in solubility due to the formation of complexes of the type $[AgX_2]^-$ (see Fig. 4). Soluble complexes are also formed with thiosulfate, cyanide, and ammonia (see Section 7.6). Solubility in organic solvents increases with increasingly nonpolar character of the chemical bond from chloride to iodide.

All three compounds are sensitive to light and are used in photographic coatings on film and paper (see Section 10.10 and → Photography). Other industrial applications of silver halides are much less important.

Silver halides are produced almost exclusively by precipitation from aqueous silver nitrate solution by the addition of solutions of

Table 6. Properties of the sparingly soluble halides and pseudohalides of silver

Property	AgCl	AgBr	AgI	AgCN [a]	Silver isocyanate Ag−N=C=O [a]	Silver thiocyanate Ag−S−C≡N [a]	Silver fulminate Ag−C≡N−O [a]	Silver azide Ag−N−N≡N [a]
M_r	143.32	187.78	234.77	133.89	149.89	165.95	149.89	149.89
Ag content, %	79.26	57.44	45.95	80.56	71.97	60.91	71.97	71.97
Crystal system [b]	cubic	cubic	hexagonal (< 150 °C) cubic (> 150 °C)	hexagonal	monoclinic	monoclinic	rhombic	rhombic
Density, g/cm^3	5.56	6.473	5.683 6.010	3.95	4.00	3.746	4.09	4.09
mp, °C	455	430	557	350 (decomp.)	335 (decomp.)	224 (decomp.)	explosive	explosive
bp, °C	1547	1533	1506					
Solubility at 20 °C, mg Ag/L								
Water	1.3	0.08	0.001	0.2		0.1	500	60
HMPT [c]	0.12	0.08				20		10
Methanol	0.03	0.003	0.0001			0.01		0.3
Liquid SO_2	3300	25	100	200		130		

[a] The isomeric compounds corresponding to the tautomeric formulas H−N≡C, H−O−C≡N, H−N=C=S, and H−O−N≡C, are not known in all cases, or the isomerism is not well defined in the solid state; the nomenclature used in the literature is contradictory [20,98,100].

[b] High-temperature and high-pressure modifications also generally exist.

[c] Hexamethylphosphoric triamide.

alkali-metal halides. Synthesis from the elements is also possible [94–96]. In the temperature range between the melting points of the metal and the halide, silver chloride, in particular, can be produced readily and economically.

7.3. Silver Oxides [2]

Silver(I) oxide [*20667-12-3*], Ag_2O, M_r 231.74, theoretical silver content 93.10 %, silver content of commercial product 93.0 %, ϱ 7.143 g/cm^3, is a dark brown powder with cubic crystal structure. It decomposes slowly at ca. 150 °C and rapidly at 300 °C.

Silver(I) oxide is formed when alkali-metal hydroxides are added to solutions of silver salts such as silver nitrate. Silver hydroxide cannot be isolated as an intermediate product (but can be detected visually as a white pricipitate before it turns dark brown), although an aqueous suspension of the sparingly soluble silver(I) oxide in water has a marked alkaline reaction. In electrolytes containing sodium hydroxide or in neutral electrolytes that do not produce sparingly soluble silver salts, silver(I) oxide can be produced by anodic oxidation of metallic silver.

Silver(I) oxide is used in considerable quantities (i.e., several hundred tons per year) in batteries (see Section 10.5). In organic synthesis, it is used to replace halogen atoms with OH groups.

Silver(II) oxide [*1301-96-8*], AgO, M_r 123.87, theoretical silver content 87.08 %, AgO content of commercial product 97.5 %, is now known to have the composition $Ag^{I}Ag^{III}O_2$. It is a gray-black powder, with a monoclinic crystal structure, which decomposes slowly to the elements above ca. 85 °C and rapidly above 100 °C. Silver(I) oxide has not been detected as an intermediate of this reaction.

Silver(II) oxide is usually prepared by the oxidation of monovalent or elemental silver by peroxodisulfate [97]. Other suitable oxidizing agents are potassium permanganate, ozone, sodium hypochlorite, or an electric current. Silver(II) oxide is a strong oxidizing agent. In contact with organic compounds, including solvents, spontaneous reaction can occur. It is increasingly used in voltaic cells (see Section 10.5), where the amount of electricity produced per mole of silver(II) oxide reacted is twice that produced per mole of silver(I) oxide, leading to economies in weight and space. The lower silver consumption means it is often more economical to use silver(II) oxide than silver(I) oxide, despite higher production costs.

7.4. Other Soluble Silver Compounds

Silver sulfate [*10294-26-5*], Ag_2SO_4, M_r 311.82, silver content 34.60 %, *mp* 657 °C, ϱ 5.45 g/cm^3, is a white, crystalline salt with rhombic structure. It is fairly soluble in water (see Fig. 10) and more soluble in sulfuric acid (see Fig. 2), forming $AgHSO_4$, which can be isolated. Silver sulfate is produced by the reaction of silver with hot, concentrated sulfuric acid.

$$2\,Ag + 2\,H_2SO_4 \longrightarrow Ag_2SO_4 + SO_2 + 2\,H_2O$$

It can also be prepared by precipitation with sulfuric acid from $AgNO_3$ solution, followed by recrystallization from dilute sulfuric acid.

Silver permanganate [*7783-98-4*], $AgMnO_4$, M_r 226.81, silver content 47.60 %, is a black-violet salt that is moderately soluble in water (see Fig. 10). It can be used as a disinfectant and oxidizing agent. It is produced as the least soluble component in the system

$$AgNO_3 + KMnO_4 \longrightarrow AgMnO_4 + KNO_3$$

or by the reaction of silver(I) oxide or silver carbonate with a solution of permanganic acid. It tends to decompose if it is completely dry.

Silver(I) fluoride [*7775-41-9*], AgF, M_r 126.88, silver content 87.79 %, *mp* 434 °C, ϱ 5.852 g/cm^3, is a white, flaky, crystalline solid with cubic crystal structure. It is hygroscopic and soluble in water. With increasing temperature, the solid phases $AgF \cdot 4\,H_2O$ and $AgF \cdot 2\,H_2O$ exist in equilibrium with the saturated solutions.

Silver fluoride is prepared by the reaction of silver(I) oxide or silver carbonate with hydrofluoric acid, or by dissolving finely divided precipitated silver powder in hydrofluoric acid and hydrogen peroxide.

Silver fluoride is of minor industrial importance but is used in preparative fluorine chemistry.

Disilver fluoride, Ag_2F, is water sensitive, and silver(II) fluoride, AgF_2, is highly reactive.

Silver tetrafluoroborate [*14104-20-2*], $Ag[BF_4]$, is used in aqueous solution as an absorbent for ethylene. It can be produced in benzene

$$AgF + BF_3 \longrightarrow Ag[BF_4]$$

or in aqueous solution

$$Ag_2O + 2\,H[BF_4] \longrightarrow 2\,Ag[BF_4] + H_2O$$

The compound is very hygroscopic and unstable to light.

Oxy Salts of Halogens. Silver chlorate [*7783-92-8*], $AgClO_3$, is moderately soluble in water, and silver perchlorate [*66259-09-4*], $AgClO_4$, is very soluble.

Soluble Organic Silver Compounds. Silver(I) acetate [*563-63-3*] (see Fig. 10), silver(I) propionate [*5489-14-5*], and silver(I) lactate [*15768-18-0*] are fairly soluble in water.

Silver(I) behenate is used in the Dry-Star copy process (Agfa).

7.5. Other Insoluble Silver Compounds [98]

Silver sulfide [*21548-73-2*], Ag_2S, M_r 247.80 %, *mp* 845 °C, ϱ 7.326 g/cm^3, silver content 87.06, forms black crystals or powders with monoclinic crystal structure (as argentite, cubic), and is the least soluble silver compound (7×10^{-14} g/L at 20 °C). It is not dissolved by the ligands ammonia and thiosulfate, but is dissolved by cyanide. Silver sulfide can be dissolved by oxidative reactions (e.g., in hot, concentrated nitric acid). It can be reduced by base metals or by dithionite in alkaline solution. The latter is used to remove the dark coating from tarnished silver.

Silver sulfide is prepared from the elements at 250 – 400 °C or by precipitation of silver by sulfide ions.

Silver carbonate [*534-16-7*], Ag_2CO_3, M_r 275.75, ϱ 6.077 g/cm^3, is a white powder that rapidly turns yellow. It has a monoclinic crystal structure and decomposes at 218 °C, liberating carbon dioxide. It is prepared from soluble silver compounds by precipitation with alkali-metal carbonate and is used in considerable amounts in the production of silver powder.

Silver isocyanate, AgNCO, M_r 149.89, can be precipitated from aqueous solution. A mainly covalent Ag – N bond is present in AgNCO. Silver cyanate, AgOCN, apparently does not exist. Unlike the isomeric silver fulminate, AgCNO (see Section 7.7), silver isocyanate is thermally stable. The nomenclature for cyanates and isocyanates is, however, not yet uniform [20, 98–100].

Silver cyanide [*506-64-9*], AgCN, M_r 133.89, *mp* 350 °C, ϱ 4.62 g/cm^3, is a white powder with a trigonal crystal structure. It is prepared from soluble silver salts by precipitation with cyanide anions. This compound is the precursor in the production of potassium dicyanoargentate(I). Silver cyanide used to be important in analytical chemistry because of its low solubility in water, and the ease and rapidity with which it can be converted into the dicyanoargentate complex.

Silver thiocyanate [*1701-93-5*], AgSCN, M_r 165.95, ϱ 3.746 g/cm^3, is a white powder with a monoclinic crystal structure that decomposes on melting. It is produced by precipitation from aqueous solution, a reaction that is also used in quantitative analysis (see Chap. 11).

7.6. Silver Complexes

Potassium dicyanoargentate(I), $K[Ag(CN)_2]$, M_r 199.01, silver content 54.20 %, ϱ 2.364 g/cm^3, forms white crystals with trigonal structure, which are very soluble in water (see Fig. 10) and fairly soluble in ethanol. It is prepared by dissolving silver cyanide in an aqueous solution of KCN, concentrating by evaporation, and crystallizing. It can also be obtained electrolytically by the anodic oxidation of silver in aqueous potassium cyanide solution. This method is analogous to that for the preparation of $K[Au(CN)_2]$ (see → Gold, Gold Alloys, and Gold Compounds, Chap. 7.1.). Potassium dicyanoargentate is the basis of cyanide plating baths (see Section 10.5) and is used extensively in the jewelry industry.

The corresponding sodium salt has similar properties and is an intermediate in the extraction of silver from ores by the cyanide process (see Chap. 4).

Thiosulfate Complexes. Several complexes are known, for example,

$$Na[Ag(S_2O_3)] = Na_2S_2O_3 \cdot Ag_2S_2O_3$$
$$Na_4[Ag_2(S_2O_3)_3] = 2\,Na_2S_2O_3 \cdot Ag_2S_2O_3$$
$$Na_3[Ag(S_2O_3)_2] = 3\,Na_2S_2O_3 \cdot Ag_2S_2O_3$$
$$Na_5[Ag(S_2O_3)_3] = 5\,Na_2S_2O_3 \cdot Ag_2S_2O_3$$

All these take part in the photographic fixing process (see Section 10.10) but have no technical importance in their own right. Their solubility in water increases as the ratio of $S_2O_3^{2-}$ to silver in the complex increases.

Diammine Complexes. The cationic complex $[Ag(NH_3)_2]^+$ is formed when most silver compounds are dissolved in aqueous ammonia. The solution is unstable (especially when produced at pH > 10, e.g., by addition of NH_4OH solution to silver(I) oxide) and can precipitate highly explosive fulminating silver (see Section 7.7).

7.7. Explosive Silver Compounds [101]

Many unstable silver compounds are known whose decomposition can occur spontaneously and explosively. This reaction is usually initiated by heat, pressure, impact, or abrasion. These unstable silver compounds are all light sensitive. Hence, high-energy flashes of light can also occasionally cause explosive decomposition. Such compounds are no longer used to any extent, or are used only for very special preparative work, partly because of the near impossibility of storing them safely. Nevertheless, they can be formed in many chemical reactions and processes and, if their existence and properties are unrecognized, can result in accidents.

Silver azide [*13863-88-2*], AgN_3, M_r 149.89, silver content 71.97 %, has colorless, needle-like crystals (*mp* 252 °C). It is an extremely explosive compound, sensitive to heat, impact, and pressure.

Silver azide is insoluble in water and can be obtained by precipitation from solutions of sodium azide and silver nitrate. It can be used to prepare the equally explosive iodine azide (IN_3):

$$AgN_3 + I_2 \longrightarrow IN_3 + AgI$$

It is less stable than lead azide and has therefore not been used as a detonator.

Fulminating silver forms black crystals that consist of a mixture of silver nitride, Ag_3N, and silver imide, Ag_2NH. It is formed when aqueous ammoniacal solutions of silver compounds are stored, and by the action of ammonia on silver oxide or silver carbonate.

An explosion can be caused by slight movement of a vessel containing an aqueous solution in which a small amount of fulminating silver has precipitated and collected at the bottom. Aqueous solutions containing $[Ag(NH_3)_2]^+$ should therefore never be stored.

Silver amide [*65235-79-2*], $AgNH_2$, is highly explosive. It is formed mainly by the reaction of silver compounds with liquid ammonia. Silver fittings or equipment should therefore not be used in contact with liquid ammonia.

Silver fulminate [*5610-59-3*], AgCNO, is the silver salt of fulminic acid, HCNO [98, 102], and is sometimes also referred to as fulminating silver (see above). It forms white needles that explode under mechanical or thermal stress, and is considerably less stable than mercury fulminate.

Silver acetylide [*7659-31-6*], Ag_2C_2, is an amorphous compound that decomposes explosively on impact or heating. On treatment with alkali-metal hydroxides, hydrolytic decomposition takes place.

Silver acetylide is prepared by passing acetylene into an ammoniacal solution of silver salts. If acetylene is passed into a neutral solution of silver nitrate, crystalline addition compounds, such as $Ag_2C_2 \cdot AgNO_3$ or $Ag_2C_2 \cdot 6\,AgNO_3$, are precipitated. Explosive silver acetylide is also formed when silver compounds come in contact with an acetylene flame. The existence of this hazard is the reason acetylene flames are not used for the excitation of silver in solution for spectral analysis. Apparatus and containers for acetylene should not include fittings or components made of silver or high-silver alloys. Conventional silver brazing alloys are safe.

Silver oxalate [*533-51-7*], $(COOAg)_2$, M_r 303.76, is a white compound that is insoluble in water and can be prepared by precipitation. It is temperature sensitive and decomposes on gradual heating, but detonates if heated rapidly.

8. Disperse Silver [18,20,58,102]

8.1. Silver Powder

Silver powder is classified as microcrystalline, spheroidal, or lamellar (flakes and bronzes), according to its particle shape. All three types, which differ in method of manufacture, have specific areas of use. The basic material is pure silver or combinations of silver and palladium.

For various applications, the following properties are important: bulk density, tapped density, and tamped density (see Table 7), sinterability, flow properties, densities achievable by compression and sintering, specific surface area, and content of impurities such as Na^+, K^+, and Cu^{2+}. These properties depend ultimately on particle shape and size, and on particle-size distribution. The size of the primary particles is not identical to the particle size measured by particle-size analysis, especially for dendritic powders in which agglomerates often occur. Earlier quality specifications were based mainly on sieve analysis, which has now largely been replaced by laser granulometry. The latter gives a value for particle size such that a defined fraction of the material (e.g., 90 %) is finer than the quoted particle size. Methods for particle-size analysis are being supplemented more and more by optical and scanning electron microscopy.

Microcrystalline Silver Powder. Several basic types of microcrystalline silver powder are available, differing in method of preparation, type of particle, and properties (see Table 8). They were used almost entirely in sintering technology for a long time. Later, they were also used in stoving preparations (e.g., metallization of ceramics) and in pastes for repairing heated rear windows of motor vehicles.

The most important method of production is precipitation by reducing agents, such as formaldehyde, potassium sodium tartrate, and hydrazine, from solutions of silver nitrate in dilute sodium hydroxide solution. Alternatively, Ag_2O or Ag_2CO_3 suspensions are sometimes reduced with these reducing agents.

Cementation with less noble metals is no longer of practical importance because of contamination of silver with the precipitating metal and wastewater problems.

Electrolytic reduction of silver nitrate solutions under conditions used in Möbius electrolysis (see Section 6.4) leads to a coarsely crystalline product that is not generally regarded as a silver powder and is marketed as "crystal silver." However, by using special methods of scraping or washing the cathode and varying the conditions, the electrolysis of silver nitrate can be made to yield a finely powdered material suitable for sintering applications. Electrolysis of an acidic medium has the advantage that base metals are not usually deposited with the silver. This allows less pure starting materials to be used [103]. Cathodic reduction of semiconducting silver(I) oxide in a medium containing sodium hydroxide leads to finely divided silver powder that is very similar to the chemically precipitated product. Silver oxide can be obtained by precipitation or by oxidation of a silver anode and deposition of the oxide on the cathode. Also, thermal decomposition of silver(I) oxide or silver carbonate gives a silver powder with a relatively small specific surface area.

Spheroidal Silver Powder. Atomization of molten silver by compressed air, inert gas, a water jet, or a rapidly rotating knife produces spheroidal silver powder, which is used mainly in sintering technology. It is characterized by relatively high bulk density, low specific surface area, and good flow properties.

Flakes and Bronzes. Lamellar particles are obtained by ball milling microcrystalline silver powder. Flakes and bronzes are incorporated into paints and adhesives, where they impart electrical conductivity to the polymer matrix due to the good particle-to-particle electrical contact.

Silver – palladium powder is used in the production of conducting electrodes in ceramic multilayer capacitors. Palladium improves the migration resistance. Several types of powder are produced by different methods, including

Table 7. Silver powder (physical properties, old system)

Origin	Production method	Particle shape	Bulk density, g/cm	Specific surface area (BET), m^2/g	Sinterability	Flow properties	Sieve analysis
Degussa	chemical precipitation	crystalline (isometric)	ca. 1.2	ca. 1.0	good	poor	< 32 µm ca. 90 %
							32 – 40 µm ca. 10 %
							40 – 60 µm ca. 1 %
							60 – 75 µm 0.1 %
							75 – 100 µm ca. 0.1 %
							> 100 µm ca. 0.1 %
	electrolysis of acidic electrolytes	crystalline (isometric to dendritic)	ca. 1.6	ca. 0.8	good	fair	< 32 µm ca. 70 %
							32 – 40 µm ca. 20 %
							40 – 60 µm ca. 10 %
							60 – 75 µm ca. 1 %
							75 – 100 µm ca. 0 %
							> 100 µm ca. 0 %
	cementation	crystalline (dendritic)	ca. 1.5	ca. 0.3	good	fair	< 32 µm ca. 50 %
							32 – 40 µm ca. 20 %
							40 – 60 µm ca. 20 %
							60 – 75 µm ca. 5 %
							75 – 100 µm ca. 5 %
							> 100 µm ca. 3 %
	atomization	spheroidal	ca. 4.2	ca. 0.06	poor	good	< 32 µm ca. 50 %
							32 – 40 µm ca. 30 %
							40 – 60 µm ca. 20 %
							60 – 75 µm ca. 1 %
							75 – 100 µm ca. 1 %
							> 100 µm ca. 0 %

Table 8. Silver powder and flakes (physical properties, new system)

Producer	Product no.	Particle shape	Tapped density, g/cm^3 *	BET surface area, m^2/g	Particle size,** µm	Impurities, ppm		
						Na^+	K^+	Cu^+
Demetron	309	microcrystalline	1.2 – 1.6	0.5 – 0.9	< 19	< 100	< 10	< 500
	328	microcrystalline	1.0 – 1.6	1.2 – 1.7	< 19	< 100	< 10	< 10
	331	microcrystalline	2.8 – 3.8	0.7 – 1.2	< 10	< 10	< 10	< 10
	F 14	flakes	2.1 – 3.2	0.7 – 1.1	< 32	< 20	< 10	< 500
	D 12	flakes	2.5 – 3.1	1.0 – 1.5	< 20	< 30	< 10	< 500
	D 25	flakes	3.0 – 4.0	1.2 – 1.8	< 16	< 150	< 10	< 10
	D 27	flakes	3.3 – 4.3	1.0 – 1.8	< 16	< 100	< 10	< 10
	D 35	flakes	3.0 – 4.0	1.0 – 1.5	< 18	< 10	< 10	< 10
Metz	Q-100	microcrystalline	0.9 – 2.0	0.55 – 1.05	< 18	< 25	< 10	< 20
	C-200 ED	microcrystalline	2.6 – 4.2	0.5 – 0.9	< 8	< 50	< 30	< 30
	EG 200	microcrystalline	1.5 – 3.0	0.15 – 0.4	< 13	< 125	< 10	< 30
	9	flakes	2.6 – 3.5	0.9 – 1.2	< 11	< 150	< 200	< 350
	15	flakes	2.8 – 4.0	0.6 – 1.2	< 10	< 50	< 10	< 30
	25	flakes	3.35 – 4.3	0.35 – 0.75	< 22	< 30	< 30	< 30
	52	flakes	3.6 – 4.65	0.4 – 0.7	< 14	< 30	< 30	< 30

* ASTM 527-85.
** Typical particle size distribution by laser granulometry; Demetron, 84 % smaller than stated particle size; Metz, 90 % smaller than stated particle size.

mixtures of silver and palladium powders (composite blends), simultaneously precipitated Ag – Pd powders (coprecipitates), and alloy powders (solid solutions). Examples of these stoving powders are given in Table 9.

8.2. Colloidal Silver [2, 20, 87, 104]

Like most noble metals, silver forms very stable colloidal solutions in water.

Table 9. Silver – palladium powders for multilayer capacitors

Producer	Product no.	Composition, %	Particle shape	Tapped density, g/cm^3	Particle size, µm	Specific surface area, m^2/g
Demetron	6427 1202	Ag 70 Pd 30	spheroidal	1.4	< 30	3.0
	6427 1102	Ag 80 Pd 20	spheroidal	0.6	< 12	3.6
	6437 0401	Ag 80 Pd 20	flakes	3.2	< 8	2.2
	6437 2401	Ag 70 Pd 30	flakes	3.4	< 7	3.6
	6437 0501	Ag 60 Pd 40	flakes	2.6	< 8	1.6
Metz	5030-1	Ag 70 Pd 30	spheroidal	1.2 – 2.2	< 5	1.5 – 2.4
	5070	Ag 30 Pd 70	spheroidal	1.5 – 3.0	< 5	1.5 – 4.0
	3030-1	Ag 70 Pd 30	spheroidal	2.1 – 3.2	< 2.6	1.2 – 2.4
	3030 flake	Ag 70 Pd 30	flakes	2.7 – 4.6	< 6.5	1.5 – 2.2
	3070	Ag 30 Pd 70	spheroidal	1.5 – 2.3	< 3.5	1.9 – 3.0

Properties typical of colloids are observed mainly for the particle-size range 0.001 – 0.1 µm. Fine-grained photographic emulsions can also consist of dispersions of colloidal dimensions. Most commercial silver colloids have an average particle size between 0.01 and 0.05 µm, corresponding to specific surface areas between 50 and 250 m^2/g silver. The use of silver sols in medicine (see Section 10.3) and for the disinfection of water (see Section 10.11) is due to the high reactivity of silver resulting from its large surface area.

As with all colloidal solutions of metals in water, the stability of the solution is maintained by the negative charge on the metal particles and their consequent mutual repulsion. The negative charge is caused by the adsorption of anions. If it is disturbed by aging or by the adsorption of electrolytes until the isoelectric point is reached, partial or complete coagulation and flocculation occur.

Aging and the effects of electrolytes can be inhibited by adding protective colloids, generally naturally occurring polymers. These can still be effective after the dispersant has evaporated, allowing the powder left behind to produce a colloidal solution on rewetting. Such reversible colloids are often preferred to colloidal solutions because of ease of handling.

Colloidal silver is generally produced by reduction of a dilute solution of silver nitrate by the usual reducing agents (e.g., formaldehyde, hydrazine) at elevated temperature. Polymers that behave both as protective colloids and contain reducing groups can also be used [105]. A silver sol can also be produced by the electrical dispersion of compact metallic silver immersed in water. Commercial silver sols usually have concentrations of ca. 1 % silver. Their color by transmitted light varies between yellow and brown, and they show green to blue fluorescence effects by reflected light. Colloidal silver powder, which is black and granular, has a silver content varying between 75 and 97 %. On redispersion in water, dark brown to black colors are obtained.

9. Silver Alloys [2, 18, 24, 106–112]

The binary alloys of silver can be divided into groups according to their alloying behavior (Fig. 11) [2, 19, 113].

9.1. Binary Silver Alloys [114–117]

Phase diagrams of binary silver alloys according to Massalski [118] are shown in Figures 12 (Ag – Al), 13 (Ag – Au), 14 (Ag – Cd), 15 (Ag – Cu), 16 (Ag – Ga), 17 (Ag – Ge), 18 (Ag – Hg), 19 (Ag – In), 20 (Ag – Li), 21 (Ag – Mg), 22 (Ag – Mn), 23 (Ag – Ni), 24 (Ag – O), 25 (Ag – Os), 26 (Ag – P), 27 (Ag – Pb), 28 (Ag – Pd), 29 (Ag – Pt), 30 (Ag – Rh), 31 (Ag – Ru), 32 (Ag – Si), 33 (Ag – Sn), 34 (Ag – Ti), and 35 (Ag – Zn). The most important binary silver alloys are described below.

Silver – Copper (Fig. 15). The Ag – Cu alloys are commercially the most important silver alloys. Almost all coinage and jewelry silver (Tables 10 and 11) consists of silver – copper alloys. AgCu eutectic alloys are widely used for vacuum brazing [19].

The two components are completely miscible in the liquid state but have limited solubil-

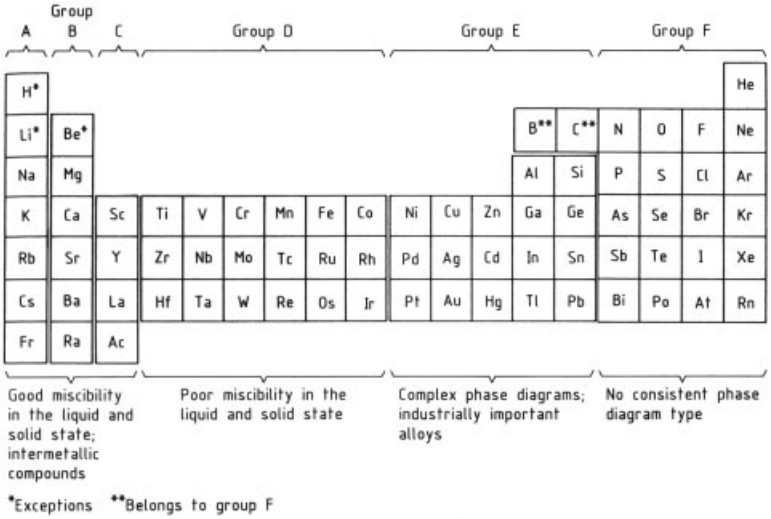

Figure 11. Grouping of binary silver alloys in the periodic table

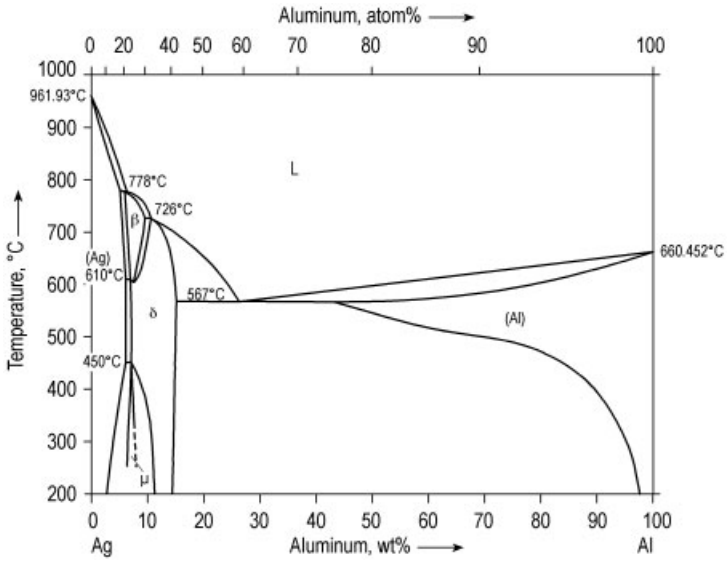

Figure 12. Phase diagrams of binary silver alloys: Ag – Al

ity in the solid state. At the eutectic temperature (780 °C), 8.8 wt % (14.1 atom %) copper dissolves in silver, and 8.0 wt % (4.9 atom %) silver in copper. At 200 °C, only 0.2 wt % copper dissolves in silver, and 8.0 wt % (4.9 atom %) silver in copper. At 200 °C, only 0.2 wt % copper dissolves in silver, and < 0.05 wt % silver in copper. At the eutectic point, the composition is 71.9 wt % (60.1 atom %) silver and 28.1 wt % (39.9 atom %) copper. Hence, in practice, all industrially important silver – copper alloys lie on the silver-rich side of the eutectic composition.

The industrial use of silver – copper alloys is based on the fact that many of their properties are satisfactory over a wide range of concentrations [2].

The color of silver is influenced by the presence of even small amounts of copper. Up to a weight ratio of 1 : 1, the alloys are almost white. Above ca. 68 % copper, they are red, and in bet-

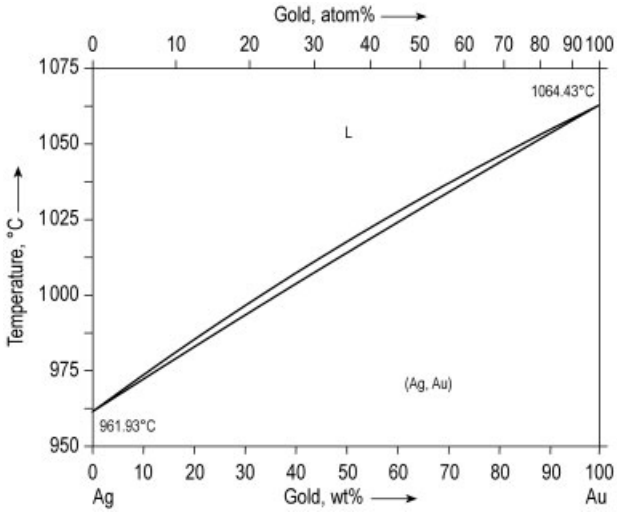

Figure 13. Phase diagrams of binary silver alloys: Ag – Au

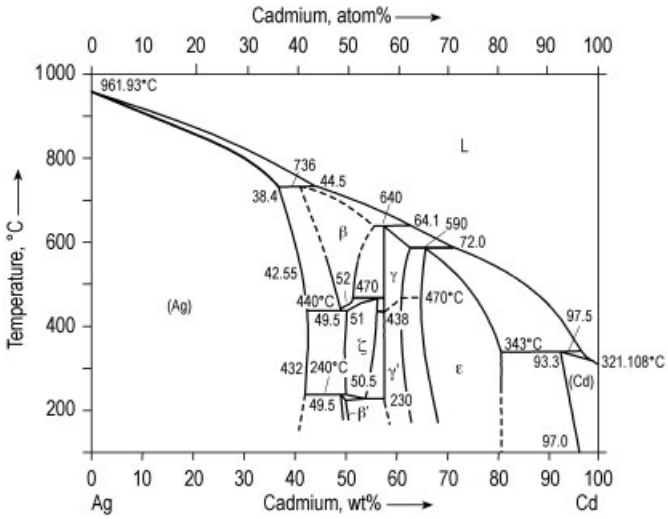

Figure 14. Phase diagrams of binary silver alloys: Ag – Cd

ween they are pale red. The shades of color, however, depend very much on the structure of the alloy and hence on its method of formation. Color variations in electrolytically deposited alloy coatings are extremely dependent on the type of electroplating bath used.

The phase relationships in the range of industrial silver – copper alloys allow the lattice structure and properties to be modified by thermal treatment. If a melt with composition Ag 925 Cu 75 (sterling silver) is allowed to cool slowly to the solidification point (915 °C) and held around this temperature, a homogeneous mixed crystal is formed in which the silver lattice positions are occupied by copper in a statistically random distribution (substitution mixed crystal). If the alloy is then quenched rapidly, the single mixed-crystal phase is retained, and the resulting alloy is ductile. If a melt of sterling silver is cooled more slowly (e.g., under the conditions of an industrial casting operation), a proportion of mixed-crystal phase is obtained together with precipitated eutectic phase. With decreasing temperature, the solubility of copper in silver decreases, and a copper-rich phase separates, leading to precipitation hardening. For

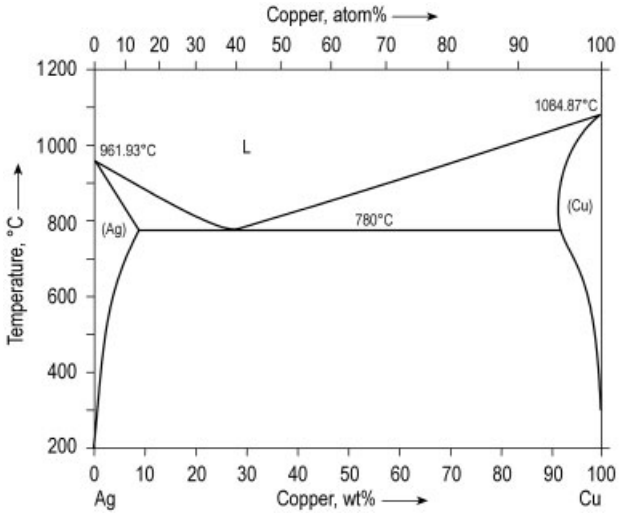

Figure 15. Phase diagrams of binary silver alloys: Ag – Cu

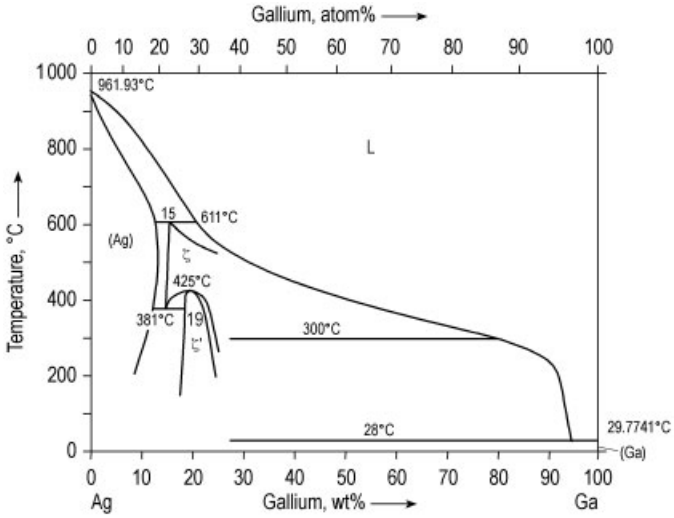

Figure 16. Phase diagrams of binary silver alloys: Ag – Ga

the same reason, ductile sterling silver with a high mixed-crystal content can be hardened by tempering at 300 °C. Ductile sterling silver is also formed by very slow cooling; the structure consists of almost pure silver and copper phases. The mechanical properties of other commercially important silver – copper alloys can be modified by similar treatment.

Silver – Cadmium. Mixed crystals are formed without miscibility gaps on the silver-rich side of the Ag – Cd system, up to a cadmium content of at most 43.2 wt % (see Fig. 14). All industrial alloys lie in this range. The existence of yellow, reddish, and violet hues in the range 40 – 60 % Cd is noteworthy.

Due to the toxicity of cadmium, the use of Ag – Cd alloys as materials for electrical contacts is now of only minor importance. They have been replaced to a large extent by silver composite materials, for example $AgSnO_2$ materials. However, to produce Ag – CdO composites (e.g., Ag – CdO: 90/10 and Ag – CdO: 85/15) by inner oxidation, the corresponding Ag – Cd alloys must first be produced [58]. Special safety precautions must be taken when the alloy is

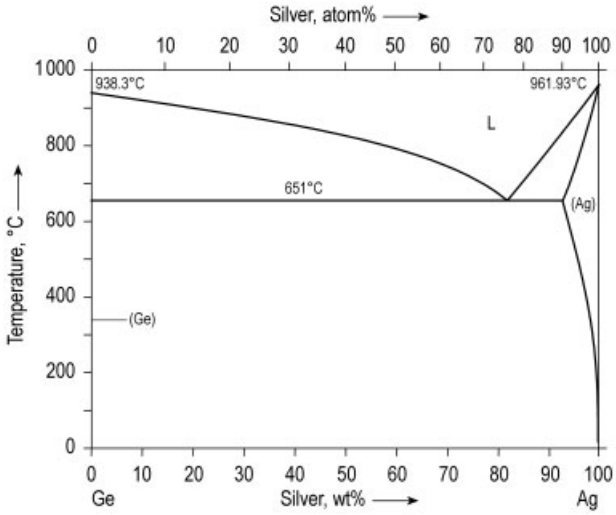

Figure 17. Phase diagrams of binary silver alloys: Ag – Ge

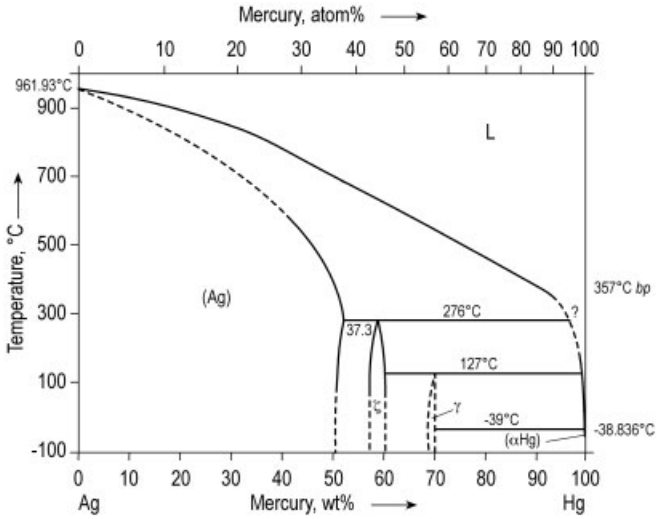

Figure 18. Phase diagrams of binary silver alloys: Ag – Hg

melted because of the high toxicity of both cadmium vapor and the readily formed cadmium oxide aerosol. Cadmium (*mp* 320.9 °C, *bp* 767 °C) has a high vapor pressure.

Silver – Gold. Silver and gold form mixed crystals without miscibility gaps (see Fig. 13). Neither separation nor transformation occurs in any part of the system. Binary silver – gold alloys are of little interest as construction materials because of their low hardness.

Silver – Palladium (Fig. 28). Apart from gold, palladium is the only metal that forms mixed crystals with silver without miscibility gaps. The alloy Ag 23 Pd 77, which is used for hydrogen diffusion membranes, exhibits extremely high hydrogen permeability and high dimensional stability.

Silver – Nickel. The extremely low mutual solubility of silver and nickel (see Fig. 23) is utilized industrially in the production of Ag – Ni composite metal by powder metallurgical processes. AgNi alloys with low Ni contents (ca.

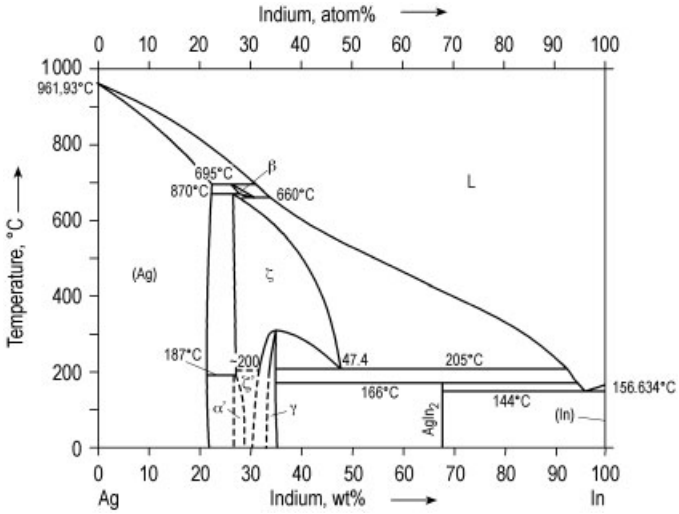

Figure 19. Phase diagrams of binary silver alloys: Ag – In

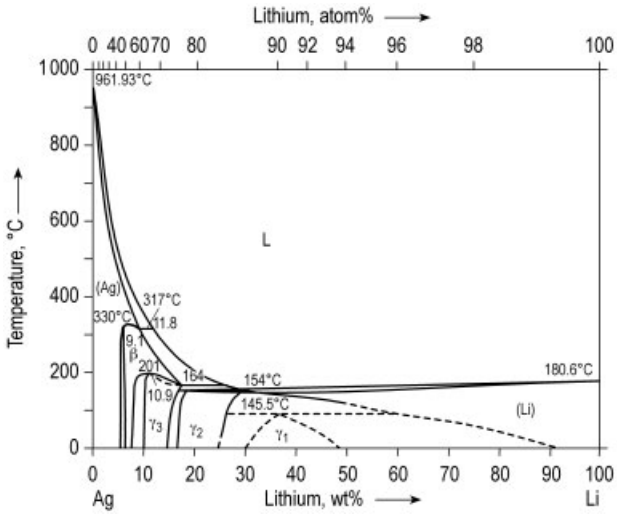

Figure 20. Phase diagrams of binary silver alloys: Ag – Li

0.15 %) can be produced by melting and casting, which lead to fine-grained silver.

Silver – Lead. Because of the eutectic in the Ag – Pb system (see Fig. 27), careful cooling of a lead alloy with a very low silver content in the region of the eutectic temperature causes relatively pure lead to separate as a solid until the content of the melt reaches 2.7 wt % (4.5 atom %) silver. This is the basis of the Pattinson process for the enrichment of silver in lead (see Section 4.2).

Other Binary Silver Alloys Used Industrially. The systems Ag – Zn (Fig. 35) and Ag – Hg (see Fig. 18) are important in silver extraction (see Section 4.2), and Ag – Hg is also important in dental technology (see Section 10.4), although ternary and even higher systems are more important.

In brazing, binary alloys of silver with aluminum, beryllium, indium, manganese, lithium, and titanium, are sometimes used as prealloys. Lithium and titanium promote wettability of the metal by the brazing alloy.

Because of the high solubility of silver in molten sodium, silver equipment or silver-containing brazing alloys must not come in contact

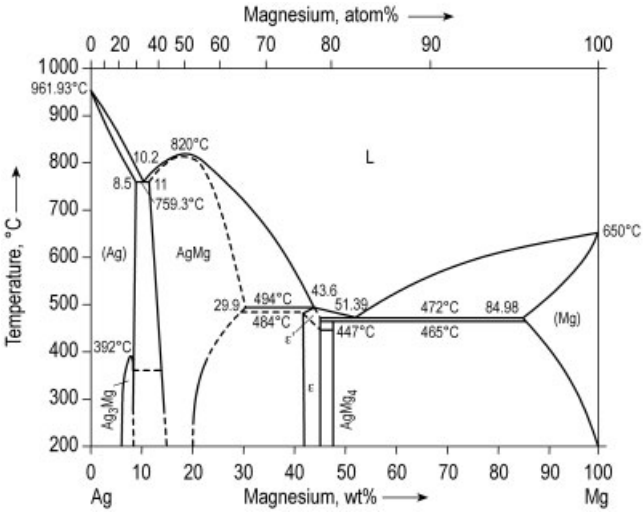

Figure 21. Phase diagrams of binary silver alloys: Ag – Mg

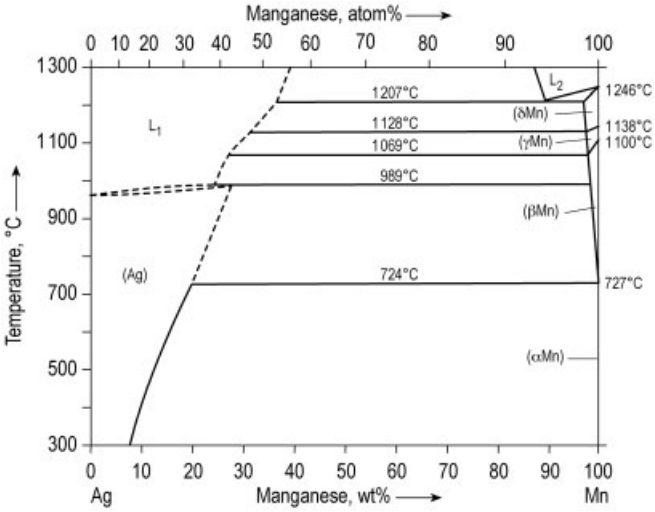

Figure 22. Phase diagrams of binary silver alloys: Ag – Mn

with sodium metal. Small quantities of Mg or Si in the silver increase its hardness.

9.2. Ternary Silver Alloys

Ternary silver alloys are of greater industrial importance than binary alloys, the only important exception being the system Ag – Cu. Alloys containing three or more components allow a wide range of mechanical, and electrical properties to be obtained. For metals that either do not alloy with silver or do so to only a limited extent in binary systems, the presence of a third component often makes alloy formation possible. A knowledge of three component alloys often leads to the experimental development of alloys containing four or more components.

The ternary alloys of silver have been investigated extensively [110–113, 119].

Silver – copper – gold is a very important system in jewelry alloys (see →Gold, Gold Alloys, and Gold Compounds, Chap. 8.2.), almost exclusively in the gold-rich region. These alloys exhibit a wide range of hardness and color.

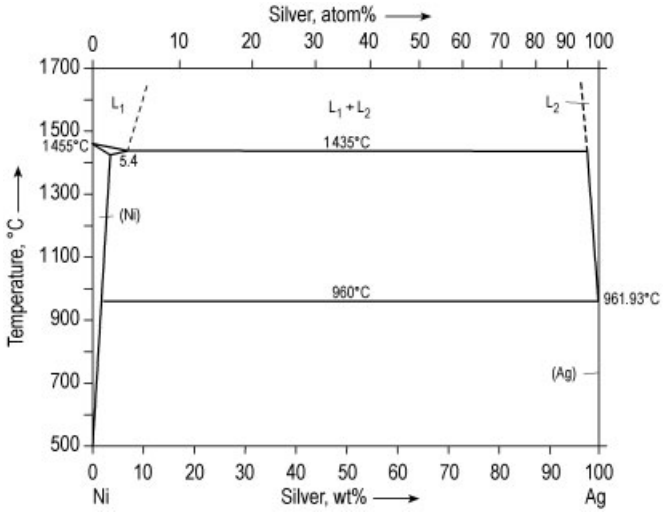

Figure 23. Phase diagrams of binary silver alloys: Ag – Ni

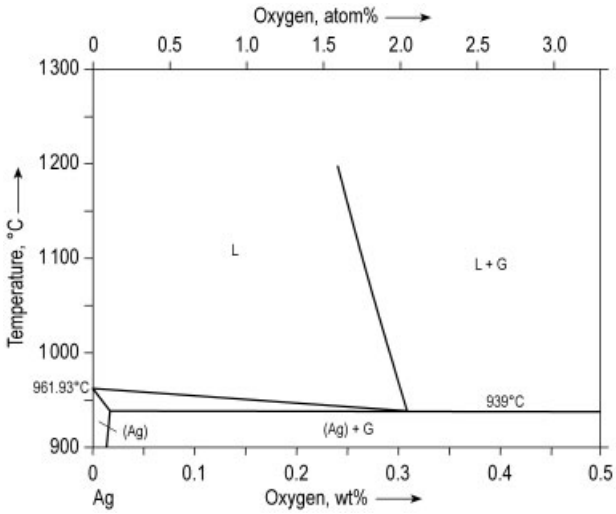

Figure 24. Phase diagrams of binary silver alloys: Ag – O

Silver – tin – mercury is the basis of the technology of dental amalgams (see Section 10.4).

Silver – Lead – Zinc. Formation of the silver-containing zinc crust in the Parkes process for the extraction of silver from ores (see Section 4.2.4) takes place essentially in this ternary system.

Silver – lead – gold alloys are present in the lead shaft furnaces of noble-metal recycling smelters. However, the entire process takes place in the molten state, so that phase relationships in the solid state are of little importance.

Silver – Copper – Phosphorus. Some hard solders with a low silver content, used without flux, are ternary systems of the type Ag – Cu – Cu_3P.

Silver – Copper – Zinc. Brazing alloys with low melting temperatures exist among the Ag – Cu – Zn alloys. L-Ag44 consist of 44 % Ag, 30 % Cu and 26 % Zn and has a melting range of 680 – 740. Even lower melting ranges can be achieved in the AgCuZnSn systems.

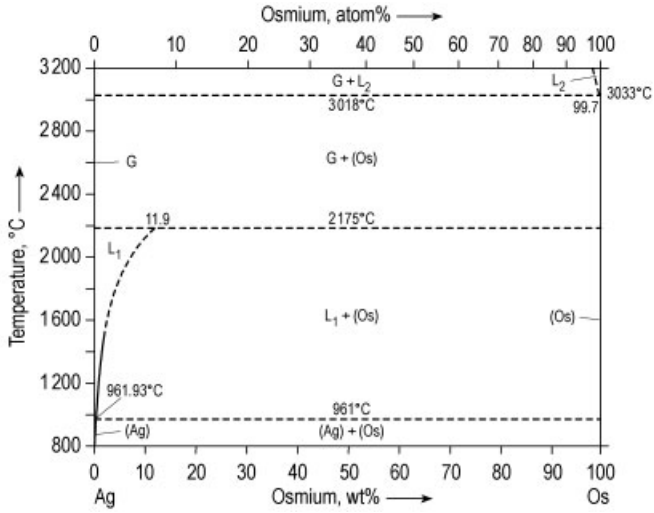

Figure 25. Phase diagrams of binary silver alloys: Ag – Os

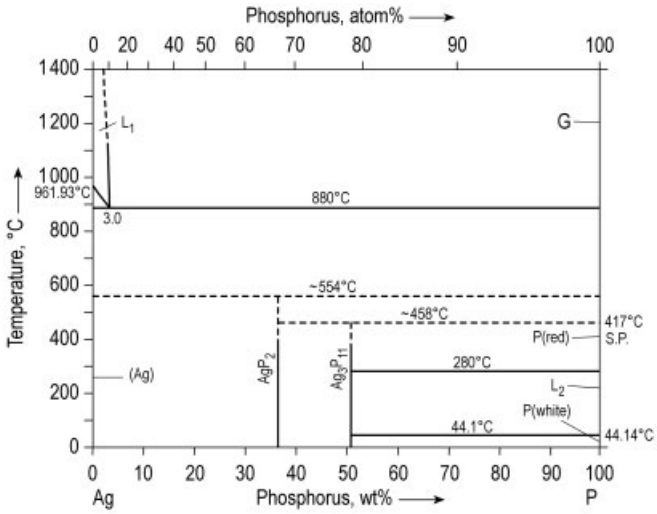

Figure 26. Phase diagrams of binary silver alloys: Ag – P

Silver – Cadmium – Indium. The alloy Ag 80 Cd 5 In 15 has been used to control nuclear reactors. It combines a high thermal neutron capture cross section with high thermal conductivity, mechanical stability, and good corrosion resistance toward hot water.

9.3. Quaternary Silver Alloys

Commercial quaternary silver alloys include brazing alloys, which belong mainly to the Ag – Cu – Zn – Sn system, and modern gold-colored dental alloys, which generally contain palladium or platinum in addition to Au – Ag – Cu [19].

9.4. Manufacturing [18, 23, 24, 58]

Fine silver is cast into marketable bars, rods, and granules by using gas- or oil-fired furnaces and gas or electric crucible furnaces. The molten metal is cast into bars in iron molds. Reducing conditions are used during melting to prevent the absorption of oxygen by molten silver, which leads to spitting on solidification (see Section 2.3), and hence to an uneven surface on the

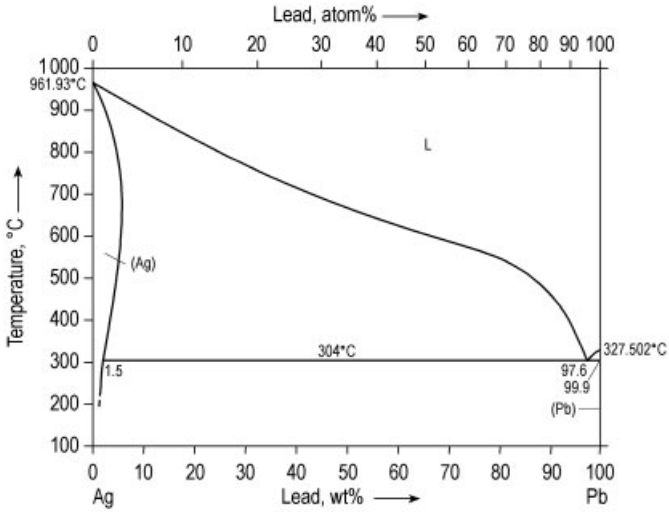

Figure 27. Phase diagrams of binary silver alloys: Ag – Pb

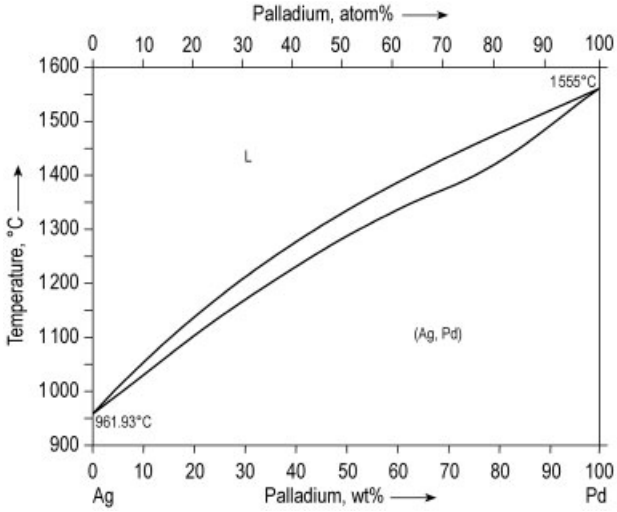

Figure 28. Phase diagrams of binary silver alloys: Ag – Pd

solid metal. In early days, the oxygen in solution in the silver was removed by stirring the melt with poles of green birch wood. Today, graphite, phosphorus, or lithium is added, or melting operations are carried out in a reducing atmosphere. Granules of fine silver are produced by pouring a thin stream of molten silver into water.

The melting of silver alloys is now carried out almost exclusively in induction furnaces. These cause a stirring effect during the melting process, which ensures thorough mixing of components. Also, the operation can readily be carried out under a protective gas to prevent uptake of oxygen. Melting under vacuum is not feasible because of the relatively high vapor pressure of silver. When molten silver alloys solidify, nonhomogeneous layer crystals nearly always form on a microscopic scale, but they can be homogenized by heat treatment. With large castings, which require a long time for cooling, liquation effects can lead to irreversible variations in composition. This causes problems in processing cast bars and billets to form sheet or wire, because the tolerances for the alloy compositions in these applications are very tight.

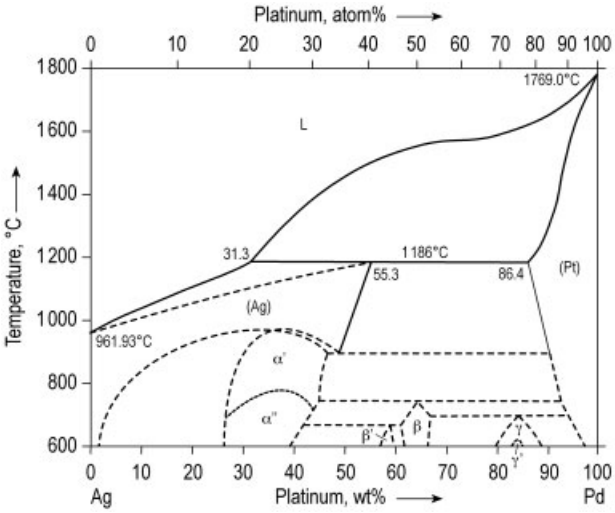

Figure 29. Phase diagrams of binary silver alloys: Ag – Pt

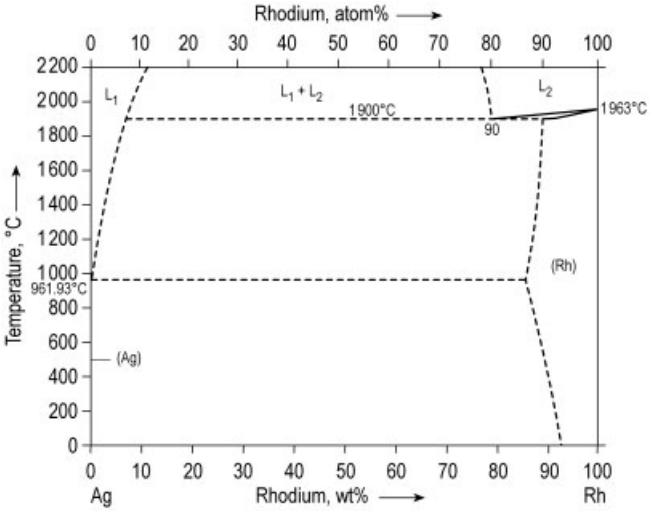

Figure 30. Phase diagrams of binary silver alloys: Ag – Rh

Silver alloys can also be shaped by continuous casting (Fig. 36). Simple profiles, as well as complex, semifinished profiles closely approximating the final cross section required (near net shape), can be produced directly from the melt. This is a major advantage because costly reprocessing is minimized. The continuous production sequence of continuous casting guarantees consistent product quality, while control of secondary cooling can prevent undesirable segregation effects. Horizontal continuous casting is used for high throughputs and large cross sections. Vertical casting plants are preferred for high-quality material and kilogram quantities.

The investment costs for continuous casting plants are very high; however, the advantages of the process have given it a dominant position in silver production technology.

Silver and silver alloys, whether produced by ingot mold or continuous casting, are converted into rods, wires, and profiles in extrusion presses. Intermediate heating is necessary, depending on the material and the extent of forming. Extrusion presses often produce directly salable end products (e.g., silver solder wire). Oth-

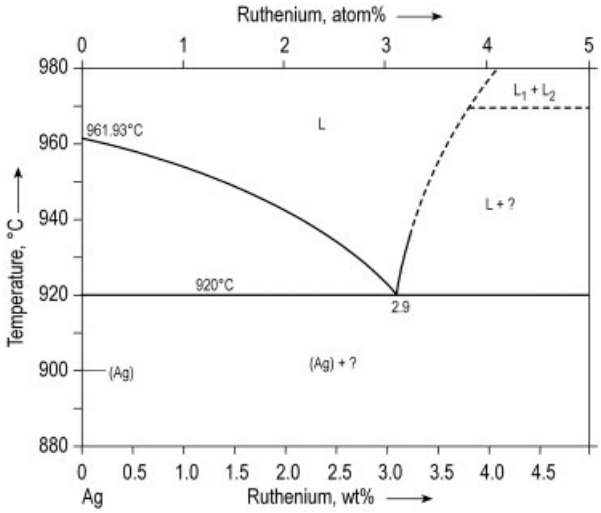

Figure 31. Phase diagrams of binary silver alloys: Ag – Ru

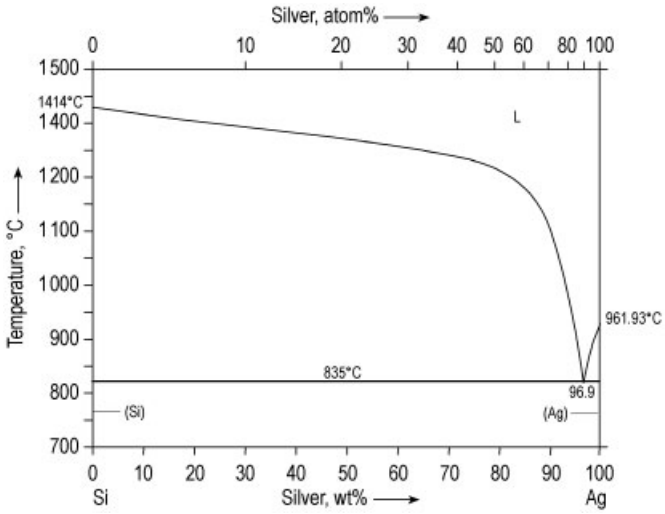

Figure 32. Phase diagrams of binary silver alloys: Ag – Si

erwise, further forming processes are required, which can include rolling, drawing, hammering, bending, pressing, stamping, forging, and weaving into gauzes. Cutting processes can also be used, including turning, milling, planing, simple engraving, and engraving with a rose engine.

For smaller items such as jewelry or fittings for chemical equipment, investment casting can be used. Usually, centrifugal casting is preferred, the molds being produced by the lost wax process. Problems of homogeneity and segregation do not usually arise here.

Thermal treatment processes not only homogenize the castings, but also restore the structure (texture) of the metal that is responsible for its hardness. Intermediate heating between forming stages must generally be carried out in nonoxidizing atmospheres for silver – copper alloys. Protective gases used include hydrogen, nitrogen, carbon monoxide, argon, or cracked gas. If silver alloys come in contact with atmospheric oxygen during thermal treatment, coatings of base-metal oxides are formed. These scale layers can be removed by pickling. If the scale con-

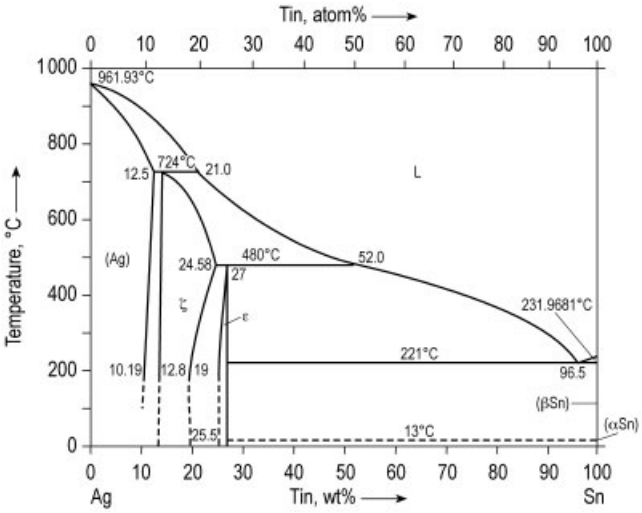

Figure 33. Phase diagrams of binary silver alloys: Ag – Sn

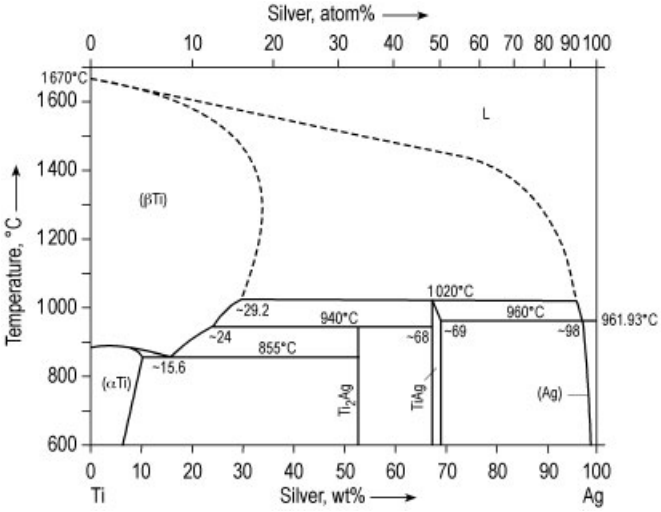

Figure 34. Phase diagrams of binary silver alloys: Ag – Ti

sists mainly of copper oxide, it can effectively be removed with 10 % sulfuric acid at 70 °C.

Sintering techniques are limited to composite silver materials.

Tula silver is silver that has been decorated by the so-called niello technique, in which a black, sulfur-containing Ag – Cu – Pb alloy is melted into an engraved design.

Silver alloys may be bonded together by the usual soldering, welding, and adhesive processes.

10. Uses [120]

10.1. Coins [1, 2, 6, 121]

The amount of coinage in circulation has undergone with fluctuations throughout the course of history. The interchange between the monetary and functional use of silver has often been influenced by the political situation. In times of war, silver was monetized to a greater extent to pay the costs, and when prosperity was restored, the metal returned to private possession in the form of silver articles. Since the growth of the industrial demand for silver, it has become de-

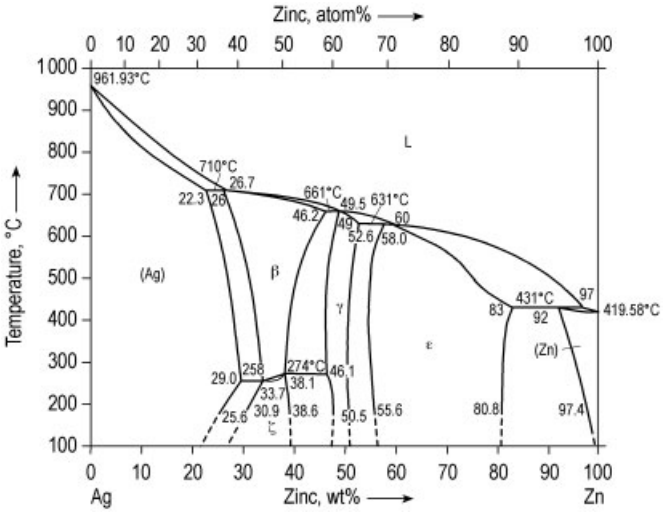

Figure 35. Phase diagrams of binary silver alloys: Ag – Zn

monetized. Nevertheless, large quantities of silver coins, withdrawn from circulation, are still used as national reserves. Even today, in some countries, small coins and banknotes are backed up by actual quantities of silver. In Germany, the circulation of silver coins reached a high point in 1880 before the introduction of paper money (see Table 10).

In ancient times, silver coins were minted from refined silver (980/1000 to 990/1000), especially in Athens and Rome. The coins of the Middle Ages usually have a fineness between 900/1000 and 990/1000. The actual value of these coins matched their face value. However, in periods of coinage debasement, so-called token coins also existed, whose metal value was less than their face value, and in which the silver content was reduced to ca. 500/1000. In the 1900s, token coins were minted generally. The silver content fluctuates between 945/1000 and 400/1000. Some silver coins are listed in Table 10.

The technique of minting silver coins has undergone many changes over time. At first, silver coins were produced partly by casting and partly by means of a hammered die. Various designs of press were then used. Today, rods or ingots are first produced by continuous casting. The ingots or rods are rolled to form plates of the exact thickness for coinage production. These plates are then heat treated in a protective gas (high-silver alloys) or in air (low-silver alloys) to remove the effects of cold work hardening. Blanks are then cut or punched out of the plates. At lower silver contents (< 800/1000) blanks are then blanched (whitened) by dipping in dilute sulfuric acid to dissolve both the thin oxide layer and the copper near the surface of the metal. This gives a lighter surface color and a pure silver appearance. After being washed, the coin blanks are dried in centrifuges, the edges are knurled in milling or crimping machines, and the edge lettering is added, this is followed by minting in a coining press or coining machine. Modern minting machines produce 8000 coins per hour.

10.2. Jewelry [18, 19, 122]

In earlier times, the manufacture of silver articles for jewelry and general use was the main outlet for silver, apart from coinage manufacture, and it is still very important. The category includes table silver such as cutlery, jugs, dishes, plates, and cups, as well as goblets, trophies, cases, church silverware, vases, lamps, and articles of jewelry such as necklaces, bracelets, broaches, pendants, and rings. Silver is also used in the fabrication of musical instruments, especially flutes. Apart from its investment and display aspects, the bactericidal action of silver has always been an important aspect of the purchase and use of silver tableware articles (see Section 13.2).

Today, silver – copper alloys with finenesses of 925/1000, 835/1000, and 800/1000 are used in

all forms of jewelry because of their favorable properties for both manufacture and use [123, 124]. Whereas fineness is now marked in thousandths, the earlier system used "Lot" units (sixteenth parts; see Table 11). Fineness was also expressed in the Russian solotnik system.

The standard materials, Ag 925, Ag 835, and Ag 800 are supplied in the form of sheets, bands, wires, rods, tubes, and profiles. The industrial process of alloy production in a vacuum oven followed by casting and cold forming has now been replaced almost entirely by continuous casting. The quality of the alloys has thereby been improved, especially with respect to homogeneity, shrinkage, gas cavities, and formability.

Tarnishing of silver is a result of the formation of silver sulfide caused by the action of hydrogen sulfide and its organic or inorganic derivatives. In the tarnishing of silver alloys, copper corrosion is significant, especially the distribution of copper in the structure. Incorporation of gold, platinum, or palladium into the alloy improves tarnish resistance but is expensive. Base-metal alloy components such as zinc, cadmium, silicon, and germanium change the color and luster without preventing tarnishing completely. Electrolytic pure silver coatings are more resistant to tarnishing than the usual silver – copper alloys. Rhodium coatings, which were sometimes used in the past, are expensive and alter the appearance, albeit only slightly. Colorless coatings (e.g., lacquers) also spoil the appearance of the silver surface.

Dipping baths for removing the tarnished layer reduce silver sulfide to fine silver, which washes off without affecting appearance. Pastes are also used that contain extremely fine abrasives and reducing agents such as sodium dithionite. Surface-active additives, usually silicone based, give some extended protection. The tarnishing of cutlery, in particular, can be prevented by wrapping it in a cloth impregnated with finely divided elemental silver. These cloths absorb sulfidic compounds from the air so that they cannot react with the compact metal surface. The most effective method of treatment, however, is still the laborious process of cleaning with a paste, which removes not only the surface layer of silver sulfide but also small surface imperfections, giving a smooth polish and restoring the original luster and warm color of the silver.

Coloring. The only important coloring processes are blanching with sulfuric acid and forming oxidized silver, which is a type of partial tarnishing.

10.3. Medicine [1,2,18,19,23,24]

Silver compounds are used only sporadically in internal medical therapy, although they were more widely used in earlier times for the treatment of epilepsy and gastric ulcers. Today, their main use is for their bactericidal action on certain pathogens. Here, they behave similarly to mercury compounds but are less toxic. Several complex compounds of silver that can be absorbed through the mucous membranes are used, mainly for disinfection of the throat. The use of very dilute solutions of silver nitrate was for a long time a legally prescribed prophylactic treatment against possible streptococcal infection of the mucous membrane of the eyes of newborn babies in Germany. Silver nitrate is a component of Fisan silver powder for the treatment of surface wounds. The local caustic effect of silver nitrate can be used to destroy proliferating tissue. A protective layer of silver chloride and silver albuminate is formed by reaction with the skin, limiting the extent of the caustic effect, which is also associated with an antiseptic and astringent action. Sulfadiazine silver(I) [AgSD, 2-sulfanilamidopyrimidine silver(I)]

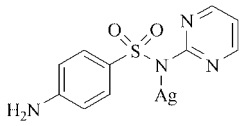

is used as a bacteriostatic compound in the treatment of life-threatening burns [120, 125].

Colloidal silver sols have similar bactericidal properties due to the oligodynamic action of silver (see Section 13.2). They are used in the form of stabilized solutions for disinfecting wounds, as are colloidal sols of sparingly soluble silver compounds.

Silver and silver alloys have long been used as bone replacements, mainly in cranial surgery. Also, silver wire is used in suturing connective tissue, and fine silver gauze is used as an implant. Here, too, the oligodynamic properties of silver

Table 10. Silver coins

Country	Year		Face value	Fineness	Silver per coin, g	Total silver content of coins issued, t
	Issue	Circulation				
United States		ca. 1895	$ 1	900	24.05	
		ca. 1895	$ 1/2	900	11.25	
		ca. 1895	$ 1/4	900	5.62	
		ca. 1895	$ 1/5	900	4.50	
		ca. 1895	$ 1/10	900	2.25	
		ca. 1895	$ 1/20	900	1.13	
	1964		$ 1/2	900	11.25	
	1965		$ 1/2	400	4.60	
	1991		$ 1	999	31.1	
Switzerland		ca. 1895	sfr. 5	900	22.50	
		ca. 1895	sfr. 2	835	8.35	
		ca. 1895	sfr. 1	835	4.175	
		ca. 1895	sfr. 0.5	835	2.083	
		ca. 1965	sfr. 5	835	12.525	
Germany	1873		M 5	900	25	
	1873		M 2	900	10	
	1873		M 1	900	5	
	1873		M 0.5	900	2.5	
	1873		M 0.2	900	1.0	
		ca. 1880	M 5/2/1/0.5/0.2	900		6000
	1924		RM 5	500	10	
	1924		RM 2	500	5	
	1924		RM 1	500	2.5	
	1936		RM 5	900	12.5	2450
	1936		RM 2	625		500
	1951 – 1974		DM 5	625	7.00	1850
	1952 – 1979		commemorative coins	625	7.00	1100
	1972		olympic games	625	9.6875	870
	1991		(200 years Brandenburger Tor)	900	15.0	

Table 11. Finenesses of some common silver alloys

Fineness in			Mark	Name	Uses
Parts per thousand	Lot ∗	Solotnik ∗∗			
800			800		cutlery, tableware
835			835		jewelry
925			925	sterling silver	tableware, cutlery
950				premier titer	high-quality cutlery
958			958	Britannia silver	high-quality cutlery
625.00	10		often none	10 lot	
750.00	12		12	12 lot	Berlin silver (mark: scepter)
781.25	12 1/2		12 1/2	12 1/2 lot	Frankfurt silver (mark: eagle)
791.66	12 lot, 12 grains				Schwäbisch Gmünd silver (mark: unicorn)
812.50	13		13	13 lot	Augsburg silver (mark: fir cone)
					Nuremberg silver (mark: N)
843.75	13 1/2		13 1/2	13 1/2 lot	Zurich silver (mark: Z)
875.00	14		14	14 lot	
888.88	14 lot, 4 grains				Venetian silver
937.50	15		15	15 lot	
875.000		84			
916.666		88			
947.916		91			
1000	16	96			medals

∗ 1 grain = 1/18 lot; 16 lot = 288 grains.
∗∗ Russian unit.

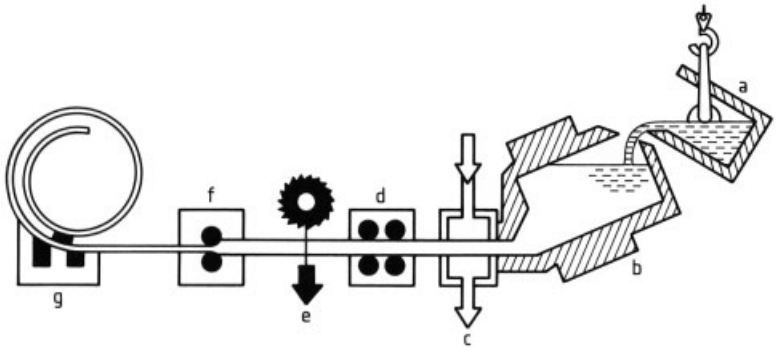

Figure 36. Continuous casting of silver
a) Pouring ladle with molten metal; b) Holding furnace; c) Continuous water-cooled mold; d) Circular saw; e) Milling machine; f) Coiling machine; g)

are beneficial. It was once common to place silver foil or even silver coins on wounds for the same reason.

10.4. Dentistry [18,23,24,58]

Silver – tin amalgams are used in dentistry to fill tooth cavities (see also → Dental Materials, Chap. 2.2.2.). The preferred alloys contain ≥40 % silver, ≤32 % tin, ≤30 % copper, ≤2 % zinc, and ≤3 % mercury. They are produced initially in the form of cast blocks, which are then powdered by machining or converted to small spheroidal particles by atomization. The dentist mixes this alloy powder with an approximately equal weight of mercury to give a stiff paste. This paste is condensed carefully into the tooth cavity, the liquid mercury acting as filling matrix and lubricant. Hardening is completed after a few hours and is accompanied by a small volume increase. In the older types of alloy, the important metallic phases present before hardening were Ag_3Sn, Sn, and Hg. During hardening, the Ag_3Sn γ-phase, which is the main component of the powdered amalgam, reacts with mercury to form a solid phase Ag_3Hg_4 (γ_1), and a Sn – Hg solid solution Sn_8Hg (γ_2). The reactions are incomplete because the amount of mercury is insufficient. Some Ag_3Sn (γ) always remains in excess.

Because of the presence of the base metal tin, the γ_2-phase is the weakest component of a filling with respect to both corrosion and mechanical properties. Formation of the γ_2-phase can be suppressed by increasing the copper content. Such formulations are known as non-gamma-2 amalgams. Despite advances in polymer-based filling materials whose color matches that of the tooth, non-gamma-2 amalgams have continued to fulfill an important function in the treatment to rear teeth where biting forces are higher.

Silver-containing alloys are also used in dental prostheses (→ Dental Materials, Chap. 2.3.). Gold-colored (gold-based) noble-metal alloys may contain 10 – 15 % silver, some white noble-metal alloys based on silver – palladium contain up to 65 % silver.

10.5. Coatings [18,19,24,126,127]

10.5.1. Silver Electroplating [128,129]

Electrolytic deposition is the most important process for producing silver coatings. The electrolytes used are alkaline baths of potassium dicyanoargentate(I) containing an excess of potassium cyanide, with potassium carbonate as the conducting salt (→ Electrochemical and Chemical Deposition, Chap. 8.17.). Soluble pure silver anodes replace the metal deposited from the electrolyte onto the cathode. For plating articles in daily use, the preferred substrate is German silver (45 – 70 % Cu, 5 – 30 % Ni, 8 – 45 % Zn), which combines good mechanical properties with good adhesion to the silver coating. Copper and nickel can also be electroplated directly. A thin initial silver coating (strike), produced in a special bath, often increases adhesion considerably. When electroplating iron, an intermediate coating of copper or nickel is neces-

sary. Articles made of plastic (mainly acrylonitrile – butadiene – styrene and polyurethane) or ceramic can be electroplated with silver if the surface is first provided with an electrically conducting coating by the electroless deposition of copper or nickel.

The luster of silver coatings can be improved by additions of xanthates or thiourea to the plating bath. Their hardness can be influenced by doping the bath with antimony, bismuth, arsenic, selenium, or lead [133]. Highly abrasion-resistant coatings, so-called dispersion coatings, are produced by dispersing ceramic materials in the electrolyte during deposition.

Coatings having different thicknesses at different parts of a workpiece can be produced by using special anodes that match the geometric shape of the workpiece and by controlling the electric field in the electrolyte with the aid of masks. This technique can be used, for example, to produce thicker deposits on cutlery at places where increased wear is expected.

Silver electroplating is used mainly in the cutlery, jewelry, and electronics industries. In total, considerably more silver is used in the manufacture of electroplated cutlery than for solid silver cutlery. The potassium dicyanoargentate(I) used in the process is the silver compound produced in greatest quantity after silver nitrate and silver(I) oxide. The quality of silver plating on cutlery is quoted in grams of silver per dozen table settings consisting of three items per setting. For the usual qualities, 60, 90, or 100 minimum coating thicknesses are specified (e.g., 36 μm for a 90-g silver plating) [127, 134]. For jewelry, coatings are considerably thinner (e.g., 5 – 10 μm). In electronics, thicknesses of a few micrometers are often sufficient.

The deposition of silver alloys usually leads to heterogeneous coatings with inferior technical properties, although significant improvements have been obtained using gold-colored Au – Cu – Ag alloy coatings, in which silver is used in place of cadmium to lighten the color [58].

Silver electroplating is carried out in batch or continuous equipment and by barrel plating. Wire-plating plants are also operated. In some cases, partial silver plating is carried out by brush plating, in which the anode and cathode are linked by a pad soaked in electrolyte.

In bulk production, continuous-flow plating plants are widely used, especially in the production of electronic components [58]. As with electroforming, their importance for gold is greater than for silver.

The need for much higher rates of electrodeposition became evident in the early 1970s when the demand for plated electronic components increased rapidly. For high-speed plating a special silver electrolyte was developed containing no free cyanide. Silver is still introduced as $K[Ag(CN)_2]$, and insoluble anodes are used. Current densities are generally between 35 and 350 A/dm^2 [130].

All attempts to develop non-cyanide silver-plating solutions have failed to provide a process that is as stable and robust as the traditional one.

10.5.2. Silver Plating by Chemical Reactions [2, 18, 126]

Deposition of silver from aqueous solutions of silver salts onto workpieces of base metals normally leads to silver coatings of inadequate quality. However, by using a solution of $K[Ag(CN)_2]$, a fairly adherent – although very thin (<1 μm) and somewhat porous%sp12;– coating can be deposited on copper by this simple method. The process is known as cementation plating.

The deposition of silver can be accelerated by contacting the metal to be plated with aluminum or zinc (contact silver plating). This plating process gives thicker, pore-free silver coatings. Contact silver plating is carried out in hot cyanide electrolytes. It is very economical for mass-produced articles, but has declined in importance.

Instead of base metals, chemical reducing agents can be added to the cyanide electrolytes [e.g., sodium hypophosphite (NaH_2PO_2), sodium borohydride, hydroquinone, or hydrazine hydrate]. The cementation reaction with the substrate metal also takes place to some extent. This plating process is improved by the addition of complexing agents such as ethylenediaminetetraacetic acid. Plastics and ceramics can also be plated by an electroless method after activating pretreatment with tin(II) chloride and palladium (II) chloride solution.

Silvering of glass, mainly for mirrors, Christmas tree decorations, and vacuum flasks, is performed with an ammoniacal silver nitrate solution to which a solution of formaldehyde and sodium hydroxide is added as reducing agent [2, 135]. Silver is deposited on the surface of the glass, usually on warming. This reduction reaction is also used in spray silvering, in which the two solutions are applied through the separate nozzles of a spray gun.

Friction plating consists of rubbing a workpiece with a moist mixture of very fine silver powder, silver chloride, and a reducing agent. This process was formerly used for silvering clock faces.

Post treatment. Silver will tarnish in time due to surface sulfide formation. Treating the silver cathodically in an alkaline chromate solution will prevent tarnishing for some time. Ornamental items can be coated with a clear lacquer. Epoxy and acrylic materials applied electrophoretically are excellent. However, these films cannot be used on components whose surface performs an electrical function.

10.5.3. Mechanical and Thermomechanical Plating [18, 19]

Mechanical plating processes are often used to produce thick silver coatings, mainly for chemical apparatus, heavy-current contacts, and coating composite semifinished products.

In roll cladding, silver is bonded to the base metal at a very thin boundary region where the two metallic crystal lattices become interlocked. Cold and hot roll cladding can be carried out.

In solder plating, a silver brazing alloy is introduced between the two surfaces to be joined, and the melting point is exceeded by a definite amount on compressing the heated metal sheets. In some systems (e.g., silver – copper), a solder-like bond is produced if the melting point of the eutectic mixture of the two metals is exceeded in the boundary region. This process is known as hot press welding or eutectic welding.

Other plating processes that can be used with silver include cold press welding, atomic hydrogen arc welding, and explosive cladding.

10.5.4. Physical Vapor Deposition.

Thin silver coatings on various substrates are produced by vacuum evaporation, by cathodic sputtering, or by ion plating. These are used in microelectronics and for decoration purposes.

A fast growing application of sputtered silver films is heat mirrors.

Thin layers of silver with thickness below $\sim 0.012\,\mu m$ are transparent for optical light waves in the visible range (380 – 780 nm) but reflect in the long-wavelength region. These layers, deposited on architectural window glasses by sputtering, reduce energy losses caused by outgoing heat radiation. Additional thin, transparent, and color-neutral oxide coatings serve as antireflection layers and to improve adhesion, corrosion resistance, and mechanical stability. Special sputtering equipment and target material conditions have been developed for large-scale batch type and continuous processing of coherent layers of homogeneous thickness [131, 132].

Figure 37. Silver pressure-contact plates for powder semiconductor devices [19, p. 487]

10.5.5. Firing Processes [18]

Ceramic articles (e.g., porcelain tableware and industrial porcelain components) and glass can be silvered by melting pasty silver-containing preparations. The resulting coatings can be soldered, polished if they are sufficientyl thick (1 – 10 μm), or used as a basis for further electroplating. The pastes consist of suspensions of very fine silver powder or organic silver compounds with powdered glass in essential oils. Heating must be carried out with addition of air

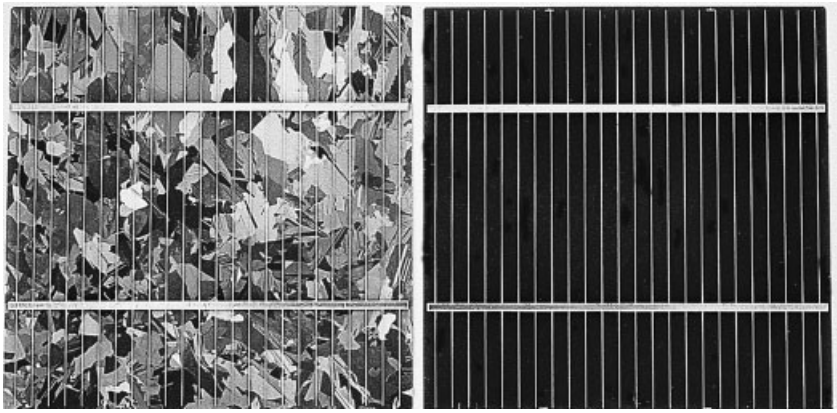

Figure 38. Silicon solar-cell with printed silver contact paths [19, p. 482]
A) polycrystalline; B) Single-crystal substrate (courtesy Nukem-Solartechnik Alzenau)

to burn off the organic components completely, preferably at 550 – 800 °C. Electronic components are often selectively silver plated by screen printing. In the mass production of glasses and porcelain tableware, decorative patterns are applied by transfer processes. Very thin layers, especially on glass, ceramic, or porcelain substrates, are also made by thermal decomposition of complex organo silver compounds (resinates) both for decoration and for microelectronic circuitry.

10.6. Electronics and Electrical Technology

10.6.1. Electronics

Silver is, due to its high electrical conductivity, the preferred conductor and electrode material. Electronic applications comprise vacuum electron tubes, active and passive semiconductor devices, electronic circuits, and components and materials needed for their production. It is used in bulk form (e.g., contact plates of rectifier and thyristor devices, (Fig. 37), as surface layers and contact pads of semiconductor chips, on silicon and gallium arsenide solar cells (Fig. 38), on lead frames (Fig. 39), packages (Fig. 40), printed circuits (Fig. 41), thik and thin-film circuits (Figs. 42, 43), capacitors (Fig. 44), and automotive windshields. Silver and silver alloys in powder form are constituents of paste preparations used for the production of conductor layers and electrodes in hybrid thick-film circuits, ceramic capacitors, and other active and passive ceramic electronic components. Silver alloys in various compositions find application as soft solders and brazes. The main applications are presented in Table 12.

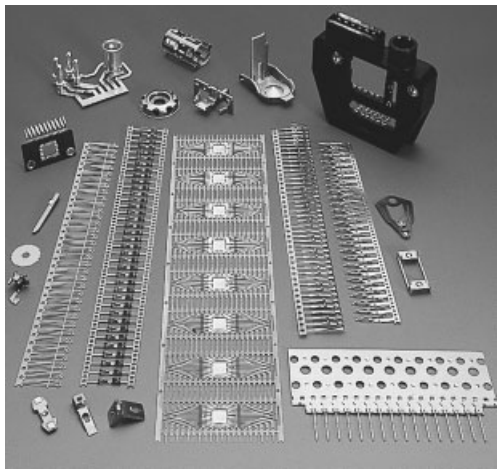

Figure 39. Lead frames and systems carrier with galvanic deposited contact layers of silver (Degussa Präzisionstechnik, Niefern)

The surface lay ers – especially for printed circuits, lead frames, and component assembly are made by galvanizing, by sputtering (e.g., for thin-film circuits), or by using powder-containing paste preparations, which are applied by screen printing and firing (thick-film circuits and electrode layers of capacitors). Typical compositions of metallizing preparations are given in

Figure 40. Ceramic packages for semiconductor power devices (courtesy Frialith, Friedrichsfeld)

Table 12. Applications of silver in electronics, [189, p. 557]

Form	Applications
Bulk	pressure-contact plates for power semiconductor devices
Bimetal	plug-in contacts, pressure-contact and soldered base plates
Surface coat	
electroplated	printed circuits boards (PCB) for high frequencies, system carrier, Mo and W base-plates, lead frames
sputtered or evaporated	contact surfaces on Si diodes lead frames
pastes, lacquers	conductor pastes for thick film circuits, potentiometer, capacitors, contact foils, screening, glass solders
solder connection	tubes, package joining components, soft solders in semiconductor devices, die-attach

Table 13. Metallizing lacquers contain silver-powder suspensions in organic lacquers. They are applied by brushing, printing, or dipping and are dried on air at moderate temperatures up to 250 °C. They are used for the production of potentiometers, Ta capacitors, flexible circuits, membrane switches, electrical shielding and other electronic devices on substrates such as plastics, glass, wood, paper or similar materials which cannot withstand temperatures above about 250 °C [182].

Multilayer capacitors are built up from alternating layers of ceramic and conductive material, usually consisting of Ag – Pd alloys or pure Pd. The layers are terminated at each end of the chip by silver coating (Fig. 45). They are produced by alternately applying layers of ceramic and of conductor pastes, then laminating together ("wet-in-wet" process), and firing in a single operation. The paste preparations are made from special powder grades of pure silver, palladium, or silver – palladium alloys with glass frits and organic binders. Table 14 lists some characteristic properties of typical powder grades.

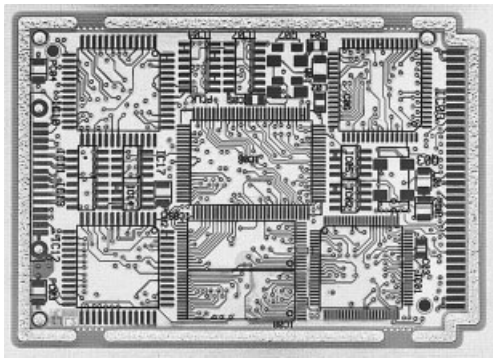

Figure 41. Multilayer printed circuit board with silver conductor network and gold contact pads for assembling with active and passive devices by surface mount technic (Schoeller Electronic, Wetter)

High-purity silver powder, as required for metallizing and electrode pastes in electronics, is produced by a two-step process:

$$2\,AgNO_3 + Na_2CO_3 \longrightarrow Ag_2CO_3 + 2\,NaNO_3$$

$$Ag_2CO_3 + HCHO \longrightarrow 2\,Ag + 2\,CO_2 + H_2$$

This process allows the alkali metal to be removed by washing before the silver carbonate is reduced with formaldehyde.

The powders are characterized by particle shape, particle size, particle size distribution, BET specific surface area [183], and tap density

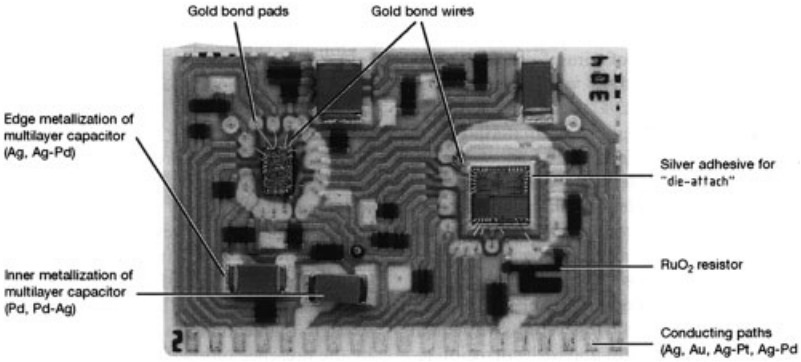

Figure 42. Thick-film hybrid circuit on ceramic substrate with conductor and resistor paths, bond pads and assembled semiconductor devices (Degussa)

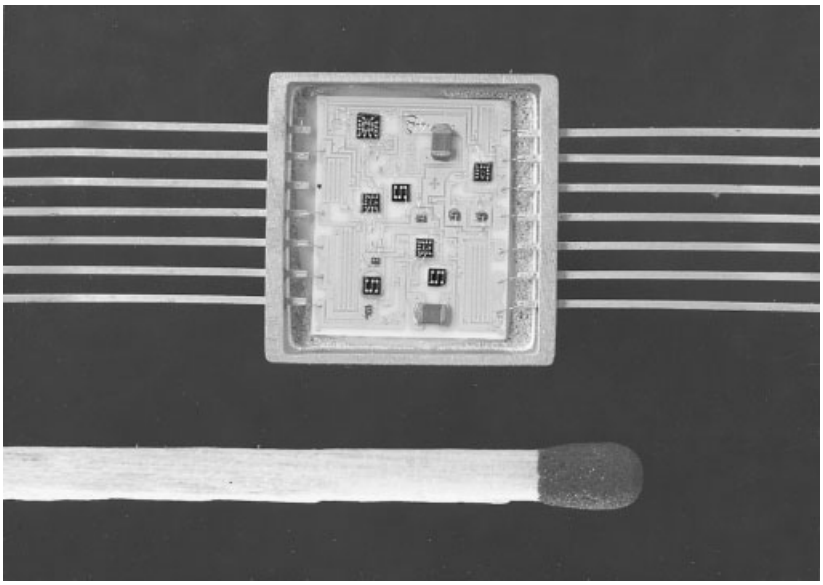

Figure 43. Thin-film hybrid circuit (power amplifier for high-frequency application) (courtesy Rohde and Schwarz, München)

Table 13. Typical composition of preparations for silver metallizing, [19, p. 496]

Type	Noble metal (NM)	Nm content, wt %	Glass frit, wt %	Organic binder, wt %	Thinner, wt %
	Ag (metallizing)	65	5	3	27
Conductive, firable	Ag, AgPd (inner electrodes)	70		5	25
	Ag (solderable)	70	5	6	19
Thick-film paste	Ag (conductor path)	70 – 80	1 – 3	3	rest.
Conductive, air-drying, thermoset	Ag, AgPt	60		5	35
Conductive, adhesive	Ag (die attach)	68	17	1	14

(ASTM 527-85 [184]). These parameters influence important production parameters such as paste rheology, sinter temperature, and density of coats after firing. Different powder production technologies lead to characteristic properties. The most common method is chemical reduction, predominantly in aqueous solution. Addition of protective colloids and compounds that promote wetting, enable the control of particle size and prevent particle conglomeration during powder deposition, Particle sizes, which also de-

Table 14. Typical silver, palladium and silver-palladium-powder grades for electrode layers of multilayer ceramic capacitors (MLCC), [189, p. 536]

Metal	Manufacturing method	Grain shape	Grain size, µm	Tap density, g/cm^3	Specific surface area (BET),m^2/g
Ag	chemical reduction	microcrystalline	0.5 – 2.0	0.8 – 5.0	0.1 – 2.0
	electrolytic deposition	dendritic	1 – 200	4.5 – 4	0.1 – 0.5
	ball milling	flakes	2 – 40	2.5 – 5	0.2 – 1.8
Pd	chemical reduction	crystalline	< 20	0.8	2.5
		spheres	< 1.2	4.0	2.3
Ag – Pd30	coprecipitation	spheres	< 1.2	3.3	2.2
		flakes	< 9	3.4	3.3

pend on the process conditions are usually in the range between 0.5 and 20 µm (Fig. 46 [185]).

Melt atomization is used for the production of silver and silver alloy powders for soft-solder and brazing pastes (Fig. 47 [185]). This is performed by mechanical dispersion or by spraying through nozzles and quenching the particles in water. High-purity spherical powders with good flow properties are formed in an inert-gas atmosphere.

Lamellar powders (flakes) are obtained by milling the deposited powders in liquid suspensions. The plate thicknesses range from 0.1 to 1 µm. The flakes are used as pigments in conductive paints and in conductive adhesives, the face contact between the flat grains leading to increased conductivity (Fig. 48 [185]).

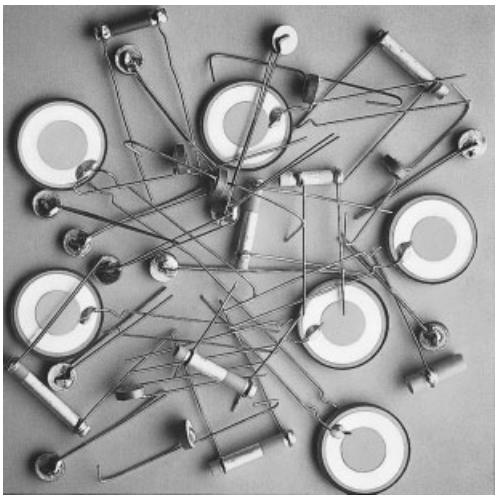

Figure 44. Plate and tube capacitors, metallized with silver preparations (Demetron)

Nanocrystalline powders with particle sizes in the range of 10 to 1000 nm and also colloidal dispersions are produced by condensation of silver vapor formed by electron beam evaporation on cool surfaces, by generating an electric arc between two silver electrodes immersed in suitable liquids, or by chemical reduction in the presence of special protective colloids. Ultrafine silver powder in an organic binder forms films with very stable electrical resistance; these are used as sensors to determine the position of magnetic tapes [186, p. 628].

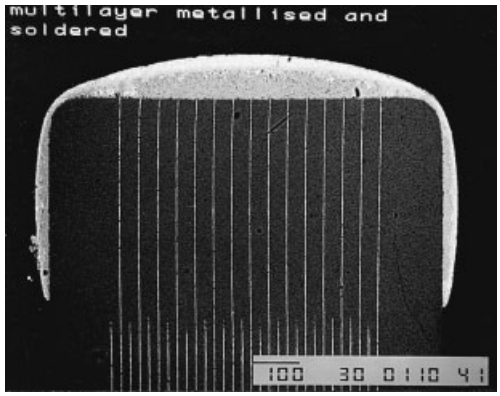

Figure 45. Multilayer ceramic capacitor (MLCC); cross section, with silver-palladium electrode layers and silver contact edge (Demetron)

Silver-based binary, ternary, and quaternary solder alloys are used for the mechanical and electrical connection of semiconductor materials with metal or ceramic components (packages, heat sinks, etc.). Addition of silver to lead – tin soft solder alloys increases the mechanical strength, improves the wetting and flow properties in the molten state, and reduces the leaching of the silver surfaces during the soldering processes. The soldering processes in electronic device and package production are predominantly carried out without flux under protective atmospheres. The solders are applied in the form of wires, foils, pastes (in preparations with an

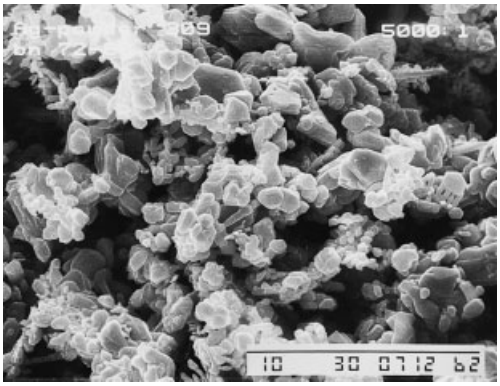

Figure 46. Ag powder made by chemical reduction (Degussa electronik)

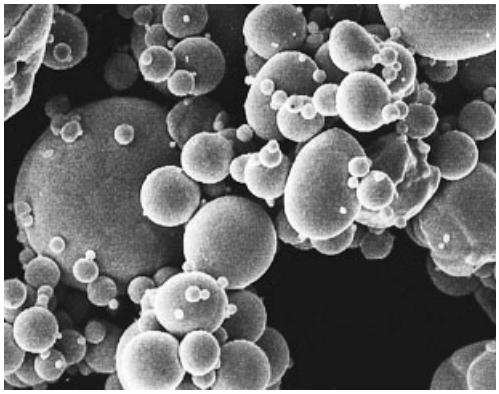

Figure 47. Ag powder made by atomization (Degussa electronik)

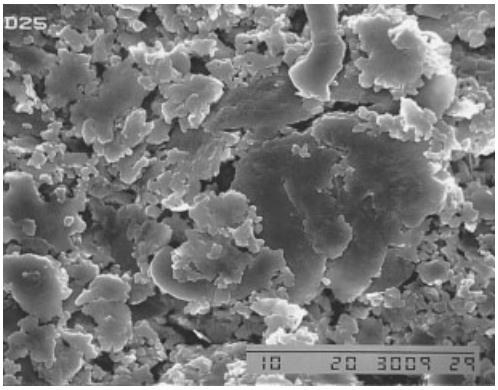

Figure 48. Ag powder flakes for conductive lacquers preparations (Degussa electronik)

Table 15. Silver-containing soft solder and vacuum brazing alloys for use in electronics, [19, pp. 556, 560]

Alloy	Composition	Melting range, °C
InPbAg	80/15/5	148 – 155
SnPbAg	62/36/2	179∗
SnAg	96.5/3.5	221∗
SnAgSb	65/25/10	228 – 395
PbAgSn	92.5/2.5/5	280∗
PbAgSn	90/5/5	292∗
PbAg	97/3	305∗
PbAgIn	92.5/2.5/5	307∗
PbAgSn	97.5/1.5/1	309∗
AgCuIn	60/27/13	605 – 710
AgCu	72/28	779∗
AgCuPd	68.4/26.6/5	807 – 810
AgCuPd	54/21/25	910 – 950
AgPd	95/5	970 – 1010

∗ Eutectic composition.

organic binder), or solder preforms (spheres, disks, or special shapes). The melting temperature range of soft solder alloys (based on lead, tin, or indium) is ca. 159 – 400 °C, and that of the brazing alloys (consisting of different compositions in the systems Ag – Cu and Ag – Cu – Pd) ca. 600 °C and above (Table 15). The ternary alloy SnSb10Ag25 has an extended melting range and higher tensile strength than the common soft solders due to the presence of low-melting phases together with hard, high-melting intermetallic compounds [187]. It is used for the die attach (connection of the semiconductor chip with the substrate) in devices which have to withstand mechanical stresses caused by temperature cycling (e.g., power devices and devices in automotive electronics).

10.6.2. Electrical Technology [2, 18, 19, 23, 24, 137]

Silver has the highest electrical and thermal conductivity of all metals, and as a precious metal it is largely corrosion-resistant. In particular, because no oxides are formed in air, silver gives extremely low electrical resistance at contact points in electrical switching devices.

Only silver sulfide, which is formed in sulfur-containing atmospheres, can hinder the flow of electric current at low contact forces, so the application of silver is concentrated in power

switching devices with their relatively high contact forces. Beside pure silver, silver – copper and silver – nickel alloys, especially the composite metals Ag – Ni, Ag – W, Ag – Mo, and the composite materials Ag – CdO, Ag – SnO_2, Ag – graphite, and Ag – WC are used (see Table 16).

The composite materials are usually produced from mixtures of the components by powder metallurgy or they can sometimes be obtained by coprecipitation followed by thermal treatment. For example, the composite system Ag – CdO can be produced by internal oxidation of the cadmium in an Ag – Cd alloy. In the case of molybdenum and tungsten, silver can be absorbed into a matrix of these refractory metals.

The choice of a specific contact material is determined by the stresses to which the contacts are subjected during making, carrying current, and current breaking.

- During making in an electric switch, the contacts bounce, and mechanical wear, transfer of the contact material from one contact to other, and arc erosion occur. There is a risk that the contacts will weld at the areas melted by the arc.
- When carrying current through closed contacts in switches and connectors, the contact resistance influences the voltage drop as well as the temperature rise and the tendency to static welding of the contacts.
- During breaking the contacts are again subjected to mechanical wear, material transfer, and arc erosion. The contact material also has an effect on arc movement and arc extinction.

While for switching currents up to 10 A the requirements can still be fulfilled by alloys like fine grain silver (AgNi 0.15) and hard silver (AgCu with 3 or 10 %), it is necessary to use composite materials for higher currents. The components, which form sintered, impregnated, or internally oxidized composite materials with the silver, are dependent on the type of switching device, its construction and the amount of current which the device must deal with under the most severe conditions. An optimum contact and switching behavior can be attached by selection of the composition, the method of manufacture, and the structure of the material for each switching device.

In Figure 49 a qualitative evaluation is given for the most important groups of materials for their contact and switching behavior by means of a simplified model. There is no optimum group materials. AgC guarantees the highest resistance to welding, but a major disadvantage is the very high erosion. AgW materials behave in the opposite way. Silver – metal oxides show predominantly positive properties and are used in nearly all switch gears with a switched current in the range of 100 – 5000 A. Because of the not completely satisfactory welding behavior above 5000 A, the materials of this group can only be used in protective and safety switches to a limited extent. AgCdO, the material used most frequently up to a few years ago, is being increasingly replaced by the more environmentally friendly $AgSnO_2$. For switching devices with a maximum switched current of 100 A, AgNi has been found appropriate, and it can be worked very easily. The contact and switching behavior of fine-grain silver and hard silver are comparable to those of pure silver, but with significantly less mechanical wear.

The areas of application for the different groups of materials are shown in Figure 50. For domestic switches, AgNi10 is the main material, and in simple switching devices in domestic installations, AgCu. Only when in both applications the contacts bounce is strong it is necessary to use silver – metal oxides like AgCdO or $AgSnO_2$. Protective switches, in which the contacts should never weld for safety reasons, have mainly AgC contacts in MCCBs and RCDs combined with Cu, and in MCBs with AgNi (20 – 40 % Ni). In auxiliary switches, such as control switches, which may only switch up to 100 A max., AgNi10 has proved best. If the switching currents are above 100 A, as in disconnectors, circuit breakers, and starters, mainly in the form of large contactors, silver – metal oxides are predominant. Here the AgCdO used up to now will be completely replaced by the newly developed material $AgSnO_2$ with 10 or 12 % SnO_2.

There are also a number of other switching devices in which silver-based materials are used for the contacts, for instance, railway switching devices or in switches and relays for motor vehicles. In electric fuses, which are also included in switching devices, silver elements are used when a high precision of the switching characteristic and good resistance to aging are required.

For fuses, silver-coated copper wires are preferred because they do not oxidize even after

Table 16. Silver-based materials for electrical contacts

Compositions, wt %										Electrical conductivity*	Characteristic properties
Ag	Cu	Cd	Ni	CdO	C	SnO_2	W	Mo	WC		
Alloys											
99.9										104	tendency to weld
92.5	7.5									88	higher mechanical strength and lower welding tendency than fine silver
90	10									88	
72	28									84	
85		15								35	reduced welding tendency
77		22.6								31	
Composite metals											
95			5							95	
90			10							87	
85			15							80	resistant to electrode burning; formable and Claddable
70			30							55	
60			40							44	
40			60							25	
90							10			92	resistant to electrode burning; large contacting forces needed
50							50			62	
35							65			51	
50								50		52	similar to Ag – W; somewhat less resistant to electrode burning
40								60		47	
Composite materials											
97.5				2.5						88	
95				5						84	very resistant to welding; intermediate coating needed for soldering
90				10						75	
85				15						65	
99.5					0.5					102	
99					1					99	extremely resistant to welding; brittle; severe electrode burning
98					2					77	
95					5					55	
90					10					35	
95						5					
92						8					slightly susceptible to welding
88						12					
65									35	57	resistant to electrode burning; very resistant to welding; difficult to work
50									50	47	
40									60	37	

*Copper = 100.

prolonged heating, and their resistance therefore remains constant. So-called superquick-acting fuses for protecting semiconductor components are made of silver wires with a defined reduction in cross section at certain points.

Silver compounds are used in the construction of electric batteries (see → Batteries, Chap. 6.1.3.2.). Silver(I) oxide is used in button cells for general electronic equipment. Silver(II) oxide, which has a higher capacity owing to divalent silver, is increasingly being used in button cells in both space and military applications. The AgCl – Zn combination is employed in emergency batteries used at sea. These start to operate automatically on contact with seawater.

The high solubility of oxygen in silver and its inertness against high-temperature reaction with ceramic oxides is used for the production of superconducting wires and tapes in single and in multifilament structure by the powder in tube technology (PIT) [188]. The high-temperature superconductors (HTSC) are based on ceramic oxides (BSCCO = $Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ with $n = 1$, 2, 3; for $n = 3$: Bi-2223). Precursor ceramic oxide powders are filled into silver tubes, which are rolled and then drawn to the final diameter. On subsequent heating in air, oxygen penetrates through the silver tube and reacts with the

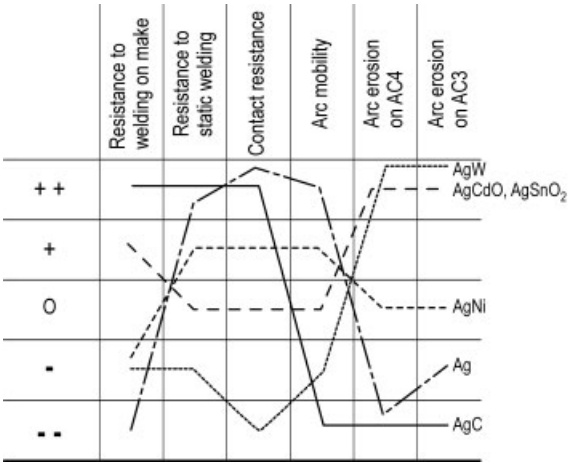

Figure 49. Contact and switching behavior of the most important types of materials; ++ very good behavior; + good behavior; o average behavior; – poor behavior; —- very poor behavior. AC3: contactors with normal load; AC4: contactors with heavy load; (by IEC-rules)

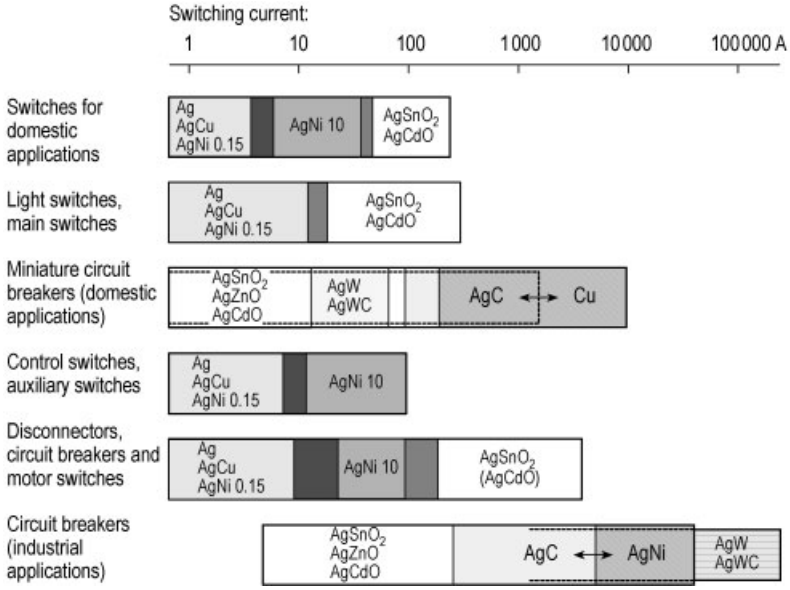

Figure 50. Areas of applications of contact materials

precursor powders to form the superconducting compounds.

10.7. Brazing Alloys [2, 18], [19, pp. 551 – 561][58, 142–146], [190, vol. 2, pp. 675 – 678], [191]

Silver-containing brazing alloys (→ Soldering and Brazing) are widely used for joining industrial metallic materials, mainly those based on copper, nickel, and cobalt alloys; tool steels; and precious metals. The main applications include brazing of construction materials, electronic components, silverware, carbide tips onto cutting tools, wear-resistant tungsten or molybdenum tips on copper alloy resistance welding electrodes, and high-temperature-resistant machine parts.

The compositions are mainly based on silver – copper alloys and contain, depending of the application type, additional amounts of one or

more of the elements Zn, Sn, Cd, Pd, Mn, and P. Table 17 gives a general survey of the most important types. Silver – manganese alloys are especially suited for soldering hard-metal and refractory surfaces (molybdenum, tungsten, tantalum). Additions of nickel or cobalt increase the strength of the bond. Ag – Pd – Cu alloys are the preferred fillers for high-temperature applications (gas turbines, supersonic aircraft, nuclear technology). For health and safety reasons, the cadmium in these alloys is being replaced as much as possible by other metals such as tin or indium.

The advantage of silver-based brazing alloys compared to those containing only base metals such as brass or German silver consists of improved workability and improved oxidation resistance during brazing. Silver-containing brazing alloys give stronger joints than brass alloys.

Compared to base metal brazing alloys like brass or German silver, the silver-based brazing alloys are more malleable. They can be drawn or rolled with high reduction between annealing steps. Silver-containing brazing alloys are mainly supplied as wires, sheets, rods, and shaped pieces of many kinds. Special preforms made of laminates of two silver – manganese outer layers on copper sheet are used to reduce mechanical stresses when materials with large differences in thermal expansion are to be joined, e.g., hard metal on steel.

Brazing alloys usually require a flux, protective gas, or vacuum. Exceptions are phosphorus-containing copper – silver brazes, which are self-fluxing on copper, silver, and tin bronze surfaces. They find application in electrical engineering and in cryotechnology, where the connections have to withstand temperature differences extending over wide ranges down to − 50 °C. For joining ceramic to metal, the ceramic surface must first be metallized with an Mo – Mn preparation (coating with pastes, drying, and firing) and then galvanically plated with nickel. Direct brazing of ceramics can be carried out without previous metallization by using the so-called active silver-brazing alloys (see Table 18). They contain 2 – 4 % titanium or zirconium. Bonding is performed by reaction of the refractory metal with the ceramic, forming an adherent bond layer.

Developments have been in the direction of flexible brazing alloys coated with flux and brazing alloys in paste form. Both of these extend the efficient use of mechanized brazing equipment, reducing the time required for brazing in mass production work. The electrical conductivity of the different silver brazing alloys depends on the silver content and ranges between 10 and 80 % IACS.

High-purity silver – copper alloys (preferably the eutectic alloy Ag – Cu28, *Mp* 779 °C, electrical conductivity ca. 77 % IACS) are needed for joining under vacuum, protective gases, or reducing gas atmospheres. The minimum purity of the components is 99.99 %. They must be free of metallic and nonmetallic constituents that have high vapor pressures at the soldering temperature (e.g., zinc, cadmium, arsenic). Palladium additions increase the mechanical strength. These alloys are predominantly applied for brazing of metallized ceramics with vacuum-tube metals and molybdenum, tungsten, zirconium, and beryllium.

The use of silver-based brazing alloys in high-tech applications is increasing. These include microelectronics, power electronics, energy technology, and aircraft and space technology. Difficult bonding problems (e.g., with titanium alloys, dispersion-hardened alloys, and hard metals) can be solved by brazing techniques, even when high strength and temperature resistance are required. This is also true of ceramic components (Si_3N_4, ZrO_2, etc.) and diamond.

Silver is also used in solders, but in lower concentration. For example, the solder L-SnAg 5 contains 3.5 % Ag and melts at 221 °C. Tin- and lead-containing solder alloys find extensive use in the production of electronic and semiconductor devices (see Section 10.6.1).

10.8. Chemical Equipment [18, 147–150]

Silver is a very useful, although expensive, material of construction in the chemical industry, having good chemical resistance, high thermal conductivity, and good mechanical and working properties.

In the chemical laboratory, silver crucibles are used mainly for alkaline fusion [149]. The presence of ca. 0.15 % nickel in the alloy prevents recrystallization on heating to red heat, which can lead to a coarse and brittle metal struc-

Table 17. Silver-containing brazing alloys

Group designation (DIN 8413)	Designation (DIN 8513, ISO 3677)	Melting range, °C	Composition, wt %								Main application	Main substrates
			Ag	Cu	Zn	Cd	Sn	P	Mn	other		
Cadmium-free universal brazing alloys	L-Ag44	680 – 740	44	30	26						solder connections operating at < 200 °C	arbitrary steel grades, copper, copper alloys
	B-Ag44CuZn											
	L-Ag45Sn	640 – 680	45	27	25	0	3					
	B-Ag45CuZnSn											
	L-Ag34Sn	630 – 730	34	36	27	0	3					
	B-Cu36AgZnSn											
Low-melting cadmium-containing universal brazing alloys	L-Ag12	800 – 830	12	48	40						solder connections operating at < 150 °C	
	B-Cu48ZnAg											
	L-Ag40Cd	595 – 630	40	19	21	20						
	B-Ag40ZnCdCu											
Manganese-containing special alloys	L-Ag30Cd	600 – 690	30	28	21	21					brazing of hardmetal mounts on tools	unalloyed and alloyed steels tungsten, molybdenum, chromium
	L-Ag49	625 – 705	49	16	23	0	0	0	7.5	4.5 Ni		
	B-Ag49ZnCuMnNi											
	L-Ag27	680 – 840	27	38	20	0	0	0	9.5	5.5 Ni		
	B-Cu38AgZnMnNi											
Phosphorus- and copper-containing brazing alloys	L-Ag15P	650 – 800	15	80	0	0	0	5			solder connections at < 150 °C	copper to copper, without flux, with flux for brass, bronze, copper alloy castings
	B-Cu80AgP											
	L-Ag2P	650 – 810	2	91.8	0	0	0	6.2				
	B-Cu92PAg											
Special brazing alloys for selected areas	L-Ag56InNi	620 – 730	56	26	0	0	0	0	0	14In, 4Ni	brazing under protective-gas atmospheres	arbitrary steel grades, copper, nickel, nickel alloys
	B-Ag56CuNi											
	L-Ag72	780	72	28								
	B-Ag72Cu											
	L-Ag85	960 – 970	85	0	0	0	0	0	15		copper-free, resistent to ammonia	chromium and chromium – nickel steel grades
	L-Ag67Cd	635 – 720	67.5	11	11.5	10					tableware jewelry	silver, silver alloys
	B-Ag67ZnCuCd											
	B-Ag68CuPd	807 – 810	68.4	26.6	0	0	0	0	0	5Pd	vacuum electron tubes, rectifiers, metallized ceramic	arbitrary steel grades, copper, copper alloys nickel and cobalt alloys, beryllium, gold, molybdenum, tungsten, titanium, zirconium, metallized ceramic
	B-Ag54PdCu	901 – 650	54	21	0	0	0	0	0	25Pd		
	B-Ag95Pd	– 970 – 1010	95	0	0	0	0	0	0	5Pd		

Table 18. Commercial active brazing alloys

Alloy	Composition, wt %						Melting range, °C	Brazing temperature, °C
	Ag	Cu	Sn	In	Ti	Pb		
CB 1	72.5	19.5		5	3		730 – 760	850 – 950
CB 2	96				4		970	1000 – 1050
CB 4	70.5	26.5			3		780 – 805	850 – 950
CB 5	64	34.5			1.5		770 – 810	850 – 950
CB 6	98			1	1		950 – 960	1030
CS 1	86	10			4		221 – 300	850 – 950
CS 2				4	4	92	320 – 325	850 – 950

ture. In preparative fluorine chemistry, silver, as well as copper, is often used for equipment parts [87].

Silver-plated articles are used in chemical process technology. These include vessels, autoclaves, pipework, heat exchangers, and sometimes small fittings. Equipment made of massive silver is uncommon, but is used for concentrating strongly alkaline solutions, for solutions of hydrochloric acid, and for fruit juices. For cost reasons, the load-bearing material is usually copper or steel, with silver cladding a few millimeters thick. In most cases, the two sheets of metal are bonded together by a combination of mechanical treatment and fusion, and the equipment is fabricated from the resulting composite. Alternatively, silver cladding can be applied to a finished steel construction. Vessels to withstand several hundred degrees Celsius are sometimes provided with a loosely fitting silver lining. Although this does not give optimum heat transfer, the difference in the expansion of the two metals does not constitute a risk of damage to the vessel. Vessels can also be plated internally by electrodeposition. Coatings >50 µm thick are generally nonporous.

Silver fittings are often machined from castings. In chemical equipment, the metal can be pure silver, fine-grained silver, or oxide dispersion-hardened silver, which has better high-temperature strength.

Filter elements of sintered spheroidal silver powder are used in the food industry, principally in fruit presses. Membranes for bursting disks to protect equipment against excess pressure can be made of 0.1 – 0.5-mm-thick silver disks. They can be used in the range of 0.1 – 40 MPa at ≤120 °C. Small tubes of palladium or Pd 77 Ag 23 are used in diffusion cells for the production of pure hydrogen. Pure silver and gold – silver alloys in wire or ring form are used in the manufacture of sealing materials for oxygen compressors and ultrahigh-vacuum equipment.

10.9. Catalysts [150]

Metallic silver is an excellent oxidation catalyst. Without doubt, the specific activation of oxygen at the metallic silver surface, forming an intermediate state, plays a part in catalytic activity. Usually, fairly coarse-grained metallic silver is used, because a finely divided product with very large specific surface area would in many cases cause total oxidation of the organic substance to carbon dioxide and water.

Ethylene oxide is produced by direct oxidation of ethylene on silver catalysts at 230 – 270 °C (→ Ethylene Oxide, Chap. 4.) [151]. These are supported on α-Al_2O_3 or silicates and have a silver content of ca. 15 %. The specific surface area of the metal is fairly low. These catalysts usually have a lifetime of several years.

Skeletal silver catalysts have also been proposed for the oxidation of ethylene to ethylene oxide [152, 153]. They are best prepared by decomposing the alloy Ag – Ca with water. However, they have not yet been used industrially.

Catalytic dehydrogenation of methanol to formaldehyde (→ Formaldehyde, Chap. 4.1.) is carried out at 600 – 720 °C over silver gauze or crystals. Catalytic oxidative dehydrogenation of isopropanol to acetone also takes place on silver catalysts. In gas-phase reactions over silver catalysts, glyoxal is produced from glycol, and acetaldehyde from ethanol. Likewise, organic amines can be dehydrated to nitriles on silver catalysts.

10.10. Photography [154] (→ Photography)

Classical silver halide photography has been subject to increasing competition from digital photography in the last few years. Nevertheless, silver halide photography will keep a considerable market share, especially in amateur photography, for the next decade.

In black-and-white photography, the light-sensitive emulsion coating consists of a gelatin layer in which silver halide crystals are suspended (two gelatin layers in X-ray films). The bromide ion is used more than chloride or iodide. The method of precipitation determines the structure of the particles in the silver halide mixture, and hence the photographic properties of the emulsion. The distribution of individual halides in the emulsion particles, the grain size, and the grain shape can be varied over a wide range, as can the bromide : chloride : iodide ratio. Properties are also influenced by physical and chemical maturing processes. Addition of special noble metal compounds enforce light sensitivity, developing properties and tuning. In the development of black-and-white film and paper, the grains that come in contact with photons are reduced by chemicals to elemental silver, to produce the black color. Silver halide grains that have not come in contact with light do not react and are dissolved by sodium thiosulfate during fixing.

In color photography, three primary colors, corresponding to the sensitivity of the human eye, must be produced. For each of the three colors, a photographically sensitive layer is required, which contains special dye components, sensitizers, and light-sensitive silver halide. On illumination, the silver halide is reduced to one of the primary colors in the layer that has been sensitized. On development, a silver image is first produced in each layer, as in black-and-white photography. An oxidation product of the developing substance is formed at the same time as elemental silver, and this couples with the dye component in each layer to form a dye, thus giving the three colored images. The three silver images are then dissolved oxidatively (bleaching), and the silver halide that has been illuminated is removed with sodium thiosulfate solution. Thus, the remaining colored image does not contain silver. Silver consumption for color photography is lower than that for black-and-white photography per unit of surface area. In all cases, high-purity silver nitrate is the starting material. Recovery of silver in color photography is a complex operation but is nevertheless economical because of the large size of most color processing laboratories.

10.11. Other Uses

Silver halides are incorporated in photochromic glasses [155, 156]. The UV component in daylight liberates elemental silver, so that the glass darkens in sunlight. As the light intensity decreases, colorless silver halide is re-formed, and the coloration decreases. The effect is used mainly in sunglasses. The change in light absorption can vary (e.g., from 10 to 50 % or from 15 to 75 %). Silver chloride in the form of colorless glass-clear films is used in special radiation detectors.

Molecular sieves charged with Ag^+ ions are used in sea rescue for producing drinking water from seawater by the reaction

$$\text{Ag-zeolite} + \text{NaCl} \longrightarrow \text{AgCl} + \text{Na-zeolite}$$

In standard sea rescue equipment, 80 g of zeolite charged with ca. 25 g of silver provides 0.4 L of drinking water.

In an increasingly used alternative to the conventional methods of removing bacteria from drinking water [157], filters containing ion exchangers or activated carbon charged with silver are used. The extent of application of this method (mainly outdoor, camping, caravaning) will depend mainly on limits to the allowable silver content of drinking water in the future. The proposed maximum of 0.01 mg/L Ag is exceeded in this process.

Colloidal silver, mostly in the form of a 1 % solution, is used for disinfection of swimming pools. Effective silver concentrations are ca. 0.1 mg/L. The advantage of the method over hypochlorite treatment consists mainly in the lack of odor. With more highly contaminated water (e.g., in open swimming pools) the effectiveness is generally inadequate.

Silver iodide, generally mixed with sodium chloride, can cause the precipitation of rain or hail if it is dispersed in the form of very small crystals with a particle size of ca. 0.1 μm. These

provide the necessary condensation nuclei. To produce early precipitation of hail, condensation nuclei are distributed in clouds in such numbers ($>100\,000/m^3$) that the hailstones formed are small and not dangerous [158]. Of more importance is the precipitation of rain in barren regions of the Earth.

11. Specifications and Analysis

11.1. Qualities and Commercial Grades

The silver bullion market is usually supplied with "good delivery" silver which is $\geq 99.9\,\%$ pure. While jewelry manufacturers may require alloys in the 80 to 92.5 % range, industrial users require purities of 99.97 %, in, for example, alloying or even 99.99 % for electrical contacts or electroplating.

For trading, the concentrations of individual impurities are in general not taken into consideration, provided their sum does not exceed the required upper limit. If agreed upon, the ASTM Standard Specification for Refined Silver (B 413-97) can be used as a guideline for the chemical composition. There exist also specifications of precious metal refiners (e.g., Degussa), characterizing the quality of their products. Special applications of silver may require particular specifications for individual metallic or nonmetallic trace elements. Table 19 illustrates the range and concentrations of other metals in the "pure" or "fine" silver grades and in the important photographic grade of silver nitrate produced for different users [193].

Silver is supplied mainly in ingot form. For good delivery quality, the ingot mass is generally ca. 31 kg (ca. 1000 oz). Small users usually prefer 1000 g fine silver ingots, although many other masses are available (e.g., cast ingots weighing 5000, 500, and 250 g; stamped bares weighing 100, 50, and 20 g; patterned plates normally 1 oz in mass, or medallions). Crafts men and some industrial users sometimes prefer silver grains, which are sold in sacks up to 20 kg content. Ingots are marked with their net mass in kilograms or troy ounces (1 troy ounce = 31.1035 g).

11.2. Sampling [162–164]

Correct sampling methods are essential for the accurate determination of silver.

Samples are taken from ingots by drilling holes diagonally, or by producing sawings by sawing across the entire cross section at various parts of the ingot. Similar methods are used for semifinished products. If possible, nonhomogeneous materials are melted and cast into ingots that can be sampled.

Waste liquids containing silver may contain sediments and deposits on the vessel walls that can be removed by filtration or dissolution, or samples can be taken after stirring.

Products in lump or powder form that cannot be melted, such as dross, are ground and sieved. The fine and coarse fractions are analyzed separately. If a metallic coarse fraction is obtained, this can be melted.

Special methods have been developed for sampling large amounts of ore, after which the above procedures are carried out.

11.3. Qualitative Analysis [160, 162, 192]

Spectrometric methods, like DC arc optical emission spectrography, flame atomic adsorption spectroscopy (FAAS) [166], inductively coupled plasma optical emission spectroscopy (ICP-OES), inductively coupled plasma mass spectrometry (ICP-MS), and X-ray fluorescence (XRF) are preferred for the qualitative detection of traces of precious metals, the qualitative investigation of small quantities of material, and the identification of precious metals in mixtures and alloys.

Amongst *chemical detection procedures,* the precipitation of silver as chloride from nitric acid solution is the most common, with as little as 0.1 μg/L detectable by the appearance of turbidity. In contrast to other sparingly soluble chlorides, silver chloride is soluble in ammonia solution.

The formation of colored complexes of silver, e.g., the redviolet complex with *p*-dimethylaminobenzylidenerhodanine or the yellow dithizonate [165], is seldom used for qualitative analysis because of possible interference by other metal ions.

Table 19. Specifications of pure silver products

Name	Silver content, %	Content of other metals µg/g	
Good-delivery silver	≥ 99.9	Cu	≤ 900
		others	≤ 100
Fine silver	≥ 99.97	Cu	≤ 250
		others	≤ 50
		including	
		Pb, Bi, Fe	≤ 10
		Se, Te	≤ 5
		total	≤ 300
Fine silver	≥ 99.99	Cu	≤ 50
		others	≤ 50
		including	
		Au, Bi, Cd, Fe, Ni, Pb, Pb, Pd, Pt, Rh, Sb, Se, Te, Zn	≤ 5
Fine silver, chemically pure	≥ 99.995	Cu	≤ 15
		others	≤ 35
		including	
		Pb, Fe	≤ 5
		Au, Bi	≤ 1
Fine silver, high purity	≥ 99.999	Al, Fe, Si, Sn	≤ 2
		Pb, Au	≤ 1
		Cd, Bi	≤ 0.5
		Cd, Mn	≤ 0.4
		total	≤ 10
Silver nitrate, photographic grade	63.5	Range of requirements of users Na, K, Ca, Mg	1 – 5
		Al, B, Cd, Cr, Cu, Fe, Pb, Mn, Ni, Sb, Se, Te, Tl, Ti, Sn, Zn	0.05 – 1
		Bi, Hg, Au, Pt, Pd, Ir, Rh, Ru	0.001 – 0.01
		nitrite	1 – 2
		sulfate	10 – 20
		chloride	1 – 5

Silver halides can be digested by fusion with $NaKCO_3$. The silver is precipitated in metallic form and can be separated by dissolving the melt in water.

Fire assay with lead as collector [160, 161] can also be used for qualitative analysis. Here, all the base metals in a powder sample are converted to slag, leaving a bead of silver and lead.

11.4. Quantitative Analysis [23, 159–162, 192]

A large proportion of the 15 000 t/a silver used by industry with a total value of ca. 4×10^9 DM is converted into silver nitrate, alloys, or into many varied forms. Calculations of the costs of production require accurate analysis of the silver contents of the products and residues. The purchase of residues and operation of recovery processes has similar needs.

Fire assay procedures (see also → Gold, Gold Alloys, and Gold Compounds, Chap. 9.3.), which have been used for hundreds of years to analyse silver and gold, are still important for the quantitative analysis of silver. *Scorification* is the fusion of the sample material with granulated lead by adding some borax or soda under oxidizing conditions in a cupellation furnace. Part of the lead remains in metallic form as precious metal collector, and another part is oxidized to PbO, which, in liquid form, dissolves the base metal oxides and forms a slag. The scorification method can be applied to the analysis of a variety of materials containing silver, such as ores, smelting products, and residues, whose silver contents are not less than ca. 0.1 %.

The aim of the *classical crucible method* is to collect the silver or, more generally, the precious metals of the sample material in a lead regulus, and at the same time scorify the base matrix components. To this end, the sample is mixed with lead oxide, a reducing agent such as charcoal or flour, and a flux as slag producer. The crucible method with lead is a procedure with a great variety of applications for enrichment and separation of precious metals. It can be applied to many metallic and nonmetallic products con-

taining silver from the ng/g up to the high percentage range.

After collecting the silver in lead, the collector metal is removed by an oxidation process (*cupellation*). Proofs are run in parallel so as to make corrections for cupel absorption and some evaporation of silver. In principle cupellation is appropriate also for silver determination in alloys. But the greater accuracy achievable with potentiometric titration means that the cupellation method is now seldom used for silver alloys.

Physical Methods. In systems of known constitution (e.g., silver brazing alloys or jewelry alloys), *X-ray fluorescence analysis* can be carried out [172–174] with an accuracy adequate for stock control.

Optical emission spectroscopy with glow discharge (GD-OES) can also be used for the quantitative determination of silver in its alloys [194].

ICP-OES of solutions with the emission lines at 328.07 and 328.29 nm is being used increasingly for precision analysis of silver. The prerequisite is the application of the principle of internal standardization, e.g., with yttrium (see → Gold, Gold Alloys, and Gold Compounds, Chap. 9.3.).

Gravimetric and Electrogravimetric Determination. Given the availability of precise physical methods, the gravimetric determination of silver by precipitation of silver chloride is no longer applied. However, the electrogravimetry of silver from cyanide solutions remains important. This method can be used for silver determinations in photographic industry scrap, which contains sparingly soluble compounds such as silver halides.

Volumetric Determination. In nitric acid or sulfuric acid solutions, silver is determined by titration with a standardized alkali halide solution with potentiometric indication of the equivalence point by using a halogenated silver electrode in combination with a $Hg/HgSO_4$ reference electrode [160–162, 167, 168, 197]. Automatic titrators which work with dynamic volume dosing are available. It is the preferred method for silver alloys, e.g., coin and jewelry alloys [195], unrefined and auriferous silver, electrotechnical alloys, recovered scrap alloys and silver catalysts, anode sludge from copper electrolysis, and others. In principle it is usable from ca. 0.1 to nearly 100 % silver. The potentiometric titration of silver has nearly completely replaced the titration method according to Gay-Lussac, in which silver in nitric acid solution is titrated with a NaCl solution with visual observation of the so-called clear point or endpoint. The potentiometric determination of silver is less subject to interferences than Gay-Lussac titration. It has become one of the most important methods of silver determination.

An alternative titrimetric method is the precipitation of Ag_2S with a solution of thioacetamide (CH_3CSNH_2). The endpoint is detected by voltammetry [169, 198]. The determination of the EMF of a silver-sensitive electrode [168, 170] should also be mentioned.

11.5. Purity Analysis

For most pure silver, the fineness is determined by subtracting the sum of the impurities from 100 %. The most important methods for the determination of trace impurities in silver are spark discharge OES (SD-OES) [196] and GD-OES [175, 194] on fused or pressed metallic samples. These methods allow rapid simultaneous determinations of the trace elements. The area of their application begins with good delivery silver (99.9 % Ag) and ranges via fine silver with 99.99 % Ag to chemically pure fine silver with 99.995 % Ag. With silver of a higher degree of purity it is better to use other methods, such as GD-MS.

Stringent purity requirements are imposed on the most important silver compound, silver nitrate, for which metallic (cationic) impurities can be checked by means of IC-OES and graphite furnace atomic absorption spectroscopy (GFAAS). Especially advantageous is the use of an ICP spectrometer with axial plasma, an echelle grating optical system, and a solid-state detector (type CID or SCD) [193]. ICP-MS is a very suitable technique for silver nitrate analysis on account of the low detection limits [193].

11.6. Trace Analysis

Detection limits of ca. 0.5 μg/g of silver in ores and rocks can be achieved by using optical spec-

trography with arc excitation with the 328.1 nm and 338.3 nm lines. If the material can be dissolved, typical detection limits are 0.001 µg/mL with FAAS [166], 0.002 µg/mL with ICP-OES, and 0.02 ng/mL with ICP-MS. XRF can also be used for trace detection, albeit with matrix-dependent detection limits.

The allowable concentration of silver in wastewater is low because of the widespread use of biological wastewater treatment. Silver(I) ions interfere with the biological equilibrium at a concentration ≥ 0.04 mg/L. FAAS or ICP-OES are widely used for wastewater analysis.

Trace analysis of precious materials in waste materials such as slags and wastewater from the precious metal industry, especially scrap recovery plants, is of great importance for checking the economic efficiency of the process.

11.7. Argentometry [167,168]

The precipitation reactions used for volumetric determination of silver can also be used for determination of ions that form insoluble precipitates with Ag^+. Titration techniques for determining chloride, bromide, iodide, cyanide, thiocyanate, etc. are generally known as argentometry.

12. Economic Aspects [15,16,23]

12.1. Production

Of all the noble metals, silver comes first in terms of annual production. Sources of supply are mine production, material from recycled scrap, and additional material from net official sales, from hedging and from implied disinvestment.

World silver supply in 1998 is summarized in the following:

Mine production	16 969 t
Recycled silver scrap	5934 t
Net official sector sales	1613 t
Hedging	158 t
Implied disinvestment	1470 t
Total	26 144 t

Figure 51 gives a survey of the proportion of quantities of silver produced by mining, ranked by country in relation to the total silver supply.

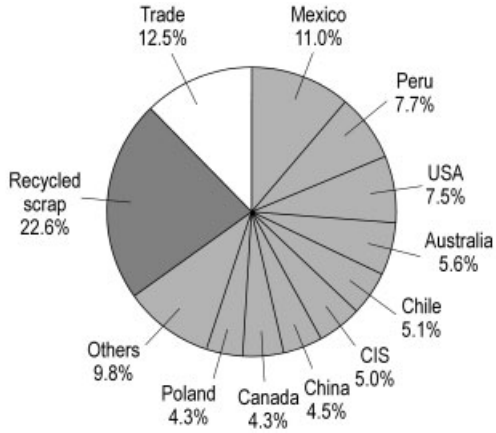

Figure 51. World silver supply in 1998 (mine production by country, scrap, trade)

Mine production is the biggest source with about 65 % of the total supply. Silver is produced from primary silver ores and as byproduct from gold, copper and lead – zinc mines (Fig. 52). The biggest mines (with output in 10^6 oz) are located in Mexico (Industrias Penoles 36.4, Grupo Mexico 16.1), Poland (KGHM Polska Mieds 35.3), Australia (BHP Minerals 18.1), U.K. (Rio Tinto 16.2), Canada (Cominco Ltd. 13.3, Noranda 11.2), Peru (Cia. de Minas Buenaventura 10.8). Their production amounted in 1998 to one-third of the total mine supply.

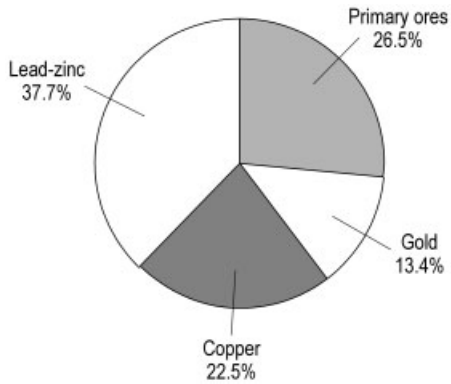

Figure 52. Silver mine production by source in 1998

The amount produced is relatively insensitive to price movements, which are sometimes large and generally of a speculative nature, because silver is mainly a side product of the extraction of base metals. Also, market price fluc-

tuations are reflected to only a limited extent in production from recycled materials. However, government sales, and sometimes purchases, are strongly dependent on price changes. Sales of silver from the CIS have a relatively small influence on the total supply situation in the Western world in spite of the considerable production in this region. Also, in comparison to gold and the platinum-group metals, large amounts of silver are frequently released to the world market from the old stocks in India, estimated at around 120 000 t. This material was accumulated many years ago from domestic production and commercial transactions. The total stocks of the Western world (excluding India) amount to ca. 80 000 t, of which approximately 50 % is owned by the U.S. government.

12.2. Demand

Figure 53 give a survey of the percentage in the most important application fields. The biggest part is the photographic sector (main product silver nitrate), with radiography (38 %), consumer papers (22 %), consumer films (15.1 %), and graphic arts (13.5 %), followed by jewelry and silverware, and electrical technology and electronics. The term "general industrial applications" in the following list refers to various fields such as the use of silver and silver compounds in catalyst, batteries, mirrors, bearings, and bactericides.

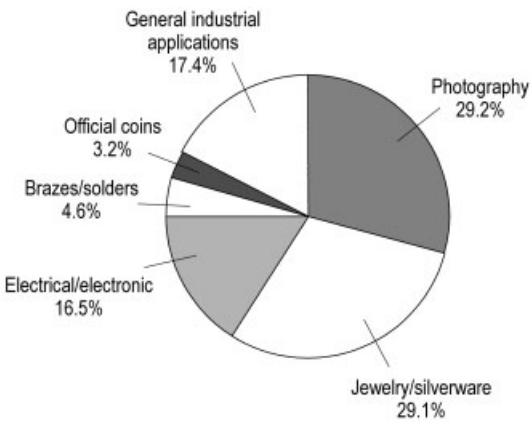

Figure 53. World demand for silver by application (percentage of total demand)

World silver demand in 1998 was as follows:

Photography	7628 t
Jewelry and silverware	7601 t
Electrical/electronic	4307 t
Brazing alloys and solders	1200 t
General industrial applications	4559 t
Official coins	849 t
Total	26 144 t

The demand for silver (Fig. 52) is fairly well matched by the supply. However, the relatively small extent of known silver reserves in ores, compared to other metals, could in the future lead to a change in stockpiling practice. In technology, an important factor in the future is likely to be a decrease in the importance of classical silver-based photography, to the benefit of electronic image-making techniques.

12.3. Silver Market and Trading

The most important silver exchanges are the New York Commodity Exchange and the TOCOM in Tokyo. The London Bullion Market Association (LBMA) provides through some of its members one spot price information per day: the London Silver Fixing.

The amounts of silver transferred yearly are often many times greater than the world's demand for silver.

On June 22, 1992, the price of silver was $ 4.30 per ounce, having remained around this low level for many years. The low point of $ 3.65 per ounce reached in 1972 was again reached in 1992.

In February 1998, the investment fund of Warren Buffet announced having bought more than 4000 t of silver in the proceeding months based on a perceived long term supply/demand deficit. The price of silver thereafter on speculative buying shot to $ 7.81 per ounce on February 6th. Failing any follow through, the silver price retreated to below $ 5.00 per ounce by August 1998.

13. Toxicology, Occupational Health, and Safety

13.1. Toxicology

Silver and its compounds are much less toxic than other heavy metals and their compounds

because silver compounds are very poorly absorbed in the gastrointestinal tract; in addition, they are rapidly converted into insoluble compounds or complexed by metallothioneine. Acute silver poisoning occurs when colloidal silver is administered parenterally in high doses [200]. Silver nitrate or silver fluoride can cause tissue damage because of their caustic properties. Gastroenteritis, diarrhea, drop of blood pressure, cramps, paralysis, and respiratory arrest were observed as poisoning symptoms. A 10 % aqueous silver nitrate solution seriously damages the eye [201].

Adverse effects (anemia, growth retardation, liver necrosis, and fatty degeneration of liver and kidneys) have been observed when animals were treated with repeated doses of silver salts [203, 204]. Implantation of silver foil or injection of colloidal silver caused local tumors in rats, probably as an unspecific reaction to foreign bodies but not a true carcinogenic effect [205].

In Germany, the MAK value for silver metal has been fixed at 0.1 mg/m^3 and for silver compounds at 0.01 mg/m^3. This low value suggests a high toxicity, but it has been set at one tenth to one hundredth of that for other heavy metals to certain prevent argyria (see below). The TLV for silver metal has been set at 0.1 mg/m^3 in the United States and other industrialized countries. BAT values for blood and urine that corresponds to the MAK values could not be determined as yet. Usual base values in these fluids are lower than 1 µg/L, even in workers with high exposures to silver dust [206, 207].

According to German law (Trinkwasserverordnung) drinking water must not contain more than 10 µg/L of silver.

Argyria. Argyria is the consequence of chronic intake of silver salts, colloidal silver, or finely divided silver metal that leads to local accumulation of silver and causes gray to brownish discoloration of the skin, especially of areas exposed to light [202]. Localized argyria is the result of direct silver intake of the skin. Generalized argyria follows absorption of silver from the gastrointestinal tract or by parenteral administration [208]. The symptoms are irreversible and cannot be treated by medication. Apart from the dark skin and local silver accumulation in internal organs, no systemic poisoning is observed. Particulary workers in the silver nitrate and silver-plating industries are at risk argyria. As a result of better workplace hygiene, the condition has become rare.

13.2. Ecotoxicology

Bacteria, some fish and other waterborne species are more sensitive to silver salts than other animals and plants.

Usually silver and its compounds do not constitute an environmental problem. However, the discharge of silver and its compounds into surface water needs special attention due to their toxicity to some species [203]. The U.S. EPA has classified the precious metals silver, gold, platinum, and palladium (and other "classical" heavy metals) as "very poisonous" and "easily available". The other platinum group metals have been assigned to the middle class (poisonous, but not easily dissoluble or rare).

Bactericidal Effect. The metabolism of bacteria is adversely affected by silver ions at concentrations of 0.01 – 0.1 mg/L. Therefore, even less soluble silver compounds, such as silver chloride, also act as bactericides or germicides, but not the much less soluble silver sulfide.

In the presence of atmospheric oxygen, metallic silver also has a bactericidal effect due to the formation of silver oxide, which is soluble enough to cause it. Bactericidal concentrations are produced extremely rapidly by adding colloidal silver, which has a high surface area, but even objects with a solid silver surface (e.g., table silver, silver coins, or silver foil) have a bactericidal effect in the absence of sulfide, which would otherwise combine with the Ag^+ ions.

The death of microorganisms such as bacteria, molds, spores, and fungi on contact with silver articles was discovered by von Nägeli (1817 – 1891), although he did not identify the cause. The effect was certainly used in antiquity, and silver drinking vessels were carried by military commanders on expeditions for protection against disease. Nägeli named the effect that he discovered the oligodynamic effect. This is now generally understood to mean the damaging effect of very small amounts of metallic cations on living cells.

13.3. Explosion Hazards [19]

Silver forms many unstable compounds (see Section 7.7) that can explode or detonate on thermal or mechanical stress, on drying, under the influence of light, or sometimes spontaneously without any known cause. This group includes several nitrogen compounds (e.g., AgN_3, $AgNH_2$, AgCNO) of silver, organic compounds (especially silver acetylide and its derivatives, also $(COOAg)_2$), and silver(II) oxide in contact with finely divided organic materials.

Since some of these compounds can be formed by reaction with elemental silver, acetylene and ammonia in the gaseous or liquid state should not be allowed to come in contact with items of equipment made of silver or its alloys. Also, silver and its compounds must be kept away from acetylene flames.

Silver salts of strongly oxidizing acids (e.g., silver chlorate and nitrate) can lead to dangerous explosions on contact with oxidizable materials (e.g., organic substances, sulfur, soot). Contact of silver nitrate with alcohol leads to immediate formation of explosive silver fulminate. Many explosions have resulted from formation of nitride in ammoniacal solutions of silver. Silver oxalate is reported to decompose at 140 °C, but explosions have occurred on drying at 80 °C.

14. References

1. R. Kerschagl: *Silber, Metallische Rohstoffe,* vol. 13, Enke, Stuttgart 1961.
2. A. Butts, C. D. Coxe: *Silver – Economics, Metallurgy, and Use,* R. E. Krieger Publ. Comp., Huntington 1975.
3. V. Tafel: *Lehrbuch der Metallhüttenkunde,* Hirzel, Leipzig 1951.
4. D. Büttner, *Naturwiss. Unterr. Phys./Chem.* **28** (1980) 129 – 140.
5. A. Binz: *Edelmetalle, ihr Fluch und ihr Segen,* W. Limpert, Berlin 1943.
6. H. Römpp: *Chemie der Metalle,* Franckh'sche Verlagshandlung, Stuttgart 1941.
7. N. H. Gale, W. Gentner, G. A. Wagner, *Metall. Numis.* **1** (1980) 3 – 49.
8. W. Gentner, H. Gropengiesser, G. A. Wagner, Mannheimer Forum 1979/80, Boehringer Mannheim.
9. W. Gentner, O. Müller, G. A. Wagner, N. H. Gale, *Naturwissenschaften* **65** (1978) 273 – 284.
10. E. Pernicka, *Erzmetall* **34** (1981) 396 – 400.
11. H.-G. Bachmann, *Erzmetall* **35** (1982) no. 5, 246.
12. C. Schnabel: *Lehrbuch der allgemeinen Hüttenkunde,* Springer, Berlin 1913.
13. G. Agricola: *Vom Berg- und Hüttenwesen,* Deutscher Taschenbuchverlag, München 1980.
14. L. Suhling: *Der Seigerhüttenprozeß,* Riederer Verlag, Stuttgart 1976.
15. W. Gocht: *Handbuch der Metallmärkte,* 2nd ed., Springer Verlag, Berlin 1985.
16. E. Schmidt, *Metall* **47** (1993) 644 – 667.
17. S. D. Dial, *Metall (Berlin)* **35** (1981) 580 – 590.
18. Degussa: *Edelmetall-Taschenbuch*, Frankfurt 1967.
19. Degussa AG: *Edelmetall-Taschenbuch,* 2nd ed., Hüthig-Verlag, Heidelberg 1995.
20. *Gmelin,* **System no. 61,** Silber, 1970/76.
21. G. H. Aylward, T. J. Findlay: *Datensammlung Chemie,* Verlag Chemie, Weinheim 1975.
22. E. Fluck, K. G. Heumann: *Periodic Table of the Elements,* VCH Verlagsgesellschaft, Weinheim 1986.
23. Fonds der Chemischen Industrie: *Edelmetalle –Gewinnung, Verarbeitung, Anwendung,* Folienserie no. 12, Frankfurt/Main 1989.
24. Degussa: *Edelmetall-Taschenbuch,* 2nd ed., Metallverlag, Berlin 1994, in press.
25. D. G. Cooper: *Das Periodensystem der Elemente,* Verlag Chemie, Weinheim 1972.
26. D. Mehdows, E. Zahn, P. Milling: *Die Grenzen des Wachstums,* Deutsche Verlagsanstalt, Stuttgart 1972 /1979.
27. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* VCH Verlagsgesellschaft, Weinheim 1988.
28. F. Pawlek: *Metallhüttenkunde,* Walter de Gruyter, Berlin 1983.
29. L. F. Trueb: *Die chemischen Elemente – Ein Streifzug durch das Periodensystem,* S. Hirzel Verlag, Stuttgart/Leipzig 1996.
30. D. F. Schutz, K. K. Turekian, *Geochim. Cosmochim. Acta* **29** (1965) 259 – 313.
31. H. Schneiderhöhn: *Erzlagerstätten,* G. Fischer Verlag, Stuttgart 1962.
32. A. Cissarz: *Einführung in die allgemeine und systematische Lagerstättenlehre,* E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1965.
33. Metallgesellschaft: *Die Welt der Metalle – Kupfer (Kupferatlas), The World of Metals – Copper,* Frankfurt 1993.

34. Metallgesellschaft: *Die Welt der Metalle – Blei, Zink,* Frankfurt 1994, in press.
35. L. van Zyl, C. P. Schreuder (eds.): *South Africa's Mineral Industry,* Department of Mineral and Energy Affaires, Bramfontein 1991/92.
36. R. Saager: *Metallische Rohstoffe von Antimon bis Zirkonium,* Bank Vontobel, Zürich 1984.
37. Degussa: Recycling – Rohstoffsicherung und Umweltschutz, Frankfurt 1981.
38. National Association of Recycling Industries: *Precious Metals,* New York 1980.
39. E. Krone, W. Dähne, *Chem. Ztg.* 1977, 421 – 425.
40. W. Kumpf, K. Maas, H. Straub, G. Hösel, W. Schenkel: *Müll- und Abfallbeseitigung (Müll-Handbuch),* Kennzahl 8572 (C. Knorr: Rückgewinnung von Silber aus photographischen Rückständen), Kennzahl 8595 (H. Renner: Behandlung und Verwertung von verbrauchten edelmetallhaltigen Katalysatoren), Erich Schmidt Verlag, Berlin 1980/81.
41. W. W. Behning, *Metall (Berlin)* **24** (1970) 794 – 796.
42. W. Dähne: *Recycling von Silber aus Fotomaterialien,* GDMB Gesellschaft Deutscher Metallhütten- und Bergleute, no. 3, lecture of 10/23, Nov. 1991 in Würzburg, Clausthal-Zellerfeld 1991.
43. D. H. Chambers, B. W. Dunning: *Silver Recovery from Aircraft Scrap,* **Report no. 8477,** Bureau of Mines, Washington 1980.
44. *Ullmann,* 3rd ed., **15,** 630.
45. K. Winnacker, L. Küchler: *Chemische Technologie,* 4th ed., vol. 6, Carl Hanser Verlag, München 1986.
46. *Kirk-Othmer,* 3rd ed. **21,** 1 – 31.
47. H. J. Bovey, S. Marks, *S. Afr. Chem. Process* **4** (1969) no. 2, 28 – 34.
48. M. Toyoda, A. Ohta, T. Shimizu: *Slime Treatment at Niihama Copper Refinery,* Metallurgical Society of AIME, New York 1976.
49. F. Habashi, *Chem. Ztg.* **86** (1962) 483 – 487.
50. R. A. Zingaro, W. C. Cooper: *Selenium,* Van Nostrand Reinhold, New York 1974.
51. W. C. Cooper: *Tellurium,* Van Nostrand Reinhold, New York 1971.
52. J. Feiser: *Nebenmetalle. Die metallischen Rohstoffe,* vol. 17, Enke, Stuttgart 1966.
53. W. Schreiter: *Seltene Metalle,* vols. II and III, Deutscher Verlag f. Grundstoffindustrie, Leipzig 1961.
54. Inko, DE-OS 2 914 439, 1979.
55. American Metal Climax, DE 2 117 513, 1971.
56. W. Truthe: Hundert Jahre Gold- und Silberscheidung nebst Gewinnung der Platinmetalle, Degussa, Frankfurt 1943, unpublished.
57. G. A. Walker: "Recovery and Refining of Secondary Silver," 9th Commonwealth Mining and Metallurgical Congress, London 1969.
58. Degussa: Metall – Forschung und Entwicklung, Frankfurt/M. 1991.
59. Degussa, DE 2 059 235, 1970.
60. Outokumpu Oy, DE 2 543 027, 1975.
61. G. Eger: *Das Scheiden der Edelmetalle durch Elektrolyse,* W. Knapp, Halle 1929.
62. Dietzel, DE 68 990, 1892; DE 82 390, 1895.
63. H. Walden, *Chem. Ing. Tech.* **36** (1964) 647 – 651.
64. Kodak: *Die Silberrückgewinnung in der Fotographie,* Stuttgart 1972.
65. Kodak, US 2 944 886, 1958.
66. Cendres & Métaux, CH 613 787, 1976.
67. Röhm GmbH: Ablösen der silberhaltigen Gelatineschicht von Filmabfällen, Company brochure, Darmstadt 1977.
68. B. Krauss, *Moderne Fototechnik* 1979, 125 – 127, 237 – 240, 433 – 437.
69. E. Mutter, *Röntgen- Laboratoriumsprax.* **13** (1960) 158 – 165, 175 – 191.
70. H. Aisenberg, U. Fritz, *Röntgenpraxis (Stuttgart)* **19** (1966) 284 – 307.
71. G. Kolf, *Chem. Lab. Betr.* **32** (1981) 43 – 48.
72. W. Grossmann, *Moderne Fototechnik* 1980, 557 – 558.
73. Degussa: TMT 15 für die Abtrennung von Schwermetallen aus Abwässern (Reihe Umwelt und Degussa), Frankfurt 1980.
74. Minnesota Mining, US 4 110 109, 1978.
75. Siemens, DE-AS 2 543 600, 1975.
76. Ciba-Geigy, US 4 186 067, 1978.
77. Comptoir Lyon-Alemand, Louyot & Co.: Procédé Purhypo, Company brochure, Paris 1979.
78. Degussa: Behandlung photochemischer Abwässer mit Wasserstoffperoxid (Reihe Umwelt und Degussa), Frankfurt 1980.
79. R. Mina: "Silver Recovery from Photographic Effluents by Ion Exchange Methods," SPSE-Symposium, Las Vegas 1980.
80. D. J. Degenkolb, F. J. Scobey, *J. Soc. Motion Pict. Telev. Eng.* **86** (1977) 66 – 68.
81. C. Ramsey: Ion Exchange for Silver Removal from Effluents, PMA Fall Seminar, 1979.
82. MacDermid Inc., US 3 242 090, 1964.
83. J. W. Hunter: "Electrochemical Principles and Refining," IPMI-Seminar, Skytop Lodge 1980.

84. Brookside Metal Comp., CH 614 238, 1974; US 3 975 244, 1976.
85. DEMAG-Elektrometallurgie GmbH, DE 1 151 128, 1964.
86. A. Prior, R. Marr, H. J. Bart in: *Raffinationsverfahren in der Metallurgie,* Verlag Chemie, Weinheim 1983, pp. 243 – 252.
87. G. Brauer: *Handbuch der präparativen anorganischen Chemie,* vol. 2, Enke, Stuttgart 1978.
88. L. Vanino: *Präparative Chemie,* vol. 1, Ferdinand Enke Verlag, Stuttgart 1925.
89. N. R. Thompson: *Comprehensive Inorganic Chemistry,* vol. 17, Pergamon Press, Oxford 1973.
90. Du Pont, US 2 614 029, 1951.
91. Du Pont, US 2 543 792, 1949.
92. Du Pont, US 2 940 828, 1957.
93. W. A. Chalkin, E. Herrmann, J. W. Norseev, I. Dreyer, *Chem. Ztg.* **101** (1977) 470 – 481.
94. JANAF, Thermochemical Tables, NSRDS-NBS 37, 1971; Supplements 1974, 1975, 1976, 1978.
95. O. Kubaschewski, C. B. Aloock: *Metallurgical Thermochemistry,* Pergamon Press, New York 1978.
96. I. Barin, O. Knacke, O. Kubaschewski: *Thermodynamical Properties Inorganic Substances,* Springer, New York 1977.
97. Varta, DT 2 430 910, 1974.
98. A. M. Golub et al.: *Chemie der Pseudohalogenide,* Hüthig Verlag, Heidelberg 1979.
99. IUPAC: *Regeln für die Nomenklatur der Anorganischen Chemie,* Verlag Chemie, Weinheim 1970.
100. A. F. Holleman, E. Wiberg: *Lehrbuch der Anorganischen Chemie,* de Gruyter, Berlin 1976.
101. A. Stettbacher, *Explosivstoffe* **2** (1954) 1.
102. W. Romanowski, S. Engels: *Hochdisperse Metalle,* VCH Verlagsgesellschaft, Weinheim, in press.
103. H. Stegemann, DE-AS P 31 196 357, 1981.
104. W. Ostwald: *Die Welt der vernachläßigten Dimensionen,* Th. Steinkopff, Dresden 1927.
105. Degussa, DE 2 302 277, 1973.
106. E. Raub: *Die Edelmetalle und ihre Legierungen,* Springer Verlag, Berlin 1940.
107. E. Jäneke: *Kurzgefaßtes Handbuch aller Legierungen,* C. Winter, Heidelberg 1949.
108. J. L. Haughton, A. Prince: *The Constitution Diagrams of Alloys. A Bibliography,* Institute of Metals, London 1956.
109. A. Prince: *Multicomponent Alloy Constitution,* Bibliography 1955 – 1973, Metals Society, London 1978.
110. A. Prince: *Multicomponent Alloy Constitution,* Bibliography 1974 – 1977, Metals Society, London 1981.
111. F. N. Rhines: *Phase Diagrams in Metallurgy,* McGraw-Hill Book Co., New York 1956.
112. W. Home-Rothery, G. V. Raynor: *The Structure of Metals and Alloys,* Institute of Metals, London 1954.
113. *Ullmann,* 4th ed., **21,** 344 – 351.
114. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* McGraw-Hill Book Co., New York 1958.
115. R. P. Elliot: *Constitution of Binary Alloys,* **1st Suppl.,** McGraw-Hill Book Co., New York 1965.
116. F. A. Shunk: *Constitution of Binary Alloys,* **2nd Suppl.,** McGraw-Hill Book Co., New York 1969.
117. W. G. Moffatt: *The Handbook of Binary Phase Diagrams,* General Electric Company, Schenectady, N.Y., 1976/77.
118. T. B. Massalski (ed.): *Binary Alloy Phase Diagrams,* **vols. 1 and 2,** American Society of Metals, Metals Park, Ohio 1986.
119. G. Petzow, G. Effenberg: *Ternary Alloys,* **vols. 1,** 2, VCH Verlagsgesellschaft, Weinheim 1988.
120. A. A. Boraiko, *National Geographic* **160** (1981) 280 – 313.
121. A. Luschin v. Ebengreuth: *Allgemeine Münzkunde und Goldgeschichte,* R. Oldenbourg, München 1971.
122. E. M. Link: *Ullstein Silberbuch 1968 – Eine Kunst- und Kulturgeschichte des Silbers,* Ullstein-Verlag, München 1968.
123. W. Londershausen, *Metall (Berlin)* **35** (1981) 465 – 466.
124. M. Rosenberg: *Goldmerkzeichen,* Frankfurter Verlagsanstalt, Frankfurt 1922.
125. A. Bult, *Pharm. Weekbl. Sci. Ed.* **3** (1981) 213 – 223.
126. E. v. Angerer, H. Ebert: *Technische Kunstgriffe bei physikalischen Untersuchungen,* Vieweg u. Sohn, Braunschweig 1964.
127. J. Fischer: *Galvanische Edelmetallüberzüge,* Leuze Verlag, Saulgau 1960.
128. H. W. Dettner, J. Elze, E. Raub: *Handbuch der Galvanotechnik,* C. Hanser, München 1966.
129. A. v. Krusenstjern: *Edelmetallgalvanotechnik,* Leuze Verlag, Saulgau 1970.
130. P. W. Meuldijk: *14. Ulmer Gespräche,* Eugen G. Leuze Verlag, Saulgau 1992, p. 81.

131. W. D. Dachselt, W. D. Münz, M. Scherer: "Optical Coatings for Energy Efficiency," *SPIE Proceedings, no. 324,* SPIE, Bellingham WA 1982, pp. 37 – 43.
132. H. H. Gläser, *Glastechn. Ber.* **62** (1989) no. 3, 93 – 99.
133. A. Keil, *Metalloberfläche* **9** (1955) A 81.
134. Deutscher Normenausschuß (ed.): Versilberte Bestecke, RAL 691 B 3, Beuth-Vertrieb, Berlin 1969.
135. G. Brauer: *Handbuch der präparativen anorganischen Chemie,* 1st ed., Enke, Stuttgart 1954.
136. Degussa, DE-OS 2 803 060, 1978.
137. A. Keil: *Werkstoffe für elektrische Kontakte,* Springer Verlag, Berlin 1960.
138. W. Böhm, N. Behrens, M. Clasing, *Metall (Berlin)* **35** (1981) 539 – 543.
139. K.-W. Jäger, B. Gegenbach, *Metall (Berlin)* **35** (1981) 534 – 539.
140. M. Poniatowski, K.-H. Schröder, E. D. Schulz, *Metall (Berlin)* **31** (1977) 1338 – 1342.
141. M. Lindmayer, *Metall (Berlin)* **34** (1980) 621 – 625.
142. DIN 8513,Blatt 2, Hartlote für Schwermetalle, silberhaltige Hartlote mit weniger als 20 Gew.- % Silber, Beuth Verlag, Berlin 1966.
143. DIN 8513,Blatt 3, Hartlote für Schwermetalle, silberhaltige Hartlote mit mindestens 20 Gew.- % Silber, Beuth Verlag, Berlin 1969.
144. T. Nesse, K.-F. Zimmermann, *Metall (Berlin)* **28** (1974) 1026 – 1029.
145. W. Mahler, K. F. Zimmermann: *Hartlöten von Kupfer und seinen Legierungen,* Deutscher Verlag für Schweißtechnik, Düsseldorf 1966.
146. Degussa: *Technik die verbindet,* vols. 1 – 26; periodical.
147. G. Reinacher, *Chem. Anlagen + Verfahren* 1972, no. 5, 44 – 46.
148. Degussa: *Silberhalbzeug,* Hanau 1980.
149. R. Bock: *Aufschlußmethoden der anorganischen Chemie,* Verlag Chemie, Weinheim 1972.
150. K. Weissermel, H.-J. Arpe: *Industrielle organische Chemie,* Verlag Chemie, Weinheim 1978.
151. S. Rebsdat, S. Mayer, J. Alfranseder, *Chem. Ing. Tech.* **53** (1981) 850 – 854.
152. A. Camborn, D. Alexander, *Can. J. Chem.* **34** (1956) 665.
153. B. M. Bogolowski, S. Kasakowa: *Skelettkatalysatoren in der organischen Chemie,* Dt. Verlag der Wissenschaften, Ostberlin 1960.
154. Fonds der Chemischen Industrie: Silber-Fotographie, Folienserie no. 23, Frankfurt/Main 1989.
155. B. J. L. Kratzer, *DOZ Dtsch. Optikerztg.* **33** (1978) 153 – 156.
156. W. Grimm, *DOZ Dtsch. Optikerztg.* **34** (1979) 173 – 180.
157. F. Krusche: *Die Trinkwassersilberung,* Verlag R. Oldenbourg, München 1957.
158. B. Federer, *Umsch. Wiss. Tech.* **79** (1979) 463 – 468.
159. H. Pohl: Studie über Probleme der chemischen Edelmetallanalyse, Forschungsbericht Nr. 15 der Bundesanstalt für Materialprüfung (BAM), Berlin 1972.
160. *Edelmetall-Analyse,* Springer Verlag, Berlin 1964.
161. *Handbuch der analytischen Chemie,* **3rd part, vol. 1,** Springer Verlag, Berlin 1967.
162. *Analyse der Metalle,* **vols. I – III,** Springer Verlag, Berlin 1961 –1975.
163. H.-M. Lüschow in Ges. Dt. Metallhütten- und Bergleute (eds.): *Probenahme, Theorie und Praxis,* Verlag Chemie, Weinheim 1980.
164. P. M. Gy: *Sampling of Particulate Materials,* Elsevier, Amsterdam 1979.
165. J. Fries, H. Getrost: Organische Reagentien für die anorganische Spurenanalyse, E. Merck, Darmstadt 1975.
166. B. Welz: *Atomabsorptionsspektroskopie,* 3rd ed., Verlag Chemie, Weinheim 1983.
167. G. Jander, K. F. Jahr, H. Knoll: *Maßanalyse,* vol. II, Sammlung Göschen vol. 1002, Walter de Gruyter, Berlin 1973.
168. Metrohm-Information, Jahrgang 1 (1971); Jahrgang 11 (1981).
169. K. Lenhardt, *Erzmetall* **28** (1975) 172 – 177.
170. A. M. Bond: *Modern Polarographic Methods in Analytical Chemistry,* Marcel Dekker, New York 1980.
171. W. Prodinger: *Organische Fällungsmittel in der quantitativen Analyse,* Enke, Stuttgart 1954.
172. H. G. Bachmann, E. Koberstein, R. Straub, *Chem. Tech. (Heidelberg)* **7** (1978) 441 – 446.
173. J. Suchomel, *Fresenius' Z. Anal. Chem.* **300** (1980) 257 – 266.
174. J. Suchomel, *Fresenius' Z. Anal. Chem.* **307** (1981) 14 – 18.
175. H. M. Lüschow in K.-H. Koch, H. Massmann (eds.): Tagungsband der 13. Spektrometertagung 1980, de Gruyter, Berlin 1981.
176. W. Diehl, *Metall (Berlin)* **23** (1967) 587 – 589.
177. B. Lange: *Kolorimetrische Analyse,* Verlag Chemie, Weinheim 1962.

178. E. Merian (ed.): *Metals and their Compounds in the Environments – Occurrence, Analysis and Biological Relevance,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1989.
179. L. Friberg, G. F. Nordberg, V. B. Vouk: *Handbook on the Toxicology of Metals,* Elsevier, Amsterdam 1979.
180. W. Dähne: MAK-Wert für Silber, Bericht über das Geschäftsjahr 1978 der Fachvereinigung Edelmetalle e.V., Düsseldorf 1979.
181. D. Henschler in Deutsche Forschungsgemeinschaft (eds.): *Gesundheitsschädliche Arbeitsstoffe,* Verlag Chemie, Weinheim 1979.
182. Präparate DEMETRON, Degussa Frankfurt/Main, DM 60-5-07-491 GB.
183. P. Brunauer, P. Emmet, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309.
184. N. Stanley-Wood, *Powder Met. Rev.* **10** (1977) no. 3, 138.
185. K. Lutz: Hagener Symp. Pulvermet., Hagen, 19th/20th November, 1998.
186. L. S. Benner, T. Suzuki, K. Meguro, S. Tanaka (eds.): *Precious Metals, Science and Technology,* International Precious Metals Institute, Allentown, PA 1991.
187. J. Gierl: Diplomarbeit, Lehrstuhl für Werkstoffkunde und Technologie, Universität Erlangen, March 1997.
188. M. Mimura et al., *Appl. Phys. Lett.* **54** (1989).
189. G. Schlamp in R. Cahn (ed.): *Material Science and Technology,* **vols. 8/9,** VCH Verlagsgesellschaft, Weinheim 1995, pp. 467 – 587. M. Fieberg, H. Renner, G. Schlamp in *Encyclopedia of Applied Physics,* VCH, Weinheim 1997, pp. 235 – 279.
190. Degussa, Metall Forschung und Entwicklung, Frankfurt/Main, 1991.
191. C. D. Coxe, A. S. Mc Donald, G. H. Sistare Jr.: "Silver-Base Brazing Alloys," in W. H. Cubberly, H. Baker, D. Benjamin (eds.): *Metals Handbook,* 9th ed., **vol. 2,** Properties and Selection: Nonferrous Alloys and Pure Metals, Metals Park, Ohio: American Society for Metals 1979, pp. 675 – 678.
192. H.-M. Lüschow: *Edelmetallanalytik, Eine Übersicht über bewährte klassische und moderne Methoden,* Heft 69 der Schriftenreihe der GDMB, Clausthal-Zellerfeld, 1993.
193. H.-M. Lüschow, H. Schwarz: "Purity Analysis of Silver Nitrate," *Erzmetall* **51** (1998) 195 – 205.
194. M. Brill: "The Quality Control of Alloys from the Precious Metal Foundry with the Aid of the Glow Discharge Lamp Emission Spectrometer," *Proc. 11th IPMI Conf.* (1989) 369 – 400.
195. ISO Standard No. 11427: "Determination of Silver in Silver Jewelry Alloys – Volumetric (potentiometric) Method," 1993.
196. H.-M. Lüschow, R. Schmitt, B. Brageser, 8. Seminar, "Spektrometrie – Metallography," Fa. Spectro Analytical Instruments GmbH, Dillingen 1996.
197. *Analyse der Metalle,* **Suppl. vol. 1,** Springer Verlag, Berlin 1980, pp. 171 – 173.
198. *Analyse der Metalle,* **Suppl. vol. 2,** GDMB-Informationsgesellschaft m.b.H., Clausthal-Zellerfeld, 1993, pp. 185 – 187.
199. *World Silver Survey 1999,* The Silver Institute Washington D.C. 20036.
200. W. Wirth, Ch. Gloxhuber: *Toxikologie,* 4th ed., Georg Thieme Verlag, Stuttgart 1985.
201. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 6th ed., Georg Thieme Verlag, Stuttgart 1986.
202. R. H. Wölbling et al., "Argyrose bei Beschäftigten in der silberverarbeitenden Industrie," *Arbeitsmed. Sozialmed. Präventivmed.* **23** (1988) 293.
203. H. G. Seiler, H. Sigel (eds.): *Handbook of Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988.
204. H. G. Seiler, A. Sigel, H. Sigel (eds.): *Handbook on Metals in Clinical and Analytical Chemistry,* Marcel Dekker, New York 1994.
205. I. M. Smith, B. L. Carson: *Trace Metals in the Environment,* **vol. 2,** Ann Arbor Sci. Publ., AnnArbor, MI 1977.
206. G. D. DiVincenco, C. J. Giordano, L. S. Schriever: "Biological Monitoring of Workers Exposed to Silver," *Int. Arch. Occup. Environ. Health* **56** (1985) 207.
207. G. Perelli, G. Piolatto: "Tentative Reference Values for Gold, Silver, and Platinum: Literature Data Analysis," *Sci. Tot. Environ.* **120** (1992) 93.
208. E. Merian (ed.): *Metals and their Compounds in the Environment,* VCH Verlagsgesellschaft, Weinheim 1991.

Single Cell Proteins

Wolfgang Babel, UFZ Umweltforschungszentrum Leipzig – Halle GmbH, Leipzig, Federal Republic of Germany (Chaps. 1, 2, 3, 7)

Hans-Dieter Pöhland, Leipzig, Federal Republic of Germany (Chaps. 4, 5)

Konrad Soyez, Universität Potsdam, Zentrum für Umweltwissenschaften, Potsdam, Federal Republic of Germany (Chaps. 4, 6, 7)

1. Introduction 1
2. Substrates and Organisms 2
3. Efficiency Considerations in the Selection of Substrates and Microorganisms 4
4. Technology and Equipment 8
5. Products and Uses 10
6. Economic Aspects 16
7. Outlook 19
8. References 20

1. Introduction

The present world population stands at more than 5×10^9 people, about 10 % of whom go hungry. More than 10^9 people in developing countries lack sufficient food to maintain their health and productivity. A clear challenge thus exists not only to overcome current food shortages, but also to provide sufficient food for the population of tomorrow. Actually, food as such is not the problem, because carbohydrates and fats are plentiful. The real problem is proteins (→ Proteins), which supply essential amino acids that cannot be replaced by other substances (→ Amino Acids). A diet deficient in these amino acids results in severe physical damage and disease.

At present, ca. 20 % of the dietary protein for the world's population comes from animal sources, ca. 55 % from grain cultures, ca. 17 % from oilseeds, and ca. 8 % from fruits and vegetables. In general, animal protein has more nutritional – physiological value than plant protein. With less animal protein in the diet, a greater quantity of total protein must be consumed. Therefore, because of an inadequate supply of animal protein (only ca. 9 – 30 % of the total), people in developing countries are forced to consume considerably more plant protein than is readily available.

In the search for new and alternative protein sources, and above all to facilitate the production of protein-rich food under difficult climatic conditions, single- cell organisms showed promise of becoming an attractive solution. In comparison with animals and plants, microorganisms are characterized by higher rates of protein synthesis and greater protein content. They are also less demanding than animals with regard to nutrients. For instance, whereas microorganisms can accept inorganic sources of nitrogen, animals require organically bound nitrogen. The same principle applies to sources of carbon. Furthermore, industrial-scale production of microorganisms can be carried out in reactors using continuous processes that are independent of climatic and weather conditions and require little space (→ Biotechnology). Like animal and plant protein, protein from microorganisms offers both advantages and disadvantages from a qualitative standpoint.

In 1966, C. Wilson at the Massachusetts Institute of Technology coined the term "single cell protein" (SCP) [1] as a way of describing dried cells from bacteria, yeasts, fungi, algae, and higher fungi produced on a large scale as a source of protein for food and feed. The terms *microbial biomass*, *microbial cell substance*, and *bioprotein* are used synonymously.

Unicellular organisms have played an important part in human nutrition for thousands of years: yeasts, especially *Saccharomyces cerevisiae*, in the preparation of wine and beer as well as in baking (→ Yeasts), and bacteria, for instance of the genus *Lactobacillus*, in cheesemaking (→ Cheese, Processed Cheese, and Whey) and in the preservation and processing of natural products [2]. Both of these applications take ad-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_165

vantage of the biocatalytic potential of microorganisms. The utilization of microorganisms as primary producers of proteins and vitamins is a concept of the twentieth century [3], the root of which can be found in the "Vienna process" for the production of baker's yeast (A. J. MAUTNER, ca. 1850).

The first methods for producing SCP were developed in Germany, and are closely associated with the names of J. SCHECKENBACH, M. DELBRÜCK, and W. HENNEBERG. Originally, molasses (World War I) and later sulfite waste liquors served as the raw material [4]. The first yeast factory based on sulfite waste liquor was opened in Björneberg, Finland, in 1928. Conditions suitable for the continuous production of yeast from sulfite waste liquor were established by H. FINK, J. KREBS, and R. LECHNER in the years following 1937. Large-scale production facilities were developed by H. SCHOLLER (mammoth pump system) and M. E. PEUKERT (Phrix system) in the case of wood hydrolyzates, and by W. CLAUS (Waldhof system) for beechwood sulfite waste liquor [5]. Fermenters used today are still based on these three fundamental types. Fodder yeast has been produced industrially in Germany since 1940 on the basis of waste products, including molasses and distillery residues. In 1944, six large-scale plants were producing a total of about 15 000 t/a of dry yeast. (These plants were designed for an annual production of 80 000 t [6].) At this time it was discovered (F. JUST and W. SCHNABEL) that yeast is also capable of growing on hydrocarbons. This finding set the stage for development in the early 1960s of industrial processes based on hydrocarbons (J. C. SENEZ and A. CHAMPAGNAT). The first large pilot plant designed around diesel fuel was operated by British Petroleum in Lavera, France [7–9]. Two 100 000 t plants were erected in Italy, but these were never put into operation [9]. From 1983 to 1990, fodder yeast was produced from petroleum distillate in Schwedt near Frankfurt/Oder, former German Democratic Republic [10]. In the former Soviet Union, production of SCP, yeast, and bacteria from both hydrocarbons and carbohydrates amounted to more than 1×10^6 t/a. In the 1970s, another raw material was introduced for the production of SCP, methanol. Imperial Chemical Industries led in this development and operated a large-scale production facility for several years [7, 11].

A number of processes for the industrial-scale production of SCP have been developed (Table 1), but only a few are currently operating, essentially for economic reasons (see Chap. 6).

2. Substrates and Organisms

In accordance with the elemental composition of cellular material ($C_aH_bO_cN_d \ldots$), its biosynthesis (i.e., growth and multiplication) requires the elements carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur, as well as calcium, magnesium, sodium, etc., (see Chap. 4). Organisms are able to incorporate the structure-forming elements either in their free or molecular form (O_2, H_2, N_2) or in combination with other elements (as NO_3^-, $C_xH_yO_z$, NH_4^+). Carbon is not utilized in its elemental form, but is assimilated as carbon dioxide or in some more or less reduced state ($C_xH_yO_z$).

With CO_2 as the sole source of carbon for growth and multiplication (autotrophic nutrition), an additional energy donor is required (e.g., H_2, Fe^{2+}, or thiosulfate for chemolithotrophic organisms, light in the case of phototrophic organisms). Carbon dioxide is the source of carbon for algae, cyanobacteria, and *Knallgas* bacteria (e.g., *Alcaligenes eutrophus*). The latter were considered as possible sources of SCP in the 1960s [14]. On the other hand, carbon compounds of the type $C_xH_yO_z$ ordinarily serve simultaneously as sources of energy (chemoorganoheterotrophic nutrition), and they are the basis of industrial processes involving bacteria, yeasts, and molds. The synthesis of SCP by growth and multiplication of microorganisms proceeds under aerobic conditions, because yield coefficients are higher than under anaerobic conditions.

Substrates of the type $C_xH_yO_z$ can be of agricultural, forest (carbohydrate), or petrochemical (hydrocarbon) origin. Accordingly, two groups of raw materials can be distinguished: renewable and fossil. The use of renewable (conventional) sources of carbon as substrates for biotechnological processes has the longest tradition. These originate photoautotrophically as starch or cellulose, and are usually made available for biotechnological conversion in the form of waste prod-

Table 1. SCP processes and companies that have exploited them (industrial and pilot scale) [12, 13]

Substrate	Organism	Country and producer	Trade name	Production capacity process scale, $\times 10^3$ t/a	Status
Methanol	*Methylophilus methylotrophus*	United Kingdom, ICI Ltd., Billingham	Pruteen	70	shut down
	Methylophilus methylotrophus	Mexico, Pemes	Amberex	50	planned
	Methylomonas clara	Germany, Hoechst/Uhde, Gelsenberg	Probion	1	pilot plant, irregular
	Methylomonas methanolica	Sweden, Norsk Hydro and AB Marabou Sundbyberg	Norprotein	4.5 ∗	irregular
	Acetobacter methanolicus	Germany, Institute of Biotechnology, Leipzig			pilot plant, has been shut down
	Pseudomonas C	Israel, Hebrew University, Jerusalem			pilot plant
	Pseudomonas sp./yeast	Japan, Mitsubishi Gas Chemicals Co., Niigata		5	irregular
Ethanol	*Candida* (Torula) *utilis*	United States, Pure Culture Products, Minnesota	Zyest, Torutein	7	in service
	C. ethanothermophilum, C. acidothermophilum	Japan, Mitsubishi Gas Chemicals Co., Niigata		0.1	in service
	Hansenula anomala	Germany, Schick-Chemie Technik GmbH		1.5	irregular
	Yeast	Czech Republic, Sovnaft, Kojetin N. Moravia		4	in service
Gas oil	*Candida tropicalis*	France, British Petroleum, Bouches-du-Rhône	Toprina L	16	irregular
	C. guilliermondii	Soviet Union, Government Companies	Fermosin	1000	
Petroleum distillate	*Candida maltosa*	former German Democratic Republic, Petrolchemisches Kombinat Schwedt	Fermosin	60	shut down
Spent sulfite liquor	*Paecilomyces varioti*	Finland, Tampela and Finnish Pulp and Paper Research Institute, Jamsankoski	Pekilo	10	in service
	Candida utilis	Finland, Metsaluton Sellulosa Oy, Aanekoski	Silva	10	in service
Whey	*Kluyveromyces fragilis*	United States, Amber Laboratories, Wisconsin,	Amber, Nutrex	5	in service
		United States, Knudsen Milk Products, Los Angeles	Wheast	0.75	in service
	Kluyveromyces fragilis; Torulopsis bovina, T. sphaerica	France, Bel Fromageries Industries, Vendôme	Protibel	8	in service
Potato water	*Endomycopsis fibuliger; Candida utilis*	Sweden, Swedish Sugar Co., Eslov	Symba	10	irregular
Food-grade glucose	*Eusarium graminearum*	United Kingdom, Ranks Hovis McDougall Ltd., High Wycombe	Mycoprotein	0.1	in service
n-Alkanes	*Candida maltosa*	Italy, Liguichimica Biosintesi SpA., Saline de Montebello	Liquipron	100	irregular
	C. lipolytica	United Kingdom, British Petroleum, Grangemouth	Toprina G	4	irregular
	C. lipolytica	Sardinia, Italoprotein, BP-Anic, Sarroch	Toprina G	100	irregular
	C. tropicalis	United States, Gulk R & D Co. Wasco, California	Petroprotein	50 ∗	irregular
	C. paraffinica	Japan, Dainippon Ink & Chemicals Inc., Tokyo	Viton	120	irregular
	C. pichia	Rumania, Dainippon and Petrom	Roniprot	60	
	Acinetobacter cerificans	United States; Switzerland, Esso-Nestlé	Protecel	0.2 ∗	irregular

∗ Capacity in 10^3 L/a.

ucts from the chemical and food industries or from farming (sulfite waste liquor, molasses, distillery residues, or whey). The easily assimilated substances in these materials are sugars, low molecular mass alcohols, and acids. In the group of fossil (nonconventional) raw materials and petrochemicals, the easily assimilated carbon compounds that have attracted the most attention are the longer-chain *n*-alkanes, but also methane and methanol, where the latter can be prepared in very pure form from natural gas, synthesis gas, or cellulose (via synthesis gas).

The most important microorganisms so far exploited for industrial production are listed in

Table 2; many others have also been identified and tested for their suitability. However, processes are often not based on a single defined microbial species, and the nutrients utilized are often chance mixtures of substances (waste) that do not have a constant composition. In principle, both bacteria and yeasts (molds) are potential candidates for the chemoorganoheterotrophic synthesis of SCP. As a rule, bacteria grow more rapidly and efficiently than yeasts, and they provide a higher content of crude protein, sulfur-containing amino acids, and lysine – but also of nucleic acids, which is unfavorable from a nutritional – physiological viewpoint (see Chap. 5). Bacteria are only about one-tenth the size of yeast cells, and they often have an unpleasant odor. Yeast culture is better established.

The development of any bioindustrial process involving cells or organisms starts with selection. If the organisms are to be used as a catalyst for the synthesis of metabolites (e.g., ethanol, citric acid, amino acids, penicillin, etc.), known as "overflow production", the objective of selection is defined qualitatively by the target product itself. In the case of SCP it is protein that is, strictly speaking, the aim and purpose of the process. Any microorganism is a potential candidate if it is capable of growing on the projected raw material. Factors such as pH and temperature, or prototrophy (demand for supplementary materials, e.g., vitamins, called *suppline*), can all be taken into account in the selection program. One cannot select for attainable protein content and protein quality, or for growth efficiency, but selection for growth rate is possible. Because there is a correlation between growth rate and growth efficiency [15], the most rapidly growing microorganisms are the ones subjected to further screening. These are usually also the ones that hold their own in the competition for substrate in an unprotected process.

3. Efficiency Considerations in the Selection of Substrates and Microorganisms

Efficiency of growth and multiplication are reflected in growth-yield coefficients expressed, e.g., in grams of dry weight formed per gram of substrate consumed. A knowledge of efficiency limits is a valuable starting point in the choice of substrates, the selection of organisms, and the optimization of processes. The maximum possible yield of biomass depends on the substrate and, for a given substrate, on the organisms selected (i.e., it is genotypically determined; see Table 3).

A yield can be calculated in advance [16, 17] on the basis of a division of the growth and multiplication processes along metabolic lines as follows:

Synthesis of a central precursor, which serves as the starting point for synthesis of all cell constituents

$$C_xH_yO_z \longrightarrow \text{Central precursor} + E_a\ (+2H + CO_2^{a1})$$

Synthesis of cellular material, anabolism

$$\text{Central precursor} + NH_3 + E\ (+2H) \longrightarrow C_aH_bO_cN_d + CO_2^{a2}$$

Energy (E) for the synthesis of cellular material can be made available during the synthesis of the precursor (assimilation, E_a). If this amount of energy is insufficient for maximum conversion of precursor carbon to cellular material, then more substrate must be consumed (dissimilated) solely for the purpose of supplying energy (E_d):

Generation of "biological" energy by complete oxidation of the substrate

$$C_xH_yO_z + O_2 \longrightarrow CO_2^d + E_d + H_2O$$

(CO_2^a = assimilatory CO_2 loss; CO_2^d = dissimilatory CO_2 loss; $E_{a,d}$ represent ATP and/or energetically valuable reduction equivalents, and 2H stands for bioenergetically valueless reduction equivalents; $E = E_a + E_d$)

The fictitious cell molecule $C_aH_bO_cN_d$ can be regarded as $C_4H_8O_2N$ in the case of bacterial biomass and $C_6H_{10}O_3N$ for yeast biomass. Differences in these specifications that arise from biomasses of differing composition can be neglected [16], because their effects on cell yield calculations are small, although the energy expenditure required for synthesis of the primary components (proteins, carbohydrates, lipids, and nucleic acids) varies greatly. (The synthesis of 1 g of protein or 1 g of nucleic acid consumes about three times as much energy as the synthesis of 1 g of carbohydrate [16, 18].) For protein and carbohydrate content ranges of

Table 2. Substrates and microorganisms frequently used or considered for SCP synthesis

Substrate	Yeasts	Bacteria	Algae
Carbohydrates			
Molasses	*Candida utilis*		
	Saccharomyces cerevisiae		
Spent sulfite liquors	*C. utilis*		
	C. tropicalis		
	Paecilomyces varioti		
	Trichosporon sp.		
Whey	*C. intermedia*		
	Kluyveromyces fragilis		
	Torulopsis bovina		
	T. sphaerica		
Hydrocarbons	*Candida guilliermondii*	*Acinetobacter cerificans*	
	C. lipolytica	*Pseudomonas* sp.	
	C. maltosa		
	C. paraffinica		
	C. pichia		
	C. tropicalis		
	Lodderomyces elongisporus		
Ethanol	*Candida acidothermophilum*		
	C. ethanothermophilum		
	C. utilis		
	Hansenula anomala		
Methanol	*Candida boidinii*	*Methylomonas clara*	
	Hansenula polymorpha	*Methylomonas methanolica*	
	Pichia pastoris	*Methylophilus methylotrophus*	
	P. pinus	*Pseudomonas C*	
		Acetobacter methanolicus	
Methane (natural gas)		*Methylococcus* sp.	
		Methylosinus sp.	
Carbon dioxide			*Chlorella* sp.
			Dunaliella sp.
			Micractinium sp.
			Scenedesmus sp.
			Spirulina sp.

40 – 55 % and 20 – 35 %, respectively, the calculated yields of yeast dry mass are 0.35 – 0.38 g per gram of methanol, and 0.46 – 0.49 g per gram of glucose [19].

If the expenditure of energy for the synthesis of cell substance from a central precursor is regarded as approximately constant and independent of substrate, then the amount of energy supplied in the course of assimilation (E_a) is decisive in establishing whether the potential growth efficiency based on carbon metabolism can actually be achieved, or whether the growth yield is energy limited.

E_a is determined by the substrate and, for a given substrate, by the metabolic pathway, as shown by the balanced equations for synthesis of a central precursor (3-phosphoglyceric acid, PGA, in this case):

$$0.5\ \text{Glucose} \xrightarrow{\text{EMP}} 1\ \text{PGA} + 1\ \text{NAD(P)H}$$

$$0.5\ \text{Glucose} \xrightarrow{\text{ED}} 1\ \text{PGA} + 1\ \text{NAD(P)H} - 1\ \text{ATP}$$

$$3\ \text{Formaldehyde} \xrightarrow{\text{serine}} 1\ \text{PGA} - 3\ \text{ATP} + 1\ \text{FADH}_2$$

$$3\ \text{Formaldehyde} \xrightarrow{\text{HuP}} 1\ \text{PGA} + 1\ \text{NAD(P)H} - 2\ \text{ATP}$$

(EMP = Embden – Meyerhof – Parnas pathway; ED = Entner – Doudoroff pathway; serine = serine pathway, isocitrate-lyase positive variant; HuP = hexulose phosphate pathway, fructose-1,6-bisphosphate variant.)

The experimentally obtained yield coefficients (Y values) of ca. 0.35 g or 0.5 g of bacterial dry weight per gram of methanol (serine and hexulose phosphate pathways, respectively) are undoubtedly attributable to this factor (see Table 3) and not to differences in the efficiency

Table 3. Theoretical and experimental growth yield coefficients [grams of biomass dry weight per grams of substrate (Y) or per grams of oxygen ($Y_{X/O}$)] [16]

Substrate	Bacteria					Yeasts				
	$Y_{85\,\%}$	$Y_{75\,\%}$	Y_{exp} %	Y_{calc} %	$Y_{X/O}$	$Y_{85\,\%}$	$Y_{75\,\%}$	Y_{exp} %	Y_{calc} %	$Y_{X/O}$
Methane	1.35				0.62					
				0.58 [a] (36.5 %)	0.18					
				0.72 [b] (45 %)	0.23					
			0.9 [c]		0.32					
Methanol	0.68				1.14	0.64				1.04
				0.6 [b] (75 %)	0.85				0.41 (55.2 %)	0.45
			0.5		0.6			0.38		0.40
Ethanol		0.83		0.83	0.85		0.78		0.78	0.78
			0.8		0.78			0.77		0.76
Acetic acid		0.64			2.95		0.6			2.57
				0.41 (47.7 %)	0.77				0.39 (49.2 %)	0.76
			0.4		0.77			0.35		0.60
Glucose	0.72				7.08	0.68				5.55
				0.52 (61.4 %)	1.42				0.51 (63.4 %)	1.40
			0.5		1.24			0.5		1.34
Hexadecane		1.35		1.35	0.82		1.27		1.27	0.75
			1.1		0.85			0.9		0.41

Y = grams of dry substance per gram of substrate at carbon conversion of 85 or 75 %; Y_{exp} = observed yields; Y_{calc} = yields calculated according to the Y_{ATP} concept [16–18]; numbers in parentheses: extent of carbon conversion based on the hypothetical cell molecule $C_4H_8O_2N$ for bacteria and $C_6H_{10}O_3N$ for yeast. [a] Serine pathway. [b] Hexulose phosphate pathway. [c] Mixed culture.

of energy recovery via electron-transport phosphorylation.

In the case of glycolytic substrates (C_{3n}), the carbon conversion efficiency is ca. 85 %, and for nonglycolytic substrates (gluconeogenetic, C_{2n}), ca. 75 % [19]. Therefore, all substrates for chemoorganoheterotrophic growth are more or less energy deficient (see Table 3). Ethanol is an exception, and probably provides a surplus of energy [20]. Based on the ratio $Y_{exp}/Y_{100\,\%}$, ethanol would be the substrate of choice for the production of SCP if all possible substrates were freely available. This is not necessarily the case, however, when the cost of raw materials, possible productivities, and growth-yield coefficients based on oxygen (Table 3) are included in the economic considerations.

In principle, yields in excess of the Y_{calc} values listed in Table 3 are possible. One approach to this end is to combine substrates in such a way that, in the course of their simultaneous utilization, a minimum amount of carbon is dissimilated for the purpose of supplying energy. This is apparently the case with the mixture sucrose – acetic acid – ethanol (Table 4), in which ethanol cannot possibly compensate for the energy deficits of sucrose and acetic acid. Instead, as in the case of a methanol – glucose mixture, the observed auxiliary substrate effect must be a result of metabolic channeling ("division of function") [22].

Another approach is provided by genetic engineering. In the case of *Methylophilus methylotrophus*, replacing the energetically more elaborate glutamine synthetase – glutamate synthase cycle for the fixation of ammonia nitrogen by glutamate dehydrogenase led to a ca. 5 % increase in the growth yield [23].

An even more pronounced effect (up to 30 %) is to be expected from manipulation of the carbon metabolism (e.g., if the pyrroloquinoline quinone (PQQ) dependent methanol dehydrogenase in bacteria [24] or the energetically worthless methanol oxidase in yeasts were to be replaced by a nicotinamide adenine dinucleotide (NAD) dependent methanol dehydrogenase [16]).

The actual objective is, of course, the production of protein, but more syntheses take place in the course of chemo- and phototrophic nutrition. Other essential structural and functional components of cellular and organismic systems are the carbohydrates, lipids, and nucleic acids mentioned above. Irrespective of whether microorganisms synthesize their cell substance from carbon dioxide or hydrocarbons, for example, es-

Table 4. Growth yield coefficients with *Candida* and *Saccharomyces* species on single substrates and a substrate mixture [21]

Substrate	g Dry substance per g substrate	g Dry substance per g carbon	Carbon conversion efficiency, %
Sucrose	0.5	1.19	60
Acetic acid	0.35	0.88	44
Ethanol	0.75	1.44	72
Mixture *			
Expected		1.14	57
Obtained		1.35 **	67.5

* Mixture ratio: 0.7 g sucrose – 1.0 g acetic acid – 0.6 g ethanol.
** Auxiliary substrate effect: 118 % compared to the additive (expected) value.

sentially the same macromolecules or polymers are always formed despite chemically very diverse precursor carbon skeletons. In all cases, optimally growing and multiplying microorganisms (unlike resting cells) are composed of relatively constant proportions of the primary components listed. This fundamental observation is not affected by the fact that bacteria differ from yeasts in, e.g., their (crude) protein content.

Thus, protein content would appear to be independent of substrate. However, there are indications that protein content increases with substrates that are richer in energy [25], and can also be increased by improving the energy – carbon ratio [26]. For a given organism – substrate combination, the protein content is highest (with continuous growth according to the chemostat principle) when the growth rate is limited by the carbon source.

Protein quality (amino acid composition) is a function of genotype. It may also be phenotypically influenced by the substrate and, in the broadest sense, by nutritional or culture conditions. This is because, depending on the substrate, enzymes with variable amino acid spectra may serve to catalyze the same event. Nevertheless, it is highly improbable that such changes in enzyme content would lead to a significant increase in a particular amino acid.

In the following considerations, the essential amino acid lysine will serve as an example to substantiate this assumption. In principle, the lysine content of the total protein in bacteria or yeasts could be increased either by increasing the lysine content of individual proteins (enzymes), or by introducing an additional protein consisting exclusively of lysine, whereby this polylysine would need to represent a relatively high fraction of the total protein present. Lysine is coded by the triplets AAA and AAG. Synonyms for lysine arise through mutation (transition, transversion) in the codons for arginine, asparagine, glutamine, threonine, and glutamic acid (one base, two triplets each), and isoleucine and methionine (one base, one triplet each). This means that point mutations could increase the lysine content in defined proteins, especially at the expense of arginine (16), asparagine (13), glutamine (9), threonine (7), and glutamic acid (6). (The numbers in parentheses indicate the frequency of observed replacements.) Glycine and proline, as well as alanine and serine, have also been found to be exchanged for lysine in homologous proteins, although in these cases mutations are required in more than one base.

The lysine content of proteins varies between ca. 2 g and 12 g per 100 g of protein [i.e., ca. 35 to 90 lysine residues for every 1000 amino acids (molecular mass ca. 110 000) in a protein (enzyme)]. If the lysine content of a specific protein is to be increased by 1 g/1000 g, ca. 6 mutations are required based on 1000 positions. Since the lysine content of the total protein is to be increased, exchange of this type would be necessary within each of at least 1000 different proteins, or at least in one or several of the proteins representing a high proportion of the total protein content. Another requirement for success is that any mutation or exchange must not be lethal; in other words, it must result in functioning, isodynamic enzymes. If the lysine content were to be increased at the expense of the above-mentioned amino acids, then there would certainly be an observable effect on growth efficiency. In addition, increased incorporation presupposes a higher rate of synthesis, which means not only that the genetic code must be "prepared" for the increased incorporation, but the metabolic regulation must allow it (feedback regulation).

4. Technology and Equipment

SCP synthesis is achieved through the growth and multiplication of microorganisms. Establishing an optimal set of conditions (phenotypic optimization) requires that the performances (in terms of yield, rate, etc.) of the microorganisms must be investigated first on a laboratory scale as a function of such conditions as pH, temperature, and residence time. Decisions can also be made at this stage about whether a protected (sterile) or open process is preferable, and whether it should be carried out in batch, fed-batch, or continuous mode. In general, continuous processes are preferred for SCP synthesis. Chemostatic cultivation – in which residence time is adjusted via the carbon substrate feeding rate – is optimal with respect to the target function. Gradients in the substrate supply must be avoided, because these constitute disturbances, and would lead to reduced yields.

For microorganisms that grow effectively in the neutral pH range, protection (air filters, sterilization of the nutrient solution, etc.) is essential to prevent contamination by unwanted organisms.

Table 5. Inorganic additives for a typical nutrient solution used to culture microorganisms

Compound	Additive – dry substance ratio, mg/g
NH_4OH	80 – 135
H_3PO_4	15.00
KOH	10.00
$MgSO_4 \cdot 7\,H_2O$	2.00
$CuSO_4 \cdot 5\,H_2O$	0.07
$CoSO_4 \cdot 7\,H_2O$	0.02
$MnSO_4 \cdot 5\,H_2O$	0.06
H_3BO_4	0.05
$FeSO_4 \cdot 7\,H_2O$	0.14
$ZnSO_4 \cdot 7\,H_2O$	0.07
$Na_2MoO_4 \cdot 2\,H_2O$	0.02

Semitechnical and pilot-scale experiments provide the data required for scale-up, together with process parameters important in the design of a large-scale plant.

An industrial production facility usually consists of separate stations for the following functions:

1) Preparation of the substrate and the nutrient medium
2) Substrate conversion (by growth, fermentation)
3) Separation and concentration of the product
4) Evaporation and drying
5) Product formulation

Not all substrates require "preparation". If the raw materials contain starch and cellulose, preparation involves digestion (acid- or enzyme-catalyzed hydrolysis), pH adjustment, and, if necessary, removal of growth-inhibiting substances (e.g., SO_2 or furfural from sulfite waste liquor [5]).

Preparation of the nutrient medium involves mixing and dissolution. All substances that the production organism cannot itself synthesize must be added with the nutrient solution. In most cases, the microorganisms are prototrophic, so only the carbon and energy source is of an organic nature. Essential ions and their required levels are determined from cell analyses (Table 5).

The synthesis of cell substance (fermentation, growth) takes place in fermenters, and involves two basic process engineering operations:

1) Stirring and mixing of a multiphase system (solid – liquid – gaseous; i.e., microorganisms, aqueous fermentation solution, and air)
2) Transport of oxygen in the dissolved state from the gas bubbles, through the fermentation liquid, to the microorganisms

Supplying dissolved oxygen to the microorganism cell (i.e., the rate of oxygen transfer) is usually the growth rate-limiting step that determines the efficiency of an overall fermentation system.

Stirred-tank, tubular loop, and deep jet reactors [27] have all been employed for the production of SCP because they provide high oxygen-transfer and dispersing efficiencies [28]. Deep jet fermenters transport up to 12 kg of O_2 per ton of fermenter content per hour without pressure, and 50 kg of O_2 per ton of fermenter content per hour at pressures up to 0.1 MPa. Reactor volumes up to several thousand cubic meters are feasible. The choice of conditions must of course take into account technological and reaction engineering criteria, but also biological factors such as the fragility of microorganism cells in the shear field of a stirrer, cell response to inhomogeneities in the fermenter (gradients

Figure 1. Facility for the production of Fermosin [former VEB Petrochemisches Kombinat Schwedt/Oder, German Democratic Republic, in operation from 1985 to 1990; output 55 500 t/a; equipped with two IZ (Ingenieurtechnisches Zentralbüro Böhlen) air-jet fermenters with a reactor volume of 2200 m^3]

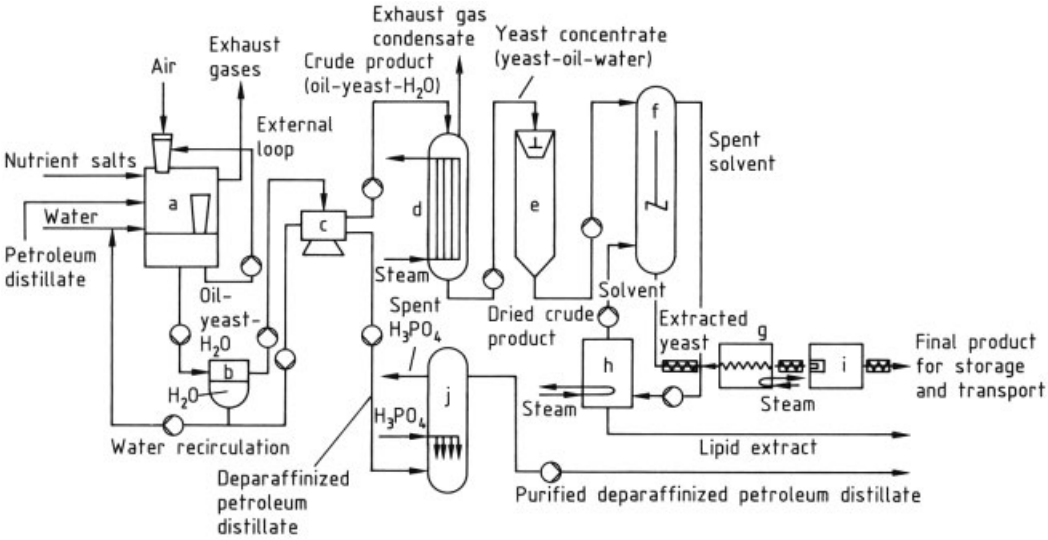

Figure 2. Process scheme for the synthesis and recovery of Fermosin [10]
a) Fermentation; b) Emulsification; c) Separation; d) Evaporation; e) Drying; f) Extraction; g) Solvent separation; h) Solvent reprocessing; i) Pelletization; j) Purification of deparaffinized petroleum distillate

with respect to substrates in the broadest sense, and to pH, temperature, etc.), and the behavior of cells in the inner and outer circuits of the fermentation system.

Typical biomass concentrations in a reactor range from 20 – 50 kg of biomass dry matter per ton of fermenter content (medium), with a maximum of 150 kg/t. Standard separation techniques are employed (e.g., flotation, sedimentation, separation, centrifugation, decantation, filtration) depending on the physical properties of the microorganism cells and the suspen-

sion (density, size, floc diameter). The result of separation is a suspension containing 15 – 22 % dry matter. Further concentration to ca. 25 % is achieved by evaporation. The most common drying methods are spray, fluidized-bed, and freeze drying. The product is delivered to the consumer either as a powder or pelletized. Purification and further refinement (e.g., extraction) are also included in the overall process.

Control of the fermentation (i.e., growth) requires monitoring the pH, temperature, and oxygen and ion contents of the solution, as well as the CO_2 concentration in the exhaust air. To some extent, biosignals (e.g., levels of cytochrome c, NAD^+ – NADH, etc.), are also employed [28]. The type of control depends on the selected organism – substrate system, with possible types being run-off and real-time control.

Basic technology for SCP production is sized according to the availability of raw materials and the intended product output. Facilities designed for the production of Fermosin are illustrative of a typical large-scale plant (Figs. 1 and 2) [10, 29]; see also Table 1.

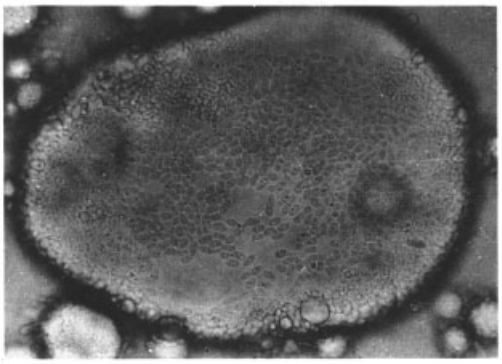

Figure 3. Cells of the production strain *Candida maltosa* on a droplet of petroleum distillate (original size ca. 3×6 µm)

In the Fermosin process, a highly productive strain of the yeast *Candida maltosa* (Fig. 3) is propagated continuously on a straight-run petroleum distillate (boiling range 240 – 360 °C) at 32 °C and pH 4.2, with a residence time of 5 h. The growth stage is followed by decantation (removal of 60 % of the process water, ca. 90 % of which is returned to the process), multistage separation (oil – water, biomass – water), evaporation, spray drying, and extraction (with a methanol – ethanol – water mixture). The main product is Fermosin (for quality information, see Table 6 and also Tables 10, 11, 12). Other products include a biolipid extract containing fatty acids, phosphatides, glycerides, sterols, and waxes [35], and a deparaffinized petroleum distillate, which can be used as a low-temperature diesel fuel.

Apart from the production of SCP in large industrial plants, attention should also be directed in the future to small, decentralized, village-level plants with high productivity. Such plants could satisfy localized feed requirements, and simultaneously exploit locally produced raw materials and industrial or agricultural waste (biogas, low-grade petroleum, waste methanol, distillery residues, and whey). They offer economic advantages because of savings in transport costs, potential local use of SCP production wastes (CO_2 in greenhouses, waste heat), and direct application of process effluent SCP (liquid feeding; for cost projections see Chap. 6). Opportunities for SCP production also exist in less developed regions. An example is a facility for utilization of the fruit from *Ceratonia siliqua* by the fungus *Aspergillus niger* [30]. Process steps involved in this case are hot-water extraction of the sugar, addition of ammonium sulfate and dihydrogenphosphate, batch fermentation, thickening in a vacuum rotary filter, and drying in a ring dryer. A 3 m^3 stirred-tank reactor is used, and the result is a storable product [30].

5. Products and Uses

The final product of an SCP process is biomass. Apart from proteins, which are the primary constituents, it contains nucleic acids, lipids, and carbohydrates (Table 6), as well as vitamins and coenzymes (Table 7). This gives rise to market possibilities beyond the original objective (Table 8).

Moisture, ash, and crude protein content are important parameters for characterizing the commercial product. A crude protein assay (nitrogen content×6.25) encompasses all nitrogen-containing compounds (proteins, nucleic acids, etc.); the reported ash value includes all mineral components as their oxides (Table 9). Protein concentrations range from 45 – 80 %. Apart from strain- and process-dependent differences, bacteria have higher crude protein values (65 –

Table 6. Cell compositions of various strains of microorganisms compared with animal feeds (as a percentage of biomass dry matter, average values) [30–34] ∗

	Moisture	Ash	Crude protein (N×6.25)	Lipids	Carbohydrates	Nucleic acids	Phosphate
IUPAC recommendation	<10	<10	>45	n.d.	n.d.	n.d.	n.d.
MBC requirement ∗	<10	<10	>55	<8	n.d.	<10	n.d.
Microorganisms/growth medium							
Yeast/carbohydrates (fodder yeast plants in the former German Democratic Republic)	7.0	8.0	45.0 – 54.0	2.1	20.5	7.4	1.6
Yeast/ethanol: Torutein (Amoco Foods, United States)	6.0	9.0	52.0	6.0		10.0	2.1
Yeast/methanol	5.0	6.1	46.0 – 61.0	5.0		6.2	
Bacteria/methanol (acidophiles) (Institute of Biotechnology, Leipzig)	10.8	6.1	69.0	6.6	22.9	8.7	1.3
Bacteria/methanol (neutrophiles) (Institute of Biotechnology, Leipzig)	3.6	9.5	81.0	6.8	8.7	12.4	1.2
Bacteria/methanol: Pruteen (neutrophiles), (ICI, United Kingdom)	10.0	8.2	74.0	8.5		14.0	2.4
Bacteria/methanol (neutrophiles) (Hoechst/Uhde)		8.0	80.0	9.0		12.0	
Bacteria/methane (Institute of Biotechnology, Leipzig)	6.0	7.0	65.7	4.6		8.0	2.2
Bacteria/methane (Shell, United States)			60.0 – 70.0	10.0			
Yeast/*n*-alkanes: Toprina (BP, United Kingdom/France)	4.6	6.3	60.0 – 65.0	8.1		8.0	1.5
Yeast/*n*-alkanes (Kanegafuchi, Japan)	4.5	9.8	53.0 – 62.0	3.2	27.9		2.0
Yeast/diesel fuel Fermosin (Schwedt, former German Democratic Republic)	6.8	7.5	68.0 – 72.0	4.0	22.0	6.6	
Fungi/spent sulfite liquors: Pekilo (Finland)	4.0	5.5	55.0 – 60.0	1.0	25.0	11.0	2.0
Algae (*Chlorella*)	7.0	8.3	55.0	7.5	17.8	4.5	
Fish meal	10.0	15.5	63.0 – 69.0	2.6	5.9		2.8
Soy extract, coarse meal	12.0	5.7	45.0 – 49.0	3.0	34.6	3.0	0.7

n.d. = not declared. ∗ Medical – Biological Committee of the former Soviet Union – German Democratic Republic for the assessment of SCP.

Table 7. Vitamin and coenzyme concentrations in various strains of microorganisms (mg of vitamin per 100 g of biomass dry matter) [31]

Microorganism	Thiamine (B_1)	Riboflavin (B_2)	Pyridoxine (B_6)	Nicotinic acid	Pantothenic acid	Biotin	Folic acid	Choline	Inositol	B_{12}
Fodder yeast/hydrolyzate	0.5 – 2.0	4.0 – 12.7	1.0 – 2.0	40 – 55	6 – 10	0.06 – 0.23	n.r.	250 – 450	120 – 480	
British Petroleum protein	0.4	18.0	2.5	43	12	n.r.	0.64	830	327	
Paecilomyces: Pekilo	0.6	6.0	2.0	45	4	0.2				
Scenedesmus – *Chlorella* mixture	1.2	2.7	1.1	11	5	0.2	73	n.r.	n.r.	0.1

n. r. = not reported.

Table 8. Potential uses of microbial biomass

Cell component	Cell mass content, %			Process	Product applications
	Bacteria	Yeast			
Whole cells Protein	≤80	40 – 60	purification, separation, digestion, degradation (enzymatic, chemical)	feed	animal feed
				food protein, intracellular enzymes	food additives (curd, sausage, cheese)
Nucleic acids	10 – 15	5 – 8	separation, enzymatic degradation	RNA, DNA, nucleotides	flavor enhancers
Carbohydrates	15 – 20	20 – 35	decomposition, modification (enzymatic, chemical)	saccharides, sugars	biosubstrates, emulsifiers, surfactants, sweeteners
Lipids ∗	15 – 20	10 – 30	extraction, hydrolysis	glycerides, fats, phosphatides, vitamins (D, A)	food, cosmetics, chemicals, emulsifiers, pharmaceuticals

∗ Dependent on substrate and growth conditions.

Table 9. Comparison of the inorganic components of various strains of microorganisms (average values) [30–34, 36]

Microorganisms/Feed	Dry substance, g/kg				Dry substance, mg/kg			
	Ca	Mg	Na	K	Fe	Cu	Zn	Mn
Yeast/carbohydrates (fodder yeast plants in the former German Democratic Republic)	n.r.	2.2	1.3	32	390	24	123	100
Bacteria/methanol (acidophiles) (Institute of Biotechnology, Leipzig)	0.15	2.0	0.2	7	300	35	42	75
Bacteria/methanol (neutrophiles) (Institute of Biotechnology, Leipzig)	1.2	5.3	1.2	17	340	44	170	100
Bacteria/methanol: Pruteen (neutrophiles), (ICI, United Kingdom)	0.7	2.2	15.0	2	150	21	26	
Bacteria/methane (Institute of Biotechnology, Leipzig)	0.1	2.7	0.5	5	370	380	270	90
Yeast/*n*-alkanes: Toprina (BP, United Kingdom/France)		2.2		2.7	250	5	n.r.	25
Yeast/diesel fuel: Fermosin (Schwedt, former German Democratic Republic)		1.5		8.5	460	83	695	158
Fungi/spent sulfite liquors: Pekilo (Finland)		1.4		17	100	20	180	130

n.r. = not reported.

85 %) than, for example yeasts because of higher nucleic acid concentrations (Table 6), which correlate directly with their high growth rate [15]. Protein quality is determined to a great extent by the concentrations and proportions of essential amino acids (Table 10).

If SCP is to be used as food, or even as feed, proof must be provided of its safety (e.g., lack of toxicity, mutagenicity, carcinogenicity, pathogenicity) and its organoleptic suitability. SCP can be used as animal feed without further treatment. However, it must always be mixed with energy-rich fodder (e.g., grain). Such mixtures should be well balanced from a nutritional–physiological and energetic standpoint. The material composition of microbial biomass assures proper balance in a mixture containing 5 – 20 mass % SCP.

Factors important with respect to the nutritional – physiological value of SCP include the true protein concentration, digestibility, protein efficiency ratio (PER), net protein utilization (NPU), and biological value (BV) (Table 11). The biological value is determined by the amino acid composition and the concentrations of essential amino acids. Even these factors are

Table 10. Comparison of amino acid compositions (grams of amino acid per 16 grams of nitrogen) for various strains of microorganisms [30–34, 36]

	Lysine	Cysteine	Methionine	Histidine	Arginine	Asparagine	Threonine	Serine	Glutamic acid	Proline	Glycine	Alanine	Valine	Isoleucine	Leucine	Tyrosine	Phenylalanine	Tryptophan
IUPAC recommendation	6.5	1.4																
MBC requirement *	7.0		$\sum > 2.0$															
Microorganisms/feed																		
Yeast/carbohydrates (fodder yeast plants in the former German Democratic Republic)	7.3	0.9	1.5	2.1	4.1		4.3						4.4	4.6	6.9	2.8	4.2	
Yeast/ethanol: Torutein (Amoco Foods, United States)	6.7	0.7	2.1	2.1	5.5	8.5	4.9	4.4	16.3	2.9	4.3	5.2	5.3	4.3	6.7	3.2	4.1	
Yeast/methanol	8.1	0.7	1.5	2.4	5.6	10.5	5.2	5.1	14.1	n.r.	5.2	6.1	6.2	5.1	8.3		5.0	
Bacteria/methanol (acidophiles) (Institute of Biotechnology, Leipzig)	5.2	0.5	2.6	2.1	6.3	8.4	4.1	3.0	9.2	3.6	5.3	6.7	5.2	3.7	7.3	3.3	3.5	1.3
Bacteria/methanol (neutrophiles) (Institute of Biotechnology, Leipzig)	5.4	0.5	2.2	1.4	3.1	8.2	3.8	3.0	8.8	2.9	5.1	6.8	4.7	3.9	6.1	2.6	3.1	
Bacteria/methanol: Pruteen (neutrophiles), (ICI, United Kingdom)	6.2	0.7	2.4	1.8	4.7	8.8	4.7	3.1	10.4	3.2	5.4	6.9	5.7	4.7	7.2	3.1	3.6	
Bacteria/methanol (neutrophiles) (Hoechst/Uhde)	6.2		$\sum 5.1$	2.0	4.6		4.0				4.8		4.9	4.5	6.3		$\sum 5.1$	

Table 10. (Continued)

	Lysine	Cysteine	Methio-nine	His-tidine	Arg-inine	Aspar-agine	Threo-nine	Serine	Glu-tamic acid	Proline	Glycine	Alanine	Valine	Iso-leucine	Leucine	Tyrosine	Phenyl-alanine	Tryp-to-phan
Bacteria/methane (Institute of Biotechnology, Leipzig)	5.8	0.7	2.6	2.4	5.7	9.2	4.9	3.8	10.1	4.5	5.3	7.5	5.9	4.5	7.6	3.4	4.5	1.4
Bacteria/methane (Shell, United States)	5.7		$\sum$ 3.3	2.2	6.2	n.r.	4.6	n.r.	n.r.	n.r.	n.r.	n.r.	6.5	4.3	8.1	8.4		
Yeast/*n*-alkanes: Toprina (BP, United Kingdom/France)	7.8	0.9	1.6	2.1	5.0	10.0	5.4	5.1	12.1	3.7	4.5	5.8	5.8	5.3	7.8	1.3	4.8	
Yeast/*n*-alkanes (Kanegafuchi, Japan)	7.1	1.4	0.9	1.6	4.6	n.r.	5.1	n.r.	n.r.	n.r.	n.r.	n.r.	5.3	4.5	7.0	2.3	3.9	
Yeast/diesel fuel: Fermosin (Schwedt, former German Democratic Republic)	7.6	0.8	1.6	2.5	4.7	10.5	5.8	5.3	11.3	4.2	4.7	5.9	5.9	5.3	8.2	3.7	5.0	
Fungi/spent sulfite liquors: Pekilo (Finland)	6.6		1.6				4.5						5.2	4.1	7.2		3.8	
Fish meal	4.2		$\sum$ 2.6	1.5	3.6	n.r.	3.0	n.r.	n.r.	n.r.	n.r.	n.r.	3.7	3.2	5.0		$\sum$ 5.2	
Soy extract, coarse meal	2.8		$\sum$ 1.3	1.1	3.2	n.r.	1.9	n.r.	n.r.	n.r.	n.r.	n.r.	2.3	2.2	3.4		$\sum$ 3.8	
Spirulina platensis	5.0	0.9	2.6	1.7	6.9	9.5	5.8	5.7	13.1	4.3	5.1	8.0	7.0	6.5	9.2	1.6	4.6	

n.r. = not reported. ∗ Medical – Biological Committee of the former Soviet Union – German Democratic Republic for the assessement of SCP.

Table 11. Nutritional – physiological values of various strains of microorganisms [31, 36, 37]

Microorganisms	Digestibility, % (t = true) [a]	PER [b], %	NPU [c], %	BV [d], %
Yeast/carbohydrates	84 – 86	0.9 – 1.4		45
(fodder yeast plants in the former German Democratic Republic), *Candida* sp. (mixed)				88 [e]
Bacteria/methanol (acidophiles)	95 – 100 (t)			55 – 75
(Institute of Biotechnology, Leipzig), *Acetobacter methanolicus*				75 – 88 [d]
Bacteria/methanol (neutrophiles)	80 – 84			77 – 84
(Institute of Biotechnology, Leipzig), *Pseudomonas* sp.				
Bacteria/methanol: Pruteen	87 – 91	1.7 – 2.5		79
(neutrophiles), (ICI, United Kingdom)				
Bacteria/methane	90 – 95 (t)			55 – 65
(Institute of Biotechnology, Leipzig), *Methylocyclus curvata*	95 [e] (t)			80 – 88 [e]
Yeast/*n*-alkanes	88 – 93 (t)	1.8 – 2.2	57 – 64	71
(BP, United Kingdom)				
Yeast/diesel fuel: Fermosin	70 – 90			44
(Schwedt, former German Democratic Republic)				90 [e]
Spirulina maxima	76 – 77	1.8	46 – 50	60 – 65
Fish meal	85 (t)	1.9	66	79
Soy extract, coarse meal	85 (t)	1.8	60	70
Casein	99	3.0		100

[a] The digestibility of a protein may be expressed either as a "true digestibility," in which case allowance is made for fecal nitrogen loss on a protein-free diet, or as an "apparent digestibility," in which no such correction is made.

[b] $\mathrm{PER} = \mathrm{protein\,efficiency\,ratio} = \frac{\mathrm{weight\,gain}}{\mathrm{protein\,consumed}}$

[c] $\mathrm{NPU} = \mathrm{net\,protein\,utilization} = \frac{(\mathrm{body\,nitrogen\,of\,test\,group}) - (\mathrm{body\,nitrogen\,of\,group\,fed\,nonprotein\,diet})}{\mathrm{nitrogen\,consumed}}$

[d] BV = biological value = (resorbed nitrogen from food) − (urinary nitrogen − endogenous urinary nitrogen)

[e] +3 % methionine.

insufficient for total SCP evaluation, however. Quality guidelines and additional criteria (e.g., for such toxicologically relevant substances as heavy metals, benzopyrenes, and hydrocarbons, see Table 12), have been established by IUPAC and the Protein Advisory Group (PAG) of the United Nations Food and Agricultural Organization (FAO) [38].

In order to determine the nutritional – physiological value of biomass, experiments are conducted in stages according to established protocols. Laboratory animals (preferably rats) are used in the first stage and farm animals in the subsequent stages. These experiments are designed to monitor feed intake, nitrogen balance, and increase in live mass (see Table 13), as well as reproduction and fertility.

In principle, SCP can also be used directly in human nutrition. The value of the product is comparable to meat products because of its high protein content and favorable amino acid composition. However, this use of SCP requires that it also comply with other nutritional – physiological criteria [39, 40] (e.g., content of nucleic acids, fatty acids, and contaminants), that its formulation be appropriate, and that the costs be lower than those of rival plant products (e.g., soy protein).

6. Economic Aspects

The price of SCP varies from 260 to 1030 €/t (1993). For any given set of process parameters (see Table 14), production costs are a function mainly of the price of raw materials, energy, and plant. Process-dependent cost factors include substrate cost, operating expenses, and depreciation. Sample calculations are provided in Tables 15 and 16. Despite process-dependent differences, the cost of the substrate has the greatest influence on the profitability of an SCP process. This represents about one-third of the total cost if petroleum distillates are the carbon source, and

Table 12. Comparison of toxicologically relevant inorganic and organic components of various strains of microorganisms (average values) [33]

	mg per kg of dry substance					µg per kg of dry substance	g per kg of dry substance	
	Pb	Hg	As	Se	Cd	3,4-Benzopyrene	Hydrocarbons	Aromatic hydrocarbons
IUPAC recommendation	5	0.1	5	n.d.	n.d.	3.0	n.d.	0.5
MBC requirement ∗	≤5	≤0.1	≤2.0	n.d.	1.0	3.0	5.0	n.d.
Microorganisms/feed								
Yeast/carbohydrates (fodder yeast plants in the German Democratic Republic)	3.3	0.11	1.0–4.8	0.9	5.3	1.1		
Bacteria/methanol (acidophiles) (Institute of Biotechnology, Leipzig)	1.8	0.01	0.03	0.07	0.02			
Bacteria/methanol (neutrophiles) (Institute of Biotechnology, Leipzig)	1.2	0.02	0.03	0.08	0.01	1.4		
Bacteria/methanol: Pruteen (neutrophiles) (ICI, United Kingdom)	0.05	0.01	0.01	0.4	0.1	0.3		
Bacteria/methane (Institute of Biotechnology, Leipzig)	0.56	0.01	0.90	0.03	0.30			
Yeast/*n*-alkanes: Toprina (BP, United Kingdom/France)	1.9		0.13			1.6	3.4	0.1
Yeast/diesel fuel: Fermosin (Schwedt, German Democratic Republic)	1.5	0.15	0.10	0.08	0.15	1.0	2.3	0.45
Fungi/spent sulfite liquors: Pekilo (Finland)	1.0		0.5	0.05		n.r.	0.2	0.1

n.d. = not declared.
∗ Medical – Biological committee of the former Soviet Union – German Democratic Republic for the assessment of SCP.

with other carbon sources it may approach one-half of the total [36]. The rate of consumption of raw materials, including technology-dependent losses, tends to be specific to a particular type of organism and process. Consequently, the choice of microorganisms and technology has a significant influence on product cost. Minimum costs attributable to biological substrate consumption alone are on the order of 150 €/t for petroleum distillate – yeast SCP and 400 €/t for methanol – bacterial SCP. For this reason, SCP prices below ca. 510 €/t are currently (1993) impossible to achieve.

Energy consumption also contributes significantly to operating costs. Energy is required for gas transfer, especially for supplying oxygen to the microorganisms (air compression), but also for cooling (withdrawal of biological process heat), sterilization (steam generation, especially in protected processes), and processing (drying). Oxygen consumption is coupled to substrate consumption, and both are in turn dependent on process design. The costs associated with biomass separation vary with biomass concentration in the fermentation medium and with the size of the biomass cells to be separated. (Because of their smaller cell size, the effort entailed in separating bacteria is substantially greater than that for yeasts.) Capital costs are determined by the demand for technical equipment, which depends on the number of process steps and the equipment layout, as well as by the size of the plant. For a given level of output, these factors decrease with increasing productivity of the process and the space – time yield of the fermenter, respectively. The most favor-

able scale economically for a large facility for converting nonreplenishable (fossil) raw materials lies between 50×10^3 and 300×10^3 t of dry matter per year [36].

An objective choice among various potential technologies, microorganisms, and process layouts is difficult because of the complexity of the process as a whole. It will almost certainly be a compromise, reached only after attempting to visualize potential processes in their optimum form on the basis of suitable evaluation criteria and biological and reaction engineering models that reflect current economic, technical, and physicochemical parameters.

Table 13. Fattening experiments with pigs and chickens [10] (pigs: fattening period 26 – 115 kg live mass; chickens: fattening period up to 56 d of life)

Feed *	Control group	Test groups		
Pig				
Feed yeast, Fermosin, g/kg DS		75		
Molasses/yeast, g/kg DS	75			
Increase in live mass, g animal^{-1} d^{-1}	706	734		
Feed expenditure, kg DS/kg LMI	3.18	3.05		
Chicken				
Feed yeast, Fermosin, g/kg DS		50	100	125
Fish meal, g/kg DS	50	20		
Soy extract, coarse meal, g/kg DS	165	165	125	85
LM on 56 d of life, g	1538	1602	1556	1503
Feed expenditure, kg DS/kg LMI	2.53	2.49	2.39	2.52
Relative	100	98	94	100

* DS = dry substance; LM = live mass; LMI = increase in live mass.

With regard to substrate requirements, a short residence time is advantageous because it is linked to low specific substrate and oxygen consumption. However, rapid growth also leads to lower concentrations of pure protein and higher concentrations of nucleic acids, which in turn impose greater demands for purification. On the other hand, at a given constant output, the reactor size requirement decreases with increasing productivity, reducing both investment costs and debt service. Figure 4 shows that when all the technological factors are considered, the economic optimum for the yeast – petroleum-distillate system does not coincide with the lowest level of biological oxygen consumption. Operating temperatures >45 °C are not economical for this particular process.

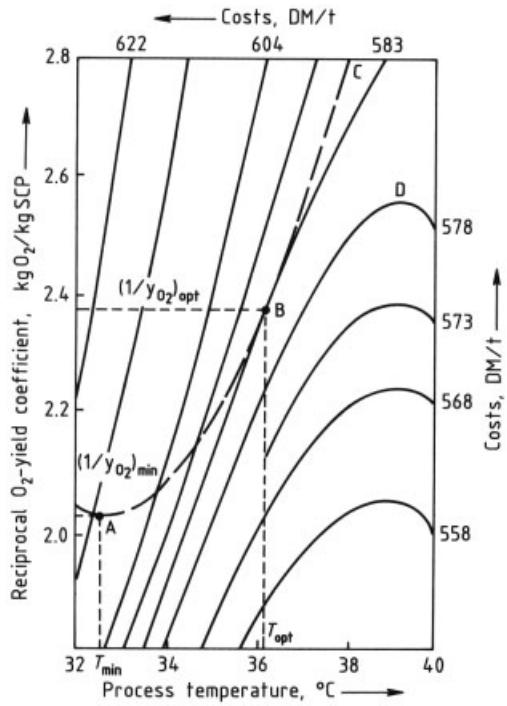

Figure 4. Dependence of fermentation costs (relative values) on specific oxygen consumption and process temperature [45]

A cost comparison alone is not sufficient for evaluating the efficiency of SCP as a source of fodder, or for comparing various single cell proteins with each other and with competing products. In fact, such a cost analysis gives a false picture. The products in question differ in their compositions, and these differences are strain- and process-specific, which in turn has an impact on feed value. The situation becomes clearer when comparisons are based not on SCP production costs, but rather on required expenditure for generating a unit of animal product, because this factor does not vary appreciably with respect to quality. The overall cost for a balanced feed ration then becomes an easily manageable stan-

Table 14. Selected technical data for various SCP processes

Consumption of	Substrate, microorganism				
	Petroleum distillate, yeast [a] [10]	Gas oil [41]	*n*-Alkanes, yeast [42]	Ethanol, yeast [43]	Starch, fungus [44]
Substrate, t/t	1.1	1.1	1.05	1.17	
Basic materials and auxiliary agents, kg/t					
NH_3	177.2	160	206	120	110
H_3PO_4	84.8	86	[c]	70	55
KCl	61.4	[b]	[c]		30
$MgSO_4$	22.0	[b]	[c]		
$CuSO_4$	0.3		[c]		
$ZnSO_4$	1.3	[b]	[c]		
$MnSO_4$	0.9	[b]	[c]		
NaOH	34.0		[c]		
$FeSO_4$		[b]	[c]		
Gasoline	10.1				
Ethanol	4.3				
Operating resources					
Process water, m^3/t	16.4		28.3	40	
Cooling water, m^3/t	1224.3		564	850	400 – 600
Electrical energy, kW · h/t	2460		2042	1500	1000 – 1400
Steam, high pressure, t/t	2.0		0.2	1.0	
Steam, low pressure, t/t	4.6		0.82		

[a] Fermosin process.
[b] $\sum = 32$
[c] $\sum = 275$

dard for comparison, and it should always be invoked as a target parameter in the optimization of production technologies. This can also serve as an objective guide with respect to such research goals as selecting suitable microorganisms for potential SCP applications.

Table 15. Investment cost structure for production of SCP from natural gas, %

Plant segment	Share of the costs	
	Total plant	Main plant
Chemical station	0.6	1.3
Fermentation	12.1	26.5
Separation	19.3	42.2
Evaporation	1.1	2.4
Drying	9.1	19.9
Storage – transport	3.5	7.7
Main plant	45.7	100.0
Recooling system	3.7	
Steam generation	1.6	
Compressor station	1.8	
Residue removal	3.0	
Common area	5.5	
Auxiliary and secondary plant	38.7	
Total investment cost	100.0	

A balanced feed ration must take into account all the factors that are of nutritional – physiological importance. Apart from crude protein content these include levels of amino acids, vitamins, and minerals, among others. In practice, particular attention is directed toward digestible crude protein, energy content, and content of the amino acids lysine, methionine, and cysteine [46]; see Table 17. Ration calculations show that for a given expenditure on SCP production, biomass with a higher lysine content – even if it has less crude protein and sulfur-containing amino acids – results in lower ration costs, and should be given priority. At an SCP price of 510 €/t, the ration price for the production of 1 t of pork from mixtures of SCP and wheat is 2 % lower for a biomass containing 48 g of lysine and 688.5 g of crude protein per kilogram of biomass dry matter than it is for a biomass with only 44 g of lysine but 731 g of crude protein per kilogram of biomass dry matter. A 10 % increase in the lysine content of a microbial biomass (corresponding to 1 g/100 g of protein) leads to a 2.5 % decrease in the price of the total ration. This explains the economic interest in high-lysine biomass products. It also sets realistic parameters on research into increasing the lysine content. In view of the limited prospects for success through genetic engineer-

Table 16. Relative prime costs for the production of SCP from petroleum distillates (percent of total prime costs for production)

Type of cost	Process stage					
	Fermentation and chemicals	Separation	Evaporation and drying	Extraction	Formulation and waste treatment	Total
Substrate	29.3					29.3
Nutrient salts and auxiliary agents	12.9	1.8		0.9	0.7	16.3
Energy	10.2	2.1	6.4	3.3	2.5	26.5
Depreciation	8.3	2.6	2.9	8.5	7.6	27.9
Production prime cost	60.7	6.5	9.3	12.7	10.8	100.0

Table 17. Measures of the quality of SCP and agricultural animal feeds, as well as ration prices for balanced rations

Quality value *		SCP	Soybean	Corn	Ration prices
Dry substance, kg/t (OP)	950	958	880	870	
Digestible crude protein, kg/t (DS)	688.5	731	447	87	
Lysine	48	44	31.4	2.5	
Methionine and cysteine	22	25	17.2	4.1	
Energetic feed value, kEF_S/kg (OP)	530	530	685	720	
Price, €/t	510			97	638
		489		97	638
		510		97	638
			178	97	437

* OP = original product; DS = dry substance; kEF
$_S$: energetic feed units for pig fattening [45]; ration prices in DM per ton of animal product.

ing it would appear advisable to invest more effort in supplementing food rations with microbially produced lysine.

7. Outlook

Biotechnologically produced protein must still demonstrate its ability to compete with protein produced agriculturally – which it is currently (1993) unable to do. Competitiveness with soybean, the most important agricultural protein feed, would be established only if soybean prices were to increase to about three times their current level (ca. 171 €/t in 1993). At present, the best opportunities for exploiting available technological know-how in the production of SCP are associated with decentralized plants that couple waste utilization with SCP production in a local setting [47]. However, future shortages of plant protein could make the industrial production of SCP viable.

The feed requirement of the EC is an indication of the potential involved here: over many years (1980 – 85) this group of countries has consumed an average of about 80×10^6 t/a of mixed feed with a mean crude protein content of 21 %, corresponding to 16.7×10^6 t/a of crude protein. The alternative use of a mixed feed containing 15 mass % SCP would require 12×10^6 t of SCP. Depending on the nature of the biomass, this is equivalent to between 6×10^6 and 8.2×10^6 t of microbial crude protein, representing 36 – 50 % of the total feed crude protein.

At an average production output of 100 000 t/a of biomass each, this would require 120 production facilities. The associated substrate requirement would be in the range of 12×10^6 t (*n*-alkanes) to 30×10^6 t (methanol), levels comparable to present world production figures for ammonia. This clearly illustrates that industrial production of SCP would entail a considerable restructuring of materials management, which could in turn be realized only over a long period of time.

8. References

1. N. S. Scrimshaw in R. J. Mateles, S. R. Tannenbaum (eds.): *Single Cell Protein I,* MIT Press, Cambridge 1968.
2. H.-J. Rehm: *Industrielle Mikrobiologie,* Springer Verlag, Berlin 1980.
3. G. Borgstrom: *Principles of Food Science,* vol. 2, "Food Microbiology and Biochemistry," Food and Nutrition Press, Westport, 1976.

4. F. Reiff, R. Kautzmann, H. Lüers, M. Lindemann (eds.): *Die Hefen,* Verlag Hans Carl, Nürnberg 1960.
5. O. Wurz: *Die Sulfitzellstoffherstellung und ihre Nebenerzeugnisse,* Leykam Verlag, Graz 1948.
6. J. F. Saeman, E. G. Locke, G. K. Dickerman, *Pap. Trade J.* **123** (1946) 38 – 46.
7. J. H. Lichfield, *Science (Washington D.C.)* **219** (1983) 740 – 746.
8. A. H. Rose (ed.): "Microbial Biomass," *Economic Microbiology,* **vol. 4,** Academic Press, London 1979.
9. "Italienische Regierung stoppt Proteinproduktion," *Eur. Chem.* **10** (1977) 220.
10. J. Bauch et al., *Chem. Tech. (Leipzig)* **30** (1978) 284 –286.
11. J. S. Gow, J. D. Littlehailes, S. R. L. Smith, R. B. Walter in S. R. Tannenbaum, D. I. C. Wang (eds.): *Single Cell Protein II,* MIT Press, Cambridge 1976.
12. I. Goldberg: *Single Cell Protein,* Springer Verlag, Berlin 1985.
13. U. Scharf, M. Schlingmann, *Int. Z. Lebensm. Technol. Verfahrenstech.* **4** (1983) 302 – 306.
14. H. G. Schlegel, R. M. Lafferty, *Adv. Biochem. Eng.* **1** (1971) 143 – 168.
15. F. Bergter: *Wachstum von Mikroorganismen,* G. Fischer Verlag, Jena 1983.
16. W. Babel in M. Hecker, W. Babel (eds.): *Physiologie der Mikroorganismen,* G. Fischer Verlag, Stuttgart 1988.
17. C. Anthony in D. E. F. Harrison, I. J. Higgins, R. Watkinson (eds.): *Hydrocarbons in Biotechnology,* Heyden, London 1980.
18. A. H. Stouthamer, *Antonie van Leeuwenhoek* **39** (1973) 545 – 565.
19. W. Babel, R. H. Müller, *Appl. Microbiol. Biotechnol.* **22** (1985) 201 – 207.
20. R. H. Müller, W. Babel, *Acta Biotechnol.* **8** (1988) 249 – 258.
21. E. Lippert, R. Fuchs, S. Gabel, DD 235 805, 1986.
22. R. Müller, K.-D. Markuske, W. Babel, *Z. Allg. Mikrobiol.* **23** (1983) 375 – 384.
23. M. J. Worse, J. D. Windaas, E 0 035 831, 1981.
24. C. Anthony, *Acta Biotechnol.* **3** (1983) 261 – 264.
25. G. Otto, G. Klappach, D. Weichert, H. J. Pitzler, DD 240 026, 1987.
26. W. Babel, R. H. Müller, DD 262 872, 1988.
27. L. Jagusch, W. Schönherr, *Chem. Tech. (Leipzig)* **24** (1972) 68 – 73.
28. M. Moo-Young (ed.): "Comprehensive Biotechnology," in C. L. Cooney, A. E. Humphrey (eds.): *The Principles of Biotechnology – Engineering Considerations,* vol. 2, Pergamon Press, Oxford 1985.
29. W. W. Michalowa et al., DD 105 825, 1981.
30. U. Faust in P. Präve, U. Faust, W. Sittig, A. D. Sukatsch (eds.): *Fundamentals in Biotechnology,* VCH Verlagsgesellschaft, Weinheim 1987.
31. H. Gohlke, G. Matthes, J. Voigt, Zentralinstitut für Ernährung, Potsdam – Rehbrücke 1979, unpublished.
32. E. Faulhaber, Institut für Biotechnologie, Leipzig 1981, unpublished.
33. S. Martin, W. Mai, H.-P. Wittig, E. Faulhaber, Institut für Biotechnologie, Leipzig 1983, unpublished.
34. A. A. Pokrowski et al., *Chem. Tech. (Leipzig)* **30** (1978) 368 – 371.
35. B. Voigt et al., *Chem. Tech. (Leipzig)* **31** (1979) 409 – 413.
36. *Ullmann,* 4th ed., **19,** 518 – 527.
37. G. Henk, T. Heinz, *Arch. Tierernähr.* **29** (1979) 67 – 91.
38. PAG-Guidelines, Protein-Calorie Advisory Group (PAG) of the United Nations System. United Nations, N.Y. 10017, 15.3.1976. Revised PAG/UNU Guidelines for the Development of Novel Protein Sources, Guidelines no. 6, 7, 12, 15, published in: UNU Food and Nutrition Bulletin vol. 5, no. 1 and 2, 1983.
39. "Kriterien und Spezifikationen zur Charakterisierung und Bewertung von Einzellerproteinen (Single Cell Protein) zur Nutzung in Lebensmitteln für die menschliche Ernährung," *Mitt. X. der Kommision zur Prüfung von Lebensmittelzusatz- und Inhaltsstoffen,* DFG, Deutsche Forschungsgemeinschaft, VCH Verlagsgesellschaft, Weinheim 1987.
40. Council Directive: On the Fixing of Guidelines for the Assessment of Certain Products used in Animal Nutrition, Official Journal no. L 126/23-126/27 (83-228-EEC), 18.4.1983.
41. A. Einsele, H. Dellweg (eds.): "Biomass, Microorganisms for Special Applications, Microbial Products I, Energy from Renewable Resources," in H. J. Rehm, G. Reed (eds.): *Biotechnology,* vol. 3, VCH Verlagsgesellschaft, Weinheim 1983.
42. *Hydrocarbon Process.* (1977) 224.
43. *Hydrocarbon Process.* (1977) 225.
44. *Tappi J.* **65** (1982) no. 11.

45. M. Prause, K. Soyez, *Acta Biotechnol.* **2** (1982) 73 – 77.
46. M. Beyer et al.: *DDR-Futterbewertungssystem. Kennzahlen des Futterwertes und Futterbedarfs für Fütterung und Futterplanung,* VEB Deutscher Landwirtschaftsverlag, Berlin 1988.
47. K. Soyez, D. Theiss, M. Prause: "Ecologic Bioprocessing – Chances in New Applications," *Proc. 3rd Internat. Workshop on Strategies for Closed Cycle Production,* Graz 1992.

Single Crystal Growth → Crystal Growth
Sintered Materials → Powder Metallurgy and Sintered Materials

Sintered Steel and Iron

Iron, Powder Metallurgy and Sintered Materials, and Steel are separate keywords.

Henri Cohrt, Schunk Sintermetalltechnik GmbH, Gießen, Federal Republic of Germany

Martin Enders, Schunk Sintermetalltechnik GmbH, Gießen, Federal Republic of Germany

1. Introduction 1
2. Iron and Steel Powders 4
2.1. Powder Production 4
2.1.1. Reduced Metal Powders 4
2.1.2. Atomized Powders 5
2.1.3. Electrolytic Iron Powders 7
2.1.4. Carbonyl Iron Powders 8
2.2. Powder Properties 8
3. Forming and Sintering 11
3.1. Forming 11
3.1.1. Gravity Sintering 11
3.1.2. Uniaxial Pressing 11
3.1.3. Cold Isostatic Pressing 13
3.1.4. Metal Injection Molding 13
3.1.5. Powder Rolling 15
3.1.6. Slip Casting 15
3.2. Sintering 15
3.3. Consolidation Processes Involving Simultaneous Application of Pressure and Heat 17
3.3.1. Powder Forging 17
3.3.2. Hot Isostatic Pressing 17
3.3.3. Extrusion 18
3.3.4. Spray Compacting 18
3.3.5. High-Velocity Compacting 19
4. Secondary Operations 19
5. Strength Properties 20
6. References 22

1. Introduction

The manufacture of semifinished products and structural parts by the production and processing of metal melts has been known since ancient times. A second, less well-known technology has also developed, in which semifinished products and structural parts with near net shapes, or that are ready for installation, are manufactured from metal powders (Powder Metallurgy and Sintered Materials); often, metal melts are not involved in this manufacturing process.

Historical Aspects. The industrial production of metallic bodies (predominantly high-melting metals) by powder technology can be dated to the beginning of the nineteenth century [1]. The powder metallurgy (PM) manufacture of iron and steel articles began in about 1930. The first developments concerned sintered magnets (German Reichspatente 306772 and 679594). Porous iron bearings and machine parts made of sintered iron were first produced around 1935 [2], [3]. Production engineering advantages and lack of skilled labor for the machining work were the driving forces for this development [3]. The subsequent rapid development of iron and steel PM is due to many factors, which include

1) Reduction of machining
2) High raw material utilization
3) Low energy requirements
4) Large variety of geometrical shapes combined with narrow tolerances
5) Good reproducibility of geometrical – and mechanical – physical properties
6) Outstanding suitability for mass production

These advantages lead to a favorable cost situation in the mass production of mechanically stressed, geometrically complex structural parts with narrow tolerances. The development of the property profile can be illustrated by following the improvement in tensile strength since 1940 (Fig. 1) [4].

In addition to structural parts manufactured by pressing and sintering, PM ferrous semifinished products have been developed since ca. 1960. This required the parallel development of special PM processes and alloying techniques. As a result, very highly alloyed materials with outstanding properties have been obtained [5–14]. Forming techniques for structural parts with near net shapes have also undergone consider-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_185

able improvements and advances [15–22]. Great efforts are being made to develop an extensive range of composite structures [23–27].

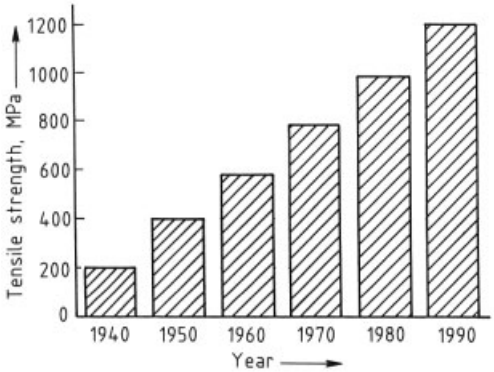

Figure 1. Development of the tensile strength of sintered steel [4]

PM materials and parts made of sintered steel and iron are now used in all areas of daily life, as demonstrated by the following examples:

- Porous materials and parts for maintenance-free, self-lubricating bearings
- Porous stainless steels for filters and ventilation units in the chemical industry
- Medium-density low-alloy and stainless steels for inexpensive moldings for fittings, domestic and handicraft tools, locks, hardware items, automobile construction, and domestic appliances
- Medium-density alloys of pure iron, iron – silicon, and iron – phosphorus for magnetic applications
- High-density low-alloy steels for highly stressed parts in automobile construction (synchronizing rings, oil pump wheels, gear and transmission parts, valve drive parts, etc.)
- High-purity materials without residual porosity for electromagnetic applications
- Stainless steel tubes and profiled rods for process technology
- Creep-resistant steels without residual porosity for turbine construction
- Wear-resistant carbide tool steels for tool construction and components subject to wear in processing plants

Further developments and improvements in powder production (e.g., very fine powders with a high sintering activity), forming (e.g., metal injection molding), heat treatment (e.g., continuous belt sintering furnaces at $\leq 1300\,^{\circ}C$), and many other areas will expand the application range of PM products. These process technology developments are accompanied by intensive research activity [28].

Economic Aspects. Worldwide consumption of iron and steel powders in 1991 for PM products was approximately 420 000 t. North America accounted for ca. 170 000 t, Asia 130 000 t, and Western Europe 90 000 t. Structural parts made of sintered iron and steel are now used in nearly all areas of daily life. The consumption of PM structural parts in various industrial sectors is illustrated in Figure 2 for North America, Japan, and Europe. The overwhelming importance of the automobile industry as a consumer of PM components can be seen in Figure 3 [29]. In the United States about 11.1 kg, in Japan 6.2 kg, and in Western Europe about 4.7 kg of PM structural parts are used per automobile [30].

Definitions and Standards. A detailed summary of PM standards is given in [31]. According to DIN 30 900 the term powder metallurgy is defined as the "production of metallic powder and the fabrication of parts from this powder with or without the addition of nonmetallic powders by forming and sintering." Accordingly, the PM production process consists of three basic steps: preparation of the starting materials, densification, and consolidation by sintering.

In the metal-processing industry the expressions sintered iron and sintered steel are understood to mean semifinished products and structural parts that are made from unalloyed or alloyed iron powders by PM processes.

2. Iron and Steel Powders

2.1. Powder Production

A range of technologies are available for producing iron and steel powders that satisfy current high-quality requirements. The importance of individual techniques has changed during the course of development. In contrast to other PM

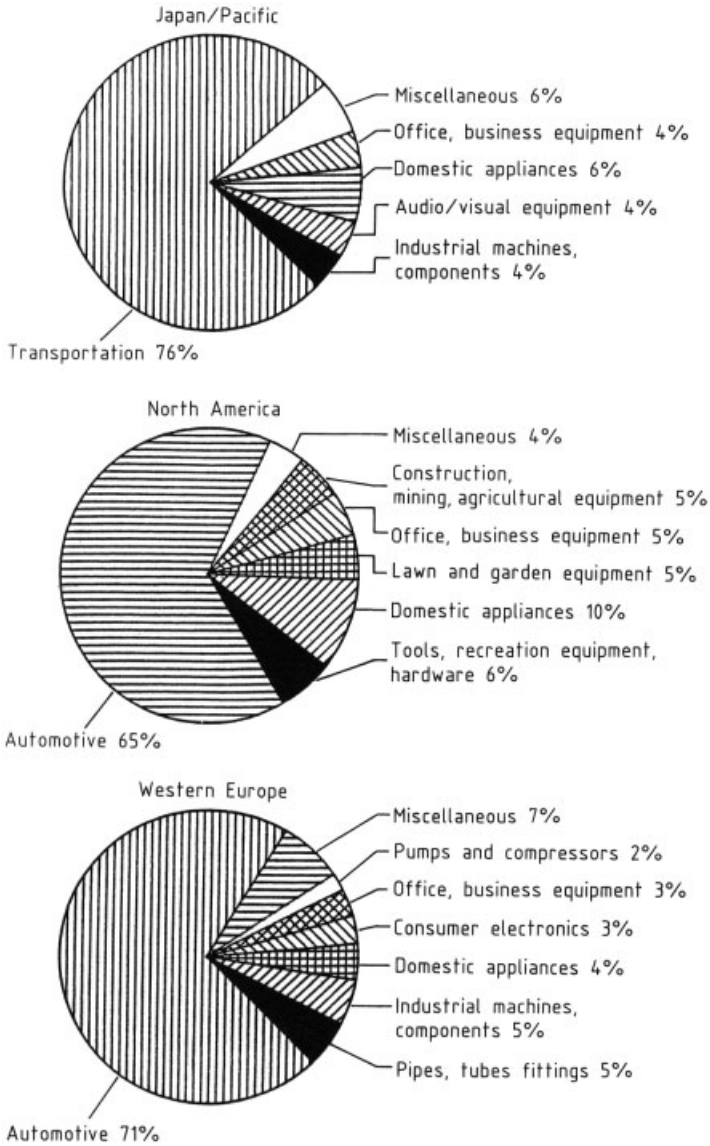

Figure 2. Consumption of PM iron and steel products in Japan, North America, and Western Europe according to area of use

products, mechanical comminution is not a primary process for the production of iron and steel powders.

2.1.1. Reduced Metal Powders

Reduced metal powder is produced from metal compounds by chemical reduction. Roller scale (oxide) or iron ore may be used as starting material for iron powder.

Pyron Process. The Pyron process enjoys a certain popularity in the United States. Roller scale (mainly Fe_3O_4), which is produced in rolling mills during the processing of unalloyed steels, is first ground and then freed of nonmagnetic constituents by using a magnetic separator. Common minor constituents such as manganese or silicon are simultaneously removed in the form of their nonmagnetic oxides. The puri-

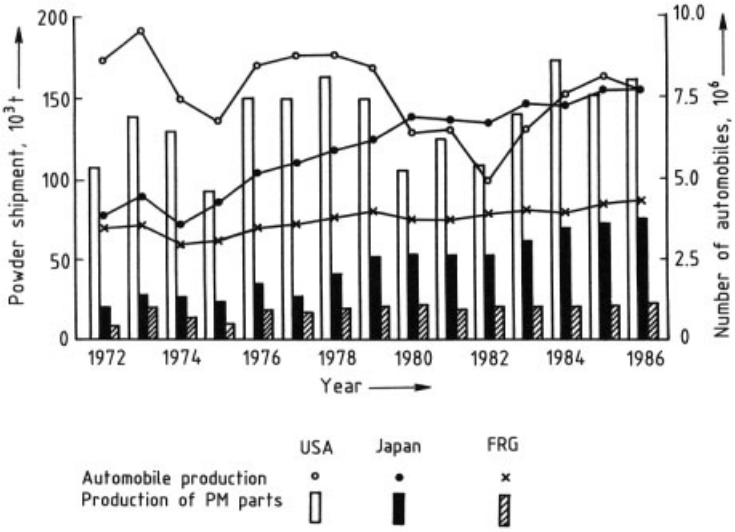

Figure 3. Production of sintered steel and automobiles in the United States, Japan, and the Federal Republic of Germany

fied oxide powder is roasted in air and reduced in a stream of hydrogen at 870 – 1040 °C.

Grinding, followed by size classification, yields powder particles that are largely permeated by micropores of some µm in diameter (a characteristic of all reduced metal powders). Pyron powders have a relatively low apparent density of 2.3 – 2.5 g/cm^3, but as a rule can be easily densified [32], [33].

Sponge-Iron Powder. The Höganäs process (→ Iron) has become extremely important for the production of reduced metal powders [34–37]. The Höganäs process yields sponge-iron powder and still supplies about 40 % of the worldwide demand for iron powder, despite the success of atomization processes (Section 2.1.2). The production of sponge-iron powder is illustrated in Figure 4. Extremely pure magnetite ores (e.g., as found in Keruna, northern Sweden) are used as raw materials. The ore is comminuted, screened, and purified by using magnetic separators (a, d). A reduction mix of coke granules and calcium oxide is prepared in parallel (a – c). The beneficiated ore is packed into a retort (e, p) whose silicon carbide jacket is lined with a layer of the reduction mix. The retort also contains a central core of the reduction mix. Reduction is carried out over 24 – 40 h at 1050 – 1200°C. The following main reactions take place:

$$Fe_3O_4 + CO \rightarrow 3\ FeO + CO_2$$

$$FeO + CO \rightarrow Fe + CO_2$$

Reduction is controlled by the Boudouard equilibrium

$$CO_2 + C \rightleftharpoons 2\ CO$$

Desulfurization occurs during reduction, whereby the lime absorbs the sulfur and is carried out in order to improve the properties of the product. The energy required for reduction is generated in tunnel furnaces that are several hundred meters long (f).

The reduction product is a relatively strongly sintered crude sponge. Adhering slag is removed from the sponge, which is then processed into powder by repeated grinding, purification using magnetic separators, and screening (j – l). The crude powder is subjected to final reduction in an ammonia atmosphere in a continuous furnace (m). The reduction temperature must be sufficiently low (500 – 600 °C) to avoid resintering the powder particles. The sponge-iron powder product has an apparent density similar to that of Pyron powder. The pores in the powder particles are, however, coarser.

2.1.2. Atomized Powders

Pig-Iron Scale Process. Iron powder is obtained by atomization of a pig-iron melt containing ca. 3.5 % carbon, which is melted in a

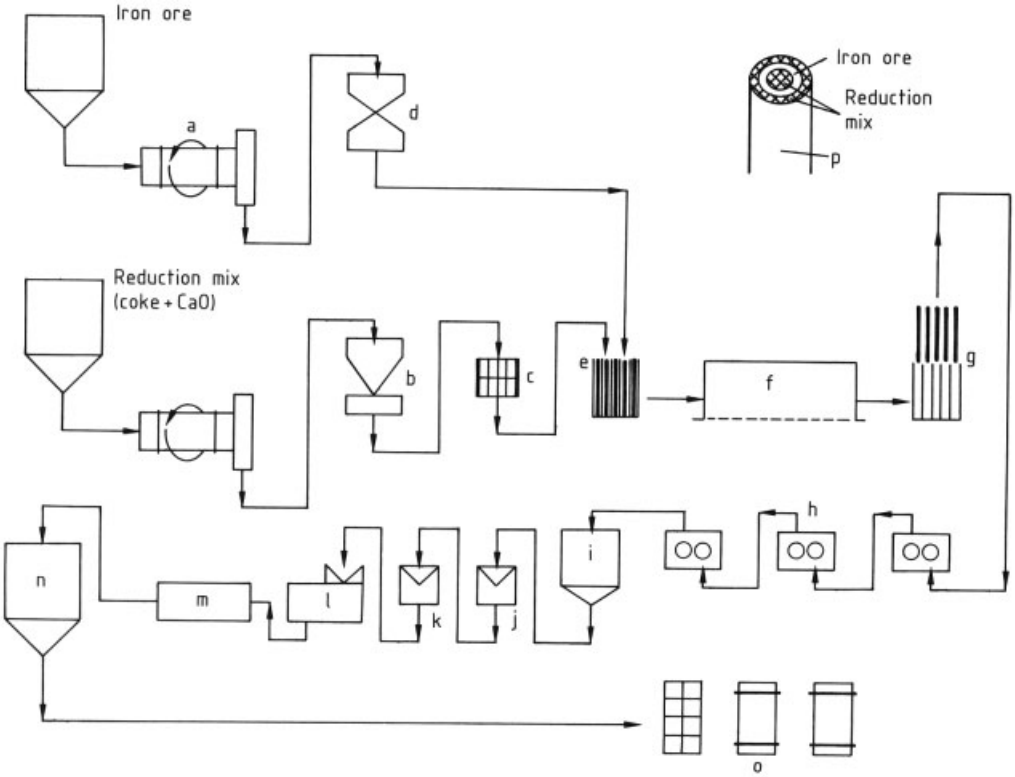

Figure 4. Production of sponge-iron powder (with permission from Höganes)
a) Drying; b) Crushing; c) Screening; d) Magnetic separation; e) Charging in ceramic (SiC) tubes; f) Reduction in tunnel kilns, ca. 1200 °C; g) Discharging; h) Coarse crushing; i) Storage silo; j) Crushing; k) Magnetic separation; l) Grinding and screening; m) Annealing in belt furnace, ca. 800 – 900 °C; n) Equalizing; o) Automatic packing of pressing, welding, and cutting powder; p) Enlarged view of ceramic tube

cupola furnace [38], [39]. Air is injected under high pressure through an atomization nozzle onto the free-falling, liquid pig-iron jet. The pig-iron melt is broken up and solidifies into small droplets that are collected in water. Atomization leads to surface oxidation of the particles (oxygen content ca. 6 %). Reductive annealing with the optional addition of a reducing gas, converts the powder into a low-carbon, low-oxygen cake that is then ground. The powder is screened and attains its technological properties after final annealing. Its structure and pressing properties are very similar to those of sponge-iron powder.

The process is generally carried out in a modified form in which water under high pressure is used as atomization medium instead of air [38–41].

Water-Atomized Powders [42]. The technology of iron powder production is increasingly changing over to the water atomization process (Fig. 5), which uses iron scrap as starting material [43], [44]. The process steps are very similar to those of the pig-iron process, but more efficient electric arc furnaces had to be developed to reach the high melting points of low-carbon scrap iron. The atomized powders contain about 1.5 % oxygen. Dewatering (f) is followed by drying (g) and reduction under a protective gas (e.g., H_2 or NH_3). Unalloyed powders produced by this process have a relatively high apparent density combined with a good pressability. The content and size of slag particles are considerably lower than those of the previously discussed powder production methods. This is beneficial for the manufacture of high-strength structural parts as the inclusion of slag particles

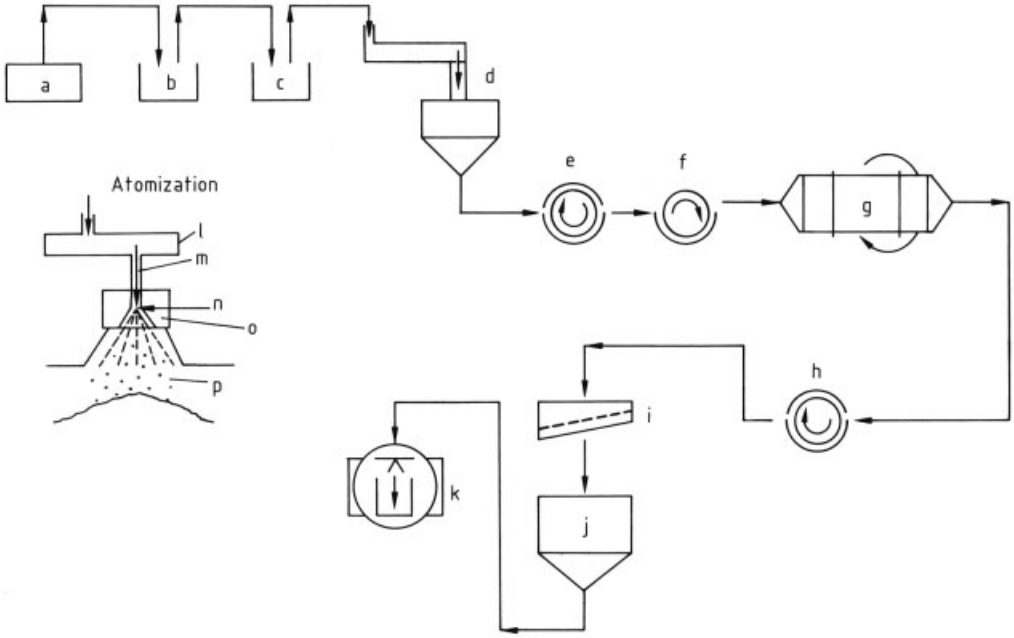

Figure 5. Production of water-atomized iron powder (with permission from Höganäs)
a) Selected scrap iron; b) Arc furnace; c) Liquid steel; d) Atomization; e) Wet magnetic separation; f) Dewatering; g) Rotary dryer; h) Dry magnetic separation; i) Screening; j) Equalizing; k) Weighing and automatic packing; l) Tundish; m) Steel stream; n) High-pressure water; o) Nozzle; p) Atomized iron powder

in the product diminishes the toughness and near resistance properties.

Water atomization permits the production of high-alloy ferrous powders [e.g., the austenitic powders X3CrNiMo 17.13(AISI 316 L), X2CrNi 19.11 (AISI 304 L)]; the martensitic powder X10 Cr13 (AISI 410); and the ferritic powder X6Cr13 (AISI 430). High-speed steels and hard alloys with a high cobalt content (stellites) can also be subjected to water atomization. Postcomminution processes (e.g., the cold stream [45] and vacuum reduction processes [46]) often have to be employed in the case of high-alloy powders; they yield products with particularly low oxygen content [47]. Water atomization processes provide a substantially larger variation in particle shape, particle-size distribution and possible alloy states, compared to the other processes discussed thus far. The particle sizes can be reduced to $< 50\,\mu m$ by increasing the water pressure. Mean particle diameters of ca. $10\,\mu m$ are achieved on a production scale [48], [49]; these powders are produced for special purposes (e.g., metal injection molding).

Oil-Atomized Powders. In the water atomization of alloys with elements having an affinity for oxygen (e.g., Cr and Mn), increased oxygen uptake can cause problems. An attempt was therefore made to provide an effective remedy by using oil as the atomization medium instead of water [49–51]. The reduction in oxygen content was, however, offset by an increased carbon content, which had to be lowered by special annealing treatment. For reasons of cost, oil atomization has not gained general acceptance and is only of minor importance.

Gas-Atomized Powders. Inert gases (nitrogen, argon, and in rare cases, helium) can be used instead of liquid atomization media for the production of special powders. For cases in which the oxygen loading of the powders is particularly critical, atomization may also take place in vacuo. The principle of inert gas atomization is illustrated in Figure 6 [52]. The chemical composition of the powder remains essentially the same as in the melt. Subsequent reduction is no longer necessary to obtain lower oxygen content. The oxygen concentration and the amount

of ceram-ic inclusions (which may be due to slag formation and the melt feed system) can be kept low by special plant technology [52], [53]. Attempts have been made to use liquefied low-temperature gases as atomization media, but this technique has not yet proved economical [54].

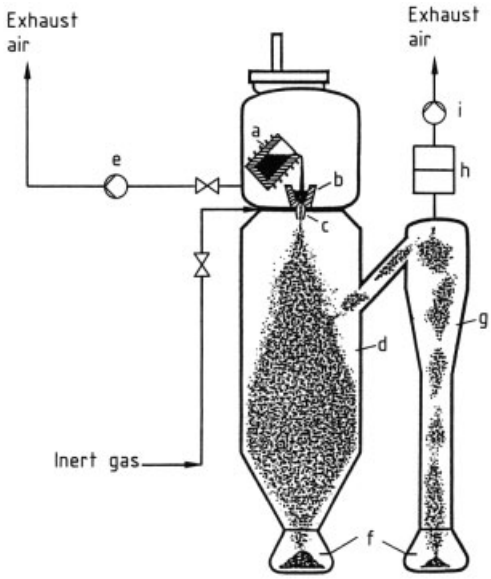

Figure 6. Production of gas-atomized powders (with permission from H. C. Starck)
a) Induction tilting furnace; b) Tundish; c) Atomizer; d) Atomization chamber; e) Vacuum pumps; f) Powder collectors; g) Cyclone separator; h) Superfine filter; i) Ventilator

The predominantly spherical particle shape of gas-atomized powders and their high content of alloying elements mean that they cannot be processed by uniaxial pressing; special processing technologies are generally required. A great deal of effort is being expended on the optimization of gas atomization; a summary of the mechanisms of gas (and water) atomization is given in [42]. Standing ultrasonic waves are being used to promote the disintegration of the melt jet [55]. An attempt is also being made to produce special properties, in particular low oxygen content at small particle sizes, by combining gas atomization with water atomization [56–58].

2.1.3. Electrolytic Iron Powders

Extremely pure iron powders are required for special applications (e.g., materials and structural parts with special electromagnetic properties). One method of producing such powders is electrolytic deposition, in which iron plates are suspended as anodes in a bath of iron and ammonium sulfates. The iron powders are deposited on special steel cathodes at an anodic current density of about 215 A/m^2. The powder is dried, subjected to reductive annealing under hydrogen or ammonia, ground, screened, and then passed for further processing [59]. Electrolytic deposition yields a platelet-shaped powder.

2.1.4. Carbonyl Iron Powders

Another means of producing extremely pure iron powder is the carbonyl process, in which iron is reacted with carbon monoxide at $\leq$20 MPa and 200 °C to form iron pentacarbonyl, $Fe(CO)_5$, which is liquid under these conditions (→ Iron Compounds). The pressure is reduced, and $Fe(CO)_5$ vaporizes and decomposes at ca. 250 °C:

$$Fe(CO)_5 \rightarrow Fe + 5\,CO$$

Decomposition occurs from the vapor phase and results in spherical, generally very fine iron powder (diameter $<$10 µm) with a shell-like structure. The powder usually has a carbon content of 0.8 – 0.9 %. It also contains a relatively large amount of nitrogen and oxygen, and is therefore extremely hard. The carbon, nitrogen, and oxygen contents can be drastically reduced by reduction with hydrogen. The removal of these elements yields an extremely pure, very soft, spherical powder. Typical metallic impurities include nickel $<$ 100 ppm, chromium $<$ 150 ppm, molybdenum $<$ 20 ppm, silicon $<$ 10 ppm, and manganese $<$ 10 ppm.

This powder is used for electromagnetic applications because of its extremely high purity. The spherical grain shape makes it particularly suitable for metal injection molding [60–62].

2.2. Powder Properties

The metal powders are characterized by physical, chemical, and technological properties that influence processing and also determine the subsequent properties of materials and structural parts. These properties have to be considered in the choice of raw materials.

Physical Properties. The geometrical characteristics of powder particles are determined by the particle shape, particle size, and particle-size distribution. The particle shape depends largely on the production process and process parameters (Fig. 7). The particles are usually irregular; only in a few exceptional cases is the shape approximately spherical. In addition to the outer surface, powder particles also have internal surfaces, as in the case of porous sponge-iron powder (Fig. 7 A). Simplifying assumptions therefore generally have to be made when describing the particle geometry.

The *specific surface area* (m^2/g) is the sum of all (readily accessible) surfaces, referred to the mass of a given amount of powder. Since the sintering activity is reflected in the form of the particles, surface energy, the specific surface is very important. Suitable measurement methods include gas absorption (e.g., the BET method) [63] and are described elsewhere (→ Particle Size Analysis). The *size of powder particles* is generally defined as the diameter of a sphere enveloping the particle. It can be determined by screening methods, in which the mesh corresponds to the spherical powder particle diameter that can still just pass through the screen (ISO 4497). Sedimentation and diffraction methods using laser light or X-rays are also based on a spherical model of powder particle shape. Screening, sedimentation, and diffraction methods are used to determine the *particle-size distribution.* For further details of particle-size measurement, see → Particle Size Analysis and [32].

Often, a powder sample can be described sufficiently by means of an *average particle size.* Measurement with the Fisher subsieve sizer is based on the flow resistance exerted by a powder charge in an airstream. The measured flow resistance is compared to the resistance of a spherical charge with corresponding average powder particle diameters. The apparatus is very easy to assemble and simple to operate. Interpretation of the measurement requires care, however, since the powder sample may contain particles that deviate considerably from the measured, average particle diameter.

Description of the *particle shape* is substantially more difficult and is frequently limited to a qualitative, sometimes subjective, description. ISO 3252 defines nine different particle shapes.

1) Acicular → needle shaped
2) Angular → sharply edged or roughly polyhedral
3) Dendritic → of branched shapes
4) Fibrous → having the appearance of regularly or irregularly shaped threads
5) Flaky → plate-like
6) Granular → approximately equidimensional but of irregular shape
7) Irregular → lacking any symmetry
8) Nodular → of rounded irregular shape
9) Spheroidal → roughly spherical

More detailed, quantitative descriptions can be obtained only with the aid of statistical methods and quantitative image analysis [64].

The *hardness* of the powder, measured for example as the Vickers hardness, HV 0.1 plays an important role in subsequent processing. Hardness characterizes the plastic behavior of a material, and important conclusions about the behavior during forming can be obtained from this parameter. Alloying, thermal treatment, and strain hardening all contribute to the hardness of a metallic material, which thus reflects the overall production process. Under certain circumstances the strain hardening state is drastically influenced by screening and mixing after the powder has been annealed.

Chemical Properties. The chemical properties of the powders ultimately determine material specifications after processing and have a decisive influence on sintering behavior. A critical factor is whether the alloying elements are distributed homogeneously in the powder particles, or whether the powder is a mixture of different powdered elements or alloys. The principles of alloy formation are illustrated in Figure 8 [65]. Powders may be of four types:

1) Unalloyed powders consisting of particles of a single element.
2) Powder mixtures consisting of different elementary or alloy powders (Fig. 8 A). The final alloy is formed during the sintering process.
3) Partially alloyed powders consisting of alloying elements that are bound in elementary form (Fig. 8 B) or as master alloys (Fig. 8 C) and penetrate the base powder by diffusion. The alloy is formed during the sintering process.

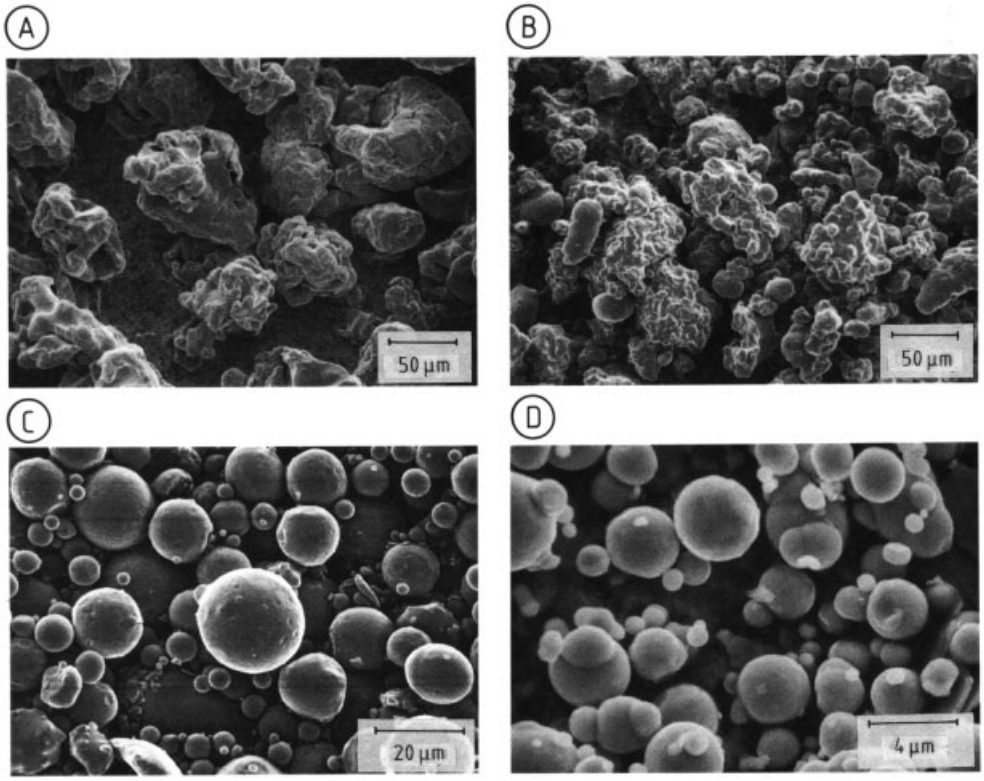

Figure 7. Typical powder shapes obtained in various production processes
A) Sponge-iron powder; B) Water-atomized iron powder; C) Inert-gas-atomized powder; D) Carbonyl iron powder

4) Completely alloyed powders, in which all alloying elements are homogeneously distributed in the powder particles (Fig. 8 D). The alloy is formed during production of the powder.

Oxygen may be present as adsorbed water on the powder surface, as dissolved oxygen in the powder particle, or as chemically bound oxygen in the form of oxides in the powder. Excess oxygen can adversely affect the sintering behavior of the powder. The reaction of oxygen with carbon results in loss of carbon from the alloy, which is then no longer available for improving hardness in heat treatment. Reaction with oxygen renders metallic alloying elements ineffective in the metallurgical sense or can lead to formation of stable oxide skins on the particle surfaces that interfere with sintering by preventing diffusion.

The oxygen content can be measured as annealing loss under hydrogen (ISO 4491T2, 4491T3). The oxygen bound as nonreducible oxides (Al_2O_3, SiO_2, CaO, etc.) is not, however, included in this measurement. The total oxygen content can be measured by melting in a graphite crucible at $\geq$2000 °C. Oxygen is released by carbothermic reduction of the metal oxide and reacts with the carbon of the crucible to form carbon monoxide. The amount of carbon monoxide formed is determined by IR spectrometry (ISO 4491T4).

With regard to work safety, the potential health hazards and the tendency to self-ignition (pyrophoricity) should be kept in mind when handling metal powders [66], [67].

Technological Properties. Metal powders are characterized by technological parameters that describe the special behavior and properties of the powders during processing.

The *apparent density* describes the specific weight of a powder when it is poured under the action of gravity. It is therefore used for

the volumetric metering of powder into molds (ISO 3923).

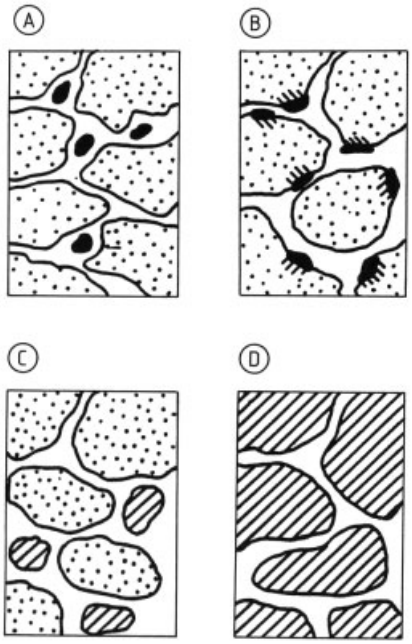

Figure 8. Types of alloy formation
A) Powder Mixture of different elements or alloys; B) Alloy powder bonded to the base powder by diffusion processes; C) Powder mixture using a master alloy; D) Prealloyed powder

In the axial pressing technique the mold is generally filled by flow under the force of gravity. Flow behavior is characterized by a standard test in which the time taken for 50 g of powder to flow through a funnel is measured (*flow rate*) (ISO 4490). The flow rate is largely determined by the particle size and the particle-size distribution.

Compressibility is determined by using a standard round die that is filled under gravity. The density of the compact during pressing is measured as a function of a given force of pressing and describes the compaction behavior of the powder (ISO 3927).

Springback describes the elastic behavior of the powder compact after pressing and during ejection from the pressing die. It is thus a measure of the tendency of a powder metal compact to form cracks during ejection from the die. Springback is sometimes also measured as a function of the specific force of pressure (ISO 4492).

The *ejection friction* is a measure of the stress exerted on the die due to friction between the powder compact and the die wall. It provides information on die wear.

The *green strength* is measured in a bending test on samples of rectangular cross section and shows of the mechanical stability of powder compacts. It is used to evaluate the handling of compacts (ISO 3995).

The *dimensional change* of a powder compact during sintering is a particularly important parameter required for the design and dimensioning of dies. The dimensional change may be positive or negative, depending on the compaction state, alloying state, and sintering parameters (atmosphere, temperature, and time). Zero dimensional change is often desired. The dimensional change is defined as (ISO 4492)

$$\frac{L - L_0}{L_0}$$

where
L = dimension after sintering
L_0 = die dimension

3. Forming and Sintering

Forming and sintering can be carried out either consecutively or in a single process.

3.1. Forming

3.1.1. Gravity Sintering

Gravity sintering is the simplest forming process in powder metallurgy. Metal powder is poured into heat-resistant molds where it is sintered. The sintered article therefore adopts the geometry of the mold. The mold material must have sufficient thermal stability at the temperatures employed for sintering and must not react with the metal powder. These requirements mean that gravity sintering is rarely used for sintered iron and steel. Experiments on the sintering of stainless steel filters have been published [68]. In medical technology, similar methods are used for applying porous coatings to solid implant materials. The coating is said to improve adhesion of the implant to the bone tissue [69], [70].

3.1.2. Uniaxial Pressing

By far the largest amount of iron and steel powder is processed into moldings by uniaxial pressing. Before the shaping process the powder is conditioned, which comprises

1) Combining the alloying powder constituents
2) Blending and homogenizing the powder mixture
3) Incorporating a suitable lubricant

The purposes of these stages are to produce an optimum alloy, to obtain appropriate technological properties, and to establish homogeneous material and structural-part properties after sintering. The lubricant dramatically influences the technological properties of the powder. It has a temporary function and must be removed from the formed part by vaporization before sintering takes place. The choice of lubricant is thus a compromise between properties relevant for pressing and the vaporization behavior; the aim is to achieve maximum vaporization and optimum lubrication. Metallic soaps such as zinc stearate or lithium stearate may be used; they have very good lubricating action but leave a residue on vaporization.

The trend toward the use of stearic acid, paraffin, or microcrystalline waxes is increasing. These substances vaporize without leaving a residue but generally have a poorer lubricating action than metallic soaps [71].

The powder is metered under gravity from a storage container into a die whose base is closed by a lower punch. The powder column in the die is compacted by movement of the lower punch and an upper punch.

The powder compact is then ejected from the die. The pressing tool returns to the starting position and the process is repeated. Further details are given in [65], [72], and → Powder Metallurgy and Sintered Materials, Chap. 5.1.

A wide range of shapes (Fig. 9) can be produced by using suitably designed dies and punches [73]:

1) Nonrotationally symmetrical parts can be formed by suitably contouring the die
2) Multiple-height parts can be pressed by providing shoulders in the upper and lower punches
3) Differences in pressing height can be increased by using split punches, the individual punch elements being controlled separately
4) Within certain limits, undercuts can be produced by means of split dies [74], special press control devices are required for this purpose

The rapid development of microelectronics has given a substantial impetus to design in uniaxial pressing. Modern process computer technology and sophisticated hydraulic systems allow control of very complicated pressing tools incorporating multiply split punches [75] (multiplate adapter technique) [76]. This means that the quality of complicated formed parts has improved, and the geometrical limits on freedom of design have substantially broadened.

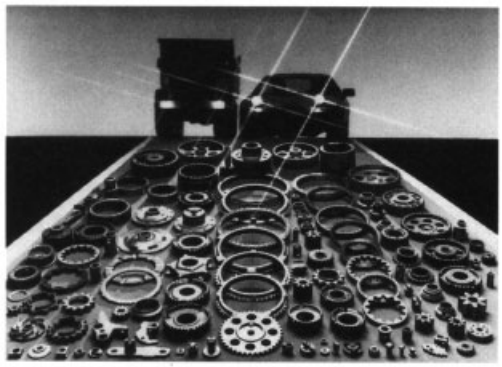

Figure 9. Sintered steel and iron parts with widely differing shapes produced by axial pressing (with permission from SMK)

The powder particles are assumed to exhibit an elastic – plastic behavior. Density gradients are formed in the powder compact due to friction between powder particles and between the powder charge and the tool surfaces. The density distribution in a nickel powder compact pressed from one direction is illustrated in Figure 10 and also applies to iron powders. Levels of equal density are thus roughly conical. Pressing from two directions is generally used in practice and produces a symmetrical, mirror-image density distribution. Consequently, undesirably high density gradients may occur in very long moldings. The ratio of the height of the part to its diameter or wall thickness has an important influence on the formation of density gradients. Density gradients can be avoided at a height – diameter ratio of ca. 2.5. Substantially

higher ratios can be achieved with parts that do not require a high mechanical strength. Density gradients lead to gradients in the pore structure of the sintered material and thus to gradients in its properties. Furthermore, parts may become distorted.

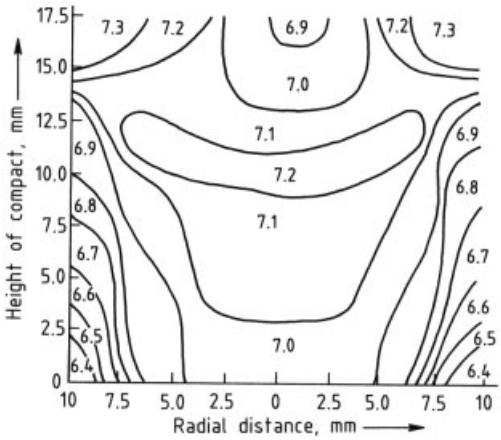

Figure 10. Density distribution in a cylindrical nickel powder compact pressed from the top only

The cohesion of powder particles in ductile iron powders is effected by mechanical forming of the powder particles. This leads to mutual anchorage in which adhesion forces become effective. The elastic – plastic properties of the material, the morphology of the powder particles, and the choice of lubricants play decisive roles with regard to compressibility. A compaction diagram for iron powders is illustrated in Figure 11. The compaction curves become asymptotic with increasing compacting pressure. The asymptotic value is less than the density of the corresponding compact material (i.e., the metal powders cannot be completely compacted). The reasons for this are the increasing friction between powder particles, the increasing contact area between powder particles, and the cold hardening of the powder. If powders are to be compressed to extremely high densities (e.g., for strength reasons), further densification can be achieved by using double pressing. The compact is first pressed and then subjected to intermediate annealing to reverse the cold hardening of the powder particles and promote the formation of initial metallurgical contacts between powder particles. The annealing temperature should be chosen such that recrystallization processes can take place in the powder, but alloying elements (particularly carbon) in the powder mixture do not dissolve. Annealing is followed by a second pressing step and a sintering step at high temperature.

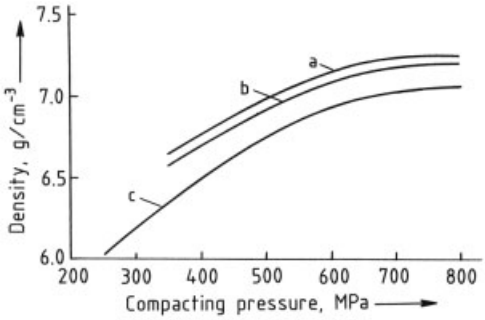

Figure 11. Compaction diagram for iron powders
a) Water-atomized iron powder; b) Water-atomized prealloyed Fe – Mo powder; c) Sponge-iron powder

Computer methods (finite-element analysis) and sophisticated measuring techniques have been used to improve understanding of the processes occurring in uniaxial pressing in order to optimize the pressing procedure [77–79]. Such investigations may also serve to optimize tools with regard to the friction between the die wall and powder charge [80].

3.1.3. Cold Isostatic Pressing

Cold isostatic pressing overcomes the limitations of uniaxial pressing in the processing of slightly ductile materials and the production of very large parts. Here, compaction is effected by means of a liquid medium that acts on the powder from all directions. The resulting (quasi) hydrostatic pressure conditions give a substantially more favorable density distribution in the powder compact, thus allowing production of very large bodies with complicated shapes.

The powder is packed into a tight, elastic mold of similar shape to the desired end contour. The mold is closed and placed in the pressure medium. Densification results by raising the pressure at room temperature. The pressure is then reduced, and the mold removed from the pressure medium. The formed part is removed from its sheath and sintered.

In the wet-bag method the elastic mold is filled outside the press and closed. In the dry-bag

method the mold is a constituent part of the pressure chamber and the powder is packed directly into the mold. The dry-bag method permits a quasi-hydrostatic pressure loading and is therefore particularly suitable for cylindrical bodies.

Pressure vessels with diameters ≤ 2 m and heights ≤ 4 m are available. The pressures employed are ≤ 600 MPa [81], [82].

By using cores, hollow bodies of complicated geometries (e.g., with internal threads) can also be produced. Cold isostatic pressing is used to produce large, stainless steel filter candles for process engineering [83], semifinished products and near net shapes from high-speed steel, and cylinder liners for large diesel engines [81], [84].

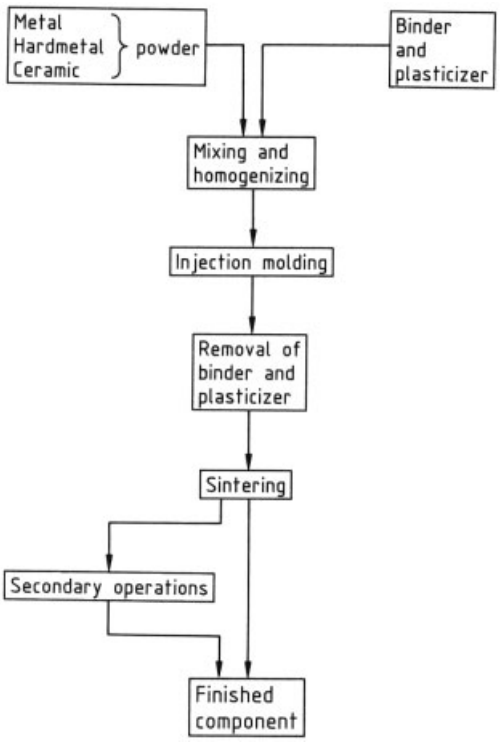

Figure 12. Processing scheme for metal injection molding

3.1.4. Metal Injection Molding

Injection molding has only recently been employed for sintered iron and steel. This process combines the design freedom that is a feature of plastics technology with the material properties of high performance steels, thereby providing the designer with completely novel possibilities [17], [19], [85–87].

Metal injection molding (Fig. 12) is similar to the injection molding of thermoplasts or thermosets, except that molding is followed by further treatment. Very fine metal powders are homogeneously mixed with thermoplastic binders and plasticizers. To achieve good flowability and thereby ensure optimum processibility, spherical powders are preferred. Inert-gas-atomized powders are frequently employed. However, iron carbonyl powders are normally used for low-alloy steels.

The thermoplastic binder gives the material an injection-moldable consistency and binds together the metal powder particles in the green body. This adhesion is necessary since the powder particles are not deformed during injection molding, in contrast to the case in uniaxial pressing. The metal powder constitutes 50 – 70 vol % of the molding material. At processing temperature the material has a consistency similar to that of toothpaste; it is processed in reciprocating screw injection molding machines at 150 – 250 °C. The injection-molded parts possess all the typical geometrical features of the finished product but have a larger volume because of their binder content.

Most of the binder and plasticizer are then removed from the molding by thermal decomposition and vaporization. The green body is converted into a porous, sensitive, but still manipulable body. Binder residues must remain in the formed part until the first sintered contacts are formed in the subsequent sintering process.

Sintering may be carried out in conventional protective gas atmospheres or under vacuum. The individual powder particles coalesce, which results in a loss of porosity and decrease in volume of the body. Shrinkage is dimensionally uniform and between 10 and 17 %. The degree of shrinkage is governed by the proportion of binder and plasticizer in the feedstsock. Purely metallic parts with a low residual porosity (generally < 5 vol %) are obtained. Metal injection molding produces parts that are ready for installation or whose properties can be modified for special applications by means of secondary operations. Examples are shown in Figure 13.

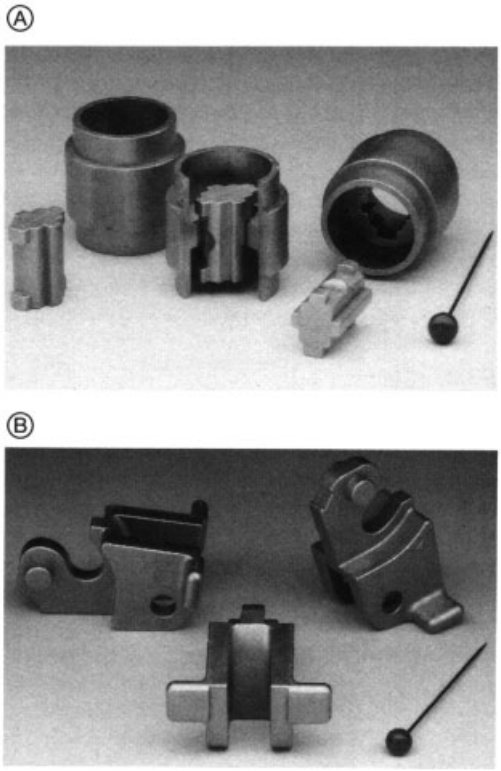

Figure 13. Components produced by metal injection molding
A) Components for security locks; B) Parts made of low-alloy steel for automobile construction

Experience has shown that economical production is generally possible with an annual production of $\geq 100\,000$ parts. In special cases, smaller output may also be feasible. Metal injection molding can be used successfully in applications where

1) Expensive mechanical processing can be reduced
2) Materials are to be used that cannot be machined or can only be machined with difficulty
3) The processing of expensive materials is necessary (high material yield by recycling sprue residues)
4) Integrated construction methods are to be employed (fewer joining operations)

A wide range of alloys can be processed by this technique since it is not restricted by powder properties. It is especially useful for very hard powders. The process has been applied to

1) Low-alloy steels
2) Acid- and alkali-resistant steels
3) Cold work steels and high-speed steels
4) Iron based magnetic materials with Si, Co, or Ni as alloying elements
5) Materials with low thermal coefficients of expansion produced from Fe – Ni alloys

3.1.5. Powder Rolling

Powder rolling (roll compacting) is a continuous process for producing semifinished products (wires and strips) for electrical and magnetic applications. The powder is compacted between two rolls into endless wires and strips before passing through the sintering oven. Several annealing and rolling steps can also be performed in succession for postcompaction and modification of structural properties [88–90].

3.1.6. Slip Casting

Slip casting is a method adopted from ceramics technology. Powder is suspended in a suitable solvent to form a casting slip that is then poured into porous molds. Slip casting of iron and steel alloys is not of any great technical importance. However, attempts have been made to use this process for one-off and small-series runs for producing complicated parts made of high-quality materials [71], [91].

3.2. Sintering

Principles. The structure and properties of the material are established during sintering. F. Thümmler has defined sintering as follows: "Sintering is a thermally activated material transport process that results in intensification of particle contacts and/or in a change in porosity and pore geometry, with a decrease in free enthalpy. It may involve a liquid phase" [65]. The term "thermally activated processes" is understood to mean diffusion processes. The driving forces are surface tension in single-phase systems (pure iron and completely alloyed systems), as well as concentration differences in powder mixtures and partially alloyed powders.

Surface, grain boundary, and volume diffusion occur as material transport mechanisms in solid-state sintering. Viscous slip processes and rearrangements due to rotation of individual particles or collections of particles may also make a substantial contribution. In the case of liquid phases, the above sintering mechanisms are supplemented by capillary forces [92–94]. If liquid phases occur in completely alloyed powders during sintering, this is described as "supersolidus sintering" [95], [96].

Solid-state and liquid sintering are described in detail elsewhere (→ Powder Metallurgy and Sintered Materials, Chap. 6.1.). In general, three sintering stages are distinguished:

1) Growth of particle contacts and formation of metallurgical boundaries between the individual powder particles (shrinkage is slight)
2) Formation of a coherent pore skeleton (shrinkage increases)
3) Formation of isolated pores, pore contouring, and elimination of pores (further shrinkage occurs, ultimately resulting in complete densification [65])

Figure 14 shows the dimensional changes produced by sintering as a function of the compacting pressure for sintered iron materials under various process conditions.

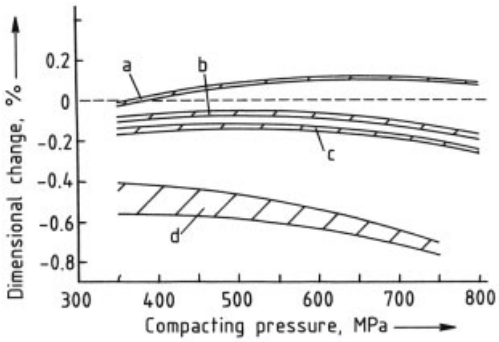

Figure 14. Shrinkage behavior of sintered iron alloys a) Fe + 1.5 % Cu, 1.75 % Ni, 0.5 % Mo, 0.6 % C partially alloyed (1120 °C, 30 min, endogas); b) Sponge-iron powder, 0 % C (1120 °C, 30 min, 75 % N_2, 25 % H_2); c) Water-atomized iron powder, 0 % C (1120 °C, 30 min, 75 % N_2, 25 % H_2); d) Fe + 0.6 % P, 0 % C (1120 °C, 30 min, 75 % N_2, 25 % H_2)

Pure iron powders and iron – carbon alloys are sintered in the solid state. Liquid-phase sintering is often also employed to promote densification. Elements that melt below the sintering temperature or that form low-melting eutectics with iron (e.g., Cu and P) are added to the iron powder. The formation of liquid phases is often transient; they disappear due to the reaction of the melt with the solid component. In the case of iron – copper alloys, volume changes resulting from solution processes and the formation of diffusion porosity can be used to a limited extent to compensate for shrinkage [92], [97], [98]. Initial swelling is then often observed prior to shrinkage [17]. The control of liquid-phase sintering is substantially more difficult than solid-state sintering because wetting is sensitive to interference (caused, for example, by oxides on powder surfaces) [99–104].

Sintering Atmospheres. The protective gas atmospheres have substantial influence on sintering kinetics. Unsatisfactory control of the oxidation potential of the sintering atmosphere can result in oxidation of the powder. Oxide formation greatly inhibits sintering by preventing diffusion; elements with a high affinity for oxygen (e.g., Cr and Mn) are bound so that they can no longer perform their metallurgical functions. A high oxidation potential is generally due to a high content of residual water vapor in the sintering atmospheres. The dew point of the gas is used to evaluate the water content. The carbon balance of the material can also be greatly influenced by the protective gas atmosphere, which may have a carburizing (excess CO) or a decarburizing (excess H_2O or CO_2) action.

Besides preventing oxidation, the protective gas also reduces any oxides already present. Carbon monoxide and hydrogen are primarily responsible for reduction. Nitrogen has a neutral function, although the formation of nitride should be kept in mind (e.g., with stainless steels) [105]. Nitride can bind with the alloying element Cr resulting a loss of powder corrosion resistance.

Temperature also has a decisive influence on the effect of the gases. In industrial sintering furnaces, special attention should be paid to conditions in the heating and cooling zones. Oxidizing conditions are desired in the dewaxing zone, especially with continuous sintering furnaces, to ensure complete removal of the lubricants from the moldings [71], [106–108].

Protective gas mixtures consisting of 75 % hydrogen and 25 % nitrogen are produced by

decomposing ammonia. This and other N_2-H_2 ratios are also obtained by mixing pure nitrogen and hydrogen. Atmospheres of $CO-CO_2$ are generally prepared by partial or complete combustion of propane, natural gas, butane, fuel oil, gasoline, etc. The term endogas or exogas is used, depending on whether energy is supplied to or released by the protective gas during its production. Endogas is used to control the carbon balance of the materials; its composition varies: 30–50 % H_2, 18–25 % CO, 28–48 % N_2, ≤ 0.5 % CO_2, ≤ 0.5 % CH_4, and ≤ 0.5 % H_2O. The dew point is between 0 and ca. −20 °C [71].

Subatmospheric pressures are used in the production of high-speed steels, magnetic materials, and low-alloy steels. Reducing conditions generally prevail in a vacuum. Also, dissolved gases (e.g., O_2 and N_2) can be removed from soft magnetic materials. The oxidation potential in vacuo is substantially lower than in protective gas atmospheres [109].

Industrial Sintering Plants. The majority of moldings produced from sintered iron and steel are sintered in continuous belt or walking-beam furnaces (→ Powder Metallurgy and Sintered Materials).

In belt sintering furnaces the sintered material passes through the furnace at up to 1150 °C on a heat-resistant steel belt. Attempts are being made to construct belt furnaces with ceramic belts and muffles that can be operated at ≤ 1300 °C [110].

In walking-beam furnaces the sintered material is transported through the furnace in boats or trays by means of a moving-beam system. Sintering can be performed at higher temperatures (≤ 1350 °C). The protective gas flows around the heating elements (Mo or W wires). Such furnaces cannot be operated with carburizing atmospheres because of the risk of carbide formation, which would destroy the heating elements, and the further risk of decarburization of the sintered material.

In all sintering furnaces the lubricant is expelled from the compacts by vaporization and combustion in a preheating (dewaxing zone), which proceeds the high-temperature sintering zone. A water-cooled cooling zone follows the sintering zone.

Vacuum sintering plants are very often used for materials with special properties, such as high-speed steels. Although these furnaces operate batchwise, they are highly flexible with regard to process control. Combined operation under both a protective gas and reduced pressure is possible. Such sintering plants are extremely cost intensive and are therefore used only for high-grade materials. The use of vacuum sintering plants for low-alloy steels has, however, also been reported [108].

3.3. Consolidation Processes Involving Simultaneous Application of Pressure and Heat

3.3.1. Powder Forging

Uniaxial pressing and the double press–double sintering technique have their limits in terms of achievable material density. Material densification is further increased by sinter forging, in which the formed parts are sintered in continuous furnaces and finally postconsolidated (forged) at 700–1100 °C in tempered pressing tools that are similar in construction to the powder pressing tools. Tool temperatures are 150–350 °C. Forging can be carried out directly by utilizing the sintering heat, or after intermediate cooling and reheating in an induction furnace.

The time during which the formed parts come in contact with the relatively cold tool must be minimized to avoid freezing the pores in subsurface regions. Short handling times in the ambient atmosphere also help to protect the surfaces against excessive oxidation. These requirements necessitate a high degree of automation.

Materials produced by powder forging have high tensile strengths (ca. ≥ 1000 MPa) combined with high elongations at break of (≥ 10 %). Low-alloy steels containing manganese, chromium, nickel, molybdenum, and carbon as alloying elements (total content ≤ 3 %) are used. Unalloyed iron–carbon materials can also be processed. Powder forging is used for highly stressed parts such as clutch rings, synchronizing rings, connecting rods, and pocket chain wheels [23], [24], [110–113].

3.3.2. Hot Isostatic Pressing

Hot isostatic pressing is a subsequent development of cold isostatic pressing. The parts are normally consolidated at high temperature ($\leq 1500\,°C$) and high pressure ($\leq 100\,MPa$) in water-cooled pressure vessels (Fig. 15). More extreme operating parameters may be used for special applications. The units are generally equipped with a rapid cooling device to achieve short cycle times [114], [115]. The batch is generally assembled on a frame outside the unit, which is lifted into the pressure vessel. Inert gases (N_2, Ar, or He) are used as pressure medium. Production units with a batch diameter <150 cm and a batch height <370 cm are in operation.

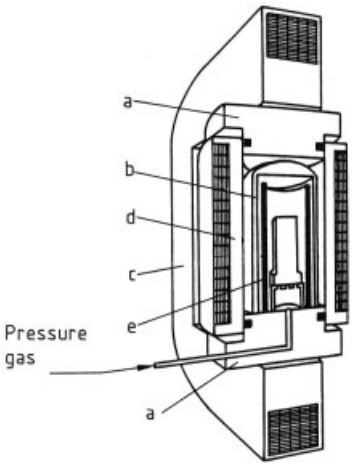

Figure 15. A heated pressure vessel with a protective hood for hot isostatic pressing (with permission from ABB)
a) Closure; b) Insulation; c) Frame; d) Pressure vessel; e) Heater

The powder usually has to be packed into heat-resistant capsules that must be evacuated before closure. The capsules must:

1) Permit shrinkage of the powder without crack formation
2) Not exhibit excessive fold formation or prevent shrinkage at the edges
3) Not react excessively with the powder
4) Be easily removed from the formed part

In an alternative approach the pressure vessel is first evacuated to sinter the powder in vacuo until the pores are closed. Complete densification then takes place under relatively moderate pressures ($\leq 20\,MPa$). This variant can generally be used only with liquid-phase sintering alloy systems since a rapid closure of the pores occurs [20], [116–118]. Care should be taken to ensure that adverse reactions do not occur between the PM materials and impurities in the pressure medium. The use of carefully purified gases is essential [119], [120].

Finally, a process has been developed in the United States in which a quasi-hydrostatic pressure is exerted by an inert powder charge [21], [22].

Attempts are being made to optimize parameters by taking the densification mechanisms into account [121–124]. Because of the complex process technology, hot isostatic pressing is economically viable only for materials with extreme material properties. Properties can be obtained that are not accessible with melt metallurgy:

1) Segregation-free structure (no solidification processes)
2) Extremely homogeneous distribution of very fine-grained carbides in high-speed steels
3) Avoidance of fracture-initiating pores
4) Production of very large parts with near net shapes (e.g., gas turbine parts) that are manufactured from poorly processible, expensive materials

The main advantages of hot isostatically pressed materials are the reduction in the critical defect size and the more uniform property values [14], [123], [125], [126]. Hot isostatic pressing is used to process high-speed steels, tool steels, and heat-resistant steels for turbine construction that are highly mechanically, thermally, and corrosively stressed.

3.3.3. Extrusion

In special cases, ferrous powder is processed into semifinished products under high pressure at ca. 1000 °C by shaping nozzles. Pipes and profiled rods of corrosion-resistant steels are produced in this way for the petrochemical industry, desalination plants, furnaces, and industrial heat exchangers. Inert-gas-atomized powders are packed into sheet-metal capsules and compacted by cold isostatic pressing. Subsequent extrusion is carried out in ram extruders using silica as a lubricant [127–129]. High-

speed steels are also processed by this technique [130].

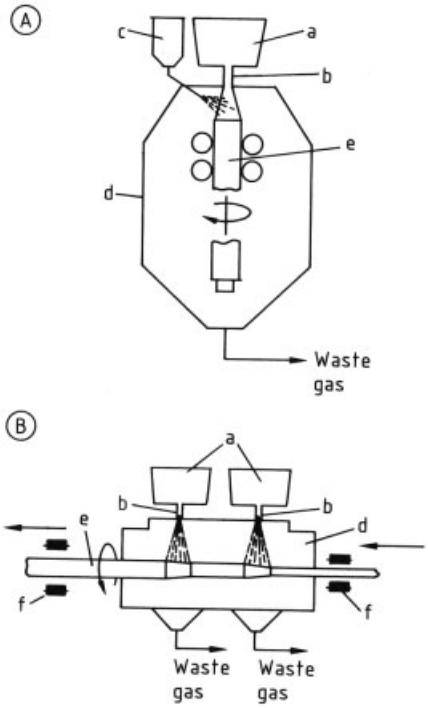

Figure 16. Schematic diagram of spray compacting (with permission from Mannesmann)
A) Production of bolts and washers; B) Production of tubes
a) Tundish; b) Gas atomizer; c) Particle injection; d) Spraying chamber; e) Product; f) Drive

3.3.4. Spray Compacting

Spray compacting is a more recent development. A melt is atomized by a gas jet into droplets, which however do not cool completely. Instead, the liquid or doughy droplets strike a substrate on which they solidify by rapid dissipation of heat. Compacts in the form of bolts, pins, tubes, or two-dimensional products frabricated from steel (Fig. 16) and other metal alloys can be produced in this way. The materials do not exhibit segregation and have a fine-grain microstructure, due to the rapid solidification of the melt droplets [131], [132].

3.3.5. High-Velocity Compacting

Many attempts have been made to subject a wide range of powders to extremely high pressure in order to obtain simultaneous compaction and sintering [133], [134]. Shock waves generated by detonation of explosives [135], pulsed magnetic fields [136], or highly accelerated projectiles [137] have been used. However, these methods are not employed to any great extent on an industrial scale.

4. Secondary Operations

Only in very few cases can PM parts be employed directly after sintering. As a rule, various refining processes also have to be carried out with the intention of modifying the geometrical, mechanico-physical, chemical, and possibly optical and decorative properties of the formed parts for their intended application.

Sizing. Geometrical tolerances of sintered parts are reduced by means of a sizing tool, similar to a powder pressing tool. Their subsurface regions are shaped slightly by means of a punch and thereby lightly compacted. Sizing can be confined to selected geometrical features or can cover the entire part. In the case of porous sintered bearings the topography of the surface is established while at the same time the tribological demands are considered [32], [71]. In the double press – double sintering technique, pressing is carried out after the first sintering stage with a view to obtain a substantially higher density. This is referred to as postcompaction (instead of sizing).

Machining. If specific tolerance ranges are exceeded due to process or material factors or if geometrical features cannot be obtained (nondemoldable undercuts, tool-damaging contours), machining operations (e.g., drilling, grinding, milling, honing, and threading) may be carried out. Chip-breaking additives (manganese sulfide, boron nitride) may be used to improve the machinability of the materials. These additives generally have no significant influence on strength properties [138–141].

Infiltration and Impregnation. The porosity of PM materials often interferes with their use. In the United States, sintered steels are often infiltrated with copper to increase their strength,

particularly the impact strength, and also to impart impermeability to both liquid and gaseous media. Lead infiltration is also performed for tribological reasons [142–144].

If sintered steels are to be case-hardened or galvanically coated, the porosity can be sealed with water glass. This prevents carburization in the interior of the material and interfering influences during electrodeposition, respectively [145]. Impregnation with plastics is also known. Sintered bearings are impregnated with oils and fats to give them a maintenance-free, lifetime lubrication [32].

Surface Treatment. The influences of porosity should be considered in the surface treatment of sintered metals. If necessary, these influences can be minimized by sealing.

Metallic coatings can be deposited electrolytically by galvanic processes (zinc, tin, cadmium, copper, nickel, and chromium). Cadmium is seldom used because of environmental problems. Nickel coatings are applied by chemical reaction in aqueous solutions. Some metals (e.g., zinc, chromium and aluminum) are applied to sintered metals by diffusion processes [146].

Coatings Applied from the Gaseous Phase. Steam treatment is an inexpensive, widely employed technique for producing corrosion-resistant Fe_3O_4 coatings on sintered iron and steel. The sintered parts are treated for ca. 2 h in a saturated steam atmosphere at ca. 550 °C, resulting in the formation of a dense, blue-black Fe_3O_4 layer. Steam treatment is sometimes performed at the same time as the tempering treatment after hardening [147], [148]. Steam treatment can also be used to a certain extent to close porosity.

Gas nitriding, plasma nitriding, and nitrocarburizing can be carried out to improve wear resistance if the porosity has already substantially decreased [149–151].

Surface Strengthening. As in the case of melt metallurgy materials, the mechanical properties (particularly dynamic properties) of sintered steels can be improved by the generation of compressive forces in subsurface regions. Surface strengthening can be achieved by surface rolling or shot peening [151–155].

Surface Remelting. Attempts have been made to strengthen sintered steel by using high-energy laser beams for remelting. Remelting leads to a completely densified surface layer up to 1 mm thick. This process may be suitable for providing sintered steel parts with extremely strong surfaces in localized areas [156].

Heat Treatment. Sintered steels are heat treated to obtain specific strength properties. The end result is influenced by the alloy composition, the atmosphere, and temperature control. The porosity must be considered if heating is accompanied by carburizing or nitrocarburizing treatment. To prevent complete carburizing with case hardening, the porosity may have to be sealed [145], [157–160].

Joining of Sintered Iron and Steel. Solid joints between parts made of sintered steel or between parts made of sintered steel and melt metallurgical materials are used as positive-, nonpositive-, and material-locking joints. A wide range of methods is available for material-locking joints. All inert-gas welding processes may be used. Recent developments have made uniaxially pressed materials accessible to electron-beam and laser welding [156], [161]. Direct-current path resistance welding is a very elegant method, to increase the local current density at the joining site; hump-shaped projections can be produced on the sintered part with specially designed tools [25], [162].

Joints by brazing are difficult to produce in porous sintered steels since the pores may absorb the brazer due to capillary forces. The brazer should therefore be carefully selected to suit the application [26], [27], [163]. Recent developments have been aimed at producing joints by means of diffusion processes during sintering [163], [164].

Frictional welding is also used. The parts to be joined are pressed together and moved relative to each other. The resulting frictional heat is utilized to produce the weld joint [165], [166].

5. Strength Properties

Widely differing structures of sintered steels are formed, depending on the choice of raw material, alloy composition, compaction and sintering processes, and secondary operations. The structures of six arbitrarily selected materials having densities between 6.9 and 7.4 g/cm^3 give some idea of possible microstructures (Fig. 17).

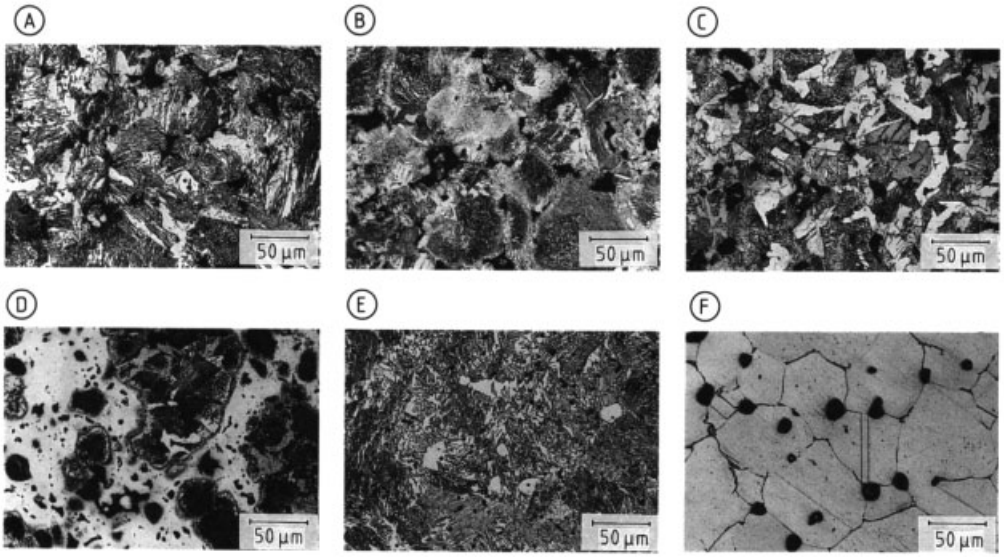

Figure 17. Microstructure of six PM alloys
A) Fe – C; B) Fe – Cu – C; C) Fe – Cu – P – C; D) Fe – Cu – Ni – Mo – C; E) Fe – C; F) Fe – Cr – Ni – Mo (AISI 316 L)
A – D: Die pressed; E, F: Injection molded

A comprehensive collection of structure photographs of PM materials is given in [167]. The structure – property relationships of fully dense sintered materials are the same as those of melt metallurgical materials. Discussion is therefore limited to the special relationships applying to uniaxially pressed materials, in which porosity is the primary influencing factor. The special effects of alloying elements on strength occur only at relatively high densities.

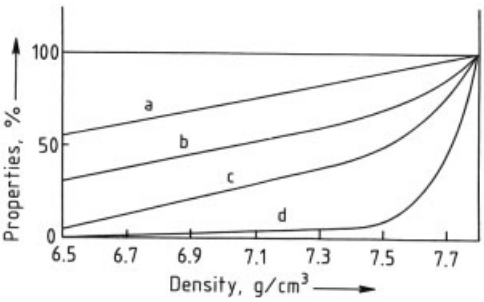

Figure 18. Schematic diagram showing the relationship between the strength and density of sintered steels
a) Tensile strength; b) Bending fatigue limit; c) Elongation; d) Impact strength

An analytical description of the relationship between porosity and mechanical properties is extremely difficult because of the formation of complex three-dimensional stress states [168–170]. Most current knowledge is based on empirical results.

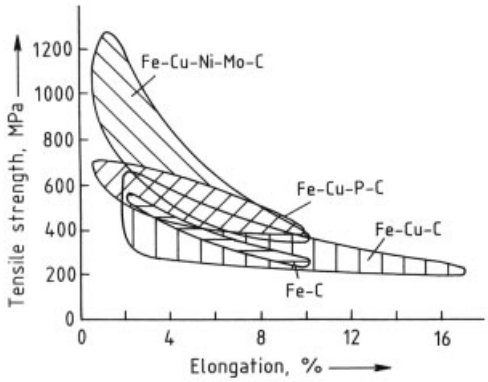

Figure 19. Tensile strength as a function of elongation for various sintered steel alloys (test method ISO/DIS 6892, specimen ISO/DP 2740)

The relationships between material density and strength properties are shown in Figure 18. Tensile strength is a linear function of density [65], but the bending fatigue limit and elongation at break increase disproportionately at high material densities. These relationships must always be taken into account in selecting materials. The tensile strength, 0.2 % yield

strength, and impact strength are shown for some important alloy systems in Figures 192021, respectively. The variations are due to variations in the alloy composition, the pressing and sintering conditions, and the final heat treatment conditions. These diagrams clearly illustrate the widely varying properties of uniaxially pressed sintered steels. DIN 30 912 specifies minimum strength values for standard sintered steels.

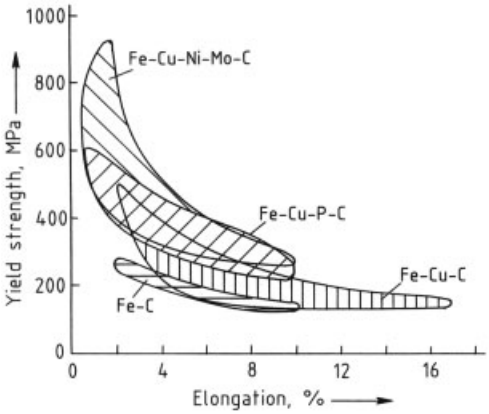

Figure 20. Yield strength as a function of elongation for various sintered steel alloys (test method ISO/DIS 6892, specimen ISO/DP 2740/1973)

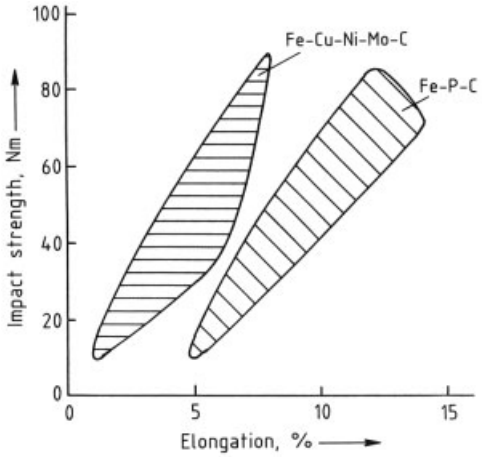

Figure 21. Impact strength as a function of elongation for various sintered steel alloys (test method ISO/DIS 140.4, specimen ISO/DP 5754)

Continuous improvements in the strength of uniaxially pressed sintered steels mean that they can now also be used for dynamically highly stressed parts. This development has been accompanied by detailed investigations of strength and fracture behavior under vibratory stress. Under certain conditions, sintered steels containing the alloying elements copper, nickel, carbon, and phosphorus have fatigue strengths and fracture-mechanical values equivalent to those of compact steels and cast iron containing spheroidal graphite. This behavior is attributed to the crack-delaying action of the pores and the resultant notch insensitivity. Subsequent surface strengthening and heat treatment (e.g., nitrocarburizing) substantially improve the fatigue strength. The measured values provide the basis for fatigue strength dimensioning of PM parts [171–176]. This knowledge has permitted uniaxially pressed materials to be incorporated in critical components used for automobile construction. Sinter-forged connecting rods have been manufactured for a fairly long time in prototype-size and small-batch production runs. In the United States and Japan they have already been used in large-scale engine production. Developmental work is being carried out worldwide on "joined camshafts" with cams made from PM steel. Relatively expensive sintered high-alloy cams are also being mass produced. Intensive development is being carried out in cheaper, sintered, low-alloy cams.

6. References

1. A. von Mussin-Puschkin: "Eine neue Methode die Platina zu schmieden," *Allg. J. Chem.* **4** (1800) 411 ff.
2. F. Skaupy: *Metallkeramik – Die Herstellung von Metallkörpern aus Metallpulvern,* Verlag Chemie, Berlin 1930.
3. R. Kieffer, W. Hotop: *Pulvermetallurgie und Sinterwerkstoffe,* Springer Verlag, Berlin 1943.
4. F. Thümmler, Schunk Werkstoffe GmbH, Berichte aus Forschung und Entwicklung, no. 41, Heuchelheim 1988, pp. 25 – 35.
5. B. L. Mordike, K. U. Kainer, D. Burchards: "Herstellung und Verarbeitung schnell abgeschreckter Pulver auf Eisenbasis," *DGM Fortbildungsseminar Pulvermetallurgie,* Karlsruhe 1986.
6. K. Isonishi, M. Tokizane: "Superplastic Deformation of a P/M Ultrahigh Carbon Steel," *Int. J. Powder Metall. Powder Technol.* **25** (1989) no. 3, 187 – 196.

7. H. D. Hederich: "Properties and Applications of Iron-Base ODS Alloys" in E. Arzt, L. Schulz (eds.): *New Materials by Mechanical Alloying Techniques,* DGM Informationsgesellschaft, Oberursel 1989.
8. P. R. Brewin, P. D. Nurthen, B. Toloui: "Developments in Hard Wearing Components," *Met. Mater.* (1988) Aug., 470 – 473.
9. W. Graf, K.-J. Krämer, J. Pötschke, W. Weiglin: "The Production and Processing of High-Quality Powder Metallurgy Materials," *Powder Metall Int.* **23** (1991) no. 4, 246 – 252.
10. I. Ucok, T. Ando, N. J. Grant: "Structure and Properties of Spray Formed Stainless Steel," *Int. J. Powder Metall. Powder Technol.* **27** (1991) no. 3, 237 –247.
11. J. Pötschke, W. G. Weiglin: "Pulvertechnologisch Hergestellte Hochtemperaturwerkstoffe und Bauteile," *VDI Ber.* **4** (1987) no. 600, 311 – 344.
12. T. Hollstein, G. Schumacher, C. J. Franklin, K. Torssell: "Mechanisches Verhalten von pulvermetallurgisch hergestellten Dampfleitungskomponenten aus 9 – 12 % Cr-Stählen." *2. Sympossium Materialforschung,* Dresden, Aug. 26 – 29, 1992.
13. J. J. Fischer, R. M. Haeberle, G. Barten: "Markstellung mechanisch legierter Werkstoffe," *Ingenieurwerkstoffe* **4** (1992) no. 7/8, 16 – 19.
14. Robert Zapp Werkstofftechnik GmbH & Co KG, EP 0 149 210 B1, 1992.
15. L. F. Pease III, R. J. Sansoucy: "Powder Injection Molding," *Advances in Powder Metallurgy,* **vol. 2,** MPIF and APMI, Princeton 1991.
16. H. Cohrt, M. Enders, L. Kramer: "Large Scale Production of Stainless Steel Gears by Metal Injection Molding," *Powder Metallurgy World Congress,* San Francisco, June 21 – 26, 1992.
17. K. F. Hens, D. Lee, S. T. Lin, R. M. German: "Integrity of Complex Shape Products by Powder Injection Molding," *Powder Metall Int.* **23** (1991) no. 1, 15 – 21.
18. M. Jeandin, J. P. Trottier, J. L. Kontny, Y. Bienvenu: "A Two Stage Compaction Process for Superalloy Powders: Presintering and HIP," *Met. Powder Report* **39** (1984) no. 9, 495 – 500.
19. "Advances in Powder Consolidation at Ceracon," *Met. Powder Report* **45** (1990) no. 4, 39 – 40.
20. B. L. Ferguson, A. Kuhn, O. D. Smith, A. Hofstatter: "Hot Consolidation of Porous Preforms Using 'Soft' Tooling," *Int. J. Powder Metall. Powder Technol.* **20** (1984).
21. G. Bockstiegel: "Powderforging – The Development of the Technology and its Acceptance in North America, Japan and West Europe" in H. Huppmann et al. (eds.): *Powder Metallurgy in Science and Practical Technology,* vol. 2, Verlag Schmidt, Freiburg 1986, pp. 101 – 122.
22. W. J. Huppmann: "The Technical and Economic Development of Powder Forging, *Powder Metall Int.* **24** (1992) no. 3, 186 – 193.
23. T. Tabata, S. Masaki, H. Suzuki, B. Q. Zhu: "Bond Strength of Bonded Compostits Preforms," *Int. J. Powder Metall. Powder Technol.* **25** (1989) no. 1, 37 –42.
24. K. Okimoto, T. Satoh: "Joining of Iron Powder Compacts by an Infiltration Method," *Int. J. Powder Metall. Powder Technol.* **23** (1987) no. 3, 163 – 170.
25. R. Wähling: "Löten von Sinterteilen," *Herbstsymposium Pulvermetallurgie 1992,* Hagen 1992.
26. B. Krentschner, V. Arnhold, K. Hummert: "Pulvermetallurgie – Eine Verfahrenstechnik zur Herstellung von Verbundwerkstoffen und Verbundbauteilen," *VDI Ber.* **3** (1989) no. 600, 215 – 230.
27. C. Nissel: "Verbundwerkstoffe und Verbundbauteile durch Heißisostatisches Pressen" in G. Ondracek (ed.): *Verbundwerkstoffe – Technologie und Prüfung,* vol. 2, DGM-Verlag, Oberursel 1990, pp. 73 – 94.
28. P. K. Johnson: "P/M Applications Diversifing into New Markets," *Int. J. Powder Metall.* **28** (1992) no. 3, 233 – 241.
29. *Fachverband für Pulvermetallurgie,* Hagen 1990.
30. *Global Assesment of the Current Status and Longer Range Prospects for the Powder Metallurgy Buisness,* Gorham Advanced Materials Institute, Maine 1990.
31. DIN Deutsches Institut für Normen e.V.: *DIN-Taschenbuch 247 – Pulvermetallurgie,* Beuth Verlag GmbH, Berlin 1991.
32. F. V. Lenel: *Powder Metallurgy – Principles and Applications,* MPIF, Princeton 1988, pp. 21 ff.
33. Powder Metallurgy OEEC, *Technical Aid Mission, Paris* **141** (1955) 146 – 149.
34. E. Sieurin, *Stahl Eisen* **31** (1911) 1391 – 1394.
35. E. Sieurin, *Stahl Eisen* **32** (1912) 830 – 841.

36. S. Eketorp, *Jernkontorets Ann.* **129** (1945) 703 – 721.
37. *ESSO Oilways* **23** (1957) 5.
38. G. Naeser, H. Steffe, W. Scholz, *Stahl Eisen* **68** (1948) 346 – 353.
39. M. Michalke, W. Scholz, *Chem. Ing. Tech.* **41** (1969) 16 – 22.
40. J. M. Capus, W. F. Fossen: "The Powder Metallurgy Industry in Canada," *Int. J. Powder Metall. Powder Technol.* **18** (1982) 335 – 346.
41. S. I. Hulthen, *Int. J. Powder Metall. Powder Technol.* **17** (1981) 345 – 346.
42. W. Kaysser: "Prinzipien der Herstellung optimierter Metallpulver durch Schmelzzerteilung," *2. Symposium Materialforschung,* Dresden 1991.
43. J. F. Watkinson, *Powder Metall.* **1/2** (1958) 13 – 23.
44. P. U. Gummeson, *Powder Metall.* **15** (1972) 67 –94.
45. R. P. Hervey, *Met. Powder Report* **35** (1980) 14 – 16.
46. Powdrex, DE-OS 2 813 018, 1978 (D. R. Gui).
47. K. Ogura, S. Takajo, N. Yamato, Y. Maeda: "Low Oxygen Alloy Steel Powder by Vacuum Annealing" in W. Kaysser, W. J. Huppmann (eds.): *Horizons of Powder Metallurgy,* part I, Verlag Schmidt, Freiburg 1986, pp. 37 – 40.
48. T. Takeda: *Proc. Lecture Meeting at National Research Institute for Metals,* Japan 1984, p. 1.
49. S. Takajo: "Innovations in Ferrous Powders and Their Production" in H. Huppmann et al. (eds.): *Powder Metallurgy in Science and Practical Technology,* vol. 2, Verlag Schmidt, Freiburg 1986, pp. 11 – 40.
50. E. Tamura et al., *Proc. Japan Soc. Powder and Powder Met.* **52** (1983) p. 92.
51. I. Karasuno et al.: "Some Properties of Oil Atomized Low Alloy Steel Powders Containing Chromium" in [47] , p. 53.
52. M. J. Hohmann: "New Powder Production Methods for Reactive Materials," *DGM Fortbildungsseminar Pulvermetallurgie,* Dresden 1992.
53. M. Hohmann, K. Mosblech, N. Ludwig: "Modern System for Ceramic-Free Powder Production," *Powder Metallurgy World Congress,* San Francisco 1992.
54. J. Bäcker, W. Graf, J. Pötschke, W. Kavasters: "Die Herstellung von Metallpulver durch Verdüsen der Schmelze mit tiefkaltem Flüssiggas," *Metall (Berlin)* **45** (1991) .
55. M. Hohmann, S. Johnsson, E. Lierke: "Metal Pow der Production by Ultrasonic Capillary or Standing Wave Technique," *International Powder Metallurgy Conference,* Orlando 1988.
56. C. P. Ashdown, I. G. Bewley: "Tandem Atomization Process for Producing Ultrafine Metal Powder," *International Powder Metallurgy Conference,* Orlando 1988.
57. N. N.: "Company Profile," *Powder Metall Int.* **24** (1992) no. 4, 256 – 257.
58. P. R. Sahm, L. Kallien, D. Storck: "Schmelzverdüsung metallischer Legierungen," *2. Symposium Materialforschung,* Dresden 1991.
59. *Handbook of Metals,* 9th ed., **vol. 7,** ASM, Metals Park, Ohio 1978, pp. 93 – 96.
60. D. I. Blömacher: *Carbonyl Iron Powders: Its Production and New Developments,* MPR (1990) Feb., pp. 117 – 119.
61. J. E. Japka: "Iron Powder for Metal Injection Molding," *Int. J. Powder Metall. Powder Technol.* **27** (1991) no. 2, 107 – 114.
62. J. E. Japka, R. M. German: "Carbonyl Iron Powders for Metal Injection Molding – Recent Research Developments," *Adv. Powder Metall.* **2** (1992) 15 – 32.
63. S. Brunnauer, P. H. Emmnett, E. Teller, *J. Am. Chem. Soc.* **60** (1938) 309.
64. E. E. Underwood: *Quantitative Stereology,* Addison-Wesly Publishing Company, Berkley 1970.
65. G. Zapf, K. Dalal: "Die Metallpulver – Herstellung und Eigenschaften," in *Fachverband für Pulvermetallurgic* (ed.): *Die Pulvermetallurgie – Vorlesungsreihe,* Hagen 1986.
66. R. P. Beliles: "Toxicity of Metal Powders" in [59] , 201 – 208.
67. C. J. Dahn: "Explosivity and Pyrophoricity of Metal Powders" in [59] , 194 – 200.
68. H. Danninger, G. Jangg, M. Giahi: "Production of Stainless Steel Filters by Liquid Phase Activated Gravity Sintering," *Mat. Wiss. u. Werkstofftech.* **19** (1988) 205 – 211.
69. R. M. Pillar, H. U. Cameron, I. McNab: "Porous Surface Layered Prothetic Devices," *Biomed. Eng.* **10** (1975) 126 – 131.
70. J. S. Hirschhorn, J. T. Reynolds: "Powder Metallurgy Fabrication of Cobalt Alloy Surgical Implant Materials" in E. Korostoff (ed.): *Research in Dental and Medical Materials,* Plenum Press, New York 1969, pp. 137 – 151.
71. W. Schatt: *Pulvermetallurgie und Verbundwerkstoffe,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1977, pp. 58 – 59.
72. Dorst, company brochure, Kochel am See 1990.

73. P. Beiss: "Möglichkeiten und Grenzen der Formgebung metallischer Pulver durch Matrizenpressen" in H. Kolaska (ed.): *Fortschritte bei der Formgebung in Pulvermetallurgie und Keramik,* VDI-Verlag, Düsseldorf 1991, pp. 7 – 32.
74. L. L. Confer, J. P. Straub: "Shape Fundamentals: Rigid Tool Compaction" in [59] , pp. 322 – 328.
75. G. Bockstiegel, J. Hewing: "Kritische Betrachtung des Schrifttums über den Verdichtungsvorgang beim Kaltpressen von Pulvern in starren Preßformen," *Archiv. Eisenhüttenwes.* **36** (1965) 751 –767.
76. E. Fischer: "Optimization of the Compacting Process of P/M Products," *Powder Metall Int.* **20** (1988) no. 1, 32 – 34; *Powder Metall Int.* **20** (1988) no. 2, 27 – 30.
77. H. Kolaska, P. Schulz, P. Beiss, E. Ernst: "Untersuchungen zum Matrizenpressen" in [73] , pp. 50 – 74.
78. E. Ernst, F. Thümmler, P. Beiss, V. Arnhold: "Untersuchungen zur Einführung keramischer Preß- und Kalibriermatrizen" in [73] , pp. 75 – 93.
79. A. B. van Groenou, R. C. D. Lissenburg: "Inhomogeneous Density in Die Compaction: Experiments and Finite-Element Calculation," *J. Am. Ceram. Soc.* (1983) 156 – 158.
80. E. Ernst, F. Thümmler, P. Beiss, V. Arnhold: "Untersuchungen zur Einführung Keramischer Preß- und Kalibriematrizen" in [73] , pp. 75 – 39.
81. P. Lindskog: "Das kaltisostatische Pressen bei der Formgebung metallischer Pulver – Wet- und Dry-bag-Verfahren," in [73] , p. 14.
82. T. Kanda, Y. Sakashita, Y. Narukawa, T. Naol: "Recent Developments and Applications of HIP and CIP Equipment," *4th Int. Conf. on Isostatic Pressing,* London 1990.
83. A. Fischer, V. Arnhold: "Pulvermetallurgische Bauelemente in der Verfahrenstechnik, *Chem. Ing. Tech.* **62** (1990) no. 4, 263 – 270.
84. P. R. M. Skoglund: "Tolerances for PM Cylinder Liners," *Met. Powder Report* (1989) 350 – 354.
85. P. Barth: *Pulvermetallurgisches Spritzgießen – Technologie zur Herstellung komplexer metallischer Bauteile,* Verlag TüV Rheinland, Köln 1988.
86. R. M. German: *Powder Injection Molding,* MPIF, Princeton 1990.
87. H. Cohrt, M. Enders: "Pulvermetallurgisches Spritzgießen – Zielsetzung, Werkstoff- und Bauteileigenschaften," *VDI Ber.* **852** (1991) 441 – 456.
88. ASM Committee on Roll Compacting: "Roll Compacting of Metal Powders" in [59] , pp. 401 – 409.
89. J. H. Tundermann: "The Roll Compaction of Iron Powder," Ph. D. Thesis, University College, Swansea 1967.
90. G. Naeser, F. Zirm, *Stahl Eisen* **70** (1950) 995 –1003.
91. A. Ruder, H. P. Buchkremer, R. Hecker, D. Stover: "Wet Powder Pouring – A New P/M Net Shape Process," *Powder Metallurgy World Congress,* San Francisco 1992.
92. W. Schatt: *Sintervorgänge – Grundlagen,* VDI-Verlag, Düsseldorf 1992.
93. W. Schatt, M. Hinz: "On the Generalization of Defect-Activated Sintering," *Powder Metall Int.* **20** (1988) no. 6, 17 – 20.
94. F. J. Esper: "Flüssigphasensintern und Korngrenzenphasen bei Sinterstählen und Keramik," *Powder Metall Int.* **20** (1988) no. 6, 43 – 48.
95. W. A. Kaysser: "Optimiertes Sintern metallischer und keramischer Werkstoffe" in H. Kolaska (ed.): *Konsolidierung und Wärmebehandlung von Sinterwerkstoffen,* Verlag Schmidt, Freiburg 1990, pp. 21 –48.
96. R. German: "Supersolidus Liquid Phase Sintering," *Int. J. Powder Metall.* **26** (1990) no. 1, 23 – 34; *Int. J. Powder Metall.* **26** (1990) no. 1, 35 – 44.
97. H. S. Nayar, M. Berlinger, Y. T. Chen: "Studies in Dimensional Change in Iron/Cu Parts in Different Zones and Under Different Atmospheres," *Powder Metallurgy World Congress,* San Francisco 1992.
98. A. Griffo, R. M. German, H. S. Nayar: "Powder Selection and Sintering Pathways for Zero Dimensional Change in Fe-2Cu-0a.8C," *Powder Metallurgy World Congress,* San Francisco 1992.
99. T. Pieczonka, A. Cias: "The Necklace Structure in Sintered Materials Fe-P and Fe-Cu-P," *Powder Metall. Int.* **22** (1990) no. 2, 13 – 15.
100. H. Miura, Y. Tokunaga: "The Effect of Phosphorus Additions on the Structure and Mechanical Properties of Sintered Iron Compacts," *Int. J. Powder Metall.* **21** (1985) no. 4, 269 – 282.
101. L. E. G. Cambronero, J. M. Torralba, J. M. Ruiz-Pieto: "Microstructure and Properties of PM Cu-P-Steels," *Powder Metall. Int.* **22** (1990) no. 2, 26 –29.

102. A. Cias, J. Frydych, T. Pieczonka: "Liquids Phase Sintering of Iron Phosphorus Systems," *Powder Metall Int.* **21** (1989) no. 3, 13 – 15.
103. A. Cias, J. Frydych, T. Pieczonka: "Liquids Phase Sintering of Iron Phosphorus Systems," *Powder Metall. Int.* **21** (1989) no. 3, 13 – 15.
104. P. Lindskog: "The Effect of Phosphorus Additions on the Tensile, Fatigue, and Impact Strength of Sintered Steels Based on Sponge Iron Powder and High-Purity Atomized Iron Powder," *Powder Metall.* **16** (1973) 374.
105. K. Frisk, C. Lindberg, A. Johansson: "Nitrogen Pick Up During Sintering of Stainless Steel," *Powder Metallurgy World Congress,* San Fancisco 1992.
106. D. Warga: "Kontrollierte Ofenatmosphären beim Sintern" in [95] , pp. 75 – 92.
107. T. Phillips: "Trouble-Shooting Guide for Sintering Furnace Atmospheres," *Int. J. Powder Metall.* **26** (1990) no. 3, 245 – 252.
108. D. W. Smith, J. Mashl: "Effects of Preheat Atmosphere and the Oxygen Content and Properties of P/F 4650 Steel," *Powder Metallurgy World Congress,* San Francisco 1992.
109. K. H. Moyer, R. Jones, D. Thorne: "Vacuum Sintering – A Way to Improve Ductility and Toughness of Structural P/M Parts at a Right Price," *Powder Metallurgy World Congress,* San Francisco 1992.
110. H. I. Sanderow, B. Cole: "Properties of Ferrous Alloys Using a Ceramic Belt High Temperature Sintering Furnace," *Powder Metallurgy World Congress,* San Francisco 1992.
111. M. Weber: "Verfahrensvergleich und Wirtschaftlichkeit bei der Herstellung von Pleueln – Guß, Schmieden, Pulverschmieden, CFK, PM-Aluminium" in [73] , pp. 119 – 140.
112. M. Weber, L. Albano-Müller, W. J. Huppmenn: "Truck Synchronizer Rings Produced by Powder Forging," *SAE-Technical Paper* **no. 860153,** 1986.
113. K. Imahashi, C. Tsumuki, I. Nagasre: "Development of Powder Forged Connecting Rods," *SAE-Technical Paper* **no. 841221,** 1984.
114. J. H. Moll: "Full Density P/M Shapes for Defense Applications" in K. M. Kulkarni (ed.): *Powder Metallurgy for Full Density Products,* MPIF, Princeton 1987, pp. 163 – 188.
115. P. E. Price, S. P. Kohler: "Hot Isostatic Pressing of Metal Powders" in [59] , pp. 419 – 443.
116. A. C. Nyce: "Containerless HIP'ing of P/M Parts: Technology, Markets, and Economics" in [114] , pp. 123 – 148.
117. H. Kolaska, K. Dreyer: "Literaturstudium: Sinter/HIP-Technologie und Gasdrucksintern" in H. Kolaska (ed.): *Sinter/HIP-Technologie,* Verlag Schmidt, Freiburg 1987.
118. R. Ekbom: "Hot Isostatic Pressing – Powder Metallurgy in Gas and Steam Turbines," *Mater. Des.* **11** (1990) 37 – 42.
119. D. Schuler, O. Jung, B. Frisch, W. R. Thiele: "Einfluß gasförmiger Reaktionsprodukte beim Sintern und heißisostatischen Pressen von Hartmetallen und Schnellarbeitsstählen" in [117] , pp. 113 – 132.
120. E. Grallath, H. M. Ortner: "Argon-Spurenanalyse in Sinter/HIP- und HIP-Teilen sowie in unter Argon gemahlenen Pulvern" in [117] , pp. 133 – 152.
121. K. T. Kim: "Hot Compaction Equation for Metal Powders," *Int. J. Powder Metall.* **24** (1988) no. 1, 31 – 38.
122. A. M. Laptev, N. V. Samarov, S. V. Podlesny: "Theory and Optimization of Hot Isostatic Pressing," *Powder Metall. Int.* **22** (1990) no. 2, 23 – 25.
123. W. A. Kaysser, M. Aslan, G. Petzow: "Verdichtung und Gefügeentwicklung beim HIPen" in [117] , pp. 67 – 86.
124. E. Arzt: "Optimierung der HIP-Parameter under Berücksichtigung der Verdichtungsmechanismen," *Powder Metall Int.* **17** (1985) 39 – 42.
125. K. E. Meiners, J. C. McCall: "Recent Developments in HIP Technology" in [114] , pp. 99 – 148.
126. R. Danzer, G. Hachl, F. Jeglitsch, H. Konvicka: "Reduction of Scattering of Mechanical Properties and of the Probability to Failure by HIPping" in [21] , pp. 677 – 680.
127. T. Hisada, K. Kusaka, K. Hayashi: "AISI 316 L Stainless Steel Pipe Produced by Powder Extrusion Process" in [21] , pp. 325 – 334.
128. G. Guntz, J. P. Peter, D. Vuillaume: "ANVAL P/M Tubes" in [21] , pp. 335 – 338.
129. A. Muggeo, G. Pitois, D. Vuillaume: "Studies of the Extrusion Consolidation in the ANVAL P/M Process" in [21] , pp. 339 – 342.
130. R. J. Causton, J. J. Dunkley: "Extrusion of Powder Metallurgy High-Speed Steel" in [114] , pp. 53 –76.
131. A. G. Leatham, R. G. Brooks, M. Yaman: "The Osprey Process for the Production of Spray-Deposited Roll, Disk, Tube and Billet Preforms" in [114] , pp. 535 – 558.

132. R. Flender, K. Wünneberg, R. Fuchs, D. Zebrowski: "Erzeugung von Flachprodukten aus Stahl und Sonderlegierungen mit dem Sprühkompaktierverfahren," *2nd Sympossium Materialforschung,* Dresden 1991.
133. R. M. Rusnak: "Energy Relationship in High Velosity Compacting of Copper Powder," *Int. J. Powder Metall. Powder Technol.* **12** (1976) no. 2, 91 – 99.
134. E. Y. Gutmanas: "High Pressure Compaction and Cold Sintering of Stainless Steel Powders," *Powder Metall Int.* **12** (1980) no. 4, 178 – 182.
135. R. Prümmer: *Explosivverdichtung pulvriger Substanzen,* Springer Verlag, Berlin 1987.
136. V. Mironov: "Densification of Powder with Magnetic Impulses," *Planseeber. Pulvermetall.* **24** (1976) 175 – 190.
137. J. W. Hagemeyer, J. A. Regalbuto: "Dynamic Compaction of Metal Powders with a High Velosity Impact Device," *Int. J. Powder Metall.* **4** (1968) no. 3, 19 – 25.
138. P. J. James: "The Machinability of Sintered Steels," *Powder Metall. Int.* **22** (1990) no. 6, 19 – 23.
139. J. S. Agapiou, M. F. DeVries: "Machinability of P/M Materials," *Int. J. Powder Metall.* **24** (1988) no. 1, 47 – 58.
140. Y. T. Chen, R. Baudisch, L. Ceri: "Free-Machining P/M Alloy Optimization Using Statistical Analytical Techniques – The Effect of MnS Content and Particle Size," *Powder Metallurgy World Congress,* San Francisco 1992.
141. D. S. Madan: "Effect of MnS on Properties of High Performance P/M Alloys and Applications," *Powder Metallurgy World Congress,* San Francisco 1992.
142. M. Svilar, E. Klar: "Impact Strength and Fatigue Properties of Copper Infiltrated P/M Steel," *Int. J. Powder Metall.* **22** (1986) no. 2, 105 – 112.
143. W. K. Fricke: "Copper Infiltration of Water Atomized Iron Powder," *Powder Metallurgy World Congress,* San Francisco 1992.
144. C. G. Goetzel: "Infiltration" in [59] , pp. 549 – 566.
145. H. J. Mäurer, W. Bodden: "Effect of Waterglass Impregnation Upon Hardenability of Porous Sintered Steel," *Powder Metall Int.* **15** (1983) no. 3, 122 – 125.
146. L. Pleney, M. Eudier: "Secondary Operations for PM Parts" in W. Huppmann, W. Kaysser, G. Petzow (eds.): *Powder Metallurgy 1986 – State of the Art,* Verlag Schmidt, Freiburg 1986, pp. 123 – 136.
147. P. Beiss: "Steam Treatment of Sintered Parts," *Proceedings of the Int. Conf. on Powder Metallurgy,* London 1990, pp. 488 – 496.
148. L. Ledoux: "Mechanical Behavior of Steam-Treated Sintered Iron," *Proceedings of the Int. Conf. on Powder Metallurgy,* London 1990, pp. 353 – 358.
149. M. Rosso, G. Scavino: "Studies on Gas-Nitrocarburizing Processes of Sintered Ferrous Materials: Influence of Treating Parameters" in [21] , pp. 369 – 373.
150. K.-T. Rie, T. Lampe, S. Eisenberg: "Surface Treatment of P/M Materials by Plasma Nitrocarburizing" in [21] , pp. 361 – 364.
151. W. Rembges, W. Zimmermann: "Influence of Material and Density upon Ionitriding of P/M Parts," *Powder Metallurgy World Congress,* San Francisco 1992.
152. Y. Takeya, T. Hayasaka, M. Suzuki, A. Kamada: "Surface Rolling of Sintered Gears," *SAE Tech. Pap. Ser.* **no. 820234,** 1984.
153. D. Tyler: "Ferrous PM-Parts Improved by Controlled Shut Peenig," *Met. Powder Report* **11** (1988) 773 – 774.
154. C. M. Sonsino, J. Tengzelius, G. Schlieper: "Influence of As-Sintered Material Strength on the Improvement of Fatigue Behavior by Surface Rolling," *Proceedings of the Int. Conf. on Powder Metallurgy,* London 1990, pp. 497 – 508.
155. H. Steindorf: "Schwing- und Wälzfestigkeitseigenschaften von Sinterstählen unter optimierten Festwalzbedingungen," *Fortschr. Ber. VDI Reihe 5,* **245** (1991) .
156. K. Schekulin, R. Jenter: "Oberflächenbearbeitung mit CO_2-Laser," *Laserbearbeitung von Sintereisen,* BMFT-Forschungsbericht zum Verbundprojekt, 1992.
157. D. Pohl: "Untersuchungen über Zähigkeitssteigernde Wärmebehandlungen an Sinterstählen" in [95] , pp. 93 – 113.
158. V. Arnhold, T. Mayr, R. Wähling: "Wärmebehandlung von Sinterstahl" in [95] , pp. 129 – 159.
159. V. B. Phadke, B. L. Davies: "Precipitation Hardening in Sintered Iron Copper Alloys," *Powder Metall. Int.* **9** (1977) no. 2, 64 – 66.
160. H. Mäurer: "Influence of the Microstructure on the Hardenability of Porous P/M Steels" in [23] , pp. 411 – 415.
161. Y. Murai, E. Takahashi, H. Isako: "Development of the Technology of Laser

Welded Application for Sintered Steel Parts," *Powder Metallurgy World Congress,* San Francisco 1992.
162. H. B. Boyer: "Secondary Operations Performed on P/M Parts and Products" in [59] , pp. 451 – 462.
163. B. Clausen, K. A. Torsen: "Joining of Iron Based P/M Components by Brazing and Diffusion Bonding," *Powder Metallurgy World Congress,* San Francisco 1992.
164. J. Sakai, K. Asaka: "P/M Diffusion Bonding" in [21] , pp. 717 – 721.
165. DVS-Merkblatt 2909, part 1, Deutscher Verband für Schweißtechnik e.V., Düsseldorf 1989.
166. H. Horn: "Reibschweißen von Werkzeugstahl mit Wolframcarbid-Hartmetallen," *Industrieanzeiger* **111** (1989) no. 63, 14 – 16.
167. W. Huppmann, K. Dalal: *Metallographic Atlas of Powder Metallurgy,* Verlag Schmidt, Freiburg 1986.
168. G. F. Bocchini: "The Influence of Porosity on the Characteristics of Sintered Materials," *Int. J. Powder Metall.* **22** (1986) no. 3, 185 – 202.
169. G. Ondracek: "Porosität und Eigenschaften von Sinterwerkstoffen," *DGM Fortbildungsseminar Pulvermetallurgie,* Karlsruhe 1986.
170. H. Danninger, G. Jangg, B. Weiss, R. Stickler: "The Influence of Porosity on Static and Dynamic Properties of P/M Iron," *Proceedings of the Int. Conference on Powder Metallurgy,* London 1990, pp. 433 – 439.
171. F. J. Esper, C. M. Sonsino, G. Leuze: "Design Criteria for PM Parts Submitted to Fatigue," *Modern Developments in Powder Metallurgy,* **vol. 15 – 17,** MPIF, Princeton 1985, pp. 719 – 737.
172. I. Bertilsson, B. Karlsson: "Cyclic Deformation Behavior and Low Cycle Fatigue of Sintered Steel," *Powder Metall Int.* **20** (1988) no. 1, 13 – 20.
173. C. M. Soncino, F. J. Espers: "Bemessungskriterien für schwingbeanspruchte Bauteile aus Sinterstahl," *Konstruktion* **37** (1985) no. 1, 1 – 10.
174. C. M. Sonsino: "Einfluß verschiedener Wärmebehandlungsverfahren auf die Schwingfestigkeit von Sinterstahl" in [95] , pp. 113 – 128.
175. S. Klumpp, D. Eifler, O. Vöhringer, E. Macherauch: "Cyclic Deformation Behavior of Sintered Steels," *Powder Metallurgy World Congress,* San Francisco 1992.
176. A. Piotrowski, I. D. Eiffler: "High-Cycle Fatigue Behavior of Sintered Steels," *Powder Metallurgy World Congress,* San Francisco 1992.

Sisal → Cellulose

Skeletal Muscle Relaxants

Barry Dubinsky, The R. W. Johnson Pharmaceutical Research Institute, Spring House, PA 19477-0776, United States

Jeffery B. Press, The R. W. Johnson Pharmaceutical Research Institute, Spring House, PA 19477-0776, United States

1. Peripherally Acting Muscle Relaxants 1
1.1. Nondepolarizing Drugs (Competitive Neuromuscular Blocking Drugs) . . . 2
1.2. Depolarizing Relaxants 5
1.3. Agents Acting Directly on Skeletal Muscle 6
2. Centrally Acting Muscle Relaxants . . 7
2.1. Treatment of Acute Muscle Spasms . 7
2.2. Treatment of Chronic Neurological Spasticity 10
3. References 11

Skeletal muscle relaxant drugs (muscle relaxants) are used in a variety of clinical situations. Choice of the individual agent is made with regard to its mechanism of action and its intended use [1–4].

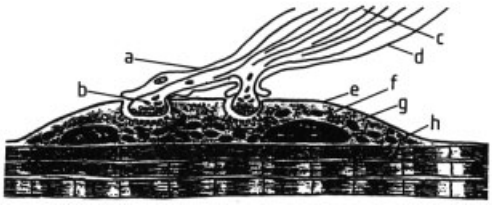

Figure 1. A schematic showing a longitudinal section of a neuromuscular junction (motor endplate)
a) Schwann cell; b) Axon terminal in synaptic trough; c) Axoplasm; d) Myelin sheath; e) Sarcolemma; f) Sarcoplasm; g) Muscle cell nucleus; h) Myofibrils

1. Peripherally Acting Muscle Relaxants

Neuromuscular blocking agents are typically administered intravenously in hospitalized patients and at usual therapeutic doses; they are devoid of central effects including analgesic activity. Their site of action resides in the periphery at specialized receptors for the neurotransmitter, acetylcholine (ACh). At the motor nerve terminal, the impulse releases minute amounts of ACh (ca. 10^{-17} M) from intraneuronal storage vesicles in a Ca^{2+}-dependent process. ACh diffuses across a junction (synapse) to its receptor (the motor endplate) on skeletal muscle to form a receptor complex that opens channels in the cell membrane. This in turn, increases the permeability of the endplate membrane to sodium and potassium ions. The released ACh is rapidly inactivated (hydrolyzed) by acetylcholinesterase, an enzyme that is localized in high concentration in the vicinity of the motor endplate. A schematic diagram of a neuromuscular junction illustrating the motor endplate can be found in Figure 1.

If sufficient ionic sodium is conducted into these cells, a wave of membrane depolarization (action potential) is generated away from the endplate. Transmission of the action potential along the surface of the membrane of the muscle cell and the intracellular release of calcium ions from the sarcoplasmic reticulum leads to muscular contraction. Calcium regulates the contractile process, and adenosine triphosphate is the cellular energy source.

Neuromuscular blocking agents are principally used as adjuncts in surgical anesthesia to obtain satisfactory muscle relaxation, especially of the abdomen, which allows reduced doses of the anesthetic. They also may be administered for orthopedic and other procedures to facilitate muscular manipulation. They may be used during electroshock therapy to prevent bone fractures or compression fractures of the spinal column. Tubocurarine may also be employed in the diagnosis of myesthenia gravis.

The degree of muscle relaxation depends on the dose. The onset of action is rapid, and the sensitivity of muscles to peripheral muscle relaxants varies as a function of the size of the muscle

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_209

and depends on whether the muscle is physiologically fast or slow. The muscles of the eye and face are the first to be paralyzed, followed by the muscles of the neck, larynx, and pharynx. The subsequent paralysis of the extremities and abdominal muscles causes loss of reflex muscular contractions. Only then does paralysis of the intercostal muscles and diaphragm ensue while consciousness is maintained.

1.1. Nondepolarizing Drugs (Competitive Neuromuscular Blocking Drugs)

The curare alkaloids represent the prototypic competitive neuromuscular blocking agents. Curare has been used on arrow tips for hundreds of years by native South Americans to kill animals for food. Death results from muscular paralysis. The mechanism of action of curare was first unraveled in the nineteenth century by Claude Bernard. Tubocurarine and other competitive neuromuscular blocking agents bind to the postjunctional motor endplate receptor making it insensitive to ACh. Since magnesium ions have a direct depressant effect on skeletal muscle, the dose required for neuromuscular blockade is reduced when plasma magnesium concentration is elevated.

These drugs can be competitively antagonized by inhibitors of acetylcholinesterase, such as neostigmine or edrophonium chloride, that increase the concentration of ACh in the motor endplate region. These acetylcholinesterase inhibitors may slow the heart and stimulate secretions; this may be controlled by atropine. They should be used with caution in asthmatics or patients with bronchoconstriction.

Alcuronium dichloride [*15180-03-7*], 4,4′-didemethyl-4,4′-di-2-propenyltoxiferinium dichloride, $C_{44}H_{50}Cl_2N_4O_2$, M_r 737.81, *mp* 85 – 90 °C. Prepared by either (1) reaction of the non-indole nitrogen of Wieland – Gumlich aldehyde with allyl chloride and dimerization to the diallyl toxiferene chloride or (2) dimerization of Wieland – Gumlich aldehyde to nortoxiferene and subsequent quaternization with allyl chloride [5].

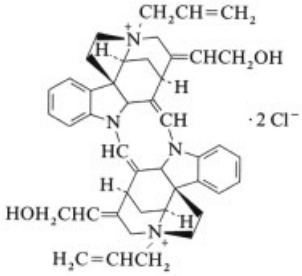

Alcuronium dichloride has a short duration of action (15 – 20 min) and produces less blockade of autonomic ganglia and less histamine release than *d*-tubocurarine. Thus, it produces little or no bronchoconstriction or decrease in blood pressure. Its paralytic effects are intensified by the inhalation anesthetic, halothane.

Trade Names. Alloferin, Toxiferin (Roche).

Atracurium besylate [*64228-81-5*], 2, 2′-[1,5-pentanediylbis[oxy(3-oxo-3, 1-propanediyl)]] bis[1-[3, 4-dimethoxyphenyl)methyl]-1, 2, 3, 4-tetrahydro-6, 7-dimethoxy-2-methylisoquinolinium] dibenzenesulfonate, $C_{65}H_{82}N_2O_{18}S_2$, M_r 1243.5, *mp* 85 – 90 °C. The "reversed ester analogue of Doxcurium" is prepared by Bischler–Napieralski cyclization and sodium borohydride reduction to the 1,2,3,4-tetrahydroisoquinoline followed by Michael-type addition to the bisacrylate ester and alkylation with methyl benzenesulfonate [6].

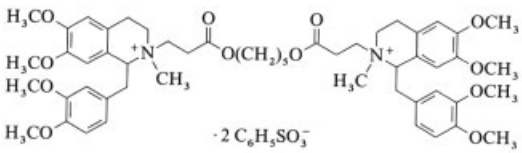

Complete recovery from atracurium besylate takes ca. 30 min but might be longer under halothane anesthesia. It produces less blockade of autonomic ganglia than alcuronium dichloride with minimal histamine release. When the recommended dose is exceeded, moderate histamine release and a significant fall in blood pressure have been observed.

Trade Name. Tracrium (Wellcome).

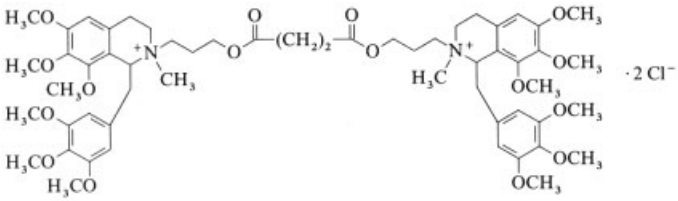

Doxacurium chloride [*106819-53-8*], *trans, -trans*, 2, 2′-[dimethylene-bis-(carbonyloxytrimethyl-ene)-bis-(1, 2, 3, 4-tetrahydro-6,7,8-trimethoxy-2-methyl-1-(3, 4, 5-trimethoxybenzyl)isoquinolinium] dichloride, $C_{56}H_{80}Cl_2N_2O_{16}$, M_r 1108.16. Prepared by reaction of *trans-N*-3-hydroxypropyl-5′,8-dimethoxyaudanosiunium chloride with succinyl chloride [7] (see above).

Doxacurium chloride is a long-acting agent with considerable variability in duration between patients. Its duration is more prolonged in the elderly. The time required for 25 % recovery is highly dose-dependent and may be as long as 160 min after an initial 0.08 mg intravenous dose. Doxacurium chloride does not promote the release of histamine or affect cardiovascular parameters, is not metabolized, and is excreted unchanged in urine and bile. Its paralytic effects are intensified by the inhalation anesthetics, halothane, isoflurane, and enthrane.

Trade Name. Nuromax (Wellcome).

Fazadinium bromide [*49564-56-9*], 1,1′-azobis[3-methyl-2-phenylimidazo [1,2-α] pyridinium] dibromide, $C_{28}H_{24}Br_2N_6$, M_r 604.35, *mp* 218 – 220 °C. Prepared by heating 2-(2-acetylhydrazino)pyridine and 2-bromopropiophenone, hydrolyzing in 24 % hydrobromic acid, and oxidizing with aqueous bromine [8].

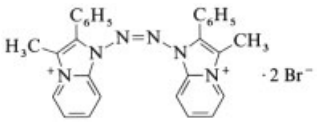

Fazadinium bromide has a rapid onset of action and is completely eliminated in patients with compromised renal function.

Trade Name. Fazodon (Robbins).

Gallamine triethiodide [*65-29-2*], 1,2,3-tris(2-triethylammonium ethoxy)benzene triiodide, $C_{30}H_{60}I_3N_3O_3$, M_r 891.54, *mp* 152 – 153 °C. Prepared by reaction of pyrogallol with diethylaminoethyl chloride and subsequent alkylation with ethyl iodide [9].

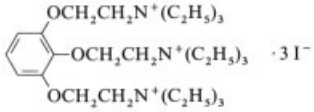

Gallamine triethiodide produces immediate paralysis lasting ca. 30 min after intravenous administration. It does not cause bronchoconstriction or release of histamine and does not affect autonomic ganglia. Since it is excreted unchanged by the kidneys, its plasma clearance is decreased in renal impairment. Its paralytic effects are intensified by diethyl ether or methoxyflurane.

Trade Names. Flaxedil (Specia, May & Baker, Lederle, Rhône-Poulenc), Relaxan (Gea), Tricuran (Deutsches Hydrierwerk).

Metocurine iodide [*7601-55-0*], dimethyltubocurarine iodide, 6,6′,7′,12′-tetramethoxy-2,2,2′,2′-tetramethyltubocuraranium diiodide, $C_{40}H_{48}I_2N_2O_6$, M_r 906.63, *mp* 257 – 267 °C. Prepared by reaction of tubocurarine with methyl iodide.

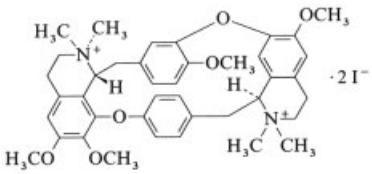

Metocurine iodide possesses similar pharmacologic properties to *d*-tubocurarine but is two to three times more potent. Thus, it should be used

with caution as an anesthetic adjunct. It does not produce the autonomic ganglionic blockade caused by *d*-tubocurarine.

Trade Name. Metubine iodide (Lilly).

Pancuronium bromide [*15500-66-0*], 1,1′-[3,17-bis(acetoxy)-androstane-2,16-diyl] bis [1-methylpiperidinium] dibromide, $C_{35}H_{60}Br_2N_2O_4$, M_r 732.68, *mp* 212 – 215 °C. Prepared by reaction of 2α, 3α: 16α, 17α-diepoxy-17β-acetoxy-5α-androstane with piperidine, and subsequent alkylation with methyl bromide [10].

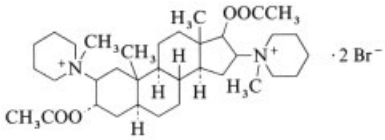

Pancuronium bromide is approximately five times more potent than *d*-tubocurarine. It has minimal effects on the circulatory system and produces little histamine release or hormonal action. Complete recovery from pancuronium bromide takes ca. 65 min but this might be longer under halothane, enthrane, or isoflurane anesthesia. It should be used with caution in patients receiving chronic tricyclic antidepressant therapy because of a potential risk of ventricular arrhythmias.

Trade Names. Mioblock, Pavulon (Organon).

Pipecuronium bromide [*52212-02-9*], pipecurium bromide, 4,4′-[3,17-bis(acetoxy)androstane-2,16-diyl]bis(1,1-dimethylpiperazinium) dibromide dihydrate, $C_{35}H_{62}Br_2N_4O_4 \cdot 2H_2O$, M_r 787.77, *mp* 262 – 264 °C. Prepared by reaction of 2α, 3α: 16α, 17α-diepoxy-17β-acetoxy-5α-androstane with piperazine and subsequent alkylation with methyl bromide [11].

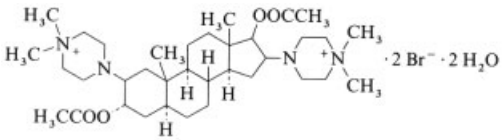

Pipecuronium bromide is a long-acting agent with an onset and duration similar to those of pancuronium bromide. Since it is eliminated primarily unchanged by the kidneys, its clinical duration is prolonged in patients with impaired renal function.

Trade Name. Arduan (Gedeon Richter, Organon).

Tubocurarine chloride [*57-94-3*], *d*-tubocurarine chloride, 7′,12′-dihydroxy-6,6′-dimethoxy-2,2′,2′-trimethyltubocuraranium chloride hydrochloride, $C_{37}H_{42}Cl_2N_2O_6$, M_r 681.66, *mp* 274 – 275 °C. Prepared by monomethylation of tubocurine [12].

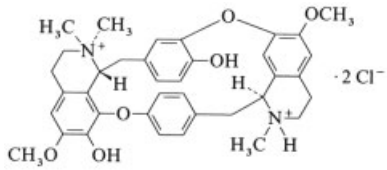

Tubocurarine chloride has a duration of action of ca. 30 min. It should be administered intravenously by slow injection (ca. 1.5 min). Its potency is increased with more rapid rates of administration. Its paralytic effects are intensified by diethyl ether, halothane, isoflurane, and enthrane. Potency is expressed as units: 3 mg tubocurarine chloride is equivalent to 20 units. Since it releases histamine and clinical doses block autonomic ganglia it has a liability for hypotension and tachycardia.

Trade Names. (*d*-Form) Delacuratine (Squibb), Intocostrin-T (Squibb), Jexin (Duncan Flockhart), Tubadil (Endo), Tubarine (Burroughs Wellcome), Tubocurarine Chloride (Abbott, Lilly, Squibb).

Vecuronium bromide [*50700-72-6*], 1-[3,17-bis(acetyloxy)-2-(1-piperidinyl)androstan-16-yl]-1-methylpiperidinium bromide, $C_{34}H_{57}BrN_2O_4$, M_r 637.75, *mp* 227 – 229 °C. Prepared by reaction of 2α,3α: 16α,17α - diepoxy - 17β - acetoxy - 5α - androstane with piperidine and subsequent alkylation with methyl bromide [10].

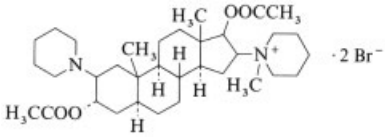

Vecuronium bromide is about one-third more potent than pancuronium. Complete recovery from vecuronium bromide takes ca. 45 – 65 min but this might be longer under halothane, enthrane, or isoflurane anesthesia. It has less effect on the circulatory system than pancuronium and produces little histamine release or hormonal action.

Trade Name. Norcuron (Organon).

1.2. Depolarizing Relaxants

Depolarizing agents produce a persistent depolarizing effect at the motor endplate that is similar to that of ACh (phase I block), followed by a secondary nondepolarizing block (phase II block). The initial depolarization results in transient muscular twitching, followed by flaccid paralysis due to the blockade of neuromuscular transmission. In contrast to the competitive neuromuscular blocking drugs, administration of acetylcholinesterase inhibitors does not reverse motor paralysis and may even intensify it if given shortly after the depolarizing block.

The depolarizing agents are not degraded by acetylcholinesterase leading to their accumulation at the motor endplate. When increasing doses of these drugs are continuously administered, the motor endplate gradually repolarizes due to receptor desensitization. This phase II nondepolarizing block may now be reversed by the administration of an acetylcholinesterase inhibitor. Reversal should not be attempted unless (1) a peripheral nerve stimulator is used to confirm that the block is of the nondepolarizing type and (2) enough time has elapsed to permit hydrolysis of the agent by plasma pseudocholinesterase. Administration of fluorinated hydrocarbons increases the incidence of conversion from phase I to phase II.

Decamethonium bromide [*541-22-0*], decamethylenebis(trimethylammonium) dibromide, $C_{16}H_{38}Br_2N_2$, M_r 418.36, *mp* 268 – 270 °C. Prepared by reaction of 1,10-dibromodecane with trimethylamine in a sealed tube [13].

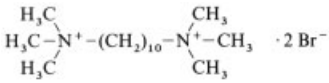

Decamethonium bromide is not appreciably metabolized and is mainly excreted by the kidneys. The acetylcholinesterase inhibitor, neostigmine, does not reverse its effect but the ganglion blocker, pentamethonium iodide, usually does. The mechanism of action of decamethonium changes after 20 – 30 min (phase II block), and neostigmine becomes an antagonist. Decamethonium bromide is no longer marketed in the United States.

Trade Name. Syncurine (Burroughs Wellcome).

Hexafluorenium bromide [*317-52-2*], hexamethylenebis(9-fluorenyldimethylammonium bromide), $C_{36}H_{42}Br_2N_2$, M_r 662.59, *mp* 188 – 189 °C. Prepared by reaction of 9-bromofluorene with dimethylamine and subsequent reaction with 1,6-dibromohexane [14].

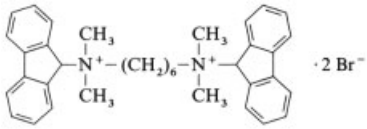

Hexafluorenium bromide is a plasma cholinesterase inhibitor that is used in combination with succinylcholine chloride (see below). It prevents the muscular twitches induced by succinylcholine chloride.

Trade Name. Mylaxen (Wallace).

Succinylcholine chloride [*71-27-2*], suxamethonium chloride, 2,2′-[(1,4-dioxo-1,4-butanediyl)bis(oxy)]bis[*N,N,N*-trimethylethanaminium] dichloride, $C_{14}H_{30}Cl_2N_2O_4$, M_r 361.30, *mp* 190 °C. Prepared by reaction of ethylene bromohydrin with succinic acid and subsequent reaction with trimethylamine in a sealed tube [15]. It can also be prepared in one step from succinic anhydride and choline chloride [16].

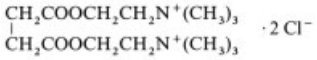

Succinylcholine chloride has a rapid onset and short duration of action (5 – 10 min). It is rapidly inactivated by plasma pseudocholinesterase to form succinylmonocholine which is ultimately reduced to succinic acid and choline. If succinylcholine is given over a long period, the neuromuscular block may change to a phase II block which may be antagonized by neostigmine.

Trade Names. Anectine chloride (Burroughs Wellcome), Lysthenon (Chemie Linz), Midarine (Burroughs Wellcome), Quelicin chloride (Abbott), Scoline (Duncan Flockhart), Succicuran (Deutsches Hydrierwerk), Succinyl-Asta (Asta-Degussa), Succostrin chloride (Squibb-Marsam).

1.3. Agents Acting Directly on Skeletal Muscle

Dantrolene and the cinchona alkaloids, including quinine, exert a direct effect on skeletal muscle excitability. They are used to reduce muscle tone rather than to induce paralysis.

Dantrolene sodium [*14663-23-1*], 1-[[5-(4-nitrophenyl)furfurylidene]amino] hydantoin, $C_{14}H_{10}N_4O_5Na$, M_r 314.56, *mp* 229 – 231 °C. Prepared by coupling 4-nitrophenyldiazonium chloride with furfural and subsequent reaction with 1-aminohydantoin [17].

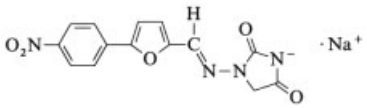

Dantrolene sodium reduces muscle tone by interfering with the release of ionic calcium from the sarcoplasmic reticulum [18]. Since this is more pronounced in fast muscle fibers, it has a preferential effect on reflex tone. The drug is administered orally for clinical spasticity but is of limited benefit in chronic spasticity because of side-effects, including drowsiness, dizziness, and weakness. It may be administered orally or intravenously to sensitive individuals to prevent or manage signs of malignant hyperthermia induced by neuromuscular blocking drugs or anesthetics during surgery. Dantrolene has a potential for hepatotoxicity.

Trade Names. Dantamicrin (Röhm Pharma), Dantrium (Norwich Eaton).

Quinine sulfate [*804-63-7*], 6′-methoxycinchonan-9-ol sulfate, $2C_{20}H_{24}N_2O_2 \cdot H_2SO_4$, M_r 746.93. It can be isolated from Cinchona bark, however, natural product synthesis has been reported by Woodward and others [19], [20].

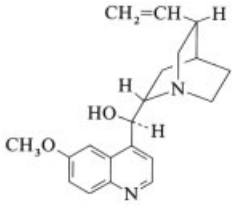

Quinine sulfate has a complex action on skeletal muscle that includes a curare-like effect on the motor endplate and a direct reduction of tonus by altering intracellular distribution of ionic calcium. It is used clinically in the prevention and treatment of nocturnal recumbency leg cramps and in the treatment of myotonia congenita. It is well absorbed after oral administration. Quinine should not be given to patients with myesthenia gravis or to patients who have received neuromuscular blocking drugs.

Trade Names. Legatrin (Scholl, Inc.), Quin-260 (Major), Quinamm (Merrell Dow), Quine (Reid-Rowell), Quinine sulfate (Lilly), Quinite (Reid-Rowell).

2. Centrally Acting Muscle Relaxants

In contrast to the peripherally acting neuromuscular blocking drugs, the muscle relaxants acting on the central nervous system (CNS) are prescribed to outpatients for acute muscle spasms, as well as for the treatment of chronic spasticity. *Acute muscular spasms* are induced by reflexes caused by local pathology (e.g., inflammation of the muscles and/or joints, local trauma) or

by sprains and strains. An acute spasm is usually localized to a particular muscle or muscle group, is self-limiting, and benefits from pharmacotherapy, bed rest, and adjunctive physical therapy.

Chronic neurological spasticity, however, refers to a specific disorder of muscle tone which accompanies upper motor lesions anywhere from the cerebral cortex to the spinal cord. This disorder results from neurological diseases including cerebral palsy, multiple sclerosis, spinal cord injury, and stroke. Spasticity consists of a state of muscle hypertonicity, i.e., an increase of normal muscle tone or resistance to stretching; an intact spinal stretch reflex arc (stretch receptors, afferent nerve, spinal cord centers, motor nerve; → Neuropharmacology, Chap. 3.2.1.) is essential. Excessive muscular tone or constant state of muscular contraction may be present even when the patient is at rest but worsens during voluntary motion. This is due to excessively active reflex contractions whenever the patient stretches a muscle. In severe cases, the reflex contraction may spread, or radiate up and/or down the spinal cord so that the movement of a single limb produces involuntary contraction of the muscles of other extremities, or can even involve the head and neck. *Cerebral spasticity* results from a loss of descending inhibition or from uncontrolled activity in bulbospinal pathways regulating the reflex arc. However, *spinal spasticity* results from the loss of supraspinal inhibitory control of flexor withdrawal reflexes.

2.1. Treatment of Acute Muscle Spasms

A number of drugs are used to relieve locally occurring skeletal muscle spasms. While some of these agents have spinal and/or supraspinal interneuronal blocking properties, many of them produce sedation at therapeutic doses. The mechanism responsible for their clinical effects, however, is unclear. With the exception of diazepam, these agents are not used for the treatment of chronic neurological spasticity. They are administered as part of a regimen of bed rest and physical rehabilitation for painful musculoskeletal conditions.

Carisoprodol [*78-44-4*], *N*-isopropyl-2-methyl-2-propyl-1,3-propanediol dicarbamate, $C_{12}H_{24}N_2O_4$, M_r 260.33, *mp* 92–93 °C. Prepared by reaction of 2-methyl-2-propyl-1,3-propanediol with phosgene and ammonium hydroxide, then with isopropyl isocyanate [21].

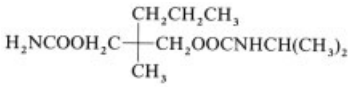

Carisoprodol has an onset of action of ca. 30 min and a duration of 4–6 h. It is administered orally. Presumably, it acts by inhibiting interneuronal activity in the spinal cord and the brainstem reticular formation. Clinically effective doses are accompanied by drowsiness or dizziness; its mechanism may involve sedation. Its CNS depressant effects are additive with those of ethanol and other psychotropic agents. Carisoprodol has a low potential for drug dependence.

Trade Names. Caprodat (Ferrosan), Carisoma (Wallace), Flexartal (Clin-Comar-Byla), Mioril (Rossini), Rela (Schering), Relasom (Rafa), Soma (Wallace), Sanoma (Heilit).

Chlorphenesin carbamate [*886-74-8*], 3-(4-chlorophenoxy)-1,2-propanediol-1-carbamate, $C_{10}H_{12}ClNO_4$, M_r 245.68, *mp* 89–91 °C. Prepared by reaction of 3-(4-chlorophenoxy)-1,2-propanediol with phosgene and subsequent reaction with ammonium hydroxide [22].

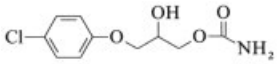

Chlorphenesin carbamate is well absorbed after oral administration and has a duration of ca. 4 h. Its use is similar to carisoprodol.

Trade Names. Maolate (Upjohn), Rinlaxer (Upjohn, Taischo).

Chlorzoxazone [*95-25-0*], 5-chloro-2-hydroxy-benzoxazole, $C_7H_4ClNO_2$, M_r 169.58, *mp* 191–191.5 °C. Prepared by hydrolyzing 2-amino-5-chlorobenzoxazole in 1 mol/L HCl [23].

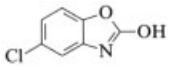

Chlorzoxazone has an onset of action of about 1 h and a duration of 3 – 4 h. It is administered orally. The metabolized drug is excreted in the urine as a glucuronide conjugate.

Trade Names. Biomioran (Bioindustria), Paraflex (McNeil), Solaxin (Eisai).

Cyclobenzaprine hydrochloride [*6202-23-9*], 5-(3-dimethylaminopropylidene)-dibenzo[*a,e*]cycloheptatriene hydrochloride, $C_{20}H_{22}ClN$, M_r 311.85, *mp* 216 – 218 °C. Prepared by Grignard addition of 5*H*-dibenzo[*a,d*]cycloheptanone with γ-dimethylaminopropyl chloride and subsequent elimination of water [24].

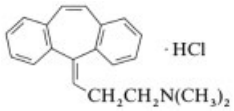

Cyclobenzaprine hydrochloride has an onset of action of about 1 h and a duration of 12 – 24 h. It is administered orally. It reduces muscular tension principally by inhibiting interneuronal activity in the brainstem reticular formation but also has activity at the spinal cord level [25]. It is chemically related to the tricyclic antidepressants and shares some of their pharmacological properties. Its CNS depressant effects are additive with those of ethanol and other psychotropic agents. It should not be administered in combination with monoamine oxidase inhibitors.

Trade Names. Flexeril (Merck, Sharp and Dohme), Yurelax (Organon).

Diazepam [*439-14-5*], 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2*H*-1,4-benzodiazepin-2-one, $C_{16}H_{13}ClN_2O$, M_r 284.76, *mp* 125 – 126 °C. Prepared by numerous synthetic routes including reduction and methylation of 7-chloro-5-phenyl-3*H*-1,4-benzodiazepin-2(1*H*)-one 4-oxide [26].

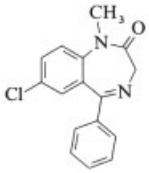

Diazepam is administered orally for the relief of acute muscle spasms and for the treatment of chronic spasticity in patients with spinal cord lesions and cerebral palsey. It acts at supraspinal levels in the reticular formation to enhance presynaptic inhibition of the spinal reflex arc [27]. Secondarily, it acts at the spinal level to augment interneuronal inhibition of the reflex. See also → Psychopharmacological Agents, Chap. 4.1.; → Sedatives, Chap. 3.1.

Trade Names. Cercine (Takeda); Horizon (Yamanouchi); Lorinon (Hoechst); ValCaps, Valium, Valiquid (Roche), Valrelease.

Meprobamate [*57-53-4*], 2-methyl-2-propyl-1,3-propanediol dicarbamate, $C_9H_{18}N_2O_4$, M_r 218.25, *mp* 105 – 106 °C. Prepared by reaction of 2-methyl-2-propyl-1,3-propanediol with excess phosgene and ammonium hydroxide [28].

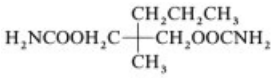

Meprobamate is rapidly absorbed after oral administration and has a duration of 3 – 4 h. Its primary clinical use is for daytime sedation in anxious patients, but it is occasionally administered orally as part of a regimen of bedrest and physical rehabilitation for painful musculoskeletal conditions. Mechanism of action and effects are similar to those of carisoprodol. See also → Psychopharmacological Agents, Chap. 4.2.

Trade Names. Aneural (Wyeth), Ayeramate (Ayerst), Equanil (Wyeth, Clin-Comar-Byla), Equatrate (Wyeth), Meprospan (Wallace), Meprotabs (Wallace), Miltown (Lederle, Wallace, Iribsa, Horner), Quaname (Wyeth), Quanil (Wyeth), Wyseals (Wyeth).

Metaxalone [*1665-48-1*], 5-(3,5-dimethylphenoxymethyl)-2-oxazolidinone, $C_{12}H_{15}NO_3$, M_r 221.25, *mp* 121.5 – 123 °C. Prepared by heating urea and 3-(3′,5′-dimethylphenoxy)-1,2-propanediol to 200 °C [29].

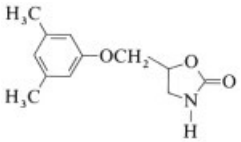

Metaxalone has an onset of action of ca. 1 h and a duration of 4 – 6 h. It is administered orally. Its clinical mechanism of action may involve sedative activity. Its CNS depressant effects are additive with those of ethanol and other psychotropic agents.

Trade Name. Skelaxin (Carnrick).

Methocarbamol [*532-03-6*], 3-(2-methoxyphenoxy)-1,2-propanediol 1-carbamate, $C_{11}H_{15}NO_5$, M_r 241.24, *mp* 92 – 94 °C. Prepared by reaction of 3-(2-methoxyphenoxy)-1,2-propanediol with phosgene and ammonium hydroxide [30].

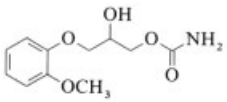

Methocarbamol has an onset of action of 30 min and a duration of ca. 4 h. It is administered orally for treatment of painful musculoskeletal conditions. A parenteral formulation is available for intravenous or intramuscular administration. It is used to control muscular spasticity resulting from tetanus. Its clinical mechanism of action is unclear but may involve its sedative activity. Its CNS depressant effects are additive with those of ethanol and other psychotropic agents.

Trade Names. Lumirelax, Miolaxene (Merrel Dow, Lepetit), Robaxin (Robins), Traumacut (Brenner), Tresortil (Gea).

Orphenadrine [*83-98-7*], *N,N*-dimethyl-2-(2-methyl-α-phenyloxy)ethylamine, $C_{18}H_{23}NO$, M_r 269.37; citrate salt, $C_{24}H_{31}NO_8$, M_r 461.51. Prepared by reacting 2-methylbenzhydryl chloride with dimethylaminoethanol [31].

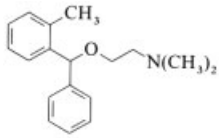

Orphenadrine citrate has a clinical duration of 4 – 6 h. It is administered orally or parenterally (intravenous or intramuscular). It reduces muscular tension principally by inhibiting interneuronal activity in the brainstem reticular formation, but its mechanism of action may be related to its analgesic activity. It has anticholinergic activity and should be used with caution in patients with cardiac disease.

Trade Names. (citrate salt) Banflex (Forest), Norflex (Rikker).

2.2. Treatment of Chronic Neurological Spasticity

The basic unit for the CNS integration of skeletal muscle tone is the reflex arc located in the spinal cord (→ Neuropharmacology, Chap. 3.2.1.). Specialized stretch receptors situated within muscle spindles (specialized muscle fibers) transmit information about muscle deformation to the spinal cord. The incoming signal activates spinal α-motor neurons innervating the stretched muscle to produce an increase in its tonus. Within the spinal cord, one (monosynaptic) or more synapses (polysynaptic) may be involved in the reflex path. Higher brain systems "fine tune" the stretch reflex by adjusting the sensitivity of the muscle spindle units through γ-motor neurons which adjust their "gain".

Multiple factors produce spasticity and it comes as no surprise that there is no satisfactory treatment. Drug therapy is limited in its efficacy by the interplay of the functional units of the nervous system: when one unit is favorably influenced by the drug, other units may be depressed or activated. Furthermore, nonselective reduction of muscle tone might not be in the

patient's best interest if abolition of spastic stiffness merely unmasks severe weakness. Reduction of muscle tone may only be indicated where the hyperexcitability interferes with function or hinders rehabilitation or nursing care. Sedation and other side effects which limit dosage regimens to suboptimal levels are another limitation of available therapy.

Baclofen [*1134-47-0*], γ-amino-β-(4-chlorophenyl)butyric acid, $C_{10}H_{12}ClNO_2$, M_r 213.67, *mp* 206 – 208 °C. Prepared by reduction of β- cyano-β-(4- chlorophenyl)propanoic acid [32].

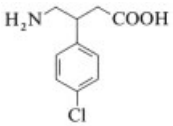

Baclofen has a rapid onset of action and a duration of ca. 4 h. It is used for relief of flexor spasms and concomitant pain, clonus, and muscular rigidity resulting from multiple sclerosis [33]. Patients with complete spinal transections respond as well as those with partial spinal lesions. It depresses mono- and polysynaptic transmission in the spinal cord. It has affinity as an agonist at a subtype of the γ-aminobutyric acid receptor ($GABA_B$). Drowsiness, insomnia, dizziness, weakness, ataxia, and mental confusion are adverse side effects that limit its clinical usefulness. Sudden withdrawal has been associated with hallucinations. Baclofen is excreted unchanged by the kidneys.

Trade Names. Baclon (Medica), Gabalon (Daiichi), Lioresal (Ciba-Geigy), Myospan (Lagap).

Chlorpromazine hydrochloride [*69-09-0*], 2- chloro-10-(3-dimethylaminopropyl)phenothiazine hydrochloride, $C_{17}H_{20}Cl_2N_2S$, M_r 355.33, *mp* 194 – 196 °C. Prepared by reaction of 2- chlorophenothiazine with dimethylaminopropyl chloride [34].

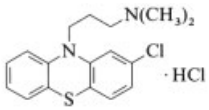

Chlorpromazine hydrochloride has antispastic activity [35] when administered parenterally (intramuscular or intravenous) and is indicated for the relief of muscle spasms associated with tetanus. Its efficacy in spasticity is overshadowed by its sedative and other CNS side effects. It inhibits descending noradrenergic brain stem pathways which regulate fusimotor discharge [36].

Trade Names. Largactil (Rhône-Poulenc), Promacil (Knoll), Promexin (Meiji), Thorazine (SKF).

Tizanidine [*51322-75-9*], 5- chloro-4-(2-imidazolin-2-ylamino)-2,1,3-benzothiadiazole, $C_9H_8ClN_5S$, M_r 253.70, *mp* 221 – 223 °C. Prepared by reaction of 4-(2-imidazolin-2-ylamino)-2,1,3-benzothiadiazol-4- cyanoamine with ethylene diamine and subsequent chlorination [37].

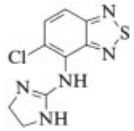

Tizanidine is marketed in 23 countries worldwide for the treatment of spasticity. It is being evaluated in multiple sclerosis patients in the United States by Athena Neurosciences.

Trade Names. Sirdalud, Ternelin (Sandoz-Wander).

3. References

General References

1. P. Taylor in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmocological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 166.
2. J. M. Cedarbaum, L. S. Schleifer in A. G. Gilman, T. W. Rall, A. S. Nies, P. Taylor (eds.): *The Pharmacological Basis of Therapeutics,* 8th ed., Pergamon Press, New York 1990, p. 463.
3. B. R. Olin et al. (eds.): *Drug Facts and Comparisons,* J. B. Lippincott, St. Louis 1991.
4. E. R. Barnhart: *Physicians' Desk Reference,* 45th ed., Medical Economics Company, Oradell 1991.

Specific References

5. Roche, US 3 089 373, 1963 (A. Boller, H. Els, A. Furst).
6. J. B. Stenlake et al., *Eur. J. Med. Chem. Chim. Ther.* **16** (1981) 515.
7. H. O. J. Collier, J. M. Z. Gladych, B. Macauley, E. P. Taylor, *Nature (London)* **182** (1958) 1424.
8. Allen & Hanburys, US 3 773 746, 1973 (D. Jack, E. E. Glover).
9. Rhône-Poulenc, US 2 544 076, 1961 (K. Fourneau).
10. W. R. Buckett, C. L. Hewett, D. S. Savage, *J. Med. Chem.* **16** (1973) 1116.
11. Z. Tuba, *Arzneim.-Forsch.* **30** (1980) no. 1, 342.
12. J. Naghaway, T. O. Soine, *J. Pharm. Sci.* **68** (1979) 655.
13. A. T. Blomquist, B. F. Hallam, A. D. Josey, *J. Am. Chem. Soc.* **81** (1959) 678.
14. C. J. Cavallito, A. P. Gray, E. E. Spinner, *J. Am. Chem. Soc.* **76** (1954) 1862.
15. J. Walker, *J. Chem. Soc.* 1950, 193.
16. C.-H. Wang, H.-H. Tso, *Org. Prep. Proced. Int.* **11** (1979) 93.
17. H. R. Snyder, Jr., C. S. Davis, R. K. Bickerton, R. P. Halliday, *J. Med. Chem.* **10** (1967) 807.
18. W. B. Van Winkle, *Science (Washington D.C.)* **193** (1976) 1130.
19. R. B. Woodward, W. E. Doering, *J. Am. Chem. Soc.* **66** (1944) 849.
20. G. Grethe, H. L. Lee, T. Mitt, M. R. Uskokovic, *J. Am. Chem. Soc.* **100** (1978) 589.
21. B. J. Ludwig, L. S. Powell, F. M. Berger, *J. Med. Chem.* **12** (1969) 462.
22. Upjohn, US 3 214 336, 1965 (H. E. Parker).
23. McNeil, US 2 895 877, 1956 (D. F. Marsh).
24. F. J. Villani, C. A. Ellis, C. Teichman, C. Bigos, *J. Med. Pharm. Chem.* **5** (1962) 373.
25. N. N. Share, *Pharmacologist* **18** (1976) 183.
26. L. H. Sternbach, E. Reeder, *J. Org. Chem.* **26** (1961) 4936.
27. A. C. Przybyla, S. C. Wang, *J. Pharmacol. Exp. Ther.* **163** (1968) 439.
28. B. J. Ludwig, E. C. Piech, *J. Am. Chem. Soc.* **73** (1951) 5779.
29. Robins, US 3 062 827, 1962 (C. D. Lunsford).
30. Robins, US 2 770 649, 1956 (R. S. Murphy).
31. A. F. Harms, W. T. Nauta, *J. Med. Pharm. Chem.* **2** (1960) 57.
32. CIBA, CH 449 046, 1968 (H. Keberle, J. W. Faigle, M. Wilhelm).
33. R. F. Jones, D. Burke, J. E. Marosszeky, J. D. Gillies, *J. Neurol. Neurosurg. Psychiatry* **33** (1970) 464.
34. Rhône-Poulenc, US 2 645 640, 1953 (P. Charpentier).
35. J. V. Basmajian, A. Szatmari, *Neurology* **5** (1955) 856.
36. P. H. Ellaway, J. E. Pascoe, *J. Physiol. (London)* **197** (1968) 8P.
37. Wander-Sandoz, US 3 843 668, 1974 (P. Neumann).

Skin Cosmetics

Shampoos and other hair care products are treated under → Hair Preparations

Günther Schneider, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chaps. 1, 2, 3, 6)

Sven Gohla, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chap. 3, Sections 4.2.2, 4.3 – 4.7)

Jörg Schreiber, Beiersdorf AG, Hamburg, Federal Republic of Germany (Sections 3.3, 4.2.2, 4.3, 4.4, 4.5, 4.6, 4.7)

Waltraud Kaden, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chap. 4)

Uwe Schönrock, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chaps. 5, 12)

Hartmut Schmidt-Lewerkühne, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chaps. 7, 8, 14, 15)

Annegret Kuschel, formerly Beiersdorf AG, Hamburg, Federal Republic of Germany (Chaps. 9, 10)

Xenia Petsitis, formerly Beiersdorf AG, Hamburg, Federal Republic of Germany (Chap. 10)

Wolfgang Pape, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chap. 13)

Hellmut Ippen, Dermatology Clinic, Universität Göttingen, Federal Republic of Germany (Chap. 11)

Walter Diembeck, Beiersdorf AG, Hamburg, Federal Republic of Germany (Chap. 13.2)

1. Introduction . . . 3
1.1. Cosmetics and the Skin . . . 3
1.2. The Function of Skin Cosmetics . 3
1.3. History . . . 3
2. Product Development . . . 3
2.1. Choice of Raw Materials . . . 3
2.2. Tests for Stability and Effectiveness . . . 4
3. Emulsions . . . 4
3.1. Introduction and Definition . . . 4
3.2. Composition and Classification . 5
3.2.1. Emulsifiers . . . 5
3.2.2. Lipophilic Constituents . . . 5
3.2.3. Hydrophilic Constituents . . . 6
3.2.4. Additives . . . 6
3.2.5. Special Ingredients . . . 6
3.3. PIT-Emulsions . . . 6
3.3.1. General Introduction and Definition 6
3.3.2. Phase Behavior of Oil–Surfactant–Water Systems . . . 7
3.3.3. Influence of Temperature on the Phase Behavior of an Emulsion . . . 7
3.3.4. Effects of Formulation Variables on Phase Inversion Temperature . . . 7
3.3.5. Nature of the Phase Inversion Area 7
3.3.6. Physicochemical Characterization of the Phase Inversion Temperature 7
3.3.6.1. Electrical Conductivity Measurement . . . 7
3.3.6.2. Interfacial Tension Measurement . . 7
3.3.7. Emulsification Production Process and Emulsification Routes . . . 8
3.3.8. CAPICO Method (Calculation of Phase Inversion in Concentrates) . . 8
3.3.9. Applications . . . 8
3.4. Stability and Product Safety . . . 8
3.5. Cosmetic Properties . . . 9
3.6. Production . . . 9
4. Gels . . . 9
4.1. Lipogels . . . 9
4.2. Hydrogels . . . 9
4.2.1. Polymer Gels . . . 9
4.2.2. Hydrodispersion Gels . . . 10
4.3. Lecithin Organogels . . . 10
4.4. Gel-Emulsions . . . 10
4.5. Surfactant Gels/Detergent Gels . 11
4.6. Microemulsion Gels . . . 11
4.7. Water-in-Silicon Oil Gels . . . 11
5. Oils . . . 11
5.1. Ingredients . . . 12
5.2. Production . . . 12
6. Lipsticks . . . 12
6.1. Lip Protection . . . 12
6.2. Composition . . . 12
6.3. Testing . . . 12
6.4. Production . . . 13
7. Shower and Bath Preparations . . 13
7.1. Composition . . . 13
7.1.1. Surfactants . . . 13
7.1.2. Auxiliaries . . . 14

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_219

7.2. Product Forms and Types 14
7.3. Production 15
8. Deodorants and Antiperspirants . 15
8.1. Composition 15
8.2. Application Forms 16
8.3. Testing 16
8.4. Production 16
9. Body Powders 16
10. Decorative Body Agents, Cosmetics, and Makeup 17
10.1. Types of Decorative Cosmetics . . 17
10.2. Colorants 18
10.3. Production 18
11. Sunscreens 18
11.1. Composition 19
11.2. Evaluation 19
11.3. Effectiveness and Production . . . 26
11.4. Toxicology and Environmental Protection 27
12. Stability Testing 27
12.1. Storage Conditions 28
12.2. Test Methods 28
13. In Vitro Compatibility Testing . . 29
13.1. Introduction 29
13.2. Test Methods 29
14. Economic Aspects 30
15. Legal Aspects 30
16. References 31

1. Introduction

1.1. Cosmetics and the Skin

Skin cosmetics are agents used in the care, cleansing, and decoration of human skin. The skin [1], with a surface area of ca. $2.0\,m^2$ in adults, constitutes the outer boundary of the human body with its environment. Skin is the largest human organ, and it fulfills various important functions. In addition to protecting against injury, environmental influences, dehydration, and desiccation, the skin serves to combat infection, regulate body temperature, and transmit sensory stimuli. The *epidermis* (outer skin) forms the actual protective covering. It consists of several layers: stratum basale, stratum spinosum, stratum granulosum, and stratum corneum (horny layer) as the outermost layer [2–7], which is an important permeability barrier.

Underlying the epidermis is a vascularized dermis that provides support and nutrition for the dividing cells in the epidermis (see Fig. 1). The dermis also contains nerves and appendages: sweat glands, hair follicles, and sebaceous glands. Nails are also considered as skin appendages. The subcutaneous fat is included as a third, deeper layer of the skin.

1.2. The Function of Skin Cosmetics

Skin cosmetics act on the epidermis or its outermost layers, but they are of great significance from the standpoint of function and appearance. Skin care measures should leave the skin's barrier function unaffected or even reinforced, and cleansing should inflict as little skin damage as possible. Decoration serves to visually enhance the skin color and structure. Skin cosmetics clearly differ from dermatological preparations, which induce effects below the epidermis in deeper layers on the skin, or even systemically (→ Dermatotherapeutic Agents). Dermatological preparations are in fact drugs used for treating skin diseases (dermatoses).

1.3. History

Preparations for the care, cleansing, and decoration of the skin have long been known [8]. For example, castor oil [*8001-79-4*] was used several thousand years ago in ancient Egypt as a protective balm. Among the Romans, Galen of Pergamon (A.D. 129 – 199) developed one of the first precursors of modern skin creams from a mixture of beeswax, olive oil, and rosewater (aqua rosae). Milestones in the preparation of ointment bases include the introduction of Vaseline [*8001-25-0*] by CHESEBOROUGH, as well as purified lanolin [*8027-32-5*] by LIEBREICH in the second half of the nineteenth century. The first stable water-in-oil skin cream was produced in 1911 (Nivea Creme).

2. Product Development

2.1. Choice of Raw Materials

A great many raw materials can be used in the formulation of cosmetics [9–11]. Choosing the

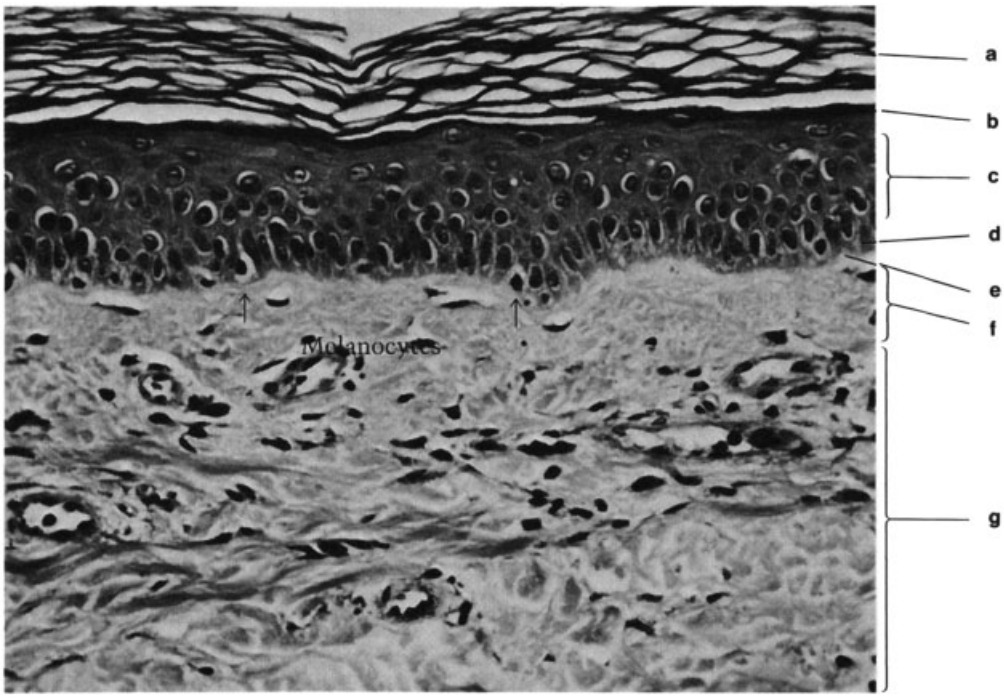

Figure 1. Subdivisions of the epidermis and dermis (× 256) (by courtesy of Little, Brown and Company, Boston, Mass.) a) Stratum corneum; b) Stratum granulosum; c) Stratum spinosum; d) Stratus basalis; e) Dermoepidermal junction; f) Papillary dermis; g) Reticular dermis

raw materials is an important first step, and it involves a balanced and thoughtful consideration of the following diverse factors:

Aesthetics
Availability
Color
Cost
Environmental compatibility
Functionality
Government regulations
Mutual compatibility with respect to other ingredients
Patent status
Processing characteristics
Skin and mucous membrane compatibility
Skin feel
Smell
Stability (chemical, physical, microbial)

Cosmetic raw materials are seldom pure substances; most are instead mixtures of a wide variety of individual components.

2.2. Tests for Stability and Effectiveness

Stability tests for skin cosmetics are of three types: physical, chemical, and microbial (see Section 12.2). Bacteria and fungi might conceivably contaminate an end product during either production or filling, but also later, during use, so microbiological evaluation must be conducted with extreme care [12], [13]. Of fundamental importance are the cosmetic properties of the product itself – that is, its aesthetic characteristics – as well as the state of the skin during and after product use. Evaluation is conducted within the framework of organoleptic and visual (i.e., sensory subjective) appraisal [14]. Objectively detectable changes in skin state are monitored by biophysical measurements. Many different methods are available [15], [16], including determination of the skin surface profile [17] and measurement of transepidermal water loss (TEWL) [18], [19].

3. Emulsions (→ Emulsions)

3.1. Introduction and Definition

Emulsions play a central role in cosmetics chemistry. Advances in cosmetic technology have opened the way to preparing otherwise impracticable combinations of constituents. Variation of rheological – tactile properties makes available a broad range of application forms (lotions, milks, creams, etc.) [20].

An emulsion is generally described as a thermodynamically unstable system consisting of at least two immiscible liquids, one of which is dispersed in fine droplet form in the other [21–23]. Unless an emulsion is stabilized, the two immiscible liquids tend to separate. Stabilization is achieved by incorporating emulsifiers into the interface, and through the use of additional polymers, solids, and liquid-crystalline phases [24]. The aggregation states of both the dispersed and continuous phases vary from highly liquid to semisolid [25]. Cosmetic emulsion systems thus cover the range from lotions (low viscosity) to semisolid preparations (ointments and creams) [26].

3.2. Composition and Classification

Cosmetic emulsions can be of three basic types: water-in-oil (W/O), oil-in-water (O/W), and bicoherent (e.g., multiple emulsions). Emulsions of the O/W and W/O type predominate in human skin care applications [27]. Different types of products (creams, lotions, milks, pastes, etc.) are preferred depending on the application site (e.g., body, face), and the intended use (care, cleansing, or decoration). Cosmetic preparations normally contain 5 – 50 % fat-like constituents (oils, fats, waxes), 1 – 10 % emulsifiers, and 50 – 90 % hydrophilic constituents (water and water-soluble ingredients) [28].

3.2.1. Emulsifiers

Emulsifier systems are constructed from appropriate lipophilic and hydrophilic molecular subunits. Proper selection of the components is critical to the stability of an emulsion. The choice is complicated by the fact that special characteristics of all the various substances in the system must be taken into account. The *hydrophilic – lipophilic balance* (HLB) is useful in selecting emulsifiers for the development of skin cosmetics [29–31] (→ Emulsions, Chap. 2.8., → Emulsions). Emulsifiers can be subdivided on the basis of their hydrophilic components into ionic (anionic, cationic, amphoteric) and nonionic materials. Common examples include:

1) Mixtures of fatty acids and their salts (e.g., stearic acid [*57-11-4*] and sodium stearate [*822-16-2*])
2) Ethoxylated fatty alcohols (e.g., polyoxyethylene (20) cetyl – stearyl ethers [*68439-49-6*])
3) Hydrogenated ethoxylated castor oils (e.g., polyoxyethylene (7) hydrogenated castor oil [*61788-85-0*])
4) Partial esters of fatty acids and modified fatty acids with glycerol, polyglycerol, and sorbitol (e.g., glyceryl monostearate [*123-94-4*], sorbitan monooleate [*1338-43-8*], and their ethoxylates)
5) Orthophosphoric acid esters (e.g., cetyl phosphate [*3539-43-3*])
6) Silicone surfactants (e.g., dimethicone copolyol [*64365-23-7*])

Detailed literature is available on the chemical structure, mode of action, and choice of emulsifiers for cosmetic products [32], [33].

3.2.2. Lipophilic Constituents

The lipophilic phase is normally a mixture of oils, fats, and waxes. It constitutes the "heart" of a cosmetic emulsion, since such product-defining properties as skin care and skin feel are strongly influenced by the composition of this phase. For example, it is responsible for "spreadability," (i.e., the tendency of an oil component to spread over the skin) and the "oily" feel of a cosmetic emulsion. Night creams are generally formulated to contain more fat and less spreading oil than day creams, which often contain easily spreading oils and more water. A wide range of oils, fats, and waxes is available commercially, corresponding to the central importance of these substances. These are divisible generally into the following groups:

1) Solid, semisolid, and liquid hydrocarbons (e.g., mineral oil [*8012-95-1*], petrolatum [*8009-03-8 (NF)*])
2) Natural oils and waxes (e.g., wheat-germ oil [*8006-95-9*], beeswax [*8012-89-3*])
3) Guerbet alcohols (e.g., 2-octyldodecanol [*5333-42-6*])
4) Partially hydrogenated and modified triglycerides (e.g., partially hydrogenated coconut oil, octanoic – decanoic acid triglyceride)
5) Fatty acid esters (e.g., cetyl palmitate [*540-10-3*])
6) Isostearic acid derivatives (e.g., isostearyl isostearate [*41669-30-1*])
7) Silicones (e.g., dimethicone [*9006-65-9*], [*63148-62-9*], [*9016-00-6*])

3.2.3. Hydrophilic Constituents

The aqueous phase of cosmetic emulsions generally contains desalinated water, which must be of high quality microbially. Other constituents normally include water-binding substances (humectants and moisturizers), such as the polyols, glycerol [*56-81-5*] and sorbitol [*50-70-4*], ethoxylated polyols, and hydrolyzed proteins. So-called natural moisturizing factor (NMF) – a mixture of amino acids, salts of pyrrolidonecarboxylic acid and lactic acid, urea, and other substances – is also used [34].

3.2.4. Additives

Additives important for product stabilization and the performance of cosmetic emulsions include:

1) Antioxidants to stabilize oxidation-sensitive oils, fats, and waxes
2) Preservatives (e.g., methylparaben [*99-76-3*], phenoxyethanol [*122-99-6*]) to protect against microbial infection of the product (consumer protection)
3) Chelate binders (e.g., ethylenediaminetetraacetic acid) to trap metal traces
4) Co-emulsifiers (e.g., cetyl alcohol [*36653-82-4*] and aluminum stearate [*7047-84-9*])
5) Acids and bases to regulate pH
6) Thickeners (e.g., one of the carbomers [*9007-16-3*], [*9003-01-4*], [*9007-17-4*], [*76050-42-5*]; cellulose derivatives)
7) Dyes, colored pigments, and white pigments, principally in decorative cosmetics
8) Light filters (sunscreens; see Chap. 11)
9) Perfume oils

3.2.5. Special Ingredients

Special ingredients, including so-called natural substances, are increasingly finding a place in face-care products [35]. The group of natural substances includes plant extracts (e.g., *Aloe vera* gel), natural biopolymers such as chitin [*1398-61-4*], collagen [*9007-34-5*], and hyaluronic acid [*9004-61-9*], and vegetable oils (e.g., evening primrose oil). Because of their sometimes problematic skin compatibility and tendency to sensitization, plant extracts must be tested thoroughly before being introduced into cosmetic emulsions.

The list of special ingredients also covers topically applied *vitamins*, particularly of provitamin B_5 (D-panthenol), and vitamin E and their derivatives, alone [36], or mixed with vitamin C and its derivatives.

More recently, special *carrier systems* (e.g., liposomes) have become important, especially in face care products. Liposomes are hollow spheres bounded by one (unilamellar) or several (multilamellar) lipid double membranes and containing one or more aqueous phases. Fat-soluble substances can be incorporated into the double membranes, and water-soluble substances can be enclosed within the spheres [37]. Phosphatidylcholine (PC) is thought to be the active agent in liposome sheath material from the standpoint of skincare. Liposomes can be regarded as combining the functions of both carriers and skin-related special ingredients. Microcapsules filled with active constituents have also been used as protection, sustained release, and transport vesicles [38].

3.3. PIT-Emulsions

3.3.1. General Introduction and Definition

Stabilization of an emulsion can be achieved by the addition of polymers, solids, or compounds that form liquid crystalline phases (see

Section 3.1). These ingredients increase the viscosity of the emulsion, thereby improving the stability of the final product.

A new low energy emulsification method called phase inversion temperature (PIT) method allows the formulation of finely dispersed low viscosity O/W emulsions without addition of the above-mentioned stabilizers. Emulsions can be thermally inverted by heating an oil-in-water (O/W) to a water-in-oil (W/O) emulsion (→ Emulsions, Chap. 1.). The inversion range is designated as phase inversion temperature [39].

3.3.2. Phase Behavior of Oil–Surfactant–Water Systems

The phase behavior of a ternary system consisting of oil, surfactant, and water can be represented in a triangular phase diagram. To study the influence of temperature on this ternary system a so-called "phase prism" is required, where the base specifies the composition and the vertical edge is the temperature axis [40], [41]. A section through the surfactant corner at a fixed water – oil ratio gives a temperature – surfactant concentration diagram that explains the different phases with respect to the temperature (so-called Kahlweit fish) [39].

3.3.3. Influence of Temperature on the Phase Behavior of an Emulsion

Heating a suitable mixture of an oil, an ethoxylated O/W surfactant (additionally a lipophilic cosurfactant) and water to 60 – 85 °C transfers an O/W emulsion to a W/O emulsion. During the heating process (O/W → W/O) and after the cooling process (W/O → O/W) the mixture shows a phase inversion. The resulting O/W emulsion (PIT emulsion) has a different droplet size distribution curve and smaller oil droplets than the O/W emulsion before the phase inversion process. PIT emulsions often have a bluish-white appearance.

Increasing the temperature of an O/W emulsion decreases the affinity of the hydrophilic ethylene oxide groups of the O/W surfactant to the water molecules. Heating of a PIT emulsion makes the hydrophilic O/W surfactant more hydrophobic, thereby increasing the solubility of the emulsifier in the oil phase and thus facilitating the formation of a W/O emulsion at higher temperatures.

3.3.4. Effects of Formulation Variables on Phase Inversion Temperature

The temperature at which phase inversion occurs depends on a number of variables [42] such as:

- Nature of the ethoxylated O/W surfactant (e.g., ethoxylated fatty alcohols, fatty acids, sorbitan esters, glycerol esters, or castor oils)
- Concentration of the ethoxylated O/W surfactant
- Nature and concentration of the W/O surfactant (glycerol esters, sorbitan esters, cetyl alcohols)
- Type and polarity of the oil phase

3.3.5. Nature of the Phase Inversion Area

The colloidal structure of the emulsion in the phase inversion range depends on the concentration of the surfactant (or the concentration of the water phase) in the composition. Several phases at 60 – 85 °C have been detected such as W/O microemulsions, lamellar liquid crystals, or bicontinuous microemulsions [43], [44]. In the phase inversion area the hydrophilic and hydrophobic properties of the O/W surfactant are balanced and a minimum interfacial tension between the oil phase and the water phase can be observed.

3.3.6. Physicochemical Characterization of the Phase Inversion Temperature

The phase inversion in emulsions is often accompanied by sharp changes in the physical properties, since different colloidal phases are involved in the process.

3.3.6.1. Electrical Conductivity Measurement

Cooling down a W/O emulsion to an O/W emulsion via a PIT process can be followed

by measuring the conductivity, since the external oil phase of the hot W/O emulsion does not contribute to the conductivity [45]. The increase in conductivity during cooling in a conductivity – temperature diagram shows the beginning of the phase inversion process of the composition.

3.3.6.2. Interfacial Tension Measurement

The interfacial tension is dependent on the temperature, measurement with a spinning drop tensiometer usually yield very low tension values (10^{-3} mN/m) in the PIT range [46]. At a constant concentration of the oil phase, the minimum level of the interfacial tension depends on the relative proportion of the O/W to W/O surfactant, the total amount of both surfactants, and the ethoxylation degree of the O/W surfactant.

3.3.7. Emulsification Production Process and Emulsification Routes

The knowledge of the phase behavior of an emulsion composed of an oil, an O/W surfactant (additionally a cosurfactant), and water allows the formulator to develop different strategies to prepare a PIT emulsion.

Hot – Cold Process. A preconcentrate is formulated where only part of the water phase is heated with the other components to the PIT range or above the PIT [43], [47]. Dilution of the hot mixture with cold water gives the final composition. This route is advantageous when temperature sensitive actives or perfumes are contained in the final product since in this case the cooling process of the hot W/O emulsion to the final PIT emulsion is faster than in the hot – hot process.

Hot – Hot Process. In the one-step process, all components are mixed together and heated above the phase inversion temperature [43], [47]. Cooling the composition to room temperature gives the finely dispersed bluish white product.

The droplet size is usually in the range of 100 to 250 nm in both processes. Preparing an emulsion suitable for phase inversion without heating the composition above the PIT usually gives a coarse emulsion with a droplet size distribution around 1000 nm [43].

3.3.8. CAPICO Method (Calculation of Phase Inversion in Concentrates)

Hydrocarbons as an oil phase in PIT emulsions show a linear dependence of the phase inversion temperature on the alkyl chain length (alkane carbon number, ACN) [39], [48], [49].

The interaction of the emollient and the O/W surfactant can be described by the hydrophilicity factor (HF) of a surfactant. Determination of the HF factor of an O/W surfactant is possible by measuring the temperature of phase inversion of oils with different alkyl chain lengths (different ACN values). The intersection of the graph with the ordinate in a PIT-ACN diagram gives the PIT_o value [39].

$$\mathrm{PIT} = \mathrm{PIT}_0 + (\mathrm{SF} \times \mathrm{ACN})$$

where SF is the sensivity factor (slope of the graph).

The advantage of the CAPICO method lies in the fact that it is possible to determine the expected temperature of phase inversion or the optimal surfactant : cosurfactant ratio for a desired PIT before preparing the product since many of the ACN, SF, and PIT_o values are published [39].

3.3.9. Applications [50]

Deodorants or antiperspirant pump sprays are usually water – alcohol solutions with a deodorant or antiperspirant active. The alcohol content is not suitable for consumers with sensitive skin, especially if the axillae are shaved.

Low viscous and sprayable deodorant PIT emulsions for the European market have been developed by several companies in the last few years. These products give a nice skin feeling and are free of alcohol.

Suncare Products. The PIT technology has been used to formulate sprayable suncare products. The homogenous distribution of the emulsion droplets facilitates a good distribution of the UV filters on the body. Furthermore, the consumer's hands do not come into direct contact with the product, making application more easy.

3.4. Stability and Product Safety

Emulsions are thermodynamically unstable structures. Depending on relevant legal statutes market-ready cosmetic emulsions are expected to be stable for several years at various temperatures; comprehensive testing is necessary to satisfy these requirements (see Chap. 12). Furthermore, cosmetic emulsions may not induce toxic responses when applied in accordance with instructions. Therefore, new developments have to be thoroughly tested in terms of their compatibility and dermatology tests, before the product is introduced to the market (see Chap. 13).

3.5. Cosmetic Properties

In addition to the aforementioned properties, public acceptance is an essential criterion for a cosmetic emulsion, particularly with respect to products for facial use. Consumer acceptance is generally based on two factors:

1) Objective effectiveness of the product. This is established by skin evaluations based on biophysical criteria, e.g., degree of moisture and roughness of skin (see Section 2.2).
2) Subjective effectiveness of the product (cosmetic elegance). The cosmetic elegance and tactile characteristics of a cosmetic emulsion can be determined by suitable test panels [51], [52].

3.6. Production

Generally speaking, the production of cosmetic emulsions involves three steps:

1) Preparation of the various phases (fat, water, thickeners, etc.) and then combining them in an appropriate sequence
2) Emulsification (i.e., comminution of the components by stirring)
3) Homogenization, achieved by further disintegration of the fat and water droplets in a homogenizer

A suitable formulation is first developed on a 1 – 2 kg scale (bench scale) using raw materials of the types described above. The procedure is then adapted to a 10 – 100 kg (pilot-plant) scale for further study prior to moving to full production scale (500 kg to several tons). Descriptions of appropriate scaleup and processing procedures can be found in [53].

4. Gels

Gels are enjoying increasing popularity. They correspond well to the prevailing "ethos" of modern cosmetology in that they combine functionality and elegance with an attractive image. The variety of available gel types permits a gel formulation to be adapted to suit nearly every cosmetic product group.

Gels are systems consisting of at least two components, generally a solid dispersed colloidally in a liquid dispersing agent [52] (→ Colloids). The dispersed particles form spatial networks in the dispersing agent that are stabilized by means of van der Waals forces, whereby the dispersing agent acts as a solvate sheath [20]. Gels can be of various types depending on their chemical composition. *Hydrogels* contain water as the dispersing agent, whereas in *lipogels* this role is played by oils.

4.1. Lipogels

Lipogels are preferred whenever skin-protecting effects are desired. Lipogels adhere to skin for a long time, and they provide a water-repelling effect. For this reason they are particularly suitable for hand care products and in sunscreen preparations for water sport enthusiasts [20], [54].

Depending on the intended use, lipogels are prepared from semisolid fats and waxes mixed with vegetable oils, longer- chain paraffins together with volatile silicones (e.g., cyclomethicone), or colloidal swelling agents (Aerosil, bentonite) in conjunction with mineral or vegetable oils. The firmness of a gel is determined by the content of microcrystalline components. Transparent gels result when the oil component has roughly the same refractive index as the incorporated solid (e.g., mixtures of esters of higher fatty acids or paraffin oil with Aerosil).

4.2. Hydrogels

Hydrogels can in turn be subdivided into polymer gels and hydrodispersion gels.

4.2.1. Polymer Gels

Polymer gels are widely used in skin- care cosmetics, particularly for fat- and emulsifier-free formulations. They readily release moisture, at the same time producing a marked cooling effect on the skin due to latent heat of vaporization. Several types of polymers have been employed:

Inorganic gel-forming agents, principally the colloidal silicic acid known as Aerosil and the swelling clay bentonite [55–58].

Natural organic gel-forming agents, including disaccharides and polysaccharides, as well as proteins of vegetable and animal origin (e.g., agar-agar or gelatin).

Semisynthetic organic gel-forming agents, such as biosynthetic rubbers (xanthanes, dextrans), starch fractions and starch derivatives (pectins, dextrins), and cellulose derivatives (cellulose ethers of widely varying types) [55], [59].

Synthetic organic gel-forming agents, the most important of which is polyacrylic acid (Carbopol), which forms a gel only after neutralization with an inorganic or organic base [55], [60].

A particular gel-forming agent is selected on the basis of its structure-forming properties, its compatibility with other raw materials and active constituents, and its behavior and function with respect to skin. It is important to note in this context that cellulose ether gels dry on the surface of the skin to form thin films, whereas polyacrylate gels are absorbed [61], [62].

Surface-active polyacrylates (Pemulen types) are increasingly being used to produce so- called quick-breaking emulsions. These are lipid- containing, creamlike gels that are highly sensitive to electrolytes; even the electrolyte content of normal skin is sufficient to destabilize them immediately (i.e., to break up the emulsion). The result is a full-bodied gel that is, nevertheless, rapidly absorbed. Detailed and comprehensive information is available from various distributors on the handling of gel-forming agents (suitable concentrations, dissolution behavior, nature of the dispersion, etc.). The testing process for gels from the development stage to marketing is similar to that of other cosmetic products (see Chaps. 12 and 13). Hydrogels are susceptible to microbial infection, and they are extremely light sensitive, especially polyacrylate gels, which means that special precautions must be taken during formulation and packaging development.

4.2.2. Hydrodispersion Gels

Hydrodispersion gels contain an additional oil phase. These emulsifier-free systems with oils and actives are sometimes used for sun-care products that are free of adverse reactions for consumers during a sun bath.

4.3. Lecithin Organogels

Lecithin organogels are made up of a three-dimensional network of entangled flexible tubular reverse micelles [77]. Lecithin organogels are prepared by addition of small amounts of water into an oil – lecithin mixture. At first spherical reverse micelles are formed. By further addition of water the inverted micelles undergo a one-dimensional growth to tubular aggregates ("wormlike micelles", polymer-like micelles) [78]. The result is the formation of a gel of a low-viscous, oily composition. The water channels and the oil phase of such systems are used to encapsulate different actives [79].

Lecithin organogels are also referred to as microemulsion gels (see Section 4.6). However, this classification is not justified (see page 11).

4.4. Gel-Emulsions

Product with a high amount of the internal phase are also known as high internal phase ratio emulsions (HIPRE) [80]. Due to their high water content, viscosity, and translucence they are also referred to as gel-emulsions. The viscosity is achieved by a large internal phase volume that exceeds 0.74, which corresponds to that of the closed-packed-sphere packing [81]. These gel-emulsions are formed in ternary mixtures of water, nonionic surfactant, and oil. They can contain up to 99.5 % of water with a very low content of surfactant (0.1 %) [82]. Some systems are

described as water-in-(water-in-oil microemulsions) emulsions. The structure of these emulsions is similar to that of a foam in which the gas is substituted by water and the continuous phase is a microemulsion of the water-in-oil type.

4.5. Surfactant Gels/Detergent Gels

The incorperation of a larger amount of a surfactant (> 10 %) into water or an oil phase transfers the low viscous micellar or reverse micellar solution into other colloidal phases. The phases that can occur upon increasing the surfactant concentration are classified as hexagonal, lamellar, inverted hexagonal, or reversed micelles [83], [84]. This phase sequences is not observed with every surfactant. Sometimes only a few phases occur upon increasing the surfactant concentration. A cubic phase can additionally be present in a system between micellar – hexagonal, hexagonal – lamellar, lamellar – inverted hexagonal or inverted hexagonal, and inverted micellar phases. The viscosity of these systems is often quite different from a free flooding gel to a stiff gel. Some of the former show the ringing gel effect when apping a glas vial containing such a surfactant – water mixture. Surfactant gels were classified as microemulsion gels in the past.

4.6. Microemulsion Gels

Definition of Microemulsions. A microemulsion is a thermodynamic stable, clear dispersion of two immiscible liquids (water and oil) containing an appropriate amount of surfactants and cosurfactants [85]. The viscosity of an oil-in-water and a water-in-oil microemulsion is usually very low. Both systems contain globular structures (oil nanodroplets or water nanodroplets). Bicontinuous microemulsions with an equal amount of water and oil are often more viscous (gel-like appearance) [86]. In this case the structure is not globular, the phase consists of relatively large interconnected and highly disordered oil and water domains, separated by the film of the surfactant monolayer. The analysis of the phase behavior of various surfactant – oil – water compositions leads to a better understanding of the colloidal phases of such compositions. Some of the systems described as microemulsions or microemulsion gels in the past are nowadays classified differently, e.g., lecithin organogels (Section 4.3) or surfactant gels (Section 4.5).

W/O-microemulsion gels are obtained by mixing water. AOT (an anionic surfactant) and an oil phase with gelatin as a polymer [87]. The gelatin is dissolved in water nanodroplets (nanogel droplets). At a specific polymer concentration the gelation of such a system occurs by a crosslinking process of the nanodroplets via gelatin helix – helix interactions in the oil phase. Another approach to these systems are hydrophilically modifies oil-soluble polymers. The formation of a polymer-mediated three-dimensional network of nanodroplets explain the gel character of such formulations.

O/W-microemulsion gels are obtained by adding a hydrophobically modified water-soluble polymer (cetylhydroxyethyl cellulose, PEG-150 distearate) to a low viscous microemulsion. The concentration of surfactants is very low (< 10 %).

4.7. Water-in-Silicon Oil Gels

By matching the refractive index of the water and oil phase it is possible to achieve transparent products [90], [91]. They are not classified as W/O microemulsion gels because they are not thermodynamically stable. In the US, deodorant gels are often based on this technology.

Silicone formulation aids (Dow Corrning 3225 C) are mixtures of cyclomethicone and dimethicone copolyols helping to solubilize the antiperspirant active. Low surface tension of cyclomethicones faciliates good spreading of a product on the skin and reduces the tackiness of antiperspirant active.

5. Oils

Cosmetic oils have been developed to satisfy a variety of needs, including:

Skin cleansing	makeup removers, baby oils
Skin care	moisturizers
Skin protection	UV protection (through the use of oil-soluble UV filters), protection against frostbite, protection against excessive dehydration from swimming
Special purposes	massage oils, bath additives, hair- care oils

Cosmetic oils are usually single-phase systems that are liquid at ambient temperature [65], [66].

5.1. Ingredients

The basic components of cosmetic oils are vegetable oils, synthetic fatty acid esters, hydrocarbons, silicone oils, mineral oil, and triglycerides of various chain lengths. Depending on the intended use, active ingredients may be incorporated, such as plant extracts, oil-soluble vitamins, and UV filters. Where necessary, antioxidants or UV filters are added to impart stability. In contrast to typical single-phase oils, hydrophilic oils contain a specific amount of an oil-in-water emulsifier to induce formation of an emulsion on contact with water. Typical emulsifiers include fatty alcohol polyglycol ethers, sorbitol fatty acid polyglycol ethers, and phosphoric acid esters.

5.2. Production

The production of cosmetic oils is quite simple when the components are anhydrous. Turbidity may develop with water- containing components, and for this reason, such ingredients (especially mineral oil) must be dehydrated by stirring under vacuum at 75 – 80 °C. After dehydration, the ingredients are combined with constant stirring. Temperature-sensitive ingredients must be added at $\leq$ 60 – 65 °C.

6. Lipsticks

6.1. Lip Protection

The skin of the lips [67] has a tendency to dry and develop small cracks, especially when exposed to cold, dry air [68], which then makes this particular skin extremely sensitive to other external influences. Lips must therefore be accorded special care and protection.

Apart from this consideration, the purely decorative aspect of lipstick or of lips of course plays a significant role [69]. Because of their convenience and functionality lipsticks occupy the dominant market position in the care and decoration of the lips.

The ideal lipstick should be easy to apply, and it should have a neutral taste. Light pressure on the stick should produce a film of fat that adheres well to the lips, leaving them smooth and supple, but the film should not be greasy, sticky, or subject to smudging. In the case of decorative lipsticks, the goal is an indelible coloring of the lips coupled with firm adherence of the coloring matter. Unfortunately, the requirements for stability conflict with the ideal application properties. A lipstick must also be mechanically stable and temperature resistant, and it must not "sweat" (i.e., there should be no visible separation of small oil droplets) [70]. With decorative lipsticks it is necessary to pay special attention to the stability of the coloring agents.

6.2. Composition

Lipsticks are normally anhydrous mixtures of fat, oil, and wax [71–73]. Water- containing formulations are an exception [74]. Uncolored lipstick base compositions can be subdivided roughly into the following groups of raw materials:

1) Liquid oils (e.g., mineral oil, castor oil)
2) Semisolid constituents (e.g., petrolatum, lanolin)
3) Solid constituents (e.g., ceresin, beeswax)
4) High-melting waxes [e.g., carnauba wax (from *Copernicia prunifera*)]

Microcrystalline paraffin waxes (e.g., ozokerite and ceresin), because of their oil-binding properties and relatively low thermoplasticity, are important constituents of high-quality lipstick bases (→ Waxes). In combination with high-melting waxes they constitute the framework of traditional lipsticks. Carnauba wax, one of the hardest and highest-melting (*mp* 83 – 86 °C) natural waxes, is particularly important in this respect, since it substantially increases the strength of a lipstick.

6.3. Testing

Lipsticks are evaluated on the basis of utility, appearance, color, ability to produce uniform col-

oration, adhesion, fragrance, and taste. Other criteria include the results of storage tests at various temperatures, heat stability, weight loss, fracture stability, and cohesion of the lipstick composition in its holder.

6.4. Production

Once an elegant and stable lipstick formulation has been devised it is crucial that a simple and unproblematic production process be developed. Normally the fat component is melted and cast directly. Important parameters include, among others, the casting and cooling temperatures [75].

7. Shower and Bath Preparations

(→ Soaps)

Cleansing of the skin is certainly one of the oldest human customs. Along with hygiene, religious and cultural factors have often played an important role in cleansing activities. Consumers expect more from modern bath and shower products than mere cleansing, including refreshment, stress relief, stimulation, and grooming.

7.1. Composition

7.1.1. Surfactants

The most important constituents of cleansing products are surfactants, with the task of reducing the surface tension of water so that thorough wetting can occur and dirt can be solubilized and removed from the skin (→ Surfactants; → Laundry Detergents, Chap. 3.1.). A choice of surfactants is made on the basis of consumer expectations; that is, the product must not only cleanse effectively, it must also foam well even when charged with dirt, sebum, and grease. These requirements are met largely by anionic and amphoteric surfactants [76]. Nonionic surfactants have so far been used to only a minor extent in special products because of their unsatisfactory foaming power, although increasing attention has recently been paid to this group of substances.

Anionic Surfactants. The most important anionic surfactants include alkyl sulfates and alkyl ether sulfates; alkyl and olefin sulfonates, as well as sulfosuccinates, and sulfoacetates; alkylethercarboxylic acids; and fatty acid – protein condensates.

Alkyl sulfates develop a good foam, are resistant to hard water, and display powerful cleansing and degreasing action. Unfortunately, they also exhibit a high protein-binding capacity and poor mucous membrane compatibility, and for these reasons, they are of limited suitability for body- cleansing purposes, and must be mixed with other, milder surfactants.

Alkylether sulfates are widely used in Europe in the production of shower and bath preparations. In particular, compounds with a $C_{12}-C_{14}$ alkyl chain and an average degree of ethoxylation of 2 – 3 mol of ethylene oxide are extremely water soluble, foam strongly even in hard water, and can be thickened with electrolytes (e.g., sodium chloride).

Sulfosuccinate and sulfoacetate esters produce a creamy foam and have outstanding skin and mucous membrane compatibility. Their disadvantage, however, is that they tend to hydrolyze. They are therefore used preferentially in neutral to weakly acidic products.

Protein – fatty acid condensates and alkylethercarboxylic acids are, because of their excellent skin compatibility, used particularly for especially mild preparations.

Amphoteric surfactants [77] can be divided into two main categories: the alkylbetaines and the alkylamphoglycinates and amphopropionates. Betaines, such as cocamidopropylbetaine [*61789-40-0*]

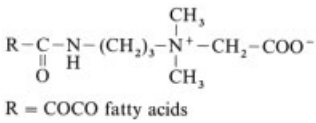

produce creamier and more stable foams than anionic surfactants. When combined with anionic surfactants they also have a thickening effect. Alkylamphoglycinates and alkylamphopropionates show even better compatibility, but their foam characteristics are poorer.

Nonionic surfactants usually foam very little, in contrast to the types mentioned above. Examples include the fatty alcohol ethoxylates or fatty acid polyglycerides, which are among the mildest surfactants that exhibit a detectable level of skin substantivity. Alkylpolyglucosides have also been known for a long time, but only recently have they been used to a significant extent. They display good foam and cleansing properties, as well as good skin compatibility.

7.1.2. Auxiliaries

The production of cosmetically attractive, functional, and hygienically effective shower and bath preparations requires the addition of such auxiliaries as thickening agents, foam stabilizers, fat-restoring agents, turbidity agents, complexing agents, and preservatives.

Thickening agents serve to adjust the viscosity, at the some time contributing to the stabilization of pearl-luster or turbidity agents. Various substances are available for this purpose, depending on the formulation and the desired effect (e.g., fatty acid alkanolamides, cocamide MEA, see structure below). Electrolytes, and (to some extent) amphoteric surfactants cause micelles to become larger, thereby increasing the viscosity.

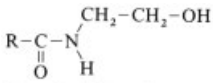

R = alkyl, $C_{12}-C_{14}$

Other substances, such as poly(ethylene glycol) fatty acid esters or hydroxyalkyl cellulose and polyacrylate polymers, create three-dimensional structure, which tends to thicken the aqueous phase.

Foam stabilizers (→ Foams and Foam Control) increase the "creaminess" and stability of foams. The main examples are fatty acid monoalkanolamides and dialkanolamides, polymers, protein – fatty acid condensates, and proteins, all of which have been mentioned previously.

Fat-restoring substances act either by minimizing fat removal from the skin or by restoring lipids to it. Monofatty acid esters of glycerol and diglycerol are particularly suitable for this purpose because of their amphiphilic character and their structural relationship to the natural cutaneous lipids.

Pearl-luster and turbidity agents are often added to shower and bath preparations to enhance properties such as "creaminess" and "richness" consistent with consumer demand. Higher fatty alcohols, insoluble fatty acid alkanolamides, ethylene glycol distearates, monoesters of polyhydric alcohols, and polystyrene emulsions are used for this purpose.

Complexing and sequestering agents form chelates with a wide variety of cations, (e.g., calcium, magnesium, and iron or other heavy metals), thereby suppressing possible decomposition reactions in the product (e.g., Haber – Weiss reactions [78]).

Preservatives are intended to prevent microbial spoilage. The substances chosen and the levels at which they are introduced are largely a function of product composition. The more hydrophobic a preservative, the greater is the danger that it will be trapped in surfactant micelles and thus no longer be effective against microorganisms. For this reason, comprehensive microbiological tests are necessary in each case.

7.2. Product Forms and Types

Cleansing products from the cosmetics industry can be subdivided into liquid and solid preparations. Liquid cleansing preparations are further subdivided into products for the shower, foam baths, cream baths, and oil baths. The latter consist mainly of lipids. Foam and shower products are almost exclusively surfactant preparations. Cream baths occupy an intermediate position. New products in the form of concentrates reflect growing ecological concerns, since their compactness produces savings in packaging materials and transportation. In principle, these new products are formulated in the same way as other liquid preparations, although the raw materials and production methods differ from those of conventional forms. Solid cleansing products include, in addition to soaps (→ Soaps), bath salts,

powders, and tablets. The principal constituents in this case are sodium hydrogencarbonate and sodium metaphosphate. Talcum, silicates, and starch are added as binders. Gelatins and pectins serve as dusting agents.

7.3. Production

Shower and bath preparations are mass-produced products, so the production methods must permit homogeneous mixtures to be prepared rapidly. If all the raw materials are pumpable at room temperature and large amounts of finished product are required, continuous production is employed, with outputs of 5 t/h. All the components are introduced in-line into static mixers via reciprocating pumps. Consistent product quality is ensured by means of measuring and regulating circuits. A batch process is preferred for processes involving more time- consuming mixing, as in the incorporation of solids or the melting of partial phases. In addition to effective, thorough mixing, care is required to ensure that a foam-free product results. Simple, propeller-type stirring equipment is therefore unsuitable. Guided jet mixing turbines have proved effective for this purpose.

8. Deodorants and Antiperspirants

Body odors served originally as a vehicle for facilitating social communication. Thus many animals secrete from special scent glands substances known as pheromones that strongly influence sexual and territorial behavior. Evolution has resulted in this form of communication becoming much less important for humans as evidenced by a severely atrophied human sense of smell. A marked exception to this rule is axillary odor. This very intense odor is now regarded as extremely unpleasant, and it is unacceptable in modern society. One approach to preventing the spread of body odor is the use of deodorants.

The Origin of Body Odor. Body odor is produced by the action of bacteria on apocrine sweat. Living conditions are ideal for such microorganisms in the moist and warm environments of armpits and similar parts of the body [79]. Various gram-positive bacteria are implicated in the decomposition of sweat, which is itself odorless. Two main types of body odor can be assigned to specific types of bacteria. Kligman [80] ascribes the weaker "acid" odor to micrococci, and the "sharp" pungent odor to lipophilic diphtheroids.

8.1. Composition

As the name implies, a *deodorant* is intended to make body odor imperceptible, and various approaches can be adopted to achieve this end.

Antiperspirants reduce the secretion of sweat by partially blocking the outlets of the eccrine sweat glands [80]. This effect is achieved mainly by the use of astringents, such as aluminum salts. The basic aluminum chlorides, such as aluminum chlorohydrate (Chlorohydrol), have proved especially suitable. Mixtures of aluminum and zirconium salts are also used, especially in the United States. These are more effective than Chlorohydrol, but are prohibited for use in aerosols because of their inhalation toxicity.

Antibacterial agents are also important aids in effectively combating body odor. Specific antimicrobial agents are chosen for their selective action against odor-producing bacteria. Examples include *triclosan* (2,4,4′-trichloro-2′-hydroxydiphenylether) [*3380-34-5*], *chlorhexedine* [1,6-di(4- chlorophenylbiguanido)hexane] [*3697-42-5*], and *TCC* (3,4,4′-trichlorocarbanilide) [*101-20-2*].

Doubts were raised in Europe in 1987 concerning the safety of triclosan because of dioxin contamination, so attention has increasingly been redirected to older, well-known odoriferous substances with antibacterial action: essential oils, typically including such constituents as clove oil, mint oil, or thyme oil. Because of their strong intrinsic smell these substances can be used only to a limited extent. An interesting active substance from plants is the terpene alcohol *farnesol* (3,7,11-trimethyl-2,6,10-dodecatrien-1-ol) [*4602-84-0*], which combines good antimicrobial properties with a weak intrinsic smell. *Nisin* [*1414-45-5*] is a peptide with a very selective effect on the lipophilic diphtheroids [81].

Odor absorbers are compounds with the ability to bind odor-forming components. They act by lowering the partial vapor pressure of the bound substance, limiting its progagation rate. Suitable active ingredients include complex zinc salts of ricinoleic acid (12-hydroxy-9,10-octadecenoic acid) [*141-22-0*] or mixtures of special fragrances with a neutral smell.

Odor maskers have existed since the invention of perfumes. They are used to hide body odor, and represent the "classic" method of dealing with it. Even today the presence of perfume contributes substantially to the effectiveness of a deodorant.

8.2. Application Forms

Despite the controversy surrounding propellants, *aerosols* still account for a large part of the deodorant market. Aerosols are alcoholic solutions of active ingredients that have been introduced into their containers under pressure, generally in conjunction with a liquefied gas that serves as a propellant, all in such a way as to form a homogeneous solution. When an aerosol is sprayed the propellant gas vaporizes suddenly and violently, breaking the alcohol solution into very fine droplets (aerosol formation, → Hair Preparations). Hydrocarbons such as propane, butane, and isobutane have largely replaced the chlorofluorocarbons (CFCs) formerly used as propellants. In addition to the alcoholic sprays, there also exist dry deodorant sprays, which are generally antiperspirant suspensions.

Pump atomizers, whose contents consist essentially of aqueous alcoholic solutions, are becoming increasingly popular. The solubility of perfume oils and other hydrophobic substances decreases with increasing water content, so emulsifiers and solubilizers must generally be added to these formulations.

Roll-on dispensers are designed to hold liquid deodorants formulated to have a specific viscosity, thereby ensuring that the contents will be transferred satisfactorily to the skin by the action of a roller ball. Both liquid gels and oil-in-water emulsions have been developed to meet this requirement.

Deodorant sticks must satisfy two conditions: on the one hand they must be hard and dimensionally stable, but at the same time the contents must be transferable to the skin with a light, gentle pressure. Transparent sticks based on soap – glycol gels have such properties. In the presence of long-chain fatty acid sodium salts (soaps), the lower glycols form clear gels with the capacity to absorb both alcohol and water. The *powder stick* represents another type of deodorant. In addition to talcum and aluminum salts, its basic formulation includes fats, waxes, and volatile silicone oils, the latter usually in substantial proportion.

8.3. Testing

To check the effectiveness of potential deodorant active constituents or new formulations, microbiological in vitro methods are invoked first [82]. The effectiveness of an odor absorber can be determined by headspace analyses. After these preliminary tests, consumer-oriented sniff tests are generally conducted. Here each tester is required to use two different products under specified conditions, in most cases for two weeks. The intensity of residual body odor is then determined at prescribed intervals [83].

8.4. Production

The production of deodorants is relatively simple, since all the necessary components can easily be mixed. However, production must always be carried out under explosion-proof conditions because of the alcohol content. Aerosols require special technical equipment for filling and sealing the containers and charging them with propellant.

9. Body Powders

A body powder is an aggregation of solid particles intended as a medicinal or cosmetic preparation for external application to the skin. One function of the particles is to offer mechanical protection, but they also confer a drying and cooling effect and serve as a base for active constituents. The cooling effect is achieved by added stearates or starch. Colloidal silica, precipitated silicic acid, and magnesium carbonate serve to

enhance the drying effect. The surface slip of a powder is increased by the addition of talcum, and adhesion is improved by introducing lubricating components such as petrolatum, paraffin oils, wool wax, or stearates. The properties of powder bases have been described in detail in [84]. Their production is discussed in Section 10.3.

Most modern body powders provide drying or deodorant action. Talcum is the usual base, supplemented with a bacteriostatic agent. Products of this type are available for use on the feet, hands, and axillary regions of the body. Years ago, vulnerary or healing powders were also common cosmetic items, but today these are considered to fall in the medicinal sector, one notable exception being baby powders.

10. Decorative Body Agents, Cosmetics, and Makeup

The desire to appear beautiful and attractive is deeply rooted in human nature, and has affected behavior for thousands of years, even though the means for achieving beauty and the very concept itself have changed constantly over the course of time. More than a thousand years before Christ the Egyptians were already tinting their skin, applying rouge, and highlighting their eyebrows. Present-day concepts of beauty are based on a desire for the face to appear as natural and well groomed as possible. Shades considered appropriate for this purpose vary greatly depending upon current fashion and the type of person involved.

The term decorative cosmetics designates a large variety of different product groups, for example lipsticks, emulsions, gel based formulations and powders. The main difference to skin care products is their content of color pigments (between 3 and 50 %), depending on the formula type and the desired optical effect.

10.1. Types of Decorative Cosmetics

Foundations. The aim of this product type is to freshen and highlight the skin tone and smooth out slight irregularities. In some countries these facial cosmetics are also termed “makeup”.

Foundations are mainly O/W emulsion-based products (see Section 3.2) containing about 5 – 10 % color pigments, usually iron oxides. The consistency of these products can vary between very liquid and creamy.

A *compact foundation* is much more solid and typically poured in a godet after preparation. The formulation of compact foundations is based on a wax-oil base, that contains powdery filling agents such as talcum and 10 – 15 % color pigments (iron oxides and pearlescent pigments). These foundations have mostly a higher coverage than emulsion-based products and have to be spread with a sponge on the skin.

Decorative Eye Cosmetics. They are intended to impart expression to the eyes. The most common eye products are eye shadow powders, mascaras, eyeliners, and pencil products for eyelids and eyebrows.

Eyeshadow powder is a combination of pressed powdery raw materials such as talc, zinc oxide, inorganic and pearlescent pigments, and about 5 % of easily spreadable oils. It is pressed in metal godets and applied with a little sponge or brush.

Mascaras are creamy O/W emulsions containing mostly black iron oxides (up to 15 %). They are applied with a special brush directly on the eyelashes to color them and make them look more expressive.

Liquid eyeliners are applied with a very tiny brush directly on the eyelid above the lashes. Their basic formula is similar to that of mascara, except that the viscosity is much lower to facilitate application.

Examples for *pencil products* are: Eyeliners; kajal, which can be directly applied in the inner eyelid area; or eyebrow pencils which are applied on the eyebrows. The inner part of these pencils is a poured or extruded lead, based on waxes, oils, and color pigments; the outer part consists of wood or plastic.

Lipsticks. The most common and fashion related decorative product is the lipstick. A large variety of colors are available to impart more emphasis and expression to a women’s face.

A modern lipstick is composed of suitable mixture of high and low melting waxes; inorganic, organic, and pearlescent pigments; and a

composition of oils. Castor oil is the most commonly used oil because of its glossy effect on the lips and easy dispersion of the color pigments.

Lipsticks are discussed in detail in Chapter 6.

Nail Varnish/Nail Polish. To highlight the fingernails and toenails, these quick-drying lacquer formulations are available in a very large variety of shades. Nail polishes are anhydrous compositions of film formers, plasticizers, and a mixture of suitable solvents. As color components inorganic and organic pigments are used as well as pearlescent pigments to impart the lacquer a metallic shine.

Nitrocellulose is traditionally used as the base for nail polish because of its ideal film forming properties. As plasticizer camphor and dibutyl phthalate are frequently present. Polyacrylates or polymethacrylates are employed to add luster. Solvents are used to control flow, viscosity, evaporation rate, and drying time of the polish. The most widely used solvents are butyl acetate and ethyl acetate.

10.2. Colorants

The following colorants are the ones most widely utilized:

1) White pigments – talcum, zinc oxide, kaolin, titanium dioxide
2) Inorganic pigments, e.g., titanium dioxide, iron oxide, ultramarine, manganese violet, or chromium oxide
3) Organic pigments, e.g., Quinoline Yellow, D&C Red 6/7, Brilliant Blue FCF
4) Nacreous or pearlescent pigments (producing a pearl-gloss effect), e.g., mica, titanium dioxide-coated mica, iron oxides-coated mica, (nacreous pigment; → Pigments, Inorganic, Chap. 4.3.1.3.), bismuth oxychloride.

Colored pigments should have a particle size of 50 – 100 µm to prevent them from penetrating into deeper layers of the epidermis [86].

All approved colorants have been tested for their compatibility with the intended application (see, for example, the German Cosmetics Statutory Regulation, KVO). Additional regulations apply with respect to heavy-metal content, ancillary colorants, and starting materials.

10.3. Production

Emulsion bases are prepared as described in Chapter 3.

For production of *emulsion bases* the pigments have to be dispersed homogeneously in the water or oil phase of the emulsion before the emulsifying process is started. The process chosen is depending on the type of formula. The further production stages are similar to those described in Section 3.6.

For the production of *powders* color and pearl pigments as well as powdery raw materials such as talcum and zinc oxide are mixed and milled until the phase is homogeneously dispersed. Oils which have a binding capacity for the powder are then added and the whole phase is mixed again until uniform. The powder is subsequently pressed into godets.

For the production of *lipsticks* oils and waxes are melted in a vessel. Then predispersed color pigments are added to the hot phase and stirred until uniform. This bulk is filled into molds and cooled down quickly. The lipstick bullets are then ready to be filled into the lipstick mechanism.

11. Sunscreens

Sunscreens in the broadest sense provide protection against the harmful effects of light (i.e., electromagnetic radiation with a wavelength of between 1 and 10^6 nm) [87, 88]. In the technical and industrial sector they are used to protect materials such as plastics or other products that can be discolored or decomposed by light (→ Plastics, Additives, Chap. 3.).

They are also of significant economic importance. In Germany, for example, annual production value exceeds 100×10^6 DM, and involves 35 producers [89]. In 1991, the annual production value for Europe was ca. $\$954 \times 10^6$ and by 1995 is expected to reach ca. $\$1130 \times 10^6$ [90].

Direct damage due to light is restricted largely to the skin (including the hair and nails) and the eyes. Those light wavelengths which, on account of their high quantum energy, preferentially initiate photochemical reactions in body constituents are absorbed in the outermost layers of the body. The effects of light can be divided into four main groups:

1) *Acute Skin Damage.* Sunburn (erythema) with subsequent post-inflammatory (indirect) pigmentation and regenerative thickening of the horny layer ("sun callus"). Ultraviolet B radiation (UV-B, 280 – 315 nm) with maximum effect at ca. 297 nm is mainly responsible for sunburn. Since the lower wavelength limit of terrestrial sunlight is between 295 and 305 nm, the erythema maximum of sunlight is somewhat higher (300 – 310 nm). In the medical and industrial sector, for example in UV sterilization or materials testing, protection against ultraviolet C (< 280 nm) from artificial UV sources is also important. UV-C also produces erythema, but has a smaller penetration depth.

2) *Chronic Skin Damage.* Premature skin aging with elastotic degeneration of connective tissue, formation of precancerous tumors and, ultimately, malignant tumors (basal cell carcinomas, spinocellular carcinomas, and possibly a type of melanomas). UV-A (315 – 390 nm), UV-B, and probably also the shorter-wavelength visible light are most important in the formation of such tumors on account of their substantially higher penetration depth.

3) *Light-Induced Skin Disorders.* These include exogenous or endogenous dermatoses, such as xeroderma pigmentosum or phototoxic and photoallergic reactions. UV-A is generally a more important causative factor than other light wavelengths.

4) *Light-Induced Ocular Damage.* Visible and UV light may affect the conjunctiva, retina, and lens (cataracts). The eyelids generally provide sufficient protection against visible light, but artificial protection must be used to shield the eyes against UV wavelengths. In recent years, the field of ophthalmological sunscreen preparations has largely been ignored as a result of the development of suitable protective glasses, so this group of sunscreens is not discussed here.

11.1. Composition

A light-protective effect is achieved by producing a layer over, on, or in the horny layer (stratum corneum) of the skin. This layer protects the living cells in the epidermis from the damaging components of light by reflection, scattering, or absorption. The disruption of photochemical primary and secondary reactions in the skin has up to now been exploited only in counteracting light hypersensitivity caused by disturbances in porphyrin metabolism: internally administered β- carotene is employed as a radical scavenger.

Clothing has a major light-protective action on the skin. The color of the clothing has little effect, whereas its thickness and nature, and the condition of the textile fibers, play a decisive role. Rough natural fibers, including artificial silk (rayon), have a substantially greater light-protective effect than the smoother synthetic fibers. Either inorganic pigments, (e.g., zinc oxide, or titanium dioxide), micabased pigments, or more recently, micronized pigments, are used as sunscreens on and in the horny layer. They act as "physical" light filters, mainly by reflection and scattering. When used in large amounts, they have the disadvantage of forming a conspicuous layer on the skin. For this reason light filters that act by absorbing specific wavelengths ("chemical" light filters) have so far been most important. Employed in a suitable base, they penetrate the outer layers of the stratum corneum, producing a layer 5 – 30 µm thick that specifically absorbs UV-B and possibly also UV-A light. A great many substances with maximum absorbances from 300 – 310 nm (for UV-B) or 330 – 360 nm (for UV-A) are possible candidates for use as UV filters. However, the number of compounds tested or actually used is still very small (Table 1) due to the fact that most potentially suitable substances are either too expensive or too toxic. Special technological requirements are also placed on UV filters, the most important being light stability and the absence of marked fluorescence in the visible region. Suitable substances must penetrate as little as possible into the living skin (epidermis), and should display a certain degree of "substantivity" with respect to keratin (analogous to the substantivity of a dye).

Table 1. Light-protective substances, tested or in actual use

Name	CAS registry no.	formula	Other name	Listed in CTFA [a]	Listed in Blaue Liste [b]	Listed in INN [c]
UV-B filtering substances						
4-Aminobenzoic acid	[*150-13-0*]	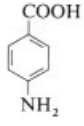	Colipa S 1	PABA	UV 530	
2,3-Dihydroxypropyl-4-aminobenzoate	[*136-44-7*]	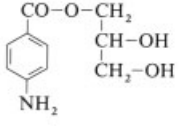	Colipa S 6	glyceryl PABA	UV 532 (e.g., Escalol 106)	
2-Ethylhexyl 4-dimethyl-aminobenzoate	[*21245-02-3*]	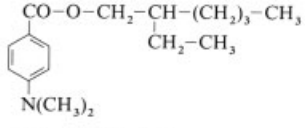	Colipa S 8	octyl dimethyl PABA	UV 535 (e.g., Escald 507, Eusolex 6007, Padimate 0, Uvasorb DMO)	
Ethyl 4-[bis((2-hydroxy-propyl)amino]benzoate	[*58882-17-0*]	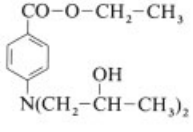	Colipa S 2	ethyl dihydroxy-propyl PABA	UV 537 (e.g., Amer-screen P)	Roxadimate
2-Ethylhexyl 4-methoxy-cinnamate	[*5466-77-3*]	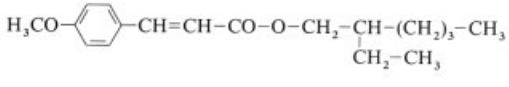	Colipa S 28	octyl methoxy-cinnamate	UV 445 (e.g., Absol EHM, Neo-Heliopan AV, Parzol MCX, Sunarome OMC)	
2-Ethylhexyl salicylate	[*118-60-5*]	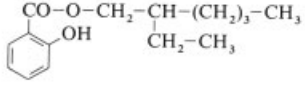	Colipa S 13	octyl salicylate	UV 322 Sunarone (e.g., Heliopan)	

[a] [91]. [b] [92]. [c] [93].

Table 1. (Continued)

Ethylpoly(oxyethylene) 4-{bis-[2-hydroxyethyl)poly(oxyethylene)]amino}benzoate	[*116242-27-4*]	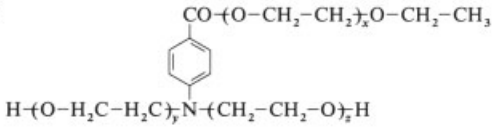	Colipa S 3	PEG-25-PABA	UV 539 (e.g., Lusanthan 25, Uvinul P25)
3-(4-Imidazolyl)acrylic acid	[*104-98-3*]	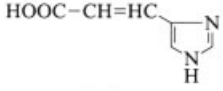	Colipa S 46	urocanic acid	UV 823 (e.g., urocaninic acid)
Isopentyl 4-methoxycinnamate	[*71617-10-2*]	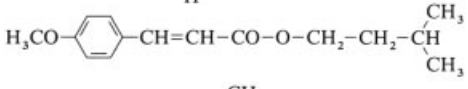	Colipa S 27	isoamyl *p*-methoxy cinnamate	UV 444 (e.g., Neo-Heliopan 1000)
4-Isopropylbenzyl salicylate	[*94134-93-7*]	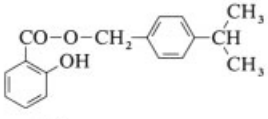	Colipa S 16	isopropylbenzyl salicylate	UV 328 (e.g., Megasol)
α-(2-Oxo-3-bornylidene)-toluene	[*36065-10-8*]	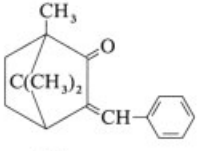	Colipa S 61	3-benzylidene camphor	UV 121 (e.g., Ultracyd)
α-(2-Oxo-3-bornylidene)-4-toluene sulfonic acid	[*56039-58-8*]	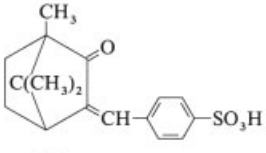	Colipa S 59		UV 124 (e.g., Mexoryl 5L)
α-(2-Oxo-3-bornylidene)-*p*-xylene	[*36861-47-9*]	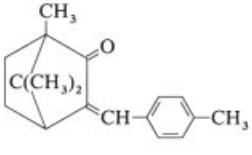	Colipa S 60	4-methylbenzylidene camphor	UV 122 (e.g., Eusolex 6300)

Table 1. (Continued)

2-Phenyl-5-benzimidazole sulfonic acid	[*27503-81-7*]	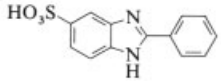	Colipa S 45	phenylbenzimidazole sulfonic acid	UV 854 (e.g., Eusolex 232, Neo-Heliopan, Hydro, Novantisol)	
3,3,5-Trimethylcyclohexyl salicylate	[*118-56-9*]	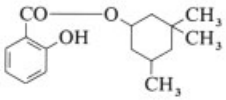	Colipa S 12	homosalate	UV 324 (e.g., Filtrosol, Homomenthyl salicylate)	Homosalate
N,N,N-Trimethyl-α-(2-oxo-3-bornylidene)-*p*-toluidinium methylsulfate	[*52793-97-2*]	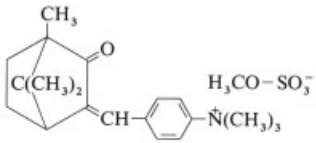	Colipa S 57		UV 127 (e.g., Mexoryl SK)	
2,4,6-Tris{4-[(2-ethylhexyl)-oxycarbonyl]-phenylamino}-1,3,5-triazine	[*88122-99-0*]	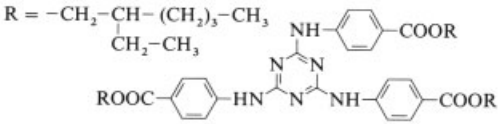	Colipa S 69	octyl-triazone	UV 536 (e.g., Uvinul T 150)	

Table 1. (Continued)

UV-A filtering substances						
1-(4-*tert*-Butylphenyl)-3-(4-methoxyphenyl)-1,3-propanedione	[*70356-09-1*]			butyl methoxydibenzoylmethane	UV 664 (e.g., Parsol 1789)	Avobenzene
1-(*p*-Cumenyl)-3-phenyl-1,3-propanedione	[*63250-25-9*]		Colipa S 64	isopropyl dibenzoylmethane	UV 653 (e.g., Eusolex 8020)	
5-Benzoyl-4-hydroxy-2-methoxybenzene sulfonic acid	[*4065-45-6*]		Colipa S 40	benzophenone-4	UV 623 (e.g., Uvinul MS-40, Uvistat 112)	Sulisobenzone
2-Hydroxy-4-methoxybenzophenone	[*131-57-7*]		Colipa S 38	benzophenone-3	UV 613 (e.g., Cyasorb UV 9, Eusolex 4360, Neo-Heliopan BB, Uvinul M-40, Uvistal 24)	Oxibenzone
2-Hydroxy-4-methoxy-4′-methylbenzophenone	[*1641-17-4*]		Colipa S 39	benzophenone-10	UV 614 (e.g., Uvistat 2211, Uvistat L)	Mexenone
3,3′-(1,4-Phenylidenedimethine)bis(7,7-dimethyl-2-oxobicyclo(2,2,1)heptane-1-methane sulfonic acid	[*90457-82-2*]		Colipa S 71		UV 135 (e.g., Mexoryl SX)	

11.2. Evaluation

Determination of Protection against Erythema (UB-B Protection). In most countries sunscreens are classed as body care agents (skin care products) or cosmetics, although in the United States they are regarded as over-the-counter drugs. This classification appears justified because sunscreens, in contrast to most cosmetics, must have an objectively assessable effect.

The *mean protection factor* proposed by R. SCHULZE [94] was adopted as a scale for objective evaluation a few years after its publication in 1956. This factor is defined as the arithmetic mean of individual protection factors measured on of at least twenty test subjects with skin types representative of those expected for subsequent users. The individual protection factor is determined by measuring the light dose (exposure time) that produces a defined erythema in the untreated skin of a test subject. At the same time this measurement is also made on skin treated with the test product. The protection factor is the ratio of these two exposure times:

$$\text{Protection factor} = \frac{\text{MED}_{\text{Prot.}}}{\text{MED}_{\text{Unprot.}}}$$

MED = exposure time required to produce a specific erythema reaction
Prot. = for protected skin
Unprot. = for unprotected skin

If, for example, an exposure time of 2 min with the chosen light source (sun or lamp) is required to produce erythema in unprotected skin, and an exposure of more than 8 min is required to produce a recognizable erythema in the protected skin after ca. 24 h, the protection factor for the test subject and the test preparation is 4. In practice this means that, after application of the test product, sunbathing time can be increased roughly four-fold before sunburn becomes apparent.

Measurements are conducted according to fixed protocols (e.g., DIN 67 501 or FDA). The test subject's body is divided by horizontal adhesive strips into areas of skin measuring ca. 10×150 mm. Every other area is then treated with a specific amount of the test preparation. These horizontal areas are next divided by vertical adhesive strips into areas of ca. 1 cm^2. After the onset of exposure to radiation of a specific wavelength the vertical rows of skin regions are covered successively from left to right at specific time intervals. The adhesive strips are subsequently removed, and after ca. 24 h the erythema threshold (the first region with a defined erythema test reaction) is identified and associated with the exposure time required to produce this reaction.

The radiation source, the composition of the group of test subjects, and the accuracy of the erythema threshold readings are the most important factors in assuring reproducibility of the results.

Sunlight would, of course, be the most reliable source. However, this is impractical because the necessary exposure time (minimum erythema dose, MED) is seldom less than 10 min even for sensitive skin. Furthermore, the total amount of available sunlight seldom reaches 15 times the MED on any one day, even in the vicinity of the equator. For these reasons (higher) protection factors cannot be determined in sunlight. A wide range of artificial light sources is therefore used instead. The results will approximate practical conditions (i.e., exposure to sunlight) only if the relationships between the light wavelengths and intensities (UV-B, UV-A, visible, and near IR) correspond to those prevailing in sunlight. With this proviso, it appears to be of relatively minor importance whether the radiation source displays a continuous or a band spectrum.

Methods used in Europe (DIN), the United States (FDA), and Australia (Standards Association of Australia, SAA) to determine the protection factor differ in the choice of radiation source (and thus in the spectral composition of the light) [95], [96]. Comparison of the values obtained by these methods is further complicated by the fact that different amounts of preparation are applied, and MED determinations are sometimes conducted on widely separated regions of the skin (possibly even with different light sensitivities), or on different days. This means, for example, that values determined by the United States method are generally higher, particularly with powerful sunscreens, than those obtained with the DIN method.

As a result of the tendency to proclaim ever higher protective factors in product advertisements, coupled with problems in reproducibility, attempts are being made to replace numerical protection factors with more straightforward protection *categories*, which are in turn based on the experimentally derived numerical values:

Protection factor	Classification
2–4	weak
4–7	moderate
7–10	strong
>10	very strong (sun blocker)

The practical significance of such a classification is illustrated in Figure 2 [97]. This particular assignment of sunscreen classes reflects more clearly the wish of the user to develop a "sporty" suntan than to follow strict medical considerations.

Those areas of the skin regularly exposed to sunlight, i.e., the face, neck, and the back of hands, are protected best with a sun blocker, since this is the only way of achieving effective prophylaxis against skin aging and light-induced cancer.

Determination of Protection against UV-A. Protection against UV-A has become increasingly important. The starting point in the development of "broad-spectrum sunscreens" (preparations effective against both UV-B and UV-A) was a search for "therapeutic sunscreens," since UV-A often plays an initiating role in light-induced skin disorders. Interest in broad-spectrum sunscreens for preventing light-induced skin aging and its consequences increased along with knowledge about the importance of UV-A in damage to the deeper layers of the skin (i.e., connective tissue and blood vessels in the dermis).

Experimental testing of such prophylactic action is impossible, however, because the skin changes of interest develop only over the course of years. Indirect methods must therefore be used to detect UV-A protective action. Use in this context of subjects with a pronounced (pathological) skin reaction to UV-A exposure is obviously excluded, not only for ethical reasons, but also because only very few such patients exist. Testing based on UV-A-induced erythema is generally impractical as well, because of the very high UV-A doses required.

Objections also exist to evaluation via the inhibition of UV-A-induced "direct pigmentation" [99], which involves photooxidation of the skin pigment melanin. This pigment is seldom present in sensitivity type I subjects (Fig. 2), and is only present in small amounts in type II subjects. Such tests are therefore possible only on skin types III and IV, which are rare among northern Europeans, and thus hardly representative. Furthermore, the melanin that is photooxidized by UV-A is located predominantly in the epidermis, so it is almost impossible to draw any conclusions about the extent to which UV-A may have penetrated the dermis.

An alternative indirect method, namely UV-A-induced photosensitization in the presence of an appropriate sensitizing agent, is now the most common approach to determining protective effects against UV-A. The primary photosensitization reaction takes place mostly in the dermis, and manifests itself in an erythematous reaction for which a threshold time can be measured, as in the case of UV-B-induced erythema.

This measurement can be performed by treating the whole test region of the skin with a 0.15 % solution of the photosensitizer 9-methoxy-7*H*-furo(3,2-*g*)(1)benzopyran-7-one (8-methoxypsoralene [*298-81-7*] in 30 % 2-propanol. After evaporation of the solvent, the skin is divided into areas as in the UV-B method (see above). Illumination with black-light tubes (e.g., Philips TLK 40W D9N) begins 30 min after application of the photosensitizer. The erythema reaction, which is pathogenetically quite different from UV-B erythema, can be measured, at the earliest, after 48 h. A UV-A protective factor is assigned in the same way as in the UV-B method. Investigation of >100 products showed protection factors significantly above 10 for certain dibenzoylmethane derivatives.

Determination of Water Resistance. The need for water-resistant sunscreens results from the fact that many preparations rapidly lose their effectiveness in the presence of sweat and during swimming and water sports. This is dangerous for swimmers subject to high solar intensity since UV-B radiation passes almost unattenuated through water.

There is no universally accepted test method for water resistance. Investigations have shown that the rinsing of test areas on seated subjects

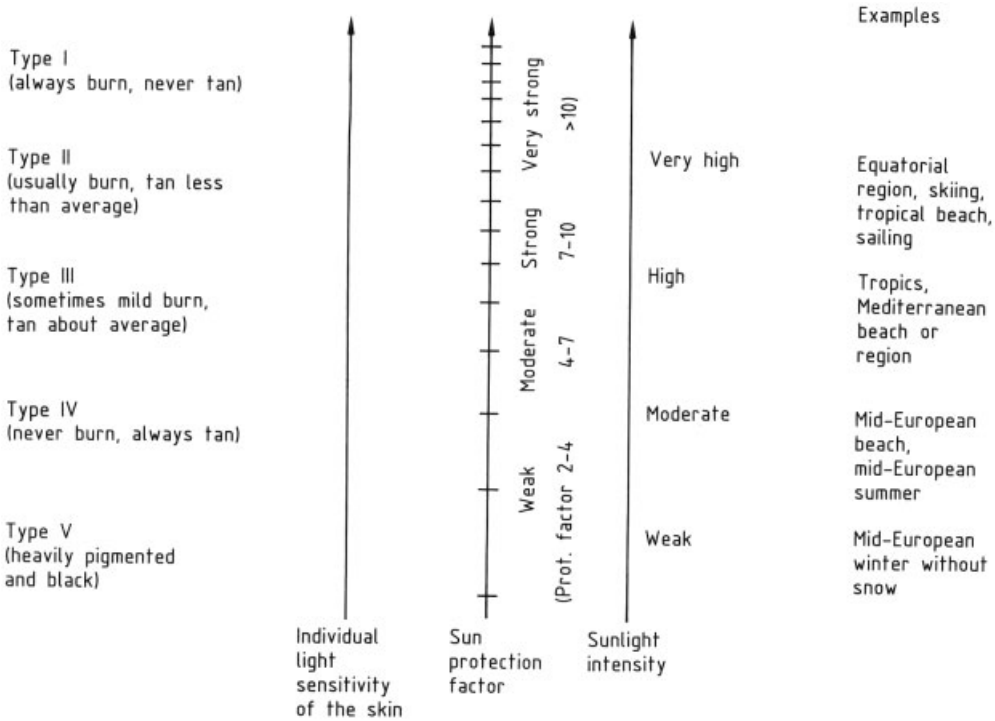

Figure 2. Nomogram for choosing a sunscreen [97], taking into consideration the sensitivity of an individual's skin and the intensity of sunlight [98].

with tap water at 21 °C for 20 min provides reproducible results. Seawater generally produces higher water-resistance values than tap water, presumably because of a salting-out effect in the sunscreen film on the skin surface.

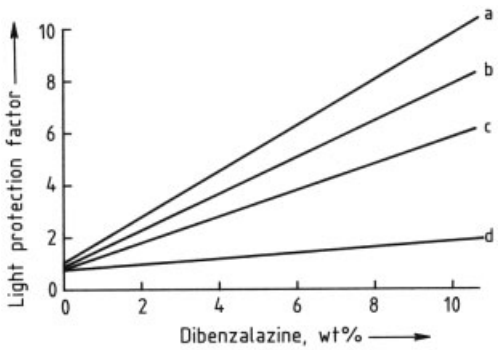

Figure 3. Dependence of light-protective action on the cosmetic base
a) Petrolatum; b) Water-in-oil emulsion; c) Oil-in-water emulsion; d) Oil

11.3. Effectiveness and Production

Several factors may be responsible for the observed discrepancy between physically measured UV-absorbance values and biological protective action on human skin: the layer structure may differ from one preparation base to another, penetration of the filter substance into the horny layer may vary, and some filter substances may react with certain constituents in the horny layer.

Systematic investigations reveal connections between the filter substance, cosmetic base materials, and biological protective action which have so far not been satisfactorily explained. For example, dibenzalazine [*588-68-1*] (EusolexR 6653), previously used as a light filter, showed very different increases in protective action with increasing concentration depending on the base (Fig. 3). Similar results were obtained with 4-methoxycinnamates [100]. With emulsions the protective effect increases with the water content [101]. The very weak protective action of oils is easily explained, because oils do not form a homogeneous layer on the skin [102]. Figure 4

demonstrates that the protective effect increases more significantly with filter content if an oily microemulsion is used as the base instead of an oil. Diffusion of filter substance from the base into the horny layer appears to be an important factor.

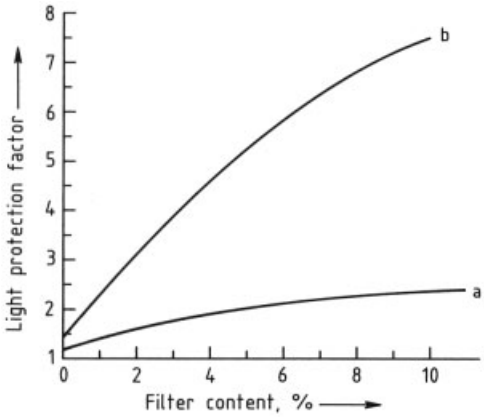

Figure 4. Improvement in light protection when an oil base is replaced by an oily microemulsion: liposoluble 2-phenyl-5-methylbenzoxazole in oil (a) and a transparent emulsion (b) [101]

Virtually all base types used in the body care sector lend themselves to sunscreen preparation, the most popular being emulsions. The precise formulations of such products are generally proprietary secrets. For information on generic formulations see [87]. Reference should be made to [103–105] for details of analysis.

11.4. Toxicology and Environmental Protection

An important reason for the small number of filter substances currently in use is the requirement for detailed, comprehensive toxicological investigations, because sunscreen preparations are generally applied repeatedly, and over a large area of the skin. Even if the concentration of a given filter substance is only 5 %, 1 – 5 g of the substance would be applied to the skin in a single application of 20 – 100 g of finished product. Absorption of only 1 % of the applied filter substance implies an uptake of 10 – 50 mg by the body after a single application. The increasing popularity of "physical" light protection (see above) is therefore welcome on medical grounds, although micronized pigments must also be subjected to thorough toxicological testing.

From the medical standpoint, sunscreen application requires careful risk – benefit evaluation despite the proven value of these agents in preventing or reducing acute and chronic skin damage. This is particularly applicable to products for pediatric use and extremely active products. The regular application of strong sunscreens may also influence photosynthetic vitamin D formation [106], [107] and immunological processes in the skin [108], [109].

Weighing the risks and benefits is easiest when a sunscreen leads to undesirable skin reactions [110]. Allergic and photoallergic eczemas demand lifelong avoidance of the causative agents. In the case of a photoallergy this means avoiding dibenzoylmethane derivatives, which remain indispensable today as UV-A filters. Successfully avoiding a filter substance that has proved to be an eczematogen obviously requires awareness of its presence in a commercial preparation [111].

Finally, modern sunscreen toxicology requires raising the question of the impact of these products on the environment, especially in view of the large quantities of sunscreens available on the market (estimated production during 1991 in Germany alone was 5000 t [89]). It is therefore important to inquire into the ultimate fate of these materials – in chlorinated swimming-pool water, for example. Chlorination of the benzene ring of methoxycinnamates is certainly conceivable [112].

12. Stability Testing

Stability information supplied in conjunction with a particular product must be regarded as binding with respect to each individual production batch. For this reason the establishment of reliable stability parameters constitutes an important goal during the systematic development of a new cosmetic product.

Stability projections are always based on many individual test results obtained gradually over the entire course of product development. The overall procedure can be divided into the following stages:

1) Stress and induced aging tests

2) Long-term testing
3) Routine follow-up tests during production
4) Tests for the detection of the effect of changes introduced into the production process

All subsequent procedures, from the selection of test batches up to stability assessment, apply to each of the above stages. These are, however, of differing importance in the projection of stability of a product.

12.1. Storage Conditions

In establishing appropriate storage parameters, a clear distinction must be made among conditions applicable to stress and induced aging tests, long-term tests, routine follow-up tests, and tests related to changes during manufacture.

Whereas the storage-temperature and humidity conditions for long-term testing are normally based on typical climatic conditions in the countries in which a product is to be marketed, stress and induced aging tests are usually performed on samples stored at minimum and maximum temperatures and/or humidity levels for the markets of interest, or under constantly varying conditions. Storage under diffuse daylight conditions or with exposure to ultraviolet radiation can also affect product quality, especially with transparent or translucent packaging. Stability testing is sometimes necessary under these conditions as well.

12.2. Test Methods

The choice of stability-testing methods is usually based on scientific knowledge of the products in question as well as on manufacturing experience. Methods routinely invoked can be divided into several categories.

Organoleptic Studies. Organoleptic (sensory) studies are normally conducted to check the following product characteristics as a function of storage time and storage conditions: appearance, odor, and in-use behavior.

Package Compatibility Tests. Some cosmetic formulations are capable of reacting with components of the packages in which they are stored. Organoleptic studies, as well as tests for loss of product weight and physical changes of the packaging material (distortion, softening, or hardening), are useful in this context [65]. The following phenomena are frequently monitored: (1) water or oil leakage, (2) phase separation, (3) changes in color, (4) changes in odor, (5) weight loss (especially in plastic containers), and (6) physical changes in the packaging material.

Physicochemical Studies. The long list of scientific publications dealing with physicochemical characterization of cosmetic formulations makes it worthwhile to consider separately the study of cosmetic emulsions, detergent-based products, gels, and sticks [113–116].

Emulsions. Instability of a cosmetic emulsion is usually characterized by increased flocculation and coalescence of the dispersed phase, followed by creaming, which may lead ultimately to phase separation. Many attempts have been made to characterize these phenomena by physical or physicochemical means. The methods used include:

1) Microscopic analysis of age-dependent changes in droplet size distribution [117]
2) Study of time-, temperature- and force-dependent induction of phase separation through centrifugation, especially as a function of storage time [118], [119]
3) Rheological characterization on the basis of flow curves, oscillation measurements (frequency sweep, strain sweep, time sweep, temperature sweep), and creep studies all of which are capable of distinguishing between stable and unstable emulsions [120–122]
4) Thermoanalysis based on differential scanning calorimetry and thermogravimetry [123]

Use of the following techniques has also been described: laser scattering [124], [125], pulsed nuclear magnetic resonance [126], turbidometric measurements [127], and measurement of changes in conductivity or in dielectric properties [128–131].

Detergent-Based Cosmetics. Apart from rheological characterizations, the detection of color changes and changes in distribution of the opacifier(s) employed are often important in the study of detergent-based cosmetics. Inhomogeneity

of the opacifier can be detected by examining the following parameters [132]: fluctuation in reflected light intensity, turbidity, and gloss. Changes in color are usually determined by simple photometric techniques.

Gels. Gels are also subject to rheological changes, but color changes and, in the case of clear gels, increasing turbidity can cause problems. Methods useful for monitoring color changes and turbidity have already been discussed.

Sticks. Deodorant sticks or lipsticks may exhibit any of the following undesirable characteristics after storage at certain temperatures: deformation, change in adhesiveness with respect to the cartridge material, crystallization, or leakage of oil or water.

Microbiological Studies. All cosmetic products are subject to microbial contamination. Aqueous cosmetics may promote the growth of yeasts, molds, and bacteria unless appropriate preservatives are included in the formulations. Dry cosmetics (mascara, eye shadow, etc.) are also subject to contamination through repeated use. Testing the efficacy of preservatives is very important in assuring the safety and stability of cosmetic products. Various test methods have been suggested for estimating microbial stability [133–136].

Chemical Studies. For products containing active ingredients, quantitative assay of these components is worthwhile for various storage periods during the overall test period. Autoxidation and hydrolysis are the most general problems encountered during storage. Quantitative study often makes it possible to predict rates of potential chemical changes during storage.

13. In Vitro Compatibility Testing

13.1. Introduction

Body care agents and toiletries are found in virtually every household. They are used often, sometimes over long periods of time and in large quantity and for a wide range of purposes on the skin, hair, and mucous membranes. Some remain in contact with the body for prolonged periods and at varying levels of intensity. For this reason, consumers expect the products to be safe when used according to instructions and for the intended purposes. The obligation on the part of legislators and cosmetic producers to protect the consumer is reflected in a comprehensive legal framework. For example, several laws and regulations apply to cosmetics distributed in Germany. These include the Chemicals Law with respect to raw materials evaluation, the Food and Consumer Goods Law, and the Cosmetics Regulations (including several appendices), which are based on the European Cosmetic Directive and its amendments [127–142], as well as recommendations from the Federal Health Office and the Scientific Committee on Cosmetology of the European Community [143], [144]. The combination of this binding legislative framework with extensive public discussion on the banning of animal experiments, particularly for cosmetics, has led to a serious dichotomy that is forcing a reexamination of conventional means of ensuring product safety.

Against the background of the manufacturers' responsibilities on the one hand and consumer expectations of product safety on the other, producers of cosmetics are making concerted attempts to develop new conceptual approaches toward satisfying the increasingly stringent requirements for the safety of their products, but doing so without recourse to animal experiments on the finished products.

13.2. Test Methods

In several EC countries and in the United States, producers are engaged in intensive efforts to validate test methods that do not involve animal experimentation, but still permit adequate evaluation of finished products, especially in terms of local compatibility with the skin, hair, and mucous membranes. Much current emphasis is on new tests for ocular mucous membrane and skin irritation. Numerous alternative test methods have been proposed [145], and the problem now is to select those that best satisfy important criteria, including ease of handling and processing, reliability (i.e., applicability to all relevant groups of materials and products), sensitivity, and reproducibility. Through joint efforts on the part of those responsible, and with support from

bodies such as the European Center for the Validation of Alternative Methodologies (ECVAM, Italy), the Center for Documentation and Evaluation of Substitutes and Supplements for Animal Testing (ZEBET, Germany), the Fund for the Replacement of Animal Testing in Medical Experimentation (FRAME, United Kingdom), and the Center for Alternatives to Animal Testing (CAAT, United States), access is now available to various means of validating these alternative methods [147–149], which must also be accepted by the regulatory bodies involved. International studies [146], [150–152] have led to the following methodological approaches:

1) *Biochemical and physicochemical properties* of constituents and products can be used to evaluate specific interactions [150].
2) *Cytotoxicity tests* on isolated blood cells [147], [149] or permanent cell lines [145], [152], [150] are extremely sensitive and, for relevant classes of substances, show good correlation with acute ocular irritation. Another advantage of these processes is that they can be partly automated.
   Furthermore, cytotoxicity tests have been successfully used for in vitro photoirritation testing. As a result of a joint EU/COLIPA validation study of in vitro phototoxicity tests the 3T3 Neutral Red Uptake Phototoxicity Test has been validated [150], [153], [154], [156] and proposed by the Scientific Committee for Cosmetic Products and Non-food Products (SCCNFP) as the standard method for testing the UV light absorbing cosmetic ingredients or mixtures of ingredients for phototoxic potential [157], [158].
3) *Hen egg tests on the chorioallantosis membrane* (HET-CAM) [155], [157]. Macroscopic changes caused by the action of mild to moderate eye irritants are readily visible in the sensitive vascular system of incubated chicken eggs, appearing as dilation and enhanced permeability of the small blood vessels [155].
4) *Organotypical skin cultures* and *excised skin* are particularly valuable for answering questions involving the skin barrier. For in vitro dermal absorption/penetration studies with excised skin standardized methods are available [159], permitting the evaluation of numerous biological parameters. Three-dimensional skin cultures have also been used and are currently prevalidated for in vitro skin irritation and photoirritation testing. Such techniques are suitable for substances that are unstable under UV-A or UV-B radiation and that, as a consequence, might induce irritation [161].
5) *Skin sensitization potentials* cannot yet be examined in vitro. Currently, research on the suitability of cytokines (especially IL-1β) as endpoint for a discrimination between sensitizers and irritants and on biotransformation (prohapten activation) is carried out in projects sponsored by COLIPA/SCAAT and ECVAM.

Known toxicological safety data must be used in conjunction with results of the in vitro tests to provide a basis for evaluating a complete cosmetic preparation prior to dermatological testing. Every product undergoes final evaluation in dermatological tests consistent with good clinical practice (GCP), a prerequisite for introducing the product onto the market.

14. Economic Aspects

Body and skin care products have always been – and are now perhaps even more than ever – very important for bodily well-being. Relatively precise figures are available for the cosmetic markets in the United States, Japan, and Europe, as shown in Table 2. Market volumes for the United States and European markets are comparable. If its smaller population is taken into account, Japan also appears prominently in the picture. In Europe the largest markets in the cosmetics sector are in France, Italy, and Germany. Face care products play a predominant role in every country. However, Japan, with an annual turnover of more than 6×10^9 €, leads in the facial products sector by a wide margin.

15. Legal Aspects

Cosmetics are regarded as consumer products, and are therefore subject to legal provisions that protect the consumer against misrepresentation and damage to health. In Germany, for example, cosmetics are defined as preparations used externally on persons for purposes of cleansing,

Table 2. The international market for cosmetic products

	Market volume, 10^9 €		
Product category	United States	Japan	Europe
Skin and face care	5.6	6.8	8.7
Body care	3.8	1.0	4.7
Hair treatment agents	5.5	4.2	6.2
Perfume	2.4	0.2	3.2
Decorative cosmetics	3.2	3.1	3.3
Other	2.6	1.4	4.1
Total	23.1	16.6	30.3

care, or influencing appearance or body odor (§ 4 of the Food and Consumer Goods Law, LMBG [139]). In 1976, these provisions were expanded by the introduction in all EC states of the EC guideline "Cosmetic Agents" [140]. In Germany this guideline and the LMBG were finally combined in 1977 to form the Cosmetics Statutory Regulation (KVO) [138], which lists forbidden constituents as well as substances permitted with certain restrictions (e.g., preservatives, coloring agents, and UV filters). A few EC member states have regulations that deviate from EC guidelines. In France and Spain, for example, both qualitative and quantitative information must be provided for publicly marketed active constituents. It is not legally necessary in Europe to label the ingredients on the package, but manufacturers are increasingly doing so nonetheless. In the United States, a distinction is made between cosmetics and over-the- counter (OTC) drug products. As in the EC, certain constituents such as UV filters and coloring agents are explicitly included in lists of regulated substances. OTC preparations must also satisfy certain additional requirements, the details of which depend on the product category. According to U.S. law, all constituents of a cosmetic formulation must be listed on the package [160].

In Japan all cosmetic constituents are included in official registry lists. Almost 600 substances are described in great detail in book-length publications covering specifications, production, and intended use. Other raw materials are listed in the Japanese Cosmetic Ingredient Dictionary (JCID) [161]. All new substances must be reported to the relevant authorities. Depending on the product type and intended use, cosmetics are subdivided into 78 different groups, further distinguished according to whether they are cosmetics or quasi drugs. The latter are permitted to contain only active constituents that have been authorized in this category. Legal provisions in many countries are based on laws applicable in the United and the European Community. For example, regulations in Austria, Norway, and Turkey are similar to the EC guidelines. In Latin America and Southeast Asia, cosmetics must be registered with the appropriate authorities. Other countries such as Sweden require that cosmetic products be reported to a national poisons office before they are introduced on the market.

16. References

1. G. F. Odland in L. A. Goldsmith (ed.): *Physiology, Biochemistry and Molecular Biology of the Skin,* 2nd ed., Oxford University Press, Oxford 1991, pp. 3 – 62.
2. P. W. Wertz, D. T. Downing in L. A. Goldsmith (ed.): *Physiology, Biochemistry and Molecular Biology of the Skin,* 2nd ed., Oxford University Press, Oxford 1991, pp. 205 – 236.
3. D. T. Downing et al., *J. Invest. Dermatol.* **88** (1987) 2 s – 6 s.
4. P. M. Elias, *Drug. Dev. Res.* **13** (1988) 97 – 105.
5. H. E. Jass, P. M. Elias, *Cosmet. Toiletries* **106** (1991) no. 10, 47 – 53.
6. G. Imokawa, H. Kuno, M. Kawai, *J. Invest. Dermatol.* **96** (1991) 845 – 851.
7. D. T. Downing, *Cosmet. Toiletries* **106** (1991) no. 12, 63 – 69.
8. F. E. Wall in M. S. Balsam et al. (eds.): *Cosmetics Science & Technology,* 2nd ed., **vol. 3,** J. Wiley & Sons, New York 1974, pp. 37 – 161.

9. J. M. Nikitakis, G. N. McEwen, J. A. Wenninger (eds.): *International Cosmetic Ingredient Dictionary,* 4th ed., The Cosmetic, Toiletry and Fragrance Association, Washington, DC, 1991.
10. C. Fox, *Cosmet. Toiletries* **108** (1993) no. 8, 47 – 65.
11. T. Mitsui, Y. Kanda, *Cosmet. Toiletries* **106** (1991) no. 9, 73 – 81.
12. P. A. Geis, *Dev. Ind. Microbiol.* **29** (1988) 305 – 315.
13. J. J. Kabara: "Cosmetic and Drug Preservation, Principles and Practice," in E. Jungermann (ed): *Cosmetic Science and Technology Series,* Marcel Dekker, New York 1984.
14. H. Stone, J. L. Sidel, *Cosmet. Toiletries* **101** (1986) no. 3, 45 – 50.
15. T. M. Kajs, V. Garstein, *J. Soc. Cosmet. Chem.* **42** (1991) no. 4, 249 – 271.
16. V. Rogiers, M. P. Derde, G. Verleye, D. Roseeuw, *Cosmet. Toiletries* **105** (1990) no. 10, 73 – 82.
17. U. Hoppe, G. Sauermann, R. Lunderstädt, *J. Soc. Cosmet. Chem.* **36** (1985) 105 – 123.
18. G. E. Nilsson, *Med. Biol. Eng. Comput.* **15** (1977) 209 – 218.
19. J. Pinnagoda, R. A. Tupker, T. Agner, J. Serup, *Contact Dermatitis* **22** (1990) 164 – 178.
20. P. H. List: *Arzneiformenlehre,* 4th ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1985.
21. A. K. Reng, J. M. Quack: "Rezeptierung kosmetischer Emulsionen mit speziellen anwendungstechnischen und physikalischen Eigenschaften," *Parfüm. Kosmet.* **66** (1985) no. 3, 139 ff.
22. G. M. Eccelston in J. Swarbrick, J. C. Boylan (eds.): *Encyclopedia of Pharmaceutical Technology,* **vol. 5,** Marcel Dekker, New York 1992, pp. 137 – 188
23. P. Walstra in P. Becher (ed.): *Encyclopedia of Emulsion Technology,* **vol. 1,** Marcel Dekker, New York 1983, pp. 57 – 127.
24. S. E. Friberg, M. L. Hilton, L. B. Goldsmith, *Cosmet. Toiletries* **102** (1987) no. 2, 87 – 98.
25. M. M. Breuer in P. Becher (ed.): *Encyclopedia of Emulsion Technology,* **vol. 3,** Marcel Dekker, New York 1983, pp. 385 – 424.
26. A. Martin, J. Swarbrick, A. Cammarata: *Physikalische Pharmazie, Methoden der Arzneibereitung – Ihre physikalischen und physikalisch-chemischen Grundlagen,* 2nd ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1980.
27. R. G. Harry (ed.): *Harry's Cosmeticology,* 7th ed., Leonard Hill Books, London 1982, pp. 50 – 73.
28. H. N. Bhargava, *Drug Dev. Ind. Pharm.* **13** (1987) no. 13, 2363 – 2387.
29. W. C. Griffin, *J. Soc. Cosmet. Chem.* **1** (1949) 311 –326.
30. W. C. Griffin, *J. Soc. Cosmet. Chem.* **5** (1954) 249 –256.
31. K. Shinoda, H. Kunieda in P. Becher (ed.): *Encyclopedia of Emulsion Technology,* **vol. 1,** Marcel Dekker, New York 1983, pp. 337 – 367.
32. M. W. Steed in R. G. Harry (ed.): *Harry's Cosmeticology,* 6th ed., Leonard Hill Books, London 1973, pp. 691 – 726.
33. W. C. Griffin in M. S. Balsam, E. Sagarin (eds.): *Cosmetic Science and Technology,* 2nd ed., **vol. 3,** Wiley Interscience, New York 1974, pp. 573 – 629.
34. G. Padberg, *J. Soc. Cosmet. Chem.* **23** (1972) 271 – 279
35. C. Fox, *Cosmet. Toiletries* **102** (1987) no. 6, 27 – 48.
36. P. Mayer, W. Pittermall, S. Wallat, *Cosmet. Toiletries* **108** (1993) no. 2, 99 – 109.
37. O. Braum-Falco, H. C. Korting, H. J. Maibach (eds.): *Liposome Dermatics,* Springer Verlag, Berlin 1992.
38. A. Huc, C. Bufferaut, E. Perrier, *Seifen, Öle, Fette, Wachse* **118** (1992) no. 18, 1147 – 1149.
39. T. Förster, W. von Rybinski, A. Wadle, *Adv. Colloid Interface Sci.* **58** (1995) 119 – 149.
40. M. Kahlweit, R. Strey, *Angew. Chem.* **97** (1985) 655 – 669.
41. M. Kahlweit, *Tenside Surf. Det.* **30** (1992) 83 – 89.
42. T. Förster, F. Schambil, H. Tesmann, *Int. J. Cosmet. Sci.* **12** (1990) 217 – 227.
43. A. Wadle, T. Förster, W. von Rybinski, *Colloids and Surf. A* **76** (1993) 51 – 57.
44. R. Hofmann, T. Förster, W. von Rybinski, A. Wadle, *Progr. Colloid Polym. Sci.* **98** (1995) 106 – 110.
45. F. Schambil, F. Jost, M. J. Schwuger, *Progr. Colloid Polym. Sci.* **73** (1987) 37 – 47.
46. T. Förster, F. Schambil, W. von Rybinski, *J. Disp. Sci. Technol.* **13** (1992) no. 2, 183 – 193.
47. T. Förster, H. Tesmann, *Cosmetics & Toil.* **106** (1991) 49 – 52.
48. T. Förster, W. von Rybinski, H. Tesmann, *Int. J. Cosm. Sci.* **16** (1994) 84 – 92.
49. A. Ansmann, R. Kawa, E. Prat, A. Wadle, *SÖFW* **120** (1994) 158 – 161.
50. A. Wadle, A. Ansmann, B. Jackwerth, H. Tesmann, *Parfuem. Kosmet.* **77** (1996) 250 – 254.

51. H. Stone, J. L. Sidel, *Cosmet. Toiletries* **101** (1986) no. 3, 45 – 50.
52. G. V. Civilli, C. A. Dus, *Cosmet. Toiletries* **106** (1991) no. 5, 83 – 88.
53. P. G. I. Lauffer in R. G. Harry (ed.): *Harry's Cosmeticology,* 7th ed., Leonard Hill Books, London 1982, pp. 757 – 799.
54. L. Träger: *Chemie in der Kosmetik,* 2nd ed., Hüthig Verlag, Heidelberg 1989, pp. 46.
55. R. Y. Lochhead: "Encyclopedia of Polymers and Thickeners for Cometics," *Cosmet. Toiletries,* **103** (1988) 12.
56. B. W. Müller: "Kieselsäure und Silicate", in *Hagers Handbuch der Pharmazeutischen Praxis,* 4th ed., vol. VII B, Springer Verlag, Berlin 1977, pp. 232 ff.
57. Chr. Verder: "Betonit – flüssige Stützwand aus Kristallen," *Bild Wiss.* **12** (1975) 50.
58. U. Hofmann: "Aus der Chemie der hochquellfähigen Tone (Bentonite)," *Angew. Chem.* **68** (1956) 53.
59. L. Grosse: "Cellulose und Cellulosederivate," *Hagers Handbuch der Pharmazeutischen Praxis,* 4th ed., vol. VII B, Springer Verlag, Berlin 1977, pp. 54 ff.
60. P. H. List: "Polyacrylate," *Hagers Handbuch der Pharmazeutischen Praxis,* 4th ed., **vol. VII B,** Springer Verlag, Berlin 1977, pp. 398 ff.
61. K. Schöffling: "Glasklar gequollen," (Hydrogele und hydrophile Gelbildner), *PTA Heute* **4** (1990).
62. K.-D. Bremecker, K. Seidel, A. Böhner: "Polyacrylatgele," *Dtsch. Apoth. Z.* **130** (1990) no. 8.
63. H. Drs. Meijer: "Tenside und Wasser," *Seifen, Öle, Fette, Wachse* **116** (1990) no. 7.
64. Prof. Dr. E. Nürnberg, N. Pollinger: "Transparente Tensidgele," *Tenside Deterg.* **23** (1986) 1.
65. K. Schrader: *Grundlagen und Rezepturen der Kosmetika,* 2nd ed., Hüthig Verlag, Heidelberg 1989.
66. H. Fey, J. Otte: *Wörterbuch der Kosmetik,* 3rd ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1991.
67. R. L. Diamond, W. Montagna, *Arch. Dermatol.* **112** (1976) 1235 – 1244.
68. C. Zugerman, *Cutis* **38** (1986) 116 – 120.
69. S. Romm, *Clin. Plast. Surg.* **11** (1984) no. 4, 571 – 581.
70. A. C. Deweck, *Cosmet. Toiletries* **96** (1981) no. 1, 29 – 32.
71. M. G. de Navarre (ed.): *The Chemistry and Manufacture of Cosmetics,* 2nd ed., **vol. 4,** Continental Press, Orlando, Fla., 1975, pp. 767 – 840.
72. M. S. Balsam, E. Sagarin (eds.): *Cosmetic Science and Technology,* 2nd ed., **vol. 1,** Wiley-Interscience, New York 1974, pp. 365 ff.
73. R. G. Harry (ed.): *Harry's Cosmeticology,* 6th ed., Leonard Hill Books, London 1973, pp. 140 – 156.
74. Shiseido Co. Ltd., US 3 957 969, 1976 (Y. Fujiyama, Y. Kenda, M. Matsuda).
75. A. C. Dweck, C. A. Burnham, *Int. J. Cosmet. Sci.* **2** (1980) 143 – 173.
76. R. G. Harry (ed.): *Harry's Cosmeticology,* 6th ed., Leonard Hill Books, London 1973, pp. 609 – 623.
77. Y. A. Shchipunov, E. V. Shumilina, *Progr. Colloid Polym. Sci.* **106** (1997) 228 – 231.
78. P. Schurtenberger, Q. Peng, M. E. Leser, P. L. Luisi, *J. Colloid Interface Sci.* **156** (1993) 43 – 51.
79. F. Dreher, P. Walde, P. L. Luisi, P. Elsner, *Skin Pharmacol.* **9** (1996) 124 – 129.
80. H. Kunieda, Y. Fukui, H. Uchiyama, C. Solans, *Langmuir* **12** (1996) 2136 – 2140.
81. C. Solans, I. Carrera, R. Pons, P. Erra, C. Azemar, *Cosmetics & Toilettries* **108** (1993) 61 – 64.
82. R. Pons, I. Carrera, P. Erra, H. Kunieda, C. Solans, *Colloids and Surf. A* **91** (1994) 259 – 266.
83. J. M. Seddon, N. Zeb, R. H. Templer, R. N. McElhaney, D. A. Mannock, *Langmuir* **12** (1996) 5250 – 5253.
84. K. Fontell, *Colloid Polym. Sci.* **268** (1990) 264 – 285.
85. C. Malcolmson, M. J. Lawrence, *Colloids and Surfaces B: Biointerfaces* **4** (1995) 97 – 109.
86. A. Kabalnov, B. Lindman, U. Olsson, L. Piculell, K. Thuresson, H. Wennerström, *Colloid Polym. Sci.* **274** (1996) 297 – 308.
87. C. Quellet, H. F. Eicke, *Chimia* **40** (1986) 233 – 238.
88. E. Nürnberg, W. Pohler, *Progr Colloid & Polymer Sci.* **69** (1984) 48 – 55.
89. E. Nürnberg, W. Pohler, *Progr Colloid & Polymer Sci.* **69** (1984) 64 – 72.
90. C. Fox, *Cosmetics & Toilettries* **111** (1996) 53 – 69.
91. C. Fox, *Cosmetics & Toilettries* **112** (1997) 31 – 48.
92. H. P. Fiedler et al. (eds.): *Blaue Liste – Inhaltsstoffe kosmetischer Mittel,* 2nd ed., Editio Cantor, Aulendorf 1992.
93. "Proposed Names Lists," *International Nonproprietary Names for Pharmaceutical Substances,* World Health Organization, Geneva 1992, nos. 1 – 66.

94. R. Schulze, *Parfüm. Kosmet.* **37** (1956) 310, 365.
95. K. L. Gabriel et al., *J. Toxicol. Cutan. Ocular. Toxicol.* **6** (1987) 357 – 370.
96. J. J. P. Morton, E. G. Murphy: "Sun Protection Factors: Regulatory Issues and Methods of Determination" in E. M. Jackson (ed.): *Photobiology of the Skin and Eye,* Marcel Dekker, New York 1986, pp. 85 – 104.
97. H. Ippen, *Medizinische* 1958, 200 – 203.
98. I. B. Fitzpatrick et al. (eds.): *Dermatology in General Medicine,* 3rd ed., McGraw-Hill, New York 1987, p. 1509.
99. K. H. Kaidbey, A. Barnes, *J. Am. Acad. Derm.* **25** (1991) 262 – 266.
100. G. A. Erlemann et al., *Parfum. Kosmet.* **54** (1973) 263.
101. E. Charlet, P. Finkel, *Aerztl. Kosmetol.* **9** (1979) 368.
102. H. Ippen, U. Perschmann, *Arch. Derm. Forsch.* **236** (1970) 207 – 216.
103. M. O. Masse, M. Herpol-Borremans, *Int. J. Cosm. Sci.* **13** (1991) 303 – 316.
104. R. Matissek, *Z. Lebensm.-Unters. Forsch.* **179** (1984) 279 – 287, 432 – 439.
105. F. P. Tomasella, P. Zuting, L. J. Love, *J. Chromatogr.* **587** (1991) 325 – 328.
106. L. Y. Matsuoka, J. Wortsman, B. W. Hollis, *J. Am. Acad. Derm.* **22** (1990) 772 – 775.
107. A. A. Skolnick, *J. Am. Med. Ass.* **265** (1991) 3218 – 3220.
108. M. C. van Praag et al., *J. Invest. Dermatol.* **97** (1991) 629 – 633.
109. V. E. Reeve, M. Bosnic, C. Boehm-Wolcox, R. D. Ley, *J. Invest. Dermatol.* **97** (1991) 624 – 628.
110. S. Schauder, *Zschr. Hautkr.* **66** (1991) 294 – 318.
111. S. Schauder: *Göttinger Liste lichtfilterhaltiger Hautpflegepräparate in Deutschland,* Berliner Med. Verlagsanstalt, Berlin 1991.
112. J. T. Dorlakoglu, R. Kickuth, *Environm. Contaminat. Toxicol.* **37** (1986) 866 – 873.
113. G. M. Eccleston, *Cosmet. Toiletries* **101** (1986) 73 –92.
114. T. de Vringer, J. G. H. Joosten, H. E. Junginger, *Colloid & Polymer Sci.* **265** (1987) 448 – 457.
115. L. M. Greenstein, R. A. Bolomy, *J. Soc. Cosmet. Chem.* **22** (1971) 161 – 177.
116. E. Nürnberg, N. Pollinger, *Tenside Deterg.* **23** (1986) 25 – 35.
117. P. Finkle, H. I. Draper, J. H. Hildebrand, *J. Am. Chem. Soc.* **45** (1923) 2780 ff.
118. E. R. Garrett, *J. Soc. Cosmet. Chem.* **21** (1970) 393 – 415.
119. R. D. Vold, R. C. Groot, *J. Phys. Chem.* **66** (1962) 1966 ff.
120. W. Cavigalli, *Kosmetika (Zürich)* **5** (1977) 17 ff.
121. J. B. Ward, J. F. Kinney, H. Y. Soad, *J. Soc. Cosmet. Chem.* **25** (1974) 437 ff.
122. P. Becher (ed.): *Encyclopedia of Emulsion Technology,* vol. 1, Marcel Dekker, New York 1983.
123. W. F. Hemminger, H. K. Cammenga: *Methoden der Thermischen Analyse,* Springer Verlag, Berlin 1986.
124. T. A. Iranloye, *Int. Lab.* 1985, Nov./Dec., 68 – 71.
125. G. W. Jo Lee, M. J. Groves, *Powder Technol.* **28** (1981) 49 – 54.
126. J. F. Trumbetas, J. A. Fisriti, R. J. Sims, *J. Am. Oil Chem. Soc.* **54** (1977) no. 10, 433 – 434.
127. M. Frenkel, R. Schwartz, N. Garti, *J. Dispersion Sci. Technol.* **3** (1982) 2, 195 – 207.
128. M. Fender, O. Tannert, *Seifen, Öle, Fette, Wachse* **14** (1989) 459 – 461.
129. M. Mrukot, M. Schmidt, *Parfum Kosmet.* **12** (1976) 337 – 343.
130. F. Schambil, S. Paasch, R. M. Sax, *Seifen, Öle, Fette, Wachse* **15** (1989) 515 – 521.
131. R. M. Sax, G. Schön, S. Paasch, M. J. Schwuger, *Prog. Colloid Polym. Sci.* **77** (1988) 109 – 114.
132. D. J. Miller, A. K. Reng, W. Skrypzak, *Int. J. Cosmet. Sci.* **10** (1988) 161 – 170.
133. K. H. Wallhäuser, *Parfum. Kosmet.* **6** (1981) 173 –181.
134. D. S. Orth, *Cosmet. Toiletries* **96** (1981) 43 – 52.
135. G. K. Muberry, D. R. Entrup, J. R. Agin, *Cosmet. Toiletries* **102** (1987) 47 – 54.
136. D. S. Orth, *Cosmet. Toiletries* **104** (1989) 49 –64.
137. Gesetz zum Schutz vor gefährlichen Stoffen (Chemikaliengesetz) from Sept. 16, 1980 (BG-Bl. I) pp. 1718 – 1728.
138. Verordnung über kosmetische Mittel (Kosmetikverordnung) from June 29, 1985 (BG-Bl. I) pp. 1082 – 1121.
139. Gesetz über den Verkehr mit Lebensmitteln, Tabakerzeugnissen, kosmetischen Mitteln und sonstigen Bedarfsgegenständen (Lebensmittel- und Bedarfsgegenständegesetz), Bundesgesetzblatt from Aug. 15, 1974, pp. 1946 – 1966.
140. Council Directive of July 27, 1976 on the Approximation of the Laws of the Member States relating to Cosmetic Products, (76/768/EEC), Official Journal of the

European Communities no. L 262, 27. 9. 1976, pp. 169 – 200.
141. Council of Europe: "Cosmetic Products and their Ingredients," 2nd ed., [Dok. P-SG (84)43], Strasbourg 1984.
142. Verordnung über gefährliche Stoffe (Gefahrenstoffverordnung) from Aug. 26, 1986 (BGBl.) pp. 1470 – 1487.
143. Empfehlungen des Bundesgesundheitsamtes zur Prüfung der gesundheitlichen Unbedenklichkeit von kosmetischen Mitteln, Bundesgesundheitsblatt **24** (1981) pp. 96 – 98.
144. Wissenschaftlicher Ausschuß für Kosmetologie der EG-Kommission, Leitlinien für Toxizitätstests von Inhaltsstoffen kosmetischer Mittel, Brüssel 1983.
145. M. Balls et al.: ECVAM Workshop 7," *ATLA* **23** (1995) 398 – 409.
146. H. Spielmann et al., *ATLA* **24** (1996) 741 – 858.
147. W. J. W. Pape et al., *Toxicol. In Vitro* **13** (1999) 343 – 354.
148. M. Balls et al., *ATLA Altern. Lab. Anim.* **18** (1990) 313 – 336.
149. W. J. W. Pape, U. Hoppe, *Skin Pharmacol.* **4** (1991) 205 – 212.
150. W. J. W. Pape et al., in: *Altern. Toxicol.* **10** (1994) 203 – 212.
151. M. Balls et al., *Toxicol. In Vitro* **9** (1995) 871 – 929.
152. P. G. Brantom et al., *Toxicol. In Vitro* **11** (1997) 141 – 179.
153. H. Spielmann et al., *Toxicol. In Vitro* **12** (1998) 305 – 327.
154. H. Spielmann et al., *ATLA* **26** (1998) 679 – 708.
155. N. P. Lüpke, *Food Chem. Toxicol.* **23** (1985) 287 – 291.
156. S. D. Gettings et al., *In Vitro Toxicology* **4** (1991) 247 – 288.
157. Commission Directive 2000/33/EC.
158. SCCNFP/0103/99 Final, 20 January 1999.
159. H. Schäfer, T. E. Redelmeier: *Skin Barrier,* Karger, Basel 1996.
160. FDA, Fair Packaging and Labeling Act (FPLA). CTFA Labeling Manual, CTFA, Washington 1982.
161. Nippon Keshohin genryoshu (Japan Cosmetic Ingredient Dictionary), Yakuji Nippo, Tokyo 1989.

Smectites → Clays
Snail Control → Molluscicides
SNG → Gas Production

Soaps

Cleansing Agents, Laundry Detergents, and Metallic Soaps are separate keywords.

KLAUS SCHUMANN, Henkel KGaA, Düsseldorf, Federal Republic of Germany

KURT SIEKMANN, Henkel KGaA, Düsseldorf, Federal Republic of Germany

1. Introduction 1
2. Historical Aspects 1
3. Raw Materials 3
4. Production 4
4.1. Basic Principles 4
4.1.1. Saponification of Neutral Oils 5
4.1.2. Saponification of Fatty Acids 6
4.1.3. Saponification of Fatty Acid Methyl Esters 6
4.2. Industrial Production of Soap Bases 7
4.2.1. Kettle Saponification 7
4.2.2. Continuous Saponification Processes 7
4.2.3. Batch Processes 11
4.3. Drying 12
4.4. Storage 15
4.5. The Finishing of Soap Base 15
5. Characteristic Features of the Most Common Types of Soap 17
5.1. Toilet Soaps 17
5.2. Other Soap Products 17
6. Syndets 18
7. Environmental Protection 19
8. Quality Control and Analysis 20
9. Economic Aspects 21
10. Toxicology 21
11. References 22

1. Introduction

In everyday language, soaps are sodium and potassium salts of fatty acids: saturated or unsaturated aliphatic carboxylic acids, most of which occur in nature (→ Carboxylic Acids, Aliphatic; → Fatty Acids). The acids almost always have unbranched carbon chains ($>C_8$), exhibit surface-active properties in the form of their alkali salts in aqueous solution, and constitute classical anionic surfactants (→ Surfactants). Salts of the same acids with other metals are called metallic soaps. These possess considerably different physicochemical and industrial properties and, for this reason, are discussed separately (→ Metallic Soaps).

2. Historical Aspects

The history of soap begins with our forefathers learning to prepare a type of washing paste by melting fat together with wood ashes. An early recipe for soft soap is recorded on clay tablets from Mesopotamia, the region between the Euphrates and Tigris, dating to ca. 2500 B.C.

In 79 A.D. PLINY the Elder asserted in his *Historia Naturalis* that soap was invented by the Gauls, but the Gauls most probably acquired the recipe from the Egyptians by way of the Phoenicians, because soap was already known in Egypt (as *Andjir*) in Coptic times, during the first century A.D. Teutonic peoples later adopted this soap product from the Gauls and improved upon it continuously. According to PLINY, "sapo" was prepared both as a liquid potassium-soap product and in a solid form. Especially high quality soaps were obtained from goat tallow and wood ashes.

Teutonic soap was a popular export article, and was used in Rome also as a hair dye and pomade. Soap from Germania was highly praised by GALEN, the personal physician of Emperor MARCUS AURELIUS, in the second century A.D. He was apparently the first to recognize the value of soap as a washing and cleansing agent. However, GALEN's propaganda success was limited. While the men took his advice, Roman ladies were sceptical, and continued to wash with plain water. In the third century, the Roman Emperor SEPTIMIUS SEVERUS had greater psychological success in promoting soap. He followed the counsel of his famous doctor SERENUS SAMMONICUS, advising women around him to cleanse their faces with crushed herbs dissolved in oil as a way of making themselves more at-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_247

tractive to men. These examples show, incidentally, that even in early history cosmetics played a significant role in peoples' lives.

In 385 A.D. THEODOROS PRISCIANUS recommended using soap for the washing of hair. He also provides the first record of the "saponarius", or soap-maker.

Domestic soap-making was broadened to include active trade in soap toward the end of the eighth century under the emperor CHARLEMAGNE, and soap-making became a recognized occupation within the imperial domains. During the ninth century, the Spanish spread the art of soap-making into the lands surrounding the Mediterranean, especially in the areas of Marseilles, Venice, Genoa, and Alicante.

In the fourteenth century, German soap-makers united to form their own guilds (Augsburg, 1324). The soap-maker's trade was frequently combined with that of the candlemaker. A soap-maker's apprentice spent three years in training before being allowed to take the apprentices' final examination. The premium for an apprenticeship was 100 Thaler. If this premium could not be paid, the apprenticeship period was increased to six years. Before taking the master's examination, journeyman soap-makers were required to perfect their skills by traveling from city to city, finally preparing their own batch of soap. The composition of the soap was deemed worthy of a master craftsman if it took on a speckled appearance in the kettle.

As early as the fifteenth century, Portuguese and Spanish missionaries introduced soap into Japan. Presented as a precious gift to the rulers, it was hoped that the soap would invoke a sympathetic response toward missionary work in their territories [1].

The art of soap-making reached its high point in the period before the Thirty Years' War, but the general spread of soap as a commodity occurred only later. For a long time the use of soap remained restricted, because it was regarded as a luxury item. Only after the introduction of large-scale soda production (by the Le Blanc and Solvay processes) and widespread import of fats and oils from the tropics was the foundation laid for the industrial production of soap, permitting it to become a common consumer article.

Although soap-making has not been practiced as a skilled trade for some time, fine soap today is more varied and inexpensive than ever before. Indeed, in the technically advanced twentieth century it is impossible to imagine everyday life in the absence of soap. Even at the beginning of the twenty-first century, the future of the "good old soap" seems quite secure.

A wide assortment of soap products has been developed during the course of time. Based on use and appearance, it is possible to distinguish the following product groups:

1) Toilet soaps (including special soaps)
2) Hard soaps, soap flakes
3) Shaving soaps or creams
4) Soft soaps
5) Liquid soaps

Twenty years ago, hard soaps, soap flakes, and soft soaps were still very important in the areas of textile washing and general household cleansing. Today, their place is frequently taken by modern powdered and liquid household detergents and cleansers (→ Laundry Detergents), and the use of liquid soaps in dispensers remains rather restricted.

By contrast, toilet and fine soaps have maintained an undisputed place in the market. Cream or cosmetic soaps, luxury soaps, deodorant soaps, children's soaps, skin-protecting soaps, and special medicinal soaps are all included in this category. All are derived from a high-quality soap base which, by the addition of various active ingredients, can be endowed with specific grooming, cosmetic, or deodorizing properties. Indeed, almost unlimited variations with respect to form, color, and fragrance satisfy almost every imaginable consumer desire [2].

3. Raw Materials

Numerous naturally occurring plant and animal fats and oils can be used as the starting materials in soap-making, especially beef tallow, palm oil, coconut oil, and palm kernel oil. Fatty acid compositions for these neutral fats are given in Table 1. Peanut oil and olive oil as well as lard are also utilized to a lesser extent. Table 2 provides typical carbon chain-length distributions for fatty acids in soap itself. The fatty acid composition of the soap-making materials varies depending on the intended purpose of the soap. Normally, 75 – 85 % beef tallow and 15 – 25 %

coconut oil constitute the main components in soap-making. C_{12} and C_{14} soaps lather quickly, but they produce an unstable, coarse-bubble foam. On the other hand, C_{16} and C_{18} soaps lather rather slowly, but lead to a stable, fine-bubble foam. Apart from fatty acid composition, the quality of a soap is also a function of the purity of the raw materials, their physical and chemical characteristics, and their color and odor.

Triglycerides are the most important starting materials for the making of soap (Fig. 1), because they are easily transported and are not subject to quality losses during transport. Free fatty acids may also serve as raw materials, but the distillates employed must be as fresh as possible. Such distillates are not always available in a suitable quality and at a reasonable price. Transport and storage factors play an important role in the soap industry. No general answer can be given to the question of whether triglycerides should be preferred over free fatty acids as starting materials for the production of soap, but recovery of the valuable byproduct glycerol must be a part of the consideration (see Section 4.1.).

The use of fatty acid methyl esters as soap-making materials has also been discussed, and recently put into practice as well (e.g., by the Lion Corp., Japan); like triglycerides, these substances can be transported and stored without any reduction in quality. As intermediates in the synthesis of fatty alcohols (→ Fatty Alcohols, Chap. 2.3.3.1.), alkanolamides, and ester sulfonates, these compounds are associated with substantial production capacities, which are constantly increasing. However, from a worldwide standpoint, such esters currently play only a minor part in the production of high-quality toilet soaps.

Sodium hydroxide is acquired primarily in the form of a 50 % solution from chlor-alkali electrolysis (→ Chlorine, Chap. 4.; Electrochemistry, Chap. 4.1.), as is potassium hydroxide. The soda sometimes required for the reaction with fatty acids is the usual anhydrous commercial product. Potassium carbonate is used in small amounts for the manufacture of soft soap, as is potassium chloride. In saponification, the ratio of caustic to fat (acid) is adjusted such that no more than a 0.05 % excess of caustic is present in the finished soap.

Apart from sodium hydroxide solution, sodium chloride (common salt, rock salt) is the most important inorganic auxiliary agent. To prevent the formation of unwanted greasy, insoluble, alkaline-earth soaps, the NaCl should contain minimal amounts of calcium and magnesium salts, a consideration that applies to potassium chloride as well.

Other additives and auxiliary agents are discussed elsewhere in the article.

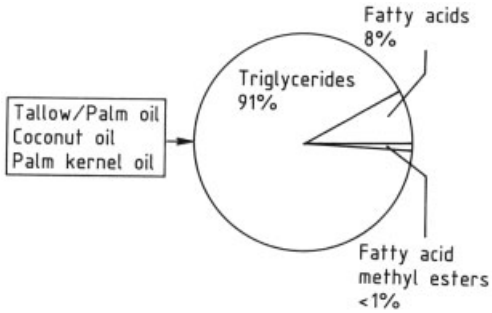

Figure 1. Raw materials utilized in soap-making [3]

4. Production

In principle, the production of soaps comprises the following steps:

1) Production of a suitable soap base
2) Mixing and homogenization of the soap base with additives, such as perfume oils, coloring matter, and skin-grooming substances (refatting agents)
3) Extrusion
4) Cutting to rough form
5) Shaping
6) Packaging

4.1. Basic Principles

In soap-making terminology, "saponification" is understood to mean simply the formation of soap. If a charge of suitable fats and oils is saponified with sodium hydroxide solution, the procedure is called "neutral oil saponification". If the charge consists of distilled fatty acids, the process is called "fatty acid saponification". The saponification of fatty acid methyl esters with sodium hydroxide solution is called "fatty acid

Table 1. Fatty acid compositions of the most important neutral fats used in soap-making (%)

	Coconut oil	Palm kernel oil	Palm oil	Beef tallow
Octanoic acid (caprylic acid)	6 – 9	3 – 10		
Decanoic acid (capric acid)	6 – 10	3 – 14		
Dodecanoic acid (lauric acid)	44 – 51	37 – 52		
Tetradecanoic acid (myristic acid)	13 – 18	7 – 17	1 – 6	2 – 8
Hexadecanoic acid (palmitic acid)	8 – 10	2 – 9	32 – 47	24 – 33
Octadecanoic acid (stearic acid)	1 – 3	1 – 3	2 – 8	14 – 29
Octadecenoic acid (oleic acid)	6 – 8	11 – 23	40 – 52	39 – 50
Octadecadienoic acid (linoleic acid)	1 – 3	1 – 3	5 – 11	1 – 5

Table 2. Typical carbon chain-length distribution of fatty acids in soaps (%)

Chain length	%	Chain length	%
C_8	0.5	C_{16}	24.6
C_{10}	0.5	C_{18}	5.6
C_{12}	7.3	C_{20}	56.5
C_{14}	5.0		

methyl ester saponification". The individual processes can be represented by the following equations:

1) Neutral oil saponification

$$\begin{array}{l} CH_2OCOR^1 \\ | \\ CHOCOR^2 \\ | \\ CH_2OCOR^3 \end{array} + 3\,NaOH \longrightarrow \begin{array}{c} R^1CO_2Na \\ + \\ R^2CO_2Na \\ + \\ R^3CO_2Na \end{array} + \begin{array}{l} CH_2OH \\ | \\ CHOH \\ | \\ CH_2OH \end{array}$$

Neutral oil or neutral fat — Lye — Soap — Glycerol

R = fatty acid alkyl group

R = fatty acid alkyl group

2) Fatty acid saponification

a) Fatty acid production by steam cleavage

$$\begin{array}{l} CH_2OCOR^1 \\ | \\ CHOCOR^2 \\ | \\ CH_2OCOR^3 \end{array} + 3\,H_2O \longrightarrow \begin{array}{c} R^1CO_2H \\ + \\ R^2CO_2H \\ + \\ R^3CO_2H \end{array} + \begin{array}{l} CH_2OH \\ | \\ CHOH \\ | \\ CH_2OH \end{array}$$

Neutral oil or neutral fat — Fatty acid

b) Fatty acid saponification

$$RCO_2H + NaOH \longrightarrow RCO_2Na + H_2O$$

Fatty acid — Soap

3) Fatty acid methyl ester saponification

a) Transesterification

$$\begin{array}{l} CH_2OCOR^1 \\ | \\ CHOCOR^2 \\ | \\ CH_2OCOR^3 \end{array} + 3\,CH_3OH \longrightarrow \begin{array}{c} R^1CO_2CH_3 \\ + \\ R^2CO_2CH_3 \\ + \\ R^3CO_2CH_3 \end{array} + \begin{array}{l} CH_2OH \\ | \\ CHOH \\ | \\ CH_2OH \end{array}$$

Neutral oil or neutral fat — Methanol — Fatty acid methyl ester

b) Saponification

$$RCO_2CH_3 + NaOH \longrightarrow RCO_2Na + CH_3OH$$

Fatty acid methyl ester — Lye — Soap

Neutral oil and fatty acid saponifications have been applied successfully for decades, and lead to products of equivalent quality. This should not be surprising: the production of a perfect soap requires that the soap base be extremely pure and, above all, free of unpleasant side odors. In principle, it is irrelevant whether the necessary purification and washing operations are achieved by fat splitting and distillation of the split fatty acids or by saponification of the neutral fats to a soap base that is subsequently purified by repeated closing, salting out (opening), separation, and washing.

While the splitting of fats by steam with subsequent distillation of the fatty acids entails considerable expense, fatty acid saponification can be accomplished fairly easily with modern continuous processes. Moreover, processing the resulting glycerol water for the recovery of the

valuable glycerol is straightforward and environmentally friendly compared with the treatment of very salty lye from neutral oil saponification.

Both saponification processes can be conducted either discontinuously in steam-heated open kettles, analogous to the classical boiling process, or continuously in closed saponification systems. Saponification of fatty acid methyl esters is carried out in continuously operated, closed systems, because in this case economic and ecological factors require that methanol be recycled. Neutral oil saponification and fatty acid saponification give rise to soap bases with fatty acid contents $\leq$63 %.

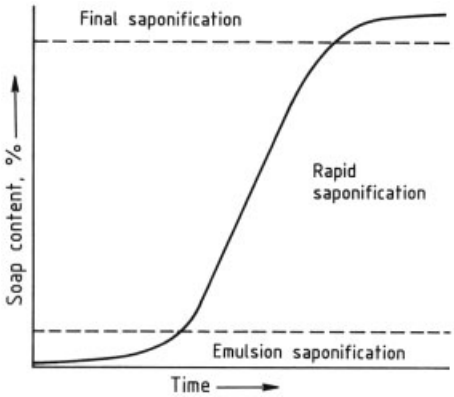

Figure 2. Soap formation from neutral fats, plotted as a function of time (adapted from [5])

4.1.1. Saponification of Neutral Oils

In the saponification of natural oils and fats with aqueous sodium hydroxide (e.g., in kettle saponification) the triglyceride and caustic soda reactants are neither mutually soluble nor miscible. Achieving a satisfactory reaction rate requires intensive mixing, which is accomplished by emulsification. In practice, small amounts of soap form very rapidly through neutralization of the free fatty acids that are always present in fats. This soap then acts as an emulsifier. In this process a specific limiting concentration of sodium hydroxide must never be exceeded. This limit corresponds to the sodium hydroxide concentration above which dissolved soap precipitates in the form of flakes [4], rendering it no longer available as an emulsifying agent. Two groups of fats can be distinguished with respect to their limiting sodium hydroxide concentrations:

1) Soft fats, with a limiting sodium hydroxide concentration of 14 – 18 % (e.g., coconut oil, palm kernel oil)
2) Hard fats, with a limiting sodium hydroxide concentration of 4 – 6 % (e.g., beef tallow)

As suggested by Figure 2, the saponification of neutral oils comprises the following three stages [5]:

1) *Emulsion Saponification*. The process of saponification proceeds slowly until formation of a fat emulsion or dispersion is complete.
2) *Rapid Saponification.* The reaction rate increases as more soap is formed until a freeflowing liquid soap paste is produced. Vigorous reaction continues while most of the fat is transformed into soap.
3) *Final Saponification*. This reaction occurs in a homogeneous, pulpy mass.

In this final stage the reaction rate is so low that extended boiling with excess sodium hydroxide solution is required to saponify the last traces of fat. The viscosity of the reaction mixture is reduced as necessary by adjusting the electrolyte concentration. In practice, this requires adding sodium chloride, either as a solid or in solution. However, this "reduction" of the paste must be carried out very carefully, because too much electrolyte leads to a separation into two phases, where the upper phase contains most of the soap and the lower phase almost all the electrolyte. The nigre thus floats on the lye phase below, a phenomenon known as "salting out".

The nigre can be purified by repeated salting out, separation of the lye, and renewed closing, thereby reducing the electrolyte concentration. In the final operation, "fitting", addition of water and electrolyte produces a three-phase equilibrium state: neat soap, nigre, and lye. The neat soap is further processed and the precipitated nigre is either added to the next batch to assist in emulsification or it is processed separately.

4.1.2. Saponification of Fatty Acids

If appropriate fatty acid distillates are used as raw materials instead of oils and fats, the saponification step corresponds to a neutralization.

Here again, the presence of soap as an emulsifier is absolutely essential. In fatty acid saponification, a soda solution can be used instead of sodium hydroxide solution.

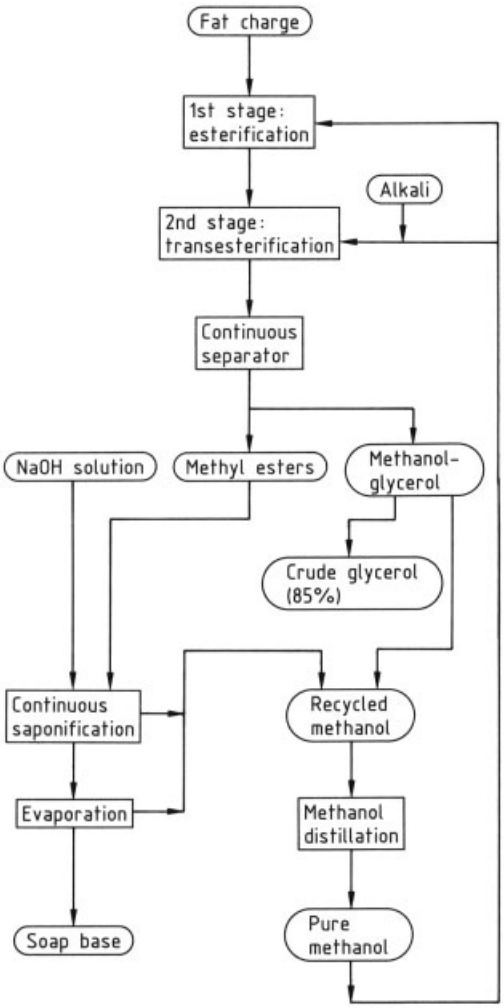

Figure 3. Flow diagam of the Lion E-S Process for the production of soap base

4.1.3. Saponification of Fatty Acid Methyl Esters

As mentioned previously, the saponification of fatty acid methyl esters must be carried out in a closed apparatus, because the methanol must be recycled.

Fatty acid methyl esters are obtained by transesterification of fats with methanol. The Lion E-S process, developed by the Lion Fat & Oil Co. in Tokyo, Japan [6], is illustrated schematically in Figure 3.

In the first step the fat charge is treated with methanol in the presence of an ion-exchange resin, resulting in esterification of the free fatty acids normally present. The second step is a transesterification with methanol in the presence of alkali, followed by recovery of excess methanol. The reaction mixture is then transferred to a continuously operated separator, where the phases are separated. The lighter phase consists of fatty acid methyl esters, and the heavier phase contains crude glycerol, some soap, and impurities from the oil. The methyl esters are saponified continuously with sodium hydroxide solution. The resulting crude soap is treated with steam to remove traces of methanol and then processed in the usual manner.

4.2. Industrial Production of Soap Bases

A generalized flow diagram for the industrial production of soap is provided in Figure 4. In principle, both batch and continuous processes can be used. Kettle saponification, a batch process, is described separately here for historical reasons.

4.2.1. Kettle Saponification

The classical boiling process with live steam in deep, open kettles or pans was for many decades the most important method of soap-making, and the process is still employed in some soap factories (Fig. 5). The pans are conical, cylindrical, or rectangular in shape. The most common type of pan is very wide, but relatively shallow in the cylindrical part, tapering to a deep cone. Inside the pan is a distributing ring introducing live steam, and sometimes a heating coil for indirect heating, together with a siphon and a swing pipe for removing the neat curd. The lye and nigre exit via a bottom valve at the point of the cone. The batch is mixed by the action of the steam.

Boiling kettles with stirrers have also proved useful. The cylindrical part of a kettle of this type is tall and narrow. Smaller kettles were occasionally equipped with “crutchers” in the form of perforated plates that could be moved rhythmically up and down through the liquid soap by means of a rod and crankshaft assembly. Stirrers and crutchers served only to intensify the

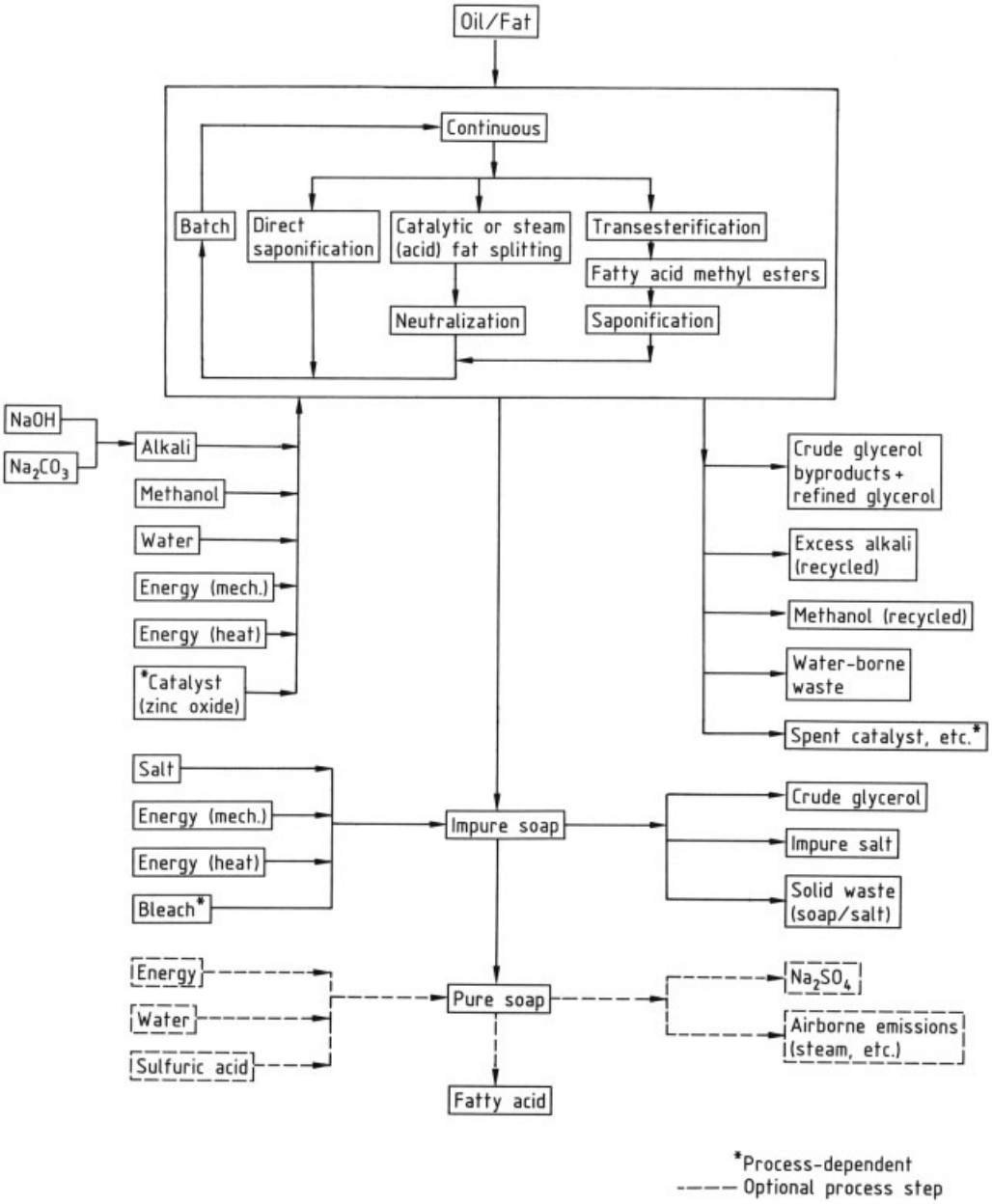

Figure 4. Generalized flow diagram for the production of soap base

mixing effect. Live steam was introduced here also to induce boiling. Boiling in open kettles is effective for the saponification of both neutral oils and fatty acids. The size of the charge varies considerably, ranging from 500 kg to more than 20 t. Large boiling kettles may have a volume $\geq 100\ m^3$.

4.2.2. Continuous Saponification Processes

Boiling in an open kettle is an elaborate process, and one that is not easy to carry out. The quality and uniformness of the brew is largely a function of the care, manual skill, and practical experience of the soap-maker. This situation was

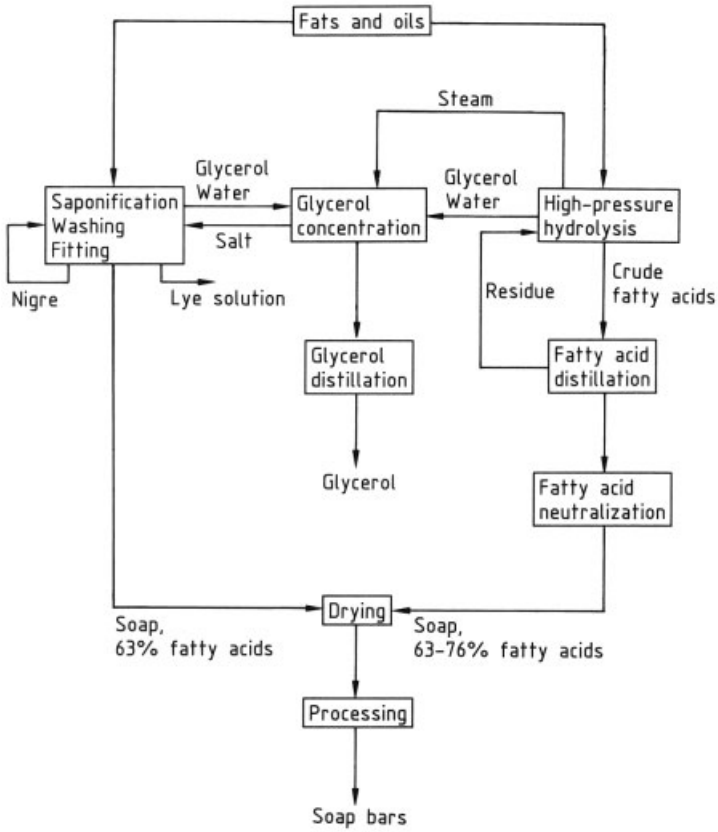

Figure 5. Flow diagram for kettle saponification [7]

recognized by experts decades ago. After World War II, closed continuous systems increasingly displaced conventional boiling processes, and highly developed process engineering and modern measuring and control technology have now completely changed the face of the soap industry. Nevertheless, replacement of the old soap kettles occurred only gradually, as industry apparently found it difficult to part with the traditional soap-maker's most important piece of equipment. Continuous production lines for processing soap bases into the usual cakes have been standard equipment in soap factories for many decades, though open boiling pans are still seen occasionally. A schematic representation of a Weber & Seeländer type DVN kettle saponification system is shown in Figure 6. It is impossible to describe in this article all the processes (roughly ten in number) available for continuous saponification, and the selection that appears is not meant to imply any value judgment. Every process offers advantages and disadvantages with respect to both soap production and environmental compatibility.

De Laval Process. The De Laval Centripure process illustrated in Figure 7 is an important continuous saponification method that involves centrifuges. Saponification itself is accomplished in a saponification column that functions as a loop reactor. Fat and caustic soda are fed into a circulating stream of liquid soap. To control the pressure, the liquid soap is withdrawn at the top of the column in an amount equal to that of newly formed soap, with fresh brine solution having previously been introduced. The principle of the loop reactor is illustrated in Figure 8. The overall system (Fig. 7) is controlled via the viscosity by measuring the pressure differentials between the suction and delivery sides of the circulating pumps (g). This surprisingly sensitive approach to constant composi-

tion (C – C) control relies on the fact that the viscosity of a soap solution depends on the electrolyte concentration. Soap is very viscous at low electrolyte concentrations, but the viscosity decreases with increasing amounts of electrolyte, passes through a minimum, and then increases again. Changes in viscosity are detected as pressure differences by a differential pressure-sensing instrument, which in turn controls the lye dosing pump. In this manner, alkali concentrations can be adjusted with an accuracy of 0.01 %.

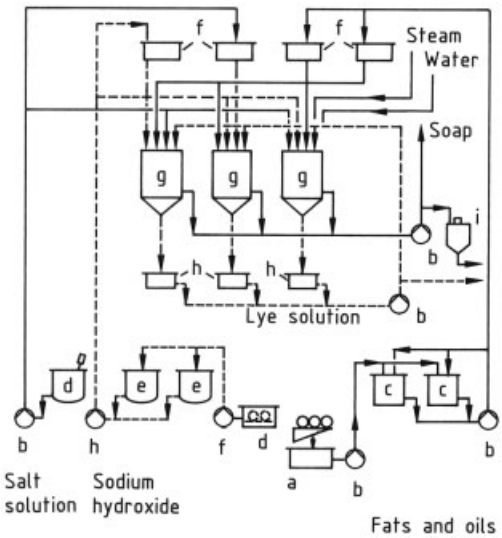

Figure 6. Kettle saponification (Weber & Seeländer, Type DVN)
a) Drum discharge facility; b) Pumps; c) Storage vessels; d) Dissolving vessel; e) Adjusting vessels; f)Dosing vessels; g) Boiling kettles; h) Lye vessels; i) Soap receiving vessel

The saponification column is suitable for both neutral oil and fatty acid saponification. Use of ca. 50 % caustic soda for saponification delivers soap with a 73 % fatty acid content to two centrifugal extractors (c) connected in series, where glycerol is washed out and curd soap is separated. In the first washing stage, lye from the second stage is added to the concentrated soap, which withdraws sufficient water from the wash lye to yield a 63 % curd soap. The lye flow from the first washing stage is diverted to glycerol recovery. In the second stage of extraction the curd soap is washed with fresh electrolyte solution together with nigre from the separators of the fitting stage (e). As noted previously, lye from the second stage flows back to the first extraction stage, which means that soap and electrolyte solution move counter to each other. The fitting step takes place in a second column (d) that receives the curd soap flow from the second wash stage. Curd and nigre are isolated in the subsequent separators (e). Separated nigre is returned together with fresh electrolyte solution to the second wash stage, while the curd is transferred to a drying facility. Similar to kettle saponification, the De Laval process usually produces a 63 % soap base. (All reported soap concentrations refer to the fatty acid content.) However, if the charge consists of fatty acids, the elaborate centrifugal steps become unnecessary and the process can be controlled in such a way (at 135 °C and a pressure of 3.5 bar) that a high-concentration 73 – 74 % soap is formed directly. The drying of such a soap is correspondingly less energy-intensive (see Section 4.3).

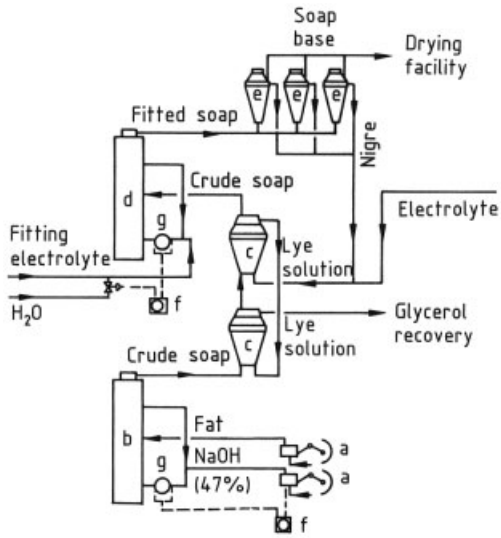

Figure 7. Schematic representation of the De Laval Centripure process
a) Dosing pumps; b) Saponification column; c) Extractors; d) Fitting column; e) Separators; f) "Constant composition" controls; g) Circulating pumps

Meccaniche Moderne Process. The heart of this process (Fig. 9) is a four-compartment autoclave (a) in which the main saponification reaction occurs at ca. 120 °C and 2 bar. The resulting soap is recycled once and then led into a mixer – cooler (c), which also consists of four compartments, each with an axially arranged

mixer. Here the saponification is completed. The first two compartments of the mixer – cooler are cooled with water, reducing the temperature of the reaction mixture to 90 °C. This decreases the solubility of the soap in the lye; two phases form (soap and lye), which are subsequently separated in a gravity separator. The process is used for the saponification of both fats and oils [8].

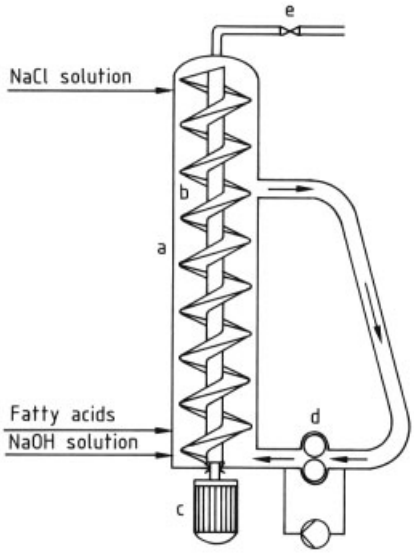

Figure 8. Loop reactor for the De Laval process a) Saponification column (loop reactor); b) Screw; c) Screw drive; d) Circulating pump; e) Reducing valve

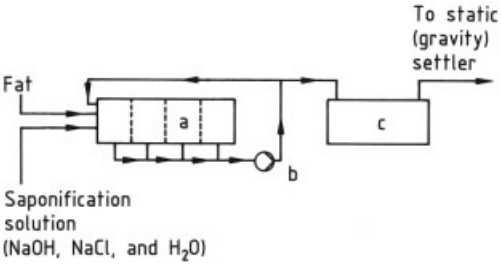

Figure 9. Saponification system used in the Meccaniche Moderne process [8]
a) Four-compartment autoclave; b) Circulating pump; c) Mixer – cooler

Ballestra Continuous Process for the Saponification of Fatty Acid Methyl Esters. This process is based on four steps:

1) Pretreatment of oils and fats
2) Production of fatty acid methyl esters
3) Saponification of the esters
4) Glycerol recovery

Further details and a schematic representation are provided in [9].

Konti Saponification, (Weber & Seeländer, Type KVN). The continuous Sela saponification process for preparing household or toilet soap from neutral fats or neutral oils is illustrated schematically in Figure 10.

The process begins in levelling vessels (a) that are charged from raw material storage tanks, as well as levelling in separate vessels for recycled, glycerol-enriched lye. Control of the four process steps is accomplished through multiple-head dosing pumps (b), (c), and (h), where each dosing component can be individually controlled.

Saponification is carried out at 110 – 120 °C by introducing fats and caustic soda into a saponification cycle consisting of saponification reactor 1 (d), a feed pump (e), and a homogenizer (f). The reactor and circulating system are heated to facilitate start-up and shut-down. The reactor can also be cooled to dissipate heat generated during the saponification. Saponified product is fed into reactor 2 (g) for cooling, where it can also be neutralized with fatty acid. Like stage 1, this stage also consists of a reaction vessel with its associated feed pump and circulation system. The soap leaves this stage at ca. 90 °C. As a rule, saponification is adequately controlled by manual sampling for adjustment of the alkali content.

In the washing stage, soap and glycerol are separated by liquid – liquid countercurrent extraction. Washed soap leaves the extractor (i) from the top, after which it is fed into the fitting stage. A glycerol – water mixture is withdrawn at the bottom of the extractor and fed into the lye levelling vessel via a float valve. From here it is recycled to the saponification stage by way of two dosing pumps. Enriched glycerol – water is simultaneously withdrawn for evaporation by means of a third pump.

In the next stage the curd soap is fitted in a fitting reactor by the addition of fitting solution (NaOH). Like the cooling stage, this stage also consists of a reaction vessel (j) with a feed pump and a circulating system. The two phases, soap curd and nigre, are heated to ca. 95 °C, leaving the reactor as a mixture.

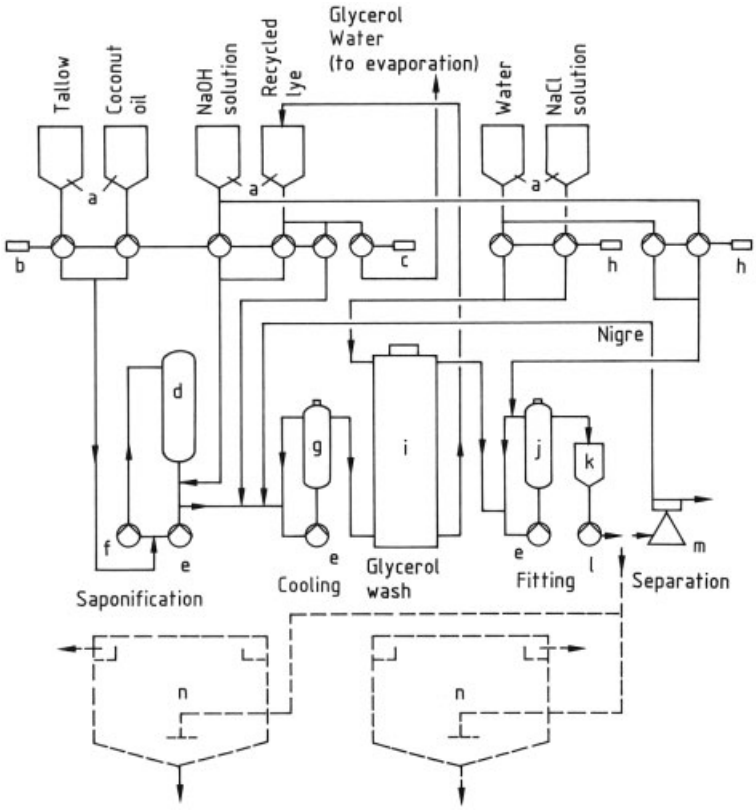

Figure 10. Continuous saponification system (Weber & Seeländer, Type KVN)
a) Level-control dosing vessels; b) Dosing pump with five dosing heads; c) Dosing pump with a single dosing head; d) Saponification reactor (reactor 1); e) Feed pumps; f) Homogenizer; g) Cooling reactor (reactor 2); h) Dosing pump with two dosing heads; i) Extractor; j) Fitting reactor; k) Leveling vessel; l) Pump; m) Centrifugal separator [alternative: n) Settling vessel]

Continuous separation of the two phases is achieved by means of a centrifuge (m), with nigre and lye being recycled to the process. Alternatively, phase separation can be accomplished in heated boiling kettles (n), each kettle being left to settle undisturbed for ca. 24 h. With two such kettles, heating and 24-h settling can be alternated, providing another approach to continuous operation with nigre and lye recycling.

For details concerning the widely practiced Mazzoni soap-making process, see [10].

4.2.3. Batch Processes

Saponiflex Process (Eirich). This modern soap-making process, developed by Eirich, permits fully automatic operation. Saponification, drying, cooling, and vigorous incorporation of all the required additives for a finished soap take place within the confines of the Eirich reaction mixer. The process, represented schematically in Figure 11 [11], is applicable to finished soap production from fats and oils, fatty acids, fatty acid methyl esters, and even syndets (see Chap. 6).

The principal raw materials–in this case fatty acids–are stocked in heated tanks (a), which maintain the components in their liquid state. Tank (b) is a storage tank for 50 % sodium hydroxide or a dissolving tank equipped with a stirrer for flakes of caustic soda. Other starting materials include coloring agents and miscellaneous additives [tanks (c)]. A computerized weighing system is employed, with dosing effected by triggering devices (d). Fatty acids are introduced first, followed by sodium hydroxide solution. Other additives are incorporated dur-

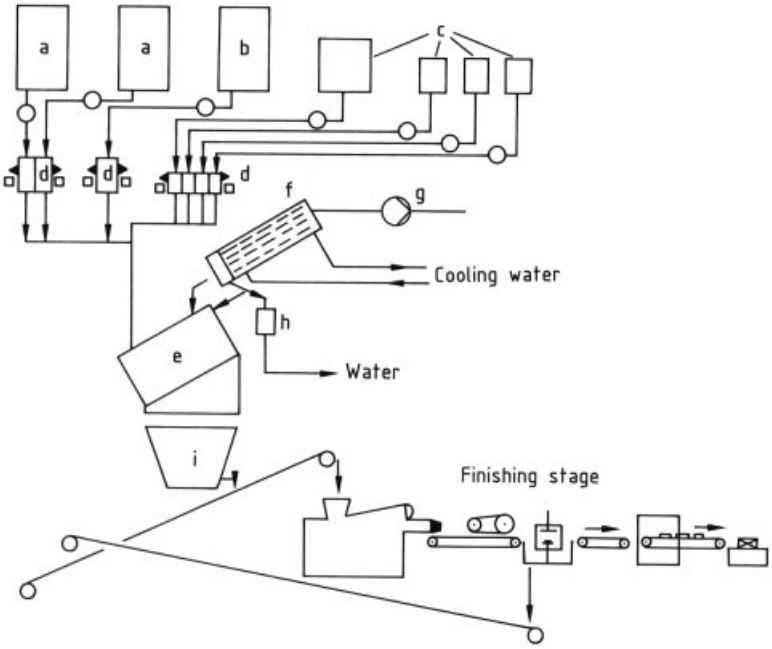

Figure 11. Schematic representation of the Saponiflex process (Eirich)
a) Heated fatty acid storage tanks; b) Storage or dissolving tank for NaOH solution; c) Storage tanks for additives; d) Scales for weight-triggered dosing; e) Mixing reactor; f) Heat exchanger; g) Vacuum pump; h) Scale for metered removal of water; i) Receiving hopper with discharge unit

ing the reaction phase. The temperature in the mixing reactor (e) rises to >100 °C. Cooling is initiated at the appropriate time by application of a vacuum. Evaporating water condenses in a multitube heat exchanger (f) and returns to the mixture. In a subsequent drying phase, the appropriate amount of condensate is collected in a scale (h) and removed. No supplemental energy is required for saponification or drying.

Temperature-sensitive components (e.g., perfume oils) are added after the cooling and drying processes.

The Eirich mixing reactor (Fig. 12) consists of a rotating vessel for mixing the charge, a stationary combination device (baffle and bottom – wall scraper), and a rapidly rotating stirrer. The use of a rotating vessel together with stationary components causes the material to be circulated continuously and mixed both vertically and horizontally. The rotating mixer is responsible for ensuring intensive preparation and dispersal of any agglomerates, as well as bringing the various components into contact with one another. Sodium hydroxide solution is mixed very quickly with the saponifiable components. The entire process is monitored and controlled automatically via a programmable control system that includes automatic recording, balancing, and alarm functions.

Numerous other batch processes exist as well, including semiboiled saponification systems from Weber & Seeländer (types DHA and DHT).

4.3. Drying

Fitting results in a liquid base soap with a fatty acid content of 62 – 63 %, but material with a 78 – 80 % fatty acid content is required for the manufacture of high-quality soap. The fatty acid content can be increased by drying. Household and toilet soaps are dried to water contents of ca. 24 % and 12 %, respectively. At the same time, the soap must be cooled to 20 – 30 °C to provide the strength necessary for further processing. The development after World War II by the Italian firm G. Mazzoni S.P.A. of vacuum spray systems for drying base soaps led

ultimately–despite initial start-up problems–to a technical revolution in the soap industry, and vacuum drying remains the drying process of choice. Evaporated water and soap are expanded under a pressure of 20 – 40 mbar, which simultaneously cools the soap to 20 – 30 °C. Water vapor is removed from the vacuum system by condensation.

Figure 12. Cutaway view of the Eirich mixing reactor

Vacuum Dryers. Figure 13 illustrates the design of a vacuum drying facility (Weber & Seeländer) [12]. Soap base from a saponifier is pumped through filters (a) into a feed tank (c) equipped with level control. Soap is then fed continuously into the dryer at a rate established by a dosing pump (d). The first stage is a heat exchanger (e) designed for a steam pressure of 9 bar absolute and a soap base inlet temperature of 80 °C. The soap is cooled in an evacuated spray chamber (f) to the temperature at which water boils at that pressure, after which it is sprayed onto a rotating cone (Fig. 14). The cone is firmly supported from outside the zone that comes into contact with the soap; the inner surface, which is in contact with the soap, is very accurately machined. This arrangement facilitates safe and virtually maintenance-free removal of the dried soap by means of a flexible scraper. Vacuum-tight soap discharge from the vacuum chamber is accomplished with a worm-type extruder (g). In the case of household soap, the dryer discharge normally takes the form of soap strands that emerge, for example, from a duplex vacuum extruder. Water evaporated during the drying process is passed through centrifugal dust collectors (h) to remove accompanying soap dust, after which it is condensed in a vent- or suface-type condenser (k). High-temperature cooling water systems require vapor compression with a steam-jet compressor (j) prior to condensation in order to achieve an equilibrium water-vapor pressure in the condenser

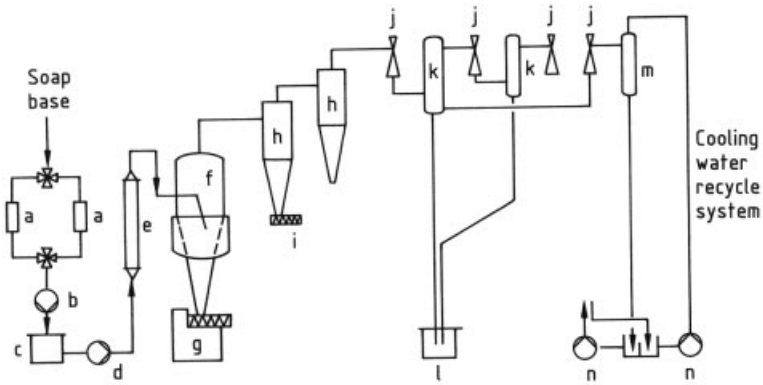

Figure 13. Schematic repesentation of a Weber & Seeländer vaccum drying and cooling system
a) Soap filters; b) Filter pump; c) Feed tank; d) Dosing pump; e) Heat exchanger; f) Evacuated spray chamber; g) Simplex screw extruder; h) Dust collectors; i) Soap-dust discharge screw; j) Steam-jet vapor compressers; k) Condensers; l) Water collection vessel; m) Vacuum flash cooler; n) Pumps

corresponding to the temperature of the cooling water. Withdrawal of noncondensable gases and air that may have leaked in is achieved either with a steam-jet device or mechanical pumps. A similar system from the Italian firm Impianti macchinesaporne affine S.R.L. (i.m.s.a.) is illustrated in Figure 15 [13].

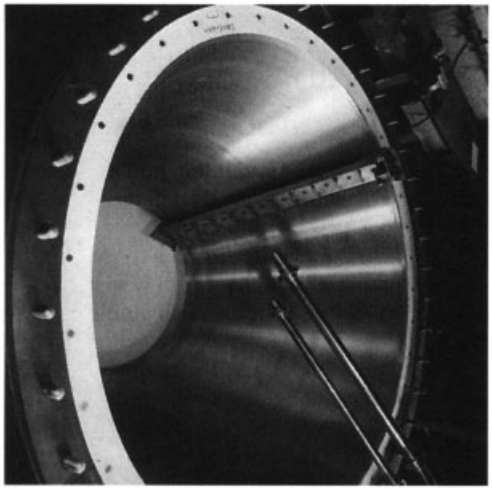

Figure 14. Rotating-cone spray dryer (Weber & Seeländer)

Alternatives to vacuum dryers include atmospheric-pressure dryers and cooling rolling mills, or vacuum dryers preceded by pressure release devices. The latter type of combination dryer provides fatty acid concentrations >68 % with simultaneous vacuum cooling, and it entails a minimal amount of cooling water for vapor condensation. In this case, the vacuum system serves only to effect cooling; the heat of condensation is essentially equivalent to heat removed in cooling the soap, so the demand for cooling water is correspondingly small. Water evaporation in this case occurs primarily in the prior atmospheric stage.

4.4. Storage

Soap flakes or noodles from the drying system are generally stored in bins before further processing. If the soap were to be processed immediately, soft bars would result depending on the soap-making oils and fats and the extent of drying, and these would tend to stick to the die of the soap press or cause problems in the packing machines.

Storing soap noodles in silos poses no problems if properly done. However, moisture diffusing outward from the interior of the ribbons cannot be absorbed at the surface, and may settle on the ribbon surfaces producing an effect known as "sweating". Wooden bins are ideally suited to the storage of soap noodles and flakes, because the porosity of wood permits moisture to escape and evaporate on the outside of the bin.

The most convenient shape for a soap silo is a box-like room with a cone to facilitate discharge of the soap to the processing machines. Belts, screws, and elevators are generally used for transport.

4.5. The Finishing of Soap Base

The processes described above give rise to a granular, pourable soap base that serves as the starting material for producing finished soap. This material has a water content of ca. 14 %, and it contains ca. 0.5 % NaCl as well as a maximum of 0.05 % free alkali, which provides some protection against rancidity. In the case of kettle saponification, the soap base still contains 0.5 – 1 % glycerol. The material has a white to yellowish color and a neutral odor. The following finishing steps are required to produce a final product:

1) Incorporation of additives for optimizing the application characteristics, producing soap with special properties, and fragrance adjustment
2) Plodding
3) Cutting into unfinished pieces
4) Forming
5) Packaging

A schematic representation of the finishing operations for producing bars of soap is provided in Figure 16.

Two methods are available for mixing and homogenizing toilet soaps:

1) Trough mixing, followed by "pilling" on a set of rolling mills. Modern equipment for this purpose has been in use for many decades. Three- and four-roll mills are most common.

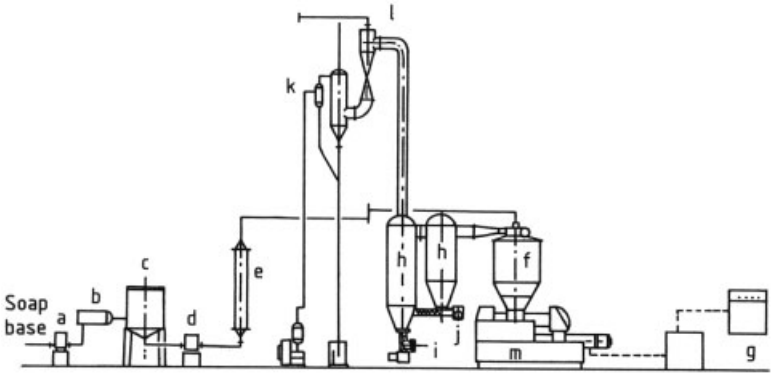

Figure 15. Schematic representation of a drying and cooling system (i.m.s.a.)
a) Soap filtration pump; b) Filter; c) Feed tank; d) Feed pump; e) Heat exchanger; f) Atomizer; g) Water chiller; h) Dust separators; i) Vertical powder extruder; j) Powder compacter; k) Barometric condenser; l) Booster; m) Duplex plodder

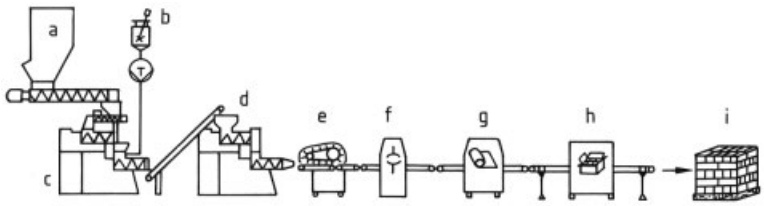

Figure 16. Finishing operations for the production of soap cakes [14]
a) Soap granulate silo; b) Storage vessel with dosing pump for color, fragrance, and additives; c) Mixer; d) Soap plodder and log extruder; e) Log cutter; f) Forming press; g) Automatic bar packer; h) Automatic cartoning unit; i) Palleting

2) Mixing and homogenization in a screw device. This process reached technical maturity only after World War II, with pioneering work again carried out by the Italian company Mazzoni.

Unlike trough mixers which process batches of soap, screw mixers (often called automatic refining machines) operate continuously. Mixing and plodding are usually accomplished in separate devices, although some machines combine the two functions.

Soap base is fed into a discontinuous mixing machine via (automatic) scales or calibrated star wheels. The appropriate small amounts of additives can be weighed by hand and added selectively to each batch. In a continuous mixing machine, perfume oils, skin-protecting agents, coloring matter, and other additives are combined as a slurry that is introduced continuously with a dosing pump.

The pelletized soap mass is then plodded, generally with a two-stage vacuum extruder. A vacuum worm extruder for this purpose consists essentially of a series of worm extruders with vacuum chambers located between each pair of worms. In the simplest case, a duplex machine like that shown in Figure 17, the plastic soap mass is precompressed in the first screw extruder (b), where it is also homogenized once again, after which it is converted in the usual way to soap noodles by extrusion through a perforated plate (die) coupled with a rotating cross-shaped cutter (c). The noodles fall into a vacuum chamber where they are lifted by the main screw (d) and extruded under high compression though a die at the head of the extruder (e). What emerges is a compact, bubble-free, glossy strand. Any breakdown in the vacuum chamber interferes with proper degassing, which can lead to bubbles in the soap strand that crack during use.

The strand is subsequently cut into unfinished pieces, which must closely approximate the desired dimensions for the finished soap bars. Machines equipped with cutting chains are now standard (e.g., machines similar to those from i.m.s.a., as shown in Fig. 18).

The unfinished pieces are stamped into their final shape in soap presses. High-speed presses

are normally used on high-volume production lines. The form of the stamp is established by designers, taking into account legal considerations such as the minimum net weight for a soap bar.

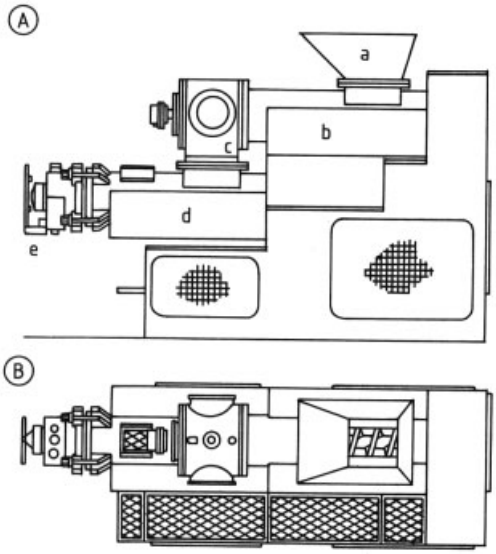

Figure 17. Duplex single-screw extruder (Weber & Seeländer)
A) Side view; B) Top view a) Soap hopper; b) Ribbon extruder; c) Ribbon cutter; d) Primary vacuum extruder; e) Log discharge

The finished bars are packaged and then placed in storage. The packaging machine [15] is usually located directly behind the press, and is adjusted to conform to its working cycle. Here the finished product is wrapped in packaging materials chosen to provide suitable protection during transport. Automatic cartoning machines are used to pack the product into boxes, which are then loaded onto pallets.

Occasionally it becomes necessary to subject a soap to interim storage before packaging to permit secondary hardening to take place.

5. Characteristic Features of the Most Common Types of Soap

5.1. Toilet Soaps

Toilet soaps are used for body cleansing and personal hygiene. Based on turnover and distribution they represent the most important type of soap. Toilet soaps are sodium soaps made from high-grade fats, such as coconut oil, beef tallow, palm oil, and olive oil, with a fatty acid content of ca. 80 %. Good foaming properties require that the starting materials contain 20 – 50 % coconut oil. Toilet soaps are slightly superfatted, in that they contain 1 – 10 % free fatty acids, fatty alcohols, glycerol, lanolin, petrolatum, cocoa butter, almond oil, wheat germ oil, or fatty acid ethanolamides [16]. These superfatting agents ensure the refatting of freshly cleansed skin. Cosmetic dyes are introduced for coloring purposes at a level of $\leq$0.1 % in accordance with government regulations. Patented processes permit the preparation of marbled or striped soaps. Added brilliance is achieved with titanium dioxide. Complex mixtures of natural and synthetic perfume oils are also employed (0.5 – 5 %, with a trend toward more). Special effects are achieved with substances such as honey, camomile extracts, milk powder, or abrasives. Antioxidants and optical brighteners are also used occasionally. Antioxidants (e.g., butylated hydroxytoluene) and complexing agents [e.g., ethylenediamine tetraacetic acid (EDTA) and (1-hydroxyethane-1,1-diphosphonic acid [*2809-21-4*] (HEDP)] prevent a soap from becoming rancid. The role of the complexing agents is to bind catalytically with active heavy metals.

Soap bars are subject to repeated contamination with bacteria from water or washed from the skin [17], so they have sometimes been regarded as "hygienically questionable." Bacterial survival in toilet soaps, especially those containing special additives like proteins, can be countered by addition of a preservative. It has been found in practice, however, that preservation of this type is usually required only in exceptional cases.

5.2. Other Soap Products

Luxury Soaps. For soaps in this category the typical toilet soap charge is supplemented with a high proportion ($\leq$ ca. 5 %) of perfume oils.

Deodorant soaps contain an active ingredient such as 3,4,4′-trichlorocarbanilide (trichlocarban) to inhibit the growth of skin bacteria responsible for the formation of odor-causing

substances from perspiration (→ Skin Cosmetics, Chap. 8.).

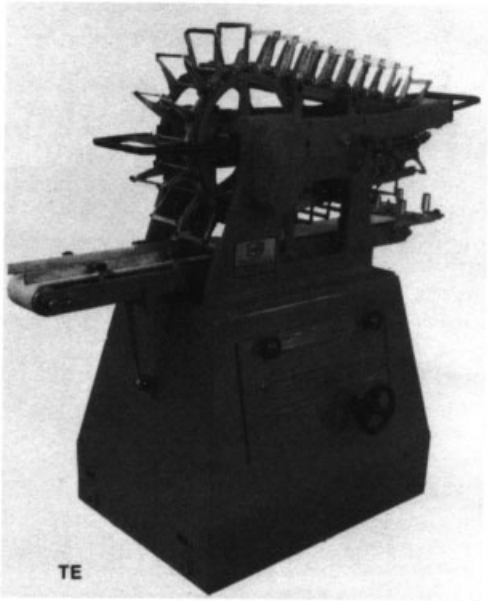

Figure 18. Example of a soap-cutting machine (i.m.s.a. Model TE)

Skin-protecting soaps are perfumed only slightly and contain refatting substances as well as additives that not only reduce the scouring effect of the soap, but also positively influence the state of the skin after washing (e.g., proteins, milk components).

Cream Soaps. Cream toilet soaps are exceptionally rich in refatting substances. The fat charge frequently contains a high proportion of coconut oil.

Transparent soaps are partially transparent, because special substances (e.g., glycerol, ethanol, sugar) have been added to prevent crystallization.

Abrasive soaps contain abrasive substances (e.g., quartz, sand, almond bran) to help remove stubborn dirt.

Floating Soaps. These soaps have a density <1 g/cm^3 and therefore float on water. They are made by working air into the soap during production, or by creating an air pocket within the soap bar.

Liquid soaps are potassium soaps with a fatty acid content of ca. 15 – 20 %. The fat charge includes coconut oil in addition to other vegetable oils. Liquid soaps are preferred for use in soap dispensers, because their discharge rate is easily controlled. A typical formulation for a liquid soap (in wt %) is as follows [18]:

Potassium salt of coconut oil fatty acids	ca. 15
Potassium salt of oleic acid	3
Glycerol	5
Refatting agents	5
Pearly-luster agents	2
Special additives, complexing agents, lime soap dispersing agents, perfume oils and dyes, water	balance

Baby Soaps. These toilet soaps are well superfatted and only slightly perfumed, if at all. They also contain special additives such as camomile constituents.

Shaving soaps and creams are mixtures of sodium and potassium soaps to which 5 – 10 % of a moisturizer (glycerol) is added. Shaving soaps command a much smaller share of the market than shaving creams and especially shaving foams (→ Hair Preparations, Chap. 8.3.). The following is a typical formulation (in wt %) for a shaving soap [19]:

Fatty acids, alkali salts (with 20 – 25 % originating from coconut oil or palm kernel oil)	ca. 80
Glycerol	5 – 10
Water	8 – 10
Perfume oils	1.5 – 2.5

Hard Soaps. These are sodium soaps made from fats and oils similar to those used for toilet soaps. Their fatty acid content is 60 – 65 %. Hard soaps are used mainly for washing the hands and occasionally for washing clothes.

Soft Soaps. This is a collective name for soaps with a paste-like consistency. They share the characteristic of having been saponified with potassium hydroxide. A distinction can be made between smooth, transparent, and granular products. The colors range from white to yellow, brown, or green. Soybean oil, olive oil, and

peanut oil can all be used for the fat charge. The fatty acid content is ca. 40 %. These soaps are intended for floor cleaning and similar tasks.

Soap flakes and soap needles are highly dried sodium soaps with a fatty acid content of ca. 85 %. They are used primarily for industrial purposes, especially in laundries and for textile finishing.

Table 3. Basic formulation for the manufacture of syndet bars [21]

Component	Content, %
Fatty alkyl sulfates ∗	30 – 60
Builders/plasticizers: mono-, di-, and triglycerides, paraffins, microwaxes, fatty alcohols	30 – 50
Fillers: starch, talcum, zinc oxide, titanium oxide	0 – 15
Refatting/superfatting agents: lecithins, lanolin, oils, alkanolamides	1 – 7
Active ingredients: D-panthenol, vitamin E, amino acids, disinfectants	1 – 3
Water and moisturizers: lactates, amino acids, sorbitol, glycerol	1 – 5
Coloring agents	0 – trace
Perfumes	0 – 1

∗ Alternative surfactants: (1) fatty alkyl sulfosuccinates, 10 – 35 %, and fatty alkyl sulfates, 10 – 40; (2) fatty acid isethionates, 10 – 40 %, and fatty alkyl sulfates, 10 – 40 %; (3) fatty acid isethionates, 30 – 50 %, and fatty alkyl sulfosuccinates, 0 – 20 %; (4) potassium salts of fatty alkyl sulfonates, ca. 70 %.

6. Syndets

The term "syndet" is derived from *syn*thetic *det*ergents (→ Laundry Detergents), and refers to soap-free products for skin-care and skin-cleansing applications. In the broadest sense, bath and shower preparations also belong to this group (→ Skin Cosmetics, Chap. 7.). However, the term syndet is usually restricted to cosmetic products in bar form [20].

Soaps and syndets share one common characteristic: they contain surfactants as their main components. Soaps are salts of strong bases and weak acids, so they always give an alkaline reaction in aqueous solution, and lose their surfactant properties under neutral or acidic conditions. This is not true with syndets, which can be "pH controlled". Soaps react with hardness-causing ions, especially calcium ions, to form slightly soluble, wash-inactive salts called lime soaps (→ Metallic Soaps) with a consequent loss of surfactant properties. Syndets do not form insoluble lime soaps. Their cleansing power can be adjusted for a particular application by appropriate choice of the surfactants, permitting skin cleansing under neutral to acidic conditions. Both solid (bar) and liquid formulations are available.

Constituents. Solid syndets contain the following principal ingredients [21]:

1) Wash-active substances
2) Builders/plasticizers
3) Fillers
4) Perfume oils/coloring agents

Basic formulations for the preparation of solid syndets are provided in Table 3.

Liquid products (Table 4) differ by the inclusion of such additional substances as thickening agents, refatting agents, or substances to increase compatibility, active ingredients, preservatives, and antioxidants.

Table 4. Basic formulation for the manufacture of liquid syndets [21]

Component	Content, %
Ether sulfates, organic neutralized (56 %) ∗	30 – 40
Protein fatty acid condensates ∗	0 – 5
Betaines ∗	0 – 5
Protein hydrolyzates	0 – 5
Thickening agents	1 – 8
Pearly-luster agents	0 – 5
Refatting agents	1 – 7
Active ingredients	1 – 5
Coloring agents	0 – trace
Perfume oils	0 – 1
Preservatives, antioxidants	0.05 (0) – 1

∗ Alternative formulations with respect to the first three components (surfactants): (1) betaines, 10 – 40 %, and fatty alcohol ether sulfosuccinates, 10 – 30 %; (2) betaines, 10 –40 %, and protein fatty-acid condensates, 5 – 15 %; (3) protein fatty-acid condensates, 20 – 50 %, and olefin sulfonates, 0 – 20 %; (4) fatty alcohol ether sulfates, 25 – 45 %, and ether carboxylates, 5 – 15 %; (5) fatty alcohol ether sulfosuccinates, 20 – 35 %, and fatty alcohol sulfoacetates, 10-15 %.

Production. Solid syndets are made on standard soap production lines. Liquid preparations can be mixed in vessels equipped with high-speed rapid stirrers, anchor stirrers, or similar devices, starting with raw materials that are either liquid at room temperature or liquified by heating.

Outlook. In the future, syndets are expected to press forward vigorously into the field of skin care because, unlike soaps, they lend themselves to a widely diversified range of products with respect to chemical composition, cleansing power, pH, and skin-care properties. Details regarding skin compatibility are best sought in the technical literature [22].

7. Environmental Protection

A simplified ecological balance for soap is presented in Figure 19. Used soap collects in wastewater, where it is degraded biologically either in the water itself under aerobic conditions or anaerobically in waste-treatment sludge. Unlike other anionic surfactants, the degradability of a soap cannot be evaluated by measuring primary degradation, because soaps show no methylene blue activity. For this reason, soap levels must be quantified by other means (e.g., gas chromatography) after various periods of degradation [23].

In 1991, The Netherlands' Organization for Applied Scientific Research (TNO) Institute of Environmental Sciences published a report on the biodegradability of soaps from fatty acids with different carbon chain lengths [24]. This report concluded that even after short time periods a high level of primary bacteriological degradation is achieved in all cases. The requirement for a biodegradation of >60 % (e.g., in the closed-bottle test) is readily met within short test periods, indicating that soaps are characterized by excellent degradability. All the evidence points to rapid degradation in waste-treatment plants as well. Both short- and long-chain soaps are degraded completely under aerobic conditions.

For information on the effects of syndet preparations on the human environment see [25].

8. Quality Control and Analysis

Paragraph 4 of the German Food and Commodities Law describes cosmetic articles as "substances and preparations made from materials intended for human use, either externally or in the oral cavity, for the purpose of cleansing, care, influencing appearance or body odor, or imparting odoriferous sensations", in contrast to substances whose primary purpose is to "alleviate or eliminate disease, injury, physical damage, or pathological complaints". Therefore, based on this definition, ordinary soaps and syndets must be regarded in Germany as cosmetic agents. This generalization does not apply to medicinal soaps, which are not discussed here. It is mandatory that soaps and syndets be free of potential hazards to human health. In Germany they are also subject to the requirements of the statute on cosmetics [26], which regulates the nature of permissible ingredients. Coloring agents and allowed preservatives are covered by the "Blue List" [27].

A typical toilet soap granulate should be consistent with the following analytical specifications immediately after production:

Bound fatty acids	79.0 ± 1 %
Free NaOH	0.05 ± 0.02 %
NaCl	0.4 ± 0.1 %
Loss on drying	12.5 ± 1 %
Complexing agent	positive test

Product appearance and odor are also tested. Recognized analytical procedures include the uniform methods for testing fats, fat products, and related substances issued by the German Society for Fat Science (Deutsche Gesellschaft für Fettwissenschaft e. V.), Münster/Westfalen. Analytical characteristics for a typical bar of toilet soap include:

Appearance:	color corresponding to a standard sample
Odor:	characteristic, identical to that of a standard sample
Bound fatty acids:	78.5 ± 1.5 % (based on a saponification number of 219 ± 2)
Free alkali:	0.04 ± 0.02 %
Petroleum ether extract:	1.5 ± 0.2 %
Smoothness:	remains smooth during washing with luke-warm water
Cracking:	no cracking in a wet-crack test

Information on quality control for syndet-containing cleansing agents is available from [28]. In this case the most important quality characteristics are convenience, foam development, foam quality, swelling behavior, economic factors, tendency to crack, and compatibility with the skin.

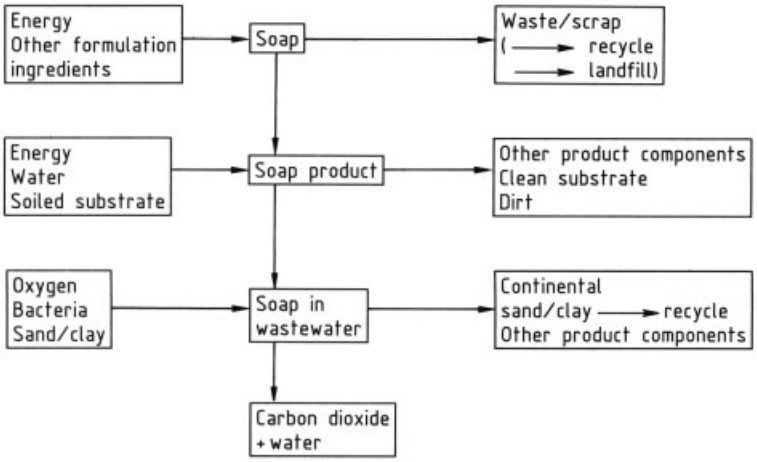

Figure 19. Simplified ecological balance for soap

9. Economic Aspects

With a production of 8.9×10^6 t in 1990, soaps still ranked first in worldwide statistics for surfactants [29]. They account for 56 % of total surfactant production, primarily because soap is very economical and can be produced inexpensively. The use of soap as a builder in liquid detergents has recently become somewhat more important in the Western European market, although its use as an antifoaming agent continues to decrease (→ Laundry Detergents, Chap. 3.4.3.). However, soap is not a suitable alternative to the usual surfactant-based, heavy-duty detergents. Soap is used to wash clothes in less than 1 % of German households.

On the whole, soap production has remained relatively stable in the past few years. The 1990 worldwide rate of increase (2.2 %) slightly exceeds the increase in population (1.8 %). Toilet soaps, with a production volume of 2.7×10^6 t/a, display a significantly lower growth rate (1 %). The consumption of soap in 1990 was 0.68×10^6 t in Western Europe, 0.70×10^6 t in the United States, and 0.19×10^6 t in Japan, giving a total of 1.57×10^6 t. Figure 20 shows the ten most important soap-producing countries (1990). Current per capita soap consumption in Western Europe is about 2.1 kg, and the world average is 0.5 kg.

The syndet market share is largest in the United States. In Western Europe, pH-regulated soap bars and liquid syndets in the form of washing lotions are steadily gaining importance.

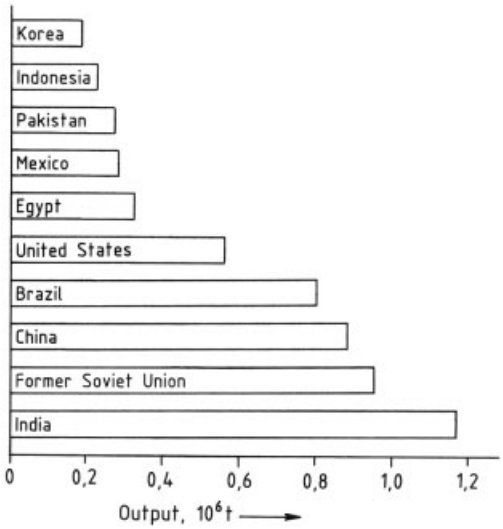

Figure 20. Overview of production data for the ten leading soap-producing countries [28]

10. Toxicology

Considered from the point of view of a cleansing and cosmetic agent, soap comes in contact only with the exterior of the body. Experience indicates that daily use of the product is unlikely to be a source of problems. From a toxicological and dermatological perspective, excessive application or prolonged contact of the skin with soap solutions can result in protein denaturing and the release of amino acids. The emulsifying action of a soap is responsible, on the one hand, for the

desired cleansing action, but it can also lead to the removal of natural oils from the horny layer of the skin; overtaxing the skin's regenerative capacity can have deleterious effects on its barrier functions. Excessive application may result in increased swelling and barrier-function disturbances that in turn lead to eczema-like manifestations. Allergic reactions to soap products have also been observed, but specific allergies to soap itself are rare. Individual reports in the literature point to perfume oil additives as the most common cause of problems. Irritation may also accompany the action of high concentrations of soap powder or soap solution on eye tissue. Injection of soap solutions can cause necrosis and hemolysis. Depending on the dose, oral ingestion of soap (typically by small children) may lead to local irritation of the mucous membranes, nausea, vomiting, and foam buildup, which can in turn induce aspirative pneumonia. Gastrointestinal disorders such as flatulence and diarrhea have also been reported. Generally speaking, however, the quantity of soap ingested is so small that risks of serious or prolonged damage to the health are insignificant.

11. References

1. B. W. Werdelmann, *Fett Wiss. Technol.* **89** (1987) 475 – 480.
2. H. Verbeek: "Historical Review" in J. Falbe (ed.): *Surfactants in Consumer Products,* Springer Verlag, Berlin 1987, pp. 3 – 4.
3. J. Knaut, H.-J. Richtler: "Soap and Personal Care Products: An Important Outlet for Oleochemicals" in T. H. Applewhite (ed.): *Proc. World Conference on Oleochemicals into the 21st Century,* American Oil Chemists' Society, Champaign, Ill. 1990, pp. 27 – 37.
4. K. Stanzl, *Parfüm. Kosmet.* **71** (1990) 696 – 700.
5. E. Woollatt: *The Manufacture of Soaps, Other Detergents and Glycerine,* Ellis Horwood, Chichester 1985, pp. 198 ff.
6. Chiyoda Chemical Eng. and Constr. Co., Ltd., "Lions E-S Process" Company brochure 1965.
7. Weber + Seeländer, "Anlagen zur Verseifung von Neutralfett und Fettsäure," Company brochure, Helmstedt.
8. E. Woollatt: *The Manufacture of Soaps, Other Detergents and Glycerine,* Ellis Horwood, Chichester 1985, pp. 179 ff.
9. G. R. Whalley, *HAPPI Household Pers. Prod. Ind.* **28** (1991) 81 – 86.
10. *Ullmann,* 4th ed., **21,** 215 – 216.
11. Maschinenfabrik G. Eirich, "Saponiflex, Das kompakte System für mehr Flexibilität und Wirtschaftlichkeit bei der Seifenherstellung," Company brochure, Hardheim.
12. Weber + Seeländer, "Trocken- und Kühlsysteme," Company brochure, Helmstedt.
13. Impianti macchine sapone affine S.R.L. (i.m.s.a.) Company brochure, Busto Arsizio (VA), Italy.
14. W. Schneider, G. Thor: "Hautwasch- und Reinigungsmittel" in W. Umbach (ed.): *Kosmetik – Entwicklung, Herstellung und Anwendung kosmetischer Mittel,* Thieme Verlag, Stuttgart 1988, p. 66. W. Schneider, G. Thor: "Skin Washing and Cleansing Preparations" in W. Umbach (ed.): Cosmetics and Toiletries–Development, Production and Use, Ellis Horwood, Chichester 1991, p. 56.
15. L. Spitz: "Soap Packaging Machinery and Materials" in Spitz (ed.): *Soap Technology for the 1990's,* American Oil Chemists' Society, Champaign, Ill. 1990, pp. 292 – 320.
16. G. Kolaczinski: "Seifen: Chemische Bestandteile" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigung mit Syndets–Chemische, ökologische und klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 18 – 23.
17. T. Kamps, B. Meyer, *Seifen öle Fette Wachse* **117** (1991) 710 – 713.
18. W. Schneider, G. Thor, Hautwasch- und -reinigungsmittel in W. Umbach (ed.): *Kosmetik – Entwicklung, Herstellung und Anwendung kosmetischer Mittel,* Thieme Verlag, Stuttgart 1988, pp. 57 – 68. W. Schneider, G. Thor: "Skin Washing and Cleansing Preparations" in W. Umbach (ed.): Cosmetics and Toiletries – Development, Production and Use, Ellis Horwood, Chichester 1991, pp. S. 49 – 57.
19. A. Pond: "Rasiermittel" in W. Umbach (ed.): *Kosmetik – Entwicklung, Herstellung und Anwendung kosmetischer Mittel,* Thieme Verlag, Stuttgart 1988, pp. 163.
20. K. Schumann: "Syndet–Ein Begriff im Wandel der Zeit" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigung mit Syndets–Chemische, ökologische und Klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 13 – 17.

21. W. Schadenböck: "Zusammensetzung marktüblicher Syndet-Zubereitungen" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigung mit Syndets–Chemische, ökologische und Klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 31 – 38.
22. B. König: "Einsatz von Hautreinigungsmitteln auf Syndetbasis in der Praxis–Die Sicht des Allgemeinmarktes" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigung mit Syndets–Chemische, ökologische und Klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 235 – 238.
23. H. Brüschweiler, F. Schwager, H. Gämperle: "Seife-Waschverhalten, Wascheffekte und Abbaubarkeit", *Seifen öle Fette Wachse* **114** (1988) 301 – 307.
24. TNO-report R91/072: A Review and Evaluation of the Biodegradability of Soaps, TNO Institute of Environmental Sciences, Delft 1991.
25. H. H. Rump: "Beeinflussung der menschlichen Umwelt durch Syndet-Zubereitungen" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigung mit Syndets–Chemische, ökologische und Klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 217 – 224.
26. Kosmetik-Verordnung of 16. 12. 1977, BGBl. I, 21. Dec, 1977, p. 2589 including last revision of 25. March, 1991.
27. H. P. Fiedler et al. (eds.): *Blaue Liste,* Editio Cantor, Aulendorf 1989.
28. K. Stanzl: "Qualitätskontrolle von syndethaltigen Reinigungsmitteln" in O. Braun-Falco, H. C. Korting (eds.): *Griesbach Konferenz Hautreinigug mit Syndets–Chemische, ökologische und Klinische Aspekte,* Springer Verlag, Berlin 1990, pp. 211 – 216.
29. H.-J. Richtler, J. Knaut, *Seifen öle Fette Wachse* **117** (1991) 545 – 553.

Soda → Sodium Carbonates

Sodium Amide

LUDWIG LANGE, Degussa AG, Frankfurt am Main, Federal Republic of Germany

WOLFGANG TRIEBEL, Degussa AG, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Properties 1
2.1. Physical Properties of a Solution of Sodium in Liquid Ammonia 1
2.2. Physical Properties of Sodium Amide 2
2.3. Chemical Properties 3
2.3.1. Chemical Properties of $Na-NH_3$ Solution 3
2.3.2. Chemical Properties of Sodium Amide 4
3. Production 6
4. Safety Aspects 7
5. Uses 8
6. Transport Regulations 10
7. References 10

1. Introduction

In 1811, DAVY and GAY-LUSSAC reported: "If sodium is heated in gaseous ammonia, an olive-green compound is formed. Ammonia is absorbed, and an amount of hydrogen is evolved equal to the amount that would be evolved if the same amount of sodium were added to water" [1].

The reaction

$$2\,NH_3 + 2\,Na \rightarrow 2\,NaNH_2 + H_2$$

proceeds quantitatively according to BEILSTEIN and GEUTHER when oxygen-free, dry ammonia is passed over heated sodium for 6 – 7 h. As the reaction proceeds, greenish blue droplets form on the metal surface [1].

In 1863, T. WEYL discovered that alkali metals dissolve in liquid ammonia to give colored solutions. Since then, these solutions [2], with their unusual properties, have been intensively investigated [3].

This is also true of sodium amide [4–6], a compound whose industrial exploitation is described in detail in [7], [8].

2. Properties

2.1. Physical Properties of a Solution of Sodium in Liquid Ammonia

Sodium dissolves in liquid ammonia without evolution of hydrogen to give a deep indigo blue color [9]. The characteristic properties of this solution are its color, its electrical conductivity, and its magnetic susceptibility [2]. The conductivity of the dilute solution is about an order of magnitude higher than that of completely dissociated salts in water. With increasing concentration, the conductivity first decreases, reaching a minimum at 0.04 mol/L, and then increases to values typical of liquid metals. Dilute solutions are paramagnetic, with a susceptibility corresponding to one free electron per metal atom. The susceptibility decreases with increasing concentration. The solution is diamagnetic in the region of the conductivity minimum, and then becomes slightly paramagnetic at higher concentrations. The changes in properties with concentration can be explained on the basis of the equilibrium reaction between liquid ammonia and dissolved sodium [2]:

$$Na_{(dissolved)} \rightleftharpoons Na^{+}_{(dissolved)} + e^{-}_{(dissolved)}$$

Gibson and Argo first coined the expression "solvated electrons" for the "state of excess electrons" in liquid ammonia [3]. It has been suggested that, in liquid ammonia, alkali metal anions, e.g., sodium anions, could exist with a centrosymmetric structure with both the valence electrons in an *s* orbital [3], [10].

On the basis of the fact that the solution $Na-NH_{3(liq.)}$ has a much lower density than liquid ammonia and hence a higher volume than would be expected from the sum of the volumes of the metal and the solvent, concepts of the structure of such solutions have been developed

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_267

[2], [3], [11], [12]. The increase in the apparent molar volume must be due to the unusual state of solution of the electrons, since volume increases of this order are not observed for solutions of other substances. On the contrary, when ionic substances are dissolved, there is a considerable volume decrease due to electrostriction [13].

The increase in equivalent conductivity by several orders of magnitude when the concentration of alkali metals (e.g., sodium) in liquid ammonia increases to >0.5 mol/L is associated with a transition from a "nonmetallic" to a "metallic" state. In the concentration range where the conductivity is highest, a miscibility gap is observed at low temperature. The $Na-NH_3$ solution splits into a dilute "nonmetallic" phase and a concentrated "metallic" phase, both in equilibrium with sodium [14]. The denser phase is blue and has a lower metal concentration than the less dense bronze-colored phase [15]. The miscibility gap is illustrated in Figure 1 [13], [15].

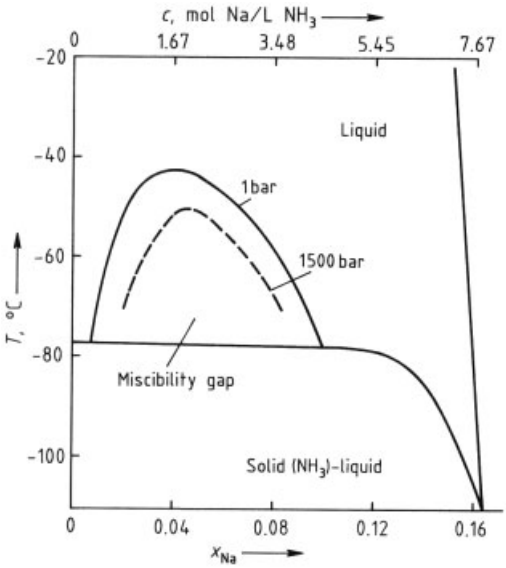

Figure 1. Excerpt from the $Na-NH_3$ phase diagram showing the eutectic and the miscibility gap, which is displaced to lower temperature and higher concentration on application of pressure

The observed physical and chemical properties of solutions of alkali metals in liquid ammonia have not yet been explained by a single model [11], [13], [16].

2.2. Physical Properties of Sodium Amide [17]

The solubility of alkali metal amides, including sodium amide, in liquid ammonia is low. Some figures for the solubility of sodium amide in liquid ammonia (in g $NaNH_2$/100 g NH_3) are given below [18]:

T, °C	−20	−10	0	+10	+20
Solubility	0.114	0.130	0.140	0.149	0.163

The solubilities of predominantly ionic solids increase with increasing dielectric constant of the solvent [6]. Figure 2 shows the values of the dielectric constants of ammonia and water as a function of temperature. Water is the best-known protic solvent. The dielectric constants of ammonia are much lower than those of water. Hence, ammonia is a less effective solvent for inorganic compounds [6].

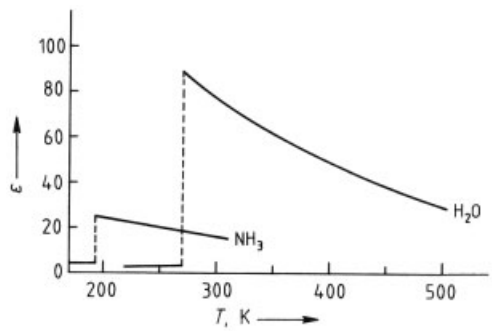

Figure 2. Dielectric constant ε as a function of temperature for ammonia and water [6]

Since sodium amide does not form a solid ammoniate, single crystals of $NaNH_2$ can be obtained from liquid NH_3 at room temperature. The properties and structure of $NaNH_2$ have been extensively investigated [6].

Sodium amide, $NaNH_2$, M_r 39.02, *mp* 208 °C, $\varrho_4^{25} = 1.39$ g/cm^3, enthalpy of formation at 25 °C $\Delta H^0 = -113.8 \pm 2$ kJ/mol, forms colorless, strongly birefringent, anisotropic crystals with a rhombic face-centered crystal structure, and a lattice energy of 774 kJ/mol. The lattice consists of Na^+ and NH_2^- ions [19], [20].

A 0.01 mol/L solution of $NaNH_2$ at − 33.5 °C is 3.8 % dissociated into Na^+ and NH_2^- ($K_{\text{dissoc.}} \approx 2.65 \times 10^{-5}$ [17]).

2.3. Chemical Properties

Solutions of sodium in liquid ammonia are versatile reducing agents and can also be used for elimination reactions. Some examples of these reactions are [13], [21]:

1) Redox reactions
2) Reduction of alkynes
3) Reduction of aromatic systems (Birch reduction)
4) Dehalogenation and decyanidation reactions

Sodium amide has a wide range of uses in chemistry. Comprehensive reviews of its reactions are available in [2], [5], [6], [19], and [21–23]. A few examples of these are:

1) Deprotonation of organic compounds
2) Dehalogenation
3) Reactions with primary and secondary amines
4) Initiation of the anionic polymerization of ionic chains
5) Amination of pyridine (Chichibabin reaction)
6) Cleavage of ketones (Haller – Bauer reaction)

2.3.1. Chemical Properties of Na – NH_3 Solution

Solutions of sodium in liquid ammonia are metastable at equilibrium pressure and room temperature, and decompose gradually, forming sodium amide [6].

Redox Reactions. Solutions of alkali metal salts in ammonia develop a blue color at the cathode on electrolysis due to the formation of solvated electrons. The thermodynamically favored formation of hydrogen at the cathode is completely prevented by a high voltage excess. The reaction inhibition that leads to the hydrogen voltage excess is possibly caused by the low rate of reaction of the solvated electrons with ammonia [13].

Thermodynamic considerations, including the overvoltage observed in liquid ammonia at the hydrogen and nitrogen electrodes, have been invoked to explain [2], [24] why it is possible to work in liquid ammonia both with powerful reducing agents, and with powerful oxidizing agents. Examples of these reactions are given in [1], [2]. The solvated electrons, which have a large negative redox potential, can react with many dissolved substances [13]. Thus, with the exception of the ions of the alkali metals, alkaline-earth metals, and divalent rare earth metals, all metal cations are reduced by solvated electrons.

Anions with completely filled electron shells, such as F^-, Cl^-, Br^-, I^-, CN^-, CNO^-, CNS^-, N_3^-, and OH^-, do not react with solvated electrons. Also, the following nonmetallic oxo-anions, which are in their highest oxidation state, are inert toward solvated electrons: $B_4O_7^{2-}$, CO_3^{2-}, HCO_3^-, ClO_4^-, HPO_4^{2-}, PO_4^{3-}, SO_4^{2-}. However, high reactivity is shown by the metal oxo-anions such as MnO_4^- and $Cr_2O_7^{2-}$, and by all the oxo-anions of nitrogen, bromine, and chlorine (except ClO_4^-) [13].

An example of the reduction of transition metal complexes to unusually low oxidation states is provided by the following equation [2]:

$$Fe(CO)_5 + 2\,Na \xrightarrow{NH_3/-33\,°C} Na_2[Fe^{-II}(CO)_4] + CO$$

High-pressure ammonolysis opens up new synthetic possibilities, for example, for some hydroxides, hydrosulfides, and chalcogenides [6].

Reduction of Alkynes. Only internal triple bonds can be reduced by sodium in liquid ammonia. Terminal alkynes are converted to acetylide ions under these conditions:

$$RC{\equiv}CH + e_{\text{diss.}} \rightarrow RC{\equiv}C^- + 1/2\,H_2$$

Isolated double bonds in organic compounds are not reduced, but allenes, conjugated dienes, and aromatic rings are [21].

Reduction of Aromatic Systems (Birch Reduction) [21]. If aromatic rings are reduced by sodium in liquid ammonia, usually in the presence of an alcohol, 1,4-addition of hydrogen results to give a nonconjugated cyclohexadiene in good yield. When substituted benzenes are

reduced, often a single product is formed. Thus, the reduction of *o*-xylene is a good route to 1,2-dimethyl-1,4- cyclohexadiene [25]. Naphthalene can be regioselectively hydrogenated to 1,4-dihydronaphthalene. If an excess of sodium is used, 1,4,5,8-tetrahydronaphthalene is obtained [26]. Reduction of thiophene by sodium in liquid ammonia in the presence of methanol gives 2,3- and 2,5-dihydrothiophene [26].

Dehalogenation and Decyanidation. Instead of the more commonly used zinc and magnesium, $Na-NH_3$ can be used to dehalogenate vicinal dihalides [13], [21].

Sodium in liquid ammonia can also be used to remove the cyanide group from alkyl nitriles. If the nitrile compounds also contain, e.g., trityl, benzyl, phenyl, or tertiary alkyl groups, the reaction with $Na-NH_3$ solution gives high yields. Considerably lower yields are obtained with primary or secondary alkyl groups. It is thought that the solvated electrons in the $Na-NH_3$ system are responsible for the reaction mechanism, which proceeds via free radicals and then carbanions [13], [21].

2.3.2. Chemical Properties of Sodium Amide

Sodium amide is a strong base [19], [27]. The reaction of ammonium salts with metal amides in liquid ammonia proceeds analogously to the neutralization of acids by bases in water, for example

$$NH_4NO_3 + NaNH_2 \rightarrow NaNO_3 + 2\,NH_3$$

Metal amides such as sodium amide behave as bases in the ammonia system, like metal hydroxides in an aqueous system. Hydrolysis corresponds to ammonolysis in ammonia. The proton affinity of ammonia, i.e., the energy released when a proton bonds to an NH_3 molecule, is 875 kJ/mol, which is much greater than the proton affinity of water (762 kJ/mol) [19]. Correspondingly, the proton affinity of NH_2^- is considerably higher than that of OH^-. The proton affinity of OH^- in the gas phase is 1634 kJ/mol, but in water this decreases to 1188 kJ/mol due to hydration of OH^-. Bases A^- whose effective proton affinity in water is >1188 kJ/mol are completely converted to the corresponding acid HA with formation of OH^-. For this reason, NH_2^- cannot be studied by dissolving its salts in water, because OH^- ions are formed quantitatively and the NH_2^- ions are completely protonated to NH_3 [27].

This behavior is known as the "leveling effect" of the solvent. Any base that is strong enough to react with water completely to form OH^- ions is leveled. Thus, OH^- is the strongest base that can exist in water [27]. Therefore, depending on the temperature and the solubility of the reactants and end products, liquid ammonia or inert hydrocarbons can be used as media for reactions with sodium amide [17].

Deprotonation. In liquid ammonia, sodium amide acts as a strong base, like NaOH in water. It deprotonates terminal triple bonds to form strongly nucleophilic acetylide anions [25]:

$$RC{\equiv}C^- + R'X \rightarrow RC{\equiv}C{-}R' + X^-$$

By using a series of reactions, it is possible to build up longer chains from shorter ones, and to prepare *cis*- or *trans*-alkenes

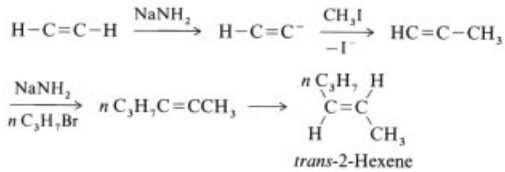

Another example of the basicity of $NaNH_2$ is its reaction with di- and triphenylmethane, which are converted into the corresponding carbanions [25]:

$$(C_6H_5)_2CH_2 \xrightarrow[-NH_3]{NaNH_2} (C_6H_5)_2CH^-Na^+$$

$$(C_6H_5)_3C{-}H + NaNH_2 \overset{-NH_3}{\rightleftharpoons} (C_6H_5)_3C^-Na^+$$

Dehalogenation. Sodium amide can be used to prepare alkynes from geminal or vicinal dihalides [25]:

$$\left.\begin{array}{l} RCX_2CH_3 \\ RCHXCH_2X \\ RCH_2CHX_2 \end{array}\right\} + 2\,NaNH_2 \longrightarrow R{-}C{\equiv}C{-}H$$

$$R{-}C{\equiv}C{-}H \xrightarrow{NaNH_2} RC{\equiv}C^-Na^+ + NH_3$$

$$RC{\equiv}C^-Na^+ + H_2O \rightarrow RC{\equiv}CH + NaOH$$

As the reaction product is the salt of an alkyne, this method can be used to prepare terminal alkynes, even if the triple bond is capable of migration.

$$n\,C_{14}H_{29}CHBrCH_2Br \xrightarrow{NaNH_2} n\,C_{14}H_{29}C{\equiv}CH$$

Inner alkynes can be isomerized to terminal alkynes by reaction with sodium amide at 150 °C [25]:

$$n\,C_5H_{11}C{\equiv}CCH_3 \xrightarrow[150\,^\circ C]{NaNH_2} n\,C_6H_{13}C{\equiv}C^-Na^+ \xrightarrow[-\,NaOH]{+\,H_2O} n\,C_6H_{13}C{\equiv}CH$$

Under normal conditions, unactivated aromatic halides hardly react with nucleophilic reagents, but they are often reactive toward very strong bases [28]. For example, chlorobenzene can be smoothly converted to aniline by reaction with sodium amide in liquid ammonia at − 35 °C:

$$C_6H_5Cl + H_2N^- \longrightarrow C_6H_5NH_2 + Cl^-$$

The corresponding reaction of *p*- chlorotoluene gives both *m*- and *p*-aminotoluene

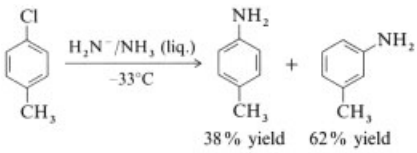

The *o*-isomer is not formed. The two reaction products do not interconvert under the reaction conditions. The fact that the amide ion deprotonates benzene in liquid ammonia and that this proceeds rapidly in the position *ortho* to halide substituents leads to the conclusion that the amide does not behave as a nucleophilic reagent, but as a base. In this reaction, the hydrogen atom in the *o*-position to the substituent chlorine atom is attacked rather than the carbon atom [28].

Reactions with primary and secondary amines are used in the synthesis of pharmaceuticals [29]. Although lower aliphatic amines [5] do not react with sodium amide at room temperature, aromatic amines undergo a metallation reaction with sodium amide in liquid ammonia or benzene:

$$NaNH_2 + H_2NPh \xrightarrow{C_6H_6} NaNHPh + NH_3$$

Molten $NaNH_2$ attacks primary aliphatic amines, forming cyanamide and a mixture of hydrogen and hydrocarbons. According to [5], it is possible to metallate a primary aliphatic amine with an alkali metal amide in toluene at 50 – 100 °C. A large number of secondary and tertiary heterocyclic amines have been synthesized by the reaction of alkyl and heterocyclic halogen compounds with sodium amide [23]. Tertiary amines containing 2-dialkylaminoethyl groups were tested for antihistamine properties. An example of such a synthesis is given below [23]:

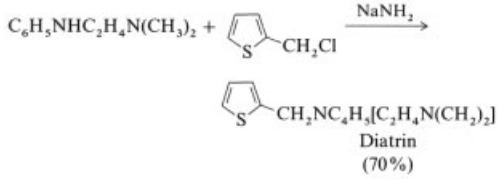

Initiation of Anionic Polymerization. If sodium amide reacts with a suitable alkene (R″ = COOR, CN), the corresponding carbanion is formed, and this initiates the anionic polymerization [26]:

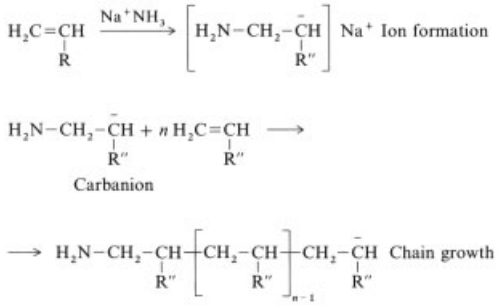

Chain termination is caused by cations [26].

There are detailed accounts of the anionic polymerization of vinyl monomers [37–39] and

of methyl methacrylate [40], [41] in the literature. The preparation of electrically conducting polymers is described in [42, 43].

Amination of Pyridine. The amination of pyridine (Chichibabin reaction) is an example of the nucleophilic reaction properties of sodium amide [44]:

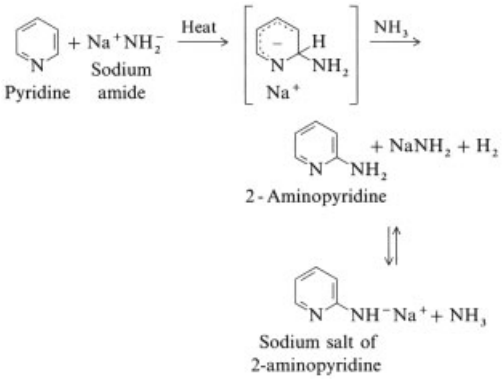

Cleavage of Ketones. The cleavage of ketones by sodium amide is known as the Haller – Bauer reaction [21]:

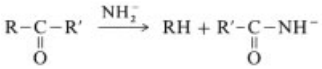

This reaction can normally only be used with ketones that do not enolize. The ketones are mainly of the type $ArCOCr_3$, and the reaction products, R_3CCONH_2, are not readily obtainable by other methods.

3. Production

Sodium amide is produced from ammonia and sodium using the reaction conditions described by BEILSTEIN and GEUTHER [1] for the reaction

$$2\,NH_3 + 2\,Na \longrightarrow 2\,NaNH_2 + H_2$$

It was known even at that time that the following conditions were essential [1]:

1) Absence of air in the ammonia and reactants
2) Minimum possible water content in the ammonia

During the reaction, ammonia is passed from below through the pre-heated molten sodium in a reactor [45]. The reaction temperature is 300 – 450 °C, preferably between 350 and 400 °C. Decomposition of the molten sodium amide [19] can take place in two ways. If the total gas pressure exceeds the dissociation pressure of sodium hydride, the reaction is

$$NaNH_2 \longrightarrow NaH + 0.5\,N_2 + 0.5\,H_2$$

and below this dissociation pressure, it is

$$NaNH_2 \longrightarrow Na + 0.5\,N_2 + H_2$$

In addition to hydrogen and nitrogen the exhaust gas also contains ammonia [19].

In general, two or more reactors containing alkali metal are arranged in series, so that the unreacted ammonia gas from the first reactor passes through reactors downstream, which reduces the ammonia loss. The hydrogen liberated on formation of the sodium amide passes out of the reactor along with the excess ammonia. The gas mixture passes through a water scrubber to remove ammonia, and the hydrogen that remains is burnt. The aqueous solution of ammonia is processed to enable the ammonia to be totally or partially recycled [46].

When indoxyl is produced by the Heumann – Pfleger process [7], a water-free, ammonia- containing gas mixture is obtained which is used to produce sodium amide. The mixture can contain > 50 % hydrogen. The dust generated in the indoxyl process is removed by filtration, and the gas mixture is passed into the molten sodium through a ring-shaped distributor. The formation of sodium amide is complete after 4 h at 340 – 370 °C; the yield is 97 % [46].

Catalysts [22] strongly accelerate the reaction of sodium with ammonia. The most effective catalysts are iron oxides and the chlorides and nitrates of Fe, Co, and Ni, which are soluble in liquid ammonia. The catalysts added in amounts of < 0.001 mol/g Na, are reduced to the metals. The catalytic action of platinum is reported in [22].

The disadvantage of iron or metal salt catalysts is that removal of the catalyst by filtration is difficult. The catalyst is usually left in the reaction product, although this can cause blockages in valves, pipes, and pumps [47]. In a process for producing alkali metal amides from liquid ammonia and alkali metals, this problem is avoided

by using an alkali metal azide as catalyst. The presence of the azide in the liquid ammonia increases the solubility of the alkali metal amide 20–30 fold. An amide yield of 99.3 % is reported [47]. Residual azide catalyst can be separated from the amide by filtration or centrifugation and recycled.

Other methods of preparation, including laboratory methods for the production of high-purity $NaNH_2$, are described in [19], [22], [23].

4. Safety Aspects

Sodium amide is hygroscopic, and decomposes in moist air [17]. The prolonged action of atmospheric oxygen produces oxidation products such as sodium hyponitrite, $Na_2N_2O_2$, sodium hyponitrate, $Na_2N_2O_3$, etc., [48]. Sodium amide must therefore be stored with exclusion of oxygen and moisture. Careless handling and storage can lead to explosions.

If the skin or mucous membranes come into contact with sodium amide, they must be treated with copious quantities of water.

Sodium amide is flammable. For extinguishing fires, water, carbon dioxide, or halons must not be used under any circumstances, but salt, sodium carbonate, or dry sand are suitable. On heating in the absence of air, $NaNH_2$ decomposes at >400 °C, forming sodium, nitrogen, hydrogen, and small amounts of ammonia.

Sodium amide should be used as soon as possible after manufacture. Products with a yellow or brown appearance must not be used, and must be destroyed. In case of doubt concerning the safety of the material, the melting and burning test should be carried out. For this, ca. 1 g sodium amide is heated in a spoon spatula over a flame. If the material does not melt and burn quietly, but with decrepitation, it is not safe to use. For disposal small amounts of sodium amide can be suspended in dioxan [49], and ethanol can then be added dropwise (for larger quantities, isopropanol is added first). When ammonia evolution ceases, the alkoxide formed is carefully decomposed with water and neutralized, and is then washed into the drainage system with large quantities of water [49]. Larger quantities can be disposed of by burning in a thin layer, e.g., on charcoal, or by spraying with water in a large vessel at a safe distance and in a suitable location.

Sodium amide reacts violently with water, forming sodium hydroxide solution and ammonia. The decomposition products, ammonia and sodium hydroxide, can cause very severe caustic effects, with permanent damage, especially to the skin and mucous membranes [17]. The eyes and respiratory tract are particularly affected. Safety precautions must be taken to protect against danger to the health from contact with ammonia [50]. The precautions for NaOH are those appropriate for caustic materials [51].

5. Uses

An important use for sodium amide is the Heumann–Pfleger indigo synthesis (→Indigo and Indigo Colorants, Chap. 4.), in which phenylglycine salts undergo ring closure in an alkaline melt with liberation of ammonia to form indoxyl [7], [46].

In a state-of-the-art process [52], indoxyl sodium is produced continuously by mixing the molten alkali metal salt of phenylglycine with a molten mixture of sodium amide and alkali metal hydroxide at 230–250 °C in a multiple-screw extruder. The temperature of the molten mixture rises to 260–340 °C in the mixing and reaction zone. The liberated ammonia is drawn off from the degassing zone of the extruder, and can be used directly for the production of sodium amide [52].

The synthesis of carotenoids and vitamins A and D involves the Wittig reaction [26], which can be carried out with sodium amide. In the reaction of methylenetriphenylphosphorane or its derivatives with substituents in the methylene group with carbonyl compounds, the carbonyl oxygen is exchanged with the methylene group, so that triphenylphosphine oxide and the corresponding alkene are formed [53]. For example, in the alkaline medium at room temperature, the reaction of methylenetriphenylphosphorane with benzophenone forms 1,1-diphenylethylene and triphenylphosphine oxide [26]:

$$(C_6H_5)_3P{=}CH_2 + O{=}C(C_6H_5)_2 \longrightarrow (C_6H_5)_3P{=}O + H_2C{=}C(C_6H_5)_2$$

When methyltriphenylphosphonium bromide in a suitable solvent is treated with an equivalent of a strong base B (e.g., phenyllithium, sodium ethoxide) the base deprotonates the salt, and an

equilibrium reaction then takes place forming methylenetriphenylphosphorane [53]:

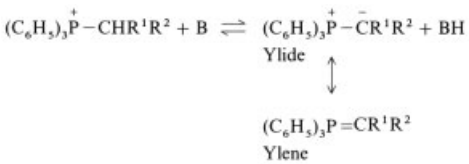

WITTIG et al. [53] suggested that sodium amide in liquid ammonia could be used to form the ylide. The ammonia is then replaced with ether or THF [54].

When lithium alkyls are used as base [55], the corresponding lithium halide is formed, which lowers the stereoselectivity of the reaction. The development of "salt-free" ylide solutions, prepared, for example, by using sodium amide, allowed stereoselective Wittig reactions to be carried out [55]. According to [21], [56], if the base does not contain lithium, the ylide is said to be prepared under salt-free conditions.

It became apparent that most triphenylphosphonium salts with a negative charge on the α-carbon atom react even with weaker bases than phenyllithium, yielding the corresponding ylides, usually in good yields [57], [58]. Suitable proton acceptors include alkali metal alkoxides and sodium amide [57], [59].

When sodium ethoxide or methoxide is used as a proton acceptor [53], [58] with the corresponding alcohol as solvent, the alcoholate anion removes a proton from the phosphonium salt. The methylenetriphenylphosphorane is formed in solution according to the equation:

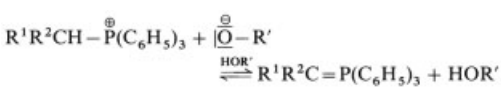

The ylides can be prepared by mixing the phosphonium salt with a stoichiometric amount of powdered sodium amide [21], [60]. The ylide can be liberated by adding a layer of diethyl ether or tetrahydrofuran [52]. The alkene can then be obtained in good yield by reaction with an aldehyde or ketone [60].

Sodium amide is used in the production of pharmaceuticals and intermediates. The papers and patents quoted in [29] can be divided into:

1) Reactions of sodium amide with secondary or primary amino groups
2) Reactions with active C – H groups
3) Addition reactions

1) Phenothiazine (or its derivatives) is first reacted with sodium amide, and then with an organic halogen compound:

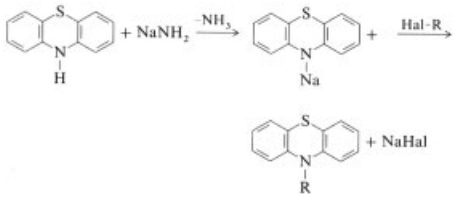

The derivatives of phenothiazine are used as ganglion blockers, local anaesthetics, and particularly antispasmodics [30].

The reaction with other secondary amines, e.g., azepine derivatives, substituted aminopyridines, or anilines to give derived products can be represented as follows [31]:

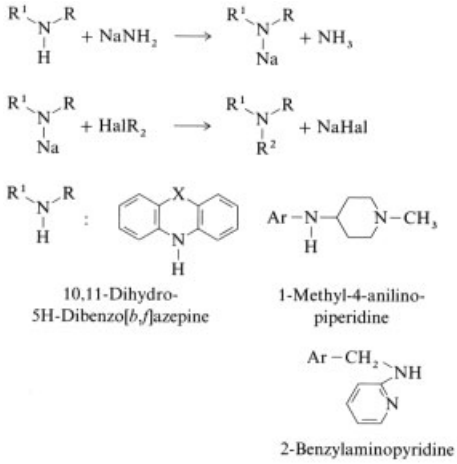

The products derived from the resulting azepine derivatives have antiallergic, sedative, or antispasmodic properties.

Only a few examples exist for the reaction of $NaNH_2$ with primary amines, 2-aminopyridine, or acid amides according to the scheme [32]:

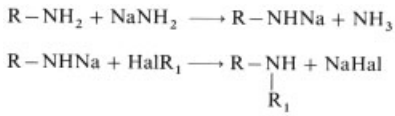

2) There are many reports dealing with the reaction of phenylacetonitrile and its derivatives [23] with $NaNH_2$, sometimes followed by reaction with a halogen compound [33]:

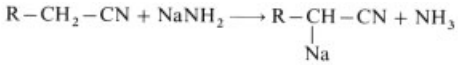

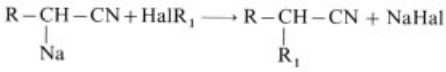

Antitussive, antihistamine, and analgesic effects have been claimed for this product group.

The introduction of another substituent into the CH group of diphenylacetonitrile can be achieved by the following reaction scheme:

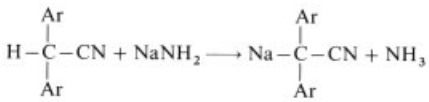

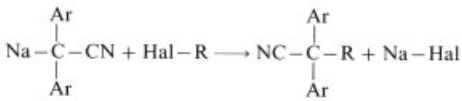

These reactions are mainly used as stages in the manufacture of antispasmodics, analgesics, and antidiarrhoics [34].

The deprotonation of alcohols by sodium amide is the basis for the preparation of antihistamines, antispasmodics, and an antitussive. These reactions take place according to the following scheme [35]:

$R'OH + NaNH_2 \longrightarrow R'ONa + NH_3$

$R'ONa + R''Cl \longrightarrow R'OR'' + NaCl$

3) An addition reaction of 2-chlorobenzophenone with acetonitrile [23] takes place in the presence of $NaNH_2$ as base [4], [36]:

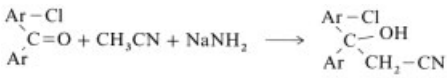

A further use for sodium amide is in halogen – NH_2 exchange reactions in aromatic chemistry [61]. These reactions between halogenated aromatics and alkali metal amides are carried out in liquid ammonia and apparently involve aryne intermediates. For example, when 1-bromo-2,4-dichlorobenzene reacts with $NaNH_2/NH_3$ in the molar ratio 1 : 2 – 1 : 7, 3,5-dichloroaniline is produced in 70 % yield:

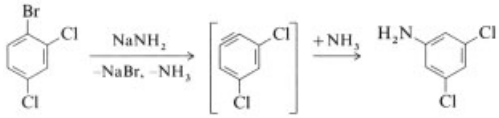

Under comparable conditions, *m*-chloroaniline can be obtained from *o*-dichlorobenzene (yield: >70 %, purity: >98 %) [61].

Sodium amide is also used as a starting material for the production of sodium azide (→Hydrazoic Acid and Azides, Chap. 4.1.) [2]. In this process, powdered $NaNO_3$ is added to molten $NaNH_2$ at 175 °C, or N_2O is passed through molten sodium amide at 190 °C:

$3\,NaNH_2 + NaNO_3 \longrightarrow NaN_3 + 3\,NaOH + NH_3$

$2\,NaNH_2 + N_2O \longrightarrow NaN_3 + NaOH + NH_3$

In the latter reaction, the $NaNH_2$ can be produced in situ from liquid ammonia and sodium [2]. Sodium azide catalyzes the reaction [47].

6. Transport Regulations

During the storage and transportation of sodium amide, moisture and air must be excluded by inertizing with a protective gas (e.g., nitrogen) or protective liquid (e.g., inert hydrocarbons) [62].

The transport regulations (1993) classify sodium amide as follows:

European land transport (RID/ADR):	Class 4.3, No. 19 b.
Sea transport (IMDG Code):	Class 4.3, UN No. 1390 PG.II
Air transport (ICAO-TI/IATA-DGR):	Class 4.3, UN No. 1390 PG. II

Apart from the maximum allowable amounts per package, the packaging regulations for the various carriers have been harmonized. In general, the packages must be UN approved and hermetically sealed against water vapor.

For transport by passenger aircraft (unlike freight-only aircraft), only assembled packages are permitted.

Intermediate bulk containers (IBCs) may not normally be used for the land or air transport of sodium amide, but metal IBCs are permitted for sea transport if they are UN approved and hermetically sealed.

Movable containers (tanks) are not permitted for sodium amide.

7. References

1. *Gmelin,* 8th ed., **System no. 21,** p. 253.
2. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* VCH Verlagsgesellschaft, Weinheim 1988.
3. P. P. Edwards, *Adv. Inorg. Chem. Radiochem.* **25** (1982) 135 – 185. W. Weyl, *Ann. Phys. (Leipzig)* **121** (1864) 601. G. Lepoutre, M. J. Sienko (eds.): "Metal-Ammonia Solutions," Colloque Weyl I, Benjamin, New York 1964. J. J. Lagowski, M. J. Sienko (eds.): "Metal-Ammonia Solutions," Colloque Weyl II, IUPAC Butterworth, London 1970. J. Jortner, N. R. Kestner (eds.): "Electrons in Fluids," Colloque Weyl III, Springer Verlag, Berlin 1973. Colloque Weyl IV, *J. Phys. Chem.* **79** (1975) no. 26. Colloque Weyl V, *J. Phys. Chem.* **84** (1980) 1065.
4. *Ullmann,* 4th ed., **17,** pp. 143 – 145.
5. M. F. Lappert, P. P. Power, R. C. Srivastava: *Metal and Metalloid Amides,* Ellis Horwood, John Wiley & Sons, New York 1980.
6. H. Jacobs, D. Schmidt: "High Pressure Ammonolysis in Solid-State Chemistry," in E. Kaldis (ed.): *Current Topics in Material Science,* vol. 8, North-Holland Publishing Company, Amsterdam 1982, p. 583.
7. G. Dittrich, H. Offermanns, H. Schlosser, Jr., *ChET, Chem. Exp. Technol.* **3** (1977) 151 – 162.
8. Degussa AG: "Im Zeichen von Sonne und Mond," Company brochure, Frankfurt/Main 1993.
9. In ref [1] , p. 257; *Ann. Phys.* (*Leipzig)* **121** (1864) 611; **123** (1864) 350; *Proc. R. Soc.* **21** (1873) 145; *Chem. N.* **22** (1870) 217.
10. S. G. Bratsch, J. J. Lagowski, *J. Phys. Chem.* **88** (1984) 1086 – 1089.
11. U. Schindewolf, *Angew. Chem. Int. Ed. Engl.* **17** (1978) 887 – 901.
12. F. A. Cotton, G. Wilkinson, P. L. Gaus: *Grundlagen der anorganischen Chemie,* VCH Verlagsgesellschaft Weinheim 1990.
13. U. Schindewolf, *Angew. Chem.* **80** (1968) 165 – 208.
14. U. Schindewolf, G. Lang, K. W. Böddeker, *Z. Phys. Chem. Neue Folge* **66** (1969) 86 – 91.
15. W. L. Jolly, "Metal-Ammonia Solutions," *Progr. Inorg. Chem.* **1** (1959) 239 – 240.
16. O. J. Foust: *Sodium-NaK Engineering Handbook,* **vol. 1**, Gordon and Breach, New York 1972.
17. *Ullmann,* 4th ed., **17,** p. 157.
18. In ref [1] , p. 926, suppl. vol. 3.
19. R. Juza, *Angew. Chem.* **76** (1964) 290 – 300.
20. R. Juza, H. H. Weber, K. Opp, *Z. Anorg. Allg. Chem.* **284** (1956) 73. A. Zalkin, D. H. Templeton, *J. Phys. Chem.* **60** (1956) 821.
21. J. March: *Advanced Organic Chemistry,* 3rd ed., John Wiley & Sons, New York 1985.
22. *Gmelin,* **System no. 21 Na, suppl. vol. 3,** p. 947.
23. R. Levine, W. C. Fernelius, *Chem. Rev.* **54** (1954) 449 –573.
24. In ref. [2] , p. 543. W. L. Jolly, C. J. Hallada: "Flüssiger Ammoniak," chap. 1, in T. C. Waddington (ed.): *Non-Aqueous Solvent Systems,* Academic Press, London 1965, pp. 1 – 45.G. W. A. Fowles: "Anorganische Reaktionen in flüssigem Ammoniak," chap. 7, in C. V. Colburn (ed.): *Developments in Inorganic Nitrogen Chemistry,* Elsevier, Amsterdam 1966, pp. 522 – 576.
25. A. Streitwieser, Jr., C. H. Heathcock: *Organische Chemie,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1990.
26. H. Beyer, W. Walter: *Lehrbuch der organischen Chemie,* 22nd ed., Hirzel Verlag, Stuttgart 1991.
27. D. F. Shriver, P. W. Atkins, C. H. Langford: *Anorganische Chemie,* VCH Verlagsgesellschaft, Weinheim 1992.
28. P. Sykes: *Reaktionsmechanismen der organischen Chemie, eine Einführung,* 9th ed., Verlag Chemie, Weinheim 1988.
29. A. Kleemann, J. Engel: *Pharmazeutische Wirkstoffe, Synthese, Patente, Anwendungen,* 2nd ed., Georg Thieme Verlag, Stuttgart 1982.
30. In ref. [29] pp. 5, 8, 25, 42, 128, 199, 252, 320, 334, 383, 423, 468, 505, 523, 564, 577, 595, 670, 676, 690, 703 – 705, 729, 732, 735, 765, 768, 769, 776, 779, 847, 876, 879, 880, 919, 920.
31. In ref. [29] pp. 59, 73, 94, 192, 194, 224, 231, 273, 291, 316, 395, 485, 562, 658, 868, 908, 928.
32. In ref. [29] pp. 126, 773, 835.
33. In ref. [29] pp. 116, 180, 196, 295, 296, 332, 373, 377, 482, 499, 697, 702, 707, 713, 939, 940.
34. In ref. [29] pp. 311, 328, 337, 396, 504, 573, 650, 786.
35. In ref. [29] pp. 78, 198, 206, 217, 234, 341, 437, 613.
36. In ref. [29] p. 226.
37. P. Caubere, S. Raynal, G. Ndebeka, S. Lecolier, *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **23** (1982) no. 1, 190 – 191.

38. P. Caubere, S. Raynal, G. Ndebeka, S. Lecolier, *Polymer* **23** (1982) no. 1, 190 – 191.
39. S. Raynal, S. Lecolier, G. Ndebeka, P. Caubere, *Polymer* **23** (1982) no. 2, 283 – 286.
40. S. Raynal et al., *J. Macromol. Sci. Chem.* **A 18** (1982) no. 3, 313 – 322.
41. S. Raynal, W. Bergeret, J. Sledz, *J. Polym. Sci. Polym. Chem. Ed.* **21** (1983) no. 8, 2601 – 2605.
42. BASF, DE-OS 3 118 630, 1982 (H. Naarmann, K. Penzien, J. Schlag, P. Simak).
43. BASF, EP-A 44 411, 1982 (V. Muench, H. Naarmann, K. Penzien).
44. R. T. Morrison, R. M. Boyd: *Organic Chemistry,* 4th ed., Allyn and Bacon, Boston 1983.
45. Degussa, DE 117 623, 1900; DE 316 137, 1914; US 1 359 080, 1915.
46. BASF, DE 26 38 223 C 2, 1982 (E. Tolksdorf).
47. Canadian Industries Ltd., OT 24 00 907, 1977 (P. J. Astrauskas, G. M. Blondin, S. H. Chen, B. Jung).
48. L. Bretherick, Sunbury Research Centre British Petroleum Comp.: *Handbook of Reactive Chemical Hazards,* Butterworth, London 1979, p. 956. F. W. Bergstrom et al., *Chem. Rev.* **12** (1933) 61. G. R. Krüger et al., *Inorg. Synth.* **8** (1966) 15. R. N. Shreve et al., *Ind. Eng. Chem.* **32** (1940) 173. S. Sandor, *Munkavédelem* **6** (1960) 4 – 6, 20.
49. D. Martinetz: *Immobilisation, Entgiftung und Zerstörung von Chemikalien,* Verlag Harri Deutsch, Frankfurt/Main 1981, pp. 220, 232.
50. Sicherheitsmaßnahmen und Gesundheitsgefahren beim Umgang mit wasserfreiem Ammoniak, Berufsgenossenschaft der Chemischen Industrie, Heidelberg 1991. Bundesanstalt für Arbeitsschutz (Wirtschaftsverlag, NW) (ed.). Schriftenreihe der Bundesanstalt für Arbeitsschutz – Gefährliche Arbeitsstoffe – GA 32, Gasatlas: Ammoniak C-4 bis C-7, Verlag für Neue Wissenschaft, Bremerhaven 1989.
51. Merkblatt M 004, 3/92, ZH 1 – 229; M 053, 12/91, Berufsgenossenschaft der Chemischen Industrie, Heidelberg 1982.
52. BASF, DE 26 43 877 A 1, 1976 (K. Eger et al.).
53. U. Schöllkopf, *Angew. Chem.* **71** (1959) 260 – 273.
54. G. Wittig, H. Eggers, P. Duffner, *Justus Liebigs Ann. Chem.* **619** (1958) 10.
55. In ref. [26] p. 178; H. J. Bestmann, O. Vostrowsky, "Selected Topics of the Wittig Reaction in the Synthesis of Natural Products," *Top. Curr. Chem.* **109** (1983) 85.
56. H. J. Bestmann, *Angew. Chem. Int. Ed. Engl.* **4** (1965) 586.
57. H. Pommer, *Angew. Chem.* **72** (1960) 811.
58. H. Pommer, P. C. Thieme: *Top. Curr. Chem.* **111** (1983) 109, 165 – 188.
59. BASF, DE 1 003 730, 1954; 943 648, 1954 (G. Wittig, H. Pommer). BASF, DE 1 026 745, 1956 (H. Pommer, G. Wittig, W. Saruecki).
60. M. Schlosser, B. Schaub, *Chimia* **36** (1982) 396.
61. S. Rittner, K. Warning in H. Harnisch, R. Steiner, K. Winnacker (eds.): *Chemische Technologie,* 4th ed., **vol. 6,** Carl Hanser Verlag, München 1982, p. 200.
62. R. Neureiter, Degussa AG, Company brochure Frankfurt/Main 1993.

Sodium and Sodium Alloys

Alfred Klemm, Max Planck Institut für Chemie (Otto Hahn Institut), Mainz, Federal Republic of Germany

Gabriele Hartmann, Degussa AG, Frankfurt am Main, Federal Republic of Germany

Ludwig Lange, Degussa AG, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Properties 2
3. Occurrence 7
4. Production 7
4.1. Chemical Processes 7
4.2. Electrolytic Processes 7
4.2.1. Development of Molten Salt Electrolysis of NaOH and NaCl 7
4.2.2. Description and Operation of NaCl Molten Salt Electrolytic Cell 9
4.2.3. Properties of the Electrolyte 13
4.2.4. Other Electrolytic Processes 13
5. Safety 14
6. Uses 15
7. Sodium Alloys 17
8. Packaging and Transport 19
8.1. Forming Sodium Metal 19
8.2. Transport Regulations 19
8.3. Break-Bulk Packaging 19
8.4. Portable Tanks 21
9. Quality Specifications and Analysis 21
10. Health and Safety 22
11. Economic Aspects 22
12. References 22

1. Introduction

Sodium is the best-known alkali metal. The amount of sodium metal produced industrially exceeds that of all other alkali metals. It is a member of group 1 of the periodic table, with the relative atomic mass 22.9898 and the atomic number 11.

Sodium has only one naturally occurring isotope, ^{23}Na. Several artificial isotopes are also known [1]. The isotope ^{24}Na is used as a tracer in industrial and medicinal applications.

Few basic chemicals are used in as many areas as sodium metal. Elemental sodium can be considered both a fine chemical whose properties make it indispensable for industrial applications ranging from pharmaceuticals to heavy inorganic chemicals, and a readily liquefied metal with a wide range of metallurgical and physical uses. A large number of products derived from sodium metal are used in the chemical industry.

History. The German name *Natrium* is derived from the Egyptian *neter,* meaning sodium carbonate [2–4], while the English name sodium is derived from soda [3]. Sodium was discovered in 1807 by Davy, who electrolyzed molten sodium hydroxide in a platinum dish, by contacting it with a platinum wire using a voltaic pile as the source of electricity [3].

However, this electrolytic process could not be used for industrial production at that time, because no economic source of direct current was available. Powerful d. c. generators first became available after the discovery of the dynamo principle by W. von Siemens in 1867.

The history of the industrial production of sodium, which extends over more than 100 years, can be divided into four periods, in which four different processes were used [5].

Thermochemical reduction processes were used between 1854 and 1890. A mixture of sodium carbonate, wood charcoal or coal, and quicklime was heated to $>1100\,^{\circ}$C, and the sodium vapor formed was condensed (Sainte. Claire Deville, Paris, 1854). An annual production of 5 – 6 t of sodium was achieved by this process. In a later development, sodium carbonate was replaced by sodium hydroxide, and the wood charcoal or coal by a mixture of iron and coal (Castner, 1886). Using this process, the Aluminium Company Limited in Oldbury produced ca. 150 t/a sodium between 1888 and 1890 [4].

Between 1891 and 1920, sodium was produced by electrolysis of molten sodium hydrox-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_277

ide, to which sodium chloride and sodium carbonate were added to lower the melting point (Castner process). The Castner process [6] (see Section 4.2.1) was the only important industrial process for producing sodium during this period.

In the 1900s, attempts were made in Switzerland (Ciba), Germany (Degussa), and the United States (Rössler & Hasslacher Chemical Co.) to use the cheaper starting material sodium chloride in a molten salt electrolysis process. In 1921, Downs succeeded in constructing a cell for electrolyzing sodium chloride [7], the basic principle of which is still used today. The electrolyte used from ca. 1925 onward for industrial production was a mixture of sodium chloride and calcium chloride. Since ca. 1950, a modified Downs cell has been used, with an electrolyte consisting of a ternary mixture of NaCl, $BaCl_2$, and $CaCl_2$. This process has since become established worldwide (see Section 4.2.2).

The industrial use of sodium has also undergone a change. The only demand for large-scale production of sodium in the mid-1800s came from the aluminum industry, where it was used for reducing aluminum trichloride. However, around 1890, aluminum producers began to switch to a molten salt electrolysis process. Sodium was also used in the production of sodium cyanide (for extracting gold from its ores) and sodium peroxide. It was later used in the production of lead- containing antiknock agents for gasoline and special metals such as titanium.

2. Properties

Physical Properties. The properties of sodium are directly related to its structure [8]. The electronic configuration of sodium is $1s^2\,2s^2\,2p^6\,3s^1$. The atomic volume is therefore large, and the cohesive forces in the sodium lattice are relatively low. Thus, like all alkali metals, sodium has a lower density than the alkaline-earth metals, as well as lower hardness; lower melting and boiling points; and lower energies of sublimation, vaporization, and dissociation [5], [9].

Sodium crystallizes from the melt with a body- centered cubic crystal lattice. It has typical metallic properties, including high thermal and electrical conductivity, the electrical conductivity of sodium at 20 °C being about one-third that of copper. It is liquid over a wide temperature range from 97.8 to 881.4 °C and has a fairly low vapor pressure, which is an important requirement for production in an electrolytic cell [10].

Alkali metals give characteristic flame coloration. Their outer electrons can easily be raised to an excited state, which is the basis for their analytical determination by flame photometry and atomic absorption spectroscopy. Sodium and sodium salts color the flame of a Bunsen burner yellow. The sodium D doublet at 589 and 589.6 nm is due to the $3s^1 - 3p^1$ transition in sodium atoms formed by the reduction of Na^+ in the flame [9].

Sodium is monatomic in the vapor state, although dimers and tetramers have also been reported [10].

Some physical properties of sodium are listed below [8], [10]:

Atomic number	11
Relative atomic mass	22.98977
Melting point	97.83 °C
Heat of fusion	113 J/g
Volume expansion on melting	2.7 %
Boiling point	881.4 °C
Heat of vaporization	3750 J/g
Atomic radius in metallic lattice	189.6 pm
Ionic radius	95 – 98 pm
Ionization potential	5.139 V
Standard electrode potential	2.714 V
Electrical resistivity at 293 K	$4.77 \times 10^{-8}\ \Omega \cdot m$
Critical temperature	2508.7 ± 12.5 K
Critical pressure	25.64 ± 0.02 MPa
Critical density	206 kg/m^3
Specific heat capacity at 293.15 K	1.22 J g^{-1} K^{-1}
Thermal conductivity at 293.15 K	132.3 W m^{-1} K^{-1}
Density	
at 293.15 K	0.968 g/cm^3
at 393.15 K	0.927 g/cm^3

The density (kilograms per cubic meter) of solid sodium in the range 273 – 370.98 K is given by the following equation [8]:

$$\varrho_T = 972.5 - 20.11 \times 10^{-2}\,(T - 273.15) \times 10^{-4}\,(T - 273.15)^2$$

For liquid sodium in the range 370.98 – 1644.24 K, the following equation applies [8]:

$$\varrho_T = 1011.8 - 0.22054T - 1.9226 \times 10^{-5}T^2 + 5.637 \times 10^{-9}T^3$$

The specific heat capacity ($J\,g^{-1}\,K^{-1}$) as a function of temperature at 273.15 – 370.98 K is given by [8]

$$c_p = 1.1987 + 6.4894 \times 10^{-4}\,(T - 273.15) + 1.053 \times 10^{-5}\,(T - 273.15)^2$$

In the range 370.98 – 1173.15 K, the following equation applies [8]

$$c_p = 1.4361 - 5.8024 \times 10^{-4}\,(T - 273.15) + 4.621 \times 10^{-7}\,(T - 273.15)^2$$

The thermal conductivity ($W\,m^{-1}\,K^{-1}$) from 370.98 to 1173.15 K is given by [8]

$$\lambda = 91.8 - 4.9 \times 10^{-2}\,(T - 273.15)$$

The vapor pressure (kilopascals) at 300 – 2500 K is [8]

$$\log p = 10.182516 - (5693.8776/T) - (1.0948 \log T) + 8.5874946 \times 10^{-5}T$$

The viscosity ($mPa \cdot s$) at < 773 K is given by [8]

$$\eta = (0.1235 \pm 0.0018) \cdot (\varrho/1000)^{1/3} \cdot \exp\frac{(697 \pm 9)}{1000T}$$

Above 773 K, the equation is [8]

$$\eta = (0.0815 \pm 0.0013)\,(\varrho/1000)^{1/3} \cdot \exp\frac{(1040 \pm 19)}{1000T}$$

The surface tension (newtons per meter) in the range 370.98 – 1300 K is given by [8]

$$\sigma = 206.7 \times 10^{-3} - 1.0 \times 10^{-4}\,(T - 273.15)$$

Figure 1 shows some properties of sodium as a function of temperature.

Chemical Properties. The reactivity of sodium can be illustrated by the following equations [12], [13]:

$$2\,Na + H_2 \rightleftharpoons 2\,NaH$$

$$4\,Na + TiCl_4 \rightleftharpoons Ti + 4\,NaCl$$

$$2\,Na + 2\,NH_3 \longrightarrow 2\,NaNH_2 + H_2$$

$$2\,Na + O_2 \longrightarrow Na_2O_2$$

$$2\,Na + 4\,S \longrightarrow Na_2S_4$$

$$2\,Na + X_2 \longrightarrow 2\,NaX\ (X = \text{halogen})$$

$$Na + ROH \longrightarrow RONa + 1/2\,H_2$$

$$2\,Na + 2\,RX \longrightarrow R{-}R + 2\,NaX$$

(X = halogen, R = alkyl group)

The high reactivity of sodium is a result of its electronic configuration [14]. Sodium, like the other alkali metals, has only one *s* electron in its outer shell, which is very weakly bound due to shielding from the atomic nucleus provided by the completely filled rare gas shells.

1) Sodium reacts with hydrogen at ca. 200 – 300 °C to form sodium hydride. In NaH, hydrogen is present as hydride ion. In organic syntheses, sodium hydride has little reducing action [15], its most important chemical property being high basicity. Sodium hydride reacts as a strong base not only with all acids, but also with all organic and inorganic compounds that contain an active hydrogen atom. Sodium hydride is more strongly basic than sodium amide [16]. However, in molten NaOH or eutectic salt baths, NaH behaves as a strong reducing agent toward metallic salts and oxides [17]. The system $Na - H_2 - NaOH - Na_2O$ has been investigated extensively [18], [19].
2) Titanium was first produced in large quantities in 1910 by HUNTER, who reacted titanium tetrachloride with sodium in an evacuated steel bomb. In 1956, ca. 1500 t of titanium sponge, representing ca. 10 % of the total world output, was produced by ICI by reducing titanium tetrachloride with sodium [20]. Also, Reactive Metals, Inc., of Ashtabula, Ohio, until recently, used sodium as the reducing agent in the production of titanium. However, by far the largest amount of titanium has been produced by the magnesium-based Kroll process. More recently, the electrolysis of $TiCl_4$ has become important.

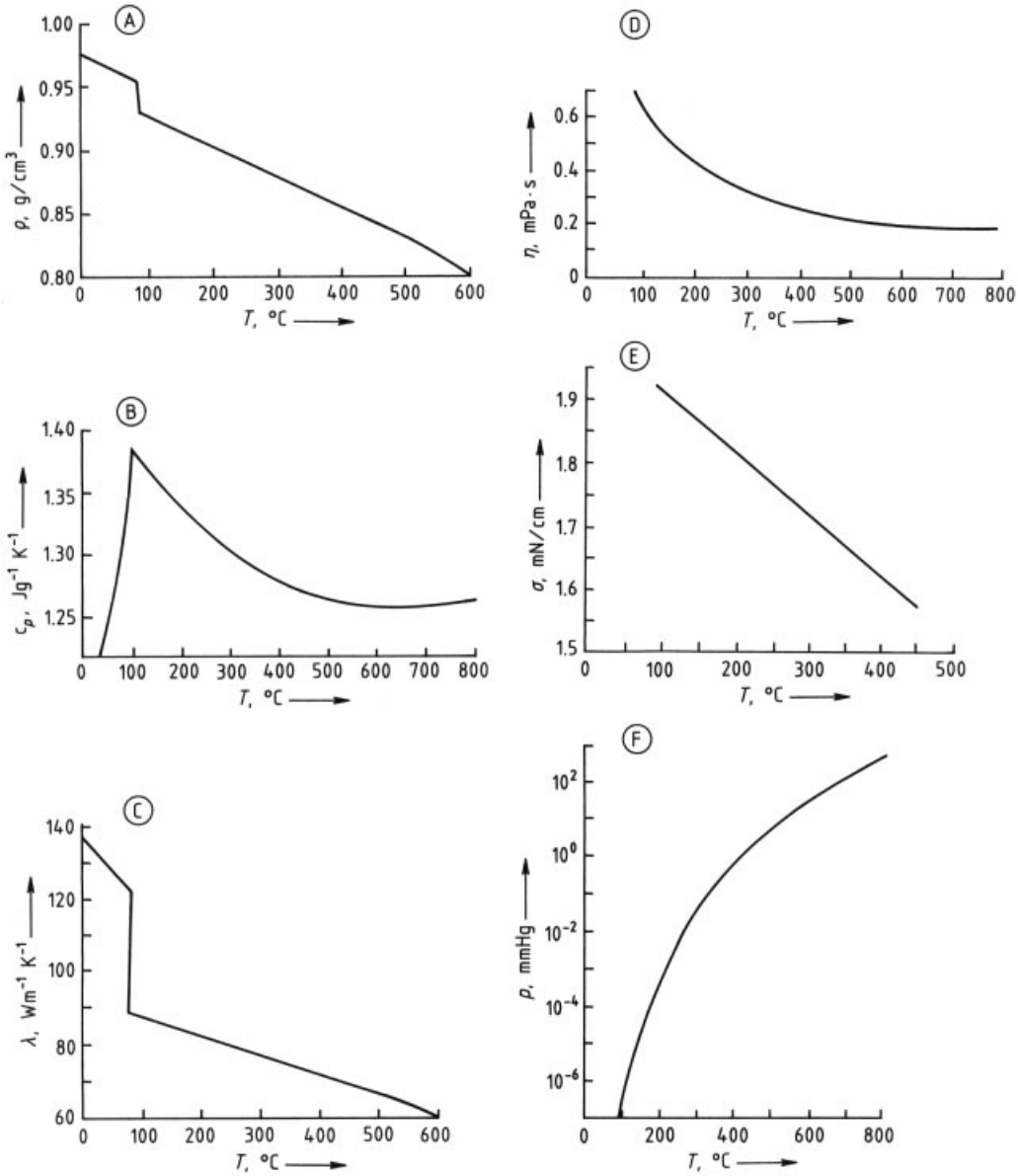

Figure 1. Some properties of sodium as a function of temperature [11]
A) Density; B) Specific heat capacity; C) Thermal conductivity; D) Viscosity; E) Surface tension; F) Vapor pressure

The Hunter process is representative of many reactions in which the halides of most of the metals in periods 4 – 6 of the periodic table can be reduced with sodium to the metal, with formation of the corresponding sodium salt. Examples include the production of aluminum by using sodium as the reducing agent (Deville and Bunsen process, 1854) [4] and of thorium from thorium tetrachloride [4], [21–23] or K_2ThCl_6 [24]. The elements potassium, beryllium, magnesium, boron, and aluminum can also be produced by this method. The reduction of KCl by sodium appears to go against the normal order of reactivity (K > Na). At 850 – 880 °C, the following reaction takes place:

$$Na\,(gas) + K^+\,(liquid) \rightleftharpoons Na^+\,(liquid) + K\,(gas)$$

Since potassium is more volatile, it is more readily distilled, so the equilibrium is shifted to the right and the reaction proceeds [9].

3) Sodium amide is produced industrially by the reaction of molten sodium with ammonia at 300 – 400 °C.
Sodium dissolves in liquid ammonia to give a blue color that changes to a coppery red with increasing concentration. Solutions of alkali metals in liquid ammonia are powerful and selective reducing agents. Aromatic systems are reduced smoothly to cyclic mono- or dialkenes, and alkynes are stereospecifically reduced to *trans*-alkenes [9].
Sodium does not react with molecular nitrogen under normal conditions. Hence, nitrogen is a suitable protective gas, even at high temperature, for the production, storage, and transport of sodium [10].
Sodium azide, NaN_3, can be prepared by the reaction of dinitrogen oxide with sodium amide at 200 °C. The existence of Na_3N is very doubtful. The heavier alkali metals do not appear to form compounds of this type, possibly for steric reasons [9].
Prior to the 1960s (i.e., before modern catalytic methods for producing hydrogen cyanide became available [12]), sodium cyanide was produced via sodium amide and sodium cyanamide by the Castner process [12]:

$$\begin{array}{l} 2\,Na + 2\,NH_3 \longrightarrow 2\,NaNH_2 + H_2 \\ 2\,NaNH_2 + C \longrightarrow Na_2N_2C + 2\,H_2 \\ Na_2N_2C + C \longrightarrow 2\,NaCN \\ \hline 2\,Na + 2\,NH_3 + 2\,C \longrightarrow 2\,NaCN + 3\,H_2 \end{array}$$

Metallic silver and gold form complexes with NaCN in a smooth reaction under mildly oxidizing conditions. This is used in the extraction of these metals from ores or gold-bearing sand [12].
With carbon, sodium forms compounds of the acetylide type, M_2C_2, and with graphite it forms a layered compound [8]. Sodium acetylide can be produced by the reaction of sodium with acetylene or by the thermal decomposition of $NaHC_2$ in a vacuum. Sodium acetylide can be used in the preparation of alkynes.
The reaction between graphite and sodium gives an intercalation compound with the approximate formula $C_{64}Na$ [8].
Sodium formate or oxalate can be produced in the reaction between carbon dioxide and sodium. Burning sodium reacts explosively with solid carbon dioxide; thus CO_2 extinguishers are unsuitable for sodium fires [10]. Alkali metals in liquid ammonia react with carbon monoxide to form colorless crystals of $Na_2C_2O_2$, which contain linear groups, NaOC≡CONa, linked together to form chains [9].

4) Sodium reacts with oxygen to form the following oxides: monoxide (Na_2O), peroxide (Na_2O_2), and hyperoxide (NaO_2). The ozonide NaO_3 is also known [25].
The reactions to form the monoxide and the peroxide are exothermic. The peroxide is manufactured in a rotating drum reactor. In the first stage, sodium reacts with air to form sodium monoxide, which reacts with oxygen or oxygen-enriched air in the second stage to form sodium peroxide [25].
The structure and chemical properties of the oxygen compounds of alkali metals have been investigated thoroughly [26].
In normal atmospheric air, molten sodium can ignite at >115 °C, but higher temperature is necessary in dry air. Combustion is exothermic, with the evolution of much smoke, and an intense white-yellow light is emitted with the formation of Na_2O and Na_2O_2 [10].
Sodium reacts with water to form sodium hydroxide and hydrogen. If the hydrogen evolved mixes with air, the gas mixture formed can explode with great violence [10].
The reaction of excess sodium with high-pressure steam in the absence of air gives sodium hydroxide and sodium hydride [8].
The heats of reaction of the individual reaction steps are given below [8]:

$$Na_{(l)} + H_2O_{(l)} \longrightarrow NaOH_{(s)} + 1/2\,H_2,$$

$$\Delta H^0_{298} = -145\,kJ/mol$$

$$Na_{(l)} + H_2O_{(g)} \longrightarrow NaOH_{(s)} + 1/2\,H_2,$$

$$\Delta H^0_{298} = -185.4\,kJ/mol$$

$$2\,Na_{(l)} + H_2O_{(g)} \longrightarrow NaOH_{(s)} + NaH_{(s)},$$

$$\Delta H^0_{298} = -\,195.3\ \text{kJ/mol}$$

$$Na_{(l)} + 1/2\,H_2 \longrightarrow NaH_{(s)},$$

$$\Delta H^0_{298} = -\,56.4\ \text{kJ/mol}$$

$$Na_{(l)} + NaOH_{(s)} \longrightarrow Na_2O + 1/2\,H_2,$$

$$\Delta H^0_{298} = +\,10.7\ \text{kJ/mol}$$

$$2\,Na_{(l)} + H_2O_{(g)} \longrightarrow Na_2O_{(s)} + H_2,$$

$$\Delta H^0_{298} = -\,128.1\ \text{kJ/mol}$$

$$2\,Na_{(l)} + NaOH_{(s)} \longrightarrow Na_2O_{(s)} + NaH_{(s)},$$

$$\Delta H^0_{298} = -\,45.7\ \text{kJ/mol}$$

5) Molten sodium reacts with stoichiometric amounts of molten sulfur under an inert atmosphere to form disodium tetrasulfide.
$2\,Na + 4\,S \longrightarrow Na_2S_4$
Disodium tetrasulfide is used industrially for the production of organofunctionalized silanes [27] by reaction with 3-chloropropyltrialkoxysilanes. In a nucleophilic reaction, the tetrasulfide anion displaces the chlorine atoms and links two propyltrialkoxysilane molecules with formation of a sulfur bridge [27].

$$2\,(RO)_3{-}Si{-}(CH_2)_3{-}Cl + Na_2S_4 \longrightarrow (RO)_3Si{-}(CH_2)_3{-}S_4{-}(CH_2)_3{-}Si(OR)_3 + 2\,NaCl$$

6) Halogens react with sodium with varying degrees of vigor. With fluorine, ignition takes place at room temperature, but no reaction occurs in liquid chlorine. With gaseous, dry chlorine, negligible reaction takes place at room temperature, but molten sodium burns in a stream of chlorine to form sodium chloride. This reaction takes place to a small extent in the Downs cell. Sodium is fairly stable toward bromine. Sodium and iodine can be melted together without reaction [10].

7) Lower and primary alcohols react readily with sodium to form alkoxides. Higher and branched alcohols generally react very slowly due to the decreasing acidity of the hydroxyl group. The reaction velocity can be increased by raising the temperature and by using finely divided sodium. Alkoxides are condensation agents of high basicity, used in the alkylation of carbon, nitrogen, and oxygen atoms (e.g., the mono- and dialkylation of malonic esters, acetoacetic esters, cyanoacetic esters, benzyl nitrile) and for base- catalyzed ring closure reactions to form nitrogen- containing heterocyclic compounds. Alkoxides are used as synthetic building blocks in the Williamson ether synthesis [11].

8) The reactions of alkali metals with alkyl halides (Wurtz reaction) are heterogeneous and probably very complex [28]. Alkanes are formed when the alkylsodium compound initially formed displaces the halogen atom from an alkyl halide molecule:

$$R{-}X + 2\,Na \longrightarrow R^-Na^+ + NaX$$

$$R^- + RX \longrightarrow R{-}R + X^-$$

Organometallic compounds of group IV A metals can be prepared by a variation of this synthetic method. For example, tetraethyllead can be prepared by reacting chloroethane with a lead sodium alloy [29]:

$$4\,C_2H_5Cl + 4\,Na{-}Pb \longrightarrow (C_2H_5)_4Pb + 3\,Pb + 4\,NaCl$$

3. Occurrence

Sodium is strongly electropositive, very reactive, and does not occur in nature in the elemental state, but always in cationic form in salts. Sodium is the seventh element in order of abundance in the Earth's crust (2.27 wt %). It occurs in several minerals in association with other alkali metals. Extensive deposits of relatively pure sodium salts were formed by the evaporation of prehistoric lakes, a process that can still be observed [e.g., in the Great Salt Lake (Utah) and the Dead Sea].

Industrially important naturally occurring sodium compounds include rock salt (sodium

chloride), trona (sodium carbonate), Chile saltpeter (sodium nitrate), mirabilite (sodium sulfate), and borax or kernite (sodium borate). Sodium chloride (NaCl) is the most used raw material in the inorganic chemical industry [9].

4. Production

4.1. Chemical Processes

Sodium can be obtained from many of its compounds by high-temperature reduction [4]. In the older thermochemical methods, sodium carbonate or sodium hydroxide was reacted with charcoal or coal, sulfur, or iron. Other compounds of sodium, such as the chloride, sulfide, sulfate, and cyanide were also used. A serious disadvantage of thermochemical methods was that reverse and secondary reactions occurred. Despite these problems, patents were applied for, covering many variations of the thermochemical methods of production [10]. In a Du Pont patent, sodium carbonate or sodium hydroxide was reduced with coal. The gases formed in the reaction were passed through molten tin, which absorbed the sodium vapor. Sodium was then recovered from the tin in an additional process step. According to a patent of Pechiney S. A., NaCl vapor is passed over calcium carbide at 800 – 1000 °C, and the sodium formed is collected in oil. In a Dow Chemical Co. process, sodium carbonate is reduced with coal at > 1200 °C. The sodium vapor formed is condensed in liquid sodium [10].

4.2. Electrolytic Processes

4.2.1. Development of Molten Salt Electrolysis of NaOH and NaCl

Using the same principle by which Davy discovered sodium, Castner developed a process for producing the metal on a large scale from molten sodium hydroxide in an electrolytic cell (Fig. 2). This process was operated at Niagara Falls (United States) and Rheinfelden (Germany), and was the only practical method for sodium production from 1881 to the mid-1920s. Cheap hydroelectric power was available at both locations.

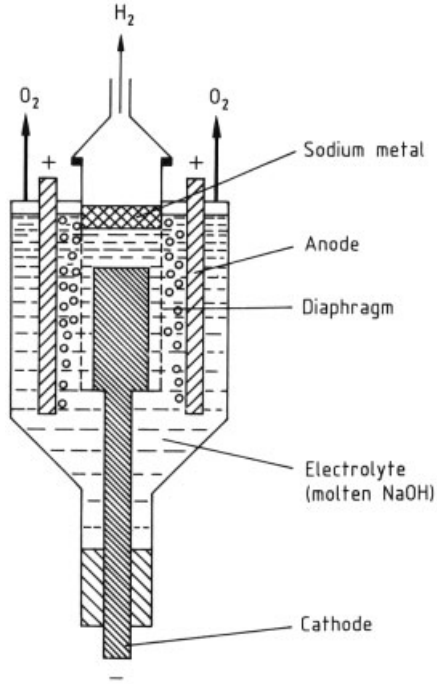

Figure 2. Production of sodium by the Castner process (1891)

The Castner process yields sodium, hydrogen, and oxygen:

$$4\,Na^+ + 4\,e^- \longrightarrow 4\,Na \text{ (cathode)}$$

$$4\,OH^- \longrightarrow 2\,H_2O + O_2 + 4\,e^- \text{ (anode)}$$

$$2\,H_2O + 2\,Na \longrightarrow 2\,NaOH + H_2$$

The overall reaction is:

$$4\,NaOH \longrightarrow 2\,Na + 2\,NaOH + H_2 + O_2$$

Thus the current efficiency cannot exceed 50 %. In practice, current efficiencies of <45 % were achieved.

Before sodium can be produced by this method, sodium hydroxide must first be obtained (e.g., by electrolysis of aqueous sodium chloride). Much research was therefore aimed at producing sodium by direct electrolysis of sodium chloride, which has the following advantages over sodium hydroxide as a starting material:

1) Higher current efficiency during sodium production

2) Cheaper raw material
3) Easier handling
4) Production of chlorine as byproduct

Its disadvantages are:

1) The need to install a plant for chlorine purification and sometimes also for chlorine liquefaction
2) Increased technical problems associated with chlorine
3) Higher melting point

Many cells were developed at this time, but only the Downs cell solved the problems encountered with other cells. With the Downs cell, the foundations were laid for large-scale sodium production. Along with the Hall aluminum cell and the Dow magnesium cell, this is one of the most successful molten salt electrolysis cells ever developed [30].

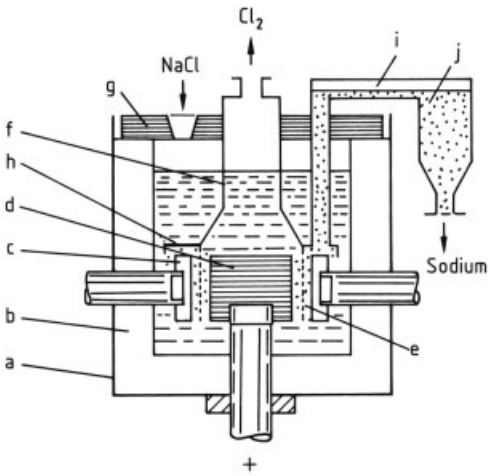

Figure 3. The Downs cell (original form)
a) Steel vessel; b) Ceramic lining; c) Iron cathode; d) Graphite anode; e) Wire gauze diaphragm; f) Dome for collecting chlorine gas; g) Insulating lid; h) Channel with riser pipe for collecting sodium; i) Sodium overflow; j) Sodium holding vessel

In the Castner cell (Fig. 2) a cylindrical nickel anode surrounds the copper cathode concentrically. If NaOH is replaced by a molten salt containing NaCl, the collection and containment of the chlorine evolved cause problems, which are solved only when the positions of the anode and cathode are reversed, as in the Downs cell in which the anode is surrounded concentrically by the cathode.

Figure 3 is a schematic of the Downs cell in its original form [13], [31]. It consists of a lined steel vessel in which the graphite anodes pass through the floor, and the two electrical conductors connected to the cathodes pass through the walls. The anode is surrounded by the cylindrical iron cathode.

The Downs process used a mixture of 60 wt % $CaCl_2$ and 40 wt % NaCl at an operating temperature of 560 – 585 °C. This composition represents a compromise between the melting point of the salt mixture and the sodium content. The eutectic mixture, which contains 66.8 wt % $CaCl_2$ and 33.2 wt % NaCl, melts at ca. 505 °C. Although increasing the $CaCl_2$ content reduces the melting point of the binary mixture, it also decreases the concentration of sodium in the cathode space due to the equilibrium

$$2\,Na + CaCl_2 \rightleftharpoons 2\,NaCl + Ca$$

The operating temperature of ca. 580 °C is well below the temperature of maximum solubility of sodium in the molten salt. The density of the salt is 1.89 – 1.94 g/cm^3. Along with sodium metal, calcium (ca. 3.5 – 4 wt %) is deposited at the cathode. Because of the high melting point of calcium (804 °C) and its low solubility in sodium (see Fig. 6), the alloy phase Na – Ca can accumulate in the sodium and cause blockages in the equipment for continuous removal of sodium from the cell. These solid deposits must be scraped off the walls of this equipment. Most of the calcium can react with NaCl in the melt (by the above reversible reaction), again forming calcium chloride. A further quantity of calcium is removed from the sodium by filtration at 110 °C. The sodium purified in this way contains <400 ppm calcium. Some of the early cell types used for sodium production (e.g., the Castner, Ciba, McNitt, and Downs cells) are described in detail in [4], [13].

4.2.2. Description and Operation of NaCl Molten Salt Electrolytic Cell

Large sodium production plants have 40 – 50 d. c. electrolytic cells connected in series to achieve optimum efficiency. The voltage of each cell depends on the operating conditions (current, temperature, melt composition, sodium production rate) and changes with time.

Electrolytic cells are usually operated at up to 45 kA, with outputs of up to 800 kg/d sodium. For problem-free operation, the individual electrolytic cells, with their associated equipment, must be electrically insulated from neighboring cells and from the ground. The graphite used to make the anodes in the electrolytic cell is resistant to nascent chlorine in the chloride melt even at high temperature, and its electrical conductivity is adequate at the operating temperature. However, all oxygen- containing compounds must be removed from the salts used to prevent corrosion of the graphite. Salts of heavy metals and salts of aluminum and magnesium also cause problems in molten salt electrolysis.

Because of the requirement for high-purity raw materials, extensive facilities are required to purify the salt used in the electrolytic process [5]. Sodium chloride is first dissolved in water to give a saturated brine. Barium chloride and sodium carbonate are added, mainly to remove $CaSO_4$. The precipitate of $BaSO_4$ and $CaCO_3$ forms a sediment together with clay minerals present in the crude salt. The filtered solution is then evaporated to crystallize the salt, which is recovered by centrifugation and dried to a water content of <600 ppm, which is important for optimum operation of the electrolytic cells [5]. A process for the purification of sodium chloride by recrystallization alone, without the use of precipitating agents, has been reported [32]. A schematic of a sodium production process is given in Figure 4.

At the melting point of sodium chloride (808 °C), chlorine would severely corrode all the cell components that come in contact with it. Also, the vapor pressure of sodium is high at this temperature (50 kPa), and its solubility in molten salt is also high (4.2 wt %). Therefore lower melt temperatures must be used. A ternary salt mixture is normally employed ($NaCl - BaCl_2 - CaCl_2$). The temperature of the molten salt in the cell is ca. 600 °C.

From this molten salt mixture, ca. 1 % calcium is deposited on the cathode with the sodium. The crude sodium removed from the cell also contains 0.3 wt % calcium oxide, 0.3 wt % sodium oxide, and small quantities of entrained molten salt. A product of marketable quality [10] is made by cooling the sodium to ca. 120 °C and removing the impurities that separate with the aid of a candle filter equipped with gauze. The filter cake obtained consists of calcium, sodium, chlorides, and oxides. Various methods of recovering calcium or calcium alloys from the filter cake have been proposed, but none has had commercial success [1], [33], [34].

In modern cells (Fig. 5) the cathode and anode are subdivided into four pairs of electrodes. Four steel tubes, welded together, surround the four cylindrical graphite anodes concentrically. The cell design has been developed over the course of time [13], [35–40].

The anodes usually take the form of a hollow cylinder with slits in the wall [41], [42]. Alternatively, anode blocks can be made with conical central holes, which increases the proportion of graphite and decreases the resistance along the length of the anode [43].

The four steel cathodes are connected to the electrical supply by two conducting rods that pass through the side walls of the cell. The annular gap between the anode and the cathode is < 50 mm wide. A cylindrical iron mesh is placed in this gap, forming a diaphragm to separate the products, sodium and chlorine. The cylindrical diaphragm is attached to a device that collects sodium and chlorine from the four electrode pairs. Above the cathode space (between the cathode and the diaphragm) is a channel with a closed roof to collect the liquid sodium that rises through the melt in the form of droplets. Four shafts lead gaseous chlorine from the four anode spaces upward into a dome. Chlorine collects in the dome above the surface of the melt and is removed via an outlet at the top.

The chlorine pipe leading from the top of the dome is fitted with an access cover that is used to monitor the extraction of chlorine from the cell into the collection device. This cover can be provided with a tongue-and-groove seal to prevent intake of air into the chlorine [44].

The roof of the sodium- collecting device usually slopes upward, so that sodium flows smoothly into the vertical riser pipe. The device for collecting the sodium and chlorine may also include a tube for promoting circulation of the melt. The chlorine bubbles that form and rise at the graphite anodes improve the circulation of the melt in the anode space. The rising gas bubbles cause the surface of the melt to rise in the chlorine collection hood, so that the salt level is higher inside the hood than outside. Circulation of the molten catholyte is improved by the up-

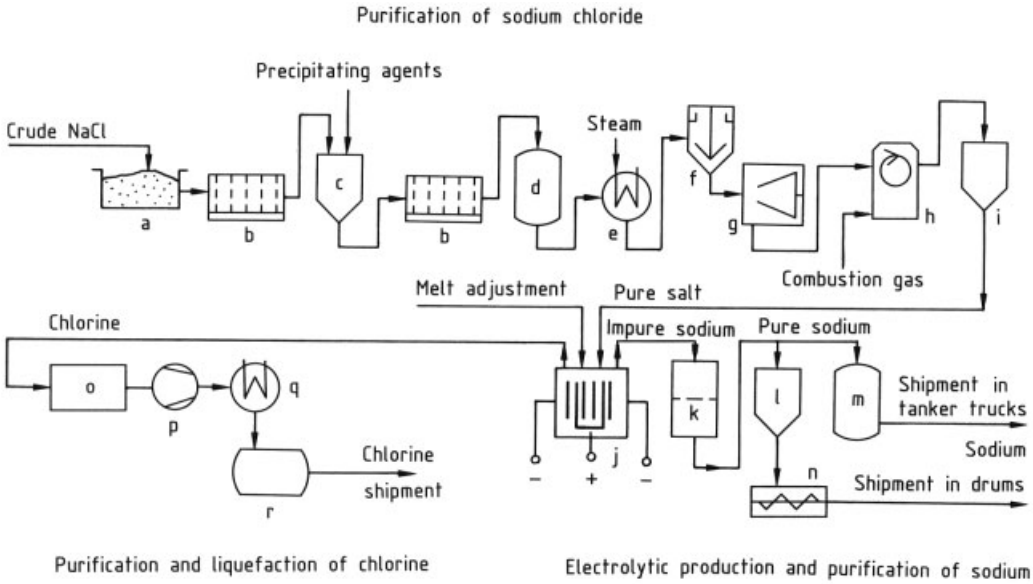

Figure 4. Flow sheet of sodium production
a) Dissolution vessel; b) Filter; c) Precipitation vessel; d) Vessel for pure brine; e) Evaporator; f) Thickener; g) Centrifuge; h) Rotary dryer; i) Pure salt bunker; j) Electrolysis cell; k) Sodium filter; l) Sodium storage; m) Sodium tank; n) Casting; o) Chlorine purification; p) Chlorine compressor; q) Liquefier; r) Chlorine storage vessel

ward flow of the sodium. Downward flow of the catholyte takes place by natural convection at the inactive outer walls of the cathodes. By including a circulation tube, the flow of the catholyte can be improved, enabling the electrolyte from the chlorine- collecting hood to flow downward into the central cathode space [45].

Sodium flows from the riser pipe of the collecting device into a closed collection vessel [6]. The calcium dissolved or suspended in the sodium separates as a solid on the inner walls of the riser pipe. Blockage is prevented by means of an agitator inside the riser pipe that scrapes calcium from the walls. Since the density of calcium is higher, it sinks down through the sodium to the salt level, where partial reaction with NaCl of the ternary salt mixture occurs to form calcium chloride and sodium. Various designs of this agitator mechanism in the riser pipe have been described [46], [47]. Calcium, oxides (e.g., Na_2O_2 and CaO), and entrained melt settle at the bottom of the collection vessel, which is therefore equipped with a stirrer [48]. A design has also been described in which this stirrer is combined with the agitator in the riser pipe [49].

A characteristic of another possible design [49] is, that the open overflow is replaced by an overflow pipe between riser pipe and collecting vessel (see Fig. 5).

If the overflow pipe is at the correct height, the hydrogen liberated occasionally at the cathode and the chlorine diffusing into the cathode space do not enter the collector. The gases that collect above the sodium in the gas space of the riser pipe are removed by a stream of inert gas (e.g., nitrogen) [50].

An overview of the procedures for starting up a Downs cell is given in [51].

The salt used for the initial melting operation can be a solid salt mixture or a powdered, solidified melt. Melting is achieved by short-circuiting the electrodes with so- called starter blocks. A number of these starter blocks are wedged between the anode and cathode, and the cell is then filled with salt mixture. In order to melt the mixture, a low electric current is passed through the cell. When sufficient salt has melted, the starter blocks are removed, the current is increased to the normal operating level, and the diaphragm is inserted.

In an improved technique, the cell is filled with solid powdered electrolyte or salt mixture until the level is just below the top of the electrodes [52]. The part of the cathode that is not co-

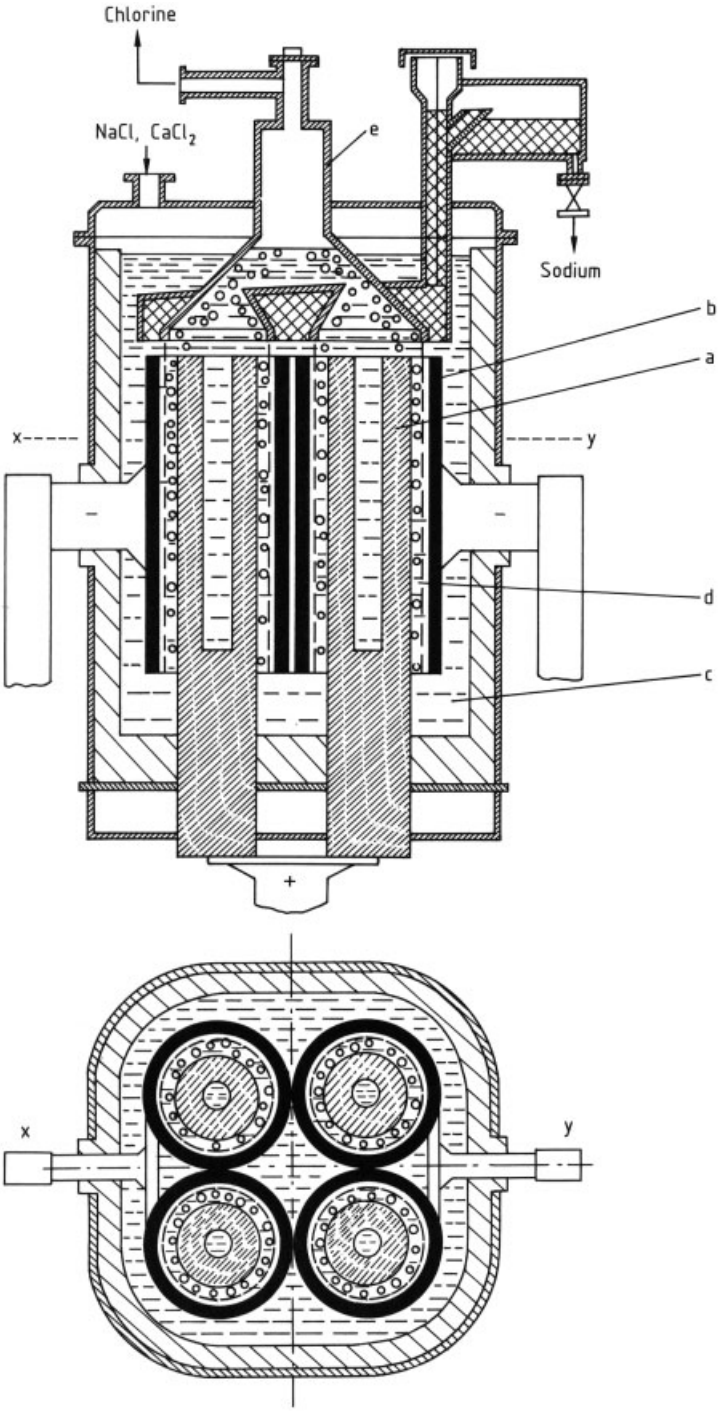

Figure 5. Modified Downs cell
a) Anode; b) Cathode; c) Molten salt; d) Diaphragm; e) Chlorine dome

vered is then heated to a temperature well above the melting point of the salt mixture. If a small amount of liquid melt is introduced from another cell, a low electric current can be passed between the electrodes. While solid salt mixture is being added, the electric current is slowly increased until the cell is filled with molten salt mixture to the correct height. The diaphragm is then introduced. A further improvement in the procedure for starting up a cell has now been developed, in which the cell is preheated, and sufficient liquid melt from neighboring cells is added to enable the diaphragm to be introduced and the electric current to be applied immediately to the cell [51]. Sodium production can then begin without the need for the complicated procedures described above.

Cells normally have a capacity of ca. 8.4 t molten salt with a density of ca. 2.7 g/cm^3. Cells have a working life of about three years. They are normally taken out of commission when the cell voltage and hence the energy consumption become too high due to wear of the anodes. Operating data for a Downs molten salt cell are [5]:

Operating temperature	590 – 610 °C
Lifetime of cell	1100 – 1200 d
Cell voltage	6.5 – 7 V
Cell current	$\leq$ 45 kA
Current efficiency	80 – 90 %
Consumption of electrical energy	9.8 – 10.5 kW · h/kg Na

The voltage required for the electrolysis of sodium chloride is ca. 3.4 V, which is ca. 50 % of the cell voltage.

In the course of a working life, various changes in the operating conditions of the Downs cell take place that lead to increased energy consumption. These changes are mainly a result of loss of graphite from the anodes due to corrosion and erosion, and buildup of impurities in the melt, leading to the accumulation of precipitated material on the cell floor [53]. A cell that consumes ca. 280 kWh per day to produce 700 kg sodium uses ca. 116 kWh for decomposing the NaCl to sodium and chlorine, which corresponds to a power efficiency of ca. 41 %. Heating the charge of NaCl from room temperature to the operating temperature requires 9.9 kWh, and an additional 11.1 kWh is required for melting, amounting to ca. 8 % of the total power consumption. To maintain a constant melt temperature, a cooling power of 143 kWh is required (i.e., ca. 51 % of the total power consumption). The difficulty in operating a cell consists in maintaining an optimum balance of the continually changing rate of heat production due to the internal resistance of the cell, recombination of sodium and chlorine to form sodium chloride, and the rate of removal of excess heat. Here, care must be taken that the flow pattern and hence the cell resistance are not disturbed by the adjustments made to control the cooling [53].

Electricity consumption figures for large installations for the production of sodium by molten salt electrolysis are given in [54]. The consumption of electrical energy for the electrolysis is 9.8 – 10.5 kWh/kg pure sodium, depending on the length of time that the cell has been in operation. If the associated equipment (ventilation, heating of vessels, and equipment) is included, this increases to 11.5 – 12.0 kWh/kg pure sodium. If a plant for purifying the raw salt and a plant for the purification and liquefaction of the chlorine produced is taken into account, the total energy consumption is 12.0 – 13.0 kWh/kg pure sodium.

4.2.3. Properties of the Electrolyte

The electrolyte consists of a mixture of NaCl, $CaCl_2$, and $BaCl_2$. Some properties of these salts are listed in Table 1.

The melting point of the ternary eutectic of these salts is 450 °C. However, the eutectic mixture is not suitable as a working electrolyte since the $CaCl_2$ content of ca. 50 wt % would result in an unacceptably high proportion of calcium in the sodium product. A mixture of 28 wt % NaCl, 25 wt % $CaCl_2$, and 47 wt % $BaCl_2$ (51 mol % NaCl, 24 mol % $CaCl_2$, 25 mol % $BaCl_2$) is suitable. This mixture melts at 570 °C, and the calcium content of the sodium produced is ca. 1 wt %.

If the rate of sodium chloride addition varies during electrolysis, the NaCl content of the mixture and its melting point fluctuate, but the $BaCl_2$ – $CaCl_2$ weight ratio remains practically constant at 1.88. For example, melting points of the mixture are 550 °C for 26 wt % NaCl and 600 °C for 32 wt % NaCl.

The current efficiency would be 100 % if pure sodium were deposited at the cathode, and no

Table 1. Physical properties of salts used in the Downs cell

Property	NaCl	$CaCl_2$	$BaCl_2$
M_r	58.45	110.99	208.27
Melting point, °C	801	782	963
Density of molten salt at *mp*, g/cm^3	1.6	2.1	3.2
Electrical conductivity of the molten salt at *mp*, Ω^{-1} cm^{-1}	3.6	2.0	2.0

back reaction of sodium with chlorine occurred to form sodium chloride. The diaphragm prevents sodium droplets formed at the cathodes from reacting with the chlorine bubbles formed at the anodes.

However, another possible mechanism for the back reaction, associated with the slight solubility of sodium in the melt, causes the melt to become a partial electronic conductor [55]. The dissolved sodium dissociates into sodium ions and electrons [56]. The electrons and surrounding cations form color centers, called F- centers, that impart a bluish appearance to the melt.

If dissolved sodium is present in the cathode compartment, electrons migrate in the electric field from the cathode across the diaphragm where they react with sodium ions to form sodium. This sodium recombines with the gaseous chlorine present in the anode compartment, which decreases the current efficiency.

To remove dissolved metal (mainly sodium) from the electrolyte, two methods of regeneration have been proposed. In both of these, regeneration of the melt takes place outside the cell. The melt can be melted repeatedly and allowed to solidify, so that the dissolved metals coagulate and separate from the salt melt [57], or the dissolved metals can be oxidized by passing air through the salt melt [58].

Under some circumstances (e.g., during a plant shutdown), the sodium production rate of a cell may have to be reduced for a period of time without interrupting the ohmic resistance heating of the cell. To achieve this, 0.1 – 0.5 wt % aluminum oxide can be added to the melt [59]. Solid bridges of aluminum form between the cathode and the diaphragm, and the latter starts functioning as the cathode. Recombination of sodium with chlorine is no longer prevented and the current efficiency is drastically decreased. Since the melting point of aluminum (660 °C) is above the operating temperature of the cell the aluminum bridges do not melt. To restore the current efficiency to its normal value, the bridges are broken by shaking the diaphragm or replacing it.

4.2.4. Other Electrolytic Processes

Sodium can be deposited electrolytically on a molten lead or mercury cathode to form an alloy [31]. Using this principle, Ashcroft has designed a double cell, in whose primary cell an amalgam is formed by electrolysis of aqueous NaCl with a mercury cathode. Alternatively, molten sodium chloride is electrolyzed with a molten lead cathode. The alkali-metal alloy is then transferred to the secondary cell where the molten salt electrolyte (*mp* ca. 200 °C) is $NaNH_2$, NaCN, or an NaOH – KOH mixture. In the secondary cell, the alloy is anodically decomposed into its constituents, whereby the alkali metal is deposited on the cathode and the alloying metal is returned to the primary cell. However, the Ashcroft process could not compete with other electrolytic processes.

In the Szechtman process, a sodium – lead alloy is produced by electrolysis of molten sodium chloride with a molten lead cathode. The sodium is then recovered from the alloy formed by distillation [31].

Sodium can be recovered by electrolysis of the amalgam produced in the chlor-alkali electrolysis process. Several electrolytic methods have been developed for recovering sodium from the amalgam, but they have not been used commercially (e.g., the Tekkosha process [10, vol. 17, p. 151]).

Reference [60] describes the use of a solid ceramic that conducts sodium ions (as used in the sodium – sulfur battery) as a divider for a two- compartment cell. This method allows low-melting sodium salts or salt mixtures to be used to produce sodium [1, vol. 21, p. 193], [60–64]. This type of cell should give current efficien-

cies approaching 100 % with 30 % lower energy consumption, as well as producing high-purity sodium. The process has, however, not been operated on a large scale [1].

5. Safety

If appropriate regulations and special safety precautions are observed, sodium can be handled safely in inert systems or even in the open air, provided its temperature is <115 °C [11]. The possibility of leakage from pipework and vessels containing molten sodium must be guarded against, because spontaneous ignition can occur at >115 °C [65]. Effectively fighting sodium fires requires a detailed knowledge of the combustion properties of sodium [66]:

1) In sodium fires in vessels in the open air, the temperature in the reaction zone is 900 – 1000 °C. Even in unfavorable conditions, the temperature of the sodium remains below its boiling point. Hence, high rates of combustion, such as those observed with boiling liquids, do not occur with sodium.
2) For any equipment carrying molten sodium, a special system of containment must be provided that safely retains sodium escaping from a leak and effectively prevents contact with air so that a fire cannot develop.
3) The preferred material for insulating sodium- carrying pipework and vessels is rock wool or slag wool with a low content of free silica. For higher temperature, ceramic fibers are suitable. Fibrous insulation made from molten materials is less hygroscopic than powdered or foamed porous material and is therefore preferable. The action of burning sodium on concrete has been studied in detail [67], [68].

The question of an optimum extinguishing agent has been discussed in the literature. The provision of inert-gas atmospheres such as nitrogen or argon in areas at risk from fire is generally too expensive. For pure sodium fires, which take place at the surface, a special expanded graphite is a very suitable fire-fighting material [69]. This powder floats on liquid sodium and thus prevents access of oxygen. Other powders, however, sink in the sodium and hence fail to provide a covering layer. The amount of such powder used on a vessel of molten sodium must therefore always be sufficient to absorb all the sodium and shield it against air [66]. To extinguish small spillages of sodium that have caught fire, powders specially developed and permitted for metal fires (Class D fires), as well as dry salt, dry sand, and anhydrous sodium carbonate, are suitable [11]. Water, foam, halons, CO_2, and fire-extinguishing powders of Classes A, B, and C must not be used under any circumstances [11].

Sodium burns to form sodium oxide (Na_2O) and sodium peroxide (Na_2O_2). With surface fires, 10 – 30 % of the sodium forms an atmospheric aerosol with a typical particle size of 0.001 – 0.002 mm. They combine with atmospheric moisture within a few seconds to form sodium hydroxide. This is followed by reaction with carbon dioxide to form sodium carbonate or hydrogencarbonate, which takes about 3 – 5 min [70–72].

To clean equipment contaminated with sodium, the sodium- containing residues are first removed mechanically and then disposed of (e.g., by controlled combustion). The remaining sodium can then be converted into hydroxide by steam. The excess steam renders the reaction zone inert [11]. The removal of sodium by melting and cleaning out with steam and an alcohol – water mixture (isopropanol with 5 % water) is described in [73]. Small amounts of sodium in the laboratory can be converted to alkoxides, first with isopropanol, then with ethanol, and finally with methanol. Only when hydrogen evolution has ceased should the solution be treated with water. Very small amounts of sodium can be disposed of by storage under paraffin oil in an open beaker, which results in gradual conversion to the hydroxide.

6. Uses

Sodium Compounds. Some sodium compounds are important because of their nucleophilic properties.

Sodium hydride [74] is produced from sodium and hydrogen at 200 – 350 °C (→ Hydrides, Chap. 2.3.). Large quantities are reacted with trimethyl borate to form sodium borohydride

$$4\,NaH + B(OCH_3)_3 \longrightarrow NaBH_4 + 3\,NaOCH_3$$

which is a very good hydrogenation and reducing agent (→ Hydrides, Chap. 3.4.). Its reducing power can be controlled by selection of the solvent and by adding ionic compounds (→ Reduction). Sodium hydride reacts with the following compounds to give their sodium salts [108]: alcohols, phenols, thiols, amines, ketones, alkynes, and carboxylic acids.

Sodium amide (→ Sodium Amide) [74], which is produced from sodium and ammonia, is used mainly to synthesize indigo by the Heumann – Pfleger reaction and to produce sodium azide for the explosives industry. In the Heumann – Pfleger indigo synthesis, sodium amide is used to cyclize phenylglycine, forming indoxyl in very high yield [12], [74]. The reaction is carried out in molten sodium amide at 180 – 200 °C.

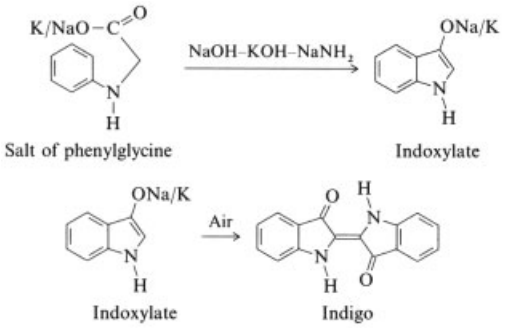

Sodium azide is synthesized by reacting molten sodium amide with dinitrogen oxide:

$$N_2O + 2\,NaNH_2 \longrightarrow NaN_3 + NaOH + NH_3$$

On an industrial scale, liquid ammonia is used as a solvent. In a variation of the method, sodium and ammonia are used without isolating $NaNH_2$ [9]:

$$3\,N_2O + 4\,Na + NH_3 \longrightarrow NaN_3 + 3\,NaOH + 2\,N_2$$

Sodium alkoxides [74] are used for *C*-, *N*-, and *O*-alkylation reactions (e.g., of malonic esters, acetoacetic esters, cyanoacetic esters, and benzyl nitrile) and in ring-closure reactions to form nitrogen-containing heterocycles. Alkoxides are used as synthetic building blocks in the Williamson ether synthesis.

Organometallic Compounds. The production of organometallic compounds is an important use of sodium. Organometallic compounds of group IV B metals can be prepared by a modified Wurtz synthesis [74].

Organotin compounds can be produced by the reaction

$$8\,Na + SnCl_4 + 4\,RCl \longrightarrow SnR_4 + 8\,NaCl$$

which, however, is used less often than other methods. Tributyltin oxide is an effective wood preservative:

$$SnCl_4 + 6\,Na + 3\,C_4H_9Cl \longrightarrow (C_4H_9)_3SnCl + 6\,NaCl$$

$$2\,(C_4H_9)_3SnCl + H_2O \longrightarrow (C_4H_9)_3Sn{-}O{-}Sn(C_4H_9)_3 + 2\,HCl$$

Tetraethyllead, $PbEt_4$, is produced by reacting a sodium – lead alloy with chloroethane (→ Lead Compounds, Chap. 3.). Use of the antiknock agents [75] $PbEt_4$ and $PbMe_4$ is being increasingly restricted as a result of environmental legislation.

Synthesis of Vitamin A. The preferred method for the production of vitamin A involves the Wittig reaction, i.e., the reaction of methylenetriphenylphosphorane with benzophenone to give triphenylphosphine oxide and 1,1-diphenylethylene [76]. (→ Sodium Amide, Chap. 5.) Triphenylphosphine is produced from chlorobenzene, sodium, and phosphorus trichloride [76].

$$3\,C_6H_5Cl + 6\,Na + PCl_3 \longrightarrow (C_6H_5)_3P + 6\,NaCl$$

Since the expensive triphenylphosphine is irreversibly converted to triphenylphosphine oxide in the Wittig reaction, ways have been sought to reuse the latter. An economic process has been developed in which the phosphine oxide is reacted with phosgene to give the dichloride, which is coproportionated with phosphorus [76].

$$(C_6H_5)_3P{=}O + COCl_2 \xrightarrow{-CO_2} (C_6H_5)_3PCl_2$$

$$3\,(C_6H_5)_3PCl + 2\,P \longrightarrow 3\,(C_6H_5)_3P + 2\,PCl_3$$

Metallurgy. Metals such as titanium, zirconium, tantalum, hafnium, and niobium can be produced in high purity by reacting their chlorides with sodium. Pure titanium was first prepared in 1887 by two Swedish chemists, using the reaction

$$TiCl_4 + 4\,Na \longrightarrow Ti + 4\,NaCl$$

This method of preparation became known as the Hunter process. The reaction of sodium with titanium tetrachloride is usually carried out by placing the two reactants in an evacuated steel bomb and initiating the reaction by passing an electric current through a wire. Other methods of production are now available (e.g., a process involving direct electrolysis of titanium tetrachloride).

Small amounts of sodium are used in lead refining, for optimizing aluminum – silicon melts and improving the properties of bronzes [11]. Sodium treatment [77] of Al – Si melts prevents anomalous crystallization, characterized by the precipitation of needle-like or lamellar silicon [78–80]. Sodium leads to supercooling processes [81], [82] that result in rounded eutectic silicon, normal fine eutectic solidification, and improved material properties, especially tensile properties [83].

Sodium – Sulfur Batteries. Rechargeable batteries with liquid sodium as the anode, sulfur – sodium polysulfide as the cathode, and sodium-doped aluminum oxide as the solid electrolyte are available for testing in the automobile industry [5], [84].

The two active materials, sodium and sodium polysulfide, are separated by a solid electrolyte of β-aluminum oxide, which conducts sodium ions. The operating temperature of the battery is 300 – 350 °C [85]. The most important component of a sodium – sulfur cell is the tube of solid electrolyte that separates the two molten electrodes. This material must be highly permeable to sodium ions and have a long lifetime to give a sodium – sulfur battery with a high power density and long working life [86].

Sodium as a Heat-Transfer Medium. The alkali metals, particularly sodium, were first used on a large scale as heat-transfer media by the nuclear energy industry. Liquid sodium is used worldwide as a coolant in fast breeder reactors (→ Nuclear Technology, Chap. 4.1.2.). Power stations may contain several thousand tonnes of the metal. Both its nuclear properties and its very favorable thermal properties have led to the choice of sodium to cool fast breeder reactor cores, where the heat flux is high. Alkali metals are also used to a limited extent as heat-transfer media outside the nuclear industry [87]. In a solar energy plant in the Mojave Desert, California, sodium is used to store energy and transfer it from the collector to the steam generator [88].

7. Sodium Alloys

According to Mott's rule [89], the miscibility of alkali metals with other metals in the liquid state depends on the ratios of the atomic radii, vapor pressures, and electronegativities of the dissolved metal and the solvent metal. Furthermore [8], the interaction of alkali metals with elements of group A (*s*- and *p*-block) of the periodic table differs from their interactions with the elements of group B (*d*- and *f*-block). Many elements of group A form intermetallic compounds with the alkali metals, unlike the elements of group B, which interact to a much lesser extent. The transition metals, including lanthanides and actinides, have very low solubilities in molten alkali metals. The metals of groups I B (Cu, Ag, and Au) and II B (Zn, Cd, and Hg) and the platinum-group metals are exceptions to this. In group I B, the solubility increases from copper to gold, the very low solubility of copper being similar to that of the transition metals.

Few of the many sodium alloys are of industrial importance. Some are briefly described below [90].

Sodium – potassium alloys [91] are used mainly in heat exchange systems, because of their good thermal conductivity and the wide temperature range over which they are liquid. The physical properties of two typical Na – K alloys are listed in Table 2. A review of physical properties and chemical reactions is given in [13]. Large-scale production is described in [5]. The ternary alloy, 12 wt % Na, 47 wt % K, and 41 wt % Cs, has the lowest known melting point (−78 °C) of all metallic systems [9].

Table 2. Physical properties of Na – K alloys [8]

Property	Composition, wt % K	
	44	78
Melting point, K	292.15	260.48
Boiling point, K	1098.15	1057.15
Density, kg/m^3 (373.15 K)	886	847
Viscosity, mPa · s (523 K)	0.316	0.279
Heat capacity, kJ kg^{-1} K^{-1} (473 K)	1.096	1.045
Thermal conductivity, W m^{-1} K^{-1} (473 K)	26.36	25.10

Although Na and K are chemically similar and both form face-centered cubic structures, the difference in atomic radii >15 %. The two are immiscible (i.e., no mixed-crystal formation occurs [92]). The compound KNa_2 is a so-called Laves phase, in which mixed-crystal formation is impossible because the difference in radius is too large [93].

Sodium – calcium alloys are byproducts of the molten salt electrolytic production of sodium from the binary salt mixture $NaCl - CaCl_2$ and the ternary mixture $NaCl - CaCl_2 - BaCl_2$. Calcium is only partially miscible with sodium. Calcium – sodium alloys containing 25 – 65 wt % sodium can be separated by high-temperature vacuum distillation [13]. Figure 6 shows the sodium – calcium phase diagram [94].

Sodium – Lead Alloys. Various methods have been described for the production of sodium – lead alloys because of their importance in the production of alkyl lead compounds [13]. One possibility is to melt the two metals together. Another is to deposit sodium electrolytically on molten lead cathodes [10]. Sodium and lead are completely miscible in the liquid state above 400 °C [13]. The following sodium – lead compounds are known.: $NaPb_3$, NaPb, Na_9Pb_4, Na_5Pb_2, and $Na_{15}Pb_4$ [95].

Sodium Alloys with the Coinage Metals. Copper has very low solubility in sodium [91], and no sodium – copper compounds have been reported. Although copper is resistant to sodium up to 300 – 400 °C, it is not recommended as a construction material for vessels or pipes containing sodium [13].

Silver has limited solubility in sodium [13], [91]. The existence of the compound $NaAg_2$ has been reported.

With gold, sodium forms the compound $NaAu_2$ (*mp* 1005 °C), which is the most stable compound in the Na – Au system [95]. It can be prepared from gold powder and sodium at 250 °C under high vacuum [13]. Sodium – gold alloys exhibit photoelectric properties and can be used in photocells [10].

Sodium Alloys with Group II B Metals. Zinc, cadmium, and mercury are very soluble in alkali metals [91]. The solubility in sodium increases from zinc to cadmium. The compounds $NaZn_{13}$ and $NaCd_2$ have been reported [91], [95]. Sodium and mercury form the following intermetallic compounds: $NaHg_4$, $NaHg_2$, Na_7Hg_8, NaHg, Na_3Hg_2, Na_5Hg_2, and Na_3Hg [95], [96]. A liquid amalgam containing ca. 1 % sodium is an intermediate product in chlor-alkali electrolysis and is used to produce sodium hydroxide solution. Amalgams with higher sodium content (1.25 %) are viscous liquids. Still higher sodium contents result in silvery-white crystalline solids [10]. The physical properties, chemical reactions, and production of sodium amalgam are described in [13].

Many attempts have been made to utilize the reducing properties of sodium amalgam in chemical syntheses [97]. Sodium dithionite is produced on a large scale by the reduction of SO_2 with sodium amalgam. Other uses of sodium amalgam include the production of sodium sulfide, reduction of salicylic acid to salicylaldehyde, and reduction of nitrobenzene to azobenzene. These reactions are no longer of industrial importance. The reaction of sodium amalgam with alcohols in a vessel filled with graphite spheres to form alkoxides is used to a limited extent [97].

Sodium amalgam formation can be used to recover mercury from mercury-containing but-

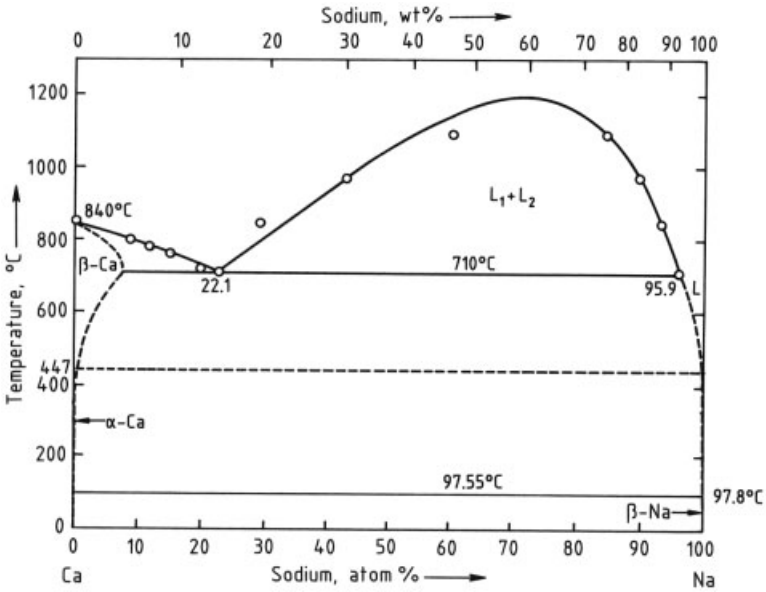

Figure 6. Calcium – sodium phase diagram [94]

ton cells. After treatment with molten sodium, the nonextractable solids are removed, and the amalgam is decomposed electrolytically into its constituents with a solid electrolyte made of sodium-ion- conducting β-Al_2O_3 [98].

8. Packaging and Transport

8.1. Forming Sodium Metal

Alkali metals are often marketed in shaped pieces of various sizes and weights. However, sodium is supplied to the customer mainly in the form of cast blocks produced by pouring the molten metal into containers and allowing it to cool and solidify. The shaped pieces are produced primarily by melting and casting the liquid metal. Alkali metals, especially sodium, are also pressed into variously shaped billets while in the solid state. The most serious disadvantage of these processes is the formation of impurities by reaction of the metal with air and atmospheric moisture. Also, the high pressure required for pressing the solid metal leads to wear. Pressing solid alkali metals is always a batch operation. The output of the machine is therefore limited, and the labor costs are high compared to a continuous process.

The production of alkali metal billets by continuous casting under pressure has also been described. In the presence of a lubricant the molten metal is forced by a pump through a die cooled to below the melting point of sodium [99].

In the extrusion molding process described in [100], molten metal is charged batchwise into the so- called feeder and, after cooling, until it is in a hot plastic state, is formed into an ingot, which is then extruded to form a billet by means of a plunger.

In the technique described in [101], two separate sets of machinery are used for extrusion of the alkali metal. The molten metal is fed into a gear pump, which delivers liquid metal. In a cooler, which is separate from the pump, the liquid alkali metal is cooled below its melting point and pressed into a solid shape under pressure from the gear pump (ca. 15 MPa).

In a state-of-the-art process [102], liquid alkali metal is fed into a twin-screw extruder, which is cooled such that the temperature falls to slightly below the melting point of the metal just before the extrusion die. The alkali metal is thus solid but sufficiently ductile for extrusion. The extrusion pressure at the outlet is 3 – 5 MPa.

The alkali-metal billet produced can be cut by an automatic cutting device, sometimes under a protective gas, to produce lengths suitable

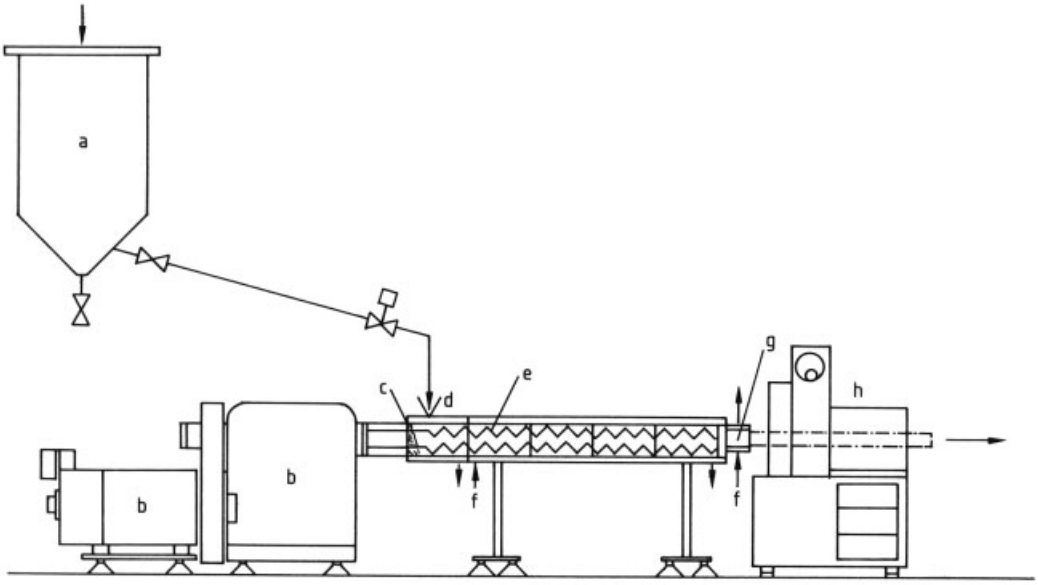

Figure 7. Schematic of an extrusion plant for the production of alkali-metal billets [102]
a) Sodium storage tank, heated; b) Drive; c) Sealing layer of sodium; d) Feeding nozzle; e) Twin-screw extruder; f) Heating and cooling jacket; g) Extrusion die; h) Cutting machine

for packing. Figure 7 shows a schematic of an extrusion plant for the production of alkali-metal billets [102].

After manufacture, sodium is in the molten state and can be filled immediately into freight containers. The following types are available:

1) Rail tank cars with a capacity of up to 45 t sodium
2) Road tank vehicles with a capacity of up to 20 t sodium
3) Iso-containers with a capacity of up to 20 t sodium
4) Special containers with a net capacity of up to 5 t sodium, and smaller packaging units in the form of drums of various types

Sodium is usually transported in the solid state for safety reasons and therefore must be remelted before it is emptied from the containers. This method of shipment allows the metal to be delivered in a very pure form. However, special handling equipment at the unloading site is required.

8.2. Transport Regulations

Sodium metal is a Class 4.3 hazardous material for all kinds of transport [103]. Detailed transport classifications read:

Road and Rail (Europe), ADR/RID: Class 4.3, No. 11 a
Inland Waterways (Europe), ADNR and ADN: Class 4.3, No. 11 a
Sea Cargo (Worldwide), IMDG Code: Class 4.3, UN No. 1428, PG. I
Air Cargo (Worldwide), ICAO-T I and IATA-DGR: Class 4.3, UN No. 1428, PG. I

The European transport regulations were harmonized with the UN-recommendation-based regulations for sea and air transport of January 1, 1993. Thus sodium metal may be transported worldwide by using a unified packing system. However, different weight limits and packaging requirements are specified for various methods of transport.

8.3. Break-Bulk Packaging

For all kinds of transport UN-approved PG. I packaging has to be used. PG. I means the highest test standard for performance-oriented packaging. The packaging must have no venting device.

Road and Rail Transport in Europe. Sodium metal must be hermetically sealed, so that neither moisture nor any other substance can penetrate the packaging and come in contact with the contents. The provision to cover the sodium with an inert gas or protective liquid is no longer required.

The following types of packaging are permitted: single packaging such as drums and canisters with removable head (made of steel or plastic) with a maximum capacity of 450 L or a net weight of 400 kg, as well as combination packaging (such as glass containers of 5-kg capacity, plastic containers of 30-kg capacity, or metal containers of 40-kg capacity) inside outer packages such as wooden, cardboard, or metal boxes with a maximum permissible net weight of 400 kg.

European Inland Waterways. Packaging regulations for European inland waters are identical with those for road and rail transport.

Sea Transport. In contrast to land transport, a distinction is made between lumps and solid fused goods (i.e., those poured into the container in a molten state) in sea transport.

Lumps have to be packed in hermetically-sealed (vapor-tight closure) packaging and must be completely covered by a protective liquid (flash point $> 50\,^{\circ}$C) or covered with nitrogen. The following types of packaging are permitted: Steel drums with removable head with a maximum net weight of 250 kg, or combination packaging such as glass bottles (max. 1 kg gross) packed in wooden boxes (max. 5 kg gross), metal boxes or cans (max. 15 kg net) packed in wooden boxes (max. 115 kg gross), or cardboard boxes (max. 40 kg gross).

Solid fused sodium must be packed in effectively closed (liquid-tight closure) drums with removable head with a maximum net weight of 250 kg.

Air Transport. Sodium metal may be transported only by cargo aircraft; transport by passenger aircraft is not permitted. The net weight per package must not exceed 50 kg. Special conditions and permitted types of packing are described in packing instruction No. 418 of the air transport regulations ICAO-T I/IATA-DGR[12].

8.4. Portable Tanks

Portable tanks include rail tank cars, road tank vehicles, and demountable tanks. Rail tank cars have a capacity up to ca. 45 t; tank containers and road tanker trucks, up to ca. 25 t.

Tanks for transporting sodium metal must be tested to a minimum pressure of 0.4 MPa. The temperature on the tank shell must not exceed 70 °C during transportation. Insulating material must be nonflammable. The atmosphere inside the tank must be inerted by an inert gas.

Road and Rail Transport in Europe. Although the transport of molten sodium metal is allowed, many producers allow the fused sodium to solidify the outer layer, so that escape of liquid sodium is prevented in case of leakage. Openings below the surface level of the material are permitted.

Portable tanks for sodium metal must undergo testing at eight-year intervals after the first certification, including testing of the condition of the inside and outside of the vessel and pressure testing with a liquid. In addition, a leakproofness test of the tank with its equipment and a check of the satisfactory operation of the equipment must be carried out at least at four-year intervals.

Transport by European Inland Waterways. Regulations for inland waterways are identical with those for the road and rail transport of movable large containers.

Sea Transport. IMO tank type 1 containers with a minimum wall thickness of 6 mm and a minimum test pressure of 0.4 MPa must be used. IMO tank type 1 tanks have a capacity of $>$450 L, with a maximum permitted operating pressure of 175 kPa or above. Road tank vehicles and rail tank cars used to transport sodium

on seagoing vessels have to comply with the IMO tank type 1 provisions too. In contrast to the land transport regulations, the sea transport regulations do not permit openings below the surface level to the material, i.e., the tank must be equipped with top openings. For sea transport, the sodium metal must be in the solid state.

Table 3. Trace impurities in sodium: Typical analysis figures and methods of determination [11]

Element	Concentration, ppm	Methods of determination
Aluminum	0.5	AAS, Zeeman
Arsenic	<1	AAS, Zeeman
Barium	<5	AAS, Zeeman
Boron	<1	Photometric
Calcium∗	450	Titrimetric
Calcium∗	5	AAS
Carbon	2	Dissolve, filter, GC
Carbon∗∗	7	Vacuum distill, GC
Chlorine	4	Nephelometry
Chromium	<0.5	AAS, Zeeman
Cobalt	<0.5	AAS, Zeeman
Copper	<0.5	AAS, Zeeman
Iron	4	AAS, Zeeman
Lead	<0.5	AAS, Zeeman
Lithium	<1	AAS
Magnesium	<1	AAS, Zeeman
Nickel	<0.5	AAS, Zeeman
Oxygen∗∗	7	Amalgam process, AAS
Phosphorus	<0.5	Photometric
Potassium	70	AAS
Silicon	<1	AAS, Zeeman
Silver	<1	AAS, Zeeman
Strontium	5	AAS, Zeeman
Sulfur	<4	AAS
Tin	<1	AAS, Zeeman
Zinc	0.3	AAS, Zeeman

∗ Specification: normal grade <600 ppm, nuclear grade <10 ppm.
∗∗ Owing to the high affinity of sodium for atmospheric oxygen, moisture, and carbon dioxide (formation of oxides, hydroxide, and carbonate), the figures given apply only to sodium that has come in contact with dry inert gas atmospheres (O_2 <1 ppm, H_2O <1 ppm) during filling, storage, and handling.

9. Quality Specifications and Analysis

Sodium is marketed with a purity of >99 wt %. Depending on the production method, standard-quality sodium can contain between 400 and 600 ppm of calcium [11]. Purification of sodium is described in [8].

Methods based on filtration and cooling depend on the low solubility of many elements in liquid sodium just above its melting point. Impurities remain in the filtration residue or settle out in solid form in cool traps.

The affinity of some of the impurities in sodium for oxygen is also utilized in purification processes since the oxides formed are easily removed from the sodium. For example, calcium is reacted with Na_2O or Na_2O_2 at 300 – 400 °C, and the calcium oxide formed is filtered off [5], [104].

The trace elements in sodium are listed in Table 3, with typical analysis figures and analytical methods [11]. Sodium can also be obtained in special grades [105].

10. Health and Safety

Sodium can be handled safely without endangering the health of personnel concerned, provided appropriate precautions are taken to deal with possible hazards [106]:

1) Spontaneous ignition of sodium in moist air at >115 °C
2) Severe caustic effects of the combustion products of sodium
3) Fire explosion associated with uncontrolled chemical reaction between sodium and acid, water, aqueous solutions, alcohol

Sodium must always be stored, transported, and handled under dry and inert conditions. Direct contact with the skin must be prevented by wearing gloves and clothing made of flame-resistant material. The eyes are protected by spectacles, goggles, or shields in all circumstances.

Information and advice about necessary technical, organizational, and personal protective measures can be obtained from company publications [11], [107] and leaflets available from chemical industrial trade associations [106]. Methods for safe working with sodium in the laboratory are described in detail in [8], [11].

11. Economic Aspects

Until recently, sodium was used mainly for the production of lead- containing antiknock agents for gasoline and for the production of titanium. However, the increasing use of lead-free gasoline and the manufacture of titanium by

other processes have led to a drastic reduction in sodium capacity worldwide [108], [109]. In 1979, world sodium capacity, excluding the former Eastern-bloc states, was estimated at 250 000 t/a [10]. By 1986, this had decreased to 200 000 t/a [5]. In 1993 it is expected to be ca. 80 000 t/a. The most important sodium producers in the Western world, with their 1993 capacities are listed below:

Du Pont, Niagara Falls, United States	51 000 t/a
Associated Octel, Ellesmere Port, United Kingdom	20 000 t/a
Métaux Spéciaux, Pomblière, France	12 000 t/a

12. References

1. *Kirk-Othmer,* 3rd ed., **21,** p. 181.
2. Hofmann, Rüdorff: *Anorganische Chemie,* 21st ed., Vieweg & Sohn, Braunschweig 1973, p. 420.
3. C. Berger, H. Fickenfrerichs, R. Peper, W. Jansen: *Prax. Naturwiss. Chem.* **40** (1991) no. 6, 2.
4. M. Sittig: *Sodium, its Manufacture, Properties and Uses,* Reinhold Publ. Co., New York 1956, pp. 1 – 9.
5. *Winnacker-Küchler,* 4th ed., **4,** p. 326.
6. H. Y. Castner, US 452 030, 1891.
7. Rössler & Hasslacher Chem. Co., US 1 501 756, 1921 (J. C. Downs).
8. H. U. Borgstedt, C. K. Mathews: *Applied Chemistry of the Alkali Metals,* Plenum Press, New York 1987.
9. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* 1st ed., Verlag Chemie, Weinheim 1988, pp. 97.
10. *Ullmann,* 4th ed., **17,** p. 143.
11. Degussa, Geschäftsbereich Chemie: Natrium, Eigenschaften, Handhabung, Anwendung, Company brochure, Ch 621-0-2-1089H, Hanau.
12. G. Dittrich, H. Offermanns, H. Schlosser, Jun., *ChET Chem. Exp. Technol.* **3** (1977) 151 – 162.
13. "The Alkali Metals," *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* **vol. 2, suppl. 2, part 1,** Longmans, London 1961.
14. Linus Pauling: *Chemie, Eine Einführung,* 7th ed., Verlag Chemie, Weinheim 1967.
15. P. Caubère, *Angew. Chem.* **95** (1983) 597 – 611.
16. Degussa, Geschäftsbereich Chemie: Natriumhydrid, Eigenschaften, Reaktionen, Anwendung, Company brochure, C 25-0-1-772 Vol, Hanau.
17. O.-A. Neumüller: *Römmps Chemie-Lexikon,* 8th ed., Franck'sche Verlagsbuchhandlung, Stuttgart 1985.
18. A. Klemenc, E. Svetlik, Z. Anorg. Allg. Chem. **269** (1952) 153.
19. K. M. Myles, F. A. Cafasso, *J. Nucl. Mater.* **67** (1977) 249 – 253.
20. *Ullmann,* 3rd ed., **17,** p. 415.
21. J. J. Chydenius *Ann. Phys.* (*Leipzig)* **2** (1863) no. 119, 43 – 56.
22. W. v. Bolton, *Z. Elektrochem. Angew. Phys. Chem.* **14** (1908) 768 – 770.
23. D. Lely, Jr., L. Hamburger, *Z. Anorg. Chem.* **87** (1914) 209 – 228.
24. L. F. Nilson, *Ber. Dtsch. Chem. Ges.* **15** (1882) 2537 –2547.
25. *Ullmann,* 4th. ed., **17,** p. 711.
26. N. Korber, W. Assenmacher, M. Jansen, *Prax. Naturwiss. Chem.* **40** (1991) no. 6, 18.
27. U. Deschler, P. Kleinschmit, P. Panster, *Angew. Chem.* **98** (1986) 237 – 253.
28. A. Streitwieser, C. H. Heathcock, *Organische Chemie,* Verlag Chemie, Weinheim 1980, pp. 192 – 194.
29. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* 4th ed., John Wiley & Sons, New York 1980.
30. National Distillers and Chem. Corp.: *Sodium Monography,* Company brochure, New York.
31. *Winnacker-Küchler,* 3rd ed., **6,** 81.
32. Nippon Soda, JP 59 050 021, 1982.
33. The Dow Chemical Comp., US 3 265 492, 1963 (P. R. Juckniess).
34. Ethyl Corp., US 2 660 517, 1952 (F. L. Padgitt).
35. R. E. Hulse, US 2 130 801, 1938.
36. Ethyl Corp., US 2 944 950, 1957 (G. O. Hayes).
37. Du Pont, US 3 037 927, 1962 (C. T. Gallinger).
38. J. S. Honea, US 2 770 364, 1956.
39. Ethyl Corp., US 2 770 592, 1956 (A. L. Feutress).
40. Olin Mathieson Chem. Corp., US 3 085 967, 1960 (G. T. Motock).
41. Ethyl Corp., US 2 755 244, 1952 (C. C. Harvey).
42. Du Pont, US 2 921 894, 1957 (T. O. Callaghan).
43. Du Pont, US 3 544 444, 1967 (V. J. Reilly).
44. Du Pont, US 4 376 028, 1981 (R. D. Marczewski, F. J. Ross).

45. Du Pont, US 3 118 827, 1961 (S. E. Eckert, F. J. Ross).
46. ICI, US 2 861 938, 1955 (R. D. Glascodine).
47. Ethyl Corp., US 3 463 721, 1967 (L. L. Harris).
48. Du Pont, US 4 305 671, 1980 (F. J. Ross).
49. Du Pont, FR 1 567 424, 1967 (C. T. Gallinger).
50. Du Pont, US 4 092 228, 1977 (F. J. Ross).
51. S. Crowther, US 3 776 823, 1967.
52. National Distillers and Chem. Corp., US 2 913 381, 1967 (W. F. McFadyen, C. E. Buterbaugh).
53. K. A. Ruppert, Degussa AG, Company information, Frankfurt/Main 1992.
54. K. H. Meier, Degussa, Company report, Frankfurt/Main 1992.
55. M. A. Bredig in M. Blander (ed.): *Molten Salt Chemistry,* Interscience, New York 1964, pp. 367 – 425.
56. N. H. Nachtrieb in J. Prigogine, S. A. Rice (eds.): *Advances in Chemical Physics,* **vol. 31,** John Wiley, New York 1975, pp. 465 – 480.
57. Du Pont, US 3 558 451, 1971 (R. E. Svadlenak).
58. Du Pont, US 3 560 353, 1971.
59. Du Pont, US 4 139 428, 1979 (A. J. Dean, F. J. Ross).
60. Ford Motor Co., US 3 488 271, 1966 (J. T. Kummer, N. Weber).
61. Du Pont, US 4 203 819, 1978 (S. A. Cope).
62. Du Pont, US 4 133 728, 1978 (S. A. Cope).
63. Du Pont, US 4 089 770, 1977 (C. H. Lemke).
64. ICI, US 3 607 684, 1967 (A. T. Kuhn).
65. F. Huber, P. Menzenhauer, W. Peppler, W. Tell: "Untersuchung des Brandverhaltens von flüssigem Natrium," *KFK Nachr.* **4** (1972) no. 3, 9 – 13.
66. F. Huber, P. Menzenhauer, W. Peppler, W. Till, Verhalten von Natrium-Bränden und Erprobung von Schutzsystemen, KFK – 1970, Kernforschungszentrum Karlsruhe, April 1974.
67. W. Cherdron: *Jahrestagung Kerntechnik 1989,* Düsseldorf, p. 231.
68. W. Cherdron, W. Schütz: *Jahrestagung Kerntechnik 1993,* Köln.
69. P. Menzenhauer, G. Ochs, W. Peppler, Erprobung eines Löschpulvers auf der Basis von Graphit für Flüssigmetallbrände, KFK – 2525, Kernforschungszentrum Karlsruhe, Nov. 1977.
70. F. Huber, P. Menzenhauer, W. Till, *KFK – 1970,* Kernforschungszentrum Karlsruhe, April 1974.
71. W. Cherdron, S. Jordan, W. Lindner, *KFK – 4358,* Kernforschungszentrum Karlsruhe, Nov. 1990.
72. S. Jordan et al., *Nuclear Technol.* **81** (1988) no. 2, 183.
73. P. Menzenhauer et al., "Experience with Cleaning of Sodium-Wetted Components and Decontamination at Nuclear Research Centre Karlsruhe," *IAEA Meeting on Sodium Removal and Decontamination,* Richland, WA, USA, Aug. 1978, pp. 30 – 37.
74. E. Bilger, L. Lange, *Prax. Naturwiss. Chem.* **40** (1991) no. 6, 12.
75. J. F. Cordes, *Chem. Unserer Zeit* **3** (1977) no. 11, 65.
76. H. Pommer, *Angew. Chem.* **89** (1977) 437 – 443.
77. K. W. Stoffregen, Fa. Dr. Riedelbauch & Stoffregen GmbH, Company report, Meisenheim/Glau.
78. A. Pacz, DT 117 773, 1925.
79. E. Dunkel, *Giesserei* **58** (1971) no. 24, 782.
80. Metallgesellschaft, "Silumin, die Al-Gußlegierung," Company brochure, Frankfurt/Main.
81. G. Gürtler, *Z. Metallk.* **44** (1953) 503.
82. S. Justi, K. Körber, K. Löhberg, *Giessereiforschung* **24** (1972) no. 1, 37 – 44.
83. R. Irmann: *Aluminium in Sand und Kokille,* Aluminium Verlag, Düsseldorf 1953, p. 104.
84. Asea Brown Bovery AG, "Hochenergiebatterie und Elektroauto," Company brochure no. DHB 1 310 881 (1987), DHB 1 253 871 (1988), Heidelberg.
85. F. Beck, *Chem. Ing. Tech.* **54** (1982) no. 9, 809 – 817.
86. W. Fischer, F. J. Rohr, *Chem. Ing. Tech.* **50** (1978) no. 4, 303 – 305.
87. H. U. Borgstedt, G. Frees, *Werkst. Korros.* **38** (1987) 725 – 737.
88. K. Förster, G. Ruloff: "Flüssigmetalle als Wärme-übertragungsmittel," *CAV,* Konradin Verlag, Leinfelden – Echterdingen 1988, p. 182.
89. B. W. Mott, *Philos. Mag.* **2** (1957) no. 8, 259 – 283.
90. P. E. Potter, M. H. Rand in R. V. Ohse (ed.): *Handbook of Thermodynamic and Transport of Alkali Metals,* Blackwood, Oxford 1985, pp. 914 – 924.
91. H. U. Borgstedt, C. Guminski, *Metals in Liquid Alkali Metals,* IUPAC Solubility Data Series, in press.
92. D. F. Shriver, P. W. Atkins, C. H. Langford: *Anorganische Chemie, ein weiterführendes Lehrbuch,* VCH, Weinheim 1992.

93. E. Riedel: *Allgemeine und anorganische Chemie,* 5th ed., De Gruyter, Berlin 1990.
94. A. D. Pelton, *Bull. Alloy Phase Diagrams* **6** (1985) 35 – 37.
95. O. J. Foust: *Sodium Na – K, Engineering Handbook,* Gordon & Breach, New York 1972.
96. C. Hirayama, Z. Galus, C. Guminsky, "Metals in Mercury," in A. S. Kertes (ed.): *Solubility Data Series,* vol. 25, International Union of Pure and Applied Chemistry, in press.
97. H. Hund, F. R. Minz in H. Harnisch, R. Steiner, K. Winnacker (eds.): *Chemische Technologie,* 4th ed., **vol. 2,** Carl Hanser Verlag, München 1982, p. 400.
98. Kernforschungszentrum Karlsruhe, DE 4 034 137 C 1, 1990 (H. U. Borgstedt).
99. E. Hermann: *Handbuch des Stranggießens,* Aluminium Verlag, Düsseldorf 1958, pp. 764 – 765.
100. Metall Invent S.A., DT 2 457 423, 1974.
101. Unilever, GB 1 236 233, 1968 (W. H. De Groot, M. H. Hilder).
102. Degussa AG, DE 3 016 173, 1980 (W. Hinrichs et al.).
103. R. Neureiter, Degussa AG, Company report, Frankfurt/Main 1992.
104. Métaux Spéciaux S.A. Paris, DT 2 454 051, 1974 (B. Batoux, A. Laurent-Atthalin, M. Salmon).
105. Métaux Spéciaux S.A., "Sodium," Company brochure, 92087 Paris la Défense, France.
106. O. Pfeffer: Berufsgenossenschaft der chemischen Industrie, Merkblatt Natrium, M 019,5/84, Jedermann-Verlag, Heidelberg 1984.
107. Du Pont, Sodium Data Sheet, Chemical Safety Data Sheet, Wilmington, Delaware 19898 SD-47, Sodium, Manufacturing Chemists Association, Washington, D.C. 1978.
108. A. Huber: "*Sodium (Metal) – United States,*" (E. H. Data Summary, Chemical Information Servicies, Stanford Research Institute, California 1990.
109. G. Busch: "Sodium Makers Adjust to Changing Market Place," *Chem. Mark. Rep.* 240 (1991) 27.

Sodium Borohydride → Hydrides
Sodium Bromide → Bromine Compounds

Sodium Carbonates

CHRISTIAN THIEME, Solvay Alkali GmbH, Solingen, Federal Republic of Germany

1.	**Sodium Carbonate**	1
1.1.	**Introduction**	1
1.2.	**Properties**	2
1.3.	**Sodium Carbonate Minerals**	4
1.4.	**Production**	4
1.4.1.	Ammonia – Soda Process	4
1.4.1.1.	General and Theoretical Principles	4
1.4.1.2.	Technology	6
1.4.2.	Modified Ammonia – Soda Process	11
1.4.3.	Production from Trona and Other Sodium Carbonate-Containing Minerals	12
1.4.4.	Other Processes	13
1.4.5.	Wastewater Problems in Sodium Carbonate Production	15
1.5.	**Quality Specifications**	15
1.6.	**Producers**	16
1.7.	**Storage and Transport**	17
1.8.	**Uses**	17
1.9.	**Toxicology**	18
1.10.	**Economic Aspects**	18
2.	**Sodium Hydrogencarbonate**	18
2.1.	**Properties**	18
2.2.	**Production**	19
2.3.	**Uses and Quality Specifications**	19
3.	**References**	20

Based on the corresponding article in Ullmann's, 4th edition, written by Gerhard Braune and Helmut Scheider

1. Sodium Carbonate

1.1. Introduction

Sodium carbonate (soda ash) is the neutral sodium salt of carbonic acid. It is one of the most important raw materials used in the chemical industry, and has been known by humans since ancient times. Sodium carbonate had many uses even then (cleaning and glass manufacture) for which it is still indispensable. Sodium carbonate from the deposits in Wadi Natrun (4 % Na_2CO_3, 25 % $NaHCO_3$) was traded in ancient Egypt.

Production of the alkaline materials sodium carbonate and potassium carbonate (which were recognized rather late as chemically distinct) was carried out from ancient times until the 1800s by the combustion of marine and land vegetation, followed by calcination at red heat and leaching of the ash. The term soda ash originates from this process.

The soda ash obtained in this way was a low-percentage material; e.g., Spanish barilla contained 25 – 30 % Na_2CO_3, varec from Normandy 3 – 8 %, and Scottish kelp 10 – 15 %. The soda ash produced from plants was very expensive, and the processes used were too primitive for mass production and consumed vast quantities of vegetation.

Large-scale production was possible only after the development of a new process, named after its inventor LEBLANC (1742 – 1806). In 1790, the first plant was built in Saint-Denis, France. Initially, production was 300 kg/d. In the following years, the soda industry developed rapidly in England due to the increasing demand for soda ash to clean cotton from the colonies. In 1878, production in England was 8 t per week, and in the rest of the world, 4 t a week.

However, in the second half of the 1800s, the Leblanc soda ash process experienced serious competition from the Solvay process. In the first half of the 1800s, investigations were already being carried out into the industrial use of the double decomposition reaction of NaCl with NH_4HCO_3, but without success. However, the Belgian ERNEST SOLVAY (1838 – 1932) brought the development of the process to a successful conclusion during 1861 – 1865. The first plant was started up in 1865 in Couillet, Belgium, with an initial production of 1.5 t/d. The first Solvay plant in Germany was built in 1880 in Wyhlen on the upper Rhine.

The technically and economically superior Solvay process displaced the Leblanc to such an extent that by the early 1900s only a few Leblanc plants were in production. The last Leblanc works ceased operation around 1923. Since

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_299

then, the Solvay process has remained predominant. Since World War II, soda ash has been increasingly obtained from trona, especially in the United States (see Section 1.4.3).

1.2. Properties

Physical Properties. In addition to anhydrous sodium carbonate, Na_2CO_3, the following hydrates exist: sodium carbonate monohydrate, $Na_2CO_3 \cdot H_2O$; sodium carbonate heptahydrate, $Na_2CO_3 \cdot 7\,H_2O$; and sodium carbonate decahydrate, $Na_2CO_3 \cdot 10\,H_2O$. The physical properties of these compounds are listed in Table 1. Solubilities of the hydrates of sodium carbonate are shown in Figure 1. The relationship between the concentration of an aqueous solution of Na_2CO_3 and its pH at 25 °C is shown in Figure 2.

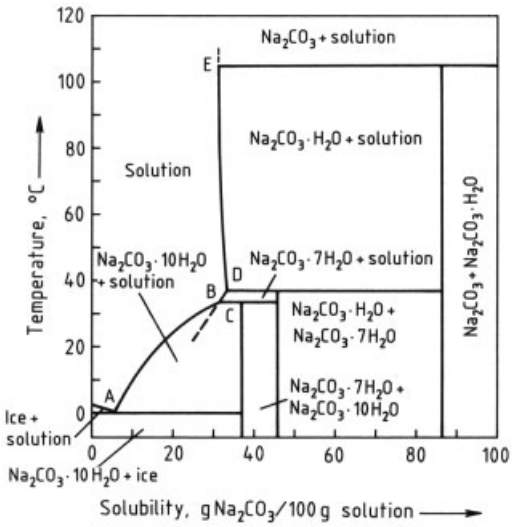

Figure 1. Solubilities in the system $Na_2CO_3 - H_2O$ [1] Temperatures and concentrations of the two-salt points and the cryohydric point

Chemical Properties. *Thermal Behavior.* The thermal decomposition of sodium carbonate to sodium oxide and carbon dioxide in a vacuum in the absence of chemically activating substances such as water vapor begins at ca. 1000 °C. The dissociation pressure is ca. 200 Pa and increases to 10.3 kPa at 1450 °C [3].

Reaction with Water and Carbon Dioxide. Sodium carbonate is hygroscopic. In air at 96 % R.H. its weight can increase by 1.5 % within 30 min [4].

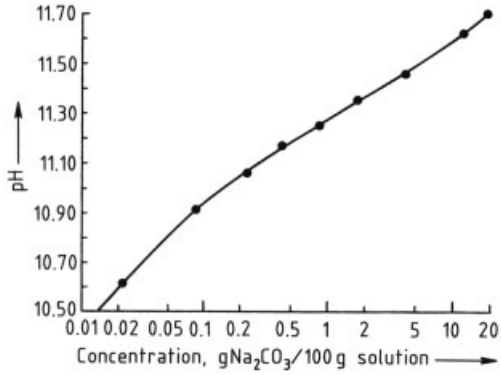

Figure 2. pH values of aqueous solutions of Na_2CO_3 at 25 °C [2]

Concentration, g Na_2CO_3/100 g solution	0.0055	0.0217	0.0884	0.2246
pH at 25 °C	10.274	10.612	10.919	11.063
Concentration, g Na_2CO_3/100 g solution	0.4376	0.8609	1.806	4.241
pH at 25 °C	11.171	11.254	11.357	11.453
Concentration, Concentration, Na_2CO_3/100 g solution	12.034	18.555		
pH at 25 °C	11.617	11.699		

If sodium carbonate is stored under moist conditions, its alkalinity decreases due to absorption of moisture and carbon dioxide from the atmosphere

$$Na_2CO_3 + H_2O + CO_2 \longrightarrow 2\,NaHCO_3$$

Water vapor reacts with sodium carbonate above ca. 400 °C to form sodium hydroxide and carbon dioxide [5].

Reaction with Elements. Sodium carbonate reacts exothermically with chlorine above ca. 150 °C to form NaCl, CO_2, O_2, and $NaClO_4$ [6].

The elements platinum, gold, vanadium, titanium, zirconium, aluminum, molybdenum, tungsten, and iron (1200 °C) are attacked by fused Na_2CO_3, with liberation of CO_2 and formation of complex metal oxide – sodium oxide compounds. At higher temperature, the Na_2O formed reacts with excess metal to form sodium metal vapor [7]. Sodium carbonate reacts very

Table 1. Physical properties of sodium carboante and its hydrates

	Anhydrous	Monohydrate	Heptahydrate	Decahydrate
Formula	Na_2CO_3	$Na_2CO_3 \cdot H_2O$	$Na_2CO_3 \cdot 7\,H_2O$	$Na_2CO_3 \cdot 10\,H_2O$
M_r	105.99	124.00	232.10	286.14
Density at 20 °C, g/cm^3	2.533	2.25	1.51	1.469
Melting point, °C	851	105 [a]	35.37 [b]	32.0
Heat of fusion, J/g	316			
Specific heat capacity at 25 °C, J g^{-1} K^{-1}	1.043	1.265	1.864	1.877
Heat of formation, J/g	10.676			
Heat of hydration, J/g		133.14	646.02	858.3
Crystal structure	monoclinic	rhombic	rhombic – bipyramidal	monoclinic – pseudohexagonal
Refractive indices n_α, n_β, n_γ	1.410, 1.537, 1.544	1.420, 1.506, 1.524		1.405, 1.425, 1.440
Heat of solution[c], J/g	− 222	− 79.6	197	243

[a] In its own water of crystallization.
[b] Incongruent.
[c] Integrated enthalpy of solution at infinite dilution.

Two-salt point	T, °C	Solubility, g Na_2CO_3/100 g solution	Solid phases
A	−2.1	5.93	ice + $Na_2CO_3 \cdot 10\,H_2O$
B	32.0	31.26	$Na_2CO_3 \cdot 10\,H_2O + Na_2CO_3 \cdot 7\,H_2O$
C	32.96	33.35	$Na_2CO_3 \cdot 10\,H_2O + Na_2CO_3 \cdot H_2O$ (metastable)
D	35.37	33.21	$Na_2CO_3 \cdot 7\,H_2O + Na_2CO_3 \cdot H_2O$
E	105 ± 5	31.15	$Na_2CO_3 \cdot H_2O + Na_2CO_3$

slowly with copper and nickel at high temperature (> 1500 °C).

1.3. Sodium Carbonate Minerals

Whereas the production of sodium carbonate from the ashes of plants that grow in salty soil near the sea is only of historical interest, extraction from soda-containing minerals, especially trona, is of increasing importance.

The most important deposits are located in Wyoming (Green River District), California (Owens Lake and Searles Lake), East Africa (Magadi Lake), and Lower Egypt (Wadi Natrun).

The proportion of Na_2CO_3 in the salt mixture in such deposits seldom exceeds 40 %. The accompanying compounds include sodium hydrogencarbonate, borax, sodium sulfate, sodium chloride, and potassium chloride.

For example, the concentrations of dissolved salts in the water of Owens Lake are [8]

9 % Na_2CO_3	1 % $Na_2B_4O_7$	16 % NaCl
1.2 % $NaHCO_3$	3 % Na_2SO_4	1.5 % KCl

Deposits of Na_2CO_3-containing minerals occur worldwide:

Trona	$Na_2CO_3 \cdot NaHCO_3 \cdot 2\,H_2O$
Natron	$Na_2CO_3 \cdot 10\,H_2O$
Thermonantrite	$Na_2CO_3 \cdot H_2O$
Hanksite	$2\,Na_2CO_3 \cdot 9\,Na_2SO_4 \cdot KCl$
Pirssonite	$Na_2CO_3 \cdot CaCO_3 \cdot 2\,H_2O$
Gaylussite	$Na_2CO_3 \cdot CaCO_3 \cdot 5\,H_2O$

Only trona is of commercial importance. These Na_2CO_3-containing minerals were formed from the original rock by the erosive action of air, water, heat, and pressure, followed by chemical changes caused by the action of atmospheric carbon dioxide. The carbonate-containing salts formed were leached by water and then concentrated and crystallized by evaporation.

1.4. Production

1.4.1. Ammonia – Soda Process

1.4.1.1. General and Theoretical Principles

The ammonia – soda process is based on the formation of sparingly soluble sodium hydrogencarbonate (known in the soda industry as bicarbonate) by reacting sodium chloride with

ammonium hydrogencarbonate in aqueous solution:

$$NaCl + NH_4HCO_3 \rightleftharpoons NaHCO_3 + NH_4Cl$$

The process includes the following stages:

1) Production of a saturated salt solution
   $NaCl + H_2O$
2) Burning of limestone or chalk (the CO_2 liberated is used in stage 4)
   $CaCO_3 \longrightarrow CaO + CO_2$
3) Saturation of the salt solution with ammonia
   $NaCl + H_2O + NH_3$
4) Precipitation of bicarbonate by the introduction of carbon dioxide (from stages 2 and 6)
   $NaCl + H_2O + NH_3 + CO_2$
   $\longrightarrow NH_4Cl + NaHCO_3$
5) Filtering and washing of precipitated bicarbonate
6) Thermal decomposition of bicarbonate to sodium carbonate (the CO_2 evolved is recycled to stage 4)
   $2\,NaHCO_3 \longrightarrow Na_2CO_3 + H_2O + CO_2$
7) Production of milk of lime
   $CaO + H_2O \longrightarrow Ca(OH)_2$
8) Recovery of ammonia by distillation of the mother liquor from stage 4 (bicarbonate precipitation) with milk of lime (the ammonia liberated is recycled to stage 3)
   $2\,NH_4Cl + Ca(OH)_2$
   $\longrightarrow 2\,NH_3 + CaCl_2 + 2\,H_2O$

The liquid remaining after distillation is usually discarded in its entirety, because only a small fraction can be used for calcium chloride production, depending on demand.

The process can be represented by the following simplified equation:

$$2\,NaCl + CaCO_3 \longrightarrow Na_2CO_3 + CaCl_2$$

In aqueous solution, the reaction proceeds from right to left due to the low solubility of calcium carbonate. The role of ammonia is to promote the formation of sodium bicarbonate via the intermediate ammonium bicarbonate.

The reaction that takes place in aqueous solution

$$NaCl + NH_4HCO_3 \rightleftharpoons NaHCO_3 + NH_4Cl$$

is determined by the solubilities of the individual salts in the presence of other components of the reciprocal salt pairs. These solubility relationships have been investigated by FEDOTIEFF [9].

The basic principles can be explained by using the 15 °C isotherm (Fig. 3), based on the values given in Table 2. The figure is the orthogonal projection of a three-dimensional diagram (i.e., a four-sided pyramid whose edges represent the coordinates of NaCl, NH_4Cl, NH_4HCO_3, and $NaHCO_3$). This diagram shows the ranges of existence of the individual solid phases that are in equilibrium with their saturated solutions at 15 °C:

Area a – I – P_2 – II:	NaCl
Area II – P_2 – P_1 – III – b:	NH_4Cl
Area P_1 – III – c – IV:	NH_4HCO_3
Area I – P_2 – P_1 – IV – d:	$NaHCO_3$

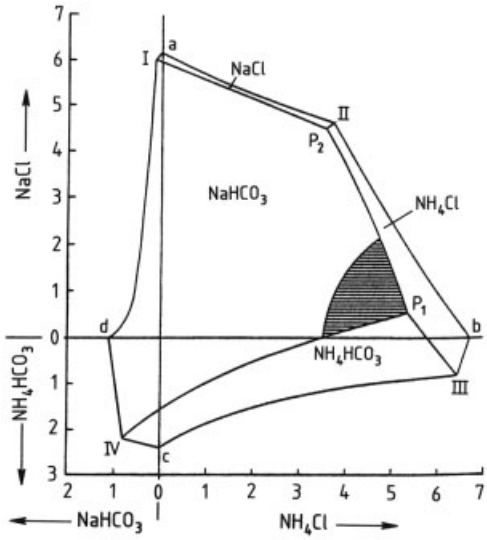

Figure 3. The 15 °C isotherm for the reciprocal salt pair $NaCl + NH_4HCO_3$ (concentrations in mol/1000 g H_2O are shown on the ordinate and abscissa)

Along the lines I – P_2, II – P_2, P_2 – P_1, P_1 – III, and P_1 – IV, two solid phases coexist, and at the points P_1 and P_2, three solid phases coexist. These two points are thus monovariant because in the presence of three solid phases, a change in concentration is possible only by changing the temperature.

The point P_1, at which the three salts NH_4Cl, NH_4HCO_3, and $NaHCO_3$ coexist, is of special importance for the ammonia – soda process, because this point has the highest NH_4Cl concen-

Table 2. Reciprocal salt pair $NaCl + NH_4HCO_3 \rightleftharpoons NaHCO_3 + NH_4Cl$ 15 °C isotherm, monovariant points (mol/1000 g H_2O)

Point	NaCl	$NaHCO_3$	NH_4Cl	NH_4HCO_3	Solid phase
a	6.12				NaCl
b			6.64		NH_4Cl
c				2.36	NH_4HCO_3
d		1.05			$NaHCO_3$
I	6.06	0.12			$NaCl + NaHCO_3$
II	4.55		3.72		$NaCl + NH_4Cl$
III			6.40	0.81	$NH_4Cl + NH_4HCO_3$
IV		0.71		2.16	$NaHCO_3 + NH_4HCO_3$
P_1	0.51	0.93	6.28		$NH_4Cl + NH_4HCO_3 + NaHCO_3$
P_2	4.44	0.18	3.73		$NH_4Cl + NaCl + NaHCO_3$

tration in the entire range. If point P_1 is reached by the precipitation of bicarbonate, it represents the maximum attainable sodium yield. Nevertheless, working so that neither the point P_1 nor the lines $P_1 - P_2$ and $P_1 - IV$ are actually reached is advisable, to avoid coprecipitation of $NaHCO_3$, NH_4Cl, or NH_4HCO_3. The coprecipitation of NH_4Cl has a particularly detrimental effect on the quality of the soda, because it is converted to NaCl on calcination:

$$NH_4Cl + NaHCO_3 \longrightarrow NaCl + NH_3 + CO_2 + H_2O$$

or

$$2\,NH_4Cl + Na_2CO_3 \longrightarrow 2\,NaCl + 2\,NH_3 + CO_2 + H_2O$$

An end point that is as close as possible to P_1 but lies within the hatched area in Figure 3 is best.

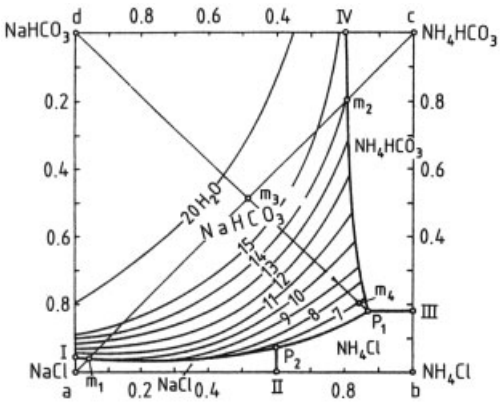

Figure 4. The 30 °C isotherm for the reciprocal salt pair $NaCl + NH_4HCO_3$
On the ordinate and abscissa of the square diagram, the amount of each salt is shown as a molar fraction of the total salt content of the solution.

The course of bicarbonate precipitation can be followed clearly in the square diagram of JÄNECKE [10]. Figure 4 shows the 30 °C isotherm of the system in such a diagram. This diagram does not represent absolute solubilities, but percentages of salts in the total salt content of the solution. Thus, the corners represent pure salts or solutions of a single salt, and the lines joining two corners represent mixtures of two salts. Points within the square represent mixtures or solutions of three salts. The solubilities can be found from the diagram if the appropriate water value is given for each point of equilibrium (the salt value being 1 kmol in all cases). This can be shown by means of lines of equal water content (isohydric lines).

The points, lines, and areas in Figure 4 have the same significance in principle as those in the 15 °C isotherm diagram of Figure 3. The corner point (a) corresponds to a solution of pure sodium chloride. If NH_4HCO_3 is added (by passing carbon dioxide and ammonia into the solution), the mixture moves from point (a) toward point (c) (NH_4HCO_3). In the case of saturation, systems represented by points on line $m_1 - m_2$ are in equilibrium with $NaHCO_3$ as solid phase.

In the ammonia – soda process, point (a) represents a solution that is not completely saturated (9.7 instead of 9.01 kmol H_2O/kmol NaCl). Point (m_3) on line a – c is reached by the introduction of CO_2 and NH_3 in an equimolar ratio. At this point, the water content necessary to give a saturated solution is ca. 17.5 kmol/kmol dissolved salt (see isohydric lines in Fig. 4). However, the proportion of water introduced with the brine, which is further reduced by the reaction of NH_3 with CO_2 to form NH_4HCO_3, is only ca. 7 kmol/kmol dissolved salt. This leads to the precipitation of sodium bicarbonate.

The precipitation process consists of separation of the supersaturated system (m_3) into the solid pure component $NaHCO_3$ (point d) and the saturated mother liquor (m_4).

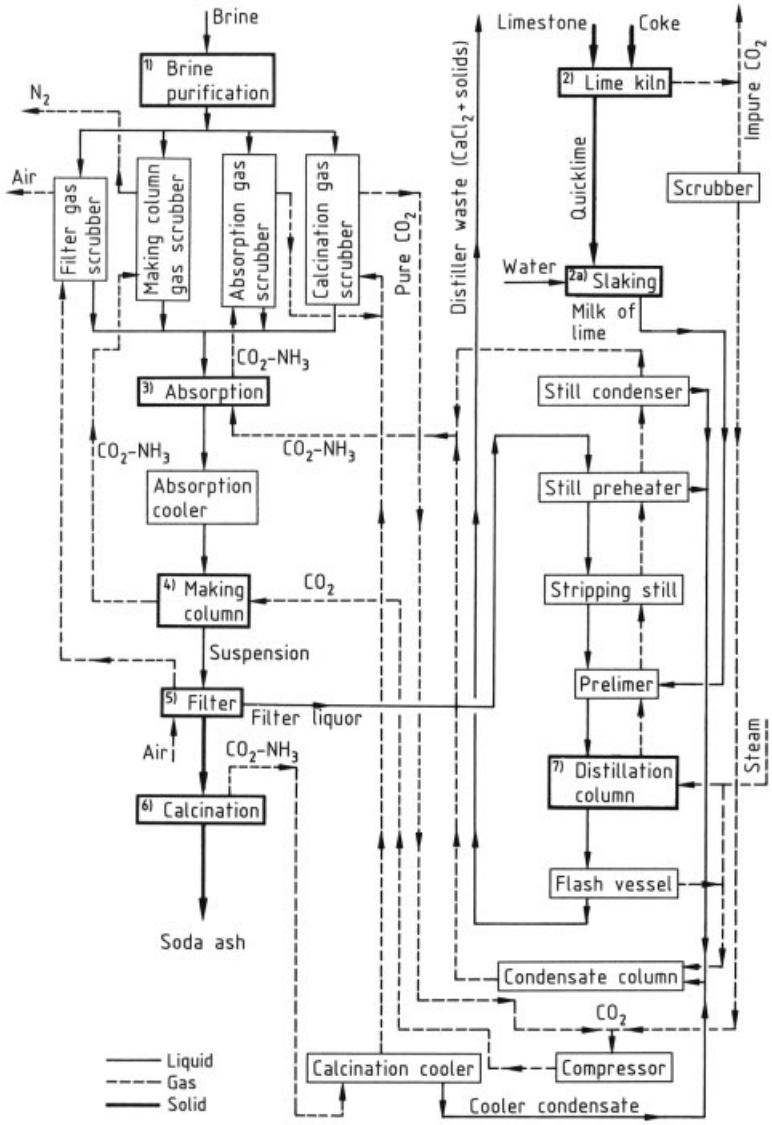

Figure 5. Flow sheet of the ammonia–soda process

In the Solvay process the degrees of utilization of Na^+ and NH_4^+ are both important. These can be calculated from the composition of the mother liquor (determined by chemical analysis) by using the equations:

$$U_{Na^+} = \frac{100\ (Cl^- - Na^+)}{Cl^-}$$

$$U_{NH_4^+} = \frac{100\ \left(NH_4^+ - HCO_3^-\right)}{NH_4^+}$$

In these equations, Cl^-, Na^+, NH_4^+, and HCO_3^- denote the concentrations in moles per kilogram of H_2O, and U_{Na^+} and $U_{NH_4^+}$ denote the percentage utilization of Na^+ and NH_4^+. The values for U_{Na^+} and $U_{NH_4^+}$ depend on the starting concentrations, and maximum Na^+ utilization does not correspond to maximum NH_4^+ utilization. These relationships are illustrated by the values given in Table 3.

Table 3. Percentage utilization of Na^+ and NH_4^+ in the reaction of NaCl with NH_4HCO_3 at 15 °C and various initial concentrations

Initial concentration of solution, mol/100 g H_2O		Point	Solution produced, g mol/1000 g H_2O				Utilization, %	
NaCl	NH_4HCO_3		HCO_3^-	Cl^-	Na^+	NH_4^+	Na^+	NH_4^+
8.188	3.730	P_2	0.18	8.188	4.62	3.730	43.58	95.17
7.658	4.552		0.31	7.658	3.39	4.552	55.73	93.19
7.128	5.450	↓	0.51	7.128	2.19	5.450	69.28	90.64
6.786	6.272	P_1	0.92	6.786	1.44	6.272	78.78	85.33
6.000	5.640		0.99	6.000	1.34	5.640	77.67	82.45
5.403	5.210	↓	1.07	5.403	1.27	5.210	76.49	79.46
5.025	4.919		1.12	5.025	1.23	4.919	75.52	77.23
4.000	4.135	IV	1.30	4.000	1.16	4.135	71.00	68.56

For economic operation of the process, the utilization of Na^+ should be high, rather than that of NH_4^+, since unconsumed NH_4^+ is recycled. Therefore, in practice, Na^+ utilization is between 72 and 76 %, and NH_4^+ between 69 and 71 %. However, note that consideration of phase rule theory gives no information about kinetic effects such as rates of reaction, nucleation, and crystallization.

A comprehensive treatment of the phase rule theory is given in [11, pp. 125 – 160].

1.4.1.2. Technology

The Solvay ammonia – soda process is shown schematically in Figure 5. The main operational stages are shown in the numbered bold rectangles:

Stage 1: Brine Purification. The sodium chloride used as a raw material in the Solvay process can be obtained by conventional or solution mining [12].

The brine always contains inorganic impurities that lead to production problems and impair the quality of the final product. For example, calcium and magnesium ions can cause scaling in the equipment and pipework due to the formation of basic carbonates during ammonia absorption, which can seriously interfere with the process. Purification of the brine, which should be as near saturation as possible to achieve optimum Na^+ utilization, is best carried out by the lime – soda process. Magnesium ions are precipitated as hydroxide, and calcium ions as carbonate:

$$Mg^{2+} + Ca(OH)_2 \longrightarrow Mg(OH)_2 + Ca^{2+}$$

$$a^{2+} + Na_2CO_3 \longrightarrow CaCO_3 + 2\,Na^+$$

The reagents used in this precipitation are prepared by dissolving soda ash or milk of lime in brine (Na_2CO_3 concentration: 65 – 80 kg/m^3; $Ca(OH)_2$ concentration: 170 – 185 kg/m^3). Reaction times, crystallization rates, and settling rates are improved if the brine with the added reagent solutions is mixed with a recycled seed slurry. Under these conditions, precipitates of calcium carbonate and magnesium hydroxide can be produced with settling rates up to 1 m/h, and in some cases even 3 m/h.

Stage 2: Limestone Burning and Lime Slaking. The limestone must be of very high purity (SiO_2: < 3 %; $Fe_2O_3 - Al_2O_3$: < 1.5 %). From it, both carbon dioxide for the carbonation stage and calcium oxide for the distillative recovery of ammonia are produced by calcination with coke (see → Lime and Limestone).

The quantity of limestone used per tonne of soda ash produced is determined solely by the amount of calcium oxide required for ammonia recovery. Hence, the excess carbon dioxide is either released to the atmosphere or supplied to other consumers.

The consumption of 95 % pure limestone varies between ca. 1100 and 1200 kg/1000 kg sodium carbonate, and is therefore approximately 10 – 20 % higher than the theoretical value.

In the system $CaCO_3 \rightleftharpoons CaO + CO_2$, the equilibrium pressure of CO_2, p_{CO^2} varies logarithmically with absolute temperature T

$$\log p_{CO_2}\ (\text{bar}) = -\frac{8580}{T\,(\text{K})} + 7.26$$

Thus the CO_2 vapor pressure at ca. 900 °C is 0.1 MPa. However, at this temperature, decomposition is too slow. If the rate of decomposition is increased by increasing the temperature dead-burnt limestone may be produced, which does not react with water. The best calcination temperature, therefore, lies between 1050 and 1100 °C.

In soda ash plants, calcination is carried out in well-insulated shaft furnaces heated with coke [11, pp. 76–83]. Furnace gas CO_2 contents of 40–42 vol % can be obtained. If liquid or gaseous fuels are used, the CO_2 content of the gas is lower.

The hot, dust-laden gas is cooled and washed by direct contact with water. It then passes to the CO_2 compressors from which it is delivered to the carbonation stage.

Quicklime from the kiln is reacted with used cooling water (50–65 °C) in horizontal rotating drums to produce a suspension of $Ca(OH)_2$ in water (milk of lime) that is as concentrated as possible. In favorable circumstances, the CaO content of the milk of lime can reach 5.5 mol/L.

Before milk of lime is used in the distillative recovery of ammonia, coarse sand and unburnt limestone are removed by sieves, mixed with new limestone, and recycled to the kiln.

Stage 3: Ammonia Absorption. Gas from the distillation stage (stage 7) contains NH_3 and CO_2, and is saturated with water vapor. In the absorption stage, this gas is absorbed by purified brine. Here, the total amount of ammonia required in the process and ca. 20 % of the required carbon dioxide are added. The ammoniacal brine obtained contains ca. 85–90 kg/m^3 NH_3 and 40–50 kg/m^3 CO_2. The NaCl content is reduced from ca. 300–260 g/L due to the increase in the specific volume of the liquor caused by the addition of ammonia and by dilution with water.

The reactions that occur during absorption are strongly exothermic and can be represented by the following equations:

$$NH_3 + H_2O \longrightarrow NH_4OH\ \Delta H = -34\,900\,\text{kJ}$$

$$2\,NH_4OH + CO_2 \longrightarrow (NH_4)_2CO_3 + H_2O\ \Delta H = -90\,850\,\text{kJ}$$

The absorption towers are equipped with effective internal or external coolers to remove the heat of reaction and the sensible heat of the distillation gas (total: ca. 1.84 GJ/t sodium carbonate). A typical absorption plant consists of a tower containing eight bubble trays and an external intermediate cooler (Fig. 6).

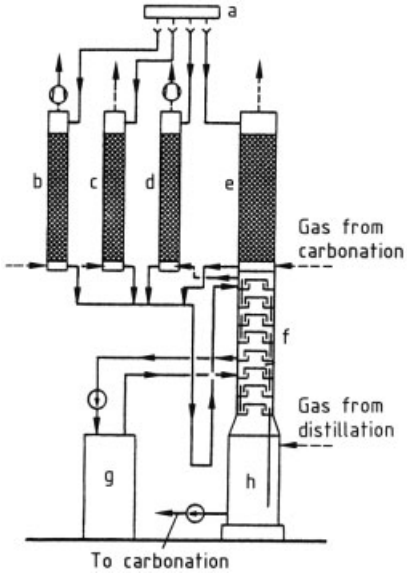

Figure 6. Absorption equipment with gas scrubbers
a) Brine distributor; b) Calcination gas scrubber; c) Filter gas scrubber; d) Absorption gas scrubber; e) Making column scrubber; f) Absorption; g) Absorption cooler; h) Reservoir

Stage 4: Precipitation of Bicarbonate (Making). The precipitation of sodium bicarbonate

$$NaCl + NH_4HCO_3 \longrightarrow NaHCO_3 + NH_4Cl$$

is an exothermic process in which ammoniacal brine from the absorption stage is reacted with carbon dioxide from the lime kiln (stage 2) or from the calcination stage (stage 6). In most ammonia–soda plants, cast iron columns of the type shown in Figure 7 are used for this reaction. In this double-tray column, which is equipped with tubular coolers in the lower part, ammoniacal brine is carbonated in countercurrent flow. The cooling water, also in countercurrent flow, must remove ca. 1.42 GJ/t of heat.

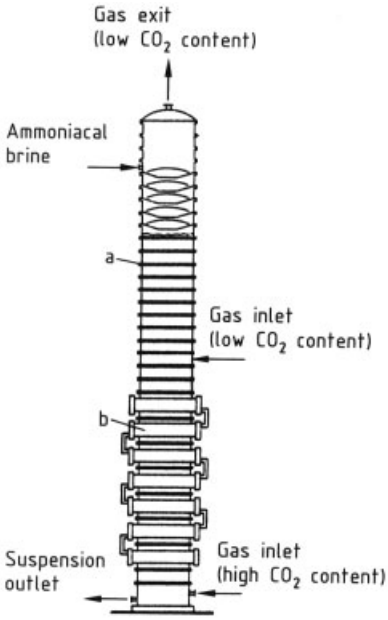

Figure 7. Making column
a) Double tray; b) Cooling tube bundle

Making columns become coated with the precipitated bicarbonate, which covers the inner parts including the cooling tubes, and must therefore be cleaned after 3 – 4 d of operation. This is carried out by washing with fresh ammoniacal brine while a stream of carbon dioxide gas is bubbled through (precarbonation). The washing liquor is then fed in parallel to the remaining columns, where it is carbonated in two stages. In the upper part of the columns (above the tube bundles), so- called weak gas (lime kiln gas) is used for carbonation, whereas in the lower part, more highly concentrated strong gas – a mixture of kiln gas and calcination gas – is used.

A distribution system for the liquid and gas streams enables each column to be used for either cleaning or making. In general, one out of four or five columns is a cleaning column.

A few soda ash plants still use the Honigmann process, in which carbonation is carried out in three to five vessels. CO_2- containing gas is bubbled through a central dip tube into the ammoniacal brine. Gas with the highest CO_2 content (calcination gas) is supplied to the vessel with the most strongly carbonated ammoniacal brine, and weak gas is fed to the vessel that contains fresh solution.

The temperature profile in both column and vessel processes exhibits a maximum of 50 – 60 °C. The suspension of bicarbonate in mother liquor leaves the carbonation stage at 30 °C.

Stage 5: Filtration of Bicarbonate. Bicarbonate is generally separated from the mother liquor in continuous rotary vacuum filters or band filters, and sometimes in centrifuges. Mother liquor adhering to the bicarbonate crystals is washed off with condensate produced during the production process or with softened water. The quantity of washing water required depends on the particle size of the precipitated product and varies between 0.3 and 1 m^3/t of bicarbonate.

Crude bicarbonate has the following approximate composition: $NaHCO_3$, 75.6 %; Na_2CO_3, 6.9 %; NH_4HCO_3, 3.4 %; $NaCl + NH_4Cl$, 0.4 %; H_2O, 13.7 %.

Centrifuged products contain only 7 – 9 % water. Calcination of 1 t of crude bicarbonate yields ca. 520 – 560 kg of soda ash.

Stage 6: Calcination of Bicarbonate. Thermal decomposition of crude bicarbonate to carbonate liberates carbon dioxide, ammonia, and water vapor. The product is technical-grade soda ash, which contains sodium chloride. The main reaction is

$$2\,NaHCO_3 \longrightarrow CO_2 + H_2O + Na_2CO_3$$

and side reactions are

$$NH_4HCO_3 \longrightarrow NH_3 + CO_2 + H_2O$$

$$NaHCO_3 + NH_4Cl \longrightarrow NH_3 + CO_2 + H_2O + NaCl$$

The heat requirement for the calcination of crude bicarbonate is 0.92 GJ/t soda, not including that required for drying the moist, filtered product. The total heat requirement for calcination in rotary calciners is 3.7 GJ/t soda. This heat is supplied at ca. 180 °C; the decomposition temperature of $NaHCO_3$ is 87.7 °C at a total $CO_2 - H_2O$ pressure of 0.1 MPa.

Calcination is generally carried out in rotary calciners. These are heated either externally with oil, gas, or coal, or internally with steam, which passes through heating tubes in the calciner countercurrent to the flow of solid material. Condensing steam gives much better heat

transfer than combustion gas, so that the output from steam-heated rotary calciners is considerably higher than that of calciners heated by combustion processes. The advantages of steam-heated calciners compared with those heated by combustion include the absence of thermal damage to the calciner, smaller space requirements and hence lower equipment costs, and ca. 15 % less fuel consumption.

Gases extracted from the calciners contain CO_2, NH_3, H_2O, and some leakage air, together with large amounts of dust (sodium carbonate and bicarbonate), which is removed in cyclones or electrostatic filters. Before this gas is used in the carbonation stage, it is cooled and washed with brine and water until it is free of ammonia. The condensate produced at this point is fed to the ammonia recovery stage.

If the soda ash is required in the form of solution or hydrated crystals, wet calcination can be used. This is carried out in a tower with steam in countercurrent flow. In this thermal decomposition, a sodium carbonate – sodium bicarbonate equilibrium is established, which limits the yield to 85 – 88 %.

Stage 7: Recovery of Ammonia. Ammonia, present in the filtrate as ammonium carbonate, hydrogencarbonate, sulfate, and chloride, is recovered by distillation followed by absorption. Both $(NH_4)_2CO_3$ and NH_4HCO_3 are completely decomposed at 85 – 90 °C, liberating all of the carbon dioxide and a small part of the ammonia from solution. Liberation of ammonia from NH_4Cl and $(NH_4)_2SO_4$ requires reaction with milk of lime:

$$2\,NH_4Cl + Ca(OH)_2 \longrightarrow 2\,NH_3 + 2\,H_2O + CaCl_2$$

$$(NH_4)_2SO_4 + Ca(OH)_2 \longrightarrow 2\,NH_3 + 2\,H_2O + CaSO_4$$

This reaction must be preceded by thermal decomposition of $(NH_4)_2CO_3$ and NH_4HCO_3 to avoid precipitation of calcium carbonate and consequent consumption of additional milk of lime. Free ammonia is displaced from solution with low-pressure steam.

A distillation plant is shown schematically in Figure 8. Distillation gas, which contains NH_3 and CO_2, is cooled to ca. 55 – 60 °C in the still condenser and still preheater. At this temperature, no danger exists of blocking the pipework with crystallized salts formed by ammonia and carbon dioxide. The stripping still removes CO_2 and NH_3 by thermal decomposition of $(NH_4)_2CO_3$ and NH_4HCO_3, and in the prelimer, NH_4Cl and $(NH_4)_2SO_4$ are reacted with milk of lime. In the distillation column the ammonia content of the liquor is lowered to ca. 0.5 kg/t sodium carbonate by steam treatment. The CO_2 – NH_3-containing condensate from the gas cooler and heat exchanger is recycled to the stripping still or distilled in a special condensate tower.

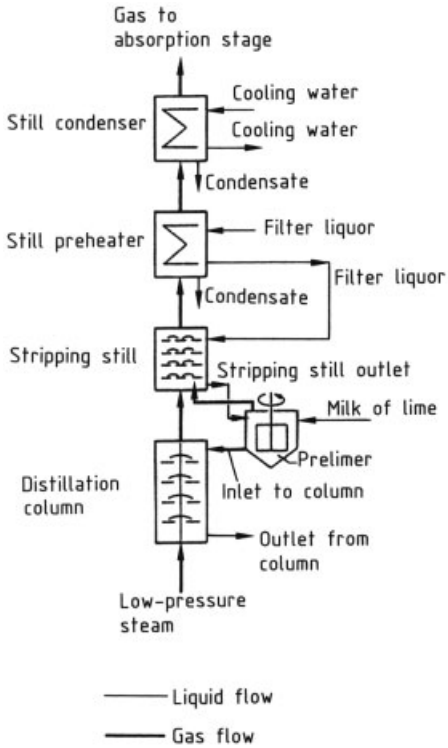

Figure 8. Flow diagram of stripping equipment

1.4.2. Modified Ammonia – Soda Process

Many variants of the ammonia – soda process exist [20]. Here, only the dual process, developed and operated in Japan, is described. This process combines soda ash production with ammonium chloride production. The importance of the process in Japan is due to the high cost of imported rock salt and use of the ammonium chloride as a fertilizer.

In the dual process [13] (see Fig. 9), ammonia is absorbed by the bicarbonate mother liquor, and solid sodium chloride is added. On cooling, ammonium chloride separates, is recovered in centrifuges, and is then dried in rotary dryers with air at 150 °C. The mother liquor is recycled to the carbonation towers where sodium bicarbonate is precipitated.

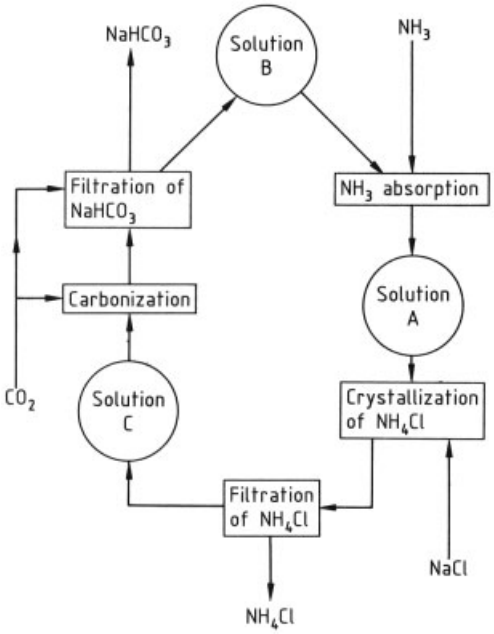

Figure 9. Flow diagram of the dual process
Solution A: 4.1 mol/L NH_4Cl + 1.05 mol/L NaCl; ϱ = 1.108 g/cm^3; t = 20 °C
Solution B: 3.45 mol/L NH_4Cl + 1.1 mol/L NaCl; ϱ = 1.110 g/cm^3; t = 5 °C
Solution C: 1.86 mol/L NH_4Cl + 3.73 mol/L NaCl; ϱ = 1.187 g/cm^3; t = 10 °C

Important differences exist between equipment used for the conventional Solvay process and that used for the ammonium chloride – soda process. In the dual process, ammonia is not recovered; hence no distillation equipment is required. Also, lime kilns are not required if other sources of CO_2 are available. Because of the extremely corrosive properties of NH_4Cl-containing solutions, they are produced in equipment made of special cast iron or plastics. As the mother liquor is recycled, special attention must be paid to the water balance of the system. The amount of water introduced into the system (e.g., for washing sodium bicarbonate and ammonium chloride) must be controlled continuously to maintain the correct quantity and composition. The quantity of sodium carbonate produced in 1970 by the dual process amounted to ca. 4.7 % of world production [14].

1.4.3. Production from Trona and Other Sodium Carbonate-Containing Minerals

The production of soda ash from trona in the United States has become significant since World War II. The raw material is either a naturally occurring brine containing sodium carbonate (e.g., from Searles Lake in California) or a trona deposit (e.g., from the Green River region of Wyoming).

Sodium carbonate brines are pumped from the lake and treated with CO_2 gas to precipitate sodium bicarbonate, which is filtered, washed, and calcined to form sodium carbonate.

In another process, the liquor is evaporated in solar ponds until trona ($Na_2CO_3 \cdot NaHCO_3 \cdot 2\,H_2O$) crystals separate. The trona contains ca. 45 % Na_2CO_3, 35 % $NaHCO_3$, 2 % NaCl, 1.5 % Na_2SO_4, 0.1 % SiO_2, 0.2 % insolubles, and 16 % H_2O. Calcination yields a product containing 95 – 96 % sodium carbonate.

In 1937, a large deposit of trona was discovered in the Green River region of Wyoming, but its significance was not recognized at first. Drilling work, which continued from 1944 to 1947, gave the first indication of the size of the deposit, which was subsequently extracted by the companies FMC, Stauffer Chemical, Allied Chemical, Texas Gulf (after 1976), and Tenneco Oil (after 1982) [15]. At first, mined trona (ca. 40 000 t/a) was simply calcined to produce crude sodium carbonate, but by 1952 an FMC plant was producing 300 000 t/a of purified sodium carbonate. Drilling work during the last 30 years has revealed reserves of $> 50\,000 \times 10^6$ t, mainly at depths of 250 – 450 m.

Natural trona from these deposits is a colorless or white mineral with a density of 2.13 g/cm^3 and a Mohs hardness of 2.5 – 4. The deposits generally contain > 90 % trona, with halite (NaCl), shortite ($Na_2CO_3 \cdot 2\,CaCO_3$), pyrite (FeS_2), shale, and other rock minerals. The United States now obtains its sodium carbonate entirely from trona; production capacity in 1991 was 9.88×10^6 t [15].

Processing is carried out mainly by the monohydrate process (see Fig. 10). Trona is crushed to < 6 mm and calcined at 300 °C in a rotary furnace to remove water of crystallization and convert the bicarbonate to carbonate. The calcined material is agitated with water to dissolve the crude soda ash. Insoluble components such as silica, sodium silicate, and gangue are removed by classifiers. The insoluble residues are then extracted with water. The concentrated liquor containing dissolved sodium carbonate is transferred to a thickener where the remaining finely divided insoluble residues are precipitated with flocculating agents. The clarified concentrated liquor is further purified by filter presses, and residual organic impurities are removed with activated carbon. The almost saturated solution is then evaporated in multiple-effect evaporators to give a 30 % slurry of $Na_2CO_3 \cdot H_2O$. This is followed by concentration to a 60 % solids content in hydrocyclones and centrifugation. The centrifuged monohydrate crystals, which still contain 2 – 8 % water, are calcined in rotary furnaces, while the filtrate is recycled to the process. The calcined product is ground to the required particle size and then sent to the storage silo.

In the sodium sesquicarbonate process, trona is first subjected to a multistage solution process to remove impurities. After vacuum crystallization, the sesquicarbonate crystals produced are separated from the mother liquor by centrifugation and calcined in rotary furnaces.

The largest trona deposit outside the United States is located at Magadi Lake in Kenya. This has an area of 65 km^2 and a thickness of 30 m, and is extracted by open-cast mining. Since 1990, the production capacity of the Magadi-Soda Co. (a subsidiary of Penrice Soda Products of Australia) has been 275 000 t/a [15–18].

1.4.4. Other Processes

Leblanc Process. The Leblanc process, which is now of only historical interest, is based on the following reactions:

$$2\,NaCl + H_2SO_4 \longrightarrow Na_2SO_4 + 2\,HCl$$

$$Na_2SO_4 + 2\,C \longrightarrow Na_2S + 2\,CO_2$$

$$Na_2S + CaCO_3 \longrightarrow Na_2CO_3 + CaS$$

The first plant using this process began operating in 1790 with a production of 300 t/d. By 1870, England had 40 factories and Germany 20. From 1885, Leblanc soda ash production declined because of competition from the more economical Solvay process, which also resulted in a higher-quality product. In England, the United Alkali Co. ceased production by the Leblanc process in 1916, and in Germany, Chemische Fabrik Rhenania in Stolberg, the last company using the process, operated until 1923.

A major disadvantage of the Leblanc process compared to the Solvay process is that it involves mainly solid-phase reactions and consumes large amounts of energy (at red heat or just below). The waste products calcium sulfide and hydrochloric acid are another disadvantage. Calcium sulfide caused both atmospheric and water pollution.

Production from Nepheline. In the former Soviet Union, pottery clay is produced from nepheline; the process also yields some sodium carbonate [19]. Finely ground nepheline is intimately mixed with very pure limestone, and the mixture is sintered. The following reaction takes place:

$$(Na, K)_2O \cdot Al_2O_3 \cdot 2\,SiO_2 + 4\,CaCO_3 \longrightarrow (Na, K)_2O \cdot Al_2O_3 + 2\,(2\,CaO \cdot SiO_2) + CO_2$$

The clinker produced is ground with water to dissolve the sodium and potassium aluminates. The aluminate solution is then treated with CO_2-containing gas (e.g., combustion gas), to precipitate aluminum oxide hydrate and give a solution of alkali-metal carbonates:

$$(Na, K)_2O \cdot Al_2O_3 + CO_2 + 3\,H_2O \longrightarrow 2\,Al(OH)_3 + (Na, K)_2CO_3$$

The solution is evaporated and separated into fractions containing various proportions of sodium and potassium carbonate, which differ in solubility. The bicarbonates present in these solutions are converted to carbonates by treatment with alkali-metal hydroxide solution before evaporation.

In the process for converting nepheline to aluminum oxide, 4 t of nepheline gives ca. 1 t of sodium carbonate.

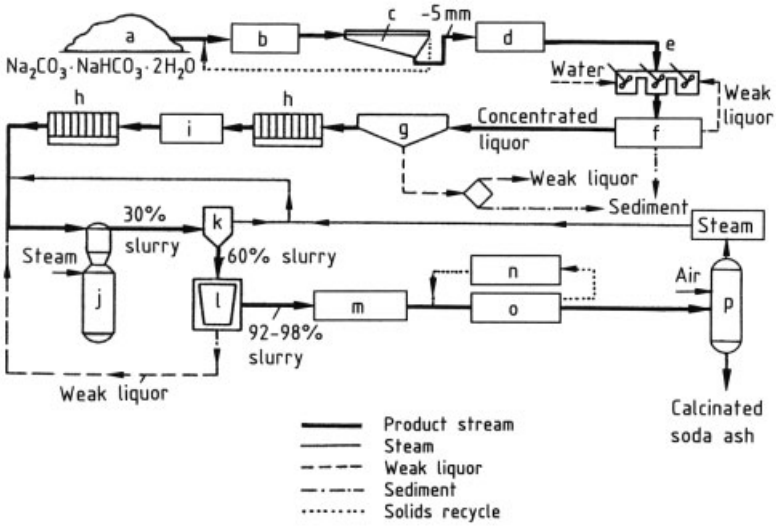

Figure 10. Conversion of trona to calcined soda ash (simplified flow diagram)
a) Trona storage; b) Crusher; c) Screen; d) Rotary furnace; e) Agitated dissolver; f) Classifier; g) Thickener; h) Filter press; i) Activated carbon filter; j) Vacuum crystallizer; k) Cyclone; l) Centrifuge; m) Dryer; n) Grinding mill; o) Classifier; p) Product storage

Other Proposed Modifications of the Ammonia – Soda Process. Attempts to produce sodium carbonate by other methods have been unsuccessful [20], [21]. A few proposals are mentioned here, although these have not been adopted industrially.

1) According to [22], recovery of ammonia by distillation can be avoided by using an intermediate salt; hydrochloric acid is produced in addition to sodium carbonate. The process consists of two stages. In the first, sodium bicarbonate is precipitated from an ammoniacal solution of sodium sulfate:
$2\,NH_3 + 2\,CO_2 + 2\,H_2O + Na_2SO_4$
$\longrightarrow (NH_4)_2SO_4 + 2\,NaHCO_3$
In the second stage, sodium sulfate is regenerated from the mother liquor produced in the above reaction. Here, either pure ammonium sulfate is precipitated in the first step,
$Na_2SO_4\,(s) + (NH_4)_2SO_4\,(aq)$
$\longrightarrow Na_2SO_4\,(aq) + (NH_4)_2SO_4\,(s)$
or a double salt is produced:
$2\,Na_2SO_4\,(s) + (NH_4)_2SO_4\,(aq) + 4\,H_2O$
$\longrightarrow Na_2SO_4\,(aq) + Na_2SO_4 \cdot (NH_4)_2SO_4 \cdot 4\,H_2O\,(s)$
The solid ammonium sulfate obtained is heated with sodium sulfate:
$(NH_4)_2SO_4 + Na_2SO_4$
$\longrightarrow 2\,NaHSO_4 + 2\,NH_3$
Ammonia is thereby recovered and recycled to the process.
Then, $NaHSO_4$ is fused with NaCl, to produce hydrogen chloride:
$NaHSO_4 + NaCl \longrightarrow Na_2SO_4 + HCl$
2) The following variation of the Solvay process is proposed in [23]: Ammonia from the bicarbonate mother liquor is recovered by reaction with magnesium hydroxide, not calcium oxide. Magnesium chloride and sodium chloride are present in the distillation liquor after removal of ammonia. Magnesium chloride is recovered as the monohydrate by evaporation and then decomposed:

$$MgCl_2 \cdot H_2O \xrightarrow{200\,^\circ C} MgOHCl + HCl$$

$$MgOHCl \xrightarrow{500\,^\circ C} MgO + HCl$$

The hydrogen chloride liberated is sold as such or after oxidation to chlorine, and magnesium oxide is recycled to the process. Apart from the fact that the hydrogen chloride produced is difficult to market, doubts exist that the decomposition of $MgCl_2$ is feasible in practice because of corrosion problems.
3) The following proposal was made by Blumberg et al. [24], [25]. In the equation
$NaCl + H_2O + CO_2 \rightleftharpoons NaHCO_3 + HCl$

the reaction proceeds to the right if hydrogen chloride is removed. This is achieved by reaction with a $C_8 - C_{12}$ tertiary amine dissolved in *n*-pentanol. The amine is regenerated by reaction with magnesium hydroxide. The 26 % solution of $MgCl_2$ produced is thermally decomposed to hydrochloric acid and magnesium oxide, which is recycled to the regeneration stage.

The starting material is brine containing 24 – 26 % NaCl and CO_2 dissolved at a partial pressure of ca. 0.3 MPa. Sodium bicarbonate is precipitated as a solid salt and recovered from the bottom of the vessel. The liquid separates into two phases: a solution of NaCl, which is recycled, and organic solvent saturated with hydrogen chloride. The equations are as follows:

$$NaCl + CO_2 + H_2O + NR_3 \rightleftharpoons NaHCO_3 + NR_3 \cdot HCl$$

$$NR_3 \cdot HCl + 1/2\, Mg(OH)_2 \rightleftharpoons 1/2\, MgCl_2 + NR_3 + H_2O$$

$$1/2\, MgCl_2 + H_2O \rightleftharpoons 1/2\, Mg(OH)_2 + HCl$$

The process is problematic because it results in pollution of the environment with amines and organic solvents. Another problem is the hydrogen chloride byproduct.

4) The combined production of sodium carbonate, chlorine, and a mixture of ammonium and sodium nitrates has been proposed in [26]. Sodium chloride is oxidized by excess nitric acid, to produce sodium nitrate and chlorine:

$$6\, NaCl + 6\, HNO_3 + 3/2\, O_2 \longrightarrow 6\, NaNO_3 + 3\, Cl_2 + 3\, H_2O$$

The sodium nitrate solution is neutralized with ammonia and then reacted with carbon dioxide to form sodium bicarbonate.

After the bicarbonate is filtered off, a mixture of ammonium and sodium nitrates that can be used as a fertilizer is recovered from the filtrate.

1.4.5. Wastewater Problems in Sodium Carbonate Production [27]

In the recovery of ammonia by distillation with milk of lime, an end liquor (ca. 7 – 10 m^3/t sodium carbonate) is produced at its boiling point. This liquor has the following approximate composition: 120 – 180 g/L $CaCl_2$, 1.0 – 1.7 g/L $Ca(OH)_2$, 50 – 75 g/L NaCl, 1.0 – 1.5 g/L $CaSO_4$, and 20 – 30 g/L suspended solids. These solids originate mainly from impurities in the limestone and coke, and include uncalcined or dead-burnt limestone. Also, some solid material results from the reaction of milk of lime with residual carbon dioxide not removed in the stripping still.

The solid material consists of ca. 67 % $CaCO_3$, 10 % $Mg(OH)_2 + Ca(OH)_2$, 10 % $CaSO_4 \cdot 2\, H_2O$, 10 % SiO_2, and 3 % $Fe_2O_3 + Al_2O_3$. It is very finely divided and therefore difficult to remove. The liquor is often clarified in large settling basins or ponds. However, mud with a high chloride content is produced, which presents disposal problems. Therefore, the end liquor is frequently diluted with used cooling water and discharged directly into the receiving water body. This method of disposal is generally recognized as safe because calcium carbonate, which is the main solid component in the wastewater, reacts with carbon dioxide in the water to form calcium bicarbonate solution.

The amount of $CaCl_2$ in the end liquor is given by the overall equation for the Solvay process:

$$2\, NaCl + CaCO_3 \longrightarrow Na_2CO_3 + CaCl_2$$

This shows that the reaction produces 1 mol $CaCl_2$ in the end liquor for each mole of Na_2CO_3 produced. The end liquor also contains some NaCl resulting from incomplete reaction. Removal of chloride from the end liquor should be considered only if at least some of it can be used to produce calcium chloride. However, this is carried out in only a few sodium carbonate plants, because the demand for this chemical is relatively low.

1.5. Quality Specifications

The product obtained by the ammonia – soda (Solvay) process is, after calcination, an anhydrous, technical-grade sodium carbonate. The two main commercial forms are light soda ash and dense soda ash, which differ in particle size and bulk density. Light soda ash has a bulk density of 0.5 – 0.6 kg/L; dense soda ash 1.05 – 1.15 kg/L.

The quality of the product is normally assessed by the following tests:

1) Loss on heating at 250 °C
   Method: Drying of sample at 250 °C for 2 h in a crucible furnace
2) Water-insoluble material
   Method: Determination of residues insoluble in water at 50 °C
3) Total alkalinity expressed as Na_2CO_3
   Method: Titration with 1 M hydrochlorid acid, with methyl orange as indicator
4) Sodium bicarbonate ($NaHCO_3$)
   Method: Winkler titration – conversion of $NaHCO_3$ to Na_2CO_3 by addition of a known amount of NaOH; precipitation of total CO_2 as $BaCO_3$; back titration of excess NaOH with HCl (indicator: phenolphthalein)
5) Sodium carbonate (Na_2CO_3)
   Calculation: Total alkalinity (method 3) minus alkalinity due to $NaHCO_3$ (method 4)
6) Sodium chloride (NaCl)
   Methods:
   a) Mohr's titration against 0.1 M $AgNO_3$, with K_2CrO_4 as indicator
   b) Argentometrically with potentiometric determination of the end point
7) Sodium sulfate (Na_2SO_4)
   Method: Gravimetrically by precipitation of the sulfate as $BaSO_4$
8) Iron oxide (Fe_2O_3)
   Method: Colorimetric (photometric) determination of the concentration of Fe(II)–dipyridyl complex
9) Water (H_2O)
   Calculation: Weight loss at 250 °C (method 1)minus wight loss due to $CO_2 + H_2O$ determined from decomposition of $NaHCO_3$ ($NaHCO_3$ by method 4).

A typical analysis is given below:

Na_2CO_3	99.6 %	Fe_2O_3	0.002 %
NaCl	0.15 %	CaO	0.01 %
Na_2SO_4	0.02 %	MgO	0.02 %

In addition to its chemical analysis, the product is also characterized by particle size and bulk density.

For soda ash produced from trona, the following are typical analysis figures:

Na_2CO_3	99.6 %	Fe_2O_3	0.001 %
NaCl	0.035 %	CaO	0.01 %
Na_2SO_4	0.1 %	MgO	0.003 %

The difference in particle-size distribution between light and dense soda ash is shown by the following value:

	Light ash (%)	Dense ash (%)
> 2 mm	0.1	0.2
2 – 1 mm	0.5	5
1 – 0.5 mm	1	26
0.5 – 0.25 mm	10	50
0.25 – 0.125 mm	30	14
0.125 – 0.063 mm	38	4
< 0.063 mm	20.4	0.8

Soda ash of the quality described above is suitable for most applications and is usually adequate for the food industry.

Other, special grades include dense granulated soda and briquetted soda. So- called washing soda crystals, the decahydrate $Na_2CO_3 \cdot 10\,H_2O$, is now of minor importance. This is produced from calcined sodium carbonate by dissolving it in boiling water until the density is 30 – 40° Bé. Hard crystals are obtained only if the solution contains some sulfate, and a small percentage of this substance is therefore added to the solution. The solution is clarified by allowing solid impurities to sediment. The solution is then cooled to 38 °C in crystallizing troughs to obtain soda crystals. A typical analysis of the commercial product is 36.6 % Na_2CO_3, 0.03 % NaCl, 0.15 % Na_2SO_4, 63.22 % H_2O.

1.6. Producers

Western Europe producers of soda ash are listed below:

Country	Company	Location
Germany	Solvay Deutschland GmbH	Rheinberg, Heilbronn, Bernburg
	Chemische Fabrik Kalk GmbH *	Köln-Kalk
	Matthes & Weber GmbH	Duisburg
	Sodawerk Stassfurt GmbH	Stassfurt
Belgium	Solvay S.A.	Couillet
France	Solvay S.A. Rhône-Poulenc	Dombasle La Madeleine
Italy	Solvay S.A.	Rosignano
Netherlands	AKZO	Delfzijl
Portugal	Soda Povoa SARL (Solvay subsidiary)	Povoa de Santa Iria
Spain	Solvay S.A.	Torrelavega
United Kingdom	Penrice Soda Products Ltd.	Winnington

* Closure planned for end of 1993.

In Eastern Europe plants in the following countries are producing soda ash by the Solvay process: Bulgaria, Poland, Romania, the former Yugoslavia, and the CIS.

The United States has the following production plants:

FMC Corporation	Green River, Wyoming
General Chemical Corp.	Green River, Wyoming
North American Chemical	Argus, California
Rhône-Poulenc	Green River, Wyoming
Solvay Minerals (formerly Tenneco) in partnership with Asahi Glass	Green River, Wyoming
Texasgulf Chemical Co.	Green River, Wyoming

The most important producers outside Europe and the United States are China, India, and Japan.

1.7. Storage and Transport

Calcined soda ash is hygroscopic, gradually taking up moisture and carbon dioxide during transportation and storage. This converts the surface of sodium carbonate to the bicarbonate and leads to a weight increase of up to 17 %. Soda ash is therefore best stored in closed, dry areas.

Soda ash is supplied either bulk in silo wagons that can be discharged pneumatically or in paper sacks.

Soda crystals (washing soda) can effloresce or dissolve in the water of crystallization, depending on the temperature and relative humidity. They melt at 32 – 33 °C. At lower temperature, they are stable, provided the relative humidity is within the limits listed below:

T, °C	0	10	15	25
R.H., %	59 – 98	64 – 96	68 – 94	76 – 87

At lower humidity, soda crystals effloresce, and at higher humidity the surface becomes wet.

1.8. Uses

Soda ash is an important raw material for the chemical industry. Its sodium content results in fluxing properties that make it important in the glass and silicate industries. It is used to neutralize inorganic and organic acids or acidic salts, and to maintain a constant pH in processes where acids are liberated. It is also utilized in the production of sodium salts (e.g., phosphates, nitrates, chromates, citrates, tartrates, salts of fatty acids).

Soda ash can be used in aqueous solution to remove sulfur dioxide from process gas or off-gases, forming sodium sulfite and sodium bicarbonate:

$$Na_2CO_3 + SO_2 \longrightarrow Na_2SO_3 + CO_2$$

$$CO_2 + Na_2CO_3 + H_2O \longrightarrow 2\,NaHCO_3$$

$$2\,Na_2CO_3 + SO_2 + H_2O \longrightarrow Na_2SO_3 + 2\,NaHCO_3$$

These reactions are also important in the production of paper pulp by the sulfite process.

Other uses of sodium carbonate are as follows:

Chemical industry:

production of bleaching agents, borax, chromates and dichromates, fertilizers, dyes, fillers, tanning agents, industrial cleaning agents, catalysts, cryolite, adhesives, metal carbonates, sodium nitrate, perborates, phosphates, silicates, ultramarine pigments, soluble silicates, etc.

Detergent and soap industry: manufacture of detergents and saponification of fats.

Petrochemicals: neutralization.

Pulp and paper Industry: cooking wood; neutralization, cleaning, bleaching, and treatment of recycled paper.

Artificial silk industry: deacidification of artificial silk.

Textile industry: dyeing, bleaching, and finishing of wool and cotton.

Coke ovens, gasworks, and hydrogenation plants: gas purification (desulfurization).

Iron and steel industry: removal of sulfur and phosphorus from pig iron, cast iron, and steel; ore beneficiation, flotation agents, and fluxing agents.

*Heavy-metal industry:*digestion and beneficiation of ores (e.g., of antimony, lead, chrome, cobalt, nickel, bismuth, and tin).

Glass industry: raw material for the glass melt and for reacting with sand.

Aluminum industry: reaction with bauxite.

Ceramics industry: production of refractory materials and glazes.

Enamel industry: as a flux.
*Food industry:*hydrolysis of proteins, production of margarine and starch, and softening of sugar beet juice.
Various branches of industry: water treatment, metal degreasing.
Environmental protection:

1) Purification of flue gases by injecting sodium carbonate or sodium hydrogencarbonate (dry process).
2) Regeneration of acidic lakes by the introduction of briquettes of sodium carbonate, so that organic sediments exhibit an alkaline reaction over a long period.

1.9. Toxicology

The lowest known lethal dose of sodium carbonate is 4000 mg/kg (rat, oral). Toxic effects do not occur under normal working conditions. For humans, oral ingestion of > 15 g is potentially lethal. In case of ingestion, vomiting should not be induced, but copious amounts of water and dilute lemon juice or vinegar (two tablespoons per glass of water) should be drunk. The stomach can be carefully irrigated for 15 min at most, with the usual precautions, but use of this treatment for longer periods is strongly contraindicated (danger of perforation). Medical help should be sought.

Because of its alkaline reaction, sodium carbonate has a irritating effect on the skin and mucous membranes. If skin is attacked, the affected part should be thoroughly washed with copious quantities of water, and a nonirritant dressing applied if necessary. If eyes are splashed with sodium carbonate, the eyelids should be held open while the eye is treated for several minutes with running water or saline solution. Medical help should be sought from a specialist.

1.10. Economic Aspects

The production of soda ash increased continuously after World War II until 1975.

World production figures (in 10^6 t) are given below:

1900	1.5	1965	16.6
1913	2.85	1970	20.0
1932	5.0	1975	24.9
1940	7.0	1980	28.9
1951	11.0	1985	29.4
1960	12.7	1990	32.3

Regional figures are listed in Table 4. The proportion of soda ash produced from trona has increased continuously in the United States from 15 % in 1960, to 70 % in 1975, and 100 % since 1986. In 1991, 28 % of worldwide soda ash capacity was based on trona or other minerals containing sodium carbonate [15].

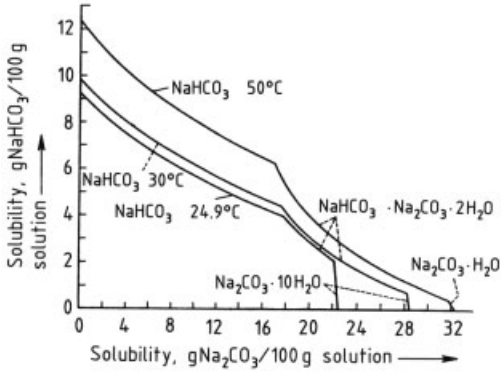

Figure 11. Solubilities in the system $H_2O - NaHCO_3 - Na_2CO_3$ at 24.9 °C, 30 °C, and 50 °C [29]

2. Sodium Hydrogencarbonate

2.1. Properties

Some physical properties of sodium hydrogencarbonate are given below:

M_r	84.007
Density	2.22 g/cm^3
Specific heat capacity (25 °C)	87.7 kJ mol^{-1} K^{-1}
Enthalpy of solution	− 18 kJ/mol
Enthalpy of formation	950 kJ/mol
Refractive indices	1.380, 1.500, 1.586
Dielectric constant (25 °C)	4.39

Solubility properties are listed in Table 5. Solubility is lower in the presence of sodium carbonate.

In the system $H_2O - NaHCO_3 - Na_2CO_3$, the double salt trona $NaHCO_3 \cdot Na_2CO_3 \cdot 2\,H_2O$ occurs at > 21.3 °C. The isotherms at 24.9 °C, 30 °C, and 50 °C are shown in Figure 11, which

Table 4. Soda ash production (10^3 t) by region [15]

Region	1975	1980	1985	1990
Western Europe	5879	6747	5966	5790
North America	6819	7857	8149	9481
Latin America	588	744	774	790
Africa	101	233	270	287
Asia	1955	2332	2648	3580
Australia	150	250	300	300
Centrally planned economies	9498	10 742	11 390	12 101

indicates the ranges of existence of the solid phases at these temperatures [29], [30].

On heating, sodium hydrogencarbonate decomposes into sodium carbonate, carbon dioxide, and water:

$$2\,NaHCO_3 \longrightarrow Na_2CO_3 + CO_2 + H_2O$$

Dissociation pressures of carbon dioxide at equilibrium are as follows:

T, °C	30	50	70	90	100	110
p_{CO^2}, kPa	0.825	3.99	16.013	55.102	97.236	166.596

At room temperature, sodium hydrogencarbonate is fairly stable because of its low rate of decomposition. The aqueous solution has an only slightly alkaline reaction due to the small extent of hydrolysis.

2.2. Production

Sodium hydrogencarbonate is an intermediate in the ammonia – soda process. However, because of the content of ammonium salts, a product that satisfies the quality requirements of consumers (mainly in the food industry) cannot be obtained by drying crude bicarbonate. Moreover, a sufficiently pure product cannot be obtained by recrystallization. Therefore, it is necessary to start with an aqueous solution of sodium carbonate which is obtained either by dissolving calcined soda ash or by decomposing crude bicarbonate with steam. This is then filtered and carbonated with pure concentrated carbon dioxide, with cooling to remove the heat of reaction:

$$Na_2CO_3 + CO_2 + H_2O \rightarrow 2\,NaHCO_3 \quad \Delta H = -43.67\,kJ$$

As carbonation proceeds, sodium hydrogencarbonate precipitates. It is recovered by centrifuging and then dried with hot air (e.g., in tray dryers).

Table 5. Solubility of sodium hydrogencarbonate in water

T, °C	Solubility	
	g/100 g solution	g/100 g H_2O
– 2.35	6.26	6.68
0	6.4	6.9
10	7.6	8.2
20	8.7	9.6
30	10.0	11.1
40	11.3	12.7
50	12.7	14.5
60	14.2	16.5
80	16.5	19.7
100	19.1	23.6

2.3. Uses and Quality Specifications

Sodium hydrogencarbonate is used in the manufacture of baking powder, as a medicament for neutralizing stomach acid, as a component of effervescent powders, in animal feed, and as a dry-powder fire extinguisher.

The following analysis is typical for the food-grade product: $NaHCO_3$, 99.7 %; Na_2CO_3, 0.2 %; NaCl, 0.004 %; H_2O, 0.05 %; insolubles, 0.003 %.

Sodium hydrogencarbonate tends to form lumps in the presence of moisture or under pressure. On heating to > 60 °C, appreciable decomposition to sodium carbonate occurs. The bicarbonate readily picks up odors from its surroundings and should therefore be stored under dry, airtight conditions.

3. References

1. *International Critical Tables,* **vol. III, p. 372, vol. IV, pp. 237, 259,** McGraw-Hill, New York 1928.
2. L. Lortie, P. Demers, *Canadian J. Res.* **18 (B)** (1940) 164.
3. A. F. Newkirk, J. Aliferis, *Anal. Chem.* **30** (1958) 982–984.
4. W. L. W. Ludekens, T. Thirunamachandran, *Chem. Ind. (London)* 1954, 1265–1266.
5. C. Kröger, E. Fingas, *Z. Anorg. Allgem. Chem.* **212** (1933) 257–268, 264.
6. F. Ishikawa, F. Murooka, H. Hagisawa, *Abstr. Rikawagaku-Kenhyu-jo Iho* **5** (1932) 23–24; *Sci. Rep. Tohuku Imp. Univ.* **22** (1933) 1179–1196.
7. E. G. Bunzel, E. J. Kohlmeyer, *Z. Anorg. Allgem. Chem.* **254** (1947) 1–30.
8. Te-Pang Hou: *Manufacture of Soda,* 2nd ed., Reinhold Publishing Corp., New York 1942, p. 17.
9. P. P. Fedodieff, *Z. Anorg. Chem.* **17** (1904) 1644–1659.
10. E. Jänecke, *Z. Angew. Chem.* **20** (1907) 1559–1564.
11. Z. Rant: *Die Erzeugung von Soda nach dem Solvay Verfahren,* F. Enke-Verlag, Stuttgart 1968, pp. 76–83, 125–160.
12. H. Börger, *Bergbau Rundschau* **4** (1952) 84–88.
13. M. Imamura, *Ind. Eng. Chem.* **54** (1962) no. 2, 35–39.
14. United Nations Economic and Social Council, *ECE/CHEM* **14** (1976) Sept. 20, p. 5.
15. *The Economics of Soda Ash,* 6th ed., Roskill Information Services, London 1991.
16. B. M. Coope, P. W. Harben: "Soda Ash," in P. W. Harben (ed.): *Raw Materials for the Glass Industry,* Industrial Minerals Consumer Survey, 1977.
17. *Ind. Chem.* **31** (1955) no. 370, 547.
18. *Ind. Min.* 1975, no. 96, Sept., 23–27.
19. N. L. Talmud, *Khim. Promst. Moscow* 1961, 226–232.
20. *Ullmann,* 3rd ed. 12, p. 656.
21. *Winnacker-Küchler,* 2nd ed., **1,** pp. 210–213.
22. E. Stiers, DE-OS 1 592 039, 1967.
23. S. Lynn et al., US 3 792 153, 1972.
24. R. Blumberg, J. E. Gai, K. Hajdu, *Proceedings International Solvent Extraction Conference 1947,* **vol. 3,** Soc. Chem. Ind., London 1974, pp. 2789–2802.
25. Israel Mining Industries Institute for Research and Development, Israel, IL 33 552, 1969.
26. E. Pischinger, St. Bursa, H. Koneczny, J. Straszko, *Przem. Chem.* **54** (1975) 593–594.
27. F. Meinck, H. Stooff, H. Kohlschütter: *Industrie-Abwässer,* G. Fischer-Verlag, Stuttgart 1968, pp. 245–248.
28. Industrial Minerals Glass Survey 1977, pp. 39–49.
29. *Gmelin,* 8th ed. **System no. 21,** pp. 764–766, 1358–1363.
30. A. E. Hill, L. R. Bacon, *J. Am. Chem. Soc.* **49** (1927) 2487.

Sodium Borohydride → Hydrides
Sodium Bromide → Bromine Compounds

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Sodium Chloride

Standard Article

Gisbert Westphal[1], Gerhard Kristen[2], Wilhelm Wegener[3], Peter Ambatiello[4], Helmut Geyer[5], Bernard Epron[6], Christian Bonal[7], Georg Steinhauser[8]

[1]Solvay Salz GmbH, Solingen, Federal Republic of Germany
[2]Solvay Salz GmbH, Solingen, Federal Republic of Germany
[3]Südwestdeutsche Salzwerke AG, Heilbronn, Federal Republic of Germany
[4]Salzbergwerk Berchtesgaden, Federal Republic of Germany
[5]Salzgewinnungsgesellschaft Westfalen mbH, Ahaus, Federal Republic of Germany
[6]Compagnie des Salins du Midi et des Salines de l'Est, Paris, France
[7]Compagnie des Salins du Midi et des Salines de l'Est, Paris, France
[8]Atominstitut der Österreichischen Universitäten, Vienna University of Technology, Austria

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a24_317.pub3
Article Online Posting Date: April 30, 2005

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. History
2. Properties
3. Formation and Occurrence of Salt Deposits
4. Production
4.1. Mining of Rock Salt
4.2. Production of Crude Brine by Mining Methods
4.3. Controlled Solution Mining
4.3.1. Drilling and Construction of the Borehole
4.3.2. The Process of Solution Mining
4.3.3. Planning of the Extraction Process
4.3.4. Other Systems
4.3.5. Equipment
4.3.6. Storage Caverns
4.4. Production of Sea Salt
4.4.1. The Raw Material: Sea Water
4.4.2. The Main Factors Governing Production of Sea Salt
4.4.3. Production Stages in a Modern Salt Field
5. Production of Pure Salt by Evaporation of Brine
5.1. Purification of Crude Brine
5.2. Evaporation of Brine
5.3. Recrystallization
5.4. Other Process Steps
5.5. Salt Quality
6. Economic Aspects
6.1. Commercial Grades
6.1.1. Salt for Food
6.1.2. Treated Salt
6.1.3. Industrial salt
6.1.4. De-icing Salt

6.2. Packaging, Transport, and Storage
6.3. Economic Data

[Top of Page]

1. History [1-6]

Gisbert Westphal

The history of mankind is also a history of salt. Although it is not clear how man learnt of its effects and its savory properties, salt has been mentioned in literature for thousands of years.

In the Odyssey, HOMER wrote: “travel . . . until you meet mortals who do not know the sea and who never eat food seasoned with salt.” Elsewhere he refers to salt as “holy.” CASSIODORUS said: “Gold can be dispensed with, but not salt.” Other notable historical sources that refer to salt include PLUTARCH, PLINY, HERODOTUS, the Old and New Testaments, and ancient Chinese writings.

Before the industrial revolution in the 1800s, salt was used almost entirely as a foodstuff and as a preservative. Its importance led people of all cultures to strive to find ways of obtaining this important, life-sustaining substance. The coastal regions of the Earth were the favored areas for salt production. Possibly the oldest mine in the world was worked in ca. 1000 B. C. at Hallstatt, Austria which gave its name to the “Hallstatt Culture.” In Germany, salt was probably first won in Schwäbisch Hall in ca. 500 B. C. The well-known salt deposit in Wieliczka, Poland was first mentioned in the 1100s. Wherever natural brine was available or early methods of brine production were practiced, a large number of salt works came into being between the 1100s and 1700s.

Salt was used universally, but was not everywhere available from natural sources, so that trading rapidly became important. The extent to which salt was used as an article of commerce is shown by the appearance of the famous salt roads in China, Africa, and Europe, and by the foundation of towns such as Salzburg, Salzach, Hall, Hallein, and Heilbronn (Celtic: *hal* = salt). Also, salt was used as currency (hence the word “salary”) and as source of tax revenue. Ancient Rome had a salt tax, and salt monopolies were established in many countries and continue to exist in Austria and Switzerland. Starting in 1765, Prussia demanded an annual salt payment from its citizens, and in Germany a tax of € 60/t was levied on all salt used in food. The total revenue from this in 1990 was € 24×10^6. This tax was discontinued at the beginning of 1993, following the establishment of the European single market.

With the advent of the industrial revolution, the use of salt as a chemical raw material increased rapidly. Industrial salt production, whether by evaporation in salt works or from mined rock salt, continually increased in importance.

Another important process of recent decades is controlled solution mining.

Although salt is no longer the cause of warlike confrontations, it has certainly lost none of its fascination, whether as a life-giving condiment or as an indispensable, versatile raw material.

Many museums, e.g., the German Museum in Munich and the German Salt Museum, Lüneburg, have comprehensive exhibitions on the history of salt.

[Top of Page]

2. Properties [7]

Gerhard Kristen

Sodium chloride, NaCl, M_r 58.443, is a colorless salt with good solubility in water. Chemically pure NaCl crystallizes from aqueous solutions in well-formed cubes, which under the influence of surface tension often grow together into funnel-shaped, hollow, square-based pyramids. In the presence of impurities, octahedra or dodecahedra are sometimes formed. Crystallization from hydrochloric acid solution gives long, fibrous, needle-shaped crystals.

During crystallization, small amounts of water can be trapped in holes in the crystals. When this vaporizes on heating, it causes the crystals to explode with audible decrepitation. In salt from natural deposits, inclusions of gases such as methane, carbon dioxide, and hydrogen sulfide can occur.

In the crystal, the Na and Cl ions alternate. The ions of each type form a face-centered cubic lattice, in which each ion is surrounded octahedrally by six ions of the other type at a distance of $a/2$ (lattice constant a = 0.56273 nm). The modulus of elasticity perpendicular to the surface of the cube is 41 074 MPa. Under high pressure, slow flow takes place. Sodium chloride is highly transparent to light of wavelength between 200 nm (ultraviolet) and 15 µm (infrared). Ionizing radiation causes lattice defects (color centers) which give the salt a blue color. On heating to ca. 250 °C, this color disappears.

Some important physical properties are listed below and in Tables 1-3.

Melting point	801 °C
Boiling point	1465 °C
Density at 25 °C	2.1615 g/cm^3

Mohs hardness	2 – 2.5
Brinell hardness	14 HB
Dielectric constant	5.9
Refractive index n_D^{20}	1.5443
Electrical resistivity at 20 °C	4.6×10^{16} Ω cm
Electrical resistivity at 100 °C	1.38×10^{13} Ω cm
Thermal conductivity at 17 °C	0.072 W cm^{-1} K^{-1}
Linear coefficient of expansion	40.5 µm/mK
Specific heat capacity at 25 °C	850 J kg^{-1} K^{-1}
Enthalpy of formation at 25 °C	–410.9 kJ/mol
Entropy at 25 °C	72.36 J mol^{-1} K^{-1}
Latent heat of fusion	0.52 kJ/g
Latent heat of evaporation	2.91 kJ/g
Viscosity of saturated aqueous solution	1.93 mPa · s
Density of molten NaCl at 801 °C	1.549 g/cm^3
Viscosity of molten NaCl at 850 °C	1.29 mPa · s
Surface tension of molten NaCl at 850 °C	110 mN/m
Electrical conductivity of molten NaCl at 850 °C	3.7 S/cm

Table 1. Solubility of NaCl in water

Temperature, °C	g NaCl/100 g H_2O	NaCl, wt %	Density of solution, g/cm^3	NaCl content of solution, g/L
0	35.76	26.34	1.2093	318.5
20	35.92	26.43	1.1999	317.1
40	36.46	26.71	1.1914	318.2
60	37.16	27.09	1.1830	320.5
80	37.99	27.53	1.1745	323.3
100	39.12	28.12	1.1660	327.9
180	44.9	30.99		

Table 2. Density of aqueous solutions of NaCl

NaCl, wt %	Density, g/cm^3				
	0 °C	20 °C	40 °C	60 °C	80 °C
4	1.03038	1.02680	1.01977	1.0103	0.9988
8	1.06121	1.05589	1.04798	1.0381	1.0264
12	1.09244	1.08566	1.07699	1.0667	1.0549
16	1.12419	1.11621	1.10688	1.0962	1.0842
20	1.15663	1.14779	1.13774	1.1268	1.1146
24	1.18999	1.18040	1.16971	1.1584	1.1463

Table 3. Vapor pressure of aqueous solutions of NaCl (in kPa)

Temperature, °C	NaCl content

	5 %	10 %	15 %	20 %	25 %
0	0.59	0.57	0.55	0.51	0.47
20	2.26	2.18	2.09	1.97	1.81
40	7.13	6.88	6.58	6.20	5.72
60	19.26	18.58	17.78	16.76	15.53
80	45.75	44.16	42.49	39.97	37.09
100	97.89	94.43	90.44	85.52	79.67

Thermodynamic properties of aqueous solutions are given in [8].

In the system NaCl – H_2O (Fig. 1), only one hydrate, NaCl · 2 H_2O, exists. It crystallizes as monoclinic, thin, bevelled platelets that decompose to solid sodium chloride and saturated brine at 0.15 °C. Below this temperature it is the stable solid phase. However, it crystallizes so slowly that, on rapid cooling, the phase diagram follows the broken curve in Figure 1. The metastable eutectic point reached is ca. 1.3 K below the true cryohydric point at – 21.12 °C. Its property of depressing the freezing point of water enables sodium chloride to be used in freezing mixtures and as a de-icing salt. The saturated aqueous solution boils at 108.7 °C.

Sodium chloride has good solubility in aqueous ammonia (Table 4). At low temperatures, fine white needles of composition NaCl · 5 NH_3 separate out. Addition compounds with urea, glucose, and sucrose also exist. Solubilities in pure methanol and pure ethanol at 25 °C are 1.31 and 0.065 g NaCl/100 g solvent, respectively. Sodium chloride is also soluble in glycerine, and dissolves in high pressure steam, the solubility being 0.1 % at 17 MPa, and 0.6 % at 19.5 MPa. Chemically pure sodium chloride is not hygroscopic, but if magnesium salts are present as an impurity it becomes so. Sodium chloride solutions are corrosive to base metals, and therefore dilute solutions of NaCl are used in corrosion tests (e.g., DIN 50 907). Solutions of pure NaCl are more corrosive than those that also contain salts of Mg or Ca [10]. Iron and steel are attacked only fairly slowly by pure NaCl solutions, but much more quickly when the solutions contain oxygen. Corrosion by NaCl solutions is accelerated by the presence of local electrolytic cells [11]. Mixtures of organic acids with sodium chloride, (e.g., in perspiration) are highly corrosive [12].

Table 4. Solubility of sodium chloride in aqueous ammonia at 20 °C [9]

[NH_3], g/L	15.38	30.06	62.56	92.97
[NaCl], g/L	303.46	294.05	273.01	257.60
ϱ, g/cm^3	1.186	1.175	1.147	1.124

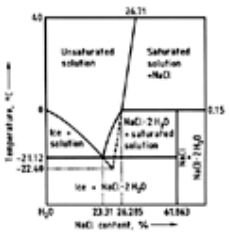

Figure 1. Phase diagram of H_2O – NaCl system (schematic, not to scale)

[Top of Page]

3. Formation and Occurrence of Salt Deposits [13]

Gerhard Kristen

When rocks are chemically weathered, salts are dissolved, carried into natural running waters, and collect in the oceans and inland lakes and seas. Where evaporation is predominant, the salts become concentrated until the solutions become supersaturated and crystallization occurs.

The proportion of sodium chloride in the salts of seawater is ca. 78 %. Although the oceans contain the largest quantity of dissolved salts, the majority of salt is bound in the form of solid deposits. In Germany alone, the quantity is estimated to be ca. 100 000 km^3.

The chemical-sedimentary genesis of salt deposits has been accepted ever since OCHSENIUS (1877) proposed his bar theory. Inspired by the contemporary example of the Kara-Bugas bay (Caspian Sea), he suggested that bodies of seawater became almost completely isolated from the open sea by bars. Then, due to the evaporation that took place in arid climatic regions, the seawater became concentrated. The sparingly soluble carbonates and sulfates were the first to

crystallize, followed by the more soluble chlorides.

In the case of the most important European salt deposit, the Zechstein Formation, communication with the North Sea has been proved by fossil finds, thus demonstrating the marine origin of the Zechstein salts, which extend over ca. 500 000 km^2 from England to Central Poland, and from Denmark to Thuringia/Hesse (Germany). The bar theory can satisfactorily account for even the 800 – 1000 m thickness of the Zechstein salt deposit, and is also valid, in modified form, for marine salt deposits. A depth of 100 m seawater produces a salt layer only 1.5 m thick on complete evaporation.

It is now believed that, for the European Zechstein salt deposit, the salt lakes were connected to the ocean by an extended saturation shelf lake where considerable amounts of $CaSO_4$ were deposited. The minor and major salt lakes were filled with salts during four largely identical deposition cycles, each having the sequence clay, calcium – magnesium carbonates and calcium sulfate, sodium and potassium chlorides, and ending with a recessive transition [sodium chloride, calcium sulfate (anhydrite), and clay] to the next cycle. The concentrations of the various soluble salts reached saturation successively as the seawater became gradually more concentrated, so forming the geographical sequence of salts. This explains why the composition does not correspond to that of normal seawater at any point in the salt profile, why all the types of salt are not deposited everywhere, why salts of differing solubilities are precipitated simultaneously at different saturation regions, and why the chlorides of potassium and magnesium are deposited only at the centers of the lakes.

The intermediate layers of clay sediments in each cycle are the result of influxes of fresh seawater. Eolian (airborne) transportation of clay minerals into the salt basin can also occur.

The depths of the brine in the deposition basins have been calculated from the bromine contents for each phase of deposition. The results obtained were: first cycle (Werra region) 330 m, second cycle in the Stassfurt region 860 m, and third cycle in the Leine region 95 m. In some cases the salt basin dried out during the third cycle. Sediments of younger formations covered and protected the Zechstein salts from dissolution.

In the North West German depression zone, a layer of overlying rock up to 5 km thick was formed. The pressure due to this denser rock, together with the interior heat of the Earth, caused metamorphosis of the salt, which became plastic and migrated. It accumulated in zones of deformation and ascended, forming salt domes (diapirs) which sometimes extended as far as the earth's surface. These were slowly dissolved by groundwater, leading to the formation of very flat secondary intermediate layers next to the overlying rock.

The upward movement of the salt during the formation of salt domes deformed the salt layers into folds, with predominantly steeply inclined strata. Under the thinner overlying rocks at the edge regions, the flat deposits remained undisturbed.

Marine rock salt deposits occur in most geological formations (see also ⟶ Potassium Compounds). In the following list, countries printed in italics have salt deposits that are economically exploited. Connected regions are separated by commas, and disconnected regions in a country by a semicolon. Countries and regions are separated by periods.

Cambrian. Canada: NW territories. Australia. *Siberia. Iran. Pakistan.*

Ordovician. Canada and United States: Williston Basin with parts of Manitoba and Saskatchewan (Canada) and parts of Dakota, Montana, and Wyoming (United States).

Silurian. United States and Canada: Salina Basin with Maryland, *Michigan, New York, Ohio*, Pennsylvania, w Virginia (United States), and Ontario (Canada).

Devonian. United States and Canada: Williston Basin with North Dakota, Montana (United States), and Alberta, Manitoba, Saskatchewan (Canada). CIS: *Siberia; Byelorussia,* Baltic, *Dnieper-Donets Basin.* Australia.

Carboniferous. Canada and United States: Williston Basin with New Brunswick and Nova Scotia (Canada), Virginia (United States), Paradox Basin with Colorado and Utah (United States). Brazil: Amazon Basin.

Permian. American Permian Basin: Colorado, *Kansas*, Oklahoma, *New Mexico*, and Texas; American Supai Basin: Arizona, New Mexico; American Williston Basin. Mexico: Chihuahua. Peru: E Cordillera. CIS: *Kama Basin*; *Donets Basin*, Volga Basin, Caspian Basin; Baltic. Bulgaria. Greece.

European Zechstein. England, *the Netherlands, Denmark, Germany, Poland.*

New Red Sandstone. Germany, *the Netherlands, Portugal.*

Muschelkalk. *Germany, France, Switzerland.*

Keuper. Germany, *England, France, Spain,* Algeria, Libya.

Triassic Alpine. *Germany, Austria.*

Triassic, unstructured. Mexico: Tehuantepec. Bulgaria. Greece. Morocco. Tunisia.

Jurassic. United States: *Idaho*, Utah; Gulf Coast: *Texas, Louisiana, Mississippi.* Cuba. *Chile. Yemen.* Kuwait. Morocco. Tadzhikistan. Germany.

Cretaceous. United States: Florida. Bolivian Cordillera, Brazil: Sergipe-Alagoas Basin, *Argentina, Colombia*. Angola, Congo, Gabon, Morocco, Algeria. Tadzhikistan.

Eocene. United States: Green River Basin Wyoming. Belize. Spain. Yugoslavia.

Oligocene. France, Germany. *Spain. Anatolia*, Caucasus Mountains, Iran, Iraq.

Miocene. *Czechoslovakia, Polish, Russian,* and *Romanian Carpathian foreland, Yugoslavia, Anatolia,* Cyprus, Syria, Iraq, Iran, Golf of Suez, *Algeria, Sicily, Tuscany, Spain.* Sudan. *Dominican Republic.*

Pliocene. United States: Nevada, Utah. Jordan, Israel.

[Top of Page]

4. Production

Wilhelm Wegener, Peter Ambatiello, Helmut Geyer, Bernard Epron, Christian Bonal

4.1. Mining of Rock Salt

Rock salt has been mined in Europe for 3000 years. A salt deposit near ground level in the Eastern Alps was developed by tunnelling and worked by excavation around 1000 B.C.

Salt-bearing regions are revealed by the presence of surface springs of saline water, and these were the areas where the possibility of mining was always investigated by sinking shafts. The main precondition for success is the presence of dry overlying rock, and it was this circumstance that enabled the first German salt mine to be opened in 1825 near Schwäbisch Hall. In other parts of Europe, salt was mined long before this, e.g., in Poland before 1000 A.D., and in England since the 1600s.

Total world production of solid salt by conventional mining and brine by solution mining is ca. 190×10^6 t. Annual production of salt in the Federal Republic of Germany in 1989 was ca. 13×10^6 t.

The purity of the salt mined from rock salt deposits is between 90 and 99 % NaCl, and sometimes higher. The other minerals present in the rock salt are mainly clay and anhydrite, often intimately intergrown.

Rock salt and potassium chloride have the same marine origin, often occurring together in a single deposit, and have essentially the same mechanical strength properties. Hence, the development and mining of rock salt deposits is similar to that of potash mineral deposits.

Mine Shafts. Access from the surface to the salt deposit is usually by vertical shafts 5 m in diameter. There must be at least two shafts, so that in case of damage to one of them, the underground workforce has an emergency route to the surface.

To sink the shaft, the rock is broken up by drilling and explosives. The rock debris is loaded into skips by grabs, and taken to the surface. In most cases, the upper layers of the overlying rock are water-bearing, and sometimes consist of unstable loose rock and moving sand capable of exerting pressure. Therefore, before sinking the shaft, these strata must be solidified and stabilized. This can be achieved by freezing via holes drilled around the area where the shaft is to be sunk. The freezing equipment must be kept in place for the duration of the sinking of the shaft.

Another method of preventing the movement of water in the overlying strata and immobilizing loose rocks is the injection of cement, synthetic resins, or other hardening materials via boreholes. This technique is also suitable for sealing the mine shaft against the ingress of water and for stabilizing the rock at the bottom of the shaft.

As rock salt is a water-soluble mineral, it is especially important to ensure that the lining of the shaft is watertight when it is sunk through water-bearing strata. In older shafts, the sealing system consists of ring-shaped sections of cast iron or steel bolted together. Newer shafts have a watertight lining consisting of a welded sheet-steel cylinder, an inner hollow cylinder of reinforced concrete, and an outer asphalt-filled ring.

Material is hoisted up the shaft in 20 t skips at rates of up to 1000 t/h. These travel between wooden or steel guide rails, or in a few cases between tensioned guide cables.

If the salt deposit appears as an outcrop, or is at low depth (< 100 m), the salt can be extracted by drift mining (tunnelling) or by means of an inclined shaft. In these cases, the salt can be transported to the surface with conveyor belts.

Mining Methods. As with any mining operation, the method of extracting salt depends on the thickness and spatial formation of the deposit.

A characteristic feature of salt mining is that the haulage and ventilation tunnels servicing the excavation chambers are driven through rock salt, whereas in other types of mine they are often coated in the adjacent rock. In salt mining this is generally not necessary, as the deposits are usually thicker. Furthermore, drilling and tunnelling into the adjacent rock should be avoided as far as possible to prevent ingress of water or gas from these strata into the mine.

Mining in Level Deposits. In most salt mines, a systematic extraction procedure is possible due to the uniformity of the deposit over large areas.

In salt deposits with horizontal or gently inclined seams, the standard method is room and pillar mining. In this method, the extraction process produces large chambers with rectangular cross sections of ca. 50 – 400 m^2 and lengths of up to 500 m. The parallel extraction chambers are separated from each other by rock salt pillars left behind during mining. These pillars must be of such dimensions that they can carry the weight of the overlying rock. Generally, additional support is not required, and the ground above is protected against subsidence.

The pressure exerted by the overlying rock increases with depth, so that the salt pillars must be wider. For example, in a salt mine at a depth of 200 m, the pillars are 15 m wide and the chambers 15 m; at a depth of 500 m, the pillars are 30 m wide and the chambers 20 m; and at a depth of ca. 700 m, the pillars are 50 m wide, and chambers 22 m. The salt pillars result in a ca. 40 – 70 % loss of product, depending on their dimensions.

The salt extraction process starts with the construction of a central haulage tunnel. At right angles to this, tunnels leading to the working faces are driven. The tunnels are developed upward or downward to form chambers if the deposit is so thick that it extends above or below them. The salt pillars produced by this mining technique are usually not cut through.

If the deposits have a workable thickness of < 10 m, the pillar and chamber method with long pillars is often changed to one with short, square pillars (room and pillar system). Here, the transverse tunnels usually have the same height and width as the long ones.

Fresh air is provided by an inlet ventilation road above the haulage tunnel. The used air is removed through return air galleries and by the main ventilating fan, which is usually underground, and blown through the exhaust shaft to the surface.

The mining of the salt during the construction of the tunnels and in the extraction chambers is by drilling and the use of explosives. Other possible methods, such as cutting the salt by part-face heading or full-section cutting machines, are used where there are special restrictions, e.g., where the winning process must be vibration free.

Various drilling and blasting processes are used. In the undercutting method, the working face is undercut to a distance of 4 – 5 m. The cut, 15 – 20 cm wide, is produced by a machine resembling a chain saw. An electro-hydraulic mobile drilling machine drills 35 – 38 mm diameter holes for the explosive. In another process, a wide hole up to 7 m in length and 0.45 m in diameter is drilled into the middle of the working face to provide an initial extension space for the salt at the moment of explosion, a function that is performed in the undercutting process by the undercut.

The extraction chambers are created by widening the access tunnels. If these are driven through the lower part of the deposit, this is performed by the use of explosives in long inclined boreholes directed upwards. If the tunnel runs through the upper part of the deposit, the salt is extracted from the floor, i.e., by drilling long holes downwards into the salt.

Figure 2 shows a vertical section through a South German mine where the rock salt deposit is 20 m thick and 200 m deep. The haulage and working roads are located in the lower part of the deposit, and the main ventilation road in the upper part. In the main extraction phase, the salt that has been broken up by explosives is left lying in the space created, and the final position of the roof of the chamber is created by blasting, working from the top of the heap of loose salt. After scaling the roof can then be collected without the danger of roof falls.

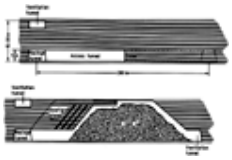

Figure 2. Vertical section through a South German salt mine

The most commonly used explosive is ANFO (ammonium nitrate fuel oil mixture) in the form of loose prills. It is charged pneumatically into the boreholes.

Due to the strength of the salt rock, it is possible in most mines to keep the mined space open without supports. At great depths, where the pressure is high, or where breakup of the deposit leads to danger to the miners from roof falls of salt, rock bolts are used.

Mining Steeply Inclined Deposits. Steep to vertical rock salt deposits are generally mined by the stepped face method, as used, for example, in North German mines at depths of 400 – 850 m, with working levels at vertical intervals of 100 – 250 m and intermediate sublevels at intervals of 15 – 20 m. Between the sublevels, there are working faces which are drilled and blasted, beginning with the lowest, using long boreholes. In one mine, the openings produced are 20 m wide, 42 m high, and 100 m long, and, in another, 20 m wide, 100 m high, and 60 m long.

To maintain stability of the rock, salt masses must also be left behind using the grid wall system, the vertical pillars being 10 – 30 m thick, and the horizontal pillars between the sublevels, 8 m thick.

Loading and Haulage. To load the mined material after blasting and to transport it to the first crusher, diesel-powered front loaders with bucket capacities of up to 18 t are used (LHD: Load, Haul, Dump System). For economic reasons, there should be a straight road between loading and discharging points not longer than ca. 300 m. For longer distances, the interrupted system is more economic, with special loading equipment and dumper trucks of up to 50 t capacity.

After crushing the salt to < 300 mm using equipment near to the working face, it is transported by band conveyors to the hoisting shaft.

In mines where the salt has a purity of > 99 % NaCl, the salt can be marketed directly after grain-size classification.

Where the natural purity is lower (e.g., 94 % NaCl in South German mines), impact mills are used in the first processing stage. As rock salt is more brittle than the clay and anhydrite inclusions, it is more readily size reduced. A screening operation follows, separating the two size fractions, which increases the NaCl content by 2 – 3 % compared to the crude salt. Further purification of the salt to > 99 % NaCl for industrial use is carried out in heavy media hydrocyclones.

Utilization of the Chambers. Salt mining is mainly carried out without back-filling, i.e., the chambers produced during mining need not be backfilled with other material to support the roof of the mine and increase the load-bearing capacity of the salt pillars.

Where processing of the rock salts produces residues that must be disposed of, these are usually dumped in the empty chambers. Hence salt mines do not produce heaps of tailings at ground level.

In some cases, empty chambers are utilized for disposing of waste materials, e.g., wastes from combustion processes and the chemical industry.

4.2. Production of Crude Brine by Mining Methods

The principle of all mining methods for the production of crude brine (NaCl solution) is the use of “wet mining”.

The rock salt, which may be present in a pure state (e.g., geologically undisturbed crystalline salt deposits) or in a mixed mineral in finely divided form (e.g., in alpine salt deposits), is dissolved by the action of fresh water on the rock formation and converted to a concentrated salt solution (brine with an NaCl content of 312 g/L or 27 %).

In all processes, the brine is produced in underground excavations made by conventional mining or solution mining [14-20]. The extraction processes result in formation of chambers [21]. Between the chambers in which the brine is produced and extracted, pillars of considerable size are left behind to maintain stability of the rock [22], [23]. Figure 3 gives a survey of brine mining methods.

Figure 3. Classification of brine mining methods

Mining of brine is no longer widely used. It is employed in alpine salt deposits and in a few rock salt mines. The most widely used and most economic process today is solution mining (see Section Controlled Solution Mining) [14], [15]. In a few cases, subsurface solution mining is also carried out at depths of 100 – 140 m [24], [25].

Water flushing (intermittent or continuous dissolution in chambers) is the extraction process usually used in the alpine salt deposits in West Germany and Austria [14], [15], [18-20]. Tunnels, shafts, and chambers are usually first produced by mining techniques. A single worked stratum is divided into separate districts of a given size in which several extraction chambers are set up. A block of definite dimensions (length, breadth, and height) is allocated to each chamber based on a predetermined degree of extraction. The block encompasses a space of irregular configuration produced by the extraction process and the pillars left between the extracted spaces, which are also of irregular shape. An initial space is first created, which is a funnel-shaped (Germany), or circular chamber (Austria). A shaft or tunnel is first driven from the main tunnel or a transverse tunnel to the center of the intended block at the bottom of the worked strata. A shaft is then sunk by mining techniques from which the initial chamber of a specific size is created at the lowest point by drilling and explosive methods (Austria). In Berchtesgaden, Germany, the initial space (Fig. 4) required for excavation is created by drilling and water-injection processes. An important feature of this process [20], [22] is that the initial space is funnel-shaped, and is generated by controlled dissolution, starting from a wide borehole, so that the insoluble components of the deposit can be raised by airlifting equipment and removed by a hydraulic pump as a solid – brine mixture. The production of the initial space is usually divided into the following stages (Fig. 4): sinking a borehole for examination of the deposit in the region to be excavated, vertical development by sinking a wide borehole, further widening of the borehole by controlled dissolution, forming a funnel-shaped space at a depth extending from 100 to 125 m. During this phase, the space normally attains a volume of 3000 – 5000 m^3, depending on the quality of the deposit, with a height of 4 – 6 m and a diameter of 30 – 32 m. The creation of the initial space completes the preparation for the extraction process. The space is next extended further, conditions now having been established for continuous, high yield, cost effective brine production. The initial space (chamber) is slowly filled with fresh water, which dissolves salt from the side walls and roof. The insoluble components, such as clay, anhydrite, and polyhalite sink to the floor of the chamber. Fresh water is continuously added, raising the liquid level, so that water remains continuously in contact with the roof of the chamber until the water in the chamber is completely saturated. The dwell time depends on the salt content of the rock, the size of the chamber, the ease of dissolution of the salt, and the surface area of the chamber roof, and is usually between 10 and 30 days. After complete saturation has been attained (312 g/L), the brine is removed either by pumping out of the excavated space (if working below the tunnel level), or by draining off to the level below. The space is then refilled with fresh water. This procedure (filling, saturation, emptying), is known as intermittent water flushing, and is repeated until the chamber roof area has reached ca. 3000 – 3200 m^2. The chamber is then ready for continuous brine production by the continuous water flushing process. The water flows at a rate appropriate for the parameters of the deposit, with simultaneous removal of concentrated brine by draining or pumping. The brine production is now almost invariably carried out below the main haulage level (Germany and Austria), i.e., working upwards from the bottom. Each day, 1 – 1.5 cm rock salt is dissolved from the surface of the roof of the chamber. The chambers take on a funnel-shaped to cylindrical configuration (Fig. 4). Chamber heights in Austria [19], [24], [25] are 20 – 30 m (normal working) or 50 – 80 m (deep working), depending on the

diameter. In Berchtesgaden, Germany, the exploitable height of the chamber is 100 m. The initial space is created in the region between 100 m and 125 m below the level of the tunnel.

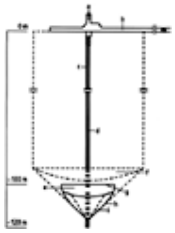

Figure 4. Production of the initial extraction chamber by drilling and water flushing, and brine production operating below the main tunnel (Salzbergwerk Berchtesgaden, Germany)

a) Drilling chamber; b) Main tunnel; c) Borehole (76 mm); d) Wide borehole (670 mm); e) Initial extraction space; f) Developed extraction space; g) Chamber roof; h) Angle of boundary; i) Angle of repose of solid residues

To calculate the efficiency of extraction of the deposit, the salt content of the rock and the size of the rock mass assigned to the particular operation point (cavern), including the horizontal safety pillars, must be taken into account. In Berchtesgaden, Germany, when the size of the horizontal and vertical pillars, the average salt content of alpine deposits (50 %), and the usable height of the chamber (100 m) are taken into account, the total quantity of brine available per chamber is 1.1×10^6 m^3. This corresponds to an extraction efficiency of ca. 12 % [20], [22]. In Austrian salt extraction operations, the following total available quantities of brine have been quoted, depending on the type of extraction chamber [24] and [25]: 200 000 – 400 000 m^3 for normal operations, and 700 000 – 1 000 000 m^3 for deep operations.

The *work face dissolution process*, in which dissolution is carried out in chamber-like tunnel sections, is similar in principle to the intermittent water flushing process, and is also not widely used. In some mining operations in Germany and Poland [15], [17] it is still operated alongside other brine production processes (e.g., solution mining using boreholes drilled from the surface).

To obtain the brine, transverse tunnels are driven one above the other on both sides of the main tunnel (Fig. 5). The transverse tunnels are linked together at their ends by a borehole. Passing from one tunnel to the next, the borehole is widened to a shaft of 1 m diameter and provided with an outlet to the lower tunnels. About 10 m below each tunnel, a tunnel for brine production is driven from the widened borehole, a work face is produced by cutting upwards at the end, and spraying equipment is installed at that point. This creates a cut which moves backwards at a fixed height until it meets the widened borehole. After the initial work has been finished, the working is filled with water or unsaturated brine. The water or brine remains in the working until saturation is reached. The concentrated brine is then removed via the lower work face. In contrast to the water flushing process described above, the salt is only removed from the work faces until the planned chamber width is produced.

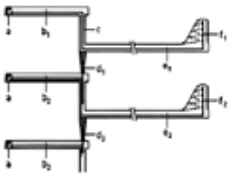

Figure 5. Production of brine work-face dissolution

a) Access tunnels; b_1) – b_3) Transverse tunnels for work faces 1 – 3; c) Borehole widened to make a shaft; d) Brine removal; e) Transverse tunnels; f_1), f_2) Spraying equipment

The production of brine by means of water jets is only used to a small extent, being employed in a few cases to solve special problems, e.g., to speed up excavation during intermittent water injection at the development stage, or to regulate the configuration of the chamber during the water flushing process in large chambers.

4.3. Controlled Solution Mining

Controlled solution mining is the modern, economical method of extracting rock salt from underground deposits, and contrasts with conventional mining (see Section Mining of Rock Salt). However, for this method certain preconditions must be fulfilled, especially with regard to the geology of the deposit (type, formation, and depth of the salt). Economic operation of the solution mining process requires that the deposit—whether a diapir (salt dome) or a flat stratum of salt—should fulfill the following geological conditions:

1. Adequate thickness and extent of the salt deposit
2. High purity of the rock salt without inclusions of potassium chloride
3. Largely undisturbed stratification at medium depth

In addition, an adequate supply of water must be available.

4.3.1. *Drilling and Construction of the Borehole*

The drilling and completion of the trial boring and the design and installation of the brine wells (also known as the cavern wells) closely resemble those of petroleum and natural gas wells (see ⟶ Oil and Gas). The method of operation and the geological conditions determine the size, type, and position of the casing that protects the walls of the borehole against the effects of permeable strata, and prevents solid material from the rock from falling into the borehole. A standard pattern for the casing diameters in a cavern well (see Fig. 6) is as follows:

1. 0 – 15 m: 24.5 inch pipe
2. 0 – 300 m: 16 inch pipe
3. 0 – 20 m (max.) beyond the projected roof of the cavern in the deposit: 11.75 inch pipe

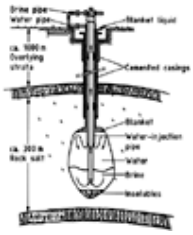

Figure 6. Method of brine production

The casings described in 2) and 3) above are cemented up to ground level.

Geophysical borehole measurements and tests on the cores obtained during sinking of the borehole not only give information about mineralogical composition, but also yield physical data about the overlying rock and the salt deposit. The most important characteristics of the overlying rock are its thickness and strength. Also, to enable the dimensions of the cavern and the proposed extraction process to be decided upon, the strength and thickness of the salt deposit and the solubilities of its component parts must be known.

The outer casing must be installed and cemented with special care, and each individual joint as well as the whole casing must be tested for tightness. After this, the well is completed by inserting two concentric pipes (of diameter between 8.625 and 5.5 inches), the final preparative work being the assembly of the well head and making all the pipe connections (Fig. 6).

4.3.2. *The Process of Solution Mining*

As shown in Figure 6, fresh water is pumped into the borehole through the annular space between the two inner pipes. The water dissolves salt from the walls of the cavern. The solution sinks as its salt content increases, and the saturated brine formed then rises to the surface in the central pipe due to the applied pressure.

Dissolution of salt from the roof is prevented by injecting a protective liquid or gas blanket through the annular space between the outer casing and the pipe carrying the fresh water. This blanket forms a layer above the water due to its lower density, so protecting the roof. Insoluble materials collect at the bottom of the cavern.

The desired cavern shape is produced by stepwise adjustments to the depth of the water-injection pipe and the level of the blanket [26-29].

The extraction process is controlled by continuously measuring the salt content of the brine, the blanket pressure, and the quantity of brine extracted. The cavern space produced is calculated from these data by computer, and the spatial extent of the cavern is continuously determined by echometric surveying. The actual cavern produced can then be continuously compared with the planned cavern.

4.3.3. *Planning of the Extraction Process*

The safety of caverns can be ensured if pillars of adequate width are left between neighboring caverns and if there is a thick enough mass of solid salt at the top of each cavern. The width of the pillar and the thickness of the roof depend on the planned cavern dimensions, the cavern shape, and, in particular, on the mechanical properties of the salt deposit. The latter have been investigated in detail [29-35].

Furthermore, correlations and formulas have been derived that enable the forces produced in the rock to be calculated from the size and shape of the cavern, the interior pressure produced by the solution mining process or subsequent use for storage purposes, the depth and thickness of the overlying rock, and the mechanical properties of the surrounding rock formations. The decrease in the volume of large caverns (convergence) caused by elastoplastic properties of the rock salt are determined from formulas, confirmed by data from practical experience [30-39].

Irrespective of the widths of the pillars calculated with these equations, the wells in a brine field should be arranged in a hexagonal pattern (Fig. 7), which gives optimum utilization of the available area.

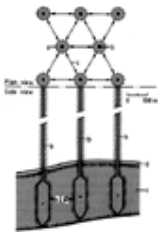

Figure 7. Hexagonal arrangement of caverns in a brine field with a level salt deposit

a) Cavern (diameter*D* = 60 – 100 m); b) Borehole; c) Pillar of stable rock salt (thickness of pillar: 2 *D*); d) Upper layer of anhydrite

The distance between the brine wells is the sum of pillar width, the maximum cavern diameter, and a safety factor to allow for deviation of the borehole from the vertical. If a borehole must be relocated due to the situation at ground level, the hexagonal arrangement below ground can be maintained by adjusting the drilling direction (Fig. 8).

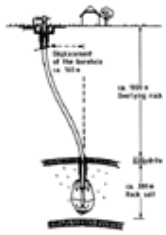

Figure 8. Directional drilling in the production of a cavern

4.3.4. *Other Systems*

Individual boreholes for extracting brine can generally only be operated economically in thick salt deposits or salt domes. However, thin salt deposits can also be exploited economically in some cases. Here, two or three boreholes are linked together by injecting water to dissolve the salt and to fracture it by hydraulic action [28], [40], [41]. Fresh water enters via the first borehole, and the brine is collected from the second or last borehole. Here also, dissolution of the roof is prevented by a blanketing medium. The floor is dissolved first, then the roof. With thin seams of salt, the strength and impermeability of the overlying strata are very important.

Control of this method of solution mining is more difficult than with individual brine wells. The stability of the cavern spaces cannot be guaranteed, as there is a risk of uncontrolled dissolution processes. Thus, this method of solution mining is only used in countries with few salt deposits where there are sparsely populated areas above thin salt strata.

4.3.5. *Equipment*

The equipment requirements normally consist of

1. Equipment for supplying water, sometimes including a pipeline
2. A pumping station with a distribution system delivering water to the individual solution mining operations
3. A pipeline for transporting the brine to the consumers
4. A central system for supervising and controlling the entire processing equipment

The water is injected by high-pressure centrifugal pumps at a pressure determined by the difference in the densities of the two media (water and brine) and the friction losses in the pipes. The salt at the working face dissolves and passes through the central pipe as saturated brine to the well head and into the piping system. The low residual brine pressure at this point is sufficient to deliver it to the pumping station, where delivery pumps force the brine into the pipeline that takes it to the consumer industries.

4.3.6. *Storage Caverns*

The caverns remaining after extraction of the sodium chloride can be used for storage of mineral oil, mineral oil products, and gases such as natural gas and ethylene.

Emergency reserves of mineral oil or mineral oil products can be injected through a pipeline into the brine-filled caverns by a high-pressure pump. In place of the two concentric pipes used in solution mining, a brine removal pipe is suspended in the cavern with its open end just above the sump region. The mineral oil is forced into the annular space between the outer pipe and the brine removal pipe and displaces the brine, which rises through the brine removal pipe to the surface and is fed into the brine network, and thence to the production plant. To discharge the oil, the process is reversed.

Caverns in rock salt are also suitable for gas storage. For example, natural gas is stored to compensate for seasonal and daily variations in demand. The conversion of a cavern to a gas-storage system is more complex than conversion to liquid storage. To protect the outer piping against pressure and temperature variations during operation, a protective pipe is fitted. The annular space between this pipe and the outer pipe is pressurized so that the welded pipes are under a constant load. For the first filling operation with gas, a brine extraction pipe is lowered to the bottom of the cavern. The gas in the annular space between the protective pipe and the vertical central pipe is compressed to a high pressure, and displaces the brine from the cavern until the latter is completely empty. The gas storage system operates by the principle of compression-decompression. The gas in the cavern required to maintain the minimum residual pressure required for stability of the cavern is called the cushion gas, but when it is filled up to its maximum operating pressure, it is known as the working gas.

Descriptions of cavern storage systems can be found in [42-45]. A worldwide information exchange service is provided by the Solution Mining Research Institute (SMRI), Woodstock, Illinois.

4.4. Production of Sea Salt

Introduction. In his *Historia Naturalis*, PLINY THE ELDER already referred to the technique of producing salt by the natural evaporation of sea water. Even today, this method of exploiting the oceans, seas, salt lakes, and chotts is still widespread. It is practiced in nearly all regions of the world where climate and topography allow.

While some methods have changed very little, such as those of the “Paludiers” working the salt marshes along the Atlantic coast of Portugal and France, other modern methods are now applied, combining productivity with strict environmental protection. This applies to the vast majority of salt fields in the Mediterranean basin, on the west coast of Australia, and Guerrero Negro, the world's largest salt field in Mexico at the southernmost tip of Baja California.

4.4.1. *The Raw Material: Sea Water*

The water of the seas and oceans contains all the known elements, most of them present in small amounts [46]. Sodium chloride is the most important compound in terms of concentration, averaging 28 g/L.

Sea water also contains significant amounts of magnesium, sulfur, and calcium, the other important sea-water-based process being the production of magnesium compounds.

The salinity (grams of salt per kilogram of sea water) of ocean and sea water varies with location and depth [47]. This dispersion, due either to dilution by precipitation or drainage pattern, or to local overconcentration caused by strong evaporation, is more evident in inland seas.

The average salinity is 3.5 %, corresponding to a relative density of 1.026. It reaches 4.1 % in the Red Sea and is only 3 % in the Baltic Sea.

The salt mixture in sea water has the following typical composition (in wt %) [47]:

NaCl	77
$MgCl_2$	10
$MgSO_4$	6
$CaSO_4$	3.9
KCl	2

4.4.2. *The Main Factors Governing Production of Sea Salt*

To produce sea salt economically, the amount of water evaporating must exceed the amount of precipitation (rain or dew) for a long and continuous period of at least three months during the year.

Table 5 lists the main types of climate compatible with the production of sea salt. In Taiwan, a borderline case, sea salt production calls for a maximum of ingenuity. In Western Australia, by contrast, the climate is ideal, notwithstanding the risk of cyclones, and sea salt can be produced all year round.

Table 5. Typical climatology of solar saltfields [47]

	South of France	Australia	Indonesia	Taiwan
Annual evaporation *, mm	1700	3600	1850	1660
Annual rainfall, mm	550	300	1300	1400

* Fresh water.

In practice, evaporation at a given production site is measured over fresh water with an evaporometer. Since the total concentration of dissolved salts is known, the evaporation is determined approximately from an empirical curve, an example of which is shown in Figure 9 [48].

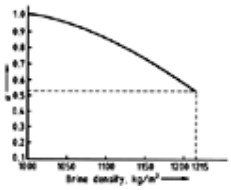

Figure 9. Evaporation coefficient α as a function of density

1 cm of evaporation over fresh water corresponds to α cm for the body of water in question; α normally ranges from 0.95 for sea water to 0.55 for saturated brines depositing their salt.

In addition to a favorable climate, a particular topography is required, and the ground must be as impermeable as possible to minimize brine seepage.

The pond areas containing the brines or saline solutions must also be as level as possible, ideally below sea level, and stepped in elevation. This reduces the number of intermediate pumping stations required and helps to optimize the design of the ponds, dikes, channels, and culverts making up the salt field.

These conditions are often satisfied in the deltas of large rivers. The Rhône delta in France's Camargue is one example, and is the location of two large salt fields: Aigues-Mortes and Salin de Giraud.

Concentration of Sea Brines. After being pumped from the sea, the sea water passes through the salt field from pond to pond. As it passes through the ponds, the NaCl concentration in the sea water rises from 28 g/L to roughly 260 g/L, corresponding to an increase in relative density from 1.026 to 1.215. At this point, the brine begins to deposit its salt. Most

of the calcium carbonate ($CaCO_3$) and calcium sulfate ($CaSO_4 \cdot 2\ H_2O$) has already crystallized before this point, as shown in Figure 10, while the magnesium salts continue to become concentrated without crystallizing.

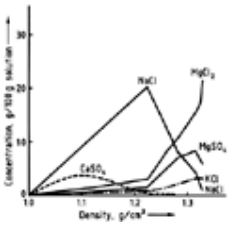

Figure 10. Brine composition as a function of density [49]

Values for $CaSO_4$ concentration are multiplied by a factor of ten.

Crystallization of Sodium Chloride. When the brines approach the NaCl saturation point, they are fed through feed channels into the crystallizers, which represent the production units. The floors of the crystallizers must first be carefully prepared by compacting and levelling to ensure easy separation of the salt from the ground and a uniform layer of brine. The surface area of these crystallizers generally represents ca. 10 % of the total area of the salt field.

From this point on, the salt worker's job is to monitor the brine strengths and purge the bitterns from the ponds when the concentration of secondary salts, particularly magnesium salts, is liable to lower the quality of the salt. The ponds are then filled with fresh or pure brine with a low magnesium salt content.

The salt is produced in the form of a layer of tangled crystals whose density and grain size depend on the characteristics of the bittern and, especially, on the turbidity or cloudiness. If the brine is turbid, the suspended fine particles act as seed crystals. If, however, the brine is clear and well settled, the number of seed crystals is greatly reduced, and crystal growth occurs in preference to germination.

The thickness of the salt layer varies from one site to another, and also, at a given site, from one year to another, depending on weather conditions. In the South of France where the production period runs from May to September, the thickness at the end of the season averages 10 cm. At other latitudes, layers may be over 20 cm thick.

4.4.3. *Production Stages in a Modern Salt Field*

Figure 11 shows the main components of the infrastructure of a modern salt field, from sea water intake to the production facilities for processing and packaging the salt.

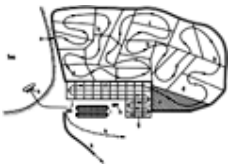

Figure 11. Modern salt field [47]

a) Sea water intake; b) Pumping station; c) Dikes; d) Condensers; e) Brine storage tanks; f) Crystallizers; g) Bitterns area; h) Washing plant; i) Salt storage areas; j) Packaging plant; k) Dispatch

The production process features the following stages:

1. Seawater pumping: the salt field is supplied with sea water by a pumping station. In some parts of the world, such as Australia, the salt fields are gravity fed with sea water in accordance with the tides. Several million cubic meters of sea water are required annually to produce several hundred thousand tonnes of sodium chloride (the Aigues-Mortes salt field in France pumps $50 \times 10^6\ m^3$ of sea water for an average annual output of 400 000 t).
2. Concentration of sea water in the ponds: in this stage, the water flows through shallow ponds and is allowed to concentrate until approximately 90 % of the initial water content has evaporated. The brines are moved pond to pond by gravity or by pumping as dictated by the local topography.
3. Deposition of NaCl in the crystallizers: for a salt field with a single production season during the year, the crystallizers are first filled with brine kept from the previous year, which has been protected from rain dilution in deep storage ponds. A few weeks later, the old brine is replaced by the new concentrated brine leaving the ponds.
   By the time the relative density of the brine reaches 1.215, when NaCl crystals start to form, the brine has already precipitated 80 % of the calcium sulfate in the form of gypsum. The bitterns are then removed before the specific density reaches 1.260 to avoid codeposition of magnesium salts. Where bitterns are present in sufficiently large quantities, and if the climate allows, the bitterns can be concentrated on a bittern salt field to extract the magnesium bromide, magnesium sulfate, and even magnesium chloride (⟶ Magnesium Compounds).
4. Salt harvesting: in batch-operated salt fields, salt harvesting begins at the end of the salt production period. For continuous salt fields, harvesting is performed at regular intervals to ensure that the level of the salt does not exceed the operating height of the harvesting machines.
   The harvester is generally carried by a tractor of the type often used for civil engineering projects. The salt is first scooped off the floor of the crystallizer by a large blade up to 3.6 m wide pushed along by the tractor. It is then lifted by slotted or bucket conveyors. The salt is then evacuated and poured into trailers by a conveyor belt.
5. Transport, washing, and storage: the harvested salt is loaded onto trucks or tractor-drawn trailers for transport to a washing plant where most of the impurities are removed.
   Washing removes all solid impurities from the salt (substances insoluble in water) along with the accompanying bitterns (solutions of magnesium or potassium compounds). The salt is kept in suspension in saturated brine. The salt – brine mixture is then dewatered and centrifuged. The salt travels up an inclined gantry conveyor and is poured on top of one of the huge salt stockpiles that are a familiar feature of salt production sites.
6. Processing and packaging: The salt reclaimed from the salt stockpile may be washed a second time, which

increases the sodium chloride content to 99 %. The salt is then generally dried, crushed, and screened, and anticaking agents are added. Other substances added at this stage include iodide or fluoride in salt for human consumption. The salt is then packaged in various containers (e.g., boxes, polyethylene bags), as required by the consumer.

[Top of Page]

5. Production of Pure Salt by Evaporation of Brine

Georg Steinhauser

Salt can be produced by evaporation of saturated brine (for brine production, see Sections Production of Crude Brine by Mining Methods and Controlled Solution Mining). Brine is generally obtained by dissolving crude salt that contains many impurities. The production of high-purity salt requires chemical purification of the brine. The high purity is the major difference to rock salt (see Table 6). Purification is also necessary to avoid scale formation in the heat exchangers and the evaporation equipment. Additionally, the use of pure salt in industrial processes causes much less waste than the use of rock salt. On the other hand, the purification of salt itself causes wastes, because of the foreign ions in the crude salt.

Table 6. Percentage compositions of rock salt from from various sources

	Bad Ischl, Austria [66]	Drang, Mandi District, India [67]	Fairport, Ohio, USA [68]	Huitrin, Argentina [69]	Khanguet el Melah, Algeria [69]	Lyons, Kansas, USA [68]	Pugwash, Nova Scotia, Canada [70]	Stassfurt, Germany [71]
NaCl	97.00	69.50	97.26	99.27	97.80	95.94	94.5	92.4
KCl							0.02	0.05
K_2SO_4	0.59							
$MgCl_2$		0.08	0.04	0.04	0.004	0.10		0.03
$MgSO_4$	*		0.02		0.070			0.05
$CaCl_2$			0.02	0.09		0.01		
$CaSO_4$	1.60	0.54	2.2	0.44	0.680	2.06	3.0	7.4
Na_2SO_4	*		0.01					
$NaNO_3$					0.003			
Insolubles		28.70	0.3		0.830	1.68	2.2	0.07
Water	0.16	1.18**	0.15	0.14	0.612	0.18	0.2	

* The combined amount of sodium sulfate and magnesium sulfate is 0.65 %.
** Includes water and other salts.

A summary of possible processes for the production of salt is shown in Figure 12.

Figure 12. Processes for the production of salt (after [72])

5.1. Purification of Crude Brine

Brine purification is an important step in the production of sodium chloride for industrial purposes and nutrition. Magnesium and calcium ions must be removed from the saturated sodium chloride brine to avoid scale formation and also impurities such as hygroscopic magnesium chloride in solid sodium chloride.

On the other hand, these ions must be removed for several technical reasons, e.g., for membrane chlorine-alkali electrolysis. In the Solvay process, brine purification is essential as well (see ⟶ Sodium Carbonates – Technology). During carbonization they would form scales such as basic and neutral carbonates or the triple salt Northupite, $NaCl \cdot MgCO_3 \cdot Na_2CO_3$ [73] or $2\ MgCO_3 \cdot 2\ Na_2CO_3 \cdot 2\ NaCl$ [74], in the production units and therefore cause damage.

The most common and most problematic impurities in crude salt are the sulfates, chlorides, and, to some extent, the carbonates of calcium and magnesium, and the triple salt polyhalite ($K_2SO_4 \cdot 2\ CaSO_4 \cdot MgSO_4 \cdot 2\ H_2O$), but also the water which is used for dissolving rock salt contributes to the content of impurities in the crude brine. Most of the magnesium compounds are very soluble in crude brine. Furthermore, the crude brine is saturated with calcium sulfate in many cases. Sodium chloride brine significantly enhances the solubility of calcium sulfate [75]. This effect is called “salting-

in".

The principal impurities of crude brine are therefore calcium, magnesium, and sulfate ions, but also strontium ions (usually dissolved in relatively low concentrations in crude brine; see Table 10), all of which take part in scale formation. Potassium ions are irrelevant for the formation of incrustations in the evaporation equipment due to their high solubility. They only play a role as impurities in the final high-purity product [76].

Table 10. Compositions of crude and purified brine, and mother liquor (in g/L, unless otherwise noted)

	Mixed crude brine from three Austrian mines [86]	Crude brine, Bernburg, Germany, values calculated from [71]	Some analyses of purified brine [76]	Mother liquor, Ebensee, Austria [86]	Max. limits for ultrapurified brine for membrane chlorine-alkali electrolysis [87]
Na^+	117.82	125.52		120.85	
K^+	2.487	0.57		46.26	
Mg^{2+}	1.500	0.05	0.0002		**
					**
Ca^{2+}	0.809	1.53	0.005*		
Sr^{2+}	0.025		0.004*		< 400 ppb
Ba^{2+}					< 500 ppb
Fe^{2+} and Fe^{3+}					1 ppm
Cl^-	183.24	191.08		189.97	
Br^-	0.065				
I^-					< 200 ppb
SO_4^{2-}	8.69	3.66		50.24	5.4 g/L
HCO_3^-	0.102				
CO_3^{2-}				1.02	

* Concentrations at 14 °C.
** The combined concentration of Mg^{2+} and Ca^{2+} must be < 20 ppb.

The following chemical methods are used to purify crude brine [50].

Purification with Lime or Caustic Soda and Soda Ash. Traditional brine purification is performed in the Schweizerhalle process, named after a Swiss saline. It consists of two steps:

1. Addition of calcium hydroxide to precipitate magnesium ions as magnesium hydroxide and produce gypsum due to the increase in the concentration of calcium ions in a solution already saturated with calcium sulfate:

   $$Mg^{2+} + SO_4^{2-} + Ca(OH)_2 \rightarrow Mg(OH)_2 + CaSO_4$$

   Usually, solid calcium oxide is preferred over milk of lime, because it does not dilute the saturated sodium chloride solution further. Since lime is hydrated much more slowly in brine than in water, it is important to use highly reactive, fine-grained lime. However, if a higher degree of utilization of the calcium oxide is desired, milk of lime with a content of about 20 % calcium oxide can be used as well [76]. Where cheap sodium hydroxide is available, it can be used instead of calcium hydroxide, thus avoiding additional input of calcium ions. In this step, an excess of base is always added to reach a pH of about 11.3.
2. The addition of soda ash to precipitate the remaining calcium ions as calcium carbonate:

   $$Ca^{2+} + Na_2CO_3 \rightarrow CaCO_3 + 2\,Na^+$$

   The crystallization of calcium carbonate can also be achieved by purging the alkaline brine with carbon dioxide, easily available in the form of combustion gas, which contains about 10 – 14 vol % carbon dioxide. Although a minimum concentration of dissolved calcium ions is obtained at a pH of about 9.5, purging is usually stopped at a pH of approximately 11.2, thus avoiding the formation of soluble calcium hydrogencarbonate, which can hardly be removed [76].

After these two reactions, magnesium and calcium ions are removed almost quantitatively (see Table 10). The insoluble compounds are decanted and filtered after each purification step. The resulting basic sludge is generally considered to be useless and is therefore suspended in water and pumped back into abandoned caverns of the salt mine, or washed into rivers, lakes, or the sea. There it covers the ground and thus may threaten the equilibrium of the aquatic ecosystem.

A typical value for the amount of sludge is 15 kg per m^3 of crude brine or 50 – 55 kg per ton of evaporated salt. In Austria, for example, the composition of the water- and salt-free brine purification mud is 58.8 % calcium sulfate, 23.4 % calcium carbonate, 16.8 % magnesium hydroxide, and 1.0 % clay minerals, which are impurities of the limestone [77].

Recently, the use of this sludge as a lime fertilizer has been discussed [77-79]. However, three aspects must be considered:

1. The brine purification mud must be washed thoroughly to remove most of the chloride. A fertilizer should contain < 1 % chloride [79].
2. The sludge must have a sufficiently low water content (≤ 28 wt %) if conventional fertilizer-distributing machines are to be used [79].
3. Such fertilizer does not contain significant amounts of the three major macronutrients nitrogen, phosphorus, and potassium. Earlier studies on the growth of plants on similar wastes show that growth is limited primarily by the low content of phosphorus [80].

Sulfate is not removed completely from the crude brine in this purification process. Salt is obtained from the purified brine by evaporation, which is stopped prior to the onset of sodium sulfate crystallization. The residual solution is called “mother liquor” and contains sodium and potassium chloride, sulfate, and carbonate (due to the excess of soda ash in the second reaction; see Table 10). The saturation concentration of sulfate depends on the composition of the purified brine, but normally the value is about 48 g of sodium sulfate per kilogram of mother liquor. Brine with a higher potassium content allows a higher sulfate concentration in the mother liquor without crystallization of an alkali metal sulfate [76]. After removing the mother liquor, it can be fed to an expansion evaporator, where a second fraction of pure sodium chloride can be produced with an extent about 50 kg of sodium chloride per cubic meter of mother liquor. The residual solution can be used for the production of sodium sulfate, or, in combination with purified brine, for the production of soda ash, where sodium sulfate is also an appropriate sodium source. Since a high sulfate concentration in the brine for the Solvay process causes formation of gypsum scales during the recovery of ammonia with milk of lime (see ⟶ Sodium Carbonates – Technology), this strategy is not entirely unproblematic.

Purification in a Circulating System. Here, the brine is usually treated in two stages. In the first stage, crude brine, mother liquor, and calcium hydroxide suspension or sodium hydroxide solution are thoroughly mixed. This causes magnesium hydroxide, calcium sulfate, and a little calcium carbonate to precipitate. The crystallization of calcium carbonate is caused by the content of carbonate ions in the mother liquor. It is also formed if the crude brine contains hydrogencarbonate ions, due to the following equilibrium:

$$HCO_3^- + OH^- \rightleftarrows CO_3^{2-} + H_2O.$$

Other problematic metal ions are precipitated in addition to the alkaline earth metal ions. After removal of these solids, sodium hydroxide solution and/or sodium carbonate solution are sometimes added to the clear brine, which is saturated with calcium hydroxide, in the second stage. Additionally, carbon dioxide (usually in the form of combustion gas) is blown through, whereby calcium carbonate crystallizes and is removed. The clarified brine is fed to the evaporator and evaporated until sodium sulfate just fails to crystallize. The salt is recovered, the major part of the mother liquor (more than 80 %) is recycled to the first stage of the brine purification process, and the residue must be disposed of if it can not be made use of in a way described above. In practice the ratio between mother liquor and pure brine is about 1:5.

Split-Precipitation. A novel process [81, 82], property of Solvay S.A., can help make brine purification more ecologically acceptable by reducing the amount of brine purification mud. Instead of substituting magnesium by calcium ions, the first step is adding soda ash to the crude brine:

$$Mg^{2+} + Ca^{2+} + 2\,Na_2CO_3 \rightarrow MgCO_3 + CaCO_3 + 4\,Na^+$$

“$MgCO_3$” is the simplified notation for “basic magnesium carbonate” (“magnesia alba”), nonstoichiometric formula 4 $MgCO_3 \cdot Mg(OH)_2 \cdot 4$ or 5 H_2O [83] or 5 $MgO \cdot 4\,CO_2 \cdot x\,H_2O$ (x = 5 – 12) [84].

A certain percentage of magnesium ions is removed in the form of basic magnesium carbonate in this reaction and need not be precipitated with calcium hydroxide. This helps save milk of lime and furthermore reduces the input of calcium ions and consequently the amount of sludge. However, magnesium cannot be removed quantitatively in this step due to the solubility of precipitated (basic) magnesium carbonate in concentrated sodium chloride brine.

Therefore, in the next reaction milk of lime and, if used, mother liquor (“SO_4^{2-}”), are added to the brine to remove the remaining magnesium ions:

$$Mg^{2+} + SO_4^{2-} + Ca(OH)_2 \rightarrow Mg(OH)_2 + CaSO_4$$

Of course calcium carbonate is precipitated in this reaction as well, because of the remaining carbonate ions dissolved in the brine. In the final reaction calcium ions are precipitated in the traditional process using sodium carbonate. The precipitates are removed from the brine by decantation and/or filtration after each purification step. Before the sludge is disposed of into the waters, it is suspended in water and thoroughly stirred. Basic magnesium carbonate (coming from the first reaction) and calcium sulfate (coming from the second reaction) react within a short time:

$$MgCO_3 + CaSO_4 \rightarrow MgSO_4(aq.) + CaCO_3$$

This reaction is accelerated by remaining sodium chloride in the slurry [85].

Result: one solid, namely calcium carbonate (instead of two) and magnesium sulfate (soluble and nontoxic).

The resulting magnesium sulfate solution is diluted in the water and does not accumulate on the ground like the sludge. Especially if the waste waters are disposed of into the sea, the magnesium sulfate solution is of no ecological significance.

Since magnesium ions are removed from the crude brine in two different steps, this process was called Split-Precipitation. This process, especially the final step, helps reduce the volume of sludge by approximately 30 wt % compared with traditional brine purification, the quality of the purified brine staying the same. The reduction of the amount of mud is the main (ecological) advantage. If the extent of sodium chloride or soda ash production is legally regulated, limiting the amount of waste sludge, this process helps increase production.

The composition of the sludge is another ecological advantage: The main constituent is no longer gypsum, but about 50 % calcium carbonate, which is almost identical with the sediments in alpine regions. Less basic magnesium hydroxide is produced and washed into the waters, which is a further ecological advantage of this novel process.

Furthermore, if this process is used for the Solvay process, it must be kept in mind that soda ash production generates large amounts of calcium chloride wastewaters. The wastewaters of this novel brine purification process contain dissolved magnesium ions. Since calcium and magnesium are biological antagonists, the natural Ca/Mg equilibrium is partly restored.

Split-Precipitation causes only a very small increase in soda ash consumption, which is the main cost factor in brine purification, so it appears to be cost-efficient.

Design of Brine Purification Plants. The brine purification processes described above were formerly carried out as batch operations, but today continuous processes are more common. Continuous processes require an almost constant composition of the crude brine. On the other hand, they use smaller reactor vessels than batch processes. In both methods, the impurities are allowed to settle and are then removed. Sedimentation is accelerated by adding flocculating agents, usually anionic polyacrylates Flocculants – Chemical Structure and Flocculants – Uses. In the continuous process, the solid precipitates are removed in thickeners. Thus, the whole process consists of three stages:

1. Mixing with flocculants
2. Flocculation stabilization
3. Sedimentation and decantation of the solids

Continuous brine purification can be carried out in one reactor vessel as well [76]. In this case, the conical vessel is filled with crude brine and mother liquor from the bottom and milk of lime from the top. The formation of gypsum causes caustification. Sodium hydrogencarbonate is added and, due to the equilibrium with hydroxide ions, carbonate ions are formed, which precipitate calcium carbonate. The precipitates settle, and clarified brine is taken from the top. The resulting sludge contains about 60 % solids and is removed from the bottom. The main disadvantage of this process is that the brine has to be kept at a temperature between 80 and 100 °C.

Quality of the Purified Brine. Purified brine still contains low concentrations of problematic ions [76] (Table 10).

A minimum concentration of magnesium ions in brine is reached with a hydroxide concentration of about 0.003 mol/L in the brine. The solubility of calcium and strontium ions is temperature-dependent. Soda ash is usually used with 0.003 – 0.004 mol/L in excess.

The low-concentration alkaline earth metals in the purified brine form primarily carbonate scales in the evaporators, with the following composition: 71.5 % calcium carbonate, 22.2 % strontium carbonate, 2.1 % magnesium carbonate. These carbonate incrustations must be removed with acids. To avoid this problem, ion exchangers can be used to remove the remaining alkaline earth metal ions. This strategy is mainly used for the production of ultrapure brine for membrane chlorine-alkali electrolysis (for some limit values, see Table 10 and ⟶ Chlorine – Brine Purification).

Another method is keeping these ions dissolved by using complexing additives, usually polyphosphates or phosphonates. Such antiscalants prevent the formation of insoluble carbonates. The alkaline earth metal ions then accumulate in the mother liquor [76].

For the production of food-grade salt, the purified brine is filtered after the purification process to remove very small particles.

5.2. Evaporation of Brine

In an evaporative crystallizer, the salt normally crystallizes at 50 – 150 °C, the solubility of sodium chloride increasing only slightly with temperature (see Table 1). In the majority of cases purified brine is evaporated. Crude brine can be used for the production of industrial salt by the gypsum slurry process, in which a certain quantity of crystalline calcium sulfate hemihydrate is maintained in the suspension of salt crystals. Calcium sulfate precipitates deposits on these hemihydrate crystals and hence scale formation on the heating surfaces or the walls of the vessel is reduced. It was found helpful to raise the operating temperature to 150 °C when utilizing this process to prevent scale [88].

Salt was formerly produced by evaporation in open pans. Due to the high heat consumption and relatively poor performance, this method is meanwhile only used for the production of special sodium chloride products [50], [51]. These pans have an evaporation area of up to 180 m^2. In Germany, one saline, Luisenhall, probably the last one in Europe, still uses this method. Evaporation in open pans has the advantage that brine purification is not absolutely essential. This method of evaporation takes place at a temperature of about 85 °C and produces large, hollow pyramidal crystals [72], [76].

Heat is used most efficiently in multistage evaporation and vapor recompression processes Evaporation – Types of Evaporators and their Energy Consumption. Multistage evaporation units Evaporation – Multistage Evaporation always produce finer grained salt, they are robust and uncomplicated in operation, but their construction is expensive and their cooling water consumption high. For newly constructed plants, this type of evaporator is only economic with a volume of production of at least 80 – 100 t/h [78].

In the multistage evaporation process, heat is used more efficiently by compressing the vapors from the first evaporator to heat the second one. The pressure above the liquid decreases from unit to unit. The brine is preheated in preheaters which are heated with vapor and condensate from the previous evaporation unit. Three to seven stages can be used, but normally a multistage evaporation facility consists of five or six stages [72].

The heat of the vapor can be used in the same evaporator if the vapor is compressed to a higher pressure Evaporation – Evaporation with Vapor Recompression [52]. This vapor recompression system is widely used where cheap electrical energy is available. In this case, only one evaporator is required.

Hybrid systems are also used, which can be operated in multistage or vapor- compression evaporators [89], [90]. There are also combined methods in which high-pressure steam is passed through backpressure turbines, and the exhaust steam heats a multistage evaporation plant. The mechanical power is available for compressing the vapors.

In general, three types of crystallizers are used [53], [54]:

Type 1: Evaporators with forced circulation and external heating (see ⟶ Evaporation – Designing Evaporators for Fabrication, Fig. 11, type B)

Type 2: Evaporators with internal heaters and a circulating pump in the central pipe producing forced circulation (see ⟶ Evaporation – Designing Evaporators for Fabrication, Fig. 11, type A).

Type 3: Oslo crystallizers (Fig. 13)

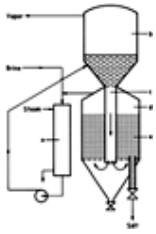

Figure 13. Oslo evaporator

a) Heat exchanger; b) Vacuum evaporator; c) Down pipe; d) Crystallizer; e) Bed of crystals

In Oslo crystallizers, recirculating brine and fresh brine are heated in the heat exchanger and evaporated in the evaporator. The supersaturated crystal-free brine passes down the central pipe into the crystallizer, and from there passes upwards through the bed of crystals.

The choice of crystallizer depends on the desired crystal size. If crystal size is not important, Type 1 crystallizers are generally used. These produce cubic crystals with 50 % > 400 µm (sieve analysis). If somewhat coarser salt is desired, Type 2 crystallizers are used, giving cubic crystals with rounded corners, with 50 % > 650 µm (sieve analysis). Type 3 crystallizers produce a granular product, with lenticular or spherical particles, about 1 – 2 mm in size. Evaporation in open pans produces the largest, hollow pyramidal crystals with an edge length of up to 24 mm.

5.3. Recrystallization

The International Recrystallizer process takes advantage of the fact that the solubility of calcium sulfate in brine decreases with increasing temperature, while that of sodium chloride increases slightly. The process is applicable when solid raw material like rock salt, solar, or waste salt is to be refined to obtain a fairly pure product [50], [55].

The process consists of three stages:

1. Dissolution: Crude salt is heated with brine, which is a kind of mother liquor of this process, from the evaporator to about 105 °C in a dissolving vessel by passing steam into it, which condenses in the brine. The sodium chloride dissolves, but not the calcium sulfate, because this mother liquor is already saturated with calcium sulfate.
2. Filtration: Removal of calcium sulfate crystals and other solid impurities from the brine.
3. Crystallization: The clear brine is pumped to the crystallizer, which is a flash evaporator with forced circulation but without heaters. On expansion in this evaporator, evaporation and cooling take place. This causes sodium chloride to crystallize. On cooling, the mother liquor becomes subsaturated with respect to calcium sulfate, so that pure sodium chloride is produced. The salt suspension is removed from the evaporator, and the mother liquor is returned to the dissolution vessels. Only trace amounts of calcium sulfate appear on the salt crystals, due to mother liquor that has dried on their surface. By washing the filter cake before drying, the content of calcium sulfate in the final product can be decreased to less than 5 ppm.

Heat exchangers are generally not used in this process, which can be carried out in one or several stages.

The principal advantages of recrystallization are as follows:

1. Low steam consumption per ton of sodium chloride
2. Production of a fairly high purity product
3. Low initial costs

5.4. Other Process Steps

Manufacture of Special Salt Crystals. Salt crystals with special shapes can be produced by adding substances to the brine that affect crystal growth in the evaporation equipment. The addition of soluble hexacyanoferrate(II) complexes in amounts of about 10 mg/kg produces dendritic crystals. This salt has a very low bulk density of ca. 0.8 g/cm^3. Other suitable additives include nitrilotriacetic acid trisodium salt, nitrilotriacetamide, nitrilotripropionamide, sodium hexametaphosphate, and cadmium chloride [56], [57].

Coarse salt is also produced by compressing fine salt to form tablets or briquettes, or into compacted sheets that are then broken up to form coarsely granular salt [50].

Construction Materials. In plants producing food-grade salt, corrosion [50], [58], and therefore contamination of the product must be avoided under all circumstances. Thus, for newly built plants high installation costs are accepted rather than high maintenance costs. In this case, noble materials like 316L stainless steel are used not only for the construction of the main equipment but also for pumps, pipes, valves, and instrumentation. Other materials which can be used are: Monel (nickel – copper) alloy plated steel (especially for the evaporator itself), titanium and titanium alloys, and copper – nickel alloys (both used for the heat exchanger tubes). Rubber or plastic linings can also be used, especially poly (vinylidene fluoride) (K. F. Polymer).

Removal, Washing, and Drying of the Crystals. The crystal suspension is removed from the bottom of the conical evaporators. To completely remove the sulfate-containing mother liquor, which is purged continuously, the crystals are washed with purified brine, which enters the evaporator at the bottom, or in a pusher centrifuge. Hydrocyclones (⟶ Hydrocyclones) or decanters are used to separate the salt crystals from the solution before centrifuging [78]. If washing is not required, a screw centrifuge can be used as well. After centrifuging, the salt still contains ca. 2 – 3 % moisture. Further drying can be carried out in rotary driers, but pneumatic conveying systems using hot air or fluidized-bed driers are now more common (see ⟶ Drying of Solid Materials – Drying Methods and Dryer Types). At the end of the fluidized-bed dryer there is usually a cooling zone into which cold air is blown, or a separate cooler can be used. The dried salt is often passed through a classification process.

5.5. Salt Quality

Typical analyses of salt produced by evaporation of crude and purified brine are given in Table 7. Anticaking agents are added to the salt, such as sodium, potassium, or calcium hexacyanoferrate(II) in a concentration of 5 – 15 mg/kg, or free-flow agents like finely grinded calcium and magnesium carbonate, sodium silicoaluminate, silicon dioxide, or calcium silicate. Sometimes both types of anticaking agents are combined. Purity requirements for salt are continually increasing, especially for industrial applications such as chlorine-alkali electrolysis.

Table 7. Typical analyses of salt from evaporation processes

	From crude brine [50]	From purified brine
NaCl, %	99.7	99.95
SO_4^{2-}, mg/kg	1900	300
Ca^{2+}, mg/kg	900	7
Mg^{2+}, mg/kg	20	0.5

[Top of Page]

6. Economic Aspects

Georg Steinhauser

6.1. Commercial Grades

6.1.1. *Salt for Food.*

Salt intended for human consumption is known as cooking salt or table salt. Apart from its function as a seasoning in foods, it is of vital importance for the human organism. The average daily salt requirement for humans is 6 g [91]. In Germany, the fraction of total salt consumption used for food is only approximately 3 %. In 2003, the consumption of table salt was about 420 000 t. The largest use in the food industry is for food preservation. Only ca. 100 000 t is sold in small packages for household use [59], [92].

Caking of table salt is prevented by adding anticaking agents (see Section Salt Quality).

Table salt is used as a carrier for various active substances. Iodized table salt is important worldwide in the prevention of diseases of the thyroid gland. Fluoridated table salt is used to prevent tooth decay. A water-soluble fluoride is used for fluoridation. In Germany, only iodate is permitted for iodization of salt. In many other countries iodide is also accredited. However, both ions are equal with respect to their physiological properties. The amount of iodine added to the salt varies from country to country: in Germany, for example, 15 – 25 mg iodine per kilogram of salt is added [91], whereas in

Switzerland the amount is 20 – 30 mg iodine per kilogram of salt. In both countries, a maximum of 250 mg fluoride is added to fluoridate one kilogram of salt.

Iodization and fluoridation are performed by spraying a concentrated solution of the particular sodium or potassium salt over the crystalline sodium chloride and thorough mixing. This process can be carried out continuously or as a batch operation. Usually drying is not necessary, since one tonne of salt is sprayed with only about one liter of solution [93].

In Europe, ca. 80 – 90 % of all meat products are preserved with nitrite pickling salt, which consists of sodium chloride and various amounts of sodium nitrite [59]. In Germany, 0.4 – 0.5 % sodium nitrite is added [94]. Due to the combination with salt, nitrite can hardly be overdosed, because then the meat product would become too salty.

6.1.2. *Treated Salt*

Salt that is used neither as a foodstuff nor for the chemical industry is denatured in many countries for tax reasons. The consumption of this type of salt in Germany in 2003 was about 700 000 t.

Salt that is not suitable for human consumption or food production is produced on demand by the addition of denaturants such as eosin, patent blue, or soda ash.

Treated salt is used in water softening, textile treatment, the leather industry, animal feed production, and in many other fields.

Of the above, the regeneration of water-softening ion exchangers is a very important application [60]. Water softening is essential owing to the widespread occurrence of hard water, and so far there is no alternative to the long-established and well-proven method based on ion exchangers and salt.

6.1.3. *Industrial salt*

Industrial salt is salt that is used as raw material for producing other compounds [61], [62]. The main applications of industrial salt in the chemical industry are in chlor-alkali electrolysis (⟶ Chlorine, ⟶ Sodium Hydroxide) and sodium carbonate (⟶ Sodium Carbonates) production.

The products obtained form the basic raw materials for innumerable other products, including glass, dyes, washing and cleaning agents, medicines, plastics, aluminum, and paper (Fig. 14).

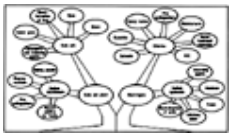

Figure 14. Family tree of salt chemistry [63]

In Germany in 2003, about 6.2×10^6 t rock salt was used, and about 6.7×10^6 t in the form of brine.

As with all forms of salt, industrial salt is subject to tax laws, but it may be used without being denatured if a salt license has been issued. For technical reasons, no anticaking agents are added to industrial salt.

6.1.4. *De-icing Salt*

Rock salt and sea salt can be used for de-icing road surfaces to give safe driving conditions for traffic in winter. The chemical processes (hydration) that take place when de-icing salt comes into contact with snow and ice on the roads in winter effectively eliminate slippery conditions [64], [95].

The effect of de-icing salt on the environment has been the subject of controversy. However, in the wake of mounting evidence from scientifically based investigations, there is now little opposition to the sparing application of de-icing salt to ensure the safety of traffic, and the frequency of accidents caused by snow and ice can be drastically reduced.

The development and use of suitable technology have contributed greatly to decreasing the amount of de-icing salt used and hence to reduction of environmental pollution. This technology makes use of modern electronic equipment for automatically adjusting the salt spreading process to the width of the road and the speed of the gritting vehicles, and the spreading of moist salt.

For the spreading of moist salt, the dry salt is wetted with a 5 – 30 % solution of sodium chloride, calcium chloride, or magnesium chloride. Moistening prevents windblown losses and gives a more rapid de-icing effect [64]. For this purpose, the distillation wastewater of the Solvay Process (⟶ Sodium Carbonates – Ammonia – Soda Process) can be used as well, because it contains about 115 kg calcium chloride and about 50 kg sodium chloride per cubic meter of solution.

In practise, the ratio between de-icing salt and solution is usually 70/30. In this case, dry salt is stored in the salt-dispensing vehicle and moistened with the solution instantaneously when dispersed. Another method is to store a moist mixture of 95 % salt and 5 % solution in a tank in the vehicle and spread this mixture.

In the 1960s, at least 40 g/m^2 de-icing salt was required, but subsequent technical developments have enabled an optimum de-icing effect to be achieved with only 10 – 20 g/m^2. In 2003, ca. 1.7×10^6 t of de-icing salt was produced in Germany.

6.2. Packaging, Transport, and Storage

Salt is supplied both loose and in packaged form. Salt for human consumption is available in big bags, in plastic and paper sacks, and in small packets. Denatured salt and de-icing salt are supplied loose or in packages, but salt intended for large-scale industrial use is only supplied in loose form.

Salt is transported by canal and river boats, or by road in tipping trucks, flat trucks, or silo trucks.

6.3. Economic Data

The world reserves of salt are very large. Economic and subeconomic deposits of salt are substantial in principal salt producing countries. The oceans contain an estimated amount of 5×10^{16} t of mineral salts [96], which is a practically inexhaustible supply of salt. The world production of salt exceeded 200×10^{6} t in 1996 for the first time [72].

A list of the most important salt producers in Germany follows:

esco – european salt company GmbH & Co. KG, Hannover (Bergwerk Bernburg; Saline Bernburg; Bergwerk Borth; Saline Borth; Bergwerk Braunschweig-Lüneburg; Bergwerk Hattorf)
Saline Luisenhall GmbH, Göttigen (Saline Luisenhall)
Südsalz GmbH, München (Bergwerk Berchtesgaden; Saline Bad Reichenhall; Saline Bad Friedrichshall)
Südwestdeutsche Salzwerke AG, Heilbronn (Bergwerk Heilbronn)
Wacker-Chemie GmbH, München (Bergwerk Stetten)

Figure 15 shows production and consumption figures in Germany for the years 2002 and 2003, respectively, classified according to use.

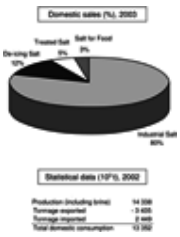

Figure 15. Domestic salt production and sales in Germany in 2002 and 2003, respectively [97]

Figure 16 shows the statistical development of salt works and salt mines in Germany.

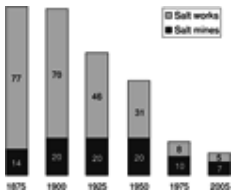

Figure 16. Statistical development of salt works and salt mines in Germany [97]

Table 8 shows the development of salt production in Europe, and Table 9 lists world production figures.

Table 8. Production of sodium chloride in Europe (in 10^3 t) [65], [97], [98]

Country	1960	1970	1980	1990	2000	2002*
Austria	532	543	696	675	401	401
Bulgaria	96	149	96	93	1 700	1 900
Czechoslovakia	185	286	805	227		
Slovakia					99	100
Denmark		181	420	552	605	610
EC/EU countries (total)	22 544	40 306	46 446	37 216	42 056	42 061
France	4 118	6 239	7 830	6 610	7 000	7 000
Germany					12 600	14 338
Democratic Republic	1 968	2 408	3 448	4 125		
Federal Republic	3 969	10 512	12 655	11 585		
Greece	107	125	133	150	150	150
Italy	2 866	4 841	5 806	4 430	3 600	3 600
Netherlands	1 208	3 166	3 818	3 650	5 000	5 000
Poland	2 145	3 200	3 998	4 060	4 307	4 200

Portugal	236	442	582	648	600	600
Romania	1 152	3 155	5 772	4 260	2 267	2 250
Soviet Union	8 300	13 700	16 094	14 700		
Armenia					30	30
Azerbaijan					4	4
Belarus					311	300
Russia					3 200	2 800
Turkmenistan					215	215
Ukraine					2 287	2 300
Spain	1 533	2 264	3 867	3 380	3 200	3 200
Switzerland	164	368	417	254	300	300
Turkey				1 890	2 126	2 100
United Kingdom	6 539	10 128	7 887	6 430	5 800	5 800
Yugoslavia	166	281	413	375		
Bosnia and Herzegovina					50	50
Croatia					34	37
Serbia and Montenegro					78	42
Slovenia					2	2
Europe (total)	35 285	61 992	74 813	68 142	57 437	58 706

*Estimated.

Table 9. World production of sodium chloride (in 10^3 t) [65]

Region/country	1960	1970	1980	1990	2000	2002*
Africa and the Middle East	1 878	2 371	2 690	5 393	8 686	9 330
Egypt	575	452	701	989	2 400	2 400
Republic of South Africa	279	468	625	728	346	431
North and Central America	29 872	57 009	56 848	57 337	69 433	63 901
Canada	3 311	5 359	7 748	11 300	12 164	12 313
United States	25 481	45 896	40 378	36 940	45 510	40 290
South America	2 260	4 982	7 519	9 370	13 137	13 531
Argentina	605	1 056	1 107	834	1 000	1 000
Brazil	1 017	2 013	4 230	5 370	6 074	5 600
Asia and Oceania	22 940	30 704	36 341	42 641	61 900	64 601
China	14 200	17 600	19 048	20 000	31 280	32 835
India	3 782	6 160	8 829	9 500	14 453	14 503
Australia	519	3 385	5 859	7 230	8 778	9 887
New Zealand	19	58	7	80	60	70
Europe	35 285	61 992	74 813	68 142	57 437	58 706
Western Europe	21 273	38 813	44 112			
Eastern Europe	14 012	23 179	30 701			
World total	92 773	160 501	184 077	183 000	212 000	210 000

*Estimated.

[Top of Page]

References

1. W. Botsch: *Salz des Lebens*, Kosmos Gesellschaft der Naturfreunde, Franckh'sche Verlagshandlung, Stuttgart 1976.
2. H. Seidel, R. Woller: *Das Geschenk der Erde*, ECON Verlag, Düsseldorf 1980.

3. M. J. Schleiden: *Das Salz, seine Geschichte, seine Symbolik und seine Bedeutung im Menschenleben*, Verlag Wilhelm Engelmann, Leipzig 1875.
4. S. Falkenberg: *Salz ist Leben, seine Bedeutung für Ihre Gesundheit*, Ariston Verlag, Genf 1987.
5. P. Meier: *Die Würze des Lebens, Salz*, Schweizer Verlagshaus, Zürich 1983.
6. J.-F. Bergier: *Die Geschichte vom Salz*, Campus Verlag, Frankfurt/New York 1989.
7. D. W. Kaufmann: *Sodium Chloride*, Reinhold Publ. Co., New York 1960, pp. 587 ff.
8. *J. Phys. Chem. Ref. Data* **11** (1982) 15 – 81; **13** (1984) 1 – 102. Links *Gmelin*, 8th ed., **System No. 21**, pp. 309 ff. *Landolt-Börnstein*, 6th ed., **2**, part 8, 46 ff.
9. Te-Pang Hou: *Manufacture of Soda*, Reinhold Publ. Co., New York 1942, p. 118.
10. S. K. Coburn, *Corrosion* **10** (1954) no. 5, 1.
11. W. W. Bradley, *Corrosion (Houston)* **11** (1955) 383.
12. DECHEMA-Werkstofftabelle, Blatt Nr. DWT 1043 ff,3. Bearbeitung, DECHEMA, Frankfurt/Main 1962.
13. S. J. Lefond: *Handbook of World Salt Resources*, Plenum Press, New York 1969.
14. G. Spackeler: *Lehrbuch des Kali- und Steinsalzbergbaues*, VEB Wilhelm Knapp Verlag, Halle (Saale) 1957, pp. 243 – 254.
15. H. Jendersie: *Kali- und Steinsalzbergbau*, vol. **2**, VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1969, pp. 150 – 154.
16. G. Brückner, *Freiberg. Forschungsh.* **A 418** (1967) 11 –46.
17. G. Brückner, *Bergakademie* **10** (1961) 625 – 629.
18. P. Ambatiello, *5th Symp. on Salt*, **vol. 2**, The North. Ohio Geol. Soc. Corp., Cleveland, Ohio 1978, pp. 213 – 218.
19. K. Thomanek, *5th Symp. on Salt*, **vol. 2**, The North. Ohio Geol. Soc. Corp., Cleveland, Ohio 1978, pp. 171 – 182.
20. P. Ambatiello, *Glückauf* **118** (1982) 1183 – 1188.
21. G. Dorstewitz, C. H. Fritzsche, H. Prause, *Glückauf* **95** (1959) 1245 – 1250.
22. P. Ambatiello, *Felsbau* **8** (1990) 180 – 184.
23. G. Feder, M. Hoscher, *Felsbau* **5** (1987) 114 – 149.
24. K. Thomanek, *BHM Berg Hüttenmänn. Monatsh.* **127** (1982) 381 – 389.
25. M. Hoscher, *BHM Berg Hüttenmänn. Monatsh.* **126** (1981) 234 – 238.
26. K. Schmidt, *Tech. Mitt.* **68** (1975) 363 – 368.
27. H. Gomm, J. Hieblinger, G. Kühn, *Erdöl Erdgas Z.* **93** (1977) 83 – 90.
28. H. K. Röhr, *Erdöl Kohle Erdgas Petrochem.* **22** (1969) 670 – 679, 734 – 741.
29. S. Meister, G. Kuhr, *Erdöl Erdgas Z.* **88** (1972) 248 –257.
30. H. Borchert, W. Dreyer: Gutachten über die Standfestigkeit und Konvergenz von Kavernen im Salzstock, Mineralog.-petrograph. Inst. d. TU Clausthal, 10 Dec. 1967.
31. W. Dreyer, *Kali Steinsalz* **5** (1971) 473 – 478.
32. S. Serata in J. L. Rau (ed.): “Continuum Theory and Model of Rock Salt Structures,” 2nd Symp. on Salt, Northern Ohio Geol. Soc., Cleveland 1965.
33. R. W. Durie, F. W. Jessen, *Soc. Pet. Eng. J.* 1964 183 –190.
34. W. Dreyer: *Gebirgsmechanik im Salz*, Enke Verlag, Stuttgart 1974.
35. K.-H. Lux: *Gebirgsmechanischer Entwurf und Felderfahrungen im Salzkavernenbau*, Enke Verlag, Stuttgart 1984.
36. F. Schober, A. Sroka: “Die Berechnung von Bodenbewegungen über Kavernen unter Berücksichtigung des zeitlichen Konvergenz- und Gebirgsverhaltens,” *Kali Steinsalz* **8** (1983) March, no. 10.
37. A. Sroka, F. Schober: “Die Berechnung der maximalen Bodenbewegungen über kavernenartigen Hohlräumen unter Berücksichtigung der Hohlraumgeometrie,” *Kali Steinsalz* **8** (1972) Aug., no. 8.
38. A. Sroka: “Abschätzung einiger zeitlicher Prozesse im Gebirge,” Schriftenreihe Lagerstättenerfassung und -darstellung, Bodenbewegungen und Bergschäden Ingenieurvermessung Kolloquium, Leoben, 15 – 16 Nov. 1984.
39. A. Sroka, A. Hartmann: “Die überwachung von Speicherkavernen mittels geodätischer und physikalischer Beobachtungen,” Proceedings 5th. International Congress, International Society of Mine Surveying Harrogate, 9 – 13 Sept. 1985.
40. H. von Schonfeldt, L. Remoline in A. H. Coogan (ed.): “Mechanism of Solution and Cavity Control in a Two-Well-System with Emphasis on Jetting,” 4th Symp. on Salt, **vol. 2**, Northern Ohio Geol. Soc. Inc., Cleveland 1974, pp. 189 – 202.
41. K. Henderson: “Methods of Joining Two or More Wells for Brine Production,” in [40], pp. 211 – 218.
42. H. Rischmüller, *Erdöl Erdgas Z.* **88** (1972) 240 –248.
43. H.-G. Haddenhorst *et al.*, *Erdöl Erdgas Z.* **90** (1974) 154 – 161, 197 – 204.
44. H. Bresson, *Rev. fr. énerg.* **26** (1969) 527 – 543.
45. M. Klafki, H.-J. Kretzschmar, W. Menzel, W. Schreiner: “Aussolung von Gasspeicher-Kavernen in mächtigen Salzlagerstätten,” *Neue Bergbautech.* **20** (1991) March, no. 3.
46. E. F. Armstrong, L. M. Miall: *Raw Materials from the Sea*, Chemical Publishing Co., Brooklyn 1946.
47. Compagnie des Salines du Midi et des Saline de l'Est (C.S.M.E.: *Le Sel*, Presses Universitaires de France, Paris.
48. Laboratoire C.S.M.E., Aigues-Mortes, France, personal communication.
49. Central Research Institute, Japan Monopoly Corporation, Figures for Salt Production, Tokyo 1954.
50. D. W. Kaufmann: *Sodium Chloride*, Reinhold Publ. Co., New York 1960.

51. *Ullmann*, 3rd ed., **12**, 674.
52. J. V. Rozycki: **vol. 2**, in [40], pp. 425 – 431.
53. C. M. van 't Land, K. J. A. de Waal in J. W. Mullin (ed.): *Industrial Crystallisation: Proceedings of the Sixth Symposion on Industrial Crystallization*, Plenum Press, New York 1976, pp. 449 – 459.
54. C. M. van 't Land, B. G. Wienk: in [53], pp. 51 – 60.
55. L. J. Theilgard, 1st Symp. on Salt, Northern Ohio Geol. Soc. Inc., Cleveland 1963, pp. 417 – 423.
56. M. A. van Damme-van Weele: *Influence of Additives on the Growth and Dissolution of Sodium Chloride Crystals*, PhD thesis, Technische Hogeschool Twente, Drukkerij Smit, Hengelo 1965.
57. E. G. Cooke, **vol. 2**, in [32], pp. 259 – 268.
58. Salt Institute, Recent successes and Failures in Combating Corrosion in Salt Plants, Seminar Chicago, IL, June 1977.
59. G. Reimer, C. Thieme in K. Winnacker, L. Küchler, H. Harnisch (eds.): *Chemische Technologie*, 4th ed., vol. **2**, Hanser Verlag, München 1982, pp.492 – 494.
60. VDS: *Weiches Wasser – Komfort und Kalkschutz*, 2nd ed., Bonn 2003.
61. W. Botsch: *Salz des Lebens*, Franckh'sche Verlagshandlung, Stuttgart 1976, pp. 69 ff.
62. VDS: *Salz II – Rohstoff in der Chemie*, Hanser Verlag, München 1984, pp. 36 – 38.
63. Deutsche Solvay-Werke GmbH: *Das Salz der Chemie*, Solingen 1983, pp. 28 – 29.
64. VDS: *Winterdienst – Verkehrssicherheit und Umweltschutz*, 2nd ed. Bonn 2003.
65. U.S. Bureau of Mines: *Minerals Yearbook 2002*, Reston, Virginia, 2002.
66. O. Schauberger, *Arch. f. Lagerst.forsch. Geol. B.-A.* **7** (1986) 217 – 254.
67. M. Hoscher, *BHM Berg Hüttenmänn. Monatsh.* **134** (1989) 230 – 234.
68. Morton Salt, Chicago, USA, personal communication, 2004.
69. S. J. Lefond: *Handbook of World Salt Resources*, Plenum Press, New York 1969.
70. The Canadian Salt Company Limited, Quebec, Canada, personal communication, 2004.
71. A. Hiltscher, C. Heinrich, *Kali Steinsalz* **11**, (1994) no. 7, 211 – 217.
72. H. Flachberger, *BHM Berg Hüttenmänn. Monatsh.* **143** (1998) 115 – 123.
73. *Winnacker-Küchler*, 4th ed., **2** (1982) 503.
74. *Gmelin*, 8th ed., System no. 27, Magnesium, Teil A, p. 78.
75. J. D'Ans, D. Bredtschneider, H. Eick, H.-E. Freund, *Kali Steinsalz* **1** (1955) no. 9, 17 – 38.
76. H. Schwaiger, *BHM Berg Hüttenmänn. Monatsh.* **143** (1998) 128 – 132.
77. H. Flachberger, K. Krenn, *BHM Berg Hüttenmänn. Monatsh.* **144** (1999) 234 – 240.
78. K. Krenn, *BHM Berg Hüttenmänn. Monatsh.* **143** (1998) 124 – 127.
79. H. Flachberger, K. Krenn, *BHM Berg Hüttenmänn. Monatsh.* **149** (2004) 138 – 143.
80. J. A. Lee, B. Greenwood, *Biol. Conserv.* **10** (1976) 131 – 149.
81. Solvay S. A., AT 6 412 U, 2002 (G. Steinhauser).
82. Solvay S. A., AT 7 006 U, 2003 (G. Steinhauser).
83. J. Falbe (ed.): *Römpp-Lexikon Chemie*, 10th ed., **vol. 4**, Thieme, Stuttgart 1998, p. 2485.
84. *Ullmann*, 3rd ed., **12** (1960) 117.
85. E. Fleischer: "Über das Verhalten der kohlensauren Magnesia zu Gyps bei Gegenwart von Kochsalzlösung," *J. pr. Ch.* **114/6** (1873) 273.
86. Solvay Österreich GmbH, Ebensee, Austria, personal communication, 2002.
87. Wacker-Chemie GmbH, Burghausen, Germany, personal communication, 2004.
88. J. Ravoire, P. Charuit, in [40], pp. 455 – 460.
89. R. Winkler, *Escher-Wyss-Mitteilungen* **34** (1961) nos. 2 - 3, 10 – 12.
90. R. O. Walter, P. Kondorosy, *Escher-Wyss-Mitteilungen* **48** (1975) no. 2, 3 – 9.
91. Verein Deutsche Salzindustrie e.V. (VDS): *Salz als Lebensmittel*, Bonn 2002, pp. 18, 15.
92. esco – european salt company GmbH & Co. KG, Hannover, Germany, personal communication, 2004.
93. Vereinigte Schweizerische Rheinsalinen, Schweizerhalle, Switzerland, personal communication, 2004.
94. VDS: *Kontrollierte Sicherheit in der Fleischverarbeitung – Nitritpökelsalz*, Bonn 1991, p. 6.
95. B. Ruess, *Strasse Verkehr* **6** (1998) 2 – 8.
96. K. H. Wedepohl, *Kali Steinsalz* **11** (1995) no. 10, 311 – 315.
97. Verein Deutsche Salzindustrie, Bonn, Germany, personal communication, 2004.
98. European Commission: *European minerals yearbook – Second edition 1996–97*, Office for Official Publications of the European Communities, Luxembourg 1998, pp. 213 ff.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Sodium Compounds, Organic

Rolf Ackermann, Degussa AG, Frankfurt am Main, Federal Republic of Germany

Ludwig Lange, Degussa AG, Frankfurt am Main, Federal Republic of Germany

1. Introduction 1
2. Structure and Properties 1
3. Preparation 2
4. Uses 3
5. References 3

1. Introduction

The discovery of the sandwich structure of ferrocene [bis(cyclopentadienyl)iron] in 1952 provided the impetus for fundamental changes in research in the field of organometallic compounds. One method of preparing ferrocene is by reaction of cyclopentadienylsodium with an iron halide or complex halide in a polar solvent such as THF [1].

2. Structure and Properties

The properties, reactions, and preparation of organosodium and –potassium compounds are described in detail in [2].

The organosodium and organopotassium compounds are generally ionic, in contrast to the corresponding lithium compounds. The ionic character of the metal – carbon bond increases with increasing electropositivity and atomic radius of the metal [3]. The chemical reactivity of the ionic organometallic compounds depends mainly on the carbanion. The carbanions can be divided into the following categories [4]:

1) Carbanions with localized charge: the negative charge is restricted to a single carbon atom (e.g., H_3C^-)
2) Carbanions with delocalized charge: the negative charge is distributed over two or more carbon atoms (e.g., $(C_6H_5)_3C^-$)
3) Radical carbanions: these are aromatic molecules that have acquired one (sometimes two) electrons (e.g., $C_{10}H_8^{\cdot -}$)

1) Organosodium and organopotassium compounds that contain carbanions with localized charge are extremely reactive and exhibit salt-like properties such as high melting point, low vapor pressure, and very low solubility in nonpolar solvents such as pentane or hexane [5]. Methylpotassium, for example, has a crystalline structure of the nickel arsenide type [4], with a trigonal prismatic arrangement of six potassium ions around a central methyl carbanion [3], [6]. The thermal stability of the sodium and potassium compounds is in general lower than that of the corresponding lithium compound, the potassium compounds often being unstable at room temperature. They generally ignite spontaneously on exposure to air and react violently with water.
2) Typical properties of organosodium and -potassium compounds of carbanions with delocalized negative charge include intense color, solubility in ether, and higher stability than that of the simple alkyl compounds, although they are also air and moisture sensitive. These properties are a result of the delocalization of the negative charge over the entire organic ligand [3]. Carbanions can be planar (sp^2 hybridization) or pyramidal (sp^3 hybridization). All the intermediate configurations are also possible, depending on the nature of the substituents. If one or more groups bonded to the carbanion stabilize the ion by conjugation, the sp^2 configuration is favored, since a planar arrangement ensures maximum overlap of the *p*-orbitals. This is true of the bonds of the central carbon atom in the triphenylmethyl carbanion [7]:

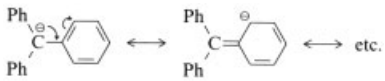

If triphenylmethyl chloride is reacted with sodium in an inert solvent, the chlorine

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_341

atom is displaced by sodium, and the deep red, ether-soluble compound tritylsodium is formed:

$$(C_6H_5)_3CCl \xrightarrow{Na/Hg} (C_6H_5)_3CNa$$

The solution of triphenylmethylsodium is used as an extremely strong base, the carbanion being a very good proton acceptor [7].

The cyclopentadienyl anion [7] produced when a hydrogen atom is cleaved from cyclopentadiene is stabilized by delocalization of the six π-electrons over all five carbon atoms to give a quasi-aromatic structure.

$C_5H_6 + Na \longrightarrow C_5H_5Na + 1/2\,H_2$

Another example is fluorenylpotassium [3]:

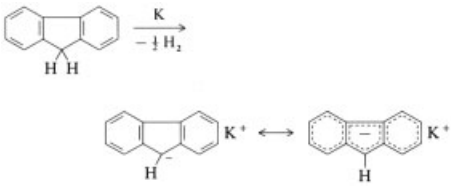

Owing to the charge delocalization, the effective radius of the anion is large, so that the coulombic interaction between cation and anion is weak. Therefore, salts of these anions are weakly dissociated in polar aprotic solvents such as hexamethylphosphoramide $[(CH_3)_2N]_3PO$ [8].

3) Direct transfer of an electron from the metal atom to the organic ligand results in formation of a radical anion [4]:

$C_{10}H_8 + K \longrightarrow K^+\,C_{10}H_8^{\cdot\,-}$

Radical ions are only stable at low temperature [9]. They decompose on heating to room temperature, whereby metalation of the organic group with loss of hydrogen can take place.

Sodium naphthalide and related compounds are strong reducing agents. As they are very soluble in ethers, they are often used in preference to sodium, which is insoluble. The resulting homogeneous reaction is usually faster, and is easier to control than a heterogeneous reaction between dissolved components and solid sodium [8].

3. Preparation

Under normal conditions, organosodium and organopotassium compounds cannot be obtained from alkyl halides and sodium or potassium metal [10]. With these metals, the main products are alkanes (Wurtz reaction) [5]. In hydrocarbon solvents alkyl halides react with sodium to give sodium alkyls as intermediates, which react with further alkyl halide to give the alkane [7]. In the case of alkyl chlorides, special methods are available to largely suppress the condensation reaction and direct the reaction towards the formation of organosodium compounds [3]. Halogenated alkanes can react explosively with highly reactive metals such as sodium or potassium. The commonly used method of removing traces of moisture from solvents by treatment with sodium must therefore never be used with halogenated alkanes [11].

Organosodium and -potassium compounds are accessible by three types of reaction:

Metalation. [3] Acidic hydrogen atoms can be replaced by sodium or potassium. The reaction can be carried out using the metal itself or a metal hydride, amide, alkoxide, or alkyl.

$$Ph{-}C{\equiv}CH + Na \longrightarrow Ph{-}C{\equiv}CNa + 1/2\,H_2$$

$$Ph_2CH_2 + NaNH_2 \longrightarrow Ph_2CHNa + NH_3$$

Halogen atoms can also be substituted:

$$CH_2{=}CHCl + 2\,Na \longrightarrow CH_2{=}CHNa + NaCl$$

The planar aromatic cyclopentadienyl ion $C_5H_5^-$ can be obtained by reductive cleavage of the C–H bond in cyclopentadiene, for which either metallic Na or NaH can be used [8]:

$$C_5H_6 + Na \xrightarrow{THF} Na[C_5H_5] + 1/2\,H_2$$

Transmetalation. [3] The transfer of an organic radical from one metal atom to another (e.g., sodium or potassium) is another method of producing organosodium or organopotassium compounds:

$$Me_2Hg + 2\,Na \longrightarrow 2\,NaMe + Hg$$

$$(CH_3CH_2)_2Hg + 2\,Na \longrightarrow 2\,CH_3CH_2Na + Hg$$

Electron Transfer. [3] The first product of the addition of sodium or potassium to an unsaturated organic system is a radical ion, which can react with a further metal atom to give a dianion. Radical anions are usually prepared by reacting a suitable hydrocarbon (e.g., anthracene) with sodium or potassium (e.g., in the form of an amalgam) at low temperature:

$$C_{14}H_{10} + Na/Hg \xrightarrow[-80\,^{\circ}C]{\text{Solvent}} Na^{+}\,C_{14}H_{10}^{\bullet -}$$

4. Uses

The organosodium compounds are of little industrial importance, apart from some that are used as catalysts. The so-called alfin catalysts for the copolymerization of butadiene with styrene or isoprene consist of allylsodium, iso-C_3H_7ONa, and NaCl.

Alkali-metal organometallics occur as intermediates in the reaction of halobenzenes with alkyl halides in the presence of sodium in ether solution, for the production of alkylbenzenes. The Wurtz–Fittig reaction takes place in two stages [10].

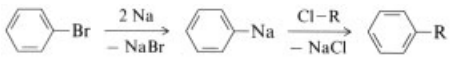

The yield of alkylbenzene depends mainly on whether the aromatic-bound halogen is more rapidly replaced by sodium than the aliphatic-bound halogen.

Further, phenylsodium must react more rapidly with the alkyl halide than with the aryl halide, i. e., the carbon–halogen bond in the alkyl halide must be more readily polarizable than that in the aryl halide.

Triphenylphosphine and tetraethyllead can be prepared by using modified Wurtz–Fittig reactions.

$$3\,C_6H_5Cl + PCl_3 + 6\,Na \longrightarrow (C_6H_5)_3P + 6\,NaCl$$

$$4\,PbNa + 4\,ClC_2H_5 \longrightarrow Pb\,(C_2H_5)_4 + 4\,NaCl + 3\,Pb$$

Sodium naphthalide, $Na^{+}\,[C_{10}H_8]^{-}$, in ether is often used as a strong reducing agent. The blue solution of sodium and benzophenone in THF is an effective reagent for removing traces of oxygen from nitrogen [12].

Cyclopentadienylsodium readily takes part in a metathetical reaction with many halides of *p*-block elements, forming σ- or π-bonded cyclopentadienyl compounds. Cyclopentadienylsodium reacts with transition metal halides to form metallocenes, for example:

$$2\,Na[C_5H_5] + MnCl_2 \xrightarrow{THF} Mn[C_5H_5]_2 + 2\,NaCl$$

The bis-cyclopentadienyl complexes of Fe, Co, and Ni can also be readily prepared by this method [8].

Organosodium and organopotassium compounds initiate the anionic polymerization of alkenes. Alkenes undergo insertion into the metal–carbon σ-bond, forming a new organopotassium compound with a longer chain [3]:

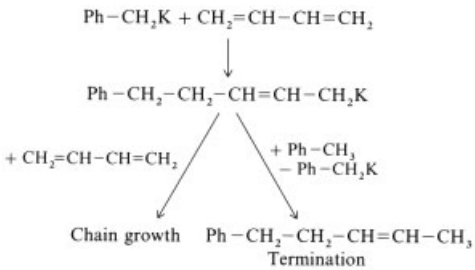

Owing to their great reactivity and capacity for spontaneous ignition in air, these compounds are used in situ as intermediate products and are seldom isolated.

5. References

1. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* 1st ed., VCH Verlagsgesellschaft, Weinheim 1988, pp. 407, 1419.
2. M. Schlosser in W. Foerst (ed.) : *Neuere Methoden der präparativen organischen Chemie,* **Vol. V,** Verlag Chemie, Weinheim 1967.
3. *Ullmann,* 4th ed., **17,** 231–232.
4. J. S. Thayer: *Organometallic Chemistry, an Overview,* VCH Verlagsgesellschaft, Weinheim 1988.
5. G. Wilkinson: *Comprehensive Organometallic Chemistry,* vol. 1, Pergamon, Oxford 1982.
6. E. Weiss, G. Sauermann, *Chem. Ber.* **103** (1970) 265.

7. P. Sykes: *Reaktionsmechanismen der organischen Chemie, eine Einführung,* 1st ed. Verlag Chemie, Weinheim, 1964; 9th ed. 1988.
8. D. F. Shriver, P. W. Atkins, C. H. Langford: *Anorganische Chemie, ein weiterführendes Lehrbuch,* VCH Verlagsgesellschaft, Weinheim 1992.
9. E. de Boer, *Adv. Organomet. Chem.* **2** (1965) 115; E. T. Kaiser, J. K. Kevan, *Radical Ions,* Wiley-Interscience, New York 1968.
10. H. Beyer, W. Walter: *Lehrbuch der Organischen Chemie,* 22nd ed., Hirzel Verlag, Stuttgart 1991.
11. Ch. E. Mortimer: *Chemie,* 5th ed., Thieme Verlag, Stuttgart 1987.
12. F. A. Cotton, G. Wilkinson: *Advanced Inorganic Chemistry,* J. Wiley & Sons, London 1980.

Sodium Hydroxide

CETIN KURT, Bayer MaterialScience AG, Leverkusen, Germany
JÜRGEN BITTNER, Bayer MaterialScience AG, Leverkusen, Germany

1. Introduction 1
2. Properties 1
2.1. Properties of Sodium Hydroxide . . . 1
2.2. Properties of Sodium Hydroxide Solution 2
3. Production 3
3.1. Production of Sodium Hydroxide Solution 3
3.2. Production of Solid Sodium Hydroxide 6
3.3. Forming 6
3.4. Specifications 8
3.5. Storage, Packaging, and Transportation 8
4. Uses 9
5. Safety Precautions 11
6. Economic Aspects 11
7. References 12

1. Introduction

Pure sodium hydroxide, NaOH, is a colorless solid. It does not occur in nature, but is manufactured on a large scale from fairly readily obtainable raw materials and is used in numerous chemical processes. Because of its corrosive action on many substances, it is known as caustic soda.

Sodium hydroxide solution is one of the oldest man-made chemicals. The reaction of sodium carbonate with calcined limestone (causticization of soda) was already known in early Afro-Oriental cultures. Alabaster vessels containing 3 % sodium hydroxide solution have been found in Egyptian tombs dating from 3rd century B.C. near the Pyramid of Cheops. The first written records of caustic soda production during the early years A.D. came from Egypt and India.

2. Properties

Because pure sodium hydroxide has a high affinity for water, the physical and chemical properties of the pure substance are difficult to determine. Many of the data are obtainable only by extrapolating values for impure sodium hydroxide. The same is true for the concentration-dependent physical properties of aqueous solutions of sodium hydroxide, since these solutions have a strong tendency toward supersaturation.

2.1. Properties of Sodium Hydroxide

Some physical properties of solid sodium hydroxide are as follows:

M_r	40
Appearance	white, crystalline
Density (liquid, 350 °C)	1.77 g/cm^3
mp (soda and water free)	322 ± 2 °C
Heat of fusion	6.77 kJ/mol
Boiling point at 0.1 MPa	1388 °C (calculated)
Specific heat capacity at 20 °C	3.24 J kg^{-1} K^{-1}

Chemical Properties. Pure sodium hydroxide is strongly hygroscopic. It dissolves in water with liberation of heat and forms six defined hydrates (see Figs. 1 and 2, and Table 1). When transporting and storing sodium hydroxide solution, the containers must be heated or insulated because the melting points of some of the hydrates are much greater than 0 °C. In the presence of moisture, NaOH reacts readily with atmospheric carbon dioxide to form sodium carbonate. Sodium hydroxide reacts with carbon monoxide under pressure and in the presence of moisture to yield sodium formate.

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_345.pub2

The high affinity of sodium hydroxide for water causes a reduction in water vapor pressure (e.g., to 0.36 kPa for 50.3 wt % NaOH at 30 °C). Sodium hydroxide is therefore a very effective drying agent. It is fairly readily soluble in methanol and ethanol.

Table 1. Stability of the hydrates of sodium hydroxide

Formula	Stability range, °C
$NaOH \cdot 7H_2O$	− 23.5 to − 28.3 [a]
$NaOH \cdot 5H_2O$	− 12.2 to − 31.9 [a]
α-$NaOH \cdot 4H_2O$	+ 7.6 to − 21.4 [a]
β-$NaOH \cdot 4H_2O$ (unstable)	− 1.7 to − 14.5 [a]
$NaOH \cdot 3.5H_2O$	+ 15.6 to − 8.8 [b]
$9NaOH \cdot 28H_2O$ (unstable)	+ 2.7 to − 2.5 [a]
$NaOH \cdot 2H_2O$	0 to + 12.5 [c]
$NaOH \cdot H_2O$	+ 4.5 to + 64.3 [b]

[a] Unstable melting point.
[b] Clearly defined melting point.
[c] Melting point can be determined only by extrapolation.

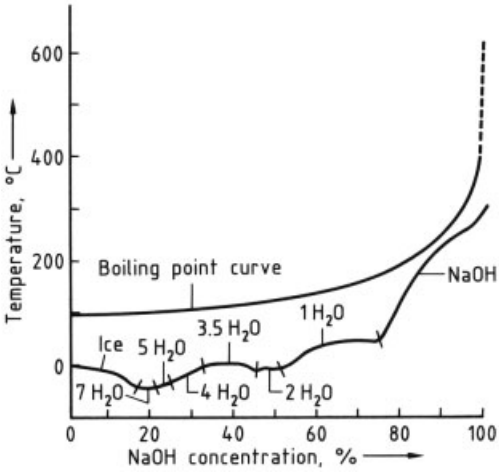

Figure 1. The NaOH – H_2O system

Anhydrous sodium hydroxide reacts very slowly with most substances. For example, it attacks many metals only slightly at room temperature (e.g., Fe, Mg, Ca, and Cd). However, corrosion rates increase rapidly with increasing temperature. More noble metals such as nickel, silver, gold, and platinum are attacked only slightly even when heated, especially if oxidizing atmospheres are excluded. Anhydrous sodium hydroxide does not react with dry carbon dioxide.

2.2. Properties of Sodium Hydroxide Solution

Physical Properties. The concentrations of the aqueous solutions in equilibrium with solid NaOH at various temperatures are given below:

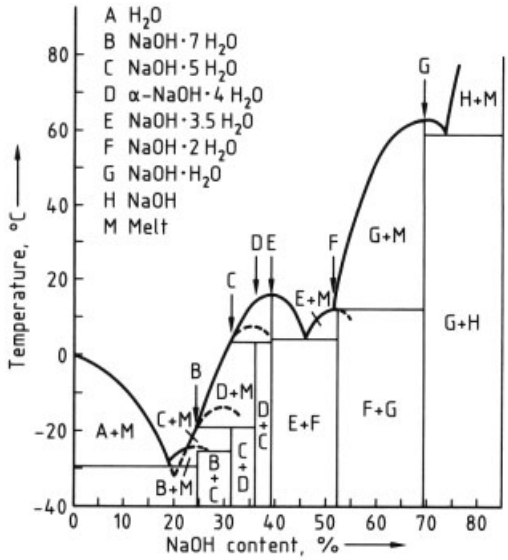

Figure 2. Melting diagram of the NaOH–H_2O system [1]

t, °C	10	18	30	40	53	64
wt % NaOH	49.8	51.0	53.3	55.4	59.7	69.0

The heat of solution is ca. 44 kJ/mol at 18 °C. The surface tension at 20 °C increases from 7.46×10^{-2} N/m for a 5 wt % solution to 0.1 N/m for a 35 wt % solution [2]. A 1 mol % solution has a surface tension of 7.43×10^{-2} N/m at 20 °C, which decreases to 6.23×10^{-2} N/m at 90 °C.

The density of aqueous sodium hydroxide solution in the concentration range 0 – 20 wt % can be calculated from the formula:

$$d_1 = d_2 + (1.16027 \times 10^{-2} - 2.511 \times 10^{-5} t + 1.0222 \times 10^{-7} t^2) p - (1.0817 \times 10^{-5} - 3.6748 \times 10^{-7} t + 2.034 \times 10^{-10} t^2) p^2$$

where

d_1, g/cm^3 = density of solution at t °C
t, °C = temperature

p, wt % = g NaOH/100 g solution
d_2, g/cm^3 = density of water at t °C

The densities of more concentrated aqueous sodium hydroxide solutions at 20 °C are as follows:

[NaOH], wt %	Density, g/cm^3
20	1.219
30	1.328
40	1.430
50	1.525

The boiling point curve is shown in Figure 1. In the concentration range 0 – 60 wt % NaOH, the boiling point increases approximately linearly with temperature.

NaOH concentration, wt %	5.9	23.1	33.8	48.3	54.6
Boiling point, °C	105	110	120	140	150

A 5 g/L solution of sodium hydroxide in water is almost completely dissociated. Activity coefficients are in the range $a = 0.784$ (0.1 M at 10 °C) and $a = 3.922$ (15 M at 70 °C).

Chemical Properties. Amphoteric metals such as zinc, aluminum, tin, and lead are attacked by dilute sodium hydroxide solution at room temperature. Iron, stainless steel, and nickel are fairly resistant.

3. Production

Sodium hydroxide is produced commercially in two forms: as 50 wt % solution (the most common form) and in the solid state (caustic soda) as prills, flakes, or cast shapes.

3.1. Production of Sodium Hydroxide Solution

Electrolysis of Sodium Chloride. Sodium hydroxide solution is produced industrially mainly by the electrolysis of sodium chloride. This yields sodium hydroxide solution, chlorine, and hydrogen in the mass ratio 1: 0.88: 0.025 in accordance with the following overall equation:

$$2\,NaCl + 2\,H_2O \rightarrow 2\,NaOH + Cl_2 + H_2$$

In the early 1980s, the membrane process was introduced, the other processes in operation at that time being the amalgam and diaphragm processes. Chlor-alkali production in 2004 was divided between the different processes as shown in Table 2.

Membrane technology continued to gain market share in 2004. Based on announced capacity changes, membrane processes will account for about 44 % of the total installed chlor-alkali capacity by 2009. Membrane technology saves approximately 30 % electrical power and is the best available technique, so it will be increasingly used in new units.

In Japan, by 1988 all plants based on the amalgam process had been either shut down or converted to the membrane process. Some Japanese diaphragm plants were also replaced by membrane plants for reasons of both cost and quality. Details of the three electrolytic processes are given in → Chlorine.

Treatment of the liquor from the electrolytic cell is shown schematically in Figure 3. In the amalgam process, sodium hydroxide solution is produced from sodium amalgam and water over a graphite catalyst at 80 – 120 °C. The 50 wt % solution produced is very pure. It is cooled, mercury is removed by centrifugation or filtration through cartridge or pressure plate filters, and it is then sold without further treatment.

In the diaphragm cell, the depleted sodium chloride solution from the anode compartment is transferred quantitatively to the cathode compartment, so that the cell liquor produced there contains ca. 18 wt % sodium chloride and 10 – 15 wt % sodium hydroxide. During evaporation to give a 50 wt % NaOH solution, all the impurities carried over with the depleted sodium chloride solution are removed to their limiting solubility by fractional crystallization; thus the sodium chloride content can only be reduced to ca. 1 – 1.5 wt %. In the DH process of the PPG Company, further purification is carried out by extraction with anhydrous liquid ammonia. This gives a product comparable to that from the amalgam process (rayon quality), but the process is seldom used because of cost. This is also true of a Japanese process in which a relatively pure sodium hydroxide solution is obtained by cooling the 50 wt % diaphragm liquor and recovering the crystals of pure $NaOH \cdot 3.5\,H_2O$ formed. Details of the evaporation and purification processes are given in → Chlorine. Methods of optimizing the various evaporation systems

Table 2. Distribution of chlor-alkali production among the three most widely used processes (percentage of installed capacity in 2004)

Process	North America	Western Europe	Japan
Amalgam	19	44	0
Diaphragm	39	7	3
Membrane	25	34	82
Others	17	15	15

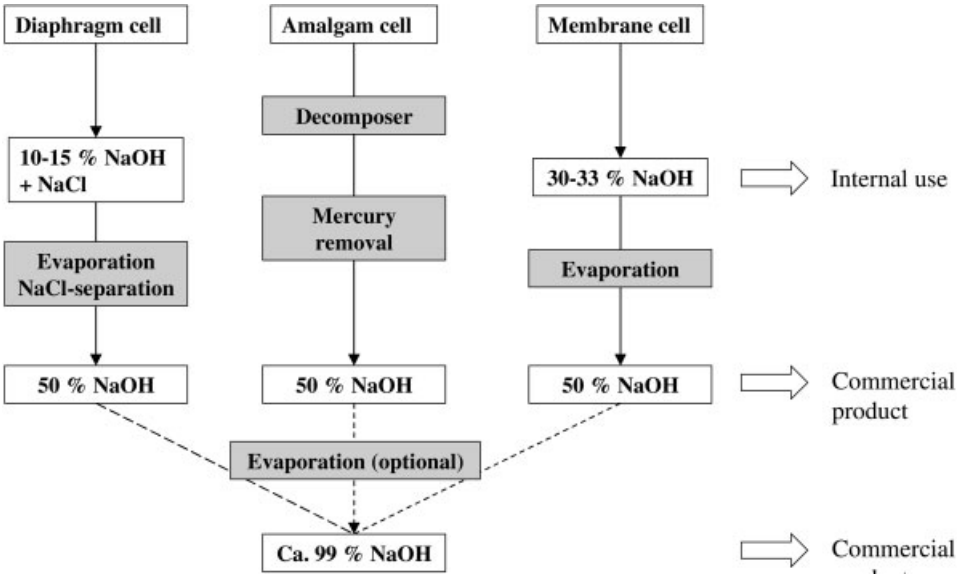

Figure 3. Treatment of the sodium hydroxide solution from electrolysis cells

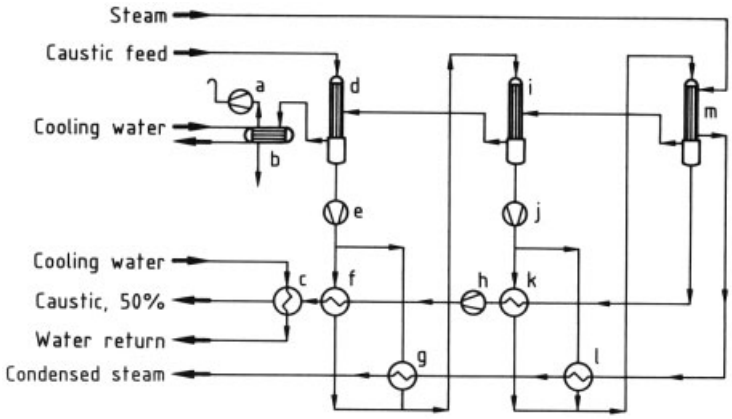

Figure 4. Membrane cell caustic evaporation (Bertrams process)
a) Vacuum pump; b) Condenser; c) Caustic cooler; d) First evaporator; e) Caustic pump; f) First caustic heat exchanger; g) Second heat exchanger; h) Caustic pump; i) Second evaporator; j) Caustic pump; k) Second caustic heat exchanger; l) First heat exchanger; m) Third evaporator

used for treating liquor from the diaphragm cell are described in [3].

In the membrane process, a hydraulically impermeable membrane prevents mixing of the electrolytes. This process produces a high-purity, virtually chloride-free sodium hydroxide solution comparable to that of the amalgam process. However, the chloride content increases by a factor of ca. 10 if operation of the cell is interrupted, due to diffusion of Cl^- ions through the membrane. The sulfate and chlorate present in the anolyte also affect the concentration of these impurities in the caustic liquor from the membrane cell, since the cation-selective membrane has an appreciable residual permeability for anions. Most membrane cells give optimum economic performance at a NaOH concentration of 30 – 35 wt %. An ion-exchange membrane has now been developed that enables a 50 wt % sodium hydroxide solution to be produced in the cell [4], although a considerable increase in voltage is required. Such membranes are viable only if steam costs are high and electricity costs low [5]. The 30 – 33 wt % sodium hydroxide solution is usually concentrated in a two- or three-stage falling-film evaporator (depending on steam costs and plant size) to give a commercial 50 wt % solution (see Fig. 4). For a three-stage evaporator, the steam consumption is 0.55 t per ton NaOH (100 %), and for a two-stage evaporator, 0.71 t per ton NaOH (100 %), with sodium hydroxide solution and steam in counter current flow.

The evaporation of membrane cell liquor (unlike that of diaphragm cell liquor) does not produce any solid material that can foul the heat exchanger or erode the pipework. The pickup of metals during evaporation is ca.1 wt ppm if appropriate construction materials are used (nickel, stainless steel). The various evaporation techniques are described in [3], with special reference to the concentration of membrane cell liquor.

As shown in Table 3, the three electrolytic processes have different energy requirements. The membrane process consumes the least electrical energy (2200 – 2500 kW · h per ton NaOH) but requires an additional 200–300 kW · h per ton NaOH in the form of steam to produce a 50 wt % solution. The diaphragm process consumes less electrical energy than the amalgam process but requires 700–900 kW · h per ton NaOH as steam for the evaporation stage and therefore has the highest total energy requirement.

Because of the high consumption of electrical energy, much research has been aimed at reducing the cell voltage. A ca. 1 V reduction can be achieved by using an oxygen diffusion cathode (as in a fuel cell) in the membrane cell process, although in this case no byproduct hydrogen is produced. This technology is gaining increasing interest. Cell design, oxygen diffusion cathode, and integration of the cathode into the cell design are priority points of development. Also the investigation of operation parameters and the periphery of the cells are in focus. Currently (in 2005) these cells have not been used on a production scale [7 – 9].

The electrolysis of fused NaCl at 350 °C with a β-Al_2O_3 diaphragm is also of interest. In laboratory-scale cells, a power consumption of 2350 kW · h per ton NaOH can be attained [10]. This technique, too, has yet to be used on a commercial scale.

Other Processes. The causticization of sodium carbonate solution is the oldest method of producing sodium hydroxide and was the only method available until the introduction of electrolysis. Although it had almost disappeared from use, it is now of special interest for companies with access to synthetic sodium carbonate from the Solvay process or to natural sodium carbonate (e.g., trona, $Na_3H(CO_3)_2 \cdot 2\,H_2O$). In this process, a hot, ca. 12 % solution of sodium carbonate is mixed with quicklime. The calcium carbonate that precipitates according to the equation.

$$Na_2CO_3 + CaO + H_2O \rightarrow 2\,NaOH + CaCO_3$$

is removed, and the ca. 12 % solution of sodium hydroxide is evaporated in several stages. The impurities that precipitate, mainly NaCl and Na_2SO_4, are filtered off. Further details on the causticization of sodium carbonate are given in [11].

In a combination of electrolytic and dialysis cells, sodium salts can be split into the corresponding acids and alkalis, e.g.,

$$Na_2SO_4 + 2\,H_2O \rightarrow 2\,NaOH + H_2SO_4$$

This reaction can be carried out in a two-chamber cell with a cation- or anion-exchange

Table 3. Energy requirements for the production of 1 t of sodium hydroxide (as 50 wt % solution) by the three electrolytic processes

Energy, kW · h/t NaOH [a]	Process		
	Amalgam	Diaphragm	Membrane
Electricity [b]	2800 – 3200	2500 – 2600	2200 – 2500
Steam (equivalent) [c]	0	700 – 900	90 – 180
Total	2800 – 3200	3200 – 3500	2290 – 2680
Relative energy costs, %	90	100	70

[a] The energy requirement given is based on the production of 1 t NaOH, 0.88 t Cl_2, and 0.025 t H_2.
[b] In dependence on the current density.
[c] 1 t steam is taken to be equivalent to 400 kWh of electrical energy.

membrane; in a three-chamber cell with a cation-exchange and an anion-exchange membrane; or in a multichamber cell with a combination of cation, anion, and bipolar membranes [12 – 15] (→ Sodium Sulfates, Chap. 1.4.1). However, in all industrially operated cells, the product is a dilute sodium hydroxide solution (15 – 30 wt %), which also contains some of the salt used as starting material. For this reason, this type of process is most suitable for treating sodium-salt containing wastewater from processes in which the dilute sodium hydroxide solution obtained can be recycled for neutralization purposes.

An alternative method of producing sodium hydroxide solution for small paper pulp plants, the ferrite recovery process, is described in [16]. Here, waste liquor containing sodium salts and organic substances is evaporated, and the residue is mixed with Fe_2O_3 and calcined. The sodium ferrite formed is decomposed by water to give NaOH and Fe_2O_3.

3.2. Production of Solid Sodium Hydroxide

Solid sodium hydroxide (caustic soda) is obtained by evaporating sodium hydroxide solution until the water content is $< 0.5 – 1.5$ wt %. The most efficient utilization of energy is achieved with multistage equipment. A flow diagram of a plant constructed by Bertrams, Basel (Switzerland), is shown in Figure 5.

A 50 wt % solution is vacuum evaporated in a preconcentration unit, by using the heat of the vapor from the main concentrator, until the concentration reaches ca. 60 wt %. In the second stage, a concentration of ca. 99 wt % is attained by heating with a molten salt heat-transfer medium ($NaNO_2 – NaNO_3 – KNO_3$) at $>$ 400 °C in a falling-film evaporator. The product is fed to a flash evaporator that operates as a gaslift pump. Molten NaOH is raised by vacuum into a heated riser pipe, causing the remaining water to evaporate. The rising steam bubbles lift the melt to the highest point of the system, where the steam is pumped off. The almost anhydrous melt passes through a downcomer pipe into the holding vessel, from which it is pumped to the conditioning plant. The holding vessel, pumps, and pipes must be heated to prevent premature crystallization.

3.3. Forming

When the NaOH melt is cooled and formed, both the solidification point of NaOH (322 °C) and that of the monohydrate, NaOH · H_2O (62 °C), must be passed through quickly to prevent caking due to unsolidified monohydrate.

Solid sodium hydroxide is supplied in the form of flakes, prills, cast blocks, and less commonly as tablets, briquettes, or granules. Flakes have a bulk density of ca. 0.9 kg/dm^3, a thickness of ca. 1.5 mm, and a diameter of 5 – 20 mm, and are easily broken so that some dust is always formed on handling. Prills have a bulk density of $>$ 1 kg/dm^3 and consist of spheres with a diameter of 0.1 – 0.8 mm (microprills) or 0.5 – 2.5 mm.

Flakes. To produce flakes, molten NaOH is fed into a trough in which an internally cooled drum rotates. The molten material solidifies on the surface of the partially immersed drum and is scraped off by a knife, which breaks up the sheet of NaOH into flakes. These are taken to a small intermediate storage silo and packed as soon as possible in sacks or steel drums. The material

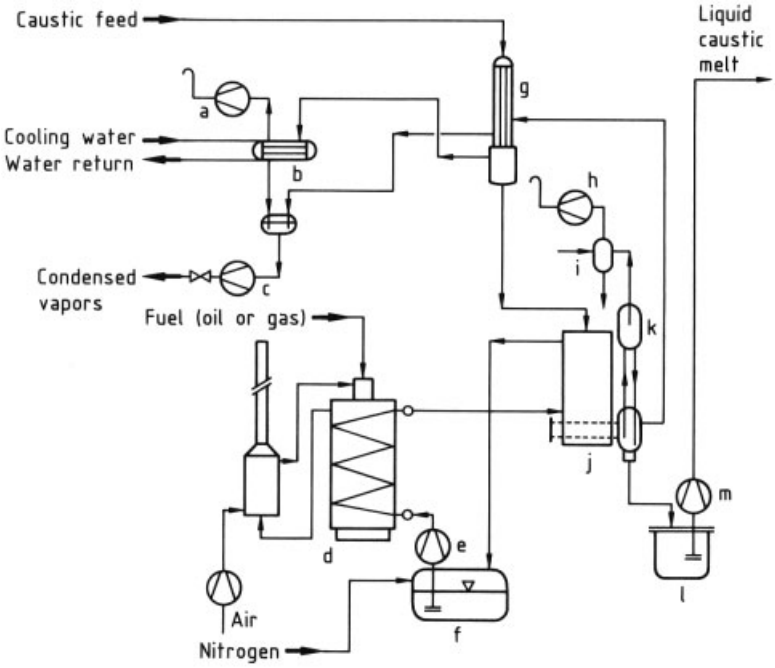

Figure 5. High-Concentration unit for caustic soda (Bertrams process)
a) Vacuum pump; b) Condenser; c) Condensate pump; d) Furnace unit; e) Salt melt pump; f) Heated salt melt tank; g) Preconcentration unit; h) Vacuum pump; i) Condenser; j) Falling-film concentrator; k) Flash evaporator; l) Heated caustic melt tank; m) Caustic melt pump

tends to bridge due to its particle shape, so that it cannot be stored in large silos or be transported by compressed air.

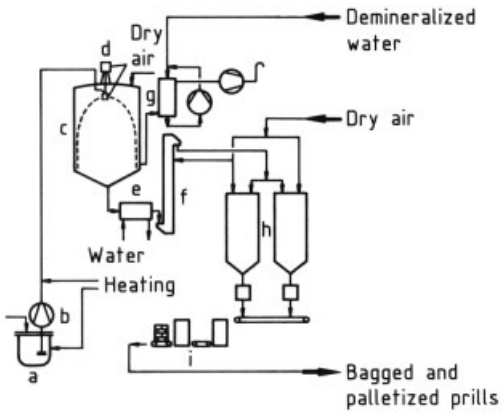

Figure 6. Production of sodium hydroxide prills (Bertrams process)
a) Caustic melt tank; b) Caustic melt pump; c) Spray tower; d) Rotating spray system; e) Product cooler; f) Elevator; g) Exhaust air treatment; h) Silo; i) Bagging and pelletizing

Prills are produced by spraying the molten material at ca. 360 °C to form droplets. The spray equipment used can be nozzles (outdated), or a spinning disk or basket. The finely divided melt forms spheres of fairly uniform diameter due to surface tension. These are cooled to ca. 250 °C as they fall down a shaft in which air flows upward. The prills are collected in a funnel from which they run into a cooling drum where they are cooled to ca. 50 °C. The finished product is stored in silos (see Fig. 6).

Unlike flakes, prills are free flowing. Also, they do not cake as readily and thus can be stored in silos, transported in containers or other large vessels, or moved by compressed air, because they are almost dust free.

Cast Blocks. Molten NaOH is directly cast into iron drums or similar containers. The drums are externally cooled intensively by spraying with water, so that a solid film of sodium hydroxide forms immediately on the walls of the container, preventing further contamination by the iron.

Special Shapes. Tablets, pills, or pellets are produced by the droplet process. Molten NaOH drops from a special device made of silver onto a slowly rotating cooled table or conveyor belt (pill machine). The product has a very low content of metallic impurities, so that this process is especially suitable for producing small quantities (e.g., for analytical purposes).

3.4. Specifications

A review of the grades of 50 wt % sodium hydroxide solution produced by the various processes is given in → Chlorine. Diaphragm liquor cannot be used in all applications (e.g., for viscose manufacture or various organic processes) because of the amounts of chloride, chlorate, and heavy metals present.

When corrosion-resistant materials are used for sodium hydroxide production (nickel and stainless steel), the pickup of metal (Ni, Fe) is < 1 ppm. Solid caustic soda should have a pure white appearance. The most important impurities in solid sodium hydroxide are listed in Table 4.

3.5. Storage, Packaging, and Transportation

Iron, stainless steel, and nickel are suitable materials for containers used for transport and storage, as well as for pipelines. However, corrosion resistance is not the sole criterion for the choice of material. Stress corrosion cracking, changes in the crystal structure, and pitting can cause sudden material failure, so that a more expensive material is often chosen for safety reasons.

Vessels for holding aqueous solutions of sodium hydroxide are usually constructed of alkali-resistant steel. However, at > 50 °C or in the presence of oxygen, iron is slowly attacked by sodium hydroxide solution. For a high-quality product (e.g., with iron content $<$ 1 mg/kg), cladding (rubber or plastic coating) is required. At < 50 °C the solution can be transported by pipes of alkali-resistant steel or plastic (e.g., PVC). Iron pipes are always kept full to prevent corrosion due to ingress of air and consequent black coloration of the solution. At higher temperature, coated materials such as rubber-lined steel, austenitic Cr – Ni steel (e.g., DIN 1.4541), nickel, Hypalon, epoxy resin, polypropylene, or glass fiber reinforced composites are suitable. In the latter case, an inner lining must be used to prevent contact between the glass fiber and the solution. Heat exchangers are usually constructed of nickel, nickel alloys, or stainless steel because of the high temperature. For equipment handling highly concentrated solutions, nickel is the most commonly used material because of its outstanding corrosion resistance. Metals that have been in contact with solutions are difficult to weld, so that repairs to pipework and equipment which have carried sodium hydroxide solutions are difficult. A comprehensive review of materials that can be used with sodium hydroxide is given in [17].

Pipelines carrying 50 wt % sodium hydroxide solution in cold regions are provided with thermostatic heating. Solutions are transported by road, rail tanker, or ship.

Caustic soda prills are stored in steel or stainless steel hoppers with a capacity of 30 – 50 m^3. The hoppers have conical bases with an angle of 25° to the vertical and are provided with external hammering equipment. Even when the product is stored under dry air, a storage time of 3 – 6 d should not be exceeded to ensure that the prills do not cake under their own weight. Solid sodium hydroxide is usually packed in 25 – 50-kg hermetically sealed polyethylene sacks. For transport, the sacks are stacked on wooden palettes and shrink wrapped in polyethylene. Caustic soda prills are also carried in 50- or 200-kg steel drums, loose in 1000-kg-capacity steel or flexible containers, or in transportable hoppers. In the closed system delivery (CSD) method of the PPG Company, containers coated with epoxy resin are transported by road and rail. Pneumatic transport is carried out with dry air and causes very little damage to the product [18].

4. Uses

Most users of sodium hydroxide require dilute aqueous solutions.

The sodium hydroxide produced in Europe (9.86×10^6 t/a NaOH) and in North America in 2004 was divided among various areas of use,

Table 4. Product quality: typical percentage analyses of sodium hydroxide from chlor-alkali electrolytic processes

	Amalgam process	Membrane process	Diaphragm process
NaOH	99 – 99.5	99 – 99.5	98 – 99
NaCl	0.03 – 0.05	0.01 – 0.02	1 – 2
Na_2CO_3	0.2 – 0.4	0.2 – 0.4	0.65
Na_2SO_4	0.03	0.001 – 0.01	0.035
Ca	0.002	0.00002	0.001
Mg	0.0005	0.00002	0.0004
Fe	0.001	0.0001 – 0.001	0.0006 – 0.0015
Al	0.001	0.00002	0.0005
Ni	0.0002 – 0.0004	0.0002 – 0.0004	0.0002 – 0.0004
Hg	0.000001 – 0.00001	Hg free	Hg free
mp, °C	322	322	322

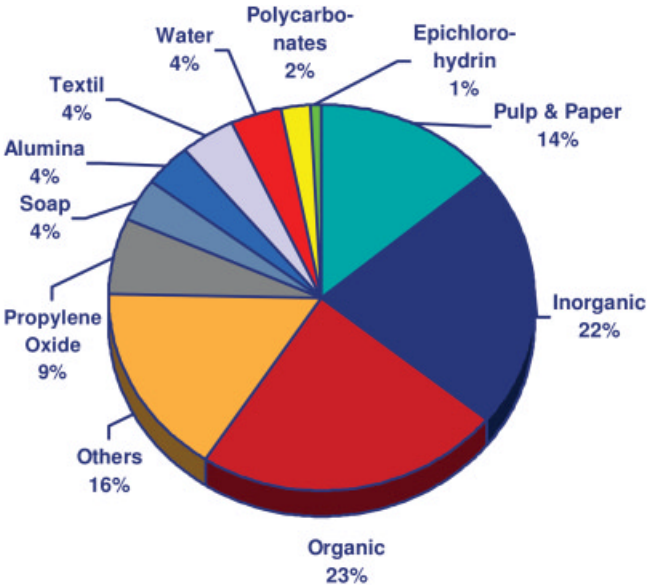

Figure 7. Demand of sodium hydroxide in Europe in 2004 broken down by major end uses (Source Bayer MaterialScience AG)

as shown in Figure 7 and 8. More than 56 % of production is used in the chemical industry:

1) In inorganic chemistry, sodium hydroxide is used in the manufacture of sodium salts, for alkaline ore digestion, and for pH regulation.
2) The organic chemical industry uses sodium hydroxide for saponification reactions, production of nucleophilic anionic intermediates, etherification and esterification, basic catalysis, and the production of free organic bases. Sodium hydroxide solution is used for scrubbing waste gases and neutralizing wastewater.
3) In the paper industry (→ Paper and Pulp, Chap. 1.3.1,→ Paper and Pulp, Chap. 1.3.2), sodium hydroxide solution is used for cooking wood (removal of lignin).
4) The textile industry uses sodium hydroxide solution to manufacture viscose and viscose staple fibers (→ Cellulose, Chap. 3.1). The sodium hydroxide solution used must contain only traces of chloride ions (rayon quality). The surface of cotton can be improved by treatment with sodium hydroxide solution (mercerization).
5) Considerable quantities of sodium hydroxide are used for sodium phosphate production in the detergent industry. Soaps are manufactured by the saponification of fats and oils with sodium hydroxide solution, and detergents are produced from organic sulfonic acids and sodium hydroxide.
6) In the aluminum industry, sodium hydroxide is used mainly for the treatment of bauxite.

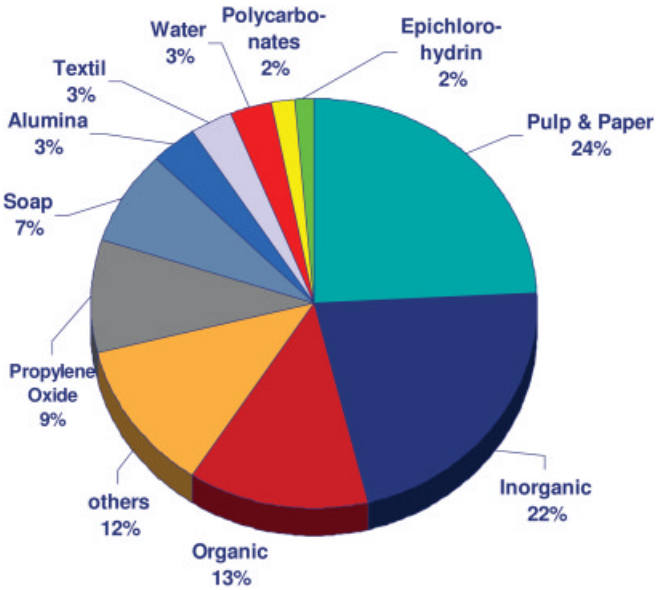

Figure 8. Demand of sodium hydroxide in North America in 2004 broken down by major end uses (Source Bayer MaterialScience AG)

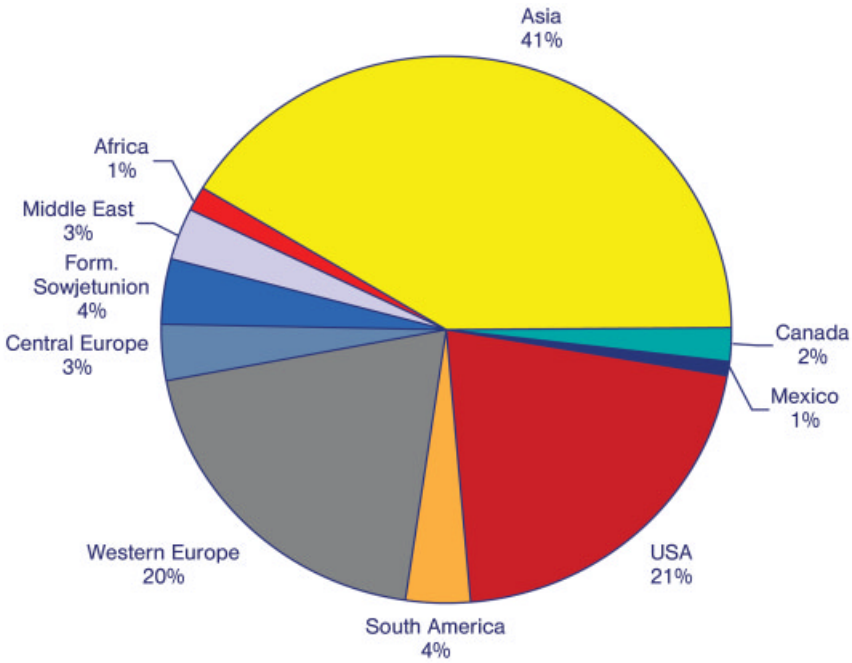

Figure 9. World production capacity of sodium hydroxide by region (total: 59 557 000 t in 2004) (Source Bayer MaterialScience AG)

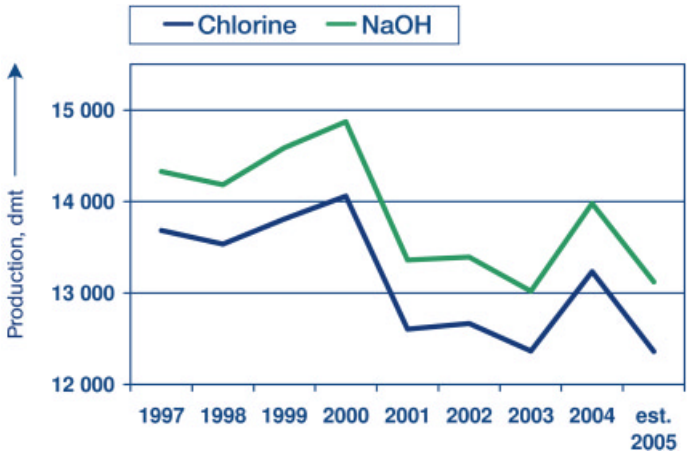

Figure 10. U.S. production of chlorine and sodium hydroxide solution (Source Bayer MaterialScience AG)

7) Waterworks use dilute sodium hydroxide solution to regenerate ion exchangers for water purification and wastewater treatment.
8) Other users of sodium hydroxide include electroplating technology, the natural gas and petroleum industries, the glass and steel industries, and gold extraction (cyanide leaching). In the food industry, sodium hydroxide is used for degreasing, cleaning, and for peeling potatoes.

Consumption of caustic soda in pulp and paper accounts for about one quarter of global caustic demand. The pulp and paper markets are relatively mature markets, so demand growth will be rather slow. The remaining 75 % of global NaOH demand is inclined to follow the health of the local manufacturing base, which is tied to the local gross domestic product (GDP).

5. Safety Precautions

Because of its strongly caustic action, all contact with sodium hydroxide during handling must be prevented by the use of suitable protective equipment (protective goggles, safety gloves, and, if needed, dust masks). Any clothing or shoes that come in contact with sodium hydroxide must be removed immediately and can be reused only after thorough cleaning. First aid to affected parts of the body consists of washing with copious amounts of water. The skin can be neutralized with a dilute solution of a buffered weak acid (e.g., acetic acid – sodium acetate). For irrigation of the eyes, a special hygienic eyewash solution is used. The product poses no direct environmental risk and has therefore been removed from the list of chemicals (e.g., in the United States) whose emission must be reported to local and national environmental authorities.

6. Economic Aspects

Chlor – alkali plants are operated to meet chlorine demand; therefore outputs are typically connected to chlorine production. Chlor – alkali operating rates will be on a steady climb, as global economics growth is stable.

Global chlorine demand is primarily driven by production of commodities such as PVC, used, for example in construction parts, and isocyanates, going into construction parts, furniture, shoe industry, consumer durables, and automobiles. These markets follow the local GDP trend, thus more mature economies will typically have a lower GDP and a lower growth in chlorine demand for these end uses.

The world production capacity of sodium hydroxide in 2004 is estimated to be approx. 60×10^6 dry metric tons (dmt) per year. Today 41 % of sodium hydroxide is produced in Asia (Fig. 9). In contrast the chlor – alkali production in the US has shown a decreasing tendency in recent years due to increasing energy costs and a smaller share in production of vinyl compounds (Fig. 10). Even though the same trend holds true for Western Europe, the production remains on a constant level. From 2004 to 2009 the global

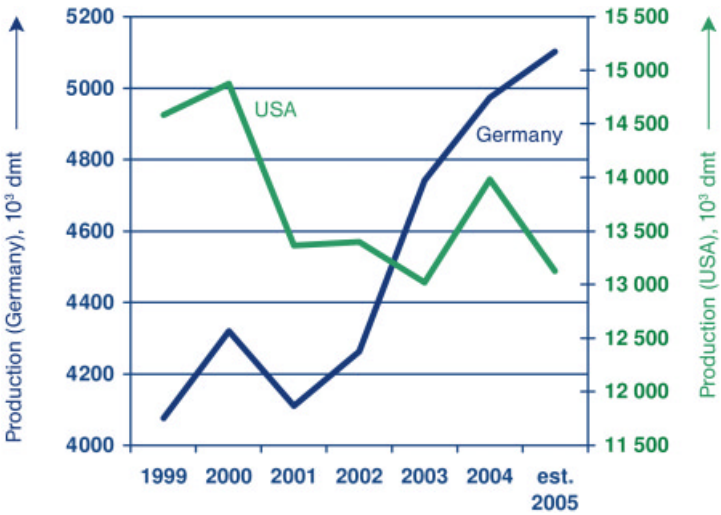

Figure 11. Sodium hydroxide production in Germany and the United States, 1999 – 2005 (Source Bayer MaterialScience AG)

market for caustic soda will increase by 2.8 % on average (demand in 2004 is estimated to be approx. 51×10^6 t, 100 %). Growing markets are textiles, water treatment, alumina production, and pulp and paper.

In order to show the different growth rates in the USA and Germany, the development of the US market is shown in Figure 10. and a comparison of the two markets in Figure 11.

7. References

1. S. U. Pickerring, *J. Chem. Soc.* **63** (1893) 890.
2. *Gmelin*, 8th ed., System no. 21.
3. S. A. McCluney, J. W. Van Zee, *J. Electrochem. Soc.* **136** (1989) 2556 – 2564.
4. C. B. Kelly, I. L. Walker, Staff of the Chlorine Institute, Inc. (eds.): *The Second Chlor-Alkali Symposium, World Chlor-Alkali: An Industry in Transition, Proceedings*, Washington, DC 1990.
5. D. C. Brandt, paper presented at the Chlorine Institute Meeting, Washington, DC, April 1989.
6. R. Winkler, *Chem. Ind.* **36** (1984) 152 – 155.
7. D. Hoormann, J. Jörissen, H. Pütter, *Chem. Ing. Tech.* **77** (2005), no. 9, 1363 – 1376.
8. Bayer MaterialScience AG, DE 101 48 599, 2001.
9. Toagosei, Mitsui Chem., Kaneka, Chlorine Engineers, EP 104 1176, 1998.
10. M. Kamaludeen et al., *Transactions of the SAEST* **21** (1986) no. 1, 53 – 56.
11. *Ullmann*, 3rd ed., **12**, pp. 659 – 660.
12. S. Sridhar, *Chem. Ing. Tech.* **61** (1989) 428 – 429.
13. H. v. Plessen et al., *Chem. Ing. Tech.* **61** (1989) 933 –940.
14. G. Kreysa, *Chem. Ing. Tech.* **62** (1990) 357 – 365.
15. J. Jörissen, K. H. Simmrock, *J. Appl. Electrochem.* **21** (1991) 869 – 876.
16. G. Venkoba, Rao, *IPPTA* **24** (1987) no. 3, 20 – 29.
17. I. K. Nelson in B. J. Moniz, W. I. Pollock (eds.): *Process Industries Corrosion*, NACE, Houston, Tex. 1986.
18. CSD, PPG Industries, Company Brochure, Pittsburgh, Pa.

Sodium Sulfates

HELMOLD VON PLESSEN, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Sodium Sulfate 1
1.1. History and Natural Occurrence . . 1
1.2. Properties 2
1.3. Extraction and Production 4
1.3.1. Crystallization and Calcination 4
1.3.2. Production as a Byproduct of the Chemical Industry 5
1.3.3. Production from Natural Sources . . . 6
1.3.4. Production of Sodium Sulfate 8
1.4. Exploitation of Byproduct Sodium Sulfate 9
1.4.1. Electrochemical Decomposition of Sodium Sulfate 9
1.4.2. Conversion Processes 11
1.4.3. Reduction Processes 11
1.5. Uses 11
1.6. Quality Specifications and Analysis 11
1.7. Occupational Health and Environmental Protection 12
1.8. Economic Aspects 12
2. Sodium Hydrogensulfate 12
3. References 13

1. Sodium Sulfate

1.1. History and Natural Occurrence

Sodium sulfate has been known since the 1500 s. Its use in the form of spa water or as a salt goes back even further. It was first described in 1658 by J. R. GLAUBER. It was known as sal mirabile Glauberi [1], was prepared from common salt and sulfuric acid, and was used medicinally as a laxative. Many kinds of spa water and the salts obtained from them contain sodium sulfate (e.g., Karlsbad salt, which contains 44 % sodium sulfate) [2].

Processes for the manufacture of soda and hydrochloric acid that developed in the 1800 s involved the production of sodium sulfate (salt cake) [3] (see Section 1.3.4). This method has now declined in importance compared to extraction from natural deposits and production as a byproduct of chemical processes. Byproduct sodium sulfate must be removed from wastewater for reasons of environmental protection, which allows largely closed recirculating systems to be established (e.g., for spinning baths in the viscose-fiber industry [4], [5]).

Sodium sulfate not only occurs in salt deposits in ancient geological formations but can also be produced on an industrial scale from continuously formed reserves in salt lakes in Canada, the United States, South America, the Commonwealth of Independent States, and other countries (see Section 1.3.3).

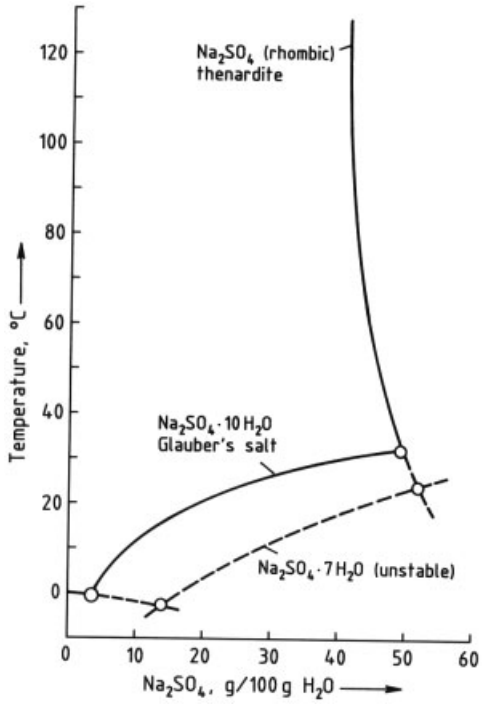

Figure 1. Solubility of sodium sulfate

Anhydrous sodium sulfate occurs naturally as thenardite, sometimes in high purity, and Glauber's salt occurs as mirabilite, $Na_2SO_4 \cdot 10\ H_2O$. In marine salt deposits and in the crystalline deposits that are being produced in salt lakes at the present time, many double salts containing

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_355

Table 1. Properties of thenardite and Glauber's salt

Property	Thenardite Na_2SO_4	Glauber's salt $Na_2SO_4 \cdot 10\,H_2O$
M_r	142.040	322.190
Water of crystallization, %	0	55.914
mp, °C	884.0 (decomp. > 884 °C)	"melts" at 32.384 °C in its own water of crystallization (decomp. > 32.384 °C)
Crystal habit	rhombic	monoclinic
Unit cell dimensions	0.586, 1.23, 0.982	1.143, 1.034, 1.290 (angle = 107 ° 40′)
No. of molecules in unit cell	8	4
Refractive indices n_D	1.464, 1.474, 1.485	1.394, 1.396, 1.398
Density, g/cm^3	2.697	1.468
Specific heat capacity c_p at 300 K, J kg^{-1} K^{-1}	899.2	1825.7
Heat of fusion, kJ/kg	147.98	214.52
Heat of solution at 18 °C, kJ/kg	16.2	− 243.8
Enthalpy of formation, kJ/mol	− 1385	− 4327
Entropy $S°$, Jmol^{-1} K^{-1}	149.60	588.2
Dielectric constant (at 10^3 Hz)	7.80	7.90
Mohs hardness	2.7	1.5 – 2

sodium sulfate occur such as astrakhanite, $Na_2SO_4 \cdot MgSO_4 \cdot 4\,H_2O$; glaserite, $Na_2SO_4 \cdot 3\,K_2SO_4$; glauberite, $Na_2SO_4 \cdot CaSO_4$; loeweite, $6\,Na_2SO_4 \cdot 7\,MgSO_4 \cdot 15\,H_2O$; d'Ansite, $MgSO_4 \cdot 9\,Na_2SO_4 \cdot 3\,NaCl$; vanthoffite, $MgSO_4 \cdot 3\,Na_2SO_4$; burkeite $Na_2CO_3 \cdot 2\,Na_2SO_4$; and hanksite, $KCl \cdot 2\,Na_2CO_3 \cdot 9\,Na_2SO_4$. Hence, the industrial production of sodium sulfate can be combined with the production of common salt, soda, or potash salts [6–10].

1.2. Properties

Physical Properties. Table 1 compares the properties of thenardite (anhydrous sodium sulfate) and Glauber's salt (the decahydrate). The heptahydrate, $Na_2SO_4 \cdot 7\,H_2O$, is unstable (Fig. 1) and can be detected on cooling saturated solutions to < 12 °C [11–13].

Table 2 lists the solubility of sodium sulfate in water [11], [12], [14–16]. Sodium sulfate is polymorphic; above 240 °C, four other modifications are formed from thenardite (Table 2) [14]. In the sodium sulfate – water system, a transformation occurs at a temperature that has been determined with great accuracy (32.384 °C) and is suitable as a fixed point. Above this temperature, anhydrous thenardite crystallizes from saturated solutions, and below it, Glauber's salt (Fig. 1).

The decrease in solubility of sodium sulfate resulting from the presence of sodium chloride can be utilized in the industrial production of sodium sulfate from aqueous solutions. This effect is shown by the three-component system sodium sulfate – sodium chloride – water (Table 3) [2], [17], [18].

The system sodium sulfate – sulfuric acid – water shows first an increase and then a decrease in the solubility of sodium sulfate with increasing sulfuric acid concentration [2], [15], [18]. The reaction to form sodium hydrogensulfate is of industrial importance (see Chap. 2).

The low solubility of sodium sulfate in organic solvents or mixtures of these with water can be used in the precipitation crystallization of thenardite [5], [15], [19] (Section 1.3.1).

Chemical Properties. In moist air, anhydrous sodium sulfate takes up water of crystallization with a considerable increase in volume, which reaches a factor of 4.17 on complete conversion to the decahydrate (Glauber's salt) and can lead to damage if the change occurs in cracks in buildings. Buildings constructed with cement, especially those made of concrete, can be destroyed by solutions containing sodium sulfate. Glauber's salt is not hygroscopic. It effloresces and is converted to an almost anhydrous sulfate after a few days in a cold current of air.

On heating below its melting point (884 °C), thenardite decomposes slowly. Above 1200 °C, the weight loss of the molten material can be several percent. If sodium sulfate is added to molten glass, sodium silicates are formed, and

Table 2. Solubility of sodium sulfate in water

Temperature, °C	g Na_2SO_4 in		Temperature, °C	g Na_2SO_4 in	
	100 g H_2O	100 g Solution		100 g H_2O	100 g Solution
−1.2	3.9 a∗	3.8	100	42.2 c	29.7
0	4.5 b	4.3	110	42.0 c	29.6
5	6.2 b	5.8	150	42.2 c	29.7
10	9.0 b	8.3	200	44.1 c	30.6
15	13.2 b	11.7	240	44.9 c, d	31.0
18	16.4 b	14.1	250	44.0 d	30.5
20	19.0 b	16.0	260	41.8 d	29.5
22	22.1 b	18.1	270	38.8 d	28.0
24	25.8 b	20.5	280	35.2 d	26.0
25	27.8 b	21.8	290	30.8 d	23.5
30	41.2 b	29.2	300	24.8 d	19.9
32.4	49.7 b, c	33.2	310	18.5 d	15.6
35	49.1 c	32.9	320	13.1 d	11.6
40	48.1 c	32.5	330	7.3 d	6.8
50	46.4 c	31.7	340	4.2 d	4.0
60	45.3 c	31.2	350	2.4 d	2.3
70	44.2 c	30.7	360	0.9 d	0.9
80	43.2 c	30.2	382	0.38 d	0.38
90	42.6 c	29.9			

∗ Solid phases: a = $Na_2SO_4 \cdot 10\,H_2O$ + ice; b = $Na_2SO_4 \cdot 10\,H_2O$; c = Na_2SO_4, rhombic (thenardite); d = γ- Na_2SO_4, hexagonal (transforms into rhombic β-Na_2SO_4 on cooling below 240 °C, and into monoclinic α-Na_2SO_4 on cooling below 180 °C: this is unstable in aqueous solution and transforms into thenardite).

sulfur dioxide and oxygen are released (see Table 6).

Table 3. Solubility (%) of sodium sulfate in water and sodium chloride solutions

Temperature, °C	NaCl content of solution, %				
	0	10	14	17	20
10	8.25 a∗	3.9 a	3.3 a	3.2 a	3.3 a
15	11.7 a	6.4 a	5.5 a	5.2 a	5.5 a
17.5	13.7 a	8.1 a	7.4 a	6.8 a	7.1 c
25	21.8 a	15.7 a	14.9 a, b∗∗	11.8 b	9.2 c
30	29.9 a	19.2 a, b∗∗	14.2 b	10.8 b	8.5 c
33	33.2 a, b	18.9 b	14.2 b	11.2 b	8.9 c
50	31.7 b	18.3 b	13.7 b	10.3 b	6.7 c
75	30.4 b	17.3 b	13.0 b	9.9 b	8.1 c
100	29.7 b	16.7 b	12.8 b	10.0 b	7.8 c

∗ Solid phases: a = $Na_2SO_4 \cdot 10\,H_2O$; b = Na_2SO_4 (thenardite); c = mixtures with NaCl.
∗∗ The transformation temperature of 32.384 °C is depressed to 17.9 °C by NaCl.

1.3. Extraction and Production

Of the world production of sodium sulfate (ca. 4.6×10^6 t in 1988), approximately 50 % is a byproduct of the chemical industry, the remainder being extracted from natural deposits. World reserves of sodium sulfate are estimated at 3400×10^6 t [20], [21].

The extraction of sodium sulfate from wastewater from the chemical industry has a beneficial effect on the environment and has the further advantage that it is independent of weather conditions, in contrast to production from natural sources [5], [20]. Furthermore, the production plant is usually close to the consumer of the sodium sulfate, whereas production from natural deposits often takes place in remote regions.

Glauber's salt is not a commercial product, owing to its high content of water of crystallization (56 %) and consequently high transportation costs, its low melting point (32.4 °C), and its tendency to effloresce. However, it is an important intermediate product in many processes.

1.3.1. Crystallization and Calcination

The principle of the crystallization process is to cool the solution to < 20 °C, causing Glauber's salt, $Na_2SO_4 \cdot 10\,H_2O$, to crystallize [2], [22]. Multistage vacuum-cooled crystallizers are used, usually of the horizontal type.

Glauber's salt is recovered and calcined to form anhydrous sodium sulfate (thenardite) in the next stage. Glauber's salt "melts" in its own water of crystallization at 32.38 °C, forming a saturated solution of sodium sulfate that contains anhydrous sodium sulfate as a solid phase (37 % of the sodium sulfate from the Glauber's salt). In older processes, the saturated sodium sulfate solution, after removal of the precipitated sodium sulfate, was once again vacuum cooled, and the Glauber's salt that crystallized was again recovered and melted. If Glauber's salt was again crystallized from the residual mother liquor, up to 85 % of the total Glauber's salt could be recovered as anhydrous sulfate.

This energy-intensive process was superseded by evaporative crystallization of the saturated sodium sulfate solution obtained on melting Glauber's salt [23–28]. The energy consumption was reduced further by using multistage evaporation, in which the vapor was employed as heating steam for the following stage, which was operated at lower pressure. Further savings resulted from the use of mechanical or thermal compression of the vapor [29–31].

The formation of encrustations in the heat exchangers presented a problem, especially in the heat exchangers of the Glauber's salt melting vessels because of the decreasing solubility of sodium sulfate above 32 °C [2], [32]. For this reason, unheated melting vessels were often used, in which Glauber's salt was melted by mixing with approximately 15 times the amount of hot, saturated sodium sulfate solution from the evaporation stages [26].

The formation of encrustations can also be prevented if the heating medium for the melting vessel has such a low temperature (e.g., vapor from the last evaporation stage) that the temperature increase of the circulating melt solution in the heat exchanger is limited to 1 – 3 °C per pass [33].

In another process, part of the sodium sulfate solution is evaporated at 108 – 114 °C and pressures of up to 0.15 MPa absolute [34]. In this temperature range, the solubility curve of sodium sulfate exhibits a minimum, so that crust formation in the heat exchanger of the evaporator is greatly reduced. The vapor from this first stage is used as the heating medium for the evaporator of the second stage, which operates with vapor compression.

The purity of the sodium sulfate produced can be controlled by varying the amount of mother liquor removed at the calcination stage. Here, the saturation concentrations of the dissolved impurities must not be exceeded [35]. These processes can then yield a very pure sodium sulfate (> 99.9 %). The mother liquor can be fed to a biological wastewater treatment plant or worked up to yield Glauber's salt [36].

Figure 2 shows a flow diagram of a plant for the production of sodium sulfate from wastewater [5].

The formation of encrustations during the thermal dehydration of Glauber's salt can be largely prevented by the use of air, spray, or fluidized-bed dryers, which have no solid heat-exchange surfaces. For example, in the Holland evaporator, predried Glauber's salt is heated by combustion gases at 870 – 980 °C in a rotary dryer to produce thenardite. If the Glauber's salt is melted and thenardite is separated, the mother liquor can be spray-dried with hot combustion gases. The specific energy consumption can be reduced if molten Glauber's salt in the form of a crystal slurry is sprayed into a rotary dryer heated to 150 – 400 °C. In fluidized-bed drying, thenardite is produced with hot air at 450 °C, giving a waste gas temperature of 80 – 100 °C. The fines are recirculated, producing a dust-free sodium sulfate with a grain size that can be controlled up to 4.8 mm.

In these processes, the entire water of crystallization is evaporated, so that the impurities in Glauber's salt remain in the sodium sulfate [2], [37–39].

Sodium sulfate can be precipitated as thenardite from aqueous solutions by adding water-soluble organic solvents such as methanol [2], [40–43]. This process is used in the treatment of solutions containing Na_2SO_4 and sulfuric acid from the cellulose industry [44], [45].

Sodium sulfate can also be obtained from Glauber's salt by using submerged combustion heaters. However, the burner must be cleaned or changed frequently due to corrosion and encrustation [2], [20].

Sodium sulfate can be separated from mixtures of solid salts by flotation. This process is being investigated for the treatment of salt mixtures from natural deposits [46]

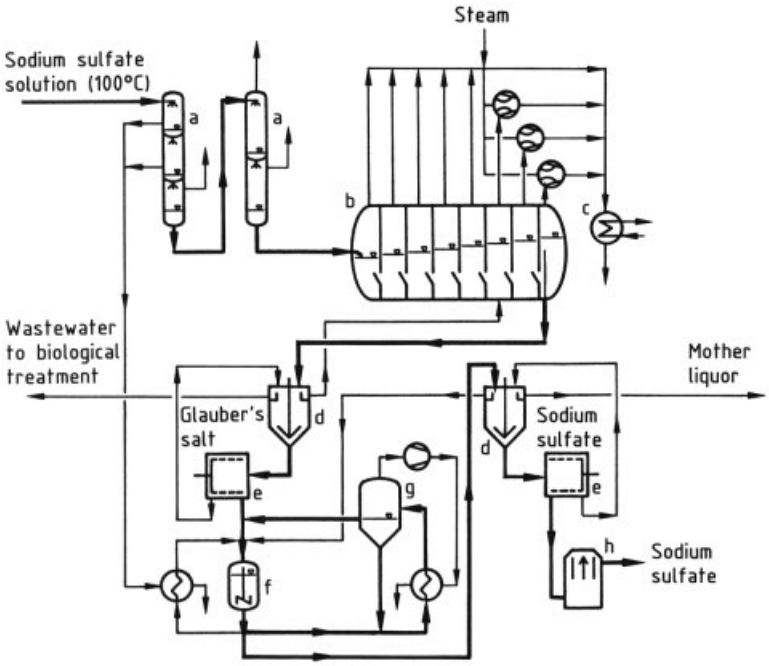

Figure 2. Production of sodium sulfate by crystallization and calcination
a) Evaporative cooler; b) Vacuum cooling crystallizer; c) Cooler; d) Thickener; e) Centrifuge; f) Melting vessel; g) Evaporator; h) Dryer

1.3.2. Production as a Byproduct of the Chemical Industry

Solutions of sodium sulfate are produced in numerous reactions of sodium compounds with sulfuric acid or with sulfur dioxide and oxygen. These solutions must be treated to prevent water pollution. Sodium sulfate is obtained from them by crystallization and calcination. The quantities of sodium sulfate produced as a byproduct from various sources depend very much on the state of the market for the main products and the state of the technology. In spite of this, total production of byproduct sodium sulfate has hardly changed since 1980 [20].

From Viscose-Fiber Spinning Baths. Fibers are produced by spinning viscose in a sulfuric acid precipitation bath in which the following reaction takes place:

$$2\,\text{Cell}-\text{OCS}_2\text{Na} + \text{H}_2\text{SO}_4 \longrightarrow 2\,\text{Cell}-\text{OH} + 2\,\text{CS}_2 + \text{Na}_2\text{SO}_4$$

Cell = cellulose

The precipitation bath contains ca. 10 % H_2SO_4, 20 % Na_2SO_4, and 1 % $ZnSO_4$. The spinning process produces ca. 1 kg of sodium sulfate for each kilogram of viscose fiber. Sulfuric acid must therefore be added continuously to the bath, and sodium sulfate removed [2], [4].

The treatment process for the bath consists of multistage evaporation, followed by cooling of the concentrated solution from 50 °C to 0 – 5 °C in a crystallizer. Glauber's salt is removed by centrifuges, and the mother liquor is recycled to the spinning bath. The zinc sulfate dissolved in the spinning bath does not crystallize. Thenardite containing 99.6 – 99.9 % sodium sulfate is produced by calcination of Glauber's salt in evaporating crystallizers with partial recycle of the mother liquor [23], [47–50]. In Western Europe in 1988, 305 000 t of sodium sulfate was produced from viscose-fiber manufacture, and in the United States in 1989, 89 000 t [20].

From Sodium Dichromate Production. Sodium sulfate is produced by the acidifica-

tion of sodium chromate solution with sulfuric acid according to the equation:

$$2\,Na_2CrO_4 + H_2SO_4 \longrightarrow Na_2Cr_2O_7 + Na_2SO_4 + H_2O$$

If the reaction is carried out above 32.4 °C, thenardite (> 99 % Na_2SO_4) precipitates. This still contains 0.2 % dichromate, which can be removed by dissolving, reducing with SO_2, and filtering off chromium (III) oxide. If the sodium sulfate is used in the sulfate cellulose (Kraft) process, the dichromate is reduced and therefore does not cause problems [21], [51,52].

Quantities of sodium sulfate obtained from sodium dichromate production in 1988 – 1989 were 70 000 t/a in western Europe and 110 000 t/a in the Unites States [20].

From Ascorbic Acid Production. In the synthesis of ascorbic acid by the Reichstein – Grüssner process, sulfuric acid is used in the production of the diacetone-L-sorbose intermediate. The sulfuric acid is neutralized with sodium hydroxide solution in a later stage of the process [53]. On removal of the solvent, a saturated solution of Na_2SO_4 is produced, from which thenardite (> 99.9 % Na_2SO_4) is obtained by evaporative crystallization. For each tonne of ascorbic acid, 1.5 t of sodium sulfate is produced. In 1988 – 1989, 45 000 t/a Na_2SO_4 was produced in this way in western Europe, and 17 000 t/a in the United States [20].

From Other Processes. Sodium sulfate is also produced in various quantities from a large number of other chemical processes (Table 4).

1.3.3. Production from Natural Sources

The most important production plants for sodium sulfate from natural sources are located in Canada, Spain, Mexico, Turkey, the United States, and the CIS. The output from these countries in 1987 – 1989 was ca. 2×10^6 t/a [20].

Owing to the varying compositions of the salt mixtures in the deposits, many types of processes for the production of sodium sulfate and other compounds have been developed [2], [46], [72,73]. A review of some important production methods is given here.

Production in Canada. The Canadian deposits are located in Alberta and Saskatchewan, and are in the form of shallow lakes with only a few inlets. These contain sodium sulfate in the form of brines or deposits of mirabilite ($Na_2SO_4 \cdot 10\,H_2O$). Mirabilite dissolves from the deposits to an extent that depends on annual temperature variations, amount of water entering the lake, and evaporation. Various processes are used to recover the product.

Brine is pumped into shallow basins, from which some of the water evaporates during the summer. In winter, mirabilite crystallizes (ca. 600 kg/m^3 brine) and collects at the bottom of the basin. Excess brine is pumped back into the lake. Mirabilite is then collected by means of drag lines or other equipment and taken for further treatment.

Other operations use floating dredges to mine permanent deposits of mirabilite from the lake beds. The slurry of mirabilite and brine is sent by pipeline for further treatment.

In a third method, brine is produced by dissolving the salt in situ with hot water, which is injected via a large number of boreholes into the body of salt lying under the brine. The solution produced is taken to the next stage of treatment, which consists of precipitating crude salt in cooling towers [2], [20], [72], [74,75].

In some cases, dry deposits of mirabilite are mined conventionally, and the crude salt can then be transported to the processing plant.

The crude salt is charged into evaporators, where it "melts" in its own water of crystallization. The crystal suspension is then evaporated with hot combustion gases and transported to rotary dryers by belt conveyors.

In operations that send a brine for processing, Glauber's salt is crystallized by cascading the brine in contact with cold air [20], [75]. Entrained clay slime can be removed in centrifuges. The Glauber's salt melt is then evaporated by the submerged combustion of natural gas. Sodium sulfate (97 – 99.8 % Na_2SO_4) with a controlled grain size is produced by grinding and screening. Total production in Canada in 1987 – 1988 was 312 000 t/a sodium sulfate [20].

Production in Spain. Glauberite deposits are mined in the provinces of Burgos, Toledo, and Madrid. The Castellar mine at Toledo produces a crude salt containing 67 % Na_2SO_4,

Table 4. Sodium sulfate as a byproduct of chemical processes

Process	Reaction	References
Production of boric acid	$Na_2B_4O_7 + H_2SO_4 + 5\,H_2O \longrightarrow 4\,H_3BO_3 + Na_2SO_4$	*
Production of chlorine dioxide	$2\,NaClO_3 + SO_2 \longrightarrow 2\,ClO_2 + Na_2SO_4$	[44], [54–56] **
Production of hydroxylamine	$2\,HON(SO_3)_2(Na, NH_4) + 4\,H_2O \longrightarrow (HONH_2)_2 \cdot H_2SO_4 + Na_2SO_4 + 2\,NH_4HSO_4$	[57], [58]
Production of lithium carbonate	$Li_2SO_4 + Na_2CO_3 \longrightarrow Li_2CO_3 + Na_2SO_4$	[20]
Production of resorcinol or naphthol	$C_6H_4(ONa)_2 + 2\,NaHSO_3 + 2\,H_2SO_4 \longrightarrow C_6H_4(OH)_2 + 2\,Na_2SO_4 + 2\,SO_2 + 2\,H_2O$	[20], [59], [60]
Production of methionine	$2\,CH_3S{-}CH_2CH_2CH(NH_2){-}COONa + H_2SO_4 \longrightarrow 2\,CH_3S{-}CH_2CH_2CH(NH_2){-}COOH + Na_2SO_4$	[20]
Production of formic acid	$2\,HCOONa + H_2SO_4 \longrightarrow 2\,HCOOH + Na_2SO_4$	[20], [61]
Neutralization of waste sulfuric acid	$2\,NaOH + H_2SO_4 \longrightarrow Na_2SO_4 + 2\,H_2O$	[62–65]
Roasting and chlorination of pyrite cinder	$2\,Cu_2S + 4\,FeSO_4 + 8\,NaCl + 5\,O_2 \longrightarrow 4\,Na_2SO_4 + 4\,CuCl_2 + 2\,Fe_2O_3 + 2\,SO_2$	[2], [20], [23], [66]
Desulfurization of flue gases	$2\,Na_2SO_3 + O_2 \longrightarrow 2\,Na_2SO_4$; $2\,Na_2SO_3 + SO_3 \longrightarrow Na_2SO_4 + Na_2S_2O_5$ $2\,Na_2SO_3 + Na_2S_2O_5 \longrightarrow 2\,Na_2SO_4 + Na_2S_2O_3$	[20], [67–71]

* See → Boric Oxide, Boric Acid, and Borates, Chap. 3.3.
** See also → Chlorine Oxides and Chlorine Oxygen Acids, Chap. 5.2.1., → Chlorine Oxides and Chlorine Oxygen Acids, Chap. 5.2.2., → Chlorine Oxides and Chlorine Oxygen Acids, Chap. 5.2.3.

19 % $CaSO_4$, and 12 % clay and shale, from which sodium sulfate can be produced in various ways [2], [7], [76], [77]. For example, sodium sulfate can be leached from the crude salt at 35 – 40 °C (stage 1). The residue contains slime (clay) and glauberite ($Na_2SO_4 \cdot CaSO_4$), which is stable above 29 °C. Below 29 °C, this solid material yields additional sodium sulfate solution, which is recycled to the first stage. Thenardite is produced from the leachate by vacuum evaporation. The three deposits yielded 510 000 t of sodium sulfate in 1988 [20].

Production in the United States, Mexico, Turkey, and the Commonwealth of Independent States. Sodium sulfate is obtained from natural deposits in California, Texas, and Utah.

Brine containing sodium carbonate is obtained by sinking wells into the deeper layers below the salt crust of Searles Lake in California. The brine is treated with carbon dioxide to precipitate sodium hydrogencarbonate, from which sodium carbonate is obtained. The filtrate is then cooled to 15 °C to precipitate borax, which is filtered off. Further cooling to 5 – 6 °C causes Glauber's salt to crystallize. The decahydrate still contains some finely divided borax, which can be removed by hydraulic screening. Only Glauber's salt with a grain size of > 0.8 mm is dehydrated to form thenardite. The fine crystals are used as crystallization seeds [2], [20], [72], [78]. In 1989, 218 000 t of sodium sulfate was produced.

In Texas, brine from deposits close to the surface is treated. In Seagraves, Glauber's salt is separated from brine saturated with sodium chloride at −7 to −10 °C. The product obtained is melted and evaporated by submerged gas burners to produce thenardite. The thenardite suspension is centrifuged and dried to yield a product containing 99.7 % Na_2SO_4 [2], [20], [72], [79]. In 1989, ca. 140 000 t of sodium sulfate was produced by this method [20].

A sodium sulfate production plant in Utah uses brine from the Great Salt Lake. The brine is pumped into solar evaporation ponds with a surface area of ca. 80 km^2. Each year, stepwise crystallization of sodium chloride, potassium salts, and Glauber's salt takes place. These are removed by scrapers and transported by road vehicles for treatment. The crude salt is dissolved, the saturated solution is filtered, and sodium chlo-

ride is added to precipitate thenardite, which, after filtration and drying, contains $>94\,\%$ Na_2SO_4. Recrystallization gives thenardite containing $>99.7\,\%$ Na_2SO_4 [2], [6], [20], [80]. In 1989, 25 000 t of Na_2SO_4 was produced [20].

Sodium sulfate production figures for Mexico, Turkey, and the former Soviet Union are listed in Table 5.

1.3.4. Production of Sodium Sulfate

In addition to production methods based on byproduct material or natural deposits of sodium sulfate, processes are also operated in which sodium sulfate is manufactured from compounds that contain sodium and sulfate.

Production from Sodium Chloride and Sulfuric Acid (Mannheim Process). The thermal reaction

$$2\,NaCl + H_2SO_4 \longrightarrow Na_2SO_4 + 2\,HCl$$

which takes place via the intermediate formation of sodium hydrogensulfate, was developed as part of the Leblanc sodium carbonate process [3]. The reaction produces sodium sulfate and hydrogen chloride, which are of comparable economic importance [20].

This process is often carried out in Mannheim furnaces [2], [21], [63], [81, 82], which have a diameter of 6 m and produce 25 t/d sodium sulfate and 37.5 t/d hydrochloric acid (31 %) by addition of 93 – 96 % sulfuric acid to sodium chloride. One tonne of sulfate requires 3.6×10^6 kJ (88 kg of fuel oil), 65 kW · h, 1.4 m^3 of process water, and 26 m^3 of cooling water. The product contains 97 – 99.7 % Na_2SO_4. The small amounts of free sulfuric acid in the product are neutralized during cooling in a rotary cooler by the addition of sodium carbonate. The commercial product is obtained by grinding and screening.

The Climax Chemical Co., United States, operates a plant in which sodium chloride and sulfuric acid are reacted by the Cannon fluidized-bed process. Production costs are lower than with the furnace process [2], [20], [83].

In 1988 – 1989, ca. 390 000 t/a sodium sulfate was produced from sodium chloride and sulfuric acid [20].

Sodium sulfate can be produced from dilute sulfuric acid and sodium chloride in a spray evaporator [62], [84].

Production by the Hargreaves Process. Up to 1983, the Morton Chemical Co. in Louisiana, United States, operated the Hargreaves process, which is based on the reaction

$$4\,NaCl + 2\,SO_2 + O_2 + 2\,H_2O \longrightarrow 2\,Na_2SO_4 + 4\,HCl$$

This process is of interest owing to its low energy consumption compared to other production processes. Mannheim furnaces were used for the reaction, but rotary kilns or fluidized-bed reactors have been proposed [2], [20], [82], [85–87].

Production by Metathesis. Several processes have been developed for the conversion of sodium chloride to sodium sulfate, for example,

$$2\,NaCl + MgSO_4 \longrightarrow Na_2SO_4 + MgCl_2$$

Of these processes, the reaction of sodium chloride with kieserite ($MgSO_4 \cdot H_2O$) is the most important, because of all the crude salts from German potash production, kieserite is the cheapest source of sulfate.

To obtain a good yield of Glauber's salt, it is important to have the optimum $NaCl/MgSO_4$ ratio in the starting liquor. In the five- component system $Mg-Na_2-SO_4-Cl_2-H_2O$, an optimum point exists for the composition of the starting liquor. During dehydration of Glauber's salt, sodium chloride is added to the mother liquor, causing thenardite to precipitate. The mother liquor can then be used to produce the starting liquor. The process produces thenardite containing 98.7 – 99.5 % Na_2SO_4 [2], [6], [40], [88–90]. In 1988, ca. 118 000 t of Na_2SO_4 was produced [20].

1.4. Exploitation of Byproduct Sodium Sulfate

Sodium sulfate is produced in a number of chemical processes (see Section 1.3.2) and must be utilized or disposed of. Hence, processes for converting sodium sulfate into its constituents (sodium hydroxide and sulfuric acid) or into other useful chemicals are of interest [5], [20].

Table 5. Production of sodium sulfate in Mexico, Turkey, and the Soviet Union in 1987 – 1988

Country	Starting material	Sodium sulfate production, t/a	References
Mexico	Brine from salt lakes	420 000	[20]
Turkey	Crude salts from salt lakes	157 000	[20]
Soviet Union	Water from the Caspian Sea, etc.	375 000	[20], [72]

1.4.1. Electrochemical Decomposition of Sodium Sulfate

The electrochemical decomposition of sodium sulfate takes place according to the overall reaction:

$$2\,Na_2SO_4 + 6\,H_2O \longrightarrow 2\,H_2SO_4 + 4\,NaOH + 2\,H_2 + O_2$$

The anodic decomposition of water produces H^+ ions and oxygen

$$2\,H_2O \longrightarrow 4\,e^- + 4\,H^+ + O_2$$

and the corresponding cathodic reaction produces OH^- ions and hydrogen

$$2\,H_2O + 2\,e^- \longrightarrow 2\,OH^- + H_2$$

If sodium hydroxide solution and sulfuric acid are required, reverse reaction between H^+ and OH^- must be prevented. This can be achieved by using a mercury cathode or a diaphragm cell [5].

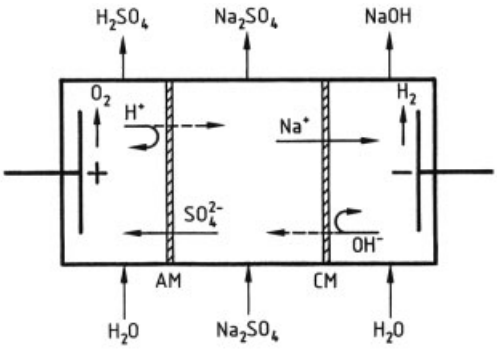

Figure 3. Decomposition of sodium sulfate by electrodialysis

The use of cation- and anion-selective ion-exchange membranes has led to new developments. Figure 3 shows the principle of Na_2SO_4 decomposition by electrodialysis in a cell divided into three zones by a cation-exchange membrane (CM) and an anion-exchange membrane (AM). The sulfate solution flows through the central zone. Sodium hydroxide and hydrogen are formed in the cathode space, and sulfuric acid and oxygen in the anode space. Sulfuric acid (5 – 15 wt %) and sodium hydroxide (15 – 20 wt %) are obtained with an energy consumption of 3500 – 4000 kW · h/t NaOH and a current efficiency of 60 – 80 %. The consumption of electrical energy can be reduced by using the reaction gases H_2 and O_2 in a fuel cell, or by the use of electrodialysis cells with gas-diffusion electrodes [91], [92].

Further improvements in electrodialysis are achieved by using anion-exchange membranes composed of styrene – divinylbenzene copolymer with strongly basic quaternary ammonium groups and a hydronium ion blocker. This gives improved current efficiencies and more concentrated sodium hydroxide and sulfuric acid [93].

Figure 4 shows a two- compartment cell divided by a cation-exchange membrane. The sulfate solution is charged into the anode space where it becomes enriched in sulfuric acid; only salt- containing acid can be produced. The sodium hydroxide in the cathode space is almost free of sulfate [94]. As the acid concentration increases and the salt concentration decreases, more H^+ ions contribute to the transport of current. Therefore, the current efficiency depends not only on the selectivity of the membrane and the sodium hydroxide concentration in the catholyte, but also on the H^+/Na^+ ratio in the anolyte. High current efficiencies are obtained only with an excess of Na^+ ions. For example, if the molar H^+/Na^+ ratio in the anolyte is 1 : 1, 5 – 10 % sulfuric acid and 15 – 25 % sodium hydroxide are obtained with current efficiencies of 65 – 70 %. Electrolysis of highly concentrated sodium sulfate solutions (up to 39 % Na_2SO_4) gives sulfuric acid concentrations of 10 – 20 %

and a sodium hydroxide content of 15 – 20 % with current efficiencies of 65 – 80 % [95].

In the Sulfomat process, sodium sulfate is electrolyzed to give sodium sulfate- containing sulfuric acid and caustic soda solution. The latter is used for absorption of SO_2 in flue gas desulfurization, giving a solution of sodium sulfite. Acidification with the Na_2SO_4-containing sulfuric acid gives SO_2 and reforms sodium sulfate, which is electrolyzed [96]. Also, spinning bath solutions from the viscose-fiber industry can be regenerated electrochemically [97], [98].

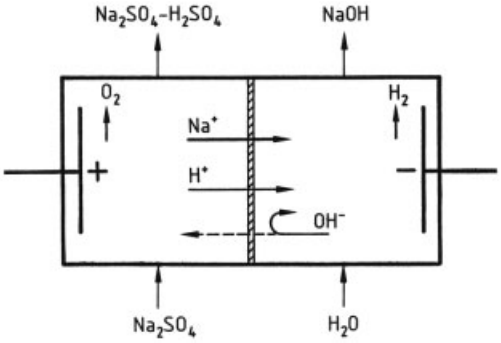

Figure 4. Electrolysis of sodium sulfate solution

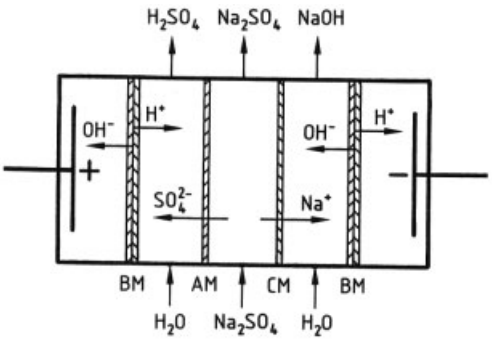

Figure 5. Electrodialysis of sodium sulfate with bipolar membranes

A development by Allied Signal of electrodialysis with bipolar membranes is of interest due to the low energy consumption [99–102]. In these bipolar membranes (BM), a cation-exchange membrane is bonded to an anion-exchange membrane [103]. Water molecules diffuse to the phase boundary CM – AM, where they dissociate in the electric field. The H^+ and OH^- ions diffuse in opposite directions. If the bipolar membranes are combined with CM and AM, as shown in Figure 5, the sulfate can be decomposed without the formation of hydrogen and oxygen. A large number of individual BM – AM – CM elements can be fitted between the two end electrodes. Decomposition of H_2O at the electrodes is then negligible. The only energy required is the dissociation energy of the water (0.022 kW · h/mol). Laboratory investigations showed an energy consumption of 1450 – 1700 kW · h/t NaOH when 12 – 16 % sodium hydroxide and 5 – 10 % sulfuric acid are produced. Allied Signal supplies cells for pilot- and full-scale operation.

1.4.2. Conversion Processes

Metathesis with Hydrogen Chloride. Owing to the high cost of electricity in the sodium sulfate electrolysis process, early attempts were made to bring about metathesis with hydrogen chloride:

$$Na_2SO_4 + 2\,HCl \rightleftharpoons 2\,NaCl + H_2SO_4$$

If the reaction is carried out in 40 – 60 % sulfuric acid, sodium chloride precipitates and can be obtained in 99.2 % yield (with 40 % H_2SO_4 at 30 °C) [5], [52][104–106]. The concentrations must be such that enough HCl is taken up by the reaction mixture. A stoichiometric amount of HCl is sufficient for the reaction to proceed. Of this, approximately one-third distils off during the process, and must be recycled to the reactor.

The sodium chloride obtained can be used for the chlor-alkali electrolytic process [105]. The 40 – 60 % sulfuric acid produced in the industrial process contains 1 – 2 % residual Na_2SO_4 [105] and is preferably recycled to the process in which the sodium sulfate is formed or used for reactions with sodium compounds, such as the production of chlorine dioxide from sodium chlorate, whereby Na_2SO_4 is reacted continuously with HCl in a decantation – washing column [107].

The metathesis of Na_2SO_4 with HCl in dilute sulfuric acid was carried out in 1951 in a pilot plant [105]. Carrying out the metathesis at 800 – 900 °C in a melt allows pure sulfuric acid and pure sodium chloride to be obtained [108].

Conversion to Potassium Sulfate. The reaction

$$Na_2SO_4 + 2\,KCl \longrightarrow K_2SO_4 + 2\,NaCl$$

has been thoroughly investigated [5], [77], [109], [110]. It occurs in two stages: In the first stage, glaserite ($Na_2SO_4 \cdot 3\,K_2SO_4$) is formed, which is decomposed by more potassium chloride to form K_2SO_4 and NaCl in the second stage:

$$Na_2SO_4 \cdot 3\,K_2SO_4 + 2\,KCl \longrightarrow 4\,K_2SO_4 + 2\,NaCl$$

The reactions can be carried out at 20 – 40 °C. If Glauber's salt is used, temperatures of > 32 °C are preferred. In a modern plant, K_2SO_4 yields of 98 – 99 % are obtained [111], [112]. The reaction – crystallization process is carried out in a loop-type bubble column [113].

The process has been operated since1984in a Russian plant that consumes 40 000 t/a Na_2SO_4. The Potash Corporation of Saskatchewan (PCS) has operated an experimental plant since 1985 in Cory, Canada, with a consumption of 25 000 t/a Na_2SO_4 [20], [112]. Single-stage processes for carrying out this reaction have been proposed [114], [115].

1.4.3. Reduction Processes

Sodium sulfate can also be used in reductive processes (Table 6).

1.5. Uses

World consumption of sodium sulfate in 1987 was ca. 4.6×10^6 t [20].

The most important consumers were the cellulose-fiber industry and manufacturers of glass and detergents (Table 7).

Other users are from a wide range of industries [2], [20], including dyeing technology [132], electrochemical metal treatment [133], animal feeds [2], heat accumulators [20], and sponge manufacture [20].

1.6. Quality Specifications and Analysis

The highest standards of quality and whiteness are required for the detergent, textile, and animal feed industries. For use as a refining agent in glass smelting, the iron content of sodium sulfate must be low. For the sulfate cellulose (Kraft) process, a purity of 96 – 98 % is adequate. Typical analysis figures are given in [2].

The specification in DAB 9 [134] sets the following maximum impurity levels in anhydrous sodium sulfate: chloride 450 ppm, calcium 450 ppm, iron 90 ppm, arsenic 5 ppm, heavy metals 45 ppm, and water 5 %.

Standard methods of analysis of sodium sulfate for industrial use are as follows:

ISO 3234-1975	weight loss on heating to 110 °C
ISO 3235-1975	fraction insoluble in acid
ISO 3236-1975	chloride content (mercurimetric)
ISO 3237-1975	sulfate content, calculation and gravimetric method (as barium sulfate)
ISO 3238-1975	calcium content by the EDTA complexometric method
ISO 3239-1975	iron content by photometry with 1,10-phenanthroline
ISO 3240-1975	acid or alkali content
ISO 3241-1975	pH by potentiometric method
ISO 5142-1977	sodium sulfate content, gravimetric method
ISO 5994-1979	calcium content by atomic absorption spectroscopy

1.7. Occupational Health and Environmental Protection

Sodium sulfate is a permitted food additive [135] and is a common component of spa waters [2].

Inhalation of sodium sulfate dust causes irritation of the mucous membranes, and prolonged skin contact has a drying effect. The laxative effect of oral ingestion is well known and is used medicinally.

Sodium sulfate in wastewater can be recovered as thenardite (Section 1.3.2). However, if the solutions are very dilute (< 5 % Na_2SO_4), the crystallization process is energy intensive. In such cases, treatment of the wastewater by a membrane process [136], a precipitation process [137], or by biochemical methods [36] can be considered.

1.8. Economic Aspects

Demand for sodium sulfate in the Western industrialized nations has decreased during the 1980s, but further reduction is expected to be balanced by increasing consumption in other parts of the world. The cellulose and glass industries and detergent manufacture account for ca. 80 % of the

Table 6. Use of sodium sulfate for reduction processes

Process	Reaction	References
Leblanc process	$Na_2SO_4 + 2\,C + CaCO_3 \longrightarrow Na_2CO_3 + CaS + 2\,CO_2$	[5], [11], [116]
Production of sodium sulfide	$Na_2SO_4 + 2\,C \longrightarrow Na_2S + 2\,CO_2$ $Na_2SO_4 + CH_4 \longrightarrow Na_2S + CO_2 + 2\,H_2O$	[5], [20], [117–120]
Production of aluminum oxide (Peniakoff process)	$2\,Al_2O_3$ (bauxite) $+ 2\,Na_2SO_4 + C \longrightarrow 4\,NaAlO_2 + 2\,SO_2 + CO_2$ $2\,NaAlO_2 + CO_2 + 3\,H_2O \longrightarrow 2\,Al(OH)_3 + Na_2CO_3$	[121–125]
Digestion of phosphate rock for fertilizer production	$Ca_3(PO_4)_2 + 2\,Na_2SO_4 + 8\,C \longrightarrow CaNa_4(PO_3S)_2 + 2\,CaO + 8\,CO$	[126], [127]
Production of sodium silicate (water glass)	$2\,Na_2SO_4 + 2\,SiO_2 + C \longrightarrow 2\,Na_2SiO_3 + 2\,SO_2 + CO_2$	[5]

sodium sulfate consumed. Environmental regulations have a great influence on the market for sodium sulfate, affecting both its production and its consumption. Owing to the large natural reserves of sulfate and the large quantities produced as a byproduct of the chemical industry, prices are not expected to increase greatly. The most important producing countries in 1988 were the United States 16 %; the Soviet Union 14 %; Spain 13 %; Mexico 10 %; Canada 7 %; Japan 6 %; and Belgium 6 %. The remaining 28 % came from a number of sources [20].

Table 7. Consumption of sodium sulfate in1987 (in 10^3 t/a)

Use	Consumption		References
	Western World∗	United States	
Cellulose industry	714	271	[20], [128], [129]
Glass production and others	456	128	[20], [130]
Detergents	1736	354	[20], [131]

∗ Including United States.

2. Sodium Hydrogensulfate

Sodium hydrogensulfate [138] occurs as the monohydrate, $NaHSO_4 \cdot H_2O$, in the system sodium sulfate – sulfuric acid – water [15], [18] or exists as the solid phase $NaHSO_4$ at a sulfuric acid concentration of 62 %. The monohydrate is converted to the anhydrous salt at 58.45 ± 0.05 °C.

The thermal decomposition represented by the equation

$$2\,NaHSO_4 \rightleftharpoons Na_2S_2O_7 + H_2O$$

takes place near the melting point, which can be determined only approximately (ca. 183 °C) with a water vapor pressure of 2500 Pa (25 mbar). Conversion to sodium disulfate is complete after heating for ca. 4 h at 240 – 260 °C. Sodium disulfate decomposes above 400 °C to form sodium sulfate with liberation of sulfur trioxide.

Production. Sodium hydrogensulfate is formed as an intermediate in the production of hydrochloric acid by the Mannheim process (Section 1.3.4) and is a byproduct of the manufacture of chromium(VI) oxide [51], [56], [63], [139]. The reaction between sulfuric acid and sodium chloride may be carried out by the old process using cast iron retorts heated with gas or oil, or by the use of submerged combustion equipment in brick-lined vessels [82], [140]. The addition of a stoichiometric amount of sulfuric acid to sodium sulfate or Glauber's salt at 200 – 280 °C produces molten sodium hydrogensulfate [141]. The same reaction can be utilized to produce sodium hydrogensulfate in aqueous solution at 80 °C directly [141].

Uses. Sodium hydrogensulfate is used in the manufacture of household cleaners, in the textile industry, for the regeneration of rubber, for metal surface treatment, and as a flux for low-melting metals and their alloys [2].

3. References

1. J. F. Henkels: *Pyritologia oder Kieshistorie,* Großische Handlung, Leipzig 1754, p. 705.
2. *Ullmann,* 4th ed. **17,** 211 – 229.
3. G. Lunge: *Handbuch der Soda-Industrie,* 3rd ed., **vol. 2**, Vieweg, Braunschweig 1909.

4. H. Furkert in F. Ebel (ed.): *Chemische Berichte,* Reichsamt für Wirtschaftsausbau, Berlin, January 1943, pp. 28 – 33.
5. H. v. Plessen et al., *Chem. Ing. Tech.* **61** (1989) 933 –940.
6. Great Salt Lake Minerals and Chemicals Corp., US 3 575 688, 1969.
7. Mannesmann, DE-OS 3 721 667, 1987.
8. Intermountain Res. Dev. Corp., US 4 519 806, 1982.
9. Techemet, ZA 87/09 599, 1987.
10. *Ullmann,* 4th ed., **13,** 450.
11. *Gmelin,* 8 th ed., Natrium (1928).
12. *Gmelin,* 8 th ed., Natrium, Erg.bd. (1964 – 1970).
13. In [12], part 3, 1117.
14. In [12], part 3, 1092 – 1095.
15. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Metal-Organic Compounds,* 4th ed., **vol. 2,** American Chemical Society, Washington, D.C., 1965, pp. 1121 – 1123.
16. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* **vol. 1/1,** Pergamon, London 1963, pp. 126 – 127.
17. A. Seidell: *Solubilities of Inorganic and Metal-Organic Compounds,* 3rd ed., **vol. 1,** Van Nostrand, New York 1940, pp. 1234 – 1235.
18. H. L. Silcock (ed.): *Solubilities of Inorganic and Organic Compounds,* **vol. 3/2,** Pergamon, Oxford 1979, pp. 827 – 848, 971 – 972.
19. R. E. Vener, A. R. Thompson, *Ind. Eng. Chem.* **41** (1949) 2242 – 2247; **42** (1950) 171 – 174.
20. *The Economics of Sodium Sulphate 1990,* 6th ed., Roskill Information Services, London 1990.
21. SRI International: *Chemical Economics Handbook,* May 1982, p. 771, 1000 G-T.
22. R. C. Kerr et al., *Nature (London)* **340** (1989) no. 6232, 357 – 362.
23. In [12], part 1, 231 – 234.
24. *Winnacker-Küchler,* 4th ed., **1 – 7,** 1981 – 1986.
25. *Winnacker-Küchler,* 4th ed., **2,** 76 – 78.
26. Metallgesellschaft AG, DE 1 150 368, 1960.
27. Standard-Messo, DE 1 467 244, 1963.
28. Standard-Messo, DE 1 519 911, 1965.
29. W. Linstaedt, G. Hofmann, *Chem. Ind. (Düsseldorf)* **31** (1979) 329 – 333.
30. R. J. King, *Chem. Eng. Prog.* **80** (1984) no. 7, 63 – 69.
31. D. J. Neville, *Cryst. Precip. Proc. Int. Symp.,* 1987, pp. 109 – 116.
32. T. Messing, W. Wöhlk, *Chem. Ing. Tech.* **45** (1973), 106 – 109.
33. Metallgesellschaft AG, DE-OS 2 904 345, 1979.
34. Ebner & Co, DE-OS 2 246 568, 1972.
35. Standard-Messo, DE-OS 1 519 910, 1965.
36. R. Hürzeler, G. Stucki, *GWF Gas Wasserfach: Wasser Abwasser* **131** (1990) 231 – 244.
37. P. G. Rueffel: *Proc. 3rd Symp. Salt 1969* (Publ. 1970) p. 429 – 451.
38. Ebara Infilco, DE-AS 1 567 941, 1967.
39. Escher-Wyss, DE-OS 2 543 721, 1975.
40. Wintershall AG, DE-AS 2 029 984, 1970.
41. Multifibre Process, AT 3 55 517, 1978.
42. W. Fleischmann, A. Mersmann, *Process Technol. Proc. 1984,* **no. 2 (Ind. Cryst.),** pp. 165 – 170.
43. W. Fleischmann, A. Mersmann, *Chem. Ing. Tech.* **57** (1985) 178 – 179.
44. D. Lobley, K. L. Pinder, *Waste Treat. Util. Proc. Int. Symp. 2nd* 1980 (Publ. 1982) pp. 63 – 70.
45. G. Mina-Mankarios, K. L. Pinder, *Can. J. Chem. Eng.* **69** (1991) 308 – 324.
46. M. R. Bhatt et al., *Res. Ind.* **37** (1992) no. 1, 34 – 35.
47. T. Kobayashi, *CEER Chem. Econ. Eng. Rev.* **8** (1976) no. 6, 42 – 48.
48. M. Driesen, *Br. Chem. Eng.* **15** (1970) 1154 – 1158.
49. Politechnika Szczecinska, DD 265 390, 1986.
50. *Chem. Abstr.* **107** (1987) 24670x.
51. Allied Chem. Corp, DE-OS 2 151 323, 1970.
52. Occidental Chem. Corp., US 5 393 503, 1991.
53. *Ullmann,* 4th ed., **23,** 688.
54. Kema Nord AB, US 4 678 653, 1985.
55. Tenneco Canada, US 4 780 304, 1987.
56. Lundberg, US 4 483 740, 1983.
57. G. Gurato et al., *Chim. Ind. (Milan)* **69** (1987) no. 12, 140 – 147.
58. G. Gurato et al., *J. Cryst. Growth* **66** (1984) 621 – 631.
59. *Winnacker-Küchler,* 4th ed., **6,** 182 – 183.
60. E. M. Arnoldov et al., *Sov. Chem. Ind. (Engl. Transl.)* **11** (1979) 549 – 550.
61. Polioli, EP-OS 245 719, 1986.
62. FIAT, Microfilm AA-216, Frames 8 – 15, 65 – 156 (PB. 73556).
63. J. B. Hunter, EP-OS 65 412, 1981.
64. B. Asano, M. Olper, *Proc. 43rd Ind. Waste Conf.* 1988 (Publ. 1989) 45 – 50.
65. R. M. Reynolds et al., *Proc. World Symp. Metall. Environ. Control 1990,* 119th TMS Annu. Meet., pp.1001 – 1056.
66. *Ullmann,* 3rd ed., **11,** 185 – 187.
67. Balcke-Dürr AG, EP-OS 301 452, 1987.

68. Chemieanlagenbaukombinat Leipzig-Grimma, DD 293 098, 1987.
69. Metallgesellschaft AG, DE-OS 3 616 508, 1986.
70. D. J. Wahl, *Umwelt* **20** (1990) no. 3, L 4 – L 10.
71. S. Weiß, H. Wanko, *Chem. Tech. (Leipzig)* **43** (1991) 177 – 188.
72. In [12], part 1, 214 – 224.
73. G. Lesino et al., *Sol. Energy* **45** (1990) no. 4, 215 –219.
74. W. A. MacWilliams, R. G. Reynolds, *CIM Can. Inst. Mining Met. Bull.* **66** (1973) no. 734, 115 – 119. Alberta Sulphate, CA 1 007 563, 1975.
75. P. L. Broughton, *Ind. Miner. (London)* 1977 no. 121, 51 – 59.
76. MINERSA, ES 504 459, 1981.
77. O. K. Yanateva et al., *J. Appl. Chem. USSR (Engl. Transl.)* **52** (1979) 388 – 389.
78. Kerr-McGee Chem. Corp., US 4 291 002, 1977; US 4 478599, 1982.
79. W. I. Weisman, *Chem. Eng. Prog.* **60** (1964) no. 11, 47 – 49.
80. P. Behrens, *Bull. Utah Geol. Min. Surv.* 1980, no. 116, 233 – 238.
81. *Chem. Ing. Tech.* **37** (1965) 666.
82. *Ullmann,* 3rd ed., **15,** 71 – 73.
83. C. W. Cannon, US 2 706 145, 1955.
84. IG Farbenindustrie AG, DE 694 409, 1936.
85. FIAT, Microfilm AA-216, Frames 174 – 188 (PB. 73556).
86. In [12], part 1, 228 – 229.
87. Morton Int., DE-AS 1 293 139, 1961.
88. *Winnacker-Küchler,* 3rd ed., **1,** 150 – 151.
89. In [12], part 1, 227.
90. H. Picht, *Hung. J. Ind. Chem.* **17** (1989) no. 2, 179 –183.
91. J. Jörissen et al., *DECHEMA-Monogr.* **124** (1991) 21 – 35.
92. J. Jörissen et al., *J. Appl. Electrochem.* **21** (1991) 869 – 876.
93. BASF, EP-OS 449 071, 1990.
94. B. Pieper, *Fortschr. Ber. VDI Reihe 3* **132** (1987).
95. ICI, EP-OS 532 188, 1991.
96. K. Tomiie, Y. Hayashi, *New Mater. New Processes Electrochem. Technol.* **1** (1981) 284 – 288.
97. H. Kau, *DECHEMA-Monogr.* **98** (1985) 291 – 298.
98. Lenzing AG, EP-OS 399 993, 1989.
99. K. Nagasubramanian et al., *J. Membr. Sci.* **2** (1977) 109 – 124; **3** (1978) 71 – 83.
100. Allied Corp., US 4 116 889, 1978.
101. Allied Corp., US 4 140 815, 1979.
102. K. N. Mani et al., *Desalination* **68** (1988) 149 – 166.
103. H. Strathmann et al. *DECHEMA-Monogr.* **125** (1992) 83 – 100.
104. BASF, DE 868 742, 1941.
105. C. F. Linström, *Chem. Tech. (Leipzig)* **4** (1952) 375 –380.
106. K. Altenbach, Hoechst AG, personal communication, Frankfurt/Main 29.11.1991.
107. Erco Industries, DE-OS 2 856 504, 1978.
108. C. F. Hoppe, Dissertation Dortmund 1995, Aachen: Shaker Verlag, 1996
109. *Gmelin,* 8th ed., Kalium, Erg.bd. (1970), 26 – 37.
110. H. Scherzberg et al., *Chem. Tech. (Heidelberg)* **20** (1991) no. 6, 41 – 48.
111. Forschungsinstitut Leningrad, DE 2 820 445, 1978.
112. U. Neitzel, *Kali Steinsalz* **9** (1986) 257 – 261.
113. G. Z. Zelmanov, V. G. Terlo, *Sov. Chem. Ind.* **18** (1986) no. 5, 46 – 48.
114. Graessor Contractors, GB 2 270 308, 1992.
115. Mannesmann, DE 4 340 839, 1993.
116. FIAT, Microfilm H-45, Frames 7834 – 7842 (PB. 70041).
117. BIOS Final Report no. 855.
118. Waagner-Biro, Chemiefaser Lenzing, AT 349 494, 1976.
119. Rockwell International, EP-OS 91 948, 1981.
120. J. H. Cameron et al., *Ind. Eng. Chem. Fundam.* **24** (1985) 443 – 449.
121. BIOS Final Report no. 288.
122. H. L. Murray: Office Publication Board, Report No. 223 (*Bibl. Sci. Ind. Rep.* **1** (1946) 7).
123. Metallgesellschaft AG, EP-OS 540 072, 1991.
124. H. Rüter in F. Ebel (ed.): *Chemische Berichte,* Reichsamt für Wirtschaftsausbau, Berlin Dec. 1941, p. 201 – 208.
125. FIAT Final Report no. 989.
126. Metallgesellschaft AG, GB 473 520, 1936; 513 744, 1937.
127. BIOS Final Report no. 107.
128. *Ullmann,* 4th ed., **17,** 555 – 556.
129. *Kirk-Othmer,* 3rd ed., **19,** 395 – 406.
130. In [24], **3,** 117.
131. Henkel, EP-OS 302 403, 1988.
132. *Ullmann,* 4th ed., **22,** 643, 660.
133. Allegheny Ludlum, US 4 824 536, 1988.
134. Deutsches Arzneibuch, 9th ed., Deutscher Apotheker-Verlag, Stuttgart, Govi, Frankfurt/M. 1986, pp. 1079 – 1080.

135. Zusatzstoff-Verkehrsordnung vom 10.07.1984, BGBl. I, no. 30 (1984), 897 – 901, and Anlage 2, p. 64.
136. S. N. Gaeta, *Chem. Sep., Dev. Sel. Pap. Int. Conf., Sep. Sci. Technol. 1st* 1986, no. 2, 95 – 106.
137. Walhalla-Kalk, EP 250 626, 1986.
138. In [12], part 1, 226, 246 – 247; part 3, 1149 – 1159.
139. *Ullmann,* 4th ed., **9,** 612 – 613.
140. Hoechst, DE-OS 2 810 693, 1978.
141. Mannesmann, DE-OS 3 723 292, 1987.

Sodium Sulfide → Sulfides, Polysulfides, and Sulfanes

Solar Technology

Heat Storage Media, Photovoltaic Cells, Semiconductors, and Wind Energy are separate keywords. Hydrogen Technology → Hydrogen, Chap. 9.

JOACHIM LUTHER, Universität Oldenburg, Fachbereich Physik, Oldenburg, Federal Republic of Germany (Chap. 1)

MICHAEL NAST, Deutsches Zentrum für Luft- und Raumfahrt, Institut für Technische Thermodynamik, Stuttgart, Federal Republic of Germany (Chap. 2)

M. NORBERT FISCH, Universität Stuttgart, Institut für Thermodynamik und Wärmetechnik, Stuttgart, Federal Republic of Germany (Chap. 2)

DIRK CHRISTOFFERS, Institut für Solarenergieforschung Hameln/Emmerthal (ISFH), Emmerthal, Federal Republic of Germany (Chap. 3)

FRITZ PFISTERER, Universität Stuttgart, Institut für Physikalische Elektronik, Stuttgart, Federal Republic of Germany (Chap. 4)

DIETER MEISSNER, AQR, Jülich, Federal Republic of Germany (Chap. 5)

JOACHIM NITSCH, Deutsches Zentrum für Luft- und Raumfahrt, Institut für Technische Thermodynamik, Stuttgart, Federal Republic of Germany (Chap. 6)

1. Introduction 2
1.1. Concerning the Importance of Solar Energy Conversion 2
1.2. Terrestrial Solar Radiation 4
1.3. Methods of Solar Energy Conversion 5
1.4. Matching of Solar Energy Production and Consumption 5
2. Solar Thermal Energy 6
2.1. Development and Historical Aspects 6
2.2. Solar Thermal Energy Conversion 8
2.3. Technical Utilization of Solar Energy 9
2.3.1. Low-Temperature Applications . . . 9
2.3.1.1. Design and Function of Flat Collectors 10
2.3.1.2. Systems for Heating Water 12
2.3.1.3. Systems for Room Heating 14
2.3.1.4. Other Low-Temperature Applications 15
2.3.2. Medium-Temperature and High-Temperature Applications 17
2.3.2.1. Concentrating Collectors 17
2.3.2.2. Solar Thermal Power Plants 18
2.3.2.3. Solar Process Heat 19
3. Passive Utilization of Solar Energy 19
3.1. Introduction 19
3.2. Building Design 21
3.2.1. Site Analysis 21
3.2.2. Design 21
3.2.3. Heat Conservation 22
3.3. Elements of Passive Utilization . . 23
3.3.1. Collection 23
3.3.2. Storage and Distribution 26
3.3.3. Shading and Cooling 26
3.3.4. Daylight 27
3.4. Performance of a Solar House . . 27
4. Photovoltaic Systems 27
4.1. Principle of Operation, PV Materials, and State-of-the-Art Solar Cells 27
4.2. General Structure of PV Systems 28
4.3. Photovoltaic Generators 29
4.3.1. Module Technology and Specifications 29
4.3.2. Tracking and Concentration 30
4.4. Power Conditioning Systems . . . 31
4.5. Storage 32
4.6. Stand-Alone PV Systems 32
4.7. Grid-Connected PV Systems . . . 32
4.8. Design Considerations and Modeling 33
4.9. Operational Experience 33
4.9.1. PV Module Performance 33

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_369

4.9.2. PV System Efficiency, Annual Energy Output 33
4.10. Economic Aspects, Market, Energy Payback, Prospects 34
4.10.1. Stand-Alone PV Systems 34
4.10.2. Grid-Connected PV Systems 34
4.10.3. Energy Payback Time 34
4.10.4. Multifunctional Use of PV Modules 35
5. Photoelectrochemical Solar Energy Conversion 35
5.1. Principles of Photoelectrochemical Solar Energy Conversion . . . 36
5.2. Dye-Sensitized Photoelectrochemical Solar Cells 39
5.3. Solar Wastewater Detoxification . 40
5.4. Direct Photoelectrochemical Water Splitting 41
6. Economic Aspects 43
7. References 48

1. Introduction

1.1. Concerning the Importance of Solar Energy Conversion

About 78 % of the world's current energy supplies are based on fossil energy sources; the remaining primary energy is provided by hydropower (5 %), firewood/biomass (13 %) and nuclear energy (4 %) [1]. The following structurally inherent problem areas of such an energy supply will become increasingly evident:

1) Exponential growth of CO_2 levels in the atmosphere, leading to the threat of a worldwide climatic change (anticipated mean temperature increase of 2 – 4 K for a doubling of CO_2 concentration [2]). The World Climate Conference held in Toronto in 1988 already recommended a global reduction in CO_2 emissions of 50 % by the middle of the next century.
2) Cheap fossil energy sources will in the long term become only limitedly available; furthermore, these resources, particularly oil, are distributed very unevenly worldwide.
3) Exploitation of nuclear energy does not appear to be socially acceptable.
4) The problem of energy provision in the Third World remains unsolved, particularly in decentralized regions (one example: at the present time more than 10^9 persons have no access to electrical energy for lighting, heating, cooking, TV, etc. [3]).

If it is assumed that the exploitation of fossil energy sources will have to be drastically reduced in the future and that nuclear energy (including fusion) will not be a feasible option for the worldwide provision of energy, two alternative energy sources will have to be vigorously pursued:

1) Rational use of energy
2) Renewable energy sources

In European industrialized countries it is acknowledged that the virtual resource of "efficient energy utilization" represents potentially about 50 % of present-day energy consumption [4].

The following simple calculation illustrates the considerable potential of terrestrial solar energy utilization:

The total mean solar radiant flux at sea level is 10^{17} W. If it is assumed that, on average, this solar input is converted with an efficiency of 10 % into final energy, then with a global areal utilization of 0.5 %, 5×10^{13} W of technically utilizable power will be available. With a future world population of 10^{10} people this would mean an average value of 5 kW per head. This corresponds roughly to the current mean value of energy consumption in western Europe.

This numerical example is simply intended to illustrate the potential of renewable energy sources with the example of the direct utilization of solar energy. Water power, wind energy, biomass (principally utilization of waste), etc., will contribute substantially to a nonfossil fuel energy supply.

The next development stage of the energy supply system, avoiding the aforementioned problem areas, could therefore have the following outline structure:

1) More intelligent handling of energy
2) Extensive utilization of renewable energy sources
3) Residual utilization of fossil energy sources

A residual utilization of fossil resources at an ecologically acceptable level is envisaged in

order to be able to compensate to some extent variations, e.g., in solar energy supplies (see Section 1.4).

Obviously, thorough investigations must be carried out to ensure that such a new energy system does not lead to problems. In this connection, some characteristic features of solar energy conversion will be mentioned:

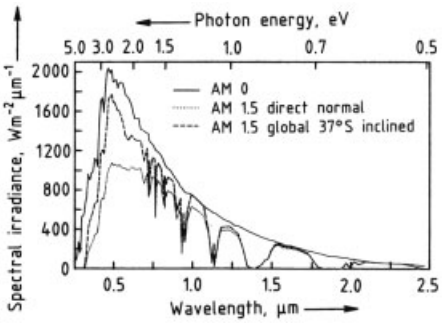

Figure 1. Spectral irradiance as a function of wavelength and photon energy. The parameter AM is explained in Figure 2

Pollutants. Solar energy conversion is not based on chemical or nuclear fuels, and therefore in principle no harmful waste products such as CO_2, NO_x, radioactive waste, etc., are produced during operation of the energy converters. Potential environmental pollution may occur in the context of a solar energy technology – as is common with any technical production – in the manufacture of technical components (e.g., solar cells). The relevant processes must be suitably optimized and the actual production stage must be carefully controlled. The recyclability of the extensive converter systems must be taken into account.

Areal Requirements. Solar energy conversion is characterized by a relatively small energy flux density (see Section 1.2), which inevitably means a large-area energy conversion technology. The requisite areas can, however, be realized to a considerable extent on existing buildings and transportation facilities (in Germany ca. 10 % of the total area); additional energy conversion areas would have to be distributed among a large number of small units.

Change in Global Energy Fluxes. The albedo (nondirectional reflectance) for solar radiation at the converter sites would on average be reduced by the introduction of a large-scale energy conversion program which leads in principle to local heating. Quantitatively, this change in atmospheric energy fluxes is at most a few thousandths; the effects will therefore be negligible compared to other anthropogenic effects, as long as largely the installation of distributed small-scale converter systems is adopted.

1.2. Terrestrial Solar Radiation

The radiation properties of the sun and the geometry of the solar-terrestrial system result in an irradiance of

$$E_{\mathrm{sc}} = 1367 \pm 1\,\% \,\mathrm{W/m^2}$$

at the top of the Earth's atmosphere (at the time of the equinoxes) [5]; where E_{sc} is termed the solar constant. The associated extraterrestrial spectrum (Fig. 1) can be satisfactorily approximated by a Planck radiator with $T = 5777$ K. The extinction of the solar radiation in the atmosphere is roughly a function of the parameter "air mass" (AM) (Fig. 2). Spectra for AM 0 and AM 1.5 are shown in Fig. 1.

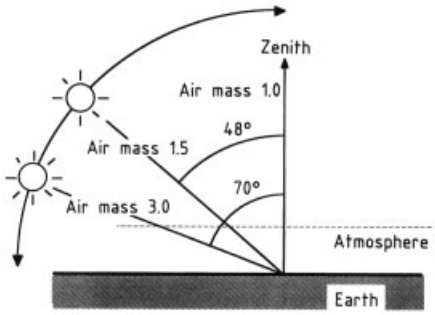

Figure 2. The parameter "air mass" (AM) is defined as the relative path length of solar beam radiation through the atmosphere (AM = 0 characterizes extraterrestrial situations)

The shape of the terrestrial solar spectrum has a considerable influence on the conversion efficiency of technologies that require photons with energies above a threshold value (photovoltaics, photochemistry).

In a clear atmosphere and under AM 1.5 conditions, maximum irradiances of about

$$E_{\max} \approx 1000\,\mathrm{W/m^2}$$

occur on the Earth's surface. Mean values of the solar radiant flux densities for three typical sites are given in Fig. 3. The relatively low energy densities in the range of a few hundred watts per m^2 are the reason why a large-scale solar energy utilization will always also be a "large-area technology" (see Section 1.1).

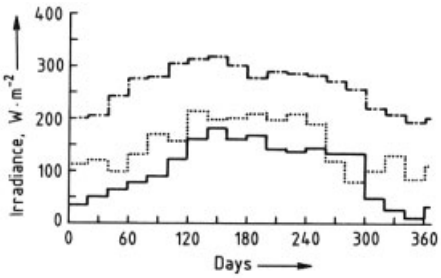

Figure 3. Means values of solar irradiance for Phoenix, United States (33.4 °N, upper curve), Ispra, Italy (45.8 °N), and Hamburg, Germany (53.6 °N, lower curve). Data for south-facing tilted planes (tilt angle equal to the latitude of the site); averaging period: 20 d

The mean annual radiation level at a radiation-favorable desert site and at a site in central Europe differ only by a factor of 2 (Fig. 3). Substantial differences are found however in the seasonal variation of the radiation intensity: The maximum to minimum mean value is 1.5 : 1 and 15 : 1, respectively. These seasonal changes in the solar energy supply necessitate the use of long-term energy storage systems or complementary, energy-producing systems that can be regulated (e.g., residual, fossil-fueled plants), at least in higher latitudes (see Section 1.4).

On account of the scattering of light in the atmosphere the global (total) radiation E_g at the Earth's surface consists of a diffuse component E_d and a beam component E_b. The E_d/E_g ratio for the three sites in Figure 3 is illustrated in Figure 4. Because of thermodynamic principles, the diffuse radiation – as long as the frequency of the light is not changed – cannot be concentrated. This means that conversion technologies that use lenses or mirrors to concentrate the radiation can be used only under unfavorable conditions at higher latitudes, in which the E_d component is as a rule more than 50 %.

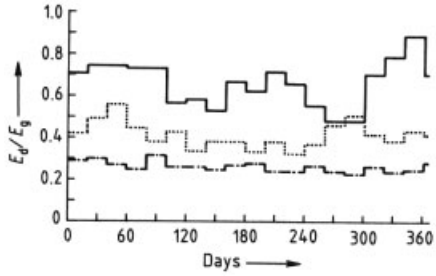

Figure 4. Ratio of diffuse radiation to global radiation for the data sets shown in Figure 3 (upper curve: Hamburg, lower curve: Phoenix, middle curve: Ispra)

1.3. Methods of Solar Energy Conversion

The methods that can in principle be used to convert solar radiation into technically utilizable energy may be subdivided into three categories:

1) Photothermal conversion
2) Photovoltaic conversion
3) Photochemical conversion

In *photothermal methods* solar radiation is absorbed by optically broad-band absorbers (typically: $\lambda_{abs} = 0.3 - 2\,\mu m$); the energy of the photons is converted into thermal energy. Depending on the quality of the insulation of the absorbers, fluid output temperatures of about 20 – 150 K above ambient temperature can be achieved with good energy efficiencies without optical concentration. Systems with moderate concentration reach useful temperatures of up to 800 K, while highly-concentrating methods reach temperatures of 1500 K and above.

Solar thermal energy conversion has broad areas of application, ranging from domestic heating, to industrial process heat generation, e.g., for cooling, distillation and drying; electrical energy can be provided on a large scale through solar thermal power-stations. *Passive solar energy conversion* for heating buildings, in which parts of structural bodies serve as radiation absorbers and energy storages, is technically particularly elegant.

Photovoltaic (PV) energy conversion involves the direct conversion of photon energy into excitation energy of electrical charge carriers. In semiconductor solar cells, e.g., of silicon, current densities of 300 A/m^2 at a voltage of

0.5 V can be achieved with optically nonconcentrating systems.

Applications of photovoltaic energy converters range from consumer products (pocket calculators), through decentralized autonomous power supplies (telecommunications, domestic power supply, drinking water pumps) to utility-grid-coupled PV units for power generation in conjunction with conventional power stations.

Technically perfected systems already exist for photothermal and photovoltaic energy conversion. There are areas of application in which the use of these techniques is already economically viable. In contrast to this, *photochemical* energy conversion is still in the basic research stage. With this type of conversion the chemical potential of reactants is increased by the direct absorption of photons (for example photosynthesis). A main problem of photochemical conversion is the low energy of the photons of the solar spectrum compared to most chemical excitation energies (see Fig. 1).

Further interesting ways of using solar energy can be obtained by combining the three methods discussed above. An example is the coupling of PV electricity and solar process heat in a high-temperature electrolysis process (reduction of the electrolysis voltage via heat input).

Reference may be made to [5] with regard to a thermodynamic treatment – particularly an exergy analysis – of the various energy conversion processes.

1.4. Matching of Solar Energy Production and Consumption

All energy supply systems have to solve the problem of the temporary matching of consumption and production. The conventional methods for solving this problem are:

1) Control of production
2) Load management
3) Influencing the structure of consumption
4) Storage devices

In the conventional heating and transportation sector there is a substantial temporary decoupling – at least when fossil, readily storable energy carriers are used – between the provision of energy and its utilization. In the electricity supply sector and in the utilization of waste heat (e.g., from combined heat and power units) the aforementioned methods have to be used more intensively.

In contrast to end energy production from fossil resources, energy provision from solar sources cannot be regulated. The variations in the energy supply are, however, not of a purely random nature, and indeed are largely governed by the deterministic diurnal and seasonal fluctuations.

Just as conventional energy systems have matched consumption – via the measures (2) and (3), to a production that is as constant as possible over time – a future energy supply system based largely on solar sources consumption would have to be adapted to the deterministic time behavior of the solar energy supply. Two simple examples of possible measures are the large-scale introduction of electric cars with charging phases preferably during the midday period, and the use of cooling units with an integrated cold store; both the vehicle charging stations as well as the operation of the cooling units could be remotely controlled from energy distribution centers.

In the case of electricity supply the random fluctuations in solar energy can be counteracted by a linked arrangement of spatially dispersed energy converter systems. The correlation of radiation fluctuations at various locations decays to a first approximation according to $\exp(-a \cdot r/T)$, where r is the distance between two converters, T is the period of the fluctuations, and a is a constant [6]. (This formula applies to roughly homogeneous climatic regions and periods of less than a few hours). In real large-area networks (1000×1000 km) random variations in the hour range are consequently largely statistically compensated. A precondition for this is of course a bidirectionally operating electricity grid. The remaining fluctuations could be covered, at least statistically, by energy-weather forecasts [7]; load management and the control of complementary, for example, fossil fueled power-stations, could thus be considerably simplified.

In situations of high shares of fluctuating solar electricity in utility grids (in central Europe from 20 – 30 % [4]), storage systems or extensive load management techniques are required to trap the solar energy production peaks. In ad-

dition to water power reservoirs, hydrogen systems in particular are suitable for this purpose [8].

A large-scale energy distribution is difficult to implement in the heat supply sector. On the other hand, in this sector the construction of storages ranging from small day reservoirs up to seasonal reservoirs is technically more favorable than in the electricity supply sector. Storages based on sensible heat (water) belong to the state of the art, whereas chemical heat reservoirs (e.g., sorption and phase transition reservoirs) are in the developmental stage [9].

On the basis of the measures outlined above for compensating or overcoming solar radiation fluctuations, future energy supply systems – including those for industrialized countries in central latitudes – are conceivable, which utilize for example the following energy sources [4]:

1) Rational energy utilization 40 %
2) Renewable energy sources 30 %
3) Utilization of fossil energy sources 30 %

The figures given are simply percentage values for scenarios that appear reasonable at the present time, and do not represent limiting values. In particular, the residual utilization of fossil energy sources could be reduced still further if large-scale energy storages or extensive load management strategies were used.

2. Solar Thermal Energy

2.1. Development and Historical Aspects

In the early advanced civilizations, the sun was worshipped as a divine being who provided light and warmth. Around 700 B.C. vestal virgins lit the sacred fire by using focused sun rays and the Incas prepared their holy meal with the help of simple reflectors. About 3000 years ago, the ancient Egyptians illuminated the underground passages and burial chambers of their kings with sunlight by using complicated mirror systems. Solar architecture originated in Greece: The philosopher SOCRATES (399 B.C.) imagined his ideal house to be cool in summer and warm in winter. While planning residential buildings, the Greek town planners and architects in the 4th and 5th century B.C. took into account that the path of the sun in summer is different to that in winter. They also constructed the heat storing masses required for the passive utilization of solar energy in the form of solid inner walls made of clay bricks.

The Greek philosopher ARCHIMEDES (287 – 212 B.C.) is supposed to have set fire to hostile Roman ships in 212 B.C. in the harbor of Syracuse by concentrating sun rays with mirrors. It was not until the 17th and 18th century that physicists, mathematicians, and engineers started scientific experiments on the utilization of solar energy. The renaissance of the focusing mirror began with ATHANASIUS KIRCHER (1601 – 1680) who in 1639 proposed that focusing mirrors with large focal lengths could be constructed from several plane mirrors. In 1686, the German mathematician E. W. VON TSCHIRNHAUS (1651 – 1708) successfully produced metal mirrors of ca. 160 cm diameter from polished sheet copper. With these mirrors, he was able to boil water, melt tin and lead plates in a few minutes, and vitrify bricks and clay pieces. In the 18th century, the French scientist G. L. LECLERC CONTE DE BUFFON (1707 – 1788) conducted extensive experiments with large reflectors of different sizes made from ca. 130 to 360 individual plane mirrors. To prove ARCHIMEDES's experiment, he used his mirror system to burn wood from a distance of 50 to 60 m. The next hundred years saw the continuation of experiments with focusing mirrors and the construction of various types of solar ovens.

In Tours in 1866, the French mathematics professor A. B. MOUCHOT was successful in realizing the first solar thermal power plant (a steam engine driven by solar energy). In the years that followed, he improved the reflectors and receivers and developed automatic tracking systems. At the World's Fair in Paris in 1878, MOUCHOT and his assistant A. PIFRE presented a printing press which was driven by solar energy. Sunlight was concentrated on an evaporator by using a concave mirror with a diameter of 5 m. The steam formed was expanded in a steam engine which drove the printing press (Fig. 5).

The Swiss natural scientist H. B. DE SAUSSURE (1740 – 1799) constructed the first hot box in 1767. It consisted of five glass boxes of different sizes that were placed one inside the other on a black wooden table and exposed to the sun.

He measured a temperature of 87.5 °C, and in later experiments 110 °C, in the innermost and smallest glass box. DE SAUSSURE did not invent the complete flat-plate solar collector because he did not utilizably withdraw the heat from inside the box. The hot box was further developed by other researchers in the 19th century and used to cook fruit and meat. In 1872, the English engineer C. WILSON distilled saline water with glass-covered vessels, in the hot sun of Las Salinas (Chile), this system produced about 19 000 L of fresh water per day.

The French engineer C. TELLIER (1828 – 1913) used a solar collector for the first time to run a solar thermal pump. The heat carrier ammonia was evaporated in ten flat-plate collectors which were made of iron absorbers placed in glass-covered wooden boxes. In 1904, the Americans H. W. WILLSIE and J. BOYLE continued with the idea of a "collector motor" and developed a solar machine driven by ammonia vapor with an output of 6 hp (4.5 kW).

It is not possible to say who was the inventor of the first solar system for the *heating of water*. In 1891, C. M. KEMP of Baltimore, Maryland applied for a patent for a solar collector for the preparation of hot water. He called the first integrated collector storage system "Climax". It consisted of four cylindrical water containers which were painted black and placed side by side in a glass- covered wooden box. The patent was sold in California in 1895. Despite several further developments, the problem remained that after a sunny day the water in the storage tanks was cold the next morning. This problem was solved only in 1909 when W. J. BAILEY built a solar system with a flat collector and a separate storage tank in a small suburb of Los Angeles. These early systems utilized the thermal buoyancy of the water heated in the collector for the transport of heat. Water was directly passed through the absorber. The absorber consisted of copper tubes with a diameter of 16 mm which were soldered parallel to each other and side by side on a copper plate. The 120 L storage tank was heat insulated with pulverized limestone. After the first frost damages, these single- cycle systems were further developed into two- cycle systems. A mixture of alcohol and water flowed through the collector tubes and indirectly heated the water in the storage tank via an internal heat exchanger. Solar water heaters were sold in larger numbers in California and later in Florida. However, electricity and cheap fossil fuels in the 1930s and 1940s brought a temporary end to this boom. In the 1950s and 1960s, solar technology was further developed in Israel and Australia (e.g., selective absorber layer) and cost-effective thermosiphon systems as well as collector storage systems were sold in large numbers. In Europe, the United States, and Japan, increased attention was paid to the technical utilization of solar energy after the oil crisis of 1974.

For further details about the historical aspects and the pioneers of solar technology, see [10–14].

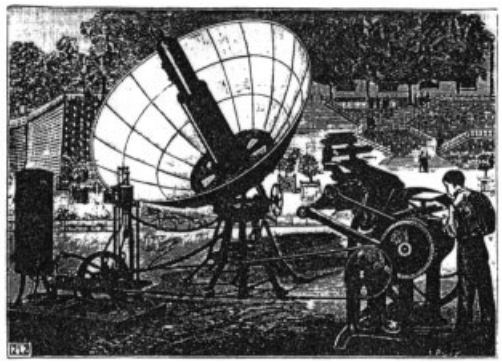

Figure 5. The solar driven printing press of A. B. MOUCHOT and A. PIFRE (1878)

2.2. Solar Thermal Energy Conversion

The energy released by the sun reaches the Earth in the form of electromagnetic radiation in the wavelength range of ca. 280 to 4960 nm. The spectral distribution approximately corresponds to a black body with a temperature of about 5800 K (Fig. 6). About 90 % of the extraterrestrial radiation is in the range of visible light and near infrared (300 – 1500 nm). The intensity of the radiation at the outer layer of the Earth's atmosphere is on average 1353 W/m^2 (solar constant) with fluctuations of about ± 3 %. The radiation striking the Earth's surface is diminished by absorption, scattering, and reflection in the atmosphere. The maximum values for the radiation measured at sea level are about 1000 W/m^2. The radiation available through an entire year is of importance for technical solar utilization.

The total radiation striking a horizontal surface is called the "annual sum of the global radiation." Figure 7 shows the worldwide distribution of the average annual global radiation, the values being between 800 and 2250 kWh m^{-2} a^{-1} depending on the place and climate. Some typical examples are presented in Table 1. The annual sum can also be characterized by an average power thought to be constant through all the hours of the year. These fictitious values and the annual useful energy attainable for a system efficiency of 35 % (typical value for solar thermal utilization) are given in Table 1.

Table 1. Typical insolation data for various cities

City	Annual sum of global insolation, kWh m^{-2} a^{-1}	Average power, W/m^2	Attainable energy for $\eta = 35\,\%$, kWh m^{-2} a^{-1}
Stockholm	920	105	322
Hamburg	978	110	342
Munich	1140	130	399
Rome	1500	170	525
Almeria	1750	200	613
Cairo	2100	240	735
Barstow (USA)	2250	260	788

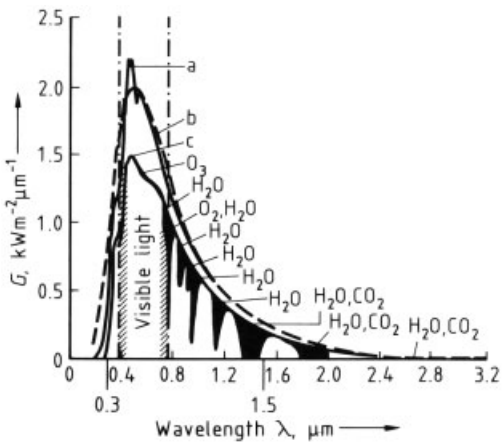

Figure 6. Spectral distribution of the extraterrestrial solar radiation (a) and emission of a black body (b) at 5762 K [15]; (c) Solar radiation at sea level

Global radiation consists of two different parts. One part comes directly from the sun without a change in direction (*direct radiation*). The other part comes from all directions after single or multiple deflections (*diffuse radiation*). The direct radiation can be directly concentrated by reflection and refraction using mirrors and lenses, respectively. The diffuse radiation can be concentrated to only a very limited extent. For this reason, direct radiation is qualitatively more valuable than diffuse radiation. The energy density is increased by the concentration of radiation, which, in turn, results in higher temperatures on absorption. The highest possible degree of concentration is about 46 000 and leads to the maximal temperature of ca. 5800 K produced in a collector by solar radiation.

2.3. Technical Utilization of Solar Energy

The temperature requirements determine the main features of the construction of a collector. The complexity increases with increasing temperature demand. For low temperature applications the *flat plate collector* is used. Simple absorbers without transparent covers are sufficient for temperatures below 30 °C. For temperatures up to 100 °C collectors with one or two transparent covers and a selective absorber surface are used in practice. To provide considerable amounts of energy above 150 °C the flat plate collector must be evacuated. Medium- and high-temperature applications require temperatures higher than 150 °C. Above this limit sufficient efficiencies are only reached by *concentrating collectors* that concentrate the rays of the sun to the absorber. For highly concentrating devices only the direct part of the radiation can be used, for lower degrees of concentration the diffuse radiation can also be utilized in part.

2.3.1. Low-Temperature Applications

The solar collector will be described in detail. Since the supply of solar energy does not coincide with the demand for heat for room heating and water heating, the collected solar energy must be stored. This energy storage will be described in → Heat Storage Media. For solar systems in the low-temperature range, there is, in practice, no alternative to the storage medium water.

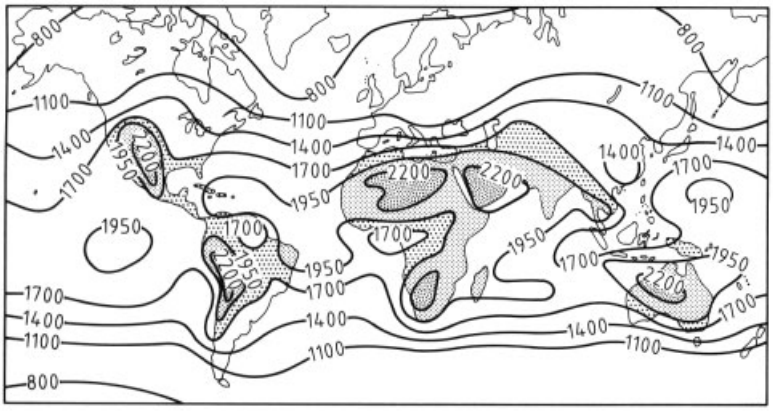

Figure 7. Global radiation chart of the world with isopyrene [16]

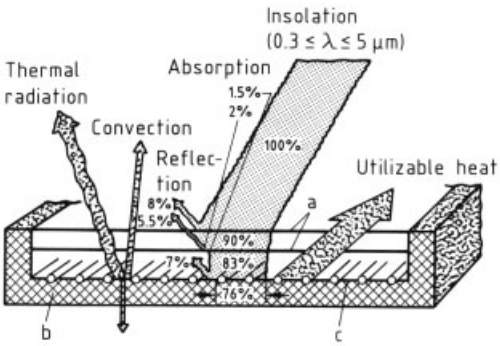

Figure 8. Basic design of a flat collector and photothermal conversion
a) Covering; b) Thermal insulation; c) Absorber plate

2.3.1.1. Design and Function of Flat Collectors

The basic design and physical transport processes of a flat-plate collector are shown in Figure 8. The most important component is the absorber (c) with a surface that is as completely absorbing for solar radiation as possible. There should be the best possible thermal contact between the absorber plate and a heat-transfer medium (e.g., water, oil, or air) that flows past the plate. The heat generated on absorption is given off to this medium and, thus, withdrawn from the collector (utilizable heat in Fig. 8). To reduce heat losses and improve collector efficiency, the absorber is provided with one or several transparent covers (a) at the front, the sides and back being heat insulated (b). Figure 9 shows various types of liquid- cooled flat collectors which are not described here in detail.

The *absorber* usually consists of a good heat conducting metal (e.g., copper, aluminum). The absorber should have the highest possible absorptance α in the spectrum of solar radiation (α is the ratio of the absorbed radiation to the total incoming radiation). Ideally, this requirement is met by a black body which completely absorbs radiation of all wavelengths λ [$\alpha(\lambda) = 1$]. Real absorber surfaces reach absorption degrees of 0.8 to 0.98. According to the Kirchhoff Law, $\alpha(\lambda) = \varepsilon(\lambda)$, i.e., a good absorber is also a good emitter of radiation (the emittance ε is defined as the ratio of the amount of energy emitted at one wavelength range to the amount that would be emitted by a black body at the same temperature). The spectral distribution of the heat radiation emitted by the absorber depends on its temperature and is defined by the Planck Law. Figure 10 shows that the solar radiation and the heat radiation of an absorber (ca. 80 °C in Fig. 10) cover different spectral ranges. This wavelength difference is made use of by selective absorber layers, which exhibit a high degree of absorption in the solar radiation range and a low degree of emission (same as degree of absorption) in the infrared range. The spectral degree of absorption of black (a) and white paints (c) and of a selective coating (b) is shown in Figure 11. Various selective absorber layers exist which can be classified as interference, semiconductor, and structure filters. In current practice the manufacturing process is shifting gradually from electroplating to sputtering. The values reached for advanced

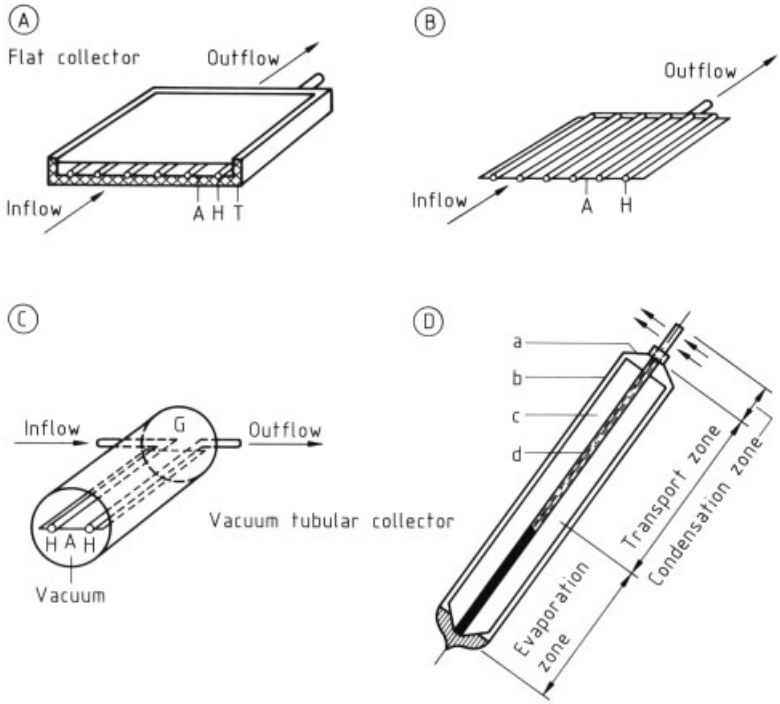

Figure 9. Types of liquid- cooled flat collectors
A) Flat collector; B) Absorber; C) Vacuum tubular collector (flow-through principle); D) Vacuum tubular collector with heat exchanger tube
a) Glass – metal connection; b) Evacuated glass tube; c) Absorber; d) Heat exchanger tube with working fluid
A = absorber; G = glass tubes; T = thermal insulation; H = heat carrier

layers are 0.95 for the absorptance and simultaneously 0.05 for the emittance.

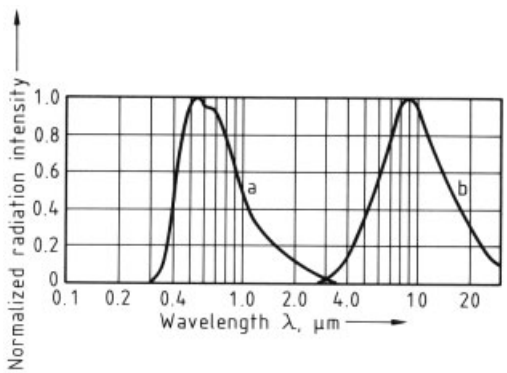

Figure 10. Normalized spectral intensity distribution for a black body at 5800 K (solar radiation, a) and 350 K (heat radiation of the absorber, b)

The *covering* of flat-plate collectors consists of low-iron glass or transparent plastics in the form of stretched films or capillary structures (honeycombs or transparent heat insulation). The covering should have the highest possible transmission ratio in the spectrum of solar radiation, but it should also be as impermeable as possible to the heat radiation of the absorber. Glass meets these requirements very well, while plastic sheets [e.g., polyethylene, poly(vinyl fluoride)] exhibit optical windows in the range of heat radiation (see Fig. 12).

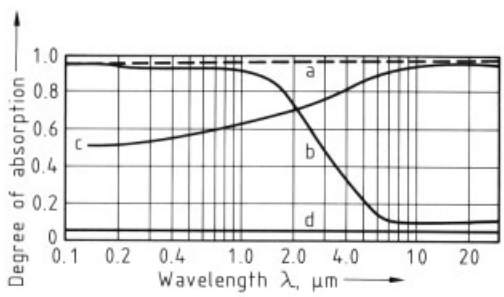

Figure 11. Spectral course of the degree of absorption of various surfaces
a) Black paint; b) Highly selective coating; c) White radiator paint; d) Mirror

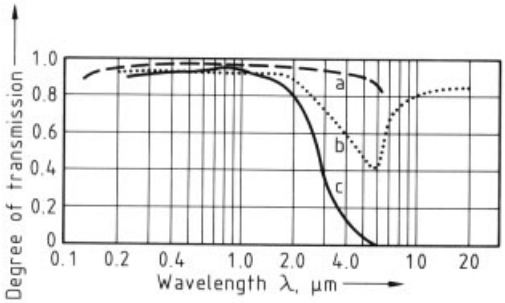

Figure 12. Spectral course of the degree of transmission of various transparent materials
a) Quartz glass; b) Polyethylene film; c) Normal window glass

The transmission of glass panes can be further improved by a treatment of the surface. By etching, the refractive index of the surface layer is reduced to lower the reflectance and increase the transmittance. Structuring of one side of the pane also results in an increased transmittance. Both effects are especially effective at low angles of incidence.

The heat attainable with a collector is described by the *efficiency* η. It is defined as the ratio of utilizable heat energy released by the collector to the incoming radiation energy. Under steady-state conditions, the following efficiency equation can be derived from the collector heat balance:

$$\eta = (\tau_G \alpha_A) - k_m(\theta_{A,m} - \theta_{amb}) / E_{glob,C}$$

where $(\tau_G\ \alpha_A)$ is the product of transmittance and absorptance (the optical collector efficiency, where G = glass and A = absorber); k_m the mean heat transfer coefficient between the absorber plate and the ambient air; $\theta_{A,\,m}$ and θ_{amb} the mean absorber and ambient air temperature, respectively; and $E_{glob,\,C}$ the global radiation in the plane of the collector. k_m increases slightly with temperature.

The collector efficiency is plotted versus the operating parameter $[(\theta_{A,\,m} - \theta_{amb})/E_{glob,\,C}]$. The typical course of the characteristic curve of collector efficiency for a certain incoming radiation, ambient temperature, and wind speed is shown in Figure 13. The *y* axis intercept corresponds to the optical efficiency $(1 - \tau_G\ \alpha_A)$ and the slope of the characteristic curve to the overall heat transfer coefficient (k_m). The curvature of the curve is due to radiative heat losses that increase over-proportionally with temperature (T^4).

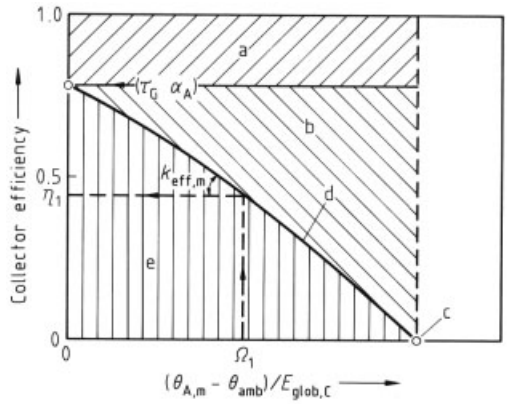

Figure 13. Characteristic curve of collector efficiency η
a) Optical losses/$E_{glob,\,C}$; b) Thermal losses/$E_{glob,\,C}$; c) Stagnation point, Ω_{stag}; d) Characteristic efficiency curve; e) Utilizable energy/$E_{glob,\,glob,\,C}$

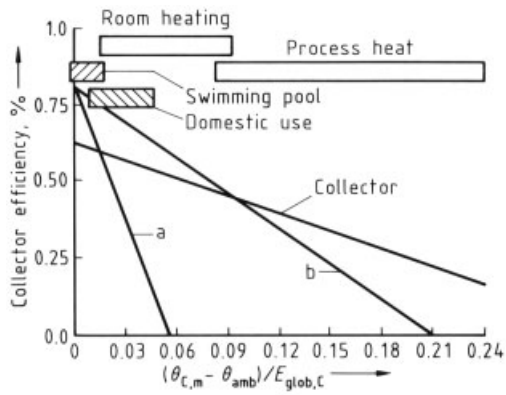

Figure 14. Typical characteristic curves of efficiency for an absorber (a), flat-plate collector (b, one glass cover, selective surface), and vacuum tubular collector (c)

For practical purposes the equation for the collector efficiency must refer to the mean fluid inlet and outlet temperature. Therefore, the above shown preliminary formula has to be changed, taking into account the heat resistance between the collector fluid and the absorber plate. The final result is given by

$$\eta = F'(\tau_G \alpha_A) - F' k_m(\theta_{C,\,m} - \theta_{amb}) / E_{glob,\,C}$$

where $F' = k_0/k_m$ with k_0 the heat transfer coefficient between the fluid and the ambient air. The typical characteristic curves of efficiency for an absorber, flat-plate collector (with one covering and selective absorber), and a vacuum tubular

collector are shown in Figure 14. Different fields of application (e.g., solar heating of swimming pools or water for domestic use result in varying ranges in the operating parameters. The type of collector to be selected for various applications to achieve the highest possible efficiency factor (same as useful energy) can be approximately determined from Figure 14.

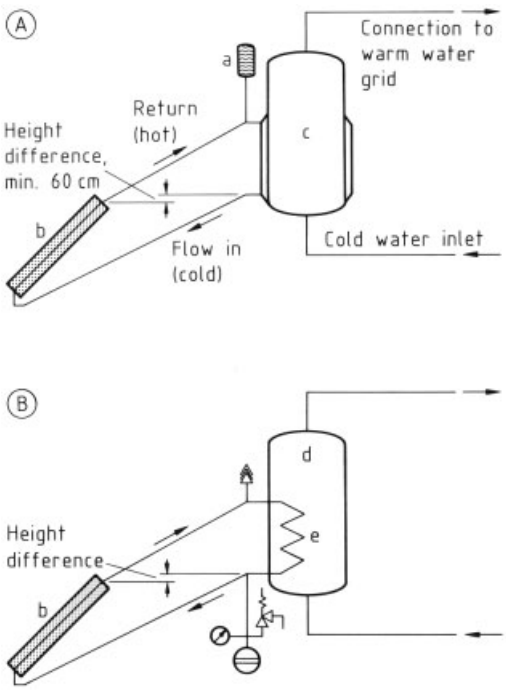

Figure 15. Natural circulation or thermosiphon system with open (A) and closed (B) collector circuits
a) Open expansion tank; b) Collector; c) Double jacket reservoir; d) Reservoir; e) Heat exchanger

2.3.1.2. Systems for Heating Water

A solar water heating system consists of a *collector field* (which can comprise one or, in the case of large systems, many collectors), one or several heat reservoirs, and, depending on the system, heat exchangers, piping, pumps, additional heating, as well as regulation and control equipment. A system functions satisfactorily only when all the components are well suited to each other. Knowledge of the warm water requirements (amount and time profile) and the local weather conditions is required for the dimensioning of a solar system. At present, the following types of system are common:

1) Thermosiphon or natural circulation system (Fig. 15)
2) Integrated collector – storage system (Fig. 16 A)
3) Forced-circulation or "pump" system (Fig. 17)

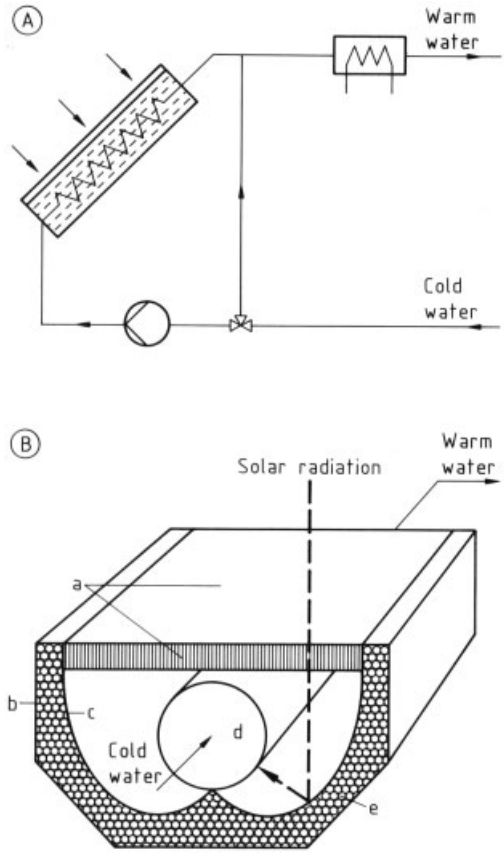

Figure 16. A) Schematic for the integration of an ICS system; B) Integrated storage – collector with transparent heat insulation
a) Transparent heat insulation; b) Housing; c) Mirror; d) Water reservoir; e) Opaque heat insulation

In the case of a *thermosiphon system*, the heat reservoir must be placed higher than the collector. These plants can be built as open (Fig. 15 A) or closed (Fig. 15 B) systems. The circulation of the heat carrier between collector and reservoir occurs as a result of temperature-based differences in the density of the medium. Since the heat carrier (water or water – glycol mixture) is heated in the collector, it is lighter than that in the reservoir. Circulation starts as soon as the

buoyant pressure is larger than the pressure loss of the entire cycle. Systems of this type can be cheaply built and easily assembled. They require no foreign energy (passive system) and are operated without control elements. Systems based on the natural circulation principle are predominantly used on a small scale (up to 10 m^2 of collector surface area) and are commonly seen on the flat roofs in Mediterranean countries. In Central and Northern Europe the elevated position of the reservoirs poses problems to their integration in the buildings. The Mediterranean arrangement is not accepted, mainly because of aesthetic reasons.

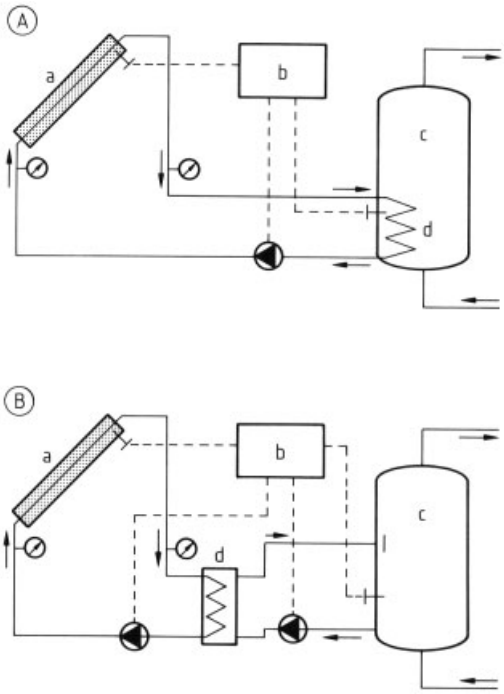

Figure 17. Schematic of a solar water heating system with internal (A) and external (B) heat exchanger in the collector cycle
a) Collector; b) Temperature control; c) Reservoir; d) Heat exchanger

In *integrated collector – storage (ICS) systems*, reservoir and collectors are accommodated in the same housing. The heat losses from the reservoir, especially through the transparent covering, must be kept as low as possible (see Section 2.1, "Climax" made by C. M. Kemp). This is necessary, to withdraw warm water for as long as possible after a sunny day and to protect the reservoir from freezing during cloudy winter days. An example of an ICS system with transparent heat insulation (capillary structure made of polycarbonate) as covering is shown in Figure 16 B.

In *forced- circulation systems*, the heat from the collector is transported to the reservoir by a heat carrier (water or water – glycol mixture for protection against frost damage). A pump is employed to circulate this medium. Instead of a heat exchanger installed in the reservoir (Fig. 17 A), which is usually the case in smaller systems, an external heat exchanger (Fig. 17 B) can also be used, as required in large systems. Forced- circulation systems have the advantage that reservoir and collector can be arranged in any manner. Simple regulation (temperature control) is sufficient to control the system. The circulating pump is switched on when the collector field can supply usable energy to the reservoir, i.e., when the temperature difference between collector and reservoir (at the level of the heat exchanger) exceeds a threshold value which can be set at the control unit. For the climatic conditions and normal warm water consumption (40 L at 50 °C per day and person) in Germany, the following design can be recommended for a solar supported water heating system: collector area of 1.0 to 1.5 m^2 per person and water storage volume of 50 – 80 L per square meter of collector area.

The literature covering collectors and thermal solar systems is extensive. The textbooks by Duffie and Beckman [17] and Kreith and Kreider [18] should be mentioned in particular.

2.3.1.3. Systems for Room Heating

In Central and Northern Europe, relatively large parts of the total final energy consumption (e.g., 30 % in Germany) are used for the heating of buildings. Long-term storage is essential because the seasonal course of the solar energy supply and the consumption of heat for room heating move in opposite directions. Therefore, cost-effective storage systems that store the energy from the summer into the winter are necessary. Increased constructional thermal insulation of buildings (e.g., heating energy index of 40 – 50 $kWh\,m^{-2}\,a^{-1}$) is an economic requirement for the active utilization of solar energy for room heating. Numerous attempts have been

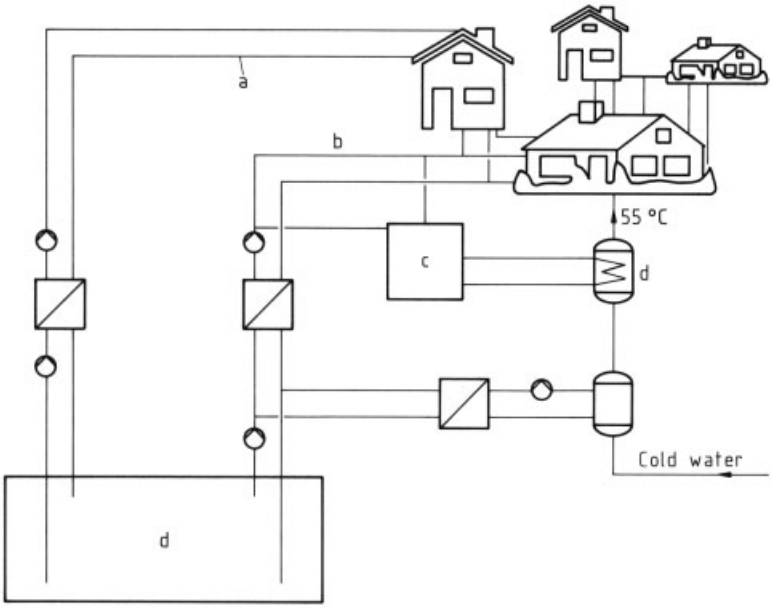

Figure 18. Solar-supported district heating system
a) Collector pipe; b) Heating water distribution grid; c) Heating boiler; d) Warm water distribution grid

made to heat individual houses mainly or even completely with solar energy. However, these normally oversized solar systems can only be recommended if only small contributions to the heat demand are required. Solar-supported district heating systems are technically and economically more suitable. Figure 18 shows the concept in which a larger number of houses are connected to a district heating system. A central long-term heat reservoir (d) is charged to temperatures of up to 95 °C in summer by decentralized collector fields on the roofs of the buildings. In the heating period, the stored heat is transported from the reservoir (in the form of warm water) to the buildings via underground piping. Pilot plants, primarily in Sweden, have shown that this concept functions and the investment costs could be reduced by more than half compared to small solar systems. International research [19] has shown that reasonable designed systems with a seasonal heat storage are able to cover 60 – 80 % of the heat requirements. In this case it is essential that at least 100 to 200 residential units are connected so that a storage volume of several 1000 m^3 becomes necessary. This storage volume is required to reduce the heat losses from the reservoir and to reduce the specific building costs. Figure 19 shows the storage concepts studied in pilot projects which use water, rock, or soil as the storage medium.

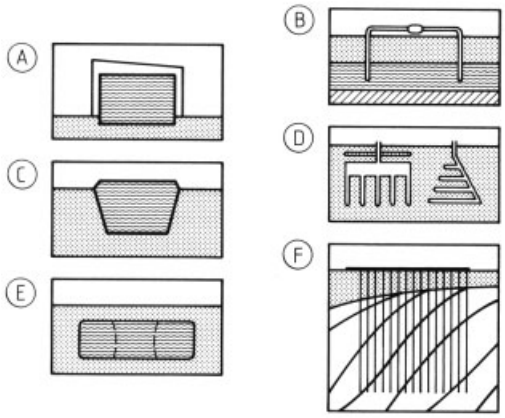

Figure 19. Storage concepts for the long-term storage of low-temperature heat
A) Tank; B) Aquifer; C) Ground water pit; D) Horizontal pipe coils, vertical soil sondes; E) Cavern; F) Vertical rock boreholes

Solar heating plants with collector areas up to 8000 m^2 (Marstall, Denmark) and storage volumes for hot water of 100 000 m^3 (rock cavern, Sweden) have been build successfully. Actual

developments concentrate rather on a decrease of cost than on an improvement of technique [20]. Some collectors are equipped with a prefabricated roof module where the wooden roof beams are already integrated in the collector. This collector is installed instead of the conventional roof in new buildings. The additional cost is about 100 Euro per square meter collector. One part of the storage research concentrates on cost-effective constructions that do not depend on special geological conditions such as rock caverns. Problems encountered with well known and cheap materials like ordinary concrete are their high coefficient of vapor diffusion and their sensitivity to temperature stress.

Solar-supported district heating has a very high potential. In Germany more than one third of the total heat demand for heating and hot water supply could be met by solar-supported district heating, even if stringent ecological and economic constraints are taken into consideration. In future cost of seasonal storages and large collector fields will decrease. Taking this into account, the solar potential can be exploited at specific costs ex heat reservoir between 0.03 and 0.10 Euro/kWh [21].

2.3.1.4. Other Low-Temperature Applications

Apart from the heating of water, solar collector systems can, in principle, be used wherever heat for domestic use is required at low temperature ($< 150\,°C$). Despite the current low prices of gas and oil, the solar heating of *open-air swimming pools* is an application that is almost profitable. Inexpensive plastic absorbers (e.g., made of polypropylene or ethylene – propylene – diene rubber) without transparent covering and without heat insulating on the back are used. The pool water flows directly (without a heat exchanger) through the absorber and a monovalent mode of operation (i.e., without a conventional heating system) is possible. The typical system costs are in the range of € 51 (\$ 56) to 102 (\$ 111) per m^2 of absorber and the solar energy yields are between 200 and 250 kWh/m^2 per swimming season.

With the help of air collectors, *warm air* is produced which can be used to dry agricultural or other products. Application of warm air in air heating systems is also possible because of the relatively low temperature level needed. Rock beds or the inner walls, floors, and ceilings (Hypocaustum system) of the building serve as heat storage material. A simple and cost effective system was developed in Canada (Solarwall, Toronto). In this system, the collector consists of a perforated, dark colored facade (ca. 2 % of holes) through which external air is sucked and heated.

Another interesting application is the *thermal desalting of sea water* with simple collector systems. In the simplest case, sea water is heated under a transparent covering by solar radiation and partially evaporated. The water vapor condenses on the inclined covering and runs into a collecting gutter. Large-scale solar distillators were erected rather early. More recent plants are to be found, for instance, on various Greek islands, e.g., Patmos, where the favorable radiation conditions permit the production of 26 000 L of fresh water per day with ca. 8760 m^2 of distillation area.

Highly efficient flat-plate collectors (e.g., vacuum tubular collectors), which achieve temperatures between 100 and 150 °C with efficiencies of 30 – 40 %, are suited to the operation of *continuous absorption refrigeration plants* which use the working fluids lithium bromide – water. Apart from these continuous systems, periodically operating adsorption systems, e.g., based on calcium chloride and ammonia, also exist. The refrigerant (ammonia) is expelled from the fixed-bed absorbent (calcium chloride) with the help of solar energy.

Another application of highly efficient collectors is the *solar thermal pump*. Refrigerant vapor is produced (110 – 130 °C) by using collectors and expanded in a piston motor which drives a water pump. Prototypes (Dornier, Friedrichshafen) are available with a hydraulic pump output of ca. 1 kW and a degree of utilization (daily average) of ca. 3 %.

2.3.2. Medium-Temperature and High-Temperature Applications

2.3.2.1. Concentrating Collectors

Solar radiation must be concentrated to produce heat at high temperature. This concentration is

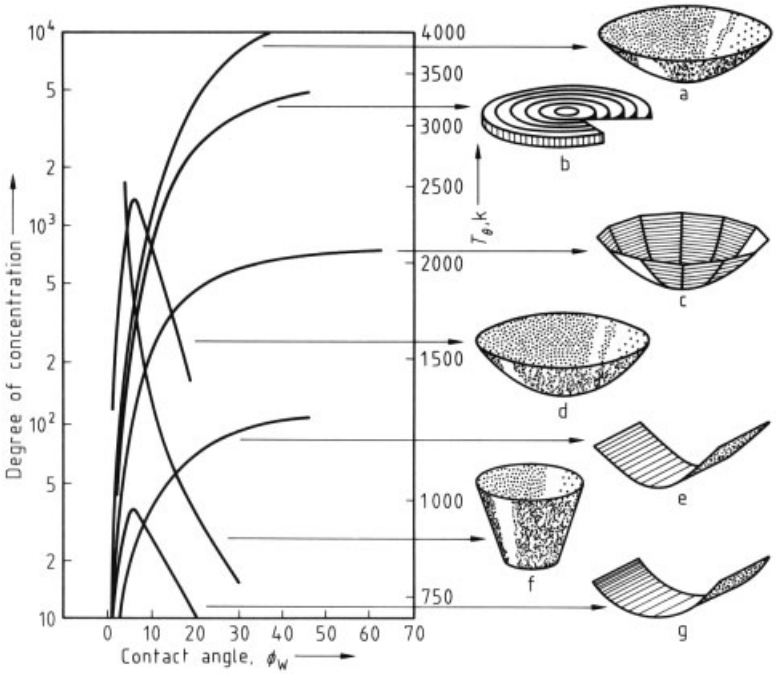

Figure 20. Concentration degrees of various reflectors and the maximum theoretical absorber temperature as a function of the contact angle [25]
a) Paraboloidal mirror; b) Fresnel lens, $n = 100$; c) Screen mirror, $n = 100$; d) Spherical mirror; e) Paraboloid-type cylinder mirror; f) Conical mirror; g) Circular cylindrical mirror
n = Number of segments

generally achieved with mirrors. The attainable temperature depends on the concentration degree C by which the incoming radiation is intensified (C is the ratio of the aperture area of the optical device to the area of the solar image in the focal plane). The degrees of concentration, in turn, depend on the quality of the optical components and on the type of concentration. For various types of mirrors (reflectors), the theoretical concentration degrees and maximal absorber temperatures as a function of the contact angle are shown in Figure 20. For comparison, the curve of a Fresnel lens is also shown. The maximum theoretical absorber temperature T_A is calculated according to the following equation: $T_A = T_S\ (C/C_{max})^{1/4}$ where $C_{max} = 46\,211$ and $T_S = 5762$ K. In the case of one-dimensional concentration, the radiation is collected at the focal line. To attain higher degrees of concentration, two-dimensional concentration is necessary. In this case, the radiation is collected at the focal point [17], [18], [23], [24]. The degrees of concentration technically obtained today lie between 40 and 80 in the case of linear concentrators (e.g., parabolic troughs). Values of up to 2500 are achieved with concentrators that are two-dimensionally effective (e.g., paraboloidal mirrors). Tower power plants (see Section 2.3.2.2) with biaxial tracking heliostats have been built with $C = 200$ to 700.

The schematic design of a *concentrating collector* is shown in Figure 21. The collector consists of two parts, the reflector that focuses the radiation and the absorber. The two parts can be firmly joined together and can follow the sun together or they can be separated, as in the case of tower power plants (see Section 2.3.2.2). The incident rays must always be kept focused to the absorber. Therefore, in contrast to flat-plate collectors, tracking is required in the case of concentrating collectors. *Biaxial tracking* is necessary for paraboloidal mirrors. For parabolic troughs that are mounted horizontally *monoaxial tracking* is sufficient to obtain good results. Two alignments of the trough are possible. For locations near the equator the north-south direc-

tion and for higher latitudes the east-west alignment give the highest energy yields.

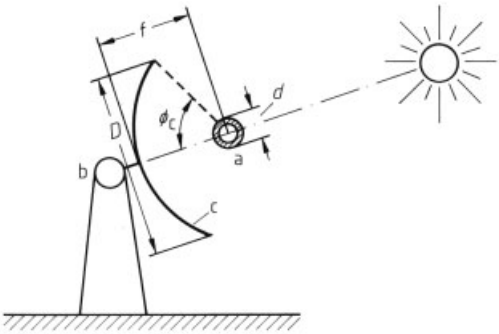

Figure 21. Basic design of a concentrating collector
a) Absorber; b) Tracking device; c) Paraboloid reflector D = Aperture, m; ϕ_c = Contact angle, °; f = Focal length; d = Absorber diameter

The absorber in the focal plane of a concentrating collector should absorb the solar radiation, convert it to heat as completely as possible, and transfer it to the fluid heat carrier. According to the constructional design, a distinction is made between *open receivers* (Fig. 22 A), with and without transparent covering (e.g., an evacuated glass tube around cylindrical absorber tubes in the case of parabolic trough collectors) and *cavity receivers* (Fig. 22 B) which have only a small opening for radiation. Another concept uses so-called "volumetric absorbers" made of highly porous materials. The concentrated solar radiation is absorbed within the structure ("in the volume") and the heat is transferred to the medium that is passed through the absorber.

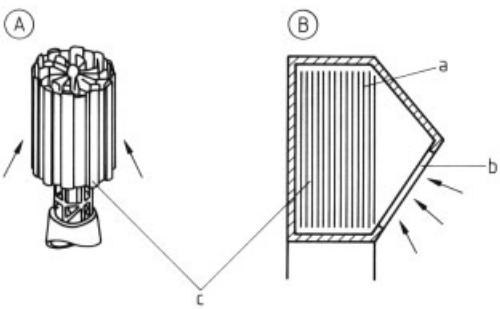

Figure 22. Types of absorbers for solar tower power plants
A) Open absorber; B) Hollow-space absorber; a) Absorber tubes; b) Aperture; c) Heat exchange tubes

In a concentrating collector, the most important losses include:

1) Incomplete reflection of the mirror
2) Surface faults on the reflector
3) Orientation errors caused in tracking
4) Reflection at the absorber
5) Heat losses (radiation, convection) at the absorber

2.3.2.2. Solar Thermal Power Plants

The indirect conversion of radiation energy to electricity via the production of thermal and mechanical energy occurs in solar power plants. The thermodynamic cycle corresponds to that of a conventional steam or gas turbine power plant. A distinction is made between farm and tower plants, depending on the type of collector employed. In a *farm plant*, each collector has its own absorber. It uses either line focusing parabolic troughs or point focusing parabolic dishes. The parabolic troughs are connected to a central machine unit (Rankine cycle) via a piping system that is usually operated with the heat transfer medium thermooil at a maximum temperature of 400 °C (Fig. 23). For the parabolic trough no heat transfer is necessary if the absorbed energy is directly transformed to electricity in a Stirling engine that is mounted near the focus.

Solar tower plants possess many biaxial mirrors (heliostats) that follow the sun. The mirrors reflect the solar radiation onto a common central absorber (receiver) which is on top of a tower (Fig. 24). A heat transfer medium (steam, air, molten salt) transports the solar heat to the machine unit, usually a Rankine cycle.

Central receiver plants using closed (pressurized) volumetric receivers can drive gas turbines (Brayton cycle) or combined cycle power plants. The high conversion efficiency of combined cycles will allow for solar-to-electricity efficiencies of over 25 %. This technology is presently in the R&D phase.

Since 1973, about 35 solar farm power plants with ca. 1 885 000 m^2 of mirror area and an electrical output of ca. 300 MW have been built. The farm and tower plants (500 kW electrical output each) in Almeria, Spain studied in the course of an international research and development collaboration are shown in Figure 25. Between 1985 and 1989, a series of large-scale farm plants with an electrical output of 30 and 80 MW

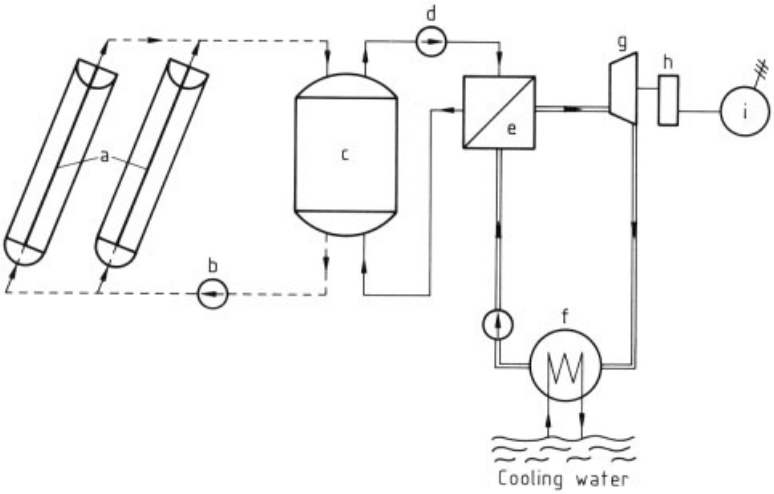

Figure 23. Schematic of a solar plant based on the farm concept
a) Collector field (parabolic troughs); b) Field pump; c) Reservoir; d) Reservoir pump; e) Evaporator; f) Condenser; g) Screw expander; h) Gear; i) Generator

were put into operation in California's Mojave desert. The net annual efficiency factors of these plants are around 9 % and the investment costs between \$ 3500 and 4000 per kW_e. Presently projected plants of this type will have annual conversion efficiencies of 13 – 15 % and a cost of \$ 2000 – 3000 kW_e. The combination of parabolic troughs and combined cycles can lead to investment costs of less than \$ 1000 kW_e. Up to now only six tower power plants have been built with a total output of about 16 MW. At present, the largest power plant based on the tower principle (Solar Two) with an electrical output of 10 MW is also to be found in the Californian desert near the city of Barstow. The net annual efficiency factor is ca. 6 % and the specific investment costs of this pilot plant are ca. \$ 14 000 per kW_e. Presently projected plants of this type will have an annual conversion efficiency of 13 – 15 % and costs of \$ 3500 – 4500/kW.

2.3.2.3. Solar Process Heat

The temperature distribution for industrial process heat in Germany is shown in Figure 26. It is obvious that there are two foci for the heat requirements, in the temperature ranges 100 – 400 °C and 1200 – 1500 °C. Temperatures up to 400 °C can be covered by parabolic trough collectors. Tower receivers cover up to 1200 °C, temperatures of 3000 °C are technically possible with biaxial tracked paraboloidal mirrors. The applications range from simple solar cookers for developing countries (to reduce the consumption of firewood) to solar melting furnaces. The largest solar furnace having a thermal output of 1000 kW was built in 1970 in Odeillo in the French Pyrenees (see Fig. 27). The maximal attainable temperature is almost 4000 °C.

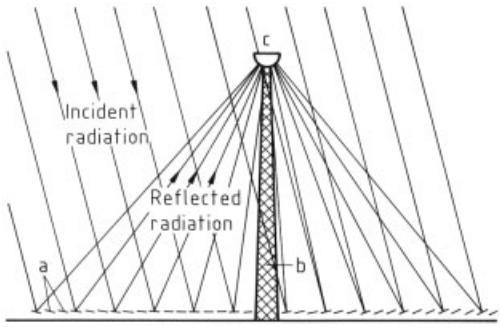

Figure 24. Principle of the tower power plant
a) Heliostats; b) Tower with evaporator; thermal power machine and generator; c) Absorber

3. Passive Utilization of Solar Energy

3.1. Introduction

In recent times, highly developed air-conditioning technology has increasingly been

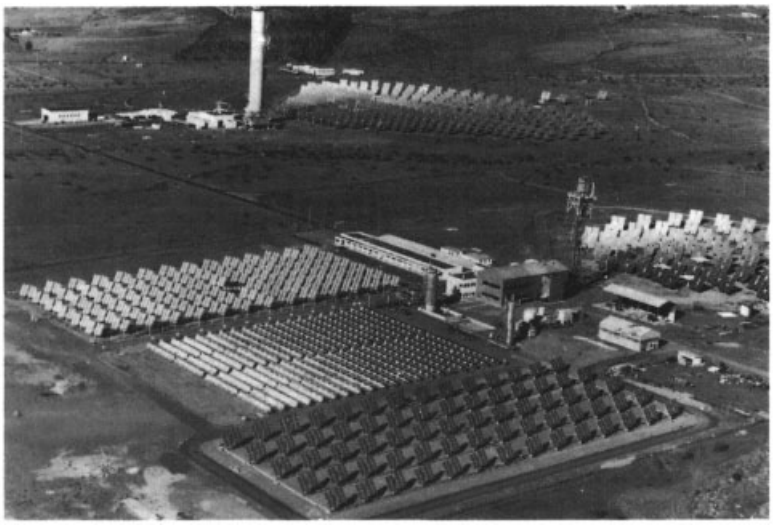

Figure 25. Pilot power plants in Almeria (Spain) based on the farm and tower concept

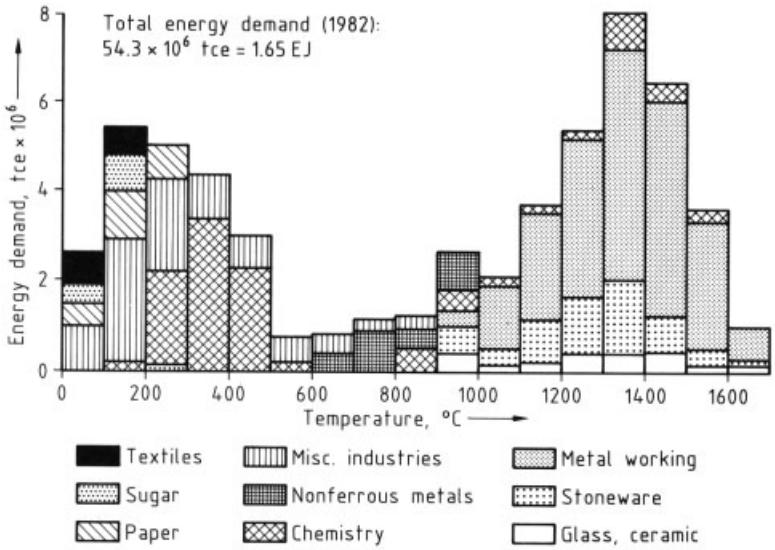

Figure 26. Typical temperature distribution for process heat in Germany (1982) [24]
tce = tons of coal equivalent

used to provide thermal comfort, without asking whether the building substance has been constructed in accordance with the climatic conditions. In contrast, it is more meaningful to think about the traditional methods of construction which arose through optimal adaptation to the climate of a particular region. Buildings of this type, which can, of course, be constructed by using modern technology and materials, have very low energy requirements. They conserve resources of fossil fuels, contribute to environmental protection, and are also more economical.

The most important measures taken to minimize energy losses in buildings and the possibilities for the passive utilization of solar energy are described in this chapter. The radiation gains in solar buildings of the type shown in Figure 28 can be substantial. This house is characterized primarily by a pleasant climate and low heat re-

quirements. In other cases, especially in commercial buildings, a reduction in electrical lighting by the use of daylight and a decrease in ventilation or cooling costs can also be important. Passive solar measures are most efficient if they are considered during the planning stage of the building. However, even in the case of existing buildings, the heat balance can be substantially improved through renovations. An example of a successful renovation of a multifamily dwelling is shown in Figure 29. The facade of this house was changed into a solar facade.

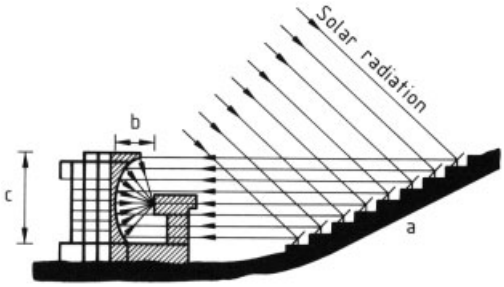

Figure 27. Schematic of the radiation path in the solar furnace in Odeillo
a) Movable plane mirror; b) Focus, 18.40 m; c) Concave mirror

Figure 28. South view of the ISFH experimental building in Emmerthal, Germany
Right: solar house with conservatory and transparent insulated walls. Left: reference building

Passive utilization of solar energy refers to the direct transformation of radiation energy into heat by the structure of the building, i.e., by means of transparent and storage building components. In comparison with active measures (e.g., heating of water for domestic use by using collectors, see Chap. 2), none or very little auxiliary energy is required for additional systems. The principles involved in the passive utilization of solar energy are: collection of radiation energy, storage of the collected energy, and utilization of the stored energy in a controlled manner. Windows, sun spaces, solar walls, and air collectors represent the most important components. Several aspects have to be considered before selecting passive components. These include location factors, shape of the building, orientation, room arrangement, ventilation, and insulation.

Figure 29. A renovated multifamily dwelling building in Freiburg, Germany
The building has been provided with transparent insulation (Project leader: Fraunhofer Institute of Solar Energy Systems, Freiburg.)

3.2. Building Design

3.2.1. Site Analysis

An energy-conscious building design requires an exact analysis of the climatic factors and the conditions of the site. The most important climatic factors are solar radiation, temperature, humidity, wind, and precipitation. Average values determined over longer periods of time can be obtained from the weather service or from the relevant literature. The annual average temperature and solar radiation measured in Hamlin (Germany) are shown in Figure 30. However, the architect must also take the local conditions (microclimate) into account. In the heating period, shading of the building can considerably reduce the utilizable solar radiation. In summer, shading can help to protect the building from overheating. The most favorable location for the utilization of solar energy is an incline facing the south, although the free selection of a building site is seldom possible. Periods of shade (ob-

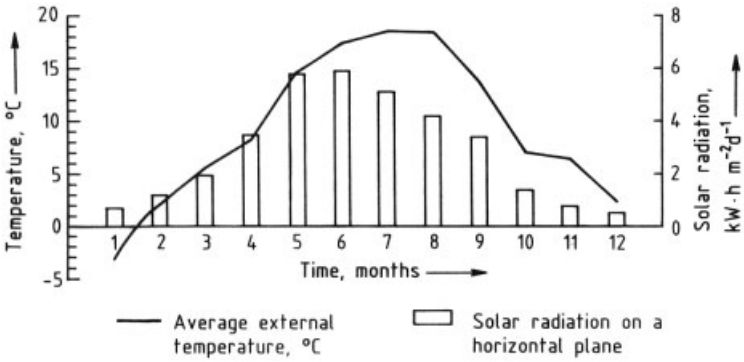

Figure 30. Average monthly temperature and solar radiation (global, horizontal) in Hamlin, Germany

struction profiles) can be determined with a solar chart (Fig. 31). The topology of the site, the vegetation, and the buildings in the neighborhood have a strong influence on the speed and direction of the wind. Wind enhances transmission losses by cooling the building envelope and increases infiltration through leaks. For this reason, wind barriers reduce the heat requirement during the heating period.

3.2.2. Design

A compact building has a smaller outer surface area and fewer heat losses than a strongly structured building with the same volume. When designing the building, this must be taken into account by providing the smallest possible surface-to-volume ratio.

Passive solar buildings should be built as far as possible in the east – west direction. They should have large apertures toward the south. Windows in the south facade achieve the highest solar gains so that their annual average total heat balance is positive.

Rooms that are not continuously lived in and are not heated can be located as buffer rooms in the temperature gradient between the inside and the outside (zoning). Staircases, garages, storerooms, and porches can be considered for this purpose. Buffer rooms exert their thermal effects on all sides of the house.

Furthermore, efforts should be made to provide effective daylighting and uniform heating by means of natural convection and ventilation. A cross section of the solar houses shown in Figure 28 is presented in Figure 32. In these houses, the principles of zoning and the creation of buffer zones have been consistently applied. The location on a slope facing the south was exploited by a split-level design.

A reduction of the outer surface area can also be achieved by combinations to form larger living units. In comparison with detached one-family houses, the room-heat requirements of one unit in a residential building with several stories can be reduced by 50 %.

3.2.3. Heat Conservation

The utilization of solar energy in buildings only makes sense when basic measures of heat conservation have been taken. The heat losses in a building are made up of transmission losses through the building envelope and infiltration losses. Therefore, good insulation of the outer envelope represents the highest energy saving potential; attics and floors above unheated cellars are to be included in the heat conservation concept. The thermal transmittance (U value) of the walls should be less than $0.3\,\mathrm{W\,m^{-2}\,K^{-1}}$. It is advantageous to equip windows with heat-conserving glass.

The exchange of air during the heating period means a continuous removal of heated air and, consequently, energy. Unintentional ventilation through badly fitting windows, doors, and cracks, etc., must be avoided. The necessary supply of fresh air can be controlled more efficiently by means of a ventilation system than by opening the windows. The energy consumption can be reduced still further by means of a heat re-

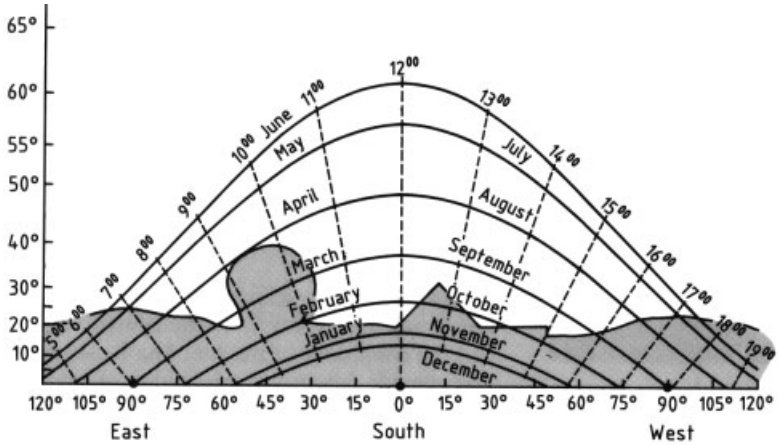

Figure 31. Solar chart for a location of 52° north with obstruction profile. The sun path lines apply to the 21st of each month

covery system. Air exchange rates of $0.5\,h^{-1}$ to $1\,h^{-1}$ are sufficient in most cases.

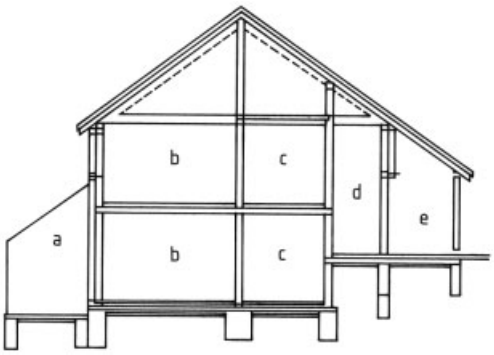

Figure 32. Cross section of the ISFH solar house
a) Conservatory; b) Living room; c) Kitchen; d) Storage room; e) Porch

3.3. Elements of Passive Utilization

3.3.1. Collection

Windows. With exposure of windows to solar radiation, considerable heat gains can be attained. As basic elements of buildings, windows represent the simplest system for the passive utilization of solar energy.

Incident solar radiation can pass through glass almost unimpaired (Fig. 33). On striking walls, ceilings, floors, and items of furniture, the radiation energy is converted to heat (direct utilization). The heat capacities of these items of construction should be large to prevent an excessive increase in temperature. The heated areas emit part of the absorbed energy. Since the emitted radiation is in the infrared region, it cannot get through to the outside, because infrared radiation cannot penetrate glass. Therefore, the heat remains in the room (greenhouse effect). The absorbed energy is distributed in the building by means of air convection.

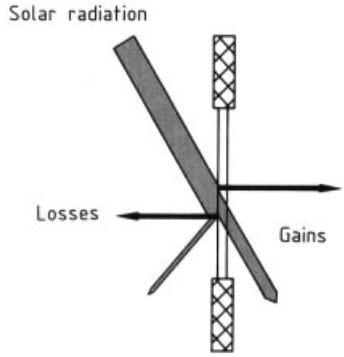

Figure 33. Radiation balance of a window pane

The solar gains from vertical windows depend on the orientation, the largest gains being achieved from windows facing the south. They always have a positive heat balance, i.e., the solar gains are on an average larger than the heat losses. Vertical glass surfaces prove to be especially favorable. The intensive summer radiation comes in at a steep angle and the irradiance is decreased. This, in turn, reduces the danger of overheating in summer. The desired winter sun-

light enters at a flatter angle resulting in a higher irradiance of the window area. The dimensions of the windows depend on the radiation supply of the site, the heat requirements of the building, and on the size of its storage mass.

Windows should be equipped with heat-conserving double glazing. As a result of a special coating with a low infrared emittance and an argon filling, this glazing has losses that are only about half as high as those of simple double glazing. The U value of heat-conserving glazing is $1.3\,\mathrm{W\,m^{-2}\,K^{-1}}$. Shutters, curtains, etc., provide additional insulation during the night.

Sun Spaces. An attached sun space or conservatory (Fig. 34) fulfills many functions: it heats up when insolated, it can give off excess heat to adjoining rooms, and it forms a buffer room against fluctuations in the outside temperature. It can also be an attractive living room in the warmer months.

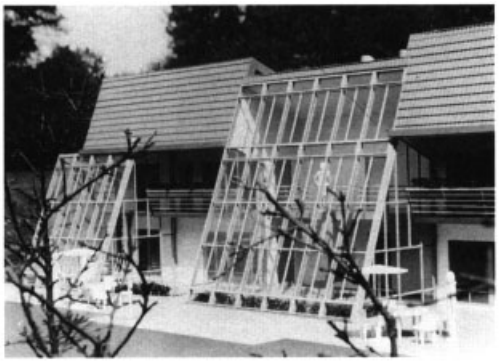

Figure 34. South view of an achitectural model with distinct sun spaces

The prerequisite for efficient functioning is that the conservatory remains unheated. If plants are grown in this conservatory, the supply of heat must be limited to frost protection. The connecting wall and the floor serve as thermal stores and heat-conserving glazing increases the utilization potential. To guarantee its function as a buffer room, the connecting wall should not be permanently kept open. In the case of sufficient insolation, the fresh air for the main house can be preheated in the conservatory. To exclude overheating, shading elements and an adequate ventilation of the conservatory are required.

Solar Walls. In the simplest case, a solar wall is a glazed outer wall which is painted black and as far as possible faces the south (Fig. 35). It should consist of a material that is a good thermal conductor with a high heat-storing capacity, e.g., concrete or solid stone masonry. The single or double glazing is attached about 10 cm in front of the wall. It reduces heat losses to the outside air. Solar radiation is absorbed on the outer surface of the wall, the heat is transferred to the inner surface of the wall by conduction, and it is then emitted into the room by means of radiation and convection. Roofs and walls that do not have transparent insulation conduct inwards only a negligibly small part of the heat generated by the absorbed radiation.

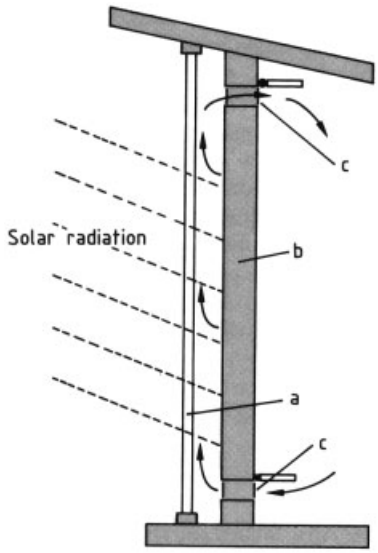

Figure 35. Sketch of a solar wall
a) Glass cover; b) Storage wall; c) Ventilation flaps

The insulating effect of the glass pane and, consequently, the effectiveness of the solar wall can be substantially increased by the additional provision of transparent insulation materials [27], [28] made of polycarbonate or polyacrylate (Fig. 36 B). The insulating layer consists of tubes or honeycombs with a diameter of ca. 2 mm to 2 cm arranged vertically to the front pane; the layer thickness is up to 10 cm. Solar walls should also be provided with shading facilities.

The variations in temperature occurring on the outside of the solar wall are greatly suppressed inside. The peak temperature appears at

the inner surface with a delay of several hours (Fig. 37). It is also possible to start air circulation by opening ventilation flaps which allows an immediate heat supply to the living rooms.

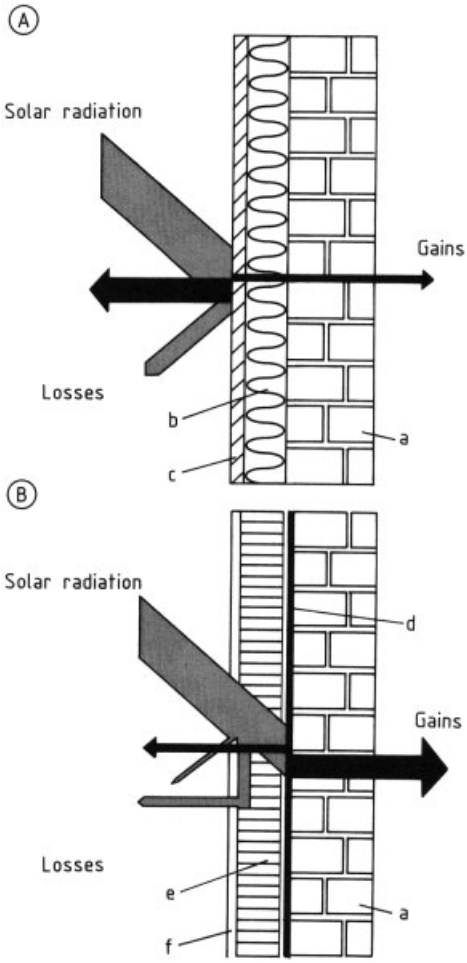

Figure 36. Structure and energy flow diagram of an opaque insulated wall (A), compared with a transparent insulated wall (B)
a) Wall; b) Insulation; c) Weather protection; d) Absorbent paint; e) Transparent insulation; f) Glass pane

Air Collectors. In air collectors, the energy of the solar radiation is directly converted to thermal energy of the air; no additional storage masses are required. The collection of radiation energy occurs at suitable points, e.g., in the roof (Fig. 38). The lamella improve the heat transfer between the absorber and the air. The heated air is transported through ducts to rooms which do not receive sufficient radiation. A distinction is made between operation by circulating air and *fresh air* operation. Air circulation can be achieved by natural convection, if the collector is suitably arranged, or with the help of blowers. Air collectors require a highly transparent covering with good thermal properties.

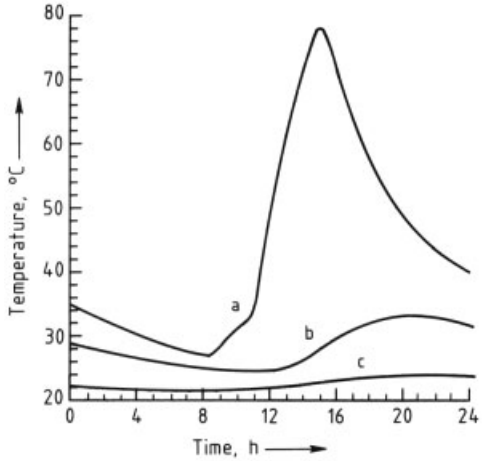

Figure 37. Daily pattern of the surface temperature of a transparent insulated solar wall and the air temperature in the adjoining unheated room. The solar radiation on the vertical was 5.6 kWh m^{-2} d^{-1}, the average outside air temperature was 2.9 °C
a) Outer surface temperature; b) Inner surface temperature; c) Room air temperature

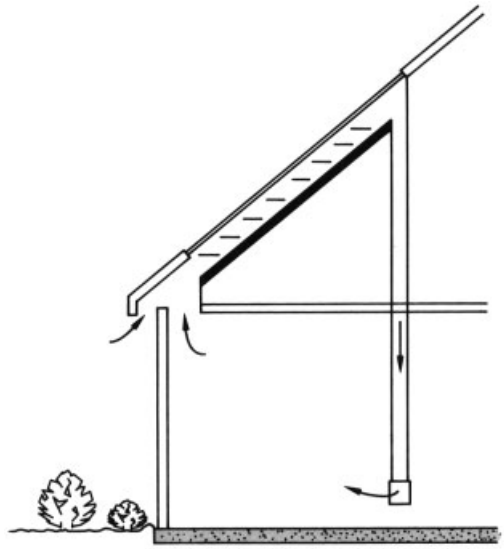

Figure 38. Air collector in the roof
The air is circulated mechanically

3.3.2. Storage and Distribution

The utilization of solar energy must be adjusted to a very irregular energy supply. If the energy of the incoming radiation is to be utilized over and above the immediate requirements, it must be stored. Fluctuations in the room temperature are suppressed and delayed by the storage of heat.

The surface of storage masses that are directly irradiated by the sun heat up and conduct a part of the energy inwards (*primary storage*). A storage mass can also undergo convective loading (*secondary storage*) if heated air passes its surface. Heat transfer then continues until the surface temperature of the building component is the same as the air temperature.

Apart from the utilization of construction elements already present in the house as thermal stores, additional elements such as *water tanks* can be used. Heat stores outside the house have also been tested. The efficiency of a store depends on a series of parameters. The most important are its heat capacity, the velocity of heat penetration, and the size of the exposed surface.

Careful planning is required to achieve distribution of the collected energy through natural convection currents (Fig. 39). In many cases, mechanical ventilation systems cannot be avoided. Moreover, they have the advantage that the consumption of stored heat can be controlled and thus be restricted to times of need.

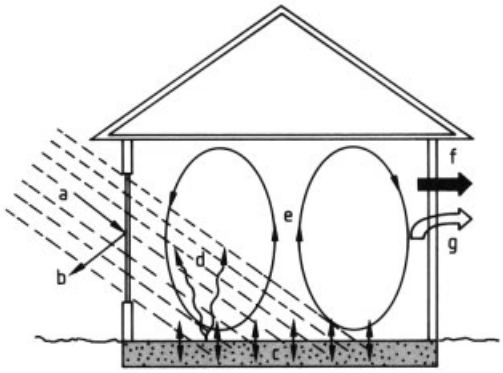

Figure 39. Schematic representation of the passive utilization of solar energy: collection, storage, and distribution
a) Solar radiation; b) Reflection; c) Charge and discharge; d) Infrared radiation; e) Convection; f) Transmission losses; g) Infiltration losses

3.3.3. Shading and Cooling

Efficient solar buildings always hold the danger of overheating in summer. In fact, in hot climatic regions, the avoidance of solar-energy input into the house represents the central aspect of the building design. If large apertures are to be used for solar heating as and when necessary, shading equipment must be provided. Apart from fixed eaves, movable elements such as blinds, shades, awnings, or folding shutters have proved useful. External shading equipment is more effective than internal equipment.

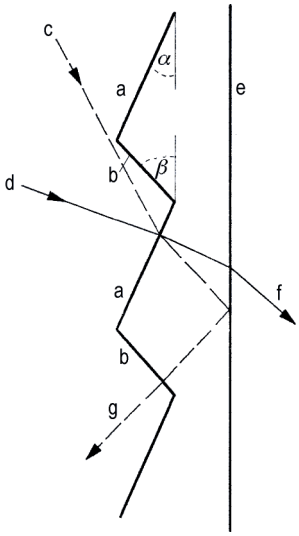

Figure 40. Section of a prismatic pane (h) for seasonal shading of buildings with refracting surfaces (a, b) and the course of two principal rays. Incident sun rays (c, d) are directed to the rear surface (e) in such as way that they are either transmitted (f) or reflected. Reflection occurs when the angle of incidence is higher than the critical angle of total reflection.

Great importance is attached to the use of windows with thermochromic or electrochromic coatings in the future [29]. They can be employed to realize an intelligent building envelope, which automatically controls its optical transmission properties in accordance with the energy requirements. On reaching a threshold temperature, *thermochromic coatings* switch from transparent to diffusely scattering. On the other hand, *electrochromic coatings* switch from permeable to absorbing on application of an electrical field. A very effective seasonal shading that is responsive to the sun's position can be realized with prismatic panes (Fig. 40). These

panes are suitable for window or facade elements, preferably in applications where no clear view is required, and especially in front of transparently insulated solar walls. The direct radiation received on a vertical south-facing facade is reduced to 10 % on clear summer days, while 90 % is transmitted on clear winter days [34].

Apart from shading, adequate ventilation must be provided. In particular, airing at night can contribute to the cooling of storage masses in the house. Mechanical ventilation systems can also make use of the coolness of the ground by leading the fresh air into the building through pipes laid in the ground.

3.3.4. Daylight

Solar design implies avoidance of artificial lighting whenever daylight is possible [30]. However, it is not only the saving of electricity that is important. Daylight increases visual comfort, enhances natural colors, and increases contrasts. Moreover, direct light permits contact with the outside world.

Suitable reflectors, e.g., in skylights, are employed for homogeneous lighting, even in internal rooms.

3.4. Performance of a Solar House

Information about the performance of solar houses is available in a large number of test reports on controlled passive solar systems [31]. Architects and civil engineers can refer to these reports when working out a concept for a construction project, selecting and dimensioning solar components, or for an estimation of the economics. Figure 41 shows the monthly consumption of heating energy and the solar savings of the ISFH solar house for the heating period of 1992/1993.

Highly developed computer programs, which allow the simulation of buildings to be planned, are available commercially. Weather data for all possible climatic regions are supplied with these programs. The simulations reveal room temperature patterns and heating or cooling requirements. Plans can be optimized stepwise by using test runs with varying parameters. A review of the programs available can be found in [32], [33].

4. Photovoltaic Systems

Since the development of the first solar cell in 1954 [35], annual photovoltaic (PV) module production in 1992 had increased to ca. 60 MW (peak) [36]. The main fields of application are:

1) The "power sector", i.e., stand-alone systems for lighting, communication, water pumping, village electrification in developing countries, etc., as well as grid- connected PV systems of the kW to the multi-MW range.
2) The "consumer electronics sector", i.e., small-scale PV systems powering, e.g., pocket calculators and wrist watches. The market share of this sector is ca. 20 %.

These figures indicate that photovoltaics have established themselves as an "additional" energy source with good prospects for the future, even if limited to the present fields of application. Up to now PV systems have not been able to achieve the status of a real "alternative" energy source. As before, the high cost prevents photovoltaics from successfully entering the field of large grid-connected central station power plants.

Modularity, the possibility to construct PV energy supplies from the mW to the multi-MW size, is the likely specific property of PV that has enabled its relatively wide distribution.

Tracking and *concentration* have been increasingly discussed and investigated, because:

1) Prices of PV modules have not decreased to values predicted 10 years ago
2) Concentrator solar cells are now available on the market

According to P. Maycock [37], concentrating PV modules show good cost perspectives (between conventional and thin film cells) up to the year 2010.

4.1. Principle of Operation, PV Materials, and State-of-the-Art Solar Cells

Solar cells make use of the photovoltaic effect in semiconductor $p-n$ junctions or related structures. In principle, the energy of light quanta is

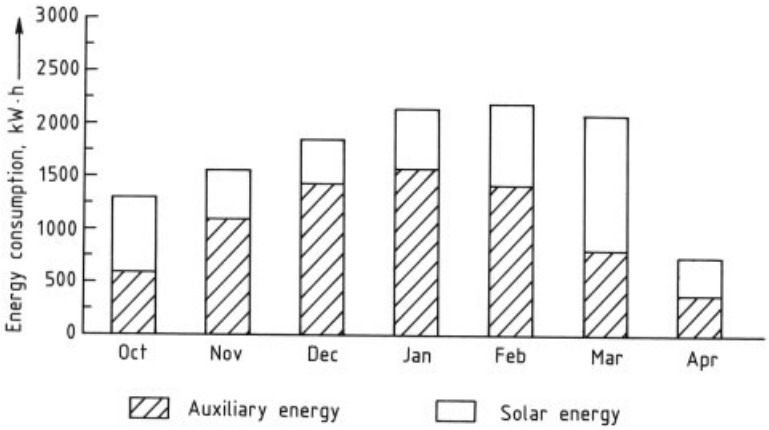

Figure 41. Monthly heating energy requirement and solar savings of the ISFH solar house

transferred to the electronic subsystem of the semiconductor (and not to the lattice!) and can be subtracted from the device directly as electrical energy. For details of the operation principle and of the state-of-the-art, see → Photovoltaic Cells.

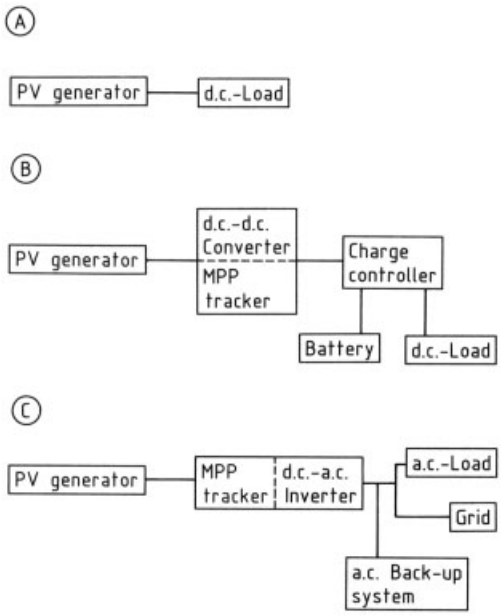

Figure 42. Photovoltaic systems
A) Simple d.c.-system; B) Sophisticated d.c.-system; C) System with a.c.-load or grid- connected system

Concerning materials for photovoltaics, *crystalline silicon* (c-Si) is still the most widely used in the field. In the laboratory efficiencies as high as 24 % have been obtained.

The most attractive alternatives are materials that enable the production of thin-film solar cells. Among these, *amorphous silicon* (a-Si) is playing the dominant role, especially in the consumer electronics sector. Efficiency values obtained in the laboratory are 13 – 14 %, however, with a-Si, modules, they are degrading during operation to values well below 10 %. Due to these stability problems a-Si has not been used in the power sector up to now.

Other promising materials are $CuInSe_2$ (*CI S*), *CdTe*, and related compounds. Efficiencies of 15 % have been achieved [38], [39], and stability problems are unknown, however, large-scale industrial production has not yet started.

Further alternatives are GaAs and related *III – V- compounds*. Efficiencies close to 30 % have been achieved with one-sun illumination and 34 % with concentrated sunlight [40], however, due to the high production cost the application of III – V compounds might be limited to concentrating PV systems.

4.2. General Structure of PV Systems

Structures of photovoltaic systems are shown in Figure 42, starting in A with the simplest system, a d.c. -load coupled directly to the PV generator. A more sophisticated system, as shown in Figure 42 B, includes maximum power point (MPP) tracking that provides impedance matching between the dynamic resistance of the solar generator and the load, and a d.c. – d.c. converter

for voltage adaption. Storage facilities can also be optionally included in the system.

The connection of PV generators to a.c. loads or to the public grid requires a power conditioning unit (PCU) consisting of at least a d.c. – a.c. inverter (see Fig. 42 C). In most cases the inverter also includes the function of an MPP tracker; optional back-up and storage units may be added to the system.

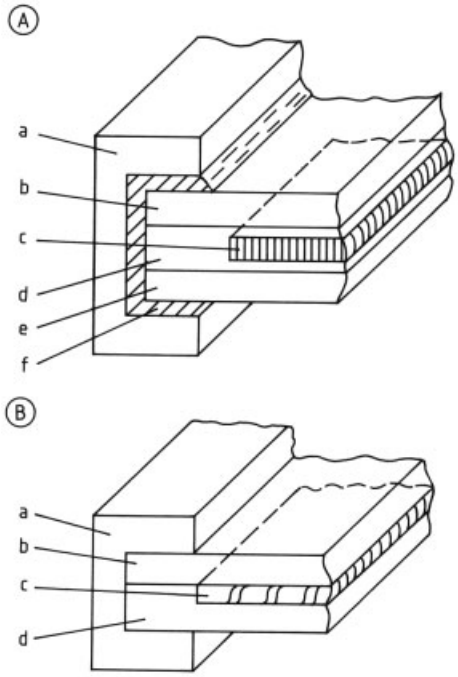

Figure 43. Present constructions of flat-plate modules with A) Discrete; B) Thin-film solar cells
a) Frame; b) Front cover; c) Solar cells; d) Embedding material; e) Back cover; f) Gasket

4.3. Photovoltaic Generators

The smallest PV unit which is easy to handle for the user, is a *module*. A number of modules mechanically assembled to a larger unit and electrically interconnected is, according to IEC standards, named *array*. Large generators may consist of a number of arrays or even *array fields*.

4.3.1. Module Technology and Specifications

The construction elements of modules with discrete or thin-film solar cells are shown in Figure 43. The requirements for the front cover (b) – low reflectance, good optical transmission, chemical durability, abrasion resistance, mechanical stability, and simple processing – are fulfilled by low-iron glass with a typical thickness of 2 – 3 mm, that is tempered or chemically hardened. The frame of a module (a) is formed by aluminum or stainless steel profiles.

Two materials have proven to be suitable for embedding the solar cells (d): poly(vinyl butyral) (PVB) and ethylene – vinyl acetate copolymer (EVA). Both are thermoplastics. EVA has a better light transmission than PVB, but effort is required to obtain good adhesion to glass and solar cells.

The back cover (e) of a photovoltaic module consists either of a glass plate or a back sheet system of Tedlar [poly(vinyl chloride)] – aluminum foil – Tedlar and protects the solar cells from environmental influences.

Two additional design objectives exist:

1) To keep the operating temperature of the solar cells as low as possible
2) To choose optimal electrical interconnection schemes

The importance of the first goal results from the fact that the efficiency of solar cells decreases with increasing temperature. Since module temperature increases with irradiance, the module efficiency in practice decreases with increasing irradiance [41]. This has also been revealed by outdoor tests carried out at ZSW, Baden-Württemberg (see Fig. 44 and [41]).

The second objective covers aspects such as shading losses and "hot spots". These occur when, e.g., in a string of series- connected cells one cell is partially shaded or damaged. Without protection measures the shaded cell becomes reverse biased and acts as a load. The dissipated power in this load heats the cell. This so- called hot spot can irreversibly damage a module.

Cell damage can be avoided and the output power may be increased if a shunt or bypass diode is connected across the shaded cell. A similar effect occurs when cells or modules are connected in parallel. In this case, a reduction of power losses can be achieved by a blocking diode connected in parallel to the shaded cell.

For designing and modeling of PV plants, correct and reliable data on the performance

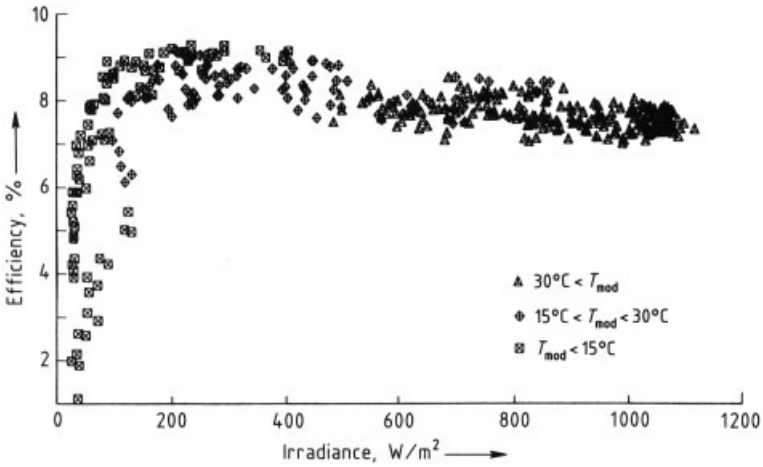

Figure 44. Temperature dependence of efficiency of PV modules on illumination according to outdoor tests at ZSW (Centre for Solar Energy and Hydrogen Research), Baden-Württemberg

of PV modules are required. Furthermore, fair competition among module producers should be ensured. Consequently, there was a strong need for standardization of the efficiency rating of PV modules. At present, the nameplate rating of PV modules has to be performed under test conditions according to the IEC Test-Specification 503:

1) Irradiance 1000 W/m^2
2) AM 1.5 spectrum according to IEC Document 904
3) Solar cell temperature 25 ± 1 °C

These Standard Test Conditions (STC) are internationally accepted, together with the agreement that the nameplate rating must be within ± 10 % of the STC value.

Since the energy generation of PV modules depends on the site-specific meteorological conditions, computer models have been developed for the simulation of PV power production from meteorological data and PV module measurements and ratings (see, [42], [43]).

4.3.2. Tracking and Concentration

Methods of PV *array tracking* including fixed-tilted orientation are depicted in Figure 45. The selection of a method depends, among other aspects, on the geographical latitude of the site: horizontal N – S-axis tracking (D) is suitable for sites near the equator, vertical axis (B) suits for high geographical latitudes, and polar axis (inclined axis) (C) and two-axes tracking (A) can be applied everywhere.

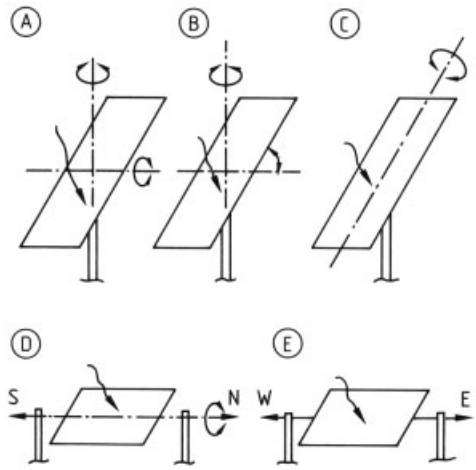

Figure 45. Methods of PV array orientation
A) Two-axes tracking; B) Vertical-axis tracking; C) Polar axis tracking; D) Horizontal-axis tracking; E) Fixed, tilted

Using average monthly readings of global and diffuse irradiation and appropriate irradiance models which are calculated on an hourly basis, the annual irradiance gain by tracking has been calculated for Weihenstephan (near Munich, Germany) [43]:

Two-axes tracking:	+ 33 %
Inclined-axis tracking (40°):	+ 30 %
Vertical-axis tracking (53 °):	+ 30 %
Elevation tracking:	+ 6 %

The degrees in parentheses indicate the inclination angle of tracking axis (40°) and of modules (53°) on the tracker. Elevation tracking – not shown in Figure 45 – is a tracking mode following only the elevation angle of the sun (e.g., tracking around the east – west-axis).

Concentrated photovoltaics are possible with various technologies. An overview is given below

Point-Focus Devices:
- Fresnel lens
- Parabolic dish

Linear-Focus Devices:
- Linear Fresnel lens
- Parabolic trough

Nontracking Devices:
- Compound parabolic concentrator (CPC)
- Other nonimaging devices

Point- and linear-focus Fresnel concentrators have been mostly used. Typical concentration ratios are ca. 25, with the tendency to higher ratios. A schematic of a conventional concentrator module package is shown in Figure 46. In Figure 47 linear- and point-focus concentrator modules on a test facility are shown. To achieve high concentration ratios, the cusp in Figure 46 is improved to a "second concentrator" mounted directly to the surface of the solar cell.

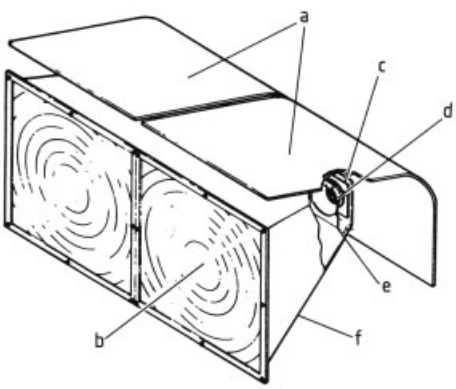

Figure 46. Concentrator module package with Fresnel lenses
a) Reflux- cooler panel; b) 24 inch (61 cm) square Fresnel lens; c) Cell mount; d) Cusp; e) Shield; f) Fiber-glass housing

A second strategy is *soft concentration* by, e.g., mirrors mounted at the side of the PV modules. This technology utilized ordinary "one-sun modules" (standard-type modules designed for application without concentration) in connection with simple plane mirrors ("V-trough"). The main idea behind this concept is that both the modules and the plane mirrors are available as mass products. Figure 48 shows the world's largest PV plant at Carrisa Plains, United States, using this technology. At the Carrisa Plains site, a "browning" (discoloration) of the encapsulant in the PV modules occurred due to the high operating temperature of the modules.

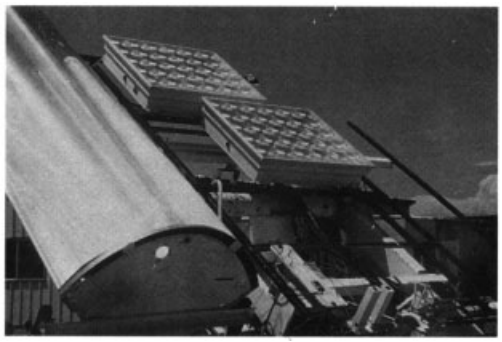

Figure 47. Concentrator modules with linear- and point-focus Fresnel lenses on a test facility in the United States

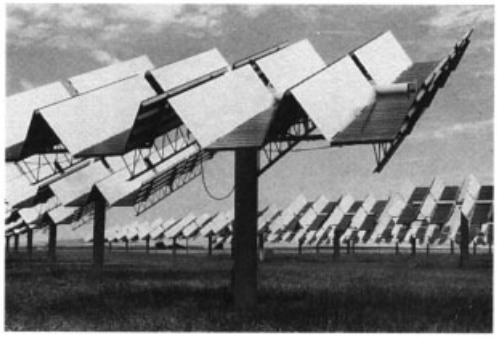

Figure 48. 6 MW (peak) PV power plant at Carrisa Plains, United States, using V-trough concentrator technology

4.4. Power Conditioning Systems

The operational units of a power conditioning system are shown in Figure 42. The design characteristics of PV systems are:

1) Systems with d.c.- load
   a) d.c. – d.c. Converter with fixed input voltage (simple, inexpensive)
   b) Controlled d.c. – d.c. converter providing also MPP tracking

2) Systems with a.c.-load or grid-connected systems
   a) d.c. – a.c. Inverter with constant d.c.-voltage input
   b) d.c. – a.c. Inverter also performing MPP tracking

d.c. – a.c. Inverters can be designed to operate:

1) Externally commutated (for grid connection), or
2) Self-commutated (for stand-alone applications, but also usable for grid connection)

Technical details of these power conditioning components are given in [44]. The central component is the d.c. – a.c. inverter. The most common types are:

1) *Square wave generator* (thyristor bridge inverter). Produces high harmonic distortion of a.c.-output, filter required; however, simple, robust, suited for high power applications.
2) *Pulse-width-modulated (PWM) inverter*
   a) with low-frequency transformer
   b) with high-frequency transformer

   Properties: line- and self-commutated versions, low harmonic distortions, higher cost, applicable only to lower power ranges (< 50 kW). However, due to progress in IGBTs (Isolated Gate Bipolar Transistors), PWM inverters can now be realized up to the power range 1 MW.

Inverter efficiencies are close to 90 % at nominal power. The main problems are their lower efficiencies at partial load and the losses occurring during standby, start-up, and shutdown.

4.5. Storage

For energy storage predominantly lead – acid and nickel – cadmium batteries are used because of their reliability, availability, and flexible sizing. The main disadvantages are their high cost, low-energy-density, and the need for careful treatment. The number of possible cycles during their lifetime decreases drastically with the depth of discharge. Advanced electrochemical cells could improve energy density, self discharge rate, and cycle life. Two promising systems are $Na - \beta\text{-}Al_2O_3 - S$ and $Zn - ZnBr_2 - Br_2$.

Figure 49. 4.5 kW (peak) PV system at "Unterkrummenhof", Black Forest, Germany

4.6. Stand-Alone PV Systems

For many applications, especially in remote places far away from the public grid, PV is an economic power source. In the dry, sunny regions of the world there is a tremendous need for water pumping (drinking water and irrigation), village electrification, and other applications which could be met by photovoltaics more economically than by other methods.

Even in industrialized countries a large market for photovoltaics exists to meet electricity needs in remote areas; for example, the powering of buoys, emergency telephones at highways, or weekend houses. This market is even considered to the several megawatts per year in Europe. Figure 49 shows the PV system of the "Unterkrummenhof", a hiker's inn in the Black Forest in Germany. The 4.5 kW(peak) PV system, equipped with storage battery and diesel back-up, has been designed by the Fraunhofer Institute for Solar Energy Systems (ISE), Freiburg.

4.7. Grid-Connected PV Systems

The present situation can be characterized as follows:

1) Large grid-connected PV power plants were installed in the United States from the late 1970s onward (see Fig. 48).

2) Numerous systems were erected in Europe in the early 1980s, sponsored by the Commission of European Communities. Meanwhile, local authorities and utilities are financing and operating PV power plants. Figure 50 shows the PV power station at the Lake of Neurath, erected by RWE Energie AG, Essen, Germany.
3) Even though energy from PV is far from being competitive with energy supplied from the public grid, the interest in installing grid-connected PV in the private sector has increased, stimulated in several countries by financial support from the government, e.g., by the "1000-Roofs-PV-Programme" in Germany.

Figure 50. 360 kW(peak) PV power plant of RWE Energy AG at Lake of Neurath/Grevenbroich, Germany (photograph RWE)

4.8. Design Considerations and Modeling

The design of a PV plant starts with selection of site and the type of system. The site properties influence whether fixed-tilted, flat-plate tracking, or concentrating systems will be chosen. A critical plant design parameter is the *voltage level:* a high nominal voltage keeps ohmic losses low, but increases cost for high-voltage d.c. hardware and requires additional safety measures. Because of decreasing inverter efficiency at partial loads it may be useful to split the d.c. – a.c. inverter into several units with lower nominal power.

About 15 – 20 design models for PV systems have been developed and are being refined. A summary is given in [45]. Representative examples are:

1) *SOMES* (for stand-alone systems including PV, wind turbines, battery storage, and back-up energy source) [46]
2) *PVFORM* (for stand-alone and grid-connected systems; high accuracy claimed) [47]
3) *ASHLING* (for all types of systems; extensively validated on component level) [48]

4.9. Operational Experience

Detailed performance evaluations of grid-connected PV plants installed in the United States and in Europe have been reported, see [49–51]. They are summarized in the following:

4.9.1. PV Module Performance

Crystalline Si modules are well-developed industrial products of high reliability and low degradation of efficiency (<1 % per year). Lifetimes of 20 – 30 y can be assumed, however, confirmation of this is still required.

Amorphous Si modules have not yet proven their suitability for large-scale power production.

The *practical field efficiencies* of crystalline Si PV modules are 10 – 20 % lower than manufacturer's nameplate d.c. power ratings.

4.9.2. PV System Efficiency, Annual Energy Output

The total *balance of systems* (BOS; all components except PV modules), electrical losses caused by module and array mismatch, wiring, shadowing, dirt, inverter, and control system, are in the range of 15 %.

The *annual energy output* of the entire PV plant is ca. 70 – 80 % of the product of annual insolation and manufacturer's rated module efficiency for crystalline Si-based systems.

Small grid- connected systems can energetically operate as efficiently as large centralized systems.

Using *tracking of arrays*, the annual energy output can be increased by up to 30 %.

4.10. Economic Aspects, Market, Energy Payback, Prospects

Stand-alone and grid-connected systems must be clearly distinguished when discussing the economics of PV. The multifunctional use of PV modules (PV roofs, PV facades, etc.) must also be considered.

4.10.1. Stand-Alone PV Systems

As mentioned before, the field of application of stand-alone PV systems is huge in both developing and industrialized countries. The world market is expected to be between 100 and 400 MW by the year 2000 and might be only limited by the financial capabilities in Third World countries. Energy payback, in this context, must not necessarily be considered a decisive problem. The primary concern is often the availability of energy.

4.10.2. Grid-Connected PV Systems

The economics of PV power plants are dominated by the installation cost; there is no fuel consumption, and maintenance cost can be assumed to be low as compared to investment cost.

The prices of PV modules, in contrast to predictions, have not continuously decreased, they even increased from values ca. $ 4/W(peak) in the late 1980s to ca. $ 5 – $ 7/W (peak) in 1992. The total cost of PV systems in 1991/1992 was in the range of $ 10/W(peak). Characteristic figures are given in Figure 51.

The "annuity method" (constant capital cost throughout the depreciation/operation period of the PV plant) may be used to estimate the cost of the energy produced by PV power plants. According to this method (see, [41]) the present PV energy costs in Central Europe are ca. $ 1.2/kWh. At sites with higher insolation, e.g., in the southwest United States, these costs are ca. 40 c/kWh. For cost-competitive "bulk" power generation, PV-system cost would need to decrease to ca. $ 1/W(peak), if conventional energy cost remains in the present range. This goal might be achievable only with thin-film solar cells (see also [37]).

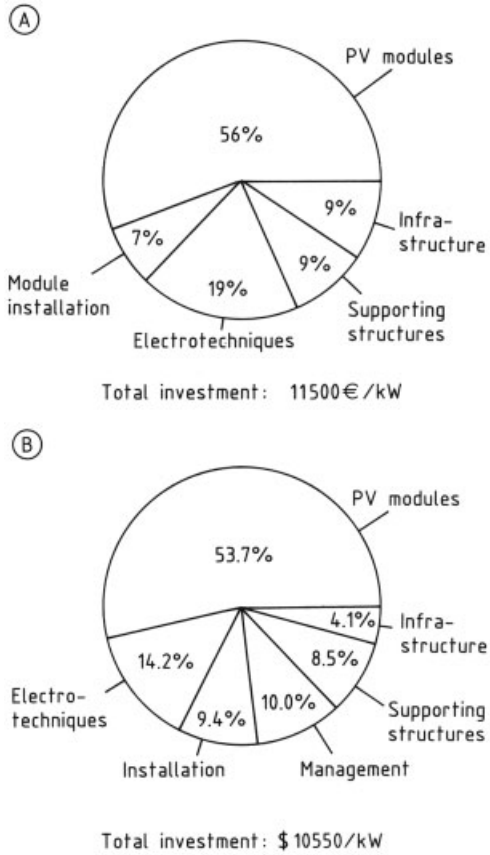

Figure 51. Investment cost of PV power plants
A) 340 kW(peak) plant at Kobern – Gondorf, Germany [52];
B) 300 kW(peak) plant at Austin, Texas [53]

As can be seen from Figure 51 the BOS cost is nearly as high as that of PV modules. The reduction of BOS cost, therefore, is of equal importance and presents a technical challenge that will be as difficult to meet as the lowering of PV-module cost.

4.10.3. Energy Payback Time

Only a limited number of studies allow an insight in methods, basic parameters, and conditions of the calculation of energy payback time. These have been summarized in [54]. For complete grid-connected systems – PV generator, supporting structure, and inverter – the energy payback times for different solar cell techniques and supporting structures, including tracking and concentration, are in the range of three to five years. These results indicate, that – also with

respect to energy payback – a clear winner of PV systems technology cannot be determined at present. Nevertheless, a tendency favoring thin-film cells seems to exist.

4.10.4. Multifunctional Use of PV Modules

The economic situation of photovoltaics can be considerably improved by multifunctional use of PV modules as, e.g., with PV roofs and PV facades. In these cases additional cost-reducing factors exist:

1) No ground consumption by PV, no ground cost
2) Low cost for supporting structures
3) No extra building for power conditioning systems and controls
4) Vicinity to consumer of electrical energy

For PV roofs stand-off mounting is preferred at present, however, the final goal should be roof integration of solar modules which saves at least the cost of conventional roofing tiles.

The situation is even more favorable for PV facades. Conventional facades are often considered not only to be a purely functional element of a building, but also an architectural ornament. Remarkably high investments, $< \$\,2000/\mathrm{m}^2$, are being made thus allowing PV facades to easily compete. If the total cost of a PV facade is assigned to the constructional part, then the energy-generating part can be considered as almost free of charge; a highly favorable prospect.

5. Photoelectrochemical Solar Energy Conversion

Besides the semiconductor solid-state contacts used in photovoltaics, the semiconductor – electrolyte contact can also be used for conversion of light energy into electric power [55–63]. However, in semiconductor electrochemistry, a charge transfer from the electrode to the electrolyte has to take place in addition to the current flow through an external circuit. This charge transfer can be utilized by inducing a battery reaction. Here, a direct conversion of solar into chemical energy is possible (*Redox Solar Battery*) [64]. A direct conversion of solar energy into a *storable fuel* is also possible by direct water electrolysis, carbon dioxide reduction, etc. For example, the direct water splitting reaction by solar energy as first reported in 1972 [65] resulted in increased interest in this type of energy conversion and in the number of research papers published in this field.

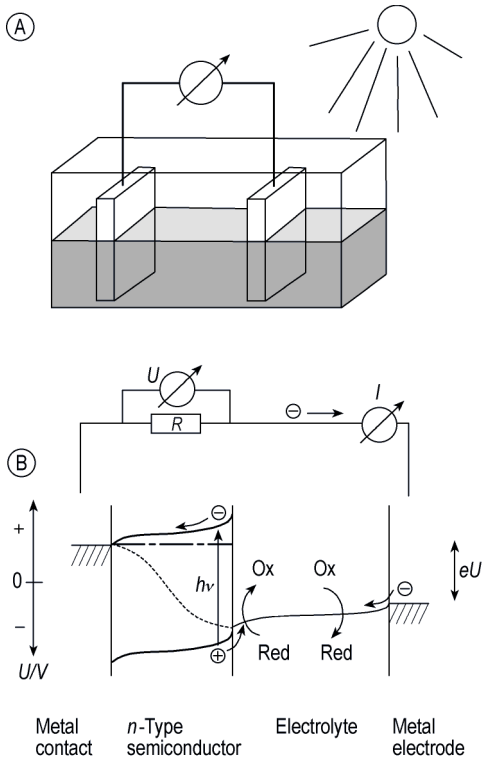

Figure 52. Regenerative photoelectrochemical (PEC) solar cell
A) Principal setup; B) Electron energy diagram for an *n*-semiconductor – electrolyte – metal configuration

The history of photoelectrochemistry dates back to 1839 when E. Bequerel discovered a photovoltaic effect illuminating a platinum electrode covered with silver halide in an electrochemical cell [66]. However, for improved investigation of the "Bequerel photovoltaic effect" [67] the growth techniques to produce high quality semiconductor crystals had to be developed (→ Semiconductors, Chap. 4.2.). The investigations of the germanium electrochemistry began in the mid-1950s [68] which led to the development of basic theories during the early 1960s. In particular, Heinz Gerischer and Rüdiger Memming must be mentioned. However, despite

research for more than 30 years, many fundamental problems still remain to be solved [69].

Photoelectrochemistry has found a new interest in two fields known for many years but which have made progress since 1990; dye sensitization and solar detoxification. The efficiency of the conversion of light to electricity by solar cells with a large band gap semiconductor *sensitized by organic metal complexes* has been continuously improved. Solar efficiencies of at least ca. 10.5 % have been achieved which, in connection with very cheap preparation techniques for the very rough surfaces required, could eventually make them superior to conventional solar cells [70]. *Solar detoxification* has reached a large-scale testing phase although many fundamental problems remain to be solved. Fundamental research is also needed in the promising fields of *direct conversion of solar light into storable fuels*, where the possibility of catalysis is still an open question.

5.1. Principles of Photoelectrochemical Solar Energy Conversion

The basic principles of semiconductor electrochemistry are described in several review papers (e.g., [55], [57], [60], [61]) and a few text books [56], [58], [62], [63]. The important equations and fundamentals have been summarized and an appropriate terminology was recommended by IUPAC [55]. The only difference between a photoelectrochemical and a photovoltaic device is that in photoelectrochemistry (PEC), a semiconductor – electrolyte junction is used as the active layer instead of the solid-state junctions in photovoltaics (see Fig. 52). In both cases, a space charge region is formed where contact formation compensates the electrochemical potential differences of electrons on both sides of the contact. The position of the band edges of the semiconductor at the interface can be assumed in a first approximation to be dependent only on the pH of the solution and independent of the potential (Fermi level) of the electrode or the electrolyte. Hence, upon illumination, *electrons in the conduction band* of the semiconductor reach the electrode surface and can reduce redox couples whose redox potentials are located below the conduction band (i.e., their electrochemical potential U is more positive). *Holes in the valence band* can oxidize species with a redox potential more negative than (i.e., above) the band edge of the valence band at the interface. However, the thermodynamic requirements as described by the positions of the quasi Fermi levels have to be fulfilled in addition (compare Fig. 54).

Using these principles two types of PEC cells can be built. If the redox system oxidized at one electrode is reduced (regenerated) at the counter electrode, the *regenerative PEC cell* can convert light energy into *electrical energy* (Fig. 52, Table 2). If, however, the two half cells are separated, e.g., by an ion exchange membrane, two different reactions can be used to store *chemical energy* directly (Fig. 53, Table 3). Here, the amount of energy stored is determined by the difference in free energy of the two products. The direct water splitting is an especially interesting example which leads to the formation of hydrogen and oxygen with a thermodynamic free energy difference of 1.23 eV.

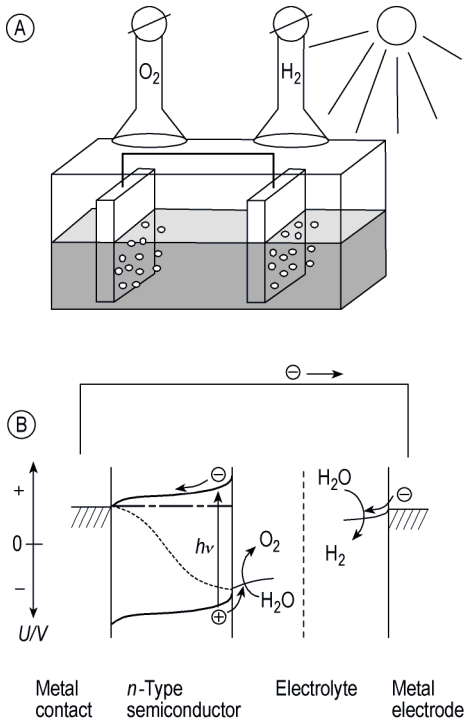

Figure 53. Direct solar water splitting using semiconductor electrodes
A) Principal setup; B) Electron energy diagram for an *n*-semiconductor – electrolyte – metal configuration

Table 2. Energy conversion efficiencies using regenerative photoelectrochemical cells

Semiconductor	Eg, eV	Redox couple	Solvent	η, %	Ref.
n-CdSe	1.7	$Fe(Cn)_6^{3-/4-}$	NaOH/NaCN	16.4	[73], [124]
n-Si/Pt in SiO_x	1.1	Br^-/Br_2	H_2O	14.9	[101]
n-WSe_2	1.2	I^-/I_3^-	H_2O	> 14	[78]
n-Si	1.1	Fc/Fc^+	CH_3OH	14	[102], [103], [81], [104]
n-Cd(Se, Te)	1.6	$Se^{2-}/Se_n{}^{2-}$	H_2O	12.7	[74], [105]
n-GaAs	1.4	$Se^{2-}/Se_n{}^{2-}$	NaOH	12	[75], [106], [107]
n-Si/Pt in SiO_x	1.1	Br^-/Br_2	H_2O	12	[108]
n-$CuInSe_2/In_2O_3$	1.01	I^-/I_3^-	H_2O	11.7	[82]
p-InP	1.3	V^{2+}/V^{3+}	HCl	11.5	[77]
n-Si	1.1	Br^-/Br_2	H_2O	11.4	[79], [80]
n-GaAs	1.4	Fc/Fc^+	CH_3CN	11	[76]
poly-n-$CulnS_2$	1.01	I^-/I_3^-	H_2O/H^+	9.7	[109]
n-CdS	2.5	I^-/I_3^-	CH_3CN	9.5	[71]
n-$MoSe_2$	1.2	I^-/I_3^-	H_2O	9.4	[110]
n-CdSe	1.7	$S^{2-}/S_n{}^{2-}/Se^{2-}$	NaOH	6.9	[72]
n-FeS_2	0.95	I^-/I_3^-	H_2O	2.8	[83]

Fc = ferrocene derivative.

Table 3. Energy conversion efficiencies using chemical storage photoelectrochemical cells

Semiconductor	Eg, eV	Redox couple	Solvent	Counter electrode	Storage as	η, %	Ref.
n- + p-GaP	2.3	$Eu^{2+/3+}$/HQ/Q	NaOH	C	redox couples	?	[111]
n-CdSe	1.7	I^-/I_3^-	NaOH	C	Ag/AgS	?	[112]
n-Cd(SeTe)	1.6	I^-/I_3^-	NaOH	CoS	Sn/SnS	11.3	[113], [114]

HQ/Q = hydroquinone/quinone

Table 2 summarizes results obtained for electrical energy production in regenerative photoelectrochemical solar cells on a laboratory scale. Table 3 gives corresponding results for the few yet published so-called photoelectrochemical solar batteries. Here the efficiency includes already all losses for the storage reactions. Since the cells have not been tested in all cases under standard conditions, the efficiencies are only an approximate guide. However, since comparably insignificant amounts of money and man power have been expended, and all experiments were performed under very crude laboratory conditions, these results are quite remarkable and prove the high quality of the semiconductor – electrolyte junction.

The energy diagrams as shown in Figures 52 and 53 display n-type semiconductor – electrolyte – metal configurations. However, p-type materials and a combination of both types (n-semiconductor – electrolyte – p-semiconductor) can also be used. As depicted in Figure 54, the latter type appears similar to $p-n$, the other configurations similar to semiconductor – metal Schottky junctions.

In photoelectrochemistry the currents of the majority carriers (i.e., electrons in the n layer and holes in the p layer) are, in most cases, determined by the charge transfer kinetics to the redox species in solution. These currents are loss currents for solar energy conversion since they diminish the needed minority carrier currents flowing in the opposite direction. However, in photoelectrochemistry are in general smaller than those found for semiconductor – metal contacts since, generally in solid state contacts, charge transfer at the interface is infinitely fast [84]. In principle this decrease in majority carrier transfer rates at the semiconductor-electrolyte interface is the reason for the high conversion efficiencies of PEC cells as shown in Table 2.

The results obtained for the *direct water splitting reaction* have been much less promising. The following conditions must be fulfilled for a direct solar energy conversion into a storable fuel: The photovoltage (splitting of the quasi Fermi level under illumination) must be larger

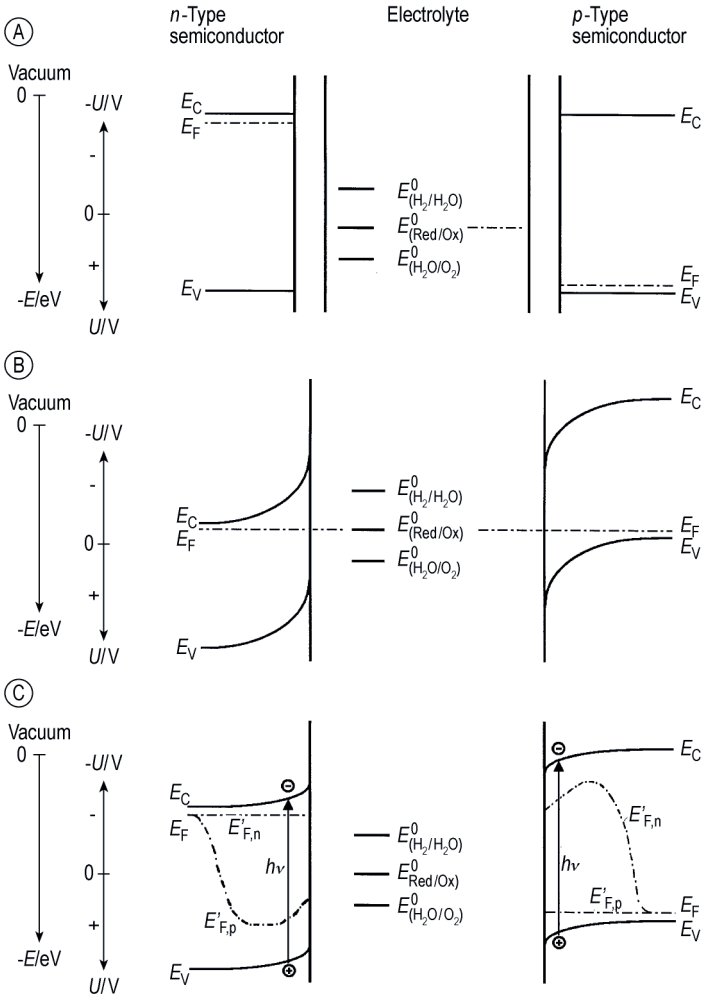

Figure 54. Formation and use of a depletion layer and it's electric field when a semiconductor connects to an electrolyte
E = electron potential; $-E$ = electron energy with respect to the vacuum level; E_C = lower conduction band edge; E_V = upper valence band edge; E_F = Fermi level; $E'_{F,n}$ and $E'_{F,p}$ = quasi Fermi levels of electrons and holes, respectively, under illumination ($h\nu$); E^0 = standard redox potentials of the redox couples indicated
A) n- and p-type semiconductor of the same kind before contacting the electrolyte; B) Equilibrium situation in the dark with both semiconductors forming a space charge layer at the surface compensating the chemical potential difference of electrons in the three materials; C) Photovoltage formation (splitting of the quasi Fermi levels $E'_{F,n}$ and $E'_{F,n}$) under illumination enabling direct water splitting

than the voltage necessary to drive the electrolysis – i.e., the free energy stored in the products (1.23 eV for water splitting under standard conditions) – plus the necessary overvoltages to run the reactions at a given current density (see Fig. 55). From photovoltaics, the photovoltage is considerably smaller than the band gap (ca. 0.5 V) due to recombination processes. In addition, for direct water splitting the band positions at the semiconductor – electrolyte interface have to be located at an energetically suitable position. The conduction band edge needs to be located negative (on an electrochemical scale) enough above the reduction potential of water (E_{H_2/H^+} plus the necessary overvoltage), the valence band edge positive enough below the oxidation potential of water (E_{OH^-/O_2} plus the necessary overvoltage) to enable the charge transfer.

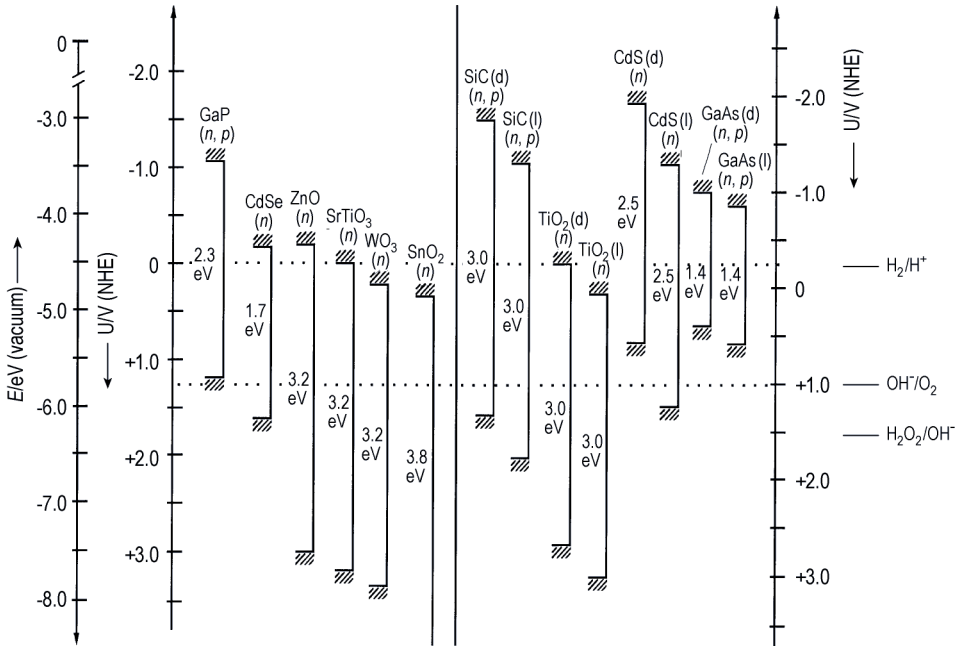

Figure 55. Band edge positions of semiconductors as determined in photoelectrochemical experiments in aqueous solutions The values on the left are literature data whereas the values obtained in the dark (d) and under illumination (l) in the right half of the diagram were determined in the author's laboratory
For each material the band gap between the lower conduction and the upper valence band edge is shown as it is fixed at the semiconductor – electrolyte interface. For comparison, the vacuum energy scale as used in solid state physics and the electrochemical energy scales, with respect to a normal hydrogen electrode (NHE) and an saturated calomel electrode (SCE) as reference points, are shown as well as the standard redox potentials of water in acidic solution

As shown in Figure 55 for an arbitrary selection of semiconductors, this requirement is fulfilled by only few of the classical semiconductors even if it is assumed that the necessary overvoltages are zero – although they are usually very high at semiconductor electrodes. These extraordinarily high overvoltages are probably the reason why even among these few suitable semiconductors only strontium titanate ($SrTiO_3$) is really able to directly split water.

However, concerning the problem of necessary overvoltages, latest results obtained with multiple nano contacts have shown that a key problem of direct water splitting has actually been solved, i.e., that of catalyzing the charge transfer of minority carriers without increasing the rate of majority carrier losses at the interface. As described above these loss currents are detrimental in the case of semiconductor-metal contacts. Therefore, using metals as catalysts in order to improve the reaction kinetics for hydrogen or oxygen formation in principle will always lead to low solar efficiencies, if real Schottky-type contacts are formed. This problem was recently overcome by depositing nano-sized metal particles as *multiple nano contacts* onto the semiconductor surface [115], [116]. This now may open new ways to again approach the "stuff that photochemists' dreams are made of", the photochemical decomposition of water [117].

5.2. Dye-Sensitized Photoelectrochemical Solar Cells

Charge transfer from light-excited organic molecules to a semiconductor with a large band gap has long been known (→ Photography, Chap. 2.2.1.) and is the basis of photography. Here, silver halide crystallites are sensitized by organic dye molecules. Although many studies have been performed on these materials the actual mechanism of dye sensitization remained unclear until electrochemists started investiga-

tions in electrolyte solutions using single crystalline semiconductor electrodes [85]. As a result [86] it became apparent that only the first monolayer of dyes directly adsorbed on the semiconductor surface is able to inject charge carriers into the semiconductor. As shown in Figure 56 light of a wavelength too long to be absorbed by a large-band-gap semiconductor electrode can be absorbed by a dye molecule (D) on the surface. If the excited electron (on D∗) is injected into the conduction band of the semiconductor (leading to the oxidized dye D^+), it can be used to create a photovoltage (eU) in a typical photoelectrochemical setup consisting of a semiconductor electrode and a metal counter electrode connected via an external load and immersed into an electrolyte. Here, the current through the electrolyte is carried by a redox couple (red/ox) which determines the potential of the counter electrode and also reduces the oxidized dye molecule (D^+) back to its ground state (D).

Although the quantum yield of this process can be higher than 90 % [87] the small overall light absorption of a dye monolayer limits the photocurrent efficiency with respect to the incident light to a value well below 1 %. Therefore, a few attempts have been made to increase the surface area by using sintered electrode materials [88]. However, these attempts were quite unsuccessful.

A breakthrough only became possible after Stalder and Augustynski succeeded in preparing highly porous TiO_2 electrodes in a sol – gel process [89]. Using these electrodes a surprisingly high light to current conversion efficiency was obtained [90]. Improvements of this system with respect to electrode preparation, type of organic ligands of the ruthenium complex used as sensitizing dye, and the solvent used for the I^+/I_3^- redox couple led to solar efficiencies of about 11 % [118], [119].

Since further improvements seem to be possible, the sensitized photoelectrochemical solar cell based on a sol – gel semiconductor and transparent conductive oxides (TCO) on glass is very promising because the production costs are expected to be low. Therefore it may become the first practical application of photoelectrochemical solar energy conversion. Nevertheless, important problems still have to be solved such as the scale-up of today's laboratory processes. Furthermore, the reported stability and efficiency has to be proven in practical applications.

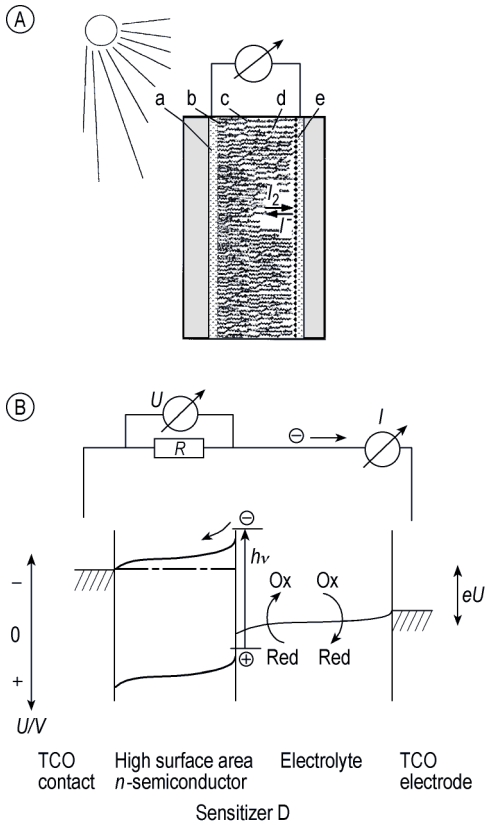

Figure 56. Sensitized photoelectrochemical cell
A) Principal setup; a) Transparent conductive oxide (TCO); b) n-Semiconductor; c) Sensitizing dye; d) I^-/I_2-Electrolyte; e) Platinum covered TCOB) Electron energy diagram for a n-semiconductor – electrolyte – TCO configuration

5.3. Solar Wastewater Detoxification

The first large-scale applications of the semiconductor – water interface have been tested in the United States and in Spain. Researchers of Sandia National Laboratory were testing the use of TiO_2 suspensions in a large-scale decontamination plant in New Mexico; Lawrence Livermore National Laboratory performed field experiment in California where in phase II of the experiment two strings of concentrating trough reactors were operating, each about 35 m long. In this program ca. 160 m^2 of

solar concentrator enables treatment of toxic organic wastewater with a maximum flow rate of 100 L through transparent tubes mounted at the focus of a reflecting parabolic trough containing the photocatalysts (Fig. 57). Powered by energy from the ultraviolet end of the solar spectrum, the catalyst produces oxidizing agents in the water (D^+) that react with the contaminants; the hazardous chemicals are converted to compounds such as water, carbon dioxide, and dilute acids that are easily neutralized. In order to maintain in electrical neutrality, a reduction process has to occur in addition (forming A^-), in most cases the reduction of oxygen.

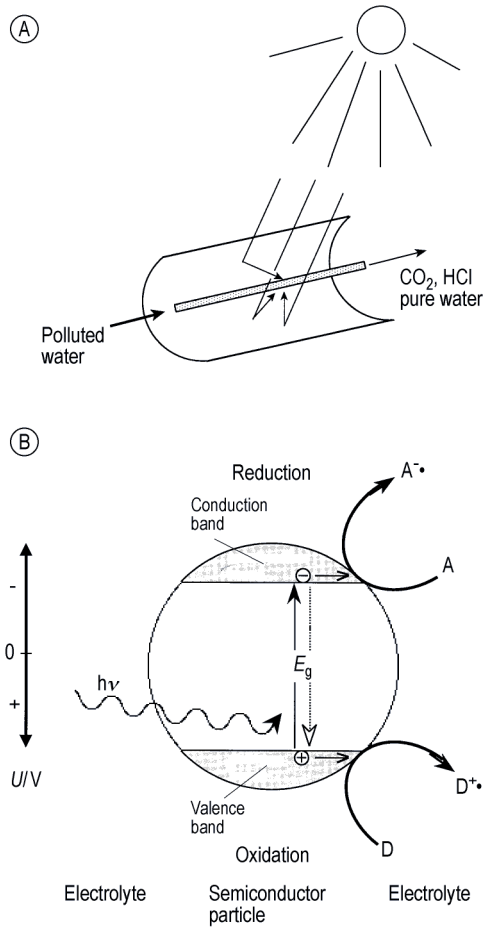

Figure 57. Solar detoxification
A) Principal setup; B) Very rough energy scheme of a semiconductor particle

A smaller but somewhat broader research program was conducted in the European solar test center Almeria (SOTA) in Spain. Among other solar high temperature and solar chemistry projects in the Plataforma Solar de Almeria there were also investigations concerning the solar detoxification of polluted wastewaters. European Community collaboration in this field has been extended by the Avicenne program to include several other Mediterranean countries, e.g., Israel and Morocco.

All large-scale programs so far are conducted in concentrating systems (see Section 4.3) even though several studies showed a decreased photocatalyst efficiency at higher illumination intensities [91].

Starting in Japan, niche applications of photocatalytic decomposition of organics [120] found widespread industrial interest, e.g., for deodorization, sterilization, disinfection and window cleaning. In Japan several companies are already producing bathroom tiles based on TiO_2 photocatalysts that work under illumination with room light or, e.g., filter papers for air conditioners under UV lamp illumination. Since TiO_2 films can be prepared colorless and transparent, all lamps illuminating traffic tunnels in Japan will be provided with glass covers which are photocatalytically self cleaning from exhaust deposits.

However, the use of TiO_2 catalysts limits the use of solar light to the far UV range of the solar spectrum and thus to efficiencies below 5 %. Therefore the question arises whether other materials can be used. Since the intermediate formation of OH radical seems to be important, mixed oxides of titanium and iron were investigated. Indeed, the absorption of light can be shifted considerably toward the visible spectrum by using these catalysts. Further search for other catalyst materials is indicated [121], [122].

5.4. Direct Photoelectrochemical Water Splitting

Despite some sensational papers on direct photosplitting of water using visible light [92], all of the systems reported so far have failed upon more critical examination [93]. Although the corresponding papers have never been withdrawn most experts agree that to date no direct water splitting has been achieved in a nonassisted process using visible light. Even with TiO_2 a bias voltage is necessary to achieve si-

multaneous evolution of hydrogen and oxygen. This concept of *photoassisted electrolysis* can of course be applied to nearly all electrode materials. In principle, the same concept is used if working and counter electrodes are immersed in electrolytes of different pHs and redox potentials or if semiconductor arrays of several contacts in series are combined to obtain the decomposition voltage in a several photon process [123].

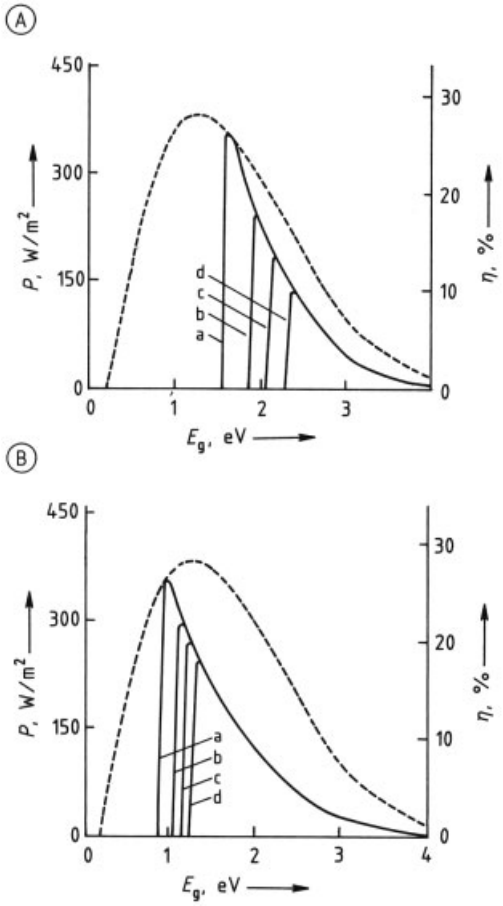

Figure 58. Theoretical solar (AM 0) efficiency of photoelectrochemical water splitting versus the band gap of the semiconductors used [123]
A) Using one semiconductor and a metal counter electrode; B) Using *p*- and *n*-type materials of identical band gaps as the two electrodes
Curves are calculated for (a) zero, (b) 300, (c) 500, and (d) 700 mV combined overvoltages for both oxygen and hydrogen evolution. Envelope gives the data for an optimum *p*–*n* photovoltaic device

As shown in Figure 58, a higher efficiency than that obtainable in a one-photon water-splitting system is achievable using two photons per transferred electron in a system with two effective junctions of semiconductors with a smaller band gap. The only successful approach to use a combination of *p*- and *n*-type semiconductor–electrolyte contacts for water splitting has been that of Inoue and Yamase using the large-band-gap semiconductor SiC [95]. Unfortunately, even this material is not stable against anodic dissolution.

As demonstrated for many systems, most modifications of the semiconductor–electrolyte interface needed to improve the charge-transfer kinetics finally lead to solid-state semiconductor junctions immersed into an electrolyte [96]. The only practical system which has ever come close to production, the Texas Instrument's *p*–*n*-silicon solar arrays for direct splitting of HBr, is a multijunction solid state device, as pointed out by Bard and coworkers [97].

It was already recognized in the early days of photoelectrochemistry [98] that metals deposited at the semiconductor–electrolyte interface may act as catalysts for charge-transfer reactions. However, it has only recently become possible to investigate the metal-modified semiconductor–electrolyte interface in more detail due to the development of very advanced in situ techniques. If the modification of the semiconductor–electrolyte interface by metals is really necessary for the catalysis of hydrogen or oxygen evolution, and if this modification leads to the formation of a Schottky junction, then the good quality of the photoelectrochemical interface is traded against a low-quality semiconductor–metal junction as pointed out above. The crucial question, therefore, is whether the advantages of both types of contacts may be combined and the disadvantages avoided. This is indeed the case as has been shown only very recently [80] on the basis of ideas developed by Y. Nakato, H. Tsubomura, and coworkers [79], [99]. They proposed that by decreasing the size of metal islands contacting a semiconductor surface, the barrier height of the semiconductor–electrolyte interface surrounding the metal may increase the otherwise too small barrier of the semiconductor–metal Schottky junction. Thereby the high loss currents that normally occur at the semiconductor–metal junction may be avoided. This has now been implemented by Meissner and coworkers us using GaAs

electrodes modified by multiple nano contacts (MNCs) formed from colloidal metal solutions [100]. By carefully avoiding all free metal during the MNC deposition a dramatic potential shift of the photocurrent onset for hydrogen evolution at *p*-GaAs has been realized without increasing the majority carrier loss currents.

6. Economic Aspects

The economic efficiency of an industrial system can most simply be defined as the ratio of costs to effectiveness. The effectiveness of an energy system is generally considered equal to the energy produced by the system. Economic efficiency is thus the ratio of costs to energy produced, that is, the unit energy production cost. This quantity allows a comparison between different energies to be made. The most important components of the energy production cost are (1) capital costs as a function of depreciation life of facilities and interest rates, (2) maintenance, upkeep, and other operating costs including personnel, and (3) fuel costs. The cost structure of solar technologies is generally characterized by high capital costs and low operating costs. Usually no costs are incurred for fuel, unless biomass is used. Economically, the use of solar technology means the substitution of capital for future fuel costs. Investment costs and financing conditions essentially determine the costs, while the present and future prices of fuel saved basically define the economic utility.

Any assessment of the costs of solar technologies must take account of the present level of development (most such technologies are novel). Solar technologies can be classified in three groups [127]:

1) Systems now on the market
2) Systems ready to go on-line
3) Systems still in development

The first group comprises *facilities fully developed* and already in commercial use; the costs and capabilities of these are accurately known. They are manufactured on an industrial scale, though mostly in relatively small quantities. This category includes wind-energy plants ($300-1000\,\mathrm{kW_e}$), small photovoltaic modules ($\leq$ a few $\mathrm{kW_{peak}}$), passive solar systems for new buildings, solar swimming pool heating, solar collectors for water heating, solar-thermal trough collector power plants that have been operate successfully in California since 1986, and last but not least biomass plants in the $\mathrm{MW_{th}}$-size for district heating and cogeneration with steam turbines. If long depreciation periods and low interest rates are assumed, systems of this category are now economically efficient. In order to further improve the competitiveness of these systems, continued development work aims at lowering the manufacturing costs by introducing more rational production methods and increasing the number of units produced. A prerequisite for this is a substantial, self-sustaining growth in demand for solar energy technologies, and given the very low prices for other energies now in effect, this is possible only with governmental support or a change in energy prices.

The group of systems *ready to go on-line* include the following, which have been tested on a pilot scale and demonstrated in individual facilities: larger wind-energy plants ($> 1000\,\mathrm{kW_e}$); photovoltaic power plants (up to $\mathrm{MW_{peak}}$-size); large thermal collector systems on buildings and on the ground, with seasonal storage for heating and hot water (solar district heating); passive systems with transparent thermal insulation; solar process heat systems; paraboloid mirror systems with Stirling engines or gas turbines, and biomass gasification plants to feed engines or gas turbines. At present these systems are not economically efficient in comparison with fossil fuels. Experience based on the operation of test and demonstration facilities is not, by and large, sufficient to indicate accurately the capacity of a plant or guarantee completely reliable service. The technical and economic development objectives for these individual technologies are well-defined and are certainly considered attainable. Essential for the future of these approaches, in particular, is to carry out a sufficient number of demonstration projects under realistic service conditions and to open the commercial markets.

The *third category* of systems, which are, at present, definitely in the development stage, includes large photovoltaic power plants in the multimegawatt range; novel solar cells; trough collector power plants with direct steam generation; solar thermal tower plants; and hydrogen technologies such as fuel cells. These systems cannot be operated economically yet. Their technical and economic potential, however, is undisputed.

Table 4. Investment costs and specific electricity costs for renewable electricity generating systems* [127], [128], [134–136]

	Year								
	1997			2010			2020		
Technology	1 €/kW$_e$	2 €/(kWh/a)	3 €/kWh	1 €/kWe	2 €/(kWh/a)	3 €/kWh	1 €/kW$_e$	2 €/(kWh/a)	3 €/kWh
Location: Germany									
Hydropower plants > 1 MW	2050	0.417	0.042	2045	0.417	0.042	2050	0.417	0.042
Small hydropower plants	4090	1.257	0.118	4090	1.172	0.009	4090	1.105	0.093
Wind turbines	1202	0.6166	0.095	767	0.383	0.059	716	0.358	0.055
PV polycryst. Si, kW-range	8692	10.060	0.978	378	4.112	0.400	2556	2.691	0.262
PV thin film, kW-range	-	-	-	2300	2.501	0.245	1534	1.628	0.159
Cogeneration with biomass-Residues	2812	0.562	0.041	2812	0.562	0.041	2812	0.562	0.041
-Energy plants	2812	0.562	0.133	2812	0.562	0.115	2812	0.562	0.092
Location: South Europe/North Africa									
Solar thermal plant, 200 MW$_e$	2147	0.955	0.095	2659	0.591	0.057	2301	0.511	0.052
PV power plant, 200 MW$_e$	7669	3.335	0.405	3579	1.556	0.189	1790	0.778	0.095

∗ All figures are mean values of a bandwidth of technology size
1 = Investment cost per unit capacity (cost of PV is referred to peak output)
2 = Investment cost per unit generated (= 1 divided by rated hours per year)
3 = Specific cost of electricity (interest rate 6 %, depreciation period 20 years)

Many scenarios for national and world energy supply strategies accord great importance to solar technologies and regenerative forms of energy in general [127, 129–133]. Hydroelectric power, biomass, wind power, geothermal energy, and all technologies for the direct utilization of solar radiation are important options for the future energy economy. They meet both the needs of developmental policy and global environmental requirements. Solar technology costs that are achievable in the medium to long term are therefore significant in the economic assessment, at least as much as current costs. Table 4 summarizes present and future investment and specific electricity costs for some of the most important solar electricity generation technologies.

Figure 59 [135] gives an overview of typical heat production costs (in 0.005 €/kWh$_{th}$) from biomass, biogas, and solar thermal collector systems of different size. Their heat production costs have to be compared with the costs of heat production from natural gas or oil, which are presently about 0.05 €/kWh. Biomass plants utilizing residues or cheap wood are competitive already with heat production from oil and gas. Present solar-thermal collector systems are far too expensive, but they show a large cost reduction potential with growing mass production. Wind energy is an impressive example of technical improvements that have been introduced with simultaneous cost reduction within only a few years (Table 5). The same development can be observed with photovoltaic modules (Figure 60) [134]. Since 1978 the world market of photovoltaic modules has increased at an annual growth rate of about 15 % and has surmounted the 100 MW$_{peak}$/a level in 1997. In the same period the module prize has fallen from about \$ 25 /W$_{peak}$ in 1978 to about \$ 4/W$_{peak}$ at present. Since the time when photovoltaic modules were first introduced for satellite power supplies in the 1970s, costs have even dropped by more than a factor of 100. The trough collector power plants erected in California from 1985 to 1990 (Table 6), too, showed a sharp decline of electricity costs within only a few years. New trough collector power plants of 200 MW$_e$ size built up now could generate solar electricity with \$ 0.1/kWh in an entirely solar mode, and with \$ 0.07/kWh in a hybrid mode where gas produces one half of the total annual electricity (see also Table 4).

In view of the usual intense focus on costs, the economic analysis of solar technology actually reduces to a pure cost analysis, a narrowing that often fails to deal with the complex assessment

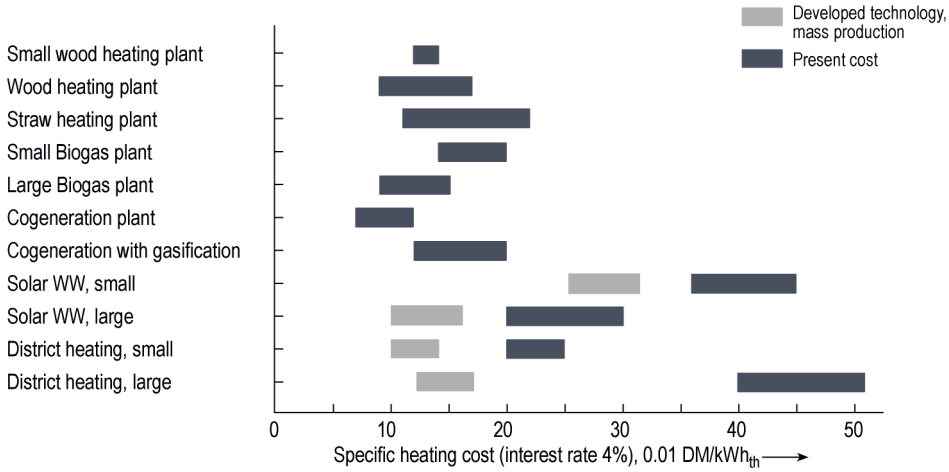

Figure 59. Specific heat production cost of biomass, biogas, and solar thermal collectors (interest rate 4 %, depreciation period 15 a; cogeneration with credit for electricity production)

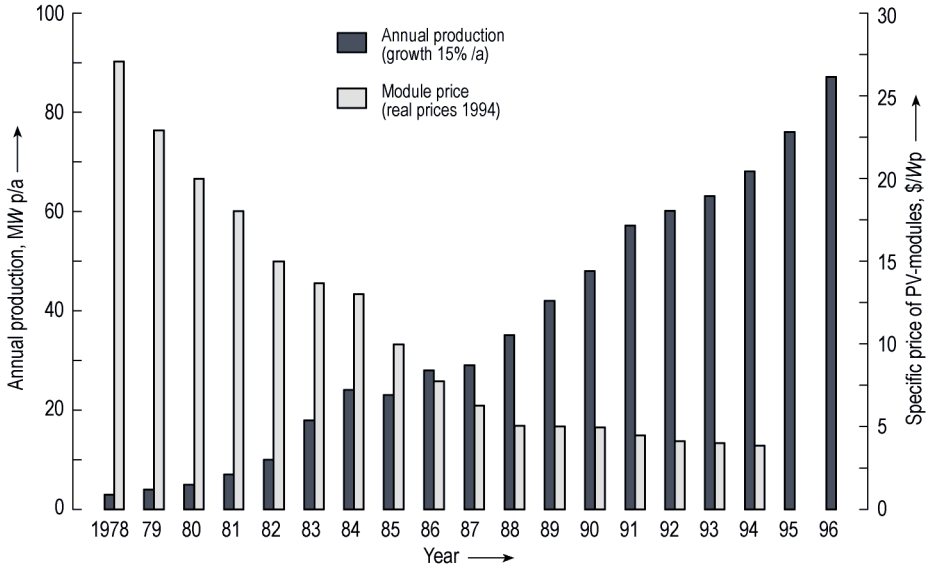

Figure 60. Global market and prices of photovoltaic modules since 1978 [134]

problem. The question of economic efficiency is always a matter of judgment and depends heavily on the situation in which the decision is made as well as the possible alternatives. Alternatives to solar technologies, in most cases, are conventional energy systems based on fossil fuels. A comprehensive evaluation of technical alternatives would have to take in the following points quantitatively or at least qualitatively:

Economic role of the operator
- Private household
- Private company
- Power supply utility
- Municipality
- Other

Purpose of power generation
- Own use
- Sale of power, heat, steam, etc.

Analysis of systems as to energy conservation
- (energy service)

Table 5. Development of plant technology and costs of wind energy [135]

Plant size, kW/a	150 (1990)			300 (1991)			600 (1993)			1500 (1997)	
Rotor diameter, m	23			31			44			63	
Mean annual wind velocity m/s (10 m)	5	6	5		6	5		6	5		6
Annual electricity production, GWh/a	0.31	0.44	0.58		0.87	1.20		1.80	2.91		4.70
Investment costs (€/kW$_e$)	1200			1020			870			840	
Specific electricity costs[a] (€/kWh)	0.146	0.107	0.127		0.08	0.102		0.069	0.09		0.064

[a] Interest rate 6 %, depreciation time 10 years, maintenance, insurance 4 % of investment per year.

These points must be added to the assessment from the viewpoint of the individual enterprise. On the overall economic level, not only the internal costs and constraints considered in present-day business calculations, but also the external or social costs of energy systems and the general economic conditions must be considered [144]. These include:

1) Emission of air pollutants
2) Emission of greenhouse gases
3) Generation of wastes that are difficult or impossible to dispose of
4) Potential hazards due to the facility as well as upstream and downstream processes
5) Noise emissions
6) Land area required
7) Import dependence
8) Reliability of supply
9) Renewability of resources

If damages and other costs are borne by the national economy without having any cost impact on the polluter, the result will be nonoptimal allocation of resources. The state will then have to ensure that competitive imbalances are offset by imposing the external costs. This can be done in two ways: indirectly through regulations that force the operator of an energy facility to take measures to comply with certain limits (an example is the *TA-Luft* air pollution regulation), or directly through user fees and/or taxes (such as fees on carbon dioxide emissions from fossil fuel systems, now under discussion). The external costs of solar technologies are only a small fraction of those of the conventional alternatives when the latter can be estimated at all. This neglection of external costs gives solar technologies a substantial competitive disadvantage, which hinders their introduction in the market.

A number of publications have given projected costs for solar energy generation [142], [126], [127], [128], [131], [136]. Table 7 gives an example. If the goal is that solar energy utilization grows rapidly up to the year 2010 (13 % share), the unit energy cost level would have to double relative to the present for solar technologies to be competitive. Solar heat from a variety of sources would then be, on average, twice as expensive as fossil fuels today (1998). The costs of a solar energy mix (hydro, wind, biomass, biogas, and photovoltaic) would be ca. 50 % higher

Table 6. Technical data and power generation costs for trough collector power plants (Solar Electricity Generating System = SEGS plants I – VIII) in California since 1985 [139], [137], [140].

	Plant							
Data	I	II	III	IV	V	VI	VII	VIII
First full year on line	1985	1986	1987	1987	1988	1989	1989	1990
Net power, MWe	13.8	30.0	30.0	30.0	30.0	30.0	30.0	80.0
Collector field area, 1000 m^2	83	189	230	230	233	188	194	464
Turbine efficiency, %	31.5∗	29.4	30.6	30.6	30.6	37.5	37.5	37.5
Energy output, GWh/a∗∗	30.1	80.5	92.8	92.8	91.8	90.9	92.6	253.4
Energy cost, c/kWh	30.0	22.9	18.0	18.0	18.0	15.5	13.8	12.5

Table 7. Energy generation costs [a] of renewable forms of energy with increasing utilization [126], [143] (€/kWh_{th} or €/kWh_e; 4 % interest rate)

	1988 [b]	2010	2030	2050
Electricity, €/kWh	0.046	0.070	0.114	0.143
Heat, (average)	0.042	0.081	0.157	0.165
Electricity	0.090	0.138	0.223	0.280
Direct heat	0.040	0.077	0.089	0.104
Hydrogen			0.315	0.271
Fuel	0.100	0.200 [c]	0.270 [d]	0.278 [d]
Final energy costs, (average)	0.064	0.123	0.175	0.220
Relative to 1988	1.0	1.92	2.73	3.44
Total final energy consumption (1988 = 1)	1.0	0.69	0.65	0.55
Absolute energy costs (1988 = 1)	1.0	1.33	1.78	1.89
Gross domestic product (GDP) (1988 = 1)	1.0	1.52	1.90	2.28
Ratio energy costs: GDP (1988 = 1)	1.0	0.88	0.94	0.83
Share of renewable energies in final energy consumption, %	2	13	36	69

[a] Includes generation, transport, storage, and distribution.
[b] "Electricity" includes generation and distribution costs; "heat" includes mixture price of fuel oil and gas; "fuels" includes mixture price of gasoline and diesel; fossil energies include consumption taxes but not value-added tax.
[c] Fossil.
[d] Liquid hydrogen from hydro and solar.

than present costs of electricity from conventional sources.

Even as the costs for the individual plants involved in solar technology decline, the growing use of solar energy will result in further cost increases because of greater expenditures on storage and conversion of solar power into a distributable form. If renewable forms of energy gain a 70 % share in the year 2050, the unit cost level will be a factor of 3.4 higher than today (1998). If unit energy expenditures are put together with the final energy consumption figures in this scenario, the absolute outlays for energy will rise only ca. 30 % by 2010 (relative to the present) and 80 – 90 % in the longer term. The fraction of gross domestic product (GDP) spent on energy will remain largely constant, since the GDP will rise further in the projection period. New investigations confirm these basic statements [127], [143]. It can thus be concluded that solar energy systems are affordable for industrialized economies.

Energy price rises now planned, and thus those foreseeable for the economy, will mobilize great potentials for more conservative energy use and will promote innovations that may have beneficial effects not only on the energy industry but on the economy as a whole. Increases in energy prices decreed by the government result from passing on external costs, as mentioned above. As long as the sum of taxes and fees imposed does not exceed the environmental damage costs, a national economy in which solar energies are widely used and environmental pro-

tection is actively pursued will not sustain any economic injury. In the long term, conservation of resources and the environment through the use of solar power, together with the social benefits of such technologies, will represent a major benefit that will justify such expenditures.

7. References

1. G. R. Davis: "Energy for Planet Earth," *Sci. Am.* **263** (1990) no. 3, 21 – 27.
2. WMO/UNEP: Intergovernmental Panel on Climate Change (IPCC), supplement, 1992.
3. W. Palz: "Solar Energy for the World: A Global Photovoltaic Action Programme," Proc. ISES Solar World Congress, Budapest 1994.
4. J. Luther, J. Nitsch: *Energieversorgung der Zukunft,* Springer Verlag, Berlin 1990.
5. R. Sizmann: "Solar Radiation Conversion," in C. Winter et al. (eds): *Solar Power Plants,* Springer Verlag, Berlin 1991, pp. 17 – 81.
6. H. G. Beyer, J. Luther, R. Steinberger-Willms: "Reduction of Fluctuations in Lumped Power Output from Distantly Spaced PV-Arrays," Proc. ISES Solar World Congress, Denver 1991, pp. 321 – 326.
7. D. Heinemann, J. Luther: "Utilization of Weather Forecasts and Climatic Information for the Control of Renewable Energy Systems," Proc. 8th Photovoltaic Solar Energy Conference, Florence, 9 – 13 May, 1988, pp. 249 – 253.
8. C. J. Winter, J. Nitsch (eds): *Hydrogen as an Energy Carrier,* Springer Verlag, Berlin 1988.
9. R. Sizmann: "Thermal Storage Systems – Present Status and Trends," in *Advances in Solar Energy Technology,* Proc. of the Biennial Congress of the International Solar Energy Society, Hamburg 1987.
10. A. Mouchot: *La Chaleur Solaire et ses Applications Industrielles,* Gauthiers-Villars, Paris 1869.
11. O. Kausch: *Die unmittelbare Ausnutzung der Sonnenenergie,* Carl Steinert Verlag, Weimar 1920.
12. H. Rau: *Heliotechnik – Sonnenenergie in praktischer Anwendung,* Udo Pfriemer Verlag, München 1975.
13. A. B. Meinel, M. P. Meinel: *Applied Solar Energy,* 2nd ed., Addison Weslay Publ. Comp., Reading 1977.
14. R. Gööck: *Die großen Erfindungen – Wind Wasser Dampf Sonne,* Sigloch Edition, Künzelsau 1989.
15. S. L. Valley: *Handbook of Geophysics and Space Environment,* McGraw-Hill, New York 1968.
16. B. Stoy: *Wunschenergie Sonne,* 3rd ed., Energie-Verlag, Heidelberg 1980.
17. J. A. Duffie, W. B. Beckmann: *Solar Engineering of Thermal Processes,* 2nd ed., John Wiley & Sons, New York 1991.
18. F. Kreith, J. F. Kreider: *Principles of Solar Engineering,* Hemisphere Publishing, London 1978.
19. J. O. Dalenbäck (ed.): "*Central Solar Heating Plants with Seasonal Storage: Status Report*" Document D14 : 1990, Swedish Council for Building Research, Stockholm 1990.
20. M. N. Fisch, M. Guigas, J. O. Dalenbäck: *Large-Scale Solar District Heating – Status and Future in Europe, Proceedings EuroSun '96,* DGS-Sonnenenergie Verlag, München 1996 pp. 157 – 170.
21. M. Nast, J. Nitsch: "Möglichkeiten und Grenzen Solarer Nahwärme", *VDI Berichte Nr.* **1182** (1995) 725 – 741.
22. R. Kübler, M. N. Fisch, E. Hahne: "Central Solar Heating Plants with Seasonal Storage: Results of Site Specific Study in Germany," *Proceedings Solar World Congress,* Pergamon Press, London 1991, pp. 1505 – 1510.
23. W. C. Dickinsen, P. N. Cheremisinoff (eds.): *Solar Energy Technology Handbook,* Marcel Dekker Inc., New York 1980.
24. C. J. Winter, R. L. Sizmann, L. L. Vant-Hull: *Solar Power Plants,* Springer Verlag, Berlin 1991.
25. M. Kleemann, M. Meliß: *Regenerative Energiequellen,* Springer Verlag, Berlin 1988.

General References

26. P. Archards, R. Gicquel (eds): *European Passive Solar Handbook,* Publication no. EUR 10 683 of the Commission of the European Community, Brussels 1986. J. R. Goulding, J. O. Lewis, T. C. Steemers (eds): *Energy Conscious Desgin, A Primer for Architects,* Publication no. EUR 13 445 of the Commission of the European Community, Brussels 1992. J. R. Goulding, J. O. Lewis, T. C. Steemers (eds): *Energy in Architecture, The European Passive Solar Handbook,* Publication no. EUR 13 446 of the Comission of the European Community, Brussels 1992. M. S. Sodha et al.: *Solar Passive Building,*

Science and Design, Pergamon Press, New York 1986. *ASHREA Handbook, 1985 Fundamentals,* American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, Tullie Circle, 1985. O. Humm: *Niedrigenergiehäuser, Theorie und Praxis,* Verlag ökobuch, Staufen b. Freiburg 1990. W. Koblin, E. Krüger, U. Schuh: "Bau- und Wohnforschung," *Handbuch, Passive Nutzung der Sonnenenergie,* Schriftenreihe 04 des Bundesministers für Raumordnung, Bauwesen und Städtebau, Bonn 1984.

Specific References

27. V. Wittwer: "Transparent Insulation Materials and Their Application in Active and Passive Systems," *Int. J. Solar Energy* **11** (1992) 117 – 134.
28. K. Jahn, D. Christoffers, K. R. Schreitmüller: "Thermal Performance of a Transparently Insulated South Facade," *Proc. of the 3rd Eur. Conf. on Architecture,* Florence 1993.
29. E. Boy, S. Meinhardt: "A New Thermo – Optical Regulation System – Suited for Transparent Insulation?," *Proc. of the 2nd Int. Workshop on Transparent Insulation,* Freiburg 1988.
30. T. C. Steemers, C. den Ouden (eds.): "Daylighting, Atria, Cooling," *Proc. Building 2000 Workshops,* Commission of the European Community, Brussels 1988.
31. *PROJEKT MONITOR, A Series of Case Studies Illustrating Passive Solar Architecture in the European Community,* School of Architecture, University College Dublin, Richview, published monthly.
32. D. van Hattem (ed.): "Non Nuclear Energies," *Workshop on Future Building Energy Modelling,* Commission of the European Community, Brussels 1988.
33. B. Stanzel, E. Hahne: "Comparison of Four Simulation Programs for the Passys Test Cell," *Tagungsber. Int. Sonnenforum 6th* **1** (1988) 317 – 321.
34. D. Christoffers, "Seasonal Shading of Vertical South Facades with Prismatic Panes," *Sol. Energ.* **57** (1996) Nov., N. 5, 339 – 343.
35. D. M. Chapin, C. S. Fuller, G. L. Pearson, *J. Appl. Phys.* **25** (1954) 676.
36. P. D. Maycock, *PV News* **12** (1993) no. 2, Feb.
37. P. D. Maycock, *PV News* **11** (1992) no. 8, Aug.
38. H. W. Schock, F. Pfisterer: "Photovoltaics in Europe – Recent Developments" in R. Noufi (ed.): *AIP Conf. Proc. 268, "Photovoltaic Advanced Research & Development Project",* Am. Inst. of Physics, New York 1992, p. 511.
39. R. Curry (ed.), *Photov. Insider's Rep.* **XI** (1992) no. 8, Aug.
40. L. M. Fraas et al., "Over 30 % Efficient Tandem Gallium Cell Assemblies for Use with Concentrated Sunlight," *Techn. Digest of the 5th Int. Photov. Sci. and Eng. Conf. (PVSEC-5),* Kyoto, Japan 1990, p. 89.
41. W. H. Bloss, W. Knaupp, F. Pfisterer, "Photovoltaic-Testanlage Widderstall," *VDI Ber.* **851** (1991) 135 – 149.
42. W. Knaupp, "Power Rating of Photovoltaic Modules from Outdoor Measurements," *Conf. Rec. 22nd IEEE Photov. Spec. Conf.,* Las Vegas 1991, p. 848.
43. S. Nann, *Sol. Energy* **45** (1990) 385.
44. G. Chryssis: *High Frequency Switching Power Supplies: Theory and Design,* McGraw-Hill, New York 1984.
45. L. Keating: "Computer Simulation and Expert Systems for Photovoltaic Applications," Ph. D. Thesis, NMRC University College, Cork (IR) 1992.
46. C. Blok, E. Terhorst: "SOMES – A Simulation . . .", Technical Report, University of Utrecht, 1987.
47. D. F. Menicucci, J. P. Fernandez: "User's Manual for PVFORM, a Photovoltaic System Simulation . . .", Technical Report, SAND85-0376, Sandia Nat. Laboratories, Albuquerque 1988.
48. S. McCarthy, L. Keating, G. T. Wrixon, "Development of ASHLING, a PV Plant Simulation Program", Technical Report, NMRC Univ. College, Cork (IR) 1990.
49. K. W. Stokes, Technical Report EPRI-AP-5229, Electric Power Research Institute, Palo Alto, CA, 1987.
50. G. Blaesser, K. Krebs: "Summary of PV Pilot Plant Monitoring Data 1984 – 1985," *Proc. 7th E.C. PV Solar Energy Conf.* , Sevilla 1986, D. Reidel, Dordrecht 1987, p. 84.
51. Proc. Execut. Conf., *Photov. Systems for Electric Utility Appl.,* Taormina, Italy, Dec. 1990 (OECD, Paris, 1992).
52. U. Beyer et al.: "Photovoltaikanlage Kobern – Gondorf . . .," *Proc. 5 Nat. Symp. PV Solarenergie,* Staffelstein 1990, p. 59.
53. J. E. Hoffner et al., *Proc. 1987 Ann. Meeting Am. Sol. En. Soc. (ASES),* Portland, Oregon, 1987, p. 121.
54. W. H. Bloss, F. Pfisterer, "Photovoltaische Systeme–Energiebilanz und CO2-Reduktionspotential", *VDI Ber.* **942** (1992) 71 – 87.

General References

55. A. J. Bard, R. Memming, B. Miller, *Pure & Appl. Chem.* **63** (1991) 569 – 596.
56. D. Meissner in D. Meissner (ed.): *Solarzellen, Physikalische Grundlagen und Anwendungen in der Photovoltaik,* Vieweg, Braunschweig 1993, pp. 137 – 150.
57. R. Memming in E. Steckhan (ed.): *Electrochemistry II, Topics in Current Chemistry,* **vol. 143,** Springer Verlag, Berlin 1988, pp. 79 – 112.
58. S. R. Morrison: *Electrochemistry at Semiconductor and Oxidized Metal Electrodes,* Plenum, New York 1980.
59. Y. V. Pleskov: *Solar Energy Conversion, a Photoelectrochemical Approach,* Springer Verlag, Berlin 1990.
60. K. Rajeshwar, L. M. Peter, A. Fujishima, D. Meissner, M. Tomkiewicz: *Photoelectrochemistry,* **Proceed. Vol. 97 – 20,** The Electrochem. Soc., Inc., Pennington, NJ, 1997.
61. H. Tsumbomura, H. Kobayashi, *Crit. Rev. Solid State Mater. Sci.* **18** (1993) 261 – 326.
62. H. J. Lewerenz, H. Jungblut: *Photovoltaik, Grundlagen und Anwendungen,* Chapter 5, Springer Verlag, Berlin 1995.
63. K. Rajeshwar, L. M. Peter, A. Fujishima, D. Meissner, T. Tomkiewich: "Photochemistry," *The Electrochemical Society Proceedings,* **vol. 97-20,** Pennington, NJ 1997.

Specific References

64. R. Memming, W. Gissler in *Proceedings of the European Conference on Solar Cells,* Luxembourg 1977.
65. A. Fujishima, K. Honda, *Nature* **238** (1972) 37 – 38.
66. E. Bequerel, *C. R. Des. Scéances Naturelles* **9** (1939) 561.
67. R. Williams, *J. Chem. Phys.* **32** (1960) 1505 – 1514.
68. W. H. Brattain, G. C. B. Gobrecht, *Bell Syst. Tech. J.* **34** (1955) 129.
69. D. Meissner in T. N. Veziroglu (ed.): *Hydrogen Energy Progress IX,* **vol. 1,** M. C. I, Paris 1992, pp. 517 – 526.
70. B. O'Regan, M. Grätzel, *Nature* **353** (1991) 737 – 740.
71. K. Nakatani, S. Matsudaira, H. Tsubomura, *J. Electrochem. Soc.* **125** (1978) 406.
72. A. Heller et al., *J. Electrochem. Soc.* **125** (1978) 1156.
73. S. Licht, D. Peramunage, *Nature* **345** (1990) 330, *J. Electrochem. Soc.* **139** (1992) L23 – L26, 15.
74. S. Licht et al., *Appl. Phys. Lett.* **46** (1985) 608.
75. K. C. Chang et al., *Science* **196** (1977) 1097. B. A. Parkinson, A. Heller, B. Miller, *J. Electrochem. Soc.* **126** (1979) 954.
76. L. G. Casagrande, N. S. Lewis, S. M. Kelso, *170th Electrochem. Soc. Meeting,* San Diego 1986; Abstr. no. 607 (1986).
77. A. Heller, B. Miller, F. Thiel, *Appl. Phys. Lett.* **38** (1981) 232.
78. R. Tenne, A. Wold, *Appl. Phys. Lett.* **47** (1985) 707.
79. Y. Nakato, K. Ueda, H. Yano, H. Tsubomura, *J. Phys. Chem.* **92** (1988) 2316.
80. H. Tsubomura, Y. Nakato, *Nouv. J. de. Chimie* **11** (1987) 167 – 171.
81. M. Rosenbluth, N. S. Lewis, *J. Am. Chem. Soc.* **108** (1986) 4689.
82. D. Cahen, Y. W. Chen, *Appl. Phys. Lett.* **45** (1984) 746.
83. H. Tributsch in M. Schiavello (ed.): *New Trends and Applications of Photocatalysis and Photoelectro-chemistry for Environmental Problems,* D. Reidel, Dordrecht 1988, p. 297.
84. D. Meissner, R. Memming, *Electrochim. Acta* **37** (1992) 799 – 809.
85. H. Gerischer, H. Tributsch, *Ber. Bunsenges. Phys. Chem.* **72** (1968) 437.
86. R. Memming in S. G. Davison (ed.): *Progress in Surface Science,* **vol. 17,** Pergamon, Oxford 1983, pp. 7 – 74.
87. B. A. Parkinson, M. T. Spitler, *Electrochim. Acta* **37** (1992) 943 – 948.
88. N. Matsumura, Y. Nomura, H. Tsubomura, *Bull. Chem. Soc. Jpn.* **50** (1977) 2533.
89. C. Stalder, J. Augustynski, *J. Electrochem. Soc.* **126** (1979) 2007.
90. J. Desilvestro et al., *J. Am. Chem. Soc.* **107** (1985) 2988 – 2990.
91. D. Bahnemann, D. Bockelmann, R. Goslich, *Solar Energy Mat.* **24** (1992) 564 – 583.
92. E. Borgarello, J. Kiwi, E. Pelizzetti, M. Visca, M. Grätzel, *Nature* **289** (1981) 158. K. Kalyanasundara, M. E. Borgarello, D. Duonghong, M. Grätzel, *Angew. Chemie* **93** (1981) 1012.
93. A. Harriman, G. Porter, P. Walters, *J. Photochem.* **19** (1982) 183 – 187. D. Meissner, R. Memming, B. Kastening, D. Bahnemann, *Chem. Phys. Lett.* **127** (1986) 419 – 423.
94. R. Reineke in A. Winsel (ed.): "Elektrochemie in Energie- und Umwelttechnik", *Dechema Monorgr.* **124** (1991) 623 – 627.
95. T. Inoue, T. Yamase, *Chem. Lett.* (1985) 869 – 872.

96. O. Kaselev, J. M. Turner, *J. Electrochem. Soc.* **145** (1998) 3335 – 3339; E. Rocheleau, E. L. Miller, A. Misra, *Energy & Fuels* **12** (1998) 3 – 10.
97. J. R. White, F. R. F. Fan, A. J. Bard, *J. Electrochem. Soc.* **132** (1985) 544 – 550.
98. F. Möllers, H. J. Tolle, R. Memming, *J. Electrochem. Soc.* **121** (1974) 1160.
99. Y. Nakato, H. Tsumomura, *Electrochim. Acta* **37** (1992) 789 – 907.
100. A. Meier, I. Uhlendorf, D. Meissner, *Proc. Symp. on Electrochemical Technology Applications in Electronics,* The Electrochem. Soc., Washington 1993, in press. and in B. D. Struck (ed.): *IEA Program of Research and Development on the Production of Hydrogen from Water,* Annual Progress Report 1993, KFA Jülich, in press.
101. Y. Nakato, H. Tsubomura, *Electrochim. Acta* **37** (1992) 897 – 907.
102. N. S. Lewis, *Appl. Phys. Lett.* **45** (1984) 1095 – 1097.
103. M. L. Rosenbluth, C. M. Lieber, N. S. Lewis, *Appl. Phys. Lett.* **45** (1984) 423 – 425.
104. H. Kobayashi, Y. Kogetsu, K. Ikezumi, Y. Nakato, *J. Electroanal. Chem.* **371** (1994) 53 – 58.
105. G. Hodes, *Nature* **285** (1980) 29 – 30.
106. B. A. Parkinson, A. Heller, B. Miller, *Appl. Phys. Lett.* **33** (1978) 521 – 523. *J. Electrochem. Soc.* **126** (1979) 954 – 960.
107. A. Heller, J. Lewerenz, B. Miller, *Ber. Bunsenges. Phys. Chemie* **84** (1980) 592 – 595.
108. A. T. Howe, *J. Chem. Soc. Chem. Commun.* (1983) 1407 – 1408.
109. H. J. Lewerenz, H. Goslowsky, K. D. Husemann, S. Fiechter, *Nature* **321** (1986) 687.
110. G. Kline, K. Kam, D. Canfield, B. A. Parkinson, *Solar Enrgy. Mater.* **4** (1981) 301 – 308.
111. W. Gissler, R. Memming, Proceedings Europ. Conf. on Solar Cells, Luxembourg, 1972, pp. 425 – 435.
112. G. Hodes, J. Manassen, D. Cahen, *Nature* **261** (1976) 403 – 404.
113. S. Licht, J. Manassen, *J. Electrochem. Soc.* **134** (1987) 1064 – 1070.
114. S. Licht, G. Hodes, R. Tenne, J. Manassen, *Nature* **326** (1987) 863 – 863.
115. A. Meier, D. Meissner in J. H. Fendler, I. Dékány (eds.): *Nanoparticles in Solids and Solutions,* Kluwer, Dordrecht 1996, pp. 421 – 449.
116. A. Meier, D. Meissner, *Electrochim. Acta* **40** (1995) 1523 – 1535; R. Hiessen, D. Meissner, *J. Phys. Chem.* **102** (1998) 6549 – 6557.
117. G. Porter, *Nature* **288** (1980) 320.
118. W. Wettling, *Phys. Bl.* **53** (1997) 1197 – 1202.
119. D. Meissner, *Photon* 1997,März/April 24 – 27.
120. A. Fujishima, L. A. Nagahara, K. Ajito, K. Hashimoto, *Electrochim. Acta* **39** (1994) 1229 – 1236.
121. M. R. Hoffmann, S. T. Martin, W. Choi, D. W. Bahnemann, *Chem. Rev.* **95** (1995) 69 – 96.
122. H. Kiesch et al., *Angew. Chem.* **110** (1998) 3201 – 3203.
123. E. S. Smotkin et al., *J. Phys. Chem.* **91** (1987) 6 – 8.
124. D. Peramunage, S. Licht, *Solar Energy* **52** (1994) 197 – 204.

General References

125. M. Kaltschmitt, A. Wiese (eds.): *Erneuerbare Enerfien–Systemtechnik, Wirtschaftlichkeit, Umweltaspekte,* 2nd ed., Springer Verlag, Berlin 1997.
126. J. Nitsch et al.: *Bedingungen und Folgen von Aufbau strategien für eine solare Wasserstoffwirtschaft.* Untersuchung für die Enquete-Kommission "Technikfolgen-Abschätzung und -Bewertung," Dt. Bundestages, Bonn, June 1990.
127. G. Altner, H.-P. Dürr, G. Michelsen, J. Nitsch: *Zukünftige Energiepolitik,* Economica Verlag, Bonn 1995.
128. M. Kaltschmitt, A. Wiese (eds.): *Erneuerbare Energien – Systemtechnik, Wirtschaftlichkeit, Umweltaspekte,* 2nd ed., Springer Verlag, Berlin 1997.
129. J. Nitsch, J. Luther: *Energieversorgung der Zukunft. Rationelle Anwendung und erneuerbare Quellen,* Springer Verlag, Berlin 1990.
130. H. G. Brauch (ed.): *Energiepolitik – Handlungskonzepte zu erneuerbaren Energien und zur rationellen Energienutzung,* Springer Verlag, Berlin 1997.
131. T. B. Johansson, H. Kelly, A. K. N. Reddy, R. H. Williams (eds.): *Renewable Energy – Sources for Fuels and Electricity,* Island Press, Washington DC, 1993.
132. Umweltbundesamt: *Nachhaltiges Deutschland – Wege zu einer dauerhaft-umweltgerechten Entwicklung,* Erich Schmidt Verlag, Berlin1997.

Specific References

133. Enquete Kommission "Schutz der Erdatmosphäre" des Dt. Bundestages (ed.): *Energie,* **vol. 3 (in two volumes),** Economica-Verlag, Bonn 1995.
134. F. Staiß: *Photovoltaik – Technik, Potentiale und Perspektiven der solaren Stromerzeugung,* Vieweg Verlagsgesellschaft, Braunschweig – Wiesbaden 1996.
135. J. Nitsch, O. Langniß: "Kosten und Märkte regenerativer Energieanlagen" in *VDI-Bericht* **1361,** VDI-Verlag, Düsseldorf 1997, p. 3 – 28.
136. F. Trieb, B. Milow, J. Nitsch: "Einführung solarthermischer Kraftwerke auf den Weltenergiemarkt," *Energiewirtschaftliche Tagesfragen* **48** (1998) 6, 392 – 397.
137. H. Klaiß, F. Staiß (eds.): *Solarthermische Kraftwerke für den Mittelmeerraum,* **vol 1+2,** Springer Verlag, Berlin 1992.
138. DIW Berlin, ISI Karlsruhe: *Kostenaspekte erneuerbarer Energiequellen in der BRD und auf den Exportmärkten,* BMFT, Berlin, Feb. 1991.
139. M. Geyer, H. Klaiß: "194 MW Solarstrom mit Rinnenkollektoren," *Brennst. Wärme Kraft* **41** (1989) no. 6, 288 – 295.
140. G. Cohen, D. W. Kearney, R. G. Cable: "Recent Improvements and Performance Experience at the Kramer Junction SEGS Plants" in *VDI-Bericht* **1200,** Anhang, VDI-Verlag, Düsseldorf 1995.
141. Enquete-Kommission Vorsorge zum Schutz der Erdatmosphäre des Dt. Bundestages (ed.): *Energie und Klima,* **vol. 7, Energieszenarien,** Economica Verlag, Bonn, Müller, Karlsruhe 1990.
142. P.-M. Nast: "Solarkollektoren und solare Nahwärmesysteme" in Enquete-Kommission Vorsorge, zum Schutz der Erdatmosphäre, Dt. Bundestages (ed.): "Erneuerbare Energien," *Energie und Klima,* vol. 3, Economica Verl., Bonn; Müller, Karlsruhe 1990.
143. O. Langniß, J. Luther, J. Nitsch, E. Wiemken: *Strategien für eine nachhaltige Energieversorgung – ein solares Langfristszenario für Deutschland,* DLR, Stuttgart, FhG, Freiburg, October 1997.
144. O. Hohmeyer: *Soziale Kosten des Energieverbrauchs,* 2nd ed., Springer Verlag, Berlin 1989.

Soldering and Brazing

Welding and Cutting is a separate keyword.

Harald Krappitz, Degussa AG, Hanau, Federal Republic of Germany

1. Introduction 1
2. Theory 2
3. Soldering and Brazing Processes . . . 3
3.1. Soldering Bits and Blocks 3
3.2. Soldering Baths (Dip Soldering) 3
3.3. Wave Soldering 4
3.4. Reflow Soldering 4
3.5. Flame Soldering and Brazing 4
3.6. Induction and Resistance Brazing . . 4
3.7. Salt-Bath Brazing 4
3.8. Furnace Brazing 4
4. Soldering and Brazing Materials . . . 5
4.1. Soft Solders 5
4.2. Brazing Filler Metals 7
5. Fluxes and Protective Atmospheres . 8
6. Materials of Construction Suitable for Soldering and Brazing 9
7. Design and Strength Properties of Soldered and Brazed Joints 10
8. Safety Aspects 11
9. References 11

1. Introduction

Soldering and brazing are thermal processes for joining and coating materials, in which a liquid phase is produced by melting a solder or a brazing filler metal or by diffusion to the boundary surfaces (e.g., to form a eutectic). The temperature remains below the solidus temperature of the base material (DIN 8505, Part 1).

Alloys or pure metals are used as solders or brazing alloys. Where their liquidus temperature is <450 °C the process is called soft soldering, and where it is >450 °C it is known as *brazing*. In *high-temperature brazing*, the operation is carried out in the absence of a flux and air, and the bonding metal has a liquidus temperature >900 °C.

A major difference between soldering and welding is that only the filler material (i.e., the solder or brazing alloy), and not the base material, is melted. This means that solders must have an appreciably lower melting range than the base material. In industrial soldering and brazing, groups of specially developed alloys are usually used (see Chap. 4).

The molten filler metal must come into intimate contact with the base material to allow formation of metallic bonds. Oxide layers exist on all industrial metal surfaces and must therefore first be broken down and removed. When soldering or brazing in air, this is done by covering the area to be joined with a flux in which the oxides dissolve at the working temperature. When brazing in the absence of air, i.e., in a protective gas or vacuum, the oxide layers can be cracked under suitable conditions without a flux.

If the base material has a clean metallic surface, it can be wetted by the liquid filler metal provided that solid solutions or intermetallic compounds are formed. The filler metal then spreads over the area to be joined, and solidifies to form a strong bond with the base metal. The lowest surface temperature at which this process takes place is known as the *working temperature*, and is a characteristic parameter for each filler metal.

Brazability denotes the capability of a workpiece to be manufactured to given specifications by a brazing process (DIN 8514). It is a function of

1) The suitability of the base and filler materials needed for a brazing operation
2) The stability of the brazed joint as a function of design
3) The possibility of including a brazing operation in the manufacturing process

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_427

These factors can influence each other. If only one of these conditions is not completely satisfied, the brazability of the workpiece may be jeopardized.

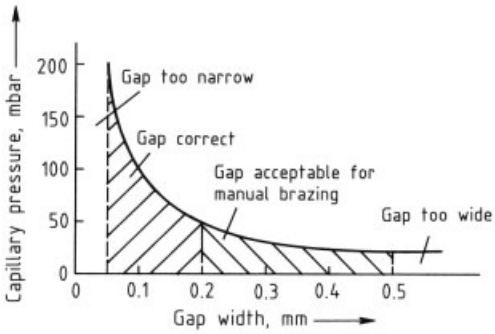

Figure 1. Capillary pressure of a silver solder at 650 °C as a function of gap width

2. Theory

Soldering and brazing are processes in which materials are thermally joined by melting a supplementary material (solder or brazing alloy); the base material is not, however, melted. This definition is not adequate because the liquid filler metal cannot wet the substrate unless at least three other conditions are satisfied:

1) The joining surface and the filler metal must be bright (i.e., clean and free of oxides)
2) The joining surface and the filler metal must be at or above the working temperature
3) At least one component of the filler metal must react exothermically with the substrate to form an alloy or compound

The surface of the workpiece to be joined is heated as uniformly as possible until the working temperature is exceeded. There should be an even temperature distribution; this is not necessary for welding.

If the conditions for wetting are satisfied, a mixed crystal or a compound is formed between the substrate and the liquid filler metal, starting in a region of only a few atomic layers. The liberated energy is sufficient to overcome the surface tension of the drops of solder, allowing them to spread. As systems always tend towards a state of minimum potential energy, the filler metal spreads and fills the gap between the jointing faces of the workpieces to be bonded. The filler metal spontaneously penetrates into narrow gaps against the force of gravity, the force responsible for this is termed *capillary tension*. The magnitude of the capillary attraction depends mainly on the width and geometry of the gap to be joined. In the case of a parallel gap, it is inversely proportional to the width (Fig. 1). If the cross section of the capillary is triangular (e.g., when a knurled bolt is brazed into a drilled hole), the capillary pressure is approximately three times the filling pressure for a parallel-sided gap.

If a flux is used during brazing or soldering, the minimum gap width is determined by the minimum volume of flux which must be available in the gap to ensure dissolution of the oxide layer on the surface. Practical experience shows that this is assured for gap widths as low as 0.05 mm. If brazing is carried out in a protective gas or vacuum, gaps can be smaller, and can be so narrow that the surfaces come into contact. In this case the capillary pressure is therefore high and ensures spontaneous, reliable filling of the gap with solder. The minimum gap width is then determined by the metallurgical processes, including alloy formation, that take place between the filler metal and the substrate [7]. Isothermal solidification of the filler metal may occur before the gap is completely filled, especially in narrow gaps.

The maximum gap width is determined by the capillary pressure necessary for the liquid filler metal to penetrate against the force of gravity, and to stay in place. If the molten filler metal is to flow spontaneously into the gap (mechanized brazing or furnace brazing), the width must not exceed 0.2 mm. In the case of manual soldering or brazing, a gap as high as 0.5 mm is permissible, since the operator is able to correct irregularities. Indeed, gap widths exceeding 0.5 mm can only be filled with filler metal manually, and require special skills. In this case, the surfaces to be joined are not heated evenly, heat is applied to the workpiece at individual points, as with gas welding. This technique is known as braze welding to distinguish it from gap brazing. It is not of great practical significance compared with gap brazing, apart from its use for installing galvanized steel pipes.

Apart from its use as a bonding technique, soldering as well as brazing is used for coating surfaces with filler metal or other components. The coating can be used for protection

against corrosion or abrasion. Alternatively, it can be employed as a method of constructing or shaping workpieces (e.g., uneven surfaces in the steel bodywork of an automobile can be smoothened by coating with soft solder, eroded turbine blades can be repaired by coating with high-melting brazing alloys).

In the production of abrasive tools or nonslip surfaces, particles of hard materials (e.g., tungsten carbide, ceramics, or diamond) are embedded in a braze matrix and applied to the surface. Nonmetallic hard materials can be wetted with special filler metals [8].

Abrasion-resistant coatings can be produced by brazing the surface with very hard materials (e.g., nickel-based brazing alloys). Abrasion resistance can be further improved if finely divided hard materials such as tungsten or chromium carbides are incorporated into the coating [9].

3. Soldering and Brazing Processes

Soldering and brazing processes can be classified on the basis of different criteria, for example according to the type of filler metal, as described in Chapter 1: soft solders, brazing filler metals or high-temperature brazing alloys. Classification according to coating by brazing and bonding by brazing has also been mentioned (Chap. 2). Soldering and brazing processes are described in this chapter according to their energy source.

3.1. Soldering Bits and Blocks

When soft soldering, the parts to be joined can be raised to the required temperature by heating with a soldering bit or "iron," which is usually made of copper. The bit generally has a pointed end whose geometry matches the shape of the workpiece. Heat transfer takes place via the molten solder and the flux. The flux can be present as a filling inside a "cored" solder, or added separately. Soldering with a bit is usually performed manually, but automatic devices are also known.

In block soldering heat is applied evenly to the surface of the workpiece. The parts to be joined are usually presoldered, and are placed on a heated metal block (e.g., a hotplate) until the soldering temperature is reached. This method is useful for soldering heavy-section components to thin metal sheet [10].

3.2. Soldering Baths (Dip Soldering)

In soldering baths, which are usually heated electrically, the liquid solder transfers the heat. A flux is first applied to the workpiece, which is then immersed in the soldering bath. In the bath solder is applied simultaneously at several joints in a single dipping operation. Baths are therefore used in electronics as a simple means of automating soft soldering (e.g., motor armatures and simple electronic circuits). The soldering temperature is precisely controlled by adjusting the temperature of the bath. This process is mainly used with soft solders, but there are some applications in which baths of brazing filler metal are used (e.g., for preplacing filler metal for a subsequent brazing operation). These baths are covered with a layer of molten flux to protect them from oxidation.

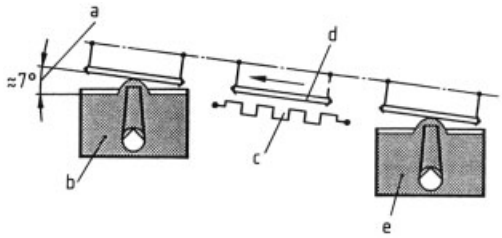

Figure 2. Principle of wave soldering
a) Pull-through angle; b) Solder bath with solder wave; c) Drying zone; d) Printed circuit board; e) Flux bath (fluxing station with foam wave)

3.3. Wave Soldering

In wave soldering a flow pattern is produced in a soft solder bath by a pump in such a way that a stationary wave is formed on the surface of the bath. This soldering process is typically used for mass production of circuit boards for electronic components. The wave soldering bath is usually preceded by a flux bath with a similarly generated surface wave (Fig. 2).

3.4. Reflow Soldering

The reflow process involves the remelting of solder which has been preplaced on the area to

be soldered before final soldering takes place. Heating is carried out in various ways, e.g., by high-temperature oil, steam, or IR radiation. In contrast to the soldering methods described in Sections 3.1, 3.2, 3.3, the placing of the solder and the heating of the workpiece take place at different times. The process has become more important with the increasing miniaturization of electronic components. For example, a solder – binder paste, with or without a flux, is applied by screen printing in exactly controlled amounts and conforming exactly to the shape of the workpiece. The advantages of this procedure are exploited in the surface-mounted devices (SMD) technique, in which the components do not have connector wires, but are soldered flat onto the surface of a circuit board on which the solder has been preplaced.

3.5. Flame Soldering and Brazing

In flame soldering and brazing, heating is carried out by the combustion of a gaseous mixture (e.g., natural gas, propane, or acetylene with air or oxygen). Special burners are used either directly or indirectly to heat the areas to be joined to the working temperature. For soft soldering, indirect heating is preferable because a direct flame can destroy the soft-solder flux.

The flame should be adjusted so that it is neutral or slightly reducing. An excess of oxygen must be avoided because of possible oxidation of the joining surfaces.

The size of the burner should be adjusted to the dimensions and thermal properties of the workpiece. During manual soldering or brazing, the burner is moved to provide even heating of the surfaces to be joined, whereas in automated soldering or brazing the workpiece is usually moved.

3.6. Induction and Resistance Brazing

These brazing methods are frequently used in the tooling industry to braze cemented carbide tips of cutting tools. In these brazing processes, electrical energy is converted into heat due to the electrical resistance or contact resistance of the workpiece. In induction brazing, alternating currents are induced in the surface regions of the workpiece. The depth to which these induced currents penetrate into the workpiece decreases with increasing frequency.

Resistance brazing can be carried out by direct resistance heating of the items to be joined, or by indirect heating via special electrodes.

3.7. Salt-Bath Brazing

In salt-bath brazing, the parts to be joined are heated by immersing them in a bath of molten salt (e.g., alkali metal fluorides) that usually possesses fluxing properties. The filler is placed on the parts to be joined before immersion in the bath. This method is widely used for brazing aluminum heat exchangers. Advantages include precise temperature control and even temperature distribution within the workpiece. After brazing, the salt residues must be removed by thorough washing to prevent corrosion; this is a disadvantage of the process from the environmental point of view. The salt-bath process is therefore being replaced by furnace brazing.

3.8. Furnace Brazing

In furnace brazing, the parts to be joined are heated by radiation and by the convection of hot gases in an electrically heated furnace. As manual handling during the brazing process is not possible, the workpieces must be assembled and secured and the filler metal put in position in a preliminary operation. Special equipment is often used to ensure accurate positioning of the surfaces being joined; other possibilities include fixing the parts by tack welding or using parts with a self-centering design. Shaped pieces of filler metal (preforms) can be put in position as a preliminary operation, or the filler metal can be applied to the gap in the form of paste.

Although fluxes are sometimes used in furnace brazing, one of the main advantages of the process is that the use of fluxes can be avoided when suitable protective gas atmospheres are used. The protective gases can be either reducing atmospheres (e.g., H_2 or $H_2 - N_2$ mixtures) or inert atmospheres (e.g., Ar or He), which in some cases must satisfy extremely high purity specifications that depend on the filler metal and the base material. Removal of flux residues is, of course, superfluous when parts are brazed in a protective gas without fluxes. Also, the production of scale on the workpieces is prevented

by the protective gas atmosphere. Brazing in a protective gas can be carried out in continuously operating furnaces in which a conveyor belt of heat-resistant woven wire passes through a tunnel furnace. Brazed parts can therefore be economically mass produced.

Flux-free brazing can also be carried out under vacuum. The quality of the vacuum required is determined by the filler and base materials, and can be between 10 Pa (10^{-1} mbar) and 10^{-4} Pa (10^{-6} mbar). Vacuum brazing can only be carried out batchwise in closed chamber furnaces.

A further advantage of all furnace brazing processes is that precise and reproducible temperature control results in very consistent brazing. Scaling is also minimized by the even temperature distribution of the workpieces. For these reasons, the process is growing in technical and economic importance.

4. Soldering and Brazing Materials

Some properties and uses of selected solders and brazing alloys are given in Table 1.

4.1. Soft Solders

Soft solders have a liquidus temperature below 450 °C. They comprise three groups, which have different areas of application:

1) Lead – tin and tin – lead solders
2) Special soft solders
3) Soft solders for aluminum-based materials

Lead – Tin and Tin – Lead Solders. These solders constitute by far the majority of the soft solders used in industry. The contain 2 – 90 % tin, solders containing 40 – 60 % tin being the commonest. See also → Lead Alloys.

Solders with a high lead content require relatively high soldering temperatures, but permit high operating temperatures, sometimes above 200 °C, depending on strength requirements (e.g., in automobile radiators or heaters). Although these solders are low in price, their soldering properties (i.e., wetting and flow properties) are inferior to those of the high-tin solders. The high-lead solders are mainly applied by flame soldering owing to their relatively high soldering temperature (up to 300 °C and above).

The liquidus temperature of tin – lead solders decreases to ca. 180 °C with increasing tin content, improving their soldering properties. For general-purpose applications, soft solders containing 35 – 50 % tin are used. If a further reduction in soldering temperature is required, the eutectic tin–lead solder (ca. 60 % tin) is used. Its main application is for soldering electronic circuits in a bath [11]. In some solders, part of the tin is replaced by antimony. This improves the mechanical strength of the solder, but the soldering properties are not quite as good.

Special Soft Solders. Special soft solders are used for special applications in which the properties of the lead – tin solders are inadequate. These can be based on zinc – cadmium alloys, and sometimes contain silver; they can be used at high operating temperature (>200 °C).

Lead-free tin – silver or tin – copper solders are used for soldering applications in the food industry, in the installation of copper pipework, and in the refrigeration industry.

Soft Solders for Aluminum-Based Materials. Zinc-containing solders are mainly used for soft soldering aluminum, and contain cadmium, tin, or aluminum as an additional alloying element. As aluminum is always difficult to wet owing to the formation of a very stable oxide layer, these solders are applied with special fluxes or special soldering techniques (e.g., ultrasonic soldering) [12].

4.2. Brazing Filler Metals

Heavy-Metal Brazing Filler Metals. Silver brazes have the widest industrial application of all the heavy-metal brazing alloys. See also → Silver, Silver Compounds, and Silver Alloys, Chap. 10.7.. Their operating temperatures lie between 610 and 870 °C, depending on the silver content and the other alloying elements. Alloys with higher silver content permit lower brazing temperature, which offer both technical and economic advantages. They also contain copper and zinc, and sometimes tin or cadmium. When working with cadmium- containing brazing filler metals, special precautions must be taken for safety reasons (see Chap. 8).

Table 1. Properties and uses of selected soldering and brazing alloys

Designation	Melting temperature range, °C	Typical solders (DIN designation)	Solidus temperature, °C	Liquidus temperature, °C	Main areas of application
Lead – tin soft solders (DIN 1707)	185 – 325	L-Sn60Pb	183	190	electrical industry
		L-Sn50Pb	183	215	
Special soft solders (DIN 1707)	145 – 395	L-SnPbCd18	145	145	metallized ceramics
		L-SnAg5	221	240	refrigeration engineering, copper pipes for cold and hot water
		L-SnSb5	230	240	refrigeration engineering, copper pipes for cold and hot water
		L-CdZnAg2	270	280	electrical industry, magnetic materials
		L-CdAg5	340	395	electrical industry, magnetic materials
Aluminum soft solders (DIN 8512)	200 – 350	L-CdZn20	270	280	radiator construction, electric cables
General-purpose, cadmium-free brazing alloys (DIN 8513)	730 – 1100	L-Ag44	680	740	water fitting, electrical engineering
		L-Ag12	800	830	transmission manufacture
		L-CuZn40	890	900	galvanized steel tubes, steel furniture
		L-CuNi10Zn42	890	920	brazing mass-produced components under protective gas atmosphere
		L-CuSn	825	990	brazing mass-produced components under protective gas atmosphere
		L-Cu	1083	1083	brazing mass-produced components under protective gas atmosphere
General-purpose, low-melting brazing alloys containing cadmium (DIN 8513)	610 – 800	L-Ag40Cd	595	630	gas fitting
		L-Ag30Cd	600	690	electrical engineering, refrigeration engineering, suppliers to automobile industry
Special brazing alloys containing manganese (DIN 8513)	690 – 1020	L-Ag49	625	705	carbide tools
		L-Ag27	680	830	carbide tools
		L-CuMn10Co *	980	1030	carbide tools
Brazing alloys for special applications (DIN 8513)	730 – 960	L-Ag56In14Ni	620	730	stainless-steel water pipes
		L-Ag67	700	730	silver cutlery, jewelry
		L-Ag72	780	780	vacuum technology
		L-Ag70ZnCd *	710	730	ammonia pipes
		L-Ag85	960	970	ammonia pipes

Table 1. (Continued)

Designation	Melting temperature range, °C	Typical solders (DIN designation)	Solidus temperature, °C	Liquidus temperature, °C	Main areas of application
Phosphoric brazing alloys for copper materials (DIN 8513)	710–800	L-Ag15P	650	800	electrical industry
		L-Ag2P	650	810	copper pipes for cold and hot water
		L-CuP6	650	810	copper pipes for cold and hot water
Aluminum brazing alloys (DIN 8512)	560–600	L-AlSi12	575	590	heat exchangers
High-temperature, nickel-base brazing alloys	900–1200	L-Ni7	890	890	nuclear and reactor construction, turbine construction
		L-Ni5	1080	1135	nuclear and reactor construction, turbine construction
		L-Ni2	970	1000	turbine construction
High-temperature, precious-metal-base brazing alloys	810–1700	SCP1	807	810	radio and electron tubes
		SCP4	901	950	radio and electron tubes
		VH950	950	950	
		PN1	1237	1237	turbine and reactor construction

∗ Not standardized.

Phosphorus-containing brazing filler metals can be used to join copper to copper in air without a flux. This is used in the construction of electrical machinery, in copper pipe installation, and in other areas.

High-melting, silver-free brazing alloys are used in applications where high brazing temperature can be tolerated. Copper and copper–tin alloys are mainly applied in a furnace under a protective gas or in vacuum, but brass and "German silver" brazing alloys (CuZn and CuZnNi, respectively) are mainly applied with flux in air for bonding steel components.

Light-Metal Brazing Alloys. The most common alloy for brazing aluminum is the eutectic aluminum brazing alloy containing 12 % silicon (operating temperature: 590 °C). A rod of alloy can be used with a suitable flux for manual brazing. The industrial brazing of aluminum (e.g., for automobile radiators) is often carried out with sheet aluminum precoated with brazing alloy. The assembled parts are then brazed, with or without a flux, in special furnaces with very precise temperature control [13].

High-temperature brazing filler metals are used when the brazed joint must have a high strength at high operating temperature (ca. $\geq$ 400 °C). This is usually associated with a need for oxidation-resistance. If the brazed joints have to be highly corrosion resistant, high-temperature brazing alloys are often used.

Nickel-base brazing alloys are preferred (→ Nickel Alloys, Chap. 8., → Nickel Alloys, Chap. 9.). They contain chromium with varying amounts of boron, silicon, or phosphorus to reduce the melting range. The nickel-base brazing alloys are brittle, and hence cannot be produced in the form of ribbon or wire; they are usually supplied as powder or paste. However, some nickel-base filler metals can be produced in the form of amorphous ribbons by rapid solidification by the "melt-spin" process. During the brazing process, suitable heating can cause diffusion of boron and silicon into the substrate, resulting in ductile brazed joints of high strength [14], [15].

In addition to the nickel-base filler metals, gold- and palladium-containing brazing alloys (rarely platinum) are available in the form of wire, ribbon, or preforms. Due to their

high noble-metal content, brazing alloys of this type are relatively expensive. An example of an industrially important gold brazing alloy is Au82Ni18, which is used in the construction of engines and in aerospace applications [16]. The palladium- containing silver–copper alloys may also belong to the group of high-temperature brazing alloys, depending on their working temperature. Their most important use is the brazing of metallic substrates to metallized ceramics in power-electronic components.

Unmetallized ceramics, diamond, and glass can be brazed with special "active brazing" alloys [17]. These filler metals contain elements that promote wetting, e.g., titanium or zirconium, and can directly wet nonmetallic inorganic materials.

As well as steels and nickel-base materials, titanium alloys are now being increasingly used as materials of construction for the manufacture of highly stressed components. These materials can be brazed with grades of silver-base alloys designed for vacuum application, but also with titanium-base brazing alloys that contain elements such as copper and nickel to reduce the melting range. If brazing is performed correctly, butt joints can be made with strengths comparable to those of the base material [18].

High-temperature brazing alloys and some of the standard type are frequently used in a vacuum furnace. Vaporization of alloying elements and impurities during brazing is prevented by using brazing alloys specially designed for vacuum use. These alloys do not contain components with a high vapor pressure, and their purity specification is tighter.

5. Fluxes and Protective Atmospheres

Fluxes are organic or inorganic substances that remove oxides and other harmful layers from the surface to be joined by soldering or brazing. They thereby provide the conditions that allow wetting by the molten filler metal.

Soft-Solder Fluxes. The fluxes for soft soldering are classified according to their effectiveness in removing oxide layers from metallic surfaces. The oxides on the substrate are of varying stability; for example, the soldering of copper requires a less "aggressive" flux, while a more active flux is required for chrome steels. However, increasing activity of the flux is accompanied by increasing corrosiveness, so that the flux must be completely removed after soldering by washing with water or solvents, especially if it is of the more active type.

Typical active components of soft-solder fluxes include hydrochloric acid, zinc chloride, ammonium chloride, citric acid, organo halogen compounds, and natural and modified resins.

Brazing Fluxes. The fluxes used for brazing aluminum are classified according to the corrosive action of their residues. Hygroscopic fluxes must be scrupulously removed by washing with water after brazing due to their corrosive properties. In contrast, water-insoluble residues can be left on the brazed area provided that they do not interfere with subsequent processing, e.g., painting or plating. The active temperature range (i.e., the range within which the flux can remove the oxide layer from the surface to be brazed) is relatively narrow for both groups of fluxes used for brazing aluminum (ca. 500 – 600 °C).

In the case of heavy-metal brazing fluxes, it is necessary to work within the active temperature range of the particular flux being used. The choice of flux depends mainly on the brazing temperature, which in turn depends on the brazing alloy. As the residues from nearly all these fluxes are corrosive, they must be removed after brazing.

Copper, copper – tin, or silver can be brazed in air without fluxes if phosphorus- containing copper-based brazing alloys are used. The phosphorus has a similar effect to that of a flux.

Brazing Atmospheres. With the above-mentioned exception of phosphorus- containing, copper-based brazing alloys, flux-free brazing requires a special atmosphere such as a protective gas or vacuum (see Section 3.8). Above a certain temperature, the atmosphere decomposes the oxide layer on the surface of the workpiece and prevents it from reforming. The required temperature depends on the quality of the brazing atmosphere, and is determined by the metal – metal oxide equilibrium curves (Fig. 3). Different compositions and purities of brazing atmospheres are thus required for each brazing

alloy and base material. The brazing of low-alloy steels on a production basis can be carried out very economically in a continuous furnace, e.g., in protective gas atmospheres of incompletely combusted natural gas or propane. However, high-alloy special steels or oxygen-sensitive materials (titanium or zirconium) require very pure atmospheres with extremely low levels of oxygen and water vapor; good brazing results can normally only be obtained in high vacuum (10 to 1 mPa, i.e., 10^{-4} to 10^{-5} mbar) [19].

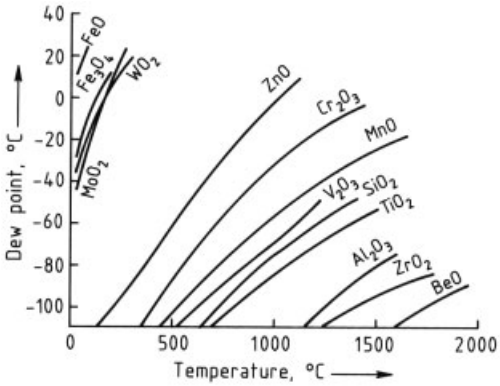

Figure 3. Dew points of metal – metal oxide equilibria as a function of temperature

6. Materials of Construction Suitable for Soldering and Brazing

An important advantage of brazing as a joining technique is that it permits the bonding different types of material. For example, ceramic and steel can be very strongly bonded by using special brazing alloys and brazing processes.

Materials of construction can be divided into three main groups according to their brazability [20]:

1) Materials that can be brazed using general-purpose brazing alloys, general-purpose fluxes, and standard processes (e.g., copper and nickel alloys, iron, steels, cobalt, and noble metals)
2) Materials requiring special brazing alloys and/or special fluxes, but not special processes (aluminum, hard metals, stellite, chromium, molybdenum, tungsten, tantalum, and niobium)
3) Materials that can only be brazed with special brazing alloys and special processes (titanium, zirconium, beryllium)

Nonmetallic inorganic materials (e.g., ceramics, diamond, and graphite can be brazed with special brazing alloys (see Section 4.2) [21], [22].

The above examples also apply to soft soldering to a limited extent. Some materials of construction must be regarded as impossible to soft soldering on an industrial scale (e.g., beryllium and cemented carbides).

7. Design and Strength Properties of Soldered and Brazed Joints

Dimensioning. In capillary soldering or brazing, the molten filler metal flows into the gap by capillary attraction and solidifies, thus bonding the surfaces to be joined. Normally, the gap width should be between 0.05 mm and 0.2 mm (see Chap. 2). If materials with different thermal expansion coefficient are bonded together, care must be taken that the desired gap width exists at the soldering or brazing temperature – this may require a narrower or wider gap at room temperature.

In addition to gap size, the dimensions of the area to be joined must be established for highly stressed soldered or brazed joints. It must be ensured that brazed joints can be subjected to as much stress as the substrate material. With overlapping joints, this is the case when the length of the overlap corresponds to 3 – 6 times the wall thickness of the weaker material to be joined.

The surface area of the joint can be dimensioned according to the theory of strengths of materials. For quasistatic loads at room temperature, the tensile strength R_m should be ca. 200 N/mm^2 and the shear stress at failure τ_B ca. 100 N/mm^2. The strength of brazed joints can be considerably higher in practice, but they should be determined, if possible, by standard tests (DIN 8525).

The amount of solder or brazing alloy required to make a joint can be determined from the maximum gap volume, with an addition of ca. 30 % to allow for the formation of fillets.

Joint Design. In order to optimize soldered or brazed joints, the following rules for design should be observed:

Table 2. Tensile strength (MPa) of joints made with brazing alloys and steels measured according to DIN 8525, Part 1

Filler metal	Base material (steel)			
(DIN 8513)	St 37	St 50	St 60	Stainless
L-Ag44	390 ∗	480	520	530
L-Ag40Cd	410 ∗	530	640	520
L-CuZn40	360		420	

∗ Fracture in base material.

1) The gaps to be soldered or brazed should have parallel sides of defined width.
2) The filler metal should be located so that, on melting, it comes into immediate contact with the surfaces to be joined. In some cases, the filler metal can be placed in the gap in the form of ribbon or specially shaped pieces (preforms) prior to joining.
3) If possible, the workpieces should be arranged so that they locate automatically. A three-line press fit or a knurled press fit can be used for joining shafts or wheels.
4) It must be ensured that the flux can be displaced by the molten filler metal so that flux inclusions are avoided (Fig. 4).
5) Sudden changes in cross section should be avoided in highly stressed brazed joints.
6) Any machining marks on the surfaces to be joined should be orientated in the direction of the flow of the filler metal.

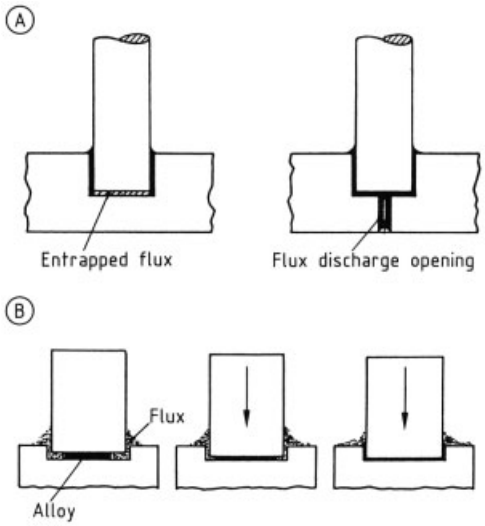

Figure 4. Prevention of flux inclusions
A) Escape of flux through a flux discharge opening; B) Brazing with preplaced shim of alloy

Joint Strength. Even at room temperature, joints made of normal soft solders usually have low strength ($<10\,N/mm^2$) due to creep [23]. Owing to their poor strength, soft-soldered joints are unsuitable for stressed structures under static load unless additional mechanical support is provided, e.g., by means of a form fit. Soft-soldered joints are thus generally used on account of their sealing or electrically conductive properties.

The tensile strength of brazed joints made with copper and copper alloys often exceeds the strength of the base material. Table 2 lists the strength of joints made in steel. The high strength values enable brazed joints to be employed for stressed structures. A decrease in strength is observed above 200 °C.

8. Safety Aspects

Safety aspects of soldering and brazing are covered by national regulations, only a few examples can be mentioned here [24], [25].

Effective ventilation of the workplace is very important to remove hazardous vapors or dusts derived from the fluxes and the solders. When using cadmium- containing solders, special safety precautions must be observed because the cadmium oxide fume produced is classified as a carcinogen. If possible, cadmium-free solders should be used.

Before using fluxes and filler metals, the safety data sheets provided by the manufacturer or supplier should be consulted for details concerning safe handling. Appropriate protective clothing and equipment should be used. For brazing, safety spectacles should be worn for protection against spurting, dazzle, and heat radiation. Further details are given in national regulations. When working with a flame, suitable precautions against injury from burns should be taken.

The safe handling of flammable gases for flame brazing and furnace brazing requires specialist knowledge, and should therefore only be carried out by trained personnel.

9. References

General References

1. *Brazing Handbook,* 4th ed., American Welding Society, Miami 1991.
2. *Soldering Manual,* 2nd ed., American Welding Society, Miami 1977.
3. H. R. Brooker, E. V. Beatson: *Industrial Brazing,* revised by P. M. Roberts, Butterworth & Co, London 1975.
4. K.-F. Zimmermann: *Hartlöten – Regeln für Konstruktion und Fertigung,* Deutscher Verlag für Schweiß-technik, Düsseldorf 1968.
5. M. M. Schwartz (ed.): *Source Book on Brazing and Brazing Technology,* American Society for Metals, Ohio 1980.
6. W. Müller: *Metallische Lotwerkstoffe – Arten, Eigenschaften, Verwendung,* Deutscher Verlag für Schweißtechnik, Düsseldorf 1990.

Specific References

7. *Brazing Manual,* American Welding Society, Miami 1975.
8. W. Malikowski et al, EP 87 113 767.5 – 2106, 1987.
9. H. Krappitz, U. Mürrle, J. Nauber: "Production and Characteristics of Wear Resistant Coatings, Using Plastic Bonded Carbide and Brazing Preforms" *DVS-Berichte,* DVS-Verlag, Düsseldorf 1988, vol. 112, pp. 45 – 49.
10. C. J. Thwaites: *Capillary Joining – Brazing and Soft-Soldering,* Research Studies Press, Chichester 1982.
11. R. J. Klein Wassink: *Soldering in Electronics, Electrochemical Publications Ltd.,* Ayr 1984.
12. *Aluminum Soldering Handbook,* Aluminum Association Inc., New York 1976.
13. H. Klock, H. Schoer: *Schweißen und Löten von Aluminiumwerkstoffen,* Deutscher Verlag für Schweiß-technik, Düsseldorf 1977.
14. E. Lugscheider: "Hochtemperaturlöten, Stand und Entwicklungstendenzen, – Lote", *Schweißen und Schneiden* **32** (1980) H. 5, 171.
15. E. Lugscheider, H. Krappitz: "The Influence of Brazing Conditions on the Impact Strength of High-Temperature Brazed Joints", *Welding J. (Miami),* **65** (1986) *Weld. Res. (Miami)* 261 – 267.
16. J. Colbus, K.-F. Zimmermann: "Properties of Gold-Nickel Alloy Brazed Joints in High Temperature Materials," – *Gold Bull.* **7** (1974) no. 2, 42 – 49.
17. E. Lugscheider, H. Krappitz, M. Boretius: "Fügen von Hochleistungskeramik untereinander und mit Metall" *Tech. Mitt.* **80** (1987) H. 4, 231 – 238.
18. E. Lugscheider, L. Martinez: "Brazing of Ti Al 6 V 4 with Al-, Au- and Ti-Base Filler Metals," *18th International AWS Brazing Conference,* Chicago 1987.
19. I. V. Etchells, P. M. Roberts: *Brazing Furnaces, Welding & Metal Fabrication* 1991, pp. 300 – 312.
20. J. Kinzel: "Lassen sich alle Werkstoffe löten?," *Praktiker* **39** (1987) 477 – 481.
21. M. M. Schwartz: *Ceramic Joining,* American Society for Metals, Ohio 1990.
22. "Herstellen von Keramik/Keramik- und Keramik/Metall-Verbindungen durch Aktivlöten" *DVS-Merkblatt 3102* Deutscher Verlag für Schweißtechnik, Düsseldorf 1993.
23. G. Zwingmann, K.-F. Zimmermann: "Zeitstandscherfestigkeit von Weichlötverbindungen bei erhöhter Temperatur an Messing Ms 58 unter Verwendung von Sonderweichloten," *Metall* **31** (1977) no 7, 746 –750.
24. ANSI/ASC Z 49.1 – 1983: Safety in Welding and Cutting, American Welding Society, Miami 1983.
25. "Schweißen, Schneiden und verwandte Verfahren" in *Unfallverhütungsvorschrift Abschnitt 15,* Jedermann-Verlag, Heidelberg 1990.

Solvents

DIETER STOYE, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany

1. Definitions 2
2. Physicochemical Principles 3
2.1. Theory of Solutions 3
2.2. Dipole Moment, Polarity, and Polarizability 7
2.3. Hydrogen Bond Parameters 8
2.4. Solvation 9
2.5. Solvents, Latent Solvents, and Non-Solvents 9
2.6. Dilution Ratio and Dilutability . . . 9
2.7. Influence of Molecular Mass on Solubility 11
2.8. Dissolution and Solution Properties 11
3. Physical and Chemical Properties . 13
3.1. Evaporation and Vaporization . . . 13
3.2. Hygroscopicity 15
3.3. Density and Refractive Index 15
3.4. Viscosity and Surface Tension . . . 16
3.5. Vapor Density 17
3.6. Thermal and Electrical Data 17
3.7. Flash Point, Ignition Temperature, and Ignition Limits 17
3.8. Heats of Combustion and Calorific Values 19
3.9. Chemical Properties 19
4. Toxicology and Occupational Health 20
4.1. Toxicology 20
4.2. Occupational Health 21
5. Environmental and Legal Aspects . 23
5.1. Environmental Protection 23
5.2. Laws Concerning Dangerous Substances 25
5.3. Fire Hazard 26
5.4. Waste 26
6. Purification and Analysis 28
7. Uses 28
7.1. Solvents in Paints 28
7.2. Solvents in Paint Removers 31
7.3. Solvents in Printing Inks 31
7.4. Extraction 31
7.5. Extractive Distillation 31
7.6. Chromatography 32
7.7. Solvents for Chemical Reactions . . 32
7.8. Solvents for Recrystallization 32
7.9. Solvents in Film Production 32
7.10. Solvents for Synthetic Fibers 33
7.11. Solvents for Rubber, Plastics, and Resin Solutions 33
7.12. Solvents for Degreasing 33
7.13. Solvents for Dry Cleaning 34
7.14. Solvents in Aerosol Cans and Dispensers 34
8. Economic Aspects 34
9. Solvent Groups 34
9.1. Aliphatic Hydrocarbons 37
9.2. Cycloaliphatic Hydrocarbons 39
9.3. Terpene Hydrocarbons and Terpenoids 41
9.4. Aromatic Hydrocarbons 50
9.5. Chlorinated Hydrocarbons 51
9.6. Alcohols 52
9.7. Ketones 56
9.8. Esters 59
9.9. Ethers 62
9.10. Glycol Ethers 63
9.11. Miscellaneous Solvents 66
10. References 68

1. Definitions

Solvents are compounds that are generally liquid at room temperature and atmospheric pressure; they are able to dissolve other substances without chemically changing them. The liquid mixture formed on dissolving a substance (*solute*) in a solvent is termed a *solution*. The molecules of the solution components interact with one another. Solutions are obtained by mixing liquid, solid, or gaseous components with liquids, the liquid always being termed the solvent. When two liquid components are combined, it is arbitrary which of the two components is considered to be the solvent, and which the solute; the liquid component present in excess is usually termed the solvent. Accordingly, plasticizers that are used for flexibilization in plastics processing and paint production may also be regarded as solvents. Plasticizers differ from solvents, however, with regard to their technological significance. A good plasticizer should have

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_437

a very low volatility and thus permanently affect the dissolved substance. An ideal solvent should, in contrast, have a high volatility, so that it can evaporate as rapidly as possible to leave the dissolved substance (e.g., in a paint film). The boundary between plasticizers and solvents is not clear cut – some high-boiling solvents of very low volatility exert a flexibilizing effect over a prolonged period.

A solvent should generally have the following properties [2];

1) Clear and colorless
2) Volatile without leaving a residue
3) Good long-term resistance to chemicals
4) Neutral reaction
5) Slight or pleasant smell
6) Anhydrous
7) Constant physical properties according to the manufacturers' specification
8) Low toxicity
9) Biologically degradable
10) As inexpensive as possible

Inorganic substances (e.g., hydrogen sulfide, ammonia, sulfur dioxide, hydrogen fluoride, and hydrogen cyanide) that are used as solvents in special applications, mainly at low temperature or under pressure, will not be discussed here.

2. Physicochemical Principles

2.1. Theory of Solutions

During dissolution the solvent acts on the substance to be dissolved to increase its state of distribution. Dissolution results in the formation of real solutions, colloidal solutions, or dispersions depending on the size of the particles that interact with the solvent molecules.

In *real solutions* the diameter of the dissolved particles is ca. 0.1 nm, and is thus of the order of magnitude of the free molecules. Real solutions are formed by most inorganic and organic compounds of low molecular mass. They are clear, physically homogeneous liquids.

In *colloidal solutions* the diameter of the dissolved particles is ca. 10 – 100 nm. Colloidal solutions are generally clear to weakly opalescent liquids, but exhibit inhomogeneities as regards some physical properties (e.g., the Tyndall effect).

In *dispersions* the diameter of the particles is larger than in colloidal solutions. Dispersions are turbid to milky liquids consisting of at least two phases.

Intermolecular Forces. During the dissolution of a substance (A) in a solvent (B) the forces of attraction between the molecules of the pure components ($K_{\mathrm{A-A}}$ and $K_{\mathrm{B-B}}$) are destroyed, and new forces are simultaneously formed between the solvent and substance molecules:

$$K_{\mathrm{A-A}} + K_{\mathrm{B-B}} \longrightarrow 2\,K_{\mathrm{A-B}}$$

A substance is generally readily soluble in a solvent if the forces of attraction in the pure substance are of the same order of magnitude as the forces of attraction in the pure solvent. A substance is generally insoluble in a solvent if the forces of attraction between its molecules are significantly higher or lower than in the pure solvent. In this case more energy is required to overcome the forces of attraction in the pure components than is released on formation of the solution. This is the explanation of the rule of thumb "Like dissolves like" (*similia similibus solvuntur*). The intermolecular forces of attraction differ – they are strongest in crystalline solids, weaker in amorphous solids and liquids, and weakest in gases. Intermolecular forces are classified according to their physical nature (Table 1) [3–6].

Table 1. Intermolecular forces

Type of force	Interaction between	Temperature dependence
Ionic	ions	weak
Ion–dipole	ions and dipoles	weak
Directional	permanent dipoles	strong
Induction	permanent and induced dipoles	weak
Dispersion	atomic dipoles	weak
Hydrogen bonds	groups of molecules	strong

Ionic (Coulomb) Forces. Forces of attraction between ions of opposite charge are termed ionic or Coulomb forces. The force with which two ions 1 and 2 attract one another depends on their electrical charges e_1 and e_2 and the distance r between them:

$$K_{\mathrm{II}} \approx -\frac{e_1 \cdot e_2}{r^2}$$

Coulomb forces are responsible for the stability of ionic crystals (e.g., NaCl). When such a compound is dissolved in a polar solvent (dipole moment μ), dissociation and simultaneous solvation of the ions occur. The force of attraction between the ions is now inversely proportional to the dielectric constant of the solvent, and is thus reduced. New ion – dipole forces are formed as a result of the attraction of the permanent dipoles of the solvent by the ions:

$$K_{\mathrm{ID}} \approx -\frac{e_{\mathrm{I}} \cdot \mu}{r^3}$$

The distance between the solvated ions in the solution generally changes only slightly with temperature and depends on the thermal expansion coefficient of the solution. The forces between the ions are therefore only slightly temperature dependent.

Dipole – Dipole Forces. Dipole – dipole (directional) forces are forces of attraction between molecules with a finite, permanent overall dipole moment. The forces of attraction resulting from the dissolution of a polar molecule (μ_1) in a polar solvent (μ_2) are given by [7]:

$$K_{\mathrm{DD}} \approx -\frac{\mu_1 \cdot \mu_2}{r^4}$$

The distance between the dipoles depends largely on the position of the poles in the molecule (i.e., on steric molecular influences) and on thermal vibrational movements. The force of attraction between the dipoles accordingly decreases sharply with increasing temperature.

Induction forces are produced as a result of interactions between permanent dipoles and induced dipoles. The electric field of a molecular dipole leads to charge displacement in the neighboring molecule and thus to the induction of a dipole.

The magnitude of the induced dipole moment μ_{ind} depends on the magnitude of the permanent dipole moment μ and on the polarizability α of the second molecule [8]. Induction forces are only slightly temperature dependent:

$$\mathrm{K}_{\mathrm{DDi}} \approx -\frac{\alpha \cdot \mu^2}{r^7}$$

Dispersion (London – Van der Waals) Forces. Dispersion forces are formed by mutual induction of atomic dipoles due to the electromagnetic field between the nucleus and electrons of the atom.

Dispersion forces therefore depend on the displaceability of the electrons in the atoms, the polarizability α, and the availability of the electrons, i.e., on the ionization potential *IP* [9]:

$$K_{\mathrm{Dis}} \approx -\frac{\alpha^2 \cdot IP}{r^7}$$

Since charge displacement is independent of the thermal motion of the particles, dependence of the dispersion forces on temperature is weak over the distance between the attracting partners. Dispersion forces act in all atoms and molecules.

Hydrogen Bonds [10–14]. Hydrogen bonding forces exist in substances that have hydroxyl or amino groups (e.g., water, alcohols, acids, glycols, and amines). These molecules act as hydrogen donors and thus form a bond with hydrogen acceptors (e.g., esters and ketones). Water, alcohols, and amines act both as hydrogen donors and acceptors. Very weak hydrogen bonds also exist in halogens and sulfur. Hydrogen bonds are highly dependent on the mutual orientation of the molecules and thus on the temperature.

Thermodynamic Principles [15]. Forces of attraction act between the molecules of the pure components and between the different molecules in the solution. If the forces of attraction in the solution are greater than those in the pure components, dissolution is accompanied by a decrease in the internal energy of the system. The process is exothermic and heat is released. If, however, the forces of attraction between the molecules of the pure components are greater than those in the solution, the internal energy of the system is increased with absorption of heat. In a closed system, this endothermic dissolution process is accompanied by cooling. In open systems heat is absorbed from the surroundings.

Most dissolution processes are endothermic and are thus promoted by a temperature increase: the solubility has a positive temperature coefficient. Exothermic dissolution processes have a

negative temperature coefficient (i.e., the solubility decreases with rising temperature). Miscibility gaps frequently occur above certain temperatures (Fig. 1).

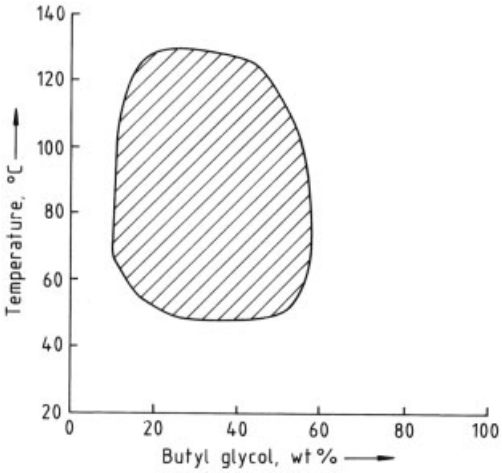

Figure 1. Miscibility gap for mixtures of butyl glycol and water at atmospheric pressure

Why do endothermic dissolution processes occur spontaneously? The Gibbs – Helmholtz equation provides the answer:

$$\Delta G = \Delta H - T\Delta S$$

The reaction enthalpy ΔH is negative in exothermic processes, and positive in endothermic processes. The reaction entropy ΔS, which is a measure of the change in the state of disorder of the system, is always positive since the entropy of the solution (mixture) is always greater than that of the pure components.

The Gibbs free energy ΔG must be a negative quantity in all spontaneously occurring processes. In exothermic processes ΔH is negative: dissolution occurs spontaneously with release of heat, and is "enthalpy determined." In endothermic processes ΔH is positive; these processes also proceed spontaneously, but the temperature of the system falls. On account of the large positive reaction entropy the negative contribution $-T\,\Delta S$ exceeds the positive reaction enthalpy ΔH, and ΔG is again negative. The process is "entropy determined." If, however, ΔH is positive and greater than $T\,\Delta S$, no dissolution takes place: the system remains heterogeneous.

Dilution processes are used very often, for example in the paint industry, and are always entropy-determined [16].

Systems with a negative temperature coefficient of solubility are mainly enthalpy-determined. The interaction between the partners is often due to weak hydrogen bonds, which become weaker than the forces between the pure components on increasing the temperature. The result is a decrease in solubility. Examples of enthalpy-determined systems with a negative temperature coefficient are summarized in Table 2.

Table 2. Aqueous solution systems with a negative temperature coefficient of solubility

Solvent	Solubility in water	g/100 g
Methyl butyl ether	1.5 (10 °C)	0.9 (25 °C)
2-Methylfuran	18.2 (10 °C)	13.9 (25 °C)
Isoamyl alcohol	2.6 (22 °C)	2.2 (36 °C)
Ethyl acetate	10.0 (0 °C)	7.7 (25 °C)
Butyl glycol	∞ (40 °C)	11.0 (60 °C)

The change in the total energy of the system can thus be negative or positive. This change in energy is termed heat of solution or mixing, or solution or mixing enthalpy [17], [18].

Systems that are miscible with one another without any change in temperature are termed athermic systems [19], [20]; they include the following solvent pairs:

Benzene – trichloromethane
Bromobenzene – chlorobenzene
Methanol – ethanol
Toluene – tetrachloromethane
Methyl acetate – ethyl acetate
Aniline – water

Negative heats of mixing (i.e., warming of the system) are found in the following solvent pairs:

Water – acetone
Water – methanol
Water – propanol
Trichloromethane – acetone

Positive heats of mixing (i.e., cooling of the system during mixing) occur with the following solvent pairs:

Trichloromethane – carbon disulfide
Acetone – carbon disulfide
Gasoline – carbon disulfide

Cohesive Energy Density and Solubility Parameters. As a result of attractive or cohesive forces the molecules in pure solvents have a cohesive energy that has to be expended in molecular separation processes (e.g., dilution, evaporation, or addition of another substance). The cohesive energy can be calculated from the enthalpy of vaporization ΔH_V and the work that is required to expand the vapor against the atmosphere (volume work) [21]. The cohesive energy per unit volume, i.e., the *cohesive energy density*, is defined as [17], [22], [23]:

$$e = \frac{(\Delta H_V - RT)}{V}$$

The cohesive energy density e is a specific temperature-dependent parameter for each solvent and represents a direct thermodynamic measure of the forces of attraction in solvents. It can be used to evaluate the solvency of the solvents if the cohesive energy densities of the substances to be dissolved are also known.

The cohesive energy density of a solvent A is altered when it is mixed with another solvent B. Two interaction pairs A – B are formed from each interaction pair A – A and B – B. The new cohesive energy density of the pair A – B in solvents of very low polarity being approximately equal to the geometric mean of the cohesive energy densities of the pure components [17], [22], [24–26]:

$$e_{AB} = (e_{AA} \cdot e_{BB})^{1/2}$$

The square root of the cohesive energy density is termed the *solubility parameter* δ. It is also a measure of the intermolecular forces in pure substances. Solvents with comparable solubility parameters have similar interaction forces and are therefore readily miscible and mutually soluble. In order to take account of not only the dispersion forces but also of the polar forces and hydrogen bonds, the solubility parameter is resolved into the following components [27], [28]:

$$\delta^2 = \delta_D^2 + \delta_P^2 + \delta_H^2$$

If the three solubility parameter components for dispersion forces δ_D, dipole forces δ_P, and hydrogen bonds δ_H are plotted on a three-dimensional space diagram (Fig. 2), a system is obtained in which a vector $\rightarrow \vec{\delta}$ is defined for each solvent. The vector describes the solvent's solubility and miscibility behavior [29], [30]. Solvents that lie close to one another in this space diagram (i.e., whose vector difference $\Delta\delta$ is small) have similar solution properties and often a similar chemical structure. Solvents that are far apart on the diagram differ greatly in their chemical and physical characteristics; they are generally immiscible [31–33]. The solubility parameters as well as their components are shown for some solvents in Table 15.

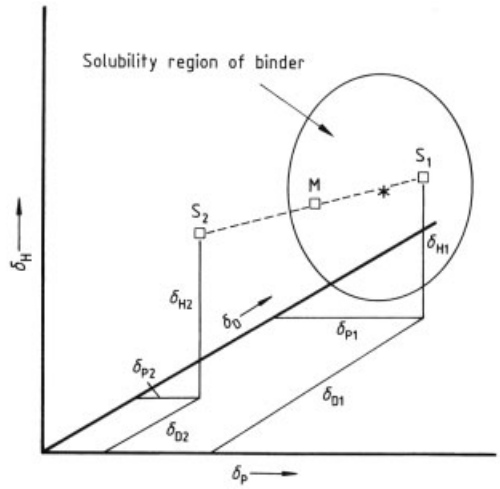

Figure 2. Solubility parameter diagram
S_1 = solvent; S_2 = non-solvent; M = solvent mixture

The predictions that can be made for solvent mixtures on the basis of the solubility diagram are not strictly valid because they involve thermodynamic simplifications and empirical parameters, and disregard temperature effects [30], [34]. A strong warning must therefore be given against the uncritical use of solubility parameters. Nevertheless, description of the solvents with the aid of the solubility parameter concept often provides useful information about their solvency and reveals similarities which can otherwise only be characterized empirically (e.g., latent solvents or dilutability).

Solvent Miscibility. As a rule solvents are readily miscible if the difference in their solubility parameters $\Delta\delta$ is $\leq 4-6$ units. Solubility is generally limited if the differences are larger (Table 3).

In the case of strongly polar (ionic) forces or hydrogen bonds, the solubility parameter con-

Table 3. Miscibility of solvents as a function of the difference in solubility parameters $\Delta\delta$

Solvent A	Solvent B	$\Delta\delta$, $(J/cm^3)^{1/2}$	Miscibility
Butyl acetate	methyl isobutyl ketone	0.2	good
Ethyl glycol acetate	xylene	0.2	good
Tetrachloroethylene	xylene	1.0	good
Ethanol	butanol	2.7	good
Methanol	dimethyl sulfoxide	3.3	good
Ethylene carbonate	dimethyl sulfoxide	3.7	good
Methyl glycol	hexane	7.2	poor
Cyclohexane	dimethyl sulfoxide	9.6	poor
Hexane	methanol	14.8	moderate

cept is of limited applicability. For example, the following pairs of solvents are only partially miscible despite the small difference in their solubility parameters:

Nitromethane – ethyl glycol	3.5
Ethanol – methyl iodide	2.0
Decane – nitrobenzene	4.7
Ethanol – naphthalene	5.9

On the other hand, many solvent pairs are readily miscible although the difference between their solubility parameters is large:

Dimethyl sulfoxide – benzene	7.6
Nitromethane – tetrachloromethane	8.2
Dimethylformamide – hexane	9.9
Benzene – methanol	10.9
Ethanol – hexane	11.1
Methanol – tetrachloromethane	12.1
Ethylene carbonate – butyl acetate	12.7
Water – acetone	27.4
Water – tetrahydrofuran	28.7

In some cases the resolution of the solubility parameter into the three components δ_D, δ_P, and δ_H [32], [33], [35] is more suitable for explaining the experimental facts. For example, although the difference between the solubility parameters of the partially miscible solvent pair nitromethane – ethylene glycol is indeed small ($\Delta\delta = 3.5$), both solvents differ greatly as regards their solubility parameter δ_H ($\Delta\delta_H = 22.0$). Many empirical results (e.g., the good miscibility of benzene and methanol) cannot, however, be explained by this approach.

Solubility of Polymers [36–43]. Most polymers are readily soluble if their solubility parameters are comparable to those of the solvent in question. The upper limit for good solubility is a difference in the solubility parameters of 6 units (Table 4). The following selected systems are not, however, mutually soluble despite the small solubility parameter difference $\Delta\delta$:

Poly(vinyl chloride) – trichloromethane	1.6
Poly(vinyl chloride) – toluene	2.5
Polystyrene – heptane	2.9
Poly(vinyl acetate) – hexane	4.3
Poly(vinyl acetate) – diethyl ether	3.7
Cellulose – dimethyl sulfoxide	5.3

Furthermore, poly(vinyl acetate) is readily soluble in methanol and amylose is readily soluble in water, even though the differences in the solubility parameters are large ($\Delta\delta = 9.8$ and 22.5 units, respectively).

2.2. Dipole Moment, Polarity, and Polarizability

In order to describe their solubility properties, solvents are often subdivided according to their polarity, i.e., polar and nonpolar solvents. Since the term polarity cannot be defined unambiguously in physical terms, such a classification is not meaningful. The term polarity includes parameters such as dipole moment, hydrogen bonding, polarizability, entropy, and enthalpy. The dipole moment μ of a substance is a molecular property resulting from the vector sum of bond-dipole moments. Highly symmetrical molecules (e.g., tetrachloromethane and benzene) accordingly have no dipole moment; other aromatic hydrocarbons and dioxane exhibit only very small dipole moments. Less symmetrical molecules with strong bond dipoles have dipole moments between 1.6 and 1.9 D (alcohols, esters, glycol ethers), glycols and ketones have higher values (2.3 – 2.9 D). The solvents with

Table 4. Miscibility of polymer–solvent pairs as a function of the difference in solubility parameters $\Delta\delta$

Polymer	Solvent	$\Delta\delta$ $(J \cdot cm^3)^{1/2}$	Miscibility
Polyacrylonitrile	ethylene carbonate	0.6	good
Polystyrene	ethyl acetate	0.8	good
Poly(vinyl chloride)	tetrahydrofuran	1.5	good
Polyisobutene	hexane	1.5	good
Polyacrylonitrile	dimethylformamide	4.9	good
Poly(vinyl acetate)	ethanol	7.0	moderate
Polyisobutene	ethanol	9.8	poor
Polystyrene	methanol	11.0	poor

the largest dipole moments (3.7 – 5.0 D) are ethylene carbonate, nitropropane, dimethylformamide, and dimethyl sulfoxide (Table 15).

The dissolution behavior of a solvent cannot be predicted solely on the basis of its dipole moment. For example, dioxane ($\mu = 0.4$ D) is a very good solvent and has a comparable solvency to dimethyl sulfoxide ($\mu = 4.0$ D).

Dipole–dipole and induction forces in solvents or solutions decrease with increasing molecular mass of the solvent [44]. Since this effect is not reflected in the dipole moment of the solvent, a polarizability parameter P is used to describe the dipole – dipole interaction forces [35]. This parameter can be calculated from the ionization potential IP, polarizability α, and dipole moment μ [45]:

$$P = \frac{2\mu^4}{3k \cdot T \cdot \varepsilon}$$

where $\varepsilon \approx \alpha \cdot IP$.

The dipole moment μ is also used in conjunction with the solubility parameter δ as a coordinate in solubility diagrams to take account of the influence of variations in polarity [45–47].

The polarizability α of an electrically neutral compound is a measure of the displaceability of charge carriers within the molecule. The greater the polarizability, the stronger are the dipoles in duced by an external electromagnetic field; the magnitude of the polarizability and strength of the dispersion forces are thus related to one another.

2.3. Hydrogen Bond Parameters [48]

The strengths of hydrogen bonds in solvents have been divided into three classes [49–53]:

1) Solvents with weak hydrogen bonding (hydrocarbons, chlorinated hydrocarbons, nitro compounds, nitriles)
2) Solvents with moderately strong hydrogen bonding (ketones, esters, ethers, aniline)
3) Solvents with strong hydrogen bonding (alcohols, carboxylic acids, pyridine, water, glycols, amines)

This classification allows solvents to be described in terms of their solvency for polymers by means of solubility parameters, and also permits hydrogen bond forces to be taken into account. This is done by determining the solubility limits of the polymer in solvents belonging to each of the three classes. Three solubility parameter regions for a polymer are thus obtained. Attempts have also been made to use corrections to take account of the variations in solubility parameters caused by hydrogen bonding [54–57].

The hydrogen bond parameters γ provide a relatively accurate, simple characterization of solvents that can form hydrogen bonds [58], [59]. The parameter (Table 15) is determined from the shift of the oxygen – deuterium vibration frequency in the IR spectrum obtained when deuteromethanol is dissolved in the relevant solvent.

Solvents that undergo hydrogen bonding may act as proton donors or acceptors [10], [48]:

1) Proton donors (e.g., trichloromethane)
2) Proton acceptors (e.g., ketones, esters, ethers, aromatic hydrocarbons)
3) Combined proton donors and acceptors (e.g., alcohols, carboxylic acids, primary and secondary amines, water)
4) No hydrogen bonding (e.g., aliphatic hydrocarbons)

No hydrogen bonds exist in solvent mixtures comprising solely proton acceptors; hydrogen bonds are only formed in the presence of a proton donor and result in an increase in miscibility [48], [60].

2.4. Solvation

When a substance (solute) is dissolved in a solvent or solvent mixture, the forces of attraction between the solute molecules decrease because the solvent molecules penetrate between the solute molecules and finally surround them as a layer. This process, termed solvation, results in distribution of the solute in the solution at a molecular level. The strength of the solvation forces and the number of solvent molecules in the solvation layer depend on the solubility parameter, dipole moment, hydrogen bonding, polarizability, and the molecular size of the solvent and solute. The number of solvent molecules in a solvent – solute complex is determined by the degree of solvation β. Values for cellulose nitrate – ketone systems are given in Table 5 [61]. The degree of solvation β increases as the size of the solvent molecule decreases. It also increases with increasing solubility parameter δ. In the cellulose nitrate – ketone systems, hydrogen bond forces and dipole forces have little influence because both the binder and the solvent molecules are proton acceptors, and few hydrogen bonds are formed. Furthermore, the dipole moments of the solvents are roughly equal. In other cases, however, hydrogen bonding and dipole moments can greatly influence the degree of solvation.

Table 5. Connection between molecular mass of ketone solvents M, degree of solvation (for cellulose nitrate) β, solubility parameter δ, hydrogen bond parameter γ, and dipole moment μ

Solvent	M	β	δ	γ	μ
Methyl amyl ketone	114.18	1	8.5	7.7	2.7
Methyl propyl ketone	86.13	3	8.9	8.0	2.7
Methyl ethyl ketone	72.12	7	9.3	7.7	2.7
Acetone	58.08	12	10.0	9.7	2.9

2.5. Solvents, Latent Solvents, and Non-Solvents

True (active) solvents dissolve a given substance at room temperature; the solubility parameters of the solvent and the solute are similar.

Latent solvents cannot dissolve a substance by themselves. They can, however, be activated by adding a true solvent or a non-solvent. Latent solvents include solvents with extremely high or low solubility parameters or hydrogen bond parameters, such as alcohols, water, acetals, higher molecular mass esters, or ketones. In the solubility parameter space diagram, the latent solvents lie outside the solubility regions of the substances to be dissolved. On mixing with true solvents or non-solvents, mixed solvents are obtained whose solubility parameters fall in the region of the substances to be dissolved (Fig. 2).

Non-solvents are unable to dissolve the substance in question. Their solubility parameters and hydrogen bond parameters lie outside the solubility regions of the substances to be dissolved.

The expressions solvent, latent solvent, non-solvent only apply to a specific substance to be dissolved. A solvent can be a non-solvent for another substance and *vice versa*.

2.6. Dilution Ratio and Dilutability

If a non-solvent is added dropwise to a cellulose nitrate solution, the cellulose nitrate eventually separates as a precipitate or gel. The volumetric ratio of non-solvent to solvent that is still tolerated in the solution is termed the *dilution ratio*. The dilution ratio is an empirical, dimensionless quantity that provides information on the solvency of a solvent or solvent mixture. In order to determine the dilution ratio toluene or butanol is used as the diluent (non-solvent).

Table 6 shows the dilution ratios for various diluents and solvents. As a rule the dilution ratio decreases with increasing boiling point of the solvent; however, the solvent of lowest molecular mass in a homologous series (e.g., methanol, methyl acetate, methyl glycol) often behaves differently.

Mixtures of two solvents frequently have a higher dilution ratio than the pure components: ether – alcohol and toluene – ethanol mixtures are better solvents for cellulose nitrate than any of the pure solvents; ethyl acetate and butyl acetate are better solvents for cellulose nitrate – synthetic resin combinations if they contain a proportion of ethanol or butanol, respectively. The solubility parameters of the mixed solvents are more comparable to those of the binder than those of the pure solvents.

Table 6. Dilution ratios for cellulose nitrate

Solvent (volume ratio)	Diluent (non-solvent)			
	Toluene	Xylene	Aliphatic hydrocarbon	Butanol
Methyl acetate	3.0		0.9	
Ethyl acetate	3.4	3.3	1.0	8.4
Propyl acetate	3.0	2.0	1.2	
Butyl acetate	2.7	2.7	1.5	8.2
Ethyl glycol acetate	2.5	2.5	1.0	7.5
Acetone	4.5	3.9	0.6	7.0
Methyl ethyl ketone	4.5	3.3		
Methyl isobutyl ketone	3.5	3.7		
Cyclohexanone		7.5	1.0	
Methyl glycol	4.7	2.9	0.2	
Ethyl glycol	5.0	4.7	1.0	6.9
Butyl glycol	4.0	3.2	2.3	
Methanol	2.5	2.0	0.3	
Butyl acetate/butanol (9/1)	3.5			
Butyl acetate/butanol (5/5)	5.5			
Methyl ethyl ketone/ethanol (85/25)	5.0			
Ethyl acetate/ethanol (1/1)	3.5			

The temperature dependence of the dilution ratio is shown in Table 7. The solvency of solvents consisting of small molecules increases with increasing temperature, whereas solvents consisting of large molecules behave in the opposite way. In other experiments, the dilution ratio generally decreases with increasing temperature, as has been demonstrated in cellulose nitrate solutions in butyl acetate, ethyl glycol, and methyl isobutyl ketone. The reason for this is that a cellulose nitrate solution adopts the gel state with a rise in temperature. The gel reverts reversibly to the liquid state when the temperature is lowered. However, for thermodynamic reasons the solvency of a solvent always increases with rising temperature. There is thus no meaningful connection between the dilution ratio and solvency of a solvent.

Table 7. Dependence of the dilution ratio on temperature (solute = cellulose nitrate; diluent = toluene)

Solvent	−10 °C	20 °C	50 °C
Ethyl acetate	2.48	2.58	2.62
Butyl acetate	2.74	2.70	2.61
Amyl acetate	2.66	2.52	2.26
Octyl acetate	1.85	1.74	1.44

As expected, the dilution ratio falls with rising cellulose nitrate concentration in the solution (Table 8).

Table 8. Concentration dependence of the dilution ratio of cellulose nitrate solutions in butyl acetate

Cellulose nitrate, g	Butyl acetate, mL	Toluene, mL	Dilution ratio
1.0	20	75	3.75
2.0	20	71	3.55
3.0	20	67	3.35
4.0	20	63	3.15

The dilution ratio reveals some similarities between the solvency of different solvents. However, most measurements are restricted to cellulose nitrate, and extrapolation of the results to other binders seldom provides correct results. Furthermore the dilution ratios, unlike solubility parameters, are unable to explain several phenomena (e.g., the enhancement of the solvency of a solvent produced on adding a non-solvent).

The apparent or theoretical *dilutability* denotes the dilution ratio that just results in a saturated solution. The true or practical dilutability predicts that this solution to which diluent has been added can still be applied satisfactorily, for example in the paint sector clear varnish films can form from these solutions after drying.

The limiting concentrations in a ternary system can be represented not only by a space diagram, but also by a three-coordinate system in the form of an equilateral triangle (Fig. 3).

The corners of the triangle represent the pure (100 %) solvents A, B, and C. The point X on

the side A – B gives the composition of a solvent mixture comprising B – X % of solvent A and A – X % of solvent B. The composition of a point P within the triangle can be determined by constructing perpendicular lines from P onto the sides of the triangle to obtain the points of intersection a, b, and c. The lengths P – a, P – b and P – c give the percentage contents of the solvents A, B, and C, respectively.

The sum of these lengths is taken as equal to 100 %. Solubility regions can thus be determined in ternary systems by measuring the quantitative miscibility of the solvent A on dilution with solvents B and C or B – C mixtures.

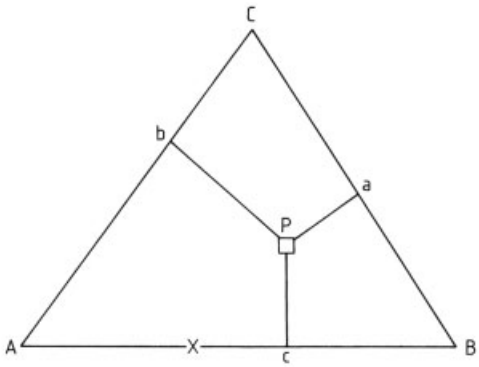

Figure 3. Triangular coordinates for determining the composition of a ternary mixture of solvents A, B, and C (for description, see text)

2.7. Influence of Molecular Mass on Solubility

The solubility of chemically related compounds decreases with increasing molecular mass since the intermolecular forces of interaction increase. For example, benzene is completely miscible with ethanol, whereas anthracene and ethanol are only partially miscible. The influence of molecular mass on solubility is particularly evident in macromolecules. For example, alcohol, acetone, and acetic acid readily dissolve styrene, but not polystyrene; vinyl acetate dissolves in saturated hydrocarbons and ether, whereas poly(vinyl acetate) does not. Cellulose is insoluble in alcohols, poly(ethylene glycol) is insoluble in ethers, poly(vinyl chloride) is insoluble in vinyl chloride, and polyacrylonitrile is insoluble in acetonitrile, even though good solubility would be expected on account of the chemical relation between the polymers and monomers.

On account of their very large molecular mass, highly cross-linked polymers do not dissolve in solvents, even at elevated temperature. They are, however, swollen by solvents depending on the nature and density of the cross-linking sites [62], [63].

2.8. Dissolution and Solution Properties

Rate of Dissolution. Section 2.1 dealt with the thermodynamic equilibrium between the solvent and substance to be dissolved; the mixing time was disregarded. Dissolution proceeds, however, at a finite rate which depends on the specific surface area of the substance to be dissolved, its degree of crystallisation, its rate of diffusion into the solvent, and on the temperature. Often a substance is considered to be sparingly soluble if its rate of dissolution is low. This assumption is, however, incorrect [64].

Miscibility Gap. Some solvent pairs are miscible with one another in all proportions only above or below a critical dissolution temperature. The region in the concentration – temperature diagram in which both partners are immiscible with one another is termed a miscibility gap [17], [65], [66] (Figs. 1 and 4). Hexane and nitrobenzene are completely miscible only above 19.5 °C (Fig. 4 A). Other systems with an upper critical dissolution temperature include:

Aniline – water (168 °C),
2-Chlorobenzoic acid – water (126.2 °C),
Water – phenol (68.8 °C)
Hexane – aniline (65.9 °C)
Cyclohexane – methanol (45.6 °C)

Systems with a lower critical dissolution temperature include:

Triethylamine – water (17 °C, Fig. 4 B)
Ethylpiperidine – water (7.5 °C)

Systems with an upper and lower critical dissolution temperature include nicotine – water (60 and 218 °C, Fig. 4 C).

The miscibilities of various solvents with water are given in Table 16.

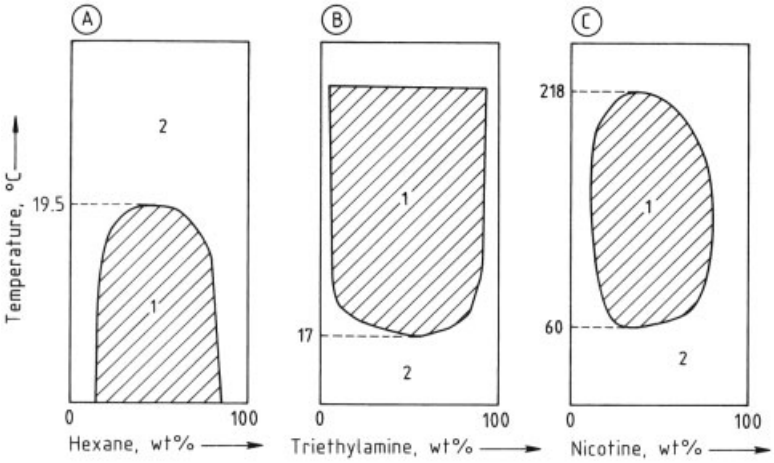

Figure 4. Two-phase solvent systems with miscibility gaps
A) Hexane – nitrobenzene; B) Triethylamine – water; C) Nicotine – water
1 = two-phase region; 2 = one-phase region

Physical Properties of Mixtures. The physical properties of an ideal mixture (e.g., vapor pressure) are obtained from the arithmetic mean of the properties of the individual components. In most mixtures, however, behavior is not ideal and the vapor pressure is greater or less than the sum of the values of the components. The reason for this is that the interaction forces in the pure components are different from those in the mixture. If the molecular forces of interaction in the pure components are lower than in the mixture, the vapor pressure of the mixture is lower than the sum of the pressures of the components (e.g., acetone – trichloromethane). In most cases, however, the forces of interaction in the pure components are higher than in the mixture, with the result that the vapor pressure increases on mixing (e.g., benzene – ethanol).

Similar variations in the densities and surface tensions of mixtures are also known. The density of a mixture of bromo- and chlorobenzene is equal to the arithmetic mean of the densities of the individual components, whereas the density of trichloromethane – diethyl ether is about 1.5 % above the mean value. The surface tensions of the systems diethyl ether – benzene and benzene – carbon disulfide are equal to the calculated values. The surface tension of methanol – ethyl iodide is greater and that of benzene – ethanol and acetone – carbon disulfide are less than the calculated values.

Azeotropes. Molecular associations between the components of a mixture can result in systems that have a constant boiling point at a given concentration. The boiling point may be lower or higher than that of the individual components. Such mixtures are termed azeotropic mixtures or azeotropes. Low-boiling and high-boiling mixtures are denoted as minimum or maximum azeotropes, respectively. Examples of low-boiling azeotropes [67], [68] are benzene – water (*bp* 69.4 °C) and benzene – ethanol (*bp* 67.8 °C). Acetone and chloroform form a high-boiling azeotrope (*bp* 64.7 °C).

The composition of the liquid phase of the azeotrope is the same as that of the vapor phase with which it is in equilibrium. It depends on temperature and pressure, and azeotropically boiling mixtures are not necessarily azeotropically evaporating mixtures. However, azeotropically evaporating mixtures often form even at room temperature; their composition differs from that of the azeotropically boiling mixtures. Azeotrope formation can be suppressed by lowering the boiling point (vacuum distillation) and the components can then be separated if the vapor pressures of the pure components respond differenly to a change in temperature. This is always the case if the heats of vaporization of the pure components differ widely.

The formation of low-boiling azeotropes is a fortunate phenomenon for the paint technologist

because water and solvents then evaporate more rapidly than usual (azeotropes, see Table 17). However, these azeotropic mixtures also have several disadvantages:

1) The flash point is lower
2) The explosion limit is lower
3) The higher evaporation rate may have a negative effect on the paint surface
4) Paint flow may be poor because the azeotropic mixture constantly evaporates

Systems Containing More Than Two Components. As in binary systems, the behavior of systems containing more than two components can be understood on the basis of intermolecular forces and solubility parameters. Water and tetrachloromethane have widely differing solubility and hydrogen bond parameters, and are therefore immiscible. Added acetone dissolves partly in the aqueous phase due to hydrogen bond formation, and partly in the tetrachloromethane phase due to dispersion and induction forces. Twice as much acetone dissolves in the aqueous phase as in tetrachloromethane. On increasing the acetone concentration a homogeneous solution is obtained. The added solvent thus acts as a *solubilizer* for the two immiscible solvents.

Solvents having average solubility and hydrogen bond parameters are generally suitable as solubilizers, particularly ketones and glycol ethers. Butyl glycol, diglycol, and triglycol are often used because they contain hydrophilic and hydrophobic groups which effect miscibility between partners having widely differing solubility and hydrogen bond parameters. Very small amounts of a solubilizer are often sufficient to homogenize a heterogeneous solvent mixture. The two-phase system comprising hexane and methanol (mass ratio 2 : 1) is rendered homogeneous by adding 0.5 wt % dichloromethane.

Addition of solvents as solubilizers is important in the dissolution of polymers. Poly(vinyl acetate) is insoluble in pure ethanol, but dissolves fairly readily in ethanol containing 3 % water. Cellulose triacetate is sparingly soluble in pure trichloromethane, but is readily soluble in trichloromethane containing 3 % methanol. Many paint resins dissolve more readily in aromatic-containing naphtha fractions than in pure naphtha. Poly(vinyl chloride) is sparingly soluble in acetone and carbon disulfide, but is more readily soluble in a mixture of the two solvents.

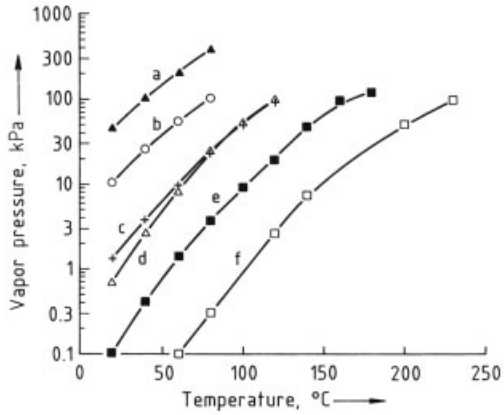

Figure 5. Vapor pressure curves of solvents
a) Dichloromethane; b) Ethyl acetate; c) Butyl acetate; d) Butanol; e) Butyl glycol; f) Butyl diglycol

3. Physical and Chemical Properties

see Tables 23 and 24.

3.1. Evaporation and Vaporization

Solvents are classified according to their boiling point ranges into:

1) Low boilers: $bp < 100\,°C$
2) Medium boilers: bp 100 – 150 °C
3) High boilers: $bp > 150\,°C$

The boiling point of a liquid is defined as the temperature at which the vapor pressure of the liquid reaches 101.3 kPa. Thermal energy is consumed in the evaporation of a liquid and is extracted from the surroundings, resulting in cooling.

A knowledge of the boiling point ranges of solvents is important in distillation and extraction processes. However, in the paint and adhesive sectors a knowledge of solvent volatility below the boiling point is more important. Below the boiling point the liquid is in equilibrium with its vapor, the vapor pressure p_s is < 101.3 kPa and is determined from the vapor pressure curve according to the Clausius – Clapeyron equation (integrated form):

$$\ln p_S = A - \frac{\Delta H_V}{RT}$$

where A is a general constant, ΔH_V the molar enthalpy of vaporization, R the gas constant, and T the absolute temperature. The higher the temperature, the greater the vapor pressure (Fig. 5); ΔH_V increases with increasing boiling point of the solvent according to the Pictet – Trouton rule:

$$\frac{\Delta H_V}{T_S} = \text{constant}$$

where T_S is the absolute boiling point.

There is no general correlation between the evaporation rates and boiling points of solvents. Solvent volatility generally decreases with increasing boiling point if the solvents are chemically related [69]. Solvents that tend to form hydrogen bonds (e.g., water, alcohols, amines) are less volatile than other solvents with the same boiling point since energy has to be supplied to rupture the hydrogen bonds before the transition to the vapor state occurs. The evaporation rate w of solvents is obtained empirically according to the Knudsen equation [70], [71]:

$$w = K P_V \sqrt{\frac{2\pi RT}{M}}$$

where the transmission coefficient $K \approx 1$, P_V is the vapor pressure, M is the molecular mass, and ω is expressed in g/cm^{-2} s^{-1}.

The evaporation rate of solvents depends on:

1) Vapor pressure at the processing temperature
2) Specific heat
3) Enthalpy of vaporization
4) Degree of molecular association
5) Rate of heat supply
6) Surface tension
7) Molecular mass of the solvent
8) Atmospheric turbulence
9) Atmospheric humidity

Since these factors all depend on one another, it is impossible to give a theoretical prediction of the rate of evaporation [70–76]. In practice, the evaporation time of a given amount of solvent is determined experimentally under identical external conditions and compared with that of diethyl ether, or in some countries with that of butyl acetate.

The results are termed *evaporation numbers* (Table 9) and are referred to diethyl ether = 1. Evaporation numbers are listed in Table 24.

Solvents can be subdivided into four groups on the basis of their evaporation numbers:

1) High volatility: < 10
2) Moderate volatility: 10 – 35
3) Low volatility: 35 – 50
4) Very low volatility: > 50

A similar volatility classification of solvents is adopted in the United States with butyl acetate = 1 being chosen as the reference:

Class 1: rapid evaporation, > 3.0
Class 2: moderate evaporation, 0.8 – 3.0
Class 3: slow evaporation, < 0.8

If a substance with a relatively high molecular mass is dissolved in a solvent, its rate of evaporation decreases [77–80]. Solvents evaporate more rapidly from cellulose ester and cellulose ether solutions than from polymer or resin solutions. The volatility of a solvent in a polymer film depends on the difference between the solubility parameters of the solvent and the polymer. The polymer occupies a certain space in the solubility parameter diagram with coordinates δ_D, δ_P, and δ_H. Solvents that can dissolve the polymer are situated outside it (see Fig. 2). A solvent whose parameter lies close to the center of the polymer solubility space interacts strongly with the polymer; the solvent therefore evaporates slowly from the drying polymer film, and may even be retained. Solvents that interact weakly with the polymer are situated at the boundary of the polymer solubility space (i.e., in the transition region between solvents and non-solvents) and therefore evaporate more readily from the polymer film [33].

Solvents should not only evaporate rapidly after the application of a paint film, but should also promote the formation of a coating as follows:

1) Prevent blushing and absorption of condensed water vapor from the atmosphere, which is formed as a result of cooling due to solvent evaporation
2) Control the paint flow
3) Prevent precipitates and turbidity in the paint film

Table 9. Evaporation of glycol ethers (percentage evaporated as a function of time at 23 °C)

Solvent	Time, h										Evaporation
	10	15	20	25	50	100	150	200	300	400	number (diethyl ether = 1)
Ethyl glycol	40	58	77	93							43
Butyl glycol					30	60	92				119
Ethyl diglycol					0	3	11	18	34	51	1200
Butyl diglycol							0	3	6	9	> 1200

4) Prevent film shrinkage

A paint formulation therefore always contains several solvents: low boilers, medium boilers, and a small proportion of high boilers. The vapor pressure of the individual solvents in the mixture is described by Raoult's law:

$$p_i = f_i \cdot x_i \cdot p_i^{\circ}$$

where p_i is the partial pressure of component i, x_i the mole fraction of component i, and p_i° the vapor pressure of the pure component. The activity coefficient f_i, which may be larger or smaller than 1, takes account of the real behavior of the solvent (i.e., deviations from the ideal state).

3.2. Hygroscopicity

Some solvents, particularly those containing hydroxyl groups (e.g., alcohols and glycol ethers) are hygroscopic: they absorb moisture from the atmosphere until an equilibrium is reached. The equilibrium water content of the solvent depends on the relative atmospheric humidity and temperature. The hygroscopic properties of glycol ethers have been investigated in detail (Table 10).

3.3. Density and Refractive Index

In addition to the boiling curve, the density and refractive index are used to assess and, with some restrictions, to determine solvent purity.

The *density* of a solvent is generally measured at 20 °C and referred to the density of water at 4 °C (relative density d_4^{20}). The densities of most organic solvents decrease with increasing temperature (Fig. 6) and are less than that of water, but halogenated hydrocarbons are denser than water. The relative densities (d_4^{20}) of homologous esters and glycol ethers decrease with increasing molecular mass, whereas those of ketones and alcohols increase:

Esters		*Alcohols*	
Methyl acetate	0.934	Methanol	0.791
Ethyl acetate	0.901	Ethanol	0.789
Propyl acetate	0.886	Propanol	0.804
Butyl acetate	0.881	Butanol	0.810
Amyl acetate	0.876	Amyl alcohol	0.815
Glycol Ethers		*Ketones*	
Methyl glycol	0.966	Acetone	0.792
Ethyl glycol	0.931	Methyl ethyl ketone	0.805
Propyl glycol	0.911	Methyl propyl ketone	0.807
Butyl glycol	0.902	Amyl methyl ketone	0.816

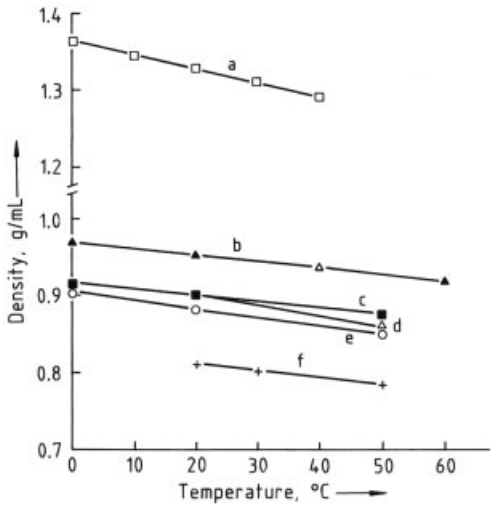

Figure 6. Temperature dependence of solvent densities
a) Dichloromethane; b) Butyl diglycol; c) Butyl glycol; d) Ethyl acetate; e) Butyl acetate; f) Butanol

The *refractive index* n_D is measured in a refractometer with a sodium vapor lamp (Na-D lines, 589.0 and 589.6 nm). The value of the refractive index [46], [47], [81] is largely de-

termined by the hydrocarbon skeleton of the substance in question. Aliphatic esters, ketones, and alcohols have refractive indices between 1.32 and 1.42. In homologous series the refractive index increases with increasing length of the carbon chain, and decreases with increasing branching. Cycloaliphatic and aromatic structures increase the refractive index (n_D^{20}), as does the incidence of functional groups:

Methyl acetate	1.3610	Cyclohexanol	1.4667
Ethyl acetate	1.3718	Benzyl alcohol	1.5390
Propyl acetate	1.3844	Methylbenzyl alcohol	1.5270
Isopropyl acetate	1.3773	Toluene	1.4955
Butyl acetate	1.3961	*p*-Xylene	1.4956
Isobutyl acetate	1.3898	Tetrahydronaphthalene	1.5443
Methanol	1.3290	Ethyl glycol	1.4075
Ethanol	1.3619	Ethylene glycol	1.4310
Propanol	1.3859	Diethylene glycol	1.4460
Isopropyl alcohol	1.3772	Dichloromethane	1.4234
Butanol	1.3994		
Isobutanol	1.3960		

The refractive index decreases with increasing temperature (Fig. 7).

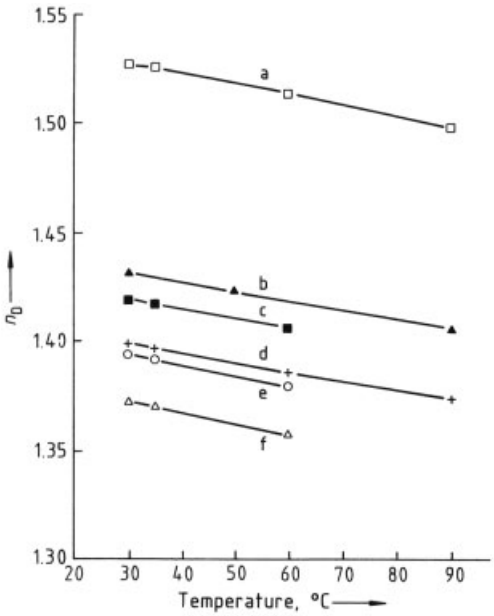

Figure 7. Temperature dependence of the refractive index of solvents
a) Methylbenzyl alcohol; b) Butyl diglycol; c) Butyl glycol; d) Butanol; e) Butyl acetate; f) Ethyl acetate

Table 10. Hygroscopicity of glycol ethers expressed as equilibrium water content (wt %) of solutions in contact with air of different relative atmospheric humidities (101.3 kPa, 25 °C)

	Relative atmospheric humidity, %						
Solvent	40	50	60	70	80	90	100
Ethyl glycol	5	8	12	20	31	48	77
Butyl glycol	4	7	11	18	27	37	49
Ethyl diglycol	6	8	14	20	30	40	53
Butyl diglycol	3	6	9.5	17	25	37	49
Ethylene glycol	18	26	37	47	60	74	89

3.4. Viscosity and Surface Tension

The *viscosities* of homologous series of solvents (Table 23) increase with increasing molecular mass. Solvents bearing hydroxyl groups have higher viscosities due to the formation of hydrogen bonds. The solvent viscosity greatly influences solution viscosity (see Section 7.1). The viscosity decreases with increasing temperature (Fig. 8).

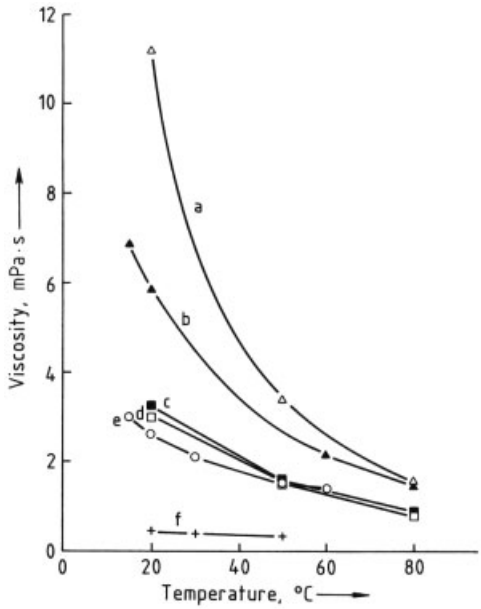

Figure 8. Temperature dependence of solvent viscosities
a) Methylbenzyl alcohol; b) Butyl diglycol; c) Butyl glycol; d) Butanol; e) Isophorone; f) Ethyl acetate

The surface tension of paint solvents is of importance for the rate of evaporation, for the formation of the coating surface, and as for the wetting of the substrates, extenders, and pigments.

The *surface tension* of a solvent is related to the cohesive energy density and internal pres-

Table 11. Thermal and electrical properties of solvents

Solvent	Thermal conductivity (20 °C), $Wm^{-1}\,K^{-1}$	Cubic expansion coefficient, $K^{-1}\times10^{-3}$	Dielectric constant (20 °C)	Electrical conductivity, S/cm
Cyclohexane			2.01	1.9×10^{-14}
Xylene		0.9	2.3	10^{-5}
Dichloromethane	0.159	1.3	7.0	4.3×10^{-11}
Tetrachloroethylene	0.16	1.08	2.5	5.6×10^{-4}
Ethanol	0.171	1.1	22.4	1.4×10^{-9}
Butanol		0.79	18.2	9.1×10^{-9}
Butyl acetate	0.14 (30 °C)	1.21	4.5	5.2×10^{-11}
Ethyl glycol acetate	0.143 (25 °C)	1.12	9.0	1.2×10^{-8}
Methyl ethyl ketone		1.27	18.4	3.6×10^{-9}
Methyl isobutyl ketone		1.15	13.1	3×10^{-9}
Butyl glycol	0.162 (30 °C)	0.92	9.2	4.3×10^{-7}
Butyl diglycol	0.139	0.85	11.0	2×10^{-8}
Diethyl ether	0.132		4.34	4×10^{-13}
Tetrahydrofuran	0.14	1.29	7.6	0.5×10^{-8}

sure of the liquid. A relationship can be derived between the solubility parameter δ (which is defined as the square root of the cohesive energy density) and the surface tension γ_o [82]. For polar and nonpolar solvents the following relationship applies:

$$\delta = 2.1K\left(\frac{\gamma_o}{V^{1/3}}\right)^a$$

where V is the molar volume and K and a are constants ($K \approx 3.6$, $a \approx 0.56$). This empirical relationship is extremely satisfactory (Table 18).

3.5. Vapor Density

The vapor density is the mass of solvent vapor per cubic meter of air (kg/m^3) that is in equilibrium with the liquid at 101.3 kPa. The vapor density thus corresponds to the solvent content in the atmosphere at saturation and is temperature dependent. The relative vapor density of solvents d_S is referred to the density of air and can be calculated according to:

$$d_S = \frac{M_S}{M_{air}}$$

where M_S is the molecular mass of the solvent, M_{air} is the mean molecular mass of air (= 28.95 g/mol).

In the ideal case, the relative vapor density is temperature independent. Relative vapor densities of some solvents follow:

Toluene	3.18
Xylene	3.67
Ethanol	1.59
Butanol	2.56
Ethyl acetate	3.04
Butyl acetate	4.01
Methyl ethyl ketone	2.49
Ethyl glycol	3.11
Dichloromethane	2.93

3.6. Thermal and Electrical Data

The dielectric constant and thermal conductivity decrease with increasing temperature, whereas the specific heat increases. The thermal conductivities, cubic expansion coefficients, dielectric constants, and electrical conductivities of various solvents are listed in Table 11. Critical data of solvents and the technical use of supercritical liquids are described in [83].

3.7. Flash Point, Ignition Temperature, and Ignition Limits

The temperature at which a solvent vapor–air mixture is ignited by a naked flame is termed the *flash point* of the solvent. The flash points of solvents increase with decreasing vapor pressure and thus with increasing molecular mass and boiling point. They form the basis for their classification according to flammability (danger classes, transport regulations, Section 5.3). Flash points of some solvents follow:

Methyl acetate	−13 °C	Acetone	−19 °C
Ethyl acetate	−12 °C	Methyl ethyl ketone	−14 °C
Propyl acetate	12 °C	Methyl propyl ketone	22 °C
Butyl acetate	22 °C	Methyl butyl ketone	23 °C
Amyl acetate	34 °C	Methyl isobutyl ketone	15 °C
Ethyl glycol acetate	52 °C	Methyl amyl ketone	49 °C

The absolute flash point T_F of a solvent is related to its absolute boiling point T_S according to the following empirical equation [84]:

$$T_F = 0.736 T_S$$

Flash points calculated according to this equation are sufficiently accurate (deviation ± 5 °C) in the case of hydrocarbons, ketones, and esters. The equation is, however, unsatisfactory in the case of solvents containing hydroxyl groups such as alcohols, glycols, and glycol ethers (deviation up to ± 30 K).

A relationship between the flash point and vapor pressure p has also been derived:

$$T_F = a \log p + b$$

where a is a general constant and b is a substance-specific constant.

For safety reasons numerous attempts have been made to raise the flash point of solvent mixtures [85]. Azeotropes having a lower flash point than those of the pure components can be formed on mixing solvents; use of solvent mixtures to raise the flash point is therefore extremely problematic. The flash point of toluene can, for example, be significantly raised above 21 °C by adding chlorinated hydrocarbons [86], [87]. However, the flash point is first lowered instead of raised if dichloromethane is added. Non-flammability is achieved only above a minimum addition of dichloromethane.

The flash point of a solvent mixture is not identical to that of its most flammable component. When solvents with widely differing hydrogen bond parameters are mixed (e.g., alcohol–hydrocarbon), the flash point is significantly reduced. On the other hand, the flash point of a mixture of chemically related solvents lies between those of the individual components [88–90]. Methods have been developed for calculating the flash point of solvent mixtures and solutions, activity coefficients are used to account for nonideal behavior [91].

Flash point determinations on mixtures of xylene with polar solvents led to the following conclusions [92], [93]:

1) Small quantities of solvents containing carbonyl groups (esters, ketones) reduce the flash point, larger amounts increase it again
2) The flash point is raised by adding 4 wt % of butanol or 8 wt % of isobutanol
3) Small additions of water do not significantly influence the flash point
4) Addition of ethanol sharply reduces the flash point; ethyl glycol acts in the same way, although to a lesser extent

In order to evaluate the flammability of water-thinnable paints, a knowledge of the increase in the flash points of water-miscible solvents after addition of water is important. These mixtures no longer ignite above a specific ratio of water to organic solvent (following results at 25 °C):

1) Isopropyl alcohol – water mixtures are still combustible in a 40/60 ratio, but not combustible in a 30/70 ratio
2) Propanol – water mixtures are still combustible in a 70/30 ratio, but not combustible in a 60/40 ratio

Flash points of solvent – water mixtures are given in Table 12.

Table 12. Flash points (°C) of solvent – water mixtures

Flash point	Solvent to water ratio			
	1 : 0	9 : 1	1 : 1	1 : 9
Ethanol	10	19	25	51
Isopropyl alcohol	12	19	24	41
Ethyl glycol	46	57	72	
Methyl ethyl ketone	−9	−7	−6	5–9
Butyl glycol	60	> 65		

A solvent vapor is ignited not only by a naked flame, but can spontaneously ignite when the solvent vapor – air mixture has reached the *ignition temperature*. A distinction is made between the gas ignition temperature and drop ignition temperature, depending on whether the measurement is made by determining the gas temperature or by allowing the solvent to drop onto a hot surface of known temperature [94]. The value of the ignition temperature is used

to group solvents into temperature classes (ignition groups). For example, VDE Regulation 0171 specifies five groups:

T 5	100 – 135 °C
T 4	135 – 200 °C
T 3	200 – 300 °C
T 2	300 – 450 °C
T 1	> 450 °C

An explosion is a self-propagating, particulary rapid combustion process that is initiated by an ignition and proceeds without external energy or air. The explosion of a solvent vapor – air mixture is possible only within a specific solvent concentration range, defined by the lower and upper ignition limits (explosion limits) which are specified in volume percent or grams per cubic meter of solvent in the solvent vapor – air mixture at 101.3 kPa.

The temperatures at which the concentrations of solvent vapor in equilibrium with the liquid reach the lower or upper ignition limit are obtained from the vapor pressure curves of the solvents; these temperatures are termed the lower and upper explosion points. The lower explosion points roughly correspond to the flash points of the solvents [95–99].

3.8. Heats of Combustion and Calorific Values

The combustion of organic solvents releases energy since organic compounds generally have higher energy contents than their combustion products. The combustion energy, also termed heat of combustion or gross calorific value, is given in kilojoules per mole or kilojoules per kilogram.

The gross calorific value refers to liquid water of reaction, whereas the net calorific value refers to the water vapor that is formed. The two calorific values thus differ by the contribution of the enthalpy of vaporization of the water formed during combustion.

The heats of combustion of alcohols, esters, and glycol ethers are very much smaller than those of hydrocarbons (Table 19). Electronegative substituents generally reduce the combustion energy in the order $NH_2 > Cl > OH > SH$. The larger the number of substituents in the molecule, the lower the combustion energy. A knowledge of the calorific values of solvents is of great industrial importance because they decisively influence the size and cost of combustion and solvent incineration plants [100], [101].

3.9. Chemical Properties

A high chemical resistance is an important prerequiste for the use of a liquid as a solvent. Aliphatic and aromatic hydrocarbons are chemically inert and thus satisfy this requirement extremely well.

Alcohols too are chemically very resistant but they dissolve alkali metals, alkaline-earth metals, and aluminum to form salts. Under certain conditions alcohols can be converted into carboxylic acids by powerful oxidizing agents. They are, however, stable toward atmospheric oxygen. On account of their reaction with isocyanates to form urethanes, solvents containing hydroxyl groups (e.g., alcohols, glycols, and glycol ethers) must not be used as solvents for polyurethane paints [102], [103].

On prolonged storage most ethers and glycol ethers form peroxides with atmospheric oxygen. In the case of glycol ethers this leads to acid formation due to oxidation which can be prevented by adding stabilizers.

Esters and ketones are chemically very resistant under normal conditions, especially in the paint industry. However, it must be remembered that esters can be hydrolyzed to form alcohols and acids. For ethyl acetate the equilibrium constant K at 40 °C is 2.51, and at 100 °C is 2.56.

$$K = \frac{[\text{Ethyl acetate}] \cdot [H_2O]}{[\text{Acetic acid}] \cdot [\text{Ethanol}]}$$

In a neutral medium hydrolysis begins extremely slowly, but is autocatalyzed by the acetic acid product. The hydrolysis kinetics proceed according to:

$$\frac{dx}{dt} = k\,(a - x)\,\sqrt{x}$$

where k is the reaction constant, a is the initial concentration of the ester, and x is the acid concentration. For example, when 44 g of ethyl acetate is dissolved in 1 L of water at 20 °C, about 15 g is hydrolyzed after 130 days. Under acid or base catalysis the reaction rate is 10^5-fold or 10^8-fold higher, respectively.

The hydrolysis rate of the ester depends on its chemical structure. The rates of alkaline hydrolysis of isobutyl and butyl acetates are greater at room temperature than that of *sec*-butyl acetate; *tert*-butyl acetate is even more stable under alkaline conditions (iso $\approx n > sec >> tert$). The rates of acid hydrolysis of the butyl acetates decrease in the order iso $> n > tert > sec$.

Chlorinated hydrocarbons can release hydrogen chloride in the presence of bases or metals, but stabilizers added by the manufacturer ensure high chemical resistance. Highly stable solvents are nitro compounds, dimethylformamide, and dimethyl sulfoxide, which, however, often become dark on prolonged storage in air.

Before using a solvent it is advisable to become acquainted with the chemical properties so that interactions can be anticipated and explained. A comprehensive literature exists on solvent effects in chemical reactions [3], [103], [104].

4. Toxicology and Occupational Health

4.1. Toxicology

Acute and Chronic Toxicity. Solvents act with different intensities on human, animal, and plant organisms. Their effects depend on the amount of solvent and the exposure time. Under short-term exposure to high solvent doses acute damage may occur, whereas the absorption of smaller amounts over a longer period leads to chronic damage and sensitization. The chronic effects are more dangerous since they are accompanied by an acquired tolerance, with the result that they are often not detected early enough.

In order to evaluate the solvent toxicities, their LD_{50} values (oral administration to rats) are listed in Table 24. Since most cases of solvent poisoning are caused by inhalation of solvent vapors, the LC_{50} values are also given.

Inhaled solvent vapors pass via the lungs and blood circulation into the body, where they accumulate in tissues with high lipid content (e.g., nerves, brain, bone marrow, adipose tissue, liver, and kidneys) [105]. The cells can either be damaged by the solvents or by their decomposition products. Solvents can also pass into the body via cutaneous or, more rarely, gastrointestinal absorption [106], [107].

Symptoms of acute solvent poisoning include dizziness, drowsiness, headache, loss of consciousness, and narcotic effects which are attributed to disturbances of the central nervous system. Chronic poisoning is initially undetectable, but subsequently causes damage to organs that are specific for each solvent [108–111].

Solvents have two effects on the skin: (1) they dissolve the natural fatty layer, the skin therefore cracks and microorganisms and dirt particles can penetrate more easily and cause infection; (2) solvents can act directly to cause inflammation or blistering.

The following solvents are absorbed very readily through the skin and pass into the body: aniline, benzene, butyl glycol, butyl glycol acetate, dimethylacetamide, dimethylformamide, dioxane, ethyl glycol, ethyl glycol acetate, ethylbenzene, isopropyl glycol, carbon disulfide, methanol, methyl glycol, methyl glycol acetate, methylcyclohexane-2-one, 4-methyl-2-pentanol, nitrobenzene, nitrotoluene, isopropylbenzene, 1,1,2,2-tetrachloroethane, and tetrachloromethane.

After sensitization of the skin or respiratory system, allergic reactions may occur but vary widely, depending on the individual susceptibility. Observance of the MAK or TLV values is no guarantee against the occurrence of such reactions [112]. Turpentine oil is the only solvent that has attracted attention because of its ability to trigger hypersensitivity reactions of an allergic nature.

Carcinogenicity. In Germany, carcinogenic solvents are classified as "III Carcinogenic Substances" in the Technische Regeln für Gefahrstoffe TRG 900 (Technical Regulations Covering Dangerous Substances).

The solvents are subdivided into Section A (substances shown to be unequivocally carcinogenic) and Section B (substances justifiably suspected of being potential carcinogens). Section A is divided into Subsection A 1 (substances that according to experience may cause malignant tumors in humans) and Subsection A 2 (substances which have clearly been shown to be carcinogenic in animal experiments).

Section A 1 includes benzene; Section A 2 includes 1,2-dichloroethane, hexamethylphosphoric triamide, and 2-nitropropane; Section B includes aniline, chloroethane, dichloromethane, 1,4-dioxane, 1,1,2,2-tetrachloroethane, tetrachloroethylene, tetrachloromethane, 1,1,2-trichloroethane, trichloroethylene, and trichloromethane.

No maximum allowable concentrations (MAK values) are specified for substances in Section A since no concentrations can be regarded as harmless. If the use of these substances is technically necessary, special monitoring and protective measures must be adopted (see Section 4.2). Particular care should also be exercised when handling substances in Section B.

Teratogenicity and Embryotoxicity. In Germany, substances that are potentially teratogenic or embryotoxic are classified in four groups (see Table 24):

Group A. Teratogenic and embryotoxic effects have been demonstrated.

Group B. Teratogenic and embryotoxic effects are likely. Examples include dimethylformamide, ethyl glycol, ethyl glycol acetate, methyl glycol, methyl glycol acetate, 2-methoxy-1-propanol, 2-methoxy-1-propyl acetate, carbon disulfide, toluene, and trichloromethane. TRGS 609 give details of limitations on use, substitutes, and alternative processes for methyl and ethyl glycols and their acetates.

Group C. There is no risk of teratogenic and embryotoxic effects provided the MAK values are observed. Examples include butyl glycol, butyl glycol acetate, cyclohexanone, dimethylacetamide, hexane, isopropyl glycol, 1-methoxy-2-propyl acetate, styrene, tetrachloroethylene, tetrahydrofuran, 1,1,1-trichloroethane, and trichloroethylene.

Group D. An evaluation is still not possible. Examples include aniline, methyl ethyl ketone, dichloromethane, diethyl ether, ethanol, ethyl formate, methyl formate, methanol, methyl acetate, *N*-methylpyrrolidone, nitrobenzene, isopropyl alcohol, tetrachloromethane, and xylene.

Mutagenicity denotes damage to male and female gametes that can lead to genetic changes in subsequent generations. Such damage may include gene mutations and changes in the structure and number of chromosomes. Mutagenic substances are subdivided into:

Group 1. Substances for which a mutagenic effect has been detected in humans

Group 2. Substances for which a mutagenic effect has been detected in suckling females in animal experiments

Group 3. Substances for which damage to the genetic material of the gametes has been detected in humans or in animal experiments

4.2. Occupational Health

Personal Safety Precautions. When working with solvents or solvent-containing preparations, contact with the skin and mucous membranes should be avoided. Protective goggles and gloves should be worn and the skin protected with skin cream. Wet articles of clothing should be changed immediately. Inhalation of solvent vapors should be avoided. The guidelines and codes of conduct published by the industrial and trade unions should be observed.

Odor Threshold. Most solvents have a characteristic odor. Human perceptibility and sensitivity to solvent vapors depends on habituation, which varies markedly from one person to another. Odors that are regarded as pleasant in small concentrations may be considered intolerable at high doses and under constant exposure. Other vapors that are initially considered objectionable may subsequently be regarded as tolerable [113]. It is therefore impossible to give objective rules for determining when an odor becomes objectionable. Odor intensity is subdivided into four levels:

1) Imperceptible
2) Weakly perceptible
3) Moderately perceptible
4) Highly perceptible

The odor threshold is the vapor concentration in a cubic meter of air (mg/m^3) that is just still perceptible; it lies between levels 1 and 2 [114]. The odor threshold values of some solvents are given in Table 24.

When handling solvents, the odor should be regarded as a preliminary warning sign. Relatively high solvent concentrations may irritate the mucous membranes. It must be pointed out, however, that some mildly smelling solvents may present more of a health hazard than other, strongly perceptible products.

MAK Values. The toxicity of solvent vapors at the workplace has been investigated in animal experiments and by observing and monitoring humans. The MAK value (Maximale Arbeitsplatzkonzentration) is the maximum permissible concentration of a substance in the atmosphere at the workplace that is generally not injurious to the health of the employees and is not regarded as intolerable by the latter, even after repeated and long-term exposure (generally 8-h exposure per day). As a rule, MAK values are average values over a period of a day or a working shift. Since the actual concentration of the working substances in the inhaled air frequently fluctuates, exposure peaks are specified for individual categories according to short-term value, short-term duration, and frequency per working shift:

Category I. Locally irritant substances: $2 \cdot \mathrm{MAK}$

Category II. Substances that act after absorption with an onset of action ≤ 2 h: $2-5 \cdot \mathrm{MAK}$

Category III. Substances that act after absorption with an onset of action > 2 h: $10 \cdot \mathrm{MAK}$

Category IV. Very weak potential action: $2 \cdot \mathrm{MAK}$

Category V. Substances with an intense odor: $2 \cdot \mathrm{MAK}$

Ventilation equipment at the workplace should be designed and dimensioned so that the MAK values of the individual solvents are not exceeded.

No MAK values can be specified for a number of carcinogenic and mutagenic substances. Some of these substances are, however, unavoidable and exposure to them cannot be excluded. Technische Richtkonzentrationen TRK (lowest technically feasible levels) have been set for these substances to minimize the risk of a health hazard at the workplace. The TRK value for benzene and 2-nitropropane is 5 ppm (TRGS 102).

Since the MAK values apply only to pure substances in the workplace atmosphere, an evaluation index I_{MAK} has been defined in TRGS 403 according to the follwing formula for mixtures of solvent that occur in the majority of cases at the workplace.

$$I_{\mathrm{MAK}} = \sum_{i=1}^{n} \frac{C_i}{\mathrm{MAK}_i}$$

TLV Values. In the United States, the equivalent of the MAK value is the threshold limit value (TLV). The TLV value is that concentration of a substance in the air to which virtually all workers can be exposed daily without any harmful effects. This value is subdivided into TLV-TWA (time-weighted average concentration), TLV-STEL (short-term exposure limit), and TLV-C (ceiling limit). TLV-TWA applies to a normal 8-h working day or a 40-h week, TLV-STEL is the maximum concentration for an exposure time of 15 min, and TLV-C is the concentration that should at no time be exceeded. The TLV values published by the ACGIH are recommendations, whereas the PEL values (permissible exposure limit) specified by OSHA (Occupational Safety and Health Administration) are enforced by law. The safety limits are recommended to OSHA by NIOSH (National Institute of Occupational Safety and Health). Solvent TLV values for the USA and several European countries are listed in Table 13.

Other Limiting Values. In Scandinavia, solvent-containing products are identified with a YL value (yrkeshygieniskt luft behov = industrial hygiene air requirement). This value states how many cubic meters of air are required to dilute the amount of solvent contained in 1 L of product to such an extent that the concentration is below the TLV value. The YL value can be calculated for each solvent by multiplying the YL factor by the percentage content of the respective solvent in the liquid. A PWA factor (**p**aint technology **w**ork hygiene **a**ir requirement) is allocated to the solvents on the basis of their YL value. The PWA number is in turn used as a basis for calculating the potential health hazard of a paint, and serves to specify the protective measures and working clothes that are to be used when applying paints (Swedish Work Safety Office Publication No. 463 and State Factory Inspectorate Publication No. 464, of 3rd August 1982).

A harmfulness factor has been proposed to quantify the harmfulness of solvents [115–117]:

$$\text{Harmfulness factor} = \frac{10\,000}{\mathrm{TLV} \cdot \text{relative evaporation number}}$$

The harmfulness factor takes into account not only the actual health-damaging action of the

Table 13. Threshold limit values of some countries (1991, ppm)

Solvent	Denmark	France	Germany	Italy	The Netherlands	UK	Sweden	USA (TLV-PEL)
Cyclohexane	300	300	300	300	300	300	50	300
Diethyl ether	400	400	400	400	400	400	100	400
Tetrahydrofuran	200	200	200	200	200	200	20	200
Ethyl acetate	300	400	400	400	400	400	200	400
Pentyl acetate	100	100	100	100	100	100	50	100
Methyl ethyl ketone	100	200	200	200	200	200	100	200
Methyl isobutyl ketone	50	50	100	50	50	50	20	100

solvent, but also the time for a health-damaging concentration to accumulate if the product leaks or is split.

5. Environmental and Legal Aspects

5.1. Environmental Protection

Being volatile substances, all solvents inevitably evaporate and pass as harmful substances into the atmosphere [118–121]. They may also affect waterways, lakes, rivers, ground water, and soil. Anthropogenic organic emissions total ca. 10^7 t/a in the EC; of this, 34 % stems from the use of solvents and 53 % from the transport and traffic sector (Corinair Study 1985, Inventaire européen des émissions).

United States. The first environmental protection measures were implemented in particularly threatened, highly industrialized regions with a high density of motorized traffic. In the Los Angeles district of California, exposure of certain organic solvent vapors to solar radiation can lead to the formation of smog [122–124]. In the United States, Public Law 84 – 159 of 1955 passed by the Federal Departments of Health, Education, and Welfare was the basis for local clean air directives. In Rule 66 and Rule 442 (district of Los Angeles), aromatic hydrocarbons (xylene, tetrahydronaphthalene, toluene, and ethylbenzene) and branched- chain ketones (e.g., isophorone, mesityl oxide, methyl isobutyl ketone, methyl isopropyl ketone, and diacetone alcohol) are classified as photochemically reactive, smog-producing solvents. Aliphatic hydrocarbons and nitro compounds are considered to be photochemically nonreactive. Efforts are being made to reduce the solvent content of paints and use photochemically nonreactive solvents or water. Clean air measures in the United States have become more stringent as a result of further laws passed by the EPA, including the Air Pollution Control Act 1962, Clean Air Act 1963, Air Quality Act 1967, National Environmental Policy Act 1969/1970/1975, Clean Air Act 1970 and 1990 (CAA) [125–128]. The implementation of these acts in the federal states is governed by executive provisions (e.g., New York State Rule 187). Since 1989, emissions have been regulated and reduced in stages by restricting the use of solvent- containing coatings in certain areas of application; for example, specific directives apply to the use of paints in automobile repair workshops, air-drying industrial paints, wood paints, swimming-pool paints, and marine paints. Special regulations apply to new industrial sites (e.g., new automobile factories).

Europe.

Germany. An anti-emission law (Immissionsschutzgesetz) was passed by the state of North Rhine – Westphalia, Germany, in 1968. The purpose of this law and various executive directives was to improve living conditions and the atmosphere in the Ruhr district. In 1974, the individual laws of German states (Länder) were unified in the form of the Bundesimmissionsschutzgesetz (Federal Anti-Emission Law) and the Technische Anleitung zur Reinhaltung der Luft, TA Luft (Clean Air Regulations) of 27 Feb., 1986, which has since become known throughout Europe [129–132].

United Kingdom. Environmental Protection Act with Air Pollution Control APC (1990) [133].

Italy. Law 203 of the Ministry for the Environment (1988) including guidelines (1990) that restrict solvent emissions in certain application sectors (e.g., the automobile industry).

The Netherlands. Guideline KWS 2000 to reduce aromatic hydrocarbon emissions by ca. 50 % of the 1981 level by the year 2000.

France. Guideline to reduce emissions of volatile organic compounds (VOC) by ca. 30 % by the year 2000.

Austria. Anti-emission law to reduce the use of solvents, particularly aromatic hydrocarbons [134].

European legislation will be standardized by EC directives that are currently being drafted. The draft guideline on Integrated Pollution Prevention and Control (IPPC) restricts and controls emissions produced during the erection of new technical plants and the expansion of existing plants. The draft of a Directive to Limit Emissions of Organic Solvents entails a "solvent management plan." Its aim is the stepwise replacement of ecologically and industrial-hygienically "critical" by less critical solvents or by solvents with a low ozone-formation potential (POCP = photochemical ozone creation potential) [135]. In the final stage the use of solvents will be restricted and, where technically possible, solvent-containing coating systems will be replaced by environmentally friendly alternatives such as waterborne paints [136–140]. The POCP values of solvents are listed below:

Perchloroethylene	1	Toluene	55
Dichloromethane	1	Butanol	55
Methyl acetate	3	*sec*-Butanol	55
Trichloroethylene	7	Ethyl glycol acetate	60
Methanol	10	Butyl glycol acetate	60
Isopropyl alcohol	15	Diethyl ether	60
Diacetone alcohol	20	Ethoxypropyl acetate	65
Acetone	20	Tetrahydrofuran	70
Ethyl acetate	20	Ethyl glycol	75
Ethanol	25	Butyl glycol	75
Cyclohexane	25	Isophorone	80
Methoxypropyl acetate	30	Methoxypropanol	80
Isobutyl acetate	35	Diisobutyl ketone	80
Methylcyclohexane	35	Ethoxypropanol	85
Isobutanol	40	*p*-Xylene	90
Methyl ethyl ketone	40	1,3-Diethylbenzene	95
Hexane	40	Tetramethylbenzene	110
Butyl acetate	45	Trimethylbenzene	120
Propanol	45	Diethylmethylbenzene	120

In the TA Luft organic substances (including solvents) are subdivided into three classes. The following concentration limits must not be exceeded in emissions:

Class I.	Substances with a mass flow ≥ 0.1 kg/h: 20 mg/m^3
Class II.	*Substances with a mass flow ≥ 2 kg/h: 100 mg/m^3*
Class III.	Substances with a mass flow ≥ 3 kg/h: 150 mg/m^3

If organic compounds of several classes are present, the mass concentration in the waste gas must not exceed 150 mg/m^3 with a total mass flow ≥ 3 kg/h.

The classes contain the following solvents:

Class I. Aniline, 1,2-dichloroethane, 1,4-dioxane, nitrobenzene, phenol, 1,1,2,2-tetrachloroethane, tetrachloromethane, trichloromethane, furfural.

Class II. Butyl glycol, chlorobenzene, cyclohexanone, dimethylformamide, diisobutyl ketone, ethyl glycol, ethylbenzene, furfuryl alcohol, isopropylbenzene, carbon disulfide, methyl glycol, methyl acetate, methyl cyclohexanones, methyl formate, styrene, tetrachloroethylene, tetrahydrofuran, toluene, 1,1,1-trichloroethane, trimethylbenzenes, xylenes.

Class III. Acetone, alkyl alcohols, methyl ethyl ketone, butyl acetate, dibutyl ether, dichloromethane, diethyl ether, diisopropyl ether, dimethyl ether, ethyl acetate, ethylene glycol, diacetone alcohol, methyl isobutyl ketone, *N*-methylpyrrolidone, paraffin hydrocarbons (except methane), pinenes.

Organic substances not listed above should be allocated to the classes containing substances that they most closely resemble as regards their environmental effects. The toxicity, degradability, degradation products, and their odor intensity should be borne in mind.

The degradation of solvents by microorganisms in water, wastewater, effluent, and in clarification plants differs. In Germany solvents have been therefore classified as non, weakly, slightly, moderately, or strongly water-polluting [141]. In general, the discharge of solvents into water and waterways should be avoided to prevent contamination of rivers and ground water. The majority of solvents in wastewater and treatment plants are degradable if handled properly [142]. Detailed investigations have been carried out to assess the toxicity of solvents toward fish, which have been embodied OECD test standards [143], [144].

International measures have been adopted to restrict the use of products (mainly chlorofluorocarbons, CFCs) responsible for the decomposition of the ozone layer in the stratosphere. EC regulation No. 584/91 of March 4, 1991 regulates the import, export, production, and use of chlorinated hydrocarbons, brominated compounds, tetrachloromethane, and 1,1,1-trichloroethane. The production of these solvents is to be phased out in stages by the year 2000 [145].

5.2. Laws Concerning Dangerous Substances

In order to harmonize the legal and executive directives of the member states, the Council of the European Community has issued directives on the "Classification, Packaging and Labelling of Dangerous Substances" (67/548/EC, Basic Directive, Official Journal of the European Communities No. 196, p. 1, ff, 1967) together with numerous adaptations to take account of technical progress. The Council Directive of June 7, 1988 on the approximation of the laws of the member states relating to classification, packaging and labelling of dangerous preparations" applies to the identification of preparations (mixtures of substances) (88/379/EC, Preparations Directive, Official Journal of the European Communities No. L 187, pp. 14 – 30, 1988). It replaces the Solvent Directive and the Paint Directive (73/173/EC and 77/728/EC). Dangerous substances and solvents are listed and classified as harmful, toxic, irritant, corrosive, etc. Danger symbols are specified for each substance and information is given concerning special risks (R phrases) and safety recommendations (S phrases). The list of substances is constantly being revised and expanded [146], [147].

The EC directives and the wording of the dangerous substances directive are being continually updated. In order to assess the potential danger of a solvent the current wording of the directive should therefore be consulted.

On the basis of the EC directives the Verordnung über gefährliche Stoffe, GefStoffV (Directive on Dangerous Substances) of August 26, 1986 was issued in Germany (Bundesgesetzblatt I, p. 1470) in the wording of June 5, 1991 (Bundesgesetzblatt I, p. 1 – 1218).

Classification criteria such as danger of explosion, fire promoting, highly flammable, readily flammable, flammable, highly toxic, toxic, harmful, corrosive, irritant, carcinogenic, teratogenic, and mutagenic are defined in this directive. Instructions are given concerning the packaging and labelling of substances and preparations. The directive also contains instructions on the handling of dangerous substances (protective measures, prohibition of use, official regulations) and on health monitoring (preventive medicine, health data file). Individual paragraphs deal with legislation for the protection of young workers, pregnant women, and nursing mothers, and regulations governing trades, businesses, and working hours. Appendices I – VI include regulations concerning classification and identification; directives covering carcinogenic substances, aliphatic chlorinated hydrocarbons, lead and antifouling paints; provisions on working in rooms and containers; prescribed medical examinations; and a list of substances together with legally binding information on their labelling.

The German Chemikaliengesetz, ChemG (Chemicals Law) was passed on September 18, 1980 (Bundesgesetzblatt I, p. 1718) and deals with protection against dangerous substances. Newly developed substances must be registered with the authorities and subjected to toxicological testing. Similar regulations exist in other countries. Toxicological and ecotoxic data are also being compiled for chemicals that already exist and are listed in the European Inventory of Existing Chemical Substances (EINECS).

The Toxic Substances Control Act (TSCA) in the United States has been effective since 1977 and is similar to the German Chemicals Law. The TSCA specifies that chemicals must be assessed as regards their risks and registered before production and use.

The Technische Regeln für Gefahrstoffe, TRGS (Technical Regulations for Dangerous Substances) specify safety, industrial medicine, hygiene, and ergonomic requirements for introducing and handling dangerous substances.

The following regulations are of particular importance for solvents: recommended concentration (TRK) for dangerous substances (TRGS 102, limits e.g., benzene, 2-nitropropane); TRGS 150 covering direct skin contact with dangerous substances; TRGS 402 for the

determination and evaluation of the concentrations of dangerous substances in the atmosphere of working areas; TRGS 403 for the measurement of mixtures of substances in the workplace atmosphere; and regulations covering restrictions on use, alternative substances, alternative processes (e.g., TRgA 502 for carbon tetrachloride, tetrachloroethane, pentachloroethane; TRGS 609 for methyl and ethyl glycols and their acetates), the MAK values and biological workplace tolerance values (TRGS 403, 900).

5.3. Fire Hazard

Most organic solvents are readily volatile and combustible, and their vapors form explosive mixtures with air. The fire hazard of solvents depends on their volatility and flash point (Section 3.7).

In Germany, the Verordnung über brennbare Flüssigkeiten, VbF (Directive on Combustible Liquids) of June 5, 1970 divides solvents into danger classes:

Class A I:	water-insoluble, flash point $< 21\,^{\circ}C$
Class A II:	water-insoluble, flash point $21 - 55\,^{\circ}C$
Class A III:	water-insoluble, flash point $55 - 100\,^{\circ}C$
Class B:	flash point $< 21\,^{\circ}C$ in the case of liquids that are miscible in all proportions with water at $15\,^{\circ}C$

In the United States substances are subdivided according to their flash points:

1) Flammable	
Class I:	flash point $< 100\,^{\circ}F$ ($38\,^{\circ}C$)
Class IA:	flash point $< 73\,^{\circ}F$ ($22.8\,^{\circ}C$), $bp < 100\,^{\circ}F$ ($38\,^{\circ}C$)
Class IB:	flash point $< 73\,^{\circ}C$ ($22.8\,^{\circ}C$), $bp > 100\,^{\circ}F$ ($38\,^{\circ}C$)
Class IC:	flash point $> 73\,^{\circ}F$, $bp < 100^{\circ}F$
2) Combustible, flash point $> 100\,^{\circ}F$	
Class II:	flash point $> 100\,^{\circ}F$ and $< 140^{\circ}F$ ($60\,^{\circ}C$)
Class III:	flash point $> 140\,^{\circ}F$ ($60\,^{\circ}C$)
Class III A:	flash point $> 140\,^{\circ}F$ and $< 200\,^{\circ}F$ ($93.3\,^{\circ}C$)
Class III B:	flash point $> 200\,^{\circ}F$ ($93.3\,^{\circ}C$)

Safety regulations concerning storage and transport are specified for each danger class.

Combustible liquids are also classified according to their flash point in other national and international regulations, e.g., ADN, ADNR, ARD, GGVS, IATA, IMDG, IMO, and RID.

The following regulations must be observed when combustible solvents are stored or handled in closed spaces:

1) Prohibition of naked flames
2) Prohibition of smoking
3) Provision of adequate ventilation
4) Provision of fire-extinguishing equipment
5) Protection against spark formation due to static charges, percussion, or impact
6) Provisions covering the installation of electrical plant and equipment in explosion-harzard workshops (e.g., VDE 0165)
7) Provisions covering explosion-proof electrical equipment and material (e.g., Part 1 – VDE 0171)

Storage and filling vessels must be earthed and connected to one another so as to ensure electrical conduction during solvent transfer to avoid spark formation due to electrostatic charges. The ignition groups of the solvents should be taken into account when installing and wiring electrical fittings and equipment.

5.4. Waste

Recycling and Waste-Gas Purification. Solvent recycling is becoming increasingly important for environmental reasons. The following methods may be employed to recover a solvent from a solvent vapor–air mixture but are still uneconomical due to their high costs [148–159]:

1) Condensation on cold surfaces
2) Adsorption on a solid adsorbent, e.g., activated charcoal
3) Absorption in high-boiling liquids
4) Membrane processes [160], [161]

Removal of solvent from a solvent vapor–air mixture is nowadays even more economically efficient than recovery methods, and can be effected by [162–165]:

1) Thermal incineration
2) Catalytic incineration
3) Biofilters (biosolve method)
4) Biowashers
5) Membrane reactors

On account of the rapid poisoning of the catalyst in the catalytic incineration method, the thermal incineration has been most widely employed. The combustion heat is used, for example, to preheat the air in paint shop dryers [166], [167].

The recycling of paint coagulates and slurries has been intensively investigated [168], [169].

Waste Disposal [170], [171]. Solvent residues, solvent waste, and solvent-containing preparations should be stored only in air- and liquid-tight containers in special dumps.

Recent directives and legislation require that solvent-containing waste be properly burnt in special incinerators. Numerous directives are concerned with waste prevention and elimination. In the United States, the observance of these provisions is monitored by the EPA on the basis of the Resource Conservation and Recovery Act (RCRA) of 1976, which defines the term waste. Legislation has been strengthened through the Hazardous and Solid Waste Amendments (HSWA) of 1984 which deals with the itemization, characterization, and utilization of waste; the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA, Superfund) of 1980; and the Superfund Amendments and Reauthorization Act (SARA) of 1986. The aim of this legislation is to recover and recycle the waste to the process in which it is used and thus avoid the formation of residues.

The treatment of waste in Germany is regulated by provisions (Abfallgesetz, TA Abfall, Sonderabfallbestimmungsverordnung, Abfall- und Reststoffüberwachungsverordnung) that require product manufacturers and processors to describe residues in terms of their chemical and physical properties. They also cover the transportation of waste, proof of and permission for waste disposal, prohibition of mixing wastes, and assigning appropriate treatment techniques to specific types of waste [172–175].

Waste is classified by code numbers to facilitate handling and define potential risks:

- Code No. 553 for solvent wastes, with subgroups for individual solvents
- Code No. 554 02 for solvent-containing slurries
- Code No. 555 with subgroups for paints and coating materials
- Code No. 597 03 for distillation residues

According to the TA Abfall of March 12, 1991, all solvent-containing waste must be disposed of thermally in pyrolysis units or in special waste incinerators equipped with flue-gas cleaning systems. The residues are then stored underground or in special waste dumps. Since combustion capacity is insufficient, medium- to longterm measures will have to be taken to avoid and recycle waste. Waste production has to be avoided as far as possible in the industrial and commercial sectors, instead, waste should be recycled. Any unavoidable waste should be safely disposed of (usually by combustion).

An OECD guideline on "Hazardous Waste from the Production and Use of Organic Solvents" is currently (1992) being drafted. Waste prevention, reduction, recovery, and recycling are preferred to the combustion and the biological and physicochemical treatment of unavoidable waste.

The problem of wastewater is closely coupled with the processing of solvent-containing paints, adhesives, and other preparations.

In Germany, for example, the conditions for the discharge of dangerous substances into lakes, rivers, and waterways have become more stringent due to the Wasserhaushaltsgesetz, WHG (Water Management Law) of September 23, 1986 [Bundesgesetzblatt I (1986) No. 50, p. 1529 ff.]. "Dangerous substances" are registered in a constantly updated list compiled by the Ministry of the Environment. All plants and factories that produce effluent containing dangerous substances covered by the Abwasserherkunftsverordnung, AbwHerkV (Directive Concerning the Origin of Effluents) of July 3, 1987 must satisfy the requirements of the state of the art.

The WHG lays down instructions for handling "water-hazardous" substances in § 19 g – 19 l. The substances are compiled in a catalog, published in the Ministerialblatt (ministerial bulletin) 1985 and its supplements. They are subdivided into classes (Wassergefährdungsklassen, WGK) according to their water-hazard potential:

WGK 0:	Generally not water hazardous, e.g., ethanol, acetone
WGK 1:	Slightly water hazardous, e.g., aliphatic compounds, ethylbenzene, alcohols, ketones, ethers, esters
WGK 2:	Water hazardous, e.g., aromatic compounds, dichloromethane, carbon disulfide
WGK 3:	Highly water hazardous, e.g., tetrachloroethylene

The water hazard class of mixtures is determined according to data obtained from the mixture. If such data do not exist, the component of the highest water hazard class is decisive. Paints are classed in WGK 2 on account of their solvent contents.

6. Purification and Analysis

Purification. Distillation is most commonly used for purifying solvents. Solvents with different vapor pressures can be separated from one another by fractional distillation. Azeotropic mixtures can be separated by extractive or azeotropic distillation (e.g., addition of benzene to a water – ethanol mixture), by chemical reaction of a component (e.g., addition of acetic anhydride to an ethanol – ethyl acetate mixture), or by altering the pressure during distillation.

Further methods for purifying solvents include freezing out water from solvents that are partially miscible with water, extracting water-soluble constituents from water-immiscible solvents by shaking with water, and the use of adsorbents (e.g., activated charcoal).

Analysis. Solvent purity is assessed by means of gas chromatography [176–182], physical properties, water content, evaporation residue, and acid, saponification, and hydroxyl numbers [183]. Color and smell are also evaluated.

Standardized tests are employed in the analysis of solvents (see Table 20).

Further special analytical methods are discussed in [184–187].

7. Uses

7.1. Solvents in Paints

See also → Paints and Coatings.

Composition of Paint-Solvent Mixtures. The composition of a paint-solvent mixture is governed by the application conditions, drying temperature, and drying time of the paint. The solvent mixture in a paint that undergoes physical drying at room temperature contains ca. 45 % low boilers, ca. 45 % medium boilers, and ca. 10 % high boilers.

True solvents and latent solvents are present in such a ratio that the paint dries to give a clear film without haze. Low boilers accelerate drying, whereas medium boilers and high boilers are used to produce a flawless surface. Oven-drying paints, stoving enamels, and coil coatings are applied at relatively high temperature and contain a large amount of high boilers and only a small amount of readily volatile solvents, if at all, because they may cause the paint to "boil" during stoving.

The nature of the solvents in the mixture also depends on the type of binder. In order to obtain rapid drying combined with low solvent retention, the solvent mixture should be formulated so that its solubility and hydrogen-bonding parameters lie at the boundary of the binder solubility range. On the other hand, the parameters of the solvent mixture should be similar to those of the binder to ensure satisfactory flow. Finding a sensible compromise is difficult and requires a great deal of experimental effort. It is advantageous if the non-solvents which, according to the solubility parameter concept, accelerate drying, are more volatile than the solvents, which remain behind and improve the flow [33].

The volatility and dissolution properties of the solvent mixture should be adjusted [188] so that its parameters move from the solubility boundary to the solubility center of the binder during evaporation. It must, however, be borne in mind that the solids concentration increases during evaporation of the solvent and that the paint temperature increases or decreases, thus altering the solubility range of the binders. Numerous studies have been published on the evaporation of solvent mixtures from paint films [189].

Paint Viscosity. The viscosity of a paint depends on:

1) The nature of the binder
2) The solvent composition
3) The binder concentration
4) The pigment content
5) The temperature

In homologous solvent series the paint viscosity generally increases with increasing molecular mass of the solvent. Since the solvency power of a solvent decreases with increasing molecular mass, a relationship between solvency power and paint vicosity seems likely.

Comparison of solvents from different homologous series shows, however, that the viscosity is not generally dependent on the solvency power [190], [191]. The viscosity of a binder solution is determined by diverse binder – solvent interactions [192–196]. It is also influenced by the intrinsic solvent viscosity, the degree of uncoiling of the binder, the molecular mass of the binder, hydrogen bonding between binder and solvent molecules, as well as solvation and hydrogen bonding between binder molecules and between solvent molecules. Accordingly, the viscosity of a paint is also not generally at its lowest when the solubility parameter values of the solvent mixture coincide with those at the center of a binder-solubility region [33]. A binder is most strongly uncoiled (i.e., has its maximum volume) in the center of its solubility region. Solvents that have the solubility parameters of the binder solubility center therefore often form paints with a particularly high viscosity [197], [198]. In order to obtain low-viscosity paints, solvent additives are used to displace the solubility parameters of the solvent mixture toward the boundary of the binder solubility region; however, turbidity or a sharp increase in viscosity may be expected if this boundary is exceeded. The viscosity of paints that contain alcohol solvents or binders with hydroxyl groups can be reduced by adding small amounts of non-solvents (e.g., white spirit); on dilution viscosity-increasing hydrogen bonds are apparently ruptured. Relationships between viscosity and the hydrogen bond parameters of binders and solvents are described in [199].

The viscosity η of a paint is reduced on raising the temperature according to the Arrhenius equation,

$$\frac{1}{\eta} = A\mathrm{e}^{-E/RT}$$

in which A and E are material constants [200]. Resultant sagging phenomena of the paint from vertical surfaces must be prevented by using solvents that form hydrogen bonds and by making the paint thixotropic.

Paints have widely differing viscosities that depend on their application conditions; low-viscosity paints are processed by dipping and spraying methods, while high-viscosity paints are used in casting, rolling, and hot-spraying methods. The correct choice of solvents serves to optimize the paint properties [201], [202].

Solvents in High-Solids and Waterborne Paints. In high-solids (low-solvent) paints, small amounts of auxiliary solvents are used to reduce the viscosity, as well as to optimize degassing and the flow properties [203]. Butyl acetate and butanol are mainly used to reduce the viscosity; in combination with glycol ethers and glycol ether acetates these solvents also improve flow properties and degassing. The viscosities (flow time in seconds, DIN-4 cup) of pigmented, high-solids paints after addition of 10 % of a solvent are listed below [204]:

Butyl acetate	78
Butanol	80
Ethyl glycol	86
Xylene	92
Ethyl glycol acetate	103
Butyl glycol	119
SOLVESSO 100	126
BUTOXYL	133
Ethylbutanol	138
Cyclohexanone	141
Butyl glycol acetate	148
Diacetone alcohol	169

Solvents for high-solids paints [205–207] cannot be selected according to the solubility parameter concept with the necessary degree of certainty. This is because the high-solids binders have a low mean molecular mass and are thus soluble in virtually all solvents with the exception of white spirit. Consequently, a boundary cannot be specified for the binder solubility region in the solubility parameter–hydrogen bonding parameter diagram and the influence of the solvents on the binder–solubility interactions cannot be estimated with sufficient accuracy [208].

In general it may be said that solvents of low intrinsic viscosity strongly reduce the viscosity of high-solids paints [209]. High-boiling solvents with a high solvency must also be used to obtain good flow properties [210].

Waterborne paints contain auxiliary solvents as solubilizers in amounts of 2 – 15 %, depending on the binder. These solubilizers are water-miscible solvents or solvents that become water-miscible in all proportions in the presence of the binder [211], [212]. The most important are listed below:

Glycol Ethers [213], [214]. Ethyl glycol, isopropyl glycol, propyl glycol, butyl glycol, isobutyl glycol, ethyl diglycol, butyl diglycol, 1-methoxy-2-propanol, 1-ethoxy-2-propanol, 1-isopropoxy-2-propanol, 1-propoxy-2-propanol, 1-butoxy-2-propanol.
Alcohols [215], [216]. Ethanol, propanol, isopropyl alcohol, butanol, isobutanol, *sec*-butanol, *tert*-butanol.

Butanol by itself is not miscible in all proportions with water, but its water miscibility is unlimited in the presence of paint binders. Butanol is an extremely effective solvent in waterborne paints, although it has the disadvantage of a somewhat more pungent smell than glycol ethers. The auxiliary solvents in waterborne paints promote solubilization of the binder and water, reduce the viscosity maximum that occurs on dilution with water, and yield smooth-flowing, flawless paint surfaces [217–227].

Auxiliary solvents and film-forming auxiliaries are also used as flow promoting agents in aqueous dispersions (emulsion paints). Propylene glycol acts not only as a solvent, but is hygroscopic and thus ensures a sufficiently high water content in the coating until a smooth surface has formed.

Pigment Wetting. The wetting of pigments is influenced both by the binder and the solvent. Good pigment wetting is important for pigment grinding in the finished paint product. The solubility parameters and, more commonly, hydrogen bond parameters of the binders and solvents influence pigment wetting [197], [198], [228], with the result that good or bad dispersibility, flocculation, and leafing effects may occur [33].

Blushing, Gloss, and Flow Properties. When the solvents evaporate from a paint the latter cools. At high atmospheric humidity water droplets condense if the temperature of the paint surface is below the dew point. This water is absorbed and homogeneously distributed in paints that contains solvents that are able to absorb water (e.g., ethanol or glycol ether). If the paint does not contain such solvents the water remains on the surface as a visible white haze (blushing). Blushing disappears if the paint contains solvents that form volatile azeotropes with water (e.g., aromatic hydrocarbons or butanol).

The gloss of paint coats is significantly enhanced if true solvents (high boilers) are the last solvents to evaporate in the paint formulation. Glycol ethers, in particular, improve the gloss due to their effect on the flow properties of the paints. A paint should dry to form a smooth, flat film without any surface structure as a result of the coalescence of the paint particles. Unsatisfactory paint flow leads to surface defects known as the orange-peel effect, honeycomb structure, and fish eyes. These defects may be attributed to physical factors (e.g., to a change in the surface tension of the paint during solvent evaporation) and associated eddy formation in the paint film [192], [229–232]. Rapid evaporation occurs preferentially at the film surface causing the surface tension to increase more markedly than in the interior of the film. The resultant eddies in the film must be prevented by using slowly evaporating solvents with a good solvency for the binder. Additives that reduce the surface tension, e.g., wetting agents or silicone oils, also have a beneficial effect.

Mechanical Properties and Solvent Retention. Solvents significantly influence the mechanical properties of a paint for the following reasons:

1) Solvents influence the molecular structure of the film by aligning the binder molecules or preventing ordering
2) Solvents react to some extent with reactive, multicomponent catalyzed lacquers containing a binder component and thus exert an internal flexibilizing effect
3) Solvents are retained by the paint film and exert an external flexibilizing effect

Glycol ethers have a flexibilizing action on saturated polyester resins that cross-link with melamine – formaldehyde resins; some functional groups in the melamine resin are apparently blocked by the glycol ether. The flexibilization of paint films by retention of solvents occurs in copolymer paints and has been investigated by gas chromatography [233] and radiotracer methods [234]. Dry, hardened paint films absorb solvents to varying degrees [235], [236]. Dichloromethane swells almost all paints films which is the reason for its good paint-removing action. The swelling action of aromatic hydrocarbons on polyester stoving enamels cross-linked with

melamine resin depends on the monomer composition of the saturated polyester resin [63] and the degree of cross-linking after stoving [237].

Thick-Coat Paints and Multicoat Paints. Paints for anticorrosion systems that are applied by airless methods are prepared on the basis of, for example, vinyl chloride copolymers. Thick, high-solids coats can be obtained in a few stages with short drying times between coats. The solvents should deaerate the paint quickly and produce rapid drying and good flow. These properties can only be obtained by accurately coordinating the solvent components. In the case of multicoat paints the solvent composition of the top coat must be selected such that the primer is not swollen by inwardly migrating solvents. Mild true solvents (e.g., alcohols and glycol ethers), diluents, and nonsolvents should be used for this.

7.2. Solvents in Paint Removers

Solvents used for paint removal are able to dissolve or considerably swell physically drying binders (e.g., vinyl chloride copolymers, cellulose nitrate, polyacrylates) and chemically cross-linked coatings (e.g., oil-based paints, dried alkyd resins, cross-linked polyester – melamine resins, cross-linked epoxy and isocyanate coatings) [238]. A combination of dichloromethane with low-boiling ketones or esters is particularly suitable. Small amounts of high-boiling solvents with a low volatility (e.g., tetrahydronaphthalene, solvent naphtha, methyl benzyl alcohol, or benzyl alcohol) are added to these mixtures to retard evaporation and increase the solvency. Modern paint removers do not contain chlorinated hydrocarbons, they are formulated on the basis of high boilers (e.g., dimethylformamide, dimethyl sulfoxide, propylene carbonate, and *N*-methylpyrrolidone) in combination with alcohols and aromatics, or consist of aqueous, frequently alkaline or acidic systems. See also, → Paints and Coatings, Chap. 6.

7.3. Solvents in Printing Inks

Solvents in printing inks (→ Printing Inks) must readily dissolve the binder or resin and must not attack the printing rollers [239]. Acidic or sulfur-containing solvents must not be used with copper rollers. Aliphatic and aromatic hydrocarbons that attack rubber are unsuitable for offset inks. Alcohols, glycol ethers, and glycol ether esters are most suitable for gravure and flexographic printing.

The volatility of the solvents must be adapted to the application process, the drying process, and the drying time during printing. Solvents are rarely used in offset, letterpress, and typograhical printing, whereas rapidly evaporating solvents are employed in flexographic and gravure printing.

7.4. Extraction [240–242]

The distribution of dissolved substances in two solvent phases is employed on a large scale in the industrial separation of mixtures of substances (→ Liquid – Liquid Extraction; → Liquid – Solid Extraction). Examples are the removal of unsaturated constituents from vegetable oils with furfural or methanol, the purification of animal and vegetable oils with liquid propane, and the removal of waxes from lubricants with liquid propane or ketones. Penicillin is similarly concentrated with methyl isobutyl ketone, and aqueous glycerol is purified with xylene. Preparative and analytical separations are also performed by liquid – liquid extraction. Inorganic salts can be extracted from aqueous solutions with suitable solvents, such as ethers, ketones, and esters. This method is particularly efficient for metal halides and nitrates, e.g., the separation of uranium compounds from aqueous solutions or the fractional extraction of rare earths.

7.5. Extractive Distillation [243], [244]

In extractive distillation a solvent is added to the mixture to be separated, its boiling point is higher than that of the components of the mixture. In the case of a binary mixture, the added solvent must interact more strongly with one of the components to lower its volatility. The other more volatile component can thus be distilled off, leaving the added solvent and the higher-boiling component at the bottom of the column. The added solvent (entrainer) must be miscible

with the mixture at all temperatures, concentrations, and pressures. See also → Distillation and Rectification, Chap. 6.5.

A mixture of cyclohexane (*bp* 80.8 °C) and benzene (*bp* 80.1 °C) can, for example, be separated by distillation after adding aniline, since the interaction between benzene and aniline is greater than that between cyclohexane and aniline. Azeotropic mixtures can be separated similarly by extractive distillation (e.g., water – ethanol by adding glycerol). Hydrocarbons with similar boiling points can be separated by extractive distillation in the presence of polar liquids (nitrobenzene, phenol, furfural).

7.6. Chromatography

Chromatography of a mixture on a stationary phase using a liquid eluent is a form of solid – liquid extraction (adsorption chromatography). The distribution of components between the surface of the carrier material and the solvent depends on their chemical structure and the nature of the carrier material and the migrating solvent [11], [245–247]. The solvents used in adsorption chromatography are arranged in an eluotropic series, the eluting action of the solvents increases with increasing polarity as follows: pentane, petroleum ether, hexane, heptane, cyclohexane, tetrachloromethane, trichloroethylene, benzene, dichloromethane, trichloromethane, diethyl ether, ethyl acetate, pyridine, acetone, propanol, ethanol, methanol, and water.

Liquid – liquid extraction is also possible in which distribution takes place between a liquid eluent and a liquid (generally water) that is adsorbed on the surface of the carrier material (distribution chromatography).

7.7. Solvents for Chemical Reactions

Solvents for chemical reactions must be chemically inert under the reaction conditions. The following solvents have proved suitable [241]:

1) *Hydrogenation*: alcohols, glacial acetic acid, hydrocarbons, dioxane
2) *Oxidation*: glacial acetic acid, pyridine, nitrobenzene
3) *Halogenation*: tetrachloromethane, tetrachloroethane, dichlorobenzene, trichlorobenzene, nitrobenzene, glacial acetic acid
4) *Esterification*: benzene, toluene, xylene, dibutyl ether
5) *Nitration*: glacial acetic acid, dichlorobenzene, nitrobenzene
6) *Diazotization*: ethanol, glacial acetic acid, benzene, dimethylformamide
7) *Coupling reactions of diazonium compounds*: methanol, ethanol, glacial acetic acid, pyridine
8) *Grignard reaction*: diethyl ether
9) *Friedel – Crafts reaction*: nitrobenzene, benzene, carbon disulfide, tetrachloromethane, tetrachloroethane, 1,2-dichloroethane
10) *Dehydration*: benzene, toluene, xylene
11) *Sulfonation*: nitrobenzene, dioxane
12) *Dehydrohalogenation*: quinoline
13) *Decarboxylation*: quinoline
14) *Acetal formation*: hexane, benzene
15) *Ketene condensation*: diethyl ether, acetone, benzene, xylene

The rate and course of the reaction often depend on the nature of the solvent [3], [248], [249]. Solvent effects have also been described in radical polymerizations [250].

7.8. Solvents for Recrystallization

Solvents for recrystallization should readily dissolve the substance to be purified as well as the impurities on heating; on cooling the solution the pure substance then crystallizes out, whereas the impurities remain in solution. If the temperature dependence of the solubility of the substance in the solvent is only slight, it can be precipitated by adding a non-solvent to the hot solution. The purification effect of recrystallization is greater the more slowly the solution is cooled and accordingly the more slowly the crystals form.

7.9. Solvents in Film Production

Diethyl ether – ethanol mixtures are used as solvents in the production of cellulose nitrate films. Acetone is most suitable for cellulose acetate or cellulose acetate butyrate films, and dichloromethane for cellulose triacetate films.

7.10. Solvents for Synthetic Fibers

Solvents used for the preparation of spinning solutions (dopes, → Fibers, 3. General Production Technology, Chap. 2.1.4.) should be inexpensive, easily recoverable, volatile, have as low a viscosity as possible, and should readily dissolve the high molecular mass polymers to be spun. The following solvents have proved suitable: dichloromethane – methanol for cellulose triacetate, tetrahydrofuran or carbon disulfide – acetone for non-postchlorinated poly(vinyl chloride), acetone for postchlorinated poly(vinyl chloride), and dimethyl sulfoxide or propylene carbonate for polyacrylonitrile.

7.11. Solvents for Rubber, Plastics, and Resin Solutions

Aliphatic hydrocarbons, in some cases mixed with aromatic hydrocarbons, are suitable for rubber solutions.

Plastics and resin solutions usually contain highly or moderately volatile solvents. The solutions are used for example to coat articles with a plastic, as adhesive solutions [251], as nail polish, and for textile impregnation. Esters, ketones, alcohols, or cyclic ethers are suitable depending on the intended use.

Table 14. Development of consumption of solvents in Western Europe (as percentage of total)

	1980	1986	1990	1995	2000
Oxygen-containing solvents	36.5	45	51	58	65
Aliphatics	28.5	22	20.5	19	18
Aromatics	20.5	20	19	17	15.5
Chlorinated hydrocarbons	14.5	13	9.5	6	1.5
Total consumption 10^6 t	5.1	4.75	4.7	4.15	3.3

7.12. Solvents for Degreasing

Ketones or chlorinated hydrocarbons, especially trichloroethylene, perchloroethylene, or dichloromethane, are used to degrease metals [252], [253]. The chlorinated hydrocarbons used to degrease light metals are stabilized to prevent the metal-catalyzed elimination of hydrogen chloride. The solvent mixture [254–257] for metal degreasing are being replaced by aqueous systems [258], [259] due to more stringent environmental protection requirements.

Table 15. Solubility parameters δ, dipole moments μ $(J/cm^3)^{1/2}$ and hydrogen bond parameters γ

Solvent	δ	δ_D	δ_P	δ_H	μ	γ
n-Hexane	14.9	14.7	0	0	0	0
n-Heptane	15.1	15.1	0	0	0	0
Cyclohexane	16.8	16.8	0	0	0	0
Toluene	18.2	18.0	1.4	2.0	0.4	4.5
Xylene	18.0	17.8	1.0	3.1	0.4	4.5
Ethyl benzene	18.0	17.8	0.6	1.4	0.6	1.5
Cumene	17.1	17.1	0.4	0		
Styrene	19.0	18.6	1.0	4.1	0	1.5
Dichloromethane	19.8	18.2	6.1	6.3	1.2	1.5
1,2-Dichloroethane	20.1	18.8	5.3	4.1	1.8	1.5
Methanol	29.7	15.1	11.3	22.9	1.7	18.7
Ethanol	26.0	15.1	8.0	20.1	1.7	18.7
1-Propanol	24.3	15.1	6.1	17.6	1.7	18.7
2-Propanol	23.5	15.3	6.1	17.2	1.7	18.7
1-Butanol	23.3	16.0	6.1	15.8	1.8	18.7
Isobutanol	21.9	15.3	5.7	15.8	1.8	17.9
2-Ethylhexanol	19.4	16.0	3.3	11.9	1.7	18.7
Cyclohexanol	23.3	17.4	4.1	13.5	1.7	18.7
Diacetone alcohol	18.8	15.8	8.2	10.8	3.2	13.0
Acetone	20.5	15.6	11.7	4.1	2.7	9.7
Methyl ethyl ketone	19.0	16.0	9.0	5.1	2.8	7.7
Methyl isobutyl ketone	17.2	15.3	6.1	4.1	2.7	7.7
Diisobutyl ketone	16.0	16.0	3.7	4.1	2.7	8.4
Cyclohexanone	19.8	17.8	7.0	7.0	3.0	11.7
Mesityl oxide	18.4	16.4	7.2	6.1	2.7	9.8
Isophorone	19.2	16.6	8.2	7.4		14.9
Methyl acetate	18.0	14.5	5.8	9.1	1.6	8.4
Ethyl acetate	18.6	15.1	5.3	9.2	1.7	8.4
Butyl acetate	17.4	15.8	3.7	6.3	1.8	8.8
Isobutyl acetate	17.0	15.1	3.7	7.6	1.9	8.8
Ethyl glycol acetate	17.8	16.0	4.7	10.6	2.3	9.4
Butyl glycol acetate	17.9	15.8	3.7	7.7		10.3
Ethyl diglycol acetate	19.0	16.0	5.4	8.7		
Butyl diglycol acetate	18.6	16.2	4.6	7.9		
Methoxypropyl acetate	18.1	15.3	4.6	8.5		
Ethoxypropyl acetate	17.9	15.4	4.1	8.1		
Ethylene carbonate	30.1	19.4	21.7	5.1	4.9	4.9
Propylene carbonate	27.2	20.1	18.0	4.1	5.0	4.9
Diethyl ether	15.6	14.3	5.1	2.0	1.1	13.0
Methyl *tert*-butyl ether	15.6	14.4	3.4	5.0		
Tetrahydrofuran	19.2	16.8	6.8	7.2		12.0
Dioxane	20.3	17.6	8.6	4.1	0.4	9.7
Ethyl glycol	20.3	16.2	9.2	14.3	2.1	13.0
Propyl glycol	22.4	16.2	5.7	14.4		
Butyl glycol	18.2	16.0	6.3	12.1	2.1	13.0
Ethyl diglycol	19.6	16.2	7.6	12.3	1.6	13.0
Butyl diglycol	18.2	16.0	7.0	10.6	1.6	13.0
1-Methoxy-2-propanol	19.8	15.6	8.2	14.3		
1-Ethoxy-2-propanol	22.1	15.8	5.7	14.3		
1-Isopropoxy-2-propanol	21.1	15.6	5.0	13.4		
1-Isobutoxy-2-propanol	20.6	15.7	4.4	12.6		
Methyl dipropylene glycol	18.1	16.4	8.0	11.6		
Dimethylformamide	24.8	17.4	13.7	11.3	3.8	11.7
Dimethyl acetamide	22.1	16.8	11.5	10.2	2.0	12.3
Dimethyl sulfoxide	26.4	19.2	13.3	12.1	4.0	7.7
Carbon disulfide	20.5	20.5	0	0	0	0

Table 15. (Continued)

Solvent	δ	δ_D	δ_P	δ_H	μ	γ
1-Nitropropane	20.6	16.3	11.9	4.1	3.6	
2-Nitropropane	20.3	16.2	12.1	4.1	3.7	2.5
Nitrobenzene	20.5	17.6	12.3	4.1	4.3	2.8
N-Methyl-2-pyrrolidone	21.4	18.4	12.6	7.4		

7.13. Solvents for Dry Cleaning

Aliphatic hydrocarbons or perchloroethylene are used for the chemical cleaning of textiles (→ Dry Cleaning, Chap. 2.). Resins are often added to the dipping baths to hydrophobize the textiles [260].

7.14. Solvents in Aerosol Cans and Dispensers

Solvents used in aerosol cans (spray cans for hair lacquer, cleansing agents, paints) must, of course, dissolve the substances that are to be sprayed. They must also be miscible with the propellants without causing the dissolved substance to precipitate. Previously, chlorofluorohydrocarbons were mainly used as propellants but they have a damaging effect on the stratospheric ozone layer and have now been largely replaced by alternatives (e.g., hydrocarbons such as butane, diethyl ether, fluorocarbons, carbon dioxide) [261].

8. Economic Aspects [262]

Solvents are mainly used for the following purposes:

1) For cleaning and degreasing: electronic components, metal surfaces, dry cleaning of textiles
2) As a raw material: for preparing other products; in the paint, adhesives, and plastics industries; for producing resin solutions
3) As a carrier or dispersion medium: pharmaceutical products, agrochemicals

Future solvent consumption is expected to decrease on account of worldwide environmental restrictions. There will be a shift from aromatic and aliphatic hydrocarbons to oxygen-containing solvents because the latter have a lower ozone formation potential and better dissolution properties. Table 14 shows the anticipated development in consumption in Europe up to the year 2000.

9. Solvent Groups

Important information and data on individual solvents are listed in the following tables:

Table 15: Solubility parameters, dipole moments, and hydrogen bond parameters
Table 16: Miscibility with water
Table 17: Azeotropes of the most important solvents
Table 18: Relationships between solubility parameter and surface tension
Table 19: Gross calorific values
Table 20: Standards for solvents and solvent tests
Table 21: Trademarks and trade names
Table 22: Abbreviations
Table 23: Physical properties
Table 24: Safety and physiological data

Table 16. Miscibility of solvents with water (wt % at 20 °C)

Solvent	Solvent in water	Water in solvent
Hexane	0.53	0.1
Tetrahydronaphthalene		0.2
Dipentene		0.72
Toluene	0.035	0.05
p-Xylene	0.02	0.02
Ethylbenzene	0.02	0.02
Styrene		0.04
Methanol	∞	∞
Ethanol	∞	∞
Propanol	∞	∞
Isopropyl alcohol	∞	∞
Butanol	7.5	19.7
Isobutanol	8.4	16.2
sec-Butanol	12.5	44.1
tert-Butanol	∞	∞
Hexanol	0.58	7.2
Trimethylcyclohexanol	0.19	4.0
Cyclohexanol	3.6	3.6
Methylbenzyl alcohol	2.9	5.8
Ethylene glycol	∞	∞
Methyl glycol	∞	∞
Ethyl glycol	∞	∞
Propyl glycol	∞	∞
Butyl glycol	∞	∞
Ethyl diglycol	∞	∞
Methoxypropanol	∞	∞
Methyldipropylene glycol	∞	∞
Nitroethane	4.5	0.9
1-Nitropropane	1.4	0.5
2-Nitropropane	1.7	0.6
Diethyl ether	6.9	1.2
Dibutyl ether	0.3	0.2
Methyl *tert*-butyl ether	4.8	1.3
Tetrahydrofuran	∞	∞

Table 16. (Continued)

Solvent	Solvent in water	Water in solvent
Dioxane	∞	∞
Methyl acetate	24.0	8.0
Ethyl acetate	6.1	3.3
Isopropyl acetate	2.9	1.9
Butyl acetate	0.83	0.62
Isobutyl acetate	0.67	1.65
Ethyl glycol acetate	23.5	6.5
Butyl glycol acetate	1.5	1.7
Cyclohexyl acetate	0.2	0.5
Butyl glycolate	7.5	25.0
Propylene carbonate	21.4	7.5
Acetone	∞	∞
Methyl ethyl ketone	26.0	12.0
Methyl isobutyl ketone	2.0	2.4
Diisobutyl ketone	0.04	0.42
Cyclohexanone	2.3	8.0
Isophorone	1.2	4.3
Trimethylcyclohexanone	0.3	1.4
Diacetone alcohol	∞	∞
Dichloromethane	2.0	0.16
1,1,1-Trichloroethane	0.44	0.05
Trichloroethylene	0.1	0.02
Tetrachloroethylene	0.02	0.01
Dimethylformamide	∞	∞
Dimethyl sulfoxide	∞	∞

Table 17. Azeotropes of the most important solvents

Solvent A	Solvent B	Azeotrope	
		bp (101.3 kPa), °C	Composition A/B, wt %
Hexane	water	61.6	94.4/5.6
	methanol	50.0	73.1/26.9
	ethanol	58.7	79/21
	isopropyl alcohol	61.0	78/22
	isobutanol	68.3	97.5/2.5
	acetone	49.8	41/59
	methyl ethyl ketone	64.2	63/37
	ethyl acetate	65.0	58/42
Cyclohexane	water	69.8	91.5/8.5
	methanol	54.2	63/37
	propanol	74.3	80/20
	isobutanol	78.1	86/14
m-Xylene	water	94.5	60/40
Methanol	methyl acetate	54.0	19/81
	ethyl acetate	62.3	44/56
	acetone	55.7	12/78
	methyl ethyl ketone	63.5	70/30
	dichloromethane	39.2	8/92
	trichloroethylene	60.2	36/64
Ethanol	water	78.2	95.6/4.4
	hexane	58.7	21/79
	toluene	76.7	68/32
	methyl ethyl ketone	74.8	34/66
	dichloromethane	41.0	3.5/96.5
	trichloroethylene	70.9	27/73
Butanol	water	93.0	55.5/44.5
	hexane	67.0	97/3

Table 17. (Continued)

Solvent A	Solvent B	Azeotrope	
		bp (101.3 kPa), °C	Composition A/B, wt %
	cyclohexane	79.8	10/90
	toluene	105.6	27/73
	p-xylene	115.7	68/32
	isobutyl acetate	114.5	50/50
	methyl isobutyl ketone	114.35	30/70
Isobutanol	water	89.9	66.8/33.2
	hexane	68.3	97.5/2.5
	toluene	100.9	44.5/55.5
	p-xylene	107.5	83/17
	isobutyl acetate	107.6	95/5
tert-Butanol	water		88.2/11.8
2-Ethylhexanol	water	99.1	20/80
Cyclohexanol	water	97.8	20/80
	α-pinene	149.9	35.5/64.5
	m-xylene	143.0	14/86
	ethyl acetate	153.8	5/95
Ethyl glycol	water	99.4	28.8/71.2
	dibutyl ether	127.0	50/50
	butyl acetate	125.8	35.7/64.3
Butyl glycol	water	98.8	20.8/79.2
Ethyl acetate	water	70.4	91.8/8.2
	hexane	65.0	42/58
	methanol	62.1	51.4/48.6
	ethanol	71.8	69/31
	isopropyl alcohol	74.8	77/23
Butyl acetate	water	90.7	73/27
	isopropyl alcohol	80.1	48/52
	butanol	117.6	32.8/67.2
	ethyl glycol	125.8	64.3/35.7
Methyl ethyl ketone	water	73.5	89/11
	hexane	64.2	37/63
	cyclohexane	72.0	40/60
	isopropyl alcohol	77.5	68/32
	ethyl acetate	76.7	22/78
Methyl isobutyl ketone	water	87.9	75.7/24.3
Isophorone	water	99.5	16.1/83.9
Tetrahydrofuran	water	64.0	94.7/5.3
Dichloromethane	water	38.1	98.5/1.5
Trichloroethylene	water	72.9	93.4/6.6
Tetrachloroethylene	water	87.1	84.1/15.9
1,1,1-Trichloroethane	water	65.0	95.7/4.3

Table 18. Relationships between solubility parameter and surface tension

Solvent	Surface tension, mN/m	Molecular volume, *V*	Solubility parameter, calculated, $(J/cm^3)^{1/2}$	Solubility parameter, actual, $(J/cm^3)^{1/2}$
Dichloromethane	29.0	64.0	22.4	19.8
Trichloroethylene	32.0	90.9	22.1	18.8
Tetrachloroethylene	31.7	102.2	21.5	19.0
Toluene	28.5	105.5	20.2	18.2
Xylene	28.3	121.5	19.6	18.0
Ethylbenzene	29.0	121.2	19.9	18.0
Methanol	22.6	40.5	21.2	29.7
Ethanol	22.3	58.4	19.6	26.0

Table 18. (Continued)

Solvent	Surface tension, mN/m	Molecular volume, V	Solubility parameter, calculated, $(J/cm^3)^{1/2}$	Solubility parameter, actual, $(J/cm^3)^{1/2}$
Isopropyl alcohol	21.7	76.5	18.4	23.5
Butanol	24.6	91.4	19.6	23.3
Isobutanol	23.0	92.4	19.1	21.9
Acetone	23.7	73.4	19.3	20.5
Methyl ethyl ketone	24.6	89.6	19.2	19.0
Methyl isobutyl ketone	23.9	125.3	17.7	17.2
Cyclohexanone	34.5	103.9	22.5	19.8
Methyl acetate	24.8	79.5	19.7	19.6
Ethyl acetate	23.9	97.9	18.5	18.6
Butyl acetate	25.5	132.0	18.2	17.4
Isobutyl acetate	23.7	133.4	17.4	17.0
Ethyl glycol acetate	29.5	136.0	19.6	17.8
Ethyl glycol	28.2	96.9	20.4	20.3
Butyl glycol	29.0	131.2	19.6	18.2
Ethyl diglycol	31.8	135.5	20.5	19.6
Butyl diglycol	31.5	169.7	19.5	18.2
Diethyl ether	16.5	103.8	14.9	15.6
Dioxane	36.5	85.2	24.1	20.3

Table 19. Gross calorific values of solvents

Solvent	Gross calorific value, MJ/kg
Hexane	– 44.6
Heptane	– 47.7
Toluene	– 42.6
Xylene	– 43.1
Methanol	– 22.5
Ethanol	– 29.7
Butanol	– 36.0
Isobutanol	– 36.0
Methylbenzyl alcohol	– 35.9
Methyl acetate	– 22.1
Ethyl acetate	– 25.0
Isobutyl acetate	– 31.2
Butyl acetate	– 30.6
Ethyl glycol acetate	– 25.9
Propylene carbonate	– 14.3
Ethyl glycol	– 28.3
Butyl glycol	– 32.8
Ethyl diglycol	– 27.7
Butyl diglycol	– 31.7
Tetrahydrofuran	– 36.7
Methyl *tert*-butyl ether	– 38.2
Acetone	– 30.8
Methyl ethyl ketone	– 33.9
Methyl isobutyl ketone	– 37.2
Isophorone	– 38.1
Dichloromethane	– 7.1
Tetrachloroethylene	– 4.9

9.1. Aliphatic Hydrocarbons

Aliphatic hydrocarbons, naphthas, gasoline, or paraffin hydrocarbons are chemically inert and are thus very stable solvents [263], [264]. Aliphatic hydrocarbons exhibit a good solvency for mineral oils, fatty oils (with the exception of castor oil), waxes, and paraffin. They also dissolve rubber, polyisobutene, molten polyethylene, poly(butyl acrylate), poly(butyl methacrylate), and poly(vinyl ethers). However, most other polymers, polar resins, cellulose derivatives, and most paint binders are insoluble. Resins and binders with a low polarity dissolve less readily in aliphatic hydrocarbons than in aromatic hydrocarbons.

Aliphatic hydrocarbons are subdivided according to their boiling point ranges into special boiling point hydrocarbons (SBP, DIN 51 631), white spirit (DIN 51 632), and petroleum ether (DIN 51 630) (see Tables 20 and 23).

Special boiling point hydrocarbons are used for quick-drying paints, dipping solutions, and quick-drying adhesives. Their flash point is below 21 °C, which means that they must be used in explosion-proof areas. Petroleum ether is a special petroleum distillation fraction (*bp* 40 – 60 °C), and is used as a solvent in the chemical industry.

White spirit is used predominantly in the paint industry as a solvent or diluent for oil-based, alkyd resin, chlorinated rubber, and some

Table 20. Standards for solvents and solvent tests

Standard	Subject area	Standard	Subject area
DIN 51 405 DIN 51 413 ASTM E 260 ASTM D 3545	gas chromatography	DIN 53 245 ASTM D 1 152 ISO R 1 387 ISO R 1 388 ASTM D 770 ISO R 756 ASTM D 304 ISO R 755 ASTM D 1 007 ISO 2496 ASTM D 2 635 ASTM D 2 627 ISO 2517 ISO 1843	alcohols methanol ethanol isopropyl alcohol butanol *sec*-butanol methyl isobutyl carbinol diacetone alcohol higher alcohols
DIN 51 408	chlorine content, combustion	DIN 53 246 ISO R 1386 ASTM D 302 ASTM D 657 ASTM D 303 ASTM D 1 718 ASTM D 343	acetic acid esters ethyl acetate isopropyl acetate butyl acetate isobutyl acetate ethyl glycol acetate
DIN 51 423 DIN 53 491 ASTM D 1 218 ASTM D 1 747	refractive index	DIN 53 247 ASTM D 329 ISO R 757 ASTM D 740 ISO 2497 ASTM D 1 153 ISO 2499 ISO 2500	ketones acetone methyl ethyl ketone methyl isobutyl ketone ethyl isoamyl ketone
DIN 51 630	petroleum ether	DIN 53 248	turpentine oil, pine oil
DIN 51 631 ISO 1250	SBP spirits/aliphatic hydrocarbons	DIN 53 249	dipentene
DIN 51 632	light naphthas	DIN 53 401 ASTM D 464 ISO 2114 ISO 3681	saponification number
DIN 51 633 ASTM D 2 600	benzene and homologues	DIN 53 402 ASTM D 1 980 ASTM D 1 613 ASTM D 664 ISO 3682	acid number
DIN 51 755 ASTM D 56 ISO 1523	flash point AP (closed cup)	DIN 53 409 ASTM D 1 209 ISO 6271	color haze unit
DIN 51 757 ASTM D 891 ASTM D 1 298 ASTM D 4 052 ISO 1306	density	DIN 53 977	trichloroethylene
DIN 51 758 ASTM D 93 ISO 3679	flash point PM (closed cup)	DIN 53 978	tetrachloroethylene
DIN 51 777 ASTM D 1 364	water content	DIN 55 680 ISO 1516	danger class

Table 20. (Continued)

Standard	Subject area	Standard	Subject area
ASTM D 1 744 ASTM E 203 ISO 760			
DIN 51 794 ASTM D 2 155 ASTM E 659	ignition temperature	DIN 55 681	trichloroethylene, stability
DIN 51 848	precision	DIN 55 685	GC alcohols
DIN 52 900 ISO 11 014	safety data sheet	DIN 55 686 ASTM D 3 545	GC acetic acid esters
DIN 53 015 DIN 53 177 DIN 53 214 ASTM D 1 200 ASTM D 445 ISO 1342	viscosity	DIN 55 687	GC ketones
DIN 53 169	solvents, general	DIN 55 688	GC ethylene glycol ether
DIN 53 170 DIN 53 249 ASTM D 3 539	evaporation number	DIN 55 689	GC propylene glycol ether
DIN 53 171 DIN 51 751 ASTM D 1 078 ASTM E 133 ISO 4626	boiling curve/boiling-point range	DIN 55 690 ASTM D 3 128 ASTM D 331 ASTM D 330	ethylene glycol ether methyl glycol ethyl glycol butyl glycol
DIN 53 172 ASTM D 1 353 ASTM D 2 109 ISO 3251	dry residue	DIN 58 752	solvents for optics
DIN 53 173 ASTM D 2 192	carbonyl number	DIN ISO 6162 ISO 6162	iodine color number
DIN 53 174	solvent mixtures	DIN 53 241	iodine number
DIN 53 240 ASTM D 1 957 ASTM E 222 ISO 4629	hydroxyl number	DIN 13 310 ASTM D 971	interfacial surface tension
DIN 53 242	raw materials	DIN 16 513	pure toluene for gravure inks
DIN 51 376 ASTM D 92	flash point (Cleveland), open cup	DIN 51 750 ASTM D 270 ISO 1512 ISO 842	sampling

vinyl chloride copolymer paints; light napththas have flash points above 21 °C. Varnish makers' and painters' (VMP) naphtha (*bp* 100 – 150 °C) is used mainly as diluent.

9.2. Cycloaliphatic Hydrocarbons

The solvency of cycloaliphatic hydrocarbons is between that of aliphatic and aromatic hydrocarbons. They have a high solvency for fats, oils, oil-modified alkyd resins, styrene-modified oils and alkyd resins, bitumen, rubber, and other polymers. Polar resins (e.g., urea --, melamine --, and phenol – formaldehyde resins), as well as alcohol-soluble synthetic resins and cellulose esters are, however, insoluble.

Cycloaliphatic hydrocarbons are miscible with most other solvents, but are insoluble in water.

Table 21. Trade marks and trade names of solvents

Solvent	Chemical name	Manufacturer
Acintene	dipentene, pinene	DRT, France
Algofrene	fluorinated hydrocarbons	Montedison
Anol	cyclohexanol	BASF
Anon	cyclohexanone	BASF
Arcosolv	propylene glycol ethers, esters	Arco
Arizole Pine Oil, Dertol	pine oil	DRT, France
Bepetan	aliphatic hydrocarbons	BP Chemicals
Bisol TA	xylene/butanol/butyl acetate	BP Chemicals
Butoxyl	3-methoxybutyl acetate	Hoechst
Carbitol	diglycol ethers, esters	Union Carbide
Cellosolve	glycol ethers, esters	Union Carbide
Chlorothene	chlorinated hydrocarbons	DOW
Citrosol, Citrosolv	terpene hydrocarbons, limonene	Krems Chemie, Austria
Cobersol	aliphatic hydrocarbons	Cölner Benzin Raffinerie
Comsol	1-nitropropane/nitroethane	Angus Chem.
Dalpad	alcohol	DOW
DBE	dimethyl esters of succinic, adipic, and glutaric acids	DuPont
Dekalin	decahydronaphthalene	Henkel
Delifrene	fluorinated hydrocarbons	Montedison
Depanol	terpene hydrocarbons	Hoechst
Dional	fluorinated hydrocarbons	Hoechst
Dioxitol	diethylene glycol ethers, esters	Shell
Diproxitol	dipropylene glycol ethers, esters	Shell
Dowanol	glycol ethers and ether esters	DOW
Driveron	methyl *tert*-butyl ether	Hüls
Ektapro	propionic acid esters, ethers	Eastman Chemical
Ektasolve	glycol ethers, ether acetates	Eastman Chemical
Emkanol	glycol ethers	PCUK, France
Estasol	dibasic dimethyl esters	Corless Chem., UK
Exxate	acetates	Exxon Chemical
Exxsol	aliphatic hydrocarbons	Exxon Chemical
Frigen	fluorinated hydrocarbons	Hoechst
GB ester	butyl glycolate	Wacker
Genklene	chlorinated hydrocarbons	ICI
Hexalin	cyclohexanol	Henkel
Hexamethapol	hexamethylphosphoric acid triamide	
HI-SOL	aromatic hydrocarbons	Ashland Chem.
Hydrosol	aliphatic hydrocarbons	DHC Deutsche Hydrocarbures
Imnulyt	dioxolane derivative	
IMS	industrial methylated spirit (ethanol)	BP Chemicals
Isanol	butanol/isobutanol mixture	BASF
Isopar	isoparaffins	Exxon Chemical
Kristallöl	aliphatic hydrocarbons	Shell
Lactonal	ethyl lactate	Boehringer Ingelheim
Lobosol	ester–ketone mixture	BP
Solvent APV	ethyl diglycol	Hüls
Lusolvan FBH	dibutyl ester mixture	BASF
M-pyrol	*N*-methylpyrrolidone	GAF

Table 21. (Continued)

Solvent	Chemical name	Manufacturer
Nappar	naphthenes	Exxon Chemical
Ni-Par	nitropropane	Angus Chem.
Norpar	*n*-paraffins	Exxon Chemical
Oxitol	ethylene glycol ethers, esters	Shell
Polysolvan IB	isobutyl isobutyrate	Hoechst
Polysolvan O	butyl glycolate	Hoechst
Proglyde	propylene glycol diether	DOW
Proposol	propylene glycol ethers, esters	Union Carbide
Proxitol	propylene glycol ethers, esters	Shell
Purasolv	lactates	Purac Biochem BV, The Netherlands
Pyranthon A	diacetone alcohol	BASF
Rhodiasolv	dimethyl ester mixture	Rhône Poulenc Chem., France
Sangajol	aliphatic hydrocarbons	Shell
Sextol	cyclohexanol	
Sexton	cyclohexanone	
Shellsol	aliphatic/aromatic hydrocarbons	Shell
Sinarol II	aliphatic hydrocarbons	Shell
Solketal	dioxolane derivative	
Solvenon I	isobutyl formate/isobutanol mixture	BASF
Solvenon PM	methoxypropanol	BASF
Solvent naphtha hydrosol	aromatics	Deutsche Hydrocarbures
Solvesso	aromatics	Exxon Chemical
Special naphtha	aliphatic hydrocarbons	Many
Starsol	dimethyl ester of glutaric/adipic acids	Chemoxy International Ltd., UK
Sulfolan		Shell
TCD-keton	ketone	Hoechst
Tecsol	ethanol	Eastman Chemical
Terapin	aliphatic hydrocarbons	Shell
Light naphtha	aliphatic hydrocarbons	Many
Tetralin	tetrahydronaphthalene	Henkel
Texacar	ethylene/propylene carbonates	Texaco
Texanol	ester-alcohol	Eastman Chemical
Texsolve	aliphatic/aromatic hydrocarbons	Texaco
UCAR ester EEP	ethyl-3-ethoxypropionate	Union Carbide
Varsol	aliphatic hydrocarbons	Exxon Chemical

Cyclohexane [*110-82-7*] is a water-clear, colorless liquid with a gasoline-like smell; it is miscible with most organic solvents except methanol, dimethylformamide, and solvents of similar polarity (→ cyclohexane).

Methylcyclohexane [*108-87-2*] is similar to cyclohexane but less volatile (→ cyclohexane, chap. 7.2.).

Table 22. Abbreviations for solvents

Abbreviation	Name
BDG	butyl diglycol
BDGA	butyl diglycol acetate
BG	butyl glycol
BGA	butyl glycol acetate
BLO	butyrolactone
DAA	diacetone alcohol
DBE	dibasic esters
DEE	diethyl ether
DEG	diethylene glycol
DEK	diethyl ketone
DIBK	diisobutyl ketone
DMAC	dimethylacetamide
DMI	1,3-dimethyl-2-imidazolidinone
DMF	dimethylformamide
DMK	acetone
DPB	dipropylene glycol butyl ether
DPG	dipropylene glycol
DPM	dipropylene glycol methyl ether
DPMA	dipropylene glycol methyl ether acetate
EDG	ethyl diglycol
EDGA	ethyl diglycol acetate
EEP	ethyl 3-ethoxypropionate
EG	ethyl glycol
EGA	ethyl glycol acetate
IBA	isobutanol
IBIB	isobutyl isobutyrate
IMI	3-isopropyl-2-methylimidazole
IO	3-isopropyl-2-oxazolidinone
IP	isophorone
IPA	isopropyl alcohol
IPE	diisopropyl ether
IPG	isopropyl glycol
IPP	isopropoxypropanol
MBK	methyl butyl ketone
MDG	methyl diglycol
MEG	monoethylene glycol
MEK	methyl ethyl ketone
MG	methyl glycol
MGA	methyl glycol acetate
MHK	methyl hexyl ketone
MIAK	methyl isoamyl ketone
MIBK	methyl isobutyl ketone
MIPK	methyl isopropyl ketone
MOX	3-methyl-2-oxazolidinone
MPK	methyl propyl ketone
NBA	*n*-butanol
NMP	*N*-methyl-2-pyrrolidone
NPA	*n*-propanol
PB	butoxypropanol
PDG	propylene diglycol
PG	propylene glycol
PM	methoxypropanol
PMA	methoxypropyl acetate
PP	propoxypropanol
PTB	*tert*-butoxypropanol
SBA	*sec*-butanol
TPM	tripropylene glycol methyl ether

1,2,3,4-Tetrahydronaphthalene [*119-64-2*] (tetralin) is an aromatic-cycloaliphatic hydrocarbon. It is a colorless liquid with a naphthalene-like odor, insoluble in water, and miscible with all common organic solvents (→ Naphthalene and Hydronaphthalenes, Chap. 5.). It dissolves fats, oils, linoxyn, rubber, waxes, asphalt, bitumen, pitch, tar, phenol, naphthalene, iodine, sulfur, etc., and is used on a large scale in painting work, and in floor wax and shoe polish production. It also dissolves colophony, Congo copals, glyptal resins, coumarone resins, ketone–formaldehyde resins, and aminoplasts. It imparts good flow properties to paints and produces high-gloss, smooth film surfaces. It is autooxidative and thus acts as an oxygen carrier in drying oils.

Decahydronaphthalene [*91-17-8*] (decalin) is a colorless solvent with a pungent odor and fairly high volatility, its solvency is somewhat lower than that of tetrahydronaphthalene (→ Naphthalene and Hydronaphthalenes, Chap. 5.).

9.3. Terpene Hydrocarbons and Terpenoids

Turpentine oil [*8006-64-2*] (DIN 53 248). Only pure ethereal oil obtained from the distillation of the resinous secretion of living pine trees, and from which no valuable constituents (e.g., pinene) have been extracted, may be used as balsam terpentine oil (RAL, Sheet 848 C). All turpentine oils obtained in any other way must be specially labelled with details of their source. Oils of turpentine from different countries differ in composition. American and Greek oils contain predominantly D-pinene, whereas French, Spanish, and Portuguese oils contain L-pinene.

Balsam turpentine oil is a highly mobile liquid with a characteristic odor. On prolonged storage the boiling point falls, and resinification may subsequently occur. Turpentine oil has a good solvency for fats, oils, waxes, and hydrocarbon resins. It is miscible with ethanol and aliphatic and aromatic hydrocarbons. Its dissolution properties are better than those of white spirit. It is used in addition to white spirit in oil-based and alkyd resin paints. The higher boiling constituents of wood turpentine oil contain a large amount of terpineol and are treated as pine oil. Pine oil is used as a high-boiling solvent in stoving enamels.

Table 23. Physical properties of solvents

Solvent	M	bp range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Aliphatic Hydrocarbons								
SBP spirit	~86	40–80	<−20	0.666	1.377	0.5	33.0	−30
SBP spirit	~100	64–138	<−20	0.708	1.399	0.6	12.5	−30
SBP spirit	~99	80–110	<−20	0.710	1.398	0.6	8.7	−20
SBP spirit	~112	100–140	<−20	0.740	1.418	0.8	4.0	−2
White spirit	~132	144–165	<−20	0.743	1.415	0.7	4.0	25
Petroleum ether	82.2	42–62	−150	0.653	1.370	0.46 (25 °C)	31.0	<0
Pentane	72.2	35–38	−130	0.632	1.358	0.45 (25 °C)	65.0	<−20
Hexane	86.2	65–70	−95	0.664	1.372	0.46 (25 °C)	20.0	−22
Heptane	100.2	94–99	−91	0.684	1.387	0.62 (25 °C)	8.5	−4
Isododecane	170	164–188	−81	0.752	1.420	1.34	0.31	46
Cycloaliphatic Hydrocarbons								
Cyclohexane	84.2	80.5–81.5	6.5	0.778	1.426	0.94	10.4	−17
Methylcyclohexane	98.2	101–103	−126	0.769	1.423	0.68	5.1	−4
Ethylcyclohexane	112.2	132	−111.3	0.79				19
Tetrahydronaphthalene	132.2	200–209	−31	0.969	1.539	2.2	0.024	74
Decahydronaphthalene	138.2	185–195	−30.4	0.888	1.475	2.4	0.29	57
Terpenes and Terpenoids								
Wood turpentine oil	140	150–180	<−40	0.860	1.465	1.48	0.44	35
Root turpentine oil	140	155–170	<−30	0.862	1.465			34
Balsam turpentine oil	140	150–180	<−30	0.861	1.471		0.44	32
Pine oil	140	179–230		0.883–0.936	>1.5			
α-Pinene	136.2	154–156	<−40	0.863	1.466	55	0.57	33
β-Pinene	136.2	164–166	−50	0.872	1.476		0.61	
Dipentene	136.2	174–181		0.864	1.474	1.7	0.22	50
D-Limonene	136.2	176	−96.9	0.843	1.473	3.5		49
Aromatic Hydrocarbons								
Benzene	78.1	79.8–80.8	5.5	0.879	1.501	0.65	10.1	−11
Toluene	92.1	110–111	−95	0.873	1.499	0.61	2.9	6
Xylene	106.2	137–142	−25	0.874	1.498	0.65	0.9	25
Ethylbenzene	106.2	135.5–136.5	−95	0.867	1.496	0.67	1.0	23
Cumene	120.2	152.9	−96	0.862	1.492	0.73	0.5	46

Table 23. (Continued)

Solvent	M	bp range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Mesitylene	120.2	164.7	− 44.7	0.870	1.499	0.71	0.8 (38 °C)	46
Pseudocumene	120.2	169.3	− 43.8	0.880	1.505		0.7 (38 °C)	52
Hemellitene	120.2	176.1	− 25.3	0.900	1.513		0.5 (38 °C)	53
Cymol	134.2	174 – 177	− 25	0.857	1.493	0.83		
Styrene	104.2	145	− 31	0.907	1.547	0.76	0.71	31
Chlorinated Hydrocarbons								
Dichloromethane	84.9	40.2	− 96	1.326	1.424	0.44	47.5	
Trichloromethane	119.4	61.2	− 64	1.488	1.445	0.58	21.0	
Ethyl chloride	64.5	12.5	− 138	0.924	1.379	0.96	133.0	− 43
Isopropyl chloride	78.6	36.2	− 117	0.859	1.378	0.29		− 36
1,2-Dichloroethane	99.0	83 – 84	− 36	1.254	1.444	0.74	8.2	13
1,1,1-Trichloroethane	133.4	74.0	− 32	1.332	1.438	1.1	13.3	
Trichloroethylene	131.4	86.7	− 86	1.446	1.477	0.58	7.3	
Perchloroethylene	165.9	121	− 23	1.629	1.506	0.88	1.9	
1,2-Dichloropropane	113.0	97.0	− 100	1.155	1.442	0.85	5.6	12
Alcohols								
Methanol	32.0	64.6	− 98	0.791	1.329	0.61	12.8	10
Ethanol	46.1	78.3	− 114	0.789	1.361	1.19	5.9	13
Propanol	60.1	97.2	− 127	0.804	1.386	2.26	1.9	23
Isopropyl alcohol	60.1	82.4	− 88	0.786	1.377	2.20	4.2	13
Butanol	74.1	117.7	− 80	0.810	1.399	3.00	0.67	35
Isobutanol	74.1	107.7	− 108	0.802	1.396	3.76	1.2	28
sec-Butanol	74.1	99.5	− 115	0.807	1.397	4.21	1.6	21
tert-Butanol	74.1	82.5	24.3	0.774	1.382	3.4 (30 °C)	4.1	14
Amyl alcohol	88.2	137 – 139	− 78.5	0.815	1.410	3.34	0.37	49
Isoamyl alcohol	88.2	130	− 134	0.810	1.408	4.37	0.30	44
Hexanol	102.2	157.6	− 57	0.820	1.418	4.3 (25 °C)	0.09	74
Methylisobutylcarbinol	102.2	131.4	− 90	0.808	1.411	4.1	0.30	41
2-Ethylbutanol	102.2	149.5	− 15	0.833	1.421	5.6	0.16	58
Isooctyl alcohol	130.2	184 – 190		0.832	1.431	10	0.04	75
2-Ethylhexanol	130.2	183 – 185	− 76	0.833	1.432	11.9	0.02	82
Isononanol	144.3	200 – 207		0.844	1.440	19	0.04	93
Isodecanol	158.3	214 – 222		0.838	1.440	18		102
Diisobutylcarbinol	144.2	178	− 65	0.812	1.423	2.8		74
Cyclohexanol	100.2	160.8	20	0.949	1.465	62.5	0.06	67

Table 23. (Continued)

Solvent	*M*	*bp* range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Methylcyclohexanol	114.2	160–195	− 38	0.925	1.465	26.2		68
Trimethylcyclohexanol	142.2	190–196	*cis* 37.3, *trans* 57.3	0.861 (60 °C)	1.439 (60 °C)	5.8 (60 °C, *cis*)		> 76
Benzyl alcohol	108.1	205.2	− 15	1.043	1.539	0.61	0.002	96
Methylbenzyl alcohol	122.2	200–205	21.5	1.008	1.525	11.2	0.003	90
Furfuryl alcohol	98.1	170.5		1.128	1.486	4.7		65
Tetrahydrofurfuryl alcohol	102.1	177	< − 80	1.062	1.451	5.49 (25 °C)	0.03	80
Diacetone alcohol	116.2	168	− 56	0.946	1.424	3.6	0.10	67
Ketones								
Acetone	58.1	56.2	− 95	0.791	1.359	0.33	24.1	− 19
Methyl ethyl ketone	72.1	79.6	− 86	0.805	1.379	0.40	0.1	− 14
Methyl propyl ketone	86.1	102.3	− 86	0.810	1.390	0.50	3.6	10
Methyl isopropyl ketone	86.1	92	− 92	0.803	1.389	0.48	8.6	5
Methyl butyl ketone	100.2	127.2	− 57	0.808	1.403	0.63	1.3	23
Methyl isobutyl ketone	100.2	114–117	− 84	0.802	1.396	0.61	2.15	14
Methyl amyl ketone	114.2	147.0–153.5	− 33	0.818	1.408	0.81	0.5	46
Methyl isoamyl ketone	114.2	141–148	− 74	0.813	1.407	0.73	0.6	41
Methyl heptyl ketone	142.2	194	− 9	0.820	1.422			60
Diethyl ketone	86.1	100–102	− 41	0.814	1.391	0.47	1.3	4
Ethyl butyl ketone	114.2	147.6	− 39	0.816	1.408	0.76	0.53	46
Ethyl amyl ketone	128.2	160.6	− 58	0.822	1.412	0.80	0.27	44
Diisopropyl ketone	114.2	124.4		0.806	1.401	0.63	0.5	19
Diisobutyl ketone	142.2	169.0	− 46	0.806	1.414	1.05	0.23	47
Cyclohexanone	98.1	153–156	− 26	0.946	1.451	2.2	0.35	43
Methylcyclohexanone	112.2	168	< − 60	0.919	1.442	1.78	0.07	48
Dimethylcyclohexanone	126.2	175–195		0.910				61
Trimethylcyclohexanone	140.2	188.8	− 10	0.888	1.445	2.54	0.1	64
Mesityl oxide	98.2	128.0	− 59	0.855	1.446	0.6	1.2	31
Isophorone	138.2	210–216	− 8	0.920	1.477	2.6	0.04	85
Acetyl acetone	100.1	140.5	− 23	0.973	1.451	0.6	1.0	35
Esters								
Methyl formate	60.1	31–32	− 99	0.978	1.341			− 18
Ethyl formate	74.1	54–55	− 80.5	0.927	1.359	0.41		− 19
Butyl formate	102.1	96–110	− 92	0.889				
Isobutyl formate	102.1	97.5	− 96	0.875	1.386	0.67	4.4	
Methyl acetate	74.1	55–57	− 98	0.932	1.361	0.37	22.6	− 13

Table 23. (Continued)

Solvent	M	bp range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Ethyl acetate	88.1	76–77	−84	0.900	1.372	0.45	10.3	−5
Propyl acetate	102.1	101–102	−92	0.888	1.385	0.59	3.1	10
Isopropyl acetate	102.1	88.4	−73	0.872	1.377	0.57	5.8	4
Butyl acetate	116.2	124–128	−77.9	0.880	1.394	0.74	1.11	23
Isobutyl acetate	116.2	114–118	−99	0.871	1.390	0.69	1.8	19
sec-Butyl acetate	116.2	112.2	−98.9	0.871	1.388	0.65	2.2	19
Amyl acetate	130.2	146–149	−100	0.876	1.403	0.92	0.6	40
Isoamyl acetate	130.2	142.1		0.874	1.401		0.6	23
Hexyl acetate	144.2	164–176		0.874	1.410	1.22 (25 °C)	0.16	59
Heptyl acetate	158.2	189–202		0.874	1.416	1.24 (25 °C)		66
2-Ethylhexyl acetate	172.3	192–205	−80	0.872	1.420	1.5	0.06	76
Cyclohexyl acetate	142.2	174.5	−72	0.970	1.440	2.27	0.9	58
Benzyl acetate	150.2	213–216	−52	1.055	1.506			93
Propylene glycol diacetate	160.1	190–192		1.057	1.414	3.0		87
Methyl propionate	88.1	75–82	−88	0.918	1.377			2
Ethyl propionate	102.1	95–99	−74	0.888	1.386 (15 °C)			8
Propyl propionate	116.2	121–123	−76	0.881	1.393	0.7		22
Butyl propionate	130.2	130–147	−90	0.870	1.404			38
Pentyl propionate	144.1	165–168		0.872	1.408	1.0		51
Ethyl butyrate	116.2	115–125	−98	0.880	1.393			22
Propyl butyrate	130.2	136–143	−95	0.872	1.398 (25 °C)			37
Butyl butyrate	144.2	159–162	−92	0.872	1.405	0.97	0.5	50
Isobutyl butyrate	144.2	142–156		0.860	1.403			46
Amyl butyrate	158.2	165–170	−73	0.864	1.414 (25 °C)			59
Methyl isobutyrate	102.1	99–101	−85	0.895				−1
Ethyl isobutyrate	116.2	109–111	−88	0.869				10
Isopropyl isobutyrate	130.2	119–121		0.844				15
Isobutyl isobutyrate	144.2	147.7–150.1	−80	0.855	1.400	1.02	0.4	49
Methyl lactate	104.1	145		1.085	1.414	2.9	0.30	55
Ethyl lactate	118.1	154.5	−25	1.036	1.412	2.44	0.2	48
Isopropyl lactate	132.2	157		0.98		2.9	0.10	55
Butyl lactate	146.2	188.0	−46	0.950	1.420	3.8	0.05	61
Butyl glycolate	132.2	182.0	−17	1.013	1.425	4.40	0.13	68
Methyl glycol acetate	118.1	142–148	−65	1.003	1.402	1.43	0.5	46
Ethyl glycol acetate	132.2	152–160	−62	0.975	1.405	1.35	0.3	52

Table 23. (Continued)

Solvent	*M*	*bp* range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Butyl glycol acetate	160.2	190–198	−65	0.945	1.415	1.8	0.04	75
Ethyl diglycol acetate	176.2	210–222	−25	1.011	1.421	2.8	0.013	107
Butyl diglycol acetate	204.3	235–250	−32	0.978	1.426	3.6	0.0013	108
Methoxypropyl acetate	132.2	143–150	−65	0.967	1.403	1.23	0.53	44
Ethoxypropyl acetate	146.2	158		0.941	1.405	1.4	0.23	54
3-Methoxybutyl acetate	146.2	166–172	−60	0.953	1.409	1.43	0.4	60
Ethyl 3-ethoxypropionate	146.2	165–172		0.947		1.0	0.23	58
Dibasic ester	160	196–225	−20	1.092	1.422	2.6	0.02	> 75
Ethylene carbonate	88.1	246.7	36.4	1.322 (0 °C)	1.415 (50 °C)	1.68 (50 °C)	0.001	145
Propylene carbonate	102.1	243.4	−54	1.205	1.422	2.76	0.004	130
Butyrolactone	86.1	204	−42	1.124	1.436	2.0	0.09	98
Glycol Ethers								
Methyl glycol	76.1	124.5	−85	0.966	1.402	1.98	0.83	38
Ethyl glycol	90.1	134–137	< −80	0.930	1.408	2.08	0.5	43
Propyl glycol	104.2	149.5–153.5	−90	0.911	1.414	2.42	0.13	49
Isopropyl glycol	104.2	142.8	< −60	0.907	1.410	2.6 (23 °C)	0.35	46
Butyl glycol	118.2	168–172	< −70	0.901	1.419	3.26	0.1	67
Hexyl glycol	146.2	208	−50	0.887	1.429	4.4	< 0.10	91
Phenyl glycol	138.2	244–250	−37	1.110	1.538	29.0		121
Methyl diglycol	120.2	194.2	−85	1.021	1.424	4.0	0.03	90
Ethyl diglycol	134.2	196–205	−90	0.990	1.429	4.38	0.02	94
Butyl diglycol	162.2	224–234	−68	0.956	1.431	5.85	0.01	105
Hexyl diglycol	190.3	240–265	−40	0.934	1.437	8.0		129
Methyl triglycol	164.2	245–255	−44	1.049	1.439	7.3	0.001	133
Ethyl triglycol	178.2	255.4	−19	1.04	1.440	7.8	< 0.001	120
Butyl triglycol	206.3	255–295	−35.1	0.985	1.439	9.42	< 0.001	130
Butyl tetraglycol	250.3	300–340		1.015	1.446		< 0.001	170
1-Methoxy-2-propanol	90.1	119–122	−96	0.922	1.403	1.89	1.12	32
Ethoxypropanol	104.1	130–138	−90	0.904	1.404	2.1	1.3	42
Isopropoxypropanol	118.1	139–145		0.877	1.407	2.3		41
Butoxypropanol	132.2	170	−102	0.884	1.417	3.38	0.08	63
Isobutoxypropanol	132.2	172	−51	0.888			0.11	
tert-Butoxypropanol	132.2	151	−34	0.872	1.411	3.3	0.6	45
Phenoxypropanol	152.2	241–246	−29	1.064	1.524	34	< 0.001	120
Methyl dipropylene glycol	148.2	187–192	−83	0.957	1.422	3.71	0.025	80

Table 23. (Continued)

Solvent	*M*	*bp* range (101.3 kPa), °C	Solidification point, °C	Density d_4^{20}, g/mL	Refractive index, n_D^{20}	Viscosity (20 °C), mPa · s	Vapor pressure (20 °C), kPa	Flash point, °C
Isopropyl dipropylene glycol	178.2	198 – 208		0.917	1.420	4.3		87
Butyl dipropylene glycol	192.2	229	– 75	0.91		4.6	0.006	113
Methyl tripropylene glycol	206.3	235 – 251	– 78	0.967	1.428	6.16	0.004	121
Butyl tripropylene glycol	248.3	274	< – 75	0.92		8.0	< 0.001	136
Diglycol dimethyl ether	134.2	160		0.944	1.407	2.0	1.5	57
Dipropylene glycol dimethyl ether	162.2	175	< – 71	0.904			0.07	65
Ethers								
Diethyl ether	74.1	34.5	– 116	0.714	1.353	0.24	58.8	– 40
Diisopropyl ether	102.2	68.3	– 85.5	0.724	1.368	0.32	21.0	– 22
Dibutyl ether	130.2	142.2	– 95	0.769	1.399	0.7	0.65	25
Methyl *tert*-butyl ether	88.2	55.3	– 109	0.740	1.369	0.36	27.1	– 28
Tetrahydrofuran	72.1	66	– 108	0.888	1.407	0.61	17.3	– 21.5
Dioxane	88.2	101.1	11.8	1.034	1.423	1.23	4.0	5
2,2-Dimethyl-4-hydroxymethyl-1,3-dioxolane	132.2	190		1.068	1.438			90
1,2-Propylene oxide	58.1	33.9		0.830	1.363	0.38	59.5	– 37
Miscellaneous Solvents								
Dimethyl acetal	90.1	56 – 65		0.839				– 28
Dimethylformamide	73.1	153.0	– 61	0.949	1.430	0.92	0.6	57
Dimethylacetamide	87.1	165.5	– 20	0.959	1.435	0.92 (25 °C)	0.2	70
Dimethyl sulfoxide	78.1	189.0	18.5	1.100	1.478	1.99 (25 °C)	0.06	95
Tetramethylene sulfone	120.2	285	– 36	1.173	1.483			177
Carbon disulfide	76.1	45.5 – 46.8	– 111	1.263	1.625			– 30
Furfural	96.1	160 – 165	– 36.5	1.155	1.526		0.15	57
Nitromethane	61.0	101.2	– 29	1.139	1.381	0.65	3.7	33
Nitroethane	75.1	114.0	– 90	1.052	1.391	0.68	2.08	29
1-Nitropropane	89.1	131		0.994				34
2-Nitropropane	89.1	120.3	– 93	0.992	1.394	0.77	1.72	27
N-Methylpyrrolidone	99.1	202 – 205	– 24.4	1.028	1.470	1.8	0.04	91
N-Ethylpyrrolidone	113.1	200	< – 70	0.993	1.466	3.5		93
N-Cyclohexylpyrrolidone	167.3	284	12	1.026	1.495	11.5	< 0.005	> 100
N-(2-Hydroxyethyl)pyrrolidone	129	295	20	1.139	1.495	53		> 100
1,3-Dimethyl-2-imidazolidinone	114.1	225	8.2	1.060	1.470	1.9		107
Hexamethylenephosphoric triamide	179.0	233	6					

Table 24. Safety and physiological data of solvents

Solvent	Evaporation no. (diethyl ether = 1)	Evaporation no. (butyl acetate = 1)	Ignition temperature, °C	MAK 1991 ppm	MAK 1991 mg/m^3	MAK 2001 ppm	MAK 2001 mg/m^3	TLV 1991–1992 ppm	TLV 1991–1992 mg/m^3	Odor threshold, ppm	LD_{50}, rat, oral, mg/kg	LC_{50}, rat, inhalation, $\times 10^{-2}$ ppm
Pentane	< 1	13	285	1000	3000	1000	3000	600	1770	900	> 24 (mL/kg)	> 90 000/2 h Mouse
Hexane	1.4	8.4	240	50	[c] 180	50	180	50	176		24–49 (mL/kg)	54 000–74 000
Heptane	3.0	3.3	215	500	2100	500	2100	400	1640	150	> 25 (mL/kg)	14 000–16 000
Octane		1.23		500	2400	500	2400	300	1400	200		
Cyclohexane	3.4	5.60		200	700	200	700	300	1010	2	1297	
Methylcyclohexane		2.99	285	200	2000	200	2000	400	1610			
Turpentine oil	38	< 0.005	252	100	[f] 560	-	-	100	556			
Tetrahydronaphthalene	200		425			-	-				2860	
Toluene	6.1	2.0	535	100	[b] 390	50	190	50	147	40	5000	40/4 h
Xylene	13.5	0.76	525	100	[d] 440	100	440	100	434	0.4–20	4300	63.5/4 h
Ethylbenzene	8.8	0.84	435	100	[e] 440	-	-	100	434	125	3500	40/4 h
Cumene		0.14	424	50	[e] 245	50	250	50	246	40 (ppb)	1400	80/4 h
Styrene	16	0.536	480	20	[c] 85	20	86	50	213	0.1	5000	240/2 h
Dichloromethane	1.8	25.0	662	100	[d,h] 360	-	-	[g] 50	[g] 174	550	2136	880/0.5 h
1,1,1-Trichloroethane	2.6	6.00	537	200	[c] 1080	200	1100	350	1910	300	10 300	10/4 h (LCLo)
Trichloroethylene	3.1	4.9	460	50	[c,h] 270	-	-	50	269	200	7193	80/4 h (LCLo)
Perchloroethylene	6.0	2.59		50	[c,h] 345	-	-	50	339	320	8850	40/4 h (LCLo)
Methanol	6.3	4.1	463	200	[d,e] 260	200	270	200	262	6000	5628	640/4 h
Ethanol	8.3	2.4	419	1000	[d] 1900	500	960	1000	1880	6000	7060	380/10 h
Isopropyl alcohol	11	1.5	425	400	[d] 980	200	500	400	983	600	5045	160/8 h
Butanol	33	0.47	360	100	300	100	310	50	152	33	790	50/5 h
Isobutanol	25	0.85	390	100	300	100	310	50	152	80	2460	80/8 h (LCLo)
Isoamyl alcohol	62		365	100	360	100	370	100	361	15	1300	
Methylisobutylcarbinol	66	0.28	305	25	[e] 100	25	110	25	104		2600	20/4 h
2-Ethylhexanol	690	0.0018	330			50	270			0.2	800	
Cyclohexanol	150	0.08	300	50	200	50	210	50	206	50 (ppb)	2060	
Methylcyclohexanol	800		296	50	235	-	-	50	234	300		
Furfuryl alcohol		0.04	391	50	200	10	41	10	40	9	275	2.33/4 h
Diacetone alcohol	150	0.15	624	50	240	20	96	50	238	3	4000	7/8 h
Acetone	2.0	5.6	540	1000	2400	500	1200	750	1780	300	9750	
Methyl ethyl ketone	3.3	3.7	514	200	[d] 590	200	600	200	590	30	3400	20/4 h

Table 24. (Continued)

Solvent	Evaporation no. (diethyl ether = 1)	Evaporation no. (butyl acetate = 1)	Ignition temperature, °C	MAK 1991 ppm	MAK 1991 mg/m^3	MAK 2001 ppm	MAK 2001 mg/m^3	TLV 1991–1992 ppm	TLV 1991–1992 mg/m^3	Odor threshold, ppm	LD_{50}, rat, oral, mg/kg	LC_{50}, rat, inhalation, $\times 10^{-2}$ ppm
Methyl propyl ketone	4.8	2.3	449	200	700	-	-	200	705	8	3730	20/4 h
Methyl butyl ketone	13	0.98	424	5	21	5	21	5	20	32	2590	
Methyl isobutyl ketone	7	1.4	460	100	400	20	83	25	104	8	2080	40/4 h (LCLo)
Diisobutyl ketone	48	0.2	345	50	290	-	-	25	145	50	5750	20/4 h (LCLo)
Cyclohexanone	41	0.25	420	50	[c] 200	-	-	25	100	50	1620	20/4 h
Methyl cyclohexanone	53	0.18	595	50	[e] 230	50	230			0.5	2140	
Mesityl oxide	8.2	0.88	340	25	100	25	100	15	60	30	1120	10/4 h
Isophorone	330	0.023	462	5	28	2	11	5	28	0.7	2330	18.4/4 h (LCLo)
Methyl acetate	2.1	9.5	501	200	[d] 610	200	610			200	3700	
Ethyl acetate	3	4.2	460	400	1400	400	1500	400	1440	50	6100	16/8 h
Propyl acetate	4.8	2.0	457	200	840	100	420	200	835	20	9800	80/4 h
Butyl acetate	12	1.0	370	200	950	100	480	150	713	15	14 000	20/4 h
Isobutyl acetate	5.8	1.7	480	200	950	-	-	150	713	4	15 000	80/4 h
Amyl acetate	15	0.8	380	100	525	50	270	100	332	20	7400	52
Methyl glycol acetate	21	0.35	400	5	[b,e] 25	5	25			0.9	3390	70/4 h
Ethyl glycol acetate	57	0.20	345	20	[b,e] 110	5	27			0.5	2900	15/8 h (LCLo)
Ethyl diglycol acetate	> 1200	0.008	360			-	-				11 000	
Butyl glycol acetate	137	0.03	375	20	[c,e] 135	20	130				3200	
Butyl diglycol acetate	> 3000	0.002	290			-	-				6500	
Methoxypropyl acetate	34	0.32	354	50	[c] 275	50	270				> 8500	> 23.9/6 h
Ethoxypropyl acetate	70	0.19				-	-				> 5000	
Ethyl-3-ethoxypropionate		0.12	377			-	-					
Butyrolactone	> 1000	0.03				-	-				1580	
Ethyl lactate	80	0.22	294			-	-				> 2000	> 50
Dibasic ester	4800	0.008	370			-	-				17 000	10.7/1 h
Methyl glycol	34	0.53	290	5	[b,e] 15	5	16			90	2460	20/4 h
Ethyl glycol	43	0.38	235	20	[b,e] 75	5	19			50	3000	40/4 h (LCLo)
Propyl glycol	75	0.2	235			20	86				3089	> 21.3/6 h
Butyl glycol	119	0.082	240	20	[c,e] 100	20	98			0.5	1480	4.5/4 h
Ethyl diglycol	1200	0.017	220			-	-				5540	> 170/12 d
Butyl diglycol	3750	0.004	225			-	100				6560	> 30/7 h
1-Methoxy-2-propanol	22	0.75	290	100	375	100	370				5660	55/4 h
Ethoxypropanol	33	0.46	280			-	-					
Diethyl ether	1	28.0	180	400	1200	400	1200			0.3	1215	730/2.5 h
Diisopropyl ether	1.6	8.1	405	500	2100	200	850			0.06	8470	3000/2 h (LCLo)

Table 24. (Continued)

Solvent	Evaporation no. (diethyl ether = 1)	Evaporation no. (butyl acetate = 1)	Ignition temperature, °C	MAK 1991 ppm	MAK 1991 mg/m^3	MAK 2001 ppm	MAK 2001 mg/m^3	TLV 1991–1992 ppm	TLV 1991–1992 mg/m^3	Odor threshold, ppm	LD_{50}, rat, oral, mg/kg	LC_{50}, rat, inhalation, $\times 10^{-2}$ ppm
Methyl *tert*-butyl ether	1.6	8.4	460			50	180			267	3866	850/4 h
Tetrahydrofuran	2.2	6.3	212	200	[c] 590	50	150	200	590	40	3000 (LDL_0)	780/2 h
Dioxane	7.3	2.17	266	50	[d,e,h] 180	20	73	25	90	200	7120	
Dimethylformamide	120	0.20	445	20	[b,e] 60	10	30	10	30	90	2800	
Dimethylacetamide	172	0.138	400	10	[c,e] 35	10	36	10	36		4300	
Carbon disulfide	1.8	10.90	102	10	[e] 30	5	16	10	31			
Furfural	75		391	5	[e] 20	-	-				127	1.53/4 h
Nitroethane	21	1.2		100	310	100	310				1100	
1-Nitropropane		0.9		25	90	25	92	25	91	300	800	
2-Nitropropane	10	1.5	428	[g] carcinogenic		–	-	[g] 10	[g] 36	200		15.13/5 h
N-Methylpyrrolidone		0.04	287	100	[d] 400	19	80				3600	

[a,b,c,d] Pregnancy groups. [e] Danger of skin resorption. [f] Danger of sensitization. [g] III A2 Carcinogenic group. [h] III B Carcinogenic group.

Root turpentine oil is obtained by steam distillation of wood chips and shavings with a high resin content. Its properties are similar to those of sulfite cellulose oil of turpentine, which occurs in wood pulp production. Root turpentine oil is a colorless liquid with similar properties to balsams turpentine oil.

Wood oil is obtained by dry distillation of chips and shavings with a high resin content. It is a yellow, acidic liquid with an unpleasant odor. On exposure to air and also after purification it becomes discolored and the odor becomes more unpleasant. It is only suitable as a solvent for dark paints and lacquers.

Terpenoids are obtained by distillation of turpentine oils. α-Pinene [*80-56-8*] is used in the preparation of camphor. Higher-boiling constituents include pine oil [*8000-41-7*] (terpineol; trade names: Depanol, Solvenol, Terposol, Yarmor). Hydroterpine is prepared by hydrogenation of pine oil and is used as a substitute for turpentine oil. Terpenoids are used as high-boiling solvents in oil-based and alkyd resin paints. Dipentene [*138-86-3*] prevents skin formation in air-drying paints. D-Limonene [*5989-27-5*] see [265].

9.4. Aromatic Hydrocarbons

Compared to aliphatic hydrocarbons, aromatic hydrocarbons (DIN 51 633) have a higher solvency for oils, castor oil, oil-modified alkyd resins,styrene-modified oils and alkyd resins,saturated polyester resins, polystyrene, poly(vinyl ethers), polyacrylate and polymethacrylate esters, poly(vinyl acetate), vinyl chloride and vinyl acetate copolymers,and many low-polarity resins.

Aromatic hydrocarbons are used as diluents in solutions of cellulose nitrate, cellulose esters, and ethers with true solvents such as esters and ketones. Rubber, polyisobutene, and molten polyethylene also dissolve in them. Poly(vinyl chloride), solid polyethylene, polyamides, and shellac are, however, insoluble or only swell.

Toluene [*108-88-3*] is mainly used in cellulose nitrate lacquers; in heat- curing paints based on urea--, melamine--, or phenol–formaldehyde resins; in alkyd resin paints; and in paints based on chlorinated rubber, polystyrene, polyacrylates, or poly(vinyl acetate) (→ Toluene).

Mixtures of toluene with esters are used to dissolve vinyl chloride copolymers and postchlorinated poly(vinyl chloride) [266].

Xylene (mixture of *o*-, *m*-, and *p*-isomers, [*56004-61-6*], [*108-38-3*], [*41051-88-1*]) generally contains very small amounts of toluene and larger amounts of ethylbenzene (→ Xylenes). Xylene is the most important aromatic solvent in the paint industry. It dissolves poly(vinyl acetate) only in combination with alcohols or glycol ethers, but is otherwise similar to toluene as regards dissolution properties.

Ethylbenzene [*100-41-4*] (phenylethane) has a water- clear appearance and characteristic odor (→ Ethylbenzene). It is miscible with practically all organic solvents, but is insoluble in water. Ethylbenzene is of limited importance as a paints solvent. Its main use is as an industrial starting material for the production of styrene by catalytic dehydrogenation (→ Styrene). Catalytic oxidation of ethylbenzene with air in the presence of heavy-metal oxides yields acetophenone and phenylmethylcarbinol. It also improves the antiknock properties of fuels for Otto engines.

Cumene [*98-82-8*] is the principal constituent of heavy naphtha which is the feedstock for phenol and acetone synthesis by the Hock process (→ Hydrocarbons, Chap. 3.7.). It is also a byproduct in the production of sulfite pulp.

Styrene [*100-42-5*] is a colorless liquid that acts as a solvent for unsaturated polyester resins (→ Styrene). Styrene is stabilized by the manufacturer to avoid polymerization on storage. Feedstock for chemical syntheses (polymers and copolymers, styrene-modified alkyd resins and oils).

9.5. Chlorinated Hydrocarbons

Chlorinated hydrocarbons have a better solvency than the corresponding nonchlorinated compounds for most resins, polymers, rubber, waxes, asphalt, and bitumen. Chlorinated hydrocarbons are miscible with other organic solvents, but are insoluble in water. They have a sweetish odor. Increasing the number of chlorine substituents reduces the combustibility and improves the solvency, but also increases the toxicity.

All chlorinated hydrocarbons may decompose under the action of light, air, heat, and water. Decomposition can be reduced but not completely prevented by adding stabilizers. On account of their health hazard, some chlorinated hydrocarbons may no longer be used as conventional solvents, e.g., tetrachloromethane, tetrachloroethane, and pentachloroethane. Dichloromethane, trichloroethylene, perchloroethylene, and 1,1,1-trichloroethane are increasingly being replaced for reasons of industrial hygiene and environmental protection, particularly of water.

Dichloromethane [*75-09-2*] (methylene chloride) is a colorless, highly volatile, neutral liquid with a characteristic odor (→ Chlorinated Hydrocarbons, Chap. 1.). It is insoluble in water but miscible with organic solvents. It has a very good solvency for many organic substances, such as fats, oils, waxes, and resins. Bitumen, rubber, chlorinated rubber, polystyrene, postchlorinated poly(vinyl chloride), vinyl chloride copolymers, polyacrylates, and cellulose esters are also soluble. The solubility spectrum can be expanded by adding other solvents. A mixture of methanol or ethanol and dichloromethane is a good solvent for cellulose ethers and acetyl cellulose. Cellulose nitrate is, however, insoluble.

Dichloromethane is used as a noncombustible solvent and extractant for oils, fats, waxes, fish oil, etc. from industrial and animal products, as well as caffeine, hops, castor oil, cocoa butter, and ethereal oils from substances of plant origin. It is used in the deparaffination of petroleum and the azeotropic dewatering of solvents. Dichloromethane is a constituent of paint-removal pastes and baths, but is increasingly being replaced by aqueous systems. It is used as a solvent in the production of cellulose-acetate-based or cellulose-acetobutyrate-based films and is employed industrially in the leather, metal, rubber, adhesives, and plastics industries.

Trichloromethane [*67-66-3*] (chloroform) is a solvent having a powerful narcotic effect and is not widely used (→ Chlorinated Hydrocarbons, Chap. 1.).

Tetrachloromethane [*56-23-5*] (carbon tetrachloride, "Tetra"), is a colorless, neutral, nonflammable liquid (→ Chlorinated Hydrocarbons, Chap. 1.). It is toxic and forms phosgene, hydrogen chloride, and chlorine when heated. It dissolves fats, oils, resins and waxes, but not poly(vinyl acetate). Tetrachloromethane is used

in the production of rubber-adhesive solutions, for diluting sulfur monochloride in rubber vulcanization, and in the leather and shoe industries.

1,2-Dichloroethane [*107-06-2*] (ethylene chloride) is one of the most stable chlorinated hydrocarbons, with a good solvency for fats, oils, resins, rubber, bitumen, asphalt, and tar (→ Chlorinated Hydrocarbons, Chap. 2.3.). Highly toxic, the odor threshold is higher than the MAK value. 1,2-Dichloroethane is used in the production of building protection agents, roofing felts, and cold asphalt. It is used as an extracting agent for alkaloids and ethereal oils.

1,1,2,2-Tetrachloroethane [*79-34-5*] is a good solvent for resins, rubber, and cellulose acetate (→ Chlorinated Hydrocarbons, Chap. 2.7.). It is used only to a limited extent on account of its potential health hazard.

1,1,1-Trichloroethane [*71-55-6*] (methylchloroform) is insoluble in water, but miscible with organic solvents (→ Chlorinated Hydrocarbons, Chap. 2.4.). It dissolves fats, oils, resins, waxes, bitumen, and asphalt. It is used to degrease metals; as a solvent in the paint, adhesives, and plastics industries; as well as an extracting agent for fats and ethereal oils. It is being increasingly replaced for environmental protection reasons by other systems.

1,2-Dichloroethylene [*75-35-4*] (vinylidene chloride) is a highly volatile liquid with a low flammability although the hot vapors can ignite (→ Chlorinated Hydrocarbons, Chap. 3.2.).

Trichloroethylene [*79-01-6*] ("Tri") is insoluble in water, but is miscible with organic solvents (→ Chlorinated Hydrocarbons, Chap. 3.4.). It dissolves fats, oil, waxes, rubber, and many resins; it is used as a solvent and extracting agent. In paints, trichloroethylene raises the flash point, and is used as a solvent in paint removal pastes. It is increasingly being replaced by other systems for environmental protection reasons.

Perchloroethylene [*127-18-4*] (tetrachloroethylene, "Per") is a colorless, water-insoluble, nonflammable solvent having a very good solvency for greases, fats, waxes, oils, bitumen, tar, and many natural and synthetic resins (→ Chlorinated Hydrocarbons, Chap. 3.5.). It is used in chemical cleaning systems, to degrease light and heavy metals, to degrease pelts and leather (tanning), extraction of animal and vegetable fats and oils, and textile dyeing (solvent for dye baths). It is also used in the preparation of secondary chemical products (fluorinated hydrocarbons, trichloroacetic acid, etc.). Perchloroethylene is increasingly being replaced by other systems for environmental protection reasons.

1,2-Dichloropropane [*78-87-5*] is a solvent for bitumen, asphalt, and tar in the production of road-surfacing materials, building protective agents, and roofing felts (→ Chlorinated Hydrocarbons, Chap. 4.2.).

Chlorobenzene [*108-90-7*] is a colorless, neutral liquid with a weak, benzene-like odor (→ Chlorinated Hydrocarbons, Chap. 8.1.). It is insoluble in water and miscible with organic solvents. Chlorobenzene has a good solvency for fats, oils, resins, polymers, binders, rubber, and chlorinated rubber. Cellulose ethers dissolve in the presence of small amounts of alcohols; cellulose nitrate is insoluble. Chlorobenzene is a solvent in the production of bitumen and asphalt coatings for building protection.

9.6. Alcohols

On account of their hydroxyl groups alcohols (DIN 53 245) differ from aliphatic, aromatic, and chlorinated hydrocarbons by having a higher polarity and a stronger tendency to form hydrogen bonds. The relationship between the nonpolar hydrocarbon chain and the hydroxyl group is decisive for the solvency of alcohols [267]. Lower alcohols accordingly exhibit a pronounced solvency for strongly polar resins such as shellac, copal, glyptal resins, urea --, melamine --, and phenol – formaldehyde resins, cellulose nitrate, slightly ethylated ethyl cellulose, and poly(vinyl acetate). Lower alcohols are, however, poor solvents for nonpolar substances such as fats, oils, oil-modified alkyd resins, styrene-modified oils, and hydrocarbon resins.

The solvency for polar substances decreases with increasing hydrocarbon chain length, propanol and higher alcohols are therefore no longer particularly good solvents for cellulose nitrate and poly(vinyl acetate). Higher alcohols are mainly used as diluents, their solvency is substantially improved by combination with the corresponding acetate esters. On account of their gentle dissolution properties, higher alcohols are

highly suited for topcoats that are applied to a primer; the primer is not softened and lifted by these alcohol solvents. In the painting of plastics, alcohols produce only weak superficial swelling of the plastic, resulting in good adhesion without softening of the plastic. In the roller coating of paper alcohols prevent the rubber rollers from swelling, and bleeding of the coating material into the paper, high-gloss prints and paper coatings are therefore obtained.

Methanol [*67-56-1*] (methyl alcohol) is a clear, colorless liquid with a characteristic odor (→ Methanol). It is hygroscopic and miscible in all proportions with water as well as with many organic solvents. It is less soluble in fats and oils, and is only partially miscible with aliphatic hydrocarbons. Numerous inorganic substances (e.g., many salts) are soluble in methanol. Methanol has a good solvency for polar resins, cellulose nitrate, and ethyl cellulose. However, oil-modified alkyd resins and polymers, with the exception of poly(vinyl acetate), poly(vinyl ethers),polyvinylpyrrolidone and polymethacrylamide, are insoluble.

Methanol is used for methylation, and is the alcohol most easily esterified. It is used in the preparation of formaldehyde and methyl esters (e.g., dimethyl terephthalate, methyl methacrylate, methyl formate). It is also employed as a solvent for cellulose nitrate, colophony, shellac, and urea resins in the explosives and paint industries. Furthermore, it is used as an antifreeze, fuel, and extracting agent. Methanol is toxic although cases of poisoning are extremely rare if it is correctly used.

Ethanol [*64-17-5*] (ethyl alcohol) is available as an ethanol – water azeotrope and in anhydrous form. Both forms are supplied completely or partially denatured. Complete denaturation is effected by adding methyl ethyl ketone. Toluene, petroleum ether, and special gasolines are usually used for partial denaturation.

Ethanol is a colorless, clear liquid with a characteristic, pleasant odor (→ Ethanol). It is miscible in all proportions with water and readily miscible with many organic solvents (e.g., ethers, hydrocarbons, acids, esters, ketones, carbon disulfide, glycols, and other alcohols). Ethanol dissolves castor oil, cellulose nitrate with a low nitrate content, polar resins, and polymers. Ethanol in combination with aromatic compounds dissolves cellulose acetate. Mixtures of ethanol, aromatic hydrocarbons, and water are good solvents for some polyamides. Ethanol is extensively used in the chemical and pharmaceutical industries. It is employed as a raw material for many chemical syntheses (e.g., esterification, as an ethylating agent, and reaction medium). Ethanol is an excellent solvent, diluent, and extracting agent for fats, oils, paints, and natural substances, such as aroma substances, fragrances, perfumes, and pharmaceutical preparations. On account of its pleasant odor and weak substrate solvency it is employed in the formulation of paper and plastics coatings and paints, as well as for flexographic inks. It is also used in the form of methylated spirit for cooking or heating.

Propanol [*71-23-8*] (1-propanol) is a colorless liquid that is soluble in water and miscible with organic solvents (→ Propanols). It has better dissolution properties than ethanol for fats and oils, and dissolves polar resins in the same way as ethanol. Cellulose nitrate and poly(vinyl acetate) are, however, almost insoluble. For economic reasons propanol is of only limited use as a solvent, and is a starting material for esters.

Isopropyl alcohol [*67-63-0*] (2-propanol, isopropanol) is a colorless liquid that is miscible in all proportions with water and with commonly used organic solvents (→ Propanols). It forms binary and ternary azeotropes with water and many organic solvents.

Isopropyl alcohol is extensively used in cosmetics, particularly for hair and skin lotions, and in pharmacy for preparations intended for external use. A further large area of use is the paint and printing ink industry. Isopropyl alcohol is added to fuels to prevent the icing up of the carburetor and increase the octane number. It is an important feedstock for the chemical industry, for example, in the production of acetone (particularly in the United States), esters, plasticizers, and ethers. Isopropyl alcohol is also important as a solvent (e.g., for recrystallization and extraction) and as a moistening agent for cellulose nitrate. It is furthermore used in the aerosol sector.

Butanol [*71-36-3*] (1-butanol) is a colorless neutral liquid that has a limited miscibility with water and a characteristic odor (→ Butanols). It is miscible with organic solvents. Butanol has a high solvency for most known natural and synthetic resins, fats, oils, linseed oil, saturated

polyesters, and poly(vinyl acetate). It considerably increases the dilutability of cellulose nitrate solutions with non-solvents. Cellulose esters, cellulose ethers, chlorinated rubber, poly(vinyl chloride), vinyl chloride copolymers, and polystyrene are not dissolved by butanol.

Butanol is one of the most important additives in the production of cellulose nitrate lacquers. It ensures haze-free drying and prevents blushing of drying lacquer films. Butanol improves not only the flow of cellulose nitrate lacquers but also acts as a solubilizer between the individual lacquer constituents. It is used to prevent stringing in the production of spirit varnishes. Although most alkyd resins are only sparingly soluble in butanol, even small additions of butanol to alkyd resin or oil-based paints considerably decrease viscosity and improve flow and coatability. Butanol is used in electrostatically sprayed paints to regulate the conductivity. It can be employed to a limited extent as a diluent in vinyl chloride copolymers. Butanol serves as a viscosity-reducing auxiliary solvent in waterborne paints and prevents foaming of colorants (e.g., in the paper industry). It is suitable for moistening cellulose nitrate wool and is used to a large extent as a feedstock in the production of plasticizers and urea – and melamine – formaldehyde resins.

Isobutanol [*78-83-1*] (2-methyl-1-propanol) is a colorless, neutral liquid with a characteristic odor (→ Butanols). It exhibits limited miscibility with water. Most organic solvents are miscible in practically all proportions with isobutanol. Isobutanol readily dissolves most natural and synthetic resins. Waxes dissolve satisfactorily only on heating. Cellulose esters and ethers, natural rubber, neoprene, chlorinated rubber, and polymers such as polystyrene and poly(vinyl chloride) are insoluble in isobutanol.

Isobutanol prevents blushing in drying paint films, and also improves flow and gloss. When added in amounts of 5 to 10 % it reduces the viscosity of oil-based paints, alkyd paints, and cellulose nitrate lacquers. Isobutanol can replace butanol as a moistening agent for cellulose nitrate. It is also used in the production of spirit varnishes and printing inks, and as a raw material for the production of plasticizers.

sec-Butanol [*78-92-2*] (2-butanol) has a peppermint-like odor and a considerably higher water solubility than butanol (→ Butanols). It is used to a limited extent as a solvent, e.g., in paint-removal pastes, spirit varnishes, and printing inks. On account of its good water absorption it is also suitable as a cosolvent in waterborne paints.

tert-Butanol [*75-65-0*] (2-methyl-2-propanol) is a colorless liquid that solidifies at room temperature (→ Butanols). It is soluble in most organic solvents such as alcohols, ethers, ketones, esters, and hydrocarbons. It is miscible in all proportions with water. The azeotrope with water is a liquid at room temperature.

tert-Butanol is widely used as a solvent and solubilizer because of its extremely high resistance to oxidizing agents and halogens in comparison with other alcohols. On account of its good solvency for waxes it is used as a wax remover and as an additive for the deparaffination of oils. *tert*-Butanol is widely used as an alkylating agent for aromatic compounds and phenols, isomerization of the *tert*-butyl group does not occur. *tert*-Butanol is employed in the preparation of *tert*-butyl hydroperoxides, which are used as polymerization catalysts. It also acts as a stabilizer in chlorinated hydrocarbons and as a denaturing agent for ethanol. *tert*-Butanol is used in the production of aromas (including *tert*-butyl esters and *tert*-butyl aromatics) and antioxidants.

Amyl Alcohol. Of the eight isomers of amyl alcohol, three have asymmetrically-substituted carbon atoms (→ Pentanols). The amyl alcohol obtained from fusel oil consists of 3-methylbutanol [*123-51-3*] (isoamyl alcohol) and 2-methylbutanol [*137-32-6*] (active amyl alcohol); the mixture is sold as fermentation amyl alcohol or technical grade isoamyl alcohol. Isoamyl alcohol has similar dissolution properties to butanol. It is used as a solvent, and as a raw material for acetate esters, plasticizers, and lubricant components. Amyl alcohol containing *n*-pentanol (1-Pentanol [*71-41-0*]) as the principal constituent is used on a small scale as a solvent, like butanol. It has a better oil solubility and is therefore used as an additive in lubricating and hydraulic oils. It is used as an extracting agent for uranium salts and bituminous constituents of coal tar. It also serves as a raw material in the production of plasticizers and amyl xanthogenates (flotation auxiliaries for sulfidic ores). Pentasols are commercially available mixtures of primary and secondary pentanols.

Hexanol. The most important isomers are hexanol [*111-27-3*], 2-ethyl-1-butanol [*97-95-0*] and methylisobutylcarbinol [*108-11-2*](4-methyl-2-pentanol) (→ Alcohols, Aliphatic, Chap. 2.4.1.). Hexanol is a high boiler and is used to improve the flow and surface properties of paints; it is also used as a solvent for fats, waxes, and dyes.

2-Ethylhexanol [*104-76-7*] is a colorless, waterwhite liquid with a characteristic odor (→ 2-Ethylhexanol). It is practically insoluble in water, but is miscible with all usual organic solvents. It is a good solvent for many vegetable oils and fats, for some dyes, and for many synthetic and natural paint raw materials.

2-Ethylhexanol is esterified to produce plasticizers for various uses. The esters of 2-ethylhexanol with phthalic, adipic, azelaic, and phosphoric acid are particularly important. Maleate and acrylate esters are used in copolymerization for the internal plastification of copolymers. Di-2-ethylhexyl esters of adipic, sebacic, and decanedicarboxylic acids are base oils used for the production of synthetic, thermally-stable lubricants, in particular for jet engines and gas turbines. 2-Ethylhexyl sulfate and diethylhexyl sulfosuccinate are used as wetting and leveling agents in the textile industry. Ethoxylates of this alcohol are used as raw materials for emulsifiers, detergents, and cleansing agents. Esterification with hydrochloric acid yields 2- chloro-2-ethylhexane, which is used in the production of tin- containing stabilizers and as an intermediate for introducing the 2-ethylhexyl group into other compounds. 2-Ethylhexanol is a defoaming agent for aqueous systems in the paper, ceramics, and textile industries. It also serves as a grinding auxiliary for pigments, acts as a surface impregnation agent, and facilitates the dispersion of pigments in nonaqueous solvents. As a high-boiling solvent, small amounts of 2-ethylhexanol are added to stoving enamels to improve flow and gloss formation.

Benzyl alcohol [*100-51-6*] is miscible with organic solvents apart from aliphatic hydrocarbons (→ Benzyl Alcohol). It dissolves cellulose esters and ethers, fats, oils, alkyd resins, natural and synthetic resins, and colorants. Polymers – with the exception of lower poly(vinyl ethers) and poly(vinyl acetates) – are not dissolved. Small amounts of benzyl alcohol improve the flow and gloss paints, delay the evaporation of other solvent components, and have a plasticizing effect in physically drying paints. It is used in ballpoint pen inks. Benzyl alcohol reduces the viscosity in two-pack epoxy resin systems.

Methylbenzyl alcohol [*98-85-1*] (1-phenylethanol, 1-phenylethyl alcohol, phenylmethylcarbinol) is an almost colorless, neutral liquid that has limited miscibility with water and a weak, bitter, almond-like odor. It has a high solvency for alcohol-soluble cellulose nitrate, cellulose acetate, and cellulose acetobutyrate; for many natural and synthetic resins; and for fats and oils. In contrast to benzyl alcohol, it is miscible with white spirit.

Methylbenzyl alcohol can be used like benzyl alcohol, and can be employed advantageously in stoving enamels. In cellulose nitrate and acetyl cellulose lacquers methylbenzyl alcohol helps to improve flow and film formation and prevents blushing at relatively high atmospheric humidity levels. It has also proved to be a useful additive in paint-removal agents on account of its dissolution properties and long evaporation time. The solvency of methylbenzyl alcohol for colorants is similar to that of benzyl alcohol.

Cyclohexanol [*108-93-0*] has a camphor-like odor with a water solubility of 2 %. It is miscible with other solvents and dissolves fats, oils, waxes, and bitumen, but not cellulose derivatives (→ Cyclohexanol and Cyclohexanone).

Cyclohexanol is used in cellulose nitrate combination lacquers and oil-based paints. It delays drying, prevents blushing, and improves flow and gloss. In topcoats and clear varnishes, cyclohexanol can prevent undesirable dissolution of the primer coats. Cyclohexanol is used to remove paraffin from mineral oil; as a solvent in waxing, cleansing, and polishing agents; and as a wetting agent for spray liquids.

Methylcyclohexanol [*583-59-5*] is marketed as a mixture of isomeric methylcyclohexanols with a camphor-like odor that is insoluble in water but miscible with all organic solvents (→ Cyclohexanol and Cyclohexanone, Chap. 8.3.). Its dissolution properties are similar to those of cyclohexanol. On account of its fat-dissolving properties, methylcyclohexanol improves the adhesion of paints to articles that cannot be completely degreased before being painted.

3,3,5-Trimethylcyclohexanol is a cycloaliphatic alcohol that is available as a mixture of *cis – trans* isomers [*933-48-2*], [*767-54-4*] or in its *cis* form (→ Cyclohexanol and Cyclohexanone, Chap. 8.4.). Trimethylcyclohexanol is insoluble in water, but miscible in all proportions with most organic solvents. It has an odor reminiscent of menthol and readily dissolves waxes, resins, fats, and oils. Trimethylcyclohexanol is used in the production of special plasticizers. Bis (trimethylcyclohexyl) phthalate is for example a "hard" plasticizer for poly(vinyl chloride) with favorable elasticity and extracting properties. Trimethylcyclohexanol esters of salicylic acid act as UV absorbers and are used in sunscreen oils. The esters of mandelic acid are spasmolytics, the ester of *cis*-trimethylcyclohexanol being said to have a lower toxicity. The esters formed from trimethylcyclohexanol and polychlorophenoxyacetic acids are intermediates in the preparation of synthetic hormones. Trimethylcyclohexanol is used under the name homomenthol as a substitute for camphor and menthol. Trimethylcyclohexanol is used in the rubber industry on account of its ability to coagulate latex. It also acts as an antifoaming agent (e.g., in hydraulic fluids, textile soaps, and emulsion paints). Oxidation of trimethylcyclohexanol with nitric acid yields trimethyladipic acid, a starting product in the production of plasticizers, lubricants, dinitriles, diamines, glycols, and diisocyanates.

Furfuryl alcohol [*98-00-0*] is a colorless liquid that dissolves cellulose nitrate and many other paint binders (→ Furan and Derivatives, Chap. 4.).

Tetrahydrofurfuryl alcohol [*97-99-4*] is a colorless liquid with a weak odor that is miscible with water and organic solvents except aliphatic hydrocarbons. It dissolves cellulose nitrate and acetate, chlorinated rubber, shellac, resin esters, and many other binders (→ Furan and Derivatives, Chap. 7.).

Diacetone alcohol [*123-42-2*] (4-hydroxy-4-methyl-2-pentanone) is an almost odorless ketone alcohol that is weakly acidic as a result of rearrangement to the enol form (→ Acetone, Chap. 11.2.). It is miscible with water and organic solvents except aliphatic hydrocarbons. It acts as a good solvent for cellulose esters and ethers, alcohol-soluble resins, castor oil, and plasticizers. Poly(vinyl acetate) and chlorinated rubber are partially dissolved or swollen. Polystyrene, poly(vinyl chloride), vinyl chloride copolymers, damar resins, resin esters, rubber, bitumen, mineral oils, ketone resins, and maleate resins are insoluble. Diacetone alcohol is used as a high boiler in stoving enamels to improve flow and gloss.

9.7. Ketones

See also → Ketones.

The ketones used as solvents are water clear, highly mobile liquids having a characteristic odor (DIN 53 247). They are highly volatile and chemically very stable. On account of the carbonyl group ketones are hydrogen acceptors and have an outstanding solvency. The lower ketones dissolve polar resins, fats, oils, and less polar substances. The higher ketones have a more pronounced hydrocarbon character and are particularly good solvents for nonpolar resins, polymers, and copolymers.

Acetone [*67-64-1*] (2-propanone, dimethyl ketone) is miscible with water and organic solvents (→ Acetone). It has excellent solvency for cellulose esters and ethers, cellulose ether esters, ketone resins, ketone–formaldehyde resins, vinyl chloride copolymers, polyacrylates, polystyrene, chlorinated rubber, fats, natural and synthetic oils, and resins. Poly(vinyl chloride) and poly (methyl methacrylate) are swollen. Polyacrylonitrile, polyamides, and rubber are insoluble. Acetone is used as a highly volatile solvent in fast-drying cellulose lacquers, in adhesives, as an extracting agent and detergent, as a cellulose nitrate swelling agent in the production of smokeless gun powder, as a solvent in the production of acetate silk and cellulose acetate films. It is a solvent for acetylene in steel cylinders (dissolved acetylene).

Methyl ethyl ketone [*78-93-3*] (MEK, 2-butanone) has a similar solvency to acetone (→ 2-Butanone). Cellulose acetate is only slightly soluble, natural and synthetic waxes are insoluble. Methyl ethyl ketone is often used instead of ethyl acetate in paints (especially in wood paints), although its more pungent odor mitigates against its widespread use.

Methyl propyl ketone [*107-87-9*] (MPK, 2-pentanone) and *methyl isopropyl ketone*

[*563-80-4*] (MIPK, 3-methyl-2-butanone) are comparable to methyl ethyl ketone, but have a somewhat lower solvency (→ Ketones, Chap. 2.1.). They are more expensive than methyl ethyl ketone and are therefore of only minor importance as solvents.

Methyl butyl ketone [*591-78-6*] (MBK, 2-hexanone) is slightly soluble in water, miscible with organic solvents. Being a medium boiler it dissolves cellulose nitrate, vinyl polymers and copolymers, and natural and synthetic resins. It increases the dilutability with non-solvents and diluents. As a paint solvent, methyl butyl ketone is important only in hot-spraying and coil-coating paints. It has a very low MAK value and is recommended as a solvent because it is photochemically inactive and does not, therefore, contribute to the formation of "photosmog."

Methyl isobutyl ketone [*108-10-1*] (MIBK, 4-methyl-2-pentanone) is a colorless liquid with a sweet odor (→ Ketones, Chap. 2.2.). It is partially miscible with water, but is completely miscible with organic solvents.

Methyl isobutyl ketone is a solvent for many natural and synthetic resins such as cellulose nitrate, poly(vinyl acetate), vinyl chloride copolymers, epoxy resins, most acrylic, alkyd, ketone, coumarone and indene resins, aminoplasts, phenoplasts, rubber and chlorinated rubber, colophony, colophony esters, elemi, manila and damara (dewaxed) resins, gum sandarac and copal esters, fats, and oils. Methyl isobutyl ketone has a poor solvency, however, for most poly(vinyl acetate) resins, poly(vinyl chloride), cellulose esters and ethers, shellac, and natural waxes. Methyl isobutyl ketone is used in the paint industry as a universal medium boiler. It imparts good flow and gloss properties to cellulose nitrate lacquers, improves blushing resistance, and permits the production of highly concentrated solutions containing high proportions of inexpensive diluents. In combination with alcohols and aromatics, methyl isobutyl ketone is an important constituent of practically all epoxy resin paint formulations. Being a good solvent for low molecular mass PVC and vinyl chloride copolymers, it is used to prepare low-viscosity solutions with a high dilutability for aromatic hydrocarbons. Methyl isobutyl ketone is used as a medium-boiling component in, for example, the production of stamping varnishes for steel, tinplate sheet metal, or aluminum, as well as for printing PVC films. Methyl isobutyl ketone reduces the viscosity of alkyd resin paints and is used in all types of acrylic paints. It has proved very useful as an anhydrous and hydroxyl group-free solvent in polyurethane paints.

Methyl isobutyl ketone is advantageously used for the preparation of pesticide concentrates, including natural (pyrethrin) and synthetic insecticides and herbicides. It is a versatile extracting agent and is used to dewax mineral oils, and in the refining of tall oil, and the purification of stearic acid. Methyl isobutyl ketone is a particularly important extracting agent for metals and metal chlorides as well as for separating niobium and tantalum ores. It plays an important role in the purification and extraction of penicillin and other antibiotics. Methyl isobutyl ketone is used for the concentration of aqueous formaldehyde solutions by azeotropic distillation under reduced pressure. Isopropyl alcohol can be freed from ethanol or butanol by extractive distillation with methyl isobutyl ketone. Finally, a mixture of equal parts of methyl isobutyl ketone, ethyl acetate, and gasoline (*bp* 74.5 –79.5 °C) is used as a denaturing agent for ethanol.

Methyl isobutyl ketone has become increasingly important as a chemical feedstock. The oxime is used as an antiskinning agent in oxidatively drying paints, and the peroxide is used to cure unsaturated polyester resins. Methyl isobutyl ketone is an intermediate in the preparation of imidazolones and acetal esters. It is also used to block amines (ketimines) that act as moisture-initiated latent epoxy resin hardeners. Finally, alkaline autocondensation at elevated temperature yields higher molecular mass products that are employed as fragrances and perfumes.

Methyl amyl ketone [*110-43-0*] (MAK, 2-heptanone) and *methyl isoamyl ketone* [*110-12-3*] (MIAK, 5-methyl-2-hexanone) are high boilers with good solvency (→ Ketones, Chap. 2.5.). They are comparable to methyl isobutyl ketone as regards dissolution properties.

Ethyl amyl ketone [*541-85-5*] (EAK, 5-methyl-3-heptanone) is insoluble in water, but miscible with organic solvents (→ Ketones, Chap. 3.5.). It is a high boiler with good solvency and improves the flow properties of paints.

Dipropyl ketone [*110-43-0*] (DPK, 4-heptanone) is a high boiler with a sweet odor.

Diisopropyl ketone [*565-80-0*] (DIPK, 2,4-dimethyl-3-pentanone) is a high-boiling solvent used in the production of cellulose nitrate emulsion lacquers for coating leather, chlorinated rubber paints, and diluents for poly(vinyl chloride) organosols [268] (→ Ketones, Chap. 3.2.).

Diisobutyl ketone (DIBK) is a colorless, low-viscosity liquid, consisting of a mixture of two isomeric dimethyl heptanones (2,6-dimethyl-4-heptanone [*108-83-8*] and 2,4-dimethyl-6-heptanone [*19549-80-5*]) (→ Ketones, Chap. 3.3.). It is practically immiscible with water, but is miscible in all proportions with all common organic solvents. It is a high boiler with good solvency properties for cellulose nitrate, vinyl resins, waxes, and many natural and synthetic resins.

Diisobutyl ketone is used as a high boiler in cellulose nitrate lacquers. It acts as a leveling agent and has outstanding properties as a solvent for cellulose nitrate emulsion lacquers used to coat leather. Diisobutyl ketone is used as a solvent for vinyl chloride copolymers and as a dispersant for poly(vinyl chloride) organosols. It also serves as an extracting agent and feedstock for chemical syntheses.

Mesityl oxide [*141-79-7*] (4-methyl-4-penten-2-one) has a strong, peppermint-like odor (→ Ketones, Chap. 5.3.). It is insoluble in water, but miscible with organic solvents. It has a good solvency for cellulose nitrate, vinyl chloride and vinyl ether copolymers, polyacrylates, and many resins.

Cyclohexanone [*108-94-1*] (Hexanone) is insoluble in water, miscible with organic solvents (→ Cyclohexanol and Cyclohexanone). It is a high boiler with very good solvency for cellulose nitrate, cellulose ethers and esters, colophony, shellac, alkyd resins, natural and synthetic resins, chlorinated rubber, rubber, vinyl polymers and copolymers, polystyrene, ketone and ketone–formaldehyde resins, fats, oils, waxes, blown oils, and bitumen. It is used in paints and leather lacquers [269].

Methylcyclohexanone [*1331-22-2*] is an industrial isomer mixture. It is comparable to cyclohexanone as regards solvency and miscibility, but does not dissolve cellulose acetate (→ Cyclohexanol and Cyclohexanone, Chap. 8.3.).

Dimethylcyclohexanone is an industrial, high-boiling mixture of *cis* and *trans* isomers; [*54396-54-2*] and [*54396-55-3*], respectively. It is comparable to methylcyclohexanone as regards solvency and miscibility.

Trimethylcyclohexanone [*873-94-9*] (TMC-one, 3,3,5-trimethylcyclohexanone) is a saturated cyclic ketone (→ Cyclohexanol and Cyclohexanone, Chap. 8.4.). It is a colorless high boiler with an aromatic odor reminiscent of menthol. Trimethylcyclohexanone is only moderately miscible with water, but is miscible in all proportions with all organic solvents. It is chemically closely related to isophorone. Trimethylcyclohexanone dissolves cellulose nitrate, low molecular mass PVC grades, poly(vinyl acetate), vinyl chloride–vinylacetate copolymers, chlorinated rubber, alkyd resins, unsaturated polyester resins, epoxy resins, acrylic resins, etc.

In paints it is used as a leveling agent for air- and oven-drying systems. It counteracts bubble and crater formation and improves flow and gloss. Vinyl paints based on low molecular mass poly(vinyl chloride) or vinyl chloride copolymers that are used to coat metals or plastics, exhibit good storage stability without gelling when trimethylcyclohexanone is added, even with high contents of diluents. Trimethylcyclohexanone is used in combination with suitable diluents as a temporary plasticizer in the processing of poly(vinyl chloride). In pastes consisting of poly(vinyl chloride) and plasticizers it serves as a thinner with a low gelling tendency. Trimethylcyclohexanone is used as a solvent in cellulose nitrate emulsion lacquers for coating leather and as a cosolvent for pesticide formulations. It is a starting material for various syntheses. Oxidation yields peroxides which are used as catalysts for the polymerization of unsaturated compounds with a long pot life. Trimethylcyclohexanone oxime prevents skin formation in air-drying paints. It can be used in the synthesis of polymers such as polyamides, polyesters, or polyurethanes. A substituted cyclic monoamine (3,3,5-trimethylcyclohexylamine) is formed by aminating hydrogenation, which is used as an anticorrosive agent, in cosmetic articles, or for special chemical syntheses.

Isophorone [269], [270] is an unsaturated cyclic ketone. It consists of α-isophorone [*78-59-1*] (3,5,5-trimethyl-2-cyclohexen-1-one), which contains about 1 – 3 % of the isomer β-isophorone [*471-01-2*] (3,5,5-trimethyl-

3-cyclohexen-1-one) (→ Ketones, Chap. 5.4.). Isophorone is a stable, water-white liquid with a mild odor that is miscible in all proportions with organic solvents. It dissolves many natural and synthetic resins and polymers, such as poly(vinyl chloride) and vinyl chloride copolymers, poly-(vinyl acetate), polyacrylates, polymethacrylates, polystyrene, chlorinated rubber, alkyd resins, saturated and unsaturated polyesters, epoxy resins, cellulose nitrate, cellulose ethers and esters, damar resin (dewaxed), kauri, waxes, fats, oils, phenol--, melamine--, and urea–formaldehyde resins, as well as plant protection agents. However, isophorone does not dissolve polyethylene, polypropylene, polyamides, polyurethanes, polyureas, polyacrylonitrile, polycarbonate, cellulose triacetate, or sandarac.

Isophorone is used in the paint industry as a high-boiling solvent for physically- and oven-drying paints. It improves flow, gloss, adhesion, and the wetting of the substrate and pigments.

Isophorone is used as a high-boiling solvent in cellulose nitrate emulsion lacquers for coating leather. In printing inks for plastic substrates that are based on vinyl resins isophorone reduces the viscosity at high solids contents, it also improves the adhesion due to its swelling ability for poly(vinyl chloride) and other plastics. In the formulation of plant protection agents isophorone is used as a solvent having a high dilutability, good emulsifiability, and good emulsion stability. It is employed in adhesives on account of its universal solvency for many thermoplastic polymers. Isophorone is used in admixture with low-boiling compounds particularly for bonding articles made of poly(vinyl chloride) or polystyrene. The special chemical structure of isophorone with its keto group and an olefinic double bond means that isophorone can be used for the synthesis of a large number of compounds and products including solvents, plasticizers, glycols, amines, diamines, diisocyanates, and dicarboxylic acids.

9.8. Esters

See also → Esters, Organic.

Esters are clear liquids that often have a pleasant, fruity odor. They are neutral and very stable, but can be hydrolyzed on heating in the presence of strong acids or bases. They are less polar than the corresponding alcohols. Esters have an extremely good solvency for polar substances; with increasing carbon chain length in the alcohol and acid groups the solvency for polar products decreases, but increases for less polar products. Lower esters are partially water-soluble. The acetates are the most important ester solvents for the paint industry (DIN 53 246). Formates are used less because they have a more powerful physiology and are more readily hydrolyzed. Propionates, butyrates, and isobutyrates are of minor importance on account of their intense fruity odor. Certain esters of glycolic and lactic acids, as well as dimethyl esters of dicarboxylic acid mixtures are also important.

Methyl formate [*107-31-3*] is a colorless, water-soluble liquid. It dissolves cellulose nitrate, cellulose acetate, oils, and fatty acids.

Ethyl formate [*109-94-4*] is partially miscible with water. It dissolves cellulose nitrate, cellulose acetate, and many resins.

Butyl formate [*592-84-7*] is slightly soluble in water. It dissolves cellulose nitrate, fats, oils, many polymers, and chlorinated rubber. It does not, however, dissolve cellulose acetate.

Isobutyl formate [*542-55-2*] has dissolution properties similar to those of butyl formate. It is a constituent of commercially available mixed solvents for paints.

Methyl acetate [*79-20-9*] is partially miscible with water and readily miscible with most organic solvents. It has a good solvency for cellulose esters and ethers, colophony, urea--, melamine--, and phenol–formaldehyde resins, poly(vinyl acetate), alkyd resins, ketone resins, and other resins. It does not dissolve shellac, damar resin, copal, or poly(vinyl chloride). Methyl acetate is used as a highly volatile solvent by itself or mixed with alcohols or other esters to reduce the viscosity of paints, as a substitute for acetone for adhesives, and as a softening agent for stiff shoe caps in the shoe industry.

Ethyl acetate [*79-20-9*] is a colorless, neutral liquid that is partially miscible with water and has a pleasant, fruity odor. It has a good solvency for cellulose nitrate, cellulose ethers, chlorinated rubber, poly(vinyl acetate), vinyl chloride copolymers, polyacrylates, polystyrene, fats, oils, and many natural and synthetic resins (alkyd resins, saturated polyesters,

ketone resins). Cellulose acetate is, however, dissolved only in the presence of small amounts of ethanol. Poly(vinyl chloride) is insoluble.

Ethyl acetate is one of the most important solvents for quick-drying paints (e.g., wood paints based on cellulose nitrate). It is also commonly employed in isocyanate paints. It increases the dilutability with non-solvents and diluents. Ethyl acetate is also used in adhesives, leather lacquers, leather impregnating agents, as an extracting and cleansing agent, as a fragrance in the perfume industry, in polishes, in the production of glazed paper and transparent paper, and in organic syntheses.

Propyl acetate [*109-60-4*] and *isopropyl acetate* [*108-21-4*] [271] are only slightly soluble in water. They are less volatile than ethyl acetate and have an improved solubility for less polar resins. Cellulose acetate is insoluble.

Butyl acetate [*123-86-4*] is a colorless, neutral, water-immiscible liquid with a pleasant, fruity odor. It has a good solvency for cellulose nitrate, cellulose ethers, chlorinated rubber, postchlorinated poly(vinyl chloride), poly(vinyl acetate), polyacrylates, polymethacrylates, vinyl chloride copolymers, polystyrene, natural and synthetic resins, alkyd resins, fats, and oils. Cellulose acetate is insoluble.

Butyl acetate is the most important moderately volatile solvent for the paint industry. Its volatility is sufficiently high for it to evaporate rapidly from the paint film, but is low enough to prevent cratering, blushing, and flow disorders. The solvency of butyl acetate is increased considerably by adding butanol. Butyl acetate is generally used together with aromatic hydrocarbons in paint formulations. The product has a low intrinsic viscosity and is therefore suitable as an auxiliary solvent in high-solids paints; it is also the most widely used solvent in polyurethane paints.

Isobutyl acetate [*110-19-0*] is a colorless, neutral liquid with a pleasant, fruity odor. It is miscible with organic solvents, but immiscible with water. Isobutyl acetate has a high solvency for cellulose nitrate, colophony, damar resin, ketone and ketone–formaldehyde resins, maleate resins, urea and melamine resins, and phenolic and alkyd resins. Polymers such as polystyrene, poly(vinyl ethers), poly(vinyl acetate), polyacrylates, chlorinated rubber and vinyl chloride copolymers, as well as fats, greases, and oils are readily dissolved. Postchlorinated poly(vinyl chloride) is less soluble. Polyisobutene, cellulose ethers, polymethacrylates, poly(vinyl butyrals), natural rubber, and manila copal are swollen. Shellac, cellulose acetate, cellulose acetobutyrate, poly(vinyl chloride), and poly(vinyl formal) are insoluble.

Isobutyl acetate has a higher volatility and lower flash point than butyl acetate, and is therefore used in the production of quick-drying cellulose nitrate lacquers. On account of its low water content it can be used as a solvent and diluent for polyurethane paints. Isobutyl acetate is used as a viscosity-reducing auxiliary solvent in low-solvent paints.

sec-Butyl acetate [*105-46-4*] is comparable to isobutyl acetate, but is highly volatile and less thinnable. It is of only minor importance as a solvent.

Amyl acetate [*628-63-7*] and *isoamyl acetate* [*123-92-2*] are comparable to butyl acetate as regards solvency and dilutability, but are less volatile. They have a rather intense odor and are used in nail polishes.

2-Ethylhexyl acetate [*103-09-3*] is a high-boiling solvent with good solvency properties. It is used in stoving enamels.

Octyl acetate is a high-boiling mixture of isomers that is used in stoving enamels.

Nonyl acetate is a high-boiling mixture of isomers that is used in stoving enamels.

Hexyl acetate [*142-92-7*] has a good solvency for resins, polymers, cellulose derivatives, fats, and oils. Cellulose acetate is insoluble. Small amounts of hexyl acetate are used as a leveling agent in paints.

Cyclohexyl acetate [*622-45-7*] is very slightly miscible with water, but completely miscible with common organic solvents. Its solvency properties are comparable to those of amyl acetate. Cyclohexyl acetate dissolves oils, fats, resins, waxes, cellulose nitrate, cellulose tripropionate and acetobutyrate, alkyd resins, unsaturated and saturated polyester resins, phenolic resins and aminoplasts, poly(vinyl chloride), vinyl chloride copolymers, poly(vinyl acetate), poly(vinyl ethers), epoxy resins, and acrylic resins, basic dyes, blown oils, crude rubber, metallic soaps, shellac, and bitumen.

In metal paints cyclohexyl acetate improves wetting and flow properties and prevents blistering. On account of its low water miscibility

and high surface tension, cyclohexyl acetate promotes the formation of stable emulsions when used as a cosolvent (e.g., in pest control agents and cellulose nitrate emulsion lacquers for coating leather). It is used for the production of pigment pastes on account of its good solvency for basic dyes. It is also employed as a fragrance component in the perfume industry.

Benzyl acetate [*140-11-4*] is an aromatic smelling, water-insoluble liquid. It is used as a high boiler in printing inks.

Methyl glycol acetate [*110-49-6*] (2-methoxyethyl acetate) is a neutral liquid with a mild odor. It is miscible with water and organic solvents and has a good solvency for many natural and synthetic resins. It does not dissolve poly(vinyl chloride), rubber, polyisobutene, poly(vinyl butyrate), shellac, and damar resin. Methyl glycol acetate imparts good flow properties and high dilutability to paints and prevents blistering and blushing. On account of its teratogenic properties it is being replaced by other solvents in paints.

Ethyl glycol acetate [*111-15-9*] (2-ethoxyethyl acetate) is a colorless, neutral liquid that is partially miscible with water and readily miscible with organic solvents. It has a slight odor. Ethyl glycol acetate dissolves many natural and synthetic resins. Cellulose acetate is swollen. Poly(vinyl chloride), rubber, and polyisobutene are insoluble. Its use in paints is restricted on account of its teratogenic properties.

Butyl glycol acetate [*112-07-2*] (2-butoxyethyl acetate) is insoluble in water, but miscible with organic solvents. It is a high boiler that is used to improve flow properties in stoving enamels. It has a good solvency for slightly polar resins and polymers, fats, and oils.

Ethyl diglycol acetate [*111-90-0*] [2-(2-ethoxyethoxy)ethyl acetate] is used as a high boiler to improve the flow properties and prevent blistering in industrial paints, automotive finishes, and coil coating of metals.

Butyl diglycol acetate [*112-17-4*] [2-(2-butoxyethoxy)ethyl acetate] is a high boiler that is used as a temporary plasticizer, as a leveling agent in stoving enamels and in printing inks, as a film-forming auxiliary in acrylate copolymer dispersions, and to reduce the viscosity of poly(vinyl chloride) plastisols.

1-Methoxypropyl acetate [*108-65-6*] and *2-methoxypropyl acetate* are pleasantly smelling liquids that are partially soluble in water, and miscible with organic solvents. Their dissolution properties and evaporation behavior are comparable but not identical to those of ethyl glycol acetate. 1-Methoxypropyl acetate is less toxic than methyl and ethyl glycol acetates; 2-methoxypropyl acetate is teratogenic.

Ethoxypropyl acetate [*54839-24-6*] is an isomeric mixture containing 1-ethoxypropyl acetate as the principal component. It is less toxic than ethyl glycol acetate, with which it shares many physical properties. Ethoxypropyl acetate differs from ethyl glycol acetate however as regards dissolution properties.

3-Methoxybutyl acetate [*4435-53-4*] (butoxyl) is a weakly smelling liquid that is partially soluble in water and miscible with organic solvents. Dissolution properties resemble those of ethyl glycol acetate, but volatility is lower.

Propionate esters are comparable to acetate esters having the same number of carbon atoms as regards dissolution, mixing, and evaporation properties. Propionates have a stronger odor than acetates and are therefore of only limited importance as solvents.

Ethyl 3-ethoxypropionate [*763-69-9*] is a high-boiling solvent that is a less toxic alternative to ethyl glycol acetate, with a comparable property spectrum.

Butyrate esters are comparable to acetate esters with the same number of carbon atoms as regards dissolution, mixing, and evaporation properties, but have a more intense odor, in some cases reminiscent of pineapple or banana. Butyl butyrate [*109-21-7*] is important as a high-boiling solvent.

Isobutyrate esters are comparable to butyrate esters, but have a slightly lower boiling point and higher volatility. Isobutyl isobutyrate [*97-85-8*] has some importance as a high-boiling solvent.

Ethyl lactate [*97-64-3*] is miscible with water and most organic solvents. It is a slowly evaporating solvent with a mild odor. It dissolves cellulose nitrate, cellulose acetobutyrate, strongly polar resins, and polar polymers, [e.g., poly(vinyl acetate) and polyacrylates]. Ethyl lactate has limited solvency for slightly polar resins, polystyrene, poly(vinyl chloride), and vinyl chloride copolymers. It is used as a powerful high-boiling solvent that can be diluted to a high degree with non-solvents.

Butyl lactate [*138-22-7*] has a good solvency for less polar resins and polymers and good dilutability with aliphatic hydrocarbons. It does not dissolve cellulose acetate. Butyl lactate is a high-boiling solvent that is used in small amounts as a leveling agent.

Butyl glycolate [*7397-62-8*] (butyl hydroxyacetate) is comparable to butyl acetate as regards dissolution properties and evaporation behavior. It has a good solvency for fats, oils, alkyd resins, cellulose nitrate, and many resins. It has a high dilutability with alcohols and aromatic hydrocarbons.

Butyl glycolate reduces the viscosity of alkyd resin paints. It improves the flow properties and gloss in air-drying and oven-drying paints, especially in coil coating and leather lacquers. It is also used to lower the film-forming temperature in poly(vinyl acetate) dispersions.

Dimethyl adipate [*627-93-0*], *glutarate* [*1119-40-0*], and *succinate* [*106-65-0*] are used as a high-boiling mixture with good dissolution properties in stoving enamels, coil-coatings, and in the automobile industry.

Ethylene carbonate [*96-49-1*] is soluble in water and organic solvents and solid at room temperature. It is a high boiler that is insoluble in gasoline and turpentine oil. It has a high solvency for polyacrylonitrile, polyamides, glycol terephthalates, and poly(vinyl chloride). Cellulose nitrate and cellulose acetobutyrate only dissolve in the presence of alcohol or esters. Ethylene carbonate is used for polyacrylonitrile spinning solutions.

Propylene carbonate [*108-32-7*] is liquid at room temperature. It is a high boiler with a solvency resembling that of ethylene carbonate. It dissolves polyacrylonitrile only in combination with ethylene carbonate. Propylene carbonate is used in paints as a high-boiling solvent and film-forming auxiliary, especially in poly(vinyl fluoride) and poly(vinylidene fluoride) systems. It is also employed as an auxiliary in the pigment and dye industry.

Butyrolactone [*96-48-0*] has a mild odor with good dissolution properties. It is a high-boiling solvent that is used to improve properties (→ Butyrolactone).

9.9. Ethers

See also → Ethers, Aliphatic.

Straight-chain aliphatic ethers are water-clear, highly mobile liquids with a high volatility and characteristic, anesthetizing odor. On account of their high volatility they are of only minor importance in paint technology. Ethers are usually used as extracting agents in chemical syntheses and for isolating natural substances. They are highly flammable and form explosive peroxides in the presence of oxygen, especially on exposure to light.

Cycloaliphatic ethers are better solvents for binders and are therefore more important in paint technology.

Ethylene and propylene glycol ethers are extremely good solvents (DIN 55 999). For toxicity reasons monoethylene glycol ethers are being replaced in the paint, dye, and printing ink industries by other solvents, including the corresponding monopropylene glycol ethers.

Diethyl ether [*60-29-7*] (DEE, ethoxyethane, ether) is a clear, colorless, highly mobile liquid with a low boiling point and a characteristic pleasant odor. It is miscible in all proportions with common organic solvents such as alcohols, ketones, hydrocarbons, esters, and ethereal oils. Diethyl ether has an evaporation number of 1.

Diethyl ether is widely used in the chemical industry as a solvent, extracting agent, and reaction medium on account of its chemical resistance in acid and alkaline media, its low boiling point, and favorable dissolution properties (e.g., for organometallic compounds). Diethyl ether mixed with ethanol is an excellent solvent for cellulose nitrate. It is used in large amounts in the gelatinization of cellulose nitrate and preparation of collodion solution. Ether is employed in medicine as an anesthetic, but this use is declining.

Diisopropyl ether [*108-20-3*] is miscible with most organic solvents, but immiscible with ethylene glycol and glycerol. It is only slightly water soluble. Diisopropyl ether is an excellent solvent for natural oils, mineral oils, and waxes. It is used as an extracting agent and reaction medium in chemical and pharmaceutical syntheses.

Dibutyl ether [*142-96-1*] and *di-sec-butyl ether* [*6863-58-7*] are very slightly soluble in water but miscible with organic solvents ex-

cept ethylene glycol and glycerol. They are used as extracting agents and reaction media in the chemical and pharmaceutical industries.

Methyl tert-butyl ether [*1634-04-4*] (MTB) is a combustible, clear, colorless liquid with a low viscosity and characteristic, turpentine-like odor (→ Methyl Tert-Butyl Ether). It is miscible with organic solvents, and is only slightly soluble in water. It forms azeotropic mixtures with some solvents such as methanol. MTB does not undergo autoxidation and, in contrast to other ethers, it does not form peroxides with atmospheric oxygen. It improves the antiknock properties when added to motor gasoline.

Tetrahydrofuran [*109-99-9*] (oxolane, tetramethylene oxide, oxacyclopentane, THF) [272] is a colorless, highly volatile liquid having an outstanding solvency (→ Tetrahydrofuran). Under prolonged exposure to air it tends to form peroxides. Tetrahydrofuran is completely miscible with most organic solvents and water. It forms azeotropic mixtures with a number of solvents. Tetrahydrofuran dissolves practically all plastics apart from certain polyamides (e.g., polyamide 6 and 12) and polytetrafluoroethylene.

Tetrahydrofuran is used alone or mixed with other solvents to prepare poly(vinyl chloride) and polyurethane solutions for coating films, sheets, and textiles (leather imitates). The special surface properties of video, computer, and recording tapes are obtained by using appropriate coatings that are prepared from tetrahydrofuran- containing solutions. Offset printing plates are produced by coating them with light-sensitive compounds that are applied from tetrahydrofuran- containing solutions. Tetrahydrofuran is also highly suitable as a solvent for printing inks based on poly(vinyl chloride). Its outstanding solvency for many plastics makes tetrahydrofuran a particularly suitable solvent for a number of special finishes that are used to coat plastics. Tetrahydrofuran has proved to be especially useful in cold cleansers for eliminating impurities and contaminants that can otherwise only be removed with difficulty.

Tetrahydrofuran is extremely suitable for extracting active agents from natural or synthetic raw materials. Substances can be separated on the basis of the specific solvency properties of tetrahydrofuran [e.g., insoluble polyamides and soluble poly(vinyl chloride)]. On account of its good dissolution properties for poly(vinyl chloride) tetrahydrofuran can be used alone or with other solvents to bond PVC pipes and for solution welding of PVC sealant films. Solutions of PVC in tetrahydrofuran are employed for gap-filling bonding of PVC moldings. Tetrahydrofuran is used as a stripper to remove coats of paint. It also serves as a reaction medium in the preparation of chemical intermediates with organometallic compounds.

1,4-Dioxane [*123-91-1*] has a slight odor reminiscent of butanol, and is miscible with water and organic solvents (→ Dioxane). It is a good solvent for cellulose derivatives, polymers, chlorinated rubber, and resins [273].

Metadioxane [*505-22-6*] (2,2-dimethyl-4-hydroxymethyl-1,3-dioxolane) is a high-boiling solvent with an excellent solvency for resins and polymers. It is used in printing inks, adhesives, and in the textile sector.

9.10. Glycol Ethers

Methyl glycol [*109-86-4*] (2-methoxyethanol, ethylene glycol monomethyl ether) has a slight odor, and is miscible with water and organic solvents except aliphatic hydrocarbons. It is a very good solvent for many natural and synthetic resins. It does not dissolve fats, oils (except castor oil), damar resin, rubber, bitumen, hydrocarbon resins, polystyrene, poly(vinyl chloride), and vinyl chloride copolymers. On account of its teratogenic properties it is being replaced as a solvent in the paint and colorants sector by other solvents or solvent mixtures.

Ethyl glycol [*110-80-5*] (2-ethoxyethanol, ethylene glycol monoethyl ether) is a neutral, colorless liquid with a slight odor, and is miscible in all proportions with water and organic solvents. Ethyl glycol has a high solvency for many natural and synthetic resins and binders. Damar resin, rubber, bitumen, cellulose acetate, polystyrene, and poly(vinyl chloride) are insoluble. Its use as a solvent in the paint and colorants industry is restricted on account of its teratogenic properties.

Propyl glycol [*2807-30-9*] (2-propoxyethanol, ethylene glycol monopropyl ether) and *isopropyl glycol* [*109-59-1*] (2-isopropoxyeth-

anol, ethylene glycol monoisopropyl ether) correspond to ethyl glycol as regards dissolution and miscibility properties, but evaporate more slowly and have a better solvency for slightly polar resins. They are less toxic than ethyl glycol and are therefore increasingly used as alternatives.

Butyl glycol [*111-76-2*] (2-butoxyethanol, ethylene glycol monobutyl ether) is a neutral, colorless liquid with a very weak, pleasant odor. It is miscible with water at room temperature, but has a miscibility gap with water at higher temperatures. It is miscible with organic solvents, and a very good solvent for cellulose nitrate, cellulose ethers, colophony, shellac, chlorinated rubber, polyacrylates, polymethacrylates, alkyd resins, phenol --, area --, and melamine – formaldehyde resins, oils, fats, and synthetic resins. Damar resin, rubber, bitumen, polystyrene, and poly(vinyl chloride) are insoluble. Poly(vinyl acetate), poly(vinyl butyral), postchlorinated poly(vinyl chloride), and cellulose acetate are sparingly soluble or swellable. Butyl glycol is widely used in solvent- containing paint systems to improve the flow and surface quality of coatings. It is increasingly used in water-thinnable paints on account of its solubilizing action.

Methyl diglycol [*111-77-3*] [2-(2-methoxyethoxy)ethanol, diethylene glycol monomethyl ether] is miscible with water and most organic solvents. It is a good solvent for cellulose nitrate, colophony, phenol --, urea --, and melamine–formaldehyde resins, alkyd resins, alcohol-soluble maleate resins, ketone – formaldehyde resins, poly(vinyl butyral), poly(vinyl acetate), poly-(vinyl ethers), epoxy resins, castor oil, linseed oil, and turpentine oil. However, it does not dissolve damar resin, rubber, bitumen, polyisobutylene, polystyrene, waxes, mineral oil, polyacrylates, and polymethacrylates. Small amounts are used as a low-volatility solvent in paints to improve the gloss and flow, in the printing ink industry, and for ballpoint pen pastes.

Ethyl diglycol [*111-90-0*] [2-(2-ethoxyethoxy)ethanol, diethylene glycol monoethyl ether] is a neutral, colorless, practically odorless liquid. It is miscible in all proportions with alcohols, esters, ethers, ketones, and water. It is partially miscible with aromatic and aliphatic hydrocarbons. Ethyl diglycol dissolves many paint resins, such as cellulose nitrate, chlorinated rubber, shellac, colophony, phenolic resins, urea– and melamine – formaldehyde resins, polyacrylates, esters, poly(vinyl acetate), as well as alkyd, ketone, and maleate resins. However, it does not dissolve cellulose acetate, poly(vinyl chloride), vinyl chloride copolymers, polystyrene, polymethacrylates, asphalt, and rubber. On account of its low solidification point and low viscosity at low temperatures, ethyl diglycol is a constituent of many brake fluids. It is used in the paint industry to improve gloss and flow properties, and thereby increase the dilutability. Ethyl diglycol can be used to predissolve alcohol-soluble dyes and is a constituent of many wood stains. This solvent is also used in printing inks and as a cleaning agent in offset printing. It is a constituent of Indian inks and ballpoint pen pastes. Ethyl diglycol phthalate is an excellent plasticizer for vinyl resins. Ethyl diglycol is used as a solvent in the printing and dyeing of textile fibers and fabrics. It also acts as a solubilizer for detergent raw materials in aqueous solutions and in drilling and cutting fluids. In the determination of the saponification value of waxes, resins, and fats, ethyl diglycol has the advantage that it can substantially accelerate saponification because it is a high-boiling solvent.

Butyl diglycol [*112-34-5*] [2-(2-butoxyethoxy)ethanol, diethylene glycol monobutyl ether] is a clear, colorless, neutral liquid with a pleasantly mild odor. It is miscible with water and organic solvents, including aliphatic compounds. Butyl diglycol has a high solvency for cellulose nitrate, cellulose ethers, chlorinated rubber, poly(vinyl acetate), polyacrylates, and some oils, as well as for many synthetic resins, natural resins, and dyes. Polystyrene, poly(vinyl chloride), fats, and most oils are not dissolved.

Butyl diglycol is used as a high-boiling solvent to improve gloss and flow properties. On account of its high evaporation number even additions of $< 5\,\%$ considerably improve paint properties, without noticeably increasing the stoving time. In cellulose nitrate and cellulose ether lacquers even smaller amounts are effective. In emulsion paints and cold-hardening paints butyl diglycol improves the coatability and enhances the surface gloss. On account of its good solvency for resins and dyes, butyl diglycol is used with other low-volatility solvents for coatings

and printing inks. It is used as a solubilizer in mineral oil products.

Ethyl triglycol [*112-50-5*] [2-[2-(ethoxyethoxy) ethoxy]ethanol, triethylene glycol monoethyl ether] is an almost colorless, neutral, mild-smelling liquid with a low hygroscopicity. It is soluble in water and most organic solvents, but is only partially miscible with aromatic and aliphatic hydrocarbons. Ethyl triglycol dissolves cellulose nitrate, shellac, colophony, ketone resins, maleate resins, chlorinated rubber, alkyd resins, and many other paint resins. It does not dissolve cellulose acetate, poly(vinyl chloride), vinyl chloride copolymers, fats, oils, and rubber.

The uses of ethyl triglycol are similar to those of ethyl diglycol. It serves as a solubilizer for incompatible liquids in the production of insecticides, and in the detergent industry in handwash pastes. It is employed in printing inks and, in small amounts, in wood paints to prevent the wood fibers from standing perpendicular of the surface during painting. Ethyl triglycol is used in hydraulic and brake fluids because its viscosity remains largely constant under temperature fluctuations. It is also used as an intermediate in the production of plasticizers.

Butyl triglycol [*143-22-6*] [2-[2-(butoxyethoxy)ethoxy]ethanol, triethylene glycol monobutyl ether] is an almost colorless, neutral, weakly smelling liquid. It is soluble in water and most organic solvents, but is only partially miscible with aromatic and aliphatic solvents. Its dissolution properties are comparable to those of butyl diglycol. Butyl triglycol dissolves cellulose nitrate, cellulose ethers, chlorinated rubber, poly(vinyl acetate), colophony, many other natural and synthetic resins, as well as drying greases and oils. Bitumen, nondrying fats and oils, cellulose acetate, poly(vinyl chloride), and rubber are, however, insoluble. Butyl triglycol serves as a solubilizer for mutually incompatible liquids, and is used in the production of household and metal cleansing agents as well as for wood preservation. It is employed in cutting and hydraulic oils. Butyl triglycol is used in the production of Indian inks, writing inks, and ballpoint pen pastes. In the paint industry it is suitable as a high-boiling solvent in stoving enamels, as a leveling agent, and as an additional solvent in wood paints, in order to prevent wood fibers standing proud from the surface. Butyl triglycol is a useful solvent for the formulation of pharmaceutical and cosmetic preparations and in the leather auxiliaries industry.

Butyl tetraglycol [2-[2-(butoxyethoxyethoxy)-ethoxy]ethanol, tetraethylene glycol monobutyl ether] is a neutral, practically odorless liquid that is soluble in water and most organic solvents, but only partially miscible with aromatic and aliphatic solvents. Its properties are comparable to those of butyl diglycol. Butyl tetraglycol dissolves cellulose nitrate, cellulose ethers, chlorinated rubber, poly(vinyl acetate), colophony, and many other natural and synthetic resins. However, it does not dissolve nondrying fats and oils, cellulose acetate, poly(vinyl chloride), and rubber. Butyl tetraglycol is used as a solubilizer on account of its good solvency for organic solvents, water, and natural and synthetic resins. It is also a constituent of hydraulic and brake fluids.

Diethylene glycol dimethyl ether [*111-96-6*] is a hygroscopic, weakly smelling liquid that is miscible with water and organic solvents. It dissolves vinyl chloride copolymers, postchlorinated poly (vinyl chloride), polymethacrylate, polystyrene, polychloroprene, and cellulose acetate. It does not dissolve rubber and polyethylene, and swells poly(vinyl chloride). Diglycol dimethyl ether also dissolves sodium borohydride, other covalent inorganic metal compounds, and sodium. It is used as a solvent in chemical reactions involving metals and organometallic compounds.

Methoxypropanol [*107-98-2*] (1-methoxy-2-propanol). Commercially available methoxypropanol also contains small amounts of 2-methoxy-1-propanol. It is a colorless, neutral liquid with a weak pleasant odor that is miscible in all proportions with water and organic solvents. The properties of methoxypropanol are largely comparable to those of ethyl glycol [274]. It has a somewhat higher volatility and, as a solvent component in paints and printing inks, improves the wetting of some pigments and colorants. It has a good solvency for cellulose nitrate, cellulose ethers, chlorinated rubber, poly(vinyl acetate), poly(vinyl butyral), ketone and ketone–formaldehyde resins, shellac, colophony, phenol--, melamine--, and urea–formaldehyde resins, alkyd resins, polyacrylates, polymethacrylates, castor oil, linseed oil,

and some vinyl chloride copolymers. However, it does not dissolve copal resins, rubber, bitumen, cellulose acetate, polystyrene, poly(vinyl alcohol), and poly(vinyl chloride). On account of its mild, pleasant odor, methoxypropanol is extremely suitable for cellulose nitrate lacquers used to coat wood, paper, and metals. Being a moderately volatile solvent, it improves paint penetration, flow properties, and the gloss of paint coats; it also prevents blushing and formation of fish eyes and blisters. Addition of methoxypropanol does not delay the drying of paint systems. It increases the dilutability with inexpensive non-solvents, reduces and stabilizes the paint viscosity, and improves the coatability of paints. Methoxypropanol only slowly swells dry paint coats, so that it facilitates painting over previous coats. It is used as a solvent in printing inks (flexographic, gravure packaging printing) to improve their gloss and flow. On account of its good dissolution properties for dyes and ability to wet pigments, it intensifies the color of the inks.

Ethoxypropanol is a mixture of 1-ethoxy-2-propanol [*1569-02-4*] and 2-ethoxy-1-propanol. Its physical properties are very similar to those of ethyl glycol, but is a less toxic alternative.

Isopropoxypropanol consists of a mixture of 1-isopropoxy-2-propanol [*3944-36-3*] and 2-isopropoxy-1-propanol [*3944-37-4*]. It is a weakly smelling liquid with a relatively high boiling that is miscible with water and organic solvents. Its dissolution properties, evaporation properties, and uses are similar to those of butyl glycol. It is an effective solubilizer and used as an auxiliary solvent in high-solids and waterborne paints.

Isobutoxypropanol is a mixture of 1-isobutoxy-2-propanol [*26447-43-8*] and 2-isobutoxy-1-propanol. It is a weakly smelling liquid with a high boiling point that is miscible with water and organic solvents. Its dissolution properties, evaporation properties, and uses are similar to those of butyl glycol. Isobutoxypropanol is a highly effective solubilizer and is used as an auxiliary solvent in high-solids and waterborne paints. Propylene glycol *tert*-butyl ether is similar to isobutoxypropanol [275].

Methyl dipropylene glycol [*34590-94-8*] [2-(2-methoxypropoxy)propanol, dipropylene glycol methyl ether] (isomer mixture) is a high-boiling solvent. It is waterclear and miscible in all proportions with water, and has only a slight intrinsic odor. It readily dissolves cellulose nitrate, cellulose esters and ethers, as well as many natural and synthetic resins. On account of its slow evaporation rate small amounts of methyl dipropylene glycol are added to paints to control their flow properties, evaporation rate, and dilutability with non-solvents. It improves the coatability of cellulose nitrate lacquers, chlorinated rubber paints, and alkyd resin paints. Methyl dipropylene glycol is also suitable as a film-forming auxiliary in dispersions and as an auxiliary solvent in waterborne paints. It is highly effective as a solubilizer in cleansing agents and as a solvent in writing inks, pastes, and printing inks.

Methoxybutanol [*30677-36-2*] (3-methoxy-1-butanol) is a mild-smelling liquid that is miscible with water and organic solvents. It has a good solvency for cellulose nitrate, cellulose esters, poly(vinyl butyral), ketone resins, phenol --, urea --, and melamine – formaldehyde resins, alkyd and maleate resins, plasticizers, fats, and drying oils. However, it does not dissolve mineral oils, waxes, rubber, chlorinated rubber, cellulose acetate, polyisobutene, polystyrene, poly(vinyl chloride), and vinyl chloride copolymers. It is used like butanol as a slow-evaporating solvent.

9.11. Miscellaneous Solvents [276]

Methylal [*109-87-5*] (formaldehyde dimethyl acetal, dimethoxymethane) is a pleasantly smelling, extremely volatile solvent. It dissolves polystyrene, poly(vinyl acetate), vinyl chloride copolymers, acrylates, methacrylates, and synthetic and natural resins. It is used in paints and lacquers, adhesives, and aerosols. Also it is an important extracting agent for natural substances, essences, and oils. It serves as a reaction medium in the chemical industry and as an intermediate for chemical syntheses.

Dimethylacetal [*534-15-6*] (acetaldehyde dimethyl acetal, 1,1-dimethoxyethane) is a neutral liquid that is miscible with water and organic solvents. It dissolves cellulose nitrate, cellulose ethers, poly(vinyl acetate), polyacrylates, polymethacrylates, poly(vinyl ethers), some vinyl chloride copolymers, and synthetic and

natural resins. It does not dissolve poly(vinyl chloride), polystyrene, chlorinated rubber, or cellulose acetate. It is used in the production of paints, adhesives, and shoe- cap stiffeners.

N,N-Dimethylformamide [*68-12-2*] (DMF) [277] is miscible with water and organic solvents except aliphatic hydrocarbons (→ Formamides). It is a good high-boiling solvent for cellulose esters, cellulose ethers, poly(vinyl chloride), vinyl chloride copolymers, poly(vinyl acetate), polyacrylonitrile, polystyrene, chlorinated rubber, polyacrylates, ketone resins, and phenolic resins. Alkyd resins and resin esters are partially soluble. Dimethylformamide does not dissolve polyethylene, polypropylene, urea–formaldehyde resins, rubber, or polyamides. It is used as a solvent in printing inks, for polyacrylonitrile spinning solutions [278], and as a solvent in the synthesis of acetylene.

N,N-Dimethylacetamide [*127-19-5*] (DMA) is miscible with water and organic solvents. It has a very good solvency for many resins and polymers. It is used in the production of acrylic fibers, films, sheets, and coatings, and as a reaction medium and intermediate in organic syntheses.

Dimethyl sulfoxide [*67-68-5*] (DMSO) is a clear, colorless liquid that is miscible with water and organic solvents except aliphatic hydrocarbons (→ Sulfones and Sulfoxides). It is a good high-boiling solvent for cellulose esters, cellulose ethers, poly(vinyl acetate), polyacrylates, vinyl chloride copolymers, polyacrylonitrile, chlorinated rubber, and many resins. It is used in polyacrylonitrile spinning solutions, in paint-removal agents, as a film-forming auxiliary in dispersions, as an extracting agent, and as a reaction medium in organic syntheses [279], [280].

Tetramethylene sulfone [*126-33-0*] (sulfolane) is a waterclear liquid that is miscible with water and most organic solvents except aliphatic hydrocarbons. It dissolves cellulose acetate, cellulose nitrate, cellulose ethers, poly(vinyl chloride), vinyl chloride copolymers, poly(vinyl acetate), polystyrene, and many resins. Its solvency is lower than that of dimethyl sulfoxide. Tetramethylene sulfone is of only minor importance as a solvent.

Carbon disulfide [*75-15-0*] is a clear, highly mobile liquid that is highly refractive and very flammable (→ Carbon Disulfide). It is insoluble in water, but generally miscible with organic solvents. It has a good solvency for fats, oils, waxes, rubber, and many resins. Cellulose esters are insoluble, vinyl polymers dissolve only in mixtures with ketones or esters. Carbon disulfide is used in the production of silk, rubber solutions, and poly(vinyl chloride) spinning solutions. It is of no importance for paints, plastics solutions, or as a fat-extracting agent.

2-Nitropropane [*79-46-9*] is a colorless, nonhygroscopic liquid with a mild odor (→ Nitro Compounds, Aliphatic, Chap. 6.4.). It dissolves cellulose nitrate, cellulose ethers, alkyd resins, chlorinated rubber, poly(vinyl acetate), vinyl chloride copolymers. Poly(vinyl chloride), colophony, polyacrylonitrile, waxes, rubber, and shellac are insoluble. It is used as a cosolvent in paints to improve pigment wetting, flow properties, and electrostatic processing; it also reduces the paint drying time. 2-Nitropropane is classified as carcinogenic.

N-Methylpyrrolidone [*872-50-4*] (NMP) has a fairly mild, amine-like odor, and is miscible with water and most organic solvents (→ 2-Pyrrolidone, Chap. 3.). It has a good solvency for cellulose ethers, butadiene – acrylonitrile copolymers, polyamides, polyacrylonitrile, waxes, polyacrylates, vinyl chloride copolymers, and epoxy resins. It is used in paint removers and stripping paints to reduce paint viscosity, and to improve the wettability of paint systems. *N*-Methylpyrrolidone is also employed for the extraction of hydrocarbons, and as a solvent in the synthesis of acetylene.

Hexamethylphosphoric triamide [*680-31-9*] [281] is a basic, highly polar, nonflammable solvent with a very good solvency. Its dissolution properties are comparable to those of dimethyl sulfoxide and dimethylformamide. It is used for selective dissolution of organic compounds, as a reaction medium in chemical syntheses, as an antifreeze, and as an antistatic agent.

1,3-Dimethyl-2-imidazolidinone [*80-73-9*] (DMI) is a colorless, high-boiling, highly polar aprotic solvent of low toxicity with good chemical and thermal stability. It is miscible with water and most organic solvents. It is used as a reaction medium in preparative organic chemistry, in the production of polyamides, as a solvent in the production of dyes, agrochemicals, and pharmaceuticals and in the electronics industry.

It is also a constituent of nail-polish, ballpoint pen pastes, and coating compositions.

10. References

General References

1. T. H. Durrans, O. Merz: *Lösungsmittel und Weichmachungsmittel,* W. Knapp Verlag, Halle 1933. T. H. Durrans: *Solvents,* Chapman and Hall, London 1971. H. Gnamm, W. Sommer: *Die Lösungs- und Weichmachungsmittel,* Wissenschaftliche Verlagsgesell., Stuttgart 1958. O. Fuchs: *Gnamm-Fuchs, Lösungsmittel und Weichmachungsmittel,* 8th ed., **vols. 1, 2,** Wissenschaftliche Verlagsgesell., Stuttgart 1980. L. Scheflan, M. B. Jacobs: *The Handbook of Solvents,* van Nostrand, Toronto 1953. I. Mellan: *Handbook of Solvents,* Reinhold Publ. Co., New York 1957. I. Mellan: *Source Book of Industrial Solvents,* **vol. 2, 3,** Reinhold Publ. Co., New York 1957, 1959. I. Mellan: *Industrial Solvents Handbook,* 2nd ed., Noyes Data Corp., Park Ridge, N.Y. 1977. E. W. Flick: *Industrial Solvents Handbook,* 3rd ed., Noyes Data Corp., Park Ridge 1985. E. W. Flick: *Industrial Solvents Handbook,* 4th ed., Noyes Data Corp., Park Ridge 1991. O. Jordan: *Chemische Technologie der Lösemittel,* Springer Verlag, Berlin 1932. A. Krause: "Lösungsmittel" *Handbuch der Nitrocelluloselacke,* **part 1,** Wilhelm Pansegrau Verlag, Berlin 1955. H. Guinot: *Solvants et Plastifiants,* Dunod, Paris 1948. I.G. Farbenindustrie: *Lösungs- und Weichmachungsmittel,* Frankfurt 1937. E. Schwarz: *Wie setzen wir die Verluste an flüchtigen Lösungsmitteln herab?,* Allgemeiner Industrieverlag Knorre & Co., Berlin 1943. F. Fritz: *Die wichtigsten Lösungs- und Weichmachungsmittel,* VEB Verlag Technik, Berlin 1957. J. A. Riddick, W. B. Bunger: "Organic Solvents," *Techniques of Chemistry,* **vol. 2,** Wiley-Interscience, London 1970. M. R. J. Dack: "Solutions and Solubilities", *Techniques of Chemistry,* vol. 8, J. Wiley & Sons, London 1976. E. Krell: *Einführung in die Trennverfahren,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1975. K. Thinius: *Anleitung zur Analyse der Lösungsmittel,* 2nd ed., J. A. Barth-Verlag, Leipzig 1957. H. Haase: *Statische Elektrizität als Gefahr,* Verlag Chemie, Weinheim 1968. R. W. Tess: *Solvents Theory and Practice,* American Chemical Society, Washington 1973. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* **vol. 3,** Verlag W. A. Colomb, Berlin-Oberschwandorf 1976, pp. 17 – 129. A. F. M. Barton: *Handbook of Solubility Parameters and Other Cohesion Parameters,* CRC Press, Boca Raton, Fla., 1983. J. Roire: *Les Solvants,* Erec, Puteaux 1989. R. Lambourne: *Paint and Surface Coatings, Theory and Practice,* Ellis Horwood, Chichester 1987, pp. 195 –211. M. Ash, I. Ash: *Dispersants, Solvents and Solubilizers,* **vol. 2,** Edward Arnold, London 1988. K. Dören, W. Freitag, D. Stoye: *Wasserlacke – Umweltschonende Alternative für Beschichtungen,* Verlag TüV Rheinland, Köln 1992, pp. 54 – 56, 107 – 117.

Specific References

2. J. Laakmann, G. Hasselkuß, *Ind. Lackierbetr.* **58** (1990) no. 5, 185 – 186.
3. C. Reichardt: *Solvents and Solvent Effects in Organic Chemistry,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1988.
4. O. Fuchs, *Farbe Lack* **71** (1965) no. 2, 104 – 112.
5. O. Fuchs, *CLB Chem. Labor Betr.* **36** (1985) no. 8, 378 – 380.
6. I. Vavruch, *CLB Chem. Labor Betr.* **35** (1984) no. 11, 536 – 543.
7. W. H. Keesom, *Phys. Z.* **22** (1921) 129, 643; **23** (1922) 225.
8. P. Debye, *Phys. Z.* **21** (1920) 178; **22** (1921) 302.
9. F. London, *Z. Phys. Chem. Abt. B* **11** (1931) 222; *Z. Phys.* **63** (1930) 245.
10. G. C. Pimentel, A. L. McClellan: *The Hydrogen Bond,* W. H. Freeman, London 1960.
11. W. A. P. Luck, *Naturwissenschaften* **54** (1967) 601.
12. A. A. Orr, *J. Paint Technol.* **47** (1975) no. 607, 45 – 49.
13. P. L. Huyskens, M. C. Haulait-Prison, L. D. Brandts Buys, X. M. van der Borght, *J. Coat. Technol.* **57 (n 724)** (1985) 57 – 67.
14. P. L. Huyskens, M. C. Haulait-Pirson, X. van der Borght, *Farbe Lack* **92** (1986) no. 12, 1145 – 1149.
15. B. Ziolkowsky, *Seifen öle Fette Wachse* **96** (1970) no. 22, 801 – 806.
16. M. L. Huggins, *J. Paint Technol.* **41** (1969) 509.

17. J. H. Hildebrand, R. L. Scott: *The Solubility of Non-Electrolytes,* 3rd ed., Reinhold Publ. Co., New York 1950.
18. H. Hildebrand, *J. Am. Chem. Soc.* **51** (1929) 66; *Nature London* **168** (1951) 868.
19. E. A. Guggenheim, *Proc. R. Soc. London, Ser. A* **183** (1944) 203, 213.
20. H. Mauser, G. Kortüm, *Z. Naturforsch. A* **10 A** (1955) 317.
21. K. L. Hoy, *J. Paint. Technol.* **42** (1970) no. 541, 76–118.
22. G. Scatchard, *Chem. Rev.* **8** (1931) 321.
23. J. H. Hildebrand, *J. Am. Chem. Soc.* **57** (1935) 866.
24. A. E. Rheineck, K. F. Lin, *J. Paint Technol.* **40** (1968) no. 527, 613–616.
25. H. Ahmad, M. Yaseen, *J. Oil Colour Chem. Assoc.* **60** (1977) 99–103.
26. H. Ahmad, *J. Oil Colour Chem. Assoc.* **63** (1980) 2.
27. C. M. Hansen, *J. Paint Technol.* **39** (1967) 104.
28. Shell Chemical, Technical Bulletin, "Solvent System Design," Rotterdam 1978, "Solubility Parameters," Rotterdam 1978, "Solvent Power," Rotterdam 1979.
29. E. B. Bagley, T. P. Nelson, J. M. Scigliano, *J. Paint Technol.* **43** (1971) no. 555, 35–42.
30. A. Jayasri, M. Yaseen, *J. Coat. Technol.* **52** (1980) no. 667, 41–45; *J. Oil Colour Chem. Assoc.* **63** (1980) 61–69.
31. C. M. Hansen, *Ind. Eng. Chem. Prod. Res. Dev.* **8** (1969) 2.
32. C. M. Hansen, *J. Paint Technol.* **42** (1970) 660.
33. J. H. Meyer zu Bexten, *Farbe Lack* **78** (1972) no. 9, 813–822.
34. B. Marti, *Schweiz. Arch. Angew. Wiss. Tech.* **33** (1967) no. 10, 297–304.
35. R. C. Nelson, V. F. Figurelli, J. G. Walsham, G. D. Edwards, *J. Paint Technol.* **42** (1970) no. 550, 644–652.
36. J. P. Teas, *J. Paint Technol.* **41** (1969) no. 516, 19–25.
37. ASTM D 3132–72,Standard Test Method for Solubility Range of Resins and Polymers, reapproved 1976.
38. D. W. van Krevelen: *Properties of Polymers, Cohesive Properties and Solubility,* Elsevier, Amsterdam 1976, pp. 129–159.
39. C. M. Kok, A. Rudin, *J. Coat. Technol.* **55** (1983) no. 704, 57–62.
40. M. V. Ram Mohan Rao, M. Yaseen, *Farbe Lack* **91** (1985) no. 9, 810–812.
41. G. Narender, M. Yaseen, *J. Coat. Technol.* **61** (1989) no. 773, 41–51.
42. J. K. Rigler: "Maßgeschneiderte Lösemittel für Polymere," Chemietag des Vereins österreichischer Chemiker, Graz 1977.
43. O. Lückert, *Farbe Lack* **75** (1969) no. 2, 128–138.
44. E. F. Meyer, R. E. Wagner, *J. Phys. Chem.* **70** (1966) 3162.
45. J. L. Gardon, *J. Paint Technol.* **38** (1966) no. 492, 43–57.
46. H. A. Papazian, *J. Am. Chem. Soc.* **93** (1971) 5634.
47. C. F. Holmes, *J. Am. Chem. Soc.* **95** (1973) 1014.
48. R. C. Nelson, R. W. Hemwall, G. D. Edwards, *J. Paint Technol.* **42** (1970) no. 550, 636–643.
49. H. Burrell, *Interchem. Rev.* **14** (1955) 31.
50. H. Burrell, *Off. Dig. Fed. Paint Varn. Prod. Clubs* **27** (1955) 726.
51. H. Burrell, *J. Paint Technol.* **40** (1968) no. 520, 197–208.
52. E. P. Liebermann, *Off. Dig. Fed. Soc. Paint Technol.* **34** (1962) 30.
53. M. Dyck, P. Hoyer, *Farbe Lack* **70** (1964) no. 7, 522–532.
54. E. B. Bagley, S. A. Chen, *J. Paint Technol.* **41** (1969) 494.
55. T. B. Epley, R. S. Drago, *J. Paint Technol.* **41** (1969) 500.
56. C. M. Hansen, K. Skaarup, *J. Paint Technol.* **39** (1967) 505.
57. *Kirk-Othmer,* 2nd ed., Suppl. vol., 889.
58. J. D. Crowley, G. S. Teague, J. W. Lowe, *J. Paint Technol.* **38** (1966) no. 496, 269–280; **39** (1967) 19; *Farbe Lack* **73** (1967) no. 2, 120–131.
59. I. Brandrup, E. H. Immergut: *Polymer Handbook,* 2nd ed., Interscience, New York 1974.
60. P. A. Small, *J. Appl. Chem.* **3** (1953) 71.
61. O. Fuchs, *Dtsch. Farben Z.* **23** (1969) 114.
62. O. Fuchs, *Kunststoffe* **43** (1953) 409.
63. U. Biethan, K. H. Hornung, *X. FATIPEC Congress Book,* Montreux 1970, pp. 277–281.
64. E. Bagda, *Colloid Polym. Sci.* **255** (1977) 384–392.
65. I. Prigogine, A. Bellemans, V. Mathot: *The Molecular Theory of Solutions,* North Holland Publ., Amsterdam 1957.
66. B. A. Wolf, J. W. Breitenbach, J. K. Rigler, *Angew. Makromol. Chem.* **34** (1973) 177–182.
67. L. H. Horsley: *Azeotropic Data, Advances in Chemistry Series,* Am. Chem. Soc., Washington, D.C., 1952.

68. L. H. Horsley: *Azeotropic Data, II, Advances in Chemistry Series,* Am. Chem. Soc., Washington, D.C., 1962; III, Washington, D.C., 1973.
69. C. H. Fisher, *J. Coat. Technol.* **63** (1991) no. 799, 79 – 83.
70. T. E. Gilbert, *J. Paint Technol.* **43** (1971) no. 562, 93 – 97.
71. G. M. Sletmoe, *J. Paint. Technol.* **42** (1970) no. 543, 246 – 249.
72. G. Marwedel, O. Hauser, *Farbe Lack* **54** (1948) no. 6, 115 – 118; no. 8, 175 – 181.
73. G. Marwedel, *Farbe Lack* **56** (1950) no. 11, 485 – 493; **59** (1953) 276 – 282, 317 – 320; **63** (1957) no. 10, 481 –487, no. 11, 542 – 548.
74. Shell Chemie, company brochure, "Voraussage der Verdunstungseigenschaften von Lösungsmittelgemischen," Frankfurt 1970, "Verdunstung von organischen Lösemitteln aus Oberflächenbeschichtungen," Frankfurt 1979.
75. A. L. Rocklin, *J. Coat. Technol.* **48** (1976) no. 622, 45 – 47.
76. H. L. Jackson, *J. Coat. Technol.* **58** (1986) no. 741, 87 – 89.
77. A. A. Sarnotsky, *J. Paint. Technol.* **41** (1969) no. 539, 692 – 701.
78. J. Ramsbotham, *J. Oil Colour Chem. Assoc.* **62** (1979) 359 – 366.
79. G. Rosen, L. M. Andersson, *J. Oil Colour Chem. Assoc.* **73** (1990) no. 5, 202 – 211.
80. J. F. Rankl, M. Huber, G. Grienberger, G. Bonn, *J. Oil Colour Chem. Assoc.* **73** (1990) no. 8, 334 – 336.
81. C. F. Böttcher: *The Theory of Electric Polarisation,* Elsevier, Amsterdam 1952.
82. L.-H. Lee, *J. Paint Technol.* **42** (1970) no. 545, 365 –370.
83. D. C. Busby et al., *J. Oil Colour Chem. Assoc.* **74** (1922) no. 10, 362 – 368.
84. W. Ormandy, E. Craven, *J. Inst. Pet. Technol.* **8** (1922) 145.
85. A. Blom, *Farben Ztg.* **36** (1931) 873.
86. J. Rybicky, J. R. Stevens, *J. Coat. Technol.* **53** (1981) no. 676, 40 – 42.
87. H. A. Wray, *J. Coat. Technol.* **56** (1984) no. 717, 37 –43.
88. K. G. Probst, *J. Paint. Technol.* **41** (1969) no. 539, 670 – 679.
89. W. H. Ellis, *J. Coat. Technol.* **48** (1976) no. 614, 45 –57.
90. H. Magdanz, H. Wacha, *Plaste Kautsch.* **25** (1978) no. 12, 707 – 712.
91. J. L. McGovern, *J. Coat. Technol.* **64** (1992) no. 810, 33 – 38, 39 – 44.
92. R. Smith, *J. Paint Technol.* **44** (1972) no. 565, 38 – 42.
93. Kansas City Society for Paint Technol., *J. Paint Technol.* **44** (1972) 565.
94. E. Oehley, *Chem. Ing. Tech.* **26** (1954) 97.
95. Le Chatelier, *Compt. Rend.* **126** (1898) 1344.
96. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin 1992.
97. Kühn-Birett: *Merkblätter Gefährliche Arbeitsstoffe,* Ecomed, Landsberg 1992.
98. Sorbe: *Sicherheitstechnische Kenndaten chemischer Stoffe,* Ecomed, Landsberg 1992.
99. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* 2nd ed., Deutscher Eichverlag, Braunschweig 1970.
100. P. H. Noll, *Ind. Lackierbetr.* **40** (1972) 354.
101. *Ind. Lackierbetr.* **44** (1976) 320.
102. J. J. Engel, T. J. Byerley, *J. Coat. Technol.* **57 (n 723)** (1985) 29 – 33.
103. S. D. Seneker, T. A. Potter, *J. Coat. Technol.* **63 (n 793)** (1991) 19 – 23.
104. S. F. Thames, P. C. Boyer, *J. Coat. Technol.* **62 (n 784)** (1990) 51 – 63.
105. H. Greim, W. Dessau: *Kombinationswirkung organischer Lösungsmittel. Toxische Wirkung auf Leber und Nervensystem,* Schriftenreihe der Bundesanstalt für Arbeitsschutz, Dortmund 1985.
106. L. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, New York 1965; *Registry of Toxic Effects of Chemical Substances,* National Institute for Occupational Safety and Health, Rockville 1976.
107. D. Henschler: *Gesundheitsschädliche Arbeitsstoffe, Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten,* Deutsche Forschungsgemeinschaft, VCH Verlagsgesellschaft, Weinheim 1991.
108. H. Altenkirch, *Adhäsion* (1984) no. 10, 9 – 11.
109. K. Kampmann, R. D. Henkler: "Arbeitsmedizinische Untersuchungen von lösemittelexponierten Chemiearbeitern in einer Lackfabrik," *Lack im Gespräch,* **vol. 9,** Informationsdienst Deutsches Lackinstitut, Frankfurt/M 1986.
110. P. Grasso, M. Sharrat, D. M. Davies, D. Irvine, *Fd. Chem. Toxicol.* **22** (1984) no. 10, 819 – 852.
111. R. J. Lewis: *Sax's Dangerous Properties of Industrial Materials,* 8th ed., vol. 1 – 3, Van Nostrand Reinhold, New York 1992.

112. *Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte,* Deutsche Forschungsgemeinschaft, Bonn 1992.
113. K.-H. Cohr, J. Stokholm, *Dan. Kemi* **58** (1977) no. 10, 316 – 322.
114. VDI-Guideline VDI 3881 Olfactometry, Odour Threshold Determination, part 1: "Fundamentals," part 2: "Sampling," part 3: "Olfactometers with Gas Jet Dilution," part 4: "Instructions for Application and Performance Characteristics."
115. A. H. Walter, *Farbe Lack* **77** (1971) no. 10, 1009 –1011; **80** (1974) 23.
116. A. H. Walter, *Adhäsion* (1975) 178.
117. A. H. Walter, *Dtsch. Farben Z.* **22** (1968) 235.
118. F. Mutter, *Eurocoat* (1990) no. 11, 690 – 695.
119. H. Haagen, *Farbe Lack* **91** (1985) no. 9, 795 – 801.
120. W. Lohrer, G. Pahlke, *Ind. Lackierbetr.* **53** (1985) no. 3, 75 – 78.
121. Globus, *Chem. Rundsch.* **43** (1990) no. 16, 2.
122. G. W. Fiero, *J. Paint Technol.* **40** (1968) no. 520, 222 –228.
123. H. H. G. Jellinek, F. Flajsman, F. J. Kryman, *J. Appl. Polym. Sci.* **13** (1969) 107 – 116.
124. National Paint, Varnish and Lacquers Assoc., Inc.,: "Paint Industry Smog Chamber," *Final Technical Report on the Role of Solvents in Photochemical Smog Formation,* Washington, D.C., 1970.
125. K.-H. Weinert, *Ind. Lackierbetr.* **47** (1979) no. 6, 216.
126. *Patty,* 4th ed., **1, part B,** 195 – 258.
127. W. S. Zimmt: *Der Einfluß von Lösungsmittel-Emissionsregelungen auf die amerikanische Farbenindustrie,* Ivry sur Seine 1970.
128. Eastman Chemical Co., Formulator's Notes No. M-6.1, Kingsport 1977.
129. R. Kreisler, W. Zöllner: *Merkblatt zur TA Luft,* 6th ed., Verband der Lackindustrie, Frankfurt/M 1990.
130. W. Lohrer, G. Pahlke, *Ind. Lackierbetr.* **49** (1981) no. 4, 135 – 140.
131. R. Schaaf, *Ind. Lackierbetr.* **53** (1985) no. 4, 111 –112.
132. L. M. Keinert, *Sicherheitsingenieur* (1975) no. 10, 1 – 8.
133. K. R. Smith, *J. Coat. Technol.* **62 (n 788)** (1990) 71 –81.
134. H. Aichinger: *Die österreichische Lösungsmittelverordnung,* Bundesministerium für Umwelt, Jugend und Familie, Wien 1991. pp. 179, 492.
135. D. C. Heslinga, P. J. H. Builtjes, M. G. M. Roemer, A. C. Baart: TNO-Report, A Consultancy Study on Volatile Organic Compounds–The POCP-Concept in Relation To Oxygenated Solvents, Apeldoorn/Delft 1992.
136. U. Zorll, *Farbe Lack* **97** (1991) no. 4, 382 – 383.
137. U. Sayler, *Ind. Lackierbetr.* **42** (1974) no. 11, 417 –420.
138. H. Schene, *Z. Ind. Fertigung* **64** (1974) 282 – 287.
139. W. Lohrer, *Oberfläche JOT* **32** (1992) no. 4, 108 – 119.
140. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* 2nd ed., van Nostrand Reinhold, New York 1983.
141. *Katalog wassergefährdender Flüssigkeiten der Länderverordnungen über das Lagern wassergefährdender Flüssigkeiten (VLwFl).*
142. H.-J. Erberich, *Farbe Lack* **90** (1984) no. 2, 146.
143. OECD Guidelines for Testing Chemicals, Paris 1981/1987: no. 201 – 209 Effects on Biotic Systems, no. 301 Ready Biodegradability, no. 302 Inherent Biodegradability, no. 303 Simultation Test, no. 304 Biodegradability in Soil, no. 305 Bioaccumulation, no. 401 – 482 Health Effects.
144. DIN 38 405,Deutsches Einheitsverfahren (DEV) zur Wasser-, Abwasser- und Schlammuntersuchung.
145. E. Montorsi, *Pitture Vernici Eur.* **68** (1992) no. 1, 29 – 31.
146. Richtlinie des Rates der Europäischen Gemeinschaft zur Angleichung der Rechts- und Verwaltungsvorschriften für die Einstufung, Verpackung und Kennzeichnung gefährlicher Stoffe, Grundrichtlinie 67/548 EWG, Amtsblatt der Europäischen Gemeinschaften no. 196, pp. 1 ff (1967).
147. L. H. Keith, D. B. Walters: *Compendium of Safety Data Sheets for Research and Industrial Chemicals,* part 1 – 2, VCH Publishers, New York 1985/1989.
148. Gebrüder Sulzer AG, Lösungsmittel aus der Abluft, Winterthur 1991.
149. H. E. Koenen, *Chem. Rundsch.* **43** (1990) no. 45, 19.
150. *Oberfläche JOT* **32** (1992) no. 2, 54.
151. P. Toepke, *Farbe Lack* **97** (1991) no. 8, 700 – 701.
152. Buss-SMS GmbH, Butzbach, *Chem. Rundsch.* **43** (1990) no. 45, 4.

153. The Weir Group Ltd., Glasgow, *Ind. Lackierbetr.* **45** (1977) no. 11, 439.
154. B. Hamill, *Paint Resin* **39** (1985) no. 5, 34 – 37.
155. A. Schmidt, M. Ulrich, *Chem. Ing. Tech.* **62** (1990) no. 1, 43 – 46.
156. D. Rasenack, ILV-Mitteilungen des Fraunhofer-Instituts für Lebensmitteltechnologie und Verpackung **8 – 9,** (1984) 116 – 133.
157. OFRU-Recycling, *Oberfläche JOT* **23** (1983) no. 10, 8.
158. W. Schnabel, *Oberfläche JOT* **32** (1992) no. 6, 40 – 42.
159. H. Erler, K. Hermanns, M. Karthaus, *Coating* (1992) no. 4, 120 – 122.
160. R. W. Baker, Yoskioka, J. M. Mohr, A. J. Khan, *J. Membr. Sci.* **31** (1987) 259.
161. H. Strathmann, C.-M. Bell, K. Kimmerle, *Pure Appl. Chem.* **58** (1986) no. 12, 1663 – 1668.
162. VDI-Richtlinie 3674, Abgasreinigung durch Adsorption, Oberflächenreaktion und heterogene Katalyse. VDI-Richtlinie 3675, Abgasreinigung durch Adsorption. VDI-Richtlinie 2442, Abgasreinigung durch thermische Verbrennung. VDI-Richtlinie 3476, Abgasreinigung durch katalytische Verfahren. VDI-Richtlinie 3477/3478, Biologische Abluftreinigung. VDI-Richtlinie 2588, Emissionsminderung: Beschichtung von metallischen Oberflächen mit organischen Stoffen. VDI-Richtlinie 2280, Auswurfbegrenzung: Organische Verbindungen, insbesondere Lösemittel. VDI-Richtlinien 2460, 3481, 2457, Emissionsmeß-technik organische Stoffe.
163. S. P. P. Ottengraf et al., *Bioprocess Eng.* **1** (1986) 61 –69.
164. D. A. Carlson, C. P. Leiser, *J. Water Pollut. Control Fed.* **38** (1966) 829.
165. U. Bäuerle, K. Fischer, D. Bardtke, *Staub-Reinhalt. Luft* **46** (1986) 233 – 235; CJB Developments Ltd., *J. Oil Colour Chem. Assoc.* **75** (1992) no. 11, 439 –441.
166. G. Heck, G. Müller, M. Ulrich, *Chem. Ing. Tech.* **60** (1988) no. 4, 286 – 297.
167. H. Buri et al.: *Lösemittelemission,* Verein zur Förderung der Wasser- und Lufthygiene, Zürich 1990.
168. P. Murr, *Oberfläche JOT* **32** (1992) no. 3, 34 – 39.
169. J. Halbartschlager, *Oberfläche JOT* **32** (1992) no. 3, 28 – 33.
170. R. F. Piepho, H. Schuster, *Ind. Lackierbetr.* **58** (1990) no. 5, 172 – 174.
171. *Patty,* 4th ed., 1, part B, 259 – 287.
172. E. Mink, *Oberfläche JOT* **31** (1991) no. 3, 22 – 26.
173. J. Diebold, *Oberfläche JOT* **30** (1990) no. 11, 857 –859.
174. H. Deuster, *Ind. Lackierbetr.* **58** (1990) no. 9, 325 –326.
175. K. Panzer, *Ind. Lackierbetr.* **58** (1990) no. 9, 323 –324.
176. M. B. Hundert, *J. Paint Technol.* **40** (1968) no. 516, 33 – 48.
177. D. G. Anderson, *J. Paint. Technol.* **40** (1968) no. 527, 549 – 557.
178. G. G. Esposito, *J. Paint. Technol.* **40** (1968) no. 520, 214.
179. L. Rohrschneider, *Anal. Chem.* **45** (1973) 1241 – 1247.
180. D. Groß, H.-J. Kretzschmar, H. Weigel, *Farbe Lack* **82** (1976) no. 8, 690 – 692.
181. K. L. Olson, C. A. Wong, L. L. Fleck, D. F. Lazar, *J. Coat. Technol.* **60** (n 762) (1988) 45 – 50.
182. K.-H. Pannwitz: *Ein neuer Probenehmer für organische Lösemitteldämpfe,* Dräger-Publikationen, Lü-beck 1990.
183. DIN-Arbeitskreis "Lösemittel, " *Farbe Lack* **97** (1991) no. 2, 127 – 129; *J. Oil Colour Chem. Assoc.* **75** (1992) no. 4, 154 – 156.
184. L. R. Snyder, *J. Chromatogr.* **16** (1964) 55 – 88.
185. M. Gruenfeld, *J. Paint Technol.* **42** (1970) no. 543, 237 – 242.
186. R. Prakash, I. C. Shukla, *J. Oil Colour Chem. Assoc.* **63** (1980) 370 – 372.
187. D. Downing, *J. Oil Colour Chem. Assoc.* **74** (1991) no. 8, 284 – 285.
188. Shell Chemicals: Thinner Formulation for the Paint Industry, Technical Bulletin, Rotterdam 1981.
189. New York Society for Paint Technol., *J. Paint Technol.* **41** (1969) 692.
190. J. K. Rigler, B. A. Wolf, J. W. Breitenbach, *Angew. Makromol. Chem.* **57** (1977) no. 854, 15 – 27.
191. W. H. Ellis, Z. Saary, D. G. Lesnini, *J. Paint Technol.* **41** (1969) no. 531, 249 – 258.
192. G. Marwedel, *Dtsch. Farben Z.* **26** (1972), no. 7, 8, 365 – 418.
193. E. Bagda, *Dtsch. Farben Z.* **32** (1978) no. 10, 372 –378.
194. K. M. A. Shareef, M. Yaseen, *J. Coat. Technol.* **55 (n 701)** (1983) 43 – 52.
195. A. Toussaint, I. Szigetvari, *J. Coat. Technol.* **59 (n 750)** (1987) 49 – 59.

196. R. A. Waggoner, F. D. Blum, *J. Coat. Technol.* **61 (n 768)** (1989) 51 – 56.
197. A. Vinther, P. Sörensen, *Farbe Lack* **77** (1971) no. 4, 317 – 325.
198. M. Camina, D. M. Howell, *J. Oil Colour Chem. Assoc.* **55** (1972) 929.
199. P. Fink-Jensen, *Farbe Lack* **68** (1962) 155.
200. H. A. Dierstead, J. Turkevich, *J. Chem. Phys.* **12** (1944) 24.
201. A. R. H. Tawn, *Farbe Lack* **75** (1969) no. 4, 311 –318.
202. H. P. Schreiber, *J. Paint Technol.* **46** (1974) no. 598, 35 – 39.
203. G. Holtmann, *Schweizer Maschinenmarkt* **43** (1978) 52 – 55.
204. W. Andrejewski, J. Dörffel, D. Stoye, Der Lichtbogen, company brochure, Hüls AG, no. 175, Marl 1974.
205. G. Walz, Kunstharz Nachr. **33** (1974) no. 6, 30 – 39.
206. G. Sprinkle, *J. Coat. Technol.* **53 (n 680)** (1981) 67 –72.
207. G. P. Sprinkle, *Mod. Paint Coat.* (1983) no. 4, 44 – 48.
208. W. H. Ellis, *J. Coat. Technol.* **55** (n 696) (1983) 63 – 72.
209. J. R. Erickson, *J. Coat. Technol.* **48** (n 620) (1976) 58.
210. A. Goldschmidt, *Farbe Lack* **84** (1978) no. 9, 675 –680.
211. J. J. Stratta, P. W. Dillon, R. H. Semp, *J. Coat. Technol.* **50 (n 647)** (1978) 39 – 47.
212. D. Stoye, *Coating* (1987) no. 6, 223.
213. J. Spauwen, *J. Oil Colour Chem. Assoc.* **71** (1988) no. 2, 47 – 49.
214. W. Appelt, *Farbe Lack* **96** (1990) no. 3, 200 – 201; *J. Oil Colour Chem. Assoc.* **73** (1990) no. 5, 200 – 201.
215. M. R. C. Gerstenberger, D. K. Kruse, *Farbe Lack* **90** (1984) no. 7, 563 – 568.
216. F. Müller, *Ind. Lackierbetr.* **58** (1990) no. 2, 48 – 52.
217. A. L. Rocklin, *J. Coat. Technol.* **50 (n 646)** (1978) 47 –55.
218. A. L. Rocklin, D. C. Bonner, *J. Coat. Technol.* **52 (n 670)** (1980) 27 – 36.
219. H. Magdanz, K. Berger, G. Schumann, *Farbe Lack* **75** (1969) no. 3, 221 – 235.
220. A. L. Rocklin, G. D. Edwards, *J. Coat. Technol.* **48 (n 620)** (1976) 68 – 74.
221. P. M. Grant, *J. Coat. Technol.* **53 (n 677)** (1981) 33 –38.
222. L. O. Kornum, *J. Oil Colour Chem. Assoc.* **63** (1980) 103 – 123.
223. T. Imai, K. Tsubouchi, *J. Coat. Technol.* **52 (n 666)** (1980) 71 – 78.
224. R. F. Eaton, F. G. Willeboordse, *J. Coat. Technol.* **52 (n 660)** (1980) 63 – 70.
225. A. Jones, L. Campey, *J. Coat. Technol.* **56 (n 713)** (1984) 69 – 72.
226. W. J. Blank, *J. Coat. Technol.* **61 (n 777)** (1989) 119 –128.
227. J. Dörffel, *Farbe Lack* **81** (1975) 10 – 15.
228. P. Sörensen, *Dtsch. Farben Z.* **24** (1970) 473.
229. H. Jebsen-Marwedel, G. Marwedel, *Farbe Lack* **66** (1960) no. 6, 314 – 330.
230. G. Marwedel, *Farbe Lack* **66** (1960) no. 7, 379 – 385.
231. G. Marwedel, H. Jebsen-Marwedel, *Dtsch. Farben Z.* **24** (1970) no. 3, 103 – 116.
232. R. Haug, *Dtsch. Farben Z.* **23** (1969) no. 10, 467 –472.
233. J. E. Weigel, E. G. Sabino, *J. Paint Technol.* **41** (1969) no. 529, 81 – 88.
234. W. Z. Borer, R. L. Bartosiewicz, P. J. Secrest, *J. Paint Technol.* **43** (1971) no. 559, 61 – 64.
235. W. Götze, *IX. Fatipec Congress Book* **no. 2,** Brussels 1968,79 – 83.
236. D. Y. Perera, D. van den Eynde, *J. Coat. Technol.* **55 (n 699)** (1983) 37 – 43.
237. D. Stoye, K.-H. Haneklaus, *Fette Seifen Anstrichstoffe* **4** (1972) 217.
238. L. M. Keinert, *Oberfläche* (1975/1976), no. 12/1.
239. S. Knödler, *Papier (Darmstadt)* **35** (1981) no. 10 A, 74 – 79.
240. O. Jordan: *Chemische Technologie der Lösungsmittel,* Springer Verlag, Berlin 1932, pp. 91 – 95.
241. A. K. Doolittle: *The Technology of Solvents and Plasticizers,* Wiley-Interscience, New York 1954.
242. L. Alders: *Liquid-Liquid Extraction,* Elsevier, Amsterdam 1955.
243. A. Weissberger: *Technique of Organic Chemistry,* **vol. 4,** Interscience, New York 1965. E. S. Perry, A. Weissberger: *Distillation,* Interscience, New York 1965.
244. *Winnacker-Küchler,* **vol.** 4, 5.
245. K. Randerath: *Dünnschichtchromatographie,* Verlag Chemie, Weinheim 1965.
246. E. S. Perry, A. Weissberger: *Technique of Organic Chemistry,* vol. 12; J. G. Kirchner: *Thin-Layer Chromatography,* Interscience, London 1967.
247. F. Cramer: *Papierchromatographie,* Verlag Chemie, Weinheim 1962.
248. E. S. Amis: *Solvents Effect on Reaction Rates and Mechanism,* Academic Press, New York 1966.

249. C. A. Bunton: *Nucleophilic Substitution at a Saturated Carbon Atom,* Elsevier, Amsterdam 1963, p. 111.
250. L. Küchler: *Polymerisationskinetik,* Springer Verlag, Berlin 1951.
251. W. C. Aten: *Einsatzmöglichkeiten von Computern bei der Formulierung von Lösemittelgemischen für Klebstoffe,* Technische Akademie, Esslingen 1984.
252. P. Goerlich, *Ind. Lackierbetr.* **43** (1975) no. 11, 383–386.
253. G. Ohlendorf, *Ind. Lackierbetr.* **51** (1983) no. 2, 49–54.
254. H. Gölz, *Oberfläche JOT* **32** (1992) no. 2, 56–58.
255. K.-H. Adams, *Ind. Lackierbetr.* **58** (1990) no. 1, 5–7.
256. J. Kitlar, *Oberfläche JOT* **32** (1992) no. 1, 50–51.
257. G. Ohlendorf, *Ind. Lackierbetr.* **51** (1983) no. 2, 49–54.
258. E. Schmidt, *Oberfläche JOT* **30** (1990) no. 3, 28–31.
259. W. Hater, *Ind. Lackierbetr.* **59** (1991) no. 5, 156–158.
260. B. Rohland, W. Majewski, H. Krüssmann, *WRP* (1989) no. 5, 46–47.
261. P. L. Bartlett, *Seifen Fette öle Wachse* **113** (1987) no. 13, 437–442.
262. Frost & Sullivan, *Polym. Paint Colour J.* **178** (1988) no. 4207, 57.
263. M. R. C. Gerstenberger, D. K. Kruse, *Farbe Lack* **90** (1984) no. 1, 30–34.
264. W. H. Ellis, C. D. McLaughlin, *J. Coat. Technol.* **52 (n 662)** (1980) 61–72.
265. E. Montorsi, *Pitture Vernici Eur.* **68** (1992) no. 11, 85–90.
266. J. R. Kelsey, D. Lyons, S. R. Smithey, *J. Oil Colour Chem. Assoc.* **74** (1991) no. 8, 278–287.
267. J. Falbe, B. Cornils, *Fortschr. Chem. Forsch.* **11** (1968) no. 1, 101–134.
268. C.-D. Mengler, *Chem. Z.* **95** (1971) 309–316.
269. C. Parducci, *Riv. Colore Verniciatura Ind.* **10** (1977) no. 114, 329–332.
270. G. S. Salvapati, M. Janardanarao, *J. Sci. Ind. Res.* **42** (1983) 261–267.
271. M. Peignier, P. Lasnet de Lanty, R. Zimmermann, *Farbe Lack* **86** (1980) no. 11, 973–975.
272. P. I. Smith, "The Use of Tetrahydrofuran in Coating Formulations" *Ind. Finish. and Surf. Coat.* (1971).
273. W. Stumpf: *Chemie und Anwendung des 1,4-Dioxans,* Verlag Chemie, Weinheim 1956.
274. J. E. Hamlin, *Paint Resin* **37** (1983) no. 5, 42–47.
275. R. A. Heckman, *Mod. Paint Coat.* (1986) no. 6, 36–42.
276. H. Kittel, *Adhäsion* (1987) no. 5, 33–34.
277. H. W. Lang, *Chem. Ztg.* **84** (1960) no. 8, 239–247.
278. S. S. Pizey: *Synthetic Reagents,* vol. 1, Wiley-Interscience, New York 1974.
279. D. Martin, H. G. Hauthal: *Dimethylsulfoxid,* Akademie Verlag, Berlin 1971.
280. Gaylord Chemical Corp., Technical Bulletin Dimethylsulfoxid, Slidell/USA.
281. H. Normant, *Angew. Chem.* **79** (1967) 1029; *Angew. Chem. Int. Ed. Engl.* **6** (1967) 1046.

Sorbic Acid

Erich Lück, Hoechst Aktiengesellschaft , Frankfurt/Main, Federal Republic of Germany
Martin Jager, Hoechst Aktiengesellschaft , Frankfurt/Main, Federal Republic of Germany
Nico Raczek, Hoechst Aktiengesellschaft , Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Physical Properties 1
3. Chemical Properties and Derivatives . 2
4. Production 3
5. Environmental Protection 3
6. Quality Specifications 4
7. Analysis 4
8. Storage and Transportation 4
9. Legal Aspects 4
10. Mode of Action and Uses 5
11. Economic Aspects 6
12. Toxicology 6
13. References 7

1. Introduction

The *trans,trans* isomer of 2,4-hexadienoic acid is known as sorbic acid [*110-44-1*].

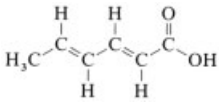

The name is derived from the scientific name of the mountain ash, *Sorbus aucuparia* Linnaeus. In 1859 Hofmann obtained an oil with a characteristic odor from the juice of unripe rowan berries by distillation. He named the most important component of this oil parasorbic acid. Parasorbic acid [*10048-32-5*] is the δ-lactone of sorbic acid. Strong acids or alkalis convert it into the isomeric sorbic acid [3]. The parasorbic acid content of unripe rowan berries is ca. 0.1 % [4].

Sorbic acid occurs in the form of 2-sorboyl-1,3-dimyristin [*7175-63-5*] in the fatty deposits of certain aphids [5].

Doebner established the structure of sorbic acid in 1890 [6], and in 1900 also described the first total synthesis [7].

In 1939/40 Müller [8] in Germany and Gooding in the United States [9] independently discovered the antimicrobial activity of sorbic acid. Industrial supplies of sorbic acid and potassium sorbate became available during the 1950s, first in the United States, shortly thereafter in Germany, and later in Japan. This opened the way to a broad range of applications in the preservation of food, animal feeds, pharmaceuticals, cosmetics, and other non-food applications. Because it is harmless from a physiological standpoint and exhibits favorable sensory characteristics sorbic acid has in the meantime become the leading preservative for foods.

2. Physical Properties

Sorbic acid crystallizes in needles or plates with a weak characteristic odor and a slightly acidic taste; *mp* 132 – 135 °C; *bp* 228 °C (decomp.); n_D^{20} 1.4248, density 1.204. Sorbic acid is soluble in 100 mL of water to the extent of 0.16 g at 20 °C, 0.58 g at 50 °C, and 3.9 g at 100 °C. The solubility in anhydrous lower molecular mass alcohols and anhydrous acetic acid is ca. 11 – 12 g per 100 mL. Liquid fats dissolve 0.5 – 1 g of sorbic acid per 100 mL.

Sorbic acid begins to sublime above 60 °C. The vapor pressure at 20 °C is <0.001 kPa; at 100 °C it is 0.25 kPa, and at 120 °C 1.3 kPa. Sorbic acid is volatile in steam without decomposition. This property is important for its isolation from foods for analytical purposes.

The dissociation constant of sorbic acid is 1.73×10^{-5} at 25 °C, and the pK_a is 4.76. Its pH-dependent dissociation behavior is important in the use of sorbic acid as a preservative. Only the undissociated form has antimicrobial activity.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_507

3. Chemical Properties and Derivatives

The chemical behavior of sorbic acid is determined by the carboxyl group and the conjugated double-bond system. Reactions at the carboxyl group correspond to those of other carboxylic acids; salts, esters, and other acid derivatives are easily prepared by the usual methods. The carboxyl group can be reduced selectively with lithium aluminum hydride to give sorbyl alcohol, 2,4-hexadien-1-ol [*17102-64-6*] [10]. Unlike sorbic acid itself, the alkali salts are readily water soluble, and they are used preferentially for the preservation of systems containing water.

Potassium sorbate [*24634-61-5*], prepared by dissolving sorbic acid in aqueous potassium hydroxide, is of great practical importance. It is sold commercially as a white powder and in granulated form. The stability of potassium sorbate is strongly dependent on its water content, which must be kept well below 0.5 %. At room temperature ca. 140 g potassium sorbate can be dissolved in 100 mL of water. Saturated potassium sorbate solutions contain ca. 58 wt % potassium sorbate.

Calcium sorbate [*7492-55-9*], with a very low water solubility, can be produced by the reaction of sorbic acid with calcium hydroxide in water. Because of its particularly high resistance to oxidation, calcium sorbate is suitable for the production of packaging materials with a preserving effect [11].

The *lower alkyl esters* of sorbic acid also exhibit antimicrobial properties [12]. Unlike sorbic acid, these are also active in neutral and weakly alkaline media. In contrast to sorbic acid and its salts, the sorbate esters have a pronounced odor. The esters of sorbic acid with long chain fatty acids are described as suitable for bread preservation because, in contrast to normal sorbic acid they do not inhibit the growth of the bakery yeasts. However, nowadays a special type of sorbic acid has been developed that can also be used for this purpose.

Reactions of sorbic acid involving the double-bond system are often complex, frequently resulting in mixtures of products. This is due to the differing reactivities of the two double bonds and to side reactions and such subsequent transformations as isomerizations, rearrangements, double-bond migration, and polymerization.

In the case of addition reactions and partial hydrogenation, preferential attack is usually observed at the 4,5 double bond. Even with long reaction times chlorine addition is incomplete, leading to mixtures of various chlorohexenoic acids with an average chlorine content of 38 – 48 wt %. Addition of bromide in the presence of organic solvents leads to 2,3,4,5-tetrabromohexanoic acid [*62284-99-5*]. The major product of partial bromination in aqueous medium is 4,5-dibromo-2-hexenoic acid [*19147-46-7*] [13].

Pure crystalline sorbic acid is surprisingly resistant to air oxidation given that it is doubly unsaturated. At room temperature it remains unchanged for years, although its stability in the solid state depends very much on purity. The shelf life is reduced considerably by traces of solvents, heavy metals, or isomeric hexadienoic acids, which can form as side products during sorbic acid synthesis. The same applies to potassium sorbate. Sodium sorbate [*7757-81-5*] is unstable in the solid state and is therefore not produced industrially. In air, but not in solution, sodium sorbate is converted very rapidly into the sodium salt of 4,5-epoxy-2-hexenoic acid [14]. The oxidation of aqueous solutions of sorbates or solutions of sorbic acid in organic solvents gives rise to a large number of carbonyl compounds, although this reaction depends upon the presence of a large quantity of oxygen is strongly increased by exposure to sunlight [15], [16]. In solutions, the presence of oxygen causes oxidative degradation which may result in brown discoloration [83]. Oxygen-free solutions of sorbic acid remain unchanged even upon exposure to daylight [17]. In intermediate-moisture foods slow degradation of sorbic acid becomes apparent after storage for several months in the presence of air [18]. Potassium dichromate oxidation of acidic, aqueous sorbic acid solutions leads to reproducible yields of malonaldehyde [*542-78-9*]. The analytical determination of sorbic acid acid is based on a color reaction of malonaldehyde with 2-thiobarbituric acid (see Chap. 7).

Sorbic acid that has been incorporated into foods is at least as stable as the important constituents of the foods themselves (e.g., vitamins, flavors, and aromatic substances). The theoreti-

cal potential for sorbic acid to undergo autoxidation thus poses no problems in practice with respect to the preservation of foods.

4. Production

The first described synthesis of sorbic acid consisted of a condensation of *trans*-2-butenal [*123-73-9*] with malonic acid [*141-82-2*] in pyridine [19]. It is named Doebner synthesis after its inventor. Many other syntheses are based on similar principles [94–100]. Most are of no commercial interest because of low yields, expensive starting materials, or high production cost. This also applies, for example, to a process practiced temporarily on an industrial scale in the 1950s and 1960s in the United States by Union Carbide, which involved the oxidation of 2,4-hexadienal [*142-83-6*] to sorbic acid in the presence of catalysts [20–22]. This process was abandoned largely because it gave up to 20 % of the isomeric hexadienoic acids as well, which are much less stable to storage than sorbic acid and could be removed only through an expensive purification process. 2,4-Hexadienal can also be oxidized to sorbic acid by microorganisms [23], [101].

An alternative preparation of sorbic acid involves isomerization of 2,5-hexadienoic acid [*38867-16-2*] by boiling with aqueous alkali. 2,5-Hexadienoic acid is in turn available by treatment of allyl chloride [*107-05-1*] with acetylene [*74-86-2*], carbon monoxide, and water in the presence of tetracarbonylnickel as catalyst [24–26]. The acidic cleavage of 5-vinyl-γ-butyrolactone [*21963-38-2*], obtained from 1,3-butadiene [*106-99-0*] and acetic acid in the presence of redox catalysts, is of no industrial importance [27–31]. Production of sorbic acid by the addition of carbon dioxide to 1,3-pentadiene [*2004-70-8*] in the presence of nickel complexes has also been described [32].

Another common process starts with ketene [*463-51-4*] and 2-butenal [*123-73-9*] [33]. In the presence of salts of divalent transition metals as catalysts, ketene and 2-butenal react at 20 – 80 °C to give a polymeric ester of 3-hydroxy-4-hexenoic acid [*26811-78-9*] with a molecular mass ≥ca. 2000. This polyester can be cleaved to give sorbic acid in good yield [34] by either bases or acids (e.g., hydrochloric acid [35]);

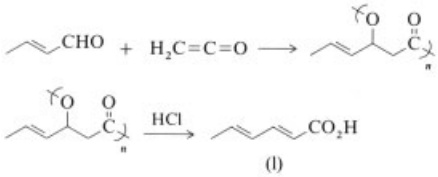

alternatively, cleavage can be effected by metal-complex catalysts, [36].

The polyester is also suitable for the production of various sorbic acid derivatives. For example, in the presence of acidic esterification catalysts it can be reacted with alcohols to give sorbate esters [38].

The crude sorbic acid must be purified, generally by recrystallization from an aqueous solution [39], from a mixture of water and an alcohol or acetone [40], or from organic solvents that are immiscible with water, such as methyl acetate [41]. Treatment with organic solvents [42] and purification by steam distillation [43] have also been described.

5. Environmental Protection

Sorbic acid shows antimicrobial activity against many fungi, yeasts, and bacteria (see Chap. 8). However, this is only the case if the population of the target organisms is small. Many microorganisms are capable of degrading and metabolizing sorbic acid when present in large numbers [44], including gram-negative sulfate-reducing bacteria [45]. It is for this reason that sorbic acid is readily degradable in soil. Sorbic acid also shows high degradability (95 % within 6 d) in the Zahn – Wellens test (OECD 302 B). Sorbic acid and the sorbates are not hazardous with respect to water and are not subject to any hazardous materials classification.

The acute fish toxicity (LC_{50} for the zebra barbel) is very low (>1000 mg/L after 48 – 96 h). Thus, sorbic acid is incorporated in the lowest German water hazard class (WGK"0") [148].

At the normal application concentrations of 0.05 – 0.2 %, sorbic acid displays very little toxicity to plants. Neither the roots nor the aboveground portions are damaged, from which it can

be concluded that the ecotoxicity of the substance generally must be very low.

In the absence of local regulations to the contrary, sorbic acid and sorbates can be disposed of or incinerated along with household waste in the same way as foods.

6. Quality Specifications

Sorbic acid and sorbates are used almost exclusively in foods, animal feeds, cosmetics, and pharmaceuticals, and quality specifications are therefore adapted to the legal requirements applicable to these areas. In the EC the regulation of 2. December 1996, concerned specifically with additives is valid, which is based on an EC directive. The purity standards of the third edition of the Food Chemicals Codex are applicable in the United States. JECFA, a joint committee of experts from the United Nations Food and Agriculture Organization (FAO) and the World Health Organization (WHO) has also developed a set of purity specifications. Furthermore, sorbic acid and potassium sorbate are cited in various pharmacopeias.

All these requirements specify a purity of at least 99 % and limit the content of aldehydes, arsenic, and toxicologically relevant heavy metals. Neither sorbic acid nor its salts should show signs of discoloring during 90 min of heating at 105 °C.

7. Analysis

Steam distillation is a common method for the quantitative isolation of sorbic acid from the investigated material. For qualitative detection and quantitative determination the red coloration is used that sorbic acid produces with 2-thiobarbituric acid after oxidation with potassium dichromate to malonaldehyde. As a polyunsaturated compound, sorbic acid displays a pronounced absorption maximum at ca. 260 nm (depending on the pH of the solution), which can be likewise used for quantitative determination [84]. Various methods are described and published that utilize HPLC for analysis. For analyses in special foods see [85–88], for determination of sorbic acid in foods and cosmetics in general see [89–91].

Standardized methods of detecting sorbates (GC, TLC, and HPLC) have been published in the revised edition of the Swiss Foodstuffs Manual (1992). A method exists for detecting sorbates in liquid sweeteners in accordance with § 35 of the Federal German food law (L. 57.22.99). Rather unconventional techniques of detecting sorbic acid such as ion chromatography or capillary isotachophoresis [92] have not become established so far in routine use. X-ray structural analysis of sorbic acid has also been described [93].

8. Storage and Transportation

Sorbic acid and its salts can be stored for several months. They should be stored in a cool, dry environment with protection from light. This applies especially to finely powdered sorbates.

Sorbic acid and sorbates are not subject to any hazardous goods classification. They can be stored and transported in the same way as ordinary foods.

9. Legal Aspects

Sorbic acid is permitted worldwide as a preservative for foods. Approval lists differ considerably from country to country, however, because of differences in eating habits. In the United States sorbic acid and its salts (which includes the sodium, potassium, and calcium salts) are classified as "generally recognized as safe" (GRAS). (Sodium sorbate is not produced industrially and is generally not used because of its instability). In the EC sorbic acid is included as number E 200 in the list of approved preservatives, which also recognizes potassium sorbate (E 202) and calcium sorbate (E 203).

Sorbic acid is also permitted as an additive for animal feeds and cosmetics. Sorbic acid and potassium sorbate are listed in various pharmacopeias and are thus recognized as acceptable preservatives for pharmaceutical products as well.

10. Mode of Action and Uses

Certain potential uses for sorbic acid described in the patent literature (e.g., as a raw material

for chemical reactions or polymerizations) have remained unimportant in practice because of the relatively high cost of the substance. Sorbic acid and the sorbates are used almost exclusively as preservatives. Because of their lack of toxicity the main areas of application are in foods, animal feeds, tobacco, cosmetics, and pharmaceuticals, as well as in packing materials for these substances and in other products that come into contact with human or animal skin in some way. Sorbic acid, potassium sorbate, and especially calcium sorbate are also used in some technical applications [46].

Some details of the antimicrobial activity of sorbic acid remain unexplained [47], [48]. Sorbic acid inhibits many enzymes in the microbial cell, namely enolase and lactate dehydrogenase, which are involved in the carbohydrate metabolism [105]. Sorbic acid interferes also, though not very specifically, with enzymes of the citric acid cycle including succinate dehydrogenase, fumarase, aspartase, malate dehydrogenase, isocitrate dehydrogenase, and α-ketoglutanate dehydrogenase [106–108]. Sorbic acid also impairs the catalase/peroxidase system [109], [110] and enzymes containing a sulfhydryl group [111]. The primary centers of attack in the cell may differ in bacteria, yeasts, and fungi. In other reports, it has been increasingly assumed that sorbic acid is also active against the cell wall [49]. Thus, sorbic acid inhibits the absorption of amino acids in bacteria. Proton-flow into the cell also increases as a result of partial destruction of the cell membrane, requiring the cell to expend additional energy to compensate for the resulting potential differences [50]. Resistance in the true sense of the word (i.e., an increase in the limiting inhibiting concentration under the influence of subthreshold amounts of sorbic acid) is not observed in either *Echerichia coli* [51] or fungi [52]. Primarily the undissociated sorbic acid enters the cell wall of microorganisms, i.e., for food preservation the undissociated proportion of sorbic acid is most effective. Owing to its low dissociation constant sorbic acid, unlike other preservative acids, can also be employed for preserving weakly acid foods with a high pH. There are reasons to believe that also dissociated sorbic acid has an antimicrobial action. However this action is about 100 times weaker than that of the undissociated acid [112], [113].

Sorbic acid is a microbiostatic and not a microbiocidal agent. Sorbic acid activity is directed mainly toward yeasts and molds, including organisms responsible for aflatoxin formation [53]. Among bacteria, those that are catalase positive are inhibited more strongly than those that are catalase negative, with aerobic bacteria affected the most. However, bacteria can also be suppressed by a combined action of sorbic acid and other inhibiting factors [114–125]

Some microorganisms are capable of incorporating sorbic acid into their metabolism and thereby degrading it [44]. This is only possible, however, if sorbic acid is present below the threshold concentration and the microorganism population is very high. This means that sorbic acid is suitable for the preservation only of hygienically clean ingredients with low microorganism counts.

An important incentive for the use of sorbic acid and sorbates in foods is their virtual neutrality with respect to taste and odor, together with the fact that sorbic acid does not react with food components because of its chemical inertness.

The application concentration of sorbic acid is usually 500 – 2500 ppm. Higher concentrations are required the higher the pH of the product to be preserved. Low concentrations suffice if the system to be preserved is subject to other protective factors, such as low pH. In the case of wine ca. 50 – 200 ppm sorbic acid or a corresponding amount of potassium sorbate is adequate because the effectiveness is enhanced by the wine's low pH and the sulfur dioxide and ethanol present.

The favorable partition coefficient of sorbic acid between oil and water is important in the preservation of fat emulsions such as margarine or mayonnaise and cosmetic lotions. Thus, a relatively high proportion of the sorbic acid in an emulsion remains in the aqueous phase, which is alone susceptible to microbiological attack [54]. Sorbic acid is the agent of choice for the preservation of butter and margarine with a low fat content.

Sorbic acid and its salts are very important for preventing the development of mold in cheeses of all types. The activity of sorbic acid against agents responsible for mycotoxin formation plays a special role in this context [55]. Depending on the type of cheese, sorbic acid can be added to the cheese itself (cottage cheese, pro-

cessed cheese), potassium sorbate can be added to the salt bath, or the cheese surface can be treated with a solution of suspension of sorbic acid, potassium sorbate, or calcium sorbate (hard cheese). The role of fungistatic packaging materials and coating materials containing sorbic acid or its salts is diminishing [56].

Undesirable mold growth on hard sausages can be suppressed by treatment with 10 – 20 % potassium sorbate solution [57]. Replacement of the nitrite used in meat processing by sorbic acid has been widely discussed. While sorbic acid is inactive towards *Clostridium* species in neutral medium, (i.e., under optimal growth conditions) it does suppress the formation of toxins very effectively at lower pH (i.e., pH ca. 5.5) in concentrations of 0.1 – 0.2 %. Sorbic acid can be considered as a good alternative to nitrite from a microbiological viewpoint provided sodium chloride is present as well. However, the red coloration and the pickling aroma preferred in cured meats do not develop in the absence of nitrite, which has prevented the successful use of sorbic acid in this application [58–61]. The use of potassium sorbate in combination with ascorbic or others acids gives a good and long curing [126–129].

The addition of potassium sorbate to fermenting vegetables leads to a particularly pure fermentation, because sorbic acid inhibits the undesirable proliferation of molds more than it does the desired lactic acid fermentation [62], [63]. Sorbic acid is also used for preserving all types of fruit products, including prunes, fruit pulps, marmelades and jams, fruit juices, jellies, and candies. Potassium sorbate in combination with sulfur dioxide is very important for stabilizing wine against undesired after-fermentation.

Baked goods can be protected against mold formation by the addition of sorbic acid to the dough. In yeast products (e.g., bread) sorbic acid may disrupt the action of the yeast, but a special form of the acid, characterized by a coarser particle size, shows virtually no detrimental effect on yeast activity [64].

Sorbic acid and potassium sorbate are acquiring increasing importance in the preservation of cosmetic products [130] (especially baby care products). Because of their good efficacy, low toxicity, and good skin compatibility they are favorably estimated in the literature [131–136].

In accordance with the "hurdle theory" [137] a combination of sorbates with modern physical preservation methods [138], [139] or other preservatives can be advantageous [140].

11. Economic Aspects

Sorbic acid and its salts, particularly potassium sorbate, have been produced industrially since the mid-1950s, primarily in the United States, Japan, and the Federal Republic of Germany. In the mid-1990s world demand exceeded 30 000 t/a.

12. Toxicology

The LD_{50} value of sorbic acid is ca. 10 g per kilogram body weight for rats [65]. Other authors give an LD_{50} of 7.4 g [141], 8.7 g [142], or 7.5 g [143] per kilogram body weight. As an acid, sorbic acid causes irritation to the mucous membranes, but skin irritation is observed only with very sensitive individuals.

Sorbic acid was first introduced as a food preservative at a time when extensive toxicological testing for new food additives was already mandatory. For this reason numerous studies have been published on acute, subacute, subchronic, and chronic tolerance (for a summary see [1, vol. II, pp. 13 – 31], [2, pp. 205 – 224] and [66], [67]). Sorbic acid has probably undergone the most thorough toxicological investigations of any preservative.

The allergenic potential of sorbic acid is extremely low, since, as a low molecular substance, it cannot cause an antibody response. Moreover, formation of covalent bonds with proteins, which might result in immediate hypersensitivity, is not known [144]. Pseudo-allergic reactions to sorbic acid as a food additive are relatively rare [145–147].

The presence of 10 % sorbic acid in rat feed causes increased growth and an increase in liver weight in the subchronic test, but no other damage [68]. The increase in liver weight observed at these high dose levels is interpreted as hypertrophy caused by excess work [69]. The increase in weight can be traced back to sorbic acid's caloric value.

Rats tolerate the lifelong addition of 5 % sorbic acid to their feed [69]. Carcinogenesis studies (rats, oral exposure, two years) have

shown that low concentrations of sorbic acid (1.5 – 10 % in the feed) produce no abnormalities compared with control animals with respect to growth, blood count, and the state and function of 12 internal organs. Adding 10 % sorbic acid to the feed leads to somewhat less weight gain and an increase in the size of the thyroid, liver, and kidneys [70]. Similar results have been obtained with mice [71].

Sorbic acid is not teratogenic [72]. All mutagenicity and genotoxicity investigations have also produced negative results [67], [73–77]. Sorbic acid is also not carcinogenic [70], [71].

Sorbic acid is metabolized like other fatty acids; degradation occurs via the typical β-oxidation pathway [69], [78, 79].

The allergy-producing potentials and phototoxicities of both sorbic acid and potassium sorbate are extraordinarily low. Just one case is of allergic reaction has been reported (a type 4 allergic reaction with a low impact on the patient) [147], although sorbic acid has been used for more than 40 years all over the world. The CIR (Cosmetic Ingredient Review) Expert Panel of the Cosmetic, Toiletry, and Fragrance Association (CTFA) has classified both substances as "safe" at concentrations and under conditions typical for cosmetic agents [80].

In the case of sodium sorbate, which is not utilized in foods because of its tendency to undergo oxidation, more recent studies have demonstrated a low level of genetic toxicity potential in vitro (e.g., micronuclei induction, sister chromatid exchanges). No analogous effects are observed with sorbic acid or potassium sorbate [81], [82], [135], [149].

13. References

General References

1. E. Lück, M. Jager: *Antimicrobial Food Additives - Characteristics, Uses, Effects* 2nd ed., Springer Verlag, Berlin 1996.
2. J. N. Sofos: *Sorbate Food Preservatives,* CRC Press, Boca Raton, Fla., 1989.

Specific References

3. A. W. Hofmann, *Justus Liebigs Ann. Chem.* **110** (1859) 129 – 140.
4. E. Letzig, W. Handschack, *Nahrung* **7** (1963) 591 – 605.
5. J. H. Bowie, D. W. Cameron, *J. Chem. Soc.* 1965, 5651 – 5657.
6. O. Doebner, *Ber. Dtsch. Chem. Ges.* **23** (1890) 2372 – 2377.
7. O. Doebner, *Ber. Dtsch. Chem. Ges.* **33** (1900) 2140 – 2142.
8. BASF, DE 881 299, 1939 (E. Müller).
9. The Best Foods, US 2 379 294, 1940 (C. M. Gooding).
10. R. F. Nystrom, W. G. Brown, *J. Am. Chem. Soc.* **69** (1947) 2548.
11. E. Lück, *Dtsch. Lebensm. Rundsch.* **50** (1962) 353 – 357.
12. Dragoco, DE 1 139 610, 1961 (G. Nowak).
13. R. Fittig, J. B. Barringer, *Ann. Chem. Pharm.* **161** (1872) 307 – 328.
14. E. H. Farmer, A. T. Healey, *J. Chem. Soc. (London)* 1927, 1064 – 1067.
15. T. Sabalitschka, H. Marx, *Naturwissenschaften* **46** (1959) 648.
16. L. Pekkarinen, P. Rissanen, *Suom. Kemistil. B* **39** (1966) no. 3, 50 – 56.
17. L. Pekkarinen, *Z. Lebensm. Unters. Forsch.* **139** (1968) 23 – 29.
18. D. A. Ledward, *Food Addit. Contam.* **7** (1990) 677 – 683.
19. O. Doebner, *Ber. Dtsch. Chem. Ges.* **33** (1900) 2140 – 2142.
20. Union Carbide, DE 1 043 314, 1953 (H. C. Chitwood, B. T. Freure).
21. Union Carbide, US 2 930 801, 1953 (A. E. Montagna, L. V. McQuillen).
22. Hoechst, DE 1 244 768, 1963 (H. Fernholz, G. Jacobsen).
23. Nippon Gosei, US 4 997 756, 1988 (M. Hasegawa, Y. Honda).
24. Montecatini, BE 571 889, 1957 (G. P. Chiusoli).
25. Montecatini, BE 586 166, 1958 (G. P. Chiusoli).
26. Montecatini, BE 590 945, 1959 (G. P. Chiusoli).
27. Nippon Gosei, US 4 022 822, 1974 (Y. Tsujino, M. Miyashita, T. Hashimoto).
28. Monsanto, EP 55 936, 1981 (J. P. Coleman, R. C. Hallcher, D. E. McMackins).
29. Monsanto, EP 83 914, 1981 (J. P. Coleman, R. C. Hallcher).
30. Monsanto, EP 83 304, 1981 (R. C. Hallcher).
31. J. P. Coleman et al., *Tetrahedron* **47** (1991) 809 – 829.
32. H. Hoberg, D. Schaefer, *J. Organomet. Chem.* **255** (1983) C 15 – C 17.
33. Hoechst, DE 1 042 573, 1956 (H. Fernholz, E. Mundlos).

34. Hoechst, DE 1 049 852, 1956 (H. Fernholz, E. Mundlos).
35. Nippon Gosei, DE 2 526 716, 1975 (K. Sekiyama, Y. Taga, S. Fujita).
36. Hoechst, DE 2 165 219, 1971 (H. Hey, H.-J. Arpe).
37. Hoechst, DE 1 059 899, 1956 (H. Fernholz).
38. Stamicarbon, US 3 056 830, 1961 (S. Koopal, U. Verstrijden, W. Pesch, J. J. M. Deumens).
39. Hoechst, DE 2 331 668, 1973 (H. Fernholz, H.-J. Schmidt, F. Wunder).
40. Hoechst, DE 1 618 357, 1967 (O. Probst, G. Roscher, H. Oehme).
41. Wacker, DE 2 103 051, 1971 (H. Liberda, G. Künstle).
42. Monsanto, EP 153 292, 1984 (H. C. Brown, R. P. Crowley, D. N. Heintz, J. R. Ryland).
43. Chemcell, CA 879 082, 1969 (R. S. Smith, E. L. Jeans).
44. J. L. Kinderlerer, P. V. Hatton, *Food Addit. Contam.* **7** (1990) 657 – 669.
45. S. Schnell, C. Wondrak, G. Wahl, B. Schink, *Biodegradation* **2** (1991) 33 – 41.
46. E. Lück, *Food Addit. Contam.* **7** (1990) 711 – 715.
47. J. N. Sofos, F. F. Busta, *J. Food Prot.* **44** (1981) 614 – 622.
48. J. N. Sofos, M. D. Pierson, J. C. Blocher, F. F. Busta, *Int. J. Food Microbiol.* **3** (1986) 1 – 17.
49. T. Eklund, *J. Appl. Bacteriol.* **48** (1980) 423 – 432.
50. T. Eklund, *J. Gen. Microbiol.* **131** (1985) 73 – 76.
51. H. Lück, E. Rickerl, *Z. Lebensm. Unters. Forsch.* **109** (1959) 322 – 329.
52. E.-M. Lukas, *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Hyg., Abt. 2* **117** (1964) 485 – 509.
53. K. H. Wallhäußer, E. Lück, *Dtsch. Lebensm. Rundsch.* **66** (1970) 88 – 92.
54. E. Becker, I. Roeder, *Fette Seifen Anstrichm.* **59** (1957) 321 – 328.
55. K. H. Wallhäusser, E. Lück, *Dtsch. Lebensm. Rundsch.* **66** (1970) 88 – 92.
56. E. Lück, *Dtsch. Lebensm. Rundsch.* **50** (1962) 353 – 357.
57. L. Leistner, I. Y. Maing, E. Bergmann, *Fleischwirtschaft* **55** (1975) 559 – 561.
58. M. C. Robach, J. N. Sofos, *J. Food. Prot.* **45** (1982) 374 – 383.
59. J. N. Sofos, F. F. Busta, *Food Technol.* **34** (1980) no. 5, 244 – 251.
60. J. N. Sofos, F. F. Busta, C. E. Allen, *J. Food Prot.* **42** (1979) 739 – 770, 1099.
61. E. Lück, *Fleischwirtschaft* **64** (1984) 727 – 733.
62. R. N. Costilow, W. E. Ferguson, S. Ray, *Appl. Microbiol.* **3** (1955) 341 – 345.
63. E. Lück, *Ind. Obst Gemüseverwert.* **51** (1966) 410 – 414.
64. Hoechst, EP 75 289, 1981 (E. Lück, K. H. Remmert).
65. H. J. Deuel, R. Alfin-Slater, C. S. Weil, H. F. Smyth, *Food. Res.* **19** (1954) 1 – 12.
66. R. Walker, *Food Addit. Contam.* **7** (1990) 671 – 676.
67. British Industrial Biological Research Association: *Sorbic Acid and its Common Salts, Toxicity,* Profile, 1987.
68. G. E. Demaree, D. W. Sjogren, B. W. McCashland, F. P. Cosgrove, *J. Am. Pharm. Assoc. Sci. Ed.* **44** (1955) 619 – 621.
69. K. Lang, *Arzneim. Forsch.* **10** (1960) 997 – 999.
70. I. F. Gaunt, K. R. Butterworth, J. Hardy, S. D. Gangolli, *Food Cosmet. Toxicol.* **13** (1975) 31 – 45.
71. R. J. Hendy et al., *Food Cosmet. Toxicol.* **14** (1976) 381 – 386.
72. Food and Drug Research Laboratories: *Teratologic Evaluation of FDA 73-4, Potassium Sorbate; Sorbistat in Mice and Rats,* PB-245 520, National Technical Information Service, US Department of Commerce, Springfield, Mass., 1975.
73. Litton Bionetics: *Mutagenic Evaluation of Compound FDA 73-4, Potassium Sorbate,* PB-245 434, National Technical Information Service, US Department of Commerce, Springfield, Mass., 1974.
74. R. Münzner, C. Guigas, H. W. Renner, *Food Chem. Toxicol.* **28** (1990) 397 – 401.
75. R. Jung, C. Cojocel, W. Müller, D. Böttger, E. Lück, *Food Chem. Toxicol.* **30** (1992) 1 – 7.
76. D. Schiffmann, J. Schlatter, *Food Chem. Toxicol.* **30** (1992) 669 – 672.
77. J. Schlatter et al., *Food Chem. Toxicol.* **30** (1992) 843 – 851.
78. H. J. Deuel et al., *Food. Res.* **19** (1954) 13 – 19.
79. M. Fingerhut, B. Schmidt, K. Lang, *Biochem. Z.* **336** (1962) 118 – 125.
80. Cosmetic, Toiletry and Fragrance Association (CTFA), *J. Am. Coll. Toxicol.* **7** (1988) 837 – 880.
81. J. Schlatter et al., *Food Chem. Toxicol.* **30** (1992) 843 – 851.
82. D. Schiffmann, J. Schlatter, *Food Chem. Toxicol.* **30** (1992) 669 – 672.
83. B. R. Thakur, R. K. Singh, S. S. Arya, *Food Rev. Int.* **10** (1994) 71 – 91.

84. F H. Lückmann, D. Melnick, *Food Res.* **20** (1955) 649 – 654.
85. W. Flak, R. Schaber, *Mitt. Klosterneuburg* **38** (1988) 10 – 16.
86. J. Kantasubrata, Imamkhasani, *ASEAN Food J.* **6** (1991) 155 – 158.
87. S. Küppers, *J. Assoc. Off. Anal. Chem.* **71** (1988) 1068 – 1071.
88. F. Olea Serrano, I. Lopes, N. Revilla., *J. Liqu. Chromatogr.* **14** (1991) 709 – 717.
89. C. Reifschneider, C. Klug, M. Jager, *Seifen Öle Fette Wachse* **120** (1994) 650 – 654.
90. U. Hagenauer-Hener, C. Frank, U. Hener, A. Mosandl, *Dtsch. Lebensm. Rundsch.* **83** (1990) 348 – 351.
91. L. Bui, C. Kooper, *J. Assoc. Off. Anal. Chem.* **70** (1987) 892 – 896.
92. J. Karovicova, J. Polonski, P. Simko, *Nahrung* **35** (1991) 543 – 544.
93. P. Cox, *Acta Crystallogr.* **C50** (1994) 1620 – 1622.
94. R. Joly, C. Amiard, *Bull. Soc. Chim. France* (1947) 139.
95. E. Philippi, *Monatsh. Chem.* **51** (1929) 277.
96. N. G. Poljanskij, *Chim. Prom.* (1963) no. 1 20.
97. G. M. Robinson, R. Robinson, *Nature (London)* **151** (1943) 195.
98. J. Klein, E. D. Bergmann, *J. Am. Chem. Soc.* **79** (1957) 3452.
99. L. Canonica, T. Bacchetti, *Gazz. Chim. Ital.* **83** (1953) 1043.
100. Tarchorminskie Zalkady Farmaceutiyczne "Polfa", PL 47 632 (I. Nagrodzka, B. Chechelska, B. Morawski, L. Brzechffa).
101. Nippon Gosei, DE 3 930 338 A1, 1990 (M. Hasegawa, Y. Honda).
102. Nippon Gosei, JP 05 032 583 A2, 1993 (I. Kakimoto, Y. Takehiko).
103. Nantong, CN 1 112 915 A, 1995 (Ding, Caifeng, Yao, Jungsheng).
104. Traian Vuia, RO 102 186 B1, 1991 (S. Tiberiu Mircea, S. Tiberiu Ladislau, S. Zlatimir).
105. J. J. Azukas, PhD Thesis Michigan State Univ. (1962).
106. H. J. Rehm, *Zentralblatt Bakteriol. Parasiten Infektion Hyg. II. Abt. 2* **121** (1967) 492 – 502.
107. G. K. York, *Food Res.* **20** (1955) 60 – 65.
108. H. J. Rehm, P. Wallhöfer, *Naturwissenschaften* **51** (1964) 13 – 14.
109. H. Lück, *Biochem Z.* **328** (1957) 411 – 419.
110. H. Lück, *Z. Lebensm. Unters. Forsch.* **108** (1958) 1 – 9.
111. W. Martoadiprowito, J. R. Whitaker, *Biochem. Biophys. Acta* **77** (1963) 536 – 544.
112. H. J. Rehm, E. M. Lukas, *Zentralblatt Bakteriol. Parasiten Infektion Hyg. II. Abt. 2* **117** (1963) 306 – 318.
113. T. Eklund, *J. Appl. Bakteriol.* **54** (1983) 383 – 389.
114. G. K. York: "*Studies on the Inhibition of Microbes by Sorbic Acid*," PhD Thesis University of California, Davis 1960.
115. L. O. Emard, R. H. Vaughn, *J. Bacteriol.* **63** (1952) 487.
116. T. A. Bell, J. L. Etchells, A. F. Borg, *J. Bacteriol.* **63** (1959) 573.
117. H. J. Rehm, *Z. Lebensmittel Unters. Forsch.* **115** (1961) 293.
118. M. Lerche, G. Reuter, *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Hyg. I. Abt. Orig* **179** (1960) 354.
119. M. L. Couceiro, B. Regueiro, *An. Bromatol. (Madrid)* **14** (1962) 263.
120. A. A. Lentzner, M. A. Toom, M. N. Woronina, M. E. Mikelsaar, *Lab Delo* (1967) 301.
121. R. H. Vaughn, L. O. Emard, *Bacteriol. Proc.* **54** (1954) 25.
122. G. K. York, R. H. Vaughn, *Food Res.* **20** (1955) 60.
123. H. Raj, *Canad. J. Microbiol.* **12** (1966) 191.
124. S. M. Finegold, E. E. Sweeny in: *Antimicrobial Agents Chemotherapy,* 1961, p. 911.
125. F. W. Beech, J. G. Carr, *J. Sci. Food Agric.* **11** (1960) 38.
126. Rhenus Rheinische Getränke Industrie Binz & Binz GmbH, DE 1 692 110, (P. Flesch, G. Bauer).
127. Armour Co., US 4 277 508, (U. Sato, A. Miller, L. J. Zimont.)
128. Rhenus Rheinische Getränke Industrie Binz & Binz GmbH, DE 1 119 640, (P. Flesch, J. Hader). US 3 099 566, (P. Flesch, J. Hader).
129. Rhenus Rheinische Getränke Industrie Binz & Binz GmbH, DE 1 155 313, (P. Flesch, J. Hader).
130. C. Reifschneider, C. Klug, M. Jager: "Sorbic acid - The gentle alternative," in: *Proceedings of the Preservatec Conference 1995,* Ziolkowsky Verlag, Augsburg pp. 1 – 15.
131. K. Schrader, *Dragoco Report* **17** (1970) 3.
132. E. Birggal, *Kosmetik Parfum Drogen Rdsch.* **17** (1970) 1/2, 5.
133. R. Goldschmidt, *Seifen Öle Fette Wachse* **96** (1970) A2, 7.
134. R. Woodford, E. Adams, *Amer. Perfumer Cosmetics* **85** (1970) no. 3, 25.
135. M. A. Kassem, A. G. Mattha, *Pharmac. Acta Helvetiae* **45** (1970) 345.
136. A. Schunk, *Seifen Öle Fette Wachse* **122** (1996) 136 – 144.

137. L. Leistner, *Fleischwirtschaft* **66** (1986) 10 – 15.
138. T. Vardag, M. Jager, *Food Marketing Technology* **10** (1996) 58 – 67.
139. T. Vardag, M. Jager, in: *Proceedings of the Preservatec Conference 1996,* Ziolkowsky Verlag, Augsburg pp. 33 – 40.
140. T. Augustin et al., *Getränkeindustrie* **6** (1996) 407.
141. H. F. Smyth, C. P. Carpenter, *J. Ind. Hyg. Tox.* **30** (1948) 63 – 68.
142. I. Sado, *Nippon Fisagaku Zasshi* **28** (1973) 463 – 476.
143. J. A. Troller, R. A. Olsen, *J. Food Sci.* **32** (1967) 228 – 232.
144. S. Vieths, K. Fischer, L. I. Dehne, K. W. Bögl, *Ernährungs Umschau* **41** (1994) 140 – 143, 186 – 190.
145. L. Rosenhall, *Eur. J. Respir. Dis.* **63** (1982) 410 – 419.
146. M. Hannuksela, T. Haahtela, *Allergy* **42** (1987) 56 – 575.1
147. M. Häberle, *Ernährungs Umschau* **36** (1989) 8 – 16.
148. C. S. Koch, *Seifen Öle Fette Wachse* **120** (1994) 655 – 660.
149. F. Würgler, J. Schlatter, P. Maier, *Mutat. Res.* **283** (1992) 107 – 111.

Sorbitol → Sugar Alcohols
Sorbose → Vitamins
Soybean Oil → Fats and Fatty Oils
Spandex → Fibers, 4.Synthetic Organic

Spasmolytics

EDWIN BUNGARDT, Pharmakologisches Institut für Naturwissenschaftler der J. W. Goethe-Universität, Frankfurt/Main, Federal Republic of Germany

ERNST MUTSCHLER, Pharmakologisches Institut für Naturwissenschaftler der J. W. Goethe-Universität, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Parasympatholytics (Neurotropic Spasmolytic Agents) 2
2.1. Muscarinic Receptors 3
2.2. Therapy 3
2.3. Therapeutic Agents 5
2.3.1. Tropane Alkaloids and Related Substances 5
2.3.2. Substances with Diverse Chemical Structures 6
3. Musculotropic and Neurotropic–Musculotropic Spasmolytic Agents 9
3.1. Musculotropic Spasmolytic Agents 10
3.2. Neurotropic – Musculotropic Spasmolytic Agents 12
4. Miscellaneous Agents 15
5. Future Prospects 15
6. References 15

1. Introduction

Spasmolytic agents are substances that relieve spasms of smooth muscle. This effect can be achieved by compounds that attack the parasympathetic nervous system (parasympatholytic = neurotropic spasmolytic agents) or by substances that act directly on the smooth muscles (musculotropic spasmolytic agents). Like the heart and glands, the smooth muscle of the intestines, blood vessels, and eyes is innervated by the autonomic (visceral) nervous system (→ Neuropharmacology). The peripheral autonomic system consists of three parts: the sympathetic, the parasympathetic, and the enteric nervous systems (Fig. 1). The sympathetic and parasympathetic peripheral nerve pathways are each composed of two neurons. The first (preganglionic) neuron conducts impulses from the central nervous system to an autonomic ganglion; here transmission is effected to the second (postganglionic) neuron, leading ultimately to the organ of response.

Transmission of impulses in all the ganglia is mediated by the neurotransmitter acetylcholine which especially activates nicotinic receptors as well as secondary muscarinic M_1/M_2 receptors (see below and → Neuropharmacology, Chap. 5.1.) located on cell bodies of the postganglionic neurons and neurons of the enteric nervous system. Blockage of the nicotinic receptors is achieved by means of *ganglia blockers*, most of which are quaternary ammonium salts. The transmission of postganglionic neuron impulses to the organ of response occurs, in sympathetic fibers, via noradrenaline as the neurotransmitter together with adrenoceptors. In parasympathetic fibers, however, acetylcholine is the transmitter and the receptors are exclusively muscarinic (see Section 2.1).

The enteric nervous system includes neurons in the walls of the gastrointestinal tract from the beginning of the esophagus to the sphincter ani internus, as well as neurons of the pancreas and the bile duct. It controls motility and secretion in all the various sections of the food passage. The enteric nervous system is capable of carrying out its function without any influence from the central nervous system. Indeed, the sympathetic and parasympathetic systems exert only modulating effects. In many neurons, especially neurons of the enteric nervous system, there are other transmitter substances in addition to the classical neurotransmitters noradrenaline and acetylcholine. These substances are released either alone or in combination with noradrenaline and acetylcholine (cotransmission) [1–4]. Examples of acetylcholine cotransmitters include vasoactive intestinal polypeptide (VIP) and sub-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_515

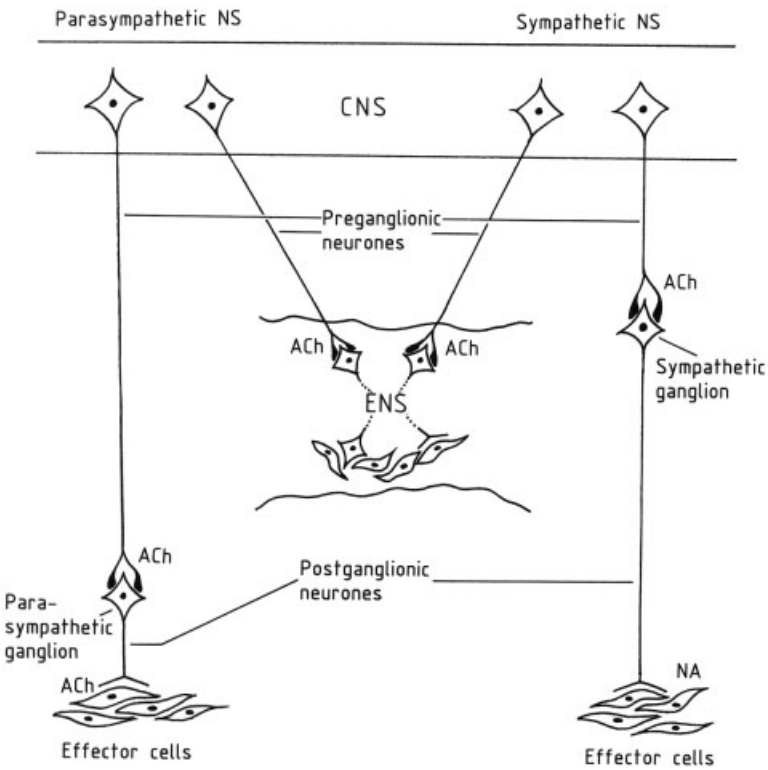

Figure 1. Simplified scheme of the autonomic nervous system
NS = nervous system; CNS = central nervous system; ENS = enteric nervous system; ACh = acetylcholine; NA = noradrenaline

stance P, which belongs to the group known as tachykinins.

2. Parasympatholytics (Neurotropic Spasmolytic Agents)

Parasympatholytics block the acetylcholine-mediated transmission of impulses from postganglionic parasympathetic fibers to the organ of response by competitive antagonism at the muscarinic receptors. Due to the resulting relaxing effect on the smooth muscles, these substances are also called *neurotropic spasmolytic agents*.

2.1. Muscarinic Receptors

It has been possible with the help of functional pharmacological investigations and radio-ligand binding studies to identify four subtypes of muscarinic receptors ($M_1 - M_4$). This heterogeneity of the muscarinic receptors was confirmed in an impressive manner by the cloning, sequencing, and expression of five muscarinic receptor genes ($m_1 - m_5$). A great similarity was found between the binding properties for muscarinic ligands coupling to various effector systems, and the localization of pharmacologically characterized native $M_1 - M_4$ receptors, on the one hand and the molecularly defined $m_1 - m_4$ receptors (the $m_1 - m_4$ gene products) on the other. Pharmacological characterization of the (molecularly defined) m_5 receptor has not yet been accomplished ([5]; for reviews see [6–9]). All the subtypes known to date belong to the "superfamily" of membrane-bound receptors coupled to G-proteins (see review articles [10], [11]).

Muscarinic receptors are widely distributed in the central and peripheral nervous systems as well as in the parasympathetically innervated organs of response. All five subtypes occur in the central nervous system. M_1 and M_2 receptors have been found in the sympathetic and parasympathetic ganglia and in the interneu-

ronal synapses of the enteric nervous system; these modulate the neurotransmission that is primarily mediated at these sites via nicotinic receptors. The function of peripheral effector organs or cells is influenced largely by the presence of postsynaptic M_2 and M_3 receptors: the heart possesses predominantly M_2 receptors which, when stimulated, cause a decrease in both the heart rate and the force of contraction. M_3 receptors on gland cells mediate increases in secretion. The smooth muscle of the respiratory tract, gastrointestinal tract, and urinary bladder contains a heterogeneous population of M_2 (high density) and M_3 (low density) receptors. However, the receptor that mediates contraction of these organs is, in all probability, the M_3 subtype, while the function of the M_2 receptors on these organs is still not understood ([12]; for review articles, see [6–9]). Figure 2 shows the largely proved intracellular reaction sequence that leads finally to contraction of smooth muscle after stimulation of the M_3 receptors [6], [13]. After binding the endogenous agonist acetylcholine (ACh), the receptor stimulates phospholipase C via a G-protein (G). This cleaves a special membrane phospholipid (phosphatidylinositol-4,5-diphosphate, PIP_2) to give inositol-1,4,5-triphosphate (IP_3) and diacylglycerol (DAG). The Ca^{2+} ions released by IP_3 from the sarcoplasmic reticulum (SR) trigger contractions of the contractile apparatus of the cell. Phosphorylation of IP_3 to IP_4, and the protein kinase C activated by DAG probably cause an additional Ca^{2+} influx from the extracellular space.

Muscarinic antagonists compete with the endogenous agonist acetylcholine for its binding site on the receptor (i.e., they also possess affinity). However, in contrast to acetylcholine, they do not trigger any successive reactions (i.e., they lack intrinsic activity). This is the mode of action for neurotropic spasmolytic agents.

However, none of the substances described below possesses the desirable property of subtype-selective blockage of the M_3 receptors that mediate contraction. The largely nonselective blockage of all muscarinic receptor subtypes (including, for example, cardiac M_2 receptors) explains some of the side effects and treatment restrictions associated with classical parasympatholytics. For the control of spasms of smooth muscle, the development of subtype-selective antagonists would open the way to more efficient therapy with fewer side effects (see Chap. 5).

2.2. Therapy

Parasympatholytics are indicated for:

1) Spasms of the smooth muscle of the gastrointestinal tract (e.g., in the case of stomach and duodenal ulcers, spastic obstipation, and pylorospasm)
2) Spasms of the bile and urinary duct (e.g., gall-stone and kidney-stone colics)
3) Spasms in the region of the female genital organs (e.g., severe dysmenorrhea)
4) Spasms in the region of the respiratory tract (especially in bronchial asthma; see → Antiasthmatic Agents)
5) Urinary incontinence as a result of detrusor instability and "neurogenic bladder"

Parasympatholytics are used not only as antispasmodic agents, but also in the treatment of Parkinson's disease (→ Parkinsonism Treatment), in bradycardic arrhythmias, and in the preparation of patients for anesthesia. Although classical parasympatholytics were used frequently in the past for the treatment of stomach and duodenal ulcers, they have in the course of time been replaced by pirenzepine [*28797-61-7*], an antagonist affecting mainly M_1 receptors, as well as other ulcer drugs (histamine-H_2 antagonists and inhibitors of H^+/K^+-ATPase; see → Antiulcer Drugs).

Parasympatholytics containing tertiary nitrogen are rapidly absorbed when administered orally. On the other hand, the absorption of quaternary ammonium compounds is poor. Achieving an adequate effect with these compounds frequently requires parenteral application or oral doses that are substantially higher in comparison with i.v. administration. The main advantage of quaternary ammonium compounds is that they are virtually incapable of crossing the blood – brain barrier because of their low lipid solubility; therefore, unlike the tertiary compounds, they are centrally inactive. However, the quaternary compounds often also act as antagonists of nicotinic receptors. Therefore, when high doses are administered, symptoms of ganglia blockage can supervene on the effects of muscarinic-receptor blockage.

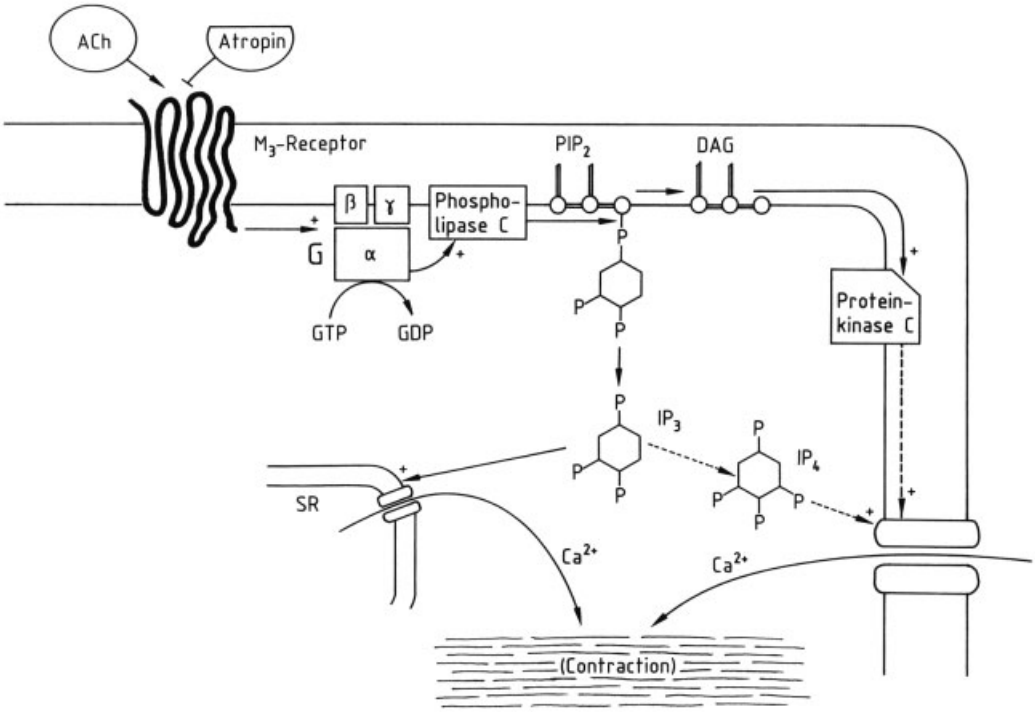

Figure 2. Schematic representation of the intracellular reaction sequence which, after stimulation of the M_3 receptors, ultimately causes contraction of the smooth muscles [modified after K. Starke and D. Palm (1992)].

As a result of the absence of receptor and organ selectivity in classical parasympatholytics, the following side effects must be expected:

1) Increase in the cardiac rate
2) Reduced secretion of saliva
3) Dilation of the pupils and disturbances of visual accommodation
4) Disturbances of urination

Parasympatholytics are contraindicated in narrow-angle glaucoma and prostate adenoma. In the presence of coronary heart disease, these substances may be administered only in doses that do not increase the cardiac rate.

2.3. Therapeutic Agents

Parasympatholytics can be divided chemically into tropane alkaloids and related substances (atropine, homatropine, scopolamine, and their quaternary derivatives, see→ Alkaloids, Chap. 4.) and substances with diverse chemical structures.

2.3.1. Tropane Alkaloids and Related Substances

Atropine [*51-55-8*], 3α(1α*H*,5α*H*)-tropanyl-(±)-tropate, (±)-hyoscyamine, $C_{17}H_{23}NO_3$, M_r 289.38, (±)-form *mp* 114–116 °C, (−)-form *mp* 108.5 °C, $[\alpha]_D^{20}$ −21.0° (EtOH); methobromide [*2870-71-5*], $C_{18}H_{26}BrNO_3$, M_r 384.33, *mp* 222–223 °C; methonitrate [*52-88-0*], $C_{18}H_{26}N_2O_6$, M_r 366.42, *mp* 163 °C; sulfate [*55-48-1*] (anhydrous), [*5908-99-6*] (monohydrate), $(C_{17}H_{23}NO_3)_2$, H_2SO_4, H_2O, M_r 694.82, *mp* 190–194 °C.

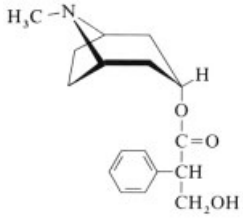

Atropine is a tropane alkaloid that occurs in *Atropa belladonna L.*, *Datura stramonium L.*, and other *Solanaceae* species, and it is still

one of the most important parasympatholytic agents (for synthesis, see [14] and → Alkaloids, Chap. 4.).

Trade Names. Atropine: Atropinol (Winzer, FRG); borate: Borotropin (Winzer, FRG); methonitrate: Ansudoral (Basotherm, FRG), Tonaton (Luitpold, FRG); sulfate: Atropen (Survival, USA), Atropisol (Cooper Vision, USA), Atropocil (Edol, Portugal), Atrosol (Liba, Turkey), Isopto-Atropine (Alcon: Sweden, UK, USA), Liotropina (Sifi, Italy), MIN-I-JET Atropinsulfat (Braun, FRG), Skiatropine Blache (Novopharma, Switzerland).

Butropium bromide [*29025-14-7*], 8-(4-butoxybenzyl)-3α-[(*S*)-(−)-tropoyloxy]-1αH,5α *H*-tropanium bromide, $C_{28}H_{38}BrNO_4$, M_r 532.53, *mp* 166 – 168 °C. Synthesis: the alkaloid hyoscyamine (→ Alkaloids, Chap. 4.1.) is quaternized with 4-butoxybenzyl bromide in 2-propanol [15]. White crystals are obtained, which have a bitter taste and are readily soluble in water. Used in Japan, butropium bromide is a quaternary derivative of atropine with atropine-like peripheral effects.

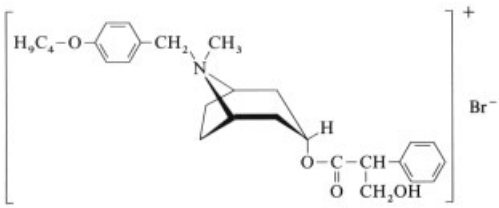

Trade Name. Coliopan (Eisai, Japan).

Butylscopolaminium bromide [*149-64-4*], *N*-butylscopolaminium bromide, hyoscine butylbromide, *N*-butyl-6β,7β-epoxy-3α-(−)-tropoyloxy-1αH,5αH-tropanium bromide, $C_{21}H_{30}BrNO_4$, M_r 440.40, *mp* 142 – 144 °C, $[\alpha]_D^{20}$ −20.8° (3 % in water). Synthesis: the alkaloid scopolamine (→ Alkaloids, Chap. 4.1.) is quaternized as described in [16] to produce white, odorless, light-sensitive crystals, which have a bitter taste and are highly soluble in water and ethanol. Butylscopolaminium bromide is a frequently used quaternary derivative of scopolamine; it was introduced to therapy in 1950.

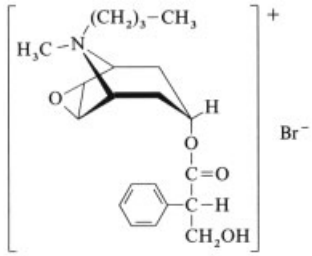

Trade Names. Buscapina (Boehringer Ingelheim,Spain), Buscopan (Boehringer Ingelheim: Belgium, Denmark, FRG, Greece, Italy, the Netherlands, Portugal, Switzerland; Delagrange, France), Buscotek (Tek, Turkey), Hyosin (Otsuka, Japan), Moryspan (Beppu, Japan), Vagotrope S (Infar, Portugal).

Homatropine methylbromide [*80-49-9*], 8-methyltropinium bromide mandelate, $C_{17}H_{24}BrNO_3$, M_r 370.29, *mp* 191 – 192 °C (partial decomp.). Homatropine, which was first used therapeutically in 1883, has gained wide acceptance as a locally applied mydriatic, but the *N*-quaternary methylbromide [17] is seldom used today.

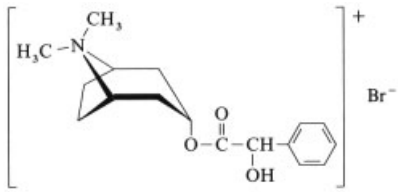

Trade Names. Nopar (Uni-Pharma, Greece), Novatropina (Asta-Medica, Italy).

Octatropine methylbromide [*80-50-2*], *N*-methyl-*O*-(2-propylvaleryl)tropinium bromide, $C_{17}H_{32}BrNO_2$, M_r 362.37, *mp* 329 °C; methylchloride $C_{17}H_{32}ClNO_2$, *mp* 289 °C. For synthesis, see [18].

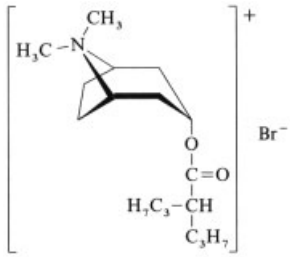

Trade Names. Valpin (Endo, USA; Sankyo, Japan).

Trospium chloride [*10405-02-4*], 3α-benziloyloxyspiro[nortropane-8,1′-pyrrolidinium] chloride, 3α-hydroxyspiro[1αH,5αH-nortropane-8,1′-pyrrolidinium] chloride benzilate, $C_{25}H_{30}ClNO_3$, M_r 427.97, *mp* 255 – 257 °C (decomp.). It is prepared by heating the benzilic acid ester of nortropine with 1,4-dichlorobutane in acetonitrile [19].

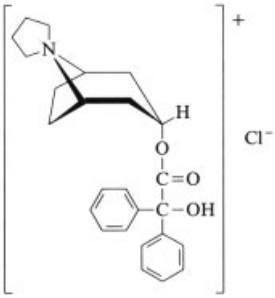

Trade Names. Relaspium (Isei, Japan), Spasmex (Pfleger, FRG), Spasmo-lyt (Madaus, FRG); present in Spasmo-Urgenin (Madaus, FRG).

2.3.2. Substances with Diverse Chemical Structures

Clidinium bromide [*3485-62-9*], 3-benziloyloxy-1-methylquinuclidinium bromide, $C_{22}H_{26}BrNO_3$, M_r 432.38, *mp* 240 – 241 °C. Synthesized by the reaction of diphenylglycolyl chloride with 3-hydroxyquinuclidine, followed by quaternization with methyl bromide. 3-Hydroxyquinuclidine is in turn formed by the reduction of 3-oxoquinuclidine [20]. White, slightly light-sensitive crystals, are formed which are soluble in water and ethanol. Clidinium bromide was first used therapeutically in 1953. It is administered either alone or in combination with chlordiazepoxide in the symptomatic treatment of peptic ulcers and other gastrointestinal functional disorders.

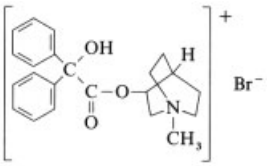

Trade Names. Quarzan (Hoffmann-La Roche, USA), contained in Clipoxide (Schein, USA), Corium (ICN, Canada), Librax (Hoffmann-La Roche: Australia, Canada, FRG, USA), Libraxin (Hoffmann-La Roche, UK).

Diphemanil methylsulfate [*62-97-5*], 4-benzhydrylidene-1,1-dimethylpiperidinium methylsulfate, $C_{21}H_{27}NO_4S$, M_r 389.50, *mp* 194 – 195 °C. For synthesis, see [21], [22]. It is soluble in water (1 : 33), ethanol, and chloroform. Diphemanil is also used externally as a 2 % cream in the treatment of profuse perspiration.

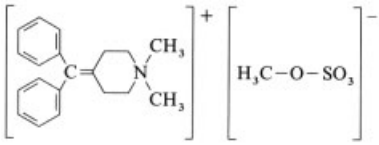

Trade Names. Prantal (Schering-Plough, Italy; Unicet, France).

Emepronium bromide [*3614-30-0*], (±)-*N*-ethyl-*N*,*N*-dimethyl-(4,4-diphenyl-2-butyl) ammonium bromide, $C_{20}H_{28}BrN$, M_r 362.37, *mp* 204 °C. For synthesis, see [23], [24]. Emepronium bromide is used in the racemic form mainly in the treatment of urinary incontinence and spasms of the bladder and ureter. The enantiomers of this substance show practically no receptor- or stereoselectivity in vitro [25].

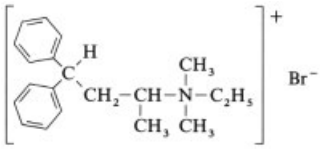

Trade Names. Cetiprin (Globopharm, Switzerland; Kabi Pharmacia: Austria, Denmark, Sweden; KabiVitrum: Finland, UK), Detrulisin (Glaxo-Allen, Italy); as the carrageenate: Uro-Ripirin (Fresenius, FRG).

Fenpiverinium bromide [*125-60-0*], 1-(3-carbamoyl-3,3-diphenylpropyl)-1-methylpiperidinium bromide, $C_{22}H_{29}BrN_2O$, M_r 417.41,

mp 177 – 179 °C. The synthesis is described in [26].

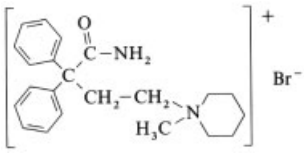

Glycopyrronium bromide [*596-51-0*], (±)-3-(α-cyclopentylmandeloyloxy)-1,1-dimethylpyrrolidinium bromide, glycopyrrolate, $C_{19}H_{28}BrNO_3$, M_r 398.36, *mp* 193.2 – 194.5 °C. Synthesized by the esterification of α-cyclopentylmandelyl chloride with 1-methyl-3-pyrrolidinol, followed by quaternization [27]. It is soluble in water (1 : 4.2) and ethanol (1 : 30), slightly soluble in chloroform, and practically insoluble in ether.

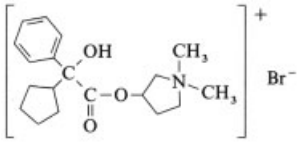

Trade Names. Robinul (Robins: Finland, Switzerland, Sweden, UK, USA; Brenner/Efeka, FRG; Kaken, Japan; Rorer, the Netherlands; Wyeth, Denmark), Tarodyl (Lundbeck, Finland), Tarodyn (Lundbeck, Sweden).

Hexocyclium methylsulfate [*115-63-9*], 4-(β-cyclohexyl-β-hydroxyphenethyl)-1,1-dimethylpiperazinium methylsulfate, $C_{21}H_{36}N_2O_5S$, M_r 428.59, *mp* 200 – 210 °C. For synthesis, see [28].

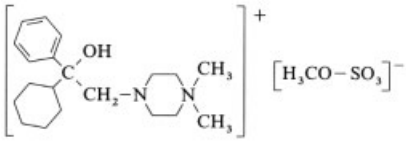

Trade Name. Tral (Abfar, Turkey; Abbott, USA).

Isopropamide iodide [*71-81-8*], (3-carbamoyl-3,3-diphenylpropyl)diisopropylmethylammonium iodide, $C_{23}H_{33}IN_2O$, M_r 480.42, *mp* 198 – 201 °C (decomp.). Formed by the reaction of diphenylacetonitrile with diisopropylaminoethyl chloride, followed by hydrolysis of the nitrile to the amide function and quaternization [29]. White to yellowish, odorless, light-sensitive crystals are obtained, which have a bitter taste and are soluble in water (1 : 50), ethanol (1 : 10), and chloroform (1 : 5), but only very slightly soluble in ether. This quaternary anticholinergic, which is mainly found in combination preparations, exhibits atropine-like peripheral effects. The antispasmodic and antisecretory effect of a single dose of 5 – 10 mg lasts up to 12 h.

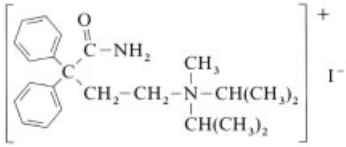

Trade Names. Darbid (Smith Kline & French: Canada, USA), Priamide (Janssen: Belgium, the Netherlands), Tyrimide (Smith Kline & French: Australia, UK); present in Combid (Smith Kline & French: Canada, USA), Prochlor-Iso (Schein, USA), Pro-Iso (Geneva, USA), Stelabid (Smith Kline & French: Australia, Canada, UK).

Mepenzolate bromide [*76-90-4*], 3-benziloyloxy-1,1-dimethylpiperidinium bromide, $C_{21}H_{26}BrNO_3$, M_r 420.37, *mp* 228 – 229 °C (decomp.). Made from benzilic acid and 3-chloro-1-methylpiperidine, followed by quaternization of the resulting intermediate with methyl bromide [30]. It is a white or yellowish powder, soluble in water (1 : 110), absolute ethanol (1 : 120), and methanol (1 : 8).

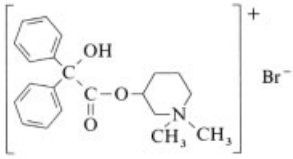

Trade Names. Cantil (Sigma, Australia; Draco, Sweden; MCP, UK; Merrel Dow: Belgium, USA; Roger Bellon, France), Cantril, Colibantil, Colum (manufacturer unknown, Italy),

Tralanta (Sawai, Japan), Trancolon (Fujisawa, Japan), Trokonil (Taiyo, Japan).

Methanthelinium bromide [*53-46-3*], *N*,*N*-diethyl-*N*-methyl-2-(9-xanthenylcarbonyloxy) ethylammonium bromide, $C_{21}H_{26}BrNO_3$, M_r 420.36, *mp* 175 – 176 °C. Prepared from 9-xanthenecarboxylic acid and 2-diethylaminoethanol [31]. Methanthelinium bromide has been largely replaced in clinical practice by its more potent derivative propantheline bromide.

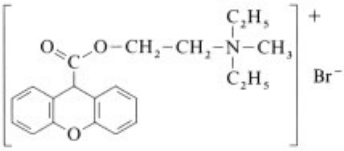

Trade Names. Banthine (Searle, USA), Vagantin (Palmicol, FRG).

Otilonium bromide [*26095-59-0*], *N*,*N*-diethyl-*N*-methyl-*N*-{2-[4-(2-octyloxybenzamido)-benzoyloxy]ethyl}ammonium bromide, $C_{29}H_{43}BrN_2O_4$, M_r 563.57, *mp* 166 – 168 °C. For synthesis, see [32].

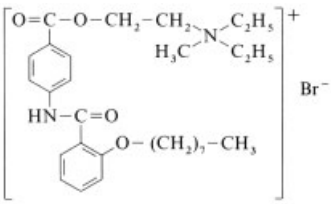

Trade Names. Spasmoctyl (Menarini, Spain), Spasmomen (Menarini, Italy).

Oxapium iodide [*6577-41-9*], (±)-1-[(2- cyclohexyl-2-phenyl-1,3-dioxolan-4-yl)methyl]-1-methylpiperidinium iodide, cyclonium iodide, $C_{22}H_{34}INO_2$, M_r 471.42, *mp* 195 – 197 °C (α-form, from isopropanol or chlorobenzene), *mp* 150 – 152 °C (β-form, from chlorobenzene). For synthesis, see [33].

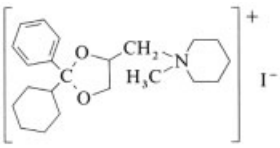

Trade Name. Esperan (Toyama, Japan).

Oxyphenonium bromide [*50-10-2*], *N*,*N*-diethyl-*N*-[2-(α- cyclohexylmandeloyloxy) ethyl]- *N*-methylammonium bromide, $C_{21}H_{34}BrNO_3$, M_r 428.41, *mp* 189 – 194 °C. For synthesis, see [34].

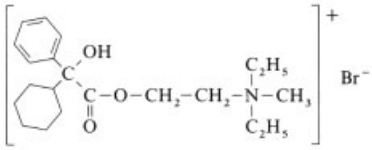

Trade Names. Antrenyl (Ciba: Belgium, the Netherlands, Switzerland, Turkey, USA), Spasmophen (Polfa, Poland), Subranyl (Vita, Greece)

Parapenzolate bromide [*5634-41-3*], 4-benziloyloxy-1,1-dimethylpiperidinium bromide, $C_{21}H_{26}BrNO_3$, M_r 420.37, *mp* 231 – 232 °C. For synthesis, see [35]. Parapenzolate bromide is a structural isomer of mepenzolate bromide, with similar activity.

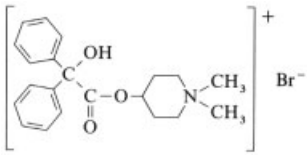

Trade Name. Spacine (Unicet, France).

Prifinium bromide [*4630-95-9*], 3-(diphenylmethylene)-1,1-diethyl-2-methylpyrrolidinium bromide, $C_{22}H_{28}BrN$, M_r 386.39, *mp* 216–218 °C. For synthesis, see [36].

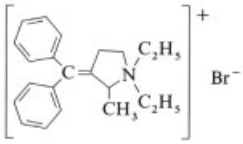

Trade Names. Padrin (Fujisawa, Japan), Riabal (Logeais, France; IBI, Italy).

Propantheline bromide [*50-34-0*], *N*,*N*-diisopropyl-*N*-methyl-*N*-[2-(9-xanthenecarbonyloxy)-ethyl]ammonium bromide, $C_{23}H_{30}BrNO_3$, M_r 448.42, *mp* 159 – 161 °C. Prepared by reacting xanthene-9- carboxylic acid chloride with 2-diisopropylaminoethanol,

followed by alkylation with methyl bromide [31]. White, odorless crystals, freely soluble in water, ethanol, and chloroform. Propantheline bromide was introduced to clinical practice in 1953, and is used in the treatment of peptic ulcers, visceral spasms, and enuresis.

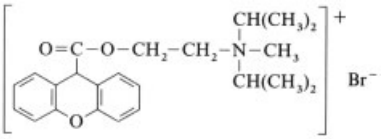

Trade Names. Corrigast (Heumann, FRG), Ercoril (Ercopharm: Denmark, Finland), Pro-Banthine (Searle: Belgium, Denmark, France, Greece, the Netherlands, Sweden, Switzerland, USA).

Tiemonium iodide [*144-12-7*], (±)-4-[3-hydroxy-3-phenyl-3-(2-thienyl)propyl]-4-methylmorpholinium iodide, $C_{18}H_{24}INO_2S$, M_r 445.38, *mp* 189 – 191 °C; methylsulfate $C_{19}H_{27}NO_6S_2$, M_r 429.56. For synthesis, see [37].

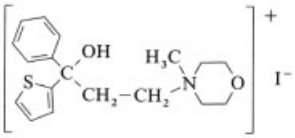

Trade Names. Visceralgina (SIT, Italy), Visceralgine (Exel Pharma, Belgium; Remek, Greece).

Tridihexethyl chloride [*4310-35-4*], (±)-*N*-(3-cyclohexyl-3-hydroxy-3-phenylpropyl)-*N*-,*N*-,*N*-triethylammonium chloride, $C_{21}H_{36}ClNO$, M_r 353.97, *mp* 196 – 202 °C; iodide [*125-99-5*], $C_{21}H_{36}INO$, M_r 445.44, *mp* 179 – 184 °C. For synthesis, see [38].

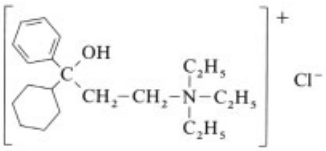

Trade Name. Pathilon (Lederle, USA).

3. Musculotropic and Neurotropic–Musculotropic Spasmolytic Agents

Independent of autonomic innervation, smooth muscle can also be relaxed by direct action on the muscle cells. Drugs that act in this manner are named musculotropic or papaverine-like spasmolytic agents, after the opium alkaloid papaverine, which was the first known example.

Beginning in 1930, systematic modification of the papaverine molecule led to numerous other substances, including ethaverine and moxaverine, that are clearly superior to papaverine in potency but less toxic. Substances introduced later into clinical practice deviated more and more from the structure of the opium alkaloid (e.g., mebeverine, camylofin, and hymecromone). Interestingly, it was assumed as early as 1934 that papaverine-like spasmolysis depends less on the structure of the substance than on its ability to reduce surface tension [39], [40]. The mechanism of action of papaverine is not yet fully understood. Effects that have been detected include nonselective inhibition of phosphodiesterases [41] with a subsequent increase in the intracellular concentration of cyclic adenosine monophosphate (cAMP), a sodium-dependent cAMP-mediated Ca^{2+} efflux [42], the inhibition of extracellular Ca^{2+} influx (calcium channel antagonism), interaction with the intracellular release and storage of Ca^{2+} (papaverine and thiphenamil [43]), and inhibition of adenosine reuptake carriers [44]. Terolidine and oxybutynin exhibit not only parasympatholytic properties, but also marked calcium-channel blocking properties [45], [46]. In this respect they resemble the "calcium antagonists" used in hypertension and coronary insufficiency (→Calcium Antagonists).

The relaxing effect of musculotropic spasmolytic agents extends to all smooth muscle; i.e., to the vascular smooth muscle as well as the bronchial and intestinal smooth muscle. Consequently, preparations from within this group of substances are indicated in the treatment of spasms of the gastrointestinal tract, uterus, bronchi, bile duct, and urinary passage. As a result of the nonselective relaxing effect on smooth muscles described above, cardiovascular side effects (fall in blood pressure, arrhythmias) must

be anticipated, especially with intravenous administration.

Numerous substances, such as oxybutynin or camylofin, occupy an intermediate position between parasympatholytics and musculotropic spasmolytics. Since they show both parasympatholytic properties and the ability to relax smooth muscles directly they are also called *neurotropic – musculotropic spasmolytic agents*.

3.1. Musculotropic Spasmolytic Agents

Papaverine [*58-74-2*], 6,7-dimethoxy-1-(3,4-dimethoxybenzyl)isoquinoline, $C_{20}H_{21}NO_4$, M_r 339.38, *mp* 147 °C; hydrochloride [*61-25-6*], $C_{20}H_{22}ClNO_4$, M_r 375.84, *mp* 220 – 225 °C; sulfate [*32808-09-6*], $(C_{20}H_{21}NO_4)_2\ H_2SO_4$, M_r 866.90. Papaverine is a companion alkaloid (ca. 0.8 %) of opium. Numerous syntheses have been described (→ Alkaloids, Chap. 4.2.). The hydrochloride is soluble in water (1 : 30 – 1 : 40), ethanol (1 : 120), and chloroform, but practically insoluble in ether. As a result of its weak effect, papaverine is seldom used today as an antispasmodic agent, but it has regained importance in the treatment of erectile malfunctions (intracavernous self injection, SKAT therapy).

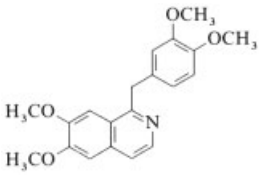

Trade Names. Cerebid (Saron, USA), Myobid (Laser, USA), Opdensit (Pharma Stroschein, FRG), Panergon (Mack/Illertissen, FRG; Pfizer, Switzerland), Pavabid (Marion, USA), Pavakey (Key, USA), Paveron (Linden, FRG).

Ethaverine [*486-47-5*], 6,7-diethoxy-1-(3,4-diethoxybenzyl)isoquinoline, ethylpapaverine, $C_{24}H_{29}NO_4$, M_r 395.48, *mp* 99 – 101 °C; hydrochloride [*985-13-7*], $C_{24}H_{30}ClNO_4$, M_r 431.94, *mp* 186 – 188 °C (decomp.). For synthesis, see [47] and [48]. Ethaverine is the tetraethyl homolog of papaverine.

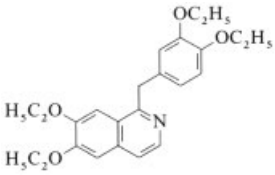

Trade Names. United States: Circubid (Merchant), Eta-Lent (Roger), Ethaquin (Ascher), Ethatab (Glaxo), Ethaverol (Pharmex), Isovex (US Chemical), Pavaspan (Jamieson-McKames), Spasodil (Rand); Elzym (Luitpold, FRG), Olren (Heumann, FRG).

Moxaverine [*10539-19-2*], 1-benzyl-3-ethyl-6,7-dimethoxyisoquinoline, Eupaverin, $C_{20}H_{21}NO_2$, M_r 307.38, *mp* 78 – 79 °C; hydrochloride [*1163-37-7*], $C_{20}H_{22}ClNO_2$, M_r 343.84, *mp* 208 – 214 °C (decomp.). For synthesis, see [49].

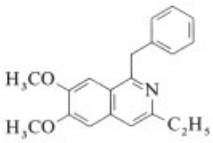

Trade Names. Eupaverin (Merck: Sweden, Switzerland), Eupaverina (Bracco, Italy), Kollateral (Ursapharm, FRG).

Benzyl mandelate [*890-98-2*], (±)-benzyl 2-phenylglycolate, $C_{15}H_{14}O_3$, M_r 242.27, *mp* 93.6 °C. Colorless, odorless crystals are formed from reaction of the sodium salt of mandelic acid with benzyl chloride, or by esterification of the acid [50]. The crystals are insoluble in water but readily soluble in ethanol, acetone, ether, and chloroform. Like ethyl, isopentyl, and methyl mandelates, benzyl mandelate exhibits weak spasmolytic and analgetic effects. Benzyl mandelate is present in certain combination preparations, and is used mainly in Germany.

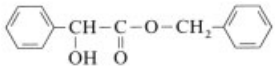

Trade Names. Doppel-Spalt N (Much, FRG), Govil (Stada, FRG), Tispol (Woelm, FRG).

Caroverine [*23465-76-1*], 1-[2-(diethylamino) ethyl]-3-(4-methoxybenzyl)-2(1*H*)-quinoxalinone, Spadon, $C_{22}H_{27}N_3O_2$, M_r 365.46, *mp* 69 °C, *bp* 202 °C (1.33 Pa); hydrochloride, $C_{22}H_{28}ClN_3O_2$, M_r 401.90, decomposition at 188 °C. For synthesis, see [51].

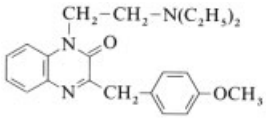

Trade Name. Spasmium (Donau-Pharmazie, Austria).

Flopropione [*2295-58-1*], 2′,4′,6′-trihydroxypropiophenone, $C_9H_{10}O_4$, M_r 182.17, *mp* 175 – 176 °C (anhydrous). Made by reacting benzene-1,3,5-triol with propionitrile in the presence of HCl and $ZnCl_2$, followed by hydrolysis of the resulting ketimine hydrochloride [52]. It is soluble in ethanol and ether and slightly soluble in cold water. Introduced in 1963, flopropione is used primarily in Japan in the treatment of gastrointestinal spasms and as a therapeutic agent for the liver and gallblader.

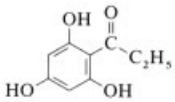

Trade Names. Japan: Bilup (Hishiyama), Cospanon (Eisai), Padeskin (Ohta), Pellegal (Mitsui), Spasmoril (Marupi), Spasmorin (Tokyo Tanabe), Supanate (Nippon Shinyadu); Labrodax (Rhodia, Argentina, Belgium).

Hymecromone [*90-33-5*], 7-hydroxy-4-methylcoumarin, $C_{10}H_8O_3$, M_r 176.16, *mp* 194 – 195 °C; sodium salt, $C_{10}H_7NaO_3$, M_r 198.14. It is made by a Pechmann reaction of resorcinol with ethyl acetoacetate [53], or as described in [54], [55]. Hymecromone is used as a choleretic agent and in the treatment of spasms of the bile duct.

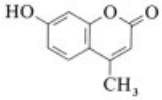

Trade Names. Bilicanta (Boehringer Mannheim, Spain), Cantabiline (Lipha/Sophar, Belgium; Aron-Medicia, France), Cholonerton (Dolorgiet, FRG), Cholspasmin (Lipha, FRG), Gallo-Merz Spasmo Hymecromon (Merz, FRG), Himecol (Kissei, Japan).

Mebeverine [*3625-06-7*], (±)-5-ethyl-6-methyl-7-(4-methoxyphenyl)-5-azaheptyl 3,4-dimethoxy-benzoate, 4-[ethyl(*p*-methoxy-α-methylphenethyl)-amino]butyl veratrate, $C_{25}H_{35}NO_5$, M_r 429.54; hydrochloride [*2753-45-9*], $C_{25}H_{36}ClNO_5$, M_r 466.02, *mp* 105 – 107 °C (from 2-butanone). For the preparation of mebeverine, *N*-ethyl-*p*-methoxy-α-methylphenethylamine is reacted with succinyl chloride, the intermediate is reduced with lithium aluminum hydride, and the resulting alcohol is esterified with veratryl chloride [56]. Used therapeutically since 1961, mebeverine is employed in the treatment of colon irritabile, spastic colitis, etc.

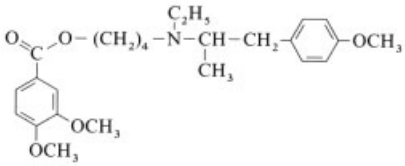

Trade Names. Colofac (Cilag, Australia; Duphar, South Africa, UK; Kali-Pharma, Austria), Duspatal (Duphar: FRG, the Netherlands, Portugal), Duspatalin (Duphar: Belgium, France, Greece, Spain; Ferrosan, Denmark; Kali-Duphar, Switzerland).

Tiropramide [*55837-29-1*], (±)-2-benzamido-3-[4-(2-diethylaminoethoxy)phenyl]-*N,N*-dipropyl-propionamide, $C_{28}H_{41}N_3O_3$, M_r 467.66, *mp* 65 – 67 °C; hydrochloride, $C_{28}H_{42}ClN_3O_3$, M_r 504.12. For synthesis, see [57].

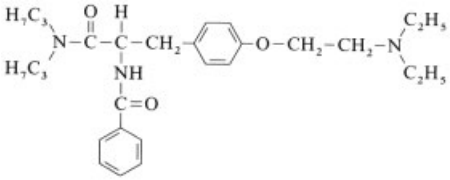

Trade Names. Alfospas (Opfermann, FRG; Rottapharm, Italy), Maiorad (Delta, Portugal; Merck Sharp & Dohme, Italy).

Trimebutine [*39133-31-8*], (±)-2-dimethylamino-2-phenylbutyl 3,4,5-trimethoxybenzoate, $C_{22}H_{29}NO_5$, M_r 387.48, *mp* 78 – 80 °C; hydrochloride, $C_{22}H_{30}ClNO_5$, M_r 423.95, *mp* 184 °C; hydrogenmaleate [*34140-59-5*], $C_{26}H_{33}NO_9$, M_r 503.50, *mp* 105 – 106 °C. The water-soluble hydrochloride is formed by heating 2-phenyl-2-dimethylaminobutanol with 3,4,5-trimethoxybenzoyl chloride in benzene [58].

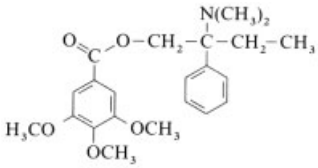

Trade Names. As the maleate: Debridat (Armstrong, Argentina; Lab. Jouveinal, France; Sigmatau, Italy), Polibutin (Juste, Spain).

3.2. Neurotropic – Musculotropic Spasmolytic Agents

Bietamiverine [*479-81-2*], 2-diethylaminoethyl (2-phenyl-2-piperidine)acetate, $C_{19}H_{30}N_2O_2$, M_r 318.45, *bp* 65 °C (133.3 Pa); dihydrochloride [*2691-46-5*], $C_{19}H_{32}Cl_2N_2O_2$, M_r 391.38, *mp* 194 – 195 °C. For syntheses, see [59], [60].

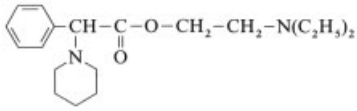

Camylofin [*54-30-8*], 3-methylbutyl 2-[(*N*-2-diethylaminoethyl)amino]-2-phenylacetate, $C_{19}H_{32}N_2O_2$, M_r 320.46, *bp* 174 – 178 °C (199.95 Pa); dihydrochloride [*5892-41-1*], $C_{19}H_{34}Cl_2N_2O_2$, M_r 393.4, *mp* 174 – 178 °C; camylofin bis(noramidopyrine mesylate), $C_{45}H_{66}N_8O_{10}S_2$, M_r 943.31. For synthesis, see [61], [62].

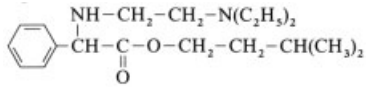

Trade Names. Avacan (Bosnalijek, Yugoslavia; Etem, Turkey); present in Spasmo-Urolong (Thiemann, FRG).

Dicyclomine [*77-19-0*], 2-diethylaminoethyl 1-cyclohexyl-1-cyclohexanecarboxylate, dicycloverine, $C_{19}H_{35}NO_2$, M_r 309.5; hydrochloride [*67-92-5*], $C_{19}H_{36}ClNO_2$, M_r 345.94, *mp* 164 – 166 °C. Obtained by conversion of bicyclohexyl-1-carbonitrile to the ethyl ester, followed by transesterification with 2-diethylaminoethanol [63], or as described in [64], [65]. The hydrochloride is a white, almost odorless, crystalline substance, soluble to the extent of ca. 25 % in water. Dicyclomine exerts a papaverine-like and antimuscarinic effect, but is clearly weaker than atropine. This substance was introduced as a therapeutic agent in 1947, and is used in the treatment of spasms in the region of the gastrointestinal tract.

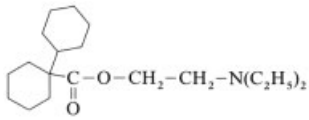

Trade Names. Atumin (Merrel, FRG), Bentyl (Lepetit, Italy; Lakeside, USA), Bentylol (Merrel Dow, Canada, Spain), Icramin (Toho Iyaku, Japan), Merbentyl (Merrel Dow, Australia, UK).

Dihexyverine [*561-77-3*], 2-piperidinoethyl 1-cyclohexyl-1-cyclohexanecarboxylate, $C_{20}H_{35}NO_2$, M_r 321.49; hydrochloride [*5588-25-0*], $C_{20}H_{36}ClNO_2$, M_r 357.96, *mp* 200 °C. For synthesis, see [66], [67].

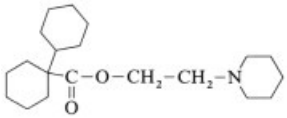

Trade Name. Spasmodex (Crinex, France).

Diisopromine [*5966-41-6*], *N*,*N*-diisopropyl-3,3-diphenylpropylamine, $C_{21}H_{29}N$, M_r 295.45; hydrochloride [*24358-65-4*], $C_{21}H_{30}ClN$, M_r 331.91, *mp* 175 – 176 °C. For synthesis, see [68], [69]. Diisopromine is often used with sorbitol in the treatment of spasms in the region of the bile ducts, and to increase choleresis.

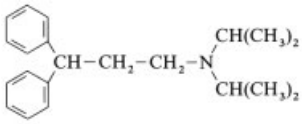

Trade Names. Megabyl (Janssen, France; Pharmacal, Switzerland), Polagol (Polfa, Poland).

Drofenine [*1679-76-1*], 2-diethylaminoethyl 2-cyclohexyl-2-phenylacetate, $C_{20}H_{31}NO_2$, M_r 317.48, *bp* 158 °C (19.99 Pa); hydrochloride [*548-66-3*], $C_{20}H_{32}ClNO_2$, M_r 353.94, *mp* 145 –147 °C. The cyclohexylphenylacetic acid present in drofenine as the diethylaminoethyl ester can be made from benzyl cyanide and cyclohexyl bromide or by hydrogenation of diphenylacetic acid or benzilic acid [70]. Drofenine hydrochloride is readily soluble in water, and very slightly soluble in ethanol and ether. Introduced as a therapeutic agent in 1938, drofenine is the spasmolytic component in several combination preparations.

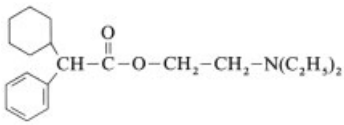

Trade Names. Spasmo-Cibalgin S, Spasmo-Cibalgin-compositum S (Ciba Geigy, FRG).

Flavoxate [*15301-69-6*], 2-piperidinoethyl 3-methyl-4-oxo-2-phenyl-4*H*-1-benzopyran-8-carboxylate, $C_{24}H_{25}NO_4$, M_r 391.45; hydrochloride [*3717-88-2*], $C_{24}H_{26}ClNO_4$, M_r 427.91, *mp* 232 – 234 °C. For synthesis, see [71], [72]. Preparation via transesterification of flavonecarboxylate esters using tetraisopropyl titanate or bis(acetylacetonato)zinc complexes as catalysts is described in [73]. Flavoxate is used for the symptomatic relief of certain disorders of the lower urinary tract and for the relief of vesico-urethral spasms resulting from the use of instrumentation (e.g., catheters) or surgery.

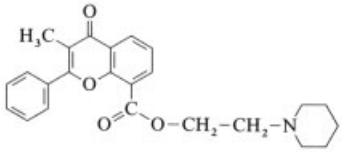

Trade Names. Bradalone (Japan), Genurin (Recordati, Italy), Spasuret (Farmitalia, FRG), Urispadol (Pharmacia, Denmark, Norway); Urispas (Byk Gulden, Austria; Cedona/Byk, Belgium; Cilag, Portugal; Galenica, Switzerland; Smith Kline & French, USA; Syntex, UK).

Oxybutynin [*5633-20-5*], (±)-4-diethylamino-2-butinyl α-cyclohexyl-α-phenylglycolate, $C_{22}H_{31}NO_3$, M_r 357.49; hydrochloride [*1508-65-2*], $C_{22}H_{32}ClNO_3$, M_r 393.97, *mp* 129 – 130 °C. For syntheses, see [74], [75]. (±)-Oxybutynin is used in managing the urgency and incontinence associated with neurogenic bladder. Anticholinergic activity resides mainly in the (*R*)-enantiomer; oxybutynin displays additional nonstereoselective effects for Ca^{2+}-channel antagonism as well as spasmolytic and local anesthetic properties [46].

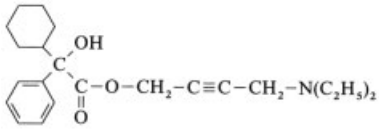

Trade Names. Cystrin (Leiras, Finland), Ditropan (Debat, France; Marion, USA; Scharper, Italy; Therapia, Switzerland), Dridase (Byk Gulden, the Netherlands; Pharmacia, FRG).

Pitofenone [*54063-52-4*], methyl 2-[4-(2-piperidinoethoxy)benzoyl]benzoate, $C_{22}H_{25}NO_4$, M_r 367.43, *mp* 168 – 169 °C; hydrochloride $C_{22}H_{26}ClNO_4$, M_r 403.91. The synthesis is described in [76].

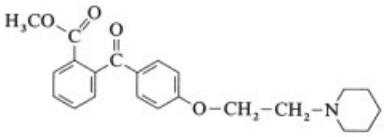

Trade Name. Contained in Litalgin (Leiras, Finland).

Propiverine [*60569-19-9*], 1-methyl-4-piperidyl 2,2-diphenyl-2-propoxyacetate, $C_{23}H_{29}NO_3$, M_r 367.49; hydrochloride $C_{23}H_{30}ClNO_3$, M_r 403.99. For synthesis, see [77]. Propiverine is used mainly in the treatment of micturition problems associated with neurogenic bladder. The possible metabolite *O*-dealkylpropiverine has more than one-hundred times the affinity of propiverine for muscarinic receptors, and is more potent than atropine [78]. For further details of the pharmacology, see [79].

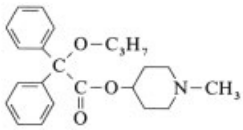

Trade Names. Mictonetten, Mictonorm (Apogepha Arzneim., FRG).

Terolidine [*15793-40-5*], (±)-*N*-*tert*-butyl-4,4-diphenyl-2-butylamine, $C_{20}H_{27}N$, M_r 281.44, *bp* 130 – 132 °C (133.3 Pa); hydrochloride [*7082-21-5*], $C_{20}H_{28}ClN$, M_r 317.90, *mp* 178 – 180 °C. Prepared by the reductive condensation of 4,4-diphenyl-2-butanone with *tert*-butylamine in the presence of $AlCl_3$ [80]; for an alternative synthesis, see [81]. (±)-Terolidine is used for relief of urinary frequency, urgency, and incontinence in patients with motor-urge incontinence and neurogenic bladder. The racemate displays both anticholinergic and calcium-channel blocking properties [45].

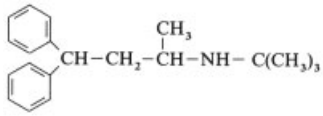

Trade Names. Mictrol (Globopharm, Switzerland; KabiVitrum: Austria, Denmark, the Netherlands, Sweden), Micturin (KabiVitrum, UK).

Thiphenamil [*82-99-5*], *S*-2-diethylaminoethyl diphenylthioacetate, tifenamil, thiphenum (USSR pharmacopoeia), $C_{20}H_{25}NOS$, M_r 327.50; hydrochloride [*548-68-5*], $C_{20}H_{26}ClNOS$, M_r 363.96, *mp* 129 – 130 °C. For synthesis, see [82]. Thioesters like thiphenamil and its *N*,*N*-dimethyl or *N*,*N*-dipropyl derivatives [83], [84] are pharmacologically similar to the corresponding oxygen compounds, and are used primarily in Eastern Europe.

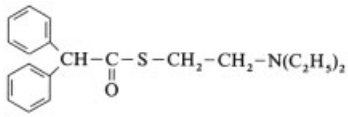

Trade Name. Trocinate (Poythress, USA).

4. Miscellaneous Agents

In certain special cases, substances may be employed which are ordinarily indicated for other conditions (e.g., hypertension, coronary insufficiency, or bronchial asthma). For example, α- or α_1-adrenoceptor antagonists, such as phentolamine [*50-60-2*], phenoxybenzamine [*59-96-1*], or prazosin [*19216-56-9*], are effective in ureteral-stone colic, ureteral spasms, and conditions involving increased α-adrenergic tone of the sphincter of the urinary bladder. The β_2-adrenoceptor agonists used in asthma therapy (e.g., fenoterol [*13392-18-2*]) are employed to relax the uterus in premature labor (tocolytic agents). Organic nitrates, such as glycerol trinitrate [*55-63-0*], useful in coronary heart disease, cause relaxation not only of the vascular smooth muscle, but also of the bronchial smooth muscle and smooth muscle of the gallbladder and bile ducts. There-

fore, glycerol trinitrate is occasionally administered by infusion in the treatment of gall-stone colic and lead colic. Indomethacin [*53-86-1*] and other inhibitors of prostaglandin biosynthesis (nonsteroidal antiphlogistics, → Anti-inflammatory – Antirheumatic Drugs, Chap. 3.) are used in the treatment of spasmodic dysmenorrhea. For additional information on the mechanisms of drug action see → Pharmaceuticals, General Survey and Development.

5. Future Prospects

The discovery of the heterogeneity of the muscarinic receptors and the pharmacological, biochemical, and molecular-biological characterization and localization of five receptor subtypes has opened new perspectives for specific therapeutic influence on muscarinic receptor-mediated functions. Spasms of smooth muscle may be subject to more effective treatment, and with fewer side effects, with highly selective M_3 antagonists. The first prototype substance is represented by *p*-fluorohexahydrosiladifenidol [85–87], a (relatively) M_3-selective antagonist used in experimental pharmacology. A new class of potentially selective M_3 antagonists has been described [88]. Compounds of the newly developed group of potassium- channel openers [89], [90], such as cromakalim [*94470-67-4*] and pinacidil [*60560-33-0*], also exert relaxing effects on smooth muscles. Certain substances in this group are already in clinical trials as antihypertonic and antiasthmatic agents. Nonpeptidic tachykinin-receptor antagonists [91] represent another potential source of new spasmolytic agents.

6. References

1. J. B. Furness et al., *Trends NeuroSci. (Pers. Ed.)* **15** (1992) 66 – 71.
2. T. Hökfelt et al., *Prog. Brain Res.* **68** (1986) 33 – 70.
3. G. Burnstock, *Arch. Int. Pharmacodyn. Ther.* **304** (1990) 7 – 33.
4. I. von Kügelgen, K. Starke, *Trends Pharmacol. Sci.* **12** (1991) 319 – 324.
5. F. Dörje et al., *J. Pharmacol. Exp. Ther.* **256** (1991) 727 – 733.
6. M. P. Caulfield, *J. Pharmacol. Exp. Ther.* (1993) in press.
7. E. C. Hulme, N. J. M. Birdsall, N. J. Buckley, *Annu. Rev. Pharmacol. Toxicol.* **30** (1990) 633 – 673.
8. R. R. Levine, N. J. M. Birdsall (eds.): "Subtypes of Muscarinic Receptors V", *Life Sci.* **52** (1993) 405 – 597.
9. R. R. Levine, N. J. M. Birdsall (eds.): "Subtypes of Muscarinic Receptors IV", *Trends Pharmacol. Sci.* **10** (Suppl.) (1989) 1 – 119.
10. J. Lameh et al., *Pharm. Res.* **7** (1990) 1213 – 1221.
11. T. P. Kenakin, *Drugs* **40** (1990) 666 – 687.
12. F. Dörje, A. I. Levey, M. R. Brann, *Mol. Pharmacol.* **40** (1991) 459 – 462.
13. M. P. Walsh, *Biochem. Cell Biol.* **69** (1991) 771 – 800.
14. G. Schwenker et al., *Chem. Ber.* **99** (1966) 2407 –2408.
15. Eisai, DE 1 950 378, 1970 (S. Tanaka, K. Hashimoto). US 3 696 110, 1972 (S. Tanaka, K. Hashimoto). JP 22 175, 1972 (S. Tanaka, K. Hashimoto).
16. C. H. Boehringer, DE 856 890, 1950 (F. Adickes). Sandoz, GB 764 227, 1956; GB 765 821, 1957.
17. Merck, DE 145 996, 1902.
18. Endo Labs., US 2 962 499, 1960.
19. Pfleger, US 3 480 626, 1967; DE 1 194 422, 1963.
20. L. H. Sternbach, S. Kaiser, *J. Am. Chem. Soc.* **75** (1953) 6068. Hoffmann-La Roche, GB 694 067, 1953;US 2 648 667, 1955 (L. H. Sternbach).
21. N. Sperber et al., *J. Am. Chem. Soc.* **73** (1951) 5010.
22. Schering Corp., GB 740 689, 1955; Tanabe, JP 1954-7585, 1956.
23. Aktiebolaget Recip., SE 136 606, 1952 (S. Carlsson) *Chem. Abstr.* **48** (1954) 729 a.
24. E. L. May, E. Mosettig, *J. Org. Chem.* **13** (1948) 459.
25. T. P. Friebe, Ph. D. Thesis, Universität Frankfurt Main 1992.
26. Hoechst, GB 708 859, 1954.
27. B. V. Franco, C. D. Lunsford, *J. Med. Pharm. Chem.* **2** (1960) 523. A. H. Robins Co., US 2 956 062, 1960 (C. D. Lunsford).
28. Abbott, US 2 907 765, 1959 (A. W. Weston).
29. Janssen, DE 1 003 744, 1955.
30. J. H. Biel et al., *J. Am. Chem. Soc.* **77** (1955) 2250. Lakeside, US 2 918 408, 1957 (J. H. Biel).

31. R. A. Robinson, J. W. Cusic, *J. Org. Chem.* **16** (1951) 1921. Searle, US 2 659 732, 1953.
32. Menarini Soc., US 3 536 723, 1970.
33. Toyama, JP 6 585, 1966; JP 16 475, 1970.
34. K. Hoffmann, H. Schellenberg, *Helv. Chim. Acta* **30** (1947) 292. Ciba, CH 259 948, 1949.
35. S. B. Coan et al., *J. Am. Chem. Soc.* **78** (1956) 3701. Schering Corp., GB 788 126, 1957.
36. Sadao Oki, JP 22 462, 1965; JP 17 015, 1965.
37. Cerm, FR 1 306 610, 1962; GB 953 386, 1964.
38. M. Harfenist, E. Magnien, *J. Am. Chem. Soc.* **78** (1956) 1060. Burroughs Wellcome, US 2 698 325, 1954. Cyanamid, US 2 913 494, 1959.
39. B. von Issekutz et al., *Naunyn-Schmiedeberg's Arch. Exp. Pathol. Pharmakol.* **176** (1934) 8.
40. A. Engelhardt, H. Wick, *Arzneim. Forsch./Drug Res.* **10** (1960) 763 – 767.
41. R. E. Weishaar, J. A. Bristol: "Enzymes & Other Molecular Targets" in C. Hansch et al. (eds.): *Comprehensive Medicinal Chemistry,* vol. 2, Pergamon Press, Oxford 1990, pp. 501 – 514.
42. N. Sunagane et al., *Jpn. J. Pharmacol.* **38** (1985) 133 –139.
43. Y. Kimoto, R. Kessler, C. E. Constantinou, *J. Urol. (Baltimore)* **144** (1990) 1497 – 1499.
44. J. A. Jacobson: "Membranes & Receptors" in C. Hansch et al. (eds.): *Comprehensive Medicinal Chemistry,* vol. 3, Pergamon Press, Oxford 1990, pp. 601 –642.
45. L. Noronha-Blob et al., *Eur. J. Pharmacol.* **201** (1991) 135 – 142.
46. L. Noronha-Blob, J. F. Kachur, *J. Pharmacol. Exp. Ther.* **256** (1991) 562 – 567.
47. Chem. Pharm. AG Bad Homburg, DE 574 656, 1930. E. Wolf, Chinoin, FR 719 638, 1931. E. Wolf, US 1 962 224, 1934.
48. Chinoin Gyogyszer es Vegyeszeti Termekek Gyara Rt., HU 150 372, 1963 (Z. Meszaros et al.); *Chem. Abstr.* **60** (1964) 2915 c; HU 14 090, 1977 (I. Turcsan et al.); *Chem. Abstr.* **88** (1978) 136 466 f.
49. Orgamol S. A., FR 1 362 765, 1964; *Chem. Abstr.* **62** (1965) 536 f. Orgamol S. A., NL-A 6 411 317, 1965; *Chem. Abstr.* **63** (1965) 11 522 a.
50. Much, DE 1 005 948, 1954.
51. Donau-Pharmazie, US 3 028 384, 1962 (H. Zellner et al.); *Chem. Abstr.* **57** (1962) 841 b; DE-OS 1 804 328, 1969 (H. Zellner); *Chem. Abstr.* **71** (1969) 70 642 a.
52. Atlantic Res. Inst., DE 2 519 990, 1975.
53. H. von Pechmann, *Chem. Ber.* **16** (1883) 2119. A. Russel, J. R. Frye, *Org. Synth.* **21** (1941) 23. L. L. Woods, J. Sapp, *J. Org. Chem.* **27** (1962) 3703 –3704.
54. M. Mazzei et al., *J. Heterocycl. Chem.* **18** (1981) 863 –868.
55. R. T. Taylor, R. A. Cassel, *Synthesis* **8** (1982) 672 – 673.
56. Philips-Gloeilampen, GB 1 009 082, 1962; BE 609490, 1962; DE 1 126 889, 1962 (T. Kralt et al.).
57. Rotta Res. Lab., DE-OS 2 503 992, 1975 (F. Makovec et al.); *Chem. Abstr.* **83** (1975) 193 728 p; US 4 004 008, 1977.
58. Lab. Jouveinal, FR 1 344 455, 1962.
59. Nordmark, DE 859 892, 1950; Beiersdorf, DE 1 014 120, 1957 (H. Ruhkopf); *Chem. Abstr.* (1960) 430 d.
60. R. B. Moffet, C. A. Hart, *J. Am. Pharm. Assoc. Sci. Ed.* **42** (1953) 717.
61. Asta, DE 842 206, 1950.
62. B. K. Edwards, A. A. Goldberg, A. H. Wragg, *J. Pharm. Pharmacol.* **12** (1960) 179 – 186.
63. C. H. Tilford et al., *J. Am. Chem. Soc.* **69** (1947) 2902 – 2905. Merrel, US 2 474 796, 1949.
64. Norwich Pharmacal Co., US 3 527 792, 1970 (F. F. Ebetino); *Chem. Abstr.* **73** (1970) 98 483 q.
65. Kawaken Fine Chemicals Co., JP-Kokai 82 122 051, 1982;JP-Kokai 58 118 546 [83 118 546], 1983;DE-OS 3 213 255, 1983 (H. Kagawa, Y. Kimizuka); *Chem. Abstr.* **97** (1982) 181 824 t; *Chem. Abstr.* **100** (1984) 67 894 w and 102 813 t.
66. M. Kopp, B. Tchoubar, *Bull. Soc. Chim. Fr.* 1952, 84 – 85.
67. M. Verrier, *Bull. Soc. Chim. Fr.* 1959, 1863 – 1865.
68. N. V. Nederl. Comb. voor Chem. Ind. and Janssen, N. V., NL 87 517, 1958; GB 808 158, 1959.
69. Zaklady Farmaceutyczne Motor, PL 44 893, 1961 (W. Daniewski, A. Daniewski); *Chem. Abstr.* **58** (1963) 3353 a.
70. Ciba, CH 215 775, 1938; CH 219 301, 1938; US 2 265 184-5, 1941.
71. P. Da Re et al., *J. Med. Pharm. Chem.* **2** (1960) 263.
72. Recordati S. A., US 2 921 070, 1960 (P. Da Re); GB 1 343 119, 1974 (P. Da Re); DE-OS 2 059 296, 1972 (E. Sianesi); *Chem. Abstr.* **76** (1972) 140 532 u. Seceph S. A., US 3 350 411, 1967 (P. Da Re); *Chem. Abstr.* **68** (1968) 87 159 j.

73. Mitsui Petrochem. Ind. Ltd., JP-Kokai 61 109 786 [86 109 786] and 61 130 283 [86 130 283], 1986 (N. Hirowatari); *Chem. Abstr.* **106** (1987) 4747 d and 18 241 w.
74. Mead Johnson & Co., GB 940 540, 1963; *Chem. Abstr.* **60** (1964) 2909 d.
75. Kodama Ltd., JP-Kokai 02 03 635 [90 03 635], 1990 (T. Kasama, R. Kadota); *Chem. Abstr.* **113** (1990) 23 118 k.
76. Hoechst, US 2 681 340, 1954.
77. J. Klosa, G. Delmar, *J. Prakt. Chem.* **16** (1962) 71 – 84. Ch. Starke, G. Thomas, J. Friese, DD 106 643, 1974; *Chem. Abstr.* **82** (1975) 155 841 s.
78. J. Riotte, E. Mutschler, *Arzneim. Forsch./Drug Res.* **37** (1987) 300 – 302.
79. A. Haruno, *Arneim. Forsch./Drug Res.* **42** (1992) 1459 – 1461.
80. Richter, Gedeon, Vegyeszety Gyar Rt., HU 32 331, 1984 (K. Harsanyi, I. Szabadkai); *Chem. Abstr.* **102** (1985) 5688 q.
81. Aktiebolaget Recip., GB 923 942, 1963; *Chem. Abstr.* **59** (1963) 9881 d.
82. W. P. Poythress, US 2 390 555, 1943. Sterling Drug, US 2 510 773, 1946 (R. O. Clinton). R. O. Clinton, U. J. Salvador, *J. Am. Chem. Soc.* **68** (1946) 2076.
83. S. S. Liberman, *Farmakol. Toksikol (Moscow)* **19** (1956) 10; *Chem. Abstr.* **51** (1957) 7587 d.
84. S. S. Liberman, *Med. Promst. SSSR* **11** (1957) 43. N. A. Kopylova et al., SU 109 348, 1957; *Chem. Abstr.* **52** (1958) 11 292 f and 10 580 c.
85. G. Lambrecht et al., *Eur. J. Pharmacol.* **152** (1988) 193 – 194.
86. G. Lambrecht et al. in [9] pp. 60 – 64.
87. R. Tacke et al., *J. Organomet. Chem.* **417** (1991) 339.
88. C. Kaiser et al., *J. Med. Chem.* **36** (1993) 610 – 616.
89. A. H. Weston, G. Edwards, *Biochem. Pharmacol.* **43** (1992) 47 – 54.
90. S. Duty, A. H. Weston, *Drugs* **40** (1990) 785 – 791.
91. K. J. Watling, *Trends Pharmacol. Sci.* **13** (1992) 266 – 269.

Specialty Plastics

HANS-GEORG ELIAS, Michigan Molecular Institute, Midland, MI 48640, United States

1. Introduction . . . 1
2. Polyphenylenes . . . 1
2.1. Chemical Structure . . . 1
2.2. Polybenzenes . . . 2
2.3. Polymers from Higher Aromatic Hydrocarbons . . . 2
2.4. Acetylene Group Containing Polymers . . . 2
3. Poly(*p*-Xylylenes) . . . 3
3.1. Introduction . . . 3
3.2. Synthesis . . . 4
3.2.1. Monomer Synthesis . . . 4
3.2.2. Polymerization . . . 5
3.3. Structure . . . 6
3.4. Manufacture . . . 7
3.5. Properties . . . 7
4. Polybenzocyclobutenes . . . 9
5. Ionomers and Related Polymers . . . 9
5.1. Overview . . . 9
5.1.1. Introduction . . . 9
5.1.2. Polyelectrolytes . . . 10
5.1.3. States of Ions . . . 11
5.2. Acidic Copolymers . . . 12
5.2.1. Introduction . . . 12
5.2.2. Structure and Properties . . . 12
5.2.3. Thermoplastic Acidic Copolymers . . . 13
5.2.4. Elastomeric Acidic Copolymers . . . 14
5.3. Ionomers . . . 14
5.3.1. Introduction . . . 14
5.3.2. Structure . . . 15
5.3.3. Properties . . . 16
5.3.4. Ethylene – Methacrylic Acid Ionomers 18
5.3.5. Fluorinated Ionomers . . . 20
5.3.6. Sulfochlorinated Polyethylene . . . 20
5.3.7. Sulfonated EPDM Polymers . . . 21
6. References . . . 21

1. Introduction

In this article, "specialty plastics" are defined as those plastics which do not fit into other keywords, either because they are prepared from unusual monomers or because they possess unique structures and/or properties. However, the term "specialty plastics" is in general often used as a synonym for "high-performance plastics", which are a special class of engineering plastics (technical plastics, technoplastics; see → Plastics, General Survey, Chap. 1.2.1.). Engineering plastics provide the engineer with a combination of useful properties, whereas high-performance plastics often exhibit only one outstanding property albeit one which surpasses the corresponding property of engineering plastics, e.g., a high continuous service temperature.

"Specialty plastics" are also not necessarily identical with either "functional plastics" or "functional polymers." Functional plastics are those plastics which are useful for one application only; examples are ethylene – vinyl alcohol copolymers with high contents of vinyl alcohol units, which are exclusively used as gas and aroma barriers in the packaging industry. The term "functional polymer" or "functionalized polymer", on the other hand, refers to polymers with functional chemical groups, i.e., specific chemical reactivities.

The polyphenylenes and poly(*p*-xylylenes) described in this article are typical functional plastics since they are used for only one purpose: coatings. The group of polymers known as ionomers, on the other hand, comprises functional plastics, engineering plastics, elastomers, and thermoplastic elastomers.

2. Polyphenylenes

2.1. Chemical Structure

Polyphenylenes are oligomeric to polymeric compounds composed of phenylene units and related units such as naphthalene or anthracene; these units are connected to each other via *ortho*, *meta*, and/or *para* links. Their molecular architecture thus ranges from linear to

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_529

highly branched compounds; they may or may not become cross-linked on further processing. Polyphenylenes are also known as polyphenyls, oligophenyls, oligophenylenes, or polybenzenes [1–7].

Polyphenylenes exhibit two remarkable properties. As compounds of low hydrogen content and free of readily activated chemical bonds, they show good thermal, oxidative, and hydrolytical stability. As aromatic compounds, they have the potential of becoming electrically conducting materials. Commercial developments are thus along two lines: polyphenylenes are either prepared as thermosetting resins with high service temperatures or as base materials for electrically conducting polymers.

The two classes of polyphenylenes differ in their chemical structures. Polyphenylenes for conductors are linear polymers with phenylene units interconnected in 1,4-position (PPP [*25190-62-9*]), whereas polyphenylenes for thermosets are highly branched oligomers [*9033-83-4*], often with reactive end groups such as acetylene (schematically shown as PXP). This section is exclusively concerned with the latter type of polymers; poly(1,4-phenylenes) are discussed in → Polymers, Electrically Conducting, Chap. 5.5.

PPP

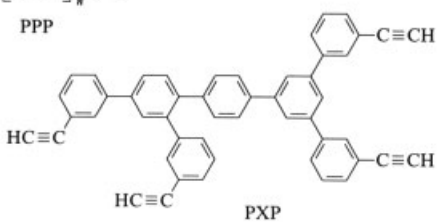

PXP

2.2. Polybenzenes

Melting temperatures of simple poly(1,4-phenylenes) increase strongly from biphenyl $H(C_6H_4)_2H$ (70 °C) to hepta(1,4-phenylene) $H(p\text{-}C_6H_4)_7H$ (ca. 540 °C). Poly(1,4-phenylenes) were thus expected to be very temperature stable compounds. All classic stepwise syntheses of such polyphenylenes are, however, too expensive for the industrial manufacture of plastics but may be utilized for electrically conducting polymers [1], [2], [6], [7].

The first industrial branched polyphenylenes were prepared by oxidative cationic polymerization of benzene with stoichiometric amounts of aluminum trichloride/copper dichloride, iron trichloride, molybdenum pentachloride, etc. [3–5], [8–10]. The arylation reaction proceeds via radical cations [7], [11]. The polymers still contain 1 – 4 % chlorine, < 5 % oxygen, 0.1 – 0.2 % copper, and 0.3 – 0.6 % aluminum after purification [5].

The resulting brown to black masses have densities of ca. 1.25 g/cm^3, tensile strengths of 65 MPa, Young's moduli of 160 MPa, elongations at break of 10 %, and linear expansion coefficients of ca. 3×10^{-5} K^{-1}. They melt above 530 °C and decompose above 700 °C. Their melt viscosities are so high that they cannot be processed by injection molding or similar fabrication techniques. The polymers can, however, be molded at high pressure; the moldings can then be machined. Such materials are produced as Eimac 221 by Eitel – McCollough, United States.

2.3. Polymers from Higher Aromatic Hydrocarbons

The oxidative cationic polymerization of terphenyl isomers, with or without addition of other compounds such as biphenyl, reduces the probability of cross-linking reactions as compared to benzene. The resulting branched prepolymers are soluble; their solutions in chloroform or chlorobenzene are used as impregnating varnish for laminates. Filled prepolymers can be processed in the melt. The prepolymers are cross-linked to give insoluble and unmeltable polymers by the addition of catalysts such as *p*-toluenesulfonic acid, sulfuryl chloride, or boron trifluoride etherate. These techniques are mainly being developed by Hughes Aircraft Co.

2.4. Acetylene Group Containing Polymers

Other types of branched polyphenylenes are the so-called H-resins of Hercules [12–15]. These

prepolymers are obtained by cyclotrimerization of *p*-diethynylbenzene (DEB) in aromatic solvents [16] or by copolycyclotrimerization of DEB with phenylacetylene [17]; catalysts are $AlR_3/TiCl_4$, $NiCl_2/P(C_6H_5)_3/NaBH_4$, or $CoBr_2/(C_6H_5O)_3P$.

The prepolymers (B-stage resins) are oligophenylenes with the schematic structure PXP (see Section 2.1). For B-stage resins from the polymerization of DEB by $TiCl_4/(C_2H_5)_2AlCl$ at 86 % monomer conversion, a number-average molar mass of 2300 g/mol and a mass-average molar mass of 8060 g/mol were found. Higher DEB conversions resulted in gelation.

These polymers contain acetylene end groups that form cross-links under the action of titanium(IV) chloride/diethylaluminum chloride, nickel(I) acetylacetonate, and similar catalysts. The chemical structure of these cross-links is unknown; they are probably not benzene rings.

B-Stage H-resins dissolve in aromatic hydrocarbons, chlorinated hydrocarbons, ketones, and cyclic ethers. They are insoluble in water, aliphatic hydrocarbons, and aliphatic alcohols. The resins can be processed with simultaneous cross-linking by molding, injection molding, extrusion, and powder coating; different grades are provided for these purposes. Molding cycles are 2 – 5 min at 160 – 170 °C [13]. Molding stability starts at 160 °C but improved mechanical stabilities and better solvent resistance require postcuring at 230 – 300 °C.

The cured resins are hard, nonporous materials which are resistant to hot solvents, acids, bases, brine, superheated steam, and salt melts. They become brittle in molten alkali metals. Continuous service temperatures are 200 – 300 °C in air and ca. 400 °C in an oxygen-free atmosphere; the limiting oxygen index is 55 %. Slow carbonization starts at ca. 500 °C.

Molded parts show flexural moduli of several gigapascal and high flexural strengths (Table 1). These properties can be enhanced by fillers such as talc, silica, or asbestos. They are often preserved over a wide temperature range.

Unmodified H-resins exhibit low elongation and low adhesion to polar substrates. The adhesion of H-resins to metal surfaces can be improved by the use of polymer blends and polymeric fillers such as epoxy resins, phenolic resins, polysulfones, and polyimides.

H-Resins serve as corrosion-resistant coatings, for example, for pipes and processing equipment for hot sour petroleum or gas wells, pump linings, high-temperature sodium and lithium batteries, and high-temperature processing equipment. They are also used as nonflammable coatings for electric and electronic components and as self-lubricating coatings for metals.

3. Poly(*p*-Xylylenes)

3.1. Introduction

Poly(*p*-xylylenes), PPX, are polymers of the unstable *p*-xylylenes (*p*-quinodimethanes) PX; they are produced industrially from the dimeric *p*-xylylenes DPX:

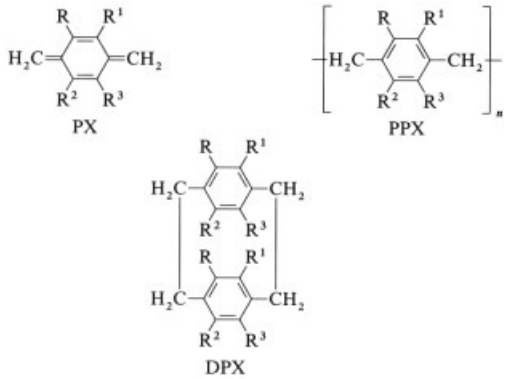

The syntheses, structures, and properties of PPX are described in [18–25]. Although PPXs with many different substituents R, R^1, R^2, and R^3 have been prepared in the laboratory (see the list in [23]), only three of them are industrially important. These polymers are based on DPX as monomer and have no (N) or one chlorine (C) substituent, or are disubstituted (D) with chlorine:

PPX-N ($R = R^1 = R^2 = R^3 = H$) [*25722-33-2*]
PPX-C ($R = R^2 = R^3 = H$; $R^1 = Cl$) [*9052-19-1*]
PPX-D ($R = R^3 = H$; $R^1 = R^2 = Cl$) [*52261-45-7*]
The systematic names are:

Table 1. Properties of cured H-resins

Physical property	Temperature, °C	Property value after filling with		
		none	50 wt % talcum	50 wt % silica
Flexural modulus, GPa	23	4.8	9.7	11.4
	250		6.3	5.9
	300	3.5		
	350		6.8	4.6
Flexural strength, MPa	23	<69	<69	<83
	250		<62	62
	350	<55	<55	46
Barco hardness	23	85		
Taber abrasion (17) (1000 g load)	23	0.0015		
Volume resistivity, Ω cm	23	10^{17}		
Dissipation factor (60 Hz)	23	0.002		
	100	0.0009		

PPX-N	poly(1,4-phenylene-1,2-ethanediyl)
PPX-C	poly[(chloro-1,4-phenylene)-1,2-ethanediyl]
PPX-D	poly[(dichloro-1,4-phenylene)-1,2-ethanediyl]

The designations monochloro and dichloro PPX for technical products are misleading since these polymers do not have one or two chlorine substituents on each ring. Monochloro and dichloro rather refer to the average number of chlorine ligands per ring, i.e., there is a distribution of chlorine substituents amongst the rings; each ring in PPX-C or PPX-D may bear between 0 and 4 chlorine substituents.

Poly(*p*-xylylenes) from the Gorham process (see below) are known as parylenes. Coatings produced by this process are correspondingly named parylene N, parylene C, and parylene D. The sole producer of dimeric *p*-xylylenes DPX-N, DPX-C, and DPX-D is Union Carbide, which also custom-coats parts with PPX-N, PPX-C, and PPX-D.

3.2. Synthesis

3.2.1. Monomer Synthesis

Two synthetic routes have been used commercially for the synthesis of dimeric *p*-xylylenes DPX, also known as [2.2]-paracyclophanes: direct pyrolysis and Hofmann elimination. In direct pyrolysis, *p*-xylene X is pyrolyzed at ca. 950 °C in the presence of steam. The pyrolysis dehydrogenates *p*-xylene X to *p*-xylylene PX-N which is converted into di-*p*-xylylene DPX-N on quenching with liquid xylene at ca. 20 °C:

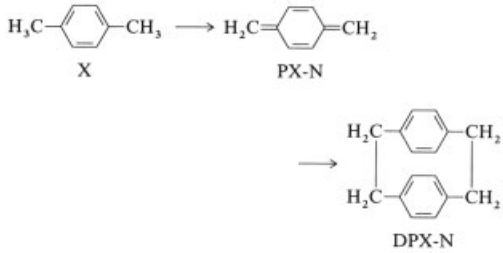

A major byproduct is the polymer PPX-N. Its simultaneous formation can be minimized by using an excess of liquid xylene (Ruggli – Ziegler dilution principle), controlled rates, etc. The polymeric byproduct has no commercial value since it can be neither processed nor recycled.

The Hofmann elimination proceeds under far milder conditions and does not need special equipment, unlike the direct pyrolysis. α-Bromo-*p*-xylene is first quaternized and the resulting product converted into the quaternary ammonium hydroxide; amine elimination then leads to DPX-N:

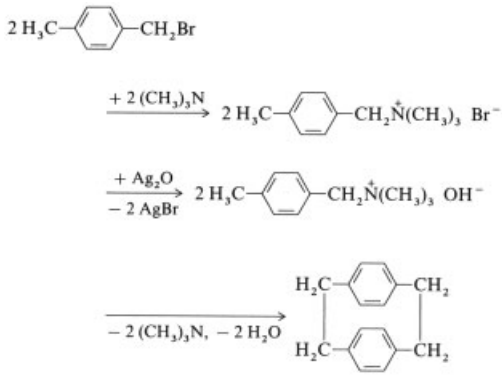

The yields of DPX-N vary between 17 and 50 %, depending on the reaction conditions (e.g., use of polymerization inhibitors).

Chlorine-containing dimeric *p*-xylylenes are produced industrially by chlorination of DPX-N using standard procedures for aromatic ring chlorinations. Chlorination gives mixtures of unreacted DPX-N with various chlorinated compounds (from mono to tetra substitution per ring) and their isomers.

3.2.2. Polymerization

Poly(*p*-xylylenes) can be synthesized by the Szwarc or the Gorham process. In the Szwarc process [26–28], *p*-xylene vapor is passed rapidly (residence time < 1 s) in vacuum (< 100 Pa) through quartz tubes heated to 800 – 1150 °C. The pyrolysis gases are subsequently quenched to temperatures below 25 °C. The mechanism was assumed [26–28] and later confirmed [29–32] to consist of the formation of monoradicals (either spontaneously or by radical transfer), followed by disproportionation to give *p*-xylylene, and finally polycombination of the diradicals to give poly(*p*-xylylene):

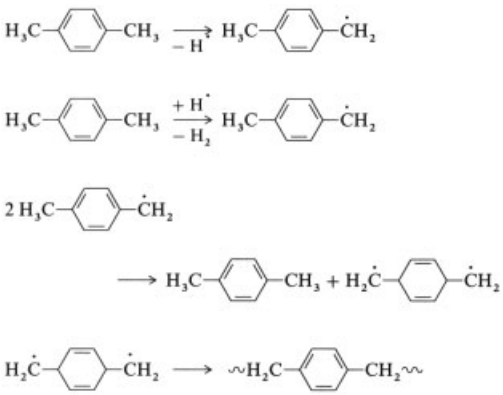

The diradical and not the *p*-xylylene is the intermediate species since the former is stable in the gas phase between 120 °C and 200 °C and in liquid phases (such as quenched hexane solutions) at low temperatures (half-life 21 h at −78 °C). The resulting poly(*p*-xylylene) is slightly cross-linked; it always contains 10 – 25 wt % of low molar mass impurities. Furthermore, since the yield never exceeds 25 %, the process was technically and economically useless.

The Gorham process, on the other hand, leads to colorless, linear (uncross-linked) polymers that are free of low molar mass compounds [33], [34]. The process uses di-*p*-xylylene DPX as monomer and proceeds in three stages: (1) DPX is vaporized at ca. 200 °C and 133 Pa, (2) on heating to 680 °C at ca. 67 Pa, DPX dissociates quantitatively [21] to *p*-xylene which, after cooling to 25 °C and 13 Pa, (3) polymerizes to poly(*p*-xylylene) PPX on the surfaces of parts in a deposition chamber at < -70 °C:

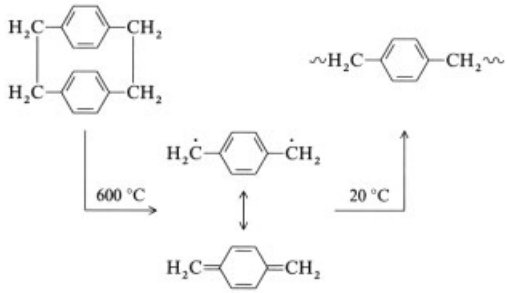

p-Xylylene PX can be represented by quinoid (diamagnetic) or benzoid (paramagnetic) structures. The vapor of PX is diamagnetic [24].

The polymerization is thus probably started by paramagnetic diradical structures; such species may be the dimeric diradical from a one-sided ring-opening of DPX-N, the benzoid structure of PX, or the dimeric diradical from the coupling of two *p*-xylylene molecules. The polymerization seems to be propagated by addition of the diamagnetic quinoid structures to the growing macrodiradicals and not (or only slightly) by the polyrecombination of macrodiradicals. The polymerization proceeds as living polymerization with radical concentrations of $(5-10)\times 10^{-4}$ mol of free radicals per mole of xylene units of the final polymers.

Polymer yields are 100 % at −50 °C and drop rapidly at temperatures above ca. 0 °C (Fig. 1). Molar masses are almost independent of polymerization temperature in the range −50 °C to +50 °C [35].

The Gorham process can be applied to substituted [2.2]-paracyclophanes, provided the threshold condensation temperatures for the condensation on surfaces are not exceeded. These threshold temperatures are 30 °C for poly(*p*-xylylene), 60 °C for poly(2-methyl-*p*-xylylene), 90 °C for poly(2-chloro-*p*-xylylene),

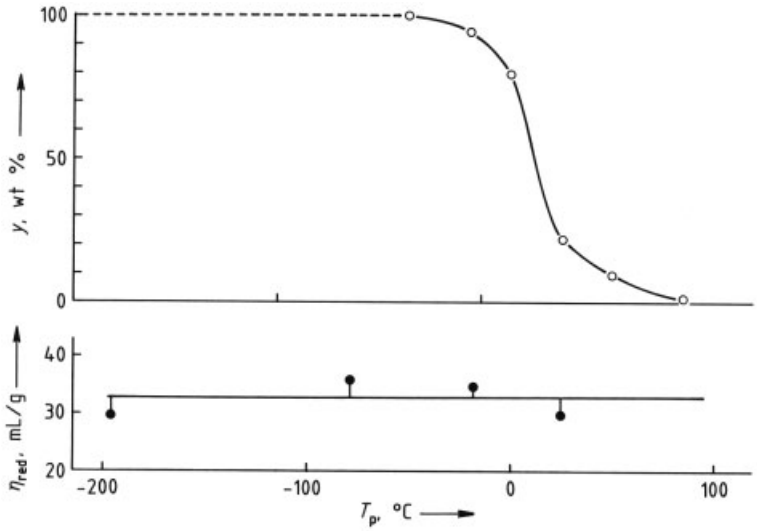

Figure 1. Dependence of polymer yield γ and reduced viscosity η_{red} (1×10^{-3} g/mL polymer in α-chloronaphthalene at 238 °C) of poly(*p*-xylylenes) on the polymerization temperature T_p (data from [35])

and 130 °C for poly(2,5-dichloro-*p*-xylylene) [34]. The physical meaning (other than the phenomenological) of the threshold temperature is unclear since the most obvious interpretation as thermodynamic ceiling temperatures is made difficult by reports of very high estimated thermodynamic ceiling temperatures of more than 200 °C for a gas pressure of 0.08 torr (11 Pa) [35] and calculated ceiling temperatures of 1149 °C (gas) and 3035 °C (liquid) for PX [25]. However, such calculations suffer from the absence of knowledge about the true nature of the polymerization: true equilibria between all species (monomers, monomeric diradicals, macrodiradicals), equilibria between monomers and macrodiradicals only, or even kinetically-controlled reactions.

Poly(*p*-xylylenes) can also be obtained by other methods [25], [36], [37], none of which seem to be used industrially since the monomers are too expensive and/or the polymers are obtained in low yield, with low molar masses, or with many low molar mass impurities.

3.3. Structure

Chemical Structure. Poly(*p*-xylylenes) from the Gorham process are linear polymers, whereas those from the Szwarc process are slightly cross-linked. The number-average molar mass, as calculated from the free-electron concentration of living PPX-N polymers [33], is ca. 200 000 – 400 000 g/mol. Osmometry of soluble fractions in chloroform gave 24 000 g/mol [24].

The commercial chlorinated monomers DPX-C and DPX-D are complex mixtures with 0 – 8 chlorine groups per molecule and 0 – 4 chlorine substituents per ring. On pyrolysis and subsequent vapor deposition polymerization, DPX-C with an average of one chlorine substituent per ring thus gives mainly base units $-CH_2-(p-C_6H_3Cl)-CH_2-$ besides considerable amounts of base units $-CH_2-(p-C_6H_4)-CH_2-$ and $-CH_2-(p-C_6H_2Cl_2)-CH_2-$. Chlorinated units may be arranged head-to-head, head-to-tail, and tail-to-tail so that the resulting constitutional copolymers also show regio-isomerism.

Physical Structure. PPX-N has two crystal modifications [38], [39]. The monoclinic α-modification is formed between −17 °C and +30 °C, while the triclinic β-modification exists below −78 °C and above 220 – 260 °C. The unit cell of the α-modification consists of two monomer units. The phenylene units are parallel to each other; the angle between the plane of the benzene rings and the methylene chain is 90°.

The $\alpha \rightarrow \beta$ transition at 220 °C is accompanied by a marked increase in brittleness. At 270 °C, a reversible smectic transition is observed.

Vapor deposition polymerization is carried out below the melting and near the glass temperatures of the polymers (see Table 2); any crystallization must thus occur simultaneously with the polymerization on surfaces. At polymerization temperatures between 0 °C and 60 °C, the α-modification is formed, whereas mixtures of α and β are found for polymerization temperatures above 80 °C.

The degree of crystallinity of parylene N has been reported to be 57 % [25]. The density of amorphous PPX-N, as calculated from the experimental density of 1.11 g/cm^3 and the theoretical density of the α-modification of 1.185 g/cm^3 (from the unit cell), is thus 1.070 g/cm^3.

3.4. Manufacture

Parylenes are exclusively prepared and applied by vapor deposition polymerization of DPXs on surfaces (Gorham process). Total production is estimated to be several tonnes per year.

Film formation by vapor deposition polymerization differs from other coating techniques in that the coating is formed from the gaseous phase rather than from a liquid one (no surface tension effects on wetting) and that the film grows from the substrate upwards and not downwards as in drying films. The films are self-adherent; coupling agents are usually not necessary.

The technique produces pinhole-free coatings of uniform thickness even on complex, three-dimensional substrates with hidden areas. Films have thicknesses between 0.050 and 100 µm. They are flexible even at low temperatures and possess good dimensional stability.

These features, together with good electrical and mechanical properties and low gas and water vapor permeability, make parylenes especially suitable for surface coating of electrical and electronic components such as printed and integrated circuits, capacitors, transistors, and miniature electrical components. Parylenes are also applied for particle encapsulations, thin membranes, lubricants, and long-term implantation devices.

3.5. Properties

Properties of poly(*p*-xylylenes) are listed in Table 2.

Solubility. Poly(*p*-xylylenes) from the Gorham process are sparingly soluble. PPX-N dissolves to 1.5 % and PPX-C to 10 % in methylene chloride at 25 °C [25]. The solubility of PPX-N in xylene is ca. 10 % at 140 °C. PPX-N and PPX-D can be dissolved in chlorinated biphenyls or in benzyl benzoate at temperatures above 200 °C, and PPX-C in 1-chloronaphthalene at 150 °C [33]. Cooling the solutions to below 150 – 200 °C leads to gel-like precipitation of the polymer. Repeated dissolution and precipitation cause degradation.

Aromatic and halogenated aromatic compounds swell parylenes at room temperature, but stress corrosion has not been observed. Parylenes are resistant to concentrated sulfuric acid and 5N nitric acid.

Thermal Properties. Melting temperatures of symmetrically substituted poly(*p*-xylylenes) PPX-N and PPX-D are higher than those of unsymmetrically substituted PPX-C. The glass temperature of PPX-N is fairly low (T_g = 13 °C) [25], and lower than the earlier reported 80 °C [21]. Similar low glass temperatures can be expected for PPX-C and PPX-D; earlier reported glass temperatures of 80 – 110 °C are definitely too high since no sudden upswing of relative permittivity (dielectric constant) ε_r due to the onset of segmental motions is observed in the temperature interval 20 – 190 °C, only a continuous increase of ε_r from 3.0 to 4.7.

Since the glass temperature is lower than ambient temperature, structural changes occur upon aging and/or annealing of parylenes; freshly deposited polymers have other properties than old ones.

The stability of parylenes is determined by the dibenzyl groups, which are easily oxidized and thermally cleaved. The continuous service temperature of PPX-N is thus only 100 °C in air and 220 °C under an inert atmosphere (see Table 2).

Table 2. Properties of poly(*p*-xylylenes) [25], [34], [40], [41]

Property	Property values of		
	PPX-N	PPX-C	PPX-D
Density at 23 °C, g/cm^3	1.11	1.29	1.42
Refractive index n_D^{23}	1.661	1.639	1.669
Water uptake (24 h, 23 °C), %	0.01	0.06	< 0.1
Melting temperature, °C	420	290	380
Glass temperature, °C	13		
Continuous service temperature (air), °C	100	130	
Continuous service temperature (inert), °C	220		220
Linear expansion coefficient (25 °C), K^{-1}	6.9×10^{-5}	3.5×10^{-5}	
Specific heat capacity (25 °C), J g^{-1} K^{-1}	1.3	1.0	
Thermal conductivity (25 °C), kW m^{-1} K^{-1}	12	8.2	
Flexural modulus (23 °C), MPa	2450	2800	2800
Modulus of elasticity, MPa			
at 23 °C	2400	3200	2800
at 200 °C	170	170	170
Tensile strength (23 °C), MPa	45	70	75
Yield strength (23 °C), MPa	42	55	60
Elongation at break (23 °C), %	< 30	200	< 10
Rockwell hardness R (23 °C)	85	80	
Relative permittivity (23 °C)			
60 Hz	2.65	3.15	2.84
1 kHz	2.65	3.10	2.82
1 MHz	2.65	2.95	2.80
Dissipation factor (23 °C)			
60 Hz	0.0002	0.020	0.004
1 kHz	0.0002	0.019	0.003
1 MHz	0.0006	0.013	0.002
Surface resistivity, Ω (23 °C)			
50 % RH	1×10^{13}	1×10^{14}	5×10^{16}
90 % RH	9×10^{11}	7×10^{11}	
Volume resistivity (50 % RH, 23 °C), Ω cm	1.4×10^{17}	8.8×10^{16}	2×10^{16}
Dielectric strength at 23 °C, kV/mm	260 – 280	145 – 220	200 – 215
Permeability (at 25 °C)			
10^{14} *P*, cm^2 s^{-1} Pa^{-1}			
N_2	0.35	0.020	0.20
O_2	1.76	0.32	1.44
Cl_2	3.32	0.016	0.025
CO_2	9.64	0.35	0.58
He	18.8	7.4	
H_2	24	5.0	
H_2S	35.7	0.58	0.065
SO_2	85.1	0.49	0.21
H_2O	428	150	

Mechanical Properties. Parylenes have tensile strengths and elastic moduli (Table 2) similar to those of other engineering plastics. The unusually high elongation at break of PPX-C refers to freshly deposited films; such films have high dimensional stability if subjected to high relative humidities. Films remain tough and flexible down to −165 °C. Since further crystallization takes place on annealing, density, hardness, and abrasion resistance also increase.

Permeability. Gas permeabilities (O_2, N_2) of parylenes are similar to those of poly(ethylene terephthalate), and water vapor permeabilities compare favorably with those of high-density polyethylene. PPX-C and PPX-D have also low permeability for SO_2 and H_2S.

Electrical Properties. Parylenes from the Gorham process have low relative permittivity (dielectric constant) and dielectric loss (dissipation factor) due to their chemical structures (only carbon and hydrogen, no hydrophilicity, no dipoles) and the absence of low molar mass impurities and catalyst residues. Dielectric strengths are high compared to other polymers, which is in part due to the thinness of the test films (0.025 mm) since dielectric strengths

decrease with increasing thickness and dielectric strengths of other polymers are usually measured at thicknesses of 3.18 mm.

4. Polybenzocyclobutenes

Polybenzocyclobutenes are polymers and copolymers of various benzocyclobutenes [42–46]. On heating, the parent compound, benzocyclobutene (BCB), is converted to the unstable *o*-quinodimethane, OQDM (*o*-xylylene, OX), which polymerizes thermally to polybenzocyclobutene (PBCB):

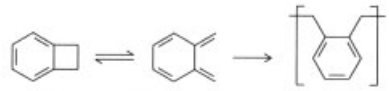

BCB [*694-87-1*] OQDM, OX PBCB (**1**) [*51774-83-5*]

The intermediate is not a biradical but has an *o*-quinone structure (ground-state singlet) according to theoretical calculations. This intermediate is the true monomer; the polymer can thus also be called a poly(*o*-quinodimethane) or a poly(*o*-xylylene).

Syntheses of benzocyclobutene derivatives start with either benzocyclobutene or 3-bromobenzocyclobutene. Monobenzocyclobutenes (functionality 2) such as **2** lead to linear polymers, and bis(benzocyclobutenes) (functionality4) such as **3–7** to cross-linked polymers.

The polymerization proceeds without formation of volatile products. It does not require catalysts or initiators and it is first order in benzocyclobutene groups. The polymerization temperature of BCB is 200 – 250 °C; it is lower for substituted benzocyclobutenes **2–7**, regardless of whether the groups are electron donating or electron withdrawing.

The polymerization mechanism of the bis-benzocyclobutene monomers **3–7** is not known with certainty: intermolecular reactions **1** may either lead to random cross-linking or to ladder-type structures.

All polymers of **2–7** have high thermal stability. The good mechanical properties (Table 3) are retained at temperatures up to 200 – 250 °C.

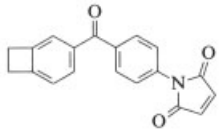

2
Monomer: [*112903-01-2*]; Polymer: [*112903-02-3*]

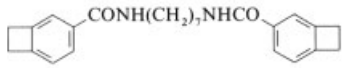

3
Monomer: [*99716-25-3*]; Polymer: [*99716-26-4*]

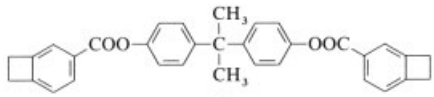

4
Monomer: [*99716-21-9*]; Polymer: [*99716-22-0*]

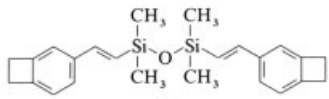

5
Monomer: [*117732-87-3*]; Polymer: [*124221-30-3*]

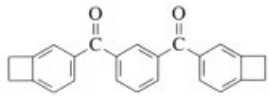

6
Monomer: [*137662-13-6*]; Polymer: [*137662-14-7*]

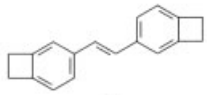

7
Monomer: [*67237-37-0*];

5. Ionomers and Related Polymers

5.1. Overview

5.1.1. Introduction

Terms such as ion-containing polymer, ionic polymer, ionomer, polyelectrolyte, etc., are commonly used in polymer literature. None of these terms are defined by international standardization bodies and their meaning often differs from similar ones in other branches of science.

The term "ion-containing polymer" in its literal sense does not specify whether ions are part of the polymer molecules or whether they are extraneous to the polymer molecules, for example,

Table 3. Physical properties of polymers from a benzocyclobutene 2 and various bisbenzocyclobutenes 3 – 7

Property	Property values of polymer					
	2	**3**	**4**	**5**	**6**	**7**
Young's modulus (25°C), MPa		2500	2660			
Shear modulus (25°C), MPa		940	1000			
Flexural modulus (25°C), MPa	3550 ∗			3310	4250	5150
Flexural strength (25°C), MPa	200				152	
Elongation at break (25°C), %	6				4.1	
Glass temperature, °C	310	> 270	> 270	> 350	310	> 350
Expansion coefficient ($\times 10^5$), K^{-1}		4.1	8		3.5	
Water absorption (48 h, 100°C), %		1.4	0.9	0.25		0.87
Relative permittivity (1 MHz)				2.57		2.7
Dissipation factor (1 MHz)				0.0008		0.0004

∗ 30°C

catalyst residues or other impurities. In the polymer literature, the term "ion-containing polymer" is used as a shorthand notation for "ion-containing polymer molecules"; i.e., it refers exclusively to ionic structures attached to polymer molecules.

Polymer molecules with ionic structures are often called ionic polymers. They are usually subdivided into two groups [47]: water-soluble and water-insoluble. Water solubility is commonly shown by polymer molecules with high contents of ionic groups; these polymers are called polyelectrolytes (Section 5.1.2). Polyelectrolytes may, for example, contain basic groups such as $-NH_2$ or acidic groups such as $-COOH$.

The terminology is less clear for water-insoluble polymers with small contents of ionic groups. If these groups are acid groups such as $-COOH$ or $-SO_3H$, the polymers are termed acidic but never ionic. The term "ionic" is reserved here for polymers with small proportions of salt groups such as $-COONH_4$, $-COONa$, or $(-SO_3)_2Zn$. The designation of these polymers as "ionic polymers" is furthermore not meant to imply that the ionic structures are indeed present as ions, i.e., with charge separation (Section 5.1.3).

The incorporation of small amounts of acidic and/or ionic groups into apolar backbone chains causes these groups to associate ("aggregate") in the solid state. The resulting intermolecular physical cross-links lead to higher moduli and strengths. The cross-links dissociate at higher temperatures and the polymers can then be processed conventionally.

5.1.2. Polyelectrolytes

The ionic groups of polyelectrolytes may be in side chains as in **8** and **9** or in backbones as in **10–12**:

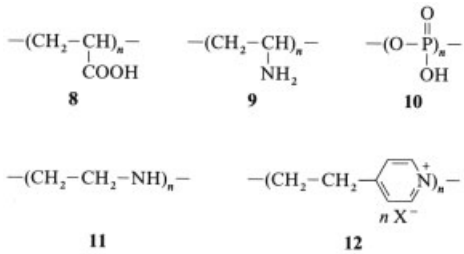

Polyelectrolytes may be polyacids such as poly(acrylic acid) **8** and poly(phosphoric acid) **10**, or polybases such as poly(vinyl amine) **9** and poly(ethylene imine) **11**. In water and other highly polar solvents, they dissociate into polyions and oppositely charged counterions. Upon dissociation, polyacids lose protons and become polyanions (e.g., $\sim[CH_2-CH(COO^-)]_n\sim$). The salts of these polyanions are thus called polysalts (e.g., $\sim[CH_2-CH(COO^-Na^+)]_n\sim$). Polybases can quaternize by adding protons, methyl groups, etc., and become polycations; the quaternization may be in the substituent as in **9** or in the main chain as in **11**. Polymers which only exist as polycations but not as polybases are called ionenes, an example of which is **12**, the polymer of 4-vinylpyridine by polymerization via the 1,6-positions.

Polyacids, polybases, polyanions, and polycations contain many acidic, basic, anionic, and cationic groups per molecule. They are to be distinguished from macroacids, macrobases, macroanions, and macrocations, which have only one such group (as in, e.g., living anionic polymerizations with monofunctional initiators), and from macrodiacids, etc., with two such functional groups, etc.

Polyampholytes are polymers whose chains carry both acidic and basic (or anionic and cationic) groups; an example is the copolymer of acrylic acid $CH_2{=}CHCOOH$ and acrylamide $CH_2{=}CHCONH_2$. Complexes of polyacids and polybases are called polyelectrolyte complexes or symplexes; they may or may not be water soluble, depending on their chemical structure (see also [47]).

5.1.3. States of Ions

In science "ion" denotes an electrically charged atom or molecule (cation or anion) which shows ionic conductivity as a free entity (in solution or as gas) or in its assemblies (in crystals or melts). Each ion possesses an oppositely charged counterion due to the principle of electroneutrality; the electric charges are separated. In polymer science, "ion-containing" does not necessarily refer to charge-separated entities, especially not in the solid state; it is merely necessary that these entities may be able to generate ions under favorable (polar) conditions.

Under less polar conditions, ions can form ion pairs consisting of oppositely charged ions bound together by Coulomb interactions but not by covalent bonds. In contact ion pairs, these ions are in direct contact. In the presence of polar solvents, the ions may be also more loosely held together in solvent-separated ion pairs; such ion pairs may share one solvent molecule or two or more.

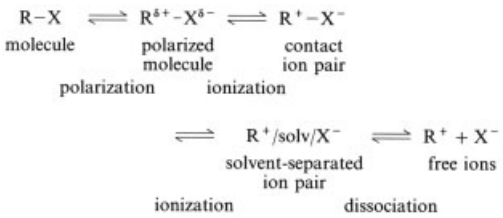

Triple ions arising from Coulombic interactions such as $\sim\sim COO^-/M^{2+}/^-OOC\sim\sim$ from two carboxylate groups $\sim\sim COO^-$ and a metal ion M^{2+} are especially stable ionic structures since the sum of the repulsion between two anions and the attraction between the cation and two anions results in a negative potential energy [48]. The same reasoning applies to higher ion aggregates; such structures are responsible for the stability of ion crystals such as NaCl. Triple ions are sometimes called triplets although the latter term should refer exclusively to multiplets of ion pairs.

Ion pairs can associate to form tightly aggregated ion multiplets consisting of associated ion pairs (2 ion pairs), triplets (3 ion pairs), etc.:

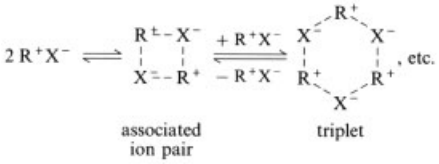

The maximum number of ion pairs in such multiplets should be six to eight if the multiplets are spherical. Ion multiplets do not contain polymer chain segments; sizes of spherical multiplets cannot exceed ca. 0.6 nm because of geometrical constraints on the multiplet structures by the chain segments to which the multiplets are attached. Because of their small size, they are not separate phases in solid polymers; they thus directly affect matrix properties such as segment mobilities. No size restriction is imposed on nonspherical geometries.

Multiplets may aggregate into ion clusters [49] which contain significant amounts of polymer segments in addition to the ionic structures [50], [51] (for a review of these and other theories, see [52]). The domains of these clusters are thus considerably larger than those of multiplets; their size is limited to 2 – 10 nm by the elastic forces generated by the backbone chains. These clusters approach the size of true phases. Clusters thus do not markedly affect matrix properties except that they may behave as cross-links with reinforcing action.

Free ions and solvent-separated ion pairs conduct electricity, but contact ion pairs, contact multiplets, and ion clusters do not. Most ion-containing polymers are electrical insulators

in the solid state; since obviously no ions are present, they really should not be called ion-containing polymers.

5.2. Acidic Copolymers

5.2.1. Introduction

Acidic copolymers are defined as copolymers of a parent monomer (or several thereof) and a small amount of an acidic comonomer. The polymers may be thermoplastic or elastomeric, depending on the glass transition temperature of the parent monomer segments.

The first acidic elastomers were copolymers of butadiene with small amounts of various unsaturated acids, patented by I. G. Farbenindustrie in 1930 [53]. In 1950, B. F. Goodrich started to introduce a series of acidic elastomers from butadiene, acrylonitrile, and small amounts of acrylic acid [54], [55]; the trade name Hycar now refers to polymers of different compositions and grades with many different CAS numbers. These and other acidic elastomers are also known as carboxylated elastomers or carboxylated rubbers; they are produced by many companies [56].

Hycar rubbers can be neutralized with zinc oxide or zinc salts; these neutralized acidic polymers were the first elastomeric ionomers. However, the ionomeric structures served only as reinforcing entities since the rubbers were still conventionally vulcanized to covalently cross-linked elastomers.

Neutralization of suitable acidic copolymers without chemical cross-linking leads to thermoplastic elastomers. The first thermoplastic elastomers were the cured Hypalon polymers of Du Pont. Hypalons are based on sulfochlorination of polyethylenes [57] and are available in many grades. The curing of these polymers by metal compounds leads to ionomers which behave as thermoplastic elastomers.

In the mid-1960s Du Pont began production of the first thermoplastic acidic copolymers from ethylene (E) with small amounts of methacrylic acid (MAA). These copolymers were not commercially available but were converted in-house to ionomers (Section 5.3.6).

Thermoplastic acidic EAA copolymers of ethylene (E) and up to 15 wt % acrylic acid (AA) were first commercially manufactured in various grades by Dow Chemical and Union Carbide (Section 5.2.3). They were followed by Du Pont's E-MAA copolymers with up to 15 mol % MAA (Nucrel).

Du Pont also manufactures a 1 : 1 copolymer of ethylene and methyl acrylate with small amounts of (meth) acrylic acid (Vamac) (Section 5.2.4).

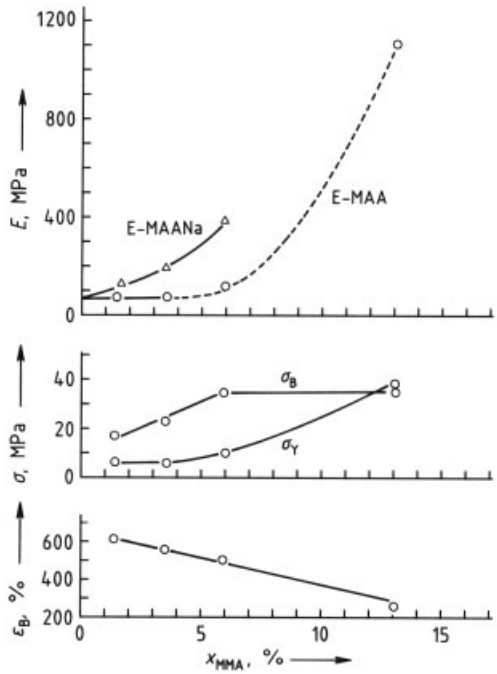

Figure 2. Tensile modulus E, yield strength σ_Y, tensile fracture strength σ_B, and elongation at break ε_B of ethylene – methacrylic acid copolymers (○) and their sodium salts (▲) (100 % neutralization) as function of the mole fraction x_{MAA} of methacrylic acid units (experimental data from [58])

5.2.2. Structure and Properties

Acidic copolymers can be amorphous or partially crystalline, depending on the crystallizability of the segments of the parent monomer units. The crystalline domains are composed of lamellae formed by folded chains of parent segments and do not contain acidic units, the latter being confined to the amorphous regions where they may be present as dimers. Associated acidic units do not form separate phases. Acidic copolymers are thus one-phase materials

if amorphous, and two-phase polymers if crystalline.

Properties of such polymers depend on intramolecular and intermolecular interactions between their segments and groups. COOH groups in polyethylene segments tend to associate; the associated groups act as cross-links. At low COOH content, this association can hardly take place since the COOH groups are not free but bound to polyethylene; the geometrical constraints imposed upon the groups reduce the probability of association. Since the crystallinity of the polymer is not markedly affected, Young's moduli of ethylene – methacrylic acid copolymers remain constant up to 4 – 5 mol % of methacryclic acid groups. At higher COOH group concentrations, intermolecular association is increasingly possible. Intermolecularly associated groups act as cross-links and the moduli and yield strengths increase sharply (Fig. 2).

Polyethylene yields easily and exhibits stress-softening. The increasing association of COOH groups impedes the flow of segments, and the elongation at breaks decreases. Tensile fracture strengths of ethylene – methacrylic acid copolymers increase with increasing COOH content, eventually becoming independent of the proportion of COOH groups.

5.2.3. Thermoplastic Acidic Copolymers

Syntheses. EAA copolymers of ethylene (E) and up to 15 wt % acrylic acid (AA) are produced by Dow Chemical and Union Carbide in various grades. Du Pont manufactures an EMAA copolymer of ethylene and <15 mol % methacrylic acid MAA (Nucrel [59]) and an 1: 1 EMA copolymer of ethylene and methyl acrylate (MA) with small amounts of (meth)acrylic acid (Table 4).

None of the industrial polymerization processes have been disclosed. They are most likely high-pressure free-radical polymerizations. Since the *Q, e* values of ethylene ($Q_E = 0.015$; $e_E = -0.200$) and methacrylic acid ($Q_{MAA} = 2.34$; $e_{MAA} = 0.65$) are unfavorable for random arrangements of comonomeric units, copolymerizations are probably limited to ca. 10 % conversion in order to prevent the formation of long blocks of methacrylic acid units. Calculation of copolymerization parameters r from $r_i = (Q_i/Q_j) \cdot \exp[-e_i(e_i - e_j)]$ gives $r_{MAA} = 2.07$ and $r_E = 0.234$ and thus number-average sequence lengths $N_{MAA} = 1 + r_{MAA}$ ([MAA] / [E]) of methacrylic acid units of only $N_{MAA} \approx 1.23$ if the monomer ratio is $[MAA]/[E] = 90/10$. *Q,e* values of acrylic acid or copolymerization parameters of the ethylene – acrylic acid system have not been published but they should be similarly unfavorable.

Structure and Properties. High-pressure radical polymerization of ethylene leads to polyethylenes with a few short and long branches. The more or less random distribution of fairly bulky acid or salt units along the chains reduces the chain lengths of crystallizable segments of such polyethylenes. Melting temperatures thus decrease with increasing size of the substituents, e.g., for ethylene polymers with ca. 3 mol % comonomer units from 115 °C (LDPE with no comonomer units) to 104 °C (ethylene – acrylic acid), 95 °C (ethylene – sodium acrylate), and 90 °C (ethylene – methyl acrylate). Since the physical cross-linking caused by such comonomer units shortens the segments between cross-links, glass transition temperatures increase from −80 °C (LDPE) to −35 °C (ethylene – methyl acrylate), 0 °C (ethylene – acrylic acid), and 50 – 60 °C (ethylene – sodium acrylate).

The density of EAA copolymers increases linearily with the content of acryl acid units AA, reaching 0.95 g/cm^3 at 15 wt % AA units. Tensile yield strengths first decrease with increasing AA content, pass through a minimum at 12 – 15 wt % AA units, and increase again. The fracture strength exhibits a maximum at approximately the same AA content. Elongations at break are ca. 600 %.

Ethylene – acid copolymers show outstanding toughness since incorporated acrylic acid units reduce the formation of spherulites by polyethylene segments and simultaneously promote intermolecular association. Toughness is maintained at low temperatures; the brittleness temperature is ca. −70 °C for polymers with ca. 3 – 9 wt % acid units.

Ethylene acid copolymers show excellent adhesion to metals, glass, and paper due to the polar COOH groups and the low melt viscosities of the polymers. They also adhere well to polyeth-

Table 4. Industrial copolymers of ethylene and (meth)-acrylic acid

Parent monomers	Comonomers	Trade name	Manufacturer
Ethylene	acrylic acid	EAA 469	Dow Chemical
Ethylene	acrylic acid		Union Carbide
Ethylene	methacrylic acid	Nucrel	Du Pont
Ethylene + methyl acrylate	(meth)acrylic acid	Vamac	Du Pont

ylene because of their polyethylene segments. Adhesion to other plastics can be improved by oxidation of EAA and EMAA surfaces.

Processing. EAA and EMAA can be processed similarly to LDPE. Since the acid groups are somewhat corrosive to metals, corrosion-resistant metals or nickel or chromium plating are recommended for processing equipment.

Certain ethylene acid copolymer grades can be dispersed in hot aqueous ammonia solutions to give dispersions stable at room temperature. They lose ammonia upon drying and form surface coatings with the same properties as copolymers processed otherwise.

Applications. Present applications comprise inner laminate layers on aluminum foil for pouches (towelettes, condiments) and tubes (toothpaste), cable sheathing, pipe coatings, and laminates on metals and glass fiber for automotive and building applications, and on polyurethane foams for carpet underlay.

FDA regulations for direct food contact with ethylene acid copolymers exist for EAA resins containing up to 10 % acrylic acid units and EMMA resins with up to 20 % methacrylic acid units.

Prices of resins are between those of low-density polyethylenes and Surlyn A ionomers.

5.2.4. Elastomeric Acidic Copolymers

Since 1976, Du Pont has offered Vamac, a copolymer with ethylene units $-CH_2-CH_2-$ and methyl acrylate units $-CH_2-CH(COOCH_3)-$ in the molar ratio 1: 1 and a small amount of undisclosed cure site units $-R(COOH)-$. The polymer is probably synthesized by free-radical emulsion polymerization. Two carbon-black-filled masterbatches (e.g., type B-124 with 24 parts SRF carbon black and auxiliaries per 100 parts polymer) and two neutral masterbatches (e.g., type N-123 with 23 parts siliceous earth per 100 parts polymer) are available [60].

All Vamac grades can be vulcanized by diamines or hexamethylene diaminocarbamate, and N-123 also by free radicals. Vulcanized Vamacs are EE – EH elastomers according to ASTM D-2000 and SAE J-200 classifications, i.e., elastomers with continuous service temperatures up to 175 °C (E) and swelling between 20 % (H) and 80 % (E) after 70 h in ASTM oil no. 3. Vamac is resistant to hot oils, hydrocarbons, glycols, and engine cooling fluids but not to brake fluids. Its weatherability is excellent.

The heat distortion of Vamac is surpassed only by the more expensive fluoroelastomers and silicones; however, Vamac has better mechanical properties. After 30 min vulcanization at 177 °C, the rebound elasticity of B-124 is 35 %, the tensile strength 14 MPa, the ultimate elongation 440 %, and the 100 % modulus 2.9 MPa. After 7 d at 177 °C, the modulus increases to 5.3 MPa and the tensile strength to 16.5 MPa, whereas the elongation decreases to 240 %, all probably due to an after-vulcanization.

Vamac is used for tubes and seals and as a damping material. It can replace sulfochlorinated polyethylene in cables and electrical systems since it burns with little smoke of low toxicity.

5.3. Ionomers

5.3.1. Introduction

The name ionomer was originally introduced by Du Pont for (partial) salts of copolymers of olefins with small amounts of carboxyl group containing monomers, such as ethylene – methacrylic acid copolymers [61]. The term ionomer is now used as a generic term for thermoplastic and elastomeric hydrocarbon polymers with

Table 5. Industrial ionomers

Parent monomeric unit	Ionomeric unit	Ions	Trade name	Manufacturer
Thermoplastics				
Ethylene	methacrylic acid	Zn, Na	Surlyn A	Du Pont
Tetrafluoroethylene	perfluoro vinyl ether A ∗		Nafion	Du Pont
Tetrafluoroethylene	perfluoro vinyl ether B ∗		Flemion	Asahi Glass
Styrene	styrene *p*-sulfonic acid			Exxon
Elastomers and thermoplastic elastomers				
Ethylene	post-sulfochlorination		Hypalon	Du Pont
Ethylene + propylene+ nonconjugated diene	post-sulfonation		Thionic	Uniroyal
Butadiene + acrylonitrile	acrylic acid	Zn	Hycar	Goodrich

∗ Perfluoro vinyl ether: $CF_2{=}CF[O(CF_2CF(CF_3)O)_mR]$ with $R = (CF_2)_2SO_3Na$ (A) or $R = (CF_2)_nCOONa$ (B).

small amounts of partially or completely neutralized pendant acid groups which show typical ionomer properties (Table 5). Total acidic group contents (free or neutralized) are usually under ca. 10 mol %. The syntheses, structures, properties, and applications of ionomers have been described in many books [62–64], reviews [50], [65–71], handbook articles [72–74], and symposia [75–77].

The acidic monomer units of such copolymers may be directly introduced by copolymerization as in ethylene/acrylic acid copolymers, or subsequently by chemical transformation of a primary copolymer as in the sulfonation of copolymers of ethylene, propylene, and a nonconjugated diene (Table 5). In ionomers, acidic monomer units are furthermore partially or completely converted into their salts. Polymer literature restricts the term "ionic" to these salts: polymer molecules with salt groups such as $-COONa$ or $-COONH_4$ are called ionic, whereas those with groups such as $-COOH$ or $-SO_3H$ are not considered ionic but are sometimes referred to as acidic copolymers. Ionic and acidic copolymers with small proportions of ionic and/or acidic groups are always insoluble in water.

At present, only three thermoplastic ionomers (Surlyn A, Nafion, Flemion), two thermoplastic elastomers (Hypalon, Thionic), and many carboxylated rubbers (Hycar and others) are manufactured. Many other ionomers have been reported in the literature, e.g., carboxylic, phosphonic, sulfonic, and thioglycolic salts of polypentenamers [78], carboxylated and sulfonated polystyrenes [79], [80], and telechelic poly(isobutylene sulfonate) ionomers [81].

5.3.2. Structure

The first thermoplastic ionomers were copolymers of ethylene and methacrylic acids, partially neutralized with zinc or sodium compounds [82]. These Surlyn A polymers of Du Pont have much better clarity and tensile properties than comparable polyethylenes, which stimulated much research.

The enhanced properties of Surlyn A polymers were first thought to be due to decreased crystallinity and the presence of cross-linking triple ions of one bivalent metal cation (Zn^{2+}) and two carboxyl anions. It was subsequently shown that crystallinities are not markedly lower than those of the parent polymers, and that the good transparency was due to the absence of spherulites. These property changes also occur with monovalent metal cations (such as Na^+) so that other ionic cross-links must be present; these were later found to be ion multiplets and ion clusters.

Ionomers may be amorphous or partially crystalline, depending on the crystallizability of the segments of the parent monomer units. Amorphous ionomers are two-phase polymers. The continuous phase of these ionomers is amorphous and composed of parent segments with interdispersed ion pairs and ion multiplets. The discontinuous phase of amorphous ionomers consists of ion clusters of ion pairs and ion multiplets with interdispersed polymer segments. Crystallizable ionic copolymers possess crystalline regions as a third phase in addition to the amorphous phase and the ion clusters [49] (Fig. 3).

Crystalline domains do not contain acidic or ionic units; the segments in these domains are chain-folded. Acidic units are confined to amor-

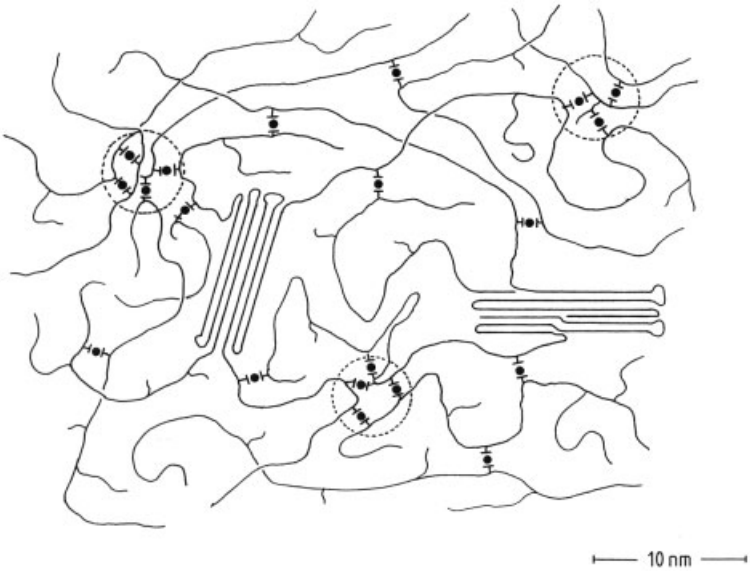

Figure 3. Schematic representation of the structure of crystallizable ionomers with crystalline (lamellar) domains, amorphous regions with ion triplets (and multiplets), and ion clusters with interdispersed chain segments (e.g., zinc-neutralized low-density ethylene – methacrylic acid copolymers)

phous regions, whereas ionic units are either in amorphous regions or in ion clusters. The amorphous regions contain polymer segments with acidic or ionic units, acidic units as dimers, and ionic units as ion pairs and/or ion multiplets in a single phase. Ion clusters are separate phases that contain ion multiplets and polymer segments. The exact structure, size, and size distribution of amorphous and crystalline phases, ion clusters, ion multiplets, ion pairs, and associations of acidic groups are not known.

5.3.3. Properties

Ionic Group Content. Ionomers behave either as physically cross-linked thermoplastics or as thermoplastic elastomers, depending on the segment mobility of the parent monomer units. In amorphous polymers, the mobility is determined by the glass transition temperature, and in crystalline polymers also by the melting temperature, both relative to the service temperatures. Melting temperatures of crystalline polymers are mainly affected by the segments of the parent units and only secondarily by acidic or ionic units, which merely reduce the segment lengths of crystallizable units.

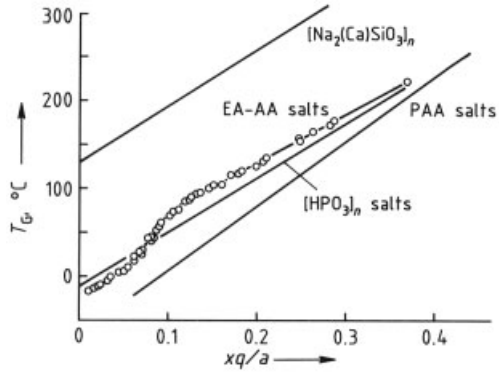

Figure 4. Glass transition temperature T_G as function of the linear charge density xq/a where x is the "metal carboxylate content" (probably in mol % of ionic units), q the cation charge (in "units of electron charges"), and a the distance between centers of charge (probably in 0.1 nm) [85]
○ $= T_G$ of ion clusters of ethyl acrylate – acrylic acid copolymers EA – AA with Li, Na, K, Cs, Ca, and Ba [84]
– $= T_G$ of amorphous domains of $[Na_2(Ca)SiO_3]_n$ glasses, salts of $[HPO_3]_n$, and salts of poly(acrylic acid) PAA

However, the glass transition temperatures T_G reflect the structure and mobility of noncrystalline phases, i.e., amorphous phases and ion clusters. Ionomers thus show two glass transition temperatures [61].

Ion pairs and ion multiplets act as cross-linking agents in amorphous phases. The higher

the concentration of these physical cross-links, the shorter are the segments between cross-links. Cooperative movements of the units in segments are affected if the segment lengths are shorter than $N_c = 30-50$ chain atoms. At $N_c < 30-50$, glass transition temperatures of amorphous phases increase linearily with the concentration of ionic units (Fig. 4).

Glass transition temperatures of ion clusters are strongly influenced by ion pair and multiplet formation. Ions of an ion pair can only move away from each other above T_G; the required work W_{el} must be proportional to kT_G. This work is given by the electrostatic interaction F_{el} and the internuclear distance a. Because $W_{el} \sim kT_G \sim \int F_{el} da$, one obtains $TG \sim q_{cat} q_{an}/a$, where $q$cat and q_{an} are the cation and anion charges. The anion charge is constant in ionomers and thus $TG \sim q_{cat}/a \sim q/a$ [83]. In copolymers, only a fraction x of the monomeric units are ionic; the glass transition temperature of ionomers must be therefore proportional xq/a [84] (Fig. 4). The glass transition temperatures of ionic clusters usually show a sigmoidal dependence on either xq/a or x.

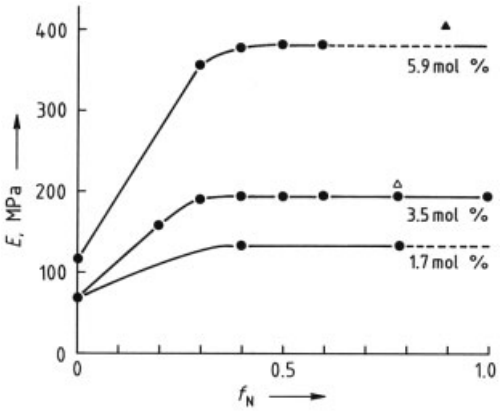

Figure 5. Young's modulus E of ethylene/methacrylic acid copolymers with 1.7, 3.5, and 5.9 mol % methacrylic acid units as function of the degree of neutralization fN of COOH groups by NaOH (●), $Sr(OH)_2$ (▲), or $Cu(OH)_2$ (△) Reproduced with permission from [58].

Effect of Neutralization. Elastic moduli of ethylene – methacrylic acid polymers with COONa groups are higher than those of the same polymers with COOH groups (Fig. 2). Increasing neutralization of COOH groups thus leads to an increase of moduli up to degrees of neutralization of $f_N \approx 30$ mol % (Fig. 5); moduli become constant at higher degrees of neutralization, which indicates that no new cross-linking sites are formed beyond this concentration.

The situation is quite different for partially neutralized copolymers of ethylene and acrylic acid. Tensile moduli first increase with higher degrees of neutralization f_N, but do not become constant as in ethylene/methacrylic acid copolymers; instead they decrease if $f_N > 0.3$ (Fig. 6). Fracture strengths σ_B increase monotonically with higher degrees of neutralization; they do not become constant. The variation of modulus and fracture strength with degree of neutralization seems to be independent of the type of monovalent cation, whereas the elongation at break depends on the ion type, especially at high degrees of neutralization. These effects are probably due to differences in the structure and/or concentration of ion clusters; experimental data are not available.

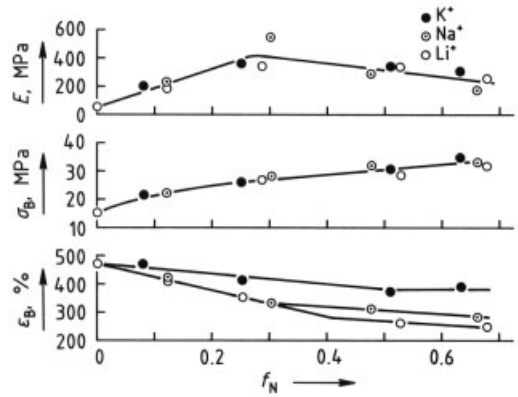

Figure 6. Variation of tensile modulus E, fracture strength σ_B, and elongation at break ε_B with degree of neutralization f_H of ethylene–acrylic acid copolymers (experimental data from [85])

Ionic domains are present in neutralized ethylene – methacrylic acid copolymers at temperatures up to 300 °C, far above the crystallite melting temperature, according to small-angle X-ray and neutron-scattering experiments. The number of ionic domains increases with the degree of neutralization f_N. Since they represent physical cross-linking sites, they are responsible

for an increase of melt viscosity with increasing neutralization, i.e., a decrease in melt volume index (melt flow index) MVI (Fig. 7). The size of these ionic domains is not affected by the type of monovalent ion nor by the applied pressure as the similar slopes of the functions log MVI = f (f_N) show.

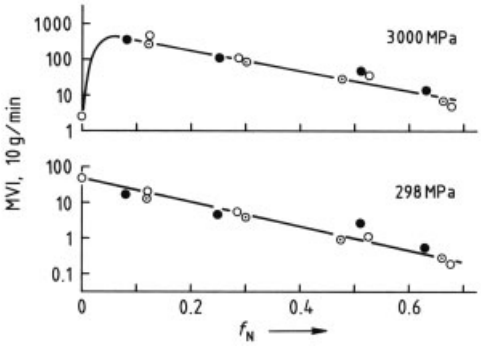

Figure 7. Effect of the degree of neutralization by Li^+ (○), Na^+ (⊙), and K^+ (●) on the melt volume index MVI (at 298 and 3000 MPa) (melt flow index) of copolymers of ethylene with 15 wt % acrylic acid (experimental data from [85])

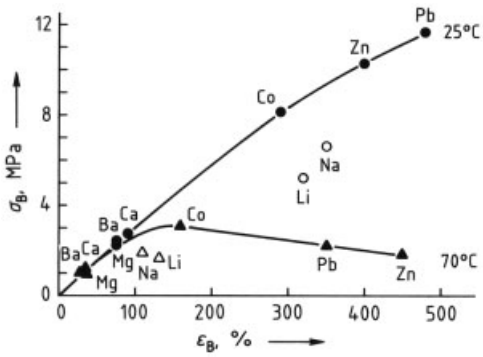

Figure 8. Tensile fracture strength σ_B as a function of elongation at break ε_B of salts of sulfonated EPDM polymers (32 milli-equivalents of sulfone groups per 100 g polymer) at 25 °C (●,○) and 70 °C (△,▲) (experimental data from[86]) ●, △ = Bivalent metals; ○, ▲ = Monovalent metals

Influence of Counterion Type. The association of salt groups depends on the coordination number of the counterions and not on its valency. Monovalent ions such as Li^+, Na^+, and K^+ and bivalent ions such as Mg^{2+}, Ca^{2+}, and Zn^{2+} cause multiplet and cluster formation (Fig. 3). The effects are however quantitatively different: at 25 °C and the same elongations at break, bivalent ions lead to higher tensile fracture strengths than monovalent ions (Fig. 8). At this temperature, higher strengths are accompanied by higher elongations for the two series of bivalent and monovalent metal cations. In each series, these properties increase with increasing atomic number of the metal.

At the higher test temperature of 70 °C, ionomers with bivalent metals of group 2 of the periodic table (Mg, Ca, Ba) still follow the $\sigma_B = f(\varepsilon_B)$ relationship for 25 °C but the members of higher groups (Co, Zn, Pb) now not only lead to much lower fracture strengths but to a decrease in strength with increasing elongation at break (Fig. 8). These changes indicate that the structures of physical cross-linking sites formed by metals with higher group numbers change at higher temperatures.

5.3.4. Ethylene – Methacrylic Acid Ionomers

Many thermoplastic ionomers have been reported in the literature but only three have been commercialized: Surlyn A (Du Pont) for films, coatings, and molded articles, and two perfluorinated ionomers for membranes (Nafion from Du Pont, Flemion from Asahi Glass).

Synthesis. Surlyn A (Du Pont) is a copolymer of ethylene with ca. 10 wt % methacrylic acid; 50 % of the COOH groups are present as the sodium salt [*25608-26-8*] or zinc salt [*28516-43-0*]. CAS registers these copolymers as "2-propanoic acid, 2-methyl-, polymers," i.e., as polymers of methacrylic acid, although methacrylic acid is clearly the minority component.

The commercial synthesis of Surlyn A polymers has not been disclosed. Most likely, it is a high-pressure free-radical copolymerization of ethylene and methacrylic acid. The acid groups are then partially neutralized by rolling the polymers with metal acetates, alkoxides, carbonates, or hydroxides with release of acetic acid, alcohols, carbon dioxide, or water, respectively.

Commercial products contain either sodium or zinc. Polymers with ammonium derivatives or lithium can also be processed but these polymers are not commercial. Magnesium- and barium-containing polymers possess such high viscosities that they can no longer be processed.

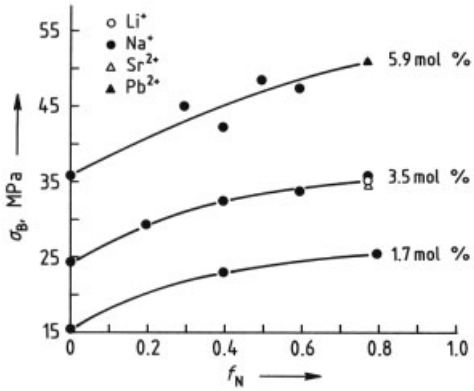

Figure 9. Tensile fracture strength σ_B of copolymers of ethylene with various mole fractions x_{MAA} of methacrylic acid as function of the degree f_N of neutralization
Reproduced with permission from [58].

Structure. Surlyn A polymers have a multiphase structure consisting of crystalline regions with chain-folded polyethylene segments, amorphous regions containing ion multiplets, and ion clusters with interdispersed polyethylene segments.

The introduction of acid/salt groups into polyethylene chains prevents the formation of spherulites but does not markedly change the crystallinity. Surlyn A polymers therefore have good optical clarity (haze $< 2\,\%$), especially the low-crystallinity grades. Transparency increases with increasing degree of neutralization; it becomes constant at neutralizations $>$ca. 30 – 50 %.

Surlyn A polymers have chemical structures similar to those of alkali metal salts of long-chain fatty acids. Since the latter are nucleating agents, similar behavior can be expected from Surlyn A polymers. Upon annealing, the crystallinity of these polymers does indeed increase. Spherulites form at temperatures above ca. 90 °C, leading to haziness.

Properties. Some properties of Surlyn A resins are listed in Table 6. Crystalline regions and ion clusters all act as physical cross-linking sites for the rubbery, ion-multiplet-reinforced amorphous domains; Surlyn A polymers thus have the characteristics of high-impact multiphase polymers. Tensile strengths are higher than those of comparable polyethylenes and increase with the proportion of acid units and the degree of neutralization (Fig. 9). Stiffness and hardness are similar to those of LDPE. The various grades retain impact strengths up to their brittleness temperatures, which are between −100 and −130 °C. The continuous service temperatures of ca. 70 °C are fairly low. The elongation at break is ca. 350 %, and the abrasion resistance is excellent.

The physical cross-linking of chains by ion clusters causes high melt viscosities (low melt volume indices) which are much higher than those of comparable polyethylenes. Melt strengths are high. The cross-linking also reduces the creep under constant load.

At room temperature, Surlyn A ionomers are insoluble in all common organic solvents. They dissolve at temperatures above ca. 70 °C in 60/40 vol/vol mixtures of toluene/isobutanol, toluene/methyl isobutyl ketone, or tetrachloroethylene/butanol, or in a 70/20/10 mixture of decahydronaphthalene/diethylene glycol dimethyl ether/dimethyl sulfoxide. The polymers have good resistance to stress corrosion, except by certain combinations of detergents and alcohols. The permeability for oils and fats is low, while water vapor permeabilities correspond to those of low-density polyethylene. The polymers are readily degraded by light; UV absorbers extend outdoor service lives to ca. 2 years. Surlyn A polymers burn slowly when in contact with flames.

Carboxyl groups provide Surlyn A polymers with good adhesion to metals, glass, and paper; the adhesive properties depend on the carboxyl group content. Adhesion on impact is good, as e.g., required by the painted surfaces of molded golf balls.

Applications. The combination of good impact strength, high melt viscosity, high melt strength, good transparency, and relatively low continuous service temperatures is ideal for the manufacture of extremely thin packaging films. Such films have high tear strengths and are especially well suited for the packaging of sharp objects. Co-extruded ionomer – LDPE films are used for sterilizable packaging of foods and pharmaceuticals. Coatings of zinc-containing Surlyn A grades make glass bottles splinter-resistant. The combination of toughness and good abrasion resistance recommends the polymers for shoe soles and heels. Other applications

Table 6. Properties of Surlyn A Resins

Property	Na Resin	Zn Resin
Density, g/cm^3	0.94	0.94
Melting temperature, °C	94	95
Heat deflection temperature (455 kPa), °C	40	41
Thermal expansion coefficient, K^{-1}	14 × 10-5	16 × 10-5
Melt flow index, g/10 min	1.3	5.5
Flexural modulus, MPa	220	130
Tensile strength at yield, MPa	12.4	8.3
Tensile strength at break, MPa	29	21.4
Elongation at break, %	450	500
Tensile impact strength, kJ/m^2	1160	925
Notched impact strength, J/m	610	no break
Hardness (Shore D)	60	54

are sheathing of electric cables and wires, automotive parts, and floor mats. Molded articles can be directly painted.

Closed-cell Surlyn A foams with densities of 50 – 700 kg/m^3 can be produced by extrusion or injection molding. Applications include carpet backings, ski boots, hockey sticks, helmets, and tennis court foundations.

5.3.5. Fluorinated Ionomers

Precursors of fluorinated ionomers are manufactured by free-radical copolymerization of tetrafluoroethylene (TFE) with perfluorovinylethers (FVE)

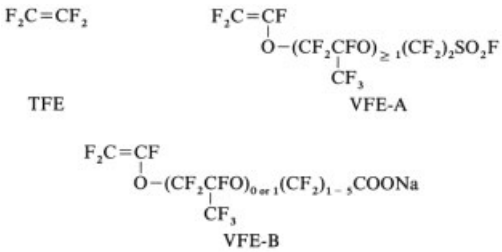

in molar ratios of 5 – 10 mol TFE per mole PVE [87], [88]. VFE-A is synthesized by reaction of hexafluoropropylene oxide (from the oxidation of TFE) with the sultone from the addition of sulfur trioxide to tetrafluoroethylene. Copolymers of TFE and VFE-A are known as XR Resins (Du Pont) whereas copolymers of TFE and VFE-B are called Flemion (Asahi Glass).

XR Resins can be melted and fabricated into membranes and tubes by conventional processing methods for thermoplastics. The $-SO_3F$ end groups are then saponified by hot caustic soda to give $-SO_3Na$ end groups, which in turn are converted into sulfonic acid groups with, for instance, nitric acid.

Du Pont manufactures articles with free sulfonic acid groups as films, tubes, or laminates under the trade name Nafion. Similar carboxylated fluoro ionomers are produced by Asahi Glass under the trade name Flemion. These polymers are permselective [89]: they allow the passage of cations but not of anions and are therefore used as membranes for electro-chemical processes, e.g., for the production of sodium hypochlorite. The exchange capacity of such resins is ca. 0.85 mmol/g, the fracture strength of resins with ca. 25 % water content ca. 21 MPa, and the elongation at break ca. 150 %. The strength can be increased by reinforcement with fabrics made from polytetrafluoroethylene fibers [90].

5.3.6. Sulfochlorinated Polyethylene

Simultaneous chlorination by Cl_2 and sulfonation by SO_2 of polyethylenes in homogeneous solution under the action of free-radical initiators or light leads to polymers with $-CH_2-$, $-CHCl-$, and $-CHSO_2Cl-$ groups in the approximate ratio 100 : 30 : 1. The polyethylene segments of these polymers are too short to crystallize, and the polymers thus behave as elastomers because of the low glass transition temperature (actual data unknown). Du Pont manufactures chlorosulfonated polyethylenes (CSM) in various grades (different CAS numbers) under the name Hypalon. Apart from a handbook article [91], little has been published about CSM polymers.

CSM polymers can be cured by way of the SO_2Cl groups [92]. The white polymers need no reinforcing fillers and can be used for many colored products. The presence of polar Cl and SO_2Cl groups provides the elastomers with excellent resistance to oil and ozone compared to standard synthetic rubbers. The maximum service temperature is ca. 120 °C.

5.3.7. Sulfonated EPDM Polymers

Synthesis and Structure. EPDM polymers are produced by Ziegler – Natta copolymerization of ethylene (E), propylene (P), and a small amount of a nonconjugated diene, preferably 5-ethylidene-2-norbornene (EN). The amorphous, rubbery polymers contain the monomeric units E, P and EN; the latter can be sulfonated by $H_2SO_4/(CH_3CO)_2O$ to sulfo units SEN [73]:

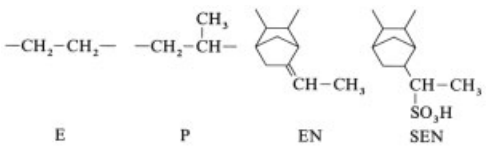

Sulfo-EPDMs can be neutralized with metal alkoxides or other bases, either in solution with subsequent precipitation of the neutralized polymer or by reactive milling of the polymer in an extruder. These thermoplastic elastomers were developed by Exxon and are now being commercialized by Uniroyal under the trade name Thionic.

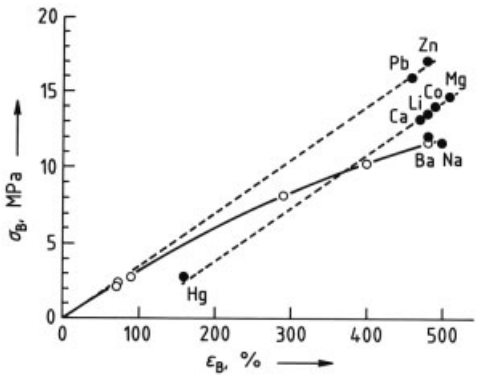

Figure 10. Tensile fracture strength σ_B as function of elongation at break ε_B of sulfonated EPDM ionomers before (○–○, bivalent ions only) and after (●– – –●) plasticization with zinc stearate (experimental data from [86])

Properties. Mechanical properties of sulfo-EPDM ionomers depend strongly on the type and concentration of cations (Fig. 8). Tensile strengths of these ionic polymers are typically much more enhanced over those of the acidic base polymers as compared to other ionomers.

The strong associations caused by ion clusters decrease with increasing temperature (Fig. 8). Even then the melt viscosity is very high at the usual processing temperatures. The Mg, Ca, Co, Li, Ba, and Na ionomers of Figure 8 typically have apparent melt viscosities between 55×10^4 Pa · s (Mg) and 51×10^4 Pa · s (Na) at 200 °C and a shear rate of 0.88 s^{-1}; for unknown reasons, melt viscosities of Pb and Zn ionomers are much lower (33×10^4 and 12×10^4 Pa · s, respectively [86]. Lead and zinc ionomers also exhibit melt fracture at higher shear rates (88 and 147 s^{-1}) than the other ionomers ($< 0.88\ s^{-1}$).

The high melt viscosities can be lowered by ionic-domain "plasticizers" such as zinc stearate [86]. Such additives interact preferentially with the ionic clusters, lessen the bonding in the cluster region and thus decrease the melt viscosity. Despite its designation as local or specific "plastification", the phenomenon is not a plastification in the scientific meaning of the word since glass transition temperatures are not affected. "Plasticized" zinc and lead sulfo-EPDMs show also higher fracture strengths than the corresponding unplasticized ionomers (Fig. 10). The phenomenon seems to be due to a reduction of the strength of the cross-linking ion clusters upon shearing the melt

6. References

1. M. S. Shartsberg, I. L. Kotlyarevskii, *Usp. Khim.* **29** (1960) 1439 – 1473; *Chem. Revs. USSR* **29** (1960) 662.
2. W. Ried, D. Freitag, *Angew. Chem.* **80** (1968) 932 – 942.
3. A. H. Frazer: *High Temperature Resistant Polymers,* Interscience, New York 1968, pp. 38 – 49.
4. P. Kovacic, F. W. Koch: "Polyphenylenes," *Encycl. Polym. Sci. Technol.* **11** (1969) 380 – 389.
5. J. G. Speight, P. Kovacic, F. W. Koch, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 5** (1971) 295 – 386; *Rev. Macromol. Sci.* **6** (1972) 295 – 386.

6. K.-U. Bühler: *Spezialplaste,* Akademie Verlag, Berlin 1978, pp. 136 – 157.
7. H. Naarmann, *Angew. Makromol. Chem.* **109/110** (1982) 295 – 338.
8. N. Bilow, L. J. Miller, *J. Macromol. Sci. Chem.* **A 3** (1969) 501 – 525.
9. G. Ensor, *Br. Polym. J.* **2** (1970) 264 – 269.
10. H.-G. Elias, F. Vohwinkel: *Neue polymere Werkstoffe für die industrielle Anwendung,* 2. Folge, Hanser Verlag, Munich 1983, pp. 78 – 81; New Commercial Polymers 2, Gordon and Breach, New York 1986, pp. 75 –79.
11. G. G. Engstrom, P. Kovacic, *J. Polym. Sci. Polym. Chem. Ed.* **15** (1977) 2453 – 2468.
12. L. C. Cessna, Jr., H. Jabloner, *J. Elastomers Plast.* **6** (1974) 103 – 113.
13. T. M. Bednarski, J. H. Del Nero, R. H. Mayer, J. A. Hagan, *SPE Ann. Techn. Papers* **21** (1975) 90 – 93.
14. J. E. French, *ACS Coat. Plast. Prepr.* **35** (1976) no. 2, 72 – 76.
15. H.-G. Elias: *New Commercial Polymers 1969 – 1975,* Gordon and Breach, New York 1977, pp. 26 – 27.
16. J. Jabloner, L. C. Cessna, Jr., *ACS Polymer Prepr.* **17** (1976) no. 1, 169 – 174.
17. V. A. Sergeev, V. K. Shitikov, Yu. A. Chernomordik, V. V. Korshak, *ACS Polym. Preprints* **16** (1975) no. 1, 328 – 332.
18. L. A. Errede, M. Szwarc, *Q. Rev., Chem. Soc.* **12** (1958) 301 – 320.
19. H. L. Lee, D. G. Stoffey, K. Neville: *New Linear Polymers,* McGraw-Hill, New York 1967, pp. 83 – 100.
20. A. H. Frazer: *High Temperature Resistant Polymers,* Interscience, New York 1968, pp. 53 – 69.
21. S. N. Nowikow, I. E. Karbasch, A. N. Prawednikow, *Vysokomol. Soedin. Ser. B* **16** (1974) 292 – 295.
22. M. Szwarc, *Polym. Eng Sci.* **16** (1976) 473 – 479.
23. L. Baldauf, C. Hamann, L. Libera, *Plaste Kautsch.* **25** (1978) 61 – 64.
24. *Kirk-Othmer,* 3rd ed., **24,** 744 – 771.
25. W. F. Beach et al., "Xylylene Polymers," *Encycl. Polym. Sci. Eng.* **17** (1989) 990 – 1025.
26. M. Szwarc, *Nature (London)* **160** (1947) 403.
27. M. Szwarc, *Discuss. Faraday Soc.* **2** (1947) 46 – 49.
28. M. Szwarc, *J. Chem. Phys.* **16** (1948) 128 – 136; *J. Polym. Sci.* **6** (1951) 319 – 323.
29. M. H. Kaufman, H. F. Mark, R. B. Mesrobian, *J. Polym. Sci.* **13** (1954) 3 – 20.
30. J. R. Schaefgen, *J. Polym. Sci.* **15** (1955) 203 – 219.
31. L. A. Errede, B. F. Landrum, *J. Am. Chem. Soc.* **79** (1957) 4952 – 4955.
32. L. A. Errede, F. DeMaria, *J. Phys. Chem.* **66** (1962) 2664 – 2672.
33. W. F. Gorham, *ACS Polymer Preprints* **6** (1965) no. 1, 73 – 83.
34. W. F. Gorham, *J. Polym. Sci. [A-1]* **4** (1966) 3027 – 3039.
35. S. Kubo, B. Wunderlich, *J. Polym. Sci. Polym. Phys. Ed.* **10** (1972) 1949 – 1966.
36. S. D. Ross, D. J. Kelley, *J. Appl. Polym. Sci.* **11** (1967) 1209 – 1215.
37. K.-U. Bühler: *Spezialplaste,* Akademie Verlag, Berlin 1978, pp. 157 – 176.
38. C. J. Brown, A. C. Farthing, *J. Chem. Soc.* 1953, 3270 – 3278.
39. M. Tsuji: "Electron Microscopy," in C. Booth, C. Price (eds.): *Polymer Characterization,* vol. 1, in G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science,* Pergamon Press, Oxford 1989, pp. 813 – 821.
40. Company Literature, Union Carbide, New York 1965 ff.
41. M. Spivack, *Rev. Sci. Instrum.* **41** (1970) 1614 – 1616.
42. Dow Chemical, US 4 540 763, 1985 (R. A. Kirchhoff).
43. F. E. Arnold, L. S. Tan, 31st International SAMPE Symposium (April 1986), pp. 968 – 976.
44. Dow Chemical, US 4 711 964, 1987 (L. S. Tan, F. E. Arnold).
45. R. A. Kirchhoff et al., *J. Macromol. Sci. Chem.* **A 28** (1991) 1079.
46. R. A. Kirchhoff, K. Bruza, C. Carriere, N. Rondan, *Makromol. Chem., Macromol. Symp.* **44/45** (1992) 531.
47. H.-G. Elias: Makromoleküle, 5th ed., vol. 1, Hüthig and Wepf, Basle 1990.
48. N. Ise, *Angew. Chem.* **98** (1986) 323 – 333; *Angew. Chem. Int. Ed. Engl.* **25** (1986) 323 – 333.
49. R. Longworth, D. J. Vaughan, *Nature (London)* **218** (1968) 85 – 87.
50. A. Eisenberg, *Macromolecules* **3** (1970) 147 – 154.
51. B. Dreyfus, *Macromolecules* **18** (1985) 284 – 292.
52. K. A. Mauritz, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 28** (1988) 65 – 98.
53. I.G. Farbenindustrie, DE 667 163, 1930 = FR 710 903, 1930, GB 360 822, 1930 (E. Konrad, W. Bock).

54. B. F. Goodrich, US 2 724 707, 1950 = GB 707 425, 1950; US 2 662 874, 1950 = DE 943 259, 1950 (H. P. Brown).
55. H. P. Brown, *Rubber Chem. Technol.* **30** (1957) 1347 – 1386; **36** (1963) 931 – 962.
56. For a list of trade names of carboxylated elastomers up to 1974, see D. K. Jenkins, E. W. Duck, in [62]
57. R. R. Warner, *Rubber Age (N.Y.)* **71** (1952) no. 2, 205 – 221.
58. R. W. Rees, D. J. Vaughan, *ACS Polymer Preprints* **6** (1965) no. 1, 296 – 303.
59. Anon., *Modern Plastics International* **13** (April 1983) 54.
60. J. F. Hagman, J. W. Crary, Ethylene – Acrylic Elastomers, *Encycl. Polym. Sci. Eng.,* 2nd ed., **1** (1985) 325 – 334.
61. R. W. Rees, *Mod. Plast.* **42** (1964) no. 1, 209 – 210; R. W. Rees, D. J. Vaughan, *ACS Polym. Preprints* **6/1** (1965) 287 – 295.
62. L. Holliday (ed.): *Ionic Polymers,* Halstead Press, New York 1975.
63. A. Eisenberg, M. King (eds.): *Ion Containing Polymers,* Academic Press, New York 1977.
64. A. Eisenberg, M. Pineri (eds.): *Structure and Properties of Ionomers,* D. Reidel Publ., Den Haag 1987.
65. E. P. Otocka, *J. Macromol. Sci. Rev. Macromol. Chem.* **C 5** (1971) no. 2, 275 – 294; *Rev. Macromol. Chem.* **6** (1971) 275 – 294.
66. C. R. Lantman, W. J. MacKnight, R. D. Lundberg, *Ann. Rev. Mater. Sci.* **19** (1989) 295 – 317.
67. W. J. MacKnight, T. R. Earnest, Jr., *J. Polym. Sci. Macromol. Revs.* **16** (1981) 41 – 122.
68. R. Longworth in A. D. Wilson, H. J. Prosser (eds.): *Dev. Ionic Polym.* **1** (1983) 53 – 172.
69. W. J. MacKnight, R. D. Lundberg, *Rubber Chem. Technol.* **57** (1984) 652 – 663.
70. M. R. Tant, G. L. Wilkens, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 28** (1988) 1 – 64.
71. J. J. Fitzgerald, R. A. Weiss, *J. Macromol. Sci. Rev. Macromol. Chem. Phys.* **C 28** (1988) 99 – 185.
72. N. L. Zutty, J. A. Faucher, S. Bonotto: "Ionomers," *Encycl. Polym. Sci. Technol.* **6** (1967) 420 – 431.
73. R. D. Lundberg, Ionic Polymers, *Encycl. Polym. Sci. Eng.,* 2nd ed., **8** (1987) 393 – 423.
74. G. W. Lantman, W. J. MacKnight, R. D. Lundberg: "Ionomers," in C. Booth, C. Price (eds.): *Polymer Properties,* vol. 2, G. Allen, J. C. Bevington (eds.): *Comprehensive Polymer Science,* Pergamon Press, Oxford 1989.
75. A. Eisenberg (ed.): *Ions in Polymers* (ACS Adv. Chem. Ser. 187) ACS, Washington, D.C. 1980.
76. A. Eisenberg, F. E. Bailey (eds.): *Coulombic Interactions in Macromolecular Systems (ACS Symp. Ser. 302)* ACS, Washington, D.C. 1986.
77. L. A. Utracki, R. A. Weiss (eds.): *Multiphase Polymers: Blends and Ionomers (ACS Symp. Ser. 395)* ACS, Washington, D.C. 1989.
78. K. Sanui, W. J. MacKnight, R. W. Lenz, *Macromolecules* **7** (1974) 101 – 105.
79. A. Eisenberg, M. Navratil, *Macromolecules* **6** (1973) 604 – 612.
80. R. D. Lundberg, H. S. Makowski, *ACS Polymer Preprints* **19** (1978) no. 2, 287 – 291.
81. Y. Mohajer et al., *Polym. Bull.* **8** (1982) 47 – 54.
82. Du Pont, US 3 264 272, 1966 (R. W. Rees); Du Pont, US 3 404 134, 1968 (R. W. Rees).
83. A. Eisenberg, *Macromolecules* **4** (1971) 125 – 128.
84. H. Matsuura, A. Eisenberg, *J. Polym. Sci. Polym. Phys. Ed.* **14** (1976) 1201 – 1209.
85. H. Longworth in L. Holliday (ed.): *Ionic Polymers,* Wiley-Interscience, New York 1975, pp. 69 – 172.
86. H. S. Makowski, R. D. Lundberg, L. Westerman, J. Bock in [75] pp. 3 – 20; R. D. Lundberg, *Encycl. Polym. Sci. Eng.,* 2nd ed., **8** (1987) 393 – 423.
87. D. J. Vaughan, *Innovation [Du Pont]* **4** (1973) no. 3, 10.
88. M. F. Hoover, G. B. Butler, *J. Polym. Sci. Polym. Symp.* **45** (1974) 1 – 37.
89. A. Eisenberg, H. J. Yeager (eds.): *Perfluorinated Ionomer Membranes* (ACS Symp. Ser. 180) ACS, Washington, D.C. 1982.
90. W. Grot, *Chem. Ing. Tech.* **44** (1972) 167 – 169.
91. G. D. Andrews, R. L. Dawson, *Encycl. Polym. Sci. Eng.,* 2nd ed., **6** (1986) 513 – 522.
92. J. T. Maynard, P. R. Johnson, *Rubber Chem. Technol.* **36** (1963) 963 – 974.

Spices → Flavors and Fragrances
Spices → Seasonings
Spirit of Turpentine → Turpentines

Spirits

DENNIS C. WATSON, Chivas Brothers Limited, Keith, Banffshire, Scotland

1.	**Introduction**	1
2.	**Types of Spirit Drinks**	3
3.	**Typical Production Processes**	7
3.1.	**Cereal-Based Processes**	7
3.1.1.	Malt Whisky Production	7
3.1.2.	Grain Whisky Production	10
3.2.	**Pot-Distilled Rum**	11
3.3.	**Cognac**	12
4.	**World Production Figures**	13
5.	**Taxation**	13
6.	**Legislation**	14
7.	**Analysis**	15
8.	**References**	15

1. Introduction

Spirits are distilled alcoholic beverages made from ferments of plant carbohydrates, with the exception of fermented whey. Their origins are lost in history, but it is reputed that alchemists in the ancient civilizations of China, Egypt, and Greece had knowledge of distillation and may have produced distilled spirits from existing alcoholic ferments such as ales and wines. The knowledge of distillation spread to Western Europe in 1000 – 1100 A.D. from Arabian practitioners, possibly by soldiers returning from the Crusades or via the Moors in Spain. However, spirit production remained mainly in the hands of monks and alchemists until 1400 – 1500. Thereafter, there was a gradual introduction into the general population. In the rural economies of those times, small-scale distillation became widespread, and most landowners and many smaller farmers possessed distillation equipment. The relatively crude spirit was commonly termed aqua vitae, the water of life. Local translation and modification of this name led to the modern terms whisky and vodka. Industrialization brought the commercialization of distilled beverages, while European colonization and increased trading links resulted in the establishment of spirit production in most parts of the world, except where there were legal prohibitions or inadequate supplies of raw materials [1–3].

Early distillation equipment consisted of a vessel in which the alcoholic ferment was heated. The evaporated alcoholic vapor passed to a long tube connected to the head of the heating vessel. Condensation of the spirit occurred as the vapors passed through the cool tube. Originally the condensation tube was air cooled, but this was improved by immersing a coiled tube, known as a worm or a serpentine, in water. This type of still is known as a pot still or an alembic. The essential features are embodied in the traditional Charentaise still (Fig. 1).

Pot stills were normally operated in a batch fashion, but in the early 1800s a number of methods of continuous distillation were introduced. The most successful proved to be the Coffey still, invented by the Irishman, AENEAS COFFEY. His still, outlined diagrammatically in Figure 2, consisted of two columns, the rectifier and the analyzer, which is also known as a beer still. The latter is essentially a stripping column, in which alcohol and other volatiles are continuously evaporated from a feedstock of fermented material, introduced at the top of the column, by a flow of steam from the bottom. The vapors pass to the rectifier, in which the analyzer feedstock is preheated by traveling through a coiled tube. A series of distillations – i.e., condensation on the coiled tube and revaporization – occurs up the length of the second column. The least volatile compounds are most easily condensed while the most volatile are rapidly revaporized, resulting in a gradient up the column equating to increased volatility. The spirit product, consisting mainly of ethanol, is drawn off the column near the top. This series of "purifying" distillations is known as rectification, from which the name of the column derives. The principles of the Coffey still have been further refined through the use of sieve and bubble-cap plates [3–5].

Continuous distillation produced lighter-flavored spirits and was more economical than

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_551

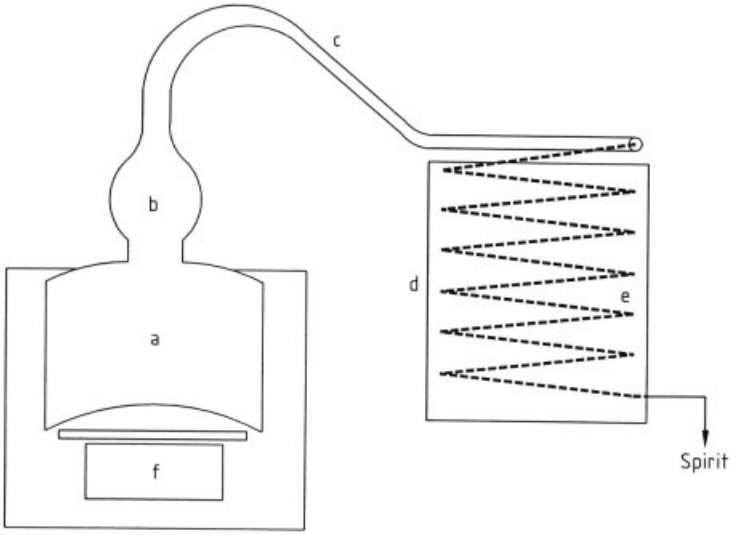

Figure 1. Schematic of a traditional Charentaise alembic still
a) Still body; b) Still head; c) Swan neck; d) Condenser water tank; e) Serpentine condenser; f) Coal-, wood-, or gas-fired furnace

pot distillation. The latter is now used principally in the production of heavily flavored spirits such as Scotch malt whisky, Cognac, and certain rums. In a number of cases, attempts have been made to protect the traditional character of certain spirits by restricting the maximum alcohol concentration of the distillate, thus prohibiting excessive rectification (e.g., Bourbon whiskey). Such products are usually made by single- column continuous or pot distillation [2], [3], [5].

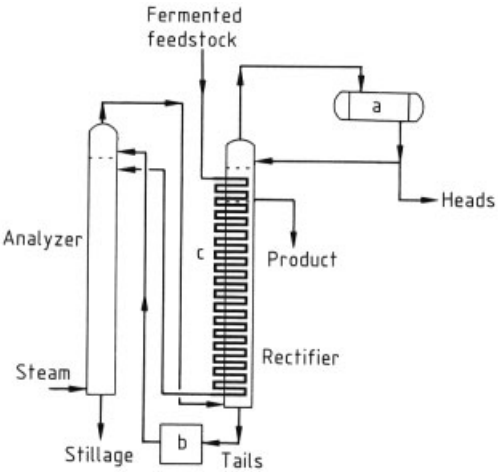

Figure 2. Flow diagram of a Coffey still (adapted from internal publication of Joseph E. Seagram & Sons, Inc.)
a) Condenser; b) Fusel oil decanter; c) Wash preheating coils

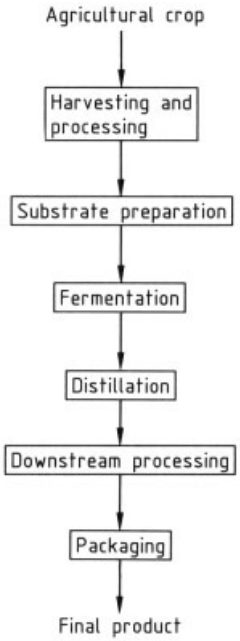

Figure 3. Flow sheet of major steps in distilled alcoholic beverage production

Besides the engineering developments, the range of raw materials used was extended as knowledge of spirit production spread. Any source of carbohydrate which could be economically grown was utilized. Regional spirit

types developed to reflect the abundance of local crops, and gradually the modern categories of products were established. All processes have the basic steps outlined in Figure 3. Variations at each stage have led to the differences between and within categories, resulting in the wide array of spirit drinks available worldwide. Most are consumed in their area of production, but a few are distributed internationally, such as Cognacs and Scotch Whisky [4], [6].

2. Types of Spirit Drinks

There are two distinct divisions in the distilled alcoholic beverages. In the first, the product is characterized by flavors emanating from the raw materials and subsequent fermentation, even though some rectification may occur during distillation. The spirit is often further modified by aging, termed maturation, in wooden casks. This division includes categories such as whiskies, rums, wine spirits, brandies, and tequila. The second group contains products which gain their main characteristics from downstream processing, such as flavor addition. Although the spirit base may be one from the first grouping, highly-rectified neutral spirits are frequently used. The individual products are categorized by the exact nature of the additional processing. This division includes gin, aquavit, liqueurs, and anis.

Throughout the world, there are numerous types of local spirits, which have similarities to the major categories described below. There is also some variation in local definitions and terminology; therefore the following sections mainly follow the conventions of the EC Spirit Drink Regulations [7].

These regulations define a spirit drink as:

1) An alcoholic liquid intended for human consumption
2) Having particular organoleptic properties and a minimum alcoholic strength of 15 vol %, except for egg liqueurs/advocaat/avocat/Advokat which are allowed a minimum of 14 vol %
3) Produced
   a) either directly by the distillation, with or without added flavorings, of natural fermented products and/or by the maceration of vegetable substances and/or the addition of flavorings, sugars or other sweetening products (listed in the regulations) and/or other agricultural products to ethyl alcohol of agricultural origin and/or to distillate of agricultural origin and/or to spirit (as defined in the regulations)
   b) or by the mixture of a spirit drink with one or more other spirit drinks; ethyl alcohol of agricultural origin, distillate of agricultural origin or spirit; one or more alcoholic drinks; or one or more drinks

Whiskies. Whisky, or whiskey, as it is spelled in Ireland and the United States, is a distilled alcoholic beverage, made from a mash of cereals fermented by yeast, and which has been matured in wooden casks. The best known whiskies are those made in Scotland, Ireland, the United States, Canada, and Japan, although local whiskies are made in many cereal-growing regions. The EC definition stipulates a maturation period of at least three years in wooden casks not exceeding 700 L in capacity. Distillation must be carried out at less than 94.8 vol %, whereas in the United States the maximum is 95 vol %.

The UK definition for Scotch Whisky restricts the wooden casks to oak and prohibits the use of added enzymes, which are permitted elsewhere. Malted barley, to which only the whole grains of other cereals may be added, must be included in the mash [8]. Scotch malt whisky is made from a mash of malted barley and is pot-distilled, whereas the mash for Scotch grain whisky includes ca. 85 % of unmalted cereal, usually wheat or maize, and is column distilled. The individual processes are described in greater detail in Chapter 3. Although numerous malt whiskies and a few grain whiskies are marketed, over 95 % of all Scotch whisky sales are of blended whisky, that is, a mixture of malt and grain whiskies.

Japanese and Irish whiskies are made in a similar way to Scotch, although the Irish make some pot-distilled product from a mash that includes a proportion of unmalted cereals. There are no restrictions on the use of added enzymes. North American whiskies are made from mashes in which unmalted cereals predominate and they are column-distilled. In the United States, bourbon and rye whiskies are made from fermented mashes of not less than 51 % maize and 51 %

rye, respectively. These whiskies must be distilled at an alcoholic strength ≤ 80 vol %, thus restricting the degree of rectification, and stored at not more than 62.5 vol % alcohol in charred new oak casks. If they are aged for at least two years, they may be termed straight whiskey, as are other whiskies made under the same restrictions from wheat, malted barley, and malted rye. American blended whiskey contains ≥ 20 vol % of straight whiskey (calculated at 50 vol % alcohol) blended with other whiskey and/or grain neutral spirits. The latter are distilled at or above 95 vol % alcohol [2], [3], [9].

Grain spirit is an EC designation for a type of spirit drink produced by the distillation of a fermented mash of cereals and having organoleptic characteristics derived from the raw materials used. In Germany and other German-speaking, traditional producing regions, the term "Korn" or "Kornbrand" may be used if the grain spirit is obtained without any additive. The designation "grain brandy" may be applied, if the grain spirit is distilled at < 95 vol % alcohol [7]. The flavor of individual products is influenced by the actual cereals in the mash and they have some resemblance to unaged whisky spirits.

Rum is a spirit produced by distilling fermented sugar cane juice, syrup, or molasses. Some countries, such as Finland, allow the use of fermented beet molasses, but this is not generally recognized. Rum-like products may go under other names, such as "cane" in South Africa and "cachaca" in Brazil. Production is mainly centered in the Caribbean and other sugar cane growing areas. In some developing countries, local molasses spirits are sold as "whiskies" and "brandies" rather than their true designation as rum.

Light-bodied rums are distilled on multicolumn continuous distillation systems, whereas heavy-bodied rums are produced by using pot stills, particularly in Jamaica and Guyana. Jamaican law prohibits the manufacture of rum other than as part of a sugar refinery complex. The maximum distillation strength in EC regulations is 96 vol % alcohol, whereas the United States has a maximum of 95 vol %. There is no general requirement for maturation, although many rums do undergo aging, in which case the cask size should not exceed 700 L nominal capacity. Products marketed as "dark rums" have normally been aged, but some of their color may arise from the addition of caramel [2], [3], [10].

Wine Spirit. The EC designation "wine spirit" refers to a spirit drink produced by the distillation of wine or redistillation of a wine distillate at a strength of < 86 vol % alcohol. The distillate should contain a quantity of volatile substances, other than ethanol and methanol, ≥ 125 g per hectoliter of absolute alcohol. If the product is aged, it can continue to be termed "wine spirit" but must meet the aging regulations stipulated for brandies.

This category includes unaged products of the eau-de-vie de vin type. However, also included are matured spirits which have, until now, been more usually called brandies, such as cognacs and armagnacs. The latter are produced by single column distillation, although a few distillers use pot stills, similar to the Charentaise alembic (Fig. 1) [6], [7].

Brandy is a term used widely to describe distillates made from the fermentation of fruits or their musts. It is normally prefixed by the fruit name but, in the case of grapes, is used alone. In the United States, all fruit brandies must be distilled at < 95 vol % alcohol and those from grapes at < 85 vol %. If the latter limit is exceeded, the spirit is termed neutral brandy. US law stipulates a minimum aging period of two years in oak casks [2].

The EC designation for "brandy" and its German equivalent "Weinbrand" is restricted to grape distillates and is complimentary to the wine spirit designation. The spirit drink is produced from wine spirit, whether or not it is blended with wine distillate distilled at < 94.8 vol % alcohol, provided that the latter does not exceed 50 % of the final product. Aging should take place for a minimum of six months in oak casks with a capacity of < 1000 L or for a minimum of one year in larger oak receptacles. Cognacs and armagnacs also meet this definition. The distinction is that brandies, containing the lighter, double- or multiple- column distillates, cannot be defined as wine spirits [7].

Fruit Spirits. The products which are known in many countries as fruit brandies,

excluding those from grapes, are included in this EC designation. Distillation must be at < 86 vol % alcohol, similar to wine spirits, but the volatile substances must equal or exceed 200 g per hectoliter of absolute alcohol. Fruit spirits may also be produced by macerating partly fermented or unfermented fruits in neutral spirit or some other distillate at not less than 100 kilograms of fruit per 20 L of absolute alcohol, followed by distillation. This category includes well-known drinks, such as Kirschwasser, and numerous regional "eau-de-vies" [7], [11].

Cider Spirit, Cider Brandy, and Perry Spirit. This category includes spirit drinks which meet the requirements for fruit spirits and are made exclusively by the distillation of cider or perry, ferments of apple and pear, respectively. The best-known product is probably calvados.

Grape Marc Spirit. The debris left in the grape press can be fermented and then distilled to produce a product known as grape marc or grape marc spirit. The designation "grappa" may be used for such products made in Italy. The EC limits distillation to < 86 vol % alcohol and stipulates that the volatile substances, other than ethanol and methanol, should be equal to or exceed 140 g per hectoliter of absolute alcohol.

Fruit Marc Spirits. Fruit marc spirits are made in a similar way to grape marc spirit by the fermentation and subsequent distillation of fruit debris or marc. The maximum distillation strength is the same, but the minimum volatile substances content is 200 g per hectoliter of absolute alcohol.

Tequila is made in the Mexican region of that name from a fermented substrate derived from the heads of the agave cactus, *Agave tequilana* Weber. The cactus is rich in the polysaccharide inulin which is broken down on heating to yield fructose monomers. The substrate may be supplemented with sugar up to 30 % of the total. Distillation is by two-stage pot distillation, but the final product may contain up to 50 % neutral spirits. There is no legal maturation requirement, but many tequilas are aged in oak casks [2], [3].

Vodka. The original vodka of Russia and Eastern Europe was essentially the local version of "aqua vitae" made from any readily-fermentable material and distilled in pot stills. With time, the process developed to include charcoal filtration to remove "impurities", and in the 1800s continuous column distillation was introduced. Although more refined, many vodkas from this region retain some flavors from the raw materials and the fermentation, and many contain added flavors [6].

The vodka of Western Europe and North America is made from neutral spirits, i.e., highly-rectified spirit produced by multicolumn continuous distillation. In the United States, the minimum distillation limit is 95 vol % alcohol and the spirit must be further treated with charcoal or activated carbon or redistilled so that it has no character, aroma or taste other than ethanol, before it can be termed vodka. If pronounced flavorings are added, they must be included in the product description. The EC regulations are similar. Nonpronounced flavorings may be used to give a more mellow taste. The neutral spirits must, however, be distilled at a higher strength and meet the EC specification.

The EC specification for neutral spirit (ethyl alcohol of agricultural origin) [7] is as follows:

1)	Organoleptic characteristics	no detectable taste other than that of raw material
2)	Minimum alcoholic strength	96.0 vol %
3)	Maximum level of residues (in grams per hectoliter of 100 vol % alcohol)	
	Total acidity, as acetic acid	1.5
	Esters, as ethyl acetate	1.3
	Aldehydes, as acetaldehyde	0.5
	Higher alcohols, as 2-methyl-1-propanol	0.5
	Methanol	50
	Dry extract	1.5
	Volatile bases containing nitrogen, as nitrogen	0.1
	Furfural	not detectable

The methanol limit may, in future, be reduced to 20 grams per hectoliter of 100 vol % alcohol.

The raw material can be any source of fermentable substrate, provided it is of agricultural origin (e.g., cereals, potatoes, cane and beet molasses, fruits, and whey, a byproduct of the dairy industry). Some countries, such as Finland, restrict the raw materials to those traditionally used in the original vodka-producing region, namely cereals and potatoes [2], [7], [12].

Fruit Spirit Drinks. Although similar in name to fruit spirits, this EC designation applies to beverages made by macerating fruit in neutral alcohol or other spirits without further distillation. Other flavors may be added but the characteristic taste and color of the drink must come from the fruit used. The minimum proportion of fruit is 5 kg per 20 L of absolute alcohol [7], [11].

Juniper-Flavored Spirit Drinks. A number of distilled beverages are produced by flavoring neutral spirit and/or grain spirit and/or grain distillate with juniper (*Juniperus communis*) berries. The most internationally famous are gin and genever, although there are many localized products, such as German "Steinhäger" (Steinhagen, Gütersloh). Other approved flavorings may be used, but the organoleptic characteristics of juniper should be discernable, except in the case of genever. In contrast, gin is required to have a predominant juniper taste. The term "distilled gin" is restricted to products obtained by redistilling neutral spirit over or with juniper berries and other botanicals and flavorings. Gin stills are modified pot stills. In some, the botanicals are held in perforated trays in the head of the still through which the alcoholic vapors pass, whereas in others the botanicals are loaded into the body of the still with the neutral spirit. "London gin" is a type of distilled gin rather than a geographical designation. There is no general requirement for aging, but some gins are matured in wooden casks, particularly some traditional genever brands and some premium gins [6], [7], [12].

Caraway-Flavored Spirit Drinks. Caraway (*Carum carvi*) is widely used to flavor spirit drinks, particularly in Northern Europe, such that the EC has a separate category for such beverages. Internationally, Danish "akvavit" is the best-known. However, in Scandinavia other drinks carry this name, or its synonym "aquavit", but are not flavored with caraway, and the term should be regarded as covering a range of flavored and unflavored distillates akin to Russian vodkas.

The EC designation stipulates that for a caraway-flavored spirit drink to carry the akvavit or aquavit name, they must be flavored by a distillate of plants or spices; that is, essential oils are prohibited. The characteristic flavor must be attributed to distillates of caraway and/or dill (*Anethum graveolens* L.) seeds [6], [7].

Aniseed-Flavored Spirit Drinks. Aniseed-flavored spirits are produced by flavoring distillates, particularly neutral spirit, with natural extracts of star anise (*Illicium verum*), anis (*Pimpinella anisum*), fennel (*Foeniculum vulgare*), or any other plant which imparts primarily the characteristic aniseed aroma and taste. The beverage may be produced by (1) maceration or distillation, (2) redistillation of the alcohol in the presence of the aniseed-flavored seeds or plants, (3) addition of distilled extracts of the plants, or (4) a combination of these methods. Although other seeds and natural plant extracts may be used, the drinks should have a predominant aniseed taste.

The characteristic flavor of products called "anis" must be derived exclusively from anise, and/or star anise, and/or fennel. The term "distilled anis" may be used if at least 20 % of the product's alcohol has been distilled in the presence of the seeds.

Aniseed-flavored drinks called "pastis" must also contain natural extracts of liquorice root (*Glycyrrhiza glabra*). They, therefore, contain colorants known as chalcones as well as glycyrrhizic acid, which has an EC limit of 0.05 – 0.5 g/L. The EC also limits sugar to < 100 g/L and anethole to 1.5 – 2 g/L.

"Ouzo" is an aniseed-flavored spirit drink produced exclusively in Greece by blending alcohols flavored by distillation or maceration of aniseed and possibly fennel seed, mastic from a lentiscus indigenous to the island of Chios (*Pistacia lentiscus Chia* or *latifolia*), and other seeds, plants, and fruits. The alcohol flavored by distillation must represent at least 20 % of the product's alcohol and the distillation must be carried out in traditional discontinuous copper stills of ≤ 1000 L capacity at 55 – 80 vol % alcohol. The product must be colorless and have a sugar content of ≤ 50 g/L [6], [7].

Arrack and **raki** are terms used throughout the Middle East, the Balkans, and southeast Asia to describe a wide range of local distilled beverages made from various raw materials, such as grapes, dates, palm sap, figs, molasses, rice, and other cereals. It is often flavored with aniseed and has some resemblance to anis.

Unflavored products could be classified in other categories; for example, molasses-based arracks are akin to rums [6].

Bitters. Bitter-tasting spirit drinks are produced by flavoring neutral spirit with approved flavoring such that the product has a predominant bitter taste. They are most widely known as pick-me-ups or hangover cures, but they are also consumed as aperitifs diluted with ice and soda, or neat as digestifs. They may also be included in cocktails; for example Angostura bitters is an essential ingredient of pink gin. The alternative name "amer" is also used [6], [7].

Liqueurs and Cordials. Cordial is the American term for products called liqueurs elsewhere. They are made by flavoring and sweetening neutral spirit or a spirit drink or a combination thereof and may also contain added agricultural products, such as cream, milk, egg, fruit, wine, or flavored wine. In the United States, the minimum sugar content is 2.5 wt % and the product may be designated "dry" if the level is less than 10 wt %. The EC regulations stipulate a minimum of 100 g/L expressed as invert sugar and restrict the use of the term "crème de" followed by the name of a fruit or the raw material used, excluding milk products, to liqueurs containing at least 250 g/L of sugar. The blackcurrant liqueurs named "crème de cassis" must contain at least 400 g/L of sugar. Gentian liqueurs, prepared with gentian as the sole aromatic substance, and cherry liqueurs, made solely with alcohol which is cherry spirit, are allowed reduced minimum sugar contents of 80 g/L and 70 g/L, respectively.

Egg liqueur, commonly known as advocaat or similar names, is regulated in the EC to contain egg yolk, egg white, and sugar or honey. The egg yolk content must exceed 140 g/L and the sugar or honey must be at least 150 g/L. Similar products of lower egg yolk content, to a minimum of 70 g/L, are termed liqueur with egg [2], [6], [7].

3. Typical Production Processes

There is such a vast range of process variations in the distilled alcoholic beverage industry that it is not possible to effectively cover them here. Instead, a few examples are described to illustrate the principles used in some of the major categories.

3.1. Cereal-Based Processes

Cereals are the major raw materials used to produce international categories of spirit drinks, such as whiskies, many more localized categories, such as Korn, and a large fraction of the neutral spirits or cruder distillates which are further processed, by flavoring or otherwise, into numerous categories of drinks (e.g., vodka, gin, etc.). There are two distinct classes of distillery processes typified by those used in the Scotch whisky industry.

3.1.1. Malt Whisky Production

Sugars are present in cereal seeds in the form of the storage polysaccharide, starch. The latter is not readily fermented and must be broken down into more easily utilizable subunits. The traditional method to achieve this was by the process known as malting, which was most successfully achieved with barley. Malting involves three steps: steeping, germination, and kilning. During steeping, water is added, usually in stages with air rests between, until the moisture content is ca. 45 %. This rehydrates the grain, particularly the endosperm area in which the starch is stored in granules, and stimulates the embryo plant to commence germination. The embryo secretes gibberellic acid, a plant hormone, which induces a layer of cells, known as the aleurone layer, to excrete a range of hydrolytic enzymes into the endosperm. Some enzymes are also secreted by the embryo portion of the grain.

The initial enzyme actions are of a cytolytic function, in which glucanase enzymes digest the glucan cell walls of the endosperm and protease enzymes hydrolyze the protein matrix contained therein, releasing the enclosed starch granules. Carbohydrase enzymes, of which α-amylase, β-amylase, and limit dextrinase are the most important, start to hydrolyze the starch. Nutrients released by the enzymes are used for embryo growth. The maltster must balance the breakdown of the endosperm, known as modification,

against the loss of potentially fermentable carbohydrate.

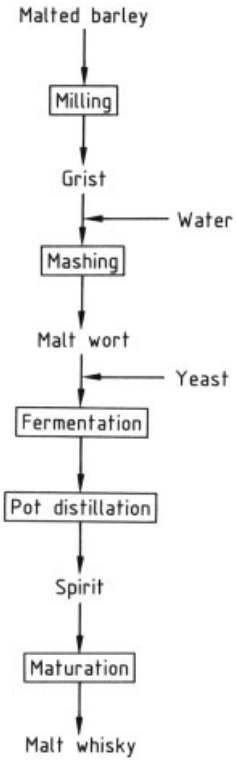

Figure 4. Flow sheet of Scotch malt whisky process

When modification is judged to be sufficient, germination is arrested by drying or kilning the barley gently in a stream of warm air. Traditionally, the heat for kilning in many parts of Scotland was obtained from peat fires. On burning, peat releases a phenolic smoke which impregnates the grains as they dry. Many of these volatiles pass through subsequent processing stages to impart a phenolic or peaty taste to the whisky. To maintain this characteristic, many malted barleys are still peated, although the peat smoke may be introduced into the drying air stream from an ancillary furnace rather than being the main source of heat. The malted barley is dried to 3.5 – 5 % moisture. The enzymes are inactive at this level, although some of the most heat-labile are largely destroyed by the heat of kilning. Besides killing the embryo and suspending enzyme activity, the malted barley is also prepared for easy storage and subsequent milling [13–17].

The major steps of the malt whisky process after malting are outlined in Figure 4. Prior to milling, dust, stones, metal, and other extraneous materials are removed. The mills are most frequently of the 4-roller type, but modern plants have 5- and 6-roller mills. The malted barley is ground to produce a grist, which contains about 10 % fine flour and 20 % coarse husk fractions, with the rest as medium-sized fragments. Care must be taken that the milling is neither too fine nor too coarse, as both can result in poor mashing performance, by affecting, respectively, malt wort run-off or starch extraction and hydrolysis.

The grist is made into a mash by mixing with hot water, to give an initial temperature of ca. 65 °C, in a mixing apparatus as it is discharged into the mash tun. Sometimes known as a lauter, this vessel has a false bottom, with apertures about 0.5 mm wide, through which the solid-free malt wort is separated from the rest of the mash. After discharge into the tun, the mash is mixed. By this time, the starch has been liquified (gelatinized) and thereafter is rapidly hydrolyzed by the amylase enzymes. At the temperature of the mash, limit dextrinase and, to a lesser extent, β-amylase are inactivated within a short period of time, as are many of the other barley enzymes, such as the proteases and glucanases. However, to maximize conversion, the mash is normally allowed to sit for 20 – 60 min. During this time, the husk settles on the false bottom to form a filtration medium through which the malt wort is drawn, thus clarifying it. The worts are drained by gravity into a balancing tank, from whence they are pumped via a heat exchanger to the fermentation vessels, termed washbacks. Further quantities of water, named sparges, of higher temperature are added to the mash batchwise or by continuous spraying. Only the strongest of these extracts is pumped to the fermenters, while the rest are collected for recycle in the next mash [16], [18].

Although the use of added microbial enzymes is prohibited in the Scotch whisky process, this is not the case in other countries. Added amylases and/or glucanases can aid conversion and wort run-off, especially if poorer quality, less well modified malted barley is used. Similar additions, usually of glucoamylase, to the fermentation stage can help maximize yields. In Ireland, similar styles of mashing are sometimes

carried out with malted barley and unmalted cereals, most usually barley or oats.

The traditional fermentation vessels are wooden vats with removable, segmented lids and slightly-angled bottoms to ease discharge, but modern distilleries have totally-enclosed, cylindrical, stainless-steel vessels with conical bottoms. They are not attemperated in Scotland, but in warmer climates this can be a necessity. In the uncontrolled system, the initial temperature is ca. 21 – 22 °C which rises to about 33 °C due to the heat generated during fermentation [19–21].

The yeast is added during washback filling, once the temperature has stabilized. Specialized distilling strains of *Saccharomyces cerevisiae* are used, singly or in combination, frequently supplemented with waste brewery yeasts [22], [23]. The sugar content of the initial wort is 120 – 150 g/L, with the main components being maltose ca. 55 %, maltotriose ca. 15 %, glucose ca. 8 %, fructose ca. 7 %, sucrose ca. 2 %, and the rest as oligosaccharides and minor sugars. The main nitrogen source is amino acids from barley protein hydrolysis. The initial free amino nitrogen concentration is ca. 0.3 g/L. The fermentation is relatively rapid due to a combination of the characteristics of the yeast strains, a relatively high inoculation level (10 – 20 million yeast cells/mL of wort), the wort composition, and the moderately high temperature. The main yeast fermentation is normally complete in ca. 40 – 48 h [20], [21].

The fermented wort is called wash, although the term beer is used in some countries. It has an alcoholic strength of ca. 7 – 7.5 vol %. It undergoes pot distillation, in copper stills, of a similar general design to that illustrated in Figure 5, although some distilleries use traditional worm condensers. The wash is first distilled to produce a distillate of ca. 18 – 20 vol % alcohol, known as low wines. A second distillation takes place from which three fractions are collected; the first runnings (foreshots or heads), the main spirit product, and the last runnings (tails or feints). The heads and tails are collected and mixed with the low wines to charge the next distillation. The middle fraction is filled into oak casks for maturation into whisky. In Ireland and a few Scottish distilleries, triple rather than double-distillation is practised, with similar recycling of heads and tails fractions [5], [24].

A major aim of the malt whisky process is to produce desirable flavors as well as ethanol. The new distillate contains alcohols, esters, carbonyl compounds, acids, sulfur compounds, phenolics when peated malt is used, and traces of nitrogen compounds. A few, such as the peaty flavors, may emanate from the raw materials, but the majority, or their precursors, are formed during fermentation by the action of yeast and other microbes which may be present [25], [26]. It is now known that many of the reactions, such as esterification, continue during distillation, often catalyzed by the copper of the still construction [27]. They also continue during maturation, but at a slower rate. However, many other changes occur as the spirit ages. Lignin and hemicellulose breakdown products, such as vanillin and sugars, are extracted by ethanol. Oxygen permeating through the cask pores participates in oxidation reactions, but there is also a loss of ethanol by evaporation amounting to ca. 2 % of the contents per year in cool climates, such as in Scotland, and higher in warmer environments. Besides the chemical reactions, various physical changes take place, such as alterations in solubilities and the formation of chemical complexes. All of the reactions, extractions, and interactions combine to convert the relatively harsh, new spirit into the more mellow and smooth, matured product [28].

3.1.2. Grain Whisky Production

The basic steps of the Scotch grain whisky process, as illustrated in Figure 6, differ in a number of ways from malt whisky production. The majority of the cereal mash is composed of unmalted cereals, usually maize or wheat, but in similar processes elsewhere, barley or rye. These are cooked in batch or continuous cookers at atmospheric or elevated pressures and at temperatures of 100 °C or higher to gelatinize the starch. Traditionally the unmalted grains were milled, but this is not necessary if pressure cooking is followed by sudden release to atmospheric pressure, which results in the grains' "exploding". The gelatinized mash is cooled to 63 – 65 °C and mixed with malted barley grist of high amylolytic enzyme content. Some distillers use green malt (i.e., unkilned malted barley), which

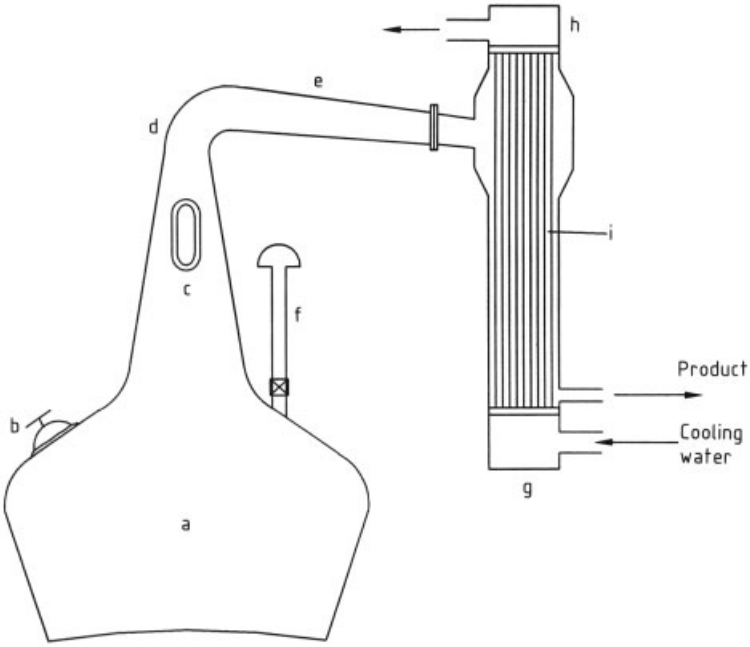

Figure 5. Schematic of Scotch malt whisky pot still
a) Still body; b) Access door; c) Sight glass; d) Still head; e) Lyne arm; f) Vacuum-release vent; g) Tube condenser; h) Water jacket; i) Condensing tubes

retains a higher proportion of enzymes, due to the lack of heat destruction in the kiln.

Although a few distillers continue to lauter the mash to extract a solid free wort, the majority now discharge the whole mash into the fermenters [9], [18], [29]. The fermentation sugars are of a similar composition to malt wort, although more concentrated, but the fermentation rate tends to be slower due to lower concentrations of assimilable nitrogen. The fermented wash (10 – 12 vol % alcohol) is continuously distilled in two- column stills, similar to the Coffey still (Fig. 2), or in more refined multicolumn systems. Many of the secondary, volatile compounds are rectified out, so that the new spirit is much lighter than its malt whisky counterpart. With both types, the maturation casks are normally of Spanish or American oak [9].

In other countries, microbial enzymes are used instead of malted barley to achieve the conversion of starch to fermentable sugars, although 2 – 3 % malt may be retained for flavor purposes, compared with the usual level of 12 – 15 % used in Scotland. Heat-stable bacterial α-amylases may be added as the mash is heated in the cooker, to aid liquification. Fungal α-amylases and glucoamylases are added on cooling to achieve conversion. In some cases, the latter is added just prior to, or at the time of discharge into the fermenters, in which most of the starch conversion occurs [2], [3]. Many of the commercial enzymes contain carbohydrases other than glucoamylase or α-amylase and may also have protease activity [29]. The process is essentially the same in the production of grain neutral spirits, except that the spirit is highly rectified in multicolumn stills. Other starch sources, such as potatoes, can be processed in a similar manner [12].

More highly-flavored grain spirits, such as bourbon and other American straight whiskies, are produced by using single- column stills, termed beer stills, often followed by redistillation in a continuously fed pot still, called a doubler (Fig. 7). Wash, or beer as it is termed in the United States, is fed into the still at the top of the stripping section, usually consisting of about 14 plates. The vapors pass through an entrainment plate that prevents the rise of solids, into a rectifying section of 3 or 4 plates. The doubler

residues are recycled into the beer still [2], [3], [30].

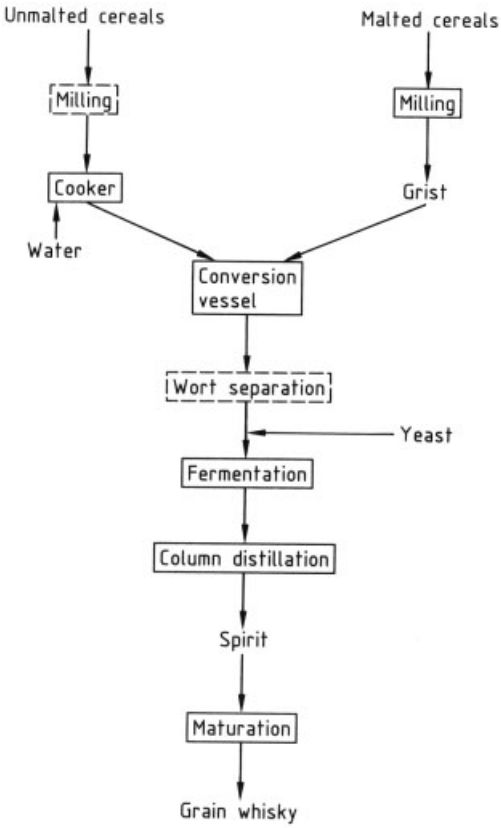

Figure 6. Flow sheet of Scotch grain whisky process
Steps within broken lines are not universal.

3.2. Pot-Distilled Rum

Continuous distillation is the prevalent process in rum manufacture. However, traditional, pot distillation is used especially in the English and French speaking West Indies, to produce highly-flavored distillates, which may be sold as such, but are most commonly used in blending. The main flavor components are esters and the rums are traded according to their ester contents. Some of the most strongly flavored rums are produced in Jamaica, where the addition of limings, a scum from molasses, and dunder, the residue after molasses wash distillation, are added to the fermentation. These materials support a population of bacteria and wild yeasts which produce higher concentrations of esters or their precursors during storage and subsequent fermentation compared with standard distilling yeasts.

The traditional pot distillation system (Fig. 8) consists of a series of vessels. The vapors from the main pot pass through the others, often termed retorts, in which condensation and redistillation occur. Tails and heads from the final condensation are recycled to the first and second retorts, respectively. This system concentrates the esters. In Guyana, some of the pot still bodies are made of wood, although the heads and condensers are of copper, which catalyzes reactions such as esterification [6], [10], [31].

3.3. Cognac

Traditional spirits with geographic designations often have national rules governing the production processes. True cognac can only be produced in the Charente and Charente-Maritime departments as well as two small zones in the Deux Sèvres and Dordognes departments of France. For sub-appellation purposes and distillate classification, there are six distinct production areas: Grande Champagne, Petite Champagne, Borderies, Fins Bois, Bons Bois, and Bois Ordinaires. The grapes are white varieties, of which at least 90 % should be Ugni Blanc, Folle Blanche, and Columbard, while the rest can be from Blanc Ramé, Jurançon Blanc, Montils, Sémillon, and Sélect. The continuous Archimedes press is prohibited and no addition of sugar to the grape must is allowed.

The traditional fermentation was by natural microflora, but cultivated yeasts, sold in a dried form, are now frequently used. Distillation must be by double-distillation in the directly fired Charentaise still, with a total capacity of ≤ 30 hL with a maximum working volume of 25 hL. In addition to the basic outline in Figure 1, many stills also have a wine heater through which the pipe from the swan's neck to the serpentine passes, thus preheating the next charge of wine. In the first distillation, a heads fraction is collected for recycle. In the second, heads and tails fractions are recycled. There is some variation in the recycle practices between distillers, but there is a restriction on the maximum strength of distillation, which must not exceed 72 vol % alcohol. All distillation must be complete by the

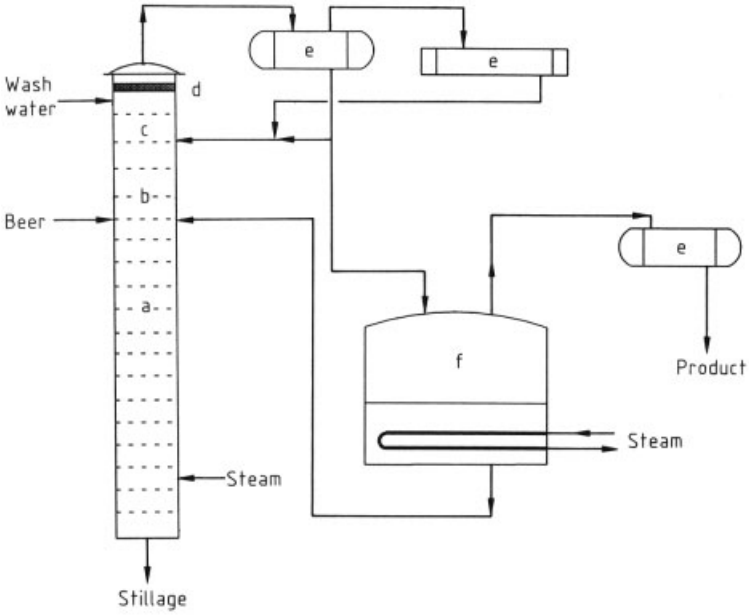

Figure 7. Schematic flow diagram of bourbon beer still and doubler (adapted from internal publication of Joseph E. Seagram & Sons, Inc.). a) Beer still stripping section; b) Entrainment plate; c) Rectifying section; d) Copper demister; e) Condensers; f) Doubler

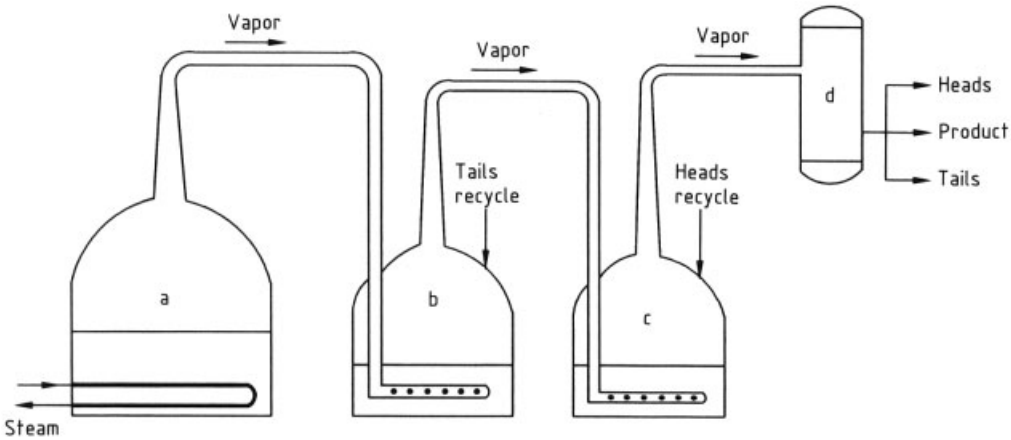

Figure 8. Schematic flow diagram of Jamaican pot still system for the production of highly-flavored rums (adapted from internal publication of Joseph E. Seagram & Sons, Inc.)
a) Wash still; b) First retort; c) Second retort; d) Condenser

31st March after harvest, restricting the distillation period to about 5 months.

Maturation must be carried out in Limousin or Tronçais oak casks. As with other brandies and matured wine spirits, maturation was aided by the use of oak extracts, but this practice is becoming less common in Cognac. The French regulations prohibit the use of additives in the final product, other than water and sugar or caramel and restrict the minimum strength to 40 vol % alcohol [6], [32].

4. World Production Figures

It is difficult to be exact in estimating the quantitiy of spirit drinks produced throughout the

world, because of the prolific number of minor categories and local products, as well as the general lack of data for many markets. In addition, there is a time lag between new spirit distillation and final product packaging with aged drinks. The consumption figures listed below are compiled within these limitations. On this basis, the total world production of distilled spirit was ca. 5.8×10^9 L of 100 vol % alcohol in 1992.

Estimated spirit drinks consumption in 1992 (in 10^6 L 100 % alcohol) was as follows:

Whiskies	
Scotch	275
Canadian	90
Japanese	90
US including bourbon	108
Irish	8
Others	68
Gins/Geneva	
Geneva	18
Gin	122
Rums	
Dark	60
Light	160
Vodka	830
Korn	32
Aquavit	6
Tequila	25
Wine Spirit/Brandy	
Cognac	40
Brandy	220
Eau de vie	25
Other miscellaneous categories	450
Local Spirits	
China	1300
Korea	420
Japan	230
South America	630
South East Asia	360
Africa and others	240

5. Taxation

Alcoholic beverages are taxed in many countries as a source of Government revenue, via excise taxes, and in some cases value-added and other taxes. Spirits are frequently taxed at higher levels than competing products, such as beers and wines. For example, in the United States, excise tax levels (1st January 1991) set the rate for spirits at $ 13.50 per US proof gallon (approximately $ 7.14 per liter of absolute alcohol), whereas the rate for wine was $ 1.07 per wine gallon (approximately $ 2.36 per liter of absolute alcohol for a wine of 12 vol % alcohol). In addition, in many countries imported spirits may face discriminatory taxation intended to protect domestic spirits and/or wines, or simply to minimize importation of luxury goods.

In the run-up to the EC Single Market, efforts were made to harmonize the excise duty throughout the European Community. However, at the start of 1993, the agreed levels were a compromise, rather than true harmonization. Wine had a zero minimum rating, whereas the minimum rate for beer was ECU 187. A minimum rate of ECU 550 was set for spirits, but existing rates between ECU 550 and ECU 1000 could not be reduced, whilst those above ECU 1000 could not be lowered below ECU 1000. The rates in the individual member states are listed in Table 1, highlighting the wide variation throughout the Community. There is a provision that the minimum rates will be reviewed at two-year intervals, but an unanimous decision is required to implement any change. It is therefore unlikely that the rates will be harmonized, at least in the short-term, on a basis of equality of taxation per unit of alcohol.

Table 1. Excise taxation rates in EC countries (1st January 1993) in ECU per hectoliter of 100 vol % alcohol

Country	Beer	Still Wine	Spirits
Belgium	328	318	1510
Denmark	1054	1253	3973
France	56	29	1124
Germany	130	0	1247
Greece	184	0	600
Ireland	2262	2422	2623
Italy	439	0	606
Luxembourg	100	130	904
Netherlands	389	327	1381
Portugal	273	0	682
Spain	78	0	564
United Kingdom	1555	1607	2780

6. Legislation

Legislative requirements vary and only the most basic guidance is given here. In many countries alcoholic beverages are included with other food and drink and must meet local food laws. There is a growing need for a quality system based on HACCP (Hazard Analysis and Critical Control Points), organized, documented, and managed to ISO 9000 standard. Regional and national definitions must also be met. Some of the

requirements are outlined in Chapter 2, but definitions also extend to packaging and labeling. For example, minimum strengths for individual categories are laid down in the EC Spirit Drinks Regulations [7], but national laws may stipulate higher values. For example, in France the minimum strength for cognac is 40 vol % alcohol, whereas the EC minimum for wine spirits is 37.5 vol %. Besides the consumer protection of weights and measures legislation, there are often prescribed bar measures and packaging volumes (e.g., the standard bottle size in the EC is now 70 cL) but a range of other volumes are also prescribed.

Labeling must be such that correct category names are used and that other information is correct. Stated age is usually that of the youngest spirit, although exceptions may be made where traditional solera systems are used, in which blending of old and young spirits is carried out during aging. Specific terms, associated with aged spirits, may only be used if the product complies with the appropriate regulations; e.g., for wine spirits in the EC, the terms VS, VSOP, and XO denote a minimum aging period of 3, 4, and 6 years, respectively. In many countries, the label should denote the actual volume and alcoholic strength of the product. In future, this may be extended to include a list of ingredients, which is already required in a few markets. Health warnings, regarding dangers of excessive consumption and risks to unborn children, are mandatory in the United States.

Some definitions limit the content of naturally occurring compounds which are potential health risks, such as methanol in fruit-derived spirits and hydrocyanic acid in stone-fruit spirits [7]. Some countries have additional limits for other naturally occurring compounds, such as nitrosamines in malt-derived spirits and urethane. Specific legislation may also exist for contaminants such as heavy metals. For example, the UK limit for lead in most spirits is 0.2 ppm compared with the general food limit of 1 ppm. The final product is also subject to consumer protection legislation.

7. Analysis

The primary measurements of the quality of spirit drinks are aroma, taste, and visual appearance. The use of nosing panels and sophisticated statistical analysis of their sensory comments, has been helpful in flavor and marketing research, but may be cumbersome for routine quality control [33], [34]. Chemical and physical analytical measurements supplement the traditional organoleptic quality judgements. Gas and liquid chromatography have been of great use in identifying many of the aroma compounds in distilled beverages, and more than 1000 are now known [35].

In 1992, the EC approved methods of analysis for neutral spirit [36], [37], and methods relevant to specific spirit drinks are planned. Elsewhere, particularly in North America, the AOAC Official Methods of Analysis are the basis of most methods [38].

Alcohol is measured by hydrometers, pycnometers, and density meters. Gas chromatography can be quite precise, but is not accepted by many excise authorities. Products containing high levels of dissolved solids, which can obscure direct alcohol measurement, must be distilled prior to analysis.

Color and clarity are measured in the EC neutral spirit methods by visual comparison with a water standard. In other spirits, color is determined by spectrophotometry at 430 or 525 nm. Clarity is measured in a nephelometer against a silicate standard.

Certain metals may be measured by traditional wet chemical methods, as are other ions, but atomic absorption spectrophotometry is the preferred method in well-equipped laboratories.

In neutral spirits, the permanganate clearing time is a general measure of the absence of impurities such as aldehydes. It is determined by monitoring the time taken for the color of the sample, after addition of potassium permanganate, to match a color standard. Aldehydes are measured, as acetaldehyde, by comparing the color formation after addition of Schiff's reagent. Higher alcohols, expressed as 2-methylpropan-1-ol, are determined by measuring the color formation after reacting the sample with an aromatic aldehyde, such as salicylaldehyde or *p*-dimethylaminobenzaldehyde, in hot dilute sulfuric acid. Acidity is determined by titration against sodium hydroxide. Esters, expressed as ethyl acetate, are determined by measuring color complexes at 525 nm after reacting the neutral spirit with hydroxylamine hydrochloride in al-

kaline solution, followed by iron(III) ions in acid solution. Methanol is measured by gas chromatography (GC) [36], [37].

Although GC methods are not precise enough to measure the low levels of higher alcohols and esters found in neutral spirits, they are the preferred and most accurate method for the determination of volatiles such as higher alcohols, aldehydes, esters, and fatty acids in other distillates. Traditional wet chemical methods can underestimate the total higher alcohol concentration because they do not detect *n*-propanol. In highly-flavored spirit, GC methods can give a truer measure of ester content, as many are of higher molecular mass than the ethyl acetate standard used in wet-chemical methods. Volatile compounds derived from fruits and other botanicals, such as terpenes, can also be measured by GC methods. However, products containing high levels of nonvolatile materials, such as sugars, require prior distillation [38].

GC coupled to sophisticated detectors, such thermal energy analyzers (TEA) or mass spectrophotometers is essential for the detection of trace compounds such as nitrosamines and urethane [38]. Liquid chromatography is now used for measuring sugars and other less-volatile materials, such as vanillin and other lignin-breakdown products in aged spirits. Carbon-14 measurement by scintillation counters is used to authenticate ethanol of fermentation origin as opposed to that derived from fossil fuels [36], [37].

8. References

1. A. H. Rose: "History and Scientific Basis of Alcoholic Beverage Production," in A. H. Rose (ed.): *Economic Microbiology,* vol. 1, Academic Press, London 1977, pp. 1 – 41.
2. *Kirk-Othmer,* 3rd ed., **vol. 3,** pp. 830 – 863.
3. L. Bluhm: "Distilled Beverages," in H.-J. Rehm, G. Reed (eds.): *Biotechnology,* vol. 5, Verlag Chemie, Weinheim 1983, pp. 447 – 475.
4. M. S. Moss, J. R. Hume: *The Making of Scotch Whisky,* James and James, Edinburgh 1981.
5. V. F. Larsen: "Alcohol Recovery," in D. R. Berry, I. Russell, G. G. Stewart (eds.): *Yeast Biotechnology,* Allen and Unwin, London 1987, pp. 501 – 531.
6. T. Lord: *The World Guide to Spirits, Liqueurs, Aperitifs and Cocktails,* Macdonald and Jane's Publishers, London 1979.
7. Council Regulation (EC) no. 1576/89 in *Official Journal of the European Communities* **32** (1989) L 160, 1.
8. The Scotch Whisky Order 1990, UK Statutory Instrument 1990, no. 998.
9. T. P. Lyons, A. H. Rose: "Whisky," in A. H. Rose (ed.): *Economic Microbiology,* vol. 1, Academic Press, London 1977, pp. 635 – 692.
10. M. Lehtonen, H. Suomalainen: "Rum," in A. H. Rose (ed.): *Economic Microbiology,* vol. 1, Academic Press, London 1977, pp. 595 – 633.
11. Commission Regulation (EEC) no. 1014/90 in *Official Journal of the European Communities* **33** (1990) L 105, 9.
12. A. C. Simpson: "Gin and Vodka," in A. H. Rose (ed.): *Economic Microbiology,* vol. 1, Academic Press, London 1977, pp. 537 – 593.
13. A. M. MacLeod: "The Physiology of Malting", in J. R. A. Pollock (ed.): *Brewing Science,* vol. 1, Academic Press, London 1979, pp. 146 – 232.
14. D. E. Briggs, J. S. Hough, R. Stevens, T. W. Young: *Malting and Brewing Science,* 2nd ed., vol. 1, Chapman and Hall, London 1981.
15. G. H. Palmer: "Cereals in Malting and Brewing," in G. H. Palmer (ed.): *Cereal Science and Technology,* Aberdeen University Press, Aberdeen 1989, pp. 61 – 242.
16. G. N. Bathgate: "Cereals in Scotch Whisky Production," in G. H. Palmer (ed.): *Cereal Science and Technology,* Aberdeen University Press, Aberdeen 1989, pp. 243 – 278.
17. G. N. Bathgate, R. Cook: "Malting of Barley for Scotch Whiskies," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 19 – 63.
18. G. D. Wilkin: "Milling, Cooking and Mashing," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 64 – 88.
19. D. C. Watson: "The Development of Specialised Yeast Strains for Use in Scotch Malt Whisky Fermentations," in G. G. Stewart, I. Russell (eds.): *Current Developments in Yeast Research,* Pergamon Press, London 1981, pp. 57 – 62.
20. D. R. Berry: "The Physiology and Microbiology of Scotch Whisky Production," in M. E. Bushell (ed.): *Progress in Industrial*

Microbiology, Elsevier, Amsterdam 1984, pp. 199 – 243.
21. M. Korhola, K. Harju, M. Lehtonen: "Fermentation," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 89 – 117.
22. D. C. Watson: "Distilling Yeasts," in *Developments in Industrial Microbiology,* vol. 25, Society for Industrial Microbiology, Arlington 1984, pp. 213 – 220.
23. D. C. Watson: "Yeasts in Distilled Alcoholic Beverage Production," in A. H. Rose, J. S. Harrison (eds.): *The Yeasts,* 2nd ed., vol. 5, Academic Press, London 1993, pp. 215 – 244.
24. D. Nicol: "Batch Distillation," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 118 – 149.
25. D. C. Watson: "Factors Influencing the Congener Composition of Malt Whisky New Spirit," in J. R. Piggott (ed.): *Flavour of Distilled Beverages: Origin and Development,* Ellis Horwood, Chichester 1983, pp. 79 – 92.
26. D. R. Berry, D. C. Watson: "Production of Organoleptic Compounds," in D. R. Berry, I. Russell, G. G. Stewart (eds.): *Yeast Biotechnology,* Allen and Unwin, London 1987, pp. 345 – 368.
27. D. C. Watson: "A Laboratory Apparatus for Distillation Research," in F. G. Priest, I. Campbell (eds.): *Current Developments in Malting, Brewing and Distilling,* Institute of Brewing, London 1983, pp. 249 – 255.
28. K. Nishimura, R. Matsuyama: "Maturation and Maturation Chemistry," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 235 – 263.
29. D. C. Watson, *CRC Crit. Rev. Biotechnol.* **2** (1985) 147 – 192.
30. R. J. Panek, A. R. Boucher: "Continuous Distillation," in J. R. Piggott, R. Sharp, R. E. B. Duncan (eds.): *The Science and Technology of Whiskies,* Longman Scientific & Technical, Harrow 1989, pp. 150 –181.
31. J. A. P. L'Anson, *Process Biochem.* **6** (1971) no. 7, 35 – 39.
32. R. Cantagrel: "A Scientific Approach to Quality Control for Cognac Spirits," in J. R. Piggott, A. Paterson (eds.): *Distilled Beverage Flavour,* Ellis Horwood, Chichester 1989, pp. 117 – 132.
33. P. Dürr: "Sensory Analysis of Alcoholic Beverages," in J. R. Piggott, A. Paterson (eds.): *Distilled Beverage Flavour,* Ellis Horwood, Chichester 1989, pp. 33 – 40.
34. M. L. Arnott: "Computerised Collection and Statistical Analysis of Descriptive Sensory Profiling Data," in J. R. Piggott, A. Paterson (eds.): *Distilled Beverage Flavour,* Ellis Horwood, Chichester 1989, pp. 65 – 74.
35. L. Nykänen, H. Suomalainen: *Aroma of Beer, Wine and Distilled Alcoholic Beverages,* Kluwer Academic Publ. Group, Dordrecht 1983.
36. Commission Regulation (EEC) no. 1238/92 in *Official Journal of the European Communities* **35** (1992) L 130, 13.
37. Commission Regulation (EEC) no. 2009/92 in *Official Journal of the European Communities* **35** (1992) L 203, 10.
38. K. Helrich (ed.): *Official Methods of Analysis of the Association of Official Analytical Chemists,* 15th ed., AOAC, Arlington, 1990.

Stains, Microscopic

UTE SCHMIDT, Merck KGaA, Darmstadt, Federal Republic of Germany

ERIK K. W. SCHULTE, Anatomisches Institut, Fachbereich Medizin, Johannes-Gutenberg-Universität, Mainz, Federal Republic of Germany

1. **Introduction** 1
2. **Historical Survey** 1
3. **Characteristics of Microscopic Stains** 2
4. **Nomenclature, Sources of Supply** . . . 6
5. **Classification of Biological Stains for Microscopy** 6
5.1. **Basic or Cationic Stains** 8
5.2. **Acid or Anionic Stains** 11
5.3. **Mordant Stains** 11
5.4. **Oil-Soluble Stains** 11
5.5. **Chromogenic Reactions** 12
6. **Enzyme Histochemical Reactions** . . . 12
7. **Immunohistochemical Staining Methods** 13
8. **References** 14

1. Introduction

Microscopic stains are used for distinguishing visually between different cellular and tissue components in animal and plant material. Optical differentiation of this type is essential for the detection and accurate identification of cellular and tissue elements. Microscopic stains are most widely used in histology, cytology, and microbiology, but they are also valuable in the examination of textile fibers, paper, and other industrial products, as well as in unusual assignments like distinguishing individual paint layers in a work of art.

2. Historical Survey

At least one important microscopic stain, carmine, was employed for dyeing textiles in the days of ancient Rome and Greece. The substance was isolated from insects (*Coccus ilicus* and *C. cacti*) and prepared for use by heating with various salts (e.g., alum). Carminic acid is the actual coloring constituent of carmine. As early as 1770 carmine was utilized by HILL for microscopic purposes, and EHRENBERG (1838) later used it in the examination of protozoa. HARTIG (1854) employed carmine to highlight various parts of plants. VON GERLACH (1858) employed an ammoniacal carmine solution to demonstrate animal blood vessels, and in the process he discovered that various cellular elements had different affinities for the dye. The first modern histological stain, one that is still important today, was an extract of campeachy wood *(Haematoxylon campechianum)* – hematoxylin. BÖHMER (1865) developed a procedure for increasing the staining potential of hematoxylin with the help of alum, which acts as a mordant. Other early stains that are still employed include the alum hematoxylins of DELAFIELD (1885), EHRLICH (1886), MAYER (1891), and HARRIS (1900); the iron hematoxylins of WEIGERT (1892), BENDA (1886), and HEIDENHAIN (1892); and MALLORY's (1891) hematoxylin solution with phosphomolybdic acid.

The first synthetic aniline dye, aniline black, was prepared by RUNGE (1834). In the years that followed, many new aniline dyes were synthesized, some of which are still used for microscopic purposes. BÖTTCHER (1869) introduced the principle of differentiation with alcohol after staining with carmine or rosaniline. FISCHER (1876) used aqueous and especially alcoholic solutions of eosin for staining various tissues. Between 1875 and 1880, numerous new, more complex aniline dyes (e.g., triphenylmethane dyes) were employed for double staining. EHRLICH (1879) first classified dyes as acid or basic, showing that these differed in both their chemical properties and their histological applications. An extensive study of the oxidation products of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a24_567

methylene blue was carried out by BERNTHSEN (1885).

The first trichrome staining method for connective tissue was developed by VAN GIESON (1889). This was followed by the introduction of many important staining methods, including the use of methylene blue (UNNA, 1891), triple staining (FLEMMING, 1891. BENDA, 1891), methyl green – pyronin (PAPPENHEIM, 1899), and Sudan III (DADDI, 1896). Between 1875 and 1895, fundamental cytological phenomena were discovered and described with the help of staining (e.g., mitosis, cell maturation, and fertilization). At first, fixation and staining were carried out as a single step, but the introduction of thin sections made it necessary to perform fixation separately. A mixture of osmic/chromic acid (FLESCH, 1879) or osmic/acetic acid (FLEMMING, 1882) was used as a fixative, permitting very effective demonstration and preservation of cell structures. Triple aniline staining (e.g., with safranine, gentian violet, and orange G), described by FLEMMING (1891), was introduced to cytology and histology in the 1890s.

A new class of dyes – the phthalocyanines– was developed in 1907. These also belong to the category of acid and basic dyes. A particularly well-known member of the group is alcian blue 8GX. Apart from slight alterations, the staining techniques of ROMANOWSKY (1891), JENNER (1890), LEISHMAN (1901), WRIGHT (1901), MAY – GRÜNWALD (1902), and GIEMSA (1902) are still employed today. All involve methylene blue, its oxidation products, and eosin in various combinations. PAPPENHEIM (1912) utilized the MAY-GRÜNWALD solution and subsequently the GIEMSA solution in "universal panoptic staining". Solutions of methylene blue and eosinate were also perfectly suited for the identification of parasites in blood. WITTEKIND (1976) found that azure B is the true coloring constituent of polychrome methylene blue, which consists of a mixture of methylene blue, azure A, B, and C, and thionin.

Attempts to stain living material were described as early as the seventeenth century. The first experiments were conducted with the naked eye, initially with insoluble colored substances and later with staining solutions (e.g., saffron, methylene blue, and neutral red). The development of colloidal acid dyes (e.g., trypan red, trypan blue, and Evans blue) early in the twentieth century led to the introduction of vital staining. This method has been used to study, for example, tissue cultures, cells in flowing blood, and nerve tissue.

CONN (1941) developed a completely new staining method for the localization of antigens using labeled antibodies. Diagnostic use of this method became possible after the discovery of special fluorescent dyes (e.g., fluorescein isothiocyanate, FITC) and ways to bind them to antisera. The binding of antibodies to horseradish peroxidase represented another remarkable step forward, because this enzyme can be visualized after it has been coupled to antigens present in a tissue (NAKANEU; PIERCE, 1967).

3. Characteristics of Microscopic Stains

A microscopic stain is a colored compound that is capable of binding to a substrate (see Table 1). Depending on the number of stains used, a distinction can be made between *simple* and *multiple* staining. Multiple staining is divided in turn into *successive* and *simultaneous* staining, depending on the manner in which the stains are applied. An other useful distinction is that between *diffuse* and *selective* staining. In contrast to diffuse staining, selective staining involves exclusive or especially conspicuous coloration of particular components. Frequently, the staining process involves such a high degree of chemical specificity that the stain can be used as a histochemical reagent.

Stains are called *metachromatic* if they induce in certain constituents of the substrate a color different from that of the stain itself. In *orthochromatic* staining, the shade of the colored tissue is identical to that of the staining solution. Metachromasia is based on a shift in the absorption wavelength of a stain as a consequence of its binding to the substrate with concurrent formation of oligomers of the stain ("absorption units"). Thiazines, oxazines, azines, and xanthenes are among the dyes that exhibit metachromatic effects.

Two characteristic groups of atoms can be differentiated within a stain molecule: the *chromophore* and the *auxochrome*. Chromophoric

Table 1. Microscopic stains in biology and medicine

Dye (Colour Index no.)	CAS registry no.	Principal uses
Acridine dyes		
Acridine orange (C.I. 46 005)	[*10127-02-3*]	fluorescent staining of DNA and RNA (differential stain) for flow cytometry, supravital staining in cytology
Acriflavine	[*8048-52-0*]	fluorescent stain for aldehydes (pseudo-Schiff reagent)
Anthraquinone dyes		
Alizarin red S (C.I. 58 005)	[*130-22-3*]	embryonic bone (fetal ossification), calcium, bone
Carmine (C.I. 75 470)	[*1390-65-4*]	chromatin (mainly as borax carmine), glycogen (Best's stain)
Carminic acid (C.I. 75 470)	[*1260-17-9*]	chromatin
Nuclear fast red (C.I. 60 760)	[*6409-77-4*]	chromatin (as aluminum lake)
Arylmethane and diarylmethane dyes		
Auramine O (C.I. 41 000)	[*2465-27-2*]	fluorescent staining of acid-fast bacteria, fluorochrome Schiff's reagent
Triarylmethane dyes		
Anionic		
Acid fuchsin (C.I. 42 685)	[*3244-88-0*]	connective tissue (van Gieson's, Cason's, Mallory's stain)
Brilliant blue R (C.I. 42 660)	[*6104-59-2*]	stoichiometric staining of proteins
Fast green FCF (C.I. 42053)	[*2353-45-9*]	cyptoplasmic counterstain (Papanicolaous's stain), nuclear histone proteins, arginine-rich sites
Light green SF (C.I. 42 095)	[*5141-20-8*]	cytoplasmic counterstain (Papanicolaou's stain), nuclear histone proteins, arginine-rich sites
Cationic		
Brilliant green (C.I. 42 040)	[*633-03-4*]	bacteria, components of culture media
Crystal violet (C.I. 42 555)	[*548-62-9*]	bacteria (Gram's stain), chromatin, Holzer's stain for neuroglia
Malachite green (C.I. 42 000)	[*2437-29-8*]	bacteria and fungi, plant cell nuclei and pollen
Methyl green (C.I. 42 585)	[*7114-03-6*]	chromatin (DNA quantitatively), fungi, gonococci
New fuchsin (C.I. 42 520)	[*3248-91-7*]	Schiff's reagent (Feulgen and PAS reaction), bacteria (Ziehl – Neelsen), elastic fibers (resorcin – fuchsin stain)
Pararosaniline (C.I. 42 500)	[*6035-94-5*]	Schiff's reagent (Feulgen and PAS reaction), elastic fibers (aldehyde fuchsin), bacteria (Ziehl – Neelsen)
Rosaniline (C.I. 42 510)	[*632-99-5*]	Schiff's reagent (Feulgen and PAS reaction), bacteria (Ziehl – Neelsen)
Victoria blue B (C.I. 44 045)	[*2580-56-5*]	chromatin (used in cytology, histology, botany), glial fibers in neurohistology, bacteria, elastic fibers
Azine dyes		
Azocarmine B (C.I. 50 090)	[*25360-72-9*]	connective tissue (azan stain)
Janus green B (C.I. 11 050)	[*2869-83-2*]	supravital stain for mitochondria
Azo dyes		
Alizarin yellow GG (C.I. 14 025)	[*584-42-9*]	indicator in bacterial media
Biebrich scarlet (C.I. 26 905)	[*4196-99-0*]	connective tissue (in trichrome stains), nuclear histone proteins, arginine-rich sites
Evans blue (C.I. 23 860)	[*314-13-6*]	vital stain for tumor cells, retrograde staining of axons
Orange II (C.I. 15 510)	[*633-96-5*]	cytoplasmic counterstain in histology
Orange G (C.I. 16 230)	[*1936-15-8*]	cytoplasmic counterstain in many techniques, keratin
Congo red (C.I. 22 120)	[*573-58-0*]	amyloid, cytoplasmic counterstain
Trypan blue (C.I. 23 850)	[*72-57-1*]	vital staining of cytoplasm
Bismarck brown Y (C.I. 21 010)	[*10114-58-6*]	cytoplasmic counterstain (Papanicolaou's stain)
Oil red O (C.I. 26 125)	[*1320-06-5*]	lipids in histological material and in electrophoresis
Ponceau S (C.I. 27 195)	[*6226-79-5*]	connective tissue (in van Gieson's stain), gel electrophoresis
Sudan red 7 B (C.I. 26 050)	[*6368-72-5*]	lipids in histological material
Sudan black B (C.I. 26 150)	[*4197-25-5*]	fat in histological material, bacteria, Golgi apparatus
Benzimidazole dyes		
Hoechst 33 258	[*23491-44-3*]	fluorochrome for *q*-banding of chromosomes

Table 1. (Continued)

Diazonium salts		
Fast blue B (C.I. 37 235)	[*14263-94-6*]	α-naphthyl acetate in monocytes and *T* lymphocytes, coupled tetrazonium reaction for proteins
Fast blue RR (C.I. 37 155)	[*14726-29-5*]	coupling agent for enzymes (esterases and phosphatases)
Hydroxyketone dyes		
Hematoxylin (C.I. 75 290)	[*517-28-2*]	chromatin (most widely used stain for cell nuclei), applied as metal – complex stain with aluminum, iron, chromium
Indigo dyes		
Indigo carmine (C.I. 73 015)	[*860-22-0*]	cytoplasmic counterstain, connective tissue
Oxazine dyes		
Brilliant cresyl blue (C.I. 51 010)	[*81029-05-2*]	reticulocytes, Heinz bodies, chromatin
Celestine blue (C.I. 51 050)	[*1562-90-9*]	chromatin (substitute for hematoxylin)
Cresyl fast violet	[*10430-47-4*]	Nissl substance, Barr bodies

groups absorb light in the visible region of the spectrum and confer upon the stain its color, which represents the complementary color of the absorbed light. Typical chromophores are atomic groupings containing double bonds, such as C=O, C=S, C=N, N=N, N=O, NO_2, and various conjugated systems. The involvement of several adjacent chromophores is also common. Auxochromes deepen the color and fulfill the additional important function of binding the stain to the substrate. These substituents may be present either in ionized form or as neutral functionalities that are capable of forming covalent bonds with the substrate or coordinate bonds with a metal ion (*mordant*). Stains often contain several auxochromes. The counterions of auxochromes are usually H^+, Na^+, or Cl^-. Stains with positively charged auxochromes (e.g., primary, secondary, and tertiary amines) are called *basic* (better: *cationic*) stains. The *acid* or *anionic* stains contain negatively charged auxochromes (carboxylate, sulfonate, or phenolate groups). Stains with both anionic and cationic groups in the same chromophore, which are thus electrically neutral overall, are called *amphoteric*.

Organic compounds containing conjugated double-bond systems that form large, relatively rigid, planar structures are often capable of fluorescence. The nature of the linear arrangement and the presence of one or several substituents can significantly influence the fluorescent properties of such a stain.

In *progressive* staining, a substance is colored with a staining solution until a particular desired color intensity has been achieved. In *regressive* staining, the tissue is intentionally overstained, and excess stain is subsequently removed (e.g., by dissolving it in water, alcohol, or acidified alcohol) in a differentiation process.

In principle, true staining is achieved only after binding has occurred between the stain and the substrate. Physical factors that influence the staining process include the size of the dye particle, the porosity of the substrate, and diffusion effects. Many different types of physical and chemical linkage can develop between a stain and a substrate, ranging from adsorption to covalent bonding. The simplest form of staining is based on electrostatic attraction between oppositely charged ionized groups on the stain and the substrate, but numerous histochemical procedures result in the formation of covalently bound colored material derived from reactive groups in the stain and the tissue. Such a stain cannot easily be leached out. Van der Waals forces are involved in the direct staining of cellulose and elastin. Hydrophobic interactions are responsible for the penetration of oil-soluble stains into lipids, and for their retention in hydrophobic tissue. Metal ions (e.g., aluminum, iron, chromium, tungsten, molybdenum, and copper), known as *mordants*, are capable of forming stain – metal complexes with many stain molecules. Ring-shaped chelates often play an important role in such stain – metal complexes. A mordant can be applied prior to staining, mixed with the stain, or applied after staining. A complex that contains excess stain relative to mordant is called a *lake*, although most staining methods instead involve excess mordant (e.g., staining with iron or aluminum hematoxylin). The chief benefit

Table 1. (Continued)

Dye (Colour Index no.)	CAS registry no.	Principal uses
Cresyl violet acetate	[*10510-54-0*]	Nissl substance, chromatin, Heinz bodies
Gallocyanine (C.I. 51 030)	[*1562-85-2*]	Nissl granules, nucleic acids (quantitatively)
Nile blue A (C.I. 51 180)	[*3625-57-8*]	vital staining of transplanted tissue, lipids (after oxidation to Nile red)
Nile red	[*7385-67-3*]	lipid stain, fluorochrome in flow cytometry
Orcein	[*1400-62-0*]	connective tissue (elastic fibers) (Darrow's, Fraenkel's stain), HB_S-antigen in histology
Phenazine dyes		
Neutral red (C.I. 50 040)	[*553-24-2*]	vital stain for cytoplasmic organelles, bacteria, Nissl substance, chromatin (Twort's stain)
Safranine O (C.I. 50 240)	[*477-73-6*]	chromatin, bacteria (Gram's stain)
Phthalocyanine dyes		
Alcian blue (C.I. 74 240)	[*33864-99-2*]	glycosaminoglycans in histology
Luxol fast blue MBSN (C.I. 74 180)	[*1328-51-4*]	myelin sheaths in histology
Thiazine dyes		
Azure A (C.I. 52 005)	[*531-53-3*]	blood and bone marrow smears (constituent of polychrome methylene blue), reticulocytes
Azure B (C.I. 52 010)	[*531-55-5*]	blood and bone marrow smears (constituent of polychrome methylene blue), reticulocytes
Azure C (C.I. 52 002)	[*531-57-7*]	blood and bone marrow smears (constituent of polychrome methylene blue), reticulocytes
Methylene blue (C.I. 52 015)	[*61-73-4*]	blood and bone marrow smears (constituent of polychrome methylene blue), bacteria, vital staining of axons
New methylene blue N (C.I. 52 030)	[*6586-05-6*]	reticulocytes
Thionin (C.I. 52 000)	[*78338-22-4*]	Nissl granules, chromatin, metachromatic substances, Barr bodies, Feulgen reac- tion
Toluidine blue O (C.I. 52 040)	[*92-31-9*]	chromatin, metachromatic substances, Nissl substance
Thiazole dyes		
Thioflavin T (C.I. 49 005)	[*2390-54-7*]	fluorescent stain for viruses
Xanthene dyes		
Anionic		
Rose bengal (C.I. 45 440)	[*632-69-9*]	cytoplasmic counterstain, eosinophilic granules, bacteria
Eosin Y (C.I. 45 380)	[*548-26-5*]	cytoplasmic counterstain in many techniques, eosinophilic granules
Erythrosin B (C.I. 45 425)	[*568-63-8*]	cytoplasmic counterstain
Fluorescein isothiocyanate (C.I. 45 350 : 1)	[*3326-32-7*]	fluorescent antibody marker
Phloxine B (C.I. 45 410)	[*18472-87-2*]	cytoplasmic counterstain
Cationic		
Pyronin Y (C.I. 45 005)	[*92-32-0*]	RNA (quantitative) in combination with methyl green
Rhodamine B (C.I. 45 170)	[*81-88-9*]	fluorescent staining of lectins, fluorochrome for peroxidase, fluorochrome for bacteria
Chromogenic reagents		
Nitro blue tetrazolium	[*298-83-9*]	histochemical determination of dehydrogenases and alkaline phosphatase
Tetranitro blue tetrazolium	[*42798-98-1*]	histochemical determination of dehydrogenases

of histological staining with mordant stains is that colors developed in the tissue are not subject to extraction with water, alcohol, or weak acids, all of which are used in counterstaining, where counterstaining means that different cellular structures are stained different colors (e.g., cell nuclei blue and cytoplasm orange). Strong acids decolorize the tissue, an effect that can be exploited for differentiation purposes.

4. Nomenclature, Sources of Supply

Most dyes are referred to by trivial names that provide no insight into their chemical nature

or use. The third edition of the *Colour Index* [39], identifies dyes and other colored compounds by number, and describes their colors and fields of application. Natural dyes are listed separately. The trade name of a dye mixture typically reflects the dominant or most important constituent present. Few dyes are supplied as pure compounds; instead, they are mixed with inactive fillers, usually sodium chloride, and they frequently consist of more than one colored substance. Chromatographic and spectroscopic techniques are used to test the quality of a dye.

The Biological Stain Commission was founded in the United States in 1920 to certify samples of stains for quality and suitability in specific applications. Certified stains always carry a label stating the content, the Colour Index number, and the batch number. The criteria applied to industrial dye and stain standards, which also serve as a basis for DIN and National Committee for Clinical Laboratory Standards (NCCLS) specifications, are listed below.

Dyes:

1) Chemical definition: generic name, chemical name molecular structure, Colour Index (C.I) number, CAS registry number
2) Relative molecular mass of the dye molecule
3) Counterion of the dye
4) Wavelength (λ) or wavenumber (ν) of absorption maximum
5) Decadic molar extinction coefficient
6) Maximum water content of the batch
7) Absence of solvent
8) Impurities:
   Maximum content of inorganic impurities
   Maximum content of colored organic impurities (as determined by TLC and/or HPLC)
   Maximum content of colorless impurities

Stains:

1) Colored components (dyes) in the stain:
   Types of dyes present (including counterions)
   Dye concentrations
   Types and amounts of colored and colorless impurities
2) Concentration of inorganic salts
3) Chemical description of the solvent
4) pH of the stain (for aqueous solutions)
5) Stability of the stain (shelf-life, expiration date)

See Table 2 for physical and chemical properties of stains listed in Table 1.

The following manufacturers produce dyes for microscopic staining:

Australia	ICI Australia Operations Pty Ltd.
Canada	Northern Pigment
China	China National Chemicals Import & Export Corp.
France	Wackherr, Les Colorants
Germany	BASF AG
	Bayer AG
	Merck KGaA
	Hoechst AG
	Sigma Chemie GmbH
India	India Dyes Stuff Industries Ltd.
	National Dyes
Japan	Dainippon Ink and Chemicals
	Mitsubishi Chemical Industries Ltd.
	Sumitomo Chemical Co. Ltd.
Spain	Fabricacion National de Colorantes S.A.
Switzerland	Ciba-Geigy Ltd.
United Kingdom	BDH Ltd.
	ICI RLC
United States	Aldrich Chemical Co.
	Carolina Color & Chemical Co.
	Eastman Chemical Products
	EM Science

5. Classification of Biological Stains for Microscopy

Biological stains have traditionally been classified by their users as basic, acid, mordant, or oil-soluble. For summary lists of representative staining procedures, see Tables 3, 4, 5, 6.

5.1. Basic or Cationic Stains

Basic or cationic stains are used to stain nuclei and most of the anionic structural elements of the cytoplasm and extracellular space. The stains are applied in acidified solution that guarantees ionization of their primary, secondary, and tertiary amino groups. Cationic stains are capable of binding to three types of anions present in tissues: phosphate residues of nucleic acids, sulfate esters of various macromolecular carbohydrates, and carboxylate ions from carbohydrates and

Table 2. Physicochemical data for stains listed in Table (according to [13])∗

Dye	λ_{max}, nm	Formula	Mr	Color of stained entity
Acid fuchsin	546 (W)	$C_{20}H_{17}N_3Na_2O_9S_3$	585.55	magenta
Acridine orange	489 (W)	$C_{17}H_{19}N_3$	369.94	yellow-orange (f)
Acriflavine	462 (M)	$C_{14}H_{14}N_3$	330.81	yellow (f)
Alcian blue 8GX	615 (W)	$C_{56}H_{68}Cl_4CuN_{16}S_4$	1298.88	blue
Alizarin red S	556 (W)	$C_{14}H_7NaO_7S$	360.28	red-violet
Alizarin yellow GG	362 (M)	$C_{13}H_8N_3NaO_5$	309.22	yellow (f)
Auramine O	432 (W)	$C_{17}H_{21}N_3$	303.84	yellow (f)
Azocarmine B	516 (W)	$C_{28}H_{17}N_3Na_2O_9S_3$	681.63	reddish
Azure A	633 (W)	$C_{14}H_{14}ClN_3S$	291.80	blue-gray
Azure B	641 (E)	$C_{15}H_{16}ClN_3S$	305.83	blue
Azure C	616 (W)	$C_{13}H_{12}ClN_3S$	277.78	blue
Biebrich scarlet	505 (M)	$C_{22}H_{14}N_4Na_2O_7S_2$	556.49	scarlet red
Bismarck brown Y	457 (E/W)	$C_{18}H_{18}N_8 \cdot 2HCl$	419.32	dark brown
Brilliant blue R	585 (E)	$C_{45}H_{44}N_3NaO_7S_2$	825.99	bright blue
Brilliant cresyl blue	622 (E/W)	$C_{17}H_{20}ClN_3O$	385.96	blue
Brilliant green	625 (E/W)	$C_{27}H_{34}N_2O_4S$	482.65	green
Carmine	528 (DM)	$C_{44}H_{37}O_{27}AlCa$	1118.78	red
Carminic acid	495 (M)	$C_{22}H_{20}O_{13}$	492.40	red
Celestine blue	642 (W)	$C_{17}H_{18}ClN_3O_4$	363.80	blue
Congo red	497 (W)	$C_{32}H_{22}N_6Na_2O_6S_2$	696.67	red
Cresyl fast violet	604 (E/W)	$C_{17}H_{14}ClN_3O$	311.77	violet
Cresyl violet acetate	596 (E)	$C_{18}H_{15}N_3O_3$	321.34	violet-red
Crystal violet	590 (W)	$C_{25}H_{30}ClN_3$	407.99	violet
Eosin Y	530 (E)	$C_{20}H_6Br_4Na_2O_5$	691.88	red
Erythrosin B	525 (W)	$C_{20}H_6I_4Na_2O_5$	879.87	red
Evans blue	611 (W)	$C_{34}H_{24}N_6Na_4O_{14}S_4$	960.82	blue
Fast blue B	371 (W)	$C_{14}H_{12}Cl_2N_4O_2 \cdot ZnCl_2$	475.46	blue
Fast blue RR	319 (M)	$C_{15}H_{16}N_2O_3$	272.31	blue
Fast green FCF	624 (E/W)	$C_{37}H_{34}N_2Na_2O_{10}S_3$	808.86	green
Fluorescein isothiocyanate	490 (W)	$C_{21}H_{11}NO_5S$	389.39	yellow (f)
Gallocyanine	620 (E)	$C_{15}H_{13}ClN_2O_5$	336.73	gray-blue
Hematoxylin	292 (M)	$C_{16}H_{14}O_6$	302.29	pale yellow
Indigo carmine	608 (W)	$C_{16}H_8N_2Na_2O_8S_2$	466.36	blue
Janus green B	660 (E/W)	$C_{30}H_{31}ClN_6$	511.07	bluish green
Light green SF	630 (W)	$C_{37}H_{34}N_2Na_2O_9S_3$	792.86	green
Malachite green	614 (W)	$C_{46}H_{50}N_4 \cdot C_2H_2O_4$	927.03	green
Methylene blue	661 (W)	$C_{16}H_{18}ClN_3S$	373.90	blue
Methyl green	633 (W)	$C_{27}H_{35}Cl_2N_3$	472.51	green
Neutral red	540 (E/W)	$C_{15}H_{17}ClN_4$	288.78	red
New fuchsin	554 (E)	$C_{22}H_{24}ClN_3$	365.91	reddish violet
New methylene blue N	630 (W)	$C_{18}H_{22}ClN_3S$	416.05	blue
Nile blue A	633 (E/W)	$C_{40}H_{40}N_6O_6S$	732.84	dark blue
Nile red	553 (M)	$C_{20}H_{18}N_2O_2$	318.38	reddish violet
Nitro blue tetrazolium	260 (W)	$C_{40}H_{30}Cl_2N_{10}O_6$	867.80	blue (dye precursor)
Nuclear fast red	535 (W)	$C_{14}H_8NNaO_7S$	357.28	red
Oil red O	518 (T)	$C_{26}H_{24}N_4O$	408.51	red
Orange G	475 (W)	$C_{16}H_{10}N_2Na_2O_7S_2$	452.38	orange-red
Orange II	483 (W)	$C_{16}H_{11}N_2NaO_4S$	350.33	orange-red
Orcein	579 (W)			yellowish
Pararosaniline	542 (E)	$C_{19}H_{18}ClN_3$	323.73	red-violet
Phloxine B	548 (E/W)	$C_{20}H_2Br_4Cl_4Na_2O_5$	829.66	red
Ponceau S	520 (W)	$C_{22}H_{12}N_4Na_4O_{13}S_4$	760.57	red
Pyronin Y	550 (W)	$C_{17}H_{19}ClN_2O$	302.75	red
Rhodamine B	543 (M)	$C_{28}H_{31}ClN_2O_3$	479.02	bluish red

Table 2. (Continued)

Dye	λ_{max}, nm	Formula	*M*r	Color of stained entity
Rosaniline	544 (E)	$C_{20}H_{20}ClN_3$	337.86	red-violet
Rose bengal	548 (W)	$C_{20}H_2Cl_4I_4Na_2O_5$	1017.65	deep pink
Safranine O	530 (E/W)	$C_{20}H_{19}ClN_4$	350.85	red
Sudan black B	598 (E)	$C_{29}H_{24}N_6$	456.55	black
Sudan red 7 B	533 (T)	$C_{24}H_{21}N_5$	379.47	red
Tetranitro blue tetrazolium	279 (M)	$C_{40}H_{28}Cl_2N_{12}O_{10}$	907.64	greenish yellow (f)
Thioflavin T	412 (W)	$C_{17}H_{19}ClN_2S$	318.87	yellow-green
Thionin	599 (E)	$C_{12}H_{10}BrN_3S$	308.23	blue
Toluidine blue O	626 (W)	$C_{15}H_{16}ClN_3S$	305.83	blue
Trypan blue	607 (M)	$C_{34}H_{24}N_6Na_4O_{14}S_4$	960.82	blue
Victoria blue B	604 (E)	$C_{33}H_{32}BF_4N_3$	562.90	blue

∗ f = fluorescent, E = ethanol, M = methanol, T = toluene, W = water, DM = dimethylsulfoxide.

proteins. The dissociation of carboxylic acids is suppressed in solutions with high concentrations of hydrogen ions, whereas sulfate groups ionize even in strongly acidic medium. A suitable acid of moderate strength is phosphoric acid. A basic protein such as hemoglobin contains an excess of free amino groups relative to carboxyl groups. For this reason, carboxyl groups in such a protein bind cationic stains only in alkaline solution. The staining properties of cationic stains are thus influenced in a fundamental way by the pH of the solution: in acidic solution, only structures rich in sulfated carbohydrates are stained, but all structures are stained in neutral or alkaline solution. In their principal application (i.e., nuclear staining), cationic stains are used in dilute aqueous solution (0.1 – 0.5 %) acidified to pH 3 – 5.5 with glacial acetic acid or a suitable buffer.

Structures susceptible to cationic stains in acidic solution are called *basophilic*. When nuclei are stained with cationic dyes, basophilia is clearly apparent at sites of accumulation of cytoplasmic RNA (e.g., Nissl bodies in the neurons) and in structures that contain sulfated carbohydrates (e.g., mast cell granules, cartilage matrix, and many secretory products). Cationic stains do not stain the nucleus specifically.

The following stains can be used for both nuclear staining and counterstaining: toluidine blue (C.I. 52 040), neutral red (C.I. 50 040), methyl green (C.I. 42 585), and polychrome methylene blue. In alkaline or neutral solution cationic stains facilitate rapid, one-step survey staining.

Differential staining of DNA (blue-green with methyl green) and RNA (pink to red with pyronin Y) can be accomplished by simultaneous application of the two cationic stains methyl green and pyronin Y (C.I. 45 005). The selectivity of staining is a function of the structures of the stains and their associated staining kinetics. Alcian blue 8GS, 8GX, or 8GN (C.I. 74 240) can be used at pH 2.5 to demonstrate acidic muco substances. The advantage of using alcian blue in carbohydrate histochemistry is that this dye does not stain nuclei or cytoplasmic deposits of RNA in sections.

Cationic stains are also important for bacteriological staining. Crystal violet (C.I. 42 555) is utilized in various histological and cytological staining methods, and it is also employed in the gram technique. In this technique, bacterial cells are first stained with crystal violet, then treated with the mordant iodine, and finally counterstained (e.g., with safranin), a procedure that permits a distinction to be made between gram-positive (dark blue) and gram-negative (red) organisms. Basic fuchsin (pararosaniline, C.I. 42 500), a well-known histological dye used in the periodic acid Schiff (PAS) reaction in Feulgen staining and the Gomori aldehyde – fuchsin method, is also employed in the detection of various mycobacteria. Cell material in this case is stained with a phenolic solution of fuchsin and counterstained with methylene blue. Acid-fast rods appear bright red against the blue-stained background material.

Cationic stains such as acridine orange (C.I. 46 005) and oxazines can act as *fluorochromes.* Both living and fixed cells absorb these stains from dilute solutions. Acridine orange is used in supravital fluorescent staining, fluorescent stain-

Table 3. Staining procedures in cytology and histology

Specimen	Fixation	Staining technique (substrate)	Dyes and stains used
Cytology			
General cytology (smears, imprints, serous effusions)	wet fix: ethanol spray	Papanicolaou, H & E *	aluminum hematein, eosin Y, fast green FCF, Bismarck brown Y, orange G
	air dry: methanol	H & E,∗ Giemsa	for Giemsa see below
Blood and bone marrow smears	air dry: methanol	Standard Giemsa, May – Grünwald – Giemsa	azure B, eosin Y, azure A, B, C, methylene blue, eosin Y
Histology			
Cell nuclei	formaldehyde	Mayer's or Harris's hematein	aluminum hematein (eosin Y as counterstain)
	formaldehyde	Domagk	nuclear fast red, aniline blue
	formaldehyde	Twort	neutral red, fast green
Cytoplasmic organelles			
Mitochondria	formaldehyde – $CaCl_2$ (Baker)	Heidenhain	iron hematein
Lysosomes	formaldehyde	hexazonium – pararosaniline	pararosaniline
Endoplasmic	formaldehyde	Nissl	thionin
		methylene blue	methylene blue
reticulum			
Keratin	cytology: ethanol	Papanicolaou	thionin
	histology: formaldehyde	Congo red	methylene blue
Connective tissue			
Elastic fibers	formaldehyde	Weigert's resorcin – fuchsin	pararosaniline
	formaldehyde	Gomori's aldehyde – fuchsin	pararosaniline
	formaldehyde	Taenzer	orcein
Collageneous fibers	Bouin	Goldner's trichrome stain	iron-hematein, acid fuchsin, ponceau, orange G, light green SF, azophloxine
Cartilage	Helly	Delafield's hematein	hematein
	Schaffer	Mayer's resorcin – fuchsin	pararosaniline
	formaldehyde	metachromatic substances	toluidine blue
Bone	formaldehyde	Goldner's trichrome stain	iron hematein, acid fuchsin, ponceau 2 R, light green SF
	formaldehyde		alcian blue 8 GX, basic fuchsin
	formaldehyde	Dahl's calcium stain	alizarin red S
Muscle			
Smooth	formaldehyde	Goldner's trichrome stain	alizarin red S
Skeletal	formaldehyde	Delafield's hematein	hematein
	formaldehyde	Goldners's trichrome stain	hematein
Heart	Susa	Heidenhain	azocarmine B, aniline blue
Nerve tissue			
Soma	formaldehyde	Nissl stain	toluidine blue O
Axons		vital staining	methylene blue
	Müller's fixative	Gerlach	carmine
Myelin	formaldehyde	Klüver – Barrera	luxol fast blue MBSN
	formaldehyde	Weil	hematein
Neuroglia	formaledehyde	Holzer	crystal violet

∗ H & E = hematoxylin – eosin.

ing in exfoliative cytology, and in the staining of nucleic acids. Differentiation of the nucleic acids (DNA green, RNA red) can also be achieved with acridine orange. Another cationic fluorochrome, atebrin (no C.I. number), exhibits green fluorescence. It is used in the staining of mitotic chromosomes, in the demonstration of Y chromatin in interphase nuclei, and for blood components and blood protozoa.

5.2. Acid or Anionic Stains

Acid or anionic stains indicate positively charged cytoplasmic and extracellular material. Anionic stains are also used in acidic solution. Most tissue structures are stained nonspecifically if anionic stains are used alone. Histological counterstaining is one of the most frequent applications of anionic stains, the most common stain being eosin (eosin Y, C.I. 45 380,

Table 4. Staining procedures in histochemistry

Substrate	Tissue	Dyes and stains used
Enzymes		
Aminopeptidase	cryostat sections	fast blue B
Dehydrogenase	cryostat sections	nitro blue tetrazolium
α-Naphthyl acetate esterase	formalin-fixed tissue	fast blue B
Peroxidase		
Acid phosphatase	formalin-fixed tissue	hexazonium – pararosaniline
Alkaline phosphatase	formalin-fixed tissue	fast blue RR
Amyloid	formalin-fixed tissue	congo red
Carbohydrates		
Glycogen	ethanol-fixed tissue	carmine, Schiff's reagent (PAS)
Mucopolysaccharides		
Acid mucopolysaccharides	Bouin fixed tissue	alcian blue 8 GX
Neutral mucopolysaccharides	formalin-fixed tissue	Schiff's reagent (pararosaniline)
Keratin	ethanol in cytology	orange G (Papanicolaou stain)
	formalin-fixed tissue	cresyl fast violet hematein
Lipids		
Neutral lipids	formalin-fixed tissue	Sudan red 7 B, Sudan black B, oil red O
Phospholipids	formalin-fixed tissue	Sudan black B
Nucleic acids		
DNA	formalin-fixed tissue	Schiff's reagent (pararosaniline), gallocyanine chromalum, methyl green – pyronin Y
RNA	formalin-fixed tissue	gallocyanine chromalum, methyl green – pyronin Y
Proteins		
Proteins	formalin-fixed tissue	brilliant blue R, fast blue B
Basic histone proteins	formalin-fixed tissue	fast green FCF
Inorganic substrates		
Calcium	formalin-fixed tissue	alizarin red S

Table 5. Staining procedures in microbiology

Microorganisms	Fixation	Staining technique	Dyes used
Bacteria	heat	Löffler, Gram	methylene blue, crystal violet – safranine
Corynebacterium diphtheriae	heat	Neisser	methylene blue, crystal violet, Bismarck brown Y
Mycobacterium tuberculosis	heat	Ziehl – Neelsen	carbol fuchsin, methylene blue
M. leprae	heat	Ziehl – Neelsen	carbol fuchsin, methylene blue
Spores	heat	Rakette	malachite green – eosin Y, safranine
Fungi	formaldehyde	Hotchkiss – McManus	Schiff's reagent – light green SF
Protozoa	dry smears (methanol)	Giemsa	azure B, eosin Y
Viruses	formaldehyde	Herzberg	Victoria blue 4 R

Table 6. Staining procedures in fluorescence microscopy

Specimen	Fixation	Staining technique	Dyes used
Bacteria	heat	Strugger	acridine orange
Protein	formalin	immunofluorescence	fluorescein isothiocyanate
Lipids	formalin	Smith	Nile blue A
DNA	intravital	Pearse	acridine orange
Mitochondria	intravital		acridine orange
Fungi	intravital		acridine orange
Mycobacterium tuberculosis	heat	Truant	auramine O – rhodamine B

eosin B, C.I. 45 400). Eosin is bound to the substrate by ionized cationic groups of proteins–usually an amino group of lysine or a guanidino group of arginine, amino acids that are present in almost all proteins. Differential staining effects are achieved only if two or more stains are applied either simultaneously or consecutively. For example, in a mixture of methylene blue and eosin, nuclei, cartilage, and collagen are stained blue, while the cytoplasm is stained red by the eosin. The various techniques available for staining collagen utilize mixtures of stains drawn from one or two stain groups. A mixture of two anionic stains is employed in van Gieson staining: one colors collagen and the other the cytoplasm, as well as muscle fibers and erythrocytes. In trichrome staining, two, three, or (less often) four anionic stains are applied in conjunction with phosphoric or phosphomolybdic acid. This causes collagen, cytoplasm, and erythrocytes all to be stained differently. Elements such as cartilage, fibrin, and secretory granules also acquire characteristic colors. Anionic stains used for marking collagen can also be empolyed to some extent in staining polysaccharides. Neutral polysaccharides, such as glycogen, may be revealed by staining according to the Best's carmine technique. Amyloid can be stained with Congo red (C.I. 22 120) and other direct techniques. Congo red is also capable of simultaneously staining amyloid and hydrophobic constituents (e.g., elastic fibers). Elastin is normally stained only slightly with anionic stains (e.g., eosin) but strongly with compounds such as orcein [C.I. (1st ed.) 1242] or aldehyde fuchsin.

Examples of common anionic stains are orange G (C.I. 16 230) and metanil yellow (C.I. 13 065). These azo dyes are used in combination with other anionic stains to stain cytoplasm. Biebrich scarlet (C.I. 26 905), another azo dye, has many applications (e.g., counterstaining, collagen staining, or selective protein staining in alkaline solution).

5.3. Mordant Stains

Mordant stains are used primarily in the staining of nuclei. Well-known representatives of this group are the natural dyes hematein and brazilein, which are derived from hematoxylin (C.I. 75 290) and brazilin (C.I. 75 280), respectively, by oxidation. Hematein is the active component of hematoxylin, and is formed by partial oxidation, also called maturation. Oxidation can take place either slowly through contact with atmospheric oxygen or by addition of an oxidizing agent. Hematein staining involves the use of aluminum or iron as a mordant.

Staining is partially prevented by chemical blockage from lysine and arginine residues. For good nuclear staining it is essential that the acidified solution contains more metal ions than hematein. Although both progressive and regressive staining are possible with aluminum hematoxylin, only progressive staining is appropriate with iron hematoxylin (e.g., Weigert's hematoxylin). In histochemical and other special staining techniques it is preferable that the nucleus is stained red. This can be done with an aluminum mordant and carminic acid (carmalum), aluminum mordant and brazilein (brazalum), or aluminum mordant and nuclear fast red (C.I. 75 470).

5.4. Oil-Soluble Stains

Oil-soluble stains (*lipochromes*) are used for demonstrating the presence of fats and lipids. Colored, nonpolar compounds dissolve in tissue and cellular lipids, causing them to become microscopically visible. Fat stains are effective for coloring lipids because they are more soluble in the lipids than in their solvents. The first soluble stains used for the coloration of fats were Sudan III (C.I. 26 100), Sudan IV (C.I. 26 105), and oil red O (C.I. 26 125). These stains dissolve in hydrophobic fats that are liquid at the temperature at which staining is carried out. Sudan black B (C.I. 26 150) is a suitable stain for all lipids. Lipochromes dissolve in lipids only at temperatures above the lipid melting point. Lipids containing unsaturated fatty acids melt above 60 °C, so preliminary treatment with bromine water is required prior to staining with Sudan black B. Certain lipoproteins in which the lipids are "masked" by proteins must be subjected to acid hydrolysis before staining.

5.5. Chromogenic Reactions

The most important chromogenic reactions are the Feulgen reaction for staining nuclear DNA

and the periodic acid Schiff (PAS) reaction for staining of tissue polyssaccharides.

With both techniques aldehyde groups are stained which are released by acid hydrolysis of DNA in case of the Feulgen reaction and by the oxidation of polyssaccharides for PAS. Both procedures use the Schiff reagent which is an aqueous solution of basic fuchsin or pararosanilin.

A standard feulgen protocol for reproducible quantitative staining of DNA for image cytometry has been recommend by the European Society for Analytical Cellular Pathology, ESACP [43–48]

6. Enzyme Histochemical Reactions

An enzyme is a protein that functions as a catalyst. Catalysis is triggered by the binding of an enzyme molecule to a molecule of one of the reactants, which acts as substrate. The substrate may be either a neutral organic compound or an ion. An important feature of enzymatic catalysis is that it proceeds at room temperature and approximately neutral pH. Various histochemical methods for staining enzymes are used in conjunction with blood and bone marrow smears to identify normal and abnormal leukocytes, and for differential staining of B and T lymphocytes. Enzymes display high specificity for their substrates and for the particular reactions they catalyze. For example, phosphatases catalyze the hydrolysis of esters of phosphoric acid. *Acid phosphatases* are typical lysosomal constituents, and can be demonstrated in the cells of kidney tubules and the prostate gland, as well as in phagocytes and degenerated cells. *Alkaline phosphatases* are present in the cytoplasm of leukocytes (neutrophils and basophils) and their precursors, but not in myeloblasts. In myeloid leukemias, the abnormal circulating leukocytes contain little or no alkaline phosphatase. The fact that this enzyme catalyzes hydrolysis of the phosphate of β-hydroxynaphthoic anilide (Naphthol AS) at high pH is utilized for its detection. The Naphthol AS molecule that is released couples immediately with a diazonium salt present in the incubation medium, giving rise to an insoluble azo dye.

Various *esterases* that catalyze the hydrolysis of α-naphthyl acetate can also be detected histochemically. These are present in both monocytes and T cells. The enzymes themselves are quite heat resistant and can be detected in material even after fixation. In this technique the α-naphthol formed on hydrolysis is coupled with diazotized pararosaniline. The final product is a stain composed of one pararosaniline molecule and three molecules of α-naphthol.

Dehydrogenases are detected by histochemical reactions with various tetrazolium salts. Tetrazolium salts are heterocyclic compounds that are converted to insoluble, colored formazans by reduction. The most frequently used reagent is nitro blue tetrazolium chloride (nitro BT); the resulting formazan is insoluble in fats and can bind to proteins.

Peroxidases catalyze the oxidation of various substances by hydrogen peroxide, including reduced coenzymes, fatty acids, amino acids, reduced cytochromes, and many other compounds. This reaction leads to an oxidized donor and two water molecules starting from a donor molecule and a molecule of hydrogen peroxide. Many organic compounds (e.g., amines, phenols, and the leuco compounds of dyes) can serve as donors. Peroxidase activity can be detected in the granules of myeloid leukocytes, in various neurons, and in secretory cells. The peroxidase histochemical method is based on the catalytic reaction between hydrogen peroxide and substances that yield insoluble products after oxidation. One of the oldest techniques utilizes benzidine as the donor, which is introduced along with hydrogen peroxide into the incubation medium, causing the benzidine to be oxidized to a blue substance. The most commonly used donor today is 3,3′-diaminobenzidine tetrahydrochloride (DAB). Spontaneous oxidation of this amine is very limited, but an insoluble, amorphous, brown substance is formed rapidly in the presence of peroxidase. This unstable substance reacts immediately with more DAB to form polymers containing the quinone and indamine chromophores. The reaction can be made more sensitive by substituting tetramethylbenzidine (TMB) as the donor, which leads to a blue product. The activities of certain enzymes (e.g., cytochrome oxidase and catalase) can be responsible for false positive results in the histochemical peroxidase reaction.

7. Immunohistochemical Staining Methods

The objective of immunohistochemical techniques is the detection and identification of antigens in cells and tissue sections by means of specific antibodies. Both *monoclonal* and *polyclonal* antibodies can be used for routine applications. Apart from appropriate specificity and low cross-reactivity, one of the main requirements for a suitable antibody is its ability to recognize antigens in tissues that have been fixed in formalin or embedded in paraffin. This requirement is frequently not met by antibodies against glycoproteins, which are diagnostically quite important. When antigen molecules encounter molecules of their antibodies, *antigen – antibody complexes* may be formed. The interaction of an antibody with its antigen can be revealed very specifically by suitable labeling of one of the reaction partners (e.g., with fluorescent dye or an enzyme). The detection of antibody that is bound to an antigen (*primary antibody*) can be accomplished in various ways, and by use of a marker coupled to the primary antibody either directly or indirectly through selected ligands. The function of the marker substance is to make the antigen – antibody complex visible. The most common marker system in immunohistochemical applications is horseradish peroxidase, which is susceptible to several histochemically stable, high-contrast color reactions. Thus, peroxidase in the presence of hydrogen peroxide is used to oxidize soluble, colorless chromogens to insoluble, colored end products.

In the *direct fluorescence antibody method*, labeled antibody is reacted with its tissue antigen, which is present in reactive form. Conditions are selected such that the antibody, which carries a fluorescent label, binds strongly only to the antigen. In *indirect* immunofluorescence, a fluorescent immunoglobulin is caused to react with specific antibodies as a way of detecting the presence of antigens in tissue. Fluorochromes used for these purposes include fluorescein isothiocyanate (FITC), tetramethylrhodamine isothiocyanate (TRITC), and lissamine rhodamine B 200 sulfonyl chloride (RB 200 SC). Excitation in the UV or blue region of the spectrum results in yellowish green (FITC) or red (TRITC or RB 200 SC) fluorescence.

The principle of the *direct enzyme labeling methods* is analogous to that of direct immunofluorescence. In this application enzymes are used as antibody markers in place of fluoro dyes. The enzymes employed are horseradish peroxidase, alkaline phosphatase, glucose oxidase, and galactosidase. The enzymes are developed with various substrates (chromogens) in such a way that differently colored products are formed, thereby revealing visually the presence of antigen – antibody complex and, consequently, a local immune response. In the case of peroxidase, the most frequently applied marker enzyme, the chromogens used are 3,3′-diaminobenzidine and 3-aminoethylcarbazole (AEC). DAB reacts with H_2O_2 to give a brown product that is insoluble in alcohol, whereas AEC gives a red, alcohol-soluble product. If alkaline phosphatase is used as the marker enzyme, the color reaction is produced, for example, by Naphthol AS MX phosphate together with fast red (red) or fast blue (blue). Enzymes can be coupled directly to antibodies (conjugates), or unconjugated methods can be applied, in which case the marker enzyme is fixed at the site of the antigen – antibody reaction through (antienzyme) antibodies. Conjugated methods are the easiest to carry out, but they have the disadvantage of lower sensitivity.

The indirect method is analogous to indirect immunofluorescence, in that it uses an enzyme as the marker. In the unconjugated antibody – enzyme-bridge method, a bridge-forming antibody is linked between an antibody directed against a particular marker enzyme and an unconjugated primary antibody directed against the antigen to be detected. The primary antibody and the antibody directed against the marker enzyme must originate from the same animal species, and the bridging antibody is also directed against immunoglobulins from this species. In order to detect the immune reaction (and antigen in the tissue), enzyme is added, this becomes bound by the enzyme antibody, and the combination is then detected histochemically.

A modification of the above method is the *enzyme – antienzyme complex technique*, in which antienzyme antibody and marker enzyme are added not successively, but in the form of a soluble enzyme – antienzyme complex. The peroxidase – antiperoxidase complex (PAP complex) and the alkaline

phosphatase – antialkalinephosphatase complex (APAAP complex) have proved particularly useful. The PAP complex is usually employed in the case of polyclonal antibodies from rabbit and goat, and the APAAP complex in monoclonal antibodies from mouse. With a monoclonal primary antibody of this type from mouse and an APAAP complex formed with monoclonal (mouse) antibodies against alkaline phosphatase, an antimouse immunoglobulin antibody is used as the bridging antibody. The concentration of the marker enzyme at the site of the antigen (hence the sensitivity of the test) can be increased by repeated incubation with the complex (multistep method).

The *(strept)avidin – biotin method* is based on the ability of the egg-white glycoprotein avidin to physically bind four molecules of the vitamin biotin. Avidin, which is isolated from chicken egg whites, can be replaced by streptavidin, a protein produced by *Streptomyces avidinii*. As in the PAP or APAAP method described above, three reagents are required. The first is the primary antibody, directed specifically against the antigen to be determined. Next is a secondary antibody, which can bind to the first, and this is conjugated with biotin. The third reagent is either an enzyme coupled to (strept)avidin or a preformed (strept)avidin – biotin – peroxidase complex. Free sites on the avidin molecule permit binding between the biotin and the secondary antibody. The enzyme (and therefore the desired antigen) is made visible with a suitable chromogen. Since several complexes can bind to a single avidin molecule, it is possible to increase the peroxidase concentration at the antigen site, thereby increasing the sensitivity of the method.

8. References

General References

1. E. N. Abrahart: *Dyes and Their Intermediates,* Pergamon Press, Oxford 1968.
2. R. L. M. Allen: *Colour Chemistry,* Nelson, London 1971.
3. J. D. Bancroft, H. C. Cook: *Manual of Histological Techniques,* Churchill – Livingstone, Edinburgh 1984.
4. J. D. Bancroft, A. Stevens: *Theory and Practice of Histological Techniques,* 2nd ed., Churchill – Livingstone, Edinburgh 1982.
5. P. Böck: *Romeis Mikroskopische Technik,* Urban & Schwarzenberg, München 1988.
6. G. Clark: *Staining Procedures used by the Biological Stain Commission,* 4th ed., Williams and Wilkins, London 1981.
7. G. Clark, F. H. Kasten: *History of Staining,* 3rd ed., Williams & Wilkins, London 1983.
8. Y. D. Clonis, T. Atkinson, C. J. Bruton, C. R. Lowe (eds.): *Reactive Dyes and Enzyme Technology,* Macmillan, London 1987.
9. H. C. Cook: *Manual of Histological Demonstration Methods,* Butterworth, London 1974.
10. R. A. B. Drury, E. A. Wallington: *Carleton's Histological Technique,* 5th ed., Oxford University Press, Oxford 1989.
11. M. T. Elghetany, A. Saleem, *Stain Technol.* **63** (1988) 201 – 212.
12. J. Elias: *Principles and Techniques in Diagnostic Histopathology,* Noyes Publications, Park Ridge 1982.
13. F. J. Green: *The Sigma-Aldrich Handbook of Stains, Dyes and Indicators,* Aldrich Chemical Company, Milwaukee 1990.
14. E. Gurr: *The Rational Use of Dyes in Biology and General Staining Methods,* Leonard Hill, London 1965.
15. E. Gurr: *Staining Animal Tissues, Practical and Theoretical,* Leonard Hill, London 1962.
16. H. Harms: *Handbuch der Farbstoffe für die Mikroskopie,* Staufen-Verlag, Kamp-Lintfort 1965.
17. R. W. Horobin: *Histochemistry, An Explanatory Outline of Histochemistry and Biophysical Staining,* Gustav Fischer, Stuttgart 1982.
18. R. W. Horobin: *Understanding Histochemistry: Selection, Evaluation and Design of Biological Stains,* Ellis Horwood, Chichester 1988.
19. L. Kass, *Stain Technol.* **65** (1990) 211 – 230.
20. J. A. Kiernan: *Histological and Histochemical Methods, Theory and Practice,* Pergamon Press, Oxford 1990.
21. R. D. Krug, P. A. Takes, *J. Histotechnol.* **14** (1991) 31 – 34.
22. R. D. Lillie: *H. J. Conn's Biological Stains,* 9th ed., Williams and Wilkins, Baltimore 1977.
23. R. D. Lillie, H. H. Fullmer: *Histopathologic Technique and Practical Histochemistry,* 4th ed., McGraw-Hill, New York 1976.
24. H. Lyon (ed.): *Theory and Strategy in Histochemistry. A Guide to the Selection and Understanding of Techniques,* Springer Verlag, Berlin 1991.

25. H. Lyon, E. K. W. Schulte, P. Hoyer, *Histochem. J.* **21** (1989) 508 – 513.
26. A. G. E. Pearse: *Histochemistry, Theoretical and Applied,* 4th ed., Churchill – Livingstone, Edinburgh 1985.
27. J. M. Polak, S. van Noorden: *Immunocytochemistry, Modern Methods and Application,* 2nd ed., Wright, Bristol 1986.
28. G. B. Proctor, R. W. Horobin, *Stain Technol.* **63** (1988) 101 – 111.
29. E. A. Schenk, E. H. Stotz, *Stain Technol.* **63** (1988) 65 – 68.
30. E. K. W. Schulte, *Analyt. Cellul. Pathol.* **3** (1991) 167 –182.
31. E. K. W. Schulte, *Histochemistry* **95** (1991) 319 – 328.
32. E. K. W. Schulte, H. Lyon, P. Prento, *Histochem. J.* **23** (1991) 241 – 245.
33. E. K. W. Schulte, D. H. Wittekind, *Stain. Technol.* **64** (1989) 253 – 256.
34. E. K. W. Schulte, D. H. Wittekind, *Histochemistry* **91** (1989) 321 – 331.
35. E. K. W. Schulte, D. H. Wittekind, *Analyt. Quant. Cytol. Histol.* **11** (1989) 131 – 139.
36. E. K. W. Schulte, D. H. Wittekind, *Analyt. Quant. Cytol. Histol.* **12** (1990) 149 – 156.
37. E. K. W. Schulte D. H. Wittekind, V. Kretschmer, *Histochemistry* **12** (1988) 427 – 433.
38. D. C. Sheehan, B. B. Hrapchak: *Theory and Practice of Histotechnology,* 2nd ed., C. V. Mosby Co, St. Louis 1980.
39. Society of Dyers and Colourists and the American Association of Textile Chemists and Colorists (eds.): *Colour Index,* 3rd ed., Lund Humphries, London 1971.
40. R. R. Tubbs, G. N. Gebhardt, R. E. Petras: *Atlas of Immunohistology,* American Society of Clinical Pathologists Press, Chicago 1986.
41. L. L. Vacca: *Laboratory Manual of Histochemistry,* Raven Press, New York 1985.
42. L. Wagner, C. P. Worman, *Stain Technol.* **63** (1988) 129 – 136.
43. H. Lyon, E. Schulte, B. de Leenheer, *Histochem. J.* **24** (1992) 217 – 219.
44. H. Lyon, E. Schulte, B. de Leenheer, *Histochem. J.* **24** (1992) 233 – 235.
45. F. L. Carson: *Histotechnology,* ASCP Press, Chicago 1990.
46. R. C. Taylor, *Biotechnic and Histochemistry* **67** (1992) 323 – 326.
47. R. C. Taylor, *Biotechnic and Histochemistry* **67** (1992) 110 – 117.
48. A. Böcking, F. Giroud, A. Reith, *Analytical Cellular Pathology* **8** (1995) 67 – 74.

Starch

James R. Daniel, Department of Foods and Nutrition, Purdue University, West Lafayette, Indiana 47907, United States (Chap. 1 and 2)

Roy L. Whistler, Whistler Center for Carbohydrate Research, Purdue University, West Lafayette, Indiana 47907, United States (Chap. 1 and 2)

Harald Röper, Solingen, Germany (Chap. 3)

Barbara Elvers, Hamburg, Germany

1. Native Starch 1
1.1. Raw Materials 2
1.2. Molecular Structure and Composition 2
1.3. Biosynthesis and Structure 2
1.3.1. Biosynthesis of Amylose, Amylopectin and Involved Enzymes . 2
1.3.2. Structure of the Starch Granule 2
1.4. Physicochemical Properties 3
1.4.1. X-Ray Patterns 4
1.4.2. Rheological Properties 5
1.4.3. Water Activity 6
1.4.4. Amylose Complexes 6
1.5. Composition of Native Starches . . . 7
1.6. Industrial Starch Production Processes 7
1.6.1. Corn and Sorghum Starch 7
1.6.2. Wheat Starch 8
1.6.3. Potato Starch 9
1.6.4. Rice Starch 10
1.6.5. Tapioca Starch 10
1.6.6. Arrowroot Starch 11
1.6.7. Sago Starch 11
1.7. Uses 11
2. Modified Starches 11
2.1. Physically Modified Starches 11
2.2. Chemically Modified Starches 12
2.2.1. Acid-Modified Starch 12
2.2.2. Oxidized Starch 14
2.2.3. Cross-Linked Starch 15
2.2.4. Starch Esters 17
2.2.5. Starch Ethers 19
2.2.6. Cationic Starch 20
3. Economic Aspects 21
4. References 22

1. Native Starch

Starch [*9005-25-8*], the principal source of dietary calories to the world's human population, has many chemical and physical characteristics that set it apart from other food components and give it numerous applications. Food processors and other industrial users should be familiar with starch's structural and behavioral characteristics so as to make use of its properties. The chemistry and technology of starch has been reviewed in detail [1].

Unlike other carbohydrates and edible polymers, starch occurs as discrete particles called starch granules. These particles are unique to the plant source and, therefore, starch from every plant type is different in appearance, properties, and particle size distribution. The size distribution of wheat starch granules during endosperm development has been reported [2]. Microscopists can readily identify the plant source of most commercial starches.

Starch in granular form is generally composed of two types of molecules, amylose and amylopectin. *Amylose* [*9005-82-7*] is a predominantly linear (1,4)-α-D-glucan (Fig. 1), some amyloses possessing a few 1,6 branches. *Amylopectin* [*9037-22-3*] (Fig. 2) is a branched, bushlike structure containing both (1,4)-α-D linkages between D-glucose residues and (1,6)-α-D branch points. Normal starches contain ca. 75 % amylopectin molecules. Amylopectin is a much larger molecule than amylose as shown in Table 1. It is these characteristic molecular sizes, amylose – amylopectin ratios, and granular structures which give each type of starch its unique properties.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_001.pub2

Table 1. Physicochemical properties of amylose and amylopectin

Property	Amylose	Amylopectin
Molecular mass	50 00 – 200 000	one to several million
Glycosidic linkages	mainly (1,4)-α-D-	(1,4)-α-D-, (1,6)-α-D-
Susceptibility to retrogradation	high	low
Products of action of β-amylase	maltose	maltose, β-limit dextrin
Products of action of glucoamylase	D-glucose	D-glucose
Molecular shape	essentially linear	bush-shaped

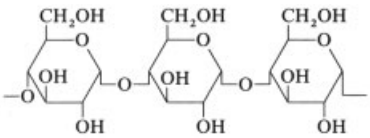

Figure 1. Representative structure of linear amylose

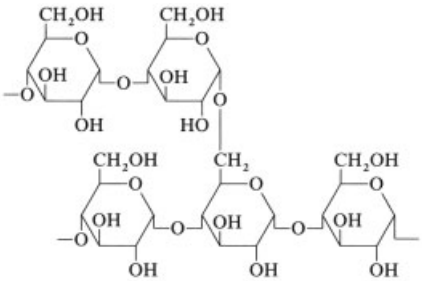

Figure 2. Representative structure of amylopectin, including (1,6)-α-D branch point

1.1. Raw Materials [1, 3 – 5]

Plants grown for starch production include corn, sorghum, wheat, rice, tapioca, arrowroot, sago, and potato. Other noncommercial sources of starch include quinoa and amaranth, whose starches have been isolated and examined with respect to chemical and physical characteristics and possible uses [6 – 9].

1.2. Molecular Structure and Composition

The molecular structure and composition of amylose, amylopectin, and starch granules is largely described in Sections 1, 1.3.2, 1.4.1, 1.4.4, and 1.5 as well as Tables 1–4. Purified amylose forms strong, transparent films and fibers, having properties similar to cellulose [*9004-34-6*], except for the greater swellability of amylose and its alkaline solubility. Amylose molecules have degrees of polymerization (DP) of 350 – 1000 compared to amylopectin which has a DP of several thousand.

Table 2. Diameter and gelatinization temperature of starch granules

Source	Mean diameter, µm	Gelatinization temperature, °C
Corn	15	62 – 71
Wheat	20 – 22	53 – 64
Rice	5	65 – 73
White potato	33	62 – 68
Sweet potato	25 – 50	82 – 83
Tapioca	20	59 – 70

1.3. Biosynthesis and Structure

1.3.1. Biosynthesis of Amylose, Amylopectin and Involved Enzymes

Normal starch, containing both amylose and amylopectin, is biosynthesized by two enzymes, a *chain-lengthening enzyme,* which is apparently identical for both amylose and the linear portions of amylopectin chains, and a *branching enzyme* that gives rise to amylopectin. Evidence suggests that a branching enzyme requires an amylose chain of ca. 35 – 40 units before it can transfer a portion of that chain to form a branch on another starch molecule.

When starch granules are examined under normal light or by scanning electron microscopy, growth rings are visible around the hilum (see below) which may become wider in successive layers as the granule grows. Some evidence points to the beginning of amylopectin molecules at an inner lamella edge and extending to the outer edge.

1.3.2. Structure of the Starch Granule

Individual starch molecules that are formed by the action of starch synthetase nucleate at a point called the *hilum* (the botanical center of the granule). These starch molecules are produced and deposited in a radial, spherocrystalline arrangement. Synthetase enzymes cling to the surface of

starch granules although there is no membrane. The surface consists of end units of starch molecular chains deposited side-by-side during granular growth. Linear amylose molecules occur interspersed among the branched amylopectin molecules. Linear amylose molecules tend to form helices which are often double. The amylopectin molecules are thought to exist in a so-called "tassel-on-a-string" structure (see Fig. 3) [10 – 14]. In this structure short chains of 12 – 15 D-glucopyranosyl units occur about every 25 units of the main chain, which may consist of more than a thousand D-glucopyranosyl units.

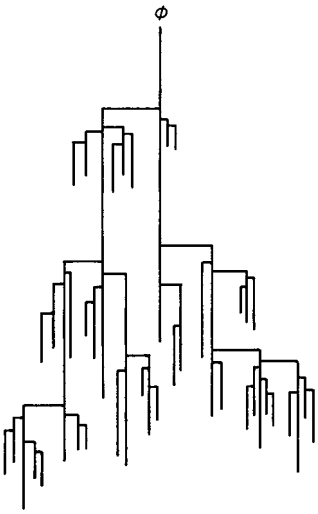

Figure 3. Tassel-on-a-string model of amylopectin (ϕ = reducing end group)

The spherocrystalline radial arrangement of starch molecules in the granule is inferred from the Maltese cross that is seen on microscopic examination of starch granules under polarized light [1] (see Fig. 4). The center of the cross is at the hilum. X-Ray patterns of starch granules show that cereal starches possess an *A pattern*, which indicates chains of amylose or outer chains of amylopectin in antiparallel double helices separated by interstitial water. Tuber and root starches produce *B patterns,* with up to 30 % water present in sheets and columns. Legume starches possess C X-ray patterns, which are a mixture of A and B patterns. These patterns result primarily from the otherwise uncomplexed amylose molecules or long outer chains of amylopectin. Amylopectin molecules seem to be oriented with their reducing ends interior to the granule, as would be expected from outward biosynthetic growth. In addition to X-ray evidence, crystallinity is also evident from measurements by differential scanning calorimetry. This method measures the energy for melting crystalline regions when heated [15].

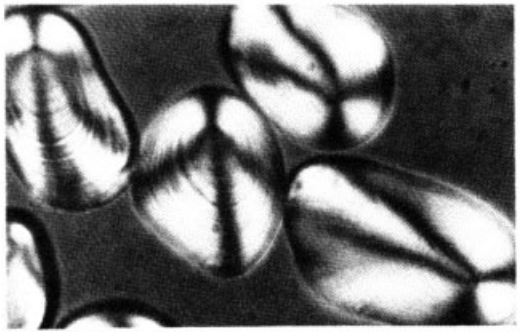

Figure 4. Micrograph of potato starch in polarized light (magnification ca. ×250)

1.4. Physicochemical Properties

Undamaged starch granules are insoluble in cold water, but can reversibly absorb water and swell when their aqueous dispersion is heated. In water at 25 °C, the percentage increase in granule diameter ranges from 9.1 % for normal (native) corn starch to 22.7 % for waxy (≈100 % amylopectin) corn starch. As temperature is increased, the molecules in the starch granule vibrate vigorously, breaking intermolecular bonds, thus allowing increased interaction with water molecules. This penetration of water combined with the increased separation of the starch chains increases randomness and decreases crystallinity in the granule. Continued heating in the presence of free water results in complete loss of crystallinity, as evident by loss of birefringence of the starch granules and changes in the X-ray diffraction pattern. The temperature range over which birefringence disappears is called the *gelatinization temperature range* (see Table 2). Gelatinization usually occurs over a small temperature range with larger starch granules generally gelatinizing first

and smaller granules later. During gelatinization, granules swell extensively (Fig. 5). Thus, a 1 % starch slurry in cold water has low viscosity, but on heating a thick paste is produced wherein almost all the water has entered the granules and combined with them. This causes the granules to press tightly against each other. The viscosity of the paste results from resistance to the flow of the swollen granules that now occupy nearly the entire volume. Highly swollen granules are fragile and disintegrate on mild stirring. This causes a significant decrease in paste viscosity.

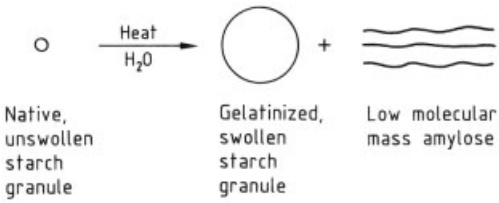

Figure 5. Representation of starch gelatinization process

The swelling and gelatinization of wheat, barley, and maize starch granules have been studied with respect to the effect of amylose, amylopectin, and lipid content [16]. A similar investigation has been performed with rice starch [17]. In the case of wheat, barley, and maize starch it was concluded that granular swelling during gelatinization was mainly a function of amylopectin content. Gelatinization has been studied by scanning electron microscopy and a honeycomb-like structure has been noted on gelatinization [18]. Visual characteristics of starch pastes have been investigated, concentrating specifically on clarity and whiteness [19]. Potato starch paste exhibited the greatest clarity, wheat and corn were intermediate in clarity, and high amylose corn paste was of very low clarity.

Starch swelling rates are greatly increased at pH 10.0 and above. At low pH, the peak viscosity (see Section 1.4.2) of starch pastes is markedly reduced, along with a rapid loss of viscosity on cooking. This is due to extensive hydrolysis of starch to form nonthickening dextrins (low molecular mass oligosaccharides derived from starch).

1.4.1. X-Ray Patterns

The linear nature of corn [20] and potato [21] amylose was proposed in 1940 and confirmed [22] three years later. However, evidence now exists that some amyloses may contain a few very long branches.

X-ray diffraction has been useful in examining the structure of both intact starch and amylose. Starch naturally occurs in three crystalline modifications designated A (cereal), B (tuber), and C (smooth pea and various beans). Starch which is precipitated from solution, or complexed with organic molecules, adopts the so- called *V structure* (Verkleisterung). Amylose also exists in A, B, C, and V structures. Marchessault and coworkers [23] studied the chain conformation of B-amylose and proposed that *B-amylose* is helical with an integral number of α-D-glucopyranosyl residues per turn. The density of B-amylose suggests a helical structure with six α-D-glucopyranosyl residues per unit cell and 3 – 4 molecules of water of hydration. Sixfold helices are energetically favored and have a hydrogen bond between the C-2 hydroxyl group of one α-D-glucopyranosyl unit and the C-3 hydroxyl group of the succeeding sugar unit. Left-handed helices are slightly favored over right-handed ones. It was thus concluded that solid-state B-amylose exists as a left-handed sixfold helix. Amylose triacetate, which exists in a left-handed helix, forms B-amylose upon saponification [24]. Calculations show that *V-amylose,* having a fiber repeat of about 80 pm, is more stable as a left-handed helix than as a right-handed one [25]. The easy conversion of V-amylose to B-amylose [26] implies that no extensive molecular reorganization accompanies the reversal of chain chirality. It seems likely that amylose, in the solid state, exists as a left-handed helical polymer with six α-D-glucopyranosyl units per turn.

A model proposed by Kreger [27] in which amylose exists as a threefold helix has been examined [28]. Although this model is consistent with some of the amylose X-ray data, it has not received wide acceptance. It should be noted that the model of the starch granule of Frey – Wyssling [28] employs radially oriented low-molecular-mass amylose as a threefold helix. This explains the granule's positive birefringence and proposes a higher-molecular-

mass amylose (DP > 280) as a sixfold helix running tangentially to the granule surface.

1.4.2. Rheological Properties

For thickening foods, sizing and coating papers, textile sizing, use in drilling muds, adhesive formulations, and many other applications, starch gelatinization is required. When starch is cooked, flow behavior of the slurry changes significantly as the suspension becomes a mixture of swollen granules, partially disintegrated granules, and molecularly dispersed granule contents (mostly amylose). The cooked product is termed a starch paste and transition from a granular suspension to a cooked paste is accompanied by an increase in apparent viscosity. Although the term viscosity is generally used, starch pastes are non-Newtonian and a better description is that pastes develop resistance to deformation or flow.

On cooling, paste consistency increases because side-by-side molecular associations form junction zones that increase paste resistance to deformation or flow. Therefore, when cooled, pastes may remain fluid or form gels with considerable strength. Starch manufacturers and users both require techniques for judging starch quality and flow behavior in its applications. For these reasons measurements of viscosity, gel rigidity, and gel strength have been devised.

A *Brabender viscoamylograph* provides information on gelatinization as well as on the properties of the cooled paste. This instrument records the torque needed to counteract the viscosity that develops when a starch suspension is subjected to heating and cooling regimes. The temperature at which the first significant rise in viscosity occurs depends on starch concentration. It is usually higher than the gelatinization temperature that is determined by loss of birefringence. Viscosity is measured in arbitrary units (Brabender units) that reflect paste consistency; for this reason, Brabender results are described in terms of paste properties and pasting temperatures at specific times. The Amylograph Handbook describes the theory, construction, applications, sources of error, troubleshooting, and adjustment of Brabender units [29].

A typical Brabender curve is shown in Figure 6. Six important points on the Brabender curve are generally recognized:

1) *Pasting Temperature.* Indicates onset of paste formation; varies with starch type or modification and with additives in the starch slurry.
2) *Peak Viscosity.* Cited regardless of the temperature at which the peak is attained.
3) *Viscosity at 95* °C. Related to ease of cooking the starch.
4) *Viscosity after 1 h at 95 °C.* Shows cooked paste stability, or lack of stability under relatively low shear.
5) *Viscosity at 50* °C. Measures the setback (increase in viscosity) that occurs when the hot paste is cooled.
6) *Viscosity after 30 min at 50* °C. Related to stability of cooked paste under simulated use conditions.

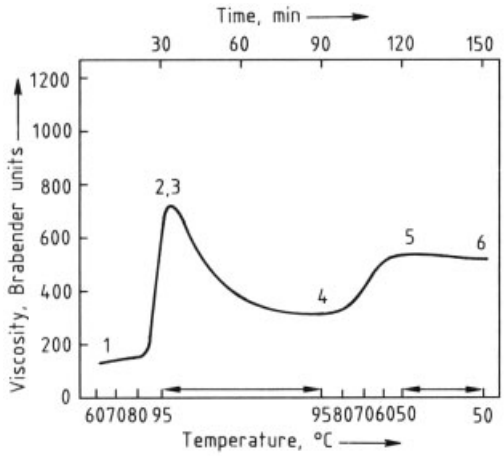

Figure 6. Representative Brabender amylograph curve
For explanation of points 1 – 6 text

Other chemicals in the starch dispersion may have a significant effect on gelatinization and gelation. This is especially true in the use of starch in foods. High *sugar* concentrations lower the rate of starch gelatinization, the peak viscosity, and the gel strength. Disaccharides are more effective in delaying gelatinization and reducing peak viscosity than monosaccharides [30]. Sugars may decrease gel strength by exerting a plasticizing action and interfering with formation of starch junction zones. *Lipids,* such as the triacylglycerol fats and oils, and lipid-related materials, such as mono- and diglycerides may also affect starch gelatinization (Section 1.4.4). Wheat starch contains up to 1 % lipids consisting

mainly of lysophosphatidylcholines of palmitic and linoleic acids which are principally complexed with amylose. Such complexes retard starch hydration and lower peak viscosity on gelatinization. These complexes also are more slowly hydrolyzed by starch amylases.

Because of the essentially neutral character of starch, low concentrations of *salts* have little effect on gelatinization and gelation. Exceptions include potato amylopectin, which contains some phosphate groups, and manufactured ionic starches. With these starches salts may either increase or decrease swelling, depending on conditions [30].

Starch-thickened *acidic* foods (pie fillings, salad dressings) usually have pH values in the range 4 – 7 which has little effect on starch gelatinization. At lower pH values starch hydrolysis occurs, yielding nonthickening dextrins (starch breakdown to oligosaccharides). To avoid thinning in acidic starch-thickened foods, a *cross-linked starch* (see Section 2.2.3) may be selected. Since cross-linked molecules are quite large, extensive hydrolysis is required to decrease viscosity significantly.

1.4.3. Water Activity

In foods and other systems, water is not just a medium for reaction, but is also an active ingredient used to control reactions, texture, and general physical and biological behavior. However, total water content is not the important factor, but rather the availability of water (water activity, a_w). Water activity may be defined as the ratio of water vapor pressure over a sample p to the water vapor pressure over pure water p_0 under the same conditions

$$a_w = \frac{p}{p_0}$$

Alternatively, it may be expressed as the equilibrium relative humidity *ERH* of the atmosphere surrounding a sample divided by 100

$$a_w = \frac{ERH}{100}$$

Water activity is affected by strong water-binding agents or hydrophilic parts of other molecules. If water-binding agents are present in starch preparations, water activity is lower and gelatinization is inhibited to varying extents. Strong water-binding constituents retard starch gelatinization by binding and holding the water needed for gelatinization.

1.4.4. Amylose Complexes

Helical amylose molecules can entrap other organic molecules, such as fatty acids as well as inorganic molecules such as iodine. Such complexes are called inclusion compounds and the amount of iodine binding is a measure of amylose content. Amylose can be separated from amylopectin by addition of *n*-butanol to a hot dispersion and allowing the mixture to cool slowly. On cooling, crystals of butanol – amylose separate and can be removed by filtration or centrifugation [31].

Lipids and emulsifiers, such as mono- and diacylglycerol, affect starch gelatinization. Fats that complex with amylose retard granule swelling. Fat, in the absence of emulsifiers, decreases the temperature at which maximum viscosity occurs but does not affect the maximum viscosity. Corn starch normally gelatinizes to produce a maximum viscosity at 92 °C, but in the presence of 9 – 12 % fat, maximum viscosity occurs at 82 °C.

Monoacylglycerols with acyl chain length of 16 – 18 carbon atoms increase the gelatinization temperature. They also decrease the temperature of maximum viscosity, the temperature of gel formation, and the gel strength. Fatty acids or the fatty acid chains of monoacylglycerols form inclusion complexes with amylose, and possibly with longer outer chains of amylopectin (see Fig. 7). Such complexes are not easily leached from the granule and inhibit entry of water into the starch granule. The lipid – amylose complex also limits retrogradation by interfering with the formation of junction zones between starch molecules.

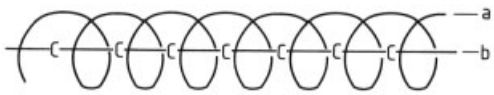

Figure 7. Complexation of a fatty acid-like chain with helical amylose segment (a = helical amylose; b = fatty acid chain)

1.5. Composition of Native Starches

Normal starches contain about 25 % linear amylose (Table 3). Particular varieties of corn, called *high-amylose corn*, produce starches with amylose contents up to 85 %, although most commercial varieties contain ca. 65 % amylose. High-amylose starches are difficult to gelatinize, often requiring temperatures in excess of 100 °C at > 1 bar. At the other extreme, some starches consist only of branched *amylopectin* (Table 3). These include waxy corn, waxy barley, and waxy or glutinous rice. They produce clear cooked pastes, resembling those of root or tuber starches. These starch pastes are fairly stable to retrogradation due to their lack of amylose.

The term *retrogradation* denotes spontaneous changes during the aging of starch solutions, pastes, or gels that lead to an increase in the degree of order. Retrogradation is crystallization due to the formation of hydrogen bonds between hydroxyl groups of neighboring starch molecules. Retrogradation kinetics of numerous starches have been studied by differential scanning calorimetry and rheology [32, 33]. Potato and pea starches are prone to retrogradation while waxy rice and modified waxy corn are least susceptible.

1.6. Industrial Starch Production Processes

1.6.1. Corn and Sorghum Starch

Corn (maize, *Zea mays* L.) has been grown for 5000 – 7000 years and was probably initially cultivated in Central Mexico. Early corn was quite different from modern varieties, which were derived by both random and scientific selection. Corn types available include popcorn, sweet corn, dent corn, flint corn, and flour corn. Corn's availability at low prices, its storability, and 70 % starch content has greatly expanded its use in commercial production of starch.

Grain sorghum (*Sorghum bicolor* M.), was probably initially cultivated 5000 – 7000 years ago in Eastern Africa. Although sorghum resembles corn in chemical composition, it is generally considered to be inferior for food, feed, and industrial uses. Sorghum can be cultivated effectively in semi-arid regions and is milled similarly to corn.

A single kernel of corn is a caryopsis (a berry) borne on an ear. This berry matures about 60 days after pollination and corn is typically harvested in late summer or early fall. In sorghum, the seeds are borne on a terminal, bisexual head, or rachis.

Corn (and sorghum) wet milling consists of several major steps (Fig. 8): (1) cleaning the grain; (2) steeping; (3) milling and fraction separation; (4) starch processing; and (5) starch drying.

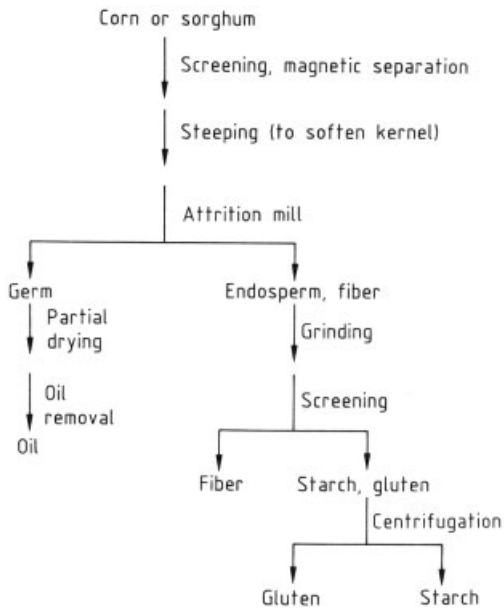

Figure 8. Flow sheet for corn (sorghum) starch production

Cleaning. Grain is first cleaned by screening and by magnetic means to remove foreign material including metal.

Steeping. The grain is steeped to soften the kernel and facilitate its separation into components. This process involves steeping in countercurrent water containing 0.10 % sulfur dioxide at 48 – 52 °C for 30 – 40 h. Sulfur dioxide combats growth of spoilage organisms and maximizes the starch yield. Sulfur dioxide also reacts with the protein matrix to disperse the glutelin by breaking disulfide bonds and releasing extra starch. Proper steeping temperature

Table 3. Comparison of composition and properties of some commercial starches

	Starch content, wt %	H_2O content, wt %	Protein content, wt %	pH	Amylose content, wt %	Amylopectin content, wt %	Specific gravity
Corn							
Normal	88	11	0.35	5	28	72	1.5
Waxy	88	11	0.28	5	0	100	1.5
High-amylose					65 – 85	15 – 35	
Sorghum	88	11	0.37	5	28	72	1.5
Tapioca		12		6.3 – 6.5	16	84	
Arrowroot					21	79	
Sago					26	74	
Potato	80	17 – 18	trace		20	80	
Wheat					30	70	
Rice							
Normal		12	0.37		20 – 30	70 – 80	1.5
Waxy		12	0.13		0	100	1.5

is critical to promote the growth of lactic acid bacteria. Growth of these bacteria lowers the steep pH, restricting the growth of other organisms. Steep water pH is typically ca. 4.0, and is highly buffered. Steep water is concentrated, dried by absorption on corn fiber, and sold as cattle food. Alternatively, it is further processed to remove phytic acid or sold as nutrient for commercial fermentation to produce, for example, antibiotics.

Milling and Fraction Separation. After steeping, grain is milled to obtain separation of components. Steeped grain is first put through an *attrition mill* (Bauer type) to release the oil- containing germ which is removed by centrifugation in hydrocyclones. Next, the residual grain is *ground* between stones to release starch granules from their containment in proteinaceous cells. The slurry is *screened* to separate the corn hull or seed coat which is recovered as a fibrous mass. The remaining starch slurry is more strongly centrifuged to separate starch granules that are washed free of protein and salts and are dried in continuous driers. The isolated germ is partially dried prior to hexane extraction or mechanical removal of corn oil.

Centrifugation. Usually the removal of starch is performed in centrifuges or hydrocyclone batteries that can separate not only the starch but also the insoluble gluten fraction of protein.

Gluten is less dense (1.1 g/cm^3) than starch (1.5 g/cm^3) and is thus easily separated by centrifugal forces. Starch thus produced contains 1 – 2 % protein, which may be removed by a second starch washing with water and centrifugation to give a starch with a final protein content of < 0.38 %. In the case of *sorghum*, starch and gluten are more difficult to separate because the sorghum pericarp is fragile compared to corn. Separation is improved if the kernels are dehulled before milling. This produces a cleaner separation of starch from gluten, giving a starch of low color and protein content.

The final starch slurry can be dried and sold as unmodified starch, or the starch can be modified (see later sections) and dried, or gelatinized and dried. It also can be acid or enzyme hydrolyzed to a syrup which may be used directly or fermented to produce ethanol [*64-17-5*] or other chemical feedstocks.

Drying. Starch is usually dried by flash drying. In this method starch slurry is centrifuged or filtered and the moist starch is introduced at the bottom of a stream of rapidly moving hot air (93 – 127 °C). Drying is rapid and the starch is collected in cyclones. Drying parameters can be used to control both bulk density and particle size. Evaporation of the water requires so much energy that the wet-milling industry is the second most energy intensive food industry.

1.6.2. Wheat Starch

Wheat has been utilized by humans for 8000 – 9000 years. Its cultivation probably began in what is now Iraq. The first instance of commercial manufacture of wheat starch was in England in the 1500s. Wheat is mainly used in human food, while on the contrary much of the corn

grown is used in animal feed, and the world annually produces more wheat than corn. Wheat is harvested by mechanical means and wheat starch is separated by one of several processing methods such as the Martin, Batter, Fesca, ammonia, or the acid process, by wet wheat milling, whole wheat fractionation, the Rasio, or the hydrocyclone process. The first two processes are most common.

Martin Process. The Martin (dough ball) process, invented in 1835, uses wheat flour as its raw material (Fig. 9). Wheat flour and water (2 : 1) are blended and the dough is allowed to rest, hydrate, and strengthen the gluten matrix. The dough is then kneaded while washing with water to remove the starch. The residual gluten is partially dried by roller compression and then flash dried. Starch slurry from the dough washer is screened to remove any remaining gluten and then fine screened to remove bran. The starch slurry is centrifuged to produce a high solids starch cake which is flash dried. Starch is dried to 10 – 12 % moisture and contains ca. 0.3 % protein.

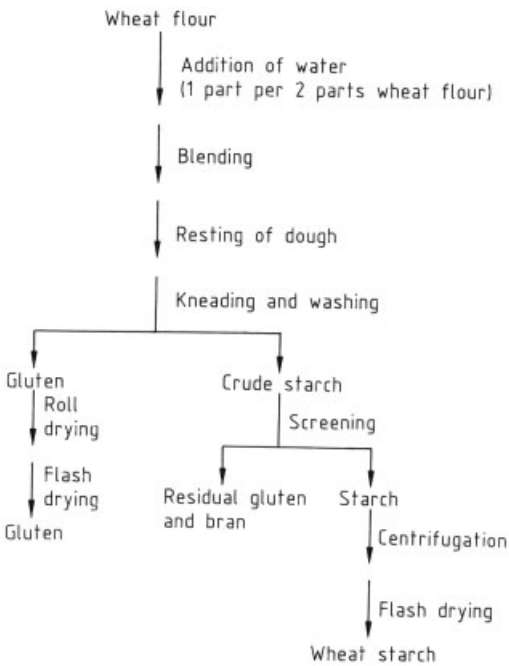

Figure 9. Flow sheet of the Martin process

Batter Process. The Batter process is derived from the Martin process but differs in the way the dough is formed and the way it is subsequently treated (Fig. 10). Wheat flour and water (ca. a 1 : 1 ratio) are mixed and the produced dough is allowed to rest for 30 min. The dough is then vigorously mixed with twice its weight of water. Gluten is separated from this mixture by screening and bran is removed from the starch by additional screenings. Centrifugal separation of the starch slurry separates the large starch granules (*type A starch*) from the small granules, damaged granules, and pentosan complexes (*type B starch*). Type A starch is essentially pure and may be dewatered and flash dried. Type B starch is processed again in a similar way but is of lower purity. Large wheat starch granules are extensively used in making carbonless copy paper where the granules serve as agents to prevent premature bursting of the ink microcapsules.

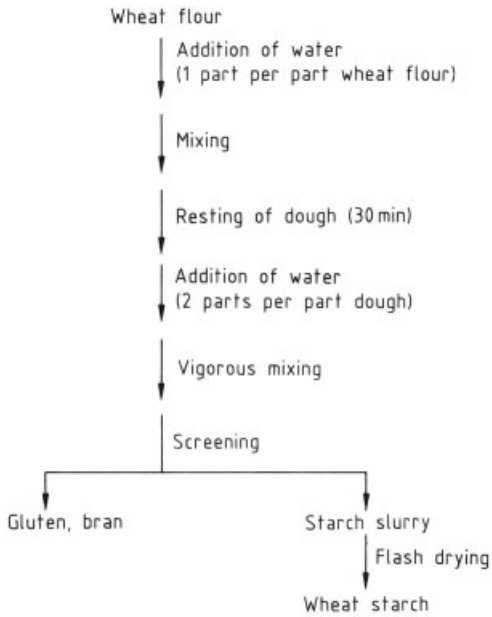

Figure 10. Flow sheet of the Batter process

1.6.3. Potato Starch

Potato starch is produced in large quantities in Europe, especially in the Netherlands, Poland, and Germany. Potato starch is usually isolated

from cull potatoes, surplus potatoes, or waste streams from potato processing plants (Fig. 11).

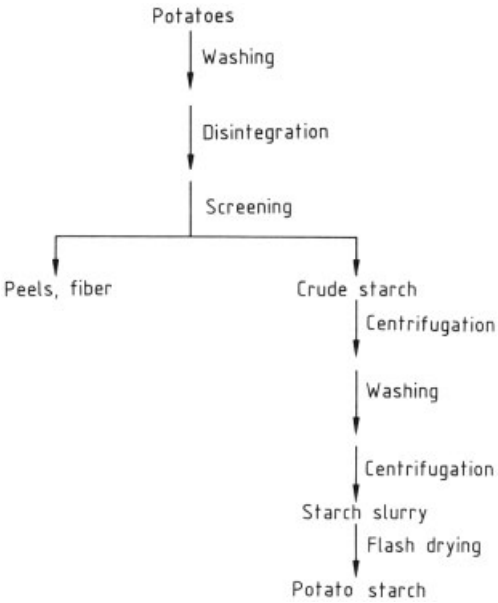

Figure 11. Flow sheet for potato starch production

Cull or surplus potatoes are washed with water to remove dirt and other foreign matter. They are then disintegrated by a saw blade rasp or hammermill and the mashed product is screened to remove peels and fiber. During this process water containing sulfur dioxide is used to preserve color and inhibit enzymatic browning. Screening separates starch from pulp. The pulp can be reground and a second extraction performed to obtain a total starch yield of 12 – 19 %, based on raw potatoes.

Starch isolated as described above is reslurried to remove water solubles and then partially dried in a continuous centrifuge. This is followed by another washing and centrifugation to obtain a purified slurry of potato starch. The slurry obtained from these operations is filtered and flash dried (< 175 °C inlet temperature). The potato starch, dried to 17 – 18 % moisture, is subsequently screened and packaged.

1.6.4. Rice Starch

Rice (*Oryza sativa* L.) cultivation probably originated in Asia where the crop is still very important. Rice is an important commercial crop in India, China, the United States, and Italy. In the United States more than 90 % of the rice crop is produced in Arkansas, Louisiana, Texas, and California. Rice grain consists of a hull surrounding an edible berry, or caryopsis. The outer fibrous layers of the caryopsis conceal a starchy endosperm and the germ. The endosperm consists principally of starch granules and protein. Little matrix protein is evident in rice endosperm.

Broken rice is reacted with 0.3 – 0.5 % sodium hydroxide solution to soften the grain and assist in protein removal. The alkali treated grain is hammermilled and the broken cell walls are removed by screening. The rice starch slurry is purified by centrifugation and dried.

1.6.5. Tapioca Starch

Tapioca starch is produced by processing the tuberous roots of the tapioca plant, also called cassava, manioc, or yuca (*Manihot esculenta*), a native of many equatorial regions. Cassava roots may be "sweet" (containing $\leq$50 mg of HCN per kilogram fresh root) or "bitter" (containing $\geq$250 mg of HCN per kilogram fresh root). Sweet root varieties are grown for food starch production while bitter varieties are used for other industrial purposes. In both cases, hydrogen cyanide content is eliminated or reduced during processing.

Root harvesting is usually manual and yields vary from 11 200 – 44 800 kg/ha. Typical root composition is 70 % water, 24 % starch, 2 % fiber, 1 % protein, and 3 % fats, minerals, and sugars. Tubers need to be processed within 48 h of harvest. They are first washed to remove dirt and then the outer skin is removed. The root is chopped into small pieces and passed through a rasp disintegrator. In this process hydrogen cyanide is released by natural enzymes and removed by washing.

Cassava root pulp is washed on screens. The starch passes through but the fibers are retained. The fiber fraction is used in fertilizer or cattle feed. After screening the starch slurry is centrifuged to separate the starch from fiber fines and other soluble material. Starch collected in this manner can be reslurried and recentrifuged to obtain greater purity. Sulfur dioxide (0.05 %)

is usually added to process water to inhibit microbial growth. Slurried starch from the purification process is dewatered by centrifugation or filtration followed by drying on a drum, belt, or tunnel, or by flash drying. Flash drying is most common. Tapioca starch is mainly used unmodified but can be modified as is done with corn starch. The viscosity of tapioca starch dispersions is a function of plant variety, geographical conditions, harvest time, root age, soil fertility, rainfall, and manufacturing practices during starch production.

1.6.6. Arrowroot Starch

Arrowroot starch comes from the root of the tropical perennial *Maranta arundinacea.* Roots are harvested after 6 – 12 months growth and contain 20 % or more starch, extracted as described for tapioca starch. The outer skin of arrowroot root must be completely removed to prevent off- color and off-taste in the starch. Arrowroot starch is mainly a product of China, Brazil, and St. Vincent in the West Indies.

1.6.7. Sago Starch

Sago starch is a granulated product which has been originally obtained from stems of several species of palms, principally *Metroxylon, Arenga,* and *Mauritia sp.* Trees of eight or more years old are used. Principal production occurs in Sarawak and New Guinea. Cut palm trunks are split and the pith removed and kneaded in water to release the starch which is screened to remove fiber, filtered, and dried. A single palm trunk can yield 100 – 200 kg sago starch. Sago starch granules are 20 – 60 µm in diameter.

Nowadays sago is also produced from tuberous roots of other tropical plants, such as cassava. Some comparative properties of commercial starches are given in Table 3.

1.7. Uses

Uses in Foods. Underivatized (native) starch is widely used in foods, both as an isolated, pure product and as a constituent of cereal grains, where it thickens and forms gels. It performs this function by undergoing gelatinization (see Section 1.4). In principle, the function of starch is to act as a "water sink", absorbing much of the available free water and providing desirable structure and texture in many foods. Starch may also be cooked and dried to provide a cold water dispersible, pregelatinized product which is the basis of many instant gel and pudding products.

Pea starches have been proposed for use in preparation of gels requiring ca. 50 % less starch than comparable corn starch gels, in production of extruded products, in production of pulpy products via freeze – thaw technology, and in production of instant starches with cold water swelling and gelling properties [34].

The viscoelastic behavior of starch gels as a function of time has been reported [35].

Starch may be hydrolyzed and/or isomerized to produce corn syrup sweeteners, corn syrup solids, or high fructose corn syrup (HFCS) (→ Glucose and Glucose-Containing Syrups). Alternatively, it may be fermented to produce ethanol and other industrial chemicals (→ Ethanol).

Other Uses. Unmodified starch also finds applications as textile sizings and adhesives, and is useful as a wet-end adhesive in the paper industry. It is used in oil field applications (in its pregelatinized form) where it functions as a suspending agent and maintains drilling fluid viscosity and decreases fluid loss.

In the pharmaceutical industry starch is used as a tablet binder, and in cosmetics rice starch in particular is employed as a dusting powder.

2. Modified Starches

Starch may be modified in numerous ways, both physically and chemically. These modifications are discussed in the following sections. Some properties of various corn starch derivatives are shown in Table 4.

2.1. Physically Modified Starches

Pregelatinized starch is a common ingredient in foods as well as other industrial applications. It is commercially prepared by putting a starch slurry in a trough between two counterrotating rolls internally heated by steam. The starch slurry is

heated above the gelatinization temperature and is then drawn downward between the nip of the hot rolls to form a thin film. The film dries as the rolls turn and is scraped off and ground to produce a fine powder. Pregelatinized starch can also be prepared by spray drying a gelatinized starch paste. Pregelatinized starch quickly rehydrates in cold water. In foods it can be uniformly incorporated without heating to provide thickening, binding, and other properties. Pregelatinized starch is an essential component in foods in which cooking is not employed, such as in instant puddings, pie fillings, and cake frostings.

Table 4. Properties of various corn starch derivatives

Type	Gelatinization temperature range, °C	Distinguishing properties
Normal	62 – 72	poor freeze – thaw stability
Waxy	63 – 72	little retrogradation
High-amylose	66 – 92	granules less birefringent than those of normal starches
Acid-modified	69 – 79	decreased hot-paste viscosity compared to unmodified starches
Cross-linked	higher than unmodified, dependent on degree of cross-linking	reduced peak viscosity, increased paste stability
Acetylated	55 – 65	excellent paste clarity and stability, good freeze – thaw stability
Phosphate, monoesters	56 – 66	reduced gelatinization temperature, reduced retrogradation
Hydroxypropyl	58 – 68 (DS 0.04)*	increased paste clarity, reduced retrogradation, good freeze – thaw stability

* DS = degree of substitution.

Pregelatinized starches have been incorporated into extruded oat-based snack foods where they provide well expanded, crisp products and lead to reduced product rancidity. Incorporation of starch in these foods also reduces wear and tear on the cooker extruder [36].

Pregelatinization of starch typically increases ease of digestibility, although not necessarily in a linear fashion [37]. Although pregelatinized starches are rapidly digested, they do not differ in their calorific value from other starches except for the case of high-amylose starch, which is so difficult to solubilize that any ungelatinized granules may be unavailable for digestion. Some precooked starches give low food value to some animals because the high viscosity produced on eating may induce satiety so that the animal eats less [38, 39].

2.2. Chemically Modified Starches

Starch molecules, both in free and granular form, are subject to chemical modification. Thus, starch is modified in various ways to produce acid-modified (thin-boiling), oxidized, cross-linked (e.g., by formation of distarch phosphates or adipates), partially esterified (either as carboxylate esters or phosphate esters), or partially etherified starch, or converted to cationic derivatives. These modifications, normally present at very low concentrations in starch preparations, produce dramatic differences in the physical and chemical properties of starch and lead to a multitude of uses in both food and nonfood applications.

2.2.1. Acid-Modified Starch

Treatment of starches such as dent corn starch with acid at a temperature below the gelatinization temperature produces virtually no change in granular appearance [40], and only a small change is noted in granular birefringence patterns [41, 42]. However, changes are produced that affect the gelation of acid-modified starch such as a lower hot paste viscosity [43 – 50], lower intrinsic viscosity [51], lower iodine affinity [51], and an increased solubility in water at temperatures just above the gelatinization temperature [45, 52, 53].

Acid treatments of gelatinized starch date back to 1811, when Kirchoff produced D-glucose [*50-99-7*] from starch by acid hydrolysis. Treatment of ungelatinized starch was reported by Naegeli in 1874 [54] who found that such treatments dissolved much of the starch granule and left a residue of short starch molecules (DP ca. 25) (Naegeli dextrins). Commercial production of acid-modified starch dates to between 1897 and 1901 [55, 56].

Acid-modified, or thin-boiling starch has a lower molecular mass, or degree of polymerization than native starch [57, 58]. For example, the DP of potato starch, which is initially 1630, drops to 990 after 4 h treatment with 0.2 N HCl at 45 °C.

Acid-modified starch is prepared today by heating a starch slurry with 36 – 40 % solids content to 40 – 60 °C (below the gelatinization

temperature of the starch) with hydrochloric acid for one to several hours. When the desired degree of hydrolysis is achieved, as judged by the viscosity of the cooked modified starch, the acid is neutralized, and the granular modified starch is filtered, washed, and dried. Sometimes other acids are used such as sulfuric acid, with different reaction conditions.

The change in iodine binding capacity (see Section 1.4.4) or iodine affinity (which reflects the amount of amylose or linear starch material) after acid treatment varies with the type of starch. With corn starch only a little change in iodine affinity occurs for moderate acid conversions (relative product fluidity in the range 40; fluidity = 1/viscosity). This result is taken as an indication of preferential acid attack on the amylopectin component of the starch.

As the molecular size of starch decreases due to acid hydrolysis, the number of reducing end groups increases. This is measured commercially by the alkali number which increases as the extent of hydrolysis increases (there is an increase in fluidity) [59] (see Table 5). As acid hydrolysis proceeds, starch solubility in hot water increases [53].

Table 5. Relation of fluidity of starch dispersions to alkali number [59]

Relative fluidity	Alkali number*
20	14.5
40	15
90	41.5

* Alkali number = milliequivalents of alkali consumed per 10 g of calculated dry starch during a standard 1 h digestion in 0.1 M sodium hydroxide at 100 °C.

Acid treatment causes hydrolysis of glycosidic bonds in the starch molecules

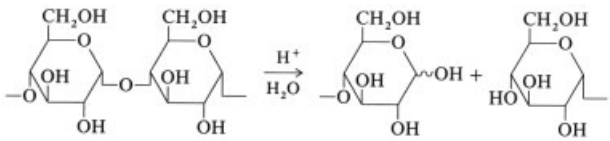

For *dispersed starch* the (1,4)-α-D linkages are more susceptible to acid cleavage than the (1,6)-α-D branch points [60]. However, in *granular starch*, portions containing many of the (1,4)-α-D linkages (amylose, long outer chains of amylopectin) are present in the crystalline regions of the granule. Hence, the (1,6)-α-D linkages become more accessible and susceptible to acid hydrolysis. For hydrolysis of potato starch, moisture conditioning without swelling may facilitate hydrolysis [58]. This, with other data, suggests that acid conversion occurs in two steps:

1) An initial attack on the amorphous regions, rich in amylopectin, especially the vulnerable (1,6)-α-D branch points
2) A slower attack on the more highly organized and crystalline regions of amylose and amylopectin

Uses. Acid-modified starches are employed in *textile manufacture* where they function as warp sizes [61, 62] to increase yarn strength and abrasion resistance during weaving. These uses have somewhat declined due to increased use of poly(vinyl alcohol). Starches with different fluidities are used for specific types of yarns. Thus, low fluidity starch is used for heavy yarns and high fluidity starch is used for light yarns. Starch concentrations of 10 – 12 % are usual for such sizing applications.

Acid-modified starch is also used in the manufacture of *gypsum board* for dry wall construction where it serves to bond plaster and paper together.

Acid-modified starches find use in *starch gum confectionery*, an example being candy orange slices. To produce proper texture, the acid-modified starch is boiled together with sugar, corn syrup, and water to form highly concentrated pastes which form very firm gels on cooling. Not surprisingly, the granule structure is almost completely disintegrated in the cooking process. The composition of a typical starch gum confectionery is given below (for the definition of the term dextrose equivalent, see → Glucose and Glucose-Containing Syrups, Chap. 4):

Sucrose	45.4 kg
Corn syrup (dextrose equivalent 63)	68.1 kg
Acid-modified starch (fluidity 70)	31.8 kg
Water	26.5 L

Production of such confectioneries begins with heating of corn syrup, water, and acid-modified starch to just below boiling in a steam kettle. When the starch has been dispersed, sugar is added and the mixture is heated to 103 °C and the heat removed. The water content of the cooked paste is adjusted to 22 – 23 % and the mixture is put through an injection cooker at 140 °C. Flavor and color are added and the paste poured into molds and allowed to set to a firm gel. As a final step, the molded gel is coated with sucrose using a sugar sander.

Thin-boiling starch is also applied in *paper manufacture* as a paper size where it increases the strength of Kraft linerboard and improves bleached board printability. It is typically applied at 66 – 71 °C and pH 7.5 – 8.0. A high solid, acid-modified carrier starch is used to increase the production rate of corrugated board [63].

2.2.2. Oxidized Starch

Oxidation of starch has a long history beginning with Liebig [64] who used chlorine to oxidize starch in 1829. Somewhat later Lieben and Reichardt [65] used bromine. These oxidations by halogens, and the hypohalites produced by their reaction with water, may proceed in four different ways, producing distinct products.

1) *Oxidation of reducing end aldehyde to carboxyl groups*, producing aldonic acid end units; namely, D-gluconic acid end units.

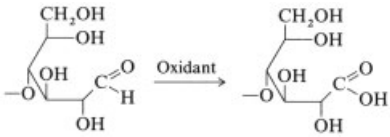

2) *Oxidation of the C-6 Methylol Group to a Carboxyl Group*. D-Glucuronic acid was isolated and identified from starch oxidized by bromine [66]. The oxidation probably proceeds by way of the 6-aldehyde derivative.

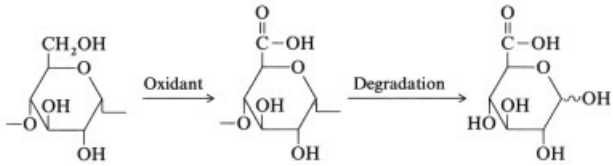

3) *Oxidation of Starch Secondary Hydroxyl to Ketone Groups.* This reaction is shown below for oxidation of the 3-OH to a carbonyl group. Bromine oxidation of starch [67] resulted in a product from which an oxime could be prepared, thus indicating the presence of ketone groups.

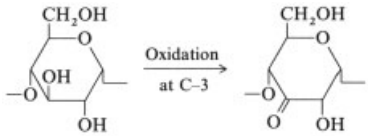

4) *Oxidation of 2,3-Glycol Units to Dialdehyde and Dicarboxylic Acid Units* (*Glycol Cleavage*). Evidence for this reaction was first obtained by Farley and Hixon [66]. Oxidative attack varies with the location of hydroxyl groups and with the nature of the oxidant. Oxidation at C-1, C-2, C-3, and C-6 determines the properties of the resulting starch product.

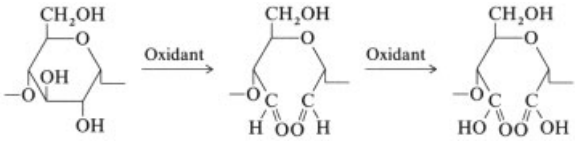

Common sodium hypochlorite-oxidized starch is granular in form with few surface changes noted by scanning electron microscopy. It retains its birefringence and granular appearance [68]; it is whiter because of the bleaching action of sodium hypochlorite. Oxidation occurs mainly in the amorphous regions of the granule. Oxidized starch is less stable to heat than native starch, tending to turn yellow or brown upon heating due to further reaction of carboxyl or aldehyde groups produced by the oxidation [69].

Heated aqueous dispersions of oxidized starches have lower intrinsic viscosities which depend on the extent of oxidation [68] and have a lower gelatinization temperature range than native starches [70, 71]. The aqueous dispersions have greater clarity than those of unmodified starch and exhibit a lower tendency to gel or setback [72].

Production. Commercially a controlled amount of chlorine gas is bubbled into dilute aqueous alkali to produce hypochlorite ions. The amounts of available sodium hydroxide and chlorine determine which of the four types of oxidation will principally occur. Reaction is performed on a rigorously stirred starch slurry of about 33 – 44 % solids content. The pH of the slurry is adjusted to 8 – 10 with sodium hydroxide and chlorine added to produce hypochlorite with 5 – 10 % available chlorine. Temperature is maintained at 21 – 38 °C by controlling the rate of hypochlorite addition or by external cooling. When the desired degree of oxidation is obtained (usually determined by viscosity analysis) the pH is decreased to 5 – 7 and excess chlorine is destroyed by reaction with reducing agents such as sodium bisulfite or sulfur dioxide. The modified starch is isolated by centrifugation or filtration and is washed and dried.

Uses. Oxidation introduces an increasing number of carbonyl and carboxyl groups into the starch molecules, which in turn cause changes in the chemical and physical properties [73, 74]. Most commercial oxidized starches contain about 1.1 % carboxyl groups. These groups play a predominant role in determining the properties of the starch derivative which is used mostly in the *paper industry*. 80 – 85 % of oxidized starch is used in paper production where it acts as a coating binder for high-solid pigment coating colors. In this application, pigment compatibility is important and the starch should not affect the rheology or the water binding characteristics of the coating color [70, 75].

Oxidized starch is used also as a surface size in paper and paperboard manufacture. Here it seals pores, improves surface strength, and provides ink holdout. As a wet-end additive in paper manufacture, oxidized starch reduces pigment retention by acting as a pigment dispersant [76].

Oxidized starches have a long history of use in *textile manufacture* where they function primarily as warp sizes but also in finishing and printing. As warp sizes they provide abrasion resistance for yarn but are now being replaced by synthetic polymers.

Oxidized starch has been used in *laundry finishing* and in making *construction materials* such as wallboard and ceiling tiles. In these applications, starch provides sizing and adhesive characteristics.

Lightly oxidized starches have application in *breading batters* for deep fried foods where they give good adhesion of the breading to the food [77].

Periodate oxidized starch is a good substitute for formaldehyde or glutaraldehyde as *hardening agents* in gelatin-immobilization of yeast cells [78].

2.2.3. Cross-Linked Starch

Cross-linking of starch reinforces the intermolecular binding by introducing covalent bonds to supplement natural intermolecular hydrogen bonds. Cross-linking restricts granule swelling, decreases peak viscosity on cooking, and increases the stability of the gelatinized granule. Most cross-linked starches are quite lightly derivatized, containing ca. one cross-link for every 100 to 3000 anhydro-D-glucosyl units (degree of substitution, DS 0.01 – 0.003).

Cross-linked starch granules also appear unchanged by microscopic investigation but they are more resistant to swelling than unmodified starch. This is especially evident when their dispersions are placed in a Brabender viscoamylograph or similar instrument. Cooked cross-linked starch granules are less fragile than native starch granules and more resistant to high shear, elevated temperatures, or low pH [79]. At low degree of substitution an increase in peak viscosity is observed when measured by a Brabender amylograph but with high DS no viscosity peak occurs when aqueous dispersions are gelatinized. Loss of peak viscosity is caused by constrained swelling of the granule due to extensive cross-linking of the starch molecules. If the degree of cross-linking is sufficient, gelatinization will not take place at 100 °C or even in an autoclave. Cross-linked starches of low DS (< 1) have improved textures; they are less rubbery and stringy than pastes and gels made from unmodified native starches. Cross-linking is sometimes performed in combination with oxidation [80] , phosphorylation [81], hydroxyalkylation [81], or esterification [82].

Production. Cross-linking reactions are usually performed on starch slurries which are

then reacted with specific bi- or polyfunctional reagents such as phosphorus oxychloride, sodium trimetaphosphate, epichlorohydrin, or mixed anhydrides such as those of acetic and adipic acid. Slurry reaction temperatures may range from 25 to 50 °C. After cross-linking is complete, the cross-linked starch is isolated by filtration, washed to remove excess reagents and salts, and dried.

The reaction of starch with *phosphorus oxychloride* to produce a distarch phosphate is shown below:

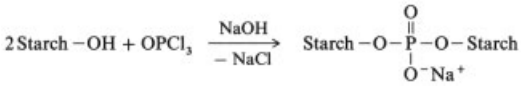

Typically, the reaction is carried out at pH 8 to 12 and with 0.005 to 0.25 % phosphorus oxychloride. Frequently, sodium chloride or sodium sulfate is added to the slurry to inhibit granule swelling and starch leaching from the granule [83]. When the concentration of phosphorus oxychloride is increased from 0.005 – 0.25 % to 1 % or more, the resulting starch phosphates are highly cross-linked.

Epichlorohydrin is commonly used [80] to cross-link starch under alkaline conditions, as shown below:

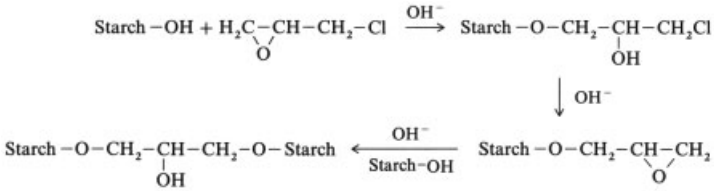

Common *mixed anhydride* cross-linking agents are those of acetic and dicarboxylic acids, such as adipic acid [82]. Other reagents used to cross-link starch include (but are not limited to) vinyl sulfone, diepoxides, cyanuric chloride, toluene-2,4-diisocyanate, *N,N*-bismethylene-bisacrylamide, phosgene, imidazolides of polybasic carboxylic acids, and aldehydes such as formaldehyde, acetaldehyde, or acrolein [84].

Properties. Cross-linked starches are less soluble in hot water than native starches even though the degree of molecular cross-linking is very low [85]. Material extractable by dimethyl sulfoxide (DMSO) is also reduced by cross-linking (see Table 6) [52].

Table 6. Solubility of waxy and cross-linked waxy sorghum starch in dimethyl sulfoxide at 25 °C [107]

Starch	Time, h	Solubility, %
Waxy	20	100
Cross-linked waxy	45	20

Cross-linking of starch at low degrees has little effect on in vitro digestibility. The calorific value of the cross-linked derivative is the same as that of the unmodified, native starch [86].

Cross-linked starch provides a stable, high viscosity paste that may be subject to high temperature, high shear, or low pH.

Uses. Cross-linked starches with a low DS are extensively used in *foods*. In extrusion cooking, high degrees of cross-linking are necessary to prevent complete granular disintegration. When a short, salve-like consistency is desired, cross-linked waxy corn, potato, or tapioca starch, which is usually also phosphorylated, acetylated, or hydroxypropylated, is used. Cross-linked starch provides texture in salad dressings due to its stability to both low pH and the high shear experienced during homogenization. These derivatized starches are also used in batter mixes for deep-fried foods, fruit pie fillings, puddings, cream style corn, soups, gravies, sauces, and baby foods [79, 87, 88]. In some food systems the starch imparts a desirable pulpy texture [89].

Cross-linked waxy starches are constituents of cake mixes where they improve cake volume, crumb texture, and increase shelf-life, due to decreased rate of becoming stale [90]. Acid-converted cross-linked starches are dispersible in cold water and form gels [91], they are thus used in instant puddings.

Cross-linked starches are used also in antiperspirants [92] and textile printing pastes where

they produce high viscosity and a short, cohesive texture [93]. They have been employed in oil-well drilling muds, printing inks, charcoal briquette binders, fiberglass sizing [94], and textile sizing [95].

2.2.4. Starch Esters

Carboxylate Esters. Starch acetates [*9045-28-7*] have been known for more than 100 years [96]. Those of high DS are soluble in organic solvents with a low polarity but because of the branched amylopectin molecules they are not economically competitive with similar cellulose derivatives. Material with a lower degree of substitution (0.3 – 1.0) is water soluble, whereas most commercial products are even more lightly substituted (DS 0.01 – 0.2). In anhydrous media, the acetylation reaction is very sluggish and, in addition, the granules require prior swelling to allow penetration by the acetylation reagents. Starch acetates can be produced in many ways. Commercial processes typically employ acetic anhydride, with or without catalysts.

Reaction with Acetic Anhydride Alone. Starch does not react with acetic anhydride at ambient temperature and only slightly at 90–140 °C. Cooking and disruption of starch granules followed by ethanol precipitation increases reactivity but acidic catalysts cause starch degradation [99].

Reaction with Acetic Anhydride [*108-24-7*] *in Pyridine* [84]. Starch granules treated with acetic anhydride alone at 20 °C for 5 months showed no reaction. Pyridine treatment [97] rendered the starch granule reactive [98]:

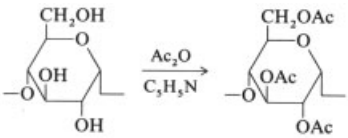

Reaction with Acetic Anhydride and Glacial Acetic Acid. Reaction requires the addition of an acidic catalyst such as sulfuric acid, *p*-toluenesulfonic acid, perchloric acid, or phosphoric acid. The yield of acetylated (and degraded) starch acetate increases at elevated temperature.

Reaction with Acid Anhydride in Dimethyl Sulfoxide (DMSO) [100]. Starch derivatives of acetic, propanoic, and butanoic anhydrides have been prepared up to a DS of 0.08. Triethylamine serves as a catalyst and acid scavenger.

Reaction with Carboxylic Acids. Starch does not react with aqueous acetic acid but with *glacial acetic acid* at 100 °C for 5 – 13 h [101], to give a product with 3 – 6 % content of acetyl groups. Treatment of starch with *concentrated formic acid* leads to gelatinization and simultaneous esterification [102].

Acetylation with Ketene. This reaction [103] can produce starch with an acetyl content of 2.2 – 9.4 %. The reaction is typically performed in acetic acid, diethyl ether, or acetone with an acid catalyst, the most effective of which are sulfuric acid and *p*-toluenesulfonic acid.

These reactions typically lead to a considerable degree of acetylation in the starch granule. More lightly substituted starches (low DS starch acetates) can be produced, e.g., by reaction of starch with *acetic anhydride* in the presence of a base such as 3 % sodium hydroxide [104]. The reaction is carried out at pH 7 – 11 and alternative bases such as other alkali metal hydroxides, calcium hydroxide, or sodium carbonate can be employed. Low DS starch acetates can be produced by reaction of granular starch with *vinyl acetate* [*108-05-4*] [105]:

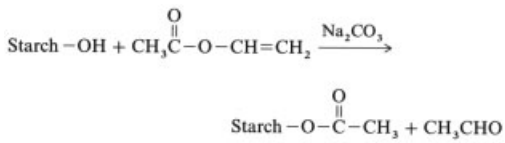

This reaction requires basic catalysts, such as alkali metal hydroxides, amines, or sodium carbonate. Acetylation is sometimes combined with hydroxypropylation to produce starches having excellent freeze – thaw stability [106].

Properties. Starch acetates with degrees of substitution of ca. 0.5 have roughly the same solubility as normal starch. Starch esters are readily cleaved by alkali, for example, acetylated granular corn starch of 1.8 % acetyl content can be deacetylated in 4 h at 25 °C at pH 11. At low DS (< 1), microscopic examination shows no difference between the granule of starch ester and the native starch granule. Low DS starch acetates have a lower gelatinization temperature, an increased hot cooked paste viscosity, easier

dispersion on cooking, and a decreased setback or final gel strength.

Higher DS starch acetates, with up to 15 % acetyl content, are soluble in water at 50 – 100 °C [107, 108]. When the acetyl content exceeds 40 %, they are only soluble in organic solvents with lower polarity (aromatic hydrocarbons, ketones, nitroalkanes, etc.). Such highly substituted derivatives are insoluble in water, diethyl ether, aliphatic alcohols, and aliphatic hydrocarbons [108].

Starch acetates with a high degree of substitution (DS 2 – 3) have increased densities, specific rotations, and melting temperatures.

Uses. Low DS starch acetates have major use in *foods* where they provide thickness, body, and texture. Certain food applications require a starch derivative with a bland taste, stability to low pH, stability to high shear and temperature, and freeze – thaw stability. Acetylation and/or cross-linking provide such qualities.

Starches containing *0.5 – 2.5 % acetyl content* (the FDA permits up to 2.5 % acetyl content) provide gels of exceptional brilliance and clarity. Such starch acetates are used in canned, baked, and dry foods such as instant gravies and pie fillings. They are used in baby foods and cream pie filling to provide texture. Products which are frozen and thawed (perhaps several times) before use also benefit from a content of starch acetates. Such products include frozen pies and frozen gravies.

Low DS starch acetates are also used in *nonfood applications* such as warp sizing in textiles to provide good yarn adhesion, tensile strength, and flexibility. Starch acetate is easily removed due to its solubility in water [109]. Starch acetates have also been used as fiberglass forming sizes [110] and in the paper industry as a surface size where it improves printability and gives uniform surface strength, porosity, solvent resistance, and abrasion resistance [109]. Such starch esters have also found use in gummed tape formulations.

High DS starch acetate derivatives have fewer uses but have been employed in thermoplastic molding [111] and in films as plasticizers [112].

Phosphate Esters. Because of their ionic character, cooked pastes of starch phosphate monoesters are "thinned" or decreased in viscosity when salts are added. Paste viscosity is also affected by pH [113].

Starch phosphate esters react with aluminum, titanium, and zirconium salts to form insoluble precipitates [114]. Starch phosphates also react with cationic dyes such as methylene blue [115, 116]. Starch phosphate esters are chemically compatible with gelatin, vegetable gums, poly(vinyl alcohol), and polyacrylates [117, 118].

Introduction of a few *distarch phosphate groups* which cause chain cross-linking greatly increases the acid stability of the phosphorylated starch derivative. In starch monophosphate esters, 28 % of the phosphate ester groups are present at C-2, 9 % are at C-3, and 63 % at C-6 of the corresponding D-glucose units. The pK_1 ranges between 1 and 2, and pK_2 between 6 and 7.

Production. Starch phosphate esters can be prepared by reaction of starch with inorganic phosphates such as sodium tripolyphosphate [*7758-29-4*] [117]. The reagent produces phosphate esters with a DS up to 0.01:

$$\text{Starch}-\text{OH} + \text{Na}_5\text{P}_3\text{O}_{10} \longrightarrow \text{Starch}-\text{O}-\overset{\overset{\displaystyle\text{O}}{\|}}{\underset{\underset{\displaystyle\text{O}^-\text{Na}^+}{|}}{\text{P}}}-\text{O}^-\text{Na}^+ + \text{Na}_3\text{HP}_2\text{O}_7$$

To do so, starch is impregnated with phosphate salts in an aqueous slurry. Starch and phosphate are mixed together for 10 – 30 min, then filtered. Typically 40 – 60 % of the phosphate salt is retained by the starch. The pH is carefully adjusted to 5 – 8.5, if the pH is too high, distarch phosphate cross-links will be formed. The final starch phosphate salt is heated at lower temperature (60 °C) to remove moisture, then heated at 120 – 170 °C to effect phosphorylation. This drying – reaction process can be performed by belt or spray drying [119].

Starch phosphates with a *low degree of substitution* can be produced by extrusion of starch with sodium tripolyphosphate [120]. Maximum degree of substitution (DS = 0.0053) is obtained at 200 °C, tripolyphosphate concentration $\geq$ 1.4 g/100 mL, and pH 8.5. Pastes of these starch phosphates have exceptional clarity.

Properties. Cooked aqueous dispersions of starch phosphates compared to natural starch

have increased gel clarity, increased viscosity, a more cohesive texture, and gels that are stable to retrogradation [117].

Uses. Starch phosphate monoesters find application in paper, textiles, adhesives, as scale inhibitors, in flocculation, and numerous other applications. In the *paper industry,* starch phosphates are used as wet-end additives which improve paper strength and filler retention [121], while in *textiles* these esters act as sizes for polyester – cotton, polyester – rayon, and pure cotton yarns [118]. They are also used as thickeners in textile printing inks because they improve ink penetration into cotton.

In *adhesives,* these esters improve storage stability and promote rapid bond strength. Use of 10 mg/L of starch phosphates prevents or inhibits *scale-forming deposits* [122]. *Coal washery tailings* can be flocculated by starch phosphates, in *food* they function as effective oil-in-water emulsifiers [123] as well as good thickeners and stabilizers for puddings [124].

Starch phosphates have also been used in *medicinal films* for the treatment of skin wounds and burns [125]. The film promotes rapid healing and reduces incidence of infection.

2.2.5. Starch Ethers

The two major types of commercially important starch ethers are hydroxyethyl and hydroxypropyl ethers.

Production. The *hydroxyethyl derivatives* are typically produced by reaction of ethylene oxide [*75-21-8*] with starch under alkaline conditions:

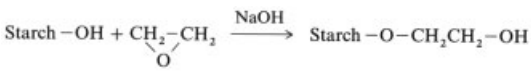

A starch slurry of 35 – 45 % solids content is made alkaline with sodium hydroxide, purged with nitrogen gas, and then ethylene oxide is added. Depending on the DS desired, the reaction is conducted for a while at 25 – 50 °C and the hydroxyethylated product is filtered, washed to remove excess reagent and salts, and dried. The mechanism apparently involves abstraction of a hydroxyl proton from a starch molecule by alkali, followed by S_N2 attack of the starchate anion on the highly strained oxirane ring. Commercial hydroxyethyl starch has a DS of about 0.2. Most of the newly formed ether groups (76 – 85 %) are formed at the oxygen bound to C-2, because the order of hydroxyl reactivity is 2-OH > 3-OH ≈ 6-OH [126].

Hydroxypropyl derivatives are prepared in a similar manner as hydroxyethyl derivatives, following the same mechanism with most substitution occurring at 2-OH. To an aqueous starch slurry of 32 – 38 wt %, sodium hydroxide is added to a concentration of 0.5 – 1.0 % (based on the dry weight of the starch). Sodium sulfate (5 – 15 %) is sometimes added to prevent starch swelling. Propylene oxide [*75-56-9*] is added in the amount of 5 – 10 % of the dry weight of starch and the reaction carried out at 52 °C for ca. 24 h. The hydroxypropyl starch is filtered, washed, and dried.

Properties. Hydroxyethyl [*9005-27-0*] and hydroxypropyl [*9049-76-7*] starch are the commercially most important starch ethers. At low degree of substitution the physical properties of these ethers are similar to those of low DS starch acetates. As the DS increases the gelatinization temperature range decreases [127], but rate of granule swelling, clarity of the dispersion [128], and dispersion cohesiveness increases. The tendency to gel or setback is lowered, and solubility in methanol or ethanol is increased [129].

It is possible to prepare films from hydroxypropyl high-amylose corn starch which are water-soluble, transparent, and oxygen impermeable [130]. Hydroxyalkylation results in decreasing digestibility at increasing degree of substitution.

Lightly substituted (low DS) hydroxyethyl or hydroxypropyl starch is similar in properties and appearance to unmodified starch. An increasing DS of either derivative decreases inter-starch hydrogen bonding which lowers pasting temperature and inhibits retrogradation. Hydroxyethyl starch pastes are more stable to added salts including those present in hard water than are the cationic or anionic starches. Hydroxyethyl ether linkages are stable over a wide pH range.

Uses. *Hydroxyethyl starch* is used in the *paper industry* as a surface sizing and coating [109] where it provides strength, stiffness, and ink

holdout. It is also applied in coatings for pigmented paper [131] to achieve good printing quality. Hydroxyethyl starch is also a wet-end additive in papermaking [121] and has been used in many *adhesives* such as bag pastes, case sealing, and label and envelope adhesives because of its excellent water holding and film forming properties [129].

In *textiles,* low DS starch ethers are employed as warp sizing. In *foods,* starch ethers (which are also usually cross-linked) provide thickness and texture and the ability to hold water at low temperature. They are used in milk-based products [132] and salad dressings [133]. The good freeze – thaw stability of the starch ethers permits their use in frozen puddings, sauces, gravies, and fruit pie fillings [134]. Hydroxypropyl, high-amylose starches produce edible, water-soluble food films [135].

2.2.6. Cationic Starch

Cationic starches include tertiary aminoalkyl ethers, quaternary ammonium ethers, aminoethylated starches, cyanamide derivatives, starch anthranilates, and cationic dialdehyde starch (although the last three are commercially less important).

Tertiary Aminoalkyl Derivatives. These derivatives are produced by reaction of the appropriate chloroalkane tertiary amine with starch under alkaline conditions, followed by acidification to produce the desired cationic derivative:

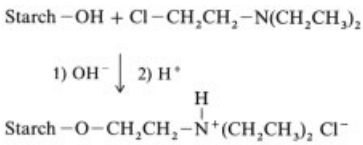

Reactions are conducted at pH 10.5 – 12 and 25 – 50 °C. The mechanism involves the reaction of starch with a nitrogen heterocycle to produce the desired product. Substitution mainly occurs at C-2 [136].

Quaternary Ammonium Starch Derivatives. These derivatives are obtained by reaction of starch with 2,3-epoxypropyltrimethylammonium chloride [137]:

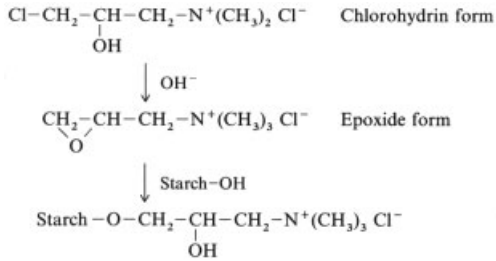

Starch is reacted with this reagent in a 40 – 46 % slurry at pH 11.0 – 12.0. The ratio of alkali to epichlorohydrin is ca. 2.8 : 1 and excess reagent is removed by vacuum distillation or solvent extraction.

Aminoethylated Starches. These derivatives are produced by reaction of starch with ethyleneimine [*151-56-4*] [138] followed by acidification:

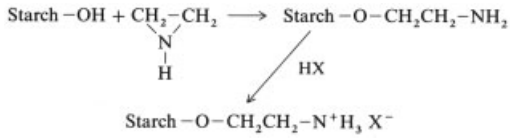

This reaction is performed with a starch slurry and typically carried out at 75 – 120 °C [139].

Cyanamide Derivatives. Cyanamide [*420-04-2*] and its dialkyl derivatives react with granular starch at pH 10 – 12 to produce iminocarbamate derivatives which can be acidified to give cationic starch derivatives [140]. This reaction is conducted under alkaline conditions

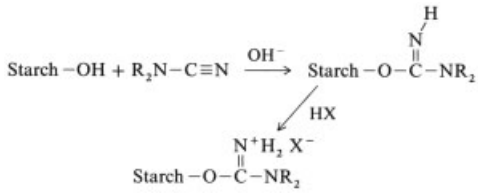

Starch Anthranilates. Starch can be reacted with isatoic anhydride at pH 7.5 – 9.0 [141].

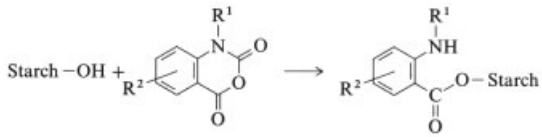

Diazotization of the resultant amine with nitrous acid produces the desired cationic starch derivative

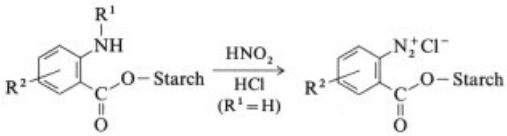

Cationic Dialdehyde Starch. Starch can be converted into a cationic form by initial oxidation to dialdehyde starch followed by reaction with hydrazine [*302-01-2*] or hydrazides (the reaction with betaine hydrazide hydrochloride is shown below) [142]

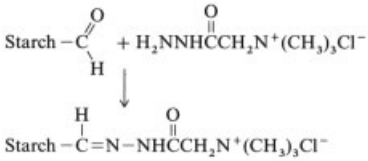

Granular starch is dispersed in an aqueous slurry (15 % solids) and reacted with a hydrazine or hydrazide derivative (3 – 5 % concentration) at an initial pH of 4.5. The pH drops to between 2.5 and 3.2 as the reaction proceeds at 90 – 95 °C for 2 – 3 h.

Properties. The gelatinization temperature range of cationic derivatives decreases as the DS increases [143]. Such cationic starch dispersions have increased stability and clarity compared to dispersions of the native starches from which they are derived. Also, zeta potential measurements show that these cationic starch derivatives have a net positive charge at a pH of between 4 and 9 [144].

Cationic starches react chemically with anionic dyes such as Light Green SF Yellowish or Acid Fuchsin [115]. On electrophoresis these derivatives migrate toward the cathode [145]. Commercial cationic starches typically have nitrogen contents of 0.2 – 0.4 % [146].

Uses. Cationic starches are used in the paper industry as wet-end additives and sizings, as coating binders for clay, as warp sizing agents in textiles, as flocculation agents for clay, titanium dioxide, coal, iron ore, silt, anionic starch, and cellulose, and for many other uses.

As a wet-end additive in *paper*, cationic starches increase the retention of fines and improve paper strength while also increasing drainage [147]. Cationic starches may also be used as paper sizing agents [144].

As a *coating binder,* cationic starch promotes the binding of clay to fiber [148]. In *textiles* it is a warp sizing agent [149] and provides good lubrication and yarn abrasion resistance. The ability of cationic starch to flocculate fines is seen in its use in treating raw primary sludge [150]. Its ionic nature makes it a useful *emulsion breaker* for either water-in-oil or oil-in-water emulsions [151].

3. Economic Aspects

Global Starch Production [152]. 99 % of the global starch production of is from maize, manioc/cassava, wheat and potatoes, with 74 % maize starch, 10 % tapioca starch, 8 % wheat starch and 7 % potato starch. Starch is predominantly produced in the United States, the EU, Japan, and Thailand. Production figures of the EU are given in Table 7.

Table 7. Starch production in the EU 2005 [152]

	EU*
Raw material input, 10^6 t	
Maize	7.0 (31.2 %)
Potato	8.7 (38.4 %)
Wheat	6.9 (30.4 %)
Tapioca	
Others	
Total	22.6
Starch produced, 10^6 t	
Maize starch	4.4 (46 %)
Potato starch	1.7 (18.1 %)
Wheat starch	3.4 (35.9 %)
Total	9.6

* Association des Amidonniers et Féculiers (AAF), 2006.

In small production units, mainly in Asia, starches are separated also from other starch-containing materials. These materials include: amaranth, arrowroot, banana, canna, cow cockle, faba/mung beans, kouzou, lentils, lotus roots, quinoa, sago palm, sorghum, sweet potatoes, taro, water chestnut, wild rice, and yam.

Starch Production in the EU (Table 7). The starch industry in Western Europe comprises 24 different companies with 68 plants, which produced 9.6×10^6 t of starch in 2005 with a share of 46 % maize starch (4.4×10^6 t), 35.9 % of wheat starch (3.4×10^6 t) and 18.1 % of potato starch (1.7×10^6 t).

EU Consumption of Starch and Starch Derivatives (Table 8). In 2005, 9.0×10^6 t of starch products were consumed in the EU, split into 5.1×10^6 t of starch hydrolysates (starch sweeteners), including isoglucose and polyols (57 %); 2.1×10^6 t of native starches (23 %) and 1.8×10^6 t of modified starches (20 %). 57 % of these 9.0×10^6 t were consumed in the food and 43 % in the nonfood sector.

Table 8. Consumption of starch and starch derivatives by sector in the EU in 2005*

Sector	Consumption	
	10^6 t	%
Food	5.13	57
Confectionery and drinks	2.7	30
Processed food	2.43	27
Nonfood	3.87	43
Paper and corrugating	2.52	28
Chemicals fermentation and other industrial products	1.26	14
Feeds	0.09	1
Total	9.0	

* Source: AAF 2006.

The breakdown by application areas/market sectors was as follows: sweets and drinks (confectionery, drinks, fruit processing) 30 %, processed food (convenience food, bakery, food ingredients and food preparations, dairy products and ice cream) 28 %, paper and corrugating (papermaking, corrugating and paper processing) 28 %, chemical, fermentation, and other industrial products 14 %.

Starch Production in the United States [153]**.** In 2005, 3.3×10^6 t of maize starch were produced in the USA and 3.03×10^6 t of starch products (corn starch, modified starches, dextrins) and 14.8×10^6 t of refinery products (glucose syrups, high fructose corn syrups, dextrose, corn syrup solids, maltodextrins, and fructose) were shipped in the United States. High fructose corn syrup (HFCS) 42 accounted for 4.7×10^6 t, HFCS 55 for 6.0×10^6 t of these 14.8×10^6 t. 136.6×10^3 t corn starch, 191×10^3 t modified starches derived from corn starch and 636×10^3 t of refinery products (glucose syrups, high fructose corn syrups, glucose, corn syrup solids, dextrins, and fructose) were exported.

Starch Production in Thailand [154]**.** Thailand is by far the largest producer of tapioca starch in the world. The other tapioca producing countries, such as Vietnam and Indonesia, are only of minor importance. In 2006 3.5×10^6 t of tapioca starch were produced in Thailand, 2.3×10^6 t were exported, the rest used for domestic supply. Of the exported 2.3×10^6 t 1.67×10^6 t were native (unmodified) and 638×10^3 t modified starches. The rest was exported as sago (granulated starch).

4. References

1. R. Whistler, J. BeMiller, E. Paschall (eds.): *Starch: Chemistry and Technology*, 2nd ed., Academic Press, New York 1984.
2. D. Bechtel, I. Zayas, L. Kaleikau, Y. Pomeranz, *Cereal Chem.* **67** (1990) 59.
3. G. Inglett (ed.): *Corn: Culture*, Processing, Products, AVI Publishing, Westport 1970.
4. Y. Pomeranz, L. Munck (eds.): *Cereals: A Renewable Resource*, Amer. Assn. Cereal Chemists, St. Paul 1981.
5. R. Whistler, J. Daniel in O. Fennema (ed.): *Food Chemistry*, Marcel Dekker, New York 1985, pp. 69 – 137.
6. R. Singhal, P. Kulkarni, *Starch/Stärke* **42** (1990) 5.
7. R. Singhal, P. Kulkarni, *Starch/Stärke* **42** (1990) 102.
8. O. Paredes-Lopez, M. Schevenin, D. Hernandez-Lopez, A. Carabez-Trejo, *Starch/Staerke* **41** (1989) 205.
9. A. Cortella, M. Pochettino, *Starch/Stärke* **42** (1990) 251.
10. Z. Nikuni et al., *Mem. Inst. Sci. Ind. Res., Osaka Univ.* **26** (1969) 1.
11. Z. Nikuni, *Staerke* **30** (1978) 105.
12. D. French, *J. Jpn. Soc. Starch Sci.* **19** (1972) 8.
13. D. French in R. Whistler, J. BeMiller, E. Paschall (eds.): *Starch: Chemistry and Technology*, 2nd ed., Academic Press, New York 1984, pp. 183 – 247.
14. A. M. Donald et al.: "Internal Structure of Starch Granules Revealed by Scattering Studies," in P. J. Frazier, P. Richmond (eds.): *Starch Structure and Functionality*, Royal

Society of Chemistry, London 1997, pp. 172 – 179.
15. L. Slade, H. Levine in R. Millane, J. BeMiller, R. Chandrasekaran (eds.): *Frontiers in Carbohydrate Research*, Elsevier, New York 1989, 215.
16. R. Tester, W. Morrison, *Cereal Chem.* **67** (1990) 551.
17. R. Tester, W. Morrison, *Cereal Chem.* **67** (1990) 558.
18. L. Jing-ming, Z. Sen-lin, *Starch/Stärke* **42** (1990) 96.
19. S. Craig, C. Maningat, P. Seib, R. Hoseney, *Cereal Chem.* **66** (1989) 173.
20. K. Meyer, M. Wertheim, P. Bernfeld, *Helv. Chim. Acta* **23** (1940) 865.
21. K. Meyer, P. Bernfeld, W. Hohenemeser, *Helv. Chim. Acta* **23** (1940) 885.
22. W. Hassid, R. McCready, *J. Am. Chem. Soc.* **65** (1943) 1157.
23. J. Blackwell, A. Sarko, R. Marchessault, *J. Mol. Biol.* **42** (1969) 379.
24. A. Sarko, R. Marchessault, *J. Am. Chem. Soc.* **89** (1967) 6454.
25. V. Rao, P. Sundararajan, C. Ramakrishnan, G. Ramachandran in G. Ramachandran (ed.): *Conformation of Biopolymers*, vol. **2**, Academic Press, London 1967, p. 721.
26. F. Senti, L. Witnauer, *J. Am. Chem. Soc.* **70** (1948) 1438.
27. D. Kreger, *Nature* **158** (1946) 199.
28. A. Frey-Wyssling, *Am. J. Bot.* **56** (1969) 696.
29. W. Shuey, K. Tipples: *The Amylograph Handbook*, Am. Assoc. Cereal Chem., St. Paul 1980.
30. E. Osman in R. Whistler, E. Paschall (eds.): *Starch: Chemistry and Technology*, vol. **II**, Academic Press, New York 1967, pp. 163 – 215.
31. M. Glicksman, R. Sand in R. Whistler, J. BeMiller (eds.): *Industrial Gums*, Polysaccharides and Their Derivatives, Academic Press, New York 1973, pp. 197 – 263.
32. P. White, I. Abbas, L. Johnson, *Starch/Stärke* **41** (1989) 176.
33. P. Roulet, W. MacInnes, D. Gumy, P. Würsch, *Starch/Stärke* **42** (1990) 99.
34. R. Stute, *Starch/Staerke* **42** (1990) 178.
35. C. Biliaderis, J. Zawistowski, *Cereal Chem.* **67** (1990) 240.
36. *Food Manufacture*, Sept. 1990, p. 21.
37. M. Wooton, M. Chaudry, *J. Food Sci.* **45** (1980) 1783.
38. S. Fleming, *J. Food Sci.* **47** (1981) 1.
39. S. Fleming, J. Vose, *J. Nutr.* **109** (1979) 2067.
40. R. Walton: *Comprehensive Survey of Starch Chemistry*, Chemical Catalog Co., New York 1928.
41. H. Brown, G. Morris, *J. Chem. Soc.* **55** (1889) 449.
42. D. French in R. W. Kerr (ed.): *Chemistry and Industry of Starch*, 2nd ed., Academic Press New York 1950, p. 158.
43. G. V. Caeser, E. E. Moore, *Ind. Eng. Chem.* **27** (1935) 1447.
44. W. Gallay, A. C. Bell, *Can. J. Res. Sect. B* **14** (1936) 360.
45. W. Gallay, A. C. Bell, *Can. J. Res., Sect. B* **14** (1936) 381.
46. J. Katz, *Textile Res. J.* **9** (1939) 146.
47. H. Schopmeyer, G. E. Felton, US 2 319 637, 1943.
48. W. Bechtel, *Cereal Chem.* **24** (1947) 200.
49. W. Bechtel, *J. Colloid Sci.* **5** (1950) 260.
50. E. Mazurs, T. Schoch, F. Kite, *Cereal Chem.* **34** (1957) 141.
51. S. Lansky, M. Kooi, T. Schoch, *J. Am. Chem. Soc.* **71** (1949) 4066.
52. H. Leach, T. Schoch, *Cereal Chem.* **39** (1962) 318.
53. H. Leach, L. McCowen, T. Schoch, *Cereal Chem.* **36** (1959) 534.
54. C. Naegeli, *Ann. Chem.* **173** (1874) 218.
55. B. Bellmas, DE 110 957, 1897.
56. C. Duryea, US 675 822, 1901.
57. R. Kerr, *Stärke* **4** (1952) 39.
58. J. Cowie, C. Greenwood, *J. Chem. Soc.* (1957) 2658.
59. W. Gallay, *Can. J. Res. Sect. B* **14** (1936) 391.
60. M. Wolfrom, A. Thompson, C. Timberlake, *Cereal Chem.* **40** (1963) 82.
61. *Posselts Textile J.* **21** (1917) 9.
62. W. Cathcart, *Textile World* **59** (1921) 2896.
63. C. Musselmann, E. Bovier, US 4 014 727, 1977.
64. J. Liebig, *Ann. Phys. Chem.* **15** (1829) 541.
65. A. Lieben, E. Reichardt, *Dtsch. Chem. Ges.* **8** (1875) 1020.
66. F. Farley, R. Hixon, *Ind. Eng. Chem.* **34** (1942) 677.
67. G. Felton, F. Farley, R. Hixon, *Cereal Chem.* **15** (1938) 678.
68. J. Schmorak, M. Lewin, *J. Polymer Sci., Part A* **1** (1963) 2601.
69. V. Prey, S. Fischer, *Staerke* **28** (1976) 125.
70. D. Lucas, C. Fletcher, *Paper Ind.* **40** (1959) 810.
71. V. Prey, S. Fischer, S. Klinger, *Stärke* **28** (1976) 259.

72. R. Mellies, C. Mehltretter, F. Senti, *J. Chem. Eng. Data* **5** (1960) 169.
73. J. Schmorak, D. Mejzler, M. Lewin, *Stärke* **14** (1962) 278.
74. J. Potze, P. Hiemstra, *Stärke* **15** (1963) 217.
75. C. Cairns, *TAPPI* **57** (1974) 85.
76. H. Brill, *TAPPI* **38** (1955) 522.
77. Anheuser-Busch Inc., US 3 767 826, 1973 (J. Fruin).
78. E. Alteriis, P. Parascandola, V. Scardi, *Starch/Stärke* **42** (1990) 57.
79. O. Wurzburg, C. Szymanski, *Agr. Food Chem.* **18** (1970) 997.
80. U.S. Department of Agriculture, US 2 989 521, 1961 (F. Senti, R. Mellies, C. Mehltretter).
81. Corn Products Co., US 2 801 242, 1957 (R. Kerr, F. Cleveland Jr.,
82. National Starch and Chemical Corp., US 2 935 510, 1960 (O. Wurzburg).
83. American Maize Products Co., US 2 328 537, 1943 (G. Felton, H. Schopmeyer).
84. M. Rutenberg, D. Solarek in R. Whistler, J. BeMiller, E. Paschall (eds.): *Starch: Chemistry and Technology*, 2nd ed., Academic Press, New York 1984, pp. 324 – 326.
85. F. Kite, T. Schoch, H. Leach, *Bakers Dig.* **31** (1957) 42.
86. World Health Organization, *Food Additive Series* 1974, no. 5, 329, 345.
87. A. E. Staley, DE 2 541 513, 1976 (R. Van Schanefelt, J. Eastman, M. Campbell).
88. National Starch and Chemical Corp., US 3 052 545, 1962 (J. Ducharme, H. Black Jr., , S. Leith).
89. National Starch and Chemical Corp., US 3 579 341, 1971 (N. Marotta, P. Trubiano).
90. American Maize Products, US 3 346 387, 1967 (J. Evans, C. MacWilliams).
91. National Starch and Chemical Corp., US 4 229 489, 1980 (C. Chin, M. Rutenberg).
92. Unilever, DE 2 837 088, 1979 (D. Chaudhuri, M. Stebles).
93. A. E. Staley, US 3 069 410, 1962 (C. Smith, J. Tuschoff).
94. PPG Industries, US 3 887 389, 1975 (J. Hedden).
95. Penick and Ford, US 3 438 913, 1969 (E. Hjermstad).
96. W. Jarowenko in N. Bicales (ed.): *Encyclopedia of Polymer Science and Technology*, vol. **12**, Interscience, New York 1970, p. 787.
97. R. Lohmar, C. Rist, *J. Am. Chem. Soc.* **72** (1950) 4298.
98. J. Mullen, E. Pacsu, *Ind. Eng. Chem.* **34** (1942) 1209.
99. R. Whistler, *Adv. Carbohyd. Chem.* **1** (1945) 279.
100. M. Rutenberg, W. Jarowenku, L. Ross, US 3 038 895, 1962.
101. C. Cross, E. Bevan, J. Tranquair, *Chem. Ztg.* **29** (1932) 2083.
102. H. Roberts in R. Whistler, E. Paschall (eds.): *Starch: Chemistry and Technology*, vol. **II**, Academic Press, New York 1967, chap. 13.
103. E. Middleton, US 1 682 220, 1928.
104. O. Wurzburg in R. Whistler (ed.): *Methods in Carbohydrate Chemistry*, vol. **4**, Academic Press, New York 1964, p. 286.
105. C. Smith, J. Tuschoff, US 2 928 828, 1960.
106. S. Takahashi, C. Maningat, P. Seib, *Cereal Chem.* **66** (1989) 499.
107. E. Degering in R. Kerr (ed.): *Chemistry and Industry of Starch*, 2nd ed., Academic Press, New York 1950, chap. 10.
108. C. Burkhard, E. Degering, *Rayon Text. Mon.* **23** (1942) 416.
109. A. Harsveldt, *Chem. Ind. (London)* 1961, 2062.
110. National Starch and Chemical Corp., US 3 481 771, 1969 (A. Doering).
111. J. Mullen, E. Pacsu, *Ind. Eng. Chem.* **35** (1943) 381.
112. R. Whistler, G. Hilbert, *Ind. Eng. Chem.* **43** (1944) 911.
113. R. Kerr, F. Cleveland, US 2 884 412, 1959.
114. R. Whistler, J. Daniel in R. Whistler, J. BeMiller, E. Paschall (eds.): *Starch: Chemistry and Technology*, 2nd ed., Academic Press, New York 1984, chap. 6.
115. E. Snyder in R. Whistler, J. BeMiller, E. Paschall (eds.): *Starch: Chemistry and Technology*, 2nd ed., Academic Press, New York 1984, pp. 661 – 674.
116. D. Christianson et al., *Cereal Chem.* **46** (1969) 372.
117. R. Hamilton, E. Paschall in R. Whistler, E. Paschall (eds.): *Starch: Chemistry and Technology*, vol. **II**, Academic Press, New York 1967, chap. 14.
118. Benckiser-Knapsack, Höchst, DE 2 426 404, 1975 (A. Kling, W. Traud, W. Hansi, H. Jalke).
119. R. Kerr, US 2 884 413, 1947.
120. E. Salay, C. Ciacco, *Starch/Stärke* **42** (1990) 15.
121. B. Hofreiter in J. Casey (ed.): *Pulp and Paper Chemistry and Chemical Technology*, 3rd ed., Wiley, New York 1981, p. 1475.
122. J. Benckiser, GB 1 233 637, 1971.
123. National Dairy Products Corp., GB 938 717, 1963.

124. M. Tessler, US 3 719 662, 1973.
125. American-Maize Co., US 3 238 100, 1966 (H. Meyer, R. Milloch, V. Shreeram, T. Tsuzuki).
126. H. Merkus, J. Mourits, L. deGalan, W. deJong, *Stärke* **29** (1977) 406.
127. S. El-Hinnaway et al., *Starch/Stärke* **34** (1982) 112.
128. T. Schoch, *TAPPI* **35** (1952) no. 7, 22 A.
129. E. Hjermstad in R. Whistler, J. BeMiller (eds.): *Industrial Gums*, Academic Press, New York 1973, pp. 601 – 615.
130. W. Roth, C. Mehltretter, *Food Technol.* **21** (1967) 72.
131. A. Harsveldt, *TAPPI* **45** (1962) 85.
132. National Starch and Chemical Corp., US 3 628 969, 1971 (R. Vilim, H. Bell).
133. Merck & Co., US 4 105 461, 1978 (J. Racciato).
134. General Foods Corp., US 3 669 687, 1972 (A. D'Ercole).
135. American-Maize Products Co., US 3 427 951, 1969 (F. Mitan, L. Jokay).
136. T. Shiroza, K. Furihata, *Agric. Biol. Chem.* **46** (1982) 1425.
137. E. Tasset, US 4 464 528, 1984.
138. R. Kerr, H. Neukom, *Stärke* **4** (1952) 255.
139. Dow Chemical, US 3 846 405, 1974 (J. McClendon).
140. H. Prietzel, *Stärke* **22** (1970) 424.
141. S. Parmerter, US 3 620 913, 1971.
142. USDA, US 3 087 852, 1963 (B. Hofreiter, G. Hamerstrand, C. Mehltretter).
143. E. Paschall in R. Whistler, E. Paschall (eds.): *Starch: Chemistry and Technology*, vol. **II**, Academic Press, New York 1967, pp. 403 – 422.
144. D. Greif, L. Gaspar in W. Reynolds (ed.): *Dry Strength Additives*, TAPPI Press, Atlanta 1980, pp. 95 – 117.
145. D. Greif, L. Gaspar in W. Reynolds (ed.): *Dry Strength Additives*, TAPPI Press, Atlanta 1980, chap. 4.
146. J. Marton, T. Marton, *TAPPI* **59** (1976) 121.
147. J. Marton, *TAPPI* **63** (1980) 87.
148. E. Mazzarella, L. Hickey, *TAPPI* **49** (1966) 926.
149. Hubinger, US 3 793 310, 1974 (L. Elizer).
150. D. Halabisky, *TAPPI* **60** (1977) 125.
151. Chemed Corp., US 4 088 600, 1978 (T. Tutein, A. Harrington, J. Jacob).
152. Fachverband der Stärke-Industrie e.V.: Zahlen und Fakten zur Stärke-Industrie Ausgabe 2006, Berlin 2006.
153. Corn Refiners Association: 2006 Corn Annual, Corn Refiners Association, Washington, D.C; http://www.corn.org.
154. http://www.thaitapiocastarch.org//trade.asp.

Steel

Dieter Schauwinhold, Düsseldorf, Germany (Chap. 1, Sections 6.2.2.1.1, 6.2.2.1.2, 6.2.2.1.5, 6.2.2.3, 6.2.2.4, 7.1)

Manfred Toncourt, VDEh, Düsseldorf, Germany (Chap. 2)

Rolf Steffen, VDEh, Düsseldorf, Germany (Section 3.1)

Dieter Janke, TU Bergakademie Freiberg, Eisenhütten-Institut, Freiberg, Germany (Section 3.2)

Klaus Schäfer, GMT, Georgsmarienhütte, Germany (Section 3.3, 3.4)

Hatto Jacobi, Mannesmannröhren-Werke AG, Duisburg, Germany (Sections 3.5.1 – 3.5.3)

Rudolf Hammer, Thyssen Stahl AG, Düsseldorf, Germany (Sections 3.5.1 – 3.5.3)

Robert Hentrich, Krupp Stahl AG, Siegen, Germany (Section 3.5.4)

Lothar Kucharcik, Verein Deutscher Giessereifachleute (VDG), Düsseldorf, Germany (Section 3.5.5)

Herbert Wiegels, RWTH Aachen, Institut für Bildsame Formgebung, Aachen, Germany (Section 4.1)

Horst M. Aichinger, VDEh, Düsseldorf, Germany (Section 4.2.1)

Friedhelm Sänger, Universität–Gesamthochschule Duisburg, Germany (Sections 4.2.2.1 – 4.2.2.3)

Harold Walter, Dortmund, Germany (Sections 4.2.2.1 – 4.2.2.3)

Rolf Kümmerling, Mannesmannröhren-Werke AG, Mühlheim/Ruhr, Germany (Section 4.2.2.4)

Ulrich Uhl, Hattingen, Germany (Section 4.2.3.1)

Werner Decker, Industrieverband Deutscher Schmieden e.V., Hagen-Ernst, Germany (Section 4.2.3.2)

Roger Pankert, Union Miniere, Balen, Belgium (Section 4.3, Chap. 5)

Hans Hougardy, Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany (Sections 6.1.1, 6.1.3)

Winfried Dahl, RWTH Aachen, Institut für Eisenhüttenkunde, Aachen, Germany (Sections 6.1.2.1, 6.1.2.3)

Hans Jürgen Grabke, Max-Planck-Institut für Eisenforschung, Düsseldorf, Germany (Section 6.1.2.2)

Ulrich Kalla, VDEh, Düsseldorf, Germany (Section 6.2.1)

Gerhard Kalwa, Kaarst, Germany (Sections 6.2.2.1.3, 6.2.2.1.4, 6.2.2.1.6 – 6.2.2.2.3)

Reinhard Winkelgrund, Stahl-Informations-Zentrum, Düsseldorf, Germany (Section 7.2)

Heinz-Lothar Bünnagel, Wirtschaftsvereinigung Stahl, Düsseldorf, Germany (Chap. 8)

Volker Brückmann, Wirtschaftsvereinigung Stahl, Düsseldorf, Germany (Chap. 8)

1. Introduction 4
2. History 8
2.1. From Prehistoric Times to the Middle Ages 8
2.1.1. Native Iron 8
2.1.2. Iron from Ores 8
2.1.3. Technology of Iron Production . 9
2.2. From the Middle Ages to the 1800s 9
2.2.1. Bloomery Furnaces 9
2.2.2. Blast Furnaces 9
2.2.3. Metal Shaping 10
2.3. The Industrial Age 11
2.3.1. Pig Iron Production 12
2.3.2. Steel Production 12
2.3.2.1. Crucible Steel 12
2.3.2.2. Puddled Steel 13
2.3.2.3. Ingot Steel 13
2.3.3. Rolled Steel Production 14
2.3.3.1. Puddled Steel and Steam Power 14
2.3.3.2. Ingot Steel and Mass Production 14
3. Crude Steel Production 16
3.1. Raw Materials 16
3.1.1. Hot Metal 16
3.1.2. Scrap 16
3.1.3. Sponge Iron 20

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_063

3.1.4. Lime 21
3.2. Physical and Chemical Fundamentals 22
3.2.1. Thermodynamics 22
3.2.1.1. Metallic Systems Based on Iron 22
3.2.1.2. Slag Systems 26
3.2.1.3. Metal – Slag Reactions 26
3.2.1.3.1. Refining Reactions 26
3.2.1.3.2. Dephosphorization 28
3.2.1.3.3. Desulfurization 28
3.2.1.3.4. Deoxidation 29
3.2.1.4. Metal – Gas Reactions 30
3.2.2. Kinetics and Mass Transfer . . . 31
3.2.2.1. Kinetics of Heterogeneous Reactions 31
3.2.2.2. Flow and Mass Transfer Coefficients 32
3.2.2.2.1. Metal – Glas Phase Boundaries with Nonturbulent Flow 32
3.2.2.2.2. Metal – Slag Phase Boundaries with Nonturbulent Flow 32
3.2.2.2.3. Solid (Metal or Slag) – Liquid Phase Boundaries with Nonturbulent Flow 33
3.2.2.2.4. Mass Transfer with Turbulent Flow 33
3.3. Production Processes 33
3.3.1. Oxygen-Blowing Processes . . . 34
3.3.1.1. LD and OBM Processes 35
3.3.1.2. Other Developments 38
3.3.2. Electric Steel Process 38
3.3.2.1. Special Processes 40
3.3.2.2. The Siemens – Martin Open Hearth Process 41
3.3.3. Production of Stainless Steels . 42
3.4. Secondary Metallurgy 43
3.4.1. Steel Treatment at Atmospheric Pressure 43
3.4.1.1. Deoxidation 44
3.4.1.2. Desulfurization 45
3.4.1.3. Dephosphorization 45
3.4.1.4. Temperature Control 45
3.4.2. Vacuum Treatment 47
3.4.2.1. Decarburization 48
3.4.2.2. Nitrogen Removal 48
3.4.2.3. Dehydrogenation 48
3.5. Casting and Solidification . . . 49
3.5.1. Fundamentals 49
3.5.2. Ingot Casting 54
3.5.3. Continuous Casting 58
3.5.3.1. The Growth of Strand Casting . 58
3.5.3.2. Productivity 59
3.5.3.3. Mold Oscillation and Shell Formation 61
3.5.3.4. Control of Molten Steel Level and Gap Lubrication 62
3.5.3.5. Support Roller Apron and Secondary Cooling Zone 62
3.5.3.6. Automation 62
3.5.3.7. Quality 63
3.5.3.8. Horizontal Strand Casting 63
3.5.3.9. Hot Charging or Direct Rolling 64
3.5.3.10. Casting Thin Slabs 64
3.5.3.11. New Processes for Hot Steel Strip Production 65
3.5.3.12. Rationalization 66
3.5.4. Consumable Electrode Remelting Processes 66
3.5.4.1. Remelting Processes 66
3.5.4.2. General Aims of Remelting . . . 67
3.5.4.3. Remelting Apparatus 68
3.5.4.4. Results and Applications 68
3.5.5. Cast Steel and Cast Iron 68
3.5.5.1. Cast Steel 68
3.5.5.2. Cast Iron 70
4. Forming 73
4.1. Fundamentals 74
4.1.1. Causes of Plasticity 74
4.1.2. Workability 74
4.1.3. Cold Forming 75
4.1.4. Hot Forming 75
4.1.5. Forming Processes 76
4.1.6. Aims of Forming Technology . 76
4.1.6.1. Combination of Casting and Forming 77
4.1.6.2. Thermomechanical Treatment . 77
4.1.6.3. Simulation of Forming Processes 77
4.1.6.4. Quality Information Systems . . 77
4.2. Hot Working 78
4.2.1. Reheating Furnaces 78
4.2.2. Hot Rolling 81
4.2.2.1. Semifinished Products 81
4.2.2.2. Flat Products 82
4.2.2.3. Long Products 86
4.2.2.4. Tubes 93
4.2.3. Forging 101
4.2.3.1. Open Die Forging 101
4.2.3.2. Closed Die Forging 103
4.3. Cold Rolling 105
4.3.1. Introduction 105
4.3.2. Pickling 106
4.3.3. Rolling 107
4.3.4. Annealing 108
4.3.5. Skin-Pass Rolling 110
4.3.6. Stainless Steels 110
4.3.7. Future Developments 111
5. Surface Coating 111
5.1. Introduction 111

5.2. Hot-Dip Coating 112
5.2.1. Plant and Processes 112
5.2.2. Other Coating Processes 115
5.3. Electrolytic Coating (Electroplating) 118
5.3.1. Plant and Processes 118
5.3.2. Electroplating Cells 118
5.3.3. Coating Types 119
5.4. Vacuum Vapor Deposition . . 120
5.4.1. Plant and Processes 120
5.4.2. Further Developments 120
5.5. Coil Coating 121
5.5.1. Plant and Processes 121
5.5.2. Coating Systems 122
5.6. Roll-Bonded Cladding 123
5.6.1. Principles 123
5.6.2. The Process 123
5.6.3. Variations 123
5.7. Summary 124
6. Uses 125
6.1. Fundamentals 125
6.1.1. Microstructure 125
6.1.1.1. Introduction 125
6.1.1.2. Phase Diagrams of Iron Alloys . 126
6.1.1.3. Definitions 126
6.1.1.4. Microstructures in Plain Carbon Steels 128
6.1.1.5. Microstructures of Alloy Steels 131
6.1.2. Properties 132
6.1.2.1. Mechanical Properties 132
6.1.2.1.1. Behavior Under Unidirectional Loads, at and Below Room Temperature 132
6.1.2.1.2. Behavior Under Cyclic Loading 154
6.1.2.1.3. Behavior at Higher Temperatures 156
6.1.2.2. Chemical Properties 160
6.1.2.2.1. Introduction 160
6.1.2.2.2. Uniform Corrosion 160
6.1.2.2.3. Atmospheric Corrosion 163
6.1.2.2.4. Passivation 163
6.1.2.2.5. Pitting Corrosion 164
6.1.2.2.6. Crevice Corrosion 164
6.1.2.2.7. Intergranular Corrosion of Stainless Steels 164
6.1.2.2.8. Stress Corrosion Cracking 165
6.1.2.2.9. Hydrogen Adsorption and Hydrogen Embrittlement 165
6.1.2.2.10. Oxidation of Iron 165
6.1.2.2.11. Oxidation of Carbon Steels and Low-Alloy Steels 166
6.1.2.2.12. High-Temperature Steels 166
6.1.2.2.13. Effects of Chlorine in Oxidation 167
6.1.2.2.14. Sulfidation of Iron and Steel . . 168
6.1.2.2.15. Carburization 168
6.1.2.2.16. Nitriding 169
6.1.2.2.17. Decarburization, Denitriding, and Hydrogen Attack 169
6.1.2.3. Physical Properties 170
6.1.2.3.1. Pure Iron 170
6.1.2.3.2. α-Iron Solid Solutions 173
6.1.2.3.3. γ-Iron Solid Solutions 174
6.1.2.3.4. Other Effects of Structure 174
6.1.3. Heat Treatment 174
6.1.3.1. Time – Temperature Transformation Diagrams 174
6.1.3.2. Heat Treatment of Transformable Steels 175
6.1.3.2.1. Austenitizing 175
6.1.3.2.2. Transformations 176
6.1.3.2.3. Effect of Dimensions 177
6.1.3.2.4. Quenching and Tempering . . . 177
6.1.3.2.5. Spheroidized Annealing 178
6.1.3.2.6. Thermomechanical Treatment . 178
6.1.3.2.7. Surface Hardening 179
6.1.3.3. Heat Treatment Without Transformation 180
6.1.3.3.1. Recrystallization 180
6.1.3.4. Simulation of Heat Treatment . 181
6.2. Uses 181
6.2.1. Classification of Steel and Steel Products 181
6.2.1.1. General 181
6.2.1.2. Definitions for the Classification of Grades of Steel 182
6.2.1.3. Definitions of Steel Products . . 183
6.2.1.4. Designation Systems for Steel . 183
6.2.2. Types of Steel 183
6.2.2.1. Steels where the Mechanical Properties are Essential 183
6.2.2.1.1. Normal- and High-Strength Structural and Reinforcing Steels 184
6.2.2.1.2. Steels Suitable for Heat-Treatment and Surface Hardening 191
6.2.2.1.3. Steels for Elevated Temperature Applications 195
6.2.2.1.4. Low-Temperature Steels 199
6.2.2.1.5. Wear-Resistant Steels 200
6.2.2.1.6. Steels for Pipelines 203
6.2.2.2. Steels with Special Chemical Properties 206
6.2.2.2.1. Stainless Steels 206
6.2.2.2.2. Heat-Resistant Steels 210
6.2.2.2.3. Steels for Hydrogen Service at Elevated Temperatures and Pressures 213
6.2.2.3. Steels with Special Physical Properties 215

6.2.2.3.1. Soft Magnetic Steels 215
6.2.2.3.2. Other Steels with Special Physical Properties 217
6.2.2.4. Steels with Special Processing Properties 218
6.2.2.4.1. Steels for Flat Products for Cold Forming 218
6.2.2.4.2. Steels Suitable for Cold Extrusion and Cold Heading 220
6.2.2.4.3. Machining or Free-Cutting Steels 222
7. Environmental Protection . . . 223
7.1. Environmental Aspects of Steel Production and Processing . . 223
7.1.1. Production of Steel and Steel Products 223
7.1.2. Steel Processing and Steel in Use 225
7.2. Steel Recycling 227
7.2.1. The Tradition of Steel Recycling 227
7.2.2. Types of Scrap 227
7.2.3. Scrap Processing 230
7.2.3.1. Separation 230
7.2.3.2. Shearing Machines 230
7.2.3.3. Presses 230
7.2.3.4. Shredders 231
7.2.4. Factors Influencing Recycling . 231
7.2.4.1. Alloy Steels 231
7.2.4.2. Surface Coatings 231
7.2.4.3. Pretreatment of Used Steel Products 232
7.2.5. Economic and Logistic Aspects 233
7.2.5.1. Scrap Consumption 233
7.2.5.2. World Market for Scrap 233
7.2.5.3. Infrastructure and Logistics . . . 234
8. Economic Aspects 234
8.1. World Steel Production, Consumption, and Trade 234
8.2. Steel Intensity and Weight Saving in Steel 237
8.3. Capital Investment and Subsidies 238
8.4. Future Prospects 240
9. References 241

1. Introduction

Nowadays the term steel is understood to include not only all forgeable iron-based materials, but also all highly alloyed metallic materials in which the element iron (Fe) is an important component, but which are not necessarily forgeable. With few exceptions, the carbon content is $< 2\,\%$.

For over 3286 years, steel has made a major contribution to human development, e.g., in tools for cultivating the soil and processing stone and almost all other materials, as a construction material for steel and reinforced concrete structures, in transport technology, for the generation and distribution of energy, for the fabrication of machinery and equipment (including equipment for the manufacture of plastics), in the household, and in medicine. It remains, for the forseeable future, by far the most important material for the maintenance and improvement of our quality of life.

The outstanding importance of steel is the result of its ready availability and its versatility. The earth's crust contains ca. 5 wt % iron, making it the fourth most abundant element after oxygen (46 %), silicon (28 %), and aluminum (8 %). Rich deposits of iron ores are available in many parts of the world (→ Iron, Chap. 2.1.1.). Moreover, the free energy required to isolate iron from its oxidic ores is less than half of that required for aluminum [1].

The versatility of steel is due to the polymorphism of the iron crystal and its ability to alloy with other elements, forming solid solutions or compounds. The microstructure of steel in a finished component can be adjusted by means of the chemical composition, the forming conditions and a wide variety of possible heat treatments. The attainable tensile strength ranges from ca. 300 N/mm^2 for deep-drawing sheet steel (e.g., for automotive body parts that are difficult to draw) to >2000 N/mm^2 for critical components in aircraft. Tensile strengths as high as 2600 N/mm^2 are achieved in 0.15 mm diameter drawn wire for steel cord used in radial tires.

Cryogenic steels with high strength and good toughness at very low temperatures are used for the transport and storage of liquefied gases at temperatures of $<-200\,°C$. Other steels with good properties at temperatures of 650 – 700 °C and above are used in power station equipment and gas turbines.

Highly developed soft magnetic steels are essential in the construction of transformers. Steel is also used to make permanent magnets. Non-

magnetizable steels have also been developed for use in electrical technology, shipbuilding, and physics research. Wear-resistant steels are used in rock-crushing machines and in industrial stirring equipment. Machine tools, used for metal cutting, require steels of the highest possible hardness to endow stability to the cutting edge. Other steels with very good machinability have been developed and are used for the economic manufacture of complex turned parts, or for mass production on high-speed automated equipment. Chemically resistant steels are essential in the chemical and foods industries, as well as in household equipment. For the majority of steel grades – more than 2500 are available today – very good welding properties are important, and here steel has an advantage over competing materials.

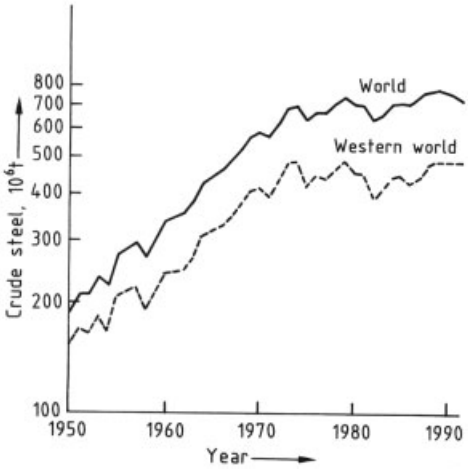

Figure 1. World crude steel production 1950 – 1991

Modern knowledge of controlling the microstructure of steel, and hence its properties, offer opportunities to match steel products to new sets of requirements [2].

Unlike brick or concrete buildings, steel structures can be dismantled relatively easily. Furthermore, almost 100 % of the steel can be recovered from steel-containing products and can be remelted to yield steels of similar or higher quality. In this respect, iron and steel are superior to all competitive materials.

The great importance of steel in the world's economy is also exemplified by production figures. In the early 1900s, total world production of steel was less than 35×10^6 t/a. In 1940 it was 140×10^6 t/a. The figures for the period after 1950 (Fig. 1) indicate a surprisingly large growth in world crude steel production after World War II. Up to the mid-1970s, this was mostly due to those developed countries with the greatest rate of economic growth, such as Japan, the six founding countries of the European Community, and also the former Soviet Union. In the United States, growth had already ceased by the mid-1960s due to market saturation.

A rapid increase in steel production also took place in some Latin American countries, continuing until the mid-1980s. In many countries in Asia, Africa, and the Middle East, new steel industries were built up, or existing capacity was increased. The developments in some Asian countries during the last 16 years are remarkable (see Table 1).

Table 1. Development of crude steel production in Asian countries∗

Country	Production, 10^3 t/a		
	1975	1985	1991
China	23 903	46 700	70 400
South Korea	1 994	13 539	26 000
Taiwan	680	5 088	11 000
India	7 991	11 140	17 100

∗ Source: International Iron and Steel Institute.

The relationship between total economic development and steel production can be seen clearly from these figures. Figure 1 also shows that world production of crude steel has stagnated since the mid-1970s, apart from the usual market fluctuations. A sharp rise in crude steel production in countries with rapidly increasing industrialization contrasts with zero or even negative growth in steel production in countries that are already highly industrialized (e.g., the United States). Several reasons can be suggested:

The steel consumption of a country is determined by the "specific market demand", i.e., production plus imports minus exports of steel products expressed in kilograms crude steel per capita. In industrialized countries, this reaches saturation at ca. 600 ± 100 kg, which can only be exceeded in special circumstances. If there are sections of industry with extremely high steel consumption (shipbuilding or automotive industries), and if these are strongly export oriented,

a value of >700 kg per capita can be reached because the exported finished products do not remain in that country. Figure 2 shows that the saturation value of 600 kg was reached as early as 1950 in the United States. Other industrialized countries reached this figure in the 1970s, and this reduced the growth of steel production. Also, the expansion of the steel industry in countries which had been importing rolled steel made it impossible for steel exports from the industrialized countries to grow. Extrapolation of the Figure 2 into the 1980s is problematical because the figure "kilogram crude steel per capita" is now no longer comparable with figures up to ca. 1975 due to technical developments (described below) in rolled steel products.

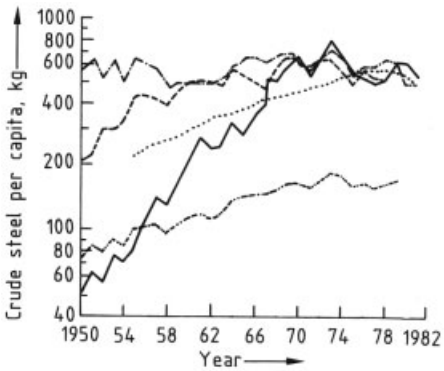

Figure 2. Specific market demand for crude steel (production + imports − exports) [1]
–·– = United States; – – – – = Federal Republic of Germany; ····· = former Soviet Union; —·— = Japan; — ··— = World

From the second half of the 1960s, use of the continuous casting process in steelworks increased, approaching 20 % by 1975 worldwide. By 1991, the proportion of continuously cast steel reached 67 % worldwide. This process gives an increase in yield of ca. 12 – 15 % of crude steel for rolled steel products, compared with ingot casting. Hence, up to 15 % less crude steel is required to produce the same amount of rolled steel product for the steel processing industry. Other improvements in yield productivity in steel plants and in steel- consuming industries also reduced the consumption of crude steel without jeopardizing the use of steel for a particular end product.

Table 2 lists all countries with a crude steel output of more than 10^6 t/a during 1991 and 1992 and gives the percentage of continuously cast steel.

Table 2. Countries with a crude steel output $>10 \times 10^6$ t/a (1991 and 1992) ∗ with portion of continuously cast steel shown

Country	Production, 10^6 t/a		Portion of continuously cast steel, %	
	1991	1992	1991	1992
Former Soviet Union	132.8	116.8	17.7	17.1
Japan	109.6	98.1	94.4	95.4
United States	79.7	84.3	75.7	78.9
China	70.4	80.0	26.5	30.0
Germany	42.2	39.7	89.5	92.0
South Korea	26.0	28.0	96.4	96.8
Italy	25.1	24.9	95.1	96.1
Brazil	22.6	23.9	56.0	58.6
France	18.4	18.0	95.0	95.2
India	17.1	18.1	14.3	
United Kingdom	16.5	16.1	85.5	87.0
Canada	13.0	13.9	83.6	86.5
Spain	12.9	12.3	91.8	93.2
Czechoslovakia	12.1	11.1	17.0	21.9
Belgium	11.3	10.3	92.2	93.9
Taiwan	11.0	10.7	94.6	94.9
Poland	10.4	9.8	8.6	9.1

∗ Source: International Iron and Steel Institute 1992.

In 1978, however, for the countries represented in Figure 2, the percentages of continuously cast steel were:

Former Soviet Union	9.5 %
Japan	46.2 %
United States	15.2 %
Federal Republic of Germany	38.0 %

The large differences between the 1978 and 1991 figures, means that comparison of the number of kilograms crude steel per capita is of questionable value.

Below a continuously cast steel portion of ca. 85 – 90 % "kilograms crude steel per capita" is an indication of the state of modernization of the steel industry of a country. Above this range, the production program and the proportion of steel products not produced by continuous casting (e.g., steel castings, heavy forgings) play an important part in crude steel production. Only countries with a crude steel output in 1991 of max. 10^6 t/a (e.g., Switzerland and New Zealand) produce 100 % of their steel by continuous casting.

A further reason for the absence of the usual growth in crude steel production in the 1960s in

the industrialized countries, must be attributed to the development of higher strength steel grades. In the last 15 – 20 years, these grades of steel have been increasingly used by steel processors. Especially in the motor vehicle industry, high-strength steels have been used to improve constructional design through weight reduction. Consequently, there is a decrease in the amount of steel used, in addition to that achieved with the replacement of steel by competing materials such as aluminum or plastics.

When assessing the technological and economic significance of steel, it is interesting to examine the different finished steel products and how their share of the total production of rolled steel has changed over the years. There are two groups: "long products" (i.e., beams, steel shapes, steel bars, railway-track material, piling sections, and wire rod) and "flat products" (i.e., plates, sheets, strip-steel and universal-plate; see also Section 6.2.1).

In 1870 3×10^6 t of railway-track material accounted for 43 % of total world production of rolled steel products (Fig. 3), while in 1991 the output of railway-track material, which underwent a threefold increase to 9×10^6 t, accounted for barely 1.2 % of total production. In modern industrialized countries, 60 – 70 % of rolled steel production consists of flat products, mainly hot- and cold-rolled strip and sheet, coated sheet, tinplate and electric sheet, all with thicknesses in the range of 0.15 – 3 mm. The production of cold-rolled stainless steel strip has increased sharply during the last 25 years, as shown in Figure 4.

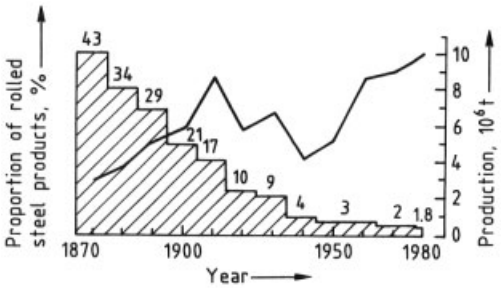

Figure 3. World production of steel for railway-track material lines
Solid line: total production of railway-track material; Hatched area: railway-track material as percentage of the total world production of rolled steel

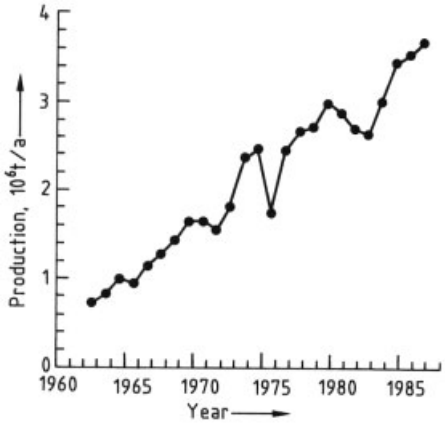

Figure 4. Production of cold-rolled stainless steel strip by western countries

There is intensive world trade in products of the steel industry, especially among major steel producing countries. Only in a few cases does this involve special products or special steel grades, for which some producers have a leading position. International competition is severe, and producers must offer consistently high-quality products. The large producers have therefore built up comprehensive quality assurance systems over many years, conforming to ISO Standard 9002 and local regulations such as ASME in the United States, Lloyds Register in the United Kingdom, and TüV in Germany. The aim is to control all steps of the manufacturing process, and to use statistical methods to ensure that the finished product has the desired properties, dimensional tolerances, and a defect-free surface finish. The intensive testing of finished products that was formerly carried out can then be largely dispensed with, and replaced by the testing of random samples, the results of which form part of the control loop.

Such quality-assured production is a prerequisite for the direct processing of sheet coils on automatic press lines without any inspection (even for exposed autobody parts).

A further important aid to international trade in steel is the existence of common standard specifications and terms of delivery. Specifications for steel products have been formulated by ISO over many years. However, these are often only the first steps toward harmonization of the various interpretations and technical possibilities. In Europe, since 1986, EN Standards for all steel products and testing methods have been

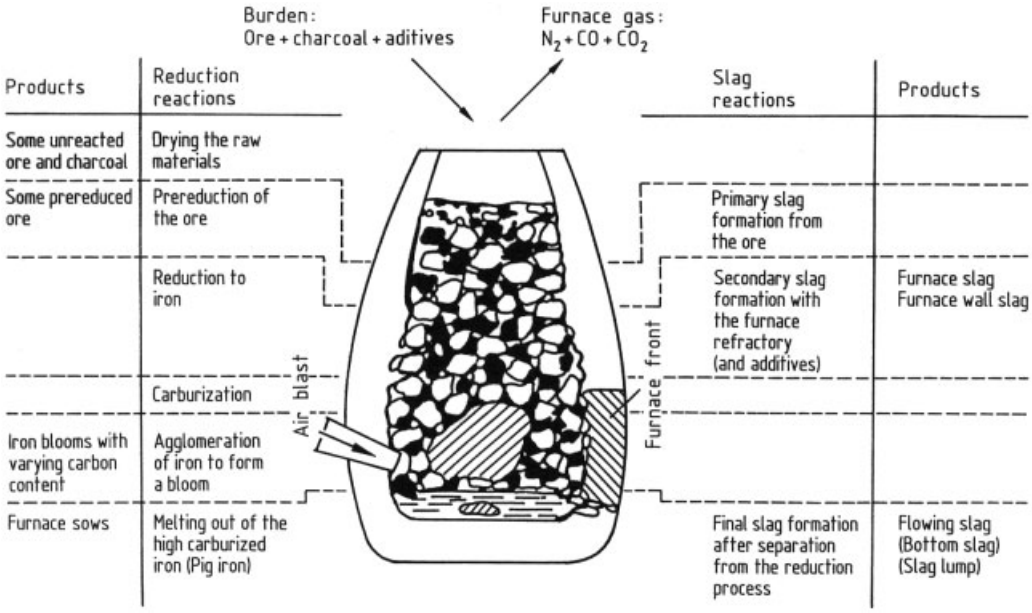

Figure 5. Furnace reactions for the production of iron by the direct method in a bloomery hearth furnace (after D. HORSTMANN)

compiled by the ECISS (European Committee for Iron and Steel Standardization). These standards must be adopted via CEN (Comité Européen de Normalisation) into the national standards of all European nations. They will certainly achieve importance outside Europe, too.

2. History

2.1. From Prehistoric Times to the Middle Ages

2.1.1. Native Iron

Since the earliest times, humans have used iron, the second most abundant metal in the earth after aluminum. It was first discovered in the form of the native element, which occurs only rarely as tellurian iron of volcanic origin, and mainly as meteoric iron. The Egyptians, Sumerians, Khatti, and Hittites must have had knowledge of its origin, as they called it ore, or metal, from heaven. Meteoric iron can be identified from its high nickel content (ca. 5 – 25 %), and the typical Widmanstätten structure, formed on solidification. The oldest known examples of worked meteoric iron are beads from Gerzeh (3500 B.C.) and a dagger from Ur (3000 B.C.)

2.1.2. Iron from Ores

It is not known when, where, or how iron was first deliberately produced from its ores. It is often said that the technique of extracting iron was developed from observations made during copper extraction, and that the idea suggested itself simultaneously in many places with a long tradition of copper smelting. A geographical analysis can be carried out from linguistic similarities, archeological finds, and documents. According to present knowledge, iron was first smelted in eastern Asia Minor and northern Mesopotamia around 2000 – 1500 B.C., and possibly also in the plains north of the Caucasus. As the Chinese were the first to produce a high-carbon, high-phosphorus (up to 7 %) liquid castable iron, it is conceivable that they also succeeded in winning iron from its ores at a very early date. Around 500 B.C., knowledge of iron winning spread through Asia Minor to North Africa, parts of Asia, and all of Europe.

2.1.3. Technology of Iron Production

At first, the iron was smelted in a bloomery fire, later in a bloomery hearth furnace, and crucibles. The iron ore was charged with charcoal, and later, according to the type of gangue material, with other fluxes, into a hearth or the shaft of a furnace. The charcoal was burned by a blast caused either by a natural draught (convection) or by hand- or foot-operated bellows. The rising combustion gases dry the materials (Fig. 5), and reduce the ore at ca. 900 – 1100 °C, forming a primary slag of wustite (FeO). From the gangue, a molten liquid slag forms, from which, in a further reduction stage, pure solid iron separates out, with formation of secondary slag. Since no appreciable carbonization of the iron takes place, it does not liquefy, even in the lower part of the furnace (temperature ca. 1200 –1300 °C). The product, the so- called bloom, consists of iron with slag inclusions and residual charcoal. These were separated, and the blooms were reheated and forged into bars. Up to the late Middle Ages iron, or rather steel, was obtained exclusively by this process directly from the ore, except in China. As a result of improved bellow technology, the furnace volume grew. About 900 years ago, this development culminated in bloomery furnaces with reached heights of 4 m, thanks to the wind pressure from waterwheel-powered bellows.

In addition to the technique of iron winning, knowledge of the properties of iron and how to modify them also developed empirically and involved little understanding of cause and effect, even up to the 1700s. It was recognized at an early stage that iron could be hardened by heating (with carburization) followed by rapid cooling. As early as the 900s B.C. crucible furnaces were used for carburizing iron in Gerar, Palestine. Moreover, the technique of combining high- and low- carbon steels was used to achieve a desired combination of properties, such as toughness and hardness, e.g., in the production of Damascus steels.

2.2. From the Middle Ages to the 1800s

In the 900s, technology was revolutionized by the waterwheel, with far-reaching effects on iron production. Water power became more important than the availability of ore when choosing the location for iron smelters. Thus, the preferred smelter location changed from the mountains, with proximity to ore and wood, to the river valleys, with their availability of flowing water.

The blast produced by water-driven bellows enabled larger furnaces to be operated. These were originally bloomery furnaces, but became flowing furnaces in the 1100s and 1200s. Water power was also used to operate tail hammers, lift hammers, etc. This development took place almost simultaneously in the alpine regions, and in western and northern Europe, which is not surprising in view of the extensive trade in medieval Europe. The historical development of iron and steel production equipment, and the growth of total world steel production, is summarized in Figure 6.

2.2.1. Bloomery Furnaces

In principle, bloomery furnaces were simply an increase in furnace size. They were up to 6 m high, operated batchwise, and produced a bloom up to 100 kg after ca. 15 h operation. The earlier furnaces produced a bloom of only 20 kg. Refining the blooms is a two-stage operation: a metallurgical treatment in the refining furnace (slag removal, homogenization, and, if necessary, decarburization); followed by hammering into bars, and a mechanical mixing operation in which bars of steel of various qualities are hammered together, forming a fairly uniform product. Bloomery furnaces, like the small Corsican and Catalan forges that used the same principle, were still producing a small fraction of the total production up to the 1800s.

2.2.2. Blast Furnaces

Together with the use of water power, the main revolution in iron smelting in Europe was the discovery of the two-stage indirect production of steel – the method still used today. In the first stage, liquid pig iron with a high carbon content is produced in the blast furnace. In the second stage, this pig iron is decarburized and refined to steel in a finery fire. The improved air pressure and flow rate enabled furnace temperatures $> 1200\,°C$ to be achieved. The furnace volume

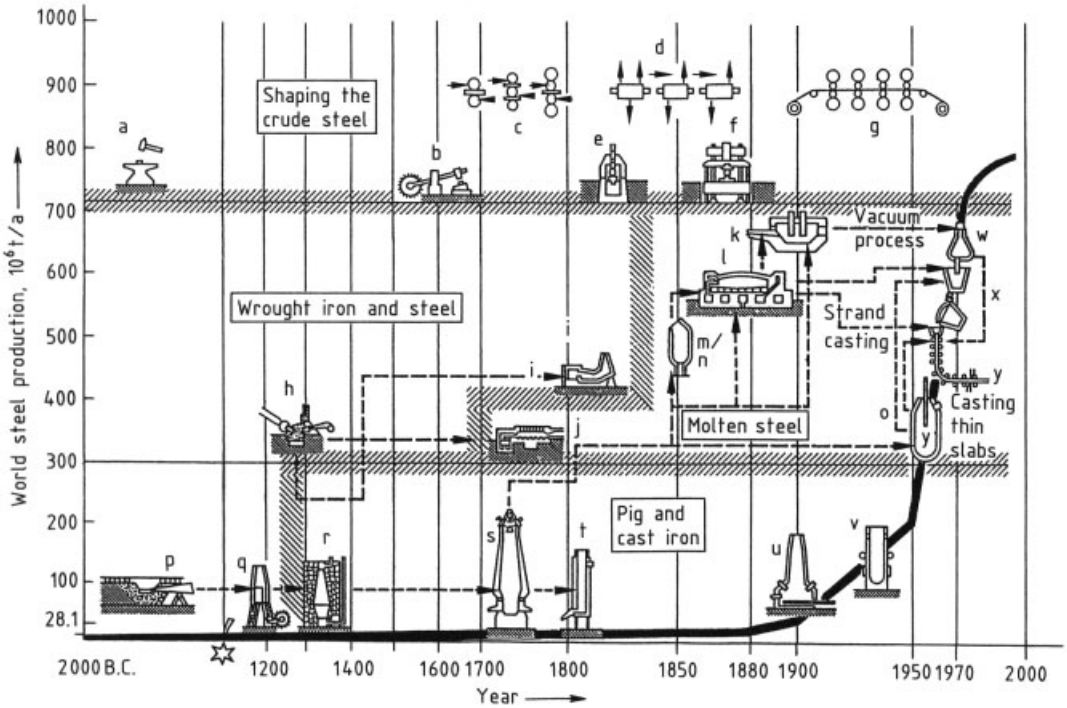

Figure 6. Development of steel production processes and forming, with a graph of annual steel production [13]
a) Hammer and anvil; b) Tail hammer; c) Two-, three- and four-high rolling stands; d) Open rolling mill; e) Steam hammer; f) Forging press; g) Continuous rolling mill; h) Finery fire; i) Puddling furnace; j) Crucible furnace; k) Electric furnace; l) Open hearth furnace; m) Bessemer converter (acid lining); n) Thomas converter (basic lining); o) Oxygen steelmaking converter; p) Finery fire; q) Bloom furnace; r) Charcoal blast furnace; s) Coke blast furnace; t) Cupola furnace; u) Electric shaft furnace; v) Low shaft furnace; w) Vacuum degassing; x) Continuous casting; y) Thin slab casting

could be increased, leading to a longer dwell time for the burden as it passed down the furnace shaft. This led to improved reduction of the iron ore (the yield increasing from 50 to > 85 %), and carburization of ca. 2.5 – 5.0 %. The lowering of the melting point led to the production of liquid pig iron for the first time, and considerably improved output. The furnaces could be charged without interrupting the process, and the liquid iron was directly cast into finished products (furnace plates and cannon balls) or bars for sale or captive processing in the finery shown. Here, the crushed pig iron was smelted in the finery hearth and decarburized under an oxidizing slag. Steel blooms were formed from the molten metal as it thickened, and these were then hammer forged into semifinished products. The furnace workers were able in some degree to control the properties of the product, but their empirical knowledge was mainly limited to processing certain types of ore or pig iron. An example is Osemund iron, which was ductile and, therefore, used for wire production. Steel production and processing centers were established in Austria, France, Belgium, United Kingdom, Germany, Sweden, and Russia. Iron production in Europe increased from 20 000 – 30 000 t/a (when small blast furnaces were introduced in ca. 1200 A.D.) to ca. 150 000 t/a at the inception of coke-based metallurgy in the mid-1700s.

2.2.3. Metal Shaping

Up to the 1500s, iron was processed only by forging and drawing, but, as the indirect method of iron production became widespread, water-powered rolling mills came into use. However, production rates of the blast furnace, the finery fire, and the hammer forge were roughly comparable, and hot rolling was unable to produce a product comparable to forged steel. Therefore,

hot rolling was only occasionally used for the treatment of forged material. Its most important application emerged later in the so-called iron-splitting works, in which strips were produced from steel sheet (e.g., for nail production). These operations were common in almost all iron-producing countries until the mid-1600s.

The range of rolled products increased during the early 1700s. In Sweden, Christoph Polhems was able to produce rolled steel with various profiles for knife blades and files. In England, John Hanbury produced thin steel sheet on a pillar rolling mill for tinplate manufacture. This type of equipment was being used for the manufacture of steel sheet for steam boilers by 1764 at the latest. In the same year, John Paynet obtained a patent for the preforming of flat bar by grooved rolls, and John Westwood was awarded a patent for cold rolling in 1783. In 1778, the Frenchman Fleur described a four-stand rolling mill with closed passes which he used for wire production. Proposals for continuous rolling were also published.

Apart from the iron bearings, rollers, cutting rings, and spacers, the rolling mills were constructed of wood. In the mid-1700s, gear wheels and pillars were all of cast iron. Energy was supplied by one or two waterwheels, depending on the size of the rolling mill. The upper and lower rolls were generally driven separately, but pinion stands were also used. The mechanical and thermal stresses on the rolls were large. At first the rolls were made by welding a hardenable steel sheet around the forged roll body. Later, cast iron rolls were used. A further development was the casting of steel around a bar which was used as roll neck.

Rolling mills did not yet replace hammer forges, as their use was limited by the quality of the engineering and the availability of power.

2.3. The Industrial Age

The 1700s saw the development of mechanical power (steam engines), transport (railways, bridges), and the increasing use of machinery, all of which was made possible only by the almost complete replacement of wood by steel as a construction material. The resulting enormous increase in demand provided the impetus for the development of new methods for bulk steel production.

Iron and steel were produced exclusively from charcoal up to the 1700s, but by the end of the century, processes had been developed that enabled charcoal to be replaced by coal. These were coke-based metallurgy and the crucible steel and puddling process, supplemented by the development of the hot rolling process. They formed the basis for the economic mass production of this universal construction material, which satisfied the broad requirements of the approaching industrial era.

During the preindustrial epoch, most of the improvements in steel production methods took place on the European mainland, but the inventions that led to the industrial revolution were made in Britain. These are linked with the names Abraham Darby II (1711 – 1763) who introduced the first true coke-based blast furnace (1735) and the first use of steam engines (1742) to steel production; Benjamin Huntsman (1704 –1776) – crucible steel casting (1742); Henry Cort (1740 – 1800) – development of the puddling process (1783 – 1784); Jean Beaumont Neilson (1792 – 1865) – hot air blast (1829); James Nasmyth (1808 – 1890) – steam hammers (1839), first installed at Creusot (1841); Henry Bessemer (1813 – 1898) – acid air refining process (1856 – 1860); Sidney G. Thomas (1850 – 1885) and Percy C. Gilchrist (1851 – 1935) – basic refining process (1878); Wilhelm Siemens (1823 1892) and Friedrich Siemens (1826 – 1904) – regenerative heating (1861). Important innovators in other countries include Faber du Faur (1786 – 1855) – blast heating (1832); Pierre Emile Martin (1824 – 1915) – open hearth process (1864); Rainer Daelen (1813 – 1887) – universal mill stand (1848); Charles Morgan (1831 – 1911) – rolling mill technology (from 1865); George Bedson – continuous rolling mills (1860); Reinhard Mannesmann (1856 – 1922) and Max Mannesmann (1857 – 1915) – rolling seamless tubes (1890); Paul Hérault (1863 – 1914) – electric arc furnace (1900).

These innovations were gradually introduced into the most important iron-producing countries, including the United Kingdom, Belgium, France, Germany, Austria, and the United States, all of which made their own contributions to

the further development of iron and steel technology. Historically, the last pioneering inventions are the basic oxygen steelmaking process, secondary metallurgy, and continuous casting, which were developed in the 1940s and 1950s.

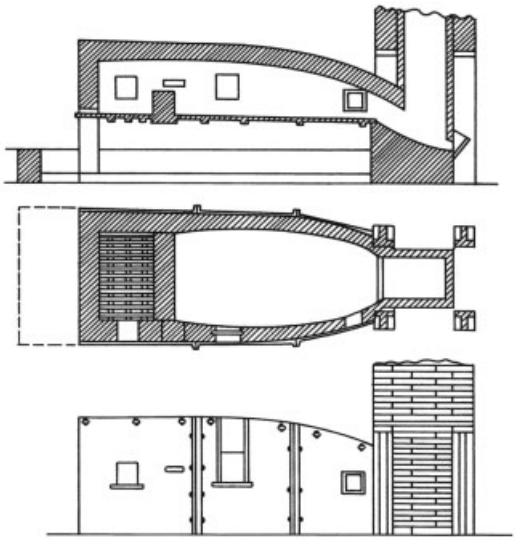

Figure 7. Puddling furnace with iron hearth and cast iron shell (1830) [14]

2.3.1. Pig Iron Production

The cradle of coke-based metallurgy was Coalbrookdale, where Abraham Darby I (1678 – 1717), who had invented the process of casting iron in sand molds in 1707, probably began the production of pig iron using coke in 1709. His son, Abraham Darby II, was the first to operate a blast furnace fueled entirely with coke in 1735. Coke-based metallurgy spread to the rest of the United Kingdom only slowly at first (in 1750, 3 of 74 blast furnaces were coke based; in 1760, 14 of 78; in 1791, 85 of 107). The real breakthrough came with the development of the puddling process for converting pig iron into steel, the production rate matching that of the iron-making process. In mainland Europe, these new technologies were adopted in a dilatory fashion. The first coke-based blast furnaces were operated in France by Creusot in 1785, and in Germany in Gleiwitz (1796). Whereas pig iron production in the U.K. was almost exclusively based on coke in the early 1800s, Germany reached this stage only at the end of the century.

From the early 1800s, coke-based metallurgy was accompanied by research and development, with increases in productivity and profitability. Developments were mainly concerned with the design and size of blast furnaces, including blast and power supplies, increasing the efficiency of energy utilization, and improving yield and quality through better understanding of the metallurgical processes, enabling the control of the blast furnace process to be improved, culminating in computer control.

This development can be demonstrated by data for selected blast furnaces (Table 3).

2.3.2. Steel Production

After the invention of the pig iron furnace, forgeable and hardenable iron (steel) was produced both from almost carbon-free wrought iron and from highly carburized pig iron.

Wrought iron, in the form of soft iron bar, was carburized to steels with a higher carbon content from the early 1600s (originally in 1601 in Nuremberg by Paulus Hannibal) by heating in carbon-containing powders with exclusion of air. This so-called cemented steel had a wide range of uses, e.g., for springs, knives, and files. In the 1700s, the steel industry in Sheffield was founded on the use of this process for the treatment of imported Swedish iron bars. Here, 205 cementation furnaces produced almost 80 000 t steel in 1862. The last of these went out of production in 1951. This cemented steel (blister steel), which was inhomogeneously surface-carburized and blistered, was homogenized by hammering out (sometimes repeatedly) bundles of cemented steel rods to form "shear steel". Although this led to some equalization in material properties, this steel met the increasing quality requirements only to a limited extent.

2.3.2.1. Crucible Steel

The first decisive step toward production of nearly homogeneous and slag-free steel was taken by Benjamin Huntsman in 1740, with the production of liquid steel by melting together cement steel with added bar iron and

Table 3. Development of blast furnaces

Year and location	Hearth area, m^2	Coke consumption per tonne iron, kg	Iron output, t/d
1796 (Gleiwitz)	0.3	3500	1 – 2 (later 4)
1801 – 1815 (U.K.)		2500	5 – 7
1856 (Hasslinghausen)	3.6	1600	20 – 23
1880 (United States)	8.8	1510	120
1901 (United States)	15.3	1000	464
1929 (Bruckhausen)	33.2	740	1100
1993 (Schwelgern 2)	174.9	480 ∗	10 600

∗ Fuel consumption.

glass in refractory crucibles. The productivity of this process was later increased by improvements to the crucible material (graphite). Crucible steel opened up completely new perspectives. The steel was both castable and forgeable. By pouring the contents of several crucibles into one mold, ingots and castings of large mass were produced for the first time. As alloying was possible, steel developed into a universal construction material. As this process was more economical than the production of shear steel, it spread relatively rapidly.

2.3.2.2. Puddled Steel

Pig iron from the small furnaces that later developed into blast furnaces was cast, either directly or after remelting, in a cupola furnace to give finished products (e.g., machine parts, construction elements, household articles) or was decarburized to forgeable steel. The steel produced by decarburizing coke pig iron tended to be brittle (owing to sulfur and phosphorus) compared with that produced from charcoal pig iron. The low productivity of decarburization in open charcoal-fired finery fire became a problem, owing to the increasing rate of production of pig iron; steel producers sought a more effective fining process that would operate with coal or coke. The breakthrough was achieved by Henry Cort with his patented method (1783 – 1784) for producing wrought steel – the puddling process. This consisted of remelting the pig iron and fining with the combustion gases from the coal- or coke-fired reverberatory furnace (Fig. 7), followed by mechanical slag removal and welding together of the lumps of steel by means of a steam hammer or squeezer to form blooms of very variable chemical analysis. These were rolled in a rolling mill to form "crude" flat iron. To produce a homogeneous rolled steel, fagots of flats were placed together, heated in a welding furnace, and rolled. This step could be carried out several times. The process met all requirements of the time, i.e., increase in productivity (tenfold), replacement of charcoal by coal, and reduction of production costs. It spread relatively rapidly into all iron-producing countries. In Germany, the first puddling plant was started in the Rasselstein iron works in 1824. The last puddling operation ceased in the 1940s.

2.3.2.3. Ingot Steel

With time the puddling process also became unable to meet the increasing demand for steel. In 1856, Henry Bessemer invented his autothermic air refining process, which made it possible for the first time in the history of iron production to convert pig iron directly into liquid steel, increasing productivity by a factor of 70 (Fig. 8). In the Bessemer process, molten pig iron is blown with air from below in a tiltable converter. The molten metal is decarburized by oxidation, which also removes silicon and manganese. The heat liberated causes the temperature to rise well above the melting point. As the original lining of the converter was acidic, only low-phosphorus pig iron could be treated, and the use of the process was therefore limited. This problem was solved in 1878 by Sidney G. Thomas and Percy C. Gilchrist by the use of basic dolomite as converter lining. Thus, phosphorus could be removed by means of a basic slag, which later became an important fertilizer (Thomas slag). The Thomas process was widely used, especially on the European mainland, where the ores, e.g., minette (oolitic iron

ore), often have a high phosphorus content. In Germany, the first Thomas steelwork was operated in 1879. The last one was shut down only recently.

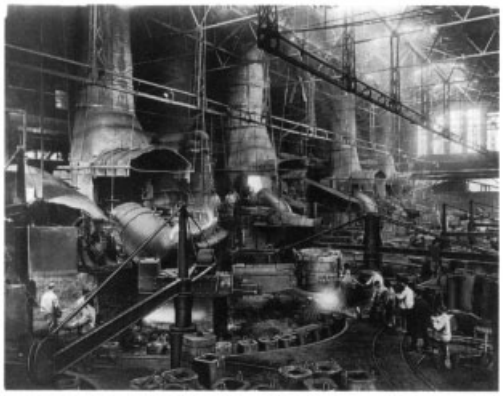

Figure 8. Bessemer steel works with circular casting pit (Friedrich Krupp, Historical Archive, 1970)

Ingot steel production by air refining became the most important steel production process worldwide with the introduction of the basic oxygen (LD, Linz – Donawitz) process in 1952.

In parallel with the air-refining processes, the allothermic hearth process was also developed. In 1864, F. and W. Siemens together with E. and P. Martin succeeded in melting pig iron and steel scrap, using regenerative heating in a reverberatory furnace. The Siemens – Martin process at first achieved great importance for the recovery of scrap. Later, pig iron was melted with iron ore, with or without the addition of steel scrap. Until the breakthrough achieved by the oxygen refining process, this was the most important steel production process, accounting for up to 60 % of all steel production.

After the 1850s, electrical energy became available, and attempts were made to build electric hearth melting furnaces. The electric arc furnace developed by Héroult was especially successful, and was first operated full-scale in Germany in 1906. Although it was initially used for the production of special steels, it is now increasingly used only for steel melting, owing to the availability of cheap electricity, and the introduction of secondary metallurgy.

2.3.3. Rolled Steel Production

2.3.3.1. Puddled Steel and Steam Power

Following the discovery of (1) the puddling process, which enabled iron to be converted to the finished product by a rolling operation; and (2) the introduction of the steam engine, the technology of the rolling mill became an integral part of steel production.

Although this new technology was quickly adopted and further developed in England, this only happened in a few places, owing to the high cost of steam engines. As early as 1784, J. Wilkinson started the first steam-driven rolling mill in his iron works in Bradbury, and replaced the hammer forge by a blooming mill with self-adjusting rolls. This was patented in 1792. There followed improvements in rolling equipment, by making it more powerful (the production of cast machine components of everincreasing size, and the development of techniques for gearing and driving). This can be seen from the dimensions of the rolls, which reached diameters of 300 mm and body widths of 1200 mm.

On the basis of these new methods, England's iron and steel industry occupied a leading position throughout the 1800s. Continental iron producers were able to adopt the new technology only later, sometimes only by industrial espionage or by the enticement of specialist personnel. The major economic incentive was the immense demand for railway track.

The plan view (Fig. 9) of the rolling mills at Couillet (1835) shows the layout of a typical works of that time. The puddling and rolling plants formed a single unit. The steam engines were at the heart of the ironworks. They drove the puddling plant and rail mill, the three-stand sheet mill, the six-stand bar mill, a two-stand mill for rolling – cutting, and an auxiliary plants. This works produced ca. 10 000 t rolled steel in 1840.

2.3.3.2. Ingot Steel and Mass Production

The further development of rolling mills took place against the background of the developing mass production of ingot steel. The first

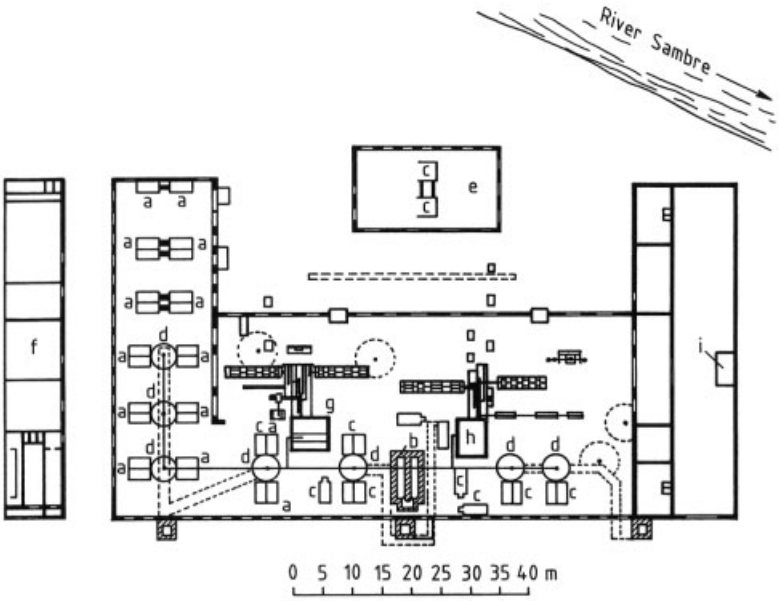

Figure 9. Ground plan of rolling mill at Couillet (1835) [15]
a) Puddling furnace; b) Coal-fired boiler; c) Welding furnaces; d) Waste heat boilers; e) Finishing shop; f) Workshops; g) Machine 1; h) Machine 2; i) Stores

product became the cast steel ingot. Developments took place particularly in drive technology and the construction of specialized rolling stands and mills (and also auxiliary machinery and reheating furnaces which are not considered in this article). The realization of reversible operation of rolling stands driven by steam engines was the trigger for the further improvement of the driving machinery. In the 1870s, hydraulic friction coupling for reversible operation was successfully designed and constructed. However, almost 30 years of development were required before the problem of reversing the steam-driven rolling mill engine was solved by Clemens Kisselbach in 1891. The use of electrical driving machinery, which began in 1900, increased, especially after the invention of the reversible electrically driven rolling mill engine by Ilgner in 1902.

The design and construction of rolling mills and plant was driven by the demand for mass production, with increasing throughput, size, and range of products. The American, John Fritz, in 1857, eliminated the need for reversible rolls by designing and successfully operating a three-high rolling stand for rail production. In 1848, Rainer Daelen built the first universal rolling stand, which found wide application, especially in the United States. Using this equipment, finishing of I-beams was first achieved in 1866. In 1883, I. J. Seaman and H. Sack separately applied for patents for universal rolling mills for rolling beams. Wire and rod steel products were rolled in open rolling trains. The introduction of mechanical loopers in 1877 led to a considerable increase in performance. At the end of the 1860s, G. Bedson constructed the first continuous finishing train for wire, with 16 alternately horizontal and vertical rolling stands. This type of train was further developed, mainly by C. Morgan in the United States. In 1905, in Gary, Indiana, the first train operating completely continuously from ingot to finished product was commissioned. From the early 1800s, a large number of proposals for the production of seamless tubes by rolling were made. However, the breakthrough was an invention by M. and R. Mannesmann, in which a hollow body was produced by skew rolling (1886), with subsequent expansion to form a seamless thin-walled tube in the Pilger step-by-step-type seamless tube rolling mill (1890). The development of flat rolling aimed to increase the dimensions of strip and sheet, the weight of charge,

and productivity. The United States led the way, and the first continuous hot strip mill went into operation during the 1920s. These mills constitute as important a development for modern steel production as the oxygen steelmaking process, secondary metallurgy, and continuous casting.

3. Crude Steel Production

3.1. Raw Materials

The following iron-bearing materials are used for crude steel production (Table 4) [16]:

1) Molten iron from blast furnaces (hot metal), mainly used for steel production by the basic oxygen furnace (BOF) process
2) Scrap, sponge iron, and solidified blast-furnace iron (pig iron), mainly used for melting steel in electric furnaces

Another important raw material is lime for slag formation. Alloying elements, such as chromium, nickel, molybdenum, titanium, vanadium, and niobium, and elements used as deoxidizing agents (e.g., manganese, silicon, and aluminum) are not discussed in detail here.

The most important iron-bearing material is hot metal. Of the total world crude steel production of 770 – 720 t/a in the early 1990s, ca. 60 % was produced from hot metal obtained by reducing iron ore, and ca. 40 % from steel scrap.

Steel scrap is of increasing importance as a raw material because of the lower primary energy consumption and lower CO_2 emissions, compared with the blast furnace – converter production route. Figure 10 compares the total specific primary energy consumption per tonne of steel produced from ore with that for steel produced from scrap [17].

3.1.1. Hot Metal

Hot metal is produced by reduction of iron ore in a blast furnace. Modern furnaces produce 5000 – 12 000 t/d. Immediate transfer of the hot metal to the converter for refining to crude steel requires the blast furnace and the steelworks to be in close proximity.

Iron ore is charged to the blast furnace in lump form or as agglomerated fine ore (sintered or pelletized), and reduced. Coke and injected coal are used as reducing agents, and these react with the hot air blast, liberating energy. The molten product, known as hot metal, separates from the liquid slag. It can be pretreated before transfer to the steelworks, to reduce the content of undesired accompanying elements such as sulfur, silicon, and phosphorus.

Various types of hot metal are produced, depending on the composition of the ore fed to the blast furnace (burden) and the method of furnace operation. They include low-phosphorus hot metal, (Thomas) hot metal (high in phosphorus), and foundry hot metal (high in silicon). The chemical compositions of several grades of hot metal are listed in Table 5.

Low-phosphorus hot metal is important for steelmaking. Typical temperatures for tapping the blast furnace are 1350 – 1450 °C. Carbon content is typically high (4 – 4.5 %), as is sulfur content, these elements being picked up from the coke and the injected coal. A pretreatment of the hot metal, e.g., with lime – magnesium mixtures, reduces the sulfur content from ca. 0.040 % to < 0.010 %.

Pig iron is also used to a limited extent as an iron-bearing material for steel production involving melting; the hot metal, after tapping from the blast furnace, is cooled in sand molds to form pigs which can be used like scrap in steel production.

Hot metal is produced not only by the blast furnace method, but also in coke-free processes in smaller units by smelting reduction, e.g., the Corex process, which produces 1000 – 2000 t/d [18].

3.1.2. Scrap

Steel scrap is traded worldwide [19]. It is mainly melted in electric furnaces, but to a limited extent is also added as a cooling agent to converters used in steel production by blowing processes.

In 1992, ca. 350×10^6 t steel scrap was used worldwide for steel production, the specific scrap consumption being ca. 445 kg per tonne of crude steel. Specific scrap consumption differs widely between countries, depending on the structure of the steel industry [20], [21]. In Italy, where the proportion of electric steel production

Table 4. Properties of hot metal, sponge iron, and scrap [16]

	Hot metal	Sponge iron	Scrap
Raw materials	poor to rich ores, low or high in phosphorus	ores with low gangue content	internal scrap production scrap collected scrap
Availability	almost unlimited	limited to low gangue ores or pellets	depends on steel consumption and crude steel production
Production and processing	classification, sintering, pelletizing blast furnace	classification or pelletizing shaft furnace rotary furnace retort fluidized bed	shears (cutting – burning)
Primary energy consumption (GJ/t)	14	12 (shaft furnace)	0.6 (shredder)
State of aggregation	liquid	solid	solid
Storage and transport	severely limited	special techniques may be necessary (e.g., briquetting)	good
Type of furnace using this material	oxygen-blowing furnace (BOF) (electric furnace)	electric furnace (as cooling agent in oxygen furnace)	electric furnace Siemens – Martin furnace (as cooling agent in oxygen furnace)
Price	depends on production costs	depends on production costs, market price, or agreed price	market price
Chemical properties			
Composition	constant	constant	variable
Unwanted elements	sulfur phosphorus (depends on application)	none phosphorus (depends on application)	depends on process
Type of gangue		depends on ore	
Physical properties	liquid	bulk material	variable shredded scrap (in bulk)
Bulk density		$> 1.2\ t/m^3$	fluctuates around $1.2\ t/m^3$ depending on type of scrap
Thermal conductivity		low	high

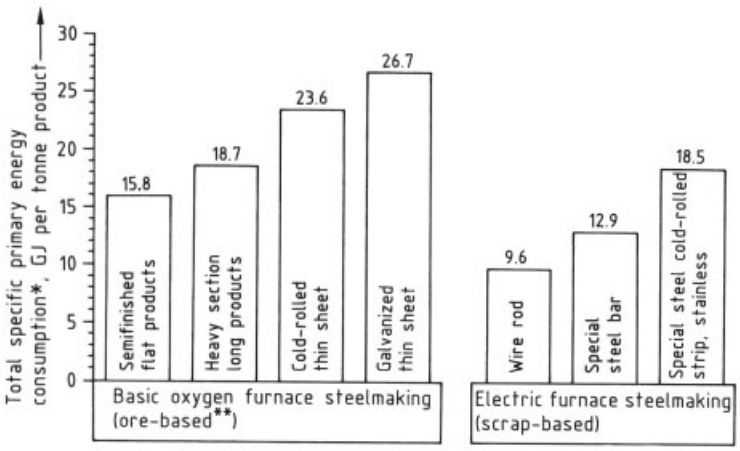

Figure 10. Comparison of total specific primary energy consumption for steel products made from ore and scrap [17]
∗ Including primary energy consumption for generating electricity in thermal power stations, and energy-intensive materials (oxygen, steam, compressed air, cooling water, etc.). ∗∗ Including added scrap for cooling.

Table 5. Characteristic compositions of grades

Type of pig iron	C, %	Si, %	Mn, %	P, %	S, %
Low-phosphorus	4.0 – 4.5	0.30 – 0.70	0.20 – 0.70	0.05 – 0.12	0.03 – 0.06
High-phosphorus	3.2 – 4.0	0.30 – 0.70	0.25 – 1.20	1.50 – 2.20	0.03 – 0.06
Foundry	3.5 – 4.2	1.80 – 2.50	0.70 – 1.00	0.50 – 0.70	0.03 – 0.06

is 50 %, the proportion of scrap used, ca. 650 kg per tonne of crude steel, is naturally larger than in Germany (ca. 380 kg/t), which has 21 % electric steel production (Fig. 11).

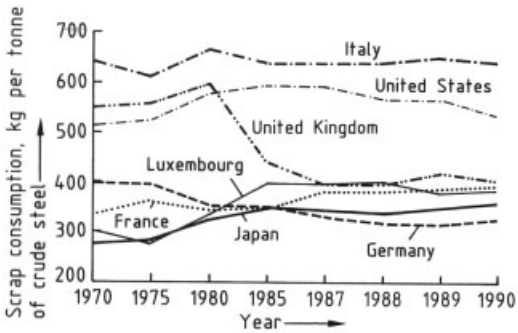

Figure 11. Average specific scrap consumption for crude steel production in various countries [21]

The total quantity of scrap handled on world markets is ca. 50×10^6 t/a. A survey of the most important exporting and importing countries is given in Table 6. The price of scrap fluctuates widely, depending on supply and demand. Figure 12 gives an example of the fluctuations in the price of a particular grade of scrap in the Ruhr region.

Table 6. Balance of exports and imports for several countries in 1991 [21]

	Export balance, 10^6 t	Import balance, 10^6 t
United States	8.28	
Germany	7.63 ∗	
United Kingdom	3.12	
France	2.31	
Japan	1.48 ∗	
European Community	1.96	
Taiwan		2.19
South Korea		3.49
Spain		4.27
Turkey		4.44
Italy		6.03
Developing countries		10.30

∗ Data for 1992.

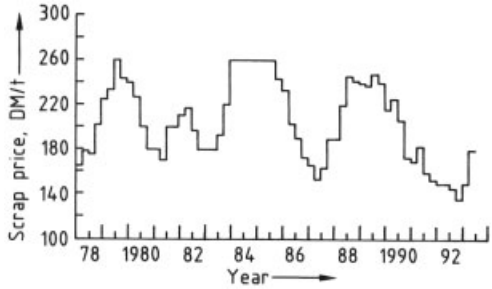

Figure 12. Variations in scrap prices since 1978 [20]

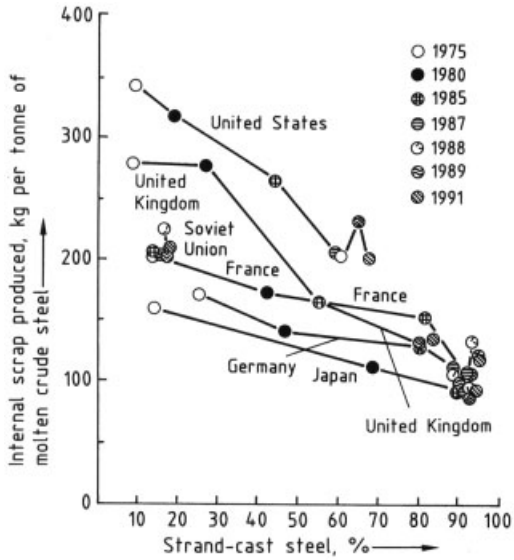

Figure 13. Variations in the specific amount of internal scrap produced in various industrialized countries as a function of the percentage of strand casting [22]

Scrap is classified according to its origin: internally recycled scrap; production scrap from the steel processing industry, also known as new scrap; and collected scrap, also known as old scrap.

Internally Recycled Scrap. Improvements in the utilization of materials and technologi-

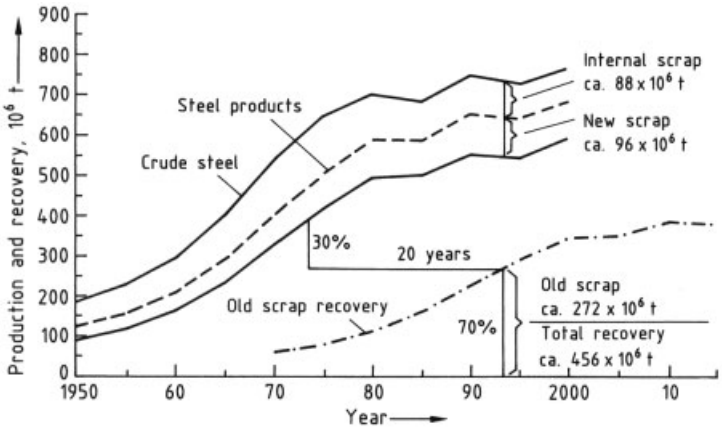

Figure 14. World crude steel production, with a forecast of old scrap availability up to 2015 [20]

cal changes, e.g., the introduction of the strand casting process, have steadily reduced the availability of internally recycled scrap (Fig. 13) [22].

New and Old Scrap. To meet the requirements of steelworks, internally recycled scrap is supplemented by the purchase of new and old scrap. To maintain steel production quality, the quality of these purchased materials is controlled by lists of scrap grades. Criteria include the metallic iron content (e.g., $> 92\,\%$), the lump size, the bulk density (e.g., $> 0.9\,\mathrm{t/m^3}$), and the density of bundles (e.g., $> 1.3\,\mathrm{t/m^3}$).

Current proportions are ca. 50 % from old scrap, 15 – 20 % from new scrap, and 30 – 35 % from internally recycled scrap.

The quantity of scrap available worldwide depends on the quantity of crude steel produced, although there is a time delay. Whereas internally recycled and new scrap are immediately returned to the system, an average of ca. 70 % of used steel products are returned as old scrap after ca. 20 years [19], [20]. On the basis of this assumption, H. W. Kreutzer has shown how the total world supply of old scrap has changed with time (Fig. 14). In line with steel production figures which increased steadily in former years, the availability of old scrap at first continues to increase, but then levels out after ca. 20 years.

It is important that scrap should not introduce unwanted elements into the steel, thereby impairing its quality. Pretreatment of the scrap by size reduction and sorting, e.g., with the aid of a shredder, is therefore very important. It is essential not to introduce elements that cannot be removed from the crude steel, or whose removal is costly, e.g., copper, nickel, molybdenum, or tin. In some steel products, a copper content of 0.40 % is tolerable, but in high-quality products, e.g., thin steel strip, the copper content is normally $< 0.01\,\%$. Therefore, when recycling scrap automobiles, copper- containing components such as electric motors and cables must be stripped out before scrapping. On the other hand, when zinc- coated car bodies are recycled, the zinc becomes concentrated in the steelworks dust, and can be easily recycled, if the levels of zinc in the dust are high enough (ca. $> 20\,\%$). So far, this has been achieved in electric steelworks owing to the high input of scrap [23]. The zinc is recovered from the dust in nonferrous metal smelting works.

Table 7. Residuals in steel and its raw materials [24]

	Cu + Sn + Ni + Cr + Mo, %
Typical maximum residual limits of carbon steels	
Tinplate for draw and iron cans	0.12
Extra deep drawing quality sheet	0.14
Drawing quality and enameling steels	0.16
Commercial quality sheet	0.22
Fine wire grades	0.25
Special bar quality	0.35
Merchant bar quality	0.50
Typical residual content of charge materials	
Direct reduced iron	0.02
Pig iron	0.06
No. 1 factory bundles	0.13
Bushelling	0.13
No. 1 heavy melting	0.20
Shredded auto	0.51
No. 2 heavy melting	0.73

Table 7 shows the maximum tolerable total content of Cu, Sn, Ni, Cr, and Mo for a number of steel products, and some typical concentrations of these elements in the raw materials pig iron, sponge iron, and five different types of scrap.

3.1.3. Sponge Iron

Another solid raw material for the steel industry is sponge iron, also known as DRI (direct reduced iron). Because of its purity and suitability for controlled feeding, it is becoming of increasing importance.

Sponge iron is produced by reduction of iron ore in the solid phase, by means of (1) reduction gas (obtained from natural gas with varying hydrogen and carbon monoxide content); or (2) coal. Of the direct reduction processes listed in Table 8, the Midrex and HyL processes are the most important. Gas reduction processes account for 92 % of world sponge iron production. Direct reduction processes are operated where cheap energy is available (natural gas, coal), e.g., Mexico, Venezuela, Saudi Arabia, Iran, Malaysia, Indonesia, India, and South Africa.

Table 8. Processes for the direct reduction of iron ores

Reduction by gases	Reduction by solids
Shaft furnace:	Rotary kiln:
Midrex	SL/RN
HyL III	Krupp – Codir
Fluidized bed:	Rotary hearth furnace:
FIOR	Inmetco
Iron carbide	Fastmet

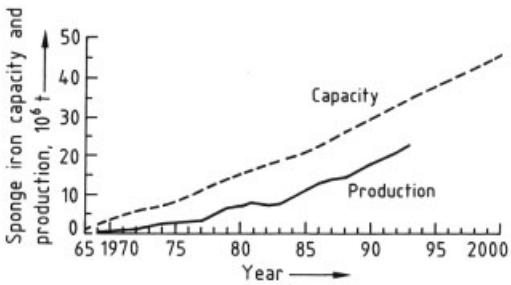

Figure 15. Growth of production capacity and production of sponge iron [25]

Figure 15 shows the growth of the capacity and production rate of sponge iron [25]. In 1993, total production was 23.9×10^6 t while capacity was ca. 35×10^6 t, i.e., production plants were only 68 % utilized. A further increase in world sponge iron capacity of at least 12×10^6 t/a is forecast for the year 2000.

The sponge iron produced is mainly melted in adjoining steelworks, so that amounts available on the world market are limited. Approximately 3.4×10^6 t sponge iron was exported in 1992 [26]. However, there is an increasing trend toward the construction of "merchant plants", e.g., in Malaysia and Venezuela.

These types of sponge iron could be of great interest for reducing levels of unwanted elements by replacing some of the scrap used in the production of scrap-based, high-grade steel products, especially flat products produced in minimills.

Table 9. Chemical composition of sponge iron [25]

	Midrex	SL/RN
Fe, %	91 – 93	90.4 – 93.7
Metallic Fe/total Fe, %	92 – 95	92 – 93
SiO_2, %	2.0 – 3.5	2.4 – 5.8
Al_2O_3, %	0.5 – 1.5	1.8
CaO, %	0.2 – 1.6	0.05 – 0.3
MgO, %	0.3 – 1.1	≤ 0.05
C, %	1.0 – 2.5	0.1 – 0.2

Because trade in these materials has so far been small, no lists of grades or general regulations exist for sponge iron supplies, such as apply to scrap. Important characteristics include the total iron content, the ratio of metallic iron to total iron, the gangue content, and the carbon content. As all the gangue constituents report to the sponge iron, ores low in gangue, or specially produced pellets, are used. Otherwise, additional energy must be used to melt the gangue when melting sponge iron. The metallic iron/total iron ratio should, therefore, preferably be > 92 %. The carbon content of sponge iron from the gas reduction process is 1 – 2.5 %, which is higher than the typical value for sponge iron produced by the coal reduction process (0.1 – 0.2 %). A high carbon content can help to reduce the consumption of electrical energy for melting if the carbon is burned with air or oxygen to form carbon monoxide, which is post- combusted in the melting plant. New developments in direct reduction involve the production of iron carbide, which should give a carbon content of > 5 % of the DRI [27].

Table 9 gives typical figures for the composition of a Midrex sponge iron (gas reduction in a shaft furnace) and an SL/RN sponge iron (solid state reduction with coal in a rotary kiln). Table 10 gives typical figures for bulk density, lump density, and porosity of sponge iron produced from lump or pelletized ore, and for briquetted sponge iron [16].

Table 10. Properties of sponge iron [16]

Form	Bulk density, t/m^3	Lump density, g/cm^3	Porosity, %
Lumps	1.9	2 – 3.5	70 – 50
Pellets	1.7		
Briquettes	2.5	4.5 – 6	35 – 15

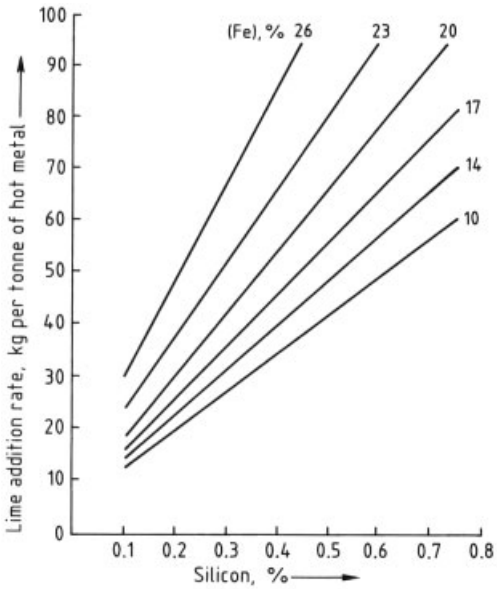

Figure 16. Required rate of addition of lime as a function of the silicon content of pig iron and iron content of slag [29]

When transporting and storing porous sponge iron, attention must be paid to its tendency to reoxidize exothermically, owing to the large surface: volume ratio. The heat given off can lead to ignition of the sponge iron and to a fire, e.g., during transport. This reoxidation behavior, typical of many types of sponge iron, can be characterized by determining the ignition temperature, the rusting behavior, and the evolution of hydrogen in a moist atmosphere. Safety precautions must always be taken for the transport, loading, unloading, and storage of sponge iron. Regulations and other information relevant to transport by sea have been published by insurance companies. The tendency to reoxidation can be reduced by passivation of the sponge iron. The best method of limiting transport hazards is briquetting.

3.1.4. Lime

Lime is used as a reagent and slag former in steel production. It is the main constituent of the most metallurgically effective slags that combine with the unwanted elements in the steel.

Lime (CaO) is formed on calcination of limestone ($CaCO_3$) at ca. 1000 °C. Depending on the calcination temperature, calcination time, dwell time in the furnace, and type of gas flow through the furnace, various qualities of CaO can be produced, e.g., soft- or hard-burned lime [28]. The quality requirements of the steel industry include high purity (94 – 97 % CaO), well-defined chemical composition, optimum grain size, and rapid dissolution in the slag. Calcined (soft-burned) dolomitic limestone is also used in steel production.

In steel production by top blowing with oxygen, the hot metal composition is of the greatest significance when determining the quantity of lime to be added for refining in the converter. The amount required to produce a slag that is slightly supersaturated with respect to lime is illustrated in Figure 16, which also shows the iron content of the slag [29]. As well as decreasing the silicon and sulfur content of hot metal, the changes brought about by the introduction of the combined blowing process in the mid-1980s had a further effect in reducing lime content [30], [31]. By bottom blowing with an inert gas to give agitation during oxygen blowing, the reactions approach equilibrium much more closely, and hence a better utilization of the slag is achieved. Figure 17 illustrates this trend toward a reduction in specific lime consumption, using German BOF steelmaking as an example. Between 1970 and 1990, specific lime addition was reduced from ca. 70 to ca. 45 kg per tonne of crude steel.

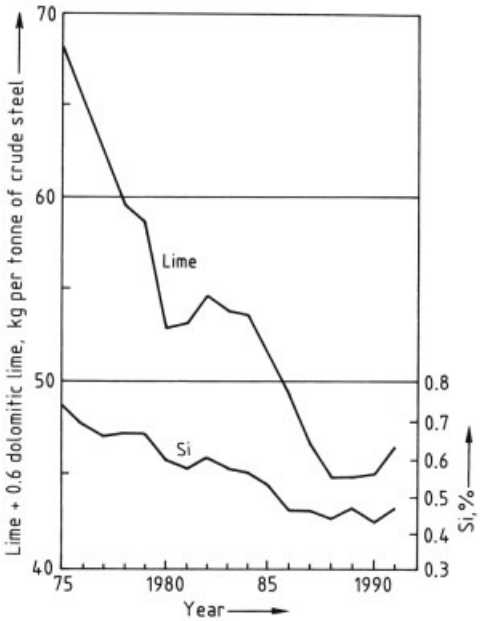

Figure 17. Variation in the specific consumption of lime and the silicon content of low-phosphorus pig iron in an oxygen steelmaking plant in the Federal Republic of Germany (1975 – 1990)

In electric steel production with scrap addition, the process in the 1970s was mainly to add an oxidizing slag, and then to change to a refining slag. Today, both low- and medium-alloyed steels are produced under one slag. The oxidizing treatment is followed by a reduction phase by the slag. All the other metallurgical operations have been moved to the later processing stages of the secondary metallurgy. Specific lime consumption in steel production in the electric arc furnace is 35 – 40 kg/t. Soft-burned lime usually has lump size 10 – 30 mm, and should have a sulfur content < 0.02 %. With sponge iron instead of scrap, the amount of slag increases, because of the silica introduced with the gangue materials in the sponge iron. The amount of lime added depends on the amount of SiO_2 that is to be reacted. The slag basicity (CaO/SiO_2) should be 1.5 – 2.

Modern developments in secondary metallurgy have led to new process steps in the operations that take place between the steelworks and the strand casting plant. Slags, which have become reactive contact phases, are now at the heart of metallurgical research and development. As there are many aims of secondary metallurgy, the slags used can differ widely. Lime-containing slags mostly consist of reactive $CaO - CaF_2 - Al_2O_3$ mixtures for steel treatment in the ladle furnace, for intensive blowing in the ladle, or as a basic covering slag. The compositions of desulfurizing slags are ca. 55 – 60 % CaO, 5 – 10 % CaF_2 and Al_2O_3. Typical lime addition rates are, therefore, 10 – 15 kg/t. As well as the intensive stirring process with synthetic slag, the injection of $CaO - CaF_2$ mixtures at the rate of 3 – 5 kg/t is common.

Metallurgical lime must fulfill special requirements for utilization of waste slag from the steelworks [32]. Steelworks slags, when used as hydratable components of building materials, should contain only such levels of lime and free magnesium oxide that volume stability can be absolutely guaranteed. For this reason, the relevant specifications limit the free lime content to < 7 wt % or < 4 wt %, according to the intended use (volume of intergranular free space).

The interests of the steelworks and the slag users coincide with respect to lime [32]. The steelworks require a readily soluble lime for metallurgical reasons. The slag user has problems with volume stability if the solid slag contains lime that failed to dissolve when it was liquid. Therefore, reactive, soft-burned lime should be used for slag utilization reasons.

3.2. Physical and Chemical Fundamentals

3.2.1. Thermodynamics

3.2.1.1. Metallic Systems Based on Iron

In steelmaking, the most common accompanying elements in molten iron are C, Si, Mn, P, S, and the gaseous elements H and N. In the reactions that take place when oxygen is brought into contact with the molten materials, the concentrations of the unwanted elements in the iron are reduced to the levels required in the steel produced. The thermodynamic data for metallic systems based on iron, such as solubility and activity, are of fundamental importance for the control of these reactions.

The state of carbon in Fe – C alloys can be seen from the phase diagram (Fig. 18) [33]. Carbon dissolves in both molten and solid iron. Its

solubility in the latter varies according to the crystal structure of the iron. Although carbon dissolves to only a very small extent in α-Fe and δ-Fe, its solubility in γ-Fe at 1153 °C is 2.1 %. The solid solution of γ-Fe – C is known as austenite. The solubility of carbon in molten iron is given by the equation [34], [35]:

$$[\%\mathrm{C}]_{\max} = 1.30 \times 2.57 \times 10^{-3} t$$
$$t = 1152 - 2000\,^{\circ}\mathrm{C} \tag{3.1}$$

The solubility of carbon in molten iron is also influenced by other elements present (Fig. 19) [34].

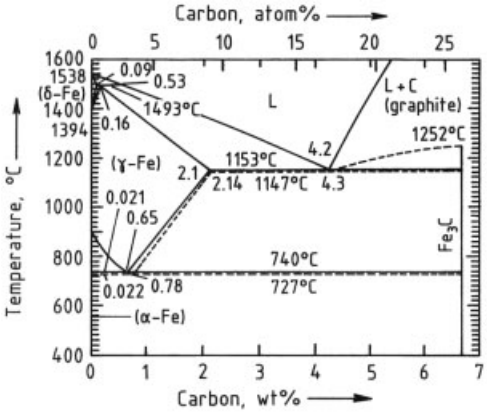

Figure 18. Phase diagram for Fe – C alloys

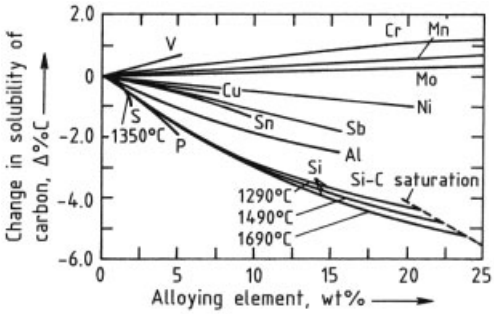

Figure 19. Effect of alloying elements on the solubility of carbon in molten iron alloys [34]

The effect of other elements on the activity of the carbon in molten iron can be represented by the interaction parameters $e_{\mathrm{C}}^{\mathrm{C}}$ and $e_{\mathrm{C}}^{\mathrm{X}}$. The value of $e_{\mathrm{C}}^{\mathrm{C}}$ at 1600 °C is 0.14 [36]. The activity of the carbon is only slightly dependent on temperature. The effect of other elements on the activity coefficient of carbon is shown in Figure 20 [36].

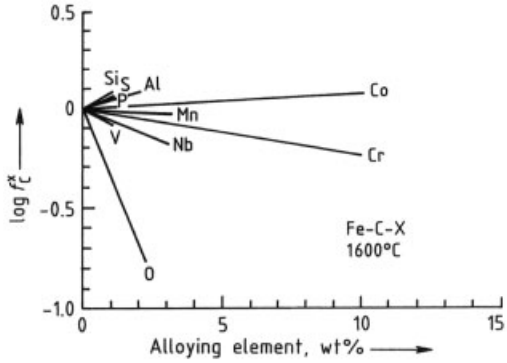

Figure 20. Effect of alloying elements on the activity coefficient of carbon in molten iron alloys [36]

The Fe – O phase diagram is shown in Figure 21 [37]. The liquid region, in which molten iron is in equilibrium with molten iron(II) oxide, is relevant to the steelmaking process. The solubility of oxygen in molten iron is limited by the precipitation of molten FeO in the reaction:

$$\mathrm{Fe\,(l) + O \longrightarrow FeO\,(l)}$$

The solubility of oxygen varies with temperature according to the equation [38]:

$$\log\,[\%\mathrm{O}]_{\max} = -\frac{6320}{T} + 2.734 \tag{3.2}$$

in the range 1530 – 1700 °C. Equation (3.2) corresponds to the line BB′ in Figure 21.

There have been many experimental determinations of oxygen activities in the system Fe – O and of the effect of othe elements X in the system Fe – O – X. The temperature dependence of $e_{\mathrm{O}}^{\mathrm{O}}$ is [36]:

$$e_{\mathrm{O}}^{\mathrm{O}} = -\frac{1750}{T} + 0.794 \tag{3.3}$$

The effect of various alloying elements and other additives on the activity coefficient $f_{\mathrm{O}}^{\mathrm{X}}$ at 1600 °C is shown in Figure 22 [39].

Hydrogen and nitrogen can have adverse effects on the properties of solid and liquid metals. During the process of steelmaking, hydrogen

from atmospheric moisture usually enters the steel. The most important source of nitrogen during steelmaking is atmospheric air. Nitrogen is also absorbed during oxygen blowing, the N_2 content of the O_2 used being ca. 3 %.

The solubility of H and N in iron obeys the Sievert square root law:

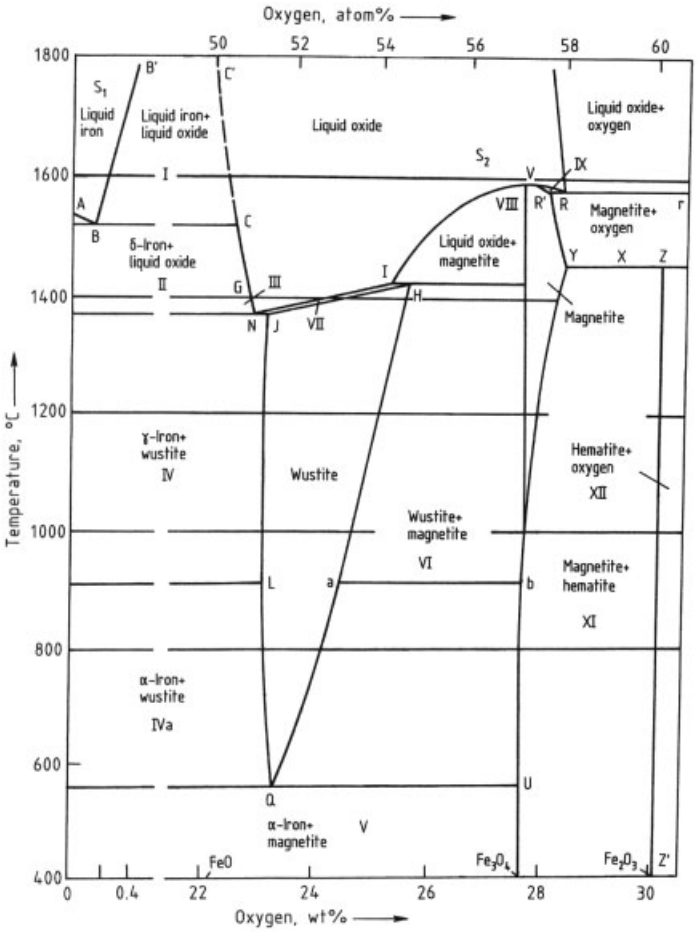

	T, °C	O_2, wt%	P_{CO_2}/P_{CO}
A	1536		
B	1528	0.16	0.209
C	1528	22.60	0.209
G	1400[a]	22.84	0.263
H	1424	25.60	16.2
I	1424	25.31	16.2
J	1371	23.16	0.281
L	911[a]	23.10	0.447
N	1371	22.91	0.282
Q	560	23.26	1.05
R	1583	28.30[b]	
R′	1583	28.07[b]	
S	1424	27.64	16.2
V	1597	27.64[c]	
Y	1457	28.36[b]	
Z	1457	30.04[b]	
Z′		30.06	

[a] Value for pure iron.
[b] $P_{O_2} = 0.1$ MPa. [c] 5.75 kPa.

Figure 21. Phase diagram for Fe – O [37]

$$[\%\mathrm{H}] = K_{\mathrm{H}} \sqrt{p(\mathrm{H_2})} \tag{3.4}$$

$$[\%\mathrm{N}] = K_{\mathrm{N}} \sqrt{p(\mathrm{N_2})} \tag{3.5}$$

where p represents partial pressure.

The effect of temperature on solubility is reflected in variations of the equilibrium constants K_{H} and K_{N}. The solubility of hydrogen and nitrogen in pure iron at partial pressure 0.1 MPa (1 bar) is shown as a function of temperature in Figures 23 and 24 [40]. For hydrogen, the solution process is endothermic over the whole temperature range 400 – 1800 °C. Nitrogen dissolves exothermically in the γ-phase.

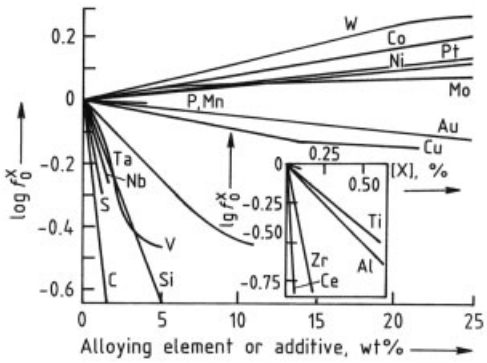

Figure 22. Effects of alloying elements and other additives on the activity coefficient of oxygen in molten iron alloys at 1600 °C [39]

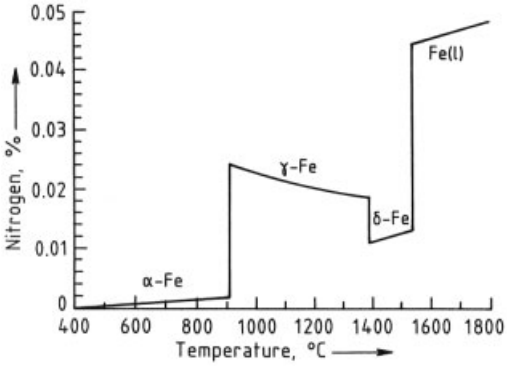

Figure 23. Effect of temperature on the solubility of hydrogen in pure iron at 0.1 MPa [40]

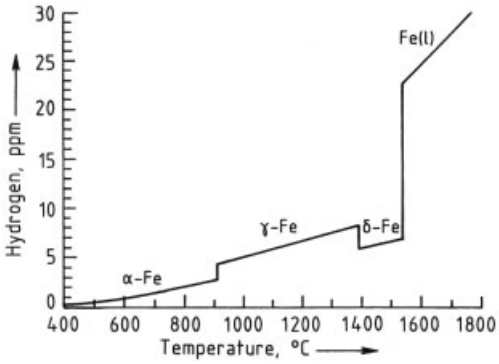

Figure 24. Effect of temperature on the solubility of nitrogen in pure iron at 0.1 MPa [40]

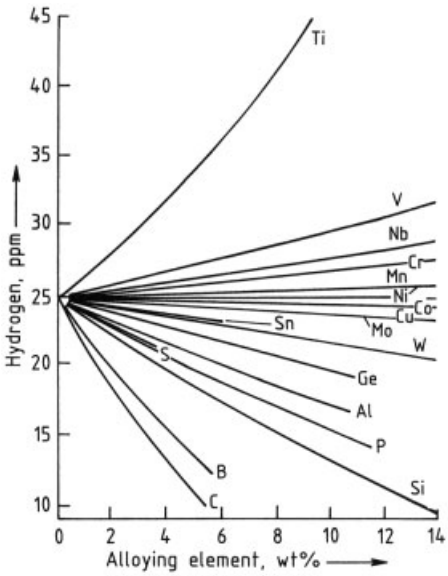

Figure 25. Effect of alloying elements on the solubility of hydrogen in iron alloys at ca. 1600 °C and $p(H_2) = 0.1$ MPa [41]

The solubility of hydrogen and nitrogen is changed by the presence of a third element, mainly by affecting the activity coefficient of the dissolved gas (Figs. 25, 26, 27, 28) [41–43].

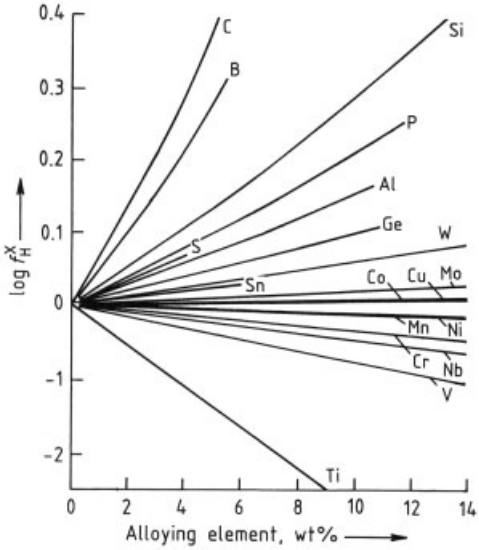

Figure 26. Effect of alloying elements on the activity coefficient of hydrogen in molten iron alloys at ca. 1600 °C and $p(H_2) = 0.1$ MPa [41]

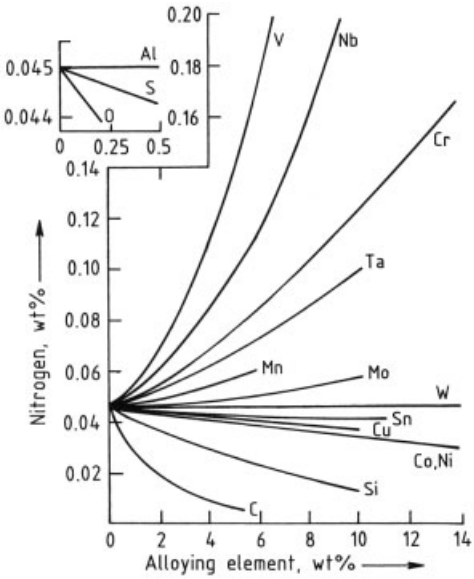

Figure 27. Effect of alloying elements on the solubility of nitrogen in molten iron alloys at ca. 1600 °C and $p(N_2) = 0.1$ MPa [42]

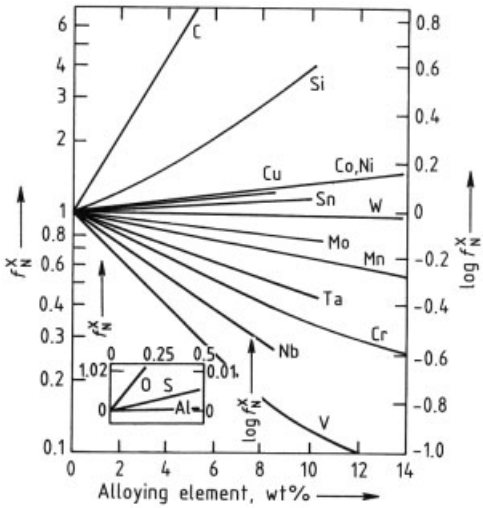

Figure 28. Effect of alloying elements on the activity coefficient of nitrogen in molten iron alloys at ca. 1600 °C and $p(N_2) = 0.1$ MPa [43]

3.2.1.2. Slag Systems

In the steelmaking process, slags are formed as products of the refining reaction and by the addition of materials such as lime. It is extremely important that the slag forms quickly, enabling the important reactions between metal and slag, e.g., desulfurization and dephosphorization, to proceed until the required low levels of sulfur and phosphorus are obtained in the finished steel.

Calcium silicate slags are important in the metallurgy of steelmaking when the raw materials have a low phosphorus content. The essential system is $FeO - CaO - SiO_2$ (Fig. 29) [44]. In general, the total of these three components is $> 80\,\%$, the remainder consisting mainly of MnO, with smaller amounts of P_2O_5, MgO, Al_2O_3, and Cr_2O_3. Figure 29 shows the solid phases in equilibrium with the melt at various surfaces of the precipitate. Continuous lines separate the saturation surfaces of individual solid phases from each other, at which more than one phase is precipitated on cooling. The isotherms are shown as broken lines. For steelworks slags, the most important region includes the stability zones of the liquid slags, dicalcium silicate, tricalcium silicate, and lime. An important isothermal diagram for steel manufacture is that at 1600 °C (Fig. 30) [45], which shows lines of equal FeO activity for the equilibrium with metallic iron. In the area a – b – c – d – e – f, the slags are liquid at 1600 °C. All other areas consist of more than two phases. The activity of FeO gives a measure of the tendency of the slag to oxidize, and from it the corresponding oxygen content of the melt can be deduced. The FeO activity is also important in desulfurization and dephosphorization.

3.2.1.3. Metal – Slag Reactions

3.2.1.3.1. Refining Reactions

In the refining treatment of molten pig iron to make steel, carbon is removed along with other elements, e.g., Si and Mn, by blowing the molten metal with oxygen. The thermodynamics of the reactions are independent of the process used; differences between the various processes affect only the kinetics of the individual reactions. Typical concentration changes in the steel melt during oxygen blowing in a basic oxygen converter are shown in Figure 31 [46]. The main aim is to reduce the carbon content to the desired value in the shortest possible time. However, the time must be long enough to enable the slag to form, the desired tapping temperature to be achieved, and phosphorus and sulfur to be removed from the system until the desired levels are reached.

Decarburization is achieved by the reaction:

$$C + O \rightleftharpoons CO \tag{3.6}$$

The CO bubbles promote homogenization of the melt, and the elements H and N dissolved in the melt are picked up in the bubbles as they ascend. The equilibrium constant of the decarburization reaction [47] is given by:

$$\log K = \frac{1160}{T} + 2.003 \tag{3.7}$$

In practical decarburization processes, a simplified expression for the equilibrium between C and O is used

$$[\,\%C]\,[\,\%O] = 0.0025\,p\,(CO) \tag{3.8}$$

at 1600 °C.

Silicon is rapidly removed during the oxidation process, owing to its high affinity for oxygen, and combines with the lime in the slag; oxidation takes place by the reaction:

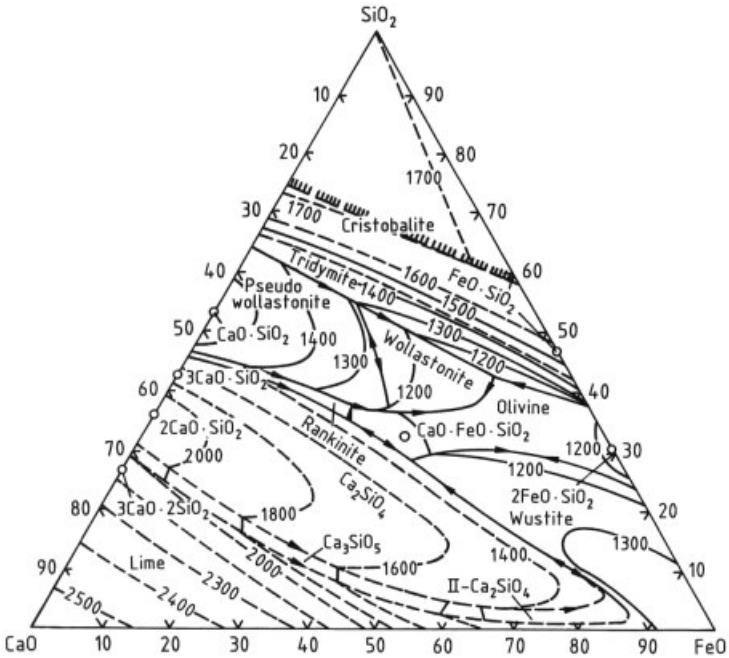

Figure 29. Phase diagram for CaO – FeO – SiO_2 [44]

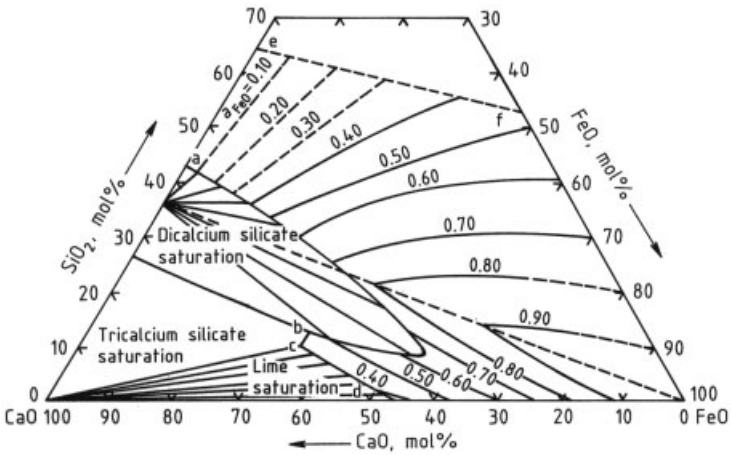

Figure 30. Phase diagram for CaO – FeO – SiO_2 with lines of equal activity for FeO at 1600 °C [45]

$$Si + 2\,O \rightleftharpoons SiO_2 \tag{3.9}$$

The equilibrium constant for this reaction is:

$$\log K = \log \frac{a_{SiO_2}}{a_{Si}\,a_O^2} = \frac{30110}{T} - 11.4 \tag{3.10}$$

The exothermic oxidation of silicon provides a large proportion of the process heat. The oxidation product SiO_2, together with FeO, forms the first slag. This reduces the melting point of the added lime to such an extent that a reactive liquid slag can be formed. Therefore, the amount of CaO added and the amount of slag formed are mainly determined by the Si content.

Manganese is removed at the beginning of the oxidation process, mostly in parallel with silicon. Oxidation of manganese can be represented by the equation:

$$[Mn] + [O] \rightleftharpoons (MnO) \tag{3.11}$$

The equilibrium constant is given by

$$\log K = \log \frac{a_{\mathrm{MnO}}}{a_{\mathrm{O}} a_{\mathrm{Mn}}} = \frac{12590}{T} - 5.53 \qquad (3.12)$$

Toward the end of the silicon oxidation, the increased melt temperature leads to the reduction by carbon of the manganese oxide in the slag, according to the equation:

$$(\mathrm{MnO}) + [\mathrm{C}] \rightleftharpoons [\mathrm{Mn}] + \mathrm{CO} \qquad (3.13)$$

However, at the end of the blow, the manganese content decreases continuously, owing to oxidation of the manganese, mainly by FeO in the slag:

$$(\mathrm{FeO}) + [\mathrm{Mn}] \rightleftharpoons (\mathrm{MnO}) + \mathrm{Fe(l)} \qquad (3.14)$$

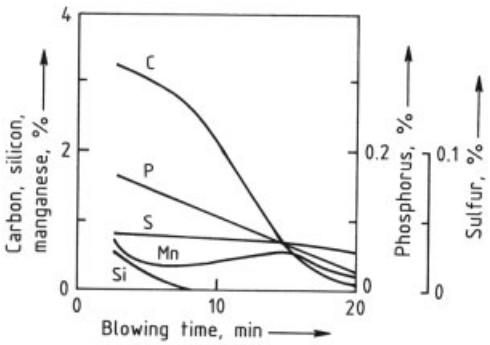

Figure 31. Changes in melt composition during the blow in a basic oxygen steelmaking converter (idealized) [46]

3.2.1.3.2. Dephosphorization

The conditions in steelmaking processes favor dephosphorization of iron, which takes place by oxidation of phosphorus and combination with a basic slag. The reaction proceeds as follows:

$$2\,[\mathrm{P}] + 5\,[\mathrm{O}] + 4\,(\mathrm{CaO}) \rightleftharpoons (4\,\mathrm{CaO} \cdot \mathrm{P_2O_5}) \qquad (3.15)$$

The equilibrium constant is given by:

$$\log K = \log \frac{a_{4\mathrm{CaO}\cdot\mathrm{P_2O_5}}}{a^4_{\mathrm{CaO}}\, a^2_{[\mathrm{P}]}\, a^5_{[\mathrm{O}]}} = \frac{61100}{T} - 23.3 \qquad (3.16)$$

This expression shows that the conditions for oxidative dephosphorization are: (1) oxidizing slag (high FeO activity); (2) high activity of lime in the slag; and (3) low temperature, as reaction (3.15) is exothermic. These conditions are particularly well satisfied at the start of the process.

In low-phosphorus pig iron, the type most commonly produced today, the slag formed during steelmaking consists mainly of CaO, SiO_2, and FeO_n, with P_2O_5 content ca. 1 – 2 %. The phosphorus equilibrium in this type of slag has been thoroughly investigated. In Figure 32, lines of equal phosphorus distribution at 1600 °C are shown for the system $\mathrm{CaO} - (\mathrm{SiO_2} + \mathrm{P_2O_5}) - (\mathrm{FeO}_n + \mathrm{MnO})$, with 6 % MnO and 2 % MgO. As expected, the lines are approximately parallel to the lime saturation line.

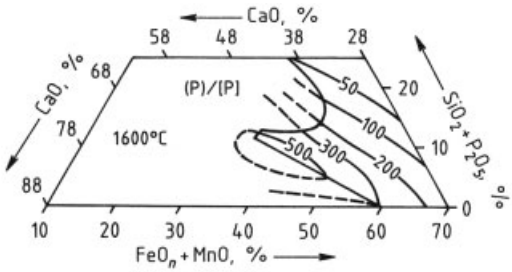

Figure 32. Lines of equal phosphorus distribution in the system $\mathrm{CaO} - (\mathrm{SiO_2} + \mathrm{P_2O_5}) - (\mathrm{FeO}_n + \mathrm{MnO})$ [48]

3.2.1.3.3. Desulfurization

Unlike dephosphorization, desulfurization is a reduction reaction. Since the oxidizing conditions in the converter do not favor the removal of sulfur from the melt, desulfurization is mainly carried out by treatment of the pig iron or of the metal in the ladle.

Desulfurization can be carried out by means of the precipitation reaction that occurs when elements with a high affinity for sulfur are added, either by taking up the sulfur into the slag phase, or by transferring it into the gas phase.

The affinity for sulfur of a desulfurizing agent is indicated by the solubility product of its sulfide in molten iron (Fig. 33) [49], the most stable of these being calcium sulfide. Although the alkaline earths and rare earths form relatively stable sulfides, their oxides are much more stable; the sulfides are formed only if the dissolved oxygen has been removed. The most commonly used desulfurizing agents are added to the steel in the form of alloys (CaSi, CaC_2, CaAl, etc.), by

special techniques. The desulfurization reaction can be represented by the typical equation:

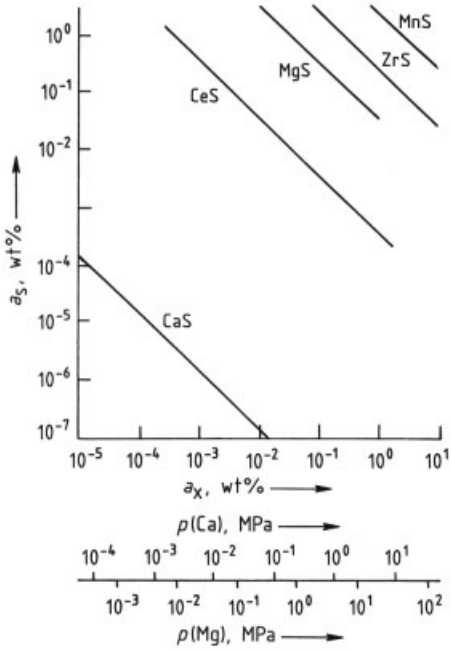

Figure 33. Equilibria between sulfur and various metallic elements in molten iron [49]

$$CaC_2(s) + [S] \longrightarrow CaS(s) + 2\,[C]$$

for which

$$\Delta G^\circ = -359\,000 + 109.4\,T\ \mathrm{J/mol} \tag{3.17}$$

Desulfurization by slag treatment is represented by the following equation, according to the ionic theory for slags:

$$(O^{2-}) + [S] \rightarrow (S^{2-}) + 2\,[O]$$

$$\Delta G^\circ = 71965 - 38\,T\ \mathrm{J/mol} \tag{3.18}$$

The most favorable thermodynamic conditions for slag desulfurization are: (1) high basicity of the slag; (2) low oxygen content in the slag or melt; and (3) high temperature.

Figure 34 shows the partition of sulfur between aluminum-containing steel and $CaO - Al_2O_3 - SiO_2 - MnO$ slag containing 5 % MgO at 1625 °C [50]. Slags of this type are used for the desulfurization of steel in ladle metallurgy.

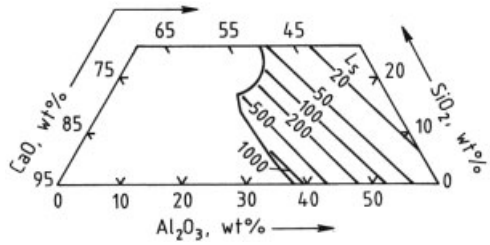

Figure 34. Sulfur partition coefficient L_S between aluminum-containing steel ($a_{Al} = 0.07\,\%$) and $CaO - Al_2O_3 - SiO_2 - MgO$ slags containing 5 % MgO at 1625 °C [50]

The sulfur in pig iron can also be removed by Na_2CO_3. The desulfurization reaction is:

$$Na_2CO_3 + [S] + 2\,[C] \longrightarrow (Na_2S) + 3\,CO(g) \tag{3.19}$$

or

$$Na_2CO_3 + [S] + [Si] \longrightarrow (Na_2S) + CO(g) + (SiO_2) \tag{3.20}$$

The Na_2S formed can be oxidized by air:

$$2\,(Na_2S) + 2\,O_2(g) \longrightarrow 2\,SO_2(g) + 2\,(Na_2O) \tag{3.21}$$

3.2.1.3.4. Deoxidation

At the end of the refining process, there is a considerable amount of oxygen dissolved in the molten steel; this can be removed by adding elements with a sufficient affinity for oxygen (precipitation deoxidation), or by slags which take up an amount of oxygen from the melt corresponding approximately to the equilibrium distribution (diffusion deoxidation).

The reaction between oxygen and a metal M in precipitation deoxidation can be represented by the equation:

$$x\,[M] + y\,[O] \longrightarrow (M_xO_y) \tag{3.22}$$

The equilibrium constant is:

$$K = \frac{a_{M_xO_y}}{a_M^x\, a_O^y} \tag{3.23}$$

In practice, either the deoxidation constant $K' = a_M^x\, a_O^y$ or the solubility product $K'' = [\,\%M]^x\,[\,\%O]^y$ is usually used. Small values of K' or K'' correspond to high deoxidizing power.

The relative deoxidizing powers of various elements are shown in Figure 35 [51]. The activities, or concentrations, of oxygen are plotted against those of the deoxidation elements on a logarithmic scale. The three deoxidation elements most commonly used, in increasing order of effectiveness, are Mn < Si < Al. Aluminum is one of the most powerful deoxidation elements.

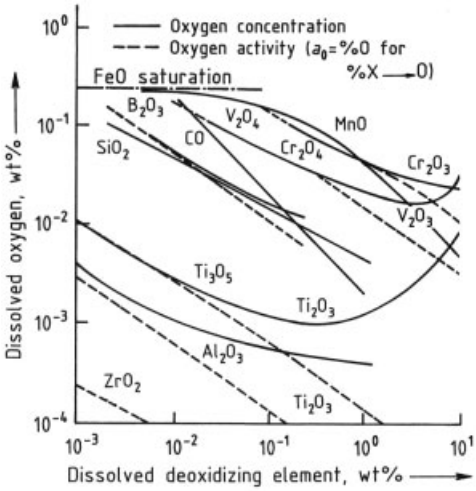

Figure 35. Deoxidation equilibria in molten iron at 1600 °C [51]

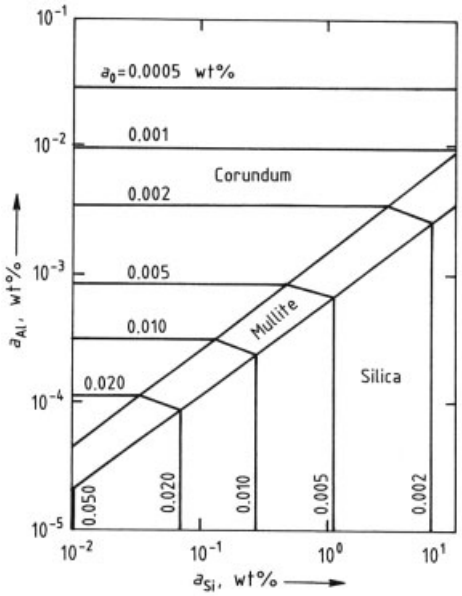

Figure 36. Deoxidation diagram for Al – Si – O at 1600 °C [49]

Deoxidation elements are often used in combination; the advantage of this lies partly in lowering the thermodynamic activity of the oxides formed, owing to compound formation, or dilution in the slag, and partly in the formation of oxidic slags with lower melting points, which separate more readily from the steel melt, or can be shaped at the rolling temperature. Figure 36 is a graphical representation of deoxidation with Al – Si. Various deoxidation products (alumina, mullite, and silicate) are formed, depending on the $[a_{Al}]/[a_{Si}]$ ratio.

In diffusion deoxidation, no inclusions are formed in the melt, but longer reaction times are necessary to remove the oxygen. The reaction can be represented by:

$$[O] + Fe(l) \longrightarrow (FeO) \tag{3.24}$$

Decreasing the FeO content of the slag leads to removal of oxygen from the melt. Diffusion deoxidation is being replaced by precipitation deoxidation.

3.2.1.4. Metal – Gas Reactions

If there is no slag on the surface of the molten metal, or if gas bubbles are formed within the melt, metallurgical reactions between the metal and gas phases take place. These include decarburization and degassing, either under reduced pressure, or by bubbling an inert gas such as argon through the melt.

As mentioned in Section 3.2.1.1, the concentrations of the gases dissolved in the iron depend on their partial pressures (Fig. 37). Pressure reduction can cause hydrogen and nitrogen to be removed from the melt, but this is not the case with oxygen. Although the solubility of oxygen varies with pressure according to the equation:

$$[\%O] = K_O \sqrt{p(O_2)}$$

it is not feasible to remove oxygen simply by pressure reduction, as the required oxygen partial pressure is less than that attainable in vacuum vessels (10 Pa). However, it is possible to use vacuum in conjunction with reaction (3.6) in the melt between carbon and oxygen to form the gaseous deoxidation product CO. The equilibrium constant:

$$K = \frac{p(CO)}{a_C a_O}$$

explains the strong pressure dependence. Figure 38 shows the equilibrium between oxygen and carbon as a function of p(CO) [52]. A decrease in p(CO) leads to simultaneous decarburization and deoxidation.

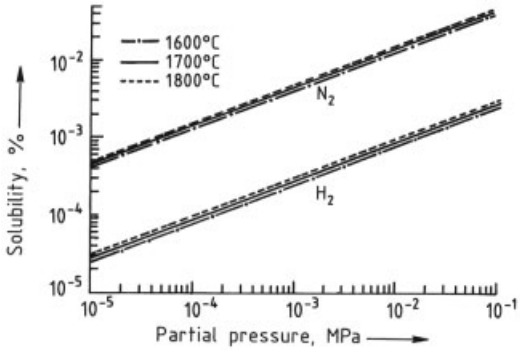

Figure 37. Solubility of nitrogen and hydrogen in molten iron [40]

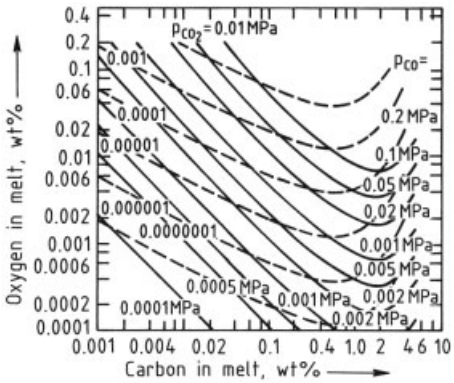

Figure 38. Equilibria between carbon and oxygen dissolved in molten iron with CO and CO_2 at various pressures and 1600 °C [52]

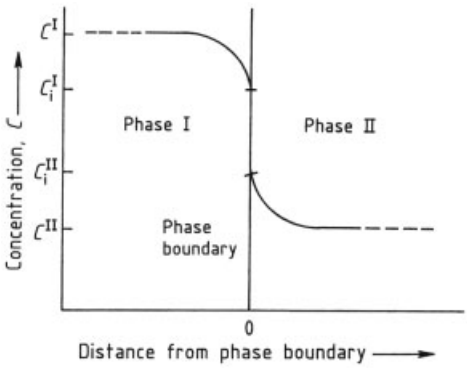

Figure 39. Schematic concentration profiles across a phase boundary

Dissolved hydrogen and nitrogen can also be removed without vacuum degassing by bubbling an inert gas through the melt. As the partial pressures of the hydrogen and nitrogen in the bubbles of inert gas are almost zero, hydrogen and nitrogen diffuse from the melt into the bubbles, owing to the concentration difference, and ascend with them.

3.2.2. Kinetics and Mass Transfer

3.2.2.1. Kinetics of Heterogeneous Reactions

In the metallurgy of steel production, the kinetics of the heterogeneous reactions are of interest because the individual homogeneous phases are usually in local thermodynamic equilibrium, so that macroscopics reactions do not occur. Heterogeneous reactions occur between the various phases that are not in thermodynamic equilibrium with each other. Figure 39 shows the concentration profile of a substance in two phases in contact with each other. The total reaction is made up of several steps:

1) Transport of the reactants to the phase boundary
2) Chemical reaction at the phase boundary
3) Transport of the reaction products from the phase boundary

The overall rate depends on the rate of the slowest step, known as the rate-determining step. If the overall process is considered, the mass flux density j is:

$$j = \left(C^{\mathrm{I}} - \frac{C^{\mathrm{II}}}{K}\right) / \left(\frac{\mathrm{I}}{\beta_{\mathrm{I}}} + \frac{\mathrm{I}}{k} + \frac{\mathrm{I}}{K\beta_{\mathrm{II}}}\right) \tag{3.25}$$

where C^{I}, C^{II} are concentrations in phases I and II, K is the equilibrium constant of the heterogeneous reaction, β_{I}, β_{II} are mass transfer coefficients, and k is the reaction rate constant. The equation:

$$\frac{\mathrm{I}}{\beta_{\mathrm{tot}}} = \frac{\mathrm{I}}{\beta_{\mathrm{I}}} + \frac{\mathrm{I}}{k} + \frac{\mathrm{I}}{K\beta_{\mathrm{II}}} \tag{3.26}$$

defines the overall mass transfer coefficient β_{tot}. Therefore, from Equation (3.25):

$$j = \beta_{\mathrm{tot}} \left(C^{\mathrm{I}} - \frac{C^{\mathrm{II}}}{K}\right) \tag{3.27}$$

The mass transfer coefficients β_{I} and β_{II} are obtained from theory of mass transfer by using experimental data. The rate constant k is obtained

from the chemical kinetics of phase boundary reactions for metallurgical systems. These rate constants are usually greater than the mass transfer coefficients of the transport processes that take place before and after the chemical reactions. The overall rate is therefore determined by mass transfer.

Equation (3.27) gives the rate of mass transfer to a phase boundary when the mass transfer coefficient is known. To obtain an expression for the variation with time of the concentration of a dissolved substance in the mother phase, a mass balance must be carried out which takes account of the nature of the phase contact and the characteristic parameters for emulsified systems [53].

3.2.2.2. Flow and Mass Transfer Coefficients

Mass transfer into a phase is made up of diffusion and convection. In the interior of the phase, sufficient flow takes place to give a uniform concentration. Concentration gradients exist only near the phase boundary. As there is no flow at this boundary layer, mass transport takes place by diffusion, and Fick's first law applies:

$$j = D\frac{C^{\mathrm{I}} - C^{\mathrm{I}*}}{\delta_{\mathrm{N}}} \tag{3.28}$$

where the asterisk indicates the phase boundary.

If this is combined with the flow equation $j = \beta(C^{\mathrm{I}} - C^{\mathrm{I}}*)$, the equation $\beta = D/\delta_{\mathrm{N}}$ is obtained. Here, δ_{N} denotes the thickness of the boundary layer, and D the diffusion coefficient of the overlying material. To determine β, it is necessary to determine the boundary layer thickness, whose value depends on flow conditions. Theoretical analysis and experimental investigation of the various flow fields that occur in metallurgical processes yield a formula for calculating the mass transfer coefficient in dimensionless form.

3.2.2.2.1. Metal – Glas Phase Boundaries with Nonturbulent Flow

The flow field is represented in Figure 40. The flow from the interior of the melt is first deflected at a stagnation point, e.g., the top of a column of bubbles or the middle of an inductively agitated melt, flows parallel to the upper surface, and is again deflected at a later stagnation point, e.g., the wall, and flows back into the interior. In this way, the surface is continuously renewed. The diffusion process takes place between the gas phase and the metal flowing to the surface. A mean mass transfer coefficient for this condition can be calculated in the form of a Sherwood number Sh, as a function of the Reynolds number Re and the Schmidt number Sc:

$$Sh = \frac{2}{\sqrt{\pi}} Re^{1/2} Sc^{1/2} \tag{3.29}$$

Where $Sh = \frac{\beta l}{D}$, $Re = \frac{ul}{\nu}$, $Sc = \frac{\nu}{D}$; β is the mean mass transfer coefficient (cm/s), l the flow length (cm), u the flow rate (cm/s), ν the kinematic viscosity of the melt (cm^2/s), and D the diffusion coefficient of the overlying material (cm^2/s).

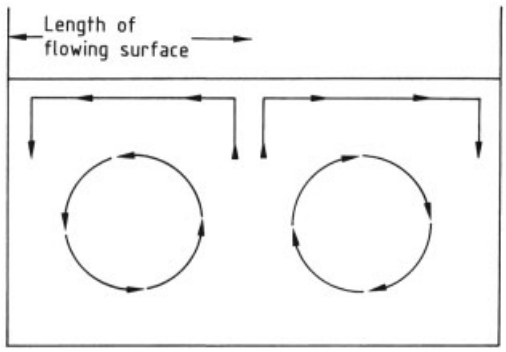

Figure 40. Mass transfer with continuous renewal of the free surface (schematic)

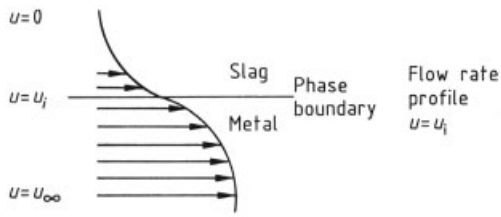

Figure 41. Flow field at a liquid – liquid phase boundary

3.2.2.2.2. Metal – Slag Phase Boundaries with Nonturbulent Flow

This flow field is represented in Figure 41. Friction takes place between the slag and the metal. The phase boundary moves at a velocity u_{i} during mass transfer. In the determination of the mass transfer, momentum transfer

due to friction must be taken into account. Equations for the flow film thicknesses on the metal and slag sides are first obtained. By including the condition that the shearing forces on each side of the phase boundary must be equal, the velocity of the phase boundary obtained and the mass transfer calculated.

3.2.2.2.3. Solid (Metal or Slag) – Liquid Phase Boundaries with Nonturbulent Flow

This condition represents, e.g., the processes of dissolution of alloys, or the solution abrasion of refractory materials. Mass transfer takes place on the liquid side. As shown in Figure 42, the flow rate, which is u_∞ within the melt, falls to zero at the boundary layer, and both the flow boundary layer and the diffusion boundary layer must be considered on the side of the melt. The mass transfer coefficient can be obtained by theoretical calculations from the equation:

$$Sh = 0.664 Re^{1/2} Sc^{1/3} \tag{3.30}$$

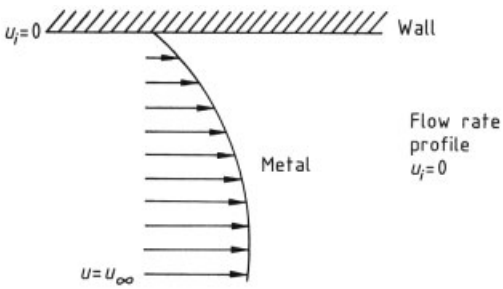

Figure 42. Mass transfer at a wall

3.2.2.2.4. Mass Transfer with Turbulent Flow

In metallurgical vessels, turbulent flow is usually present, caused by ascending CO bubbles and/or blowing with oxygen. In turbulent flow, turbulence spheres at the phase boundary contribute additional mass transfer by their flow components perpendicular to the phase boundary. This mass transfer can be regarded as turbulent diffusion, leading to an increased value of the diffusion coefficient. To calculate the mass transfer at a solid wall with turbulent flow, the following equation can be used [54]:

$$Sh = 0.037 Re^{0.8} Sc^{0.2} \tag{3.31}$$

At free surfaces [53], [54]:

$$Sh = 0.32 \left(\frac{D \varrho u_\tau^3}{\sigma_{\text{equiv}}} \right)^{1/2} \tag{3.32}$$

where D is the diffusion coefficient, ϱ the density of the melt, σ_{equiv} the equivalent interfacial tension of the melt, and u_τ the shearing force rate for turbulent flow.

The equivalent interfacial tension differs from the normal interfacial tension in that the back pressure, caused by the irregularity of the boundary surface, is expressed as the sum of the effects of interfacial and gravitational forces [54]. For steel melts, which have high interfacial tensions, σ_{equiv} can be assumed equal to the interfacial tension of the melt. The shearing force rate u_τ is a measure of the degree of turbulence. The value must be determined by turbulence measurements. For steel melts, this can at present only be investigated with the aid of mathematical models.

Equation (3.32) is also valid for metal – slag phase boundaries with turbulent flow, if there is no turbulence in the slags, and if the full weight of the stationary slag layer on the melt is considered when calculating the equivalent interfacial tension. If there is any turbulence in the slag, this should be included. However, it is often negligible.

3.3. Production Processes

Steel production depends on the availability of raw materials such as pig iron, scrap, and sponge iron. There are two fundamentally different methods (Fig. 43) [55]: blowing oxygen into liquid iron (mainly pig iron), and smelting iron-containing materials, such as scrap and sponge iron.

A typical example of the first method is the oxygen converter. In the electric furnace, and (now of less importance) the open hearth furnace, large amounts of scrap and sponge iron can be melted together. Depending on local conditions, mixed methods can also be used, in which various proportions of scrap

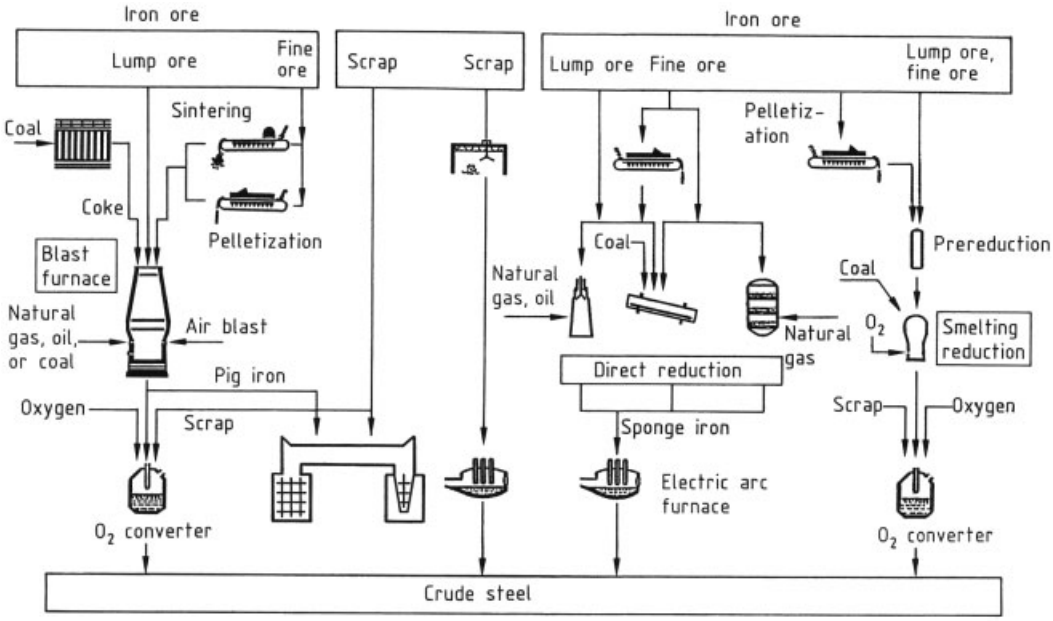

Figure 43. Crude steel production methods

and pig iron are processed, e.g., the KMS (Klöckner – Maxhütte – Stahlerzeugungsverfahren) and EOF (energy-optimized furnace) processes.

The object of all steel production processes is the removal of unwanted metallic, nonmetallic, and gas-forming elements from the raw materials, and the controlled addition of alloying elements to obtain the required characteristics of the various grades of steel.

The growth of world crude steel production, and the proportional contributions of the various processes are shown in Figures 44 and 45 [56], [57].

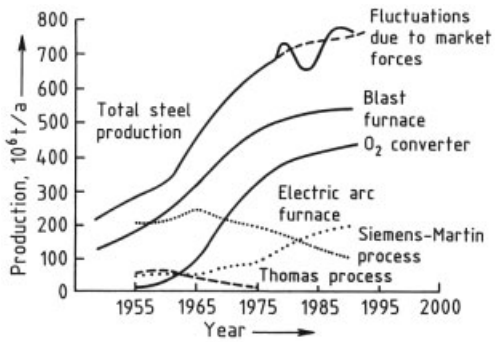

Figure 44. Growth of world pig iron and crude steel production by various processes

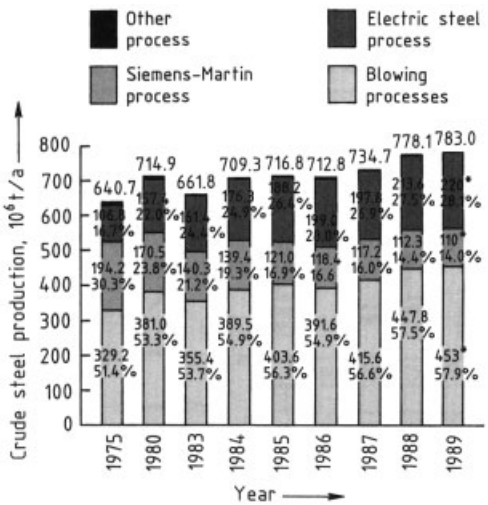

Figure 45. Growth of crude steel production, with contributions of the various production processes
∗ Estimated.

3.3.1. Oxygen-Blowing Processes

Unwanted elements in pig iron (and sometimes in scrap), principally carbon, silicon, and phosphorus, are removed by the oxidation reactions

described in Section 3.2, and transferred into the gas or slag phases (decarburization).

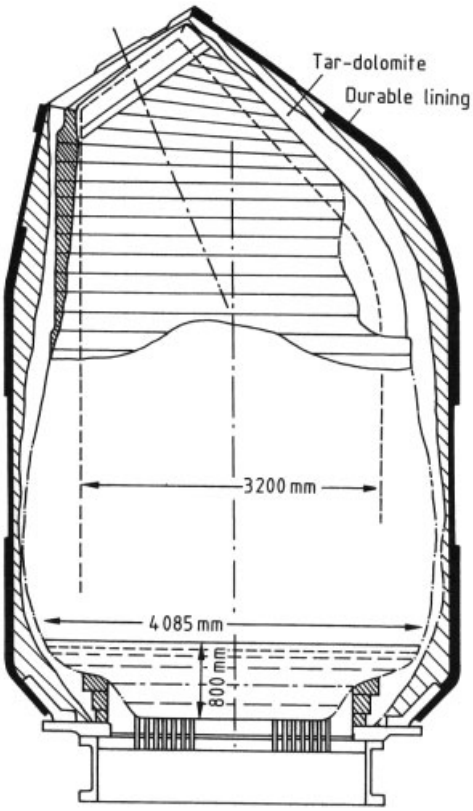

Figure 46. Section through a Thomas converter after 400 melts

The carbon monoxide produced is a source of thermal energy. After thorough purification, it is collected in gasholders and burned to provide heat for associated production plant. The oxides SiO_2 and P_2O_5 dissolve in the liquid slags formed from the added lime and the iron oxide produced during the oxidative decarburization processes.

The energy required to raise the temperature and melt the raw materials during the blowing process comes from the enthalpies of the oxidation reactions:

$$2\,C + O_2 \longrightarrow 2\,CO \quad \Delta H = -221\,kJ$$

$$2\,Fe + O_2 \longrightarrow 2\,FeO \quad \Delta H = -529\,kJ$$

$$Si + O_2 \longrightarrow SiO_2 \quad \Delta H = -878\,kJ$$

$$2\,Mn + O_2 \longrightarrow 2\,MnO \quad \Delta H = -771\,kJ$$

From ca. 1870, conversion of pig iron to liquid crude steel was carried out by blowing air, the oxidizing agent, through tuyeres in the base of a bottom-blown converter (tiltable steel vessel with refractory lining) containing a 20 – 80 t charge. Steel production processes using air blowing can be subdivided into the Bessemer process (only for low-phosphorus pig iron) and the Thomas process (capable of removing phosphorus). In the Bessemer process, the converter was lined with silica bricks. It was not possible to use a lime-containing slag to combine with the P_2O_5, owing to the acidic lining. In the Thomas process (Fig. 46), dolomite was used to line the converter, and a lime-based slag could then be used to treat phosphorus-containing pig iron [58].

The nitrogen in air dissolved to some extent in the liquid steel, and has a detrimental effect on its properties. Increasing quality requirements (aging stability, etc.) and the unfavorable economics of this process led to its discontinuation.

3.3.1.1. LD and OBM Processes

The first LD (Linz-Donawitz) converter, with a capacity of 30 t, was operated in Linz in 1952. The decarburization reaction is greatly speeded up by the use of pure oxygen, giving blowing times of 10 – 20 min. The oxidation enthalpies result in lower heat losses and enable more scrap or ore to be added. The converters have total charge 50 – 400 t, and are lined with dolomite or magnesite bricks [55].

Oxygen is injected from above through a water-cooled lance with several nozzles, onto the surface of the melt (Fig. 47). The high-pressure jets of oxygen (up to 1.2 MPa) oxidize the iron, carbon, and other elements at rates depending on their affinity for oxygen. The gas reactions cause thorough mixing of the molten materials, and this is maintained after completion of the blowing process by purging bricks built into the base of the vessel. Temperatures of 2500 –3000 °C are produced in the central reaction zone, and a reactive slag is rapidly formed from the added lime and the oxidized iron. The thermodynamic and kinetic processes are described in Sections 3.2.1 and 3.2.2. The heat and mass transfer processes take place according to the rules given in Section 3.2.2.

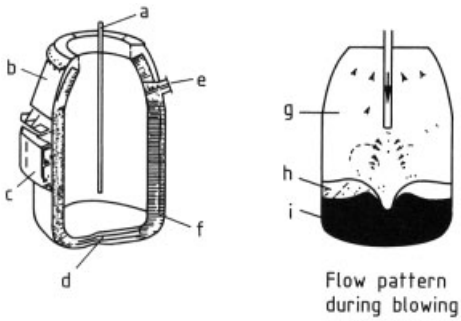

Figure 47. Top-blown oxygen converter
a) Oxygen lance; b) Converter top; c) Supporting ring; d) Converter bottom; e) Tapping hole; f) Refractory lining; g) Gas space; h) Slag layer; i) Molten metal

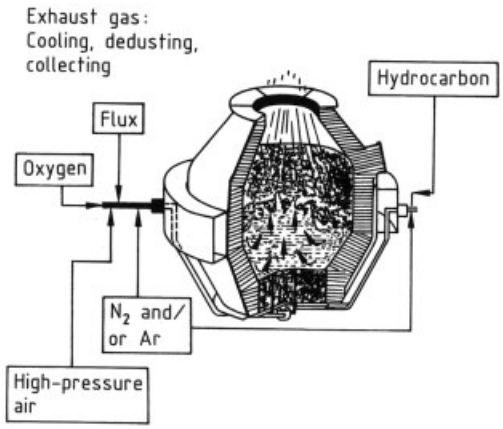

Figure 48. Section through an OBM converter
Oxygen is injected from below, through tuyeres cooled by hydrocarbons blown into the melt

The OBM (oxygen-blowing technique) (Fig. 48), a modification of the air blast processes, was developed in 1968 – 1969 in Sulzbach-Rosenberg [55]. Pure oxygen is passed through bottom tuyeres into the melt. These tuyeres are highly stressed, and are stabilized by cooling them with hydrocarbons. The cooling effect is a result of the endothermic decomposition of the hydrocarbons by the hot melt. The OBM process causes more intensive mixing of the steel and the slag compared with the method of blowing in the oxygen from above. This gives improved reaction kinetics and yield, owing to the lower iron content of the slag. The advantages of the more rapid slag formation and the less violent blow can be enhanced by adding powdered lime through the bottom of the converter, along with the oxygen. However, the capacity of the OBM process for melting scrap is limited in comparison with the LD process, owing to the lesser extent of afterburning of the waste gases, and the smaller amount of iron oxidation.

The combined blowing technique in general use today (Fig. 49) has been developed from the top-blowing process and past experience of the bottom-blowing process [55]. The combined process provides:

1) Homogeneous melts due to rapid breakdown of the scrap
2) Reduction of the blowing cycle time by 25 %
3) Higher yield of iron and alloying elements
4) Better control of the chemical composition
5) Improved purity
6) Lower quantity of slag and reduced tendency for it to be ejected
7) Increased lifetime of the converter lining
8) More favorable conditions for the measuring systems for process control

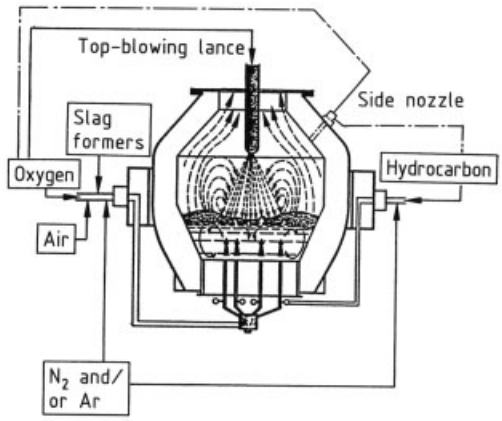

Figure 49. Combined blowing technique with top-blowing lance or side tuyere

Plant layouts do not differ greatly from one another. Figure 50 shows a side view of the converters and continuous casting plant of a steelworks [55].

Depending on the products required, the pig iron in the torpedo ladle car or filling pit is desulfurized by blowing with slag formers containing lime and/or magnesium. In Japan, for steel products with a very low phosphorus content, it is also usual to remove the silicon and phosphorus

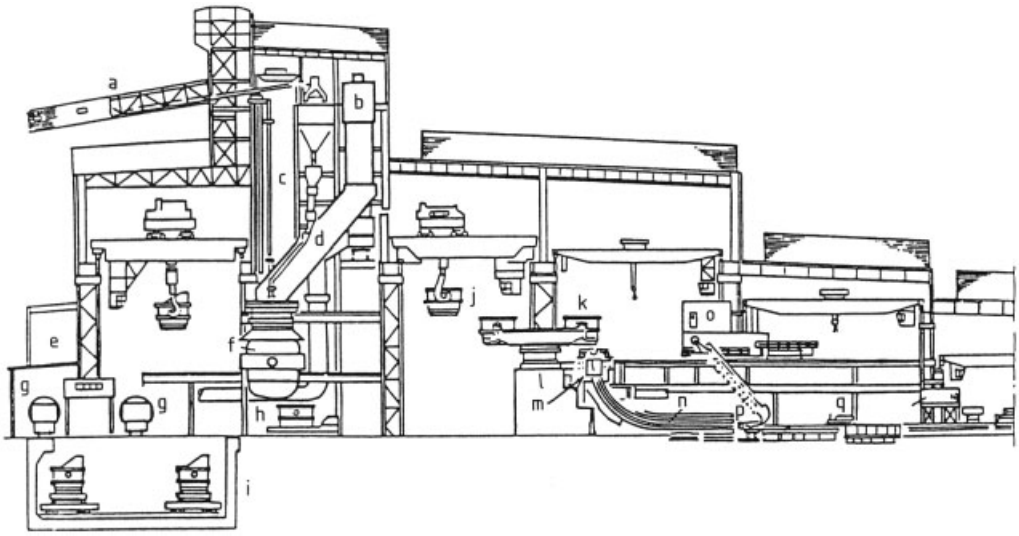

Figure 50. Converter and continuous casting plant
a) Feeder conveyor belt; b) High bunker; c) Blowing lance; d) Chute; e) Control room; f) 310 t converter; g) 600 t torpedo ladle car; h) Ladle transporter; i) Pig iron filling pit; j) Filled ladle; k) Pouring ladle; l) Ladle turret; m), n) Continuous casting machine; o) Continuous casting control room; p) Dummy bar; q) Cutting

[59]. In this case, the converter simply removes carbon from the melt (Fig. 51).

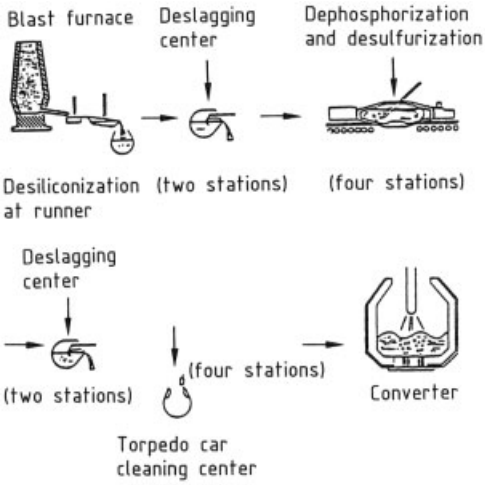

Figure 51. Hot metal pretreatment process at Mizushima works

The mathematical model of the process calculates the quantities of pig iron, scrap, and lime for the melt (based on the composition and temperature of the raw materials); the nominal tapping temperature; and the analysis of the crude steel, and starts the oxygen addition after the scrap and pig iron have been charged. Distribution of the oxygen and control of the height of the lance are computer controlled. The course of the blowing operation is set by the dynamic process control, which requires occasional fine adjustments based on "sub-lance measurements", that give the temperature and the carbon and oxygen content of the melt. Other elements may also be determined if required. Figure 52 shows how the melt temperature and composition change during the blowing operation. Here, the final composition is 0.059 % C, 0.031 % Mn, 0.018 % P, 0.019 % S, 0.003 % N, and 0.083 % O. Figure 53 shows the corresponding changes in the composition of the slag [60].

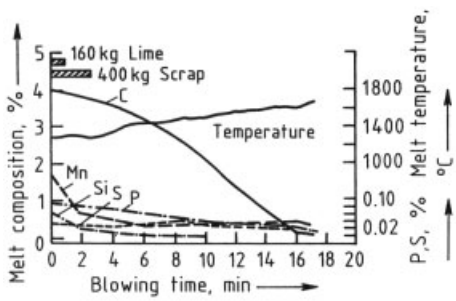

Figure 52. Changes in melt temperature and composition during blowing

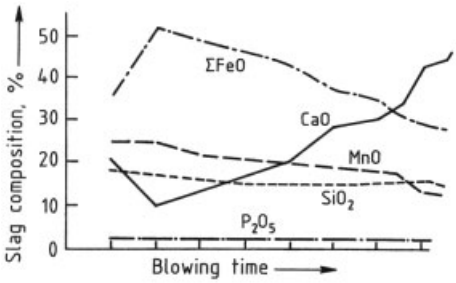

Figure 53. Changes in slag composition during blowing

At the end of the oxygen treatment, the converter is tilted and the steel is tapped into a ladle (a steel container with bottom-pouring facility and refractory lining). The steel is separated from the slag during emptying of the converter by means of a floating stopper introduced into the tapping spout. This prevents the slag, which is of lower density, from running out. Alternatively, an electromagnetic measurement in the tapping system can give a signal for the emptying process to be stopped.

To prevent emission of dust in the waste gases, all sources of dust, such as the pig iron handling and charging operations, lime feeding equipment, etc., are linked to a filtering system, usually cloth filters. The physical and chemical heat of the converter waste gases are utilized by installing a waste heat steam boiler with associated electrostatic filter or scrubber. The waste gases, at 1400 – 1600 °C, supply heat to the steam generation system, and the cooled gases, calorific value ca. 7000 kJ/m^3, are collected in a gasholder after purification. The dust collected in the filters or scrubbers can be recycled to the process, after suitable treatment.

3.3.1.2. Other Developments

Steel production processes based on an air or oxygen blast are autothermic, i.e., the energy requirement is provided by the physical and chemical heat from the pig iron. The use of scrap is limited by the composition and temperature of the pig iron, and varies between 0 and 20 %, depending on process variations and quality requirements.

Higher additions of scrap require the use of an allothermic process, with added fuel. In the KMS (Klöckner – Maxhütte – Steelmaking Process) process, additional energy is provided by adding hydrocarbons during scrap preheating, or powdered coal or coke during the blowing operation. This process is extremely flexible with regard to the metallic materials used, i.e., scrap, pig iron, and sponge iron.

In the KS process, the use of pig iron is completely eliminated. In the first phase, a carbon-rich melt is produced by melting scrap and adding carbon by a blowing process, together with an equivalent amount of oxygen. Heat transfer is optimized by afterburning the reaction gases. In the secound phase, conversion of the carbon-rich melt to steel is completed by the further addition of scrap.

In the Tula process (developed in the former Soviet Union), a carbon- containing material in lump form is added from above to a combined blowing converter. This provides the energy required for melting the scrap.

The EOF furnace (energy-optimized furnace) uses combustion of the waste gas and continuous preheating of the scrap in the hot waste gas stream to minimize the additional fuel requirements for an increased proportion of scrap in the melt. This has enabled up to 60 % steel scrap to be used.

3.3.2. Electric Steel Process

In the electric steel process, the heat required is obtained not by oxygen combustion of the accompanying elements in the pig iron, but from electrical energy. The conversion of electrical energy into heat can be achieved by an electric arc, induction, or plasma furnace. Electric steel processes are based on the use of scrap, with small amounts of solid pig iron. For a long time, the use of these processes was limited to the production of special steels, as energy consumption was high and the economics were unfavorable. Increases in the size of power stations and the capacity of electrical distribution systems have enabled batch weights to be increased, and the costs of energy, electrodes, refractory material, and capital investment to be reduced. Today, this process is second in importance only to the oxygen-blowing process in world crude steel production (see Fig. 47).

Over 90 % of all electric steel produced is by the use of the a.c. electric arc furnace [61]. Three

graphite electrodes carry the current through the furnace roof into the charge of metal. The electric arc formed melts the charge at temperatures up to 3500 °C. The furnace has the following essential components: the vessel or shell with a furnace door and a tapping hole; the roof which can be removed for charging; electrode arms which support the electrodes; tilting equipment for emptying the furnace; the furnace transformer; and the measuring and control equipment.

The melting procedure for the electric arc furnace comprises the following stages:

1) Charging
2) Melting
3) Oxidization (decarburization), with an increase in temperature
4) Tapping

The raw materials (scrap, sponge iron, pig iron, alloying elements, etc.) together with the required additives (lime, coal, ore, etc.) are loaded into special charging buckets which are then emptied into the furnace through a bottom opening. To fill the furnace, two or three charging operations are required, between which the scrap is partially melted.

The melting process begins with switching on the current and striking the arc. A supplementary blow with oxygen and fuel – oxygen mixtures accelerates melting and reduces current consumption. The duration of the melting period is determined by the electric power limit and the maximum heat load of the furnace shell.

Oxidation of elements, such as silicon, manganese, carbon, phosphorus, and sulfur begins during the melting phase as the liquid reactants, steel and slag, react with the added oxygen. The gaseous carbon monoxide formed from the reaction between the iron oxide and the carbon causes bubbling in the melt and purges the hydrogen and nitrogen. Removal of the phosphorus as calcium phosphate in the slag during the oxidation phase is important, and all the other metallurgical processes take place during the reductive secondary metallurgical treatment in the ladle or ladle furnace.

After the steel has reached the required temperature, the tap hole is opened and the furnace is tilted to empty it into the ladle below.

The development of the modern electric steel process began early in 1960, using powerful transformers and blowing with oxygen (Fig. 54) [62]. Modern furnace design and process control techniques have made it possible to increase the specific transformer power from 300 – 400 kVA/t to 800 – 1000 kVA/t. The electric furnaces of this new generation are operated at a power factor of 0.8 – 0.86, and enable tap-to-tap times of 50 – 80 min, and throughputs of > 100 t/h to be achieved (Fig. 55).

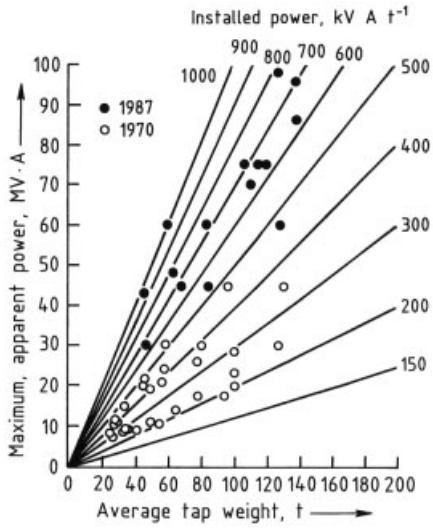

Figure 54. Development of specific installed power

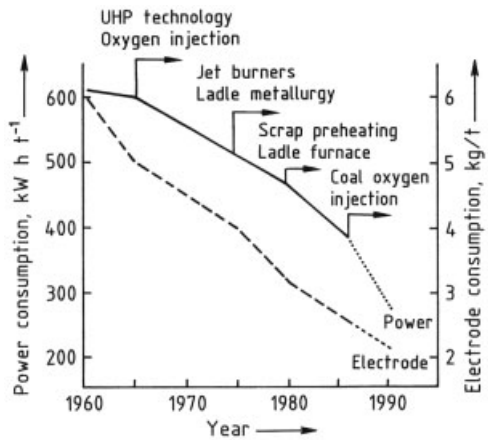

Figure 55. Effects of developments in electric arc furnace technology since 1965 on specific power and electrode consumption

The essential characteristics of modern electric steel production are:

1) Process automation

2) Enclosure of the furnace
3) Preheating of the scrap
4) Water- cooled wall and roof elements
5) Gas/oil – oxygen jet burners
6) Foam slagging
7) Current- carrying supporting arms for the electrodes
8) Electrode cooling
9) Facilities for charging through the furnace roof
10) Bottom stirring devices
11) Slag-free tapping (eccentric bottom tap hole)

Owing to their high energy input, modern electric furnaces produce considerable quantities of smoke and waste gas. To minimize environmental pollution, the furnaces are enclosed. Conditions in the workplace and surroundings can thus be maintained to a standard that meets legal requirements.

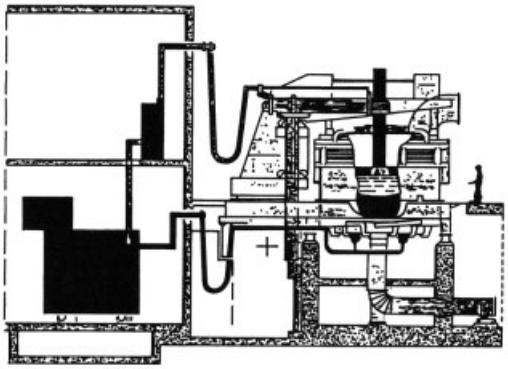

Figure 56. Direct current arc furnace
The bottom is made of electrically conducting refractory material to ensure reliable and durable continuity

The waste gases produced in the furnace are extracted through an aperture in the furnace roof, and the dust from various sources in the furnace area is collected in the upper part of the housing. This is constructed from sound-insulating elements, and reduces the noise level from the furnace to 20 – 25 dB. The waste gas is passed through a cooler which can also serve to recover the heat, and is then dedusted by cloth filters. In some operating conditions, the dust from the electric furnace can contain considerable quantities of heavy metals, such as zinc and lead, and so it is worth recovering them.

The slags produced consist of calcium silicate with 10 – 15 % iron oxide. These are quite stable, and can be used as fill for road and dam construction.

The very rapid development of the d.c. electric arc furnace began in the middle of 1985 (Fig. 56) [63]. Important advantages are:

1) Lower electrode consumption
2) Savings in electrical energy
3) Smaller effect on the electricity supply system
4) Symmetrical distribution of heating in the melt
5) Stirring effect on the melt

The central electrode becomes the cathode, and the melt the anode, the bottom of the furnace vessel being insulated from the wall. Current-carrying elements are built into the hearth, and provide an electrial connection to the melt.

The d.c. arc acts as a jet pump, directing the gases in and around the electric arc plasma toward the melt, causing efficient heat transfer from the electrode to the melt. The largest d.c. furnace operating has a capacity of 130 t, with a transformer power of 100 MV · A, and a steel output of 10^6 t/a.

An important feature of modern electric steel production is computer control. The supervision and control of the production units, combined with precise data collection, leads to improved material flow and more efficient utilization of energy and alloying metals, and ensures high and reproducible quality.

3.3.2.1. Special Processes

In foundries, induction furnaces are widely used for melting and holding steel and nonferrous alloys [63]. The water- cooled, current-carrying coil surrounds a refractory-lined crucible (Fig. 57) that holds the material to be melted. Electricity is supplied by a medium-frequency (MF) frequency converter, fed from a high- or low-voltage transformer. The current density can be 1 MW/t, and an output of 15 t/h can be obtained from an MF furnace with a capacity of 8 t.

Since the development of high-power plasma burners, it has become possible to use plasma melting furnaces for the production of special steels in 40 t batches. The energy is supplied by adjustable d.c. argon plasma burners, built into

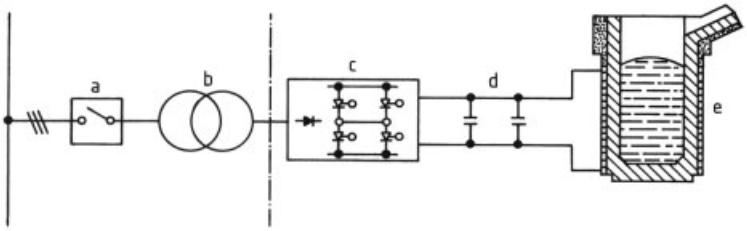

Figure 57. Electric induction furnace
a) Circuit breaker; b) Transformer; c) Frequency converter; d) Capacitor bank; e) Induction furnaceMelting capacities of 15 t/h can be achieved at current densities of 1 MW/t in an 8 t medium-frequency furnace

the furnace shell. The burner produces a plasma (ionized gas) in the electric arc, at temperatures 3000 – 5000 °C, and this melts the furnace contents. The use of inert gases for the plasma protects the melt against reactions with oxygen, nitrogen, hydrogen, etc., thereby ensuring a high yield of alloy. Carburization of the melt does not occur, as there are no graphite electrodes. Vacuum treatment of the steel is not necessary.

For the production of special alloys containing high proportions of alloying elements with a high affinity for oxygen (Al, Ti, etc.), it is common to use inductively heated or electron beam vacuum furnaces.

3.3.2.2. The Siemens – Martin Open Hearth Process

In 1960, the Siemens – Martin open hearth process (Fig. 58) produced ca. 70 % of the world's crude steel, but this rapidly declined following the growth of the oxygen-blowing and electric steel processes, and ceased operation after a few years.

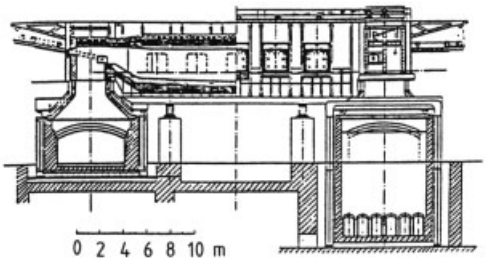

Figure 58. German – American design for a 250 t furnace for gas and oil firing [64]

It was developed for melting scrap, and was characterized by its great flexibility with respect to the proportions of pig iron and scrap used. The tank-shaped furnaces are heated by fossil fuels such as gas or oil, burned in hot air. The regenerative air preheating invented by Siemens takes place in "checker chambers" that contain refractory bricks heated by the waste gas, and are situated under the furnace. The combustion temperatures achieved enable steel scrap to be melted.

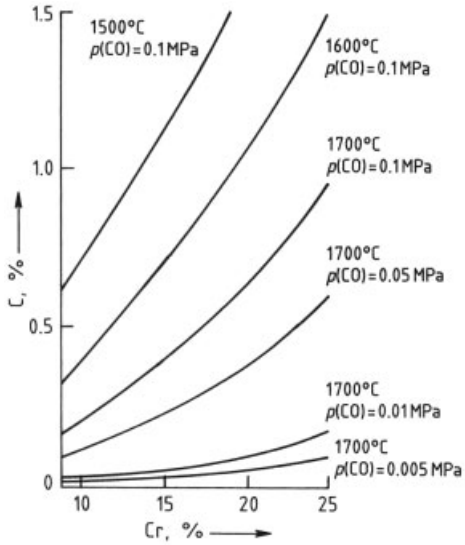

Figure 59. The Fe – C – Cr – O system

The capacity of these furnaces is 50 – 1000 t. The use of air for the combustion leads to the production of large quantities of waste gas, whose purification, sometimes including desulfurization, is difficult and expensive. Additional oxygen, introduced through lances or burners (tandem furnaces), has been used to improve operation, but this leads to severe erosion of the refractory furnace lining. The open hearth process

has in general lost its importance, as it cannot achieve either the output or flexibility of blowing processes.

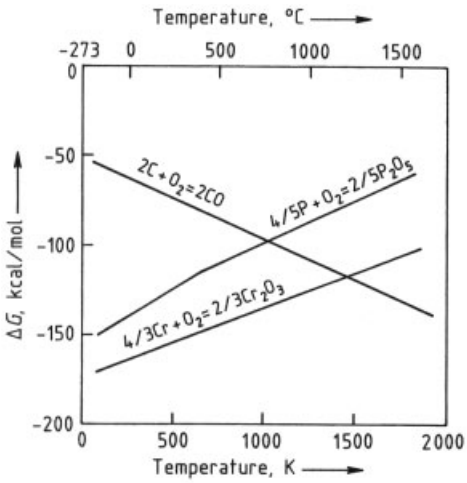

Figure 60. Free energy of oxidation of Cr, P, and C as a function of temperature

3.3.3. Production of Stainless Steels

Although world steel production is subject to very severe fluctuations dictated by market conditions, the average growth rate of stainless steels (i.e., resistant to rust, acid, and heat) between 1950 and 1985 was 6 %. In 1990, total output reached 10^7 t. Special metallurgical processes have been developed for steels with a chromium content 12 – 28 % and low carbon content [65].

Production conditions are determined by the thermodynamic equilibria of the reactions of oxygen with carbon and chromium (Figs. 59, 60).

In the first process variation, which is mainly used in western countries, an electric furnace is used to melt scrap and alloying elements. The composition of the molten metal used corresponds approximately to that of the desired steel product, apart from the carbon content. Decarburization and desulfurization are carried out, either in a converter or a ladle in vacuum, or in a combination of the two (Fig. 61).

The converters are operated by the combined blowing technique (i.e., via top lance and bottom tuyeres) to increase the rate of decarburization. Nitrogen is used as the protective gas for the bottom tuyeres and to reduce the oxygen partial pressure. Argon is used to give low carbon contens and to avoid nitrogen pick up of the steel. After decarburization, the Cr_2O_3-containing slag is reduced with silicon, and the sulfur combines with the lime-based slag. In an

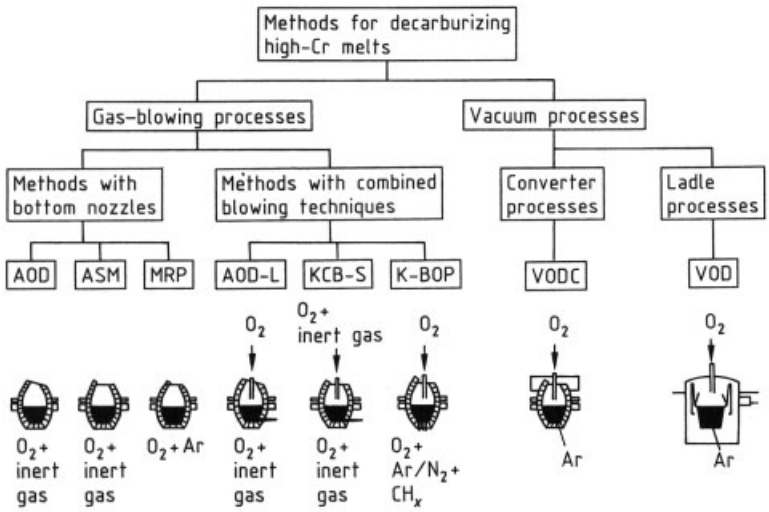

Figure 61. Methods for decarburizing high- chromium melts
AOD = argon – oxygen decarburization; ASM = argon secondary metallurgy; MRP = metal refining process; AOD-L = argon – oxygen decarburization lance; KCB-S = Krupp combined blowing stainless; K-BOP = Kawasaki basic oxygen process; VODC = vacuum – oxygen decarburization converter; VOD = vacuum – oxygen decarburization

alternative process, the premelt is poured from the electric furnace into a ladle placed in a vacuum chamber. Decarburization is then achieved by a stream of oxygen which enters via the lid of the vacuum chamber as the pressure is reduced, and is then blown through a lance into or over the melt.

In the vacuum process, the rate of decarburization reaches 0.03 %/min, a considerably lower figure than that achieved by the converter processes. The latter give high productivity and favorable economics and are therefore, used for high-volume production of crude stainless steels. Vacuum processes are especially suitable for the production of special steels with very low carbon content (30 ppm), and low levels of nitrogen and hydrogen.

A combination of both processes utilizes the high decarburization rate of the converter in the high- carbon content region, and the low carbon levels achievable by the vacuum process.

In Japan, a two-stage process is used for the melt reduction of chromium ores to produce stainless steels (Fig. 62) [66]. Two 85 t combined blowing converters are used. In the first of these, chromium ore is added to a dephosphorized pig iron bath where it is reduced to chromium with coal and oxygen. The carbon-containing chromium melt is tapped and, after slag removal, charged to the second converter, where it is decarburized. The heat of reaction enables added scrap and alloying metals to be melted. The advantage of the combined blowing converter is that, as well as passing oxygen and argon through the bottom tuyeres, it is also possible to decarburize with pure oxygen, as the tuyeres are cooled with hydrocarbons.

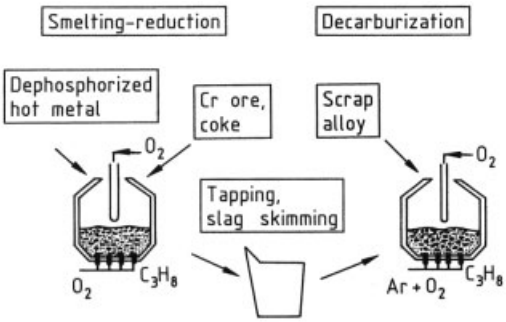

Figure 62. Two-stage stainless steel production (smelting – reduction) process

3.4. Secondary Metallurgy

The steelmaking processes described in Section 3.3, with the exception of plasma furnaces, take place in oxidizing atmospheres, and are designed to maximize the output and cost-effectiveness of crude steel production.

The secondary metallurgical treatment processes produce the specific properties required for the various applications of the steels concerned. They include:

1) Adjustment of alloy composition
2) Deoxidation
3) Homogenization of temperature and composition
4) Temperature adjustment
5) Decarburization
6) Desulfurization
7) Degassing
8) Control of inclusion shape
9) Adjustment of purity

The processes take place in the steel ladle, so that the steel production equipment is not involved, and correspond to the quality program and the production plan (Fig. 63) [55].

It is essential that the slag produced during the oxidation processes should be separated cleanly, and that the refractory lining of the ladle should not act as a source of oxygen during the reduction phase, or have a detrimental effect on the composition of the reaction slag.

To hold back the slag formed during the oxygen-blowing process, floating stoppers are used to close off the tap hole. Their function depends on the different densities of the steel and slag; alternatively, mechanical or pneumatic devices can be used. In electric furnaces, there is a curved region or syphon near the tapping hole, or an oval hearth is provided so that no slag is discharged, provided that a residue of 15 – 20 % steel is retained in the furnace.

3.4.1. Steel Treatment at Atmospheric Pressure

Secondary metallurgy requires mixing in the ladle. This is achieved by passing gases through porous bricks in the bottom of the ladle, or by injecting them through refractory- covered lances

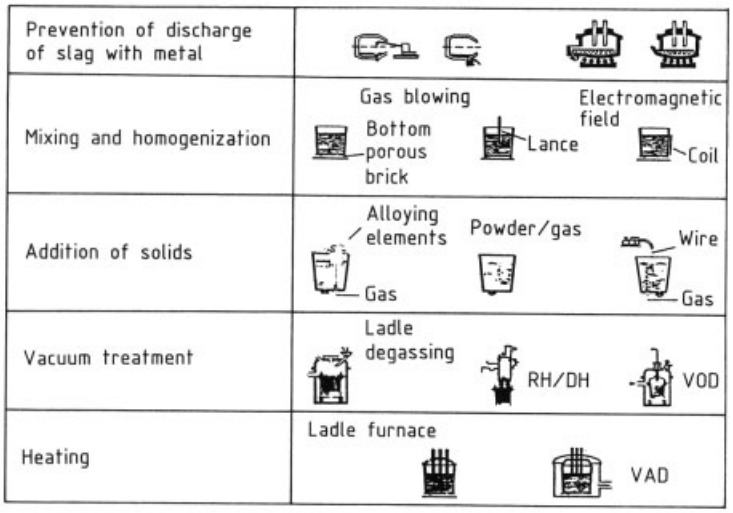

Figure 63. Secondary metallurgy processes in steel production
VAD = vacuum arc decarburization

dipping into the melt, thus producing a rising stream of bubbles. A possible alternative is to use the stirring effect of electromagnetic fields acting through the nonmagnetic walls of the ladle. Depending on the amount of mixing energy generated, this can cause circulation and homogenization of the steel under the slag layer, or mixing of the steel with the slag with chemical reaction, e.g., desulfurization. If ladle covers are used, the steel can be protected from reoxidation and falling temperature.

The solid materials can be added in lump form through openings in the ladle lid, or injected in powder form through lances, using a gas as transport medium. For exactly measured additions, solid wire can be added (e.g., aluminum), or wire filled with other substances (e.g., CaSi), which can be added almost without loss of material.

3.4.1.1. Deoxidation

As well as alloy addition, deoxidation (removal of residual oxygen) is necessary in the production of quality steels. Depending on the carbon content of the steel, it can contain 100 – 300 ppm dissolved oxygen, and this is removed by adding elements with an affinity for oxygen (e.g., silicon or aluminum, to form silica or alumina). These oxides are of lower density than steel, but their particle size also greatly affects their upward movement. The deoxidation products are encouraged to coagulate by agitating the steel in the ladle with argon or nitrogen. The ascending reaction products take up the liquid reactive ladle slag. The amount of these materials remaining in the steel determines its oxide content, and affects its mechanical properties. By using a combination of deoxidizing metals, a very low level of oxygen in the steel can be achieved (Fig. 64) [68].

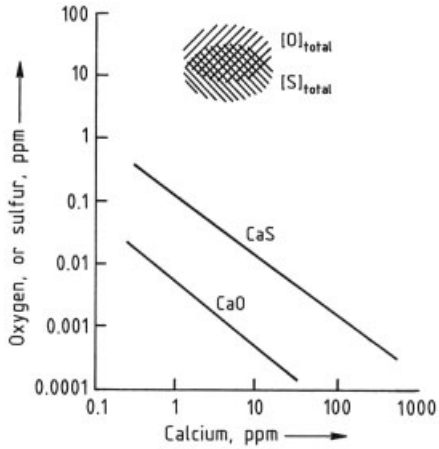

Figure 64. Equilibrium of calcium and sulfur or oxygen in liquid iron at unit activity of CaS and CaO [66], [67]

3.4.1.2. Desulfurization

For high rates of steel production, only a limited amount of desulfurization can be carried out in the blast furnace. When a blast furnace process is optimized for low fuel consumption, the sulfur distribution coefficient between the slag and the pig iron is 20, which, for current mass ratios, leads to a sulfur content of 500 ppm in the pig iron (Fig. 65). Depending on quality requirements, the first step is the injection of lime and other calcium – magnesium compounds into the pig iron ladle. The relatively poor sulfur partition coefficient (ca. 5) in the oxygen injection steelmaking process is due to the oxidizing reactions. However, if the steel has a high enough aluminum content after deoxidation, this gives good reducing conditions. If the correct desulfurizing agents are chosen (CaO, Mg, CaC_2, CaSi), or lime-saturated desulfurizing slags are used, sulfur distribution coefficients of 1000 can be achieved [69]. The desulfurization of the pig iron and steel, especially if both these processes are used, enables the required quality to be achieved flexibly and economically. The use of desulfurizing slags requires intensive mixing of the steel and the slag under reducing conditions.

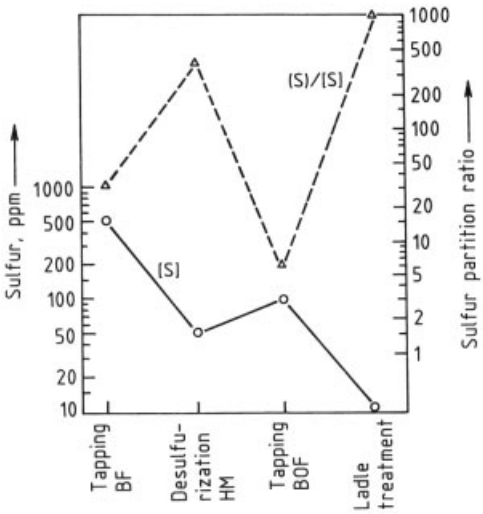

Figure 65. Process stages for adjustment of low sulfur content in liquid steel

Figure 66 shows the changes in sulfur content in steel produced on a large scale over a number of years [68]. In this example, which refers to steel used for pipes, the development of toughness and sulfur content is shown. The use of calcium compounds under certain conditions leads to the formation of spheroidal inclusions that are not deformed during the rolling process. This has enabled high-strength steels with good toughness properties to be developed.

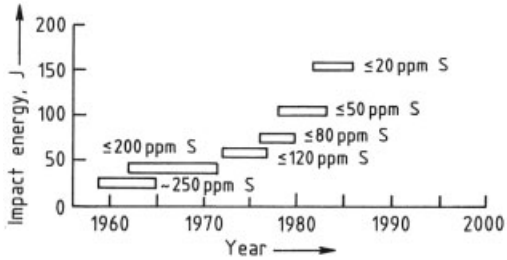

Figure 66. Development of toughness and sulfur content

3.4.1.3. Dephosphorization

Phosphorus removal is carried out under oxidizing conditions during production of the crude steel, and is influenced by the lime saturation, the P_2O_5 content, and the temperature of the slag produced during decarburization (Section 3.3). The production of low-phosphorus steel requires low-P, low-Mn pig iron, and effective removal of the P_2O_5-containing slag produced during decarburization and before deoxidation.

3.4.1.4. Temperature Control

The temperature changes during steel production affect the course of the reaction (as shown by the example of dephosphorization), and also affect the economics (wear of the refractory linings of furnaces and ladles). The secondary metallurgical operations: alloying, desulfurization, degassing, etc., require time, and therefore involve heat losses. Without ladle heating, this would necessitate the use of unfeasibly high tapping temperatures. Electrically heated ladle furnaces are increasingly used in steelworks – initially in EF steel plants and then in BOF plants – to ensure an optimum temperature regime during the process.

Figure 67 illustrates the increasing use of ladle furnaces [67]. The furnaces consist of an assembly of two or three a.c. or d.c. electrodes, and a water-cooled roof, with fume extraction

equipment, and a working door for adding materials and for the control equipment (Fig. 68) [69].

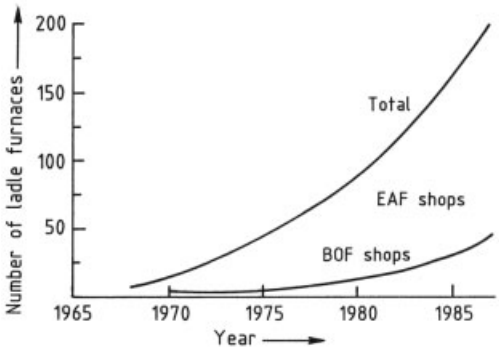

Figure 67. Ladle furnace development

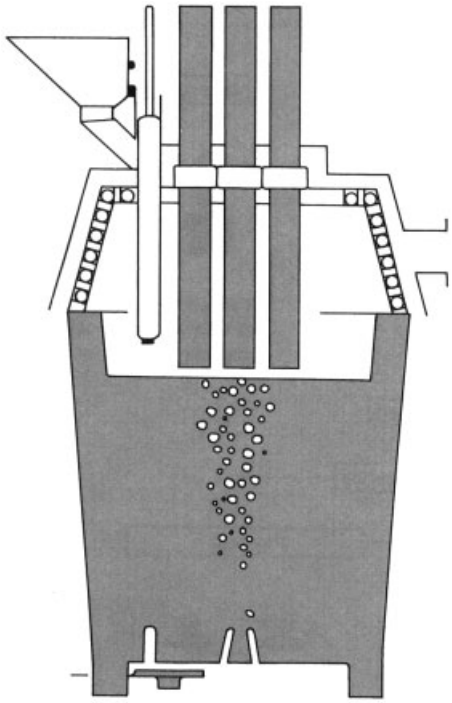

Figure 68. Typical layout for a ladle furnace

The transformer capacity is ca. 150 kV · A per tonne ladle contents, and gives a heating rate of 5 °C/min. The diameter of the pitch circle has been reduced to 600 – 700 mm in new designs to protect the refractory ladle lining. Agitation of the steel and slag to homogenize the steel, and to promote the steel – slag reactions, is very important in the operation of the ladle furnace. The electricity consumption for treatment time ca. 40 min is 30 – 40 KWh/t, and the amount of electrode material consumed is 0.3 kg per tonne of steel. With a ladle furnace, the temperature at which the steel is tapped can be reduced by 100 °C without affecting the secondary metallurgical reactions.

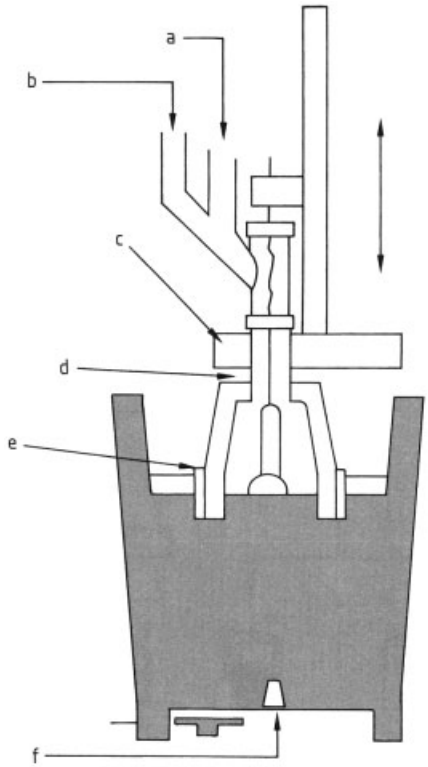

Figure 69. Equipment for the CAS process
a) Dust- collecting hood; b) Alloy chute; c) Snorkel raising and lowering unit; d) Oxygen lance; e) Bell; f) Porous plug

An alternative to the ladle furnace is the CAS/CAS – OB (Composition Adjustment by Sealed Argon Bubbling) developed by Nippon Steel [69], the IR-UT (Injection Refining-Up Temperature) process proposed by Sumitomo. The heat evolved by the oxidation of aluminum or silicon is used to heat the steel, instead of electrical energy. The process is suitable for producing flat products that are insensitive to hydrogen. After the slag-free steel has been tapped, it is agitated by inert gas, and alloyed in a bell (Fig. 69). If the temperature has decreased too much, the melt can be heated to the required pouring temperature by adding aluminum and oxidizing it

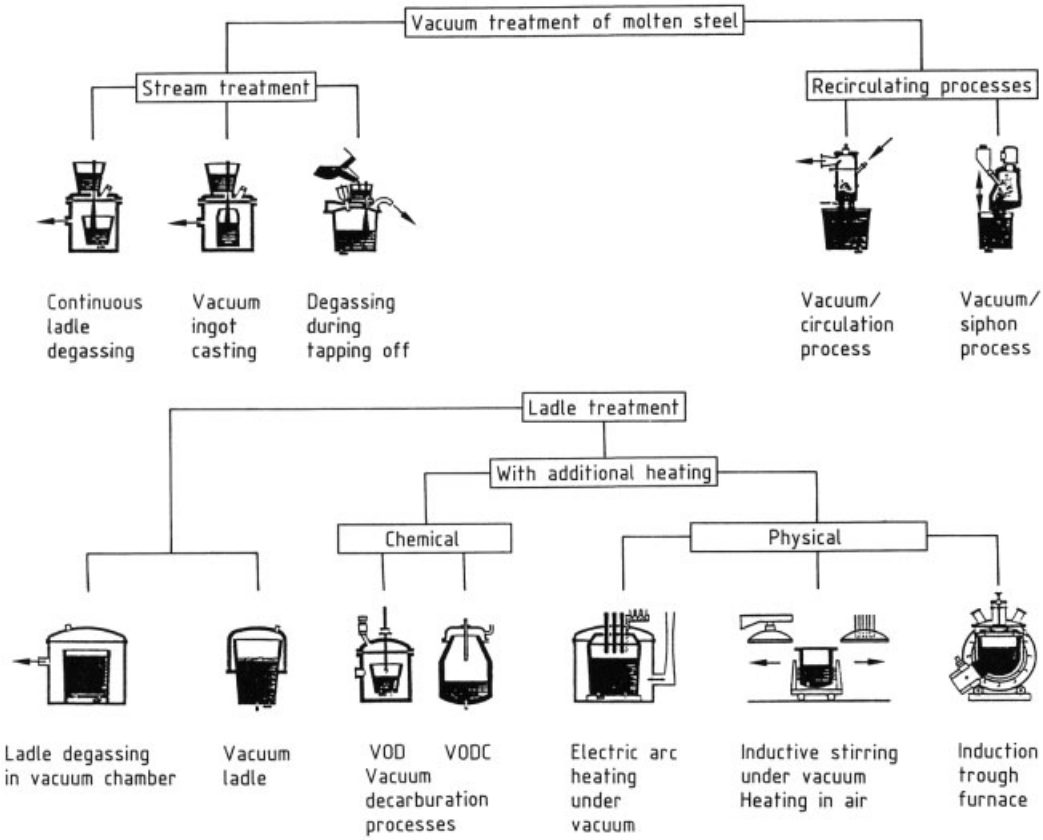

Figure 70. Vacuum treatment of steel

with oxygen. The addition of O_2 at a rate of $11\,m^3\,h^{-1}\,t^{-1}$ gives a heating rate of 7 °C/min.

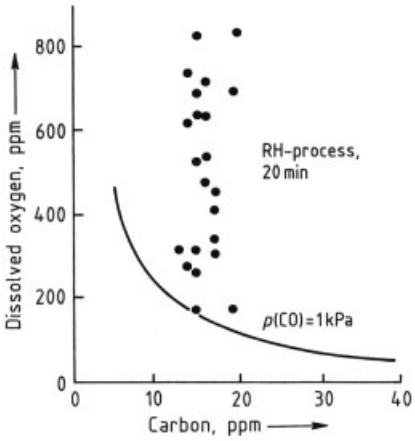

Figure 71. Final carbon content as a function of dissolved oxygen

3.4.2. Vacuum Treatment

The vacuum treatment of steel is used increasingly owing to its versatility and special advantages, and is indispensable for removal of gases dissolved in the steel, such as hydrogen and nitrogen. The reactions proceed at very low pressure (0.5 hPa). The various processes are shown in Figure 70 [55]. They can be divided into three groups: recirculating treatment of small quantities of metal; ladle treatment; and treatment of a stream of molten metal.

The recirculating processes are characterized by high throughputs. In the RH plant, two refractory- clad snorkels dip into the steel being treated; in the DH plant only one snorkel is used. Whereas in the RH process, the steel is transferred into the vacuum vessel by blowing inert gas via a nozzle (gas lift), in the DH plant the movement of the steel in the vessel is achieved

by the up-and-down motion of the vessel. In both processes, oxygen lances can be provided.

3.4.2.1. Decarburization

The decarburization reaction is strongly pressure dependent, so that when it is carried out under vacuum, fine decarburization is the result. As shown in Figure 71, carbon content < 50 ppm can be achieved reproducibly under optimal process conditions [68].

The required treatment time is strongly pressure dependent. A rapid decrease in pressure is only feasible if the vacuum vessel is fairly large, as the melt splashes severely. If the amount of oxygen dissolved in the steel is insufficient for the decarburization reaction, oxygen is blown in with a lance, liberating heat.

3.4.2.2. Nitrogen Removal

High nitrogen content has a detrimental effect on the deep-drawing properties of continuously annealed mild steel, and also reduces the resistance of tube steels to hydrogen-induced cracking (HIC). The change in nitrogen content during oxygen blowing is shown in Figure 72 [68].

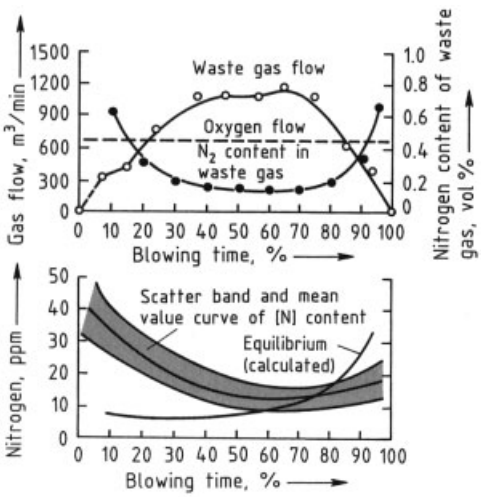

Figure 72. Nitrogen removal

If the nitrogen content in the pig iron is low, nitrogen levels of 30 ppm in plain carbon steels can be achieved. However, steel can pick up nitrogen if nitrogen-containing alloying elements are used, and nitrogen removal may then be necessary. This can be carried out by vacuum treatment under controlled conditions [68]. Nitrogen removal by ladle degassing (after desulfurization) can give a nitrogen content ca. 50 ppm (Fig. 73). Sulfur interferes with nitrogen removal.

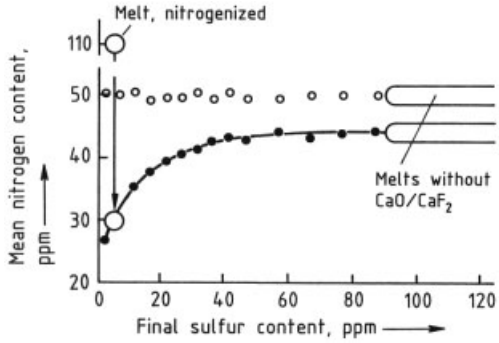

Figure 73. Mean nitrogen content before and after vacuum treatment
○ = [N] before ladle degassing; ● = [N] after ladle degassing

The ladle degassing operation gives intensive phase contact between the steel and the slag, which has a beneficial effect on desulfurization. With 10 ppm sulfur and low oxygen content, a nitrogen level of 30 ppm in the steel can be achieved, whether the starting level is 50 or 110 ppm.

3.4.2.3. Dehydrogenation

To avoid flaking (material separation) in the steel, the hydrogen content must be lower than its maximum solubility. For large cross sections, i.e., conditions favoring segregation, it is necessary to reduce the hydrogen content of the liquid steel to < 0.8 ppm. The diffusion reaction is controlled by the difference between the hydrogen partial pressure in the steel and that in the gas at the phase boundary, and therefore proceeds more rapidly as the pressure in the vessel decreases. The formation of new surfaces by intensive purging with inert gas, or otherwise agitating the steel in RH and DH equipment promotes hydrogen removal, as shown in Figure 74 [70].

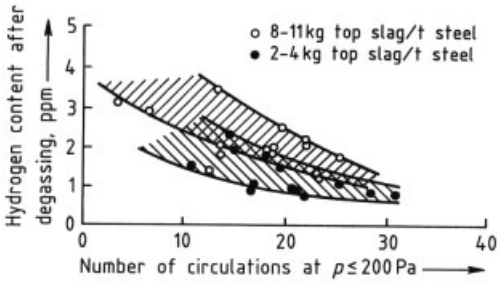

Figure 74. Hydrogen removal

The choice of steel production and secondary metallurgical processes depends on the product, the production rate, the raw material, and the energy situation. In Europe, high-quality flat products are produced by the route: blast furnace – oxygen blowing – RH or DH vacuum treatment – continuous casting. Long products are mainly produced by the route: electric furnace – ladle furnace – continuous casting. New developments such as COREX – EOF – KES and continuous casting to give cast-to-shape products, promise technological and economic stimuli, especially with respect to the environment.

3.5. Casting and Solidification

3.5.1. Fundamentals

All processes of casting and solidification depend on the basic processes of heat transfer, mass transfer, and interface kinetics. This applies not only to ingot casting, continuous casting, and foundry work, but also to remelting and spraying processes, crystal growth, the production of composite materials from a melt, and rapid solidification to give a crystalline or amorphous structure. These basic principles are valid for all metals and alloys, and the theory is very advanced [71–76].

Nevertheless, steel is special material, having its own characteristic casting and solidification properties that differ from those of both light and heavy nonferrous metals. The peculiarities of steel include: high melting point, low thermal conductivity, phase transformations of the iron, high density, and strong affinity for oxygen. The wide variety of alloy compositions of steel is also reflected in its solidification properties, i.e., the way in which the thermal processes and structure formation take place [77].

A rigorous scientific treatment of heat and mass transfer and interface kinetics cannot be given here, but their most important consequences are indicated, including solidification morphology, micro- and macrosegregation phenomena, sulfide precipitation, and the formation of inclusions by deoxidation. Thermal stress and mechanical deformation also merit consideration, as they particularly affect the quality of the product; processing techniques and mechanical engineering are tailored to suit them.

Solidification Morphology of Steel. The general criteria for the stability of a nominally planar solidification front are: high superheat, convection, low rate of solidification, low content of alloying elements, and a high partition ratio. These conditions are seldom met in casting operations in the steel industry. Much more frequently, there is "constitutional undercooling" at the solidification front, so that steel almost always solidifies dendritically.

Figure 75 shows a scanning electron micrograph of a "quasi-decanted" dendritic soldification front from the central shrinkage cavity of a continuously cast billet of high-temperature steel.

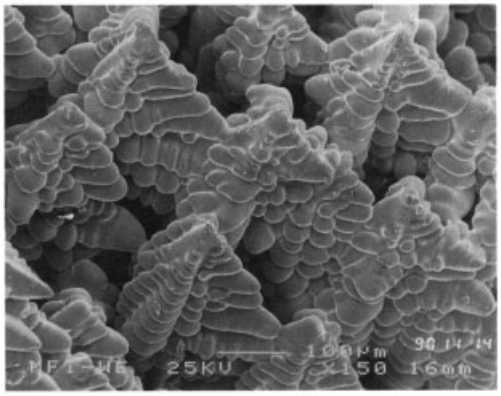

Figure 75. Dendrite morphology in center of a continuously cast billet of steel, grade 10 CrMo 910

The dendritic crystals consist of primary, secondary, and tertiary arms, all with crystallographic growth direction [100]. The dendritic growth mainly determines the spatial pattern of segregation.

Primary and secondary dendrite arm spacings in a typical continuously cast steel slab are shown in Figure 76. Plots of these spacings against distance from a surface of the slab are parabolic, with the widest spacings in the center of the slab.

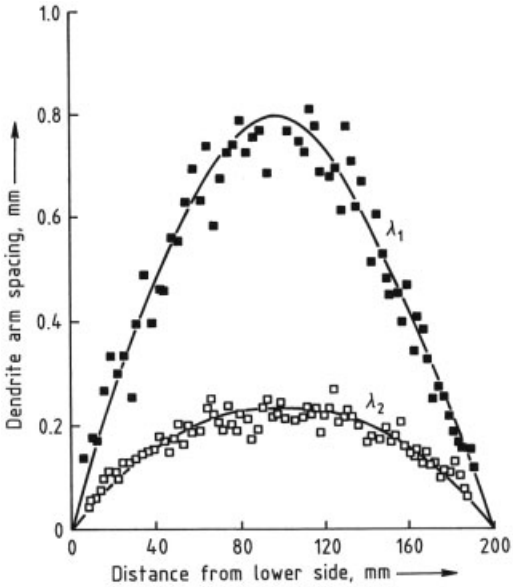

Figure 76. Variation of dendrite arm spacing in a continuously cast slab of steel, grade API X60, 0.14 % C, 1.50 % Mn

By the method of steady-state unidirectional solidification, typical values for the growth variables affecting the dendrite arm spacings can be determined [78], [79]. In the equations:

$$\lambda = c R^m G^n$$

and

$$\lambda = c \theta_f^n$$

R represents the solidification rate, G the temperature gradient, and θ_f the local solidification time.

Figure 77 shows such a correlation for the primary arm spacings in a steel with 0.15 % C and 1.44 % Mn. Further values of these parameters for selected steels are given in Table 11. These show that high-alloy steels have a finer dendritic structure than low-alloy steels. Figure 78 shows that, for four low-alloy steels (I–IV), the data are in good agreement with equations of the form:

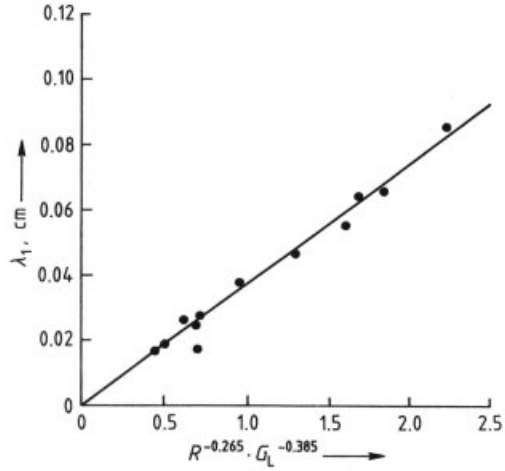

Figure 77. Correlation of primary arm spacing λ_1 with R and G_L for steel II with 0.15 % C, 1.44 % Mn $\lambda_1 = 0.0368\, R^{-0.265}\, G_L^{-0.385}$; λ_1 in cm; R in cm/s; G_L (K/cm) is the temperature gradient at the liquidus temperature

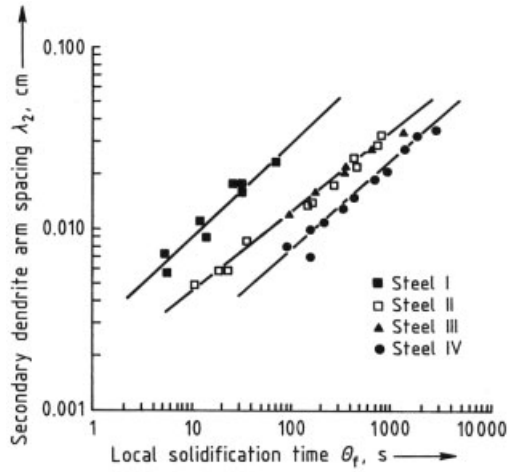

Figure 78. Correlation of secondary arm spacing λ_2 with local solidification time θ_f

$$\lambda_2 = c \theta_f^n .$$

If R, G, and θ_f are known, the segregation spacings can be predicted. Conversely, local cooling rates can be found by measuring the segregation spacings.

If molten metal is undercooled during casting, free crystallites can be generated. This can take place by heterogeneous nucleation, or by the detachment of dendrite arms, which can be

Table 11. Equations for dendrite arm spacings λ (cm) as a function of growth variables R (cm/s), G (°C/cm), and θ_f (s)

Type	Steel composition	Primary arms	Secondary arms	
I	0.09 % C, 1.36 % Mn	$\lambda_1 = 0.0367\ R^{-0.29}\ G^{-0.43}$	$\lambda_2 = 0.01487\ R^{-0.43}\ G^{-0.51}$	$\lambda_2 = 0.00280\ \theta_f^{0.507}$
II	0.15 % C, 1.44 % Mn	$\lambda_1 = 0.0368\ R^{-0.27}\ G^{-0.39}$	$\lambda_2 = 0.00645\ R^{-0.43}\ G^{-0.40}$	$\lambda_2 = 0.00168\ \theta_f^{0.43}$
III	0.59 % C, 1.10 % Mn	$\lambda_1 = 0.1900\ R^{-0.26}\ G^{-0.72}$	$\lambda_2 = 0.01500\ R^{-0.41}\ G^{-0.51}$	$\lambda_2 = 0.00158\ \theta_f^{0.44}$
IV	1.48 % C, 1.14 % Mn	$\lambda_1 = 0.2700\ R^{-0.24}\ G^{-0.72}$	$\lambda_2 = 0.01000\ R^{-0.49}\ G^{-0.51}$	$\lambda_2 = 0.00072\ \theta_f^{0.50}$
V	0.63 % C, 24.9 % Mn	$\lambda_1 = 0.0610\ R^{-0.17}\ G^{-0.36}$	$\lambda_2 = 0.00360\ R^{-0.41}\ G^{-0.37}$	$\lambda_2 = 0.00058\ \theta_f^{0.44}$
VI	0.68 % C, 28.3 % Cr	$\lambda_1 = 0.0280\ R^{-0.25}\ G^{-0.31}$	$\lambda_2 = 0.00320\ R^{-0.41}\ G^{-0.37}$	$\lambda_2 = 0.00052\ \theta_f^{0.39}$

caused by a small degree of superheat, or by inoculation, stirring, jolting, or vibration.

A suspension of crystals in molten steel is shown in Figure 79. The crystallites can redissolve and become globular in form. The globulites settle out during casting and appear, e.g., at the bottom of an ingot or the lower side or central zone of a continuously cast strand. An advantage of this equiaxed structure is that its properties are isotropic.

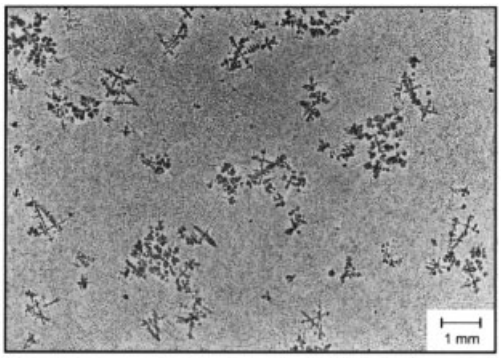

Figure 79. Formation of free crystals in a steel melt continuously cast with undercooling

Interdendritic Microsegregation. The thermodynamic basis for microsegregation is the change in concentration that accompanies a phase change from liquid to solid, and is expressed by the partition ratios k of the solute elements present in the iron. Kinetic factors include enrichment of the dissolved elements in front of the phase boundary, which depends on mass transfer in the liquid, and the extent of diffusional equilibration in the solid.

The degree of concentration enrichment at the phase boundary $C*_S$, in the case of layer crystal formation without diffusion in the solid, is given by:

$$C*_S = k\,C_0\,(1 - f_s)^{k-1}$$

where C_0 is the starting concentration, and f_s the extent of solidification in the dendritic volume element. The degree of solute rejection is, therefore, determined by k. Typical values of k for the more important elements found in steel are given in Table 12. The elements S, P, and C segregate strongly. With carbon, some concentration equilibration takes place by diffusion.

Table 12. Equilibrium partition ratio k of solute element at solid – liquid interface for δ- or γ-iron

	$k^{\delta/l}$	$k^{\gamma/l}$
C	0.20	0.35
Si	0.77	0.52
Mn	0.75	0.75
P	0.13	0.06
S	0.06	0.025

Segregation can be measured over a large area by computer- controlled automatic electron beam microanalysis. For example, Figure 80 shows a three-dimensional relief map of the manganese concentration in a 25 CrMo 4 steel [80]. The scattered high peaks are caused by MnS inclusions.

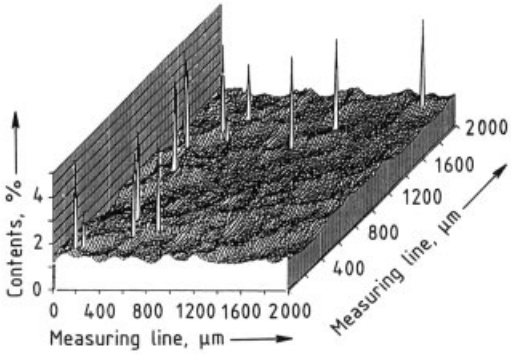

Figure 80. Concentration profile of the microsegregation of Mn in a 2.8 t ingot of steel, grade 25 CrMo 4

Modern microanalytical techniques can produce up to 10^6 point measurements within 24 h. In a typical manganese determination, the area investigated can be 80×80 mm, the number of points measured 600 000, the time required 13 h, and the beam diameter 100 µm [81].

Formation of Macrosegregations. Macrosegregations extend over considerably greater distances than dendrite branches. Typical ranges are between the edge and the center or between the top and the bottom of an ingot. The distinction must be made between enrichment and depletion. A systematic discussion of segregration phenomena, including a theoretical treatment, is given in [82].

Positive A- and V-type segregations in cast ingots result from natural convection due to density differences in the enriched liquid over the cross section of the cast material. Stable flow channels can be formed. Sometimes, V-type segregation can also occur in the center of continuously cast strands with equiaxed solidification structure.

In the two-dimensional solidification of continuously cast billets, segregation can take place along the axis of the core. This is associated with shrinkage, and is caused by the growth of dendrites that periodically form bridges that partially seal off the lower part of the liquid metal. Owing to the volume deficit, a powerful suction force is produced toward the crater end. This type of flow restriction does not occur during the one-dimensional solidification of slabs.

Segregation can also be caused by deformation of the outer shell of a strand, e.g., by nonsteady-state bulging outward and rolling back again [83]. This causes a pumping effect in the remaining liquid metal, and leads to so-called central segregation in the slab. Tensile strains at the heterogeneous solidification front cause segregated internal cracking if critical stresses are exceeded. "Soft" or "hard" reduction, to densify the central zone must, therefore, not lead to overstressing of the material below the solidus.

If there is forced convection at the solidification front, e.g., caused by rimming action during ingot casting of unkilled steel (see Section 3.5.2), or by electromagnetic agitation, a washing effect occurs. A depleted streaky layer is produced. With continuous casting, "white bands" are formed.

Depletion is also caused by sedimentation of suspended crystallites, as these are purer than the melt. The accumulation of a conical heap of crystals at the bottom of the ingot therefore leads to "negative" segregation, whereby the axial concentration of alloying elements increases toward the top of the ingot. This happens, e.g., with heavy forging grade ingots.

Sulfide Precipitation. With modern ladle metallurgical treatment, the sulfur content can be brought below 10 ppm, albeit at some cost, by transferring the sulfur into the slag. Any dissolved residual sulfur can be combined with calcium to form stable calcium sulfide, which has no harmful effects on the properties of the product.

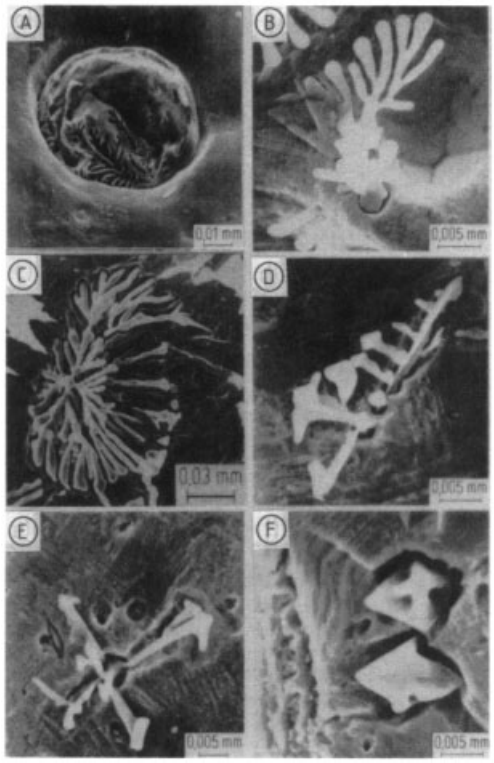

Figure 81. Variation of sulfide morphologies in as-cast steel
A) Type I; B), C) Type II; D), E) Transitional types; F) Type III

However, it is not always necessary to have the lowest possible sulfur content, so desulfurization and calcium treatment of the steel can often be omitted. The sulfur then becomes con-

centrated in the interdendritic spaces, owing to its low k value, and eventually precipitates as MnS. The course of the reaction, and hence the form of the sulfide, are determined by the degree of deoxidation and alloy content of the melt, and the local processes of solute redistribution and mass transfer. The various sulfide types in the cast product are [84], [85]:

Type I. Globular oxysulfides that were originally liquid. Formation mechanism: monotectic degeneration. Steel grades: unkilled, semikilled.

Type II. Coral-shaped chain-like sulfide colonies. Formation mechanism: eutectic. Steel grades: killed, low alloy.

Type III. Angular, crystalline sulfide particles. Formation mechanism: "divorced" eutectic. Steel grades: fully killed, high-alloy, containing C, Si, P, Cr, Zr.

The scanning electron micrographs in Figure 81 show typical examples of these morphologies. Types intermediate between II and III can have faceted shapes with branched dendrites.

The size of the sulfide inclusions depends on the local precipitation rate and hence indirectly on the rate of cooling at the solidus temperature. This correlation is illustrated by the log – log plot in Figure 82 [86].

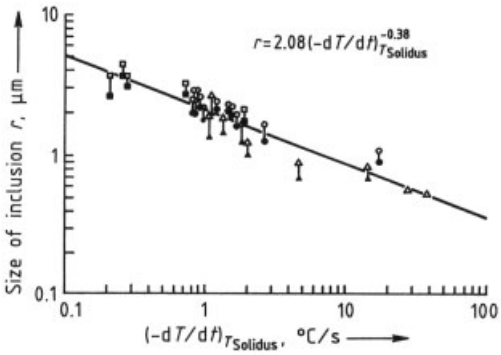

Figure 82. Mean size of sulfide inclusions in steel ingots with 0.55 % C, 2.00 % Mn, 0.009 – 0.039 % S

Formation of Inclusions by Deoxidation. In contrast to sulfide formation during solidification, oxide inclusions are precipitated in the liquid steel at an earlier stage by deoxidation. The amount of dissolved oxygen is reduced to the equilibrium value in a few seconds after the addition of aluminum [87]. As deoxidation proceeds, the inclusions agglomerate or coagulate, and mostly go into the slag. This flotation leads to a reduction in so- called total oxygen.

When aluminum is used for deoxidation, alumina inclusions are formed with a spheroidal or dendritic morphology (Fig. 83) [88]. The coral-like shapes are formed by agglomeration of particles, 2 – 5 µm in size. In steelworks these clusters are referred to as "nests" or "clouds".

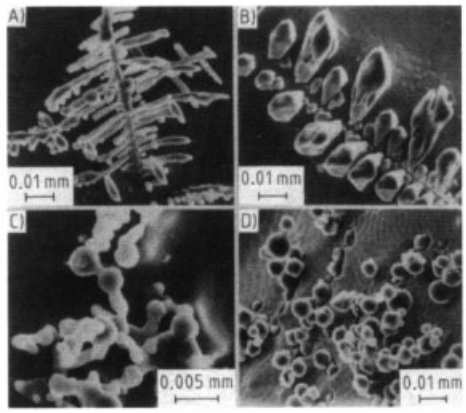

Figure 83. Growth morphologies of alumina inclusions in Al-killed steel

The various growth shapes that can be formed under different supersaturation conditions and activities of the oxygen and aluminum are shown in Figure 84 [89]. From left to right, the oxygen level decreases and the aluminum activity increases. At low oxygen levels, the spheroidal oxides initially formed become covered with radiating, stalky, or compact nitrides.

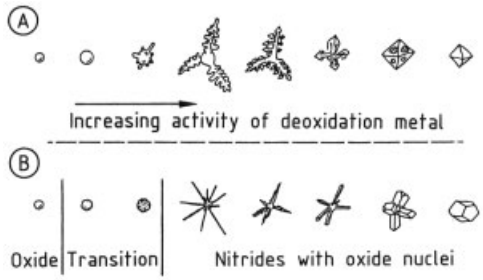

Figure 84. Schematic representation of oxide and nitride morphologies as a function of oxygen level

3.5.2. Ingot Casting

The transition of steel from liquid to solid is mainly achieved by continuous casting. Nevertheless, the older ingot casting process is still used in steelworks all over the world. The molten steel is poured into ingot molds, where it solidifies.

The steel is poured (teemed) from the refractory-lined ladle via a ceramic, closable outlet nozzle. There are two methods of filling the ingot mold. In the top-pouring method, the steel is poured from above directly into the mold, and in the bottom-pouring method, it passes down a central funnel, then through a system of refractory passageways built into a thick steel plate, and passes from below into the ingot molds standing on the plate (Fig. 85) [90].

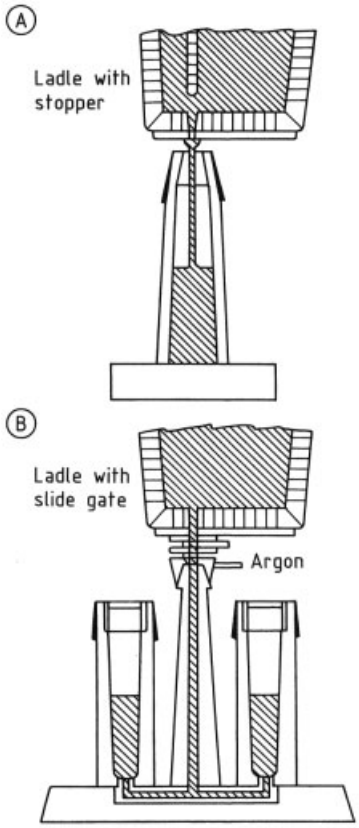

Figure 85. Top pouring (A) and bottom (B) pouring during ingot casting [90]

The molds can have a rectangular, square, circular, or polygonal interior cross section, depending on the cross section of the profile to be produced by rolling. To facilitate removal of the solid ingot, the mold tapers slightly outward, with the wider end either at the bottom (bigend-down molds) or top (big-end-up molds) (Fig. 86) [91]. Ingots can be 0.1 – 40 t for rolled steel, and up to 500 t for forged steel.

Apart from the cross-sectional shape, the type of ingot mold, and the pouring method, ingots are also characterized according to the physical, thermodynamic, and chemical processes that take place during solidification. Several gases are present in solution in liquid steel. The solvent power of the steel decreases as the temperature decreases, and as the steel solidifies. In ingot casting, oxygen is of particular significance for the solidification process. If oxygen is liberated during solidification, it reacts with the carbon in the steel to form carbon monoxide. This can be partially or completely prevented by combining the oxygen with deoxidants such as aluminum, silicon, or titanium.

If the oxygen is allowed to react with the carbon, the steel product is known as rimming steel. Partial deoxidation leads to "semikilled" steel, and complete deoxidation to "killed steel". Deoxidation to semikilled or killed steel is carried out in the ladle. There is also a combined rimming and killed steel, obtained by first allowing the melt to solidify for some time, and then deoxidizing it in the mold, after which it solidifies completely like a killed steel. This produces "ingot-killed steel".

Rimming Steel. As soon as the liquid steel comes into contact with the walls of the mold and the bottom plate, a thin skin of solid steel is formed. This grows rapidly at first, as the cold mold and bottom plate rapidly absorb heat, and, much later, release it to the atmosphere. After a short time, this compact layer of steel becomes detached from the ingot wall as the solid steel contracts. The gap formed reduces the heat transfer rate from the steel to the mold, and hence the rate of loss of heat to the surroundings. Consequently, the steel solidifies more slowly [92–96].

As the steel changes from liquid to solid, the solubility of oxygen decreases, so that it is liberated and reacts with the carbon present to form carbon monoxide. In addition to the dissolved oxygen present at the start of the pouring process, atmospheric oxygen is picked up by the

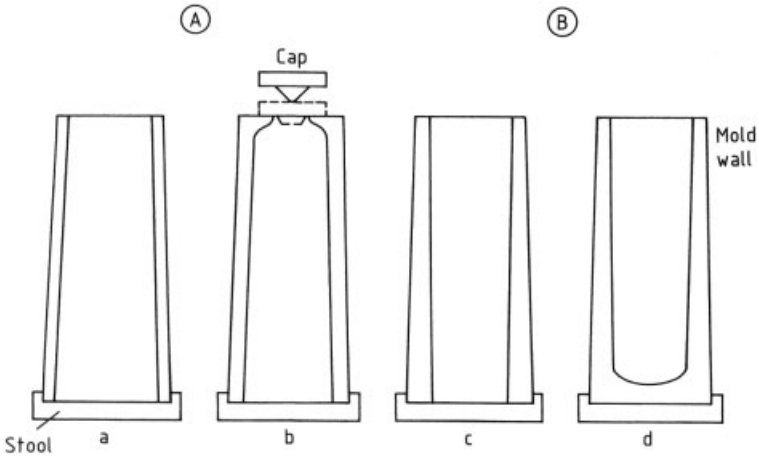

Figure 86. Cross sections (not to scale) of the principal types of ingot mold [91]
A) Big-end-down-molds: open top (a) and bottle top (b); B) Big-end-up-molds: open top (c) and closed top (d)

liquid steel during pouring, and during its subsequent solidification; this also reacts with the carbon in the steel. The CO gas bubbles are partly entrapped between the growing steel dendrites, forming so- called edge blow holes, but most of them ascend in the region of the solidification front, combining with other bubbles on their way up. This leads to the formation of a rim of bubbles at the edge of the steel, and an upward gas flow which causes an upward flow of the liquid steel. Thus, the steel in the ingot flows strongly upward at the solidification front and downward in the center (Fig. 87) [97]. This rimming action has led to the name "rimming steel".

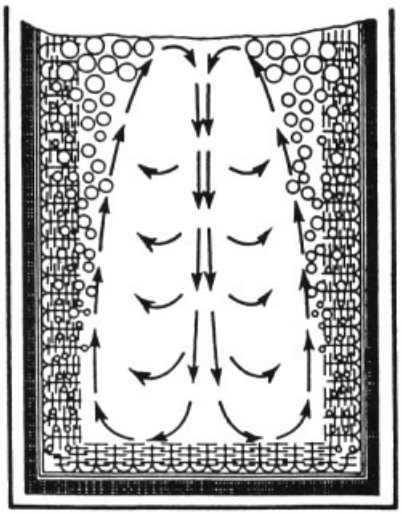

Figure 87. Flow pattern in a rimming steel ingot during solidification [97]

In the production of rimming steel, CO should be produced rapidly from the start of solidification, giving a strong flow. If the flow is weak, CO bubbles remain entrapped by the growing dendrites. This causes excessive formation of a ring of bubbles not far below the outer surface of the ingot, which results in the appearance of surface defects resembling cracks when the ingot is processed in the rolling mill.

Apart from oxygen, other elements that are less soluble in the solid state are released at the solidification front. This solute rejection is known as segregation (see Section 3.5.1). The elements that separate out are carried away by the flow of steel, and are concentrated in the remaining melt. This leads to a rim layer of greater chemical purity and enrichment of the segregated elements in the interior.

After a time, the process of gas release (rimming action) is deliberately stopped by placing a heavy steel cover on top of the ingot, or a steel plate, over which water is sprayed. The steel solidifies on the underside of the cover, and forms a dense cap. In the interior of the ingot, gas formation ceases, owing to the increase in gas pressure, and the flow of steel gradually subsides. In the part that still remains liquid, there are detached crystals of very pure iron, formed by redistribution of solute elements. These suspended solid particles sink to the lower part of the ingot, as their density is higher than that of the remaining liquid metal. This sedimentation results in an

equiaxed zone, with a reduced level of impurity elements in the lower part of the ingot. At the top of the ingot, enrichment of elements occurs, e.g., phosphorus and sulfur, which can have a harmful effect on the properties of the rolled product. This marked concentration of elements at the top of the ingot is called macrosegregation.

The ring of gas bubbles formed during the rimming action, and those formed even after the ingot has been closed off, tend to compensate for the difference between the specific volume of the liquid and that of the solid steel. The walls of these cavities have a clean and bright metallic surface, and weld together during subsequent rolling.

These processes during solidification cause the ingot of rimming steel to have a clean edge zone, a rim of bubbles in the lower region, a chemically pure equiaxed region at the bottom end of the interior, and a strongly segregated upper region (Fig. 88) [98].

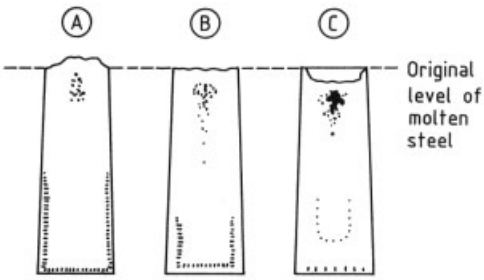

Figure 88. Distribution of bow holes in rimming steel ingots [98]
A) After slight rimming action: low oxygen content; B) After moderate rimming action: medium oxygen content; C) After vigorous rimming action: high oxygen content

Thus, the ingot of rimming steel has a very inhomogeneous structure. Owing to its clean edge region, it can be used in applications that require a good surface. However, it cannot be used for manufacturing components where a large amount of deformation is involved. This is due to its great inhomogeneity and its aging properties, owing to the residual nitrogen. It is also unsuitable if great importance is attached to welding properties [97–110].

Killed Steel. If the dissolved oxygen content in the steel is reduced by deoxidation until it is significantly less than the carbon present, no carbon monoxide gas can form during solidification, and the steel is said to be killed. In killed steel, there is redistribution of those elements whose solubility in solid steel is lower than that in liquid steel. Some of these segregated elements are enclosed between the growing crystals (dendrites). This is termed microsegregation (see Section 3.5.1).

Convection of the molten metal also takes place during solidification of killed steel. However, this is considerably less intense than in rimming steel, and takes place in the opposite flow direction. In front of the solidification interface, free crystallites move downward along with the liquid steel. The solid particles of steel settle at the bottom end of the ingot, while liquid steel flows upward in the center. As the descending crystals are chemically purer than the remaining liquid steel, a depleted conical heap forms at the bottom of the ingot. Here, the concentration of solute elements is lower than it was in the steel before pouring, whereas the concentration of the solute elements in the upper part of the ingot is correspondingly higher. Thus, there is so- called negative segregation in the lower part, and positive segregation in the upper part of the ingot. This is generally referred to as ingot segregation, and becomes more marked as the ingot cross section increases.

As well as ingot segregation, so- called A- or V-type segregation can take place when killed steel ingots solidify (see Section 3.5.1). Unsteady flow through the network of dendrites sometimes leads to the formation of channels, in which enriched liquid metal solidifies. On longitudinal cross sections of the ingot, these channels appear as A- or V-shaped segregation. The intensity of this type of segregation is dependent on the morphology of the solidified ingot. The finer the dendritic structure, the less marked is A- and V-type segregation.

To prevent the uncontrolled formation of shrinkage cavities in the interior of ingots of killed steel, the steel at the upper part of the ingot (the ingot head) is kept in the liquid state as long as possible. The liquid steel can then flow downward to compensate for the shrinkage of the solidifying region, and cavity formation is prevented.

The reservoir of liquid steel is maintained by reducing the heat transfer from the top part of the ingot to the mold by lining the latter with refractory, thermally insulating, and often slightly

exothermic plates. The loss of heat in an upward direction to the atmosphere is reduced by applying an insulating and/or exothermic powder. Figure 89 shows typical longitudinal sections of steel ingots poured into normal big-end-up and big-end-down ingot molds, with and without a feeder head, or "hot top", of this type [91]. In big-end-down ingots, through-solidification occurs earlier in the upper part than in the lower part of the ingot. This produces axial porosity, which under no circumstance should connect with the outside atmosphere, as this leads to oxide formation in the interior, preventing welding during subsequent rolling operations. Unless the top of the ingot is thermally insulated (hot topping), ingots produced in big-end-up molds form shrinkage cavities which extend deep into the ingot, so that the yield of good material is poor. An ingot from a big-end-down mold without a hot top will have axial porosity.

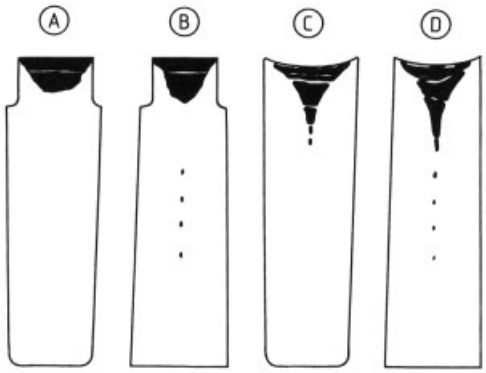

Figure 89. Types of killed ingot [91]
A) Big-end-up, hot-topped; B) Big-end-down, hot-topped; C) Big-end-up, not hot-topped; D) Big-end-down, not hot-topped

When steels are to be used in applications that require an absolute minimum of interior defects, a hot top of adequate size is essential. For example, when heavy ingots are produced for forging, and are cast big-end-up, a feeder head of thermally insulating material is placed on top, with a volume large enough to prevent formation of a shrinkage cavity in the ingot. Segregation of the remaining molten material then takes place in the feeder head inside the hot top. This part is later separated from the good part of the ingot.

As killed steel is mostly used for demanding applications, there must be as few nonmetallic inclusions as possible. Modern ladle metallurgy, combined with a very high standard of pouring practice, have brought great improvements. If any nonmetallic inclusions remain entrapped, however, these consist mainly of oxidic and sulfidic materials. These can originate from the products of deoxidation, from insulating materials, or are formed during solidification by reactions of the segregation elements. Inclusions in killed steel are mainly found entrapped in the region of the cone of sedimented equiaxed material, at the bottom of the ingot [91], [111–119].

Semikilled Steel. In this type of steel, the ratio of oxygen to carbon content is set by controlled deoxidation, so that the ingot at first solidifies, like killed steel. After a time, the carbon and oxygen become enriched in the remaining liquid, so that the solubility product is exceeded and carbon monoxide is formed. Meanwhile, the ingot head has solidified and is therefore dense, so that blow holes are formed in the interior of the ingot, owing to shrinkage (Fig. 90) [91].

Figure 90. Longitudinal section through a semikilled steel ingot [91]

Semikilled steels are not used for applications with high demands on internal purity and defect-free surface properties [91], [120–122].

Ingot killed steel was developed to combine the clean surface of rimming steel with the low degree of macrosegregation and good aging stability of killed steel. The tops of the ingot molds

are provided with insulating tiles, as with killed steel. The steel is poured undeoxidized into the mold until the insulated region contains a few centimeters of steel. A second and possibly a third mold is then filled in the same way. During this time, rimming takes place in the first mold, as well as solidification of an outer shell of steel, several centimeters thick, which consists of very clean unkilled steel, free of inclusions and segregation. The first ingot mold is then completely filled, while simultaneously adding aluminum in granular form to the stream of metal, thereby mixing it into the liquid part of the ingot. Carried by the central downward flow of rimming steel, the aluminum is transported to the lower part of the ingot. The dissolved aluminum reacts with the oxygen to form coarse Al_2O_3 particles, which rapidly ascend into the highest part of the ingot, owing to their low density. The ingot then solidifies like a killed steel. All the ingots from a heat are poured in this piecemeal fashion.

This process produces an ingot with an outer layer of clean rimming steel, ca. 30 mm thick, and a larger inner region of nonaging killed steel with little macrosegregation (Fig. 91) [123]. This steel can be used when a defect-free surface is required on a nonaging material [123].

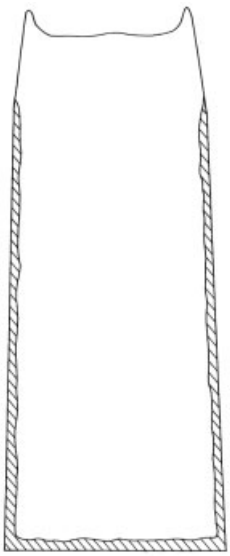

Figure 91. Longitudinal section through an ingot of steel killed in the mold [123]

3.5.3. Continuous Casting

In the historical development of continuous casting (strand casting), a breakthrough was achieved in large-scale production by S. Junghans with the invention of the oscillating mold in 1949. This became widely used during the 1970s [124], [125]. By definition, strand casting is a steady-state process in which a strand is continuously drawn out of a mold [126–130]. After the 1980s, development of the conventional continuous casting process virtually came to an end [131–143]. The rolled products produced from continuously cast strand sections are characterized by their extremely good finish and functional properties, are not subject to any particular limitations, and are suitable for a wide range of applications [144].

3.5.3.1. The Growth of Strand Casting

The growth of strand casting worldwide is illustrated in Figure 92. It has already reached $> 60\,\%$ of the total. Up to the 1980s, the competitiveness of steel producers was measured by the proportion of strand casting used. Today, many steelworks use the process exclusively. In many industrialized countries, the proportion is $> 90\,\%$, but in some regions, it is only $20\,\%$ [144].

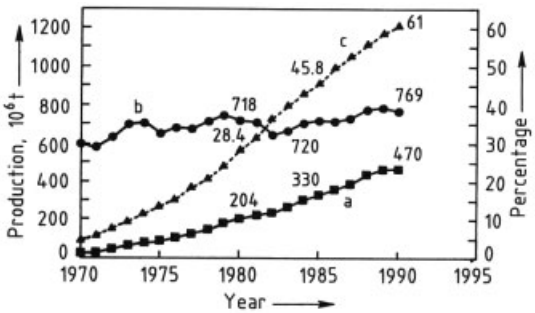

Figure 92. Development of world steel production, CC output and CC ratio 1970 – 1990
a) World CC-production; b) World crude steel production; c) Share of CC-production, %

The strand casting technique replaces individually cast ingots, blooming and slabbing mills, and semifinished mills. The yield of rolled products per tonne molten steel is increased by 10 – 18 % (see Chap. 1). This technological change has resulted in an excess of crude steel

capacity, but has greatly contributed to energy savings.

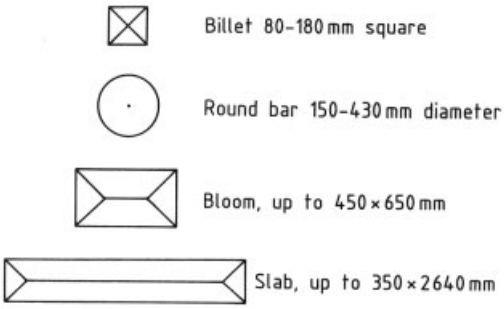

Figure 93. Typical cross sections and dimensions of continuously cast strands

Figure 93 shows the shapes and dimensions of some typical strands, with size ranges. Billets are used to manufacture wire or rod. Round bar products are used as the starting material for seamless tubes. Continuously cast blooms are used in the production of profile steel and rails. Case hardening and heat-treatable steels, cold-heading and cold-extruding steels, spring steels, chain steels, rolling bearing steels, and free cutting steels are routinely produced [145]. Slabs are converted into thick plate for pipeline or vessel construction, or processed into hot-rolled wide strip and products made from this, such as cold-rolled thin sheet, which may be surface treated for automobile construction or packaging applications.

3.5.3.2. Productivity

If strand casting accounts for 100 % of the total output of a steelworks, productivity rates become very important. The capacities of the melting, ladle treatment, and casting process must be matched as precisely as possible. If the mix of steel types produced is variable, it must be possible to react to changes with great flexibility.

Modern strand casting plants are characterized by high availability. The utilization ratio, i.e., the ratio of casting time to time available, is 40 – 90 %. In sequential casting, one melt is cast immediately after another without interruption. The ladles are mounted on a turret which swivels, raises, and lowers the ladle (Fig. 94). The molten steel passes through the following items of plant or process steps: ladle – tundish – mold – secondary cooling zone – straightening and withdrawal machine – runout rollers.

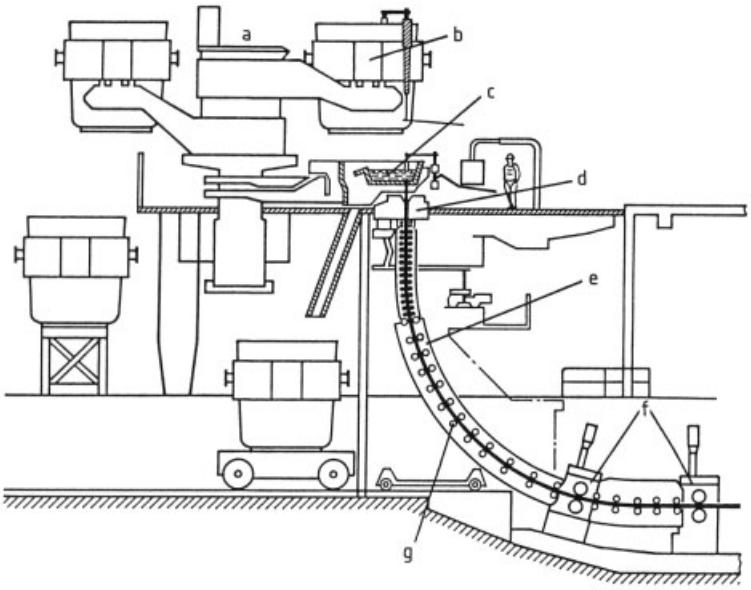

Figure 94. Bloom caster with ladle turret
a) Ladle turret; b) Ladle; c) Tundish; d) Mold; e) Secondary cooling zone; f) Straightening and withdrawal; g) Guide rollers

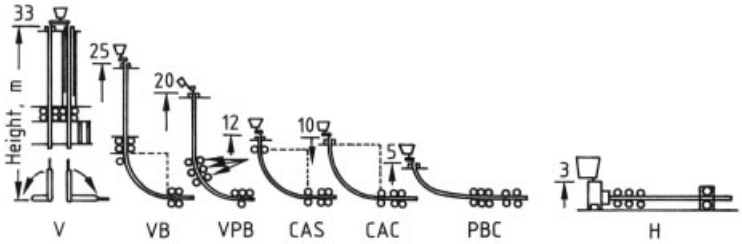

Figure 95. Principal types of continuous casting machines for steel
V = vertical; VB = vertical with bending; VPB = vertical with progressive bending; CAS = circular arc with straight mold; CAC = circular arc with curved mold; PBC = progressive bending with curved mold; H = horizontal

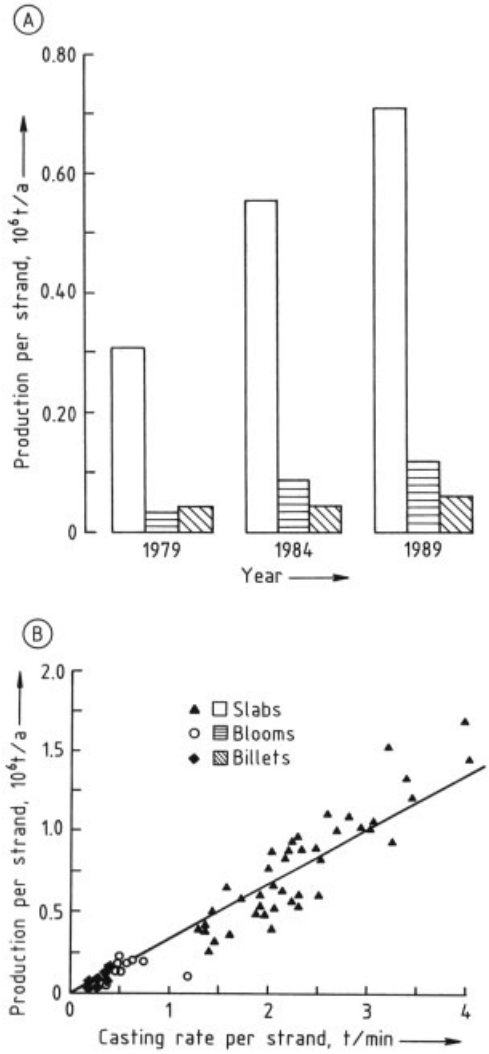

Figure 96. Annual production per strand of European casters (A) as a function of the casting rate per strand for different strand sections (B)

For high productivity, it is also necessary to have a high rate of withdrawal and a long length of the liquid pool. High outputs can only be achieved when bow type casters are used, where the strand, which still has a molten interior, is cast on a circular arc and then straightened to the horizontal position. Figure 95 shows the various types of plant used during the development of the process. Vertical plants (V) have a limited output, owing to the height required. The most usual system is the circular arc plant with a curved mold (CAC), with several straightening points. In a circular arc machine with straight mold (CAS), the strand, with its molten center, is first bent into an arc, and then straightened. The bending and straightening take place progressively. Extremely precise mechanical technology is used to minimize stressing of the cast material.

The production rate for a single strand depends on the format, dimensions, and casting speed (Fig. 96) [144]. For slabs, a rate of 5.5 t/min is possible. This must be matched to the batch time of the basic oxygen furnace. For blooms and billets, the number of strands must be increased to four or six to match this batch time.

Plant availability can be improved by: insertion of the dummy bar from above; design of mold and runout rollers as a quick change unit; flying tundish changeover by means of a swivel device; uninterrupted casting of different steels one after the other without forming a mixed zone (the old strand being joined to the new by a cage-and-anchor joint); and computer-controlled automatic width adjustment of the strand. A high standard of control technology and automation is necessary in the casting plant.

Regular inspection and preventive maintenance ensure problem-free casting.

With slabs and blooms, an average run of nine heats is usual, and with billets, five heats [144]. More flexibility in the production program is provided by a device for longitudinal cutting of slabs, especially for "jumbo" slabs, up to 2640 mm in width. Double and triple casting can also be provided for.

3.5.3.3. Mold Oscillation and Shell Formation

The most important invention in the technical development of strand casting was mold oscillation, which gives the product a good surface finish. Figure 97 shows how the molten steel from the tundish passes through a submerged entry nozzle into a slab mold. The steel flows sideways during slab casting, but vertically downward with square or round mold shapes. The pouring temperature is controlled within a very narrow range.

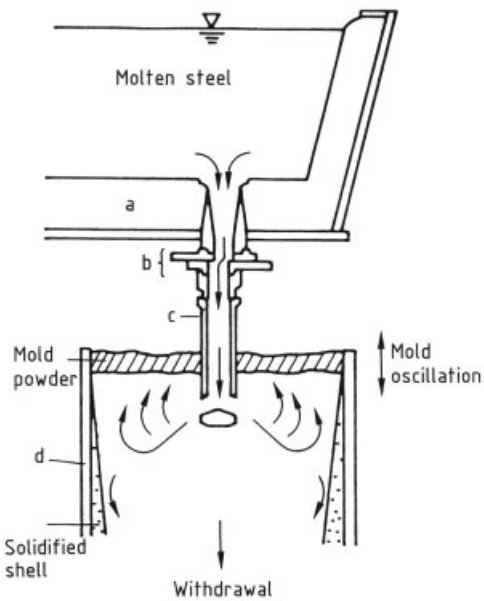

Figure 97. Steel flow from tundish into an oscillating slab mold
a) Tundish; b) Sliding gate; c) Submerged entry nozzle; d) Copper mold plate

The oscillation of the mold is illustrated in Figure 98. The movement is usually sinusoidal. Stroke and frequency are so adjusted that the mold briefly overtakes the strand during the downward movement. During this so-called negative strip time, the skin of the casting is briefly upset. Any surface damage is completely healed. The negative strip time t_N (an important control parameter) is given by the following formula:

$$t_N = \frac{1}{\pi f} \arccos - \frac{v_g}{\pi f h}$$

where t_N is in s, the frequency f is in s^{-1}, the casting speed v_g in mm/s, and the stroke h in mm. If $t_N \to 0$, tearing of the shell occurs, and if it is too large (> 0.2 s) deep oscillation marks are produced, and lubrication in the gap between the mold and the strand is impaired. When changes to the casting rate occur during unsteady phases, such as the start of casting or ladle changes, the frequency is adjusted. Amplitudes are kept constant at 2 – 6 mm, but frequencies vary around $6\,s^{-1}$. A curved mold oscillates precisely along the arc.

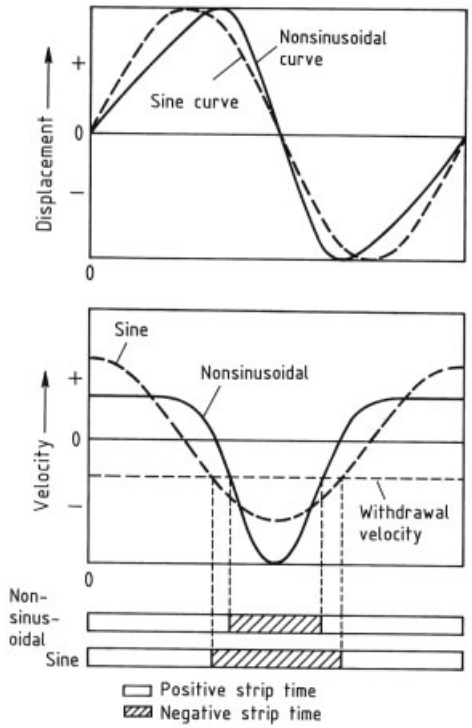

Figure 98. Mold oscillation with definition of positive and negative strip times

Molds are usually 700 – 900 mm long, and are made of high-strength, copper-based mate-

rials. To maximize operating life, the surface is coated with chrome (60 – 80 µm) or nickel (0.5 – 4.0 mm). For small cross sections, tube molds are usual, and for large cross sections, plates are mounted on stable steel frames. To accomodate the contraction of the steel, the molds taper inward. For plants with high casting rates for billets and blooms, very close geometrical tolerances must be adhered to. For sensitive steels with high sulfur content or wide solidification range, the casting speed is reduced.

3.5.3.4. Control of Molten Steel Level and Gap Lubrication

For problem-free strand withdrawal from the mold, the cycle must be optimized, the meniscus level of molten steel kept constant, and an appropriate casting powder must be used to lubricate the gap.

The flow of steel is controlled by sliding gates (Fig. 97) or stoppers. The steel level is detected with the aid of radioactive sources, ^{60}Co or ^{137}Cs, or eddy current detectors, and is controlled within ± 5 mm, or even ± 2 mm, depending on the sophistication of the control equipment, the format of the strands, and the casting rate.

The casting powders used must be appropriate for the types of steel and the casting conditions. On contact with the surface of the steel, they should melt promptly at 1100 – 1200 °C, and form a lubricating film between the strand and the mold. Penetration into the gap plays a key role in the formation of the steel surface for high withdrawal speeds [145].

Casting powders consist of CaO, SiO_2, and $Al_2O_3 \cdot CaF_2$, Na_2O, or Li_2O are added to lower the viscosity. Carbon particles in the form of powdered coke or soot retard the melting. The basicity of the slags is given by:

$$CaO/SiO_2 = 0.8 - 1.2$$

The amount of casting powder required depends on the oscillation cycle, the casting speed, the geometry of the gap, and the viscosity. Slabs are cast by means of low-viscosity slags: 0.1 – 0.3 Pa · s (1300 °C). For blooms, viscosities must be somewhat higher: 0.4 – 1.5 Pa · s (1300 °C).

Casting powders protect the molten steel from radiation heat loss, prevent contact with atmospheric oxygen, and absorb oxide inclusions that rise to the surface, especially Al_2O_3. Powders and granules (nowadays preferred) can be added automatically in the fluidized state.

3.5.3.5. Support Roller Apron and Secondary Cooling Zone

At the mold exit, the skin of the strand is sometimes quite thin, e.g., 10 – 12 mm, depending on the withdrawal velocity, and is ductile and deformable under the conditions of high temperature and ferrostatic pressure. Whereas billets and round bars are dimensionally stable, the wide faces of slabs and blooms tend to bulge, and must be supported by rollers [146].

To prevent excessive thermal stress on the rollers, they are cooled internally. The helirollers contain a spiral channel for circulating the water, and are stable and long-lasting [138]. Divided rollers with intermediate bearings are of smaller diameter and can accomodate narrower gaps between them. Where bending and straightening of the strand is carried out, deviation of the rollers from the ideal line must not exceed 1 mm. Loose-fitting and worn rollers must not cause this limit to be exceeded.

In the secondary cooling zone, the strand is cooled with water sprays. There are various jet systems to provide different flow rates of water impingement. In the extreme, the surface can be sprayed "black". In the early days, large quantities of water were used for fear of break-outs of molten steel. Since then, the water rate has been limited to the minimum required for the necessary cooling intensity. Air/water mixtures are used to achieve gentle and uniform cooling. Specific consumptions of spray water are 0.7 L per kilogram of steel for spray cooling, and 0.1 – 0.2 L/kg for air-mist cooling.

3.5.3.6. Automation

Most functions of a casting machine are automatically controlled. Figure 99 gives an overview of the individual processes [132]. Secondary cooling is dynamically controlled. The

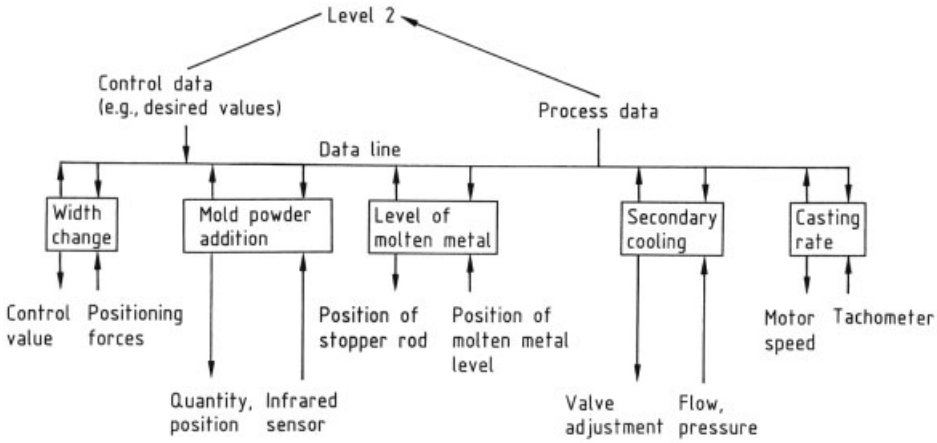

Figure 99. Instrumentation and automatic process control of a continuous casting machine

spray water is adjusted according to the casting speed to prevent undercooling. Cooling water circulation, hydraulic, and lubrication systems are computer controlled. The rate of steel casting is measured by weighing, and cutting of the strand into lengths is optimized with respect to the heads and tail pieces.

3.5.3.7. Quality

The quality of continuously cast steel as a semifinished material for rolled steel production can be assessed by four criteria:

1) Homogeneity of chemical composition
2) Uniformity of steel structure
3) Extremely high purity
4) Defect-free surface

The above requirements are met if the process is trouble free. Solidification can be controlled to be columnar or equiaxed by means of temperature control and the stirring effect of electromagnetic fields, enabling the blooms and billets to have isotropic properties throughout. With slabs, soft reduction is carried out by conical compression of the center.

Contact of air with the steel is prevented by shrouding systems and by covering the molten metal in tundish and mold with powder. Carryover of slag along with the steel from the casting ladle is detected and prevented. The flow of steel in the tundish is controlled such that oxide inclusions separate out. It is also necessary to use stable refractory materials and refined pouring techniques to ensure a high-purity product. For high-output plants, a vertical part of 2 – 3 m in height enhances the flotation of inclusions and gas bubbles.

Materials technology has elucidated the high-temperature ductility of steel from its molten state to the $\gamma - \alpha$ transformation, and mechanical engineering has provided the necessary data on permissible material stresses during casting. By progressive bending and straightening, and suitable roller distances, tensile strain at the solidification front can be kept $< 0.3\,\%$. Critical temperature zones susceptible to surface cracking are avoided by appropriate cooling. Harmful elements, e.g., N and S, are removed from the steel, or chemically combined to form stable compounds.

3.5.3.8. Horizontal Strand Casting

Horizontal strand casting has replaced ingot mold casting for several types of steel which have hitherto been regarded as unsuitable for strand casting [137], [147]. The capital cost of a horizontal strand casting plant is small. The tundish and mold form a single unit, giving complete exclusion of air. The horizontal strand requires neither bending nor straightening.

The special feature of the process is the creation of an "artificial meniscus" at the break ring, where the skin of the casting first forms

(Fig. 100) [148]. The strand is withdrawn stepwise, and the short frozen sections are welded together. The cycle comprises three phases: pulling out, pushing back, and pause. The pushing-back stage allows healing to occur. The design of the machinery must enable the large mass of the strand to be moved with very high precision with respect to both distance and time [149], [150]. Typical round bar dimensions are 150 – 350 mm in diameter.

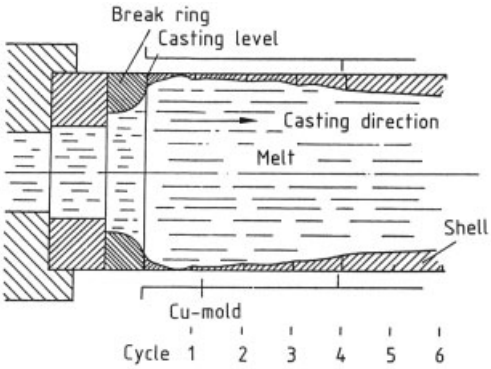

Figure 100. Creating an artificial meniscus during solidification in a horizontal casting mold

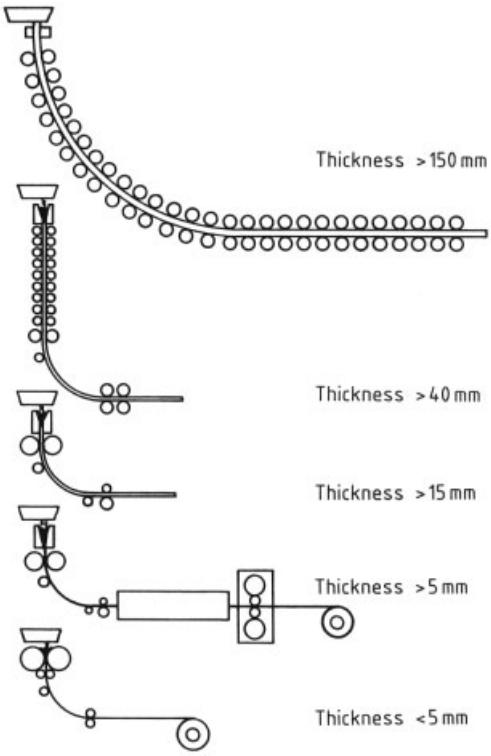

Figure 101. New processes for the production of thin slabs and hot strip

The breaking ring is made of BN, and is not wettable. Useful lives for this item of 4 – 7 h have been reported. The formation of the solidified skin is precisely controlled with the aid of thermocouples. The casting parameters and mold taper must be precisely matched. Nevertheless, the process is flexible with regard to the formats and steel types. It is of interest for small-scale operations, e.g., 3000 t/month in 100 different grades of mild, stainless, or high-temperature steel [132], [149].

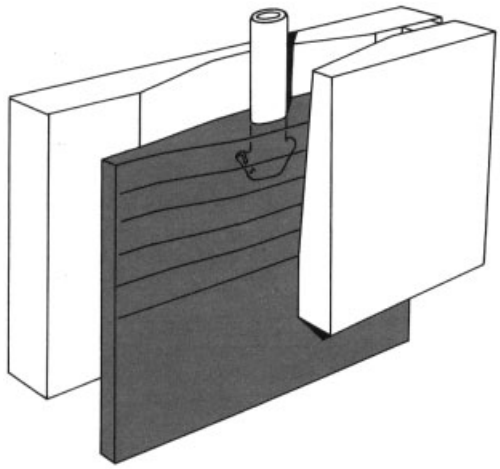

Figure 102. Schematic representation of casting mold for thin slabs

3.5.3.9. Hot Charging or Direct Rolling

Not only is conventional strand casting a fully developed process, guaranteeing a high productive capacity, but it offers the opportunity for energy savings by integration with the rolling mill, i.e., utilizing the heat of the strand by hot charging or direct rolling [139]. For this, the operation of the rolling mill must synchronize with the flow of material from the melting shop, the secondary metallurgical processes, and the strand casting plant. The temperature of the cast product is not allowed to fall below 1200 °C. The surface and the interior must be of high quality.

3.5.3.10. Casting Thin Slabs

Under the heading "near net shape casting," the development of casting technology since 1983

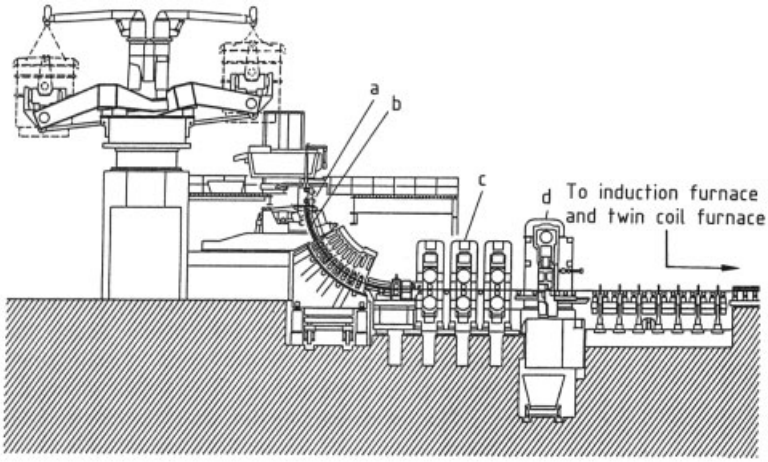

Figure 103. Caster for in-line strip production (ISP)
a) Vertical bow-type mold; b) Soft reduction; c) High reduction (three stands); d) Pendulum shear

has aimed to match the thickness of continuously cast slabs more closely to that of the end product [151], [152], i.e., to reduce the amount of metal forming required, and hence to economize on plant equipment [153–155]. Processes for the production of thin slabs or strip are illustrated schematically in Figure 101.

At the heart of thin slab technology is the funnel-shaped mold into which a narrow submerged entry nozzle with side exit openings is inserted (Fig. 102) [153]. The shell of the strand is smoothly deformed in a controlled manner until it leaves the mold, and the gentle contours keep the bending stress small. The strand leaves the mold with plane parallel surfaces in its wide dimension. Thin slab technology is otherwise similar to conventional strand casting, e.g., control of the level of molten steel, supporting rollers, secondary cooling, etc.

3.5.3.11. New Processes for Hot Steel Strip Production

For integrated "mini-steelworks", various processes have been devised for the production of hot steel strip, based on thin slab casting technology. It is generally agreed that a direct link between casting and rolling to produce a finished product in an in-line, multistand rolling mill is impossible, because the casting and rolling speeds cannot be matched. A buffer can be provided between these processes by operating them in two stages.

A compact strip production (CSP) plant consists of a casting machine, a conveying and linking system designed as a roller hearth furnace, and a hot finish rolling mill [154], [155]. The casting machine is vertical, with bending facility. The casting format is 50×1600 mm; the casting speed is 5 m/min. The strand is cut into 45 m lengths, which are loaded at nearly 1100 °C into an annealing furnace, 162 m long. The final rolling temperature is 880 °C.

In-line strip production (ISP) is shown in Figure 103. The thin slab, with a liquid core, is softly reduced in thickness from 60 to 43 mm immediately after the mold [142], [156], [157]. After through-solidification, the strand passes through three graded stands of high-reduction rolling units. Its thickness is reduced to 15 – 25 mm, and it is made into 28 t coils. Before coiling, it can be reheated by inductive heating to a temperature of 1150 °C. The coils are kept hot in furnaces, ready for the finish rolling mills. The hot steel strip has a final thickness of 1.8 –12.5 mm and can be intensively cooled.

The casting – pressing – rolling (CPR) process (Fig. 104), combines a conventional oscillating thin slab mold of the type described above, with a powerful pair of compression rollers to reduce the thickness of the strand, which has a liquid interior, to 15 – 25 mm in a single step

[158]. The hot strand is then rolled to its final dimensions at 1300 °C in a four-high stand.

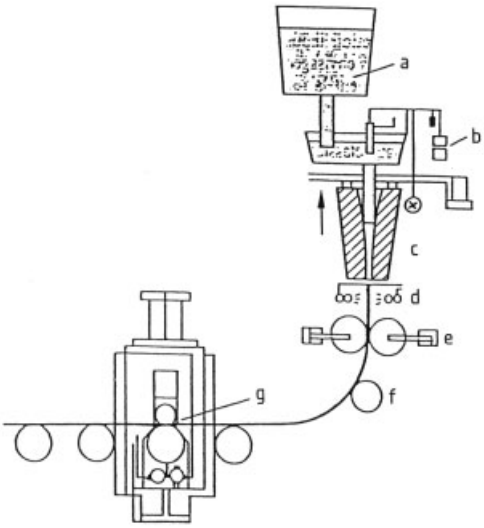

Figure 104. Casting – press – rolling process
a) Ladle; b) Tundish; c) Mold; d) Descaling; e) Press rollers; f) Bending roller; g) Straightening unit

3.5.3.12. Rationalization

The new processes of hot steel strip production may replace some of the classical "thick slab – hot strip steel mills", but not completely [155]. Hot strip steel produced by the standard method can be divided into: mild steel for cold forming and cold rolling (76.2 %); weldable construction steels (10.5 %); high-strength carbon steels (1.7 %); microalloyed weldable fine grain and tube steels (6.4 %); stainless steels (3.3 %); and silicon steels for electric sheet (1.9 %).

Assuming sufficiently good surface properties of the mild steels produced by the new casting – rolling processes, their suitability for cold forming and cold rolling is beyond question [155]. Of the total amount of hot steel strip listed above, 75 % could be produced more economically. A considerable fraction of the total production could be transferred from large integrated steelworks to mini-steelworks. The forming properties of 21 % of this total are difficult or critical. Further development work is needed; in particular, the alloy compositions of these materials must be matched to the new casting – rolling processes.

3.5.4. Consumable Electrode Remelting Processes

Remelting processes are used for upgrading the quality of steels and Fe-, Co-, and Ni-based superalloys; conventionally made metal bars are transformed, by simultaneous remelting and resolidification, into new materials with superior properties.

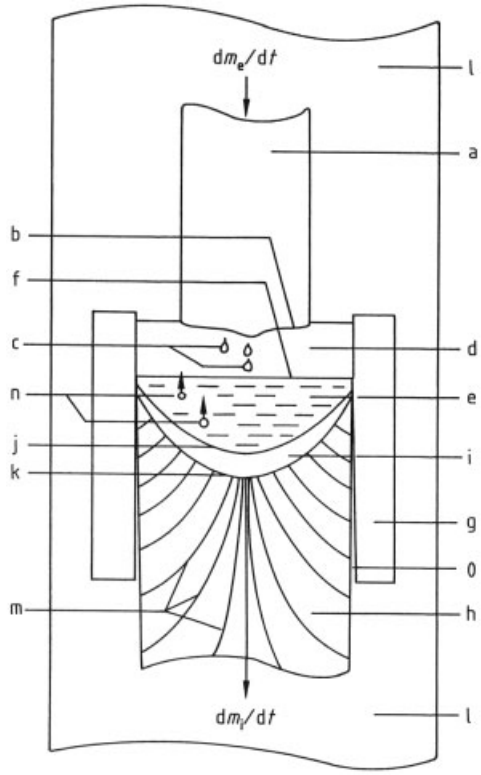

Figure 105. Consumable electrode remelting
a) Electrode = metal to be remelted; b) Electrode surface heated above *mp*; c) Metal droplets; d) Remelting environment (vacuum, liquid slag, plasma); e) Liquid metal pool; f) Surface of liquid metal pool; g) Crystallizer (water-cooled mold); h) Remelted ingot; i) Mushy zone; j) Liquidus isotherm; k) Solidus isotherm; l) External environment (vacuum, air, inert gas); m) Solidification direction; n) Inclusion in liquid pool

3.5.4.1. Remelting Processes

Figure 105 outlines the remelting process for steel and superalloys. The metal to be remelted is

first produced by primary melting, secondary refining, and casting processes, including vacuum induction (VIM) and plasma melting and casting. It is introduced into the remelting process as a so-called "electrode", which is normally shaped by direct ingot casting, with or without subsequent rolling or forging, or by continuous casting. Compound electrodes, such as bundles of bars or continuously cast billets may be used, as well as several individual electrodes, to be remelted simultaneously. Addition of deoxidizers or alloying elements (e.g., Al wire, Ti strip, high-nitrogen alloy compacts) to an electrode, to form a compound electrode, is common.

The electrode is heated at one end above the liquidus temperature so that liquid metal droplets are formed. These fall through the reactive environment of the remelting zone to the liquid metal pool, the surface of which is also heated above the liquidus temperature. The liquid metal pool is confined by a heat-extracting crystallizer, usually a water-cooled copper (or steel) mold. Heat extraction causes the liquid pool to solidify slowly into a new ingot, the "remelted ingot". A "mushy zone" is maintained between the pool and the fully solidified ingot.

The remelting rate dm_e/dt of the electrode(s) practically equals the mold-filling rate dm_i/dt of the remelted ingot. This rate is one or two orders of magnitude smaller than in conventional ingot or continuous casting. A typical value for round remelted ingots of diameter D_i is:

$$dm_i/dt = \left(1000\,\mathrm{kg\,h^{-1}\,m^{-1}}\right) D_i$$

A typical ingot for the rolling mill of mass 3000 kg and diameter 0.5 m thus needs a remelting time of 6 h, a big forging ingot (180 000 kg, 2.4 m) needs 75 h; the time is further increased by the necessary "hot topping" at the end of remelting.

3.5.4.2. General Aims of Remelting

Ingot Structure. One aim of remelting is improvement of ingot structure and homogenization of longitudinal and transverse mechanical properties.

Conventionally cast ingots and strands have a predominant direction of solidification, from the surface to the center, with consequent voids and center segregations. To achieve homogeneous and acceptable mechanical properties, high reduction during hot forming is necessary, which limits the final dimensions of the finished products. Because of their low solidification rate, remelted ingots solidify mainly from the bottom to the top (Fig. 105), which presents center voids and segregations and leads to an extremely sound and homogeneous ingot structure. Less reduction is necessary to achieve a sound product, which increases the final dimensions attainable with a given hot forming device (rolling mill, forge).

Cleanliness. The second aim of remelting is the improvement of metal cleanliness, two mechanisms being effective:

1) The heated metal surfaces and the droplets react with the reactive environment. If the reactive environment is high vacuum, degassing and vacuum deoxidation reactions take place, lowering the content of hydrogen, oxides, nitrides, and volatile metallic residual elements (Cu, Pb, Zn, etc.); if it is reactive slag, complex deoxidation, denitrogenation, and desulfurization take place, together with retention of inclusions from the electrode in the slag.
2) Large inclusions, finding their way into the liquid pool, have time to rise slowly to the top, where they are eliminated by various mechanisms. By reaction with the liquid metal, many of the rising inclusions are reduced in size; below a certain particle diameter rising stops, and these inclusions are retained in the solidifying matrix. The remelted ingot has not only a lower total inclusion content, but the maximum inclusion size of the remaining inclusions is much lower than in the electrode material.

Other Aims. Through the introduction of remelting, these processes have often had the special aim of making products previously not possible by conventional means. An important application is the production of high-nitrogen steels, with high-pressure nitrogen (e.g., 5 MPa) as the remelting environment. Another application is production of sound hollow ingots of steels which are difficult to hot-work, by remelting multiple electrodes into a common annular mold.

3.5.4.3. Remelting Apparatus

Table 13 lists the main processes. Most units for general applications are ESR units, followed by VAR units; EBR, PESR, VADER, and PAR are used for special applications. Details of remelting apparatus construction (electrode feeding; round, square, rectangular, and hollow molds; ingot withdrawal; slags; alloy addition devices; automation, etc.) and other processes are given in the references [159–166].

3.5.4.4. Results and Applications

The high cleanliness of remelted steels makes them useful for surgical implants, highly polishable pressing sheets for laminate production, etc. The homogeneity of mechanical properties is good for thick products of ultrastrength structural steels, tool steels, etc. The soundness of the ingot allows large forgings, up to 100 t or more, to be made from a much smaller ingot, compared with the conventional route. Some high-temperature alloys exhibit improved forgeability. High-nitrogen, fully austenitic PESR steels are used for production of large, nonmagnetic retaining rings for electrical power generators. Further applications are listed in [159–166].

During the last decade, steady improvements in conventional melting and casting technologies regarding cleanliness and soundness have taken over many applications; the more economic route of melting, secondary refining, and continuous casting is favored over remelting processes. For sound ingots of up to a few tonnes, the powder metallurgy route has become a competing, albeit more expensive process.

3.5.5. Cast Steel and Cast Iron

3.5.5.1. Cast Steel

Many grades of steel, of various composition, are suitable for casting. The carbon content can be as high as 2 %, but is usually 0.02 – 0.4 %. Casting consists of pouring into refractory molds, and allowing to solidify.

In the production of steel castings, all types of molds and casting processes are used, including hand- and machine-made molds, shell molds, and ceramic molds. Other processes include precision casting by the lost wax process, the full mold process, and centrifugal casting. Castable steel is melted by methods similar to those for rolling and forging steel, mostly in electric arc furnaces lined with a basic refractory, and, less commonly, in induction-heated, crucible furnaces. The processes of secondary metallurgy in oxygen/argon converters or specially designed gas-purged ladles are increasingly important.

In Germany, of all the crude steel produced, 0.6 % is cast. The proportion is similar in other industrialized countries.

Cast Steel for General Applications. These materials are used for components intended for use between − 10 and 300 °C, which are subject to moderate dynamic and impact stresses, e.g., parts for general machinery, wheels and wheel bearings, rolling equipment, and ships' anchors. The most common material in this group is mild steel, though some low-alloy steels are also used. Steels with a low carbon content have good welding properties, which has ensured a wide range of applications for steel castings. It is common to improve the welding properties by reducing the carbon content, and to compensate for this by increasing the manganese content.

Figure 106 shows the temperature dependence of the mechanical properties of unalloyed cast steel.

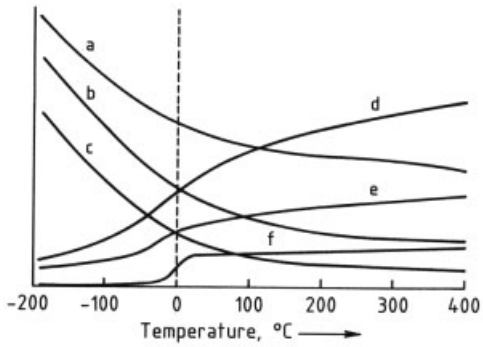

Figure 106. Mechanical properties of unalloyed cast steel as a function of temperature
a) Tensile strength; b) 0.2 % yield strength; c) Flexural fatigue strength; d) Constriction; e) Elongation; f) Notched impact strength

Table 13. Consumable electrode remelting processes

Process	Name	Energy input	Remelting environment	External environment	References
ESR	electro-slag remelting	high- current a.c. (or d.c.) resistance heating of liquid slag	liquid slag of $CaF_2 - CaO - Al_2O_3$ or similar type, heated above metal *mp*	air or inert gas	[159], [162], [163]
VAR	vacuum arc remelting	high- current d.c. arc between electrode and pool surface	low-vacuum plasma	high vacuum	[160–163]
EBR	electron-beam remelting	high-voltage electron beams hitting electrode tip and pool surface	high-vacuum plasma	high vacuum	[160], [162], [164]
PESR	pressure electro-slag remelting	see ESR	see ESR	high-pressure inert gas, mostly N_2 plus additions	[165]
VADER	vacuum arc doubleelectrode remelting	high- current d.c. arc between two electrodes	vacuum plasma	vacuum	[161–163]
PAR	plasma arc remelting	plasma beams from d.c. or a.c. plasma burner(s)	plasma gas	vacuum, normal or high-pressure inert or reactive gas	[162], [166]

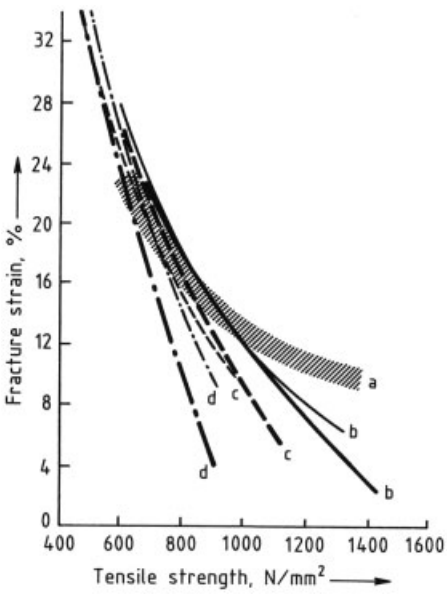

Figure 107. Correlation between fracture strain and tensile strength of tempered cast steel compared with rolled and forged steels
a) Steel values from DIN 17200; b) Tempered structure; c) Mixed structure; d) Ferritic – pearlitic structure —— = Cast steel valves from Steel Castings Handbook; —— = Experimental cast steel valves

Cast Steel for Special Applications. Cast steel was formerly at a disadvantage compared with forged and rolled steel for welded construction, in that high-strength grades with adequate welding properties were unavailable. Metallurgical advances in steel production, especially secondary metallurgy, have enabled the production of steels which, after casting and heat treatment have tensile strengths of ca. 800 N/mm^2. They are thus equivalent to rolled and forged steels, and can be welded to them (Fig. 107).

With these properties, the high-strength weldable and castable grades of steel now available are suitable for welded designs in many areas of engineering. Grades suitable for turbines and steam power stations can be used to make castings up to 100 t.

Alloying costs are not high, but they need to have a carbon content of ca. 0.2 % for good welding properties. To ensure that components of all types will have the properties desired, including welding properties, martensitic and lower bainitic structures must be produced, for uniform strength and toughness.

For plain carbon steel castings, the limit of the wall thickness for through-tempering is ca. 20 mm. By the use of suitable alloying elements, this can be extended to ca. 500 mm, although the welding properties are somewhat affected.

Hot-strength, low-temperaure, and stainless grades of cast steel, including austenitic and duplex steels, are produced in a similar way to the rolling and forging steels (see Chap. 6.2).

The tendency of low-temperature nickel steels to give a coarse-grained structure on solidification makes them less suitable for casting.

It is, therefore, recommended that Cr – Mo or Ni – Cr – Mo alloy steels should be used, if possible.

Typical grades of steel for casting have heat- and abrasion-resistance properties which cannot be matched by rolling and forging steels.

Cast steel is regarded as a high-temperature steel if it effectively resists the scaling effect of gases at > 600 °C. Heat-resistant steels can be divided into ferritic and austenitic grades, and alloys based on nickel and cobalt. Ferritic steels are alloys containing 7 – 28 % Cr and 1.7 % Si, and austenitic steels contain 18 – 30 % Cr and 10 – 37 % Ni. High-temperature steels (Fig. 108) are used for components that are highly stressed, both thermally and mechanically, and are subjected, either continuously or intermittently, to corrosive gases at ca. 600 –1150 °C. These materials have made the continuous operation of industrial furnaces an economic possibility. Other areas of use include ore treatment (roasting furnaces), and the cement, petroleum, and petrochemical industries. Heat-resistant cast steel is used for valve cages, and combustion chambers in diesel engines.

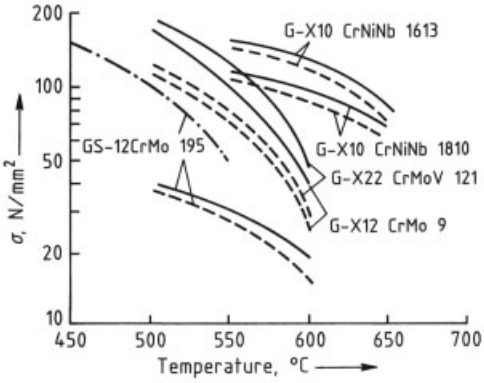

Figure 108. High-temperature cast steel for working temperatures above 540 °C
—·= $\sigma_{B/100\,000}$; – – – = $\sigma_{1/100\,000}$ –·–= DVM (Deutscher Verband für Materialprüfung der Technik) creep limit

Cast hard manganese steel (Fig. 109), which contains 1.0 – 1.4 % C and 12 % Mn, was first described in 1888, and has retained its importance up to the present day. In order to develop its maximum abrasion-resistance properties, a cold-hardening process is required, e.g., by impact or pressure working, which can increase the hardness of the abrasion-resistant surfaces from 250 – 300 to 500 HB (Brinell hardness scale). Good toughness properties can be produced by a solution annealing process at 1050 °C, with water quenching.

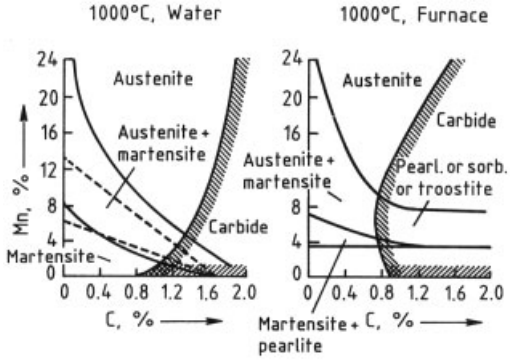

Figure 109. Structure of cast hard manganese steel as a function of cooling rate

Hard manganese steel is unsatisfactory under conditions of abrasive wear without impact and pressure work hardening, as under these conditions no cold hardening occurs.

Tempering steels are used where components and workpieces undergo mainly abrasion, but are also subjected to impact stress.

Abrasion-resistant cast hard chrome steels have the highest abrasion resistance of all cast steels. They contain 2.5 – 3.5 % C and 15 – 27 % Cr, and attain their highest abrasion resistance after hardening at 900 – 1050 °C, with either accelerated or static air quenching. This group has the lowest toughness of all the abrasion-resistant materials. Components made of this material must not be subjected to transverse stress. Hard chromium steel is used under conditions of predominantly frictional abrasion, and where impact and pressure stresses are small.

3.5.5.2. Cast Iron

Cast iron includes iron – carbon alloys with carbon content ca. 2.8 – 4 %. The carbon is for the most part not chemically bonded to the iron, but is present in elemental form.

Gray Iron. Cast iron with flake graphite, also known as gray iron, is an iron-based material in which the carbon is nearly all in the form of microscopic graphite flakes with a ferritic –pearlitic structure (Fig. 110). In Germany,

this material falls under the DIN 1691 standard, which specifies tensile strengths of 150 – 350 N/mm^2. In the United Kingdom, BS 1452 applies, and in the United States, ASTM A 48.

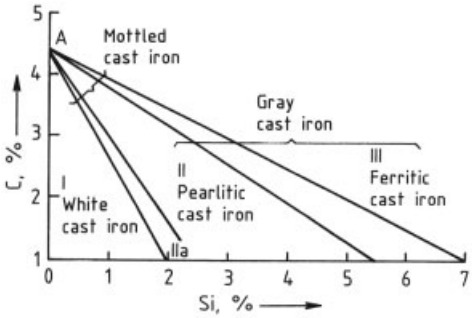

Figure 110. Structure zones of cast iron

Formation of the graphite and its primary structure, and hence the strength and hardness of the product, depend very much on the cooling and solidification rate, and hence on the wall thickness. However, heat treatment of gray iron is not normally practicable, so this relationship is less important than the correct choice of chemical composition. For gray iron components, the design can only specify the wall thicknesses and the required strength – the choice of chemical composition must be left to the foundry.

Cast iron is poured at ca. 1300 – 1450 °C. When molten, it can completely fill complicated thin-walled molds, reproducing pattern details exactly. Other important characteristics include its property of damping vibration, its good resistance to corrosion by weathering, and its good machining properties.

The amount of cast iron exceeds that of all other cast materials; ca. 40 % is used in the automobile industry. Other applications include mechanical engineering, the building industry (radiators, boilers, sanitary ware, pipes), chemical plant, shipbuilding, and mining. The former demand by steelworks for cast iron ingot molds has been greatly diminished by the introduction of the continuous casting process.

Spheroidal Graphite (SG) Cast Iron. In spheroidal graphite, also known as ductile iron or nodular iron, the free graphite in the ferritic – pearlitic matrix is almost completely in spheroidal form. This is achieved by treating the melt with magnesium. The formation of spheroidal graphite gives very good ductility, unlike the brittle properties of gray iron.

This material, first produced in 1948, is now the subject of the German DIN 1693, BS 2789 in the United Kingdom, and ASTM A 536 in the United States. Its hardness can be increased by heat treatment and tempering. Its mechanical and physical properties place it between lamellar cast iron and cast steel, though it more closely resembles cast steel.

A large consumer of SG iron is the gas and water pipe industry, which produces centrifugally cast pressure pipes and special pieces (elbows, Y-pieces, T-pieces, etc.). The automobile and general engineering industries have an even greater and increasing demand. Rollers, ingot molds, and heavy, thick-walled storage containers for the disposal of radioactive waste are manufactured from SG iron.

A special development, austempered ductile iron (ADI), also known (somewhat inappropriately) as "bainitic" cast iron, has interesting mechanical properties. When its tensile strength is 900 N/mm^2, its breaking elongation is 5 – 12 %; when its tensile strength is 1400 N/mm^2, its breaking elongation is 1 – 2 %. These materials also have very good abrasion properties. The bainitic hardening is carried out in a salt bath at ca. 300 – 500 °C for 20 min to ca. 2 h. As yet, there is no standard specification in Germany, the United States, Japan, or Sweden. A European standard is in course of preparation.

Cast Iron with Vermicular Graphite. In this type of cast iron, the graphite that separates out is mainly in vermicular form. The connection between this type of cast iron and the mineral vermiculite, a magnesium – aluminum silicate, is only conceptual. On heating and expanding, vermiculite swells up to worm-like shapes (Latin: vermis, worm). Expanded vermiculite is used in the building industry as a sound, and thermal insulator, and has no connection with cast iron.

The basic structure of cast iron with vermicular graphite can be ferritic, pearlitic, or a mixed structure. The mechanical properties of the material lie somewhere between those of gray iron and SG iron. Vermicular graphite has been known since the discovery of spheroidal graphite in 1948. Any vermicular graphite in the SG iron structure was regarded as having a harm-

ful effect on properties. Its occurrence indicated the presence of troublesome elements (principally titanium), which limited the formation of spheroidal graphite if the standing time of the melt was inadequate or excessive. This led to fading of the effects of the magnesium treatment, and imperfect formation of the graphite in the interior of thick-walled SG iron castings.

There is no German standard for this material, but local standards exist.

Alloyed Cast Iron. The properties of cast iron can be modified within wide limits by adding elements such as nickel, chromium, manganese, copper, and silicon, which change the metallic structure. Corrosion-resistant and high-temperature grades are used, e.g., in the chemical industry, in furnaces, and in the automobile industry. The cryogenics industry uses low-temperature grades that do not become brittle at temperatures far below 0 °C. The electrical industry has wide-ranging requirements for nonmagnetic materials, and materials with high electrical resistivity.

In Germany, these materials are covered by DIN 1694 (austenitic cast iron) and DIN 1695 (abrasion-resistant alloyed cast iron), in the United Kingdom by BS 3468, and in the United States by ASTM A 439.

White Cast Iron and Roll Casting. In these special grades of cast iron, the carbon in the structure is not graphite, but iron carbide. White cast iron is extremely resistant to abrasion by frictional and grinding effects. These grades are used in a wide range of rolling operations, and in grinding and mixing processes. As these materials are normally not heat treated, the mechanical properties depend only on the chemical composition, even for very heavy and thick-walled components (rolls).

Malleable Iron. Malleable iron castings acquire their special properties by a tempering heat treatment. Malleable iron is formed from an iron – carbon alloy (DIN 1692) whose chemical composition is such that, after solidification, the carbon in the metallic structure of the casting is not free graphite, but chemically combined as iron carbide, Fe_3C. In this condition, the material is hard and brittle, and almost useless. The tempering process causes a transformation of the white iron by decomposition of the iron carbide at the high temperature used. Elemental carbon is precipitated, and is then present in the structure as compact nodules or flakes.

Depending on the heat treatment, two types of malleable iron can be produced: "white heart" malleable iron in which the heat treatment removes the carbon (Fig. 111), and "black heart" malleable iron in which it does not.

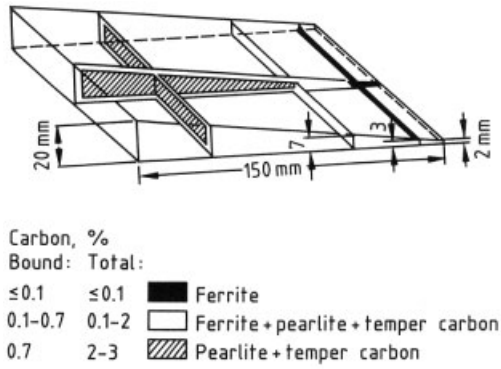

Figure 111. Distribution of carbon and development of the structure in white heart malleable iron of various wall thicknesses

Malleable iron can be produced only in castings of limited mass and wall thickness. Mass varies between a few grams and ca. 100 kg. The good machining properties and especially the uniformity and consistency of the useful properties of malleable iron make it highly suitable for production foundries.

Most of the malleable iron castings produced (65 %) go to the automobile industry. In second place comes the building industry (fittings and pipe connectors). Other consumers include the general engineering and electrical industries, and manufacturers of locks.

Melting of Cast Iron. The oldest and most widely used melting furnace in iron foundries is the cupola furnace. It was developed from the early charcoal furnaces, which were sometimes provided with a cover to keep off snow and rain.

The modern cupola furnace is a shaft furnace with height approximately six times its diameter. The inside diameter is ca. 800 – 3000 cm. Most cupola furnaces have melting capacities of 6 – 12 t (Fig. 112), but the largest cupola furnaces produce 80 t/h. Pig iron, alloying elements, scrap castings, and steel scrap, together

with coke (the fuel) and limestone to liquefy the slag and combine with the sulfur in the coke, are all charged in weighed batches into the cupola furnace.

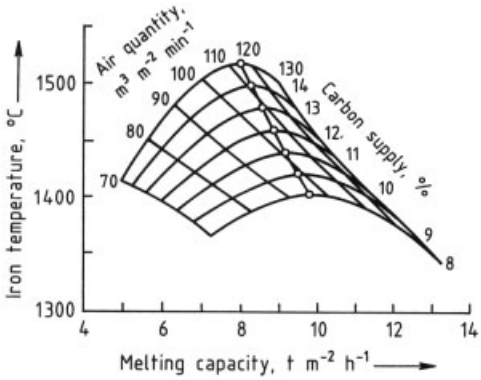

Figure 112. Iron temperature and melting rate as a function of air blast and available carbon for a cupola furnace with specific cross-sectional area 1 m^2

The cupola furnace consists of a strong cylindrical steel sheet outer cover, lined with a thick layer of refractory bricks and rammed refractory material. In the bottom part of the furnace shaft, ring-shaped nozzles (tuyeres) blow air into the incandescent layers of coke, iron, and other materials.

In the melting zone, the metallic raw materials, selected so that the molten cast iron has the desired composition when it flows from the furnace, are mixed. Molten cast iron at ca. 1500 °C and molten slag are then tapped from separate openings into separate containers. The cast iron is usually collected in a large, refractory-lined holding vessel in front of the cupola furnace. The casting ladles can then be filled from this as required. The slag flows into a cast iron skip, and, after solidification, is disposed of on a waste tip. It is sometimes granulated by adding water, or is used to produce slag wool, used as an insulating material.

Modern cupola furnaces use the hot blast system, in which the combustion air is preheated and sometimes enriched with oxygen. This increases the temperature and melting capacity.

In another type of cupola furnace, firing is by gas or oil. The furnace is provided with burners instead of tuyeres. The column of metal is supported in the shaft by a water- cooled grate, located above the burner zone. Ceramic spheres on the grate increase heat transfer by convection, and the molten iron passes over these and drips into the hearth. The cupola furnace is an economic method of melting large amounts of iron of constant or slightly varying composition. The molten iron can be continuously removed from the furnace. To an increasing extent, induction crucible furnaces are used in iron foundries, usually operating at mains frequency (50 Hz) with iron capacities up to 80 t, although most are in the range 3 – 25 t. Furnaces of this type do not absorb electrical energy when empty, and absorb very little when charged with material in small pieces, because no appreciable induced current is produced under these circumstances. Hence, low-frequency crucible furnaces must always have a quantity of molten metal in the furnace (25 % of the total crucible capacity). To start up the furnace, a precast starting block that matches the shape of the crucible is placed in the crucible by means of a crane.

Medium-frequency induction crucible furnaces, operated at 500 – 1000 Hz, are similar to the mains frequency furnaces, except that their capacities are considerably less for electrical reasons. Medium-frequency furnaces are used if rapid melting and heating, melting without starting blocks, or frequent changes of metal grade are required.

4. Forming

Metals, including steel, can be subjected to plastic deformation in the solid state by the application of external forces, i.e., they can be formed into new permanent shapes. Here, the term “forming” or “metal forming” means the intentional and controlled change of the geometry of a workpiece, usually from a simple geometry to a complex one with regard to the shape, size, accuracy and tolerances, appearance and properties. The mass, composition, and state of the material remain unchanged. If, however, the plastic limit is exceeded so that the change in geometry is not controlled, forming with defects occurs, and the desired product is not obtained.

The phenomenon of plasticity results from the crystalline structure of metals. Their engineering and industrial use depends on forming technology.

4.1. Fundamentals

4.1.1. Causes of Plasticity [171], [172]

In defect-free metal crystals, the atoms are arranged in regular lattice structures.

Plastic deformation takes place when lattice planes slide over each other under the influence of shear forces. The lattice structure remains unchanged, but stepwise slippage occurs at the crystal edges, and these ultimately lead to changes in shape. In metals produced by the usual processes, lattice defects are present (holes, interstitial atoms, dislocations, grain boundaries, twinning boundaries, and interphase boundaries). The dislocations lead to a reduction by two or three orders of magnitude in the stresses (and hence the forces and energies) needed to cause deformation, compared with an ideal, defect-free crystal. However, plastic deformation leads to an increase in the number of dislocations; increasing hardness as the extent of plastic deformation increases. The other lattice defects hinder plastic deformation, thus increasing the hardness. Grain boundaries are an exception to this, being associated with "creep", a very slow form of plastic deformation.

4.1.2. Workability

The properties that enable a material to undergo more or less plastic deformation under given circumstances are grouped together under the term "workability". The two most important characteristics that determine workability are the yield stress and the formability.

The yield stress k_f is defined as the uniaxial tensile stress required to cause plastic flow. Its magnitude, like that of other material properties, mainly depends on the structure of the material, and hence, on its chemical composition [173]. As the structure depends to a large extent on the previous history of the material, i.e., the thermal and forming processes, these processes provide an effective means of influencing the structure and hence the properties of the material (Section 4.1.6). On the other hand, these phenomena mean that the yield stress is not a constant property of the material, but changes during forming. Thus, it also depends on the instantaneous temperature, the degree of plastic deformation, and the previous changes to both these quantities.

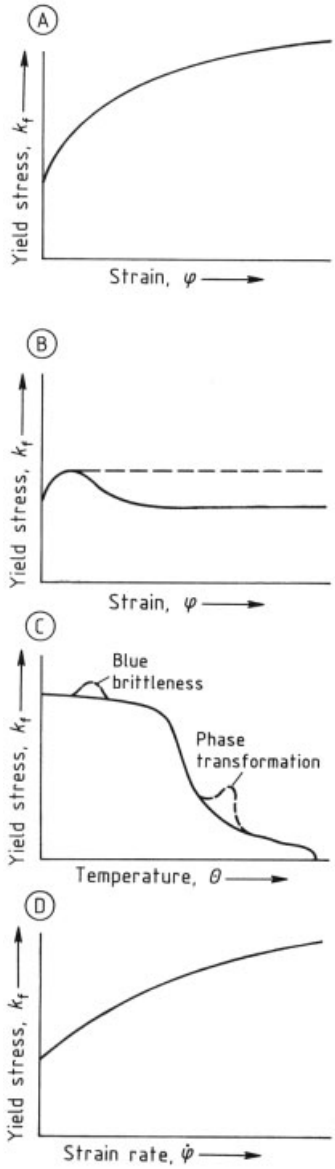

Figure 113. Stress – strain curves
A) Typical cold stress – strain curve; B) Typical hot stress – strain curve; C) Effect of temperature on yield stress; D) Effect of strain rate on yield stress at hot-forming temperature

These complex interactions make it extremely difficult to express the yield stress as a function of all the parameters that affect it, or to enable it to be determined with certainty for a given situation. Mostly one is content with a simplified view, in which yield stress is determined

and quoted as a figure that depends on the instantaneous temperature, strain, and strain rate. Attempts are being made to provide a comprehensive treatment of the problem [174], [175].

The yield stress is shown by the degree of strain φ in the flow curve, usually at constant temperature θ and strain rate $\dot{\varphi}$, although graphs of $k_f(\theta)$ and $k_f(\dot{\varphi})$ are also possible.

In cold forming, the material becomes harder with increasing strain (mainly due to an increase in dislocation density). However, the effect of the strain rate is negligibly small. As the temperature increases, the hardness and yield stress decrease. In hot forming, thermally induced hardness reduction processes take place in opposition to the hardness increase described above. This causes the stress – strain curve to remain constant above a certain amount of strain, or to decrease to a lower, approximately constant level after reaching a maximum (Fig. 113). Also, the strain rate has a much greater effect during hot forming than it does at room temperature.

Formability means the extent of plastic deformation φ_B, which a material can undergo without damage (fracture or cracking). This depends not only on the material (chemical composition and structure), but also on the temperature and stress condition during the entire forming process, and is therefore regarded as a property of the system [172]. There are at present no figures available for the quantitative indication of formability. Comparative assessments of materials are made with the aid of such quantities as the reduction of area after fracture, the number of rotations required to cause fracture in torsion tests, or the degree of deformation on fracture calculated from these figures. It has not yet been possible to use these figures to represent relationships in other forming processes. This is also due to the fact that formability is affected by temperature and deformation history. Formability increases with increasing temperature, and also with increasing compression component in the tensile state. The effect of strain rate is usually small, and the nature of the material determines whether formability increases or decreases with increasing strain rate.

4.1.3. Cold Forming

In cold forming, no heat is applied to the workpiece before forming. However, its temperature can increase during the forming process, as the energy expended in forming is almost completely converted into heat. Frictional energy at the contact surfaces of the workpiece, if established, is also converted to heat. In rapid adiabatic processes the temperature can increase by $> 100\,°C$.

The advantages of cold forming include:

For the workpiece:

- Good dimensional accuracy and close tolerances
- Good surface properties
- Very small dimensions (sheet thickness ca. 10 µm, wire diameter ca. 10 µm)
- Work hardening (if desired)

For the tool:

- Reduced wear (effective lubricants can be used, the thermal load is low, and scaling does not occur)

Generally:

- No energy is required for heating the workpiece

4.1.4. Hot Forming

The term hot forming is used when heat is applied to the workpiece before forming. In general, the temperature for hot forming is higher than the recrystallization temperature. If the temperature is too high, a coarse-grained structure can be formed, which is undesirable (Section 6.1.1).

The advantages of hot forming include:

1) Improved formability of the workpiece
2) Smaller forces required, as the yield stress of the workpiece is less
3) Large degree of deformation possible in one forming step, with consequent reduction in processing time
4) Beneficial effect on the structure, and hence on the properties of the workpiece
5) Little or no work hardening (if this is not desired)

The most significant disadvantages of hot forming are:

1) Very high cost of heating the workpiece compared with the energy required for forming
2) Capital investment for the heating furnaces
3) Very severe tool wear due to the very hard and brittle scale that forms on the surface of the workpiece
4) Reduced working life of the tools due to the thermal load and increased wear

Warm forming in the temperature region between room temperature and the recrystallization temperature is similar to hot forming. It represents a compromise between the advantages and disadvantages of hot forming (reduced yield stress, increased formability, little or no scale formation, and effectiveness of lubricants).

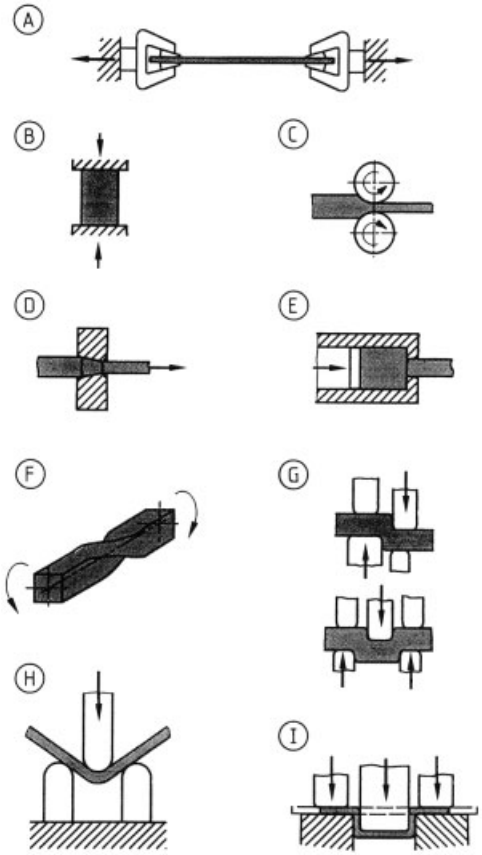

Figure 114. Principles of forming
A) Stretching; B) Upsetting; C) Rolling; D) Drawing; E) Extruding; F) Twisting; G) Shearing; H) Bending; I) Deep drawing

4.1.5. Forming Processes

Forming processes must provide a tensile state that causes plastic deformation. The yield stress of the workpiece must be exceeded as well as the friction losses, which depend not only on the coefficient of friction in the contact area between workpiece and tool, but also on the geometry of the forming zones. If the geometry of these zones is unfavorable, friction losses can cause the work of forming to be many times the ideal forming energy, e.g., if a very thin sheet is rolled between relatively large rolls.

Especially during the forming of brittle workpieces, the tensile state must have a large compression component to increase the formability of the workpiece. This has led to the development of forming processes in which compression effects predominate, such as extrusion pressing or shape rolling.

According to DIN 8580, forming is one of the most important manufacturing processes, along with casting, cutting, joining, surface coating, and modification of material properties.

In DIN 8582, forming processes are classified according to the stress state:

1) Compressive forming (mainly massive forming processes): rolling, forging, die forging, extruding
2) Tenso- compressive forming (massive forming and sheet forming processes): wire, rod, and tube drawing (through tapered tools), and deep drawing of flat sheet to produce more complex components
3) Tensile forming (of sheet and rods)
4) Bending (of sheet and rods)
5) Shear forming (mainly massive forming)

The forming processes listed above are based on the principles of stretching, upsetting, rolling, drawing, extruding, twisting, shearing, bending, and deep drawing (Fig. 114).

In DIN 8583 – 8587, the individual forming processes are defined and explained. International and national guidelines and standards that correspond to the DIN standards are not known.

4.1.6. Aims of Forming Technology

Forming technology has developed out of one of the oldest hand-finishing operations, i.e., manual

metal working as applied to jewelry, tools, and weapons, and has led to the almost completely mechanized and automated manufacturing processes of the metal industry that enable dimensionally accurate, high-value products with special properties to be manufactured economically. It competes with other processes (casting, cutting, and joining), and, with increasing scarcity of raw materials and energy, it helps to meet the demand for cost effectiveness.

4.1.6.1. Combination of Casting and Forming

In the usual method for producing steel sheet and strip, slabs 1500 – 2000 mm wide, 200 – 300 mm thick, and 10 – 15 m long are produced by casting (usually continuous casting, Section 3.5.3). After intermediate storage, the slabs are usually reheated to 1100 – 1250 °C before shaping in a rolling mill, first in the blooming train and then in the finishing train, using a total of 10 – 15 rolling passes, to give a strip 1.5 –15 mm thick. Cold rolling is then carried out if thinner products and/or special properties are required.

The improved process consists of casting "thin slabs" (ca. 50 mm thick), or even strip 2 – 20 mm thick, and (sometimes after temperature equalization) rolling this with just sufficient passes to give the desired final thickness, structure, and dimensional accuracy. This conserves the heat energy for reheating the slabs, and the mechanical energy for the rolling stages; it also leads to savings in plant investment costs.

4.1.6.2. Thermomechanical Treatment

Thermomechanical treatment (TMT, see also Section 6.1.3.2.6) is described in DIN 17 014 as a combination of forming and heat treatment aimed at avoiding reheating, reducing the time and equipment required, and improving the properties of the product. In some cases, smaller amounts of expensive alloying elements need to be used. Several processes with various objectives are used for different grades of steel. For example, with "normalizing forming" of structural steels, a structure can be produced that would otherwise require additional heat treatment. The properties of high-strength structural steels can be improved by "thermomechanical forming." Hardening from the hot-forming temperature of high-strength structural steels and the "controlled cooling" of ferritic – pearlitic steels and special steels after such a heating process leads to improvements in their properties and the avoidance of heat treatment [176].

Most large-scale TMT procedures can be divided into three groups [177]. According to the temperatures used and the point on the time – temperature – transformation curve used for the forming process, the processes are referred to as high-temperature thermomechanical treatment (HTMT), low-temperature thermomechanical treatment (LTMT), and isoforming.

4.1.6.3. Simulation of Forming Processes

The increasing capacity of computers enables technical parameters associated with forming operations to be calculated, by sophisticated computational techniques, such as the finite element method. This gives increasingly accurate forecasting of integral parameters, such as forming forces and forming power, enabling plant design and utilization to be planned. Local temperatures, stresses, and strains can also be calculated. Even at the planning stage, statements can be made about local loads on materials, so that any necessary corrective measures can be tried out if these loads represent failure conditions. The same procedure can be applied to the fill behavior of grooved rolls and forging dies.

Algorithmic optimization, supported by modern computers, enables technical objectives to be realized during the planning stage, and can, for example, help in the design of the grooves for a pass sequence. Information systems can also provide expert knowledge not formulated algorithmically.

4.1.6.4. Quality Information Systems

Particularly in multistage sequential finishing lines with quality information systems, the quality criteria achieved in the separate process steps are linked by information technology to the associated process parameters. The resulting picture of the whole process enables the causes of

defects to be identified much earlier, so that production flow can be optimized.

4.2. Hot Working

4.2.1. Reheating Furnaces

The various types of fuel-fired reheating furnaces employed in rolling mills are used to heat the rolling stock as uniformly as possible for a preset period to give a temperature on discharge appropriate for the forming process. Moreover, these heating and soaking processes must take place with optimum energy utilization, minimum cost, little scale formation, and no unwanted surface decarburization. Surface defects, such as skid marks formed during movement of the rolling stock in the furnace, must be avoided.

Furnace construction and thermal design depend on the condition of the charge material and the particular needs of the rolling process. The most usual types of continuous reheating furnaces [178–180] are the top- and bottom-fired pusher type, walking beam, and rotary hearth furnaces. The rolling stock can be of various sections, lengths, masses, and steel grades, e.g., slabs 160 – 260 mm thick, billets 100 – 330 mm edge length, and round bars 175 – 550 mm diameter. The discharge temperature is normally 1200 – 1250 °C for most hot-forming processes. Pusher type furnaces have a maximum output of 300 t/h, walking beam furnaces 450 t/h, and rotary hearth furnaces 120 t/h.

Depending on the output and complexity of reheating furnaces, the heating process is subdivided into several zones, whose temperatures are controlled separately. In the gas flow direction, there is usually a preheating zone, one or more heating zones, and a soaking zone. In pusher type and walking beam furnaces, these main zones are subdivided into upper and lower zones, and sometimes left, middle, and right zones. In general, the furnace control over the whole range of outputs and steel grades should give a uniform heating rate of the workpiece as it passes from the inlet to the exit of the furnace.

Heating is mainly by natural gas, which can vary in composition. In integrated steelworks, coke oven gas, converter gas, and mixed gases are used, or liquefied gas or heating oil.

Developments in furnace technology and heat conservation, especially during the last 10 years, have led to new or modernized reheating furnaces, operating under computer control [181–187]. The aims of computer control include: minimization of energy consumption; temperature equivalence (especially during nonsteady operation); reduction of scaling; minimization of personnel workload; increased output; matching input and output rates within a furnace group; quality assurance; and documentation. These aims can be achieved with the aid of various furnace control strategies, e.g., using the temperatures of the furnace room or of the surface of the rolling stock, and fuel flow rates and parameters derived from these. In models of this type of furnace operation, external information about processes taking place before and after reheating must also be considered. Material flows inside and outside the furnace must be monitored accurately.

Process computer control means using a supervisory computer to monitor and control the plant, implementing mathematical models to optimize the control parameters. In the process control concept for continuous reheating furnaces (Fig. 115), the actual state of the process is calculated with the aid of a mathematical model from the measured process parameters, material flow data, and physical characteristics of the fuel, combustion gas, and charge material. From this, a second mathematical model calculates the optimum control set point for the heating process. These depend on set values, e.g., of the desired cycle time and required final temperature. Disturbances and planned or unplanned interruptions are also included. A special characteristic of this kind of furnace operation strategy is the use of a volume rate of combustion air, which depends on the mixing ratio of the various fuel gases, as a setpoint for the subsidiary control circuit.

The complexity of the hardware and software depends on individual needs. Integration of the process computer to the basic automation system has been progressively simplified.

Potential Energy Savings. Continual improvements in design, construction materials, heating and combustion techniques, and thermal economics have optimized the process of heating the material, with simultaneous reduc-

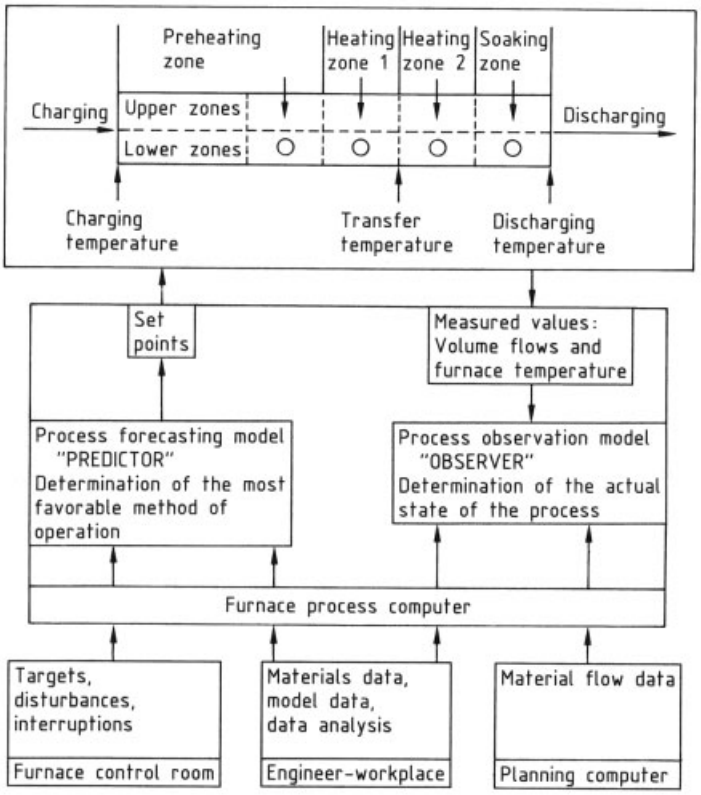

Figure 115. Data flows in a computerized process control system for a continuous reheating furnace

tions in energy consumption and energy-linked emissions.

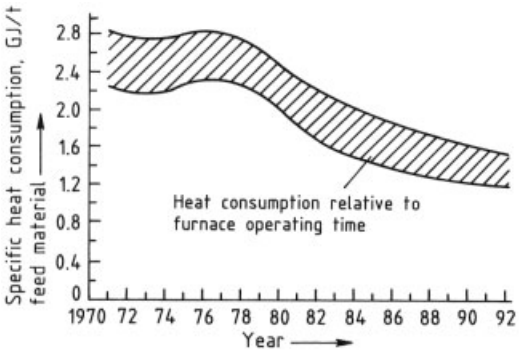

Figure 116. Development of the specific heat consumption of walking beam furnaces for heating slabs

Figure 116 shows the changes in the specific heat consumption of walking beam furnaces for heating slabs for rolling mills producing hot-rolled strip, thick plates, and beams. The heat consumption over a complete operating cycle is estimated for cost calculation, including the energy used for heating the furnaces and maintaining them at temperature. The scatter is mainly a consequence of differences in furnace layout, the extent of energy recycling by media preheating, and the benefits resulting from waste heat utilization. Heat consumption is also affected by, for example, heating requirements, fuel qualities, rolling mill operating, and the considerable fluctuations in workload associated with the economic situation.

The following discussion gives an overview of the most important steps to save energy, and enables a systematic, quantitative comparison of the individual steps to be carried out. Calculation of potential savings involves both observed figures and estimated values. Figures are based on a reheating furnace of average design and operation.

Increase in Furnace Length. The most important design characteristic is the length of the furnace, and hence the specific load

($kg\,m^{-2}\,h^{-1}$). The optimum specific load is $500-600\,kg\,m^{-2}\,h^{-1}$, depending on the heat consumption and capital cost. An unheated zone at the charging end (preheating zone), to reduce the temperature of the waste gas at the furnace exit to ca. 750 °C, increases the effectiveness of heating and hence utilization of the fuel.

Increase in Preheating Temperature of the Combustion Air. A temperature rise of 150–200 °C, i.e., up to 600 °C, leads to a saving of ca. 15 % in heat consumption. The waste gas temperature at the exit of the recuperator should not fall below 200–300 °C. This is necessary for an energy-saving natural draft, and avoids low-temperature corrosion in the waste gas system. The furnace design enables the combustion air and the preheating zone to be jointly optimized.

Subdivision of the Furnace into Individually Controlled Zones. This enables the heat consumption in the various furnace zones to be optimized, so that the charge material can be heated at as late a stage as possible. This gives an increase in flexibility, a reduction in heat consumption, and better conditions for computer control.

Uniform Distribution of Fuel and Combustion Air to the Burners: Choice of Burners. The objective is to obtain a uniform flame temperature and hence higher heat transfer, with consequent energy savings. Local overheating and scaling problems can be avoided in this way.

The burners selected should be controllable over a wide range, enabling the temperature in the furnace room to be as uniform as possible. Specially optimized burners have been developed to minimize NO_x production.

Optimization of Fuel/Air Ratio and Furnace Draft. To minimize heat consumption, it is essential to monitor the furnace draft closely, and to reduce air leakage. The ratio of fuel gas to combustion air must be controlled. A precondition for this is that the charge and discharge doors should be as leak-tight as possible. An improvement in the control of this ratio can give ca. 10 % saving in energy for each 2 % reduction in oxygen content.

Improvements in Furnace Linings. In many cases, a large reduction in heat loss through the furnace walls and top can be achieved by additional or improved insulation. A reduction in heat consumption of ca. 2–4 % is attainable.

Improvements in Insulation of Movement and Transport System. This technique is especially effective in prolonging the working life of the equipment. The utilization of waste heat from the movement and transport system for steam production reduces the amount of energy consumed by the steelworks.

Use of Process Computers to Optimize Control of the Heating Process. The introduction of a process control system reduces heat consumption by 5–10 %, and gives constant, reproducible heating and heat penetration of the charge material, especially during nonsteady furnace operation. The process computer optimizes furnace operation, especially in the event of disturbance or interruption.

Reduction of the Discharge Temperature of the Heated Product. This can reduce heat consumption by 10 % per 90 K. The use of lower product temperature is sometimes limited by metallurgical factors, and the requirements of subsequent forming operations.

Increase in Charge Temperature. An increase in the temperature of the charge material reduces heat consumption by 10 % per 140 K, with a potential increase in output. A precondition for the use of hot charge material is that its surface must be defect free. Technical and quality requirements of the material, and technical and logistic capacity of the available plant largely determine whether this considerable potential for energy saving can be realized.

Figure 117 shows a walking beam furnace for heating slabs for a medium-width strip mill, and illustrates the present state of development of the design of this type of plant with energy optimization. The advantage of this construction is its flexibility of operation. Rolling stock of very different thicknesses can be treated simultaneously. If necessary, the furnace can be emptied rapidly. Charging and discharging of workpieces can take place independently. Surface damage by the skid rails at points of contact with the rolling stock is avoided.

Characteristics of an Energy-Optimized Furnace Design. A reheating furnace should fulfill the following requirements. The furnace profile should be box-shaped. The doors should seal as well as possible, to minimize escape of flames and air leakage. To reduce skid marks, the carrier rails should be staggered as far as the furnace exit. To give uniform distribution of the fuel gas, the upper furnace should be provided with roof or side burners, and the lower furnace with side

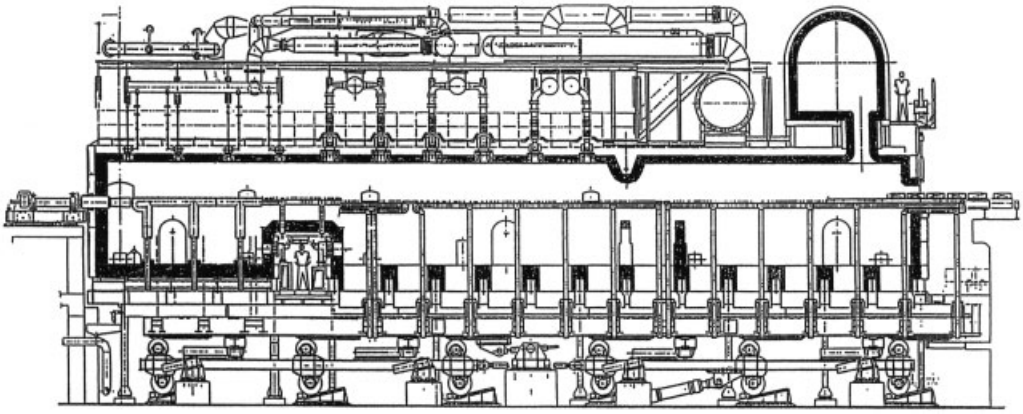

Figure 117. Longitudinal section of a top-and-bottom-fired, eleven-zone walking beam furnace for heating slabs

burners, sometimes in combination with front burners. The furnace should be subdivided into a large number of controlled zones, especially at the charge end. The possibility of shutting off pairs or rows of burners enables the heated length of the furnace to be matched to the load, with the further advantage that the other burners in the working area can be operated. When deciding on the length of the furnace and hence the load on the hearth area, the possibility of using hot charge material must be borne in mind. If the furnace is to be constructed for mixed hot and cold charge material, its length is determined by the cold charge material; the hearth area output is chosen so that, at 75 % of maximum output, the energy consumption of the furnace is a minimum. The exhaust gas temperature should then be ca. 750 °C. The temperature of the preheated combustion air should be 550 °C, so that the quality specifications of burners and pipework need not be particularly high. The furnace should be provided with equipment for oxygen monitoring, not only in the exhaust gas system, but also in the preheating zone. With such a furnace, a specific net energy consumption of < 1.1 GJ/t is possible, even with cold charge materials and discharge temperature 1250 °C. Heating should be under computer control, and appropriate for the charge material.

Future Prospects. The increasing integration of the processes of steel production, including strand casting, heating and soaking of the workpiece, and forming in the rolling mill, is leading to a unified continuous process to optimize the finished product. With flat products, further energy saving can be obtained by near net shape casting. The current practice of producing slabs up to 260 mm thick, is likely to be replaced by casting thinner slabs. The first industrial-scale plants for thin slab casting are in operation. Compared with a conventional hot-rolled strip mill, considerable energy savings can be realized by producing thin slabs that do not require much energy for the reheating furnace, or electrical energy for the rolling mill. Realization of these energy savings depends on the speed at which these new developments in the production of the steel grades demanded by the market can be implemented.

4.2.2. Hot Rolling

4.2.2.1. Semifinished Products

Definitions. Semifinished products are produced by rolling (more rarely by forging) slab ingots and blooms, and are suitable for rolling into flat and long finished products, and also for making forgings. They can also be produced by continuous casting, followed by rolling or forging.

Slab ingots are products with an approximately rectangular cross section, of width at least twice the thickness. Their mass is 10 – 30 t. Slab ingots are mainly used for the production of flat semifinished products.

Blooms, or ingots with square or rectangular cross section, are used for the production

of rolled semifinished products. Their width is less than twice their thickness, and their mass 3 – 10 t.

Semifinished products for rolling are classified according to their cross-sectional shape (EN 10 079) into:

1) Square products with > 50 mm sides
2) Rectangular products with a cross-sectional area > 2 500 mm^2 and width/thickness < 2
3) Flat products normally > 50 mm thick, width/thickness > 2
4) Profiled semifinished products for producing large profiled finished products, the cross-sectional shape being preformed

Production of Semifinished Products. Semifinished products are produced in slabbing mills and blooming mills. The raw materials are heated to ca. 1300 °C in soaking pits, and formed in 15 – 30 rolling operations (passes) through a heavy two-high reversing stand. The forming process is accompanied by densification and grain refinement of the structure produced on solidification. The rolls have diameter 1000 –1250 mm, the length of the roll body being 2500 –3000 mm. The rolls are provided with box grooves of various widths, and with a plain cylindrical section for rolling flat products (see Fig. 125 A). The material to be rolled is moved along roller tables, positioned with the aid of manipulators, and fed into the roll gap. The material is rotated repeatedly by 90° about its long axis during the series of passes.

There are differences between slabbing stands and blooming stands associated with the product, the most important being the greater height of the stands used for rolling slabs, owing to the necessary lift height of the upper roll and the grooving of the rolls.

The rolls in slabbing stands, mainly used for producing flat semifinished products, have a wide, cylindrical working surface designed to accommodate the widest semifinished slab to be produced, and also two or three box grooves.

Another rolling system is the universal slabbing stand. Working pairs of rolls are arranged horizontally and vertically in succession. The vertical rolls form the narrow sides.

The rolls for blooming stands have a much narrower cylindrical part, and a larger number of box grooves.

The torque required for the forming operation is usually provided by electric motors. These can be twin-drive motors, i.e., one for each roll, or the torque can be transmitted from one common motor by means of a pinion stand.

The rolled semifinished products are cropped with shears at both ends, and divided into ordered lengths. Cropping losses, associated with both forming and metallurgical effects, amount to ca. 10 – 15 wt %. Surface defects are removed by flame deseaming (scarfing).

Square and rectangular semifinished products with cross-sectional areas < ca. 40 000 mm^2 are then usually rolled at 1050 – 900 °C, without intermediate reheating, in the stands of a semifinished-product mill until the desired dimensions are obtained. These trains of rolls are open, semicontinuous, or continuous, as required (see Fig. 126).

With the growth of continuously cast products, the number and importance of slabbing, blooming, and semifinished-product mills is decreasing.

4.2.2.2. Flat Products

Definitions. Flat products have a rectangular cross section with width several hundred times the thickness. The surface is technically smooth, though sometimes a regular pattern of raised or depressed areas can be produced (checker plate, button plate, or stud plate).

This group of products includes:

1) Wide flat steel plate
2) Plate
3) Hot-rolled strip

Wide flat plate is a flat product with thickness > 4 mm, the upper limit being ca. 60 mm. The width is 150 – 1200 mm. EN 91 gives a specification for the edge sharpness. Since wide flat plate is now mainly cut from plate or hot strip steel, the cut or burned edge must meet this specification. The rolled product formerly obtained with a universal roll stand (sharp edges produced by vertical rolls) is seldom used.

Hot-rolled flat products supplied in sheets with minimum width 600 mm are known as sheet metal. The minimum thickness is 3 mm and the upper limit is ca. 150 mm, these products being known as thick or heavy plate. Plate up to ca.

20 mm thick can be cut from hot-rolled strip, if the width does not exceed ca. 2 m. For sheets < 3 mm thick, known as thin sheets, this is almost always done.

The maximum possible width depends on the width of the rolls. This dimension is used when naming the rolling mill, e.g., a 4.2 m heavy plate rolling mill.

Hot-rolled strip is defined as a product which, after the last rolling pass and after passing through a cooling stage, is formed into a coil. As with plate rolling mills, the width of the rolls is an essential characteristic. The maximum width of this type of product is ca. 2 m. For example, a plant is referred to as an 88 inch hot-rolled strip plant, the roll width being quoted in inches by international convention.

The European standard EN 10 079 makes the distinction between hot-rolled wide strip (width > 600 mm) and strip steel (width < 600 m, formerly medium and narrow strip).

The second group of these products can also be produced by cutting wide strip longitudinally.

Another very important characteristic is the specific coil mass in kilograms per millimeter strip width. This is derived from the maximum possible diameter of the coil, and is 20 – 30 kg/mm.

As rolled wide strip mills exceed medium and narrow strip mills in both number and output, the following description of hot-rolled strip relates to the first product.

Production Methods: General Features. Hot-rolled flat products are produced by hot forming flat semifinished products (roughly rolled or continuously cast slabs). Exceptionally (e.g., for heavy plate > 100 mm thick), slab ingots > 30 t are used because of their large mass, and because the required degree of deformation can be achieved. This raw material is heated to ca. 1300 °C.

Hot-rolled flat products are produced almost exclusively on four-high stands. Depending on the thickness of the slabs, the plain working rolls have diameters of ca. 700 mm for hot-rolled strip and 1000 mm for heavy plate, and are provided with considerably larger backup rolls (ca. 1500 – 2500 mm diameter) to prevent severe bending, or even breakage, of the rolls. In this way, the greater stretching action obtainable with thinner rolls can be utilized. The torque necessary for the forming action is transmitted from the drive motors to the working rolls.

In heavy plate rolling mills, where the semifinished slabs used are very wide and there is a large reduction in thickness at each pass, large rolling forces and hence torque values are required. The working rolls have one drive motor each (twin drive). This is also the case for the cogging (roughing) stands of hot-rolled wide strip mills. When torques permit (e.g., in the finishing stands of hot-rolled wide strip mills), pinion stands are used.

Production Process: Plate. The rolling process for producing a single sheet is illustrated in Figure 118. The required width of the sheet is first produced from the slab by deformation in the width direction (finished width plus the extra width at the side required by the process). The rolling stock is then turned by 90°, and stretched in the longitudinal direction to the required thickness. The edges of the sheet do not undergo a fixed amount of deformation, i.e., they are free spreaded.

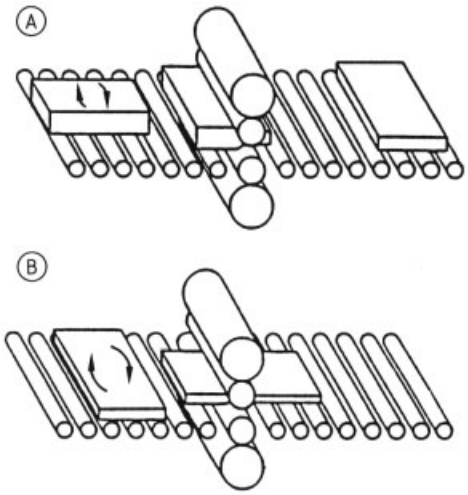

Figure 118. Principle of rolling single sheets
A) Rolling to required width; B) Rolling to required length or thickness

The shaping process in a four-high stand is achieved by many passes between the working rolls, the thickness being reduced at each pass by operating the screw-down gear, and reversing the direction of rotation by reversing the main drive (reversing mill). Heavy plate mills can have a

cogging stand and a finishing stand (two-stand reversing mills).

The rolls in heavy plate mills have maximum width ca. 5 m. Plates of this width are necessary for the production of large, longitudinal-welded pipes. Narrow sheets are usually obtained by cutting double-width sheets longitudinally. The length of a rolled single sheet depends on the thickness and maximum mass of the slab, and can reach ca. 40 m.

Under the influence of the rolling force, the rolls bend (in spite of the backup rolls), so the middle of a sheet is usually thicker than its edges. To reduce this thickness difference, a bending device is used as well as other technical measures.

Variation in rolling forces can lead to varying expansion of the stand, compensated for in modern plants by a hydraulic adjusting device (gap adjuster).

In the production of heavy plate for high-strength constructional steels, thermomechanical treatment is increasingly important [188].

In a conventional rolling mill, the output capacity is fully utilized, so that rolling temperatures increase with larger plate thicknesses, but, with thermomechanical treatment, the final rolling temperature is fixed for a given raw material with a predetermined degree of deformation in the final rolling phase [189].

Normalizing forming (also known as high-temperature mechanical treatment, HTMT) is characterized by an end-rolling temperature higher than the recrystallization temperature of austenite. The limitation to the rolling capacity imposed by the cooling phase on the roll table after cogging down must be set against the energy saving due to the lower heating temperatures and the omission of normalizing.

In thermomechanical rolling (low-temperature mechanical treatment, LTMT) the final rolling phase is carried out below the recrystallization temperature of austenite. The considerable increase in recrystallization temperature for this process is achieved by small additions of titanium or niobium, often in combination with vanadium (microalloyed steels). [190]. The necessary lower temperatures of the final rolling, and the high yield strengths of the austenite (which no longer recrystallizes), lead to high stresses on the plant and reduced production rates compared with conventional rolling [191].

The temperature – time characteristics of the three rolling methods are compared in Figure 119. [192].

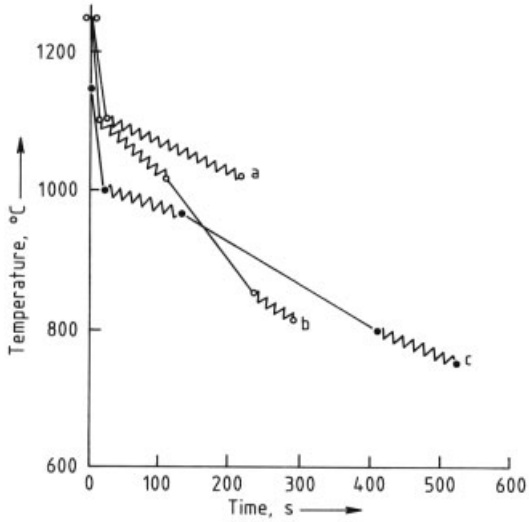

Figure 119. Temperature – time curves
a) Conventional rolling; b) Normalizing forming; c) Thermomechanical rolling

After the rolled single sheets are cooled on cooling beds, they are usually mechanically cropped or automatically cut by burning to the required dimensions (usually rectangular). The cut edges can be prepared for welding, and the rolled sheet can be supplied in a ready- curved shape.

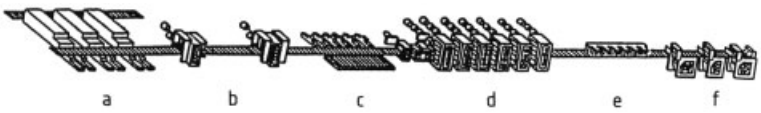

Figure 120. Design of a hot strip mill [197]
a) Slab-heating furnace (pusher-type heating furnace or walking beam furnace); b) Cogging stands; c) Intermediate roller table with transfer skid; d) Finishing train; e) Cooling roller table; f) Coilers

Depending on the layout of the mills, the rolling program, and the dimensions of the products, heavy plate mills can have outputs of $(0.4-1)\times 10^6$ [193].

Production Process: Hot-Rolled Wide Strip [194], [195]. The raw material used for hot strip rolling is mainly continuously cast slabs, ca. 250 mm thick and, depending on the width, up to 40 t mass, the length being 10 – 12 m. Very rarely, rolled semifinished slabs are used.

The slab is first rolled on one or more reversing cogging stands, producing a semifinished strip, 30 – 50 mm thick. Depending on the width of the hot strip, it may be necessary to reduce the width of the slab in the cogging stands. This "upsetting" of the edges is carried out by vertical rolls, or sometimes by a press.

The semifinished strips are usually rolled on continuous finishing trains to give hot strip ca. 1.5 – 20 mm thick (Fig. 120). Exceptionally, reversing stands are used. For this, the four-high stand is provided with two coilers installed in holding furnaces (Steckel mill) [196]. Strip up to ca. 4 mm thick is the starting material for producing cold-rolled strip. The thicker strip is uncoiled on hot strip cropping lines, and cut into single sheets.

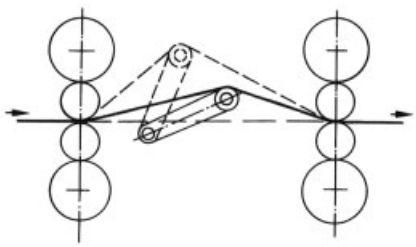

Figure 121. Loop control

The continuous finishing train has 5 – 7 four-high stands whose rotation speeds must be controlled in accordance with the continuity law: cross section × speed = constant. The control parameters can be either the length of strip between two stands (control by looping, Fig. 121), or, for stiff products, the armature currents of the drive motors (minimization of traction force). In the control sequence, the fixed point is the last stand, which has a predetermined constant speed. From this "controlling stand", the entire group of finishing stands is controlled in the reverse direction to that of the rolling operation. This system is also linked to the control of the coiling equipment.

It is essential to maintain a specified final rolling temperature for constant mechanical properties over the length of the strip. As the end of the semifinished strip is rolled in the finishing line some time after the beginning of the strip, it is much cooler, an effect intensified by the water jets used to remove scale. This temperature difference is evened out by the continuous increase in rolling speed over the length of the strip. Final rolling speeds > 15 m/s can be reached, raising the temperature of the strip, owing to the increasing rates of deformation. The final rolling temperature is thus kept constant over the length of the strip.

The variation in temperature along the semifinished strip can also be evened out by means of a coil box. The semifinished strip is formed by bending rollers into a coil on a mandrel or "boom". This is swivelled into the uncoiling position, so that the end of the semifinished strip becomes the beginning of the strip fed into the finishing train [198].

Except for the techniques described above for evening out the thickness differences between the edge and the middle of the strip during hot rolling, an axial shift can be achieved by the use of working rolls, provided with a special profile grinding ability. The hydraulic systems used are known as continuous variable crown (CVC) or universal profile control (UPC) [199], [200].

Because of the variety of materials, precise temperature control is also necessary when coiling the hot strip. This has an important effect on the structure and hence the mechanical properties, which must be suitable for the later treatment processes. This requirement is met by controlling the water rate in the cooling zone, of length 120 – 180 m. Depending on the material, coiling temperatures can be 550 – 780 °C.

After removal from the coiling boom, the coils are mechanically bound, and transferred to the cold rolling mill or to the hot strip finishing shop.

The annual output of hot strip mills can exceed 4×10^6 t, depending on the type of mill and the rolling program.

Since the mid-1980s, the technique of continuous casting thin slabs has enabled hot-rolled wide strip to be produced in considerably smaller production units (Section 3.5.3.11).

Steckel rolling mills have outputs of (400 000 – 700 000 t/a, mostly high-alloy strip (austenites and ferrites) [201].

4.2.2.3. Long Products

Definitions. Long products have a constant cross-sectional area over their whole length, but are usually of variable thickness over the cross section. Long products with a square or rectangular cross section are exceptions to this rule. These have a considerably smaller width/thickness ratio than flat products. Hot-rolled products are usually supplied in standard sizes, to agreed technical specifications. Long products (Figs. 122, 123, 124) include:

1) Wire rod: formed into coils while hot. Nominal diameter or thickness usually > 5 mm. Round, oval, square, rectangular, hexagonal, octagonal, half-round, etc.
2) Bar (solid section): round, rectangular, hexagonal, octagonal, or other shapes. Flat bar of rectangular section, rolled on all four surfaces, thickness > 5 mm, width < 150 mm.
3) Concrete reinforcing or prestressing steel, ribbed or profiled: supplied as bar or wire rod.
4) Permanent way products: rails, sleepers, sleeper clips, fish plates, crane rails, streetcar (tramway) rails (Fig. 122).
5) Sheet piling products: shaped so that continuous walls can be assembled by means of interlocking or clamping devices (Fig. 123).
6) Profiles for mine supports: channel or I-section, with a more pronounced flange taper than with normal I-section (Fig. 124).
7) Large I-section: beams with narrow or medium flanges and web height > 80 mm.
8) Large H-section: beams with broad flanges and web height > 100 mm.
9) Large U-section (channel) with web height > 80 mm.
10) Small I-, H-, and U-section with web height < 80 mm.
11) Angle section: equal-sided or unequal-sided. Rounded edges.
12) Equal-sided T-section.
13) Bulb flat section: flat on one side with a thickened zone on the other.
14) Special sections: usually produced in small amounts.

Figure 122. Permanent way products (railways)

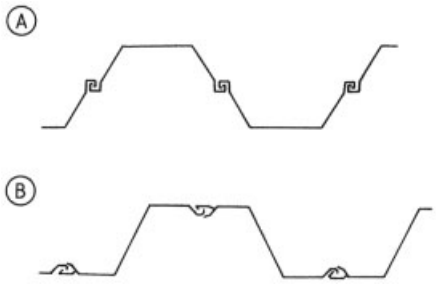

Figure 123. Sheet piling products, shaped so that continuous walls can be assembled by interlocking or clamping devices
A) Larsson profile; B) Z-profile

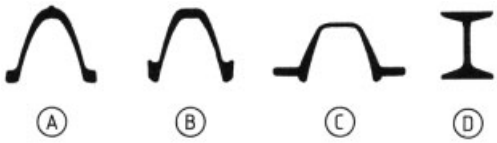

Figure 124. Profiles for mine supports (U- or I-section with a more pronounced inner flange taper than normal I-beams)
A) Bell section; B) TH section; C) TH section with jointing flange; D) Mine beams

Equipment for Producing Long Products. Most long products are formed between grooved rolls in two-high or three-high stands. Parallel-flanged beams are produced in universal stands, which can also be used for rolling rails.

Rolling equipment with grooved rolls is characterized by the number of rolls and the roll diameter in millimeters.

The roll passes are the spaces formed between the pairs of rolls by concentric grooves in the roll bodies (Fig. 125). A series of passes, of continuously changing size, to produce the desired final section stepwise, is known as the roll pass design. A system of repeating roll pass shapes (e.g., diamond/square or oval/round) is known as a roll pass sequence.

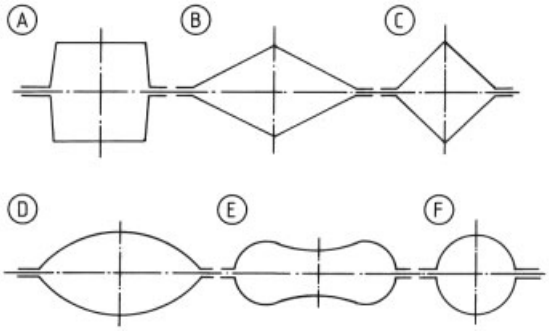

Figure 125. Roll pass shapes
A) Box groove; B) Diamond groove; C) Square groove; D) Oval groove; E) Swedish groove; F) Round groove

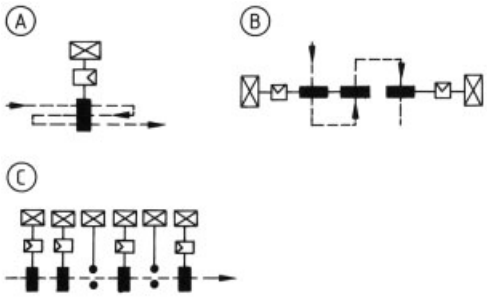

Figure 126. Rolling mills
A) Reversing mill; B) Open mill; C) Continuous mill

The shape of each pass in a roll pass design must be such that a dimensionally accurate product can be rolled, utilizing the flow properties of the material (stretching and spreading) and the available number of roll passes.

The wide range of shapes and sizes of long products requires a large number of roll pass designs, and a wide range of rolling mill layouts. Depending on the production program, this requires a comprehensive stock of rolls, which can include several thousand sets.

The term "rolling mill" includes all the equipment required for rolling the material, and the ancillary equipment for feeding, descaling, edging, guiding, hot cutting, and coiling.

Rolling mills are classified according to

1) The arrangement of the stands and the route taken by the material being rolled (e.g., open or continuous mills)
2) The roll diameter and/or the product (e.g., a 550 mm mill for medium-sized products)

A reversing mill (Fig. 126 A) has one or two stands, through at least one of which the rolling stock passes several times.

In an open mill (Fig. 126 B), several stands are arranged side by side on one axis. These are linked to each other and to the common drive motor by spindles and gears. Two or more strands of steel can be rolled simultaneously in different stands.

In a staggered mill, the stands are arranged with their axes displaced with respect to each other. The rolls can be driven individually, or several stands can have a common drive motor.

In a continuous mill (Fig. 126 C), the stands are arranged in line, so that the strand is rolled in several stands at the same time. For small sections, several strands can be rolled side by side. Depending on the product and/or the material, continuous mills can have horizontal stands only, or alternate horizontal and vertical stands.

Combinations of one or more reversing mills, or an open mill with groups of continuous stands are known as semicontinuous mills.

In continuous mills, the ratio of the rotation speeds of the rolls in the various stands must satisfy the continuity law cross section × speed constant. Deviations give rise to longitudinal forces in the rolled strand, either compression (causing loops), or tension. These forces have a harmful effect on the process and/or the cross section of the rolled product.

Two systems are used to control the longitudinal forces:

Loop control
Control by minimizing tension

In loop control, variation of the size of a loop (a buildup of rolled product between two stands) is used to correct the rotation speed of one of the stands involved. Loop control can be used only for flexible rolled products.

If minimal tension control is used, rolling takes place with only very small longitudinal forces. Physical quantities (rolling forces, rolling torques, and armature currents) are measured and the information is stored before the rolled material enters the next stand; changes in these measured quantities give a measure of the longitudinal forces, and trigger a change in the rotational speed of one of the two stands [202–205].

At the end of the rolling process, the rod-shaped rolled product travels over roll tables to cooling equipment, through which it is moved by various mechanical devices, to give a temperature suitable for processing in the finishing shop.

Finishing operations include straightening and cutting the rod-shaped product, testing for surface and interior defects, marking, and bundling. With some grades of steel, heat treatment is also carried out.

Rolling mills are named according to the size of the section and the mass per meter of long products. They are classified as follows:

Rolling mills for heavy-section products (heavy profile mills and heavy beam mills)
Rolling mills for medium-section products
Light-section and wire-rod rolling mills

The limits in this classification system are not precisely defined.

Rolling Mills for Heavy-Section Products. These mills produce large, heavy, profiled long products, i.e., > ca. 25 kg/m. Products include:

1) Ails
2) Sheet pile
3) Structural steel for mines
4) Beams and channel > ca. 150 mm
5) Bulb plate steel > ca. 180 mm wide

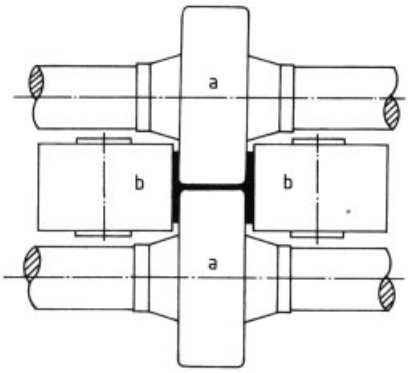

Figure 127. Roll arrangement for making parallel flanged beams
a) Horizontal rollers; b) Vertical rollers

Heavy rolling mills can produce 300 000 – 800 000 t/a. They usually have a cogging or breakdown stand, and, depending on the plant production program, two- or three-high stands, and/or universal stands.

Two- or three-high stands, in which the metal is enclosed and formed by only two rolls are not suitable for rolling parallel flange beams, owing to flow of the metal and wear of the grooves; these products are rolled on universal stands (Fig. 127).

The raw materials for heavy-section rolling mills include:

1) Ingots cast in steelworks
2) Blooms and slabs produced by continuous casting
3) Continuously cast slabs divided longitudinally
4) Rolled or continuously cast rough profiles (dog bones)

The mass of metal is 3 – 10 t, depending on the product and the rolling program. Continuous casting is becoming increasingly important, owing to the higher yield and better surface.

The ingots are heated to ca. 1300 °C, and rolled on a blooming stand to produce a square or rectangular section, or, in the case of broad-flanged beams or sheet pile, to give a roughly profiled product. The rolls of the blooming stand have diameter ca. 1000 – 1200 mm, and are provided with several box-shaped grooves. By screwing down the top roll and turning over (edging) the rolling stock, square or rectangular sections can be rolled in these grooves.

The final section is formed in 7 – 11 shaping passes. Two to four stands are needed, depending on the width of the rollers and the width of the stock being rolled. Roll diameters are 750 – 950 mm. An example of this type of shaping process is shown in Figure 128, which shows the grooves in two stands for rolling rails, the first stand having seven different grooves for seven passes, while the second is designed for four passes, with two grooves being used for each of the last three passes [233].

For parallel-flanged beams, the cogging stand first produces a rough profile, similar to the section produced by the first pass of the subsequent universal mill. Shaping continues in this main stand and a flange-edging stand.

Universal stands have two pairs of rolls, horinzontal and vertical. The web is formed by the driven horizontal rolls, and the flanges by idle vertical rolls. Both pairs of rolls are adjustable,

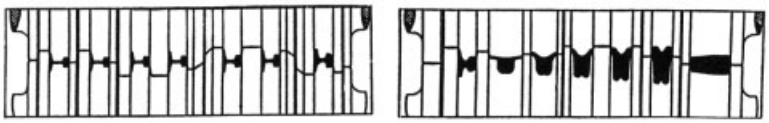

Figure 128. Roll pass design for making rails

so that the thickness of the web and flanges is reduced at each pass. The dimensional accuracy of the flange edges, which are stretched without restriction, is adjusted by passing through a flange-edging stand.

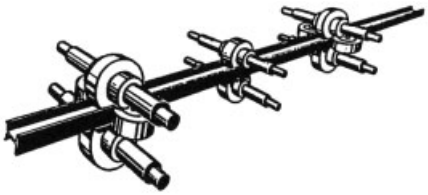

Figure 129. Reversing tandem group for rolling parallel-flanged beams

Forming can also be carried out in a reversing tandem group. This consists of two universal stands with an intermediate flange-edging stand (Fig. 129).

Universal beam rolling mills are operated as single-purpose mills, but can also be operated as open trains, with the stands arranged parallel to each other, and with a common cogging stand [207].

The plant can be designed as a combined rolling mill (Fig. 130), where complete stands can be removed from the rolling line by changeover equipment. Universal and two-high stands can be interchanged. This gives great flexibility in the rolling program, with rapid changeover [208].

Thermomechanical treatment of rolled steel products is mainly used for beams intended for high-stress applications. The temperature

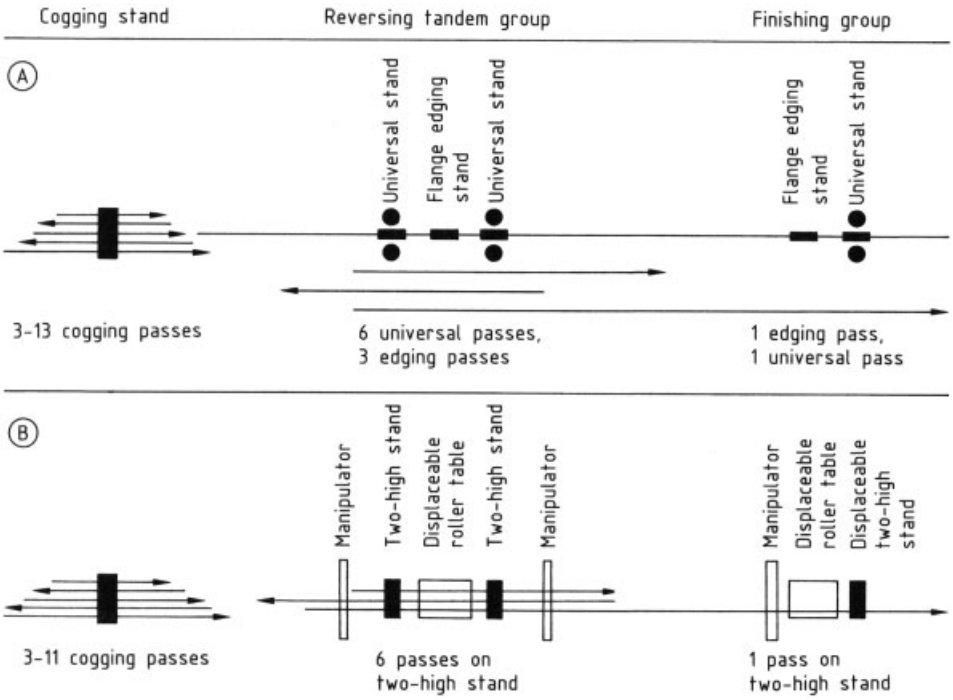

Figure 130. Operating system in a combined profile rolling mill
A) For beams; B) For other sections

changes during a thermomechanical rolling operation are shown in Figure 131, and compared with conventional rolling. Special attention must be paid to the differences in wall thickness. The thinner webs and thicker flanges can form different structures, owing to temperature differences, resulting in differences in mechanical properties across the section. To limit these temperature differences, the rolled product is selectively water- cooled in the area of the transition zone from the flange to the web. A high degree of low-temperature deformation is required, but limits are imposed by the mechanical loads that the rolls and stands can support [209], [210].

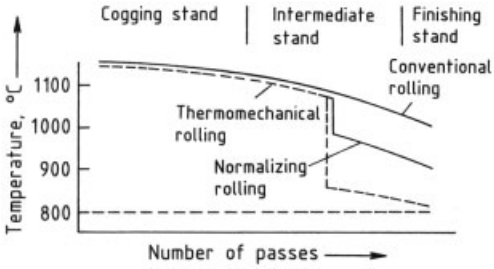

Figure 131. Temperature changes while rolling beams

Another current development is to combine thermomechanical rolling with heat treatment of the rolled product after the last forming pass. This is known as the quenching and self-tempering (QST) process, and consists of rapidly cooling the rolled profile with water. This quenching is interrupted when the interior of the rolled product is still hot enough to reheat the surface region to ca. 600 °C, producing a structure which gives the product a high yield strength and good toughness properties [211].

The finished rolled products can be 60 – 120 m long. They pass over roller tables to the cooling bed which cools them to < 100 °C. Most products can go to the despatch store after normal finishing operations.

Rails undergo extensive quality control and further treatment where necessary:

1) Testing for inner and surface defects
2) Testing for smoothness and straightness
3) Straightening the rail ends
4) Sawing to an exact right angle and very close length tolerances

For special requirements, the rail head can be tempered. This utilizes either the heat from the rolling, or involves an induction heating step [212].

Rolling Mills for Medium-Section Products. Medium-section long products, of less mass per unit length, are linked to the production program of heavy-section mills. They include the following:

1) I-beams with web width 80 – 240 mm
2) H-beams with web width 100 – 240 mm
3) U-section (channel) with web width 80 – 240 mm
4) Equal-sided angle steel, 70 – 150 mm
5) Unequal-sided angle steel, from 60×40 mm to 150×100 mm
6) T-section: 50 – 150 mm
7) Round bar: 40 – 180 mm
8) Flat bar: 80 – 150 mm×5 – 50 mm

The rolling programs for medium-section mills include, in addition to the normal I- and U-sections light structural sections with web or flange thickness ca. 3 mm, and considerably smaller flange width than those of normal sections. These thin-walled products are known as "junior beams" and "junior channels", with web width 6 – 12 inches.

The rolling programs of medium-section product mills are often restricted to certain products, as the production of medium-section products requires plant and rolling techniques that differ in some respects. The main emphasis lies on the production of profiled sections, especially parallel-flanged beams. The products have mass 6 – 25 kg/m, but larger broad-flanged beams lie outside this ange [213].

The raw materials are rolled semifinished products made from ingots or continuously cast blooms of square or rectangular section; weighing 1 – 5 t continually cast blooms are also used.

Apart from plants designed for rolling special products in small batch sizes, e.g., special sections or high-alloy steel round bar, most rolling mills for medium-section products are designed for semicontinuous operation; there are one or two reversing cogging stands and a group of continuous stands. The material that goes to the finishing group, which consists of at least 6 stands, is first rolled on the cogging stand(s) in 3 – 8 passes. The rolls of the two-high continuous stands have diameters 650 – 450 mm.

Continuous trains give shorter throughput times compared with reversing mills, so temperature losses of the rolled products are lower. This ensures that the final rolling temperatures are high enough in the case of small and/or thin-walled profiles, which cool more quickly. When continuous finishing stands are used in conjunction with a crank-cutting device after the last rolling stand, long strands can be rolled and cut as they pass through the stands, improving output and yield. If parallel-flanged beams are included in the rolling program, it is best to have combined two-high/universal stands and flange edging stands. Figure 132 shows a plan of a typical rolling mill for medium-section steel products.

Rolling speeds are 8 – 12 m/s. The rolled strand is cut into optimized lengths, which can be > 100 m, depending on the dimensions of the cooling bed. They are then straightened in machines that can sometimes handle several strands, and passed to the finishing shop.

The outputs of rolling mills for medium-section products, depending on plant design and rolling program, are 250 000 – 800 000 t/a.

Rolling Mills for Light-Section Steel and Wire Rod. Light-section steel comprises rolled products in the form of bar, ca. 12 kg/m. The various sections and dimensions are:

1) Round bar: 8 – 40 mm
2) Square bar: 10 – 40 mm
3) Hexagonal bar: 10 – 35 mm
4) Flat bar: 20 – 80 mm × 4 – 15 mm
5) Angle steel: 20 × 20 mm to 60 × 60 mm
6) T-section steel: 20 – 50 mm
7) U-section steel (channel): 30 – 50 mm

Larger sections are produced by some manufacturers. Most products have solid sections (round, square, hexagonal, and flat); production of profiled-section products (angle, T, and channel section) is not described here.

Wire rod differs from bar in that it is formed into coils after rolling; the production process is very similar for both, so combined wire rod/bar mills are often operated. Light-section steel and wire rod form the largest proportion of hot-rolled long products. The raw material is mainly square section semifinished product with edges 100 – 150 mm long. This can be rolled either from ingots or continuously cast blooms. In the upper size ranges, material can be used direct from the continuous casting machine. The blooms weigh from a few hundred kilograms to > 2000 kg, and are heated to 1150 – 1200 °C.

The stands used for light-section steel and wire rod are arranged for semicontinuous or continuous operation. Open trains are less often used.

Instead of conventional cogging stands, high-reduction machines of various designs can be used. In one pass, these produce a stretching degree of 3 – 10, compared with 1.2 – 1.4 in conventional rolling mills [214].

For light-section products, 10 – 24 rolls passes are required, depending on the raw material and final section; for wire rod, up to 30 passes.

The diameters of the rolls are ca. 550 mm for the cogging stands, ca. 280 mm for the finishing stands. Various roll pass sequences are used for forming (diamond/square, diamond/diamond, oval/round). The thickness of the rolled steel is reduced along each axis of the section alternately; for continuous rolling in horizontal stands, the rolled strand must be twisted after each pass. The twisting operation is often avoided by the use of alternately horizontal and vertical stands (H – V mills). This twist-free rolling method, beginning at an early stage of the rolling process, is preferred for products with high-quality surfaces.

Light-section steel mills can be designed for one or two strands, wire rod mills up to four. Multistrand rolling is only possible with horizontal stands. For improved surface quality and dimensional accuracy of the product, the finishing stands, and even the intermediate stands, are divided into several parallel lines rolling single strands. In light-section mills for round and hexagonal bar, twist-free roll blocks, described in more detail below, are used for the intermediate and finishing trains.

If thermomechanical treatment is to be carried out, it must be taken into account that in continuous rolling mills the final forming operation takes place at high speed, and the distances between the stands are short. This leads to heating of the rolled product. The temperature before the final pass can be reduced to < 800 °C by a water-cooling and equilibration stage after the intermediate rolls, sometimes accompanied by

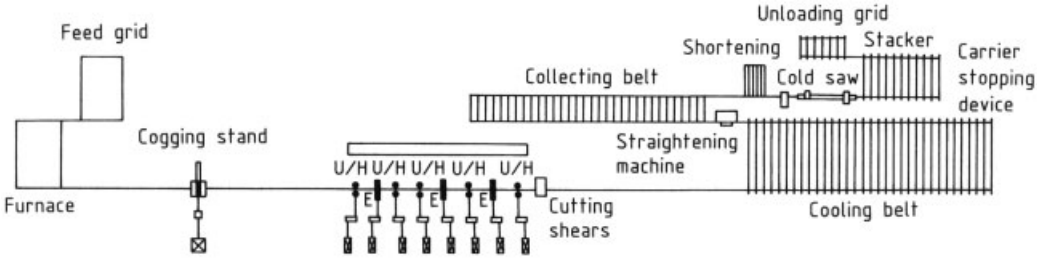

Figure 132. Plan of a rolling mill for medium-section products
U/H = combined universal/horizontal stand; E = edging stand.

reducing the temperature at which the material leaves the furnace to ca. 950 °C.

The flow rate of the cooling medium is designed to give as uniform a temperature profile as possible, and depends on the section of the product and its composition. Water cooling in the finishing roll block may also be necessary to control the final temperature [215].

The light-section steel leaves the final stand at speeds of 10 – 20 m/s, corresponding with the cycle time for the delivery of the bar or rod to the cooling bed. While the strand is in motion, it is cut with rotary shears into lengths suitable for the cooling bed, preferably of the rake type. For some special requirements, e.g., for certain concrete reinforcing products, the bars can be quenched and tempered using the rolling heat of the material.

In light-section rolling mills, the products undergo the normal finishing processes, according to the material and its intended use.

Significant improvements have been made in the technology of rolling wire rod. These include the development of twist-free roll blocks, and cooling equipment that provides accelerated or retarded cooling. For a wire rod 5.5 mm diameter, the final rolling speed has been increased from ca. 40 m/s in 1960 to > 100 m/s.

Hence, larger cross sections of continuously cast material can be used for the initial pass, without leading to an unacceptably low rolling speed in the first stand. This also enables the coil weight to be increased from 600 to 2000 kg or more [216], [217].

In twist-free roll blocks, up to 10 rolling units are arranged on a single common frame, and these are linked by a system of gears [218]. Forming can be carried out by two rolling disks with an oval/round roller pass sequence (Fig. 133), or by three rolling disks (Fig. 134 A) with the roll pass design shown in Figure 134 B [219], [220].

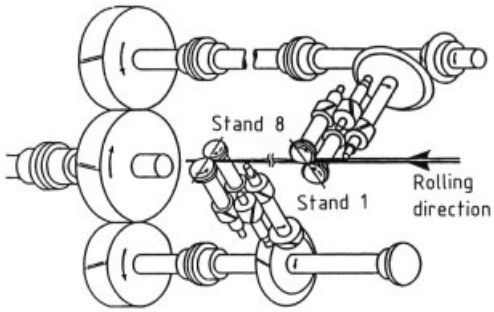

Figure 133. Two-disk roll block (drive mechanism)

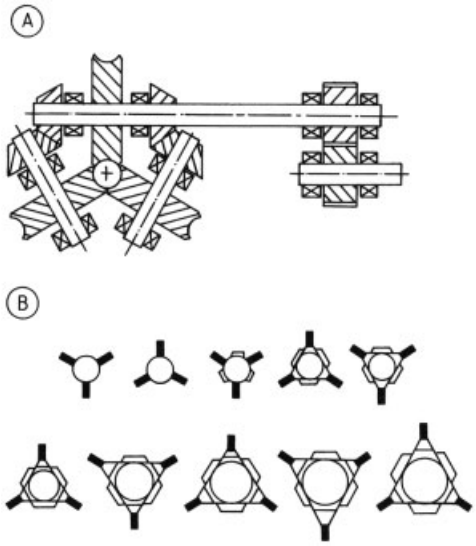

Figure 134. Three-disk roll block
A) Drive mechanism; B) Roll pass design

The disk rolls are made of highly wear-resistant hard metal and give the best possible surface finish. Roll blocks replace conventional roll stands in the finishing trains, and sometimes also in the intermediate trains. The high strand rates obtainable with roll blocks make it possible to replace multiple-strand train by stands with one or two strands.

Experience with roll blocks has enabled so-called precision roll units to be developed. These give improved dimensional accuracy, and are used in both light-section steel and wire rod mills [221].

Wire rod of diameter ca. 5 – 20 mm passes through a water- cooling zone after leaving the last stand. This consists of cooling boxes with intermediate zones to equalize the temperature between the interior and exterior of the product. Finally, the wire rod is formed into loops by a winding head, and these are spread out on a speed-regulated conveyor belt. The overlapping loops are force cooled by a controlled air flow in several zones. Cooling of the wire rod can be retarded by covering the conveyor belt. At the end of the cooling process, the wire loops are collected in coil-forming chambers, and tied into coils (Fig. 135) [222].

With this controlled cooling by water and air, the crystalline structure and condition of the scale are controlled to give the required material properties.

Wire rod of diameter ca. 20 – 40 mm is formed into rings in rotary basket (Garrett) coilers.

Capacities of light-section steel mills are 100 000 – 350 000 t/a, depending on the plant design and the rolling program (dimensions and materials). Multi-strand, high-output wire rod mills can have capacities > 700 000 t/a.

4.2.2.4. Tubes

Steel tubes can be either seamless or with a welded seam, depending on their dimensions and intended use [223]. The current dimensional standard is ISO 4200. The dimensional ranges are fixed partly for technical reasons, but manufacturing costs also play a role. Tables 14 and 15 show how production and export have changed during the last 20 years.

Table 14. World steel tube production, 1000 t

	1975	1980	1985	1990	1995
Japan	8921	12 280	12 224	10 527	8678
Other Far East	1427	3025	4017	5919	7490
China	1200	2168	3206	4337	5896
Germany	5553	5450	6247	4390	3520
Italy	3496	4140	3727	3428	3483
France	1801	2101	1735	1491	1344
United Kingdom	1528	1084	1342	1437	1134
Rest of EU∗	2207	2749	3134	3279	2974
CIS	15 967	18 170	19 354	19 512	5782
Eastern Europe	4646	5223	5289	4179	2429
United States	7466	8252	3717	4220	4932
Latin America	1954	2747	2554	2491	3147
Others	3633	4290	4646	4410	5342
Total	59 799	71 679	71 192	69 620	56 151

∗ 15 countries.

Table 15. World steel tube exports, 1000 t

	1970	1980	1985	1990	1995
Japan	4374	6306	6529	3143	2499
Other Far East	352	1144	1441	1417	1611
China	10	2	7	85	113
Germany	3525	3058	4300	2641	2202
Italy	1671	1464	1867	1158	1734
France	902	1181	1153	994	868
United Kingdom	418	283	462	456	528
Rest EU∗	1227	1862	2647	2456	2095
CIS	339	370	430	249	575
Eastern Europe	1065	1239	1384	1139	979
United States	953	427	203	430	851
Latin America	98	411	840	1285	917
Others	578	763	1301	1178	1644
Total	15 512	18 510	22 564	16 631	16 616

∗ 15 countries.

Seamless Tubes. Production of seamless steel tubes is usually carried out in three manufacturing stages [224], [225]:

1) Piercing
2) Elongating
3) Final rolling

In modern plants, one mill is used for each manufacturing stage. In older plants, each stage is usually divided into several smaller stages. The processes are named according to the elongating equipment used. The most important types of plant are:

1) Mandrel mills (Fig. 136),
2) Plug mills (Fig. 137),
3) Push bench mills (with a cross-rolling mill for piercing)

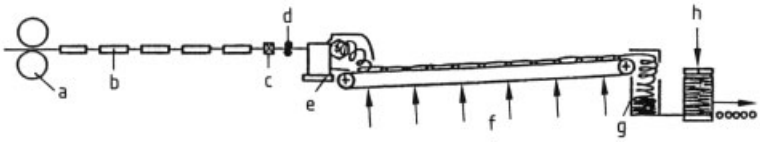

Figure 135. Wire- cooling system
a) Finishing stand; b) Water- cooling train; c) Crop shear; d) Drive unit; e) Laying head; f) Air- cooling belt; g) Coil-forming chamber; h) Coil-tying station

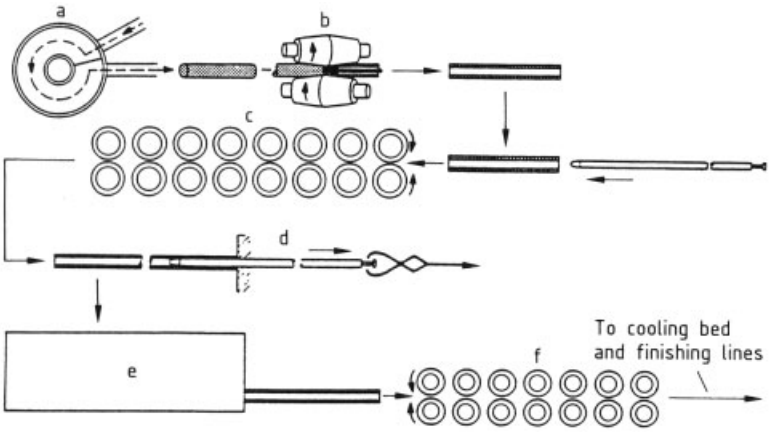

Figure 136. Mandrel mill
a) Rotary hearth furnace; b) Cross-rolling piercer; c) Mandrel mill (offset 45°); d) Stripper; e) Reheating furnace; f) Stretch-reducing mill (offset 60°)

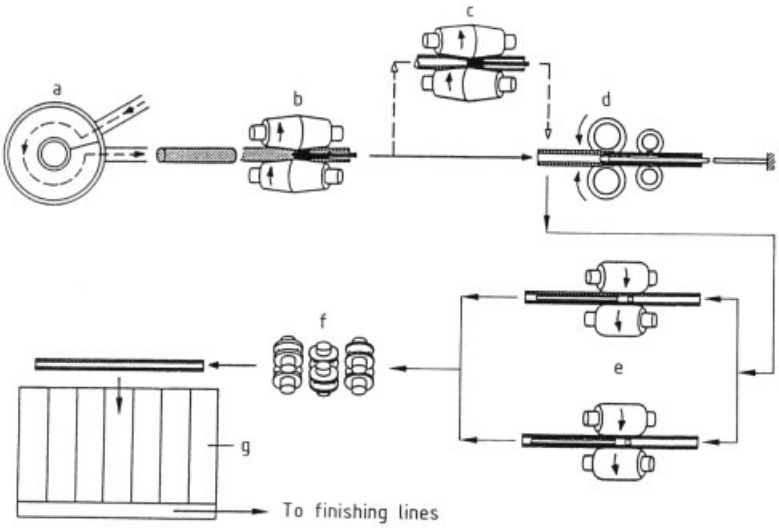

Figure 137. Plug mill
a) Rotary hearth furnace; b) Cross-rolling piercer; c) Cross-rolling elongator; d) Plug mill; e) Reeler; f) Sizing mill; g) Cooling bed

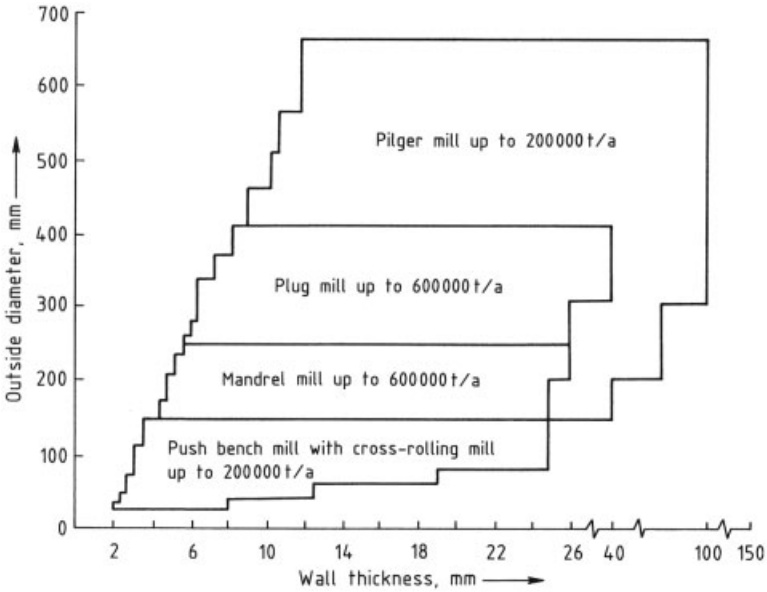

Figure 138. Operating ranges of tube rolling mills

4) Pilger mills

Product size ranges and annual outputs are given in Figure 138.

As well as the four processes, others are used for lower volume products:

1) Assel mills, for ball-bearing tubes and machinery tubes
2) Diescher mills, for medium-sized products, not thin-walled tubes
3) Extrusion presses, for high-alloy materials

Plug and mandrel mills are mainly used for processing unalloyed steels, e.g., simple structural steels, and medium-alloy materials. However, high-alloy steels, e.g., for pipework, oilfield applications, and boiler tubes, are also processed in these mills, the materials 100Cr6, X20CrMoV121, and X20Cr13 being the most difficult materials used.

Raw Materials. The raw material now used all over the world for unalloyed and medium-alloy steels is continuously cast round bar, although continuously cast square bar is also used, either rolled to form round bar, or directly as square bar. Square bar is first pierced by a press piercing mill (integral with the tube production process) to give a thick-walled hollow billet. The cross-sectional area remains approximately constant in this operation. The material for high- chrome steel tubes is still mainly round bar stock, rolled from cast billets.

Pilger mills for large diameter tubes still use cast billets.

Billet Heating. Billet heating is today almost entirely carried out in rotary hearth furnaces. As a rule, these have six zones, and are heated by natural gas or oil, the process being computer controlled [225].

Piercing. By far the most important process for the hole-piercing stage uses inclined rollers. In the method invented in 1886 by the Mannesmann brothers, holes are produced in massive round ingots (Fig. 139) by two inclined rolls rotating in the same direction, so that the ingot is rolled with a screw motion over a piercer plug to form a hollow bloom. The piercer plug is fixed to a freely rotating rod supported by a thrust bearing on the exit side.

In the cross-rolling zone in front of the piercer plug, the core of the material undergoes a continuous transition from compressive to tensile stress, this being caused by the screw motion. If this is too violent, tearing of the core takes place. Formerly, this was thought necessary, but it is now avoided as much as possible by correct design of the rolls, thus preventing interior defects [227–229]. By means of guide shoes

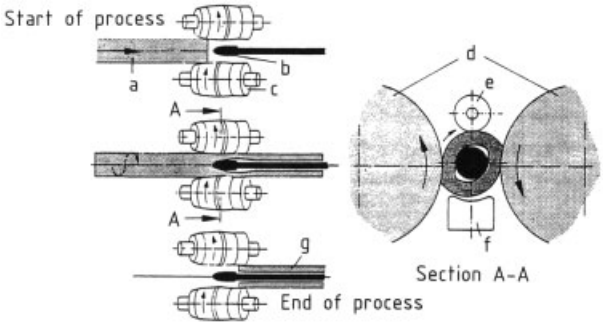

Figure 139. Piercing in a Mannesmann cross-rolling mill
a) Billet; b) Piercer plug; c) Cross-roll; d) Working rolls; e) Supporting roll; f) Guide shoe; g) Hollow bloom

(Stiefel mills, Fig. 140) or disks (Diescher mill, Fig. 141), modern cross-rolling mills are able to produce thin-walled tubes at such a high rate that they can be used, even in high-output plants such as plug and continuous mills [230]. Particularly with plug mills, where the elongating ability of the rolling mill is limited, it was necessary to roll the hollow billet in two stages. In the first cross-rolling mill, a thick-walled hollow billet was produced, and in the second mill the hollow billet was elongated in preparation for the final plug rolling process.

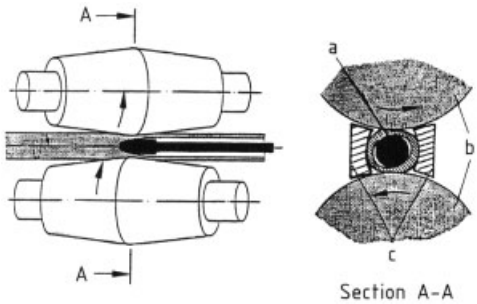

Figure 140. Cross-rolling piercer, Stiefel type
a) Piercer plug; b) Working roll; c) Guides

Elongating. Longitudinal rolling, transverse rolling, or presses are used for the elongating stage. The most important longitudinal methods are:

1) Continuous tube rolling mills
2) Plug mills
3) Push benches
4) Pilger mills

Of these, the continuous tube rolling process is the most important, as it is one of the most efficient and productive processes, and is used for products where the demand is highest. The name expresses the fact that several stands are used simultaneously. There are usually 6 – 8 two-high stands per group, at an angle of 90° to each other. The wall thickness reduction takes place between the rolls and a cylindrical lubricated mandrel (Fig. 142).

There are three types of process:

Type 1. Free-floating mandrel, with stripping of the mandrel as a separate operation
Type 2. Semifloating mandrel, where the mandrel speed is about the same as the exit speed from the first stand, with stripping of the mandrel as a separate operation
Type 3. Semifloating retained mandrel, with short working length (twice the distance between two stands); the rolled product is pulled from the mandrel by pulling equipment, and the rod is pulled back through the rolls to the entry side

For elongating by cross-rolling, the following equipment is used [231]:

1) Diescher mills
2) Assel mills
3) Planetary cross-rolling mills

Cross-rolling processes are used only for plants of small to medium capacity. Owing to certain engineering aspects of the cross-rolling process, thin-walled tubes are difficult or impossible to produce with this type of plant.

A further possibility is to use extrusion presses, but only special materials such as high-alloy steels are processed in this way.

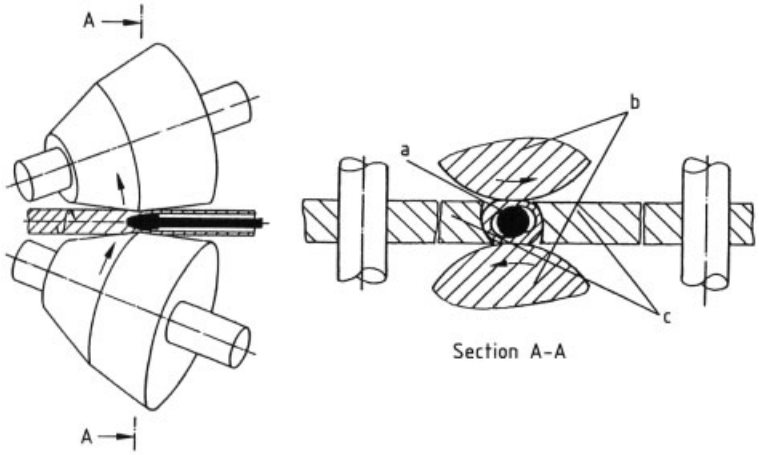

Figure 141. Cone cross-rolling piercer with Diescher disks
a) Piercer plug; b) Working roll; c) Diescher disks

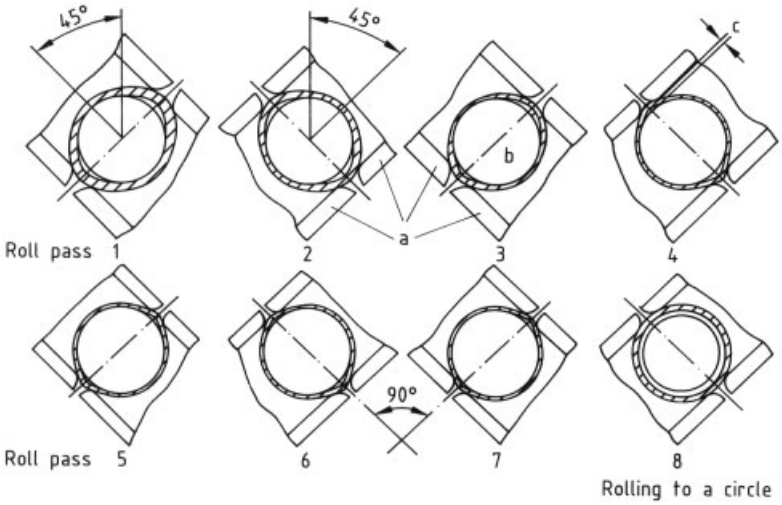

Figure 142. Arrangement of rolls and roll passes in a mandrel mill
a) Rolls; b) Mandrel; c) Clearence

Finishing. Rolling to the finished size is carried out without an inner mandrel in sizing mills or stretch-reducing mills. Sizing mills have 2 – 10 stands, and can have two or three rolls per stand. As the circumference is reduced, the wall thickness is increased by the sizing mill.

In stretch-reducing mills, which normally have three rolls per stand, and up to thirty stands (Fig. 143), an increase in wall thickness is avoided, as tensile forces are applied between the stands during the process.

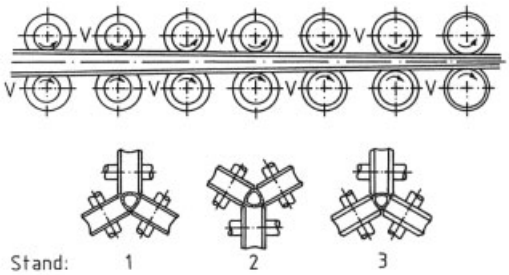

Figure 143. Stretch-reducing mill
Rolls marked V are shown staggered through 60°.

Table 16. Welding processes

Forming process	Welding method	Name of Process	Weld	Other diameter, mm
Continuous	flame pressure welding	Fretz – Moon	longitudinal	13 – 114
	resistance pressure welding	d.c.	longitudinal	10 – 20 (30)
		low frequency	longitudinal	10 – 114
		high frequency	longitudinal	20 – 600
		submerged welding	spiral	168 – 2500
		protective gas (MAG) (tack welding only)	spiral	406 – 2032
		protective gas (TIG, MIG, TP)∗	spiral/longitudinal	30 – 500/10 – 420
Individual forming	electric arc welding (fusions welding)	submerged welding	longitudinal	≥500
Three-roller bending machine C-press		protective gas (TIG, MIG, TP)∗	longitudinal	200 – 600
Individual forming U – O-press		submerged welding protective gas (MAG) (tack welding only)	longitudinal	457 – 1626

∗ High-grade steel tubes.

Further Heat Treatment. Tubes for undemanding applications can be shipped immediately after rolling. For more exacting quality requirements, especially good toughness properties, separate heat treatment is usually carried out after rolling. This can include normalizing and tempering. Normalizing mainly improves toughness, and tempering considerably improves both strength and toughness.

Research has been intensified into tempering the product during rolling, i.e., quenching the tube to $< 100\,^{\circ}\mathrm{C}$ after it leaves the final roll pass, having been heated by rolling to temperatures above A_{C3} (A_{C3} is the temperature at which the transformation of ferrite to austenite is complete). The tube then undergoes a further separate heat treatment stage.

Welded Steel Tubes. Welded steel tubes are produced with longitudinal or spiral welds. The starting material is cold- or hot-rolled steel strip, hot-rolled wide strip, or thick plate.

The manufacture of these starting materials involves considerable cost. Therefore, seamless tubes of medium wall thickness and medium diameter are often cheaper to produce than welded tubes, although production of the latter does not present any technical problems. New developments in near net dimension casting may well lead to changes.

The physical properties which the tube should possess are often already present in the rolled raw material. However, changes, especially to the yield strength, caused by the forming operations must be taken into account. A typical example is the thermomechanically rolled thick plate or hot-rolled strip used in the production of large tubes. The finished tubes can also be heat treated or cold hardened.

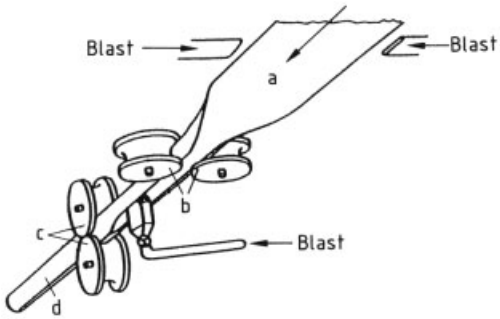

Figure 144. Fretz – Moon rolling and welding plant viewed from below
a) Heated strip; b) Forming rolls; c) Welding rolls; d) Welded tube

Forming of the flat products into tubes can be carried out hot or cold, as a continuous or batch process.

The welding processes are of two types (Table 16):

Pressure welding processes
Fusion welding processes

Pressure Welding. In the Fretz – Moon process (Fig. 144) coils of hot-rolled strip are heated to a high temperature in a tunnel furnace, and the edges of the strip are heated to welding temper-

ature by additional burners. The strip is formed by rolls into a tube with a longitudinal gap, and a two-roll compression device at 90° then reduces the gap and presses the edges together so they become welded.

The semifinished tube passes without further heating through a stretch-reducing mill, where the dimensions are reduced as required. A flying saw cuts the continuous tube into lengths, which are then cooled on a cooling bed. The welding process is controlled by a hot-tube eddy current device, which monitors the semifinished tube and is located behind the compression equipment. The tubes produced are thin-walled with diameter ca. 21.3 – 114.3 mm.

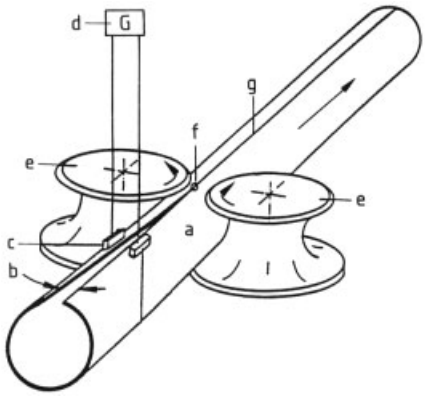

Figure 145. Conductive high-frequency welding
a) Open-seam tube; b) Gap angle at feed; c) Sliding contacts; d) Welding generator; e) Compression rollers; f) Welding point; g) Weld seam

There are three methods of resistance pressure welding, the most important of which is the high-frequency method introduced in the early 1960s. Here, too, the simultaneous effects of pressure and heat are used to weld together the edges of an open-seam tube into a tube without any added material. The heat is generated by the energy of a high-frequency alternating current, which produces a magnetic field within the conductor, increasing the ohmic resistance. This causes electrons to flow along the edges of the strip to the point where the edges meet. Energy production is greatest at this point, where the approach effect at the counter- conductor occurs.

The welding current can also be carried to the gap by conduction (Fig. 145), with sliding contacts, or by induction, using singly or multiply wound coils. Forming is by shaped rolls or roller cages. The tube outer diameter is 20 – 600 mm, with wall thickness 0.5 – 15 mm. The raw material is coiled steel strip or hot-rolled wide strip. The strip can be pickled or cold rolled, depending on the intended use.

The bulb formed both inside and outside the tube by the pressure welding process is planed or scraped off while it is still hot, and the tube is then rolled in 2 – 6 reducing stands to give a circular section of the correct dimensions. It is possible to form the round tube into a different profile, e.g., square, by profiling stands. Here, too, nondestructive testing of the welded seam is carried out continuously.

The properties of the welded seam can be improved, either by continuous heat treatment to give partial normalization, or by a separate heat treatment of the individual tube.

Fusion Welding. Fusion-welded steel pipes, mainly produced in diameters of > 457.2 mm (18 in), are used in pipeline construction. Four forming processes are used (Fig. 146), the last two being by far the most important. Welding is carried out under a powder, often in combination with tack welding under a protective gas.

In the submerged arc welding process, the electric arc is formed under a layer of slag and powder. The added material is coiled, uncoated welding wire. The welding current is fed to the wire by sliding contacts, and is carried away via the electrically earthed workpiece (Fig. 147).

In the protective gas welding process, the electric arc can be seen, although the electrodes, electric arc, and weld pool are protected from the atmosphere by a stream of inert or active protective gas. The following processes are used:

Tungsten – protective gas welding:
- tungsten – inert gas welding (TIG)
- tungsten – plasma welding (TP)
- tungsten – hydrogen welding (THG)

Metal – protective gas welding:
- metal – inert gas welding (MIG)
- metal – active gas welding (MAG)

The TIG, MIG, and MAG processes are used in tube manufacture.

Cold Working. Some seamless and welded tubes are subjected to cold working. This enables wall thicknesses and cross sections to be produced to better tolerances, improves physi-

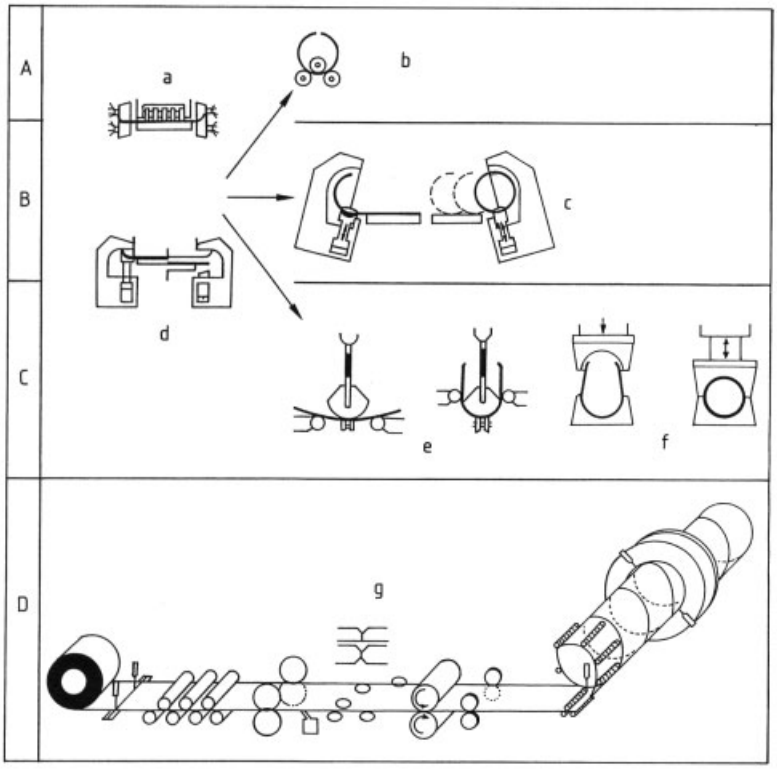

Figure 146. Pipe-forming processes
a) Bending rollers; b) Three-roller bending machine; c) C-press; d) Bending press; e) U-press; f) O-press; g) Spiral pipe process

cal properties, and, not least, enables tubes with thinner walls and smaller diameters to be produced.

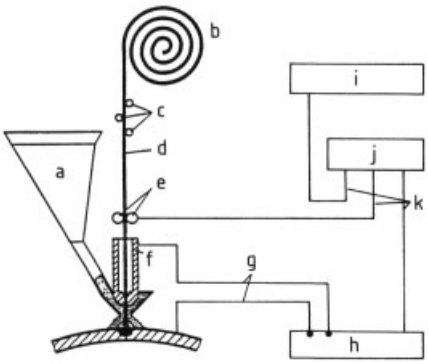

Figure 147. Submerged arc welding process
a) Powder hopper; b) Coil of welding wire; c) Straightening rollers; d) Wire electrode; e) Wire feed rollers; f) Welding current feed device; g) Welding current cables; h) Welding current supply; i) Wire feeding equipment; j) Control unit; k) Control sensor leads

Four variations of the cold drawing process are used (Fig. 148). Cold working increases the yield point and hardness of the material.

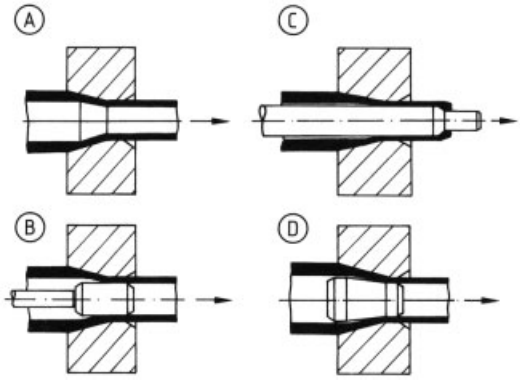

Figure 148. Cold drawing process
A) Sinking; B) Rod drawing; C) Plug drawing (fixed plug); D) Plug drawing (free plug)

A much greater degree of deformation can be produced by cold treatment with a Pilger mill (Fig. 149) than can be achieved by drawing. The rough tube is elongated over a stationary conical mandrel by two specially shaped rolls. These roll the tube, first in one direction and then the other, in a regular repeating cycle.

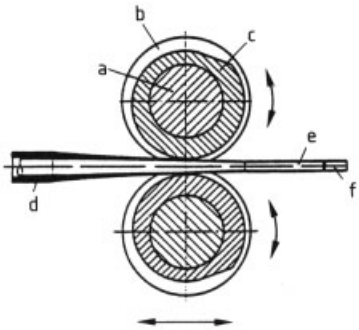

Figure 149. Cold rolling in a Pilger mill
a) Roll axis; b) Roll groove; c) Roll; d) Rough tube; e) Mandrel; f) Finished tube

There is a great reduction in diameter and wall thickness, so that any irregularities in the rough tube are effectively eliminated. A tube rolled in a Pilger mill is therefore characterized by close dimensional tolerances and extremely good surface finish.

4.2.3. Forging

4.2.3.1. Open Die Forging

Open die forgings are the products of forging plants, which can operate by compression (presses) or impact (hammers). Advantage is taken of the malleability of steel at high temperatures to produce a wide range of shapes, some very complex, from almost all grades of steel. The importance of these products lies in the fact that, not only are the design requirements of a steel product largely fulfilled by forging and machining, but so also are the quality requirements. This is a result of the improvement in steel quality brought about by forging, and the heat treatment that it entails.

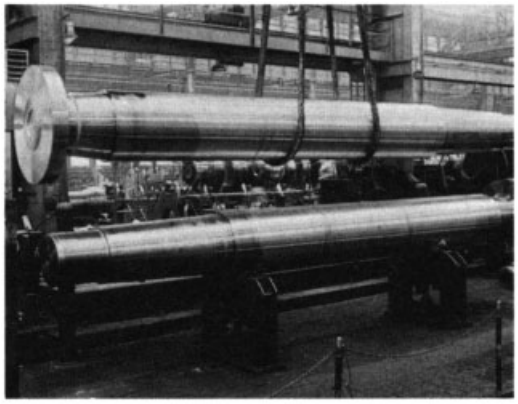

Figure 150. Bronze-clad shaft for a ship's propeller
Material 32CrMo12; Length 7400 mm; Diameter 682 mm; Mass 22.4 t

The amount of crude steel used in open die forging is small compared with the total annual output of steel, but this does not give a true indication of the economic significance of the process. This is only truly reflected if all the stages of value addition are taken into account, starting with the ingot, and finishing with the machined forging, ready for use.

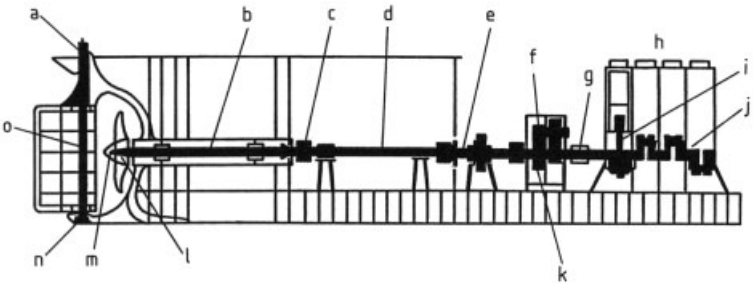

Figure 151. Schematic of a ship's propulsion system with components produced by open die forging
a) Rudder post; b) Propeller shaft; c) Coupling bolts and nuts; d) Intermediate shaft; e) Thrust shaft; f) Gears; g) Pinion shaft; h) Coupling; i) Connecting rod; j) Crankshaft; k) Wheel tire; l) Propeller key; m) Propeller retaining nut; n) Rudder shaft retaining nut; o) Rudder shaft

Customers for these products constitute a highly developed processing industry, with considerable demand potential.

Almost half the products of open die forging go to the mechanical engineering industry. Other consumers include the electrical industry, shipbuilding, diesel engines, mining, quarrying, earthmoving, the heavy chemical industry, structural steel engineering, and the iron and steel industries (Figs. 150 and 151). Smaller consumers are supplied by merchants.

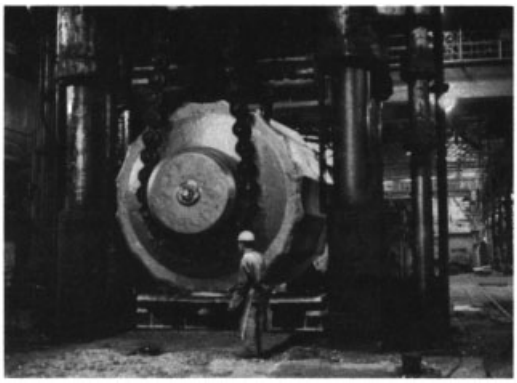

Figure 152. Forging bloom (350 t) in an 80 MN press
Material 26NiCrMo1010; Total length 5300 mm; Diameter 3700 mm

Figure 153. Forging the shaft of a water turbine on a 40 MN press
Material Ck35; Ingot mass 100 t; Forged mass 60 t; Maximum diameter 2010 mm; Total length 6000 mm

The range of steel grades used extends from unalloyed carbon steel, through low- and medium-alloy grades, to high-alloy and stainless steels. The raw shapes used in open die forging include continuously cast steel, semifinished rolled products, and large ingots (Figs. 152 and 153). The largest ingots for forging weigh 650 t (Japan), and, in Europe, 450 t (Germany). Attempts are being made in the People's Republic of China to produce an ingot of 350 t by the electroslag remelting process. The most modern methods of steel production are used to ensure that the starting materials are of the highest possible purity.

The main machines used for hot forming by open die forging include hammers, hydraulic presses (oil or water), and two- or four-hammer forging machines. All types of machine are available in a range of power ratings. The largest are open die forging presses rated at 150 MN (Italy) and 120 MN (People's Republic of China). For environmental reasons, hammers have been increasingly replaced by open die forging presses and forging machines. Ring rollers are machines for hot forming ring-shaped forged products, mentioned here for completeness. The ring roller, with its preheating furnace, upsetting press, and perforating press, form a single unit (Fig. 154).

Figure 154. Ring for rotary ball mill
Material 50CrMo4; Diameter 6000/5600 mm; Height 200 mm; Mass 5.8 t

Open die forging includes both preheating and shaping. Correct preheating is a necessary precondition for economic, problem-free forming. The preheating process used and its duration depend on the size of the forging and the composition of the steel. When forging does not involve the use of shaped tools, experience of working procedures is necessary, including the use of the correct force and power for the shaping operation, and the correct positioning of the workpiece

before each stroke of the hammer. The forming properties of a particular material depend on the nature of the material, its temperature, and the rate and extent of deformation.

The procedures used with power-driven hammers and presses, between flat or profiled saddles are the same as those used in hand forging on an anvil. A forging can sometimes be completed in one working cycle, but usually the final form is obtained gradually, after several different types of working procedure.

Forging always brings about an improvement in the structure of the raw ingot; it tends to remove defects in the ingot which were produced during manufacture, the properties being improved in all three directions. An ingot of given dimensions is given a smaller or larger cross section, i.e., it is stretched or upset. The quality of a forging is determined not only by the type and purity of the steel and its subsequent heat treatment, but also by its drawing-out ratio (initial section : final section). The fiber orientation in a forging should conform to the stresses that the forging will receive in use.

After forging, the workpiece is cooled. Depending on the grade of steel and the dimensions of the forging, this is carried out in still air, under matting, ashes or coarse sand, or in pits or furnaces. For larger forgings and special grades, heat treatment may be necessary before cooling to remove any exterior and interior cracks, giving a product that will pass an ultrasonic test. A crack-free exterior depends on the state of stress in that area, and a crack-free interior depends on effective diffusion of hydrogen, the state of stress, and the structure. A good ultrasonic test result depends on grain refining in the forming process.

Tempering in the unfinished state, i.e., before machining, is carried out, with the aim of producing particular properties (technical parameters, tensile stresses, etc.). Tempering consists of two heat treatment steps: hardening and stress relief. In addition to special equipment for thermochemical treatment to modify the surface down to various depths, equipment is available for heat treating forgings, e.g., vertical, shaft, and horizontal (bogie hearth) furnaces, and heated or unheated pits. Hardening is carried out in still air, by ventilators or compressed air, by dipping in oil or water, or by spraying with water or water – air mixtures.

Machining workshops are often located in the vicinity of the open die forging plant; the end user is often not equipped to machine workpieces to conform with the design drawing. Workpieces must be machined before the final heat treatment, depending on the grade of steel (Fig. 155). For these reasons, a link between the forging plant and the machining workshop has developed.

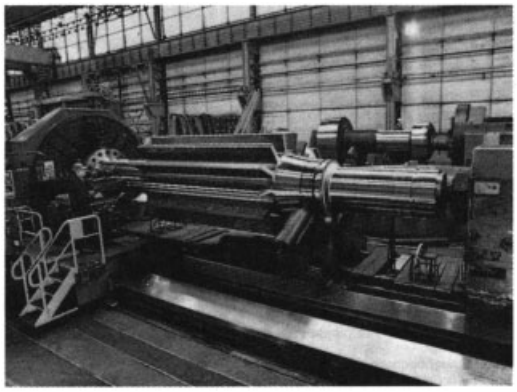

Figure 155. Generator shaft for the largest hydroelectric power station in Norway
Material Ck35; Mass 97.2 t; Ingot mass 270 t; Body diameter 2256 mm; Total shaft length 9390 mm

4.2.3.2. Closed Die Forging

History. Closed die forging is a very old process for producing identical parts. Its importance increased greatly after the invention of the steam engine. This provided the energy necessary for the closed die forging of steel (steam-driven hammers), and enabled large numbers of similar forgings to be produced.

Closed die forging is a highly developed area of industry. The countries that produce the most steel products by this technique are Japan, Germany, the United States/Canada. Each of these produces ca. 10^6 t/a; the other countries of Western Europe produce a further 10^6 t/a.

Definition. In closed die forging, the die is a block of steel, usually rectilinear, in which the "negative" shape (known as the impression) of the article that is to be produced has been machined.

The die usually consists of two halves, between which the slug that is to be forged is placed.

The two halves of the die are then closed by impact or by a press.

Closed die forging is thus a massive forming process, at the end of which the workpiece is completely enclosed by the forming tool. This differs from open die forging, in which the tool does not have the shape of the forging, or only to a limited extent, the workpiece being produced by relative movement between the tools and the slug.

Raw Materials. All metals capable of plastic deformation, including nonferrous metals and steel, can be forged, e.g., in a die.

Steel is the most frequently used material in the production of closed die forgings. The designer has a large number of standard grades of steel from which to choose the most suitable material for a particular requirement. These include ordinary mild steels, tempering steels, stainless steels, or special alloys based on nickel and cobalt (poor forming properties requiring large forces).

The starting materials for die forging are usually rolled billets (square section), bars, and flats. Billets and bars are cut into slugs, and flats are cut into cropped pieces. Slugs and cropped pieces are the raw materials for closed die forging.

Before the forging process, the raw material is heated in fuel-fired batch type, rotary-hearth, or continuous furnaces, or electric heating machines (induction or conduction), to 1100 – 1200 °C.

The material is heated to reduce its deformation stress. Hot, warm, and cold forming can be carried out, but the deformation stress is greater in the last two processes. The tools and machines available limit the degree of deformation achievable, and hence the possible range of die-forged workpieces. Both cold and warm processes give improved accuracy, but the most important process today is hot die forming.

Tools. Because the die is expensive and product specific, closed die forging is, with few exceptions, economic only for a large number of forgings.

The materials used to make the dies are mainly hot-work tool steels, which retain their strength at high temperatures. This is very important when the steel, heated to ca. 1200 °C, flows in the die impression and heats its surface.

Tool life is determined by the wear caused by flow of the raw material and stresses due to heat and pressure; it is defined as the number of forgings that can be produced before the die becomes unusable, and depends on the die wear that still gives acceptable dimensional tolerances in the forging.

The die impressions are produced by machining (turning or milling), or by spark erosion.

A large range of methods of surface treatment are used to improve the wear resistance of the die surface, and hence to increase tool life. The best known include nitriding and metal deposition by welding. Laser alloying processes have shown promise.

Forming Machines. Steel die forgings are produced with the aid of hammers (drop hammers, double-acting hammers, counterblow hammers), presses (crank presses, eccentric presses, screw presses), horizontal forging (upsetting) machines, and other special forging machines. All have their advantages, disadvantages, and special areas of use. Most forged articles can be produced with either hammers or presses.

Hammers and screw presses (fixed energy machines) differ from ordinary presses (fixed movement machines) in that the former involve successive blows of the ram onto the metal in the die, whereas, in the case of a press, the ram header has a fixed lower reversing point, and the forming process in the die must be terminated by one header stroke of the ram header. The ram of the hammer is traveling at its greatest velocity when it first strikes the metal slug, whereas the ram header of the press has a much lower velocity at this point. These different velocities lead to different deformation stresses of the metal being formed, so that a greater amount of deformation can be achieved by one header stroke of the ram header of a press than by one blow of the ram of a hammer, for machines of comparable size.

Forming Process. The forming of a cropped billet to give a finished forging can take place in one or more stages; the choice depends on several factors.

Edging of the material may be necessary as a first step if the forging to be produced has large

changes in section, or if a specific fiber orientation of the material in the forged item is required; this process may involve upsetting, stretching, or automatic roll stretching. If the raw material is a flat, the distribution of the material is determined by the shape of the cropped piece.

A preforming step is used if the work of deformation during production of a complex forging is too great for a single operation, or if the metal flow in the tool takes place so rapidly that forging defects are produced. It can also be used for economic reasons; preforming dies cost less than finishing dies, and the smaller amount of flow or material in the latter leads to a longer die life.

The final forming step is carried out in the finishing die, which has the shape of the forging being produced, but with the dimensions increased by the amount that the forging shrinks on cooling.

These forming operations can be carried out in various machines, but if no special machinery, e.g., forging roll, is necessary, they are carried out alongside each other in the main forging plant. The finished forging has a flash that must be clipped. Flash-free forging is possible, but there must be exactly the right amount of starting material in, e.g., the slug, as fluctuations in the mass can lead to variable thicknesses of the forgings. When forgings are produced with a flash these variations are mostly compensated by the flash.

After flash removal, the forging is often warm calibrated to achieve closer tolerances. The tolerances of particular parts of the forging may be further improved by a cold calibration of these parts.

Heat Treatment. The properties of the steel are usually improved by heat treatment of the forging. This can involve normalizing or quenching and tempering utilizing the heat present in the forging, or, for some grades of steel, using controlled cooling of the hot forging, so that the required mechanical properties, e.g., tensile strength, yield point, or notched bar impact test figures are achieved. The machinability of the forging can also be affected by its structure.

Dimensional Accuracy. The dimensional accuracy achievable depends on the particular forging and the degree of sophistication of the production methods. Dimensional tolerances and permitted deviations are standardized in most countries. In Western Europe, values are harmonized to conform to a Euroforge proposal of 1965. These standards are rather complex, as the rules apply to some 100 000 different types of forgings. The standards quote two quality levels, one for normal manufacture, and one for more sophisticated production methods.

It is possible to produce better dimensional accuracy than that given in the standards. The products are then known as exact forgings, or, with even better tolerances, precision forgings.

Die forgings are normally machined, at least on their functional surfaces. Machining allowances are therefore specified.

Very close dimensional tolerances, which involve high manufacturing costs, are normally only demanded if the cost of machining can be avoided as a result. The decisive factor is the cost effectiveness of the finished die forging.

Shapes and Uses of Forgings. It is estimated that ca. 600 000 different kinds of forgings are produced worldwide, differing in shape and size, and amounting to millions of items.

The range of forgings produced by hammering and pressing in two-part dies can be considerably extended by further forming in upsetting machines, or by a combination of forging and welding.

Closed die forgings can range in mass from a few grams to well over 1 t.

The main consumers of die forgings include the motor vehicle industry (ca. 70 %) and the engineering, mining, railway, shipbuilding, and aircraft industries.

4.3. Cold Rolling

4.3.1. Introduction

For many applications, the products of hot rolling are unsatisfactory, e.g., with respect to cross section, surface quality, dimensional accuracy, and general finish, so that cold rolling is necessary, i.e., reduction of the thickness of the hot-rolled strip between two working rolls without additional heating. Cold rolling of steel is mainly used in the production of light-section,

thin sheet, and stainless steel sheet of thickness 0.1 – 3.0 mm.

The production cycles in a cold rolling mill differ considerably from those in a hot rolling mill. The raw material is first descaled, rolled, and then heated. Further steps are slitting, coiling, inspection, skin-pass rolling and packing.

4.3.2. Pickling

Hot-rolled steel sheet always has a layer of scale of variable structure on its surface, depending on the hot rolling conditions (Fig. 156). The method of descaling depends on the composition of the scale. For high-grade steels, chemical descaling is carried out in hydrochloric or sulfuric acid, while stainless steels are normally pickled in a nitric acid mixture. Purely chemical descaling is often a long process, but it can be accelerated by preliminary mechanical descaling, e.g., stretching, leveling abrasive blasting, or rolling. Effective descaling requires the scale to be broken down.

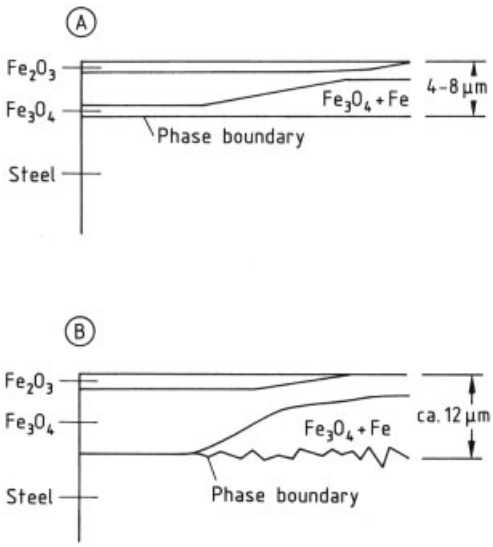

Figure 156. Structure of the scale layer on hot-rolled unalloyed or low-alloy steel
A) Coil temperature ca. 570 °C; B) Coil temperature ca. 750 °C

Pickling is carried out in continuously operating plants (Fig. 157). There are usually two payoff reels from which the strip is run. It is first cut by cropping shears to form the beginnings and ends of the strips, and the start of the band is machine welded to the end of the previous band. Great care must be taken in the production of the weld bead, as this is rolled in the cold rolling mill, and must not be torn off. So that the actual pickling process can take place continuously during this interruption to the flow, there is a storage looper between the welding machine and the pickling plant. After welding, the storage looper is refilled at ca. 2.5 times the pickling speed.

Before the actual pickling process, there is usually a stretcher – bend – leveling device which also causes mechanical descaling. This preliminary descaling is very important, as there must be a certain percentage of free iron surface in the pickling bath, so that pickling can start. Normal steels are usually pickled in 20 – 25 % sulfuric acid at 95 – 100 °C, or 15 – 20 % hydrochloric acid at 60 – 70 °C in three or four fully enclosed pickling baths. For strip speed 240 m/min, the dwell time in the bath is ca. 20 – 30 s. The pickling time depends not only on the steel composition and hot rolling conditions, but is also a function of acid concentration, temperature, and concentration of iron(II) salt, which increases as the dissolution of scale proceeds.

The chemical reactions during pickling are described in the literature [232–236]. Individual pickling reactions are given in Table 17.

Table 17. Pickling reactions

$Fe_2O_3 + Fe + 3\,H_2SO_4$	$\longrightarrow$	$3\,FeSO_4 + 3\,H_2O$
$Fe_3O_4 + Fe + 4\,H_2SO_4$	$\longrightarrow$	$4\,FeSO_4 + 4\,H_2O$
$FeO + H_2SO_4$	$\longrightarrow$	$FeSO_4 + H_2O$
$Fe + H_2SO_4$	$\longrightarrow$	$FeSO_4 + H_2$
$Fe_2O_3 + Fe + 6\,HCl$	$\longrightarrow$	$3\,FeCl_2 + 3\,H_2O$
$Fe_2O_4 + Fe + 8\,HCl$	$\longrightarrow$	$4\,FeCl_2 + 4\,H_2O$
$FeO + 2\,HCl$	$\longrightarrow$	$FeCl_2 + H_2O$
$Fe + 2\,HCl$	$\longrightarrow$	$FeCl_2 + H_2$

As the area of the clean steel surface increases, the pickling time decreases until a point known as the critical free surface is reached. For hydrochloric acid, this is ca. 15 %, and for sulfuric acid, ca. 2 %.

After leaving the acid baths, the strip is washed, dried, oiled, and coiled to form rolls of the required mass. The storage looper compensates for interruptions to the process. All pickling plants are provided with trimming shears,

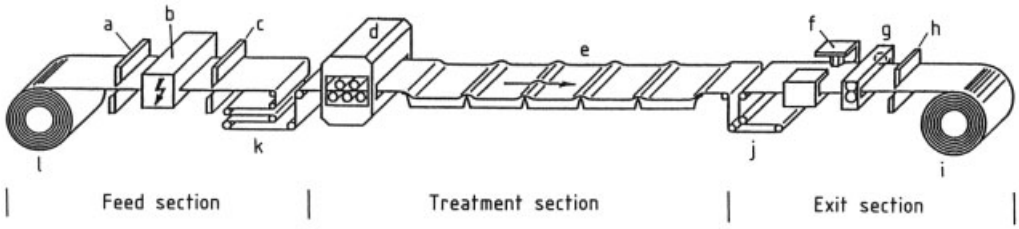

Figure 157. Continuous pickling plant
a) Cropping shears 1 (head); b) Welding/planing machine; c) Cropping shears 2 (foot); d) Stretch – bend – leveling device; e) Pickling/washing; f) Stamping device; g) Trimming shears; h) Cropping shears; i) Take-up reel; j) Exit looper 2; k) Entry looper; l) Pay-off reel

normally located at the exit, though in some plants at the inlet.

The used acid can be regenerated, in the case of sulfuric acid by precipitating iron sulfate as its heptahydrate in a crystallizer.

Hydrochloric acid can be regenerated by decomposing the iron chloride to hydrochloric acid and iron oxide, by a spray process at 450 °C, or in a fluidized bed at 850 °C [237].

4.3.3. Rolling

After pickling, the descaled hot-rolled strip is cold rolled, reducing its thickness (1.5 – 5 mm) by up to 90 %. Many types of stands are used in the production of cold-rolled strip, e.g., two-high, four-high, or larger. In reversing stands, one roll pass can be immediately followed by another by reversing the direction of the rollers. Reversing stands with many rolls exert a high pressure on the strip because the diameters of the working rolls are small. These stands are therefore mainly used for steels with work-hardening properties, e.g., stainless steels and electric sheet steels.

Some 4 – 6 four-high stands can be installed one after the other to form a so- called tandem cold rolling mill (Fig. 158). These mills have the advantage that the strip reaches its final thickness by going in one pass from the pay-off reel through all the stands. Strip speeds at the exit can reach 2400 m/min.

The gap is lubricated with an oil – water emulsion to enable a high degree of deformation (thickness reduction) to be achieved at high rolling speeds. The oil provides lubrication, and the water absorbs the heat produced in the rolling process.

The most important types of cold rolling stands are shown in Figure 159. Areas of use are a function of the work-hardening and thickness of the product strip. Working rolls with small diameters can bend in a horizontal plane, and this can affect the profile; they are therefore given side support in many types of stand.

Experiments have shown that it is not possible to change the profile in a cold rolling mill, so the hot-rolled strip profile must be produced to close tolerances. If the shape of the gap between the rolls does not correspond to the profile of the strip being rolled, flatness defects can occur (differences in the length of individual fibers) [238].

The control elements usually used to adjust the flatness are:

1) Curved working rolls, both positively and negatively. Gap correction up to 180 µm can be achieved.
2) Adjustable intermediate rolls. This system requires six stands.
3) Thermal changes to the crowning of the roll. The degree of thermal expansion of the roll body can be changed locally by adjusting the cooling, but this can only produce extremely small changes.
4) Continuously variable crown systems (universal profile control). The rolls can be adjusted to produce a convex or concave gap; the range of control here is ca. 400 µm.

With the aid of precisely controlled strip tensions between the stands and highly developed measuring and automation technology, it is possible to provide a gap shape appropriate for the shape of the strip. With this system, flat cold-rolled strip with thickness variations of only a

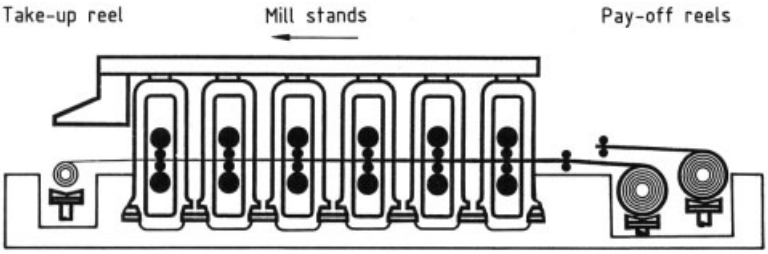

Figure 158. Six-stand tandem cold rolling mill

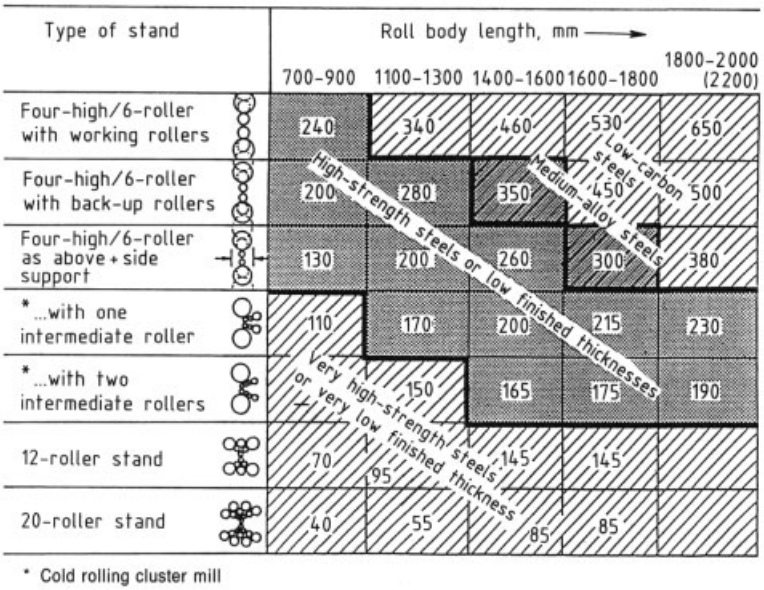

Figure 159. Types of cold rolling stand

few thousandths of a millimeter and high-quality surface finish can be produced.

Cold rolling causes work-hardening of the steel. This is reduced by an annealing stage.

4.3.4. Annealing

The cold-rolled coiled strip is stacked in a hood furnace for annealing (Fig. 160). The combustion chamber is heated by oil or gas burners. The heat passes through the protective hood into the space where the steel coils are stacked. A circulating fan provides as uniform a temperature distribution as possible. The atmosphere in conventional plants is usually HNX gas (a nitrogen – hydrogen mixture in which the hydrogen content is close to the flammability limit).

The heat passes into the coils through their outer edges, so that these areas are always hotter than the inner windings, especially during heating up.

This heat treatment causes the organic residues of the emulsion to burn off without leaving a residue, in accordance with the reactions given in Table 18 [239–241]. The strip is then heated to the recrystallization temperature, and annealed under the protective gas at ca. 700 °C. This treatment produces a complete recrystallization of the cold-rolled steel. The coils are cooled by removing the outer casing. As the annealing space is then hotter than its surroundings, heat passes out of the protective hood in the reverse direction. The outer windings of the coil cool more quickly than the inner windings. If the cooling is too rapid, tensions due to shrink-

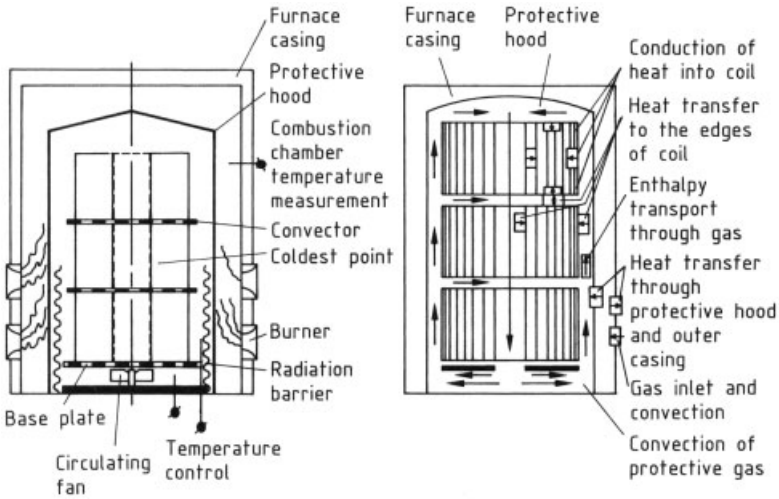

Figure 160. Hood furnace with heat transfer and gas flow indicated

ing occur, and these can cause diffusion welding of the windings (stickers).

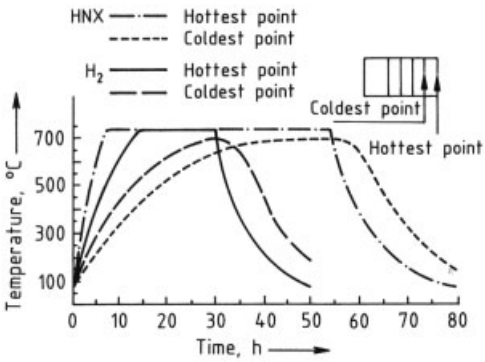

Figure 161. Comparison of annealing times in a hood furnace under HNX and H_2 atmospheres

Table 18. Annealing reactions

$CO_2 + C$	$\rightleftharpoons$	$2\,CO$	Boudouard reaction
$CO + H_2O$	$\rightleftharpoons$	$CO_2 + H_2$	Water gas reaction
$Fe + x\,CO_2$	$\rightleftharpoons$	$FeO_x + x\,CO_2$	Oxidation of iron due to CO_2/CO ratio
$Fe + x\,H_2O$	$\rightleftharpoons$	$FeO_x + x\,H_2$	Oxidation of iron due to H_2O/H_2 ratio
$C + 2\,H_2$	$\rightleftharpoons$	CH_4	Methane reaction

Annealing in this type of furnace was formerly very slow. Another serious limitation was the limited rate of cooling attainable, so that certain grades of steel were very difficult to produce in this furnace. An annealing process in a 100 % hydrogen atmosphere with improved convection has been developed more recently. This gives shorter annealing times (mainly owing to the rapid cooling), and more uniform mechani-

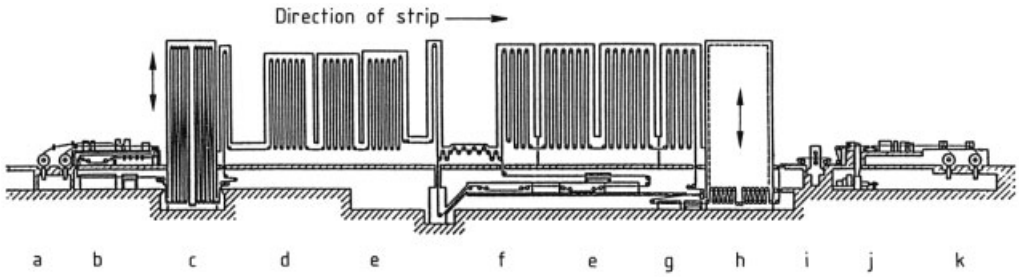

Figure 162. Continuous annealing plant
a) Pay-off reel; b) Strip cleaning; c) Entry looper; d) Heating zone; e) Tempering zone; f) Reheating zone; g) Cooling; h) Exit looper; i) Skin-pass rolling; j) Inspection/oil treatment; k) Take-up reel

cal properties. The hydrogen easily diffuses into the coil windings, and its conductivity is a factor of seven greater. Moreover, the smaller H_2 molecules enable more rapid circulation of the atmosphere (high convection, Fig. 161).

The continuous strip annealing process was developed in Japan in the 1970s (Fig. 162). At the feed end there are: a double-feeder pay-off reel, a welding machine, a cleaning zone, and a storage looper (required if coils need to be welded together to give continuous operation). The strip then passes through the heat treatment zone: heating chamber, annealing chamber, cooling unit, tempering zone, and second cooling chamber. These are followed by the exit looper, the skin pass mill, the side-cutting shears, the inspection area, the cut-to-length line, and the double-coiling reel. As the treatment time for the material is short (ca. 10 min), the annealing temperature is higher than in the batch annealing process.

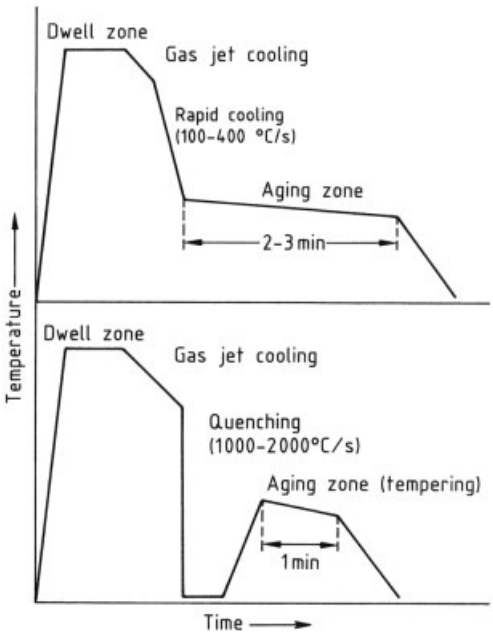

Figure 163. Heating cycles of continuous annealing

There are four continuous annealing processes which compete worldwide, differing mainly in the cooling equipment used, and the cooling rates achieved [242].

The recrystallizing annealing process eliminates the hardening produced by cold working, but the mechanical properties of the annealed strip are poor: low tensile strength, low yield point, high fracture strain, etc. Annealed thin strip has a marked upper yield point and a large extension at the lower yield point.

There are two basically different annealing cycles for continuous annealing furnaces (Fig. 163). In the first, the strip is heated to the annealing temperature, held at this temperature, cooled to the dwell temperature, held again, and then cooled to room temperature. The other alternative consists of heating to the annealing temperature, cooling with gas jets to an intermediate temperature, and quenching in water to room temperature. The strip is then brought up to the aging temperature, and finally cooled again to room temperature. Treatment at the aging temperature is not necessary for dual-phase steels.

Various steel grades can be produced in these two annealing cycles.

Quenching is produced by the rapid cooling of the strip in the continuous annealing process, so that the carbon remains in solution. For this reason, continuous annealing is very suitable for producing high-strength, dual-phase, and IF (interstitial-free) steels.

4.3.5. Skin-Pass Rolling

As stated above, the annealed material must be rolled (skin-pass rolling) to prevent flow lines owing to the distinct upper and lower yield points. This also produces the roughness which customers require. This is of great importance for deep drawing or coating the steel. For example, if complex deep-drawn parts are required, e.g., for automobile construction, the surface of the steel should be rough in order to retain the lubricant. Good coating properties require a surface with a large number of peaks, which should be randomly distributed.

4.3.6. Stainless Steels

The processing of stainless steel in a cold rolling mill is quite different (Fig. 164). Austenitic stainless steel behaves differently from ferritic stainless steel. The hot-rolled strip is first annealed. This is necessary because, after the hot rolling process, the strip cools slowly, and carbon tends to precipitate at the grain boundaries,

owing to its reduced solubility at lower temperature. Annealing causes the carbon to go back into solution. Rapid cooling of the strip then prevents the carbon from precipitation by diffusion. Although the annealing time required to bring the carbon into solution is short for austenite, a long time is necessary for ferritic steels, so that batch annealing is necessary. In the continuous furnace, annealing is carried out in an oxidizing atmosphere to produce a scale with a high oxygen content, thereby assisting the pickling process. This does not apply to high-grade steels. The strip is then quenched, by air – water cooling.

Annealing in the hood furnace is successful because of the increase in throughput obtainable under the reducing atmosphere. Considerable increases in throughput have been achieved by the use of hydrogen, even though its concentration is only 25 %.

A high proportion of the scale is removed by abrasive blasting. Although the abrasive particles are mainly spheroidal, a high degree of roughness is produced. This must be removed by rolling, another cause of the high deformation strain in high-grade steel mills.

Sodium sulfate is used in electrolytic pickling. The electrolytic pickle liquor is in effect self-regenerating.

Conventional pickling is carried out with a mixture of nitric (12 – 14 %) and hydrofluoric acid (2 – 4 %). Pickling temperatures are usually 40 – 60 °C. Unlike sulfuric acid or hydrochloric acid pickling solutions, regeneration is not possible, so that the spent acids must be rendered harmless and disposed off.

After pickling, the strip is usually rolled on "Sendzimir" rolling mills (more rarely on four- and six-high stands).

The strip is then annealed to convert the steel from the work-hardened state produced by cold rolling into a recrystallized and stress-relieved state, to enable further treatment of the strip to be carried out.

Stainless steels are to an increasing extent being annealed in bright annealing plants [243], as this process enables the bright surface produced by cold rolling, and typical of special steels, to be retained (Fig. 165).

The strip is finally rolled again (skin-pass rolling).

4.3.7. Future Developments

The extent of automation in cold rolling mills has raised the possibility of linking the individual process steps. For instance, attempts are being made to link the pickling stage with the tandem rolling operation. This requires the outputs of the two operations to match. In Japan, a cold rolling mill is being operated in which pickling, rolling, and continuous annealing have been combined, despite the fact that the outputs of the continuous annealing process and the rolling mill do not correspond. The maximum economic output of a continuous annealing plant is 800 000 t/a in the late 1980s, while tandem rolling mills can easily achieve twice this production rate.

5. Surface Coating

5.1. Introduction

Technical developments have greatly increased the output of surface- coated steel strip. The demand for improved corrosion protection is not the only driving force. The properties of coated strip, such as the yield point, expansion properties, and cold workability needed to be improved before these products could be used, e.g., in the automotive industry. New processes had to be developed to improve welding and paint application properties for large-scale industrial manufacture. These involved close cooperation between the producers and consumers of coated steel sheet. Within the steel industry, the continuous surface coating processes used include hot dipping, electrolytic techniques, vapor treatment, and coil coating.

The status of surface coating in relation to other production processes within the steel industry is shown in Figure 166.

The design of surface coating plants for wide steel strip essentially consists of the mechanical entry section, the process itself, and the exit section. The input and output sections are provided with pay-off reels, coiling reels, leveling machines, shears, welding machines, loopers, regulating and pulling rolls, etc.

This equipment produces a continuous strip, and enables treatment to be carried out at constant strip speed, under stable process conditions. Inspection areas for quality assurance are usually situated at the exit end.

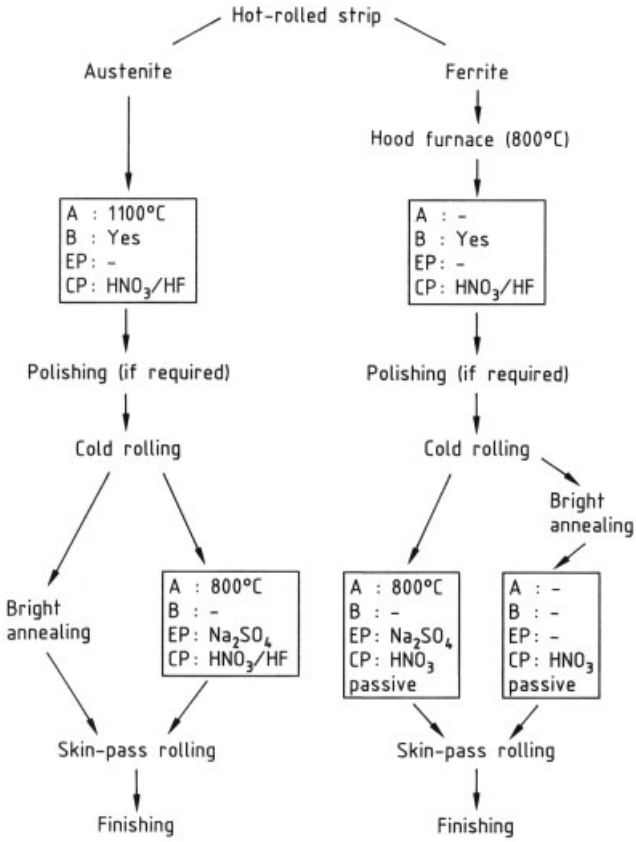

Figure 164. Flow diagram of a combined annealing/pickling line
A = annealing; B = abrasive blasting; EP = electrolytic pickling; CP = chemical pickling.

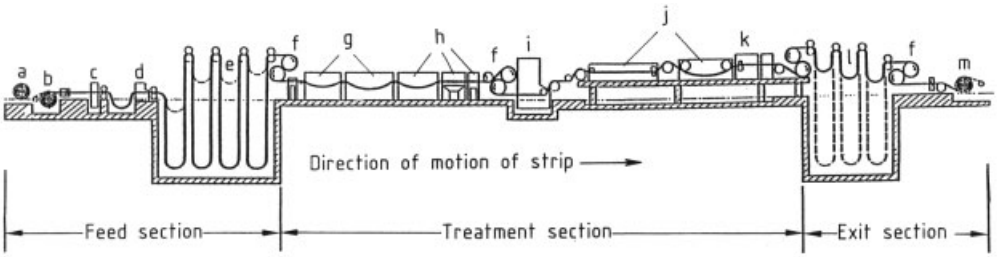

Figure 165. Annealing/pickling line for stainless steel
a) Pay-off reel 1; b) Pay-off reel 2; c) Welding machine; d) Trimming shears; e) Looper 1; f) Roll stand; g) Furnaces; h) Cooling zone; i) Abrasive blasting; j) Pickling; k) Spray washing; l) Looper 2; m) Take-up reel

As well as coated strip, the steel industry also produces clad steel sheet. This is a composite material that combines the properties of a carrier material which is usually cheap (e.g., carbon steel) with those of a metal more suitable for the conditions of use (e.g., corrosion-resistant steel).

5.2. Hot-Dip Coating

5.2.1. Plant and Processes

In the hot-dip coating process, the steel strip is continuously passed through molten metal. An

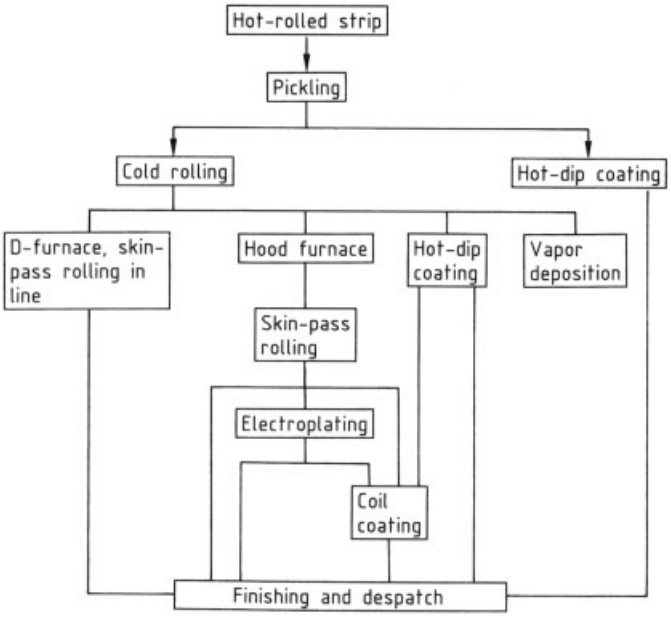

Figure 166. Surface coating in relation to other production processes

alloying reaction between the two metals takes place, leading to a good bond between coating and substrate.

Cold-rolled wide strip is usually used for hot-dip coating, but this is contaminated with rolling lubricant emulsion and of abraded iron fines. To ensure a good bond between the metal coating and the substrate, the surface of the strip must be thoroughly cleaned before dipping in the molten metal.

The important parts of the hot-dip coating process (Fig. 167) are [244]: the continuous furnace; the treatment zone where the metal coating is applied; cooling and final treatment.

Older plants usually have a horizontal continuous furnace which comprises a directly heated preheating furnace and indirectly heated reduction and holding zones with reducing H_2/N_2 atmospheres, followed by cooling zones. In the preheating furnace, the strip is rapidly heated to $> 550\,°C$, and is cleaned by burning off the oil

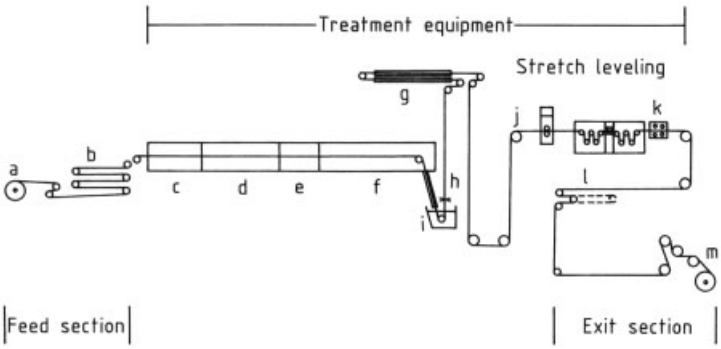

Figure 167. Hot-dip coating plant
a) Pay-off reel; b) Entry looper; c) Preheater; d) Reduction zones; e) Holding zone; f) Cooling zones; g) Strip cooling; h) Nozzle; i) Molten metal bath; j) Skin-pass rolling; k) Final chemical treatment; l) Exit looper; m) Take-up reel

emulsion residues. In the reduction and holding zones, the strip is heated to cause recrystallization or normalization, according to the grade of steel, to give the cold-rolled, work-hardened substrate material the desired mechanical properties.

For normal grades of thin steel strip, temperatures are 700 – 750 °C, and for IF (interstitial free) steels > 800 °C.

The strip is cooled in the cooling zones to a temperature slightly above that of the molten metal, with static cooling elements and rapid coolers.

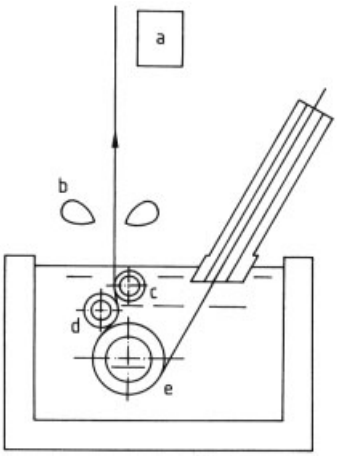

Figure 168. Control of coating thickness
a) Coating thickness measurement; b) Jet processing nozzle (slit), c) Pass line roll; d) Stabilizing roll; e) Bottom roll

After leaving the continuous furnace, the cleaned and heated strip, under a protective gas, is fed by means of a so- called snout into the molten metal (Fig. 168), which is contained in a heated metal or ceramic vessel. Immediately after this, jet processing is carried out by blowing a gaseous medium through a slit, of the same length as the width of the strip, onto the upper and lower sides of the strip. This enables different coatings to be produced on the two sides. The coating thickness (7 – 45 µm, or 50 – 300 g/m^2) is a function of gas pressure, strip speed, nozzle – strip distance, nozzle aperture, jet angle, and the viscosity and density of the metal being removed.

The thickness of the metal coating is continuously monitored, and controlled by adjustments to the jet processing equipment.

After jet processing, the liquid metal coating passes through air cooling equipment, so that it solidifies before it comes into contact with the first roller that deflects it from the vertical.

The strip then passes through more processing equipment. It is rolled in the finishing stand to improve its technical properties or to give a uniform surface appearance. The flatness of the strip is then improved by stretch – leveling equipment.

Further chemical treatment – chromate passivation and/or oil treatment, for temporary protection of the surface against "white rust" is sometimes carried out.

In new plants (Fig. 169), the continuous furnace is usually vertical, which reduces the number of contacts between the easily damaged strip and the carrier rollers. These Plants also have alkaline cleaning equipment, with brushing and rinsing zones to remove both the emulsion residues and most of the particles of abraded

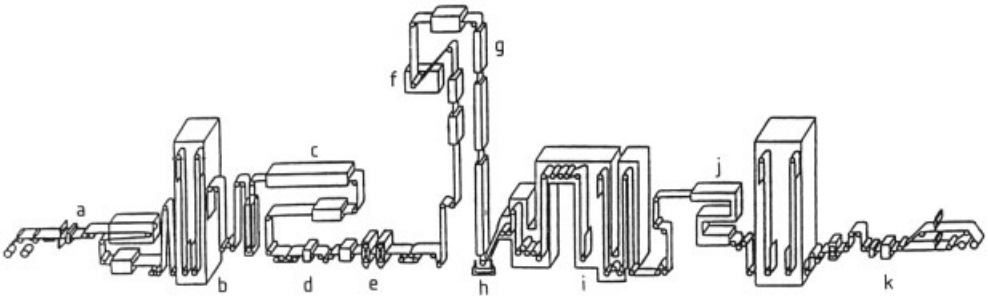

Figure 169. Hot-dip zinc coating plant at the Kawasaki Mizushima works
a) Exit section; b) Chromate passivation; c) Iron flash (electrolytic); d) Stretcher – roller leveling equipment; e) Measurement and control equipment; f) Skin-pass rolling; g) Zinc bath; h) Galvannealing furnace; i) Furnace; j) Pretreatment; k) Entry section

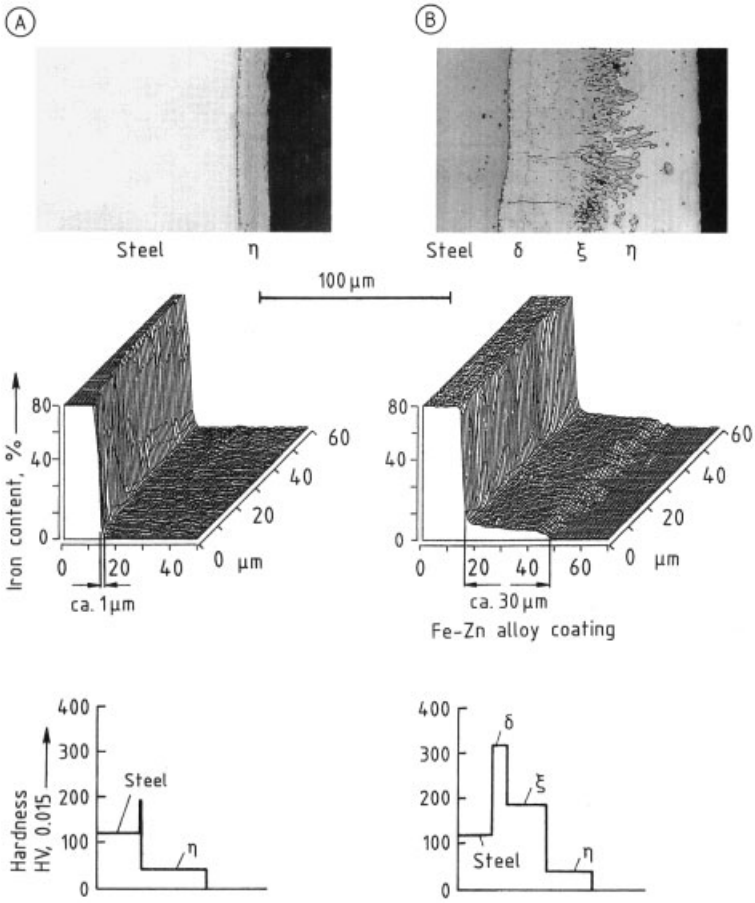

Figure 170. Coating structure, Fe content, and hardness of hot-dipped steel
A) Continuous galvanization; B) Batch galvanization

iron. The preheater in these plants is usually indirectly heated, as the emulsion residues are removed in the alkaline cleaning.

5.2.2. Other Coating Processes

Up to the late 1960s, only zinc and aluminum were used to coat steel. Corrosion resistance and working properties were later improved by using coatings based on a composite zinc – aluminum system. Table 19 lists the most important coatings, with compositions, coating thicknesses, and bath temperatures.

Zinc in the presence of moisture provides cathodic protection of the steel substrate. This is effective, even in the region of the uncoated edges, up to a steel thickness ca. 1 mm.

Aluminum has a dense outer oxide layer which reduces loss rates in various atmospheres, and gives better corrosion protection, especially at higher temperatures. The zinc – aluminum combination is selected to give the best coating properties for each application.

In the hot-dip zinc coating process, when the oxide-free steel surface is dipped, a multilayer Fe – Zn diffusion zone is formed which is then covered with a layer of zinc on removing the sheet from the bath. In batch galvanizing, uncontrolled growth of the Fe – Zn layer leads to a thick and very brittle coating with several component layers (Fig. 170) that does not readily undergo forming processes. On the other hand, a continuous zinc coating process involves short contact times between the strip and the molten

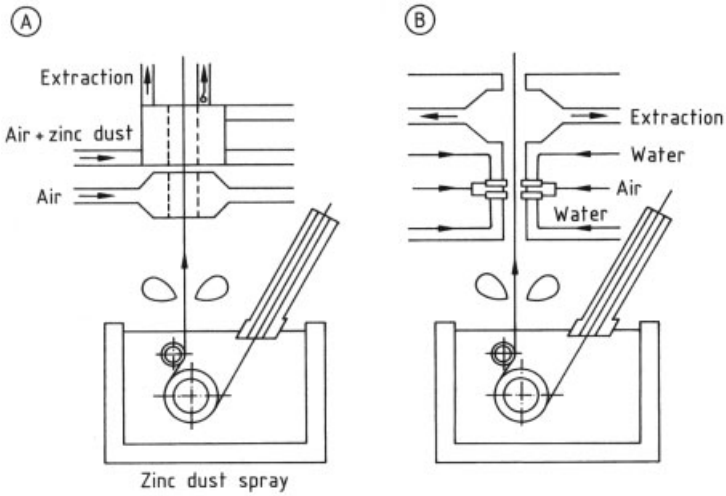

Figure 171. Control of crystal size in hot-dip zinc coatings by adding foreign nuclei
A) Zinc dust spray; B) Water – air spray

Table 19. Composition and temperature of hot-dip coating baths and coating thicknesses produced

Coating	Zn, %	Al, %	Si, %	Bath temperature, °C	Coating thickness, µm
Zinc coating	99.5				7 – 45
Monogal	99.5	0.12 – 0.30		450 – 480	7 – 15
Galvannealed	99.5				7 – 15
Galfan	95	5		420 – 450	7 – 25
Galvalume	43.4	55	1.6	600 – 620	13 – 28
Aluminum coating (type 1)		90	10	660 – 680	17 – 45

zinc. Aluminum is added to reduce the speed of the Fe – Zn reaction, leading to a relatively thin alloy layer. Compared with batch processes, this composite system permits an extreme degree of deformation without detachment of the coating.

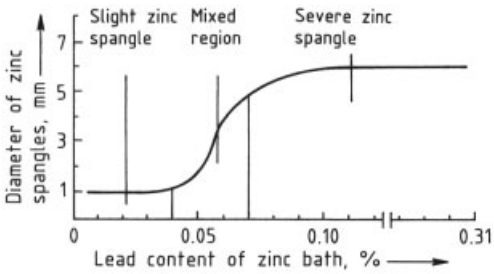

Figure 172. Zinc spangle diameter as a function of lead content of the coating bath

When the zinc coating solidifies, the well-known zinc spangle appears, caused by lead in the zinc bath (usually 0.10 – 0.15 %). Hot zinc-coated thin strip with good forming and paint application properties must have zinc crystals that are as small and uniform as possible over the length and width of the strip. Crystal size can be controlled by nucleation or by the composition of the bath.

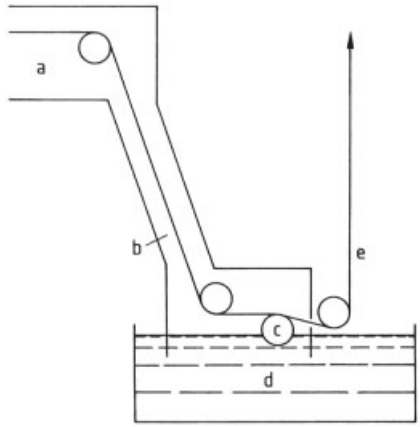

Figure 173. Coating steel strip on one side only (Nippon Steel)
a) Furnace; b) N_2 atmosphere; c) Coating roller; d) Zinc bath; e) Stripping nozzle

Control of the number of nuclei by the introduction of foreign nuclei has been used for some years (Fig. 171). Nuclei in the form of water, water vapor, water-soluble salts, or zinc dust are blown onto the liquid coating surface. The success of the operation depends on uniformity over the whole width of the strip.

The third method of reducing zinc spangle is to reduce the lead content (Fig. 172). As the solubility of lead in the zinc crystals is low, separate crystals tend to appear. If the lead content is $< 0.05\,\%$, the crystals are hardly visible macroscopically.

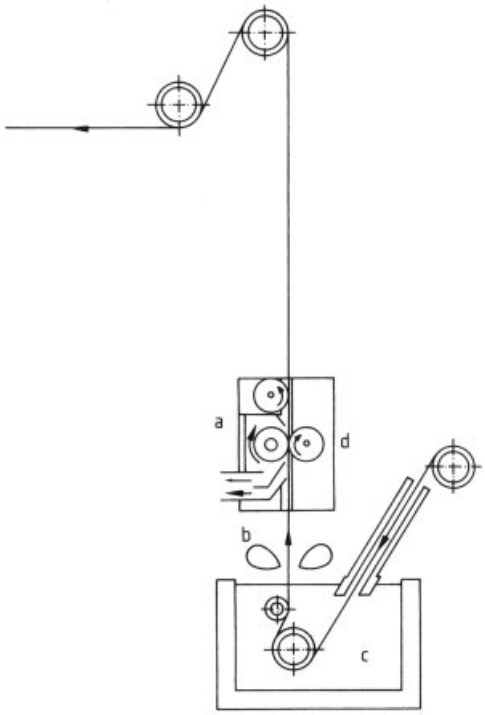

Figure 174. Hot-dip zinc coating by the Monogal process
a) Reverse-direction metal brush with zinc dust extraction; b) Nozzles; c) Zinc bath; d) Pass line roller

Steel strip coated with zinc on one side only, the other side having similar properties to those of cold-rolled strip, is required in small amounts by the automotive industry.

In the process shown in Figure 173, liquid zinc is transferred to the strip under a nitrogen atmosphere by means of a carrier roll. In the Monogal process, the zinc coating is removed from one side of the strip by brushing in the opposite direction with a metal brush (Fig. 174), and is removed by a fume extraction system. A residual layer remains on the brushed side, consisting of ca. $10\,g/m^2$ material in the form of an Fe – Zn layer. The result resembles material that has been coated by applying a hot melt to one side only.

In the Galvannealed process (Fig. 175), online treatment of the strip is carried out immediately after zinc coating. After solidification, the coating ($30 - 90\,g/m^2$, $5 - 12\,\mu m$) is heated briefly at $460 - 600\,°C$, inductively or with gas. This converts the zinc coating to a penetrating Fe – Zn alloy layer, which should have Fe content $9 - 11\,\%$ for maximum forming properties. Compared with a pure zinc coating, Fe – Zn alloy coating gives improved welding and paint adhesion properties. If the Galvannealed strip is not painted, its corrosion resistance is less good.

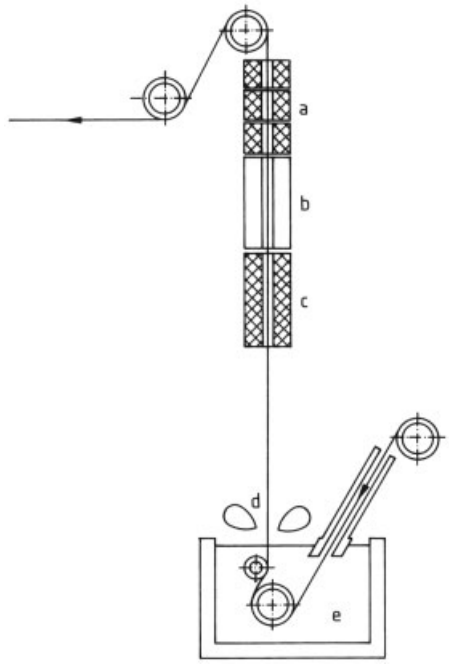

Figure 175. Hot-dip zinc coating by the Galvannealed process
a) Cooling zone; b) Holding zone; c) Heating zone; d) Nozzles; e) Zinc bath

A coating with the trade name Galfan was developed by the Centre de Recherches Métallurgiques in 1980. The zinc is alloyed with ca. $5\,\%$ aluminum to give a eutectic composition with melting point ca. $380\,°C$. Wettability of the steel strip is improved by including small

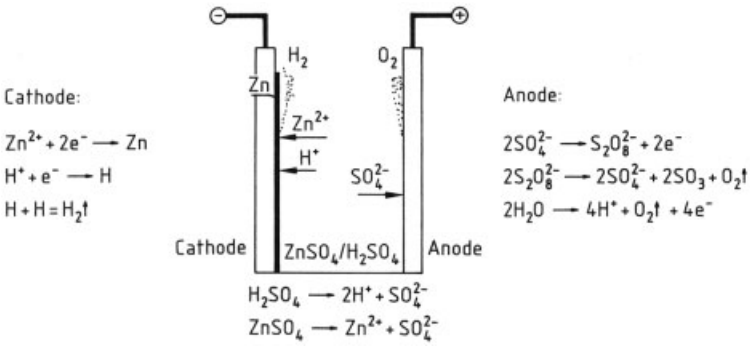

Figure 176. Electrolytic coating with zinc

amounts of cerium and lanthanum in the melt bath. The Galfan coating has a lower loss rate and improved forming porperties (for comparable coating thicknesses) when used without a paint coating in an industrial atmosphere.

An Al – Zn alloy (ca. 55 % Al, ca. 43 % Zn, and 1.6 % Si) was introduced by Bethlehem Steel in 1963 under the trade name Galvalume. It has a low loss rate, especially in maritime climates.

5.3. Electrolytic Coating (Electroplating)

5.3.1. Plant and Processes

In electrolytic coating (electroplating), the steel strip, which has been annealed and rolled (skin-pass rolling), is continuously passed through an aqueous electrolyte containing zinc ions (Fig. 176) [245]. An electric current causes the zinc ions to be deposited on the steel strip, which forms the cathode. The process takes place at 55 – 75 °C, so that there is no alloying with the steel strip, and the mechanical properties of the substrate are not affected. The desired coating thickness is achieved by controlling current and strip speed.

In a electroplating plant (Fig. 177), the equipment at the feed end is similar to that used for hot-dip coating, and only those features that differ from hot dipping are described here.

The surface of the cold-rolled strip must be clean so that an adherent, optically defect-free metallic coating is produced. Electrolytic coating plants have facilities for cleaning the strip, normally a combination of spray, brush, and electrolytic cleaning with alkaline cleaning solutions at 80 °C.

In cleaning facilities that use chemical solutions, multistage rinsing with brushing equipment is always provided, so that residues from the treatment solutions are removed as completely as possible.

After cleaning, many plants include stretch – leveling equipment to ensure that the strip is flat. This is necessary so that the distance between the anode and the strip can be kept small, while avoiding contact between the two (short circuits), thus minimizing electricity costs.

In the electrolytic pickling process, the strip is activated before the zinc coating stage. After cleaning and pickling, the strip enters the coating facility, in which the number of electrolytic cells is sufficient to enable the production rate to match that of the rest of the plant.

The strip can be treated to give additional temporary corrosion protection, most commonly by phosphating and/or chromate passivation.

A strip drier is always provided at the exit end of the treatment zones of an electrolytic coating plant. In the discontinuous part of the exit area, there are usually trimming shears to ensure the correct strip width.

5.3.2. Electroplating Cells

In steel strip electroplating plants, coating takes place in the electroplating cells. The strip is connected to the negative pole of a rectifier, and forms the cathode. The conduction rollers have

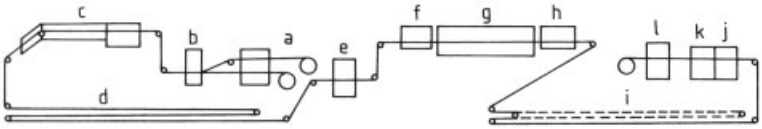

Figure 177. Electrolytic coating plant
a) Pay-off reels; b) Welding machine; c) Alkaline cleaning; d) Entry looper; e) Stretch-leveler; f) Pickling; g) Electroplating; h) Aftertreatment; i) Exit looper; j) Trimming shears; k) Oil treatment; l) Take-up reel

a highly wear-resistant cladding of special steel, and have to be cooled, owing to the high current which they carry. Compression rollers ensure good electrical contact. The anodes are situated on both sides of the strip, as near the strip as possible.

There are two types of anode:

1) Solid anodes of the metal being coated dissolve at a rate proportional to the current and must be replaced at regular intervals.
2) With insoluble anodes made of lead alloys, or titanium plates coated with noble metals. Zinc solutions must be added to the electrolyte.

To produce zinc or alloy coatings, sulfuric or hydrochloric acid electrolytes and soluble anodes are most often used. Addition of conducting salts, e.g., Na_2SO_4 or $(NH_4)_2SO_4$, ensures high conductivity, with current densities $> 100\,A/dm^2$.

Three types of cell can be used for electroplating (Fig. 178); horizontal and vertical cells are used for coating one or both sides, while radial cells coat one side only.

5.3.3. Coating Types

Zinc Electroplated Thin Steel Strip. Corrosion protection increases with increasing coating thickness, but workability and weldability deteriorate. A coating thickness of ca. 7.5 µm represents a good compromise.

Zinc – Nickel Coatings with 11 ± 1 % Nickel. This material has good corrosion properties, but mechanical properties are sometimes inferior to those of pure zinc coatings. Sealed Zn – Ni coatings (Durasteel) give improved workability.

Zinc – Iron Coatings. These give cathodic protection up to an iron content of 85 % in the coating. Anticorrosion and workability properties are good with iron content 15 – 25 %, while good paint application properties are obtained with iron content > 50 %. This has led to the development of a two-layer coating, the first with 10 – 25 % Fe, (the rest being Zn), the outer coating containing > 60 % Fe.

As well as these established types of coating, Zn – Co and Zn – Mn alloys are being investigated. Cobalt also shows interesting properties, even at low concentration, and can be plated at high current densities. Manganese is of great interest owing to the formation of a protective surface layer of Mn_2O_3.

Tin and Lead – Tin Coatings. Lead – tin coatings containing 10 % Sn (Terne) are mainly used to plate steel sheet for gasoline tanks. It has indifferent electrochemical behavior towards chemicals, but displays good corrosion behavior as well as good paint application properties for selected paint systems.

Tinplate is steel strip, 0.1 – 0.5 mm thick, with a tin coating of ca. 2.4 g/m^2. A solution of tin fluoroborate can be used as electrolyte. The thin strip, which forms the cathode, is passed between two rows of thin anodes. After coating, the strip is heated above 232 °C (*mp* tin), and quenched in water to give the brilliant luster of electrolytically produced tinplate. Strip speeds > 650 m/min can be reached. Tinplate is used in packaging (painted, foil- covered, or printed).

Electrolytic chromium- coated steel (ECCS) is a special type of chromium-plated thin steel strip, and is produced in a similar process. In many applications, ECCS has replaced tinplate.

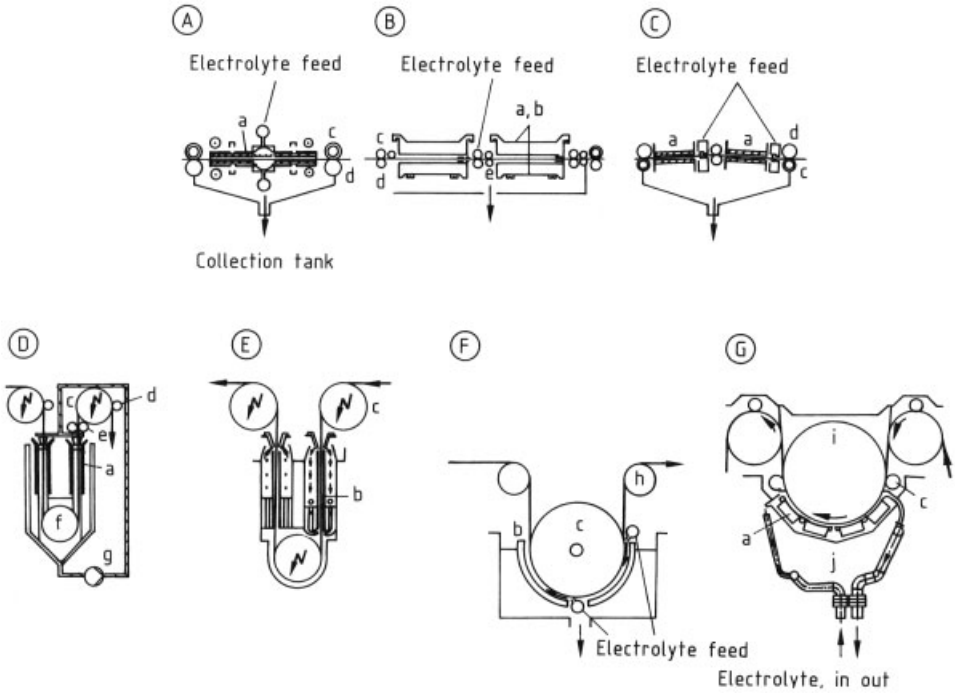

Figure 178. Electroplating cells
A) LCC-H (insoluble electrode, sulfate electrolyte, horizontal cell); B) NKK (soluble/insoluble anode, chloride/sulfate electrolyte, horizontal cell); C) Krupp/SEH (insoluble anode, sulfate electrolyte, horizontal cell); D) Gravitel (insoluble Ti anode, sulfate electrolyte, vertical cell); E) HS (soluble anode, sulfate electrolyte, vertical cell); F) Carosel (soluble anode, chloride electrolyte, radial cell); G) KC (insoluble anode, sulfate electrolyte, radial cell)a) Insoluble anode; b) Soluble anode; c) Current- carrying roller; d) Pressure roller; e) Squeeze rollers; f) Dipping rollers; g) Pump; h) Deflecting roller; i) Top roller; j) Edge mask

5.4. Vacuum Vapor Deposition

5.4.1. Plant and Processes

In the vacuum vapor deposition process, the coating metal is deposited onto the "cold" steel strip from the vapor phase [246]. Vacuum vapor deposition is not yet established on a large scale in the steel industry, and is operated at only one location so far, where a hot-dip plant has been converted to a vapor deposition plant (Figs. 179, 180).

Vaporization is simply carried out in the zinc melting vessel. The plant has only a directly heated preheater, without alkaline cleaning. This degree of cleanliness of the strip is often adequate for the vapor process. Coating takes place in two chambers, both operating at 2.66 Pa; the vaporization capacity of the first is 350 kg/h, and that of the second chamber is 500 kg/h. Molten zinc at 460 °C is pumped into the vaporization chamber where it immediately vaporizes at the low pressure. The strip must be cooled to ca. 250 °C in the vacuum chamber so that the zinc precipitates onto the cold strip. The operating pressure is reached by successive reduction of the atmospheric pressure by squeeze rollers.

While the hot-dip process is mainly suitable for producing zinc coatings with thickness $> 60\,g/m^2$, zinc vaporization produces coatings with thickness ca. $30\,g/m^2$, often more cheaply than the electrolytic process.

5.4.2. Further Developments [247], [248]

Vacuum vapor technology is mainly suitable for coating with pure metals, alloys, chemical compounds, and composite layers that cannot be produced by hot-dip or electrolytic coating. Potential coating materials under investigation are listed below:

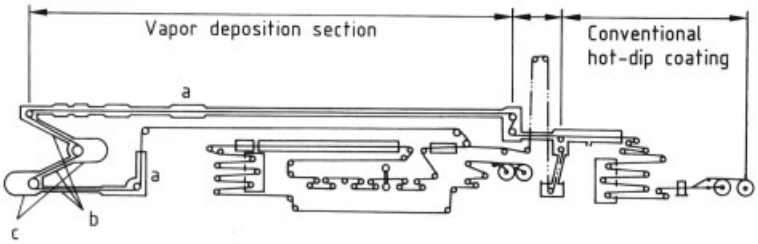

Figure 179. Zinc vapor deposition plant at the Nisshin Sakai works
a) Gas jet cooler; b) Vacuum system; c) Vaporization chambers

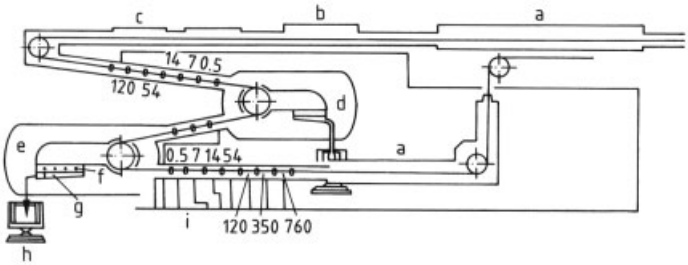

Figure 180. Vacuum system for the zinc vapor deposition plant
a) Gas jet cooler; b) N_2 gas; c) Edge heater; d), e) Vaporization chambers (2.66 Pa); f) Heating coils; g) Vaporizer; h) Zn melt; i) Vacuum pumps

Metals
- Zinc
- Aluminum
- Chromium
- Titanium
- Nickel
- Copper
- Silicon

Alloys
- Nickel – chromium
- Zinc alloys, e.g., Zn – Mg, Zn – Cr, Zn – Al, Zn – Ni, Zn – Ti

Chemical compounds
- TiN, TiC
- Al_2O_3
- SiO_x

Composites
- Al – Ti, Al – Cr
- TiN – Cr
- Zn – Al
- Zn – Zn alloys

With a sufficient number of vaporizing chambers vapor coating could be carried out at rates of 350 – 400 m/min. A vaporizing plant could be incorporated into an existing continuous annealing furnace, adding to the production capacity for coated steel strip. As only one side is coated in each chamber, it would be possible to coat either one side or both.

5.5. Coil Coating

5.5.1. Plant and Processes

In the production of coil- coated thin sheet [249], the substrate material can be cold-rolled strip, zinc- coated thin sheet (hot-dip or electroplated), aluminum strip, or special steel strip. Liquid paint or foil is applied to these carrier materials by rolling.

A coil coating plant (Fig. 181) consists of the feeding equipment, strip cleaning, chemical pretreatment, coating areas for primer and topcoat with the associated drying ovens, and the exit equipment. The strip is cleaned by methods similar to those used in zinc electroplating, but in this case the surface is then phosphated or chromated. For this, the solutions are sprayed at ca. 40 °C, under pressure. The strip is then cooled with deionized water and passivated. The chromate or phosphate coating provides corrosion protection for the whole system, and improves adhesion of the polymer coating. Before the strip

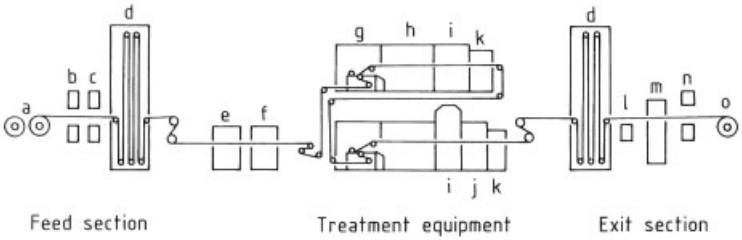

Figure 181. Coil coating plant
a) Pay-off reel; b) Cropping shears; c) Tack welding machine; d) Storage looper; e) Strip cleaning; f) Chemical pretreatment; g) Coating equipment; h) Furnace; i) Lamination equipment; j) Air cooler; k) Water quenching; l) Inspection table; m) Protective foil application; n) Edging shears; o) Coiling reel

is introduced into the coating station, this precoat is dried with hot air.

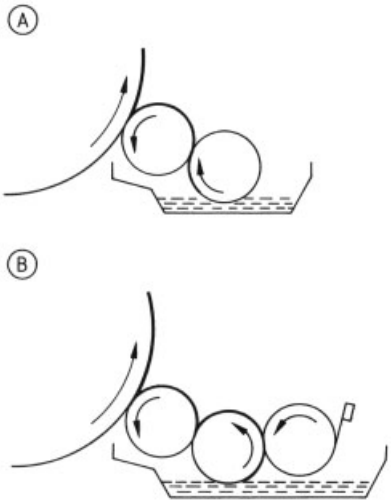

Figure 182. Typical coil coating systems
A) Two rollers, reverse rotation; B) Three rollers, complete reverse rotation

Primer or undercoat is applied at the first coating station (Fig. 182). The paint is usually applied by typical rolling machines, running in the opposite direction to the strip, one on each side of the strip, each paint roller having two or three feed rollers, depending on the coating material. As a rule, a chromium-plated dipping roller supplies the paint from a tank to a rubber-coated roller running in the opposite direction to the strip, and applies a paint film at a rate controlled by a special roller. The primer can be applied to the upper and lower side of the strip. The strip then usually passes into the first drying oven in which it hangs freely, and is heated to ca. 200 – 240 °C, depending on the coating system. The strip with its coat of primer is then cooled by air and water, and passes to the second coating station. The topcoat is applied on one or both sides, usually only the upper side. The coating is then backed in the second drying oven at ca. 240 – 260 °C. Before the strip passes to the air or water cooling area, it can be provided with a decorative and/or protective film. These laminates are applied under very precise conditions of temperature and pressure. If a decorative foil is used, an adhesive is applied instead of the top coat in the second coating station.

Coating thicknesses, which depend on the coating system, are continuously monitored over short distances by integrating thickness measuring equipment, and are maintained within very close tolerances by varying the rotation speed or the application pressure of the coating rollers.

5.5.2. Coating Systems (Fig. 183, Table 20)

The liquid coatings used are paints, based on polyester and acrylic resins, poly(vinyl chloride) plastisols, and poly(vinylidene fluoride). More recently, polyurethane paint systems have been used in special applications. Zincrometal and Inmozinc consist of thin steel sheet coated with zinc dust colors. Printed or single color poly(vinyl chloride) films are applied by rolling with an adhesive activated by hot air, and the same process is used for poly(vinyl fluoride) films. The coated or laminated material can be provided with additional protection against handling and assembly operations by means of adhesive polyethylene protective films, applied hot or cold. Coating thicknesses are usually 15 – 300 µm.

Table 20. Coating materials and thickness ranges of coating films

Coating materials	Abbreviation	Usual range of film thickness, µm
Liquid coating materials		
Standard systems		
Polyesters	SP	5 – 25
Acrylics (resin)	AY	5 – 25
Silicone-modified polyesters	SP – SI	15 – 25
Epoxy resins	EP	3 – 15
Polyurethanes	PUR	10 – 25
Poly(vinylidene fluoride)	PVDF	20 – 25
Poly(vinyl chloride) organosol	PVC(O)	30 – 60
Poly(vinyl chloride) plastisol	PVC(P)	80 – 400
Special systems		
Weldable zinc dust primer	ZP	10 – 20
Heat-resistant nonstick system	HRNS	5 – 15
Film coatings		
Polyacrylate	PMMA(F)	50 – 75
Poly(vinyl chloride)	PVC(F)	100 – 300
Poly(vinyl fluoride)	PVF(F)	38
Polyethylene	PE(F)	100 – 300

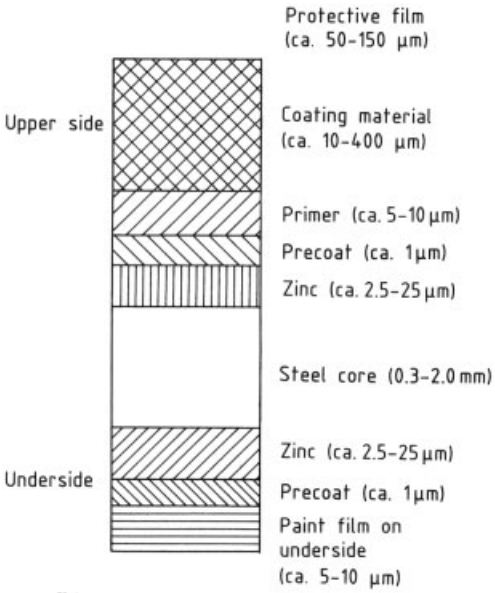

Figure 183. Cross section of coated steel strip

5.6. Roll-Bonded Cladding [250], [251]

5.6.1. Principles

In roll-bonding (cladding), two or more materials, which are usually flat, are pressure welded together. The process can be carried out with cold or preheated materials (cold roll-bonding or hot roll-bonding).

In continuous cold pressure welding of strip material, two metallic solids are bonded strongly together by high pressure, which causes plastic deformation. The adhesive bond can only be assured if the surfaces are clean and have a well defined surface structure (roughness).

The surfaces must be brought together under pressure until their distance apart is of the order of interatomic distances. As surfaces of technical quality are always covered with a very thin oxide layer, these layers must be broken by deformation. In roll-bonding, it is therefore necessary to stretch the surface, with exclusion of gas, so that new reactive surfaces are exposed by plastic flow. These are then brought into contact at interatomic distances, so that a strong bond can be formed by adhesion. The bond strength is increased by subsequent annealing homogenization.

Table 21. Substrate cladding materials for cold rollbonding

Substrate materials	Cladding materials
Iron	Corrosion-resistant steels
Steel	High-temperature steels
	Ni and Ni alloys
	Cu and Cu alloys
	Aluminum
	Noble metals
	Special metals (Ti, Ta, Mo)

5.6.2. The Process (Fig. 184)

In the production of roll-bonded steels, surface pretreatment is very important. The necessary conditions are:

1) Complete absence of grease
2) A clean, freshly activated metal surface
3) A roughened surface, produced by stretching

After pretreatment, the freshly brushed steel is fed, with the cladding material, into the roll nip, and deformed in one roll pass, the extent of deformation being 50 – 70 %.

5.6.3. Variations

In roll-bonding, a composite material is produced which often combines the special prop-

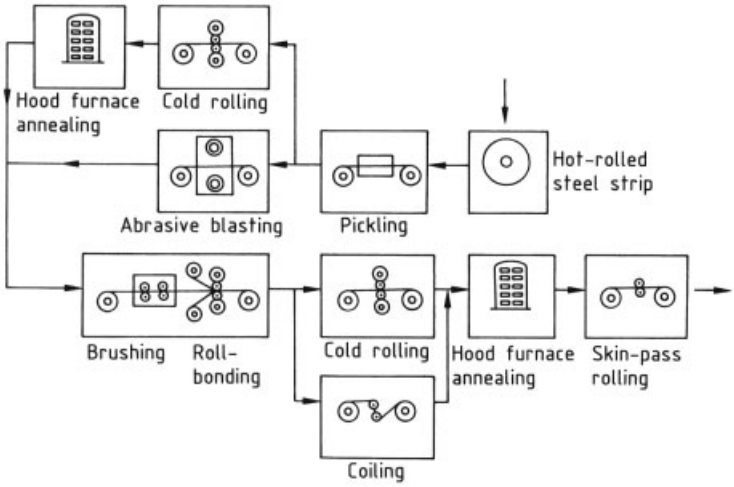

Figure 184. Production of roll-bonded steel strip

erties of the individual components in an ideal manner. For cost effectiveness, a cheap substrate material is often plated with an expensive, corrosion-resistant, acid-resistant, or heat-resistant material (Table 21).

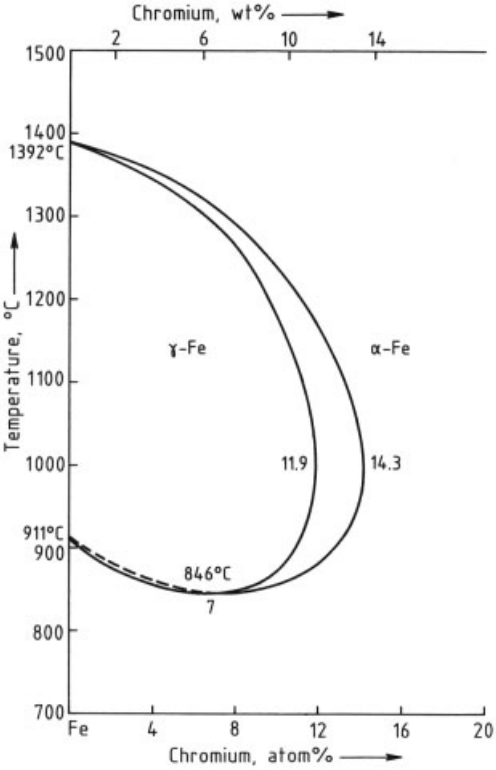

Figure 185. Phase diagram of Fe – Cr [254]

5.7. Summary

Surface-coated thin steel strip was developed to improve, e.g., the corrosion properties of the steel, or to provide a decorative appearance. The user must consider a range of properties, such as workability, welding properties, adhesive properties, and paint application properties. Each type of surface treatment has strengths and weaknesses compared with competitive processes.

Other surface treatment methods are likely to become available, including electro- and plasma polymerization, electrolytic deposition of Al and Ti from organic solutions, plasma spraying, ion implantation, and cathodic sputtering.

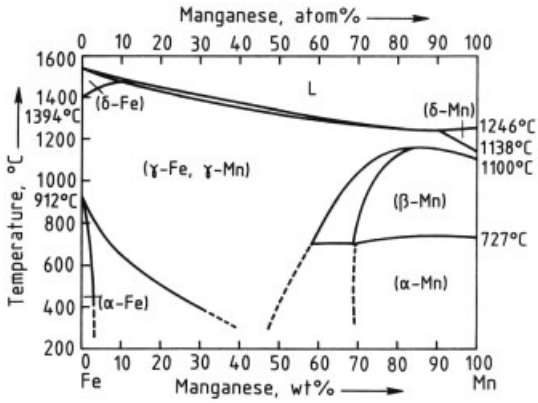

Figure 186. Phase diagram of Fe – Mn [255]

6. Uses

6.1. Fundamentals

6.1.1. Microstructure

6.1.1.1. Introduction

At room temperature, pure iron has a body-centered cubic structure. The solubility of carbon is < 0.001 wt % in this phase, which is known as α-iron. On heating, α-iron is transformed at 911 °C into the face-centered cubic γ-phase. This γ-iron can dissolve ca. 2 wt % carbon as a solid solution. On cooling this, transformation of γ-iron to α-iron occurs and carbon is precipitated as carbide Fe_3C. In the heat treatment of transformable steels, they are heated until the γ-phase is formed, so when this is cooled more or less quickly, metastable states may be produced, and the precipitation of Fe_3C, which in equilibrium is associated with the formation of α-phase, is more or less suppressed. The range of applications of steel depends on the fact that heat treatment can induce all the transitions between the metastable states, and those states effectively correspond to equilibrium conditions, so that a large number of microstructures and combinations of properties can be produced [252], [253]. Another method of modifying steel consists of adding alloying elements to stabilize the α-phase up to the melting point, or conversely to cause the γ-phase to be stable down to room temperature. These possibilities are mainly used to improve corrosion resistance or to achieve particular physical properties.

In the following discussion, important phase diagrams of the iron alloys are considered, and the microstructures and conditions needed to produce them are described. The description of the properties (Section 6.1.2) then provides an understanding of the heat treatment of steels (Section 6.1.3). In Section 6.2, the important steel groups and their properties are reviewed.

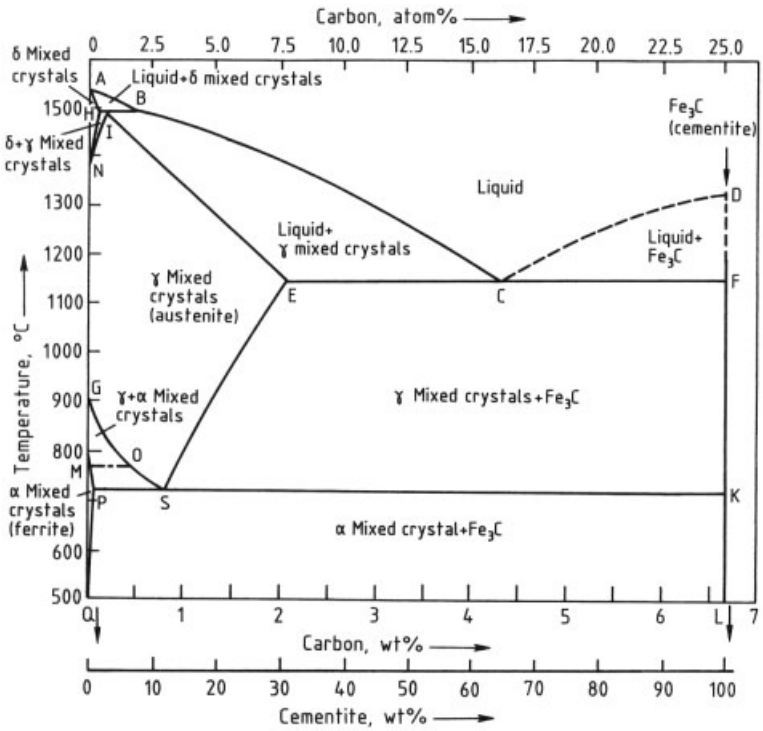

Figure 187. Phase diagram of the metastable state Fe – Fe_3C [257]

6.1.1.2. Phase Diagrams of Iron Alloys

The phase diagram of iron – chromium for low chromium content is shown in Figure 185 [254]. With chromium content > 15 wt %, γ-iron is no longer formed, and the α-iron is stable from room temperature to the melting point [255]. Figure 186 shows that with ca. 45 wt % manganese, the γ-iron is stable almost to room temperature. In commercial steels, the addition of carbon increases this stability range further. Like chromium, the alloying elements silicon and molybdenum limit the range of existence of γ-iron, and, like manganese, nickel and carbon, extend this range. However, the effect of the alloying elements is not additive. In a steel containing 18 % chromium and 10 % nickel, γ-iron is stable from room temperature to the melting point. These phase diagrams can be calculated from thermodynamic data [256], a compilation of the most important two- and three- component systems having been prepared [254], [257–262].

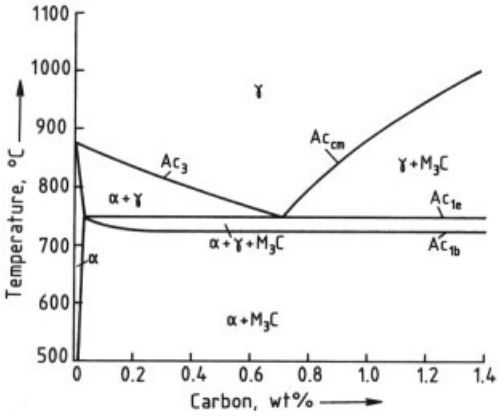

Figure 188. Quasibinary section through the three-component system Fe – M – M_3C [299]

In plain carbon steels, the metastable Fe_3C phase is of great importance (Fig. 187) [257]. This phase is stable for long periods under technical conditions of heating and cooling, especially if a solid solution $(Fe, Mn)_3C$ is formed in the presence of the usual manganese content. In Fe – Cr –C systems, $(Fe, Cr)_3C$ is stable if more than ca. 0.7 wt % chromium is present [254]. Figure 188 shows a quasibinary section through an Fe – M – M_3C system, where M represents a metal. For technical applications, the temperature limits of the phase fields during heating at 3 °C/min are measured, and are termed the Ac temperatures. The relationship of the Ac temperatures to the boundaries of the phase fields can be seen from Figure 188. Alloy steels, which are multicomponent systems containing Fe and C, can contain not only M_3C, but also the carbides M_7C_3, $M_{23}C_6$, M_2C, MC, and intermetallic phases. In these cases, the simple scheme shown in Figure 188 is not valid.

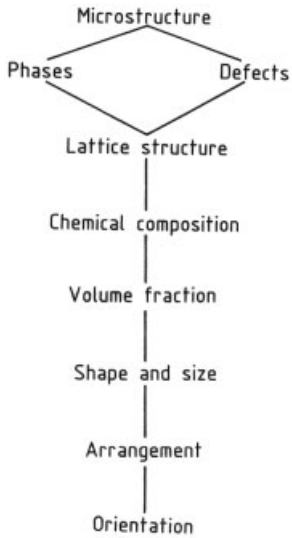

Figure 189. Parameters for the quantitative characterization of the constituents of a microstructure

6.1.1.3. Definitions

Steels are polycrystalline materials. A crystallite is a region of a phase with uniform crystal orientation. The phases which the phase diagram shows to be present or metastable states form crystallites, separated from each other by grain boundaries or phase boundaries [263]. In addition, the microstructure includes lattice defects, such as vacancies and dislocations [264], [265]. For a quantitative description of microstructure, the parameters in Figure 189 must be given for each crystallite and each defect [266–268]. A typical example of a phase is α-iron. More than one phase can exist within a crystallite region if ordered arrangements or coherent precipitates are present. The chemical composition inside a crystallite must also be given, sometimes as a function of location. The orientation must be

given for both the lattice and the spatial position of nonspherical particles. The texture within a specimen is determined by the orientation of individual crystallites [269], [270]. An example of a defect is a step dislocation [271].

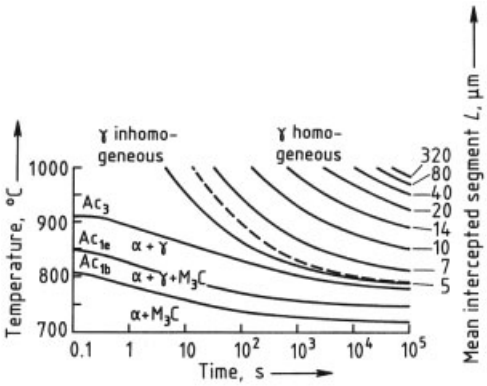

Figure 190. Time – temperature austenitization (TTA) diagram for the isothermal austenitization of Ck 45 steel, including lines of equal grain size of the austenite [299]

The geometric parameters (Fig. 189) describe a spatial arrangement of phases and defects. However, two-dimensional polished sections enable only the volume fraction and the specific surface to be determined. Three-dimensional parameters, such as particle density, require that the samples be examined in serial sections, which is very difficult [272]; in many cases, approximations are based on polished section measurements [266].

In steel microstructures, γ-iron is referred to as austenite, and α-iron as ferrite. The term cementite is used for the carbide M_3C. Sulfides, oxides, and carbonitrides, e.g., Nb(C, N), are often referred to as nonmetallic inclusions. Single-phase microstructures occur only in steels with ferrite or austenite as the stable phase. In plain carbon steels, on cooling from the austenite temperature range (Fig. 187), ferrite and carbide are formed, but, under production conditions, these are not in equilibrium. The size of such carbides ranges from a few nanometers to a few micrometers [273].

The iron – cementite phase diagram (Fig. 187) can be used as a first approximation to determine the phases present at equilibrium in plain carbon steels. In the range of carbon content below the eutectoid point S, an austenite, e.g., containing 0.3 wt % C, transforms partially to the so- called proeutectoid ferrite during cooling from 1000 °C. At temperatures below the line P – S – K, ferrite and carbide form simultaneously from the residual austenite in the eutectoid reaction. Steels of this type are known

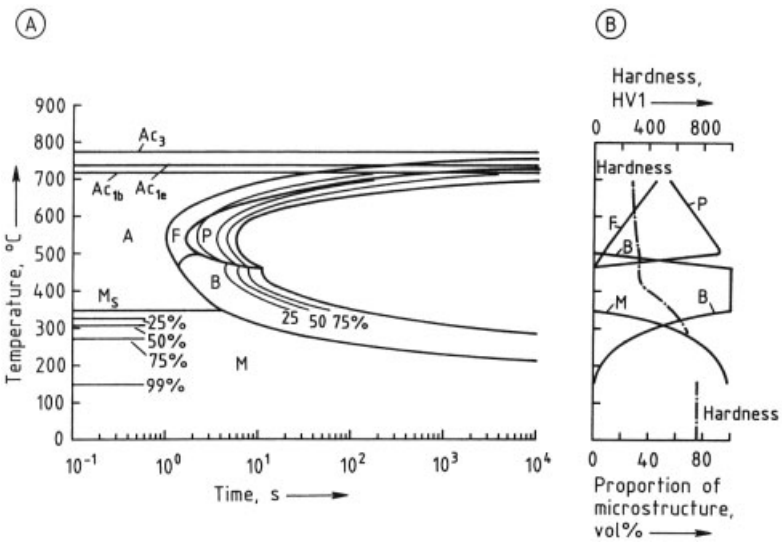

Figure 191. A) Time – temperature transformation (TTT) diagram for the isothermal transformation of Ck 45 steel [299] Austenitization: 850 °C for 15 min
B) Volume fraction of various microstructures at various temperatures and hardness after cooling to room temperatureA = austenite; F = ferrite; P = pearlite; B = bainite; M = martensite

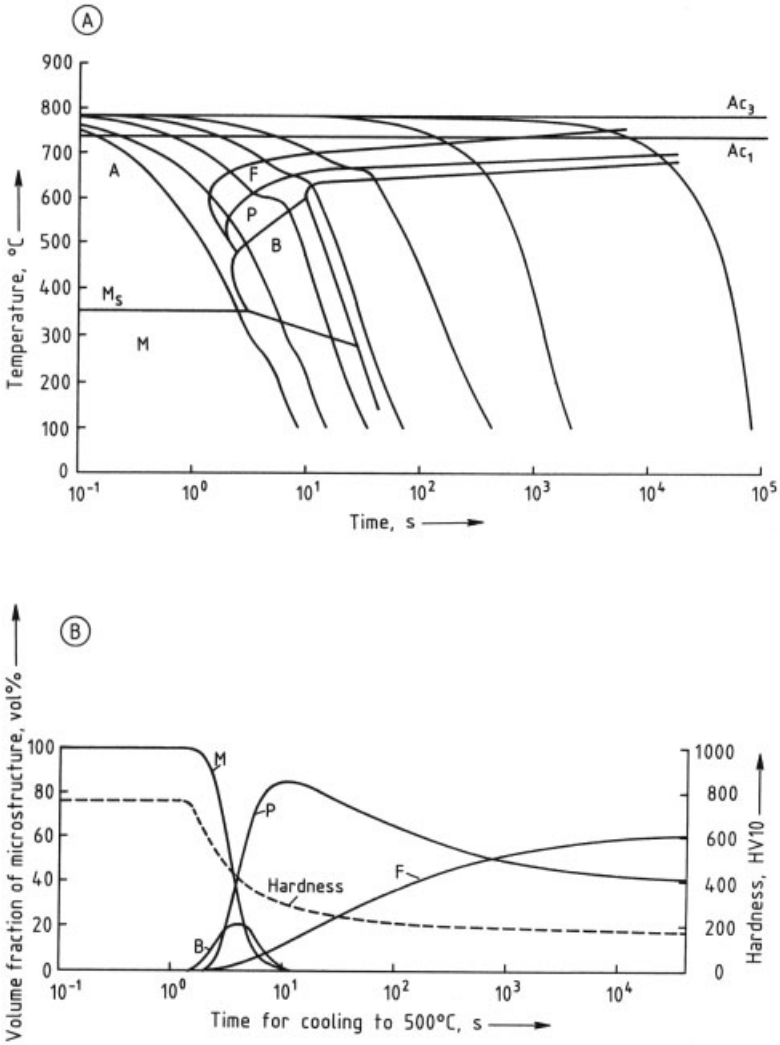

Figure 192. A) Time – temperature transformation (TTT) diagram for continuous cooling of Ck 45 steel
B) Volume fractions of the various structures [299]

as hypoeutectoid. With carbon content above the eutectoid point S (hypereutectoid steels), transformation of the austenite begins during cooling with precipitation of carbide. The carbon content at the eutectoid point is considerably altered by the addition of alloying elements [253].

6.1.1.4. Microstructures in Plain Carbon Steels

Austenite Formation. Figure 190 illustrates how austenite is formed by an arrangement of α-iron and M_3C (ferrite and carbide) in the steel Ck 45 (see Section 6.2.1) by an isothermal temperature – time treatment [274–276]. The first line Ac_{1b}, which is reached at 800 °C after 0.4 s, represents the formation of 1 vol % γ-iron (austenite). The last line Ac_3, reached at 800 °C after 2000 s, represents the formation of 99 vol % γ-iron. The area $\alpha + \gamma + M_3C$ corresponds to the three-phase equilibrium in Figure 188. Above 775 °C, the single-phase area of γ-iron (austenite) is reached only after long heating times. In practice, austenitization is carried out at temperatures 30 – 50 °C above the tem-

perature Ac_3 measured on slow heating (830 °C in the previous case). In the austenite formed, which is at first inhomogeneous, the carbon is enriched in regions where carbides have gone into solution. With increasing temperature and holding time, the austenite becomes homogeneous. With increasing austenitization temperature and time, the grain size of the austenite also increases (Fig. 190). If the austenite grain size is small, an improvement in toughness is obtained for a given yield strength (see Section 6.1.2). In practice, austenitizing is a compromise between homogeneous austenite, which requires high temperatures and long heating times, and fine austenite grain, which requires low austenitizing temperatures and short heating times.

Austenite Transformation. On cooling austenite, microstructures known as ferrite, pearlite, bainite, and martensite are formed as metastable states [276]. The formation of ferrite, pearlite, and bainite is mainly diffusion controlled. Figure 191 shows the effect of the change of the time interval between the beginning and end of the transformation in the pearlite range and the bainite range as a function of temperature. The transformations overlap [277], [278]. At 550 °C, the transformation begins after 1 s, with formation of ferrite. It ends with the transformation to pearlite, and is complete after 6 s. At 480 °C, the transformation of austenite to bainite is superseded by a transformation to pearlite. The maximum in the transformation rate to pearlite at 550 °C occurs because the nucleation rate passes through a maximum as the temperature decreases, while the growth rate (a diffusion process) continuously decreases with decreasing temperature. The martensite transformation is not diffusion controlled, so, starting at temperatures below M_s (the martensite start temperature), it proceeds at approximately sonic speed, independently of time [279].

Industrial heat treatment processes to produce a microstructure usually involve continuous cooling, for cost reasons. Figure 192 shows the time – temperature diagram for continuous cooling of Ck 45 steel. As the cooling curves are all similar, they may be characterized by quoting the duration of the cooling between 800 and 500 °C ($t_{8/5}$) or the cooling time to 500 °C. From each cooling curve, the temperatures and times can be read off for the start of formation of ferrite, pearlite, bainite, or martensite and also for the end of the transformation process [278], [280].

Ferrite. Figure 192 shows that Ck 45 steel with $t_{8/5} = 2 - 10^4$ s first forms ferrite, with carbon content fairly close to that predicted from the phase diagram. This ferrite forms either as a film on the austenite grain boundaries (Fig. 193) or as separate grains (Fig. 194) [281–283]. Its properties mainly correspond to those of pure iron.

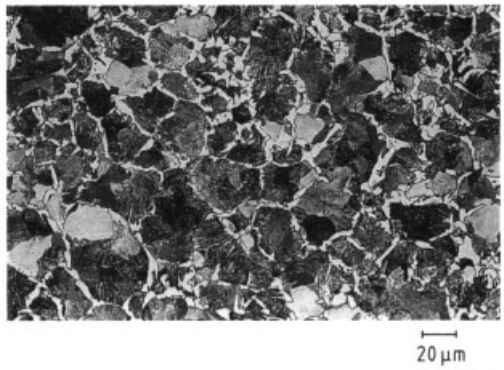

Figure 193. Ferrite (white) on the former austenite grain boundaries and pearlite (gray or black)
Ck 45 steel, 900 °C for 15 min, 530 °C in 300 s, water quench; etched with nital

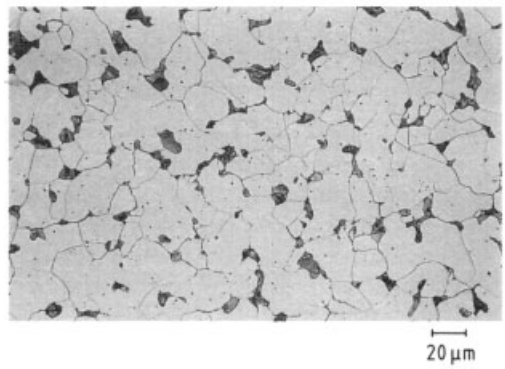

Figure 194. Globular ferrite (white) and pearlite (gray or black)
Ck 15 steel, 900 °C for 15 min, $t_{8/5} = 100$ s; etched with nital

Pearlite. After the ferrite, Ck 45 steel forms pearlite (Fig. 192), a lamellar arrangement of ferrite and cementite, formed simultaneously from the austenite (Fig. 195) [281]. This dou-

ble precipitation is much more rapid than the independent transformation of the austenite into ferrite with precipitation of cementite that would be expected from the equilibrium. The transformation to pearlite is diffusion controlled. The formation of ferrite, which has a low carbon content, causes carbon enrichment in the surrounding austenite until the formation of cementite (ca. 6.7 wt % carbon) is possible. The volume ratio of ferrite to cementite in the pearlite can fluctuate over a wide range, as can be seen from the volume fractions (Fig. 191 B). Thus, by suitable heat treatment, a microstructure of pure pearlite can be produced in steels containing 0.3 wt % carbon, just as in steels containing 1.2 wt %.

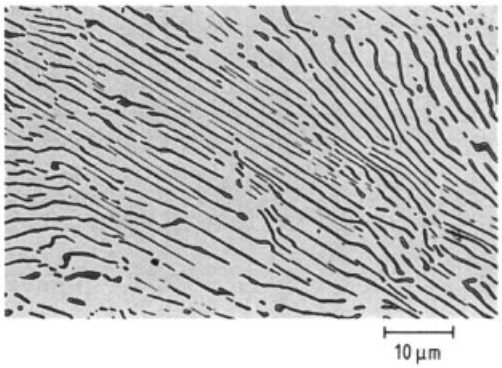

Figure 195. Pearlite, lamellar arrangement of cementite (gray or black) and ferrite (white)
C 70 steel, 800 °C for 15 min, $t_{8/5}$ = 10 000 s; etched with sodium picrate

Martensite. Martensite is formed without diffusion by an ordered movement of atoms over short distances [279]. This movement is defined by the crystal microstructures of γ-iron and α-iron. The carbon dissolved in the austenite as a solid solution remains in solution during martensite formation, which leads to very marked lattice distortion. Platelets are formed, which appear needle-like in the polished section (Fig. 196) [281]. The transformation of austenite to martensite is associated with a volume increase of 0.5 – 1 %, depending on the alloy content. Figure 192 shows that if Ck 45 steel is to form martensite only, $t_{8/5}$ must be < 1.2 s. Of all steel microstructures, martensite has the highest hardness and yield strength, although martensite with carbon content > 0.2 wt % is very brittle. For lower carbon content, so- called lath marten-site, is formed. This has very good toughness properties [284], although the hardness of martensite decreases with decreasing carbon content (Fig. 197). At a given temperature, martensite formation is incomplete. Depending on the chemical composition, steel must be cooled ca. 250 °C below M_s to transform 99 % of the austenite into martensite.

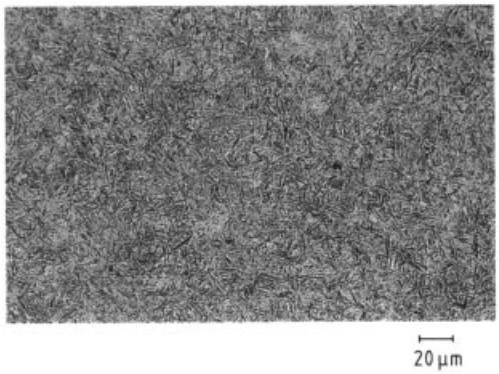

Figure 196. Martensite in Ck 45 steel
850 °C for 15 min, water quench; etched with nital

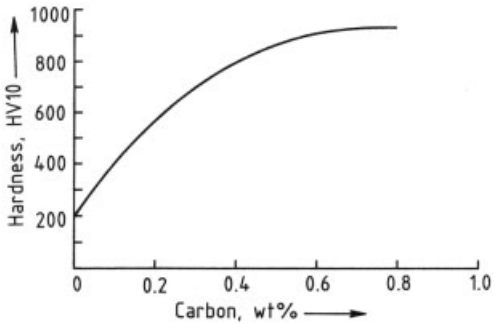

Figure 197. Hardness of martensite in plain carbon steels as a function of the content of dissolved carbon in the austenite [299]

Bainite. Bainite is formed at a temperature between that of the transformation of austenite to ferrite and pearlite and that of its transformation into martensite. The transformation is partly diffusion controlled, and partly proceeds by a lattice-shearing mechanism similar to that for martensite. By a combination of these mechanisms, the transformation of austenite to bainite can be much more rapid than the transformation to pearlite at a given temperature (see Fig. 191) [285], [286].

Bainite is an arrangement of ferrite and carbides detectable, by optical microscopy, only at high formation temperatures (Fig. 198) [281]. After formation just above M_s, the bainite is very fine (Fig. 199), and cannot be distinguished from tempered microstructures (see below). Bainite formation at a given temperature is time dependent (see Fig. 191). However, at lower M_s, the diffusion processes leading to bainite formation are slower, so that bainite formation does not proceed to completion in technically realistic times. In some steels, there are regions in which bainite formation is incomplete, even at ca. 450 °C [287].

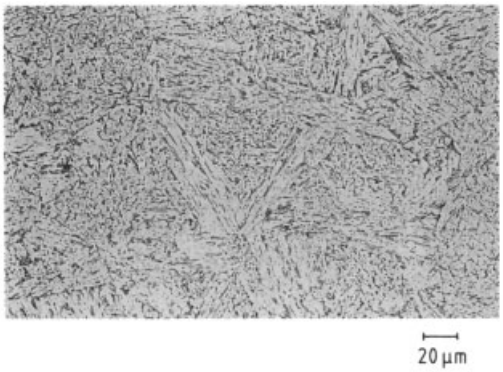

Figure 198. Coarse bainite in a steel with 0.2 % C and 2 % Mo
950 °C for 15 min, cooling to 500 °C in 100 s; etched with nital

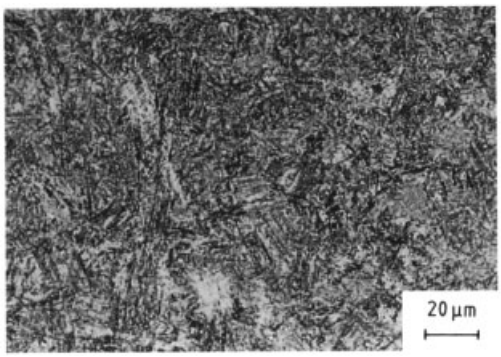

Figure 199. Fine bainite in 34 Cr 4 steel
850 °C for 15 min, cooling to 375 °C in 500 s, water quench; etched with nital

In a continuous cooling process, the various microstructures appear in sequence (Fig. 192). When Ck 45 steel is cooled to 500 °C in 8 s, ferrite is formed at 650 – 610 °C, followed by pearlite down to 575 °C, then bainite, and then martensite below 290 °C (Fig. 192). Depending on the way these microstructures grow and the temperature range involved, the various components of the microstructure are separable only rarely and with difficulty. It should be mentioned that the interparticle distances for pearlitic carbide are usually very much smaller than those for bainite formed at the same temperature (Fig. 200).

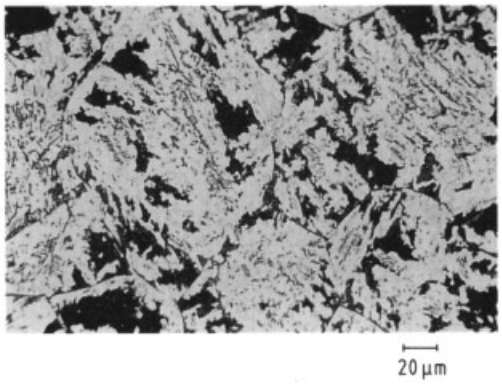

Figure 200. Pearlite and bainite in 41 Cr 4 steel [299]
840 °C for 5 min, cooling to 510 °C in 4000 s, water quench; etched with nital

Microstructures after Tempering. The carbon held in solution within the lattice on martensite formation can form carbide precipitates by diffusion if the temperature is increased [253], [276]. Precipitation takes place more rapidly and to a greater extent as the annealing temperature is increased. Precipitation of carbide reduces the lattice distortion, so that the hardness and yield strength steadily decrease and the toughness increases. Thus, choice of annealing temperature enables any desired intermediate stage to be produced with yield strength between that of a martensite and that of a soft-annealed microstructure (see Section 6.1.3.2). Annealing also gives a very uniform carbide distribution, which greatly improves the toughness of the material for a given yield strength [288] (see Section 6.1.3.2).

6.1.1.5. Microstructures of Alloy Steels

The mechanisms of microstructure formation described above also apply to alloy steels where

the volume ratios of individual phases change, particularly if carbide and nitride phases are formed, which does not occur with plain carbon steels. In tool steels, a quenched and tempered microstructure is produced with included coarse carbides which have an abrasion-resistant effect (Fig. 201). The matrix strength can be controlled by choice of annealing temperature [289]. In alloy tool steels, the hardness of chromium and tungsten carbides is also utilized. Furthermore, precipitation of carbides as very fine particles on annealing these alloys sometimes starts only at ca. 500 °C, so that, unlike the plain carbon steels, these materials can be used at high temperature [290]. The growth of alloy-rich carbides, e.g., M_7C_3 or MC, is determined by diffusion of the alloying elements, which is much slower than that of the carbon. The carbide particles remain small for a long time after they are formed, so that the yield strength remains high, decreasing only as the particle size increases. This effect, known as precipitation hardening (see Section 6.1.3.3), leads to an increase in the high-temperature creep resistance of high-temperature steels, as the precipitates are very small [289], [291].

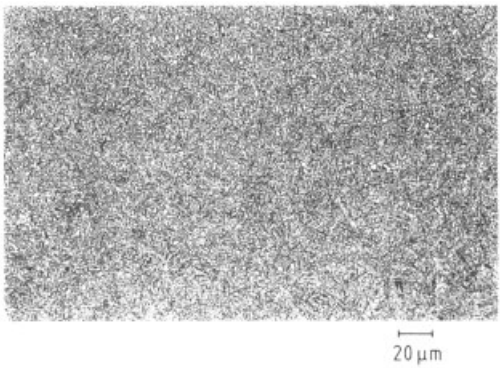

Figure 201. Quenched and tempered microstructure of C 130 tool steel
Coarse, undissolved carbides (white) in a matrix of very finely divided carbides in ferrite, 810 °C for 15 min, water quench, 150 °C for 10 min, aircooling; etched with nital

In maraging steels, precipitation of intermetallic phases leads to precipitation hardening, and increases the yield stength of a low carbon martensite to ca. 2000 MPa [292]. In nontransformable austenitic and ferritic steels, hardening by martensite formation is not possible. Thus, precipitation hardening is an important method of increasing strength, in addition to grain refinement and work hardening [293] (see Section 6.1.2).

Corrosion-resistant steels usually have a pure ferritic or austenitic microstructure [294–296]. However, duplex steels have a microstructure consisting of approximately 50 % austenite and 50 % ferrite [297]. In corrosion-resistant steels, the chemical composition of the phases and the type and arrangement of the carbides have an extremely important influence on their resistance to aggressive media [298].

6.1.2. Properties

6.1.2.1. Mechanical Properties

Steel must have high strength and good toughness properties. Following the plan of the comprehensive treatment of mechanical properties in *Werkstoffkunde Stahl* [300], [301], the first part of this section describes conditions under which high strengths can be achieved by modification of the microstructure. The discussion then moves to toughness requirements, definitions of this property for particular applications, and ways of influencing it through the microstructure. The discussion deals first with axial stress conditions at and below room temperature. If the direction of the stress is changed, special effects occur, which are described later. As steel is also used at higher temperatures, methods of controlling strength properties under these conditions are described in the last section.

6.1.2.1.1. Behavior Under Unidirectional Loads, at and Below Room Temperature [302–309]

The Stress – Strain Curve [310], [311]. At first, iron deforms elastically under an applied force. The relationship between the stress and the resulting strain (or elongation) is expressed by Hooke's law:

$$\sigma = F/S = E\varepsilon = E\,(l - l_o)\,/l_o$$

where F is the force, S the cross section, l_o the initial length, and l the length of the test piece subjected to a tensile stress. The modulus of elasticity E is 216 000 MPa, but this value depends

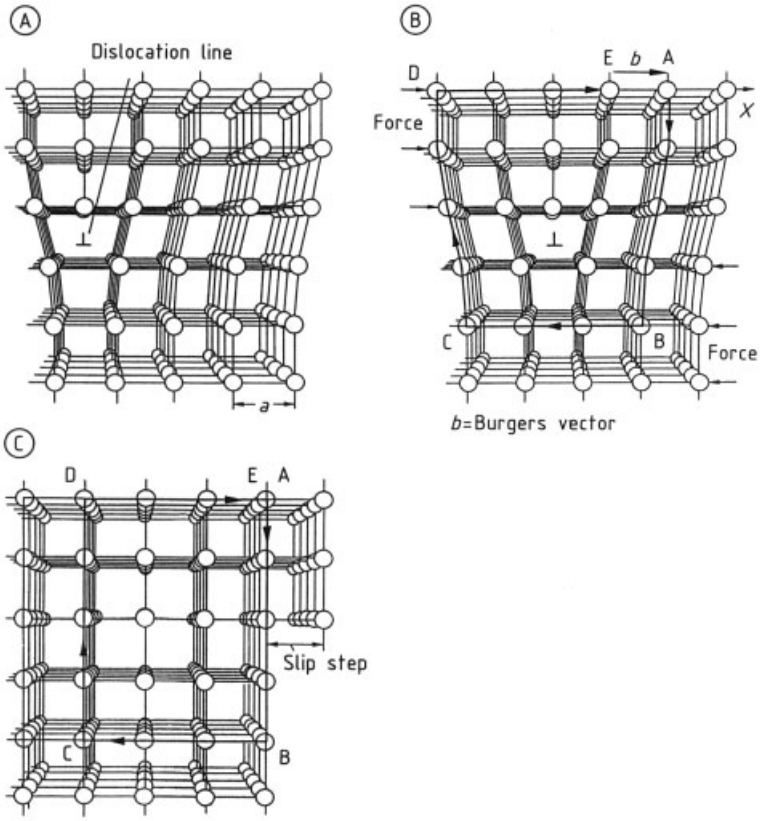

Figure 202. Movement of an edge dislocation (represented by the symbol ⊥) across a cubic lattice [314]
A) Initial state; B) Burgers circuit around a dislocation; C) Burgers circuit in an undisturbed lattice after slipping

on the orientation. A similar relationship exists between shear stress and shear strain; the proportionality factor G is the modulus of rigidity or shear modulus. The relationship between the Young's modulus E and the shear modulus is:

$$G = E / [2(1+\nu)]$$

in which Poisson's ratio ν for iron is 0.3.

In the elastic range the strain is reversible, but beyond a specific stress limit, permanent plastic deformation begins. Experiment shows that this phenomenon is associated with slip of crystal planes along the direction of greatest shear stress. The lattice structure is preserved. According to theory, this slip should require stresses a thousand times as great as the measured values in pure iron. The discrepancy is explained by the presence of linear lattice defects, i.e., dislocations [312–314]. In a linear edge dislocation (Fig. 202), slip takes place perpendicular to the dislocation line. In a screw dislocation (Fig. 203), slip is parallel to the dislocation line. With the aid of these and other types of dislocation, in which slip takes place at an angle 0 – 90° to the dislocation line, the plastic deformation of iron can be well understood. As a dislocation causes a slip of one interatomic distance, and can only move over a short distance in the lattice before the movement stops at an obstacle (Fig. 204 A), many dislocations are necessary for deformation. The number of dislocations must increase continuously during the deformation process. A mechanism for the production of dislocations is illustrated in Figure 204 C (the Frank – Read source). This is based on the fact that dislocations of opposite sign annihilate each other. In the Orowan mechanism (Fig. 204 B), the dislocation passes the obstacles, leaving dislocation loops around them. This illustrates a possible mechanism to hinder

the movement of a dislocation, which leads to an increase in strength as the stress moves the dislocation through the lattice.

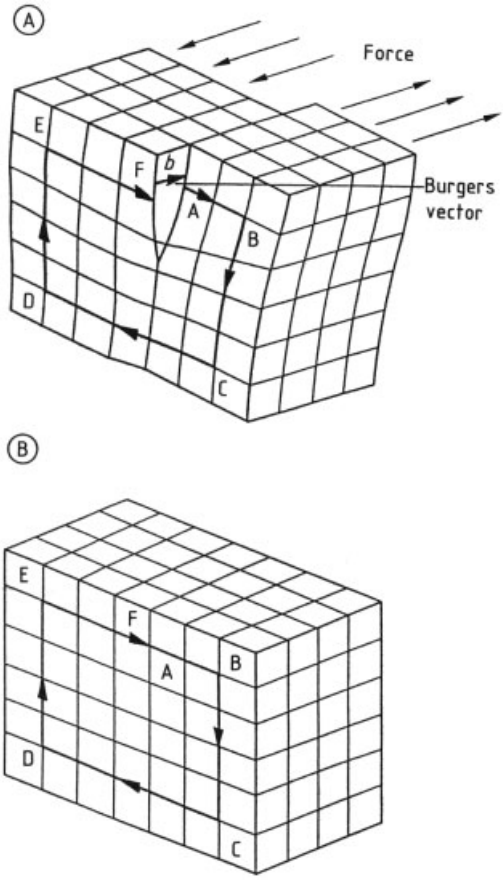

Figure 203. Burgers circuit
A) Burgers circuit around a screw dislocation; B) Burgers circuit in an undisturbed lattice

In crystals that have not been deformed, the dislocation density is $10^6 - 10^8\ \mathrm{cm}^{-2}$ and can increase to ca. $10^{12}\ \mathrm{cm}^{-2}$ on deformation. This figure can be regarded either as the number of intersections of dislocations with an area of 1 cm^2, or as the total length of the dislocation lines in a volume of 1 cm^3.

The most important method of measuring the relationship between stress and strain is the tensile test. A cyclindrical test piece is deformed by an axial tensile load at a constant rate of elongation or constant rate of increase of tensile stress (DIN 50 145, 50 146). Figure 205 shows the technical or conventional stress – strain curve (curve a). Dividing the force F by the initial cross section S_o, gives the stress $\sigma = F/S_o$, and the strain is calculated from the elongation. Initially, Hooke's law is obeyed. As plastic deformation begins, the stresses can be determined at which permanent deformations take place, e.g., the technical yield stress with a permanent deformation of 0.01 % (0.01 % offset), the 0.2 % offset yield stress $R_{p0.2}$, or the 1 % offset yield stress R_{p1}. Steel often shows a well-defined upper yield point R_{eH} and a lower yield point R_{eL}. These will be discussed later.

The transition from the elastic to the plastic range occurs when movement of dislocations takes place under the action of applied stress. The magnitude of the stress required is determined by the structure. The resistance to further deformation, i.e., the strain hardening, is the result of the increase in the number of dislocations, which prevents further movement of the dislocations. This hardening has the following effect: After the elastic limit has been exceeded, the regions which deform first, and hence have a higher stress owing to the reduction in cross section, do not flow further until the remainder of the test piece has reached the same cross section. However, as hardening decreases with increasing elongation, at a certain strain the increase in stress is no longer compensated for. Thus, further deformation takes place at a fortuitously deformed place, and necking occurs. The maximum in curve (a) of Figure 205 corresponds to this. The force F_m divided by the initial cross section S_o is referred to as the ultimate tensile strength R_m. Tensile test measurements also enable determination of the uniform elongation (i.e., without necking) A_{gl}; this is the elongation at maximum force. A is the elongation at rupture, and Z is defined as the reduction of area at rupture:

$$Z = (S_o - S_u)/S_o$$

where S_u is the cross section at rupture. Because of the necking, A depends on the geometry of the sample. The reduction of area is a significant parameter for characterizing toughness in tensile tests.

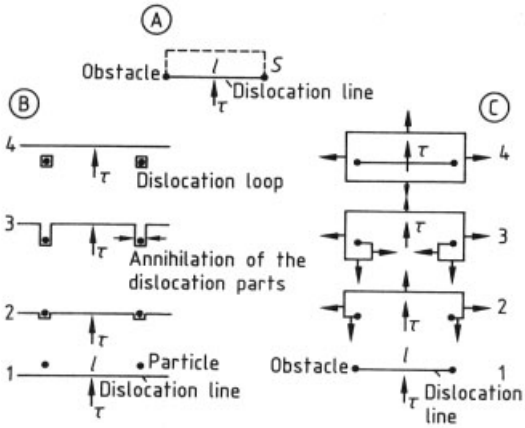

Figure 204. Behavior of dislocations at obstacles
A) Elongation of dislocation line l by the amount $2s$ under a stress; B) Bypassing of particles by a dislocation (Orowan mechanism); C) Production of dislocations (Frank – Read source)

To determine the predominant true stress in the test piece, the force must be divided by the smallest cross section S. The corresponding true elongation ε_w is obtained for elongations up to the maximum force at which the test piece deforms uniformly:

$$\varepsilon_w = \int_{l_0}^{l} (dl/l) = \ln(l/l_o) = \ln(l+\varepsilon)$$

Over the whole region:

$$\varepsilon_w = \ln(S_0/S)$$

This is due to the constant volume relationship

$$l_0 S_0 = l S$$

In Figure 205, the true stress, determined as described, is shown as curve (b); for exact calculations, it must be corrected to take account of the multiaxial stress condition when necking takes place (curve c) [315].

At maximum load, $dF = 0$, so the relationship between ultimate tensile strength and the start of necking can also be derived graphically. Figure 206 A shows stress – strain curves for steels of various strengths [316]. Figure 206 B shows the true stress – true strain curves, and the strain hardening as functions of the true strain. The condition $dF = 0$ means that, at the start of necking:

$$\sigma_w = d\sigma_w / d\varepsilon_w$$

For higher yield stresses, the start of necking is shifted to continually decreasing strain values.

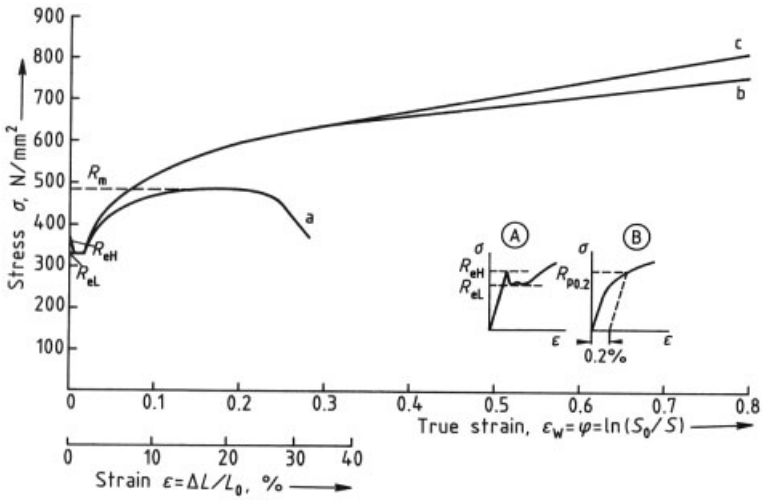

Figure 205. Stress – strain curve at room temperature of an unalloyed, fully killed steel with ca. 0.15 % C, 0.4 % Si, and 1.4 % Mn (St 52-3).
True stress as a function of true strain (a and b [315]); stress as a function of strain (c)
A) Upper and lower yield stress; B) Yield strength at 0.2 % offset

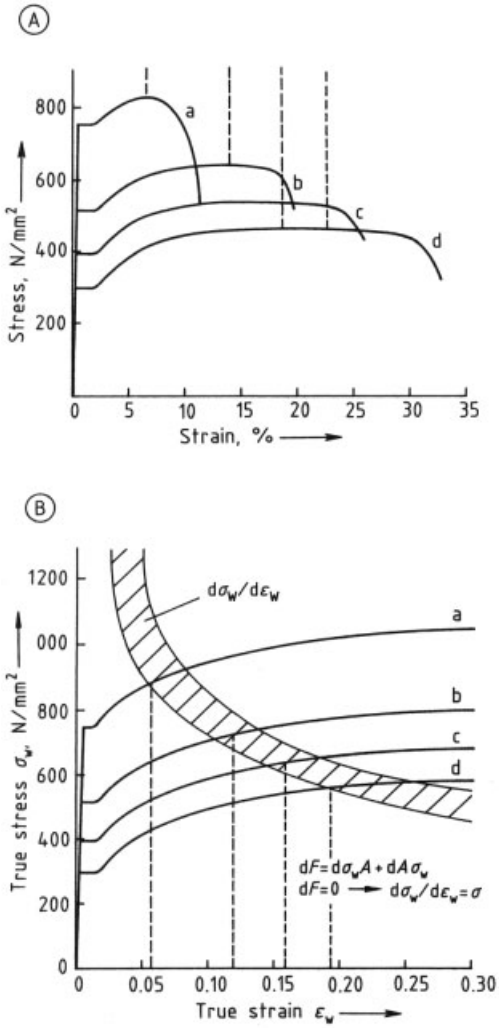

Figure 206. Stress – strain behavior of steels [316]
A) Conventional stress – strain curves; B) True stress – strain curves and strain hardening $d\sigma_w/d\varepsilon_w$
a) P690Q (StE690V); b) P460M (StE460TM); c) S355J2 (SES2-3); d) S235J2 (St37-3)

Steels often show a discontinuous transition from the elastic to the plastic region – the distinct limit of elasticity or yield point (see Fig. 205) [317]. The Hooke's law straight line rises to the upper yield point. A reduction in stress then takes place to the value for the lower yield point, where a narrow deformed region is produced in the test piece (the Lüders band) at 45° to the direction of the tensile load. This band broadens at the stress of the lower yield point, until the whole test piece is deformed by this amount. Further deformation takes place uniformly over the whole volume. The cause of this distinct yield point, according to COTTRELL [318], is blocking of the dislocations by carbon and nitrogen atoms, caused by elastic interaction of the stress fields at a dislocation and at the interstitial foreign atoms (Cottrell atmosphere). This blockage can occur if the diffusion rate of carbon or nitrogen atoms is high enough, e.g., in aging (precipitation hardening) on cooling from the rolling temperature or after heat treatment (Fig. 207), or during the deformation process in strain aging (Fig. 208) [319], [320].

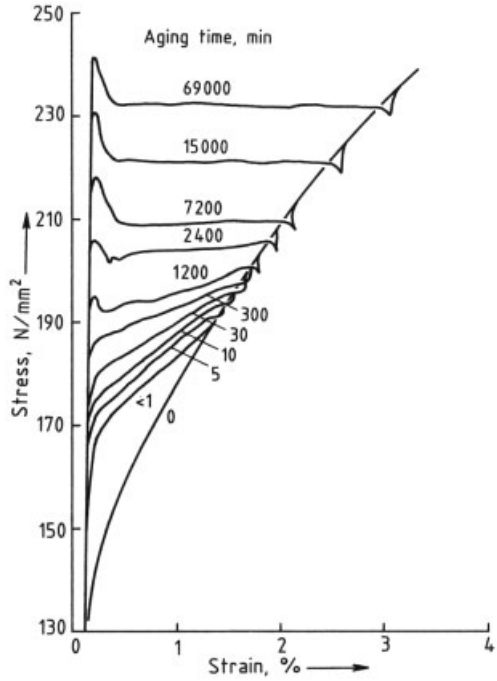

Figure 207. Stress – strain curves for an iron – carbon alloy with 0.01 % C after quenching from 720 °C and various aging times at 60 °C [319]

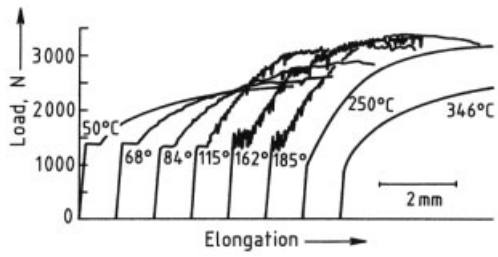

Figure 208. Load – elongation curve of a steel with 0.035 % C, 0.36 % Mn, and 0.003 % N at various temperatures
Strain rate $6.2 \times 10^{-5}\ s^{-1}$ [320]

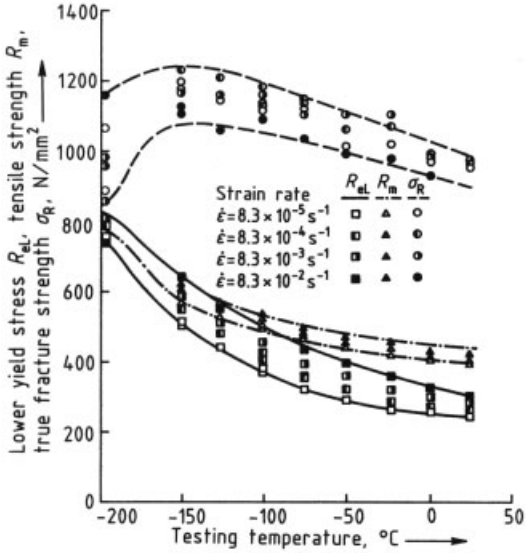

Figure 209. Effect of temperature on lower yield stress R_{eL}, ultimate tensile strength R_m, and rupture stress σ_R of a steel with 0.089 % C (C 10) and average grain diameter 16.8 µm at various strain rates $\dot{\varepsilon}$ [321]

The values determined from the stress – strain curve depend on both test temperature and strain rate. In Figures 209 and 210, the changes in stress and strain values due to changes in the loading rate below room temperature are shown as an example. Lower yield stress, ultimate tensile strength, and rupture stress increase with decreasing temperature and increasing loading rate. At first, the uniform elongation (without necking) and elongation at rupture remain approximately constant, and the Lüders elongation increases. At lower temperature, the elongation values decrease, but the strength values increase, owing to a change in the rupture mechanism, as discussed below.

If the strain rate is increased further, the result shown in Figure 211 is observed. Assuming thermally activated dislocation movement, all the points lie on a common curve if the results are plotted against activation energies. There is a narrow scatter for the yield points (Fig. 212). This does not apply to ultimate tensile strengths if these are plotted in a similar way, as, under rapid, adiabatic conditions, the energy of deformation can raise the temperature, whereas isothermal conditions predominate at lower strain rates [321].

Many attempts have been made to describe the true stress – true strain curves mathematically. LUDWIK has proposed the equation:

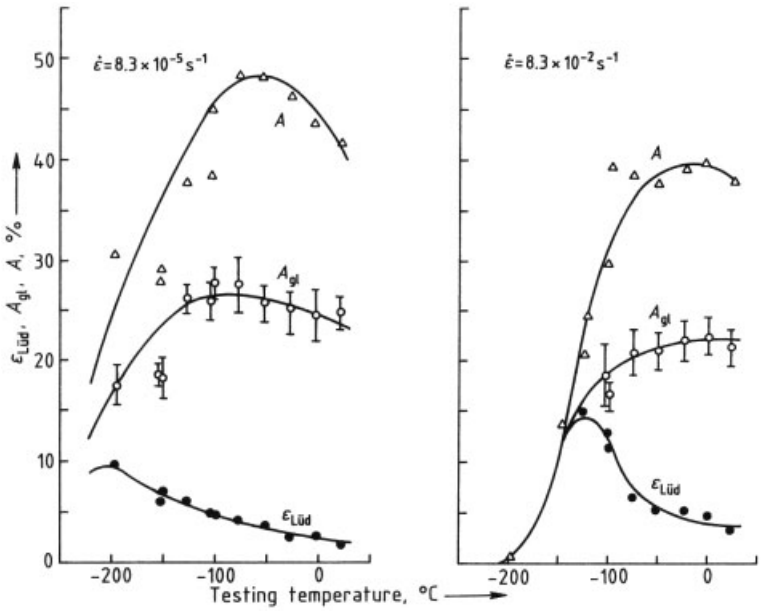

Figure 210. Effect of temperature on Lüders strain $\varepsilon_{Lüd}$, uniform elongation A_{gl}, and elongation at rupture A of the steel referred to in Figure 209 [321]

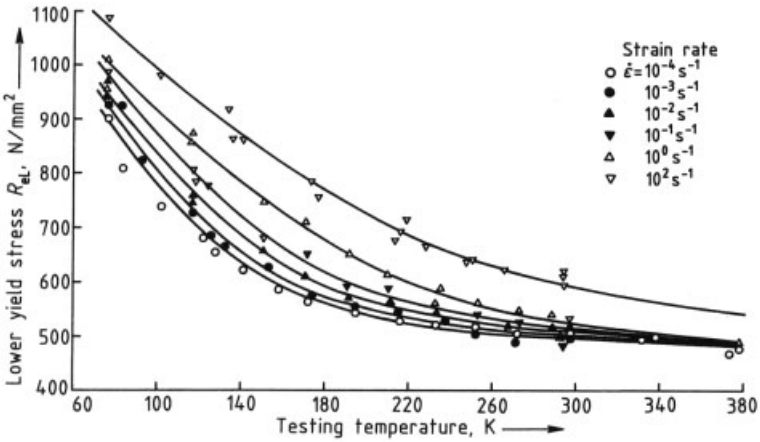

Figure 211. Effect of temperature on lower yield stress R_{eL} at various strain rates of a high-strength, fine-grained steel with 0.16 % C, 0.63 % Ni, and 0.16 % V (FeE 460) in the normalized condition [315]

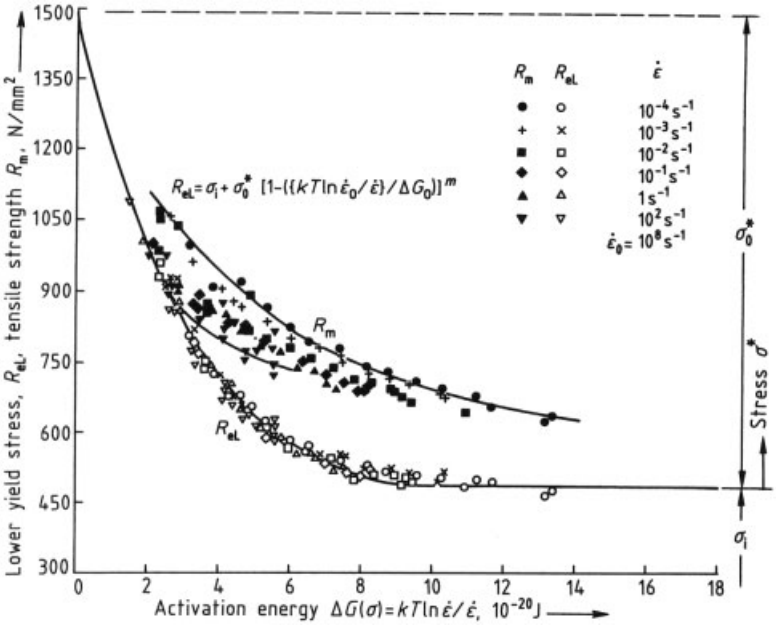

Figure 212. Lower yield stress R_{eL} and ultimate tensile strength R_m of the steel referred to in Figure 211 plotted against the activation energy ΔG for yielding at various strain rates $\dot{\varepsilon}$ [315]
σ_i = athermal fraction of yield point; $\sigma*$ = thermal fraction of yield point; $\sigma_o^* = \sigma*_{max}$

$$\sigma_w = \sigma_o + K_L \varepsilon_w^n$$

This has been simplified by HOLLOMON to:

$$\sigma_w = K'_L \varepsilon_w^{n'}$$

where σ_w is the true stress, σ_o the yield stress, K_L a constant, ε_w the true strain, and n the exponent. This formula adequately describes the properties of many steels, but n' decreases with increasing yield stress, while the strength properties do not change. The equation:

$$n' = \varepsilon_{wgl}$$

is correct, as shown by differentiation of the equation, in which ε_{wgl} represents the uniform

strain, but n' does not represent the hardening behavior.

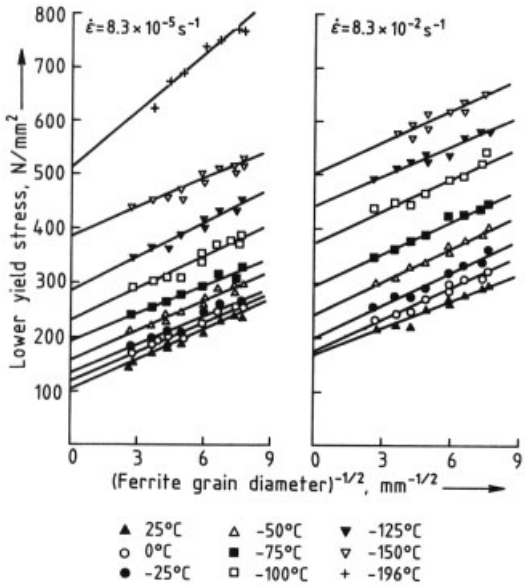

Figure 213. Lower yield stress R_{eL} plotted against the reciprocal of the square root of the average ferrite grain diameter d (mm) of a steel with 0.089 % C (C 10) at various testing temperatures and strain rates [339]

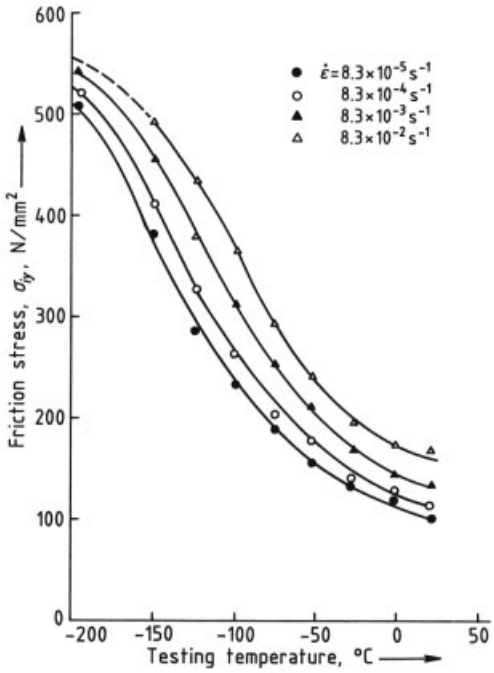

Figure 214. Effect of testing temperature and strain rate on the friction stress σ_{iy} for the steel referred to in Figure 213 [339]

Phenomenological methods can be applied to dislocation theory to obtain various mathematical formulae for the stress – strain curves, enabling approximate correlation with the microstructure [322–324].

The flow properties of steel can be investigated by other experimental methods. By upsetting a cylinder, the yield stress is determined from the ratio of the force to the instantaneous surface area as upsetting proceeds [325], [326]. Friction at the end faces leads to barrel deformation of the test piece, which can be avoided by appropriate experimental techniques. In the torsion test [327], the moment is determined as a function of the angle of rotation. The shear strain at any point on the diameter is calculated as a function of the shear stress at the outer edge. This technique is mainly used to determine properties at high temperature. Bending tests [328], [329] are used to determine strength properties of brittle materials. Strength properties can also be estimated from hardness measurements [330].

Methods of Increasing Strength [331–335]. All processes for increasing the strength of steel depend on hindering the movement of dislocations. These can be classified in accordance with their dimensions:

1) Zero-dimensional obstacles (e.g., point defects and foreign atoms)
2) One-dimensional linear obstacles (dislocations)
3) Two-dimensional surfaces (grain boundaries)
4) Three-dimensional obstacles (precipitates)

The discussion starts with the effect of grain size, as this shows very clear correlations.

Strength increases as grain size decreases [336–339]. The yield point of unalloyed steel is inversely proportional to the square root of the ferrite grain size. This is represented by the Hall – Petch equation:

$$R_{\mathrm{eL}} = \sigma_{iy} + k_y d^{-1/2}$$

where σ_{iy} represents the frictional stress, k_y the grain boundary resistance, and d the average diameter of the ferrite grains.

Figure 213 shows the lower yield point as a function of the reciprocal of the square root of the grain diameter for temperatures between room temperature and − 196 °C for two strain

rates. The linear Hall – Petch relationship is observed. Figure 214 shows friction stress as a function of testing temperature. The values of friction stress are obtained by extrapolation of the straight lines in Figure 213 onto the y-axis. The friction stress σ_{iy} represents the resistance of very large grains to dislocation motion.

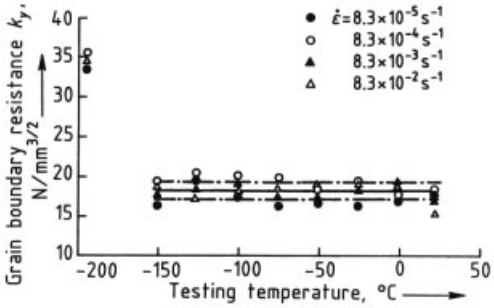

Figure 215. Effect on testing temperature and strain rate on the grain boundary resistance k_y for the steel referred to in Figure 213 [339]

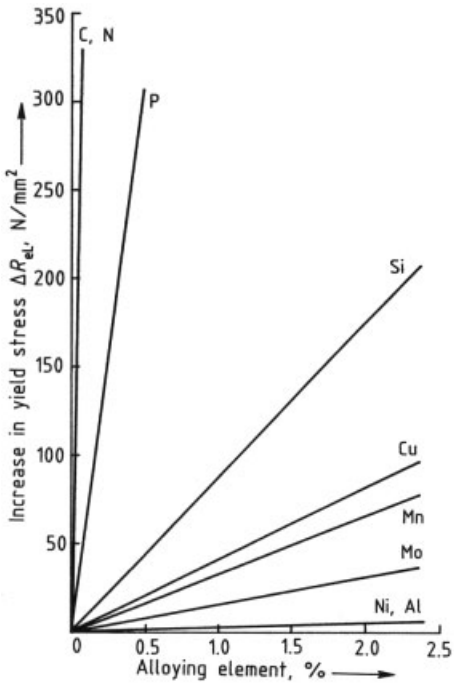

Figure 216. Effect on alloy content on the lower yield stress of ferrite at room temperature [332]

Strain rate and temperature have a considerable effect on the frictional stress. The grain boundary resistance is approximately independent of strain rate and temperature (Fig. 215). The higher value at – 196 °C is due to the change in the deformation mechanism from dislocation motion to twin formation.

The Hall – Petch equation is also valid for the yield stresses in the extended stress – strain curve. The grain boundary resistance is then smaller than at the lower yield point.

Grain refining is an important method of increasing strength, as toughness is also improved.

The mechanism of strength improvement by foreign atoms [340] is mainly interaction between the stress fields at the dislocations and the foreign atoms. Interstitial atoms have a stronger effect than foreign atoms in substitution solid solutions, as they are more firmly fixed (Fig. 216). The effect of the foreign atoms is limited by their solubility. The extreme hardness of martensite is caused on quenching by "freezing" the dissolved carbon atoms in the austenite, "folding" them into the body-centered cubic lattice; this produces an extreme degree of supersaturation with a corresponding increase in hardness.

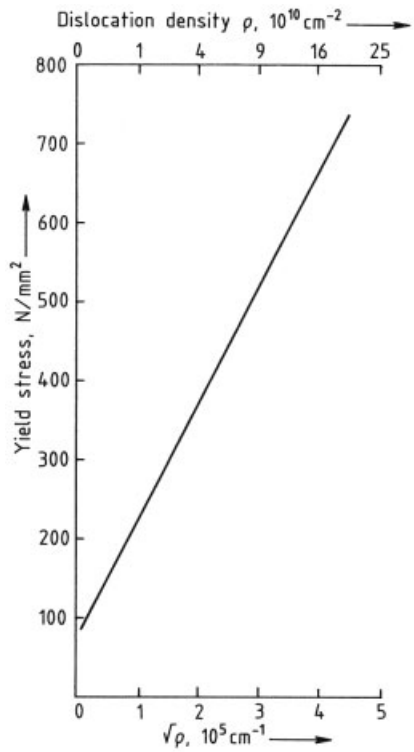

Figure 217. Yield stress of α-iron as a function of dislocation density ϱ [341]

Dislocations are also obstacles [341]. Increase in dislocation density on deformation leads to an increase in strength and hence an increase in yield stress (Fig. 217). Ductility is progressively reduced by prestraining.

Precipitates hinder the movement of dislocations very effectively [342]. Finely dispersed precipitates, which have only a slight misfit to the matrix, are sliced by dislocations. Larger precipitates are enveloped by dislocation loops (Orowan mechanism; Fig. 218). Precipitate formation with time, and the change in the yield stress are shown in Figure 219.

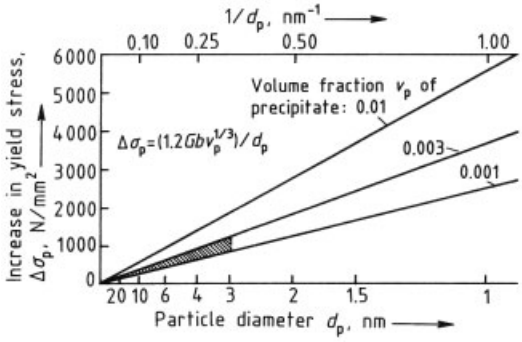

Figure 218. Calculated increase in the yield stress $\Delta\sigma_P$ of α-iron by hard particles in various volume fractions v_P Hatched area: region of microalloyed structural steels [343]

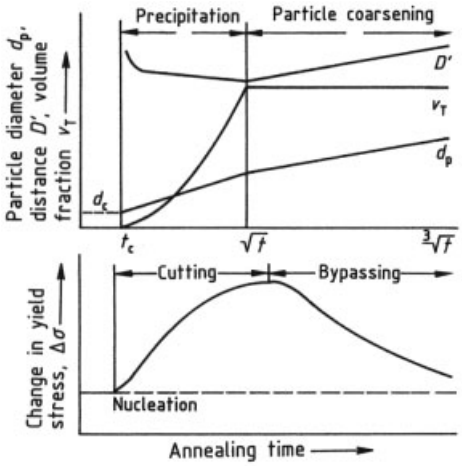

Figure 219. Formation and growth of nuclei on annealing a solid solution, and changes in yield stress with time [342]. Nuclei of the critical size d_c develop at time t_c

Different mechanisms of increasing strength may be combined. Figure 220 shows how grain refining and precipitation hardening overlap. The left lower region of the diagram shows how repeatedly austenitizing a steel (ca. 0.1 % C and 1 % Mn) reduces grain size and hence increases yield stress. The x-axis is the reciprocal of the square root of the grain diameter, so Hall – Petch straight lines are obtained. If 0.2 % Nb is added and the steel is normalized, grain size is considerably reduced and yield stress increases. Niobium carbonitrides act as obstacles to grain growth during normalization. If the steel is heated to 1100 – 1200 °C and cooled in air, considerably higher yield stresses are obtained, as these carbonitrides dissolve and are reprecipitated on air cooling. However, grain size increases, which adversely affects the toughness properties. If the material is deformed with decreasing final rolling temperature after austenite formation, finer grains with high yield stresses are obtained by dynamic recrystallization (finely dispersed carbides are precipitated before, during, and after deformation). The yield stress increases with the cooling rate. Even with small additions of microalloying elements, this thermomechanical treatment leads to high strengths by a combination of grain refining and precipitation hardening.

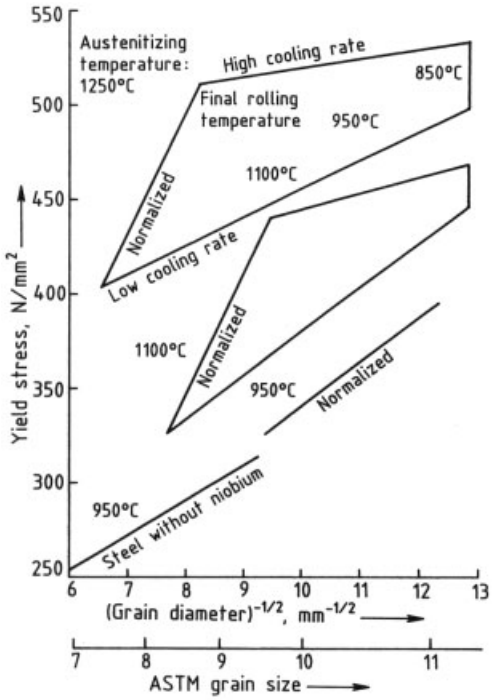

Figure 220. Yield stress plotted against grain size at various austenitizing temperatures for a steel with ca. 0.1 % C, 1 % Mn, and 0.02 % Nb [344]

Pearlite is another microstructure that hinders the movement of dislocations and limits the extent of this movement. The relationship between the 0.2 % offset yield stress and the interlamellar distance follows a Hall – Petch equation [345]. In martensite, the most important effect is solid solution hardening due to the high content of dissolved carbon. The high dislocation density caused by the flip-over process is also effective [346]. The boundaries of twins or blocks in the martensite take over the role of the grain boundaries. In bainite both flip-over ferrite and carbides are present.

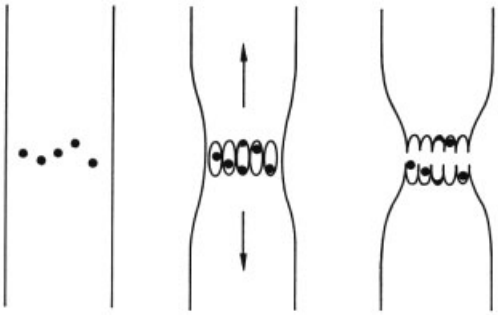

Figure 221. Development of cavities at precipitated particles or nonmetallic inclusions, leading to dimples on the fracture surface [351]

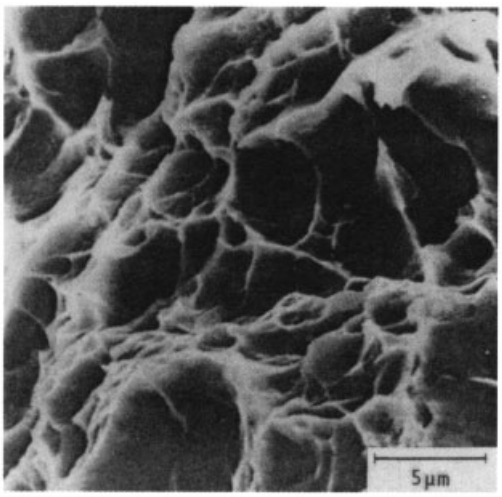

Figure 222. Scanning electron micrograph of a dimpled fracture

Similar considerations also apply to austenitic steels, but corrosion resistance must also be considered. A further possibility is the hardening of unstable austenitic steels by partial conversion to ε- or α'-martensite by prestraining, particularly at low temperature, and/or during loading [347].

The analysis of strength-increasing mechanisms is important because these mechanisms can affect toughness properties in various ways (Section 6.1.3.2). The attempt should always be made to choose a suitable combination of mechanisms.

Toughness and Fracture Behavior [304], [348–350]. *Fracture Characterization* [351–354]. When steel is in service, overloads that could eventually lead to fracture cannot be completely excluded. Not only strength, but also toughness is of great importance. For body-centered cubic steels, there are two different types of fracture: ductile fracture at high temperature, and brittle fracture at low temperature.

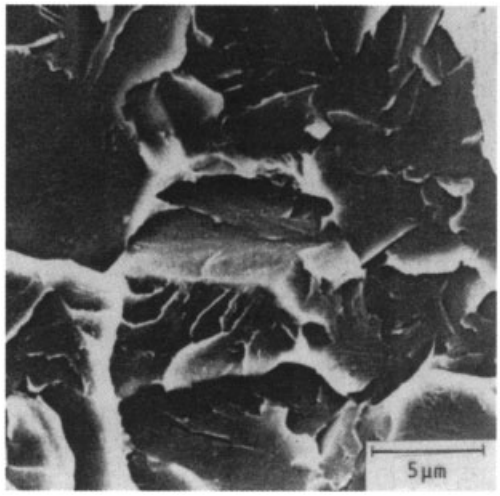

Figure 223. Scanning electron micrograph of a transcrystalline cleavage fracture

Ductile fracture takes place after plastic deformation caused by tensile stress, which can reach very high values in pure metals. In industrial materials, cavities form in the course of deformation before fracture. These start at the phase boundaries between the matrix and the inclusions or precipitates. They grow and coalesce as the load is applied (Fig. 221) [355], [356]. Typical cavities or dimples can be seen on the fracture surface, and the inclusions which led to the initiation of pore formation can often be an-

alyzed. A typical scanning electron micrograph of a dimple fracture is shown in Figure 222.

Brittle fracture (Fig. 223) takes place by transcrystalline separation of crystal faces when a critical tensile stress, the microscopic cleavage fracture stress, is reached. Macroscopic deformation cannot usually be detected. The test piece or component fails suddenly. If the grain boundaries are weakened by precipitates or concentrations of foreign atoms, intercrystalline cleavage fracture can occur by separation at the grain boundaries.

External Influences on Fracture Behavior. The most important external influences on fracture behavior are temperature, loading rate, and state of stress. The effects of temperature and loading rate are shown in Figure 224, in which the temperature – lower yield stress curve is shown as a continuous line. When the temperature is lowered, the level of microscopic cleavage fracture stress is eventually reached. This stress value is only slightly affected by temperature. Below the point of intersection, the material fails by brittle fracture. At higher temperatures, ductile fracture takes place; i.e., after the yield point has been exceeded, deformation, cavity formation, cavity growth, and fracture occur.

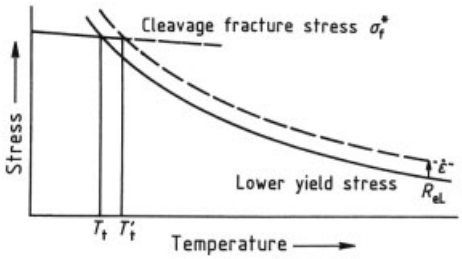

Figure 224. Effect of temperature on lower yield stress R_{eL} and cleavage fracture stress σ_f^* [357]

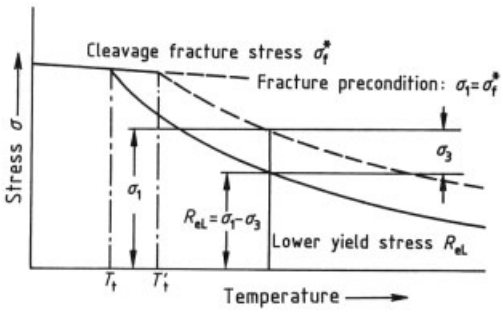

Figure 225. Effect of stress state on transition temperature [357]

At higher strain rates, the yield stress increases, but the cleavage fracture stress is almost unchanged. The transition temperature is shifted to higher values (Fig. 224), and the safe operating temperature is higher.

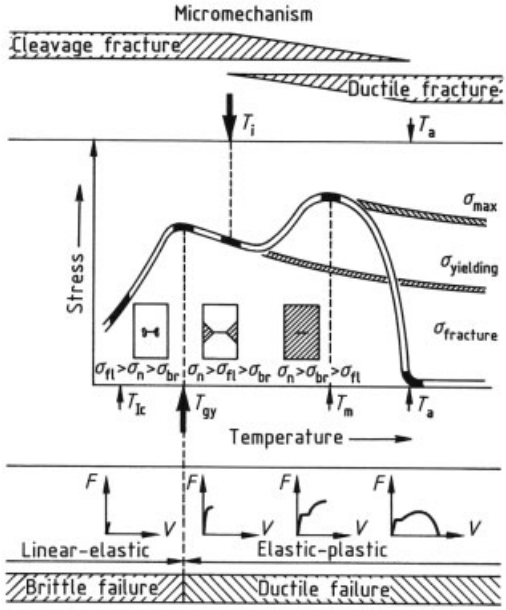

Figure 226. Effect of temperature on the behavior of test pieces and components that contain cracks [358]

The state of stress is important for fracture behavior, and this too can be shown by changes in cleavage fracture stress and yield stress. With multiaxial stress, ductile deformation begins as soon as the difference between the highest and lowest principal stress reaches the value of the yield stress (Tresca's flow criterion):

$$R_{eL} = \sigma_1 - \sigma_3$$

Figure 225 shows the effect of temperature on the highest principal stress σ_1; the following equation must always be valid at the onset of flow:

$$\sigma_1 = R_{eL} + \sigma_3$$

The fracture condition in the brittle fracture area is reached when the highest principal tensile stress is equal to the microscopic cleavage fracture stress:

$$\sigma_1 = \sigma_f^*$$

Because the stress is multiaxial, the brittle – ductile transition temperature is shifted to higher values. Multiaxial stresses in components tend to occur at notches and cracks, so fractures tend to start at these locations. Ahead of the notch or crack, a plastic zone always appears, as a principal stress on the free surface disappears. The plastic zone is larger for larger notch radii and lower yield stresses.

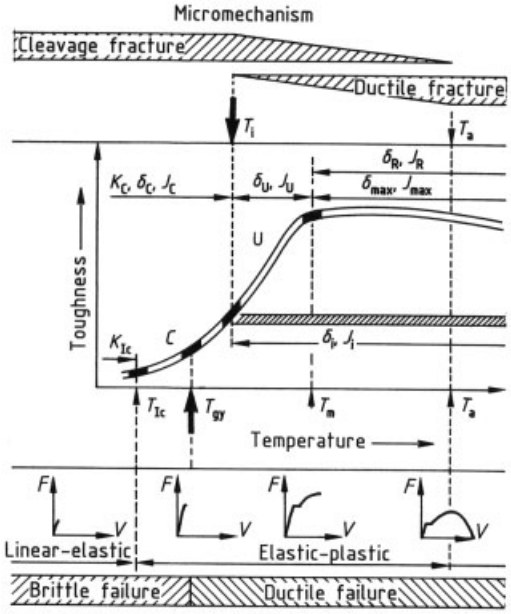

Figure 227. Effect of temperature on toughness [358]

It follows that a test piece or component that contains cracks has a characteristic behavior which depends on the testing temperature or operating temperature. At low temperatures, fracture takes place within the elastic part of the stress – strain curve, at low stress values and with small plastic zones as shown in Figure 226, which refers to an example of a center-notched flat tensile test piece [358]. With increasing temperature, the size of the plastic zone at fracture increases, until the whole section flows. As the temperature increases further, the fracture stress decreases as the yield stress decreases, but the mechanism of cleavage fracture remains. Ductile fracture occurs above the transition temperature T_i. Initially crack growth is stable, but soon reverts to unstable cleavage fracture. The effect of temperature on the onset of plastic deformation, ultimate strength, and fracture stress are also shown in Figure 226. Above T_a, stable crack growth takes place until the test piece is completely broken in two.

The corresponding figures for toughness are shown in Figure 227. The terminology is taken from the nomenclature of fracture mechanics, but the characteristic shape of the curve, with a lower plateau, a transition region, and an upper plateau is observed in all testing methods. The detailed form of the curve varies according to the shape of the test piece. Index i applies to the values at the beginning of stable crack growth, which cannot be determined with all test methods.

Test Methods [359–361]. Toughness measurements can be used to compare different steels, enabling product consistency to be monitored (quality control) or the effect of different microstructures to be observed (development of steel properties). In another group of test meth-

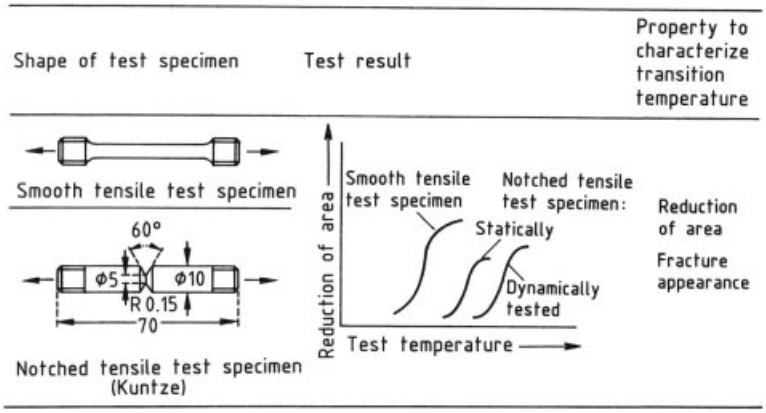

Figure 228. Test pieces and results of tensile tests to characterize toughness and fracture behavior [361]

ods, the results enable prediction of component behavior.

The result of a test can be expressed as a transition temperature which separates the brittle region from the ductile region. Alternatively, statements can be made about the degree of toughness, especially in the region of ductile fracture.

Testing methods can be divided into two groups. Small test pieces are prepared from the material being investigated, and tested at various temperatures or at the intended operational temperature. Large test pieces similar to the component can simulate operational conditions, and hence give more reliable predictions of the suitability of the material. Their use is very expensive, and is only justified for special applications; they are not described in detail here [303], [304].

Test piece shapes and the results of tensile measurements are illustrated in Figure 228. The reduction in area changes as the temperature moves from the lower plateau, characterized by brittle fracture, through the transition region to the upper plateau, with ductile fracture and stable crack growth. The lowest transition temperatures are obtained with smooth tensile test specimens at quasistatic loading rates. The transition to higher temperatures takes place on dynamically loaded notched test pieces.

Notch bending tests are of great practical importance (Fig. 229). The testing of a shape similar to that of the component (VAN DER VEEN)

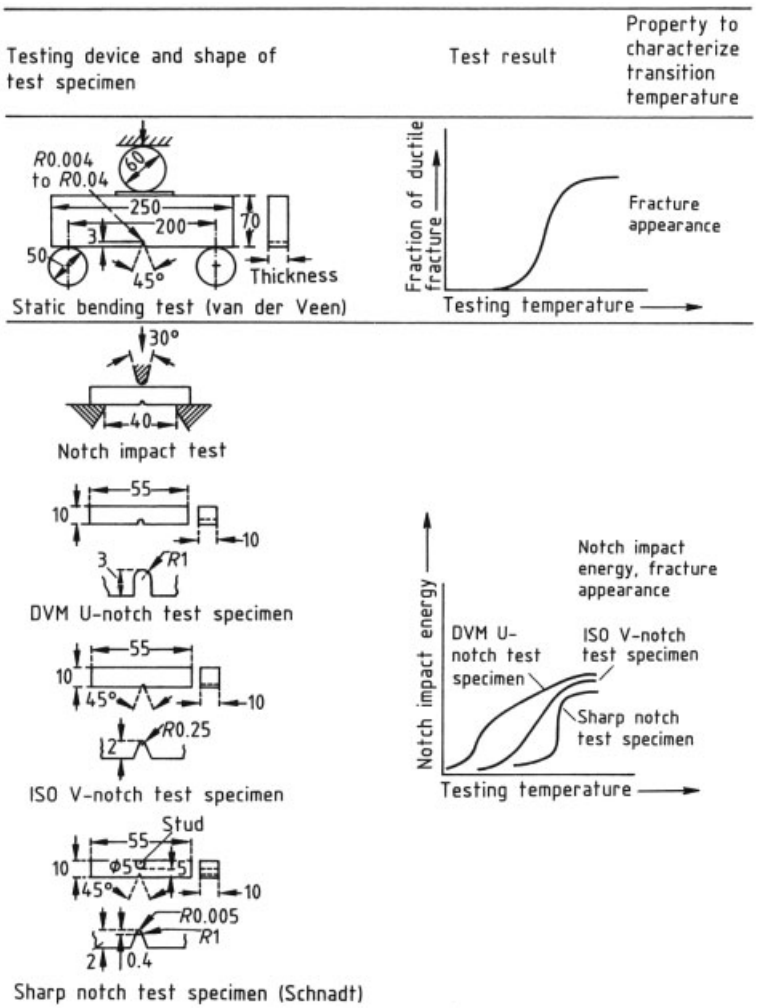

Figure 229. Testing devices, test pieces and test results in notched bending tests [361] (dimensions in mm)

is shown, as well as various proposed notched impact bending tests. The ISO V-notched bar is used inernationally. The transition from the lower to the upper plateau can be seen by plotting the impact energy, or the amount of brittle fracture surface as a function of temperature. Increasing notch acuity increases the transition temperature.

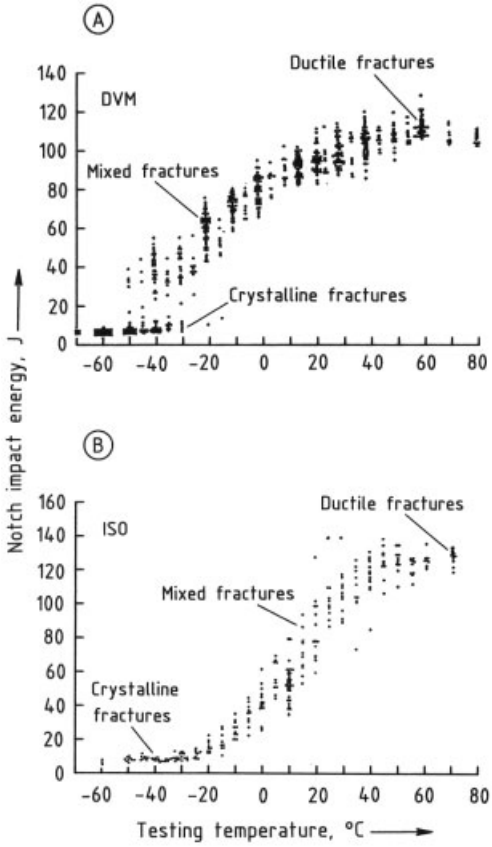

Figure 230. Notch impact energy – temperature curves for various notch shapes for an unalloyed steel with ca. 0.22 % C and yield stress $\geq$ 255 MPa (St 42-2) (dimensions in mm) A) DVM U-notch test pieces; B) ISO V-notch test pieces

To give an idea of the scatter of the results, and hence of the accuracy of all the figures, Figure 230 A shows the effect of temperature on the impact energy for DVM (Deutscher Verband für Materialprüfung) U-notch test pieces and Figure 230 B shows results for ISO V-notch test pieces. In the brittle low-temperature region, the scatter of results is small. The greatest degree of scatter, in the impact energy at constant temperature or in the temperature shift at constant energy, occurs in the transition region. In the upper plateau, the scatter decreases. Most investigations on the effect of microstructure on toughness have been carried out by measuring notch impact toughness.

The effect of the state of stress on the transition temperature has been systematically investigated. Figure 231 shows that the transition temperature for very different types of steel increases linearly with the degree of multiaxiality, defined as the ratio of the equivalent stress (VON MISES) to the largest principal normal stress. It is possible to calculate the test results for test pieces with other notch acuities.

Sharp cracks produced by fatigue in the test pieces are used, according to the concept of fracture mechanics [363–367]. It is possible to state the stress in components that contain cracks, and in corresponding laboratory test pieces, with the same parameters. The characterization, carried out by means of the stress intensity, is of crack-tip opening displacement, or by the J-integral (Fig. 232). The stress intensity is thus a measure of the state of stress ahead of the crack. The crack tip opening displacement immediately gives the local elongation at the crack tip. The J-integral is a measure of the deformation energy required for plastic deformation of the crack tip. The original definition is derived from a path-independent line integral around the crack tip. In standardized approximation methods, the J-integral can be determined by laboratory investigations from the load – elongation diagram. Whereas the application of the stress intensity K is limited to linearly elastic properties, the concept of crack-tip displacement and of the J-integral can be used, with certain limitations, over the whole range from brittle to ductile behavior in the upper plateau. In the ductile fracture region, the onset of stable crack growth is characteristic (Fig. 233). The behavior on further loading can then be expressed numerically by a crack resistance curve, either as crack-tip opening displacement or as a J-integral, depending the stability of crack growth.

The effect of loading rate on toughness can be described quantiatively with the aid of test methods based on fracture mechanics [369]. Figure 234 shows the fracture toughness of FeE 460 steel as a function of temperature for various loading rates. In the lower plateau, the toughness values are almost independent of the load-

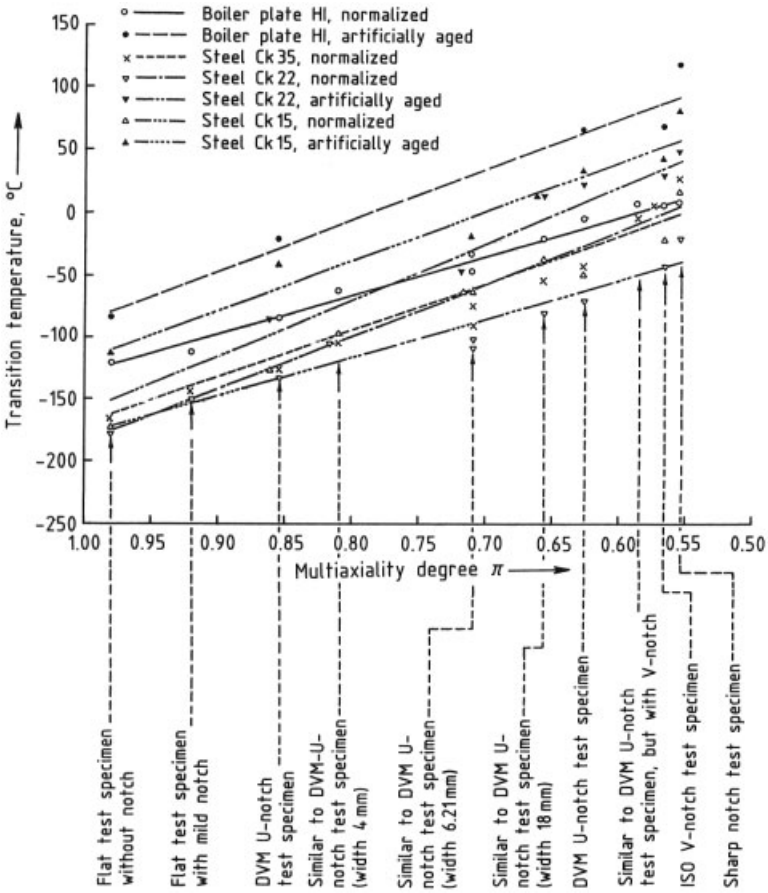

Figure 231. Effect of the degree of multiaxiality π on the transition temperature at $A_v \approx 40\,\mathrm{J/cm^2}$ with various shapes of notch impact test pieces of unalloyed steels with $\leq 0.16\,\%$ C (boiler plate HI), ca. 0.15 % C (Ck 15), ca. 0.22 % C (Ck 22), and ca. 0.35 % C (Ck 35)
Strain aging: 10 % cold working followed by annealing for 30 min at 250 °C [362]

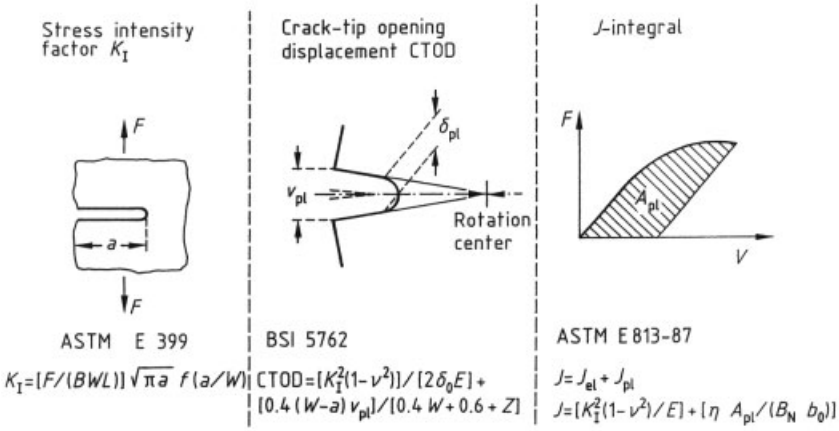

Figure 232. Fracture mechanics parameters

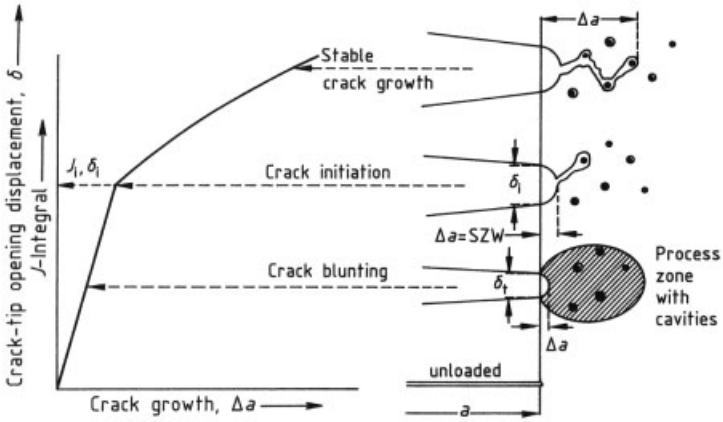

Figure 233. Crack-tip opening displacement or J-integral as a function of crack growth [355]
SZW = Stretched zone width

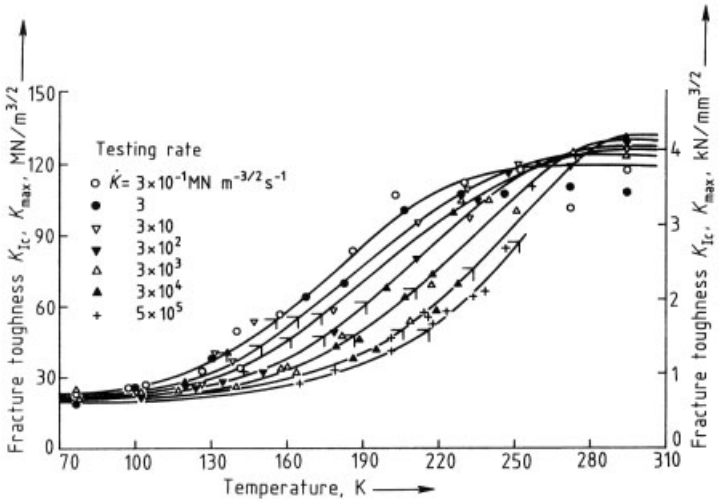

Figure 234. Effect of testing temperature on fracture toughness at various loading (testing) rates for the steel FeE 460 referred to in Figure 211 [CT (compact tension) test specimens, 13 mm thick] [368]

ing rate, while in the upper plateau, the toughness increases with loading rate. In the transition region, increasing loading rate causes a considerable shift of the transition temperatures to higher values. This result is due to the sensitivity of the elastic limit and yield stress to the loading rate. With increasing loading rate, the yield stress increases, and the plastic zone ahead of the crack becomes smaller. The transition temperature varies linearly with strain rate sensitivity (Fig. 235) [370]. The quantitative relationship is given by the formula in Figure 212. The increase in the values at the upper plateau with increasing loading rate has been established by determining crack resistance curves at high loading rates [371], [372].

Because small test pieces are not suitable for determining the properties of actual components, and test pieces that simulate components are very expensive, safety concepts have been developed that use mathematical models based on fracture mechanics [367], [373]. The most accurate, but also the most elaborate of these is stress analysis, in which the method of finite elements is used to calculate the stress in the component, expressed as the crack-tip opening displacement or the J-integral, and compared with critical characteristic material parameters

determined in the laboratory. Increasing simplification leads to the EPRI method, the "two-criteria approach" or R6 method, and finally to the CTOD design curve of DAWES and BURDEKIN. The results are on the conservative side, and the discrepancies between these and the actual failure point increase with increasing simplification. For the use of these methods the fracture mechanical parameters of the material must be determined in every case.

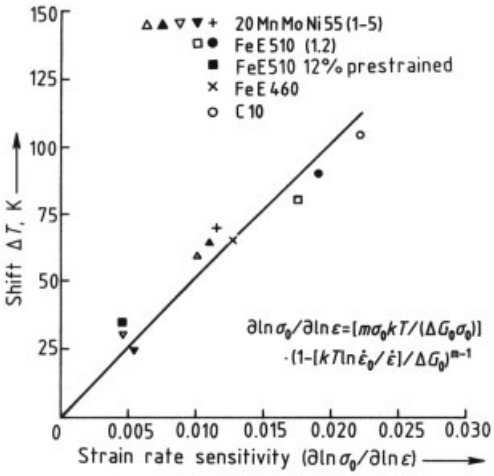

Figure 235. Shift of the fracture toughness – temperature curve by dynamic loading as a function of the strain rate sensitivity of the flow stress of various steels [370]

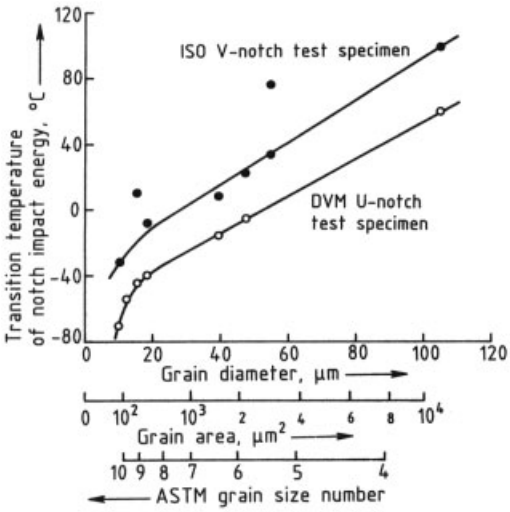

Figure 236. Decrease of transition temperature of notched impact energy (T_t at 50 % of upper plateau) as a function of grain size for an unalloyed, fully killed steel with 0.19 % C, 0.27 % Si, and 1.46 % Mn (St 52-3) [375]

Effect of Microstructure [374]. Microstructure has various effects on toughness. These have an important bearing on the behavior of steels in use, and their suitability for certain applications.

Grain refining reduces the transition temperature (Fig. 236). Similar results have been found with other test methods. The effect of a finer grain is favorable, because it increases strength (see Fig. 213) and improves toughness, as measured by transition temperature. The explanation is that the microscopic cleavage fracture stress σ_{f*} is increased to approximately four times the yield point R_{eL}, so that the transition temperature for finer-grained material not only fails to increase, but is shifted to lower values (Fig. 224) [376], [377]. This useful effect of grain refining is utilized in all modern steelmaking processes, e.g., in fine-grained structural steels and in thermomechanical treatment.

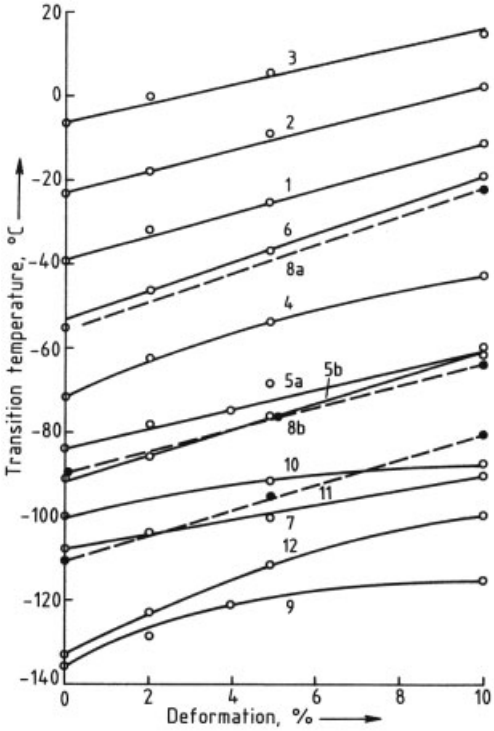

Figure 237. Transition temperature of notch impact energy (T_t at 35 J/cm^2 with DVM test specimens) plotted against degree of deformation for various steels
Curve 1, fully killed constructional steel; 2, 3, unalloyed boiler plate; 4, 5, unalloyed fine-grained steels, normalized; 6, 7, alloyed fine-grained steels, normalized; 8 – 10, alloyed fine-grained steels, quenched and tempered; 11, 12, alloyed steels, tough at low temperature [378]

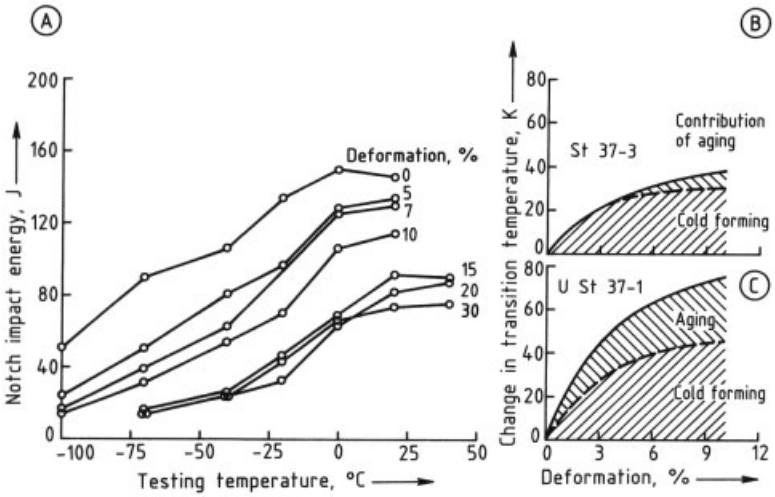

Figure 238. Effect of cold working due to increase in dislocation density on notched impact energy [379], [380]
A) Notch impact energy – temperature curve of an unalloyed, fully killed steel with ca. 0.15 % C, 0.4 % Si, and 1.4 % Mn (St 52-3); B), C) Effect of deformation degree on the transition temperature of hot-rolled wide strip of two unalloyed steels (Solid line = cold formed + aged, 250 °C, 30 min, air; Broken Line = cold formed); B) 0.14 % C, 0.1 % Al, and 0.006 % N (St 37-3); C) 0.08 % C and 0.009 % N (St 37-1)

Other processes for improving strength usually have unfavorable effects on toughness. The transition temperature increases with increasing cold working for many steels investigated. A deformation of 10 % increases the transition temperature by 20 – 30 K (Fig. 237). The effect on fracture mechanics tests can be even greater.

If cold working is combined with an aging treatment, (e.g., annealing at 250 °C with air cooling) a further shift of the transition temperature takes place by an amount that depends on the grade of steel. Figure 238 A shows the effect of temperature on the notch impact energy for various degrees of deformation. The change in the transition temperature caused by cold working alone, and the effect of an additional aging on a killed steel (Fig. 238 B) and a rimming steel (Fig. 238 C) are also shown. The considerably greater effect of strain aging on rimming steel can be seen. This is due to the higher content of free nitrogen. Figure 238 shows that, in the discussion of the sensitivity to aging of steel, the distinction must always be made between the effect of cold working alone (which always leads to an increase in the transition temperature) and that of additional aging (the magnitude of which depends on the level of dissolved nitrogen and/or carbon).

If the increase in yield stress is achieved by precipitation, toughness is usually decreased. This is also true for pearlitic structures in which increased pearlite content causes a decrease in the upper plateau and an increase in transition temperature.

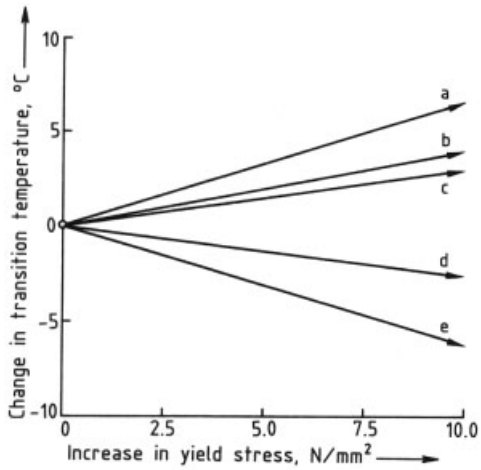

Figure 239. Effect of an increase in yield stress caused by various influences on the transition temperature of the notch impact energy [381]
a) Solid solution hardening by carbon; b) Strengthening by dislocations; c) Precipitation hardening; d) Solid solution hardening by manganese and nickel; e) Grain refining

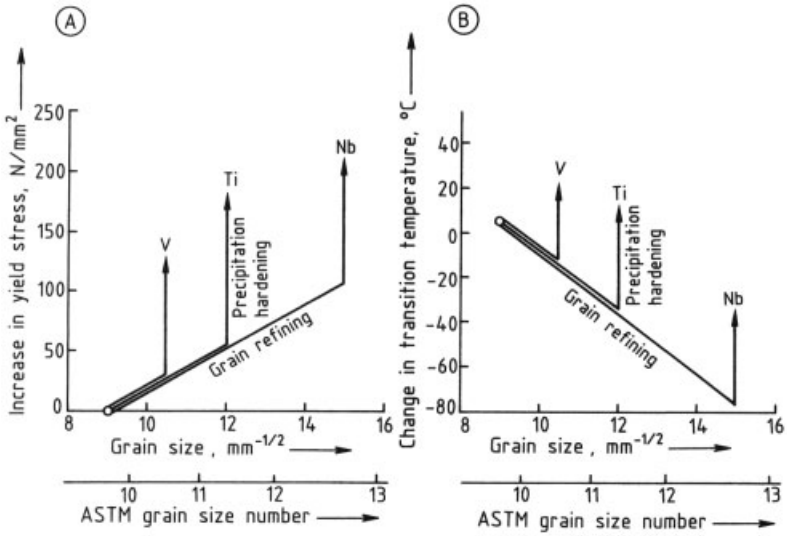

Figure 240. Contribution of grain refining and precipitation hardening to:
A) Change in yield stress; B) Change in transition temperature of notch impact energy (toughness)Unalloyed, fully killed steel, ca. 0.15 % C, 0.4 % Si, and 1.4 % Mn (St 52-3) with 0.05 % Nb, 0.1 % V or 0.1 % Ti [382]

Figure 239 gives a summary of the results, which show the relationship between increase in yield stress and change in transition temperature. Only grain refining (e) causes a significant reduction, i.e., improvement, in the transition temperature, with a simultaneous increase in yield stress. The other mechanisms of strength improvement increase the transition temperature. The effect of Mn and Ni is probably explained by transformation behavior leading to a secondary microstructure that is more finely grained.

Figure 240 shows the important relationship between the amount of added Nb, V, or Ti and the yield stress for the development of microalloyed fine-grained structural steels [331], [382]. In Figure 240 A, yield stress is plotted against the reciprocal of the square root of grain size, which should give a straight line according to the Hall – Petch equation for grain refining (see also Fig. 220). The yield strength is most effectively increased by the addition of Nb, which causes grain refining and also hardening (to ca. 200 MPa). Titanium and vanadium are less effective grain refiners. Figure 240 B shows the effect on the transition temperature. Although the addition of Nb causes grain refining and hence hardening, it also causes a reduction, i.e., an improvement, in the transition temperature. Vanadium leads to less grain refining and a less favorable transition temperature. Titanium occupies an intermediate position.

The diagram, which has been extended by many other investigations, shows that microalloying elements have various effects. A suitable combination of alloying elements can be used, depending on rolling conditions and intended operational conditions.

Accompanying elements in steel also affect its toughness properties. Sulfur is picked up from the coke in the blast furnace, but most is removed by the blast furnace slag and in the steelworks. The remaining amount (<0.02 %) leads to some deterioration of the toughness properties. Sulfur is practically insoluble in iron in the solid state and, at room temperature, is present as the sulfide, but mainly as MnS in structural steels. The sulfides are deformed during hot rolling, and mainly affect the toughness properties in the upper plateau. Figure 241 shows the relationship between the rolling conditions and the results of notch impact tests with various test piece orientations. These can be explained by assuming that the interface between the matrix and the inclusions promotes the fracture. An increase in the degree of deformation affects the differences in the direction of stress by extension of the sulfides. This is also shown in Figure 242, which shows the effect of sul-

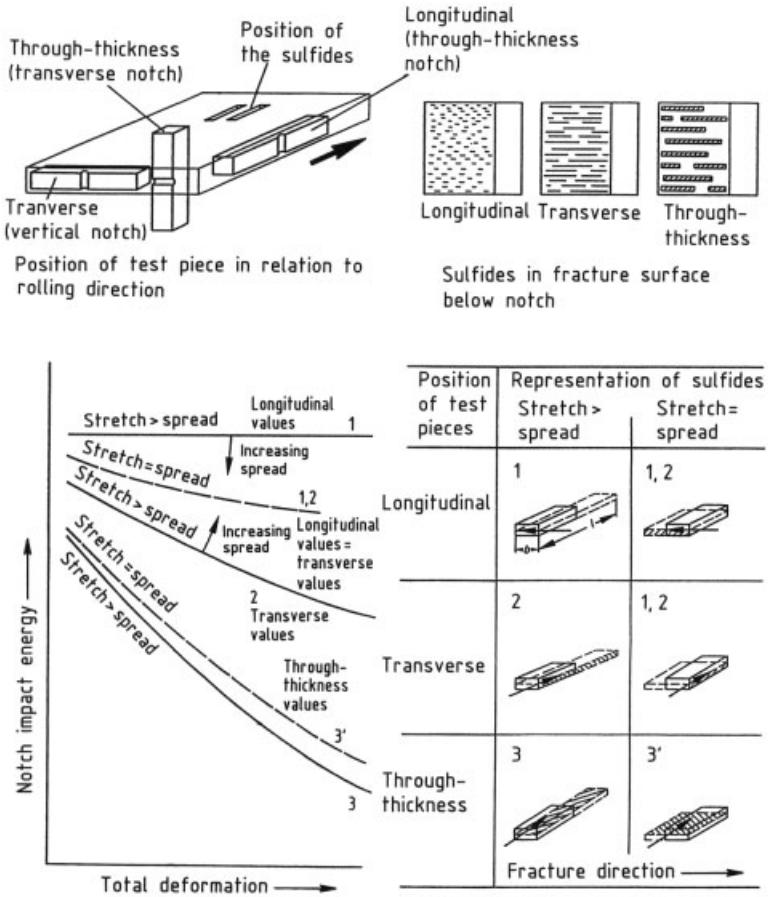

Figure 241. Relationships between notch impact energy, shape of sulfides, orientation of the test piece, and degree of deformation for various ratios of stretch to spread

fur content and that of desulfurization by CaSi treatment. This treatment not only reduces the sulfur content to $< 0.004\,\%$, but also increases the sulfide hardness, which reduces the stretching of the sulfide. The effect on the anisotropy is shown in Figure 243 A, which presents the notch toughness – temperature curves for longitudinal, transverse, and through-thickness test pieces taken from thick steel plate of conventional manufacture; the corresponding curves in Figure 243 B show considerably higher values in the upper plateau and an almost isotropic behavior after treatment with CaSi.

Sulfides have similar effects on the reduction in area in the thickness direction. Figure 244 shows typical values for the reduction of area when breaking FeE 355 steel plate in the thickness direction. It can be seen that the most frequent value of 13 % (for sulfur content 0.015 – 0.054 %) is increased to 68 % after treatment with CaSi, and the scatter of results is considerably reduced.

Phosphorus also has an unfavorable effect on toughness, mainly because of segregation at the grain boundaries. Figure 245 shows the transition temperature corresponding to 50 % ductile fractures in notch tests for 20MnMoNi55 steel as a function of phosphorus content. However, investigations have shown that phosphorus has hardly any effect on the toughness of thick (40 mm) FeE 355 steel plate. Notch impact energy – temperature curves for transverse test pieces with phosphorus content 0.002 – 0.13 % are compared in Figure 246.

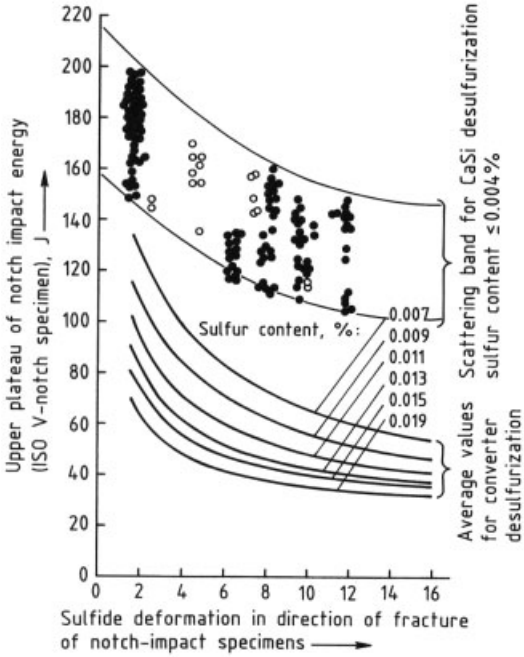

Figure 242. Effect of degree of elongation of sulfides, sulfur content, and method of desulfurization on the upper plateau of the notch impact energy (ISO V-notch specimen) for thermomechanically treated structural steels [383]
● = Nb steel; ○ = Nb – V steel

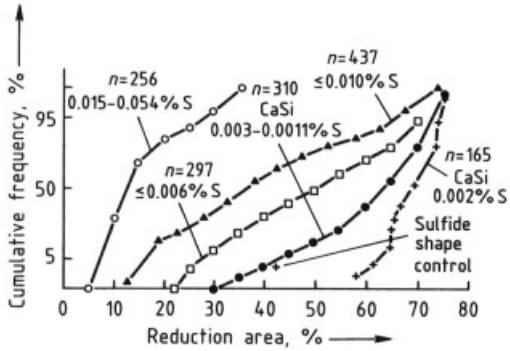

Figure 244. Effect of sulfur content and sulfide shape on the reduction in area in the thickness direction of steel plate (FeE 355) [385]

Microstructures with Optimal Combinations of Strength and Toughness [388–391]. For an optimum combination of strength and toughness, particular attention must be paid to the values in the upper plateau or the transition temperature region, depending on the intended use. Figure 247 A shows the effect of various alloying elements on yield stress and transition

(A)

Sampling	Longitudinal	Transverse	Through-thickness direction
Yield point, N/mm²	396	402	x 384 S 7.2
Tensil strength, N/mm²	624	628	x 561 S 16.0
Reduction in area, %	64	50	x 9 S 2.3

(B)

Sampling	Longitudinal	Transverse	Through-thickness direction
Yield point, N/mm²	426	439	x 423 S 4.3
Tensile strength, N/mm²	602	619	x 588 S 5.8
Reduction in area, %	77	74	x 74 S 1.9

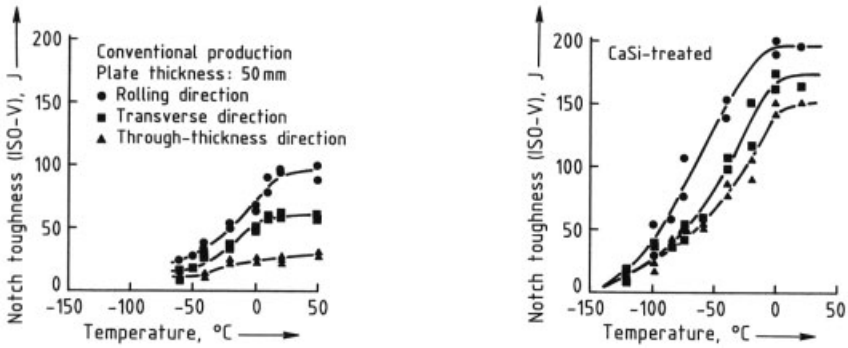

Figure 243. Notch impact energy of FG 36 steel [384]
A) Before CaSi treatment; B) After CaSi treatment

temperatures; Figure 247 B shows the effect of precipitation hardening and grain refining. Grain refining has a beneficial effect on both yield stress and transition temperature, whereas the other strength improvement mechanisms increase the transition temperature. If the transition temperatures are plotted against the yield stress, the connection between toughness and strength can be seen in zones corresponding to the various steel microstructures (Fig. 248).

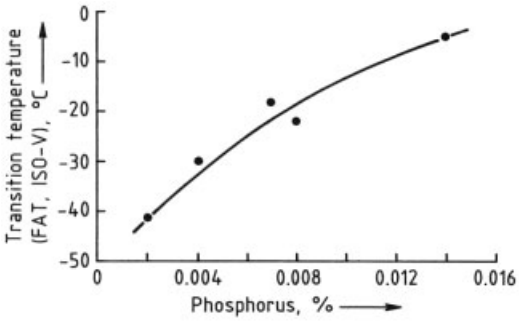

Figure 245. Effect of phosphorus on toughness of 20MnMoNi55 steel [386]

Favorable transition temperatures can be achieved by thermomechanical treatment with constant strength, or, conversely, the strength can be higher for constant transition temperature. Further optimization is possible by producing a bainitic structure, albeit at higher alloying cost, as the transformation is usually promoted by the addition of ca. 0.3 % Mo. Tempering steels show the best combination of properties, but the cost of the alloy and of the tempering process must be considered.

Figure 249 shows how the systematic development of steel products and their thermomechanical treatment has enabled steels with yield strengths up to 650 MPa and transition temperatures of $-80\,^{\circ}C$ to be produced, although these are only laboratory-scale investigations.

6.1.2.1.2. Behavior Under Cyclic Loading

Reverse Strain (Bauschinger Effect) [392]. If, after an initial forming operation in a certain direction, the direction of forming is reversed, flow begins at lower stress (Bauschinger effect) [393]. Figure 250 shows the result of reversal of the forming direction for various steels. Figure 250 A shows work hardening of the materials St 37-3, C 35, and C 70 under tensile load. If the stress direction is reversed after a given amount of prestrain, values of the crushing yield

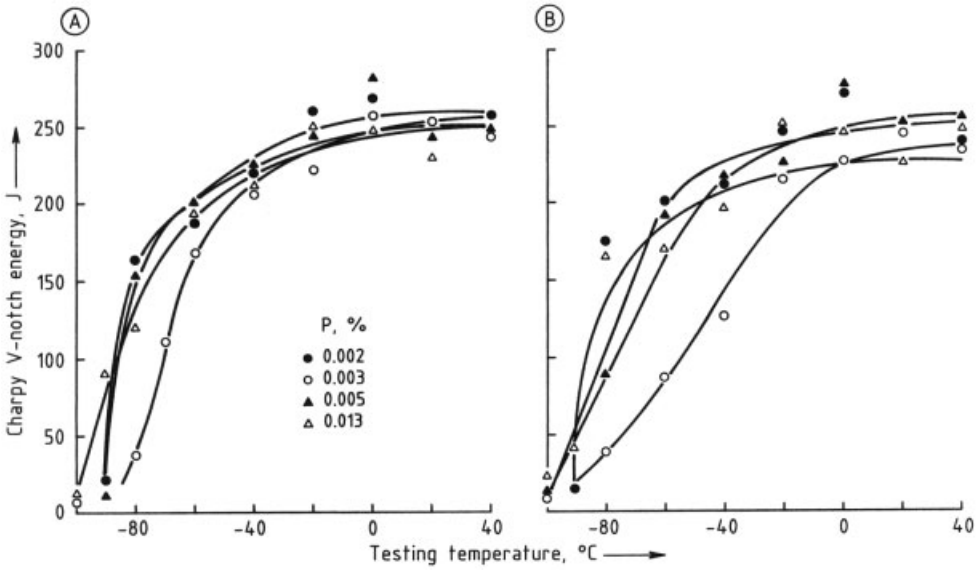

Figure 246. Effect of phosphorus on the notched impact energy of 40 mm heavy steel plate (FeE 355) [387]
A) Surface; B) Interior

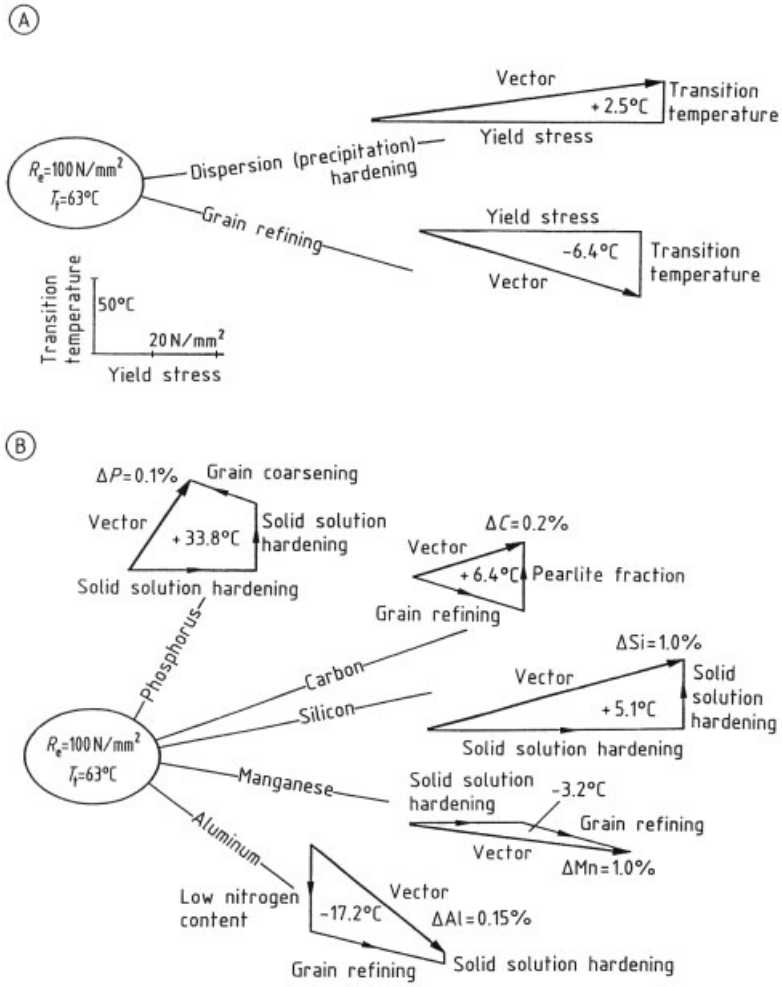

Figure 247. Methods for optimizing yield stress and transition temperature of the notch impact energy of steel [388]
Basic material: pure iron, yield stress 100 MPa, transition temperature 63 °C
A) Influence of alloying elements; B) Effect of grain refining and precipitation hardeningNumbers within vector triangles show the shift of transition temperature of the notch impact energy (T_t) given as °C per 10 N/mm^2

point $\sigma_{q0.2}$ (shown by a continuous line) are obtained. Even for small amounts of prestrain, the crushing yield point decreases steeply, and then slowly begins to increase. The effect increases with increasing carbon content. After low-stress annealing, $R_{p0.2} \approx \sigma_{q0.2}$. Figure 250 B shows prestrain by compression, followed by a tensile test.

The Bauschinger effect is due to internal (residual) stresses. In single-phase materials, stresses are formed between dislocations, and these favor movement in the opposite direction. In polycrystalline or multiphase materials, internal stresses are produced between parts of the microstructure with different strength properties. The directions of these are such that they also reduce the stress of deformation in the reverse direction.

Cyclic Loading (Fatigue) [394–397].
When very frequent reversals of stress take place in metallic materials, irreversible changes occur if the stress maxima lie appreciably below the yield point. As the number of oscillations increases, cracks begin to form. After these have grown to some extent, fracture occurs. Typical

types of stress are illustrated in Figure 251, and the results are shown in Figure 252, an extended Wöhler diagram. The number of oscillations is increased at various amplitudes, until the first strain structures appear, i.e., permanent damage begins, leading eventually to fracture. Below the fatigue strength, no failure takes place, even with a high number of oscillations, and no traces of deformation are observed below the fatigue limit. In steels with well-defined yield points, there are irreversible changes in the phase which is free of incipient cracks. The permanent elongations after oscillating stress are usually considerably less than those found by tensile tests, as individual dislocations become separated from their Cottrell atmosphere, enabling Lüders bands to develop.

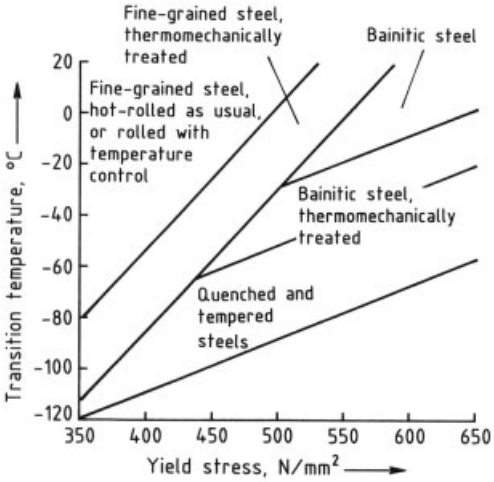

Figure 248. Yield stress and transition temperature of notch impact energy of steels of different microstructure [331]

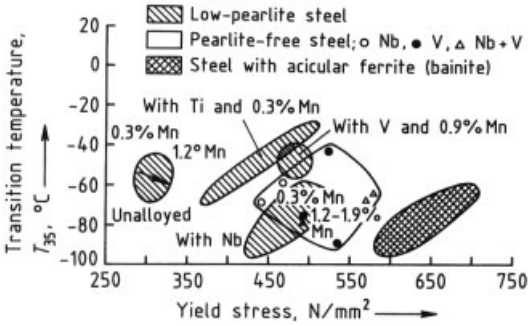

Figure 249. Mechanical properties of hot-rolled wide strip of unalloyed and microalloyed steels of different microstructure [331]
T_{35} = temperature at which the notch impact toughness decreases to 35 J/cm^2; ISO V-notch test specimens

As soon as macroscopically detectable cracks appear, the rate of crack growth follows the Paris law:

$$\mathrm{d}a/\mathrm{d}n = C\,(\Delta K)^{\mathrm{m}}$$

where da/dn is the crack propagation rate, ΔK the stress intensity amplitude, and C and m are constants. This relationship is shown in Figure 253. The crack growth rate lies within a common scatter band for many steels (Fig. 254) if corrosive effects are excluded.

The fatigue strength is proportional to the tensile strength. For tension/compression, reversed symmetric stress applies:

$$0.3 < \sigma_{\mathrm{zdW}}/R_{\mathrm{m}} < 0.45$$

Alloying elements have a similar effect on fatigue strength as on tensile strength. Figure 255 shows their effect on tensile strength R_{m}, and their effect on fatigue limit under reversed bending stresses σ_{zdW}. Purity is of great importance for fatigue strength, as cracks usually start at inclusions just below the surface. Thus, electroslag or electron-beam-melted materials have enabled the fatigue strength of roller bearing steels to be increased considerably [403]. Prestraining increases the fatigue strength if the deformation is greater than the Lüders elongation.

When evaluating components subjected to cyclic stress, the sequence of stresses is very important. Statistical analysis of the stress enables simulation of a fatigue test of a component under operational stress [394], [404]. The sequence effect is a result of the plastic zones and the internal compression stresses which they cause ahead of the crack tip. These slow down further crack growth at higher amplitudes.

6.1.2.1.3. Behavior at Higher Temperatures [405], [406]

In the range between room temperature and ca. 400 °C, interactions between interstitially dissolved carbon or nitrogen atoms and dislocations lead to dynamic strain aging, which causes jerky flow during plastic deformation in the tensile test (see Fig. 207). If the yield points for various elongations are read from this diagram and plotted against temperature, the curve shown in Figure 256 is obtained. Dynamic strain aging

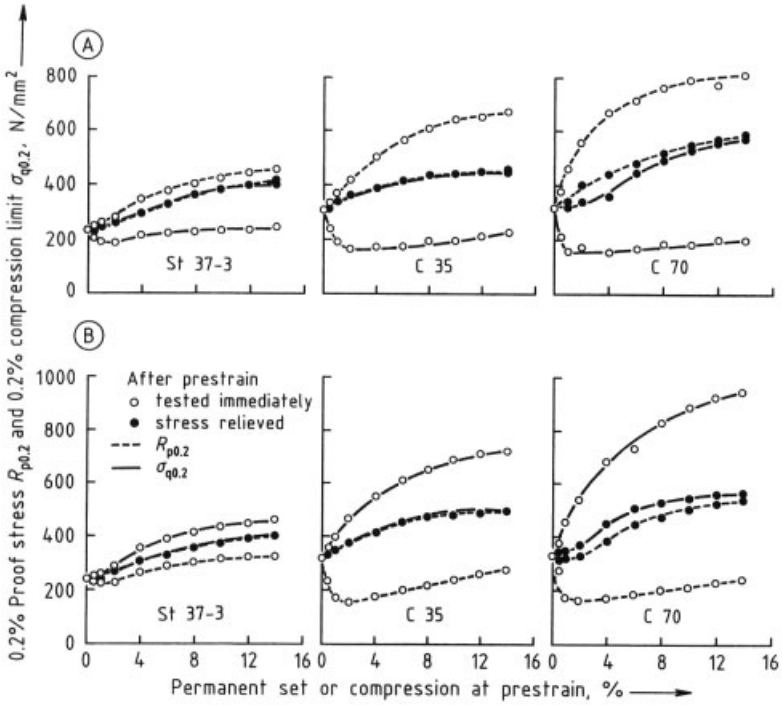

Figure 250. Plots of 0.2 % compression limit ($\sigma_{q0.2}$) and 0.2 % proof stress ($R_{p0.2}$) as a function of the extent and direction of strain before and after stress-relief annealing of various unalloyed steels [393]
A) Prestrain by tension; B) Prestrain by compression

increases the resistance to plastic deformation, and the material is hardened, especially in the medium-deformation range. The toughness decreases (blue brittleness). The maximum interaction is shifted to higher temperature with increasing strain rate [407].

In hot forming, hardness is reduced during or after the forming process, by dynamic or static recrystallization [408]. Changes in yield stress can be determined, e.g., by hot-torsion tests. Figure 257 shows how the yield stress changes with deformation temperature for strain

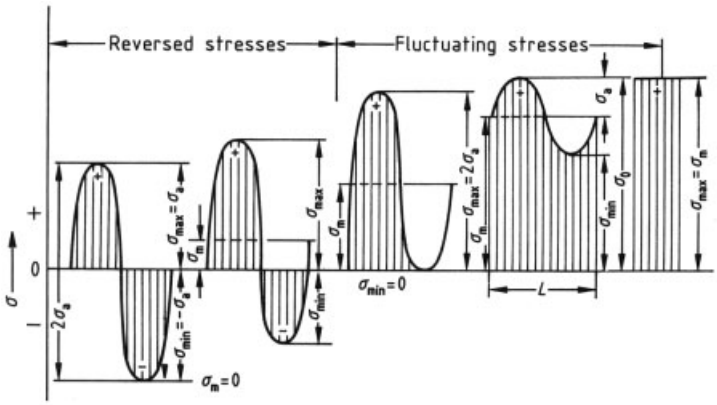

Figure 251. Ratios of mean stress to stress amplitude under cyclic loading [398]
σ_m = mean stress = static prestress; σ_a = stress amplitude; σ_{max} = maximum stress; σ_{min} = minimum stress

values of 0.05, 0.2, and 0.4, at various deformation rates. The yield stress decreases rapidly with increasing temperature, and at the transition from the body- centered to the face- centered cubic structure, an intermediate maximum occurs. Figure 258 shows flow curves at 1100 °C. After an initial hardening, the strength decreases by recrystallization until a stationary state with an approximately constant yield strength level is reached. Repeated changes from hardening to softening can lead to periodic curves.

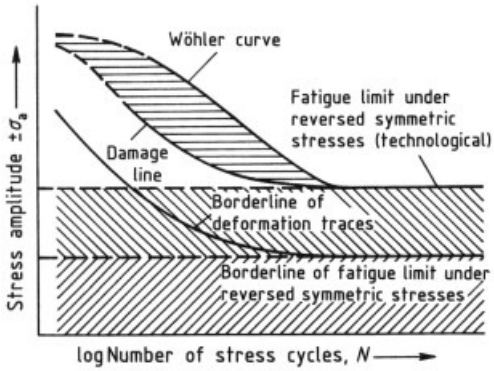

Figure 252. Extended Wöhler diagram [399], [400]

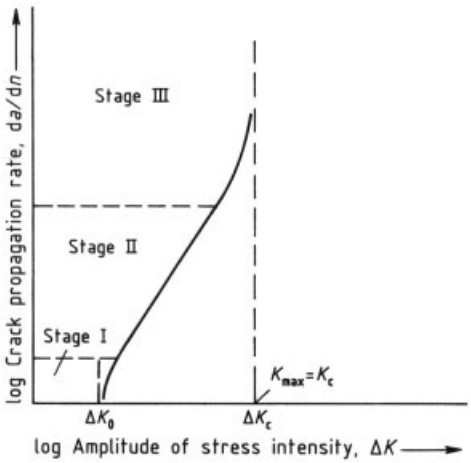

Figure 253. Effect of amplitude of stress intensity ΔK on crack propagation rate d*a*/d*n* [401]

If steel is stressed at high temperature for long periods, creep can take place, i.e., time-dependent deformation at constant stress [406]. Figure 259 A shows that the strain rate is initially high, and then constant. Toward the end of the experiment, there is an acceleration in the creep rate due to void formation and cross-sectional weakening. At higher stresses and temperatures, the phenomena occur more rapidly. Figures 259 B and C show a way of interpreting the creep curves for practical cases. Values for the stress and stress duration are used, leading to a figure for the elongation, or to fracture. This gives the maximum tolerable stress for a given component lifetime. Figures 259 D and E show a method of analyzing mechanical operations. Figure 259 D shows a plot of the logarithm of the strain rate against the logarithm of the strain. A region of constant strain rate, stationary creep, can be seen. If the strain rate is plotted against the stress on a log – log scale, a straight line is obtained which satisfies the equation $\sigma = \varepsilon^n$. This creep law enables the activation energy to be determined from measurements over a range of temperature [411].

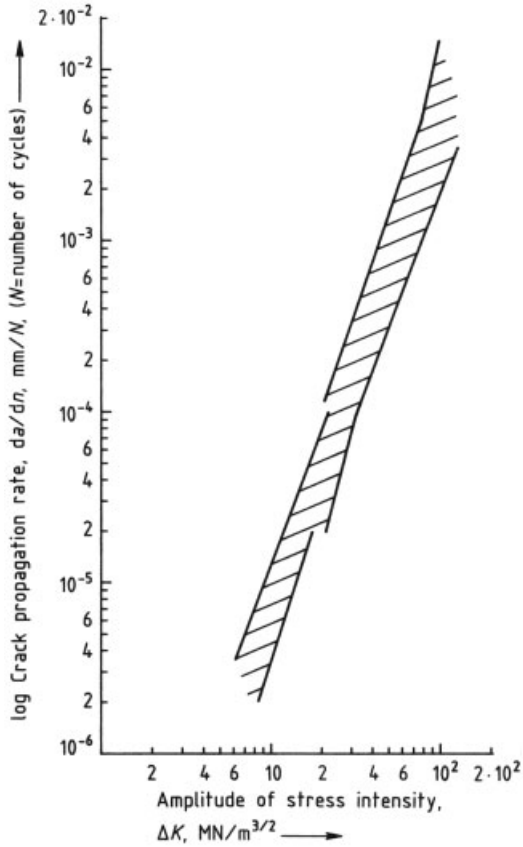

Figure 254. Effect of amplitude of propagation stress intensity ΔK on crack propagation rate d*a*/d*n* for a series of steels with very different chemical compositions and with yield stress 250 – 1680 MPa [363], [401]

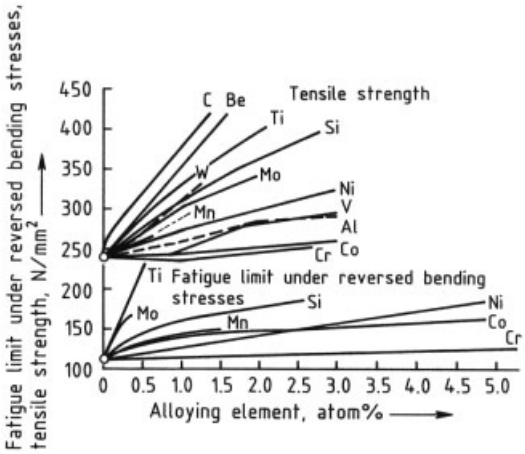

Figure 255. Effect of alloying elements on tensile strength and fatigue limit under reversed bending stress of pure iron [402]

The microstructure must be stable in hot-strength steels. Special carbides that contain molybdenum, chromium, or vanadium coagulate much more slowly than cementite, as the alloying elements also need to diffuse. This explains the principle underlying the alloys used for hot-strength steels, e.g., 15Mo3, 13CrMo4, and 14MoV63. A further increase in hot strength can be obtained with austenitic steels, as transverse slipping processes (which reduce the dislocation density and hence the strength) are hindered owing to the lower stacking fault energy.

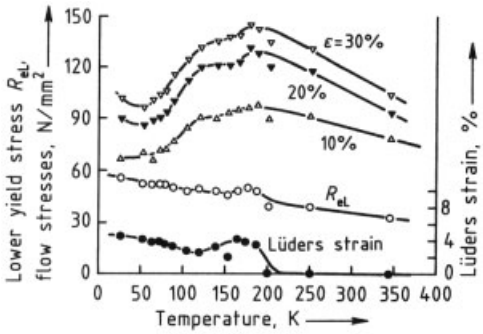

Figure 256. Effect of temperature on yield stresses for elongation $\varepsilon = 10\,\%$, $20\,\%$, and $30\,\%$, lower yield stress R_{eL}, and Lüders strain, derived from Figure 208 [320]

Creep curves of a hot-strength austenitic steel at 700 °C under various tensile stresses are shown in Figure 260 A. The interpretation in Figure 260 B shows creep limit and creep stress rupture curves. Creep stress rupture curves are determined at various temperatures (Fig. 261). The shape of these curves, which drop away, explains why extrapolation to longer times is difficult. There is also a scatter in the results associated with the materials and the test methods used. It is therefore very difficult to predict results at high temperatures and long times from tests at low temperatures and short times. Elab-

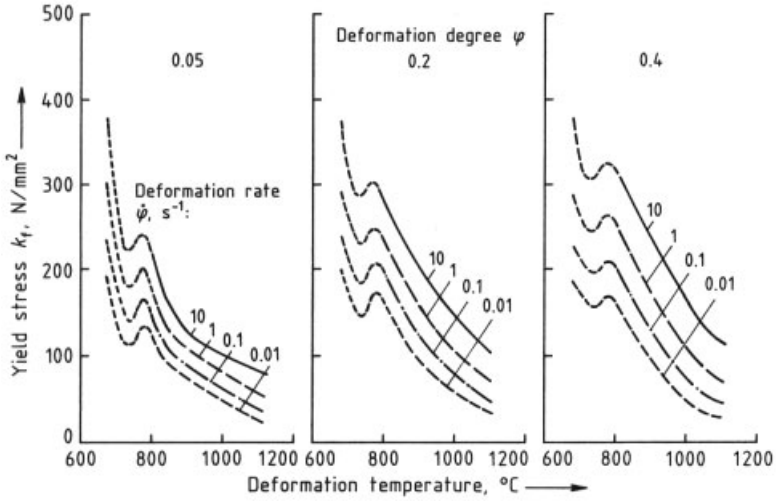

Figure 257. Yield stress as a function of the deformation temperature for various deformation rates and degrees of deformation [409]

orate and prolonged tests must be carried out to predict creep behavior with confidence.

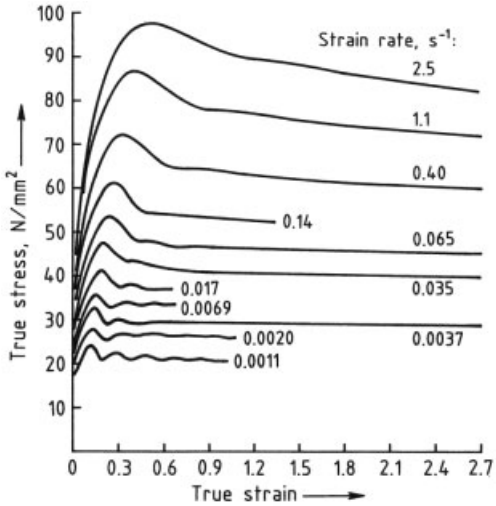

Figure 258. True stress – strain curves of a steel with 0.25 % C, deformed at 1100 °C in a hot-torsion test, i.e., in the austenitic state [410]

Steels, including austenitic steels, can be used at temperatures up to 800 °C, depending on operating conditions. Higher temperatures require the use of superalloys based on nickel or cobalt, sometimes strengthened by dispersed or fibrous materials. Ceramics provide an alternative approach.

The many varieties of steel show a broad spectrum of properties, determined by their microstructure. By the use of appropriate alloys, forming operations, and heat treatments, combinations of properties can be produced which place steel among the most important materials of modern technology. Examples of its use are given in Section 6.2.

6.1.2.2. Chemical Properties

6.1.2.2.1. Introduction

Iron is the most widely used metallic material and, alloyed with different elements, it is the basis of a wide variety of steels, with a range of mechanical properties and chemical stability. Iron itself is rather reactive, readily forming oxides, hydroxides, sulfides, etc. It undergoes an electrochemical reaction with humid air (rusting) which can rapidly consume unprotected unalloyed steel. The rust layer is not protective, unlike the thin, dense oxide layer on aluminum or zinc, but similar behavior can be attained by alloying Fe with Cr (> 12 % Cr gives stainless steels).

Iron is also very susceptible to oxidation or sulfidation at high temperature; FeO and FeS show a wide homogeneity range, correlated to high concentrations of lattice defects (cation vacancies), high diffusivities in the lattice, and high growth rates. Alloying can lead to high oxidation resistance; Cr steels or Ni – Cr steels with ca. 20 % Cr can be used up to 1000 °C, and Fe – Cr – Al alloys at even higher temperatures, since they form slow-growing oxide layers, which also give limited protection in environments which are sulfidizing, carburizing, or chloriding (heat exchangers in combustion or gasification atmospheres).

The chemical behavior of iron and steels is of greatest interest in connection with corrosion problems (→ Corrosion); in the following, reactions and corrosion processes are described only briefly, emphasizing the special points of chemical behavior of iron and steels.

6.1.2.2.2. Uniform Corrosion

Corrosion of iron in aqueous electrolytes results from the anodic dissolution process:

$$Fe \longrightarrow Fe^{2+} + 2e^- \tag{6.1}$$

and the simultaneous, independent cathodic reduction of an oxidizing agent, in most cases either hydrogen ions:

$$2\,H^+ + 2\,e^- \longrightarrow 2\,H_{ad} \longrightarrow H_2 \tag{6.2}$$

or oxygen molecules dissolved in the electrolyte:

$$(O_2)_{aq} + 2\,H_2O + 4\,e^- \longrightarrow 4\,OH^- \tag{6.3}$$

The potential dependence of reactions (6.1) and (6.2) is described by the Volmer – Butler equation, an exponential relation; oxygen reduction is mostly controlled by diffusion of molecular oxygen to the metal surface, independent of potential. The overlap of anodic and cathodic reactions is shown in Figure 262. The corrosion potential U_k is established at a negative value,

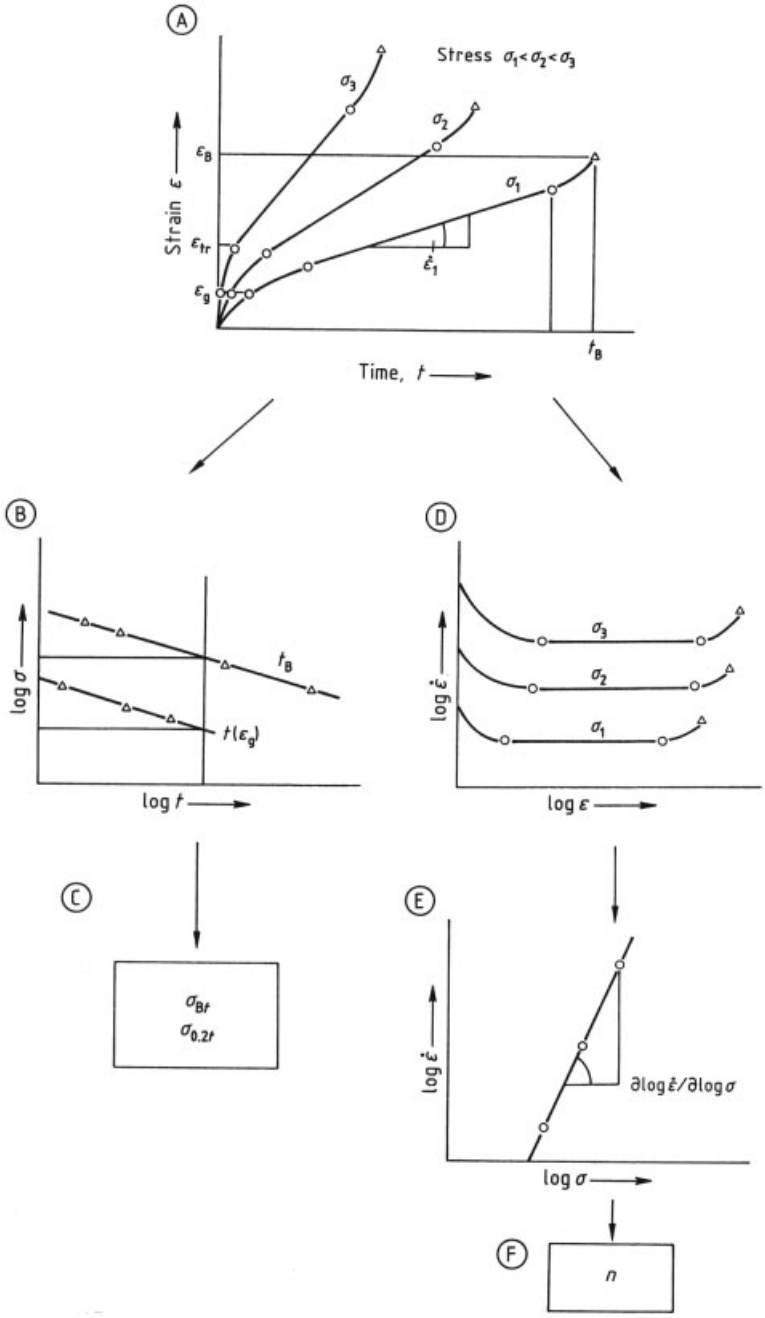

Figure 259. Interpretation of creep tests at constant temperature
Relationships between strain ε, strain rate $\dot{\varepsilon}$, time t, and stress σ [406]
A) Time – strain curves at constant stresses; B) – F) Evaluations of these curves

where the current density of anodic dissolution i_a and oxygen reduction i_k are equal. The anodic dissolution is further dependent on the concentration of defects in the metal surface, and especially for iron, is enhanced with increasing concentration of OH^- [413], and also SH^- or NH_4^+ [414].

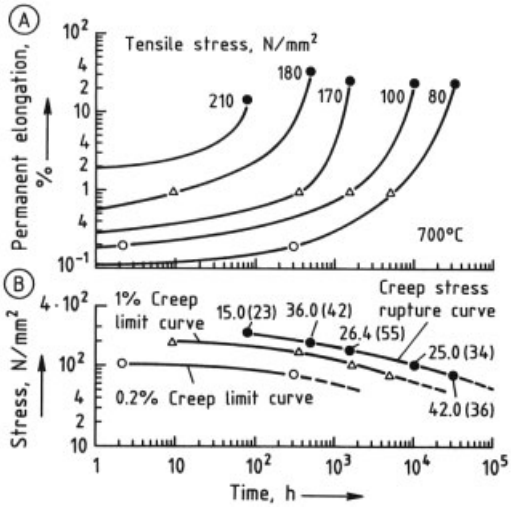

Figure 260. Evaluation of creep tests at 700 °C on a high-temperature austenitic steel with ca. 0.12 % C, 20 % Co, 21 % Cr, 3.2 % Mo, 20 % Ni, and 2.5 % W (X12CrCONi2120) [412]
A) Creep curves; B) Creep limit and creep stress rupture curves

The reduction of H^+ occurs in several steps; electron transfer occurs by the Volmer reaction:

$$H^+ + e^- \longrightarrow H_{ad} \tag{6.4}$$

The adsorbed H_{ad} atoms can react to H_2, according to the Tafel reaction:

$$2\,H_{ad} \longrightarrow H_2\,(g) \tag{6.5}$$

or the Heyrowski reaction:

$$H_{ad} + H^+ + e^- \longrightarrow H_2\,(g) \tag{6.6}$$

Adsorbed H atoms also can enter the metal, which leads to hydrogen-induced cracking (HIC) or hydrogen-induced stress corrosion cracking (HISCC).

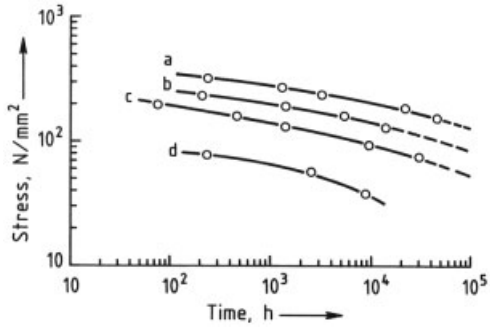

Figure 261. Creep stress rupture curves of a high-temperature Cr – Ni steel at various test temperatures [412]
a) 600 °C; b) 650 °C; c) 700 °C; d) 800 °C

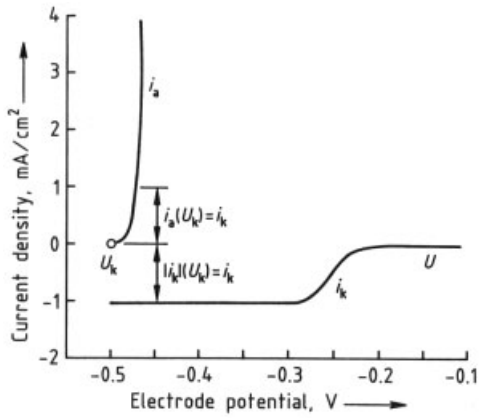

Figure 262. Electrochemistry of iron corrosion

In neutral water, the corrosion rate is determined by H^+ diffusion to the metal surface, and the anodic dissolution rate is restricted to 0.2 µA/cm^2, corresponding to 2×10^{-4} cm/a. Since the exchange current densities i_0 of the Volmer and Tafel reactions are very low, and since that of the Heyrowski reaction is negligibly small [415], iron does not corrode in neutral water free of oxygen. In contrast, the corrosion of iron in neutral water saturated with oxygen (in equilibrium with air) is rapid – the calculation assuming rate- controlling O_2 diffusion in a slowly flowing electrolyte yields 0.15 mA/cm^2, corresponding to a corrosion loss of 0.15 cm/a. This high value is not reached under natural conditions, since rust is formed, and the oxygen must diffuse through the pores of this corrosion product.

6.1.2.2.3. Atmospheric Corrosion

In corrosion by oxygen- containing electrolytes, Fe^{2+} is oxidized to Fe^{3+}, and oxides and oxyhydrates are precipitated on the metal surface:

$$4\,Fe^{2+} + 3\,O_2 + 2\,H_2O \longrightarrow 4\,FeOOH \qquad (6.7)$$

$$6\,Fe^{2+} + O_2 + 12\,OH^- \longrightarrow 2\,Fe_3O_4 + 6\,H_2O \qquad (6.8)$$

These reactions occur on the surface of the solid corrosion products and not in the electrolyte. Atmospheric corrosion (rusting) in the presence of a liquid electrolyte film [416] on the metal leads to Fe_3O_4 at the iron surface, and FeOOH and Fe_2O_3 at the phase boundary with the atmosphere. With restricted oxygen access, when all pores of the rust are filled with water, Fe^{2+} formed at the inner phase boundary [417] can react with FeOOH:

$$Fe^{2+} + 2\,FeOOH + 2\,OH^- \longrightarrow Fe_3O_4 + 2\,H_2O \qquad (6.9)$$

On drying the rust layer, allowing better access of oxygen, Fe_3O_4 is oxidized again to FeOOH. Atmospheric corrosion can be retarded by certain alloying elements: Cu, P, Cr, and Ni, which are added to rust-resistant steels, e.g., Corten. Sulfur dioxide strongly accelerates rusting of steels by initiating the reaction sequence [418]:

$$4\,Fe + 4\,SO_2 + 2\,O_2 + 4\,H_2O \longrightarrow 4\,FeSO_4 + 4\,H_2 \qquad (6.10)$$

$$4\,FeSO_4 + 6\,H_2O + O_2 \longrightarrow 4\,FeOOH + 4\,H_2SO_4 \qquad (6.11)$$

In this reaction sequence, one SO_2 molecule can transfer 15 – 150 Fe atoms into the rust [419] by its catalytic action, until the reaction is stopped by formation of insoluble basic iron sulfates.

6.1.2.2.4. Passivation

The passivation behavior of iron is shown in Figure 263 [420]. In region 1, the current – potential curve describes active anodic dissolution; according to the Butler – Volmer equation, the increase in current is exponential. In region 2, the increase is limited since the solubility limit of $FeSO_4$ is approached at the metal surface – corrosion is controlled by outward diffusion of the Fe^{2+} through the diffusion boundary layer, independent of potential. At the passivation (Flade) potential U_F the current density abruptly decreases, owing to formation of a passive layer, i.e., a very thin oxide layer, most probably γ-Fe_2O_3 [421], a few nanometers thick, depending on potential and temperature. In the passive region 3, a steady state of continuous film growth and dissolution is established, at current density i_p ca. 10^{-5} A/cm^2, corresponding to a corrosion rate 0.1 mm/a. In the transpassive region 4, enhanced iron dissolution and oxygen formation occur simultaneously. For the passivation potential of iron in the acidic region (in V):

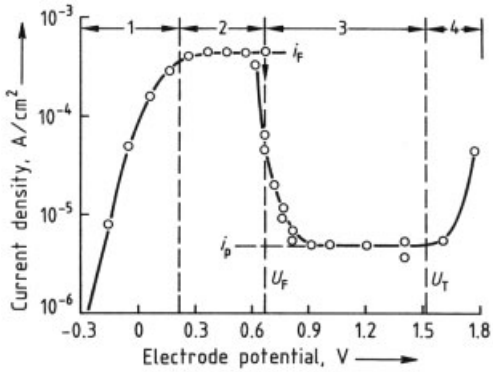

Figure 263. Electrochemistry of passivation of iron in HCl 1 mol/L at 25 °C
U_T = transpassivation potential

$$U_F = 0.5 - 0.005\,pH \qquad (6.12)$$

Iron is passivated at pH > 10 by oxygen in the electrolyte, this is important for the use of reinforcing steel in concrete, because the electrolyte in the pores of the concrete is saturated with $Ca(OH)_2$ at pH 12.5. Iron can also be passivated in strong oxidizing acids such as HNO_3.

The current – voltage curve for nickel [422] is similar to that for iron; for chromium [423] the passivation potential is shifted to more negative potentials, and current density in the passive state is $< 10^{-7}$ A/cm^2 (Fig. 264). In the passive region, Cr^{3+} is formed and very slowly dissolved, in the transpassive region, formation of chromate and dichromate with hexavalent Cr begins. This behavior is reflected in the passivation of Fe – Cr and Fe – Ni – Cr steels [424]; with increasing Cr content the passivation potential is shifted to lower values and the current density in the passive region is decreased, so stainless steels are passive over a wide range of conditions. The breakthrough potential at which Cr^{6+}

formation starts, is virtually independent of Cr content.

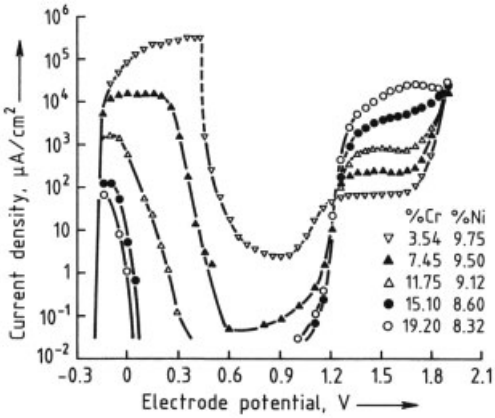

Figure 264. Electrochemistry of iron passivation at various concentrations of Cr and Ni

6.1.2.2.5. Pitting Corrosion

If a passive layer is destroyed locally, the layer may heal under passivating conditions or, in the presence of halide ions, pitting may occur, i.e., local attack with formation of hemispherical pits. During pitting, active and passive regions are stable in the immediate neighborhood. The active areas are stabilized by the potential drop, caused by a high dissolution current through the electrolyte in the pit (the bottom of the pit is at potentials below passivation potential and the surrounding passive film is at a higher potential) [425]. Pitting occurs if the value of the pitting potential U_p is exceeded, and if the corrosive ions have a critical concentration c_L. For Fe – Cr and Fe – Ni – Cr steels, the pitting potential increases with Cr and Mo content (Mo is more effective than Cr). Since then, N has been found to be even more effective [426]. Pitting can start in Cr-depleted zones, e.g., after carbide precipitation at grain boundaries [427].

6.1.2.2.6. Crevice Corrosion

If there are crevices between a metal and a second phase, or in the metal itself, the onset of corrosion is often observed in the crevice, owing to retardation of the electrolyte exchange into and out of the crevice. When the transport of oxidizing agent into the crevice is retarded, leading to depletion of, e.g., dissolved oxygen, the potential may fall below the passivation potential and iron and low-alloy steels may start to corrode. Also, the transport of corrosion products out of the crevice is inhibited, and their hydrolysis may lower the pH, e.g., by the reaction:

$$Cr^{3+} + 3\,H_2O \longrightarrow Cr(OH)_3 + 3\,H^+$$

The increasing acidity reduces the activation and pitting potentials, so active corrosion or pitting may start [428], [429].

6.1.2.2.7. Intergranular Corrosion of Stainless Steels

Ferritic chromium steels and austenitic Cr – Ni steels are susceptible to intergranular corrosion after sensitization, i.e., after precipitation of carbides at 450 – 750 °C. Precipitation at grain boundaries [430], [431] leads to chromium depletion nearby. If the potential is in the range between the passivation potentials of the steel and iron, rapid dissolution of the metal phase occurs along the grain boundaries. Potentials in this range are established in the Strauss test [432], where the sample is in contact with Cu and immersed in $CuSO_4 - H_2SO_4$ solution, so that sensitized materials undergo intergranular disintegration. In the temperature – time diagram, the conditions for sensitization are limited by curves for the onset of carbide precipitation, which tends to shorter times at higher C concentration and longer times at lower temperature, and by curves for the end of carbide precipitation, which also needs shorter times at higher temperature. After the end of precipitation, the chromium concentration equalizes and the material becomes corrosion resistant again.

This intergranular corrosion of stainless steels can be suppressed by very low carbon concentrations < 0.01 % (Fig. 265), or by addition of stabilizing elements, such as Ti and Nb, which tie up carbon in very stable carbides [433], [434]. Precipitation of TiC and NbC can be induced by heat treatment at 900 – 950 °C (stabilization), where the chromium carbides dissolve and the solubility of TiC and NbC is still very low; such treatment is important after welding of steels.

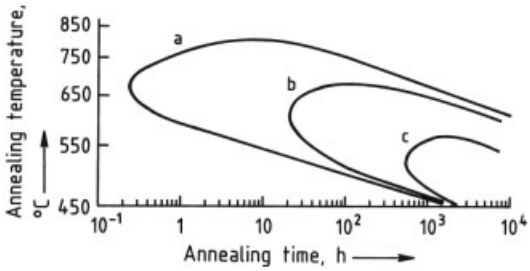

Figure 265. Annealing behavior of steel at various carbon concentrations
a) 0.04 % C; b) 0.02 % C; c) 0.01 % C

6.1.2.2.8. Stress Corrosion Cracking

The preconditions for stress corrosion cracking (SCC) are: (1)

the stress causes a shear or strain of the material; and (2)

the electrolyte induces formation of a passive or protective layer [435], [436]. The strain leads to failure of the passive layer at a slip step; there the free material is rapidly attacked under anodic dissolution. Repeated failure at this spot initiates crack growth. The electrolyte forms a new protective film at the walls of the crack and only the crack tip stays active, leading to deep cracks. Strain rate and passivation rate must obey certain relations to obtain SCC; if, for example, the passivation rate is too high, this leads to immediate passivation of the whole crack and termination of crack growth, whereas too slow passivation leads to strong truncation of the crack, or widening to a deep groove. SCC is observed in the following cases:

1) Austenitic steels in chloride solutions – transgranular crack growth [437–441]
2) Ferritic and ferritic – pearlitic structural steels in hot concentrated alkali carbonate and nitrate solutions – intergranular crack growth [442–446]

The latter type of SCC is enhanced by the presence of impurity elements, such as P, at grain boundaries, similar to the intergranular corrosion of iron and ferritic steels in nitrates at elevated potentials [446]. Under stress conditions, the selective corrosion of the crack tip causes SCC at low P concentrations and without anodic polarization.

6.1.2.2.9. Hydrogen Adsorption and Hydrogen Embrittlement

On the surface of iron or steel, immersed in aqueous solution, the interplay of reactions (6.1) – (6.6) establishes coverage with adsorbed hydrogen H_{ad}, which may correspond to a high hydrogen activity ($a_H = 1$ at $p(H_2) = 100$ kPa), depending on electrochemical potential E and pH:

$$\log a_H = E\frac{F}{2.3RT} + \mathrm{pH} = \frac{E}{0.059} + \mathrm{pH} \qquad (6.13)$$

especially if the Tafel reaction (6.5) is poisoned by inhibitors, such as arsenic acid in the electrolyte. In a neutral NaCl solution free of oxygen, the corrosion potential E of iron is -0.6 V, and hydrogen activity can reach 10^3. Correspondingly, hydrogen is absorbed into the iron matrix in much higher concentration than its solubility at $p(H_2) = 100$ kPa, which at 25 °C is ca. 3 atoms per 10^8 iron atoms [447], [448]. Not all hydrogen in steels is dissolved, i.e., distributed on interstitial sites; most is in special sites (traps) where its energy is decreased, owing to stress fields in the lattice around dislocations, inclusions, etc. [449]. High hydrogen concentration in steels causes hydrogen embrittlement, i.e., decrease of toughness and brittle fracture. This effect of hydrogen is explained by H enrichment at the crack tip, where it weakens the Fe – Fe bonds in the already strained regions at the tip, and decreases the surface energy of the newly formed surface in the crack [450]. By this mechanism the presence of hydrogen can cause hydrogen-induced cracking (HIC), and aggravate stress corrosion cracking (HISCC). High-strength steels are especially susceptible (→ Corrosion).

6.1.2.2.10. Oxidation of Iron

Iron forms three oxides, hematite Fe_2O_3, magnetite Fe_3O_4, and wustite FeO which is stable only above 570 °C [451–453]. FeO has a wide range of homogeneity, due to cation vacancies V_{Fe}^{2-}, for which the concentration is dependent on oxygen pressure. Oxidation of iron at temperatures > 570 °C yields fast-growing oxide layers composed of an outer layer of Fe_2O_3, then

Fe_3O_4, and a thick inner layer of FeO [454], [455]. The oxides grow mainly by outward diffusion of iron via cation vacancies, which move inward. Since the cation vacancy concentration is highest in FeO, the diffusivity of Fe^{2+} is high, and the growth rate is most rapid.

Owing to the high oxide growth rate, unalloyed steels cannot be applied at temperatures $> 570\,°C$, but perform reasonably well at lower temperatures, since the scale of hematite and magnetite grows slowly enough. In CO_2-CO and H_2O-H_2 mixtures at high temperatures, oxygen partial pressures are effective at which only FeO or a Fe_3O_4/FeO layer is formed. Under such conditions, the growth rate of FeO up to a thickness of about 100 µm is controlled by the phase boundary reaction on the outer surface [456–458], where the oxygen is transferred from the gases:

$$CO_2 \longrightarrow CO + O_{ad} \tag{6.14}$$

$$H_2O \longrightarrow H_2 + O_{ad} \tag{6.15}$$

In this stage of oxidation, a linear rate law applies for the increase of scale thickness:

$$\frac{dx}{dt} = k_1 \tag{6.16}$$

where the linear rate constant k_1 is a function of the partial pressures, the oxygen activity at the surface, and the distance from equilibrium [458–460]. There is a gradual transition from phase boundary reaction control to diffusion-controlled growth, where the rate is inversely proportional to the thickness x:

$$\frac{dx}{dt} = \frac{k_p}{x} \tag{6.17}$$

This results in the parabolic law:

$$x^2 = 2k_p t \tag{6.18}$$

which leads to growth rate decreasing with time. Such parabolic behavior is characteristic of most metals and alloys used at high temperatures, the lower the value of k_p the slower the growth of the scale. Values of k_p can be deduced according to the theory of C. Wagner [461] from data on the diffusivities of cations and anions or electric conductivity and transference numbers of the ions.

6.1.2.2.11. Oxidation of Carbon Steels and Low-Alloy Steels

On heat treatment and hot rolling, steels are oxidized and decarburized [462–464]. Oxidation is described by a linear law in CO_2 and a parabolic law in O_2; both reactions behave additively. At first only iron is oxidized and carbon is enriched in the metal below the scale. With continued scale growth, vacancies condense at the interface to form voids in which CO_2 and CO are formed in equilibrium with FeO. This gas diffuses outward through cracks and pores of the scale, which leads to carbon loss. Carbon diffusion through the solid oxide is impossible. After some time, a steady state is established in which the carbon supply to the interface is controlled by diffusion in the metal. Thus, after a brief oxidation, carbon enrichment is observed, but after long-term oxidation, the steel is always decarburized.

Some other more noble alloying (or impurity) elements in steels, e.g., Ni, Co, Cu, Sn, Pb, As, and Sb also enrich beneath the scale during oxidation. Enrichment of Cu, Sn, As, or Sb decreases the melting point, which can lead to difficulties in hot rolling. Specifically Cu, sometimes in combination with Sn, causes "hot shortness" by formation of a melt which penetrates into grain boundaries. If other elements are enriched beneath the scale, the iron must diffuse through this zone to the oxide – metal interface before it is oxidized. Especially for Fe – Ni and high-Ni steels [465–467], this causes inward growth of protrusions from the oxide into the metal which leads to a rugged, strongly interlocked interface; the oxide scale is strongly adherent to the metal and difficult to remove, e.g., after hot rolling. With increasing Ni content, enhanced temperature and oxygen partial pressure, the formation of FeO on Fe – Ni alloys is suppressed in favor of the more slowly growing spinel $(Fe,Ni)_3O_4$.

6.1.2.2.12. High-Temperature Steels

Steels to resist heat and oxidation must be alloyed with sufficient concentrations of Cr and/or Al. The slowly growing oxides Cr_2O_3 or Al_2O_3, which have a small degree of disorder, can form a protective scale. Iron – chromium alloys are the

base materials for many ferritic steels. The oxidation behavior as a function of Cr content is shown in Figure 266 [468]. At Cr concentrations $< 2\,\%$, the oxidation rate is increased since dissolution of Cr_2O_3 in FeO may even increase the amount of cation vacancies (Wagner – Hauffe doping effect); trivalent Cr replaces divalent Fe, this is equalized by negatively charged vacancies. With increasing Cr content wustite formation is suppressed, the spinel $(Fe,Cr)_3O_4$ is formed, for which the growth rate decreases with Cr content. At Cr content 15 –30 %, the mixed oxide $(Fe,Cr)_2O_3$ is formed on the metal surface, its growth rate shows a minimum at 20 – 25 % Cr in the alloy [469]. Many ferritic steels have Cr content 17 – 25 %; they are oxidation resistant, but there may be chromium depletion beneath the scale, and if the scale is damaged outgrowth of wustite may occur with formation of "rosettes."

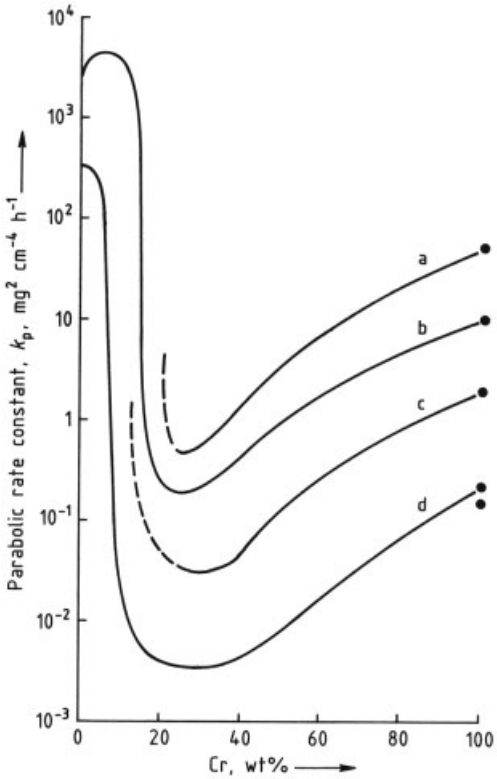

Figure 266. Oxidation kinetics as a function of Cr content a) 1200 °C; b) 1100 °C; c) 1000 °C; d) 900 °C

This effect is suppressed in the Cr – Ni steels where, instead of wustite, a Fe,Ni spinel is formed. Such austenitic steels, with 20 – 27 % Cr and 12 – 32 % Ni, are widely used for high-temperature processes, since, in contrast to ferritic Fe – Cr steels, they have higher mechanical strength and do not show embrittlement by σ-phase formation. Additional alloying with 1 – 2 % Si improves resistance to oxidation and especially carburization, since an inner film of SiO_2 below the Cr_2O_3 retards oxidation and carburization.

Iron – aluminum alloys need 6 – 8 % Al to form a protective Al_2O_3 layer [470], [471], but only at temperatures $> 900\,°C$ is the most protective α-Al_2O_3 generated. Formation of Al_2O_3 and suppression of internal oxidation is favored by chromium; Fe – Cr – Al alloys with ca. 20 % Cr and 4 – 5 % Al are used for heating elements and other highly oxidation-resistant components. The lifetime of the heating elements is greatly increased by small additions of rare earths, e.g., Ce, Y, La, which strongly improve the adherence of the Al_2O_3 scale. The use of such additives also helps to improve other high-temperature alloys, in terms of nucleation, stability, and adherence of protective scales.

Silica layers may also serve as protective scales, but Fe – Si alloys are not used as high-temperature materials. Formation of Fe_2SiO_4 and an inner layer of SiO_2 plays a role in the annealing of silicon sheet for electrical applications.

6.1.2.2.13. Effects of Chlorine in Oxidation

The presence of chlorine may cause acceleration of oxidation, so-called "active oxidation" [472–475]. Especially in waste PVC and NaCl-containing coal incineration, HCl and chlorides from combustion react with the oxide scale on steels to form chlorine

$$82\,HCl + {}^1/_2\,O_2 \longrightarrow H_2O + Cl_2 \qquad (6.19)$$

$$2\,NaCl + Fe_2O_3 + {}^1/_2\,O_2 \longrightarrow Na_2Fe_2O_4 + Cl_2 \qquad (6.20)$$

The chlorine penetrates the scale and reacts at the oxide – metal interface to form $FeCl_2$ which evaporates and oxidizes according to

$$2\,FeCl_2(g) + {}^3/_2\,O_2 \longrightarrow Fe_2O_3 + 4\,Cl_2 \qquad (6.21)$$

The chlorine partially diffuses inward and reacts again, thus establishing a circuit, in which Cl_2

catalyzes the "active corrosion" (active, since no dense oxide layer passivates the steel). This process causes problems in waste incineration and other processes in oxidizing, chloridizing atmospheres. Not only are low-alloy steels strongly attacked, but also chromia-forming high-alloy steels, due to the reaction

$$4\,NaCl + Cr_2O_3 + {}^5/_2\,O_2 \longrightarrow 2\,Na_2CrO_4 + 2\,Cl_2 \qquad (6.22)$$

The chlorine produced penetrates the scale and causes formation of volatile chlorides of the alloy components.

6.1.2.2.14. Sulfidation of Iron and Steel

The main sulfidation products of iron in sulfidizing environments are pyrrhotite FeS and pyrite FeS_2. Pyrrhotite, like FeO, is highly nonstoichiometric, owing to cation vacancies, their concentration being dependent on sulfur partial pressure and temperature [476], [477]. Accordingly, the diffusivity of Fe in FeS is high and the FeS growth rate is very rapid, even at relatively low temperature. As in the case of FeO, the linear rate law (Equation 6.15) applies for growth in $H_2S - H_2$ atmospheres [478], [479]; k_1 depends on $p(H_2S)$ and is inversely proportional to the sulfur activity on the sulfide surface. The rate- controlling step is transfer of S_{ad} from H_2S to the surface, which is retarded with increasing sulfur activity. Upon sulfidation, rate control gradually changes to solid-state diffusion and to the parabolic rate law [480]. Corrosion by H_2S is a problem in the petrochemical and chemical industries, and is not easily suppressed. Even steels with Cr content up to 17 % do not behave much better than low-alloy steels [481].

During corrosion of iron in SO_2 or $Ar - SO_2$ mixtures at 600 – 1000 °C, the phases FeO and FeS grow simultaneously on the metal surface [482–484], if the transport in the gas phase or the phase boundary reaction is rate determining. Owing to the high diffusivities of cations in FeS, lamellae of FeO and FeS grow side by side up to a layer thickness of ca. 150 µm, following a linear law. In the application of steels in flue gases, a scale of Fe_2O_3/Fe_3O_4 is stable on the steel surfaces and SO_2 should have no effect on oxidation. However, if SO_2 penetrates the scale, at the interface sulfides can be formed. The mass action law for the reaction $SO_2 = {}^1/^2\,S_2 + O_2$ predicts that very high sulfur pressures are possible at the oxide – metal interface, since $p(O_2)$ is very low there. This is especially dangerous in high-Ni alloys, since $Ni - Ni_3S_2$ forms a liquid eutectic at 645 °C, and the presence of such melts strongly aggravates corrosion.

Sulfidation of heat exchangers is a problem in coal gasifiers; chromia-forming steels have been tested at 700 – 800 °C [485–489] but fail after lifetimes of some hundreds or thousands of hours. Preoxidation has been used to improve protection, but the chromia layers generated can also fail after similar lifetimes, either by external sulfidation, i.e., growth of sulfides of Fe, Ni, and Mn on the chromia scale, or by internal sulfidation, formation of sulfides of Cr and Mn beneath the scale.

6.1.2.2.15. Carburization

Carburization in gas atmospheres is a technical process for case hardening of steels, a specified carbon content is established in the surface, and the carburized components are quenched in oil for hardening [490], [491]. Carburization can also be an unwanted process in the high-temperature corrosion of materials, e.g., cracking tubes in petrochemistry, and heating tubes or grates in industrial furnaces. In this process, carbon diffuses into the alloy and carbides are precipitated, mainly chromium carbides M_7C_3 and $M_{23}C_6$, with loss of both ductility and oxidation resistance [492], [493].

In both cases, the carbon transfer is by reactions in which CO, CH_4 or hydrocarbons decompose on the metal surface [494], [495], the transfer is very fast by the reaction:

$$CO + H_2 \longrightarrow C + H_2O \qquad (6.23)$$

less fast by:

$$2\,CO \longrightarrow C + CO_2 \qquad (6.24)$$

and very slow by:

$$CH_4 \longrightarrow C + 2\,H_2 \qquad (6.25)$$

Rate equations and rate constants for these reactions have been determined for iron and binary iron alloys. The reactions are strongly retarded by sulfur [496].

In gas carburization for case hardening, the process is jointly controlled by the phase boundary reaction and inward diffusion of carbon; this has been studied in detail, and can be controlled well, to obtain certain penetration profiles of carbon in different steels and corresponding hardness profiles [497].

The unwanted carburization of cracking tubes [492], [493], [498] is generally suppressed, or at least retarded, by the oxide layer on the high-Cr – Ni steels used, only small amounts of carbon are transferred through pores or cracks formed on creep of the tubes. However, failure occurs if the tubes are heated above 1050 °C; the coke in the tubes reacts with Cr_2O_3 to chromium carbides which are no longer protective, and relatively fast carburization takes place [493]. This may happen during decoking of the tubes in $H_2O - H_2$ mixtures, when the temperature is not sufficiently controlled. Under these conditions, materials with 1 – 2 % Si are resistant because of the protective effect of an inner SiO_2 layer beneath the chromia layer. Small amounts of sulfur in the atmosphere can also retard carburization.

Metal dusting [499–503] is the disintegration of steels into a dust of metal particles and carbon, occurring in atmospheres of $a_C > 1$ (carbon activity $a_C = 1$ means equilibrium with graphite) and in recent years has become problematic. The reaction sequence: (1) supersaturation of the metal with dissolved C; (2) formation of metastable carbide at $a_C > 1$; (3) deposition of graphite on this carbide ($a_C \to 1$); (4) decomposition of the carbide to carbon and metal particles; and (5) deposition of further graphite from the atmosphere on the metal particles. This leads to metal wastage and the formation of the loose corrosion product "coke," which is easily removed by fast gas flow in the plants, leaving pits and holes. Protection is possible by the chromia layer on high-alloy steels – their resistance depends on their ability to form oxide layers and healing. Furthermore, sulfur can protect since adsorbed sulfur effectively poisons steps (1), (3), and (5) of the reaction sequence, however, addition of sulfur is not possible with synthesis gas ($CO + H_2$) for the production of methanol, hydrocarbons, etc.

6.1.2.2.16. Nitriding

Similar to carburization, there is a technical process of nitriding for surface hardening, and an unwanted corrosion process at high temperatures which leads to nitride formation in steels and losses in ductility and oxidation resistance.

The nitrogen transfer can be from N_2, $NH_3 - H_2$ mixtures, or by plasma nitriding in a glow discharge. The latter process is performed since N_2 is rather inert; nitrogen transfer:

$$N_2 \longrightarrow 2\,N\ (\text{dissolved}) \tag{6.26}$$

is slow, with high activation energy [504]. Much faster is:

$$NH_3 \longrightarrow N\ (\text{dissolved}) + {}^3/_2\,H_2 \tag{6.27}$$

So $NH_3 - H_2$ mixtures can be used for nitriding and nitrocarburizing (with CO) at relatively low temperature. Their nitrogen partial pressures are [504] very high as calculated from the mass action law; the equilibrium values are of the order of 10^5 kPa N_2. This may lead to deterioration of the nitrided case, by void formation with condensation of dissolved N at lattice defects to molecular N_2, which can develop high pressure, corresponding to the N_2 partial pressure of the $NH_3 - H_2$ mixture, and form small bubbles in the material.

Nitriding for surface hardening leads to a layer of ε-nitride on the steels.

Nitriding of high-alloy steels happens only at high temperature > 1000 °C, e.g., on burner tubes [505], [506] and after failure of the oxide scale by cracking or spalling; mainly chromium nitrides Cr_2N are formed in the metal matrix.

6.1.2.2.17. Decarburization, Denitriding, and Hydrogen Attack

Thorough decarburization of steel is necessary for several applications, e.g., sheet for single-layer enameling, deep drawing steels, and silicon steels for use in generators and transformers. Decarburization is performed at 700 – 900 °C in wet hydrogen, and is controlled by carbon diffusion to the surface. The product $c_c D_c$ (solubility × diffusivity) is critical, and has a maximum at ca. 800 °C [507], [508]. The surface reactions are the formation of CO and CH_4, i.e., the back reactions of Equations (6.23) and (6.25), CH_4 formation being negligible.

Denitrogenation [509], [510] is also important for deep drawing and electric sheet, this is also conducted in wet H_2 at ca. 750 °C, and occurs by NH_3 formation and N_2 desorption, the back reactions of Equations (6.22) and (6.23). The presence of some H_2O accelerates the rate of NH_3 formation [511], [512], but electric sheet with Si cannot be denitrogenated in wet hydrogen since SiO_2 formation retards the process.

In processes which involve high hydrogen partial pressures at elevated temperatures (hydrogenation reactions, ammonia synthesis), considerable amounts of atomic hydrogen enter the walls of steel vessels, and the cementite in unalloyed steels is attacked and methane is formed. Dissolution of the Fe_3C and precipitation of CH_4 and H_2 at interfaces and defects in the steel cause weakening and embrittlement of carbon steels. The more stable carbides in low-alloy Cr and Cr – Mo steels are resistant to hydrogen attack in the usual temperature range for the processes in question; the stability ranges of different steels as functions of temperature and hydrogen partial pressure are compiled in Nelson diagrams [513].

6.1.2.3. Physical Properties

Knowledge of the physical properties of steel is important for controlling its structure, and hence its properties. In this Section, the physical properties of pure iron are described first, followed by those of the solid solutions of the various iron modifications, and finally some other factors that influence structure [514].

6.1.2.3.1. Pure Iron

Crystal Structure. Iron occurs both in body-centered cubic (bcc) and face-centered cubic (fcc) structures. The bcc α-phase, which is stable at lower temperatures, is ferromagnetic up to the Curie temperature $T_c = 1041$ K, and is transformed into the paramagnetic state at higher temperatures (above the A_2 point). At 1184 K, the A_3 point, α-iron is tranformed into the fcc form, γ-iron. The stability range extends up to 1665 K (A_4 point). Between 1665 and 1807 K, iron has a bcc structure, again δ-iron. Pure iron melts at 1807 K.

The crystal structures of α- and γ-iron are shown in Figure 267. The number of atoms in the unit cell is 2 in the bcc structure and 4 in the fcc structure. The number of nearest neighbors is 8 and 12 respectively, and of next nearest neighbors 6. The lattice constant of α-iron at room temperature $a_{RT} = 0.28662$ nm. Thermal expansion data are shown in Figures 268 and 269. In Figure 268, the difference in the y-axis scales for the bcc and fcc structures is due to the difference in the number of atoms per unit cell. The calculated density is $\varrho = 7.876$ g/cm^3.

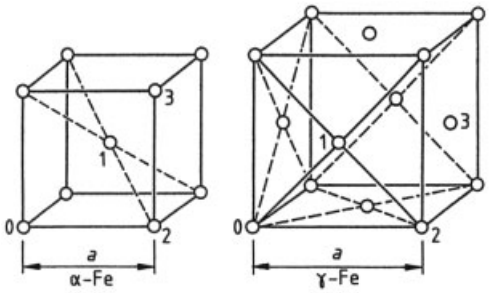

Figure 267. Crystal structure of α- and γ-iron

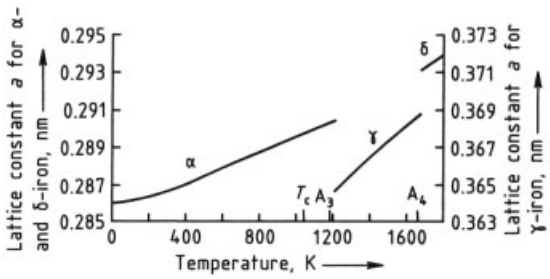

Figure 268. Effect of temperature on the lattice constant of iron

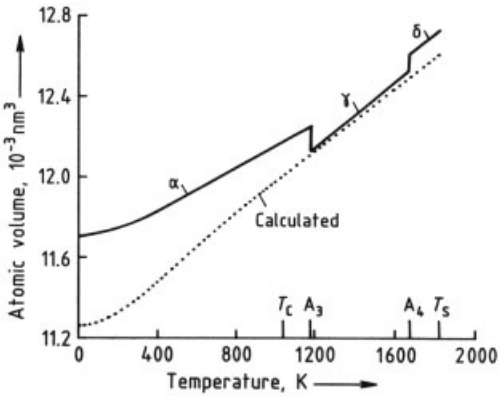

Figure 269. Effect of temperature on the atomic volume of iron

The coefficient of linear thermal expansion α (Fig. 270) increases with increasing temperature, and is higher for the fcc than for the bcc phase. The hatched area between the measured and calculated curves is due to a magnetostrictive effect in the region of the Curie temperature.

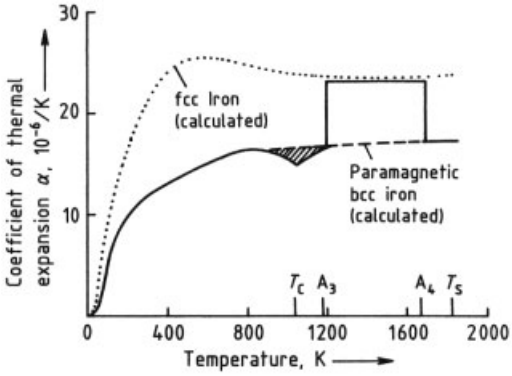

Figure 270. Effect of temperature on the thermal expansion coefficient of iron

Specific Heat Capacity. The specific heat capacity of a substance is the heat required to increase the temperature of unit mass by 1 K. The effect of temperature on its value is of fundamental importance, as the specific heat capacity is used in the calculation of enthalpy and entropy. The results of some determinations are plotted in Figure 271, with the most probable values indicated by a continuous curve [515]. The broken curves refer to the region of instability of the bcc and fcc phases.

The heat capacity is the sum of various components. There is a sharply defined maximum in bcc iron in the region of the Curie temperature. This represents the magnetic component, due to the energy required to destroy the magnetic ordering that occurs during the ferromagnetic – paramagnetic transition.

Elastic Properties. The modulus of elasticity (Young's modulus E) represents the relationship between the elastic extension or strain and an applied tension or stress σ. This obeys Hooke's law $\sigma = E\varepsilon$. The effect of temperature on E is illustrated in Figure 272 (the values were determined for polycrystalline iron). In single crystals, the modulus of elasticity is strongly dependent on the crystallographic orientation of the sample (anisotropy). This anisotropy is also apparent in materials with a well-defined texture.

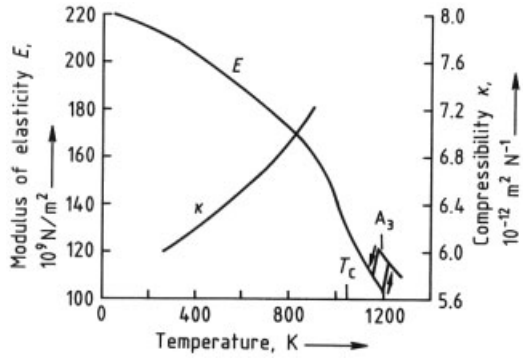

Figure 272. Moduli of elasticity and compressibility of iron [520]

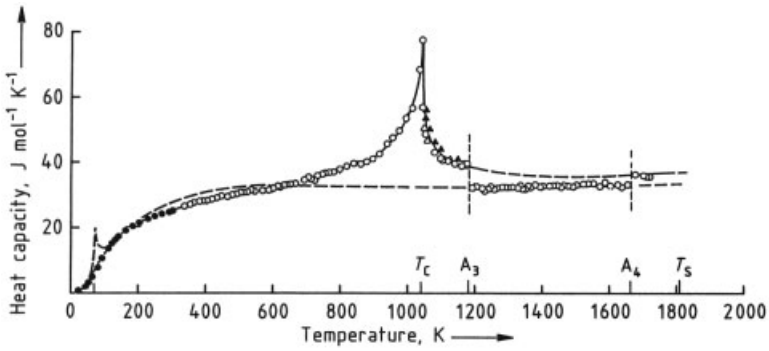

Figure 271. Effect of temperature on the molar heat capacity of iron
Solid line, most probable values; dashed line, values calculated for the instability ranges of α- and γ-iron
• According to [516]; △ According to [517]; △ According to [518]; ○ According to [519]

Volume changes under hydrostatic pressure are described by the compressibility κ, shown as a function of temperature in Figure 272.

Magnetic Properties. Individual iron atoms have a permanent magnetic moment. At temperatures above the Curie temperature, the directions of the magnetic moments are randomly distributed, and the material is paramagnetic. There is a linear relationship between field strength and polarization, represented by the volume susceptibility.

Below the Curie temperature, the magnetic moments of neighboring atoms in the bcc lattice are arranged parallel to each other, and the iron is ferromagnetic. Within the Weiss zones (magnetic domains) the moments are all in the same direction. This direction may differ from the direction within neighboring domains, so that they cancel each other.

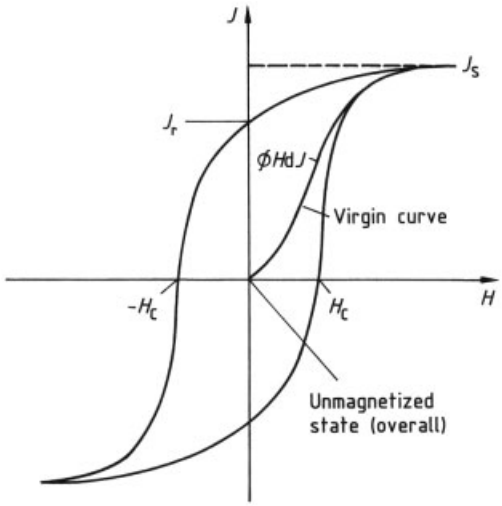

Figure 273. Magnetization curve of a ferromagnetic material
H = field strength; J = polarization ($B = J + \mu_0 H$ = induction); J_s = saturation polarization; J_r = remanence; H_c = coercive field strength; $\oint H dJ$, the area of the hysteresis loop, gives the amount of energy liberated as heat on reversal of the direction of magnetization. This is the hysteresis loss, quoted per unit mass

The properties of a ferromagnetic material can be represented by the magnetization curve (Fig. 273). With increasing field strength H, the polarization increases as the number of favorably oriented magnetic domains increases, and the magnetic moments align in the direction of the applied field. The maximum value is the saturation polarization. If the applied field is removed, the residual polarization at $H = 0$ is the remanence. The coercive field strength is the applied field which reduces the polarization to zero.

The saturation polarization varies with temperature, and disappears at the Curie temperature [521], [522]. The magnetization curve for single crystals varies with crystal orientation, being steepest in the [100] direction, and considerably less steep in the [111] direction. This effect is utilized in the manufacture of textured electroplated steel.

Electrical Properties. The high electrical conductivity of iron is due to the great mobility of the free electrons, but their movement is scattered by deviations from the periodic lattice potential or by the electrons of unfilled shells. All crystal defects, e.g., holes, dislocations, stacking faults, impurity atoms, and precipitates, lead to energy losses by the motion of free electrons, and hence to an electrical resistance that falls to the so-called residual resistance at absolute zero. According to the Matthiessen rule, resistance is made up of a temperature-dependent part, due to lattice vibrations, and a residual resistance which is independent of temperature. The effects of the magnetic and the $\alpha - \gamma$ transitions can also be seen in Figure 274.

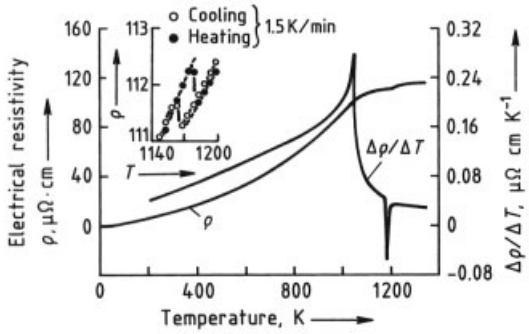

Figure 274. Effect of temperature on the electrical resistivity of pure iron

At high temperatures, thermal conductivity is a result of the same processes that lead to electrical conductivity. The thermal conductivity – temperature curve is similarly affected by the magnetic transformation as shown in Figure 275 [515], [523]. According to the Wiedemann – Franz – Lorenz law, the ratio of

thermal conductivity to electrical conductivity is directly proportional to temperature. From this relationship, thermal conductivity can be calculated from electrical conductivity even at high temperatures when direct determination is difficult.

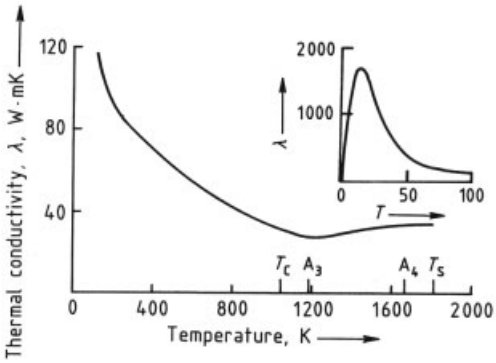

Figure 275. Effect of temperature on the thermal conductivity of pure iron

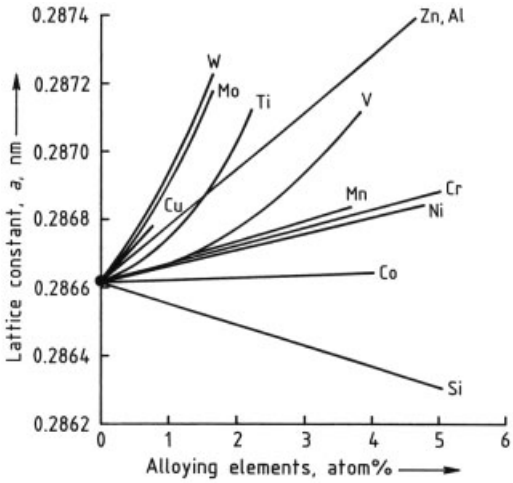

Figure 276. Effect of alloying elements on the lattice constant of α-iron

6.1.2.3.2. α-Iron Solid Solutions

As described in Section 6.2.1, alloying elements can be divided into those that restrict the γ-region, and those that stabilize the γ-phase and extend its limits [524]. The effect of alloying additives on the transformation behavior of steel can be understood only from knowledge of the effect on various properties. Almost all alloying elements cause a volume increase. Figure 276 shows their effect on the lattice constant of α-iron. Only silicon leads to a contraction. Electrical resistance increases in proportion to the concentration of the alloying element (Fig. 277). The most significant effect of alloying elements on magnetic properties is on the Curie temperature, as this affects the temperature of the transformation from γ- to α-iron. Figure 278 shows the effects of various elements. The large increase caused by cobalt is of special interest.

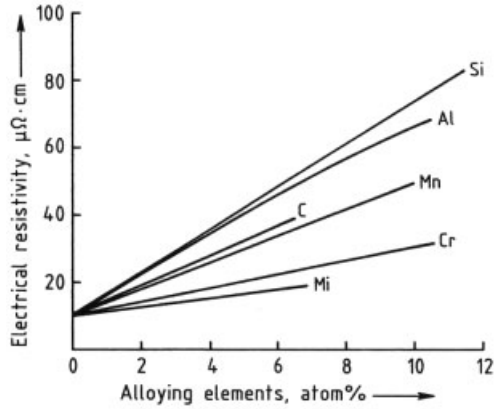

Figure 277. Increase in the electrical resistivity of α-iron by alloying elements

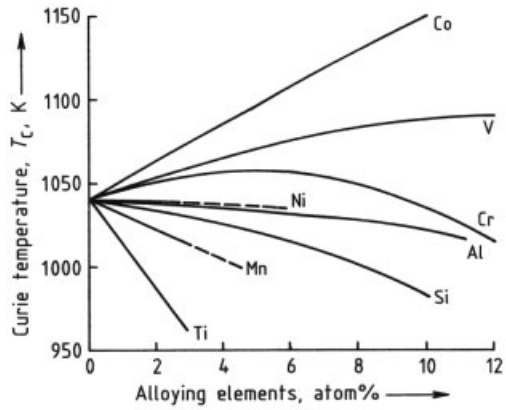

Figure 278. Effect of alloying elements on the Curie temperature of α-iron

The temperature dependence of the physical properties of α-iron mixed crystals is closely linked to their magnetic behavior, and is not significantly different from that of pure iron. For example, the heat capacity of iron – chromium

alloys is essentially constant if the temperature scale is based on the Curie temperature.

Ferritic iron – silicon alloys are of special importance, as these have low magnetization losses [521], [525]. Silicon addition increases electrical resistance, and hence reduces eddy current losses. The crystal anisotropy energies and the consequent hysteresis losses are reduced. Addition of $> 4.5\,\%$ Si prevents the $\alpha - \gamma$ transformation, and enables a coarse grain structure to be produced. Texturing allows the magnetic anisotropy to be exploited.

6.1.2.3.3. γ-Iron Solid Solutions

The magnetic properties of austenitic iron alloys are complex – as ferromagnetism, antiferromagnetism, and paramagnetism can occur. Iron – manganese [526], iron – nickel [527], and iron – nickel – manganese alloys [528] are of industrial importance.

The thermal expansion coefficients of paramagnetic γ-alloys (e.g., iron – manganese – nickel and iron – nickel – chromium alloys) are relatively high at room temperature. This is utilized in bimetallic systems [529]. Conversely, a large volume magnetostriction in magnetically ordered alloys compensates for thermally induced changes to the lattice dimensions; only a small amount of thermal expansion takes place over a certain temperature range [530]. Iron alloys containing 35 % nickel are examples of Invar alloys of this type.

6.1.2.3.4. Other Effects of Structure

In general, physical properties are influenced by lattice defects. Density is somewhat decreased by lattice defects. All lattice defects increase the residual electrical resistance, whereas lattice vibrations are the main cause of resistance at high temperatures. Spontaneous magnetization and the Curie temperature depend mainly on lattice structure and composition. However, the magnetization curve is strongly influenced by the structure; lattice defects tend to prevent movement of the domain walls, and hence increase the remanence and coercive field strength.

In multiphase structures, the different properties of the phases must also be considered. These properties combine in proportion to the volume ratios in only the simplest cases, e.g., density, heat capacity, heat of transformation, and heat of formation, provided that the phases are in a coarse state of subdivision. A systematic treatment is not yet possible. Probably the most important industrial application is represented by the "hard" or permanent magnetic materials, with high coercive field strengths and a high *BH* value.

6.1.3. Heat Treatment

6.1.3.1. Time – Temperature Transformation Diagrams

Under technical conditions, desired microstructures can be produced by isothermal transformation or continuous cooling. In practice, temperature – time treatments intermediate between these extremes are used, and the steel is sometimes also mechanically formed during cooling. The progress of such transformations is shown in time – temperature transformation (TTT) diagrams (see Figs. 191 and 192). By special measuring techniques, the effect of deformation on the transformation behavior can be measured and represented in TTT diagrams. A general representation of transformation behavior is possible only by computer simulation of the transformations of the microstructure [256], [531], [532]. The following treatment is mainly restricted to the continuous cooling process (Fig. 192).

The addition of alloying elements enables the M_S temperature and the time required for a transformation in the pearlite and bainite ranges to be shifted within wide limits [277], [278], [280]. Figures 279 and 280 show continuous cooling transformation (CCT) diagrams for a hypoeutectoid alloy steel and a tool steel [299]. With Ck 45 steel (see Fig. 192), 45 vol % ferrite and 55 vol % pearlite are formed during a cooling period of 300 s to 500 °C. When this cooling process is applied to the tool steel shown in Figure 280, it produces 100 % martensite, apart from the proeutectoid carbide precipitation (K). Therefore, another method of influencing the microstructure of a workpiece, apart from the temperature – time treatment, is by choice of its chemical composition. However, steels with the

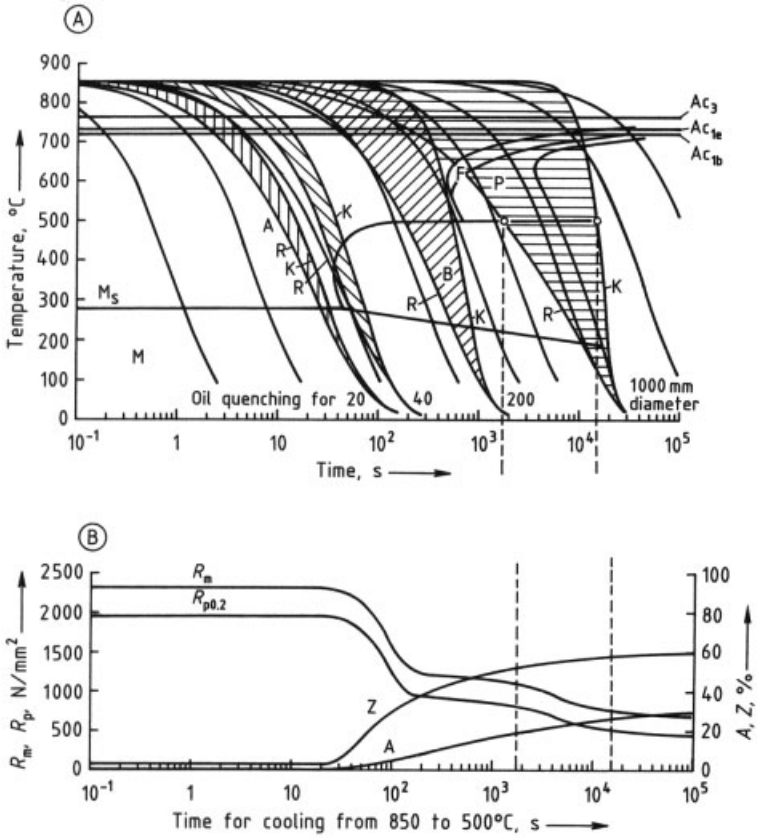

Figure 279. CCT diagram for continuous cooling of 50 CrMo 4 steel [299]
A) Temperature – time curve in the surface zone (R) and the core (K) of cylinders of various diameters for oil cooling. The surface zone extends below the surface by 10 % of the diameter; B) 0.2 % offset yield stress $R_{p0.2}$, tensile strength R_m, elongation at fracture A, and reduction in area at fracture Z for test pieces with the given durations of cooling to 500 °C

lowest possible alloy content are used in any given application, for reasons of cost [533].

6.1.3.2. Heat Treatment of Transformable Steels

6.1.3.2.1. Austenitizing

With transformable steels, the first step in many heat treatments is austenitizing. In an industrial situation, the austenitizing temperature selected for hypoeutectoid steels is 30 – 50 °C above Ac_3 (see Fig. 188), and, for hypereutectoid steels, 30 – 50 °C above Ac_1 [534–536]. These temperatures are given in the relevant standards [537], [538]. The annealing times required for the formation of a homogeneous austenite can be found from the time-temperature austenitization (TTA) diagrams [274], [275], e.g., Figure 190. The tool steels are an exception to this, and the temperatures needed to dissolve the special carbides are given in the relevant standards. These temperatures lie in the two-phase austenite + carbide region ($\gamma + M_3C$ in Fig. 188, A + K in Fig. 280). As well as M_3C, other carbides occur in these steels, e.g., M_7C_3, $M_{23}C_6$, M_2C, or MC. If these carbides remain undissolved during austenitizing, they improve wear resistance [289], [290].

The industrial austenitizing of heat treated engineering steels leads usually to the formation of a homogeneous austenite with small grain size (see Fig. 190). If low austenitizing temperatures and short times are used, the austenitization is dependent on the initial state. Thus, although a

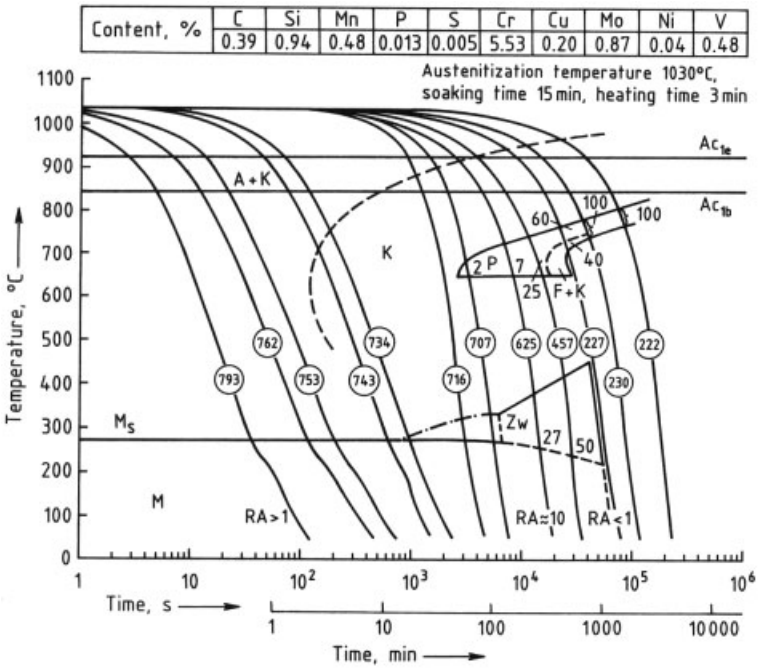

Figure 280. CCT diagram for X 38 CrMoV 51 tool steel [277].
Austenitization at 1030 °C for 15 min; RA residual austenite
In circles, the hardness at room temperature is given for each cooling curve.

microstructure produced by quenching and tempering leads to a rapid and uniform distribution of the carbon, if the starting condition is a ferritic – pearlitic microstructure corresponding to Figure 194, the carbides in the pearlitic region must first dissolve, and this carbon, now in the austenite, diffuses into the zones that consisted of low- carbon ferrite in the initial state. This requires extra annealing times, or an increase in temperature for the same annealing time, in order to obtain a homogeneous austenite. When using TTA diagrams to plan an isothermal temperature – time process, the austenitization time should start after the workpiece has been heated throughout.

6.1.3.2.2. Transformations

As industrial cooling operations are usually carried out continuously in air or liquid media, the temperature – time curve at the center and near the surface of the workpiece is different, and depends very much on the dimensions (Fig. 279) [539], [540]. The following discussion applies only to dimensions for which the stated cooling regimes are feasible (see Section 6.1.3.2.3).

Transformation in the Pearlite Range. A transformation in the lower region of the pearlite range is used in wire patenting [541]. This forms pearlite with very small interlamellar distances which has very good wire-drawing properties. As the diameter is small, the transformation can be carried out isothermally. Other applications include the production of rails, in which pearlite is formed at low temperature using the heat of the rolling [542]. In plain carbon steels, a ferritic – pearlitic microstructure can be produced by air cooling (Fig. 192). A corresponding heat treatment is known as normalization [534–536], [538], [541]. Delayed cooling is necessary with alloy steels, especially in thin sections, if a ferritic – pearlitic microstructure is desired (Fig. 279 and 280).

Transformations in the Bainite Range. In bainitization, a microstructure consisting only of bainite is produced in an isothermal transformation (Fig. 191). Whereas continuous cool-

ing cannot in principle give a 100 % bainite microstructure (see Fig. 279), microstructures formed in the upper temperature region of the bainite range have a much coarser microstructure than those formed in the lower region of the pearlite range (see Fig. 200). These coarsely bainitic microstructures usually have poor toughness properties for their yield strengths. On the other hand, microstructures formed in the lower region of the bainite range (see Fig. 199) have very favorable combinations of toughness and yield strength, often matching the properties of annealed martensite [288], [293], [541].

Transformations in the Martensite Range. Martensite can be obtained by cooling so rapidly that the formation of microstructures of the pearlite and bainite range is suppressed. For 50 CrMo 4 steel this requires a cooling time to 500 °C of ≤ 20 s (Fig. 279). Plain carbon steels must usually be water-quenched to achieve this. As martensite is the hardest of all the steel microstructures, the process of martensite formation is known as hardening [534], [535], [541]. The effect of the cross section of the workpiece is discussed in Section 6.1.3.2.3. Owing to the high rates of cooling, large temperature differences between the outer and inner regions of the workpiece occur (see Fig. 279), so that stresses and residual stresses are produced that can lead to cracking. These hardening cracks impose a limit to heat treatment processes involving rapid cooling of large or complex workpieces [543], [544].

With alloy steels, e.g., tool steels which need long time for transformation (Fig. 280), martensite can be produced by air cooling in components with small cross sections made of transformable steel, so that the limitations due to interior stresses are mostly avoided [289], [541]. Hardened steels are annealed at 100 – 150 °C, as the yield strength is hardly altered up to these temperatures, and the toughness can be considerably improved [289].

6.1.3.2.3. Effect of Dimensions

Figure 279 shows the temperature – time curves for the outer and inner zones of cylinders of various diameters. The temperature differences between these zones can be very great, especially with cooling media that cool rapidly, and this leads to the formation of different microstructures in the outer and inner zones [539], [545]. This must be borne in mind when deciding on the heat treatment process needed to give particular properties. An increase in the cooling period $t_{8/5}$ decreases the yield strength $R_{p0.2}$ and increases the elongation at fracture A and reduction in area at fracture Z (Fig. 279) [539]. Figure 281 shows how the hardness varies over the cross section of cylinders of different steels; for an oil-quenched plain carbon steel Ck 45, maximum hardness (see Fig. 197) is not reached, even near the surface. Maximum hardness can be obtained over only ca. 40 % of the section with 50 CrMo 4 steel; a purely martensitic microstructure is produced near the outer surface.

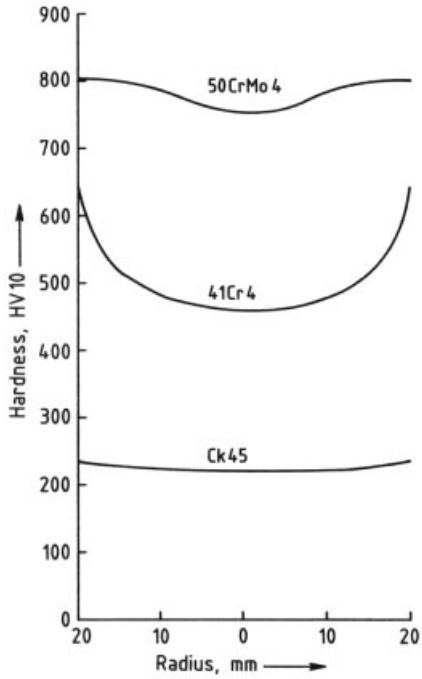

Figure 281. Hardness variations over the cross section of cylinders of diameter 40 mm fabricated from various steels after quenching in oil [299]

6.1.3.2.4. Quenching and Tempering

Quenching and tempering consists of hardening followed by annealing. As stated in Section 6.1.1.4, the tempering of martensite can give

any yield strength between that of martensite and that of a soft-annealed microstructure. The relationship between tempering temperature and yield strength is known for many grades of steel (Fig. 282) [536], [538]. Because the arrangement of carbides in the ferrite is very uniform and homogeneous after martensite is tempered, this process gives very good toughness properties for a given yield strength, considerably better than ferritic – pearlitic microstructures or microstructures produced in the upper temperature region of the bainite range [288], [293].

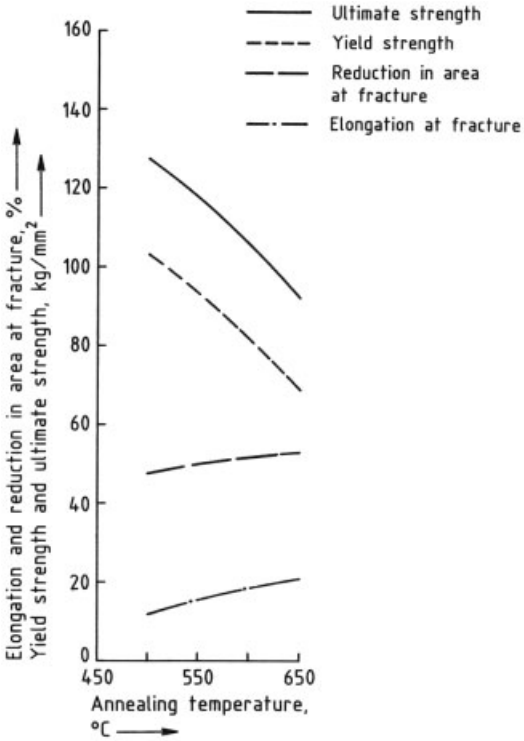

Figure 282. Variations in mechanical properties as a function of the tempering temperature of 50 CrMo 4 steel [538]

6.1.3.2.5. Spheroidized Annealing

Figures 279 and 280 show that, in high alloy steel components of small dimensions, microstructures of appreciable hardness can be produced, even with air cooling, and these can give machining problems. The carbides are often spheroidized by means of a prolonged annealing process at temperatures just below Ac_1 [546], i.e., coarse carbides are formed in the ferrite (Fig. 283). This process, spheroidize annealing, can be accelerated by an oscillation around A_1, or by a prior deformation process [536], [541], [547].

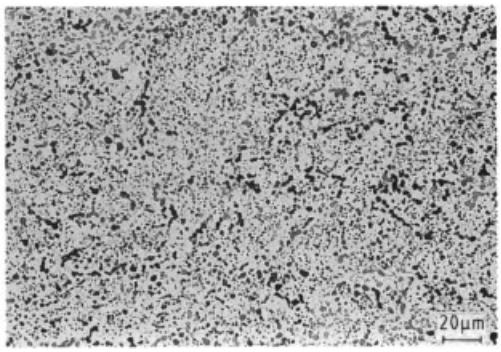

Figure 283. Microstructure after spheroidize annealing of C 130 steel
Carbides (gray – black) in ferrite (white); fine lines are ferrite grain boundaries; 810 °C for 15 min (water), and 720 °C for 8 h (air); ×500, etched with sodium picrate + nital

6.1.3.2.6. Thermomechanical Treatment

In Section 6.1.2, it was shown that high yield strengths and good toughness properties can be ensured only if the carbides are fine and homogeneously distributed. A fine-grained austenite can be produced at low austenitizing temperatures and with short annealing times, even though this leads to incomplete dissolution of the carbides (Fig. 190). As only part of the carbon in the austenite is dissolved, there is a reduction in the maximum hardness (Fig. 197). With engineering steels, which are not heat treated by the user, a very fine grain can be produced during manufacture by a hot deformation process in the austenite region, followed by recrystallization [548–552]. An example of the effect of the initial grain size and the deformation temperature is shown in Figure 284 [553]. If a fine-grained austenite of this type is cooled in a controlled manner, the very fine-grained transformation microstructures ferrite – pearlite, bainite, or even martensite are formed, depending on the cooling rate and chemical composition of the steel [554].

Another method of obtaining a very fine microstructure is to add suitable elements, e.g., Nb or Ti, to prevent recrystallization after deformation of the austenite at low temperature, and

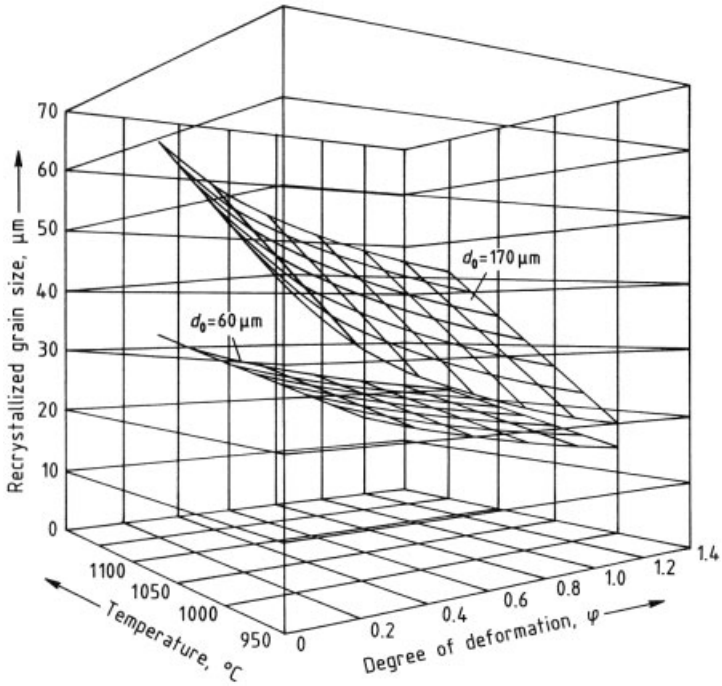

Figure 284. Static recrystallization of a steel with 0.038 % C, 1.8 % Mn, 0.28 % Mo, 0.07 % Nb, and 0.009 % N for two initial grain sizes
Deformed by one pass; $\varphi = \ln (d_0/d)$, d_0 = starting thickness, d = final thickness [553]

to allow the transformation from the uncrystallized austenite to take place [555]. This requires very precise control of the temperature – time curve, especially at the end of a rolling operation [553], [556]. The transformation of an unrecrystallized austenite proceeds more quickly than that of a recrystallized austenite [557]. The two processes can be combined. At temperatures at which rapid recrystallization of austenite takes place, e.g., slightly above Ac_3, hot deformation produces a fine austenite grain. This fine austenite is subsequently deformed at lower temperatures (e.g., at or below Ac_3) such that recrystallization of the austenite is avoided. It is then immediately cooled and the transformation initiated. The temperature – time curve and the degree of deformation must be correct for the steels concerned, and the chemical compositions of the steels must be suitable for the available plant equipment. Transformations in the pearlite, bainite, or martensite range are carried out according to the desired application properties. Elements such as Nb and Ti which are added to delay recrystallization can give precipitation hardening during the cooling [558–561].

6.1.3.2.7. Surface Hardening

For technical applications, it is often desired that the surface of a workpiece should be hard in order to provide wear resistance. However, as the load-bearing forces are comparatively small, the total section can have a low yield strength. In this case, it is reasonable to heat treat the workpiece at the surface only, giving high yield strength, while leaving the interior ductile and tough. Because of the treatment necessary, residual compressive stresses are usually produced at the surface, and this also improves fatigue resistance. Because of the internal stresses, the properties of the workpiece can no longer be described by the properties of the material only.

There are two fundamental methods of hardening the surface zone [541]:

1) Hardening by heat treatment, using a temperature – time regime derived from the transformation characteristic
2) Modification of the chemical composition in the surface zone

As shown in Figures 279 and 281, oil quenching of a workpiece, diameter 40 mm, consisting of 50 CrMo 4 steel, produces only small differences in hardness between the surface and the interior. If only the surface zone of a workpiece is austenitized and quenched, the interior temperature does not exceed Ac_1. The interior is very highly tempered, and further cooling has no further effect on its strength. Only the austenitized surface zone is hardened, so that, on tempering, a hardened and tempered high-strength surface zone is produced, surrounding a comparatively ductile core. The depth of hardening can be influenced by the temperature – time regime selected. A process used for this type of surface hardening is induction hardening [541].

In case hardening, steel is placed in a carbon-containing medium at ca. 950 °C, and carbon diffuses into the surface zone [534–536], [562], [563]. The most common technique is gas carburization, which is very easy to control. The carbon which is taken up (see Fig. 197) increases the hardness after quenching, but transformation of the austenite is delayed. However, by using an appropriate cooling medium for the dimensions of the workpiece, the two effects can be used to give the maximum hardness and depth of hardening. Surface hardening in very thin layers can be produced by nitriding, chromizing, and boronizing [534], [535], [541], [564].

6.1.3.3. Heat Treatment Without Transformation

With nontransformable austenitic or ferritic steels, the only aims of the heat treatments are to cause recrystallization to give a desired grain size, to quench a steel that has been heated to dissolve precipitated phases, and sometimes to temper after precipitation hardening. Strictly speaking, the annealing stage of the tempering process (see Section 6.1.3.2.4) is a heat treatment without transformation. However, the most important heat treatment of this type is recrystallization.

6.1.3.3.1. Recrystallization

Mechanical deformation of steel causes gliding which leads to a large increase in dislocation density, and hence to a high level of stored energy [565], [566], (see Section 6.1.2). Dislocations of opposite sign attract one another. If the temperature is high enough to permit easy movement of the dislocations, defects of opposite sign migrate and annihilate each other. The dislocation density and stored energy decrease. This process is known as recovery. Recovery proceeds more quickly at higher temperatures and for higher stored energies. In steels at 400 °C it takes place in hours, and >700 °C in seconds, although these times can vary by a factor of ten, depending on the chemical composition of the steel. If the recovery process, e.g., at 600 °C, rapidly forms small local regions with very low dislocation densities, these regions, which have high-angle grain boundaries, grow into the neighboring regions (recrystallization). The lower the degree of mechanical deformation, the higher is the temperature that leads to a measurable recrystallization rate, and the larger are the grains produced by the recrystallization (Fig. 284) [567].

For steels above 900 °C, in the austenitic region, recrystallization is often so rapid that it can take place during a deformation process. This is referred to as dynamic recrystallization, in contrast to static recrystallization (which only occurs after deformation ceases). At the end of the primary recrystallization described above, the high dislocation densities produced by the deformation in all the grains are just removed. In secondary recrystallization, grain growth proceeds further (Fig. 190), as the single crystal is the thermodynamically stable state, and large grains grow at the expense of small grains. The addition of alloying elements, e.g., Nb and Ti, can cause recrystallization in the austenite region to be slowed down so much that, after the deformation caused by rolling, the steel can be transformed without recrystallization. This is utilized in thermomechanical treatment (see Section 6.1.3.2.6). Figure 284 shows the fundamental influence of the starting size of the grains, deformation temperature, and degree of deformation, provided that the effect of Nb carbonitride precipitation, which has not been extensively researched, is not superimposed.

In the deformation of steels in the α-iron region, preferred orientations are produced. This is caused by the gliding processes that occur [565]. In accordance with the recrystallization sequence outlined above, recrystallization starts in grains with the highest energy density, and a certain orientation relative to the frame of reference during deformation, e.g., to a steel sheet. On recrystallization, grains of one particular orientation begin to grow sooner than others, which leads to a higher concentration of grains of this orientation at the end of the primary crystallization than would be expected for a random distribution. The textures so produced are used in the production of deep-drawing sheets to give favorable deformation properties of the steel. With electric sheets, these textures enable the magnetic anisotropy of the body- centered cubic lattice to impart improved electrical properties [270], [568].

6.1.3.4. Simulation of Heat Treatment

For transformable steels, there are limitless possibilities for transforming austenite in the pearlite, bainite, or martensite range, using controlled temperature – time regimes. Finding an optimum temperature – time regime to produce a desired microstructure, and hence the specified properties, requires a very large number of experiments. Models representing transformation behavior have been developed which can be used to calculate the progress of transformations [531], [569], [570]. Similarly, knowing the temperature and degree of deformation, grain sizes of the austenite after heat treatment of particular grades of steel can be calculated, so that the transformations that take place during fabrication can be simulated [571], [572]. If the mechanical properties of the individual microstructures of a steel are known, the final properties produced by a given process can also be calculated [573]. In this way, a proposed process can be simulated in advance. Also, deviations at the start of the process sequence can be rectified by changes at its end.

6.2. Uses

6.2.1. Classification of Steel and Steel Products

6.2.1.1. General

The existence of so many grades of steel and steel products makes it essential to have an agreed terminology to facilitate communication between producers, users, and traders in these materials. This is necessary not only within the European Common Market, but also for world trade. In the early 1990s, ca. 30 % of steel products were exported from the countries producing them (Table 22).

Table 22. World steel exports by product 1992 ∗

Product	Exports, $\times 10^6$ t
Ingots and semifinished material	12.9
Railway track material	0.9
Concrete reinforcing bars	5.5
Hot-rolled bars and rods	4.3
Angles, shapes and sections	10.2
Wire rod	8.0
Hot-rolled strip	1.9
Plates	9.4
Hot-rolled sheets and coils	22.4
Cold-rollled sheets and coils	17.1
Galvanized sheet	9.6
Tinplate and blackplate	4.6
Other coated sheet	3.9
Wheels (forged and rolled) and axles	0.1
Steel tubes and fittings	14.0
Drawn wire	2.1
Cold-rolled strip	1.4
Cold-finished bars and rods	1.3
Castings	0.1
Forgings	0.6
Total	130.4

∗ International Iron and Steel Institute: World Steel in Figures, 1994.

The International Standards Organization (ISO) and the European Committee for Iron and Steel Standards (ECISS) have issued standards over many years, including definitions of steel products (EN 10 079) and a system for classifying grades of steel (EN 10 020).

Designation systems for steels have been agreed in Europe (EN 10 027-1, -2), but are not yet thoroughly established.

6.2.1.2. Definitions for the Classification of Grades of Steel

Steel is a general term for materials in which the iron content is higher than that of any other element, and in which the carbon content is normally $< 2\,\%$.

According to EN 10 020, steel is divided into unalloyed and alloyed grades, according to its chemical composition, and these grades are further subdivided into principal quality groups, based on their most important properties and uses. The starting point is the minimum level of each element in the melt, specified in the standard or in the terms of delivery. The harmonized system for customs nomenclature is based on other physical quantities.

Unalloyed steels are those in which the level of an individual element in no case reaches or exceeds the limiting concentration given in Table 23.

Table 23. Concentration limits of elements for classification as base or alloy steel

	Element	Limiting concentration, %
Al	Aluminum	0.10
B	Boron	0.0008
Bi	Bismuth	0.10
Co	Cobalt	0.10
Cr	Chromium *	0.30
Cu	Copper *	0.40
La	Lanthanides (assessed separately)	0.05
Mn	Manganese	1.65 ***
Mo	Molybdenum *	0.08
Nb	Niobium**	0.06
Ni	Nickel *	0.30
Pb	Lead	0.40
Se	Selenium	0.10
Si	Silicon	0.50
Te	Tellurium	0.10
Ti	Titanium **	0.05
V	Vanadium **	0.10
W	Tungsten	0.10
Zr	Zirconium **	0.05
Others (except C, N, O, S), each		0.05

* If two, three, or four of these elements are present in concentrations less than the maxima permitted, their total concentration must not exceed 70 % of the sum of the maxima. ** The same rule applies to these elements. *** If the manganese content is quoted as a minimum, this value applies.

In an alloy steel, the content of an individual element reaches or exceeds the limiting values quoted in Table 23 in at least one case.

From general economic considerations, special regulations have been set for some product groups, e.g., weldable fine-grained structural steels, free-cutting steels, electric sheets, and rails. Only the fundamental principles of this system are described here.

Other terms are often used, e.g., microalloyed steels, which contain small additions of, e.g., Nb, V, or Ti, usually in association with a thermomechanical treatment of the steel product. Other common expressions are lowalloy steel, which contains $< 5\,\%$ alloying elements, and HSLA (high-strength, low-alloy) steel.

EN 10 020 further divides steels into the following main quality groups:

Base steels are those whose quality specifications do not require special processing during manufacture. They are generally unsuitable for heat treatment.

High-grade unalloyed steels are those for which no consistent response to heat treatment and no purity requirements with respect to nonmetallic inclusions are specified. However, because of the stresses that arise in use, there are stricter or more numerous requirements compared with base steels, e.g., with respect to their fracture toughness, their grain size control, or their formability, so the manufacture of these steels requires special care.

High-grade alloy steels are for similar applications to those of high-grade *unalloyed* steels, but contain levels of alloying elements (to give special properties) that place them among the alloy steels (Table 23). They are in general unsuitable for quenching and tempering or surface hardening.

Special carbon steels are of higher purity than high-grade steels, and contain fewer nonmetallic inclusions. They are mostly intended for quenching and tempering or surface hardening, and show a consistent response to these treatments. A very wide range of working and application properties, e.g., high-strength or hardening properties within closely controlled limits, together with high specifications for toughness, forming and welding properties, are achieved by precise control of chemical composition, and by special care in manufacture and process control procedures.

Alloy special steels have an extremely wide range of working and application properties, achieved by precise control of chemical composition, and by special methods of manufacture and process control. This group includes steels

for structural engineering, pressure vessels, as well as mechanical engineering steels, stainless and heat-resistant steels, creep-resistant steels, steels for bearings and machine tools, steels with special physical properties, etc.

Steels can be classified according to their main characteristics and applications:

Flat products for cold forming
Structural steels
Steels for pressure vessels
Concrete reinforcing and prestressed concrete steels
Steels for rails
Free- cutting steels
Steels for wire drawing
Cold-heading steels
Engineering steels
Spring steels
Tool steels
Steels for packaging (tinplate, blackplate)
Welding filler materials
Stainless steels
Bearing steels
Steels with special physical properties

6.2.1.3. Definitions of Steel Products

According to DIN EN 10 079, steel products are defined according to their shape and dimensions, and/or their appearance and surface condition. There is no overall system; several classification systems exist in Europe. Definitions of the most important products are given below, and a more detailed subdivision is given in Table 24.

Liquid crude steel is steel as produced from its raw materials and ready for pouring. A distinction is made between liquid crude steel for ingot casting or continuous casting and for steel casting.

Solid crude steel is the product obtained by the solidification of liquid steel in a mold such that the product is suitable for subsequent forming.

Semifinished products are produced by strand casting (sometimes followed by rolling, forging, or cutting), by pressure casting, or by the rolling, forging, or cutting of blooms or slabs, and are generally suitable for further forming into flat or long products.

Flat products have an approximately rectangular cross section with width much greater than thickness. The surface is usually technically smooth, but can have regularly spaced indented or raised surface patterns, e.g., floor plates.

Long products have a constant cross section along their length, which is much greater than the width and height. The cross section usually complies with a standard specificaton. The surfaces of long products are technically smooth, but these too can have regularly spaced indented or raised surface patterns, e.g., for concrete reinforcing bars.

6.2.1.4. Designation Systems for Steel

Clear designation systems for steel grades are provided by the European Standards EN 10 027 (systems for naming steels) Part 1: codes (principal symbol) and ECISS Information Circular IC 10 (additional symbols for steel); and Part 2: numbering system.

The codes are formed from a principal symbol (Table 25) and an additional symbol, supplemented by numbers.

The steel numbers are formed according to the system given in Table 26. These correspond to the numerical classification used in Germany.

6.2.2. Types of Steel

6.2.2.1. Steels where the Mechanical Properties are Essential

Of the many properties of steel, its mechanical properties are the most relevant to its use as a material for the manufacture of commercial goods [574]. In many applications, other properties are also important, e.g., cold workability, suitability for heat treatment, resistance to corrosion and high temperatures, and machining properties. The mechanical properties are mostly asked for in the condition of delivery, but in many cases, they are obtained after further heat treatment or cold working by the customer.

Table 24. Classification of steel products

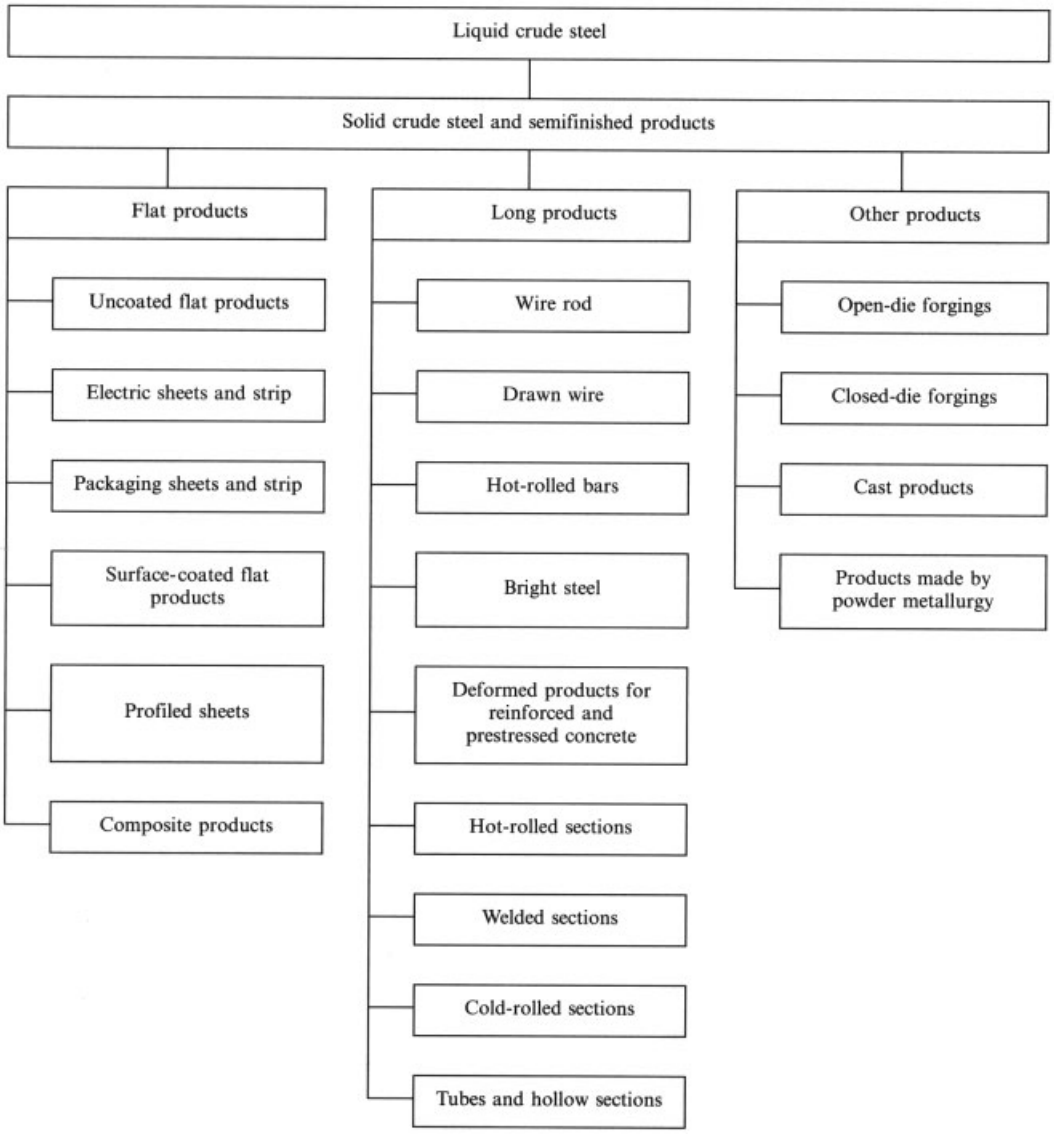

6.2.2.1.1. Normal- and High-Strength Structural and Reinforcing Steels

General. Steel grades used in structural steelwork, civil engineering, hydraulic engineering, bridge building, shipbuilding, offshore technology, pressure vessels, and many other applications in which climatic temperatures are important, are assessed by their strength and toughness at room temperature, and by their welding properties [575]. Normal-strength structural steels have minimum yield strength $< 355\,\mathrm{N/mm^2}$ in the as-rolled condition. Steels in this strength range are also supplied in the normalized condition. The usual normalizing process can be replaced by a temperature-controlled final roll pass, just above the normalizing temperature (thermomechanical treatment) which gives equivalent properties (delivery condition N). High-strength structural steels

Table 25. Principal symbols used in naming steels (DIN EN 10 027-1)

Symbol	Steel group	Characteristics
S	General structural steels (GS indicates cast steel)	Minimum yield strength, N/mm^2 for the smallest thickness range
P	Steels for pressure vessels (GP indicates cast steel)	Minimum yield strength, N/mm^2 for the smallest thickness range
L	Steels for pipelines	Minimum yield strength, N/mm^2 for the smallest thickness range
E	Engineering steels	Minimum yield strength, N/mm^2 for the smallest thickness range
B	Steels for reinforcing concrete	Minimum yield strength, N/mm^2 for the smallest thickness range
Y	Steels for prestressing concrete	Nominal tensile strength in N/mm^2
R	Rail steels	Minimum tensile strength, N/mm^2
H	Cold-rolled flat products of high-strength steels for cold forming	Minimum yield strength, N/mm^2
D	Flat products for cold forming	C = cold-rolled D = hot-rolled for immediate cold forming X = rolled condition (hot- or cold-rolled) not stated
T	Packaging sheets and strip	H = specified average Rockwell hardness for single reduced product; minimum yield strength, N/mm^2 for double-reduced product
M	Electric sheets and strip	Maximum hysteresis loss, W/kg
C	Unalloyed steels with average Mn content $< 1\,\%$, precluding machining steels (GC indicates cast steel)	Average C content $\times 100$
	Unalloyed steels with average Mn content $> 1\,\%$, unalloyed free-cutting steels, where the average content of each alloying element < 5 wt % (G indicates cast steel)	Specified average C content $\times 100$, followed by characteristic alloying elements with average content multiplied by the factor F Cr, Co, Mn, Ni, Si, W $F = 4$ Al, Be, Cu, No, Nb, Pb, Ta, Ti, V, Zr $F = 10$ Ce, N, P, S $F = 100$ B $F = 1000$
X	Alloy steels, where average content of at least one alloying element $\geq 5\,\%$ (GX indicates cast steel)	Specified average C content, followed by characteristic alloying elements and their average content
HS	High-speed steels	Average content of alloying elements in the order: W, Mo, V, Co

have minimum yield strength ≥ 355 N/mm^2. They are usually supplied in the normalized or thermomechanically rolled (TM) condition. Steels with minimum yield strength 500 – 880 N/mm^2 are usually supplied in the quenched and tempered condition, as the carbon content and alloying element content must be limited for weldability.

These steel grades are supplied in sheet, plates, beams, and merchant-bar stock. Steel grades in the quenched and tempered condition are supplied only as plates.

Steel products for incorporation in concrete are known as reinforcing steels [576]. These are of two types: those used in stress-free ferroconcrete, and prestressing steels, which are subjected to controlled prestressing, causing them to exert a compressive stress on the concrete [577]. The deadweight of the building itself and the load that it carries give rise to additional high tensile stresses in the steel rod or wire. The yield strength (and the 0.01 % offset yield strength for prestressing steels) is therefore the most important property of reinforcing steels. These steels are supplied in the as-rolled condition, though special temperature- controlled rolling and cooling is sometimes used. Quenched and tempered or cold-drawn wires are also very important prestressing steel products.

Reinforcing steels are used for buildings which are subject to the building regulations of individual countries. They require certification of public building authorities and are subjected to official continuous quality control.

Table 26. Steel groups *

	Unalloyed steels							
	Base steels		High-grade steels		Special steels	High-grade steels		Special steels
No.								Tool steels
0	00 90 base steels				10 steels with special physical properties			20 Cr
1			01 91 general structural steels with $R_m < 500$ N/mm²		11 structural, pressure vessel and engineering steels with C < 0.50%			21 Cr–Si Cr–Mn Cr–Mn–Si
2			02 92 other structural steels not intended for heat treatment with $R_m < 500$ N/mm²		12 structural, pressure vessel and engineering steels with C ≥ 0.50%			22 Cr–V Cr–V–Si Cr–V–Mn Cr–V–Mn–Si
3			03 93 steels with average C < 0.12% or $R_m < 400$ N/mm²		13 structural pressure vessel and engineering steels with special requirements			23 Cr–Mo Cr–Mo–V Mo–V
4			04 94 steels with average C ≥ 0.12% < 0.25% or $R_m \geq 400$ < 500 N/mm²		14			24 W Cr–W
5			05 95 steels with average C ≥ 0.25% < 0.55% or $R_m \geq 500$ < 700 N/mm²		15 tool steel			25 W–V Cr–W–V
6			06 96 steels with average C ≥ 0.55% or $R_m \geq 700$ N/mm²		16 tool steel			26 W excluding groups 24, 25 and 27
7			07 97 steels with higher P or S content		17 tool steel			27 with Ni
8					18 tool steel	08 98 steels with special physical properties		28 other
9					19	09 99 steels for other applications		29

* The classification of steel groups is in accordance with the classification of steels in EN 10020. The following information is provided: (a) steel group number, in upper left hand side; (b) principal characteristic of the steel group;

Alloy steels					
Miscellaneous steels	Stainless and heat resisting steels	Structural, pressure vessel and engineering steels			
30	40 stainless steel with < 2.5% Ni without Mo, Nb and Ti	50 Mn–Si–Cu	60 Cr–Ni with ≥ 2,0 < 3,0% Cr	70 Cr Cr–B	80 Cr–Si–Mo Cr–Si–Mn–Mo Cr–Si–Mo–V Cr–Si–Mn–Mo–V
31	41 stainless steel with < 2.5% Ni and Mo but without Nb and Ti	51 Mn–Si Mn–Cr	61	71 Cr–Si Cr–Mn Cr–Mn–B Cr–Si–Mn	81 Cr–Si–V Cr–Mn–V Cr–Si–Mn–V
32 high speed steel with Co	42	52 Mn–Cu Mn–V Si–V Mn–Si–V	62 Ni–Si Ni–Mn Ni–Cu	72 Cr–Mo with < 0.35% Mo Cr–Mo–B	82 Cr–Mo–W Cr–Mo–W–V
33 high speed steel without Co	43 stainless steel with ≥ 2.5% Ni without Mo, Nb and Ti	53 Mn–Ti Si–Ti	63 Ni–Mo Ni–Mo–Mn Ni–Mo–Cu Ni–Mo–V Ni–Mn–V	73 Cr–Mo with ≥ 0.35% Mo	83
34	44 stainless steel with ≥ 2.5% Ni with Mo but without Nb and Ti	54 Mo Nb, Ti, V, W	64	74	84 Cr–Si–Ti Cr–Mn–Ti Cr–Si–Mn–Ti
35 bearing steels	45 stainless steels with special additions	55 B Mn–B < 1.65% Mn	65 Cr–Ni–Mo with < 0.4% Mo + < 0.2% Ni	75 Cr–V with < 2.0% Cr	85 nitriding steels
36 Materials with special magnetic properties without Co	46 chemical resistant and high temp. Ni alloys	56 Ni	66 Cr–Ni–Mo with < 0.4% Mo + ≥ 2.0 < 3.5% Ni	76 Cr–V with ≥ 2.0% Cr	86
37 materials with special magnetic properties and with Co	47 heat-resistant steels with < 2.5% Ni	57 Cr–Ni with < 1.0% Cr	67 Cr–Ni–Mo with < 0.4% Mo + ≥ 3.5 < 5.0% Ni, or ≥ 0.4% Mo	77 Cr–Mo–V	87 steels not for heat treatment by user
38 materials with special physical properties without Ni	48 heat-resistant steels with ≥ 2.5% Ni	58 Cr–Ni with ≥ 1.0 < 1.5% Cr	68 Cr–Ni–V Cr–Ni–W Cr–Ni–V–W	78	88 high-strength weldable steels not intended for heat treatment by user
39 materials with special physical properties and with Ni	49 materials with elevated temperature properties	59 Cr–Ni with ≥ 1.5 < 2.0% Cr	69 Cr–Ni except groups 57 to 68	79 Cr–Mn–Mo Cr–Mn–Mo–V	89 high-strength weldable steels not intended for heat treatment by user

(c) R_m = tensile strength for high-grade steels. The limiting values for the chemical composition and tensile strength (R_m) are for guidance only.

Properties and Testing Methods. As a basis for the calculation of steel structures, values of yield strength and tensile strength are determined by tensile tests. If alternating mechanical stresses are expected to be applied to a component, e.g., in bridges or ships, the fatigue strength is important. Under some circumstances, fatigue strength is proportional to tensile strength. Deformability of strcutural steels under axial and multiaxial stress is assessed by tensile tests which give the percentage elongation after fracture, and sometimes the reduction in area at fracture.

Toughness and resistance to brittle fracture are determined by notched-bar impact energy tests at various temperatures, usually down to ca. − 60 °C. The values decrease more or less with temperature, sometimes irregularly. The important parameter, by international agreement, is the so- called transition temperature T_u, at which the Charpy V-notch impact energy falls to 27 J.

Many other test methods for toughness and crack arrest properties of steel have been developed, and form the basis of fracture mechanics. Examples inlcude V-notched tensile or bending test pieces taken from the full thickness of the plate, the Robertson test, which gives the crack arrest temperature (CAT) [578], or the drop weigth Pellini test, which gives the nil ductility transition (NDT) [579], [580].

Tensile tests, and the value derived from them, are also used to characterize reinforcing steels. For prestressing steels, the long-term behavior under static stress is important. This is determined either by a creep test (elongation under constant stress) or by a relaxation test (stress reduction at constant gauge length, with starting stress ca. 60 – 80 % of the measured tensile strength).

The metal-working properties are also very important. They influence the methods and costs of manufacture of components, e.g., cold or hot forming, welding, and flame cutting. Machinability is not so important for structural steels. Suitability for zinc coating may also be important. Flat products can be cold worked without difficulty by bending, bevelling, dishing, and other processes, the force required being determined by the yield strength. Changes to mechanical properties caused by work hardening must be borne in mind, as well as toughness reduction or yield strength reduction on cold forming under certain conditions (Bauschinger effect, see Section 6.1.2.1.2). After extensive cold forming, the mechanical properties can, if necessary, be restored to the starting condition by appropriate heat treatment, e.g., stress-free annealing. In hot forming, it is important to control the temperature before, during, and after the forming process if subsequent heat treatment is to be avoided. Thermomechanically rolled products cannot be heat treated, except by stress-free annealing, but this should only be carried out up to the temperature limit specified by the manufacturer.

Good welding properties are very important for structural steels. It must be possible to prepare defect-free welded joints at low cost, and these must have comparable load- carrying capacity to that of base material. The cost is influenced by the welding method, welding conditions, the choice of welding electrode, whether preheating is necessary and to what temperature. The latter depends on the cross-sectional thicknes of the weld, the grade of steel, and its chemical composition [581].

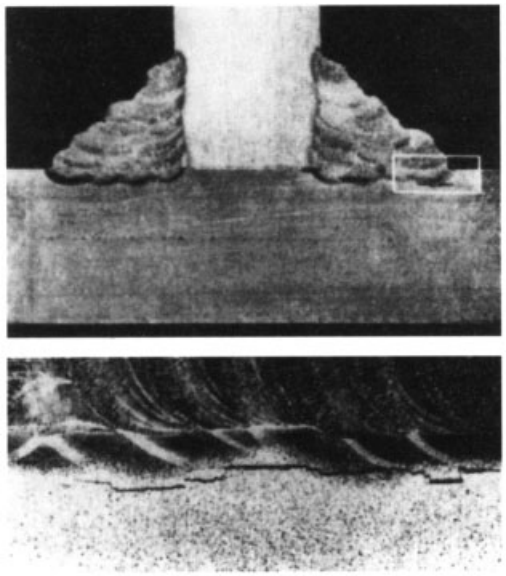

Figure 285. Lamellar tearing [583]

The chemical composition, especially the carbon content, the amount, shape, and distribution of sulfide inclusions, and the fineness of

grain determine the resistance to crack formation and propagation due to stresses arising during and after the welding process. These stresses, and the form and structure of the heat-affected zone in the base material near the weld seam, depend on welding data, e.g., heat input, working temperature, and cooling time from 800 to 500 °C ($t_{8/5}$) [582].

The overall aim is to limit the hardening in the region of the weld seam as much as possible, and to prevent grain growth in the heat-affected zone. Component design and the sequence of the various welds must ensure that the resulting stresses are as small as possible, and that there is no stress concentration (notch effect).

The test methods used to assess the suitability of a steel for welding are carried out on welded test pieces under standard conditions, and are the same as those described above for testing the base material for toughness and resistance to brittle fracture. To test the sensitivity of a steel to so-called lamellar tearing (Fig. 285) [583], when thick-walled items are to be welded, the mechanical properties are assessed by tensile tests in the thickness direction, and the reduction in area at fracture is determined (Z grades).

Grades of Steel. The most important steels in this group are the general-purpose structural steels. The starting point from which they were developed was rimming steel produced by the standard manufacturing processes of the early 1900s. In mainland Europe, basic converters were mainly used (basic Bessemer or Thomas steel), and in Britain and the United States the open hearth process (Siemens–Martin steel) was used almost exclusively. The differences in carbon content due to differences in steelmaking processes led to the production of two grades of steel with similar structures, both with a ferrite matrix, but differing in pearlite content and in the levels of other elements (P and N). The Thomas process produced a steel containing ca. 0.10 % C, the tensile strength of the rolled product being 365 N/mm^2 (St 37, now referred to as Fe 360 in EN 10 025); and the Siemens–Martin process produced a steel containing ca. 0.16 % C, with tensile strength 400 N/mm^2 like ASTM-A36 or grade A shipbuilding steel.

Rimming steels are hardly used today as structural steels, as the unavoidable segregation greatly interferes with the weldability, and they cannot be strand cast. The basic oxygen steelmaking process which predominates today results in a chemical composition which gives tensile strength ≥ 360 N/mm^2, and good weldability, no special metallurgical techniques being necessary.

In all grades of structural steel, yield strength depends on thickness. Fe 360 steel, which has minimum yield strength 235 N/mm^2 up to 16 mm thick, is still very important.

The effort to improve the efficiency of steel structures, i.e., to increase the useful load for a given mass of the load-carrying component, led in the early 1930s to the development of a weldable structural steel with yield strength ≥ 355 N/mm^2, and tensile strength ≥ 510 N/mm^2. This product, Fe 510 (EN 10 025), contains ≤ 0.2 % C, ≤ 1.6 % Mn, and ≤ 0.55 % Si, is a low-alloy steel, and is mostly produced with a fine grain, i.e., with Al added for fixation of nitrogen.

A steel intermediate between Fe 360 and Fe 510, with tensile strength ≥ 400 N/mm^2 and yield strength ≥ 235 N/mm^2 for all thicknesses, is still of great importance in shipbuilding. However, Fe 430 is little used as a general-purpose structural steel.

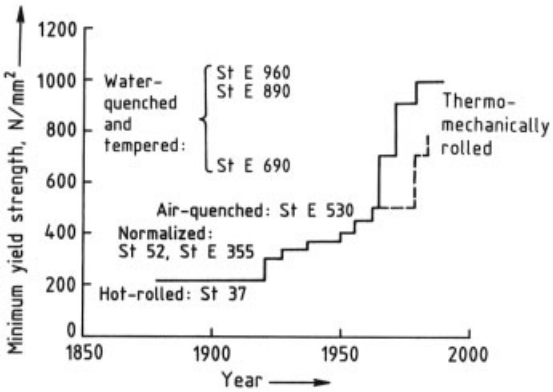

Figure 286. Development of the minimum yield strength of structural steels

Steels with improved toughness and welding properties are available in several quality groups (e.g., groups B to DD in EN 10 025), and are supplied in the as-rolled or normalized condition. These quality groups differ in the temperature

used for measuring the notched-bar impact energy, a higher minimum value for this quantity being specified for quality group DD.

In the 1960s, structural steels were further developed for use in very highly stressed components, e.g., bridges, lock gates, storage tanks, leading to the weldable fine-grained structural steels (EN 10 113), which are supplied in the normalized or thermomechanically rolled condition. These have minimum yield strength 275 – 460 N/mm^2. With reduced carbon content, these steel grades can be low-alloyed, and contain Al, with small additions of carbonitride-forming elements, e.g., niobium. This leads to improved strength properties via a very fine-grained structure and precipitation hardening.

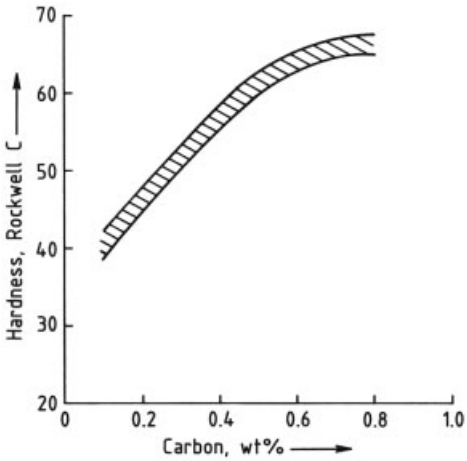

Figure 287. Maximum achievable hardness of martensite as a function of carbon content

Offshore technology, especially in the North Sea, requires steels with yield strength $\geq$ 355 N/mm^2 with further improvements in toughness and weldability, e.g., a sharp limitation on the hardness values in the heat-affected zone of a weld seam. Further reductions in carbon and sulfur content led to the costly offshore steels [584], which may be alloyed with Nb or other microalloying elements (e.g., Ti or V), sometimes with small additions of Cu and Ni. More stringent standards are specified in EN 10 225 for these steels.

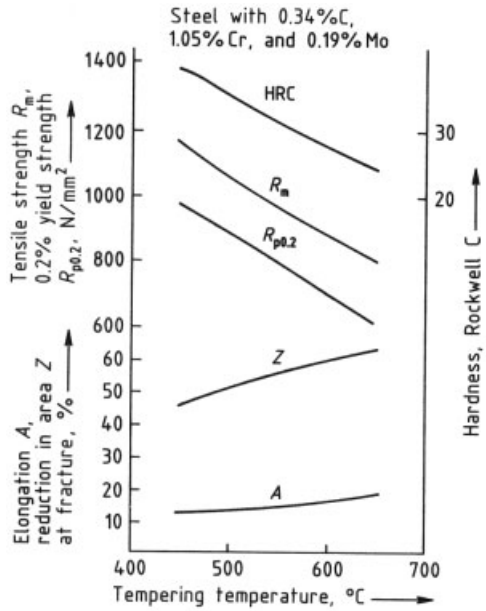

Figure 288. Tempering diagram of a heat-treatable steel after quenching from 830 °C in water
Diameter 60 mm

The development of quenched and tempered steels with good toughness and welding properties for use in structural work has resulted in steels with minimum yield strength up to 900 N/mm^2 (Fig. 286) [585]. These can be used very economically for applications where weight saving is very important, e.g., in mobile cranes, heavy commercial vehicles, and in supports for mine workings [586]. Carbon content has been reduced, and the steels are alloyed, so that, on quenching, the austenite is transformed into martensite or bainite over the whole cross section.

Within this group of water-quenched and tempered structural steels, a steel grade with yield strength $\geq$ 690 N/mm^2 is most important, as good toughness and welding properties have been achieved with only small additions of alloying elements. For mobile cranes with high carrying capacity, a steel with minimum yield strength 890 N/mm^2 is widely used. In offshore technology and shipbuilding, water-quenched structural steels with minimum yield strength 420 and 460 N/mm^2 are preferred; for general-purpose structural steelwork, the cheaper normalized or thermomechanically treated grades of steel are used.

The development of low-pearlite, thermomechanically treated high-strength structural steels, produced in the form of wide strip, up to 18 mm thick, is discussed later, as these are important for their cold forming properties as well as their mechanical properties.

Finally, the weatherproof steels must be mentioned. The presence of chromium and copper in these steels gives them improved resistance to atmospheric corrosion; they have similar structures and mechanical properties to those of Fe 360 and Fe 510 steels. On unprotected weathering, these steels form surface layers, whose structure, formation time, and protective action vary with different kinds of atmospheric action. For their successful and economic use, the design must be appropriate to the material, as specified in EN 10 155 [587].

Table 27 lists the most important structural steels, with their essential characteristics. Only a general overview of these can be given here.

The specifications for concrete-reinforcing steels have been greatly uprated compared with those for the earlier product, Monier iron. Continuing efforts to save weight have led to smaller sections, with considerably increased yield strengths. EN 10 080 specifies two products BSt 420 and BSt 500, with minimum yield strength 420 and 500 N/mm^2, respectively. These values can be achieved for different rod diameters with steels of approximately the same chemical composition by microalloying, cold working, or controlled cooling after hot rolling [588]. The stability of the structure depends on the total effect of the rods embedded in the concrete, not on the tensile properties of the individual rods. Tensile testing is therefore a statistical process; within a production batch, 5 % of the rods may fall below the minimum value in a normal distribution.

If tack or butt welding is used, the suitability of the material for these processes must be confirmed.

Steels used in prestressed concrete must have considerably better properties. Steel grades for $\leq$ 36 mm diameter rods should have 0.2 % proof strength 1080 N/mm^2, with tensile strength 1230 N/mm^2, and 0.01 % offset yield strength 950 N/mm^2. For wire, depending on the diameter, 0.2 % proof strength 1420 N/mm^2 in the quenched and tempered state and 1570 N/mm^2 in the cold-drawn state is specified. Braided or stranded cold-drawn wire is also used in prestressed concrete. Quality specifications for prestressing steels are given in EN 10 138.

6.2.2.1.2. Steels Suitable for Heat-Treatment and Surface Hardening [589]

General. The effect of structure on the properties of steel has been described in Section 6.1. The explanation for the wide range of properties that can be produced by heat treatment is that two phase transformations occur in the iron lattice between the melting point and room temperature. This enables a large number of different structures to be produced, depending on the chemical composition and the type of heat treatment (see Section 6.1.3).

In steels intended for quenching and tempering or similar heat treatment, strength and toughness properties are of overriding importance [590], as is the case with the structural steels described in Section 6.2.2.1.1. However, other properties, described as important in the

Table 27. Important normal- and high-strength structural steels

Grade	Standard specification	Delivery condition	Yield strength, N/mm^2	Tensile strength, N/mm^2	Elongation, %	Notched-bar impact energy, J *	Testing temperature, °C
S 235 J2 G3	EN 10 025	N	$\geq$ 235	340 – 470	24	27	– 20
S 355 J2 G4	EN 10 025	N	$\geq$ 355	490 – 639	20	27	– 20
S 355 NL	pr EN 10 113	N	$\geq$ 355	470 – 630	22	27	– 50
S 460 NL	pr EN 10 113	N	$\geq$ 460	550 – 720	17	27	– 50
S 690 QL	pr EN 10 137	V	$\geq$ 690	770 – 940	14	40	– 40
S 890 QL	SEW 090 **	V	$\geq$ 890	940 – 1100	11	40	– 40

* Longitudinal test piece.
** Stahl-Eisen-Werkstoffblatt.

earlier section, e.g., weldability, are not so vital. Hence, it is not necessary from the outset to consider limiting the carbon content. On the contrary, high carbon content is always considered in these steels, as their strengths need to be high, sometimes extremely high. In spite of certain transitions, the carbon content is the important difference from the weldable quenched and tempered steels described earlier.

As described in Section 6.1.3.2 this heat tretment consists of quench hardening in the first step. The material is heated until transformation to austenite is just complete and then it is quenched so rapidly that during cooling the austenite is not transformed to ferrite and perlite, but to martensite. Figure 287 shows the hardness values that can be achieved by complete transformation to martensite, as a function of carbon content.

In a second step, the hardened workpiece is tempered, i.e., reheated to a temperature appreciably below that at which the first γ solid solution reappears. This enables carbon to be released from forced solution in the martensite, and the dislocation density, which had been increased by the hardening, decreases again. This process can be very precisely controlled by means of the temperature and duration of the tempering stage, enabling the desired properties to be obtained (Fig. 288).

Quenched and tempered steels have a wide range of uses: as components of vehicles and machines, in heavy forgings, in springs and roller bearings, and in a wide range of tool steels. In general, the carbon content is 0.25 – 1 % or above. Depending on the application they are unalloyed, but mostly have alloy content $<$ ca. 5 % (low-alloy steels). Tool steels, which have a very wide range of applications and are frequently highly alloyed, do not come within the scope of this article [591].

Properties and their Measurement. For the steels discussed in this section, the distinction must be made between the application properties and the processing properties.

When steel is in use, strength, hardness, and toughness are important, but requirements vary widely according to the loading conditions in individual cases. The fatigue endurance limit is important in many applications, and a very high degree of purity with respect to nonmetallic inclusions can also be extremely important (springs, roller bearings).

Surface decarburization can be tolerated only within very narrow limits, so that, for high carbon content, great care in hot forming processes and subsequent heat treatment is necessary. With steels in which special carbides are present (Cr, Mo, V), the conditions must ensure uniform carbide distribution in the structure in all stages of steel manufacture.

The demand for these application properties means that the values determined in tensile tests are very important. The area reduction at fracture Z is of great interest when assessing these materials.

The notched-bar impact energy is usually determined at room temperature with the ISO V-notched specimen, which has replaced U-notched (e.g., DVM) test pieces. However, the DVM test piece is still used in some cases, and was predominantly used to obtain the results quoted in the literature.

The Rockwell test (HRC) is the most common method of hardness measurement, especially for high hardness values.

Metallographic investigation of purity, metal structure, and carbide distribution (if applicable) is essential with these steels.

The most important processing property is hardenability, the tendency of a steel to increase in hardness on quenching. The highest hardness value attainable on complete transformation to martensite (Fig. 287) depends on the carbon content; the extent of hardness penetration depth (partial conversion to martensite) depends on the overall chemical composition of the steel.

Of the numerous methods of testing hardenability, only Jominy's end quench test is now of practical importance [592]. The cylindrical test piece is heated to the hardening temperature (Fig. 289), and one end face is quenched by a standard water jet [593]. The extent of hardening is determined by observing a polished side surface. The hardness values at various distances from the quenched end face (Fig. 290) give a measure of the hardenability of the steel. These end quench hardness curves, which show a small amount of scatter, are the most important criteria for the assessment and classification of hardenable steels in all standard specifications. In conjunction with the TTT diagrams (see Section 6.1.3.1) for the various grades of steel, the

hardening curves facilitate the choice of the most suitable and economic steel grade for a given application and with a given section, and the conditions of the heat treatment. As the hardenability, so defined, is a function of the chemical composition and the steel making process, formulas have been developed for different alloy types which can be used to calculate end quench hardness curves [594].

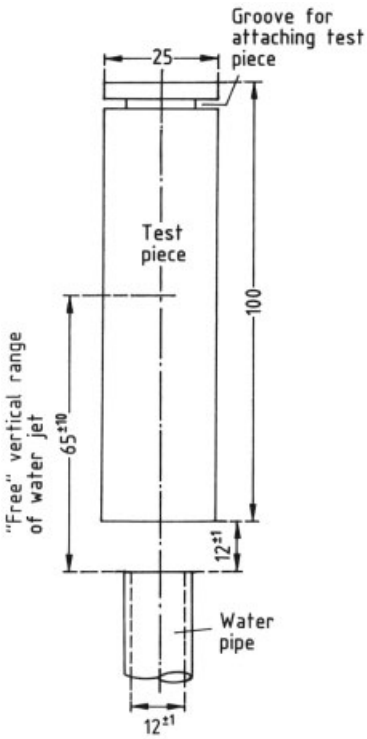

Figure 289. The end quench test

In addition to hardenability, other important processing properties include the capacity for cold shearing of semifinished products and merchant bars, machinability, and massive cold forming properties. If weldability is needed, e.g., for joining structural elements, a suitable welding process must be used, with low heat input or high energy density.

Capacity for cold shearing can be ensured only by spheroidize annealing appropriate grades of steel (> 0.4 % C), while machinability (see Section 6.2.2.4.3) can be improved by a small increase in sulfur content, and sometimes by a small addition of lead. Massive cold forming properties can sometimes be improved by adjusting the chemical composition (see Section 6.2.2.4.3).

Steel Grades. Whereas heat-treatable unalloyed steels, usually with carbon content varying stepwise from 0.2 % (2 C 22) to 0.6 % (2 C 60) are similar worldwide, alloyed heat-treatable steels have been developed in the industrialized countries to varying extents, depending on the availability and cost of the expensive alloying elements, e.g., nickel. The resulting range of heat-treatable steels specified in EN 10 083-1, ISO/R 683, and the SAE Handbook, whose areas of use sometimes overlap, cannot be described in detail here [595]. They include, in order of increasing capacity, the Cr, Cr – Mo, and Cr – Ni – Mo steel grades.

Heat-treatable steels which also contain the alloying elements vanadium or cobalt can have yield and tensile strengths $\gg$1200 N/mm^2, with good toughness properties. These are known as ultra-high-strength steels (UHS steels), and are used for components where the cross section must be as small as possible. e.g., in the aerospace industry [596]. This group also includes very highly alloyed maraging steels; these are not described further here.

For surface hardening, in which a completely martensitic structure is formed in only a controlled surface layer while the structure of the inner core is unchanged, unalloyed steels or steels alloyed with chromium only are preferred, as the hardenability curves of these materials are relatively steep.

Brief mention should be made of the development of precipitation-hardened ferritic – pearlitic steels (PFP steels), which are used for forged components, mainly in the automotive industry [597]. These microalloyed steels are subjected to a thermomechanical treatment by controlled cooling from the forging temperature, and can replace "classical" quenched and tempered steels, enabling the manufacturing costs of items such as steering stub axles, front axle components for commercial vehicles, and inlet valves for combustion engines to be much reduced by the avoidance of heating costs (Fig. 291), and by other processing simplifications [598].

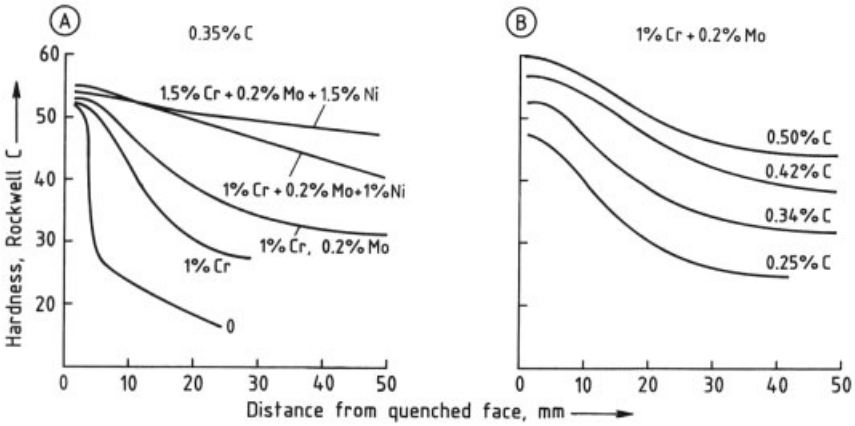

Figure 290. End quench hardness curves for heat-treatable steels of various chemical compositions
A) Carbon content approximately constant, variable alloy content; B) Alloy content approximately constant, variable carbon content

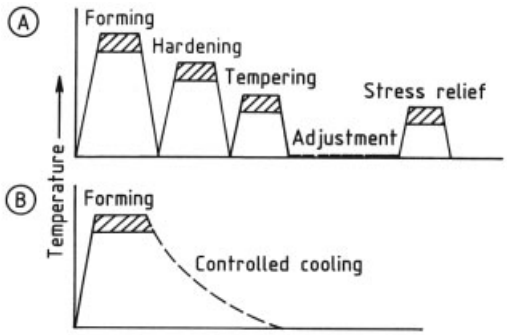

Figure 291. Processes in the production of crank shafts from various steels
A) Heat-treatable steels; B) Microalloyed pearlitic steels

These steels have carbon content ca. 0.25 – 0.45 %, with Mn content 1.2 – 1.6 %, and contain 0.1 % vanadium (or sometimes niobium). Components can have tensile strength 850 N/mm^2 (27 Mn Si V S 6), or 1000 N/mm^2 (44 Mn Si V S 6), combined with good toughness properties. Development work on these steels and their possible applications is expected to lead to further improvements.

Most spring steels are quenched and tempered steels [599]. Steels containing 1.5 – 2 % Si that were formerly widely used for springs have lost much of their importance because of their strong tendency to surface decarburization, except in cases where the surface quality has to be improved by machining after heat treatment. Heat treatable steels used for springs mostly have carbon content > 0.5 %.

Case Hardening Steels. These are designed for a heat treatment similar to quenching and tempering, but with a different objective, which can be compared to the surface hardening process [600].

Components made of these steels, with carbon content 0.10 – 0.25 %, are carburized by total or partial immersion in powder, or in a molten salt bath, or a gas atmosphere. This can give carbon content > 1 % in the surface layer. After the component has been case hardened, it is tempered at comparatively low temperature (ca. 200 °C). This produces a very hard surface layer and a less hard but very tough interior, resulting in high wear resistance, improved fatigue strength, and very good resistance to impact stresses.

The chemical composition of case hardening steels must be such that good carburization is possible without the formation of a coarse-grained structure at the temperatures used. Components made of case hardening steel are often produced in large numbers, e.g., gear wheels for motor vehicles, so there must be a high degree of uniformity within a melt and from one melt to the next. Plain carbon steels can be used, but usually low- to medium-alloy steels, containing Cr,Cr – Mn,Cr – Mo,or Cr – Ni – Mo, are utilized.

In the United States, case hardening steels are also widely used for roller bearings, but in Europe, highly through-hardening steels containing ca. 1 % C are preferred [601]. These steels are alloyed with ca. 1.5 % Cr. For larger dimensions, the Cr content is increased to ca. 1.8 %, with addition of ca. 0.3 % Mo.

Hardness, wear resistance, and fatigue strength are the most important qualities of roller bearing steels. After quench hardening, stress relief is usually carried out at < 200 °C. The high specifications for purity and metal structure, especially with respect to the carbide structure, have already been described.

6.2.2.1.3. Steels for Elevated Temperature Applications

General. Creep-resistant steels must have strength properties and ductility appropriate for the high operating temperatures expected in use. Many components, e.g., for the high-pressure systems of thermal power stations, are used at an almost constant operating temperature and constant mechanical load. For this type of design, statically determined characteristic properties are used, including the hot yield strength at up to ca. 400 °C, and the creep strain limit or creep rupture strength above this temperature. For fluctuating mechanical and thermal stresses, dynamically determined characteristic properties are used, and the number of cycles of high load (e.g., emergency shutdowns) is limited during the component lifetime. To avoid crack formation in the case of impact loads, the toughness properties must be adequate, both in the material as supplied and after it has undergone prolonged thermal exposure. A material must have good high-temperature corrosion resistance in the medium (gas, air, steam) at the operating temperature.

Properties and Their Measurement. As the temperature increases, the yield strength and tensile strength values determined in tensile tests on metallic materials decrease. At higher temperatures, the mechanical strength is so strongly time dependent that the creep strain limit and creep rupture strength can be less than the hot yield strength.

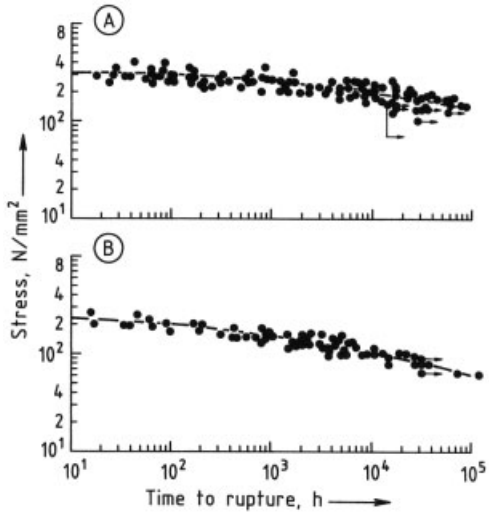

Figure 292. Creep rupture stength of 12 Cr 1 Mo V steel A) At 550 °C; B) At 600 °C

Reliable characteristic parameters for design lifetimes > 200 000 h, e.g., for power stations, can only be determined by very prolonged creep rupture tests. A time extrapolation of test results of this kind by a factor > 3 is not meaningful [602], [603]. Experiments lasting at least 8 years with a large number of test pieces from different heats at various temperatures are required to ensure that the figures for a new material are known with certainty for design purposes (Fig. 292) [604]. Creep tests on welds, in some cases in the intended environment, are also required. Tests in aggressive gases, or in reactor helium [605], greatly increase the cost compared with normal testing in air. If the hot yield strength and creep rupture strength are plotted against the test temperature for a material, the hot yield strength curve intersects that of the creep rupture strength. This intersection point (Fig. 293) [606], indicates the temperature up to which the component design can be based on the material's hot yield strength. For design temperatures above the intersection point, calculations must be based on the creep rupture strength. According to the German technical regulation for steam vessels, TRD Series 100, the lowest of the following figures is critical when calculating the highest operating temperature:

1) Hot yield strength:	$\frac{R_{p\,0.2}}{1.5}$
2) Creep strain limit:	$\sigma_{1/100000}$
3) Creep rupture strength	
at 100 000 h:	$\frac{\sigma_{B/100000}}{1.5}$
at 200 000 h:	$0.8\,\sigma_{B/200000}$

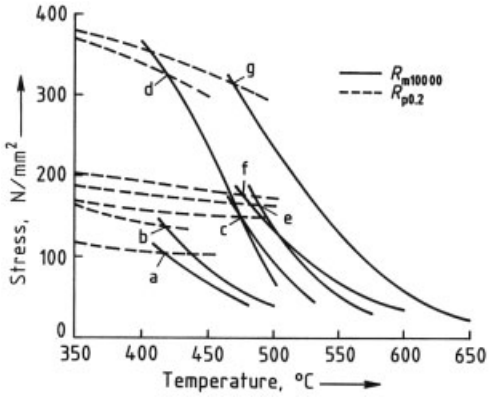

Figure 293. Hot strength values according to DIN 17 175
a) St 35.8; b) 17 Mn 4; c) 15 Mo 3; d) 15 Ni Cu Mo Nb 5; e) 13 Cr Mo 4 4; f) 10 Cr Mo 9 10; g) X 20 Cr Mo V 12 1

The minimum values of the hot yield strength and the recommended values of creep strain limit and creep rupture strength are specified as a function of temperature in the most recent issues of DIN 17 155 for heavy boiler plate, DIN 17 175 for seamless tubes, DIN 17 177 for longitudinally welded tubes, DIN 17 240 for nuts and bolts, and Stahl – Eisen Werkstoffblatt (StEW) 670 for high-temperature steels.

Similar regulations apply in other countries. In the United States, the ASME Code gives allowable stresses as a function of temperature for every material. This is based on the lowest of the following figures:

1) Tensile strength at 20 °C:	$\frac{R_m}{4}$
2) Hot yield strength:	$\frac{R_{p\,0.2}}{1.5}$
3) Creep rupture strength:	$\frac{\sigma_{B/100000}}{1.5}$

For internationally used high-temperature and creep-resistant steels, the characteristic high-temperature strength values have been harmonized by agreement between the national committees. However, for newly developed steels, the creep rupture values available from some countries should be viewed critically, as they may be based on extrapolations by a factor of $\gg 3$, and hence may be too high.

As well as the creep measurements mentioned, isothermal relaxation experiments at high temperatures are carried out for screw stock.

Toughness behavior has so far been assessed from the reduction in area and elongation at fracture measured on creep rupture specimens, and the behavior of notched creep rupture specimens. Good performance properties are postulated if the observed figures for notched test pieces lie within the same scatter band as that for unnotched pieces, or in the upper part of it (Fig. 294) [607], [608]. In future, additional figures obtained in fracture mechanical tests at room temperature and elevated temperatures are to be included in the assessment, and the possibility of embrittlement of the material on prolonged stressing of the component is to be considered.

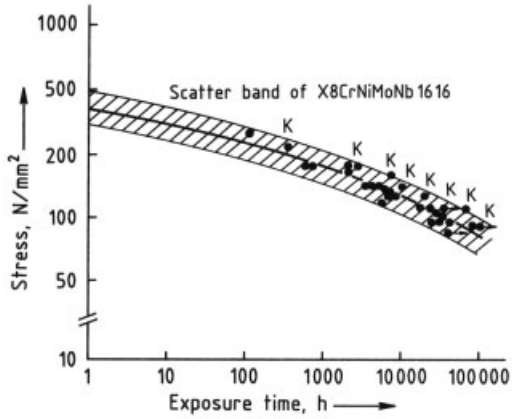

Figure 294. Creep rupture strength at 650 °C of X 3 Cr Ni Mo N 17 13 steel
K, notched sample; arrow, still flowing

Unlike components made of most other types of material, those made of creep-resistant steels are assessed for their remaining life expectancy after various periods of time. Such assessments can be carried out if the lifetime assumed in the design calculations is exceeded, if other design assumptions (pressure, temperature) are changed, or if other components have not withstood similar stresses [609], [610]. In Germany, TRD 508 specifies procedures in such cases.

Grades of Steel. *Plain carbon steels* have been used since the 1900s for temperatures up to ca. 400 °C and low mechanical stresses. To meet the increased demands on the mechanical load-bearing properties of power station components, fine-grained structural steels with improved high-temperature properties have been developed. Their carbon content has been reduced to $< 0.20\,\%$, or even $< 0.15\,\%$, to improve their toughness and to limit hardening in heat-affected zones. They contain $\leq 2\,\%$ manganese and small additions of aluminum, vanadium, niobium, and nitrogen in various combinations to give increased strength by grain refinement. Depending on the trend of development, the hot yield strength can be increased by small additions of molybdenum, nickel, and copper. This group of materials includes normalized ferritic – pearlitic steels and heat-treated bainitic steels. These steels are used for pressure-bearing parts of nuclear power stations, and also in conventional power stations for pressure feedwater pipes, cylindrical boiler shells, etc. (Fig. 295) [611]. There are also areas of application in the chemical and petrochemical industries. These steels are normally used only in the temperature range in which hot yield strength is the design criterion.

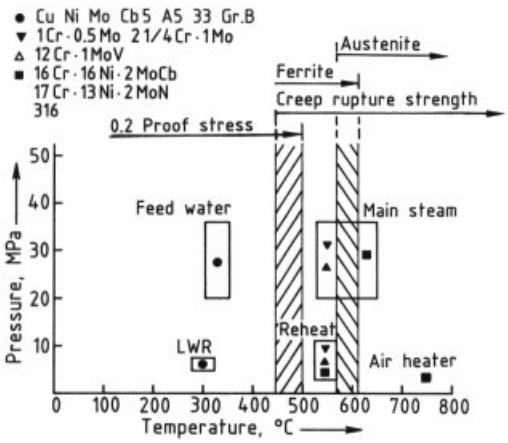

Figure 295. Areas of application of high-temperature steels
LWR = Light water reactor

Mainly as a result of the safety requirements of the nuclear industry, further developments to these steels have been directed to improving their toughness properties. Important developments have included reduction of the carbon content, and strength improvement by the addition of manganese, nickel, and molybdenum (e.g., 15 MnNi 6 3, 15 NiCuMoNb 5, and 20 MnMoNi 5 5 steels). By reducing the sulfur content and chemically bonding the remaining sulfur, a significant increase in toughness can be achieved, especially perpendicular to the direction of rolling [612].

Low-Alloy Steels for Creep Resistance. Molybdenum is the alloying element that most effectively increases creep resistance in this group of steels. The greatest effect is observed at 0.5 % Mo [613]. The oldest creep-resistant alloy steel in Germany is 15 Mo 3 steel, alloyed with 0.3 % Mo to prevent long-term embrittlement. In the United States, Mo content 0.5 % is usual in the steel grade T 1. After heat treatment, molybdenum steel has a ferritic – pearlitic structure. It is mainly used in the temperature range 400 – 500 °C.

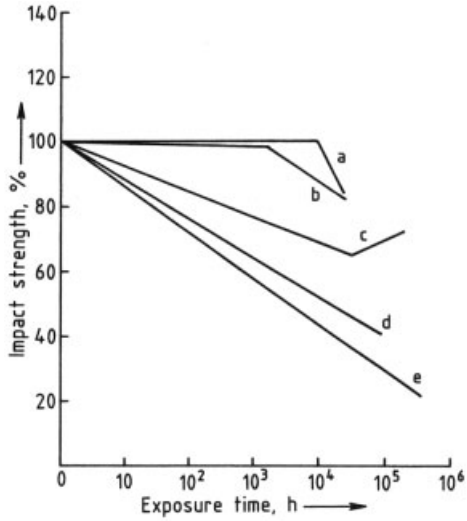

Figure 296. Change of impact strength at room temperature after long-term high-temperature exposure (550 °C)
a) 17Cr-13Ni-2Mo N; b) 16Cr-13Ni-1Mo V Cb N; c) 12Cr-1Mo V; d) 2.25Cr-1Mo; e) 1Cr-0.5Mo

The long-term embrittlement of Mo-containing steels above 440 °C can be prevented by alloying with chromium [614]. This discovery led to the development of stronger creep-resistant steels with molybdenum content up to about 1 %. The steels 13 CrMo 44 and 10 CrMo 9 10, which correspond to the United States steels T 12 and T 22, resulted from this development, and have been used successfully

all over the world in the temperature range 450 – 550 °C (Fig. 295). Depending on size, shape, and heat treatment, these steels have variable ferrite and bainite content. Chromium contributes only slightly to the creep strength, but improves scaling resistance.

Increasingly strict safety requirements are leading to high specifications for toughness properties and creep rupture strengths for these steels. However, in the course of a prolonged period of stress, the toughness of components can decrease to a greater or lesser extent according to the stability of the structure of the material (Fig. 296) [611]. This behavior cannot be changed by metallurgical treatment of the heat.

Martensitic Steels with Higher Chromium Content for Creep Resistance. As power station design develops, the tendency is towards larger units, higher temperatures, and higher pressures. This has led to highly stressed components such as live steam pipework, for which low-alloy steels with very thick walls are used. The development of steels with higher strength and higher creep resistance containing 9 – 12 % chromium, with additions of molybdenum, vanadium, nickel, and niobium, offers the possibility of using lighter and more flexible designs, even for these highly stressed areas (Fig. 297). After heat treatment, these steels have the tempered martensite structure, even with wall thickness as high as 100 mm.

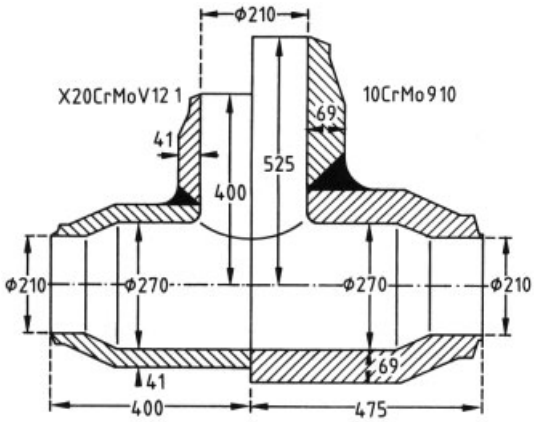

Figure 297. Dimensions of comparable T-pieces in 10 CrMo 9 10 and X 20 CrMoV 12 1

In Europe, a 12 % chromium steel with the German grade number CX 20 CrMoV 12 1 has been used for some decades with great success. Its high carbon content (0.20 %) leads to a fairly high degree of hardening on hot processing or welding, reaching ca. 600 HV. With appropriate temperature control, this steel is included in the program of all European fabricating plants. To facilitate metal-working processes, a 9 % chromium steel with carbon content reduced to 0.10 % was developed in the 1980s [615]. This reaches only ca. 420 HV after hardening [616]. In the United States, this is known as T 91 or P 91 steel, and, in Germany, X 10 CrMoVNb 91. Control of the alloy content within narrow limits and especially the precipitation of very finely divided niobium carbide gives creep rupture strengths in the important temperature range 550 – 600 °C that are 10 – 30 % better than those of X 20 CrMoV 12 1 steel.

As these martensitic steels have the advantage of lower thermal expansion and higher thermal conductivity compared with the austenitic steels used at ca. 600 °C, development programs are in progress, aimed at increasing the creep rupture strength of martensitic steels still further.

As the carbon is in the form of very stable carbides in these steels, they also meet the demanding requirement for resistance to high-pressure hydrogen for the chemical and petrochemical industries.

Austenitic Steels for Creep Resistance. Austenitic Cr – Ni steels are used for operating temperatures $\geq$ 600 °C (Fig. 295). To avoid embrittlement at the operating temperature (σ-phase), these steels are produced with higher nickel and lower chromium content than the stainless grades, e.g., 16 % Cr and 13 % Ni. Addition of N, Mo, W, V, Co, Nb, Ti, Al, or B, individually or in combination, significantly increases the creep strength of these steels.

Unstabilized austenite was at first favored for boiler construction in the United States, but niobium-stabilized austenites were later developed in Germany to reduce the risk of intercrystalline corrosion, and to improve creep resistance; they have been used with great success [607].

In order to produce crack-free welds on even larger sections, a series of unstabilized austenites with much reduced carbon content were developed in Germany at the end of the 1960s [608], [617]. Addition of ca. 0.15 % N compensated for

the lack of niobium, whose presence improves creep rupture strength (Fig. 294).

Austenites with additions of W, Ti, B, and Co were used to meet still higher specifications for creep strength and operating temperatures. For weldable components, boron content is tolerable only up to a maximum of 0.006 %. Steels with cobalt additions of 10 – 20 % [618] give the highest creep rupture strengths above 650 °C. These are used in the construction of gas turbines and power transmission systems. They represent the transition to the high-temperature nickel and cobalt alloys, both in composition and use.

Standardization. Testing procedures at high temperatures are described in EN 10 002-5. The European standard EN 10 028-2 gives a more detailed specification for high-temperature sheet steels, EN 10 216-2 for seamless tubes, EN 10 217-2 for welded tubes, EN 10 222-2 for forgings, and EN 10 213-2 for steel castings. The following United States standards give specifications for high-temperature steels: for sheet, ASTM A 202, A 203, A 204, A 225, A 299, A 302, A 387, A 515, A 516, A 517; for seamless tubes, ASTM A 106, A 161, A 192, A 209, A 210, A 213, A 271, A 335, A 369, A 376, A 426, A 430; for welded tubes, ASTM A 178, A 226, A 249, A 250, A 358, A 671, A 672, A 691; for forgings, ASTM A 105, A 182; for bolts, ASTM A 193; and for nuts, ASTM A 194.

6.2.2.1.4. Low-Temperature Steels

General. Low-temperature steels must have good toughness properties at the lowest planned operating temperature. In cryogenics, gas separation, liquefaction of hydrocarbons, space technology, superconductivity, and other areas, components are subjected to temperatures of − 200 °C and below [619]. These components and their welded joints must have good toughness properties, especially under impact stress, to avoid brittle fracture.

Properties and Tests. Yield strengths and tensile strengths, which are measured by tensile tests, increase with decreasing temperature. With ferritic steels, a decrease in toughness occurs over the narrow temperature range (transition temperature) typical of these grades of steel. At temperatures below the transition temperature, which also depends on the testing method, the test pieces show brittle fracture properties. The transition temperature is shifted to higher temperatures by an increasing degree of multiaxial stress, a sharper notch shape, and a higher rate of loading. Stable austenitic steels show a small decrease in toughness properties with falling temperature, but with no sudden drop.

Toughness measurements are carried out by means of notched-bar impact tests, preferably with the ISO V-notch test piece at a low temperature typical for the material, or at the lowest planned component temperature. The tests must be carried out not only on the base material, but also on the welded joint, the weld metal, and the heat-affected zone. For a fundamental assessment of the material, tests must also be carried out on test pieces similar in size and shape to the component.

Other important application properties include the thermal expansion coefficient and the thermal conductivity [620]. These play an important part in control of shrinkage stresses in long pipes or rigid structures on cooling to the operating temperature, and also when different materials are joined, e.g., when ferritic steels are welded with an austenitic material. Shrinkage stresses in pipework are reduced by compensators, whose capacity and number depend on the thermal expansion coefficient of the steel, the operating temperature, and the length of the pipeline. Shrinkage of austenitic steels is 50 % greater than that of ferritic steels.

Ferritic Steels. Within the range of world climatic temperatures and a little below this, i.e., down to ca. − 50 °C, normalized plain carbon fine-grained structural steels meet toughness requirements. Addition of ≤ 1.7 % Mn and ≤ 0.8 % Ni, with carbon content reduced to 0.16 %, increase both yield strength and toughness within this temperature range. These steels are suitable for propane and propene liquefied gas technology, i.e., for operating temperatures down to − 50 °C.

For lower temperatures, several nickel-containing steels have been developed. Nickel reduces the transition temperature, and favors formation of bainite and martensite. Both these effects lead to a reduction in the size of

the secondary grains. Heat treatment involving quenching and tempering further enhances this effect, and leads to a significant reduction in ductile–brittle transition temperature. Thus, steels containing 3.5 % Ni can be used at − 100 °C, and those containing 5 % Ni down to − 120 °C (below the liquefaction temperatures of carbon dioxide, ethane, and ethene). Steel containing 5.5 % Ni and 0.2 % Mo is a later development which gives improved toughness properties, enabling it to be used even at − 170 °C.

Steel containing 9 % Ni represents a special case. It has outstanding toughness properties even at − 200 °C, and is widely used in natural gas and air liquefaction. Compared with the austenitic steels used in this temperature range, 9 % Ni steel has twice the yield strength and therefore enables designs with smaller load-bearing sections to be used. A further advantage over austenitic steels is the significantly lower thermal expansion coefficient of 9 % Ni steel, enabling the number of compensators in long pipelines subjected to low temperatures to be reduced. On cooling from room temperature to − 200 °C, a 1 km pipe of austenitic steel shrinks by 3 m, while 9 % Ni steel shrinks by only 2 m. This steel also has a special structure. On normalizing by cooling with air or water, a transformation structure of low-carbon martensite with some bainite is at first formed. Subsequent tempering is carried out just above A_{c1}. This produces small amounts of stable austenite in the transformation structure, and these give the toughness properties [621].

In all low-temperature ferritic steels, the sulfur content should be minimized to give optimum toughness properties, and to reduce toughness anisotropy. With 3 % Ni steel, matching welding consumables can be used. The higher-nickel steels are welded with austenitic filler materials. These can be iron-based alloys with ca. 18 % Cr, 12 % Ni, or nickel-based materials with ca. 65 % Ni and 15 % Cr. The lower yield strength of the austenitic weld metal compared with the ferritic base material must be allowed for in designing the component. When austenitic weld materials containing 18 % Cr and 12 % Ni are used, the welded joint must not be heat treated, to avoid diffusion of carbon from the base material into the weld metal. To overcome these problems, a type of welding material containing 11 % Ni has been developed which gives good toughness properties at − 196 °C, and has strength properties that match the base material (Fig. 298) [622], [623].

Austenitic Steels. Austenitic Cr–Ni steels are suitable for operating temperatures close to absolute zero (−273 °C). To optimize toughness properties, the alloy content must be such that a stable austenitic structure is produced, and no martensite is formed by plastic deformation at these low temperatures. These steels are also used in superconductivity technology, as they have no ferromagnetic constituents, even at the lowest temperatures. This is also true for the welded joints, whose freedom from hot cracking (a common occurrence otherwise) must be ensured by including small amounts of ferrite. A stable austenitic steel with a high nickel content and additions of Mo, Mn, and N has proved successful [624]. It also shows a higher yield strength than nitrogen-free austenitic steels (Fig. 299).

Whereas austenitic Cr–Ni steels have a relatively high thermal expansion coefficient, that of austenitic steel containing 36 % Ni is extremely low, ca. 10 % of the value for normal austenites. This material also maintains its good toughness properties near absolute zero. It is therefore suitable for low-temperature components where it would be difficult to design a shrinkage compensation system, or where sudden temperature changes are to be expected. Examples of its use include membrane tanks for liquefied gas tankers and components for filling plants in rocket technology [625].

Standards. Specifications for low-temperature steels are provided by European standards EN 10 028-4 for nickel-steel sheet and EN 10 213-3 for steel castings. The United States ASTM standards include specifications for low-temperature steels as follows: for sheets, ASTM A 320, A 516, A 537, A 553; for pipes, ASTM A 333, A 334; for screws, ASTM A 320; for welded components, ASTM A 350, A 420, A 522.

6.2.2.1.5. Wear-Resistant Steels [626]

General. Wear occurs as a result of the mechanical interaction of two bodies in contact

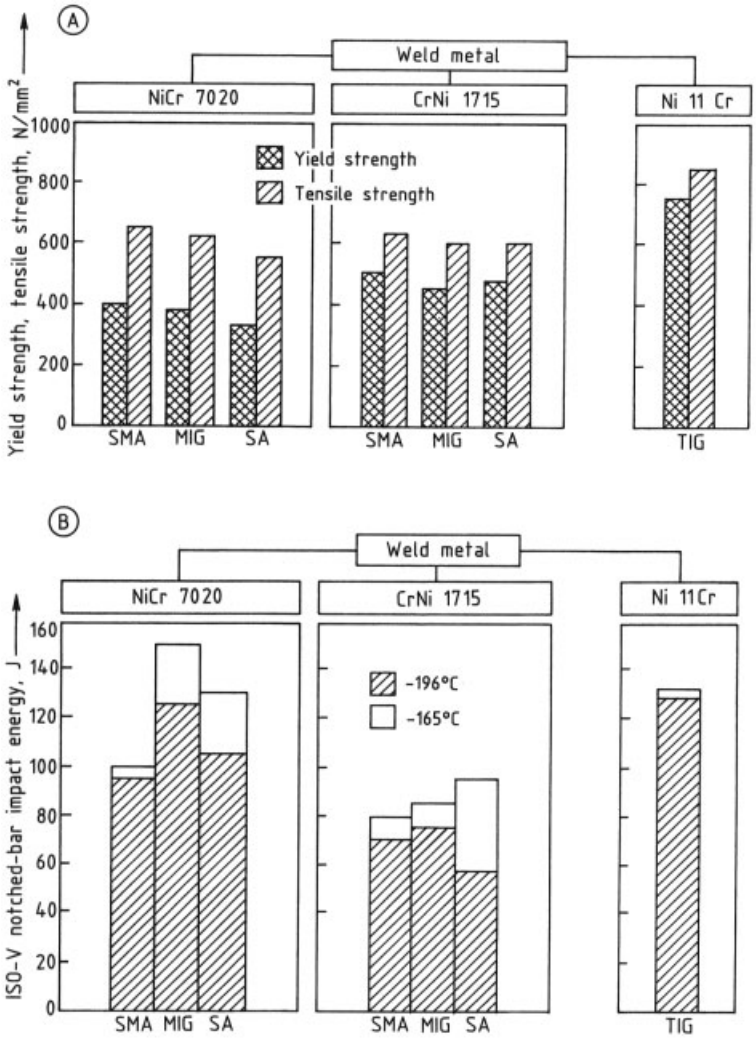

Figure 298. Properties of weld metals for X 8 Ni 9 steel
A) Strength properties; B) Notched-bar impact energySMA = shielded-metal arc welding; MIG = metal – inert gas welding; SA = submerged-arc welding; TIG = tungsten inert-gas welding

which move relative to each other, so that their surfaces become eroded and altered (see also → Abrasion and Erosion). The surface properties of a material subjected to wear, e.g., during the dressing and transporting of granular mineral materials, are therfore of the greatest importance. The nature of the wear system is also very important, e.g., dry or lubricated sliding wear, dry or lubricated rolling wear, jet wear, or other mechanisms. Thus, wear depends not only on the material of the main constructional component, but also on all the other materials that make up the system, and on other conditions. Assessment of the suitability of a material by means of laboratory wear tests is problematic [627]. Tests of this type, e.g., by rotary grinding mills or the impact of a stream of abrasive particles, should be used only for basic research. The evaluation of different iron-based materials for a particular use requires practical investigation of the actual system.

Properties and Methods of Measurement. As wear is a surface problem, surface hardness

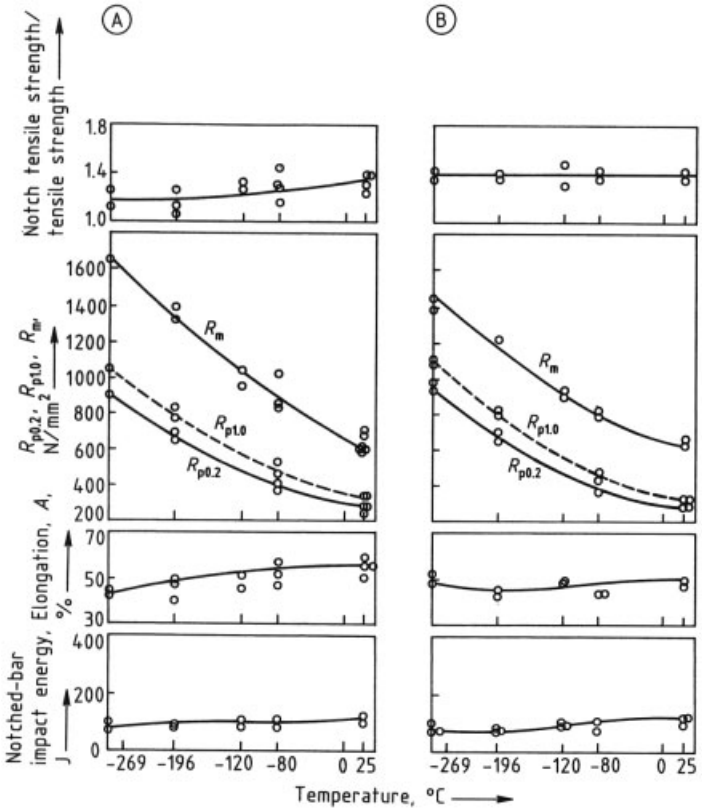

Figure 299. Low-temperature mechanical properties of plate and rod 12 – 80 mm thick
A) X 3 CrNiN 18 10 steel; B) X 3 CrNiMoN 18 14 steelTensile test pieces from plate: transverse; from rod (80 mm square): longitudinal; notched-bar impact test pieces: ISO V-notch transverse; notch-tensile test pieces: $\alpha_k = 4$

is the most important property to be measured. However, when hardness properties are equal, differences in structure can affect wear behavior [628]. Therefore, as well as surface hardness measurements, investigations of metallic structure are carried out. This gives information about the proportions of the various very hard particles in the structure, not only in the surface layer, but also in the cross section, giving the depth to which wear can be tolerated in a component. With steels that are only quench hardened, and also with quenched and tempered hypoeutectoid steels, increased carbon content leads to increased hardness values, and hence higher wear resistance properties. Carbides are responsible for wear resistance in most hypereutectoid steels which are normalized or quenched and only slightly tempered.

Grades of Steel. In quenched ledeburitic steels (> ca. 1.5 % C) and in martensitic white cast iron, resistance to abrasive wear is maximized at the expense of toughness, and hence of resistance to brittle fracture.

Components that are exposed to very severe wear often also need to withstand bending and impact stresses, so toughness and resistance to brittle fracture are also important. Plain carbon steels and alloy steels containing carbide-forming alloying elements, e.g., chromium, molybdenum, and vanadium, are used (50 Mn 7, 42 Cr Mo 4), and also hypereutectoid steels such as the roller bearing steel 100 Cr 6. Tempering treatment is used to adapt the hardness and toughness to the application. In products made of steel plate, e.g., dumper trucks, garbage trucks, concrete mixers, chutes, and bunkers, weldability is important. Quenched and tempered struc-

Figure 300. Hardening the head of a rail
A) Isotherms after inductive heating (°C); B) Macrostructure after hardening with compressed air; C) Hardness distribution (HV10)

tural steels, e.g., Fe 690, have good wear resistance.

Austenitic manganese steel with 1.2 % carbon and 12 % manganese attains its very high wear resistance only after work hardening of the surface layer by impact wear loading. The toughness of the core structure is therefore suitable for dredger buckets, crusher jaws, and similar equipment. Steel castings are often used here.

Good wear resistance properties are also important in railway track [629]. This functions as support, roadway, and as guiding device, so that rails are subjected to very high dynamic stresses [630]. Where the wheel comes into contact with the rail, considerable deformation and work hardening of the steel of the rail takes place. The steel used for rails must have sufficient welding properties for welding processes used to produce continuous rails, e.g., flash-butt welding and aluminothermic welding. The steel grade UIC 90 A, specified by the Union International des Chemins de Fer (UIC) contains ca. 0.75 % C and ca. 1 % Mn, and is used by most railway companies. It must have tensile strength $> 880\,N/mm^2$ in the as-rolled condition.

In very highly stressed sections of railway line, wear resistance is increased by heat treating the head of the rail when this grade of steel is used (Fig. 300) [631]. A fine pearlitic structure is produced in the head area of the rail, giving increased hardness and tensile strength $\geq 1175\,N/mm^2$. This heat treatment is sometimes also carried out at the ends of the rails to improve resistance to impact stresses at the expansion gap in jointed rail.

For highly stressed sections of track in difficult terrain, steel rails with as-rolled strength > 1100 or even $1200\,N/mm^2$ are used [632], [633]. These steels have increased silicon content, with ca. 1 % Cr, and are alloyed with vanadium.

6.2.2.1.6. Steels for Pipelines

General. Gas, oil, water, chemical products, and solids, such as coal and ore in slurry form, are transported through pipelines. The choice of pipeline material depends on the operating conditions and the material being transported. On one hand, high operating pressures, and on the other, temperatures as low as $-60\,°C$ for gas pipelines in arctic regions, affect the choice of material. Corrosion damage is possible on the inside of the pipeline by the material being transported, and on the outside by the surrounding medium. Field welding processes must be suitable for these steels.

Properties and Testing. Testing of the material in the direction of the main stress is carried out on cylindrical tensile samples taken from the pipe wall in a transverse direction. To avoid the Bauschinger effect (see Section 6.1.2.1.2), the sample test pieces are taken in such a way that no dressing or straightening is required. Because of the high toughness requirements, a Battelle drop weight tear test (BDWT) is often carried out as well as the usual ISO V-notch test [634]. The sharply notched test pieces used to span the entire pipe wall thickness. The temperature at which a given proportion of the fracture is ductile (e.g., 85 %) is determined.

In gas pipelines, after initiation of a crack of a critical length, instable cracks can propagate at a high constant velocity over great distances. To avoid such crack propagation, the notched-bar

impact energy should be very high, and fracture should be fully ductile at the operating temperature. To determine the required toughness values, large-scale tests are carried out in which sections of pipelines of varrying toughness are tested in 200 m long test strings. A crack is initiated in a full-scale experimental pipeline which is kept under pressure. The crack propagation rate and mode of propagation give information about the behavior of the material [635]. With increasing pipe wall stresses, notched-bar impact energies must be higher to avoid long-running cracks (Fig. 301).

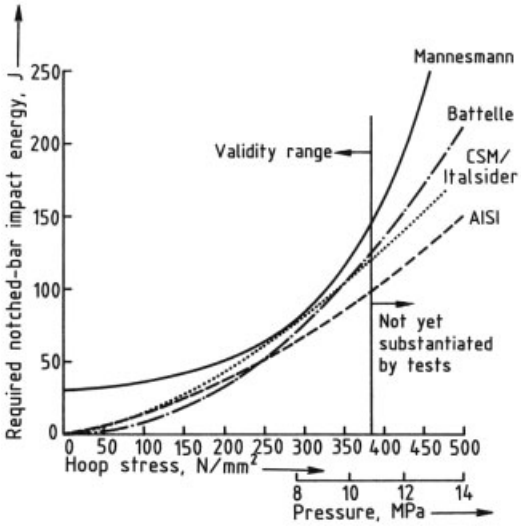

Figure 301. Required notched-bar impact energy as a function of hoop stress
Pipe diameter 1.420 mm; wall thickness 20 mm

The action of condensates containing hydrogen sulfide can lead to hydrogen-induced cracks starting at inclusions in the steel. Tests for resistance to this type of corrosion consist of keeping rectangular test pieces in a standard solution, saturated with H_2S, for 96 h, followed by metallographic inspection for the presence of hydrogen-induced cracks [636].

Whereas hydrogen-induced corrosion can take place at normal pressures, tensile stresses must be present for hydrogen-induced stress corrosion cracking to occur. This may be prevented by not exceeding a critical stress value which is in the region of 20 – 50 % of the yield strength of the material. Mechanical expansion or quenching and tempering of the tubes is also effective in reducing critical internal stresses [637].

Field weldability is of special significance when efficient welding processes are used, e.g., vertical down welding using cellulose-coated electrodes, and mechanized welding with protective gas, assessment of fabricability being the main consideration. For this purpose, a special carbon equivalent (CE) is calculated for these low-alloy steels [638]:

$$\begin{aligned}\mathrm{CE},\ \% = \mathrm{C},\ \% &+ \frac{\mathrm{Si},\ \%}{25} \\ &+ \frac{\mathrm{Mn},\ \% + \mathrm{Cu},\ \%}{20} \\ &+ \frac{\mathrm{Cr},\ \%}{10} + \frac{\mathrm{Ni},\ \%}{40} \\ &+ \frac{\mathrm{Mo},\ \%}{15} + \frac{\mathrm{V},\ \%}{10}\end{aligned}$$

This is suitable for assessing the hardening tendency of the steel. To measure susceptibility to cold cracking, the implant test is used.

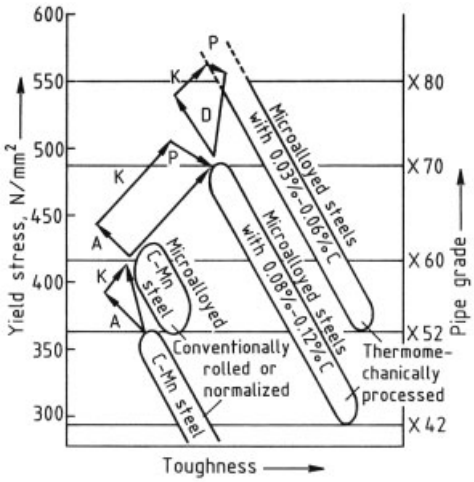

Figure 302. Development of pipeline steels
Numerical values not shown on the abscissa, as measured toughness values depend on testing method
A = precipitation hardening; K = grain refining; P = reduction of pearlite content; D = dislocation hardening

Grades of Steel. In the development of steels for pipelines, steels with minimum yield strength 420 N/mm^2 were produced by microalloying normalized manganese steels with V and Nb. Good toughness properties were also obtained by a combination of grain refining and precipitation hardening.

The trend to higher material transport rates has led to increased pipeline diameters that are

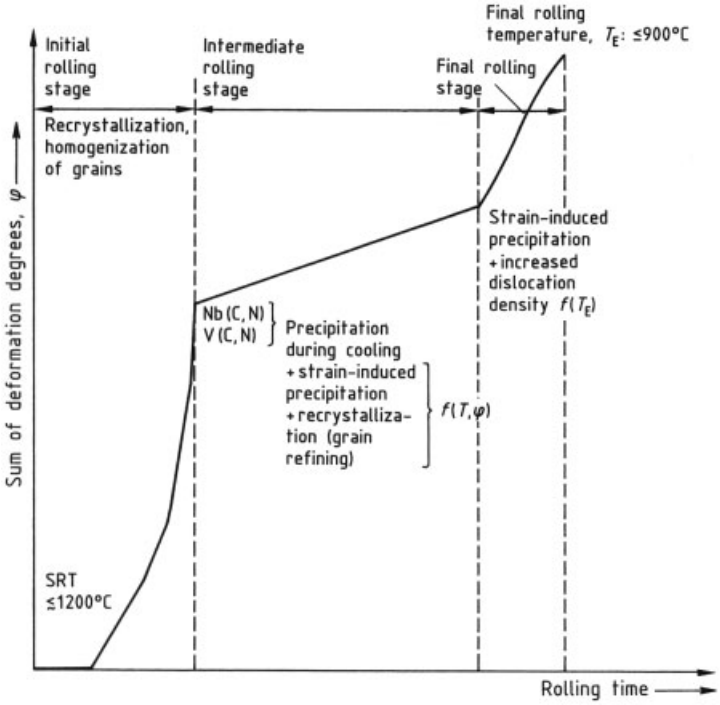

Figure 303. Physical metallurgical processes during thermomechanical rolling as a function of the sum of deformation degrees, rolling time, and rolling temperature at various stages of rolling
SRT = slab reheating temperature

only possible with welded pipes and higher operating pressures. To minimize wall thickness and hence the mass of the pipeline, materials with higher yield strengths and higher toughness values were developed (Fig. 302).

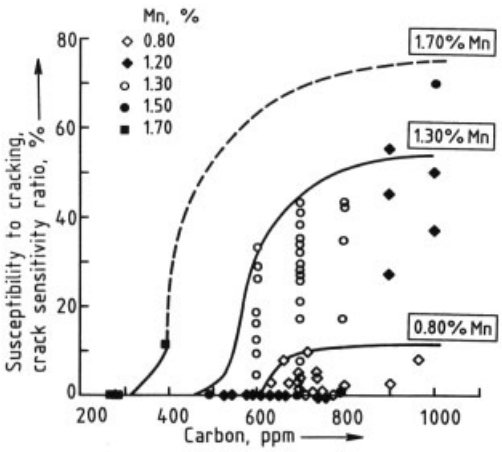

Figure 304. Effect of C and Mn content on susceptibility to hydrogen-induced cracking
Solution: H_2S 2.5 g/L, pH 3

Steels with good field weldability and minimum yield strength 552 N/mm^2 were produced by thermomechanical treatment of low- carbon, low-alloy steels (Fig. 303) [639].

Further improvement in these properties can be obtained by quenching and tempering to a bainitic structure with an even smaller grain size than that in the thermomechanically produced structure. A combination of thermomechanical treatment with quenching directly from the final rolling temperature leads to particularly good results. For special applications, water-quenched and tempered pipes or heat-treated plates for the production of pipes are used.

Resistance to sour gas (hydrogen-induced cracking) is provided by extremely clean steels. The steel must be desulfurized to < 0.003 % S, and the residual sulfur must be reacted to form sulfides that do not easily deform into the shape of flat inclusions. This prevents formation of cracks by recombination of the hydrogen (dissolved in the structure in atomic form) to form molecules at these inclusions. Also, the formation of hard macrosegregations in the steel must

be prevented, as this too can promote the formation of hydrogen-induced cracks [640]. Reduction in carbon content can also reduce the segregation of other elements that contribute to hardening, e.g., manganese, and thereby reduce the tendency to hydrogen-induced cracking (Fig. 304).

Standards. A European standard for testing resistance to hydrogen-induced corrosion is in course of preparation. The German standard DIN 17 172 and the United States API standards 5 L, 5 LS, 5 LX, and 5 LU give specifications for pipeline steels. Customers' delivery conditions often give specifications that enable the demands of particular applications to be met, and correspond to the current state of technology.

6.2.2.2. Steels with Special Chemical Properties

6.2.2.2.1. Stainless Steels

General. Stainless steels have high resistance to chemically aggressive materials. They are used under normal environmental conditions (air, water), and in corrosive chemical solutions (acids, alkalis). These steels have chromium content $> 12\,\%$, and can also contain other alloying additions of Ni, Mo, Cu, Si, Ti, Nb, or other elements.

The areas of application of these steels range from domestic equipment, motor vehicles, and architecture, to chemical equipment, energy generation, marine technology, and environmental protection.

Properties and Testing. Steels with chromium content $> 12\,\%$ are passivated in the presence of oxygen or even such weak oxidizing agents as water. The surface becomes covered with a very thin, dense layer of oxide which restricts further electrochemical reactions and hence the dissolution of metal. Mechanical damage to the passive layer heals very rapidly. In water and many other aggressive media, only an extremely slight, uniform surface corrosion of the passivated surface takes place, which is virtually unmeasurable. Under reducing conditions, e.g., in sulfuric and phosphoric acids, a stable passive coating is not formed. Even very highly alloyed stainless steels dissolve smoothly and uniformly when in this active state. The temperature and concentration of the attacking medium therefore determine the rate of dissolution of a stainless steel. As this process takes place at a constant rate, the long-term behavior can be determined from short-term corrosion tests. Corrosion rates of 0.3 mm are tolerated in many applications. The release of metallic ions into the environment or the product is sometimes unacceptable, e.g., in pharmaceutical and photochemical products, food chemistry, or surgical implants.

Whereas component lifetime can be accurately estimated if surface corrosion is uniform, local corrosion phenomena can lead to sudden unforeseen failures. This group of local corrosion phenomena includes pitting, crevice corrosion, stress corrosion cracking, and intercrystalline corrosion.

Pitting of stainless steels is caused almost exclusively by chloride ions. These cause the steel to become locally active, and the active region is anodically dissolved. The resistance of stainless steels to pitting is improved by increasing the Cr, Mo, and N content (Fig. 305) [641]. The effectiveness of these elements can be expressed as the pitting resistance equivalent (PRE):

$$\mathrm{PRE} = \mathrm{Cr}\% + 3.3\,\mathrm{Mo}\% + 30\,\mathrm{N}\%$$

The incidence of pitting also depends on other factors, e.g., stability of the microstructure, condition of the surface, flow conditions in the attacking medium.

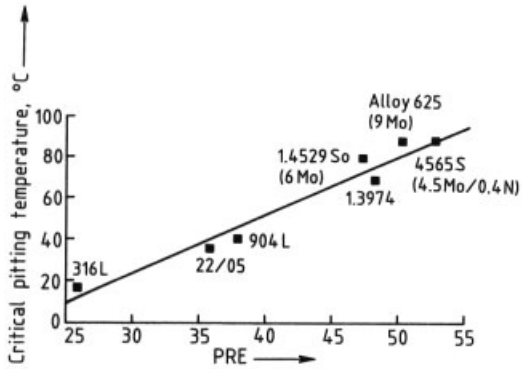

Figure 305. Correlation between pitting resistance equivalent (PRE) and the pitting resistance of stainless steels $FeCl_3$ test

Crevice corrosion is closely related to pitting. It is also caused by chloride ions, and is a modified form of pitting (Fig. 306). It can also occur in steels that are resistant to general pitting corrosion in the same attacking medium. Crevices between stainless steels and nonconductors, e.g., ceramic or plastic sealing materials, are especially damaging. Precautions should be taken against crevice corrosion at the design stage. The pitting tendency of stainless steel can be estimated by current density – potential determinations, which give a measure of pitting resistance. Estimation of resistance to crevice corrosion requires long-term tests of samples with a simulated crevice (contact with a plastic disk) in the attacking medium.

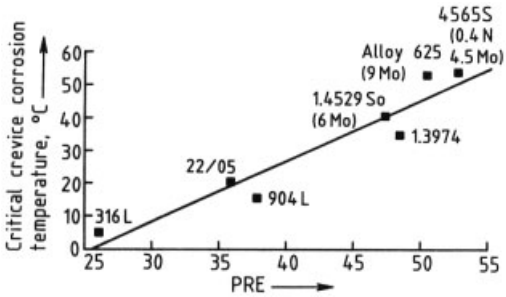

Figure 306. Correlation between PRE and resistance to crevice corrosion of stainless steels
$FeCl_3$ test

Stress corrosion cracking can cause components to fail completely within a very short time, by crack formation over the entire load-bearing section. The condition which leads to stress corrosion cracking is attack by an aggressive medium combined with mechanical tensile stress. Media which attack austenitic stainless steels include chloride- containing and alkaline solutions. In chloride- containing solutions, stress corrosion cracking often starts at previously formed pits. Alloying elements that improve pitting resistance also improve resistance to stress corrosion cracking in chloride solutions. Chloride-induced stress corrosion cracking can be almost completely controlled by high nickel content. The critical mechanical stress for this type of corrosion is very low ($50\,N/mm^2$) [642], [643], as is the temperature above which stress corrosion cracking can take place in aqueous chloride solutions (ca. 50 °C; Fig. 307) [644]. Testing for susceptibility to stress corrosion cracking is performed on mechanically stressed bend or tensile test pieces simultaneously subjected to the attack of the medium, whose composition and temperature must be kept constant for the duration of the test.

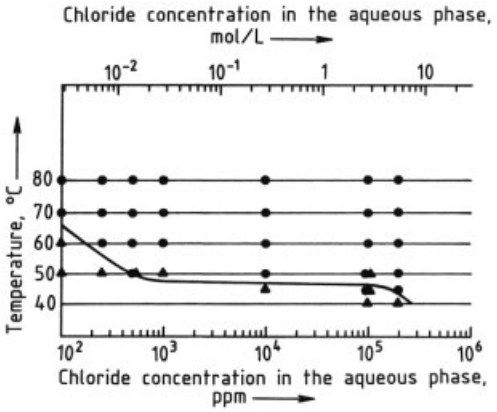

Figure 307. Effect of temperature and chloride concentration in the aqueous phase on the stress cracking resistance corrosion of U-bend specimens of 18 Cr 10 Ni steel in NaCl solutions
Isothermal storage of specimens with three-phase interface-Circles = stress crack corrosion, starting from pitting sites; diamonds = no stress crack corrosion after a test period of 1000 h

In chromium- containing steels, chromium-rich carbides $Cr_{23}C_6$ are precipitated at the grain boundaries at moderate temperatures. This causes the zone adjacent to the grain boundaries to be depleted in chromium (Fig. 308) [645]. The steel then becomes sensitized to intercrystalline corrosion. This can take place, e.g., by stress-relief heat treatment of the entire component, or in areas close to a weld. This selective corrosion takes place only in the presence of an attacking medium. It can be prevented by greatly reducing the carbon content, i.e., to $< 0.07\,\%$ or even $< 0.03\,\%$, or by bonding the carbon with Nb or Ti to form stable carbides. To determine resistance to intercrystalline corrosion, samples of the material are metallographically examined for grain boundary separation after a predetermined sensitization treatment, e.g., 30 min at 650 °C, followed by prolonged exposure to a test solution. Solutions of $CuSO_4/H_2SO_4$ or HNO_3/HF are used for these tests. The most sensitive is the Huey test in boiling HNO_3, as this is sensitive not only to chromium depletion, but also to the precipitation of nitrides and intermetallic

phases [646]. Some types of selective corrosion that seldom occur with stainless steels include "knife-line" attack (near weld seams on stabilized steels) and corrosion fatigue [647].

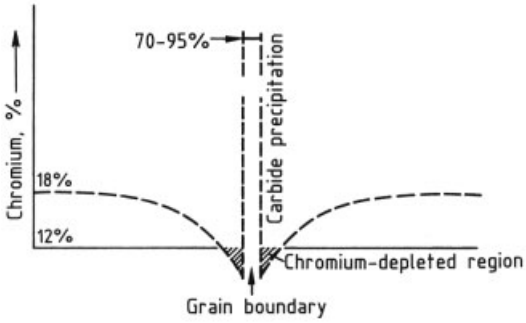

Figure 308. Chromium depletion near the grain boundary caused by precipitation of Cr-rich carbides at the grain boundary

Grades of Steel. The first classification of stainless steels by structure [648] was made in 1920, and is used today in all systematic descriptions. One of the most well known is classification by the Schaeffler diagram [649], which describes the microstructures of weld materials of chromium and Cr – Ni steels (Fig. 309).

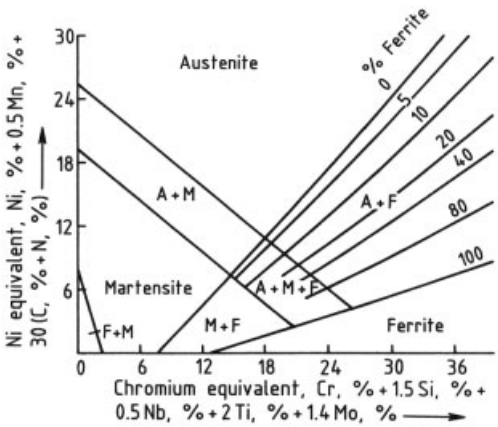

Figure 309. Effect of alloying elements on the microstructure

Ferritic Steels [650]. This group starts with 11 % Cr steels with low corrosion resistance, e.g., for vehicle construction and for containers. Further steels contain ferrite, together with larger amounts of transformation structures. Steels for motor vehicle catalytic exhaust systems contain 12 – 18 % Cr. Dishwashers and washing machine drums contain 18 % Cr. The corrosion resistance of this grade of steel can be further improved by the addition of 2 % Mo. Further improvements have led to a 30 % Cr, 5 % Mo steel. These superferrites are produced with extremely low carbon and nitrogen content, and can be further stabilized by addition of Nb. Because of their very high PRE, they can be used when very good pitting resistance properties are specified. Stabilization, i.e., bonding of all the carbon and nitrogen by Ti or Nb, is essential for all components of ferritic steels intended for welding, to prevent intercrystalline corrosion in the heat-affected zone.

Normal ferritic chromium steels have their ductile – brittle transition temperature near room temperature. As they do not undergo transformation, when they are hot rolled in large sections there is a tendency to form a coarse-grained structure which leads to a further reduction in toughness. In the superferrites, reduction of the carbon and nitrogen content causes the transition temperature to be shifted to lower temperature, and the upper shelf energy measured on the notched-bar impact specimen to be significantly improved. With increasing Cr and Mo content, 475 °C embrittlement causes the embrittlement tendency of ferritic chromium steels to increase in the range 350 – 500 °C, and precipitation of χ- and σ-phases [651] causes a similar increase in the 500 – 800 °C range. This means that these steels are not suitable for prolonged use > 300 °C, and that cooling from the final working temperature must be very rapid through this temperature range. With high Cr and Mo content, the required cooling rate can be achieved only with small sections. A superferrite containing 28 % Cr and 2 % Mo, with sheet thicknesses ≤ 15 mm, can be successfully worked [652]. This steel is used for seawater desalination plants and flue gas desulfurization plants.

Martensitic Steels. Only moderate corrosion resistance can be achieved with martensitic chromium steels because customers require chromium content 13 – 17 %, but it can be somewhat improved by adding ca. 1.5 % Mo. Higher strength and hardness can be achieved with normal martensitic chromium steels with carbon content 0.20 – 1 %.

In the quenched and tempered condition, steel grades containing 0.20 % C are used for steam turbine blades and other machine components. The higher-carbon steels are used in the hardened condition for cutting equipment, roller bearings, and components with good wear resistance properties. The toughness can be increased by stress relieving at 200 – 350 °C.

An alternative to the quenched and tempered chromium steels with 0.20 % C is provided by the nickel-containing martensitic steels in which the carbon content has been reduced to < 0.05 %, and ≤ 6 % Ni is added as the austenite former. These measures lead to considerable improvements in toughness, corrosion resistance, and welding properties.

Another group of steels with even higher strength properties has been developed from these nickel-containing martensitic steels. The addition of Al, Cu, Nb, or Ti causes precipitation hardening, i.e., intermetallic phases, leading to strengths of up to 1400 N/mm^2 in these maraging steels.

The nickel-containing martensitic and maraging steels should not be used above 300 °C, owing to the possibility of structural changes.

Austenitic Steels. The austenitic Cr – Ni steels constitute the highest proportion of stainless steels produced. As well as the basic grade containing 18 % Cr and 9 % Ni, a large number of steels with reduced carbon content and additions of N, Mo, Cu, Si, Ti, and Nb have been developed.

The Cr – Ni steels are used for kitchen equipment, interior decorations, and components of exterior architecture that are not subjected to severe corrosion. Austenitic Cr – Ni – Mo steels are preferred for greater resistance to pitting and crevice corrosion by aggressive outside atmospheres, for domestic pipework, and in automobile construction. Both these basic types of austenitic steels are widely used in the chemical industry.

Austenitic steels with high proportions of alloying elements have been developed to improve their general corrosion resistance and to prevent selective corrosion.

To prevent intercrystalline corrosion, almost all austenitic grades of steel produced today have reduced carbon content and/or are stabilized with Ti or Nb/Ta. Pitting and crevice corrosion resistance are improved by increasing the Cr, Mo, and N contents (Figs. 305 and 306). Chromium content ≤ 30 % and molybdenum content ≤ 6 % are typical of state-of-the-art austenitic steels with good resistance to chloride-containing media, sulfuric acid, and organic acids. A Cr – Ni steel with added Si is particularly resistant to concentrated nitric acid.

As the austenitic Cr – Ni steels have low yield strength (0.2 % offset value ca. 200 N/mm^2), relatively large sections are needed for high mechanical loads. This situation can be improved by the addition of nitrogen, which causes a very large increase in the yield strength of austenitic Cr – Ni steels (Fig. 310) [653]. Melting or electroslag remelting under pressure gives considerably higher nitrogen content than can be achieved by melting under atmospheric pressure, i.e., maximum 0.4 % N. Nitrogen improves both the yield strength and the austenite stability. Precipitation of carbides and intermetallic compounds such as χ-, σ-, and Laves-phases is shifted to longer annealing times by nitrogen, so that the working of Cr- and Mo-containing steels is facilitated. Under long-term exposure at high temperature embrittlement must be guarded against. Sensitization to intercrystalline corrosion by prolonged operation at high temperature must be borne in mind if periods of operation under wet corrosion conditions are expected.

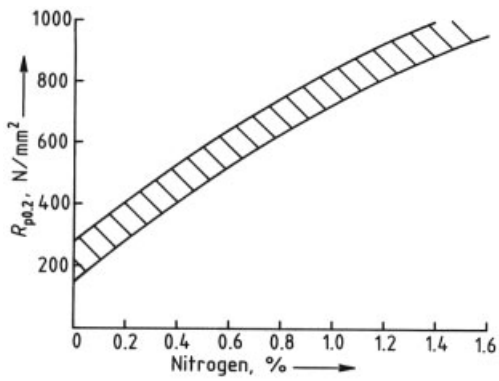

Figure 310. Effect of nitrogen content on the 0.2 % offset yield strength $R_{p0.2}$ of nitrogen-alloyed austenitic steels at room temperature

When welding austenitic steels with high Mo content, segregation and precipitation in the weld and heat-affected zone can take place, with

a reduction in corrosion resistance. This effect becomes more significant at higher Mo content (Fig. 311) [641], [654]. If part of the Mo is replaced by Cr and N, the welded joint has a slightly reduced corrosion resistance for the same PRE.

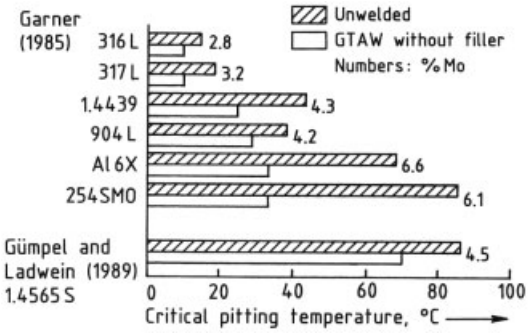

Figure 311. Critical pitting temperature of stainless steels in the solution-annealed condition and after autogenous gas tungsten arc welding (GTAW) without filler

Ferritic – Austenitic Steels. The importance of the ferritic – austenitic steels has increased considerably because of their spectrum of useful properties [655]. Their composition range, i.e., $>20\,\%$ Cr, $5-8\,\%$ Ni, and $\leq 3\,\%$ Mo, leads to a mixed structure of austenite and ferrite. This enables 0.2 % offset yield strength $>450\,\mathrm{N/mm^2}$ to be achieved at room temperature.

The higher PRE compared with standard austenitic steels gives this type of steel a considerably better resistance to chloride-induced selective corrosion. Their general corrosion resistance is maximized by adhering to certain ratios of austenite to ferrite, so the achievement of this structural state is mainly determined by the ratio of ferrite formers to austenite formers (Fig. 309). These steels also have added nitrogen. This has an austenite-forming action, increases the PRE, and decreases the precipitation of σ-phases that lead to embrittlement. When specifying possible operating temperatures, the possibility of 475 °C embrittlement must also be borne in mind. This limits the highest operating temperature to 280 °C.

The composition of the welding material must be such that the cooling process after welding gives a ferrite/austenite ratio corresponding to that of the base material.

The ferritic – austenitic steels have a lower thermal expansion and a better thermal conductivity than the austenitic steels.

The ferritic – austenitic grades of steel are used in seawater desalination plants, sour gas pipelines, and especially where their good resistance to stress corrosion cracking and corrosion fatigue is important, e.g., in chemical and petrochemical industries, especially if the components concerned are in motion.

Standards. A list of standard grades of stainless steel is planned for the European standards. The Europea standards EN 10 028-5 specify stainless steels for plate and strip for pressure vessels; EN 10 088-1 gives delivery requirements for plate and strip for general use; EN 10 088-2 for rod material for general use; EN 10 088-3 for flat products for pressure vessels; EN 10 222-5 for forgings; and EN 10 213-4 for steel castings. Standards for stainless steel tubes are also planned.

The following specifications for stainless steels are given in United States standards; for sheets, ASTM A 167, A 176, A 240; for seamless tubes, ASTM A 213, A 268, A 269, A 270, A 271, A 312, A 336, A 511; for welded tubes, ASTM A 249, A 268, A 269, A 270, A 312, A 409; for forgings, ASTM A 403.

6.2.2.2.2. Heat-Resistant Steels

General. Heat-resistant steels have good resistance to high-temperature corrosion, especially by hot gases, ash constituents, and molten salts and metals at temperatures $>550\,°\mathrm{C}$. As well as chemical resistance, these steels must also have the necessary mechanical load-bearing properties at high temperature, and must be unaffected by temperature changes. For prolonged use at high temperature, the microstructure must remain stable, and embrittlement should not occur.

Properties and Testing. The rate of scale formation (oxidation rate) on plain carbon steel increases with temperature. This rate can be significantly reduced by alloying with Cr, Al, and Si. The oxidation products of these elements form protective layers which restrict oxidation processes at the metal surface. Resistance to scaling increases with increasing proportions of

the elements that form protective layers. Small additions of Ce, Hf, or Y increase adhesion of the oxide layer to the metal surface.

The scaling resistance of a steel is measured by gravimetric determination of the weight loss due to scaling of test pieces after long periods at high temperature. Adhesion of the oxide layer can be assessed by cooling the samples a number of times during the test (Fig. 312). Components made of these types of steel are usually used not only in air, but also in atmospheres with other partial pressures of oxygen, possibly also containing reactive gaseous, solid, or liquid components. Testing should be carried out in these atmospheres in such cases.

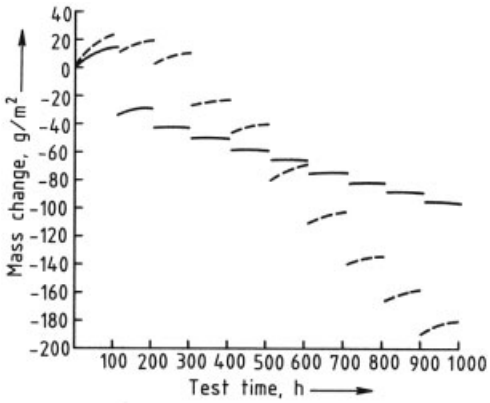

Figure 312. Mass changes on thermal cycling
Solid lines = Fe – Ni – Cr steel + Ce; dashed lines = Fe – Ni – Cr steels + Si

The tests can simulate actual operating conditions. For example, alloys for heating elements are very rapidly heated to the testing temperature by an electric current, and are then subjected to many more temperature changes. The mechanical stresses thus applied to the scale resemble those in a real situation.

As well as hot corrosion tests, high-temperature long-term testing is also carried out in order to obtain mechanical parameters for design purposes. These tests are sometimes carried out in aggressive gas mixtures, with very sophisticated equipment. Toughness tests are also carried out after long-term exposure to high temperatures to determine the embrittlement tendency of these steels. The ferritic chromium steels containing $> 13\,\%$ Cr tend to undergo 475 °C embrittlement within the range 300 –550 °C [656], and σ-phase embrittlement in the range 600 – 800 °C [657]. At higher operating temperature, existing precipitates redissolve, reversing the effects of embrittlement. In many cases, embrittlement does not affect operating behavior, as shown by experiments on valve steels [658]. However, embrittlement caused by service exposure of components must be taken into consideration when making repairs.

For valve steels, determination of fatigue limits at high temperature and an engine test for the qualification are important.

Ferritic Steels. If the specifications for mechanical load-bearing capacity are not too high, high-temperature corrosion resistance can often be obtained by the use of ferritic chromium steels. These contain 12 – 25 % Cr with additions of Al and Si. As they do not contain nickel, these steels are stable in neutral and oxidizing atmospheres, even if these have sulfur-containing components. However, they are significantly less resistant to reducing H_2S-containing atmospheres [659], [660]. If reducing gases such as CO or hydrocarbons are present, carburization of ferritic steels can occur at operating temperatures $> 500\,°C$. This carbon can then combine with a significant portion of the active chromium content, leading to a considerable decrease in scaling resistance. In gases rich in nitrogen or ammonia, scaling resistance can be reduced by combination of the nitrogen with the active Al to form AlN, especially if there is damage to the scale layer.

Whereas the usual heat-resistant chromium steels, which contain additions of $\leq 2\,\%$ Si and Al, form a protective layer consisting mainly of chromium spinels at high temperatures, chromium steels with aluminum content $> 4\,\%$ form a dense protective layer of Al_2O_3 which gives them even better scaling resistance. As these steels have high electrical resistance, they are especially suitable for use in heating elements. The low variation in electrical resistivity with temperature of these steels is also an advantage when they are used as high-temperature conductors. The higher aluminum content also improves their resistance to sulfur-containing gases and suppresses precipitation of the σ-phase, a cause of embrittlement (Fig. 313) [657]. In these steels, an Al content of 6 % should not be exceeded at 15 – 25 % Cr to prevent the formation of a coarse-grained structure.

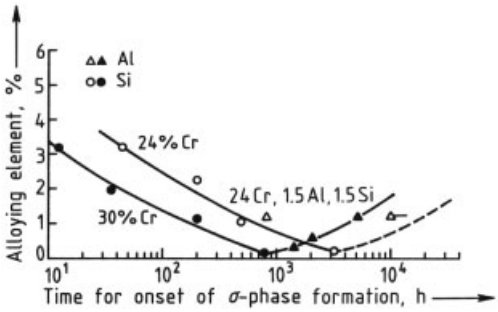

Figure 313. Effect of silicon and aluminum on the time for onset of σ-phase formation at 600 °C in steels containing ca. 24 % and ca. 30 % Cr

Steels used in the valves of internal combustion engines, which operate at ca. 550 °C, should have high hot strength and good hot corrosion resistance. They are therefore quenched and tempered steels with chromium content 9 – 18 %, and also sometimes contain higher levels of Si and additions of the special carbide formers Mo, W, and V to improve hot strength and hot wear resistance properties. These steels are oil quenched and tempered to a strength of 900 – 1100 N/mm^2. Their structure consists of tempered martensite with included carbides.

Austenitic Steels. If higher mechanical stresses are to be expected, heat-resistant austenitic Cr – Ni steels, Ni – Cr alloys, or Co – Cr – Fe alloys are used. These steels, too, mainly rely on chromium to give oxidation resistance, chromium content being 18 – 27 %. Silicon, another ferrite former, can be added at $\leq$ 2.5 %. To maintain a fully austenitic structure with this high level of ferrite formers, nickel content is 10 –36 %. This prevents the formation of δ-ferrite, which under exposure to high temperature would accelerate the precipitation of the σ-phase, which causes brittleness. The austenitic steels have a lower tendency to embrittlement then the ferritic steels, and steels with > 30 % Ni are in any case not likely to be affected by σ-phase precipitation.

Unlike ferritic steels at room temperature, austenitic steels have very good toughness properties and good deformability. Their resistance to carburizing and nitriding atmospheres is also significantly better. Steels containing > 30 % Ni or > 2 % Si are particularly stable in carburizing gases (Fig. 314) [661]. A layer of SiO_2 forms under the chromium spinel protective layer, directly in contact with the metallic surface. This strongly inhibits carburization. Nickel – chromium alloys have the highest resistance to carburization and nitridation [662]. The diffusion and solubility of carbon and nitrogen decrease with increasing nickel content.

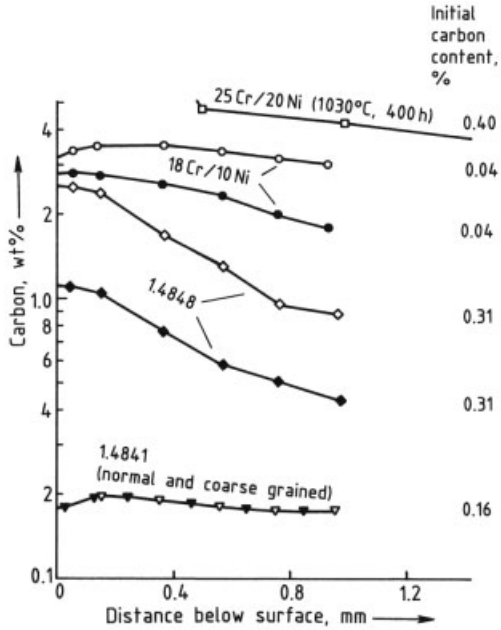

Figure 314. Carburization at 950 °C after 500 h in H_2 10 L/h + town gas 10 L/h
Carbon analysis values at various depths

If the heat-resistant steels contain small amounts of elements that are more easily oxidized than chromium, e.g., Al, Ti, or Si, internal oxidation of these elements can take place under the interface between the scale and the steel [663]. The effectiveness of these elements for improving scaling resistance or hot strength is reduced, and the load-bearing section is weakened.

If the content of these elements is so increased that a dense continuous layer of the oxide is formed, the solubility of the oxygen that diffuses into the metallic material is limited by its equilibrium concentration in the oxide layer. Because of the decrease in the concentration of oxygen from the interface region to the interior of the material, internal oxidation can take place only of those elements that are more easily oxidized than the scale-forming elements. The formation of a pro-

tective scale requires the presence of > 4 % Al or > 2 % Si.

Attack by reducing sulfur-containing gases can cause damage by internal sulfidation in these high-nickel steels, as shown in the Pourbaix – Ellingham diagram (Fig. 315) [664]. Liquid nickel sulfide is formed at > 650 °C in and under the oxide layer, which loses its protective action.

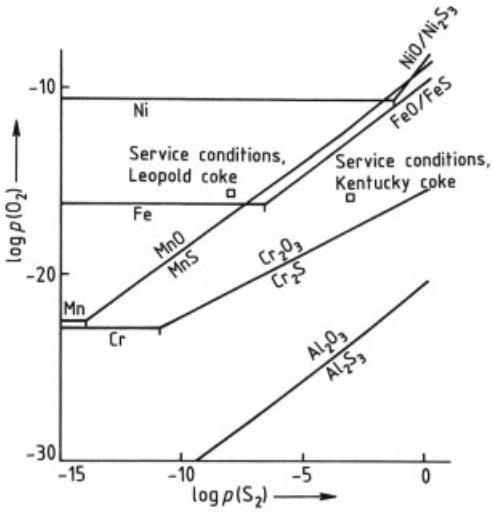

Figure 315. Stability diagram for some alloying elements and the service conditions for high- and low-sulfur coke

Chromium – nickel steels containing > 30 % Ni and 20 % Cr, and nickel alloys containing 20 % Cr are used as austenitic electrical resistance alloys. Their resistance is lower than that of ferritic electrical resistance alloys. Their higher hot strength can be an advantage for various applications in this area. However, they cannot be used at the high temperatures of the ferritic electrical resistance steels, as their scaling resistance temperature does not exceed 1200 °C.

Because of their high hot strength, austenitic valve steels are used at material temperatures of 900 °C. This means that they can be exposed to the severe conditions under which exhaust valves operate. Here, they are subjected to the strongly corrosive effects of the products of the combustion of the fuel and its additives [658]. The products of the oxidation of the organic sulfur and vanadium compounds in diesel fuel also contribute to the corrosive attack on exhaust valves. The liquid combustion residues in particular reduce the scaling resistance, while the solids can cause wear of the valve faces. The latter are often provided with cladding to improve the lifetime of the valves. Austenitic valve steels contain 0.60 % C and 0.45 % N to increase the hot strength. As well as the 20 % chromium, Mo, V, and W are added as special carbide and nitride formers. Manganese content ≤ 10 % and nickel content ≤ 20 % ensure that an austenitic microstructure is formed. Precipitation hardening, which also takes place during operation of the valve in the engine, leads to the required strength properties, including fatigue resistance, at the high operating temperatures.

Standards. The German Stahl-Eisen-Werkstoffblatt 470 gives specifications for heat-resistant steels. Electrical resistance alloys are described in DIN 17 470, and valve materials in DIN 17 480.

6.2.2.2.3. Steels for Hydrogen Service at Elevated Temperatures and Pressures

General. In high-pressure syntheses, oil refining, and other high-pressure technologies, hydrogen at high pressure and temperature can have an effect on constructional material. These materials must be resistant to attack by H_2 under the operating conditions, and must have sufficient hot strength to be stable under the high mechanical loads.

Properties and Testing. At metallic surfaces, H_2 molecules at high pressure can be split up into H atoms which can then penetrate into the metallic lattice. At > 200 °C, this atomic hydrogen can react with Fe_3C in the steel structure to form CH_4. The steel is thus decarburized where the reaction occurs and consequently loses its strength. The CH_4 formed cannot diffuse through the metallic lattice, and collects at the grain boundaries and inclusions. There, the CH_4 and the recombined H_2 molecules give rise to high pressures which can destroy the material by crack formation [665]. Remedial measures can be taken by bonding the C to form stable carbides, e.g., with the alloying elements Cr, Mo, V, W, Ti, Nb, and Zr. With increasing levels of these elements, i.e., with decreasing Fe content of the mixed carbides formed, the temperature range within which the steels are stable to the high-

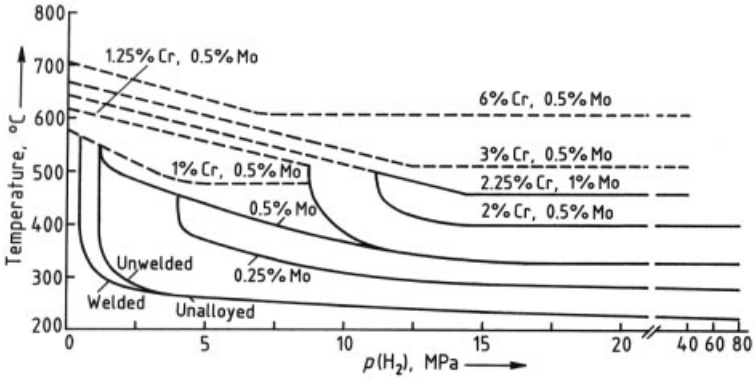

Figure 316. Limiting conditions for the stability of steels in high-pressure hydrogen (Nelson diagram)

pressure hydrogen is shifted upward. The critical temperature for the various steels has been determined by comprehensive and very costly autoclave experiments with H_2 [666]. A summary of operating experience of several decades was first produced by Nelson [667] in the form of a resistance diagram (Fig. 316), which is continually updated, and gives information for choosing suitable materials for known operating temperatures and hydrogen partial pressures [668]. The continuous lines in the diagram show the limit of resistance for the various steel compositions, i.e., for which higher temperatures or pressures would lead to internal damage by the high-pressure hydrogen. For the dashed lines, internal damage does not occur, though the use of these steels is limited by surface decarburization.

As with high-temperature steels, the hydrogen-resistant steels are also tested at high temperature by hot tensile measurements and by creep tests to establish the data used in design calculations. The scaling resistance in air or sulfur- containing gases, depending on the application, can also be important in the choice of steel.

Grades of Steel. With wall temperatures 200 – 400 °C, steels containing 0.5 – 1 % Mo and < 3.3 % Cr are used in hydrogenation plants. For resistance to higher hydrogen pressures, steels containing 5 – 9 % Cr and 0.5 – 1 % Mo are available, these being mainly used in oil refineries. If high values of creep rupture strength are required in the range 550 – 600 °C, 12 % Cr steel containing Mo and V, a material commonly used in the construction of power stations, can be used. Steels with higher chromium content also have better resistance to corrosion by sulfur-containing substances.

If the hydrogen contains N_2 and NH_3, as in the high-pressure synthesis of ammonia, nitridation of the steel at > 400 °C is likely. This produces hard, brittle nitride layers on the surface, and these readily spall off, reducing the cross section of the component. Ferritic steels cannot be used under the conditions of methanol synthesis, as the CO in the gas forms carbonyls at 120 – 300 °C, leading to the removal of material; the steel must be clad with austenitic Cr – Ni steel or copper to minimize material loss.

Heat treatment of these steels should ensure the precipitation of stable carbides to optimize stability to high-pressure hydrogen, i.e., high tempering temperatures and long soaking times. High strengths should be avoided. These procedures should also be used when heat treating welded components. Formerly, austenitic welds without subsequent heat treatment led to massive hydrogen damage in the weld-affected zones. Rapid cooling of the hot weld caused the formation of unstable higher Fe- containing carbides. This kind of damage does not occur if the weld material is identical to the material of construction, enabling subsequent heat treatment to be carried out. The best high-pressure hydrogen resistance is secured if the welded component is re-heat-treated.

Components made from these grades of steel are often very thick-walled, as operating pressures are very high (up to 100 MPa), so it is important to achieve uniform through-hardening

with a uniform microstructure and a high degree of purity. This maximizes resistance to hydrogen embrittlement.

Austenitic Cr – Ni steels have very high resistance to hydrogen embrittlement because of their high chromium content and the possible presence of other carbide-forming alloying elements, e.g., Mo, V, Ti, and Nb. They also have good corrosion resistance to other materials in the reaction gas, such as H_2S at 650 °C. In NH_3-containing gases, a thin, dense nitride protective layer is formed on austenitic Cr – Ni steels [669].

Standards. Specifications for steels resistant to hydrogen embrittlement are given in the German Stahl-Eisen-Werkstoffblatt 590, and for cast steel resistant to hydrogen embrittelement in Stahl-Eisen-Werkstoffblatt 595.

United States standards inlcude: ASTM A 200 for seamless tubes of medium-alloy steels for refineries; ASTM A 213 for seamless tubes of ferritic and austenitic steels; ASTM A 335 for seamless tubes of ferritic steels for high-temperature use; and ASTM A 336 for forgings of ferritic and austenitic steels for pressure vessels.

6.2.2.3. Steels with Special Physical Properties

In steels that have been developed simply so that certain physical properties can be utilized, mechanical properties are frequently unimportant. This does not apply to any other classes of steel. The soft magnetic steels are by far the most important members of this special group, not only from the tonnage aspect. Nonmagnetizable steels and steels with good electrical conductivity also play an important part in the economy. In steels with defined thermal expansion and special elasticity characteristics, the iron content is only ca. 50 %, exceptionally reaching 70 or 80 % [670]. In permanent magnets, the iron content is important, but most of these materials are not strictly steel products [671].

The highly developed techniques for measuring all the important physical properties cannot be treated within the scope of this article.

6.2.2.3.1. Soft Magnetic Steels [672]

(→ Magnetic Materials, Chap. 5.1.)

General. Traditionally, all materials that have coercivities of less than 1000 A/m are classified as soft magnetic. The International Electrical Commission (IEC) has proposed the following classification of soft magnetic materials:

A	pure iron
B	low- carbon steel
C	silicon steel
D	other steels
E	nickel – iron alloy
F	cobalt – iron alloy
G	other alloys
H	soft magnetic ceramics

Materials B – D are described below.

Low-Carbon Steel/Soft Iron. The term soft iron describes steels produced economically in normal steel production processes, containing a minimum of impurities, and only small additions of certain elements. Soft iron is produced in almost all product shapes for applications which require the production, amplification, or shielding of constant magnetic fields.

The various applications of soft iron require the following properties: ease of magnetization to high values, high saturation polarization, high maximum permeability, and low coercivity. The magnetic properties must change only insignificantly in the course of the working life of the component. If low core (power) losses are also required, the eddy currents associated with the good electrical conductivity must be taken into account.

For good soft magnetic properties, it is important that the iron should have an undisturbed ferritic microstructure and a large grain size. Precipitates caused by carbon or nitrogen and inclusions caused by oxygen and sulfur disturb the microstructure. Levels of these elements must therefore be minimized. The size of the precipitated particles becomes important if their diameter is comparable with the diameter of a Bloch wall (Fig. 317) [673]. The chemical bonding of the oxygen (deoxidation), sulfur, and nitrogen is carried out by adding small amounts of manganese and aluminum, which do not harm the magnetic properties. It is especially important to react any residual nitrogen with aluminum to prevent magnetic aging. As the grain size also

has a great influence on the coercivity (Fig. 318) [674], steps must be taken during further processing and heat treatment to produce a coarse grain.

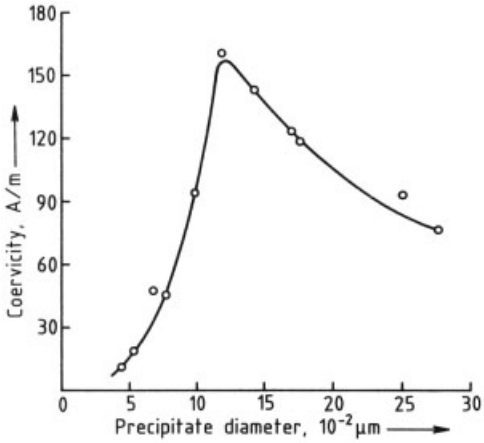

Figure 317. Coercivity as a function of average diameter of the precipitates in iron with 0.02 % C at constant volume fraction of the precipitates [673]

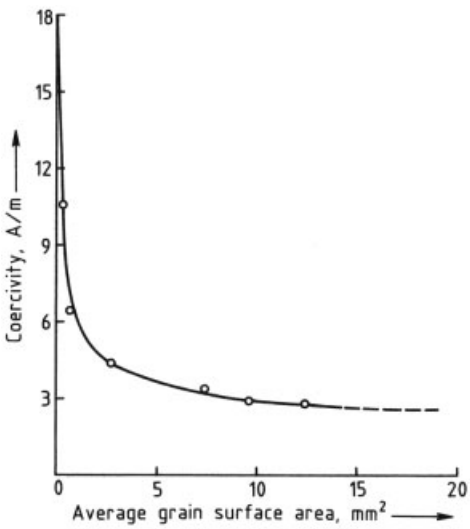

Figure 318. Coercivity as a function of average surface area of the grains for "pure" iron with 0.02 % C and 0.07 % Si [674]

An important application for soft iron is in the production of a wide variety of relays. Cold- or hot-rolled strip, merchant bar, wire, forgings, and castings can be used for this purpose. Heavy forgings, e.g., for accelerator magnets (up to 100 t) in large-scale research installations, are also produced in soft iron.

Electric Sheets. Electric sheets are mostly thinner than 0.65 mm, and are unalloyed or contain up to 4 % silicon. They are used in the construction of magnetic cores consisting of fine laminations in electrical equipment, e.g., motors, generators, transformers.

Electric sheets are produced almost entirely as cold-rolled strip manufactured on a large scale, and are usually coated on one or both sides with a thin layer of insulating material. Electric sheets must be easily magnetizable, and must have a low power loss on cyclic magnetization. Other requirements for further processing include good punching properties, a high specification for flatness, and a narrow thickness tolerance.

For small machines, unalloyed steel with low polarization is often sufficient, as the power losses and heat generation due to eddy currents in the core do not cause problems. At higher powers, the problems intensify, and the power loss becomes very important. Increasing silicon content increases electrical resistance and decreases power loss. Another approach to the problem is to reduce the thickness of the sheet.

The economic pressure for reduction in power losses led, in the 1940s, to the industrial production of the so- called Goss texture (cube on edge texture) [675], in cold-rolled electric sheets. In the Goss texture, the crystals in the microstructure are oriented such that the more easily magnetizable edge of the cube ⟨100⟩ is in the rolling direction, and the normal to the rolling direction in the plane of the sheet coincides with a face diagonal ⟨110⟩ of the cubic unit cell (Fig. 319) [676].

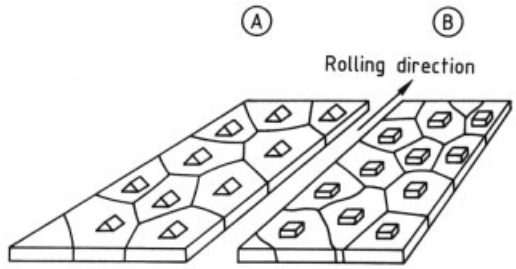

Figure 319. Positions of unit cubes in the crystal structure [676]
A) Goss texture; B) Cube texture

This development of the first grain-oriented electric sheet, with thickness 0.35 mm, resulted in greatly reduced power losses in transformers [677]. In rotary machines, this type of sheet is of limited application, and nonoriented electric sheets are nearly always used.

The silicon content is limited to ca. 3.5 % for the benefit of the cold rollability. In nonoriented electric sheets, an aluminum content of a few tenths of a percent completes the effect of the silicon. In grain-oriented electric sheets, this effect cannot be utilized, as the conditions for producing the texture require very close control of the chemical composition of the steel with respect to all elements.

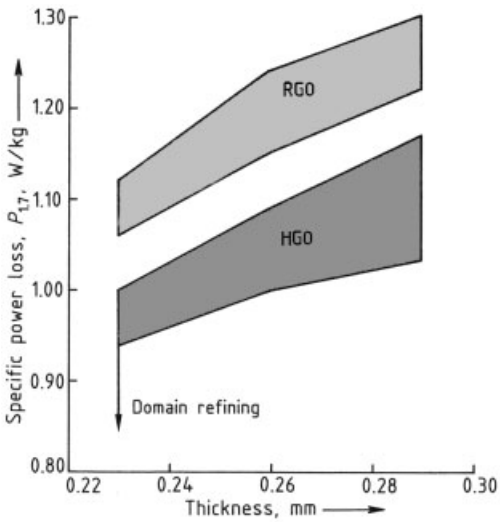

Figure 320. Relationship between specific power loss and sheet thickness [679]
RGO = regular grain oriented; HGO = high permeability grain oriented

Attempts to develop grain-oriented electric sheets with cube texture did not led to satisfactory and economically justifiable results. Further developments of regular grain oriented (RGO) electric sheet with Goss texture in the last 20 years have led to major improvements in the energy efficiency of power transformers. A first step was reduction of the sheet thickness to 0.30 mm and later to 0.27 mm. The next step was the development of high permability grain-oriented (HGO) silicon steel by changing the chemical composition of the steel to improve the degree of orientation and by using an insulating coating (see Fig. 320) [678], [679]. Further decreases in transformer power losses have been achieved by laser treatment of the sheet surface to refine the domains [680], and a further reduction of the sheet thickness to 0.23 mm. However the costs of the surface treatment and of producing the 0.23-mm product are high. Further developments on nonoriented and on grain-oriented electric steel sheet promise further energy savings.

6.2.2.3.2. Other Steels with Special Physical Properties

Steels with soft magnetic properties are also required to operate in situations where mechanical stresses are also present, e.g., generator shafts or steel sheet in laminated pole pieces. The properties of these steels, which depend on their chemical composition and microstructure, represent a compromise between opposing requirements.

Nonmagnetizable Steels [681]. These have a stable austenitic structure, i.e., there must be no $\gamma - \alpha$ transformation or martensite formation, even at low temperatures or after cold forming [682]. These steels are used to prevent harmful interactions with the magnetic field of the environment, e.g., in inductor cover rings on generator shafts, in the structures of ships in the vicinity of the compass, in equipment for low-temperature physics research, and in many other areas. These steels are often required to have good mechanical and anticorrosion properties.

The addition of a high proportion of manganese (ca. 18 %) is the simplest method, if the steel is not to be used appreciably below room temperature [683]. However, in most cases, high-alloy chromium and Cr – Ni steels are necessary. These often contain added nitrogen (ca. 0.5 %) to ensure that specifications are met.

Steels with Good Electrical Conductivity [684]. These are required for the conductor rails of electric railways, for telephone and telegraph wires, and for other current-carrying applications in high- and low-current technology. They are used for reasons of strength and sometimes cost in preference to materials with higher conductivity, e.g., copper and aluminum.

It is important to maintain low levels of impurity elements, including carbon and silicon

[685]. This must be borne in mind when selecting the raw materials to be used in steel production.

6.2.2.4. Steels with Special Processing Properties

When choosing a material for the manufacture of a particular component, the properties of the material that are important for the use of the component, the price, and the costs for manufacturing the finished component from this material are decisive. Most of the properties of steel as a construction material, and the extent to which it is possible to match these to the requirements, have been described in detail in previous sections. In the following section, groups of steel grades are described for which the mechanical properties are not the only properties of prime importance. They have been extensively developed to facilitate the manufacture of components by cold forming or machining, and to make these processes more convenient and cost effective.

6.2.2.4.1. Steels for Flat Products for Cold Forming [686]

General. Thin flat products, up to ca. 16 mm thick, especially suitable for cold forming, represent by far the greatest proportion of rolled steel products. These are mainly manufactured in the form of hot- or cold-rolled strip, with or without metallic or organic coatings (see Section 6.2.1 and EN 10 079), but also to a small extent as plate. Improvements in the grades of steel with good cold forming properties in these product shapes have led to great improvements in design, production, and manufacturing costs, especially in vehicle construction. These developments include not only mild steels, but also steels with tensile strength of up to 600 N/mm^2, the latter being associated with developments in forming techniques and die design (Fig. 321) [687]. In the interests of the style and safety of the autobody, complicated shapes can be produced from steel sheets without problems. Weight savings can be achieved by using high-strength grades of steel sheet with good formability [688]. Cold-formed parts made from thin flat products, welded to components have replaced forgings, e.g., for suspension systems, and in commercial vehicle construction.

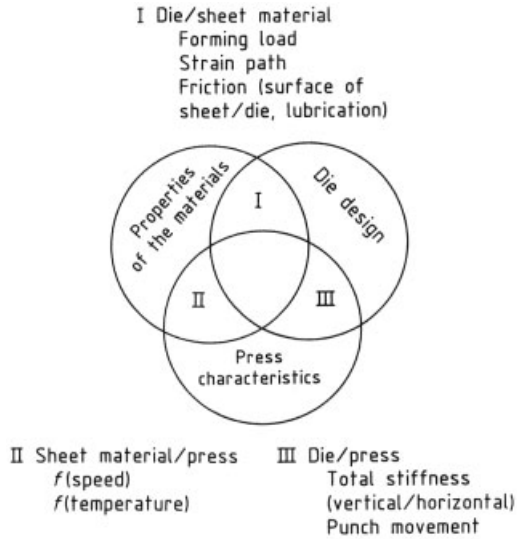

Figure 321. Factors influencing the result of a forming process

Corrosion problems in the automotive industry have practically been eliminated by galvanized steel sheet with good cold formability, following the development of suitable painting and welding processes for these sheets [689].

The cold forming properties of steel depend on its microstructure. However, the cold forming of flat products is very different from cold extrusion and cold heading of massive products. In bending and folding operations yielding of the material is easy to control; tearing must be avoided. In deep drawing and stretch forming stress conditions are considerably more complex. Actual stampings usually exhibit a combination of deep drawing and stretch forming.

Properties and Their Measurement. For deep drawing steel sheet, the most important properties for quality control are those determined from tensile tests, i.e., yield strength, tensile strength, and percentage elongation after fracture. In contrast to all other steel groups, the maximum allowable values of yield strength and tensile strength are important in this case. Whether minimum values should be quoted in the specification, as a basis for calculation by the designer, is in dispute, because cold form-

ing leads to considerable work hardening in the finished stamping.

In the last 20 years, two parameters have become of increasing interest, both determined by tensile tests: the normal anisotropic or plastic strain ratio (r value) and the work hardening exponent (n value) [690], [691]. The r value is the ratio of the change in width to the change in thickness of a tensile test specimen deformed by stretching, and is an expression of the relative tendency of the metal to yield in the width direction ($r \gg 1$) or in the thickness direction ($r<1$) [692]. If during deep drawing, the material becomes thinner, the load-bearing cross section is reduced, and the drawing stress that can be applied without tearing is limited. The r values for deep drawing should be as high as possible.

Different r values in the sheet plane at various angles to the rolling direction (i.e., the planar anisotropy Δr) are responsible for the occurrence of "earing" during pure deep drawing. The best conditions for avoiding this are provided when $\Delta r \approx 0$. The anisotropic properties of the steel sheet are a result of the hot rolling conditions and the annealing practice of the cold-rolled strip, producing the desired texture in the microstructure.

The work hardening exponent (n value) is determined from the true stress – true strain curve, and is given by the slope of this curve in log – log coordinates. The n value indicates the degree to which the material can withstand a higher stress after a given deformation than was possible in the starting condition. This work hardening of the deformation zone during stretch forming has the effect that the adjacent zone of even lower strength is included in the stretching process and necking is avoided. The r and n values are useful for the further development of steel grades, and can also be an aid in deformation analysis of difficult new deep-drawn items, using the grid measuring technique [693].

Tests that simulate forming operations, e.g., the cupping test for deep drawing or the bulge test for stretch forming, are now of little importance [694], but procedures such as the bending or double folding test (handkerchief test) are widely used for on-line monitoring. The adhesion and formability of corrosion-resistant coatings are tested by the ball impact test, or by forming profiles with various degrees of edge sharpness (lock forming test).

The surface condition of the steel sheet (produced by hot rolling, cold rolling, or surface coating) and the surface finish (smooth, matt, or rough) are of great importance for the forming process. To meet the customer's requirements, the cold-rolled sheet can be given a defined surface roughness, which is measured by contact stylus instruments.

Steel Grades and Product Shapes. The cold formability of flat products depends on the microstructure, which is determined by the chemi-cal composition, the rolling technique, and the heat treatment. A discussion of steel grades must therefore be linked with the shape of the product. The most important steel product used is coldrolled sheet, with or without coatings. Today, this is only produced from hot-rolled wide strip, nearly all produced from continuously cast material. With very few exceptions, only fully killed steel (by Al) is used. Up to 20 years ago, rimming steel was widely used in sheet production, as a defect-free surface and a higher yield was more easily achieved with this steel type at ingot casting (see Section 3.5.2). Rimming steel is still used if there is a shortage of strand casting capacity.

Unalloyed mild steels, with carbon content < 0.1 %, have the best cold formability as cold-rolled sheet. The specifications in EN 10 130 describe a series of five sheet grades with graded cold formability, improving as the tolerances in chemical composition are tightened and the specifications of the properties are increased. The fifth grade (Fe P 06), which has the highest cold formability, is a so-called IF steel (interstitial-free), with a very low carbon content (<0.02 %), in which the carbon and nitrogen are all combined with Al and Ti (or Nb) [692]. This steel sheet grade shows no yield point elongation in the stress strain curve after annealing. With the other sheet grades, the yield point elongation is removed by a final skin pass rolling (temper pass), to prevent the formation of Lüders lines in the surface after cold forming.

The same grades of steel are used for hot-rolled sheet and strip for direct use (EN 10 111) and for surface-coated cold-rolled strip (e.g., galvanized steel sheet, EN 10 142) [695]. The microstructure produced by hot strip rolling does not give quite the same cold formability. The properties of coated cold-rolled strip is influ-

enced by various heat treatments (see Chapter 5), but cold formability is practically equivalent to cold-rolled strip. Attention must be paid to the effects of the various coatings on the friction behavior in the drawing gap.

To save weight in vehicle construction and in other areas, e.g., in the building industry, grades of steel have been developed for higher-strength thin flat products which still have good cold forming properties.

Cold-rolled sheet, strip, and thin plate made of general-purpose structural steels with high strength due to the carbon content have poor cold forming properties because of the pearlite in the microstructure. The technique of increasing yield strength by cold rerolling, which increases the dislocation density, impairs the cold forming properties to an unacceptable extent. The introduction of microalloying with Nb, Ti, and V in the early 1960s, giving precipitation-hardened products with greatly reduced carbon content, led to the development of a wide range of high-strength grades of steel with good cold forming properties.

These grades of steel have only small differences in their chemical composition, and are used for cold- and hot-rolled flat products (EN 10 149). The hot-rolled flat products are usually produced by a thermomechanical process, which can be followed by normalizing if required [696]. These steels are now referred to as high-strength low-alloy (HSLA) or low-pearlite steels, and are widely used, especially in the automotive industry [697].

Other methods of improving strength properties without appreciable affecting the cold formability include solid solution hardening, e.g., by increasing the phosphorus content, and the so- called dual-phase steels, in which a pure ferritic matrix structure containing 20 – 30 % dispersed martensite islands is formed [698].

The so- called bake-hardening steels are of particular interest for automotive production [699]. These include cold-rolled uncoated or electrolytically coated sheet, which have low yield strengths and are age-resistant at room temperature as supplied. After press forming, followed by the paint-baking heat treatment (a controlled aging), the yield strength R_e increases from ca. 30 to 50 N/mm^2 (Fig. 322) [700].

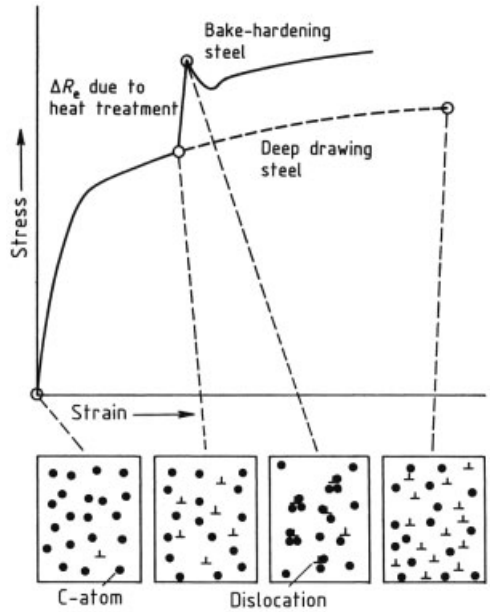

Figure 322. Mechanism of the increase in yield strength of a tensile test piece following deformation and heating (bake-hardening)

Figure 323 gives an overview of the various grades of steel for cold-rolled sheet, and includes the properties associated with cold formability and strength [701].

6.2.2.4.2. Steels Suitable for Cold Extrusion and Cold Heading [702]

General. In the cold forming of flat products, the starting material is pressed to form a hollow product or a profile, thereby increasing its surface area. However, in cold extrusion, the starting material, extending into three dimensions, is forced by pressure to yield in all directions (see Section 4.1). The drawing of wire, rod, and tubes also falls under this classification.

Steel can be very suitable for plastic deformation in which the degree of deformation is high, if its microstructure fulfills the necessary requirements. Unalloyed and alloyed grades of steel with up to ca. 1.6 % C are suitable for cold extrusion, as are high-alloy stainless steels.

The many advantages of cold extrusion, e.g., good dimensional control, are described in Section 4.1.3. The process causes work hardening, which is sometimes directly utilized in the finished product. In many applications, a final heat

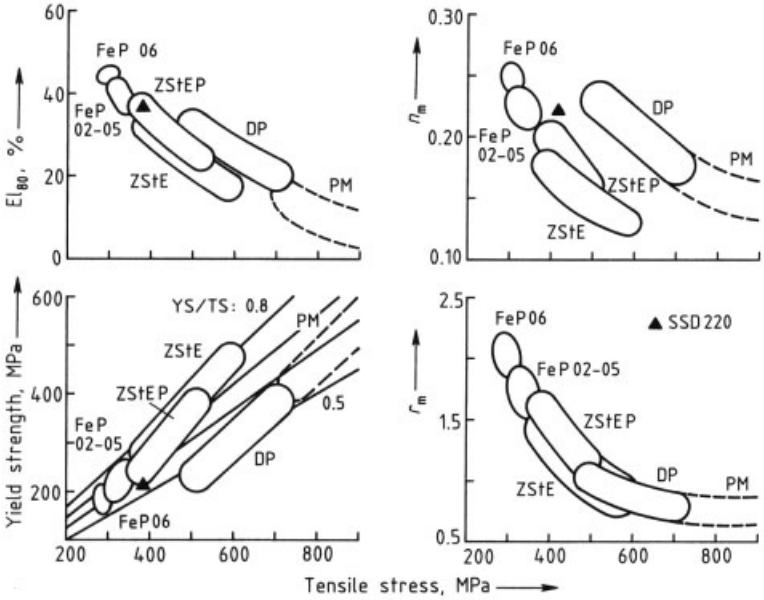

Figure 323. Mechanical properties of cold-rolled steel sheet grades
Fe P 02 – 05 = deep drawing grades acc. to EN 10 130; Fe P 06 = IF deep drawing grades acc. to EN 10 130; ZstE = HSLA – grades (microalloyed); ZSteP = high strength-grade with increased P content; DP = dual-phase-grades; PM = partially martensitic grade; SSD 220 = high Strength IF-grade (experimental).El_{80} = % elongation at fracture (gage length 80 mm)

treatment is included to give the desired structure and properties.

Properties and Their Measurement. For cold extrusion, the yield stress k_f of a material for a given degree of deformation φ is important. The relationship between these values is represented by the true stress – true strain curve (see Section 6.1.2.1). From the yield stress, the force required for the deformation and the material properties after forming can be estimated. The scatter bands for flow curves of some unalloyed and alloyed steels are shown in Figure 324 [703]. For a given chemical composition, these depend on the microstructure of the material. Understandably, ferrite has the best deformability, influenced by its chemical composition, while the cold deformability of coarsely laminated pearlite is limited. Therefore, steels containing more than ca. 0.15 % C must be heat treated, usually by annealing, to convert the cementite in the pearlite into spheroidal form. The degree of spheroidization and the size of the carbide particles can be used to estimate the formability. Very finely laminated pearlite, e.g., after quenching and tempering or wire patenting, is also very suitable.

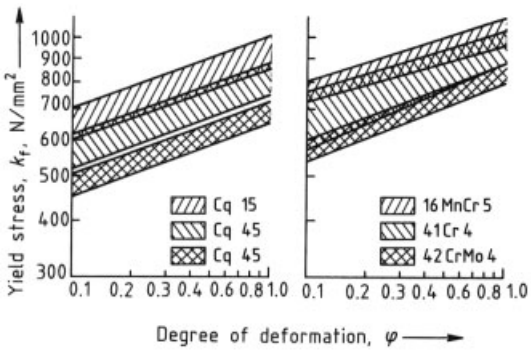

Figure 324. Scatter bands of flow curves of unalloyed and low-alloy steels after annealing to form spheroidal cementite [703]

The yield stress is usually determined by a cylindrical compression test. Here, the friction at the surfaces under compression is very important. In practice, the suitability for cold extrusion and the deformability are frequently assessed

simply from the tensile strength, the reduction in area on breaking, and the percentage elongation at fracture. It is essential that the surface of the starting material should be practically defect free for successful cold forming.

Grades of Steel. Cold extrusion and cold heading are used for the manufacture of screws, nuts, and many components used in automobile and machine manufacture, including gear wheels, shafts, piston pins, etc., as well as cold-drawn tubes and wires. The chemical compositions of the grades of steel for these applications must be adapted to cold deformability, i.e., Si and Mn content must be reduced, and undesirable tramp elements such as Cu, Ni, Mo, and P must be excluded by appropriate selection of the raw materials. Small additions of boron improve the hardenability of the steel without increasing the yield stress in the annealed state. A high degree of purity with respect to nonmetallic inclusions is also important, especially long banded inclusions should not be present.

Unalloyed mild steels not intended for heat treatment, and unalloyed and alloyed case hardening and heat-treatable steels with suitable chemical compositions, and after spheroidizing are used in large amounts for cold extrusion and cold heading.

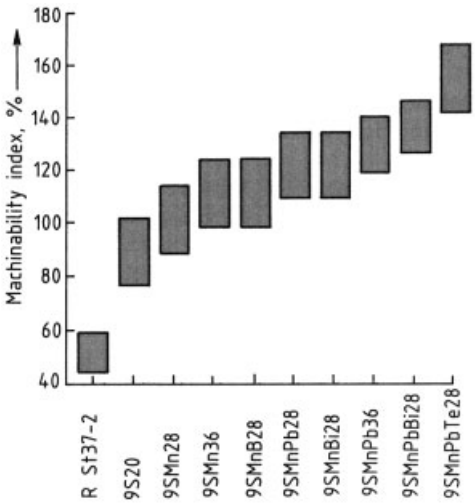

Figure 325. Machinability of free-cutting steels
Machinability index is the machined volume for tool life 8 h, based on a value of 100 % for 9 SMn 28 steel.

6.2.2.4.3. Machining or Free-Cutting Steels [704]

General. The shaping of steel to produce finished products by machining is extremely important. If we exclude sheet stamping, almost all other articles made from steel are formed into their final shape by processes involving the removal of metal, e.g., milling, turning, drilling, screw cutting, sawing, to achieve the required tolerances and surface quality. The costs of these processes are greatly influenced by the machinability of the material (Fig. 325).

Good machining properties are of great importance, especially for the mass production of machinery components, the automotive industry, and also for the operation of computer-controlled automatic lathes or production lines. Good machinability means:

1) Good tool stability at high cutting rates
2) Good breaking properties of the turnings or chippings, so that these can be removed safely
3) Dimensionally accurate and smooth machined surfaces

To meet these requirements, the so-called machining or free-cutting steels were developed some time ago. The formation of short breaking turnings and a very smooth surface was achieved by the addition of sulfur, and, more recently, other elements (Pb, Te, Bi), with a slight increase in phosphorus content. The high levels of these elements can have a harmful effect on technical properties for certain applications, so that the machining steels represent a compromise, and this must be borne in mind when they are used.

Properties and Their Measurement. The use of machining steels is based on their mechanical properties. The distinction is made between low-carbon mild steels which are not intended for heat treatment, and steels which are suitable for case hardening or quenching and tempering (EN 10 087).

Machinability is difficult to quantify, and depends on the interaction of the material of the workpiece with the material of the cutting tool under the actual machining conditions [705]. A large number of test methods have been proposed. These include long-term tests to determine the one-hour cutting speed V_{60}, m/min,

i.e., the cutting speed under prescribed conditions which causes the tool to be worn out after 60 min, and also a quick test to determine the tool failure speed V_E, m/min, with an increased cutting speed. In addition, the shape of the cuttings and the surface finish are assessed.

The distribution and size of the sulfide precipitates (which are of complex composition) in the microstructure is of great importance for the machinability. This distribution is influenced by the metallurgical processes occurring during the production of the steel, and also by the degree of hot rolling. Elements such as tellurium or selenium promote the formation of globular sulfides. Metallographic observations of the microstructure are therefore used in controlling the working method.

Grades of Steel. Low-carbon mild free-cutting steels, not intended for heat treatment, have strength properties like those of Fe 360, and their chemical composition is graded to give varying degrees of machinability. Sulfur content (up to ca. 0.4 %) is approximately ten times that of other steels. The formation of iron sulfide, which leads to "red shortness" on hot forming, is avoided if the Mn content is ca. 1.5 % higher, while the Si content is kept as low as possible (< 0.05 %). Additions of lead (up to 0.1 %) and tellurium improve machinability, but are expensive. Figure 325 shows the effect of various chemical compositions on machinability compared with Fe 360 [704].

With case hardening and heat treatable steels, which are optimized for machinability, the usual principles of alloying apply, but sulfur content is limited to 0.15 – 0.25 %. Machining steels suitable for quenching and tempering are only supplied unalloyed, with carbon content < 0.65 %.

7. Environmental Protection

7.1. Environmental Aspects of Steel Production and Processing

7.1.1. Production of Steel and Steel Products

Large quantities of raw additive and auxiliary materials are used in the production of crude steel and rolled steel products, and large amounts of energy are used in these processes. Substances in the solid, liquid, and gaseous states are potential sources of environmental pollution.

For reasons of economy, the steel industry, especially in Europe, has always attempted to use as little primary energy as possible, to limit the consumption of auxiliary materials, e.g., water, by means of complex recycling systems, and to utilize or recycle byproducts or residues, e.g., slags and mill scale. From the early 1900s, integrated steel plants, in close association with coking plants and rolling mills, have optimized the use of CO-containing blast furnace top gas according to the current state of technology. The use of blast furnace slag for cement and for road construction is very much older than the general interest in environmental pollution. In the last 20 – 30 years, the ever-increasing severity of environmental controls have led to technological developments within the steel industry which have been very successful [706]. However, some of the regulations imposed demand additional capital investment and operating costs whose justification may be dubious, especially when in some countries the regulations imposed are different and, thereby, distort competition. The following discussion is based mainly on available data for the steel industry, and particularly, individual steel plants in the western part of the Federal Republic of Germany.

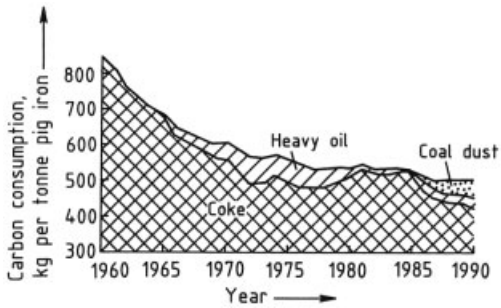

Figure 326. Carbon consumption in pig iron production

Environmental aspects of the production of pig iron in a blast furnace have already been described (→ Iron, Chap. 4.). Further numerical data on developments during the last 30 years are given here. Figure 326 shows the considerable decrease in the consumption of carbon for reduction purposes in pig iron production [706]; Figure 327 shows the continuous de-

crease in the specific amount of blast furnace slag, and its 100 % use in various applications [707]. The generation of electrical energy from expansion turbines running on blast furnace top gas (Fig. 328) is a further contribution to environmental protection [708].

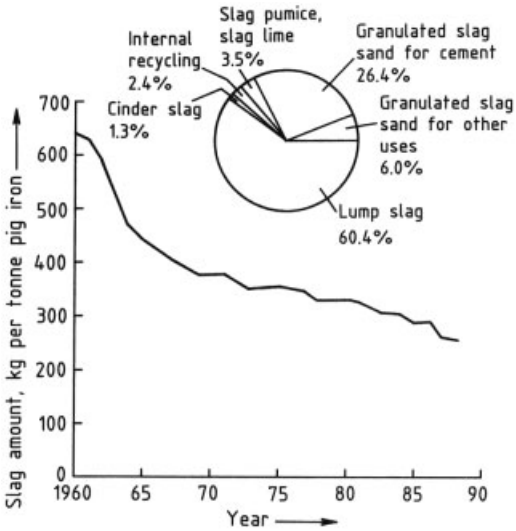

Figure 327. Development of the specific amount of blast furnace slag and its utilization

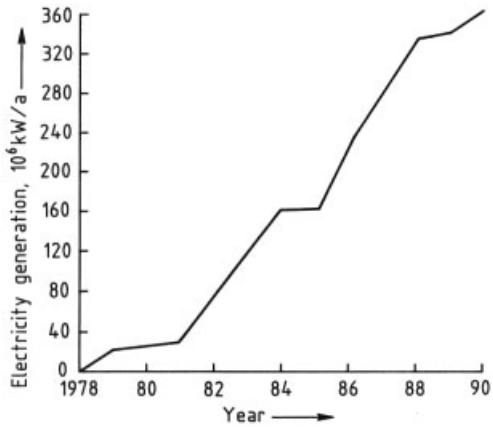

Figure 328. Electricity generation by expansion turbines from blast furnace top gas

In a steel melting shop, the amount of dust formed and the possibility of its recovery depend very much on the different steel melting processes, which have undergone many changes (see Section 3.3). The changeover to the oxygen blast process and improvements in dedusting equipment, including secondary dedusting, have led to an emission rate of only 0.7 kg per tonne crude steel in the melting shops of the German steel industry (Fig. 329) [709]. Secondary dedusting, comparable to the "cast house dedusting" at the blast furnace, involves the additional dedusting equipment at different operation points located before and after the converters, e.g., pig iron transfer pit, mixers, deslagging stands; the dedusting equipment consists of stationary or movable hoods and skirts. Figure 330 illustrates the complexity of this additional dust collection in a steel melting shop with three converters [710].

A comprehensive study has shown that the total quantity of dust and sludge collected in the steel plants in North Rhine/Westphalia is 41 kg per tonne crude steel (Fig. 331), of which more than 80 % is reused [708].

The hot waste gas formed in an oxygen converter is used to produce steam in a waste heat boiler in the primary dedusting installation [709], [710]. After further purification and cooling (Fig. 332), it is collected so that its high CO content (68 – 70 %) can be used sometimes after mixing with other gases. Figure 333 shows the significant energy generation which has been achieved in Germany [707].

Slag, an inevitable byproduct in steelworks has been further reduced, reaching 121 kg per tonne crude steel in 1991, of which 87.5 % was utilized [708].

The mill scale occurring in hot rolling mills is a pure iron oxide, and can be reused in sintering plants. However, oil-containing mill scale sludge must be pretreated. Rotary kiln installations to be used for this purpose are in the experimental stage [710].

Control of water consumption in a steel plant is also very important from the environmental point of view. All parts of the production operation involve the consumption of water, sometimes in very large quantities, e.g., for cooling the installations and the products. Attention has always been paid to the development and use of methods for saving water, e.g., the treatment and reuse of water in complex recirculating systems. In 1989, in a steelplant situated near the Rhine, only 3 % of the process water used was fresh water, i.e., 2.5 m^3 fresh water per tonne crude steel (Fig. 334) [706]; the remaining 97 % came

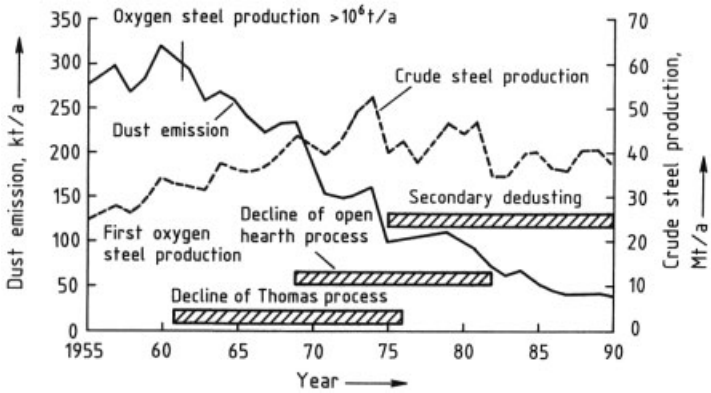

Figure 329. Development of crude steel production and dust emission by the German steel industry

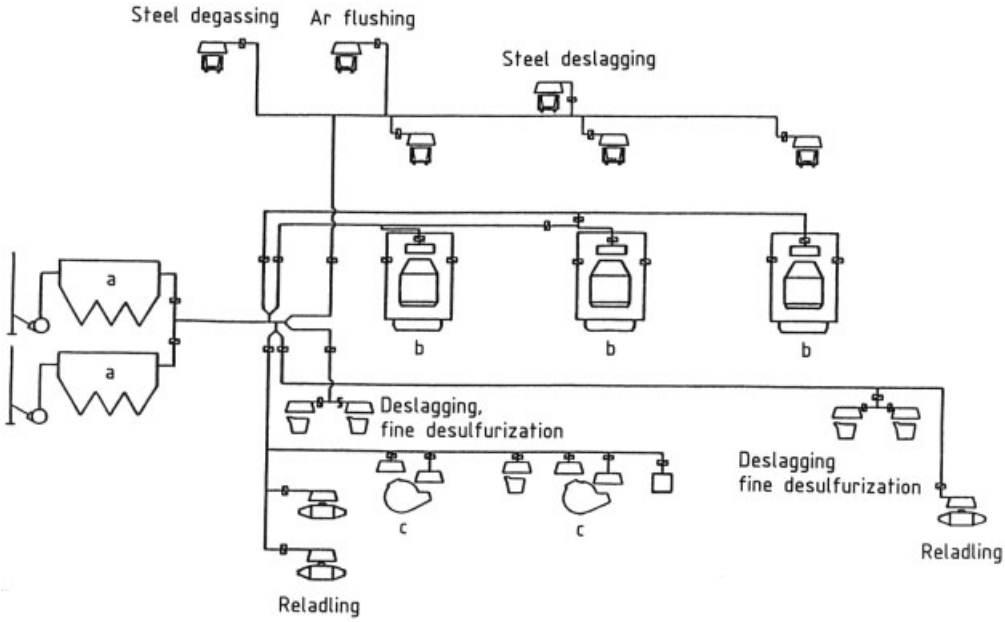

Figure 330. Dust collection in a three- converter steelshop
a) Electrostatic precipitators; b) Converters; c) Mixers

from the recirculation system. Some 83 % of the water used in the steel industry is surface water.

A high proportion of rolled steel production is cold-rolled or coated steel sheet, tinplate, black plate, and cold-rolled electric sheet. The hot-rolled strip for these finished products must be treated in continuous pickling lines to remove the adhering mill scale before the cold-rolling process (see Section 4.3). The used pickling solution is also treated in a recirculating system in which an iron salt, e.g., sulfate, crystallizes, and fresh acid is added, so that only a small fraction of the solution leaves the plant after neutralization. The iron hydroxide sludge, which contains gypsum, cannot be used any further. This also applies to the wash-water in the pickling line, but the amount of wash-water used can be much reduced by appropriate techniques.

7.1.2. Steel Processing and Steel in Use

Steel used in manufacturing as described in earlier chapters, is environmentally a very friendly material.

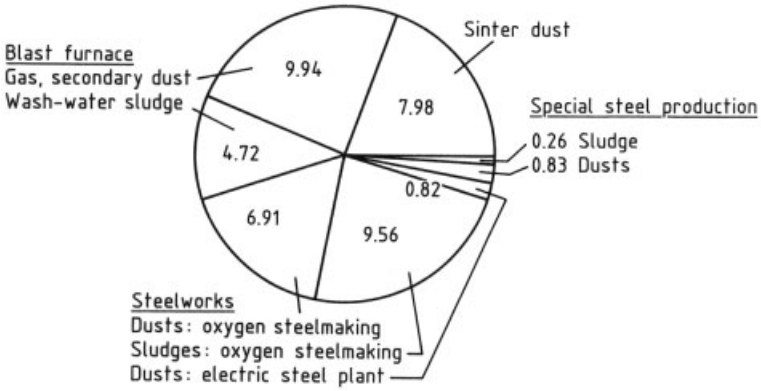

Figure 331. Origins of dusts and sludges in pig iron and steel production in North Rhine/Westphalia
Amounts in kg per tonne crude steel

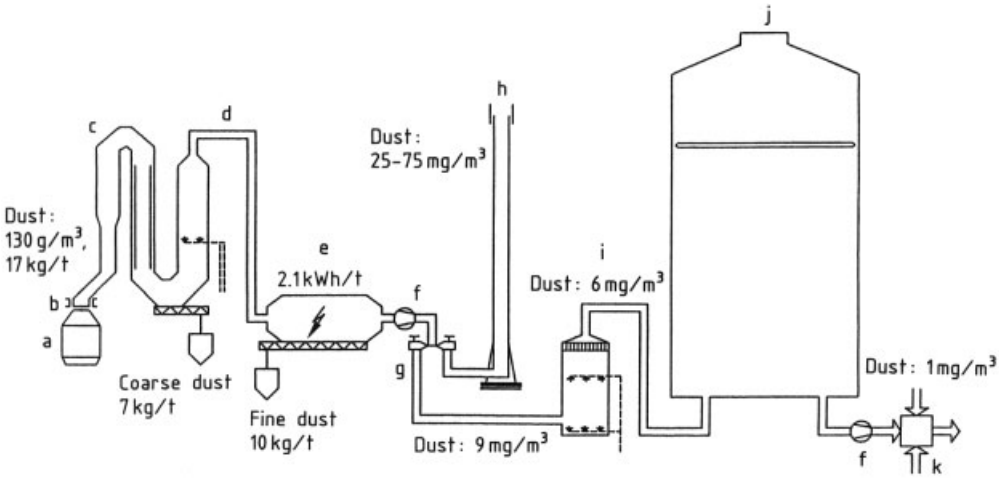

Figure 332. Converter gas cleaning and recovery system
a) Converter; b) Skirt; c) Steam boiler; d) Evaporation cooler; e) Dry electrostatic precipitator; f) Fan; g) Reversing valve; h) Flare stack; i) Gas cooler; j) Gasholder; k) Gas mixing station

The processing of steel can certainly involve sound emissions, e.g., during hammering, sheet metal working, and forging. Gases can result when flame cutting and welding steel. However, in both these cases, it is possible to protect personnel at the workplace without great cost and without causing too much inconvenience to the individual. In certain heat treatment processes, e.g., those using molten salt baths, the advice of the manufacturers of the salts or other substances used should be followed, and relevant legal requirements must be observed relating to their handling and disposal.

Environmental problems associated with the use of steel products are practically nonexistent, since steel is a highly recyclable material (see also Section 7.2). Even tinplate is environmentally a friendly packaging material, as it can be separated from waste materials or residues, e.g., from garbage incineration plants, by virtue of the ferromagnetic properties of the steel, and can then be reused.

Figure 335 illustrates the extent to which construction materials are recycled, and in the case of steel this can easily be increased further [711]. The reuse of steel scrap is of great importance for the overall economy. Approximately 60 % less primary energy is required for melting scrap for steel production, than for the smelting of ores [711].

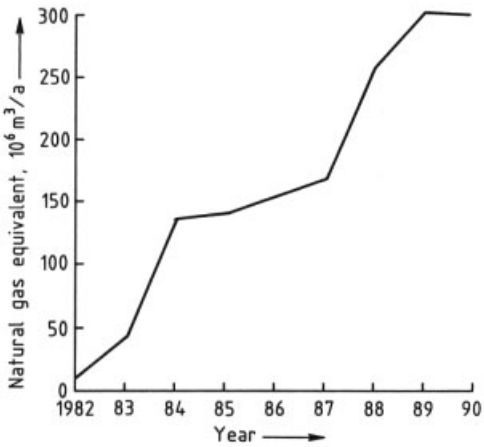

Figure 333. Energy production by the use of converter gas

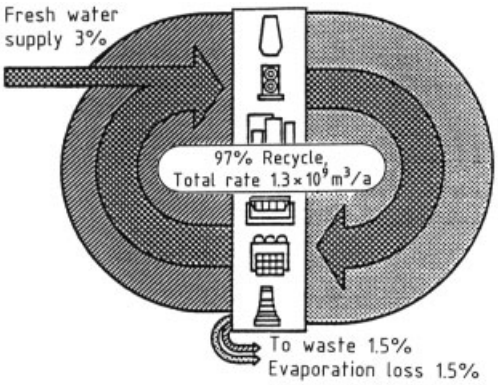

Figure 334. Internal water circulation

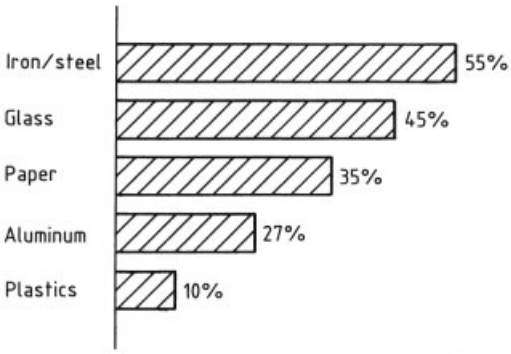

Figure 335. Extent of recycling of various materials

In the last 15 years, the use of sheet with a surface coating, mostly zinc, has increased considerably, both in automotive body manufacture and in other applications. Doubts about the use of zinc-coated steel scrap from these sources are unjustified. Most of the zinc vaporizes during crude steel production, and is recovered in the dust from the filters in the dedusting equipment. If manufacturers dispose of moderately zinc-rich scrap separately, this can be processed and used in suitable electric steel plants, and the zinc can be economically recycled. For this there must be at least 16 % Zn in the filter dust [712].

7.2. Steel Recycling

7.2.1. The Tradition of Steel Recycling

More than 3000 years ago, iron had already been the material foundation for cultural development. It has given its name to an entire epoch. The industrialization is closely linked with the development of the iron and steel industry. Almost all other industries have been directly or indirectly dependent on advances in steel production.

On 28 October 1865, the first iron was smelted by the Siemens – Martin open hearth process using 70 % scrap, so beginning the industrial utilization of scrap [713].

For over a hundred years, a technically and economically viable high-tonnage recycling system has existed for handling production waste and worn-out capital equipment and consumables made of steel and cast iron. For steel scrap, there is a completely closed system for recycling worn-out steel products, any number of times. The recycling process continually yields high quality steel grades without downgrading.

As industrialization progressed and steel production increased toward the 1890s, the demand for raw materials rose. By then, scrap was already an important substitute for the usual raw material in the iron- and steel-producing industry.

7.2.2. Types of Scrap

Listings. In order to provide precise descriptions of the physical and chemical properties required in reusable scrap, an agreed list of scrap grades has been prepared by the steel and steel recycling industry in Europe (Table 28).

Table 28. Extract from the European list of scrap grades (1 July 1995)

Category	Specification	Description	Dimensions	Density, t/m^3	Steriles [a]
Old scrap	E3	Old thick steel scrap, predominantly more than 6 mm thick in size not exceeding 1.5×0.5×0.5 m, prepared in a manner to ensure direct charging. May include tubes and hollow sections. Excludes vehicle body scrap and wheels from light vehicles.Must be free of rebars and merchant bars, free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	Thickness $\geq$ 6 mm$\leq$ 1.5×0.5×0.5 m	$\geq$ 0.6	$\leq$ 1 %
	E1	Old thin steel scrap, predominantly less than 6 mm thick in size not exceeding 1.5×0.5×0.5 m, prepared in a manner to ensure direct charging. If greater density is required it is recommended that maximum 1 m is specified. May include light vehicle wheels, but must exclude vehicle body scrap and domestic appliances.Must be free of rebars and merchant bars, free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	Thickness $<$ 6 mm$\leq$ 1.5×0.5×0.5 m	$\geq$ 0.5	$<$ 1.5 %
New scrap Low residuals, uncoated [b]	E2	Thick new production steel scrap predominantly more than 3 mm thick prepared in a manner to ensure direct charging. The steel scrap must be uncoated unless permitted by joint agreement and be free of rebars and merchant bars even from new production.Must be free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	Thickness $\geq$ 3 mm$\leq$ 1.5×0.5×0.5 m	$\geq$ 0.6	$<$ 0.3 %
	E8	Thin new production steel scrap predominantly less than 3 mm thick prepared in a manner to ensure direct charging. The steel scrap must be uncoated unless permitted by joint agreement and be free of unbound ribbons to avoid trouble when charging.Must be free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	Thickness $<$ 3 mm$\leq$ 1.5×0.5×0.5 m (except bound ribbons)	$\geq$ 0.4	$<$ 0.3 %
	E6	New production thin steel scrap (less than 3 mm thick) compressed or firmly baled in a manner to ensure direct charging. The steel scrap must be uncoated unless permitted by joint agreement.Must be free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.		$\geq$ 1	$<$ 0.3 %
Shredded	E40	Shredded steel scrap. Old steel scrap fragmentized into pieces not exceeding 200 mm in any direction for 95 % of the load. No piece in the remaining 5 % shall exceed 1000 mm. Should be prepared in a manner to ensure direct charging. The scrap shall be free of excessive moisture, loose cast iron and incinerator material (especially tin cans).Must be free of metallic copper, tin, lead (and alloys), and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.		$>$ 0.9	$<$ 0.4 %
Steel turnings [c]	E5H	Homogeneous lots of carbon steel turnings of known origin, free from excessive bushy. Should be prepared in a manner to ensure direct charging. Turnings from Free Turning Steel must be clearly identified. The turnings must be free from all contaminants such as nonferrous metals, scale, grinding dust, and heavily oxidized turnings or other materials from chemical industries. Prior chemical analysis could be required.			[f]

Table 28. (Continued.)

Category	Specification	Description	Dimensions	Density, t/m^3	Steriles [a]
	E5M	Mixed lots of carbon steel turnings, free from excessive bushy and free from turnings Free Cutting Steel. Should be prepared in a manner to ensure direct charging. The turnings must be free from all contaminants such as nonferrous metals, scale, grinding dust and heavily oxidized turnings or other materials form chemical industries.			[f]
High residual scrap	EHRB [d]	Old and new steel scrap consisting mainly of rebars and merchant bars prepared in a manner to ensure direct charging.May be cut, sheared or baled and must be free of excessive concrete or other construction material.Must be free of metallic copper, tin, lead (and alloys), mechanical pieces and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	max. 1.5×0.5×0.5 m	≥ 0.5	< 1.5 %
	EHBM [e]	Old and new mechanical pieces and components not accepted in the other grades prepared in a manner to ensure direct charging.May include cast iron pieces (mainly the housing of mechanical components).Must be free of metallic copper, tin, lead (and alloys) and pieces such as bearing shells, bronze rings and others as well as steriles, to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.	max. 1.5×0.5×0.5 m	≥ 0.6	< 0.7 %
Fragmentized scrap from incineration	E46	Fragmentized incinerator scrap. Loose steel scrap processed through an incinerating plant for household waste followed by magnetic separation, fragmentized into pieces not exceeding 200 mm in any direction and consisting partly of tin coated steel cans. Should be prepared in a manner to ensure direct charging. The scrap shall be free of excessive moisture and rust.Must be free of excessive metallic copper, tin, lead (and alloys) and steriles to meet the aimed analytical contents.Refer to points B) and C) of the general conditions.		≥ 0.8	Fe content ≥ 92 %

[a] Weight percentage of steriles not adhering to the scrap that are left in the vehicle after unloading by means of magnetic separation.[b] Coated material must be declared.[c] Free from contaminants (nonferrous metals, scale, grinding dust, chemicals, excess oil).[d] Rebar and merchant bar must be calssified separately due essentially to the copper content which could place them in a category with old scrap and new scrap of low residual grades.[e] Mechanical and engine components must be classified separately principally due to their Ni, Cr, and Mo content which could place them in a category with thick old scrap and heavy new scrap of low residual grades.[f] To date, there is no clear method to determine these values.

These defined grades apply only to carbon steel scrap. The following general requirements are included in the European list of scrap grades of 1st July, 1995. The scrap must be:

1) Prepared in a manner to ensure direct charging
2) Handleable magnetically
3) Free from elements harmful to the smelting process

Plant Scrap. Plant scrap is produced in the manufacture of steel and of rolled finished products. It is directly reused in the steelworks.

New Scrap. New scrap arises in steel processing industries as production waste. It is usually returned to steelworks and foundries by the recycling industry.

Capital Scrap. Capital scrap is composed of steel, cast steel, and cast iron originating from used commodity and industrial goods. There are three types, classified according to origin:

1) Demolition scrap
2) Automobile scrap
3) Collected scrap

The development of the amounts of plant scrap, new scrap, and capital scrap is illustrated in Figure 336, with estimated future figures, assuming that 70 % of steel consumption is returned to the material cycle after a period of 20 years.

7.2.3. Scrap Processing

7.2.3.1. Separation

Steel is used in engineering, plant construction, vehicles, building industry, shipbuilding, household appliances, packaging, etc. When products that consist mainly of steel are no longer usable, they are usually processed by the steel recycling industry. The composition and dimensions of scrap are usually such that it cannot be used in steelworks and foundries in its original state. Appropriate processes must therefore be used for:

1) Size reduction
2) Separation from other materials
3) Classification into grades

A unique characteristic of steel and iron scrap is that it can easily be recovered from accompanying materials by its magnetic properties, enabling the iron- containing constituents to be returned to steelworks and foundries in a high state of purity.

7.2.3.2. Shearing Machines

Large pieces of scrap can be size-reduced by shearing machines. Even before the World War I guillotine shears were used for this purpose.

Further developments led to the larger and faster alligator shears, and then to the hydraulic machines used today, where the scrap is charged by a crane into a feeding bed. The feeding bed is usually provided with a prepress which compresses the scrap so that even bulky material can be pushed under the blades of the shears. Size reduction of, e.g., freight cars industrial equipment, and large components, including material from shipbreaking and other heavy demolition work, is carried out with shears with a compression force of up to 2000 t and blades up to 2.50 m wide [714].

7.2.3.3. Presses

Scrap presses are used to densify clean steel scrap, e.g., production waste from rolling mills and steel processing industries, to facilitate transportation and charging, and to reduce it to the specified dimensions.

In the early 1920s, slow mechanical screw presses with large vertical boxes and small charging holes were developed. These were able to bale light collected scrap and new sheet scrap.

Mechanical baling presses were followed by high-pressure hydraulic presses with large charging boxes. This hydraulic equipment produces high-density bales of the size needed for steel production plants, e.g., converters and electric furnaces, and for the melting facilities of the foundry industry. The usual bale sizes are: $30 \times 30 \times 30$ cm and $40 \times 40 \times 60$ cm for the foundry industry, and $60 \times 60 \times 150$ cm for the steel industry [715].

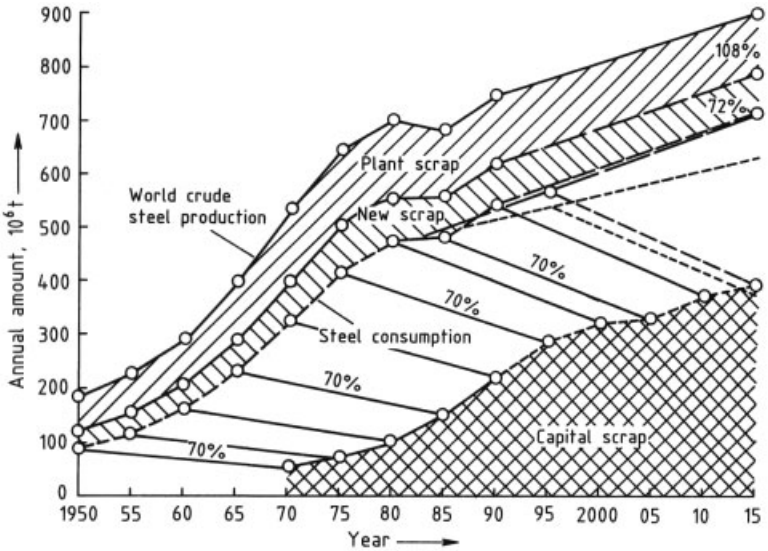

Figure 336. Development of arisings of plant scrap, new, and capital scrap, with estimated future figures

7.2.3.4. Shredders

Increasingly, scrap is no longer in pure form, but combined with other materials. Examples include old vehicles and household appliances, which contain considerable amounts of plastics, glass, rubber, nonferrous metals, and other materials. Shredders (Fig. 337) have been used by the steel recycling industry since the mid 1960s to produce automobile and collected scrap that meets the requirements of the steelworks.

A shredder functions on the principle of the swing-hammer mill. The products fed to the shredder are broken down to pieces not larger than fist-sized, and are then fed to an air separator to remove the light waste fraction, which includes plastics, textiles, and cardboard. An electromagnetic drum or band separator separates the steel scrap from the nonferrous metals and coarse impurities (heavy waste fraction). The materials removed by the air separator, which cannot, at present, be recycled (shredder residues) must be disposed of. This is expensive, and is likely to become more so. Shredding provides a means of recovering the iron and steel from used products with purity ca. 98 %, so that they can be returned to the raw material cycle.

7.2.4. Factors Influencing Recycling

7.2.4.1. Alloy Steels

In steel production, the properties of various grades of steel are achieved by adding alloying elements. Alloy steel can also be returned to the material cycle, if it is sorted according to chemical composition. The alloying elements are retained when the scrap is melted, and new alloy steels can be produced from this metal.

7.2.4.2. Surface Coatings

Steel is surface coated to give the properties required, i.e., surface quality, stress condition, and corrosion protection. Surface coatings can be either metallic (e.g., zinc) or nonmetallic. The latter include organic coatings (plastic or paints) and inorganic coatings (cement mortar or enamel) [716].

Smelting of zinc-coated steel scrap during steel production is the present state of the art. As zinc boils at 907 °C and steel melts at 1400 – 1500 °C, the zinc vaporizes from the melt and is collected as zinc oxide in dedusting plants. If the zinc content of the steelworks dust is ca. 16 %, it is profitable to recover the zinc [717], [718].

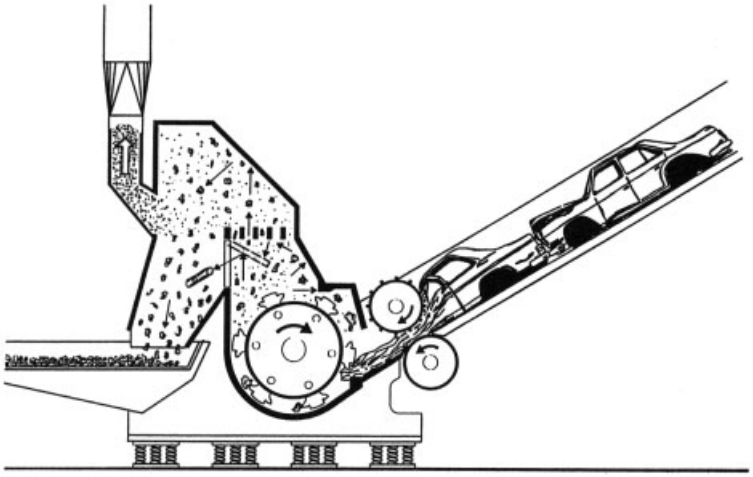

Figure 337. Operating principle of a shredder

Before scrap recovered from steel with a nonmetallic coating can be returned to steel production, it is usually processed by shredding, or cold shock treatment to remove adhering materials from the scrap, as these are harmful to the smelting process, or can lead to inadmissible emissions.

7.2.4.3. Pretreatment of Used Steel Products

Steel products and components at the end of their lives are often associated with impurities, or are attached to other materials. To prevent the unwanted materials from being returned to steel production along with the scrap, pretreatment and purification are necessary, e.g., removal of oil from chips arising from the processing of steel, which are classified as new scrap.

Pretreatment of used products that consist mainly of steel is often also necessary for disposing of or recycling the other materials in the products, e.g., the disposal of old cars. In the initial dismantling, the batteries of the old vehicles are removed to return the lead to a separate material cycle. This also prevents the lead from contaminating waste tips when the shredder residues are dumped. For the same reason, fuel, oil, gear, differential, and shock absorbers, and brake fluid and coolant are recovered before the old cars are shredded. In refrigerators, freezers, and air conditioning equipment, the cooling liquids are recovered to prevent their escape into the atmosphere.

Tinplate packaging, e.g., for foodstuffs, consists of very thin tinned steel sheet. The scrap from tinplate cans is processed by the steel recycling industry, and returned to steel production where it is used without problems. The consumption of tinplate packaging and the extent of its recovery for recycling in Germany is shown in Figure 338.

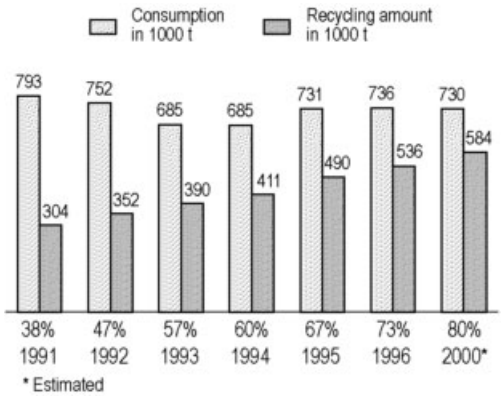

Figure 338. Consumption and partial recycling of tinplate packaging in Germany

Steel sheet packaging for commercial use, e.g., barrels or canisters, can be contaminated with organic or inorganic residues from the filling (paints, varnishes, oils, or adhesives). They must be emptied as far as possible and then pro-

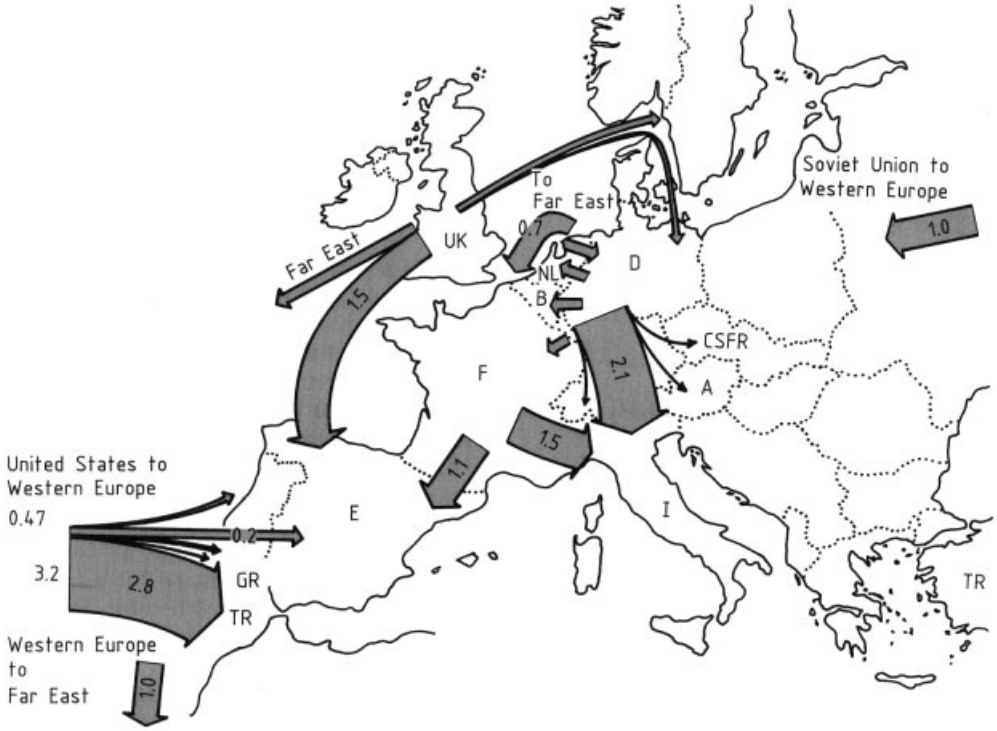

Figure 339. World trade in scrap 1990

cessed by, e.g., shredding, cold shock treatment, or centrifuging, enabling the scrap to remain in the material cycle, provided that the chemical and physical quality requirements of the steel industry are met.

7.2.5. Economic and Logistic Aspects

7.2.5.1. Scrap Consumption

According to the International Iron and Steel Institute (IISI), total world consumption of scrap, including scrap used in foundries, was 369×10^6 t in 1996. Most of this (335×10^6 t) was used in steel production. In 1996, total world crude steel production was 753×10^6 t [719]. Hence, the average consumption of scrap per tonne crude steel produced was ca. 450 kg. Electric steel plants accounted for 226×10^6 t scrap. Thus, ca. 60 % of scrap was consumed by electric steel plants.

7.2.5.2. World Market for Scrap

Scrap is an international commodity with a free price structure. Scrap arisings and consumption vary greatly from one country to another. The proportions of the various steel production processes in different countries have a considerable influence on scrap consumption, and hence on the flow of imports and exports. In electric steelworks, 100 % scrap can be used for steel production. In 1996, the average usage of scrap in German electric steelworks was 1031 kg per tonne crude steel.

In oxygen steel production, pig iron produced in the blast furnace is oxidized together with scrap in a converter. The average usage of scrap in 1996 in the Federal Republic of Germany was 181 kg per tonne crude steel for this process [720]. In the production of high quality steel grades shredded scrap for cooling is preferred because of its high purity and the small lump size.

The most important flows of scrap in world trade in 1990 are illustrated in Figure 339.

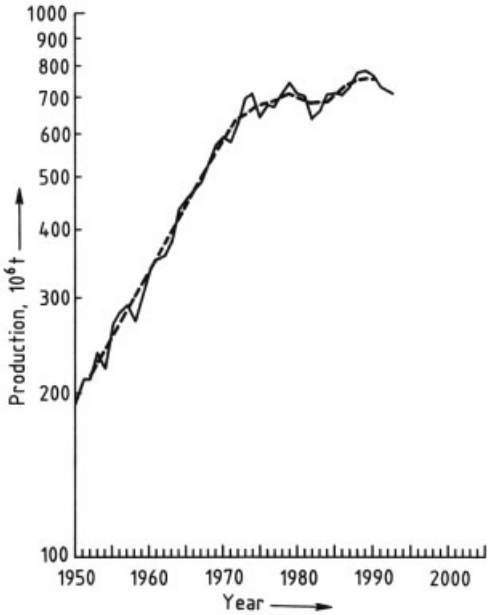

Figure 340. World steel production from 1950 (selected regions)
Smooth curve is 5-year average
Data Supplied by the Statistisches Bundesamt (German Federal Bureau of Statistics) and the International Iron and Steel Institute (IISI)

7.2.5.3. Infrastructure and Logistics

The task of the steel recycling industry is to take responsibility for iron and steel scrap, i.e., to transport, store, process, and deliver to steelworks and foundries a high-quality raw material.

The recycling system offers one particular advantage: The collecting points are installed in all countries. They collect scrap according to its origin, lump size, and impurity levels. The structure of these organizations varies:

1) Small-scale and collection service
2) Medium-sized scrap wholesale traders
3) Organizations supplying scrap to steelworks

Scrap is transported by rail, water, and road. To enable prompt delivery, the steel recycling industry has made considerable investments in handling and transportation facilities.

The demand for scrap is not constant, so the scrap recycling industry maintains stocks containing the equivalent of the purchasing requirements of steelworks for ca. 2 months. Thus, the industry has taken over an important function in balancing scrap arisings with consumption [721].

8. Economic Aspects

8.1. World Steel Production, Consumption, and Trade

From the World War II to the early 1970s, world crude steel production grew at an average annual rate of ca. 5.5 %. Since then, the rate has been only 1 % (Fig. 340). The number of steel-producing countries increased from ca. 30 to ca. 90.

The largest steel-producing companies, with their 1992 production, are shown in Table 29.

Table 29. Leading steel producers 1992

Company	Country	Production, 10^6 t
Nippon Steel	Japan	25.10
Usinor Sacilor	France	21.10
Posco	Korea	20.01
British Steel	United Kingdom	12.39
NKK	Japan	10.89
Ilva	Italy	10.60
Thyssen	Germany	10.13
Kawasaki	Japan	10.00
Sumitomo Metals	Japan	9.97
Sail	India	9.70
Bethlehem	United States	9.60
US Steel	United States	9.50

Of the total world crude steel production, 721×10^6 t in 1992, 345×10^6 t (48 %) was produced by 50 companies in the western world. There are also large steelworks in the countries of the former Soviet Union. Some of these are listed in Table 30.

Table 30. Steel production in the former Soviet Union

Steelworks	Country	Capacity, 10^6 t/a
Magnitogorsk	Russia	16.2
Chereporets	Russia	14.0
Krivoy Rog	Ukraine	13.9
Lipetsk	Russia	9.9
Nizhiny Tagil	Russia	8.3

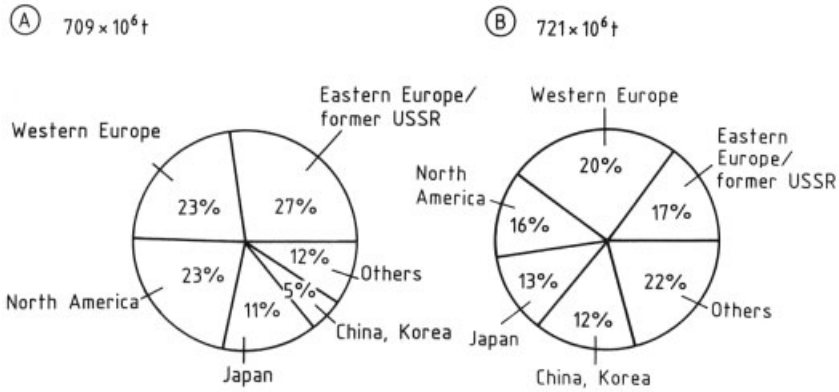

Figure 341. Steel consumption
A) 1974; B) 1992Source: IISI, Brussels

After the political changes of the early 1990s, steel production in these countries decreased considerably, i.e., from 216×10^6 t in 1987 to 124×10^6 t in 1993.

The strong growth of the world steel industry starting in the 1970s was due to the high demand for steel in western Europe, Japan, and eastern Europe. In these regions, steel consumption expanded by almost 7 % per annum. In the United States, the rate was only 2 %. Consumption by the developing countries grew rapidly from small beginnings, at > 8.5 % per annum.

In the last 30 years, considerable regional changes in steel consumption and production have occurred. Approximately one-fifth of steel consumption and 15 % of steel production is associated with the developing regions (Fig. 341). China has also moved ahead. Japan improved its position until the early 1970s, but its share of the world volume has stagnated since then. Europe and the United States are producing a considerably lower proportion than 20 years ago.

Regional shifts have increased the extent to which the developing countries have become self-sufficient, but since then, demand has grown more rapidly than supply. The traditional exporters, e.g., Japan and the European Community, have relied more strongly on their domestic markets. The United States has been a net importer since 1959. The relationship between production and market supply has developed as shown in Table 31.

Brazil has become one of the largest net exporters. Meanwhile, the contribution by eastern Europe and the former Soviet Union to the world total has decreased since the end of the communist planned economy (Fig. 342).

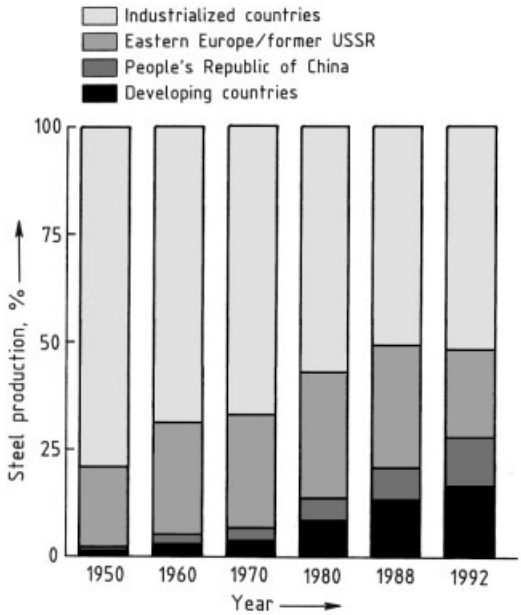

Figure 342. Crude steel production by region 1950 – 1992

World steel exports have grown more rapidly than world production (Fig. 343) because of variations in the regional distribution of demand and production. In the 1970s and early 1980s,

Table 31. Steel production and consumption 1970 – 1992

Region	Production Consumption Balance 10^6 t			Production Consumption Balance 10^6 t			Production Consumption Balance 10^6 t		
	1970			1980			1992		
European Community	146.2	132.9	13.3	141.9	119.9	22.0	129.1	115.2	13.9
Western Europe	161.9	153.9	8.0	161.2	142.1	19.1	153.8	135.3	18.5
United States	119.3	127.3	−8.0	101.5	118.6	−17.1	84.3	100.3	−16.0
Japan	93.3	69.9	23.4	111.4	79.0	32.4	98.1	84.0	14.1
Industrialized countries	397.4	374.0	23.4	406.9	366.2	40.7	366.9	340.4	26.5
Developing countries	22.1	42.2	−20.1	56.7	94.4	−37.7	117.2	147.7	−30.5
Brazil	5.4	6.0	−0.6	15.3	14.3	1.0	23.9	9.1	14.8
Eastern Europe	40.1	41.5	−1.4	61.2	59.6	1.6	32.4	21.1	11.3
Former Soviet Union	115.9	110.2	5.7	147.9	150.4	−2.5	117.0	115.0	2.0
China	18.0	22.5	−4.5	37.1	43.2	−6.1	80.0	83.1	−3.1
World	595.7	592.8	2.9	715.9	718.8	−2.9	720.5	714.8	5.7

the volume of international trade in steel tended to grow at the same rate as general world trade, but in recent years, the volume of steel trade has not increased. The increasing self-sufficiency of new producing countries tends to have a braking effect, but in the area of products and grades, and as a result of competition for customers by the manufacturing companies, active interchange across national boundaries is likely in future.

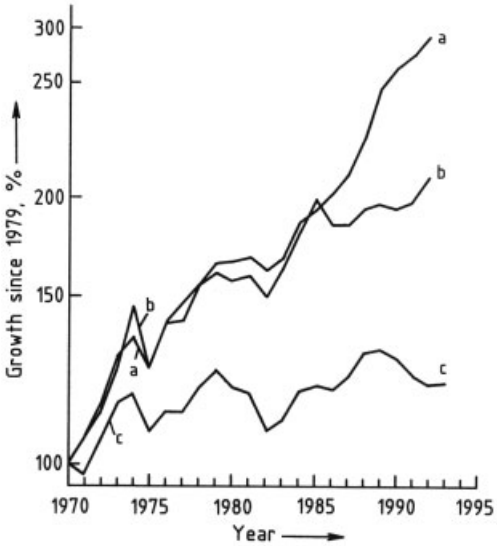

Figure 343. World steel exports, production, and trade from 1970 to 1995
1970 figures = 100 %
a) World trade (all products); b) Steel exports; c) Crude steel production
Source: IISI; Dresdner Bank

Steel exports from Japan have decreased from $> 30 \times 10^6$ t in the mid-1980s to 18×10^6 t in 1991. Most significant is the collapse of trade among the former COMECON countries – ca. 15×10^6 t reduction in the exchange volume. As shown by the high uptake by China from the international market in late 1992/early 1993, relatively large fluctuations in world steel exports are likely.

A summary of the largest exporting and importing countries in 1991 is given in Table 32.

Table 32. World trade in steel 1991

Country	Exports, 10^6 t	Country	Imports, 10^6 t
Germany	19.6	Germany	16.8
Japan	17.9	United States	14.3
Belgium/Luxembourg	14.3	France	10.3
France	12.0	Italy	10.3
Brazil	10.9	Japan	9.0
Italy	9.0	Taiwan	8.7
United Kingdom	8.0	South Korea	8.5
Korea	7.7	United Kingdom	5.6
United States	5.8	Iran	5.5
Netherlands	5.7	Netherlands	5.2
Former Soviet Union	5.4	Belgium/Luxembourg	4.7
Spain	4.8	Former Soviet Union	4.6

An important feature of the steel industry that affects investment decisions is the extent of export orientation of steel manufacturers. In the industrialized countries, the important steel consumers, e.g., engineering and motor vehicle manufacturers, export a high proportion of their steel- containing products (Table 33).

In countries with very high steel exports, e.g., Germany and Japan, the so- called indirect for-

Table 33. Direct and indirect steel exports of selected regions

	Direct export, 10^6 t rolled steel	Indirect export, 10^6 t rolled steel	Direct production, %	Indirect consumption, %
European Community				
1979	65.7	32.0	53.4	32.4
1990	70.2	37.7	58.9	35.8
Japan				
1979	30.7	12.0	30.8	17.1
1990	16.6	20.4	16.9	23.0

eign steel trade affects the trend of development of steel consumption and per capita steel consumption considerably (Table 34). For these two countries in 1990, indirect exports and imports of rolled steel resulted in a net balance, estimated at 9.3 and 19.4×10^6 t. These examples show the importance of the basic conditions for fair international trade.

Table 34. Steel consumption in selected countries 1990

Country	Consumption, kg rolled steel per capita
West Germany	482
France	285
United Kingdom	244
Sweden	336
Japan	751
Brazil	60

The demand for steel, the most important construction material for industry (including the construction industry) and handicrafts, is mainly determined by overall economic development, and hence by capital investment and external trade in steel-containing goods. The optimism which, in the early 1970s, led to investment in steel capacity was considerably dampened after the first oil crisis of 1974–75. The second large increase in the price of oil in 1979, and the increased attention to environmental problems slowed down the economic growth of the western industrialized countries, where the standard of living had reached a remarkably high level. Future growth is likely to be essentially calmer.

8.2. Steel Intensity and Weight Saving in Steel

Steel consumption depends on the state of the overall national economy, especially on investments and foreign trade. In the early stages of industrialization, steel consumption grows at a rate out of proportion to that of the national product. Later, the so-called steel intensity decreases, and steel consumption grows more slowly than the national product (Fig. 344). The most important reasons include:

1) Weight savings in steel manufacture by design optimization, making use of lighter and qualitatively superior steels. The 320 m Eiffel Tower was built of 7000 t steel in 1889, but today would require only 2000 t. A 165 m mast for a radio link station can be built today with only 210 t steel.
2) The use of different materials within the broad area of rolled steel products, e.g., greater use of flat products in place of long products, and the use of increasingly thin sheet for flat products (Fig. 345).
3) Replacement of steel by other materials, e.g., light metals or plastics, e.g., in automobiles. The proportion of steel and iron has decreased from 75 % in the 1970s to 65 % in the early 1990s. The aluminum content has grown from 3 % to 5 %, and plastics from 6 % to 10 %. It is likely that steel will continue to be the most important constituent of automobiles, remaining at ca. 60 %
4) A combination of steel with other materials is often the best solution (e.g., sandwich construction).
5) The reason for most of the reduction in specific steel consumption is the improved productivity of the steel industry itself, and not the replacement of materials.
6) With increasing maturity of a national economy, the fraction of the total national product accounted for by commercial services increases, and the fraction accounted for by manufacturing and heavy basic materials decreases. In 1960, the proportion of the West German national product accounted for by

manufacturing was 53 %, in 1980 it was 43 %, and in 1992, 37 %. Mining, basic materials, and manufactured goods contributed 32 %, 30 %, and 23 % to the turnover of manufacturing business in these years.

8.3. Capital Investment and Subsidies

The steel industry, a capital-, energy-, and raw-material-intensive industry, has always striven to improve productivity and cost effectiveness. Improvements in industrial plant design production methods, and installations for recovering or recycling waste heat, waste gases, electricity, and water have been of central importance. The high intensity of capital investment brought steep cost reductions whenever it was possible to fully utilize the capacities of very large installations. Consequently, from the 1950s to the 1970s, large integrated steelworks were central to strategic planning. Optimization of locations with respect to raw material supplies (coastal steelworks) or large marketing regions (Rhine, Ruhr) was emphasized. Technical development at that time mainly provided the measurement and control techniques essential for large blast furnaces, large converters, strand casting plants, and large rolling mills, and the electronic support for the organization of production. Innovations and improvements have contributed to reductions in the capital cost per tonne steel produced, in specific energy consumption, in raw and auxiliary materials costs, and in general running costs.

In modern steelworks, the production of 1 tonne rolled steel requires 1.1 t crude steel, whereas 20 years ago it required 1.3 t. Consequent savings of raw steel are ca. 80×10^6 t/a worldwide. There is a further possible saving of 40×10^6 t by further modernization, assuming the same volume of rolled steel.

However, intensity of capital investment compared with other branches of industry is still relatively high. In 1990, it was 170 000 € per installation, which compares with only 60 000 € in engineering (Fig. 346). Capital investment constitutes just 20 % of the cost of steel produced in West Germany. Personnel costs account for ca. 25 %. The sum of these costs is only slightly influenced by short-term market fluctuations.

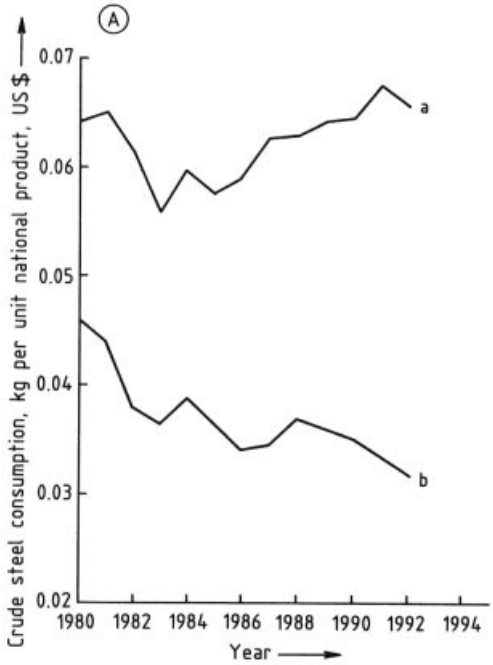

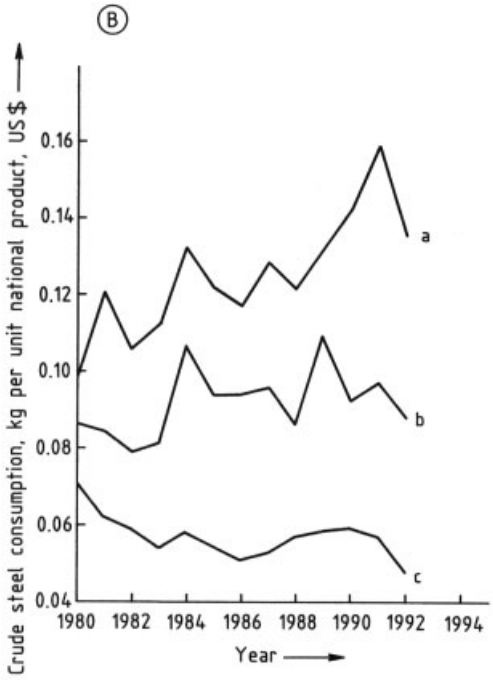

Figure 344. Steel consumption from 1980
A) Developing countries (a); Industrialized countries (b);B) Korea (a); Turkey (b); Japan (c)From calculations supplied by IISI

Large production units cannot be throttled back at will. However, as the steel market reacts extremely strongly to economic change (rising demand in boom periods, falling demand during recessions), suppliers have problems in matching supply to demand, especially in times of weak markets. The result is intense competition between suppliers for small market volumes at falling prices. As long as businesses can build up a financial reserve in good economic times, short-term falls in demand can be coped with. However, this becomes much more difficult if medium-term demand forecasts are over-optimistic.

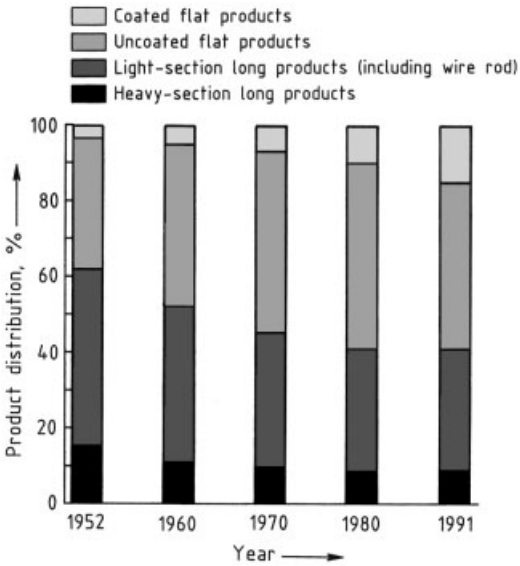

Figure 345. Distribution of rolled steel finished products in the European Community
Source: Eurostat

For a long time, so-called mini-steelworks have become the cost and price leaders among producers of light-section steel products (e.g., rod and wire rod). For this product area, the favored production scheme is electrosteel processing of scrap and small rolling mills. This reduces capital costs, and gives flexibility in matching market fluctuations.

New technology, organization systems, and the resulting increase in productivity and economic efficiency have also influenced price. Steel prices have risen considerably more slowly than those of the average industrial product, or have even fallen, and the severe fluctuations in international quotations during economic cycles have at least been no worse than those the 1970s (Fig. 347).

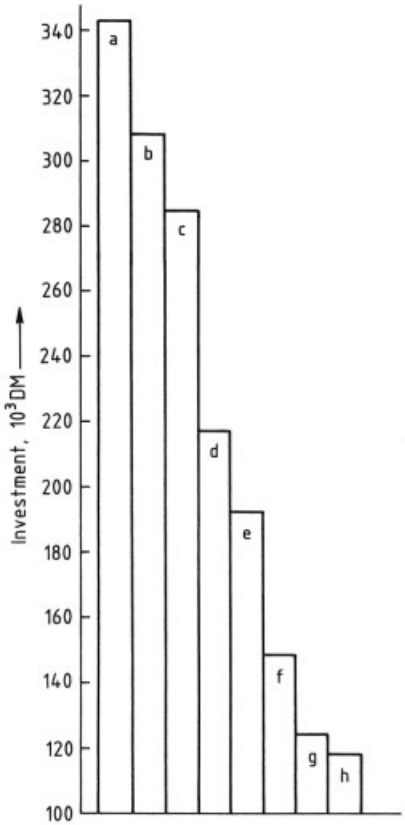

Figure 346. Capital investment per installation, West Germany 1990
a) Steel; b) Chemicals; c) Quarrying, extraction, and working up of stone and earths; d) Industry (total); e) Road vehicles; f) Food industry; g) Electrical industry; h) Mechanical engineeringSource: IW, Cologne

Table 35. Capital investment and subsidies in selected countries 1975 – 1991∗

	West Germany	France	United Kingdom	Italy
Subsidies	5	35	47	49
Capital investment	24	32	35	34

∗ Average figures, €per tonne crude steel produced

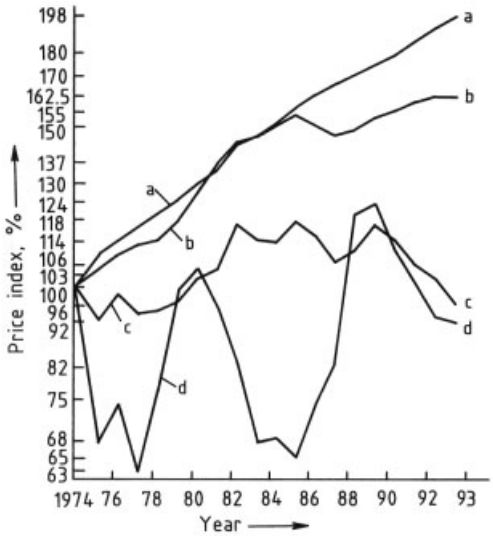

Figure 347. Price fluctuations in selected branches of the West German business economy 1974 – 1993
1985 index recalculated to set 1974 figures = 100 %
a) Producers' prices, motor vehicle construction; b) Commercial products (total); c) Rolled steel; d) Export prices, Western Europe (steel bars)
Data supplied by the Statistisches Bundesamt (German Federal Bureau of Statistics)

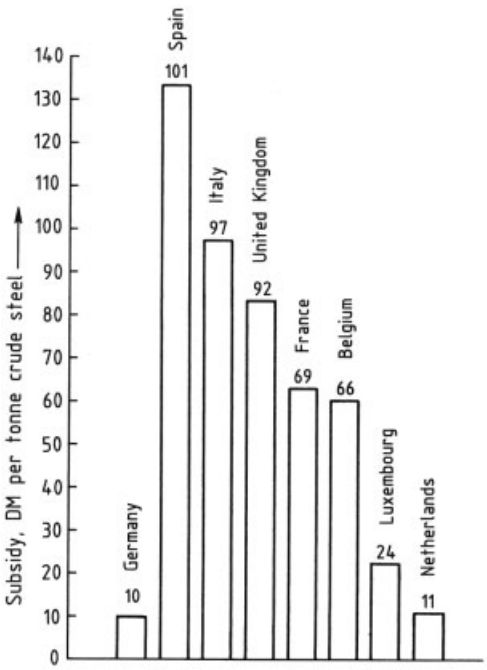

Figure 348. Distortion of competition by subsidies 1975 – 1993
Figures for Spain from 1984Source: Wirtschaftsvereinigung Stahl (Industrial Association for Steel)

There are very high barriers preventing steel-making installations with overcapacity from getting out of the market. In western Europe since the mid 1970s these have become even higher, as state intervention distorts competition, making reduction in capacity very difficult. Large subsidies have ensured survival, especially of nationalized undertakings (Fig. 348). Reduction in capacity is lagging behind the true need, and finance for capital investment in modernization has continued to be available (Table 35).

Private undertakings can only combat such subsidized competition for a limited period, if there is no countervailing policy to avoid subsidized companies increasing their market share.

The steel politics of the European Community since the middle 1970s exemplify the disastrous effects of state intervention in a market economy. The decisions of the European Commission and the Council of Ministers during the steel crisis of 1991 – 1993 show that there is a gulf between the broad political aims of a European Union and the real economic, financial, and structural politics of the member countries.

In some steel-producing countries (e.g., the United Kingdom and Brazil) governments have become unwilling to pay high subsidies, on the grounds that they block the possibilities for investment in alternative projects. Privatization of nationalized steel undertakings has gone ahead.

8.4. Future Prospects

New perspectives are opening up in the steel industry through further technical developments in casting and rolling technology. A new generation of continuous production lines, from crude steel to products close to final dimensions, may considerably reduce the specific capital costs both of flat steel products and of heavy profiles, increasing the flexibility of undertakings to match economic fluctuations. At the same time, productivity can be significantly improved. Different configurations of the supply structure are likely to become available to the large suppliers of flat steel products, who are themselves advancing these technical developments, i.e., independent small and medium manufacturing units, or a combination of several plants with an efficient source of crude iron and steel from large units.

In recent years, small and medium-sized steelworks, mainly based on scrap, have been in favor, but there are limits to this development. Increasing amounts of scrap are available for steel production, but growing world steel demand is likely to require the smelting of iron ore.

Thus, from the economic point of view, many opportunities present themselves for steel as a competitive material for the future.

9. References

1. H.-J. Engell: *Stahl – Prinzipien seines Aufbaus und Aufgaben künftiger Entwicklungen,* Jahrbuch der Akademie der Wissenschaften in Göttingen, 1982.
2. Verein Deutscher Eisenhüttenleute (ed.): "Fundamentals," *Steel–A Handbook for Materials Research and Engineering,* **vol. 1,** Springer Verlag, Berlin 1992.
3. International Iron and Steel Institute (IISI): *World steel in Figures,* Brussels 1992.
4. K. A. Zimmermann: "Renaissance der Eisenbahn – eine Realität?" *IISI – Jahrestagung,* Colorado-Springs 1978.
5. R. Grundmann, H. M. Mozek, K. H. Michel, H. Mülders: "Warmband und Kaltband aus nichtrostenden Stählen," *Stahl u. Eisen* **107** (1987) 1221 – 1227.

General References

6. J. G. L. Blumhof: *Versuch einer Encyklopädie der Eisenhüttenkunde,* **vol. 1 – 4,** G. F. Heyer, Gießen 1816 – 1821.
7. L. Beck: *Geschichte des Eisens,* vol. 1 – 5, Vieweg & Sohn, Braunschweig 1891 – 1903.
8. O. Johannsen: *Geschichte des Eisens,* 3rd ed., Verl. Stahleisen, Düsseldorf 1953.
9. R. F. Tylecote: *A History of Metallurgy,* 2nd ed., Inst. of Materials, London 1992.
10. A. Paulinyi: *Das Puddeln,* Oldenbourg, München 1987. K. C. Barraclough: "Blister Steel," vol. 1; "Crucible Steel," vol. 2; Steelmaking Before Bessemer, Metals Society, London 1984.
11. K. C. Barraclough: *Steelmaking 1850 – 1900,* Inst. of Metals, London 1990.
12. O. Johannsen: "Die geschichtliche Entwicklung der Walzwerkstechnik," in J. Puppe, G. Stauder (eds.): *Walzwerkswesen,* vol. 1, Verl. Stahleisen, Düsseldorf, Springer Verlag, Berlin 1929, pp. 252 –337.

Specific References

13. *Stahlfibel,* Verlag Stahleisen, Düsseldorf 1989, p. 135.
14. O. Johannsen: *Geschichte des Eisens,* 3rd ed., Verlag Stahleisen, Düsseldorf 1953, p. 331.
15. Valerius-Hartmann: *Handbuch der Stahleisen Fabrikation,* J. G. Engelhardt, Freiberg 1985, Fig. 1.
16. W. Rosenbleck, H. W. Kreutzer, R. Steffen, W. Schäfer, R. Willeke, *Stahl u. Eisen* **101** (1981) no. 7, 463 – 471.
17. D. Springorum, *Stahl u. Eisen* **112** (1992) no. 12, 45.
18. H. M. W. Delport, *Ironmaking Steelmaking* **19** (1992) no. 3, 183 – 189.
19. H. W. Kreutzer: "Schrotteinsatz in der Eisen- und Stahlindustrie," paper presented at "Rückstandsoptimierte Werkstoff- und Produktionstechnik am Beispiel des Automobils." RWTH, Aachen, Germany, Nov. 25, 1993.
20. R. Willeke, R. Ewers, H. W. Kreutzer, *Stahl u. Eisen* **114** (1994) no. 5, 83 – 88.
21. Verein Deutscher Eisenhüttenleute (VDEh) (eds.): *Jahrbuch Stahl 1994,* **vol. 1,** p. 216.
22. H. W. Kreutzer, *Stahl u. Eisen* **112** (1992) no. 5, 65 –69.
23. W. Ullrich, H. Schicks, *Stahl u. Eisen* **111** (1991) no. 11, 85 – 92.
24. J. W. Brown, R. L. Reddy, P. J. Salomon: "Economic Problems Related to Creation of Steelplants Using Direct Reduction", Economic Commission for Europe (ECE), Seminar given at the Steel Committee, May 16 – 20, 1983, Nordwijkerhout, Report R 30.
25. R. Steffen, H.-B. Lüngen, *Stahl u. Eisen* **114** (1994) no. 6, 85 – 92.
26. Midrex Corp.: Midrex News Release, Charlotte, N.C. 1993.
27. F. A. Stephens, *Steel Times Int.,* March 1993, 11 –12.
28. K. H. Zepter, F. Daube, *Stahl u. Eisen* **103** (1983) no. 7, 319 – 324.
29. N. Bannenberg, *Stahl u. Eisen* **111** (1991) no. 7, 71 – 76.
30. E. Höffken: "Entwicklungstrend des Oxygenstahlverfahrens," *Freiberg. Forschungsh. B* **229** (1982) 57 – 66.
31. W. Florin, R. Hammer, E. Höffken, W. Ullrich, H. Schicks, *Stahl u. Eisen* **105** (1985) no. 9, 531 –536.
32. J. Geiseler, *Stahl u. Eisen* **111** (1991) no. 1, 531 –536.

33. T. B. Massalski (ed.): *Binary Alloy Phase Diagrams,* 2nd ed., vol. 1, ASM Int, Materials Information Society, Ohio 1990.
34. F. Neumann, H. Schenck, *Arch. Eisenhüttenwes.* **30** (1959) 477 – 483.
35. M. G. Benz, J. F. Elliott, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **221** (1961) 303 – 332.
36. G. K. Sigworth, J. F. Elliott, *Met. Sci.* **8** (1974) 298 – 310.
37. L. S. Darken, R. W. Gurry, *J. Chem. Soc. A* **68** (1946) 798 – 816.
38. C. R. Taylor, J. Chipman, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **154** (1943) 228 – 247.
39. Z. Buzek, A. Hulta, *Freiberg. Forschungsh. B* **117** (1969) 59 – 73.
40. The Japan Society for the Promotion of Science, The 19th Committee on Steelmaking. Steelmaking Data Sourcebook, Gordon and Breach, New York 1988.
41. H. D. Kunze, E. Schürmann, *Gießereiforschung* **20** (1968) 35 – 48.
42. R. D. Pehlke, J. F. Elliott, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **218** (1960) 1088 – 1101.
43. M. Ohtani, *Tetsu-to-Hagané* **54** (1968) 1381 – 1407.
44. E. Schürmann, F. Winterfeld, *Thyssenforschung* **4** (1972) 118 – 132.
45. F. oeters: *Die physikalische Chemie der Eisen- und Stahlerzeugung,* Verlag Stahleisen, Düsseldorf 1964, 156 – 211.
46. R. D. Walker, D. Anderson, *Iron Steel* **45** (1972) 271 – 276.
47. S. Banya, S. Matoba in G. R. St. Pierre (ed.): *Physical Chemistry of Process Metallurgy,* part 1, Interscience Publisher, New York 1959, pp. 373 –401.
48. P. Hammerschmid, D. Janke, H. W. Kreutzer, E. Reichenstein, R. Steffen, *Stahl u. Eisen* **105** (1985) 433 – 442.
49. M. Olette, C. Gatellier, R. Vasse: "Progress in Ladle Steel Refining," *Proc. Int. symp. Phys. Chem. Iron Steelmaking* 1982, art. VII-1.
50. P. V. Riboud, C. Gatellier, *Ironmaking Steelmaking* **12** (1985) 79 – 86.
51. E. T. Turkdogan, *Arch. Eisenhüttenwes.* **54** (1983) 1 – 10.
52. E. Steinmetz, *Rodex Rundschau* 1969, 605 – 617.
53. F. Oeters: *Metallurgie der Stahlerzeugung,* Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1989.
54. V. G. Levitch: *Physicochemical Hydrodynamics,* Prentice Hall, Englewood Cliffs 1962.
55. A.-K. Bolbrinker in Verein Deutscher Eisenhüttenleute (ed.): *Stahlfibel,* Verlag Stahleisen, Düsseldorf 1989.
56. R. Weber, L.-A. Mersoletto, H.-W. gudenau, F.-H. Gradien, *Stahl u. Eisen* **110** (1990) no. 12, 98.
57. Verein Deutscher Eisenhüttenleute (ed.): *Jahrbuch Stahl,* Verlag Stahleisen, Düsseldorf 1990.
58. H. Vöge, E. Eickworth, *Stahl u. Eisen* **79** (1959) no. 23, 1716.
59. T. Shima in: *Proc. Sixth IJSC,* Nagoya 1990.
60. E. Plöckinger, W. Wahlster, *Stahl u. Eisen* **80** (1960) 407 – 416.
61. L.-P. Pesce, H. Todzy, H. Schnitzer, *Int. Steel Mat. Mag.* (1989) 591.
62. R. Heinke et al., *Steel Techn. Int.* 1988.
63. ABB Publikation: "Direct Current Arc Furnace."
64. Verein Deutscher Eisenhüttenleute (ed.): *Eisenhütte,* 5th ed., Verlag Stahleisen, Düsseldorf 1961.
65. H.-U. Lindenberg, K.-H. Schubert, H. Zörcher, *Stahl u. Eisen* **107** (1987) 1192.
66. G. K. Sigworth, J. F. Elliot, *Metal Science* (1971) 298. I. Barin, O. Knacke *Thermodynamical Properties of Inorganic Substances* Springer-Verlag, Heidelberg1993.
67. H. Legrand, M. Amblard in: *Int. Conf. Sec. Met.* 1987, p. 452.
68. F. Oeters, E. Görl, *steel research* **61** (1990) no. 9.
69. G. Stolte, R. Teworte, 33 Met. Prod. **11** (1991) 18/19.
70. D. Nolle, U. Eulenburg, A. Jahns, H. Miska in: *Int. Conf. Sec. Met.* **1987,** p. 275.
71. B. Chalmers: *Principles of Solidification,* J. Wiley & Sons, New York 1964.
72. M. C. Flemings: *Solidification Processing,* McGraw-Hill, New York 1974.
73. W. Kurz, D. J. Fisher: *Fundamentals of Solidification,* Trans Tech SA, Aedermannsdorf, Switzerland 1984.
74. B. W. Berry (ed.), *Solidification of Metals, Proc. Int. Conf.,* Brighton, 4 – 7 December 1967, The Iron and Steel Institute, London 1968.
75. *Solidification and Casting of Metals, Proc. Int. Conf. Solidification,* Sheffield, 18 – 21 July 1977, The Metal Society, London 1979.
76. The University of Sheffield and The Institute of Metals (eds.): "Solidification Processing 1987, Preprints," *Int. Conf.,* Sheffield, United Kingdom, Sept. 21 – 24, 1987.

77. G. Grünbaum et al.: *A Guide to the Solidification of Steels,* Jernkontoret, Stockholm 1977.
78. H. Jacobi, K. Schwerdtfeger, *Metall. Trans.* **7 A** (1976) 811 – 820.
79. M. A. Taha, H. Jacobi, M. Imagumbai, K. Schwerdtfeger, *Metall. Trans.* **13 A** (1982) 2131 –2141.
80. E. Schürmann, S. Baumgartl, L. Nedeljkovic, M. Tripkovic, *steel research* **58** (1987) no. 11, 498 –502.
81. K. Miyamura, I. Taguchi, H. Soga, *Trans. Iron Steel Inst. Jpn.* **24** (1984) 883 – 890.
82. K. Schwerdtfeger: *Metallurgie des Stranggießens, Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1992, pp. 171 – 202.
83. K. Wünnenberg, *Stahl u. Eisen* **98** (1978) no. 6, 254 – 259.
84. W. Dahl, H. Hengstenberg, C. Düren, *Stahl u. Eisen* **86** (1966) no. 13, 782 – 795.
85. E. Steinmetz, H.-U. Lindenberg, *Arch. Eisenhüttenwes.* **47** (1976) no. 9, 521 – 524; no. 12, 713 – 718.
86. K. Schwerdtfeger, *Arch. Eisenhüttenwes.* **41** (1970) no. 9, 923 – 937; **43** (1972) no. 3, 201 – 205.
87. F. Oeters, H.-J. Selenz, E. Förster, *Stahl u. Eisen* **99** (1979) no. 8, 389 – 397.
88. E. Steinmetz, H.-U. Lindenberg, W. Mörsdorf, P. Hammerschmid, *Stahl u. Eisen* **97** (1977) no. 23, 1154 – 1159.
89. E. Steinmetz, H.-U. Lindenberg, *Arch. Eisenhüttenwes.* **47** (1976) no. 4, 199 – 204.
90. *Ullmann,* 4th Ed., 22, 14 – 16.
91. *The Making, Shaping and Treating of Steel,* 10th ed., United States Steel Corp., Pittsburgh, Ass. Iron Steel Eng. (1985) 691 – 740.
92. G. Lepie, H. Rellermeyer: "Untersuchungen über den Erstarrungsverlauf in Gußblöcken," *Arch. Eisenhüttenwes.* **37** (1966) no. 12, 925 – 934.
93. K.-K. Aschendorff, E. Köhler, H. Schroer, B. Abel: "Seigerungen in unberuhigten Stahlblöcken," *Stahl u. Eisen* **82** (1962) no. 20, 1356 – 1366.
94. H. F. Bishop, F. A. Brandt, W. S. Pellini: "Solidification Mechanism of Steel Ingots," *J. Met.* **4** (1952) 44 – 54.
95. A. Diener, A. Drastik, W. Haumann: "Untersuchung über den Wärmeübergang zwischen Block und Kokille beim Erstarren von Stahl," *Arch. Eisenhüttenwes.* **43** (1972) 525 – 533.
96. F. Oeters, K. Sardemann: "Untersuchung zum zeitlichen Verlauf der Erstarrung in der Randzone und zur Spaltbildung zwischen Block und Kokille," *Int. Iron Steel Congr.,* Düsseldorf 1974, vol. 3, VDEh, Düsseldorf 1974.
97. H.-J. Langhammer, H. G. Geck: "Vorgänge beim Gießen und Erstarren von unberuhigtem Stahl," in: *Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1967, pp. 33 – 75.
98. W. Recknagel, H. Oppenhoff: "Was der Blasstahlwerker von seiner Arbeit wissen muß," *Stahleisen-Schriften,* 2nd ed., vol. 6, Verlag Stahleisen, Düsseldorf 1977, p. 120.
99. H. C. Vacher, E. H. Hamilton, *Trans. Am. Inst. Min. metall. Eng. Iron Steel Div.* **95** (1931) 124 –140.
100. A. Hultgren, G. Phragmén, *Trans. Am. Inst. min. metall. Eng. Iron Steel div.* **135** (1939) 133 – 244.
101. H. Rellermeyer: "Vorgänge bei der Blockerstarrung," in: *Die physikalische Chemie der Eisen- und Stahlerzeugung,* Verlag Stahleisen, Düsseldorf 1964, pp. 350 – 376.
102. H. Rellermeyer, R. Hammer: "Metallurgische Fragen beim Vergießen und Erstarren von unberuhigtem Stahl," *Stahl u. Eisen* **78** (1958) no. 22, 1505 – 1513.
103. K. Asano, T. Otiashi, *Tetsu-to-Hagané* **54** (1968) 74 – 75, 643 – 673.
104. P. Nilles: "Theoretical Study of the Solidification of Rimming Steel," *J. Iron Steel Inst. London* **202** (1964) 601 – 609.
105. P. Nilles, F. Becker, A. Thill: "Untersuchung über die Erstarrung von unberuhigt vergossenem Stahl," *Stahl u. Eisen* **85** (1965) 1025 – 1032.
106. H.-J. Langhammer, H. G. Geck: "Vorgänge beim Gießen und Erstarren von unberuhigtem Stahl," in: *Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1967, pp. 33 – 75.
107. K. Kupzog, R. Hammer, H. Rellermeyer: "Untersuchungen zur Seigerung bei unberuhigtem Stahl," *Stahl u. Eisen* **82** (1962) no. 7, 394 – 401.
108. E. Schürmann, W. Groetschel, O. Peter: "Untersuchung des Auskoch- und Erstarrungsverhaltens von unberuhigt vergossenem Stahl," *Arch. Eisenhüttenwes.* **36** (1965) 619 – 631.
109. R. Hammer, G. Lepie, Th. Kootz, K. Wick: "Einschlüsse im unberuhigten Stahl," in "Das Vergießen und Erstarren von Stahl zu

Blöcken," *Freiberger Forschungsh. B* **137,** Leipzig 1968, pp. 56 – 77.
110. "Gemeinschaftsforschung über Grundlagen des Gießens und Erstarrens," *Stahl u. Eisen* **96** (1976) no. 17, 840 – 846.
111. F. Oeters: "Blockguß," *Z. Metallk.* **63** (1972) 301 –306.
112. P. M. Macnair, *J. Iron Steel Inst. London* **160** (1948) 151 – 163.
113. A. Diener: "Strömungstechnische Probleme beim Vergießen von Stahl," *Stahl u. Eisen* **96** (1976) no. 25/26, 1337 – 1340.
114. J. R. Blank, E. B. Pickering: "The Effect of Solidification in Large Ingots on Segregation of Non-Metallic Inclusions," *The Solidification of Metals. Proc. of the Conf.,* Brighton 1967, Iron and Steel Inst., London 1968, pp. 370 – 376.
115. K. Schwerdtfeger: "Einfluß der Erstarrungsgeschwindigkeit auf die Mikroseigerung und die interdendritische Ausscheidung von Mangansulfideinschlüssen in einem Mangan und Kohlenstoff enthaltenden Stahl," *Arch. Eisenhüttenwes.* **41** (1970) no. 9, 923 – 937.
116. R. D. Doharty, D. A. Melford: "Solidification and Microsegregation in Killed Steel Ingots with Particular Reference to 1 % C, 1.5 % Cr Steel," *J. Iron Steel Inst. London* **204** (1966) 1131 – 1143.
117. F. Oeters, K. Rüttiger, A. Diener, G. Zahs: "Zur Theorie der metallurgischen Vorgänge bei der Erstarrung von Stahl," *Arch. Eisenhüttenwes.* **40** (1969) no. 8, pp. 603 – 613.
118. C. Roques, P. Martin, Ch. Dubois, P. Bastien: "Etude de l'hétérogénéité des gros lingots de forge," *Rev. Metall.* (*Paris)* **57** (1960) no. 12, 1091 –1103.
119. A. Hultgren: "A and V Segregation in Killed Steel Ingots," *Scand. j. Meetall.* **2** (1973) 217 – 227.
120. H.-D. Pantke, H. Neumann: "Vorgänge beim Gießen und Erstarren von halbberuhigtem Stahl," in: *Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1967, pp. 76 – 90.
121. H. Knüppel, F. Eberhard: "Gesetzmäßigkeiten bei der Kohlenmonoxydentwicklung teilberuhigt erstarrender Stähle," *Arch Eisenhüttenwes.* **34** (1963) no. 5, pp. 325 – 340.
122. J. Koenitzer, R. Hammer: "Der Sauerstoff im beruhigten und halbberuhigten Stahl," *Stahl u. Eisen* **83** (1963) no. 10, 569 – 577.
123. K. F. Behrens, R. Hammer: "Production of Aluminium-Killed Steel by Mold Additions of Aluminium," *Proc. Nat. Open Hearth Basic Oxygen Steel Conf.* **50** (1967) 112 – 115.
124. U. Petersen, K. G. Speith, A. Bungeroth, *Stahl u. Eisen* **86** (1966) no. 6, 333 – 353.
125. M. M. Wolf, *Steelmaking Conf.* **75** (1992) 83 – 137.
126. *Contin. Cast. Steel, Proc. Int. Conf.,* Biarritz, 31 May – 2 June 1976, The Metals Society, London 1977.
127. "Continuous Casting," *Proc. Int. Iron Steel Congr. 4th,* London, May 12 – 14, 1982.
128. *Continuous Casting '85,* London, 22 – 24 May, 1985, The Institute of Metals, London 1985.
129. CRM and VDEh (eds.): *4th International Conference Continuous Casting,* ICC '88, Brussels, 17 – 19 May 1988, Stahleisen, Düsseldorf 1988.
130. AIM (ed.): *1st European Conference on Continuous Casting,* Florence, 23 – 25 Sept. 1991, Milano 1991.
131. H. Schrewe: *Stranggießen von Stahl, Einführung und Grundlagen,* Verlag Stahleisen, Düsseldorf 1987.
132. K. Schwerdtfeger: *Metallurgie des Stranggießens, Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1992.
133. L. J. Heaslip, A. McLean, I. D. Sommerville: "Chemical and Physical Interactions During Transfer Operations," *Continuous Casting,* vol. 1, ISS-AIME, Warrendale 1983.
134. J. K. Brimacombe, I. V. Samarasekera, J. E. Lait: "Heat Flow, Solidification and Crack Formation," *Continuous Casting,* vol. 2, ISS-AIME, Warrendale 1984.
135. J. J. Moore: "The Application of Electromagnetic Stirring (EMS) in the Continuous Casting of Steel," *Continuous Casting,* vol. 3, ISS-AIME, Warrendale 1984.
136. T. B. Harabuchi, R. D. Pehlke: "Design and Operations," *Continuous Casting,* vol. 4, ISS-TMS, Warrendale 1988.
137. R. A. Heard, A. McLean: "Horizontal Continuous Casting," *Continuous Casting,* vol. 5, ISS-TMS, Warrendale 1988.
138. H. Schrewe et al., *Stahl u. Eisen* **108** (1988) no. 9, 427 – 436.
139. "Progress of the Iron and Steel Technologies in Japan in the Past Decade, III: Steelmaking," *Transactions ISIJ* **25** (1985) 627 – 710.
140. E. Höffken et al., *Stahl u. Eisen* **105** (1985) no. 22, 1167 – 1175.

141. G. Holleis, R. Scheidl, W. Dutzler, *Fachber. Hüttenprax. Metallweiterverarb.* **25** (1987) no. 10, 959–967.
142. D. Kothe, F.-P. Pleschiutschnigg, F. Boehl, *MPT Metallurg. Plant and Technol.* **13** (1990) no. 1, 12–26.
143. P. Nilles, A. Etienne, *1st European Conf. on Continuous Casting,* **vol. 1,** Florence 1991, pp. 1.1 – 1.16.
144. D. Schauwinhold, *Stahl u. Eisen* **102** (1982) no. 21, 1077 – 1082.
145. P. Riboud in K. Schwerdtfeger (ed.): *Metallurgie des Stranggießens, Gießen und Erstarren von Stahl,* Verlag Stahleisen, Düsseldorf 1992, pp. 233 – 255.
146. K. Wünnenberg, *Stahl u. Eisen* **98** (1978) no. 6, 254 – 259.
147. R. Thielmann, R. Steffen, *Stahl u. Eisen* **100** (1980) no. 7, 401 – 407.
148. P. Voss-Spilker, W. Reichelt, *Iron Steel Eng.* 1983 June, 32 – 38.
149. P. Stadler, J. von Schnakenburg, *Stahl u. Eisen* **109** (1989) no. 9, 463 – 469.
150. T. Nozaki, S. Itoyama, *Trans. Iron Steel Inst. Jpn.* **27** (1987) 321 – 331.
151. The Korean Institute of Metals, The Institute of Metals, UK, Preprints 2 (eds.): "SRNC-90 Near Net Shape Casting," Internat. Conf., RIST, Pohang, Korea, Oct. 14 – 19, 1990.
152. J.-P. Birat, R. Steffen, *MPT Metallurg. Plant Technol.* **14** (1991) no. 3, 44 – 57.
153. E. Höffken, P. Kappes, H. Lax, *Stahl u. Eisen* **106** (1986) no. 23, 1253 – 1259.
154. F. K. Iverson, K. Busse, *Stahl u. Eisen* **111** (1991) no. 1, 37 – 45.
155. W. Rohde, H. Wladika, *Stahl u. Eisen* **111** (1991) no. 1, 47 – 61.
156. H.-J. Ehrenberg, L. Parschat, F.-P. Pleschiutschnigg, C. Praßer, W. Rahmfeld, *Stahl u. Eisen* **109** (1989) no. 9/10, 453 – 462.
157. F.-P. Pleschiutschnigg et al., *MPT Metallurg. Plant Technol.* **15** (1992) no. 2, 66 – 82.
158. E. Höffken, *Stahl u. Eisen* **113** (1993) no. 2, 49 – 54.
159. W. E. Duckworth, G. Hoyle: *Electro-Slag Refining,* Chapman and Hall, London 1969.
160. H. C. Child, G. E. Oldfield, R. Bakish, A. Lawley: "Vacuum Melting" in O. Winkler, R. Bakish (eds.): *Vacuum Metallurgy,* Chap. V, Elsevier, Amsterdam 1971, pp. 517 – 642.
161. W. H. Sutton: "Progress in the Vacuum (VIM, VAR) Melting of High Performance Alloys," *Proc. Int. Conf. Vac. Metall. 7th,* Nov. 26 – 30, 1982, Tokyo, The Iron and Steel Institute of Japan, Tokyo 1982, pp. 904 – 915.
162. L. W. Lherbier: "Melting and Refining," in C. T. Sims, N. S. Stoloff, W. C. Hagel (eds.): *Super-alloys II,* Chap. 14, J. Wiley & Sons, New York 1987, pp. 387 – 410.
163. A. Choudhury: "State of the Art of Superalloy Production for Aerospace and Other Applications Using VIM/VAR or VIM/ESR," *ISIJ Int.* **32** (1992) no. 5, 563 – 574.
164. A. Choudhury, E. Hengsberger: "Electron Beam Melting and Refining of Metals and Alloys," *ISIJ Int.* **32** (1992) no. 5, 673 – 681.
165. J. Menzel, G. Stein, R. Dahlmann: "Manufacture of N-Alloyed Steels in a 20 t PESR-Furnace," in *High Nitrogen Steels HNS 90,* Aachen, Oct. 10 – 12, 1990, Verlag Stahleisen, Düsseldorf 1990, pp. 365 –375.
166. G. K. Bhat: "Plasma Arc Remelting" in J. Feinman (ed.): *Plasma Technology in Metallurgical Processing,* Chap. 13, Iron and Steel Society, Warrendale 1987, pp. 163 – 174.
167. K. roesch, H. Zeuner, K. Zimmermann: *Stahlguß,* 2nd ed., Verlag Stahleisen, Düsseldorf 1982.
168. H. Wübbenhorst, G. Engels: *5000 Jahre Gießen von Metallen,* Giesserei-Verlag, Düsseldorf 1989.
169. F. Roll: *Handbuch der Gießerei-Technik,* Springer Verlag, Berlin 1959.
170. H. G. Gerhard, O. Nickel, K. Röhrig, D. Wolters: *Legiertes Gußeisen,* **vol. 1, 1970, vol. 2, 1974,** Gießerei-Verlag, Düsseldorf.
171. A. Troost: *Einführung in die allgemeine Werkstoffkunde metallischer Werkstoffe I,* Bibliographisches Institut, Mannheim 1980.
172. Verein Deutscher Eisenhüttenleute: "Grundlagen," *Werkstoffkunde Stahl,* **vol. 1,** Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1984.
173. A. Hensel, T. Spittel: *Kraft- und Arbeitsbedarf bildsamer Formgebungsverfahren,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1978.
174. K. Karhausen, R. Kopp, *steel research* **63** (1992) no. 6, 247 – 256.
175. C. M. Sellars, *Mater. Sci. Technol.* **6** (1990) 1072 –1082.
176. W. Dahl et al., *Stahl u. Eisen* **111** (1991) no. 4, 113 –119.
177. L. Meyer, *Stahl u. Eisen* **101** (1981) no. 7, 483 – 491.
178. H. M. Aichinger et al.: "Großöfen zur Erwärmung von Walzgut-Betriebserfahrungen und Entwicklungstrends," *Stahl u. Eisen* **107** (1987) no. 18, 819 –831.

179. H. Hammer: "Konzeption und Betriebserfahrungen mit dem neuen 200-t/h-Hubbalkenofen einer Mittelbandstraße," *Stahl u. Eisen* **109** (1989) no. 22, 1095 – 1103.
180. F. Reinitzhuber, H.-P. Domels: "Einsatz neuzeitlicher Technologien an Walzwerksöfen," *Stahl u. Eisen* **110** (1990) no. 8, 139 – 149.
181. W. Evers et al.: "Energieoptimale modellgestützte Führung einer Walzwerksofengruppe," *Stahl u. Eisen* **108** (1988) no. 10, 489 – 497.
182. H. Klammer, G. Woelk, W. Brüsecke, H. J. Heidemüller: "Dynamisch-adaptive Prozeßführung eines 35-t/h-Drehherdofens mit Rechnervernetzung auf PC-Basis," *Stahl u. Eisen* **110** (1990) no. 9, 103 – 109.
183. F. Köster, H.-J. Wick, A. Funghini, B. Pastorino: "Rechnergestützte Vorgabe der Zonentempera- turen an einem Hubbalkenofen mit Industrie-Personal-Computern," *Stahl u. Eisen* **111** (1991) no. 11, 49 – 56.
184. W. Schupe: "Erfahrungen mit Prozeßrechnerführungen an verschiedenen Ofentypen und Wärmgutgeometries," *Stahl u. Eisen* **112** (1992) no. 8, 47 – 52.
185. H. Brod, G. Thiemann, H. Voigt, D. Wick: "Optimierung der Wiedererwärmungsöfen des Warmbreitbandwalzwerkes der Krupp Stahl AG," *Stahl u. Eisen* **112** (1992) no. 8, 41 – 46.
186. H. Hammer, P. Müller, D. Kostka: "Einsatz eines dynamisch-adaptiven Prozeßführungssystems an einem 200-t/h-Hubbalkenofen," *Stahl u. Eisen* **113** (1993) no. 3, 115 – 122.
187. S. Böhnisch, R. Klima, A. Kuck, D. Sucker: "Verbesserte Führung von Durchlauf-Wärmöfen durch Off-line-Simulation und autoadaption," *Stahl u. Eisen* **113** (1993) no. 10, 63 – 68.
188. Stahl-Eisen Werkstoffblatt des VDEh: SEW 082, 2nd ed., June 1989.
189. L. Meyer, *Stahl u. Eisen* **101** (1981) 483 – 491.
190. W. Bleck, L. Meyer, R. Kaspar, *Stahl u. Eisen* **111** (1991) no. 5, 51 – 56.
191. H. Baumgardt, M. Rohde, L. Schulz, *Thyssen Tech. Ber.* (1982) no. 2, 126 – 135.
192. K. Lorenz et al., *Stahl u. Eisen* **101** (1981) 593 – 600.
193. F. Weber, W. Oswald, M. Engel, G. Kopf, *Stahl u. Eisen* **107** (1987) 307 – 313.
194. H. Schmitz et al., *Stahl u. Eisen* **109** (1989) 823 – 832.
195. H. Wladika, W. Rohde, *Stahl u. Eisen* **111** (1991) no. 1, 47 – 61.
196. F. Brettbacher, F. Hirschmanner, *Fachber. Huettenprax. Metallweiterverarb.* **21** (1983) no. 10, 835 –841.
197. A. Hensel, P. Poluchin: *Technologie für Metallformung,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1991, p. 249.
198. H. G. Hüsken, K. Herwig, *Stahl u. Eisen* **103** (1983) 295 – 300.
199. Schloemann-Siemag (SMS): W 6/3141: CVC, company brochure.
200. Mannesmann-Demag-Sack (MDS): *UPC-Technologie,* company brochure.
201. G. Kneppe, W. Rohde, *Stahl u. Eisen* **113** (1993) no. 7, 71 – 79.
202. B. Berger, K.-H. Mommertz, E. Neuschütz, P. D. Pütz, *Stahl u. Eisen* **105** (1985) 1201 – 1206.
203. G. Oehlstöter, J. Jäckel, *Stahl u. Eisen* **112** (1992) 53 – 60.
204. G. Oehlstöter, W. Schwenzfeier, *BMH Berg – Hüttemaenn. Monatsh.* **135** (1990) 237 – 242.
205. E. Probst, E. Wild, BBC Nachrichten 1979 Nr. 8 S 274/82.
206. K. F. Becker, W. Krämer, A. Rank, *Stahl u. Eisen* **105** (1985) 1329 – 1332.
207. A. Hensel et al., *Stahl u. Eisen* **111** (1991) no. 11, 63 – 68
208. G. Engel, H. Feldmann, *Steel Metals Magazine* **27** (1989) no. 10, 755 – 760.
209. F. Becker, J. M. Dengler, H. Schmedders, K. Wick, *Stahl u. Eisen* **111** (1991) no. 4, 133 – 138.
210. A. Frantz, F. Becker, J. de la Hamette, *Stahl u. Eisen* **109** (1989) 497 – 502.
211. C. Panunzi, M. Rasquin, J. M. Dengler, St. Wilmotte, *Stahl u. Eisen* **110** (1990) no. 6, 39 – 43.
212. H. Bienzeisler: Berichtsband 2nd Duisburger Adjustage-Tage, Feb. 20/21, 1992, Akademie für Wissenschaft und Technik, Duisburg 1992, pp. 237 –259.
213. H. Feldmann, H. Müller, *Stahl u. Eisen* **108** (1988) 275 – 280.
214. R. Kopp, K. H. Tuke, *Stahl u. Eisen* **97** (1977) 761 –765.
215. M. Albedyhl et al., *Stahl u. Eisen* **108** (1988) 1077 –1082.
216. J. Grotepass, H. G. Limper, H. Muckli, K. Kessler, *Stahl u. Eisen* **104** (1984) 995 – 997.
217. A. Hauck, *MPT Metall. Plant Technol.* **10** (1987) 56 – 64.
218. W. Dietl, *Stahl u. Eisen* **90** (1970) 1223 – 1234.

219. H. Brauer, W. J. Ammerling, K. H. Stauss, *Der Kalibreur* **53** (1990) 3 – 35.
220. H. Brauer, *Stahl u. Eisen* **112** (1992) no. 7, 53 – 60.
221. W. Krämer, J. Schnyder, H. Feldmann, G. Kirchmann, *Stahl u. Eisen* **110** (1990) no. 6, 59 – 64.
222. A. Hensel, P. Poluchin (eds.): *Technologie der Metallformung,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1991, p. 476.
223. K.-H. Brensing, B. Sommer: "III Herstellverfahren," *Stahlrohrhandbuch,* 9th ed., Verlag Stahleisen, Düsseldorf 1982, pp. 111 – 168.
224. K. Oberem: "Herstellung nahtloser Rohre," *Stahl u. u. Eisen* **104** (1984) no. 25 – 26, 47 – 51.
225. H. Biller: "Das Walzen nahtloser Rohre – Probleme der Verfahrensauswahl," *Stahl u. u. Eisen* **106** (1986) no. 9, 49 – 55.
226. W. Schupe, G. Thaler: "überblick über den Entwicklungsstand der Ofenführung durch Rechnerunterstützung," *Stahl u. u. Eisen* **107** (1987) no. 20, 37 – 42.
227. E. Siebel: "Grundsätzliche Betrachtungen zum Schrägwalzverfahren," *Stahl u. u. Eisen* **47** (1927) no. 41, 213 – 214.
228. M. Hildebrand: "Der Einfluß der Walzenkalibrierung bei Lochwalzwerken auf den Verlauf der Formänderung und das Auftreten von Fehlern bei der Rohrherstellung," *Neue Hütte* **6** (1961) no. 4, 251 – 254.
229. A. Eicken: "Untersuchung über die wichtigsten Einflußgrößen, die beim Schrägwalzen zu Fehlern führen," Dissertation, TU Clausthal 1971.
230. K. Oberem: "Das Schrägwalzwerk als Lochaggregrat beim Rohrwalzen," *Stahl u. u. Eisen* **107** (1987) no. 16, 51 – 54.
231. R. Kümmerling: "Comparison of various cross-olling processes for elongation stages in the production of seamless tubes," *Metallurgical Plant and Technology International* **15** (1992) no. 1, 48 – 57.
232. L. Prümmer in VDEh (ed.): *Grundlagen des Beizens in Herstellung von kaltgewalztem Band,* Verlag Stahleisen, Düsseldorf 1970.
233. W. Bumbullis, G. Köhler, B. Schweinsberg: "Kaltwalzen auf flacher Bahn," *Beizeinrichtungen in Kontaktstudium Umformtechnik,* part II, seminar preprint, Verlag Stahleisen, Düsseldorf 1981.
234. B. Frisch, W. R. Thiele, *Stahl u. u. Eisen* **101** (1981) no. 9, 577 – 585.
235. B. Frisch, W. R. Thiele, D. Prediger, *Arch. Eisenhüttenwes.* **54** (1983) no. 8, 311 – 380.
236. B. Frisch, W. R. Thiele, C. Messerschmidt, *Arch. Eisenüttenwes.* **54** (1983) no. 9, 375 – 380.
237. H. J. Heimhard, G. Hitzemann, *Stahl u. Eisen* **105** (1985) 1222 – 1228.
238. E. Neuschütz: "Planheitsmessung und -regelung beim Warm- und Kaltwalzen von Bändern," *Symposium DGM "Walzen von Flachprodukten,"* DGM-Informationsgesellschaft, Oberursel 1987.
239. B. Chatelain: "Einfluß der Glühparameter auf die Gas-Metall-Reaktion bei fest gewickelten Ringen," EGKS-Bericht EUR 12 115 FR.
240. B. Chatelain, V. Leroy: "Gas-Metall-Reaktionen beim Haubenglühen, Einfluß die Oberflächeneigenschaften des Blechs," EGKS-Abschlußbericht Nr. 7210/KB/204.
241. B. Chatelain et al.: "Beurteilung des Glühens von kohlenstoffarmen Stählen mit Wasserstoff – Gas-Metall-Reaktionen," *Rev. Metall.* CII, Feb. (1989) 173 – 180.
242. R. Pankert: "Durchlaufglühverfahren für Feinbleche im Vergleich," *Stahl u. Eisen* **105** (1985) 889 ff.
243. H. W. Honervogt et al.: "Blankglühanlage für nichtrostende und säurebeständige Kaltbänder," *Stahl u. Eisen* **107** (1987) no. 9, 267.
244. J. Albrecht et al.: "Stand der Technik bei der Herstellung von Schmelztauchüberzügen," *Stahl u. Eisen* **107** (1987) no. 21, 973 ff.
245. J. H. Meyer zu Bexten, G. Gessner, K. F. Hüttebräucker, K. P. Mohr: "Elektrolytisch verzinktes Feinblech – Die Situation in Deutschland," *Stahl u. Eisen* **108** (1988) no. 21, 975 ff.
246. B. Meuthen, R. Pankert: "Stand der Breitbandverzinkung in Japan," *Stahl u. Eisen* **110** (1990) no. 6, 75 ff.
247. H. Lämmermann, K. Frommann: "Oberflächenveredelung von Band," *Vakuumbeschichtung von Stahlband,* Berichtsband, Verlag Stahleisen, Düsseldorf 1990.
248. W. Dürr, G. Höltzsch, J. Kuhn, W. Schlump: "Aus der Dampfphase beschichtetes Blech und Band," *Stahl u. Eisen* **113** (1993) no. 5, 79 – 87.
249. P. von Laer, H. Hülsmann: *Herstellung von bandbeschichtetem Feinblech – Eigenschaften,*

Verarbeitung, Anwendung, VDEh, DVV, Düsseldorf 1987.
250. P. Funke, H. R. Priebe, H. Buddenberg: "Die Untersuchung von Verfahrensparameters beim Walzplattieren," *Stahl u. Eisen* **110** (1990) no. 6, 67 ff.
251. H. Buddenberg, P. Funke, W. Gebel, R. Theile: "Herstellung und Eigenschaften NE-metallplattierter Stahlbänder für die Automobilindustrie," *Stahl u. Eisen* **111** (1991) no. 7, 93 ff.
252. Verein Deutscher Eisenhüttenleute (ed.): "Fundamentals," *Steel – A Handbook for Materials Research and Engineering,* **vol. 1,** Springer Verlag, Verlag Stahleisen, Düsseldorf 1992, p. 17.
253. G. Krauss in F. B. Pickering (ed.): *Microstructure and Transformations in Steel,* **vol. 7,** VCH Verlagsgesellschaft, Weinheim 1992, p. 1.
254. O. Kubaschewski: *Iron Binary Phase Diagrams,* Springer Verlag, Berlin 1982.
255. T. B. Massalski: "Binary Alloy Phase Diagram," *ASM Int.* **2** (1990).
256. B. Sundmann, B. Jansson, J.-O. Andersson, *CALPHAD; Comput. Coupling Phase Diagrams Thermochem.* **9** (1985) 153 – 190.
257. B. Massalski: "Binary Alloy Phase Diagrams," *ASM Int.* **1 – 3** (1990).
258. V. Raghavan: "Phase Diagrams of Ternary Iron Alloys," *ASM Int.* (1987) part 1.
259. V. Raghavan: *Phase Diagrams of Ternary Iron Alloys,* Indian Institute of Metals, Calcutta 1988, part 2.
260. G. V. Raynor, V. G. Rivlin: *Phase Equilibria in Iron Ternary Alloys,* Institute of Metals, London 1988.
261. V. Raghavan: *Phase Diagrams of Ternary Iron Alloys,* Institute of Materials Publ., Calcutta 1988, part 3.
262. V. Raghavan: *Phase Diagrams of Ternary Iron Alloys,* Institute of Materials Publ., London 1992, parts 6 A and 6 B.
263. W. C. Leslie: *The Physical Metallurgy of Steels,* McGraw-Hill, New York 1981.
264. V. L. Indenbom, J. Lothe: *Modern Problems in Condensed Matter Sciences – Elastic Strain Fields and Dislocation Mobility,* Elsevier, New York 1992.
265. K. Lücke, *Z. Metallkd.* **75** (1984) no. 12, 948 – 956.
266. H. E. Exner et al., *Prakt. Metallogr.* **30** (1993) no. 8, 389.
267. H. E. Exner, H. P. Hougardy: *Quantitative Microscopy,* DGM Informationsgesellschaft, Oberursel 1988.
268. D. J. Diaz: *Quantitative Microscopy and Image Analysis,* ASM Publ., Ohio 1994.
269. H.-J. Bunge, *Z. Metallkd.* **76** (1985) no. 2, 92 – 101.
270. M. Hatherley, W. B. Hutchinson: *An Introduction to Textures in Metals,* Institute of Materials Publ., London 1994.
271. A. H. Cottrell: *Introduction to the Modern Theory of Metals,* Institute of Metals, London 1988.
272. E. E. Underwood: *Quantitative Stereology,* Addision Wesley, London 1970.
273. *Metals Handbook,* 9th ed., vol. 9, "Metallography and Microstructures," American Society for Metals, Ohio 1985.
274. *Atlas zur Wärmebehandlung der Stähle,* vol. 3, Verlag Stahleisen, Düsseldorf 1973.
275. *Atlas zur Wärmebehandlung der Stähle,* vol. 4, Verlag Stahleisen, Düsseldorf 1976.
276. W. Pitsch, G. Sauthoff: "Microstructure of Steels," in VDEh (eds.): *Steel,* Springer Verlag, Verlag Stahleisen, Düsseldorf 1992, p. 1.
277. *Atlas zur Wärmebehandlung der Stähle,* vol. 1, Verlag Stahleisen, Düsseldorf 1961.
278. G. Vander Voort: *Atlas of Time-Temperature Diagrams for Irons and Steels,* American Society for Metals, Ohio 1991.
279. G. B. Olson, W. S. Owen: *Martensite,* ASM Publ., Ohio 1991.
280. A. Rose, H. P. Hougardy: *Atlas zur Wärmebehandlung der Stähle,* vol. 2, Verlag Stahleisen, Düsseldorf 1972.
281. H. P. Hougardy: "Microstructures of Steels. The Characterization of Ferrite, Pearlite, Bainite and Martensite, *Prakt. Metallogr.* **20** (1989) 179.
282. S. Okaguchi, H. Ohtani, Y. Ohmori, *Metall. Trans. JIM* **32** (1991) no. 8, 697 – 704.
283. S. Crusius et al., *Z. Metallkd.* **83** (1992) 729 –738.
284. A. S. Schulz-Beenken, H. P. Hougardy: "Development of High-Strength Martensite by Age-Hardening of Coherent Intermetallic Precipitates," in E. Hornbogen, H. Jost (eds.): *The Martensitic Transformation in Science and Technology,* DGM Informationsgesellschaft, Oberursel 1989, p. 473.
285. H. K. D. H. Bhadeshia: *Bainite in Steels,* Institute of Materials Publ., London 1992.
286. W. T. Reynolds, Jr., H. I. Aaronson, G. Spanos, *Metall. Trans. JIM* **32** (1991) no. 8, 737 – 746.
287. W. T. Reynolds et al., *Metall. Trans.* **21 A** (1990) 1479.

288. G. R. Speich: *Fundamentals of Aging and Tempering in Baintic and Martensitic Steel Products,* Aime Iron & Steel Soc., Warrendale 1992.
289. H. W. Rayson: "Tool Steels," in F. B. Pickering (ed.): *Materials Science and Technology,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 583.
290. S. Karagöz, H. Fischmeister, *steel research* **58** (1987) no. 8, 353.
291. J. Peterseim, G. Sauthoff, *steel research* **57** (1986) no. 2, 77 – 81.
292. E. Hornbogen, K. Kishida, *Arch. Eisenhüttenwes.* **55** (1984) no. 6, 281.
293. F. B. Pickering (ed.): "Structure-Property Relationships in Steels," in F. B. Pickering (ed.): *Materials Science and Technology,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, pp. 41, 583.
294. J. G. Parr, H. Hanson, R. A. Lula: *Stainless Steel,* ASM Publ., Ohio 1985.
295. H. Nordberg, J. Björklund (eds.): *Applications of Stainless Steel '92,* Proceeding, June 9 – 11, 1991, Stockholm/Sweden, ASM, 1992.
296. J. E. Trumann: "Steinless Steels," in F. B. Pickering (ed.): *Materials Science and Technology,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 527.
297. J. Charles, S. Bernhardsson: *Duplex Stainless Steels '91,* Les Editions de Physique, Les Ulis Cedex A 1991.
298. *Metals Handbook,*9th ed., vol. 3, "Properties and Selection: Stainless Steels, Tool Materials and Special-Purpose Metals," American Society for Metals, Ohio 1980.
299. H. P. Hougardy: *Umwandlung und Gefüge unlegierter Stähle. Eine Einführung,* 1nd ed., Verlag Stahleisen, Düsseldorf 1990.
300. W. Dahl: "Grundlagen," in *Werkstoffkunde Stahl,* vol. 1, Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1984, pp. 235 – 400.
301. W. Dahl: "Fundamentals," in *Steel–A Handbook for Materials Research and Engineering,* vol. 1, Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1992, pp. 203 – 378.
302. W. Dahl, *Stahl u. Eisen* **104** (1984) 267 – 274.
303. W. Dahl (ed.): *Grundlagen des Festigkeits- und Bruchverhaltens,* Berichte, gehalten im Kontaktstudium "Werkstoffkunde Eisen und Stahl," Verlag Stahleisen, Düsseldorf 1974.
304. W. Dahl (ed.): *Grundlagen der Festigkeit, der Zähigkeit und des Bruchs,* Berichte, gehalten im Kontaktstudium "Werkstoffkunde Eisen und Stahl," Verlag Stahleisen, Düsseldorf 1983.
305. A. Troost: *Einführung in die allgemeine Werkstoffkunde metallischer Werkstoffe,* Bibliographisches Institut, Mannheim 1980.
306. H. Böhm: *Einführung in die Metallkunde,* Bibliographisches Institut, Mannheim 1968, no. 196/196 a.
307. P. Haasen: *Physikalische Metallkunde,* Springer Verlag, Berlin 1974.
308. B. Ilschner: *Werkstoffwissenschaften,* Springer Verlag, Berlin 1982.
309. R. E. Reed-Hill: *Physical metallurgy principles,* 2nd ed., Van Nostrand, New York 1963.
310. W. Dahl, H. Rees: *Die Spannungs-DehnungsKurve von Stahl,* Verlag Stahleisen, Düsseldorf 1976.
311. P. Belche: "Einfluß von Prüftemperatur und Dehngeschwindigkeit auf das Fließverhalten von Stählen mit besonderer Berücksichtigung des Werkstoffverhaltens bei statischer Beanspruchung," Dissertation, RWTH Aachen, Aachen 1983.
312. K. Lücke in [303], pp. 5 – 18.
313. K. Lücke, G. Gottstein in [303], pp. 19 – 32.
314. A. G. Guy: *Physical Metallurgy for Engineers,* Addison-Wesley Publ. Co., Reading, Mass. 1962.
315. E. Siebel, *Ber. Werkstoffaussch. d. Fachaussch. d. VDEh* (1925) no. 71, 1 – 3.
316. W. Dahl, W. Hesse, A. Krabiell, *Stahl u. Eisen* **103** (1983) 87 – 90.
317. W. Pitsch in [303], pp. 41 – 52.
318. A. H. Cottrell: *Dislocations and plastic flow in crystals,* 3rd ed., Clarendon Press, Oxford 1958.
319. W. Dahl, E. Lenz, *Arch. Eisenhüttenwes.* **46** (1975) 119 – 125.
320. A. S. Keh, Y. Nakada, W. C. Leslie in A. R. Rosenfield et al. (eds.): *Dislocation dynamics,* McGraw-Hill, New York 1968, pp. 381 – 408.
321. A. Krabiell, W. Dahl, *Arch. Eisenhüttenwes.* **52** (1981) 429 – 436.
322. U. Reichel: "Analyse des Verfestigungsverhaltens von Stählen anhand phänomenologischer Modelle," Dissertation, RWTH Aachen, Aachen 1988.
323. U. Reichel, W. Dahl, *steel research* **59** (1988) no. 4, 171 – 178.
324. X. Fang: "Experimentelle und theoretische Untersuchungen zum Einfluß des Gefüges und der Temperatur auf das Verfestigungsverhalten von Stählen," Dissertation, RWTH Aachen, Aachen 1992.

325. H. Wiegels, R. Herbertz, *Stahl u. Eisen* **101** (1981) 1127 – 1132.
326. H. Wiegels, R. Herbertz, *Stahl u. Eisen* **101** (1981) 1487 – 1492.
327. Deutscher Normenausschuß (ed.): *DIN-Taschenbuch 19,* "Materialprüfnormen für metallische Werkstoffe," 9th ed., Berlin 1981; DIN 1602.
328. Deutscher Normenausschuß (ed.): *DIN-Taschenbuch 19,* "Materialprüfnormen für metallische Werkstoffe," 9th ed., Berlin 1981; DIN 50 110, ed. Febr. 1962.
329. A. Troost, D. Schaefer, *Konstruktion* **14** (1962) 436 – 441.
330. Deutscher Normenausschuß (ed.): *DIN-Taschenbuch 19,* "Materialprüfnormen für metallische Werkstoffe," 9th ed., Berlin 1981; DIN 50 133, Part 1, 2nd ed. Dez. 1972; DIN 50 351, ed. Jan. 1973; DIN 50 103, Part 1, 2nd ed. Dez. 1972.
331. C. Straßburger: *Entwicklungen zur Festigkeitssteigerung der Stähle unter besonderer Berücksichtigung der unlegierten und mikrolegierten Baustähle,* Verlag Stahleisen, Düsseldorf 1976.
332. F. B. Pickering: *Physical metallurgy and the design of steels,* American Society for Metals, London 1978.
333. R. W. K. Honeycombe: *Steel, microstructure and properties,* Arnold, London 1981.
334. W. C. Leslie: *The physical metallurgy of steels,* Hemisphere Publ. Co., Washington 1981.
335. H. Hornbogen, *Z. Metallkd.* **64** (1973) 867 – 870.
336. W. Dahl, H. Rees in [303], pp. 53 – 70.
337. W. Dahl in [304], vol. 1, pp. 208 – 225.
338. W. Dahl, H. Rees in [349], pp. 100 – 180.
339. W. Dahl, H. Rees, *Arch. Eisenhüttenwes.* **49** (1978) 25 – 29.
340. H. Gleiter in [303], pp. 71 – 83.
341. A. S. Keh in J. B. Newkirk, J. H. Wernick (eds.): *Direct Observation of Imperfections in Crystals,* Interscience Publishers, New York 1962 (Metall. Soc. Conf. Proc. 14), pp. 213 – 238.
342. E. Hornbogen in [303], pp. 85 – 100.
343. E. Hornbogen, *Stahl u. Eisen* **93** (1973) 822 – 826.
344. K. J. irvine, *Iron Steel* **44** (1971) no. 1, 31 – 38.
345. J. Flügge, W. Heller, E. Stolte, W. Dahl, *Arch. Eisenhüttenwes.* **47** (1976) 635 – 640.
346. P. G. Winchell, M. Cohen, *ASM Trans. Q.* **55** (1962) 347 – 361.
347. U. Reichel et al., *steel research* **60** (1989) no. 10, 464 – 468.
348. In [305], pp. 28 – 57.
349. W. Dahl, H. Rees: "Festigkeit metallischer Werkstoffe," *Berichte zum Symposium der Deutschen Gesellschaft für Metallkunde,* Bad Nauheim 1974.
350. D. Aurich in E. Macherauch, V. Gerold (eds.): *Bruchvorgänge in metallischen Werkstoffen,* Werkstofftechnische Verlagsgesellschaft, Karlsruhe 1978.
351. G. Henry, D. Horstmann: *Faktographie und Mikrofraktographie,* Max-Planck-Institut für Eisenforschung, Düsseldorf, Institut de Recherches de la Sidérurgie Française, Brussels, Düsseldorf 1979.
352. *Metals handbook,* B ed., vol. 9, American Society for Metals, Metals Park, Ohio 1974.
353. L. Engel, H. Klingele (eds.): *An atlas of metal damage,* 2nd ed., C. Hanser Verlag, Munich 1981.
354. W. Dahl, K. Kühne, H.-Chr. Zeislmair in [304], vol. 1, pp. 321 – 380.
355. M. A. M. Halim: "Einfluß der Mikrostruktur und des Spannungszustandes auf die Bruchvorgänge von technischen Stählen," Dissertation, RWTH Aachen, Aachen 1989.
356. D. Holland: "Einfluß des Spannungszustandes auf die Vorgänge beim Gleitbruch von Baustählen," Dissertation, RWTH Aachen, Aachen 1992.
357. W. Dahl, *Stahl u. Eisen* **101** (1981) 967 – 976.
358. W. Dahl: "Stahl, heutiger Stand und zukünftige Entwicklung der Eigenschaften und Prüfverfahren," *VMPA-Tagung,* Munich 1990, pp. 41 – 54.
359. J. Degenkolbe in [303], pp. 203 – 218.
360. J. Degenkolbe in [304], vol. 1, pp. 381 – 453.
361. W. Heller, K. J. Kremer, *Stahl u. Eisen* **89** (1969) 1005 – 1018.
362. K. Wellinger, D. Blind, *Bänder Bleche Rohre* **8** (1967) 550 – 556.
363. K.-H. Schwalbe: *Bruchmechanik metallischer Werkstoffe,* C. Hanser Verlag, Munich 1980.
364. H. Blumenauer: *Bruchmechanik,* Leipzig 1973.
365. D. Broek: *Elementary engineering fracture mechanics,* Sijthoff & Nordhoff, Alphen aan den Rijn 1978.
366. W. Dahl, H.-C. Zeislmair: *Anwendung der Bruchmechanik auf Baustähle,* Verlag Stahleisen, Düsseldorf 1983.
367. R. Hubo: "Bruchmechanische Untersuchungen zum Einsatz von Stählen unterschiedlicher Festigkeit und Zähigkeit," Dissertation, RWTH Aachen, Aachen 1990, *Fortschr. Ber. VDI Z. Reihe 18* **80** (1990).

368. A. Krabiell, W. Dahl, *Arch. Eisenhüttenwes.* **53** (1982) 225 – 230.
369. W. Dahl: "Einfluß der Beanspruchungsgeschwindigkeit auf Festigkeit und Bruchverhalten von Stählen," *Vortragsveranstaltung des DVM-Arbeitskreises "Bruchvorgänge,"* Aachen 1992.
370. W. Hesse: "Zur Auswirkung hoher Beanspruchungsgeschwindigkeiten auf das Versagungsverhalten von Baustählen unter einsinniger Zugbeanspruchung," Dissertation, RWTH Aachen, Aachen 1986.
371. K. Brüninghaus: "Zum Einfluß der Beanspruchungsgeschwindigkeit auf das Verhalten von Stählen bei stabiler Rißverlängerung, Dissertation, RWTH Aachen, Aachen 1988, *Fortschr. Ber. VDI Z. Reihe 18* **60** (1988).
372. K. Brüninghaus et al., *Nucl. eng. Des.* **127** (1991) 95 – 102.
373. A. G. Heuser: "Beurteilung des Versagensverhaltens von Schweißverbindungen hochfester Baustähle mit Hilfe bruchmechanischer Methoden," Dissertation, RWTH Aachen, Aachen 1988, *Fortschr. Ber. VDI Z. Reihe 18* **48** (1988).
374. Verein Deutscher Eisenhüttenleute, Kommission der Europäischen Gemeinschaften (ed.): Vergleich der Aussagen verschiedener Sprödbruchprüfverfahren, Schlußbericht zum Forschungsvertrag no. 6210-55/1/510, Luxemburg 1975 (EUR 5391 d).
375. W. Dahl, H. Hengstenberg, H. Behrens, *Stahl u. Eisen* **88** (1968) 578 – 596.
376. K. Kühne, W. Dahl, *Arch. Eisenhüttenwes.* **54** (1983) 439 – 444.
377. H. Dünnewald-Arfmann: "Anwendung des Konzepts der mikroskopischen Spaltbruchspannung zur Beurteilung des Bruchverhaltens von Stählen," Dissertation, RWTH Aachen, Aachen 1988.
378. J. Degenkolbe, B. Müsgen, *Arch. Eisenhüttenwes.* **44** (1973) 769 – 774.
379. C. Straßburger, D. Schauwinhold, *Stahl u. Eisen* **87** (1967) 792 – 798.
380. W. Dahl: "Erörterungsbeitrag zu W. heller, E. Stolte," *Stahl u. Eisen* **90** (1970) 909 – 916.
381. L. Meyer, H. de Boer, *Thyssen Tech. Ber.* **9** (1977) 20 – 29.
382. C. Straßburger, L. Meyer, F. Heisterkamp, *Bänder Bleche Rohre* **12** (1971) 153 – 159.
383. J. Degenkolbe, B. Müsgen, *Oerlikon Schweißmitt.* **37** (1979) no. 88, 27 – 35.
384. E. Spetzler, J. Wendorff, *Radex Rundsch.* 1976, no. 1, 595 – 608.
385. J. Degenkolbe, G. Kalwa, K. Kaup, *Stahl u. Eisen* **108** (1988) no. 11, 531.
386. D. Schauwinhold et al., *Proc. Int. Conf. Secondary Metallurgy,* Aachen 1987, pp. 3 – 18.
387. "Aus der Arbeit des Vereins Deutscher Eisenhüttenleute," *Eisenhüttentag 1989,* p. 4.
388. T. Glasman, *J. Iron Steel Inst. London* **203** (1965) 1038 – 1039.
389. L. Meyer, *Stahl u. Eisen* **101** (1981) 483 – 491.
390. K. Lorenz et al., *Stahl u. Eisen* **101** (1981) 593 – 600.
391. K. J. Irvine, *Iron Steel* **44** (1971) no. 1, 31 – 38; in [331].
392. W. Dahl in [303], pp. 101 – 110.
393. G. Robiller, C. Straßburger, *Mater. Prüf.* **11** (1969) 89 – 95.
394. W. Dahl (ed.): *Verhalten von Stahl bei schwingender Beanspruchung,* Berichte, gehalten im Kontaktstudium "Werkstoffkunde Eisen und Stahl, III," Verlag Stahleisen, Düsseldorf 1978.
395. D. Munz, K.-H. Schwalbe, P. Mayr: *Dauerschwingverhalten metallischer Werkstoffe,* Vieweg, Braunschweig 1971.
396. W. Günther (ed.): *Schwingfestigkeit,* VEB Verlag für Grundstoffindustrie, Leipzig 1973.
397. K. E. Hagedorn, *Stahl u. Eisen* **98** (1978) 102 – 107.
398. W. Schmidt in [394], pp. 1 – 22.
399. M. Hempel, *Arch. Eisenhüttenwes.* **38** (1967) 446 –455.
400. D. Dengel in [394], pp. 23 – 46.
401. K.-H. Schwalbe in [394], pp. 180 – 189.
402. M. Hempel, *Draht* **11** (1960) 429 – 437.
403. A. Randak, A. Stanz, W. Verderber, *Stahl u. Eisen* **92** (1972) 981 – 993.
404. Verein Deutscher Eisenhüttenleute (ed.): *Leitfaden für eine Betriebsfestigkeitsrechnung,* Verlag Stahleisen, Düsseldorf 1977.
405. W. Dahl, W. Pitsch (ed.). *Festigkeits- u. Bruchverhalten bei höheren Temperaturen,* vols. 2, Berichte, gehalten im Kontaktstudium "Werkstoffkunde Eisen und Stahl IV," Verlag Stahleisen, Düsseldorf 1980.
406. B. Ilschner: "Hochtemperatur-Plastizität," *Reine und angewandte Metallkunde in Einzeldarstellung,* **vol. 23,** Springer Verlag, Berlin 1973.
407. W. Pitsch in [405], vol. 1, pp. 149 – 179.
408. Deutsche Gesellschaft für Metallkunde (DGM) (ed.): *Warmumformung und Warmfestigkeit,* Oberursel 1976.

409. H. G. Müller in Verein Deutscher Eisenhüttenleute (ed.): *Grundlagen der bildsamen Formgebung,* Verlag Stahleisen, Düsseldorf 1966, pp. 146 – 161.
410. C. Rossard, *Mét. Corros. Ind.* **35** (1960) 102 – 115, 140 – 153, 190 – 205, 9 Taf.
411. B. Ilschner, *Z. Werkstofftech.* **2** (1971) 123 – 127.
412. A. von den Steinen: *Gasturbinen, Probleme und Anwendungen,* Verlag Stahleisen, Düsseldorf 1967, pp. 131 – 146.
413. K. E. Heusler, *Z. Electrochem.* **62** (1958) 582.
414. T. Ramachandran, K. Bohnenkamp, *Werkst. Korros.* **30** (1979) 43.
415. E. G. Dafft, K. Bohnenkamp, H.-J. Engell, *Corros. Sci.* **19** (1979) 591.
416. M. Stratmann, *Ber. Bundesnges. Phys. Chem.* **94** (1990) 626.
417. U. R. Evans, C. A. J. Taylor, *Corros. Sci.* **12** (1972) 227.
418. K. Bohnenkamp, *Arch. Eisenhüttenwes.* **47** (1976) 751.
419. G. Schikorr, *Werkst. Korros.* **14** (1963) 69.
420. G. Herbsleb, H.-J. Engell, *Z. Elektrochem.* **65** (1961) 881.
421. C. Wagner, *Ber. Bunsenges. Phys. Chem.* **77** (1973) 1090.
422. N. Sato, G. Okamoto, *J. Electrochem. Soc.* **110** (1963) 605.
423. Y. M. Kolotyrkin, *Z. Elektrochem.* **62** (1958) 664.
424. K. Osazawa, H.-J. Engell, *Corros. Sci.* **6** (1966) 389.
425. K. J. Vetter, H.-H. Strehblow, *Ber. Bunsenges. Phys. Chem.* **74** (1970) 1024.
426. J. E. Truman in J. Foct, A. Hendry (eds.): *Proc. I. Int. Conf. High Nitrogen Steels,* Institute of Metals, London 1989, p. 225.
427. R. Stefec, F. Franz, A. Holecek, *Werkst. Korros.* **30** (1979) 189.
428. N. Lukomski, K. Bohnenkamp, *Werkst. Korros.* **30** (1979) 482.
429. J. W. Oldfield, W. H. Sutton, *Br. Corros. J.* **13** (1978) 13.
430. K. Bungardt, E. Kunze, E. Horn, *Arch. Eisenhüttenwes.* **29** (1958) 193.
431. J. J. Demo, A. P. Bond, *Corrosion (Houston)* **31** (1975) 21.
432. B. Strauß, H. Schottky, J. Hinnüber, *Z. Anorg. Chem.* **188** (1930) 309.
433. I. Class, H. Gräfe, *Werkst. Korros.* **11** (1960) 529.
434. G. Herbsleb, *Werkst. Korros.* **19** (1968) 406.
435. D. L. Davidson, F. F. Lyle, *Corrosion (Houston)* **31** (1975) 135.
436. H.-J. Engell in J. C. Scully (ed.): *The Theory of Stress Corrosion Cracking in Alloys,* published by NATO Scientific Affairs Division, Brussels 1971, p. 86.
437. E. Brauns, H. Ternes, *Werkst. Korros.* **19** (1968) 1.
438. M. Ahlers, E. Riecke, *Corros. sci.* **18** (1978) 21.
439. M. Marek, R. F. Hochmann, *Corrosion (Houston)* **27** (1971) 361.
440. H. E. Hänninen, *Int. Met. Rev.* **24** (1979) 85.
441. M. O. Speidel, *Corrosion (Houston)* **33** (1977) 199.
442. A. Bäumel, H.-J. Engell, *Arch. Eisenüttenwes.* **32** (1961) 379.
443. K. Bohnenkamp, *Arch. Eisenhüttenwes.* **39** (1968) 361.
444. E. Wendler-Kalsch, *Werkst. Korros.* **31** (1980) 534.
445. W. Rädecker, B. N. Mishra, *Werkst. Korros.* **17** (1966) 193.
446. J. Küpper, H. Erhart, H. J. Grabke, *Corros. Sci.* **21** (1981) 227.
447. E. Riecke, *Werkst. Korros.* **29** (1978) 106.
448. E. Riecke, *Met. Corros. Proc. Int. Congr. Met. Corros. 8th* 1981, no. 1, 605.
449. J. P. Hirth, *Metall. Trans.* **11 A** (1980) 861.
450. R. A. Oriani, *Ber. Bunsenges. Phys. Chem.* **76** (1972) 848.
451. L. S. Darken, R. W. Gurry, *J. Am. Chem. soc.* **67** (1945) 1398.
452. J. Spencer, O. Kubaschewski, *CALPHAD Comput. Coupling Phase Diagrams Thermochem.* **2** (1978) 147.
453. B. E. F. Fender, F. D. Riley, *J. Phys. Chem. Solids* **30** (1969) 793.
454. A. Rahmel: in *Chemical Metallurgy of Iron and Steel,* ISI Publ., no. 146, London 1973, p. 395.
455. H.-J. Engell, *Acta Metall.* **6** (1958) 439.
456. K. Hauffe, H. Pfeiffer, *Z. Metallkd.* **44** (1953) 27.
457. E. T. Turkdogan, W. M. McKewan, L. Zwell, *J. Phys. Chem.* **69** (1965) 327.
458. F. S. Pettit, R. Yinger, J. B. Wagner, Jr., *Acta Metall.* **8** (1960) 617.
459. H. J. Grabke, *Ber. Bunsenges. Phys. Chem.* **69** (1965) 48.
460. H. J. Grabke, K. J. Best, A. Gala, *Werkst. Korros.* **21** (1970) 911.
461. C. Wagner, *Z. Phys. Chem. Abt. B* **21** (1933) 25; **32** (1936) 447.
462. K. Bohnenkamp, H.-J. Engell, *Arch. Eisenhüttenwes.* **33** (1962) 359.

463. H. Meurer, H. Schmalzried, *Arch. Eisenhüttenwes.* **42** (1971) 87.
464. W. Koenigsmann, F. Oeters, *Werkst. Korros.* **29** (1978) 10.
465. L. A. Morris, W. W. Smeltzer, *Acta Metall.* **15** (1967) 1591.
466. A. D. Dalvi, W. W. Smeltzer, *J. Electrochem. Soc.* **121** (1974) 386.
467. C. Wagner, *Corros. Sci.* **9** (1969) 91.
468. G. C. Wood, I. G. Wright, T. Hodgkiess, D. P. Whittle, *Werkst. Korros.* **21** (1970) 900.
469. K. A. Hay, F. G. Hicks, D. R. Holmes, *Werkst. Korros.* **21** (1970) 917.
470. W. C. Hagel, *Corrosion (Houston)* **21** (1965) 316.
471. P. tomaszewicz, G. R. Wallwork, *Corrosion (Houston)* **40** (1984) 152.
472. M. J. McNallan, W. W. Liang, S. H. Kim, C. T. King in R. A. Rapp (ed.): "High Temperature Corrosion," *NACE* 1983, 316.
473. H. J. Grabke, E. M. Müller, *Werkst. Korros.* **41** (1990) 227.
474. D. Bramhoff, H. J. Grabke, E. Reese, H. P. Schmidt, *Werkst. Korros.* **41** (1990) 303.
475. E. Reese, H. J. Grabke, *Werkst. Korros.* **43** (1992) 547.
476. G. H. Geiger, R. L. Levin, J. B. Wagner, Jr., *J. Phys. Chem. Solids* **27** (1966) 947.
477. B. Swaroop, J. B. Wagner, Jr., *Trans. Metall. Soc. AIME* **239** (1967) 1215.
478. K. Hauffe, A. Rahmel, *Z. Phys. Chem. (Leipzig)* **199** (1951) 152.
479. R. A. Meussner, C. E. Birchenall, *Corrosion (Houston)* **13** (1957) 677.
480. E. T. Turkdogan, *Trans. Metall. Soc. AIME* **242** (1968) 1665. W. L. Worrell, E. T. Turkdogan, *Trans. Metall. Soc. AIME* **242** (1968) 1673.
481. W. Auer, *Met. Corros. Proc. Int. Congr. Met. Corros. 8th* 1981, no. 2, 649.
482. A. Rahmel, *Oxid. Met.* **9** (1975) 401.
483. T. Flatley, N. Birks, *J. Iron Steel Inst.* **209** (1971) 523.
484. R. P. Salisburg, N. Birks, *J. Iron Steel Inst.* **209** (1971) 534.
485. B. A. Gordon, V. Nagarajan, *Oxid. Met.* **13** (1979) 197.
486. R. A. Perkins in R. A. Rapp (ed.): *High temperature corrosion,* NACE, Houston, Texas 1981.
487. F. H. Stott, F. M. F. Chong in D. B. Meadowcroft, M. J. Manning (eds.): *Proc. Conf. Corrosion Resistant Materials for Coal Conversion Systems,* Applied Science, London 1983, p. 491.
488. J. Stringer, "Sulfidation as Industrial Problem and Limiting Factor," Workshop on The High Temperature Corrosion of Alloys in Sulfur Containing Environments, Petten, Dec. 1985.
489. R. E. Lobnig, H. J. Grabke, *Corros. Sci.* **30** (1990) 1045.
490. F. Neumann, U. Wyss, *HTM Härterei Tech. Mitt.* **25** (1970) 253.
491. C. A. Stickels, C. M. Mack, M. Brachaczek, *Metall. Trans.* **11 B** (1980) 471.
492. H. J. Grabke, U. Gravenhorst, W. Steinkusch, *Werkst. Korros.* **27** (1976) 291.
493. A. Schnaas, H. J. Grabke, *Oxid. Met.* **12** (1978) 387.
494. H. J. Grabke, *Arch. Eisenhüttenwes.* **46** (1973) 603.
495. H. J. Grabke, *HTM Härterei Tech. Mitt.* **45** (1990) 110.
496. H. J. Grabke, *Mater. Sci. Eng.* **42** (1980) 91.
497. J. Wünning, *Härterei Tech. Mitt.* **23** (1968) 101; *HTM Härterei Tech. Mitt.* **39** (1984) 50.
498. K. Ledjeff, A. Rahmel, M. Schorr, *Werkst. Korros.* **30** (1979) 767; **31** (1980) 83.
499. G. M. Lai in: *High Temperature Corrosion of Engineering Alloys,* American Society for Metals, Materials Park, Ohio 1990, pp. 68 – 71.
500. J. C. Nava Paz, H. J. Grabke, *Oxid. Met.* **39** (1993) 437.
501. H. J. Grabke, R. Krajak, J. C. Nava Paz, *Corros. Sci.* **35** (1993) 1141.
502. H. J. Grabke, R. Krajak, E. M. Müller-Lorenz, *Werkst. Korros.* **44** (1993) 89.
503. H. J. Grabke, C. B. Bracho-Troconis, E. M. Müller-Lorenz, *Werkst. Korros.* **45** (1994) 215.
504. H. J. Grabke, *Ber. Bunsenges. Phys. Chem.* **72** (1968) 533, 541.
505. I. Aydin, H.-E. Bühler, A. Rahmel, *Werkst. Korros.* **31** (1980) 675.
506. W. Steinkusch, *Werkst. Korros.* **27** (1976) 91.
507. K. Lücke, *Arch. Eisenhüttenwes.* **25** (1954) 181.
508. A. Mayer, *Stahl u. Eisen* **83** (1963) 1169.
509. W. Oelsen, K.-H. Sauer, *Arch. Eisenhüttenwes.* **38** (1967) 141.
510. W.-D. Jentzsch, S. Böhmer, *Neue Hütte* **11** (1974) 647.
511. R. M. Hudson, *Trans. Metall. Soc. AIME* **230** (1964) 1138.
512. G. J. Kor, J. F. van Rumpt, *J. Iron Steel Inst.* **207** (1969) 1377.
513. G. A. Nelson: *Steels for Hydrogen Service at Elevated Temperatures and Pressures in Petroleum Refineries and Petrochemical Plants,* API Publication 941, Washington, May 1983.

514. W. Pepperhoff: "Physikalische Eigenschaften," in Verein Deutscher Eisenhüttenleute (ed.): *Werkstoffkunde Stahl,* vol. 1, "Grundlagen," Düsseldorf 1984, pp. 401 – 433. W. Pepperhoff: "Physical Properties," in Verein Deutscher Eisenhüttenleute (ed.): *Steel,* vol. 1, "Fundamentals," Düsseldorf 1992, pp. 379 – 411.
515. *Landolt-Börnstein,* 6th ed., **4**, 2 a, 131 – 300.
516. K. K. Kelley, *J. Chem. Phys.* **11** (1943) 16 – 18.
517. D. C. Wallace, P. N. Siddles, G. C. Danielson, *J. Appl. Phys.* **31** (1960) 168 – 176.
518. M. Braun: "über die spezifische Wärme von Eisen, Kobalt und Nickel im Bereich hoher Temperaturen," Dissertation, Universität Köln 1964.
519. W. Bendick, W. Pepperhoff, *Acta Metall.* **30** (1982) 679 – 684.
520. H. M. Ledbetter, R. P. Reed, *J. Phys. Chem. Ref. Data* **2** (1974) 531 – 618.
521. K. M. Koch, W. Jellinghaus: *Einführung in die Physik magnetischer Werkstoffe,* Deuticke, Wien 1957.
522. E. Kneller: *Ferromagnetismus,* Springer Verlag, Berlin 1962.
523. R. Kohlhaus, F. Richter, *Arch. Eisenhüttenwes.* **33** (1962) 291 – 299.
524. W. Oelsen, *Stahl u. Eisen* **69** (1949) 468 – 475.
525. R. M. Bozorth : *Ferromagnetism,* Van Nostrand, New York 1955.
526. Y. Endoh, Y. Ishikawa, *J. Phys. Soc. Jpn.* **30** (1971) 1614 – 1627.
527. L. Kaufman, M. Cohen, *J. Met.* **8** (1956) 1393 –1400.
528. H.-H. Ettwig, W. Pepperhoff, *Phys. Status Solidi A* **23** (1974) 105 – 111.
529. W. Bendick, H.-H. Ettwig, F. Richter, W. Pepperhoff, *Z. Metallkd.* **68** (1977) 103 – 107.
530. F. Richter, W. Pepperhoff, *Arch. Eisenhüttenwes.* **47** (1976) 45 – 50.
531. S. Denis, D. Farias, A. Simon, *ISIJ Int.* **32** (1992) no. 3, 316 – 325.
532. H. K. D. H. Badheshia, *Met. Sci.* **16** (1982) 159.
533. Verein Deutscher Eisenhüttenleute (ed.): "Applications," *Steel – A Handbook for Materials Research and Engineering,* **vol. 2**, Verlag Stahleisen, Springer Verlag, Berlin 1992.
534. K.-E. Thelning: *Steel and its Heat Treatment,* 2nd ed., Butterworths, London 1984.
535. G. Krauss: *Steels: Heat Treatment and Processing Principles,* ASM Publication, Materials Park, Ohio 1990.
536. J. R. Davis: "Heat Treating," *ASM Handbook,* **vol. 4**, ASM International, Materials Park, Ohio 1991.
537. P. M. Unterweiser, H. E. Boyer, J. J. Kubbs: *Heat Treater's Guide: Standard Practices and Procedures for Steel,* ASM Publication, Materials Park, Ohio 1982.
538. Vergütungsstähle – Technische Lieferbedingungen für Edelstähle, DIN EN 10 083-1,part 1, Normenausschuß Eisen und Stahl (FES) im DIN Deutsches Institut für Normung e.V., Berlin 1991.
539. H. P. Hougardy: "Heat Treatability, Heat Treatment, and Properties," in Verein Deutscher Eisenhüttenleute (ed.): "Fundamentals," *Steel – A Handbook for Materials Research and Engineering,* vol. 1, Springer Verlag, Berlin, and Verlag Stahleisen, Düsseldorf 1992, p. 461.
540. D. J. Naylor, W. T. Cook: "Heat Treated Engineering Steels," in "Constitution and Properties of Steels", *Materials Science and Technology. A Comprehensive Treatment,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 433.
541. H. Othani: "Processing – Conventional Heat Treatments," in F. B. Pickering (ed.): "Constitution and Properties of Steel," *Materials Science and Technology. A Comprehensive Treatment,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 149.
542. W. Heller, J. Flügge, *Stahl u. Eisen* **100** (1980) 1240 – 1246.
543. V. Hauk, H. P. Hougardy, E. Macherauch, H.-D. Tiet: *Residual Stresses,* DGM Informationsgesellschaft, Oberursel 1993.
544. G. Totten: "Quenching and Distortion Control," ASM Conference Book, 1992.
545. B. Liscic, H. M. Tensi, W. Luty: *Theory and Technology of Quenching,* vol. 5, Springer Verlag, Berlin 1992.
546. K. M. Vedula, R. W. Heckel, *Metall. Trans.* **1** (1970) 9 – 18.
547. H.-J. Gasterich, K. E. Hagedorn, R. Kaspar, O. Pawelski, *steel research* **64** (1993) 414 – 419.
548. R. Kaspar, P. Flüß, O. Pawelski, *steel research* **60** (1989) no. 12, 566.
549. S. Engineer, B. Huchtemann, V. Schüler, *steel research* **58** (1987) 369 – 376.
550. G. L. Dunlop, C.-J. Carlsson, G. Frimodig, *Metall. Trans.* **9 A** (1978) 261 – 266.
551. R. Kaspar, J. S. Distl, O. Pawelski, *steel research* **59** (1989) 421 – 425.

552. I. Kozasu: "Processing – Thermomechanical Controlled Processing," in F. B. Pickering (ed.): "Constitution and Properties of Steels," *Materials Science and Technology,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 183.
553. R. Kaspar, O. Pawelski, *steel research* **56** (1985) 477 – 481.
554. R. Kaspar, W. Kapellner, O. Pawelski, *HTM Härterei-Tech. Mitt.* **41** (1986) 177 – 185.
555. J. Kaspar et al., *steel research* **58** (1987) 271 –277.
556. R. K. Amin, F. B. Pickering: "Ferrite Formation from Thermo-Mechanically Processed Austenite; Austenite Grain Coarsening and the Effect of Thermo-Mechanical Processing on Austenite Recrystallization; The Charles Hatchett Award Papers," preprinted from the Proceeding of the International Conference on the Thermo-Mechanical Processing of Microalloyed Austenite, TMS-AIME, 1982.
557. R. Kaspar, W. Kapellner, O. Pawelski, *Z. Metallkd.* **78** (1987) 569 – 575.
558. S. S. Hansen, J. B. V. der Sande, M. Cohen, *Metall. Trans.* **11 A** (1980) 387 – 401.
559. R. Kaspar, A. Streißelberger, O. Pawelski: "Thermomechanical Treatment of Ti- and Nb-Mo-Microalloyed Steels in Hot Strip Rolling, Thermomechanical Processing of Microalloyed Austenite," Proceedings of the International Conference on the "Thermomechanical Processing of Microalloyed Austenite," Pittsburgh, PA, 1981, p. 555.
560. I. Weiss, J. J. Jonas, *Metall. Trans.* **10 A** (1979) 831 –839.
561. M. J. Luton, R. Dorvel, R. A. Petkovic, *Metall. Trans.* **11 A** (1980) 411 – 420.
562. P. Hodgeson: *Quenching and Carburising,* Institute of Materials Publications, London 1993.
563. J. Grosch et al.: *Einsatzhären,* vol. 356, Expert-Verlag, Renningen – Malmsheim 1994.
564. "Eigenschaften von Werkstoffen mit Verschleiß-schutzschichten," *HTM Härterei Tech. Mitt.* **37** (1982) 158 – 195.
565. R. W. Cahn, P. Haasen: *Physical Metallurgy,* North-Holland Physics Publishing, Amsterdam 1983.
566. J. A. Charles, G. C. Smith: Advances in Physical Metallurgy – A collection of invited papers presented to mark the 70th birthday of Professor Sir Alan Cottrell, Institute of Materials Publications, London 1990.
567. H. Hansen, A. R. Jones, T. Leffers: "Recrystallization and Grain Growth of Multi-Phase and Particle Containing Materials," Proceedings of the 1st Riso International Symposium on Metallurgy and Materials Science, Sep. 8 – 12, 1980, Riso National Laboratory, 1980.
568. R. C. Hudd: "Processing – Cold Working and Annealing," in F. B. Pickering (ed.): "Constitution and Properties of Steels," *Materials Science and Technology. A Comprehensive Treatment,* vol. 7, VCH Verlagsgesellschaft, Weinheim 1992, p. 219.
569. SteCal 2.0, ASM Software.
570. H. P. Hougardy, K. Yamazaki, *steel research* **57** (1986) no. 9, 466 – 471.
571. A. Kern, W. Reif, U. Schriever: "Computer Simulation of the Austenite Grain Coarsening in Microalloyed Steel Plate Using the Monte Carlo Technique," in: *Grain Growth in Polycrystalline Materials,* Int. Conf. Rome 1991, *Mat. Sci. Forum.* **94 – 96** (1992) 709 – 714.
572. A. Kern, U. Schriever: "Computer Simulation of the Austenite Grain Growth and Recrystallization During Thermomechanical Rolling of Microalloyed HSLA-steels," in: *Recrystallization,* 2nd Int. Conf. of Recrystallization and Related Phenomena, *Mat. Sci. Forum.* **113 – 115** (1993) 349 – 354.
573. A. K. J. Degenkolbe, B. Müsgen, U. Schriever, *ISIJ Int.* **32** (1992) no. 3, 387 – 394.
574. W. Dahl in Verein Deutscher Eisenhüttenleute (ed.): "Fundamentals," *Steel–A Handbook for Materials Research and Engineering,* **vol. 1, part C 1,** Springer Verlag, Berlin 1992.
575. B. Müsgen, H. de Boer, H. Fröber, J. Petersen: "Anwendung," *Werkstoffkunde Stahl,* **vol. 2, part D 2,** Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1985.
576. H. Weise, W. Krämer, W. Bartels, W.-D. Brand in [575], part D 3.
577. F. Leonhardt: *Spannbeton,* 4th ed., Springer Verlag, Berlin 1980.
578. T. S. Robertson, *J. Iron Steel Inst. (London)* **175** (1953) 361 – 374.
579. W. S. Pellini, P. P. Puzak: NRL Rep. 5920, March 15, 1963.
580. W. S. Pellini, P. P. Puzak: NRL Rep. 6030, Nov. 5, 1963.
581. D. Uwer, H. Höhne, *Schweißen Schneiden* **43** (1991) 282 – 287; IIW Doc. IX-1631-91.
582. J. Degenkolbe, D. Uwer, H. G. Wegmann, *Thyssen Tech. Ber.* **17** (1985) 57 – 73; IIW Doc. IX-1336-84.

583. J. C. M. Farrar, R. D. Dolby: *Lamellar Tearing in Welded Steel Fabrication,* The Welding Inst., Abington 1972.
584. B. Müsgen, H. Baumgardt, H. de Boer, W. Janzon: 17th Offshore Technical Conference, May 1985, Houston;OTC 5072.
585. B. Müsgen, *Metal Constr.* **17** (1985) 495 – 500.
586. B. Müsgen, J. Degenkolbe, *Bänder Bleche Rohre* **13** (1972) 504 – 510.
587. DASt-Richtlinie 007: Lieferung, Verarbeitung und Anwendung wetterfester Baustähle.
588. M. Economopoulos et al.: "Application of the Tempcore Process to the Fabrication of High Yield Strength Concret-Reinforcing Bars," *Metall. Rep. CRM* **45,** Dec. 1975.
589. G. Tacke, K. Forch, K. Sartorius, A. von den Steinen, K. Vetter, "Anwendung," *Werkstoffkunde Stahl,* **vol. 2, part D 5,** Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1985.
590. D. Schauwinhold, W. Schlüter, *Stahl u. Eisen* **104** (1984) 275 – 282.
591. S. Wilmes, H. J. Becker, R. Krumpholz, W. Verderber in: [589], part D 11.
592. W. E. Jominy in: *Hardenability of Alloy Steels,* American Society of Metals, Cleveland 1939, pp. 66 – 94.
593. Euronorm 23.
594. H. Gulden et al., *Stahl u. Eisen* **111** (1991) 103 –110.
595. Society of Automotive Engineers (eds.): *SAE-Handbook,* New York.
596. K. Vetter, E. Gondolf, A. von den Steinen in [589], part D 8.
597. S. J. Engineer, H. Gulden, H. W. Klein, V. Schüler, *Stahl u. Eisen* **108** (1988) 906 – 910.
598. J. A. M. Hertogs et al., *Int. Conf. "Steel in Motor Vehicle Manufacture,"* Würzburg 1990.
599. D. Schreiber, J. Wissemecker-Krieg in [589], part D 18.
600. In [589], part D 5.4.
601. K. Barteld, A. Stanz in [589], part D 26.
602. DIN 50 118,Prüfung metallischer Werkstoffe, Zeitversuch unter Zugbeanspruchung, Jan. 1982.
603. ISO-Standard 6303, March 1981.
604. Verein Deutscher Eisenhüttenleute (eds.): *Ergebnisse deutscher Zeitstandversuche langer Dauer,* Verlag Stahleisen, Düsseldorf 1969.
605. W. Jakobeit, J. P. Pfeiffer, G. Ullricht in Minister für Wirtschaft, Mittelstand und Verkehr in NRW (ed.): *Energiepolitik in Nordrhein-Westfalen,* vol. 14, Düsseldorf 1982.
606. W. Bendick: "Festigkeit bei höheren Temperaturen," *VDEh-Kontaktstudium, Werkstofftechnik,* **part II,** Tagungsband, Mönchengladbach, June 15/16, 1992.
607. G. Kalwa, *VGB Kraftwerkstech.* **63** (1983) no. 4, 356 – 365.
608. K. Lorenz, H. Fabritius, E. Kranz, *Stahl u. Eisen* **92** (1972) no. 9, 393 – 400.
609. H. Fabritius, H. Weber, *VGB Kraftwerkstech.* **54** (1974) no. 4, 250 – 262.
610. J. Stubbe, C. van Nelson: International Symposium "Prediction of Residual Lifetime of Constructions Operating at High Temperature," Netherlands Institute of Welding, Den Haag 1977.
611. G. P. Kalwa, K. Haarmann, K. J. Janssen: Proc. Topical Conference on Ferritic Alloys for Use in Nuclear Energy Technologies, Snowbird 1983.
612. L. Meyer, *Stahl u. Eisen* **97** (1977) 410 – 416.
613. J. Glen, R. R. Barr: "High-Temperature Properties of Steels," *Spec. Rep. Iron Steel Inst.* **97** (1967) 225 – 226.
614. A. Kirsch, F. K. Naumann, H. Keller, H. Kudielka, *Arch. Eisenhüttenwes.* **42** (1971) 353 – 357, 439 – 447.
615. V. K. Sikka, G. T. Ward, K. G. Thomas: Modified 9 % Cr-1Mo-Steel. An Improved Alloy for Steam Generator Application. Technical Program and Data Package for Use of Modified 9 %Cr-1Mo-Steel in ASME Sections I and VIII, 7 Apr. 1982.
616. F. Brühl, H. Cerjak, H. Müsch, K. Niederhoff, M. Zschau, *VGB-Kraftwerkstech.* **69** (1989) no. 12, 1225 – 1231.
617. H. Fabritius, K. Haarmann, H. Müsch, K. Schneemann, *VGB-Kraftwerkstech.* **65** (1985) no. 12, 1167 – 1176.
618. G. Lennartz, A. von den Steinen, *DEW Tech. Ber.* **9** (1969) 139 – 149.
619. B. Bersch, J. Degenkolbe, M. Haneke, W. Middeldorf, *Stahl u. Eisen* **98** (1978) 763 – 778.
620. F. Richter, W. Hemminger, E. Hanitzsch, *steel research* **62** (1991) no. 9, 421 – 425.
621. A. G. Haynes, K. Firth, G. E. Hollex, Buchan: "Properties of Material for Liquefied Natural Gas Tankage," ASTM spec. tech. publ. no. 579, Philadelphia Pa 1975, pp. 288 – 293.
622. D. Uwer, H. G. Wegmann, R. Ortmann, *Stahl u. Eisen* **108** (1988) no. 8, 383 – 388.
623. K. Agusa, A. Kamada, M. Kosha, Y. Nakono, N. Nishigama: *Matching Ferritic Filler MIG Welding of 9 % Ni Steel,* Kawasaki Steel Technical Report no. 6, Sept. 1982.

624. A. Ranak, W. Weßling, H. E. Bock, H. Steinmaurer, L. Faust, *Stahl u. Eisen* **91** (1971) 1255 – 1270.
625. G. L. Swales, A. G. Haynes: 2nd Liquefied Natural Gas Transportation Conference, Oct. 1972, The International Nickel Company of Canada, Ltd., Inco Publ. 437c, pp. 1 – 11.
626. H. Berns: "Anwendung," *Werkstoffkunde Stahl,* **vol. 2, part D 12,** Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1984.
627. E. Stolte in Verein Deutscher Eisenhüttenleute (ed.): "Fundamentals," *Steel–A Handbook for Materials Research and Engineering,* **vol. 1,** Springer Verlag, Berlin 1992.
628. K. H. Zum Gahr, *Fortschr. Ber. VDI-Z. Reihe 5* **27** (1981).
629. W. Heller, H. Schmedders, H. Klein in [626], part D 27.
630. F. Fastenrath in: *Railroad Track, Translation of "Die Eisenbahnschiene,"* Fred. Ungar Publ. Co., 1981.
631. H. Schmedders, H. Bienzeisler, K. H. Tuke, K. Wick, *Eisenbahntechn. Rdsch.* **39** (1990) 195 –199.
632. H. Schmedders, Proc. Seminar Vanadium in Rail Steels, Chicago Nov. 1979, Vanitec, London, pp. 3 – 11.
633. A. M. Sage, W. Hodgson, D. H. Stone in [632]pp. 12 – 15.
634. R. J. Eiber in American Gas Association, AGA (ed.): Symposium on Line Pipe Research, New York 1965, pp. 83 – 118: L 30 000.
635. W. W. Wiedenhoff, G. Vogt, *3 [Drei] R Rohre Rohrleitungsbau Rohrleitungstransp.* **26** (1987) no. 8, 522 – 527.
636. E. M. Moore, J. J. Warga, *Mater. Perform.* **15** (1976) no. 6, 17 – 23.
637. R. Pöpperling, W. Schwenk, *Werkst. Korros.* **31** (1980) 15 – 20.
638. C. Düren in Verein Deutscher Eisenhüttenleute (eds.): *Schweißen von Baustählen,* Verlag Stahleisen, Düsseldorf 1983, pp. 114 – 151.
639. W. W. Wiedenhoff: Symposium on the Interrelation between the Iron and Steel Industry and the Steel Consuming Sectors. Steel/SEM 3R 18 New Grades of Steel for the Production of Large Dimension Steel Pipe, Economic Commission for Europe, Geneve 1977, pp. 1 – 27.
640. W. Haumann et al., *Stahl u. Eisen* **107** (1987) 585 –594.
641. C. Gillessen, W. Heimann, T. L. Ladwein, *steel research* **62** (1991) no. 9, 412 – 420.
642. S. Bernhardsson, J. Oredsson, Martenson in R. A. Lula (ed.): "Duplex Stainless Steels," Conference Proceedings, American Society for Metals, Metals Park, Ohio, 1983, pp. 267 – 281.
643. G. Herbseb, B. Pfeiffer, *Werkst. Korros.* **35** (1984) 254 – 264.
644. G. Herbsleb, *Werkst. Korros.* **37** (1986) 354 –357.
645. G. Herbsleb, E. Szederjei, *Werkst. Korros.* **40** (1989) 651 – 660.
646. Verein Deutscher Eisenhüttenleute: *Prüfung und Untersuchung der Korrosionsbeständigkeit von Stählen,* Verlag Stahleisen, Düsseldorf 1973.
647. W. Schwenk, *Stahl u. Eisen* **89** (1969) 535 – 547.
648. B. Strauss, E. Maurer, *Kruppsche Monatsh.* **1** (1920) Aug., 129 – 146.
649. A. L. Schaeffler, *Met. Prog.* **56** (1949) Nov., 680.
650. E. Baerlecken, H. Fabritius, *Stahl u. Eisen* **78** (1958) 1389 – 1395.
651. R. Courrier, G. le Caer, *Mém. Sci. Rev. Métall.* **71** (1974) 691 – 709.
652. R. Grundmann, *Thyssen Edelst. Tech. Ber.* **12** (1986) no. 2, 176 – 180.
653. N. Arlt, H. J. Fleischer, R. Grundmann, *Thyssen Edelst. Tech. Ber.* **14** (1988) no. 1, 3 – 11.
654. A. Garner, *Met. Prog.* **127** (1985) Apr., 31 – 36.
655. R. A. Lula (ed.): "Duplex Stainless Steels," Conference Proceedings, American Society for Metals, Metals Park, Ohio, 1983.
656. E. Baerlecken, H. Fabritius, *Stahl u. Eisen* **78** (1958) 1389 – 1395.
657. K. Bungardt, H. Borchers, D. Kölsch, *Arch. Eisenhüttenwes.* **34** (1963) 465 – 476.
658. W. E. Cowley, P. J. Robinson, J. Flack, *Proc. Inst. Mech. Eng. part 2 A* **179** (1964/65) no. 5, 145 – 180.
659. E. Baerlecken, K. Lorenz, G. Kalwa, *Erdöl Kohle* **11** (1958) no. 8, 537 – 543.
660. E. B. Backensto, J. W. Sjoberg, *Pet. Ref.* **37** (1958) no. 12, 119 – 120.
661. A. Gala, *Werkst. Korros.* **26** (1975) 115 – 118.
662. J. J. Moran, J. R. Mihalisin, E. N. Skinner, *Corrosion (Houston)* **17** (1961) 191 t – 195 t.
663. G. Kalwa, *Materialprüfung* **22** (1980) no. 1, 27 – 32.
664. W. Schendler, H. Weber, *Nucl. Eng. Des.* **80** (1984) 343 – 358.
665. C. Bosch, *KDJ Z.* **77** (1933) 305 – 317.
666. F. K. Naumann, *Stahl u. Eisen* **58** (1938) 1239 –1250.

667. G. A. Nelson, *Trans. ASME* **51** (1959) 205 – 219.
668. American Petroleum Institut (API), Division of Refining, publication 941, Washington 1990.
669. U. Jäckel, W. Schwenk, *Werkst. Korros.* **22** (1971) 1 – 7.
670. H. Thoma, H. Haas: "Anwendung," in *Werkstoffkunde Stahl,* vol. 2, part D 23, Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1985.
671. H. Stäblein, H.-E. Arntz in [670, .]
672. E. Gondolf et al. in [670], part D 20.
673. L. J. Dijkstra, C. Wert, *Phys. Rev.* **79** (1950) 979 –985.
674. A. Hoffmann, *Arch. Eisenhüttenwes.* **40** (1969) 999 – 1003.
675. N. P. Goss, *Trans. Am. Soc. Met.* **23** (1935) 511 –544;US 1 965 559, 1933.
676. C. Radeloff: in "Magnettechnik," *Kontakt Stud.* **56** (1980) 13 – 32.
677. V. W. Carpenter, J. M. Jackson, US 2 287 467, 1942. V. W. Carpenter, US 2 385 332, 1945. M. F. Littmann, J. E. Heck, US 2 559 340, 1952.
678. S. Taguchi et al., *J. Magn. Magn. Mater.* **2** (1976) 121 – 131.
679. F. Bölling, Hastenrath, *Thyssen Tech. Ber.* **18** (1986) 49 – 68.
680. T. Nozawa, T. Yamamoto, Y. Matsuo, Y. Ohya, *IEEE Trans. Magn. MAG* **15** (1979) 972 – 981.
681. W. Weßling, W. Heimann in [670], part D 22.
682. G. H. Eichelmann, F. C. Hull, *Trans. Am. Soc. Met.* **45** (1953) 77 – 104.
683. H. Dietrich, *DEW Techn. Ber.* **4** (1964) 111 – 132.
684. K. Weber, H. Beck in [670], part D 24.
685. T. D. Yensen, *Trans. Am. Soc. Met.* **27** (1939) 797 –820.
686. C. Straßburger et al.: "Anwendung," in *Werkstoffkunde Stahl,* vol. 2, part D 4, Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1992, pp. 80 – 122.
687. E. Doege, *Z. Ind. Fertigung* **66** (1976) 615 – 619.
688. E.-J. Drewes et al. in *Efficiency in Sheet Metal Forming,* conf. proc., Melbourne 1984, pp. 202 –212 (Stahl u. Eisen 104 (1984) 1333 – 1338).
689. H. Burst, M. Grahle, J. Grashoff, Ch. Schnider: in "Sheet metal forming and energy conservation, " in *IDDRG-Conf. Ann Arbor, Mich.,* American Society for Metals, 1976, pp. 54 – 66.
690. W. T. Lankford, S. C. Snyder, J. A. Bauscher: *Trans. Am. Soc. Met.* **42** (1950) 1197 – 1225.
691. R. L. Whitely, D. E. Wise, D. J. Blickwede, *Sheet Met. Ind.* **38** (1961) 349 – 353.
692. D. J. Blickwede, *Trans. Am. Soc. Met.* **61** (1968) 653 – 679.
693. W. Müschenborn, H. M. Sonne, L. Meyer, *Bänder Bleche Rohre* **15** (1974) 407 – 414.
694. C. Straßburger, W. Müschenborn, G. Robiller: in *Neuzeitliche Verfahren der Werkstoffprüfung,* Verlag Stahleisen, Düsseldorf 1973, pp. 35 – 60.
695. J. Pansera, A. Jcuet in: *Internat. Conf. Steel in Motor Vehicle Manufactur Würzburg 1990,* Verlag Stahleisen, Düsseldorf 1990.
696. L. Meyer et al., *Thyssen Techn. Ber.* **8** (1976) 21 –39.
697. M. Korchynski, H. Stuart in: *Symp. Low Alloy High Strength Steels* (*Pap.)* **1970,** 17 – 27.
698. O. Maid, W. Dahl, C. Straßburger, W. Müschenborn, *Stahl u. Eisen* **108** (1988) 355 – 370.
699. R. Pradhan in [695], pp. 60 – 74.
700. M. Kurosawa, S. Satho, T. Obara, K. Tsunoyama, *Kawasaki Steel Tech. Rep.* **18** (1988) 61 – 65.
701. E. J. Drewes, B. Engl, F. J. Lenze, in [695], pp.75 – 84.
702. H. Gulden, I. Wiesenecker-Krieg: in [686], part D 6, pp. 182 – 199.
703. G. Robiller, W. Schmidt, C. Straßburger, *Stahl u. Eisen* **98** (1978) 157 – 163.
704. H. Sutter, G. Becker: in [686], part D 19, pp.478 – 499.
705. W. Knorr, H. Vöge in Verein Deutscher Eisenhüttenleute (eds.): "Fundamentals," *Steel – A Handbook for Materials Research and Engineering,* **vol. 1, part C 9,** Springer Verlag, Berlin 1992.
706. E. Schulz, Jahrestagung IISI, Montreal, Canada, 7 Oct. 1991, *Stahl u. Eisen* **112** (1992) no. 5, 43 –51.
707. K. A. Zimmermann, *Stahl u. Eisen* **111** (1991) no. 12, 36 – 40.
708. J. A. Philipp et al., *Stahl u. Eisen* **112** (1992) no. 12, 75 – 86.
709. J. A. Philipp, *World Steel & Metalworking Annu.* 1988, 101 – 104 (SPG-Coburg).
710. J. A. Philipp, *World Steel Rev.* **1** (1991) 62 – 73.
711. E. Schulz, *Stahl u. Eisen* **113** (1993) no. 2, 25 –33.
712. E. Höffken, J. Wolf, P. Kühn, *Thyssen Tech. ber.* **24** (1992) 119 – 127.
713. Verein Deutscher Eisenhüttenleute (eds.): *Stahlfibel,* Verlag Stahleisen, Düsseldorf 1989, p. 134.

714. Bundesverband der Deutschen Schrott-Recycling-Wirtschaft e.V. (ed.): *Vom Schrott zum Stahl,* Düsseldorf 1984, pp. 37, 38, 39.
715. Bundesverband der Deutschen Schrott-Recycling-Wirtschaft e.V. (ed.): *Vom Schrott zum Stahl,* Düsseldorf 1984, pp. 36, 37.
716. W. Ullrich, H. Schicks: "Aspekte zum Recycling von metallisch beschichtetem Stahl," *Stahl u. Eisen* **111** (1991) no. 11, 85, 86.
717. J.-P. Kleingarn: "Rückgewinnung von Zink aus verzinktem Stahlschrott," *Feuerverzinken* **15** (1986) no. 3, Beratung Feuerverzinken, Düsseldorf.
718. In [712], pp. 121, 124.
719. International Iron and Steel Institute, Brussels .
720. Bundesvereinigung Deutscher Stahlrecycling- und Entsorgungsunternehmen e.V., Düsseldorf.
721. Bundesverband der Deutschen Schrott-Recycling-Wirtschaft e.V. (ed.): *Vom Schrott zum Stahl,* Düsseldorf 1984, pp. 11, 12.

Sterilization → Biotechnology

Steroids

RENATE MÜLLER, Schering AG, Berlin, Federal Republic of Germany

1. **Nomenclature** 1
2. **Sterols** 3
2.1. **Cholesterol and Related Compounds** 3
2.2. **Vitamin D** 4
2.3. **Methyl Sterols** 4
3. **Bile Acids** 5
4. **Steroid Hormones** 6
5. **Sapogenins** 8
6. **Steroid Alkaloids** 9
7. **Steroid Lactones** 10
8. **References** 10

1. Nomenclature

Steroids are compounds that contain a cyclopenta[α]-phenanthrene ring skeleton which, when fully hydrogenated, is called gonan or steran.

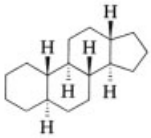

5α-Gonan or 5α-Steran

This parent structure can be modified by opening one or more bonds, by ring enlargement, or by ring contraction. The carbon atoms are numbered and the rings are indicated by letters (Fig. 1). There are angular methyl groups on C-10 and C-13. If, for instance, C-10 does not bear a methyl group, the resulting compounds are called 19-norsteroids.

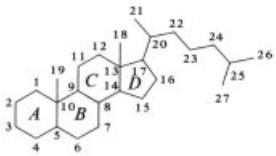

Figure 1. Steroid numbering

For chiral centers at carbon atoms 8, 9, 10, 13, and 14, the absolute stereochemistry (Fig. 2) is specified in the parent names. Rings B, C, and D are predominantly *trans* linked. All atoms or groups below the plane of the ring are represented by a dashed line and the letter α, and those above the plane are indicated by a solid line and the letter β. If the configuration is unknown, a wavy line and the letter ξ is used.

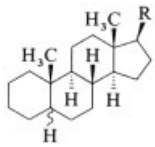

Figure 2. Steroid stereochemistry
R = Oxygen functional group or alkyl group, which may be substituted

Hydrogen atoms at 8, 9, and 14 are not shown in the structural formula as long as there is *trans* linkage. The position of the hydrogen at C-5 should always be indicated. The parent compounds are shown in Figures 3 and 4.

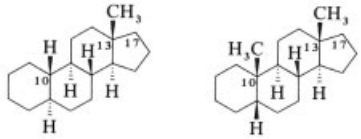

5α-Estrane or 19-Nor-5α-androstane

5β-Androstane

Figure 3. Parent compounds with no side- chain at C-17

The great variety of natural and synthetic steroid compounds results from the presence of double bonds and various substituents on the ring skeleton and the side- chain. The systematic names are formed according to IUPAC rules [1]. Trivial names are permitted for some steroids, mostly natural steroids with high biological activity.

Owing to the chair conformation of the cyclohexane ring (Fig. 5), substituents on the steroid skeleton are either approximately in the plane of the ring, equatorial (e), or vertical, axial (a). The chemical reactivity depends on this conformation [2].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_309

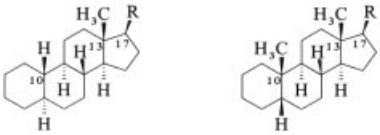

R	Configuration	5α-Series	5β-Series
		5α-pregnane	5β-pregnane
	20*R*	5α-cholane	5β-cholane
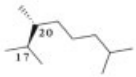	20*R*	5α-cholestane	5β-cholestane
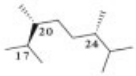	20*R*, 24*S*	5α-ergostane	5β-ergostane
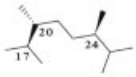	20*R*, 24*R*	5α-campestane	5β-campestane
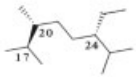	20*R*, 24*S*	5α-poriferastane	5β-poriferastane
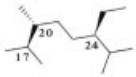	20*R*, 24*R*	5α-stigmastane	5β-stigmastane
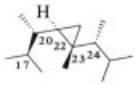	20*S*, 22*R* 23*R*, 24*R*	5α-gorgostane	5β-gorgostane

Figure 4. Parent compounds with a side-chain at C-17

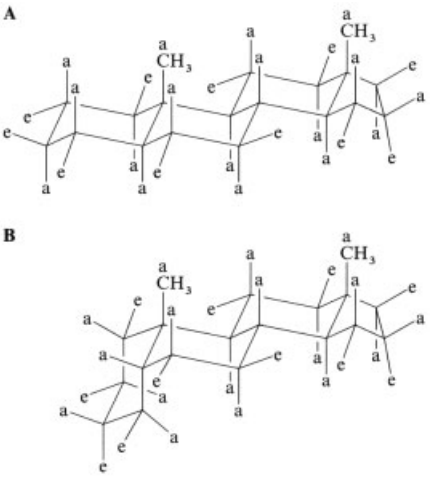

Figure 5. Conformations of steroids
A) 5α-Series; B) 5β-Series

2. Sterols

The first compounds of this class to be studied in detail were isolated from the unsaponifiable parts of animal and plant oils and fats. They are solid crystalline secondary alcohols with 26–30 carbon atoms, and can occur as such or as fatty acid esters and glycolipids. They were given the group name sterols (Greek, *stereos*, solid).

2.1. Cholesterol and Related Compounds (Figs. 6, 7)

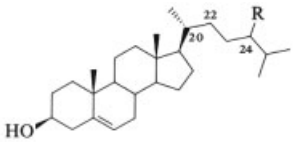

Sterol	Δ	R	Configuration	CAS No.
Cholesterol (**1**)		H	20 *R*	[*57-88-5*]
Stigmasterol (**2**)	22	Et	20 *R*,24 *R*	[*83-48-7*]
β-Sitosterol (**3**)		Et	20 *R*,24 *R*	[*83-46-5*]
Brassicasterol (**4**)	22	Me	20 *R*,24 *R*	[*474-67-9*]
Campesterol (**5**)		Me	20 *R*,24 *R*	[*474-62-4*]
24-Methylene-cholesterol (**9**)		$=CH_2$	20 *R*	[*474-63-5*]
Poriferasterol (**11**)	22	Et	20 *R*,24 *S*	[*481-16-3*]
Clionasterol (**12**)		Et	20 *R*,24 *S*	[*83-47-6*]
24-Dehydro-cholesterol (**14**)	24	H	20 *R*	[*313-04-2*]

Figure 6. Δ^5-Sterols

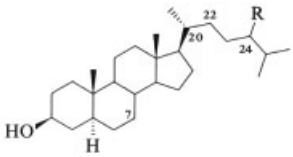

Sterol	Δ	R	Configuration	CAS No.
Spinasterol (**6**)	7,22	Et	20 *R*,24 *R*	[*481-18-5*]
Zymosterol (**7**)	8,24	H	20 *R*	[*128-33-6*]
Faecosterol (**8**)	8	$=CH_2$	20 *R*	[*516-86-9*]
Spongesterol (**10**)	22	Me	20 *R*,24 *S*	[*474-66-8*]
Chondrillasterol (**13**)	7,22	Et	20 *R*,24 *S*	[*481-17-4*]

Figure 7. 5α-Sterols

Cholesterol (**1**) is the most widely distributed sterol. Its name indicates that it is the main constituent of human gallstones. It is found in many animal tissues, in wool fat, and in plants. A survey of the literature on the occurrence, isolation, determination, metabolism, etc., is presented by Cook [3]. In humans and other mammals, cholesterol is a metabolic precursor of gallstones [4] and steroid hormones (→ Hormones, Chap. 3.1.). At a low developmental level, diverse, poorly separable sterol mixtures occur, but in the course of evolution, cholesterol becomes increasingly important until, in vertebrates, it is practically the only sterol [5]. In the 1930s, cholesterol served as the starting compound for androgen synthesis [2]. Although it was for some time replaced by sapogenins, it has become important again, owing to progress in microbiological fermentation [6]. Other important starting sterols for the partial chemical synthesis of compounds of the androstane and estrane series are stigmasterol (**2**) and β-sitosterol (**3**) (→ Hormones, Chap. 3.2.1.). Both sterols are contained, e.g., in soybean oil. Further phytosterols include brassicasterol (**4**) (→ Fats and Fatty Oils, Chap. 2.5.) and campesterol (**5**) [7], spinasterol (**6**) from spinach [8] or alfalfa [9], and the yeast sterols zymosterol (**7**) [10] and faecosterol (**8**) [7]. The sterols isolated from invertebrate sea animals exhibit structural changes, especially in the side chain. 24-Methylenecholesterol (**9**), spongesterol (**10**), poriferasterol (**11**), clionasterol (**12**), and chondrillasterol (**13**) occur in various fungi, and 24-dehydrocholesterol (**14**) occurs in the goose barnacle [2]. A cyclopropane ring in the side-chain, as in gorgosterol (22*S*,22*R*,23*R*,24*R*) [*29782-65-8*] (**15**), appears to be characteristic of a whole class of marine steroids [11].

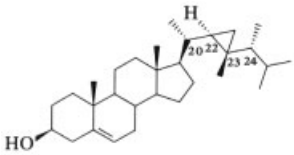

15

Calysterol [*57331-04-1*] (**16**) contains a cyclopropene ring [12].

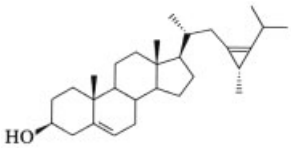

16

A list of natural steroids is given in [7, Table 19]. Some cholesterol derivatives, e.g., cholesteryl benzoate, form liquid crystals (→ Liquid Crystals, Chap. 3. → Liquid Crystals, Chap. 3.6. → Liquid Crystals, Chap. 3.7.).

2.2. Vitamin D

It was recognized around 1920 that rickets could be cured by cod-liver oil or solar irradiation. The antirachitic constituent of cod-liver oil was found to be a fat-soluble compound which was called vitamin D. It was also discovered that plant oils became antirachitic on irradiation [13], and that the potentially active compounds were in the unsaponifiable part [14], which consisted of a mixture of sterols. Whereas cod-liver oil apparently contained preformed vitamin D, the phytosterol mixture contained a provitamin. It was found [2] that ergosterol is the percursor of vitamin D_2 (ergocalciferol) (→ Hormones, Chap. 5.1.) and 7-dehydrocholesterol yields vitamin D_3 (cholecalciferol). Vitamin D_3 is metabolized to 1α,25-dihydroxyvitamin D_3 (calcitriol) in the liver and kidney. Calcitriol is the agent responsible for calcium and phosphate transport in the body. Thus, cholecalciferol and its biologically active metabolites must be regarded as vitamins as well as hormones (→ Hormones, Chap. 5.2. → Hormones, Chap. 5.3.). Since the only abundant source of vitamin D is fish-liver oil, syntheses are of great importance [2]. Attempts were made to find other antirachitic substances by irradiating sterols that differed from ergosterol only in the side-chain [15]. Methods for the synthesis of calcitriol have been developed more recently, and attempts are being made to find derivatives which are more efficient in separating the systemic effect on calcium metabolism from the antiproliferative activity (→ Hormones, Chap. 5.4.). In the last few years, side-chain analogs have been tested for possi-

ble use in some forms of cancer and in psoriasis [16–19].

2.3. Methyl Sterols

"Sterols" with additional methyl groups on the ring skeleton are found together with cholesterol in many plants and fungi, and, abundantly, in the wool fat of sheep. The best known are lanosterol [*79-63-0*] (**17**) and agnosterol [*472-29-7*] (**18**) from wool fat [20], and kryptosterol from yeast [21].

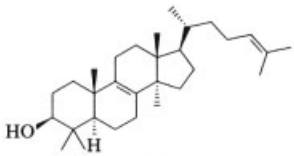

17

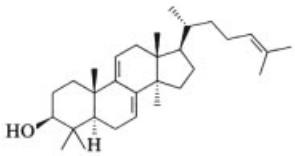

18

The identity of kryptosterol and lanosterol was established later [22]. Although it is analytically difficult to distinguish these compounds from the real sterols, they are regarded as triterpenes, which are mainly pentacyclic isoprenoid compounds with 30 carbon atoms. Lanosterol plays a key role in the biosynthesis of cholesterol. Acetate is converted to squalene in ten steps with the help of enzymes, and the transformation of squalene to cholesterol involves another 19 enzymatic steps [23, 24]. Squalene gives rise to lanosterol, which is then converted to ergosterol, an important sterol in many fungi. Substances which interfere with steroid biosynthesis are used in plant protection to control pathogenic fungi (→ Fungicides, Agricultural, Chap. 8.6.). Other compounds of the lanostane type are mainly 4,4,14α-trimethyl derivatives of cholestane or ergostane; some are acids (Fig. 8), e.g., eburicoic acid (**19**), polyporenic acid C (**20**), tumulosic acid (**21**), and pinicolic acid (**22**). They all occur in parasitic fungi that live on wood [2]. Others have a cyclopropane ring at position 9, 10 (Fig. 9), e.g., cycloartenol (**23**) from the latex of the fruit of *Artocorpus integrifolia,* cyclolaudenol (**24**) from opium, and cycloeucalenol (**25**), with only one methyl group at C-4, from eucalyptus. The significance of cycloartenol in the biosynthesis of steroids in higher plants is comparable to that of lanosterol [25].

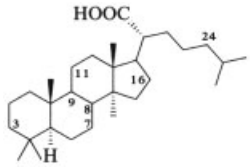

	Δ	3	16	24	CAS No.
Eburicoic acid (**19**)	8	β-OH		$=CH_2$	[*560-66-7*]
Polyporenic acid C (**20**)	7,9(11)	=O	α-OH	$=CH_2$	[*465-18-9*]
Tumulosic acid (**21**)	8	β-OH	α-OH	$=CH_2$	[*508-24-7*]
Pinicolic acid (**22**)	8,24	=O			[*466-05-7*]

Figure 8. Lanostane acids

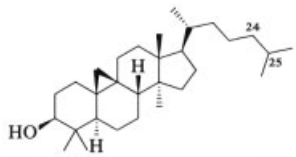

Cycloartenol (**23**)	Δ^{24}	[*469-38-5*]
Cyclolaudenol (**24**)	Δ^{25}-(24*S*)-methyl	[*511-61-5*]
Cycloeucalenol (**25**)	24-methylene	[*469-39-6*]

Figure 9. 9,19-Cyclo lanostane compounds

Compounds of the euphane and lanostane types are epimers with respect to carbon atoms 13, 14, and 17. Thus, euphol, obtained from the resin of a species of *Euphorbia*, corresponds to 13α, 14β, 17α-lanosterol. The 20-epimer is tirucallol [26].

The cucurbitacins exhibit natural bactericidal activity, and have been tested for cytotoxic and antitumor activity [27], and are regarded as kairomones (→ Biological Control). The fusidane series should also be mentioned here, fungal metabolites which have antibiotic properties and are especially active in staphylococcal infections. These compounds include fu-

sidic acid [*6990-06-3*] (**26**) (→ Antibiotics) and helvolic acid [*29400-42-8*] (**27**).

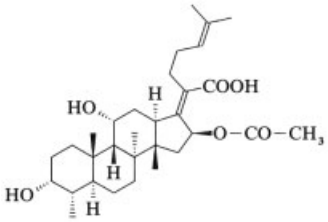

26

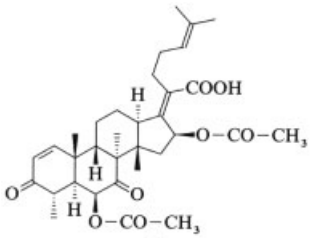

27

The stereochemical linkage, *trans-syn-trans-anti-trans*, which forces ring B into the boat form, is unusual [28].

3. Bile Acids

The bile acids are among the oldest known steroid compounds. In 1828, cholic acid was isolated in the impure state by the saponification of ox bile. The first careful studies were carried out by STRECKER in 1848 [29]. Later, it was primarily WIELAND and coworkers who studied the constitution and configuration of the bile acids. They are mostly mono-, di-, or trihydroxy derivatives of cholanic acid [2,4]. Choleinic acid, the second bile acid to be discovered, is a molecular compound consisting of 1 mol fatty acid and 8 mol deoxycholic acid [30]. Deoxycholic acid reacts with many acids, esters, alcohols, ethers, phenols, and hydrocarbons to form molecular compounds, now collectively known as cholenic acids. They are very stable, and X-ray analysis showed that they are inclusion compounds [31]. Apocholic acid (8, 14-dehydrodeoxycholic acid) behaves in the same way [32]. Only cholic acid and deoxycholic acid (from ox bile) and hyodeoxycholic acid (from pig bile) can be used as starting compounds for syntheses, because they are available in large quantities from natural sources. Even microbiological methods of cortisone production, starting with plant steroid sapogenins, have not been able to completely replace cholic acid and deoxycholic acid as base materials (→ Hormones). Whereas mammalian bile acids are derived from 5β-cholanic acid, bile acids with side- chains extended by 3 – 4 carbon atoms have been found in the bile of phylogenetically old species of toads and crocodiles, e.g., 3α-7α-dihydroxy-5β- cholestanic acid in the bile of *Alligator mississipiensis.* The bile alcohol scymnol [*6785-34-8*] (**28**) was isolated from shark bile, and has also been found in other species of fish and frogs [33].

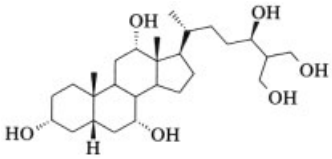

28

A survey of the bile acids found in fish, amphibia, reptiles, birds, and mammals is presented by HASLEWOOD [34–36]. Bile acids with extended side- chains are regarded as the phylogenetic precursors of mammalian bile acids [37].

The bile acids are the end-products of cholesterol metabolism in the liver [38]. First, changes occur at the ring skeleton. The side- chain is then subjected to oxidative degradation [39]. The bile acids possess a hydrophobic and a hydrophilic part. They form water-soluble micelles with lecithins and water-insoluble lipophilic food components; thus, fatty acids, monoglycerides, steroids, fat-soluble vitamins, and cholesterol are kept in solution in the bile [40]. However, if the concentration of bile acids falls below a certain value, gallstones can form. Gallstones differ in chemical composition; ca. 10 – 20 % are pure cholesterol stones [41], which can be dissolved by oral administration of chenodeoxycholic acid [*474-25-9*] (**29**) or ursodeoxycholic acid [*128-13-2*] (→ Gallbladder and Liver Therapy, Chap. 2.). Ursodeoxycholic acid can also be used in the treatment of liver cirrhosis [42] and to ameliorate the symptoms of biliary reflux disease [43]. Cholic

acid [*81-25-4*], deoxycholic acid [*83-44-3*], and dehydrocholic acid [*81-23-2*] are used as choleretics (→ Gallbladder and Liver Therapy, Chap. 1.). The sodium salt of deoxycholic acid is used on a small scale to improve the resorption of drugs, e.g., absorption of insulin through the nasal mucous membrane [40].

Chenodeoxycholic acid (Fig. 10) is produced on an industrial scale from cholic acid according to the method of FIESER [44]. Ursodeoxycholic acid can be obtained from chenodeoxycholic acid (**29**) by oxidation of the 7α-OH group to the 7-ketone, followed by reduction with alkali metal in amyl alcohol or liquid ammonia [45], by the selective microbiological hydroxylation of lithocholic acid [46], or from cholic acid [47].

4. Steroid Hormones

Estrone [*53-16-7*] was the first sex hormone to be obtained in the pure state [48]. A short time later, estriol [*50-27-1*] was isolated from the urine of pregnant women [49]. In 1934, four different teams succeeded almost simultaneously in preparing pure progesterone [50–53]. The preparation of testosterone [*58-22-0*] was achieved in 1935 [54]. At the same time, investigation of extracts of the adrenal cortex intensified. The research groups of REICHSTEIN and KENDALL introduced the practice of naming their isolated substances by letters. By 1943, as many as 28 steroids had been identified, including cortisone, cortisol, cortexolone (Reichstein S), 11-dehydrocorticosterone, corticosterone, and deoxycorticosterone [55]. Aldosterone was isolated in 1953 [56]. Industrial production of pregnenolone from diosgenin, which occurs in the root of the Mexican plant *Dioscorea*, began in 1940 [57]. The discovery in 1948 that cortisone had a strong effect on the symptoms of arthritis, initiated an intensive search for further anti-inflammatory steroids and the development of new methods for the commercial production of cortisone. The discovery of the reduction of phenol ethers to dihydro compounds (1950) made possible the production of 19-norsteroids from estrone methylether [58]. Many steroid reactions had been described by 1950, but little was known about the reaction mechanisms. BARTON showed that many rules that had been established empirically could be rationalized by the concept of axial and equatorial bonds [59]; conformation analysis gave steroid research a strong impetus. It was also recognized that UV [60] and IR spectra [61] are valuable in structural elucidation.

Another important physical quantity for the characterization of steroid compounds is optical rotation. Steroids exhibit several asymmetric centers; all naturally occurring steroids are optically active (→ Hormones, Chap. 3.1.).

Nuclear magnetic resonance sprectra, mass spectrometry, and X-ray analysis are also used for structure determination [62].

Steroid hormones are divided into three main physiological groups:

1) Sex hormones
   Estrogens (→ Hormones, Chap. 3.5.)
   Gestagens (→ Hormones, Chap. 3.7.)
   Androgens (→ Hormones, Chap. 3.6.)
2) Adrenal steroids
   Glucocorticoids
   (→ Hormones, Chap. 4.4.)
   Mineralocorticoids
   (→ Hormones, Chap. 4.9.)
3) Calcium-regulating sterols
   (→ Hormones, Chap. 5.1.)
   Vitamin D
4) Insect skin-shedding hormones

Cholesterol is the common precursor in the biosynthesis of the first three groups (→ Hormones, Chap. 3.1.). Apart from steroid compounds with hormonal activity, some compounds act as hormone antagonists. Depending on the mode of action, distinction is made between substances that occupy specific receptors and those that block hormone production. Thus, not only are antiestrogens known (→ Hormones, Chap. 3.5.4.), but also antiandrogens (→ Hormones, Chap. 3.6.3.), antigestagens (→ Hormones, Chap. 3.7.5.), and antagonists of mineralocorticoids (→ Hormones, Chap. 4.9.).

In the synthesis of steroid hormones, four processing modes are distinguished:

1) Partial chemical synthesis from natural products (→ Hormones, Chap. 3.2.)
2) Microbiological degradation of sterols (→ Hormones)
3) Total synthesis (→ Hormones, Chap. 3.4.)
4) Isolation from natural sources

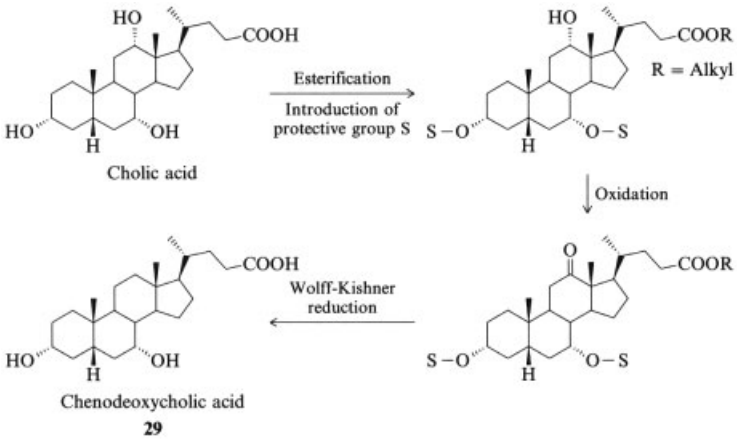

Figure 10. Production of chenodeoxycholic acid

The starting materials for partial chemical synthesis must contain the steroid nucleus, they must be convertible to steroid hormones via intermediates by simple chemical reactions, and they must be available in sufficient quantities. The most important starting substances are cholesterol, stigmasterol, β-sitosterol, and diosgenin (→ Hormones, Chap. 3.2.1.) as well as bile acids, ergosterol, and hecogenin (→ Hormones, Chap. 4.8.2.). The most important compounds obtained are dehydropregnenolone, progesterone, androstenolone, and estrone (→ Hormones, Chap. 3.2.1.). The androgenic steroids also have a more-or-less strong anabolic activity which can be intensified by chemical modification of the steroid nucleus. However, some androgenic activity always remains (→ Hormones, Chap. 3.6.4.). The glucocorticoids, another group of steroid hormones, have also been subjected to various chemical modifications. Attempts have been made to reduce the mineralocorticoid side effects. The introduction of a double bond at position 1,2 gave compounds with glucocorticoid, anti-inflammatory, and antiallergenic activity 4 – 5 times that of cortisone (→ Hormones, Chap. 4.10.). The introduction of 9α-halogen (→ Hormones, Chap. 4.10.2.), 16α-hydroxyl (→ Hormones, Chap. 4.10.3.), 6α-methyl (→ Hormones, Chap. 4.10.4.), 16α-methyl (→ Hormones, Chap. 4.10.5.), 16-methylene (→ Hormones, Chap. 4.10.6.), 6α-fluoro, 6α- chloro, and 6α, 9α-difluoro substituents (→ Hormones, Chap. 4.10.7.) produced further improvements. In 1964, Raspé and coworkers found that 16α-methylcorticoids lacking the 17-hydroxyl group are still relatively strong anti-inflammatory substances, with very low mineralocorticoid activity (→ Hormones, Chap. 4.10.8.). The 11β-hydroxyl group has been replaced by a halogen (→ Hormones, Chap. 4.10.9.). Other derivatives of well-known, active corticoids gave a series of compounds with little or no systemic activity, so these compounds can be very successful in treating dermatosis in children (→ Hormones, Chap. 4.11.). It has not been possible to find either the exact mechanism of the anti-inflammatory and immunosuppressive activity of the glucocorticoids (→ Dermatotherapeutic Agents), or a new class of compounds to replace the glucocorticoids.

The insect skin-shedding hormones are completely different. Each stage of skin shedding from larva to pupa to adult insect is subject to hormonal control. Whereas growth is controlled by the juvenile hormones (→ Biological Control, Chap. 5.1.1.), the skin-shedding hormones are steroids. In 1954, the first 25 mg of a crystalline skin-shedding hormone was isolated from the silkworm *Bombyx mori* [63]. This substance, α-ecdysone (**30**), was found to be a steroid by chemical and X-ray studies [64]. A very similar compound, β-ecdysone (**31**), also occurs in the silkworm. Other skin-shedding hormones, or zooecdysones (Fig. 11) have been isolated from insects or crustaceans,

e.g., inokosterone (**32**) and makisterone A (**33**) from crabs [65] and 26-hydroxy-β-ecdysone (**34**) from the tobacco hornworm *Manduca sereta (L)* [66]. Shortly after the elucidation of the structure of α-ecdysone (**30**), a number of steroids with skin-shedding activity, the phytoecdysones, were also found in plants [67]. They are widely distributed, and occur in relatively large quantities [68]. By 1979, a total of 47 different phytoecdysones were known, and most of the zooecdysones had been found in plants as well [69]. The first synthesis was published by Schering AG – Hoffmann LaRoche [70]. This was followed by several improvements and other starting materials were used [71–73]. The direct introduction of a 2,7-diene-6-one group from 3β-tosylate-5-ene resulted in a considerable shortening of the synthetic pathway [74].

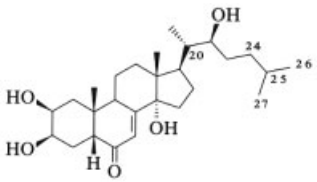

Name	20	24	25	26	CAS No.
α-Ecdysone (**30**)			OH		[*3604-87-3*]
β-Ecdysone (**31**)	OH		OH		[*5289-7-7*]
Inokosterone (**32**)	OH			OH	[*15130-85-5*]
Makisterone A (**33**)	OH	CH_3	OH		[*20137-14-8*]
26-Hydroxy-β-Ecdysone (**34**)	OH		OH	OH	[*52717-49-4*]

Figure 11. Zooecdysones

5. Sapogenins

Plant glycosides that form a soapy foam in water are called sapogenins. They are composed of an aglycon, sapogenin, and one or more sugars (→ Saponins, Chap. 3.1.). Apart from a few exceptions, the sugar moiety is bound to the steroid through a 3β-OH group. The first saponins were derived from digitalis preparations and were called digitonin and gitonin, and the corresponding aglycons, digitogenin and gitogenin. Even later, the names were often derived from the plants from which the sapogenins were isolated. Raw plant extracts were used as detergents because they are neither alkaline nor precipitable by hard water. Saponin preparations without cardiac effects have been used in fire extinguishers, and as fish poisons. Although fish are numbed or killed by saponins, they are still edible because the oral consumption of saponins is not harmful to humans. The structures of the aglycons are not as varied as those of the cardiac glycosides. These steroids belong predominantly to the spirostanes (→ Saponins). The most important is diosgenin, the starting compound for the partial synthesis of steroid hormones (→ Hormones, Chap. 3.2.1.). Hecogenin was also of importance for a time (→ Hormones).

6. Steroid Alkaloids

Steroid alkaloids occur in both plants and animals. Plant steroid alkaloids are widely distributed, and have various structures, depending on the plant species. Funtumine [*474-45-3*] (**35**), conessine, holarrhimine Δ^5[*468-31-5*] (**36**), and holarrhidine 5α-H [*82182-51-2*] (**37**) have been isolated from certain species of *Holarrhena*.

The first three compounds were considered as starting substances for the commercial synthesis of steroids [75], but without success.

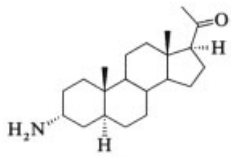

35

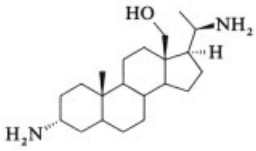

36, 37

Irehdiamine A is found in *Funtumia elastica*, and cyclobuxine D and buxocine C [76–78] in certain species of *Buxus*. The solanum alkaloids solasodine (→ Alkaloids, Chap. 14.), solanidine

[*80-78-4*] (**38**), and tomatidine [*77-59-8*] (**39**) should also be mentioned [79].

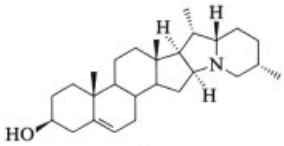

38

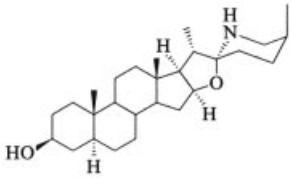

39

Solasodine and tomatidine have been converted to 16-dehydropregnenolone by the Marker process [80]. The solanum alkaloids occur in plants in the form of glycosides (→ Saponins, Chap. 3.2.).

All the alkaloids mentioned above have an essentially unchanged steroid skeleton. Attempts have been made to prepare solanum alkaloids with the help of biotechnology, but without commercial success [81].

In the veratrum and fritillaria alkaloids, ring *C* is contracted to a five-membered ring and ring D is enlarged to a six-membered ring. Hence, they are *C*-nor-*D*-homosteroids [82, 83]. Jervine and verticine (→ Alkaloids, Chap. 14.5.), as well as the ester alkaloid protoverine [*76-45-9*] (**40**) belong to this group.

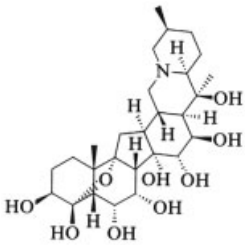

40

The ester alkaloids contain 7 – 9 oxygen atoms, esterified with organic acids, such as acetic, angelic, veratric, vanillic, or tiglic acid. Protoverine possesses vasodilative properties.

The animal steroidal alkaloids are poisonous, and are excreted though the skin. The main poisons of the salamander are samandarin [*467-51-69*] (**41**) and cycloneosamandione [*631-72-1*] (**42**).

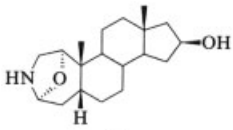

41

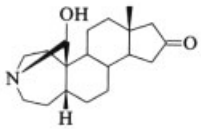

42

Samandarin poisoning leads to cramps and respiratory paralysis. In fact, salamanders can die from their own poison if it enters the circulatory system. The poison of a small, brightly colored frog, *Phyllobates amotaenia*, is used as an arrow poison in Colombia. The active substances are 20-esters of batrachotoxins with various pyrrole carboxylic acids [84]. One of the most poisonous substances known is 2,4-dimethylpyrrole-3- carboxylic acid ester. Its subcutaneous LD_{50} is 2 µg/kg in the mouse, compared to 500 µg/kg for strychnine.

7. Steroid Lactones

The steroid lactones include the cardenolides and bufadienolides, aglycons of the cardiac glycosides (→ Cardiac Glycosides and Synthetic Cardiotonic Drugs, Chap. 1., → Cardiac Glycosides and Synthetic Cardiotonic Drugs, Chap. 2.) as well as the holothurigenins, withanolides, and antheridiols. The sea cucumber excretes saponins that are toxic to fish, and exhibit hemolytic and oncolytic properties [85]. Acid hydrolysis gives the holothurigenins, e.g., seychellogenin [*24041-68-7*] (**43**).

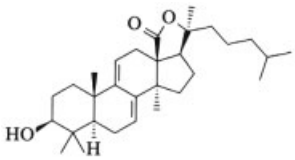

43

The withanolides are found in the Solanacae plant family [86]. The best-known representative is withaferin A [*5119-48-2*] (**44**).

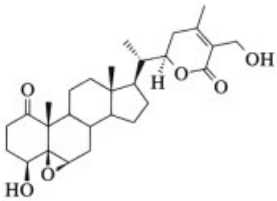

44

Antheridiol [*22263-79-2*] (**45**) is the first specifically acting steroid sex hormone to be discovered in plants [87].

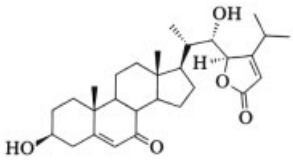

45

Antheridiol is secreted by the female plant and brings about formation of antheridial hyphae on the male plant of *Achyla bisexualis* [88], which lives in water.

8. References

1. IUPAC-IUB 1989 *Pure Appl. Chem.* **61** (1989) 1783.
2. L. F. Fieser, M. Fieser: *Steroide,* Verlag Chemie, Weinheim 1961.
3. R. P. Cook: *Cholesterol,* Academic Press, New York 1958.
4. A. Aigner, A. Bauer, *Med. Monatsschr. Pharm.* **11** (1988) 369.
5. W. Bergmann: "Evolutionary Aspects of Sterols," in [3].
6. M. Nagasawa, M. Bae, G. Tamura, K. Arima, *Agric. Biol. Chem.* **33** (1969) 1636.
7. C. J. W. Brooks in S. Coffey (ed.): *Rodd's Chem. Carbon Compounds,* 2nd ed., **vol. 2 D,** Elsevier, Amsterdam, pp. 150 – 156.
8. M. C. Hart, F. W. Heyl, *J. Biol. Chem.* **95** (1932) 311.
9. E. Fernholz, M. L. Moore, *J. Am. Chem. Soc.* **61** (1939) 2467.
10. H. Wieland, Y. Kanaoka, *Justus Liebigs Ann. Chem.* **530** (1937) 146.
11. R. L. Hale et al., *J. Am. Chem. Soc.* **92** (1970) 2179. R. L. Hale, N. C. Ling, C. Djerassi, *J. Am. Chem. Soc.* **92** (1970) 5281.
12. E. Fattorusso et al., *Tetrahedron* **31** (1975) 1715.
13. A. F. Hess, M. Weinstock, *J. Biol. Chem.* **62** (1924) 301.
14. A. F. Hess, M. Weinstock, F. D. Helman, *J. Biol. Chem.* **63** (1925) 305.
15. A. Windaus, R. Langer, *Justus Liebigs Ann. Chem.* **508** (1933) 105.
16. T. Eguchi et al., *Chem. Pharm. Bull.* **38** (1990) 1246.
17. L. Binderup et al., *Biochem. Pharmacol.* **42** (1991) 1569.
18. N. Kubadera, H. Watanabe, T. Kawanashi, M. Matsumoto, *Chem. Pharm. Bull.* **40** (1992) 1494.
19. W. H. Okamura, J. A. Palenzuela, J. Plumet, M. M. Midland, *J. Cell. Biochem.* **49** (1992) 10.
20. H. Windaus, R. Tschesche, *Hoppe-Seyler's Z. Physiol. Chem.* **190** (1930) 51.
21. H. Wieland, H. Pasedach, A. Ballauf, *Justus Liebigs Ann. Chem.* **529** (1937) 68.
22. L. Ruzicka, R. Deuss, O. Jeger, *Helv. Chim. Acta* **28** (1945) 759. L. Ruzicka, R. Deuss, O. Jeger, *Helv. Chim. Acta* **29** (1946) 204.
23. J. R. Sabine: *Cholesterol,* Marcel Dekker, New York 1977, pp. 79 – 103.
24. L. J. Goad in H. L. J. Makin (ed.): *Biochemistry of Steroid Hormones,* Blackwell Scientific Publ., Oxford 1975, p. 17.
25. R. B. Boar, C. R. Roamer, *Phytochemistry* **14** (1975) 1143.
26. F. L. Warren, K. H. Watling, *Chem. Ind. (London)* 1956, 24.
27. D. Lavie, E. Glotter, *Fortschr. Chem. Org. Naturst.* **29** (1971) 307.
28. W. F. Johns (ed.): *MTP International Review of Science,* **8**, Section 1, Butterworths, London 1973, 198; W. F. Johns (ed.): *MTP International Review of Science,* **8**, Section 2, 1976, 195.

29. A. Strecker, *Justus Liebigs Ann. Chem.* **67** (1848) 1.
30. H. Wieland, H. Sorge, *Hoppe-Seyler's Z. Physiol. Chem.* **97** (1916) 1.
31. F. Cramer: *Einschlußverbindungen,* Springer Verlag, Heidelberg 1954.
32. G. Cilento, *J. Am. Chem. Soc.* **74** (1952) 908.
33. P. Nair in D. Kritchevsky (ed.): *The Bile Acids,* **vol. 1,** Plenum Press, New York 1971; vol. 2, 1973.
34. G. A. D. Haslewood, W. M. Wootton, *Biochem. J.* **47** (1950) 584.
35. G. A. D. Haslewood, *Physiol. Rev.* **35** (1955) 178.
36. I. G. Anderson, G. A. D. Haslewood, I. D. P. Wootton, *Biochem. J.* **67** (1957) 323.
37. P. Back, *Klin. Wochenschr.* **58** (1980) 55.
38. K. Bloch, B. N. Berg, D. Rittenberg, *J. Biol. Chem.* **149** (1943) 511.
39. H. Danielsson, J. Sjovall, *Annu. Rev. Biochem.* **74** (1975) 233.
40. A. Aigner, A. Bauer, *Med. Monatsschr. Pharm.* **11** (1988) 369.
41. A. Sieg et al., *DMW, Dtsch. Med. Wochenschr.* **111** (1986) 1760.
42. R. Poupon, R. E. Poupon, Y. Calenus, *Lancet* 1987, 834.
43. A. B. Stefanicosky et al., *Gastroenterology* **89** (1985) 1000.
44. L. F. Fieser, S. Rajagapolan, *J. Am. Chem. Soc.* **72** (1950) 5530.
45. Roussel Uclaf, EP 0 072 293, 1983 (J. Bulidon, R. Demuynck, C. Pavan).
46. Yakult Honsha, EP 119040, 1984 (H. Sawada, M. Watanuki).
47. Erregierre, EP 0 063 106, 1982 (A. Bonaldi, E. Molinari).
48. A. Butenandt, U. Westphal, *Hoppe-Seyler's Z. Physiol. Chem.* **191** (1930) 127.
49. G. F. Marian, *Biochem. J.* **24** (1930) 435;1021.
50. A. Butenandt, U. Westphal, W. Hohlweg, *Hoppe-Seyler's Z. Physiol. Chem.* **227** (1934) 84. A. Butenandt, *Wiener Klin. Wochenschr.* **30** (1934) 934.
51. K. H. Slotta, H. Ruschig, E. Fels, *Ber. Dtsch. Chem. Ges.* **67** (1934) 1270.
52. W. M. Allen, O. Wintersteiner, *Science (Washington, D.C. 1883)* **80** (1934) 190. O. Wintersteiner, W. M. Allen, *J. Biol. Chem.* **107** (1934) 321.
53. M. Hartmann, A. Wettstein, *Helv. Chim. Acta* **17** (1934) 1365.
54. K. David, E. Dingemanse, J. Freud, E. Laqueur, *Hoppe-Seyler's Z. Physiol. Chem.* **233** (1935) 281.
55. T. Reichstein, C. W. Shoppee, *Vitam. Horm. (N.Y.)* **1** (1943) 345.
56. S. A. Simpson et al., *Experientia* **9** (1953) 333. S. A. Simpson et al., *Helv. Chim. Acta* **37** (1954) 1163.
57. P. A. Lehmann, A. Bolvar, R. Quintero, *J. Chem. Educ.* **50** (1973) 195.
58. A. J. Birch, *J. Chem. Soc.* 1950, 367.
59. D. H. R. Barton, *Experientia* **6** (1950) 316. D. H. R. Barton, *J. Chem. Soc.* 1953, 1027.
60. L. Dorfman, *Chem. Rev.* **53** (1953) 47.
61. W. Neudert, H. Röpke: *Steroid-Spektrenatlas,* Springer Verlag, Berlin 1965.
62. E. Heftmann (ed.): *Modern Methods of Steroid Analysis,* Academic Press, New York 1973.
63. A. Butenandt, P. Karlson, *Z. Naturforsch. B Anorg. Chem. Org. Chem.* **9 B** (1954) 389.
64. R. Huber, W. Hoppe, *Chem. Ber.* **98** (1965) 2403.
65. K. Nakanishi in K. Nakanishi et al. (eds.): *Natural Products Chemistry,* **vol. 1,** Academic Press, New York 1974, p. 527.
66. J. N. Kaplanis, W. E. Robbins, M. J. Thompson, S. R. Dutky, *Steroids* **27** (1976) 675.
67. K. Nakanishi et al., *Chem. Commun.* 1966, 915.
68. D. A. Shooley, G. Weiss, K. Nakanishi, *Steroids* **19** (1972) 377.
69. A. Prakash, S. Ghosal, *J. Sci. Ind. Res.* **38** (1979) 632.
70. U. Kerb et al., *Helv. Chim. Acta* **49** (1966) 1601.
71. A. Furlenmeier et al., *Helv. Chim. Acta* **50** (1967) 2387.
72. H. Mori, K. Shibata, K. Tsuneda, M. Sawai, *Chem. Pharm. Bull.* **16** (1968) 2416.
73. E. Lee, Y.-T. Liu, P. H. Solomon, K. Nakanishi, *J. Am. Chem. Soc.* **98** (1976) 1634.
74. D. H. R. Barton, P. G. Feakins, J. P. Poyser, P. G. Sammes, *J. Chem. Soc. C* 1970, 1584.
75. R. Pappo, *J. Am. Chem. Soc.* **81** (1959) 1010.
76. R. Goutarel, *Alkaloids (London)* **1** (1971) 382.
77. F. Kuony-Hu, R. Goutarel, *Alkaloids (London)* **3** (1973) 258.
78. F. Kuony-Hu, R. Goutarel, *Alkaloids (London)* **4** (1974) 246.
79. R. B. Herbert, *Alkaloids (London)* **3** (1973) 280.
80. K. Schreiber, *Pure Appl. Chem.* **21** (1970) 131.
81. R. Verpoorte, R. van der Heuden, W. M. van Gulik, H. J. G. ten Hoopen, *Alkaloids (N.Y.)* **40** (1991) 157.
82. R. B. Herbert, *Alkaloids (London)* **4** (1974) 383.

83. J. V. Greenhill, *Alkaloids (N.Y.)* **41** (1992) 177.
84. T. Tokuyama, J. Daly, B. Witkop, *J. Am. Chem. Soc.* **91** (1969) 3931.
85. J. S. Grossert, *Chem. Soc. Rev.* **1** (1972) 1.
86. A. B. Kundu, A. Mukherjee, A. K. Dey, *J. Sci. Ind. Res.* **35** (1976) 616.
87. G. P. Arsenault, K. Biemann, A. W. Barksdale, T. C. Mc. Morris, *J. Am. Chem. Soc.* **90** (1968) 5635.
88. T. C. Mc. Morris, A. W. Barksdale, *Nature* **215** (1967) 320.

Stilbene Dyes → Optical Brighteners
Stimulants → Psychopharmacological Agents

Strontium and Strontium Compounds

J. Paul MacMillan, Timminco Limited, Toronto, Ontario, Canada (Chaps. 1, 2, 7)

Jai Won Park, Solvay Barium Strontium GmbH, Hannover, Federal Republic of Germany (Chaps. 3 – 6, 8)

Rolf Gerstenberg, Solvay Barium Strontium GmbH, Hannover, Federal Republic of Germany (Chaps. 3 – 6, 8)

Heinz Wagner, Solvay Barium Strontium GmbH, Hannover, Federal Republic of Germany (Chaps. 3 – 6, 8)

Karl Köhler, Solvay Barium Strontium GmbH, Hannover, Federal Republic of Germany (Chaps. 3 – 6, 8)

Peter Wallbrecht, Solvay Barium Strontium GmbH, Hannover, Federal Republic of Germany (Chaps. 3 – 6, 8)

1. Introduction 1
2. Properties 1
3. Occurrence 2
4. Mining and Processing 2
5. Chemical Digestion of Ores 3
6. Strontium Compounds 3
6.1. Strontium Carbonate 3
6.2. Strontium Nitrate 4
6.3. Strontium Chloride 5
6.4. Strontium Chromate 5
7. Strontium Metal 6
8. Toxicology 7
9. References 8

1. Introduction

Strontium belongs to group 2 of the periodic table and is one of the alkaline earth metals. Strontium [*7440-24-6*], with atomic number 38, lies between calcium and barium in this group. Strontium was first discovered by A. Crawford in 1790. The discovery of strontium carbonate, which was originally thought to be barium carbonate, is associated with a lead mine located in Strontian, Scotland, hence the name strontium. In 1791 T. C. Hope confirmed Crawford's discovery, distinguishing strontium carbonate from calcium and barium carbonates.

In 1808 H. Davy [8] produced strontium metal from strontium carbonate originating from the Strontian deposit. Davy produced an oxide from the carbonate, and electrolyzed a mixture thereof with mercuric oxide to produce a strontium amalgam.

2. Properties

Strontium metal is silvery white and lustrous, but it readily forms a white surface coating when exposed to air and water. Strontium is allotropic and displays three crystal structures; at temperatures below 215 °C it is face-centered cubic. The crystal system properties are summarized in Table 1 . Some other properties are listed below:

Atomic number	38
Atomic mass	87.62
Stable isotopes, % abundance	
^{84}Sr [*15758-49-3*]	0.56
^{86}Sr [*13982-14-4*]	9.86
^{87}Sr [*13982-64-4*]	7.02
^{88}Sr [*14119-10-9*]	82.56
Density, g/cm^3	2.63
Melting point, °C	768
Boiling point, °C	1381
Heat of fusion, kJ/kg	104.7
Electrical resistivity, μΩ/cm	22.76

Table 1. Strontium crystal structures

Temperature, °C	Structure	Atomic radius, pm	Lattice constants, pm	
			a	*c*
< 215	fcc	213	607	
215 – 605	hcp	202	431	705
> 605	bcc	195	484	

Strontium is generally more reactive than magnesium and calcium and less reactive than barium. Strontium reacts with H_2O, O_2, N_2, F_2, and S producing compounds which correspond to its valence of two.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_321

Common practice for the analysis of strontium is based on the use of spectral emissions. Strontium and strontium compounds impart a bright crimson color to a flame. Some of the lines useful in the quantitative analysis of strontium are listed below (wavelength in nm):

Arc spectra	460.7, 421.6
Spark spectra	407.8, 421.6
Flame photometry	460.7, 407.8
Atomic absorption	460.7, 407.8

Strontium, while less reactive than the alkali metals, is a strong reducing agent and requires appropriate handling. Strontium reacts vigorously with water to form SrOH, liberating H_2 in the process.

3. Occurrence

Strontium is present in the earth's crust at an average concentration of 0.04 %, and is therefore 15th in abundance. In seawater it is the 10th most abundant element, having a concentration of 0.0008 % [9].

Of the naturally occurring strontium compounds, only the minerals strontianite (strontium carbonate, $SrCO_3$) and celestite (strontium sulfate, $SrSO_4$) are of economic importance.

Celestite can be formed as a direct precipitate in a series of marine salt deposits. Hydrothermal formation is also possible [10]. In addition, it seems possible that strontium sulfate could be formed by modification of primary strontium-bearing minerals (aragonite, calcite, anhydrite), whereby part of the strontium included in the crystal lattice is liberated and then precipitated as celestite by sulfate-containing solutions [11]. Strontianite can be formed both as a primary mineral from the hydrothermal phase and as a secondary mineral from, e.g., celestite [12].

Celestite occurs principally as nodules, lenses, beds and materials filling crevices in sedimentary rocks such as carbonates, gypsums, clays, and evaporites. The economic importance of celestite is considerably greater than that of strontianite.

Very large, exploited deposits occur in Spain, Mexico, Turkey, China, and Iran. There are much less important deposits in Algeria, Cyprus, Argentina, and Morocco. Mining in the United Kingdom, the traditional producer, ended in 1992. Little is yet known of the deposits and their extraction in the former eastern bloc [13], [14].

4. Mining and Processing

Celestite ores are obtained by both open-cast and underground mining. In most cases, only coarsely intergrown rich ores are worked, which can be further enriched by hand picking to give a material containing >90 % $SrSO_4$. Demands for increasing purity in chemically produced strontium salts together with the growth of mechanized mining techniques necessitate a continually increasing amount of processing.

Finely divided impurities are removed by desliming. By means of density separation, celestite ($\varrho = 3.8 - 3.9$ g/cm^3) is separated from limestone, quartz, gypsum, and dolomite, which have densities of <3 g/cm^3, giving end products containing >95 % $SrSO_4$. In Spain, a flotation plant for separating finely intergrown celestite ores has been in operation since 1990. The introduction of this technology has markedly increased the number of deposits worldwide that can be profitably worked.

Many important deposits cannot be mined because of unacceptably high levels of barium and deposit-specific trace elements that cannot be removed by available processing technologies.

Economic Aspects. Whereas strontianite was the principal starting material used in the production of strontium compounds between 1870 and 1920, it has now been almost completely replaced by celestite for this purpose.

Practically the only use for celestite is the production of other strontium compounds. It is used only to a small extent as a white filler in competition with ground barite (barium sulfate).

Table 2. Estimated production of celestite (10^3 t)

Country	1988	1991
Spain	45	70
Turkey	55	65
United Kingdom	20	2
China	20	25
Mexico	40	90
Iran	10	20
Morocco	2	4

World production of celestite in the early 1990s was estimated at ca. 250 000 – 300 000 t/a. The most important producing countries, with their production figures, are listed in Table 2.

5. Chemical Digestion of Ores

Reduction with Carbon. Celestite, flotated or ground to 0.1 – 1 mm, is chemically reduced according to the equation

$$SrSO_4 + 2\,C \longrightarrow SrS + 2\,CO_2$$

The reduction is carried out at 1100 – 1200 °C with coal in a rotary kiln, heated by fuel oil, natural gas, or coal dust in countercurrent flow. The conditions used depend in particular on the type of celestite and its impurities.

The light gray to grayish-black crude strontium sulfide (75 – 80 % SrS) produced is extracted with hot water. The insoluble solids are removed in a series of thickeners. The solution obtained is generally used to produce strontium carbonate.

Reaction with Sodium Carbonate. Ground celestite ($SrSO_4$) is converted to strontium carbonate by adding it to hot sodium carbonate solution. Numerous variations of this process have been proposed. Typically, an excess of Na_2CO_3 solution is used. The crude product so obtained is not usually acceptable without further treatment. A product sufficiently pure for most applications is obtained by dissolving the crude material in hydrochloric acid, neutralizing with sodium carbonate solution to precipitate iron and aluminum, filtering, and reprecipitating with caustic soda. This process also produces impure sodium sulfate solution as a byproduct which must be further treated or disposed of. The high consumption of sodium carbonate is a further drawback.

Other Processes. Celestite can be reacted with ammonium carbonate in a continuous countercurrent reactor to form strontium carbonate and ammonium sulfate [15]. The ammonium sulfate is treated with calcium oxide or calcium carbonate to form calcium sulfate, ammonia, and carbon dioxide, which are reused for producing ammonium carbonate.

In the Odda process for producing orthophosphoric acid, phosphate rock is reacted with nitric acid. The strontium present in many types of phosphate rock is converted to strontium nitrate, which remains in the insoluble residues if the concentrations of HNO_3 and $Ca(NO_3)_2$ are kept within certain limits. It can then be recovered by extraction with water.

6. Strontium Compounds

Like the compounds of the other alkaline earth metals, those of strontium are white or colorless. The sulfate and the carbonate are of low solubility, being intermediate between the corresponding compounds of calcium and barium in this respect.

Strontium compounds give a striking carmine red color to a flame. This property is used both for the rapid detection of strontium compounds and in pyrotechnics. Quantitative analysis of the industrial strontium compounds is usually carried out by means of titrimetry, X-ray fluorescence spectroscopy, and atomic absorption spectroscopy.

6.1. Strontium Carbonate

Strontium carbonate [*1633-05-2*], $SrCO_3$, M_r 147.63, ϱ 3.70 g/cm^3, *mp* 1497 °C (6.99 MPa), forms rhombic crystals (aragonite type) below 924 °C, and hexagonal crystals (calcite type) above this temperature. Its solubility at 25 °C is 10^{-3} g/100 mL.

Strontium carbonate is produced from strontium sulfide [*1314-96-1*], SrS.

Precipitation by Carbon Dioxide. Strontium carbonate is precipitated from the clarified solution of strontium sulfide (see Chap. 3) with carbon dioxide. Strontium carbonate and hydrogen sulfide are formed.

$$SrS + CO_2 + H_2O \longrightarrow SrCO_3 + H_2S$$

The carbon dioxide is obtained from natural wells or from industrial processes.

The reaction is carried out in gas-tight reactors, and is so controlled that the hydrogen sulfide produced has a low CO_2 content. The hydrogen sulfide is converted to sulfur in a Claus plant (→ Sulfur, Chap. 6.).

Precipitation with Sodium Carbonate. Sodium carbonate can be used instead of carbon dioxide in the above reaction.

$$SrS + Na_2CO_3 \longrightarrow SrCO_3 + Na_2S$$

In addition to strontium carbonate, a dilute solution of sodium sulfide is formed, which is concentrated to give a 60 % solution; the solidified product is marketable.

The choice of production method depends on the comparative availabilities of carbon dioxide and sodium carbonate, and the opportunities for exploiting the byproducts sulfur and sodium sulfide.

The two processes lead to products with different properties. The impurities in the strontium carbonate depend partly on the raw materials used. Also, the concentrations, temperatures, and reaction times lead to different values of bulk density, particle size distribution, and concentrations of sulfur and heavy metals.

The strontium carbonate powder is often converted into granules in a further process.

Quality Specifications. Typical specifications for strontium carbonate from the two precipitation processes are as follows:

$SrCO_3$	98.0 – 98.5 %
$BaCO_3$	1.0 – 2.5 %
Sulfur (as SO_3)	0.3 – 0.5 %
HCl insolubles	0.1 – 0.6 %
Fe_2O_3	0.005 – 0.01 %
For powder:	
Bulk density	0.4 – 1.2 g/cm^3
Fraction of particles <3 µm	95 – 100 %
For granules:	
Bulk density	1.2 – 1.8 g/cm^3
Grain size	
>850 µm	0 – 5 %
<150 µm	5 – 15 %

Transport. Strontium carbonate is stored in silos, and transported in 25 or 50 kg multi-ply paper sacks, 1 t big bags, or silo wagons. Strontium carbonate is not listed as a dangerous material in current regulations.

Uses. Strontium carbonate is by far the most important industrial compound of strontium. Its main use is in the manufacture of X-ray-absorbing glass for cathode ray tubes. Strontium carbonate is also used in the manufacture of ceramic permanent magnets (strontium ferrite), for removing lead from solutions of zinc sulfate in the electrolytic zinc process, and for the production of strontium metal, electroceramics, and oxide superconductors [14], [16]. Strontium carbonate is also the starting material for the production of other strontium compounds. For example, heating strontium carbonate with carbon in an electric furnace gives strontium oxide, which is used for production of strontium metal.

Between 1980 and 1990, world consumption of strontium carbonate increased from ca. 80 000 t/a to >150 000 t/a. Percentage consumption of strontium carbonate in 1990 was [14]:

Glass for cathode ray tubes	74
Electroceramics	12
Pyrotechnics	5
Electrolytic zinc extraction	2
Other uses	7

6.2. Strontium Nitrate

Strontium nitrate [*10042-76-9*], $Sr(NO_3)_2$, M_r 211.63, ϱ 2.986 g/cm^3, *mp* 570 °C, forms face-centered cubic crystals.

Below 29.3 °C, the solid phase in equilibrium with an aqueous solution on strontium nitrate is strontium nitrate tetrahydrate, $Sr(NO_3)_2 \cdot 4\,H_2O$. The solubility of strontium nitrate (in grams per 100 g solution) at various temperatures is as follows:

Temperature, °C	20	29.3	60	105
Solubility	40.7	47.0	48.3	51.2

Production. Strontium nitrate is obtained by dissolving strontium carbonate in 60 % nitric acid.

If the solution is maintained above ca. 30 °C during the evaporation, colorless crystals of the anhydrous compound are formed. After the carbonate, the nitrate is the strontium compound produced commercially in the largest quantities.

Uses. Strontium nitrate has long been used in pyrotechnics and as a component of signalling devices. Related devices are used in the United States and Japan as alternatives to portable electric lamps in motor trucks [14].

Packaging and Storage. Strontium nitrate is packed in 50 kg polyethylene bags. The following transport regulations apply:

RID/ADR:	Class 5.1, 22 C
IMDG Code:	Class 5.1, UN no. 1507

6.3. Strontium Chloride

Strontium chloride [*10476-85-4*], $SrCl_2$, M_r 158.53, ϱ 3.052 g/cm^3, *mp* 875 °C, forms cubic crystals. Its solubility at 20 °C is 34.5 wt %.

Production. Strontium chloride is obtained by dissolving strontium carbonate in concentrated hydrochloric acid. The hexahydrate, $SrCl_2 \cdot 6\,H_2O$ [*10025-70-4*], is formed on crystallizing below 61 °C. On dehydration, the hexahydrate dissolves in its water of crystallization at 61 °C. After passing through the di- and monohydrate stages, strontium chloride becomes fully dehydrated at 320 °C.

Uses. Strontium chloride is mainly used as an intermediate in the production of other strontium compounds. Compared with the sulfide, it has the advantage of not reacting with oxygen and carbon dioxide, which facilitates industrial handling.

Strontium chloride hexahydrate is used in toothpastes for sensitive teeth [17].

Strontium chloride is not classified as a dangerous material in transport regulations.

6.4. Strontium Chromate

Strontium chromate [*7789-06-2*], $SrCrO_4$, M_r 203.61, ϱ 3.895 g/cm^3, forms monoclinic crystals.

Strontium chromate is included in dangerous materials regulations [18] (symbol: T).

Production. Strontium chromate is produced by precipitating a solution of chromate ions with a strontium salt, or by reacting strontium hydroxide solution with chromic acid.

Uses. Strontium chromate (strontium yellow) is used as an anti-corrosion primer for aluminum and its alloys in aircraft and ship construction [19].

7. Strontium Metal

Production. Commercial strontium production relies on a variant of the Pidgeon process, exploiting the reaction proposed by GUNTZ [20]:

$$4\,SrO + 2\,Al \longrightarrow 3\,Sr + SrO \cdot Al_2O_3$$

Produced under vacuum, the strontium evolves as a gas and is transported through the production retort to a cooler end where precipitation occurs. The strontium precipitates as a solid and when the production cycle is complete the vacuum is broken and the product removed. Strontium is produced by Timminco Limited in Canada, Pechiney Electrometallurgie in France, and Cal-Stron Corporation in the United States. In addition, production may well occur in the CIS and China but the extent of production, if any, is unknown.

Strontium metal is commercially available as pure metal, and as an alloy with aluminum or aluminum and magnesium.

A typical percentage chemical analysis of commercial strontium is as follows:

Sr	99.0 min.
Mg	0.20
Ca	0.10
Ba	0.50
Na	0.07
N	0.03
Al	0.01
Fe	0.015

Transportation, Storage, and Safety. Due to its reactive nature strontium must be handled with care. The pure metal and the commercial eutectic alloy, 90 % Sr – 10 % Al, are classified as alkaline earth metals n.o.s. (not otherwise specified), UN no. 1393, for the purposes of transportation [21]. This classification is indicative of materials requiring special approved packaging for transport. These two forms of commercial strontium are packed under an inert atmosphere to prevent formation of an oxide coating. Strontium is a strong reducing agent and forms explosive hydrogen gas when in contact with water. Strontium is therefore classified as a flammable solid. Hence strontium is a Class 4.3 material, i.e., one that emits a flammable gas on contact with water. Strontium should therefore be stored in a well ventilated

dry place, avoiding all contact with water. It is recommended that protective gloves and glasses should be worn when handling pure strontium. However, commercial alloys of strontium, such as the Sr – Mg – Al master alloy or the family of aluminum – strontium master alloys which contain up to 10 percent strontium do not require these special precautions.

Uses. Metallic strontium and master alloys containing strontium have experienced significant expansion in application in modern foundry practice. In aluminum foundries strontium is used as an additive to Al – Si alloys to enhance their mechanical properties and improve the machining performance. Strontium improves the mechanical properties of Al – Si castings by changing the morphology of the eutectic silicon phase. In an unmodified Al – Si alloy this eutectic silicon phase precipitates as coarse acicular plates, which causes the casting to exhibit poor mechanical properties. Modification causes the silicon to assume a fine, interconnected fibrous morphology, resulting in higher tensile strength and greatly improved ductility. Impact resistance and machining characteristics are also enhanced through modification.

The optimum strontium level to achieve the best possible mechanical properties in hypoeutectic Al – Si alloys is 0.005 – 0.015 percent strontium [22]. The impact of strontium is also influenced by the solidification rate, with faster rates producing better properties.

Early foundry practice utilized sodium as a modification agent. In 1920, Pacz outlined a process by which the solidification structure of Al – Si alloys was modified by the addition of sodium to the melt. The sodium addition produced a fibrous eutectic structure which showed improved mechanical properties, especially elongation.

The principal drawback with sodium is that the effect of the sodium rapidly dissipates as sodium is lost by evaporation and/or oxidation during the holding of the molten alloy.

Strontium has the advantage of imparting a semi-permanent modifying effect, as it is less prone to the fading observed with sodium. This allows the production of so called pre-modified aluminum ingots which can be sold to end users. Nevertheless, many foundries for economic and logistical reasons chose to modify in house.

Strontium also improves the properties and processing of wrought aluminum alloys.

Strontium appears to act on both the mechanical and forming properties of 6xxx series wrought alloys. These improvements are reported to occur when, in the presence of strontium, the complex intermetallic compounds found in these alloys are precipitated in more favorable crystal structures [23]. Under normal or non-strontium treated conditions the formation of β-AlFeSi intermetallic is favored. The presence of this compound can contribute to surface defects upon extrusion of 6xxx series alloys. The controlled addition of strontium promotes the less detrimental α-AlFeSi leading to better surface quality in the extrusion and faster extrusion speeds.

In the 7xxx series of wrought alloys it has been reported that strontium, at levels of at least 0.005 percent, refines the intermetallic phases, including Mg_2Si, Al – Cu – Mg, and Al – Cu – Fe. When these phases are refined the alloy has enhanced toughness, finer grain size, and can be processed more quickly due to shortened homogenizing times compared to strontium-free alloys [24].

In an aluminum alloy sheet which requires high formability the presence of strontium is reported to reduce the number of intermetallic precipitates. This improves the alloy's formability and reduces the incidence of edge cracking [25].

In ferrous metallurgy strontium finds application as a component in certain foundry grades of ferrosilicon, used as inoculants in gray iron casting. The addition of 0.65 – 1.05 % strontium controls the carbon structure in an iron casting and inhibits the chill. The chill, which is in some cases a desired phenomena, is a rapidly solidified zone in a casting where the formation of mottled iron (cementite and graphite) is favored over a pearlite structure with graphite. By controlling the chill the foundry may be able to decrease the minimum section size in a casting keeping it free from the brittle carbides which would most likely form without inoculation. In addition strontium-bearing inoculants are particularly effective in helping to control internal shrinkage in these castings and thereby enhancing the pressure tightness of the casting [26].

8. Toxicology

Strontium carbonate is not considered to have any dangerous properties. There are no known occupational diseases associated with handling strontium carbonate, the most important industrial strontium compound.

The strontium ion is slightly toxic. The toxic action of strontium compounds is thus closely associated with the anion of the compound concerned. Strontium behaves similarly to calcium both chemically and biologically. The biological effects of strontium and its compounds have been described in detail [27].

Strontium is deposited preferentially in the bones and teeth of the human body. Strontium salts are not readily absorbed via the intestinal tract.

Symptoms of acute toxicity are excessive salivation, vomiting, colic and diarrhea, and possibly respiratory failure.

The LD_{50} values (oral) of various strontium salts (nitrate, hydroxide, chloride, fluoride) are >2000 mg/kg body weight [28].

Strontium nitrate was administered orally in the food of rats at a concentration of 50 mg/kg (1030 ppm daily for 8 weeks) without harmful effects. Strontium lactate at a level of 16 000 ppm, after more than one year, caused temporary cessation of growth and reduction of bone mineralization [14].

The inhalation of ca. 50 mg/m^3 strontium nitrate (4 h daily for one month, particle size mainly $< 5\,\mu$m) was less well tolerated by male rats. As well as local effects on the respiratory tract, functional and histological systemic changes took place, especially in the liver and kidneys [29].

Strontium oxide and strontium hydroxide have a strongly irritant effect on the skin and mucous membranes, especially the eyes.

No special hygienic measures are taken in the industrialized countries, except in the case of strontium chromate. The TLV/MAK value for strontium nitrate is 1 mg/m^3, and for strontium sulfate and strontium carbonate, 6 mg/m^3 [30]. However, these values are still the subject of debate.

Strontium chromate is included in the list of MAK values for carcinogenic industrial materials in Group III, A 2 [31]. The TRK of strontium chromate ($SrCrO_4$) is 0.1 mg/m^2, calculated as CrO_3.

9. References

General References

1. *Kirk-Othmer,* 2nd ed., **19,** 48 – 49.
2. *Metals Handbook,*9th ed., vol. 2, American Society of Metals, Metals Park, OH, 1979.
3. C. L. Mantell: "Strontium" in C. A. Hampel (ed.): *Rare Metals Handbook,* 2nd ed., Reinhold Publ. Co., London 1961, pp. 27 – 31.
4. J. E. Gruzleski, B. M. Closset: *The Treatment of Liquid Aluminum-Silicon Alloys,* American Foundrymen's Society, Des Plaines, IL, 1990, pp. 25 – 105.
5. *CRC Handbook of Chemistry and Physics,* 67th ed., CRC Press, Boca Raton, Flo. 1986.
6. R. D. Goodenough, V. A. Stenger: "Magnesium, Calcium, Strontium, Barium and Radium" in J. C. Bailar (ed.): *Comprehensive Inorganic Chemistry,* vol. 1, Pergamon Press, Oxford 1973, pp. 591 – 664.
7. *The Economics of Strontium 1992,* 6th ed., Roskill Information Services, London 1992.

Specific References

8. H. Davy: "Elements of Chemical Philosophy" in J. Davy (ed.): *The Collected Works of Sir Humphrey Davy,* vol. 4, Smith Elder and Company, Cornhill, 1840, p. 253.
9. B. Mason, C. B. Moore: *Grundzüge der Geochemie,* Enke Verlag, Stuttgart 1985.
10. H. Gundlach, *Geol. Jahrb.* **76** (1959) 637 – 712.
11. E. Usdowski, *Contrib. Mineral. Petrol.* **38** (1973) 177 –195.
12. H. Harder, *Beitr. Mineral. Petrogr.* **10** (1964) 198 – 215.
13. J. Griffiths, *Ind. Miner. (London)* **293** (1992) 21 –33.
14. Roskill: *The Economics of Strontium 1992,* 6th ed., Roskill Information Service Ltd., London 1992.
15. Proinsur, WO 920 16 30 A1, 1992 (A. Piedrafita Palacios).
16. H. Maeda, Y. Tanaka, M. Fikutomi, T. Asano, *Jpn. J. Appl. Phys.* **27** (1988) 209.
17. M. L. Schole et al., US 3 699 221, 1972.
18. EC Guideline 91/325, March 1, 1991, EC Number: 024-009-00-4.
19. US Department of the Interior, Bureau of Mines: Strontium – Uses, Supply, and Technology, IC 9213, Washington 1989.

20. A. Guntz, F. Benoit, *Bull. Soc. Chim. Fr.* **35** (1924) no. 4, p. 712.
21. International Maritime Organization (IMO): *International Maritime Dangerous Goods Code,* **vol. 3,** London 1990, p. 4325.
22. B. Closset, J. Gruzleski, *Metall. Trans. A* **13A** (1982) 945 – 951.
23. Alcan Research and Development, US 3 926 690, 1975 (L. R. Morris, F. B. Miners).
24. Aluminum Company of America, US 4 711 762, 1987 (W. D. Vernam, B. W. Lifka).
25. Kobe Steel Inc, JP 62-207 849, 1987 (N. Shinano).
26. *Metals Handbook,* 9th ed., **vol. 1,** American Society of Metals, Metals Park, OH, 1979, p. 21.
27. G. Thiedemann, *Wiss. Umwelt* **1** (1979) 48 – 51.
28. US Department of Health and Human Services: Registry of Toxic Effects Chemicals Substances, NIOSH-RTECS-File, Washington 1992.
29. *Patty,* 3rd ed., **2A,** 1894.
30. N. Ju. Tarasenko, Ju. V. Zjuzjukin, *Gig. Sanit.* **5** (1976) 28.
31. Deutsche Forschungsgemeinschaft (ed.): MAK- und BAT-Werte-Liste, Mitteilung 28, VCH Verlagsgesellschaft, Weinheim 1992.

Styrene

Polystyrene and Styrene Copolymers is a separate keyword.

Denis H. James, Dow Chemical, Freeport, Texas 77 541, United States

William M. Castor, Dow Chemical, Freeport, Texas 77 541, United States

1. Introduction 1
2. Physical Properties 2
3. Chemical Properties 3
4. Production 4
4.1. Catalytic Dehydrogenation of Ethylbenzene 4
4.1.1. Reaction Mechanisms 4
4.1.2. Adiabatic Dehydrogenation 5
4.1.3. Isothermal Dehydrogenation 5
4.1.4. Distillation of Crude Styrene 7
4.1.5. Oxidative Hydrogen Removal 7
4.2. Styrene – Propylene Oxide Process . 7
4.3. Styrene from Butadiene 9
4.4. Styrene from Toluene 10
4.5. Styrene from Pyrolysis Gasoline . . 10
5. Quality and Testing 10
6. Storage and Transportation 10
7. Uses and Economic Aspects 11
8. Related Monomers 12
8.1. Vinyltoluene 12
8.2. Divinylbenzene 13
8.3. α-Methylstyrene 14
8.4. Chlorostyrene 14
8.5. Vinylbenzyl Chloride 15
9. Toxicology and Occupational Health 15
10. References 16

1. Introduction

Styrene [*100-42-5*], also known as phenylethylene, vinylbenzene, styrol, or cinnamene, $C_6H_5{-}CH{=}CH_2$, is an important industrial unsaturated aromatic monomer. It occurs naturally in small quantities in some plants and foods. In the nineteenth century, styrene was isolated by distillation of the natural balsam storax [1]. It has been identified in cinnamon, coffee beans, and peanuts [2], and it is also found in coal tar.

The development of commercial processes for the manufacture of styrene based on the dehydrogenation of ethylbenzene [*100-41-4*] (→ Ethylbenzene) was achieved in the 1930s. The need for synthetic styrene – butadiene rubber (→ Rubber, 3. Synthetic, Chap. 2.1.) during World War II provided the impetus for large-scale production. After 1946, this capacity became available for the manufacture of a high-purity monomer that could be polymerized to a stable, clear, colorless, and cheap plastic (→ Polystyrene and Styrene Copolymers). Peacetime uses of styrene-based plastics expanded rapidly, and polystyrene is now one of the least expensive thermoplastics on a cost-per-volume basis.

Styrene itself is a liquid that can be handled easily and safely. The activity of the vinyl group makes styrene easy to polymerize and copolymerize. When the appropriate technology became available through licensors styrene was quickly transformed into a bulk-commodity chemical, growing to a world-wide capacity estimated at 17×10^6 t/a in 1993 and $18\text{x}10^6$ t/a in 1995 [54]. For reviews of this important industrial history see [3], [4].

2. Physical Properties

Styrene is a colorless liquid with a distinctive, sweetish odor. The most important physical properties of styrene are summarized below. Data appearing in the literature are not always consistent, varying as a consequence both of the method of measurement and the purity of the monomer. Most manufacturers make available the physical properties they regard as valid [5]. Important general physical properties of styrene monomer are as follows:

M_r	104.153
bp	145.15 °C
fp	−30.6 °C

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_329

Critical density, D_c	0.297 g/mL
Critical pressure, P_c	3.83 MPa
Critical temperature, T_c	362.1 °C
Critical volume, V_c	3.37 mL/g
Flammable limits in air	1.1 – 6.1 vol %
Flash point, Tag Closed Cup (TCC)	31.1 °C
Autoignition point	490 °C
Heat of combustion, ΔH_c, constant pressure (25 °C)∗	−4.263 MJ/mol
Heat of formation, ΔH_f	
gas (25 °C)	147.4 kJ/mol
liquid (25 °C)	103.4 kJ/mol
Heat of fusion, ΔH_m	−11.0 kJ/mol
Heat of polymerization, ΔH_p (25 °C)	−69.8 kJ/mol
Heat of vaporization, ΔH_v	
(25 °C)	421.7 J/g
(145 °C)	356.7 J/g
Volume expansion coefficient	
(20 °C)	9.783×10^{-4} °C^{-1}
(40 °C)	9.978×10^{-4} °C^{-1}
Q value	1.0
e value	0.8
Volume shrinkage on polymerization, typical	17.0 %
Solubility of oxygen (from air)	
(15 °C)	53 mg/kg
(25 °C)	50 mg/kg
(35 °C)	45 mg/kg

∗ All reactants and products are gases.

Vapor pressure is a key property in the design of styrene distillation equipment. It is essential that the same data set be used for both column design and column assessment. Vapor-pressure data we believe to be accurate are as follows:

Temperature, °C	Vapor pressure, kPa
20	0.6
50	3.2
80	12.2
100	25.7
145.2	101.3

Antoine equation:

$$\log_{10} P = [6.08201 - 1445.58/(209.43 + T)]$$

Styrene is miscible with most organic solvents in any ratio. It is a good solvent for synthetic rubber, polystyrene, and other non- cross-linked high polymers. Styrene is sparingly soluble in polyvalent hydroxy compounds such as glycol and diglycol monoethers. Styrene and water are sparingly soluble in each other (Table 1).

The composition of the azeotropic mixture at standard pressure is ca. 66 wt % styrene and 34 wt % water, and the minimum boiling point is 94.8 °C.

Important parameters for monitoring the production and use of styrene are the refractive index and the density (Table 2). Table 3 lists supplementary physical properties of styrene.

Table 1. Mutual solubilities, styrene/water, wt %

Temperature, °C	Water in styrene	Styrene in water
0	0.02	0.018
10	0.04	0.023
25	0.07	0.032
50	0.12	0.045

Table 2. Refractive index and density of styrene

Temperature, °C	Refractive index, n_D	Density, g/mL
0		0.9223
20	1.54682	0.9050
25	1.54395	0.9007
30	1.54108	0.8964
60		0.8702
100		0.8355

$n_D = 1.55830 - 0.000574\,T$

Table 3. Other physical properties of styrene

Temperature, °C	Viscosity, mPa s∗	Surface tension, mN/m	Specific heat capacity, C_p, J g^{-1} K^{-1}
0	1.039	34.5	1.6367
20	0.762	32.3	1.6907
40	0.588	30.0	1.7489
60	0.469	27.8	
80	0.385	25.6	
100	0.324	23.5	

∗ ln viscosity (mPa s) = − 4.2488 + 1170.8/T

3. Chemical Properties

The most important reaction of styrene is its polymerization to polystyrene, but it also copolymerizes with other monomers. The copolymerization with butadiene to give Buna S synthetic rubber (emulsion styrene – butadiene rubber, E-SBR) was the reaction that led initially to the development of the styrene industry. Other characteristic reactions are described below. The literature has been well summarized [3], [6], [7].

Oxidation of styrene in air is of special importance. The reaction leads to high molecular mass peroxides by way of free radicals. Styrene is also

oxidized to various other compounds, including benzaldehyde, formaldehyde, and formic acid. Other typical alkene reactions are observed with stronger oxidizing agents:

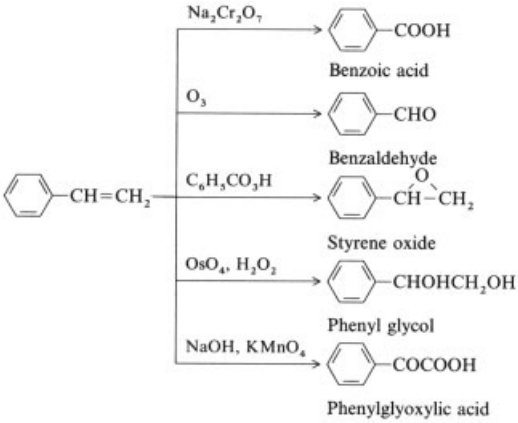

Styrene can undergo addition to the double bond of the side chain as well as substitution in the ring. Treatment with bromine in the cold gives addition, leading to styrene dibromide. This crystallizes well, and has been used both to characterize styrene and to determine it quantitatively in solution [8]. Splitting off hydrogen bromide from styrene dibromide gives α-bromostyrene, while phenylacetylene can be formed by heating with calcium oxide. Addition of chlorine forms styrene dichloride, which can be converted to α-chlorostyrene and β-chlorostyrene by elimination of hydrogen chloride.

The halohydrins are important intermediates in preparative chemistry. They react with alkali to form styrene oxide, while further hydrolysis leads to phenyl glycol. Iodohydrin is formed from styrene in the presence of iodine, mercury(II) oxide, and water. 1-Phenylethanol is an important intermediate product in the perfume industry (→ Flavors and Fragrances). It is formed by the hydration of styrene via quantitative addition of mercury(II) acetate:

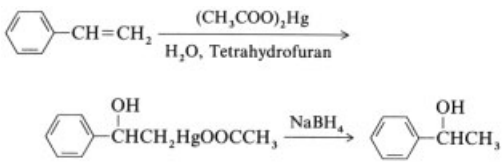

The corresponding methyl ether is obtained by addition of methanol (e.g., at 135 – 150 °C in the presence of sulfuric acid):

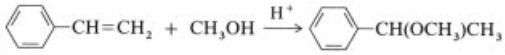

Numerous reactions of styrene with sulfur and nitrogen compounds have been reported [3]. Thus, styrene reacts with sulfur dioxide, sulfur monochloride, sodium or ammonium dithionite, mercaptans, aniline, amines, diazomethane, and sodium hydrazide to give cleavage of the C=C double bond. Heating styrene with sulfur at high temperature gives hydrogen sulfide, styrene sulfide, and diphenylthiophenes.

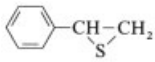

Styrene sulfide

Styrene forms solid complexes with copper and silver salts at low temperature. These compounds are suitable for the purification of styrene, or for separating styrene from mixtures with other hydrocarbons. The color of the copper complexes makes this metal and its alloys unsuitable for use in contact with styrene.

Styrene also undergoes many cyclization reactions [7]; e.g.:

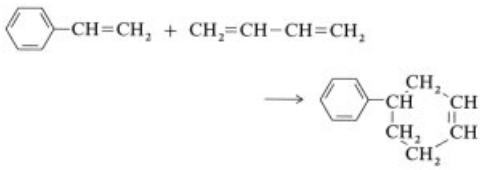

Lithium alkyls, such as C_2H_5Li, initiate a polymerization of styrene. In the presence of ethers, styrene is polymerized almost completely by sodium. Controlled polymerization (telomerization; →Organometallic Compounds and Homogeneous Catalysis, Chap. 3.2.4.) with olefins has generated a great deal of industrial interest, leading to new polymers that show promise as plasticizers, lubricants, and textile auxiliaries. Aromatic hydrocarbons such as benzene can be added to styrene in the presence

of aluminum chloride, just as they can to other olefins.

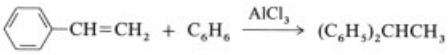

The high reactivity of the double bond, which is due to the resonance-stabilized aromatic ring, usually means that reactions must be carried out in several steps, with the double bond protected until the final step.

4. Production

Production of styrene is world-wide, and dates back more than 50 years. However, new variations on established processes are continually being developed, and new companies are building plants or buying out older producers. The feedstock for all commercial styrene manufacture is still ethylbenzene (→ Ethylbenzene). This is converted to a crude styrene that requires finishing to separate out the pure product.

4.1. Catalytic Dehydrogenation of Ethylbenzene

Direct dehydrogenation of ethylbenzene to styrene accounts for 85 % of commercial production. The reaction is carried out in the vapor phase with steam over a catalyst consisting primarily of iron oxide. The reaction is endothermic, and can be accomplished either adiabatically or isothermally. Both methods are used in practice.

4.1.1. Reaction Mechanisms

The major reaction is the reversible, endothermic conversion of ethylbenzene to styrene and hydrogen:

$$C_6H_5CH_2CH_3 \rightleftharpoons C_6H_5CH{=}CH_2 + H_2$$
$$\Delta H\,(600\,^\circ C) = 124.9\,\text{kJ/mol}$$

This reaction proceeds thermally with low yield and catalytically with high yield. As it is a reversible gas-phase reaction producing 2 mol of product from 1 mol of starting material, low pressure favors the forward reaction.

Competing thermal reactions degrade ethylbenzene to benzene, and also to carbon:

$$C_6H_5CH_2CH_3 \longrightarrow C_6H_6 + C_2H_4 \quad \Delta H = 101.8\,\text{kJ/mol}$$

$$C_6H_5CH_2CH_3 \longrightarrow 8\,C + 5\,H_2 \quad \Delta H = 1.72\,\text{kJ/mol}$$

Styrene also reacts catalytically to toluene:

$$C_6H_5CH{=}CH_2 + 2\,H_2 \longrightarrow C_6H_5CH_3 + CH_4$$

The problem with carbon production is that carbon is a catalyst poison. When potassium is incorporated into the iron oxide catalyst, the catalyst becomes self- cleaning (through enhancement of the reaction of carbon with steam to give carbon dioxide, which is removed in the reactor vent gas).

$$C + 2\,H_2O \longrightarrow CO_2 + 2\,H_2 \quad \Delta H = 99.6\,\text{kJ/mol}$$

Typical operating conditions in commercial reactors are ca. 620 °C and as low a pressure as practicable. The overall yield depends on the relative amounts of catalytic conversion to styrene and thermal cracking to byproducts. At equilibrium under typical conditions, the reversible reaction results in about 80 % conversion of ethylbenzene. However, the time and temperature necessary to achieve equilibrium give rise to excessive thermal cracking and reduced yield, so most commercial units operate at conversion levels of 50 – 70 wt %, with yields of 88 – 95 mol %.

Dehydrogenation of ethylbenzene is carried out in the presence of steam, which has a threefold role:

1) It lowers the partial pressure of ethylbenzene, shifting the equilibrium toward styrene and minimizing the loss to thermal cracking
2) It supplies the necessary heat of reaction
3) It cleans the catalyst by reacting with carbon to produce carbon dioxide and hydrogen.

Many catalysts have been described for this reaction [9]. One catalyst, Shell 105, dominated the market for many years, and was the first to include potassium as a promoter for the water-gas reaction. This catalyst is typically 84.3 % iron as Fe_2O_3, 2.4 % chromium as Cr_2O_3, and 13.3 %

potassium as K_2CO_3. It has good physical properties and good activity, and it gives fair yields [10].

More recently, increasing competition has promped manufacturers to seek new catalysts that produce higher yields without compromising activity or physical properties, or catalysts that meet specific requirements. The Süd-Chemie Group, which includes Nissan Girdler Catalyst in Japan, Süd-Chemie in Germany, and United Catalysts (UCI) in the United States, now has the major share of the catalyst market with its G-64 and G-84 types [11]. Shell also remains active through a joint partnership with American Cyanamid called Criterion Catalyst. In addition to Criterion 105, a series of new Criterion catalysts is available, including C-115 and C-025 HA. Dow Chemical and BASF manufacture their own catalysts to suit their specific needs, and there are other small producers as well. A catalyst life of ca. two years is claimed.

4.1.2. Adiabatic Dehydrogenation

Over 75 % of all operating styrene plants carry out the dehydrogenation reaction adiabatically in multiple reactors or reactor beds operated in series (Fig. 1). The necessary heat of reaction is applied at the inlet to each stage, either by injection of superheated steam or by indirect heat transfer.

Fresh ethylbenzene feed is mixed with recycled ethylbenzene and vaporized. Dilution steam must be added to prevent the ethylbenzene from forming coke. This stream is further heated by heat exchange, superheated steam is added to bring the system up to reaction temperature (ca. 640 °C), and the stream is passed through catalyst in the first reactor. Adiabatic reaction drops the temperature, so the outlet stream is reheated prior to passage through the second reactor. Conversion of ethylbenzene can vary with the system, but is often about 35 % in the first reactor and 65 % overall. The reactors are run at the lowest pressure that is safe and practicable. Some units operate under vacuum, while others operate at a low positive pressure. The steam: ethylbenzene ratio fed to the reactors is chosen to give optimum yield with minimum utility cost. The reactor effluent is fed through an efficient heat recovery system to minimize energy consumption, condensed, and separated into vent gas, a crude styrene hydrocarbon stream, and a steam condensate stream. The crude styrene goes to a distillation system. The steam condensate is steam-stripped, treated, and reused. The vent gas, mainly hydrogen and carbon dioxide, is treated to recover aromatics, after which it can be used as a fuel or a feed stream for chemical hydrogen. The complete technology is for sale from various licensors (Chap. 7).

4.1.3. Isothermal Dehydrogenation

Isothermal dehydrogenation (Fig. 2) was pioneered by BASF and has been used by them for many years. The reactor is built like a shell-and-tube heat exchanger. Ethylbenzene and steam flow through the tubes, which are packed with catalyst. The heat of reaction is supplied by hot flue gas on the shell side of the reactor – exchanger [12]. The steam : oil mass ratio can be lowered to about 1 : 1, and steam temperatures are lower than in the adiabatic process. A disadvantage is the practical size limitation on a reactor – exchanger, which restricts the size of a single-train plant to about 150×10^3 t/a, translating into increased capital for large plants.

Lurgi GmbH operates an isothermal reactor system that uses a molten salt mixture of sodium, lithium, and potassium carbonates as the heating medium. The multitubular reactor is operated at ca. 600 °C under vacuum and a steam : ethylbenzene ratio of 0.6 – 0.9. High conversion and selectivity are claimed [13]. A demonstration plant has been operated since 1985 at Mantova, Italy, by Montedison. The technology is offered for license by Lurgi, Montedison, and Deggendorfer, but so far no more units have been built.

4.1.4. Distillation of Crude Styrene

A typical crude styrene from the dehydrogenation process consists of:

Benzene (*bp* 80 °C)	1 %
Toluene (*bp* 110 °C)	2 %
Ethylbenzene (*bp* 136 °C)	32 %
Styrene (*bp* 145 °C)	64 %
Others	1 %

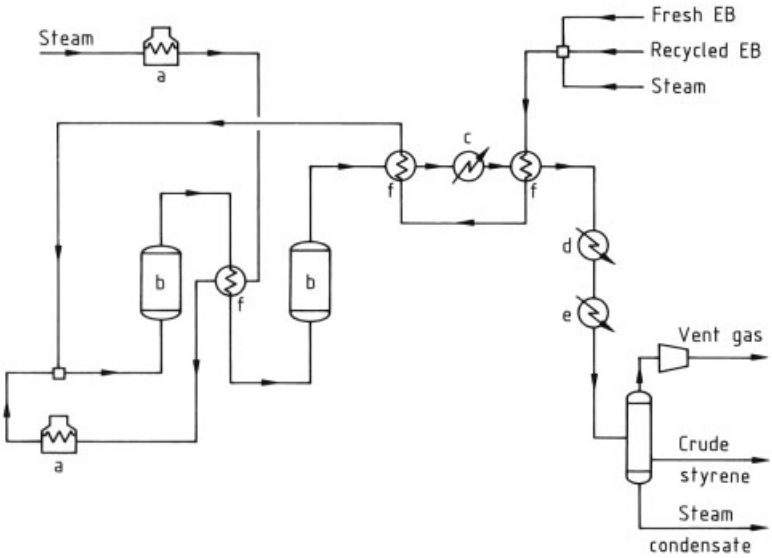

Figure 1. Adiabatic dehydrogenation of ethylbenzene (EB)
a) Steam superheater; b) Reactor; c) High-pressure steam; d) Low-pressure steam; e) Condenser; f) Heat exchanger

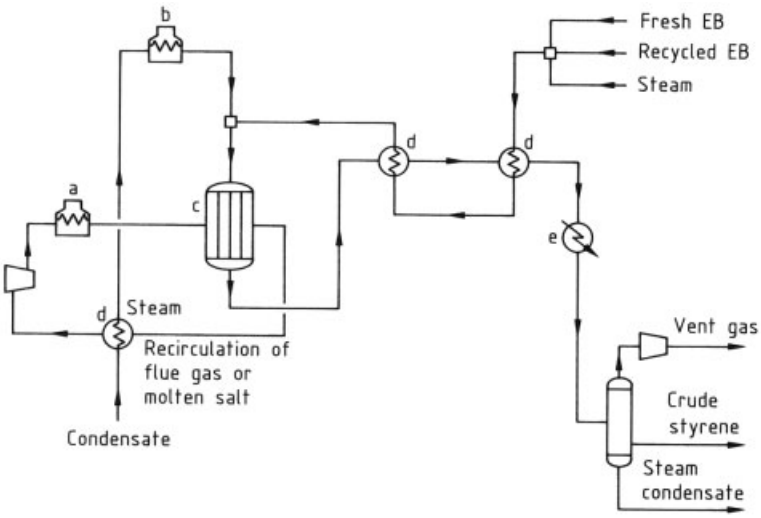

Figure 2. Isothermal dehydrogenation of ethylbenzene (EB)
a) Heater; b) Steam superheater; c) Reactor; d) Heat exchanger; e) Condenser

The separation of these components is reasonably straightforward, but residence time at elevated temperature needs to be minimized to reduce styrene polymerization. At least three steps are involved. Benzene and toluene are removed first, and either sent to a toluene dehydrogenation plant or further separated into benzene for recycling and toluene for sale. Ethylbenzene is then separated and recycled to the reactors. Finally, styrene is distilled away from the tars and polymers under vacuum to keep the temperature as low as possible. Figure 3 A shows a typical distillation train. The variant shown in Figure 3 B is the Monsanto approach, where the major split is accomplished first, followed by separation of the benzene – toluene mixture.

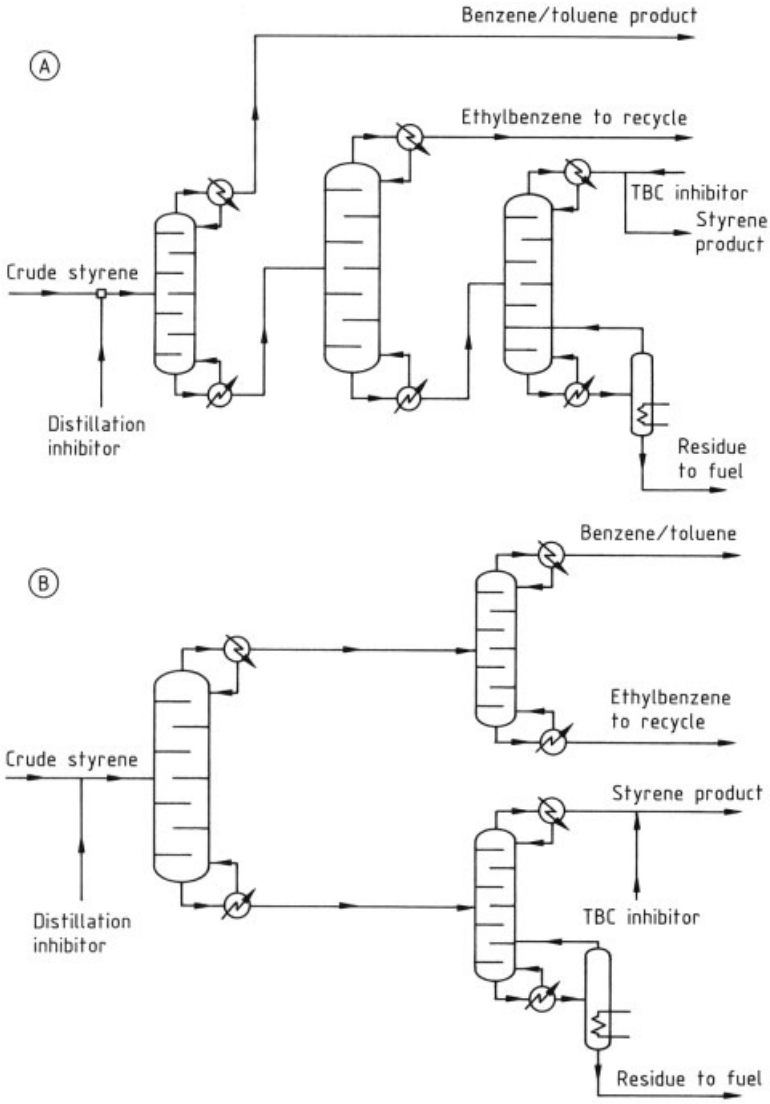

Figure 3. Distillation of crude styrene
A) Standard approach; B) Monsanto approach

Ethylbenzene and styrene, having similar boiling points, require 70 – 100 trays for their separation depending on the desired ethylbenzene content of the finished styrene. If bubble-cap trays are used, as in old plants, a large pressure drop over the trays means that two columns in series are necessary to keep reboiler temperatures low. Low-pressure drop trays, such as the Linde UCC sieve tray [14], permit this separation to be achieved in one column [15]. The most modern plants use packing in place of trays [16], [17]. This results in less pressure drop, giving a lower bottom temperature, shorter residence time, and hence less polymer (→ 12. Energy Management in Chemical Industry, Chap. 6.6., → 12. Energy Management in Chemical Industry, Chap. 7.). Sulzer has done pioneering work in the field of packings for distillation [18]. Koch, Norton, and Glitch also produce packings for this purpose.

A polymerization inhibitor (distillation inhibitor) is needed throughout the distillation

train. Sulfur was originally used, but environmental constraints make sulfur tar unacceptable as a fuel. Many new inhibitors have been marketed, usually aromatic compounds with amino, nitro, or hydroxy groups (e.g., phenylenediamines, dinitrophenols, and dinitrocresols). Uniroyal is especially active in this field. The distillation inhibitor tends to be colored and is thus unacceptable in the final product; finished monomer is usually inhibited instead with *tert*-butylcatechol (TBC) (10 – 50 mg/kg) during storage and transportation.

4.1.5. Oxidative Hydrogen Removal

Many efforts have been made to remove hydrogen from the process to favor the forward reaction. A new process is now being offered for licensing by Lummus, based on technology of UOP [19]. It employs a noble metal catalyst that selectively oxidizes hydrogen, allowing the ethylbenzene conversion to be increased to over 80 % while maintaining good styrene selectivity [20]. The process was originally named Styro-Plus, but is now called the SMART SM Process. It has performed well in a pilot unit (Mitsubishi Petrochemical, Japan) since 1985. It is mainly offered as a retrofit to existing units as a way of gaining extra capacity. It is claimed to have higher variable costs, but a lower capital cost than a conventional plant. Several licenses have been sold, but the corresponding plants are not yet operating.

4.2. Styrene – Propylene Oxide Process

The only other route to commercial production of styrene involves coproduction of propylene oxide, illustrated in simplified form in Fig. 4 [21] (→ Propylene Oxide, Chap. 4.2.). The first step is direct air oxidation of ethylbenzene at ca. 130 °C and 0.2 MPa. This gives ethylbenzene hydroperoxide (EBHP); α-methylbenzyl alcohol (MBA) and acetophenone (ACP) are also formed. Conversion is held to about 13 % in this step to minimize byproducts. The selectivity for ethylbenzene to EBHP is approximately 90 %, and the selectivity to MBA and ACP is 5 – 7 %.

$$C_6H_5C_2H_5 + O_2 \longrightarrow C_6H_5CH(OOH)CH_3 \text{ EBHP}$$

$$C_6H_5C_2H_5 + 1/2\,O_2 \longrightarrow C_6H_5CH(CH_3)OH \text{ MBA}$$

$$C_6H_5C_2H_5 + xO_2 \longrightarrow C_6H_5CO(CH_3) + \text{acids ACP}$$

Ethylbenzene hydroperoxide is then reacted with propylene in the presence of a metallic catalyst to form propylene oxide (PO) and more MBA. Liquid-phase molybdenum or heterogeneous titanium catalysts are used at ca. 110 °C and 4 MPa. The conversion of EBHP is nearly complete, with 70 – 85 % selectivity to PO and > 70 % selectivity to MBA.

$$C_6H_5CH(OOH)CH_3 + C_3H_6 \longrightarrow C_3H_6O + C_6H_5CH(CH_3)OH \text{ PO}$$

To improve yields, ACP is hydrogenated to MBA in the liquid phase at 90 – 150 °C and ca. 8 MPa. The catalyst is a mixture of ZnO and CuO. Approximately 90 % of the ACP is converted, with 92 % selectivity to MBA.

$$C_6H_5CO(CH_3) + H_2 \longrightarrow C_6H_5CH(CH_3)OH$$

Finally, the MBA is dehydrated to styrene at 250 °C and low pressure over a suitable metal oxide catalyst (often Al_2O_3).

$$C_6H_5CH(CH_3)OH \longrightarrow C_6H_5CH{=}CH_2 + H_2O$$

The patents issued to Halcon, Arco, and Shell on this process are extensive. The resulting ratio of styrene to PO is about 2.5. The process requires clean-up and purification steps to generate specification products.

The first commercial development of the process was by a joint venture company, Oxirane, formed by Halcon and Atlantic Richfield (ARCO); ARCO became the sole owner in 1980. Independently, Shell developed its own process based on similar chemistry. This route offers a way to manufacture propylene oxide without the need for chlorine as in the chlorohydrin process; (→ Propylene Oxide, Chap. 4.1.), but it is subject to market fluctuations in styrene and PO demand that do not match the stoichiometry of the process. The process requires a large capital investment and produces styrene at a higher cost than the conventional process, but credits for the coproduct PO can make the overall operation profitable. About 15 % of the world's supply of styrene is now made by this process.

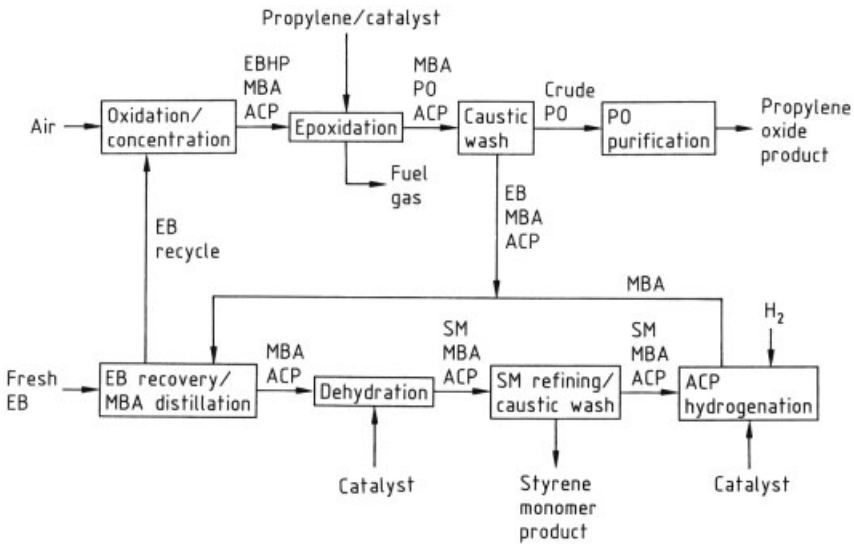

Figure 4. Styrene – propylene oxide process
ACP acetophenone; EB ethylbenzene; EBHP ethylbenzene hydroperoxide; MBA methylbenzyl alcohol; PO propylene oxide; SM styrene monomer

4.3. Styrene from Butadiene

Another route to styrene that is being heavily researched starts with the Diels – Alder dimerization of 1,3-butadiene to 4-vinylcyclohexene-1 (VCH) [*100-40-3*] (→ Butadiene, Chap. 3.):

$$2\,C_4H_6 \longrightarrow C_6H_9CH{=}CH_2$$

This reaction is exothermic, and can be accomplished either thermally or catalytically. Thermal processes [22], [23] require a temperature of ca. 140 °C and a pressure of ca. 4 MPa. The thermal approach is most suitable for use with purified butadiene. Yields of VCH are ca. 90 %. The catalytic process is based on nitrosyl halide – iron complexes, and operates at 0 – 80 °C and 0.1 – 1.30 MPa [24], [25]. The yield is almost quantitative, and purification is not difficult. This route can accept either the raw C_4 stream from a naphtha or gas-oil steam cracker, or purified butadiene.

VCH is then dehydrogenated to ethylbenzene [26], [27] or, under more severe conditions, oxidatively dehydrogenated directly to styrene [28], [29]:

$$C_6H_9CH{=}CH_2 \longrightarrow C_6H_5CH_2CH_3 \longrightarrow C_6H_5CH{=}CH_2$$

This route to styrene is not yet economically attractive, but the availability and price of butadiene in the future could make it so.

4.4. Styrene from Toluene

There have been many attempts to find a route to styrene starting from toluene (→ Toluene). Toluene is readily available, is usually at least 15 % cheaper than benzene, and is not as toxic. However, no process has become commercially competitive. Monsanto worked extensively on a process for styrene starting with air oxidation of toluene to give stilbene [30]. This used a fluidized bed of supported lead oxide catalyst:

$$2\,C_6H_5CH_3 + O_2 \longrightarrow C_6H_5CH{=}CHC_6H_5 + H_2O$$
$$\Delta H = -\,77.26\ \mathrm{kJ\,g^{-1}\,mol^{-1}}$$

Stilbene is then reacted with ethylene over a molybdenum catalyst to give styrene:

$$C_6H_5CH{=}CHC_6H_5 + C_2H_4 \longrightarrow C_6H_5CH{=}CH_2$$
$$\Delta H = -\,15.99\ \mathrm{kJ\,g^{-1}\,mol^{-1}}$$

The idea attracted interest, but the project has been formally abandoned.

Another route being researched is the alkylation of toluene with methanol over zeolite catalysts [31–33]. The selectivity with respect to

toluene is claimed to be high, but it is only about 50 % based on methanol. No commercial plants have been announced.

4.5. Styrene from Pyrolysis Gasoline

An aromatic mixture ("pyrolysis gasoline;" → Benzene, Chap. 5.2.2.) including styrene is obtained from the thermal cracking of naptha or gas-oil. Recovery of the styrene has been proposed on the basis of extractive distillation with dimethylformamide or dimethylacetamide, by adsorption, by complex formation, or by membrane separation. So far none of these processes has been commercially exploited.

5. Quality and Testing

The specifications and methods of analysis for styrene have changed through the years. Almost all manufacturers use ASTM D 2827-92 in their sales specifications, which calls for a minimum purity of 99.7 %, but much of the styrene in today's competitive market is of much higher purity. Historically, styrene purity has been determined by freezing point, the method referred to in ASTM D 2827-92. Gas chromatography (GC) is used to determine specific impurities, and in practice, most manufacturers now use GC to determine overall purity by subtracting the total impurities from 100 %. Table 4 lists ASTM specifications and typical analysis for styrene monomer.

Table 4. ASTM styrene specifications (1993)

	ASTM D 2827-92	Typical analysis	ASTM method
Purity, %	99.7	99.8	D 3799*
Color, Pt – Co, max.	10	7	D 1209
Aldehydes, mg/kg	200	30	D 2119
Peroxides, mg/kg	100	5	D 2340
Polymer, mg/kg	10	0	D 2121
Inhibitor, mg/kg	10 – 15	12	D 4590
Impurities			D 5135**

* By freezing point.
** By gas chromatography.

The major impurities in styrene monomer are a function of the process variables. Ethylbenzene content varies depending on the effort put into the main ethylbenzene – styrene distillation column. α-Methylstyrene, isopropylbenzene (cumene), *n*-propylbenzene, and minor amounts of ethyltoluene and vinyltoluene in finished styrene are a function of the separation power of the styrene still. Xylene content is influenced by the purity of the feed ethylbenzene. Phenylacetylene is produced by the dehydrogenation process in amounts depending on the catalyst used, typically 50 – 150 mg/kg. Phenylacetylene must be removed by a hydrogenation step; it is absent from material made by the styrene – propylene oxide process.

6. Storage and Transportation

Styrene is a flammable, reactive monomer. It has a flash point of 31 °C and a flammable range of 1.1 – 6.1 vol % in air. It undergoes exothermic polymerization quite readily (0.02 %/h at 25 °C) liberating 69.8 kJ per mole of reacted monomer. Failure to remove this heat from an enclosed container could theoretically produce a temperature of 300 °C; uncontrolled polymerization may therefore lead to pressure build-up in a closed container, or a potentially explosive vapor cloud. However, styrene has been used on a large scale industrially for more than 50 years, and extensive experience is available with respect to preventive measures against fire and explosion, permitting problem-free storage and safe shipping of the monomer. Data are available from all styrene producers [5], [34].

Transportation of styrene monomer is subject to regulation in all countries (e.g., DOT in the United States and the VbF guidelines in Germany). The DOT identification number is UN 2055, and the reportable quantity for a spill is 454 kg.

Leakage can easily be detected owing to the characteristic styrene odor and a low odor threshold (0.005 mg/kg in air).

To increase its shelf life, styrene monomer is inhibited, typically with 4-*tert*-butylcatechol (TBC). Hydroquinone has also been used, but it is not as effective. At a TBC level of 12 mg/kg, a shelf life of 6 months is predicted at 20 °C; this falls to 3 months at 30 °C. This inhibitor requires trace amounts of oxygen to render it effective. To ensure a margin of safety, the recommended minimum oxygen level in the monomer

is 15 mg/kg. The solubility of oxygen from the air in styrene monomer is 50 mg/kg. To prevent the occurrence of hazardous polymerization, styrene should always be kept cool, and appropriate inhibitor and oxygen levels should be maintained.

Styrene in storage is frequently padded with inert gases, which are partially soluble in the monomer. The consequences of desorption of these gases must be considered in subsequent handling. When the oxygen level in the inert gas pad is < 8 vol %, the possibility of fire or explosion is eliminated.

Storage and shipping containers for styrene may be of standard steel or aluminum. Rust acts as a catalyst for the polymerization of styrene, so inorganic zinc linings are recommended for storage tanks. Copper and brass fittings must be avoided because they can lead to discoloration of the styrene. Contamination with bases must also be avoided, as these react with the inhibitor, making it ineffective. Loading on a common carrier in a compartment next to a heated product is an unacceptable practice.

Foam is the preferred medium for fighting styrene fires. Water fog dissipates vapor clouds and provides cooling for structural supports, but rarely extinguishes the fire. Styrene monomer, with its high volume resistivity, can acquire and hold a static electric charge during transfer. Effective grounding measures must therefore be taken to eliminate uncontrolled electric discharge.

7. Uses and Economic Aspects

Styrene monomer is used as a feedstock in a variety of polymer products: thermoplastics, elastomers, dispersions, and thermoset plastics. The following breakdown represents an estimate of the distribution of the monomer, as well as some of the uses of its derivatives [35].

Approximately 65 % of the styrene produced goes into polystyrene. This is used to make a wide range of products, from toys, housings for room air conditioners, and television cabinets, to cassettes, combs, and furniture parts. Polystyrene can also be foamed to produce insulation board, loose-fill packaging, and disposable food containers.

Approximately 6 % still goes into styrene – butadiene rubber elastomers (SBR) for such uses as passenger car tires, industrial hoses, and footwear.

Approximately 7 % goes to styrene – butadiene latexes, used in tufted carpet, paper coatings, and as a component in latex paints.

Approximately 9 % goes into styrene – acrylonitrile copolymer (SAN) and terpolymers of acrylonitrile, butadiene, and styrene (ABS) (→ Polystyrene and Styrene Copolymers); SAN is used for drinking tumblers and battery cases, ABS for piping, automotive components, refrigerator doorliners, and shower stalls.

A further 7 % is combined with unsaturated polyester resins in fiberglass-reinforced boats, storage tanks, shower units, and simulated marble products.

The remainder goes to miscellaneous uses, especially involving blends with other thermoplastics for a constantly growing line of products, including ion-exchange resins and adhesives.

World capacity for styrene in 1993 was estimated at ca. 17×10^6 t/a, in 1995 ca 18×10^6 t/a, divided among the most important markets as follows:

	1993	1995
North America	35 %	34 %
Western Europe	27 %	23 %
Japan	16 %	-
Korea	7 %	-
Far East (remainder)	5 %	-
Eastern Europe	5 %	6 %
South America	4 %	2.2 %
Middle East	1 %	3.3 %

Over 80 companies are now producing styrene monomer, but the pace of announcements of new plants, shutdown of old plants, and business mergers makes this a constantly changing number.

The majority of the producing companies have licensed their technology. A very rough estimate of capacities for the various technologies is as follows:

Badger	40 %
ABB Lummus Crest	24 %
PO/Styrene	13 %
Dow Chemical	11 %
BASF	7 %
Miscellaneous	5 %

Badger, a Raytheon Company, entered the styrene business in 1955. They acquired and developed portions of the Cosden, Union Carbide, and Fina styrene technology. Lummus Crest is part of the ABB group, and has acquired and developed the Monsanto styrene process, also merging technologies and resources with UOP and Unocal. ARCO and Shell own their own styrene – PO technologies. Dow and BASF have their own dehydrogenation technologies, and have both been making styrene for over 50 years. The size and age of the industry means that it is well reported and analyzed [34–36].

World-scale single-train units have a capacity of 450×10^3 t/a, but units down to 20×10^3 t/a still find their niche.

Benzene [*71-43-2*] (→ Benzene), used for making ethylbenzene, accounts for ca. 65 % of the cost of styrene. Large swings in the price of benzene are reflected in that of styrene, which also responds to variation in supply and demand. The price of styrene has fluctuated from a high of $ 1100/t in the late 1980s to $ 450/t in 1992 [37]. High profit margins tempt new producers into the business, and because they can buy their way in quickly via the licensors, overcapacity is easily reached (as in 1993), giving rise to a classic business cycle.

8. Related Monomers

Styrene can be substituted in the side chain or in the ring to give many other compounds; only the following have attracted commercial interest.

8.1. Vinyltoluene

Vinyltoluene, [*25013-15-4*] *p*-methylstyrene, $CH_3C_6H_4CH{=}CH_2$, is a specialty monomer, with properties similar to those of styrene [38]. It is available commercially in two different isomer mixes. For many years it has been supplied as a 68/32 mixture of the *meta* and *para* isomers, reflecting the composition of the feedstock ethyltoluene (ethylmethylbenzene), from the Friedel – Crafts synthesis (→ Hydrocarbons, Chap. 3.6.). This is usually called vinyltoluene (VT). The advent of a *para*-specific route to ethyltoluene using zeolite technology has made available a 97 % *para*-isomer product called PMS [*627-97-9*] to differentiate it from the original mix. The VT mix was originally developed to compete with styrene, with the idea that its feedstock (toluene) would give it a competitive advantage over a benzene-based material. However, benzene remained cheap and available, and VT proved to be more costly to make than styrene, so it has not gained a large market share. It has nevertheless maintained a specialty position because of its unique properties. Table 5 lists the physical properties of VT and PMS.

Table 5. Physical properties of VT and PMS

	VT	PMS
M_r	118.18	118.18
bp, °C	171.2	170.0
fp, °C	− 77	− 34
Flammable limits, vol % in air	1.1 – 5.2	
Flash point, TCC, °C	52.8	
Autoignition point, °C	575	
Heat of polymerization, ΔH_p, kJ/mol (25 °C)	− 70	− 70
Heat of vaporization, ΔH_v, J/g (25 °C)	426.1	410.9
Density, g/mL (25 °C)	0.8930	0.8920
Refractive index, n_D^{25}	1.5395	1.5408
Specific heat capacity (liquid), $J\,g^{-1}\,K^{-1}$ (20 °C)	1.715	
Viscosity, mPa s (25 °C)	0.79	0.79
Surface tension, mN/m (25 °C)	31.0	34.0
Solubility in water, mg/kg (25 °C)	90	

The reactivity of VT is similar to that of styrene. Thus, it undergoes both homopolymerization and copolymerization. Polymerization can be initiated by exposure to strong acids, peroxides, perchlorates, or hypochlorites. It also reacts similarly to styrene with oxygen and halogens, although the presence of an extra side chain makes the chemistry more complex.

The dehydrogenation of ethyltoluene to vinyltoluene is similar to that described for ethylbenzene to styrene. It is a vapor-phase process, employing superheated steam and a suitable catalyst. The amount of *o*-ethyltoluene in the feed should be kept at a minimum because it can form indene in the cracker, which impairs the properties of the final polymer. Distillation of the crude product must be performed carefully under vacuum to limit the temperature and thus the rate of polymerization. A suitable polymerization inhibitor is also necessary. The complexity of the crude reactor product requires elaborate separation by distillation to give a specification

product. The monomer is stored under refrigeration at 10 °C, and it is typically inhibited with TBC (12 mg/kg) during storage and transportation.

Toxicologically, VT has been considered to be very similar to styrene, and normal precautions should be taken when handling it (see Chap. 9). The ACGHI TLV is 50 mg/kg. VT is flammable, but its vapor pressure does not produce a flammable mixture in air below 38 °C. It polymerizes more rapidly than styrene, and polymerization can be initiated by exposure to strong acids, peroxides, perchlorates, or hypochlorites. Contact between vinyltoluene and a base will remove the TBC inhibitor.

Vinyltoluene is used in copolymers and as a specialty monomer in paint, varnish, and polyester formulations.

8.2. Divinylbenzene

Divinylbenzene (DVB) [*1321-74-0*], $C_6H_4(CH{=}CH_2)_2$, is a cross-linking agent that improves polymer properties. The commercial monomer is mainly a mixture of *m*- and *p*-DVB (typical ratio 2.25 : 1), diluted with ethylvinylbenzenes. Physical properties of the 55 % commercial mixture are listed below:

M_r	130.191
bp	195 °C
fp	−45 °C
Pseudocritical pressure	2.462 MPa
Pseudocritical temperature	369 °C
Flammable limits in air	0.8 – 5.5 vol %
Flash point, TCC	65.6 °C
Autoignition point	505 °C
Heat of vaporization, ΔH_v	
(195 °C)	350.6 J/g
Vapor pressure	
(100 °C)	4.4 kPa
(140 °C)	20.0 kPa
(180 °C)	67.0 kPa
Density (25 °C)	0.9084 g/mL
Refractive index, n_D^{25}	1.5585
Surface tension (25 °C)	32.1 mN/m
Viscosity (25 °C)	1.007 mPa s
Solubility	
DVB in water (25 °C)	0.0052 %
Water in DVB (25 °C)	0.054 %

The method of manufacture is analogous to that for styrene: endothermic dehydrogenation of an isomer mix of diethylbenzenes obtained as a side-stream of ethylbenzene production (→ Hydrocarbons, Chap. 3.4.). Diethylbenzene is vaporized, diluted with superheated steam, and then passed over a catalyst at ca. 600 °C. The reactor effluent is mainly *m*- and *p*-DVB, the corresponding ethylvinylbenzenes, and unreacted diethylbenzene. Most of the *o*-diethylbenzene undergoes ring closure to naphthalene. Separation of this mixture must compete with the ready polymerization of DVB. Separation can be achieved by vacuum distillation with suitable in-process polymerization inhibitors.

Divinylbenzene is more reactive than styrene, and can homopolymerize or copolymerize with many other monomers. Because of the two vinyl groups, the resulting polymers are cross-linked. Cross-linking improves solvent resistance, heat distortion, impact resistance, tensile strength, and hardness.

DVB is very similar to styrene monomer in its toxicological properties (Chap. 9). The ACGIH TLV for divinylbenzene is 10 mg/kg (53 mg/m^3). Its single-dose oral toxicity is low. Because DVB polymerizes so readily, additional precautions must be taken to prevent runaway reactions in storage. The monomer is inhibited with TBC (ca. 1000 mg/kg), and refrigerated storage at 10 °C is recommended.

Most DVB is used in ion-exchange resins based on suspension-polymerized beads of styrene – divinylbenzene. It is also used as a cross-linking agent in specialty rubber and other plastic applications (→ Polyester Resins, Unsaturated).

8.3. α-Methylstyrene

α-Methylstyrene (AMS) [*98-83-9*], $C_6H_5C(CH_3){=}CH_2$, is a monomer with a polymerization rate much lower than that of styrene. It can be produced commercially by the dehydrogenation of isopropylbenzene (cumene), but also occurs as a byproduct in the manufacture of phenol and acetone via the cumene oxidation process, now the major source of AMS (→ Acetone, Chap. 4.1.; →Hydrocarbons, Chap. 3.8.). Physical properties of AMS are listed below:

M_r	118.18
bp	165 °C
fp	−23.2 °C

Critical density	0.29 g/mL
Critical pressure	4.36 MPa
Critical temperature	384 °C
Critical volume	3.26 mL/g
Flammable limits in air	0.7 – 3.4 vol %
Flash point, Cleveland	
Open Cup (COC)	57.8 °C
Heat of polymerization	−39.75 kJ/mol
Heat of vaporization, ΔH_v	
(25 °C)	404.6 J/g
(165 °C)	326.4 J/g
Specific heat capacity (liquid)	
(20 °C)	2.047 J g^{-1} K^{-1}
(100 °C)	2.176 J g^{-1} K^{-1}
Specific heat capacity (vapor) (25 °C)	1.2357 J g^{-1} K^{-1},
Solubility	
AMS in water (25 °C)	0.056 %
water in AMS (25 °C)	0.010 %
Density (20 °C)	0.9106 g/mL
Viscosity (20 °C)	0.94 mPa s

AMS undergoes reactions and polymerizations similar to those of styrene and vinyltoluene. However, the addition of an α-methyl group to the side chain changes its chemical reactivity. AMS monomer tends to be more stable than other vinyl aromatics. Homopolymers are best prepared by ionic catalysis, copolymers by radical polymerization.

Toxicologically, AMS is similar to styrene (Chap. 9). The ACGIH TLV is 50 mg/kg, and the STEL 100 mg/kg. AMS has a higher flash point than styrene, and it is less likely to polymerize in storage. It forms low molecular mass polymers slowly after the depletion of oxygen and/or inhibitor. The polymer thus formed is usually a viscous liquid. Uninhibited AMS in storage oxidizes slowly to acetophenone, aldehydes, and peroxides, which may affect the polymerization rate.

Homopolymers of AMS are used as plasticizers in paints, waxes, and adhesives. The monomer is also used to form a copolymer with methyl methacrylate. The copolymer has a high heat-distortion temperature, and has been approved for use in food applications. Its light color makes it useful in modified polyester and alkyd resin formulations.

8.4. Chlorostyrene

The two compounds *ortho-* and *para-* chlorostyrene [*2039-87-4*], $ClC_6H_4CH{=}CH_2$, are reactive monomers produced only in small quantities. The *para* isomer has a boiling point of 192 °C and a flash point of 60 °C. It is combustible. Chlorostyrene is more reactive than styrene, and is easily polymerized. It is typically inhibited with 3,5-di-*tert*-butylcatechol (250 mg/kg) for storage and transportation.

The monomer can be manufactured from either *o-* or *p-* chloroethylbenzene by oxidation to the corresponding alcohol, followed by dehydration. It is possible to produce pure monomeric *o-* or *p-* chlorostyrene as well as a mixture of the two.

Extensive toxicological data for chlorostyrene are not available, but it is assumed to be similar to styrene (Chap. 9), with an ACGIH TWA of 50 mg/kg.

Chlorostyrene is used to provide shorter cure times, better heat-distortion properties, and flame resistance in polyester resins.

8.5. Vinylbenzyl Chloride

Vinylbenzyl chloride (VBC) [*30030-25-2*], $CH_2{=}CHC_6H_4CH_2Cl$, is also called (chloromethyl)ethenylbenzene or α- chloromethylstyrene. It is a difunctional monomer with both a polymerizable double bond and a benzylic chlorine, and is an isomeric mixture made by the chlorination of VT. VBC has a high boiling point (98.9 °C at 1.3 kPa) and low flammability. Other physical properties are as follows:

M_r	152.62
bp (1.33 kPa)	98.9 °C
(101.325 kPa)	229 °C∗
fp	−26 to −42 °C
Flash point, COC	104.4 °C
Autoignition point	610 – 620 °C
Heat of fusion, ΔH_m	−10.04 kJ/mol
Heat of vaporization, ΔH_v (157.2 °C)	334.0 J/g
Heat of polymerization, ΔH_p (195 °C)	−0.422 kJ/mol
Solubility in water (25 °C)	730 mg/kg

∗ Extrapolated value.

The reactivity of VBC is similar to that of VT. VBC is stored under refrigeration, and inhibited with 75 mg/kg of TBC and 800 mg/kg of nitromethane [38]. Its acute oral toxicity is claimed to be low, but the monomer vapor is a strong lachrymator. VBC can be transformed into a high molecular mass homopolymer, will copolymerize with many other monomers, and is used as comonomer in a variety of specialty plastics. It is also possible to carry out reactions

on the chloromethyl group, either before or after polymerization. The chemistry of VBC has been covered in an excellent review [39].

9. Toxicology and Occupational Health

In view of the large-volume production of styrene and the widespread use of styrene-based polymers in general, and especially in food-contact use, there is much research and a huge literature on the toxicology of styrene [40], [41] (→ Polyester Resins, Unsaturated, Chap. 9.).

Styrene is slightly toxic. In rats, the acute oral LD_{50} is 500 – 5000 mg/kg, with an inhalation $LC_{50} > 3000$ mg/kg. Death from acute dermal exposure has not been reported. Styrene is a moderate skin irritant, especially if trapped against the skin by contaminated clothing. Styrene vapor at high concentration has an irritant effect on the eyes and mucous membranes of humans and animals. Controlled observations on humans have shown that no irritation occurs at a concentration of ca. 100 mg/kg [42], [43], but trace amounts of styrene in the workplace together with trace amounts of halogen can produce a lachrymator in the presence of bright sunlight.

Styrene can affect the central nervous system, producing fatigue and headache at a certain level of exposure, and narcosis if the level is sufficiently high. Signs of transient impairment of central nervous system function have been described in voluntary test subjects [44] and in workers [45] exposed to styrene in concentrations of 375 –800 mg/kg. Impaired hearing and visual coordination have been detected after exposure to styrene at ca. 100 mg/kg. At 50 mg/kg, however, any negative effects of this type were considered doubtful or barely discernible [46]. Prenarcotic symptoms have been shown in workers exposed to a TWA of 50 mg/kg. They are clearly established at a TLV of 100 mg/kg [47].

A review of eight cohort studies involving nearly 50 000 employees over 46 years [48] showed no indication of human carcinogenicity related to styrene exposure. A comprehensive review of published data [49] concluded that there is little indication that styrene has any specific developmental or reproductive toxicity.

Long-term animal studies on styrene are more difficult to analyze. Limitations in the studies often preclude definitive conclusions. However, a review by Boyd et al. [50] concludes that there is no clear evidence of a carcinogenic response related to styrene exposure. It continues to be a problem to relate high-dose results in different animals to each other, and to project these to human response.

The metabolism of styrene proceeds through styrene-7,8-oxide to mandelic acid and phenylglyoxalic acid, which are then excreted in the urine [51]. There is a correlation in humans between urinary elimination of mandelic acid and the styrene concentration in the respiratory air if the latter is ≤ 100 mg/kg (420 mg/m^3). However, the styrene concentration in the blood seems to be a more reliable exposure index than either the mandelic acid content of urine or the styrene content in exhaled air.

The mouse seems to be a very sensitive species in styrene exposure studies. Mice apparently have a greater capacity than rats or humans to form styrene-7,8-oxide, and a lesser capacity to metabolize this chemical [52].

Nevertheless, the International Agency for Cancer Research classified styrene in 1987 as a possible human carcinogen, at the same time concluding that the evidence for carcinogenicity was limited, and in humans inadequate. This classification has provoked legislation, causing some countries and several states in the US to list styrene as a possible carcinogen. OSHA, on the other hand, decided in 1989 that styrene should not be regulated as a workplace carcinogen.

For styrene monomer, as for many other reactive substances, the toxic effect clearly depends on dose. Health risks can be expected at very high concentrations, but the no-observed-effect levels (NOEL) are considerably higher than those found in a well-managed operating facility, or in the background to which users of styrene-based products are exposed. The ACGIH TLVs for styrene are 50 mg/kg (213 mg/m^3) for an 8-h time-weighted average (TWA) exposure, with a 100 mg/kg (426 mg/m^3) short-term exposure limit (STEL), defined as a 15-min TWA [53]. The MAK value is 20 mL/m^3 (85 mg/m^3).

10. References

1. E. Simon, *Justus Liebigs Ann. Chem.* **31** (1839) 265.
2. D. H. Steele, MRI Project No. 6450, Midwest Research Institute, Kansas City, Mo., 1992.
3. R. H. Boundy, R. F. Boyer, (eds.): *Styrene, Its Polymers, Copolymers and Derivatives,* Reinhold Publ. Co., New York 1952; reprinted by Hafner Publishing Corp., Darien, Conn., 1970.
4. R. F. Boyer: "Anecdotal History of Styrene and Polystyrene," in R. Semour (ed.): *History of Polymer Science and Technology,* M. Dekker, New York 1982.
5. Dow Chemical, The Dow Family of Styrenic Monomers, Form No. 115-00 608-1289 X-SAI, Midland, Mich. 1989.
6. W. S. Emerson, *Chem. Rev.* **45** (1949) 183 ff.
7. K. E. Coulter, H. Kehde, B. F. Hiscock in E. C. Leonard (ed.): *High Polymers,* Wiley-Interscience, New York, vol. XXIV, "Vinyl and Diene Monomers," part 2: "Styrene and Related Monomers," 1971.
8. K. Meinel, *Justus Liebigs Ann. Chem.* **510** (1934) 129.
9. D. L. Williams, "Styrene Catalysts, Past Present and Future," AICHE Spring Meeting, New Orleans, March 1988.
10. Shell Chemical, Technical Bulletin on Shell 005, Shell 105 Dehydrogenation Catalysts, 78 : 14 Houston, Texas 1978.
11. United Catalysts Inc., G-64 & G-84 Product Bulletin, 2nd ed. Louisville, Ky. 1987.
12. *Ullmann,* 4th ed., **22,** 298.
13. Lurgi, Low Energy Concept for Styrene Production, Technical Brochure 2173 e/4.89.
14. Union Carbide Corp., US 3 282 576, 1966.
15. J. C. Frank, G. R. Geyer, H. Kehde, *Chem. Eng. Prog.* **65** (1969) no. 2, 79.
16. "Facelift for Distillation," *Chem. Eng.* 1987, March 2, 14.
17. D. B. McMullen et al., *Chem. Eng. Prog.* **87** (1991) no. 7, 187.
18. Sulzer Bros. Inc., GB 1 020 190, 1966.
19. UOP Inc., US 4 435 607, 1984 (T. Imai).
20. K. Egawa et al., *Aromatics* **43** (1991) 5 – 6.
21. ARCO Chemical, Construction Permit Application for Propylene Oxide/Styrene Monomer Facility, Channelview, Texas, Texas Air Control Board Account No. HG-1575 W, July 1989.
22. Phillips, US 4 117 025, 1978 (T. C. Liebert, W. A. McClintock).
23. Exxon, US 2 943 117, 1960 (A. H. Gleason).
24. Phillips, US 3 377 397, 1968 (P. L. Maxfiel).
25. Phillips, US 4 144 278, 1979 (D. J. Strope).
26. BASF, US 3 903 185, 1975 (H.-H. Vogel, H.-M. Weitz, E. Lorenz, R. Platz).
27. ARCO, US 4 029 715, 1977 (R. W. Rieve, H. Shalit).
28. Maruzen Oil, US 3 502 736, 1970 (M. Sato, K. Tawara).
29. Montedison, DE 2 612 082, 1976.
30. Monsanto, US 3 965 206, 1976 (H. W. Scheeline, J. J. L. Ma).
31. Monsanto, US 4 115 424, 1978 (M. L. Unland, G. E. Barker).
32. Monsanto, US 4 140 726, 1979 (M. L. Unland, G. E. Barker).
33. Shell Oil, US 5 015 796, 1991 (L. H. Slaugh, T. F. Brownscombe).
34. R. Kuhn, K. Birett: *Merkblätter gefährliche Arbeitsstoffe,* Blatt Nr. S 25, Verlag Moderne Industrie, München 1978.
35. EPA-450/4-91-029, 11 – 12, Office of Air Quality, Research Triangle Park, N. C. T. Wett, *Chemical Business,* March 1993, 21.
36. SRI International, Styrene, Report 33 C, Menlo Park, Ca., 1993.
37. Chem Systems, Styrene/Ethylbenzene 91-9, Tarrytown, New York, 1992.
38. Dow Chemical, Speciality Monomers Product Stewardship Manual. Form No. 505-0007-1290 JB, Midland, Mich. 1990.
39. M. Camps, M. Chatzopoulos, J. P. Montheard, *J. Macromol. Sci., Macromol. Chem. Phys., Part C* **22** (1982 – 83) no. 3, 343.
40. J. Santodonato et al., "Monograph on Human Exposure to Chemicals in the Workplace; Styrene," PB 86-155 132, Syracuse Research Corp., Syracuse, N.Y., July 1985.
41. *The SIRC Review, vol. 1, no. 1 and vol. 1, no. 2.* 1275 K St., N.W., Suite 400, Washington, DC 20 005.
42. R. D. Stewart, H. C. Dodd, E. D. Baretta, A. W. Schaffer, *Arch. Environ. Health* **16** (1968) 656.
43. R. L. Zielhuis et al.: *14th Int. Congr. Occup. Health,* Madrid, 3, 1092.
44. C. P. Carpenter et al., *J. Ind. Hyg. Toxicol.* **26** (1944) no. 3, 68.
45. H. Harkonen, *Int. Arch. Occup. Environ. Health* **40** (1977) 231.
46. M. Oltramare et al.: *Editions Medecine et Hygiene,* Geneva 1974, p. 100.
47. Y. Alarie, *Toxicol. Appl. Pharmacol.* **24** (1973) 279.
48. G. Bond, K. Bodner, R. Cook, *SIRC Review* **1** (1991) no. 1, 43 – 55.

49. N. A. Brown, *Reproductive Toxicol.* **5** (1991) 3.
50. D. P. Boyd et al., *SIRC Review* **1** (1990) 5 – 23.
51. Z. Bardodej, E. Bardodejova, *Amer. Ind. Hyg. Assoc. J.* **31** (1970) 206.
52. D. H. Steele et al., *J. Agric. Food Chem.* **42** (1994) 1661.
53. 1992 – 1993 Threshold Limit Values, American Conference of Governmental Industrial Hygienists, Cincinnati, Ohio 1992.
54. Kirk-Othmer, 4th ed., **vol 22**, John Wiley & Sons (1997), p 981

Styrene-Butadiene Rubber → Rubber, 3.Synthetic
Styrene Copolymers → Polystyrene and Styrene Copolymers
Styrene Polymers → Polystyrene and Styrene Copolymers
Suberic Acid → Dicarboxylic Acids, Aliphatic

Sugar

HUBERT SCHIWECK, Südzucker AG Mannheim/Ochsenfurt, Mannheim, Germany (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22 and 23)

MARGARET CLARKE, Sugar Processing Research Institute, Inc., New Orleans, Louisiana 70124, United States (Chap. 20)

GÜNTER POLLACH, Gross-Enzersdorf, Austria

1. Terms, Distinguishing Criteria, Uses, Nutritional Physiology 2
2. History 3
2.1. Development of the Cane Sugar Industry 3
2.2. Development of the Beet Sugar Industry 4
3. Properties 5
3.1. Physical Properties of Crystalline Sugar 6
3.2. Physical Properties of Aqueous Sugar Solutions 6
3.3. Chemical Properties 7
3.4. Sensory Properties 7
4. Sucrochemistry, Chemical and Enzymatic Synthesis 7
4.1. Sucrochemistry 7
4.2. Chemical and Enzymatic Synthesis 8
4.3. Plant Biosynthesis 8
5. Methods of Analysis 9
6. Sugar Beets: Cultivation, Harvesting, Preparation 10
6.1. Cultivation 10
6.2. Chemical Composition and Beet Quality 11
6.3. Preparation 13
6.4. Delivery and Storage 13
6.5. Thick-Juice Storage 14
6.6. Weighing and Sampling 14
6.7. Internal Transport, Washing 14
6.8. Processing Wash and Flume Water 15
6.9. Slicing Beets 16
7. Production of Juice (Extraction) . 17
7.1. Principles 17
7.2. Extraction Technology 19
7.3. Extraction with Nonaqueous Solvents 20
7.4. Extraction Systems 20
7.4.1. BMA Extraction System 20
7.4.2. Extraction with the DdS Slope Diffuser 21
7.4.3. Extraction with the De Smet System 22
7.4.4. The RT-5 Continuous Extractor . . . 22
7.5. Pressing Extracted Pulp 22
8. Juice Purification 23
8.1. Lime–Carbon Dioxide Treatment 23
8.2. Other Juice Purification Processes 29
8.2.1. Simultaneous Defeco-Carbonation . 29
8.2.2. Braunschweig Juice Purification . . 30
8.2.3. Defeco - Carbonation with Intermediate or Post Main Liming 30
9. Evaporation of Thin Juice 30
10. Production of Sugar from Thick Juice 33
10.1. Production of Massecuite 33
10.2. Evaporating Crystallization 34
10.3. Mixing of Massecuites and Cooling Crystallization 36
10.4. Factors That Influence Rate of Crystallization 36
11. Separation of Sugar from Massecuite 37
12. Preparation of Refined and White Sugar 37
13. Working Schemes of Various Sugar Factories 38
13.1. Operation of a Raw Sugar Factory 41
13.2. Operation of a White Sugar Factory – Standard Liquor Process 41
13.3. White Sugar Factories with a Refinery Scheme 42
13.4. Refining 42
14. Ion - Exchange Processes in the Sugar Industry 42
14.1. Softening of Thin Juice 43
14.2. Exchange of Alkali Ions for Alkaline - Earth Ions 43
14.3. Separation of Molasses into Sugar and Nonsugar Fractions (Ion Exclusion) 44

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_345.pub2

14.4. Decolorization of Remelt Syrups and Thin Juices 44
14.5. Elimination of Ionogenic Nonsugar Substances 44
14.6. Production of Liquid Sugar 45
15. Production of Special Types of Sugar 45
15.1. Cube Sugar 45
15.2. Sugarloafs and Sugar Cones 46
15.3. Rock Candy 46
15.4. Icing Sugar 46
15.5. Instant Sugar 46
15.6. Preserving (Jelly) Sugar 46
15.7. Brown Sugar (Soft Sugar) 46
15.8. Sugar Solutions 47
15.9. Burnt Sugar and Caramel Colors (Caramels) 48
16. Storage of White Sugar 48
17. Auxiliary Facilities in a Sugar Factory 48
17.1. Steam and Energy Balance 48
17.2. Drying, Pelletizing, and Storing Pulp 49
17.3. Lime–Carbon Dioxide 50
18. Desugarization of Molasses 50
19. Sugar Yield, Energy Requirements, Processing Aids, Water, and Working Time 51
20. Cane Sugar 52
20.1. Sugarcane Cultivation 52
20.2. Composition of Sugarcane 52
20.3. Deterioration of Sugarcane 53
20.4. Harvesting and Delivery 53
20.5. Sugarcane Processing: Milling and Diffusion 55
20.5.1. Cleaning 55
20.5.2. Juice Extraction 55
20.5.3. Factory Fuel 56
20.6. Cane Factory Process 56
20.7. Direct White Sugar Factory 57
20.8. Refining 58
21. Sugar from Other Plants 59
22. Quality Demands on Sugar and Side Products of Sugar Production 60
22.1. Refined and White Sugar 60
22.2. Molasses 62
22.3. Vinasse 63
22.4. Liquid Feed Supplement 63
22.5. Beet Pulp 63
23. Economic Aspects 64
24. References 67

1. Terms, Distinguishing Criteria, Uses, Nutritional Physiology

The word sugars is the general term used in carbohydrate chemistry for all nutritive mono- and disaccharides such as glucose, fructose, galactose, maltose, and sucrose. Sugar is the technical and legal name for the disaccharide sucrose [*57-55-6*], (EINECS no. *200–334–9*). Irrespective of whether cane or beet is the source, the term sugar is used both in technology and in food laws. However, commercially a distinction is made between cane and beet sugar.

The chemical and physical properties of crystalline sugar made from cane and beet are largely identical. Slight differences exist only in the type and mass of the individual nonsugar substances that are enclosed in the sugar crystal or adsorbed on its surface (e.g., raffinose, theanderose [16], betaine, amino acids, potassium, sodium, calcium, or flavor- and taste-imparting substances). On the other hand, the $^{13}C/^{12}C$ abundance ratio (δ-^{13}C value) varies; it is ca. 11.3 ‰ for sucrose from cane and ca. 25.2 ‰ for sucrose from beet [17].

The origin of sucrose (i.e., cane or beet) can be determined unambiguously from this difference. It is attributed to the varying CO_2 fixation in the photosynthesis of sugarcane, a C_4 plant, and sugar beet, a C_3 plant [18]. However, a detection of unauthorized addition of cane white sugar to beet white sugar is limited by the variation of δ-^{13}C values.

Apart from the δ-^{13}C value, the δ-D/H and the δ-$^{18}O/^{16}O$ values are also gaining considerable importance as analytical determination tools, e.g., for detecting the unauthorized addition of sugar to fruit juice, honey, or wine (enrichment or chaptalization, after the French chemist F. Chaptal). With the help of these values, it is possible to assign a sugar sample to a definite area of cultivation. For instance, the addition of beet sugar from Michigan to orange juice from Florida, or the enrichment of must (wine) with beet sugar, can be detected unambiguously (Martin process) [19].

In the carbohydrate metabolism of the plant, sucrose plays an important role as the first nonphosphorylated compound. According to some calculations, plants synthesize 150×10^9 t of sucrose in one year as an intermediate product. In some juices from plant parts, the sucrose stored as a reserve substance accounts for up to 75 % of the solids. These plants serve as raw materials for the isolation of sugar. They are mainly sugar beet and sugarcane and, to a much lesser extent, dates [20] (Iraq, Iran, and the Near East), sweet corn [21] (Middle America), sweet sorghum [22] (Europe, North and South America), maple trees [23] (North America and Canada), and palms [24] (India).

Sugar is used predominantly in the production of food and beverages. The amounts added relative to the finished product vary considerably. They range from 70–80 % in hard boiled candies, to 40–50 % in chocolates and jams, ca. 20 % in cakes and pastries (pound cake, etc.), 10 % in carbonated beverages and soft drinks, and ca. 1 % in toast.

As a component of a formulation, sugar offers the following factors:

1) It confers sweetness on products and, after hydrolysis, participates in formation of the product-specific flavor and browning crust color (Maillard reaction, Strecker degradation, and caramelization) (→ Foods, 1. Survey)
2) It is a bulking (bodying) agent, which substantially influences the texture and mouth feeling of the products during consumption
3) Sugar is also important from a nutritional point of view

In some countries, sugar accounts for more than 10 % of the daily intake of calories (see Section 23, Table 18). Some nutritionists consider a sugar consumption of > 10 % of the daily intake of calories as too high. However, evaluation of the literature (e.g., for the U.S. FDA, United Kingdom, and Germany) has shown that except for the formation of caries, no scientific reasons exist for condemning a higher sugar consumption [25].

Sugar is used, for example, as a substrate for the production of ethanol (→ Ethanol, Chap. 5.1.1, → Ethanol, Chap. 5.1.2), citric acid (→ Citric Acid, Chap. 4.3), yeast (→ Yeasts), amino acids (→ Amino Acids), enzymes (→ Enzymes, Chap. 3.1), and antibiotics (→ Antibiotics, Chap. 5.4.2). For these biotechnological processes, sugar is often employed in the form of molasses (see Section 22.2). Sugar can be converted enzymatically to isomaltulose (→ Sugar Alcohols) and leucrose.

In Brazil, bioethanol is produced on a large scale from sugarcane raw juice (ca. 10×10^6 t/a) and used either directly as a carburetor fuel or as a gasoline–ethanol mixture (up to 20 vol % ethanol). The European Union supports the use of biofuels with the directives 2003/96/EC and 2003/30/EC. The target is to produce 5.75 % of the total fuel consumed in the EU from renewable energy sources by 2010 [26]. (→ Ethanol; → Automotive Fuels).

2. History [27]

Since time immemorial, man not only has craved salt but also had a desire for sweet food. Whereas the consumption of salt (electrolyte) is a physiological necessity, the desire for sweet food is associated with an adequate intake of calories. Sweet, naturally occurring products, such as mother's milk, honey, and fresh and dried fruits and their juices, not only are considered good and healthy, but also often have a high energy density.

Before humans were able to produce sugar from plants, honey and concentrated grape and fruit juices (which can be stored safely for a long time) were the sole sweeteners.

Rock drawings found in the Arana caves (eastern Spain, ca. 8000 B.C.) show how honey was collected from wild bees. Similar depictions of the collection of honey and the making of thickened grape must concentrates have been found in drawings in Egyptian tombs dating from 2600–1600 B.C.

2.1. Development of the Cane Sugar Industry

On the Indian subcontinent, the practice of chewing and sucking bits of cane was popular in early days. According to recent research, sugarcane has been cultivated in New Guinea since 8000 B.C. It traveled from there to Indochina,

Table 1. Main beet and cane sugar producing countries (2002–2003) [28]

Beet sugar	Raw sugar, 10^3 t	Cane sugar	Raw sugar, 10^3 t
France	5140		
Germany	4366	Brazil	22 540
United States	4046	India	21 900
Turkey	2345	China	10 200
Poland	2195	Thailand	7300
Russia	1740	Australia	5370
Ukraine	1550	Mexico	5230
United Kingdom	1538	Pakistan	3940
Italy	1532	United States (including Puerto Rico)	3594
People's Republic of China	1350	South Africa	2760
Spain	1288	Cuba	2200
The Netherlands	1112	Philippines	2160
Belgium/Luxembourg	1108	Indonesia	1750
Subtotal	29 310		88 944
Total		**118 254**	

and finally, to the Bay of Bengal, where it was crossed with native wild species in India.

Much later, about 5000 B.C., a sugar syrup was prepared by pressing or boiling cane and thickening the juice. This type of sugar production was practiced for thousands of years. The art of producing solid sugar from cane was developed in India about A.D. 200–400.

Through the campaigns of Alexander the Great in India (327 B.C.), information about sugarcane first reached Europe.

The cultivation of sugarcane spread from India to China (200 B.C.) and Java, as well as to Sicily and Spain via Persia and lower Egypt. The Egyptians had a developed sugarcane industry as early as A.D. 1000 and became the masters of their former teachers, China, Persia, and India. However, the sugarcane cultures in Syria, Egypt, and Cyprus were, for the most part, destroyed by the Turkish campaigns in the 15th century. The Portuguese Prince Henry the Seafarer transplanted sugarcane from Sicily to Madeira in 1420 and laid the foundations for a flourishing industry on the Canary Islands. On his second voyage in 1493, Christopher Columbus brought sugarcane from the Canaries to Hispaniola, which was later called Santo Domingo. Thus, the foundations were laid for the powerful sugarcane industry of the West Indian Antilles, the sugar islands, which supplied the European market for centuries. Sugarcane then traveled via Middle America to South and North America and to almost all tropical and subtropical regions.

Since the cultivation of sugarcane is very labor intensive, a flourishing slave trade had developed between Africa and America by the beginning of the 16th century. This trade brought with it social, political, and economic problems that finally led to the abolition of slavery in the second half of the 19th century.

Sugarcane was introduced into Europe in A.D. 996. As a result of the spread of trade relations to the Arab countries and, above all, the Crusades (1096–1291), sugar quickly became known and appreciated as both a medicine and a luxury article. The European sugar trade was controlled almost exclusively by the Republic of Venice. With the decline of Venice after the Thirty Years' War, the sugar trade was taken over by Holland and, later, by England.

The main cane sugar producing countries are listed in Table 1. In Europe, sugarcane is cultivated only in Spain.

2.2. Development of the Beet Sugar Industry

While studying the isolation of sugar from native plants in Berlin, Andreas Sigismund Marggraf (1709–1782) discovered in 1747 that the sugar present in beet root is chemically identical to cane sugar. He also isolated D-glucose (dextrose, grape sugar) from grape juice (→ Glucose and Glucose-Containing Syrups, Chap. 1). Publication of this important discovery laid the foundation for the European beet sugar industry. Marggraf also indicated how sugar could be isolated from beet and recognized the economic

importance of this branch of industry. However, his discovery remained virtually unused in his lifetime.

Franz Carl Achard (1753–1821) recognized that a beet with a high sugar content is necessary for economic sugar production.

Achard developed selection–breeding methods that are still valid. In addition, he perfected the technology of sugar isolation and, after overcoming many difficulties, was able to present the first sample of beet sugar to the Prussian king Friedrich Wilhelm III on January 11, 1799. In 1801 a sugar factory was built in Kunern, and it was put into operation in April 1802. After completing his campaign, Achard described his experiences in a paper published by Wilhelm Gottlieb Korn, "Instructions for the Cultivation of Beet Used in Sugar Production and for the Profitable Isolation of Sugar from Beet," which appeared in Breslau in 1803. However, the expected economic success failed to materialize. The factory burned down in 1807 and was not rebuilt [29].

Table 2. Solubility of sucrose in pure water under normal pressure, according to [34] *

t, °C	S, wt %	L_t
0	64.45	1.8127
10	65.43	1.8926
20	66.72	2.0047
30	68.29	2.1535
40	70.10	2.3450
50	72.12	2.5863
60	74.26	2.8857
70	76.48	3.2515
80	78.68	3.6901
90	80.77	4.2003
100	82.65	4.7637

* As calculated from the equation
$S = 64.447 + 8.222 \cdot 10^{-2} t + 1.6169 \cdot 10^{-3} t^2 - 1.558 \cdot 10^{-6} t^3 - 4.63 \cdot 10^{-8} t^4$. S = weight percent of sucrose, L_t = g sucrose per g water.

As a result of the continental blockade imposed by Napoleon on England in 1806 and delivery problems for the cane sugar industry resulting from the slave rebellions of 1791, some sugar factories were set up in Germany and France. However, most of them had to be closed down after a short time because they did not operate profitably. The French sugar industry fared better because it was strongly supported by Napoleon. Only the sugar factories founded after 1831 endured. The Zuckerfabrik Waghäusel/Baden of Südzucker AG, founded in 1837, was operated during a period of nearly 160 years, until 1994, and was the oldest beet sugar factory in the world in operation at the end of the 20th century.

The main beet producing countries are listed in Table 1. In the United States, the People's Republic of China, Spain, Pakistan, Japan, and Uruguay, sugarcane and sugar beet are cultivated side by side. As compared with 1990, the fraction of beet sugar declined from 38 % to 25 % of the total production and a further drop is to be expected in the future (see Section 23).

3. Properties

The conformation of crystalline sucrose is shown in Figure 1. The glucose portion has the $^4C_1$ chair form (→ Glucose and Glucose-Containing Syrups), and the fructose portion the $^4T_3$ twist conformation (→ Fructose, Chap. 2.1). The two rings are almost at right angles to each other. Strong intramolecular hydrogen bonds exist between the 6′-OH group of fructofuranose and the ring oxygen of glucopyranose (189 pm) and between the 1′-OH of fructose and the 2-O of the glucose moiety (185 pm). Since the two glycosidic hydroxyl groups are linked in etherlike fashion to each other, α-1 → β-2, sucrose is a nonreducing disaccharide.

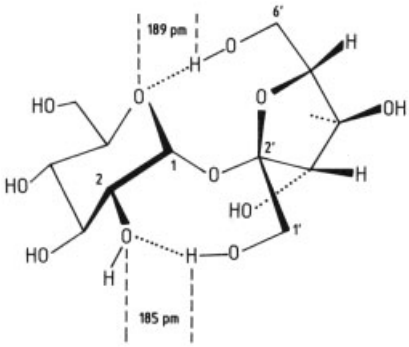

Figure 1. Conformation of sucrose

The specific rotation of sucrose in water is $[\alpha]_D^{20}$ =+ 66.529° [12] (measured at a concentration of 26 g filled to 100 cm^3 with water, but recalculated to c = 1 g/cm^3 and 1 dm of optical path length).

Table 3. Density of aqueous sucrose solutions (in g/cm^3), according to [35]

Sucrose content, wt %	Temperature, °C						
	20	30	40	50	60	70	80
10	1.038114	1.035144	1.031394	1.026965	1.021931	1.016344	1.010244
20	1.080976	1.077580	1.073501	1.068818	1.063586	1.057845	1.051629
30	1.127031	1.123203	1.118790	1.113850	1.108421	1.102532	1.096213
40	1.176518	1.172262	1.167519	1.162326	1.156707	1.150683	1.144281
50	1.229656	1.224991	1.219935	1.214505	1.208713	1.202574	1.196116
60	1.286613	1.281585	1.276253	1.270618	1.264682	1.258460	1.251986
70	1.34748	1.34217	1.33662	1.33084	1.32480	1.31855	1.31211
80	1.41220	1.40673	1.40108	1.39523	1.38917	1.38295	1.37662

Table 4. Viscosity of sucrose solutions (in mPa · s), according to [36]

Sucrose content in vacuum, wt %	Temperature, °C									
	5	10	15	20	30	40	50	60	70	80
40	11.44	9.166	7.463	6.167	4.375	3.241	2.49	1.97	1.60	1.32
50	33.18	25.21	19.53	15.43	10.11	6.991	5.07	3.81	2.94	2.34
60	159.1	110.9	79.49	58.49	33.82	21.04	14.0	9.66	6.98	5.20
65	480.7	313.1	211.3	147.2	77.29	44.36	27.5	17.9	12.4	8.81
70	2 038	1206	746.9	481.6	221.6	114.0	64.4	39.0	25.0	16.8
75*	14 428	7402	4039	2328	884.8	389.5	193	105	61.4	38.3

* Values extrapolated below the temperature range of the measurements.

The IR spectrum of sucrose exhibits characteristically sharp bands at 1010, 990, 940, 920, 870, and 850 cm^{-1} and wide bands at 680 and 580 cm^{-1}.

3.1. Physical Properties of Crystalline Sugar

At 0–95 °C, sucrose forms monoclinic anhydrous crystals that have an axial ratio of 1.2532 : 1 : 0.8895; 103° [30]. The *crystallization enthalpy* is 360 J/mol, and the *density* is 1.5879 g/cm^3 at 15 °C and 1.5871 g/cm^3 at 20 °C [31].

Sugar crystals of different morphology and shapes are formed, depending on the conditions of crystallization and the nonsugars inserted in the crystal lattice. Hence, commercially crystallized sugar occurs in many crystal forms [32].

The *melting point* of pure sucrose is 186 °C; commercially available sugar melts in the range 175–190 °C, depending on purity.

On heating a sugar melt (5–20 min) to 180–210 °C, the disaccharide bond is cleaved and the resulting monosaccharides eliminate water to yield volatile and nonvolatile aroma imparting substances, such as 2-butanone [*78-93-3*], 2-acetofuran, furanone, 5-(hydroxymethyl)furfural [*67-47-0*], levoglucosan [*498-07-7*], levoglucosone [*37112-31-5*], anhydroglucoses [*498-07-7*], difructose anhydrides [*124277-48-1*], and polymers of these monomers. The product is known as burnt sugar.

When heated above 250 °C, the carbon chain is increasingly broken up to yield small gaseous molecules, such as CO_2 and H_2, the residue remaining is sugar charcoal.

3.2. Physical Properties of Aqueous Sugar Solutions

When sucrose is dissolved in water, the three primary hydroxyl groups are hydrated first with the formation of hydrogen bonds, followed by the secondary hydroxyl groups. Up to a mass concentration of about 67 % at 25 °C, sugar molecules are hydrated with water molecules in the form of clusters [33]. For this reason, sugar is readily soluble in water. The pH of sugar solutions is 5.0–8.0. The pK values at 25 °C are pK_1 = 12.7 and pK_2 = 13.1.

At higher concentration, sugar forms viscous syrups that are easily supersaturated. Their rate of crystallization is influenced by the type and quantity of impurities present in the solution (see Fig. 8). The solubility of pure sucrose in water as a function of temperature is given in Table 2

[34]. The density of aqueous sucrose solutions as a function of concentration and temperature is listed in Table 3 [35]. Viscosity is important for the industrial processing of sucrose solutions (see Table 4) [36]. It increases with increasing concentration ($\eta = A \cdot 10 \cdot Bb/100 - b$) and decreases with increasing temperature ($\eta = A \cdot B/T$, where A and B are constants, T is the absolute temperature, and b is the solids content).

The most important physical properties of sugar and sugar solutions are summarized in [37].

3.3. Chemical Properties

As a result of the labile furan ring in the fructose portion of the molecule, sucrose can be cleaved easily into its components by hydrogen activation or enzymes. This reaction proceeds with a gain in energy (− 27.9 kJ/mol). The dextrorotatory labile β-fructofuranose released at first is rearranged quickly to give the levorotatory tautomeric mixture (→ Fructose). This is a first-order reaction. Thus, the resulting mixture of equal parts of glucose and fructose becomes levorotatory $[\alpha]_{\mathrm{D}}^{20} = -22.2°$ and is called invert sugar. The cleavage reaction is known as the inversion of sucrose.

Sucrose can be cleaved from both the fructosidic (β-fructosidases, invertase) and the glucosidic (α-glucosidases, sucrases) sides. Microorganisms are capable to produce both enzymes. Cleavage by β-fructosidases predominates in plants and cleavage by α-glucosidases in animals. The latter also applies to the enzyme system of the human mucosa. Because sucrose, being a disaccharide, cannot pass through the intestinal wall, its enzymatic cleavage is a requirement for the absorption and metabolism of sucrose.

In *food production*, the processing parameters (pH, temperature, residence time) are optimized so that the amount of invert sugar produced is just enough for the flavor and browning (color).

With acidic food and drink, attention must be paid to the fact that only traces of hydroxymethylfurfural (HMF) are formed from the fructose part of the molecule (→ Fructose, Chap. 2.3).

In the alkaline degradation of sugar during juice purification, the invert sugar present in sugar beet is degraded to organic acids (D- and L-lactic acid, acetic acid, formic acid, and saccharinic acid). Alternatively, juice purification can be carried out in such a way that the invert sugar is not broken down (cane sugar juice purification).

3.4. Sensory Properties

The sweetness of aqueous sucrose solutions is set as the standard (= 1) and is used by way of comparison to determine the relative sweetness of other substances. The intensity of sweetness of food and beverages depends on sugar concentration, temperature, and the matrix of the food (viscosity in the case of solutions and texture in the case of pasty, solid foods).

The taste profile of sugar is purely sweet and free from all secondary tastes. As with the sweetness, the course of the intensity–time curve of sucrose solution, an important characteristic for the sensation of sweetness, is used as the standard for the determination of curves of this type for other sweetening agents [38].

4. Sucrochemistry, Chemical and Enzymatic Synthesis

4.1. Sucrochemistry

Since the middle of the last century, the world's sugar production has increased greatly (Chap. 23, Table 17) and is consistently higher than the requirements of traditional markets. Furthermore, the production of sugar could be increased still further; sugar is produced in a highly pure state and is one of the cheapest bulk organic chemicals. For all of these reasons, attempts have been made since 1948 to use sugar as a raw material for chemical syntheses.

With the oil crisis in the mid-1970s, these efforts were given a strong impetus by the general realization of the limited availability of fossil fuels. Sugar presented itself as an easily renewable energy resource (see also → Fructose, Chap. 2.3). This area of sucrose chemistry was given the name "sucrochemistry." The numerous microbiological transformation products of

sucrose are not included in the field of sucrochemistry.

Processes for the use of sucrose [39] or other compounds, which are obtainable from the skeletal substances of beet and cane, as chemical raw materials can be divided into three groups:

1) Direct chemical conversion of the unchanged sucrose molecule
2) Chemical conversion of compounds made from sugar (e.g., glucose, fructose, isomaltulose, leucrose, isomalt, hydroxymethylfurfural)
3) Use of side products of sugar processing for the production of bulk chemicals, e.g., xylose or furfural (→ Furfural and Derivatives, Chap. 3) from bagasse, L-araban, L-arabinose, and pectin from beet cossettes

Extracted beet cossettes are used as substrates in solid fermentation to produce protein-enriched cossettes [protein content 19–24 % of dry substance (DS)] [40], and enzymes, such as pectinases and other hydrolases [41].

Apart from esters with fatty acids (e.g., for the production of biologically degradable surfactants and ethers), sugars or products belonging to the second group are oxidized to carboxylic acids, reduced to alcohols, or subjected to destructive hydrogenation and reductive amination.

In summary, sucrochemistry has not gained much success until now although it includes interesting compounds, e.g., 4,1′,6′-trichloro-4,1′,6′-trideoxygalactosucrose [*56038-13-2*], which is suitable as an intense sweetener.

4.2. Chemical and Enzymatic Synthesis

Despite its great economic importance, attempts to synthesize sucrose were made relatively late. In 1928, the tetraacetates of glucopyranose and fructofuranose with free glycosidic groups were chemically linked by the addition of dehydrating agents [42]. However, the resulting octaacetates did not have the proper steric conformation of sucrose (α-1 → β-2) but were the acetates of the isomeric isosucrose [43].

The first successful chemical synthesis of sucrose involved the condensation of 1,2-anhydro-3,4,6-tri-*O*-acetyl-D-glucopyranose (Brigl's anhydride) with syrupy 1,3,4,6-tetra-*O*-acetyl-D-fructofuranose, followed by deacylation. The synthesis could be improved by using crystallized 1,3,4,6-tetra-*O*-benzyl-D-fructofuranose (instead of the syrupy tetra-*O*-acetyl compound), but here too the yield of sucrose was only ca. 35 % [44].

A successful biochemical synthesis [45] used a bacterial phosphorylase (EC 2.4.1.7), which catalyzed the following reaction:

α-D-glucose 1-phosphate + D-fructose $\rightleftarrows$ sucrose + phosphate

In this way, sucrose could be prepared in the crystalline state [46]. However, the reaction equilibrium favors cleavage, reflecting the physiological significance of sucrose phosphorylase (degradation reaction).

4.3. Plant Biosynthesis

In plants, the biosynthesis of sucrose occurs in the cytosol (especially of the leaves) from the "active" forms of glucose and fructose: uridine diphosphate (UDP) glucose and fructose 6-phosphate. These two compounds are, in turn, formed from the triose phosphates produced directly in the Calvin cycle [47]:

Triose phosphate → fructose 1,6-diphosphate → glucose 6-phosphate → glucose 1-phosphate

The actual synthesis of sucrose proceeds via three enzymatic steps:

Glucose 1-phosphate + UTP
$\rightleftarrows$ UDP-glucose + PP (1)

UDP-glucose + fructose 6-phosphate
$\rightleftarrows$ sucrose 6^F-phosphate + UDP (2)

Sucrose 6^F-phosphate + H_2O
$\rightleftarrows$ sucrose + P (3)

Enzymes:
Reaction (1): UDP-glucose pyrophosphorylase (EC 2.7.7.9)
Reaction (2): Sucrose-phosphate synthase (EC 2.4.1.14)
Reaction (3): Sucrose phosphatase (EC 3.1.3.2.24)

With the cleavage of the pyrophosphate (PP) formed in reaction 1, three energy-rich phosphate bonds are hydrolyzed in the synthesis of sucrose. Hence, the equilibrium favors formation of sucrose.

At the time of storage, sucrose is transported in an energy-dependent process from the epigeal parts of the plant (source) through the phloem and into the roots [48]. The storage parenchymal cells (sink), in turn, concentrate sucrose within the tonoplasts in an energy-dependent process. Thus, the source–sink relationship and the osmotic conditions are retained [18, 49].

The assimilation efficiency in sugarcane was studied with labeled $^{14}CO_2$. C. E. Hartt [50] found that 65 % of the carbon dioxide absorbed is used for the synthesis of sucrose and that a single leaf of about 5.7 dm^2 produces 100 mg of sucrose in 1 h. Of these 100 mg, 74 mg is transported to the stalk for storage. A large plant possesses about 22 leaves with a total area of 122 dm^2, which corresponds to a sucrose production of 2.140 g/h.

5. Methods of Analysis

In the analytical methods applied in the sugar factory, a distinction is made between methods for

1) *Final product control,* which are largely stipulated in international agreements (WHO–FAO Codex Commission, pharmacopoeias, etc.) and in national food and feed regulations.
2) *In-process control,* which must be simple but easily reproducible because of the large number of analyses to be performed. For this purpose, optical methods, polarimetry, and refractometry have gained acceptance.

At first, quartz wedge compensation instruments were used exclusively as polarimeters. They were calibrated to give a rotation of 100 (100° Ventzke), corresponding to 34.657 radians, for a solution of 26.00 g of pure sucrose, weighed in air, made up to 100 cm^3 with water, and measured with sodium light (D lines) at 20 °C in a 2-dm tube. The instruments were checked with calibrated quartz plates because quartz and sucrose have about the same rotatory dispersion at wavelengths close to the D lines. With 26 g of substance per 100 cm^3, the percentage of sucrose in the sugar solution is obtained directly from the angle of rotation in degrees Ventzke. Later 100° S (International Sugar Degrees) was used, corresponding to 34.616 radians for the yellow sodium lines of 589.44 nm [51]. Lead acetate solutions were used to clarify the solution. For environmental reasons, clarification is achieved today with aluminum–zinc salts with sodium or calcium hydroxide solutions. As a further alternative to lead acetate clarification, polarimeters are being used that work in the near-infrared range, to permit a measurement of unclarified, dark solutions directly after dilution.

More recent measurements with highly purified sucrose showed certain deviations from the old scale. Thus, the hundred-degree points for various wavelengths were redetermined so that further development of instrumentation would not be hindered [52]. The hundred-degree point for the mercury isotope line ^{198}Hg is:

$$\alpha \frac{20.00^\circ\mathrm{C}}{546.2271\,\mathrm{nm}} = 40.777^\circ \pm 0.001^\circ$$

for the yellow sodium lines:

$$\alpha \frac{20.00^\circ\mathrm{C}}{589.4400\,\mathrm{nm}} = 34.626^\circ \pm 0.001^\circ$$

for the helium–neon laser:

$$\alpha \frac{20.00^\circ\mathrm{C}}{632.9914\,\mathrm{nm}} = 29.751^\circ \pm 0.001^\circ$$

for near-infrared wavelengths:

$$\alpha^{20.00^\circ\mathrm{C}}_{882.60\,\mathrm{nm}} = 14.836^\circ \pm 0.001^\circ$$

$$\alpha^{20.00^\circ\mathrm{C}}_{880.00\,\mathrm{nm}} = 14.927^\circ \pm 0.001^\circ$$

and is defined as 100°Z in the sugar polarimeter.

Another important method is determination of the solids content of juices, syrups, and crystal magmas (massecuite). In the beet sugar industry this is determined almost exclusively from the refractive index of the solutions. The refractive indices of pure sucrose solutions have been determined with great accuracy and are internationally defined [53].

Errors made in using the refractive index for determination of the solids content of impure sugar solutions increase as the nonsugar portion increases. For molasses, these errors are of the order of 1–3 % absolute, compared to values obtained by determining the water content of molasses with Karl Fischer reagent solutions (→ Gas Production, Chap. 8.2.2.3). Nevertheless, the accuracy of the refractometric method is sufficient for technical purposes.

The degree of purity p (or simply purity) is an important value in controlling the success of technological measures. Determined from the polarization, p gives the sucrose content as a percentage of the total solids content. In the raw juice obtained by extracting sugar beet, sucrose accounts for about 90 % of the solids content (i.e., the purity of raw juice is 90 %); the purity of thick juices is approximately 92 % and that of molasses ca. 60 %. This type of purity determination gives approximately correct results only in the case of intermediates in sugar beet processing. It cannot be applied to cane products because the polarimetric determination of sucrose gives incorrect results owing to the high content of invert sugar in these products. In these cases, modern analytical methods for saccharides (GC, HPLC, ion chromatography) must be used to determine the actual sucrose concentration in the products.

In the making of sugar, other important methods of analysis involve the determination of color, inorganic constituents (ash), and individual nonsucrose organics, e.g., glucose and fructose (invert sugar), raffinose, D- and L-lactic acid, citric acid, amino acids, betaine. Lately, enzymatic and chromatographic methods [GC, high-performance ion-exchange chromatography (HPIC), HPLC, high-performance anion-exchange chromatography (HPAEC)] for the determination of organic nonsugar substances have gained acceptance.

The International Commission for Uniform Methods of Sugar Analysis (ICUMSA http://www.icumsa.org), founded in 1897, meets once every four years (since 2002 biennially) to solve analytical problems on an international level. The results of these meetings are published in the Commission's proceedings (last issues: 1990, 1994, 1998, 2002, 2004). Methods for the analytical control of final and intermediate products that are accepted by ICUMSA were summarized in monographs in the past [54] and are published in the loose-leaf ICUMSA Methods Book today [55].

In addition to laboratory methods, on-line measurements of, e.g., dry substance (measured by refraction), white sugar color, pulp moisture, juice alkalinity, and juice hardness are gaining in importance.

6. Sugar Beets: Cultivation, Harvesting, Preparation

6.1. Cultivation [56]

The sugar beet, *Beta vulgaris altissima,* is a biennial plant that is, however, harvested after the first vegetation period for the production of sugar. At that time, the largest amount of sugar is stored in the fleshy root. Thus, the sugar beet is processed before it reaches physiological maturity.

The sugar beet grown today was bred by Achard from the "white Silesian beet" by means of selection. The wild beet *Beta maritima* L., still found in the coastal areas of the Mediterranean and the Atlantic, is regarded as the original form of sugar beet. Like other plants belonging to this genus (fodder beet, beetroot, and chard), the sugar beet has 2×9 chromosomes (i.e., it is diploid).

Classical breeding was based on individual selection, in which the quality of the mother plant was judged on the basis of the properties of the progeny (surface appearance, density of roots, sugar content of cell sap, etc.). The development of a new variety took 16–20 years. New breeds were largely impaired because of self-sterility and the resulting difficulty in inbreeding. A considerable step forward was achieved around 1950 with the introduction of hybrid breeding by the use of sterile male plants. Thus, inbred lines could be developed and crossed. Introduction of the haploid technique has made breeding easier. In this technique, plants are regenerated from unfertilized germ cells. These plants are 100 % homozygous after the doubling of their set of chromosomes. Consequently, in the breeding of sugar beet, improvements in quality can now be achieved much faster. With the help of genetic engineering [57], efforts have been made to make sugar beets resistant to viral diseases (e.g. Rhizomania A, B, and P types, caused by the virus BNYVV and transmitted by the fungus *Polymyxa betae*) and to herbicides, as well as to changes in the concentration of sugar or other constituents (raffinose, betaine, etc.). From the 1970s to 2000 breeding of Rhizomania resistent varieties with normal sugar yield was a very important goal and the success was a milestone in sugar beet breeding. Concepts to

develop Rhizoctonia solani resistance in a similar way are just at the beginning, and the first resistant variety was registered in 2001 [58].

Genetically monogerminal seeds are usually sown in pellet form. Fungicides and insecticides are contained in the cover mass so that further plant protection measures, with the exception of herbicide treatment, are not required during growth.

Efforts are made to achieve a plant population of 80 000–120 000/ha, depending on climatic conditions (row width is 38–50 cm; space between plants in a row is 18–22 cm in the case of cultivation without thinning and 12 cm in the case of mechanical thinning to give a desired number of plants per hectare after germination of the seeds). With the introduction of genetically monogerminal seeds, the labor expended (working hours) in sowing sugar beet was reduced from about 45 h (cultivation with thinning) to 3–11 h for cultivation without thinning, depending on the size of the field (0.5–5 ha). The total amount of labor (including preparation of the field, harvesting, and loading of the beets) decreased to the same extent. In fact, the working hours per hectare were reduced from 260 h in 1967 to 15–50 h, depending on the size of the fields and the harvesting technique used (e.g., a six-row self-propelled combine harvester). The harvesting process, subsequent mechanical removal of dirt tare, and the loading of beets must be conducted gently so that mechanical damage to the beets is as low as possible (< 600 cm^2 per 100 beets). With mechanically damaged beets, high sugar losses occur during storage, washing, and fluming. For undamaged beets sugar losses at this stage are typically 0.3–0.5 %, but for damaged beets ca. 1 %.

In Central Europe and the United States, the vegetation period is about 170–200 days (April 1 to October 15). The time required between sowing and germination is 22 days. In the case of a winter rainfall of 240 mm, the amount of precipitation during the vegetation period should be at least 360 mm. The total heat (sum of the average daytime temperatures during the vegetation period) required to ripen sugar beet is 2500–2900, with the longest possible duration of sunshine in the months of July to September.

The nutrient withdrawal per hectare for 10 t of beets including the leaves is 45 kg of nitrogen, 6 kg of phosphorus, 52 kg of potassium, 15 kg of sodium, 16 kg of calcium, and 8 kg of magnesium. Depending on the fertilization state of the soil (→ Fertilizers, Chap. 2.3.2), fertilization with 120–160 kg of nitrogen, 80–130 kg of P_2O_5, and 150–250 kg of K_2O per hectare compensates for deficiencies. In organic fertilization with farmyard manure, liquid manure, or straw, the mineral components present should also be taken into account. For fertilization, the ratio of N : P : K must be about 1 : 0.4 : 1, based on the pure nutrients, to make sure that the plant has a balanced supply of nutrients. Since the late 1980s, the beet yield (return) per hectare in Western Europe has been an average of 50.8 t, with large regional variations (e.g., Finland 29.2 t/ha, Switzerland 60.3 t/ha). The yield in the United States is somewhat lower, and that in the Commonwealth of Independent States and other Eastern European countries is only two-thirds of this amount.

Since the sugar concentration in beet does not vary greatly (18.4 % in Austria, 13.9 % in Greece), the mass of white sugar produced per hectare correlates with the beet yield per hectare. This is ca. 7.47 t/ha in Western Europe (Finland 4.20, Switzerland 9.06), 6.5 t in the United States, and 4 t/ha in Eastern European countries.

Together with the tops (yield of ca. 20–40 t/ha), which constitute a valuable feed after silage, the utilizable energy obtained per hectare of a sugar beet field is 219 GJ (100 %), compared with 89 GJ/ha for the cultivation of potatoes (40 %), 69 GJ/ha for wheat (31 %), and 114 GJ/ha for the cultivation of grain corn (52 %).

In most areas under cultivation, sugar beet is cut back when harvested. In Europe, for example, the leaves and the beet crown (ca. 15 % of the mass of the defoliated root) are separated from the root body. In the United States, the crown part that is cut off is much smaller (≈ 6 %). Despite the use of considerable mechanization in sowing, care, and harvesting, and of herbicides, fungicides, and insecticides, the cultivation of sugar beet is still one of the most labor intensive areas in agriculture.

6.2. Chemical Composition and Beet Quality

Even beets that are harvested from the same field do not represent homogeneous material. There-

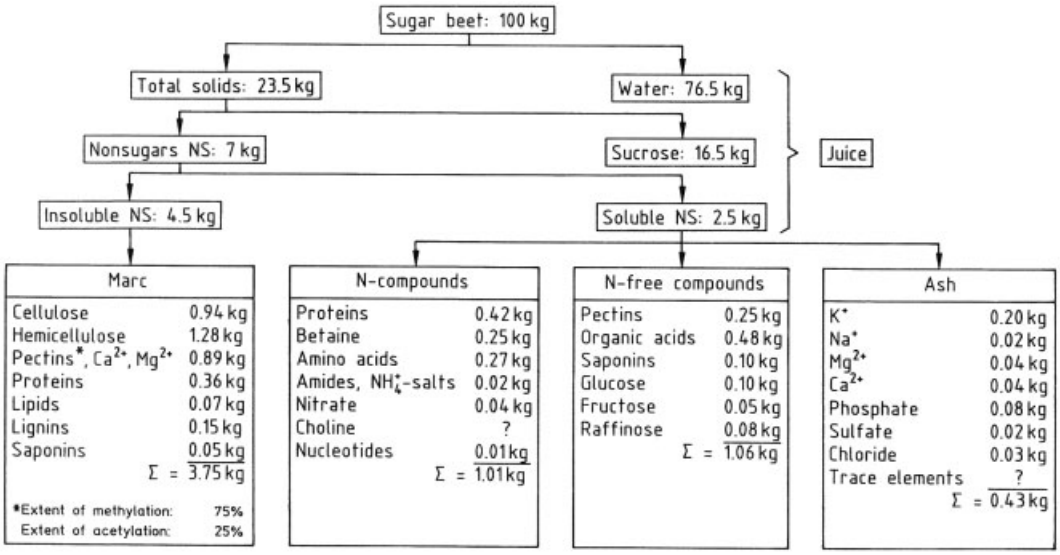

Figure 2. Average chemical composition of sugar beet

fore, data on the chemical composition of sugar beets (Fig. 2) should always be regarded as average values for normally grown beets. Also, certain criteria must be maintained as far as sampling and the size of random samples are concerned [59].

The root body consists of about 75–78 % water. Of its solids (22–25 %), sugar (sucrose) accounts for 15–18 %, raw protein 1–1.2 % (total N×6.25), raw fat 0.1 %, raw fiber 1.1–1.3 %, pectinous substances 1.0–1.2 %, nitrogen-free extractive substances (apart from sucrose) 2.8–3.0 %, and inorganic components 0.7–0.9 %. Similar figures were given in a monograph in 1998 [1]. In sugar technology, sugar beets are subdivided more usefully into "marc" and "cell juice." The marc, which accounts for ca. 4–5.5 % of the sugar beet mass, refers to the components that are insoluble in water under certain extraction conditions (i.e., beet skeletal substances consisting of cellulose, lignin, pectin, araban, and galactan). The cell juice contains, in addition to sucrose, other soluble nonsucrose substances, or nonsugar substances, including invert sugar (0.1–0.2 %), raffinose (→ Carbohydrates, Chap. 2) (0.05–0.2 %), and kestoses (trisaccharides consisting of one molecule of D-fructose and one molecule of sucrose; i.e., fructosylsucroses); a soluble pectin–protein–hemicellulose complex; organic acids (oxalic, citric, tartaric, malic, etc.); proteins, polypeptides, glutamine, glutamic acid, asparagine, and other amino acids; betaine and other plant bases; nucleotides; saponins; enzymes; vitamins; the inorganic components potassium, sodium, calcium, and magnesium as cations; and phosphate, chloride, sulfate, and nitrate as the main inorganic anions.

Of the soluble nonsugar substances in sugar beet, 30–40 wt % is eliminated during purification of the juice (with Ca^{2+}-precipitable anions, pectins, proteins, Mg^{2+}); the remainder is left in the juice and prevents complete crystallization of the sugar, leaving a final syrup, molasses. In molasses, for 1 part of nonsugar substances, about 1.5 parts of sugar are present in the dissolved state. The loss of sugar in molasses is the largest single loss (i.e., 12–18 % of the sugar content of the beet).

For this reason, efforts are being made to predetermine the expected loss of sugar in molasses and/or the purity of the thick juice, which closely correlate with each other, from beet constituents that can easily be determined analytically (content of sugar, K^+, Na^+, α-amino-N, and invert sugar). The average values of these constituents, also called quality parameters, vary greatly from year to year, even in a limited area under cultivation (e.g., southern Germany). Thus, the content of potassium is 45–60 mmol/kg of beets, of

sodium 5–10 mmol/kg, and of α-amino-N 15–32 mmol/kg. Scientific papers of different authors concerning these quality parameters were summarized and compared in 1996 [60].

In controlled cultivation experiments, the ability to influence some quality characteristics was determined analytically [61]. The main influencing factors were found to be location and weather (Table 5). The statement that the influence of the variety is, at ca. 8 %, relatively low applies only when approved, time-tested varieties are compared with each other. The differentiation between varieties that are "good" and those that are "unsuitable" for a certain area under cultivation is considerably greater.

Apart from these chemical characteristics, the mechanical properties of the root, such as breaking and bending strength, elasticity, and cutting resistance, are of prime technological importance. Consequently, they are also significant for the economic success of beet sugar production.

6.3. Preparation

After harvesting, sugar beets are stable for only a limited time and should be processed as quickly as possible. For this reason, beet sugar factories operate for only part of the year (campaign), e.g., about 80–120 days (from the end of September to the end of January) in Central or Northern Europe and the United States. In more southern countries (e.g., Greece and Italy), work starts in July/August. About 2000–25 000 t of beets are processed per day, depending on the capacity of the factory. For economic reasons, most factories produce sugar for consumption immediately (white sugar factories). Only a few factories continue to produce raw sugar first, which is then processed into white sugar in refineries or other white sugar factories.

6.4. Delivery and Storage

For practical reasons, even larger fields must be harvested at one time. The harvested beets are either stored temporarily in clamps at the side of the field or transported directly to the factory. Before the transport vehicles are loaded, the beets are mechanically precleaned to reduce the dirt tare from $\approx$15 % at harvest to $<$ 10 %.

Mechanized harvesting has the disadvantage of delivering the beets along with leaf remnants, stones, and a lot of soil, especially in wet weather. This can lead to tremendous problems during processing (fluming, washing, cutting) in the factory. In the annual processing of 500 000 t of beets, an average of 60 000 t of soil accumulates (40 000 m^3).

If the separated soil cannot be reused – after being stored temporarily to restructure its mechanical properties – the profitability of sugar production is questionable because of the high costs of its disposal in industrialized countries.

Depending on long-term weather conditions, harvesting must be completed before the rainy season (e.g., in Italy) or frost (e.g., in Finland or Montana, United States) and long before the end of the working period. Therefore, a part of the harvest must be kept temporarily in field clamps or large piles on natural or paved ground.

To keep the loss of sugar during storage as low as possible, only ripe, healthy, low-damaged, and clean beets should be stored for longer periods. Sugar losses occur through respiration, since the sugar beet requires a certain amount of energy to maintain its vital functions, and through aerobic degradation or rot caused by yeasts, molds, and bacteria. Excessive mechanical damaging and bruising of beet causes higher sugar losses during storage, due to increased microbiological attack at the damaged surface and additional sucrose utilization of damaged beet to heal the injury. The optimal storage conditions for sugar beet are obtained when beets are piled in a uniformly porous manner: that is at no point in the pile should the hollow spaces between beets be able to get clogged with dirt, weeds, or broken pieces of beet. The porosity of the sugar beet pile should permit the heat generated by respiration, ca. 3000 kJ/t of beets per day, to be dissipated by natural ventilation easily and the temperature of the entire pile to be kept at + 2 to +4 °C. If necessary this low temperature can be maintained by forced air ventilation at night when the air is cooler. In addition, a porous sugar beet pile permits the relative humidity to be kept constant (ca. 85–95 %) so that the beets do not wither excessively (water loss) [62].

Until now, only unwashed sugar beets were believed suitable for storage. However, the opti-

Table 5. Influence of different factors on yield and composition of sugar beets, in percent

	Root yield	Sugar content	Potassium content	Sodium content	α-Amino-N content
Location	22.4	12.4	21.5	37.2	30.4
Weather	35.9	54.2	37.1	42.3	35.3
Time of sowing	25.5	3.3	4.4	3.7	1.2
N fertilization	0.8	3.0	0.7	2.4	20.9
Plant population	5.8	2.3	7.6	4.7	5.2
Time of harvesting	4.2	12.4	10.0	5.7	3.8
Variety	5.4	12.4	18.7	4.0	3.2

mal conditions for storage described above are achieved more easily if sugar beets are washed and then piled to a height up to 10 m. The moisture remaining on the beet surface can be removed by forced air ventilation, which in addition has the advantage of faster temperature reduction in the pile as a result of water evaporation [63].

Despite all preventive measures, sugar losses during storage cannot be avoided completely. Depending on the temperature in the sugar beet pile (+ 2 to + 10 °C), the losses are 150–200 g of sucrose per day per tonne of beets (0.015–0.020 % per day, beet based).

Frozen beets can be stored for long periods with only minor sugar losses [64]. However, they must be processed quickly after thawing because they rot easily. In the northern regions of the United States, to prevent thawing, beets are stored in free span insulated buildings where they freeze. The frozen beets are protected against sudden warm spells.

6.5. Thick-Juice Storage

The recommendations made until now for achieving a year round working period in sugar factories by preserving the beets (freezing, drying, etc.) have not gained acceptance because of their high costs.

Since the mid-1960s, the working period in some factories (e.g., in the United States, France, and Germany) has been extended by making the beet processing factory (parts of the plant preceding the sugarhouse), including the evaporating station, larger than the sugarhouse and by temporarily storing the thick juice in floating-roof tanks [65] with a capacity of up to 100 000 m^3. In tanks without floating roof, which are available at lower investment costs, the juice surface must be protected against yeasts by paraffin, formalin, or caustic soda solution [66]. With properly adjusted storage conditions in the bulk (DS content $\approx 67\,\%$, pH ≥ 9.0, temperature $< 15\,°C$, lowest possible microorganisms content), the thick juice can be stored for years without appreciable quality losses [67], but sometimes thick juice degradation was observed [68]. The profitability of this processing mode depends greatly on local conditions (costs for energy, labor and investment) [69].

6.6. Weighing and Sampling

After the first weighing (gross − tare = net mass), one representative sample is taken, usually mechanically and at random, from every delivery. This sample is used for the determination of dirt tare and partly for top tare [70], sugar content, and for the evaluation of quality. In the case of dry unloading with a wagon or truck tipper, beets are conveyed via a dirt separator (roller grate, wire rod screen) to stacks arranged on paved areas. In wet unloading according to the "Elfa" process, beets are washed out of the vehicle or wagon and pass through flumes to the washing stage.

6.7. Internal Transport, Washing

The beets are transported from the Elfa unloading areas or beet stacks to the actual washing stage in "flumes," concrete or tiled channels laid in the ground about 0.5–1.2 m wide with a slope of 1–2 %. The beets are transported by means of water and simultaneously freed from a large part of the beet tare. Flume channels made of steel sheet are also installed above ground. Trash separators in the channels intercept leaves, straw, and weeds (a type of mechanically operated rotating rake with free-swinging long prongs is suspended from the top of the channel), and a

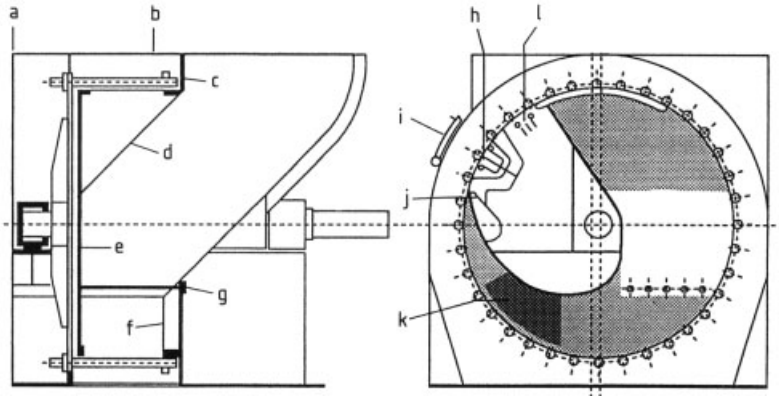

Figure 3. Cross section of a Maguin drum slicer [75]
a) Slicing disk casing; b) Outer casing; c) Beet feed hopper; d) Hinge plate; e) Fixed disk; f) Guide plate; g) Locking bar; h) Access for knife-box changing; i) Flap; j) Inspection hole; k) Blow-out connector (patented)

rock catcher removes stones (the beet flow is led over an ascending water stream whose flow rate is such that the beets remain afloat and the stones sink). The beets are lifted from the flume channel to the level of the washing facility by wheels, screws, or inclined scrapers. Alternatively, the entire beet–water mixture is pumped up by mammoth pumps or one- channel centrifugal pumps (unfavorable because heavy damage is caused to the beets). Before the beets enter the washing facility, flume water is separated from the beet–water mixture with the help of square-roll screens or vibrating screens because part of the beet tare has already been separated from the beets and is contained in the flume water.

The dirt still adhering to the beets is subsequently removed in the washing facility. The beets are washed in trough washers, drums, washers with rotating agitating arms, or on vibrating screens. *Agitating washers* consist of long troughs that are divided into individual compartments and provided with cylindrical screens and a horizontal shaft. This shaft is equipped with stirring arms, transport arms, and throwers. At the bottom, the washers have stone catching compartments and hatches to allow periodic withdrawal of wastewater. The beets are rubbed together, and thus washed, by the movement of the stirring arms.

In *vibrating washers*, the beets are transported over a wedge-wire screen, which is made to vibrate by a circular or linear vibrator. At the same time, water at a pressure of up to 1 MPa is sprayed on the rotating beets from the top. Since the root canals of the beets are also touched by the water jet, the beets obtained when vibrating washers are used for the second wash are generally cleaner than those obtained after only one washing process.

The water requirement is as follows: for fluming, ca. 500–800 %, beet based; for washing, 150–200 %, beet based; and for the operation of a single stone catcher, 70–100 %, beet based. The sugar losses during fluming and washing are 0.05–0.5 %, beet based, depending on the extent of damage to the beets. Losses of < 0.2 % were given as standard for technological leadership [71].

6.8. Processing Wash and Flume Water

The flume and wash water contain beet particles. For this reason, all of the wastewater must first pass through separators. These are mechanical screening devices with round or slit perforations (0.6–4 mm) that remove beet and leaf fragments and root hairs from the water. The intercepted material is separated into its components in cyclone washers. Larger beet particles (> 2 cm) and tails are returned to the washed beets. Smaller particles are either led to the pulp dryer or used directly as feed.

The drained flume and wash water is led to wastewater treatment plants for mechanical purification (sedimentation). The thickened soil sludge, which contains most of the beet soil, is pumped into settling tanks equipped with drainage systems for dewatering and restructur-

ing soil. After two to three years, the soil is compact enough to be reused.

Attempts to remove the beet soil from the thickened sludge directly by means of sedimentation centrifuges did not give the expected results [72].

The mechanically clarified water is reused for fluming and washing (flume and wash water cycle). In this way, only 25–30 % (beet based) of industrial water (condensate) need be added to the water cycle as make-up during the last rinsing of the beets after washing. This quantity corresponds to the amount of water pumped off with the thickened soil sludge. To prevent excessive bacterial growth in the circulating flume and wash water, the addition of $Ca(OH)_2$ to the water to give a pH of 10–12 is appropriate. At the same time, the sedimentation of solid matter in the mechanical settling tank is improved.

Since the transport water pumped off with the soil sludge has a high organic contamination (COD 5000–10 000 mg/L), it is first subjected to anaerobic degradation with an efficiency of 90–95 %. The wastewater is subsequently subjected to aerobic treatment, and nitrification or denitrification, depending on the demands made on the composition of the wastewater to be discharged (e.g., total N content) [73].

Ammonium ions enter the wastewater through the vapor condensates of the evaporator station, which have a NH_4–N content of 100–200 mg/L, depending on the glutamine content of the beets. For this reason, ammonia can be eliminated from the condensates by air stripping [74].

Since the beets (water content $\approx$ 75 %) carry sufficient water into the production process, which accumulates as condensate, a factory today can largely do without ground or surface water. The total COD burden on the wastewater of a beet sugar factory is 2–3 kg per tonne of beets, with $<$ 1 kg attributed to leadership technology [71], and in the case of cane sugar factories, 14–25 kg per tonne of cane.

6.9. Slicing Beets

From the washing facility, the beets, beet fragments, and beet tails are transported by belts or bucket elevators to the slicer hopper, which is installed above the slicers. As far as possible, this bin should accept the amount of beets processed in an hour and should always be kept full (controlled via load cells) to guarantee uniform feeding of the slicers. The beets are weighed either before the slicer hopper (with a tipping weigher) or after slicing (with a conveyer scale). The weight of the "clean beets" obtained in this manner is used to calculate internal yield.

Besides, vertical shaft slicers with horizontal cutting disks, rotary drum slicers are being used increasingly for the slicing of beets. The output of slicers is controlled by altering their speed, and the set output required is controlled by the conveyer scale in the path of the cossettes to the extraction system. This scale sends out the impulse for output control. A constant supply of beet cossettes is required for good extraction, along with a constant mass flow in the following process steps.

A typical rotary drum slicer is shown in Figure 3 [75]. The beets enter the slicing drum via a feed hopper. As a result of the comma-shaped baffles, the rotating motion of the slicing drum (20–120 rpm), and the resulting pressure of the beet mass, the beets are cut by knives screwed into the knife boxes. Three knives, with a length of 200 mm each, are mounted side by side in a knife box,resulting in a cutting length of 600mm.

If a foreign body appears (stones, metal pieces, etc.), the stone catcher door springs open and the machine automatically comes to a standstill. Drum slicers are built with an output of up to 5000 t/d.

The zigzag Königsfelder knives are used most often today (Fig. 4). These knives should theoretically produce cossettes having a V-shaped cross section. Cossettes of 3–5-mm thickness result in surfaces of 2.5–3.5 m^2 per kilogram cossette, a cossette circumference of 25–30 mm, and a cross-sectional area of 15–35 mm^2.

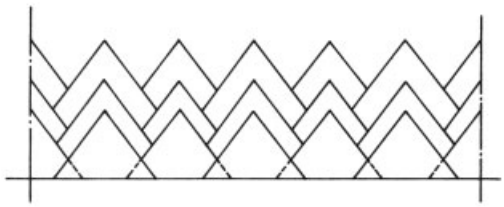

Figure 4. Königsfelder knife

For evaluation of the quality of beet cossettes, two criteria have been introduced:

1) The Silin number gives the length in meters of 100 g of beet cossettes arranged in a row. Well- cut slices give a length of 10 to 16, depending on the spacing of the knives.
2) The "Swedish number" is the quotient "weight of cossettes $\geq$ 5 cm"/"weight of cossettes $\leq$ 1 cm" and has a value of 15–25 for good cossettes.

The mechanical strength (elasticity) of sugar beets has a great influence on cossette quality [76]. The preferred quality, which is slightly dependent on the type of extraction, are small, brittle cossettes.

7. Production of Juice (Extraction)

The extraction of beet cossettes is an important process step that greatly influences sugar losses and subsequent process steps. Upon extraction under sterile conditions with as little water as possible, extensive desugarization of the beet cossettes should be achieved and a raw juice of high purity and density (solids content) should be obtained.

7.1. Principles [77]

The sugar dissolved in cell juice can pass out of the beet cossettes (solid phase) into the extraction liquid only if the cell and vacuole membranes are either opened mechanically (e.g., by producing a finely ground pulp) or denatured thermally. When beets are cut, about 5–15 % of the cells are damaged mechanically to such an extent that an exchange can occur between the cell components, including those dissolved colloidally, and the extraction liquid. The remaining intact cells must be opened irreversibly by the action of heat. Denaturation of the cells should occur quickly (i.e., at the highest possible temperature) [78]. At 80 °C, 80 % of all the beet cells are opened in 5 min; at 70 °C, about 14 min is required to achieve the same result. The transfer of sucrose and other water-soluble nonsugar substances from the denatured cossettes to the extraction liquid occurs in two phases:

1) Transport from the inside of the cossettes to the surface
2) Transport from the surface of the cossettes to the extraction liquid through the entire distance of extraction

Since this is a disperse system, Fick's second law applies to the mass transport. However, mass transport is calculated more easily according to Fick's first law.

If ds is the amount of substance that diffuses in time dt through the surface F of the cossette at a concentration difference dc along the path dr, then

$$\dot{m} = \frac{\mathrm{d}s}{\mathrm{d}t} = D \cdot F \cdot \frac{\mathrm{d}c}{\mathrm{d}r}$$

The diffusion coefficient D for beet tissue is $\approx 1 \times 10^{-5}$ cm^2/min at 70 °C. According to the Einstein relation, the diffusion coefficient is temperature dependent. Consequently, the formulas give the dependence of practical juice extraction on temperature, cossette shape (surface), amount of extraction liquid (concentration gradient), and extraction time.

According to the formula, the extraction temperature should be as high as possible to accelerate mass transport. This can be fulfilled only within certain limits. Temperatures that are too high lead to impure, poorly processible raw juices because structural substances of the cell walls (pectin–araban–protein complex) are also degraded to a considerable extent and thus become water soluble. The following temperature course represents a compromise: the cossettes are heated to 70–78 °C for a short time (about 5 min) to denature the beet cells, and extraction is carried out at 69–73 °C. Optimal temperatures depend on the growth conditions of the beet (i.e., they vary from year to year) and on the equipment. They must be determined in practical operation.

The temperature in the extraction system also influences the vital functions of microorganisms. Soil bacteria, especially mesophilic bacteria, enter the extraction system with the soil still adhering to the washed beets (0.1–0.5 %). Possibly, other strains of bacteria enter the system via the extraction water. As a result of their metabolism, these bacteria cause sugar losses (a part of the "unaccounted losses") that, in the case of larger infections, can amount to 0.1–0.2 % of sugar (beet based). The turnover of mesophilic

bacteria can be limited by avoiding longer spells at the temperatures at which they grow best (< 40 °C) or even by killing them through exposure to higher temperatures. In contrast, thermophilic organisms (e.g., *Bacillus stearothermophilus, Clostridium thermohydrosulfuricum*, now named *Thermoanaerobacter thermohydrosulfuricus* [79]) are not inhibited by the temperatures allowed for extraction. For this reason, regular treatment of the extraction system with disinfectants (→ Biocides, Chap. 2.6, → Biocides, Chap. 2.8) (e.g., hydrogen peroxide, glutaraldehyde, peracetic acid, and especially 30 or 40 % formalin) cannot be avoided. A quantity of formalin equal to 0.5–1.0 % of the hourly raw juice volumetric flow rate (i.e., 50–100 kg per 100 m^3 of raw juice per hour) should be added suddenly. The interval between two additions depends on the intensity of the infection. In general, shock disinfection must be conducted every 8–24 h. Today, after the introduction of hops β-acids (BetaStab 10A), rosin acids, and myristic acid as antibacterials for beet extraction, it is possible to avoid addition of strong disinfectants, but at higher costs [80]. Enzymatic determination of the concentrations of L- and D-lactic acid has proved useful in monitoring the microbial state [81]. L-Lactic acid is the main acid formed microbially in the extraction system. Other carboxylic acids (acetic, propionic, and butyric acids), as well as ethanol, acetone, nitrite, hydrosulfide, and hydrogen can also be formed [79, 82]. Furthermore, pH measurements in the extraction system can be used to detect infection.

The demand for cossettes with a large surface area can be fulfilled to only a certain extent; otherwise, the permeability of the cossettes packet with regard to the juice is not guaranteed. This applies especially to equipment without forced charging. The charging of the extraction equipment, which is expressed as kilograms of fresh cossettes per hectoliter of extraction volume, should be as high as possible to achieve a high relative velocity between the cossette mass and the extraction liquid. In equipment without cossette-forced charging, a filling of 65–75 kg/hL can be achieved. A filling of up to 55 kg/hL is possible in extraction systems with forced charging of cossettes. Calculation of the velocity of the extraction liquid (the juice or operating velocity) is greatly simplified as follows:

$$w = \frac{S}{q_{\mathrm{f}}} = \frac{S}{q\left(1 - \frac{F_s}{d \cdot 100}\right)}$$

where w is the flow rate in meters per minute, S is the amount of raw juice in cubic meters per minute, q_{f} is the free cross section in square meters, q is the cross section of the extraction equipment in square meters, Fs is the cossette filling in kilograms per hectoliter, and d is the relative density of the cossettes [83]. Therefore, the velocity depends on the "raw juice draft" (i.e., the mass of raw juice that goes into the factory) and is expressed in weight percent (beet based). For thermal reasons, efforts are made to attain the lowest possible raw juice draft (i.e., weight of juice/weight of beet); it is normally 105–115 % (beet based).

Furthermore, the raw juice draft influences sugar losses during extraction. The losses are expressed as the sugar remaining in the desugarized cossettes, based on the beet mass processed. These "determinable extraction losses" are 0.12–0.28 % (beet based) in continuous plants with a juice draft of 105–115 % (beet based). This corresponds to an extraction efficiency for sugar of 0.97–0.99. Extraction loss is reduced by about 0.01 % (beet based) by increasing the draft by 1 % (beet based) in each case.

Parallel to a sucrose extraction of 97–99 %, approximately 20–30 % of the total nonsugars are extracted (i.e., about 70–75 % remain in the extracted cossettes). Thus, the purity of the raw juice is 89–91 %, compared to 68–72 % in the beets. Of the nonsugar substances in the beet, about 42 % of the total nitrogen, 88 % of the protein nitrogen, 72 % of the calcium ions, 34 % of the magnesium ions, and 17 % of the alkali ions remain in the pressed pulp. The alkaline-earth and alkali ions in the pressed pulp are, for the most part, present as pectates and pectinates [concentration ca. 70 milliequivalents (meq) per 100 g solids of pressed pulp]. Thus, the degree of extraction for cations is influenced decisively by that of the pectic substances. About 95 % of the pectins introduced with the beets normally remain in the desugarized cossettes. About 95 % of the low molecular mass water-soluble anions in beet are extracted. Since part of the extracted nonsugar substances is precipitated during the juice purification step, a higher sugar yield is also achieved with better desugarization [84].

The extraction time should be short to avoid the prolonged action of heat, which increases the solubility of the structural substances of the cossettes. However, extraction times for the solid phase of 70–75 min, and up to 130 min with some equipment, are required to attain good extraction efficiency. The extraction water used to produce the juice should be free of thermophilic bacteria and preheated to 65–72 °C. The pure condensate obtained on evaporation of the thin juice is well suited to this purpose. The extraction water is adjusted to pH 5.6–5.8 with sulfur dioxide, hydrochloric acid, or sulfuric acid.

The ability of the structural substances to be pressed mechanically is improved by converting the carboxyl groups of the pectins as much as possible to calcium carboxylates. For this reason, 350–700 mg/L of calcium ions in the form of calcium chloride or calcium sulfate are usually added to the fresh extraction water [85]. In this manner, pressed pulp with a solids content of 28–35 % can be produced by purely mechanical means. Since about 90 % of the salts and acids in the extraction water pass into the raw juice and are only partially precipitated during juice purification, a high salt content in the extraction water lowers the purity of the thin juice.

The beet-based quantity of extraction water is calculated as follows:

$$\begin{aligned}&\text{Quantity of extraction water in \% (beet based)}\\&\quad = \text{raw juice draft in \% (beet based)}\\&\quad + \text{pressed pulp obtained in \% (beet based)} - 100\end{aligned}$$

Water quantities of 25–30 % (beet based) are used. Apart from the fresh extraction water, the "press water" serves as an extraction liquid. This water is obtained on pressing the desugarized "wet pulp" from 7–9 % solids content to 18–35 % solids content (pressed pulp). Since the press water contains 0.6–2.5 % sugar, it must be returned to the extraction system "below" the point of addition of extraction water. Depending on the initial and final solids content of the pressed pulp, 35–65 % (beet based) press water is obtained. The press water cycle, including the entire pulp pressing station, is very susceptible to infection. For this reason, efforts are made to have short residence times (small pump recipients, short piping, no collection boxes, etc.) and to reheat the press water to 68–71 °C as quickly as possible. The advantages of recycling the press water compared to using only fresh water are no accumulation of wastewater during extraction, a lower extraction water requirement, an increase in the pressed-pulp solids obtained by 0.3–0.4 % (beet based), higher-quality pressed pulp that contains about three times the sugar of pressed pulp obtained by using only fresh water (sugar content of 6 % compared with 2 % of the solids), and lower extraction losses.

The mathematical control of juice production includes daily determination of the extraction losses, the amount of pressed pulp obtained, and the raw juice draft in percent (beet based). In addition, the sugar content (polarization, Pol) and the dry substance (DS) of the fresh cossettes (FC), pressed pulp (P), and raw juice (RJ) must be determined analytically. Although rare, the amount of pressed pulp obtained can also be determined by weighing. According to the calculation proposed by F. Schneider [86], the amount of sugar X_1 obtained in the raw juice is determined first:

$$X_1 = \frac{\mathrm{Pol_{FC}} \cdot \mathrm{DS_p} - \mathrm{Pol_p} \cdot \mathrm{DS_{FC}}}{\mathrm{DS_p} - \frac{\mathrm{Pol_p} \cdot \mathrm{DS_{RJ}}}{\mathrm{Pol_{RJ}}}}$$

(in % Pol, beet based)

The extraction loss is then

$$X_2 = \mathrm{Pol_{FC}} - X_1$$

the raw juice draft is

$$\frac{X_1}{\mathrm{Pol_{RJ}}} \cdot 100$$

the pressed pulp obtained is

$$\frac{X_2}{\mathrm{Pol_P}} \cdot 100$$

all three in weight percent, beet based. These formulas are valid only in a closed system (i.e., when no wastewater is produced).

7.2. Extraction Technology

The *batchwise operating diffusion battery* (→ Liquid–Solid Extraction) has now been largely replaced by continuous extraction systems (see below). Diffusion batteries consist of 8–16 cylindrical vessels that have a conical top and bottom, and a net capacity up to 12 m^3. Except for two diffusers that are, in each case, being emptied or filled with fresh beet cossettes, all the others are connected in series to a battery.

The juice flows from the top to the bottom countercurrently to the slices.

The *continuous extraction process* was introduced to reduce extraction losses, to avoid wastewater (effluent water), and for purposes of rationalization. Continuous extraction systems can be divided into equipment *without cossette-forced charging,* equipment *with cossette-forced charging*, and equipment *with cossette- and juice-forced charging* [87].

In equipment *without the forced charging of cossettes,* the extraction liquid flows steadily countercurrent to the cossette mass. The cossettes are transported within the extraction area with relatively simple internals (e.g., screws). Examples of this type of equipment are the BMA tower, the Buckau-Wolf tower, and the DdS diffuser. The BMA tower and the DdS slope diffuser are described here (Sections 7.4.1 and 7.4.2).

In equipment *with forced charging of cossettes*, the cossettes are drawn through the extraction area with a special transport system, which consists of endless chains with conveying elements. The Olier system belongs to this type of equipment and is similar in design to a chain conveyor extractor (→ Liquid–Solid Extraction).

Systems with *forced charging of cossettes and juice* consist of a number of cells in which the material being extracted and the extraction liquid move cocurrently for a time. However, countercurrent extraction is obtained on the whole because the extraction liquid or the feed is, in each case, returned to the preceding cell [88]. Examples of this type of system are the RT continuous drum [89, 90] and the De Smet diffusion system (see Section 7.4.3), which is also called trickle diffusion.

The best technological results have been obtained with further development of the RT continuous diffuser. However, this system requires the highest specific investment costs and has the greatest space requirement. Throughout the world, extraction towers, DdS slope, and RT systems are most frequently installed in beet sugar factories.

For a survey of the historical development of equipment and machines for the extraction of juice, evaporation, crystallization, and separation, see [90].

7.3. Extraction with Nonaqueous Solvents

Extraction with nonaqueous solvents has not achieved commercial importance although appropriate experiments with methanol, ethanol, mixtures of these with polyhydric alcohols [92], and liquid ammonia [93] as extracting agent have been conducted repeatedly on a pilot scale. Since these processes can only use dried beet cossettes, they are not profitable because of the high costs of drying and storing cossettes, despite the possibility of year-round operation.

7.4. Extraction Systems

Extraction plants are complex. They consist of:

1) The actual extraction equipment
2) The system for denaturation of the beet cells, which is used simultaneously as a heat exchanger to obtain raw juice that is as cold as possible
3) The pulp pressing station, the resulting press water being returned to the extraction step

7.4.1. BMA Extraction System [90]

The BMA extraction system, supplied by Braunschweigische Maschinenbauanstalt AG, Braunschweig, is depicted in Figure 5.

"Fresh cossettes" from the slicer (a) are denatured in a horizontal countercurrent mixer (c) with the "tower juice", which is warmed up in the preheater (f). The cossette–juice mixture is deaerated and brought to extraction temperature before being pumped (h) into the bottom of the extraction tower (j). The tower juice cools down in contact with the cold cossettes, washes the cell juice out of the opened beet cells by convection, and is withdrawn from the system as raw juice at 20–30 °C through a front screen and via pump (g).

In the extraction tower, the cossette mass is led from the bottom to the top countercurrently to the extraction liquid with the help of suitable internals (e.g., a rotatable tube inside the tower with stirring arms that interact with baffle plates connected to the outer wall). The extraction freshwater fed at the top and press water fed

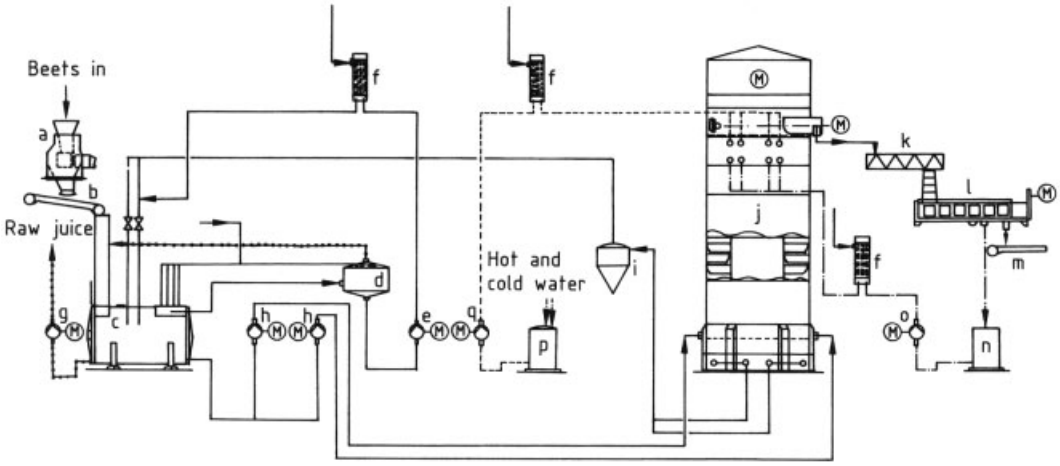

Figure 5. Flow diagram of a BMA extraction plant
a) Slicer; b) Belt conveyor with scale; c) Countercurrent cossette mixer; d) Defoamer; e) Pump for defoamed juice; f) Preheater; g) Pump for raw juice; h) Pumps for cossette–juice mixture; i) Sand separator; j) Extraction tower; k) Screw conveyor; l) Pulp press; m) Belt conveyor; n) Tank for press water; o) Pump for press water; p) Tank for fresh water; q) Pump for fresh water
—- Circulation juice; ●●●●●●● Raw Juice;———-Fresh water; – · – · – · Press Water; —- Vapor; °°°°°°° Vent

a little lower down are preheated (65–70 °C) so that the extraction tower need not be heated.

The desugarized cossettes are discharged at the top and led via screws to the pulp pressing station, where they are pressed to give *pressed pulp* with a solids content of 18–22 %, if used directly, or 22–35 % if dried.

The press water is freed of smaller beet marc particles by rotating or vibrating screen surfaces or sieve-bend screens. It is then heated in preheaters and returned to the extraction system. The pH of the press water should be 5.0–5.5. If the pH is higher, it is adjusted by addition of acid, as in the case of freshwater extraction. If the pH is lower, the press water cycle contains bacteria and must be disinfected. But after a rapid increase of energy costs in 1973, some sugar factories started to operate towers with intentional growth of microorganisms and concomitantly lower pH of press water to improve pulp pressing [91].

Extraction plants of this type were built with a maximum beet processing capacity of 12 000 t/d, but are often operated at higher than nominal rates. The extraction tower has a diameter of 11.0 m and a height of 21.6 m; the countercurrent mixer has a diameter of 6.7 m and a length of 8.5 m. The tower and circulation juice are about 1000 % (beet based) in circulation. For this reason, the circulation juice has to be heated by only ca. 2–3 K to introduce the required amount of heat. The time of extraction is 70–85 min, including 8–10 min in the scalder. At a raw juice draft of 106–115 % (beet based), extraction losses of 0.19–0.25 % (beet based) are obtained.

7.4.2. Extraction with the DdS Slope Diffuser [90, 94]

In contrast to extraction towers, the DdS slope diffuser (Fig. 6) of the De danske Sukkerfabrikker, Copenhagen, consists of a single piece of equipment, a trough (apart from the cossette pressing station). This trough is inclined at about 8°; it is ca. 25 m long, and 4–7 m wide. The cossettes are transported from the bottom to the top against the flow of extraction liquid by two parallel interworking transport screws. The trough is heated by 12 steam caps attached to its outer jacket. Because of the good heat exchange throughout the trough, this system made it possible for the first time to obtain raw juice with a temperature of 20–25 °C. At a raw juice draft of 109 to 116 % (beet based), extraction losses of 0.14–0.26 % (beet based) are obtained [95]. A disadvantage of the DdS slope diffuser is the long residence time of the cossettes in the apparatus (about 125–140 min). These plants are

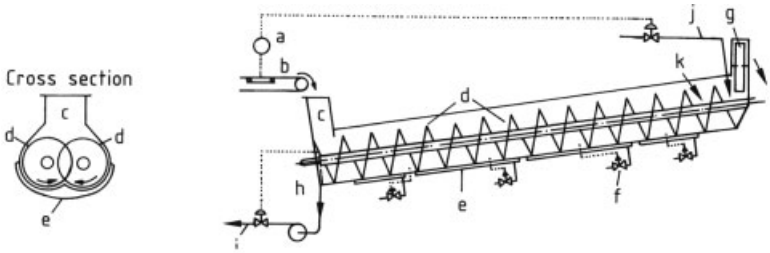

Figure 6. Schematic of DdS slope diffuser [95]
a) Belt weigher; b) Belt conveyor; c) Cossettes bin; d) Screw conveyor; e) Steam cap; f) Control valve; g) bucket wheel for discharge of desugarized cossettes; h) Screen; i) Raw juice draft; j) Extraction freshwater inlet; k) Press water inlet

built up to a beet processing capacity of 5000 t/d.

7.4.3. Extraction with the De Smet System

The De Smet system(→ Liquid–Solid Extraction) belongs to the type with forced charging of cossettes and juice. A compact cossette layer with a height of 60–70 cm is moved intermittently forward at a speed of 3.2 cm/min on the upper part of an endless, horizontal wedge-wire screen (ca. 5 m wide, total length 75 m) moving on two guide rolls. The extraction liquid trickles constantly through the entire cossette layer from the top downward in 18 partial streams. The liquid is collected under the screen in compartments and pumped into the juice gutter above the next compartment against the cossette flow. From here, the juice again trickles through the cossette layer. In this manner, countercurrent extraction is achieved on the whole, although cocurrent extraction occurs in individual steps. With a raw juice draft of 108–115 %, sugar losses of 0.14–0.30 % (beet based) are obtained [87].

7.4.4. The RT-5 Continuous Extractor

The RT-5 extractor [88 – 90] consists of a rotating cylindrical drum having a double screw system, similar to an Archimedes screw, welded to its inner wall. Thus, the residence time of the extraction liquid is very short (≈ 30 min), and the residence-time distribution is very narrow. The cossette mass is present in weighed partial amounts in each of the "compartments." With each rotation, it is lifted out of the extraction liquid and transported forward by two compartments with the help of grates (screens) (number of compartments ≈ 35). To ensure smooth operation of the drum, the lifting of cossette packets is carried out in accordance with the number of compartments, staggered in time, over a period of one rotation of the drum. This elaborate design gives excellent technological values and low extraction losses at a low raw juice draft. The maximum capacity of the RT-5 extractor is 13 000 t of beets per day, the length being 63 m, the diameter of the denaturation compartment 8.95 m, and that of the heat exchange compartment 7.5 m.

7.5. Pressing Extracted Pulp [82, 96]

Two types of presses exist: The single-screw vertical press and the swing screw horizontal pulp press. The presses consist of a cylindrical jacket with a screen pack in which a conical press screw (spindle) rotates (1–4 rpm). The spindle is equipped with spirals and a sieve covering. The ring opening at the end of the press spindle can be made wider or narrower by means of a "cone." Thus, constant pressure can be maintained with a changing output. Pressing is achieved by the fact that the passage volume is diminished progressively due to the conical form of the spindle body and the reduction of pitch of the screw. Press water is withdrawn from the screen jacket and press spindle by means of tubing. A *single-screw vertical pulp press* with a spindle diameter of 4 m and a height of 19.6 m can press cossettes from 3000 t of beets per day. Since the mid-1960s, a *swing screw horizontal press* (produced by Stord–Bartz Industri, Bergen, Norway), operated according to the double screw principle. Today (early 2000s) presses by BMA

(single screw), Babbini (double screw) and Stord (double screw) are installed with maximum capacities of 3500–5000 t/d, based on sliced beet and 30 % dry substance in pressed pulp [1]. Vertical double screw presses supplied by Babbini are among the latest developments [97].

Pressed pulp with a solids content up to 35 % can be obtained by optimization of the operating parameters and further development of construction details. Efforts are being made to attain a solids content that is still higher by double pressing or diffusive dewatering (up to 50 %). For the drying of beet cossettes, see Section 22.5. For uses, see also Section 22.5.

8. Juice Purification

The raw juice with a solids content of 13–17 % has a grayish black, opalescent appearance. Its color is caused by melanin, a blackish pigment, which is formed in the presence of air by tyrosinase (E.C. 1.10.3.1) from the tyrosine in the cell juice via dihydroxyphenylalanine (DOPA) and quinoid compounds. In addition, polyphenol–iron complexes contribute to the color of the raw juice [98].

The raw juice contains almost all the colloidally dispersed and molecularly dissolved substances (see Section 7.1) present in the cell juice of the sugar beet. It also contains the acids and salts present in or added to the extraction water, the metabolic products (L-lactic acid, acetic acid) formed microbiologically during extraction, methanol and acetic acid formed by the demethoxylation and deacetylation of pectin, and suspended particles consisting mainly of pectin–hemicellulose–proteins, macroscopic fibers, and cell residue as well as sand. The composition of the colloid fraction is largely influenced by the temperature course used in juice production (see Section 7). Raw juices have approximately the analytical values given in Table 6.

Until now, attempts to crystallize sugar directly from the raw juice without juice purification have not achieved importance [99].

The main purpose of juice purification is largely to remove the nonsugar substances contained in the raw juice and thus to increase the ratio of sucrose to total solids content (i.e., the purity of the juice) as much as possible. In this process, only those processing aids and reaction conditions should be used that, as far as possible, avoid not only the destruction of sucrose but also the conversion of nonsugar substances to compounds (e.g., colorants) that impede further processing of the juice.

Before juice purification, the fibers and cell particles suspended in the raw juice should be removed if possible. These particles would otherwise give rise to degradation products that have a negative effect on juice quality [100]. Sand should also be eliminated from the raw juice (e.g., by means of hydrocyclones).

The juices are heated to 60 °C before purification. In juice production, efforts are made to obtain a raw juice that is as cold as possible. For the first heating residual heat can be used (pan vapors of ca. 75 °C, condensates), which cannot be utilized otherwise and would save 2–4 kg of steam per 100 kg of beets. However, working with cold raw juice can result in bacterial infections and, consequently, sugar losses and filtration problems.

The heating devices used primarily are standing tubular heat exchangers and, increasingly, spiral-plate and plate-type heat exchangers. The juice in the heat exchangers should flow with a rate of less than 2 m/s so that the sludge particles agglomerated in the precipitation reactions are not destroyed again.

8.1. Lime–Carbon Dioxide Treatment

Colloidally dispersed or molecularly dissolved nonsugar substances can be precipitated out of the raw juice with additives. Of the many substances that have been tested (see, e.g., [101]), only quicklime, which was introduced more than 100 years ago, is of industrial importance. It is available everywhere, and its activity, cheapness, and easy applicability cannot be surpassed by any other substance. Furthermore, the carbon dioxide (CO_2) used today solely to remove excess lime from the juices (carbonation) is obtained as a side product (lime kiln gas) in the production of quicklime from limestone. For this reason, juice purification generally refers to lime–carbon dioxide treatment. Even ion-exchange processes have not been able to displace lime–carbon dioxide juice purification, although only 30–40 % of the nonsugar substances

Table 6. Analytical values of raw juice

Quantity	Analytical values of raw juice	Quantity	Based on 100 g DS
Solids content, %	15–18	Ash content, %	2.50
Sugar content, %	13.2–16.4	Invert sugar content, %	0.40–0.60
Purity (see Section 5)	89–91	Total nitrogen content, %	0.48
Acidity (as % CaO) against phenolphthalein	0.022	Pectin plus proteins content, %	≈0.61
pH	≈ 6.2	Cation content	35–45 meq
		Anion content	40–50 meq
		Amino acid content	10–20 meq

are removed by the latter process, whereas ion exchange eliminates up to 85 %.

An optimal juice purification with lime–carbon dioxide should simultaneously fulfill the following requirements as far as possible [102]:

1) Precipitation or flocculation of the pectin–hemicellulose–protein complex and other high molecular mass nonsugar substances (araban, galactan, etc.).
2) Precipitation of anions, such as phosphate, sulfate, citrate, malate, oxalate, polygalacturonates, that give poorly soluble calcium compounds with calcium ions, and precipitation of magnesium ions as magnesium hydroxide. The higher the content of anions precipitable with CaO in proportion to the total anion concentration, the greater is the excess of alkali ions (potassium, sodium) present in the second carbonation (see page 26) in the form of carbonate and hydrogencarbonate salts. This is connected with a low residual lime content in the thin juice. About 30 % of the total cations and 40 % of the nitrogen-free anions of the raw juice are precipitated during juice purification.
3) Alkaline degradation of invert sugar, galactose, and, partially, galacturonic acid present in the raw juice. In this process, ca. 1.9 meq of acid – 50 % of which is lactic acid (racemic mixture of D- and L-lactic acid) – is obtained from 1 mmol of saccharide. Furthermore, formic, acetic, arabonic, and saccharinic acids are formed. Lighter thin and thick juices are obtained by degradation of invert sugar in the presence of atmospheric oxygen. In this way the formation of degradation products bearing carbonyl groups is prevented, which can immediately react with the amino acids in the juice to give precursors of Maillard products [103].
4) Hydrolysis of glutamine to pyrrolidonecarboxylic acid and, in part, of asparagine to aspartic acid. All known juice purification methods fulfill this requirement incompletely because the temperatures, residence times, and pH values used during juice purification are not sufficient for glutamine hydrolysis. Almost complete glutamine hydrolysis occurs only during the first two steps in the evaporator station, with all the accompanying disadvantages. These include lowering of the pH of the thick juice and re-formation of invert sugar, with subsequent colorant formation and corrosion in the evaporator station. The ammonia generated in glutamine hydrolysis passes over with the vapor condensates (ammonia content 100–200 mg/L). The other amino acids present in raw juice are influenced only slightly during juice purification. Ornithine and lysine are preferentially incorporated into Maillard products. Small amounts of glycine and γ-aminobutyric acid are formed from threonine and serine or glutamine via pyrrolidonecarboxylic acid and its decarboxylation product [104]. γ-Aminobutyric acid further reacts with galacturonic acid or its oligomers to give tagaturon-γ-aminobutyric acid or its derivatives. These are intermediates in the Maillard reaction that gives rise to very highly colored compounds [102].
5) Other nonsugar substances that are not precipitated in the first carbonation, primarily colorants and multivalent anions (citrate, oxalate, etc.), should, if possible, be adsorbed on sludge particles or on the surface of freshly formed calcium carbonate crystals (second carbonation).
6) Prevention of chemical [102] and bacterial degradation of sucrose.

Apart from these chemical requirements, optimal juice purification must also produce an easily settleable and filterable sludge. Almost all the chemical analytical values (e.g., color, residual lime content) deteriorate when filtration problems occur, and turbid juice is obtained (with a turbidity level > 10 mg/L). In classical juice purification, safe filtration can be achieved only when part of the precipitate formed in the first carbonation is returned to the preliming stage.

The requirements for optimal juice purification presented above are influenced by process engineering parameters, such as the following:

1) The total mass of calcium hydroxide added and distributed to the individual process steps determines the activity of calcium ions and the pH of the juice in the individual juice purification steps.
   The CaO required for juice purification is 70–90 %, based on the mass of the nonsugar substances introduced with the raw juice. This corresponds to about 0.8–1.3 kg of CaO per 100 kg of beets. Of the total CaO mass, 10–12 % is added for preliming, 5–7 % to the thin juice 1 (see page 26) before the second carbonation, and the rest for main liming.
   Only the $Ca(OH)_2$ content dissolved in the juice is chemically active. Since calcium hydroxide is more soluble in cold sugar solutions than in warm solutions (35 g of CaO per liter of thin juice at 30 °C compared with 5 g of CaO per liter at 80 °C), the addition of milk of lime and the "chemical reaction steps" should be carried out at the lowest possible temperature. The rate of dissolution of calcium hydroxide in juice depends on the composition of the quicklime (magnesium, aluminate, and silicate content), the burning conditions (hard or soft burning), and the lime slaking technique used, especially the temperatures arising in the pores of the quicklime.
2) The volume of carbonation slurry or slurry concentrate returned to the preliming step. In general, the amount of precipitate returned to the preliming stage with the first carbonation slurry or slurry concentrate is such that the total CaO content in the preliming step is 8–12 g/L.
   Apart from this recycling of carbonation sludge to improve the filtration properties of the juices, the juice must be circulated, especially within the first carbonation, to achieve complete dissolving of calcium hydroxide particles. In general, circulation of six to seven times the quantity of juice via a detention vessel is sufficient. This measure also prevents calcium hydroxide particles from being sheathed by calcium carbonate crystals in the first carbonation and prevents secondary reactions during further processing of the first carbonation slurry (incrustation of the filter cloths, etc.). Furthermore this measure improves the exchange of materials during carbonation.
3) The temperature course during the individual steps of juice purification. Generally, the colors of thin and thick juices are better when the first steps of classical juice purification (preliming and main liming) are carried out at lower temperature (30–40 °C). Today, for fuel economy reasons, a temperature of 55–60 °C is generally used for preliming. Main liming is conducted at 85 °C. A temperature of 85 °C is also used for the first carbonation. If the heat loss in the first carbonation is very high, the first carbonation slurry should be heated to 85–90 °C to obtain better filtration properties. The second carbonation should be conducted at ≥ 94 °C.
4) The residence time during the individual process steps (i.e., the average residence time and the residence-time distribution in the case of continuous operation). The usual average residence time for the individual process steps is

Preliming	20 min
Cold main liming	5 min
Hot main liming	20 min
First carbonation	15 min
Addition of lime before second carbonation	5 min
Second carbonation	
Bubble column	12 min
Reaction vessel	10 min
	87 min

5) The composition of the carbonation gas (carbon dioxide and oxygen content) and the type of gas distribution in carbonation. These substantially influence the exchange of materials (gaseous–liquid and liquid–solid) and the exploitation of carbonation gas.
   In the carbonations, for precipitation of the calcium hydroxide mass mentioned above, 3.41 kg of lime kiln gas per 100 kg of beets is

required in the first carbonation, and 0.72 kg per 100 kg in the second, if the lime kiln gas has a CO_2 content of 40 vol %. For the product of the liquid side mass-transfer coefficient (k_L) times the specific interface (a) for carbonations used today, the following $k_L \times a$ values are calculated: $\approx 250\ h^{-1}$ for the first carbonation and $\approx 50\ h^{-1}$ for the second carbonation at superficial velocities of v_0 = 30–40 cm/s for the first carbonation and v_0 = 8–10 cm/s for the second.

Compared to bubble columns used in other industries, the $k_L \times a$ values for the carbonations are too low, and the superficial velocities too high. This means that the design of the actual carbonation vessels (bubble columns) is not yet optimal.

The values also imply that the installation of mechanical stirrers cannot produce an improvement in the operation of carbonation vessels, especially for the second carbonation.

The average crystal diameter and the crystal-size distribution of the calcium carbonate precipitates formed in the carbonations depend on the temperature and the calcium ion concentration in the juice. The higher the temperature, the coarser are the crystals, and the higher the calcium ion concentration, the finer is the average crystal diameter. Presently, the average crystal diameter obtained (mass distribution) is between 7 and 10 μm.

6) The manner in which carbonation sludge is separated from the juice, i.e., whether direct filtration is employed, e.g., via chamber filter presses (→ Filtration), with decanters and drum filters (→ Filtration, Chap. 8.6), with thickening filters and drum filters, with press filter automatic machines (→ Filtration), or membrane filter presses. During the contact time of the precipitated nonsugar substances with the juice, partial redissolving of the precipitate and other undesirable chemical reactions occur, including acid formation by alkaline degradation of nonsugar substances and sugar. For this reason, the separation of precipitate and juice should occur quickly.
7) The design of the equipment for individual process steps [102]. This applies not only to the reaction vessels, but also to pumping vessels, intermediate containers for juice, etc. These vessels should be designed in accordance with the principle "first in, first out."
8) The maintenance of optimal pH values during preliming, and first and second carbonation, which substantially influences the success of juice purification.

The industrial purification process consists of individual operations. Apart from the main processes described below, numerous process variations are known.

Classical Juice Purification. Classical juice purification consists of preliming, main liming, first carbonation, first sludge separation, second liming, second carbonation, second sludge separation, possibly sulfitation, and safety filtration. The process is represented schematically in Figure 7.

Preliming. In the preliming operation, raw juice is subjected to stepwise alkalinization with milk of lime (ca. 200 g of CaO per liter) up to a concentration of 0.12–0.25 wt % CaO, based on the juice. Today, progressive preliming is used almost exclusively (i.e., the pH of the raw juice is raised slowly from 6.2 to 10.8–11.4). The optimal pH to be maintained can be determined in the laboratory [105]. The optimal pH for the first carbonation is determined in a similar manner.

For preliming, horizontal U-shaped troughs [106] or vertical cylindrical vessels [102, 107], with forced charging of the juice, are used. In both cases, milk of lime is added at the end of preliming, and the pH steps in the individual compartments are achieved by back-mixing of a part of the juice in each case. The slow stepwise increase in pH from 6.2 to 11.4 has proved advantageous for effective precipitation and for stabilization of the precipitated substances because the colloids present in the raw juice have no uniform coagulation optimum.

The carbonation sludge to be recirculated should be added to the juice at pH 8–9 so that the colloids that start to precipitate are already provided with adsorption surfaces such as $CaCO_3$ crystals. In addition, the sludge particles returned are "overcarbonated" due to the lower pH, becoming dehydrated and less soluble. The more carbonation slurry concentrate or slurry is circulated, the better are the sedimentation and filtration properties of the juice. Less than the newly produced carbonation sludge should be circulated (i.e., no more than 1 volume of slurry per volume of raw juice in the case of carbonation slurry or 0.15–0.25 volumes per volume of raw

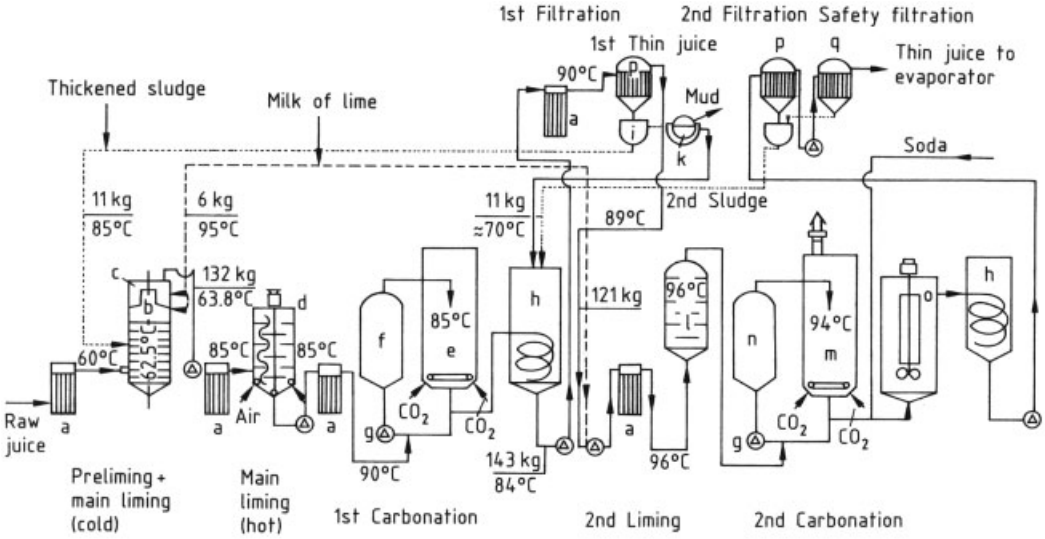

Figure 7. Schematic of juice purification
a) Preheater; b) Prelimer; c) Main limer (cold); d) Main limer (hot); e) 1st carbonation, reaction vessel; f) 1st carbonation, primary tank; g) Recirculation pump; h) Pumping tank; i) Thickening filter; k) Drum filter; l) 2nd liming tank; m) 2nd carbonation, reaction vessel; n) 2nd carbonation, primary tank; o) 2nd carbonation, holding tank; p) 2nd filter; q) Safety filter

juice in the case of slurry concentrate). The recycling causes simultaneous juice dilution, which has a favorable effect on colloid flocculation.

Most of the nonsugar substances are precipitated in the prelimer; i.e., the purity of juices ranging from normal to thin juice 3 (see Section 8.1) differs only slightly from that of the clear preliming juice. However, the preliming floc formed can be separated from the juice only with difficulty.

Only in the *Sepa process* [108] is the carbonation sludge separated after the preliming operation. In this process, the preliming precipitate is dried together with molasses on pressed pulp (see Section 22.5) and then used as feed. The preliming juice is heated to 90 °C and subjected to defeco- carbonation (see Section 8.2.1) to load the flocculated colloids. However, the CaO added should not cause the total CaO concentration in the juice to exceed 0.5–0.6 %. The carbonation slurry obtained in this way is thickened with the help of decanters, and the resulting slurry concentrate is again concentrated in self- cleaning disk centrifuges. The thick sludge with ca. 78 % settleable substances is filtered through drum filters, and the desweetened filter cake is very finely dispersed in ca. 1.8 % of molasses (beet based). The homogenized mixture is sprayed onto pressed pulp. The clear juice from the decanter and filter is again subjected to defeco- carbonation, the juice being further processed as in normal juice purification processes. This method gives strongly colored thick juices.

Main Liming. In the main liming operation, CaO is added to the juice up to an alkalinity of 0.8–1.3 wt % CaO, which results in a total CaO concentration in the main liming juice of 1.8–2.5 wt %. Apart from the chemical reactions at high alkalinity, especially the destruction of invert sugar and amide hydrolysis, calcium hydroxide causes the formation of so much calcium carbonate in the first carbonation that the slimy flocculate formed during preliming becomes enmeshed and filterable. Since 2002, lime addition is often controlled by the filterability of juice, according to the LIMOS method [109], in order to lower the gross consumption of limestone ($CaCO_3$) to values < 2.5 % (beet based).

For the main liming operation, the vertical or horizontal vessels used are connected in a cascade and provided with slow stirrers [104]. The duration of main liming depends on the temperature: the residence time of the juice should be about 25 min at 80–85 °C or 15 min at higher temperature (up to 92 °C).

As a result of the better dosing ability, CaO is now added almost exclusively in the form of milk of lime (wet liming). The dry liming pro-

cedure used earlier, in which the quicklime was slaked directly in the juice, is seldom employed [110] because local overheating could cause sucrose destruction and discoloration of the juice.

First Carbonation [111]. In the first carbonation, excess calcium hydroxide in the juice is partially converted to calcium carbonate with carbon dioxide (carbonation gas). Thus, a filter aid for enmeshing the precipitated colloids (see above) and a surface for the adsorption of nonprecipitable nonsugar substances are produced. Furthermore, part of the ammonia – formed in the main liming by hydrolysis of the amides – and other gaseous organic substances are eliminated from the juice by the passage of carbonation gas. The resulting saturation of the exit gas with water vapor is connected with a heat loss of ca. 6000 kJ per 100 kg of beets. This can, however, be avoided by means of a suitable switch in the process flow [112]. The end point of the first carbonation is reached at pH 10.8–11.4, which corresponds to an alkalinity of 0.07–0.12 % of CaO in the filtered juice (*thin juice 1*). At lower pH (overcarbonation), part of the precipitated colloids is redissolved and the quality of the juice deteriorates. The temperature of the first carbonation should be 85 °C.

To exploit the carbonation gas sufficiently (≈ 85 %), it is passed countercurrently at a juice level of 5–6 m through the juice that enters from the top.

First Sludge Separation. After the first carbonation, the first carbonation slurry, which is heated to 90–95 °C and has a sludge content of 50–60 g of solids per liter, depending on sludge recycling, is separated in decanters into clear juice and slurry concentrate. The decanter volumes required are about 100–120 % of the slurry volume. Suitable decanters are thickeners with rakes or thickeners without forced discharging of the slurry concentrate. The concentration of sludge in the slurry concentrate is 2.5–5 times that in the carbonation slurry.

Sludge separation can be simplified by the use of "thickening filters," which immediately give a filtrate free from turbidity and a greatly thickened slurry concentrate. In *thickening filtration*, the carbonation slurry is filtered directly through continuously and batchwise operating filters. Thus, the residence time of the juice in the decanter is eliminated, and no heat is lost due to cooling. The level of turbidity in the filtrate obtained is generally so low that postfiltration is unnecessary. The sludge content of the slurry concentrate from filter thickeners is five to seven times that of carbonation juice. Cloth-covered tube filters are used preferentially for thickening filtration [113].

Of the total slurry concentrate obtained, ca. 45 vol % is returned to the preliming stage, and about 55 vol % is filtered through multicompartment or single- compartment drum filters (vacuum or pressure drum filters). In this process, 20–30 m^2 of filter surface is required for 1000 t of beet processing per day, depending on the sludge content of the slurry concentrate. The capacity of vacuum drum filters is 0.5 m^3 of clear juice per square meter per hour.

The carbonation sludge is washed on drum filters with about 80–100 % of hot water (condensate), based on the sludge volume, until a sugar content of 0.1–1.0 % is attained (desweetened). Depending on the lime added in juice purification, 8–10 % (beet based) of sludge is obtained with a solids content of 45–50 %.

The decantated material (level of turbidity 50–100 mg/L) and the drum filter filtrate (level of turbidity 200–500 mg/L) are refiltered separately or together through ceramic or cloth- covered tube filters (→ Filtration) or bag filters (→ Filtration). The level of turbidity in the postfiltered juices is less than 10 mg/L. In the case of cloth- covered filters, the capacity is ca. 2–3 m^3 m^{-2} h^{-1}. If thickening filters are used, the drum filter filtrate is returned to the first carbonation.

To immediately obtain a spreadable carbonation lime cake that can be used as fertilizer lime, automatic membrane filter presses (→ Filtration) or automatic press filters (→ Filtration) [114] are being used increasingly for filtration of the slurry concentrate. These automatic machines produce "carbolime" with a solids content of > 70 % and a CaO content of > 30 %, which is a suitable fertilizer lime. The composition is given in Table 12.

Second Carbonation. In the second carbonation step, more carbon dioxide is led into the filtered juice from the first carbonation to convert the remaining calcium hydroxide to carbonate and, in addition, to transform the alkali present in the juice, mainly potassium hydroxide, into hydrogencarbonates or carbonates. In a secondary reaction, the calcium bound to organic acids or complexes is partially precipitated as carbonate.

A sufficiently low residual lime content cannot always be attained in the thin juice 2 (<30 mg CaO/100 g DS). Further softening can be achieved by adding aqueous sodium hydroxide–sodium carbonate to thin juice 1 or by ion-exchange treatment of thin juice2 (see Section 14.1).

To increase the adsorption of nonsugar substances by increasing the calcium carbonate crystal surface and to improve the filtration properties of the second carbonation juice, 0.1–0.2 % CaO, juice based (after filtration of the first carbonation juice), is added to thin juice 1 before the second carbonation. The temperature of the second carbonation should be 94–98 °C. A pH between 8.9 and 9.2 (alkalinity 0.015–0.022 % CaO) is maintained as the end point of the second carbonation, at which the remaining residual lime content in the filtered juice (*thin juice 2*) is at a minimum.

The juice reaching the second carbonation must not contain any sludge particles (i.e., it must be free of turbidity after filtration) because these particles would otherwise redissolve. Even very little residual sludge produces the juice shade characteristic of overcarbonation and a noticeable deterioration in color [115].

As in the first carbonation, cylindrical vessels with a juice level of 5–6 m are also used in the second carbonation. In contrast to the first carbonation, the second is often carried out cocurrently, juice and carbonation gas being introduced into the vessel from the bottom. The juice from the second carbonation flows through a holding tank, a cylindrical vessel with a central tube, which contains a propeller mixer. This mixer circulates the juice 20 times from the bottom to the top to mechanically remove from the juice the carbon dioxide liberated by the changing of the CO_3^{2-}–HCO_3^- equilibrium.

Control of Carbonations. The optimal end points of the first and second carbonations are determined experimentally. Addition of carbonation gas is controlled by continuous measurement of pH.

Second Sludge Separation. Since the amount of carbonation sludge obtained on filtration of the second carbonation slurry is only 2–5 % of that obtained on filtration of the first slurry (in cases where no calcium hydroxide is added in the second carbonation), slurry 2 is usually filtered directly through cloth- covered filters. Only in exceptional cases is it preseparated in decanters into clear juice and slurry concentrate. The sludge obtained is not "desweetened" in filters, but is either recirculated to the preliming stage or added to slurry concentrate 1, which is then filtered and desweetened.

Sulfitation. In some factories, the next stage is a so- called third carbonation. This comprises the treatment of the juice with gaseous sulfur dioxide to adjust the pH and brighten the color (prevention of excessive discoloration during subsequent evaporation). This step should be performed only if the pH of the thin juice is so high (pH > 9.2) that after evaporation (see Section 9) an alkaline thick juice (pH > 8.8) would be obtained. If the pH is not so high, it must be raised by the addition of soda (50–200 g per cubic meter of juice). The amount of SO_2 added should not exceed 50 g per cubic meter of juice. Otherwise, the pH of the juice can decrease sharply during evaporation, which causes corrosion. The exact amount of sulfur dioxide is usually added at the lowest point of a U-shaped pipe via a control valve. The SO_2 inlet pressure is kept constant by heating the SO_2 storage tank to a certain temperature (e.g., 30 °C $\triangleq$ 471 kPa).

Safety Filtration. If the thin juice is sulfurized, it must be filtered again before evaporation to remove precipitated calcium sulfite, which would otherwise result in strong scaling in the first stage of the evaporating station. Safety filtration is recommended, even in the absence of sulfitation. Tube filters that are precoated with a filter aid are an advantage. The level of turbidity (water-insoluble solids > 0.45 μm) of the filtered *thin juice 3* is less than 10 mg/L.

8.2. Other Juice Purification Processes

8.2.1. Simultaneous Defeco-Carbonation

Simultaneous defeco- carbonation refers to juice purification processes in which the raw juice, heated to 85–90 °C, is simultaneously treated with lime and carbon dioxide, while an exact or only slightly fluctuating pH value is maintained. In contrast to classical juice purification, in defeco- carbonation the colloids are precipitated in the presence of carbon dioxide in such a manner that the sludge particles are enmeshed immediately by calcium carbonate crystals. The

conglomerates thus formed are generally larger, better dewatered, and more solid than the sludge particles in classical juice purification. Therefore, the sedimentation and filtration properties of these sludge particles are very good. On the other hand, the resulting juices are only slightly thermostable because of incomplete destruction of invert sugar and low amide hydrolysis. They are stronlgy discolored in subsequent processing steps so that much darker thick juices are obtained.

Simultaneous defeco- carbonation is carried out according to the Dorr system, especially in the United States and the United Kingdom. The heated raw juice, together with six to eight times the amount of slurry, is first pumped into the bottom of a mixing vessel (primary tank), which is as large as the subsequent carbonation vessel. In the primary tank, a pH of ca. 10 is attained. Milk of lime (1.0–1.6 % CaO, juice based) is added as this juice mixture overflows into the actual carbonation vessel. Here, the juice is carbonated countercurrently to an optimal pH of 10.8–11.4. The finished juice is withdrawn from the bottom of the carbonation vessel. About six-sevenths of the juice is returned to the primary tank as circulated juice, and one-seventh proceeds to sludge separation. Further process steps correspond to those of classical juice purification.

8.2.2. Braunschweig Juice Purification

A further development of the Dorr system, the Braunschweig juice purification process, is used occasionally in Europe [116]. In addition to returning the slurry to the mixing vessel of the Dorr system, this process also requires the thick sludge to be recirculated to the raw juice in a separate mixing or reaction vessel located earlier in the process flow. Before carbonation at pH 10.8–11.2, the raw juice colloids pass through two separate pH steps during the thick sludge recycling (pH = 9) and the slurry recycling (pH = 10). Here, too, juices with very good sedimentation and filtration properties are obtained. The destruction of invert sugar, as in main liming, does not occur, which means that thermolabile juices are again obtained.

8.2.3. Defeco - Carbonation with Intermediate or Post Main Liming

The idea of combining the advantages of classical juice purification (attainment of thermostable juices that are free from invert sugar and have a good color) with the advantages of defeco- carbonation (good sedimentation and filtration properties of the precipitates produced) led to the development of other combination processes.

In the Braunschweig juice purification process (see Section 8.2.2), stabilization of the flocculated colloids is so good that defeco- carbonation for chemical stabilization of the juices can follow without impairing the filtration ability of the resulting carbonation slurry. For this reason, in one of the combination processes, before the separation of the precipitate, another addition of lime (main liming) to the slurry – which causes the pH of the juice to increase to > 12.5 for a certain time [117] – is carried out before a further carbonation step (pH 10.8–11.2).

In the *Novi-Sad juice purification process* (former Yugoslavia) [118], efforts are made first to flocculate the precipitable nonsugar substances in a type of defeco- carbonation; the precipitate is subsequently separated, and the juice is chemically stabilized in a main liming step. The lime is added as follows: 1.0–1.3 % of CaO is added in the defeco- carbonation, and 0.6–0.7 % of CaO (beet based) is used for chemical stabilization in the main liming step. The precipitate obtained after main liming in a normal first carbonation at pH 10.8–11.2 (i.e., mainly calcium carbonate) is separated with decanters. The resulting slurry concentrate (ca. 10–14 %, juice based) is returned to the defeco- carbonation. In this way, the entire sludge obtained can be filtered through one filtration unit [119].

9. Evaporation of Thin Juice

After purification, the thin juice has a solids content of 15–18 %. It is concentrated in a multistage process to a solids content of 68–74 % (thick juice), ca. 85–100 kg of water being evaporated per 100 kg of beets. From ca. 115–130 % (beet based) of thin juice, 25–30 % (beet based) of thick juice is obtained.

Table 7. Heat-transfer coefficients of juice (in W m^{-2} K^{-1}) as a function of total solids content

Solids range, wt %	Circulation evaporator	Falling-film evaporator	Plate evaporator
>15–25	2900–3500	2900–3500	3500–4000
>25–40	1700–2300	2300–3100	
>40–55	800–1400	1200–2300	
>55–65		800–1700	
>65	700–900	700–1000	1200–1600

Evaporated water accumulates as vapor condensate in the individual effects of the evaporator station. It contains about 6–10 $\mathrm{mmol/L}NH_4^+$, partly as $(NH_4)_2CO_3$, and varying amounts of steam-volatile organic substances [120]. The condensate is used as fresh water for extraction, as dissolving water in the sugarhouse (remelt, dilution of syrups, etc.), and as plant water for the wash and flume water cycles. The surplus must be treated biologically (see Section 6.8).

During evaporation, the juice generally undergoes only slight chemical alteration. The remaining glutamine is hydrolyzed primarily to pyrrolidonecarboxylic acid with elimination of ammonia. At the same time, the pH decreases (danger of sugar hydrolysis), and alkaline and oxidative decomposition of sugar occurs to a slight extent. Calcium salts of various acids (calcium oxalate, citrate, sulfate, and carbonate) precipitate, along with SiO_2 in some cases, due to the change in solubility conditions. To prevent the formation of scale on the evaporator tubes, calcium complexing agents (antiscaling) are added to the juice (5–10 g per cubic meter of juice). During juice evaporation, a dark color appears, owing to the formation of Maillard products [103] (condensation products from reducing carbohydrates and amino acids), Strecker degradation, and caramelization processes [103]. The discoloration [121] depends greatly on juice temperature (it should be $<$ 132 °C), H^+ activity, average residence time, and residence-time distribution.

Steam-heated evaporators of various designs (Robert evaporator, → Evaporation; falling-film, evaporator, → Evaporation; or plate evaporator) are used for juice evaporation. They have heating surfaces of normal or stainless steel tubes, that are smooth or twisted, up to 6000 m^2 per evaporator. A total heating surface of 2.0 to 2.5 m^2 is required for 1 t of processed beets per day.

The evaporators are arranged in multiple-effect evaporator stations in such a way that the saturated vapor generated in one effect is, in each case, used to heat the next effect and other heat consumers (cocurrent evaporation). The evaporator stations consist of four to six effects with a pressure drop from ca. 0.35 MPa in the calandria of the first effect to 0.12 MPa in the calandria of the last effect. In the process, the temperature of the juice falls from 130 °C to ca. 90–95 °C. The last effect operates under weak vacuum. The vacuum is produced by condensing the vapor with a 15–20-fold amount of water in jet condensers and by withdrawing the noncondensable gases (water-ring pumps, piston pumps). In industrialized countries, the heated water (condenser water) is recycled after cooling in cooling towers.

The heating surfaces must be distributed over the individual effects so that the desired thickening of the juice to 68–74 % solids content is achieved in the last effect. Also, the entire remaining saturated vapor requirement of the factory should be covered by vapor from the last two effects. The size of the heating surfaces in the individual effects must be calculated from case to case by taking into account the mass flow and the vapor consumption of the individual vapor users [122].

Guide values for the overall heat-transfer coefficient k (in W m^{-2} K^{-1}) for various solids contents in the case of rising and downflow evaporators are given in Table 7. The solids content, which correlates with the viscosity of the juice to be evaporated, greatly influences the heat transfer coefficient [121].

Because long average residence times and poor residence-time distributions in the individual evaporators are a disadvantage to the juice [strong discoloration, sugar destruction [123] up to 0.15 % (beet based)], efforts were made to decrease the entire installed heating surface by increasing the heat-transfer values or to reduce the residence time of the juice by using other evaporator designs (plate evaporator). Values for k of 7000–8000 W m^{-2} K^{-1} can be achieved for the

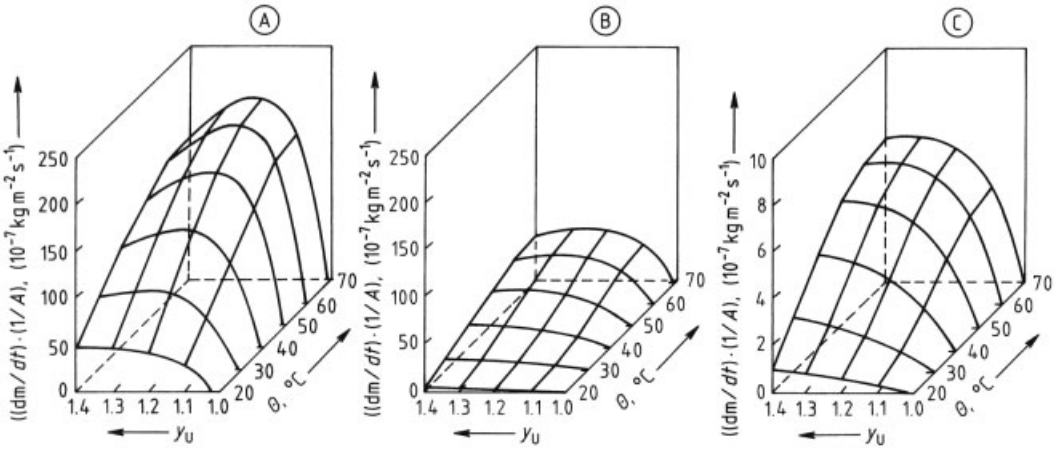

Figure 8. Rate of crystallization of sucrose as a function of supersaturation and temperature at different purities
A) Purity = 100 %; B) Purity = 80 %; C) Purity = 60 %

first effect and 2500 W m^{-2} K^{-1} for the second effect by increasing the juice velocity within the piping [e.g., by forced feeding (Buckau–Wolf continuous evaporator)]. Higher k values can be attained by using falling-film evaporators, and a smaller temperature difference is obtained between the calandria and the vapor space in the individual effects. As a result, the residence time of the juice is reduced by about half. This applies more to plate evaporators, the average residence time being reduced to 1–2 min per effect, which is only one-fifth that of a Robert evaporator station [124]. In 1992 an advantageous combination of plate evaporators and falling-film principle was introduced, and in the following years various integrated concepts were developed, with high k values and small heating surface, compared to other technology [125].

The steam requirement of the evaporator station and, therefore, the energy requirement for sugar production depend on the mode of operation of the factory. In the case of white sugar factories that do not produce refined sugar, the requirement is ca. 20–25 kg of normal steam per 100 kg of beets (54 000–67 000 kJ, corresponding to 15–18.5 kW · h). In factories with refined sugar production, this is increased by ca. 10 000 kJ, and in raw sugar factories, it is ca. 40 000 kJ. For the steam and energy economy of a sugar factory, see [126].

The precipitation and partial caking of sparingly soluble calcium compounds on the heating surfaces of the evaporators can greatly lower the heat-transfer and evaporating efficiency. In this case, the heating surfaces must be cleaned (boiled out). The equipment is treated first with a 3–5 % NaOH–Na_2CO_3 solution, then with a 3 % HCl solution containing an inhibitor, and finally with a 1–2 % Na_2CO_3 solution at the boiling temperature.

The heating surfaces of the evaporators that come in contact with the juice in the first and second effects are susceptible to corrosion. Corrosion occurs only occasionally, but to a very great extent. The causes of corrosion are disputed [127]. It is, in all probability, largely acid corrosion, which is favored by the fact that the oxide layers formed are constantly peeled off by the complexing agents (malic and citric acids) present in the juice. Acid corrosion can also be caused by the release of sulfur dioxide added to the thin juice in the vapor bubbles for a short time during evaporation. Alternatively, the metal may be attacked by an increased hydrogen-ion activity resulting from a decrease in pH of the juice by ca. 1.0–1.5 units at higher temperature [128]. Thus, the pH of thin juice should be kept high enough that the pH values of the thick juice do not fall below 8.8, measured at 20 °C. The thick juice is filtered before further processing.

10. Production of Sugar from Thick Juice

In the sugarhouse, sugar is produced from the thick juice obtained by the evaporation of thin juice. This involves crystallization (recovery of massecuite magmas) and separation of the sugar from the mother syrup (centrifugation). Both crystallization and separation of the sugar must be carried out in several steps because the technically achievable degree of desugarization in a single crystallization is limited by the maximum amount of crystals (50–60 % yield) in the magma (i.e., crystal mass) permitted for adequate movement and the increase in nonsugar substances in the mother syrup.

10.1. Production of Massecuite

From the thick juice (solids content = 68–74 %, sugar content = 61–70 %, nonsugar content = 4–7 %), multistage crystallization produces 85–92 % of crystalline sucrose without the use of ion-exchange processes or other molasses desugarization methods. The rest of the thick juice sugar (8–15 %) and practically all of the nonsugar substances in the thick juice are found in molasses, the mother syrup of the last crystallization (solids content = 84–89 %, sugar content = 50–56 %, nonsugar content = 32–40 %). If inclusion of mother syrup in the crystal and imperfections of the surface are prevented by optimal crystallization techniques, more than 99 % of all the ionogenic nonsugar substances and ca. 95 % of all the high molecular mass and surface-active nonsugar substances, including colorants, remain in the mother syrup after each crystallization step.

The syrups (thick juices, remelts) can be simplified regarded as three-component systems consisting of sugar, nonsugars, and water. If precipitation crystallization with organic solvents is disregarded, two methods are available theoretically for the industrial-scale crystallization of sucrose. Both methods are used in practice.

1) Evaporating crystallization (→ Evaporation, Chap. 6, → Evaporation, Chap. 7) is performed either isobarically or isothermally. With pure syrups, the sugar–water ratio in the mother syrup remains approximately constant during evaporating crystallization. With impure syrups ($p < 90$), however, this ratio increases because the amount of nonsugars in the mother syrup increases significantly during crystallization. Therefore, the solubility-increasing effect of the nonsugar substances also increases greatly. The solids content of the magma increases during evaporating crystallization.
2) Cooling crystallization is based on the decrease in solubility of sucrose with decreasing temperature (see Table 2). The remaining sucrose concentration in the mother syrup essentially follows the saturation line in the case of pure sugar solutions. In less pure industrial sugar solutions ($p < 85$, see Section 5), the solubility of sucrose in the mother syrup is considerably increased. In pure cooling crystallization, the solids content of the magma and the nonsugars–water ratio in the mother syrup remain constant. Cooling crystallization is used only in combination with evaporating crystallization.

Supersaturation Coefficient, Measurement, Control. The rate of cooling or evaporation during a crystallization process must be adjusted to the immediate crystallization rate (→ Crystallization and Precipitation, Chap. 4.2) to prevent secondary nucleation (→ Crystallization and Precipitation, Chap. 4.1.2) and achieve the most uniform crystal formation and distribution. The rate of crystal growth is controlled mainly by diffusion. The increase in mass of a crystal per unit time depends on the crystal surface [129]; the rate of insertion of the molecules into the crystal surface or the diffusion constant of sucrose in the solution [130]; the thickness of the diffusion boundary layer, which is between 60 and 150 µm [131]; and the supersaturation (Fig. 8). Mass transfer in the magma is influenced substantially by flow conditions [132] so that in evaporating crystallization, natural vapor-bubble mixing is supported by mechanical stirring [133].

The sucrose concentration is normally represented by the supersaturation coefficient c_1. This refers to the quotient

$$c_1 = \frac{c}{c^x}$$

where c is the sugar content of the solution in question and c^x the saturation concentration of sugar in the solution at the same temperature.

The range c_1 = 1.0–1.2 is called metastable. In this range, only crystals that are already present can grow; no new crystals can be formed. In the intermediate saturation range c_1 = 1.2–1.3, however, crystal formation can be caused by shock (air, impact, etc.). In the labile range above $c_1 \geq 1.3$, crystal nuclei are formed spontaneously. Maintaining the proper supersaturation coefficient is of great importance for practical crystallization. Since no direct method exists for continuous measurement of the supersaturation coefficient, it must be determined by indirect methods. The following parameters are suitable for this purpose: measurement of the electrical conductance, measurement of the boiling-point elevation, determination of viscosity or consistency [134], shifting of the oscillator circuit of radio-frequency sources [135] (frequency range 6–150 MHz), determination of the total solids content (e.g., radiometrically [136] or by high-frequency ultrasonic vibrations), and refractometric measurement of the solids content of the mother syrup [137]. Measurements of the viscosity or consistency of the magma (stirrability measurement [134]), total solids content of the magma [136], and density of the mother syrup have proved useful as guiding values for the entire crystallization process [137].

Attempts have been made to control crystallization via on-line measurement of the crystal-size distribution with the help of laser-diffraction spectrometers.

10.2. Evaporating Crystallization

Batchwise Process. Evaporating crystallization is carried out mostly batchwise in pans at 20–30 kPa and a massecuite temperature of 65–80 °C. The process can be divided into four sections [138]:

1) A quantity of syrup to be boiled is drawn into the pan so that the heating chamber remains covered during thickening of the syrup. The syrup is then evaporated to a supersaturation coefficient c_1 of 1.0–1.2.
2) A known number of crystals in the form of microcrystals (d' = 10–20 µm, $n \approx 2$, suspended in 2-propanol) or seed magma (d' = 180–250 µm, $n \approx 3.0$) are added to the syrup. The grain-size distribution of sugar crystals follows the Rosin–Rammler–Sperling distribution, n is a parameter of the Rosin–Rammler–Sperling distribution; high n indicates uniform grain size (→ Elutriation).

When working with microcrystals, the number of crystals added must be larger than the desired number of crystals in the entire massecuite in the pan at the end of sugar boiling. The reason for this is that part of the crystal nuclei dissolves because of the varying temperature and supersaturation conditions in individual tubing sections of the heating chamber. Table 8 lists the number of total surface of the crystals contained in 1 kg of sugar as a function of crystal size [139]. Until crystals become visible to the naked eye (interfacial length of ca. 50 µm), the supersaturation coefficient must be kept at about the same level. During this time, only small amounts of sugar crystallize, but the movement of massecuite must be maintained by means of water evaporation or mechanical stirring. For this reason, water instead of syrup is drawn into the pan in this phase. Since this type of "crystal seeding" results in a higher energy requirement, is technologically difficult to control, and is more time consuming, a seed magma has become the crystal seed of choice. This seed magma is produced separately by evaporating–cooling crystallization in a process similar to the one described here [140]. The mass of the crystal magma charge increases by a factor of 10^3 (30–50 kg of seed magma per tonne of massecuite in the pan). This mode of operation allows the use of feed syrups of higher concentration (e.g., thick juice and liquor with a solids content of 72–74 %; for details on liquor, see Section 13.2).

Table 8. Number and total surface area of crystals in 1 kg of sugar as a function of crystal size [119]

Size of crystals* L_s	Number of sugar crystals $N = 840 \times L_s^{-3}$	Total surface, m^2 $A_m = 4.21 \times L_s^{-1}$
0.2	105×10^6	21
0.4	13×10^6	11
0.6	3.9×10^6	7.0
0.8	1.6×10^6	5.3
1.2	0.49×10^6	3.5
1.5	0.25×10^6	2.8

* Mesh size of the screens in millimeters.

3) In the following phase, the crystals should grow to their final size without formation of new crystals (secondary nuclei) or conglomeration of the crystals present (agglomerated grain). To work at high supersaturations and, consequently, high rates of crystallization, the layer thickness λ (i.e., half the space between two crystals) must be less than 0.25 mm at the start and about 0.1 mm at the end of crystallization (layer thickness λ = volume of mother syrup in cubic meters per surface of sugar crystals in square meters). In addition, adequate circulation must be maintained within the massecuite in the pan. Therefore, natural vapor-bubble mixing is being increasingly supported by mechanical stirring [133]. The continuous or periodic feeding of juice and the rate of evaporation should be mutually adjusted in such a way that the requirements mentioned above are largely fulfilled. Control values here are indirect supersaturation measurements and the level in the pan.
4) Finally, the massecuite in the pan is concentrated to the proper consistency (crystal content 50–60 %) for centrifuge work. The solids content at the end of the crystallization process is 89–91 % in the case of "white" massecuites (refined massecuites, see Section 13.2, white sugar 2, raw sugar 1) and 92–96 % in the case of "brown" massecuites (i.e., raw sugar-2 and after-product massecuites, see Section 13.2).
The flowability of the magmas depends on the viscosity of the mother syrup (solids content, temperature, pH); crystal content; crystal-size distribution; and crystal form [141]. For this reason, only approximate values for the flowability of magmas can be given. The "viscosity" of refined massecuites is between 50 and 100 Pa · s, of white sugar massecuites between 90 and 170 Pa · s, of raw sugar-2 massecuites between 300 and 800 Pa · s, and of after-product massecuites between 500 and 2000 Pa · s. The end point of sugar boiling is determined by the consistency of the magma.

Despite batchwise operation, the technological course of the process is automated to a large extent [142]. A pan holds a maximum of 120 t of massecuite; the installed heating surface in square meters is about 45–60 % of the massecuite content in 100 kg (i.e., 200–300 m^2 calandria surface for a 50 t pan). The calandria are either tubular (brass or steel tubes with an inner diameter of 95 mm) or ring or hexagonally swaged calandria tubes with about 50 mm of clear space between two walls. To permit good circulation of the massecuite in the pan, a central tube is located in the middle of the calandria (e.g., for welded tubular and ring heating chambers), or an annular space is present at the outer edge e.g., for suspended calandria (dished floating calandria). In calandria with a central tube, additional circulation is possible via mechanical stirring. In this case, the massecuite is pushed through the central tube into the space below the calandria by a slowly operating mixer (60–80 rpm) [133]. The vacuum required (15–25 kPa) for boiling is produced as described in Chapter 9.

Continuous Process. Intensive efforts have enabled continuous evaporating crystallization to be developed into a controllable process on the condition that about 25 % of the seed magma (based on finished massecuite) produced batchwise is fed continuously into the process. New plants in industrialized countries, especially for "brown" massecuites (raw sugar 2, after-product), are installed as continuous plants [90, 129, 143].

Two types of equipment exist: horizontal and vertical. In *horizontal equipment*, the total crystallization distance can be divided without problem into several compartments (stirred-tank cascades, $n < 12$). With *vertical crystallizers*, a compromise must be made between the technological requirement for a good residence-time distribution (i.e., more than six compartments) and equipment costs. For this reason, vertical crystallizers usually consist of only four stirred tanks.

The individual process steps described above for batchwise crystallization are carried out one after another in a continuous process, spatially separated in individual compartments. This has both advantages and disadvantages. Thus, a crystal size distribution ($n > 3.0$) as good as that in batchwise operation has not yet been achieved because secondary crystal formation in individual compartments cannot be prevented. Unlike the batchwise process, secondary crystals cannot be redissolved, if necessary, by diluting the mother syrup (attainment of subsaturation) [144].

The advantages of continuous crystallization are as follows:

1) The temperature difference between heating steam and magma is smaller because the massecuite level is low, which makes use of low-pressure steam possible [145].
2) Individual compartments are easier to control. Continuous boiling equipment is usually operated isobarically. For vertical equipment, however, various vapor pressure steps are employed in individual chambers. In this way, evaporating–cooling crystallization is achieved (flash evaporation) [146].

Easy scaling of the equipment at the interface of magma–vapor space, especially in the case of white massecuites, is a problem that has not yet been solved [147].

10.3. Mixing of Massecuites and Cooling Crystallization

After the completion of evaporative crystallization, the massecuites are stirred continuously in mixers/crystallizers until the sugar is centrifuged off. This prevents separation of the crystal mass from the mother syrup by sedimentation. In mixing white massecuite, the temperature is usually held constant. Temperature changes, especially at the walls, must be avoided because of possible secondary nucleation and conglomerate formation.

Cooling Crystallization. Controlled cooling of the massecuite in the crystallizer during the mixing process (cooling crystallization) is generally employed for after-product massecuites (see Section 13.2) and is used only occasionally for raw sugar-2 and white sugar massecuites [148]. The advantages that can be achieved from a technological and heat economy point of view are so great that cooling crystallization will definitely be introduced in the near future for raw sugar-2 and white sugar massecuites [149].

In the case of after-product massecuites, about 85 % of the sugar crystal mass is produced by evaporating crystallization and the remainder by cooling crystallization. In cooling crystallization, the massecuite in the crystallizer is cooled by 0.5–1.5 °C/h from ca. 65 °C to 35–40 °C with constant stirring. This results in cooling crystallization times of 36–40 h. The crystallizers (→ Glucose and Glucose-Containing Syrups) used for cooling are horizontal U-shaped vessels or horizontal or vertical cylindrical vessels in which rotating or stationary cooling elements ($\approx$ 1.5–2.5 m^2/m^3 crystallizer volume) are installed. Thermostated water flows through these elements (temperature difference between the water and the massecuite in the crystallizer $<$ 15 °C). To lower the consistency, which increases with decreasing temperatures, water or syrup must be added to the massecuite during cooling in an amount that prevents the magma viscosity from rising above 3000 Pa · s.

An increase in consistency of the massecuite in the crystallizer can also be prevented by first cooling the massecuite only by 10–15 °C. Part of the massecuite (ca. 20 %) is then centrifuged. The centrifuged mother syrup is reheated, deaerated, and returned to the magma. In this manner, the crystal content of the magma is lowered by about 7 %. The crystal content increases again during further cooling, without the massecuite consistency increasing in an uncontrollable way. This mode of processing leads to a decrease in purity of 2–3 units in the molasses and, thus, higher sugar yields [150].

Cooling crystallization of other massecuites in the crystallizer, such as white sugar and raw sugar, is carried out in similarly designed crystallizers. Cooling rates up to 8 °C/h are possible, depending on product purity. The installed cooling surface per cubic meter of crystallizer volume is between 2.5 and 5 m^2. The temperature difference between cooling medium and massecuite in the crystallizer should not exceed 5 °C because of possible secondary nucleation at cooling surfaces.

10.4. Factors That Influence Rate of Crystallization

The rate of crystallization of sucrose is expressed as mass growth rate G (g m^{-2} min^{-1}) or as linear growth rate L_G (μm/min), both related by $G = 0.711 \times L_G$ [139]. The growth rate depends primarily on purity, temperature, and supersaturation (Fig. 8) [151]. Under otherwise identical

conditions, the rate is reduced by high molecular mass substances (e.g., dextrans), colored substances (e.g., Maillard products), and also by raffinose, kestoses, etc. [152]. A raffinose concentration greater than 1 % (based on magma solids) causes a marked decrease in the rate of crystallization and the formation of needle-shaped crystals, which is especially pronounced when calcium saccharate is returned to the juice purification (Steffen process, see page 50).

These problems can be avoided by cleavage of raffinose into sucrose and galactose with α-galactosidases [153]. Raffinose is usually cleaved in the raw sugar-2 green syrup. Attempts to eliminate the invertase contained, in addition to α-galactosidase, in the enzyme mixtures have been successful [154]. Apart from improving crystallization work, these processes permit better exhaustion of the molasses. However, the process is profitable only under certain conditions.

11. Separation of Sugar from Massecuite

Crystallized sugar is recovered from the massecuite mostly by batchwise centrifugation in screen-basket centrifugals [90] (top-suspended centrifuges, → Centrifuges, Filtering). This operation – including loading and unloading – is automated to such an extent that a station with several machines works practically continuously.

Continuously operated centrifuges, especially continuous conical basket centrifugals, are used only for centrifugation of after-product massecuites [156] and, to a lesser extent, for raw sugar-2 massecuites. In the raw sugar industry, pusher centrifugals are also used (→ Centrifuges, Filtering). Because of the crystal damage occurring in these centrifuges, they are only partly suited to centrifugation of consumption sugar (refined sugar, white sugar).

The first mother syrup obtained on centrifuging massecuite is called "green syrup." Molasses is obtained in the last crystallization step. For the theory of the formation and exhaustion of molasses, see [156].

If the sugar is subsequently washed in the centrifugal with hot water and steam to completely remove syrup particles adhering to the sugar crystals, a "wash syrup" is obtained. This washing is required in the production of consumption sugar (white, refined) and in the affination (washing the crystal surface free of mother syrup) of raw and after-product sugars (see Section 13.1). The wash syrup is generally collected separately. However, the wash syrup and the green syrup cannot be very cleanly separated. The separation technique was improved by introduction of a pneumatically controlled covering device in a BMA centrifugal [157]. Separation depends primarily on the purity that must be attained in the syrup to maintain the planned crystallization scheme. About 6–8 % (sugar based) of water and steam is required for washing. The water or steam wash must be started at the proper time to prevent unnecessary dissolution of sugar [158]. Before the actual washing process, washing the sugar crystals with slightly undersaturated sugar syrups of a higher temperature than that of the magma (syrup wash) to be centrifuged represents another measure to minimize the dissolution of sugar in the centrifuge.

Double- cone centrifugals are used today for the affination of brown massecuites, a syrup wash replacing the mixing process. This was previously a two-stage process (centrifugation of the mother syrup and remixing of the crystals in a syrup of higher purity than that of the original mother syrup; second centrifugation with washing of the crystals). In these centrifuges, the baskets of two continuous centrifuges are installed above each other (Fig. 9).

12. Preparation of Refined and White Sugar

Refined and white sugar leave the centrifuges with a water content of 0.5–1 %, depending on the temperature of the massecuite and wash water and the *g* attained (maximum 900) during spinning. Surface water content must be reduced to 0.03–0.05 % before further processing. For this purpose, hot-air heaters, cocurrently or countercurrently operating drum dryers (granulators), trickle-bed towers, turbo-tray dryers (→ Drying of Solid Materials, Chap. 2.1.2), or fluidized-bed dryers (→ Drying of Solid Materials) are used. The sugar can be cooled before it is discharged.

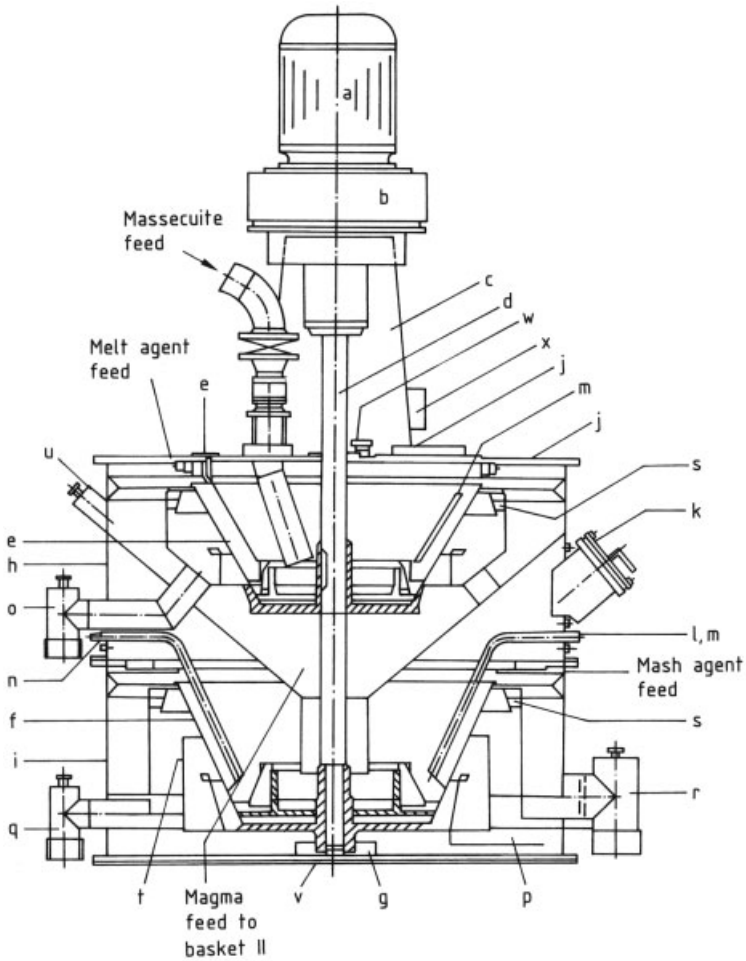

Figure 9. Double continuous centrifuge type SC 1350/1500 ODS
a) Motor; b) Bearing; c) Bridge girder; d) Basket shaft; e) Basket I; f) Basket II; g) Out-of-balance limiter; h) Enclosure top; i) Enclosure substructure; j) Enclosure cover; k) Service cover with inspection glass; l) Water washing system; m) Steam washing system; n) Thick-juice washing system for basket II; o) First molasses runoff; p) Mash runoff (affination runoff); q) Wash syrup runoff; r) Liquor runoff; s) Cascades; t) Runoff separation; u) Sample removal; v) Transport safeguard; w) Automatic oscillation controller; x) Automatic vibration controller

The next step is screening in vibrators or screening machines (→ Screening) in which the sugar crystals, lumps of > 2 mm, and fines of < 0.25 mm (which are redissolved to give remelt/liquor) are separated. In factories, all the preclassified sugar is stored in large- capacity silos without further separation or it is screened into individual fractions (e.g., size fractions 1.2–2 mm = coarse, 0.8–1.2 mm = medium, 0.25–0.8 mm = fine), and sometimes packed for sale.

In some cases, the medium-grain fraction is ground in a mill, and a 0.25–0.75-mm fraction is screened out (ground sugar). Coarse-grained sugar is used mainly in industry and fine-grained sugar is household sugar.

13. Working Schemes of Various Sugar Factories [149, 159]

Figures 10 and 11 present the crystallization and separation stages in beet sugar factories with different production programs. The starting material is thick juice of the same purity, solids con-

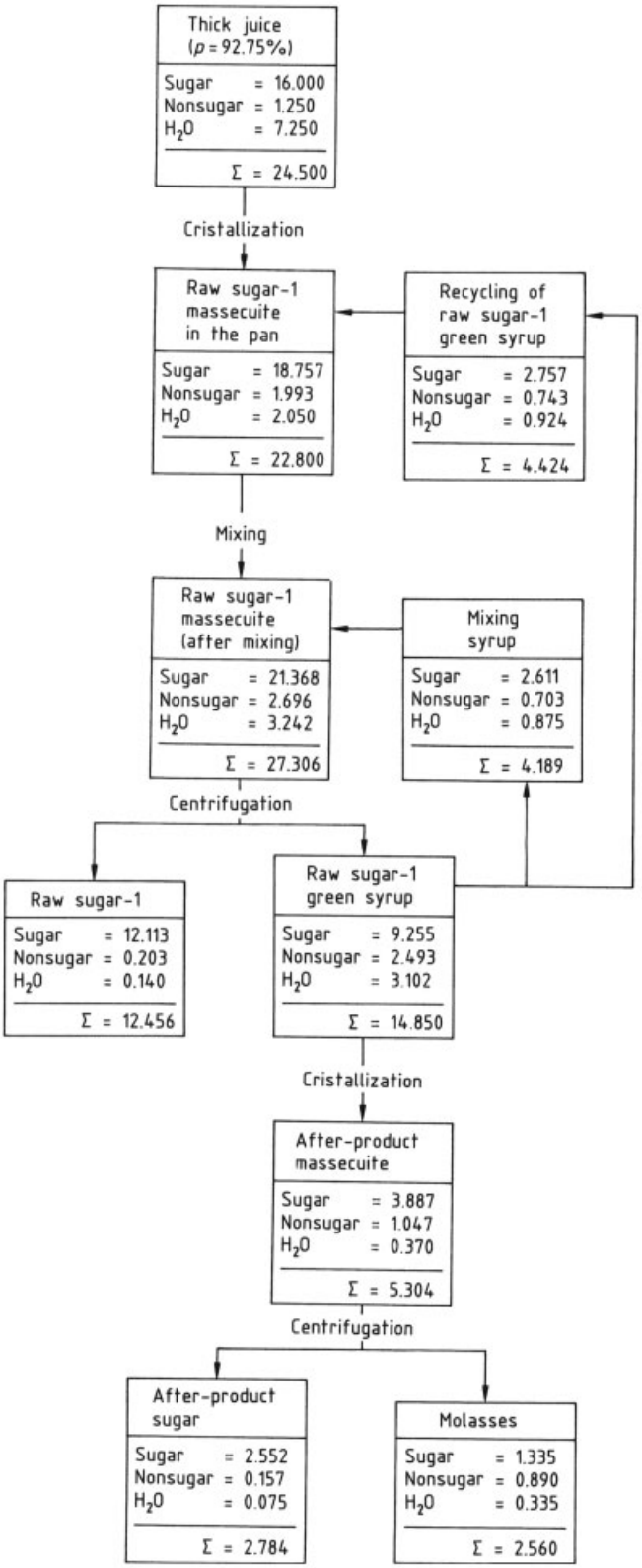

Figure 10. Crystallization schematic of a raw sugar factory with two-product operation (quantities in kilograms per 100 kg of beets)

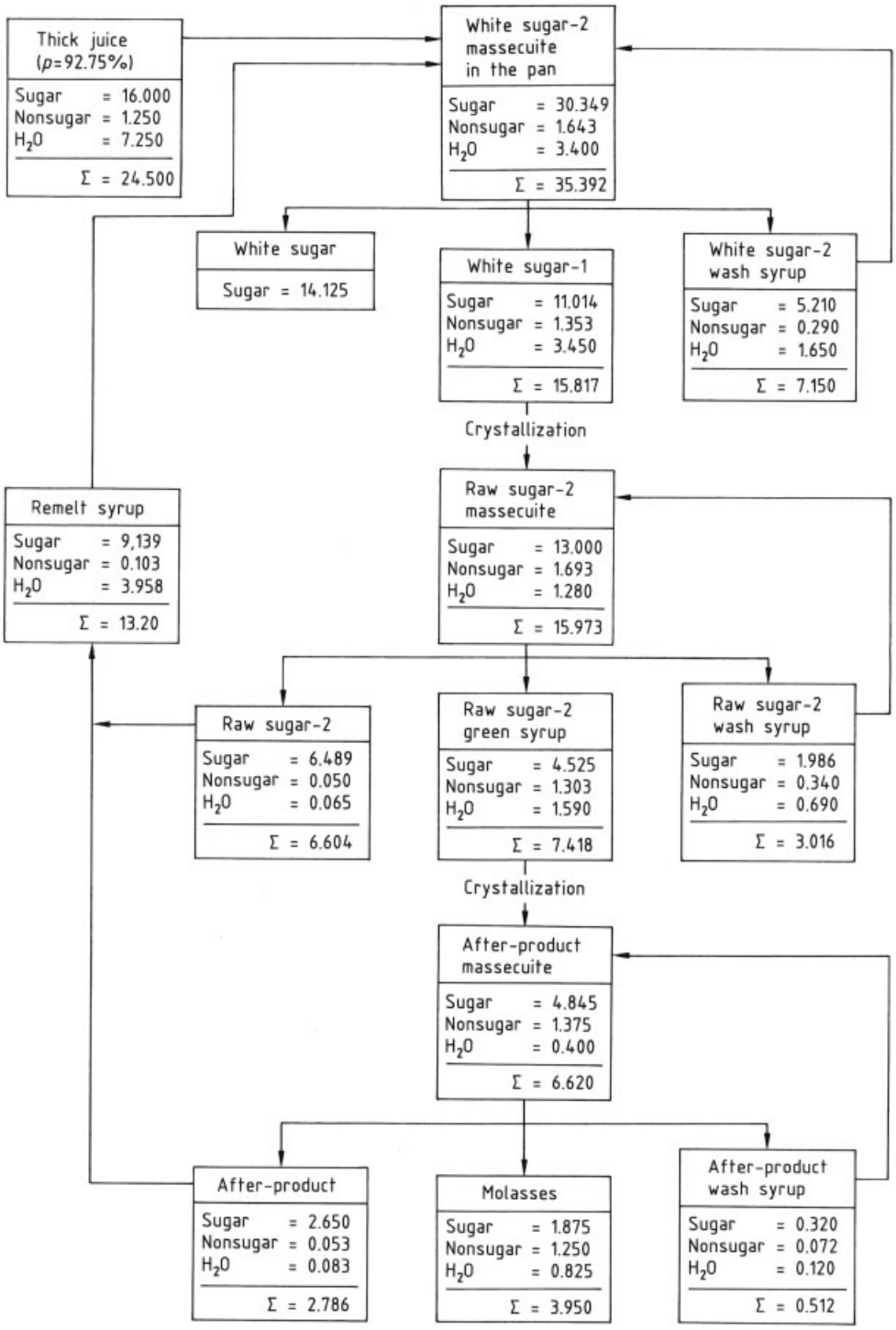

Figure 11. Crystallization schematic of a white sugar factory (quantities in kilograms per 100 kg of beets)

tent, and mass in percent (beet based). For a better comparison, the sugar loss in the molasses was assumed to be the same in all variations. The total mass of massecuites to be processed and the mass of water to be theoretically evaporated are listed in Table 9 for different modes of operation. However, the water mass to be evaporated does not take into account the additional water utilized for crystallization (see Section 10.3) and the energy requirement for heating the media to be boiled.

Table 9. Total mass of massecuites processed and mass of water evaporated in various modes of operation*

Mode of operation	Amount of massecuite, kg/100 kg beets	Theoretical amount of water to be evaporated in the sugarhouse, kg/100 kg beets
Raw sugar factory with two-product operation (see Fig. 10)	≈28	≈7.5
White sugar factory with standard liquor process (see Fig. 11)	≈60	≈13.5
White sugar factory with 50 % refined sugar production	≈70	≈19.0
Sugar factory with 100 % refined sugar production	≈77	≈21.0

* The numbers change when other amounts of sugar or nonsugar enter the sugarhouse with the thick juice.

13.1. Operation of a Raw Sugar Factory

At first, beet sugar factories were all raw sugar factories, as are most cane sugar factories to this day. The raw sugar produced is processed further to white sugar in refineries.

In the refinery (see Section 13.4) raw sugar is again mixed with a mother syrup (p = 85–89, solids content 75 %) to give a magma, centrifuged, and washed with water and steam to obtain consumption (affinated) sugar as the immediate product. Alternatively, the affinated sugar is dissolved to give a remelt syrup and subsequently boiled to refined sugar.

Raw sugar factories are operated in accordance with the two-product scheme (Fig. 10). The raw sugar-1 massecuite in the pan is boiled from thick juice with the return of 15–20 % of raw sugar-1 green syrup (p = 78–80); the runoff syrup is drawn into the vacuum pan only at the end of the boiling process. At discharge, the solids content of the massecuite in the pan is 94 %, and the temperature is ca. 74 °C. The massecuite is cooled to 50–60 °C by stirring for several hours in mixers while more raw sugar-1 green syrup is added. The purity of the mother syrup is reduced from an initial 82–84 % to 78–80 %. In a second crystallization stage, after-product massecuite is made from raw sugar-1 green syrup that is not returned. After-product massecuite is separated into after-product and molasses.

The raw sugar (*raw sugar 1*: pol = 95–98°Z, organic nonsugar = 0.7–1.2 %, inorganic nonsugar = 0.8–1.0 %, water = 1–2 %; *after-product sugar*: pol = 90–92°Z, organic nonsugar = 3–4 %, inorganic nonsugar = 1.5–2.5 %, water = 2–3 %) obtained in a beet raw sugar factory has a light-yellow to dark-brown appearance because of the adhering syrup film. It must be made alkaline (prevention of inversion and formation of colorants) to be storable for longer periods.

Beet raw sugar is not suitable for consumption because of its unpleasant taste and its salt content. However, *cane raw sugar* (see Section 20.6) can be used directly in the production of gingerbread, fruit cake, etc., because of its aromatic flavor and taste components. Raw sugar has no nutritional physiological advantages over white sugar.

13.2. Operation of a White Sugar Factory – Standard Liquor Process

The schematic of crystallization of a white sugar factory operating according to the standard liquor process is shown in Figure 11. The white sugar-2 massecuite in the pan is made from thick juice by the addition of white sugar-2 wash syrup and remelt syrup (explained below) from affinated raw sugar 2. In some cases, the affinated raw or after-product sugar is dissolved directly in thick juice. Boiling a white sugar strike takes about 2–4 h. The crystal yield is 45–55 %, and the solids content at discharge is 90–92 %.

Remelt syrup (solids content = 68–72 %) is a solution obtained by dissolving affinated raw sugar 2 and after-product sugar (see below). It is purified of color and high molecular mass nonsugar substances ($p > 99.5$ %) by activated carbon or ion-exchange resins. This syrup is heated to 88–95 °C and filtered hot. Ceramic or cloth-

covered tube filters precoated with a filter aid (silica gel and cellulose) are used for filtration.

The raw sugar-2 massecuite in the pan is subsequently boiled from white sugar-2 green syrup (p = 87–89 %) and raw sugar-2 wash syrup (p = 85–88 %). In some cases, affinated or unaffinated after-product sugar is used as seed magma. In this case, "seed magma" refers to the after-product sugar suspended in diluted raw sugar-2 green syrup after heating and deaeration to give a massecuite in the crystallizer. Since the after-product sugar has an average crystal size (d' = 150–250 µm, n>3), the crystals introduced in this way grow further to 0.6–1.0 mm. Return of the after-product sugar as "seed magma" for the raw sugar-2 massecuite in the pan gives a raw sugar-2 product that has a more uniform grain size and, thus, can be affinated more easily. Return of unaffinated after-product sugar results in constant recycling of about 10–15 % of the entire nonsugar substances between after- and raw sugar-2 massecuite.

Raw sugar-2 massecuites in the pan are thickened to a solids content of 92–93 %, the boiling time being 3–6 h and the crystal yield 48–53 %. As mentioned above, the raw sugar 2 obtained is added directly to the white sugar boiling via the thick juice or in the form of remelt syrup.

From the green syrup of raw sugar-2 massecuite and the wash syrup of after-product massecuite, after-product massecuite in the pan is obtained, which has a boiling time of 10–20 h because of the low purity of the syrup to be boiled (p = 78–80 %) and the resulting low crystallization rate (ca. 700 mg of sugar per square meter of crystal surface per minute). If ion-exchange processes (see Section 14.2) are used to attain a better sugar yield, the raw sugar-2 green syrup must have a purity of $p < 75$ %. The solids content at discharge of the massecuite is 93.5–95.5 %; the purity in the mother syrup is 62–65 %, or 50–55 % if ion-exchange processes are used. As a result of subsequent cooling crystallization, the purity of the mother syrup, the molasses, decreases to 59–62 % or, if ion-exchange processes are used, 48–53 %. The crystal yield from after-product massecuites is 45–49 %.

After-product affination is shown in the scheme as if it occurs in the same centrifuge as separation of molasses. For the operation of affination, see Section 13.1. This additional process is one reason for returning the unaffinated after-product sugar as seed magma for raw sugar-2 massecuite in the pan (see above).

13.3. White Sugar Factories with a Refinery Scheme

In white sugar factories with the crystallization scheme of a refinery, all the thick juice obtained is first boiled to raw sugar 1. This eliminates any influences on quality exerted by the nature of the thick juice. The raw sugar is washed with only a little water and then (while raw sugar 2 is returned) dissolved to give a remelt syrup, from which the first refined-sugar massecuite is boiled. The resulting mother syrup yields a second refined-sugar massecuite.

The mother syrup of the second refined-sugar massecuite is, together with the wash syrup, processed to white sugar massecuite in the pan. The last two boiling stages correspond to those shown in Figure 11. Thus, it is possible to obtain 80 % of the white sugar production as refined sugar with five crystallization steps.

13.4. Refining

The working scheme of a refinery corresponds largely to that described above, the difference being that the raw material is "foreign raw sugar" instead of thick juice. The processing of *beet raw sugar* (see Section 13.1) involves only purification of the remelt syrup with the help of bone charcoal, activated carbon, and decolorizing resins before the boiling process. The processing of *cane raw sugar* (see Section 20.6) requires a chemical purification step because of the greater content of impurities (especially dextrans). For this purpose, the remelt syrup must be diluted to a solids content of 50–60 % and subjected to juice purification with either phosphoric acid–phosphates and CaO, or CaO and CO_2 (see Section 20.8).

14. Ion - Exchange Processes in the Sugar Industry [160]

Ion-exchange processes can be used to increase sugar yield or improve sugar quality (→ Ion

Exchangers, Chap. 12.1). As early as 1843, HOCHSTETTER proposed that alkali salts of the beet contribute to the formation of molasses. Because of these salts, ca. 11–15 % of the sugar introduced with the beet cannot be obtained in the crystalline state. For molasses formation and exhaustion, see [156]. At the turn of the century, efforts were made to increase sugar yield by exchanging alkali ions for less molasses-forming alkaline-earth ions with zeolites or other inorganic ion exchangers [161]. However, only after the development of ion exchangers based on synthetic resins were ion-exchange processes applied on a large scale.

Ion-exchange processes are used primarily for

1) Softening thin juices by exchanging alkaline-earth ions for alkali ions to prevent incrustation during evaporation
2) Increasing sugar yield by
   a) Replacing alkali ions of the raw sugar-2 green syrup by Mg ions (Quentin process)
   b) Chromatographic separation (→ Ion Exchangers) of molasses into sugar and nonsugar fraction
3) Decolorization of remelt syrups, thick and thin juices
4) Production of liquid sugar (see Section 15.8)

Ion-exchange resins used in the production process must comply with food regulations in different countries.

Processes used to increase sugar yield by removing ionogenic nonsugar substances (complete deionization) have not achieved importance.

When ion exchangers are applied in sucrose-containing solutions, the fact should be noted that strongly acidic cation exchangers in the H^+ form catalyze the hydrolysis of sucrose [162]. The total inversion (heterogeneous, homogeneous), in the case of a thin juice (solids content = 17 %, pol = 15.8°Z, p = 93.0 %) and a contact time of 4 min, is 0.07 % at 14 °C; 0.2 % at 17 °C; and 1.5 % at 23 °C (based on sucrose).

Monosaccharides (glucose, fructose, etc.) and reducing disaccharides are isomerized and, at high temperature, partly degraded to acids by strongly basic anion exchangers in the OH^- form [163].

14.1. Softening of Thin Juice

The ion-exchange process most frequently used in the sugar industry is the thin-juice softening process at 60–85 °C. Since the equivalent ratio (Ca^{2+}, Mg^{2+}) : (K^+, Na^+) in thin juices to be softened is ca. 1 : 15–1 : 25 – compared with, for example, 2 : 1 in water softening – only a small operating capacity (→ Ion Exchangers) is achieved (ca. 70–80 %). For removal of 1 kg of CaO from the juice, 7–11 kg of NaCl must be used as regenerant.

To avoid accumulation of regenerant wastewater, softening plants are sometimes regenerated with diluted raw sugar-2 green syrup containing ca. 1 mol/L alkali ions (Gryllus process [164]) or directly with 1 mol/L sodium hydroxide (Akzo process [165]), the regeneration runoff being returned in front of the first carbonation. Softening of thin juice with weak cation exchangers in the acidic form and regeneration of the exchangers with sulfuric acid [166] has developed to an important wastewater saving process, because the regeneration effluents can be used as a pressing aid in beet extraction and pulp pressing.

14.2. Exchange of Alkali Ions for Alkaline - Earth Ions

The sugar yield can be increased by 0.4 to 0.8 % (beet based) by exchanging 30–50 % of the alkali ions in the syrup for alkaline-earth ions. Since alkaline-earth salts are more hydrated per equivalent in aqueous sugar solutions than the corresponding alkali salts, they bind more water in the mother syrup of the last crystallization than alkali salts. Thus, less solution water is available for the sucrose. The molasses obtained after exchange can achieve a purity of less than 50 %, compared to 60 % for normal alkali molasses.

Quentin Process. The Quentin process [167] was used widely in the beet sugar industry as long as wastewater pollution was of lower importance compared to sugar yield. In this process, alkali ions in raw sugar-2 green syrup are replaced by magnesium ions. The runoff is diluted to a solids content of 68–70 % and passed at 85–90 °C over a strongly acidic cation

exchanger loaded with magnesium ions as counterion. Since only about 2–3 m^3 of runoff can be treated per cubic meter of resin in one operation cycle, the difficulty here is to "sweeten" (i.e., displace water from the resin bed with syrup) and "desweeten" (i.e., displace syrup from the resin bed with water) in such a way that the amounts of water used are just sufficient to dilute the runoff from 78 to 70 % solids content. Direct exchange of alkali ions for magnesium ions is possible in this case because the ionic concentration in the runoff is ca. 1 mol/L and the osmotic pressure in the surrounding liquid is very high. Consequently, the selectivity of the ion exchanger for divalent ions is largely counteracted. The exchanger is regenerated with dilute magnesium chloride solution.

14.3. Separation of Molasses into Sugar and Nonsugar Fractions (Ion Exclusion) [168]

At the same time as large-scale processes were developed for separation of glucose–fructose mixtures (invert sugar, HFCS, → Enzymes, Chap. 5.2.3.2.7), efforts were made to isolate sugar from molasses by using ion exclusion (→ Ion Exchangers). In 1994, several large-scale plants with a separation efficiency of 250 t of molasses per day were in operation (columns up to 3.0 m, separation path up to 32 m, batchwise or continuous, simulated moving-bed process, ion-exchange resin in Na^+ or Ca^{2+} form). The importance of the ion exclusion process has grown rapidly, and in 1997 about 80 % of the molasses in the US beet sugar industry were treated in this way [169].

Molasses diluted to a solids content of 60 % is separated at 80 °C into a sugar (solids content 20–30 %, p = 89–92 %) and a nonsugar fraction (solids content 4–6 %, $p < 20$ %) with water as the eluent. The fractions are evaporated in several stages to a solids content of 70 %. The sugar fraction is crystallized like thick juice, and the nonsugar fraction is processed as cattle feed.

Table 10 shows the extent to which individual classes of substances are divided among sugar and nonsugar fractions. With this process, up to 80 % of the molasses sugar can be obtained as white sugar, about 10 % goes into the nonsugar fraction, and the remaining 10 % is contained in the secondary molasses.

The overall sugar recovery was improved by introduction of the "Coupled Loop" separation [170] or the two-profile FAST technology [171]. Today (2005) equipment for separation of molasses into more than two fractions is being installed in the factories, and mainly betaine is isolated as a second fraction besides sugar [172].

14.4. Decolorization of Remelt Syrups and Thin Juices

Instead of treatment with activated carbon, remelt syrups and thin juices can be treated with strongly basic macroporous anion exchangers in the chloride form at 80–85 °C. The volume of remelt syrup treated in one operating cycle is 200–300 times the resin volume, depending on the purity of the solution to be decolorized. With a specific hourly load of 2–3 volumes of remelt syrup per volume of resin, an initial decolorization of 90 % is achieved; this is reduced to 50 % with increasing running time. The average decolorization attained is 75–80 %. Neutral or alkaline (pH 12–13) 10 % sodium chloride solutions are used to regenerate the ion exchanger. About 100–150 kg of NaCl is required per cubic meter of resin [173].

14.5. Elimination of Ionogenic Nonsugar Substances

Acidic Deionizing Process [174]. Thin juice (solids content = 15–17 %) cooled to at least 14 °C, or runoff or molasses diluted to 30–40 % solids content, is passed first over a strongly acidic cation exchanger in the H^+ form and then over a weakly basic anion exchanger in the OH^- form. The cations, amino acids, and betaine are adsorbed on the cation exchanger, and nitrogen-free anions and pyrrolidonecarboxylic acid on the anion exchanger.

To prevent ion slippage as far as possible, two pairs of ion exchangers are connected in series.

The acidic deionization process eliminates ca. 85 % of all nonsugar substances. Thin juices with a purity of 97–99 % are obtained. Sugar yield can be increased to a maximum of 1.2–1.4 % (beet based).

Table 10. Division of individual classes of substances between sugar and nonsugar fractions

Components	Molasses, %	Portion of Sugar fraction, %	Nonsugar fraction, %	Percentage going into nonsugar fraction
Solids	100.00	59.00	41.00	41.00
Sugar	60.00	54.50	5.50	9.20
Glucose	0.02	0.01	0.01	
Fructose	0.06	0.02	0.04	
Raffinose	2.50	0.45	2.05	82.00
Ash	12.50	1.10	11.40	92.00
Total nitrogen	2.60	0.30	2.30	88.00
Betaine	6.80	1.40	5.40	79.00
α-Amino-N	0.60	0.10	0.50	83.00

Apart from acidic deionization, other partial and complete deionization processes have been developed to prevent the hydrolysis of sucrose. These include the exchange of alkali ions for NH_4^+ ions plus the exchange of anions in the juice for HCO_3^-–CO_3^{2-} ions, and subsequent distillation of $(NH_4)_2CO_3$; or the exchange of anions for OH^- ions with a strongly basic exchanger, followed by elimination of alkali ions with a weakly acidic cation exchanger.

14.6. Production of Liquid Sugar [175]

Crystalline sugar is dissolved to give a syrup with a solids content of 50–60 %. The syrup is filtered through a depth filter (→ Filtration, Chap. 4), decolorized at 80 °C, and then deionized completely at 20–40 °C. For the production of invert sugar syrups, an inversion resin (cation exchanger in the H^+ form) is integrated into the deionization path. The desired degree of inversion is attained by variation of temperature and residence time.

The syrup leaving the ion-exchange facility is finally evaporated in several stages to a solids content of 65–75 %. Facilities are sized to require regeneration only after 120–150 h.

15. Production of Special Types of Sugar

15.1. Cube Sugar

The invention of cube sugar dates back to 1843. Today two different processes are applied in the production of cube sugar: the poured cube process (Adant process) and the pressed cube process.

Poured Cube Process. In the Adant process, the cube massecuite with a temperature of 98–100 °C is filled into the baskets of special centrifuges in which it solidifies on cooling (12 h) to give slabs. The baskets are then placed in centrifugals; the adhering green syrup is centrifuged off; the sugar slabs are washed with wash syrup (i.e., a saturated pure sugar solution) and centrifuged dry. Slabs containing 2–3 % moisture are dried in drying chambers or channels and then cut into strips with saw blades in special machines. Subsequently, the strips are broken into cubes in clipping machines and usually packed loose. Since cube yield is poor (ca. 80 % based on dried slabs) and production is both cumbersome and labor intensive (ca. 500 working minutes per tonne of cubes in large plants), poured cubes are seldom produced in industrialized countries.

Pressed (Molded) Cube Process. In the pressed (molded) cube process, crystal sugar with a certain grain size (e.g., 0.4–0.8 mm) and a relative density of 0.82–0.84 is moistened in a mixing screw with 1.5–1.8 % of water. It is subsequently pressed into cubes by means of presses (Machines Chambon, Paris [176]) or by vibration in vibration forming machines (SSA-Process) [177]. The individual cubes are dried in continuous drying ovens with hot air or on a steel band by high-frequency heating. They are then cooled and packed immediately in an arranged manner. The hardness of the cubes produced and, thus, their mechanical stability as well as rate of breakdown (dissolution) are substantially influenced by parameters such as grain-size distribution of the sugar used, water content in the mixing path, forming pressure, and drying rate [178].

15.2. Sugarloafs and Sugar Cones

Sugar cones (sugarloafs) have only regional and seasonal importance (burnt punch). Like cube sugar, they can be made by using the pouring or pressing process with a loaf or cone form instead of slabs. In the past, pressing was performed exclusively manually. Fully mechanical plants were developed in the 1960s [179].

15.3. Rock Candy

Rock candy is available as string candy or string-free candy. *String candy* is made according to the Stoven method. In this process, an especially pure sugar solution with a solids content of 75–78 % is discharged at 95 °C into tanks containing strings stretched across them. On cooling, the sugar crystallizes on the strings in large crystals. Cooling must occur very slowly (8–9 d) according to a definite temperature program. The "massecuite" should not be shaken during cooling. *String-free rock candy* can be made by crystallization in movement (e.g., in Wulff–Bock crystallizers) or in other continuous processes (e.g., where seed crystals are placed on screens) [180]. In all these processes, exact maintenance of temperature, flow rate, concentration, pH of the sugar solution, etc., is important. Brown rock candy is obtained by adding caramel sugar (see Section 15.9) to the sugar solution.

15.4. Icing Sugar

Icing sugar is made by grinding sugar in impact or hammer mills (→ Size Reduction). The grain-size distribution follows the Rosin–Rammler–Sperling distribution ($d' = 40$–60 µm, $n = 1.8$–2.5). Because of the danger of sugar dust explosion [181], icing sugar may be produced only if special safety measures are taken (separate rooms with light walls, ceiling, etc.). Icing sugar must be conditioned carefully (heated, dried to a certain moisture content, and cooled) before packing so that it does not become hard during storage. No general agreement exists about optimal conditioning requirements [182].

To prevent hardening of icing sugar, the addition of up to 5 % starch or 1.5 % anticaking agents, (e.g., certain calcium phosphates, magnesium carbonate, amorphous SiO_2) is permitted in many countries [183].

15.5. Instant Sugar

Instant sugar dissolves very quickly and is produced from finely ground icing sugar by agglomerating it in a fluidized bed with the addition of superheated steam [184]. The resulting granulate is fine pored and unevenly formed, and has a grain size up to 5 mm and a bulk density of 0.4 kg/L. In comparison, the bulk density of icing sugar is 0.6 kg/L and that of crystal sugar 0.8–0.9 kg/L, depending on grain size.

15.6. Preserving (Jelly) Sugar

Preserving (jelly) sugar is a preparation of crystal sugar, dry pectin, tartaric acid, and/or citric acid: 0.8–1.3 % pectin, depending on the jelling strength of pectin (→ Polysaccharides, Chap. 3.4.1), and 0.4–0.7 % tartaric acid or 0.6–0.9 % citric acid are added to the sugar. Pectin with a degree of esterification of 48–50 % dissolves completely in a solution with a solids content of > 50 % and forms a stable gel [185]. Amidated pectins (→ Polysaccharides, Chap. 3.4.1) are employed for the production of preserving (jelly) sugar, which contains sorbic acid as a preservative and is used in the proportions 1 part of preserving sugar to 2 parts of fruit.

15.7. Brown Sugar (Soft Sugar)

Brown sugar is the general term for crystal sugar that contains invert sugar (muscovado, brown candy, bastard sugar). This crystal sugar has color flavoring and taste components dissolved mainly in its syrup film (water content ≤ 3.5 %, total sugar content > 93 %, color in solution up to 30 000 ICUMSA units). Soft sugar is produced by directly boiling special sugar syrups, and the sugar centrifuged off is not washed. Alternatively, it can be made by mixing white crystal sugar with prepared sugar syrups [186].

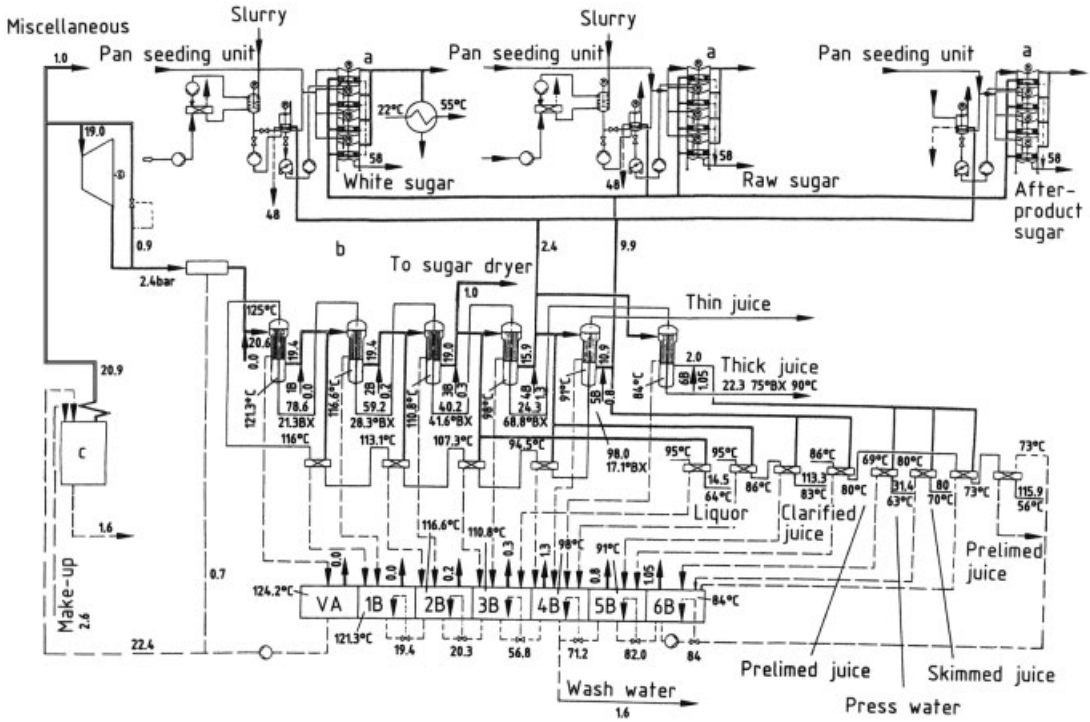

Figure 12. Beet sugar factory
a) Continuous evaporation–crystallization tower; b) Evaporator station; c) Boiler (6 MPa; 515 °C)
VA = steam boiler- condenser; 1 B = vapor of the 1st step of evaporation.

15.8. Sugar Solutions

The quality, composition, and labeling of sugar solutions are regulated in the EC [187] and, in part, in the United States [188].

Accordingly, "liquid sugar" may not contain more than 3 %, "invert liquid sugar" not more than 50 %, and invert sugar syrup must contain more than 50 %, invert sugar.

Since liquid products can be transported and distributed easily, and because of their uniform quality from a microbiological viewpoint, sugar solutions are being used increasingly in manufacturing. Their share of the market, including high-fructose corn syrup (HFCS) (→ Glucose and Glucose-Containing Syrups, Chap. 6.2), is ca. 45 % of total sugar sales in the United States and 10–15 % in other industrialized countries.

In the production of sugar solutions, white sugar is dissolved in hot, low-salt drinking water; the syrup is filtered and sterilized by high-temperature, short-time heat treatment (8 s at 138 °C). The syrup is then cooled to 15–20 °C and stored almost sterile (plate counts < 200/10 g dry substance) before transport. Any contamination with microorganisms in the air must be avoided.

Solutions containing invert sugar are prepared mainly by the ion-exchange process (see Section 14.5). However, batchwise or continuous *acid hydrolysis* is also used [189]. On heating to 85 °C and acidification with sulfuric acid to pH 2.5 (ca. 140 g of sulfuric acid per 100 kg of sugar solution), an 80 % sucrose solution undergoes inversion to ca. 70 % in ca. 90 min. To prevent further inversion, the pH must be increased subsequently with sodium carbonate (pH 4–5). If no further discoloration is to take place, the inverted solution should be cooled quickly. *Enzymatic hydrolysis* of concentrated sugar solutions with the help of immobilized invertases at their pH optimum of 4.5 (β-fructosidase, E.C. 3.2.1.26) is technically possible, but it is economically inferior to the other processes. Inverted sugar solutions are supplied with a solids

content up to 75 %. The degree of inversion must be ca. 0.65 to prevent crystallization of glucose or sucrose.

15.9. Burnt Sugar and Caramel Colors (Caramels)

Burnt sugar is made by heating (melting) sugar or sugar solutions. It has a typical caramel taste, which deteriorates with increasing color of the product. Burnt sugar is produced with a color of up to 18 000 EBC (European brewery colour) units, the caramel taste determining the product's value [190]. *Caramel colors* (caramels) are made from sugar or other edible types of sugar by adding certain permitted burning additives [191]. Caramels can be divided into four groups: caustic, alkali sulfite, ammoniac, and ammonium sulfite [192]. Different types of caramel are required to color certain foods (e.g., spirits, vinegar, malt beer, and cola beverages).

16. Storage of White Sugar

Since white sugar factories produce sugar only during the working season (three to six months per year), but sugar is consumed year-round, large amounts of the sugar produced must be stored. Consumption sugar is packed in 100-kg jute sacks or in 50-kg (25-kg) paper sacks and stored. Household sugar is packed in the factory directly in small packages of 0.5–5 kg.

Sugar should not contain dust or lumps before storage. It must be dried to 0.03–0.04 % residual moisture and cooled to 20–25 °C. Storage in high-capacity silos is becoming increasingly common [193] because it enables the delivery of bulk shipments (rail/trucks) to the manufacturing industry. In industrialized countries, industry consumes approximately two-thirds of the total sugar sold.

White sugar silos are usually vertical cylinders made of steel, prestressed concrete, or sheet steel and provided with a heatable insulated jacket [194]. Most of the silos have a capacity of 10 000–40 000 t, a diameter of 20–37 m, and a sugar head up to 40 m high, but very large silos with capacities of 80 000 t, diameters of 51 m, and a sugar head up to 60 m high were erected in 2000 [195]. Sugar silos are filled and emptied according to various principles; e.g., they are filled from the bottom upward from a central pipe with a screw system and emptied from the top downward (N. Weibull, Malmö; ABR, Wiedemann, Sarstedt). In other silos, sugar falls through the entire free height of the silo. Measurements have shown that fears of dust explosions in this mode of operation (due to sugar dust concentrations) are unjustified [181, 196]. The sugar stored in a silo of this type can be circulated constantly by removing it from the bottom and transporting it to the top. This measure, together with proper air conditioning ($T < 25$ °C, R.H. < 60 %) of the silo, prevents sugar from caking and/or binding during storage. For air conditioning and especially for adjusting the humidity, adsorption and refrigeration processes are used.

During storage, sugar releases not only moisture adhering as a syrupy film to the crystal surface, but also part of the bound moisture (in the mother syrup enclosed in crystal defects) [197].

High-capacity silos constructed like sugar silos are used for storage of pelleted dry pulp (see Section 17.2) [198].

17. Auxiliary Facilities in a Sugar Factory

17.1. Steam and Energy Balance

The steam requirement of a sugar factory is met by the production of superheated steam in boilers of various types. High-pressure boilers (6.7–10 MPa) are being used increasingly. All or the largest part of the superheated steam (<537 °C) is expanded to 0.25–0.30 MPa via turbogenerators (counterpressure machines). The electrical energy thus produced is largely consumed in the sugar factory itself, and the surplus is transferred to the public grid.

Unlike the process steam requirement, the electrical energy requirement has increased in the last years. Consequently, gas turbines are now set up in front of the boiler plant to meet electrical energy requirements.

The steam flow, vapor distribution, and condensate scheme for a white sugar factory are depicted in Figure 12. The entire condensate from the first effect of the evaporator station (b) (steam condensate from the turbines) is used

as boiler feedwater. In the case of low-pressure boilers (<4 MPa), condensate from the second effect of the evaporator station is used to cover losses. However, since this condensate can contain volatile organic substances [120] and, possibly, traces of sugar, this portion of the water is replaced by treated water in the case of high-pressure boilers.

Exploitation of the energy of boiler exit gases for drying cossettes in the beet sugar factory did not gain acceptance for decades. Only with the introduction of oil or gas burners was boiler exit gas at 150–200 °C increasingly used as "additional air" for drying pulp. In this mode of operation, the oil or gas is burned with a slight excess of air, and the 1600 °C hot burner gases are cooled subsequently with boiler exit gas to the required working temperature of the drying process (600–700 °C).

Herewith, 93–94 % of the usable amount of heat introduced with the fuel can be utilized, compared to 80–86 % in the normal mode of operation. For detailed energy balances of the entire factory, see [126].

In contrast to beet sugar factories, which must use fossil fuels for steam generation, the entire energy requirement of the *cane sugar industry* is met by burning the bagasse in specially constructed boilers (see Section 20.5.3). Since the cane industry has taken over time-tested heat-flow schemes for the production process from the beet sugar industry, a surplus of bagasse is now available that can be used elsewhere.

17.2. Drying, Pelletizing, and Storing Pulp

Drying. Good mechanical pressing of the extracted pulp is a precondition for efficient pulp drying (see Section 7.5). If 5.6 kg of dry pulp (solids content 90 %) is obtained from 100 kg of beets, the amounts of water listed in Table 11 must be evaporated during drying.

The amount of heat required for evaporation of 1 kg of water depends on the temperature gradient between the entrance and exit of the drum. The higher the combustion gas temperature at the entrance of the drum, the greater is the thermal efficiency of the plant. However, excessively high entrance temperatures lead to strong scaling of the internals and burning of the material to be dried [formation of volatile organic acids, total amount expressed as total organic carbon (TOC) and CO]. For entrance temperatures of 600 °C, ca. 3500 kJ per kilogram of water to be evaporated must be employed. In the drums, 150–180 kg of water evaporates per hour per cubic meter of drum contents [199]. Cocurrently operated rotary drum dryers with preceding firing are used for drying pressed pulp (→ Drying of Solid Materials). For measures for the efficient use of energy in cossette drying and optimal control of cossette drying plants, see [126, 200].

For predrying pressed pulp, low-temperature dryers have been employed to make use of the lower-energy heat that was previously released to the environment. Most of the four-rack belt dryers operate with hot air (≈ 60 °C), which is heated with the waste heat carriers' hot water and condensate (< 70 °C) [201].

Apart from hot-air dryers, steam dryers are now being used to dry pressed pulp. Drying is achieved with superheated steam:

1) With steam of 130 °C, which expands to 102–103 °C (≈0.1 MPa) by the uptake of water [202] or
2) With steam of 260 °C (ca. 2.6 MPa), which expands to 148 °C (0.37 MPa) [203]

The advantages of steam dryers are energy saving and the closed design, which prevents escape of gaseous emissions (odors). Besides high investment, improvable performance and considerable maintenance costs have to be taken into account as disadvantages [204].

Processes that eliminate water from pressed pulp not by drying, but by multiple-effect evaporation with the help of an entrainer (e.g., edible oil, paraffin oil) have not yet gained acceptance (Carver–Greenfield process).

Storage and Transportation. Dried pulp has a bulk density of 200–300 kg/m^3. Thus, storage and transportation of this material are very space consuming. Also, in this form, it is not flowable, is poorly miscible (with other components in the production of mixed feed), and only poorly grindable. For these reasons, dried pulp is for the most part pelletized (8–10-mm diameter, 20–30-mm length) [205] (→ Size Enlargement B, D). This increases the bulk density to 650–700 kg/m^3, and the product can

Table 11. Amount of water (in kilograms) to be evaporated in the drying of beet pulp as a function of solids content of pressed pulp

	Solids content of pressed pulp, %										
	16	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**	**36**
Based on 100 kg of beets	25.8	22.4	19.6	17.3	15.4	13.8	12.4	11.2	10.2	9.2	8.4
Based on 100 kg of dried pulp	462	400	350	309	275	246	221	200	181	165	150

be handled and ground easily. Dried-pulp pellets are stored in warehouses and, more recently, in high-capacity silos of up to 30 000 t [198]. In these silos, better conditioning of the entire space is possible and molding is virtually impossible.

17.3. Lime–Carbon Dioxide

For juice purification, 0.8–2.0 % (beet based) of CaO and 4.0 % (beet based) of carbonation gas (carbon dioxide content of 42 vol %) are required. These materials are produced in the sugar factory itself by burning limestone (2–5 %, beet based) in high-performance shaft kilns [206]. The quicklime is hydrated with water to form milk of lime (180–220 g of CaO per liter of water) and freed of oversize material by means of vibrating screens, rake classifiers, and hydrocyclones. To guarantee exact dosing, the density of the milk of lime is kept constant. Since the lime added during juice purification is fully retained in the carbonation sludge (see page 26) and is largely present as calcium carbonate, efforts have been made to completely or partially reburn the carbonation sludge.

On addition of 1.5 % of CaO (beet based), the following average composition of the carbonation sludge (solids based) can be expected: 48–50 % CaO, 0.9–1.2 % MgO, 0.3 % K_2O and Na_2O, 1.7–2.1 % PO_4^{3-}–SO_4^{2-}, 23–26 % CO_2, 2–5 % SiO_2 (sand, clay), and 10–15 % organic substances (protein, pectins, polysaccharides, citric, oxalic, and malic acids, etc.). Depending on the conditions of recalcination (temperature and residence time), decomposition of the clay or sandlike components to soluble calcium aluminates and calcium silicates can occur. This greatly reduces the quality of the reburnt lime [207]. After reburning once, 50–55 % of the dry sludge weight is recovered, and after repeated reburning, a strong enrichment of MgO, calcium aluminate, silicate, calcium phosphate, and calcium sulfate occurs so that part of the carbonation sludge must be discarded. In the United States, a rotary kiln and a circular multiple hearth furnace have been used for many years for reburning carbonation sludge [208]. Attempts have been made on a pilot scale to recalcine previously granulated carbonation lime cake [207] with a fluidized-bed system and with a high-velocity reaction chamber [209], which can process the carbonation lime in its primary particle-size distribution ($d' < 10$ µm).

Carbonation Lime Cake as Lime Fertilizer. In the past, naturally dried carbonation sludge was spread manually on fields by farmers to improve the soil structure of acidic soil. Today, only carbonation lime cake that can be spread on the ground by machines is sold. For isolation, see page 26. The composition is given in Table 12 [210].

Table 12. Fertilizer analysis of carbonation lime cake with 60–70 % solids content* (in kg/1000 kg) [210]

Nutrients	Average	Variation
Basic components calculated as $CaCO_3$	500	460–540
Organic substances	100	80–150 (180)
Total N	4	3–6
P_2O_5	9	6–18
K_2O	1	0.2–2
MgO	10	5–13 (22)

* Other constituents ca 1 kg sulfur, 1 kg sodium, and 0.9 kg boron.

18. Desugarization of Molasses

To reduce losses of sugar in molasses, chemical processes are used that are based on the fact that sucrose forms poorly soluble compounds (saccharates) with alkaline-earth hydroxides under certain conditions. The structure of these saccharates is still not understood clearly. Of the methods used (Strontian process, baryte process, Steffen process) [211], only the *Steffen process* is still important. It is used on a large scale in the United States, the CIS, Turkey, and Iran.

Steffen Process. This process can be operated batchwise or continuously [212], at ca.

10 °C (cold) or 65–80 °C (hot). If operated carefully, ca. 40 kg of white sugar can be produced from 80 kg of molasses solids.

The molasses is diluted (e.g., to a solids content of 10–15 %) and cooled to 10–14 °C. At this temperature, very finely ground CaO (<100 µm) is added slowly, with continuous stirring, until 0.7–0.9 % of free CaO is present in the solution. The precipitated tricalcium saccharate ($C_{12}H_{22}O_{11} \cdot 3\ CaO$) is filtered off with filter presses or belt filters, washed with water, and suspended in water. The filtrate containing the nonsugar substances in molasses (Steffens waste) can be used for isolation of glutamic acid, betaine, raffinose, etc., or evaporated to give vinasse (see Section 22.3). During the beet processing season, the tricalcium saccharate suspension is used instead of part of the lime milk as the precipitating agent for juice purification. Outside the working season, the tricalcium saccharate suspension is heated and calcium hydroxide is precipitated as calcium carbonate by the introduction of CO_2. The sugar solution thus obtained is subsequently evaporated and boiled.

Apart from the Steffen process, chromatographic processes are being used on a large scale for the desugarization of molasses (see Section 14.3).

Membrane and electrodialysis processes, also with ion-exchange membranes, have not gained acceptance because of the short service life of the membranes used [213].

Isolation of Glutamic Acid, Betaine, Raffinose, and Other Substances. Of all the nonsugar substances present in sugar beet, the substances mentioned above are isolated on an industrial scale. *Glutamic acid* is isolated from the spent wash from molasses desugarization processes, especially from Steffen waste (after hydrolysis of L-pyrrolidonecarboxylic acid); from the regeneration wash from deionization processes; or from the nonsugar fraction in the chromatographic separation of molasses (see Section 14.3). Processes for isolating amino acids are described in [214].

Like glutamic acid, *betaine* [215] can also be isolated from molasses on an industrial scale, today preferentially by chromatographic separation [172]. *Raffinose* [216], *inosite* [216], *galactinol,* and *kestoses* can be isolated from molasses by chromatographic separation as well.

19. Sugar Yield, Energy Requirements, Processing Aids, Water, and Working Time

The *sugar yield* is very important for the profitability of a sugar factory. In a modern mode of operation (continuous extraction process, good juice purification and crystallization) with exact operation control, the sugar losses occurring in a white sugar factory are listed in Table 13.

Table 13. Sugar losses in a white sugar factory (kg sugar per 100 kg beets)

Type of loss	Portion of sugar lost	Sugar contained in side products
Respiration loss during storage (average of 8 d) between harvesting and processing	0.12–0.20	
Losses during fluming and washing	0.10–0.25	
Extraction losses in pressed pulp		0.20–0.30
Unaccounted losses*	0.15–0.40	
Losses in molasses		1.6–2.1
Average total losses	0.6	2.1

* Sucrose losses due to bacterial activity, sucrose destruction during juice purification, evaporation, and boiling.

If the beets have a sugar content of 16.0–17.5 %, 13.3–14.6 kg of white sugar is obtained per 100 kg of beets. Since the amount of accounted and unaccounted losses does not vary greatly, the sugar yield depends primarily on the sugar content of the beets.

The sugar yield can be increased by 0.3–1.5 kg per 100 kg of beets by the use of ion-exchange or molasses desugarization processes.

The *primary energy* required for the boiler house (i.e., only for sugar production) is 20–25 kW · h per 100 kg of beets and for cossette drying, 2–3 kW · h per 100 kg of beets. Since the 1970s, the total primary energy consumption in the European sugar industry has decreased by more than 50 %.

Limestone consumption is about 2–5 % (beet based), and the coke requirement for burning limestone is 7–10 kg per 100 kg of limestone.

Other processing aids employed include acids (sulfuric, hydrochloric, SO_2), hardening agents (calcium sulfate, calcium chloride), disinfectants (formalin, hydrogen peroxide), antifoaming agents, antiscaling agents, activated carbon, and filter aids.

Theoretically, a sugar factory requires no additional water supply from ground or surface water because with the sugar beets, 0.75 m^3 of water is fed per tonne of beets into the production process as condensate and ca. 0.35–0.45 m^3 of water is required per tonne of beets for supplementing the water cycles (0.25–0.3 m^3 per tonne of beets for the flume water cycle, 0.1–0.15 m^3 per tonne of beets for evaporation water losses in the cooling water cycle). Wastewater from the flume water cycle and surplus condensate are treated biologically before they enter the wastewater receiving stream.

In practice, however, at the start of the working season the water cycles must be filled with ground or surface water.

To process 1 t of beets, including delivery and unloading of beets and storage of the sugar in silos, 10 working minutes are required during the working season, depending on the size, mechanization, and automation of the factory. On the other hand, the total work requirement per year amounts to 30–40 working minutes, including the time required for packing and dispatching, plant maintenance and repair, and general administration.

20. Cane Sugar

Production and consumption figures for cane sugar are provided in Tables 1 (Section 2.1) and 18 (Chap. 23).

20.1. Sugarcane Cultivation

Sugarcane is a giant tropical grass; *Saccharum officinarum,* the cane of commerce, is a complex hybrid of several *Saccharum* species. Constant development of new varieties through crossbreeding and selection is essential to maintain disease resistance and healthy crops with high sucrose content. Seed crossings (bi- or poly-) produce multiple interspecies varieties; selection takes place over eight to twelve years. Somoclonal variation in sugarcane has prevented replication by tissue culture; the varieties selected are increased by vegetative reproduction. The same procedure is employed in cane planting: a section of stalk with one or more buds (each node has at least one bud) is planted; roots develop; then a shoot appears. Cane is planted, manually or mechanically, in single or multiple rows, in furrows, and covered with soil. Planting season and fertilizer, pesticide, and growth regulator applications vary tremendously from the subtropical areas (Louisiana, Pakistan, Morocco) to the optimum growing tropical areas (North Queensland, Colombia, Cuba). Generally, replanting is not necessary after each harvest: buds on the plant base and roots remaining will sprout again to produce another crop, called *ratoon* or *stubble*; this ratooning is repeated until the yearly decline in yield (successive ratoons yield lower cane tonnage) is no longer economical. Ratoon crops vary from none in Hawaii, where push-rake harvesters can harvest roots with the stalks, to eight to ten in optimum regions: two to four ratoon crops are customary.

Cane pests include insects (borers, leaf hoppers, grubs, aphids) and vertebrates (rats, pigs, and elephants), which eat and trample cane. Other *diseases,* now often controlled by breeding, include mosaic, rust, smut, and ratoon stunting disease. The most common *weeds* are other grasses and therefore difficult to control with herbicides.

Sugarcane is the most efficient collector of solar energy in the plant kingdom, converting 2 % of available solar energy into chemical bonds of stored compounds, chief among them sucrose. Yield in tonnes of cane per hectare varies from 55–60 t in poor growing areas to more than 200 t for cane grown for 18–24 months in optimum areas (e.g., Hawaii). The quantity of sugar produced per hectare varies from ca. 5.0 t (Ethiopia) to ca. 26 t (Campos, Brazil).

Harvest season is generally during the cooler, drier part of the year, varying from three months (October–December) in Louisiana, to the first half of the year in most Northern Hemisphere tropics and the second half in most Southern Hemisphere tropics, to year-round in Hawaii, Colombia, and Peru.

20.2. Composition of Sugarcane

Sugarcane stalks, the portion of the plant desirable for milling, contain 73–76 wt % water and 24–27 wt % solids (10–16 wt % soluble solids, 11–16 wt % dry fiber); the inclusion of "trash"

(trash is defined as leaves, tops, roots, weeds, soil, and other extraneous matter) will increase the total solids level but decrease the sugar content of the mass, and thereby of the cane juice. The sugar (sucrose) content per tonne of cane varies from 75 kg in marginal growing areas or with overratooned or diseased cane, to more than 140 kg in very good quality cane.

The soluble solids content (in percent) in the juice is listed as follows [7]:

Sugars	75–92
Sucrose	70–88
Glucose	2–4
Fructose	2–4
Salts	3.0–4.5
of inorganic acids	1.5–4.5
of organic acids	1.0–3.0
Organic acids	1.5–5.5
Carboxylic acids	1.1–3.0
Amino acids	0.5–2.5
Other organic nonsugars	
Protein	0.5–0.6
Starch	0.001–0.050
Gums	0.30–0.60
Waxes, fats, phosphatides	0.05–0.15
Other	3.0–5.0

The major amino acids are aspartic and glutamic acids; the major inorganic salts are chlorides and sulfates of potassium, calcium, and magnesium. The level of invert sugar decreases as sugarcane matures, but increases at the expense of sucrose concentration after cane is cut (or wounded), with a concomitant decrease in pH.

Concentration of nonsugars relative to sucrose, or juice purity (see Section 5), affects the potential isolation of sucrose: in general, higher nonsugar levels or lower purity make sucrose recovery difficult and decrease yield.

Sugarcane fiber makes up 10–17 wt % of millable sugarcane. Called bagasse after the milling stage, cane fiber has three components: (1) inner stalk pith (containing most of the juice); (2) fiber in the outer stalk and small fiber bundles in pith; and (3) nonfibrous, wax-coated epidermis, in the ratio 6 : 13 : 1. Bagasse is composed of cellulose (40–60 %), hemicelluloses, especially L-xylan (20–30 %), and lignin (15–21 %) [14].

20.3. Deterioration of Sugarcane

The composition of sugarcane after it is cut (harvested) or wounded by freeze or pest damage can change drastically and rapidly. "Stale cane is an anathema to the whole industry: growers lose tonnage and processors lose sugar" [217]. Deterioration is both chemical and microbial: chemical deterioration includes hydrolysis of sucrose by invertases; the invert sugar formed continues to decompose partly during juice purification to colorants and acids, which lower the pH of the juice and increase inversion. Degradation (and sugar loss) by microorganisms invading open cane tissue includes *yeast* infection (causing production of ethanol, lactic, and acetic acids) and *bacterial* infection, especially from *Leuconostoc mesenteroides,* which converts sucrose to dextrans (long-chain polymers of glucose) and organic acids [218, 219]. If present in juice at concentrations >1000 mg/kg, dextrans can increase viscosity of juices and syrups; hinder clarification, filtration, crystallization, and centrifugation; and lower sugar yield [7, 14, 220]. The deterioration rate increases with temperature, mud on cane, and time. Deterioration, which varies in rate with variety, can best be controlled by minimizing cut-to-crush time (see Section 20.4) and processing frozen cane as rapidly as possible. Bactericide application is not practical in fields or cane yards; treatment in the factory, to control further dextran formation, with the few approved biocides has proved too expensive and impractical [220]. Clearing with steam or very hot water spraying of equipment is the recommended factory treatment.

20.4. Harvesting and Delivery

Harvesting. More than 60 % of the world's sugarcane is harvested *by hand* in the tropics, where labor is available and employment is desirable. Cane knives range from long machetes to shorter-handled Australian and Brazilian knives with hand guards. Cane leaves and tops – which contain little sugar, add weight to transport, hinder cane cutters, and wear down mill rolls – are removed first by burning the cane field and then by hand or mechanical harvesters. Cane stalks are sufficiently high in moisture so that controlled and rapid burns (fire in a 50-ha field will be complete in 3 min) incinerate only the leaves, tops, and trash. In Australia and Hawaii, cane is harvested without burning, to

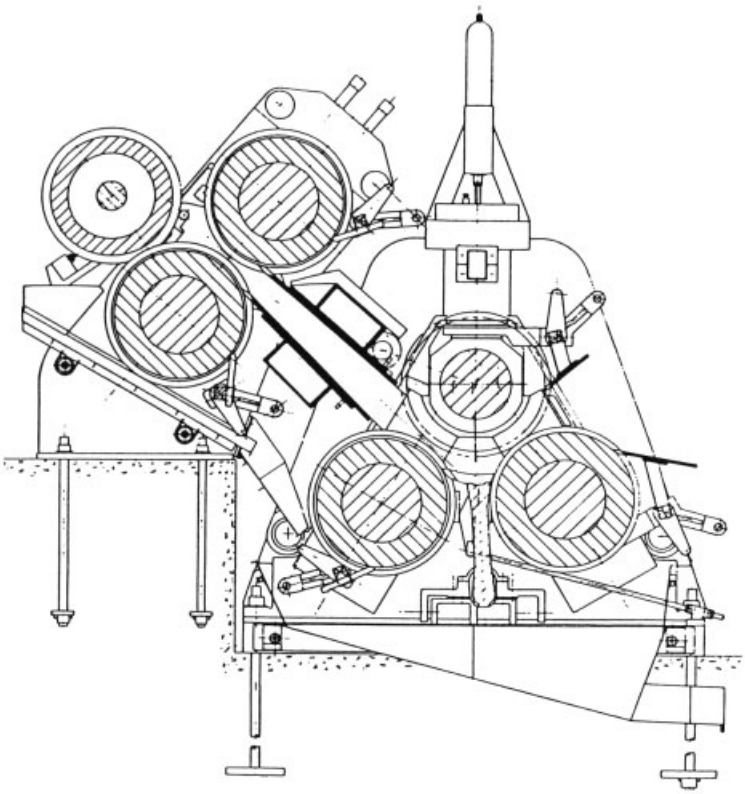

Figure 13. Constant-ratio cane mill with pressure feeder combination

provide more fiber as fuel (for electricity cogeneration at the factory) and for environmental protection [221]. Important factors in cutting are to produce clean, undamaged cane, free of trash, and to leave viable root stock in the field. *Mechanical harvesting* is found in Australia, the United States, some Caribbean and Latin American Countries, and new developing cane areas in Southeast Asia. Most common are *combine harvesters*, or chopper harvesters, developed in Australia, which cut cane stalks at the base, cut the stalk into billets [9 to 15 inches (i.e., 28–38 cm) long], blow excess leaves and trash off the billets, and drop the billets into a cane cart pulled alongside the combine harvester. In Louisiana, or where tonnage is light, *soldier harvesters* cut and top erect cane, leaving rows of whole stalks in the field, which are burnt after harvest because the canopy is too light to support a burn on standing cane. Other whole-stalk harvesters in Hawaii, where cane tonnage is heaviest, are the *V-cutter,* which cuts cane at base but not at top, and the *push-rake*, which pushes cane, including the roots, out of the ground (used on hilly areas), necessitating replanting. Under good conditions, a man can cut 0.5 t of cane per hour, and a combine harvester 30 t of cane per hour, with other mechanical systems between 15 and 30 t/h. Mechanical cutting is always more expensive than hand cutting and yields lower quality, more damaged cane and an increased portion of trash, but is increasing for sociological reasons.

Transportation. Cane *loading* in the field is accomplished by hand, grab loaders, or continuous belt loaders, into small bins or wagons, which collect at transloader stations for transfer to larger transport containers. In some areas (India, Pakistan, Southeast Asia, Africa), cane is still transported in small bullock carts. Transport by narrow gauge railroad (cheapest) continues in Australia and the Philippines; by water, in China, Southeast Asia, and Guyana; and by road, elsewhere. Chopper-harvested cane must

be shipped directly to the mill and be processed on arrival, not stored in the mill yard, to prevent serious deterioration: delivery time of less than 24 h is recommended [220].

Harvesting and shipping schedules are decided between grower and processor to ensure a constant supply of cane for the mill, and a fair distribution of maturity and quality.

Cane payment is generally based on weight, with a deduction for trash. Where payment for cane quality (weight, sucrose content, fiber, sugar yield) has been introduced (United States, South Africa, Australia, Brazil, Colombia, Philippines), the quality and efficiency of the industry have greatly improved [222].

20.5. Sugarcane Processing: Milling and Diffusion

Cane that arrives at the factory (or mill, Central, Ingenio, or sugarhouse) is either processed directly or stored temporarily (preferably <12 h) in the factory yard. Because factories operate 24 h but transport is only by day, cane is stored in delivery vehicles or piles in the factory yard for nighttime grinding. Hygiene and prompt use are important for good sugar yields. The cane is then loaded onto a moving table to enter the factory. Sugarcane factories generally run during the entire harvest season, shutting down temporarily only to clean out evaporators (see Section 20.5). Hawaiian and Australian factories close on weekends. Factories range in size from 300–400 t of cane per day (tcd) in some less developed areas, through the usual range of 4000–10 000 tcd, to the large Florida factories of 20 000–25 000 tcd, and the giant Brazilian processors of >40 000 tcd, where ethanol is produced in addition to sugar.

Cane Weighing and Testing. Cane is weighed in the delivery vehicle (other than boats) on platform scales, with a second weighing of the empty vehicle. When payment is based on cane quality, samples are taken from the truck (grab sample or, better, core sampler) [222] for analysis or from a hatch opening across the mill after the knives; the most modern system (Florida) is to sample the first mill juice from cane, on an accurately timed basis, where the truck is labeled and the label is related by a timing system to the juice from that truck's cane.

20.5.1. Cleaning

Hawaiian factories, where cane arrives with roots, mud, and rocks, first dry-clean cane to remove loose debris; then the cane is washed extensively. Louisiana factories, and others in muddy areas, wash cane on the ascending cane conveyer or feeder table as it enters the mill. Magnets above the table remove metal trash. Sand must be removed especially carefully because it can lead to erosion of downstream equipment.

20.5.2. Juice Extraction

Cane is *prepared* for extraction through chopping with sets (usually two) of revolving knives, which slice it into small pieces (1–4 inch, ca. 2–10 cm, ideally) but do not extract juice. Cane passes from knives through a shredder, often a swing-hammer mill, which mashes up the sliced cane, breaking open cells to allow increased extraction at the next stage. Juice is then extracted by either milling or diffusion.

Milling. Shredded cane passes through a crusher mill and a series of three to seven subsequent sets of three-roll mills. The rolls are 2–3 m long. These mills are built in a multitude of structural variations. Four- or five-roll mills are becoming more common. A three-roll mill with pressure feeder combination is shown in Figure 13. Continuous-pressure feed improves the speed and efficiency of extractions. Water (10–35 % by weight of cane) is applied to the milling train in a reverse direction to the cane, on the last mill, with juice recycled backward, to increase extraction in a process called "imbibition." *Crusher juice* is obtained from the first mill; it is highest in sugar (12–17 %). *Last mill* or *residual juice* is that from the final mill (1–3 % sugar). Juices from all mills are strained or screened into tanks and called *mixed* or *dilute juice*.

Diffusion (see Section 7.4). The sugarcane diffusion process has been developed from sugar beet extraction. Here, cane from the shredder must be prepared further in a fiberizer, or extended shredder, for best extraction. Because of this finer preparation, diffusion gives a higher degree of extraction (93–98 %) than milling (85–95 %); therefore, further cane preparation is increasingly used in mill trains also. Finely prepared cane enters a multicell, countercurrent diffuser of linear (→ Liquid–Solid Extraction, Chap. 2.3.2), diagonal (→ Liquid–Solid Extraction), or circular design (→ Liquid–Solid Extraction, Chap. 3.1) [7, 15]. In the diffusers, shredded cane moves countercurrent to hot water (75 °C). This system is applied for *cane diffusion*. Various combinations of sets of mills with a diffuser are called *bagasse diffusers*. Bagasse emerging from the diffuser must be dewatered to reach the approximately 50 % moisture of mill-run bagasse; at this moisture, bagasse can be fed as fuel to factory boilers.

20.5.3. Factory Fuel

Cane factories, except for those in the People's Republic of China and Cuba (where wood fiber is in short supply and bagasse is a fiber source), burn their own bagasse and are more than self-sufficient in energy, with excess energy used for cogeneration [Cuba, United States (Hawaii, Florida), Mauritius], distillery operation (Brazil, other South and Central American countries, Southeast Asia), or increasingly, a refinery at the end of the mill (a white end refinery).

20.6. Cane Factory Process

The factory process, outlined in Figure 14, is general for the production of *raw sugar,* a light-brown product that will subsequently be refined into white and brown sugar food products (see Section 13.1). Cane sugar has traditionally been processed in two stages because sugarcane cannot be stored after harvest: a crude raw sugar product was made, seasonally, in cane-growing areas and shipped to areas of major consumption (North America, Japan, Northern Europe), where it was refined into final products. Raw sugar is traded on the commodity futures markets. A white sugar product (*mill white, plantation white crystal*) has traditionally been produced at sugar factories in the tropics for local household and industrial use; it is described in Section 20.7, and traditional refineries are described in Section 20.8.

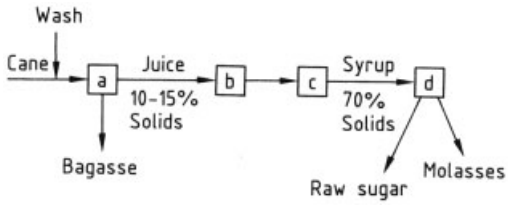

Figure 14. Schematic of sugarcane factory process for raw sugar manufacture
a) Mill or diffuser; b) Clarification; c) Multistage evaporation; d) Vacuum crystallization

Juice treatment for raw sugar manufacture includes clarification, evaporation, and crystallization. The mixed juice extracted from cane is at pH 4–5, dark colored, and turbid. For *clarification*, the juice is heated to < 75 °C to destroy invertase and other enzymes, and treated with lime (milk of lime or lime in a sugar solution, named "lime saccharate") at concentrations of ca. 0.5 kg of CaO per tonne of cane to increase the pH and stop sucrose inversion. Many different modifications of the process exist, along with various designs of clarifiers and feeders; heating and liming are sometimes done in reverse order, depending on juice purity. Lime and flocculating polymer (→ Flocculants, Chap. 2.2, → Flocculants, Chap. 2.3, → Flocculants) settle out of heated juice in the clarifier tanks, taking suspended and colloidal solids (bagasse, soil, inorganic material, waxes, gums) with them in a process called *defecation*. Clear, clarified juice is pumped off the top of the clarifier tanks; muds are pumped from the bottom and filtered on rotary vacuum dry filters, on which residual sugar is washed out. The filtrate is returned to the mixed juice or mixed with clarified juice; the latter procedure is a source of contamination unless filtrate has been clarified with a flotation clarification step, e.g., talo filtration (see Section 20.8) [223, 224].

Clarified juice, which contains about 13–15 % solids, is pumped directly to a series of

multiple-effect *evaporators*, in which the vapor from the first effect is used to heat the second, and each subsequent effect (usually three to five) is operated at vacuum and lower temperature (see Section 9). Many designs and combinations of evaporators are in use [7, 15]. Scale buildup in evaporator bodies reduces heat transfer and must therefore be removed; for this "boilout," the factory stops processing several times per season, unless it has a second set of evaporators. *Evaporator syrup* (corresponding to thick juice in sugar beet processing) of 62–69 % solids content is the product. Syrup in some modern factories is clarified in an additional step by flotation clarification to produce high-grade raw sugar for sale or for the factory's own refinery [225].

Syrup is pumped to the *crystallization station*, a series of single-effect vacuum pans; again, many designs and combinations exist [7, 15] (see also Section 10.1). In a pan, syrup is evaporated to supersaturation and seeded with fine sugar crystals, which act as nuclei for crystal growth. When crystals have grown to desired size and number, the crystal–syrup mixture (massecuite, fillmass) is discharged (the action of emptying the pan is called a "strike") into a mixer, and from there to a centrifugal (basket-type, batch centrifugal for high-grade sugars, continuous for low-grade sugars), where the mother liquor syrup is spun off ("purging") and the crystals are "washed" (i.e., sprayed with a fine stream of water in the centrifugal to remove most of their syrup coating so that they may be free flowing in bulk storage) (see Section 11).

Runoff syrup from the crystallized evaporator syrup (A-strike) is called *A-molasses*; it is boiled again to crystallization in a vacuum pan (B-strike, with runoff *B-molasses*). This serial boiling is continued usually three times (see Section 13.1), with various modifications, cuts, and cross-blending, to obtain the maximum yield of sugar. Crystallization is the cheapest and most effective purification step; sugar syrups of purity lower than massecuite are cooled in open stirred *crystallizers* at atmospheric pressure until sufficient crystals have formed for centrifugal separation. Designs for crystallizers are as numerous as those for vacuum pans [7, 15]. The lowest grade of molasses from which no more sugar crystals can be recovered economically is *blackstrap molasses*, a heavy, viscous, bitter material containing 26–40 % sucrose, 9–23 % invert sugars, and at least 80 % total solids. About 24–40 kg of molasses is produced per tonne of cane, depending on the purity of the cane juice and the processing systems.

Sugar cane juices have a roughly ten times higher invert sugar content than beet juices but their amino acid content is 50 % lower than that of beet juices. Thus, color formation can be limited by working at low temperatures and low pH values during juice purification. The invert sugar is not destroyed during all processing steps.

The A- and B-strike sugars are usually mixed to become the raw sugar of commerce, traded on a 96°Z (pol) basis. Very high pol (VHP) sugars of more than 98°Z (pol), are produced in some areas (e.g., Hawaii and South Africa) [226] by remelting B-sugars and washing the A-product extensively, sometimes after syrup clarification.

Raw sugar is cooled and dried on belts conveying it to bulk storage (dryers are not common) and is weighed into the warehouse. The custom of bagging raw sugar for storage and shipping has almost ceased.

20.7. Direct White Sugar Factory (see Section 13.2)

The characteristic reagent in the processes for direct white (mill white, plantation white) sugar production is sulfur dioxide (produced by burning sulfur in air or from liquid SO_2), which is used to inhibit nonenzymatic browning by addition either directly into the juice or into the carbonated syrup to lower the pH during sulfitation. Direct white sugar is produced by a variety of processes:

1) Settling clarification and juice sulfitation [12, 15]
2) Single or double carbonation with syrup sulfitation [12, 15]
3) Blanco Directo process (juice sulfitation; juice filtrate and syrup clarification) [223, 224]

In all these processes, powdered carbon, in pressure filters, may be used for additional decolorization. Various modifications of processes 1 and 2, using carbonation methods similar to those for beet sugar production, are usually applied in India, China, Africa, Southeast Asia, and

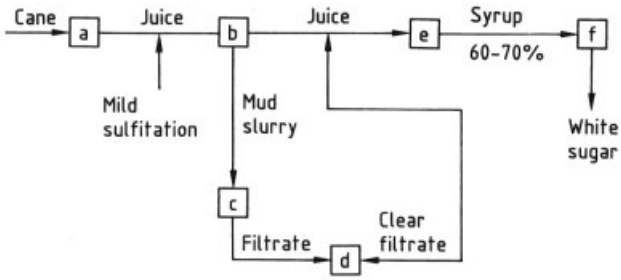

Figure 15. Schematic of process for manufacture of Blanco Directo (direct white sugar)
a) Mill; b) Clarification; c) Rotary vacuum filter; d) Filtrate clarification; e) Evaporation; f) Crystallization

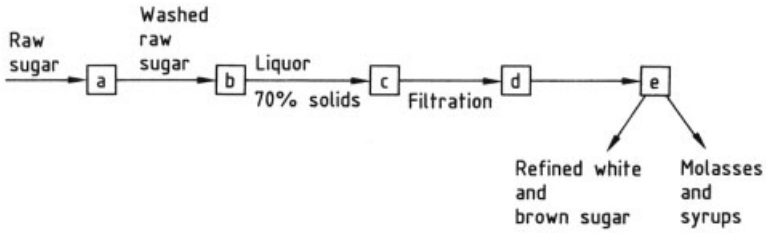

Figure 16. Schematic of raw sugar refining process
a) Mingler and affination; b) Melter; c) Clarification (phosphatation, carbonatation); d) Decolorization (granular carbon, bone charcoal, ion-exchange resin); e) Vacuum crystallization

South America for sugar for domestic production and are discussed thoroughly in older literature [12]. The above processes are supplanted in newer factories in these countries by type 3 Blanco Directo processes (see Fig. 15), which can readily convert a raw sugar factory to production of white sugar of quality higher than the standard mill white and suitable for food and beverage processors in the tropics. Many factories produce either direct white or raw sugar at different time periods [227].

20.8. Refining [228] (see also Section 13.4)

Traditional cane sugar refineries use raw cane sugar at 96 to 98°Z (pol) as input ("melt") and produce not only a range of granulated and liquid sucrose products, similar to those described in Chapters 14 and 15, but also a range of brown sugars and syrups, as well as edible syrups (molasses). The brown products have characteristic palatable cane and cane molasses flavors, not available from sugar beet. A generalized refining scheme is shown in Figure 16. Refineries are large processing plants operating around the clock typically for five (weekend shutdown) or ten days (four-day shutdown). Fuel for free-standing refineries is fuel oil, natural gas, or coal, according to local availability; a few refineries have extended their power plants to generate extra electricity for the local grid; refineries attached to raw sugar factories use the factory's excess bagasse fuel.

The quality of incoming raw sugar is paramount for efficient operation. Sucrose content (pol) is a universal quality criterion. Color, ash (inorganic), invert sugar, moisture, dextran content, and grain size are other criteria that may be included in raw sugar purchase contracts [229].

Raw sugar is weighed into the refinery from rail car, ship, or raw sugar warehouse, and conveyed to the affination station, where it is mingled with a heavy syrup (80 % solids content), then spun in basket centrifugals and washed with a spray of water to remove the added and the integral syrup coatings. The washed raw sugar is dissolved (melted) to give a washed sugar liquor of ca. 70 % solids content, which is pumped to clarification. Three types of clarification exist:

Phosphatation. Phosphate (phosphoric acid, P_2O_5, concentration up to 400 mg/kg solids) and calcium hydroxide (as milk of lime or sugar solution of lime, up to pH 7.5–8.3)

are combined with the sugar liquor in an aerated flotation clarifier (Williamson, Jacobs, or Sucrest designs). Calcium phosphate forms, occluding suspended solids and inorganics in its mass, and floats to the surface where it is scraped off by rotating blades. Clarified liquor (syrups are called liquors in refineries) is pumped out from the bottom of the clarifier [7, 14, 15, 223]. The process removes 25–40 % color, ash, and turbidity from the sugar liquor.

Talo Phosphatation. Phosphatation is performed as described above with the addition of *color-precipitating chemicals* (quaternary ammonium compounds with long-chain fatty acid substituents) and a series of mud-desweetening steps, which remove a greater amount of color (30–50 %), ash, and turbidity [223 – 225, 227]. This process is widely used in white end refineries and has almost replaced traditional phosphatation.

Carbonation (Carbonatation in the United States and the United Kingdom). In this process, lime and carbon dioxide are mixed in liquor in a two-stage process similar to that for beet sugar processing but carried out on liquor of 65–70 % solids.

Any type of clarification is followed by *filtration* through leaf-type vertical or horizontal pressure filters. Carbonated liquors, containing calcium carbonate, may require addition of diatomaceous earth as a filter precoat. Phosphatated liquors are generally filtered with the addition of diatomaceous earth as precoat and body feed.

Filtration, often a refinery bottleneck, especially with poor-quality raw sugar, is followed by *decolorization* with bone char (traditional), granular activated carbon (now most common), or ion-exchange resins, or any combination of these (see Sections 13.4, 14.4). Discussions on comparative merits and regeneration of these decolorizing systems may be found in [7, 15, 230, 231].

Decolorized liquor, or fine liquor of very pale yellow color, is evaporated further to 72–74 % solids and sent to *crystallization* in a series of vacuum pans, as with raw cane sugar and beet sugar. Refinery strikes are designated 1, 2, 3, etc. Four to six white sugar strikes are common. The lowest-grade runoff syrups are sent to a second series of pans and crystallized to improve sugar recovery. *Brown low-grade runoff syrups* and refiners' final molasses are sold for food processing, brewing, and blending to make cane syrups and edible molasses.

Refined brown sugars ("soft sugars" in the trade, see Section 15.7) are made by crystallizing sugar from a mixture of third and fourth runoff syrups and affination syrup (boiled brown sugars) or by coating white sugar crystals with a brown-liquor–caramel-syrup (painted or coated brown sugar). Table 14 shows the composition of typical raw cane sugar and of refined granulated, direct mill white, and Blanco Directo sugar.

White end refineries are referred to above. These have the advantage of using available fuel from the raw sugar factory to which they are attached, although this may require seasonal operation. Increasingly, raw sugar factories are grouping together and building a white end refinery to process their raw product [224]. Processes employed are typically melting without affination because the raw sugar can be washed well in the factory; use of color precipitants (talo phosphatation); decolorization with granular activated carbon or ion-exchange resin; and crystallization. Customarily, only white sugar is made, and low-grade syrups are returned to the attached raw sugar factory. This system seems to be the trend for the future, replacing the large free-standing refinery.

21. Sugar from Other Plants

Palm Sugar. The production of palm sugar is of some importance in India (production 170 000 t/a), Burma (production 20 000 t/a), Indonesia, and in the Philippines. Of the nine species of palm with sucrose-containing cell sap, four are cultivated in India: *Phoenix sylvestris*, *Borassus flabilliformis* (palmyra), *Cocos nucifera* (coconut), and *Caryota urens* (sago). The tapping of palm juice (neera) is accompanied by many difficulties.

After purification with lime and skimming of the floating precipitate, the palm juice is concentrated to a thick syrup by evaporation in open pans [232]. It is vigorously stirred as the sucrose begins to crystallize. This "massecuite" is then filled into clay molds or palm-leaf baskets

Table 14. Composition of typical raw, refined granulated, direct mill white, and Blanco Directo sugars [212, 219]

Compound	Raw sugar	Refined sugar (from raw sugar)	Mill white	Blanco Directo
Sucrose, °Z	97.73	99.9	99.6	99.9
Color, IU*	1800	30	100–200	20–50
Invert sugar, wt %	0.56	0.006	0.15	0.02
Ash content, wt %	0.45	0.007	0.15	0.05
Moisture, wt %	0.64	0.023	0.15	0.03
Organic nonsugars, wt %	0.62	0.014	0.40	0.03
SO_2, mg/kg			1–5	20–50

* ICUMSA unit.

through which the mother syrup can drain off. For the production of 1 t of palm sugar, ca. 2.3 t of palm syrup with a solids content of 60 % and a purity of 84 % are required [24].

Maple Sugar and Maple Syrup [23]. In the United States (especially the state of Vermont) and Canada, maple syrup and maple sugar are isolated from the juice of the sugar maple (*Acer saccharinum*). Maple syrup and maple sugar are used in the production of food syrup, preserved food, jams, and ice cream and in the fermentation of tobacco because of their special flavor. Annual production in the United States is ca. 9000–10 000 t of maple syrup and 130–150 t of maple sugar. Juice is tapped from maple trees from February to April. The tree must be about 40 years old before juice can be obtained. It then produces juice containing 2.5–3.5 % sucrose for 100 years. The juice is tapped from a hole bored in the trunk about 1 m from the ground. Before evaporation, maple juice is purified by adding milk of lime, and the precipitate is filtered off. About 0.9–1 kg of maple syrup with a solids content of 66.8 % (66.1 % carbohydrates and 0.7 % ash) is obtained per tree per year [233].

Sugar from Sweet Sorghum. Like sugarcane, sorghum (*Andropogon sorghum saccharatum*) [22] belongs to the class of grasses. It is grown in the southern parts of the United States for production of a sweet syrup and was studied as a sugar and energy crop in, e.g., Brazil, Hungary, Austria, Italy, Germany, Belgium, Greece, and Spain. As with sugarcane, the cell juice, which contains ca. 10–14 % total sugars (glucose, fructose, and sucrose), is obtained by pressing. Subsequent process steps are also similar to those used in the cane sugar industry. In Louisiana, extensive cultivation was attempted to prolong the total working time of a factory by introducing a sorghum campaign.

Date Sugar Syrup. Ripe dates contain about 30 % water. The largest part of the solids is water soluble and consists mainly of glucose, fructose, and small amounts of sucrose. The seed is a problem in processing dates to liquid sugar. Nevertheless, techniques [20] for the isolation of invert sugar from dates have been worked out and are used on a large scale in Iraq.

Sugar from Sweet Corn (*Zea Mays L.*). The isolation of sugar from the sucrose-containing cell juice of corn stalks has achieved no technical importance [21].

22. Quality Demands on Sugar and Side Products of Sugar Production

22.1. Refined and White Sugar

In the past, white sugar was divided into "refined sugar" and "white sugar." "Refined sugar" referred to a sugar obtained by recrystallization of crystalline sugar.

Today, white sugar is evaluated and labeled only according to its chemical and microbiological quality. Since white sugar is a commercial product worldwide, specifications for it exist in the form of official, internationally and nationally obligatory regulations or standards in pharmacopeias, and voluntary quality agreements between sugar manufacturers and the sugar processing branches of industry.

The sucrose content of white sugar is almost always > 99.7 %. For this reason, evaluation is based not only on the content of water, invert sugar, SO_2, heavy metals, etc., but also on the

cumulative parameters for the individual classes of impurities, such as:

1) Ash content (i.e., content of inorganic or organic salts)
2) Color in solution (i.e., content of Maillard products and colored higher molecular mass compounds)
3) Color type (i.e., optical appearance of the sugar by visual comparison with a type series)

The Codex Alimentarius [234] divides white sugar into two categories, A and B, and stipulates certain requirements for them (Table 15). Specifications valid within the EC [187], which have been accepted by many other countries, are generally used in international white sugar trading (Table 16). The requirements to be met in the United States are published in the Food Chemical Codex [188].

Table 15. Essential composition and quality factors for white sugar, according to WHO–FAO [234]

Specification	A	B
Sucrose content, °Z	>99.7	>99.5
Invert sugar content, %	<0.04	<0.1
Conductivity ash, %	<0.04	<0.1
Loss on drying (3 h at 105 °C), %	<0.1	<0.1
Color, IU *	<60	<150
Food additives, mg/kg		
Sulfur dioxide	<20	<70
Contaminants, mg/kg		
Arsenic	<1	<1
Copper	<2	<2
Lead	<0.5	<0.5

* ICUMSA units.

Identification: Dissolve 26 g of the sample in water and dilute to 100 mL at 20 °C. Determine the polarization in a saccharimeter, by using a 20- cm tube.

Further quality requirements for white sugar are as follows:

Sucrose content, °Z	99.8–100.2
Arsenic (as As), mg/kg	< 1
Color, ICUMSA units	< 75
Conductivity expressed as residue on ignition, %	< 0.15
Heavy metals (as Pb), mg/kg	< 5
Invert sugar, %	< 0.1
Lead, mg/kg	< 0.1
Loss on drying, %	< 0.1
Ash residue %	< 0.15
Specific rotation $\left([\alpha]_D^{20}\right)$	+ 65.9–66.7°

The heavy-metal contamination of white sugar is very low (Pb < 0.1 mg/kg; Cd < 0.002 mg/kg; Hg < 0.003 mg/kg within the limits of detection). The same applies to the concentration of herbicides, fungicides, and insecticides, which lie within the analytical limit of detection.

Apart from these official minimum requirements, other parameters are important for storage and industrial processing [235]. These include the total (< 0.06 %) and surface water (< 0.03 %) content [197]; the content of water-insoluble solids [236], which must be < 7 mg/kg (determined with an 8-µm filter) for some consumers; and the filth [237] (insect parts and rodent hairs). Furthermore, for the beverage industry, the saponin content (< 1 mg/kg) of white beet sugar and the polysaccharide content (dextran < 100 mg/kg) of white cane sugar are important because both cause foam and floc formation [238]. The content of betaine, amino acids, and cations (calcium) can also be of significance in individual cases [239].

For storage in small silos (up to 100 t), the emptying and internal factory transport of sugar to the places of use, dust content (< 200 µm), and grain-size distribution are of substantial importance. Sugar with a d' of 1.0–1.2 mm, $n >$ 3, and a dust content of < 1.5 % is most suitable for this purpose.

The screen analysis of sugar is evaluated according to the RRSB (→ Size Enlargement, Chap. 3.3) or to the Gaussian distribution (→ Size Enlargement, Chap. 3.3).

Although sugar is a food with one of the lowest germ contents, certain consumers (manufacturers of nonalcoholic drinks, canned food industry, etc.) demand practically germ-free sugar (total content of microorganisms < 200 per 10 g, yeast, mold < 10 colonies per 10 g etc.). Target values for the permitted microorganism burden of sugar, especially for "liquid sugar," have been put forward by the National Soft Drink Association, United States (Bottlers Test), and the National Canners Association, United States [240] (Canner's Test), and are maintained by the sugar industry on a voluntary basis. The internationally recommended microbiological methods of analysis are found in ICUMSA proceedings or in the ICUMSA Methods Book [55].

Table 16. Quality criteria for white sugar according to EC sugar market regulations (EC 1972) (grade 2 is white sugar of standard quality), according to [187]

Quality criterion	Grade*		
	1	2	3
Sucrose content (polarization), °Z (min.)		99.7	99.7
Moisture content, % (max.)	0.06	0.06	0.06
Invert sugar content, % (max.)	0.04	0.04	0.04
Color type, Brunswick unit (max.)	(2)	(4.5)	(6)
Points (0.5 unit = 1 point)	4	9	12
Conductivity ash, % (max.)	(0.0108)	(0.0270)	
Points (0.0018 % = 1 point)	6	15	
Color of solution, IU (max.)	(22.5)	(45)	
Points (7.5 units = 1 point)	3	6	
Points according to EC point system (max.)	8	22	

* Quality criteria for all grades: sound, fair and marketable quality, dry, in homogeneous granulated crystals, free-flowing.

22.2. Molasses

If molasses is used in feed, it must comply with feed regulations of individual countries with respect to composition, heavy-metal content [241], agricultural pesticide residues [242], and labeling as single or mixed feed. The same applies to other feed obtained in the production of sugar, such as pressed pulp, dried pulp, molassed dried pulp, residual molasses, etc.

If molasses is sold as a raw material (nutrient medium) for the production of ethanol, yeast, citric acid, or lactic acid, and for other chemical or biochemical processes, general trading conditions or as-agreed specifications are valid [solids content (min.) 76.3 % = 40.5°Bé].

Beet Molasses. Beet molasses accumulates in production with a solids content of 80–89 %. Apart from sucrose, the solids in molasses comprise, depending on purity, 40–50 % of nonsugar substances, which can be divided as follows [243] (all nonsugars = 100 %):

1) Other usable saccharides such as glucose, fructose, raffinose, kestoses, and galactinol (1.5–4 %)
2) Inorganic nonsugar substances: K^+, Na^+, Mg^{2+}, Ca^{2+}, Fe^{3+}, Pb^{2+}, Cl^-, SO_4^{2-}, NO_3^-, NO_2^- (ca. 33 %)
3) Nitrogen-containing organic nonsugar substances [244]: pyrrolidonecarboxylic acid, γ-aminobutyric acid, glutamic acid, aspartic acid, glycine, tyrosine, serine, alanine, leucine, isoleucine, valine, betaine, peptides, and Maillard products (42 %)
4) Nonsugar substances containing no nitrogen: citric, malic, fumaric, succinic, oxalacetic, oxalic, tartaric, glycolic, lactic, arabonic, saccharinic, formic, acetic, propionic, butyric, and glacturonic acids; araban and galactan fragments

Apart from these constituents, other compounds have been detected in the part-per-million range in beet molasses [245]. Investigations have been carried out in past years because molasses-containing substrates have occasionally suffered from inhibition of growth or metabolism, which was assumed to be caused by pesticide residues. The investigations could not verify these assumptions [246]. However, increased concentrations of volatile carboxylic acids (formic and propionic acids) and furfural-like compounds, which can arise on heat treatment of molasses in the acidic range, appear to exert inhibitory effects under certain conditions.

Cane Molasses (Blackstrap). Cane molasses has a different composition from beet molasses because the components of sugarcane cell juice are different. According to [247], normal cane molasses with a water content of 17–25 % has a total sugar content (sucrose, glucose, fructose) of 45–50 %. The content of polysaccharides (dextrans, pentosans, polyuronic acids) is 2–5 %, and that of peptides and free amino acids is 2.5–4.5 %. The content of nitrogen-free acids (aconitic, citric, malic, oxalic, mesaconic, succinic, fumaric, and tartaric acids) is 1.5–6 %. The content of cations, determined as carbonate ash, varies between 7 and 15 %.

The carbonate ash has the following composition: 30–50 % K_2O, 7–15 % CaO, 2–14 % MgO, 0.3–9 % Na_2O, 0.4–2.7 % Fe_2O_3, 1–7 % SiO_2 and insoluble substances, 7–27 % SO_4^{2-}, 12–20 % Cl^-, and 0.5–2.5 % P_2O_5.

The storage of cane molasses is problematic because of its low pH. Chemical or microbiological decomposition of cane molasses stored in tanks occurs from time to time [248].

Cane molasses can be utilized in part like beet molasses (see above), e.g., as pelletizing aids. It behaves differently as a fermentation substrate; because of its aromatic taste and odor, it gives rise to arrack and rum on alcoholic fermentation. Aconitic and itaconic acids are obtained from cane molasses [249].

Three types of cane molasses are available commercially, corresponding to the water and sugar (sucrose, glucose, and fructose) content: *superior blackstrap molasses* (water content = 23.4 %, total sugar content = 53.5 %), *blackstrap molasses* (water content = 23.5–26.4 %, total sugar content = 48.5–53.4 %), and *utility blackstrap* (water content = 26.5 %, total sugar content = 42.5–48.4 %).

High-Test Molasses. High-test molasses is not molasses in the original sense of the word, which defines it as the final syrup obtained after crystallization of sugar. It is rather a thin juice that is two-thirds inverted, only moderately purified, and concentrated to about 85 % solids. High-test molasses consists of about 15.5 % water, 27.0 % sucrose, 50.0 % invert sugar, and 2.25 % ash.

Dry Molasses. As a result of its high viscosity, molasses is difficult to handle and, consequently, cannot be used alone as feed in agriculture. To open up this market, methods have been developed for the production of both beet and cane dry molasses. The drying operation used is either spray drying [250] or drum drying [251]. To make drying possible, carriers (e.g., calcium oxide [250, 251]) or floury feeds [252] (e.g., alfalfa flour or whey powder) must be added to the molasses.

22.3. Vinasse

After the fermentable sugar in molasses has been biotechnologically used and the compounds produced have been isolated, the nonsugar substances that are not assimilated and metabolic side products remain more or less completely in the nutrient medium. In Baker's yeast fermentation, for example, ca. 35–45 % of the total nitrogen content of the beet molasses is assimilated. These substrates are being increasingly subjected to multiple-effect evaporation to give a solids content of 70 %. They are referred to as vinasse or condensed molasses solubles [249]. Vinasses have gained approval by feed regulating authorities, and are used as feed or soil conditioner [253]. Efforts are also being made to use vinasse directly for the production of biogas [249].

22.4. Liquid Feed Supplement [254]

For large animal farms (cattle, sheep), a liquid feed based on diluted molasses (60–70 %) is produced in the United States. Urea, biuret, mineral substances, trace elements, vitamins, etc., are added to optimize requirements in the basic feed and are then added to this liquid feed.

22.5. Beet Pulp

Beet pulp obtained on extraction of cossettes is pressed to an appropriate solids content, dependent on individual factory conditions. If pressed pulp is delivered directly as feed or for preservation by ensiling, it is mostly pressed to 18–22 % solids content. If pulp is dried after pressing, factories try to achieve the highest possible solids content by pressing, mostly in the range of 28–35 %, to save drying costs. Their solids, beet marc, consist of ca. 20–25 % cellulose, 20–25 % pectin, 18–23 % araban, 5–10 % galactan, 6–10 % raw protein, 5–6 % ash, and 4–5 % sugar.

In contrast to other agricultural animal feeds, pressed pulp is normally ensiled hot, at 50 °C, with the aid of thermophilic lactic acid bacteria [255]. Shrink-wrapped round bale ensiling, a well-known development in agricultural ensiling, was studied as a new way of marketing pressed pulp [256].

Dried pulp is obtained by drying pressed pulp (see Section 17.2) to a solids content of 90–92 %. If 1–3 % of molasses (based on the weight of the pressed pulp) is added to the pressed pulp before drying, *molassed dried pulp* with varying sugar content (9–30 %) is obtained. Apart

from sugar content, the concentrations of HCl-insoluble substances, calcium, and ash are controlled in individual countries according to feed regulations.

Sepa patent cossettes are obtained by drying on pressed pulp a previously homogenized molasses preliming flocculation mixture(see Section 8.1). They have a solids content of 91–92 % and consist of ca. 9.9 % raw protein, 12.7 % raw fiber, and 9.8 % raw ash of which 6.1 % is $CaCO_3$ and 0.65 % is P_2O_5. The sugar content is 17–18 %. In the case of *Torno slices*, carbonation sludge (see page 26) is dried to 3 % solids instead of the preliming precipitate.

"Steffen cossettes" are obtained by drying partially desugarized wet pulp (sugar content 28–35 %). Dried pulp with a total nitrogen content of 2.5–3.5 % is obtained directly by extraction of cossettes with liquid ammonia (see Section 7.3) or by treatment of dried pulp with liquid ammonia in pressure vessels [257]. Feed with a high NPN (nonprotein nitrogen) total nitrogen content can also be produced by addition of urea to a dried pulp–molasses mixture. The finished feed consists of ca. 60 parts of dried pulp, 25 parts of molasses, and 15 parts of urea. Another production process for amide cossettes starts directly with pressed pulp [258]. To obtain a feed with a higher NPN-level also an aqueous $(NH_4)_2SO_4$ solution (1 % based on dry substance) can be mixed with pressed pulp.

Full-quality beet cossettes are obtained by drying fresh beet slices (see Section 6.9) to a solids content of 92–94 %. Their composition corresponds to that of fresh beet.

23. Economic Aspects

The worldwide sugar economy is determined by the following factors:

1) Sugar is an agricultural product, affected by crop failure and record crops (i.e., it is available in changing supply)
2) The demand for sugar depends on eating habits, the economic situation, and population development
3) Since the 1970s, apart from traditional glucose syrups, sugar has had a serious competitor in the form of HFCS (→ Glucose and Glucose-Containing Syrups)

Today, sugar is a staple food. In industrialized countries, 10–15 % of the daily intake of calories is provided by sugar and sweeteners produced from corn or wheat. In some countries the sugar consumption per person per year has remained practically unchanged for decades. For instance, it remained virtually constant in the United States and the United Kingdom from 1920 into the 1980s.

Until 1850, the cane sugar industry alone dominated the market. About 5000 t/a was imported into Europe of which, on average, 10 kg per person per year was consumed in England, 7 in the Netherlands, 2 in France and the German "Zollverein," 1 in Austria and Italy, and 0.5 in Russia [259].

The beet sugar industry then developed in Central Europe and became a competitor of the cane sugar industry. In an effort to be independent of imports, more and more countries started their own sugar production and switched from being importing to possible exporting countries. Nowadays, more than 100 countries have their own sugar production from sugarcane or sugar beet.

Countries that have a permanent sugar requirement do not buy sugar on the open market, but often conclude long-term supply contracts with exporting countries at preferential prices. For instance, the United States awarded export quotas to individual countries, especially Cuba until 1960. Thus, the United States was the main purchaser of Cuban sugar until that time. Since Fidel Castro's revolution and until 1990, the Eastern-bloc countries, the Soviet Union, and the People's Republic of China have received most of the Cuban sugar. Sugar imports into the United Kingdom were also controlled by special regulations (Commonwealth Sugar Agreement of 1951) or arranged in agreements such as the African, Malagasy, or the Lomé Agreement (1971). In the latter, the countries of the EC were committed to buy 1.3×10^6 t of sugar per year from these countries at EC prices. For this reason, only a fraction, about 10–15 %, of world sugar production is traded on the stock markets of New York, London, and Paris.

Since at all times and in all countries, sugar has represented a source of taxes that are profitable and easy to raise, good statistics are available. Table 17 shows the development of world sugar production and cane sugar production as

percentage of total production since 1900 [260]. Worldwide sugar production has increased tenfold since 1900. At more than 50 %, the percentage of beet sugar was at its maximum before World War I. Since 1940, it has been in the range of 34 % (1980–1992) to 40 % (1960).

The reason for the favorable development of beet sugar production in 1964–1965 was widespread sugar beet cultivation in the Soviet Union after World War II. Since 1970, the percentage of cane sugar production has increased to about 75 % and has remained at that level.

Since July 1, 1968, EC members have formed a common sugar market based on decree No. 1009/67 EEC from December 18, 1967, which was valid largely unchanged until July 2006. For the main areas of surplus production in the EC, a target price and an intervention price were set for unpacked white sugar of a definite standard quality ex factory (see Table 16, grade no. 2). Raising a surcharge for imports from third countries and payment of a reimbursement on exports should compensate for price differences within and outside the EC. A desirable limitation of sugar production was achieved by establishing basic quotas with a price and sales guarantee by the EC for each sugar-producing company (A quota: the level depends on domestic requirements in the EC). These guarantees were reduced (B quota) for the amount that exceeded the basic quantity (ca. 20–30 % of the A quota). If the EC suffered financial losses due to exporting part of the quota, this was reapportioned among beet growers and sugar factories (production tax). If the sugar production of a company exceeded the entire quota amount, this sugar (C sugar) had to be sold within a year outside the EC and was the sole responsibility of the company.

Since the 1980s, development of the world sugar market has led to tougher competition among previously nationally oriented sugar industries. Antiquated factories have been shut down or modernized with high investments. As a result of transport costs, factories are now located in cultivation areas. For the same reasons, the size of factories is kept within limits.

In 2006, this tougher competition led to a reform of the EU sugar regime, which was published as decree No. 318/2006 EC of February 20, 2006 and which came into force on July 1, 2006, when the former EU directive expired. This reform was necessary especially after a complaint by Australia, Brazil, and Thailand about increased EU sugar exports with the aid of subsidies. This complaint was confirmed by a WTO panel in 2004 and again in 2005. With the new regime, the intervention system was replaced by a white sugar reference price, which is to fall by 36 % until 2009 according to decree No. 318/2006 EC. A number of ACP (African, Carribean, Pacific) and LDC (least developed countries) states as well as European countries will be unable to produce sugar at this low price. Thus, a restructuring fund was installed to encourage uncompetitive sugar producers to leave the industry, in order to reduce the pressure on the EU domestic market. For sugar beet farmers a minimum price, to fall by 20 % after three years, and a monetary compensation were fixed. A and B quotas were combined and new quotas were fixed, based on 2005/06 data [261]. The new regime will ensure a sustainable future for the remaining EU sugar production, but will cause large changes for individual countries.

Regulations for the production of sugar (prices, quotas) similar to those of the EC also exist in other countries, such as the United States, India, and Brazil.

Consumption per person in a series of countries is listed in Table 18 [262].

The world market price of sugar undergoes large fluctuations, as shown in Table 17 and Figure 17.

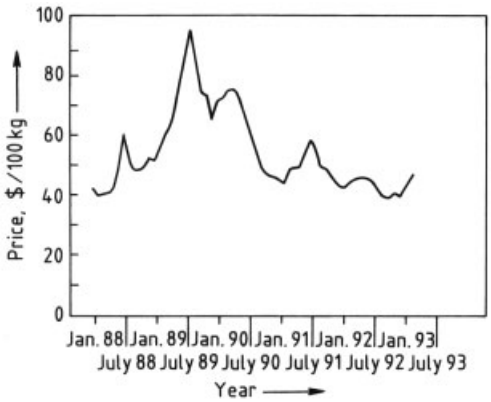

Figure 17. World market and prices for white sugar, monthly values Paris 1988–1993 [263]

To stabilize production and prices, the sugar industries of Cuba, Japan, Poland, Hungary, Bel-

Table 17. Development of worldwide sugar production [260]

	Raw sugar, 10^6 t	Cane sugar, %	Average annual quotation of the New York stock exchange, cents/lb
1900	11.26	47.0	2.88
1910	16.59	48.2	2.85
1920	16.41	70.5	11.35
1930	28.35	58.0	1.47
1940	30.86	62.3	1.15
1950	39.37	64.3	4.98
1960	61.15	60.3	3.14
1965	71.12	61.8	2.03
1970	82.29	64.2	3.68
1975	92.19	64.6	20.43
1980	98.94	66.9	16.89
1985	111.34	66.3	4.03
1990	128.48	67.1	12.55
1995	136.11	73.1	13.28
2000	143.38	74.3	8.18
2001	147.87	77.7	8.64
2002	159.99	76.7	6.84
2003	156.06 prel.	77.8	7.05
2004	155.10 prel.	76.8	7.12

gium, Germany, and Czechoslovakia signed an agreement on May 9, 1931, in Brussels (Chadbourne agreement), which stipulated a limitation of sugar production in subsequent years. The supply (production and sales of surplus) should not exceed the demand. The international sugar agreements of 1937, 1953, 1958, and 1968 had the same goal.

To increase the export proceeds of developing countries and stabilize international trade, a new international sugar agreement came into effect in 1977, which was signed by 40 sugar exporting countries and 15 sugar importing countries and was the last one containing direct market intervention instruments or so-called economic clauses [260]. This agreement was ratified by the United States in 1980 and the EC participated with constructive proposals, but was never a member of the agreement. The aims of the agreement should be achieved by export quotas in combination with reserves, which should be used in the case of increasing demand and refilled in the case of falling prices. A minimum price for raw sugar of 13 cents per pound and a maximum price of 23 cents per pound were agreed upon. The world market price was in this desired range for only a relatively short time. An agreement on the financing of reserve storehouses could not be reached. Consequently, the intended stabilization of world market prices was not achieved.

This agreement was replaced by a so called administrative agreement, ratified in 1984, 1987, and–after withdrawal of the United States–by a service agreement, signed in 1993, 1995, 1997, 1999, etc., with a possibility of indefinite extensions, each valid for two years. The aims now are merely to intensify international cooperation in sugar production and sales, to serve as a forum for international consultations on sugar and on the possibilities of promoting the world sugar economy, to facilitate trade by recording and making available information on the world market for sugar and other sweeteners, and to promote the demand for sugar for new purposes.

In the past, beet and cane sugar and, to a small extent, starch hydrolysates (glucose syrup, glucose) competed with each other on the market for nutritive sweeteners. Starch hydrolysates are used for technological reasons in the industrial production of certain foods. Since 1975, HFCS (isoglucoses, → Glucose and Glucose-Containing Syrups) has appeared on the market as a direct competitor of "liquid sugar" in the United States, Canada, Japan, and some European countries. In 1991, the production of HFCS amounted to 7.8×10^6 t (solids), ca. 7 % of the world sugar market. These amounts substantially influence the sugar market worldwide.

Apart from the market for nutritive sweeteners, a market has developed for calorie-free (intense) sweeteners (→ Sweeteners), calorie-reduced (polyols, inulin hydrolysates, synthetic

Table 18. Sugar consumption (in kilogram per person) in some industrialized and nonindustrialized countries, 1991 [262]

Area	Consumption	Area	Consumption
World	18.9	South America	38.1
Europe	35.0	Africa	12.6
North America	29.9	Asia	11.1
Industrialized countries		**Nonindustrialized countries**	
Hungary	55.6	Cuba	82.0
Iceland	49.5	Swaziland	63.1
Israel	47.2	Columbia	61.9
Czechoslovakia	47.2	Costa Rica	54.9
Ireland	45.7	Fiji	54.2
New Zealand	44.9	Gambia	52.3
Australia	44.1	Mexico	47.7
Switzerland	42.1	Panama	46.6
Austria	41.2	Surinam	44.3
Sweden	40.2	Brazil	43.7
United Kingdom	38.2	Belize	42.6
Norway	38.2	Persian Gulf	38.4
Netherlands	37.9	Lebanon	38.2
Canada	37.5	Guatemala	37.5
Denmark	36.6	Libya	37.1
France	35.7	Ecuador	35.7
Germany	35.1	Malaysia	35.1
South Africa	34.9	Jordan	33.7
Portugal	31.7	Argentina	32.1
Belgium–Luxembourg	29.3	Honduras	31.5
United States	29.1	Algeria	31.3
Italy	27.0	Syria	30.8
Spain	26.5	Egypt	29.3
Japan	21.2	Thailand	29.0
		Pakistan	28.0
		Saudi Arabia	24.0
		Tunisia	23.0
		Iran	22.4
		Yemen	21.5
		South Korea	16.6
		Kuwait	16.5
		India	13.2
		Senegal	13.0
		People's Republic of China	5.7
		Somalia	4.3
		Ethiopia	2.8
		Afghanistan	2.2
		Liberia	2.0
		Nepal	1.9
		Central African Republic	0.9
		Cambodia	0.5

polydextroses) sweeteners, and sweeteners with a definite nutritional physiological purpose [e.g., suitable for diabetics, noncariogenic sweeteners (fructose, → Fructose; polyols, → Sugar Alcohols)]. The percentages of individual product groups fluctuate greatly from country to country: for example, for intense sweeteners, ca. 10 % of the entire sweetener market is in the United States, 5 % in Germany, and < 0.5 % in developing countries [263].

24. References

General References

1. P. W. van der Poel, H. Schiweck, T. K. Schwartz (eds.): *Sugar Technology*, Verlag Dr. A. Bartens, Berlin, Engl. ed. 1998, German ed. 2000.
2. J. L. Multon (Coordin.): *Le Sucre, Les Sucres, Les Édulcorants et les Glucides de Charge dans les I.A.A.*, TEC & DOC – Lavoisier, Paris 1992.

3. O. Bia: *Tecnologia e Impianti Industrial Saccariferi*, Gruppo Ferruzzi Publishing, Eridania Béghin-Say 1992.
4. S. Marie, J. R. Piggott (eds.): *Handbook of Sweeteners*, Blackie and Son Ltd., Glasgow London 1991.
5. N. L. Pennington, C. W. Baker (eds.): *Sugar, A User's Guide to Sucrose*, Van Nostrand Reinhold, New York 1990.
6. G.-W. von Rymon Lipinski, H. Schiweck (eds.): *Handbuch Süßungsmittel*, Behr's Verlag, Hamburg 1990.
7. J. C. P. Chen, Ch. Ch. Chou (eds.): *Cane Sugar Handbook*, 12th ed., J. Wiley & Sons, New York 1993.
8. R. A. McGinnis: *Beet-Sugar Technology*, 3rd ed., "Beet Sugar Development Foundation," Fort Collins, Colorado 1982.
9. Autorenkollektiv: *Die Zuckerherstellung*, 2nd ed., VEB Fachbuchverlag, Leipzig 1980.
10. K. Vukov: *Physik und Chemie der Zuckerrübe als Grundlage der Verarbeitungsverfahren*, Akadémiai Kiadó, Budapest, russ. ed. 1972; *Physics and Chemistry of Sugar-Beet in Sugar Manufacture*, engl. ed. 1977.
11. P. M. Silin: *Technology of beet-sugar production and refining*, Israel Programme for Scientific Translations, Jerusalem 1964: russ. ed. PPI, Pishchepromizdat 1958.
12. P. Honig: *Principles of Sugar Technology*, Elsevier, Amsterdam vol. 1, 1953; vol. 2, 1959; vol. 3, 1963.
13. H. M. Pancoast, W. R. Junk: *Handbook of Sugars*, 2nd ed., AVI, Westport, CT, 1980.
14. M. A. Clarke, M. A. Godshall (eds.): *Chemistry and Processing of Sugarbeet and Sugarcane*, Elsevier, Amsterdam, 1988.
15. V. E. Baikow: *Manufacture and Refining of Raw Cane Sugar*, Elsevier, Amsterdam 1981.

Specific References

16. P. G. Morel du Boil, *Int. Sugar J.* **99** (1997) 102–106.
17. J. Bricout, J. C. Fontes, *Sugar J.* **39** (1977) 31–32.
    B. N. Smith, *Naturwissenschaften* **62** (1975) 390.
    A. Rossman, W. Rieth, H.-L. Schmidt, *Z. Lebensm. Unters. Forsch.* **191** (1990) 259–264.
18. W. R. Sponholz in G. Würdig, R. Woller (eds.): *Chemie des Weines*, Verlag Eugen Ulmer, Stuttgart 1989, pp. 46–59.
    K. Mengel: *Ernährung u. Stoffwechsel der Pflanze*, Gustav Fischer Verlag, Jena 1991.
19. L. W. Doner in H. F. Linkens, J. F. Jackson (eds.): *Analysis of Nonalcoholic Beverages*, Springer Verlag, Berlin 1988, pp. 120–133.
    G. J. Martin, B. L. Zhang, M. L. Martin, M. L. Dupuy, *Biochem. Biophys. Res. Commun.* **111** (1983) 890–896.
20. H. D. Wallenstein, B. Schubbar, *Z. Zuckerind.* **14** (1964) 621–624, 674–676.
    K. Vukov, G. Pátkai, J. Monszpart-Sényi, *Z. Zuckerind.* **27** (1977) 792–795.
    J. Ehrenberg, *Zucker* **30** (1977) 612–619.
21. J. Tschersich, *Z. Zuckerind.* **17** (1967) 638.
22. J. H. Hulse, E. M. Laing, O. E. Pearson: *Sorghum and the Millets: Their Composition and Nutritive Value*, Academic Press, New York 1980.
    P. Angelle, *Sugar J.* **44** (1982) no. 10, 17–18.
23. C. O. Willits, C. H. Hills: "Maple Syrup Producers Manual," *Agriculture Handbook*, **no. 134**, Agricultural Res. Service, USDA, Washington 1976.
24. G. W. Walawalkar, *Sharkara* **4** (1961) 20.
    N. Gopinathan, *Int. Sugar J.* **64** (1962) 9.
    T. Yamane in P. Honig (ed.): *Principles of Sugar Technology*, vol. **3**, Elsevier, Amsterdam 1963, pp. 507 ff.
25. Federal Register, 53, no. 215, Nov. 7. Washington 1988, 6.44862.
    E. H. Glinsmann, H. Irausquin, Y. K. Park (FDA), *J. Nutr. Suppl.* **116** (1986)
    Committee on Medical Aspects of Food Policy: *Dietary Sugars and Human Disease*, London 1989, pp. 1–44.
    K. H. Bässler: *Z. Ernährungswiss. Suppl.* **29** (1990).
26. E. Nägele, *Zuckerindustrie* **130** (2005) 687–693.
27. E. O. von Lippmann: *Geschichte des Zuckers*, 2nd. ed., Springer Verlag, Berlin 1929.
    J. Baxa: *Die Zuckererzeugung 1600–1850*, Fischer-Vlg., Jena 1937.
    J. Baxa, G. Bruhns: *Zucker im Leben der Völker*, Vlg. Dr. A. Bartens, Berlin 1967.
    S. W. Mintz: *Sweetness and Power, The Place of Sugar in the Modern History*, Penguin Books, New York 1986, German ed., Campus Verlag, Frankfurt 1987.
28. K. Maier, O. Baron, J. Bruhns: *Zuckerwirtschaft Europa 2005* (Statistik Welt), Verlag Dr. A. Bartens, Berlin 2004, pp. 11–15.
29. E. O. von Lippmann in W. Ostwald (ed.): *Klassiker der exakten Wissenschaft*, no. 159, F. W. Engelmann, Leipzig 1907.
30. G. Vavrinecz: *Atlas der Zuckerkristalle*, Verlag Dr. A. Bartens, Berlin 1965, p. 8.

31. Z. Bubnik, P. Kadlec, D. Urban, M. Bruhns: *Sugar Technologists Manual*, Verlag Dr. A. Bartens, Berlin 1995, p. 116.
32. H. E. C. Powers, *Sugar Technol. Rev.* **1** (1969/70) 86–190.
G. Vavrinecz: *Atlas der Zuckerkristalle*, Vlg. Dr. A. Bartens, Berlin 1965.
F. H. C. Kelly, *Sugar Technol. Rev.* **9** (1982) 271–323.
33. R. W. Hartel, A. V. Selastry, *Crit. Rev. Food Sci. Nutr.* **1** (1991) 49–112.
34. G. Vavrinecz, *Z. Zuckerind.* **12** (1962) 481–487.
ICUMSA Methods Book, Verlag Dr. A. Bartens, Berlin 2005, sect. SPS-2 (1998), pp. 1–3.
35. ICUMSA Methods Book, Verlag Dr. A. Bartens, Berlin 2005, sect. SPS-4 (1998), p. 13.
36. J. F. Swindells, C. F. Snyder, R. C. Hardy, P. E. Golden, *Nat. Bur. Stand.* **Suppl. to Circ. No. C 440**, July (1958),
as quoted in R. S. Norrish, *Sci. Tech. Surv. Br. Food Manuf. Ind. Res. Assoc.* **51** (1967).
37. M. P. Maggean, J. U. Kristott, S. A. Jones: *Physical Properties of Sugars and their Solutions*, Leatherhead Food R. A., Surrey, no. 172, 1991.
38. A. C. Noble, N. L. Matysiak, S. Bonnans, *Food Technol.* **45** (1991) 121–126.
39. L. F. Wiggins, *Carbohydr. Chem.* **4** (1949) 293.
R. Khan, *Adv. Carbohydr. Chem. Biochem.* **33** (1976) 235–294.
J. L. Hickson (ed.) Sucrose Chemistry, *ACS Symp. Ser.* **41** (1977).
M. R. Jenner in *Developments in Food Carbohydrate*, vol. **2**, Applied Science Publ., London 1980.
C. E. James, L. Hough, R. Khan, *Prog. Chem. Org. Nat. Prod.* **55** (1989) 117–184.
F. W. Lichtenthaler (ed.): *Carbohydrates as Organic Raw Materials*, VCH Verlagsgesellschaft, Weinheim, Germany 1991.
F. Descotes (ed.): *Carbohydrates as Organic Raw Materials*, VCH Verlagsgesellschaft, Weinheim, Germany 1993.
F. W. Lichtenthaler, P. Pokinskyj, S. Immel, *Zuckerindustrie* **121** (1996) 174–190.
40. A. Durand et al.: "Abstracts," *VII International Biotechnology Symposium*, New Delhi 1984, pp. 464–467.
Südzucker AG, DE 3 812 612, 1988 (P. Nigam, M. Vogel).
41. M. Bruder: Solid-State Fermentation of Sugar Beet Pulp; A New and Promising Industrial Process, British Sugar Techn. Conference, Eastbourne 1990. F. Lyvan, company brochures, 14 630 Cagny/France.
42. A. Pictet, H. Vogel, *Helv. Chim. Acta* **11** (1928) 436.
43. J. Levi, C. B. Purves, *Adv. Carbohydr. Chem.* **4** (1949) 28, 31.
44. R. U. Lemieux, G. Huber, *J. Am. Chem. Soc.* **75** (1953) 4118.
R. K. Ness, H. G. Fletscher, *Carbohydr. Res.* **17** (1971) 465–470.
45. M. Doudoroff, H. H. Barker, W. J. Hassid, *J. Biol. Chem.* **168** (1947) 725.
G. Avigad: "Plant Carbohydrates I," in F. A. Loewus, W. Tanner (eds.): *Encyclopedia of Plant Physiology*, vol. **13 A**, Springer Verlag, Berlin 1982, pp. 217–347.
46. W. Z. Hassid, M. Doudoroff, *Adv. Carbohydr. Chem.* **5** (1950) 29 ff.
47. W. Oltmann, M. Burba, G. Bolz, *Beih. Z. Pflanzenzücht.* **12** (1984) 36–41.
N. J. Kruger in D. T. Dennis, D. H. Turpin (eds.): *Plant Physiology, Biochemistry and Molecular Biology*, Longman, New York 1990, pp. 59 ff.
48. R. Giaquinta, *Ann. Rev. Plant Physiol.* **34** (1977) 347–387.
49. R. A. Saftner, R. E. Wyse, *Plant Physiol.* **66** (1980) 884–889.
50. C. E. Hartt, *Rep. Hawaii. Sugar Technol.* **22** (1963) 151, 157.
51. F. Schneider: *Sugar Analysis*, ICUMSA Methods, ICUMSA, Peterborough 1979, pp. 2–3.
52. *ICUMSA Methods Book*, Verlag Dr. A. Bartens, Berlin 2005, loose-leaf ed. 2005, sect. SPS-1, p. 3.
53. *ICUMSA Methods Book*, Verlag Dr. A. Bartens, Berlin 2005, loose-leaf ed. 2005, sect. SPS-3, pp. 1–8.
54. H.C.S. de Whalley, *ICUMSA Methods of Sugar Analysis*, Elsevier, Amsterdam-London-New York 1964.
F. Schneider: *Sugar Analysis, ICUMSA-Methods*, ICUMSA, Peterborough, 1979.
55. *ICUMSA Methods Book*, Verlag Dr. A. Bartens, Berlin 2005, last loose-leaf ed.
56. D. A. Cooke, R. K. Scott: *The Sugar Beet Crop*, Chapman & Hall, London 1993.
B. Märländer: *Zuckerrüben*, Bernhard-Patzold Druckerei & Verlag, Stadthagen 1991.
C. Winner: *Zuckerrübenanbau*, DLG-Verlag, Frankfurt/M. 1981.

57. K. Düring, *Nachr. Chem. Tech. Lab.* **41** (1993) 233.
J. B. Nichols, C. C. Dalton, G. A. Todd, N. W. Broughton, *Zuckerindustrie* **117** (1992) 797–800.
58. G. Büttner, M. E. Führer Ithurrart, J. Buddemeyer, *Zuckerindustrie* **127** (2002) 856–866.
B. Märländer, I. Rothe, *Zuckerindustrie* **130** (2005) 482–486.
59. M. Burba, E. Schulze, *Zuckerindustrie* **106** (1981) 303.
M. Burba, W. Haufe, *Z. Zuckerind.* **22** (1972) 75.
60. G. Pollach, W. Hein, G. Rösner, *Zuckerindustrie* **121** (1996) 332–344.
61. B. Märländer, *Zuckerindustrie* **118** (1993) 131.
62. M. Burba, *Z. Zuckerind.* **26** (1976) 647–658.
P. V. Schmidt: *Zuckerrüben-Lagerung, -Aufbereitung*, Verlag Dr. A. Bartens, Berlin 1990.
63. H. Schiweck, *Zucker* **23** (1970) 559–567; *Ind. Secera* **27** (1973) 3–12.
64. F. X. Kammerer, H. J. Delavier, *Z. Zuckerind.* **17** (1967) 349, 419, 464.
65. Salzgitter Stahlbau AG, DE-OS 2 461 486, 1974 (K. Kutzke, H. W. Rose).
66. G. Pollach, W. Hein, G. Rösner, *Zuckerindustrie* **124** (1999) 622–637.
67. L. Lambrechts, R. Wibon, A. Simonart, *Sucr. Belge* **87** (1967) 139.
M. Devilliers, *Sucr. Fr.* **108** (1967) 289, 359.
E. Reinefeld, *Zucker* **28** (1975) 397.
68. D. Sargent, S. Briggs, S. Spencer, *Zuckerindustrie* **122** (1997) 615–622.
69. P. Valentin, *Zucker* **27** (1974) 305–312.
P. S. Worthington, *Sucr. Belge Sugar Ind. Abstr.* **101** (1982) 403–404.
70. L. Rigo, *Zuckerindustrie* **127** (2002) 31–44.
71. H. Wunsch, *Zuckerindustrie* **130** (2005) 405–409.
72. R. Hurasky, *Zuckerindustrie* **107** (1982) 926–930.
73. C. Nähle, *Zuckerindustrie* **114** (1989) 703–705.
J. P. Lescure, P. Bourlet, D. Verrier, G. Albagnac, *Sugar Technol. Rev.* **13** (1986) 179–231.
74. H. Schiweck, C. Nähle, *Zuckerindustrie* **115** (1990) 639–674.
75. H. Hartmann, H. van Malland, R. Hies, T. von Döring, *Zuckerindustrie* **111** (1986) 634–642.
76. K. Vukov: *Physics and Chemistry of Sugar Beet in Sugar Manufacture*, Akademiai Kiado, Budapest 1977.
L. Drath, R. Strauß, H. Schiweck, *Zuckerindustrie* **109** (1984) 993–1007.
77. G. V. Genie, *Sugar Technol. Rev.* **9** (1982) 119–270.
H. Brüniche-Olsen: *Solid-Liquid Extraction*, Nys. Nordisk Farlag, Copenhagen 1962.
G. V. Genie, *Int. Sug. J.* **96** (1994) 124–128, 160–162.
78. F. Schneider, H. P. Hoffmann-Walbeck: *Zucker Beihefte* **2** (1953) 1.
79. K. J. Demeter, W. Kundrat, H. Neumann, *Zucker* **9** (1956) 5–8.
H. Klaushofer, E. Parkkinen, *Z. Zuckerind.* **15** (1965) 445–448.
G. Pollach, W. Hein, A. Leitner, P. Zöllner, *Zuckerindustrie* **127** (2002) 530–537.
80. G. Pollach, W. Hein, D. Beddie, *Zuckerindustrie* **129** (2004) 555–564.
81. H. Schiweck, L. Büsching, *Zucker* **25** (1972) 7–12.
T. Cronewitz, H. Schiweck, R. Strauß, *Zuckerindustrie* **106** (1981) 127–135.
82. G. Haska, R. Nystrand, *Sucr. Belge Sugar Ind. Abstr.* **101** (1982) 131–144.
R. Nystrand, *Zuckerindustrie* **110** (1985) 693–698.
83. F. Schneider, E. Reinefeld, *Zucker* **13** (1960) 460–471.
84. H. Schiweck, *Zucker* **24** (1971) 523–537.
85. M. Shore et al., *Int. Sugar J.* **85** (1983) 6–10, 43–48, 76–81.
T. Cronewitz, *Zuckerindustrie* **114** (1989) 30–39.
86. F. Schneider, *Zucker Beihefte* **1** (1950) 1.
87. F. Schneider, E. Reinefeld, D. Schliephake, *Chem. Ing. Tech.* **35** (1963) 567.
88. G. V. Genie, *Zuckerindustrie* **104** (1979) 827–832.
89. R. Pieck, W. Loop, *Zucker* **28** (1975) 236–241.
90. H. Eichhorn, *Zuckerindustrie* **116** (1991) 329–358.
91. F. Hollaus, G. Pollach, *Zuckerindustrie* **111** (1986) 1025–1030.
92. K. Seidel, DT 824 929, 1949.
Simco Incorp., US 2 943 004, 1958 (E. E. Hanry).
93. IG-Farben Ind., DT 596 091, 1929 (K. Vierling).
F. Uhde GmbH, DE 962 420, 1954 (G. Hingst). Südzucker AG,
Uhde GmbH, DE 1 139 800, 1959 (A. Bauer, R. Weidenhagen, H. Schiweck).
94. H. Brüniche-Olsen, *Zucker* **7** (1954) 471–475.
95. M. Tyczewski, *Zuckererzeugung* **4** (1960) 250, 282, 317, 349, 377.

96. P. Mathismoen, J. Ehrenberg, G. Lampe, S. Matusch, *Zuckerindustrie* **106** (1981) 965–981.
H. Lührs, J. Ehrenberg, T. Rosmer, H. J. Krombach, *Zuckerindustrie* **112** (1987) 571–581.
97. G. Witte, *Zuckerindustrie* **125** (2000) 311–322.
98. R. F. Madsen, W. Kofod Nielsen, B. Winstrom-Olsen, T. E. Nielsen, *Sugar Technol. Rev.* **6** (1978/79) 49–115.
99. G. Vaccari et al., *Zuckerindustrie* **117** (1992) 724–728.
100. H. Schiweck, *Zucker* **29** (1976) 549–556.
101. E. O. v. Lippmann, *Dtsch. Zuckerind.* **62** (1937) 67, 87, 105, 125, 151.
102. H. Schiweck, *Zucker* **28** (1975) 670–674; *Sugar J.* **47** (1985) 18–22; *LC Listy Cukrov.* **106** (1990) 176–186.
E. Reinefeld, D. Miehe, *Zuckerindustrie* **113** (1988) 15–20.
R. F. Madsen, *Zuckerindustrie* **113** (1988) 33–36.
P. W. van der Poel, N. H. M. de Visser, *Zuckerindustrie* **113** (1988) 22–26, *Zuckerindustrie* **115** (1990) 943–949.
103. F. H. C. Kelly, D. W. Brown, *Sugar Technol. Rev.* **6** (1978/79) 1–46.
F. Ledl, E. Schleicher, *Angew. Chem.* **102** (1990) 597–626; *Angew. Chem. Int. Ed. Engl.* **29** (1990) 565–594.
104. E. Reinefeld, K.-M. Bliesener, J. Schulze, *Zuckerindustrie* **107** (1982) 283–291, 1111–1119.
105. F. A. Baczek, V. M. Jesic, *Zucker* **27** (1974) 475–476. L. Büsching, *Zucker* **30** (1977) 595–601.
106. De Danske Sukkerfabrikker, DT 917 663, 1950.
107. Selwig & Lange Maschinenfabrik, DE 2 833 348, 1978 (H. Ahlers, P. Mosel, S. Matusch).
P. Mosel, S. Matusch, *Zuckerindustrie* **104** (1979) 1033.
S. Matusch, *Zuckerindustrie* **113** (1988) 27–29.
108. Südzucker AG, DT 1 069 539, 1958 (L. Kayer et al.).
109. G. Pollach, W. Hein, G. Rösner, *Zuckerindustrie* **128** (2003) 358–363.
110. W. Uhlenbrock, *Zuckerindustrie* **113** (1988) 30–32.
111. W. Napp, T. Cronewitz, G. Witte, *Zuckerindustrie* **112** (1987) 859–867.
S. Matusch, *Zuckerindustrie* **113** (1988) 27–29.
112. Süddeutsche Zucker-AG, DE 2 925 283, 1979 (G. Witte, H. Schiweck).
113. H.-J. Praus, *Zucker* **19** (1966) 38.
R. F. Madsen, *Sugar Technol. Rev.* **1** (1969) 61–83.
114. H. Weidner, *Zuckerindustrie* **103** (1978) 311–313.
115. O. Spengler, St. Böttger, *Z. Ver. Dtsch. Zucker Ind.* **83** (1933) 331.
116. F. Schneider, *Zucker* **12** (1959) 326.
Braunschweiger Maschinenbauanstalt AG, DT 921 980, 1952. F. Schneider, G. Fischer, G. Baumgarte DT 1 013 590, 1954;DT 1 043 979, 1957.
117. F. Schneider, G. Baumgarten DT 1 048 840, 1957. G. Schäfer, *Zuckerindustrie* **115** (1990) 371–375.
118. A. V. Brieghel-Müller: "Vortrag beim IX. Internat. Kongreß d. landwirtsch. Industrien in Rom", *Z. Zuckerind.* **2** (1952) 194.
R. F. Madsen, *Zuckerindustrie* **113** (1988) 33–36.
119. G. Quentin, *Zucker* **17** (1964) 336; **18** (1965) 623.
120. G. Steinle, *Zucker* **25** (1972) 81.
121. G. Witte, T. Cronewitz, *Zuckerindustrie* **113** (1988) 479–487.
E. Wittwer, *Zuckerindustrie* **108** (1983) 218–226.
122. S. Zagrodzki, A. Kubasiewicz, *Sugar Technol. Rev.* **5** (1977/78) 1–54.
123. K. Bohn: "Untersuchungen über die Saccharosezerstörung bei Temperaturen über 100 °C," Vortrag bei der *II. Int. Conf. on the Chemistry and Technology of Sugar,* Lódź, 14–15 June, 1973. M. Athenstedt, *Z. Zuckerind.* **13** (1963) 563. K. Vukov, G. Patkai, *Zuckerindustrie* **106** (1981) 314. K. Vukov, I. Körmendy, H. M. Loko: Paper presented at the 17. CITS-Conference, Copenhagen, 30 May–3 June, 1983.
124. H. Schiweck, *Zuckerindustrie* **116** (1991) 793–805.
125. B. Morgenroth, H. Daschmann, K. Niepoth, G. Abker, B-C. Schulze, *Zuckerindustrie* **123** (1998) 597–606.
126. K. Urbaniec: *Modern Energy Economy in Beet Sugar factories*, 1st ed., Elsevier, Amsterdam 1989.
A. Baloh: *Energiewirtschaft in der Zuckerindustrie*, vol. **1**, Verlag Dr. A. Bartens, Berlin 1991.
D. Schliephake, *Zuckerindustrie* **114** (1989) 193–202.

127. G. Trabanelli, G. Mantovani, F. Zucchi, *Sugar Technol. Rev.* **4** (1976/77) 131–208; *Sugar Technol. Rev.* **14** (1988) 1–28.
128. H. Schlegel, *Z. Zuckerind.* **13** (1963) 14.
129. K. E. Austmeyer, *Chem. Ing. Tech.* **53** (1981) 716–719.
130. D. Schliephake, *Zucker* **18** (1965) 138.
131. L. J. Kuijvenhoven, I. J. Risseeuw, *Int. Sugar J.* **85** (1983) 35–38.
132. D. Schliephake, F. A. Orlowski, *Chem. Ing. Tech.* **43** (1971) 676–682.
133. K. E. Austmeyer, K. Kipke, *Zuckerindustrie* **105** (1980) 227–234.
R. F. Madsen, *Zuckerindustrie* **105** (1980) 234–237.
P. W. van der Poel, *Zuckerindustrie* **105** (1980) 237–240.
L. Bachan, R. R. Sanders, *Proc. SASTA* **61** (1987) 65–69.
S. K. Helfels, E. J. de Jong, *Chem. Eng. Tech.* **59** (1987) 443.
134. H. Thiele, A. Langen, *Int. Sugar J.* **76** (1974) 136–146, 169–173.
W. Böhm, *Zucker* **26** (1973) 21.
D. Schliephake, K. E. Austmeyer, *Zucker* **28** (1975) 546.
135. D. J. Radford, D. J. Tayfield, M. G. S. Cox, *Int. Sugar J.* **91** (1989) 52.
136. R. Hempelmann, K. E. Austmeyer, *Zuckerindustrie* **112** (1987) 695.
137. T. Frankenfeld, *Zuckerindustrie* **113** (1988) 968.
138. G. V. Genie, *Z. Zuckerind.* **12** (1962) 557.
139. Z. Bubnik, P. Kadlec, *Zuckerindustrie* **121** (1996) 36–39.
140. H. Schiweck, M. Munir, G. Witte, *Zuckerindustrie* **112** (1987) 681–693.
K. E. Austmeyer, T. Frankenfeld, *Zuckerindustrie* **112** (1987) 36–45.
141. T. Kaga, *Seito Gijutsu Kenkyu Kaiski* **10** (1961) 27–38.
A. J. Gromkovskij, V. M. Fursov, *Izv. Vyssh. Uchebn. Zaved. Pishch. Tekhnol.* **4** (1977) 157–162;
Sugar Ind. Abstr. **39** (1977) 169.
142. E. A. Knøvl, G. R. Møller, *Sugar Technol. Rev.* **3** (1975/76) 275–309.
G. R. Møller, *Int. Sugar J.* **86** (1984) 73–79.
143. G. Witte, *Zuckerindustrie* **113** (1988) 414–420.
P. Mentzel, P. Valentin, D. Wagner, G. Witte, *Zuckerindustrie* **115** (1990) 751–760. E. E. A. Rouillard:
Proc. Annu. Congr. S. Afr. Sugar Technol. Assoc. **61** (1987) June.
St. Jeschka, M. Lorenz, *Zuckerindustrie* **126** (2001) 448–455.
144. M. Schneider, H. Schiweck, *Zuckerindustrie* **111** (1986) 907–916.
145. K. E. Austmeyer, D. Schliephake, B. Ekelhof, G. Sittel, *Zuckerindustrie* **114** (1989) 875–878.
146. O. Bia, *Zuckerindustrie* **112** (1987) 518–524.
147. K. Helfels, A. Pot, C. H. de Nie, E. J. de Jong, *Zuckerindustrie* **111** (1986) 845–850.
148. H. Eichhorn, *Zuckerindustrie* **112** (1987) 114–122.
149. H.-J. Krombach, *Zuckerindustrie* **106** (1981) 793–804.
150. H. Schiweck, *Ind. Sacc. Ital.* **73** (1980) 71–78.
R. A. Mc Ginnis, *Sugar Technol. Rev.* **5** (1978) 155–286.
R. F. Madsen, W. Kofod Nielsen, *Sugar Technol. Rev.* **7** (1979/80) 49–85.
151. D. Schliephake, B. Ekelhof, *Zuckerindustrie* **108** (1983) 1127–1138.
152. G. Vaccari, M. Mantovani, G. Squoldino, D. Aquilano, M. Rubbo, *Sugar Technol. Rev.* **13** (1986) 133–178.
153. J. O. Bara, S. Hashimoto, *Sugar Technol. Rev.* **4** (1976/77) 209–258.
J. C. Linden, *Enzyme Microb. Technol.* **4** (1982) 130–136.
154. R. Mattes, K. Beaucamp, *Chem. Unserer Zeit* **17** (1983) 54.
155. T. Cronewitz, *Zucker* **30** (1977) 662.
W. Hunter, *Int. Sugar J.* **80** (1978) 358.
156. R. A. McGinnis, *Sugar Technol. Rev.* **5** (1978) 155–286.
G. Vavrinecz, *Sugar Technol. Rev.* **6** (1978/79) 117–129.
157. G. Sittel, *Zuckerindustrie* **125** (2000) 501–507.
158. T. Cronewitz, W. Jäkel, K. Korn, *Zucker* **29** (1976) 230–237.
159. K. E. Austmeyer, R. Marwede, *Zuckerindustrie* **112** (1987) 193–201.
A. Murphy, A. G. Punter, P. D. Thompson, *Int. Sugar J.* **93** (1991) 33–38, 52–55, 79–81, 86, 95–96, 97–98.
160. K. W. R. Schönrock in K. Dorfner (ed.): *Ion Exchanges in the Sweetener Industry in Ion Exchangers*, De Gruyter, Berlin 1991.
161. A. Feldhoff, *Zentralbl. Zuckerind.* **15** (1906/07) 1307.
162. E. W. Reed, J. S. Dranoff, *Ind. Eng. Chem. Fundam.* **3** (1964) 304.
E. Berghofer, H. Klaushofer, L. Wieninger, *Zucker* **30** (1977) 187–203.
163. J. D. Phillips, A. Pollard, *Nature (London)* **171** (1953) 41.

164. J. Ponant, *Sucr. Belge Sugar Ind. Abstr.* **101** (1982) 369–380.
165. W. Pannekeet, *Ind. Aliment. Agric.* **97** (1980) 757.
P. L. Mottard, *Int. Sugar J.* **85** (1983) 233–237.
166. Amalgamated Sugar Corp., US 3 982 956, 1975 (K. W. R. Schoenrock, A. Gupta, H. G. Rounds).
167. G. Quentin, *Zucker* **7** (1954) 407; **10** (1957) 408.
168. M. Kearney: "Simulated Moving Bed Technology Applied to the Chromatographic Recovery of Sucrose from Sucrose Syrups," *Proc. of the Sug. Processing Research Conf.*, American Society of Sugar Beet Technologists, Colarado Springs 1990, pp. 291–304.
V. J. Jaro, G. Pool, K. Schwarz: "Chromatographic Separator Extract Processing," *Proc. from the 26th Biennial Meeting of ASSBT*, American Society of Sugar Beet Technologists, Colarado Springs 1991, pp. 396–414.
N. N., *Zuckerindustrie* **108** (1983) 636–638.
H. G. Schneider, J. Mikule, *Int. Sugar J.* **77** (1975) 259–264, 294–298.
M. Munir, *Int. Sugar J.* **78** (1976) 100–106.
169. H. A. Paananen, *Zuckerindustrie* **122** (1997) 28–33.
170. D. Costesso, G. Rearick, M. Kearney, *Zuckerindustrie* **125** (2000) 333–335.
171. H. Paananen, F. Rousset, *Zuckerindustrie* **126** (2001) 601–604.
172. B. Ehrenhuber, *Zuckerindustrie* **130** (2005) 463–469.
173. H. Schiweck, *Zucker* **16** (1963) 555.
D. Hervé, *Ind. Aliment. Agric.* **90** (1973) 897.
174. G. Assalini, G. Brandoli, *Sucr. Belge Sugar Ind. Abstr.* **91** (1972) 107–111.
175. H. G. Schneider, *Int. Sugar J.* **73** (1971) 69–73, 100–104.
176. G. Vernois, *Z. Zuckerind.* **13** (1963) 39–40.
L. Stelzig, *Zucker* **16** (1963) 659–661.
177. A. Birch-Jensen: *Socker Handl. 2* **9** (1964) 35.
178. P. Betak, O. Mikus, *LC Listy Cukrov.* **98** (1982) 56–61.
179. Maschinenfabrik Buckau R. Wolf AG, DT 1 167 760, 1962 (H. Ludewig).
180. H. Melcher, DT 190 613, 1906, DT 212 589, 1908, DT 239 906, 1910.
181. G. Schneider, *Zucker* **26** (1973) 524–532.
D. H. Napier, H. W. Cerda: *Tech. Pap. Proc. Ann. Meet. Sugar Ind. Technol.*, 1978, 489–513.
EC Directive 89/392/EC.
182. D. Nahler, *Zucker* **30** (1977) 2–7.
D. Roth, *Zucker* **30** (1977) 274–284, 464–470.
183. Codex Alimentarius, Joint FAO/WHO Food Standard "Powdered Sugar", CAC/RS 5–1969 (rev.).
184. W. A. Ritschel: *Die Tablette*, Editio Cantor KG, Aulendorf 1966.
N. Pintauro: "Agglomeration Processes in Food Manufacture," *Food Processing Rev.*, **no. 25**, Noyes Data Corp., Park Ridge, USA 1972.
H. Schubert, *Chem. Ing. Tech.* **62** (1990) 892–906.
185. Pfeiffer & Langen, DT 1 275 853, 1968 (A. Langen, H. Thiele).
Süddeutsche Zucker AG, DT 1 800 141, 1968 (H. Schiweck).
186. Codex Alimentarius, Joint FAO/WHO Food Standard "Soft sugars", CAC/RS 6–1969. [13], p. 118.
187. European Economic Community, Council Directive of 11 Dec., 1973 (73/437/EC), Off. J. E. C. no L 356, 27.12.73, pp. 71–78;First Commission Directive of 26 July, 1979 (79/786/EC), Off. J. E. C. no L 239, 22.9.79, pp. 24–52.
188. US Food Chemical Codex Standards-Sucrose 57–50–1 Draft Nov. 1990. US Food Chemical Codex Standards-Invert Sugar 8013–17–0 Draft Feb. 1991.
189. N. Marignetti, G. Mantovani, *Sugar Technol. Rev.* **7** (1979/80) 3–47.
R. D. Moroz, J. P. Sullivan, J. P. Troy, C. B. Broeg, *Sugar Azucar* **68** (1973) 46–52.
190. E. Thier, *Zuckerindustrie* **105** (1980) 80–82.
191. FAO *Food Nutr. Paper* **38** (1988) 31–42.
192. H. Schiweck, *Dtsch. Lebensm. Rundsch.* **76** (1980) 274–280.
193. O. Mikus, L. Budicek, *Sugar Technol. Rev.* **13** (1986) 52–123.
194. H. Schindler, B. Junckers, *Zuckerindustrie* **118** (1993) 103–106.
195. W. Graup, S. Buhrmann, H. Will, *Zuckerindustrie* **126** (2001) 331–335.
196. F. Schneider, D. Schliephake, *Zucker* **14** (1961) 575.
197. H. Schiweck, T. Cronewitz, G. Müller, *Zuckerindustrie* **103** (1978) 185–190.
198. E. Muhlack, U. Schröter, M. Kunz, W. Glanser, *Zuckerindustrie* **105** (1980) 481–484.
199. T. Cronewitz, G. Müller, B. Bissinger, F. Stadler, *Zucker* **28** (1975) 401–410.

200. A. Bausier, *Sugar Technol. Rev.* **8** (1981/82) 149–179.
D. Schröter, *Zuckerindustrie* **108** (1983) 126–135.
T. Cronewitz, *VDI Ber.* **345** (1979) 63–76; *Zuckerindustrie* **105** (1980) 129–139.
J. Spanbauer, *Eng. Food* **2** (1984) 819–825.
M. Kunz, P. Valentin, *Zuckerindustrie* **111** (1986) 741–750.
201. K. E. Austmeyer, W. Poersch, *Zuckerindustrie* **108** (1983) 861–868, 1033–1041, *Zuckerindustrie* **109** (1984) 411–419, *Zuckerindustrie* **110** (1985) 28–34.
W. Kunz, *Zuckerindustrie* **108** (1983) 868–870.
202. P. Valentin, *C. R. Assem. Gèn. Comm. Int. Tech. Sucr.* **18th** (1987) 187–216.
BMA: *Zuckerindustrie* **111** (1986) 357–358.
203. A. S. Jensen, J. Borreskov, D. K. Dinesen, R. F. Madsen, *C. R. Assem. Gèn. Comm. Int. Tech. Sucr.* **18th** (1987) 217–230; *Zuckerindustrie* **112** (1987) 886–891.
A. S. Jensen, *Zuckerindustrie* **124** (1999) 790–794, **125** (2000) 868–873.
204. A. Timmermans, *Zuckerindustrie* **126** (2001) 185–187.
205. E. Muhlack, *Zucker* **23** (1970) 589–595.
206. H. Schneidt, *Sugar Technol. Rev.* **3** (1975) 127–154.
207. H. Schiweck, J. Henatsch, F. Curtius, *Zuckerindustrie* **108** (1983) 835–847.
208. G. F. Kroneberger, *Sugar Technol. Rev.* **4** (1976/77) 3–47.
209. H. Schiweck, T. Cronewitz, F. Schoppe, *Zuckerindustrie* **104** (1979) 813–819.
210. A. P. Draycott: *Sugar-Beet-Nutrition*, Applied Science Publ., London 1972, pp. 89–91.
H. Irion: "Nährstoff-Zusammensetzung und Wirkung von Carbokalk auf Boden und Pflanzen," *Proc. 51th IIRB-Winter-Congress, Brussels* 1988, pp. 231–241.
I. P. Vandergeten: " Formes actuelles des valorisation des ecumes des sucreries en Europe," *Proc. 51. IIRB-Winter-Congress*, Brussels 1988, pp. 209–218.
211. E. M. Hartmann, *Sugar Technol. Rev.* **2** (1974) 213–252.
212. R. Vandewijer, J. Jacques, R. Pieck, *Sucr. Belge Sugar Ind. Abstr.* **91** (1972) 187–197; **93** (1974) 163–165.
213. J.-C. Giorgi, R. Gontier, B. Richard, *Sucr. Fr.* **121** (1980) 93–101.
W. Kofod Nielsen, S. Kristensen, R. F. Madsen, *Sugar Technol. Rev.* **9** (1982) 59–117.
214. W. Steinmetzer, *Zuckerindustrie* **116** (1991) 30–39.
215. Suomen Sokeri, Finland, US 4 359 430, 1982.
216. K. Sayama, T. Kamada, S. Oikawa, T. Masuda, *Zuckerindustrie* **117** (1992) 894–898.
217. J. E. Irvine in G. P. Meade, J. C. P. Chen (eds.): *Cane Sugar Handbook*, 11th ed., J. Wiley & Sons, New York 1985, p. 13.
218. B. T. Egan, *Proc. Queensl. Soc. Sugar Cane Technol.* 1964, 15–26.
219. B. T. Egan, *Proc. Queensl. Soc. Sugar Cane Technol.* 1965, 25–30.
220. M. A. Clarke (ed.): *Proc. Int. Dextran Workshop*, Sugar Process. Res. Inc., New Orleans 1984.
221. G. R. Shaw, G. A. Brotherton, *Proc. Conf. Aust. Soc. Sugar Cane Technol.* 1992, 1–7.
222. B. L. Legendre, *Sugar J.* **54** (1992) no. 2, 2–12.
223. R. B. Trott: "Clarification and Decolorization processes," in M. A. Clarke, M. A. Godshall (eds.): *Chemistry and Processing of Sugarbeet and Sugarcane,* Elsevier, Amsterdam 1988, pp. 265–291.
224. M. A. Clarke, *Sugar y Azucar Yearbook* 1991, 6–11.
M. A. Clarke, *Zuckerindustrie* **121** (1996) 505–510.
225. M. C. Bennett et al., Proc. Int'l. Soc. Sugar Cane Technol. (1977), 2797–2805.
226. J. B. Alexander, *Proc. Congr. Int. Soc. Sugar Cane Technol.* **14** (1971) 1619–1675.
227. M. A. Clarke, *Sugar y Azucar* **80** (1985) 32–50.
228. O. Lyle: *Technology for Sugar Refinery Workers*, Chapman and Hall, London 1941.
229. P. Carreno: "Symposium on Raw Sugar Quality," *Proc. Sugar Ind. Technol.* **52** (1993) 95–174.
230. M. A. Clarke, *Sugar y Azucar Yearbook* 1983, 71–93.
231. J. C. Abram, J. T. Ramage: "Sugar Refining: Present Technology and Future Development," in G. G. Birch, K. J. Parker (eds.): *Sugar: Science and Technology,* Applied Science Publishing, London 1979, pp. 49–96.
232. J. Dransfield, *Principles* **20** (1976) 83–90.
233. H. J. Delavier, *Z. Zuckerind.* **17** (1967) 136.
G. Graefe in: *Handbuch der Lebensmittel-Chemie*, vol. **5/1**, Springer Verlag, Berlin 1967, pp. 671 ff.
234. Codex Alimentarius, Joint FAO/WHO Food Standards, White Sugar CAC/RS 4, 1969.
235. W. Mauch, E. Farhoudi, *Sugar Technol. Rev.* **7** (1979/80) 87–171.
236. T. Cronewitz, W. Serkezy, *Zucker* **26** (1973) 134–138.

237. E. Hansen, *Z. Lebensm. Unters. Forsch.* **98** (1954) 405–411.
238. H. Hallanoro, J. Ahvenainen, L. Ramm-Schmidt: "Saponin, a Cause of Foaming Problems in Beet Sugar Processing and Use", *Proc. of the Sugar Processing Research Conference,* San Francisco 1990, 174–201.
M. A. Clarke, E. J. Roberts, M. A. Godshall, *Zuckerindustrie* **122** (1997) 873–877.
K. K. Foong, R. Amal, W. O. S. Doherty, L. A. Edye, *Zuckerindustrie* **127** (2002) 382–388.
239. G. Steinle, E. Fischer, *Zuckerindustrie* **103** (1978) 129–131.
H. Schiweck, A. Kolber, *Gordian* **72** (1972) 41–51.
240. P. X. Hoynak, G. N. Bollenback: *This is Liquid Sugar, Refined Syrups & Sugars*, Yonkers, New York 1966, pp. 144, 148;
Microbial Ecology of Foods, vol. **2**, Academic Press, New York 1980, pp. 779–803.
241. P. B. Koster et al., *Zucker* **28** (1975) 555–562;
Zuckerindustrie **104** (1979) 49–53;
Zuckerindustrie **106** (1981) 895–900.
242. P. B. Koster, *Zuckerindustrie* **103** (1978) 200–203.
243. H. Schiweck, *Branntweinwirtschaft* **117** (1977) 87–93.
E. Sinda, E. Parkkinen: "Problems with Molasses in the Yeast Industry," Symposium 31 Aug.–1 Sept., 1979, Helsinki, Finland, Kauppakirjapaino oy Helsinki 1980.
244. H. Schiweck, C. Jeanteur-De Beukelaer, M. Vogel, *Zuckerindustrie* **118** (1993) 15–23.
245. R. Tressel, R. Jakob, T. Kossa, W. K. Bronn, *Branntweinwirtschaft* **116** (1976) 117.
246. E. Zauner, W. K. Bronn, H. Dellweg, R. Tressl, *Branntweinwirtschaft* **119** (1979) 154.
W. K. Bronn, G. E. Kim, N. Fattohi, *Branntweinwirtschaft* **130** (1990) 134–144;
Branntweinwirtschaft **131** (1991) 262–266.
N. Fattohi, *Zuckerindustrie* **119** (1994) 554–560,
847–853.
Zuckerindustrie **120** (1995) 587–592,
677–682.
247. In [7] pp. 408–427.
248. C. E. Moyer, *Sugar J.* **37** (1975) 20–23.
F. Cordovez, *Sugar J.* **45** (1983) 23.
C. Mantovani, G. Vaccori, *Sugar. J.* **56** (1994) no. 12, 16–19.
249. J. M. Paturau: "By-products of the Cane Sugar Industry," *Sugar Ser.*, vol. **3, 2nd ed.**, Elsevier, Amsterdam 1982, pp. 76–119.
250. Süddeutsche Zucker-AG und Industriewerke Karlsruhe, DT 2 011 257, 1970 (T. Cronewitz, W. Portale, H. Schiweck, D. Lohmann).
251. Süddeutsche Zucker-AG, DT 2 142 425, 1971 (H. Schiweck, T. Cronewitz).
252. Zuckerfabrik Franken GmbH, DT 2 156 131, 1971 (F. X. Kammerer).
253. J. Debruck, W. Lewicki, *Deutsche Zuckerrübenzeitung* **18** (1982).
254. *Liquid Ingredients Handbook NFIA*, National Feed Ingredients Association, USA, 1975.
255. F. Hollaus, W. Braunsteiner, N. Kubadinow, *Zuckerindustrie* **108** (1983) 1049–1058.
256. V. Bückmann, *Zuckerindustrie* **124** (1999) 519–521.
257. H. Bergner, E. Krieghoff, *Dtsch. Landwirtsch.* **17** (1966) 546.
258. Pfeiffer & Langen, DE-AS 1 229 826, 1965.
259. *Bilder-Conversations-Lexikon für das deutsche Volk*, vol. **4**, F. A. Brockhaus, Leipzig 1841, 817.
260. F. O. Licht: *Jubiläumsausgabe, Die Weltzuckerwirtschaft 1936–1961*, Selbstverlag, Ratzeburg 1962 pp. 42–45.
F. O. Licht: *Weltzuckerstatistik*, Selbstverlag., Ratzeburg 1962, 1968, pp. 14–18.
A. Bartens/Mosolff: *Zuckerwirtschaft–Europa 2006*, Verlag Dr. A. Bartens, Berlin 2006
261. U. Nöhle, *Zuckerindustrie* **129** (2004) 703–712.
Supplement to *Zuckerwirtschaft–Europa 2006*, Verlag Dr. A. Bartens, Berlin 2006.
262. J. Berghäuser, *Ernährungsumschau* **28** (1981) 422–425,
figures changed according to letter of 09.03.93.
263. H. Schiweck, *Zuckerindustrie* **117** (1992) 785–794.

Sugar Alcohols

Hubert Schiweck, Südzucker Aktiengesellschaft, Mannheim/Ochsenfurt, Obrigheim, Federal Republic of Germany (Chaps. 1, 6 – 8)

Albert Bär, Bioresco AG, Binningen, Switzerland (Chap. 2)

Roland Vogel, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 3)

Eugen Schwarz, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 4)

Markwart Kunz, Südzucker Aktiengesellschaft, Mannheim/Ochsenfurt, Obrigheim, Federal Republic of Germany (Chap. 5)

1. Polyols, General 2
2. Xylitol 4
2.1. Physical and Chemical Properties . 5
2.2. Production 5
2.3. Specifications, Analysis, and Legal Aspects 5
2.4. Uses 5
2.5. Metabolism, Tolerance, and Safety 6
2.6. Producers 6
3. Sorbitol 6
3.1. Physical and Chemical Properties . 7
3.2. Raw Materials and Production . . . 8
3.3. Environmental Protection 10
3.4. Analysis 10
3.5. Storage and Transportation 10
3.6. Legal Aspects 11
3.7. Uses 11
3.8. Physiology, Tolerance, Toxicology . 12
3.9. Producers 12
3.10. Economic Aspects 12
4. Mannitol 13
4.1. Physical and Chemical Properties . 13
4.2. Raw Materials and Production . . . 14
4.3. Analysis 15
4.4. Uses 15
4.5. Physiology, Tolerance, Toxicology . 15
4.6. Economic Aspects 15
5. Isomaltulose and Trehalulose, Isomalt 15
5.1. Isomaltulose and Trehalulose 15
5.1.1. Physical and Chemical Properties . . . 16
5.1.2. Production 16
5.1.3. Uses 17
5.1.4. Economic Aspects 17
5.2. Isomalt 17
5.2.1. Physical and Chemical Properties . . . 17
5.2.2. Production 18
5.2.3. Uses 18
5.2.4. Economic Aspects 19
6. Lactitol 19
6.1. Physical and Chemical Properties . 19
6.2. Physiological Properties 19
6.3. Technological Properties 20
6.4. Economic Aspects 20
7. Maltitol and Maltitol-Containing Syrups 20
7.1. Maltitol-Containing Syrups 20
7.2. Maltitol 21
7.3. Economic Aspects 22
8. Erythritol 22
9. References 22

1. Polyols, General

Historical Aspects, Nutrition Physiology. In the 1920s, the production and marketing of sugar alcohols (polyols) were begun to provide people who should not eat sugar (because they suffer from diseases, e.g., diabetics) with sweet foods such as confectionery, jams, hard boiled candies, and chocolates. Numerous studies have shown that the "sweet" taste is so important that no one wants to do without sweet-tasting products [7].

In the production of these foods, polyols function as a bulking agent (→ Sugar, Chap. 1.). The sweetness of the products is regulated by the addition of intense sweeteners (→ Sweeteners) [8].

Since the metabolism of polyols in humans is largely independent of insulin, polyols can be used in the production of foods for diabetics [9].

Furthermore, polyols have been found to be largely noncariogenic because they are not a substrate for plaque and acid formation in the oral

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_413

cavity. In fact, they stimulate salivation when consumed and, consequently, promote the remineralization of enamel. For this reason, polyols are suitable for the production of noncariogenic products, especially for consumption between meals (e.g., hard boiled candies, chocolates, biscuits). These products may be labeled "does not promote tooth decay" if they fulfill the following criterion: on consumption under stipulated experimental conditions, the pH of the plaque should not fall below 5.7 for a period of 30 min [10]. For pure polyols, the noncariogenicity (nonacidogenicity) decreases as follows: xylitol $\leq$ isomalt $<$ sorbitol, maltitol $<$ mannitol $\approx$ lactitol.

A third property that substantially differentiates polyols from sugar and corn syrups is their reduced physiological calorific value [11]. In the United States, the calorific value of mannitol is 2 kcal/g (ca. 8.3 kJ/g) [12]. In the EC, a physiological calorific value of 10 kJ/g is set for approved polyols: mannitol, sorbitol, xylitol, isomalt, lactitol, and maltitol [13]. An expert commission appointed by the Dutch government [14] and in 1994 an expert panel of the Life Science Research Office (LSRO) and the Federation of the American Societies for Experimental Biology (FASEB) [15] have evaluated the literature on this subject. The results are summarized in Table 1:

Table 1. Physiological calorific value of some polyols (in kJ/g)

Sugar alcohol	Nutrition Council 1987	LSRO Expert Panel 1994
Xylitol	15.0	~ 10.0
Sorbitol	12.0	7.5 – 13.8
Mannitol	8.5	6.7
Isomalt (Palatinit)	10.0	8.4
Lactitol	8.5	6.7 – 9.2
Maltitol	12.1	11.7 – 13.4
Maltitol syrup (see Sect. 7.1)		11.7 – 13.4

The exact value for the physiological energy yield of polyols is disputed scientifically, especially the question of whether dose or adaptation dependence exists [11], [16].

The reduced calorific value is due to the fact, that polyols are only partially absorbed in the upper intestine. Thus, a large part of the ingested polyols reaches the large intestine where it is degraded by bacteria to lower fatty acids, methane, CO_2, etc. Fatty acids are partially absorbed from the large intestine and, thus, enter intermediary metabolism.

Monosaccharide and disaccharide alcohols are absorbed and metabolized differently. The monosaccharide alcohols (e.g., sorbitol, xylitol, mannitol, erythritol) are directly absorbed from the small intestine, whereas the disaccharide alcohols must be cleaved first by intestinal carbohydrases into their constituents (i.e., monosaccharides and hexitols/polyols). For this reason, the disaccharide alcohols isomalt, lactitol, and maltitol are also called second-generation sugar alcohols. The rate of cleavage of individual disaccharide alcohols by the digestive enzyme systems of the human mucosa varies [17].

On the other hand, incomplete absorption is the reason for gastrointestinal side effects (e.g., gas, flatulence, diarrhea) that may occur if polyols are ingested in single doses of more than 20 g. The tolerance of polyols depends on the form of administration, i.e., whether the sugar alcohols are consumed as aqueous solutions in form of beverages or as components of solid foods such as chocolates [18], confectionery, or hard boiled candies. In solid foods, daily doses up to 60 g are well tolerated. The tolerance of polyols is not related to body weight: If equal amounts are consumed by children and adults, children do not exhibit greater sensitivity [19]. According to the Scientific Committee for Food of the EC [20], "a laxative effect can be observed when polyols are ingested in larger amounts. An intake of 20 g/person/day probably causes no undesirable laxative problems."

Especially in sweetmeats, confectionery, chocolates, and dietary products for diabetics, sugar alcohols are food ingredients and are not considered additives in the usual sense. In these products they can account for up to 5 % of the daily food intake. Consequently, the determination of an acceptable daily intake (ADI) value in the conventional sense by using the safety factor 100 after the determination of the "no-effect level", is not practicable.

Therefore, after reviewing long-term toxicological, carcinogenic, teratogenic, and reproductive studies, the Joint FAO/WHO Expert Committee on Food Additives (JECFA) stipulated a "not-specified" ADI for the sugar alcohols xylitol, sorbitol, mannitol, isomalt, lactitol, maltitol, and for hydrogenated glucose-containing syrups [21]. The specifications and methods of analy-

sis adopted by the WHO/FAO are presented in [22] on the following pages: pp. 111 – 113 for isomalt, pp. 117 – 121 for lactitol, pp. 125 – 128 for maltitol, pp. 129 – 130 for maltitol syrups, pp. 134 – 139 for mannitol, pp. 241 – 244 for sorbitol, pp. 245 – 248 for sorbitol syrups, and pp. 267 – 271 for xylitol.

For the EC the specifications are summarized in a draft of a directive from June 1994 (EC 5757/III/93 EN-Rev.). The classification of the individual polyols varies from country to country according to their different food laws. They are sometimes regarded as food additives (e.g., within the EC [23]) and other times as food or generally recognized as safe (GRAS) substances, for example in the United States [24].

General Properties. Some of the sensory, chemical and physical properties of polyols are very different from those of sugar.

The *relative sweetness* varies from 1 for xylitol to ca. 0.4 for lactitol and mannitol. The relative sweetness of polyols is more concentration dependent than in the case of other carbohydrate sweeteners. The taste profile "sweet" is for some polyols very different from that of sugar, mainly because the enthalpy of solution of polyols differs from that of sugar (Table 2) [25–28]. Polyols with a higher positive enthalpy of solution cause a sensation of cooling (refreshment) in the oral cavity.

The *melting* (*sintering*) *points* of polyols range from 165 °C for mannitol to 92 °C for xylitol and sorbitol. When compressed to tablets, polyols with a lower melting point give products with a smooth surface. They can be ground to microcrystalline grains only with difficulty (required d' ca. 60 µm, n ca. 2; where d' is the diameter of the crystals and n a parameter of the Rosin – Rammler – Sperling distribution; → Sugar, Chap. 10.2.), as required, for example, in the production of chocolates and chewing gum.

Individual polyols differ greatly in their *hygroscopicity*. For instance, sorbitol (crystalline and aqueous solutions) and maltitol syrups are very hygroscopic, while mannitol and isomalt are practically nonhygroscopic with water activities (a_w values, → Foods, 2. Food Technology, Chap. 1.4.) lower than that of sugar.

The *solubility of polyols in water* as a function of temperature is shown in Figure 1. The solubility of the polyols sorbitol, xylitol, and maltitol at 20 °C is the same order of magnitude as that of sucrose. However, the solubility of isomalt and mannitol is significantly lower. For all polyols, the temperature dependence of solubility is greater than that of sucrose. The most important physical and chemical parameters, such as density, boiling-point elevation, and viscosities of aqueous solutions and melts, are listed in [29].

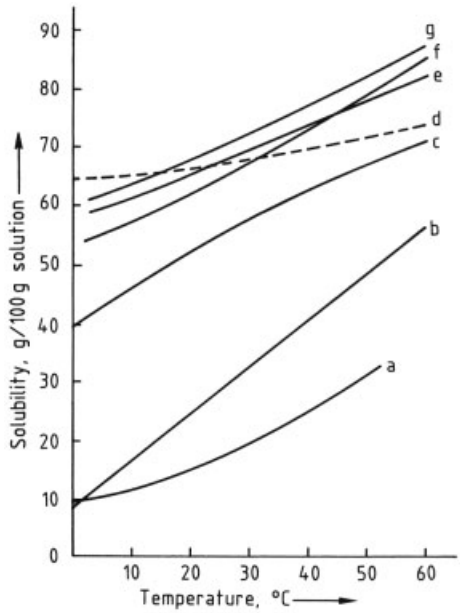

Figure 1. Solubility of polyols in water
a) Mannitol; b) Isomalt; c) Lactitol; d) Sucrose; e) Maltitol; f) Xylitol; g) Sorbitol

Chemically, sugar alcohols differ considerably from carbohydrate sweeteners because they lack a carbonyl group. Thus, Maillard reactions and the Strecker degradation are inhibited (little browning or flavor formation). In addition, mono- and disaccharide alcohols are much more resistant to enzyme systems, acids, alkalis, and the action of heat than sugar or glucose syrups.

Because of their noncariogenic properties and their reduced calorific value compared to carbohydrate sweeteners, sugar alcohols are suitable for the production of noncariogenic and reduced- calorie (light) products and products with a defined nutritional purpose (e.g., diabetics).

Table 2. Enthalpy of solution of saccharides and sugar alcohols

Compound	Enthalpy		
	kJ/mol	kJ/kg	Reference
Sucrose	6.21	18.2	[24]
Maltitol	6.7	23.0	[27]
Isomalt (Palatinit)	14.6	39.4	
Lactitol monohydrate	19.2	53.2	[26]
Sorbitol	20.2	111.0	[25]
Isomaltulose	21.66	60.2	
Mannitol	22.0	120.9	[25]
Xylitol	23.27	153.1	[24]

The sensory and food manufacturing properties of individual sugar alcohols vary substantially. Thus, proper selection of raw materials or the combination of several sugar alcohols enables sugar and glucose syrup to be replaced as bulking agents and sweeteners in all food categories in such a manner that the resulting products are acceptable to consumers.

2. Xylitol

Xylitol [*87-99-0*], $C_5H_{12}O_5$, M_r 152.15, is a naturally occurring, five-carbon sugar alcohol (pentitol) with a sweetness similar to sucrose. Xylitol is produced commercially by catalytic hydrogenation of D-xylose. It is used as a sugar substitute mainly in noncariogenic confectionery and special foods for diabetics, and as a glucose substitute in solutions for parenteral nutrition. The general characteristics, metabolic properties, and applications of xylitol have been reviewed in [30–36].

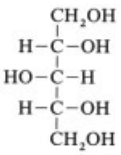

Xylitol is found in small amounts in yeast, lichen, and fungi, as well as in a variety of fruits and vegetables [35], [37], [38]. In the human body, xylitol is formed as a normal intermediate of glucose metabolism in amounts of about 5–15 g/d [39].

2.1. Physical and Chemical Properties

Xylitol forms orthorhombic crystals, *mp* 93–94.5 °C, when crystallized from ethanol or methanol [40]. A metastable, monoclinic form, *mp* 61–61.5 °C, has also been described [41]. Commercial, food-grade xylitol has a *mp* of 92–96 °C [42], [43].

Xylitol is very *soluble* in water (122.0 g per 100 g H_2O at 4 °C, 168.8 g per 100 g at 20 °C, 291.3 g per 100 g at 40 °C, and 571.1 g per 100 g at 60 °C) and sparingly soluble in absolute ethanol (1.2 g per 100 g solution at 25 °C) and absolute methanol (6 g per 100 g solution).

Xylitol is optically inactive. The refractive index n_D^{25} of a 10 % aqueous solution is 1.3471 (50 % solution: 1.4132).

Crystalline xylitol has a relatively low *hygroscopicity* (moisture absorption at 60 % R.H. after 4 d at room temperature: 0.051 %). It is stable at temperatures up to its *bp* (216 °C) and does not caramelize.

Aqueous solutions of xylitol have a low viscosity (10 %: 1.23; 20 %: 1.67; 40 %: 4.18; 50 %: 8.04; 60 %: 20.63 mPa · s). The density of a 10 % solution is 1.03 g/mL (20 %: 1.07; 40 %: 1.15; 60 %: 1.23 g/mL).

Xylitol is the sweetest of all polyols (about as sweet as sucrose; see Table 2) and has the highest positive enthalpy of solution (+ 153.1 kJ/kg).

2.2. Production

Xylitol was first obtained by E. Fischer in 1891 as a syrup by reduction of D-xylose with sodium amalgamate. Since the early 1970s, xylitol has

been produced on a commercial scale by *catalytic reduction* of D-xylose with hydrogen in the presence of Raney nickel. D-Xylose is obtained by acid- catalyzed hydrolysis of xylan-containing plant materials such as birch wood, corn cobs, and straw [44]. Since these natural raw materials also contain other carbohydrates, xylitol must be separated from other polyols by chromatographic techniques.

The synthesis of xylitol by *fermentative or enzymatic processes* is also possible in principle. However, such procedures have not been used thus far on a commercial scale. Other routes of synthesis of xylitol are merely of scientific interest [45][46][47].

2.3. Specifications, Analysis, and Legal Aspects

Specifications. Specifications for identity and purity have been laid down by the JECFA [42]. Monographs on xylitol are also included in the Food Chemicals Codex [43] and the U.S. Pharmacopoeia/National Formulary [48].

Analysis. Different methods have been used for the analysis of xylitol in food and biological samples. In most cases, high-pressure liquid chromatography yields the most satisfactory results [49][50][51].

Legal Aspects. Xylitol is approved for use in foods, cosmetics, and pharmaceuticals in more than 40 countries. In the EC, xylitol is included as E 967 in the relevant food additive directives [52]. In the United States, xylitol may be used in foods for special dietary use [53].

In the EC, the calorific value of xylitol is set at 10 kJ/g for purposes of food labeling [54]. This value takes into account the fact that xylitol, like other polyols, is incompletely absorbed [55] (see Chap. 1).

2.4. Uses

Xylitol is used as a bulk sweetener, mainly in noncariogenic confectionery and oral hygiene preparations (toothpaste and mouthwash), and less frequently in dietetic foods (food products for diabetics), pharmaceutical preparations (tablets, throat lozenges, multivitamin tablets, cough syrup), and solutions for parenteral nutrition [36], [56–59]. In noncariogenic confectionery and oral hygiene preparations, xylitol is used mainly because of:

1) Benefits in terms of dental health (xylitol is noncariogenic, inhibits certain caries-causing plaque bacteria (*Streptococcus mutans*), reduces the growth and adhesion of dental plaque, and may favorably affect the balance between de- and remineralization of incipient caries lesions). Regular consumption of xylitol-sweetened confectionery has resulted in significantly reduced increases in caries in several long-term studies of human volunteers [31], [34], [60–63].
2) Organoleptic properties (xylitol is the sweetest polyol and produces a sensation of "freshness" because of its high positive enthalpy of solution [36]).
3) Food-technology properties (relatively low hygroscopicity; good direct compressibility of certain xylitol grades; low melting point, which allows the sinter- coating of pharmaceutical tablets; and relatively high osmotic pressure, which ensures the microbiological stability of products with an aqueous phase) [36], [56], [64].

In solutions for parenteral nutrition, xylitol is used as an alternative energy source (glucose substitute) because of its largely insulin-independent metabolism. In order not to exceed the recommended infusion rate ($\leq 250\,\mathrm{mg\,kg^{-1}\,h^{-1}}$ and $\leq 3\,\mathrm{g\,kg^{-1}\,d^{-1}}$), xylitol is usually administered in combination with glucose, fructose, sorbitol, or lipids [33], [65], [66]. In amino acid solutions for parenteral administration, xylitol is used to adjust the osmotic pressure without the risk of browning (Maillard reactions) during heat sterilization and storage. In contrast to fructose and sorbitol, metabolic intolerance to xylitol is unknown.

2.5. Metabolism, Tolerance, and Safety

Metabolism. Xylitol is absorbed from the digestive tract by passive diffusion. Because of the slow rate of this process, absorption is incomplete (in rats, ca. 20 % [67], see Chap. 1).

The absorbed fraction of xylitol is metabolized mainly in the liver via the pentose phosphate shunt to intermediates of the glycolytic pathway [39]. The nonabsorbed fraction reaches the microbially colonized large intestine, where xylitol is degraded by fermentative processes to absorbable, short-chain fatty acids (mainly acetate, propionate, butyrate) and minor amounts of gas (H_2, CH_4, CO_2) [30], [55].

Tolerance and Safety (see Chap. 1). The safety of xylitol has been assessed by expert national and international groups. On an international level, the JECFA assigned a "not-specified" ADI to xylitol in 1983 [68]. The Scientific Committee for Food (SCF) of the EC arrived at a corresponding favorable decision in 1984 [69].

2.6. Producers

Xylitol is produced mainly in Finland by Xyrofin Oy and in the United States by American Xyrofin Inc.

3. Sorbitol

Sorbitol belongs to the group of naturally occurring hexitols. It was first discovered in 1868 [70] in the berries of the mountain ash in concentrations between 5 and 12 %, and is widely distributed in the flora [71]. The name sorbitol is derived from the scientific name of the mountain ash, *Sorbus aucuparia* L. The fruits of certain rosaceae are especially rich in sorbitol: plums 1.7 – 4.5 wt %, pears 1.2 – 2.8 wt %, peaches 0.5 –1.3 wt %, and apples 0.2 – 1 wt %. In fruit and leaves, sorbitol is formed as a biochemical intermediate in the synthesis of starch, cellulose, sorbose, or vitamin C [72]. In animals, sorbitol can be detected as an intermediate in the absorption of glucose [73] or the formation of fructose via glucose [74]. In 1890, E. Fischer carried out the first chemical synthesis of sorbitol by reducing glucose with sodium amalgamate [75]. The first catalytic hydrogenation was reported by V. Ipatieff in 1912 [76].

Since the 1950s, sorbitol has gained tremendous economic importance throughout the world as a food, sweetening agent, as a stabilizer of humidity, and raw material for other products [77].

3.1. Physical and Chemical Properties

Sorbitol [*50-70-4*], D-glucitol, D-sorbitol, D-glucohexane-1,2,3,4,5,6-hexaol, $C_6H_{14}O_6$, M_r 182.2, has the following structure:

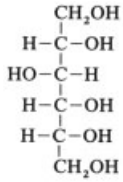

It forms γ-polymorphous, finely crystalline white crystals, that are odorless, freely flowable, and slightly hygroscopic. Sorbitol is also available in the form of water-clear syrup solutions. Sorbitol has a sweet, cooling taste; it has approximately half the sweetness of sucrose. Its physiological calorific value corresponds to that of other sugar alcohols (see Chap. 1).

Solubility. Sorbitol is readily soluble in water (256 g/100 g H_2O at 25 °C); 70 % solutions of sorbitol in water are available commercially but tend to crystallize when cooled to < 10 °C for long periods. Sorbitol is also soluble in dilute acetic acid, methanol, and warm ethanol; it is practically insoluble in organic solvents. The *melting range* of sorbitol is 92 – 96 °C; its *specific rotation* $[\alpha]_D^{25}$ is $-1.9°$ (10 g/100 mL H_2O). The optical rotation of aqueous solutions is increased greatly by the addition of complexing salts (borax, ammonium molybdate): $[\alpha]_D^{20} = +4.0$ to $+7.0°$ (10 g in 100 mL of water, borato complex). *The enthalpy of solution* is +111 kJ/kg (see Table 2). The *bulk density* of instant grades is 40 – 50 g/100 mL and that of crystallized grades 60 – 70 g/100 mL. The equilibrium moisture content is given in Table 3; the density and viscosity of aqueous solutions are listed in Table 4.

Depending on conditions (crystallization from solvents or from the melt), sorbitol crystallizes in several modifications [77], the most stable being the γ-form (*mp* 101 °C).

In neutral media, sorbitol is temperature resistant up to 150 °C and, consequently, bake and boil proof. As a sugar alcohol, sorbitol does not undergo Maillard reactions with amino acids or

Table 3. Equilibrium moisture content of sorbitol

Relative air humidity (26.8 °C), %	30	40	50	60	70	80	90
Water absorption, %	4.0	8.0	12.5	18.0	25.0	33.0	46.0

Table 4. Relative density and viscosity of aqueous solutions of sorbitol

Sorbitol content, %	5	10	25	50	60	70
Relative density	1.014	1.038	1.099	1.198	1.249	1.299
Viscosity, mPa · s	1.230	1.429	1.689	11.09	35.73	185.00

proteins at high temperature (see Chap. 1). Like other hexitols, sorbitol forms chelate complexes with some metal ions (e.g., Fe, Cu, Co, and Ni) in aqueous solution that are stable in the alkaline range [78], [79]. Thus, sorbitol inhibits the prooxidative effect exerted by trace metals in autoxidation (e.g., of fats).

Sorbitol cannot be fermented by yeast. In highly concentrated aqueous solutions (sorbitol content of ca. 50 %), sorbitol prevents the growth of bacteria by osmosis. At low concentration, sorbitol solutions serve as a nutrient medium for bacteria [80]. *Acetobacter xylinium* oxidizes sorbitol enzymatically to give L-sorbose (an intermediate in the synthesis of ascorbic acid). Under dehydrating conditions (action of acid at high temperature), sorbitol loses one molecule of water to form either 1,4-sorbitan (85 %) or 3,6-sorbitan (15 %).

On heating with concentrated sulfuric or hydrochloric acid, a second molecule of water is eliminated with the formation of 1,4: 3,6-dianhydrosorbitol = isosorbide (see Fig. 2).

If sorbitol is kept in aqueous solution for a long period at 170 °C under hydrogenating conditions in the presence of a catalyst in a weakly alkaline reaction, it isomerizes partially to form the isomeric polyols D-mannitol and L-iditol [81]. The hydroxyl groups of sorbitol, sorbitans, and sorbids can be esterified and etherifies with fatty acids; reactions of primary hydroxyl groups predominate [82], [83]. Sorbitol, sorbitan, and partially esterified products react with alkylene oxides (e.g., ethylene oxide) to form polyoxyalkylene derivatives [84].

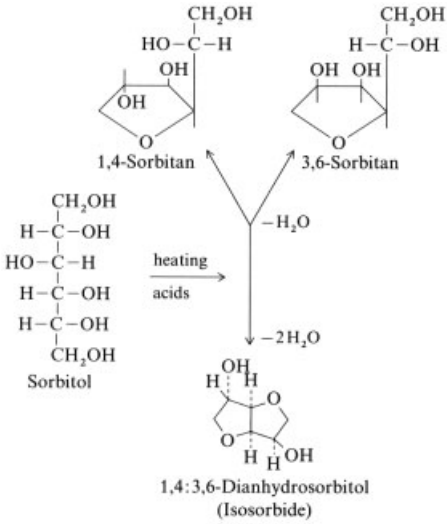

Figure 2. Anhydride formation of sorbitol

3.2. Raw Materials and Production

Raw Materials. The raw materials of choice for the production of sorbitol are glucose from starch (starting material corn, etc.) [85] and sucrose. The use of sucrose, which must first be subjected to acid hydrolysis to give invert sugar, results in the simultaneous production of mannitol. As renewable resources, both starting materials glucose and sucrose are practically inexhaustible. To produce pure sorbitol grades [86], [87], such as sorbitol powder and sorbitol instant, a highly pure glucose solution is used, which also gives the crystallizing sorbitol solutions [sorbitol liquid, solids content 69 %(min.); sorbitol 91 % (min.) of dry weight].

Noncrystallizing sorbitol solutions [sorbitol syrup, solids content 69 % (min.); sorbitol 50 % (min.) of dry weight]are made from partially hydrolyzed starch syrup [glucose content 50 % (min.)].

Production. The industrial synthesis of sorbitol is based on a process developed in 1925 that involves the catalytic hydrogenation of sugars [88]. Appropriate methods are described in the literature [89], [90]. The microbiological enzymatic production methods mentioned in the research literature use *Zymomonas mobilis* [91–94] and other microorganisms [95], [96], but are not yet economically important.

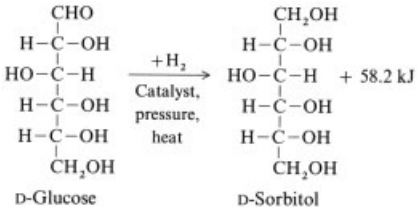

In catalytic hydrogenation, a ca. 50 % glucose solution, which is made from crystallized glucose, a starch hydrolysate (usually corn starch), or an invert sugar solution, is employed as shown in Figure 3.

Hydrogenation is performed in the presence of supported nickel catalysts (kieselguhr support) or Raney nickel. Two- and more-component catalysts have also been proposed [97], [98]. Industrially, the following processes are employed:

Batch Process. In the batch process [89], 3 – 6 % of catalyst (based on glucose) is suspended in a ca. 50 % glucose solution whose pH has been adjusted to 6. The solution is then hydrogenated with hydrogen in an agitated autoclave at 120 – 150 °C and ca. 70 bar. The time of hydrogenation is 2 – 4 h, depending on the type and amount of catalyst.

Continuous Process. In the continuous process [99–101], a glucose solution, adjusted to pH 6 and containing the catalyst, is passed continuously under hydrogen pressure through reaction tubes connected in series. A constant pressure of 180 bar and a temperature of 150 °C are maintained for a reaction time of ca. 1 h. The continuous process is more economical, but requires strictly controlled hydrogenation conditions. Hydrogenation can also be carried out on a fixed-bed catalyst based on copper oxide –nickel oxide [102] or nickel – titanium [103]. The disadvantage is the relatively short duration of reactivity of the catalyst, but the advantage is a higher-purity sorbitol. With fixed-bed catalysts, hydrolysates of potato starch can also be hydrogenated. After the reaction, the raw sorbitol solution is cooled and freed from catalyst by precipitation and filtration. It is then decolorized with filter acids (e.g., activated charcoal) and demineralized by means of mixed-bed ion exchangers to remove gluconate and nickel ions. The commercial water- clear 70 % *sorbitol solution* is obtained by evaporation of water in vacuum. *Solid sorbitol* is prepared by dehydration un-til a water-free melt is obtained. This melt is cooled and, at the same time, kneaded intensively [101]. The melt is subsequently seeded, and the crystals are removed continuously from the surface (melt crystallization) [104], [105]. Grades of powder with the desired particle size are obtained by grinding.

Instant sorbitol is made by spray drying concentrated aqueous sorbitol solution [106], [107]. Instant sorbitol is readily soluble in water, has a bulk volume that is greater by one-third than powdered sorbitol, and adsorbs finely powdered substances on its microfilament crystalline surface with cross-links in all three spatial directions [108].

In all processes, crystalline sorbitol having the stable γ-form is obtained predominantly [109].

Purity Requirements. The purity of sorbitol grades used for food must meet the requirements of the JECFA [110], the Food Chemical Codex (FCC), and EC guidelines, which largely agree. The methods stipulated here are valid for the determination of purity, and the specifications have also been adopted by pharmacopoeias in most countries.

The following requirements have been stipulated for sorbitol powder: sorbitol content 91 % (min.; based on dry weight), 98 % (min.; dry weight) of sugar alcohols of the general formula $CH_2OH(CHOH)_nCH_2OH$, where n is an integer. The nonsorbitol portion is composed mainly of mannitol [ca. 2 % (max.)], small amounts of other sugar alcohols of the general formula $CH_2OH(CHOH)_nCH_2OH$, where $n > 4$,

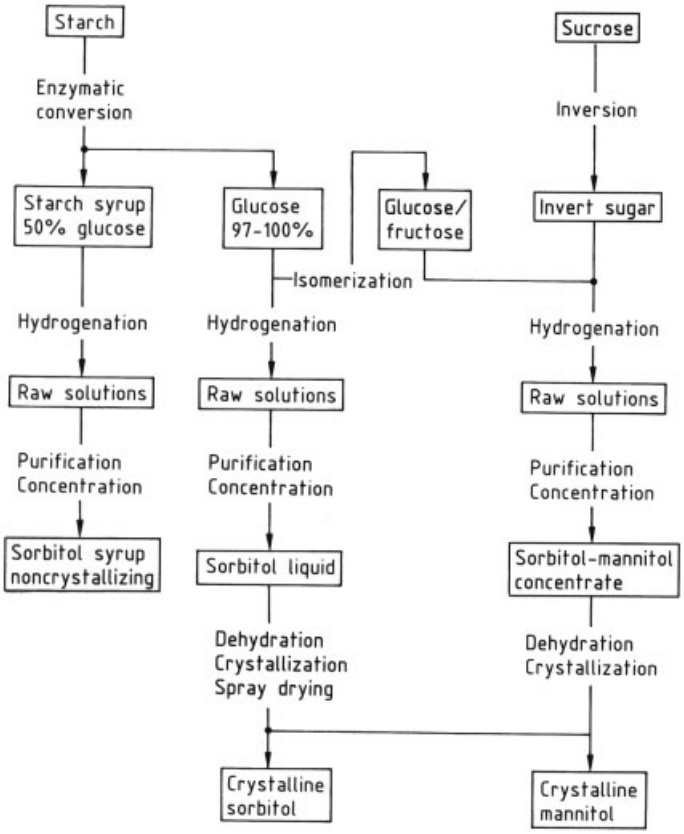

Figure 3. Production of sorbitol and mannitol

and minor amounts of hydrogenated oligosaccharides. Maximum concentrations of other constituents are given below:

Water (according to K. Fischer)	1 % (max.)
Reducing sugars (as glucose)	0.3 % of dry weight
Total sugar after hydrolysis (as glucose)	1 % of dry weight
Sulfate ash	0.1 % of dry weight
Sulfate	0.01 % of dry weight
Chloride	0.005 % of dry weight
Nickel	2 mg/kg
Arsenic	3 mg/kg
Lead	10 mg/kg
Zinc	25 mg/kg
Copper and zinc together	50 mg/kg
pH (10 % solution in water)	5 – 7

The same conditions apply to sorbitol liquid and sorbitol syrup.

3.3. Environmental Protection

Sorbitol is a globally approved food additive that is also used like sugars for the production of sweetmeats and diabetic sugar. Consequently, special environmental measures are not required for handling sorbitol. Since sorbitol is also a nutrient for microorganisms, it is microbiologically degradable and does not form hazardous intermediates or final products. In its chemical synthesis, relevant safety regulations must be observed. The collected catalysts are separately regenerated and returned for reprocessing. Unused hydrogen, which is not chemically bound, is recycled. The waste quota produced is extremely small and, therefore, can be neglected.

3.4. Analysis

Several methods are available for the qualitative and quantitative determination of sorbitol [111–115]. The proper isolation of sorbitol from samples, with subsequent purification and enrichment, is important for exact identification. An easy oxidimetric determination is based on the cleavage of C – C bonds with periodate, which is reduced to iodate in the process. In periodate oxidation, 4 mol formic acid and 2 mol formaldehyde are produced [116–118] from 1 mol of sorbitol. This method is not specific and is applicable only when no periodate- consuming accompanying substances are present. TLC [119] permits the separation of other sugars and polyols with lead tetraacetate as the developing agent. The use of GC allows substantially more accurate isolation and quantitative detection [120–123].

A very specific and exact separation can be obtained by means of HPLC [124–128], which is generally the method of choice for quick and safe product control. The enzymatic determination of sorbitol used for clinical analysis has gained acceptance for food analysis although its specificity is doubtful [115], [129–131]. In this reaction, sorbitol is oxidized to fructose by sorbitol dehydrogenase in the presence of NAD. The amount of NADH, which is produced during the course of the reaction is equivalent to the amount of sorbitol present. The NADH absorption is measured at 334, 336 or 340 nm.

3.5. Storage and Transportation

As a result of its hygroscopicity, sorbitol must be packed in tightly closed containers and protected against moisture. Even slight absorption of water causes lumping of the powdered product. In its unopened original packing, solid or liquid sorbitol can be stored almost indefinitely. The usual commercial packs contain a maximum of 50 kg of sorbitol powder or up to 270 kg of sorbitol liquid (barrels). For bulk purchasers, sorbitol can be delivered in big bags, containers, or tank trucks. Sorbitol liquid should not be stored at less than 10 °C for long periods because the resulting crystallization hinders its removal from the container. The maintenance of hygiene regulations associated with food is essential.

3.6. Legal Aspects (see Chap. 1)

In accordance with EC guidelines, sorbitol is a food additive. Consequently, its use in food must be legally controlled by member countries. In view of the harmonization of the Common Market, these regulations should be standardized. According to the German food law, food that contains $> 10\,\%$ sorbitol must carry the label "excessive consumption can produce a laxative effect." In addition, the concentration must be labeled (g/kg or wt %). Further details can be found in relevant national laws.

The use of sorbitol in foods, cosmetics, and pharmaceuticals is approved by law in more than 40 countries. In the United States, the Code of Federal Regulations (21 CFR 184.1835) acknowledges sorbitol as a GRAS substance (generally recognized as safe) and provides guidelines for its use in foods [132], [133].

3.7. Uses

The main consumer of solid and liquid grades of sorbitol is the food industry, followed by the cosmetic and pharmaceutical industries. Sorbitol serves as a moisture stabilizer and softener, sugar substitute, carrier, and general conditioning agent.

Foods. The use of sorbitol in the food industry has many advantages [134–136]. The powdered grades are employed as a sugar substitute for diabetic foods; for sugar-free items such as chocolates, dragées, tablets, chewing gum, and filled hard candy; for reduced- calorie sweetening agents whose sweetness is adjusted with intense sweeteners; and for infusion solutions used in parenteral nutrition. Sorbitol liquid and sorbitol syrup serve as softeners and moisture stabilizers for confectionery of all types [137–140]; marzipan; cakes and pastries; aroma substances; delicacies; mayonnaise and dressings; products made of fish, shrimp, and roe; baking emulsifiers; candied fruits; sauerkraut; natural and artificial skins; ice cream; enzyme preparations; tonics; and tobacco products. The concentration of sorbitol powder used is 10 – 100 %, depending on the food. Sorbitol liquid and sorbitol syrup are

employed in concentrations of 5 – 10 %. For details about special processing technologies, see [135], [136], [141].

Cosmetics. In the production of water-containing cosmetics such as creams, ointments, emulsions, lotions, gels [142–145], and especially toothpastes [146–148], sorbitol syrup is preferred to glycerol. Depending on the desired effect, the concentration required is 3 – 5 % for water stabilization and 10 – 20 % for softening effects.

Pharmaceuticals. Similar to its uses in cosmetic formulations, sorbitol can also be used in pharmaceutical preparations [149–153]. Sorbitol powder has gained acceptance as a tablet excipient [154]. As a result of its surface structure, spray-dried sorbitol is especially suitable for the uniform binding of solid agents. It can be molded directly into tablets without granulation [109], [154–156]. This type of sorbitol is also suited to the production of stable dry suspensions (e.g., with antibiotics [157]) of vitamin tablets [158] or antacid granules [159].

Medical Uses. In medicine, pyrogen-free sorbitol is used in 10 – 20 % solutions, with or without amino acids [160], as infusion solutions for parenteral nutrition [161], especially for diabetic patients [162–166]. It accelerates diuresis in osmotherapy, stimulates cholokinesis in gallbladder and liver diseases [167], and serves as a laxative in enemas administered for constipation.

Technical Uses. In technical applications sorbitol syrup is important primarily as a moisture stabilizer and softener. It prevents the embrittlement of all types of paper [168]; hydrous adhesives and (gelatin-based) glues remain flexible. In the finishing of textiles, leather, and fur, sorbitol serves as a conditioning agent. It also stabilizes the moisture in polyurethane foams [169]. About 25 % of the worldwide production of sorbitol is used in the synthesis of ascorbic acid [170].

Derivatives. Esters of sorbitan [171] with fatty acids (e.g., sorbitan monopalmitate, stearate, laurate, oleate [172], [173], and trioleate) as well as polyoxyethylene derivatives [174] of the general formula:

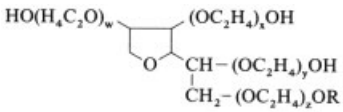

where R is a fatty acid residue (e.g., palmitoyl, stearyl), are formed from sorbitan by etherification with ethylene oxide, followed by acylation with fatty acid. These substances (e.g., Span, Arlacel, Crill, Tween, and Polysorbate) are important emulsifiers and solubilizers in cosmetic and pharmaceutical products and in the food industry [174], [175]. In sorbitan esters, the lipophilic character predominates, and in polyoxyethylene derivatives, the hydrophilic character [171].

3.8. Physiology, Tolerance, Toxicology

For a general discussion of physiological effects of polyols, see Chapter 1 and [176–183].

The absorption of sorbitol results independent of insulin; an increase in blood sugar level is not detectable in either healthy or diabetic persons. For this reason, sorbitol is an ideal sugar substitute for diabetics [184], [185]. The absorption capacity – orally administered – is 10 – 20 g/h. Metabolism occurs in the liver with the help of sorbitol dehydrogenase via fructose and subsequent glycolysis [176]. No side effects have been observed.

The acute toxicity (LD_{50}, oral) for rats is 17.5 g/kg body weight and for mice 25.7 g/kg. In intravenous administration to rats, the LD_{50} was 7.3 g per kilogram of body weight. Even under extreme conditions, no substance-specific toxic or teratogenic effects could be determined [186], [187]. Earlier suggestion that the oral intake of sorbitol causes cataract formation in the lens of the eye could not be confirmed [188].

3.9. Producers

The most important producers, with commercial products in parentheses, are as follows:

Europe: Roquette, France (Neosorb); Cerestar and Hydrierwerke Rodleben, Federal Republic of Germany and Varallo, Italy (all produce Sorbidex); E. Merck, Federal Republic

of Germany (Karion); Hefti, Switzerland (Sorbex); Cultor-Kotka, Finland; Roche-Dairy, Federal Republic of Germany and United Kingdom; Pet Nitrogen, Hungary; Pliva, Croatia; Vitamine Comp., Commonwealth of Nations.

America. Lonza, ICI, Pfizer, Etichem, and Roche, United States; Atkemix, Canada; Polybasicos, Mexico; Sad Quimica Rhodia, Argentina; Getec, Brazil; Ecuador Polychemics, Ecuador.

Asia. Towa Kasei Kogyo, Nikken, Aito, Takeda, Ajinomoto and Sanei Toka, Japan; Lucky, Korea; San Fu Chemical, Taiwan; Pure Chemicals, Thailand; Sorini, Indonesia.

Africa. Klipfontein Pure Organics, South Africa.

3.10. Economic Aspects

Worldwide production (1992) of sorbitol was estimated at about 650 000 t (calculated as 70 % solution). Of this amount, about 357 000 t (55 %) is accounted for by the sale of 70 % sorbitol solution, including sorbitol syrup, and ca. 20 % is allotted to solid grades of sorbitol (91 000 t). About 25 % of the sorbitol solutions (162 000 t) are used for the synthesis of ascorbic acid. In past years, the increase in demand was, on average, about 3 % per year. Depending on the amounts delivered, the world market price of sorbitol liquid and sorbitol syrup (70 %) is ca. \$ 0.95 –1.40/kg, and the price of solid grades is in the range \$ 1.90 – 2.50/kg. The market value of the sorbitol produced worldwide was at least \$ 900 million in 1993.

4. Mannitol

D-Mannitol is a hexitol derived from D-mannose. D-Mannitol is widely distributed in nature, occurring in olive trees, plane trees, and fruit and vegetables (e.g., pumpkins, strawberries, celery, and onions). It is the main constituent of the ash (*Fraxinus ornus L.*). The air-dried sap of this tree, manna, obtained by heating the bark, contains 30 – 50 % mannitol. PROUST isolated mannitol in the crystalline state from manna [189]. Marine algae, especially brown algae, are also rich in mannitol. They can contain 10 – 20 % mannitol, depending on the time of harvest. Furthermore, a large number of fungi and bacteria can produce this polyol from glucose, fructose, or sucrose [190].

4.1. Physical and Chemical Properties

D-Mannitol [*69-65-8*], mannite, D-mannohexane-1,2,3,4,5,6-hexaol, mannitolum, mannitolo, and manna sugar, $C_6H_{14}O_6$, M_r 182.2 , has the following structure:

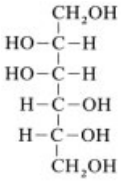

Mannitol is a white crystalline powder that is about half as sweet as sucrose. Of the commercially used sugar alcohols, mannitol is the least soluble in water (see Chap. 1, Fig. 1). Because of its positive heat of solution in water (see Chap. 1, Table 2), a cooling sensation occurs when mannitol dissolves in the mouth. This effect is, however, less pronounced than that observed with sorbitol or xylitol.

Mannitol is sparingly *soluble* in many organic solvents (1.2 g/100 mL of ethanol; 5.5 g/100 mL of glycerol), and practically insoluble in ether, ketones, and hydrocarbons. Mannitol has a *mp* of 165 – 168 °C; optical rotation $[\alpha]_D^{25} = -0.48°$ ($c = 7$ g/100 mL). The optical rotation is increased greatly by the addition of borax or ammonium molybdate, and the values become positive (see Table 5). Mannitol has a relative density $d^{20} = 1.52$; the particle-size distribution of crystalline grades is 20 – 250 µm and of granulated grades 100 – 850 µm; its bulk density is 550 – 650 g/L. The viscosities of mannitol and sorbitol solutions are comparable.

Crystalline mannitol is only slightly hygroscopic; appreciable water absorption occurs at > 90 % R.H. Mannitol, like other hexitols, forms chelate complexes with metal ions (e.g., iron, copper, and nickel) in aqueous solution [191]. The iron complexes are especially stable [192]. Even at 150 – 170 °C, mannitol does not react with proteins according to the Maillard reaction. Nitration of mannitol with nitrating acid yields mannitol hexanitrate (hexanitromannitol), which is used as an igniter for detonators and electrical fuses [193]. Like other polyalcohols, mannitol releases water when heated with concentrated sulfuric or hydrochloric acid, giving rise to isomannide (1,4 : 3,6 dianhydromannitol, $C_6H_{10}O_4$). Mannitol cannot be fermented by yeast.

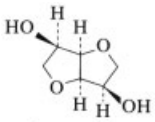

Isomannide

Table 5. Specifications and purity testing of mannitol

Specification	USP XXII	JECFA
Concentration of D-mannitol (based on the dry matter), %	96.0 – 101.5	96.0 – 102.0
Melting range, °C	165 – 169	165 – 169
Specific rotation $[\alpha]_D^{20}$	+ 137 – 145∗	+ 23 – 25∗∗
Loss on drying (105 °C, 4 h), %	≤ 0.3	≤ 0.3
pH		5 – 8
Chloride, %	≤ 0.007	≤ 0.007
Sulfate, %	≤ 0.01	≤ 0.01
Nickel, %		≤ 0.0002
Arsenic, %		≤ 0.0003
Lead, %		≤ 0.0001
Heavy metals (Pb), %		≤ 0.001
Sulfate ash, %		≤ 0.1
Reducing sugar as glucose, %		≤ 0.3
Total sugar as glucose after hydrolysis, %		≤ 1.0

∗ 1 g of mannitol in 10 mL of water after the addition of 4 % ammonium molybdate. ∗∗ 2 g of mannitol in 25 mL of water after the addition of 2.6 g of disodium tetraborate.

4.2. Raw Materials and Production

Production of mannitol by extraction of plant raw materials is no longer economically important. Only in China is mannitol still obtained by acid extraction of marine algae.

Today, mannitol is usually produced industrially (see Fig. 3) from sucrose, after inversion to glucose – fructose syrup (ratio 1 : 1), or from a glucose solution, which is also converted to high-fructose corn syrup (→ Glucose and Glucose-Containing Syrups, Chap. 6.2.) (52 % glucose; 44 % fructose, maltose, and oligosaccharides).

Theoretically, 25 % mannitol and 75 % sorbitol are formed in the neutral pH range from a 1 : 1 D-glucose – D-fructose syrup. Hydrogenation occurs under high pressure in the presence of Raney nickel catalysts. As a result of its lower solubility compared to sorbitol, mannitol is isolated from the concentrated hydrogenation solution by crystallization. The filtrate still contains considerable amounts of mannitol, which can be separated chromatographically [194].

If sucrose is used as starting material and hydrogenation is performed in alkaline solution, mannitol yields of up to 31 % can be obtained [195]. In a pure glucose solution, moreover, D-glucose can be epimerized to D-mannose before hydrogenation by the addition of an acidic solution of ammonium molybdate. Mannose is then hydrogenated directly to D-mannitol [196], [197].

Microbiological processes for the direct production of mannitol from glucose [198], [199] or fructose [200] have been described.

In 1986, the isolation of mannitol from a glucose – fructose syrup was achieved by using mannitol dehydrogenase instead of hydrogenation [201].

Mannitol syrup is not commercially available because of its tendency to crystallize.

The measures that need to be taken for environmental protection, storage, and transport are similar to those taken for sorbitol powder products (see Section 3.5).

Purity Requirements. Mannitol is listed in all the important pharmacopoeias. If used as a food additive, it must meet the requirements of the JECFA [202], the FCC, and the EC guidelines (E 421), which largely agree with each other. The methods for purity testing and the specifications are stipulated in Table 5.

4.3. Analysis

Mannitol can be detected qualitatively by thin-layer chromatography [203], [204].

Three methods are commonly used for quantitative detection of mannitol. The periodometric method [203], [204] is relatively nonspecific, HPLC analysis is very specific [205], and gas chromatography is becoming less and less important.

4.4. Uses

The pharmaceutical industry is the largest consumer of mannitol. Due to its low chemical reactivity and low hygroscopicity, it can be used in the production of tablets as a suitably inert sweet excipient for both organic and inorganic agents.

Normally, granular mannitol is manufactured by the customer; however, directly compressible grades also exist [206]. Granulation can be carried out with or without binder and, if necessary, together with active agents.

Furthermore, mannitol can be used as an osmotic diuretic [207] in the treatment of cerebral edema [208], for reducing intraocular pressure [209], and as a laxative [210].

Mannitol is suitable only for special foods because it exerts a stronger laxative effect than sorbitol, is more expensive and tends to crystallize in water- containing products.

Mannitol is used principally to powder the surface of sugar-free chewing gum; as an admixture to other polyols (sorbitol or xylitol), it is usually employed in concentrations of 5 – 10 %. Furthermore, the quality of hard caramel based on sorbitol or maltitol can be improved by the addition of mannitol. Mannitol is used for powdering and sweetening the surface of diabetic confectionery and other sweetmeats because hygroscopic polyols are unsuitable for this purpose. It is also added to instant dessert powders or polydextrose to reduce hygroscopicity and improve flowability [211], [212].

4.5. Physiology, Tolerance, Toxicology

A general discussion of the physiological properties of sugar alcohols is given in Chapter 1.

When mannitol is administered intravenously, about 80 % is excreted unchanged via the kidneys [213]. Fructose is formed as the first intermediate [214]. Absorption is insulin independent [215], so mannitol is suitable for diabetics.

As a result of slow absorption, the oral intake of single doses > 10 – 40 g can lead to diarrhea caused by osmosis. The threshold value depends on the food consumed together with mannitol [216–218].

Acute Toxicity. LD_{50} = 17.3 g/kg (oral, rat), LD_{50} = 22 g/kg (oral, mouse), and LD_{50} = 17 g/kg (intravenous, mouse). The addition of 2.5 and 5 % mannitol to the food of rats and mice, respectively, for a period of two years showed that mannitol exhibits no carcinogenic properties [219].

4.6. Economic Aspects

Producers: Roquette, France; ICI, United States; E. Merck, Federal Republic of Germany; Fin Sugar, Finland; Gretec, Brazil; Towa Kasei, Kogyo, Japan; Cerestar, Italy; Hefti, Switzerland; Test, People's Republic of China.

Approximately 20 000 t of mannitol was produced worldwide in 1992 at a price of $ 3.75 – 4.75.

5. Isomaltulose and Trehalulose, Isomalt

5.1. Isomaltulose and Trehalulose

Isomaltulose (**1**) (6-*O*-α-D-glucopyranosylfructose [*13718-94-0*]), common name palatinose, and *trehalulose* (**2**) (1-*O*-α-D-glucopyranosylfructose [*51411-23-5*]) are two of the few naturally occurring reducing disaccharide ketoses. Honey can contain both sugars but in low concentrations.

The formation of isomaltulose in the fermentation broth of *Leuconostoc mesenteroides* was first suspected in 1956 [220]. The structure was elucidated in 1957 [221]. The first industrial process for the isolation of isomaltulose was reported in 1972 [222]. The industrial preparation

of trehalulose was described for the first time in 1983 [223]. The sole raw material for the industrial preparation of both sugars is sucrose.

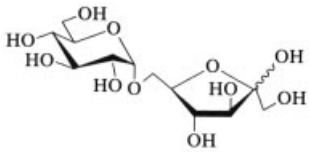

(1)

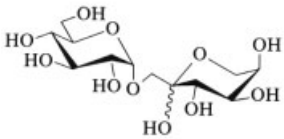

(2)

Both saccharides belong to the group of glucosylfructoses. However, unlike sucrose, they are reducing carbohydrates.

5.1.1. Physical and Chemical Properties

Isomaltulose crystallizes as the monohydrate, *mp* 123 – 124 °C. The specific rotation is $[\alpha]_D^{20} = +103°$ (calculated on a water-free basis). The tautomeric distribution (20 °C, in water) is 19.7 % of the α-furanose form and 80.3 % of the β-furanose form [224]; the solubility is 0.49 g (anhydrous isomaltulose) per g of water (20 °C).

Trehalulose. The crystallization of trehalulose has not been described until now. The specific rotation of trehalulose in water is $[\alpha]_D^{20} = +50°$ with the tautomeric forms in equilibrium (70.9 % β-pyranose form, 21.1 % β-furanose form, 5.7 % α-furanose form, and 2.3 % α-pyranose form) [224].

On hydrolysis with acids, isomaltulose and trehalulose yield 1 mol each of glucose and fructose. These compounds hydrolyze much more slowly than sucrose (Fig. 4).

On reduction of isomaltulose, the sugar substitute isomalt is obtained [222] (see Section 5.2). The reduction of trehalulose gives an isomeric mixture of 1-*O*-α-D-glucopyranosylsorbitol and 1-*O*-α-D-glucopyranosylmannitol [225].

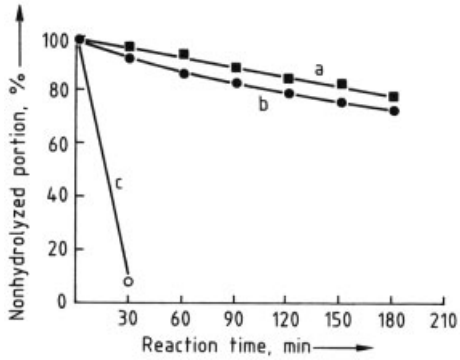

Figure 4. Hydrolysis of glucosylfructoses in acid solution (10 g/L in 0.1 mol/L HCl, 95 °C)
a) Isomaltulose; b) Trehalose; c) Sucrose

Isomaltulose and trehalulose can be converted to new intermediate products by a number of different reactions. These products are being tested intensively for use in synthesis of new consumer products [226].

5.1.2. Production

The industrial production of isomaltulose and trehalulose proceeds exclusively from sucrose by transglucosidation with the help of immobilized microorganisms that are incapable of propagation.

Isomaltulose. Although a large number of other microorganisms are also capable of forming isomaltulose, *Protaminobacter rubrum* CBS 574.77 is being used for its commercial production. The product composition (% of solids) of an industrial transglucosidation solution is shown below [227]:

Isomaltulose	79 – 85
Trehalulose	8 – 10
Sucrose	0.5 – 2
Isomaltose	1 – 1.5
Oligosaccharides, degree of polymerization ≥ 3	1 – 1.5
Fructose	2.5 – 3.5
Glucose	2.0 – 2.5

Isomaltulose can be obtained in a highly pure crystalline state by multistage evaporative – cooling crystallization.

Trehalulose. Under suitable reaction conditions, trehalulose is produced together with isomaltulose by using *P. rubrum* CBS 574.77 [223].

In addition microorganisms have been found that predominantly form trehalulose directly from sucrose [228].

5.1.3. Uses

Isomaltulose and trehalulose are only partially digested by the bacterial flora of the human mouth. Because of their pure sweet taste, they are being used increasingly, especially in Japan, to replace sucrose, glucose, and fructose in the production of sweets with low cariogenic potential.

However, in contrast to the sugar alcohols, both sugars are fully metabolized, and their use in sweets is restricted because they tend to form reversion and Maillard products. Isomaltulose is used mainly for production of the sugar substitute isomalt.

5.1.4. Economic Aspects

At present, the worldwide production of isomaltulose is ca. 25 000 t and that of trehalulose ca. 3000 t (as syrup with ca. 80 % trehalulose content of the dry substance).

Producers. Südzucker AG Mannheim/Ochsenfurt (Federal Republic of Germany); Mitsui Sugar (Japan).

5.2. Isomalt

Isomalt [*64519-82-0*] is a mixture of the diastereomers 6-*O*-α-D-glucopyranosyl-D-sorbitol (isomaltitol, 1,6-GPS) [*534-73-6*] and 1-*O*-α-D-glucopyranosyl-D-mannitol (GPM; the 1-*O*- and 6-*O*-bound mannitol derivatives are identical because of the symmetry of the mannitol part of the molecule) [*20942-99-8*]. The two isomers are formed in different ratios, depending on reaction conditions for the hydrogenation of isomaltulose [222]. Indeed, 1,6-GPS is also formed in small amounts in the catalytic hydrogenation of starch hydrolysates.

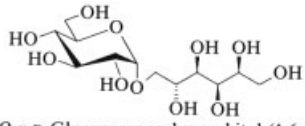

6-*O*-α-D-Glucopyranosyl-D-sorbitol (1,6-GPS)

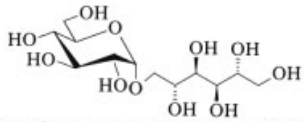

1-*O*-α-D-Glucopyranosyl-D-mannitol (1,1-GPM)

Both isomers belong to the group of disaccharide alcohols.

5.2.1. Physical and Chemical Properties

6-*O*-α-d-Glucopyranosyl-d-Sorbitol (1,6-GPS). The crystal structure of 1,6-GPS is given in [229]. The melting range is 166 – 168 °C and the specific rotation $[\alpha]_D^{20} = +92.4°$.

1-*O*-α-d-Glucopyranosyl-d-Mannitol (GPM). GPM crystallizes from an aqueous solution as the dihydrate (crystal structure see [230]). The dihydrate releases water of crystallization at 70 – 130 °C. An anhydrous modification is formed that melts at 172 °C. The specific rotation of GPM is $[\alpha]_D^{20} = +91.2°$ (calculated on a water-free basis).

Important characteristics of isomalt are as follows:

Water content	≤ 7 %
Specific rotation	+ 89.8° – + 92.2°
Sintering/melting range	145 – 150 °C
Ash content	≤ 0.05 %
Content of 1,1-GPM or 1,6-GP	each 43 – 57 %
Σ 1,1-GPM + 1,6-GPS	≥ 98 %
Σ D-Mannitol + D-Sorbitol	≤ 1.0

Its solubility in water as a function of temperature is shown in Figure 5.

With a comparable water content (e.g., ca. 2 %), isomalt, in the melt, has a lower viscosity, higher specific heat capacity, and higher boiling-point elevation than sucrose [231]. This indicates that complex formation is much lower and affects its processing technology to a certain extent.

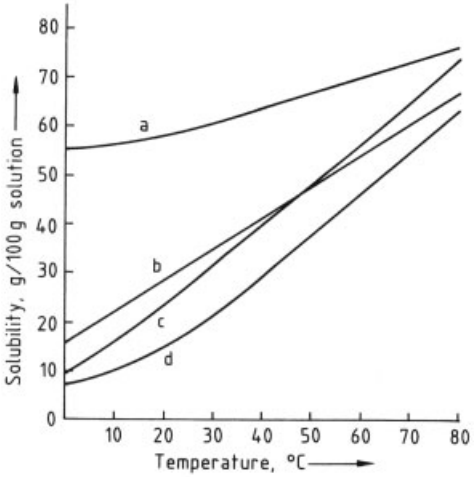

Figure 5. Solubility of GPM, GPS, isomalt, and isomaltulose in water
a) GPS; b) Isomaltulose; c) Isomalt; d) GPM

The components of isomalt are hydrolyzed by acid to give 1 mol each of glucose and sorbitol or mannitol. Compared to sucrose and similar disaccharides, the hydrolysis is clearly slower (see Fig. 6). For this reason, isomalt can be regarded as acid stable in most food applications.

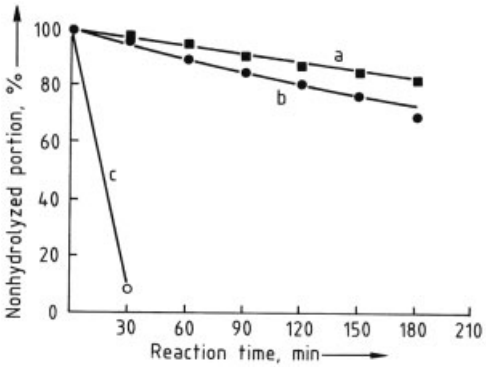

Figure 6. Hydrolysis of isomalt components and sucrose in acid solution (concentration 10 g/L in 0.1 mol/L HCl, 95 °C)
a) GPM; b) GPS; c) Sucrose

Like those of sucrose, the hydroxyl groups of the components of isomalt can be etherified and esterified [232].

5.2.2. Production

On catalytic hydrogenation, isomalt is obtained as an approximately equimolar mixture of the isomers [233]. Industrially hydrogenation is carried out as a fixed-bed process in the presence of nickel catalysts. After hydrogenation and complete deionization, isomalt is obtained as an agglomerate of the two isomalt components by using a specially developed drying process. The agglomerates consist of discrete crystallites of GPM dihydrate, anhydrous GPM, and 1,6-GPS with a primary crystal-size distribution of $< 1\,\mu m$. In processing, agglomerates ca. 0.2–2 mm in size behave like homogeneous crystallites. Since the isomalt components have a high tendency to crystallize (low solubility at low temperature, see Fig. 5), isomalt is sold exclusively as a dry product.

5.2.3. Uses

Because of its pure sweet taste (sweetening power = 0.45 relative to sucrose in ca. 10 % solution), its noncariogenicity, and its low calorific value (0.5–0.6 relative to sucrose), isomalt is almost the ideal sugar substitute for most applications in confectionery. Compared to other raw materials, isomalt is characterized by a low positive heat of solution and very low hygroscopicity (lower than sucrose). In the production of food, isomalt generally does not react with other components of the formulation. Isomalt can usually be processed in plants designed for sucrose as the raw material.

In many countries, isomalt is approved for use in food production. It is classified as a food or food additive, depending on the country. In the United States, acceptance of the application for establishing the GRAS status of isomalt has been published. Its use as a sugar substitute has been confirmed in a large number of publications from virtually all the different product lines in the confectionery industry [234].

Applications in the nonfood sector (e.g., surfactant synthesis and resin production) are also known [232].

5.2.4. Economic Aspects

Isomalt is produced by Südzucker AG Mannheim/Ochsenfurt. Current production is ca. 20 000 t/a.

6. Lactitol

Lactitol was described for the first time by SENDRENS in 1920 [235]. It was produced by KARRER and BÜCHI in 1936 by the reduction of lactose [236]. Today, lactitol is prepared commercially from pure lactose that is as protein free as possible (→ Lactose and Derivatives) by liquid-phase or fixed-bed hydrogenation with Raney nickel catalyst in a continuous or batchwise operation [237]. The conditions are 40 – 50 % aqueous neutral solution, ≥ 100 °C, and ≥ 40 bar. Then the catalyst is separated and the hydrogenated solution is purified by complete deionization with an ion exchanger (→ Ion Exchangers, Chap. 2.2.1.). After multiple-effect evaporation of the purified syrup to a solids content of ca. 75 %, lactitol is isolated by cooling crystallization. Depending on crystallization conditions, lactitol crystallizes usually as the mono- or dihydrate [238], or possibly the trihydrate [239]. Anhydrous lactitol crystals can be obtained only from ethanolic solutions [236], [240].

The specifications for commercially available lactitol monohydrate and lactitol dihydrate are given in Table 5.

6.1. Physical and Chemical Properties

Lactitol [*585-86-4*], 4-*O*-β-D-galactopyranosylsorbitol, 4-*O*-β-D-galactopyranosylglucitol, $C_{12}H_{24}O_{11}$, has a M_r 344.32, *mp* 146 °C, and optical rotation $[\alpha]_D^{25} = +14.40°$.

The density of the crystalline monohydrate is 1.44 g/cm^3; that of the crystalline dihydrate, 1.52 g/cm^3. Both hydrates have an orthorhombic crystal structure that was determined by X-ray diffraction [241].

When lactitol is heated to 170 – 240 °C, it is converted partly to anhydro derivatives (lactitan), sorbitol, and lower saccharides. The formation of lactitan is accelerated by acid (e.g., toluene-*p*-sulfonic acid) [242].

Lactitol is readily soluble in water. For the saturation curve, see Figure 1, Chapter 1; for the enthalpy of solution, see Table 1, Chapter 1. The viscosity of lactitol solutions as a function of temperature is given in [242], [243]. The viscosity of a 64 % aqueous lactitol solution at 20 °C is about 200 mPa · s. The viscosities of lactitol melts have also been determined (e.g., 2.4 Pa · s at 135 °C). For further physical parameters, see [244].

Since the molecule contains no carbonyl group, it is chemically more stable than the corresponding disaccharide (e.g., lactose). The stability of lactitol to alkali is clearly higher than that of lactose. A 10 % solution of lactitol, adjusted to pH 13 with sodium hydroxide solution, can be heated to 100 °C for 1 h without the appearance of discoloration. The stability of lactitol to acids is similar to that of lactose. On heating 10 % solutions to 100 °C for 4 h at pH 1 or 2, 5.6 % or 1.4 %, respectively, of the lactitol is hydrolyzed. No discoloration of the solutions occurs under these conditions. The nine hydroxyl groups of lactitol can be esterified with fatty acids [242]. They react with propylene oxide to form starting materials, e.g., for polyurethanes.

Enzymatic Properties. Enzymatic cleavage of lactitol by β-galactosidases (E.C. 3.2.1.23; → Enzymes) was studied by KARRER and BÜCHI [236]. They found that lactitol is cleaved only with great difficulty. In a patent specification [245], the rate of cleavage was found to be about 10 % that of lactose. More recent results have been published in [246]. The action of β-galactosidase also results in the formation of oligosaccharides (reversion products) [247].

Sensory Properties. The relative sweetness of lactitol is ca. 0.4 compared to that of a 10 % sugar solution and is strongly concentration dependent [248]. The other components of the formulation (matrix) also influence the sensation of sweetness of finished foods containing lactitol. For the enthalpy of solution and the refreshment effect, see Table 1, Chapter 1.

6.2. Physiological Properties

The general comments in Chapter 1 apply to lactitol. Review articles dealing mainly with lactitol

discuss its noncariogenicity [249]; suitability for diabetics [250]; and absorption, tolerance, and reduced physiological calorific value [251].

Lactitol can be used instead of lactulose [*4618-18-2*] (→ Lactose, Chap. 2.1.) for the treatment of hepatocerebral syndrome [252] and chronic obstipation [253].

6.3. Technological Properties

Like sucrose, lactitol is a disaccharide, and it has about the same solubility in water. For this reason, its food manufacturing properties correspond largely to those of sugar. However, the 5 – 10 % water content of lactitol represents the significant difference. It substantially influences processing parameters (e.g., in conching of chocolate) and must be taken into account when developing a formulation. For the use of lactitol in food production, see [242], [243], [254]. Lactitol is employed primarily in making confectionery, chocolates, chewing gum, sweetmeats (e.g., hard boiled candies, fruit-flavored gum, fruit jelly), sugar coatings, dragées, ice cream, and jam, and as a bulking agent for the production of tabletop sweeteners.

6.4. Economic Aspects

Producers of lacticol are: CCA Biochem B. V., The Netherlands; Cultor-Kotka, Finland.

7. Maltitol and Maltitol-Containing Syrups

7.1. Maltitol-Containing Syrups

The development of hydrogenated glucose/ corn syrups, lycasin [*9052-46-7*], goes back to the Swedish starch factory Lyckeby [255]. The composition of hydrogenated products varied according to the composition of the glucose syrup used (→ Glucose and Glucose-Containing Syrups, Chap. 6.). About 50 % of the hydrogenated products were sorbitol and maltitol, and the other 50 % were hydrogenated oligosaccharides with a DP of 3 – 20. Only after standardization of the product composition in the 1980s were these substances approved for use in accordance with food laws. All producers adhere to these specifications (Table 6).

Table 6. Specifications for commercially available lactitol mono- and dihydrate

Specification	Monohydrate	Dihydrate
Appearance	white crystalline powder	
Taste	sweet	
Odor	neutral	
Water content (Karl Fischer method), %	max. 5.5	9.5 – 10.0
Specific rotation $[\alpha]_D^{25}$	+ 14.0 – 14.6	
Melting range, °C	115 – 125	70 – 80
pH, 10 % solution	4.5 – 7.0	
Lactitol content, %	93.5 – 9.60	88 – 90
Mannitol content, %	0.3 (max.)	0.3 (max.)
Sorbitol content, %	0.7 (max.)	0.6 (max.)
Dulcitol content, %	0.2 (max.)	0.1 (max.)
Reducing sugars as glucose, %	0.04 (max.)	0.1 (max.)
Ash content, mg/kg	500 (max.)	
Heavy-metal content, mg/kg	10 (max.)	
Arsenic content, mg/kg	1 (max.)	
Nickel content, mg/kg	1 (max.)	

Table 7. Specifications of hydrogenated glucose syrups and hydrogenated high-maltose syrups

Specification	Hydrogenated glucose syrups	Hydrogenated high-maltose syrups
Trade name	Maltidex 200 Lycasin 80/55	Maltidex 100 Malbit liq.
Solids content, %	75 ± 1	≥ 74
Sorbitol content (solids based), %	6 – 8	2.5 – 3.5
Maltitol content (solids based), %	50 – 55	73 – 77
Maltotriitol content (solids based), %	≈ 10	9.5 – 13.5
Higher polyols, DP * 4 – 20 (solids based), %	≈ 30	6.5 – 14
Reducing sugar as glucose (solids based), %	0.2 (max.)	0.3 (max.)
Ash content, %	0.1 (max.)	0.1 (max.)
Nickel content, mg/kg	1 (max.)	2 (max.)
pH	6 – 7	5 – 7

* DP = Degree of polymerization.

To increase the maltitol content in syrups and, thus, adapt the physical and chemical properties

of the syrups to those of sugar solutions, high-maltose corn syrups (→ Glucose and Glucose-Containing Syrup, Chap. 7.) were hydrogenated [256], and the product composition was standardized (Table 7). Reduction and further processing of the hydrogenated solutions were carried out as described for lactitol (Chap. 6; see also [257]). For both types of syrups, no crystallization occurs even at high solids content. For this reason, they have the same food manufacturing properties in the production of "sugar-free" foods as glucose/corn syrups have in the production of normal foods. Consequently, they are primarily crystallization inhibitors, influencing the texture, moisturizing capacity, and a_w value of low-water foods [258].

The viscosity at 20 °C and a solids content of 75 % is ca. 2000 mPa · s for hydrogenated glucose/corn syrups and 1500 mPa · s for hydrogenated high-maltose syrups.

At pH 5 – 7, maltitol-containing syrups are thermostable up to 200 °C. At lower pH and > 50 °C, however, marked hydrolysis occurs. In the acid or enzymatic hydrolysis of oligosaccharide alcohols, glucose molecules are eliminated preferentially until a molecule of maltitol remains. Maltitol is more difficult to hydrolyze [259].

For these reasons, the nutrition properties of these syrups, such as noncariogenicity [260], suitability for diabetics [261], and reduced calorific value, depend greatly on the composition.

7.2. Maltitol

For the production of maltitol, high-maltose glucose corn syrups are the starting material from which pure maltose is obtained by chromatographic separation on ion-exchange resins (→ Ion Exchangers). Depending on the separation technique, maltose of varying purity (85 – 98 %) is obtained. The syrups are evaporated to a solids content of 50 % and then hydrogenated [262]. Initially, hydrogenated syrups were spray-dried together with a carrier (alginates, methyl celluloses) [263]. The products have a maltitol content of > 85 % and are very hygroscopic.

Crystallization of maltitol from water was achieved in the early 1980s [264], [265]. Continuous crystallization processes are also being used now [266].

Commercially available crystalline maltitol has the following specifications:

Appearance	white powder
Taste	sweet
Odor	neutral
H_2O content	1 % (max.)
Maltitol content (solids based)	≥ 98 %
D-Sorbitol content (solids based)	0.5 % (max.)
Maltotriitol content (solids based)	1.5 % (max.)
Reducing sugar as glucose %	0.1 % (max.)
mp, °C	ca. 147 °C
Sulfate ash content	0.1 % (max.)
Chloride content	50 mg/kg (max.)
Nickel content	1 mg/kg (max.)

KARRER and BÜCHI [267] and WOLFROM and GARDNER [268] were the first to obtain maltitol in the crystalline state from ethanolic solutions.

Physical and Chemical Properties. Maltitol [*585-88-6*], 4-*O*-α-D-glucopyranosyl-D-sorbitol 4-*O*-α-D-glucopyranosyl-D-glucitol), $C_{12}H_{24}O_{12}$, M_r 344.32, *mp* 146 – 147 °C, optical rotation $[\alpha]_D^{20} = +106.5°$. Maltitol forms orthorhombic crystals [264]. It is weakly hygroscopic in the solid state. The *viscosity* of aqueous maltitol solutions corresponds to that of sucrose solutions (→ Sugar, Chap. 4.). For further physical and chemical parameters, see [269].

In the pH range 5 – 7, maltitol and maltitol solutions are stable up to ca. 150 °C. Lower pH values, temperature > 50 °C, and longer contact times result in cleavage of the disaccharide bond [259].

Maltitol is also cleaved by enzyme systems (e.g., of the mucosa). The published values for the Michaelis constant (K_m) and the maximum reaction rate (V_{max}) differ tremendously, probably because individual researchers have worked with substrates of varying purity [270].

Sensory Properties. The relative sweetness of maltitol is generally given as 0.65 [271], although values up to 0.9 are encountered in chocolates [272]. For the enthalpy of solution, see Table 2, Chapter 1.

Physiological Properties. The general comments in Chapter 1 apply to maltitol. Review

articles deal specifically with maltitol and describe its noncariogenicity [273], suitability for diabetics [274], absorption [275], and tolerance and reduced calorific value [276].

Uses. The applications of maltitol correspond largely to those of sugar. However, its hygroscopicity and capacity for browning or flavor development differ. Maltitol can be used in the production of chocolates [272], hard and soft caramel [277], toffees, gummy bears, chewing gum, jams, ice cream, etc. A general review of uses is given in [278].

7.3. Economic Aspects

Maltitol is produced by Roquette (France), Cerestar (Federal Republic of Germany), Cultor-Kotka (Finland), and Hayashibara (Japan).

8. Erythritol

Osmophilic yeasts (e.g., *Moniliella pollinis*) can form erythritol in ca. 50 % yield from carbohydrates [279] or alkanes [280] as substrate. Arabitol, ribitol, and glycerol are formed as byproducts. The companies Nikken Chemicals [281] in Japan and Cerestar [282], [283] in Europe are working on the production and use of erythritol in the food industry.

Chemical and Physical Properties. Erythritol [*149-32-6*] has a M_r 122 and *mp* 126 °C. Its solubility in water at 20 °C is 37 wt % (58 g/100 g water); the enthalpy of solution in water is + 79.4 kJ/kg. The very high enthalpy of solution characterizes the taste profile of this substance.

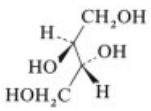

With regard to nutritional physiology, an important difference between erythritol and other polyols is that erythritol is absorbed in the small intestine to a great extent and excreted unchanged via the kidneys [285]. The excreted portion in humans is ca. 80 % of the amount consumed (for a consumption of 25 g) [286].

Because of its high excretion rate via the kidneys, the physiological calorific value of erythritol is said to be ca. 10 % that of sucrose (i.e., 0.4 kcal/kg = 1.7 kJ/g). Erythritol is also suitable for diabetics [287] and is noncariogenic [288].

9. References

General References

1. S. Marie, J. R. Piggott (eds.): *Handbook of Sweeteners,* Blackie and Son Ltd., Glasgow 1991.
2. L. O'Brien Nabors, R. C. Gelardi (eds): *Alternative Sweeteners,* 2nd ed., Marcel Dekker, New York 1991.
3. Lebensmittelchemische Gesellschaft, Fachgruppe in der GDCh (eds.): *Zuckeralkohole und Süßstoffe, Lebensmittelchemie, Lebensmittelqualität,* vol. 19, Behr's Verlag, Hamburg 1992.
4. G.-W. von Rymon Lipinski, H. Schiweck (eds.): *Handbuch Süßungsmittel,* Behr's Verlag, Hamburg 1990.
5. J. L. Multon (Coord.): *Le Sucre, Les Sucres, Les Édulcorants et les Glucides de Charge dans les I.A.A.,* TEC & DOC-Lavoisier, Paris 1992.
6. T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Science Publ. Ltd., London 1989.

Specific References

7. H. Mehnert: *Wissensch. Veröff. der Dtsch. Ges. Ern.,* vol. 20, D. Steinkopf Verlag, Darmstadt 1971, pp. 79 ff.
8. D. A. M. Mackay in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Appl. Science, London 1987, pp. 263 – 285.
9. A. Bär in N. Kretchmer, C. B. Hollenbeck (eds.): *Sugars and Sweeteners,* CRC-Press, Boca Raton 1991, pp. 131 – 150.
10. T. N. Imfeld: *Monographs in Oral Science,* vol. 11, Karger, Basel 1983.
11. K.-H. Bässler, *Ernähr. Umsch.* **36** (1989) 395 – 401. A. Bär in N. Hosoya (ed.): *Caloric Evaluation of Carbohydrates,* Research Foundation for Sugar Metabolism, Tokyo 1990, pp. 209 – 257. A. J. H. van Es, *Proc. Nutr. Soc.* **50** (1991) 383 – 390.

12. J. Beereboom in B. K. Dwivedi (ed.): *Low Calorie and Special Dietary Foods,* CRC Press Inc., Palm Beach 1978, pp. 1 – 22.
13. EC-Directive 90/496, 24.09.90, Richtlinie über Nährwertkennzeichnung von Lebensmitteln.
14. Nutrition Council: "The Energy Value of Sugar Alcohols," Recommendations of the Committee on Polyalcohols, The Hague 1987.
15. LSRO – FASEB: "The Evaluation of the Energy of Certain Sugar Alcohols used as Food Ingredients", Bethesda, MD 1994.
16. G. Livesey, *Food Sci. Nutr.* **42 F** (1990) 223 – 243. D. Krüger et al., *Experientia* **48** (1992) 733 – 740.
17. U. Nilsson, M. Jägerstad, *Br. J. Nutr.* **58** (1987) 199 – 206. S. C. Ziesenitz, G. Siebert in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Applied Science, London 1987, pp. 109 – 149.
18. A. Zumbe, R. A. Brinkworth, *Z. Ernährungswiss.* **31** (1992) 40 – 48.
19. D. M. Paige, T. M. Bayless, L. R. Davis, *Nutr. Res. (N.Y.)* **12** (1992) 27 – 37.
20. Commission of the European Communities, Directorate-General for Internal Market and Industrial Affairs, III/A/2, Scientific Committee for Food, "Report of the Scientific Committee for Food and Sweeteners," III/1316/84, CS/EDUL/27 rev., Dec. 1984.
21. FAO/WHO, JECFA Specifications for Identity and Purity, FAO Food and Nutrition Paper 28, Rome 1983. FAO/WHO Expert Committee on Food Additives, Evaluation of Certain Food Additives and Contaminants, 30. Report, WHO Technical Report Series no. 751, 25 Geneva 1987.33. Report, WHO Technical Report Series 776, Annex 2, Geneva 1989.
22. FAO/WHO, Food and Nutrition Paper 38, Roma 1988.
23. Commission of the European Communities, Counsil Directive 94/35/EC on Sweeteners for Use in Foodstuffs from June 30, 1994 (ABl L 237, from September 10, 1994, pp. 3 – 12).
24. Affirmation of G.R.A.S: "Status of Direct Human Food Ingredients: Mannitol, Sorbitol," *Fed. Reg.* **38** (1973) no. 143, 20 046 – 20 048.
25. J. N. Counsel: *Xylitol,* Applied Science Publishers, London 1987, p. 7.
26. *Kirk-Othmer,* 3rd ed., **vol.** 1, 756.
27. *Ullmann,* 4th ed., **24,** 778.
28. L. O'Brien Nabors, R. C. Gelardi (eds.): *Alternative Sweeteners,* 2nd ed., Marcel Dekker, New York 1991, p. 260.
29. Leatherhead Food RA, Scientific & Technical Surveys, J. U. Kristott, S. A. Jones (eds.): *Physical Properties of Polyols, Polydextrose and their Solutions,* **no. 173,** Surrey, June 1992.
30. A. Bär in L. O'Brien Nabors, R. C. Gelardi (eds.): *Alternative Sweeteners,* 2nd ed., Marcel Dekker, New York 1991, pp. 349 – 379.
31. A. Bär, *World Rev. Nutr. Diet.* **55** (1988) 183 –209.
32. A. Bär in N. Kretchmer, C. B. Hollenbeck (eds.): *Sugars and Sweeteners,* CRC Press, Boca Raton 1991, pp. 131 – 150.
33. M. Georgieff, L. L. Moldawer, B. R. Bistrian, G. L. Blackburn, *JPEN J. Parenter. Enteral Nutr.* **9** (1985) 199 – 209.
34. K. K. Mäkinen in T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Appl. Sci., London 1989, pp. 331 – 362.
35. K. K. Mäkinen: *Biochemical Principles of the Use of Xylitol in Medicine and Nutrition with Special Consideration of Dental Aspects,* Birkhäuser, Basel 1978, pp. 160 ff.
36. T. Pepper, P. M. Olinger, *Food Technol. (Chicago)* **42** (1988) 98 – 106.
37. K. K. Mäkinen, E. Söderling, *J. Food Sci.* **45** (1980) 367 – 374.
38. J. Washüttl, P. Riederer, E. Bancher, *J. Food. Sci.* **38** (1973) 1262 – 1263.
39. O. Touster in H. L. Sipple, K. W. McNutt (eds.): *Sugars in Nutrition,* Academic Press, New York 1974, pp. 229 – 239.
40. J. F. Carson, S. W. Waisbrot, F. T. Jones, *J. Am. Chem. Soc.* **65** (1943) 1777 – 1778.
41. M. L. Wolfrom, E. J. Kohn, *J. Am. Chem. Soc.* **64** (1942) 1739.
42. FAO Food and Nutrition Paper, no. 38, 1988, pp. 267 – 271.
43. Food Chemicals Codex, 3rd ed., National Academy Press, Washington D.C. 1981, pp. 348 –349.
44. C. Aminoff, E. Vanninen, T. E. Doty in J. N. Counsell (ed.): *Xylitol,* Applied Science Publ., London 1978, pp. 1 – 9.
45. L. F. Chen, C. S. Gong, *J. Food Sci.* **50** (1985) 226 –228.
46. D. Holland, J. F. Stoddart, *Tetrahedron Lett.* **23** (1982) 5367 – 5370.
47. J. Kiss, R. D'Souza, P. Taschner, *Helv. Chim. Acta* **58** (1975) 311 – 317.
48. U.S. Pharmacopeia XXII/National Formulary XVII, Suppl. 5, 1990, pp. 2747 – 2748.
49. E. C. Samarco, E. S. Parente, *J. Assoc. Off. Anal. Chem.* **65** (1982) 76 – 78.
50. C. Vidal-Valverde, C. Martin-Villa, B. Olmedilla, I. Blanco, *J. Liq. Chromatogr.* **8** (1985) no. 1, 75 –94.

51. D. W. Armstrong, H. L. Jin, *J. Chromatogr.* **462** (1989) 219 – 232.
52. Commission of the European Communities. Proposals for Council Directives on Sweeteners for Use in Foodstuffs and on Food Additives other than Colours and Sweeteners, (OJ no. C 206/3 – 11, 13. 8. 1992; OJ no. C 172/4 – 45, 24. 6. 1994).
53. United States Code of Federal Regulations, 21 CFR 172.395 U.S. Government Printing Office, Washington D.C., 1993, p. 14.
54. Council Directive on Nutrition Labelling for Foodstuffs (90/496/EEC), OJ no. L 276/40 – 44, 6.10.1990.
55. A. Bär in N. Hosoya (ed.): *Caloric Evaluation of Carbohydrates,* Research Foundation for Sugar Metabolism, Tokyo 1990, pp. 209 – 257.
56. C. Auty, *Intl. Food Ingred.* 1992, no. 1, 48 – 49.
57. C. Krüger, *Süsswaren* **11** (1987) 506 – 516.
58. C. Krüger, *Zucker Süsswaren Wirtsch.* **41** (1988) 360 – 365.
59. A. Askar et al., *Dtsch. Lebensm. Rundsch.* **83** (1987) 389 – 394.
60. S. M. Waaler, *Scand. J. Dent. Res.* **100** (1992) 204 –206.
61. P. Isokangas et al., *Caries Res.* **25** (1991) 444 – 448.
62. D. Kandelman, G. Gagnon, *J. Dent. Res.* **69** (1990) 1771 – 1775.
63. K. K. Mäkinen et al., *J. Dent. Res.* **72 (special issue)** (1993) 346.
64. F. Voirol, R. Etter, DE-OS 2 913 555, 1979.
65. M. Georgieff et al., *Infusionsther. Klin. Ernähr. Forsch. Prax.* **14 (suppl. 1)** (1987) 53 – 64.
66. M. Oehmke, EP-A 0 150 053, 1985.
67. B. Schmidt, M. Fingerhut, K. Lang, *Klin. Wochenschr.* **42** (1964) 1073 – 1077.
68. WHO/FAO, WHO Technical Report Series no. 696, 1983, pp. 23 – 24.
69. Commission of the European Communities. Reports of the Scientific Committee for Food, 16th Series, 1985, p. 16.
70. I. Boussingault, *J. Pharm. Chim.* **40** (1872) 36.
71. R. E. Wrolstadt, R. S. Schallenberger, *J. Assoc. Off. Anal. Chem.* **64** (1981) 91.
72. O. Touster, D. R. D. Shaw, *Physiol. Rev.* **16** (1962) 261.
73. H. G. Hers, *Biochem. J.* **66** (1957) 30.
74. M. W. Neil, D. G. Walker, F. L. Warren, *Biochem. J.* **65** (1957) 35.
75. E. Fischer, *Ber. Dtsch. Chem. Ges.* **23** (1890) 3684.
76. W. Ipatieff, *Ber. Dtsch. Chem. Ges.* **45** (1912) 3225.
77. Y. J. Park, G. A. Jeffrey, W. C. Hamilton, *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 27** (1971) 2393.
78. I. A. Millis, *Biochem. Biophys. Res. Commun.* **6** (1962) 217.
79. Kokai Tokkyo Koho, JP 63 313 571, 1988 (T. Kitayama).
80. M. Barr, L. F. Tice, *J. Am. Pharm. Assoc.* **46** (1957) 217.
81. L. Wright, L. Hartmann, *J. Org. Chem.* **26** (1961) 1583.
82. ICI, US 4 297 290, 1980 (G. J. Stockburger).
83. H. Velde, *Seifen öle Fette Wachse* **29** (1956) 778.
84. S. Ropuszynski, J. Perka, *Wiad. Chem.* **23** (1969) 297.
85. Ch. M. Upadhyay et al., *Starch/Stärke* **43** (1991) 107.
86. Roquette Frères, EP-A 168 315, 1986 (F. Devos, M. Huchette).
87. VEB Deutsches Hydrierwerk Rodleben, DD 252 003, 1987 (H. Aring et al.).
88. IG-Farben, DE 544 666, DE 554 074, 1925 (J. Müller, A. Hoffmann).
89. E. Haidegger, *Stärke* **29** (1970) 430.
90. R. Albert, A. Strätz, G. Vollheim, *Chem. Ing. Tech.* **52** (1980) 582.
91. S. L. Paterson et al., *Biocatalasis* **1** (1988) 217.
92. D. M. Kim, H. S. Kim, *Biotechnol. Bioeng.* **39** (1992) 336.
93. B. Rehr, C. Wilhelm, H. Sahm, *Appl. Microbiol. Biotechnol.* **35** (1991) 144.
94. Forschung Zentrum, Jülich, DE 3 936 757, 1991 (B. Rehr, H. Sahm).
95. Y. Tani, V. Vongsuvanlert, *J. Ferment. Technol.* **65** (1987) 405.
96. Z. Durnjak, G. Turcotte, Z. D. Duan, *Appl. Microbiol. Biotechnol.* **35** (1991) 714.
97. Pfizer Inc., US 4 694 113, 1987 (G. J. Bautier, J. D. Miceli).
98. A. G. Ivchenko, Yu. S. Khakimov, M. F. Abidova, *Uzb. Khim. Zh.* **3** (1990) 66.
99. Atlas Powder, US 2 642 462, 1953 (L. Kasehagen).
100. Hydrocarbon Res., DE 3 110 493, 1981 (Chao James C., D. Huibers).
101. ICI, EP 0 032 288, 1984 (J.-W. Du Ross).
102. E. Haidegger, *Ind. Eng. Chem. Proc. Res. Dev.* **7** (1968) 107.
103. Yu. S. Khakimov, M. F. Abidova, *Uzb. Khim. Zh.* **1** (1989) 58.
104. Roquette Frères, DE 2 350 619, 1973 (M. Huchette, H. Mangin).
105. Kabushiki Kaisha, EP 330 352, 1989 (Yuji Kunimi, Akihiko Tabata, Yatsuka Fugita).

106. Niro Atomizer, DK 133 603, 1968 (N. C. Holt, O. E. Hansen, O. Adreassen).
107. VEB Deutsches Hydrierwerk, DD 277 176, 1990 (H. Aring et al.).
108. P. C. Schmidt, *Pharm. Technol.* **7** (1983) 65.
109. S. Quinquenet et al., *Anal. Therm. Thermodyn. Chim.* **17** (1986) 304.
110. FAO-WHO Food and Nutrition Paper, nos. 28 and 38, Rome 1988.
111. A. Turner, *Analyst (London)* **89** (1964) 115.
112. A. Rotsch, G. Freese, *Dtsch. Lebensm. Rundsch.* **60** (1964) 343.
113. U. Stoll, *Mitt. Geb. Lebensmittelunters. Hyg.* **58** (1967) 56.
114. E. Benk, R. Kaa, *Brauereitechniker* **23** (1971) 26.
115. P. Meier, B. Gramm, U. Vögeli, *Mitt. Geb. Lebensmittelunters. Hyg.* **76** (1985) 478.
116. H. Auterhoff, V. Lohmann, *Dtsch. Apoth. Ztg.* **97** (1957) 904.
117. L. H. Adcock, *Analyst (London)* **82** (1957) 427.
118. *Deutsches Arzneibuch,* 10th ed., Deutscher Apotheker Verlag, Stuttgart; Govi-Verlag, Frankfurt 1991.
119. R. Klaus, J. Ripphahn, *J. Chromatogr.* **244** (1982) 99.
120. G. Manius, *J. Pharm. Sci.* **61** (1972) 1831.
121. B. Müller, G. Göke, *Dtsch. Lebensm. Rundsch.* **68** (1972) 222.
122. D. L. Sondack, *J. Pharm. Sci.* **64** (1975) 128.
123. H. D. Daniels, Ch. R. Warner, Th. Fazio, *J. Assoc. Offic. Anal. Chem.* **65** (1982) 588.
124. NF XVII, United States Pharmacopeial Convention, Inc., Rockville 1990.
125. J. Dokladalova, A. Y. Barton, E. A. Mackenzie, *J. Assoc. Offic. Anal. Chem.* **63** (1980) 644.
126. E. C. Samarco, E. S. Parente, *J. Assoc. Offic. Anal. Chem.* **65** (1982) 76.
127. R. Galensa, *Z. Lebensm. Unters. Forsch.* **176** (1983) 417.
128. L. C. Zygmunt, *J. Assoc. Offic. Anal. Chem.* **65** (1982) 256.
129. H. U. Bergmeyer, W. Gruber, J. Gutman in H. U. Bergmeyer (ed.): *Methoden der enzymatischen Analyse,* 3rd ed., **vol. 2,** Verlag Chemie, Weinheim 1974, p. 1368.
130. K. H. Bässler, K. Wagner, B. Schoenerstedt, *J. Clin. Chem. Clin. Biochem.* **16** (1979) 547.
131. L. Klingebiel, R. Grossklaus, G. Pahlke, *Z. Lebensm. Unters. Forsch.* **169** (1980) 359.
132. LSRO for FDA No. 223-75-2090, FASEB, Bethesda, Md., 1979 (p. 85).
133. Life Sciences Research Office (LSRO) for FDA No. 223-83-2020, Federation of American Societies for Experimental Biology (FASEB) Bethesda, Md., 1986 (p. 85).
134. R. Vogel: *BDL-Spectrum,* **vol. 2,** Rhenania Fachverlag, Hamburg 1986, p. 85.
135. G. W. von Rymon-Lipinski, H. Schiweck (eds.): *Handbuch der Süßungsmittel (Eigenschaften u. Anwendung),* B. Behr's Verlag, Hamburg 1991, p. 265.
136. Lebensmittelchemische Gesellschaft (eds.): *Zuckeralkohole und Süßstoffe,* **vol. 19,** B. Behr's Verlag, Hamburg 1992, p. 49.
137. H. Völker, *Süßwaren* **5** (1961) 1089; **9** (1965) 1245.
138. H. Völker, *Zucker-Süßwaren Wirtsch.* **19** (1966) 196.
139. E. Thieme, *Gordian* **72** (1972) 132.
140. R. Caliari, *Manuf. Confect.* **63** (1983) 25.
141. I. Du Ross, *Manuf. Confect.* **62** (1982) 35.
142. W. Huhn, *Riechst. Aromen* **9** (1959) 416.
143. H. C. Janowitz, *Soap Perfum. Cosmet.* **34** (1961) 56.
144. G. Reese, *ärztl. Kosmetol.* **11** (1981) 56.
145. S. I. Strianse, *Am. Perfum. Cosmet.* **77** (1962) no. 10, 31.
146. I. H. Shay, D. Griffith, *J. Dent. Res.* **39** (1960) 377.
147. Colgate-Palmolive, DE 3 701 123, 1987 (S. L. Schelm, L. J. Vellekoop).
148. Colgate-Palmolive, US 4 701 318, 1987.
149. F. Neuwald, *Pharm. Ztg.* **88** (1952) 98.
150. H. von Czetsch-Lindenwald, *Dtsch. Apoth. Ztg.* **94** (1954) 433.
151. J. Breinlich, *Dtsch. Apoth. Ztg.* **96** (1956) 709.
152. H. Guyot, A. M. Hermann, D. Leblanc, M. Draguet-Brughmans, *Drug Dev. Ind. Pharm.* **11** (1985) 551.
153. S. Bolton, R. Atluri, *Drug Cosmet. Ind.* **135** (1984) 44.
154. A. M. Basedow, A. G. Möschl, P. C. Schmidt, *Drug Dev. Ind. Pharm.* **12** (1986) 2061.
155. P. C. Schmidt, W. Vortisch, *Pharm. Ind.* **49** (1987) 495.
156. P. C. Schmidt, K. Benke, *Drugs Made Ger.* **28** (1985) 49; *Ger. Pharm. Ind.* **46** (1984) 193.
157. I. Nikolakakis, J. M. Newton, *J. Pharm. Pharmacol.* **41** (1989) 145.
158. P. C. Schmidt, *Pharm. Ztg.* **132** (1987) 1501.
159. Merck, DE 3 315 800, 1984 (A. M. Basedow, P. Lembke, R. Vogel).
160. H. Bickel, F. Matzkies, W. Fekl, G. Berg, *Dtsch. Med. Wochenschr.* **98** (1973) 2079.
161. W. Griem, K. Lang, *Klin. Wochenschr.* **38** (1960) 951.
162. H. Mehnert, *Med. Ernähr.* **11** (1970) 77.
163. M. Haslbeck, *Infusionstherapie* **1** (1974) 569.

164. G. Berg, F. Matzkies, *Klin. Wochenschr.* **53** (1975) 187.
165. U. C. Dubach, *Int. J. Vitam. Nutr. Res.* **15** (1976) 81.
166. H. F. Förster, *Z. Ernährungswiss.* **17** (1978) 210.
167. S. Wich, H. Causemann, *Med. Welt.* **2** (1962) 609.
168. L. Rothamel, *Allg. Pap. Rundsch.* **12** (1955) 526.
169. K. Hajek, *Farbe + Lack* **83** (1977) 798.
170. U. Wintermeyer: *Vitamin C, Entdeckung, Identifizierung und Synthese,* Deutscher Apotheker Verlag, Stuttgart 1981.
171. F. R. Benson in M. Schick (ed.): *Nonionic Surfactans,* Marcel Dekker, New York 1967, p. 247.
172. Kokai Tokkyo Koko, JP 61 186 347, 1986 (H. Kyono, T. Uschibori, S. Inamasu, H. Nishitani).
173. Kokai Tokkyo Koko, JP 03 115 208, 1991 (F. Sugizaki).
174. W. N. Duck, *Manuf. Confect.* **37** (1957) no. 4, 14.
175. L. R. Chislett, J. Walford, *Int. Flavours Food Addit.* **7** (1976) 61.
176. H. F. Förster, H. Mehnert, *Aktuel. Ernährungsmed.* **5** (1979) 245.
177. R. Peters, R. H. Lock, *Br. Med. J.* **5097** (1958) 677.
178. L. H. Adcock, C. H. Gray, *Biochem. J.* **65** (1957) 554.
179. D. Birkhed, *Caries Res.* **12** (1978) 128.
180. V. Harjola, H. Liesmaa, *Acta Odontol. Scand.* **36** (1978) 237.
181. J. J. Møller, *Dtsch. Zahnärztl. Z.* **32** (1977) 66 (Suppl. 1).
182. D. Birkhed, A. Bär, *World Rev. Nutr. Diet.* **65** (1991) 1 – 37.
183. 26th Report of Joint FAO-WHO Expert Comm. of Food Addit., Tech. Rep. Ser. 683, Geneva 1982.
184. H. Mehnert, *Int. J. Vitam. Nutr. Res.* **32** (1976) 295.
185. H. F. Förster, *Med. Monatsschr. Pharm.* **1** (1978) 42.
186. A. Bär in ASPEC-Inform. Sorbitol, c/o ECCO, 19 Rue de L'orme, B-1040 Brüssel (1984).
187. K. M. Mackenzie, W. W. Hauck, A. G. Wheeler, F. C. J. Roe, *Food Chem. Toxicol.* **24** (1986) 191.
188. H. Mehnert, *Dtsch. Med. Wochenschr.* **103** (1978) 229.
189. L. Proust, *Ann. Chim. Phys.* **57** (1806) 1.
190. D. Le Tourneau, *Trans. Br. Mycol. Soc.* **62** (1974) 618.
191. I. A. Millis, *Biochem. Biophys. Res. Commun.* **6** (1962) 217.
192. F. Köchel, *Mitt. Dtsch. Pharm. Ges.* **34** (1964) 153.
193. R. Meyer: *Explosivstoffe,* 6th ed., Verlag Chemie, Weinheim 1985, p. 233.
194. Suomen Sokery Oy, US 2 418 801, 1973 (A. Malaja, L. Hämäläinen).
195. Atlas Chem. Ind. Inc. US 7 38 410, 1968 J. Brandner, L. Wright, D. Wilmingn.
196. ICI, GB 1 506 534, 1977 (K. Kruse).
197. ICI, US 4 029 878, 1977 (K. Kruse).
198. US Dept. of Agriculture, US 3 427 224, 1969 (K. Smiley et al.).
199. Noda Institute for Scientific Research JP, US 3 622 456, 1971 (H. Onishi, T. Suzuki).
200. W. Soetaert, E. J. Vandamme, *Bioorg. Chem. Healthcare Technol.* **207** (1991) 249 – 250.
201. K. Kulbe, U. Schwab, U. Howaldt, K. Kümmerle, *GBF Monogr. Ser.* **9** (1986) 189 – 200.
202. FAO/WHO-Food and Nutrition Paper no. 38, Rome 1988.
203. *Deutsches Arzneibuch,* 10th ed., Deutscher Apotheker Verlag Stuttgart; Govi-Verlag GmbH, Frankfurt 1991.
204. BP 88; Her Majesty's Stationary Office, London (1988).
205. USP XXII, United States Pharmacopeial Convention, Inc., Rockville 1990.
206. Roquette Frères, EP 177 428, 1986 (M. Serpelloni, P. Lemay).
207. L. Gever, *Nursing* **15** (1985) 64 – 72.
208. F. Heppner, G. Lanner, L. Auer, *Wien. Med. Wochenschr.* **126** (1976) 105.
209. U. Hase, H. I. Reulen, *Neurochirurgia* **23** (1980) 205.
210. K. R. Palmer, A. N. Khan, *Gut* **20** (1979) no. 10, AG 28.
211. Warner Lambert, US 563 230, 1990 (S. Cherukuri, S. Faust).
212. Nabisco Brands, EP 166 603, 1986 (W. Steensen, R. Weaver).
213. S. M. Nasrallah, F. L. Iber, *Am. J. Med. Sci.* **258** (1969) 80 – 88.
214. O. Touster, D. R. D. Slaw, *Physiol. Rev.* **42** (1962) 181.
215. A. Wretland, *Nutr. Metab.* **18** (1975) Suppl. 1, 242.
216. P. S. Elias: BGA Schriften 1/88 MMV Medizin Verlag, München 1988.
217. Y. M. Wang, J. van Eys, *Annu. Rev. Nutr.* **1** (1981) 437 – 475.

218. D. R. Sounders, H. S. Wiggins, *Am. J. Physiol.* **241** (1981) no. 5, 397 – 402.
219. K. M. Abdo, J. K. Haseman, G. Boorman, *Food Chem. Toxicol.* **21** (1983) 259.
220. F. H. Stodola, E. S. Sharpe, H. J. Koepsell, *J. Am. Chem. Soc.* **78** (1956) 2514.
221. Süddeutsche Zucker-AG, DE 1 049 800, 1957 (R. Weidenhagen, S. Lorenz).
222. Süddeutsche Zucker-AG, DE 2 217 628, 1972 (H. Schiweck et al.).
223. Süddeutsche Zucker-AG, EP 0 109 009, 1983 (M. Munir et al.).
224. F. Lichtenthaler, S. Rönninger, *J. Chem. Soc. Perkin Trans. 2* 1990, 1489.
225. M. Munir, B. Schneider, H. Schiweck, *Carbohyd. Res.* **164** (1987) 477.
226. F. W. Lichtenthaler (ed.): *Carbohydrates as Industrial Raw Materials,* VCH Verlagsgesellschaft, Weinheim 1991.
227. H. Schiweck, M. Munir in M. A. Clarke (ed.): *Proc. Symp. Div. Carbohyd. Chem. of Am. Chem. Soc.,* Bartens, Berlin 1992.
228. Mitsui Sugar Tokyo und Südzucker, EP 0 483 755, 1991 (T. Sugitani et al.).
229. F. W. Lichtenthaler, H. J. Lindner, *Liebigs Ann. Chem.* 1981, 2372.
230. H. J. Lindner, F. W. Lichtenthaler, *Carbohyd. Res.* **93** (1981) 135.
231. K. Mende, H. Schiweck, *Alimenta* **5** (1991) 82; *Alimenta* **6** (1991) 115.
232. Grillo-Werke, Südzucker, WO 92/13 866, 1992 (B. Schneider et al.).
233. Bayer, DE 3 403 973, 1984 (G. Darsow, W. Biedermann).
234. P. J. Sträter in L. O. Nabors, R. C. Gelardi (eds.): *Alternative Sweeteners,* 2nd ed., Marcel Dekker, New York 1991.
235. J. B. Sendrens, *Comptes Rendus* **170** (1920) 47 – 50.
236. P. Karrer, J. Büchi, *Helv. Chim. Acta* **20** (1937) 86 –90.
237. R. Albert, A. Straitz, G. Vollheim, *Chem. Ing. Tech.* **52** (1980) 582 – 587.
238. C. V. Chemie Combinatie Amsterdam, EP 039 981, 1981 (C. F. Wignman, J. A. van Velthuijzen, H. van den Berg). Cultor Ltd., Helsinki, Finland, WO 90/06 317, 1990 (H. O. Heikkilä; V. Nurmi).
239. Tow Chemical Industry, Tokyo 100 (JP), EP 0 381 483, 1990 (S. Kawashina; H. Ide; K. Kato; M. Magora; Y. Ishii).
240. M. L. Wolfram, W. J. Burkl, K. R. Brown, R. J. Rose, Jr., *J. Am. Chem. Soc.* **60** (1938) 571 – 572.
241. J. Kivikoski, I. Pitkänen, J. Nurmi, *Carbohydr. Res.* **232** (1992) 189 – 195.
242. J. A. van Velthuijzen, *J. Agric. Food Chem.* **27** (1979) 680 – 686.
243. Jr. J. H. Blankers, *Süsswaren* **35** (1991) 91 – 94.
244. Leatherhead Food RA, Scientific & Technical Surveys, J. U. Kristott, S. A. Jones (eds.): *Physical Properties of Polyols, Polydextrose and their Solutions,* no. 173, Surrey, June 1992.
245. Maizena, DE 2 133 428, 1974 (G. Graefe; H. Gudermann; H. Müller).
246. M. Harju, *Milchwissenschaft* **43** (1988) 76 – 78, 148 – 152.
247. S. Yanahira et al., *Carbohydr. Res.* **232** (1992) 151 –159.
248. K. Hoppe, B. Gassmann, *Nahrung* **29** (1985) 417 –420. K. Hoppe, *Lebensmittelindustrie* **38** (1991) 13 –14.
249. T. H. Grenby, *Int. Dent. J.* **39** (1989) 25 – 32. T. H. Grenby in T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Applied Science London 1989, pp. 363 – 379.
250. J. Bircher, M. Bühler, K. Franz, J. A. van Velthuijzen, *Schweiz. Med. Wochenschr.* **112** (1982) 1306 –1307.
251. L. Beaugerie et al., *Gastroenterol. Clin. Biol.* **15** (1991) 929 – 932. S. P. Bird, D. Hewitt, M. I. Gurr, *J. Sci. Food Agric.* **51** (1990) 233 – 246. A. J. van Es, L. De Groot, J. E. Vogt, *Br. J. Nutr.* **56** (1986) 545 – 554.
252. O. Riggio, G. Balducci, K. Ziparo, L. Capocaccia, *Hapato gastroenterology* **37** (1990) 524 – 527.
253. D. H. Patil, D. Estaby et al., *Gut* **28** (1987) 255 –259. J. Vanderdonckt, J. Coulon, W. Denys, G. P. Ravelli, *J. Clin. Exp. Gerontol.* **12** (1990) 171 – 189.
254. C. H. den Uyl in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Applied Science, London 1987, pp. 65 – 81.
255. Lyckeby Stärkelsförädling DE 1 270 019, 1964.
256. Roquette Frères, DE-OS 2 949 783, 1979 (F. Verwaerde; I.-B. Leleu; M. Huchette).
257. M. W. Kearsley, G. G. Birch, *Stärke* **29** (1977) 425 – 429.
258. S. Z. Dziedzic, M. W. Kearsley, *Confect. Prod.* **56** (1990) 739 – 740. A. Rapaille, *Starch/Stärke* **40** (1988) 356 – 359.
259. J. Szejtli: *Säurehydrolyse glykosidischer Bindungen,* VEB Fachbuchverlag, Leipzig 1976, pp. 69, 172 –176.

260. A. J. Rugg-Gunn in T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Applied Science, London 1989, pp. 311 – 329.
261. M. W. Kearsley, G. G. Birch, R. H. P. Lian-Loh, *Starch/Stärke* **34** (1982) 279 – 283. M. L. Wheeler, S. E. Fineberg, R. Gibson, N. Fineberg, *Diabetes Care* **13** (1990) 733 – 740.
262. Hayashibara, DE-OS 1 903 074, 1969 (M. Mitsuhashi; K. Sugimato; M. Hirao); DE-AS 2 034 700, 1970.
263. Kowa, DE-OS 2 319 510, 1973.
264. Hayashibara, DE 3 146 085, 1981 (M. Hirao; H. Hijiya; T. Miyake); US 4 408 041, 1981.
265. S. Ohno, M. Hirao, M. Kido, *Carbohydr. Res.* **108** (1982) 163 – 171.
266. Roquette Frères, EP 0 202 165, 1986 (P. A. Gony et al.).
267. P. Karrer, J. Büchi, *Helv. Chim. Acta* **20** (1937) 86 –90.
268. M. L. Wolfrom, T. S. Gardner, *J. Am. Chem. Soc.* **62** (1940) 2553 – 2555.
269. Leatherhead Food RA, Scientific & Technical Surveys, J. U. Kristott, S. A. Jones (eds.): *Physical Properties of Polyols, Polydextrose and their Solutions,* no. 173, Surrey, June 1992.
270. H. J. Zunft, J. Schulze, H. Gärtner, F. K. Grütte, *Ann. Nutr. Metab.* **27** (1983) 470 – 476. C. Rosiers, F. Verwaerde et al., *Ann. Nutr. Metab.* **29** (1985) 76 – 82. H. Schiweck, *Lebensmittelchem. Gerichtl. Chem.* **41** (1987) 49 – 55. U. Nilsson, M. Jägerstad, *Br. J. Nutr.* **58** (1987) 199 – 206.
271. P. Würsch, *Ernährung/Nutrition* **11** (1987) 334. G.-W. von Rymon Lipinski, *Alimenta* **25** (1986) 119 – 122.
272. W. Weber, *Int. Z. Lebensmitteltechnik* **43** (1992) 708 – 709.
273. T. Ooshima et al., *Caries Res.* **26** (1992) 33 – 37. T. Imfeld, F. Lutz, *Swiss. Dent.* **6** (1985) 44 – 53.
274. S. Moon, M. Lee, K. Huh, K. Lee, *Hanguk Yongyang Hakhoe Chi* **23** (1990) 270 – 275. L. Wolfslehner, *Ernährung/Nutrition* **12** (1988) 788 –792.
275. T. Oku et al., *J. Nutr. Sci. Vitaminol.* **37** (1991) 529 –544.
276. L. Beaugerie et al., *Gastroenterology* **99** (1990) 712 –723. J. P. Felber et al., *J. Parenter. Enteral Nutr.* **11** (1987) 250 – 254.
277. Roquette Frères, EP 094 292, 1987.
278. I. Fabry in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Applied Science, London 1987, pp. 83 – 108. G. Celia, *Confectionery Manufacture and Marketing* **22** (1985) 25 – 26.
279. J. F. T. Spencer, *Prog. Ind. Microbiol.* **7** (1968) 1 – 42.
280. K. Hattori, T. Suzuki, *Agr. Biol. Chem.* **38** (1974) no. 3, 581 – 586; **38** (1974) no. 6, 1203 – 1208.
281. K. Wako et al.: "Development of Production Technology for Erythritol – A new Low Calorie Sweetener," XIVth International Carbohydrate Symposium, 14 – 19 Aug. 1988, Stockholm, Poster D 4.
282. CPC Intern., EP 136 805, 1984 (J.-C. de Froostenberg; B. Avalosse; K. L. Migudet).
283. J. Goossens, H. Röper in G. Descotes (ed.): *Carbohydrates as Organic Raw Materials,* **vol. 2,** VCH Verlagsgesellschaft, Weinheim 1993, pp. 27 – 38.
284. H. Röper, J. Goossens: *Starch/Stärke* **45** (1993) 400 – 405.
285. T. Oku, K. Noda, *Nutr. Res.* **10** (1990) 987 – 996.
286. M. Hiele, Y. Ghoos, P. Rutgeerts, G. Vantrappen: "Metabolism of Erythritol in Humans: Comparison with Glucose and Lactitol," *Br. J. Nutr.* **69** (1993) 169 – 176.
287. F. Bornet, F. Dauchy, A. Chevalier, G. Slama: "Etude de devenir metabolique, après ingestion chez l'homme sain, d'un nouvel edulcorant de charge basse calorie: l'erythritol," *Gastroenterol. Clin. Biol.* **16** (1992) A 169.
288. J. Kawanabe, M. Hirasawa, T. Takeuchi, T. Oda, T. Ikeda: "Non-Cariogenicity of Erythritol as a Substrate," *Caries Res.* **26** (1992) 358 – 362.

Sulfamic Acid

ADOLF METZGER, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 1
4. **Production** 2
5. **Uses** 3
6. **Toxicology** 3
7. **Sulfamates** 3
8. **References** 3

1. Introduction

Sulfamic acid (IUPAC recommendation: amidosulfuric acid) has the formula NH_2SO_3H.

Sulfamic acid is a crystalline, strong inorganic acid. It has been produced industrially for ca. 50 years. In expectation of good sales, production plants were set up in a number of industrial countries during the 1950s to 1980s. However, as a result of a decrease in demand, production difficulties, and problems with the disposal of byproducts, all the plants in Europe and the United States have been closed down. Plants are still operating in Japan (Nissan) and Taiwan (several producers).

Table 1. Solubility of sulfamic acid in water

Temperature, °C	Solubility, g/100 g	Concentration, wt %
0	14.7	12.8
10	18.5	15.6
20	21.3	17.5
30	26.1	20.7
40	29.5	22.8
50	32.8	24.7
60	37.1	27.0
70	41.9	29.5
80	47.0	32.0

2. Physical Properties

Pure sulfamic acid is a white, odorless, crystalline, nonhygroscopic compound. It is not volatile at room temperature. Some important data are given in [1], [2].

Density, 25 °C	2.126 g/cm^3
Specific heat	1.1467 J/g
Melting point	205 °C
Vapor pressure	
20 °C	0.8 Pa
100 °C	0.25 Pa
Dissociation constant	1.10×10^{-1}

Sulfamic acid is strongly dissociated in aqueous solution. The pH of a 1 % solution is 1.18. The variation in water solubility with temperature is shown in Table 1. The solubility in water can be increased by the addition of other acids or salts. Sulfamic acid is only sparingly soluble in concentrated inorganic acids, and most organic solvents.

3. Chemical Properties

Sulfamic acid decomposes on heating above 209 °C. Sulfur trioxide, sulfur dioxide, water, ammonia, and nitrogen have been detected as decomposition products [3].

Aqueous solutions of sulfamic acid hydrolyze to ammonium hydrogen sulfate:

$$NH_2SO_3H + H_2O \longrightarrow NH_4HSO_4$$

The rate of hydrolysis depends on concentration, pH, and temperature. Dilute, aqueous sulfamic acid solutions are stable at room temperature. A 10 % solution is 50 % hydrolyzed at 80 °C in 10 h [4].

Chlorine, bromine, and chlorates oxidize sulfamic acid to sulfuric acid:

$$2NH_2SO_3H + KClO_3 \longrightarrow 2H_2SO_4 + N_2 + KCl + H_2O$$

Concentrated nitric acid reacts to give dinitrogen oxide on warming:

$$NH_2SO_3H + HNO_3 \longrightarrow H_2SO_4 + N_2O + H_2O$$

Nitrous acid converts sulfamic acid completely to nitrogen and sulfuric acid, so this reaction can be used for quantitative analysis:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_439

$$NH_2SO_3H + HNO_2 \longrightarrow H_2SO_4 + H_2O + N_2$$

Reaction with thionyl chloride forms sulfamyl chloride:

$$NH_2SO_3H + SOCl_2 \longrightarrow ClSO_2NH_2 + SO_2 + HCl$$

Metal hydroxides, oxides, and carbonates are dissolved by sulfamic acid.

The reactions of sulfamic acid with organic compounds have been described in detail [5]. Sulfamic acid reacts with primary and secondary alcohols to alkylammonium sulfates:

$$ROH + NH_2SO_3H \longrightarrow ROSO_2ONH_4$$

Secondary alcohols react only in the presence of amines, which act as catalysts. Tertiary alcohols do not react. Aromatic alcohols, such as phenol, give sulfonates with sulfamic acid, phenylammonium sulfate being an intermediate:

$$C_6H_5OH + NH_2SO_3H \longrightarrow C_6H_5OSO_2ONH_4$$
$$\longrightarrow HO - C_6H_4 - SO_2 - ONH_4$$

Aldehydes form addition products with salts of sulfamic acid. Aromatics with unsaturated side-chains, e.g., styrene, are sulfonated by sulfamic acid, with retention of the side-chain double bond:

$$C_6H_5CH{=}CH_2 + NH_2SO_3H$$
$$\longrightarrow C_6H_5 - CH = CH - SO_2ONH_4$$

4. Production

Sulfamic acid is now produced solely by the urea process; the plants using ammonia and sulfur trioxide [6] have been closed down.

Following a suggestion by BAUMGARTEN, equimolar quantities of urea, sulfur trioxide, and sulfuric acid are reacted directly to sulfamic acid [7]:

$$NH_2CONH_2 + SO_3 + H_2SO_4 \longrightarrow 2NH_2SO_3H + CO_2$$

This is a strongly exothermic reaction. The process is carried out in two stages, based on the following reactions:

$$NH_2CONH_2 + SO_3 \longrightarrow NH_2CONHSO_3H$$

$$NH_2CONHSO_3H + H_2SO_4 \longrightarrow 2NH_2SO_3H + CO_2$$

Urea is stirred in excess sulfuric acid and sulfur trioxide in the first stage. The temperature should be kept below 40 °C to hinder the formation of carbon dioxide. In the second stage, the reaction product from the first stage reacts to form sulfamic acid and carbon dioxide at 50 – 80 °C in the presence of excess sulfur trioxide. After removal of the excess sulfur trioxide, sulfamic acid of purity > 95 % can be obtained [8]. High-purity sulfamic acid is obtained by recrystallization.

According to a Nissan patent, sulfamic acid is produced by the reaction of urea and oleum ($xH_2SO_4 \cdot ySO_3$). The reaction product is recrystallized wet, and the crystallized acid is separated and dried. By grinding under high pressure, preferably 600 – 1500 kg/m^2, a fine crystalline product with average particle diameter $\leq 500\,\mu m$ is obtained [9].

5. Uses

The 1995 world production capacity was ca. 96 000 metrictons [10]. Most of this is used in cyclamate production (→ Sweeteners). The use of sulfamic acid in cleaning agents for carbonate- and phosphate-containing deposits, e.g., boiler scale, is based on its ability to form readily soluble salts and its relatively low corrosive effect on metals. Sulfamic acid is widely used for cleaning machines and instruments in the paper, sugar, dairy, and brewing industries, and for removing deposits in evaporation plants, heat exchangers, and cooling systems.

In some countries, a process involving the treatment of fatty or ethoxylated alcohols with sulfamic acid to produce raw materials for waxes is used on an industrial scale [11].

6. Toxicology

Sulfamic acid is not specifically toxic, and its physiological properties are typical of a strong mineral acid. Sulfamic acid dust has an irritant

effect on the mucous membranes of the nose and pharynx, and the conjuctiva of the eyes. Oral administration of 1.6 g/kg to rats is lethal.

7. Sulfamates

The solubility of the alkaline-earth sulfamates is important for the utilization of sulfamic acid in industry [12].

Ammonium sulfamate can be obtained by neutralization of sulfamic acid with ammonia. Technical grade ammonium sulfamate was formerly produced by reaction of ammonia with sulfur trioxide under pressure.

Ammonium sulfamate is an effective softener for paper. Impregnated paper is not readily flammable [13]. Ammonium sulfamate has also been used for the removal of lime salts without leaving a residue, which would interfere with the tanning process in leather production.

8. References

1. E. S. Taylor, R. P. Desch, A. J. Catotti, *J. Am. Chem. Soc.* **72** (1950) 74 – 77.
2. E. J. King, G. W. King, *J. Am. Chem. Soc.* **84** (1962) 1212 – 1215.
3. E. Divers, T. Haga, *J. Chem. Soc.* **69** (1896) 1634.
4. J. M. Notley, *Trans. Inst. Met. Finish.* **52** (1974) 78.
5. L. F. Audrieth, M. Sveda, H. Sisler, L. J. Butler, *Chem. Rev.* **26** (1940) 49 – 94.
6. Ullmann, 4th ed., **22,** pp. 311 – 313.
7. P. Baumgarten, *Ber. Dtsch. Chem. Ges. B* **69** (1936) 1929 – 1936.
8. Du Pont, US 2 880 064, 1959.
9. *Chem. Abstr.* **109** (1988) 22,193174.
10. *Kirk-Othmer, 4th ed.,* John Wiley & Sons, **vol. 23,** 1997.
11. General Anilin and Film Corp., US 2 758 977, 1958.
12. G. B. King, I. F. Hooper, *J. Phys. Chem.* **45** (1941) 938.
13. Du Pont, US 2 723 212, 1958. F. T. Blakemore, *Pap. Film Foil Converter* **23** (1949) 29.

Sulfanes → Sulfides, Polysulfides, and Sulfanes
Sulfation → Sulfonic Acids, Aliphatic

Sulfides, Polysulfides, and Sulfanes

Ludwig Lange, Degussa AG, Frankfurt/Main, Federal Republic of Germany

Wolfgang Triebel, Degussa AG, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Physical and Chemical Properties of Compounds of the System Na_2S-S . 1
2.1. Physical Properties 1
2.2. Chemical Properties 5
3. Production of Compounds of the System Na_2S-S 7
3.1. Sodium Tetrasulfide 7
3.2. Sodium Disulfide 7
3.3. Sodium Monosulfide 8
4. Sodium Hydrogensulfide 9
4.1. Properties 9
4.2. Production 9
5. Sulfanes 10
5.1. Properties 10
5.2. Production 11
6. Uses of Sulfides and Sulfanes 12
7. Safety and Environmental Aspects . . 15
8. Economic Aspects 16
9. Transportation 16
10. References 17

1. Introduction

Sulfur – sulfur bonds occur in numerous inorganic, organic, and biologically interesting compounds. One reason for this is the high bond energy of the S – S single bond (265 kJ/mol). This is the third strongest single bond, after H_2 (435 kJ/mol) and the C – C single bond (330 kJ/mol) [1].

Of the compounds of sulfur with metals, only the alkali and alkaline earth sulfides, which contain the S^{2-} ion [2], and the corresponding hydrogensulfide, are soluble in water. The sulfides of the alkali and alkaline earth metals undergo addition of sulfur with conversion to polysulfides, which can also be regarded as salts of the sulfanes [3].

2. Physical and Chemical Properties of Compounds of the System Na_2S-S

2.1. Physical Properties

The Na_2S-S phase diagram has been widely investigated [4–6], [7], [8] with essentially consistent results being obtained [9]. Above 285°C the phase diagram (Fig. 1) contains two liquid regions: a sodium-rich, single-phase region in which the composition moves continuously towards Na_2S_5 as the sulfur content increases, and a two-phase region where two immiscible liquids – Na_2S_5 saturated with sulfur, and sulfur saturated with Na_2S_5 – coexist [5], [10]. In the sodium-rich part of the diagram, the liquidus curve between Na_2S_2 and Na_2S_4 shows an eutectic, which lies at 235°C according [5], and at 240 ± 2°C according to [11]. According to [5], the eutectic corresponds to the composition Na_2S_3; earlier investigations indicated the composition $Na_2S_{3.24}$. X-Ray diffraction experiments have shown that Na_2S_3 does not exist as a defined polysulfide in the solid state [12]. The solid product with the stoichiometric composition Na_2S_3 contains only Na_2S_2 and Na_2S_4 [13].

Although the above results show that solid sodium trisulfide is unstable, the situation is different for the molten state [14]. A melt with the composition Na_2S_3 consists of a mixture of Na_2S_3, Na_2S_2, and Na_2S_4 [14]. In spite of the undoubted presence of the S_3^{2-} ion in K_2S_3, BaS_3, etc., Raman spectroscopy shows that in molten Na_2S_3 the ion disproportionates into S_2^{2-} and S_4^{2-} ions [15]. Thus, the compositions of the known alkali polysulfides are as listed below [8]:

Li_2S	Li_2S_2				
Na_2S	α-, β-Na_2S_2		Na_2S_4	Na_2S_5	
K_2S	K_2S_2	K_2S_3	K_2S_4	K_2S_5	K_2S_6
Rb_2S	Rb_2S_2	Rb_2S_3	Rb_2S_4	Rb_2S_5	Rb_2S_6
Cs_2S	Cs_2S_2	Cs_2S_3	Cs_2S_4	Cs_2S_5	Cs_2S_6

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_443

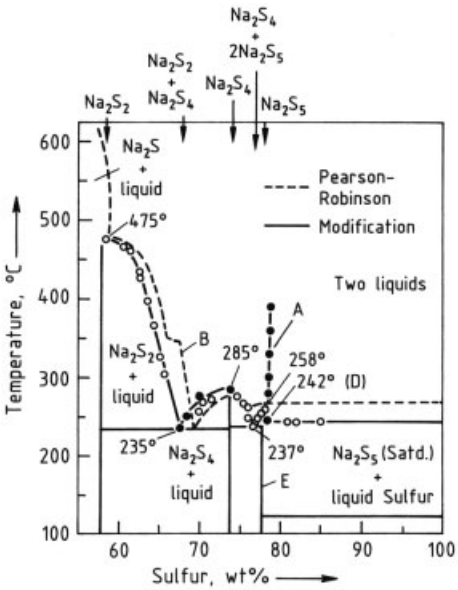

Figure 1. Phase diagram of the Na – S system between the compositions Na_2S_2 and 100 % S [7]

The following route to Na_2S_3 has been proposed on the basis of the differing solubilities of NaCl and KCl in liquid ammonia (NaCl: 3.02 g/100 g solution; KCl: 0.04 g/100 g solution) [8], [11].

$$2\,NaCl + K_2S_3 \xrightarrow{NH_3} Na_2S_3 + 2\,KCl$$

The question of whether Na_2S_3 can be synthesized was partially elucidated when it was shown that the compound can exist as an ammoniate [11]. X-Ray and thermogravimetric measurements at various temperatures on powdered samples showed that ammonia can be removed at 370 K without fundamentally changing the structure [11]. The eutectoid decomposition reaction is hindered, at least kinetically, at low temperature [11].

Experiments on the melting and crystallization processes in the sodium – sulfur cell show remarkable behavior by some sodium polysulfide compounds [16]. The melting temperatures of, e.g., polycrystalline Na_2S_4 and Na_2S_5 are reproducible for the first melting process, but in subsequent melting and crystallization cycles, the two compounds behave like the products of eutectic composition. With the exception of the disulfide, the polysulfides can form supercooled melts, and vitreous states have been observed. Some data on the melting behavior of polysulfides are given in Table 1 [16].

Sodium monosulfide, Na_2S, like Li_2S, K_2S, Rb_2S, Li_2O, Na_2O, and K_2O, crystallizes with the antifluorite structure, in which each S atom is surrounded by a cube of 8 Na atoms, and each Na atom by a tetrahedron of 4 S atoms [15]. The *mp* is 1170 ± 10°C [17], the heat of fusion $\Delta H_{fus} = 30.1$ kJ/mol [19], the standard entropy $S^{\circ}_{298} = 90.3\ J\,K^{-1}\,mol^{-1}$ [17], and the heat of formation $\Delta H_{298} = -386.6$ kJ/mol, based on [18] [19].

Sodium disulfide crystallizes in two modifications [12]. The transformation from the α- to the β-modification takes place in the temperature range 150 – 250°C [12]. The change in structure is irreversible. Products that solidify from the melt or are tempered at high temperature have the β-structure, which is isotypical with that of Li_2O_2 [20]. Sodium disulfide isolated from alcoholic solution exists in the α-modification, isotypical with the structure of Na_2O_2 [20]. The sodium disulfide was prepared in alcoholic solution by the method proposed in [21] in which an alcoholic solution of sodium tetrasulfide is reduced with sodium metal [12].

The specific heat capacity c_p of α-Na_2S_2 was found to be constant at $1071\ J\,kg^{-1}\,K^{-1}$ over the temperature range 97 – 167°C [16].

For β-Na_2S_2 in the range 191 – 267°C, c_p is $1113\ J\,kg^{-1}\,K^{-1}$ [16]. The relative density d_{20} of α-Na_2S_2 is 2.01, and that of β-Na_2S_2, 2.05. The color of Na_2S_2 is yellow to orange.

Sodium Tetrasulfide. The crystal structure of Na_2S_4 is described in [22]. The structure of the S_4^{2-} ions in sodium tetrasulfide is shown in Figure 2 [23].

The interatomic distance *d* of the sulfur atoms is 207 pm, the valency angle α is 109.8°, and the torsion angle τ is 98° [23].

According to [23], reliable IR and Raman spectra of sodium tetrasulfide were first published in [24], the Na_2S_4 spectra available until then having been falsified by the presence of lines due to the pentasulfide. The relative density of sodium tetrasulfide is $d_{20} = 2.08$ [12].

Table 1. Melting behavior of polysulfides

Polysulfide composition	1st melting		Remelting		
	mp, °C	ΔH_{fus}, kJ/mol	Remelt cycles	*mp*, °C	ΔH_{fus}, kJ/mol
Na_2S_2	474	7.48	3	474	7.48
$Na_2S_{3.0}$	234	15.62	3	234	15.62
$Na_2S_{3.3}$			2	241	23.81
$Na_2S_{4.0}$	283	28.85	3	240	25.12
$Na_2S_{4.2}$			3	207 – 234	17.30
			2	240 – 260	26.33
$Na_2S_{4.6}$			3	213 – 231	28.48
$Na_2S_{5.0}$	258	32.55	2	220 – 241	36.67
$Na_2S_{5.2}$	255	30.41	10	219 – 247	36.33
$Na_2S_{5.3}$			11	236 – 246	34.48
$Na_2S_{6.1}$	254	27.89	4	205 – 247	32.55
$Na_2S_{10.0}$	254	29.36	3	205 – 265	28.98

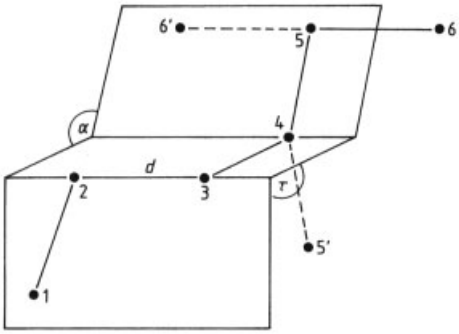

Figure 2. Structural principle of the ions S_4^{2-}, S_5^{2-}, and S_6^{2-} [23]
In S_4^{2-}, atoms 5 and 6 or 5′ and 6′ are absent, and in S_5^{2-}, atom 6 or 6′ is missing. In gg′-S_5^{2-}, the fifth atom is in position 5. Likewise, the last atom of ggg′-S_6^{2-} would be in position 6′.

Sodium Pentasulfide. Like Na_2S_2 and Na_2S_4, Na_2S_5 in the solid state forms a structure that appears uniform by X-ray diffraction, and is therefore a true chemical compound with a definite chain length [12], [25]. The structure, shown in Figure 2 as gg′−S_5^{2-}, is based on the Raman spectrum of Na_2S_5 [23]. The relative density of Na_2S_5 is $d_{20} = 2.08$ [12].

Some physical properties of anhydrous alkali metal sulfides are listed in Table 2 [26].

Molten Sodium Polysulfide. Conductivity data for some molten sodium polysulfides are plotted in Figure 3 [27]. The values are in the range expected for ionic liquids.

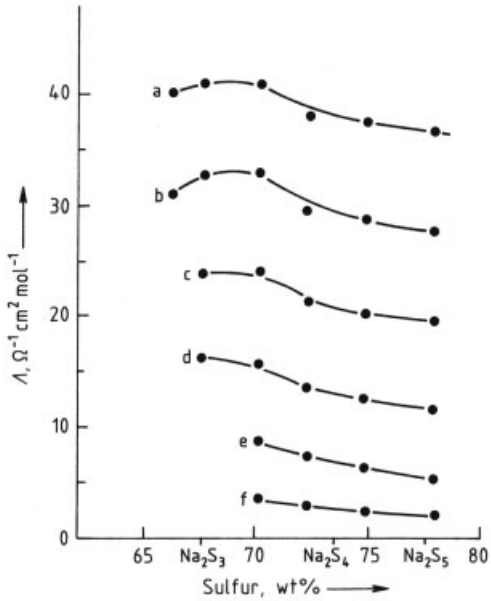

Figure 3. Equivalent conductivity plotted against wt % sulfur for fused sodium polysulfides [27]
a) 420°C; b) 380°C, c) 340°C; d) 300°C; e) 260°C; f) 220°C

At 340°C, the conductivity κ of Na_2S_4 is $0.45\,\Omega^{-1}\,\mathrm{cm}^{-1}$, which compares with

Table 2. Physical properties of the anhydrous alkali metal sulfides

Compound	ΔH_f, kJ/mol	*mp*, °C	Density, g/cm^3	H_{soln}, 0°C, kJ/mol
NaHS	−238.4	350	1.79	16.0
KHS	−265.9	455	1.68 – 1.70	0.4
RbHS	−262.1			0.0
CsHS	−264.2			−1.7
Na_2S	−376.6	1 180	1.856	65.1
K_2S	−420.0	840	1.805	94.5
Rb_2S	−349.4	530	2.912	103.4
Cs_2S	−340.6	510		114.4
Na_2S_2	−443.6	445		
K_2S_2		470	1.973	
Rb_2S_2		450	2.79	
Cs_2S_2		460	3.83	
Na_2S_3	−450.7	275		
K_2S_3		252	2.102	
Rb_2S_3		210	2.68	
Cs_2S_3		217	3.47	
Na_2S_4	−413.3	285	2.08	41.2
K_2S_4	−474.6	145		5.0
Cs_2S_4		160		
Na_2S_5		253	2.08	
K_2S_5		206		
Rb_2S_5		230	2.67	
Cs_2S_5		211		
K_2S_6		196	2.02	
Rb_2S_6		201		
Cs_2S_6		186		

1.11 Ω^{-1} cm^{-1} for $NaNO_3$ and 0.87 Ω^{-1} cm^{-1} for NaSCN at the same temperature.

The conductivity increases with decreasing sulfur content approximately twofold from Na_2S_5 to Na_2S_3. This is partly a result of the increasing ionic concentration as the proportion of sulfur decreases.

The conductivity of pure molten sulfur and of molten sulfur saturated with Na_2S_5 is shown in Figure 4. The two "types of sulfur" show only small differences in their conductivities, in contrast to the conductivity of molten Na_2S_5. This shows that the solubility of Na_2S_5 in sulfur is extremely low [27].

Density values for some sodium polysulfide melts are listed in Table 3 [28].

Comparing the densities of solid and liquid Na_2S_4 and Na_2S_5 shows that the transformation from solid to liquid is accompanied by a volume increase of ca. 8 – 10 % [28].

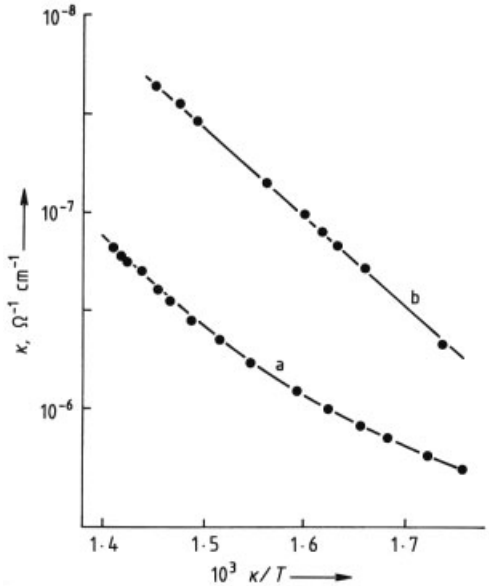

Figure 4. Temperature dependence of conductivity for liquid sulfur (a) and sulfur saturated with Na_2S_5 (b), [27]

Table 3. Values of parameters in the equation $\varrho = D + E\,(T - 600)$, describing the density of fused sodium polysulfides [28]

Melt	S, wt %	Temperature range, K	D, g/cm^3	$10^4\ E$, g cm^{-3} K^{-1}
Na_2S_3	67.6	590 – 683	1.887	−5.65
$Na_2S_{3.3}$	69.7	576 – 689	1.901	−7.96
$Na_2S_{3.7}$	72.0	563 – 669	1.926	−5.47
$Na_2S_{4.4}$	75.4	571 – 680	1.869	−6.66
$Na_2S_{4.8}$	77.0	573 – 683	1.876	−7.16

2.2. Chemical Properties

Sodium Monosulfide. The chemical properties of sodium sulfide are described in detail in [29–31]. Anhydrous Na_2S is a white solid that readily reacts with moist air to form sodium sulfite and carbonate. Pure NaHS can be obtained by the action of H_2S on anhydrous Na_2S.

Sodium sulfide is very soluble in water. The concentration of a saturated solution of Na_2S at various temperatures is given below [32]:

g Na_2S/100 cm^3	12.2	16.0	21.3	28.0	33.0	40
T, °C	0	20	40	60	80	90

The relationship between wt % Na_2S y and density x at various temperatures is given by [31]:

$$y = -98.24 + 98.28 \cdot x \text{ at } 20°C$$
$$y = -98.73 + 99.01 \cdot x \text{ at } 30°C$$
$$y = -98.40 + 99.01 \cdot x \text{ at } 40°C$$

Aqueous solutions [32] of sodium sulfide react alkaline due to hydrolysis:

$$Na_2S + H_2O \longrightarrow NaHS + OH^- + Na^+$$

Iron compounds dissolve in hydrolyzed sodium sulfide solutions, forming thio complexes with an intense brown-red color. Iron metal is also attacked by concentrated Na_2S solutions.

Na_2S solutions can dissolve sulfur up to a composition corresponding to Na_2S_5.

Sodium sulfide solutions are oxidized by NaOCl solutions; depending on the concentration of the reactants, the temperature, and the pH, various amounts of sulfur and sulfate are formed [31]. For dilute solutions of approximately constant concentration, the ratio of sulfur to sulfate formed is constant. With a large excess of sulfide, the formation of sulfur increases, and for a large excess of NaOCl, the formation of sulfate increases. Sulfate formation is favored by increased temperature. Maximum sulfate formation is observed at pH 10 [31].

Sodium sulfide is oxidized by the Karl Fischer reagent, presumably by the reaction [31]

$$H_2S + I_2 \rightleftharpoons 2\,HI + S$$

Sodium sulfide is appreciably soluble in the lower alcohols (methanol, ethanol) [32].

Sodium Polysulfides. The chemical properties of the sodium polysulfides are described in detail in [29], [30], [33], [34].

Sulfur dissolves readily in alkali metal sulfide solutions with formation of alkali metal polysulfides [34]. These solutions are strongly alkaline owing to the reaction

$$S^{2-}{}_x + 2\,HOH \longrightarrow H_2S_x + 2\,OH^-$$

The S_8 ring is broken by the thiophilic S^{2-} and SH^- ions in accordance with the equation

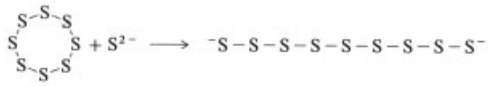

leading to the formation of nonasulfide ions which can break down to form shorter chains. At the same time, longer chains are formed from S^{2-} ions (e.g., S_2^{2-}, S_3^{2-}, etc.) which can also undergo degradation. In solutions of alkali polysulfides produced from alkali sulfide and sulfur, these reactions take place in single steps and, therefore, constitute a complex dynamic equilibrium between sulfide ions of various chain lengths. This explains why aqueous solutions of polysulfide ions of definite chain length cannot be obtained by reaction of stoichiometric amounts of sulfur with sulfide. Addition of such solutions to cold hydrochloric acid always gives a mixture of sulfanes [34]. Furthermore, of the alkali metal polysulfides M_2S_n , only the

pentasulfide dissolves without decomposition, dissociating to form S_5^{2-} ions [25]. Higher polysulfides with $n > 5$ also form pentasulfide ions with precipitation of the excess sulfur, while the lower polysulfides ($n < 5$) disproportionate into monosulfide, tetrasulfide, and pentasulfide ions, which are in pH-dependent equilibrium with each other [25], [35]:

$$4\,S_4^{2-} + H_2O \rightleftharpoons OH^- + SH^- + 3\,S_5^{2-}$$

Free polysulfide ions consist of sulfur chains [35]. The sulfur atoms of an S_3^{2-} chain are necessarily coplanar, but for longer chains there are possibilities for isomerism, as shown in Figure 5. Thus, for the structure of the S_4^{2-} ion there are two possibilities (Fig. 5 A and B), for the S_5^{2-} ion there are the possible structures (C) and (D), and for the S_6^{2-} ion, the structure (E) [35]. The equilibrium between the polysulfide ions in aqueous solution also appears to influence the production of polymeric polysulfides, represented by the general equation

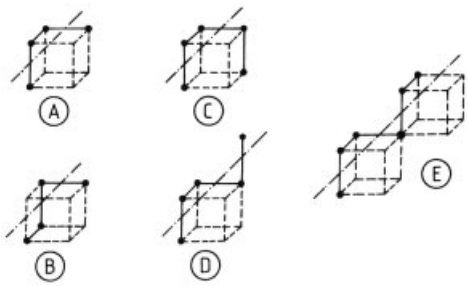

Figure 5. Idealized stereochemistry (dihedral angle equal to 90°) for polysulfide ions [35]
A, B) *d*- and 1-S_4^{2-}; C, D) *cis*- and *trans*-S_5^{2-}; E) *trans, trans*-S_6^{2-}

$$n\,\mathrm{Hal{-}R{-}Hal} + n\,Na_2S_x \longrightarrow (-RS_x-)_n + 2n\,\mathrm{NaHal}$$

This is essentially a nucleophilic exchange between the halogen and the polysulfide anion. The exact nature of the polysulfide ions that react with the dihalides is not yet clear. The polymer initially formed on addition of dihalide to aqueous sodium polysulfide contains less sulfur than expected.

The degree of sulfur substitution (sulfur grade) is the average number of sulfur atoms in each repeating unit of the polymer. However, if a polymer with, for example, a sulfur grade of 4, is treated with 1 mol Na_2S_4, the sulfur grade in the polymer decreases to 3.1, while that of the polysulfide solution increases by a corresponding amount [36].

For reactions that must be carried out under anhydrous conditions, anhydrous polysulfide reagents can be prepared according to the following reaction scheme [37]:

$$2\,LiBHEt_3 + 5/8\,S_8 \longrightarrow Li_2S_5 + 2\,BEt_3 + H_2$$

$$[Cp_2\,MCl_2] + Li_2S_5 \longrightarrow [Cp_2MS_5] + 2\,LiCl$$

(Cp = cyclopentadienyl; M = Ti, Zr, Hf, V)
The so-called Super-Hydride method has been found to be very versatile for the preparation of polysulfide complexes of transition metals [37].

3. Production of Compounds of the System $Na_2S - S$

Sodium tetrasulfide is nowadays produced from the elements in the absence of water and solvents on an industrial scale; sodium disulfide and sodium monosulfide can be prepared similarly on the laboratory scale [38–42].

A summary of the processes is given in Figure 6. The production of Na_2S_4, Na_2S_2, and Na_2S is briefly described below.

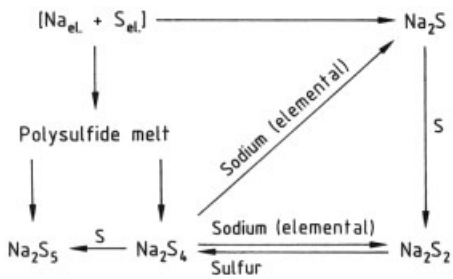

Figure 6. Some production routes to anhydrous, solvent-free sodium monosulfide and polysulfides

3.1. Sodium Tetrasulfide

Sodium tetrasulfide is now produced on an industrial scale in high purity from the elements in the absence of water and solvents.

Molten polysulfide is charged into a cylindrical reactor, with no added water or solvent [38]. Molten sodium and sulfur are added alternately and are mixed with the melt by vigorous agitation. The heat of reaction is removed by a heat-transfer oil in contact with the outer wall of the reactor.

For optimum process operation, the temperature of the polysulfide melt should be in the range 340 – 360°C.

As the molten polysulfide is extremely corrosive, the choice of construction materials is very important. The alloy AlMg appears to be sufficiently resistant. The atmosphere in the reactor is inerted with argon [38].

The preparation of the sodium polysulfides Na_2S_3, Na_2S_4, and Na_2S_5 on the laboratory scale by synthesis from the elements is described in [13]. An equimolar amount of sulfur is added to molten sodium in small amounts with agitation. The reaction product, consisting mainly of Na_2S_2, is cooled, size reduced, and then mixed at room temperature with the required amount of sulfur. The mixture is then melted. Unreacted sodium recovered from the size-reduced reaction product is then added to this polysulfide melt [13].

In another laboratory method, based on early work [5], [21], Na_2S_4 was prepared in alcoholic solution from sodium, hydrogen sulfide, and sulfur [12]. In a widely cited preparation of sodium tetrasulfide or polysulfides from Na_2S and sulfur, alkali metal sulfide, prepared by the reduction of Na_2SO_4, is added to molten sulfur [43].

3.2. Sodium Disulfide

Sodium disulfide, Na_2S_2, is produced by adding a higher sodium polysulfide and sodium, either simultaneously or alternately, to a polysulfide melt [39]. As the melting point of Na_2S_2 (474°C) is higher than the boiling point of sulfur, it is preferable to use sodium tetrasulfide. Thus, sodium disulfide free of water or solvent can be produced in an easily controlled process. The process has three variations [39]. The whole of the polysulfide can be premelted, and the sodium added to the melt until the desired stoichiometric composition is obtained, or sodium and, for example, Na_2S_4 can be added to the melt either alternately or simultaneously [39].

There are many literature references to other methods of producing sodium disulfide. For example, sodium tetrasulfide can be dissolved in alcohol and then reduced with sodium metal to give sparingly soluble Na_2S_2 [21]. It has not been possible to prove conclusively that the product so obtained, which contains 4 – 6 % alcohol, represents a defined alcoholate [12]. High-purity Na_2S_2 can be produced from the elements in liquid ammonia [12].

Small amounts of sodium polysulfide for the investigation of thermodynamic properties have been prepared by melting together Na_2S and sulfur in a small sealed capsule [16].

Sodium disulfide is produced on an industrial scale from sodium hydroxide solution and sulfur [44], and is used for the production of polysulfide polymers [36] (see Chapter 6).

3.3. Sodium Monosulfide

The most important processes for the production of Na_2S are described in detail in [29–31], [45]. Of these, the reduction of sodium sulfate with carbon is described in detail here.

Sodium sulfide is produced in large quantities by the reaction of anhydrous sodium sulfate with carbon in a process that has been used for decades. The reaction, which is carried out at 700 – 900°C, is represented by the following overall equation [32]:

$$Na_2SO_4 + 2\,C \longrightarrow Na_2S + 2\,CO_2 \quad \Delta H = +203.7\ \text{kJ/mol}$$

At ca. 730°C, the evolution of CO_2 begins, and the reaction mass becomes very fluid. On further heating, it becomes pasty, and finally solid. This behavior can be explained by means of the sodium sulfate – sodium sulfide phase diagram (Fig. 7) [32].

The process details and the operation of the reactors or furnaces are described in patents [29]. The rotating-hearth furnace (Fig. 8) operates fully continuous. The sulfate – carbon mixture is delivered by a water-cooled screw feeder (a), onto the outer part of the rotating hearth (b) [32]. As the hearth rotates, a second screw feeder (c) moves the mixture nearer to the center of the plate where the melt passes through the discharge shaft (d).

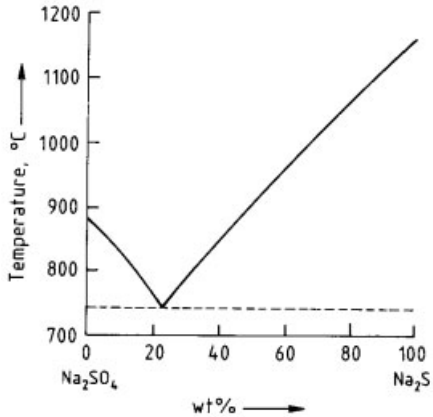

Figure 7. Phase diagram of the system $Na_2SO_4 - Na_2S$ [30]

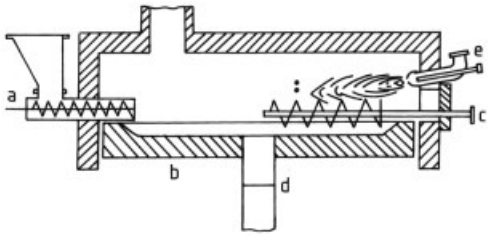

Figure 8. Rotating-hearth furnace [32]
a) Screw feeder; b) Rotating hearth; c) Transporting screw feeder; d) Discharge shaft; e) Burner

To ensure sufficient heat transfer to the reaction mixture, the temperature in the gas space of the furnace must be ca. 1100 °C. The furnace lining is strongly attacked by Na_2S.

The sodium monosulfide so produced contains excess carbon, ash components, Na_2CO_3, $Na_2S_2O_3$, and Na_2SO_3. The product is leached with water, and the undissolved components are removed by filtration. The solution is concentrated by evaporation to 60 – 62 % Na_2S. The concentrated solution, which solidifies at ca. 90°C, is converted to the solid form by cooling on rolls or in pans [32]. Depending on process conditions, the purity of the sodium monosulfide obtained can vary between 50 and 98 wt % [29]. The product can be purified in various ways, e.g., by removal of the strongly colored iron compounds from the sodium sulfide solution [46].

High-purity sodium monosulfide can be obtained by treating the reaction product with methanol to dissolve the sulfide, instead of water [47]. Unlike its hydrates, anhydrous sodium sulfide is very soluble in alcohols. The solubility of Na_2S in methanol is 160 g/L at 20 °C, whereas that of $Na_2S \cdot 9\,H_2O$ in methanol is only 30 g/L [47]. Other aliphatic, aromatic, and polyhydroxy alcohols can also be used as extractants: the solubility of Na_2S in ethanol is 90 g/L, and in ethylene glycol >200 g/L. The sodium sulfide obtained by evaporating the extractant is free of water, iron, sulfite, sulfate, and thiosulfate [47].

Sodium sulfide can also be obtained by reducing Na_2SO_4 with hydrogen (Sulfigran process) [32], methane, or other hydrocarbons [29], [32], [45].

The reduction of sodium polysulfides with sodium amalgam to give high-purity monosulfide proceeds according to the following equation [32]:

$$Na_2S + 3\,S \longrightarrow Na_2S_4$$

$$Na_2S_4 + 6\,NaHg_x \longrightarrow 4\,Na_2S + 6\,Hg_x$$

Anhydrous sodium monosulfide is obtained from moist, anhydrous sodium sulfide in a two-stage process [48].

Pure Na_2S can be prepared on a laboratory scale from the elements using liquid ammonia as solvent [12] or without the use of solvents [40–42]. Na_2S has been prepared in the laboratory in two steps:

In the first step, one of three processes was carried out in a preheated kneading machine:

1) Solid Na_2S_4 was charged into the machine and ground. Liquid sodium was then slowly added, maintaining the temperature of the reaction mixture at 120 – 250 °C by means of a heat-transfer oil [40]; or
2) Liquid sodium was charged into the machine, and solid sulfur was slowly added, maintaining the temperature of the reaction mixture at 120 – 250 °C by means of a heat-transfer oil [41]; or
3) Sulfur and sodium were added alternately to powdered sodium monosulfide, prepared by method 1 or 2, maintaining the temperature of the reaction mixture at 120 – 250 °C by means of a heat-transfer oil [42].

All of these first stages gave a blue-black powder as intermediate product.

In the second stage, the intermediate product was placed in a small reactor. The unreacted

residues of sodium and sulfur in the black intermediate product were then allowed to react at 250 – 480 °C. A white sodium monosulfide was produced from all three products of the first stage. The purity of the Na_2S so obtained was consistently >95 wt % [40–42].

4. Sodium Hydrogensulfide [32]

Sodium hydrogensulfide, NaHS, M_r 56.07, *mp* 350 °C, *d* 1.79, is a colorless to yellow solid. The anhydrous salt (95 – 98 % NaHS) is of minor importance. In industry, the following low-iron hydrated products are used [45]:

1) Sodium hydrogensulfide solution (30 % NaHS)
2) Sodium hydrogensulfide flakes (70 % NaHS)
3) Sodium hydrogensulfide powder (95 % NaHS)

4.1. Properties [32]

Sodium hydrogensulfide is very soluble in water [31]. The variation of solubility with temperature is shown below:

g NaHS/100 cm^3	41	50	55	60	65	72
T, °C	5	22	35	39	43	65

Sodium hydrogensulfide forms two hydrates, $NaHS \cdot 2\,H_2O$ and $NaHS \cdot 3\,H_2O$. The aqueous solution has an alkaline reaction.

The solid products are strongly hygroscopic, and are oxidized by atmospheric oxygen.

4.2. Production [32]

Sodium hydrogensulfide is usually obtained by the action of hydrogen sulfide on sodium sulfide or sodium hydroxide solution. If hydrogen sulfide is passed over anhydrous high-percentage sodium sulfide at ca. 300 °C, high-percentage sodium hydrogensulfide is obtained [49]. According to a patent, anhydrous sodium hydrogensulfide is obtained by the reaction of sodium metal with hydrogen sulfide in inert organic solvents [50], and high-percentage solid products can be obtained by the evaporation of aqueous solutions [51].

Hydrated products are obtained by reacting hydrogensulfide with solutions of sodium sulfide or sodium hydroxide (e.g., in spray towers). If sodium hydroxide solution is used, it should be noted that the sodium sulfide formed as an intermediate is less soluble than the end product. Problems arising from this can be avoided by a patented method [52]. Another patent [53] describes how the difference between the solubilities of sodium sulfide and sodium hydrogensulfide can be exploited by allowing hydrogen sulfide to react with sodium sulfide (60 % Na_2S) in lump form in a tower, controlling the temperature so that the sodium hydrogensulfide runs off, leaving behind the sodium sulfide as a solid.

The hot solutions obtained using pure hydrogen sulfide can be converted to solid products by cooling without intermediate treatment. If the hydrogen sulfide contains carbon dioxide, Na_2CO_3 is formed in the solutions in addition to NaHS. A process for obtaining pure products from such solutions is described in a patent [54]. Another patent describes how NaHS solutions can be produced from H_2S-containing gaseous byproducts of petroleum refining [55].

5. Sulfanes

Hydrogen polysulfides, H_2S_n, were named sulfanes by Fehér and Laue in analogy to the alkanes [56].

Present-day knowledge of this extensive class of compounds is mainly due to the work of F. Fehér et al. in the 1950's [57].

5.1. Properties

The sulfanes, H_2S_n, formally constitute the simplest class of polymers of sulfur with other atoms [33].

These unbranched sulfur chains, terminated by hydrogen atoms, are the starting compounds not only for the metallic polysulfides, but also for many covalent compounds containing sulfur – sulfur bonds, including commercial products of the Thiokol type, sulfane sulfonic acids, and sulfur-vulcanized rubber. The various polymers that contain sulfur chains have in

common that they react as electrophiles towards thiophiles [33], [58].

The sulfanes H_2S_n with $n = 2-8$ have been prepared and isolated in pure form. Accurate determination of chain length distribution in sulfane mixtures and detection of higher sulfanes (up to H_2S_{35}) first became possible with the use of NMR spectroscopy on sulfanes dissolved in benzene or benzene/CS_2 [56]. All sulfanes consist of unbranched chains of sulfur atoms, reflecting the well-known chain-forming tendency of the element. The sulfanes are reactive liquids whose densities, viscosities, and boiling points increase with increasing chain length. H_2S_2, which is an analog of H_2O_2, is colorless. The higher homologues have a yellow color which becomes more intense with increasing chain length [15].

Physical Properties [32]. Some physical and thermodynamic data are given in Table 4. More detailed information can be found in [59]. Extensive spectroscopic experiments and calculations to elucidate the structures of disulfane and trisulfane can be found in [60–62].

Chemical Properties. The sulfanes are soluble in hydrogen sulfide, benzene, and chloroform [30]. All sulfanes are vigorously decomposed by solutions of alkalis.

The sulfanes are readily oxidized. They are thermodynamically unstable, i.e., the thermodynamic equilibrium [32]

$$xH_2S_n \rightleftharpoons xH_2S + 1/8[x(n-1)]S_8$$

lies mainly to the right-hand side. Also, traces of alkali that dissolve from the surface of glass vessels can cause decomposition (i.e., sulfur precipitation, in accordance with the above equation) [15], [34].

Sulfanes dissolve rapidly and without decomposition at 120°C in molten sulfur [32]. At higher temperature, these solutions decompose, liberating H_2S. If NH_3 in the ppm range is added as a catalyst, this decomposition to H_2S and S becomes explosive. This is utilized in the Claus process for producing sulfur, to prevent a slow outgassing of the H_2S.

The decomposition to H_2S and S takes place even on slight heating. It proceeds more rapidly as the temperature increases (up to 100 °C) and is strongly accelerated by light, especially sunlight and UV light [32].

The reaction of sulfanes with reactive alkenes is described in [63]. The addition of a mixture of sulfanes H_2S_n to 2,3-dimethyl-1,3-butadiene is an exothermic reaction which proceeds readily at room temperature in nonpolar solvents:

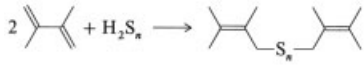

If the unsaturated compound contains two reactive double bonds, sulfane addition can be used to prepare sulfur-rich polymers [63].

Sulfanes react quantitatively with excess sulfite to form hydrogen sulfide and thiosulfate [15], [33]:

$$H_2S_x + (x-1)SO_3^{2-} \longrightarrow H_2S + (x-1)S_2O_3^{2-}$$

This reaction provides a useful method of quantitative analysis by determining the H_2S formed by precipitation as CdS, and the thiosulfate by iodometry.

The reaction with CN^- ions proceeds analogously:

$$H_2S_n + (n-1)CN^- \longrightarrow H_2S + (n-1)\,SCN^-$$

The thiocyanate formed can be determined colorimetrically as $Fe(SCN)_3$ [15], [33].

1,3-Diketones in organic solvents saturated with HCl react with sulfanes to form 1,2-dithiolium ions [34], [64].

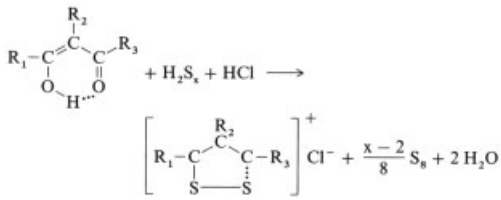

Sulfanes have been reacted with chlorosulfanes to obtain novel sulfur rings [65], [66]. Some examples are given below [66]:

$$H_2S_6 + S_4Cl_2 \longrightarrow S_{10} + 2\,HCl$$

$$H_2S_4 + S_2Cl_2 \longrightarrow S_6(50\,\%) + S_{12}(3\,\%) + 2\,HCl$$

$$H_2S_8 + S_4Cl_2 \longrightarrow S_{12}(18\,\%) + 2\,HCl$$

$$H_2S_8 + S_{10}Cl_2 \longrightarrow S_{18} + 2\,HCl$$

$$H_2S_{10} + S_{10}Cl_2 \longrightarrow S_{20} + 2\,HCl$$

Table 4. Physical and thermodynamic data for sulfanes

Sulfane	d_{20}	n_D^{20}	p^{25}, mbar	S^{25}, J/K	$C^{25}{}_{p,l}$, J mol^{-1} K^{-1}	$C^{25}{}_{p,s}$, J mol^{-1} K^{-1}
H_2S_2	1.334	1.631	149.226	165	92.528	51.078
H_2S_3	1.499	1.729	2.5935	216	123.510	74.525
H_2S_4	1.582	1.791	0.06783	264	154.074	97.971
H_2S_5	1.644	1.836	0.002527	351	184.638	120.998
H_2S_6	1.688	1.867				
H_2S_7	1.721	1.893				
H_2S_8	1.747	1.912				

5.2. Production

If aqueous solutions of alkali metal polysulfides are added to excess concentrated hydrochloric acid with cooling, and local alkalinity due to the hydrolysis of Na_2S_n is avoided, a yellow "crude oil" ($H_2S_4, H_2S_5, H_2S_6, H_2S_7, H_2S_8$) separates. This can be separated into its pure components by high-vacuum distillation or selective extraction. The lower sulfanes H_2S_2 and H_2S_3 can be obtained by distillation with cracking [34], [67], [68]. As even trace amounts of alkali on the surface of glass catalyze the decomposition of sulfanes to H_2S and S, great care is necessary in this process. Glass apparatus intended for use with sulfanes must be pretreated with gaseous HCl [59]. Because of the low thermal stability of the higher sulfanes, these compounds can only be handled at low temperature [69].

The crude oil, containing a mixture of sulfanes, can be produced from various starting materials [59]:

1) From aqueous polysulfide solutions
2) From anhydrous polysulfides
3) From hydrogen sulfide and sulfur
4) By the reaction of lower sulfanes with chlorosulfanes
5) From aqueous solutions of thiosulfate and dithionite
6) By the cathodic reduction of saturated aqueous solutions of SO_2

The third reaction indicates that liquid sulfur produced by the Claus process can contain large quantities of sulfanes [69]. Their formation in the Claus combustion chamber is essentially determined by the equilibrium between sulfur vapor, hydrogen sulfide, and sulfanes.

6. Uses of Sulfides and Sulfanes

Much of the new knowledge of the system Na_2S-S mentioned above has come from research into sodium – sulfur cells. However, the Na_2S_n compounds that are formed and decomposed in these cells are initially produced from the elements sodium and sulfur, so that the batteries that contain these cells have no economic implications for the production and use of anhydrous, solvent-free sodium polysulfide compounds [10], [70].

An established application is in the production of *sulfur dyes* [71], which are produced by the two methods (→ Sulfur Dyes):

1) The polysulfide bake method, where sulfur or sulfur – sodium sulfide is heated in the dry state with aromatic amino or nitro compounds or with nitrogen-free organic substances such as cresol, anthracene, or decacyclene
2) The polysulfide melt method, in which organic compounds are refluxed with polysulfide in an aqueous or alcoholic medium

The water-insoluble sulfur dyes are dissolved by reduction to the leuco form, generally with sodium sulfide. The leuco dyes become attached to the fibers, and the sulfur dye is then fixed on the fiber by oxidation with peroxides or perborates [71].

Anhydrous, solvent-free sodium tetrasulfide is extensively used in the production of *organosilanes*, which are used in the rubber industry:

$$2(RO)_3Si-(CH_2)_3-Cl + Na_2S_4 \longrightarrow (RO)_3Si-(CH_2)_3-S_4-(CH_2)_3-Si(OR)_3 + 2\,NaCl$$

The ethyl derivative of bis(trialkoxysilylpropyl) tetrasulfide (Si 69) is one of the most important organofunctionalized silanes [72]. Organosilanes are used in the rubber industry in combination with various fillers to produce elastomers with improved static and dynamic vulcanizate properties [73]. These silanes have two different reactive groups that react with the silanol groups of fillers during compounding in rollers or kneaders, and form chemical bonds to the polymer and cross-linking agents during vulcanization. Si 69 can also be used as a sulfur donor that gives reversion-free vulcanization (equilibrium cure system) [73].

Another example of the use of an anhydrous sulfide is in a process for producing *poly(phenylene sulfide)* [33], [74], [75] (see also → Polymers, High-Temperature, Chap. 5.). In the production of poly(phenylene sulfide), 1,4-dichlorobenzene and sodium sulfide are reacted at ca. 280°C under pressure in a polar aprotic solvent, usually *N*-methyl-2-pyrrolidone [76]. While aliphatic polysulfides can easily be prepared by nucleophilic exchange reactions of chloride with sulfide, the reaction with aromatic compounds is much more difficult because of the considerably lower reactivity of the aromatically bonded chlorine [33]. It has been suggested that formation of poly(phenylene sulfide) does not proceed like a normal polycondensation [77]. The reaction is a one-electron transfer process in which radical cations act as reactive intermediates [77]. Poly(phenylene sulfide) can be used with very high contents (up to 70 %) of fillers such as glass fiber, carbon fiber, or minerals. Its good chemical resistance (PPS is practically insoluble in all the usual solvents) and inherent flame-retardant properties enable it to be used under severe thermal and chemical conditions. Reinforced PPS can replace metals in many applications, e.g., in the construction of machinery and equipment, pump housings, and gear wheels [76].

Sodium sulfide, sodium hydrogensulfide, and sodium polysulfide are sometimes used in the same fields of application [31]. This is also true to some extent of the sulfanes, although their instability prevents their large-scale use [32]. Examples are the manufacture of polysulfide polymers and the sulfide reduction of polynitroaromatics, both described below.

Aliphatic polysulfides of the thiokol type have the general formula $(C_xH_{2x}S_y)_n$. These find wide application [74] as sealants (→ Sealing Materials, Chap. 3.) and, in combination with epoxy resins, as adhesives and mastics. These products have a marked ability to expand and contract, are elastic over a wide temperature range, and are resistant to seawater, solvents, and many chemicals. The best known commercial products are the Thiokols, which are polyalkyl- and polyalkoxypolysulfides. They are produced by the condensation of Na_2S_n with dichloroalkyls and bischloroalkylformals [79], [80]. Trichloroalkanes [81], epichlorohydrin [82], condensation products of α-monochlorohydrin formaldehyde [83], and monochloro alcohols [80] are also used as modifiers in the condensation reaction. The basic properties of these polysulfides mainly depend on the sulfur content [36]. The Na_2S_n compounds can be obtained by the reaction of NaOH with sulfur [44]. At temperatures of <110°C and with NaOH concentrations of <30 wt %, polysulfide and thiosulfate are formed according to the equation

$$10\,S + 6\,NaOH \longrightarrow 2\,Na_2S_4 + Na_2S_2O_3 + 3\,H_2O$$

whereas at higher NaOH concentrations (up to 35 wt %) and higher temperatures (100 – 320°C) sulfite is also formed [44]:

$$6\,NaOH + (2x+1)S \longrightarrow 2\,Na_2S_x + Na_2SO_3 + 3\,H_2O$$

The autoxidation of aqueous polysulfide solutions in the temperature range 23 – 40°C, the kinetics of polysulfide – thiosulfate disproportionation [84], [85], and equilibria of polysulfide ions in solution [86] should be mentioned in this connection [87]. The reaction usually proceeds very rapidly from sulfite to thiosulfate, so that only small amounts of sulfite remain. However, as sulfite crystallizes from concentrated NaOH solution, small amounts of thiosulfate remain in the solution [88].

The polymers produced by the reaction of alkyl chlorides with sodium polysulfide are obtained as high molecular mass, rubber-like products with a very limited range of uses. The polymers can be selectively cleaved by the action of sodium hydrogen sulfide on the disulfide groups, allowing the average chain length to be controlled [89].

$$-R-S-S-R + NaSH \longrightarrow R-SNa + H-S-R + S$$

The sulfur liberated in this reaction reacts with added sodium sulfite to form sodium thiosulfate [89]:

$$Na_2SO_3 + S \longrightarrow Na_2S_2O_3$$

The bis(vinylarylalkyl) polysulfides [90], obtained by reacting 4-vinylbenzyl chloride with Na_2S-S, are used as coatings and sealants in the form of copolymers with halogenated alkanes.

Aromatic amino compounds can be prepared by the reduction of polynitroaromatics with sulfide [91]. Sulfide reductions are carried out in industry with Na_2S, NaHS, Na_2S_2, or ammonium sulfide solutions. The reactions are not exactly stoichiometric, proceeding approximately in accordance with the following equations:

$$4\,Ar-NO_2 + 6\,Na_2S + 7\,H_2O \longrightarrow 4\,Ar-NH_2 + 3\,Na_2S_2O_3 + 6\,NaOH$$

$$4\,Ar-NO_2 + 6\,NaSH + H_2O \longrightarrow 4\,Ar-NH_2 + 3\,Na_2S_2O_3$$

$$Ar-NO_2 + Na_2S_2 + H_2O \longrightarrow Ar-NH_2 + Na_2S_2O_3$$

Some other important products obtained by using the sulfide reaction are 1,2-diamino-4-nitrobenzene from 2,4-dinitroaniline and NaSH; 2-amino-4-nitrophenol from 2,4-dinitrophenol and NaSH; and 4,6-dinitro-2-aminophenol (picraminic acid) from picrinic acid and NaSH [91].

To reduce friction and increase load-bearing capacity, EP (extreme-pressure) and AW (antiwear) *oil additives* are added to engine oils, transmission oils, turbine oils, and machining oils and their emulsions [92], [93]. It is known that sulfur-containing compounds in which the sulfur is in "available" form perform especially well in this application. Especially suitable are those compounds with polysulfide bridges in which the sulfur atoms are in a labile, active form, are liberated at elevated temperature, and can form metal sulfide coatings on the metal surface. These coatings usually have a lower coefficient of friction than the metal itself and thus improve frictional behavior [92], [93].

These sulfur-containing compounds can be prepared by reacting epoxides with an alkali metal disulfide or tetrasulfide [92]:

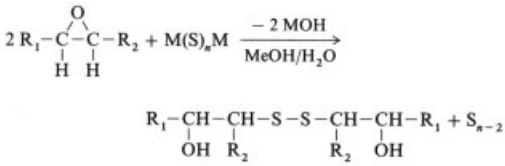

In accordance with the above reaction scheme, elemental sulfur is liberated when Na_2S_4 is used [92].

Oil additives can also be produced by the reaction of Na_2S or Na_2S_n, often in combination with elemental sulfur, with chlorinated organic compounds. These compounds can be formed by the addition of an S – Cl compound (e.g., S_2Cl_2) to a double bond (e.g., isobutene), and then condensed with Na_2S-S [94–99]. Other products may be obtained from the reaction of aromatic hydroxy compounds (e.g., 2,6-di-*tert*-butylphenol or 2,4,6-tri-*tert*-butylphenol) with aldehydes or ketones and Na_2S [100].

The reaction of 4-chlorotetrahydropyran with Na_2S in a methanol/ethanol mixture, followed by hydrogen peroxide oxidation of the resulting thiol group leads to bis(tetrahydropyranyl)disulfide, which is used as a lubricant additive [101].

Dialkylbenzyl disulfides, which are used as oil additives, are produced from a catalytically cracked naphtha fraction (boiling range 164 – 250 °C), which is chloromethylated with formaldehyde – HCl, and then reacted with Na_2S_2 [102].

In a similar process, benzyl chloride and Na_2S_2 are reacted at 80 – 90 °C for 20 h to produce dibenzyl disulfide, which is used as a component of an oil additive [103].

The production of additives for high-pressure lubricants is described in [104]. A mixture of xylol isomers is chloromethylated and then reacted with Na_2S_3 (a mixture of Na_2S_2 and Na_2S_4) to form a polysulfide.

A group of microcrystalline Al – Na silicates with various sulfur contents have been referred to as ultramarines [105]. Analysis gives an approximate empirical formula of $(Na_8Al_6Si_6O_{24}S_4)_n$, suggesting the name sodium aluminum silicate polysulfide.

The production of an *ultramarine* and a green pigment from zeolites and Na_2S_5 by heating to

>300 °C has been reported [106]. The color of the ultramarines may be due to singly charged polysulfide(−1) radical ions [107], [108].

Sodium monosulfide solutions mixed with $FeSO_4$, $Ca(OH)_2$, and a polymeric flocculating agent are used to remove copper from aqueous solutions, process liquors, and wastewater [109]. Lead can be removed by using a mixture of caustic soda solution, sodium monosulfide solution, and aluminum chloride [110].

7. Safety and Environmental Aspects [111]

Sodium sulfide and Na_2S_n are classified as irritants in GefStoffV [112], as indicated by the hazard symbol C.

The following R and S phrases are applicable:

R31: "Toxic gases produced on contact with acids"
R34: "Causes irritation"
S26: "On contact with eyes, rinse immediately with plenty of water, and seek medical help"
S45: "In case of accident or illness, seek medical help immediately. Show this label if possible."

In accordance with GefStoffV vessels containing Na_2S or Na_2S_n should carry the name of the substance in permanent printed letters, the hazard symbol "irritant", the hazard information R 31 – 34, and the safety advice S 26 – 45 [113], [114].

Aqueous solutions of Na_2S, NaHS, and sodium polysulfides must always be maintained at a pH of >10 to prevent liberation of toxic hydrogen sulfide [3]. The hazardous properties of hydrogen sulfide and its effects on humans are described in [32] (see also → Hydrogen Sulfide, Chap. 10.).

Solid sodium sulfide, sodium hydrogensulfide, and sodium polysulfide form strongly alkaline solutions when dissolved in water, and these can cause severe irritation of the eyes and skin. Goggles must therefore be worn. When working with larger quantities, impermeable protective clothing consisting of rubber boots (not leather-soled shoes) and long rubber gauntlets must also be worn.

If dust is formed, respirators with combination B/P2 type filters (color code: gray with white ring) must be worn [113].

In case of fire, respiratory equipment with an independent air supply must be used, completely sealed chemically-resistant clothing must be worn while fighting the fire, and the auxiliary equipment used must be resistant to caustic solutions [113]. Note that sulfide solutions can emit hydrogen sulfide on heating, which can represent a potential hazard during storage or in production [31].

Industrial safety data for sodium sulfide, sodium polysulfide, and sodium hydrogensulfide are given in [111], [115].

Sulfide is classified as a hazardous substance according to § 7 a of the German Water Management Act [116]. In accordance with the current state of the art of the process in which byproduct sulfide is formed, the sulfide content of wastewater has to be reduced. This can be carried out in various ways, e.g., by oxidation of the alkaline solution with H_2O_2 [117]:

$$Na_2S + 4\,H_2O_2 \longrightarrow Na_2SO_4 + 4\,H_2O$$

The limit values for the sulfide content of wastewater are taken from the Framework Wastewater Directive or a recommendation of the German wastewater technical association (ATV) for indirect dischargers. Some values (in mg/L) are listed below [118]:

Appendices to Framework Wastewater Directive (1992)	
Appendix 25 (leather manufacture, fleece treatment, leather fiber production): wastewater from soaking, liming, and deliming, including washing	2
Appendix 37 (manufacture of inorganic pigments): Source type 1.1.3 (lithopone, zinc sulfide pigment, and precipitated barium sulfate)	1
Appendix 39 (manufacture of nonferrous metals)	1
Appendix 40 (metal working and treatment)	
Source type 1.1.1 (electrolytic)	1
Source type 1.1.2 (pickling)	1
Source type 1.1.4 (burnishing)	1
Source type 1.1.7 (printed circuit manufacture)	1
Source type 1.1.8 (battery manufacture)	1
Source type 1.1.9 (enamelling)	1
Appendix 47 (scrubbing of furnace combustion gases)	0.2
34th Framework Wastewater Directive (manufacture of barium compounds)	1
38th Framework Wastewater Directive (textile manufacture)	0.1
43rd Framework Wastewater Directive (synthetic fibers): Source type 2.1.3 (manufacture of viscose spun fiber)	15
Wastewater technical association (ATV) Information sheet A 115 (yellow printing)	2

8. Economic Aspects [119]

Sodium monosulfide is supplied commercially in the form of flakes containing 60 – 62 % Na_2S.

Major producers of the monosulfide include Tessenderlo Chemie (Belgium) and Foret (Spain). There are also producers in Germany, Italy, and France.

The output of sodium monosulfide in Western Europe is estimated to be 50 000 t/a in the mid 1990s. Excluding the considerable quantities exported to the African copper mines, this consumption is distributed among the following areas of use:

Dehairing agents for the leather industry	40 %
Precipitation of heavy metal ions (chemical industry)	30 %
Production of sulfur dyes	20 %
Other uses	10 %

Sodium hydrogensulfide is supplied in the following forms:

1) In powder form with a purity of 95 %
2) In flake form with a purity of 70 – 74 %
3) As a 30 % solution

Sodium hydrogensulfide is manufactured by the sodium monosulfide producers mentioned above, and also by Goldschmidt, Rhodanid Chemie, and Leuna. The main producers in the United States are Jupiter Chemical, PPG Industries, and Ethyl Corp.

As sodium hydrogensulfide is used in the same areas as Na_2S, no reliable information is available concerning the production and consumption of NaHS in Western Europe. In the United States, about two-thirds of annual production (i.e., ca. 180 000 t/a) is of lower quality (low grade, 20 – 30 % NaHS) and is used in the cotton industry.

For many years, sodium polysulfide has been available only in solution. Sodium polysulfide solution is supplied by Goldschmidt (Germany), Witton (United Kingdom), PPG (United States), and Sankyo Kasei (Japan).

Approximately 25 000 t/a polysulfide solution, produced mainly by Morton International (United States), Toray Thiokol (Japan), and Akcros (Greiz-Dölau, Germany), is used in the production of polysulfide elastomers.

Sodium polysulfides are also available in anhydrous and solvent-free form, but the price of this material is considerably higher than that of the aqueous solution.

9. Transportation [120]

Sodium sulfide with at least 30 % water of crystallization is classified as follows:

RID/ADR	Class 8, no. 45 b
IMDG Code	Class 8, UN no. 1849, PG II
ICAO – TI/IATA – DGR	Class 8, UN no. 1849, PG II

Approved packaging for sodium sulfide has been harmonized for all means of transport. Intermediate bulk containers are permitted for land and sea transport. Portable tanks (rail tank cars, road tank cars, tank containers) are permitted for land and sea transport.

Sodium sulfide with less than 30 % water of crystallization is classified as follows:

RID/ADR	Class 4.2, no. 13 b
IMDG Code	Class 4.2, UN no. 1385, PG II
ICAO – TI/IATA – DGR	Class 4.2, UN no. 1385, PG II

Approved packaging has been harmonized for all means of transport, and must be hermetically sealed. Bags are not permitted for sea transport. Bags may be used for land transport provided they are loaded on palettes and forwarded as full load. Intermediate bulk containers are only permitted for land and sea transport. Portable tanks (rail tank cars, road tank cars, tank containers) are permitted only for land transport.

Sodium hydrogensulfide with at least 25 % water of crystallization is classified as follows:

RID/ADR	Class 8, no. 45 b
IMDG Code	Class 8, UN no. 2949, PG II
ICA – TI/IATA – DGR	Class 8, UN no. 2949, PG II

Packaging has been harmonized for all means of transport. Intermediate bulk containers and portable tanks are permitted for land and sea transport.

Sodium hydrogensulfide with less than 25 % water of crystallization is classified as follows:

RID/ADR	Class 4.2, no. 13 b
IMDG	Class 4.2, UN no. 2318, PG II
ICAO – TI/IATA – DGR	Class 4.2, UN no. 2318, PG II

Packaging has been harmonized for all means of transport, and must be hermetically sealed. Bags are not permitted for sea transport, but can be used for land transport provided they are loaded on palettes and forwarded as full load. Intermediate bulk containers are permitted for land and sea, and portable tanks only for land transport.

Sodium polysulfides are not specifically named in the transport regulations, but should be classified as follows:

RID/ADR	Class 8, no. 45 b
IMDG Code	Class 8, UN no. 1759, PG II (solid) Class 8, UN no. 1760, PG II
ICAO – TI/IATA – DGR	Class 8, UN no. 1759, PG II (solid) Class 8, UN no. 1760, PG II (liquid)

Packaging has been harmonized for all means of transport. Intermediate bulk containers and portable tanks are permitted for land and sea transport, with restrictions for solid polysulfide in portable tanks.

10. References

1. R. Steudel, *Angew. Chem.* **87** (1975) no. 19, 683.
2. F. A. Cotton, G. Wilkinson, P. L. Gaus: *Grundlagen der Anorganischen Chemie,* VCH Verlagsgesellschaft, Weinheim 1990, p. 160.
3. *Römpp,* 9th ed., **5,** 4377.
4. T. G. Pearson, P. C. Robinson, *J. Chem. Soc.* 1930, 1473.
5. N. K. Gupta, R. P. Tischer, *J. Electrochem. Soc.* **119** (1972) 1033.
6. E. Rosen, R. Tegman, *Chem. Scr.* **2** (1972) 221.
7. D.-G. Oei, *Inorg. Chem.* **12** (1973) 435.
8. D.-G. Oei, *Inorg. Chem.* **12** (1973) 438.
9. B. Cleaver, S. M. Upton, *Electrochim. Acta* **36** (1991) no. 3/4, 673 – 677.
10. W. Fischer: "Die Natrium-Schwefel-Batterie," *Elektrochem. Energietech. Entwicklungsstand Aussichten* 1981, 185 – 205.
11. P. Böttcher, *Z. Anorg. Allg. Chem.* **467** (1980) 149 –157.
12. F. Fehér, H. J. Berthold, *Z. Anorg. Allg. Chem.* **273** (1953) 144.
13. A. P. Brown, J. E. Battles, *Synth. React. Inorg. Met. Org. Chem.* **14** (1984) no. 7, 945 – 951.
14. B. Cleaver, S. J. Sime, *Electrochim. Acta* **28** (1983) no. 5, 703 – 708.
15. N. N. Greenwood, A. Earnshaw: *Chemie der Elemente,* 1st ed., VCH Verlagsgesellschaft, Weinheim 1988, p. 893.
16. G. J. Janz, D. J. Rogers, *J. Appl. Electrochem.* **13** (1983) 121 – 131.
17. (a) B. Warnqvist, *Thermochim. Acta* **37** (1980) 343 345. (b) R. Tegman, *Chem. Scr.* **9** (1976) 158 – 166.
18. E. Rosén, *K. Tek. Hoegsk. Handl.* (1960) 159.
19. E. Rosén, *Sven. Kem. Tidskr.* **76** (1964) 195.
20. H. Föppl, E. Busmann, F.-K. Frorath, *Z. Anorg. Allg. Chem.* **314** (1962) 12.
21. J. S. Thomas, A. Rule, *J. Chem. Soc.* **105** (1914) 177.
22. R. Tegman, *Acta Crystallogr. Sect. B: Struct. Sci.* **B 29** (1973) 1463.
23. R. Steudel, F. Schuster, *Z. Naturforsch. A: Phys. Phys. Chem. Kosmophys.* **32 A** (1977) 1313 – 1319.
24. H. H. Eysel, G. Wieghardt, H. Kleinschmagger, G. Weddigen, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **31B** (1976) 415.
25. H. Köpf, B. Block, *Chem. Ber.* **102** (1969) 1504 – 1508.
26. T. P. Whaley in A. F. Trothman-Dickenson (ed.): *Comprehensive Inorganic Chemistry,* **vol. 2,** Pergamon Press, Oxford 1973, p. 369.
27. B. Cleaver, A. J. Davies, M. D. Hames, *Electrochim. Acta* **18** (1973) 719 – 726.
28. B. Cleaver, A. J. Davies, *Electrochim. Acta* **18** (1973) 727 – 731. B. Cleaver: "Properties of Polysulfide Melts," in R. P. Tischer (ed.): The Sulfur Electrode, Fused Salts and Solid Electrolytes, vol. II, Academic Press, New York 1983, pp. 36 – 78.
29. *Mellor's Comprehensive Treatise on Inorganic and Theoretical Chemistry,* **vol. II,** suppl. II, part I, Longman Group Ltd., London 1961, p. 991.
30. *Gmelin,* **21,** 1049.
31. *Kirk-Othmer,* 3rd ed., **21,** 256.
32. *Ullmann,* 4th ed., **17,** 167.

33. M. Schmidt, *Inorg. Macromol. Rev.* **1** (1970) 101 – 113.
34. M. Schmidt, W. Siebert: "Sulfane," in *Comprehensive Inorganic Chemistry,* vol. 2, sect. 2.1, Pergamon Press, Oxford 1973, pp. 826 – 842.
35. A. Müller, E. Diemann, *Adv. Inorg. Chem. Radiochem.* **31** (1987) 89.
36. L. Hockenberger, *Chem. Ing. Tech.* **36** (1964) no. 10, 1046.
37. T. B. Rauchfuss, M. Draganjac, *Angew. Chem.* **97** (1985) 745 – 760.
38. Degussa, US 4 640 832, 1985.
39. Degussa, US 5 080 881, 1989.
40. Degussa, US 5 075 098, 1990.
41. Degussa, US 5 039 506, 1990.
42. Degussa, US 5 039 505, 1990.
43. I. G. Farbenind., DE 590 278, 1933 (E. Reißmann, H. Wolff).
44. Thiokol Chem. Corp., US 2 796 325, 1951 (E. R. Bertozzi).
45. Winnacker-Küchler: *Chemische Technologie,* 4th ed., **2,** 83.
46. Kali-Chemie, DE 842 789, 1950. TH. Goldschmidt, DE 920 010, 1949. R. Botson, US 1 736 741, 1927. I. G. Farbenind., DE 499 417, 1925.
47. Bayer, DE 1 016 242, 1958 (H. Beyer).
48. Sankyo Kasei, EP 0 361 998 A1, 1990 (Maeda, Kannosuke).
49. Verein. Chem. Fabr. Mannheim, DE 194 882, 1907.
50. Degussa, DE-AS 1 084 701, 1959.
51. BASF, DE-AS 1 122 045, 1960.
52. BASF, DE-AS 1 102 114, 1959.
53. Bayer, DE 814 140, 1949.
54. Solvay Process, US 2 376 433, 1944; US 2 376 434, 1944; US 2 376 435, 1944.
55. Deutsche Erdöl-AG, DE-AS 1 144 241, 1961.
56. J. Hahn, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **40B** (1985) 262 – 272.
57. F. Fehér, M. Baudler: "über das Ramanspektrum und die Konstitution des Wasserstoffdisulfides," *Z. Elektrochem.* **47** (1941) 844. F. Fehér, M. Baudler: "über eine neue Darstellungsmethode von Wasserstoffdisulfid," *Z. Anorg. Allg. Chem.* **253** (1947) 170. F. Fehér, M. Baudler: "Präparative Darstellung und Eigenschaften des Wasserstofftrisulfids," *Z. Anorg. Allg. Chem.* **254** (1947) 251. F. Fehér, M. Baudler: "über die Existenz und Darstellung von Wasserstofftetrasulfid," *Z. Anorg. Allg. Chem.* **254** (1947) 289. F. Fehér: "Neue Ergebnisse auf dem Gebiet der Polyschwefelwasserstoffe," *Angew. Chem.* **A 59** (1947) 33. F. Fehér, B. Talpay, E. Heuer: "über ein neues analytisches Bestimmungsverfahren für Polyschwefelwasserstoffe," *Z. Anorg. Allg. Chem.* **255** (1948) 316. F. Fehér, E. Heuer: "über die elektrochemische Darstellung von Polyschwefelwasserstoffen," *Angew. Chem.* **A 59** (1947) 237. F. Fehér, E. Heuer: "Zur Thermochemie der Polyschwefelwasserstoffe," *Z. Anorg. Allg. Chem.* **255** (1947) 185. F. Fehér, M. Baudler: "über das Wasserstoffpentasulfid H_2S_5 und das Wasserstoffhexasulfid H_2S_6," *Z. Anorg. Allg. Chem.* **258** (1949) 132. F. Fehér, H. Klug, L. Emmerich: "über die maß-analytische Bestimmung von Schwefelwasserstoffverbindungen," *Z. Anorg. Allg. Chem.* **260** (1949) 273. F. Fehér, H. J. Berthold: "über langkettige Glieder der Schwefelwasserstoffreihe H_2S_x," *Z. Anorg. Allg. Chem.* **267** (1952) 251. F. Fehér, W. Laue, J. Kraemer: "über die Darstellung von Schwefelölen S_xH_2 und S_xCl_2 mit wählbarer mittlerer Kettenlänge," *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **7B** (1952) 574. F. Fehér, E. Schliep, H. Weber: "über die elektrochemische Darstellung der Sulfane H_2S_n," *Z. Elektrochem.* **57** (1953) 916. F. Fehér, W. Laue, J. Kraemer: "Die Reaktion zwischen Sulfanen H_2S_n and Chlorsulfanen S_nCl_2," *Z. Anorg. Allg. Chem.* **281** (1955) 151. F. Fehér, W. Laue: "über die Darstellung von Rohsulfanen," *Z. Anorg. Allg. Chem.* **288** (1956) 103. F. Fehér, W. Laue, G. Winkhaus: "über die Darstellung der Sulfane H_2S_2, H_2S_3, H_2S_4 und H_2S_5," *Z. Anorg. Allg. Chem.* **188** (1956) 113. F. Fehér, G. Winkhaus: "über die Darstellung der Sulfane H_2S_5, H_2S_6, H_2S_7 und H_2S_8," *Z. Anorg. Allg. Chem.* **288** (1956) 123. F. Fehér, H. Weber: "Kinetische Untersuchungen über den thermischen Zerfall des Disulfans H_2S_2," *Z. Elektrochem.* **61** (1957) 285. F. Fehér, W. Laue, G. Winkhaus: "Molvolumina und Molrefraktionen der Sulfane und ihre Abhängigkeit von der Kettenlänge," *Z. Anorg. Allg. Chem.* **290** (1957) 52. F. Fehér, W. Kruse, W. Laue: "über die Viskositäten der Sulfane," *Z. Anorg. Allg. Chem.* **292** (1957) 203. F. Fehér, G. Winkhaus: "Zur Thermochemie der Sulfane: Bildungsenthalpien und Bindungsenergien," *Z. Anorg. Allg. Chem.* **292** (1957) 210. F. Fehér, W. Kruse: "über die Reaktion des Schwefelwasserstoffes mit Chlorsulfanen bzw.

Chlor oder Brom (Synthese niederer Sulfane)," *Z. Anorg. Allg. Chem.* **293** (1958) 302.
58. H. Viola, *Z. Chem.* **27** (1987) 15 – 22.
59. *Gmelin,* suppl. 4a/b, 389 – 498.
60. J. Hahn et al., *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **46B** (1991) 1338 – 1342.
61. D. Mauer et al., *Z. Naturforsch. A: Phys. Phys. Chem. Kosmophys.* **43A** (1988) 617 – 620.
62. M. Liedtke et al., *J. Phys. Chem.* **97** (1993) no. 43, 11204 – 11210.
63. J. Hahn, R. Hosch, *GIT Fachz. Lab.* **36** (1992) 727 –728.
64. H. Schulz, Thesis, Universität Würzburg 1969. M. Schmidt, H. Schulz, *Chem. Ber.* **101** (1968) 277. M. Schmidt, H. Schulz, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **23B** (1968) 1540.
65. A. P. Hagen (ed.): *Inorganic Reactions and Methods,* **vol. 5,** VCH Publishers, New York 1991, p. 27.
66. B. Meyer in H. J. Emeléus, A. G. Sharpe (eds.): *Advances in Inorganic Chemistry and Radiochemistry,* **vol. 18,** Academic Press, New York 1976, p. 287.
67. Hollemann-Wiberg: *Lehrbuch der Anorganischen Chemie,* 91th – 100th ed., Walter de Gruyter, Berlin 1985.
68. F. Fehér in G. Brauer (ed.): *Handbook of Preparative Inorganic Chemistry,* 2nd ed., Academic Press, New York 1963.
69. R. K. Kerr, E. M. Berlie, *Energy Process. Can.* **69** (1977) 48 – 50.
70. J. L. Sudworth, A. R. Tilley: "The Sodium-Sulfur Battery with β-Alumina Electrolyte," in R. P. Tischer (ed.): *The Sulfur Electrode, Fused Salts and Solid Electrolytes,* vol. VII, Academic Press, New York 1983, pp. 235 – 325.
71. Winnacker-Küchler, *Chemische Technologie,* 4th ed., **7,** 41. *Kirk-Othmer,* 3rd. ed., **22,** 168.
72. U. Deschler, P. Kleinschmit, P. Panster, *Angew. Chem.* **98** (1986) 237 – 253.
73. Degussa AG, Geschäftsbereich AC; Company brochures:"Organosilane für die Gummiindustrie" Jan.1994/Nr. 6000.0. Schriftenreihe Pigmente, Nr. 75, "Degussa-Silane" Pig. 75-2-3-1087 DD, Okt. 1987."Si 69X 50-SX50" Verstärkungsadditiv, Information für die Gummiindustrie PT68-2-3-691 Ha.
74. *Kirk-Othmer,* 3rd ed., **18,** 793.
75. *Ullmann,* 5th ed., **21,** 429.
76. H. Cherdron, F. Herold, A. Schneller, *Chem. Unserer Zeit* **23** (1989) no. 6, 181.
77. W. Koch, W. Heitz, *Makromol. Chem.* **184** (1983) 779 –792.
78. *Kirk-Othmer,* 3rd ed., **18,** 793.
79. Société Nationale des Pétroles d'Aquitaine, DE 2 157 575, 1972 (J. B. Signouret, Y. Labat, C. Esclamadon).
80. Société Nationale Elf Aquitaine S.A., DE 2 757 148, 1978.
81. J. Kausch, DD 156 709, 1982.
82. K. Negoro, T. Watanabe, *Nippon Gomu Kyokaishi* **43** (9) , 743 – 752; *Chem. Abstr.* **73** (1970) : 121333 b.
83. Toray Industries, JP 46 014 671, 1971 (S. Tanemoto, M. Takahashi).
84. W. Schulz, Universität Frankfurt/Main 1979, Ph. D. Thesis: "Untersuchungen an Verbindungen mit verschiedenartigen Sulfidbrücken"
85. W. F. Giggenbach, *Inorg. Chem.* **13** (1974) 1730.
86. W. F. Giggenbach, *Inorg. Chem.* **13** (1974) 1724.
87. R. Steudel, G. Holdt, R. Nagorka, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **41B** (1986) 1519 – 1522.
88. M. B. Berenbaum: *Encylopedia of Polymer Science and Technology,* **vol. 11,** Wiley-Interscience" pp. 425 –447.
89. G. Wilhelm, *Adhäsion* **5** (1976) 156.
90. Dow Chemical, US 4 438 259, 1984 (V. E. Meyer, T. E. Dergazarian).
91. Winnacker-Küchler, *Chemische Technologie,* 4th ed., **6,** 2.14.
92. Henkel, DE 3 604 793 A1, 1986 (G. Borggrefe).
93. *Ullmann,* 4th ed., **20,** 552.
94. E. Cooper, EP 32 281, 1981 (A. G. Papay, J. P. O'Brien).
95. Mobil Oil, EP 7 735, 1980 (A. G. Horodysky, S. Landis).
96. Lubrizol, US 4 119 549, 1978 (K. E. Davis).
97. E. Cooper, US 4 563 302 A, 1986 (P. G. Griffia, W. Y. Lam).
98. Mobil Oil, US 4 123 372, 1978 (R. F. Bridger, P. S. Landis).
99. Institut Français du Pétrole, FR 2 588 881 A1, 1987 (M. Born, L. Briquet, G. Parc).
100. Lubrizol, US 86-838 234, 1986 (C. P. Bryant, S. Q. A. Rizvi, K. B. Grover, J. N. Vinci).
101. Phillips Petroleum, US 3 759 956, 1973 (P. R. Strapp).
102. A. M. Kuliev, F. I. Gasanov, F. N. Mamedov, *Prisadki Smaz. Maslam* **2** (1969) 27 – 9; *Chem. Abstr.* **72** (8) : 33982m.
103. A. M. Kuliev, K. I. Sadykhov, R. K. Mamedova, N. S. Kerimov, S. M. Abutalybova, *Azerb. Neft. Khoz.* **53** (1973) 40; *Chem. Abstr.* **80** (10) : 50087 f.

104. Magyar Asvanyolai Es Foldgaz Kiserleti Intezet, GB 2 193 957 A1, 1988 (R. Csikos, S. Borzsonyi, P. L. Farkas).
105. R. Böttcher, S. Wartewig, W. Windisch, A. Zschunke, *Z. Naturforsch. A: Astrophys. Phys. Phys. Chem.* **23A** (1968) 1766 – 1770.
106. Toyo Soda Mfg., JP 55 071 761, 1980; *Chem. Abstr.* **93** (16) : 151794h.
107. F. Seel, H. J. Güttler, *Angew. Chem.* **85** (1973) 416.
108. F. A. Cotton, J. B. Harmon, R. M. Hedges, *J. Am. Chem. Soc.* **98** (1976) 1417.
109. Nitto Chemical Industry Co., JP 50 060 050, 1975 (I. Watanabe); *Chem. Abstr.* 83 (12) : 102984w.
110. Yuasa Battery, JP 56 058 582, 1981; *Chem. Abstr.* **95** (16) : 138034g.
111. F. Holdinghausen, D. Männig, R. Wenzl, company report, Degussa AG, Frankfurt/Main 1994.
112. Gefahrstoffverordnung, 4th ed., 12/1993, Deutscher Bundes-Verlag, Bonn 1993.
113. Kühn-Birret-Merkblätter, 13. Erg.-Lfg. 11/80-N 19-2, Ecomed-Verlagsgesellschaft, Landsberg.
114. Roth-Weller: *Sicherheitsfibel Chemie,* 5th ed., Ecomed-Verlag, Landsberg/Lech 1991.
115. Daten Chemischer Stoffe, 28. Erg.Lfg. 12/92, Ecomed-Verlag, Landsberg/Lech 1992.
116. Wasserhaushaltsgesetz (WHG) in der Fassung der Bekanntmachung vom 23. Sept. 1986, Bundesgesetzblatt I, p. 1529.
117. Degussa AG: brochure “Umwelt und Degussa” no. CH 460-4-05-290 and leaflet no. CH 337-3-05-491, Degussa AG, Hanau 1991.
118. Allgemeine Rahmen-Verwaltungsvorschrift über Mindestanforderungen an das Einleiten von Abwässer (Rahmen-Abwasser-VwV) in Gewässer, Gemeinsames Ministerialblatt Nr. 25 vom 22. Sept. 1989.
119. H. W. Sonntag, company report, Degussa AG, Frankfurt/Main 1994.
120. R. Neureiter, Degussa AG, company information, Frankfurt/Main 1994.

Sulfinic Acids and Derivatives

RÜDIGER SCHUBART, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Properties 2
3. Preparation of Sulfinic Acids and their Salts 2
3.1. Reduction of Sulfonic Acid Derivatives 2
3.2. Sulfination with Sulfur Dioxide . . . 4
3.3. Oxidation of Thiols 4
3.4. Nucleophilic Cleavage of Sulfones and Related Compounds 5
4. Reactions of Sulfinic Acids and their Derivatives 7
4.1. Alkylation of the Sulfinate Anion . . 7
4.2. Addition of Sulfinic Acids to Multiple Bonds 7
4.3. Redox Reactions of Sulfinic Acids . 7
5. Sulfinic Acid Derivatives 8
5.1. Sulfinyl Chlorides 8
5.1.1. Preparation 8
5.1.2. Reactions 9
5.2. Sulfinic Acid Esters 10
5.2.1. Preparation 10
5.2.2. Reactions 10
5.3. Thiosulfinic Acid Esters 11
5.4. Sulfinic Acid Amides 11
6. 1-Hydroxyalkanesulfinates 12
6.1. Properties 12
6.2. Production 12
6.3. Quality Specifications, Analysis . . . 13
7. Formamidinesulfinic Acid 13
7.1. Properties 13
7.2. Production 13
8. Industrial Uses of Sulfinic Acids and their Derivatives 14
9. Toxicology 15
10. References 15

1. Introduction

Sulfinic acids are synthetic chemical building blocks that also occur in living organisms. For example, allicin, *S*-2-propenyl 2-propene-1-sulfinothioate [*539-86-6*] (**1**) is a component of extract of garlic [1]. It can be prepared by oxidation of diallyl disulfide with *m*- chloroperbenzoic acid [2]:

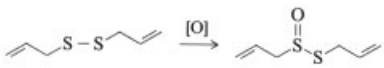

1

Allicin has antibiotic properties [3], and it has also been found to be an effective antithrombotic and an inhibitor of blood platelet aggregation [3,4]. In inflammatory processes (e.g., allergies, asthma, rheumatism, etc.) allicin causes 100 % inhibition of lipoxygenases [5] as well as inhibiting the formation of arachidonic acid [6], which plays an important role in such processes (→ Anti-inflammatory – Antirheumatic Drugs). Animal experiments have shown that cholesterol formation is also significantly lowered by allicin [7]. The biosynthesis of taurine, an important neurotransmitter in living organisms (→ Neuropharmacology, Chap. 5.2.1.), proceeds from cysteine via cysteinesulfinic acid and then hypotaurine [8]. More information in [9].

2. Properties

Physical Properties. Table 1 lists a number of aliphatic and aromatic sulfinic acids together with their CAS registry numbers, melting points, and methods of synthesis.

Chemical Properties. Sulfinic acids are compounds with the following general formula [33,34]:

Sulfur has a coordination number of 3 in these compounds, and is in the +4 oxidation

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_461

Table 1. Physical data for sulfinic acids whose methods of synthesis are described in the text (all compounds are colorless in the pure state)

Sulfinic acid	Molecular formula (molecular mass)	CAS registry no.	*mp*, °C	Method of synthesis	Reference
$C_2H_5-SO_2H$	$C_2H_6O_2S$ (94.13)	[*598-59-4*]	liquid	oxidation of ethanethiol with *m*-chloroperbenzoic acid reduction of ethanesulfonyl chloride with zinc cleavage of sulfones with thiolate	[10] [11] [12]
$H_3C-(CH_2)_3-SO_2H$	$C_4H_{10}O_2S$ (122.19)	[*55109-28-9*]	liquid	oxidation of butanethiol with *m*-chloroperbenzoic acid reduction of butanesulfonyl chloride with triethylaluminum cleavage of sulfones with sodium cyanide from butyllithium and sulfur dioxide	[10] [13] [14] [15]
$H_3C-(CH_2)_7-SO_2H$	$C_8H_{18}O_2S$ (178.30)	[*3944-71-6*]	5–8	reduction of octanesulfonyl chloride with triethylaluminum	[13]
$H_3C-(CH)_{11}-SO_2H$	$C_{12}H_{26}O_2S$ (234.40)	[*26535-63-7*]	35–36	from dodecylmagnesium chloride and sulfur dioxide cleavage of sulfones with thiolate	[16] [12]
$HO_2S-(CH_2)_4-SO_2H$	$C_4H_{10}O_4S_2$ (186.25)	[*6340-77-8*]	122–126	reduction of 1,4-butanebis(sulfonyl chloride) with sodium sulfite	[17], [18]
Cl_3C-SO_2H	$CHCl_3O_2S$ (183.44)	[*7430-24-2*]	liquid	reduction of trichloromethanesulfonyl chloride with hydrogen sulfide	[19]
F_3C-SO_2H	CHF_3O_2S (134.08)	[*34642-42-7*]	liquid	reduction of trifluoromethanesulfonic acid with hydrazine reduction of trifluoromethanesulfonic acid with zinc reduction of trifluoromethanesulfonic acid with potassium iodide in acetone	[20] [21] [22]
$4\text{-}CH_3-C_6H_4-SO_2H$	$C_7H_8O_2S$ (156.20)	[*536-57-2*]	84	reduction of 4-toluenesulfonyl chloride with triethylaluminum reduction of 4-toluenesulfonyl chloride with zinc reduction of 4-toluenesulfonyl chloride with lithium aluminum hydride from toluene and sulfur dioxide/aluminum trichloride from the diazonium salt of toluene and sulfur dioxide–copper powder	[13] [23] [24] [25] [26]
$4\text{-}(H_3C-CO-NH)-C_6H_4-SO_2H$	$C_8H_9NO_3S$ (199.23)	[*1126-81-4*]	155	reduction of *p*-acetamidobenzenesulfonyl chloride	[27], [28]
$C_6H_5-CH_2-C_6H_4-4\text{-}SO_2H$	$C_{13}H_{12}O_2S$ (232.30)	[*58661-47-5*]	70–72	intramolecular rearrangement of diarylsulfones with butyllithium	[29]
$H_3C-CO-CH_2-CH_2-SO_2H$	$C_4H_8O_3S$ (136.17)		94–95	cleavage of sulfones with sodium cyanide or thiolate	[30]
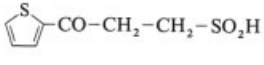	$C_7H_8O_3S_2$ (204.27)	[*65373-90-2*]	91–92	cleavage of sulfones with sodium cyanide or thiolate	[30]
$H_2N-C(NH)-SO_2H$	$CH_4N_2O_2S$ (108.12)	[*1758-73-2*]	144 (decomp.)	oxidation of thiourea with hydrogen peroxide	[31]
$HO-CH_2-SO_2Na$	CH_3NaO_3S (118.08)	[*149-44-0*]	63–64	from formaldehyde and sodium dithionite	[32]

state. Sulfinic acids have an asymmetric, pyramidal structure in which the sulfur atom constitutes a chiral center. Free sulfinic acids are relatively unstable compounds. On storage or heating they tend to disproportionate (via intermediates) with the loss of water to give sulfonic acids and S-esters of thiosulfonic acids [15, 16, 35–38]. Some sulfinic acids decompose as soon as they are formed. For example, benzo-1,3-thiazole-2-sulfinic acid desulfinates immediately upon acidification of the sodium salt, giving benzo-1,3-thiazole (**2**) [39]:

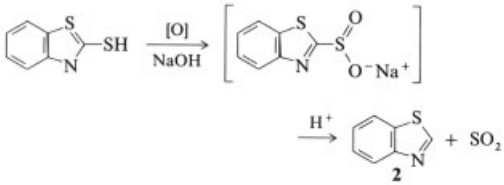

Aromatic sulfinic acids are generally more stable than long-chain aliphatic sulfinic acids, which are in turn more stable than their short-chain counterparts. Sulfinic acids are best stored as the alkali metal salts, from which they can be liberated with mineral acid [40].

Sulfinic acids are somewhat weaker acids than the corresponding sulfonic acids (→ Sulfonic Acids, Aliphatic), with pK_a values between 1.2 (aromatic sulfinic acids) and 2.2 (aliphatic sulfinic acids). Aromatic sulfinic acids generally crystallize readily as colorless compounds that are sparingly soluble in water but readily soluble in diethyl ether. Aliphatic sulfinic acids are often obtained as oily or low-melting compounds that are soluble in many organic solvents. Further information in [41].

3. Preparation of Sulfinic Acids and their Salts

Sulfinic acids can be synthesized by reduction of compounds containing sulfur in the +6 oxidation state, transformation of other sulfur +4 compounds, oxidation of sulfur +2 compounds, and nucleophilic cleavage of sulfone-like species; review in [42].

3.1. Reduction of Sulfonic Acid Derivatives

The preparation of sulfinic acids and their salts by reduction of sulfonic acid derivatives is straightforward, especially from the readily accessible sulfonyl chlorides.

$$R-SO_2Cl + 2e^- \longrightarrow R-SO_2^- + Cl^-$$

Sodium sulfite, zinc, iron, magnesium, and stannous chloride are frequently used as reducing agents [11, 18, 23, 28, 36, 40, 43–45]. The use of lithium aluminum hydride [24], alkylaluminum compounds [13], sodium in liquid ammonia [46], and electrochemical methods [47], has also been investigated. These alternative procedures often lead to high yields, but they have not been adopted on an industrial scale because of technical complications or cost. For additional synthetic information see [48].

Synthesis of 1,4-Butanedisulfinic Acid [17, 18]. A solution of 65.5 g (0.52 mol) anhydrous sodium sulfite in 310 mL of water is mixed with 128 g (0.52 mol) of magnesium sulfate heptahydrate to give a thick precipitate. This mixture is heated to 40–50 °C and 51 g (0.2 mol) of 1,4-butanebis(sulfonyl chloride) is added. The mixture is kept alkaline by the slow addition of ca. 23 g of magnesium oxide. The sulfinic acid precipitates as a magnesium salt. Acidification with dilute sulfuric acid, extraction with diethyl ether, and evaporation gives 1,4-butanedisulfinic acid, *mp* 122–126 °C.

Sodium Ethanesulfinate [11]. 436 g (3.4 mol) of ethanesulfonyl chloride is added dropwise with stirring over 3–4 h to a mixture of 500 mL of water, 500 g of ice, and 340 g (5.2 mol) of zinc dust. The temperature is maintained at +5 °C by continuous addition of 500 g of ice. The reaction mixture is stirred for 1 h and then a solution of 350 g of sodium carbonate decahydrate in 1.5 L of water at 80–90 °C is added over 1.5 h. The reaction mixture should remain distinctly alkaline. It is subsequently filtered, and the residue is washed with warm water. The filtrate and washings are evaporated to give 232 g of the sodium salt, from which the free sulfinic acid can be obtained.

3.2. Sulfination with Sulfur Dioxide

Sulfur dioxide sulfinates both aliphatic and aromatic organometallic compounds, (e.g., Grignard reagents [45], organoaluminums [49]), and particularly organolithiums [50], to give the corresponding sulfinic acid salts:

p-Dodecylbenzenesulfinic Acid [50]. A solution of 29.2 g (0.09 mol) *p*-dodecylbromobenzene in 120 mL of diethyl ether is added dropwise to 1.5 g (0.22 mol) of finely cut lithium. The reaction mixture is stirred under reflux for 1h and then cooled in an ice bath, after which sulfur dioxide is passed through the cooled solution for ca. 30 min. A further 100 mL of diethyl ether is added during the sulfination. The reaction mixture is centrifuged and the residue removed from the centrifuge vessel with dilute mineral acid. The resulting suspension is extracted with diethyl ether, the solution is evaporated, and the *p*-dodecylbenzenesulfinic acid residue in 53 % yield is dried, *mp* 54–55 °C.

1-Dodecanesulfinic Acid [49]. A solution of 70 g (0.13 mol) tridodecylaluminum in 78 mL of toluene is added dropwise over 15 min to a mixture of 60 g (0.95 mol) of sulfur dioxide and 240 g of toluene at − 45 °C. The reaction mixture is allowed to warm to room temperature and the solvent is removed by distillation in vacuo. The residue is the aluminum salt of dodecylsulfinic acid, from which the sulfinic acid is liberated in 98 % yield, *mp* 35 °C.

Sulfur dioxide also reacts with aromatic hydrocarbons and aluminum trichloride to give sulfinic acids by a Friedel–Crafts-type mechanism [25,40]. Alkanes react as well under high pressure to give the corresponding alkyl sulfinic acids [51], and alkenes undergo the same reaction at atmospheric pressure. Photochemical sulfinations of alcohols in liquid sulfur dioxide [52] and of alkenes and alkanes in the gas phase are also known [53]. Sulfination of aromatic diazonium salts with sulfur dioxide in the presence of copper salts or copper powder is often used for the synthesis of sulfinic acids [26,43]. For the preparation of sodium 2,2,2-trifluoro-1,1-dichloroethanesulfinate see [54].

3.3. Oxidation of Thiols

Thiols can be easily oxidized to sulfinic acids, but it is often difficult to limit the oxidation of sulfur to the +4 state [55]. Halogens [56], hydrogen peroxide, and peracids [40] have been used as oxidizing agents. *m*-Chloroperbenzoic acid is particularly applicable because it does not result in further oxidation to sulfonic acids or sulfones [10]:

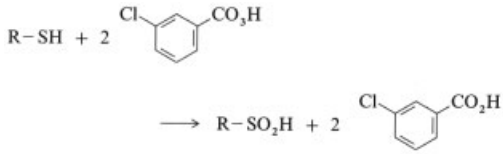

Ethanesulfinic Acid [10]. A solution of 17.2 g (0.1 mol) *m*-chloroperbenzoic acid in 200 mL of dichloromethane at − 30 °C is added slowly to a solution of 3.1 g (0.05 mol) of ethanethiol in 10 mL of dichloromethane (also at − 30 °C). The reaction mixture is allowed to stand overnight at − 30 °C and the precipitated *m*-chlorobenzoic acid is removed by filtration. The filtrate is cooled to − 80 °C and refiltered to remove residual *m*-chlorobenzoic acid, after which the solvent is evaporated in a stream of nitrogen. Analytically pure ethanesulfinic acid is obtained in 80–85 % yield.

The oxidation of 2-mercaptobenzo-1,3-thiazole with hydrogen peroxide in alkaline medium gives the sodium salt of benzo-1,3-thiazole-2-sulfinic acid, which (see Chap. 2) is converted immediately upon acidification into benzo-1,3-thiazole (**2**, 70 %) [39].

3.4. Nucleophilic Cleavage of Sulfones and Related Compounds (→ Sulfones and Sulfoxides)

Cleavage of the Sulfonyl Group at the α-Carbon Atom. The cleavage of aromatic sulfones carrying electron-withdrawing groups on the aromatic ring constitutes a general method for the synthesis of aromatic sulfinic acids [57]:

Intramolecular rearrangement (Smiles rearrangement) of the lithium salt of 2-hydroxyphenyl phenyl sulfone gives lithium 2-phenoxyphenylsulfinate (**3**) [58]:

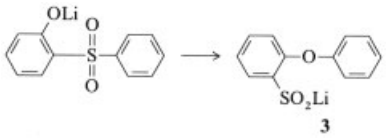

The reaction of 2-methylphenyl phenyl sulfone (**4**) with butyllithium is similar to the Smiles rearrangement [29]:

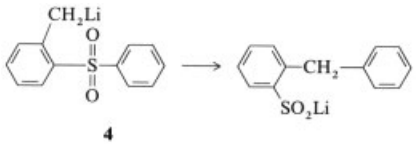

For further information see [59]; for electrochemical cleavage of sulfones see [60].

The facile cleavage of symmetrical aliphatic sulfones with sodium cyanide when both β-positions are activated by electron-withdrawing substituents is of greater interest [30]:

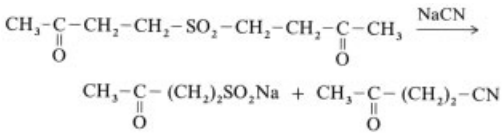

Similarly, α-disulfones are readily cleaved with sodium cyanide [14], and β- cyanosulfones with sodium thiolate (see below) [61]. Furthermore, the cleavage of phthalimidomethyl sulfones with alkoxide or thiolate proceeds smoothly to give sulfinic acids in high yield [12]:

If the cleavage is carried out using thiolate, a phthalimidomethyl sulfide is formed as a byproduct. This sulfide can be reoxidized and itself cycled through the cleavage process, a reaction that permits the high-yield conversion of any thiol to the corresponding sulfinic acid.

General Example [12]. An alkylsulfonylmethylphthalimide (0.02 mol) is added to a solution of 0.51 g (0.022 mol) of sodium ethoxide in 65 mL of ethanol under nitrogen. The reaction mixture is stirred under reflux until a homogeneous solution is formed. After evaporation to dryness the residue is extracted several times with warm benzene. The sulfinate remains as an insoluble powder in a 92 – 100 % yield, purity 94 – 98 %.

Cleavage of Thiosulfonic Acid Esters. The cleavage of thiosulfonates with thiolate [62], sulfinate [63], and hydroxyl ions [64] can also be used for the preparation of sulfinic acids:

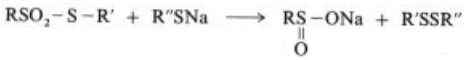

Cleavage at the β-Carbon Atom of a Sulfonyl Group. Both sulfones with β-hydrogens [65] and β-disulfones [66] undergo fragmentation with base to give the corresponding sulfinic acid salts:

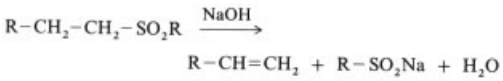

The following reaction sequence also gives high yields of sulfinic acids [67]:

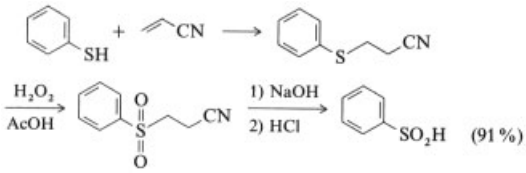

In addition, sulfinic acids can be prepared by alkylation of sulfonyl carbanions with subsequent elimination of sulfinate [68, 69]:

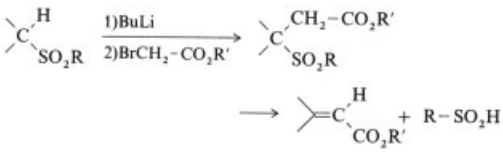

Cleavage of Sulfolenes. The ring-opening of a cyclic sulfone with a Grignard reagent represents another synthetic possibility. For example, in the case of 2-methylsulfolene (**5**) the product is a doubly unsaturated sulfinic acid [70]:

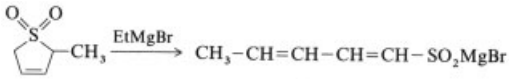

5

Addition of Thionyl Chloride to an Olefin with Subsequent Hydrolysis. Thionyl chloride adds to olefins in the presence of aluminum trichloride forming a complex that can be hydrolyzed with water to a sulfinic acid [71]:

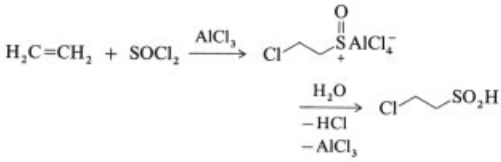

For further syntheses see [72].

4. Reactions of Sulfinic Acids and their Derivatives

4.1. Alkylation of the Sulfinate Anion

The sulfinate anion can be alkylated either at sulfur or oxygen depending on the reaction conditions, forming a sulfone or a sulfinic acid ester, respectively. Reaction with a soft alkylating agent (in the sense of the Pearson HSAB principle [73–75]) such as an alkyl halide, gives a sulfone [76]:

Hard alkylating agents, such as triethyloxonium tetrafluoroborate [77] or diazomethane [78], give sulfinic acid esters:

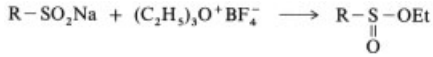

The solvent may also have a significant effect on the observed ratio of sulfone to sulfinic acid ester [79].

The alkylation reaction is an effective approach to preparing dialkyl sulfones, aralkyl sulfones, and β-hydroxysulfones [75, 80–82]. Reaction of sulfinates with epoxides leads to β-hydroxysulfones [83]; with lactones, sulfonylcarboxylic acids are formed [84]. Sulfones are also formed by the reaction of sulfinates with systems containing activated double bonds [85–87]. Formaldehyde reacts with sulfinate to give hydroxymethyl sulfone [88]. Simultaneous addition of an amine leads to an *N*-methylsulfonylamine in a reaction analogous to the Mannich reaction [89]. On the other hand, diaryl sulfones are obtained by nucleophilic substitution of haloaromatics containing electron-withdrawing substituents (e.g., 2-nitrochlorobenzene) [90]:

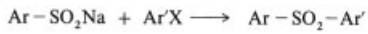

[*1-(Phenylsulfonyl)-aceto*]-*4-nitrobenzene* [82]. 244 g (1 mol) of (1-bromoaceto)-4-nitrobenzene and 164 g (1 mol) of sodium benzenesulfinate are suspended in 1 L of ethanol and boiled for 4 h. After cooling, the precipitate is separated by filtration, boiled with acetone, and filtered again. The filtrates are combined, evaporated, and cooled. The crystalline sulfone is isolated by filtration in 76 % yield, *mp* 138 °C.

For the synthesis of thioacylic sulfones see [91].

4.2. Addition of Sulfinic Acids to Multiple Bonds

Polarized double-bond systems react with aliphatic and aromatic sulfinic acids in a process similar to the Michael reaction. Acrylic acids and their derivatives [92], vinyl ketones [93], nitroolefins [94], vinyl sulfones [95], quinones [96–98], and quinone imines [98] all undergo this reaction with sulfinic acids to form sulfones; e.g.:

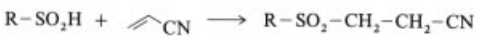

The addition of a sulfinic acid to an α-ethynyl ketone sulfone gives an α-oxo-β-sulfonylalkene in good yield [99]. A *cis*-1,2-disulfonylethylene is formed with 1-phenylsulfonylpropyne [100].

4.3. Redox Reactions of Sulfinic Acids

Sulfinic acids are readily oxidized to sulfonic acids by, for example, peroxides [101], hypochlorites [102], and oxygen [103]. α-Disulfones are formed with permanganate or cobalt(III) salts [104]. The reduction of sulfinic acids with lithium aluminum hydride gives disulfides [105]. As they are relatively unstable compounds, sulfinic acids undergo acid-catalyzed disproportionation to sulfonic acids and thiosulfonates [15, 16, 35–37].

5. Sulfinic Acid Derivatives

5.1. Sulfinyl Chlorides

5.1.1. Preparation

From Disulfides. Both aliphatic and aromatic sulfinic acids form sulfinyl chlorides. Low yields are generally obtained from the direct reaction of a sulfinic acid with thionyl chloride [106], but chlorination of a disulfide has proven to be an effective method of synthesis. The chlorination is carried out at 0 – 10 °C in acetic anhydride [107, 108]:

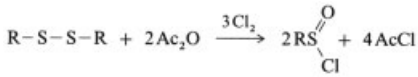

The course of the reaction is clearly indicated by the disappearance of the orange-red color of the intermediate sulfenyl chloride (RSCl).

Methanesulfinyl Chloride [107]. A mixture of 23.5 g (0.25 mol) of freshly distilled dimethyl disulfide and 51 g (0.5 mol) of acetic anhydride is chlorinated at − 10 to 0 °C. During the course of the reaction the color changes from yellow to red and then disappears, at which point the addition of chlorine is suspended, leading to 126 g of reaction mixture. This is distilled in a 45 cm Vigreux column at 2 kPa with cooling. Acetyl chloride distills at 0 °C, and at 47 – 48 °C the almost colorless methanesulfinyl chloride is obtained in 83 – 86 % yield.

Many sulfinyl chlorides decompose to varying degrees in the course of distillation, and in aromatic cases there is even a danger of explosion. Sulfinyl chlorides may also decompose slowly during storage, causing pressure to build up in a closed container.

From Trithianes. Chloromethanesulfinyl chloride (**7**) can be prepared by the chlorination of trithiane (**6**) in acetic anhydride at 0 °C [109]:

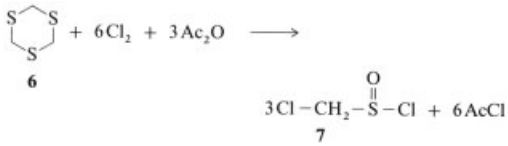

Chloromethanesulfinyl Chloride. A mixture of 46 g (0.33 mol) of *sym*-trithiane (**6**) and 102 g (1 mol) of acetic anhydride is chlorinated at 0 °C. Residual chlorine and resulting acetyl chloride are evaporated in vacuo and the chloromethanesulfinyl chloride (**7**) is distilled in 55 % yield (*bp* 42 – 62 °C at 2 kPa).

For α-heteroatom-substituted methanesulfinic acids and their derivatives see [110]; for the synthesis of trimethylammoniomethanesulfinate see [111].

From Sulfuranes ($RSHal_3$). The selective hydrolysis of sulfurane also gives a sulfinyl halide [112]:

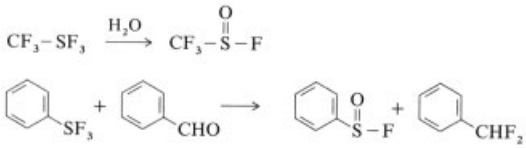

For the photolytic addition of sulfur dioxide to perfluoroalkenes giving sulfinyl fluorides see [113].

Trichlorosulfuranes [114, 115] are converted into sulfinyl chlorides in high yield in alcohol [116], carboxylic acids [117], or acetic anhydride [118]:

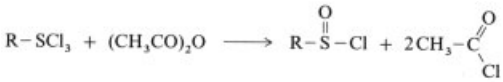

Because of the instability of sulfuranes, the compounds are usually synthesized immediately prior to use by further chlorination of the corresponding sulfenyl chloride.

From Thioesters. Another method for producing sulfinyl chlorides involves the chlorolytic cleavage of thioesters in acetic anhydride, leading to a mixture of the desired sulfinyl chloride and acetyl chloride [119]:

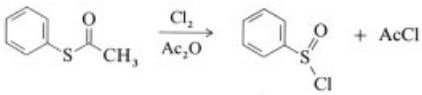

8

Phenylsulfinyl Chloride (**8**) [119]. A mixture of 30.4 g (0.2 mol) of phenyl thioacetate and 20.4 g (0.2 mol) of acetic anhydride is chlorinated at − 10 to 0 °C. The color changes to yellow, then red, and then back to yellow. The reaction mixture is heated to 80 °C under reduced pressure (3.5 kPa), until the residue begins to boil, after which it is cooled; yield of **8**: 31.5 – 31.8 g (98.1 – 99.1 %).

By the Addition of Thionyl Chloride to Alkenes. Addition of ethylene to thionyl chloride in the presence of aluminum trichloride gives the corresponding acid chloride complex, which upon hydrolysis forms 2-chloroethanesulfinic acid (**9**) [71]:

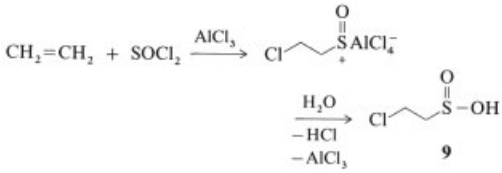

2-Chloroethanesulfinic Acid (**9**) [71]. A mixture of 32 g (0.27 mol) of thionyl chloride and 32 g (0.24 mol) of aluminum trichloride is saturated with 6 L (0.27 mol) of ethylene. The reaction mixture is hydrolyzed with water and the resulting byproducts are extracted with benzene. The residue is extracted with diethyl ether, which is then dried over magnesium sulfate and evaporated in vacuo. The yield of **9** is 98 %.

By Treatment of CH-Acidic Compounds with Thionyl Chloride. CH-acidic compounds react readily with thionyl chloride to give sulfinyl chlorides. For example, formylsulfinyl chloride can be obtained from isobutyraldehyde (2-methylpropanal) in 81 % yield [120]:

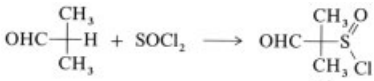

10

1-Formyl-1-methylethanesulfinyl Chloride (**10**) [120]. 72 g (1 mol) of 2-methylpropanal is

added dropwise to 360 g (3 mol) of thionyl chloride at 50 °C with cooling. As soon as the vigorous production of hydrogen chloride has subsided, the temperature is raised to 80 °C. When hydrogen chloride formation ceases completely excess thionyl chloride is removed by distillation at atmospheric pressure and the residue is distilled under vacuum, leading to 124 g, of **10**, *bp* 83 – 84 °C at 1.6 kPa.

For an analogous synthesis of adamantanesulfinyl chloride in 72 % yield see [121].

5.1.2. Reactions

Sulfinyl chlorides are important starting materials for the preparation of sulfinic acid esters (see Section 5.2) and amides (Section 5.4), as well as thiosulfinic acid esters (Section 5.3). For a transformation of sulfinyl chlorides into sulfines ($R_2C{=}S{=}O$) see [122]. Phenylsulfinyl chloride can be converted into a diarylthiosulfonate ester in dichloromethane in the presence of pyridine and 1-hydroxypyridine-2-thione [123].

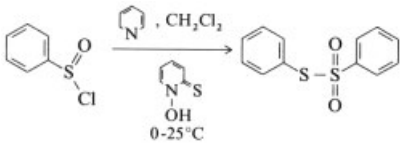

Asymmetric aromatic sulfoxides can be prepared by the Friedel – Crafts reaction of an aromatic sulfinyl chloride with a benzene derivative [124, 125].

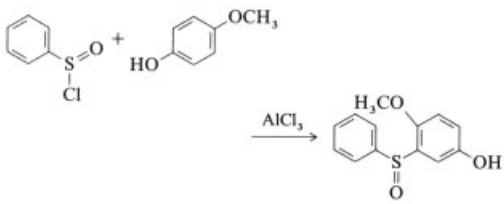

5.2. Sulfinic Acid Esters

5.2.1. Preparation

From Sulfinyl Chlorides and Alcohols. Sulfinic acid esters are readily prepared from sulfinyl chlorides and alcohols in an inert solvent [126]:

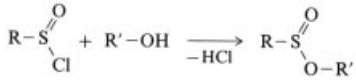

The hydrogen chloride formed can be removed by applying a vacuum [126], or by the addition of a base such as potassium carbonate [127] or pyridine [128]. For a one-pot synthesis of sulfinic acid esters (50 – 88 %) see [129].

Methyl Methanesulfinate [126]. A mixture of 47 g (0.5 mol) of dimethyl disulfide and 102 g (1 mol) of acetic anhydride is chlorinated to give methanesulfinyl chloride. The latter is then treated with 35 g (1.1 mol) of cold anhydrous methanol at – 30 °C. Hydrogen chloride is removed by applying a vacuum and the reaction mixture is heated to 30 °C until gas formation has ceased. *p*-Toluidine can be added to assist in the removal of methanesulfinyl chloride. The reaction mixture is filtered and the filtrate is distilled using an 18 cm Vigreux column in yield 71 %.

From Disulfides by Chlorination in Alcohol. If chlorination of a disulfide to give the corresponding sulfinyl chloride is carried out in an alcohol, the sulfinic acid ester is obtained directly [130]:

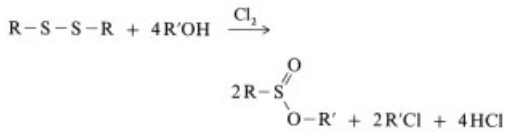

The oxidation of disulfides with lead tetraacetate in alcohol also leads to sulfinic acid esters in 30 – 90 % yield [131]. For an analogous reaction with diphenyl disulfide see [132]. Electron-withdrawing substituents on the aromatic ring inhibit this oxidation. *N*-Bromsuccinimide also acts as an oxidizing agent in the synthesis of

sulfinic acid esters (5 – 78 % yield) [133]. For cyclic sulfinic acid esters see [134].

Other Methods. The preparation of sulfinic acid esters from sulfinyl chlorides and diazoalkanes in polar solvents is described in [135] and from sulfinic acids and diazomethane in [78].

The reaction of benzenesulfinyl chloride with diazomethane is described in [109]. 4-Methylbenzenesulfinic acid reacts with diphenyldiazomethane in dichloromethane to give a sulfone as the major product, but in dimethylsulfoxide the sulfinic acid ester is formed in 100 % yield [135]. Another route involves the reaction of a sulfinate with a hard alkylating agent [78] (see Section 4.1).

Unlike carboxylic acids, sulfinic acids cannot be esterified by acid catalysis because under these conditions they disproportionate to sulfonic acids and thiosulfonates [15, 16, 35–37] (Section 4.3). However, esterification can be accomplished in a large excess of alcohol in the presence of dicyclohexyl carbodiimide [136]:

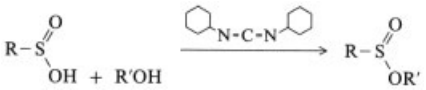

Sulfinic acid esters can also be obtained in good yield by the alcoholysis of *N*-sulfinylphthalimides [137]:

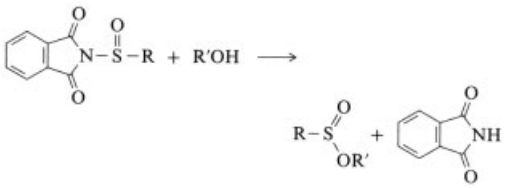

Synthesis of sulfinic acid esters by the alcoholysis of *N*-sulfinylsulfonamides is described in [138].

5.2.2. Reactions

Rearrangements. Alkane-, alkene, and alkynesulfinic acid esters undergo acid-catalyzed rearrangement to sulfones [139]; for example:

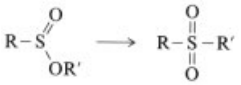

For further information see [140].

Conversion into Sulfoxides. Sulfinic acid esters react with Grignard reagents to give optically active sulfoxides [141]:

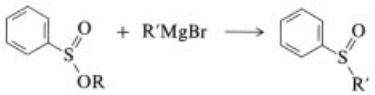

This reaction is highly stereospecific, and is, therefore, frequently used for the determination of absolute configurations.

Claisen Condensation. The Claisen condensation (→ Esters, Organic, Chap. 3.5.) of sulfinic acid esters with enolates leads to β-oxosulfoxides [142, 143], which can also be obtained by the reaction of sulfinyl carbanions with carboxylic acid esters:

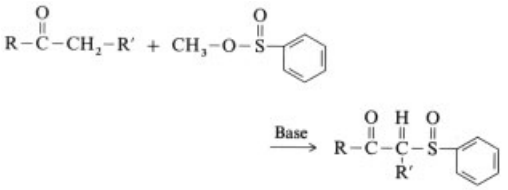

For further reactions (e.g., the transformation of sulfinic acid esters into disulfides) see [144]. Information about MS-Research of sulfinic acid esters in [145].

5.3. Thiosulfinic Acid Esters

Thiosulfinic acid esters are readily prepared from the corresponding disulfides by careful oxidation with, for example, *m*- chloroperbenzoic acid [2]. Allicin (**1**) can be obtained in good yield by this method.

The disproportionation of sulfinyl chlorides in water also gives thiosulfinic acid esters [146]. For further references see [147].

5.4. Sulfinic Acid Amides

Sulfinic acid amides can be readily produced from sulfinic acid chlorides and amines:

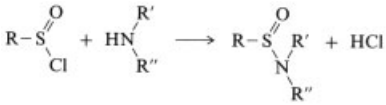

The reaction is carried out in inert, dry solvents in the presence of base or excess amine [148–150].

General Method of Preparation [149]. Sulfinyl chloride (0.2 mol) is added dropwise to a solution of 0.4 mol of amine in 150 mL of dichloromethane under nitrogen in the course of 1 h. The temperature of the reaction mixture is maintained at − 20 to − 40 °C. It is stirred for a further hour at room temperature, after which ammonium chloride is filtered off. The filtrate is concentrated by evaporation, and the product is distilled in vacuo or crystallized.

Sulfinic acid amides can also be prepared by the reaction of sulfinylamines with Grignard reagents [72, 151, 152]:

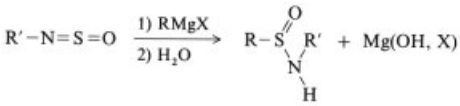

Good yields are obtained with aromatic sulfinylamines. The Grignard complex should be decomposed carefully with 10 % cold ammonium chloride solution.

For a discussion of the preparation of *N*-sulfinylnonafluorobutanesulfonamide via an ene reaction and its subsequent conversion into the sulfine – sulfone imide see [138].

Sulfinic acid amides can also be prepared directly by the oxidation of sulfenic acid amides. Thus, 2-benzo-1,3-thiazolesulfenamides can be transformed into 2-benzo-1,3-thiazolesulfinamides by careful oxidation. The latter can be used as vulcanization accelerators for natural and synthetic rubbers [150]. Sodium hypochlorite is generally used as the oxidizing agent [150]:

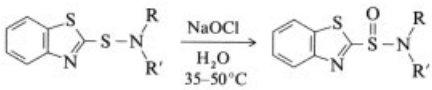

Percarboxylic acids and hydrogen peroxide are also suitable.

N,N-Dicyclohexyl-2-benzo-1,3-thiazolesulfinic Acid Amide [153]. A 15 % aqueous sodium hypochlorite solution (150 g) is added dropwise over 10 minutes to a suspension of 48.3 g of *N,N*-dicyclohexyl-2-benzo-1,3-thiazolesulfenamide in 300 mL of methanol. The reaction mixture is boiled for 1 h and then cooled to 0 °C. The product is collected by filtration, washed, and dried; *mp* 135 –137 °C.

Additional references regarding the preparation of benzo-1,3-thiazolesulfinic acid amides in yields of ca. 90 % can be found in [154].

6. 1-Hydroxyalkanesulfinates

The sodium salt of a 1-hydroxyalkanesulfinic acid was first prepared by Bazlen in 1904 [155, 156].

6.1. Properties

Sodium hydroxymethanesulfinate [*149-44-0*] is commercially available as the dihydrate $HOCH_2SO_2Na \cdot 2\,H_2O$, which crystallizes as white needles, *mp* 63 °C. Its solubility in water at room temperature is 60 g/100 g. The salt is sparingly soluble in alcohol, and can be easily obtained pure from 60 % methanol. In moist air it decomposes slowly to products with noxious odors. The water of crystallization is lost upon heating to 120 °C, and then from 125 °C the compound undergoes a very exothermic decomposition to methanethiol, hydrogen sulfide, some formaldehyde, and sulfur dioxide, during which the temperature rises to 160 °C. The residue consists of sodium hydroxymethanesulfonate, sodium thiosulfate, and sodium sulfite. In solution the decomposition is pH and temperature dependent. Sodium hydroxymethanesulfinate reacts with sodium hydrogen sulfite to give sodium dithionite ($Na_2S_2O_4$) and sodium hydroxymethanesulfonate. The hydroxyl group is acidic and susceptible to salt formation.

The monozinc salt of hydroxymethanesulfinic acid $(HOCH_2SO_2)_2Zn$ is readily soluble in water, but zinc oxidomethanesulfinate $(OCH_2SO_2)Zn$ is only sparingly soluble. The former crystallizes from water at 80 °C in anhydrous form, at 60 °C as the trihydrate, and

at 20 °C as the tetrahydrate (rhombohedric leaflets). By contrast, the sparingly soluble dizinc salt precipitates from cold solution as a trihydrate and from warm solution as a monohydrate. The amorphous and crystalline calcium salts of hydroxymethanesulfinic acid and also zinc 1-oxidoethanesulfinate have been described as well. The latter is readily cleaved, and acts as a reducing agent at 50 °C. Derivatives of other aldehydes, such as benzaldehyde and butyraldehyde, also decompose readily, reducing vat dyes even at low temperature (see Chap. 8). The nature of sodium hydroxymethane sulfinate in aqueous solution is described in detail in [157].

6.2. Production

Sodium hydroxymethanesulfinate can be prepared by reacting sodium dithionite with formaldehyde in the presence of alkali, or by catalytic reduction of sodium hydroxymethanesulfonate with hydrogen. It also results from the zinc dust reduction of sodium hydroxymethanesulfonate, which is in turn prepared in situ from zinc oxide, sulfurous acid, and formaldehyde. A variant of this process involves the reduction of aqueous sulfur dioxide (sulfurous acid) with zinc dust to zinc dithionite, which is then converted with formaldehyde into the zinc salts of hydroxymethanesulfonic and hydroxymethanesulfinic acids. The sulfonic acid portion is reduced to the desired sulfinic acid salt in the presence of zinc and sodium hydroxide:

$$Zn + 2SO_2 \xrightarrow{H_2O} ZnS_2O_4$$

$$ZnS_2O_4 + 2H_2O + 4CH_2O \longrightarrow (HO-CH_2-SO_3)_2Zn + (HOCH_2-SO_2)_2Zn$$

$$(HOCH_2-SO_3)_2Zn + 2Zn \longrightarrow 2(-OCH_2-SO_2)Zn + ZnO + H_2O$$

$$(-OCH_2SO_2)Zn \xrightarrow{NaOH} HOCH_2-SO_2Na + ZnO$$

General Synthesis. Zinc dust is suspended in 5 times its volume of water and sulfur dioxide is passed in at 30 – 35 °C until all the zinc has dissolved. Aqueous formaldehyde (30 – 40 %) is then added at 50 – 60 °C until the solution no longer reduces indigo in the cold. After addition of the calculated quantity of zinc dust, the reaction mixture is heated to 90 °C, and this temperature is maintained until virtually no more zinc compounds dissolve. Sodium hydroxide is subsequently added and the mixture is filtered. The clarified solution of sodium hydroxymethanesulfinate is evaporated in vacuo until it acquires the consistency of honey; $HOCH_2SO_2Na \cdot H_2O$ solidifies within a few hours, after which it can be ground.

The synthesis also produces zinc hydroxymethanesulfinate in soluble form, together with the zinc salt of the sulfonic acid. The former can be separated by evaporation in vacuo and subsequent drying. Zinc oxidomethanesulfinate can be obtained from zinc hydroxymethanesulfinate by the addition of zinc salts and sodium hydroxide. Other salts, such as the calcium salt, are precipitated by treating sodium hydroxymethanesulfinate with the appropriate chloride (e.g., $CaCl_2$). Other aldehydes can be used in place of formaldehyde to synthesize alternative hydroxyalkanesulfinic acids. Many aldehydes have been shown to be effective [156]. Hydroxyalkanesulfinic acids are intermediates in the reduction of aldehydes and ketones with dithionite [158].

6.3. Quality Specifications, Analysis

The industrially important salts of hydroxymethanesulfinic acid decompose in moist air. Their packaging must, therefore, provide protection against air and moisture. The products present neither a fire hazard nor a tendency to self-ignite.

Sodium hydroxymethanesulfinate is commercially available in 98 – 100 % purity; sodium hydroxymethanesulfonate and sodium sulfite are present as impurities. The commercial zinc salt contains 88 – 90 % sulfinate based on anhydrous product. The main impurities in this case are sodium sulfate and zinc hydroxymethanesulfonate. The proportion of sulfinate in the calcium salt is 70 % based on anhydrous product, with the corresponding reactants present as impurities. Other salts of industrially produced alkanesulfinic acids (e.g., sodium hydroxy-1-ethanesulfinate) are avoided in pure form because of their odors.

The content of alkanehydroxysulfinate can be determined iodometrically in the absence of air.

The method is analogous to determination of the content of dithionite, in which titration is carried out with acidic indigo carmine solution. The required temperature for the titration depends on the nature of the aldehyde.

7. Formamidinesulfinic Acid

Formamidinesulfinic acid (**11**) [*1758-73-2*], $CH_4N_2O_2S$, is used in certain branches of the textile industry [159].

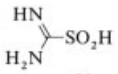

11

7.1. Properties

Formamidinesulfinic acid is a white crystalline compound, *mp* 128 °C (decomp.), in [159] *mp* 144 °C. It is sparingly soluble in water and other common solvents. The compound dissolves in alkaline solutions with decomposition, but it can be heated to 100 °C in concentrated sulfuric acid without decomposition. It reacts with formaldehyde and sodium hydroxide to form sodium hydroxymethanesulfinate, and with sodium hydrogensulfite to give sodium dithionite. Further information in the references given in [160]; for spectrochemical study see [161].

7.2. Production

Formamidinesulfinic acid was first obtained by Barnett, in 1910 by the oxidation of thiourea with hydrogen peroxide:

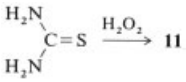

Since then a number of syntheses have been described in the literature, involving the use of hydrogen peroxide in a variety of solvents. Methanol: water (40 : 60) leads to formamidinesulfinic acid in 96.5 % yield and 98 % purity [162].

Formamidinesulfinic Acid (**11**) [162]. Thiourea (80 g) is dissolved in 800 g of methanol : water (40 : 60) at 20 °C and oxidized with 210 g of 35 % hydrogen peroxide at − 20 to − 10 °C over the course of 30 min. The reaction mixture is then stirred for a further 10 min. The product is obtained in 96.5 % yield and 98 % purity. For further information see [163].

Other synthetic procedures can be found in the literature, including oxidation of thiourea with chlorine dioxide [164], photochemical oxidation with singlet oxygen [165], and oxidation with hydrogen peroxide in water [166] or carbon tetrachloride [167]. Oxidation of substituted thioureas to the corresponding formamidinesulfinic acids has also been described [168].

Oxidimetric methods and titration with indigo in the presence of sodium hydrogencarbonate are suitable means for determining purity.

8. Industrial Uses of Sulfinic Acids and their Derivatives

Sulfinic acids and their derivatives are used in a wide range of industrial applications, serving as:

1) Catalysts for the polymerization of styrene, butadiene, and methacrylic acid esters at low temperature [10, 72], as well as for polymerization of acrylonitrile [169]
2) Catalysts for rapid and mild isomerization of disubstituted double bonds [170]
3) Surface-active agents for the pretreatment of raw cotton as a way to avoid bleaching [171]
4) Antistatic additives to facilitate dyeing of polyamides [172]
5) Agents to improve the dyeing of linear polyesters [173]
6) Additives for polyacrylonitrile solutions subject to spinning [174]
7) Plant-growth regulators [175], crop-protection agents [176, 177], and fungicides [178]
8) Fragrances and flavors for foods, perfumes, and tobacco [179, 180]
9) Activators for nitrogen- containing propellants [181] (e.g., azodicarbonamide [182])

10) Emulsifiers and dispersing agents for oil–water emulsions [183, 184], agents for floating ores [185]
11) Intermediates in the production of crease-resistant finishes [186]
12) Intermediates in the production of pharmaceuticals [187]
13) Antifogs in silver halide photography [188]
14) Agents for stimulating mucous membranes of the mouth and nose providing a cooling sensation similar to that of menthol [189]
15) Vulcanization accelerators [190, 191]
16) Reducing agents for dyes [192], as well as elsewhere in the dye industry [193]
17) Reducing agents for aldehydes and ketones [158, 194].

Several hydroxyalkanesulfinic acid salts are available commercially. For example, sodium hydroxymethanesulfinate is sold under the trade names Rongalite C [195] (BASF) and Brüggolit (Brüggemann, Germany), and as various Hydrosulfite formulations (Ciba-Geigy and Rohner). It is used mainly in direct and discharge printing for converting vat dyes into their leuco forms and as reducing catalyst for synthetic rubber thermoplastics and for bleaching polyesters. A redox potential corresponding to that of dithionite is attained only at 75–80 °C, so the reduction is carried out in a steam bath (→ Bleaching, Chap. 2.3.3.).

Zinc hydroxymethanesulfinate is sold under the trade name Decrolin (BASF) and as various Hydrosulfite preparations (Ciba-Geigy, Rohner, and Pechiney, France). The zinc salt is used in the wool industry and also as a bleaching agent for such materials as soaps, fatty acids, glycerol, sizing agents, and for corrosion color printing on polyamide, wool, silk etc.

Calcium hydroxymethanesulfinate is commercially available under the trade name Rongalite H (BASF). It is used for vat-dye printing on acetate, polyamide, silk, and wool fabrics. Such prints display exceptional stability toward storage.

Hydroxyalkanesulfinic acids derived from other aldehydes, such as acetaldehyde, benzaldehyde, or butyraldehyde, as well as products condensed with ammonia, are available under such trade names as Rongalite 2 PH–A and 2 PH–B_{fl} and Rongalite FD_{fl} (BASF). These vary in their redox potentials. For example, the redox potential of Rongalite 2 PH lies between those of Hydrosulfite and Rongalite C (sodium hydroxymethanesulfinate), and that of Rongalite FD is above Rongalite C. These products have proven useful in two-phase and fill printing as well as in the textile industry.

For the use of sulfinic acids as fungicides see [178], for other applications in the dye industry see [192, 193], and as reducing agents for aldehydes and ketones see [158, 194].

Formamidinesulfinic acid is sold under the trade names Reduktionsmittel F (Degussa) [196] GLO Lite TD (Glo-Tex Chemicals, United States), Arolite TD Concentrate (Arol Chemical Products Company, United States), Thio Urea Dioxide and tec light (Tohai Denka Kogya Co., Japan). Formamidinesulfinic acid can be used for the reduction of disulfides [197] and the hydrogenation of aldehydes [198], as a processing aid in noble-metal extraction [199], as a dye stabilizer for pyrazoline brighteners [200], as a bleaching agent in the paper industry [201], as an agent for improving silver halide photography [202], for anaerobic textile wastewater treatment [203], and for catalyzing the reactions of olefins with dibromodifluoromethane [204]. Further sulfinic acid salts for synthetic intermediates are available: 4-chlorobenzenesulfinic acid sodium salt (Rohner, Bayer) [205]; benzenesulfinic acid sodium salt (Bayer) [206], the analogous zinc salt (Rohner) [207]; 4-methylbenzenesulfinic acid sodium salt (Rohner) [208]; 4-nitrobenzenesulfinic acid sodium salt (Bayer) [209]; 4-chloro-3-nitrobenzenesulfinic acid sodium salt [210].

9. Toxicology

Published toxicological data regarding these compounds are sparse. The reported LD_{50} values are in the region of a few grams per kilogram. Benzenesulfinic acid is said to cause slight skin irritation [211, 112].

The LD_{50} value of sodium hydroxymethanesulfinate is ca. 6400 mg/kg (oral), that of zinc hydroxymethanesulfinate 1280 mg/kg oral, rat [213]. The former is slightly irritating to the mucous membranes, but not to the skin (Safety Data Sheet, Brüggolit C) [214]. The latter causes skin irritation and moderate mucous membrane irri-

tation [213]. Further LD_{50} values (oral, rat) for some sulfinic acid salts are listed below:

Formamidinesulfinic acid	1120 mg/kg skin irritant [196]
4-Chlorobenzenesulfinic acid sodium salt	> 5000 mg/kg [205]
Benzenesulfinic acid sodium salt	> 5000 mg/kg [206]
Benzenesulfinic acid zinc salt	> 5000 mg/kg [207]
4-Methylbenzenesulfinic acid sodium salt	3200 mg/kg [208]
4-Nitrobenzenesulfinic acid sodium salt	> 5000 mg/kg [209]
4-Chloro-3-nitrobenzenesulfinic acid sodium salt	> 5000 mg/kg [210]

In case of an emergency, the appropriate safety data sheets should be consulted for all these products.

10. References

1. E. Block, S. Naganathan, D. Putman, Shu Hai Zhao, *Pure Appl. Chem.* **65** (1993) 625. X. Yan, Zh. Wang, Ph. Barlow, *Food Chem.* **47** (1993) 289. H. Jansen, B. Müller, K. Knobloch, *Planta Med.* **53** (1987) 559.
2. F. Freeman, Xiao Bo Ma, R. J. San Lin, *Sulfur Lett.* **15** (1993) 253.
3. R. J. Apitz-Castro, M. K. Jain, US 4 665 088, 1987.
4. E. Block et al., *J. Am. Chem. Soc.* **108** (1993) 7045. S. F. Mohammad, St. C. Woodward, *Thromp. Res.* **44** (1986) 793.
5. H. Wagner, W. Dorsch, EP 299 424, 1989. E. Block et al., *J. Am. Chem. Soc.* **110** (1988) 7813.
6. K. K. Nippon Kokan, JP 02 204 487, 1990, *Chem. Abstr.* **114** (1991) 6523.
7. R. Gebhardt, *Lipides* **28** (1993) 613.
8. M. Tappaz, K. Almarghini, F. Legay, A. Remy, *Neurochem. Res.* **17** (1992) 849. A. J. Khodair, A. A. Swelin, A. A. Abdel-Wehab, *Phosphorus Sulfur* 2 (1976) 173.
9. N. Kharash, A. S. Arora, *Phosphorus and Sulfur* **2** (1976) 1 – 50.
10. W. G. Filby, K. Günther, R. D. Penzhorn, *J. Org. Chem.* **38** (1973) 4070. Gesellschaft für Kernforschung, DE-OS 2 322 199, 1973.
11. Fiat Final Rep. 1313 I (1948) 357.
12. M. Uchino, K. Suzuki, M. Sekiya, *Synthesis* **11** (1977) 794.
13. H. Reinheckel, D. Jahnke, *Chem. Ber.* **99** (1966) 1718. Dt. Akad. Wiss., DE-AS 1 248 656, 1965.
14. W. M. Ziegler, R. Connor, *J. Am. Chem. Soc.* **62** (1940) 2596.
15. E. Wellisch, E. Gipstein, O. J. Sweeting, *J. Org. Chem.* **27** (1962) 1810.
16. C. S. Marvel, R. S. Johnson, *J. Org. Chem.* **13** (1948) 822.
17. American Cyanamid, US 2 917 540, 1958. Du Pont, US 2 993 932, 1961; US 2 315 514, 1938.
18. M. T. Beachem et al., *J. Am. Chem. Soc.* **81** (1959) 5430.
19. U. Schöllkopf, P. Hilbert, *Justus Liebigs Ann. Chem.* 1973, 1061.
20. H. W. Roesky, *Angew. Chem.* **83** (1971) 890.
21. R. N. Haszeldine, J. M. Kidd, *J. Chem. Soc.* 1955, 2901.
22. J. B. Hendrickson, A. Giga, J. Wareing, *J. Am. Chem. Soc.* **96** (1974) 2275.
23. F. C. Whitmore, F. H. Hamilton, *Org. Synth. Coll.* **I** (1941) 492.
24. L. Field, F. A. Grunwald, *J. Org. Chem.* **16** (1951) 946.
25. E. Knoevenagel, J. Kenner, *Chem. Ber.* **41** (1908) 3315.
26. L. Gattermann, *Chem. Ber.* **32** (1899) 1136.
27. J. Thomas, *J. Chem. Soc.* **95** (1909) 342.
28. S. Smiles, C. M. Bere, *Org. Synth. Coll.* **1** (1948) 7.
29. W. E. Truce, W. J. Ray, *J. Am. Chem. Soc.* **81** (1959) 481.
30. P. Messinger, H. Greve, *Justus Liebigs Ann. Chem.* **1977,** 1457. Hoechst, DE-AS 1 129 477, 1960.
31. J. Boeseken, *Recl. Trav. Chim. Pays-Bas* **55** (1936) 1040. E. W. Tillitson, US 2 493 471, 1950.
32. BASF, DE 222 195, 1910.
33. H. Bredereck, G. Brod, G. Höschele, *Chem. Ber.* **88** (1955) 438.
34. S. Detoni, D. Hadzi, *J. Chem. Soc.* 1955, 3163.
35. H. Bredereck et al., *Angew. Chem.* **70** (1958) 268.
36. J. L. Kice, K. W. Bowers, *J. Am. Chem. Soc.* **84** (1962) 605.
37. J. L. Kice, N. E. Pawlowski, *J. Org. Chem.* **28** (1963) 1162.
38. *Houben Weyl,* 4th ed., **E 11, part 1,** pp. 64, 88, suppl.
39. J. Metzger, H. Plank, *Bull. Soc. Chim. Fr.* 1956, 1701.
40. W. E. Truce, A. Murphy, *Chem. Rev.* **48** (1951) 69.
41. C. J. Stirling, *Int. J. Sulfur Chem, B* **6** (1971) nr. 4, 277.
42. L. Field, *Synthesis* 1978, 713; 1972, 101.

43. *Houben-Weyl,* 4th. ed., **IX,** pp. 289, 303.
44. B. Lindberg, *Acta Chem. Scand.* **17** (1963) 377.
45. P. Allen, Jr., *J. Org. Chem.* **7** (1942) 23.
46. D. B. Hope, C. D. Morgan, M. Wälti, *J. Chem. Soc. C.* 1970, 270. W. E. Truce, D. P. Tate, D. N. Burdge, *J. Am. Chem. Soc.* **82** (1960) 2872.
47. L. Horner, H. Neumann, *Chem. Ber.* **98** (1965) 3462. Hoechst, DE-OS 1 493 664, 1964.
48. C. J. M. Stirling, *Int. J. Sulfur. Chem. Part. B* **6** (1971) 280. L. Field, *Synthesis* 1972, 120. L. Field, *Synthesis* 1978, 713. K. K. Anderson, *Compr. Org. Chem.* **3** (1979) 317. E. Wendschuh, K. Dölling, *Z. Chem.* **20** (1980) 122.
49. Continental Oil, BE 603 176, 1961. K. Ziegler, DE-AS 1 050 762, 1957.
50. W. E. Truce, J. F. Lyons, *J. Am. Chem. Soc.* **73** (1951) 126.
51. IG Farbenindustrie, GB 321 843, 1929.
52. J. R. Nool, P. C. van der Hoeven, W. P. Haslinghuis, *Tetrahedron Lett.* 1970, 2531.
53. F. S. Dainton, K. I. Irvin, *Trans. Faraday Soc.* **46** (1950) 374.
54. W. Huang, *Youje Huaxue* **12** (1992) 12; *Chem. Abstr.* **117** (1992) 191 273.
55. H. Berger, *Red. Trav. Chim. Pays-Bas* **82** (1963) 773.
56. I. Scheinfeld, J. C. Parham, S. Murphy, G. B. Brown, *J. Org. Chem.* **34** (1969) 2153.
57. N. Kharash, R. Swidler, *J. Org. Chem.* **19** (1954) 1704.
58. J. F. Bennet, R. E. Zahler, *Chem. Rev.* **49** (1951) 273. W. E. Truce, E. M. Kreider, W. W. Brand, *Org. React. (N.Y.)* **18** (1970) 100.
59. *Houben Weyl* **E 11, part 1,** p. 614, suppl.
60. L. Horner, H. Neumann, *Chem. Ber.* **98** (1965) 1715.
61. W. E. Truce, F. E. Roberts, Jr., *J. Org. Chem.* **28** (1963) 593.
62. A. J. Parker, N. Kharash, *Chem. Rev.* **59** (1959) 583. S. Smiles, D. T. Gibson, *J. Chem. Soc.* **125** (1924) 176.
63. D. T. Gibson, J. D. Loudon, *J. Chem. Soc.* 1937, 487.
64. G. Leandri, A. Tundo, *Ann. Chim. (Rome)* **44** (1954) 340.
65. G. W. Fenton, C. K. Ingold, *J. Chem. Soc.* 1928, 3127; 1929, 2338.
66. W. M. Ziegler, R. Connor, *J. Am. Chem. Soc.* **62** (1940) 2596.
67. W. E. Truce, F. E. Roberts, *J. Org. Chem.* **28** (1963) 593.
68. J. B. Hendrickson, A. Giga, J. Wareing, *J. Am. Chem. Soc.* **96** (1974) 2275.
69. M. Julia, B. Badet, *Bull. Soc. Chim. Fr.* 1976, 525.
70. R. C. Krug, J. A. Rigney, G. R. Tidulaar, *J. Org. Chem.* **27** (1962) 1305.
71. A. J. Titov, A. N. Baryshnikova, *Dokl. Akad. Nauk SSSR* **157** (1964) 139. *Chem. Abstr.* **61** (1964) 9396.
72. Bayer, DE 975 943, 1943.
73. G. W. Fenton, C. K. Ingold, *J. Chem. Soc.* 1929, 2338. J. Büchi, H. R. Füeg, A. Aebi, *Helv. Chim. Acta* **42** (1959) 1368.
74. C. M. Suter in: *Tetravalent Organic Sulfur Compounds,* J. Wiley, New York 1944, p. 660. C. M. Suter in: *The Organic Chemistry of Sulfur,* J. Wiley, New York 1948, p. 658.
75. G. E. Vennstra, B. Zwanenburg, *Synthesis* 1975, 519. L. Field, R. D. Clark, *Org. Synth. Coll.* **4** (1963) 674.
76. G. E. Vennstra, B. Zwanenburg, *Synthesis* 1975, 519.
77. M. Kobayashi, *Bull. Chem. Soc. Jpn.* **39** (1966) 1296.
78. F. Arndt, A. Scholz, *Justus Liebigs Ann. Chem.* **510** (1934) 62.
79. J. C. Meek, J. S. Fowler, *J. Org. Chem.* **33** (1968) 3422.
80. P. Allen, Jr., L. S. Karger, J. D. Haygood, J. Shrensel, *J. Org. Chem.* **16** (1951) 767.
81. I. B. Douglass, J. F. Ward, R. V. Norton, *J. Org. Chem.* **32** (1967) 324.
82. J. Tröger, O. Beck, *J. Prakt. Chem.* **87** (1913) 289.
83. C. C. J. Culvenor, W. Davies, N. S. Heath, *J. Chem. Soc.* 1949, 278.
84. T. L. Gresham et al., *J. Am. Chem. Soc.* **74** (1952) 1323.
85. O. Bayer, *Angew. Chem.* **61** (1949) 229.
86. J. E. Herweh, R. M. Fantazier, *J. Org. Chem.* **41** (1976) 116. H. Gilman, L. F. Cason, *J. Am. Chem. Soc.* **72** (1950) 3469.
87. W. B. Price, S. Smiles, *J. Chem. Soc.* 1928, 3154.
88. L. Field, P. H. Settlage, *J. Am. Chem. Soc.* **73** (1951) 5870.
89. H. Brederick, E. Bäder, *Chem. Ber.* **87** (1954) 129. E. Bäder, H. D. Hermann, *Chem. Ber.* **88** (1955) 41.
90. F. Ullmann, G. Pasdermadian, *Chem. Ber.* **34** (1901) 1150. C. W. Ferry, J. S. Buck, R. Baltzly, *Org. Synth.* **22** (1942) 31.
91. N. H. Nilson, C. Jacobsen, A. Senning, *Chem. Commun.* 1971, 314.
92. Hoechst, DE-AS 1 097 434, 1961. R. Kerber, J. Starnick, *Chem. Ber.* **104** (1971) 2035.

93. H. Gilman, L. F. Cason, *J. Am. Chem. Soc.* **72** (1950) 3469. IG Farbenindustrie DE 676 013, 1938.
94. L. F. Cason, C. C. Wanser, *J. Am. Chem. Soc.* **73** (1951) 142.
95. W. E. Truce, E. Wellisch, *J. Am. Chem. Soc.* **74** (1952) 5177. IG Farbenindustrie, DE 663 992, 1938.
96. H. Burton, E. Hoggarth, *J. Chem. Soc.* 1945, 468.
97. J. Walker, *J. Chem. Soc.* 1945, 630.
98. S. Pickholz, *J. Chem. Soc.* 1946, 685.
99. K. Bowden, E. A. Braude, E. R. H. Jones, *J. Chem. Soc.* 1946, 945. E. P. Kohler, G. R. Barrett, *J. Am. Chem. Soc.* **46** (1924) 747.
100. C. J. M. Stirling, *J. Chem. Soc.* 1964, 5856.
101. A. G. Davies, R. Feld, *J. Chem. Soc.* 1956, 665.
102. S. Atkin, *Anal. Chem.* **19** (1947) 816.
103. L. Horner, O. H. Basedow, *Justus Liebigs Ann. Chem.* **612** (1958) 108.
104. P. Allen, Jr., L. S. Karger, J. D. Haygood, J. Shrensel, *J. Org. Chem.* **16** (1951) 767. G. C. Denzer, Jr., P. Allen, Jr., P. Conway, J. M. van der Veen, *J. Org. Chem.* **31** (1966) 3418.
105. J. Strating, H. J. Backer, *Recl. Trav. Chim. Pays-Bas* **69** (1950) 638.
106. F. Kurzer, *Org. Synth.* **4** (1963) 937.
107. I. B. Douglass, R. V. Norton, *J. Org. Chem.* **33** (1968) 2104.
108. M. L. Kee, I. B. Douglass, *Org. Prep. Proced.* **2** (1970) no. 3, 235.
109. G. G. Venier, H.-H. Hsich, H. J. Barager (III), *J. Org. Chem.* **38** (1973) 17.
110. E. Wenschuh, *Z. Chem.* **24** (1984) 126.
111. J. F. King, S. Skanieczny, *Phosphorus Sulfur* **25** (1985) 11 – 20.
112. J. M. Shreeve, *Acc. Chem. Res.* **6** (1973) 387.
113. D. Sianesi, G. C. Bernadi, G. Moggi, *Tetrahedron Lett.* 1970, 1313.
114. K. R. Brown, J. B. Douglass, *J. Am. Chem. Soc.* **73** (1951) 57 87.
115. J. B. Douglass, K. R. Brown, F. T. Martin, *J. Am. Chem. Soc.* **74** (1952) 5770.
116. J. B. Douglass, D. R. Poole, *J. Org. Chem.* **22** (1957) 536. US 3 253 028, 1957/1962.*Chem. Abstr.* **65** (1966) 3751.
117. J. B. Douglass, B. S. Farah, E. G. Thomas, *J. Org. Chem.* **26** (1961) 1996.
118. J. B. Douglass, R. V. Norton, *J. Org. Chem.* **33** (1968) 2104.
119. Mi Lo Kee, J. B. Douglass, *Org. Prep. Proced.* **2** (1970) 235.
120. Bayer AG, DE 1 167 326, 1962. *Chem. Abstr.* **61** (1964) 1758.
121. H. Stetter, M. Krause, W.-D. Last, *Chem. Ber.* **102** (1969) 3357.
122. B. Zwanenburg, *Phosphorus, Sulfur and Silicon,* **43** (1989) 1 – 2, 1 – 24.
123. W. Sas, *J. Chem. Res. Synop.* 1993, 160.
124. Goodrich Co., US 4 032 505, 1977.
125. Goodrich Co., US 4 055 540, 1977.
126. I. B. Douglass, *J. Org. Chem.* **30** (1965) 633.
127. C. J. M. Stirling, *J. Chem. Soc.* 1963, 5741.
128. A. Heesing, M. Jaspers, I. Schwermann, *Chem. Ber.* **112** (1979) 2903.
129. M. Mikolajcyk, J. Drabowicz, *Synthesis* 1974, 124.
130. I. B. Douglass, *J. Org. Chem.* **39** (1974) 563.
131. L. Field, C. B. Hockel, J. M. Locke, J. E. Lawson, *J. Am. Chem. Soc.* **83** (1961) 1256.
132. L. Field, C. B. Hockel, J. M. Locke, *J. Am. Chem. Soc.* **84** (1962) 847.
133. P. Brownbridge, J. C. Jowett, *Synthesis* **3** (1988) 252.
134. R. St. Henion, *Org. Chem. Bull.* **41** (1969) 4.
135. J. S. Meek, J. S. Fowler, *J. Org. Chem.* **33** (1968) 3422. M. Kobayashi, H. Minato, H. Fukuda, *Bull. Chem. Soc. Jpn.* **46** (1973) 1266.
136. Y. Miyaji, H. Minato, M. Kobayashi, *Bull. Chem. Soc. Jpn.* **44** (1971) 862.
137. D. N. Harpp, T. G. Back, *J. Org. Chem.* **38** (1973) 4328.
138. A. J. Ruming, G. Kresze, *Phosphorus Sulfur* **29** (1986) 49.
139. D. Darwish, R. McLaren, *Tetrahedron Lett.* 1962, 1231. A. C. Cope, D. E. Morrison, L. Field, *J. Am. Chem. Soc.* **72** (1950) 59. C. J. M. Stirling, *Chem. Commun.* 1967, 131.
140. C. J. M. Stirling, *Chem. Commun.* 1967, 131. S. Braverman, *Int. J. Sulfur Chem. Part C* **6** (1971) 149.
141. K. K. Andersen et al., *J. Am. Chem. Soc.* **86** (1965) 5637.
142. R. M. Coates, H. D. Pigott, *Synthesis* 1975, 319.
143. M. Axelrod et al., *J. Am. Chem. Soc.* **90** (1968) 4835.
144. M. Koboyaski, A. Yamamoto, *Bull. Chem. Soc. Jpn.* **39** (1966) 2736.
145. A. J. Khodair, A. A. Swelin, A. A. Abdel – Wehab, *Phosphorus Sulfur* **2** (1976) 173.
146. K. Kondo, A. Negishi, I. Ojima, *J. Am. Chem. Soc.* **94** (1972) 5786.
147. F. Freeman, *Chem. Rev.* **84** (1984) 117 – 135.
148. I. B. Douglass, B. S. Farah, *J. Org. Chem.* **23** (1958) 805.
149. Y. H. Chiang, J. S. Luloff, E. Schipper, *J. Org. Chem.* **34** (1969) 2397.

150. Monsanto, FR 1 529 050, 1967; GB 1 139 349, 1966.
151. A. Sonn, E. Schmidt, *Chem. Ber.* **57** (1924) 1355.
152. D. Klamann, C. Sass, M. Zelenka, *Chem. Ber.* **92** (1959) 1910. Fuji Film Co., US 3 498 792, 1965.
153. Monsanto, US 3 541 060, 1967/1970.
154. Ugine Kuhlman, BE 738 304, 1968; FR 2 041 585, 1969.
155. M. Bazlen, BASF, US 855 566, 1904. M. Bazlen, *Chem. Ber.* **42** (1900) 4634.
156. M. Mulliez, C. Naudy, *Tetrahedron* **49** (1993) 2469.
157. J. S. Edgar, *Phosphorus Sulfur* **2** (1976) 181.
158. J. G. de Vries, R. M. Kellogg, *J. Org. Chem.* **45** (1980) 4126.
159. *Römpp,* 9th ed., Thieme-Verlag, 1990, p. 1426.
160. Mitsubishi Gas Chemical, EP 607 448, 1993. S. Iwasaki, T. Oka, JA 49 040 451, 1974. A. Kojima, M. Okawara, *Chem. Lett.* 1984, 2125. Bayer, DE 4 240 708, 1992. Bayer, EP 600 339, 1993.
161. D. De Filipo, G. Pouticelli, E. F. Tragu, A. Lai, *J. Chem. Soc., Perkin Trans.* **2** (1972) 1500.
162. Kuroda (Nihon) Chemicals KK, JP 50 062 934, 1975; *Chem. Abstr.* **83** (1975) 96 423.
163. M. Drifoglu, W. N. Marmer, R. L. Dudley, *Text. Res. J.* **62** (1992) 94. G. Wank, A. Mursyidi, *J. Photochem. Photobiol. A* **64** (1992) 263. G. Rabai, R. T. Wang, K. Kustin, *Int. J. Chem. Kinet.* **25** (1993) 53.
164. G. Rabai, R. T. Wang, K. Kustin, *Int. J. Chem. Kinet.* **25** (1993) 53.
165. G. Crank, A. Mursyidi, *J. Photochem. Photobiol. A* **64** (1992) 263.
166. Japan Oils and Fat Co., JP 45 017 665, 1970; *Chem. Abstr.* **73** (1970) 98 387.
167. E. Jourdan-Laforte, FR 2 040 797, 1971; *Chem. Abstr.* **75** (1971) 48 427.
168. I. I. Havel, R. Q. Kluttz, *Synth. Commun.* **4** (1974) 389.
169. Teijen Ltd, JP 47 025 464, 1972; *Chem. Abstr.* **78** (1973) 17 055.
170. T. W. Gibson, P. Strassburger, *J. Org. Chem.* **41** (1976) 791.
171. Agrotex, DL 129 567, 1977.
172. Phillips Petroleum, US 4 059 653, 1976.
173. Du Pont, GB 856 917, 1958.
174. Bayer, US 3 511 800, 1970.
175. Bayer, DE-OS 2 110 773, 1971.
176. R. Wegler: Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel, vol. 5, Springer, Heidelberg 1977, pp. 233, 503.
177. Phillips Petroleum, US 2 955 980, 1958.
178. M. Kling, US 5 270 058, 1993; *Chem. Abstr.* **120** (1994) 99 425.
179. Int. Flavors and Fragrances Inc., US 3 906 119, 1973.
180. Int. Flavors and Fragrances Inc., US 3 966 989, 1975.
181. Bayer, DE-OS 2 102 177, 1971.
182. Uniroyal Inc., EP 101 198, 1984; *Chem. Abstr.* **100** (1984) 211 426. Armstrong Co., US 4 104 301, 1978; *Chem. Abstr.* **90** (1979) 71 929.
183. Dt. Akad. Wiss., DE-AS 1 248 656, 1965.
184. Du Pont, US 2 315 514, 1938.
185. V. Petrovich, US 3 890 222, 1974.
186. American Cyanamid, US 2 917 540, 1958.
187. Hoechst, DE-AS 1 129 477, 1960.
188. Fuji Film Co., US 3 498 792, 1965.
189. Wilkinson Sword Ltd., BE 802 469, 1972.
190. Monsanto, US 3 532 693, 1967. Ugine Kuhlmann, FR 2 037 001, 1969.
191. Uniroyal Inc., FR 1 581 710, 1969; *Chem. Abstr.* **73** (1970) 16 057.
192. Sandoz Products, GB 1 591 616, 1981; *Chem. Abstr.* **96** (1982) 53 748.
193. Yu. U. Polenov et al., *Zh. Prikl. Khim. (Leningrad)* **63** (1990) 1622; *Chem. Abstr.* **114** (1991) 64 230.
194. S. K. Chung, *J. Org. Chem.* **46** (1981) 5457.
195. *Römpp,* 9th ed., Thieme Verlag, 1992, p. 3916.
196. Degussa Safety Data Sheet, 00 122/1 580 006 000/SDB/D001, 1994.
197. A. Sazasz et al., HU 55 763, 1991; *Chem. Abstr.* **116** (1992) 6562.
198. S. L. Huang, T.-Y. Chen, *J. Chin. Chem. Soc. (Taipei)* **21** (1974) 235; *Chem. Abstr.* **82** (1975) 139 916.
199. Degussa AG, DE 4 028 239, 1992; *Chem. Abstr.* **117** (1992) 94 231. Degussa AG, DE 4 028 240, 1992; *Chem. Abstr.* **117** (1992) 94 230. Sumitomo Metal Mining Co., JP 1 136 911, 1987; *Chem. Abstr.* **111** (1989) 185 579.
200. Ciba Geigy AG, EP 396 503, 1990; *Chem. Abstr.* **114** (1991) 188 036.
201. J. Maier, *Wochenbl. Papierfabr.* **12** (1993) 111;M. Hammann, WO 9 221 814, 1992; *Chem. Abstr.* **119** (1993) 74 932. Degussa AG, DE 3 923 728, 1991;*Chem. Abstr.* **114** (1991) 209 447. Interox International S.A., WO 9 207 139, 1992; *Chem. Abstr.* **117** (1992) 92 512. P. Pettit, *Apprita J.* **45** (1992) 385.
202. Konica Co., EP 552 650, 1993; *Chem. Abstr.* **120** (1994) 231 815. Y. Fu, S. Ji, B. Zhon, X. Ren, *Ganguang Kexue Yu Kuang Huaxue* **10**

(1992) 103; *Chem. Abstr.* **120** (1994) 177 952. Fuji Photo Film Co., JP 5 165 132, 1993; *Chem. Abstr.* **120** (1994) 204 455. Konishiroku Photo Ind., JP 5 224 335, 1993; *Chem. Abstr.* **120** (1994) 120 622.
203. D. L. Michelsen, M. Padaki, *Hazard. Ind. Wastes* **25** (1993) 218.
204. H. F. Wu, B. N. Huang, *Chim. Chem. Lett.* **4** (1993) 683; *Chem. Abstr.* **120** (1994) 162 932.
205. Bayer, Safety Data Sheet, 016 962/01, 1993.
206. Bayer, Safety Data Sheet, 018 604/05, 1993.
207. Bayer, Safety Data Sheet, 323 520/04, 1993.
208. Bayer, Safety Data Sheet, 378 945/02, 1994.
209. Bayer, Safety Data Sheet, 692 100/01, 1993.
210. Bayer, Safety Data Sheet, 321 668/02, 1993.
211. Bayer AG, Leverkusen, personal communication.
212. *Ullmann,* 3rd. ed., **16,** 490.
213. BASF, Ludwigshafen, Safety Data Sheets ET 00 338.
214. Brüggemann, Safety Data Sheet, Bruggolit C, SD 00/ 01 A DOC.

Sulfites, Thiosulfates, and Dithionites

José Jiménez Barberá, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chap. 1)
Adolf Metzger, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chap. 2)
Manfred Wolf, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 3)

1. Sulfites and Disulfites 1
1.1. General Properties 1
1.2. Ammonium Sulfite 2
1.3. Potassium Disulfite 2
1.4. Sodium Disulfite 2
1.5. Sodium Hydrogensulfite 2
1.6. Sodium Sulfite 3
1.7. Analysis of Alkali-Metal Sulfites . . . 3
1.8. Toxicology 3
2. Thiosulfates 3
2.1. General Properties 3
2.2. Sodium Thiosulfate 4
2.3. Ammonium Thiosulfate 5
2.4. Toxicology 7
3. Sodium Dithionite 7
3.1. Introduction 7
3.2. Properties 7
3.3. Production 8
3.4. Quality Specifications and Analysis . 9
3.5. Trade Names and Uses 10
3.6. Safety and Environmental Aspects; Storage and Transport 10
3.7. Economic Aspects 10
3.8. Toxicology 11
4. References 11

1. Sulfites and Disulfites

1.1. General Properties

Sulfurous acid is formed on passing sulfur dioxide into water:

$$SO_2 + H_2O \rightleftharpoons H_2SO_3$$

Since the equilibrium lies almost completely on the left-hand side and gas hydrates are formed on cooling the solution, isolation of the anhydrous acid is not possible. The acid dissociates in two steps:

$$H_2SO_3 \rightleftharpoons H^+ + HSO_3^- \quad K_1 = 1.6 \times 10^{-2}$$

$$HSO_3^- \rightleftharpoons H^+ + SO_3^{2-} \quad K_2 = 1.0 \times 10^{-7}$$

The dissociation constants K refer to 25 °C [1]. At higher concentrations the disulfite ion is formed:

$$2\,HSO_3^- \rightleftharpoons S_2O_5^{2-} + H_2O$$

The neutral alkali-metal sulfites M_2SO_3 and hydrogensulfites $MHSO_3$ are readily soluble in water. The latter cannot be isolated, since they are converted to disulfites on concentrating the solution.

Sulfurous acid and its salts are used as reducing agents. Aqueous solutions of halogens are reduced to hydrogen halides, iodates to iodides, and dichromates to chromium(III). However, sulfurous acid behaves as an oxidizing agent in the presence of stronger reducing agents; for example, with sulfur, thiosulfates are formed.

1.2. Ammonium Sulfite [2]

Solutions of ammonium sulfite [*10196-04-0*], $(NH_4)_2SO_3$, M_r 116.13, are produced by treating sulfur dioxide with aqueous ammonia. They are mainly used for the production of ammonium thiosulfate.

1.3. Potassium Disulfite [3]

The industrial production of potassium disulfite [*16731-55-8*], $K_2S_2O_5$, M_r 222.31, is similar to that of sodium disulfite. It is used for combatting fungi on vines and in wine production.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_477

1.4. Sodium Disulfite [4]

Sodium disulfite [*7681-57-4*], $Na_2S_2O_5$, M_r 190.1, dissolves readily in water with the formation of sodium hydrogensulfite $NaHSO_3$ (see Fig. 1) [5].

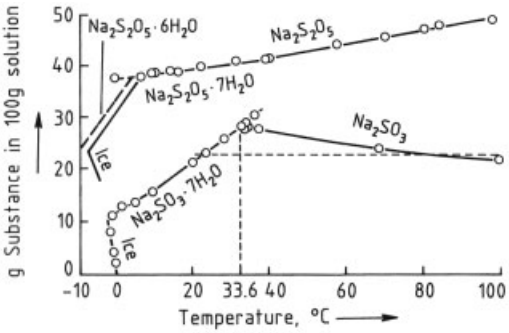

Figure 1. Solubility of Na_2SO_3 and $Na_2S_2O_5$ in water as a function of temperature

The fine crystalline salt can be stored without any problem if kept cool, dry, and with the exclusion of air. When moist it is oxidized by atmospheric oxygen to the sulfate.

Sodium disulfite is produced by treating 50–70 % sodium hydroxide or a suspension of sodium carbonate with sulfur-dioxide-containing gases in saturated sodium hydrogensulfite solution (Fig. 2) [6–9]. The sodium disulfite is removed by centrifugation and then dried. The saturated mother liquor from the centrifuge is mixed with the liquor from the waste-gas scrubber and adjusted to the usual commercial concentration.

Sodium disulfite production is coupled with the formation of sodium hydrogensulfite solution, for which, therefore, a demand must also exist. Production processes that use an anhydrous route, involving treatment of sodium sulfite or sodium carbonate with sulfur dioxide, have not achieved industrial importance.

Equipment and storage containers are predominantly made from 1.4571 grade stainless steel. Fiberglass-reinforced plastic containers are also suitable for storage. Paper or plastic sacks, steel barrels, or big bags made of synthetic fibers are used for transportation.

The commercial product has a content of ca. 97 % $Na_2S_2O_5$, 1 – 2 % Na_2SO_3, and ca. 1 % Na_2SO_4.

Sodium disulfite is used in the photography and fiber industries, for preserving foodstuffs, for wastewater treatment, and in the paper, textile, and leather industries.

1.5. Sodium Hydrogensulfite [4]

Sodium hydrogensulfite [*7631-90-5*], $NaHSO_3$, M_r 104.06, is not known in the solid state. On concentrating an aqueous solution, solid sodium disulfite crystallizes. Sodium hydrogensulfite is oxidized in air. Solutions of high purity can be produced by dissolving sodium disulfite in water. Commercial solutions are produced as the mother liquor in the production of sodium disulfite or by treating a suspension of sodium hydroxide, or sodium sulfite or sodium carbonate with sulfur dioxide until a pH of 3.5 – 4 is reached. The equipment is mainly made of 1.4571 grade stainless steel. For storing the solution, fiberglass-reinforced plastic containers are also suitable.

1.6. Sodium Sulfite

Sodium sulfite [*7757-83-7*], Na_2SO_3, M_r 126.04, is predominantly marketed as the anhydrous salt. It is produced by treating a suspension of sodium hydroxide or sodium carbonate with sulfur dioxide. The temperature is maintained at 60 – 80 °C for optimal crystal separation. The sodium sulfite suspension formed is centrifuged and the salt dried. Some sodium sulfite is obtained from scrubbing of SO_2- containing off-gases. Production processes using an anhydrous route have not yet achieved industrial importance.

Sodium sulfite is sold in various purity grades containing 90 – 98 % Na_2SO_3. The salt can be stored for long periods if kept dry, with the exclusion of air. It is transported in plastic and paper sacks. Sodium sulfite is used as a reducing agent, for the production of sodium thiosulfate, in the photography, paper, and textile industries, for treating boiler water, and as a food preservative.

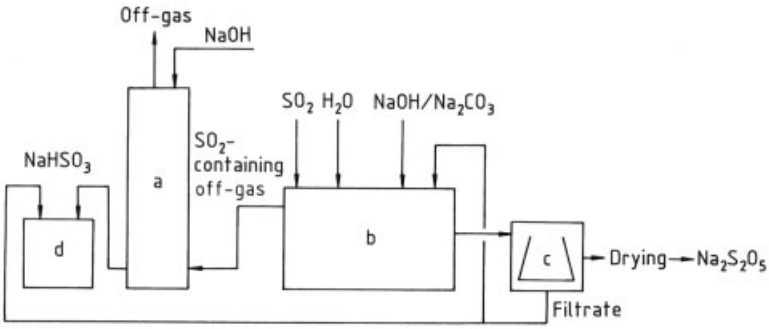

Figure 2. Production of sodium disulfite and sodium hydrogensulfite solution
a) Scrubber; b) Production plant; c) Centrifuge; d) Storage vessel for sodium hydrogensulfite solution

1.7. Analysis of Alkali-Metal Sulfites

The total sulfur dioxide content is titrated iodometrically. Sulfate is determined, after acidification of the solution with HCl and driving off the SO_2 with CO_2, either titrimetrically or as barium sulfate. To determine sulfite in sodium hydrogensulfite or sodium disulfite, the sample solution is treated with formalin and the NaOH liberated is titrated in the presence of phenolphthalein.

1.8. Toxicology

Oral administration of a few grams of sodium sulfite causes dizziness, vomiting, and diarrhoea [10]. Small doses of sodium hydrogensulfite are oxidized to sulfate in vivo, whereas larger doses cause considerable irritation of the gastrointestinal tract.

2. Thiosulfates

2.1. General Properties

Thiosulfates are the salts of thiosulfuric acid [*14921-76-7*], $H_2S_2O_3$. Whereas thiosulfuric acid is highly unstable, thiosulfates are quite stable, even in aqueous solution.

Most thiosulfates are readily soluble in water, with the exception of the moderately soluble barium salt and the sparingly soluble silver and lead salts. The soluble salts are readily crystallized. Acid salts do not exist.

The following properties are characteristic of thiosulfates: decomposition to sulfur and sulfur dioxide in acid solution:

$$S_2O_3^{2-} + 2\,H^+ \longrightarrow S + SO_2 + H_2O$$

decomposition in alkaline solution according to the equation:

$$S_2O_3^{2-} + 2\,OH^- \longrightarrow SO_4^{2-} + S^{2-} + H_2O$$

oxidation by weak oxidizing agents (I_2, Fe^{3+}, and Cu^{2+}) to give tetrathionate:

$$2\,S_2O_3^{2-} + I_2 \longrightarrow S_4O_6^{2-} + 2\,I^-$$

and oxidation by strong oxidizing agents (Cl_2, Br_2, MnO_4^-, and $Cr_2O_7^{2-}$) to give sulfate:

$$S_2O_3^{2-} + 4\,Cl_2 + 5\,H_2O \longrightarrow SO_4^{2-} + H_2SO_4 + 8\,HCl$$

The use of thiosulfates as antichlorines for removal of chlorine (e.g., from bleached materials) is based on the last-mentioned reaction. With heavy metals (e.g., silver) the thiosulfate ion forms soluble complexes $[Ag(S_2O_3)_2]^{3-}$. This property is used in photography for fixing.

Sodium and ammonium thiosulfates are important industrial chemicals.

2.2. Sodium Thiosulfate

Sodium thiosulfate is produced and sold as the anhydrous salt, and as the pentahydrate $Na_2S_2O_3 \cdot 5\,H_2O$. The hydrated product was formerly known as antichlorine, because of the earlier wide use for removal of residual chlorine.

Properties. Sodium thiosulfate [*7772-98-7*] crystallizes from aqueous solution as short, colorless prisms or elongated crystals. It effloresces in dry air, particularly at temperatures above ca. 33 °C. At 48 °C the crystals dissolve in their own water of crystallization. The solution tends to become supersaturated and can be kept in this state for weeks at room temperature. Spontaneous crystallization only occurs on supercooling to − 30 °C or below. On heating, preferably with stirring and under reduced pressure, the anhydrous salt is obtained as a heavy, white, crystalline powder. Anhydrous $Na_2S_2O_3$ is also produced by dehydration of $Na_2S_2O_3 \cdot 5\,H_2O$ in a stream of dry air at temperatures of up to 60 °C. At 223 °C sodium thiosulfate decomposes, predominantly giving sodium sulfate and pentasulfide:

$$4\,Na_2S_2O_3 \longrightarrow 3\,Na_2SO_4 + Na_2S_5$$

The pentahydrate, which exists in two modifications, is stable below 48 °C. Four modifications of anhydrous $Na_2S_2O_3$ are known [11].

Aqueous solutions of sodium thiosulfate are of neutral pH. They can be kept for months in the dark with the exclusion of air, but decompose on prolonged boiling in air. On addition of acid sulfur precipitates and sulfur dioxide is liberated. The densities of aqueous $Na_2S_2O_3$ solutions at 20 °C are listed below:

$[Na_2S_2O_3]$, wt %	ϱ, g/cm^3
1	1.0065
2	1.0148
4	1.0315
8	1.0654
10	1.0827
12	1.1003
16	1.1365
20	1.1740
24	1.2128
28	1.2532
30	1.2739
35	1.3273
40	1.3827

Table 1 lists the solubility in water; other properties are given in Table 2 and [12].

Production of sodium thiosulfate is based on the following reactions:

$$Na_2SO_3 + S \longrightarrow Na_2S_2O_3$$

$$2\,NaOH + SO_2 + S \longrightarrow Na_2S_2O_3 + H_2O$$

The raw materials are usually used in a technically pure form and consist of sulfur (solid or liquid), sodium sulfite, sodium hydrogensulfite, sodium hydroxide, and sulfur dioxide. The latter is used either as a pure gas or as a sulfur combustion gas containing ca. 18 % SO_2.

The addition of sulfur to sodium sulfite is carried out in stirred vessels. The reaction rate depends on the temperature, the excess of sulfur, and the intensity of stirring. The reaction lasts for several hours.

Polysulfide solutions are also produced to some extent as intermediates and react with disulfite to give thiosulfate [13]. Processes have also been described involving the production of sodium thiosulfate by oxidation of aqueous sodium sulfide solutions in the presence of catalysts [14–16] .

The sodium thiosulfate solutions are filtered hot, sometimes with the addition of decolorizing charcoal. The excess sulfur is recycled. The $Na_2S_2O_3 \cdot 5\,H_2O$ is crystallized by cooling the filtrate. The mother liquor still contains ca. 5 % $Na_2S_2O_3$ at 20 °C. Further sodium thiosulfate can be obtained from this mother liquor by evaporation. The crystalline salt is separated by centrifugation. The pentahydrate must be dried below ca. 40 °C because at higher temperatures water of crystallization is lost. Pneumatic-conveyor or fluidized-bed driers are used.

Anhydrous sodium thiosulfate is produced by dehydration of the pentahydrate at 60 –105 °C at atmospheric or reduced pressure. Another possibility is crystallization from sodium thiosulfate solution above 70 °C.

The standard commercial particle sizes are obtained by using continuous or batch crystallization equipment.

Quality Specifications and Analysis. For photographic purposes the degree of purity laid down in the ASA and DIN/ISO standards is usually adequate (99 % for $Na_2S_2O_3 \cdot 5\,H_2O$ and 97 % for anhydrous $Na_2S_2O_3$). The content of sodium thiosulfate is determined potentiometrically by titration of the diluted solution with iodine solution [17–19].

Both types of sodium thiosulfate are produced in Germany at Chemiewerk Bad Köstritz and anhydrous $Na_2S_2O_3$ is produced by Th. Goldschmidt AG, Werk Mannheim; Brotherton, England; and Foret, Spain, for example.

Table 1. Solubility of sodium thiosulfate in water

Solid phase	Ice		$Na_2S_2O_3 \cdot 5H_2O$			$Na_2S_2O_3 \cdot 2H_2O$			$Na_2S_2O_3 \cdot 0.5H_2O$			$Na_2S_2O_3$	
T, °C	−3.9	−11	0	20	40	48	50	60	65	70	75	80	100
wt %	15.0	30.0	34.4	41.2	50.6	62.0	62.9	67.4	69.5	70.0	70.8	71.0	72.7

Table 2. Properties of sodium thiosulfate

	$Na_2S_2O_3 \cdot 5H_2O$	$Na_2S_2O_3$
M_r	248.17	158.10
Crystal system	monoclinic	monoclinic
ϱ at 25°C, g/cm^3	1.749	2.267
Specific heat capacity at 0° C, J kg^{-1} K^{-1}	1 453	924
Heat of fusion, kJ/kg	148.1	
Heat of solution at 25°C, kJ/kg	190.6	
Heat of hydration at 18°C, kJ/mol	55.7	
Enthalpy of formation, kJ/mol	2 601	

Uses. Anhydrous sodium thiosulfate is now only added in small quantities to ammonium thiosulfate (see page 6). The hydrated salt is used as an antichlorine in bleaching, in wastewater purification, for reduction of dichromate in chromed leather production, and as a solvent for silver chloride in the chloride roasting of silver-containing minerals.

2.3. Ammonium Thiosulfate

Properties. Ammonium thiosulfate [*7783-18-8*], $(NH_4)_2S_2O_3$, crystallizes in an anhydrous form as shiny plates or sword-shaped crystals. No hydrates are known. On heating to 150 °C it sublimes with partial decomposition into $(NH_4)_2SO_3$, S, NH_3, and H_2S. Ammonium thiosulfate is hygroscopic; the stability limit is at 64 % relative humidity. It is readily soluble in water, sparingly soluble in acetone, and insoluble in alcohol. The solubility in water is lowered by the presence of NH_3. The solubilities are listed in Table 3. The densities of aqueous solutions at 20 °C are listed below:

$[(NH_4)_2S_2O_3]$, wt %	ϱ, g/cm^3
56	1.303
57	1.310
58	1.314
59	1.319
60	1.324
61	1.329
62	1.334

Table 3. Solubility of ammonium thiosulfate

T, °C	Solubility	
	in water	in NH_3-saturated water
0	60.0	
10	61.6	
20	63.1	25.8
30	64.7	38.1
40	66.3	46.5
50	67.9	58.2
60	69.4	65.9
70	71.0	70.7
80	72.6	72.4
90	74.2	

The data are for technical grades, which contain ca. 0.2 % NH_3 and small quantities of sulfate and sulfite. Aqueous ammonium thiosulfate solutions decompose slowly on heating above 40 °C with formation of sulfur and sulfate. A small quantity of sulfite stabilizes the solutions considerably. Some other properties are as follows:

100 % $(NH_4)_2S_2O_3$	
M_r	148.20
Crystal system	monoclinic
ϱ at 25 °C, g/cm^3	1.71
Heat of solution, water, 25 °C, kJ/kg	105
58–60 % $(NH_4)_2S_2O_3$ (aq.)	
Thermal conductivity, kJ m^{-1} h^{-1} K^{-1}	1.84
Specific heat capacity, kJ kg^{-1} K^{-1}	2.45
Vapor pressure, kPa	
20 °C	1.7
40 °C	5.4
60 °C	14.2

Production of ammonium thiosulfate is based on the reaction of sulfur, or occasionally ammonium sulfide and polysulfides [20], with ammonium sulfite in an aqueous ammoniacal medium. The ammonium sulfite is produced from ammonia or ammonia – water and SO_2 in the same or in a separate process step:

$$2\,NH_3 + SO_2 + H_2O \longrightarrow (NH_4)_2SO_3$$

$$(NH_4)_2SO_3 + S \longrightarrow (NH_4)_2S_2O_3$$

The sulfur dioxide is used neat (gaseous or liquid) or as sulfur combustion gas, and sulfur is used in liquid or solid form.

The rate of reaction of sulfur with ammonium sulfite depends on the temperature, the intensity of stirring, and the excess of sulfur. At 110 °C the reaction is complete after 1 – 2 h. Increased pressure with higher reaction temperatures and shorter reaction times can also be used [21]. A one-step process is also possible. Solutions containing ca. 60 % ammonium thiosulfate are produced to be sold as such, and 70 – 72 % solutions for production of the solid salt. The latter are filtered hot, sometimes with the addition of charcoal. The 60 % solutions can be filtered hot or after cooling.

To obtain the salt the filtered solution is cooled to ca. 20 – 30 °C. The crystallized salt is separated by using sieve-screw centrifuges. Drying must be carried out carefully; a product temperature of 50 °C must not be exceeded. The ca. 60 % mother liquor is recycled.

Crystalline ammonium thiosulfate tends to cake. It can be rendered free-flowing by the addition of 5 – 10 % anhydrous sodium thiosulfate (ammonium thiosulfate 90/10). The stability to storage is increased by the addition of ca. 1 – 2 % Na_2SO_3. The stability to storage of ammonium thiosulfate solutions can be increased by adding $(NH_4)_2SO_3$ or Na_2SO_3.

Ammonium thiosulfate can also be produced by the reaction of ammonium polysulfide solutions with gaseous SO_2 [22], [23]. In a process developed by the company Foret $(NH_4)_2S_2O_3$ is obtained from H_2S, NH_3, and water at ≤ 45 °C in the first step and subsequent reaction with SO_2 [23], [24].

Quality Specifications and Analysis. Commercially the degrees of purity and analysis methods laid down in the ASA, DIN, and ISO standards are valid [17–19] (see Table 4). In some cases particle size and particle-size distribution are also specified.

Table 4. Specifications for photo-quality ammonium thiosulfate

Analysis	Solid (ASA)	Aqueous solution (DIN 19 080 T82, ISO 3619)
$(NH_4)_2S_2O_3$, wt %	> 97	57 – 61
Sulfite (as SO_3), wt %	≤ 1.4	≤ 0.7
Sulfite [as $(NH_4)_2SO_3$], wt %		max. 1
H_2O, wt %	≤ 0.5	–
$(NH_4)_2S$, ppm	≤ 10	≤ 5
Insolubles, wt %	≤ 4	≤ 0.2
Combustion residue, wt %	≤ 2	≤ 0.2
Heavy metals (as Pb), ppm	≤ 20	≤ 10
Iron, ppm	≤ 50	≤ 10

Titration with iodine solution gives the combined thiosulfate and sulfite content; titration in the presence of formaldehyde gives the thiosulfate content.

Uses. Ammonium thiosulfate is used almost exclusively as a photographic fixing salt. Its advantage compared with sodium thiosulfate lies in the shorter fixing times, the increase in the efficiency of the fixing baths by ca. 50 %, and shorter washing periods. The recovery of the silver from spent fixing baths is easier. In Western Europe almost only ammonium thiosulfate is now used for fixing. By far the greatest proportion of the demand is covered by solutions.

The following are commercially available as solids:

Ammonium thiosulfate	98/100
Ammonium thiosulfate	90/10
Rapid fixing salt U3	70 % $(NH_4)_2S_2O_3$
Rapid fixing salt U3	20 % $Na_2S_2O_5$
Rapid fixing salt U3	10 % $Na_2S_2O_3$

Some producers of ammonium thiosulfate in Europe are: Th. Goldschmidt; Chemiewerk Bad Köstritz, Germany; Foret, Spain; W. Blythe & Co., United Kingdom.

2.4. Toxicology

For *sodium thiosulfate*, the LDLo for subcutaneous application in rabbits is 4000 mg/kg, and in frogs 6000 mg/kg [25].

For $Na_2S_2O_3 \cdot 5\,H_2O$, the oral TDLo in humans is 300 mg/kg (oral intake during 7 days). For dogs, the intravenous LDL_0 is 3000 mg/kg [25].

According to [26], $Na_2S_2O_3 \cdot 5\,H_2O$ has low oral toxicity. Larger doses (up to 12 g daily) have a severe laxative effect.

The following values have been determined for *ammonium thiosulfate* [25]: LD_{50} (oral, rat) 2890 mg/kg, LD_{50} (oral, guinea pig) 1098 mg/kg.

3. Sodium Dithionite [27]

3.1. Introduction

Sodium dithionite [*7775-14-6*], $Na_2S_2O_4$, is the only industrially important salt of dithionous acid ($H_2S_2O_4$), which has not been isolated. The importance of sodium dithionite lies in its powerful reducing capacity, which allows, for example, vat dyes to be reduced at room temperature. It is also used as a bleaching agent, mainly in the textile and paper industries.

STAHL first prepared dithionite unwittingly in 1718 when he treated iron with aqueous sulfur dioxide and obtained a yellow solution. In 1789 BERTHOLET showed that no hydrogen was produced in this reaction. In 1852 SCHÖNBEIN used this solution for the reduction of indigo. SCHÜTZENBERGER isolated dithionite as the dihydrate and gave it the name hydrosulfite. In 1881 BERNTHSEN established that hydrosulfite did not contain any hydrogen, but corresponded to the formula $Na_2S_2O_4$. In 1905 BAZLEN introduced the zinc-dust process for the production of anhydrous sodium dithionite. He also suggested using sodium amalgam as reducing agent.

3.2. Properties

Sodium dithionite is known as the dihydrate $Na_2S_2O_4 \cdot 2\,H_2O$, M_r 210.146, and as the anhydrous salt, M_r 174.114. The dihydrate crystallizes in thin, yellowish shiny, soft prisms of density 1.58 g/cm^3. The anhydrous salt forms monoclinic white crystals of density 2.38 g/cm^3. The solubility in water of both forms is shown in Figure 3. Solutions with anhydrous undissolved solute are labile below 72 °C. The solubility and the position of the transformation point are strongly influenced by foreign salts, particularly sodium salts, and water-miscible organic solvents.

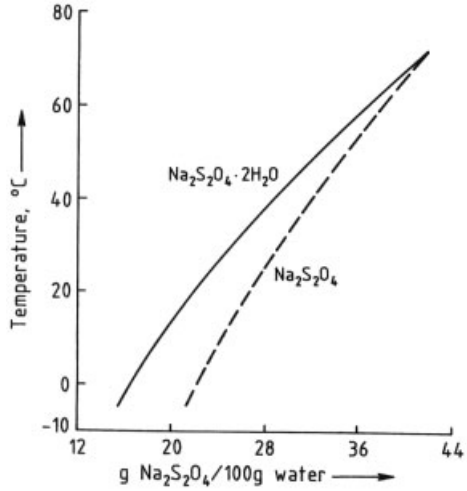

Figure 3. Solubility of sodium dithionite in water

Sodium dithionite dihydrate is very sensitive toward atmospheric oxygen in the finely crystalline state. The heat of oxidation can lead to ignition. The anhydrous salt decomposes exothermically in air on prolonged heating above 90 °C. The main decomposition products are sodium sulfate and sulfur dioxide. Above ca. 150 °C, with exclusion of air, sodium dithionite decomposes in a vigorous reaction, giving mainly sodium sulfite, sodium thiosulfate, sulfur dioxide, and a small amount of sulfur. In the absence of air, moisture only causes a small degree of decomposition. Sodium dithionite in powder form can decompose in air on contact with a small amount of water with such intense heat formation that it burns with a flame.

Aqueous dithionite solutions decompose slowly in the cold and rapidly in the warm, whereby hydrolysis to hydrogensulfoxylate and hydrogensulfite occurs first. The hydrolysis products can be trapped as sodium 1-hydroxyalkane-sulfinates and 1-hydroxyalkanesulfonates by addition of aldehydes or ketones:

$$NaO_2S - SO_2Na + H_2O \longrightarrow NaHSO_2 + NaHSO_3$$

$$NaHSO_2 + CH_2O \longrightarrow HOCH_2SO_2Na$$

$$NaHSO_3 + CH_2O \longrightarrow HOCH_2SO_3Na$$

In weakly acidic solution dithionite decomposes rapidly, especially under warm conditions, to give the thiosulfate and disulfite or hydrogensulfite, and in alkaline solution to give the sulfide and sulfite:

$$2\,Na_2S_2O_4 \xrightarrow{H^+} Na_2S_2O_3 + Na_2S_2O_5$$

$$3\,Na_2S_2O_4 + 6\,NaOH \longrightarrow Na_2S + 5\,Na_2SO_3 + 3\,H_2O$$

The decomposition in alkaline solution is accelerated by thiosulfates and polysulfides. On addition of strong acids the dithionite solution first becomes yellow-red, and after a short time complete decomposition occurs with precipitation of sulfur. The dithionite can be recovered if the solution is rapidly neutralized before the sulfur precipitates. Weak alkalis (pH 8 – 13) stabilize dithionite solutions, which can then be kept for weeks below 10 °C with the exclusion of air. In the presence of air the dissolved dithionite is converted rapidly into sulfate and sulfite at room temperature, with or without stabilizer.

The industrial applications of sodium dithionite are determined by its strong reducing properties, due to the S – S bond and the oxidation state of the sulfur in the compound.

With relatively strong oxidizing agents, such as potassium permanganate and iodine, sulfate is formed; with weaker ones, such as vat dyes, sulfite is formed. With MnO_2, dithionate can be formed.

3.3. Production

All processes for the production of dithionite start with the reduction of sulfurous acid, which can either be present in the free form or as hydrogensulfite. The production processes with zinc dust, sodium amalgam, sodium formate, sodium borohydride, and electric current as the reducing agent are important industrially.

Zinc-Dust Process. Some important producers still use the zinc-dust process, which was developed by BASF. The basic reactions are:

$$Zn + 2\,SO_2 \longrightarrow ZnS_2O_4$$

$$ZnS_2O_4 + 2\,NaOH \longrightarrow Zn(OH)_2 + Na_2S_2O_4$$

An aqueous slurry of zinc dust is treated in a stirred reactor with cooling at ca. 40 °C with liquid or gaseous sulfur dioxide to give zinc dithionite. After completion of the reaction the solution is passed through a filter press to remove unreacted zinc dust and impurities from the zinc.

The zinc is then precipitated from the zinc dithionite by adding sodium carbonate or sodium hydroxide in stirred vessels. The zinc carbonate or hydroxide is removed in filter presses. Anhydrous sodium dithionite is precipitated from the clarified sodium dithionite solution by concentration under vacuum and addition of sodium chloride at > 60 °C. It is filtered, washed with methanol, and dried at 90 – 100 °C.

Besides the evaporation process the salting out process, which was more widely used previously, is still known. In this process the dithionite is obtained from the solution by the addition of sodium chloride and methanol.

The zinc carbonate or hydroxide formed as side products can be further processed to give zinc salts or zinc oxide.

Amalgam Process. In the amalgam process, sodium hydrogensulfite is reduced to sodium dithionite in aqueous solution in a cooled, stirred vessel using the sodium amalgam of a chloralkali electrolysis cell. The sodium-free mercury is returned to the electrolysis cell where it is recharged with sodium. During reaction of the amalgam with the hydrogensulfite solution a pH of 5 – 6 must be maintained. The product is obtained by precipitation with salts or methanol or both.

Formate Process [28]. Sodium formate, dissolved in 80 % aqueous methanol, is charged to a stirred vessel. At a pressure of 2 – 3 bar sulfur dioxide and sodium hydroxide are introduced into this solution such that a pH of 4 – 5 is maintained. The reaction can be described by the following equation:

$$HCOONa + 2\,SO_2 + NaOH \longrightarrow Na_2S_2O_4 + CO_2 + H_2O$$

Under the above conditions anhydrous sodium dithionite precipitates as fine crystals. It is filtered, washed with methanol, and dried.

Sodium Borohydride Process. Sodium borohydride is stable in strong aqueous alkali and can be used in this form for the production of sodium dithionite by adding SO_2 and sodium hydroxide:

$$NaBH_4 + 8\,NaOH + 8\,SO_2 \longrightarrow 4\,Na_2S_2O_4 + NaBO_2 + 6\,H_2O$$

Electrolytic Process [29]. The electrolytic process, developed by BASF and by Olin, is used in only one small plant.

In a bipolar electrolysis cell, with the cathode and anode spaces separated by a semipermeable membrane, disulfite ions are reduced to dithionite at the cathode at pH 5 – 6. At the anode chloride ions, for example, are oxidized to chlorine. For charge equalization sodium ions pass through the membrane from the anode space to the cathode space. This process gives a dithionite solution which can be further processed as in the amalgam process.

3.4. Quality Specifications and Analysis [30]

Commercial sodium dithionite generally has a purity of ca. 88 %. It contains ca. 3 % of each of the following: sodium disulfite, sodium sulfite, sodium sulfate, and sodium carbonate. The latter stabilizes the $Na_2S_2O_4$. The total heavy metal content is generally < 20 ppm. The product from the zinc-dust process has a zinc content of up to 300 ppm.

The sodium dithionite content can most readily be determined iodometrically. It is dissolved in a neutral formaldehyde solution in a standard flask. The dithionite, which is sensitive to oxidation, reacts immediately on dissolving to give the more stable sodium hydroxymethanesulfinate, which can be titrated.

$$Na_2S_2O_4 + 2\,CH_2O + H_2O \longrightarrow HOCH_2SO_2Na + HOCH_2SO_3Na$$

$$HOCH_2SO_2Na + 2\,I_2 + 2\,H_2O \longrightarrow NaHSO_4 + 4\,HI + CH_2O$$

However, this method of analysis does not differentiate between dithionite and any thiosulfate which may be present. Since no acid is formed in the iodometric oxidation of thiosulfate, but is formed in the dithionite titration, the proportion of dithionite can be determined by subsequent titration of the solution with alkali.

3.5. Trade Names and Uses

Sodium dithionite is sold as the 88 % product, for example, under the names Hydrosulfit (BASF, Brüggemann, Germany; Prayon, Belgium; Hoechst-Celanese, USA; Hoechst-UK, UK, etc.), Blankit (BASF), Albite A (Idrosol, Italy), and Konite (Akzo, Holland). Hydrosulfit mixtures are sold, for example, under the names Blankit (with index, BASF), Blancolen (Brüggemann, Germany), Albite LA (Idrosol, Italy), Konite TP (Akzo, Holland), Zepar (Du Pont), Burmol (BASF), Arostit (Sandoz).

All uses of sodium dithionite are based on its reducing properties. It is used predominantly in the textile industry as a dyeing and printing auxiliary and as a bleaching agent in the textile and paper industries.

In dyeing and printing, sodium dithionite is used to convert insoluble vat dyes to the soluble leuco form. High-purity sodium dithionite (e.g., Blankit) is used to bleach wool, cotton, silk, bristle, straw, horsehair, coconut fiber, raffia, soaps, glues, clay, sand, bauxite, and in some countries for bleaching sugar, syrup, fruit, edible oils, edible fats, and gelatin.

For special applications in the paper or textile industries complexing agents such as trilons or phosphates, or also optical brighteners are added to dithionite-containing products.

The reducing action of sodium dithionite is also used in preparative and analytical chemistry. It can reduce azo, diazo, nitro, nitroso, and carbonyl groups.

3.6. Safety and Environmental Aspects; Storage and Transport

Above a certain dithionite concentration, mixtures in powder form decompose if subjected to prolonged exposure to high temperatures or

come into contact with water. In the case of very finely divided products this decomposition can already occur at 80 °C. Product that is decomposing must be covered immediately with dry sand or powder extinguisher, or dissolved by shovelling into large quantities of water. If sodium dithionite packed in iron drums ignites, the SO_2 gas produced must be allowed to escape by opening the container or drilling holes in the wall. The contents of the container are destroyed by throwing them into large quantities of water. The aqueous solution thus formed must be treated as it contains reducing agent and must be slowly introduced into an appropriate wastewater-treatment plant. Gloves and respiratory protection must be worn while extinguishing fires.

Dithionite is classified in the German water hazard class 1 as posing a slight hazard to water.

The wastewater from all production processes contains approximately equal quantities of reducing agent, which must be removed in a wastewater-treatment plant. The CO_2 formed in the formate process (Section 3.3) must be freed from SO_2, methanol, and thiols before being released into the atmosphere.

Commercial sodium dithionite (e.g., Hydrosulfit and various product mixtures, which contain Hydrosulfit) are spontaneously combustible hazardous goods (Class 4.2, UN no. 1384), and are therefore subject to the corresponding transport regulations [31–35].

Because of the danger of spontaneous ignition sodium dithionite and its mixtures must be stored dry and in a cool place. Storage or transport together with sodium nitrite, sodium nitrate, and ammonium nitrate is forbidden. Transport containers must always be kept closed. Product should only be removed in a dry area using dry equipment.

3.7. Economic Aspects

World consumption of sodium dithionite is ca. 300 000 t/a, corresponding to ca. 60 % of production capacity in 1990. The zinc dust process accounts for ca. 35 % of the capacity, the formate process 40 %, the amalgam process 15 %, and the sodium borohydride process 10 %.

3.8. Toxicology [36]

The LD_{50} (rat, oral) is ca. 2500 mg/kg. In the rabbit the product has no irritant effect on the skin, but has an irritant effect on the mucous membrane (rabbit's eye). Therefore, in an emergency or as first aid measures it is sufficient to wash the skin thoroughly with water and to rinse the eyes for 10 min. with running water. An eye specialist should then be consulted.

The 48-h LC_{50} in golden orfe is 10 – 100 mg/L.

4. References

1. G. Nickless: *Inorganic Sulfur Chemistry,* Elsevier, Amsterdam 1968, pp. 509 – 533.
2. *Gmelin,* system no. 23, Ammonium, 1936, 256 – 259.
3. *Gmelin,* system no. 22, Kalium, 1938, 702 – 708.
4. *Gmelin,* system no. 21, Natrium, 1928, 515 – 535; suppl. 1964 – 1970, 205 – 213, 1079 – 1090, 1876.
5. F. Foerster et al., *Z. Phys. Chem. (Leipzig)* **110** (1924) 435 – 496.
6. Hoechst, DE-OS 2 221 298, 1973 (W. Schreiber, A. Metzger, J. Scholderer).
7. BASF, DE-OS 1 186 450, 1965 (W. Spormann, J. Heinke).
8. BASF, DE-OS 1 809 507, 1970 (W. Spormann).
9. Allied Chem., DE-OS 2 611 190, 1976 (S. L. Bean).
10. M. W. Wirth: *Toxikologiefibel,* Thieme Verlag, Stuttgart 1962.
11. H. V. Benda, *Z. Naturforsch. B Anorg. Chem. Org.* **34B** (1979) 957 – 968.
12. *Gmelin,* **system no. 21**, 3rd ed., 1162 – 1176.
13. *Chem. Abstr.* **99** no. 20, 160805 p.
14. *Chem. Abstr.* **104** no. 10, 71120 r.
15. *Chem. Abstr.* **110** no. 10, 78885 b; *Chem. Abstr.* **111** no. 12, 99852 z.
16. *Chem. Abstr.* **82** no. 12, 77551 f; *Chem. Abstr.* **82** no. 4, 19032 b.
17. American Standards Association INC., Specifications PH 4, New York 1960, pp. 250 – 251.
18. ISO Norm 3619, 1976.
19. DIN 19 080, Part 82, August 1978.
20. Th. Goldschmidt, DT 1 184 330, 1963; DT 1 265 146, 1966.
21. Allied Chem., CA 82 159, 1967.
22. *Chem. Abstr.* **70** no. 20, 89265 f.

23. *Chem. Abstr.* **80** no. 14, 72539 g.
24. *Chem. Abstr.* **107** no. 20, 179463 f.
25. NIOSH: *Registry of Toxic Effects of Chemical Substances,* Cincinnati 1983.
26. W. Braun, *Vergiftungsfibel,* Thieme Verlag, Stuttgart 1980.
27. M. Schmidt, W. Siebert: *Comprehensive Inorganic Chemistry,* **vol. 2,** Pergamon Press, Oxford 1973, p. 881. G. Nickless: *Inorganic Sulfur Chemistry,* Elsevier, Amsterdam 1968, p. 519.
28. Mitsubishi Edogawa, DE 1 592 013, 1978 (Y. Yoshikawa, H. Okazaki, T. Yamaguchi). BASF, DE-AS 2 442 418, 1974 (E. Voelkl et al.).
29. BASF, DE 2 646 825, 1976 (B. Leutner et al.). Olin Corporation, US 4 836 903, 1989 (D. W. Cawlfield).
30. G. Nickless: *Inorganic Sulfur Chemistry,* Elsevier, Amsterdam 1968, pp. 220, 227.
31. Bundesgesetz Blatt II, p. 838 (1990).
32. Bundesgesetz Blatt I, p. 1001 (1990).
33. Bundesgesetz Blatt I, p. 860 (1992).
34. Bundesgesetz Blatt I, p. 1714 (1991).
35. Bundesgesetz Blatt I, p. 1221 (1990).
36. BASF: safety data sheets, 1983.

Sulfochlorination → Sulfonic Acids, Aliphatic
Sulfochlorination → Surfactants
Sulfolane → Sulfones and Sulfoxides
Sulfonamides → Chemotherapeutics
Sulfonation → Sulfonic Acids, Aliphatic

Sulfones and Sulfoxides

Kathrin-Maria Roy, Langenfeld, Federal Republic of Germany

1. Sulfones 1
1.1. Physical and Chemical Properties . 1
1.2. Synthesis 2
1.2.1. Oxidation of Thioethers (Sulfides) . . 2
1.2.2. Alkylation of Sulfinic Acid Derivatives 4
1.2.3. Addition of Sulfinic Acid Derivatives to Multiple Bonds 4
1.2.4. Rearrangement of Sulfinates 5
1.2.5. Sulfones from Sulfonyl Carbanions . 5
1.2.6. Addition of Sulfur Dioxide 6
1.3. Reactions 6
1.4. Industrial Uses 7
2. Sulfoxides 7
2.1. Physical and Chemical Properties . 7
2.2. Synthesis 8
2.2.1. Oxidation of Sulfides 8
2.2.2. Reaction of Sulfur Dioxide with Aromatics 10
2.2.3. Sulfoxides from Sulfinates 10
2.2.4. α,β-Unsaturated Sulfoxides 10
2.2.5. Optically Active Sulfoxides 10
2.3. Reactions 10
2.3.1. Reactions of α-Sulfinyl Carbanions . 10
2.3.2. Reactions with Grignard Reagents . . 11
2.3.3. Pummerer Rearrangement 11
2.3.4. Oxidation and Reduction 11
2.4. Uses 11
2.5. Dimethyl Sulfoxide 12
2.5.1. Production 12
2.5.2. Uses 13
2.5.3. Reactions 13
3. Toxicology 13
3.1. Sulfones 13
3.2. Dimethyl Sulfoxide 13
4. References 14

The author would like to thank Bayer AG for undertaking the literature research for this article

1. Sulfones

1.1. Physical and Chemical Properties

Sulfones are described by the following general structural formula, in which the groups R^1 and R^2 can be hydrocarbon substituents, including cyclic ones, of any combination:

Typical sulfones are colorless and odorless in the pure state and are very stable both thermally and chemically. They are protonated only by the strongest acid systems. Unlike sulfoxides, sulfones form relatively weak hydrogen bonds.

The presence of a sulfonyl group in the molecule is shown by strong absorptions at 1300 – 1320 cm^{-1} and 1140 – 1160 cm^{-1} in the IR spectrum. Physical properties of industrially important sulfones are listed in Table 1.

1.2. Synthesis [1, pp. 223, 1129], [2], [3, p. 171]

1.2.1. Oxidation of Thioethers (Sulfides)

Sulfones are generally synthesized by oxidation of sulfides via sulfoxides as intermediates. A common oxidizing agent in the laboratory is hydrogen peroxide in acetic acid. It is also often used in industry, but requires reaction temperatures of 70 °C up to reflux in many cases [4], [5]. In the presence of various metal catalysts, oxidation with hydrogen peroxide proceeds under milder conditions. Thus, β-hydroxy sulfides can be oxidized selectively and in good yield to the corresponding sulfones by using titanium, vanadium, molybdenum, or tungsten salts [6], as well as selenium dioxide [7], as catalysts. A very effective oxidizing agent, which gives high-purity products in high yield, is *m*-chloroperbenzoic acid (MCPBA). This reagent can often be used if the molecule contains other sensitive groups,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_487

Table 1. Physical data and molecular formulae for some industrially important sulfones

Sulfone	*mp*, °C (*bp*, °C/kPa)	Molecular formula (molecular mass)
	(285/101.3)	$C_4H_8O_2S$ (120.16)
$4\text{-}CH_3-C_6H_4-SO_2-C_6H_4-4\text{-}CH_3$	158	$C_{14}H_{14}O_2S$ (246.33)
$4\text{-}Br-C_6H_4-SO_2-C_6H_4-4\text{-}Br$	172	$C_{12}H_8Br_2O_2S$ (376.08)
$CH_3-(CH_2)_3-SO_2-(CH_2)_3-CH_3$	46	$C_8H_{18}O_2S$ (178.30)
$HO-CH_2-CH_2-SO_2-CH_2-CH_2-OH$	58	$C_4H_{10}O_4S$ (154.19)
$C_6H_5-CH_2-SO_2-CH_2-C_6H_5$	149–153	$C_{14}H_{14}O_2S$ (246.33)
$C_2H_5-SO_2-CH_2-CONH_2$	99	$C_4H_9NO_3S$ (151.19)
$CH_3-(CH_2)_2-SO_2-CH_2-CONH_2$	104	$C_5H_{11}NO_3S$ (165.21)
$C_6H_5-SO_2-CH_2-CH_2-OH$	(184/19)	$C_8H_{10}O_3S$ (186.23)
$CO_2H-CH_2-CH_2-SO_2-CH_2-CH_2-CO_2H$	222–224	$C_6H_{10}O_6S$ (210.21)
$C_2H_5-SO_2-(CH_2)_2-CH_3$	25	$C_5H_{12}O_2S$ (136.21)
$(CH_3)_2CH-SO_2-CH(CH_3)_2$	35–36	$C_6H_{14}O_2S$ (150.24)
$CH_3-SO_2-CH_3$	109	$C_2H_6O_2S$ (94.13)
$CH_3-(CH_2)_2-SO_2-(CH_2)_2-CH_3$	26	$C_6H_{14}O_2S$ (150.24)
$CH_3-(CH_2)_7-SO_2-(CH_2)_7-CH_3$	73	$C_{16}H_{34}O_2S$ (290.51)
$CH_3-SO_2-CH_2-C_6H_5$	126–127	$C_8H_{10}O_2S$ (170.23)
$CH_3-SO_2-(CH_2)_3-CH_3$	28– 30	$C_5H_{12}O_2S$ (136.21)
$CH_3-SO_2-C_6H_4-4\text{-}CH_3$	88– 89	$C_8H_{10}O_2S$ (170.23)
$4\text{-}Cl-C_6H_4-CH_2-SO_2-C_6H_4-4\text{-}CH_3$	167–168	$C_{14}H_{13}ClO_2S$ (280.78)
$CH_2=CH-CH_2-SO_2-C_6H_4-4\text{-}CH_3$	50– 52	$C_{10}H_{12}O_2S$ (196.27)
$C_2H_5CO_2-CH_2-SO_2-C_6H_4-4\text{-}CH_3$	91– 93	$C_{11}H_{14}O_4S$ (242.30)
$4\text{-}CH_3-C_6H_4-SO_2-C_6H_5$	125–126	$C_{13}H_{12}O_2S$ (232.30)
$2,2',4\text{-}(CH_3)_3-C_6H_2-SO_2-C_6H_4-4\text{-}OCH_3$	136	$C_{16}H_{18}O_3S$ (290.38)
$4\text{-}CH_3-C_6H_4-SO_2-C_6H_3-2,4\text{-}(CH_3)_2$	51–52	$C_{15}H_{16}O_2S$ (260.36)
$C_6H_5-SO_2-C_6H_4-4\text{-}C_2H_5$	92–93	$C_{14}H_{14}O_2S$ (246.33)
$4\text{-}CH_3O-C_6H_4-SO_2-CH_3$	120	$C_8H_{10}O_3S$ (186.23)
$2,4\text{-}(CH_3)_2-C_6H_3-SO_2-CH_3$	56	$C_9H_{12}O_2S$ (184.26)
$C_6H_5-SO_2-C_6H_5$	121	$C_{12}H_{10}O_2S$ (218.28)
$4\text{-}CH_3-C_6H_4-SO_2-C_6H_5$		$C_{13}H_{12}O_2S$ (232.30)
$4\text{-}CH_3-C_6H_4-CH(CH_2-NO_2)-SO_2-C_6H_4-4\text{-}NHCOCH_3$	185–187	$C_{17}H_{18}N_2O_5S$ (362.41)
$C_6H_5-CH(CH_2-NO_2)-SO_2-C_6H_5$	186–187	$C_{14}H_{13}NO_4S$ (291.33)
$4\text{-}CH_3O-C_6H_4-CH(CH_2-CO-C_6H_4-4\text{-}CH_3)-SO_2-C_6H_4-4\text{-}Cl$	158	$C_{23}H_{21}ClO_4S$ (428.94)
$C_6H_5-CH(CH_2-CO-C_6H_5)-SO_2-C_6H_4-4\text{-}N(CH_3)_2$	191–192	$C_{23}H_{23}NO_3S$ (393.51)
$C_6H_5-SO_2-CH_2-OH$	57–60	$C_7H_8O_3S$ (172.20)
$4\text{-}CH_3-C_6H_4-SO_2-CH(OH)-(CH_2)_2-CH_3$	78	$C_{11}H_{16}O_3S$ (228.31)
$CH_3-(CH_2)_{11}-SO_2-CH_2-OH$	64–66	$C_{13}H_{28}O_3S$ (264.43)

as in many pharmaceutically active substances [8]. Alkyl thiocyanates are converted to sulfonyl cyanides by MCPBA [9]

$$R-SCN \xrightarrow{MCPBA} R-SO_2-CN$$

Other oxidizing agents used include potassium permanganate [10], sodium perborate [11], and dimethyldioxirane [12]. An extremely simple, rapid process is sulfide oxidation with periodic acid in the presence of ruthenium tetroxide as catalyst. This reaction is carried out in a two-phase solvent mixture and is, therefore, also suitable for the synthesis of water-soluble sulfones [13]. A related method using osmium tetroxide in the presence of a tertiary amine *N*-oxide gives good yields [14]. In industrial processes, gaseous chlorine is also used as the oxidizing agent (see, for example [15]). The potassium peroxomonosulfate complex Oxone has also gained importance [16]. With this reagent, allyl sulfoxides, for example, can be converted into the corresponding sulfones [17]. In the presence of wet clays such as kaolin or montmorillonite, the oxidation of various sulfides in dichloromethane gives the sulfone almost quantitatively [18].

Oxidation of Phenyl Vinyl Sulfide with H_2O_2 in Glacial Acetic Acid [4]. To a solution of 19.7 g (0.145 mol) of phenyl vinyl sulfide in 70 mL of glacial acetic acid, 56 mL (0.5 mol) of 30 % hydrogen peroxide is added dropwise in such a way that the reaction temperature becomes 70 °C. The reaction mixture is subsequently heated for 20 min under reflux. After cooling, 150 mL of diethylether and 200 mL of water are added, and the organic phase is separated. The organic phase is subsequently washed with water (50 mL) and sodium chloride solution (50 mL) and evaporated, giving 18 – 19 g (74 – 78 %) of phenyl vinyl sulfone as a colorless solid, *mp* 64 – 65 °C.

Oxidation of Sulfides with Oxone in the Presence of Wet Montmorillonite [18]. Commercial clay (montmorillonite or kaolin; 10 g) is treated with 2 g of distilled water in eight portions. After each addition the mixture is shaken vigorously for several minutes, until a loose powder is formed. Then 1.2 g of this catalyst is suspended in 6 mL of dichloromethane (7 mL for the kaolin catalyst) and treated with Oxone (2.5 equivalents based on the sulfide). After addition of 1 mmol sulfide the reaction vessel is purged with argon, closed, and stirred for 2 h at room temperature. After filtration of the catalyst on a glass frit, washing with ca. 100 mL of dichloromethane, and evaporation of the solvent, the pure sulfone (> 98 % purity) is obtained almost quantitatively.

A particularly simple and rapid process allows the synthesis of sulfonyl-substituted propanones on a large scale [19]. Here, chloroacetone is reacted with various sulfides. The intermediate β-keto sulfide is not isolated but is oxidized directly with Oxone to the sulfone.

$$CH_3-CO-CH_2Cl + R-SH \xrightarrow[MeOH]{NaOH} CH_3-CO-CH_2-S-R \xrightarrow[H_2O]{Oxone} CH_3-CO-CH_2-SO_2-R$$

1.2.2. Alkylation of Sulfinic Acid Derivatives

Alkylation of the ambidentate sulfinate ion with soft alkylating agents, such as alkyl halides, is a widely used method for synthesis of sulfones:

$$R^1-SO_2^-Na^+ + R^2-X \longrightarrow R^1-SO_2-R^2$$

Aryl sulfones, in particular, in which the second sulfonyl substituent can be an alkyl [20], an alkynyl [21], a β-keto [22], or an allyl [23] group, can be produced by this process. These reactions are catalyzed, for example, by copper(I) bromide [24], phase-transfer catalysts [20], [21], [25], or ultrasound [26].

A related reaction is that of arylsulfonyl chlorides with alkyl iodides or benzyl chlorides in the presence of telluride. The reaction is based on a reactivity inversion, where the original electrophilic sulfonyl group is converted to the nucleophilic sulfinate ion by electron transfer from the telluride ion [27]:

$$Ar-SO_2Cl \xrightarrow{Te^{2-}} [Ar-SO_2^-] \xrightarrow{R-X} Ar-SO_2-R$$

Substituted vinyl and allyl sulfones are obtained in good yield by palladium-catalyzed coupling of arylsulfonyl chlorides with organotin compounds [28].

$$Ar-SO_2Cl + Bu_3Sn-CH=CH-R \xrightarrow{(Ph_3P)_4Pd} Ar-SO_2-CH=CH-R$$

The sulfonylation of various organometallic reagents with arylsulfonyl fluorides is a one-step sulfone synthesis, which gives alkyl, trimethylsilylmethyl, and allyl sulfones in good yield [29]. Cycloalkenes can be converted particularly easily into the corresponding vinyl sulfones by reaction with tetrabutylammonium sulfinate [30].

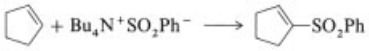

1.2.3. Addition of Sulfinic Acid Derivatives to Multiple Bonds

Isolated double bonds do not generally react with sulfinic acids or their salts to give sulfones. Iodosulfonylation and sulfonylmercuration, which both take place regioselectively [31], are exceptions. The adducts formed initially are converted by dehydrohalogenation or demercuration into the corresponding vinyl sulfones (see Figure 1).

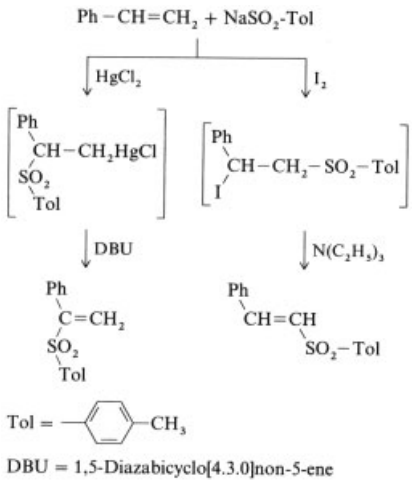

Figure 1.

Conjugated dienes [32] and α,β-unsaturated carbonyl compounds [33] can also be converted regioselectively into unsaturated sulfones by using iodosulfonylation.

The addition of sulfinic acids to polarized double or triple bonds is an excellent method of synthesizing sulfones [34]. A large number of olefins, such as vinyl aldehydes [35] and ketones [36], nitroolefins [37], acrylic acid derivatives [38], vinyl sulfones [39], quinones [40], and quinonimines [41] have been used successfully as acceptor compounds.

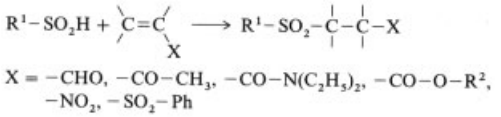

Sulfinic acid also adds readily to sulfonyl-substituted allenes and acetylenes to give disulfonyl derivatives [42].

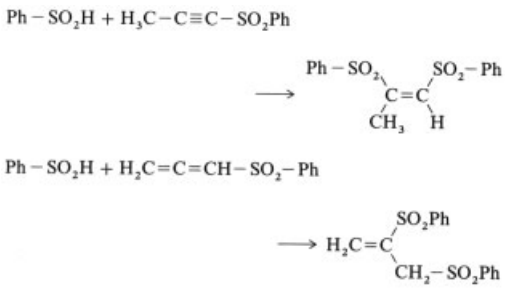

Unlike sulfinic acids, the addition of sulfonyl halides to unsaturated C–C systems is a radical reaction in which mixtures of isomers can be formed. To suppress competing polymerization reactions, the addition is carried out in the presence of copper(I) or copper(II) salts. The β-halo sulfones formed initially can be converted into vinyl sulfones by elimination of hydrogen halide [43].

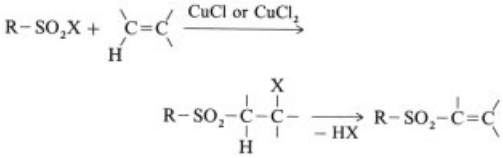

1.2.4. Rearrangement of Sulfinates

Sulfinates rearrange to sulfones under various conditions. Acids [44], tertiary amines [45], or palladium(0) complexes [46] are examples of suitable catalysts. Allyl-substituted sulfinic acid esters can also be isomerized thermally [46], [47].

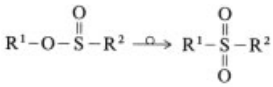

1.2.5. Sulfones from Sulfonyl Carbanions

α-Sulfonyl carbanions are formed from sulfones containing α-hydrogens by the action of bases, such as Grignard reagents, alkali-metal amides, sodium hydride, or butyllithium. These carbanions can be converted into a variety of new sulfones by alkylation, acylation, or addition to unsaturated systems.

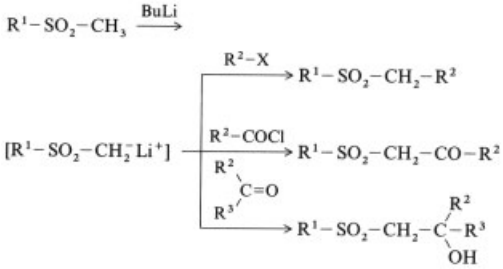

Treatment of α-sulfonyl carbanions with acid chlorides [48] or esters [49] gives β-keto sulfones, which can rearrange with allyl bromide to give α-allyl-substituted β-keto sulfones [50].

The Knoevenagel condensation of aryl sulfones with aldehydes or ketones is also frequently used. The intermediate products eliminate water to form vinyl sulfones [51] (see Figure 2).

The Knoevenagel condensation, with subsequent allyl sulfoxide – sulfenic acid ester rearrangement, is of synthetic importance in the production of γ-hydroxy-α,β-unsaturated sulfones. Here piperidine is used as the base [52] (see Figure 3).

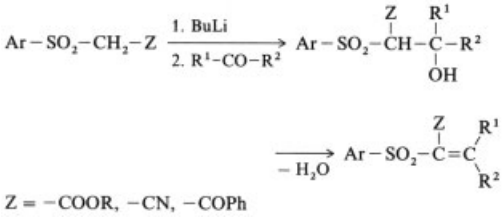

Figure 2.

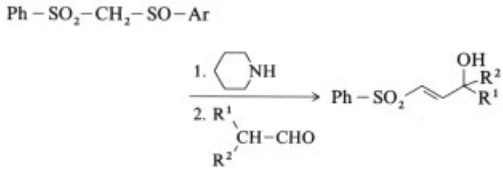

Figure 3.

1.2.6. Addition of Sulfur Dioxide

By addition of sulfur dioxide to aryldiazonium salts, aryl sulfonyl radicals are formed, which can react with olefins to give alkyl aryl sulfones [53].

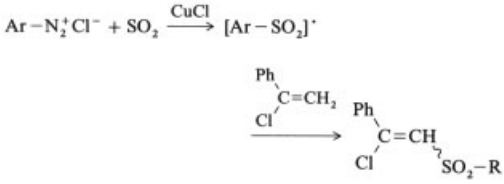

1,3-Dienes undergo stereoselective cycloaddition with sulfur dioxide to give sulfolenes. This reaction is reversible [54].

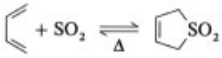

1.3. Reactions

Reductive Elimination of the Sulfonyl Group. Reductive elimination of the sulfonyl group is of great synthetic importance because sulfones are used as intermediates in many syntheses. Alkali metals in ammonia or low molar

mass amines [55], sodium amalgam [56], and aluminum amalgam [57] have been described as reducing agents. Sulfur dioxide is eliminated from unsaturated, cyclic sulfones by potassium dispersed with ultrasound [58].

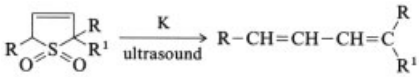

The simultaneous reduction and desulfonylation of α-alkylidene β-keto sulfones is effected under mild conditions by using sodium hydrogentelluride [59].

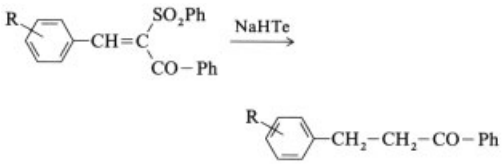

The sulfonyl group can be eliminated photochemically from β-keto sulfones by using a β-keto ester in the presence of a ruthenium(II) salt, so that many other functional groups are not attacked [60].

Elimination of Sulfinate. The base-catalyzed elimination of organic sulfinate groups from sulfones is often used preparatively. Arylsulfonyl [61] and, in particular, trifluoromethanesulfonyl groups are good leaving groups [62]:

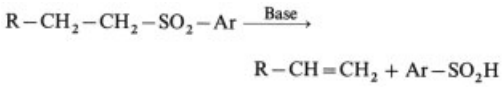

Ramberg–Bäcklund Reaction. Bases effect the elimination of sulfur dioxide from α-halo sulfones with α-hydrogens via an episulfone intermediate, to give olefins [63].

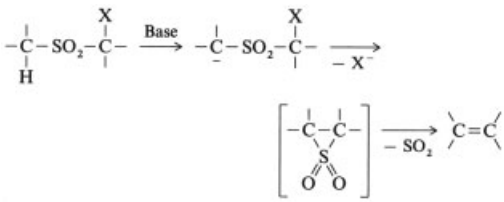

Smiles and Truce–Smiles Rearrangement. Diaryl and aralkyl sulfones undergo base-catalyzed rearrangement to form sulfinic acids [64]. In the *Smiles rearrangement* the migrating aryl group must be activated by an electron-withdrawing group.

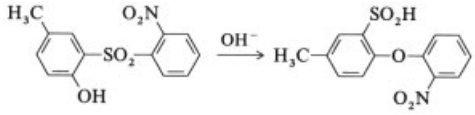

In the related *Truce–Smiles rearrangement* a nonactivated aryl or tertiary alkyl group migrates to a carbanion formed by lithiation of an *o*-methyl group [65].

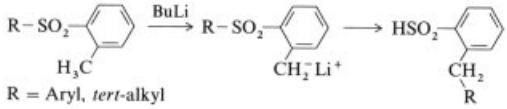

1.4. Industrial Uses

Unsaturated sulfones can be used as intermediates for pharmaceuticals and pesticides and for modification of polymers [66]. Many reactive dyes also contain unsaturated sulfonyl groups, (e.g., see [67]). The possible uses of sulfones described below are taken from the extensive patent literature. For uses of sulfolane, see → Benzene, Chap. 6.3.; → Gas Production.

Uses in Crop Protection. 1,1-Diiodomethyl sulfone [68], heterocyclic sulfones [69], and α-sulfonyloximes [70] have been described as *plant fungicides*. Certain sulfones are effective *herbicides* [71]. Others act as *insecticides* and *acaricides* [72].

Pharmaceutical Uses. Dihydropyridine sulfone derivatives are *calcium antagonists*, which are said to have a longer period of activity and fewer side effects than nifedipine [73]. Alkylsulfonyl-substituted steroid pyrazoles and thiazoles exhibit *antiandrogenic activity* [74].

Sulfonylnitromethane derivatives are aldose reductase inhibitors and can be used for the treatment of certain secondary effects of *diabetes* and *galactosemia* [75]. A range of sulfonyl-substituted benzazepines and imidazoles have

been patented as agents for the treatment of *gastrointestinal disorders* [76]. Fluoro sulfones are said to be very effective leukotriene antagonists with *antiallergic* and *anti-inflammatory activity* [77]. Aryl- or heteroaryl-substituted sulfones exhibit activity against *picornaviruses* [5] and *retroviruses* [78]. Sulfonylcyclopentenones are *antitumor agents* and promote *bone formation* [79].

Others. Piperidinyl sulfones are used as light stabilizers in polymers, particularly polyolefins [80]. Certain heterocyclic sulfones can be used as*corrosion protection agents* for metals [81]. Sulfones are used as hardeners [82] and couplers [83] in *photographic materials*. Diazodisulfones can be used in the production of *semiconductors* and as photosensitive reagents for *printing plate materials* [84]. Certain fluorinated sulfones have *nonlinear optical properties* [85]. Sulfone percarboxylic acids are *bleaching agents* with little damaging effect on colors and good storage stability [86]. *Fabric softeners* contain esters or ethers of thiodiglycol sulfones [87].

2. Sulfoxides

2.1. Physical and Chemical Properties

Sulfoxides are colorless, hygroscopic liquids or solids. They dissolve readily in protic solvents to form strong hydrogen bonds. Because of the polar nature of the S−O bond, sulfoxides form adducts with salts of most metals.

Because of the pyramidal structure of sulfoxides, unsymmetrically substituted compounds of this series form enantiomers, with the sulfur atom as the chiral center.

In sulfoxides, sulfur has the intermediate oxidation state of +4. Sulfoxides can therefore be reduced to form sulfides (thioethers) and oxidized to form sulfones by reversible redox reactions.

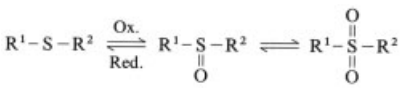

Physical properties of important sulfoxides are listed in Table 2.

2.2. Synthesis

2.2.1. Oxidation of Sulfides [1, pp. 207, 665], [2], [3, p. 121], [88], [89]

The oldest and most frequently used method for the production of sulfoxides is by oxidation of sulfides. Since further oxidation to the corresponding sulfones can occur, a range of selective reagents and methods has been developed. Oxidation with hydrogen peroxide, which has long been known, can be carried out in the presence of various catalysts. Acids, such as acetic [90], [91], perchloric [92], and sulfuric acids [93], [94] are suitable. A more recent industrial process uses 50 % sulfuric acid and an organic acid, such as formic or acetic acid, in a 1 : 1 ratio as the catalyst [93]. Similarly, sterically hindered sulfides can be oxidized selectively and in high yield with hydrogen peroxide in the presence of a catalyst mixture of sulfuric acid and a secondary or tertiary alcohol [94]. Other suitable catalysts are titanium trichloride [95], selenium dioxide [96], tellurium dioxide [97], and vanadium pentoxide [98], which can act simultaneously as an indicator of excess hydrogen peroxide. Good yields of sulfoxide are obtained by using the complex from urea and hydrogen peroxide (UHP complex) together with half an equivalent of phthalic anhydride [99]. The use of organic peroxides, such as cyclohexyl or *tert*-butyl hydroperoxide has also been known for a long time [100]. Sulfide oxidation takes place quantitatively at −78 °C using 2-methoxy-2-hydroperoxypropane produced in situ from 2,3-dimethyl-2-butene and ozone [101]. Among the organic peroxycarboxylic acids, peracetic [102], perbenzoic [103], and monoperoxyphthalic acids [104], in particular, are effective oxidizing agents. One of the best selective oxidizing agents is *m*-chloroperoxybenzoic acid, which not only gives sulfone-free sulfoxides [105], but also reacts with a broad range of substrates [106] including unstable unsaturated thioethers [4], multifunctional heterocyclic thioesters [107], thioacetals [108], disulfides [109], and penicillins [110].

Table 2. Physical data and molecular formulas for some industrially important sulfoxides

Sulfoxide	*mp*, °C (*bp*, °C/kPa)	Molecular formula (molecular mass)
$C_6H_5-CH_2-SO-CH_2-C_6H_5$	134–135	$C_{14}H_{14}OS$ (230.33)
$C_6H_5-SO-CH_3$	33–34	C_7H_8OS (140.21)
$C_6H_5-SO-C_6H_5$	69–71	$C_{12}H_{10}OS$ (202.28)
$C_6H_5-SO-CH_2-COOH$	118–119	$C_8H_8O_3S$ (184.21)
$CH_3-CH_2-SO-CH_2-CH_3$	(45–47/0.015)	$C_4H_{10}OS$ (106.19)
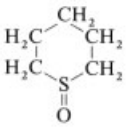	67–68	$C_5H_{10}OS$ (118.20)
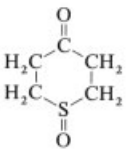	109–110	$C_5H_8O_2S$ (132.18)
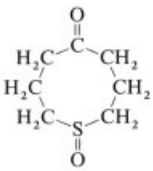	91–92	$C_7H_{12}O_2S$ (160.24)
$C_6H_5-SO-CH_2-CH_3$	(68–70/0.005)	$C_8H_{10}OS$ (154.23)
$CH_3-(CH_2)_3-SO-(CH_2)_3-CH_3$	(105/0.3)	$C_8H_{18}OS$ (162.30)
$4\text{-}H_3C-C_6H_4-SO-CH_3$	38–39	$C_8H_{10}OS$ (154.23)
$C_6H_5-CH_2-SO-CH_2-CH_3$	51–52	$C_9H_{12}OS$ (168.26)
$C_6H_5-SO-CH_2-C_6H_5$	122–123	$C_{13}H_{12}OS$ (216.30)
$CH_3-(CH_2)_2-SO-(CH_2)_2-CH_3$	(73–75/0.1)	$C_6H_{14}OS$ (134.24)
$C_6H_5-CH_2-SO-CH_3$	53–54	$C_8H_{10}OS$ (154.23)
$C_6H_5-SO-CH=CH_2$	(105–110/0.15)	C_8H_8OS (152.22)
$4\text{-}H_3CO-C_6H_4-SO-CH_3$		$C_8H_{10}O_2S$ (170.23)
$C_6H_5-SO-CH_2-CH=CH_2$		$C_9H_{10}OS$ (166.24)
$C_6H_5-CO-CH_2-SO-C_6H_4-4\text{-}CH_3$	83– 84	$C_{15}H_{14}O_2S$ (258.34)
$4\text{-}CH_3-C_6H_4-SO_2-CH_2-SO-C_6H_4-4\text{-}CH_3$	117–118	$C_{15}H_{16}O_3S_2$ (308.42)

The longest-known oxidizing agent, used for sulfide oxidation as early as 1865, is nitric acid [111]. In the presence of gold (III) salts and by phase-transfer catalysis, oxidation with aqueous nitric acid in nitromethane is selective; oxidizable groups such as vinyl, tertiary amino, hydroxy, and diol groups are not attacked [112]. Very good yields of sulfoxide are also obtained with acetyl and benzoyl nitrates [113], and with the inorganic oxidants thallium(III) nitrate [114] and ceric ammonium nitrate (CAN). Oxidation with CAN alone is limited to diaryl sulfides, since in the presence of an α-hydrogen in the substrate molecule, the sulfoxide formed undergoes a Pummerer rearrangement under reaction conditions [115]. When CAN oxidation is carried out with sodium bromate as a co-oxidizing agent [116] or by phase-transfer catalysis [117], dialkyl and aryl alkyl sulfides can also be oxidized selectively. One of the most frequently used methods for effecting the sulfide–sulfoxide transformation is oxidation with sodium metaperiodate in aqueous methanol at 0 °C [118]. Modified processes employ sodium metaperiodate on an alumina carrier [119] or water-insoluble tetrabutylammoni-

um metaperiodate in boiling chloroform [120]. Iodosobenzene [121] and *o*-iodosylbenzoic acid [122] are also effective oxidizing agents. Selective oxidation with halogen compounds, such as sodium bromite [123], benzyltrimethylammonium tribromide [124], and particularly sodium hypochlorite has been described. The latter can be used in alkaline media at pH > 10 for the production of pharmacologically active compounds [125].

In addition to sulfones, sulfoxides are also obtained by reaction of sulfides with Oxone [16]. In the presence of wet alumina, phenyl alkyl sulfides are converted to either sulfone-free sulfoxides or sulfones, depending on the quantity of Oxone used [126]. Diastereoselective oxidation with Oxone has also been described [127].

Synthesis of 4-Chlorobenzyl-*N,N*-Diethylcarbamoyl Sulfoxide [94]

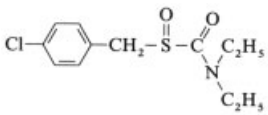

10 mol (2.575 kg) 4-chlorobenzyl-*N*,*N*-diethylthiocarbamate are dissolved in 1.5 L of toluene and treated with 0.4 L of formic acid and 0.4 L of 50 % sulfuric acid. Then 1.07 kg of 35 % hydrogen peroxide (11 mol) is added dropwise with stirring and ice- cooling in such a way that the temperature of the reaction mixture does not exceed that of the surroundings. After the mixture has been stirred for 4 h at room temperature, the organic phase is separated and neutralized with 20 % aqueous NaOH. The toluene phase is washed with water and evaporated to give the sulfoxide in 95 % yield (purity 95 %).

Synthesis of Phenyl Vinyl Sulfoxide [4]

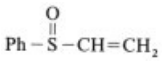

A solution of 20 g (0.147 mol) phenyl vinyl sulfide in 250 mL of dichloromethane, is cooled to − 78 °C, and a solution of 25.4 g (1.0 equivalent) of *m*- chloroperoxybenzoic acid in 200 mL of dichloromethane is added dropwise with stirring over 30 min. The reaction mixture is then warmed for 1 h in a water bath at 30 °C and poured into 300 mL of saturated hydrogencarbonate solution. The organic phase is separated, and the aqueous phase is extracted with 3×250 mL of dichloromethane. The combined organic phases are dried over magnesium sulfate, and after evaporation of the solvent, the residue is distilled. About 15 – 16 g (68 – 70 %) of the sulfoxide is obtained as a colorless liquid, *bp* 98 °C (80 Pa).

2.2.2. Reaction of Sulfur Dioxide with Aromatics

Simple aromatic compounds can be converted into the corresponding sulfoxides in good yield with sulfur dioxide and an equimolar quantity of Magic Acid ($FSO_3H \cdot SbF_5$). Unsymmetrical sulfoxides are also accessible by this route [128].

2.2.3. Sulfoxides from Sulfinates

Reaction of sulfinates with organometallic compounds, such as Grignard reagents and organocopper or organolithium reagents, is important for the production of optically active sulfoxides.

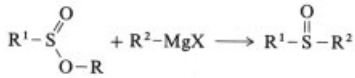

An extension of this reaction is the Claisen condensation of enolate anions with methyl arylsulfinates, giving α-keto sulfoxides [129].

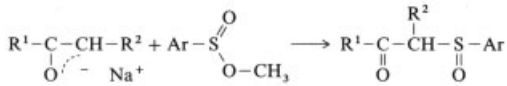

2.2.4. α,β-Unsaturated Sulfoxides

A range of vinyl sulfoxides is accessible through trapping reactions of thermally produced sulfenic acids with alkynes [130].

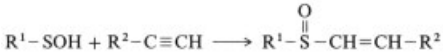

2.2.5. Optically Active Sulfoxides

Chiral sulfoxides, which play an important role in asymmetric synthesis, are produced by asymmetric synthesis, optical and kinetic separation of enantiomers, and stereospecific synthesis; for reviews, see [3], [131].

2.3. Reactions

2.3.1. Reactions of α-Sulfinyl Carbanions

Strong bases, such as butyllithium, sodium hydride, or lithium dialkylamide, deprotonate sulfoxides with α-hydrogens to give α-sulfinyl carbanions. The latter react stereoselectively with organic electrophiles to form C–C bonds [1], [2, p. 583].

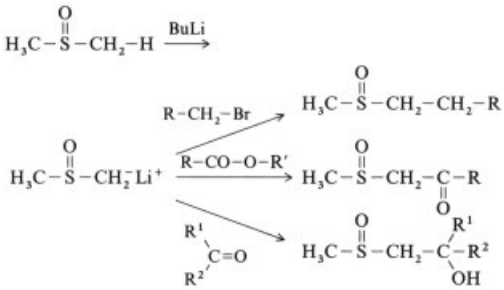

Subsequent reactions consist of reductive cleavage of the sulfinyl group to a ketone and elimination with the formation of a double bond. These processes have been used for the synthesis of various ketones and carboxylic acids [132].

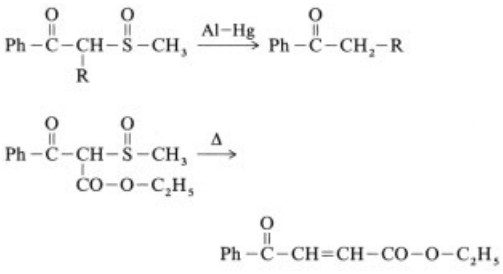

For α-halogenation of sulfoxides, which also proceeds via sulfinyl carbanions, various reagents have been described [1], [3].

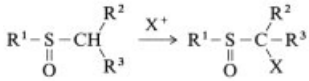

Halogenation with *N*-halosuccinimides [133] and chlorination with *N*-,*N*-dichloro-*p*-toluenesulfonamide [134] give good yields.

2.3.2. Reactions with Grignard Reagents

Ligand exchange reactions between heterocyclic sulfoxides and Grignard reagents are used for production of the corresponding heterocyclic Grignard reagents. Subsequent reaction of these organometallic derivatives with carbonyl compounds provides a route to new heterocyclic compounds [135], [136].

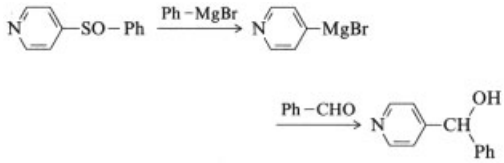

A variant of this reaction is ligand coupling. The heterocyclic Grignard compound (**2**) formed initially reacts with the educt sulfoxide (**1**) to give the coupling product **3** [137].

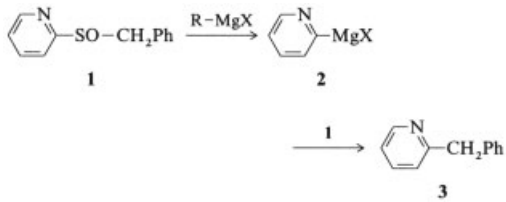

This reaction is used mainly for the synthesis of bisaryls [138], [139].

2.3.3. Pummerer Rearrangement

In the Pummerer rearrangement, sulfoxides having at least one α-hydrogen are treated with acids or strong electrophiles such as carboxylic acid anhydrides or chlorides to form α-substituted sulfides as the primary products.

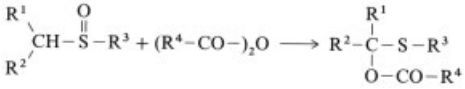

The Pummerer rearrangement is used widely in synthesis [1], [3]. For example, α-dicarbonyl compounds are formed from β-keto sulfoxides on acidic workup.

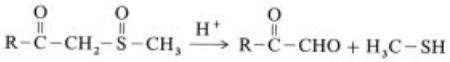

Cycloalkyl phenyl sulfoxides are converted into vinyl sulfides under conditions of the Pummerer rearrangement [140].

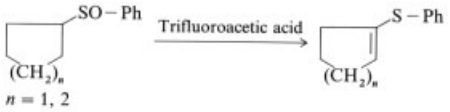

2.3.4. Oxidation and Reduction

Oxidation of sulfoxides to sulfones not only is a frequently occurring side reaction in the production of sulfoxides, but is also used for the synthesis of sulfones (see Section 2.2.1).

Reduction of sulfoxides to sulfides is achieved with hydrogen in ethanol on Pd–C catalysts [141] or selectively with dichloroborane [142] even in the presence of acid chloride, ester, nitrile, or nitro groups. Deoxygenation of sulfoxides with zinc – dichlorodimethylsilane under very mild conditions [143] and with polystyryldiphenylphosphine [144] has also been described. These reactions have no effect on a wide variety of functional groups and permit a cleaner product isolation.

2.4. Uses

Many sulfoxides are used as intermediates in the production of pharmaceuticals, plant protection agents, and other physiologically active substances.

Sulfoxides also play an important role in the synthesis of optically active substances. A selection of patents on the use of sulfoxides is given below.

Crop Protection and Veterinary Drugs
Insecticides and acaricides [145].
Bactericides and fungicides [146].
Plant growth regulators [147].
Anthelmintics in veterinary medicine [148].

Pharmaceuticals
Agents for treating gastrointestinal disorders [149].
Enantiomerically pure sulfoxides as cardiovascular medicines [150].
Sulfoxide derivatives of 5-thiaprostaglandin E_1 act as vasodilators and inhibit thrombocyte aggregation [151].
Thiocarbamoylsulfoxide derivatives exhibit antihypertensive action [152].
Imidazolyl sulfoxides are anti-inflammatory and are also said to be suitable for treating asthma and skin diseases such as psoriasis [153].
Trichlororuthenium complexes with a sulfoxide group have antitumor activity [154].
Agents for treating allergic reactions, particularly asthma [155].
Methionine sulfoxide lowers the liver toxicity of acetaminophen in humans [156].
Tetrahydrothiophen-1-oxide and its derivatives are inhibitors of alcohol dehydrogenase and therefore act as antidotes for alcohol poisoning [157].
Various sulfoxides are enzyme inhibitors [158].
Thiocarbamate sulfoxides have been described as agents for combating alcohol addiction [159].
Agents for treating sun-damaged skin [160].
Sulfoxides support the percutaneous absorption of a range of pharmaceuticals [161].

Others
Bis(2-hydroxyethyl) sulfoxide is used to treat oily hair [162].
Sulfoxides act as stabilizers for polymers and lubricating agents against heat, light, and oxygen [163].
Polyalkoxylated thioglycol sulfoxides are used as fabric softeners [164].
Use in refrigerating liquids for refrigerators and air-conditioning equipment [165].
Coupling agents in silver halide color films [166].

2.5. Dimethyl Sulfoxide

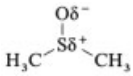

Dimethyl sulfoxide (DMSO) is a colorless, odorless, strongly hygroscopic liquid. It is readily miscible with water and other protic solvents such as alcohols. This property can be attributed to the formation of strong hydrogen bonds between the polarized S–O bond and water molecules. The strong polarity of the S–O bond, with a partial negative charge on oxygen (see formula) is shown by the high dipole moment (4.3 ± 0.1 D at 20 °C) and basicity ($pK_B = -2.7$) of dimethyl sulfoxide. Pure DMSO forms chain-like polymeric association complexes at room temperature. This gives rise to considerably higher melting and boiling points compared to dimethyl sulfide (see Table 3).

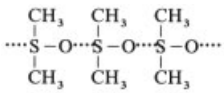

Table 3. Comparison of physical properties of dimethyl sulfide and dimethyl sulfoxide

	Dimethyl sulfide	Dimethyl sulfoxide
mp, °C	−98	18.5
bp, °C	38	189
d, g/mL	0.846	1.101

2.5.1. Production

In principle, DMSO can be produced by oxidation of dimethyl sulfide as described in Section 2.3.1. These methods are, however, often suitable only for small quantities. In industry, DMSO is produced by catalytic oxidation of dimethyl sulfide with oxygen or by oxidation with nitrogen dioxide [167]. The oxidation of dimethyl sulfide with oxygen at 7.2 MPa and 105 °C gives good yields of DMSO [168].

2.5.2. Uses

Dimethyl sulfoxide is an excellent solvent [169] that is widely used for reactions in modern chemistry. Because of its high polarity and good solvation of cations, DMSO dissolves not only a large number of organic compounds, but also many metal salts, particularly those of alkali and alkaline-earth metals. DMSO is also a good solvent for many polymers.

A range of different pharmacological properties and effects have been described for DMSO [170]. For example, it increases the capacity of the skin to absorb drugs, and exhibits both analgesic and bacteriostatic activity. It can be used for the treatment of infection by retroviruses [171] and as an anti-inflammatory agent. Furthermore, DMSO has been described as a preservative and cryo-protective agent for organ and tissue transplants [172].

In plant protection, DMSO can be used to ward off mammals (e.g., deer [173]) and as a growth stimulant [174].

2.5.3. Reactions

The so-called dimsyl anion, formed on treatment of dimethyl sulfoxide with base, can react further like other sulfinyl carbanions (see Section 2.3.1).

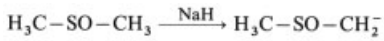

Cyclic ketones react with DMSO in the presence of KOH to give 2-methylenecycloalkanols [175].

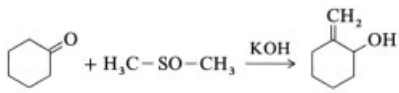

The methylation of various aromatic systems by the dimsyl anion is also of interest [176].

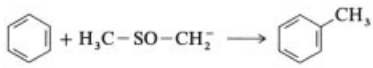

3. Toxicology

3.1. Sulfones [177–181]

The toxicity of sulfones is generally low. The sulfonyl group is relatively stable metabolically.

However, for some compounds, enzymatic reduction to the sulfide has been discussed. For the solvent *sulfolane* (tetramethylene sulfone), on oral administration to rats, an average lethal dose of 1.8 – 2.5 g/kg has been found. Various mammals show neurotoxic symptoms after oral administration, injection, inhalation, or dermal absorption of sulfolane. No information is available on the carcinogenic or teratogenic activity of sulfolane. *p-Chlorophenyl phenyl sulfone*, used as an insecticide, is a skin irritant; in animal experiments, liver and kidney damage occurs. For *2,4,5,4′-tetrachlorodiphenyl sulfone* (tetradifon, an acaricide), very low toxicity has been found in animal experiments.

3.2. Dimethyl Sulfoxide [178], [181–184]

Dimethyl sulfoxide has very low *acute toxicity*, and chronic effects are observable only at high doses. The lethal dose for humans is estimated to be ca. 10 – 12 mL/kg [177]. The toxic effects of DMSO, such as skin rash and itching after dermal application, hemolysis after intravenous infusion, and gastrointestinal disorders after swallowing, depend strongly on the concentration. DMSO – water mixtures have an irritant effect only at concentrations > 70 % DMSO.

In the mammalian organism, sulfoxides can be reduced to sulfides by tissue or microbial enzyme systems. Enzymatic oxidation to sulfones has also been described, therefore the toxic effect of sulfoxides can arise from both of these redox states.

Teratogenic and *mutagenic activity* has been described for DMSO. DMSO strengthens the action of other toxic substances due to absorption through the skin.

In handling DMSO, attention should be paid to the fact that hazardous chemical reactions can occur in some cases. Explosions have been reported as a result of thermal decomposition and on reaction with organic acid chlorides, nonmetal halides, perchloric acid, perchlorates, nitrates, and sodium hydride.

Only limited data can be found in the literature on the toxicity of other sulfoxides.

4. References

1. *Houben-Weyl,* 4th ed., **9,** 223, **E 11,** 1132.
2. S. Patai, Z. Rappoport, C. J. M. Stirling (eds.): *The Chemistry of Sulphones and Sulphoxides,* J. Wiley & Sons, New York 1988.
3. T. Durst in D. N. Jones (ed.): *Comprehensive Organic Chemistry 3,* Pergamon Press, Oxford 1979.
4. L. A. Paquette, R. V. C. Carr, *Org. Synth.* **64** (1986) 157.
5. Synphar, EP 335 646, 1989 (M. Daneshtala et al.).
6. C. S. Giam, K. Kikukawa, D. A. Trujillo, *Org. Prep. Proced. Int.* **13** (1981) 137. Sumitomo Pharmaceuticals, EP 322 800, 1989 (N. Ohashi, K. Fujimoto).
7. J. Drabowicz, P. Lyzwa, M. Mikolajczyk, *Phosphorus Sulfur* **17** (1983) 169.
8. J. B. Doherty et al., *J. Med. Chem.* **33** (1990) 2513.
9. J. M. Blanco et al., *Tetrahedron: Asymmetry* **3** (1992) 749.
10. K. Peseke, U. Schoenhusen, *J. Prakt. Chem.* **332** (1990) 679. S. T. Purrington, A. G. Glenn, *Org. Prep. Proced. Int.* **17** (1985) 227.
11. A. McKillop, J. A. Tarbin, *Tetrahedron Lett.* **24** (1983) 505. W. Zhou, L. Zhang, X. Xu, Z. Zhang, *Wuji Huaxue Xuebao* **8** (1992) 88; *Chem. Abstr.* **117** (1992) 191 428 n.
12. W. Adam, L. Hadjiarapoglou, *Tetrahedron Lett.* **33** (1992) 469.
13. C. M. Rodriguez, J. M. Ode, J. M. Palazon, V. S. Martin, *Tetrahedron* **48** (1992) 3571.
14. S. W. Kaldor, M. Hammond, *Tetrahedron Lett.* **32** (1991) 5043. W. Priebe, G. Grynkiewicz, *Tetrahedron Lett.* **32** (1991) 7353.
15. Monsanto, US 4 966 731, 1990 (Y. Chou). Monsanto, EP 480 900, 1992 (Y. Chou).
16. B. M. Trost, D. P. Curran, *Tetrahedron Lett.* **22** (1981) 1287.
17. J. Holoch, W. Sundermeyer, *Chem. Ber.* **119** (1986) 269.
18. M. Hirano, J. Tomaru, T. Morimoto, *Bull. Chem. Soc. Jpn.* **64** (1991) 3752.
19. R. Davis, *Synth. Commun.* **17** (1987) 823.
20. J. K. Crandall, C. Pradat, *J. Org. Chem.* **50** (1985) 1327. J. Wildeman, A. M. van Leusen, *Synthesis* 1979, 733.
21. Z.-D. Liu, Z.-C. Chen, *Synth. Commun.* **22** (1992) 1997.
22. A. Borchardt, A. Kopkowski, W. Wasilewski, A. Zakrzewski, *Chem. Abstr.* **118** (1992) 38 529 x.
23. A. Borchardt, H. Janota, A. Zakrzewski, *Chem. Abstr.* **114** (1987) 81 139 z.

24. Z.-Y. Yang, D. J. Burton, *J. Chem. Soc. Perkin Trans. 1,* 1991, 2058.
25. G. Bram, A. Loupy, M. C. Roux-Schmitt, J. Sansoulet, T. Strzalko, J. Seyden-Penne, *Synth.* 1987, 56.
26. G. K. Biswas, S. S. Jash, P. Bhattacharyya, *Indian J. Chem. Sect. B* **29 B** (1990) 491.
27. H. Suzuki, Y. Nishioka, S. J. Padmanabhan, T. Ogawa, *Chem. Lett.* 1988, 727.
28. S. S. Labadie, *J. Org. Chem.* **54** (1989) 2496. J.-L. Parrain, A. Duchene, J.-P. Quintard, *Tetrahedron Lett.* **31** (1990) 1857.
29. L. L. Frye, E. L. Sullivan, K. P. Cusack, J. M. Funaro, *J. Org. Chem.* **57** (1992) 697.
30. K. S. Kim, T. K. Kim, C. S. Hahn, *Chem. Abstr.* **113** (1989) 5812 c.
31. K. Inomata et al., *Chem. Lett.* 1986, 289. K. Inomata et al., *Bull. Chem. Soc. Jpn.* **60** (1987) 1767.
32. J. Barluenga et al., *J. Chem. Soc. Perkin Trans. 1,* 1987, 2605.
33. C. Najera, B. Baldo, M. Yus, *J. Chem. Soc. Perkin Trans. 1,* 1988, 1029.
34. C. J. M. Stirling, *Int. J. Sulfur Chem. Part B* **6** (1971) 277.
35. G. K. Cooper, L. J. Dolby, *J. Org. Chem.* **44** (1979) 3414.
36. J. Fayos, J. Clardy, L. J. Dolby, T. Farnham, *J. Org. Chem.* **42** (1977) 1349. P. Messinger, K. Treudler, *Arch. Pharm. (Weinheim Ger.)* **321** (1988) 441.
37. P. Messinger, *Arch. Pharm. (Weinheim Ger.)* **306** (1973) 458.
38. R. Kerber, J. Starnick, *Chem. Ber.* **104** (1971) 2035.
39. J. M. Bazavova, V. M. Neplyuev, M. O. Lozinskii, *Zh. Org. Khim.* **18** (1982) 865; *Chem. Abstr.* **97** (1982) 91 851.
40. H. Maruyama, T. Hiraoka, *J. Org. Chem.* **51** (1986) 399.
41. K. Bailey, B. R. Brown, B. Chalmers, *J. Chem. Soc. Chem. Commun.* 1967, 618.
42. C. J. M. Stirling, *J. Chem. Soc.* 1964, 5856.
43. J. Sinnreich, M. Asscher, *J. Chem. Soc. Perkin Trans. 1,* 1972, 1543. Y. Amiel, *J. Org. Chem.* **36** (1971) 3697.
44. A. C. Cope, D. E. Morrison, L. Field, *J. Am. Chem. Soc.* **72** (1950) 59. R. W. Hoffmann, W. Sieber, *Justus Liebigs Ann. Chem.* **703** (1967) 96.
45. S. Braverman, T. Globerman, *Tetrahedron* **30** (1974) 3873.
46. K. Hiroi, M. Yamamoto, Y. Kurihara, H. Yonezawa, *Tetrahedron Lett.* **31** (1990) 2619.
47. G. Büchi, R. M. Freidinger, *J. Am. Chem. Soc.* **96** (1974) 3332.
48. M. W. Thomsen, B. M. Handwerker, S. A. Katz, R. B. Belser, *J. Org. Chem.* **53** (1988) 906.
49. K. Kondo, D. Tunemoto, *Tetrahedron Lett.* 1975, 1007.
50. J. W. Lee, D. Y. Oh, *Bull. Korean Chem. Soc.* **12** (1991) 347; *Chem. Abstr.* **115** (1991) 114 077 d.
51. J. W. Lee, D. Y. Oh, *Bull. Korean Chem. Soc.* **10** (1989) 392; *Chem. Abstr.* **112** (1989) 157 769 m. D. B. Reddy et al., *Phosphorus Sulfur Silicon Relat. Elem.* **70** (1992) 325. D. Villemin, A. Ben Alloum, *Synth. Commun.* **21** (1991) 63.
52. E. Dominguez, J. C. Carretero, *Tetrahedron Lett.* **31** (1990) 2487. B. M. Trost, T. A. Grese, *J. Org. Chem.* **56** (1991) 3189.
53. V. M. Naidan, G. D. Naidan, *Zh. Obshch. Khim.* **50** (1980) 2611; *Chem. Abstr.* **94** (1981) 191 834. N. D. Obushak, E. E. Bilaya, N. J. Ganushchak, *Zh. Org. Khim.* **27** (1991) 2372; *Chem. Abstr.* **116** (1991) 235 172 k.
54. P. Robson, P. R. H. Speakman, D. G. Stewart, *J. Chem. Soc. C* 1968, 2180.
55. W. E. Truce, D. P. Tate, D. N. Burdge, *J. Am. Chem. Soc.* **82** (1960) 2872.
56. B. M. Trost, H. C. Arndt, P. E. Strege, T. R. Verhoeven, *Tetrahedron Lett.* 1976, 3477.
57. E. J. Corey, M. Chaykovsky, *J. Am. Chem. Soc.* **87** (1965) 1345.
58. T. S. Chou, M. L. You, *J. Org. Chem.* **52** (1987) 2224.
59. X. Huang, H. Zhang, *Synth. Commun.* **19** (1989) 97.
60. M. Fujii et al., *Bull. Chem. Soc. Jpn.* **61** (1988) 495.
61. T. Mandai et al., *Tetrahedron Lett.* **24** (1983) 4993. C. Herve Du Penhoat, M. Julia, *Tetrahedron* **42** (1986) 4807.
62. M. Julia, B. Badet, *Bull. Chem. Soc. Chim. Fr.* 1976, 525.
63. L. A. Paquette, *Org. React. (N. Y.)* **25** (1977) 1. D. Scarpetti, P. L. Fuchs, *J. Am. Chem. Soc.* **112** (1990) 8084.
64. R. D. G. Cooper, *J. Am. Chem. Soc.* **94** (1974) 1018.
65. W. E. Truce, E. J. Madaj, Jr., *Sulfur Rep.* **3** (1983) 259.
66. Nissan Chem. Ind., JP 63 216 860, 1987 (K. Ogura, T. Fujimo). Kurakay, EP 282 915, 1988 (T. Onishi et al.). Hoffmann-La Roche, EP 298 404, 1988 (K. Bernhard, S. Jaggli, P. Kreienbuhl, U. Schwieter). Roussel Uclaf, EP

348 254, 1989 (D. Babin, J. P. Demoute, J. Tessier).
67. Bayer, EP 418 664, 1990 (M. Hoppe, K. Herd, H. Henk, F. Stoehr). BASF, EP 492 236, 1991 (C. Marschner et al.). Hoechst, EP 489 360, 1991 (J. Dannheim).
68. Abbott Laboratories, EP 218 095, 1986 (D. S. Kenney, J. C. Kane, B. N. Devisetty).
69. Bayer, EP 486 798, 1991 (W. Brandes et al.). Duphar Int. Res., EP 301 613, 1989 (H. Dolman, J. Kuipers).
70. Hoechst, WO 92/11 237, 1991 (P. Braun, H. Mildenberger, B. Sachse, F. Zurmühlen). Bayer, EP 205 076, 1986 (C. Fest, W. Brandes, G. Hänssler, P. Reinecke).
71. Rhône-Poulenc, EP 351 332, 1989 (P. Desbordes, M. Euvrard). Sumitomo, JP 63 208 564, 1987 (O. Kirino et al.).
72. Abbott Laboratories, EP 258 878, 1987 (A. J. Crovetti, R. A. Smith, B. E. Melin, F. M. H. Casati). ICI, EP 273 549, 1987 (R. A. E. Carr, M. J. Bushell).
73. Nisshin Flour Milling, DE 3 620 632, 1986 (T. Takahashi, K. Hagihara, Y. Suzuki).
74. Sterling, EP 207 375, 1986 (R. G. Christians, M. R. Bell, J. L. Herrmann, C. J. Opalka). Sterling, US 5 053 405, 1990 (G. M. Pilling, J. P. Mallamo).
75. ICI, WO 90/08 761, 1990 (D. R. Brittain et al.).
76. Smithkline Beecham, WO 88/07 858, 1988 (W. E. Bondinell, H. S. Ormsbee). Tanabe Seyaku, EP 270 091, 1987 (Y. Honma et al.).
77. Merck Frosst Canada, EP 288 202, 1988 (H. W. R. Williams, R. N. Young).
78. Research Corporation Technologies, WO 92/06 683, 1992 (F. W. Wassmundt).
79. Teijin, EP 338 796, 1989 (A. Hazata et al.).
80. Pennwaldt, EP 225 981, 1985 (L. K. Huber, J. L. Reilly).
81. Bayer, DE 3 921 691, 1989 (W. Broda et al.).
82. Konica, JP 02 110 544, 1988 (M. Nishizeki, N. Tachibana, N. Kagawa).
83. Kodak, WO 90/13 852, 1990 (M. Crawley).
84. Wako Pure, EP 440 375, 1991 (F. Urano, M. Nakahata, H. Fujie, K. Oono).
85. Du Pont, WO 91/08 198, 1990 (L. T. Cheng, A. E. Feiring, W. Tam).
86. Monsanto, EP 267 175, 1987 (D. R. Dyrott, D. P. Getman, J. K. Glascock).
87. Henkel, WO 91/06 534, 1989 (A. Meffert et al.).
88. J. Drabowicz, M. Mikolajczyk, *Org. Prep. Proced. Int.* **14** (1982) 45.
89. M. Madesclaire, *Tetrahedron* **42** (1986) 5459.
90. O. Hinsberg, *Ber. Dtsch. Chem. Ges.* **43** (1910) 289.
91. G. Schill, P. R. Jones, *Synthesis* 1974, 117.
92. A. Ceruiani, G. Modena, P. E. Todesca, *Gazz. Chim. Ital.* **90** (1970) 383.
93. Nihon Tokushu Noyaku Seizo K. K., EP 91 052, 1983 (J. Saito, K. Shiokawa, T. Takemoto).
94. J. Drabowicz, P. Lyzwa, M. Popielarczyk, M. Mikolajczyk, *Synthesis* 1990, 937.
95. Y. Watanabe, T. Numata, S. Oae, *Synthesis* 1981, 204.
96. J. Drabowicz, M. Mikolajczyk, *Synthesis* 1978, 758.
97. K. S. Kim, H. J. Hwang, C. S. Cheong, C. S. Hahn, *Tetrahedron Lett.* **31** (1990) 2893.
98. F. E. Hardy, R. P. H. Speakman, P. Robson, *J. Chem. Soc. C* 1969, 2334.
99. R. Balicki, L. Kaczmarek, P. Nantka-Namirski, *Liebigs Ann. Chem.* 1992, 883.
100. L. Bateman, K. R. Hargrave, *Proc. Roy. Soc. London Ser. A* **224** (1954) 389.
101. P. Leriverend, M.-L. Leriverend, *Synthesis* 1987, 587.
102. Taiho Pharmaceutical Co., JP-Kokai 58 222 064, 1983 (S. Yasumoto et al.); *Chem. Abstr.* **100** (1984) 209 381 c.
103. P. Laur in A. Senning (ed.): *Sulfur in Organic and Inorganic Chemistry,* **vol. 3,** Marcel Dekker, New York 1971, p. 203.
104. B. Rajanikanth, B. Ravindranath, *Indian J. Chem. Sect. B* **23 B** (1984) 877.
105. A. L. Ternay, Jr., L. Ens, J. Herrmann, S. Evans, *J. Org. Chem.* **34** (1969) 940.
106. S. Florio, J. L. Leng, C. J. M. Stirling, *J. Heterocyclic Chem.* **19** (1982) 237. D. E. Beattie, R. Crossley, K. H. Dickinson, G. M. Dover, *Eur. J. Med. Chem. Chim. Ther.* **18** (1983) 277.
107. F. H. Walker, US 4304-916, 1981; *Chem. Abstr.* **96** (1982) 122 654 u.
108. R. Kaya, N. R. Beller, *J. Org. Chem.* **46** (1981) 196.
109. F. Freeman, C. N. Angeletakis, *J. Org. Chem.* **50** (1985) 793.
110. J. A. Webber, E. M. van Heyningen, R. T. Vasileff, *J. Am. Chem. Soc.* **91** (1969) 5674.
111. C. Märcker, *Justus Liebigs Ann. Chem.* **136** (1865) 891.
112. F. Gasparini et al., *Tetrahedron* **39** (1983) 3181. F. Gasparini et al., *J. Org. Chem.* **55** (1990) 1323.
113. R. Low, H. P. W. Vermeeren, J. J. A. van Asten, W. J. Ultee, *J. Chem. Soc. Chem. Commun.* 1976, 496.

114. Y. Nagano et al., *Tetrahedron Lett.* 1977, 1345.
115. T. L. Ho, C. M. Wong, *Synthesis* 1972, 562.
116. T. L. Ho, *Synth. Commun.* **9** (1979) 237.
117. E. Baciocchi, A. Piermattei, R. Ruzziconi, *Synth. Commun.* **18** (1988) 2167.
118. C. R. Johnson, J. E. Keiser, *Org. Synth.* **46** (1966) 78.
119. K. T. Liu, Y. C. Tong, *J. Org. Chem.* **43** (1978) 2717.
120. E. Santaniello, A. Manzocchi, C. Farachi, *Synthesis* 1980, 563.
121. K. R. Roh, K. S. Kim, Y. H. Kim, *Tetrahedron Lett.* **32** (1991) 793.
122. H. E. Folsom, J. Castrillion, *Synth. Commun.* **22** (1992) 1799.
123. T. Kageyama, Y. Ueno, M. Okawara, *Synthesis* **1983,** 815. Nippon Silica Industrial Co., JP-Kokai 2 040 354, 1990 (T. Morimoto); *Chem. Abstr.* **113** (1990) 39 528 e.
124. S. Kajigaeshi, K. Murakawa, S. Fujisaki, T. Kakinami, *Bull. Chem. Soc. Jpn.* **62** (1989) 3376. Ube Industries, JP-Kokai 2 268 150, 1990 (S. Kajisori); *Chem. Abstr.* **114** (1990) 184 990 z.
125. Zambon S.p.A., EP 125 654, 1984 (M. Meneghin); *Chem. Abstr.* **102** (1985) 95 392 r.
126. R. P. Greenhalgh, *Synlett* 1992, 235.
127. G. J. Quallich, J. W. Lackey, *Tetrahedron Lett.* **31** (1990) 3685.
128. K. K. Laali, D. S. Nagvekar, *J. Org. Chem.* **56** (1991) 1867.
129. R. M. Coates, H. D. Pigott, *Synthesis* 1975, 319. H. J. Monteiro, J. P. de Souza, *Tetrahedron Lett.* 1975, 921.
130. E. Block, J. O'Connor, *J. Am. Chem. Soc.* **96** (1974) 3929. D. N. Jones, P. D. Cottam, J. Davis, *Tetrahedron Lett.* 1979, 4977.
131. H. B. Kagan, *Phosphorus Sulfur* **27** (1986) 127.
132. M. Trost, K. K. Leung, *Tetrahedron Lett.* 1975, 4197. G. A. Russell, L. A. Ochrymowycz, *J. Org. Chem.* **34** (1969) 3624.
133. S. Iriuchijima, G. Tsuchihashi, *Synthesis* 1970, 588.
134. Y. H. Kim, S. C. Lim, H. R. Kim, D. C. Yoon, *Chem. Lett.* 1990, 79.
135. N. Furukawa et al., *Tetrahedron Lett.* **27** (1986) 3899. T. Shibutani, H. Fujihara, N. Furukawa, S. Oae, *Heteroat. Chem.* **2** (1991) 521.
136. Wako Pure Chem. Ind., JP 62 212 370, 1986 (N. Furukawa, T. Fujiwara, N. Shibuya); JP 64 003 169, 1987 (N. Furukawa, T. Fujiwara, N. Shibuya).
137. T. Kawai et al., *Phosphorus Sulfur* **34** (1987) 139.
138. S. Oae, T. Kawai, N. Furukawa, *Phosphorus Sulfur* **34** (1987) 123.
139. Wako Pure Chem. Ind., JP 03 024 058, 1989 (S. Wakabayashi, J. Uenishi, S. Daikyo).
140. P. Bakuzis, M. L. F. Bakuzis, *J. Org. Chem.* **50** (1985) 2569.
141. K. Ogura, M. Yamashita, G. Tsuchihashi, *Synthesis* 1975, 385.
142. H. C. Brown, N. Ravindran, *Synthesis* 1973, 506.
143. K. Nagasawa, A. Yoneta, T. Umezawa, K. Ito, *Heterocycles* **26** (1987) 2607.
144. R. A. Amos, *J. Org. Chem.* **50** (1985) 1311.
145. Chem. AG Bitterfeld, DE 4 032 414, 1990 (B. Biber et al.). ICI, EP 273 549, 1987 (R. A. E. Carr, M. J. Bushell).
146. Suntory, JP 3 197 483, 1989 (M. Ishiguro, R. Tanaka). ICI, EP 390 394, 1990 (F. F. Morpeth, M. Greenhalgh). BASF, DE 3 328 770, 1983 (H. Graf et al.).
147. Bayer, EP 119 457, 1984 (H. G. Schmitt, K. Lürssen, K. Wedemeyer).
148. Elf Aquitaine, EP 209 462, 1986 (J. C. Gautier, J. Komornicki, J. Foix, G. Pastor). Richter Gedeon, DE 3 506 998, 1985 (J. Kreidl et al.).
149. Hoechst, EP 298 440, 1988 (H. J. Lang, K. Weidmann, A. W. Herling).
150. Merck, DE 3 621 112, 1986 (R. Jonas, H. Wurziger, J. Pinlats, M. Klockow).
151. Teijin, JP 59 175 465, 1983 (A. Hasato, T. Tanaka, S. Kurozumi).
152. Mitsubishi, JP 3 232 853, 1990 (H. Okujima et al.).
153. Hoffmann-La Roche, US 4 973 599, 1989 (N. W. Gilman, W. Y. Chen).
154. Boehringer Biochemia Robin, EP 471 709, 1990 (E. Alessio et al.).
155. Boehringer, EP 232 820, 1987 (E. Reinholz et al.).
156. G. M. Rosen, US 4 314 989, 1982.
157. Univ. Iowa Res. Foundation, US 4 482 568, 1984 (B. V. Plapp, V. K. Chadha).
158. Merck, US 4 617 301, 1984 (A. A. Patchett, M. J. Wyvratt). Merck, US 4 670 470, 1985 (R. A. Firestone).
159. Faiman, US 5 153 219, 1991 (M. D. Faiman, B. W. Hart, A. Madan).
160. Hoffmann-La Roche, US 5 061 733, 1990 (G. F. Bryce, S. S. Shapiro).
161. Sagami Chem. Res. Centre, JP 2 036 131, 1988 (K. Matsui, T. Aoyanagi, K. Suzuki, M. Yamamura).
162. L'Oreal, US 4 888 164, 1986 (J. F. Grollier, J. Maignan).

163. Ciba Geigy, EP 473 549, 1991 (H. R. Meier, P. Dubs).
164. Henkel, DE 4 021 694, 1990 (G. Stoll, P. Daute, J. Wegener, F. Berger).
165. Lubrizol, EP 404 903, 1989 (S. T. Jolley, M. F. Salomon).
166. Fuji Photo Film, JP 59 180 559, 1983 (T. Kamio, K. Aoki, J. Arakawa).
167. L. Field, *Synthesis* 1972, 101.
168. P. E. Correa, D. P. Riley, *J. Org. Chem.* **50** (1985) 1787.
169. D. Martin, A. Weise, H.-J. Niclas, *Angew. Chem.* **79** (1967) 340. W. O. Ranky, D. C. Nelson in N. Kharash (ed.): *Organic Sulfur Compounds,* **vol. 1,** Pergamon Press, New York 1961, p. 170. A. J. Parker, *Q. Rep. Sulfur Chem.* **3** (1968) 185.
170. S. W. Jacob, R. Herschler, *Cryobiology* **23** (1986) 14. C. F. Brayton, *Cornell Vet.* **76** (1986) 61. B. N. Swanson, *Rev. Clin. Basic Pharmacol.* **5** (1985) 1.
171. Rhoderton, EP 320 271, 1988 (J. M. Vichipascu).
172. N. G. Leveskis, US 4 512 337, 1981.
173. Boehringer, DE 1 642 289, 1967 (H. Hildebrand).
174. Kemer Agric. Res. Inst., SU 1 386 071, 1985 (N. L. Voronova, V. J. Kozyakov, A. M. Yufferrov).
175. B. A. Trofimov, O. V. Petrova, A. M. Vasil'tsov, A. J. Mikhaleva, *Izv. Akad. Nauk SSSR Ser. Khim.* 1990, 1601.
176. G. A. Russell, S. A. Weiner, *J. Org. Chem.* **31** (1966) 248. J. K. Stamos, *Tetrahedron Lett.* **26** (1985) 2787.
177. W. Braun, A. Dönhardt: *Vergiftungsregister,* Thieme Verlag, Stuttgart 1982.
178. A. G. Renwick in L. A. Damani (ed.): *Sulphur- Containing Drugs and Related Organic Compounds,* **vol. 1,** Part B, Ellis Horwood Limited, Chichester 1989, p. 133.
179. Industrieverband Pflanzenschutz e.V. (ed.): *Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmitteln,* Pressehaus Bintz-Verlag, Offenbach 1982.
180. M. V. Cone: *Chemical Hazard Information Profile-Draft Report. Sulfolane. Office of Toxic Substances,* US EPA, Washington D.C., 1984.
181. *Ullmann,* 4th ed., **22,** 334.
182. C. C. Willhite, P. J. Katz, *JAT J. Appl. Toxicol.* **4** (1984) 155.
183. L. F. Rubin, *Ann. New York Acad. Sci.* **411** (1983) 6.
184. L. Roth, U. Weller: *Gefährliche chemische Reaktionen,* 13. + 14. Ergänzung, ecomed 1992.

Sulfonic Acids, Aliphatic

KURT KOSSWIG, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany

1. Properties and Uses 1
2. Production 3
3. References 4

1. Properties and Uses

Alkanesulfonic Acids. In alkanesulfonic acids RSO_3H, a hydroxysulfonyl group is bonded via the sulfur atom to a carbon atom of an alkyl group. Alkanesulfonic acids are strong acids, and the lower members up to heptane-1-sulfonic acid are completely dissociated in dilute aqueous solution. Estimated pK values in aqueous solution of some other strong acids are compared with those of methane-, ethane-, and propanesulfonic acids below [5]:

$C_3H_7SO_3H$	− 1.5
$C_2H_5SO_3H$	− 1.7
CH_3SO_3	− 1.9
p-$CH_3C_4H_4SO_3H$	− 2.7
$C_6H_5SO_3H$	− 2.8
$HOSO_3H$	− 3.0
CH_3OSO_3H	− 3.5
p-$O_2NC_4H_4SO_3H$	− 3.8
CF_3SO_3H	− 5.5
FSO_3H	− 5.6

Methanedisulfonic acid (methionic acid) [*503-40-2*], mp 96 – 100 °C, has the same strength as sulfuric acid.

From heptane-1-sulfonic acid onwards the sulfonic acids increasingly associate in aqueous solution, forming micelles [6]. Pure alkanesulfonic acids attack stainless steel, but the presence of 0.5 to 1 % sulfuric acid hinders corrosion.

The physical properties of some *n*-alkanesulfonic acids are listed in Table 1. Alkanesulfonic acids decompose on distillation at normal pressure, but can be distilled without decomposition under vacuum.

Alkanesulfonic acids darken when stored in the absence of light. They form hydrates, are hygroscopic, and dissolve readily in water. The stable alkali metal, ammonium, alkaline-earth metal, and lead salts also have high water solubility and can be recrystallized from alcohol.

Alkanesulfonic acids with nonterminal sulfonic acid groups are less stable than the alkane-1-sulfonic acids, in particular 2-methylpropane-2-sulfonic acid [*16794-13-1*], *mp* 113 °C [7].

Lower molecular mass alkanesulfonic acids, especially methanesulfonic acid, being strong, nonoxidizing acids, are used as catalysts for esterification, polymerization, and alkylation, and for the production of peroxycarboxylic acids from hydrogen peroxide and carboxylic acids [8]. Their esters are used as alkylating agents. Short- chain alkanesulfonic acids such as butanesulfonic acid, have been proposed as components of industrial cleaning agents, as they can remove both organic and mineral contamination [9].

Perfluorinated Alkanesulfonic Acids [10], [11]. Perfluoroalkanesulfonic acids, which can be obtained by electrochemical fluorination of alkanesulfonic acids, followed by hydrolysis of the resulting sulfonyl fluorides, are the strongest organic acids and are even stronger than sulfuric acid. Perfluoroalkanesulfonic acids are thermally very stable and can be distilled at normal pressure (Table 2). They form crystalline hydrates.

Table 2. Boiling points of some perfluoroalkanesulfonic acids at normal pressure

	CAS no.	*bp*, °C
Perfluoromethanesulfonic acid	[*1493-13-6*]	162
Perfluoroethanesulfonic acid	[*354-88-1*]	178
Perfluoropropane-1-sulfonic acid	[*423-41-6*]	196
Perfluorobutane-1-sulfonic acid	[*375-73-5*]	210 – 212
Perfluoropentane-1-sulfonic acid	[*2706-91-4*]	224 – 226
Perfluorohexane-1-sulfonic acid	[*355-46-4*]	238 – 239

Perfluoroalkanesulfonate anions are good leaving groups [12]. Esters of perfluoroalkanesulfonic acids are therefore used as strong alkylating agents in preparative chem-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_503

Table 1. Physical data of some alkane-1-sulfonic acids

Acid	CAS no.	*bp*/ hPa, °C	*mp*, °C	d_4^{25}
Methanesulfonic acid	[*75-75-2*]	122/1.3	20	1.4844
Ethanesulfonic acid	[*594-45-6*]	123/1.3	− 17	1.3341
Propane-1-sulfonic acid	[*5284-66-2*]	136/1.3	7.5	1.2516
Butane-1-sulfonic acid	[*2386-47-2*]	147/0.9	− 15.2	1.1906
Pentane-1-sulfonic acid	[*35452-30-3*]	163/1.3	15.9	1.1226
Hexane-1-sulfonic acid	[*13595-73-8*]	174/1.3	16.1	1.1047

istry. Trifluoromethanesulfonates are known as triflates and nonafluorobutanesulfonates as nonaflates.

Hydroxyalkanesulfonic Acids. Some hydroxyalkanesulfonic acids and their derivatives are important in industry.

The sodium salts of 1-hydroxyalkane-1-sulfonic acids are readily crystallizable compounds which are formed from aldehydes and sodium bisulfite in neutral solution. In an acidic or alkaline medium they decompose into their components. They are used for the sulfonalkylation of compounds with an active hydrogen atom, in particular sodium hydroxymethanesulfonate (from formalin and sodium bisulfite), for example:

$$RNH_2 + CH_2O + NaHSO_3 \longrightarrow RNH{-}CH_2SO_3Na + H_2O$$

Isethionic acid, 2-hydroxyethanesulfonic acid [*107-36-8*], is used as an intermediate in numerous syntheses. It is generally used in the form of its readily accessible stable sodium salt [*1562-00-1*], *mp* 192 – 194 °C, for example:

$$RCOCl + HOCH_2CH_2SO_3Na \longrightarrow RCOOCH_2CH_2SO_3Na + HCl$$

Isethionic acid is the starting material for the industrial synthesis of taurine (2-aminoethanesulfonic acid [*107-35-7*]), *mp* 317 °C dec., and *N*-methyltaurine (2-*N*-methylaminoethanesulfonic acid [*107-68-6*]), *mp* 245 °C. Taurine is a naturally occurring substance, formed in organisms by oxidation of cysteine, and is a building block of taurocholic acid. The derivatives of isethionic acid, taurine, and *N*-methyltaurine, especially their fatty acid derivatives, are important wetting, washing, and dispersion agents in body and household cleaning agents, and also in textile processing.

1,3-Propanesultone [*1120-71-4*] [13], *mp* 31 °C, *bp* 150 °C (13 hPa), d_4^{40} 1.392, the internal anhydride of 3-hydroxypropane-1-sulfonic acid [*15909-83-3*] is a strong alkylating agent, for example:

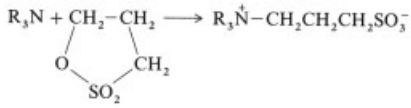

Reaction products of propanesultone with alcohols, phenols, thiols, sulfides, and amides have been proposed as tensides, leather auxiliaries, antistatic agents, and bacteriostatic agents. Cellulose becomes soluble when treated with propanesultone. Polyamide fibers treated with propanesultone can be dyed more easily. However, since propanesultone has been shown to be a strong carcinogen, it is no longer important as a synthon.

Vinylsulfonic acid, ethenesulfonic acid [*1184-84-5*] [14], *bp* 95 °C (1 Pa), is derived from isethionic acid and can only be kept below 0 °C, as it polymerizes rapidly. It is preferable to use its sodium salt in aqueous solution or its esters [e.g., methyl ester, *bp* 54 – 60 °C (30 – 50 Pa), n_D^{20} 1.4310]. The sodium salt [*3039-83-6*] or the esters of vinylsulfonic acid add to compounds with activated hydrogens (sulfonethylation) for example:

$$NaHSO_3 + CH_2{=}CHSO_2OR \longrightarrow NaO_3SCH_2CH_2SO_2OR$$

The addition products with alcohols, phenols, thiols, and amines have been proposed as plasticizers for poly(vinyl chloride), as cross-linking agents, emulsifiers, microbicides, fungicides,

and fungistatic agents. Vinylsulfonic acid and its salts are also suitable monomers for polymerization reactions [15], [16]. Esters of vinylsulfonic acid are powerful alkylating agents: phenol is converted smoothly into anisole with methyl vinylsulfonate in alkaline solution, even at room temperature.

Allylsulfonic acid, 2-propene-1-sulfonic acid [*1606-80-8*], *bp* 180 °C (0.1 Pa), is unstable and is preferably used in the form of its sodium salt [*2495-39-8*] in aqueous solution as a comonomer for polymerizations [15].

Methallylsulfonic acid, 2-methyl-1-propene-3-sulfonic acid [*3934-16-5*], is also predominantly used in the form of its aqueous sodium salt [*1561-92-8*] as a comonomer in polymerizations. It improves, for example, the dyeability of fibers, especially polyacrylonitrile fibers.

2. Production

Alkanesulfonic Acids. Alkanes do not react specifically with sulfur trioxide or oleum, alkanesulfonic acids being among a number of products. Methanesulfonic acid and sulfur trioxide give methionic acid as the only product in good yields. Starting from alkanes, alkanesulfonic acids are smoothly obtained by sulfoxidation or sulfochlorination [17] (see → Surfactants, Chap. 6.2.3.). However, mixtures of positional isomers are formed in these reactions.

Uniform, defined alkanesulfonic acids are obtained by Strecker synthesis from alkyl halides, preferably alkyl bromides, and alkali metal or ammonium sulfite, or by reaction of the corresponding Grignard reagent with sulfur dioxide and subsequent oxidation. The oxidation of thiols with nitric acid, potassium permanganate, or hydrogen peroxide/glacial acetic acid is frequently used, as is the oxidation of thiocyanates and disulfides, all of which are obtained from alkyl halides.

Starting from 1-alkenes the sodium salts of alkane-1-sulfonic acids are obtained by the addition of sodium bisulfite in a radical-initiated reaction [9], [18].

2-Hydroxyalkane-1-sulfonic acids, especially isethionic acid, are produced by the reaction of 1,2-epoxides (e.g., ethylene oxide) with aqueous sodium bisulfite. The reaction of ethylene with sulfur trioxide gives carbyl sulfate (*mp* 107 – 108 °C), which reacts with water to give ethionic acid [*461-42-7*] (sulfuric acid half ester of 2-hydroxyethane-1-sulfonic acid), and on heating to give isethionic acid [*107-36-8*] [19]:

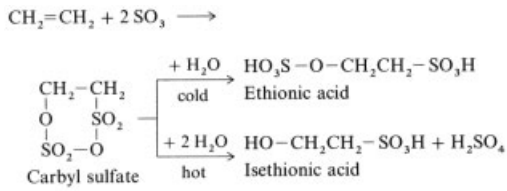

Ethionic acid is also obtained by the reaction of sulfur trioxide with ethanol.

The methyl-substituted carbyl sulfate, obtained analogously from propene and sulfur trioxide, gives 1-propene-1-sulfonic acid [*97141-05-4*] on decomposition.

1,3-Propanesultone is obtained by elimination of water from 3-hydroxypropane-1-sulfonic acid [20], whose sodium salt is formed by radical-initiated addition of sodium bisulfite to allyl alcohol [21]:

$$HOCH_2CH=CH_2 + NaHSO_3 \longrightarrow HOCH_2CH_2CH_2SO_3Na$$

3-Hydroxypropane-1-sulfonic acid is obtained on acidification.

Vinylsulfonic Acid. The alkaline hydrolysis of carbyl sulfate or ethionic acid gives the vinylsulfonate salt directly:

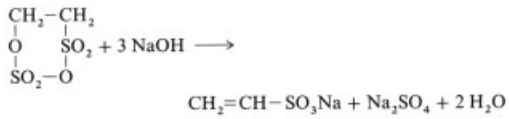

Sodium vinylsulfonate can also be produced by dehydrochlorination of sodium 2-chloroethane-1-sulfonate with aqueous sodium

hydroxide. The chloroethanesulfonate is produced by sulfochlorination of ethyl chloride and subsequent hydrolysis of the 2-chloroethane-1-sulfonyl chloride, or by treatment of sodium bisulfite with vinyl chloride [14].

Allylsulfonic Acid. Only the sodium salt of allylsulfonic acid is used in industry. It is obtained by treatment of allyl chloride with aqueous sodium bisulfite [22].

Methallylsulfonic Acid. Dimethylcarbyl sulfate cannot be obtained from isobutene and sulfur trioxide, as the methallylsulfonic acid reacts further to give methallyldisulfonic acid:

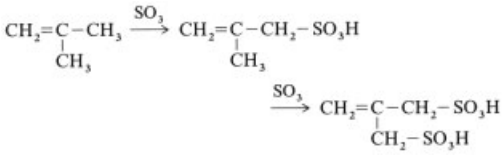

The formation of the latter can be suppressed by complexation of the sulfur trioxide, for example, with *N*,*N*-dialkylcarboxylic acid amides [23]. However, in industry methallylsulfonate is preferably produced from methallyl chloride and sodium sulfite [24].

3. References

General References

1. *Beilstein,* 324 – 329.
2. *Houben-Weyl,* **9,** 343 – 427; E **11,** 1055 ff.
3. *Kirk-Othmer,* **22,** 45 ff.
4. S. Patai, Z. Rapoport: *The Chemistry of Sulphonic Acids, Esters and their Derivatives,* J. Wiley & Sons, Chichester 1991.

Specific References

5. J. F King in [4], p. 251.
6. E. C. McBain, W. B. Dye, S. A. Jonston, *J. Am. Chem. Soc.* **61** (1939) 3210.
7. F. Asinger, P. Laue, B. Fell, G. Gubelt, *Chem. Ber.* **100** (1967) 1696.
8. L. S. Silbert, E. Siegel, D. Swern, *J. Org. Chem.* **27** (1962) 1336.
9. A. Behr, N. Döring, *Fat Sci. Technol.* **92** (1990) 335.
10. H. C. Fielding in R. E. Banks (ed.): *Organofluorine Chemicals and their Industrial Applications,* Ellis Harwood Ltd., Chichester 1979, pp. 214 – 234.
11. W.-Y. Huang, Q.-Y. Chen in [4], pp. 903 – 946.
12. P. J. Stang, M. Hanack, R. Subramanian, *Synthesis* **1982,** 85.
13. R. F. Fischer, *Ind. Eng. Chem.* **56** (1964) no. 3, 41.
14. H. Distler, *Angew. Chem.* **77** (1965) 291.
15. D. S. Breslow, *Encycl. Polym. Sci. Eng.* **6** (1986) 564.
16. D. A. Kanglas in R. H. Yocum, E. B. Nyquist (eds): *Funtional Monomers,* vol. 1, Marcel Dekker, New York 1974, pp. 489 – 640.
17. F. Asinger: *Die Petrolchemische Industrie,* Akademie Verlag, Berlin 1971, pp. 731 – 761.
18. F. Asinger, in [17], p. 757.
19. G. H. Weinreich, M. Jufresa, *Bull. Soc. Chim. Fr. Mem.* **5** (1965) 787.
20. D. K. Tolopko, E. S. Pugasch, O. E. Baranovskaja, O. I. Romanyuk, *Visn. L'viv Politekh Inst.* **181** (1984) 137; *Chem. Abstr.* **101** (1984) 90801 t.
21. *Chem. Abstr.* **104** (1986) 129455 x. O. E. Baranovskaya, A. Yu Shuter, L. Ratkovskaya, in [20] 191, (1985) 116.
22. R. Buening et al., DD 282 819, 1988; *Chem. Abstr.* **114** (1991) 186 306 e
23. Lion Corp., JP-Kokai 63 51 368, 1986 (O. Hizawa, M. Yoshida, M. Takahashi); *Chem. Abstr.* **109** (1988) 169 858 r
24. Hüls AG, DE 3 337 103, 1983 (D. J. Mueller, S. Austin); *Chem. Abstr.* **103** (1985) 106612 m.

Sulfoxidation → Sulfonic Acids, Aliphatic
Sulfoxides → Sulfones and Sulfoxides

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Sulfur

Standard Article

Wolfgang Nehb[1] and Karel Vydra[2]
[1]Lurgi AG, Frankfurt/Main, Germany
[2]Frankfurt/Main, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a25_507.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Historical Background
3. Properties
3.1. Physical State
3.2. Viscosity
3.3. Other Physical Properties
3.4. Chemical Properties
4. Raw Materials for Sulfur Production
4.1. Elemental Sulfur Resources
4.2. Sulfide Ores
4.3. Fossil Fuels
4.4. Sulfates
5. Production from Native Deposits
5.1. Flotation
5.2. Melting Out
5.3. Distillation
5.4. Flotation with Melting Out
5.5. The Frasch Process
5.5.1. Structure of “Fraschable” Sulfur Deposits
5.5.2. Outline of the Process
5.5.3. Sulfur Purity
5.5.4. Energy Consumption
5.5.5. The U.S. Frasch Industry
5.5.6. The Mexican Frasch Industry
6. The Claus Process
6.1. Occurrence of Hydrogen Sulfide
6.2. Hydrogen Sulfide Removal from Gases
6.3. History of the Claus Process
6.4. Chemistry of the Claus Process
6.5. Claus Plant Configurations
6.6. The Claus - Oxygen Process
6.7. Troubleshooting
6.8. Optimization of Sulfur Recovery Units by Use of Analyzers
7. Increased Sulfur Recovery
7.1. Selective Oxidation
7.1.1. The Superclaus Process, the Euroclaus® Process
7.1.2. The Recycle Selectox Process

7.1.3. The Once-Through Selectox Process
7.2. Tail Gas Treatment
7.2.1. Sub-Dewpoint Processes
7.2.1.1. The Sulfreen, Hydrosulfreen, and Doxosulfreen Processes
7.2.1.2. The CBA Process
7.2.1.3. The MCRC Process
7.2.2. Processes Working on Reduced Tail Gases
7.2.2.1. The SCOT and Super-SCOT Processes
7.2.2.2. The Sulften Process
7.2.2.3. Solvents for Regenerative H_2S Removal
7.2.2.4. The BSR/Selectox Process
7.2.3. Processes Working on Oxidized Tail Gas
7.2.3.1. The Wellman - Lord Process
7.2.3.2. The Clintox Process
7.2.4. Biodesulfurization; Thiopaq Process
7.3. Purification
7.3.1. Arsenic and Selenium Removal
7.3.2. Melting of Stretford Sulfur
7.3.3. Degassing of Claus Sulfur
7.4. Sulfur Management; the OmniSulf Process
8. Gas-Phase Reduction of Sulfur Dioxide
8.1. The Allied Chemicals Process
8.2. Other Processes with Gas as Reducing Agent
8.3. Processes with Coke or Coal as Reducing Agent
9. Commercial Grades and Forms of Sulfur
9.1. Quality
9.2. Commercial Forms for Bulk Transport
9.2.1. Dry Bulk Sulfur
9.2.2. Liquid Sulfur
9.2.3. Formed Sulfur
9.2.3.1. Slating
9.2.3.2. Wet Pelletizing (Water Prilling)
9.2.3.3. Air Prilling
9.2.3.4. Other Prilling Processes
9.2.3.5. Drum Granulation: The Procor GX Process
9.2.3.6. Spouted Bed Granulation
9.3. Special Forms and Grades
10. Uses
10.1. Agriculture
10.2. Construction Materials
10.3. Sulfur-Containing Batteries
10.4. Other Potential Uses

[Top of Page]

1. Introduction

Sulfur is the fifteenth most common terrestrial element, and the ninth most abundant element in the universe. It is widely distributed in nature in different compounds as well as in pure form. The mean sulfur content of the lithosphere is estimated to be ca. 0.05 %. In massive elemental form, sulfur is often referred to as brimstone.

Sulfur is of great industrial importance, though relatively little (ca. 10 %) is used in the elemental form, in the production of carbon disulfide, rubber, pesticides, pharmaceuticals, and cosmetics. By far the greatest part is used in the form of sulfuric acid and sulfur dioxide.

Elemental sulfur is recovered from volcanic and other surface deposits by a number of processes.

The sources of sulfur for the production of sulfuric acid and sulfur dioxide are often sulfur compounds produced as secondary products, e.g., as sulfur dioxide from the processing of sulfide ores of nonferrous metals.

Elemental sulfur is often obtained via hydrogen sulfide during the purification of sulfur- containing natural gas or raffination

of crude oil; after conversion of H_2S to sulfur, the latter is stored or transported to sulfuric acid plants or to purification; it is easier, safer, and cheaper to store and transport sulfur than sulfuric acid or sulfur dioxide.

Because of the increasing production of sulfur-containing byproducts, the relative contribution of mined brimstone production is decreasing: from 47 % in 1979 to 37 % in 1989.

Huge amounts of sulfur compounds contained in coal are converted in thermal power stations to sulfur dioxide which must be recovered for environmental reasons; as the product is highly dilute, it is very difficult to recover the sulfur economically in usable forms, such as the element, sulfuric acid, or concentrated sulfur dioxide; it is easier to remove it as calcium or sodium sulfate or sulfite, even though the major part cannot be marketed and has to be disposed of as a waste product.

Figure 1 shows the principal sulfur sources and processing routes. Table 1 summarizes sulfur production in the major regions of the world.

Table 1. Production of sulfur in 1992 (10^3 t sulfur)

Region	All forms	Brimstone
Europe including Turkey, without former Soviet Union	11 461	6 845
Former Soviet Union	8 331	5 181
Africa	866	145
North America without Mexico	19 045	15 904
Latin America	2 788	1 879
Middle East	4 140	4 140
South and East Asia	10 441	2 202
Oceania	300	51
World Total	57 372	36 347

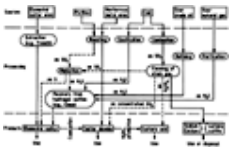

Figure 1. Main sulfur sources and processing routes

——— = Common routes; – – – – – = Possible routes

[Top of Page]

2. Historical Background

Sulfur has been known since ancient times. The Greeks called it *theion*, the Romans *sulphurium.* Genesis refers to it as brimstone — the stone that burns. Its characteristic yellow color, its combustibility, and the pungent odor of its combustion products attracted attention, and led to its use in religious ceremonies. As early as ca. 2000 B.C. the Egyptians are known to have used the gases from sulfur combustion to bleach fabrics. Certain Egyptian dyes, known to have existed in the sixteenth century B.C., could not have been prepared without the use of sulfur compounds. HOMER mentions sulfur as a disinfectant; the Romans used it as a dermatological remedy. PLINY regarded sulfur as one of the most singular kinds of earth, and an agent of great power over other substances. Sulfur was probably one of the components of so- called Greek fire.

Alchemists thought that sulfur was a fatty mineral substance that could be melted to a combustible oil. It was supposed that all combustible materials contained sulfur. In 1777, LAVOISIER classified it as an element, but DAVY, in 1808, still thought that sulfur was a compound containing oxygen, similar to resins.

In Europe, sulfur gained great importance after the Chinese invention of gunpowder was introduced in the early 1200s. The first commercial sulfur was produced in Sicily early in the 1400s. With the advent of industrialization toward the end of the 1700s, sulfur became more and more important in the form of sulfuric acid and fertilizers.

[Top of Page]

3. Properties

Sulfur is a nonmetallic element of group 16, with valence –2 to +6, atomic number 16, atomic mass 32.064. It is tasteless and odorless.

3.1. Physical State

Elemental sulfur occurs in several different allotropic forms, which differ in solubility, relative density, crystalline form, etc., [1-3].

The crystalline rhombic α-form is the most common type of solid sulfur. At room temperature, it is pale yellow; at low temperature it becomes lighter in color, almost white at the temperature of liquid air. It is stable up to 95.5 °C, at which it is transformed to β-sulfur.

The crystalline monoclinic β-form occurs as needle-like crystals, and is stable up to the melting point, 119.3 °C. It is almost colorless and is slowly converted to the α-form on cooling below 95.5 °C.

The γ-form is also crystalline monoclinic, but with a different lattice from that of the β-form. It is light yellow, nacreous, and metastable.

The amorphous μ-form exists in the liquid phase at ca. 160 °C, and is responsible for the increased viscosity at this temperature. It can be preserved as a plastic solid by instantaneous cooling of the melt, but after 1 – 2 d, it becomes hard. This form, insoluble in carbon disulfide, is an amorphous polymer, molecular mass ca. 60 000 – 150 000.

The π-form is a light yellow melt, produced at 119.3 °C from the solid α-form. This is the “ideal” melting point. After several hours, the melting point decreases to 114.5 °C, the “natural” melting point. This effect is explained by transformation of a fraction of the S_8 molecules to the liquid.

The π-form probably consists of chains of S_6, S_7, and S_x and is less soluble in carbon disulfide than the α-form.

Other crystalline modifications are also known, but are unstable at all temperatures and have not yet been studied thoroughly.

The proportions of different types of sulfur molecules in the liquid are given in Figure 2 [3].

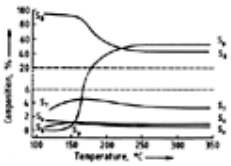

Figure 2. Composition of liquid sulfur as a function of temperature

Sulfur vapor dissociates at higher temperature, from S_8 through the intermediate species S_6 and S_4, to S_2 (Fig. 3). At temperatures above ca. 2000 °C, the gas consists only of sulfur atoms.

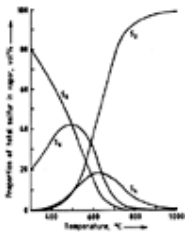

Figure 3. Composition of sulfur vapor as a function of temperature

Figure 4 shows a phase diagram for sulfur [4]. Because of the wide range of pressures, a logarithmic scale is used, which obscures the transition points from the stable to the metastable equilibrium of rhombic, monoclinic, and molten sulfur with the vapor phase.

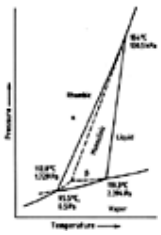

Figure 4. Sulfur phase diagram [4]

More information on the physical state of sulfur can be found in the literature [5-10].

3.2. Viscosity

The viscosity of a pure sulfur melt is shown in Section Other Physical Properties . If the melt temperature is raised, the viscosity first decreases (as is usual for liquids) from 0.017 Pa · s at ca. 120 °C to a minimum value at ca. 158 °C. Above this temperature, the viscosity rises steeply to a maximum value at 186 – 188 °C, at which the melt is dark red. Above this temperature, the viscosity decreases again. This striking phenomenon is interpreted as follows: As the temperature is increased, the S_8 rings are progressively broken up into biradicals whose concentration, at ca. 160 °C, reaches the value necessary for spontaneous polymerization to μ chains. When the temperature rises further, the long chains are increasingly cracked, so that the viscosity decreases [11].

The formation of μ-sulfur is strongly enhanced by even small amounts of impurities. Paradoxically, however, the viscosity of an impure sulfur melt increases during melting to a much smaller extent than that of a pure one; the impurities favor the formation of shorter chains of the μ-sulfur.

The viscosity of liquid sulfur is greatly reduced by the presence of hydrogen sulfide at saturation pressures of 120 – 460 kPa. The viscosity of sulfur in equilibrium with hydrogen sulfide at ca. 100 kPa has been calculated [10]. Calculated values for H_2S partial pressures in the range 0.1 –100 kPa are shown in Figure 5 [12].

Figure 5. Viscosity of liquid sulfur [12]

3.3. Other Physical Properties [1], [13-18]

The melting/solidification point of normal sulfur is 119.3 °C – 114.5 °C, depending on the mode of heating. The melting points of the various allotropes together with other physical data are listed below [1]:

Density of solid at 20 °C, kg/m^3		
Rhombic α	2070	
Monoclinic β	1960	
Nacreous γ	2050	
Amorphous μ	1920	
Melting point, °C		
	Ideal	Natural
Rhombic α	112.8	110.2
Monoclinic β	119.3	114.6
Nacreous γ	106.8	103.4
Density of liquid, kg/m^3		
115 °C	1808	
125 °C	1801	
150 °C	1780	
200 °C	1756	
250 °C	1728	
300 °C	1697	
350 °C	1666	
400 °C	1638	
425 °C	1625	
445 °C	1614	
Boiling point, °C	444.6	
Density of vapor/density of air,		
470 °C	7.837	
Critical temperature, K		1313.1
Critical pressure, MPa		11.75
Critical volume, cm^3/mol		158.0
Surface tension of liquid, mN/m		
125 °C	58.1	
200 °C	52.3	
300 °C	47.0	
400 °C	41.1	
Dynamic viscosity of liquid, Pa · s		
120.0 °C	0.017	
140.0 °C	0.008	
158.0 °C	0.0064	
160.0 °C	5.952	
180.0 °C	86.304	
187.8 °C	93.0	

200.0 °C	78.864
300.0 °C	3.72

Vapor pressure *mp* – *bp*, bar, *T* (K)

$\log_{10}(p) = (-4830/T) + 5\log_{10}T + 21.0049$

Specific heat of solid, J mol^{-1} K^{-1}

Rhombic α, 273 – 368.6 K

$C_p = 14.989 + 0.0261\,T$

Monoclinic β, 273 – 392 K

$C_p = 14.905 + 0.0291\,T$

Specific heat of liquid, J g^{-1} K^{-1}

100 °C	0.978
150 °C	1.079
160 °C	1.867 (peak)
200 °C	1.109
250 °C	1.081
300 °C	1.112
350 °C	1.133
400 °C	1.160

Specific heat of gas, J mol^{-1} K^{-1}

$C_p = A + B\,T + C\,T^{-2}$

	A	*B*×10^4	*C*
S_2	36.509	6.699	– 376 812.0
S_3	57.635	3.182	– 633 044.1
S_4	82.605	34.01	– 711 756.0
S_5	106.554	8.499	–1 645 831.0
S_6	131.842	6.699	–1 720 775.0
S_7	156.013	10.93	–2 335 397.0
S_8	181.297	9.253	–2 280 131.0

Heat of conversion, 95.5 °C, J/g		12.527
Heat of fusion, J/g		
Rhombic α,	112.6 °C	62.25
Monoclinic β,	119.0 °C	43.54

Heat of evaporation, J/g

	Uncorrected	Corrected for dissociation of S_8 in vapor
200 °C	308.6	
300 °C	289.3	
400 °C	286.4	278.0
420 °C	287.6	276.3

440 °C	290.1	274.6
460 °C	293.1	272.9

Thermal conductivity, W m^{-1} K^{-1}	
Rhombic α	
20.0 °C	0.2730
40.0 °C	0.2566
60.0 °C	0.2432
80.0 °C	0.2336
95.0 °C	0.2290
Monoclinic β	
Liquid	
100.0 °C	0.15 – 0.17
115.0 °C	0.1314
120.0 °C	0.1323
140.0 °C	0.1365
160.0 °C	0.1407
180.0 °C	0.1469
200.0 °C	0.1545
210.0 °C	0.1583

A plot of thermal conductivity versus temperature is shown in Figure 6 [1].

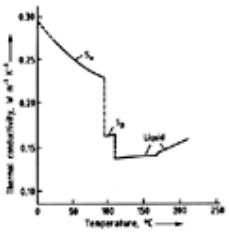

Figure 6. Thermal conductivity of solid and liquid sulfur

Sulfur has four naturally occurring stable isotopes: ^{32}S (95 %), ^{33}S (0.75 %), ^{34}S (4.21 %), and ^{36}S (0.02 %). Radioactive ^{35}S, used in tracer studies, has a half-life of 88 d. The other isotopes: ^{29}S, ^{30}S, ^{31}S, ^{38}S, and ^{39}S are all very short-lived [13].

Sulfur is insoluble in water, but soluble in a number of solvents, such as carbon disulfide, hydrated naphthalenes, and chlorinated aromatics. Disulfur dichloride (sulfur monochloride) dissolves large amounts of sulfur.

Sulfur has extremely low thermal and electrical conductivity, and acquires a strong, negative electric charge by friction. These properties are of considerable importance in its large-scale commercial transport and handling.

3.4. Chemical Properties [1], [13]

Sulfur is one of the most reactive elements; it reacts directly with most elements except iodine, gold, platinum, and the noble gases. In humid air it is weakly oxidized, forming traces of sulfur dioxide and sulfurous acid. At ca. 250 °C, sulfur ignites in air and burns with a blue flame, but the presence of sulfur dioxide significantly raises the ignition point. Finely divided sulfur reacts with metal powders, usually exothermically, to metal sulfides, e.g., iron sulfide, zinc sulfide, or cuprous sulfide. Sulfur reacts with nitrogen to form various sulfur – nitrogen compounds. In the absence of oxidizing agents it is stable to hydrochloric acid, but this stability is diminished by the presence of iron. Concentrated sulfuric acid scarcely reacts with sulfur at low temperature; at 200 °C sulfur dioxide is formed. Nitric acid, and even more so aqua regia or hydrochloric acid with oxidizing agents such as potassium chlorate or bromine, attack sulfur to form sulfuric acid. Concentrated nitric acid in the presence of bromine oxidizes sulfur in a strongly exothermic reaction. When sulfur is heated with aqueous ammonia, ammonium polysulfide and thiosulfate are formed. Boiling with aqueous solutions of alkali and alkaline-earth metals also produces polysulfides and varying amounts of thiosulfates. There is a great variety or organic sulfur compounds; some of them are essential to living organisms.

[Top of Page]

4. Raw Materials for Sulfur Production

World resources of sulfur have been estimated at 25×10^9 t: 4.0 % as elemental sulfur; 4.1 % in sulfide ores; 83.6 % in coal; 3.0 % in crude oil; 5.3 % in natural gas.

4.1. Elemental Sulfur Resources

Ores containing native sulfur are found exclusively in the upper layers of the Earth's crust and are either of sedimentary or volcanic origin. Ore bodies worthy of large-scale exploitation are restricted to a few regions, but there are also many smaller deposits of local importance.

Significant *sedimentary deposits* are located in the following areas:

1. Gulf of Mexico basin — in the states of Louisiana and Texas, and in Mexico
2. Mediterranean basin — main deposits in Sicily, but others on the Italian mainland, and on the French and Spanish coasts
3. Eastern European basin — along the Volga river and the coast of the Caspian Sea along the Caucasus mountains
4. Central European basin — north of the Carpathians in Poland
5. Central Asian basin — in Turkmenistan, Uzbekistan, Russia and Afghanistan

Deposits of *volcanic origin* are found in:

1. East Asian basin — along the Philippines, Japan, the Kuril Islands, and the Kamchatka Peninsula
2. Cordillera basin — along the western side of North and South America, through Mexico, Peru, Bolivia, and Chile

4.2. Sulfide Ores

The main deposits of pyrite are in Spain (Huelva) and Portugal (Alemtejo). Pyrite is also mined in Norway (Lökken), in Cyprus, in Germany (Westphalia), in northern Italy, and in Russia.

By far the largest proportion of copper, zinc, lead, nickel, and cobalt is obtained from sulfide ores. The most important deposits of copper ores are on the west coasts of North and South America, in Zaire, and in Zimbabwe. The main source of lead ores is the Tri-State region of the United States.

4.3. Fossil Fuels

Coal. All coals contain sulfur; the amount depends on their type and origin. In the United States, the coals from the eastern, midwestern, and southern states are predominantly high in sulfur, containing 2 – 4 %, while those from the western states mostly contain less than 1 % [19]. The sulfur content of German lignite is usually 1 – 2 %, rarely > 5 %; that of coal from the Ruhr has 0.5 – 2.5 % [20]. South African coals range from ca. 0.3 % to ca. 2.8 % [21].

Occasionally, larger amounts of pyrite (marcasite) occur in coal, which may necessitate mechanical or chemical separation. However, the major part of the sulfur is organically bound in the coal. When the coal is coked or gasified, most of the sulfur escapes as hydrogen sulfide, which has to be removed from the gases and converted into elemental sulfur. When the coal is combusted, the sulfur is converted to sulfur dioxide, and escapes to the atmosphere, if not removed from the flue gas. The quantity of sulfur discharged in this way exceeds the total amount of sulfur consumed for industrial purposes by a large margin.

Crude oil mostly contains 0.1 – 2.8 wt % sulfur. A few individual crudes contain over 3 % [22]. The sulfur is removed for the most part during crude oil processing, giving substantial quantities of hydrogen sulfide and thiols. World production of crude oil increased steadily from 2654×10^6 t in 1985 to 3019×10^6 t in 1992 [23].

Natural gas contains sulfur, mainly as hydrogen sulfide; the proportion of sulfur varies widely. Some accumulations contain no sulfur at all; in other cases the hydrogen sulfide content is extremely high. The gas from Lacq (southwestern France), for instance, contains 15 % H_2S, while some of the gases from Canada (Alberta and northeast British Columbia) contain ca. 30 % H_2S. World production of natural gas has continued to grow (equivalent to $73\,083\times10^{12}$ J in 1990 [24]).

4.4. Sulfates

Among the naturally occurring sulfates, only calcium sulfate has ever been of any real importance as a source of sulfur, and then only in the production of sulfur dioxide for sulfuric acid manufacture. The economics of this process are now so unfavorable that it is only used under very special circumstances.

Elemental sulfur has never been extracted from sulfates on any significant scale, principally because of the unfavorable energetics of the process, although experimental work has been done on both chemical [25], [26] and microbiological [27-29] methods of reducing gypsum to hydrogen sulfide.

[Top of Page]

5. Production from Native Deposits

Depending on the geology of the deposit, sulfur ores are excavated by normal open-pit or underground mining methods.

Under the special geological conditions in the deposits around the Gulf of Mexico, and in one or two other places, elemental sulfur is produced directly by fusion with hot water (Frasch process).

Various combinations of flotation, autoclaving, filtration, melting out, etc., have been proposed for extracting sulfur from mined ores [30]. In practice, only a few processes are important, and their application depends on the nature of the deposit and the economic conditions (e.g., mining rights, energy and labor costs).

The extraction methods, of which some are only of historical interest, can be classified as follows.

5.1. Flotation

Flotation can be applied easily to the concentration of sulfur ores [31], [32]. After grinding to a sufficiently fine size (> 50 % smaller than 5 µm) the sulfur is readily flotated with foaming agents such as pine oil, cresylic acid, and amyl alcohol. Flotation is enhanced by collectors in a neutral or basic slurry. When the gangue contains limestone, it is recommended that water glass or soda be added.

Flotation gives sulfur concentrates of 70 –90 %. The method has the drawbacks that the concentrate has to be dried, and that it does not produce pure sulfur directly, as do most other processes.

5.2. Melting Out

In Sicily (which until 1904 had a virtual monopoly in sulfur production [33]) melting-out processes were predominant. Sulfur melts at 112 – 114 °C, but becomes highly viscous at temperatures above 158 °C (see Section Other Physical Properties), so the permissible temperature range for melting out is small. Furthermore, sulfur is a poor heat conductor and the gangue absorbs sulfur and retains it in its capillary structure. Therefore, in most cases, the yield is not very high.

Originally, sulfur was melted out in mound-like ore piles known as *calcarelle*, built up on a small hearth equipped with a drain hole, and simply ignited. Only one-third of the sulfur was recovered; two-thirds were lost to the atmosphere as sulfur dioxide.

Calcarone. A substantial improvement to the calcarelle was achieved by covering the ore pile with soil, as practiced in charcoal-making. In the late 1800s, this was the main method in Sicily. To build a calcarone, a circular brick wall with an inclined base was erected on a mountain slope. At the lowest point, there was a duct through the wall to allow air to enter and the sulfur to run out. Coarse and fine lump ore were piled up to make a heap with diameter ≤ 30 m and height ≤ 6 m. The piles were covered with fine ore and residues from earlier operations, sealed with gypsum and clay, and ignited on the uphill side at a point somewhat below the peak. The burning rate was controlled by openings in the outer layer, and the burning front advanced slowly toward the outlet. The sulfur yield was 30 – 60 %.

Chamber Furnaces. The melting process was substantially accelerated by the introduction of the chamber furnace, developed by GILL in 1890, and consisting of two chambers. In the first of these, the sulfur residue from melted-out ore was burned, and the resulting hot gases were passed through a duct into the second chamber, where they melted out the sulfur from fresh ore. So as not to lose any time in charging and emptying, furnaces were built with five or six chambers. These were generally introduced from 1927 onward, and gave yields ≤ 78 %.

As they required neither external fuel nor power, and the capital costs were low, they were excellently suited to Sicilian conditions. It was, however, impossible to connect more furnaces in line, as in the case of the annular chamber kilns used in brickyards, because the higher temperature required to maintain the draught made the sulfur too viscous.

Melting Out with Steam. As early as 1927, steam-heated autoclaves were being used in Sicily. Cylindrical autoclaves, slightly inclined from the horizontal, with diameter ca. 1 m and length 4.75 m, were in general use. They were equipped with a removable cover at the top. Lump ore was introduced on small wagons built to fit inside the autoclave. Steam of 400 – 500 kPa was used for melting out. For ore with 25 – 35 % S, the sulfur yield was about 47 – 55 %. For this reason, only high-grade ores were used in the steam melting process, and the residues, with 13 – 14 % S, were generally passed on for further melting out in a chamber furnace.

Melting Out with Superheated Liquids. Superheated water at 150 °C, as used in the Frasch process, is too expensive on account of the need for pressure autoclaves. Without applying pressure, it is still possible to extract the sulfur, either by melting with salt solutions of high boiling point, or by leaching with inorganic or organic solvents. Examples of recommended media are concentrated solutions of calcium, magnesium, and zinc chlorides, sodium sulfide solution, and solvents such as chlorinated [34], [35] and unchlorinated hydrocarbons [36], and carbon disulfide [37-39].

5.3. Distillation

It is not practicable to melt out the sulfur content of a sulfur ore in an externally heated vessel. Sulfur is such a poor conductor of heat that, with the temperature limitations imposed by viscosity, an extremely large heating surface is required. It is theoretically more favorable to distill the sulfur out of the ore. The ore is either heated in a retort, or brought into direct contact with solid heat sources. For high-grade ores, the retort process has acquired local importance. In Chile, raw and sublimed sulfur are recovered from high-grade ores in iron retorts [40]. In Japan, ores with > 25 % sulfur mixed with pyrite are processed by a similar method. In the Yakutori process, the ore is heated by external firing with coal in small retorts (0.4 m^3) made of 13 % chromium steel, which hold 0.24 t of ore. Up to 100 retorts are coupled to form a furnace. For 1 t of recovered sulfur, 0.41 t of coal is consumed [41].

On account of the poor thermal conductivity of sulfur ores, it has been proposed to heat them with hot inert gases. In spite of many proposals and practical tests, it has proved impossible to exclude oxygen and moisture completely, and inevitably sulfur dioxide and sulfuric acid form. Under the process conditions, these are strongly corrosive, and only processes that use brick-lined equipment (e.g., rotary kilns) and recycle the sulfur dioxide- containing waste gases are suitable. In any

case, this method is not competitive at current sulfur price levels.

5.4. Flotation with Melting Out

In 1953, important sulfur deposits were discovered near Tarnobrzeg and Machov in the Piaseczno region of Poland. They are located near the surface, and can be extracted by open-pit mining, after removal of ca. 50 m overburden. The ore, 20 – 25 % sulfur, is concentrated by flotation. The concentrate, 25 % sulfur and 75 % water, is dewatered by filtration and a drum filter to a water content of 50 %. The sulfur is melted out of this concentrate with steam (ca. 5 MPa) in 50-m^3 autoclaves.

The pit water contains hydrogen sulfide (40 – 400 mg/L). This is removed by stripping with air in a packed spray tower. The stripping air is then scrubbed in a second tower with an alkaline iron (II) hydroxide suspension. The colloidal sulfur obtained is used in agriculture [42]. A similar technique of flotation and autoclaving has been used for refining volcanic sulfur in Chile [43].

5.5. The Frasch Process

By today's standards, the only economic method of extracting sulfur from native deposits is in situ melting by the Frasch process. It is the only way of obtaining marketable sulfur from a native deposit that does not necessitate two discrete operations (mining followed by processing). It was originally developed to take advantage of a very special set of geological circumstances that are virtually confined to the area around the Gulf of Mexico. Although the technique has been further developed for use on less ideally structured deposits in Poland and Iraq [44], [45], technical and economic considerations generally limit its use to the south-east (Gulf) region of the United States and the Mexican sulfur fields.

5.5.1. *Structure of "Fraschable" Sulfur Deposits*

The geological structure around the Gulf of Mexico is a deep sedimentary basin with cretaceous strata underlying tertiary deposits of sand and clay. In a number of places intrusions of rock salt from lower levels have distorted the cretaceous layers into dome-like structures. The Gulf of Mexico has about 400 such domes (Fig. 7 A). They have a diameter of a few hundred meters up to several kilometers. Immediately overlying the salt are layers of anhydrite (calcium sulfate), dolomite (calcium-magnesium carbonate), and limestone (calcium carbonate), which together form the so- called cap rock. The anhydrite and tertiary clays which lie over the limestone are effectively impervious. The theory is that, in the more porous layers of dolomite and calcium carbonate, oil once accumulated. Bacteria in the oil are then thought to have reduced sulfate from the anhydrite to elemental sulfur [46], and may also have been responsible for the virtual disappearance of the oil in these accumulations. The elemental sulfur is, therefore, dispersed through the dolomitic and limestone layers of the cap rock. The thickness of the sulfur- containing strata is 1 – 75 m and the sulfur content varies from a few percent to ca. 50 %, so only about 20 domes contain sulfur in economically significant amounts.

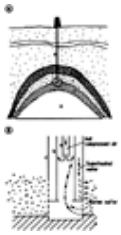

Figure 7. Sulfur extraction by the Frasch process

A) Structure of a sulfur dome

a) Borehole; b) Impervious cap rock; c) Sulfur-bearing limestone; d) Anhydrite; e) Rock salt; f) Pump region (see B)

B) Principle of the Frasch pump

a) Inner tube; b) Middle tube; c) Outer tube; d) Sulfur-bearing limestone; e) Anhydrite

5.5.2. *Outline of the Process*

The process, developed by HERMAN FRASCH between 1894 and 1903, consists of injecting large quantities of hot water directly into the deposit, and then pumping the molten sulfur to the surface. By a method akin to that used by the oil industry, a well is drilled through the cap layers and the sulfur-bearing layers down to the top of the anhydrite stratum, at a depth of ca. 50 – 800 m. Three coaxial pipes are introduced into the borehole (Fig. 7 B). The outermost pipe, with a typical diameter of ca. 200 mm, reaches to the bottom of the borehole; its lower end is perforated with slots. The middle (sulfur delivery) pipe is somewhat shorter, ending about halfway down the perforated part of the outer pipe. A collar on the end of the middle pipe closes off the annular space between the two pipes from the space below the end of the middle pipe. Water at ca. 165 °C, under sufficient pressure to keep it from boiling (2.5 – 3 MPa), is forced down the annular space between the outer and middle tubes, through the perforations in the outer tube above the collar, and into the deposit. The water penetrates into the cracks, pores, and larger voids of the sulfur-bearing limestone, and heats and melts the sulfur around the end of the pipe. The molten sulfur, of higher density than the water, collects at the bottom of the borehole. After a reservoir of molten sulfur has been established there, which requires heating for 24 h or longer, the hot water is turned off and the sulfur can flow through the bottom openings of the outer pipe into the middle pipe. As sulfur has about 1.8 times the density of water, it will only rise about halfway up the middle tube under the influence of hydrostatic pressure. Hot compressed air at about 3 MPa is injected through the innermost pipe, which extends about two-thirds of the way down the borehole, and therefore ends well below the surface of the sulfur column in the middle tube. The resulting foam of sulfur and air is very light and easily rises the rest of the way to the surface. Injection of hot water is continued to maintain the melting process. The transport air rate must correspond to the sulfur melting rate.

At the surface, the sulfur froth is deaerated and transferred to a heated storage tank, or to a sulfur-forming device, to be solidified as slates, prills, pellets, or pastilles. Because the rate at which water is pumped into the sulfur-bearing stratum is much higher than the rate of sulfur production, the pressure tends to increase, so the water must be drained from the deposit through ancillary boreholes, usually located on the flanks of the dome.

5.5.3. *Sulfur Purity*

The effective range of a borehole is 50 –800 m; its output can be as much as 300 t/d. The extracted sulfur can be quite

pure (99.7 – 99.8 %) and light yellow in color. If the sulfur is contaminated by even quite small amounts of bituminous residues from the oil associated with the genesis of the sulfur deposit, it is brown or occasionally blackish in appearance. This so- called dark sulfur may contain ca. 0.08 – 1 % of bituminous material (measured as carbon). The majority of this material is present as complex organic sulfides (carsul), which are responsible for the dark color. Because carsul is dissolved in the sulfur, it cannot be removed by simple filtration. It can, however, be removed to some extent by filtration through diatomaceous earth, which absorbs it [47]. Alternatively, the sulfur can be treated with concentrated sulfuric acid, to char and sulfonate the carsul, after which it is either mixed with diatomaceous earth to absorb the reaction products, or extracted with superheated water [48]. In general, no treatment is completely effective in removing organic impurities from dark sulfur. The most effective one, distillation [49], is hardly considered economic today as a routine means of purifying large tonnages of sulfur. In fact, dark sulfur is acceptable for many uses, except when the organic content is rather high. If dark sulfur is used for making sulfuric acid, the moisture resulting from the combustion of the organic materials may give rise to acid mists, which can cause problems with corrosion and pollution.

5.5.4. *Energy Consumption*

The greatest problem facing the Frasch mines is the cost of energy. The most important factor is the “water ratio”; depending on the geological conditions at the bottom of the borehole, 3 – 38 m^3 of water at 165 °C is required to produce 1 t of sulfur.

The U.S. industry has always used natural gas as fuel. The enormous increase in gas prices over the past few decades has forced the industry to investigate the possibilities of cheaper fuels, e.g., heavy fuel oil, and to consider measures to improve mining efficiency and economics. These include improvements in borehole techniques, as well as the relatively old concept of combining the Frasch operation with the generation of electricity, which can be sold locally [50]. Another possibility to save energy is the reuse of the hot injection water. The problem is contamination of this water by hydrogen sulfide, other noxious and corrosive sulfur compounds, and mineral constituents.

5.5.5. *The U.S. Frasch Industry*

Sulfur production on the Gulf coast of Louisiana and Texas began with Union Sulfur's mine in 1895 and, by 1963, there were 26 plants using the Frasch process. The annual production of individual plants ranged from a few hundred to several million tonnes. The Union Sulfur mine produced about 9.5×10^6 t of sulfur between 1895 and 1924. At first, all plants were built on the mainland, but some of the more recent mines have been located offshore in the Gulf of Mexico, on artificial islands in the open sea, complete with the necessary drilling rigs and boiler plants. The liquid sulfur is pumped through insulated piping, laid on the ocean bed, into storage tanks on the coast, or is directly loaded in liquid form into tankers for immediate transport to the consumer.

The largest mining project of this type is the Main Pass offshore platform, operated by Freeport Sulfur, and located off the Louisiana coast, some 140 km southeast of New Orleans. The 1.8-km platform (the largest offshore structure in the Gulf of Mexico) is designed to accommodate some 300 workers; sulfur production is specified at 5500 t/d. Main Pass, with its 68×10^6 t of proven recoverable sulfur is the largest existing sulfur reserve in North America, and is planned to be in production for at least 34 years [51].

The great challenge for the U.S. Frasch industry was the emergence of large-scale production of sulfur from the gas fields of France and Canada, which as a byproduct is priced on different criteria from Frasch sulfur. Frasch production in the United States decreased from 61 % of total sulfur production in 1980 to ca. 24 % in 1992; it is operated now by just three companies, all of which have been in the sulfur industry for many years. They have six operating mines, but only two have sufficient reserves to keep them operating through the 1990s.

Company production figures for 1987 were as follows:

- Freeport Sulfur (Garden Island Bay, Grand Isle, and Caminada), ca. 1.22×10^6 t/a.
- Main Pass, which came on stream in 1993, $(2.0 – 2.5)\times10^6$ t/a.
- Pennzoil Sulfur (Culberson mine, exploration in the Delaware Basin), ca. 1.5×10^6 t/a.
- Texasgulf, part of the Elf Aquitaine Group of France (Boiling Dome at Newgulf), ca. 0.26×10^6 t/a.

With the majority of elemental sulfur on the world market now coming from other sources, the U.S. Frasch producers can no longer control the price of sulfur, and many operations have become uneconomic. Total production reached a maximum in 1974, at 8.03×10^6 t in response to a peak in world demand. In 1987, production of Frasch sulfur was 3.2×10^6 t, accounting for 34 % of the total produced in the United States. This represents falls of 21 % and 36 % from the 1986 and 1985 figures. In 1988 and 1989, production was 3.17 and 3.89×10^6 t [52], [53].

5.5.6. *The Mexican Frasch Industry*

The Mexican Frasch industry is concentrated on three salt domes on the Isthmus of Tehuantepec, near the port of Coatzacoalcos on the Gulf of Mexico. The salt domes have been exploited for sulfur by the Frasch method for nearly 40 years, and until recently five Frasch mines were operating in the area. These mines are owned by Compania Azufrera Panamericana SA (APSA) which is 96 % state owned, and by Compania Explorado de Isthmo SA (CEDI), 34 % owned by Texasgulf and 66 % state owned.

Frasch production in Mexico was approximately constant between 1985 and 1989 (1.55×10^6 t in 1985, 1.53×10^6 t in 1989), but dropped to 1.02×10^6 t in 1991 [54-56].

The 1991 production figures were as follows:

Coachapa, Veracruz (APSA), 110 000 t/a.
Jaltipan, Veracruz (APSA), 360 000 t/a.
Otapan, Veracruz (APSA), 50 000 t/a.
Texitepec, Veracruz (APSA on behalf of CEDI), 500 000 t/a.

[Top of Page]

6. The Claus Process

6.1. Occurrence of Hydrogen Sulfide

Hydrogen sulfide per se is relatively little used in chemical processing, and is no longer very common as a byproduct of the chemical industry, but it is a common byproduct of fuel processing.

Most fossil fuels — natural gas, petroleum, and coal — contain some chemically combined sulfur. In natural gas it is present mainly as hydrogen sulfide, with only minor proportions of organic sulfur compounds. Petroleum contains a variety of organic compounds, such as thiols, alkyl and aryl sulfides and disulfides, thiophenes, and more complex condensed aromatic heterocyclic sulfur compounds. These are also present in coal, which usually contains a high proportion of inorganic sulfur as the iron sulfides pyrite and marcasite.

When crude oil is distilled at the refinery, the sulfur is partitioned between the various product fractions in proportion to their relative density; very little is found in the lightest fractions, while the heaviest distillates and the residues contain proportionately more sulfur than the original crude oil. These were formerly used as industrial fuel, but this contributed significantly to atmospheric sulfur dioxide pollution, and most industrial countries have now introduced regulations which specify a limit either on the sulfur content of fuel oil or on the amount of sulfur dioxide that may be discharged into the atmosphere from a given installation. Either way, the use of high-sulfur fuel oil is effectively outlawed, and fuel oil has to be desulfurized at the refinery. Demand for the lighter products has increased at the expense of heavy products, and refiners have responded by introduction processes which convert heavy distillates and residues into lighter products.

In both the desulfurization of fuel oil and the conversion of heavy fractions into light products, organic sulfur compounds are converted mainly into hydrogen sulfide. Much of this leaves the process in admixture with other gaseous materials, such as hydrogen (used as a reactant in some processes and produced as a byproduct in others) and byproduct hydrocarbons. Such gas streams are amalgamated with other hydrocarbon-containing off-gases to form the refinery fuel gas.

Some of the hydrogen sulfide released in the processing of heavy fractions becomes absorbed into aqueous condensates and scrubber effluents, where it associates with dissolved ammonia (from the decomposition of organic nitrogen compounds in the oil) to form ammonium sulfides. These effluents are referred to collectively as “sour water”.

For economic reasons, relatively little coal is “refined” or processed to convert it into cleaner fuels. (The majority is burned in large installations such as electricity power stations, and where restrictions on atmospheric emissions from these very large-scale producers of sulfur dioxide apply, the exhaust gases are washed or otherwise treated.) In some installations coal is gasified by partial oxidation to produce fuel gas or synthesis gas for chemical processing, and considerable tonnages of coal still go into the production of coke and byproduct fuel gas for use in steelmaking and other activities. In both cases, the sulfur appears as hydrogen sulfide in the fuel gas, and as ammonium sulfides in sour water [57].

6.2. Hydrogen Sulfide Removal from Gases

In no case is hydrogen sulfide a desirable constituent of gas streams. It is corrosive to metals, highly toxic, and very offensive smelling, and when burned produces sulfur dioxide, which possesses somewhat similar attributes. It is also a strong poison for many catalysts. It must therefore be separated from any gas stream that is to be burned as a fuel or used as a feed for a chemical process. Because of the corrosive and toxic hazards, it must be removed before it is piped over any distance, i.e., sour natural gas must be processed at the wellhead.

Because hydrogen sulfide has few industrial uses and is inherently dangerous, once it has been separated from other useful constituents of the gas mixture it is normally converted into a more useful form of sulfur. It can be burned to give sulfur dioxide and water and converted into sulfuric acid in a special type of sulfuric acid plant designed to handle moist feed gas; however, the product of choice is usually elemental sulfur, which is cheaper and easier to transport than sulfuric acid. The Claus process operates continuously and performs best under stable operating conditions. Therefore, any gas-purification process used to abstract feed gas for a Claus plant must operate continuously. Particularly in refineries, where a large number of independent process units contribute to the raw fuel gas, the composition and flow of the raw gas may fluctuate suddenly and widely. The hydrogen sulfide removal process must be capable of accommodating such fluctuations, although it is not possible to prevent the fluctuations from having some effect on the Claus feed gas flow rate and composition.

Three main types of continuous process are used for separating hydrogen sulfide from gas mixtures:

1. Direct conversion processes
2. Chemical solvent processes
3. Physical solvent processes

The first group includes wet oxidation and “wet Claus” processes. In the former, represented by the well-proven Stretford process and a number of variations, including Lo-Cat, hydrogen sulfide is absorbed into a solution which oxidizes it to elemental sulfur. This is separated physically, and the oxidizing capability of the solution is restored by blowing air through it. In the solution, the hydrogen sulfide-containing gas is mixed with stoichiometric amounts of sulfur dioxide, and passes

into a solvent in which they react to form sulfur, which is insoluble and is removed in molten form. These processes do not produce a Claus plant feed gas, and are mainly intended for processing relatively dilute hydrogen sulfide gas streams.

The chemical and physical solvent processes regenerate the hydrogen sulfide in a form acceptable in a Claus plant. The raw gas, under superatmospheric pressure, is fed into the bottom of an absorption tower and meets a countercurrent stream of cold scrubbing solvent or chemical scrubbing solution. Desulfurized gas leaves the top of the absorber. The rich solution from the absorber is warmed, depressurized, and fed into the top of a desorption column. In the chemical solvent process, the solution has to be stripped with steam to decompose the weakly bound compound formed between the hydrogen sulfide and the active constituent of the scrubbing solution. The steam is generated by passing solution from the boot of the regeneration column through a reboiler heated by low-pressure steam, and reintroducing it into the column, whereupon steam flashes off. In the physical solvent processes, much of the dissolved gas comes out of solution when the liquor is depressurized, and considerably less steam is used in the reboiler.

There are essentially two main groups of chemical solvent processes. One is based on aqueous solutions of alkanolamines (Fig. 8), including monoethanolamine (MEA), diethanolamine (DEA), diisopropanolamine (DIPA), methyldiethanolamine (MDEA), and diglycolamine (DGA). The other is based on hot aqueous potassium carbonate solutions containing additives to improve absorption performance and inhibit corrosion. These include the Catacarb, Benfield, Giammarco – Vetrocoke, and Alkazid processes.

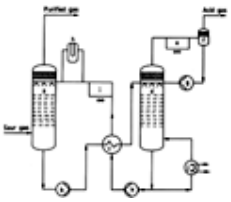

Figure 8. Amine acid gas removal system

a) Absorber; b) Rich solution pump; c) Heat exchanger; d) Regenerator; e) Condenser; f) Separator; g) Condensate pump; h) Filter; i) Cooler; j) Reboiler; k) Lean solution pump

Physical solvent processes (Fig. 9) include Selexol, using the dimethyl ether of poly(ethylene glycol) as solvent, Fluor Solvent (propylene carbonate), Purisol (*N*-methylpyrrolidone), Rectisol (methanol), and Sepasolv MPE (oligo(ethylene glycol) methyl isopropyl ethers). A hybrid physical/chemical process, Amisol, uses a solution of MEA and DEA in methanol.

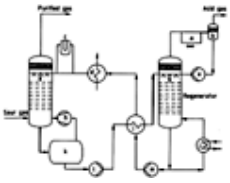

Figure 9. Selective physical H_2S absorption plant

a) Condenser; b) Separator; c) Filter; d) Cooler; e) Condensate pump; f) Absorber; g) Regenerator; h) Recycle gas blower; i) Heat exchanger; j) Reboiler; k) CO_2 flash vessel; l) Rich solution pump; m) Lean solution pump

Most hydrocarbon gases contain carbon dioxide as well as hydrogen sulfide, in proportions which depend on the origin of the raw gas. Like hydrogen sulfide, CO_2 is acidic, and is always removed to some extent in the absorption tower. To what extent this is desirable depends principally on the use to which the desulfurized gas is to be put, and this has to be taken into account in the choice of process and the composition of the solvent or scrubbing solution. In cases where the raw gas has a high CO_2 : H_2S ratio and it is necessary to remove the CO_2 to very low residual levels, a one-stage acid gas removal process may produce an acid gas with too high a CO_2 content (more than 80 vol %) for a Claus plant to handle efficiently. In that case, a two-stage process can be employed, using selective scrubbing agents, the first to remove only carbon dioxide for separate disposal, and the second to remove the remainder of the carbon dioxide and all the hydrogen sulfide. In such a system, both stages may employ a physical solvent, or the first may be physical and the second chemical, or both stages chemical.

The choice of process is further complicated by the nature and concentration of minor constituents in the raw gas, some of which may be important to remove together with the hydrogen sulfide (e.g., carbonyl sulfide, carbon disulfide, and thiols), and others which are less desirable to absorb, either because they degrade the scrubbing liquor or because they impair the efficiency of the Claus plant (e.g., ammonia, hydrogen cyanide). The choice of acid gas separation process is far from straightforward [57-59].

Direct Conversion Processes (Liquid Redox). The criteria mentioned here are not hard and fast rules, but merely indicate when a process should be considered:

1. Low Sulfur Rate. Consider using one of the liquid redox processes if the total amount of H_2S to be removed is equivalent to or less than an elemental sulfur rate of 5 t/d.
2. High CO_2 : H_2S Ratio. The point where the CO_2 : H_2S ratio is high enough to make one of the liquid redox processes economically attractive is usually below 5 % H_2S in the acid gas.
3. Problems. Common problems in liquid redox processes in the oil and gas industry include sulfur plugging and off-color sulfur [60].
4. General. All liquid redox sulfur recovery processes have the same three basic steps: absorption of the H_2S in an aqueous alkaline solution; reaction of absorbed H_2S with a metal complex, forming elemental sulfur; and reoxidation of the metal complex by air sparging.

Before reaction with H_2S, the metal is in either the +5 or the +3 oxidation state, and is reduced to +4 or +2 by the reaction. Reoxidation returns the metal to the higher oxidation state, completing the cycle. Two processes are described. The first (Stretford) employs vanadium as the primary catalyst interacting with the sulfur species; the second (Lo-Cat) is iron-based.

Stretford (Fig. 10). The Stretford process, the first liquid redox process, has been available since 1959. The original patents have expired and many companies can supply Stretford. The original developer was the North Western Gas Board (now British Gas). Stretford uses vanadium as the primary catalyst for the sulfur reaction and anthraquinone disulfonic acid (ADA) to catalyze oxygen transfer in the regeneration of the reduced vanadium. The Stretford process embodies a wealth of experience with liquid redox processing, both with respect to the variety of applications and the amount of operating time.

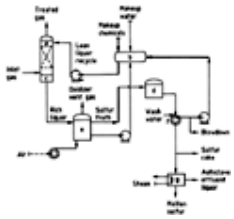

Figure 10. The Stretford process

a) Absorber; b) Pump tank; c) Delay tank; d) Slurry tank; e) Oxidizer(s); f) Filter/centrifuge; g) Melter/autoclave

Lo-Cat (Fig. 11). Lo-Cat is a chelated iron liquid redox process that was commercialized by ARI in the late 1970s and has achieved considerable market success since the late 1980s, especially in smaller units. Solution compositions are maintained by regular additions of iron, two proprietary chelates, a biocide, and a surfactant to ensure that the sulfur sinks to the bottom of the oxidizer, where it is removed as a slurry. This eliminates the need for a separate pump tank for the lean liquor. Typical plants have iron concentrations of 500 – 1500 g/m^3, determined by trade-offs between pumping costs and chemical costs. Applications where gas streams contain relatively high H_2S concentrations, such as regenerator acid gases from amine treatment, can use the ARI autocirculation unit, consisting of a single, liquid-filled vessel that replaces the absorber, reaction tank, and oxidizer. Solution circulation is achieved via fluid density differences [61].

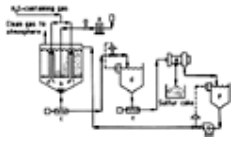

Figure 11. The Lo-Cat process

a) Air blower; b) Autocirculation scrubber; c) Slurry pump; d) Centrifuge feed tank; e) Centrifuge; f) Clarified solution return tank

Other liquid redox processes, include Unisulf (Unocal), Sulfolin (Linde) — both vanadium-based processes — and SulFerox (Shell Oil and Dow Chemical), which is an iron-based process.

6.3. History of the Claus Process

More than 100 years ago, the Claus sulfur recovery process was introduced as a means of disposing of hydrogen sulfide generated in the recovery of ammonium sulfate from gasworks ammoniacal liquor. Later, it was extended to process hydrogen sulfide from the LeBlanc soda ash process. Today, a modification of the original process is the basis of hundreds of sulfur recovery units worldwide, with individual capacities up to 2000 t/d, or more. They process hydrogen sulfide from a variety of sources; some comes from the chemical industry, but a large majority is produced in the processing of fossil hydrocarbon fuels and coal.

In the original process, the Claus reaction was carried out in a single step over a catalyst (Fig. 12). Because the heat of reaction was dissipated by radiation alone, sulfur yields of 80 – 90 % were possible only if the hydrogen sulfide space velocity was low. Processing capacity was not significantly increased by installing cooling coils within the catalyst bed, nor by recycling cooled waste gas through the reactor [62]. It was not until the 1930s that IG Farbenindustrie in Germany introduced a revolutionary development which gave improved efficiency at much higher space velocities and also allowed waste heat to be recovered as steam. This consisted of separately burning one-third of the hydrogen sulfide in a flame with the stoichiometric amount of air to convert it completely to sulfur dioxide, cooling the hot gases in a waste heat boiler, and introducing them, with the remaining two-thirds of the hydrogen sulfide, to the catalytic reactor at 200 – 300 °C

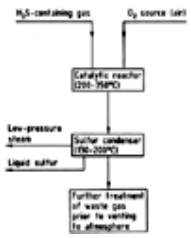

Figure 12. The Claus process

$$H_2S + 3/2\,O_2 \rightarrow SO_2 + H_2O$$

$$2\,H_2S + SO_2 \rightarrow 3/x\,S_x + 2\,H_2O$$

This was later modified so that all the hydrogen sulfide, with the stoichiometric amount of air, was passed through the

flame and waste heat boiler, followed by a sulfur condenser [57].

Because a major proportion of the sulfur was formed in the flame zone and removed in a condenser, the loading on the catalytic stage was reduced, and the conversion efficiency was correspondingly raised. In this form, the process was able, in a single stage, to convert hydrogen sulfide in concentrations of ca. 15 % or more to sulfur, with a conversion efficiency of 92 – 94 % at space velocities of 250 – 300 m^3 H_2S per h m^3 catalyst. This version has become known as the *modified Claus process* (Fig. 13).

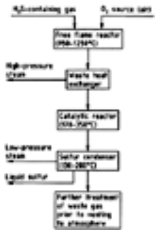

Figure 13. The modified Claus process

The efficiency is raised by adding one, two, or three further stages, in which the gases leaving the sulfur condenser of the previous stage are reheated and passed through another catalytic reactor, followed by another sulfur condenser. However, plants operating on this principle are usually referred to simply as Claus plants. It is interesting to note that the original Claus process (under the name “direct oxidation”) has become of interest again as a method of dealing with hydrogen sulfide streams that are too weak for processing by the modified process.

Despite the enormous accumulated experience in design and operation of Claus plants, the process is inherently difficult to maintain at full efficiency, yet environmental authorities in many parts of the world have set strict emission limits. Under completely stable conditions, the process can operate consistently at its design efficiency, but any perturbation may lead to a quite substantial loss of efficiency. The problem is partly due to the fact that the chemical reactions involved do not go to completion. Fluctuations in the proportions of air and hydrogen sulfide upset the balance between hydrogen sulfide and sulfur dioxide, and whichever is in excess passes through the plant unchanged. Since the exhaust gas is always incinerated to destroy hydrogen sulfide this leads to an increase in sulfur dioxide emissions. Variations in the temperature in the catalytic reactors also reduce efficiency. These may result from a change in either the flow or the composition of the acid gas. Maloperation of an upstream process can result in the appearance of impurities in the acid gas, which can have a deleterious effect on the catalyst or otherwise affect the performance of the plant.

As environmental and safety regulations have become more demanding, it has become necessary to recover sulfur at high efficiency from gas streams which, formerly, were considered too lean or too impure for the process, and would have been incinerated and discharged to the atmosphere.

The German TA Luft demands sulfur recovery rates of 97 % for recovery units with a sulfur capacity ≤ 20 t/d, 98 % for units with capacity 20 – 50 t/d, and 99.5 % for units of capacity ≥ 50 t/d.

As the chemical reactions involved in the Claus process do not go to completion, these high sulfur recovery rates are impossible to reach without the aid of special tail gas units.

6.4. Chemistry of the Claus Process

The chemistry of the Claus process is much more complicated than the simple equations given earlier suggest. This is partly due to a variety of side reactions between the sulfur- containing species and other components of the acid gas stream, such as carbon dioxide, methane, ammonia, and nitrogen from the air, and partly due to the existence in equilibrium of several molecular forms (S_2 – S_8) in the elemental sulfur vapor [63].

Although there are a number of reactions in the conversion of hydrogen sulfide to sulfur, the overall conversion can be represented by the equation:

$$H_2S + 1/2\,O_2 \rightarrow 1/x\,S_x + H_2O$$

This is, in fact, a chemical equilibrium between the forward reaction of hydrogen sulfide and oxygen to sulfur and water vapor, and the reverse reaction. The rate of forward reaction in an equilibrium of this type is roughly proportional to the product of the molar concentrations of hydrogen sulfide and oxygen, while the rate of the reverse reaction is proportional to the product of the molar concentrations of sulfur and water vapor. As the partial pressure *p* of each component is directly proportional to its molar concentration, the reaction rates are proportional to products of the partial pressures of the reacting species. At a given temperature, the rate of forward reaction is:

$$k_1\left(p_{H_2S}\right)\left(p_{O_2}\right)^{1/2}$$

and that of the reverse reaction is:

$$k_2\left(p_{S_x}\right)^{1/x}\left(p_{H_2O}\right)$$

where k_1 and k_2 are the rate constants for the two reactions at that temperature. As the hydrogen sulfide and oxygen react, their partial pressures decrease, while those of sulfur and water vapor increase; the forward reaction slows down

and the reverse reaction speeds up, eventually reaching a point at which their rates are equal:

$$k_1\left(p_{H_2S}\right)\left(p_{O_2}\right)^{1/2} = k_2\left(p_{S_x}\right)^{1/x}\left(p_{H_2O}\right)$$

Thus:

$$k_1/k_2 = K = \left(p_{S_x}\right)^{1/x}\left(p_{H_2O}\right)/\left(p_{H_2S}\right)\left(p_{O_2}\right)^{1/2} \quad (1)$$

The system is then in equilibrium and there is no apparent further change. The relation k_1/k_2 between the two rate constants is the equilibrium constant K. The values of k_1 and k_2 are profoundly influenced by temperature, though not to the same extent, so the equilibrium constant is also temperature dependent, as is the conversion efficiency.

Conventional modified Claus plants use air as the source of oxygen, and although nitrogen is essentially inert in the reaction, it dilutes the reacting gases, and therefore lowers their partial pressures. The first real study of the thermodynamics of the H_2S – O_2 – S – H_2O equilibrium was performed in 1953 by GAMSON and ELKINS [64], who produced a curve expressing the temperature dependence of the calculated equilibrium conversion of a stoichiometric mixture of pure hydrogen sulfide and air to elemental sulfur (dashed line in Fig. 14) [65].

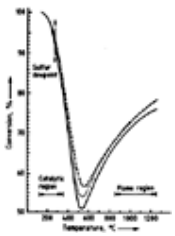

Figure 14. Equilibrium conversion of hydrogen sulfide to sulfur

Pure H_2S, stoichiometric air, 101 kPa, no sulfur removal – – – –, S_2, S_6, S_8 only [64]; . . . . ., S_2, S_6, S_8 only, recent data; ——— all S_x (S_2 – S_8), recent data

At that time, only the species S_2, S_6, and S_8 were known to exist, but since then S_3, S_4, S_5, and S_7 have been shown to be significant in certain temperature ranges (Fig. 3) [65]. Later studies [65], [66], using more up-to-date thermodynamic data and taking the existence of these species into account, have produced revised curves (dotted and solid lines in Fig. 14).

The left-hand region of the curves represents the equilibrium between H_2S and O_2. As the partial pressure of each component is proportional to its molar concentration in the gas mixture, for a given mass of sulfur, the higher the molecular mass of a species, the lower its molar concentration and, therefore, the lower its partial pressure (1 mol S_8 = 2 mol S_4 = 4 mol S_2). As Figure 14 shows, the proportion of S_8 falls with increasing temperature in favor of S_6, S_7, S_5, and, especially, S_2, and therefore the partial pressure of sulfur increases very rapidly with temperature, promoting the reverse reaction.

This effect tails off above ca. 550 °C, by which point the majority of the sulfur is S_2, and the equilibrium sulfur concentration begins to rise again, because a significant proportion of the hydrogen sulfide is beginning to decompose into its elements. This is the main mechanism by which sulfur is produced in the sulfur furnace.

When the hot gas from the burner is cooled, the sulfur concentration at first drops to the minimum represented in Figure 14, as the dissociated hydrogen and sulfur recombine, and then rises along the right-hand portion of the curve, as the oxidation of hydrogen sulfide progresses. However, as the temperature is lowered, the reaction becomes progressively slower until, below about 350 °C, it is too slow to be of practical use. At this temperature, the best theoretical yield (assuming the system has reached equilibrium) is only about 80 – 85 %. The function of the catalyst in a Claus plant is to extend the lower end of the temperature range over which the reaction takes place at a satisfactory rate to the lowest possible temperature so as to attain the highest possible conversion. The catalyst has no influence on the ultimate position of the equilibrium, only on the rate at which it is attained. Since the position of the equilibrium is a function of the partial pressures of the individual components of the gas, if the partial pressure of any one of them is disturbed, the values of the rest will adjust so as to bring the value of Equation (1) back to K. It is not possible to increase conversion efficiency by using excess oxygen, because although a great proportion of the hydrogen sulfide reacts, it forms sulfur dioxide rather than sulfur. Conversion is, however, favored, if the hydrogen sulfide is highly concentrated, and it is further increased if the nitrogen content of the combustion air is reduced or eliminated altogether. Because water is one of the products, a high moisture loading in the feed gas has a detrimental effect on conversion efficiency.

A fortunate characteristic of the reaction system is that one of the products — sulfur — is much less volatile than any of the other constituents, and can therefore be removed by condensation. This reduces its partial pressure considerably, so that if the remaining gases are reheated and passed over a catalyst, they react again to produce more sulfur. Furthermore, since the sulfur loading of the reaction gases is reduced at each stage, the succeeding catalytic stage can be run at progressively lower temperature without risk of sulfur deposition on the catalyst. This is the principle on which the design of practical Claus plants is based [57].

6.5. Claus Plant Configurations

Figure 15 shows the two main process configurations which are in use for processing acid gases in Claus plants.

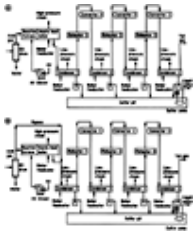

Figure 15. Modified Claus process configurations

A) Straight-through (rich feeds); B) Split-flow (lean feeds)

Straight-Through Configuration. All of the acid gas is passed through the reaction furnace and waste heat exchanger. A significant amount of sulfur is produced in the furnace, and removed in the waste heat exchanger and in the condenser before the gas enters the first catalytic converter. The amount of sulfur condensed in the waste heat exchanger depends on the pressure of the steam produced in the waste heat exchanger, and on the gas temperature at the outlet of the waste heat exchanger. The straight-through configuration is used for feed gases containing H_2S at ca. 40 mol % or more.

When the hydrogen sulfide content of the feed gas is below ca. 40 mol %, the flame temperature in a straight-through Claus plant is too low to destroy satisfactorily any carbon or ammonia compounds present in the acid gas. The sulfur produced may be discolored and there may be a risk of plugging the sulfur lines in the “cold” part of the Claus plant.

However, with acid gas and/or air preheaters upstream of the combustion furnace, or with oxygen instead of air for the partial oxidation of the hydrogen sulfide, acid gases with H_2S levels of 20 mol % and even lower can be processed.

Split-Flow Configuration. Acid gases containing H_2S at ca. 20 – 40 mol % can be processed in a split-flow configuration, in which part of the acid gas is fed directly to the first catalytic converter, and the remainder is burned in the combustion furnace. In theory, two-thirds of the acid gas may be bypassed this way, and in that extreme case essentially no sulfur is produced in the furnace, because one-third of the hydrogen sulfide is burned completely to sulfur dioxide [57]. If the acid gas contains ammonia or other contaminants, it cannot be operated in the split-flow configuration, because these contaminants are also present in the bypassed portion. However, in processing an ammonia-free acid gas and a separate ammonia-containing feed gas (e.g., from a sour water stripper containing only a small amount of hydrogen sulfide compared with that in the acid gas) this ammonia-containing gas can be totally processed to the burner and only “clean” acid gas is bypassed to the first catalytic converter.

6.6. The Claus – Oxygen Process

Operators of Claus plants are increasingly confronted with the problem of improving the throughput of existing plants. This problem has arisen from the processing of crude oils of higher sulfur content and from more stringent regulations calling for further desulfurization of certain products. Apart from higher hydrogen sulfide volumes, wider fluctuations in the concentration of the Claus feed gas are experienced, especially when several gas scrubbing units feed one Claus plant.

As the process gas volume is the main criterion for high throughput, and the hydrogen sulfide content of the Claus feed gas determines the temperature in the reaction furnace, the use of oxygen offers a solution to the partial combustion problem. If this operation mode is applied to cover temporary peaks and to overcome periods of low hydrogen sulfide concentration in the Claus feed gas, the oxygen demand is reasonable, and in many cases where oxygen is available within the plant, this demand can be met at low capital cost. The distribution grid of known oxygen suppliers has meanwhile become so dense that oxygen storage/evaporation equipment can be made available [67].

In a normal plant, air is introduced into the system to supply the oxygen necessary for reaction; a considerable volume of inert gas, primarily nitrogen, is also passed through the system. However, if the oxygen content is increased, by replacing some or all of the air with pure oxygen, the volume of inert material passing through the system is substantially reduced or virtually eliminated.

The advantages claimed for oxygen enrichment mostly stem from the reduced nitrogen content of the gas stream. One effect is that, as the degree of enrichment increases, so too does the temperature in the reaction furnace, and it is possible to treat much leaner feeds in the straight-through process configuration. Without modification, the increase in oxygen content is limited to ca. 28 %. Simulations confirm that the use of oxygen does not significantly increase the overall sulfur recovery efficiency. Theoretically, a 0.1 % increase in a three-stage Claus plant is possible.

There are now three types of technology for using oxygen in Claus plants:

1. A special burner (the Lurgi multipurpose OxyClaus burner)
2. A special burner and an additional recycle blower (the COPE process)
3. An additional special burner, combustion chamber, and waste heat boiler (the BOC process)

The Lurgi Multipurpose OxyClaus Burner. In the early 1980s, Lurgi developed a multipurpose burner (Fig. 16) that can be operated with air, air and oxygen, or pure oxygen, without shutdown. Changeover can be done from the control room, either fully automatically or manually, ensuring optimum oxygen consumption. Oxygen is burned with part of the Claus gas while the remaining Claus gas is combusted with the reaction air flow. This means that an extremely hot oxygen flame is produced, enclosed by a colder air/Claus gas flame. In the core of the oxygen flame, decomposition of hydrogen sulfide to hydrogen and sulfur with reduced H_2O formation takes place simultaneously with the reduction of carbon dioxide to carbon monoxide. The oxygen liberated by reduction of the carbon dioxide reacts in turn with the hydrogen sulfide to form sulfur dioxide.

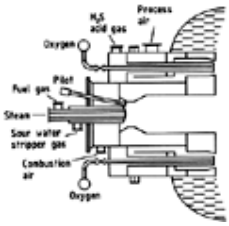

Figure 16. The Lurgi multipurpose OxyClaus burner

The decomposition reactions:

$$H_2S \rightarrow H_2 + 1/2\,S_2$$

$$CO_2 \rightarrow CO + 1/2\,O_2$$

$$H_2O \rightarrow H_2 + 1/2\,O_2$$

are strongly endothermic, and provide for temperature control in the Claus reaction furnace. Gas recycling for temperature reduction in the reaction furnace is unnecessary, even at elevated hydrogen sulfide concentration in the feed gas and with pure oxygen operation.

The Lurgi burner is also designed to combust H_2S acid gas and NH_3 acid gas, and ensures almost complete combustion of ammonia without additional energy consumption. To prevent serious plugging in the “cold” part of the Claus plant, the NH_3- containing gases have to be burned at a temperature of at least 1300 –1350 °C in air.

This occurs in a central burner muffle, which is also used to combust fuel gas for heating, when the plant is started up or shut down. Around the central burner muffle a number of individual H_2S acid gas burners with concentric process air and oxygen supplies are arranged. The oxygen lances for the individual burners are equipped with individual temperature control.

In 1985, the first Lurgi burner went on stream in a refinery in Germany, operating with ca. 50 % oxygen. The plant has been in continuous oxygen operation up to now without experiencing any problems.

By 2005, more than 32 plants equipped with the Lurgi OxyClaus burner were in successful operation. The Lurgi OxyClaus burner is the most widely used oxygen technology for Claus plants using an oxygen content of > 30 % [68].

The COPE Process. Offered by Air Products & Chemicals, Allentown, Pennsylvania, who developed it with Goar, Allison & Associates, Tyler, Texas, the COPE (Claus oxygen-base process expansion) process was the response to Conoco's need to increase capacity because of changing feedstocks. With the COPE process it is possible to operate with pure oxygen. Such a high degree of enrichment is attainable mainly because of two design features.

First, a recycle stream is used to keep the furnace temperature within safe limits; second, a special burner is used. A flow diagram for a typical three-stage Claus sulfur recovery unit after installation of the COPE process is shown in Figure 17 [69].

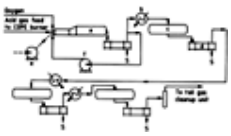

Figure 17. The COPE sulfur recovery unit

a) Reaction furnace with waste heat boiler; b) Steam reheater; c) Converter; d) Condenser; e) Air blower; f) Recycle blower

The recycle, which is the heart of the process, allows the temperature rise in the reaction furnace to be moderated by the recycle of inert components and reaction products, specifically nitrogen, carbon dioxide, and water vapor.

The recycle of inert components increases the pressure drop of the furnace – waste heat boiler – first condenser section. However, the reheater – catalytic converter – condenser sections of the plant experience a reduced pressure drop due to lower volumetric rates. The combination of oxygen enrichment and gas recycling thus permits a large increase in throughput. Increasing the feed capacity of a Claus plant with this oxygen technology results in the production of additional sulfur. Frequently, rundown lines are generously sized when the plant is designed and these do not represent a bottleneck. If replacements are necessary they generally do not represent a major expense compared with other parts of the unit.

Revamps of two sulfur recovery units at Conoco's Lake Charles refinery and two at Champlin Refinery's Corpus Christi refinery were completed in 1988. In the latter installation, debottlenecking was achieved by installing oxygen enrichment without the recycle system, although provision to add this (i.e., full COPE operation) has been made [69], [70].

The BOC Process. Through its Industrial Process Division, Shipley, West Yorkshire, BOC has supplied oxygen enrichment technology to several industries. As well as conventional oxygen enrichment technology (adding up to ca. 25 % oxygen to the process air feed), BOC has signed an agreement with The Ralph M Parsons Company, Pasadena, California, to market processes that can be operated with either oxygen-enriched air or pure oxygen; the proprietary SURE oxygen-enrichment technology provides cost-effective solutions to this demand.

The conventional configuration involves the addition of a new reaction furnace burner, reaction furnace, and waste-heat boiler upstream of the existing reaction furnace (Fig. 18).

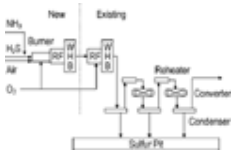

Figure 18. Conventional SURE double combustion configuration reusing existing reaction furnace (RF) and waste-heat boiler (WHB)

Gas effluent from the new waste-heat boiler is routed to the existing reaction furnace, which serves as the second thermal stage. With this configuration the SURE double combustion technology allows the capacity of the sulfur recovery unit to be expanded at considerably lower costs compared to building new air-based sulfur recovery units.

Occasionally, existing reaction furnaces and waste-heat boilers cannot be reused because of original design limitations. In these cases, a two-pass waste-heat boiler with an extended head can be designed in which the extended head serves as the second-stage reaction furnace and the second pass serves as the second waste-heat boiler (Fig. 19). This two-pass waste-heat boiler configuration effectively reduces capital cost and conserves plot space requirement.

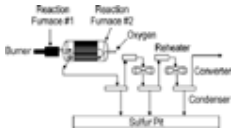

Figure 19. Conventional SURE double combustion configuration with new two-pass waste-heat boiler [71]

6.7. Troubleshooting

Operating Claus plants is straightforward — provided that the feed is according to specification, steam tracing is good, all ratio controllers are set correctly, heat exchanger tubes are not plugged, and so on. It may happen that some of these assumptions are not justified. This Section describes some problems which may occur in Claus units.

Incorrect Air Ratio. Using too much air is the easiest way to lose conversion. For optimum conversion, the H_2S : SO_2 ratio is 2 : 1. The ratio is measured in the tail gas from the final condenser. Sulfur plants are best run on closed-loop tail gas analyzer control. In practice, other methods are often needed to adjust the air supply to the reaction furnace.

If air flow to the reaction furnace is much too high, SO_3 is formed in the incinerator and forms a white plume. A yellowish plume indicates insufficient air to the front end of the sulfur plant.

The incinerator temperature can also be used to indicate the air required in the reaction furnace. A high incinerator temperature, coupled with low incinerator fuel consumption, indicates insufficient air supply to the reaction furnace. Alternately, if large amounts of fuel are required to maintain incinerator temperature, the air supply to the reaction furnace is in excess.

If the tail gas analysis shows SO_2 equal to H_2S, one can cut sulfur losses by 5 % just by reducing the air supply.

COS and CS_2. The presence of hydrocarbons and CO_2 in the acid gas promotes the formation of COS and CS_2 in the reaction furnace. These compounds can significantly contribute to SO_2 in the incinerator. One can essentially destroy both COS and CS_2 by operating the first reactor at an outlet temperature of 320 – 340 °C, so that they are hydrolyzed to H_2S and CO_2. An increase in SO_2 emissions, accompanied by a lower than normal first reactor inlet temperature, is likely due to COS and CS_2.

When to Change Catalyst. A common question put to an operating engineer by the sulfur plant supervisor is “Do we need to change catalyst during the unit turnaround?” With a little luck, this question may be asked before the plant is shut down.

If pressure drop is normal through the catalyst beds, this is a tough question. With adequate instrumentation, one can obtain a vertical temperature profile through the first catalyst bed, and then develop data to make a firm decision. For catalysts in good condition, ≥ 90 % of the heat of reaction is released in the top 200 mm. If catalyst activity is impaired, the reaction is shifted down the bed.

Damage to catalyst and reduced conversion can be a consequence of many other factors besides lost activity: carbon deposits, leaking condenser tubes, damaged support screens, sulfuric acid formation, operation at the sulfur dewpoint. These problems are, however, invariably associated with increasing pressure drop. The most common cause of lost catalyst activity is reversible — sulfur deposits due to low bed temperatures.

Carbon Deposits. The following chain of events illustrate the problem. Light hydrocarbons accidentally enter the amine regenerator, along with the rich amine. The hydrocarbon is stripped overhead. Some condenses in the reflux drum; the rest remains as a vapor and is transferred, along with H_2S, to the sulfur plant. Ten times more air is needed to oxidize 1 mol propane than 1 mol H_2S. When the light hydrocarbon vapors reach the sulfur plant, carbon black is produced in the reaction furnace:

$$C_3H_8 + 2\,O_2 \rightarrow 3\,C + 4\,H_2O$$

The carbon black is deposited on the top of the first catalyst bed. Gas flow is restricted, and a high pressure drop results. Providing sufficient combustion air to the reaction furnace could have prevented this.

The operating engineer can determine if increasing pressure drop is due to carbon accumulation on the catalyst by making the following observations:

1. Is the SO_2 concentration in the sulfur plant tail gas very low (less than 1000 mL/m^3) ? Low SO_2 is a sign of insufficient air supply to the reaction furnace.
2. Are light hydrocarbons accumulating in the amine regenerator reflux drum?

Having determined that the catalyst bed is plugged with carbon, the engineer seeks to correct the situation. Over a period of time, the carbon reacts with SO_2. Unfortunately, this reaction proceeds slowly at low temperature. Maximizing reactor inlet temperature and SO_2 levels can help. Significant (10 %) reductions in pressure drop can take weeks. Shutting down and changing the catalyst may be more practical. The best solution to this incident would have been to keep carbon black from forming in the first place.

There is only one reliable way to prevent such situations — liquid hydrocarbons must be separated from the rich amine upstream of the amine regenerator.

Leaks. A tube leak in the high-pressure steam boiler can lead to major problems. The high-pressure water can erode the metal, and the flow of water into the hot gas stream may increase rapidly.

Water quenches the sulfur-bearing gases, sulfur precipitates on the catalyst. This stops gas flow through the plant. The worst thing that can happen to a sulfur plant is a crash shutdown. Plants should be cleared of sulfur by burning natural gas instead of H_2S before a shutdown, continuing natural gas firing until the amount of liquid sulfur overflowing from the seal legs is reduced to a trickle. When the plant shuts down suddenly, precipitating sulfur solidifies in the catalyst beds, and flow through the unit cannot be re-established.

An alert operating engineer must identify boiler tube leaks before it is too late. A gradual increase in pressure drop is an early warning sign. When this happens, it is vital to check for low steam production rates from the high-pressure boiler. Another tip-off is a low gas outlet temperature from the boiler. If both steam production and outlet temperature are low and pressure drop is high, the plant must be shut down. There is a tube leak.

Water (steam) leaks also reduce conversion of H_2S to sulfur. The Claus equilibrium is shifted to the left as the H_2O partial pressure increases [72].

6.8. Optimization of Sulfur Recovery Units by Use of Analyzers

In the thermal stage of the sulfur recovery process, one-third of the H_2S gas is burned to form SO_2:

$$3H_2S + 1.5\,O_2 \rightarrow 2\,H_2S + SO_2 + H_2O \qquad (1)$$

and the remainder reacts subsequently with SO_2 to form S in the thermal and catalytic stages:

$$2\,H_2S + SO_2 \rightarrow 3/n\,S_n + 2\,H_2O \qquad (2)$$

This is often referred to as the modified Claus process.

Sulfur recovery efficiencies in conventional Claus plants vary from 94 to 97 % for a two-stage plant, and from 96 to 98 % for a three-stage plant. Higher sulfur recoveries are difficult to obtain mainly for the following reasons:

- The Claus reaction is an equilibrium reaction and thus H_2S and SO_2 conversion is limited.
- Process water produced in the Claus reaction hinders conversion and limits total sulfur recovery.
- The critical H_2S/SO_2 ratio of 2/1 is difficult to maintain in a practical manner.

The choice of process configuration (straight-through, split-flow, sulfur recycle, and direct oxidation) largely depends on the acid-gas composition and flow rate, impurities in the acid gas, past experience of designers, and economics. Most sulfur recovery plants are based on the modified Claus process.

Poor operation of the sulfur recovery unit (SRU) may lead to:

- Poor sulfur quality which makes selling difficult and lowers the selling price.
- Soot formation in the furnace or to hydrocarbon breakthrough, which leads to plugging of catalyst beds and further to limited throughput because of high differential pressure.
- Shortened catalyst lifetime due to deactivation/sulfating.
- Fouling in catalyst bed structure, heat exchangers, mesh pads, and sulfur drains.

The most important parameter after flows and temperatures is feedback control of the air/acid gas ratio, which

automatically controls the air to the furnace to stoichiometric optimum and compensates for other O_2-consuming reactants such as hydrocarbons and ammonia, often with fluctuating concentrations.

The use of tail-gas analyzers to assist in the control of combustion air to acid gas ratios in Claus sulfur recovery plants is standard in the industry. The plant tail gas is analyzed by UV spectroscopy and a corresponding output signal is determined. This signal is proportional to the percentage change required in the combustion air to provide stoichiometric concentrations of the principal reactants H_2S and SO_2. When the process is optimized and the correct stoichiometric concentrations of H_2S and SO_2 are achieved, the feedback signal — normally referred to as air demand A — is zero, and no change in combustion air is required. The simplified air demand equation is:

$$A = F([H_2S] - R_{op}[SO2])$$

where F is the plant-specific gain factor, $[SO_2]$ the concentration of SO_2, $[H_2S]$ the concentration of H_2S, and R_{op} the operating ratio (typically 2). At stoichiometric concentrations of H_2S and SO_2 (H_2S:SO_2 = 2:1) A becomes zero regardless of the factor F when R_{op} is 2. When the plant-specific gain factor is established for a particular process, the units of air demand become “per cent change required in process air”, that is, a computed air demand of +1.5 % indicates that the process has 1.5 % excess in air.

To achieve optimum performance of the Claus sulfur plant, the air demand should be kept near zero. Independent analysis of just H_2S and SO_2 is not sufficient to ensure accurate results. The sample gas, obtained from the sulfur plant waste or tail gas stream, also contains other sulfur species: COS, CS_2, and sulfur vapor. These species, while not necessarily required as outputs, must be analyzed and corrected for, as they interfere with the H_2S and SO_2 analysis. Requirements for H_2S, SO_2 analyzer are:

- Fast response
- Efficient sulfur-vapor suppression
- Trouble-free sample system
- Robust industrial design
- Accurate measurement of components
- Reliable and easy to maintain system
- Self-explanatory user interface
- Low installation cost and cost of ownership
- Competent, reliable, and customer-oriented service support

The latest generation of H_2S/SO_2 analyzers, e.g., those of the company Pier Enterprises, are based on UV spectrophotometers with broadband source/spectral detection having full spectrum analysis, high resolution (0.2-nm pixel resolution), and interferences calculated with precision for all SRU gas matrices. Wide dynamic range and very fast response are further advantages of this technology [73].

[Top of Page]

7. Increased Sulfur Recovery

7.1. Selective Oxidation

Section The Claus – Oxygen Process mentioned the use of oxygen as an important improvement in Claus technology. Another improvement is the introduction of super-catalysts in the Superclaus and Selectox processes [74].

7.1.1. *The Superclaus Process, the Euroclaus Process*

The Dutch companies Comprimo and VEG-Gasinstituut, in cooperation with the University of Utrecht, have developed a new version of the Claus process, called the Superclaus process, which increases the sulfur recovery rate of a Claus plant to > 99 %. Two versions have been developed: Superclaus-99 and Superclaus-99.5.

The Superclaus process produces elemental sulfur from H_2S gas, and is characterized by the use of a new selective oxidation catalyst in the last reactor stage. The process has the following features:

1. Overall sulfur recovery up to 99.5 % without tail gas treatment
2. Application in new as well as existing Claus plants
3. Low investment and utility costs
4. Continuous catalytic process without water condensation
5. Use of a new catalyst for direct selective oxidation of H_2S to sulfur

The new catalyst has the unique ability to convert H_2S with excess oxygen into elemental sulfur at > 90 % efficiency. Moreover, high water concentrations in the process gas hardly influence the conversion [75].

Superclaus-99. The conventional Claus process employs a series of steps consisting of: (1) sulfur conversion; (2) cooling, sulfur condensation and removal; and (3) process gas reheating before the next conversion step. Since the Claus equilibrium is involved in each conversion step, a close approximation to equilibrium is reached in the final stage. The

standard Claus process has several limitations which are due to: failure to reach equilibrium; increasing water content of the process gas with decreasing concentration of hydrogen sulfide and sulfur dioxide; formation of COS and CS_2 in the flame, which requires hydrolysis in the first Claus reactor; and the demand for exact control of the air : acid gas ratio. The Superclaus process (Fig. 20) incorporates two new concepts. These are a less sensitive and more flexible air : acid gas control, and a new catalyst for selective oxidation of hydrogen sulfide directly to sulfur, rather than to sulfur dioxide.

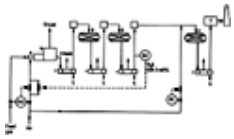

Figure 20. The Superclaus-99 process

a) Combustion chamber; b) Waste heat boiler; c) Condenser; d) Reactor; e) Selective oxidation reactor; f) Incinerator

The acid gas feed is burned in the thermal stage in such a manner that the tail gas leaving the second Claus reactor stage contains 0.8 – 1.5 mol % hydrogen sulfide. The burning step is operated under oxygen-deficient conditions so that the $H_2S : SO_2$ ratio in the furnace is much higher than the conventional value of 2 : 1. As the process gas passes through the various stages, the presence of surplus hydrogen sulfide suppresses the concentration of sulfur dioxide in the gas. As a result, the second Claus reactor outlet may contain concentrations of 0.8 mol % H_2S and 0.05 mol % SO_2. In the Superclaus stage, the gas leaving the second-stage Claus sulfur condenser is reheated and mixed with surplus air before it enters the Superclaus reactor. Here, an essentially nonequilibrium oxidation reaction occurs:

$$H_2S + 1/2\,O_2 \rightarrow 1/^n\,S_n + H_2O$$

The new Superclaus catalyst is the key to this oxidation step. Over 85 % of the hydrogen sulfide is oxidized directly to sulfur, and very little sulfur dioxide is formed. The catalyst consists of a carrier on which metal oxides are supported. The catalyst is chemically and thermally stable and has good mechanical strength. It has a specific surface area < 20 m^2/g, compared with the Claus catalyst value of 300 – 325 m^2/g. The Superclaus catalytic stage is followed by a final sulfur condenser and coalescer vessel, and then the incinerator/stack.

Superclaus-99.5. The Superclaus-99.5 process consists of a thermal stage, two or three catalytic Claus stages, a hydrogenation stage, and a selective H_2S oxidation stage (Fig. 21).

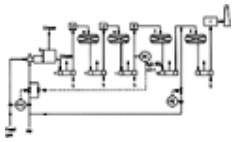

Figure 21. The Superclaus-99.5 process

a) Combustion chamber; b) Waste heat boiler; c) Condenser; d) Reheater; e) Reactor; f) Hydrogenation reactor; g) Selective oxidation reactor; h) Gas cooler; i) Incinerator

The thermal stage and the catalytic Claus reactors are operated conventionally, with an $H_2S : SO_2$ ratio of 2 : 1. The combustion air to the burner is controlled by a conventional H_2S/SO_2 tail gas analyser. As only H_2S is oxidized to elemental sulfur, other sulfur components, e.g., SO_2, COS, and CS_2, must be converted first to H_2S in the hydrogenation reactor. The gas from the hydrogenation reactor is cooled to the optimum inlet temperature of the selective oxidation reactor. The control of the $H_2S : SO_2$ ratio of 2 : 1 is less critical than for standard Claus operation. Fluctuations in the feed gas composition to the unit are absorbed by the excess air supplied to the selective oxidation stage.

The total recovery with two or three catalytic Claus reactors is 99.2 – 99.7 %, depending on the oxidation efficiency [76].

The Euroclaus process (Fig. 22) is an extension of the Superclaus process. It exhibits enhanced efficiency of the selective oxidation section and in particular a reduction of the SO_2 level in the Claus product gas.

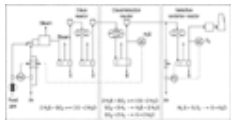

Figure 22. The Euroclaus process

The product gas from the Claus section contains some SO_2, which is not converted in the Superclaus reactor. The amount of SO_2 in Claus tail gas is usually in the range of 500 – 1200 vol ppm wet. This amount depends on the catalyst activity of the preceding Claus reactor and the amount of H_2S in the process gas (ratio H_2S to SO_2).

The recovery loss due to this Claus SO_2 is normally in the range of 0.15 – 0.35 % sulfur recovery for process gas from a second Claus reactor. Conversion of the bulk of this SO_2 by H_2 and CO in the Claus process gas according to:

$$SO_2 + 3\,H_2 \rightarrow H_2S + 2\,H_2O$$

$$SO_2 + 2\,H_2 \rightarrow 1/8\,S_8 + 2\,H_2O$$

$$SO_2 + 2\,CO \rightarrow 1/8\,S_8 + 2\,CO_2$$

will reduce these SO_2 losses to below ca. 0.10 % sulfur recovery.

A special challenge is the suppression of simultaneous COS formation from CO and S vapor. CO is normally present in the Claus process gas in the range of 0.2 – 5.0 vol % wet. At higher CO levels, say 1 vol % or more, this COS formation may be in the range of 500 vol ppm and cannot be prevented. An additional bottom layer of hydrolyzing catalyst which is active at 220 – 230°C has proven to be effective for bulk conversion of COS and also SO_2 according to [77]:

$$COS + H_2O \rightarrow CO_2 + H_2S$$

$$2\,COS + 2\,CO \rightarrow 2\,CO_2 + 3/8\,S_8$$

7.1.2. *The Recycle Selectox Process*

The Recycle Selectox process (Fig. 23) was developed by Union Oil Company of California (UNOCAL) and The Ralph M. Parsons Company. The process was developed mainly for sulfur recovery from lean acid gas streams, typically containing 40 mol % H_2S or less. The Recycle Selectox process is an all-catalytic process, which eliminates the need to burn one-third of the H_2S in a furnace. The incoming acid gas feed is preheated, mixed with recycle gas, and then with process air in the same stoichiometric amount used in the conventional Claus process. The hot, mixed gases enter the Selectox catalyst bed, where a selective oxidation of H_2S to SO_2 occurs in the upper few inches of the bed, and the Claus reaction occurs (to about 80 % completion) in the remainder. The hot gases exit the Selectox converter and enter a sulfur condenser where they are cooled, and about 65 – 70 % of the sulfur present in the feed is condensed and recovered. Effluent gases from the condenser are split, with a portion of the stream being returned to the Selectox bed inlet as recycle gas. This recycle of mainly inert gas serves to dilute the incoming feed and prevents the bed outlet temperature from becoming too high for the carbon steel structural materials in the plant. The rest of the stream (net forward flow) is reheated and passed to one or more Claus stages, the final tail gas being incinerated and vented to the atmosphere.

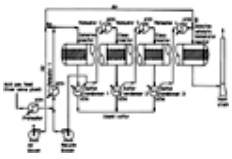

Figure 23. The Recycle Selectox process

HTM = Heat-transfer medium

Overall sulfur recoveries of 94 – 97 % are common. The Recycle Selectox process is appropriate for sulfur recovery from relatively small, lean acid gas streams [78].

7.1.3. *The Once-Through Selectox Process*

The Once-Through Selectox process (Fig. 24) is used for acid gas feed streams containing less than ca. 5 mol % H_2S. No recycle blower is required. The incoming feed is preheated, and the correct amount of air injected. The mixture enters the Selectox stage (first reactor) in the autoignition temperature range. From that point forward, the discussion given earlier for the Recycle Selectox process applies, with the exception of the recycle flow relationships. Typically, two Claus stages and a final incineration stage follow the first Selectox reaction stage.

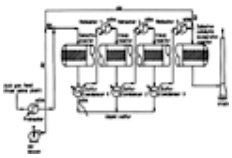

Figure 24. The Once-Through Selectox process

HTM = Heat-transfer medium

With an acid gas feed of 1 – 5 mol % H_2S, an overall sulfur recovery of 85 – 90 % can be expected [79].

7.2. Tail Gas Treatment

7.2.1. *Sub-Dewpoint Processes*

Sulfur recovery of the modified Claus process is limited by the operating temperatures in the catalytic reactor stages. The lower the operating temperatures, the higher the recovery rates.

However, operating the Claus reactors at lower temperatures, i.e., below the sulfur dewpoint, involves adsorption of sulfur on the catalyst, which has to be regenerated after a certain operation time.

So-called sub-dewpoint processes have been developed, all of which use the more favorable equilibrium at lower temperatures. These processes can be simply described as an extension of the conventional modified Claus process.

All are cyclic processes, adsorbing the sulfur produced on the catalyst. The catalyst is regenerated before it becomes significantly deactivated.

In the beginning, around 1970, the maximum achievable sulfur recovery rate with the modified Claus plus sub-dewpoint

process, was ca. 98.7 –99.2 %; in the mid-1990s, sulfur recovery rates of 99.0 – 99.9 % became possible.

The best known sub-dewpoint processes are:

1. The Sulfreen processes
2. The CBA processes
3. The MCRC processes

Common to them all is the ability of the catalyst, not only to adsorb the H_2S and SO_2 from the process gas and act as catalyst for the Claus reaction, but also to have an affinity for the sulfur produced. Product sulfur is adsorbed on the walls of the pores in the catalyst. This reduces the partial pressure of sulfur vapor and allows additional H_2S and SO_2 to react in accordance with the reaction equilibrium [80].

7.2.1.1. *The Sulfreen, Hydrosulfreen, and Doxosulfreen Processes*

Claus plants, as a rule consisting of one thermal and two or three catalytic stages, generally convert 94 – 98 % of the sulfur compounds in the feed gas into elemental sulfur. To increase the sulfur recovery rate in order to reduce SO_2 emissions, the Sulfreen processes have been developed by Lurgi, Germany; and Elf Aquitaine, France. The first Sulfreen plant was started in 1970 in Lacq, France; by 2005, more than 50 had been built.

A wide degree of flexibility is possible with the Sulfreen processes [81]. Three versions are available:

1. The Sulfreen process, based on an extension of the Claus process, where residual H_2S and SO_2 in the Claus tail gas reacts on a fixed catalytic bed at a temperature below the sulfur dewpoint. At such a low temperature (125 – 150 °C) equilibrium conversion of H_2S and SO_2 is nearly complete and sulfur recovery of 99.5 % or more can be achieved.
2. The Hydrosulfreen process, which overcomes the limitation of the Sulfreen process due to COS and CS_2 present in the Claus tail gas. It is implemented by converting these compounds to H_2S in a catalytic pretreatment step upstream of the Sulfreen reactor. This process can achieve an overall sulfur recovery of 99.5 – 99.7 %.
3. The Doxosulfreen process converts residual H_2S and SO_2 from the upstream Sulfreen step with oxygen into elemental sulfur. A sulfur recovery rate of 99.9 % can be obtained.

Sulfreen. The Sulfreen process (Fig. 25) is based on the well-known Claus reaction in which H_2S and SO_2 in the process gas are catalytically converted to elemental sulfur.

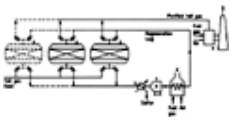

Figure 25. The Sulfreen process

a) Sulfreen reactors; b) Condenser; c) Blower; d) Heater; e) Incinerator and stack

The process operates in the gas phase, the operating conditions being those at which the tail gas leaves the upstream Claus plant. The catalyst, arranged in fixed beds, is activated alumina, with properties similar to those of Claus catalysts.

The tail gas from the Claus plant is passed at a temperature of 120 – 140 °C through one of the two reactors where, by means of the catalyst, most of the H_2S and SO_2 are converted to elemental sulfur and adsorbed on the catalyst.

The sulfur-loaded catalyst can be regenerated by using part of the Claus tail gas. The regeneration gas is heated in an indirectly fired heater. Hot flue gas from thermal incineration or directly fired heaters can also be used.

The sulfur in the hot regeneration gas is recovered in the sulfur condenser. After subsequent cooling of the catalyst bed with purified tail gas, the reactor is again ready to be switched to adsorption.

Hydrosulfreen. The Hydrosulfreen process (Fig. 26) is based on the pretreatment of the Claus tail gas by hydrolysis of COS and CS_2 to H_2S, and direct oxidation of H_2S to sulfur:

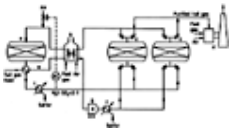

Figure 26. The Hydrosulfreen process

a) Hydrolysis/oxidation reactor; b) Heater; c) Sulfreen reactors; d) Incinerator and stack; e) Heat exchanger; f) Condenser; g) Blower

$$COS + H_2O \rightarrow CO_2 + H_2S$$

$$CS_2 + 2\,H_2O \rightarrow CO_2 + 2\,H_2S$$

$$2\,H_2S + O_2 \rightarrow 2\,S + 2\,H_2O$$

These reactions are performed by bringing the Claus tail gas directly into contact with an activated titanium oxide catalyst.

In a second process step, the remaining sulfur compounds in the process gas are removed as in the Sulfreen process.

The tail gas from the upstream Claus unit (temperature ca. 130 °C) is heated in tail gas preheaters to ca. 200 – 250 °C, and passed to the hydrolysis/oxidation reactor where the temperature rises to ca. 250 – 300 °C, owing to the exothermicity of the reactions. For gas heating, a gas – gas heat exchanger in the waste gas downstream of the final incinerator, or directly fired heaters can be used.

Doxosulfreen. The Sulfreen process is based on the Claus equilibrium, which at 130 – 140°C lies almost completely on the product side. At the outlet of the Sulfreen reactor the concentrations of H_2S and SO_2 are already greatly lowered, and the rather high concentration of water vapors favors the back reaction to a certain extent. Hence, the potential of the Claus reaction is almost exhausted. Further desulfurization in the Doxosulfreen process is therefore based on direct oxidation of H_2S with oxygen at temperatures similar to those of the Sulfreen process. The upstream plant is operated with a small excess of H_2S so that the tail gas of the Sulfreen step contains ca. 2500 ppm H_2S and 250 ppm SO_2. A controlled amount of air is mixed with the gas (O_2/H_2S ratio approximately twice the stoichiometric ratio), and the gas is passed through the catalytic reactor. On the catalyst surface, with the excess of hydrogen sulfide, the Claus reaction is further completed according to:

$$10\,H_2S + SO_2 \rightarrow 3\,S + 2\,H_2O + 8\,H_2S$$

The remaining hydrogen sulfide is oxidized to elemental sulfur:

$$2\,H_2S + O_2 \rightarrow 2\,S + 2\,H_2O$$

Depending on the type of catalyst, a small amount of sulfuric acid is formed by oxidation of SO_2.

The products are adsorbed in the pore system of the catalyst. Initially activated carbon was used as catalyst, leading to the name Carbosulfreen process. However, experience with gases of different origin and composition has shown that in some cases other catalysts such as impregnated alumina are advantageous. Therefore, Doxosulfreen was chosen as a general name for these direct oxidation processes.

The plant essentially consists of two reactors for pretreatment of the Claus tail gas (Sulfreen stage) and two reactors filled with direct oxidation catalyst (Fig. 27). One of these reactors operates in the adsorption mode and the other in the regeneration mode. The regeneration loop comprises the sulfur condensers as well as the blower and the heater for the regeneration gas. Both stages are regenerated simultaneously. The tail gas from the Claus plant is passed through the reactor containing the Sulfreen catalyst at 130 – 140°C. The gas leaving the Sulfreen unit is cooled to ca. 125°C before adding the air for oxidation. The tail gas from the Doxo reactor is passed to the incinerator, where the very small amounts of sulfur components are converted to SO_2.

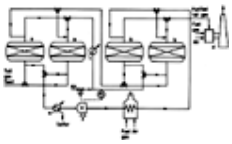

Figure 27. The Doxosulfreen process

a) Sulfreen reactors; b) Doxosulfreen reactors; c) Intercooler; d) Incinerator and stack; e) Condenser; f) Blower; g) Heater

Regeneration of the catalyst is carried out with tail gas from the Claus plant, which is heated to 300 – 330 °C by a special gas heater or by heat exchange with hot flue gas from the incinerator. This gas contains 1 – 2 % H_2S, so any sulfuric acid formed during adsorption is reduced to elemental sulfur. The gas is cooled to ca. 125 °C in a sulfur condenser, and the elemental sulfur is drawn off in liquid form. The regeneration gas is recycled. Following desorption, the reactors are cooled to operating temperature again [83].

7.2.1.2. *The CBA Process*

The first commercial installation of the Amoco cold bed adsorption (CBA) process was commissioned in September 1976 at Amoco's East Crossfield sulfur recovery facilities in Alberta, Canada.

By 1993, approximately 22 additional facilities have shown the process to be readily capable of meeting the robust processing requirements of sulfur recovery.

The CBA process offers a viable solution to the 98 – 99.5 % sulfur recoveries necessary to meet regulatory requirements in many parts of the world. The process is typically accomplished in two or more CBA catalytic reactors: one being used for the low-temperature sulfur formation reaction, with sulfur adsorption on the catalyst, while the other is being regenerated.

Regeneration is carried out by feeding hot gas to the process through the reactor to heat the catalyst, and to desorb and vaporize the sulfur.

The effluent gas from the first reactor in the Claus unit, which is operated at high temperature, is used for regeneration. After regeneration, the CBA reactor has to be cooled by means of a cool gas stream, such as CBA feed gas. When the catalyst has been adequately cooled, the converter is reused for the low-temperature Claus reaction, with adsorption of sulfur on the catalyst.

A number of equipment configurations and flow-sheets have been developed.

Four-Reactor CBA/Rotate R 1, R 2, R 3, R 4. The four-reactor CBA/rotate R 1, R 2, R 3, R 4 process configuration (Fig. 28) is similar to the majority of the four-reactor designs completed and operating to date. The feed gas to the CBA system is supplied by the front-end facilities, e.g., thermal reactor, waste heat boiler, thermal condenser, and reheater.

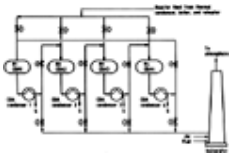

Figure 28. Four-reactor CBA/rotate R 1, R 2, R 3, R 4

Gas flows in sequence through the four reactors. The first reactor in the series operates as a Claus reactor, and is also simultaneously being regenerated. The remaining three reactors operate as CBA reactors. Once the reactor operating as the Claus reactor is regenerated and cooled, it is switched to the tail CBA position. A condenser is provided downstream of each reactor, but only one reheat gas system is required.

Of all the possibilities, this process configuration is capable of achieving the highest levels of sulfur recovery. The equipment is relatively simple, but careful design is required to balance the expected sulfur loading rate with the catalyst regeneration and process cooling capability.

Three-Reactor CBA/Rotate R 1, R 2, R 3. One of the major enhancements to the original four-reactor system is the three-reactor multiple CBA bed configuration (Fig. 29). This improved configuration consists of a single Claus reactor followed by two sub-dewpoint CBA reactors. Overall sulfur recovery approaching that of a conventional four-reactor system can be attained.

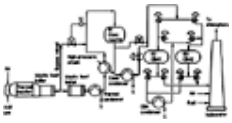

Figure 29. Three-reactor CBA/rotate R 1, R 2, R 3

Front end, thermal reaction; back end, Claus/CBA catalytic converter

In the three-reactor CBA/rotate R 1, R 2, R 3 configuration, the first reactor operates continuously in the Claus mode, and the second and third reactors operate in the CBA mode. The Claus reactor is operated at a high temperature to destroy carbonyl sulfide and carbon disulfide and to provide regeneration gas. The adsorption – regeneration CBA cycle is accomplished in two CBA reactors, functioning in cyclical fashion between the lead position and tail position. In each step, effluent gas from the lead position reactor is fed to a two-stage CBA condenser before being fed to the tail-position reactor. The two-stage CBA condenser functions both as a gas cooler and as a sulfur condenser, generating steam and also using boiler feedwater as coolant. The gas outlet temperature of the CBA condenser is controlled by the flow of boiler feedwater to the second stage CBA condenser. Low-pressure steam is generated in the first-stage CBA condenser, using boiler feedwater preheated in the second stage. Effluent from the tail-position CBA reactor is routed directly to incineration. The lead CBA reactor is in either a regeneration or adsorption mode, while the tail CBA reactor is always in an adsorption mode. Regeneration is accomplished by introducing hot Claus reactor effluent directly to the lead CBA reactor. The stripped sulfur is subsequently recovered in the CBA condenser. Once the lead CBA reactor is regenerated, the lead bed is cooled by directing the Claus reactor effluent through the Claus condenser, before flowing to the lead CBA reactor. The final step in the cycle is the adsorption – adsorption operation. In this case the lead and tail CBA reactor positions are switched [80].

7.2.1.3. *The MCRC Process*

The MCRC process is a sub-dewpoint Claus process that achieves sulfur recovery efficiencies of 99 % or more. The licensor is Delta Projects, Canada; the first MCRC sub-dewpoint Claus plant began operation in 1980.

The MCRC process is designed with either three or four catalytic converters. A higher conversion of H_2S to sulfur is achieved by operating the last converter in a three-converter plant, or the last two converters in a four-converter plant, at temperatures below the dewpoint of the sulfur produced. As the reaction proceeds and sulfur is adsorbed into the catalyst, the catalyst becomes deactivated and requires periodic regeneration or sulfur desorption and vaporization. This is accomplished on-line by using the full forward flow of the sulfur plant process stream. It is not necessary to take the converter out of the process flow for off-line regeneration, as is the case for other sub-dewpoint Claus processes, so the MCRC process requires less equipment and fewer switching valves and associated piping.

The four-converter MCRC process described below achieves sulfur recoveries as high as 99.5 %, the three-converter process can sustain average recoveries above 99 %.

Four-Converter Process. The four-converter MCRC process recovers more sulfur than the three-converter process, but more significantly, the dip in recovery during the bed switch which occurs in the three-converter process is essentially eliminated.

In the four-converter process, the last three converters alternate sequentially in the second, third, and fourth converter positions, which are in regeneration mode. In Figure 30, converters 2, 3, and 4 occupy positions 2, 3, and 4. Gas from condenser 2 is reheated in a gas – gas exchanger, and enters converter 2 to regenerate the catalyst bed. Sulfur that is desorbed and vaporized exits with the regeneration gas stream and enters condenser 3 where it condenses and separates.

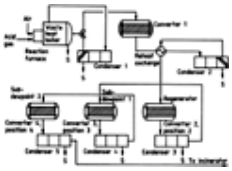

Figure 30. The four-converter MCRC process, operating step 1

Vapor from condenser 3 enters converter 3 where the first stage of sub-dewpoint conversion occurs. Most of the sulfur produced is adsorbed into the catalyst, while the remainder leaves the converter with the gas, and condenses and separates in condenser 4. Condenser 4 discharges vapor into converter 4, where the second and final stage of sub-dewpoint conversion occurs. Virtually all of the sulfur produced is adsorbed into the catalyst, and the remaining vapor cools in condenser 5, and passes to the tail gas incinerator.

When the catalyst in converter 4 reaches a certain sulfur loading, and the catalyst in converter 2 has been regenerated, the converter positions change successively (Fig. 31). Converter 2, which was in the second position (regeneration mode), is now in the third position (first sub-dewpoint mode), and converter 3, which was in the third position (first sub-dewpoint mode), is now in the fourth position (second sub-dewpoint mode), and finally, converter 4, which was in the fourth position (second sub-dewpoint mode), has switched back to the second position (regeneration mode).

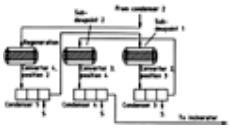

Figure 31. The four-converter MCRC process, operating step 2

Figure 32 illustrates the completion of the cycle. Converter 2 has switched from the third position to the fourth position, and converter 3 has switched from the fourth position back to the second position, while converter 4 has switched from the second position to the third position.

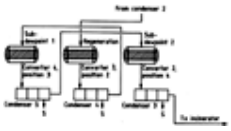

Figure 32. The four-converter MCRC process, operating step 3

To summarize, in the four-converter MCRC process, each of the last three converters switches sequentially through three alternate operating steps: regeneration, first sub-dewpoint, and second sub-dewpoint. These three steps constitute a complete cycle. When a converter completes its second sub-dewpoint step, it switches into the regeneration step to begin a new cycle. Most of the production of sulfur that occurs by sub-dewpoint Claus conversion occurs in the converter that operates in the third position (first sub-dewpoint mode). Hence, the temperature rise in this converter is relatively large. The amount of conversion that can occur in the fourth converter position (second sub-dewpoint mode) is less than for the third position, and therefore the corresponding temperature rise is low, < 3 °C. Thus, the operating temperature in the last converter is lower for the four-converter MCRC plant than for the three- converter plant, and there is a corresponding increase in recovery.

The four-converter MCRC process is inherently stable during a bed switching. There is essentially no dip in recovery because the two sub-dewpoint beds in series act as a buffer that damps the impact of the bed switch [84].

7.2.2. *Processes Working on Reduced Tail Gases*

7.2.2.1. *The SCOT and Super-SCOT, and "SCOT-like" Processes*

The SCOT Process. The Shell Claus off-gas treating (SCOT) process (Fig. 33) was introduced in the early 1970s as an attractive process, combining the well-known ADIP treating technique with a new development.

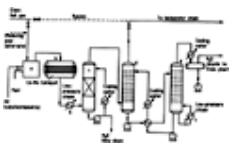

Figure 33. The Shell SCOT process

a) Inline burner; b) SCOT reactor; c) Quench tower; d) SCOT absorber; e) SCOT regenerator (stripper); f) Reboiler; g) Waste heat exchanger

In the SCOT process, nearly all sulfur compounds leaving the Claus unit are converted into H_2S over a hydrotreating catalyst (Co – Mo or Ni – Mo) at ca. 280 – 300 °C. Nearly all sulfur compounds (other than H_2S), e.g., SO_2, COS, CS_2, and elemental sulfur are converted into H_2S. Claus tail gas is heated via an in-line burner which at the same time can produce reducing gas (H_2 and CO) by substoichiometric fuel combustion. Alternatively, if available, an external source of reducing gas can be applied. In this case, another possibility is to heat the tail gas in a gas – gas heat exchanger against incinerator off-gas, or in a heat exchanger using steam to heat the Claus tail gas. The hot H_2S-containing gas leaving the reactor is cooled in two stages: first a gas cooler, producing low-pressure steam, and then a quench column, in which most

of the water in the gas is condensed and sent to a sour-water stripper.

The cooled process gas is subsequently treated in an amine absorber. A selective wash solvent is used for H_2S absorption, while most of the CO_2 in the gas passes through. The SCOT off-gas, typically containing 250 – 500 mL/m^3 H_2S, is routed to an incinerator.

The amine can subsequently be regenerated by stripping with steam or, alternatively, its carrying capacity, which is only partially employed in SCOT, can be fully utilized by feeding to an intermediate tray in the fuel gas or mainstream absorbers. The concentrated acid gases are recycled to the Claus inlet [85].

The Super-SCOT Process. The Super-SCOT version (Fig. 34 [86]) has been developed to achieve an H_2S concentration of 10 mL/m^3 or a total sulfur content < 50 mL/m^3.

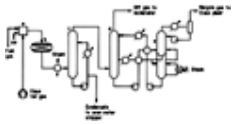

Figure 34. The Super-SCOT process

a) Reheater; b) Reactor; c) Quench column; d) Absorber; e) Regenerator; f) Gas cooler

The process is based on two-stage stripping, optionally in combination with a lower lean solvent temperature; these features can be applied separately or in combination.

In order to reach a specific solvent leanness (H_2S/amine), a specific steam rate (steam/solvent) is required. A leaner solvent will result in a lower H_2S concentration in the SCOT off-gas. However, in order to lower the H_2S concentration in the SCOT off-gas, it is necessary to regenerate the entire solvent flow to this lower leanness level.

This has led to the idea of applying two-stage regeneration in which part of the amine solvent flow is more deeply stripped. The super-lean solvent is routed to the top tray of the absorber, whereas the semi-lean solvent enter half way up the absorber.

The solubility of H_2S in amine solvents is increased when the temperature is lowered. A lower amine temperature results in a lower H_2S partial pressure of the solvent, which enables reduction of the H_2S concentration in the SCOT off-gas. The objective is to lower the total sulfur content to less than 50 mL/m^3 as well as to reduce steam consumption by 30 % compared with the standard SCOT unit. Both options have been applied successfully.

Now, after expiration of the SCOT patent some other processes, known as “SCOT-like processes' are on the market, such as Lurgi's LTGT (Lurgi Tail Gas Treating) and the HCR (High Claus Ratio) process from Siirtec Nigi.

7.2.2.2. *The Sulften Process*

The Sulften system (Fig. 35) is an amine-based tail gas reduction process, developed by Ford, Bacon & Davis, and Union Carbide which reduces the H_2S content of the vent gas to less than 10 mL/m^3.

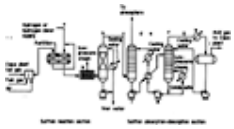

Figure 35. The Sulften process

a) Sulften combined reactors; b) Sulften contact cooler; c) Sulften absorber; d) Solution cooler; e) Solution exchanger; f) Sulften stripper; g) Sulften stripper reboiler; h) Condenser; i) Reflux accumulator; j) Sulften reactors heater; k) Sulften reaction products cooler; l) Rich pump; m) Lean pump

The process consists of two sections: the reaction section and the absorption – desorption section. The reaction section provides for the conversion of sulfur species to H_2S. Tail gas enters the process and is heated to the optimum temperature for the hydrolysis reactions to take place. The heated gas is then reacted over a fixed bed of hydrolysis catalyst. After these reactions are complete, hydrogen is added to the gas stream, which is adjusted to the optimum operating conditions for the hydrogenation reactions. The hydrogenation reactions take place over a second fixed bed of catalyst. The residue gas contains essentially no sulfur species except H_2S.

The reactor effluent is indirectly cooled by generation of low-pressure steam in a waste heat boiler. Further cooling is achieved by quenching the gas with water in a packed or trayed tower prior to the absorption – desorption section.

The absorption – desorption section consists of two towers, an absorber where the H_2S is removed from the vent gas stream, and a stripper to release the absorbed gases for recycle to the Claus unit. The treated gas is either released to the atmosphere or passed to an incinerator, depending on the emission standards required in the country where the plant is located.

The Sulften solvent is a proprietary methyldiethanolamine-based solvent which can reduce the H_2S content of the absorber vent gas to less than 10 mL/m^3, while absorbing approximately 10 % of the CO_2 present in the absorber feed [87].

7.2.2.3. *Solvents for Regenerative H_2S Removal*

Many of the solvents used in regenerative hydrogen sulfide removal processes absorb both hydrogen sulfide and carbon dioxide. However, increased energy costs and the desire for improved sulfur plant operation have made selective gas sweetening (i.e., preferentially absorbing hydrogen sulfide while “slipping” carbon dioxide into the treated stream) more attractive, and several selective solvents are now commercially available [88].

To guarantee sulfur recovery rates of 99.95 % or better for Claus plus tail gas units, special “performance” solvents are now on the market.

The development of alkanolamine solvents for acid gas treating is summarized in Figure 36. Monoethanolamine (MEA), diethanolamine (DEA), and, to a lesser extent, diglycolamine (DGA) have been used for total acid gas removal. With the dramatic increase of energy costs in the early 1970s, improved sulfur plant operation and selective sweetening were needed.

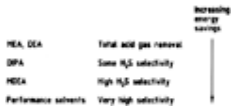

Figure 36. Evolution of gas-treatment solvents

Methyldiethanolamine (MDEA), a tertiary amine, eventually became the most common selective solvent, owing to its resistance to degradation and low tendency to cause corrosion.

The latest development in amine technology was the introduction of performance solvents, based on tertiary amines. Proprietary additives are combined with the tertiary amine to give even greater selectivity and lower energy consumption. Additives can also be used to remove organic sulfur compounds such as carbonyl sulfide [89].

The Super-SCOT process (Section The SCOT and Super-SCOT, and “SCOT-like” Processes) uses a tertiary amine-based solvent, guaranteeing < 10 mL/m^3 H_2S and < 50 mL/m^3 total sulfur components in the absorber off-gas.

The solvent used in the Sulften process (Section The Sulften Process) is a proprietary MDEA-based solvent which can reduce the H_2S content in the gas downstream of the absorber to < 10 mL/m^3. Exxon Research and Engineering has developed Flexsorb SE and SE Plus solvents which are aqueous solutions of a sterically hindered amine, specifically developed to have high hydrogen sulfide selectivity. For low-pressure service, such as sulfur plant tail gas cleanup, H_2S levels of ca. 150 mL/m^3 can be achieved with Flexsorb SE, or 10 mL/m^3 with Flexsorb SE Plus [88].

Another performance solvent, UCARSOL Solvent HS-103 (Union Carbide) is also able to achieve H_2S levels < 10 mL/m^3 in the treated off-gas stream.

7.2.2.4. *The BSR/Selectox Process*

The BSR/Selectox tail gas cleanup (TGCU) process (Fig. 37) is utilized downstream of a Claus unit to improve the overall sulfur recovery to typically 98.5 – 99.2 % [79]. Claus tail gas is heated by mixing with hot gases containing hydrogen and carbon monoxide, produced by a reducing gas generator. This gas mixture then passes through a bed of catalyst for reduction of essentially all sulfur components in the tail gas to H_2S. Effluent gases from this reactor are partially cooled by flowing through a waste heat boiler, and then cooled still further by a contact condenser which lowers the water saturation temperature of the gases to just above ambient temperature. Cooling and water removal are necessary to assure maximum conversion of H_2S to sulfur in the oxidation reactor. The hydrogenated tail gas normally contains about one-third water vapor. Removal of most of this water is essential to the process, since water opposes the Claus reaction to sulfur.

Figure 37. The BSR Selectox sulfur recovery process

The condensate is pumped to a sour water stripper for H_2S removal. The cooled, saturated gases from the contact condenser are heated by exchange with steam to ca. 160 °C, and mixed with the stoichiometric air needed for the oxidation of H_2S to sulfur, which occurs in the second fixed-bed catalytic reactor containing Union Oil's proprietary Selectox-32 catalyst. Premium quality sulfur is obtained by cooling the effluent gases from the direct oxidation reactor to ca. 130 °C, using another waste heat boiler. The cooled gases are mixed with ca. 50 % excess air, and sent to incineration [90].

7.2.3. ***Processes Working on Oxidized Tail Gas***

7.2.3.1. *The Wellman – Lord Process*

The Wellman – Lord process (Fig. 38) was developed in the United States in the late 1960s. The first commercial plant went on stream in 1970. In the following 12 years, 30 plants were built in the United States and Japan; later, Wellman – Lord activities switched to Europe, where 6 plants had been installed by 1993.

Figure 38. Wellman – Lord absorption – regeneration circuit

The Wellman – Lord process is a regenerative process. It involves in removing highly dilute sulfur dioxide from the flue gas in the absorption section, and turning it into rich SO_2 gas in the regeneration section. The recovered gas is directly processed into the end-products: elemental sulfur produced by the Claus process, concentrated sulfuric acid, or liquid SO_2. As the recovered SO_2 gas is already of high purity, these end products meet the highest quality standards.

The process employs sodium sulfite solution in the absorption – regeneration circuit. The solution selectivity removes SO_2 from the flue gas, and sodium bisulfite is formed. During the absorption and subsequent desorption processes, the pH of the solution varies only slightly.

The absorption reaction is reversed in the regeneration section. By thermally regenerating the product solution in an evaporator – crystallizer, the absorbed SO_2 is released again and, after removing the water vapor by cooling, it emerges as a concentrated gas containing 95 % SO_2 by volume. The resulting condensate is used to redissolve sodium sulfite crystallized during regeneration. The fresh solution is ready again for SO_2 absorption, and the circuit is closed.

Figure 39 shows a simplified diagram of the Wellman – Lord system. The concentrated, pure SO_2 gas recovered in the regeneration section can easily be converted into elemental sulfur.

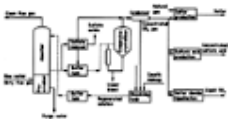

Figure 39. Wellman – Lord flow diagram

For SO_2 reduction the Redotherm process is used. The gas mixture from the Redotherm unit is further processed, together with the refinery H_2S gases, in a Claus plant. The Claus plant (present in most refineries) only needs to be extended by a Redotherm reactor upstream of the thermal Claus reactor; installation of a new Claus unit is usually unnecessary. In the Redotherm process, natural gas is preferable for reduction of SO_2, but hydrogen or town gas can also be used [91].

7.2.3.2. *The Clintox Process*

The Clintox process (Figs. 40, 41) can be described as follows. Tail gas from a Claus plant is incinerated as usual. The hot gas from the incinerator is cooled, e.g., by steam generation, heating of the purified gas, and/or by cooling water. The gas is fed to the bottom of a scrubbing column with three different sections. In the first section, the gas is cooled by direct contact with cooling water, and most of the process water is removed. The dewatered gas enters the second section of the SO_2 scrubber through a chimney tray. There the SO_2 is removed down to a residual content which may be < 1 mL/m^3. The purified gas leaves the top of the column after passing through a third step, another water scrubbing section, removing all traces of the solvent from the purified gas. The purified gas is then fed to the stack, usually after heating. The charged solvent is regenerated in a second column. At the top of the regeneration column, an SO_2 fraction is tapped off, which typically contains more than 85 % SO_2 (plus CO_2 and water). This fraction is recycled to the Claus plant, where the SO_2 reacts with the H_2S in the Claus feed and is recovered as elemental sulfur.

Figure 40. Principle of the Clintox process

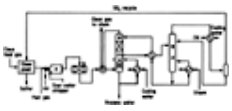

Figure 41. The Clintox SO_2 recycle process

a) Heater; b) Catalytic incinerator; c) SO_2 scrubber; d) Solvent regeneration

One of its main features is that the Clintox process uses physical scrubbing while the others all use chemical scrubbers. In a physical scrubbing process, the amount of SO_2 in the tail gas has no influence on the amount of scrubbing liquid, so high SO_2 concentrations can be scrubbed at the same cost as minor concentrations. In other words, the efficiency of the

scrubbing increases with increasing SO_2 concentrations in the tail gas, provided there are enough trays in the wash column. This has some positive effects on the Claus tail gas treatment. One is that the physical SO_2 scrubber almost completely absorbs fluctuations in the SO_2 load, short-lived or long-term, e.g., by catalyst deactivation in the upstream Claus unit. This ensures a high overall sulfur recovery rate, even if the upstream Claus unit cannot be operated optimally, e.g., owing to varying Claus feeds. Even then, sulfur recovery > 99.9 % can be achieved as a long-term average, (not just for short, favorable conditions). Claus catalyst aging can be tolerated much longer without reducing sulfur recovery. This reduces operating costs [92].

7.2.4. *Biodesulfurization; Thiopaq Process*

The Thiopaq process, which has been used to sweeten biogas since 1993, has now been further developed to remove H_2S from high-pressure natural gas, synthesis gas, and refinery gas streams. The Shell-Paques/Thiopaq process can offer a safe and cost effective alternative to liquid redox processes and amine/Claus installations with loads up to 45 t/d.

The Thiopaq family of biotechnological processes removes and converts hydrogen sulfide (H_2S) or sulfur oxides (SO_x) from gas streams, aqueous streams, or light hydrocarbons. The process operates under near-ambient conditions using safe, naturally occurring, living microorganisms as catalysts.

Sulfur compounds must be dissolved in an aqueous medium. When gas streams are treated, the gas is scrubbed with an aqueous sodium bicarbonate solution, which is regenerated in the Thiopaq bioreactor and recycled to the scrubber. In the regeneration process elemental sulfur is produced. When refinery spent caustic is treated with Thiopaq, sulfide can be converted to either sulfur or sulfate. Although the treated caustic is not regenerated, it is safe for disposal and light mercaptans are also converted.

The Shell-Paques/Thiopaq process for H_2S removal of gaseous streams offers:

- Replacement for amine treating, Claus recovery, and tail gas treating or liquid redox processes.
- Minimal chemical consumption.
- Gas treatment as well as sulfur recovery.
- 99.99 % H_2S removal for gas streams.
- 100 % conversion of sulfide in the bioreactor with 95 – 98 % selectivity to S^0.
- No deactivation.
- Applicable to H_2S concentrations of 100 ppm to 100 vol % and pressures from 1 – 75 bar. Therefore, direct treatment and amine off-gas treatment are both possible.
- High turndown ratio.

A simplified flow scheme of the Shell-Paques/Thiopaq process is presented in Figure 42.

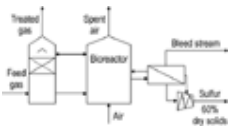

Figure 42. Thiopaq Process for the treatment of H2S-containing gas streams

In the Shell-Paques/Thiopaq process, H_2S-containing gas is absorbed in an alkaline solution under full pressure (up to 75 bar) in an absorption step. For hydrocarbon-containing gases a tray or packed column is the preferred gas/liquid contactor but gases at low pressures and without hydrocarbons can be directed immediately to the bioreactor. The dissolved sulfide is subsequently oxidized to elemental sulfur in the Thiopaq reactor:

Absorption and hydrolysis of H_2S

$$H_2S(g) + OH^- \rightarrow HS^- + H_2O$$

Biological sulfur formation

$$HS^- + \frac{1}{2} O_2 \rightarrow S^0 + OH^-$$

The caustic that is used to absorb H_2S gas is regenerated during the production of elemental sulfur. Normally less than 3.5 % of the sulfide is oxidized to sulfate. To avoid accumulation of sulfate, a continuous bleed stream from the bioreactor is required and make-up water with some caustic is needed. For installations located in dry areas the discharge of sulfate-containing aqueous streams into the environment may be unwanted.

The bleed stream can be minimized by application of membrane filtration. The membrane filter separates the sulfate ions from the other ions, leading to a sulfate-rich effluent stream. Alternatively, the production of an effluent stream can be avoided by continuously recycling part of the bioreactor content over a sulfate-reducing reductive stage to convert sulfate back to hydrogen sulfide. In 1999, an installation producing 4 t/d H_2S started up using H_2 gas as the reducing agent. By the end of 2000, 16 Thiopaq units treating biogas and one treating vent air had been licensed. In addition, a large pilot-plant testing programme was performed by Shell Global Solutions Int. and Paques Natural Solutions at the gas plant of

BEB in Großenkneten (Germany), where the process was tested for treatment of high-pressure natural gas under a variety of conditions.

Paques Natural Solutions developed the Thiopaq process for biogas applications. Together with Shell Global Solutions the development was extended to (high-pressure) natural gas and syngas applications. Simultaneously, UOP and Paques developed treatment of spent caustic and LPG [93].

7.3. Purification

There is no general method of purifying raw sulfur. The treatment used depends on the nature and concentration of the impurities and on the intended use of the sulfur. Simple mechanical impurities can be removed to some extent by melting and settling, but it is better to filter the molten sulfur in wire mesh filter presses.

7.3.1. *Arsenic and Selenium Removal*

Arsenic and selenium occur principally in sulfur produced from sulfide ores, although arsenic may also be found in sulfur recovered from hydrogen sulfide in direct oxidation processes which use an arsenical scrubbing solution. Arsenic is best removed by treating with alkalis or chlorine. Selenium is resistant to alkali treatment; to remove it the sulfur can be treated with aqua regia.

7.3.2. *Melting of Stretford Sulfur*

The aqueous sulfur suspensions produced by the direct-oxidation type gas-washing process (see Direct Conversion Processes (Liquid Redox).) can be treated by an autoclave process [94].

The suspensions contain ca. 10 % sulfur, the remainder being scrubbing solution. After thickening in a decanting centrifuge to a solid content of 43 – 57 %, the slurry is maintained in a homogeneous state in an agitated vessel, from which it is continuously conveyed to the autoclave by an eccentric screw pump. The upper part of the autoclave is divided by a tightly wound steam-heated heating coil into a central melting zone, containing an agitator, and an annular phase separation zone; there is a space in the bottom of the autoclave for sulfur to collect. The entire vessel is maintained at 140 – 150 °C and 500 kPa to keep the liquid sulfur mobile, and prevent boiling of the solution. The slurry is fed in at the top of the melting zone and flows downwards. The dispersion of liquid sulfur and aqueous scrubbing solution flows upward through the phase separation zone. As the liquid sulfur agglomerates, it falls back into the sulfur collection zone in the bottom of the autoclave. The aqueous phase leaves the vessel at the top, and is flash-cooled by depressurization. Condensate from the flash vessel is reused in the slurry homogenizing vessel or returned to the scrubbing process. The molten sulfur is released from the bottom of the autoclave into a heated sulfur collection tank. Its purity is generally ca. 95 – 98 %. The principal impurities are constituents of the scrubbing solution used in the direct oxidation scrubbing process, e.g., sodium and vanadium salts from the Stretford process [63].

7.3.3. *Degassing of Claus Sulfur*

Liquid sulfur produced in a Claus plant contains some dissolved H_2S and hydrogen polysulfides. When the temperature of the sulfur decreases, the polysulfides tend to decompose and H_2S is released. This may cause a buildup of toxic quantities during sulfur loading and unloading, and a buildup of explosive concentrations in unvented storage tanks or tankers.

For these reasons, it is often desirable to degas molten sulfur; many refineries are legally required to degas the liquid sulfur to an H_2S level < 10 mg/kg.

The solubility of hydrogen sulfide in liquid sulfur is anomalous; more hydrogen sulfide dissolves at high temperature than at low temperature (Fig. 43). This is due to the formation of hydrogen polysulfides, weakly bound polymeric compounds of hydrogen sulfide and sulfur:

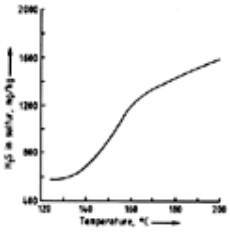

Figure 43. Solubility of H_2S in liquid sulfur at 100 kPa H_2S vapor pressure

$$H_2S + (x-1)\,S \rightarrow H_2S_x$$

The concentrations of H_2S and H_2S_x as functions of total H_2S solubility at different temperatures are shown in Figure 44.

Figure 44. H_2S and H_2S_x as functions of H_2S solubility in liquid sulfur at 100 kPa H_2S vapor pressure

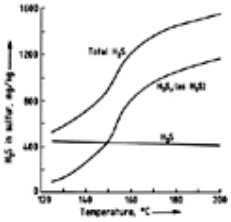

Sulfur produced at different stages in a Claus plant has different hydrogen sulfide concentrations; for example, from the thermal stage there may be 500 – 700 mg/kg S, while from a third reactor stage the concentration is only mg/kg S. An average value of 100 – 150 mg/kg S is normally used when designing degassing systems.

Decomposition of Polysulfides. The reaction whereby hydrogen polysulfides, in the liquid phase, decompose to give liquid sulfur and dissolved hydrogen sulfide, is very slow.

$$H_2S_x \rightarrow H_2S + (x-1)\,S$$

Hydrogen sulfide passes into the gas phase by physical desorption. Thus, the degassing of liquid sulfur involves both releasing the dissolved hydrogen sulfide, which can be enhanced by agitating the sulfur, and accelerating the decomposition of the polysulfides. Consequently, it is important to know both the hydrogen sulfide concentration and the $H_2S : H_2S_x$ ratio; whereas most of the hydrogen sulfide is in the form of polysulfides in freshly produced sulfur, in sulfur that is, say, 12 h old, the hydrogen sulfide and polysulfides are in equilibrium (at 150 °C they are present in roughly equal amounts).

To promote the decomposition of hydrogen polysulfides, it is possible to use a catalyst for example the Aquisulf® catalyst.

Hydrogen sulfide released from the liquid sulfur is removed by a sweep gas passed across the surface. Preferably air, but also Claus plant tail gas, nitrogen, or other inert gases can be used. The sweep gas reduces the hydrogen sulfide partial pressure in the gas space above the liquid, thereby accelerating the physical desorption of hydrogen sulfide, which in turn favors the decomposition of polysulfides [95].

Two degassing systems are briefly outlined: the Shell and the ELF/Lurgi Aquisulf processes.

The Shell Sulfur Degasification Process. In the Shell process (Fig. 45), sulfur degassing is accomplished in a section of the normal underground sulfur production pit or in a section of a normal sulfur vessel. The actual degassing takes place in the stripping column (box) where sulfur is vigorously agitated by bubbling air through the liquid sulfur. The stripping column is open at the bottom and the top, and designed so that the sulfur is circulated through the column a few hundred times per hour (sulfur lift principle). The stripping gas used in the column must be air, not N_2, CO_2, Claus tail gas, SCOT off-gas, nor any other gas which contains little or no oxygen. Air is normally supplied as slipstream from the air blower of the Claus plant.

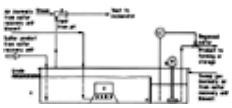

Figure 45. The Shell sulfur degassification process

a) Eductor; b) Sulfur pump; c) Pit; d) Stripping column (box); e) Weir

The presence of oxygen actually lowers the H_2S content of the liquid sulfur, probably by a direct oxidation reaction with some of the dissolved H_2S, to form elemental sulfur. The stripping air, together with an additional flow of air to the vapor space in the pit or sulfur vessel, serves as sweep gas to keep the H_2S concentration in the vapor space below 1.5 vol %. This is well below the “lower explosive limit” of H_2S in air (3.5 vol %). A steam eductor is used to remove the sweet gas, containing the released H_2S, from the pit (or sulfur vessel) and discharge it to the incinerator for disposal. The degassed sulfur product flows over a weir in the pit (or sulfur vessel) to a pumping compartment where it is pumped to product forming or storage.

The ELF/Lurgi Aquisulf degassing process (Fig. 46) consists typically of a concrete pit which is divided in two compartments by a partition wall. The first compartment is equipped with a sulfur recirculation pump and a spraying system, and the second compartment with a sulfur recirculation and a transfer pump also with a spraying system.

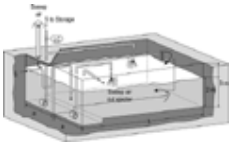

Figure 46. ELF/Lurgi Aquisulf degassing process

Sulfur to be processed is routed continuously into the first compartment of the pit. The liquid sulfur is pumped back into the pit by spraying system. Sulfur flows to the second compartment through a defined opening located at the bottom of the

partition wall. Sulfur is sprayed again in the second compartment, where degassing is completed. It is then transferred to storage by the recirculation and transfer pump. A constant level is controlled inside the pit.

From the catalyst dosing system the liquid Aquisulf catalyst in an amount of max. 20 mg/kg sulfur is routed to the degassing pit by means of a catalyst metering pump. To allow quick mixing of the catalyst with sulfur the catalyst is injected at the suction side of each sulfur recirculation pump.

The vapor phase of the pit is swept with air. Sweeping gas flows countercurrently to the liquid sulfur from the second compartment to the first through a defined hole located at the top of the partition wall. The air is routed either to the incineration or to the Claus combustion chamber.

Several safety features are incorporated to ensure that an explosive H_2S/air mixture is not formed in the pit:

- The flushing air flow is measured at the pit inlet. Should this flow be too low, degassing is automatically stopped.
- The H_2S concentration can be measured by an online analyzer.
- In case of high gas temperatures degassing is stopped.
- Fire is extinguished by snuffing steam.
- Explosion doors are installed on the pit roof.

The sulfur temperature must be at least 130 °C in the degassing pit. Depending on temperatures of inlet sulfur and of flushing air, the sulfur pit can be heated by means of low-pressure steam coils installed at the bottom of the pit. For optimal degassing conditions the sulfur temperature should be less than 150 °C. A low sulfur level automatically cuts off recirculation pumps to avoid uncovering of the steam coils and pump impellers.

7.4. Sulfur Management (Lurgi's OmniSulf Process)

Natural gas is the fastest growing primary energy source in the EIA (Energy Information Administration) 2002 forecast, and its use is projected to almost double between 1999 and 2020, providing a relatively clean fuel to meet worldwide energy demand. However, it is essential for all sulfur components and (in part) carbon dioxide to be removed before natural gas is liquefied (LNG), fed to a pipeline, or used as a feedstock for the production of base chemicals such as methanol or fertilizers. OmniSulf comprises technologies from BASF, Zeochem, and Lurgi to meet specifications according to process requirements and/or environmental regulations. It is based on Lurgi's proprietary experience as a technology licensor and engineering contractor.

Natural gas purification consists of four main process units: acid gas removal (AGR), dehydration and mercaptan removal (DMR), sulfur recovery unit (SRU), and mercury removal unit (MRU). This normally results in four different licensors, four sets of liabilities, and numerous interfaces between the involved parties that often give rise to good excuses. If one unit does not work properly, it is quite easy to blame inadequate battery limit conditions or specifications. The exception, however, proves the rule: BASF/Lurgi's “one-stop shop” sulfur management eliminates all the aforesaid constraints. One reason OmniSulf was developed was to raise the overall sulfur recovery rate to 99.5 %, but if the dictated sulfur recovery rate is not that strict OmniSulf can easily be economically adapted to customer requirements. Single-line responsibility is achieved with the following concept:

- One license contract, one overall guarantee and liability, and one transparent license fee.
- Comprehensive basic engineering completed by a single company (Lurgi).
- Tailor-made design to avoid multiplication of internal design margins.

This results in significant Capital expenditure (CAPEX) and Operational expenditure (OPEX) savings. The three partners, BASF, Zeochem, and Lurgi, rank among the leaders in their respective fields:

- OmniSulf can be licensed all over the sales gas and LNG producing world without restrictions.
- It can be designed as a full plant concept (Fig. 47), including sulfur production, but it can also be restricted to single units, for example, fuel gas treatment and acid gas re-injection.

Figure 47. OmniSulf process

The OmniSulf concept shown in Figure 47 encompasses the following proprietary key processes: Acidic components are removed using BASF's aMDEA process (AGR). The cleaned gas is routed to a DMR unit that removes moisture and mercaptans with Zeochem's special 13X zeolite technology. Where necessary, the sweet gas can be further routed to a mercury-removal unit operated with impregnated activated carbon.

The DMR is thermally regenerated at regular intervals. Mercaptans are recovered from the regeneration gas using Purisol technology. The gas can then be fed to the fuel gas system.

All gas streams containing sulfur are routed to a sulfur recovery unit (SRU). Elemental sulfur is produced in the Claus process (CLS, Lurgi Multi-Purpose Burner). A Lurgi tail gas treating and acid gas enrichment unit (LTGT) is connected to the Claus unit to boost sulfur recovery. The sulfur product is then further treated in an Aquisulf degassing process, which

achieves H2S concentrations in sulfur below 10 vol ppm. Vent gases are incinerated before being released to the atmosphere. In the Claus process, Lurgi's Multi-Purpose Burner is used for optimum operation. This unique burner system allows for stoichiometric combustion and therefore complete destruction of mercaptans in the center of the burner, while H_2S is simultaneously converted substoichiometrically as required by the Claus reaction $(2\,H_2S + SO_2 = 3\,S + 2\,H_2O)$. Conversion to elemental sulfur is completed by the reaction between the remaining SO_2 and H_2S in catalytic converters.

Due to the saturated sulfur market, acid gas re-injection is increasingly chosen as a viable alternative. The OmniSulf concept can be tailored for gas re-injection [96], [97].

[Top of Page]

8. Gas-Phase Reduction of Sulfur Dioxide

Elemental sulfur can be recovered directly from gaseous sulfur dioxide by reducing it with gases containing carbon monoxide, hydrogen, or hydrocarbons. It is also possible to reduce sulfur dioxide with solid carbonaceous material, such as coke or coal. The problem is significant in the smelting of nonferrous sulfide ores, where the large amounts of sulfur dioxide raise pollution problems if no market exists near the smelter for sulfuric acid or liquid SO_2. The reduction method can be also used for treating stack gas from power stations; however as the concentration of SO_2 is very low, a sulfur dioxide concentrator is necessary.

Several processes, based both on solid and gaseous reducing agents have been proposed, some of them have been tested in pilot plants, but only a few have been commercialized.

8.1. The Allied Chemicals Process [98], [99]

In the Allied Chemicals process (Fig. 48), the SO_2-containing gas is purified, mixed with natural gas, and preheated. The reduction step consists of one main catalytic reactor and two auxiliary reactors, which are alternatively utilized as regenerators, i.e., one of them is heated by the hot gas from the main reactor, while the other is used for preheating the gas stream to the main reactor. The conversion is carried out at 800 – 1100 °C in two steps. In the first, a portion of the SO_2 in the feed gas reacts with natural gas, yielding a mixture of elemental sulfur, hydrogen sulfide, carbon dioxide, and water vapor, according to the following equation (natural gas shown as methane):

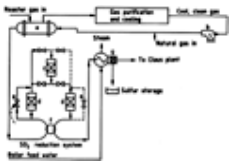

Figure 48. The Allied Chemicals process

a) Hot gas heat exchanger; b) Centrifugal blower; c) Main reactor; d) Auxiliary reactors/regenerators; e) Temperature control bypass valves; f) Four-way valve; g) Sulfur condenser;

$$2\,CH_4 + 3\,SO_2 \rightarrow 1/n\,S_n + 2\,H_2S + 2\,CO_2 + 2\,H_2O$$

Hydrogen sulfide, formed in the first step, reacts in the second step with the remaining SO_2, yielding additional elemental sulfur and water vapor:

$$H_2S + SO_2 \rightarrow 3/n\,S_n + 2\,H_2O$$

After condensing the sulfur, the gas is led to the Claus plant, where the residual sulfur dioxide, hydrogen sulfide, and carbon – sulfur compounds are converted to elemental sulfur.

The process was developed by Allied Chemicals, Morristown, New Jersey, in 1965; it was first commercialized by Falconbridge Nickel Mines during the period 1969 – 1972; several other commercial installations are known.

At present, however, only one plant is in operation: the Buschhaus power plant in Germany, where the process has been converted to continuous mode [100] (Fig. 49). The concentrated SO_2 gas is divided into two streams; one of which is preheated to the starting temperature of the catalytic reaction and conducted into the first reactor. The heat of reaction, is utilized to heat the second stream by mixing both flows. The combined gas flow then passes through the second reactor, where the residual sulfur dioxide is reduced. After the two-step condensation of sulfur, the gas is conducted to the Claus plant.

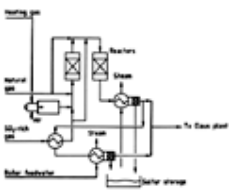

Figure 49. Modified Allied Chemicals process [100]

8.2. Other Processes with Gas as Reducing Agent

Two older natural gas reduction processes exist, both with one main and two additional catalytic stages, but differing in the temperature of the main reactor: the Asarco process, developed by the American Smelting and Refining Company in the

1940s; and the Texasgulf process, developed in the 1950s. Neither has ever been commercialized [98].

Lurgi has developed two thermal reduction processes, called Redotherm and Redomet, which use hydrogen and natural gas, respectively, as reducing agent [100], [101].

In the Redomet process (Fig. 50), the sulfur dioxide is reduced at 1000 – 1200 °C in a noncatalytic reactor to S_n and H_2S according to the following equations (natural gas shown as methane):

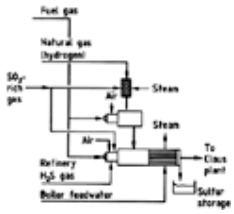

Figure 50. The Redomet process

$$CH_4 + 2\,SO_2 \rightarrow 2/n\,S_n + CO_2 + 2\,H_2O$$

and:

$$2\,CH_4 + 3\,SO_2 \rightarrow 1/n\,S_n + 2\,H_2S + 2\,CO_2 + 2\,H_2O$$

Two Redotherm reactors have been in operation at the OMV refinery at Schwechat, Austria since 1985. Each Redotherm – Claus plant has a sulfur capacity of 90 t/d; a mixture of 90 % hydrogen and 10 % natural gas is used as reducing agent.

8.3. Processes with Coke or Coal as Reducing Agent

Cominco operated a process at Trail, British Columbia, between 1935 and 1940, in which SO_2 gas with some oxygen was passed at 1300 °C through a bed of coke. Coke was charged at the top, and ash was removed on a rotary grate at the bottom of the furnace. The gas, containing CO, CO_2, S_x, and COS was conducted to a catalytic bauxite bed, where the carbonyl sulfide was converted to CO_2 and S_x by addition of sulfur dioxide. Sulfur production was ≤ 150 t/d at 99.99 % purity [98].

In the late 1960s, Foster Wheeler Energy, Livingston, New Jersey, began development of a sulfur dioxide reduction process, using coal as reducing agent. The commercial-size demonstration plant, with a Bergbau-Forschung activated-coal SO_2 concentrator as front end, was erected in 1974 in Lünen, Germany. The process, called Resox (Fig. 51), is based on the direct reduction of sulfur dioxide to sulfur or hydrogen sulfide by passing it, together with steam at 550 – 800 °C (the temperature depends on the type of coal), through a bed of rice-grain-sized coal, contained in a vertical reaction vessel. The coal moves slowly downward, countercurrent to the gas stream. The sulfur is condensed from the gas stream in the sulfur condenser. The gas is then incinerated and returned, partially to the reactor, partially to the SO_2 concentrator [98], [102], [103].

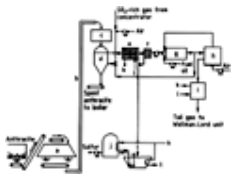

Figure 51. The Resox process

a) Anthracite storage pile; b) Conveyor/elevator; c) Anthracite feed bin; d) Resox reactor; e) Sulfur condenser; f) Entrainment separator; g) Incinerator; h) Air preheater; i) Waste heat boiler; j) Sulfur storage tank; k) Steam; l) Condensate

The Sulf-X process, which has been developed and tested in a small-scale plant by Pittsburgh Environmental and Energy Systems, uses for the reduction of sulfur dioxide a complicated conversion via a solution of ferrous sulfide – ferrous sulfate – sodium sulfide – sodium sulfate. Dry, crystalline sodium sulfate is then calcined with coke to give sodium sulfide [104].

In the late 1960s, Outokumpu Oy, Finland, developed a process for conversion of sulfur dioxide in the uptake shaft of copper or nickel flash-smelting furnaces (Fig. 52). After the gas temperature has been increased to 1300 – 1350 °C by combustion of a hydrocarbon fuel, the powdered coal is fed into the shaft; the reduced gases are cooled, cleaned, and catalyzed to convert all sulfur compounds into elemental sulfur, which is condensed by cooling. The sulfur capacity of the pilot plant, operated at the research center at Pori was ca. 5 – 6 t/d. Two industrial-scale plants were built, but neither was put into operation [105].

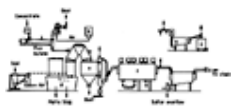

Figure 52. The Outokumpu reduction process (pilot plant)

a) Dryer; b) Air preheater; c) Fluidizing chamber; d) Flash furnace; e) Waste heat boiler; f) Electrostatic precipitator; g) Sulfur condenser; h) Hot catalyzer; i) Cold catalyzer

In 1983 – 1986, Outokumpu Oy developed a new reduction process, called Sulfred, and tested it as pilot equipment, with a

gas capacity of 2000 – 10 000 m^3/h, installed at the Outokumpu Pori Works. The process (Fig. 53) is characterized by four main steps in two closed chemical cycles. In the first step, sulfur dioxide from the flue gas is absorbed in aqueous sodium sulfide. A small bleed of the washing solution goes to the autoclave stage, where intermediate product, formed in the absorption tower, is decomposed at elevated pressure to liquid sulfur and sulfate. The molten sulfur settles to the bottom of the autoclave, where it is drawn off. The sodium sulfate solution is converted in the regeneration stage to sodium sulfide solution by adding solid barium sulfide, forming insoluble barium sulfate, which is filtered out and reduced back to barium sulfide in a rotary kiln by means of low-ash coal. The process, which has some similarity to the Sulf-X process, has not been commercialized up to now [106].

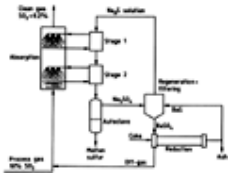

Figure 53. The Sulfred process

Many patents for different reduction processes exist, but little is known of their testing or commercialization [107-115].

[Top of Page]

9. Commercial Grades and Forms of Sulfur

9.1. Quality

Frasch sulfur is usually classified in three grades:

1. Superbright: maximum organic content 0.02 % C; guaranteed purity 99.5 %
2. Bright: maximum organic content 0.04 % C
3. Dark: maximum organic content 0.2 % C

There is also Lacq quality sulfur, which is 99.9 % pure (dry basis), and contains 0.01 – 0.03 % organics, 0.01 – 0.05 % ash, and ca. 0.5 % moisture, but no arsenic, fluorine, or bitumen.

As the greatest part of sulfur produced is converted into sulfuric acid, the international standards for this use are of great importance (Table 2).

Table 2. International standards for sulfur in sulfuric acid production

	Liquid	Solid
Purity	> 99.9 %	> 99.7 %
Carbon	< 0.02 %	< 0.05 %
Sulfuric acid	< 0.01	< 0.01
Hydrogen sulfide	< 10 mg/kg	< 10 mg/kg
Moisture		< 2 %
Color	bright yellow	
Commercially free of As, Se, Te, i.e., < 5 mg/kg of each		

9.2. Commercial Forms for Bulk Transport

9.2.1. *Dry Bulk Sulfur*

Until the end of the 1960s, the standard method of sulfur storage was vatting at the production site, the molten sulfur being poured into vast open-air rectangular shuttered frameworks, typically containing 0.25 – 1×10^6 t or even more, and left to solidify. These sulfur stockpiles have played a vital role in balancing supply and demand, being built up during periods of slack demand and drawn on at times when demand has exceeded the rate of production. Solid sulfur is quite brittle and sulfur stored in vats can be reclaimed by mechanical excavators. The larger chunks may have to be crushed, to produce a product consisting of a mixture of particles, ranging from coarse lumps to very fine dust. This material can be handled and transported in much the same fashion as ores and other bulk solids. Apart from the obvious disadvantage that a material of this kind is liable to contamination from moisture, soil, rust, or other impurities during outdoor storage and handling, the dust constitutes a major safety and environmental hazard — the danger of explosion, pollution of the surrounding area by formation of prominent yellow deposits, as well as chemical hazards to the soil and plant life. Because of this, an increasing share of sulfur is now reclaimed from the vat in the molten phase with an in situ remelter; several different systems exist, but in all of them steam is used as the heating medium. Some states have prohibited the movement of dry bulk sulfur, and techniques for remelting vatted sulfur in situ have attracted considerable interest [116], [117].

9.2.2. *Liquid Sulfur*

The majority of the elemental sulfur used in industry goes into the production of sulfuric acid, for which purpose it is needed in the molten state. This is also the state in which it is produced in all the most important primary production

methods, and because of the very low thermal conductivity of sulfur, it can easily be maintained in the liquid state for storage and transport with relatively low amounts of heating. Molten sulfur is easy and quick to load and unload; but the loading and unloading facilities require considerable investment, and they can be used only for this single material, unlike the facilities for handling solid sulfur. The major handling problem is the possible release of gaseous hydrogen sulfide, which is extremely toxic and flammable [118].

Rail is probably the most common transportation method. Liquid sulfur is typically carried in a tank-train system which can be loaded and unloaded through a single connection at rates up to 9000 L/min. This corresponds to a discharge rate for individual tank-wagons of 10 min. In addition, the single connection minimizes possible spillage, and allows single-point collection of hydrogen sulfide vapor.

Trucking is often used to transport sulfur from remote plants to a railhead. Trucks may haul from 22 t on single trailers to 35 t on dual trailers, depending on the terrain and local load restrictions. On privately owned purpose-built roads, payloads up to 87 t may be hauled, as at the Amoco-operated Whitney Canyon plant in Wyoming [119].

Where adequate waterways are available, molten sulfur is generally transported in barges, usually of 2500 t capacity and containing several insulated tanks, which like trucks, tankers, and tank-wagons contain coils heated by steam generated on board. During loading and unloading, the tanks are connected to the loading system by steam-jacketed piping, which provides loading rates up to 1120 t/h. Shipping of molten sulfur began in about 1945, and the proportion of sulfur delivered by barges increased very rapidly. The key routes are in the Mississippi – Missouri river system, and the Rhine.

Seagoing vessels, for liquid sulfur transport are 20 000 – 30 000 t deadweight tankers with isolated tanks and steam coil heating. The rapidly increasing costs of building specialized ships for sulfur transport are the main reason that the seaborne liquid sulfur trade is declining. The fleet is contracting and aging; there were 21 ships in 1992 [120].

Transporting liquid sulfur by heated pipelines within the process plant or refinery has been carried out for many years. There are also a small number of pipelines operating successfully to transport liquid sulfur over long distances, e.g., the 14.5-km undersea pipeline between the Grand Isle sulfur dome in the Gulf of Mexico and the United States mainland, and the 7-km electrically trace-heated line in the north of Germany [121], [122]. There is also a 20-km insulated and electrically trace-heated pipeline from the Berri natural gas plant to the Jubail sulfur terminal in Saudi Arabia [123].

Two main problems have tended to inhibit large-scale transport of sulfur by pipeline: the rather narrow temperature range, 120 – 150 °C, at which liquid sulfur can be transported; and the poor thermal conductivity of solid sulfur, which makes it impossible to remelt the contents of the pipeline should they solidify. A standby heating system, using electricity, steam, or hot water must therefore be available. The viscosity of liquid sulfur is almost ten times that of water, but the power requirement for a liquid sulfur pump need be only 1.8 times that required for pumping water [124].

9.2.3. *Formed Sulfur*

It is not always practical to handle and transport liquid sulfur. For many years the sulfur industry has sought a method of forming sulfur into regularly sized particles which are resistant to attrition and can be handled without the disadvantages associated with dry bulk sulfur.

Processes for forming sulfur have been used since the 1950s, when they were mainly employed in refineries and small-scale units [125]. Flaking — cooling of sulfur on a rotating steel drum — became common after it was developed by Escher Wyss. Slating, installed first by the Swedish-based steel company Sandvik, was used in relatively small-scale units. As sulfur production increased during the early 1960s, flaking became unsuitable, and air prilling, pelletizing (water prilling), and granulating processes were developed in Europe. It was not until the early 1970s, that large-scale sulfur forming facilities became widespread.

9.2.3.1. *Slating*

Slating is one of the most widely used forming techniques in terms of tonnage produced, and is also one of the easiest and least expensive. Molten sulfur is cast onto a continuous conveyor belt and cooled, directly or indirectly, with water so that it solidifies into a thin sheet, usually 3 –20 mm thick. As it reaches the end of the belt, this sheet is broken, or breaks naturally, into smaller pieces that resemble slates.

The first large-scale slating process was installed in 1970 by Vennard & Ellithorpe (now the Sulfur Services Division of Procor, Calgary, Canada) [126]. It uses a two-stage cooling system. From the first stage (tube-and-shell heat exchanger), where the sulfur is cooled to 120 °C, the sulfur is conveyed to the slating unit, which consists of an endless rubber belt with a small retaining ridge at each side. Sulfur solidifies at ca. 113 °C and is cooled to < 50 °C by immersing the belt in water.

Procor now offers a slating unit with two-stage cooling on the slating belt; the initial cooling occurs by aid of hooded fans; in the second stage the belt is submerged in cooling water at 37 °C. The slates are irregular in shape, but have uniform thickness of 5 – 10 mm. The moisture content of the slates is said to be < 1 % [127].

In the Sandvik process, molten sulfur is fed from a steam-heated feeder on the top of a moving endless steel belt, which is cooled by water sprayed against its underside. At the end of the belt, the thin, solid strip of sulfur has cooled to a temperature of ca. 90 °C, and is broken into slates (Fig. 54). As there is no direct contact with the cooling water, the product remains dry and the cooling water can be recycled [128].

Figure 54. The Sandvik slating process

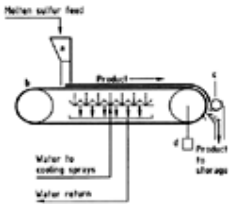

a) Steam-heated weir; b) Conveyor; c) Product breaker; d) Motor

The similar Rotoform process, also developed by Sandvik, uses a patented rotating dropformer, the Rotoformer (Fig. 55), which forms individual lenticular particles or pastilles which are laid on a steel belt, and cooled from the underside, to solidify them. The belt is much shorter than that used in the slating process. The process is economical and reliable, and environmental pollution is negligible [129]. A similar, fully automatic process, the Kaiser Rollomat, is available [130].

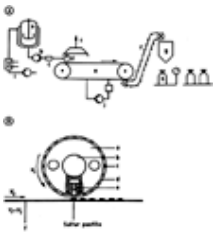

Figure 55. The Rotoform process

A) General view a) Melting/mixing vessel; b) Pump; c) Filter; d) Drop former; e) Steel belt cooler; f) Elevator; g) Intermediate storage silo; h) Bagging and weighing system; i) Thermo-oil heating system; j) Water cooling system;

B) Sectional view of the Rotoformer a) Stator; b) Product channel; c) Bore; d) Rotating shell; e) Metering bar; f) Steel belt

9.2.3.2. *Wet Pelletizing (Water Prilling)*

In the Sulpel process, developed by Elliot Associated Developments, London, and now offered by Humprey & Glasgow [131], [132], molten sulfur is dropped through specially designed nozzles into water, maintained in a vortex by swirl jets. The sulfur solidifies into spherical pellets, 0.5 – 6 mm in diameter with a bulk density 1280 – 1400 kg/m^3, which are removed from the bottom of the quench tank, dewatered, and, if necessary, dried (Fig. 56).

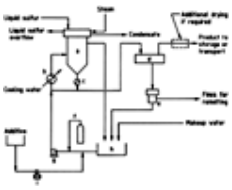

Figure 56. The Sulpel process

a) Pelletizer; b) Solution cooler (or air- cooled exchanger); c) Rotary valve; d) Dewatering screen; e) Fines screen; f) pH control (ammonia); g) Solution pump; h) Solution tank; i) Metering pump

In 1975, the Capsul process was introduced by Agra Chemicals, Canada, and Frit Industries, United States [133]. Molten sulfur is fed from a surge tank to the pelleting tank through a vortex chamber nozzle, which produces a diverging conical sheet of sulfur. Pellets from the bottom of the tank are raised by an air lift onto a dewatering screen and dried (Fig. 57). The bulk density of the pellets is 1152 kg/m^3.

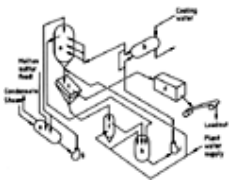

Figure 57. The Capsul process

a) Pelleting tank; b) Heat exchanger; c) Dewatering screen; d) Drier (optional); e) Molten sulfur surge tank; f) Settling tank; g) Molten sulfur pump; h) Water tank; i) Water circulating pump

Real International Marketing, Calgary, Canada, offers the RIM process, previously known as the Fletcher process. Sulfur is distributed under gravity through a heated perforated plate distributor into the quench water. The pellets, which are of irregular shape, with a size range of 1 – 10 mm and bulk density 1161 kg/m^3, are then fed to dewatering screens (Fig. 58).

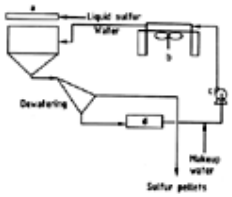

Figure 58. The RIM process

a) Forming unit; b) Cooler; c) Water pump; d) Separator

PV Commodity Systems, Canada, have developed a similar process; sulfur is distributed into quench water with the aid of a perforated plate; a drag-flight conveyor, which removes the pellets, also induces the required movement of the quench water.

Liquid Terminals, Corpus Christi, have developed a process based on the submerged spray technique. The product "nuggets", with a particle size between 10 mm and fines, are markedly less prone to dusting than is slated sulfur. The capacity of a standard commercial unit is 85 t/h [134].

Other slating and pelletizing processes are described in the literature [125], [128], [135], [136].

9.2.3.3. *Air Prilling*

The first air prilling process was developed by Outokumpu Oy, Finland, and was operated between 1962 and 1977. Sulfur is sprayed by the prilling head, located in the prilling tower, together with compressed air and cooling water. The prills are 0.5 – 3 mm diameter, with bulk density 1200 – 1300 kg/m^3.

In the 1960s, a new air-prilling process was developed by Ciech in Poland. The world rights to this technology, known as the Polish air-prilling process, are now owned by Intcan/Technoex, Canada. Liquid sulfur is filtered, and cooled in tube-and-shell heat exchangers to 125 °C. The sulfur is then transported via a feed tank to the top of the prilling tower 65 – 90 m high, 12 – 24 m diameter (Fig. 59). The prills are 1 – 6 mm diameter, with bulk density 1040 – 1200 kg/m^3 [125], [135].

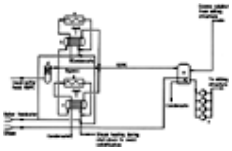

Figure 59. The Polish air-prilling process

a) Aerial condenser; b) Cooler I; c) Cooler II; d) Sulfur filter; e) Sulfur feed tank; f) Sulfur feed pumps (two spares)

9.2.3.4. *Other Prilling Processes*

Union Oil, California, has developed a process in which the product, called popcorn sulfur, is formed by spraying equal amounts of molten sulfur and water through a high-pressure nozzle directly onto a stockpile. The product, small particles, which have to be drained, is used in agriculture.

The Brazilian oil company, Petroleos Brasileiros, developed at its Gabriel Passos refinery an open-air granulation system in which the liquid sulfur is sprayed together with cooling water onto a rotary plate. The sulfur droplets produced solidify rapidly before falling onto a pile below. The water quickly drains from the pile, so that the sulfur moisture level falls to ca. 4 % after ca. 50 h. The investment cost of this simple installation is very low [137].

9.2.3.5. *Drum Granulation: The Procor GX Process*

Procor have developed the GX granulation process, based on the enlargement of "nuclei" or "seed" material in a granulation drum (Fig. 60). Liquid sulfur is sprayed through a series of nozzles to the already solidified particles, falling from the lifting flights of the drum, coating them and increasing their size. The system uses humidified air at ambient temperature as coolant. The drum is 2.4 m diameter, and 10 m long. The seeds, displaced from the drum are 1 – 6 mm diameter [127], [135], [138].

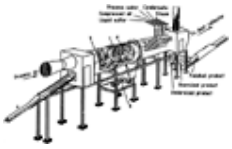

Figure 60. The Procor GX granulation unit

a) Seed feed conveyor; b) Sulfur spray header; c) Cross-flow fans; d) Humidification nozzles; e) Granulation drum; f) Dust preseparator; g) Drum drive; h) Thermocouple; i) Screen

9.2.3.6. *Spouted Bed Granulation*

PEC Engineering, Paris, offers the Perlomatic process, based on the spouted bed principle. Sulfur is sprayed into an upward current of fluidizing air in the throat of a Venturi vessel; it is cooled by the air, and granulates in the chamber above the Venturi cone (Fig. 61). Some of the granules leave the spouting bed with the stream of cooling air; the rest fall down and are removed to screen. The spherical granules produced have a size of 1 – 5 mm, and bulk density ca. 1000 kg/m^3 [139].

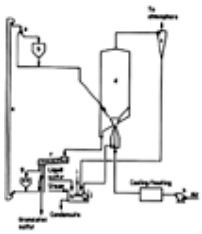

Figure 61. The Perlomatic process

a) Distribution box; b) Recycle silo; c) Cyclone; d) Perlomatic granulator; e) Recycle bucket elevator; f) Screen; g) Crusher; h) Buffer tank; i) Agitator; j) Liquid sulfur pump; k) Fan

9.3. Special Forms and Grades

There are a number of special forms of elemental sulfur, produced for particular agricultural and industrial applications. Their most important characteristic is the fineness of particle size. These special grades are generally referred to as "refined sulfur"; they account for not more than 2 % of total sulfur production.

Ground sulfur is produced by milling to a particle size of 100 – 325 mesh; it is used for protection of some agricultural products, as a fertilizer, and as a vulcanizing agent.

Granulated sulfur is broken sulfur in all grain fractions from 50 mesh to 50 mm.

Micronized sulfur, which is even finer than ground sulfur (< 5 µm), is produced either by milling or via an emulsion of sulfur in superheated water with subsequent rapid cooling; it is used as a pesticide.

Flowers of sulfur is a powdered form produced by sublimation of sulfur. It is used in rubber vulcanization, the pharmaceutical industry, and as a fungicide.

Insoluble sulfur is the µ form of sulfur, insoluble in carbon disulfide. It is produced as the residue after extraction of flowers of sulfur with CS_2, or by fast quenching of molten or vaporized sulfur with subsequent extraction by CS_2 to remove soluble allotropes. It is used in the rubber industry, and for production of high-pressure lubricants.

[Top of Page]

10. Uses

More than 85 % of the total world production of elemental sulfur is converted to sulfuric acid. More than 50 % is used for the production of fertilizers (together with pesticides, insecticides, and fungicides, agriculture consumes more than half of the sulfur produced).

The remaining 35 % of the sulfur converted into sulfuric acid goes into the production of detergents, pharmaceuticals, petroleum catalysts, synthetic resins, titanium pigments, viscose, acetates, and explosives; other uses of sulfuric acid are as a pickling agent in steel production, and leaching of nonferrous ores.

An important nonacid application of sulfur is in the manufacture of carbon disulfide, which can be synthesized either directly from the elements or by the reaction of sulfur with methane, with hydrogen sulfide as byproduct. The principal use of carbon disulfide is the production of cellulose xanthate in the manufacture of products such as viscose rayon or cellophane. Carbon disulfide is also used as a solvent for sulfur, iodine, phosphorus, and rubber.

In the sulfite process for manufacturing wood pulp for the paper industry, sulfur is burned to SO_2, which is then converted to sulfurous acid and subsequently to sulfite, a principal component of the cooking liquor for wood chips. For a variety of reasons, the use of sulfur in pulp making is decreasing [140]. Sulfur dioxide is also used for refining and bleaching in the sugar industry.

Soluble and insoluble sulfur is used as a vulcanizing agent for rubber. Other users of elemental sulfur are the pharmaceutical, cosmetics, photographic, soap, and dye industries.

A survey of sulfur use in the United States in 1989 and 1990 is given in Table 3 [141].

Table 3. Sulfur demand by end use in the United States (kt)

	1989	1990
Food and related products	*	*
Pulp and paper products	8	30
Synthetic rubber and other plastics	*	*
Agricultural chemicals	277	324
Other chemical products and industrial organic chemicals	122	187
Soaps and detergents	*	*
Petrol refining and petroleum coke products	142	104
Other industrial inorganic chemicals	188	126
Unidentified	372	538
Sulfuric acid		
Domestic sulfur	8 241	7 870
Imported sulfur	2 245	2 528
Total domestic uses	11 595	11 707

* Not available separately, included in unidentified.

10.1. Agriculture

Plant Nutrition. Sulfur is, after nitrogen, phosphorus, and potassium, the fourth major plant nutrient, and is essential for crop growth. Its vital role is to form the amino acids methionine, cystine, and cysteine, which are crucial to the formation of proteins. Sulfur reduces the quantity of nonprotein nitrogen and nitrate; it is also necessary for the formation of chlorophyll, enzymes, and vitamins [142-145].

There is undoubted evidence of sulfur deficiency in soil in some areas of the world [146-148]. A slight deficiency affects crop yield and quality, while the symptoms of severe deficiency are yellowing of the leaves and dwarfing of the plant. The main reasons for the sulfur deficiency are:

1. Increasing levels of specific agricultural production, with proportional increase of sulfur uptake
2. Shift in fertilizer practices, from ammonium sulfate and single superphosphate, to multinutrient (compound) fertilizers with little or no sulfur
3. Decrease in the atmospheric sulfur supply, owing to increasing environmental controls on sulfur dioxide emissions
4. Decline in the usage of sulfur-based herbicides and pesticides

The crop yield in sulfur-deficient areas can be improved by application of sulfur- containing fertilizers. The common dry fertilizers of this type are ammonium sulfate, gypsum, single superphosphate, and potassium sulfate; the liquid fertilizers are ammonium thiosulfate, ammonium bisulfate, and ammonium polysulfate. Sulfur can also be added to some nonsulfur fertilizers (in combination with phosphate rock, sulfur-coated urea, and potassium chloride) [148]. Sulfur – bentonite is a new type of fertilizer, typically 10 % bentonite and 90 % sulfur. The bentonite swells in contact with water and the sulfur particles disintegrate to particles of varying size, which secures the availability of sulfur to the plant over a long time period [149].

Fungicides and Pesticides. Fine elemental sulfur has been used traditionally as a fungicide. Because of the development of highly effective organic fungicides, this use of elemental sulfur is declining. Nevertheless, sulfur, generally applied by spraying, has the advantage of not being consumed by the plant; residues are washed off by rain and act as a nutrient in the soil.

Organic sulfides, sulfones, and sulfonic acid esters are used as long-acting, low-toxic acaricides (anti-mite agents) [150].

10.2. Construction Materials

Sulfur Concrete. Two types of sulfur-containing concrete can be distinguished [151-153]:

1. Hot mixed/hot poured sulfur concrete, in which plasticized sulfur is used instead of portland cement as binder; conventional materials, sand and coarse aggregate, are added
2. Sulfur-impregnated concretes, in which conventional materials such as coarse cement – water – aggregate are impregnated with liquid sulfur, which helps to increase both strength and resistance to corrosion

Sulfur concrete has excellent properties for use in corrosive environments, it sets quickly, can be used in freezing conditions, and is less porous than conventional concrete. The main drawback is its loss of structural strength above 105 °C, as well as its flammability, which excludes it from commercial and residential construction.

The United States Bureau of Mines has developed a modified corrosion-resistant sulfur concrete known as Chempruf [154]. The technology and equipment for its production and pouring is owned by National Chempruf, a United States, British, and Danish consortium, which also markets the product. Chempruf consists of 16 % sulfur cement, 38 % coarse aggregate, 38 % fine aggregate, and 8 % mineral filler. The sulfur cement itself consists of 95 % sulfur and 5 % plasticizer. Chempruf is extremely resistant to most mineral acids and salts and its strength develops very rapidly, to a level higher than that of Portland cement concrete. It has been commercially used for replacement of Portland cement curbs supporting the rails in the cell house of a United States zinc plant. The cost of Chempruf, 2 – 4 times that of Portland cement concrete, limits its use to highly corrosive environments.

Sulfur – Asphalt Compositions. The use of sulfur to modify asphaltic binders, with the aim of improving their properties, has been known for well over 100 years. The addition of sulfur to asphalt compositions increases their resistance to wear and to diesel fuel and gasoline. The viscosity of a sulfur – bitumen mix is lower than that of pure bitumen, which makes its preparation and laying easier. The main concern is the possibility of evolution of hydrogen sulfide during processing of the mix. There are essentially two preparation methods:

1. The sulfur, asphalt, and aggregate are mixed directly together
2. The asphalt binder is first blended with the aggregate, and the sulfur added subsequently

The use of sulfur – asphalt has been studied in many road trials in Canada, United States, United Kingdom, France, and Saudi Arabia [155-159], but no large-scale application of this material is known.

10.3. Sulfur-Containing Batteries

The lithium – sulfur dioxide battery, which has significant advantages over the conventional zinc – carbon battery, is manufactured and marketed in Europe by Vidor, a subsidiary of Crompton Parkinson. However, the cost of electricity generated by this type of battery is ca. 6 – 7 times that of a conventional cell [160].

The sodium – sulfur battery has about four times the storage capacity of a lead – acid accumulator, on a capacity/weight basis, and is seen as a possible energy source for electrical vehicles [161], [162]. Batteries of this type, which operate at ca. 350 °C and require high-purity sulfur, are at present under development by several companies. Because of the small quantity of sulfur per battery (ca. 80 – 90 kg per vehicle), even extensive production of such vehicles would not substantially influence the sulfur market.

10.4. Other Potential Uses

Another potential application of sulfur is the use of sulfur foams in road building in permafrost areas. The foam, with carbon dioxide as blowing agent, is laid as an insulating layer in the foundations of roads to prevent subsidence due to thawing

[163-166]. The foams apparently show very good resistance to freeze/thaw cycles. Foamed sulfur has also been proposed as sheathing for electric cables, on account of its very high resistivity.

Elemental sulfur possesses excellent resistance to attack by a wide variety of acid and salts. Trials have therefore been conducted on the use of plasticized sulfur as a protective coating for conventional concrete surfaces exposed to corrosive conditions around chemical plants, and as a sealant for earth containment dikes around petroleum product storage tanks [163], [167]. Sulfur impregnation of ceramic tiles, to reduce water absorption and increase frost resistance, has also been patented [168].

[Top of Page]

References

1. O-A. Neumüller: *Römpps Chemie-Lexikon*, 8th ed., **vol. 5**, Franckh'sche Verlagshandlung, Stuttgart 1987, p. 3748 ff.
2. R. Steudel, *Top. Curr. Chem.* **102** (1982) 149 – 176. Links
3. R. Steudel, H. J. Mäusle, *Chem. unserer Zeit* **14** (1980) 73 – 81. Links
4. M. Pourbaix, *Bull. Soc. Chim. Belg.* **53** (1944) 155.
5. R. F. Bacon, R. Fanelli, *Ind. Eng. Chem.* **34** (1942) 1043 – 1048.
6. F. Feher, *Angew. Chem.* **67** (1955) 337 – 339. Links
7. P. W. Schenk, *Angew. Chem.* **67** (1955) 344 – 347. Links
8. P. W. Schenk, *Z. Elektrochem.* **63** (1959) 1002 – 1005.
9. W. N. Tuller: *The Sulphur Data Book*, McGraw-Hill, New York 1954. J. Chao, *Hydrocarbon Process.* **59** (1980) no. 11, 217–223.
10. R. Winter, P. A. Egelstaff, W. C. Pilgrim, W. S. Howells: “The Structural Properties of Sulphur,” *Liq. Matter, Proc. Liq. Matter Conf. Eur. Phys. Soc. 1st* 1990, 215 – 218.
11. F. J. Touro, T. K. Wiewiorowski, *J. Phys. Chem.* **70** (1966) no. 1, 239 – 241.
12. J. W. Palm: “Sulphur Recovery Plant Operation,” *Sulphur* **143** (1979) July/Aug., 34 – 39.
13. *Gmelin*, System no. 9, vol. A, p. 537 ff.
14. R. C. Weast (ed.): *CRC Handbook of Chemistry and Physics*, 58th ed., CRC Press Inc., Cleveland, Ohio 1977, pp. B 274 – 275.
15. *Gmelin*, Schwefel; Teil A, Verlag Chemie, Weinheim, Germany 1953.
16. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances*, Springer Verlag, Berlin 1973.
17. O. Kubaschewski: *Metallurgical Thermochemistry*, Pergamon Press, Oxford 1967.
18. K. H. Simmrock, R. Janowsky, A. Ohnsorge: *Critical Data of Pure Substances*, Dechema, Frankfurt 1986.
19. *1977 Keystone Coal Industry Manual*, McGraw-Hill, New York 1977, pp. 561 – 719.
20. *Ullmann*, 4th ed., **14**, 477.
21. Coal, Gold and Base Minerals of Southern Africa (Apr. 1971), 79 – 85.
22. C. J. Thompson, E. M. Shelton, H. J. Coleman, *Hydrocarbon Process* **55** (1976) no. 2, 73 – 76.
23. Jahresbericht 1992 des Mineralölwirtschaftsverbandes der Bundesrepublik Deutschland, Hamburg.
24. Bundesministerium für Wirtschaft der Bundesrepublik Deutschland, 1991.
25. Elcor Chem. Co., GB 1 241 575, 1971.
26. *Sulphur* **147** (1980) 36 – 38.
27. *Chem. Eng. News* **45** (1967) no. 20, 21; **46** (1968) no. 26, 11, 12.
28. *Chem. Eng. (N.Y.)* **46** (1968) no. 1, 16 – 18; **46** (1968) no. 6, 94, 95.
29. D. Coghill, GB 1 245 759, 1971.
30. M. Gatto: “Trattamento Mineralurgico del Minerali di Zolfo”, Turin (1928).
31. T. P. Forbath, *Min. Eng.* **5** (1953) no. 881 – 885.
32. Chemical Construction Corp.: US 2 537 842, 1951 (P. J. McGauley, E. S. Roberts, S. Strelzoff).
33. D. R. Muir, J. W. Terry: “The Evolution of Sulphur,” *Sulphur '82: Proceedings of the International Conference*, vol. II, The British Sulphur Corporation Ltd., London 1982, pp. 596 – 600.
34. *Chem. Week* **103** (1968) Oct. 26, 28.
35. Brameda Resources Ltd., US 3 619 147, 1971 (C. K. Amano, C. J. Lewis).
36. Alaska Junean Gold Mining Co., US 2 934 414, 1960 (P. R. Bradley).
37. Delaware Chemical Engineering Co., US 1 963 921, 1934 (A. Nagelvoort).
38. Delaware Chemical Engineering Corp., US 2 088 190, 1937 (F. I. Du Pont).
39. Sulfan Exploration Syndicate, *Q. Bull.* **5** (1954) 43 – 46.
40. S. V. Griffith, *Min. Mag.* **64** (1941) 123 – 130.
41. Y. Fujita, *Eng. Min. J.* **157** (1954) no. 8, 104 – 106.
42. *Sulphur* **35** (1961) 36 – 39.
43. F. Benitez, *Eng. Min. J.* **157** (1956) no. 6, 96 – 98.
44. B. Zakiewicz, *Sulphur* **120** (1975) 35 – 43.
45. Z. Gatarz, J. Kirejcyk, J. Krajewsky, J. Klimek: “The Development of Sulphur Mining and Processing in Poland,” *Sulphur '82*, vol. I, The British Sulphur Corporation Ltd., London 1982, pp. 355 – 360.

46. H. G. Thode, R. K. Wanless, R. Wallouch, *Geol. Soc. Am. Bull.* **64** (1953) 1482.
47. *Sulphur* **23** (1958) 34 – 36.
48. *Sulphur* **56** (1965) 27 – 29.
49. W. W. Duecker, J. R. West: *The Manufacture of Sulphuric Acid*, Rheinhold, New York 1959, p. 48.
50. *Sulphur* **209** (1990) 29 – 34.
51. *Sulphur* **224** (1993) 35 – 37.
52. *Sulphur* **153** (1981) 22 – 27.
53. *Sulphur* **196** (1988) 19 – 22.
54. *Sulphur* **207** (1990) 28 – 33.
55. *Sulphur* **195** (1988) 9.
56. Sulphur Fact File, Cancrude Consulting Inc., Calgary, Alberta, Canada, Oct. 1992.
57. "The Status of Claus Sulphur Recovery Process," *Sulphur* 1986, Nov.–Dec., 3 – 10.
58. J. A. Lagas: "Selection of Gas Sweetening Processes," *Comprimo/Western Research Gas Sweetening and Sulphur Recovery Seminar*, Nov., Amsterdam 1981.
59. J. C. King *et al.*: "Rigorous Screening Selects Sour-Gas Process," *Oil Gas J.* **84** (1988) no. 36, 101 – 110.
60. J. C. King: "H_2S Removal by Direct Liquid Phase Oxidation Processes," *35th Annu. Southwestern Petroleum Short Course-Ass. 88*, 1988.
61. D. A. Dalrymple, T. W. Trofe, J. M. Evans: "An Overview of Liquid Redox Sulfur Recovery," *Chem. Eng. Prog.*, March, 1989, 43 – 49.
62. H. Paskall: "The Claus Sulphur Recovery Process," *Sulphur* **160** (1982) no. 38 – 43. TA Luft, C. Heymanns Verlag, Köln 1986, p. 81.
63. H. Fischer: "Sulphur," in A. I. More (ed.): *Sulphur, Sulphur Dioxide and Sulphuric Acid,* The British Sulphur Corp. Ltd., London 1984, pp. 3 – 148.
64. B. W. Gamson, R. H. Elkins: "Sulfur from Hydrogen Sulfide," *Chem. Eng. Prog.* **49** (1953) 203 – 215. Links
65. H. G. Paskall: "Capability of the Modified Claus Process," Report to the Alberta Department of Energy and Natural Resources, Western Research & Development, Calgary, Mar. 1979.
66. A. Meisen, H. A. Bennet: "Consider all Claus Reactions," *Hydrocarbon Process.* **54** (1974) no. 11, 171–174.
67. W. Nehb: "LURGI Claus Technology Using Oxygen," *Western Research Sulphur Recovery Seminar*, Budapest, Nov. 1989.
68. W. Nehb: "Claus and Oxygen Fit Together," *Western Research Sulphur Recovery Seminar*, Budapest, Sept. 1992.
69. "Oxygen Enrichment in Claus Process Sulphur Recovery Units," *Sulphur* **195** (1988) no. 3 – 4, 24 – 30.
70. B. G. Goar et al.: "The COPETM Process: A Proven Oxygen-Based Technology for Improving Claus Plant Designs and Operations," Sulfur 87, Int. Conf., Houston, Texas, April 1987.
71. "Cost Effective Options to Expand SRU Capacity Using Oxygen", presented by M. Rameshni, Parsons Energy & Chemicals Group, Inc., at Sulfur Recovery Symposium Brimstone Engineering Services, Inc., Banff, Alberta, Canada, May 2002.
72. N. P. Liebermann: *Trouble Shooting Process Operations*, PennWell Publ. Co., Tulsa, Oklahoma pp. 83–92.
73. "Optimising the Recovery of Sulphur in Oil Refineries", Matthias Franz, Pier Enterprises, Kai Torp, PPM Systems Oy, www.engineerlive.com, Health/Safety, 2004.
74. J. B. Hyne: "Sulfur Recovery – New Challenges in the Nineties," *Laurance Reid Gas Cond. Conf. 41st*, Norman, Oklahoma 1991, pp. 1 – 21.
75. J. Borsboom, J. A. Lagas: "The SUPERCLAUS Process Increases Sulphur Recovery," *22nd ACHEMA Exhibition-Congress*, Frankfurt June 1988.
76. J. Borsboom, B. G. Goar, G. Heijkoop, J. A. Lagas: "SUPERCLAUS: Performance World-Wide," *Sulphur* **220** (1992) nos. 5 – 6, 44 – 47.
77. SUPERCLAUS®/EUROCLAUS® The Latest Developments, Brimstone Symposium Banff, Canada, 2003, H. Borsboom, Jacobs Nederland B.V., Rien van Grinsven, Jacobs Nederland B.V., Peter van Nisselrooij, Gastec Technology B.V., Eric Butler, Chevron Canada Limited.
78. B. G. Goar: "Emerging New Sulfur Recovery Technologies," *Laurance Reid Gas Cond. Conf.*, Norman, Oklahoma 1989, pp. 87 – 115.
79. B. G. Goar, C. F. Maitland: "Selectox – A New Approach to Sulphur Recovery From Lean Acid Gas and Tailgas Streams," Sulphur 84 Int. Conf. 3rd, 1984, pp. 347 – 353.
80. K. Stevens, W. H. Buckhannan: "Enhanced Process Configuration for the CBA Process," *Sulphur* **225** (1993) nos. 3 – 4, 37 – 48.
81. R. Lell, J. B. Nougayrede: "Reducing Claus Plant Sulphur Emissions With SULFREEN," *Sulphur* **213** (1991) nos. 3 – 4, 39 – 46.
82. LURGI Information Brochure, SULFREEN Catalytic Reduction of Claus Plant SO2 Emission, Frankfurt 1992.
83. G. Bandel, W. Willing: "Sulfur 96" Conference, Vancouver, Canada, October 20 – 23, 1996.
84. R. E. Heigold, R. G. Holcek: *Enhanced Sub-Dewpoint Claus Process (99 % Plus)*, Delta Projects Ltd., Calgary, Alberta pp. 359 – 366.
85. W. Dam, J. A. Hugill, M. L. J. A. Wetzels: "The SCOT Process Still Very Much Alive; Nearly 20 Years After its First Introduction," *Western Research Seminar*, Budapest, Nov. 1990.
86. J. A. Lagas, M. L. J. A. Wetzels: "Recent Developments in SCOT Tail Gas Technology," Sulphur 93, Hamburg, Apr. 1993.

87. T. R. Armstrong, J. C. Bourdon, G. N. Tragitt: “The SULFTEN System: Commercial Experience With a New Tail Gas Treating Technology,” *Nat. Pet. Ref. Assoc. Annu. Meet.*, San Antonio, Texas March 1985.
88. “Regenerative H_2S Removal Processes,” *Sulphur* **218** (1992) 22 – 34.
89. W. I. Echt, C. J. Wendt, Jr. “Reduce Sulfur Emissions from Claus Sulfur Recovery Unit Tail Gas Treaters,” *AIChE Spring National Meeting Gas Processing: Sulfur Recovery and Tail Gas Clean Up*, Houston, March/April 1993.
90. D. K. Beavon, R. H. Hass: “BSR/Selectox I Sulfur Recovery Process for Claus Tail Gas Treating,” *The 5th Canadian Symposium on Catalysts*, Calgary, Alberta, Oct. 1977, pp. 76 – 84.
91. LURGI Information Brochure, Wellman – Lord Regenerative Flue Gas Desulphurization in Oil Refineries, Sulphur 93, Hamburg, Apr. 1993.
92. M. P. Heisel, F. J. Marold: “CLINTOX Claus Tail Gas Treatment,” *Gas Sep. Purif.* **1** (1987) Dec., 107–109.
93. *Sulphur* **276** (2001), Sept./Oct.
94. H. Ullrich, *VT Verfahrenstech.* **12** (1978), no. 10, 648 – 649.
95. “Degassing Liquid Sulphur: A Step Towards Safer Handling and Storage,” *Sulphur* **173** (1984) 36 – 37.
96. T. Tork, M. M. Weiss: “Natural Gas Sweetening”, *Hydrocarbon Eng.* **8** (2003) 41 – 46
97. “Modern Concepts for Natural Gas Sweetening”, presentation held by U. Koss, Lurgi AG, during “Oil & Gas” seminar at Giprogazoochistka, Moscow, October 2003.
98. Reduction of Sulphur Dioxide Emissions from Non-Ferrous Smelters, Seminar, Toronto, Ontario, Canada, Aug. 1986, 16 – 17.
99. W. D. Hunter, A. W. Michael, *AIChE Symp. Ser.* **71**, no. 148, 320 – 323.
100. U. Neumann, W. Willing, Paper at *Sulphur 93,* Hamburg, Apr. 4 – 7, 1993, pp. 97 – 98.
101. R. Lell, *Chem. Umwelt Tech.* **1988/89,**10 – 15.
102. P. Steiner, H. Jüntgen, K. Knoblauch, *Adv. Chem. Ser.* 1975, no. 139, 180 – 191. Links
103. W. F. Bischoff, P. Steiner, *Chem. Eng.* **82** (1975) no. 1, 74 – 75.
104. D. S. Moulton, *Meitan Kexue Tishu* **9** (1985) 729–738.
105. R. Malmström, T. Tuominen, *Adv. Extr. Metall Refin. Proc. Int. Symp.* 1971, 529 – 543.
106. A. Roine, T. Talonen, M.-L. Metsärina, *Proc. Copper 91-Cobre 91 Int. Symp.* **IV** (1991) Aug., 271–281.
107. Sumitomo Heavy Ind., JP 80 104 905, 1980.
108. Allied Chemical Corp., GB 1 552 886, 1977 (R. E. Sorg, R. L. Sturtevant, D. G. Frick).
109. Foster Wheeler Energy Corp., US 4 095 953, 1978 (C. Gutterman, P. Steiner).
110. Ishikawajima-Harima Heavy Industries Co. Ltd., JP 81 69 202, 1983.
111. Dow Chemical Co., EP 270 223, 1987 (C. Ayasse).
112. Mitsui Mike Kakoi Co. Ltd., JP 63 277 502, 1988 (M. Okamura, S. Chiyawantani, O. Matsunaga).
113. Hitachi Ltd., JP 58 207 925, 1983.
114. Mitsubishi Heavy Ind. Ltd., JP 03 08 413, 1991 (M. Suehiro *et al.*).
115. Exxon Research Co., GB 1 382 041, 1971 (G. Moss, R. W. Humphrey).
116. *Sulphur* **163** (1982) 22 – 28.
117. *Sulphur* **167** (1983) 34 – 37.
118. J. A. Lagas, *Hydrocarbon Process.* **61** (1982) no. 10, 85 – 89.
119. *Sulphur* **190** (1987) 16 – 19.
120. *Sulphur* **219** (1992) 22 – 25.
121. *Sulphur* **166** (1983) 28 – 33.
122. *Sulphur* **187** (1986) 14 – 17.
123. *Pipe Line Ind.* **61** (1984) no. 3, 39 – 41.
124. D. J. Cartwright, R. D. Waymouth, *J. Can. Pet. Technol.* 1971, 18 – 24.
125. *Sulphur* **142** (1979) 30 – 39.
126. *Sulphur* **87** (1970) 26.
127. *Sulphur* **202** (1989) 19 – 21.
128. *Sulphur* **188** (1987) 30 – 36.
129. *Sulphur* **163** (1982) 37 – 38.
130. *Sulphur* **210** (1990) 35 – 36.
131. *Sulphur* **80** (1969) 30 – 32, 35.
132. *Sulphur* **100** (1972) 84 – 86.
133. *Eur. Chem. News* **13th** June, 1975, p. 32.
134. *Sulphur* **126** (1976) 47 – 49.
135. *Sulphur* **208** (1990) 35 – 47.
136. B. G. Goar: *Proc. Gas Cond. Conf.* **34th** (1984) B/1 – B/26.
137. *Sulphur* **214** (1991) 42 – 44.
138. *Sulphur* **174** (1984) 32 – 35.
139. *Sulphur* **118** (1975) 40 – 41.
140. *Sulphur* **186** (1986) 14 – 16.

141. *Sulphur* **219** (1992) 19 – 21.
142. *Sulphur* **153** (1981) 37 – 39.
143. *Sulphur* **165** (1983) 35 – 41.
144. *Sulphur* **176** (1985) 14 – 16.
145. *Sulphur* **215** (1991) 19 – 23.
146. *Sulphur* **164** (1983) 22 – 25.
147. *Sulphur* **184** (1986) 17 – 20.
148. *Sulphur* **182** (1986) 28 – 30.
149. *Sulphur* **192** (1987) 15 – 16.
150. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel*, Springer Verlag, Berlin 1970, p. 541 ff.
151. *Sulphur* **147** (1980) 29 – 31.
152. *Sulphur* **161** (1982) 32 – 34.
153. A. Oshio *et al.*: *Onoda Kenkyu Hokoku* **37** (1985) no. 113, 40 – 56.
154. *Sulphur* **199** (1988) 35 – 36.
155. *Sulphur* **150** (1980) 29 – 30.
156. *Sulphur* **141** (1979) 42.
157. *Sulphur* **170** (1984) 30 – 33.
158. *Sulphur* **161** (1982) 38 – 39.
159. *Sulphur* **181** (1985) 44 – 48.
160. *Sulphur* **155** (1981) 40.
161. *Sulphur* **196** (1988) 39.
162. *Sulphur* **225** (1993) 33 – 34.
163. *Sulphur* **122** (1976) 45 – 50.
164. *Sulphur* **125** (1976) 19 – 22.
165. *Sulphur* **131** (1977) 35 – 37.
166. G. L. Woo *et al.*, *Adv. Chem. Ser.* **165** (1978) 227–240.
167. *Sulphur* **126** (1976) 44 – 45.
168. Tile Council of America, US 3 208 190, 1965 (M. J. Kakos, J. V. Fitzgerald).

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Sulfur Dioxide

Hermann Müller, Lurgi Metallurgie GmbH, Frankfurt/Main, Federal Republic of Germany

1. Introduction 2
2. Properties 2
3. Raw Materials for Sulfur Dioxide Production 4
4. Production from Elemental Sulfur . 6
4.1. Sulfur Burning Processes 7
4.2. Heat Recovery 8
4.3. Impurities and their Removal 8
4.4. Two-Stage Sulfur Combustion . . . 9
5. Production by Pyrite Roasting . . . 10
5.1. Importance of Impurities 10
5.2. Thermodynamics and Kinetics . . . 10
5.3. Roasting Processes 11
5.4. Roasting of Arsenical Pyrite 15
5.5. Production in Nonferrous Metal Sulfide Ore Processing 16
5.5.1. Copper Smelting 16
5.5.2. Nickel Smelting 19
5.5.3. Zinc Blende Roasting 19
5.5.4. Lead Smelting 21
5.6. Production by Decomposition of Sulfuric Acid and Sulfates 23
5.6.1. Waste Sulfuric Acid Processing 23
5.6.2. Sulfate Decomposition 27
5.7. Production from Waste Gases 30
5.7.1. Combustion of Hydrogen Sulfide . . . 30
5.7.2. Recovery from Flue Gases 31
6. Cleaning of Crude Sulfur Dioxide Gas Streams 32
6.1. Constituents of Roaster and Metallurgical Gases 32
6.2. Cleaning Systems 33
7. Production of Pure Sulfur Dioxide . 36
7.1. Chemical Production from Sulfur Trioxide 36
7.2. Separation from Sulfur Dioxide-Containing Gases 37
8. Construction Materials 41
9. Industrial Output and Uses 42
10. Analysis 42
11. Environmental Aspects and Toxicology 43
12. References 44

Based on the corresponding article in Ullmann, 4th ed., written by Ulrich Sander and Ulrich Rothe (English Edition: British Sulphur Corporation, 1984)

1. Introduction

Sulfur dioxide is produced industrially in greater quantities than any other single sulfur compound. It is generated as the first stage in the manufacture of virtually all the sulfuric acid used by industry, irrespective of the basic raw material. It may arguably be considered the linchpin of the sulfur sector of the heavy inorganic chemicals industry.

Sulfur dioxide is generated intentionally by combustion of elemental sulfur or by oxidation of sulfides. Since most carbonaceous or hydrocarbon fuels and the ores of many nonferrous metals contain sulfidic sulfur, vast volumes are also produced involuntarily in large fuel-consuming installations, such as power stations, and in nonferrous metal smelters. On account of its chemical and physiological properties, coupled with the scale on which it is produced, waste sulfur dioxide is now considered as one of the gravest chemical threats to the global environment. Progress has been made in controlling atmospheric discharges of sulfur dioxide from the most amenable sources – an aim towards which much of the technological development relating to sulfur dioxide has been directed in the past decade. However, sulfur dioxide pollution, particularly where it crosses national boundaries, continues to be a matter for argument and debate within and between governments and industry.

Sulfur dioxide has been known since ancient times for its bleaching and disinfecting effects [1]. These days it finds widespread use in a rather diverse range of chemical and other industries, including wood pulping and paper production, textile manufacture and food processing, where it serves as reagent, bleaching agent, preservative or solvent. However, the outstanding use of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_569

sulfur dioxide – about 98 % of the total amount used by industry – is in the manufacture of sulfuric acid.

2. Properties

Physical Properties. Sulfur dioxide, SO_2, is a colorless, nonflammable, toxic gas with a characteristic pungent smell and acid taste. Its most important physical properties are summarized below [2–8]:

Molecular mass	64.06
Melting point (101.3 kPa)	−75.5 °C
Latent heat of fusion (at *mp*)	115.6 J/g
Dynamic viscosity at 0 °C	368 Pas
Density at − 10 °C	1.46 g/cm^3
Critical density	0.525 g/cm^3
Critical pressure	78.8 bar
Critical temperature	157.5 °C
Boiling point (101.3 kPa)	−10.0 °C
Latent heat of vaporization (at *bp*)	402 J/g
Standard density at 0 °C (101.3 kPa)	2.93 kg/m^3
Density relative to air (0 °C, 101.3 kPa)	2.263
Molar volume (0 °C, 101.3 kPa)	21.9 L/mol
Standard enthalpy of formation	−4636 J/g
Specific heat capacity c_P (101.3 kPa):	
0 °C	586 J kg^{-1} K^{-1}
100 °C	662 J kg^{-1}K^{-1}
300 °C	754 J kg^{-1}K^{-1}
500 °C	816 J kg^{-1}K^{-1}
c_P/c_v (15 °C, 1013 mb)	1.29

Graphical representations of the temperature functions of the various physical properties can be found in [2,6,9]. Interpolation formulas for thermodynamic characteristics such as specific heat are given in [2,5,7,10,11]. The vapor pressure curve of liquid sulfur dioxide as a function of temperature is shown in Figure 1. Figure 1 also shows the dewpoints of gases of varying sulfur dioxide contents at different pressures. These are important in the production of pure sulfur dioxide by condensation (see 7.1).

The solubility of sulfur dioxide in water rises in proportion to the SO_2 partial pressure in good conformity with Henry's law and is increased by lowering the temperature, as represented graphically in Figure 2 [12]. In solution, sulfur dioxide is present mainly as SO_2 molecules, but Raman spectroscopy confirms the presence of minor amounts of the species HSO_3^-, $S_2O_5^{2-}$ and H_2SO_3 [5]. The latter, sulfurous acid (the anhydride of which is sulfur dioxide), exists only in aqueous solution. Aqueous solutions of alkaline compounds will absorb much more sulfur dioxide than pure water because of the formation of hydrogensulfite and sulfite ions [12].

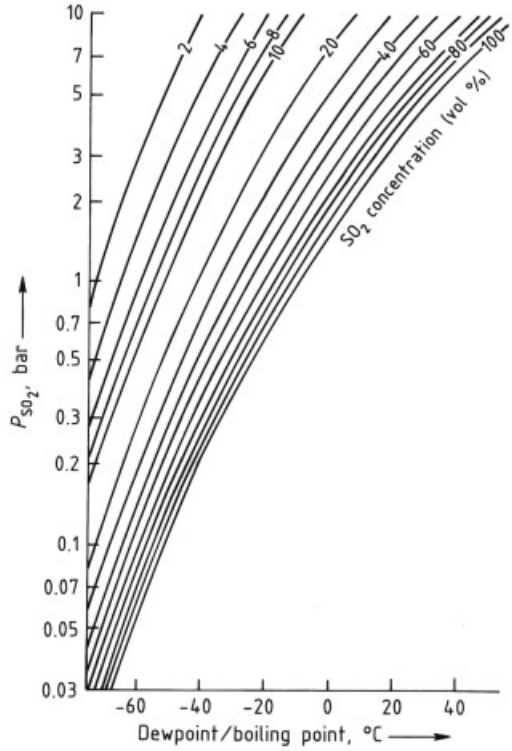

Figure 1. Vaporization curves for sulfur dioxide
Curve for 100 % SO_2 shows relationship of boiling point to equilibrium vapor pressure. Curves for impure SO_2 mixtures show dependence of dewpoint on SO_2 partial pressure

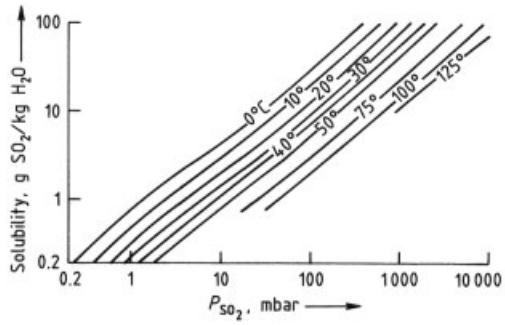

Figure 2. Solubility of sulfur dioxide in water [12]

The solubility of sulfur dioxide in sulfuric acid is of industrial importance. It first decreases with rising sulfuric acid concentration, reaching a minimum at an H_2SO_4 concentration of

ca. 85 %; beyond this concentration it increases again (Fig. 3) [13].

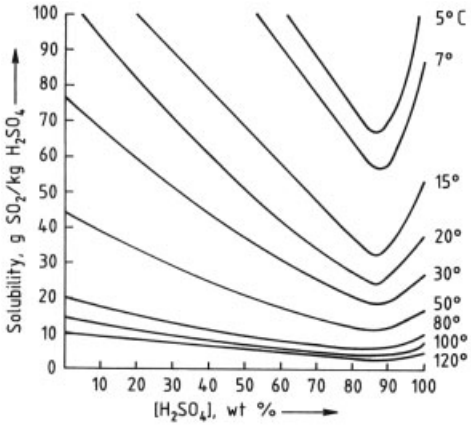

Figure 3. Solubility of sulfur dioxide in sulfuric acid at $p_{(SO_2)} = 101.3$ kPa [13]

Liquid sulfur dioxide is miscible in all proportions with liquid sulfur trioxide. Sulfur dioxide also dissolves readily in most organic liquids (e.g., methanol, ethanol, benzene, acetone, or carbon tetrachloride). With ether, carbon disulfide, chloroform, and glycol it is completely miscible, even at low temperatures [2,5].

Chemical Properties. Sulfur dioxide is very stable; thermal dissociation becomes significant only above 2000 °C. It can be decomposed by shock waves, irradiation with ultraviolet or X-rays, or by electric discharge [2,5].

The reaction of sulfur dioxide with oxygen to form sulfur trioxide is industrially the most significant of all its reactions because of its importance in sulfuric acid production. In the gas phase, it only takes place at elevated temperatures and, for a satisfactory yield of sulfur trioxide, it requires the presence of a catalyst (see → Sulfuric Acid and Sulfur Trioxide). In aqueous solution, sulfur dioxide is oxidized to sulfuric acid at low temperature by air in the presence of activated coke or nitrous gases, or by oxidizing agents such as hydrogen peroxide.

The reduction of sulfur dioxide with hydrogen, carbon or carbon compounds such as methane or carbon monoxide is also of industrial interest [2,5]. These reactions require high temperatures, or catalysts, or both. They result in mixtures of elemental sulfur and hydrogen sulfide. If carbon or a carbon compound is used as reducing agent, carbon- containing species such as carbon dioxide, carbonyl sulfide, and carbon disulfide are formed as well.

Sulfur dioxide reacts with hydrogen sulfide in accordance with the Claus reaction (Eq. 1). This is the basis of the main method by which elemental sulfur is recovered from hydrogen sulfide in industry (see also → Sulfur, Chap. 6.)

$$SO_2 + 2\,H_2S \longrightarrow 3\,S + 2\,H_2O \qquad (1)$$

Sulfur dioxide generally oxidizes metals at elevated temperatures, simultaneously forming metal sulfides and oxides [2,5].

Liquid sulfur dioxide is a relatively efficient solvent with some waterlike properties [2,3]. Polar inorganic compounds are usually insoluble or only sparingly soluble in liquid sulfur dioxide [2], whereas covalent inorganic and organic compounds are often dissolved, mostly forming stable solvates. The fact that aromatic hydrocarbons dissolve more readily than aliphatics in sulfur dioxide is exploited on an industrial scale for the extraction of aromatics from crude oil (Edeleanu process) [5].

3. Raw Materials for Sulfur Dioxide Production

Sulfur dioxide is manufactured industrially from the following raw materials:

1) Elemental sulfur
2) Pyrite
3) Sulfide ores of nonferrous metals
4) Waste sulfuric acid and sulfates
5) Gypsum and anhydrite
6) Hydrogen sulfide-containing waste gases
7) Flue gases from the combustion of sulfureous fossil fuels

Elemental Sulfur. Elemental sulfur is the most important raw material for sulfur dioxide production worldwide. For information on sources of elemental sulfur and production methods, see → Sulfur.

Today, elemental sulfur used for sulfur dioxide production is considered not only as a raw material but also, to an increasing extent, as an energy source. Modern sulfur-burning plants are

normally connected to a turbo – alternator combination for electrical power production; these are known as cogeneration units. Elemental sulfur is the raw material of choice for production of 100 % sulfur dioxide and of industrially significant sulfites, such as calcium hydrogensulfite solution, which is used in cellulose production. Of the 33.5×10^6 t of elemental sulfur produced worldwide in 1981, almost 28×10^6 t were burnt to sulfur dioxide for conversion to sulfuric acid [14]. The proportion of industrial sulfur dioxide production that is based on elemental sulfur varies in different countries. Of the total amount of sulfur-in-all-forms used for producing sulfur dioxide-containing gases for sulfuric acid production, the share of elemental sulfur was ca. 80 % in the United States, about 50 % in Germany, and probably less than 25 % in the former Soviet Union.

Pyrites. Pyrites and other iron sulfide ores still constitute today, in some countries, an important raw material basis for sulfur dioxide production, especially as the primary stage for sulfuric acid manufacture [15]. The most important iron sulfide minerals are pyrite and marcasite, both FeS_2, as well as pyrrhotite or magnetopyrite, the compositions of which correspond approximately to the formula Fe_7S_8. They occur in varying purity as sedimentary rocks in all formations, and as minor constituents in coal deposits. Pyrites are usually slate gray or sometimes greenish in color, depending on the deposit. Other constituents, apart from gangue materials that are present in varying proportions in pyritic ores include sulfides of nonferrous metals, especially of copper, zinc, and lead as well as, to a lesser extent, cobalt, nickel, manganese, bismuth, silver and gold. In some cases, pyrites may also contain arsenic, antimony, tellurium, and thallium as well as fluorides and chlorides.

The production of pyrites, either as crude pyrites or as beneficiated pyrites, has decreased drastically within the last years. This was a result of a depressed fertilizer market, lower sulfur prices, and increasing amounts of involuntarily produced sulfuric acid. Flotation pyrites are, however, obtained in markable quantities as a byproduct of nonferrous metal sulfide beneficiation plants. The sulfur content of pyrites concentrates may be as high as 50 %. Average analyses of pyrite ores of different origins are presented in Table 1.

Depending on the grain size, a distinction is made between flotation pyrites (grain size < 0.1 mm), fine pyrites (0 – 6 mm), coarse pyrites (1 – 12 mm and more), and lump pyrites (10 – 100 mm and above).

Whereas in Japan, for instance, all the pyrite ores produced are processed exclusively in the country, other countries still export pyrites as raw materials for sulfuric acid production, especially the former Soviet Union, Spain, and Norway. In 1982, the Soviet Union exported about

Table 1. Typical analyses of various types of pyrites

Country	Location	Grade*	Typical content, wt %					
			S	Fe	Cu	Zn	Pb	As
Germany	Meggen	flot.	43 – 48	37 – 40	0.03 – 0.1	0.4 – 1.2	0.4 – 0.6	0.06
Cyprus	Kalavassos	flot.	45 – 48	40 – 44	0.2 – 0.3	0.1	0.1	0.01
	Limni	flot.	44 – 46	37 – 40	0.2 – 0.3	0.1	0.1	
Finland	Outokumpu	flot.	43 – 48	40 – 44	0.2 – 05	0.1 – 1.1	0.3	0.03 – 0.06
Greece	Kassandra	flot.	47 – 49	41 – 44	0.1 – 0.3	0.6 – 1.5	0.5 – 0.7	0.5 – 1.8
Italy	Maremma	flot.	48 – 50	44 – 45	0.02	0.09	0.02	0.045
former Yugoslavia	Bor	flot.	44 – 49	41 – 43	0.25 – 0.6	0.04 – 0.1	0.04 – 0.1	0.03 – 0.07
Norway	Orkla	fine	41 – 46	37 – 40	2.2 – 2.7	1.6 – 2.0	0.05 – 0.06	0.04 – 0.08
	Folldal	flot.	48 – 50	44 – 45	0.06 – 0.07	0.07 – 0.08	0.04 – 0.05	0.03 – 0.05
Portugal	Aljustrel	fine	45 – 47	39 – 40	0.65 – 1.33	2.0 – 3.0	1.2 – 1.4	0.4 – 0.5
Romania	Baia Sprie	fine	44 – 45	37 – 40	0.6 – 0.7	4 – 5	1.0 – 1.1	0.03 – 0.05
Sweden	Boliden	flot.	49 – 50	45 – 46	0.04 – 0.15	0.2 – 0.4	0.04 – 0.05	0.05 – 0.2
Spain	Rio Tinto	flot.	48 – 49	43 – 44	0.15 – 0.4	0.08 – 1.0	0.08 – 1.0	0.26 – 0.4
	Tharsis	flot.	46 – 48	41 – 43	0.7 – 1.1	1.3 – 1.6	0.5 – 0.6	0.4 – 0.5
	San Telmo	fine	45 – 47	40 – 41	1.9 – 2.5	1.0 – 2.6	0.05 – 0.15	0.1 – 0.2
former Soviet Union		flot.	46 – 48	39 – 41	0.9 – 1.0	0.1 – 0.2	0.1 – 0.2	0.05 – 0.1

* Flot. = flotation pyrite; fine = fine pyrite.

178 000 tonnes S, Spain 145 000 tonnes and Norway 82 000 tonnes S in the form of pyrites [16]. From 1965 until about 1975, world pyrites extraction remained almost constant at an annual rate of about $10-12\times10^6$ t sulfur content. But, as industrial sulfur demand increased substantially during that period, pyrite was in fact declining consistently in its relative importance among the sulfur raw materials. Processing of iron sulfide containing gold and silver ores has increased in importance, resulting in additional involuntarily produced SO_2.

Nonferrous Metal Ores. Nonferrous metal sulfide ores are another important raw material source for sulfur dioxide production. In the pyrometallurgical processing of sulfide ores for extraction of copper, nickel, zinc, and lead waste gases containing sulfur dioxide are inevitably produced. In most countries it is no longer permissible, for environmental reasons, to discharge these gases directly into the atmosphere because of the large plant capacities used nowadays. The development of the various smelting processes is, therefore, strongly influenced by the dual needs to capture the released sulfur dioxide as completely as possible and to generate waste gases with high sulfur dioxide concentrations so that they can be processed as economically as possible [17].

The rate at which sulfur dioxide is produced in the processing of nonferrous metal sulfide ores can be estimated approximately on the basis of metal smelter production statistics, assuming characteristic mean sulfur/metal weight ratios in the ore concentrates of each metal. For 1992 a figure of ca. 20×10^6 t of SO_2 can be given. A comparison of smelter productions for the most important nonferrous metals in 1979 [18], which can be assumed to be based mainly on sulfide ores or concentrates enriched by flotation, is given in [18].

Other Materials. Sulfur dioxide is normally only recovered for environmental reasons from waste sulfuric acid and sulfates, from hydrogen sulfide-containing industrial waste gases and from power station flue gases. Calcium sulfate, in the form of natural gypsum or anhydrite, was formerly used in a few small plants as raw material for sulfur dioxide and sulfuric acid production, but on account of their high energy consumption such plants have either been shut down or modified to process waste gypsum (e.g., from phosphoric acid manufacture).

All sulfur raw materials, except for elemental sulfur and pyrites, are summarized in the statistics under the general item of "sulfur-in-other-forms". The worldwide production of "sulfur-in-other-forms" (with nonferrous metal ores playing the dominant role) amounted in 1982 to about 8.2 million tonnes, of which approximately 7.5 million tonnes were consumed in sulfuric acid production, corresponding to 18 % of the total amount of sulfur used in the production of sulfur dioxide in sulfuric acid manufacture. The most important sources of sulfur-in-other-forms accounted for the following quantities in 1982 (in 10^3 t of sulfur) [16]:

former Soviet Union	1700
United States	1378
Japan	1371
Canada	430
Federal Republic of Germany	408
Finland	221
Australia	160
Italy	152

4. Production from Elemental Sulfur

To burn sulfur in accordance with Equation (2), equimolar quantities of sulfur and oxygen are required

$$S\,(\text{rhomb}) + O_2\,(g) \longrightarrow SO_2\,(g)\quad \Delta H = -297\ \text{kJ/mol} \qquad (2)$$

With air, which contains 20.95 vol % O_2, sulfur dioxide concentrations of ca. 20.5 vol % can in theory be obtained. But in practice, to ensure complete combustion of the sulfur, air is generally supplied in excess, as the problems which might result from unburnt sulfur condensing and solidifying in colder parts of the plant are not worth risking.

As a prerequisite for complete sulfur combustion and for maintaining constant sulfur dioxide concentrations under industrial conditions, it is essential for the sulfur to be finely and evenly dispersed and intimately mixed with the combustion air during burning. The general practice to use liquid sulfur at 140 – 150 °C, at which temperature its viscosity is low enough to enable it to be sprayed through nozzles. The high

surface area of the resulting fine droplets aids rapid evaporation (the normal boiling point of sulfur is ca. 440 °C).

During the combustion process, it is S_2 molecules in the vapor phase that are actually oxidized. Evaporation of liquid sulfur initially produces mainly S_8 molecules, which are only decomposed to S_2 to an appreciable extent above 600 °C [19]. More than 60 % of the heat of reaction (about 9280 kJ/kg S) liberated in the combustion of sulfur to sulfur dioxide is theoretically required for preheating the air and sulfur and for evaporation and decomposition of the sulfur at 600 °C. Therefore, favorable combustion is dependent not only on maximum mass transfer velocity but also, very markedly, on intensive heat transfer in the initial phase.

When solid sulfur is used, the lowest ignition temperature in air under atmospheric pressure is ca. 246 °C. However, sulfur dust clouds may ignite spontaneously at temperatures as low as 190 °C, a phenomenon which has to be borne in mind when handling and storing solid sulfur [20].

Sulfur combustion in industrial oxygen or in oxygen-enriched air is generally uneconomic on account of the high cost of oxygen, and it is not currently used on an industrial scale.

The practicable upper concentration limit for the industrial production of sulfur dioxide gases by burning liquid sulfur with air is at about 18 vol % SO_2. Normally, however, the gases from sulfur combustion furnaces associated with sulfuric acid plants contain 10 – 12 vol % SO_2, corresponding to an O_2/SO_2 ratio of ca. 1:1, as required in double absorption plants (see → Sulfuric Acid, Chap. 4.1.4.1.).

Normally the combustion air used in the industrial production of sulfuric acid plant feed gases from elemental sulfur is dried with concentrated sulfuric acid. It is then compressed by a blower to1.3 – 1.5 bar to provide the driving force needed to overcome pressure drop within the rest of the plant. Drying and compression raise the temperature of the air to about 60 – 100 °C.

Theoretically, the combustion temperature is ca. 1000 °C when liquid sulfur is burnt at an initial temperature of 140 °C in dried air originally at 60 – 80 °C to produce a gas containing 10 vol % SO_2. As can be seen from Figure 4, the combustion temperature rises in direct linear proportion to the sulfur dioxide concentration. It reaches about 1600 °C when the combustion gas sulfur dioxide concentration is 18 vol %.

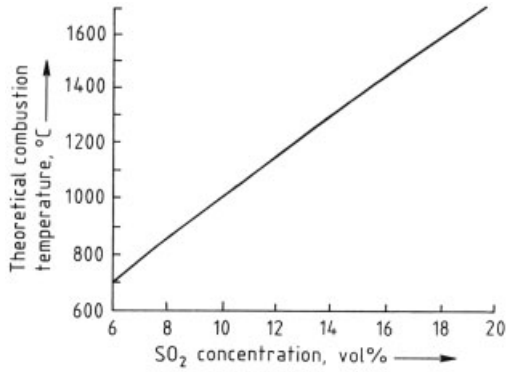

Figure 4. Theoretical sulfur combustion temperature as a function of combustion gas SO_2 concentration (air temperature = 80 °C, sulfur temperature = 140 °C)

4.1. Sulfur Burning Processes

In industry nowadays sulfur combustion furnaces with atomizing burners for liquid sulfur are used almost exclusively. Figure 5 shows the layout of a Lurgi sulfur combustion furnace. The furnace comprises a cylindrical steel shell lined with several layers of insulating and refractory bricks. It is normally arranged horizontally, but when space is limited a vertical version is available. The sulfur burner is mounted at one end of the combustion chamber and a waste heat boiler is connected directly to the gas outlet at the opposite end. The specific furnace size is very much dependent on the type and efficiency of the burner used.

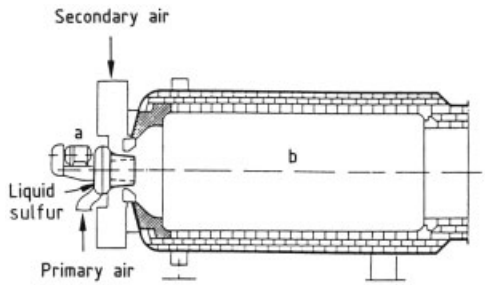

Figure 5. Lurgi sulfur furnace
a) Rotary atomizer; b) Sulfur furnace

Burner Types. There are various ways in which liquid sulfur can be atomized to produce the fine spray needed for the most efficient possible combustion. The main commercial burner types are the pressure-nozzle burner, in which sulfur is injected under pressure through a spray nozzle into the furnace chamber; the two- component burner, in which the sulfur is released through a nozzle into a turbulent jet of air; and the rotary burner, in which the sulfur is distributed centrifugally from the edge of a spinning cup into a stream of air [21, 22].

With the simple *pressure-nozzle burner*, the liquid sulfur is atomized with a single spray head. The relatively high pressure of 8 – 15 bar needed to drive this type of burner is produced by a sulfur pump of special design [23]. A steam-heated double shell around the nozzle pipe serves the purpose of keeping the sulfur molten when the burner is not operating and preventing overheating when it is. The main disadvantage of this type of atomizer is its relatively low flexibility: major changes in the throughput can only be accommodated by changing the nozzle. The smaller-diameter nozzles are also prone to clogging.

The maximum sulfur throughput of a pressure-nozzle burner is ca. 72 t/d. In the relatively large plants commonly used nowadays, it is, therefore, necessary to install several nozzle systems with separate measuring and control instruments. This also increases the flexibility of the system.

Two-component burners, in which the sulfur is chiefly atomized by the combustion air, can be used for extremely low sulfur throughputs of < 7 t/d, for which pressure-nozzle burners are inherently unsuitable. The combustion air flows at high velocity from an annular space surrounding the sulfur nozzle. The sulfur, supplied to the nozzle through steam-jacketed pipes, is entrained in the air stream and finely distributed by turbulence in the combustion air induced by deflector plates. The necessary sulfur pressure is only ca. 2 bar. Modern two- component burners are designed for sulfur throughputs of up to ca. 170 t/d. The turn-down limit is usually around 15 % of capacity, although a compact design intended for small-scale applications such as pulp mills is reported to be able to run at as little as 5 % of its nominal capacity of 5 t/d. This is because a variable-speed sulfur pump is used instead of the more normal check valve arrangement to control the flow of sulfur [24].

The *rotary burner*, represented by the Lurgi-developed Luro Burner, is nowadays the preferred type for large plants. The design of such a burner is shown schematically in Figure 6. The sulfur passes at virtually atmospheric pressure into an open cup rotating at 4000 – 6000 rpm. The sulfur film on the inner wall is dispersed from the edge by centrifugal force and is further atomized by entrainment in the so- called primary air, which emerges at high velocity through the narrow annular space between the rotating cup and the atomizer hood. "Secondary air," constituting the majority of the combustion air, enters the furnace through an annular space between the atomizer hood and the burner casing. A system of adjustable air guide vanes in this space imparts a tangential component to the flow of secondary air as it enters the conical combustion space between the burner and the main furnace chamber. The resulting spiral path of the gases through the furnace not only effectively keeps the flame central but at the same time prevents unvaporized sulfur droplets from impinging on the furnace wall.

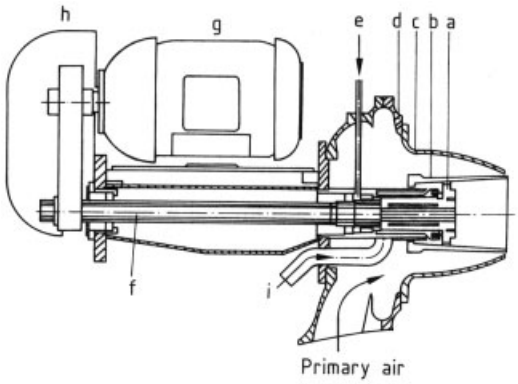

Figure 6. LURO rotary sulfur burner
a) Sulfur distributor; b) Atomizing cup; c) Atomizer hood; d) Burner housing; e) Seal air; f) Burner shaft; g) Electric motor; h) Belt transmission; i) Sulfur inlet

A highly advantageous characteristic of rotary burners is that the throughput can be varied continuously from the rated capacity down to 20 % capacity or less without appreciable change in atomizing intensity. Rotary burners can be designed for sulfur throughputs ranging from 5 to 400 t/d. The highly turbulent atom-

izing effect ensures swift and safe combustion of the sulfur, which makes it possible to design for extremely high specific furnace loads (ca. 4 – 8×10^6 kJ m^{-3} h^{-1}) and small specific furnace volumes.

4.2. Heat Recovery

The heat released in the combustion furnace is utilized in a waste heat boiler immediately downstream, generally producing high-pressure steam. In the case of a sulfuric acid plant, gases of 10 – 12 vol % SO_2 leave the furnace at ca. 1100 °C and are cooled to the converter inlet temperature of ca. 420 – 440 °C. The waste heat boiler is coupled with the waste heat system of the contact plant and, in conventional double-absorption contact plants, it accounts for ca. 58 – 60 % of the total energy recovered as high-pressure steam. On the whole, ca. 3.8 – 4.0 t of steam at 40 bar and 400 °C can be produced per tonne of liquid sulfur. If the sulfur is supplied in solid form, about 0.5 t of saturated steam at 5 bar is required per tonne of sulfur for melting, heating the sulfur filters, and tracing the sulfur piping.

4.3. Impurities and their Removal

The 10 – 12 vol % SO_2 combustion gases produced in a sulfuric acid plant contain minor amounts of sulfur trioxide (3 – 6 g/m^3) [25] and up to 20 ppm nitrogen oxides. Other impurities mainly depend on the quality of the sulfur used.

Much of the Frasch and recovered sulfur commercially available nowadays, which is supplied in Western Europe and in the United States mainly in liquid form, is adequate for direct burning without requiring additional cleaning, but various impurities do occur in sulfur, particularly where it is handled and transported in solid form.

Bitumen and hydrocarbons, which occur especially in the so-called dark sulfur grades, burn to carbon dioxide and water. The correspondingly higher moisture content of the sulfur dioxide-containing contact gases can lead to sulfuric acid mists, which are difficult to absorb. In sulfuric acid plants running on dark sulfur, additional equipment, such as candle filters downstream of the sulfur trioxide absorbers, may be necessary to avoid corrosion and emission problems.

Any *ash* in the sulfur may lead to deposits in the combustion furnace and waste heat boiler. However, the major part is entrained as fine dust in the sulfur dioxide-containing gases and passes into subsequent parts of the plant. In a sulfuric acid plant, it is mostly caught in the first catalyst bed. This leads to a progressive increase in the gas pressure drop and may also cause a loss in the sulfur dioxide conversion efficiency. Regular screening of the catalyst may be necessary. Hot gas filters were formerly placed in the gas line upstream of the converter to separate the dust. But they are hardly used any more because they add to the pressure drop and are only partially effective. Sedimentation tanks or settling pits cannot be relied on alone to remove all the solid components from the molten sulfur.

Filtration of the liquid sulfur [26] is the most effective means of separating higher ash contents, which may occur especially in solid sulfur. The sulfur is melted and heated by using low-pressure steam of about 5 bar to a temperature of ca. 140 – 150 °C, at which its viscosity is at a minimum. It is then pumped through vertical or horizontal filters of up to ca. 150 m^2 filtering area. Normally, these are of the pressure leaf type, comprising a battery of stainless steel wire gauze elements suspended in a containing vessel. Each element consists of two fine-mesh filter gauzes suspended in a ring frame and supported and held apart by a third, coarse mesh layer. The intermediate space between the two fine gauzes of each element is connected to the outlet of the filter. In operation, the outside of the filter elements is first coated with a filter aid such as kieselguhr by passing it as a suspension into the filter vessel and recirculating the effluent from the outlet back into the filter feed tank or settling pit. Once prepared in this way, the filter is ready for service. Periodically, as it becomes clogged with ash-forming material from the sulfur, the filter is cleaned and recoated.

Apart from the additional operations and equipment required, sulfur filtration poses a problem to the sulfur user of what to do with the filter cake, which contains ca. 70 % sulfur. As a result, the responsibility for sulfur filtration is increasingly being shifted to the sulfur producer.

4.4. Two-Stage Sulfur Combustion

For producing gases of high sulfur dioxide concentration and extremely low nitrogen oxides (NO_x) content, Lurgi has developed a double-stage process with substoichiometric sulfur combustion [27]. As explained earlier, the combustion temperature increases in proportion to the sulfur dioxide concentration of the combustion gas (see Fig. 4). Above the usual combustion temperature of 1000 °C, formation of nitrogen oxides increases in proportion to the temperature, even though the free oxygen content remaining in the gas is lower when the sulfur dioxide content is higher if the combustion air is not enriched. Only above a concentration of ca. 18 vol % SO_2 is the free oxygen content low enough for thermal NO_x formation to decline drastically. Nitrogen oxides formation is also the factor limiting the extent to which it is possible to preheat the combustion air in ordinary sulfur burning systems, as this raises the combustion temperature.

Nevertheless, formation of nitrogen oxides can be avoided, even at a combustion temperature of 1750 °C, if the combustion of sulfur is divided into two stages. In the first, the sulfur is burnt with a stoichiometric deficiency of oxygen. It is particularly important in that stage to ensure that the gases are extremely well mixed so that all the oxygen is consumed, and nitrogen oxides cannot form. The combustion gas, containing ca. 20 vol % SO_2 and some unconverted sulfur vapor, is cooled in a waste heat boiler to ca. 620 – 650 °C and then passed into the second stage, an afterburning chamber (see Figure 7). There, enough dried air is added to convert the remaining sulfur vapor completely to sulfur dioxide, leading to a final gas concentration of ca. 18 vol % SO_2. On account of the low concentration of sulfur vapor in the second stage, the combustion temperature is too low for nitrogen oxides to form. The gas mixture leaves the afterburning chamber at a temperature not exceeding 700 °C and is cooled in another waste heat boiler (usually designed as a steam superheater) before being passed on to, for example, a sulfur dioxide liquefying or a bisulfite plant. This type of system has been proved in years of satisfactory service in large-scale plants.

The concentration of nitrogen oxides observed in the gases containing 18 vol % SO_2 produced in this double-stage sulfur burning system are actually lower than those of the 10 % SO_2 gases originating from single-stage sulfur burning plants.

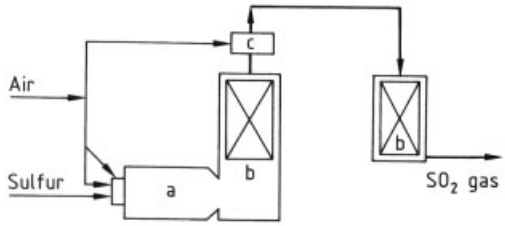

Figure 7. Lurgi substoichiometric sulfur combustion process
a) Sulfur furnace; b) Waste heat boiler; c) Afterburner

5. Production by Pyrite Roasting

5.1. Importance of Impurities

In the production of sulfur dioxide-containing gases by roasting of pyrites, the constituents occurring besides sulfur and iron in the flotation concentrates or fines exert a substantial influence on the selection of the roasting process to be applied and its economics. Pyrite contains ca. 30 – 50 % sulfur, 26 – 46 % iron and up to 2.7 % copper, 3 % zinc, 1.4 % lead, up to 10 % arsenic, and a number of other metals in small quantities (see Table 1). Some of these impurities are entrained with the sulfur dioxide in the roaster gases, and to remove them efficient gas cleaning is necessary.

Impurities are also of prime concern in disposal or utilization of the iron oxide residue of pyrite roasting, the so-called cinder. Each tonne of pyrite concentrate yields about 0.7 t of cinder. High purity is a prerequisite where the cinder is to be used for steelmaking. Cinders within the following specification are generally considered acceptable [28]: Fe 65 %, Cu 0.1 %, Zn 0.1 %, Pb 0.02 %, As 0.06 %, S 0.5 %.

If a pyrite concentrate contains arsenic in larger quantities, special roasting methods have to be used so as to obtain sufficiently low residual arsenic values in the cinder (see Section 5.4).

Nonferrous metals such as Cu and Zn, which are still present in flotation pyrites, are again found in the cinder. At high nonferrous metal

contents, the cinder can be treated by chlorination and leaching.

5.2. Thermodynamics and Kinetics

Four main reaction steps (Eqs. 3 – 6) make up the overall pyrite roasting reaction (represented by Eq. 7). The specified enthalpy values [7] refer to 298 K:

$$2\,FeS_2 \longrightarrow 2\,FeS + S_2(g) \quad \Delta H = +\,293\ \text{kJ/mol} \tag{3}$$

$$S_2(g) + 2\,O_2 \longrightarrow 2\,SO_2 \quad \Delta H = -723\ \text{kJ/mol} \tag{4}$$

$$2\,FeS + 3\,O_2 \longrightarrow 2\,FeO + 2\,SO_2 \quad \Delta H = -\,948\ \text{kJ/mol} \tag{5}$$

$$2\,FeO + 0.5\,O_2 \longrightarrow Fe_2O_3 \quad \Delta H = -\,282\ \text{kJ/mol} \tag{6}$$

$$2\,FeS_2 + 5.5\,O_2 \longrightarrow Fe_2O_3 + 4\,SO_2 \quad \Delta H = -\,1660\ \text{kJ/mol} \tag{7}$$

The endothermic decomposition of pyrites gives FeS or magnetopyrite, liberating sulfur vapor (Eq. 3). If oxygen is available in adequate quantities, the sulfur is immediately oxidized to sulfur dioxide (Eq. 4) and simultaneously the sulfur content of the FeS is converted to sulfur dioxide as well, in accordance with the equally exothermic reaction (Eq. 5). At the temperatures prevailing in the roasting process, the iron(II) oxide (FeO) formed as an intermediate phase is oxidized to magnetite (Fe_3O_4) or hematite (Fe_2O_3), depending on the oxygen ratio (Eq. 6). The overall pyrite roasting reaction (Eq. 7) is strongly exothermic, yielding ca. 13 000 kJ/kg pyrite sulfur. (The corresponding energy yield of burning elemental sulfur is only ca. 9300 kJ/kg of sulfur). The roasting rate depends on the temperature, grain size, and mineralogical composition of the pyrites as well as the oxygen partial pressure.

Iron(II) oxide and sulfide form an eutectic mixture containing 56.2 % FeS and 43.8 % FeO, which melts at 940 °C. It is not possible to increase the reaction velocity by raising the temperature above this value because any molten material deposits on the surface of the pyrite particles. This renders unreacted material inaccessible to oxygen and thus inhibits further reaction, particularly with larger particles.

The ignition temperature of pyrite lies between 350 °C and 550 °C. The finer the grain size and the higher the nonferrous metal content are, the nearer the ignition temperature is to the lower limit. A temperature of at least 800 °C is required for complete reaction; the optimum range is 850 °C to 940 °C, depending on the grain size. A few pyrites decrepitate during roasting. The fine particles thus formed influence the roasting velocity and increase dust formation.

The residual sulfur content of the cinder depends not only on the amount of sulfide remaining unconverted but also on sulfate sulfur. The latter forms chiefly when the pyrite contains nonferrous metal impurities such as copper, zinc, lead, cadmium, magnesium, and calcium, which remain in the cinder as sulfates.

The maximum attainable sulfur dioxide content of the roaster gas depends on the pyrite composition as well as the atmosphere in which it is burnt. In air it is ca. 16 vol % in the most favorable case, although with the slight excess of air usually used values of 13 – 14 vol % are common. In oxygen or oxygen-enriched air, higher sulfur dioxide concentrations can be attained. In spite of the thermal and process advantages that this would have with respect to sulfur dioxide and sulfuric acid production, oxygen is rarely used for pyrite roasting because of the high costs.

5.3. Roasting Processes

Various types of roasting equipment have been used in the past for pyrite roasting, including shaft furnaces, multiple hearth roasters, rotary kilns, and dust roasters [29, 30]. Nowadays fluid-bed roasters are preferred, as they are much superior to the other types in terms of process technology, throughput rates, and economy. As a further step, circulating fluid bed technology was introduced for roasting special ores that containing economically viable quantities of gold.

Basically a fluid-bed roaster is in the form of two upright cylinders, the upper being of a larger diameter than the lower, to which it is joined by a frustoconical section. The steel shell is lined with insulating and refractory bricks. At the lower end is a horizontal bar grate or a nozzle grate, below which is a conical wind box, at the apex of which the roasting air is introduced. In this way the roasting air is uniformly distributed over the entire cross section of the lower portion of the roaster. Above the grate, a bed of cinder

is maintained in a fluidized state by the roasting air. The pyrite is fed into the roaster directly above or in the uppermost zone of the fluid bed. The depth of the fluid bed is kept constant by an overflow. In the wider upper section of the roaster, the velocity of the gas leaving the fluid bed is reduced. This serves to stabilize the top zone of the fluid bed and reduces entrainment of solids in the gas flow. This section also acts as an afterburning chamber. The roaster gases pass from the top of the roaster to a waste heat boiler directly downstream.

Fluid-bed technology for pyrite roasting was first introduced at the beginning of the 1950s almost simultaneously by Dorr Oliver in the United States and BASF in the Federal Republic of Germany.

Dorr Oliver FluoSolids Reactor. In its original concept, the FluoSolids reactor was designed on the wet feeding principle [31,32]. Finely ground ore or flotation pyrites are mixed with water to form a viscous pyrite sludge with a 70 – 80 % solids content. This is forced continuously into the roaster by means of a pump. The height of the cinder bed is up to 1 m in the unfluidized state and 1.6 m when fluidized at a gas velocity in the bed of 0.2 – 0.6 m/s. The specific daily throughput is around 4 – 5 t/m^2 of roaster area. The bed temperature is adjusted to 600 – 900 °C by adiabatic evaporation of the water introduced with the pyrite and (where appropriate) of additional water injected separately. A later version of the FluoSolids roaster was designed for dry feeding, using indirect cooling and heat recovery [33].

The BASF – Lurgi fluid bed furnace roaster was developed [34–36] for dry pyrite feeding and designed to recover the maximum amount of heat for steam production. In this roaster the temperature of the fluid bed is kept constant by indirect cooling; the surplus heat of reaction is removed in the fluid bed by immersed cooling elements which form part of a waste heat system for the production of high-pressure steam. Figure 8 shows a typical design of a modern fluid bed furnace with waste heat system according to the Lurgi principle.

The grain size of the pyrites to be processed may be up to 10 mm, depending on the stability of the particle structure during the roasting process. Usually the particles have diameters of up to 6 mm. The coarser the grain is, the higher the gas velocity must be to maintain the bed in a fluidized state. Even flotation pyrites with grain sizes < 0.2 mm can be processed in the fluid bed roaster, but in that case the moisture content is important. Although the pyrites must be dry enough to be fed in the solid state, enough moisture must be present to ensure that the particles are sufficiently agglomerated to maintain a stable fluid bed. Otherwise the particles will be too light and will be largely blown out of the reaction in the gas flow.

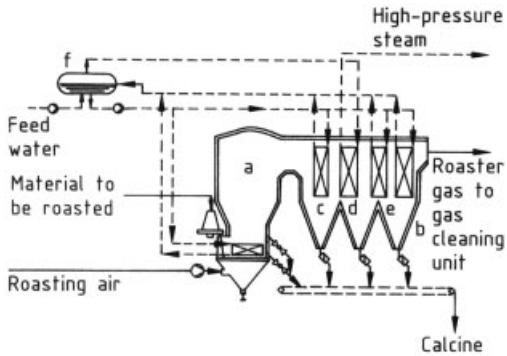

Figure 8. Fluid bed roaster with waste heat system
a) Fluid bed furnace; b) Waste heat boiler; c) Pre-evaporator; d) Superheater; e) Evaporator; f) Steam drum

To ensure uniform, controlled feeding of the pyrites, table feeders or weight- controlled variable-speed conveyors are used. The gas velocity in the fluid bed is ca. 0.8 – 2.5 m/s at 1 bar and 850 °C. In the top part of the roaster the velocity is ca. 0.5 – 1.8 m/s. In the unfluidized state, the bed has a rest depth of 700 – 800 mm.

In contrast with all the earlier roasting systems, the material distribution in a fluid bed system is very rapid, and both the composition of the bed and its temperature are very uniform. Therefore, it is possible to select and maintain the roasting process conditions within a narrow optimum range. On account of the good heat transfer characteristics (800 –1200 kJ m^{-2} h^{-1} K^{-1}) the temperature can be controlled satisfactorily by indirect cooling using immersed, water- cooled elements with a relatively small heat exchange surface area.

The bed temperature is normally maintained in the range 650 – 900 °C. In the gas space above the fluid bed, the temperature may be some-

what higher (up to 950 °C) on account of the afterburning effect when flotation pyrites are being processed. Above the fluid bed the pressure is kept slightly below atmospheric pressure to avoid leakage of sulfur dioxide to the outside. The roaster gases are drawn off from the furnace by the main gas fan of the subsequent contact plant or by an intermediate gas fan in the gas cleaning system.

Abrasion under the turbulent bed conditions and decrepitation of the pyrite particles create a substantial amount of fines, so that a sizeable amount of the solids leave the bed in the roaster gas in the form of dust. This dust is precipitated in the waste heat boiler and also in subsequent cyclones and hot gas electrostatic precipitators. When fine pyrites are roasted, the portion of cinder leaving the roaster as entrained dust may be around 50 % and in the case of flotation pyrite roasting it may even be as high as 90 % or more.

Being predominantly iron oxide, the dust has catalytic activity for the oxidation of sulfur dioxide to sulfur trioxide. This reaction is highly undesirable at this stage because any sulfur trioxide formed reacts with moisture in the gas to form sulfuric acid, which condenses in cool parts of the plant and causes corrosion. In the roaster itself the oxidation of sulfur dioxide is inhibited by the high prevailing temperature and by controlling the air supply to minimize the residual oxygen content in the roaster gas. Having a low oxygen content is of even greater importance in the waste heat boiler, where the gas is cooled to temperatures at which sulfur trioxide formation is thermodynamically more favored. It is also of importance to cool the gases here as quickly as possible to temperatures at which the reaction is kinetically inhibited. Nevertheless, some sulfur trioxide is inevitably formed, and since neither the pyrites nor the roasting air are dried, the roaster gas temperature at the outlet of the hot gas electrostatic precipitator must be maintained above the dewpoint of sulfuric acid. This limits the amount of heat that can be recovered from the gases in the waste heat boiler. Generally, the roaster gases are cooled to about 350 °C before they enter the dust separation cyclone.

Figure 9 shows a typical energy flow diagram for the roasting of flotation pyrites containing 47 – 48 % sulfur in a Lurgi fluid bed roaster system. Approximately 70 % of the total energy is recovered from the fluid bed and roaster gas in the form of high-pressure steam. This corresponds to the production of about 1.5 t of steam of 40 bar and 400 °C per tonne of 48 % S pyrite from feed water at a temperature of about 104 °C. The remaining 30 % of the total energy is represented by the heat content of the roaster gases at an exit temperature of 350 °C, the heat content of the cinder, and radiation losses. The power consumption amounts to a 25 – 30 kWh/t pyrites. Most of it is used in the roasting air fan. The sulfide sulfur content of the cinder from pyrites originally containing 47 – 48 % sulfur is about 0.5 %. The sulfur dioxide content of the roaster gases at the furnace outlet is ca. 14 – 14.5 vol %. The use of an automatic control system to maintain the oxygen content of the roaster gas at the furnace outlet at a minimum substantially reduces sulfur trioxide formation in the waste heat boiler and with it the rate of scrubber acid accumulation in the wet gas cleaning system.

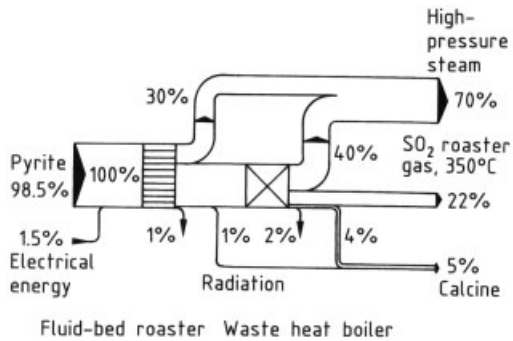

Figure 9. Energy flow diagram for fluid bed pyrite roasting

The specific capacity of Lurgi fluid bed roasters corresponds to a daily throughput of about 20 t/m^2 of roaster area. Today, fluid bed roasters are designed for pyrites throughputs of up to about 1000 t/d. Their turndown limit is around 70 – 80 %. For large plant capacities, in most cases two furnaces are connected in parallel to provide better flexibility.

During start-up, the fluid bed made up of cinder is preheated with oil burners. As soon as the bed temperature reaches ca. 600 °C, pyrites can be fed in and the oil burners are turned off.

On account of the high thermal capacity of the cinder bed, the roaster can be restarted with-

out requiring additional heating after standstills of up to about 24 h.

Circulating Fluid Bed Roasting [37]. A further step towards more efficient roasting – especially of gold-containing ores – was the introduction of the circulating fluid bed process.

The circulating fluid bed (CFB) technology is known from the calcination of hydrated alumina or clay and combustion of coal, and was adapted for roasting of gold-bearing minerals. After extensive pilot plant tests the first industrial scale plant was built by Lurgi and started up in 1989 in Australia.

The aim of roasting gold-containing minerals is to obtain a maximum gold recovery from the calcine in subsequent leaching steps. Such conditions can only be achieved within a narrow range of roasting temperatures and oxygen partial pressures in the roaster gas atmosphere. The CFB process is capable of providing such conditions.

Contrary to a "normal" stationary fluid bed, circulating fluid bed roasters are operated at higher gas velocities in the range of 3 – 6 m/s. Because of the velocity the fine solids (calcine) are entrained with the gas and in part recycled to the roaster via a cyclone.

The main features of the CFB process are:

1) Treatment of a wide range of minerals
2) High heat and mass transfer rates
3) Uniform temperature profile over the whole roaster/cyclone system
4) Accurate temperature control by cooling surfaces, direct water injection, or direct fuel injection
5) Stable oxygen partial pressure all over the system
6) Turndown ratio of 50 %

Ores with sufficient sulfur and iron content can be roasted autothermally; ores with low sulfur and iron contents can be treated with additional heating; and even the most sensitive refractory gold ores can be roasted successfully in the CFB. Air or oxygen-enriched air can be employed. The high heat and mass transfer rates guarantee almost complete combustion of sulfur and carbon. The uniformity of the temperature profile avoids the partial overheating of particles known from stationary fluid beds. Overheating of particles may cause substantial drop in gold recovery. The temperature is uniform in the whole system due to the high amount of circulated material and the effect of cooling (or heating) is transferred to the total inventory of the system. Because of the high velocities and mass transfer rates an exact and stable control of the oxygen partial pressure is possible to obtain optimum roasting results. A turndown ratio of up to 50 % offers a high operational flexibility compared to normal fluid bed roasting.

Key components of CFB plants are the roaster and the recycling cyclone. The roaster consists of a cylindrical shell with insulating and refractory bricks. At the lower end a nozzle grate is installed through which the roasting air is introduced. At the upper end the recycling cyclone is connected to the roaster. The ore is fed to the CFB roaster and roasted with air or oxygen-enriched air. A portion of the hot calcine which is separated in the cyclone is recycled to the roaster, a second portion is transferred to a fluid bed cooler where part of the heat contained in the calcine is removed. The hot gas leaving the cyclone is cooled and treated in downstream plant elements.

The associated equipment must be selected according to the specific requirements of each plant; some of such elements are listed below:

1) Dry feed or slurry feed
2) Preheating and drying of feed with recovered heat from the calcine
3) Preheating of roasting air with recovered heat from the calcine
4) Direct injection of additional fuel (if required)
5) Cooling of the reaction material by integrated cooling elements, direct water injection or recirculation of cooled calcine
6) Cooling of the calcine in a fluid bed cooler, or direct quenching
7) Heat recovery from roaster gas in a waste heat boiler or gas – gas heat exchangers
8) Further treatment of the off-gas (bag house, electrostatic precipitator, acid plant, . . .)

A flowsheet showing the key equipment and some alternatives for associated equipment is given in Figure 10 (see next page).

Roasting in a Cyclone Furnace. Pilot-plant trials have been carried out in Poland on a

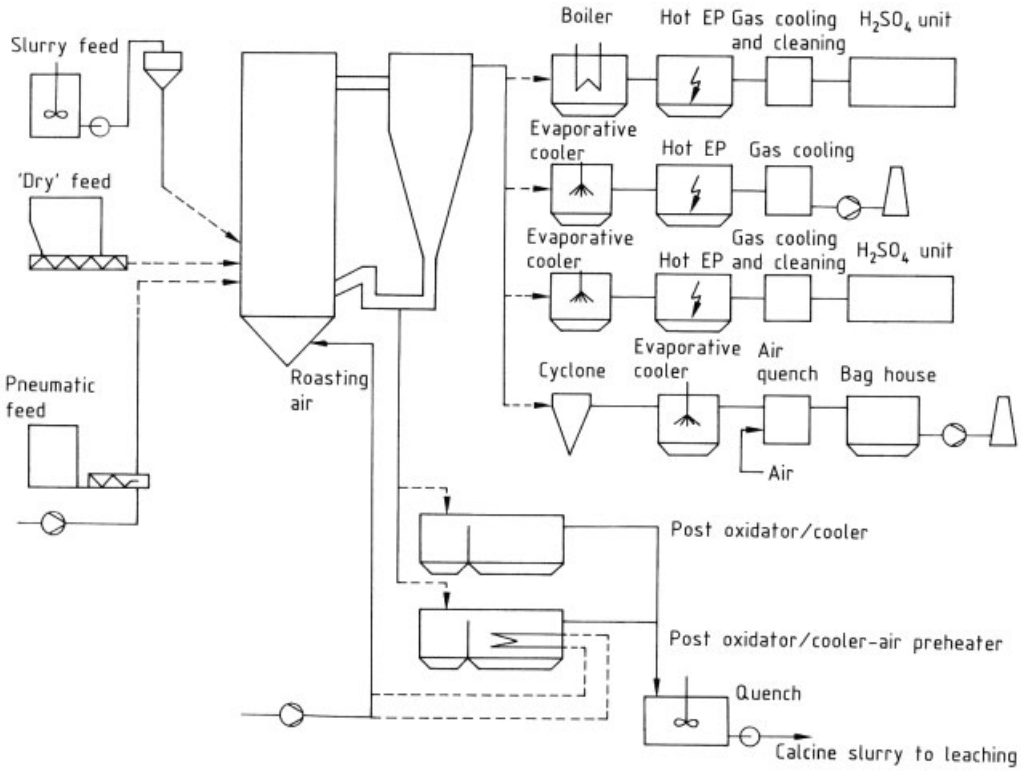

Figure 10. CFB-roasting of gold-bearing ores
EP = electrostatic precipitator

method of roasting pyrites in a cyclone furnace, which is claimed to facilitate the subsequent recovery of nonferrous metals [38]. Flotation pyrites, preheated air, and (as necessary) fuel are introduced into the upper section of a cyclone. The operating temperature is such that the burnt particles of pyrite cinder melt, coalesce, and run down the walls of the cyclone. The melt is drained into a quench bath, where it forms granules. The gas leaving the cyclone contains entrained iron oxide droplets and part of the original nonferrous metals content of the pyrites, especially zinc and lead, both as particles and vapors of oxides and metals. It is cooled in a fluid bed cooler with immersed cooling coils, which recover the heat partly as high-pressure steam and partly by preheating the furnace air. The entrained material and vapors of nonferrous metals are largely arrested by agglomeration in the fluid bed. The material overflowing from this bed is stated to contain a substantially higher concentration of nonferrous metal values than the cinder from other types of roasting furnace.

5.4. Roasting of Arsenical Pyrite

Arsenic occurs in pyrites in the form of arsenical pyrite (FeAsS; $FeAs_2 \cdot FeS_2$). Under the usual roasting conditions, with a slight excess of oxygen in relation to the stoichiometric requirement for converting the iron content of the pyrite to hematite, the majority of the arsenic is oxidized to As(V) and will react with the hematite to form stable iron arsenate ($FeAsO_4$). In this way about 70 – 80 % of the arsenic content in the pyrite feed becomes bound to the cinder. The remaining arsenic, which is oxidized to As(III) as shown in Equation (8), is volatilized and entrained by the roaster gas and must be separated in the wet gas cleaning system

$$2\,FeAsS + 5\,O_2 \longrightarrow Fe_2O_3 + 2\,SO_2 + As_2O_3 \qquad (8)$$

If the cinder is to be used in steelmaking, itmust not contain too much arsenic. In such cases high-arsenic pyrites can be processed in a two-stage

fluid-bed roasting system. In the first stage, operating under oxygen-deficient conditions, neither hematite nor iron arsenate is formed. The arsenic is thus volatilized almost completely and discharged with the roaster gases. In the second fluid bed stage, the separated cinder is oxidized to hematite. This principle is applied on an industrial scale in the Boliden roasting process and the BASF two-stage process.

Both the Boliden and BASF processes are capable of diminishing the residual arsenic content of the cinder to less than 0.05 %.

Boliden Process. The Boliden process [28] was specifically developed for processing arsenical flotation pyrites. In the fluid bed roaster, the roasting air/pyrites ratio is adjusted so that magnetite forms at ca. 850 °C. Most of this is blown out of the furnace in the roaster gas. It is separated at an efficiency of ca. 90 % in a hot-gas cyclone immediately following the furnace. The gas leaving the system contains, besides residual dust, the arsenic and ca. 10 g/m^3 of sulfur vapor. The sulfur vapor is burnt in a special combustion chamber at the inlet of the waste heat boiler with a controlled amount of additional air. More air is added at the waste heat boiler outlet to complete the combustion of residual traces of sulfur before the gases pass on to the hot gas electrostatic precipitator. The cinder separated from the gas in the hot gas cyclone is treated in a second fluid bed roaster with a controlled volume of air. In the process, the residual sulfur level is diminished substantially, and the magnetite is partly converted to hematite (Eq. 9)

$$2\,Fe_3O_4 + 0.5\,O_2 \longrightarrow 3\,Fe_2O_3 \quad \Delta H = -57.3\,\text{kJ/mol} \qquad (9)$$

As in the first roaster, the surplus heat of reaction is recovered by generating steam in cooling coils immersed in the fluid bed. The discharged cinder is cooled to ca. 200 °C in a separate fluid bed cooler, which is connected into the steam system as a preheater. Up to ca. 1.6 t of high-pressure steam at approximately 425 °C are produced per tonne of pyrites.

The arsenic is removed from the dedusted roaster gases in the wet gas cleaning system (see Section 6.2). Plants built under license by Lurgi processing up to 500 t/d pyrites have been operated successfully in Sweden, Spain, and Japan.

BASF Process. The BASF two-stage process [39] is especially suited for processing relatively coarse pyrites. Partial roasting in a first fluid bed unit at 900 °C under oxygen-deficient conditions removes only the loosely bound polysulfide sulfur of the pyrite, giving FeS.

$$FeS_2 + O_2 \longrightarrow FeS + SO_2 \quad \Delta H = -215\,\text{kJ/mol} \qquad (10)$$

Arsenic, antimony, and lead are carried out of the system with the roaster gases, mainly as volatile sulfides or oxides. Coarse ferrous sulfide cinder withdrawn directly from the bed overflow and hot dust separated in the high-efficiency cyclone are passed on to a second fluid bed unit. The iron(II) sulfide, which is virtually free from arsenic, is roasted at about 800 °C with excess oxygen to obtain hematite.

The roaster gas from the first stage leaves the hot gas cyclone with about 18 vol % SO_2. It is first passed with additional air through an afterburning chamber to oxidize any remaining sulfur vapors and residual sulfides. The gases are then cooled in the waste heat boiler. The dust separated there and in the hot gas electrostatic precipitator immediately downstream is rich in arsenic, antimony, and lead and can be treated further to recover them. But the main part of the arsenic remains in the gases as volatile arsenic trioxide and is separated in the wet gas cleaning system.

The gases originating from the second roaster stage, with about 10 vol % SO_2, are cooled in another waste heat boiler and dedusted in a cyclone followed by a hot gas electrostatic precipitator. The gas streams from the two stages are subsequently united (upstream or downstream of the wet gas cleaning system).

In the low-arsenic cinder and dust coming from the second stage, the iron is present as hematite (Fe_2O_3).

5.5. Production in Nonferrous Metal Sulfide Ore Processing

5.5.1. Copper Smelting (→ Copper, Chap. 5.3., → Copper, Chap. 5.4., → Copper, Chap. 5.5., → Copper, Chap. 5.6., → Copper, Chap. 5.7.)

Copper ore concentrates consist mainly of copper pyrites or chalcopyrite ($CuFeS_2$), the most

frequent copper mineral, but may also contain pyrite, bornite, cuprite, chalcocite and others. A typical concentrate composition is 27 % Cu, 28 % Fe, 32 % S. They are processed almost exclusively by pyrometallurgical methods. Pyrometallurgical copper extraction is based on the decomposition of complex iron – copper sulfide minerals into iron and copper sulfides, followed by selective oxidation and separation of the iron portion, and finally oxidation of the remaining copper sulfide. Virtually all the sulfur in the original concentrate ends up as sulfur dioxide. The overall process is conducted in the three discrete stages of roasting, smelting and conversion.

In the roasting stage, the higher sulfide components of the so-called green concentrate are initially decomposed thermally to iron(II) and copper(I) sulfides and elemental sulfur (Eq. 11) [40, 41]. In the prevailing oxidizing atmosphere, this sulfur is instantly converted to sulfur dioxide

$$2\,CuFeS_2 \longrightarrow Cu_2S + 2\,FeS + S \tag{11}$$

$$S + O_2 \longrightarrow SO_2$$

The iron(III) sulfide also undergoes exothermic oxidation. To attain a good separation of iron and copper in the subsequent smelting stage one must ensure that the copper does not become converted to oxide. Therefore, the amount of oxygen is limited so that some of the iron(II) sulfide remains unchanged. Under such conditions magnetite is the predominant oxide (Eq. 12) rather than hematite (cf. Eq. 7)

$$3\,FeS + 5\,O_2 \longrightarrow Fe_3O_4 + 3\,SO_2 \tag{12}$$

The roasted ore, alone or with green concentrate (depending on the concentrate composition and the type of process used), then passes to the smelting stage. Here, in the presence of a siliceous flux, it is heated by the combustion of a fuel, by electricity, or autogenously by oxidation with oxygen to a temperature at which the unconverted sulfides – mainly copper(II) sulfide – are molten. The iron oxide component reacts with the siliceous flux to form a molten slag. Being appreciably less dense, this floats on top of the heavier sulfide phase, known as copper matte. The two can thus be tapped separately.

In the conversion or blowing stage, air or oxygen is blown through the matte in two steps. In the first step (slag blowing), the residual iron sulfide portion is converted to iron(II) oxide, which is removed as a slag from the concentrated matte of copper(I) sulfide. In the second step (copper blowing), the copper(I) sulfide is converted to crude copper (blister copper):

$$Cu_2S + O_2 \longrightarrow 2\,Cu + SO_2 \tag{13}$$

The development of the copper recovery processes during the last few decades has been dominated by two principal objectives. One is the need to economize on energy by making maximum use of reaction heat liberated in the individual process steps. The other is the need to lower the volume and increase the sulfur dioxide concentration of the waste gases as much as possible so as to facilitate environmental pollution control. The trend is increasingly in favor of oxygen technology.

Off-Gas Composition from Specific Process Types. The *reverberatory furnace process* is still used to a considerable extent for smelting copper concentrates, although it suffers from the economic disadvantage of high energy consumption. The design and mode of operation of reverberatory furnaces is described in [42]. Partially preroasted concentrates and green concentrates of low sulfur content are the most frequent feed materials. Copper matte is produced with a copper content about 10 % higher than that of the feed material. The most serious technical disadvantage of this process is the large volume of combustion gas that results from the high consumption of fossil fuels. This dilutes the sulfur dioxide concentration of the total smelter off-gas to ca. 1 – 2 vol %. It is economically scarcely feasible to process this into sulfuric acid.

A comparison of the various pyrometallurgical copper extraction processes from the point of view of the waste gases produced, based on calculations of Schwartz and Eisert [43], emphasizes the problems inherent in the reverberatory furnace process.

By using oxygen-enriched air in oxyfuel burners, the sulfur dioxide content of reverberatory furnace waste gases can be raised to about 6 – 8 vol % and the specific waste gas rate is considerably diminished.

The *electric furnace process* can be used to produce copper matte when electric power is available at low cost. The connected load of such furnaces used in copper metallurgy is up

to 50 MW. The process is in many respects similar to that of the reverberatory furnace except that the specific waste gas rate can be kept relatively low if the furnace is adequately sealed. The waste gas composition depends on the feed material characteristics. The sulfur dioxide content may reach 3 – 10 vol %. Carbon dioxide and carbon monoxide also occur in small concentrations in the waste gas owing to burning of the carbon electrodes.

The *Outokumpu flash smelting process* [44] is one of the standard pyrometallurgical processes used most frequently today. The flash smelting furnace (Fig. 11) consists of an upright cylindrical reaction shaft designed for simultaneous roasting and smelting atop one end of a hearth, over the other end of which is a vertical waste gas shaft. Fine-grained dry concentrate is injected into the reaction shaft with air preheated to 450 – 750 °C or with oxygen-enriched air.

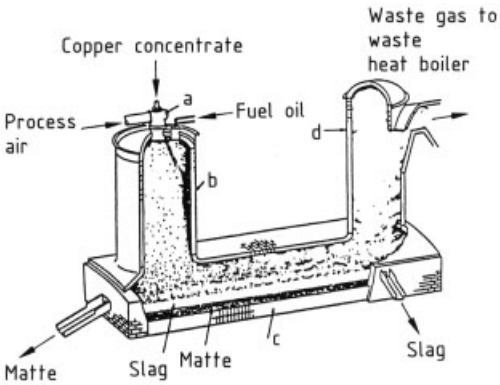

Figure 11. Outokumpu flash smelting furnace
a) Concentrate burner; b) Reaction shaft; c) Settling trough; d) Waste gas shaft

Since the process is not usually autogenous, supplementary heating is necessary, not only in the hearth but also in the reaction shaft, except when oxygen-enriched air is used. In that case, the addition of fossil fuels to the reaction shaft can be dispensed with. The droplets forming in the reaction shaft, consisting of copper matte and slag, accumulate in the hearth, while the gases leave the flash smelting furnace through the waste gas shaft at 1200 – 1300 °C.

Depending on the concentrate composition and the copper concentration of the matte obtained in the flash smelting furnace, about 50 – 70 % of the sulfur content of the concentrate is converted to sulfur dioxide. When operating the furnace without oxygen enrichment, the waste gases contain 10 – 13 vol % of sulfur dioxide, 1 vol % of oxygen, and ca. 0.05 vol % of sulfur trioxide at specific off-gas rates of about 6000 – 6500 m^3/t copper.

Most flash smelting furnaces are operated with reaction air enriched to 45 – 60 % O_2, which substantially reduces the off-gas rate and increases the SO_2 concentration. The relationships shown below are obtained when feeding a concentrate containing 21.9 % Cu, 30.3 % Fe and 32.0 % S.

O_2 content of feed air, vol %	21.0	28.0	31.9	50.0
Inlet temperature of feed air, °C	450	200	25	25
Specific off-gas volume, m^3/t Cu	6016	4593	4087	2363
Off-gas SO_2 content, vol %	13.4	17.8	19.8	28.5

Oxygen has already been used continuously on an industrial scale in the *Inco flash smelting process*. Inco of Canada developed a flash smelting furnace with a daily throughput of up to 1500 t of concentrate. After pretreatment in a fluid bed dryer, the concentrates are injected into the reaction space with 95 % oxygen. The process operates without supplementary heating. The specific off-gas rate is as low as 1090 m^3/t Cu at about 80 vol % SO_2. After adequate gas cleaning, the sulfur dioxide-containing gases are chiefly converted to liquid sulfur dioxide.

The *flame cyclone processes* for intensive smelting with oxygen represent another practical option. On an industrial scale, a combination of a flame cyclone reactor with an electric resistance furnace has been used in the *Kivcet process* developed at the Irtysh metal combine at Glubokoye in the former Soviet Union [45]. The feed material is copper concentrate or complex zinc – lead – copper concentrate, which must be of fine grain size and pourable. Tonnage oxygen is used for the smelting process. The off-gases leave the furnace at about 1250 – 1350 °C. They contain 80 – 85 vol % SO_2. The specific off-gas rate ranges from 800 to 2000 m^3/t Cu, depending on the concentrate composition.

Composition of Converter Off-Gases. Converter processes are generally used for blowing air or oxygen-enriched air through copper matte to convert it to crude copper. The blowing procedure is not continuous, nor are, therefore, the composition and volume of the off-gas. Dur-

ing slag blowing, the sulfur dioxide content of the waste gases at the converter mouth is up to ca. 10 vol %, but during copper blowing it may rise to as much as 20 vol %.

If oxygen-enriched air is used, correspondingly higher sulfur dioxide concentrations are achieved, but precautions must be taken due to the increased generation of heat. However, it is not only the continually fluctuating sulfur dioxide concentrations which are typical of converter processes but also the heavy fluctuations in the gas flow rates.

Because air leaks into the exhaust hoods of the converters, the converter off-gases are diluted considerably. The dilution factors are between ca. 1.8 and ca. 5 [46]. In the *Hoboken syphon converter* (Fig. 12), operated at Metallurgie Hoboken Overpelt in Belgium, air inleakage is effectively avoided [47].

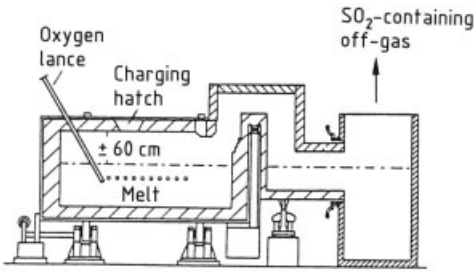

Figure 12. Hoboken syphon converter

Developments to Overcome Off-Gas Processing Problems. All sulfur dioxide-containing gas streams from copper smelting operations, including converter off-gases, must be processed to prevent more than traces of sulfur dioxide entering the atmosphere. The key to successful purification of sulfur dioxide-containing gases to the required standards is continuous stable operation of the purification process, be it a sulfuric acid plant (as is most common at smelters) or some other kind of abatement process. Gas streams which are liable to wide fluctuations in both volume and composition, such as converter off-gases, are quite incompatible with this criterion. These fluctuations can be mitigrated to some extent by conducting the conversion procedure in several converters in a phased sequential operation and combining their off-gases [48]. It is also common practice to mix them with the steadier, more concentrated gas streams from the roasting and smelting stages to produce a combined gas stream within the concentration range needed to maintain autothermal operation of a sulfuric acid plant [49]. As an example, the routing of off-gases from a modern copper smelter incorporating an Outokumpu flash smelting furnace and Pierce Smith [50] converter operation is schematically represented in Figure 13.

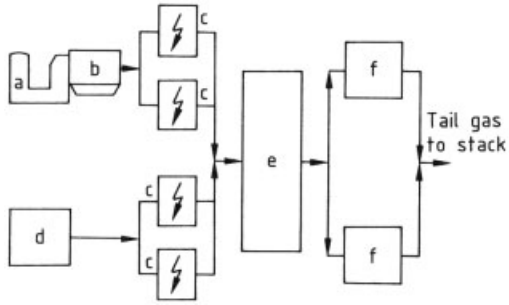

Figure 13. Sulfur dioxide production and processing in a modern copper smelter
a) Smelting furnace; b) Waste heat boiler; c) Hot electrostatic precipitator; d) Pierce Smith converter; e) Wet gas cleaning system; f) Sulfuric acid plant

The chief aim of current development work is to unite the individual stages of crude copper extraction into one continuous process, avoiding any discontinuous converter stages. This has been achieved on an industrial scale by Mitsubishi. In the *Mitsubishi process*, the individual process steps are carried out in separate furnaces which are connected with each other via launders to form a continuously operating system [51].

5.5.2. Nickel Smelting (→ Nickel, Chap. 4.)

The mineral used most commonly for nickel production is pentlandite $(Ni, Fe)_9S_8$, which often occurs in conjunction with pyrrhotite (Fe_7S_8) and chalcopyrite ($CuFeS_2$). Nickel flotation concentrates contain 6 – 12 % Ni, 2 – 10 % Cu, 25 – 35 % Fe and 25 – 32 % S.

As in copper metallurgy, the basic operations of nickel extraction are roasting, smelting, and blowing [52].

During concentrate roasting, initially pyrrhotite is oxidized (Eq. 14):

$$3\,Fe_7S_8 + 38\,O_2 \longrightarrow 7\,Fe_3O_4 + 24\,SO_2 \quad (14)$$

Roasting can be performed in sinter machines, multiple-hearth furnaces or, in modern plants, in fluid bed furnaces.

At Inco, at Copper Cliff, Ontario, Canada, ca. 1350 t/d of nickel concentrate are roasted in a *fluid bed furnace* (diameter 5.5 m, height 7.9 m). About 40 000 m^3/h of roasting air is injected through 566 nozzles; the fluid bed has a height of 1.2 m. Roasting proceeds at 650 °C. The roaster gases contain ca. 12 vol % SO_2.

The *Outokumpu flash smelting furnace* yields an off-gas containing 10 – 12 vol % SO_2 when the roasting air is preheated to 500 °C. Oxygen enrichment of the reaction air to 27 – 35 vol % O_2 allows throughput to be increased from 400 to 700 t/d of nickel concentrate and simultaneously raises the sulfur dioxide content in the furnace off-gas to 17 – 18 vol % [44].

A smelting furnace operated by Inco with tonnage oxygen achieves throughputs of 1500 t/d of nickel concentrate. From the furnace off-gases, containing about 80 vol % SO_2, liquid sulfur dioxide is recovered. The annual liquid sulfur dioxide production amounted to ca. 110 000 t in 1970 [53].

The smelting and blowing operations of most process steps of nickel smelting are technically comparable to those of copper metallurgy.

5.5.3. Zinc Blende Roasting (→ Zinc)

Zinc blende (ZnS) is the most important primary material for zinc production. Zinc blende is enriched by flotation to obtain a concentrate with a typical composition of 55 % Zn, 8 % Fe, 32 % S.

The first step in all zinc recovery processes is the roasting of zinc blende to convert zinc sulfide to zinc oxide, simultaneously removing the bound sulfur [54, 55]. In the exothermic roasting reaction, the sulfur is liberated as sulfur dioxide.

$$2\,ZnS + 3\,O_2 \longrightarrow 2\,ZnO + 2\,SO_2 \quad \Delta H = -440\,\text{kJ/mol} \quad (15)$$

To produce a calcine which can easily be leached for electrolytic zinc recovery, which accounts for about 85 % of world production, especially severe demands are made on the roasting process. The calcine should contain not more than 0.3 % sulfide sulfur and 1.8 % sulfate sulfur. The formation of zinc ferrite ($ZnO \cdot Fe_2O_3$), which would disturb the leaching process, must be avoided as far as possible by maintaining constant, optimum roasting temperatures. At the same time, the heat released in the roasting process should be utilized as efficiently as possible in steam generation. Also, to permiteconomic processing of the sulfur dioxide roaster gases, the sulfur dioxide concentration should be kept as high and steady as possible. These requirements can be fulfilled with the aid of fluid bed technology originally developed for the roasting of pyrite.

Vieille Montagne – Lurgi Process. Figure 14 shows the layout of a fluid bed roasting plant for zinc concentrate according to the process of Vieille Montagne and Lurgi, which is nowadays used worldwide [56]. The concentrate, containing ca. 8 % water, is continuously

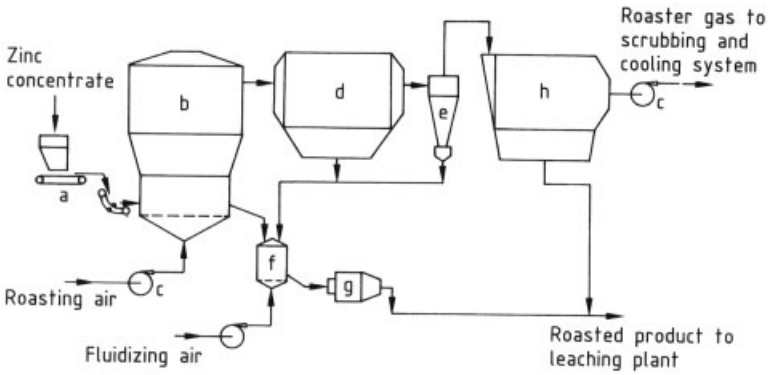

Figure 14. Vieille-Montagne – Lurgi zinc concentrate fluid bed roasting plant
a) Slinger conveyor; b) Fluid bed roaster; c) Blower; d) Waste heat boiler; e) Cyclone; f) Fluid bed cooler; g) Ball mill; h) Hot-gas electrostatic precipitator

fed into the fluid bed furnace by slinger conveyors. Roasting air is injected through a nozzle grate. The gas velocity in the fluid bed is ca. 0.5 m/s. Cooling elements incorporated in the fluid bed maintain the roasting temperature between 900 °C and 1000 °C. The dissipated heat is recovered in steam generation.

Depending on the composition of the zinc concentrate, the roaster gas at the furnace outlet contains 9 – 11 vol % SO_2 and 5 – 6 vol % O_2. A characteristic feature of fluid bed furnaces is that gas flow rates and gas compositions are relatively constant, which is advantageous from the point of view of subsequent processing. In a downstream waste heat boiler, the roaster gases are cooled to ca. 350 °C and then passed on to the hot gas electrostatic precipitators. About 1 t of high-pressure steam at 400 °C is produced per tonne of concentrate processed.

All other cooling and cleaning operations are as practised in conventional sulfur dioxide gas cleaning systems (see Chap. 6).

The zinc oxide calcine is recovered at various points of the roasting plant in the following approximate proportions:

Fluid bed discharge	20 – 30 %
Waste heat boiler discharge	30 – 40 %
Cyclone discharge	22 – 48 %
Electrostatic precipitator discharge	2 – 5 %

The precise distribution is influenced by the grain size range of the concentrate and by controlled wetting and agglomeration of the concentrate before it is fed onto the slinger belts.

Some characteristic data for the largest zinc blende fluid bed furnace based on the Vieille Montagne – Lurgi process, which has been in operation since 1975 at Electrolytic Zinc in Risdon, Australia [57] are presented below:

Grate area	123 m^2
Furnace volume	2680 m^3
Concentrate throughput (dry)	800 t/d
Roasting air rate	70 000 m^3/h
Roasting temperature	920 ± 20 °C
Roaster off-gas (dry)	67 000 m^3/h
Roaster gas SO_2 content	10 – 11 vol %
Steam generation	1 t/t of concentrate
Steam grade	40 bar/400 °C
Sulfide sulfur in calcine	0.15 – 0.20 %
Sulfate sulfur in calcine	1.4 – 1.6 %

Sanki – Dorr Process. In the Sanki – Dorr fluid bed process, the concentrate is fed to the roaster as an aqueous suspension containing ca. 78 % solids. The fluid bed temperatures are controlled by injecting additional water. This wet feeding method yields substantially less steam than dry feed processes (ca. 0.68 t per tonne of concentrate). It also produces about ten times as much dilute scrubber acid, which must be disposed of [58].

Characteristic data for a Sanki – Dorr plant operating at Mitsui Mining and Smelting, Hiroshima, Japan, are listed below:

Grate area	66 m^2
Furnace height	10.4 m
Concentrate throughput (dry)	250 t/d
Roasting air	18 960 m^3/h
Roasting temperature	930 °C
Concentrate contents	
Zn	50.6 %
Fe	9.6 %
Pb	1.04 %
Cu	0.42 %
Cd	0.28 %
S	30.7 %
Sulfide sulfur in calcine	0.22 %
Sulfate sulfur in calcine	1.63 %

The roasting air is injected into the furnace through 1025 nozzles and the fluid bed is about 1.2 – 1.5 m high.

Oxygen Enrichment. The effect on the throughput of adding oxygen in zinc blende fluid bed roasters has been tested, and several plants are now being operated successfully with oxygen-enriched air. The conditions in the phase diagram of the system Zn – Fe – S – O [59] suggest that there is a limit on the extent to which it is practical to enrich the roasting air with oxygen on account of undesirable sulfate and ferrite formation. According to information from Electrolytic Zinc Co., which has operated a 360 t/d nominal capacity fluid bed furnace with oxygen-enriched air, the optimum was at ca. 24 vol % O_2. By this means it was possible to raise the throughput by ca. 25 % and the sulfur dioxide content of the roaster gases to 12 – 12.5 vol %.

5.5.4. Lead Smelting (→ Lead, Chap. 4.2.→ Lead, Chap. 4.3.→ Lead, Chap. 4.4.)

Lead sulfide ore flotation concentrates, containing ca. 60 % Pb and 17 % S, are roasted almost exclusively in travelling-grate sinter plants.

The oxidation of the lead sulfide is initiated by chemisorption of oxygen on the particle surface; the lead sulfate formed then reacts with the remaining lead sulfide and oxygen to form basic sulfates and, finally, litharge (PbO), releasing sulfur dioxide [60, 61]

$$PbS + 2\,O_2 \longrightarrow PbSO_4 \tag{16}$$

$$PbSO_4 + PbS + 1.5\,O_2 \longrightarrow PbO \cdot PbSO_4 + SO_2 \tag{17}$$

$$PbO \cdot PbSO_4 + PbS + 1.5\,O_2 \longrightarrow 2\,PbO \cdot PbSO_4 + SO_2 \tag{18}$$

$$(x\,PbO \cdot PbSO_4) + PbS + O_2 \longrightarrow (x+2)\,PbO + 2\,SO_2 \tag{19}$$

Metallic lead can be formed in side reactions

$$PbS + 2\,PbO \longrightarrow 3\,Pb + SO_2 \tag{20}$$

$$PbS + PbSO_4 \longrightarrow 2\,Pb + 2\,SO_2 \tag{21}$$

so the sinter product frequently contains entrained lead droplets as well as litharge.

The sulfur dioxide concentrations of lead concentrate roaster gases are in the range 4 – 6 vol % under optimum operating conditions when sintering machines with gas recycling are used.

Sintering roasting of mixed zinc – lead concentrates, as used in the *Imperial Smelting shaft furnace* for simultaneous recovery of zinc and lead, gives only slightly higher sulfur dioxide concentrations of 5 – 7 vol %.

The sulfur dioxide concentration of sintering roaster gases is subject to large and irregular fluctuations and may remain at extremely low levels over extended periods. This makes autothermal operation of a double-absorption sulfuric acid plant difficult. The oxygen content of the roaster gases discharged from the sinter plant corresponds already to more than the required minimum O_2/SO_2 ratio for a contact sulfuric acid plant.

The sulfur trioxide content is influenced to a certain extent by the sintering temperature and by the presence of other materials such as iron oxide. It may amount to as much as 5 – 10 % of the sulfur dioxide content.

The dust content of lead concentrate roaster gases is ca. 5 – 10 g/m^3; in the case of zinc – lead sinter roasting it may even be as high as 20 – 30 g/m^3. Therefore, an efficient gas cleaning system is required. Sinter roasting gases from the processing of flotation concentrates also usually contain varying amounts of hydrocarbons, which originate from additives used in the flotation process. These are only partially burnt in the sintering process, and they pass almost completely through the normal gas cleaning systems, only being oxidized when they reach the dryer of the sulfuric acid plant. The colloidal carbon formed by oxidation colors the acid black. Lighter hydrocarbons even reach the converter, where they burn.

In sintering technology attempts have been made to raise the roasting capacity and increase the sulfur dioxide content of the roaster gas by oxygen enrichment [62]. Although the throughput and sulfur dioxide concentration could be raised by a small extent, there are decisive factors in the sintering process, such as the porosity of the mix and the optimum specific air load for heat transport in the hot sintering zone which limit the increase in throughput [63]. The cost of using technical oxygen is out of all proportion to the potential benefits [39], so this procedure has not been adopted in industrial practice.

Apart from these problems in processing roasting gases with low and fluctuating sulfur dioxide concentrations, dust emissions and dedusting within the sinter plant itself present particular difficulties. Development work on pyrometallurgical lead extraction has, therefore, been directed towards avoiding the complicated sintering process and carrying out the roasting and reduction procedures in a single reaction unit so as to produce waste gases with a higher concentration of sulfur dioxide. *Outokumpu Oy*, for instance, has processed lead concentrates on a pilot scale in a flash smelting furnace, applying principles analogous to those of copper technology. Off-gases with sulfur dioxide concentrations of 11 vol % have thus been obtained [64].

In Italy, a plant for the production of 80 000 t/a lead is in operation based on the *Kivcet process*, which is used in the former former Soviet Union for copper extraction. By using technical oxygen, gases containing 20 vol % SO_2 are obtained from the smelting of lead concentrates.

In the *QSL process* (the Queneau – Schuhmann process as developed further by Lurgi) [65], (see next page, Fig. 15), the lead concentrate is continuously fed into a horizontal cylindrical reactor with flue dust and flux

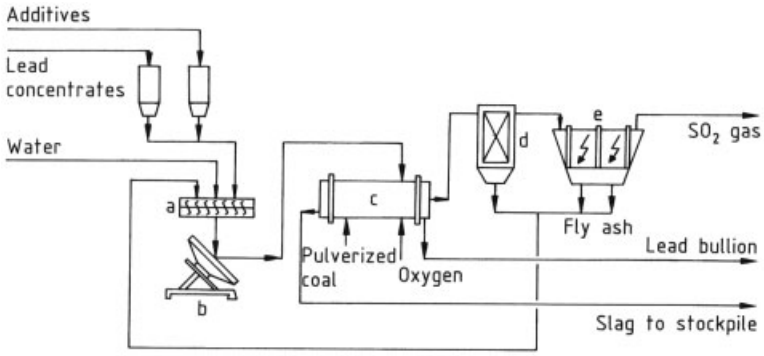

Figure 15. QSL lead smelting process
a) Mixer; b) Pelletizing disk; c) Reactor; d) Waste heat boiler; e) Hot gas electrostatic precipitator

additives in the form of green pellets. In the oxidation zone the lead sulfide is partially oxidized by injection of technical oxygen (Eq. 22)

$$2\,PbS + 2.5\,O_2 \longrightarrow Pb + PbO + 2\,SO_2 \qquad (22)$$

The lead oxide is taken up by the slag and is reduced by carbon injected into the subsequent reduction zone. The gases leaving the reactor at ca. 1200 °C reach concentrations of 15 – 20 vol % SO_2. QSL plants are in operation in Germany and Korea, with capacities of 75 000 t/a and 60 000 t/a lead, respectively.

5.6. Production by Decomposition of Sulfuric Acid and Sulfates

5.6.1. Waste Sulfuric Acid Processing [66]

Sulfur dioxide-containing gases can be produced by decomposing sulfuric acid. This procedure has been used industrially since 1932 as a means of disposing of heavily contaminated spent sulfuric acid while recovering its sulfur values. The various processes can be classified into two groups [67, 68]. Sulfur dioxide is produced either by thermal decomposition at 800 – 1300 °C or by reducing decomposition at 200 – 600 °C. The input to the process is usually spent sulfuric acid so heavily contaminated with organic substances that it cannot be restored to a useful level of purity by other processes (see → Sulfuric Acid and Sulfur Trioxide, Chap. 4.3.). In some cases the organic contaminants may be of such a nature and be present in such concentration that the acid is more of a sludge than a liquid. In that case it will have a relatively high heating value on account of the impurities, which is advantageous in supplying a portion of the heat required for decomposition. Decomposition is the most reliable way to recover high-grade sulfuric acid or oleum from such wastes, as the impurities are removed completely [69–71].

The sulfur dioxide-containing gases obtained by decomposition of spent sulfuric acid are generally converted back into sulfuric acid in a contact plant after purification (see Chap. 6). The most important prerequisite for a decomposition process, therefore, is that it should generate a gas with a high enough sulfur dioxide concentration for use in a contact plant. This means that it should contain at least 4 vol % of sulfur dioxide after its O_2/SO_2 ratio has been adjusted to the requisite value by addition of air [72]. The sulfur dioxide concentration depends chiefly on the H_2SO_4 content of the waste acid and the heating value of the oxidizable impurities in the spent acid being decomposed.

Thermal Decomposition. Gas-phase thermolysis of sulfuric acid is represented by the overall equation

$$H_2SO_4 \longrightarrow SO_2 + H_2O + 0.5\,O_2 \quad \Delta H = +202\,\text{kJ/mol} \qquad (23)$$

To ensure complete decomposition of the acid and combustion of the organic impurities, decomposition is carried out under oxidizing conditions at ca. 1000 °C [69]. This is a strongly endothermic process, and the more dilute the waste acid the more additional energy is needed to evaporate water before decomposition of sulfuric acid begins. To decompose acid of ini-

tial concentration 70 % at 1000 °C, the theoretical energy requirements are ca. 6.4×10^6 kJ/t H_2SO_4 [69]. About 46 % of this energy is required for dehydration of the 70 % sulfuric acid, evaporation of the water, heating to the maximum boiling temperature of about 340 °C, and evaporation of H_2SO_4. Another 31 % is required for decomposing H_2SO_4 into SO_2, H_2O and O_2; the remaining 23 % is required for heating the decomposition products to the reaction temperature of 1000 °C.

However, this theoretical energy demand of ca. 6.4×10^6 kJ/t H_2SO_4 applies only in the case of indirect heat transfer [69]. Furthermore, there are no construction materials for the heat exchangers that are resistant to corrosion under these reaction conditions. Therefore, the heat required for acid decomposition is in practice supplied directly by means of hot flue gases from the combustion of fossil fuels. Inevitably the heat requirements are far higher than in the case of indirect heating, as the flue gases must be heated to the decomposition temperature (1000 °C). But a large part of the additional heat can be recovered when the gases leaving the decomposition reactor are cooled from 1000 °C to about 350 °C.

The energy requirement for acid decomposition depends mainly on the composition of the spent acid. It rises steeply with decreasing H_2SO_4 concentration in proportion to the amount of water to be evaporated and it is diminished by the heating value of the organic impurities. In the case of dilute spent acids, it is always energetically more favorable to preconcentrate to ca. 60 – 75 % H_2SO_4 so as to separate the bulk of the water before decomposition rather than to decompose the dilute acid directly.

The sulfur dioxide-containing gases produced in the decomposition of sulfuric acid are heavily diluted by the combustion gases introduced for direct heating. However, the amount of flue gas can be substantially reduced by preheating the combustion air. To obtain gases from the decomposition with a high enough sulfur dioxide content for processing in a contact sulfuric acid plant when using fuel oil, the spent acid must contain at least 60 % H_2SO_4 even when the combustion air is preheated to 450 °C.

Another method of reducing external energy input and increasing SO_2 concentrations is the use of oxygen-enriched air.

Figure 16 shows the operation window for the spent acid decomposition process.

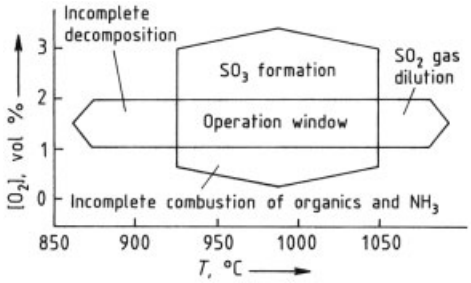

Figure 16. Operation window for the spent acid decomposition process

The decomposition of spent sulfuric acid at high temperatures was developed in the 1930s by *Lurgi* [65]. Figure 17 (see next page) shows the flow diagram of a modern spent acid regeneration plant. The decomposition unit itself consists of the empty decomposition furnace and the waste heat boiler system. Downstream of this is a gas cleaning system for further cooling and purifying the decomposition products (see Chap. 6) and a double-absorption plant producing concentrated sulfuric acid and oleum. The application of such an empty decomposition furnace is, however, limited to acids with low residue levels and low organics concentrations that do not result in excess heat beyond the thermal energy requirements of the process.

Spent acid is fed to the top of the vertical, brick-lined furnace and is finely distributed by means of a rotary atomizer [69] specifically developed by Lurgi for atomizing viscous spent sulfuric acid. The atomizing medium is cold air, injected at ca. 1.35 bar. Fuel oil or gas is burnt in several combustion chambers attached laterally to the top of the furnace. The separation of the combustion chamber from the furnace permits complete mixing and combustion of the fuel with the combustion air outside the acid decomposition zone proper. A 10 – 30 % excess of air over the amount actually required for combustion of the fuel is maintained. This is because the amount of oxygen liberated in the decomposition of the acid is not normally enough to burn the oxidizable inorganic and organic components completely. The flue gases enter the furnace tangentially on the same level as the combustion chambers at ca. 1800 °C. A part of their

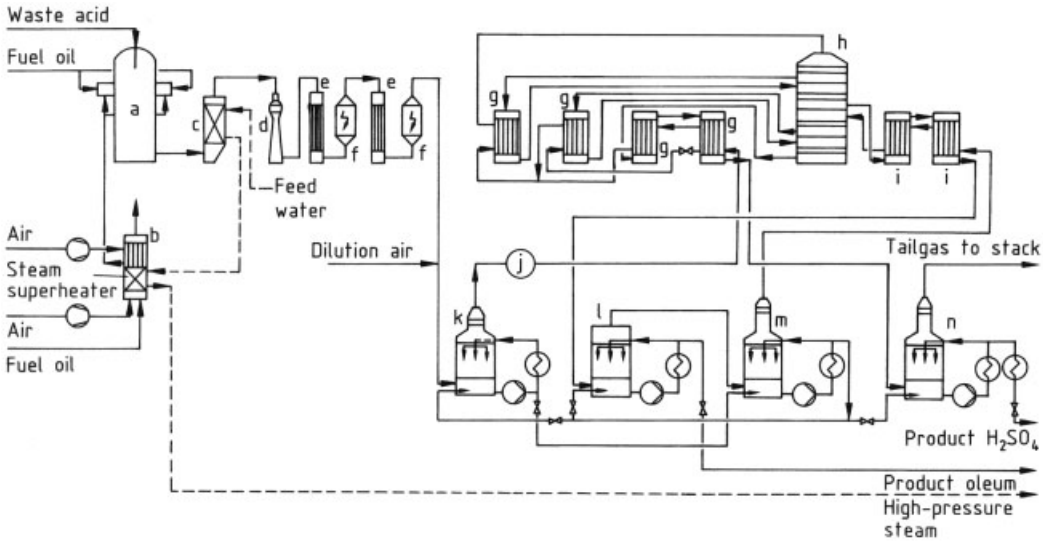

Figure 17. Sulfuric acid regeneration plant
a) Decomposition furnace; b) Air preheater; c) Waste heat boiler; d) Venturi scrubber; e) Gas cooler; f) Electrostatic precipitator; g) Gas heat exchanger; h) Converter; i) Intermediate heat exchanger; j) Fan; k) Drying tower; l) Oleum tower; m) Intermediate absorber; n) Final absorber

heat content is transferred to the furnace top by radiation, but the majority is transferred to the acid because of the intensive mixing which takes place in the vortex formed near the outlet openings of the combustion chambers. The furnace temperature of ca. 1000 °C required for acid decomposition is adjusted by controlling the fuel and combustion air rates.

The sulfur dioxide-containing gases leaving the furnace are cooled to ca. 350 °C in an energy recovery system. Depending on the intended use of the recovered energy, different waste heat systems may be used for this purpose. Figure 18 shows three typical decomposition plant concepts with different energy recovery systems [69].

The system in Figure 18 A (analogous to Fig. 17) is designed to produce high-pressure steam (60 – 65 bar). The heat content of the gases emerging from the furnace is used to generate saturated steam. The steam is then heated to ca. 450 °C in a superheater fired by fuel oil. The residual heat of the flue gas is recovered in a gas – gas heat exchanger which preheats combustion air for the decomposition furnace to about 350 °C. Enough high-pressure steam is produced by this system to power a back-pressure turbine drive on the main blower of the downstream sulfuric acid plant, which – as is usual for a "cold gas" plant – does not generate steam of its own. The low-pressure steam exhausting from the back-pressure turbine may be used to provide heat for preconcentrating the spent sulfuric acid.

This process concept is specifically suited for larger plants with capacities > 100 t/d H_2SO_4. It has, for example, been used in a Lurgi-built plant regenerating 360 t/d of 65 % waste sulfuric acid.

In the system of Figure 18 B, the decomposition furnace is followed by a gas – gas heat exchanger, in which the decomposition products are cooled indirectly from 1000 °C to ca. 350 °C. Simultaneously, the furnace combustion air is heated to ca. 450 °C. In general, only about one third of the hot air generated in this heat exchanger is required in the fuel oil burners of the decomposition furnace. The remaining two thirds can be utilized, for instance, for generating steam at 10 bar in a waste heat boiler. A waste heat system of this type has been installed in a twin-stream decomposition plant built by Lurgi for processing ca. 900 t/d of ammonium sulfate- containing spent sulfuric acid resulting from acrylic acid production.

The system of Figure 18 C has the same gas – gas heat exchanger arrangement for cooling the decomposition gas and preheating air as system B. But in this case the surplus hot air at about

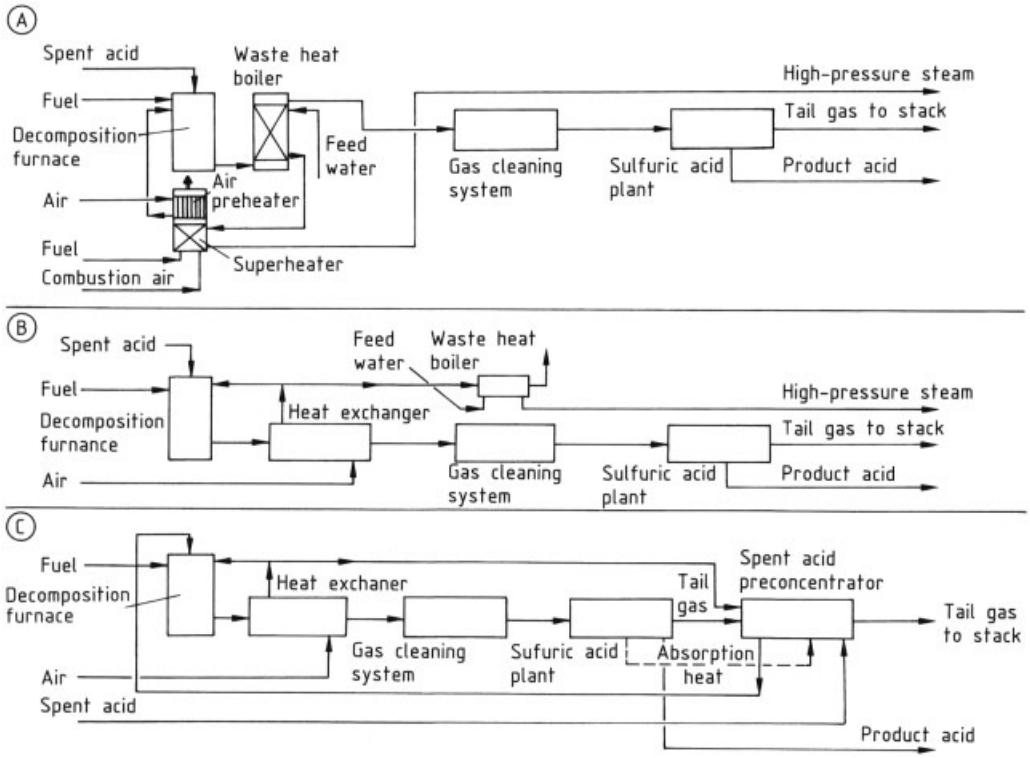

Figure 18. Typical sulfuric acid regeneration plant concepts with different energy recovery systems [19]

450 °C is used, along with the dry tail gases of the sulfuric acid plant, for preconcentrating the spent sulfuric acid. In this way, 43 % sulfuric acid can be concentrated to 70 %. The preconcentration unit and the downstream sulfuric acid plant can be integrated so as to allow waste heat from the acid plant absorption system to be utilized for preheating the contaminated acid. In this case spent acid even as dilute as 33 % H_2SO_4 can be concentrated to 70 % using only waste heat [69].

In another decomposition process, developed by *Chemical Construction Corporation* (Chemico) in co-operation with Consolidated Chemical Co. (now *Stauffer Chemical*) [73], the acid is decomposed at 1000 – 1200 °C in a horizontal furnace in which the fuel is burnt directly. For this process, exact control of the oxygen content of the furnace gases is necessary. If the gases contain less than about 2 % free oxygen, there is a risk that sulfur dioxide may be reduced to elemental sulfur, which is precipitated in the subsequent colder parts of the plant. The spent acid is injected through nozzles into the furnace. Acid sludges with a high content of organic impurities are mixed with spent acids from other sources to reduce their viscosity so that they can be atomized more easily.

Spent acids with widely varying physical and chemical properties can be decomposed in the Grillo rotary kiln process. The Grillo process (Fig. 19, see next page) is a multistage decomposition process which is particularly suitable for the decomposition of spent sulfuric acid with widely varying physical and chemical properties.

The Grillo acid decomposition plant consists of a rotary kiln decomposition reactor, a downstream intermediate combustion chamber, and two post-combustion chambers.

The waste acid is injected virtually unpressurized onto a red hot coke bed in the rotary kiln with relatively coarse atomization. The coke bed serves the following three main functions:

1) Uniform distribution of the liquid feed, which offers advantages in terms of reaction rate and uniform reaction conditions
2) Heat storage and regulation

3) Most of the solids content of the waste acid is retained in the coke bed.

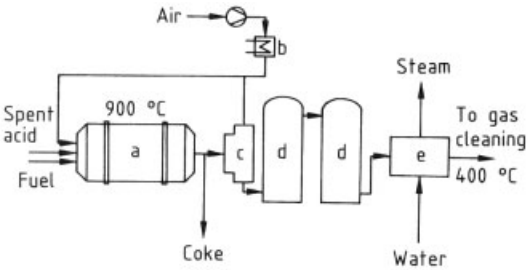

Figure 19. Lurgi Grillo decomposition process
a) Rotary kiln decomposition furnace; b) Air preheater; c) Intermediate combustion chamber; d) Post combustion chamber; e) Waste heat boiler

In the rotary kiln, the spent acid is burnt under substoichiometric conditions at 900 – 1000 °C, so that the decomposition gases contain H_2S and COS in addition to CO. By adding combustion air to the intermediate combustion chamber downstream of the rotary kiln, the organic decomposition gas components are completely oxidized at 1250 °C.

Advantages of this process, which is relatively simple to control, are the drastic reduction of the NO_x levels in the decomposition gas, the complete destruction of PCB- containing oils due to the high reaction temperatures, and the retention of the acid-borne solids in the form of coking duff.

Even when using residue oils with high PCB levels as an energy source, this process meets all environmental standards.

For treating spent acids with high residue concentrations dry metal sulfates, and all mixtures of sulfur- containing solids with spent acid, the fluid bed decomposition (Fig. 20) process is particularly suitable.

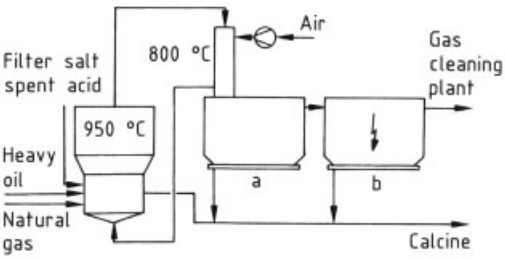

Figure 20. Decomposition in fluid bed reactor
a) Waste heat boiler; b) Hot electrostatic precipitator

The stationary fluid bed process for sulfate and spent acid decomposition has been successfully employed on a commercial scale for more than 10 years.

A further development of this process is circulating fluid bed technology, which achieves improved decomposition efficiencies due to staged reaction control in the decomposition furnace.

In thermal sulfuric acid and sulfate decomposition, staged process control will gain in significance, as this is the simplest method to suppress NO_x formation. The decomposition temperatures are in the range of 900 – 1000 °C.

Where feasible, the decomposition of spent acids/sulfates can be combined with the roasting of pyrites. The rate at which acid can be processed is limited by the available pyrite roasting reaction heat. When the spent acid is of about 70 % concentration, the ratio of the sulfur dioxide produced from pyrite to that produced from the acid is about 3 : 1. This means that the sulfuric acid output of the downstream plant is four times the H_2SO_4 content of the spent acid feed to the fluid bed reactor. To increase the decomposition capacity of the pyrite roaster, fuel oil, sulfur and oxygen-enriched air may be introduced.

Reductive Decomposition. In a process developed by *Hechenbleickner*, spent sulfuric acid is decomposed at 200 – 600 °C after mixing with a reducing material such as tar sludge or coal. The reduction product is sulfur dioxide, and coke is generated as a byproduct [68]. This process is chiefly used for processing acid sludge from lubricating oil and kerosene production plants associated with oil refineries. In the original process concept, the sludge, which contains up to 50 % organic matter and has a comparatively high calorific value, is fed through a rotary kiln in countercurrent with hot flue gas. The most important variants of this process are known under the names of *Chemico* [73, 74], *Ross – Wilde*, and *Miley* [68]. A further development of the process has been designed to process a wide range of spent acids containing from 0.5 to 60 wt % C and 40 – 90 % H_2SO_4 [75].

In all process variants of the reductive acid decomposition, the sulfur dioxide-containing raw gas obtained contains, besides carbon dioxide and moisture, varying proportions of carbon monoxide and hydrocarbons [68]. Since

the hydrocarbons, especially, are detrimental in a downstream contact plant, generally post-combustion at about 800 °C is required before the gas cleaning stage.

5.6.2. Sulfate Decomposition

Calcium Sulfate. All industrial processes for producing sulfur dioxide-containing gases by decomposition of calcium sulfate, the so-called gypsum – sulfuric acid processes, are based on the *Müller – Kühne process* [76, 77]. In this process the calcium sulfate is decomposed in a rotary kiln at 900 – 1400 °C to give sulfur dioxide-containing gases and cement clinker. The sulfur dioxide is processed to sulfuric acid in an associated contact plant. The clinker is used for making Portland cement, whose sale recovers part of the extremely high costs.

Though originally the raw materials for this process were natural anhydrite or natural gypsum, the process has been employed for decomposing waste gypsum from chemical processes, especially the phosphogypsum, produced as a waste product in the manufacture of phosphoric acid by the wet process (sulfuric acidulation of phosphate rock). In addition, coke is required as a reducing agent and additives such as clay, sand, and iron oxide are needed to convert the lime into a clinker that can be utilized in cement manufacture.

The Müller – Kühne process is based on the reductive decomposition of calcium sulfate [78] with carbon above 700 °C. Part of the calcium sulfate is reduced to calcium sulfide, which reacts above 900 °C with the remaining calcium sulfate to form calcium oxide and sulfur dioxide (Eq. 24 – 26)

$$CaSO_4 + 2\,C \longrightarrow CaS + 2\,CO_2 \tag{24}$$

$$CaS + 3\,CaSO_4 \longrightarrow 4\,CaO + 4\,SO_2 \tag{25}$$

$$4\,CaSO_4 + 2\,C \longrightarrow 4\,CaO + 4\,SO_2 + 2\,CO_2 \tag{26}$$

It is assumed that the reduction to calcium sulfide takes place not as a solid-state reaction but by way of an intermediate stage involving carbon monoxide [79]. The silica-, alumina- and iron-containing additives accelerate the release of the sulfur dioxide by reducing the decomposition temperature and by undergoing side reactions with the calcium sulfate [77]. They also react at 1200 – 1400 °C with the resulting calcium oxide to form cement clinker. The clinker quality depends essentially on exact proportioning of the carbon in relation to the calcium sulfate [78, 80]. Although the stoichiometric molar C/$CaSO_4$ ratio for decomposing calcium sulfate is 1:2, a carbon excess of about 20 % is applied in practice because part of it will burn directly and will therefore not be available for reduction [79]. If not enough carbon is present, the clinker will contain unconverted sulfate and will melt; if there is too much, the clinker will contain sulfide and will disintegrate. Under strongly reducing conditions in the bed, elemental sulfur may form and vaporizes under the prevailing conditions. It can be burnt in the gas phase merely by injecting additional air, but this is undesirable because the nitrogen from the extra air dilutes the sulfur dioxide in the product gases.

The primary materials have to be prepared specifically for the process. Gypsum must first be calcined to remove its water of crystallization and convert it to anhydrite, a process which requires additional energy. The anhydrite, coke, sand, and alumina are finely ground, dried, and fed into a mixer by an automatic metering device. In some cases, the raw meal is pelletized [81]. The average composition is 78 – 80 % anhydrite, 9 – 10 % clay, 4 – 5 % sand, and 6 – 7 % coke. The material is fed in at the upper end of the slightly inclined rotary kiln and passes in countercurrent with the hot gases to the lower, heated end, from which the clinker product is discharged.

The rotary kiln (normally about 70 m long and about 3.2 m inside diameter) is refractory lined with magnesite bricks at its lower end, in the clinkering zone, where a temperature of about 1400 °C prevails. Formerly, pulverized coal was used for heating but those plants which are still in operation preferentially use fuel oil.

The raw gas leaving the kiln at about 350 – 420 °C contains ca. 7 – 10 vol % SO_2, 0.7 – 1.6 vol % O_2, 15 – 20 vol % CO_2, and 23 vol % H_2O [78]. It is efficiently cleaned in a gas cleaning system (similar to those used at roasting plants; see Chap. 6) and is then processed in a contact sulfuric acid plant after addition of enough air to bring the O_2/SO_2 ratio to the required value. The total sulfur yield is about 85 %.

The hot cement clinker leaving the kiln is cooled with air in the clinker cooler and is then ground with about 2 % of added gypsum to ob-

tain the cement product [82]. The attainable quality compares with that of high-grade Portland cement. With this process, cement and sulfuric acid are produced in equal quantities. Per tonne of cement or sulfuric acid, the following raw material quantities are required [77]: 1.6 t anhydrite or 2.2 t natural gypsum, 0.3 t clay, 0.1 t sand, and 0.1 t coke. Furthermore, a total of 9.6 – 11.7×10^6 kJ of energy in the form of fuel oil are consumed. This includes that required in the raw material preparation system.

If phosphogypsum is used, it is much more difficult to produce high-grade cement, and the cleaning of the sulfur dioxide- containing gases is much more complicated than when natural gypsum is used [74]. On account of its high moisture content (up to 40 %), phosphogypsum requires a large amount of calcining energy. The precise contents of residual phosphate and fluoride depend on the phosphoric acid production process and the rock phosphates used, but they are practically always considerably above the very low levels allowable either in cement clinker or in sulfuric acid plant feed gas. Therefore, they have to be removed by a specific, elaborate pretreatment of the phosphogypsum. Material analyses and consumption figures for the *OSW-Krupp process*, developed by Oesterreichische Stickstoffwerke (now Chemie Linz) are shown in Table 2 [77]. Material requirements per tonne of clinker or sulfuric acid are as follows:

Gypsum (as calcined anhydrite)	1.6 – 1.7 t
Clay	0.2 – 0.7 t
Sand	0.1 t
Fuel oil	$11.7 - 13.8 \times 10^6$ kJ

Table 2. Material analyses and consumption figures for the OSW – Krupp calcium sulfate decomposition process [77]

Constituent	wt % in raw meal	wt % in clinker
SiO_2	9.33	19.35
Al_2O_3	2.8	5.81
Fe_2O_3	0.93	1.94
CaO	30.45	63.19
SO_3	42.86	2.00
F	0.15	0.30
P_2O_5	0.50	1.00
C	4.50	0
H_2O	5.26	0

The gypsum – sulfuric acid process was developed by MÜLLER and KÜHNE in Leverkusen, Germany, during World War I, from 1915 to 1917 [76] as a means of making sulfuric acid from indigenous raw materials independent of imports. It has remained of interest in countries with natural gypsum resources where, for strategic or foreign exchange reasons, it is desirable to avoid dependence on imported sulfur. Furthermore, a Müller – Kühne plant operating on recycled phosphogypsum is, in theory, an ideal means of avoiding waste disposal problems in phosphoric acid production. Nowadays, though, its very high capital costs, as well as the cost of its very substantial energy requirement, have made the economics of the Müller – Kühne process so unfavorable as to preclude its use in any but a handful of special cases. The capital costs for a complete gypsum – sulfuric acid plant are about five times as high as those of a sulfuric acid plant based on sulfur combustion. And while the gypsum-based plant is a substantial net energy consumer, a sulfur-based plant produces a substantial energy credit.

The first gypsum – sulfuric acid plant (40 t/d H_2SO_4) operated in Leverkusen from 1916 to 1931. Other plants running on natural gypsum or anhydrite were built in the United Kingdom (Billingham, Widnes, Whitehaven), France (Miramas), Germany (Wolfen, Coswig), Poland, and Austria (Linz) [77]. The French plant shut down as early as 1952 and the plants in Great Britain all closed down in 1972 – 1975. The plant in Linz (230 t/d H_2SO_4) was modified in 1969 to run on phosphogypsum. In 1972, a plant with a capacity of 350 t/d H_2SO_4, processing phosphogypsum according to the OSW – Krupp process, started up at Phalaborwa in South Africa.

Iron(II) Sulfate. Iron(II) sulfate can be decomposed into sulfur dioxide and iron oxide in fluid bed furnaces, multiple hearth furnaces, or rotary kilns [72]. Iron(II) sulfate is obtained in large quantities as its heptahydrate (green salt) during the regeneration of pickling liquors from the metals industry and in the production of titanium dioxide pigment by the sulfate process, [68, 73] which produces 3 – 4 t of green salt for every tonne of titanium dioxide. In 1978, titanium dioxide production capacity in the Western Hemisphere based on the sulfate process amounted to ca. 1.6×10^6 t/a [83]. Disposing of

the enormous amount of co-produced green salt is a major problem. Decomposition is one of the few remaining options for keeping this largely unwanted material from entering the environment.

Iron(II) sulfate heptahydrate is normally separated from sulfuric acid solutions by crystallization. In most cases, it is subjected to preliminary heat treatment to drive off the majority of the water of crystallization before it is decomposed [74], since the direct decomposition of the heptahydrate is energetically unfavorable. In this first treatment step, the heptahydrate is dehydrated at 130 – 200 °C by flue gases in spray dryers or fluid bed dryers to a monohydrate or mixed hydrate of the empirical formula $FeSO_4 \cdot 1.5\,H_2O$ [84]:

$$FeSO_4 \cdot 7\,H_2O \longrightarrow FeSO_4 \cdot 1.5\,H_2O + 5.5\,H_2O \quad \Delta H = 298\,\text{kJ/mol} \tag{27}$$

In the second step, the sesquihydrate is decomposed at ca. 900 °C

$$FeSO_4 \cdot 1.5\,H_2O \longrightarrow 0.5\,Fe_2O_3 + SO_2 + 1.5\,H_2O + 0.25\,O_2 \quad \Delta H = 305\,\text{kJ/mol} \tag{28}$$

Formerly, decomposition was conducted in sintering machines or rotary kilns with coke as reductant [73], and the gases thus obtained contained ca. 7 vol % of sulfur dioxide. Today it is common practice for iron(II) sulfate to be decomposed in a fluid bed pyrite roasting furnace at 850 °C or more. Elemental sulfur, coal, or fuel oil may be used as supplementary fuels [84]. The sulfate is fed in above the fluid bed or iron(III) oxide. The sulfur dioxide-containing gas leaving the furnace is cooled in a waste heat boiler to 350 – 400 °C and is subsequently passed on to the gas cleaning system. The cleaned gases are fed to the sulfuric acid plant. The mixture of metal sulfates and sulfuric acid resulting from the concentration of acid wastes of titanium dioxide production can similarly be processed in a fluid bed reactor [85]. In individual cases, iron(II) sulfate is also decomposed in multiple-hearth furnaces with flue gases from fuel oil or natural gas combustion [72, 85, 86].

5.7. Production from Waste Gases

5.7.1. Combustion of Hydrogen Sulfide

The main source of hydrogen sulfide-containing gases is fossil fuel upgrading. Large-volume gas streams with high hydrogen sulfide contents such as those produced in plants for natural gas purification, in crude oil refineries and in plants for the liquefaction or gasification of solid fuels are in most cases treated by the Claus process to convert the hydrogen sulfide to elemental sulfur (see also → Sulfur, Chap. 6.). With small-volume gas streams such as coke plant off-gases the hydrogen sulfide is burnt to sulfur dioxide, which is converted to sulfuric acid in a wet contact plant. Gas streams which cannot be processed by the Claus process because they contain too little hydrogen sulfide or too high a concentration of gaseous constituents that interface with the Claus process are usually incinerated to oxidize the hydrogen sulfide to sulfur dioxide. Even extremely dilute gas streams, if they contain any hydrogen sulfide or other sulfur-containing compounds such as carbonyl sulfide, carbon disulfide, and thiols, cannot be discharged to the atmosphere because of their high toxicity. The maximum allowable concentration of hydrogen sulfide that can be emitted in Germany, for example, is 10 mg/m^3. The sulfur dioxide-containing gases formed during incineration are either processed to sulfuric acid by special wet catalysis methods or, at very low sulfur dioxide concentrations, are emitted directly.

The hydrogen sulfide-containing gases are burnt with an excess of air:

$$2\,H_2S + 3\,O_2 \longrightarrow 2\,SO_2 + 2\,H_2O \quad \Delta H = -518\,\text{kJ/mol} \tag{29}$$

Depending on the calorific value of the gases and their content of air, the combustion temperature may range from 500 to 1000 °C. The calorific value is a function of the concentrations and the heats of combustion of the hydrogen sulfide and of the oxidizable accompanying compounds such as carbonyl sulfide, carbon disulfide, carbon monoxide, ammonia, hydrogen cyanide, hydrogen, and organic sulfur compounds. In cases where the combustion temperature required for complete oxidation of the noxious substances

cannot be reached, a pilot flame or supplementary natural gas or fuel oil firing must be provided. Gases with a hydrogen sulfide content of 10 vol % or more or with a corresponding calorific value can normally sustain combustion without supplementary firing.

Gases with fluctuating and, especially, extremely low hydrogen sulfide contents cannot be oxidized completely without supplementary heating. For them it is preferable to use catalytic incineration. Special catalysts are available which completely convert all hydrogen sulfide at initiation temperatures of ca. 200 °C. The essential advantage of catalytic incineration over thermal incineration of low-grade hydrogen sulfide-containing gases is that it consumes no additional fossil fuel. This in itself is a saving, and another advantage is that the sulfur dioxide-containing gas formed has a lower moisture content and is not diluted by combustion gases. In addition, the catalyst presents a large specific surface area for the reaction so that the hydrogen sulfide can be completely oxidized with a lower oxygen excess. Therefore a smaller volume of oxidized gas with a correspondingly higher sulfur dioxide concentration is formed.

The sulfur dioxide concentration is determined by the concentration of hydrogen sulfide and other sulfur compounds and the relative volume of air. Where the oxidized gases are destined for subsequent processing to sulfuric acid, the air necessary to provide the required O_2/SO_2 ratio of 1.1 – 1.3 in the contact gases is introduced directly into the hydrogen sulfide combustion zone. Starting from a gas with 20 vol % hydrogen sulfide, for instance, a contact gas with 6 vol % sulfur dioxide is obtained. The water content of the sulfur dioxide-containing gas is inherently high because the oxidation of hydrogen sulfide produces an equimolar quantity of water, to which is added water from the combustion air and water generated by combustion of any other hydrogen-containing substances present in the gas or added as auxiliary fuel.

The hydrogen sulfide-containing gases are normally burnt in a horizontal cylindrical furnace, lined with several layers of refractory bricks and equipped with special safety devices such as explosion flaps, as mixtures of air with hydrogen sulfide in the concentration range 4.3 – 45.5 vol % H_2S are explosive [87]. The hot gases leaving the furnace are preferably cooled in a fire-tube waste heat boiler to generate steam. It must be ensured that the wall temperatures do not fall below the relatively high dewpoint of the moist gases, which is partly determined by the sulfur trioxide content.

5.7.2. Recovery from Flue Gases

The waste gases produced by combustion of sulfur-containing fossil fuels such as coal or crude oil in power stations ("flue gases") have quite low sulfur dioxide contents. The sulfur dioxide content of flue gas from a plant fired by coal containing 1 % S, for instance, is about 700 mL/m^3. Consequently these gases are inherently unsuitable for processing by methods used for treating metallurgical waste gases, for example. However, on account of the large installed power of such stations the volume of flue gas is such that the emission rate of sulfur dioxide is extremely high in absolute terms and is nowadays environmentally intolerable. A 600 MW power station fired with coal containing 1 % S, for instance, emits about 2×10^6 m^3 flue gas containing ca. 4 t/h of sulfur dioxide.

Most industrial countries have introduced exacting regulations relating to environmental protection in recent years, and a large number of dedicated processes for flue gas desulfurization have been developed, some of which are already in use in the industrial countries on a large scale. In most processes, the sulfur dioxide is ultimately converted to calcium sulfate, which is partly disposed of as a waste material and partly utilized in the construction industry. Apart from these, there are the so-called regenerative processes in which the sulfur dioxide is recovered in a concentrated form from the flue gases ("rich gas") or in the form of sulfuric acid (Desonox-process). Typical regenerative processes are the BF process of Bergbau-Forschung and the Wellman – Lord process of Davy McKee (see page 38).

The *BF process* is a dry flue gas cleaning process based on sulfur dioxide sorption on activated coke with thermal regeneration [88]. A simplified flow diagram of this process is shown in Figure 21. The dedusted flue gases from the electrostatic precipitator of the power station are conveyed at about 130 °C to the sulfur dioxide

adsorber, which operates as a moving-bed reactor. The bed is composed of activated coke and moves downwards in crosscurrent with the flue gas. About 80 – 95 % of the sulfur dioxide is adsorbed on the activated coke; simultaneously, residual flue dust is arrested. The cleaned flue gas is drawn through by means of a fan, which discharges into the stack. In contrast to wet flue gas desulfurization processes, in this case the purified gas does not need reheating.

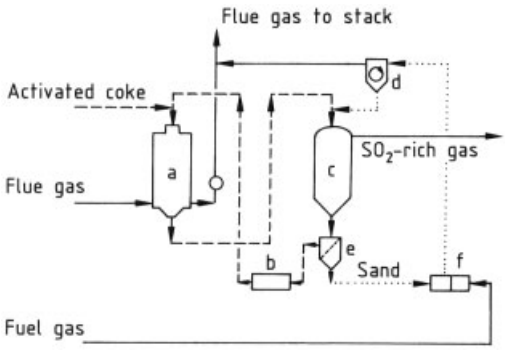

Figure 21. Bergbau-Forschung regenerative flue gas desulfurization process
a) Absorber; b) Coke cooler; c) Regenerator; d) Cyclone; e) Screen; f) Burner

The sulfur dioxide adsorbed on the activated coke is oxidized in the presence of adsorbed moisture by the residual oxygen available in the flue gas. The resulting sulfuric acid is stored in the coke pores. The laden coke is discharged continuously from the adsorber into the regenerator, where it is heated to 600 – 650 °C by the addition of hot sand at ca. 700 °C. The sulfuric acid is decomposed to water vapor and sulfur trioxide. The sulfur trioxide is immediately reduced to sulfur dioxide by the coke, which is itself partly converted to carbon dioxide. Accordingly, the regenerated sulfur dioxide contains substantial amounts of carbon dioxide and moisture.

The regenerator is designed as a moving bed reactor. The mixture of sand and regenerated activated coke discharged at the bottom is separated by means of a vibratory screen. The coke is directly cooled to about 120 – 140 °C by injected water and is recycled to the top of the adsorber. Any carbon losses due to the chemical reaction in the regenerator are compensated by the continuous addition of fresh activated coke. The sand is heated in a fluid bed by combustion of fuel gas and is conveyed pneumatically to the top of the regenerator in the hot flue gases.

Typical analysis values for the rich sulfur dioxide- containing gas discharged from the regenerator are: 25 vol % SO_2, 15 vol % CO_2, 60 vol % H_2O, and traces of N_2, SO_3, HCl, and HF.

The chlorine and fluorine concentrations in the rich gas are about 300 times as high as the concentrations in the original flue gas [88]. Before subsequent processing, the impurities must be removed in a gas cleaning system similar in principle to the type used for cleaning roaster gases (see Section 6.2). The purified and cooled rich gas contains 50 – 60 vol % SO_2 and 30 – 35 vol % CO_2, the remainder consisting of water vapor and nitrogen.

The BF process is suitable for the desulfurization of flue gases with sulfur dioxide concentrations of up to 0.4 vol % [88]. The sulfur dioxide-rich gas produced can be utilized for the production of liquid sulfur dioxide, sulfur or sulfuric acid. In 1974 a large-scale pilot plant using the process was started up in Lünen. It was designed to treat a flue gas volume of about 150 000 m^3/h, the rich gas being processed to elemental sulfur [88].

6. Cleaning of Crude Sulfur Dioxide Gas Streams

6.1. Constituents of Roaster and Metallurgical Gases

The crude sulfur dioxide-containing gases obtained by pyrites roasting, pyrometallurgical nonferrous metal extraction from sulfides, and thermal decomposition of waste sulfuric acid and sulfate must be cooled and cleaned efficiently before being processed to obtain sulfuric acid or liquid sulfur dioxide.

The impurities which may be present can be classified into four categories according to their state of aggregation and the process step of the gas cleaning system in which they are removed [43].

Primary dusts are particles discharged in solid form from the roasting or metallurgical extraction process at the so- called flue dust, purely

mechanically entrained by the waste gases. The proportion and grain size of such impurities depend on the furnace feed and the flow conditions.

Fumes are sublimed substances which are volatilized under the furnace conditions but have resolidified in the form of an aerosol by the time they enter the dedusting system inlet at 300–400 °C. Their grain size is 0.01–2 µm. Frequently they consist of zinc and lead compounds such as zinc oxide and lead sulfate or basic lead sulfate.

Volatiles are substances which are still gaseous when they enter the dry dedusting system but may precipitate in the cold areas of the gas cleaning system. These include arsenic and selenium compounds, mercury, sulfuric acid mists, and hydrocarbons.

Gaseous substances include, for instance, hydrogen chloride, hydrogen fluoride, and lower hydrocarbons. Normally, carbon dioxide and small quantities of carbon monoxide originating from the combustion gases of supplementary heating systems are not harmful.

Whereas the primary dusts and fumes can in most cases be separated mechanically at elevated temperatures in a dry gas cleaning step, the gas must be cooled and passed through a wet cleaning stage to ensure complete separation of the volatile and gaseous impurities.

6.2. Cleaning Systems

The simplified flow diagram of a Lurgi gas cleaning system for sulfur dioxide- containing gases from a fluid bed roasting plant is shown in Figure 22. The same combination of individual process steps is also used for cleaning sulfur dioxide-containing gases produced in other metallurgical processes.

Dry Dedusting. After cooling in a waste heat boiler to 320–400 °C, the furnace waste gases, laden with 50–150 g/m^3 of dust, pass through a cyclone for preliminary separation of the coarse particles. Generally, the collection efficiency is 85–90 % with a pressure drop of about 5–10 mbar.

The majority of the dust remaining in the hot gas is then removed in a hot-gas electrostatic precipitator. The gas passes between two electrodes, one of which (the collector electrode) is grounded and the other (the discharge electrode) is maintained at a sufficiently high negative potential to ionize the surrounding gas molecules, resulting in a so-called corona discharge. Dust particles passing through the corona acquire an electrostatic charge of the same polarity as the discharge electrode and are, therefore, repelled towards the collector electrode, where they are discharged.

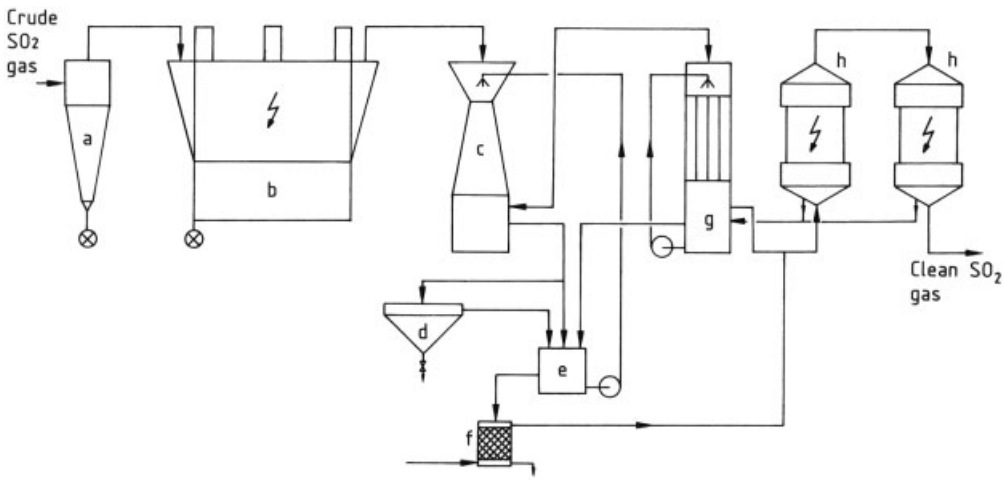

Figure 22. Gas cleaning plant (Lurgi system)
a) Cyclone; b) Hot-gas electrostatic precipitator; c) Venturi scrubber; d) Settling tank; e) Pump tank; f) SO_2 stripper; g) "Star" gas cooler; h) Wet electrostatic precipitator

Because corona discharges form much more readily around charged objects which have a very small surface area, and particularly from pointed objects, the discharge electrode is usually in the form of a wire or needle. The collector electrode, however must have as large a surface area as possible to minimize the chances of any charged particles escaping. Hot-gas precipitators are usually of the plate type, in which the gas passes through the channels between a battery of vertically mounted flat collector electrodes. Rows of discharge electrodes are suspended in the gas spaces between the plates. Dust accumulating on the surface of the collector electrodes falls by gravity into a hopper below the precipitator. The applied d.c. voltage between the discharge and collector electrodes is 10 – 80 kV. The specific power requirements are 1.5 – 2.5 kWh/1000 m^3.

Electrostatic precipitators have the considerable advantage of being able to separate more than 99 % of the dust at a pressure drop of $\leq$ 1 mbar. The residual dust content is typically 20 – 200 mg/m^3.

The size and mode of operation of the hot gas electrostatic precipitator depends not only on the volume of gas but also on its dust loading, the size distribution of the entrained particles and, above all, its resistivity. The gas velocity through the unit must be kept below the value at which significant re-entrainment of the discharged particles occurs. The optimum temperature range for operation is determined by the dust characteristics. By means of adequate insulation and temperature control, especially in the "dead" corners, it can be ensured in practical operation that the temperature does not drop below the dewpoint, which would give rise to corrosion by acid condensation [89].

The separated dust is normally recycled to the roasted material handling system.

Wet Separation. The remaining impurities are removed by scrubbing the gases. Adiabatic water evaporation simultaneously cools the gas to 60 – 80 °C so that the major part of impurities such as selenium, arsenic trioxide, and sulfuric acid condense. Some of the impurities are not washed out in the scrubber but remain in the gas in the form of minute solid or liquid particles. These are subsequently removed in wet electrostatic precipitators (see below). Gaseous impurities such as hydrogen chloride or fluoride are also separated in the scrubber, at least in part.

The scrubbing liquid is normally recirculated, so the separated impurities accumulate in it. Solid impurities can be removed as a slurry by sedimentation in a settling tank. Liquid impurities with a relatively low vapor pressure, such as sulfuric acid, increase the acidity of the scrubbing liquid, hence the term wash acid. Part of the wash acid is continuously bled off to hold the concentration in the circuit at or below a maximum of 30 – 50 % H_2SO_4. It is replaced by fresh water and recirculated condensate from the subsequent wet gas cleaning systems. In some cases it may be necessary to bleed the wash acid at a lower sulfuric acid concentration so as to hold the concentrations of other, more volatile impurities such as hydrogen chloride and hydrogen fluoride down to levels at which they are still efficiently absorbed. Arsenic (as As_2O_3) first condenses in the scrubber and then dissolves in the wash acid. The limit of solubility may be exceeded in the circulating wash acid, and crystalline arsenic trioxide can be separated from the supersaturated solution in settling tanks. The discharged wash acid is normally neutralized after air stripping to remove sulfur dioxide.

The scrubbing and adiabatic cooling procedure is usually performed in an unpacked scrubber tower, a Venturi scrubber, a radial flow scrubber, or a combined scrubbing and cooling tower of the Peabody type.

The moisture content of the gases at the temperature at which they leave the scrubber (60 – 80 °C) is too high for production of concentrated sulfuric acid in a contact plant. The gas therefore must be cooled to reduce its moisture content.

Various systems can be used for gas cooling. One method is to pass it through an additional, packed scrubbing tower in which it is cooled by circulating scrubbing liquid. The warm liquid leaving the tower is cooled in a scrubbing acid cooler and is then pumped back to the top of the tower. In another method, the gas is cooled to the requisite gas temperature (30 – 40 °C) in indirect gas coolers (star coolers). The resulting condensate is passed on to the scrubber where it is used for balancing the evaporation losses that occur there.

When the crude gas comes from the roasting or pyrometallurgical treatment of an ore concentrate with a high fluorine content, it may also be

necessary to install a so- called fluorine tower to remove any hydrofluoric acid still present after scrubbing and gas cooling. This tower is packed with a material with a high content of silica, which reacts with hydrogen fluoride to form fluosilicic acid.

The solid and liquid impurities still present in the gas after scrubbing are removed almost completely in wet electrostatic precipitators.

To ensure that no unpurified gas reaches the contact plant, even in case of a wet precipitator failure, two precipitator steps are normally installed in series (primary and secondary cleaning). The two stages together have a collection efficiency of $> 99\,\%$. At the outlet the gas is optically clear when viewed against a light source.

Two patterns of wet precipitator are in use. The older plate type is made of lead. The gas passes along the space between a battery of extremely smooth vertical plates which are the collecting electrodes. The discharge electrodes are suspended between the collector electrodes, from which they are isolated.

Nowadays, though, the plate type of precipitator has practically been replaced by the tube type, in which the collector electrodes are tubular and the discharge electrodes are suspended concentrially inside them. In these precipitators the tubes are made of plastics: the requisite conductivity is attained by preparing the inside of the tube so that the separated acid mist will form a uniform liquid film flowing downwards. In another version, the collection tubes are replaced by a plastic honey comb element.

For the best possible separation effect, the voltage must be maintained – as in the dry precipitators – just below the spark-over voltage of the gas, which may fluctuate, depending on the process. An automatic voltage control system is used for this purpose.

A narrow negative electrode is required only for establishing the corona discharge. The subsequent repulsion of charged dust particles away from the negative electrode is, in fact, more efficient if the electrode diameter is larger and there is no discharge. Some tubular precipitators are therefore designed on a two-stage principle, in which a relatively short corona discharge needle is mounted on the end of a wider, nondischarging electrode rod. There are thus discrete ionization and collection zones [90].

The wet precipitator casing may also be made from plastic, but steel casings with internal rubber-linings or homogeneous interior lead coatings are used as well.

The gas-side pressure drop of such electrostatic precipitators is ca. 2 mbar. The high-tension set requires about 0.8 kWh per 1000 m^3 gas.

The gases at the gas cleaning system outlet must be optically clear (checked by means of a light source) to guard against problems in subsequent processing. Any residual moist mists are difficult to separate, even in the dryer of a sulfuric acid plant, so they may cause substantial operational and corrosion problems.

Mercury Removal. Any mercury contained in the concentrates is volatilized almost completely, passing on to the scrubbing plant as mercury vapor. Depending on the partial pressure, metallic mercury may occur in the scrubbing system, but part of it always remains as vapor in the gas.

As the volatilized mercury still contained in the SO_2 gas will end up in the sulfuric acid production, special gas treatment has to be provided for the removal of the mercury in order to obtain acceptable mercury contents in the acid. Alternatively, if the mercury vapor remains in the gas, the sulfuric acid produced has to be treated to eliminate the mercury.

Both wet and dry methods have been used to remove mercury from roaster gas [91]. Dry methods include filtration through a porous material impregnated with selenium, with which the mercury reacts to form mercury(II) selenide, and activated carbon adsorption. Wet methods include introducing elemental selenium into the wet scrubber of the gas cleaning plant and washing with mercury(II) chloride solution to convert the mercury to insoluble mercury(I) chloride (removed from the scrubbing liquor by filtration). Scrubbing with concentrated sulfuric acid removes mercury in the form of mercuric sulfate.

If mercury is not removed from sulfuric acid plant feed gas, it is absorbed in the drying tower and therefore appears in the product acid. Though a certain amount may be precipitated in the form of mercury(II) sulfate, which can be filtered out, for some purposes it is necessary to purify the acid to a higher standard. One method [91] is to treat the acid in the drying circuit with

thiosulfate, which decomposes to give colloidal sulfur. This in turn reacts with mercury to give insoluble mercury(II) sulfide. Because this technique can only be used when the acid concentration is below 85 %, a preliminary drying tower is installed operating on acid of 70 – 85 % concentration. The main drying tower operates on full-strength product acid.

7. Production of Pure Sulfur Dioxide

One way of producing pure sulfur dioxide is to chemically reduce sulfur trioxide with sulfur. However, since sulfur trioxide is itself usually produced in a sulfuric acid plant processing moderately concentrated sulfur dioxide-containing gases, physical or chemical separation may offer a more direct and attractive route to the pure product. From an economic point of view, the higher the sulfur dioxide concentration is in the raw gas from which it is to be separated the better. Therefore, gases produced by combustion of sulfur or in pyrometallurgical processing of sulfides ores are preferred. But it is possible, by means of regenerative flue gas scrubbing methods, to produce pure sulfur dioxide from low-concentration sources such as power station combustion gases. In view of the very large volumes of these gases that are produced and the current need in many locations to remove their sulfur dioxide content for environmental reasons, this method cannot always be dismissed on simple economic grounds alone.

7.1. Chemical Production from Sulfur Trioxide

Pure sulfur dioxide is formed when sulfur and sulfur trioxide undergo the following weakly exothermic reaction [92]:

$$S + 2\,SO_3 \longrightarrow 3\,SO_2 \quad \Delta H = -99\,\text{kJ/mol} \tag{30}$$

Practical processes using this reaction are distinguished by the form in which the reactants are converted.

Despite its age, the process developed in Germany by *IG Farbenindustrie* [93] is still in use today. Solid sulfur is fed into a reactor containing oleum at ca. 110 °C.

Simultaneously, preheated oleum with 26 – 30 % free SO_3 is fed in at the bottom of the reactor. The oleum rises in countercurrent to the slowly descending sulfur, leaving the reactor at the top with a residual free SO_3 content of 1 – 5 wt %. The sulfur dioxide gas formed in the reactor still contains 5 – 10 vol % SO_3, which is converted to further sulfur dioxide in a column filled with lump sulfur. The last traces of sulfur trioxide are finally removed from the sulfur dioxide by scrubbing with 98 % sulfuric acid. The dilute oleum leaving the reactor flows to the oleum tower of the sulfuric acid plant for reconcentration. As it contains impurities introduced in the sulfur, it must be discharged partially and replaced by fresh oleum.

In modern processes, the sulfur trioxide is not fed as concentrated oleum but in dilute liquid form, and the sulfur is preferably molten. The reaction medium is either oleum of extremely low SO_3 content [94], a liquid mixture of sulfur dioxide and sulfur trioxide under pressure [95], liquid sulfur [96] or oleum of 25 % strength [95].

Figure 23 (see next page) shows the flow diagram of the *Schweizerhall* process [97]. Molten sulfur and liquid sulfur trioxide in stoichiometric proportions are agitated in a reactor with oleum of ca. 25 % strength. The reactor bottom is equipped with a double shell, through which cooling water is pumped to provide temperature control. In normal operation, only slight cooling is required. For start-up, steam heating is necessary. The reaction temperature is adjusted to maintain a sufficiently high pressure of the sulfur dioxide formed that it liquefies in a water- cooled condenser without further refrigeration. The sulfur dioxide leaving the reactor is passed through a secondary reactor filled with solid sulfur, in which the residual sulfur trioxide is converted completely to sulfur dioxide, before entering the condenser. The sulfur dioxide produced by this process is of high purity.

The principal advantages of this process are its relatively low capital cost and the simplicity of operation. The sulfur dioxide production costs are essentially determined by the sulfur price and (especially) the cost of liquid sulfur trioxide, which are normally very high. For 1t of liquid sulfur dioxide, ca. 20 kWh of electric power and around 30 m^3 of cooling water at 20 °C are required.

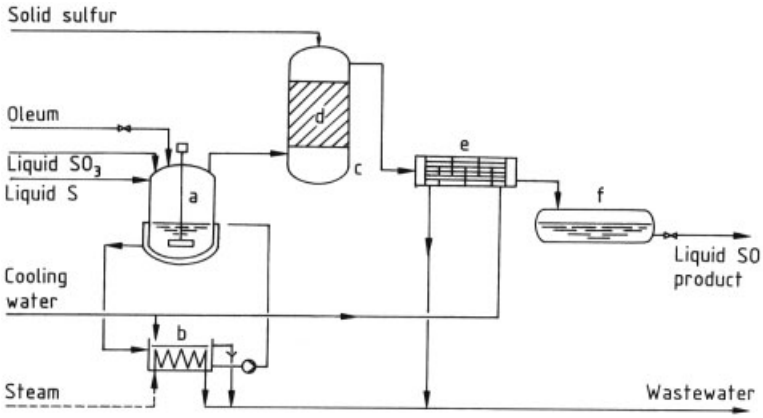

Figure 23. Schweizerhall liquid sulfur dioxid process
a) Reactor; b) Thermostat; c) Post-reactor; d) Lump sulfur; e) SO_2 condenser; f) Tank for liquefied SO_2

7.2. Separation from Sulfur Dioxide-Containing Gases

Condensation processes for the production of pure sulfur dioxide are based on the separation of sulfur dioxide from gases by liquefaction. In accordance with the vapor pressure equilibrium between gaseous and liquid sulfur dioxide (see 2), low temperatures, or high pressures, or both are required for this purpose.

In a plant operated by *Bayer*, roaster gases with 6 – 7 vol % of sulfur dioxide are compressed to ca. 5 bar and then passed through a number of countercurrent coolers. These are cooled by liquid ammonia and by the waste gas of the process, the temperature of which has been reduced to − 70 °C by stepwise flashing. The sulfur dioxide content of the waste gas is ca. 0.7 vol % [98]. The energy consumption per tonne of liquid sulfur dioxide of a plant designed for a production of 25 – 30 t/d of liquid sulfur dioxide is ca. 680 kWh.

The so-called *partial condensation* process is quite commonly used. Normally, 30 – 60 % of the fed sulfur dioxide is condensed. Partial condensation plants are frequently operated in connection with sulfuric acid plants, as the residual sulfur dioxide content of the depleted gas is in most cases still relatively high, and so it requires subsequent treatment.

A simplified diagram of this process is shown in Figure 24. The sulfur dioxide is partially condensed under normal pressure by cooling in indirect exchange with an evaporating coolant. The ratio of the volume of liquid sulfur dioxide produced to the residual sulfur dioxide content in the gases is determined by the sulfur dioxide vapor pressure at the condensation temperature, which is controlled in the coolant circuit.

Dry, sulfur dioxide-rich gases are pre-cooled by the cold residual gases in a heat exchanger and passed on to the condenser. The temperature in the sulfur dioxide condenser is held between − 70 and − 40 °C by evaporation of the coolant in the shell. The depleted gases leaving the condenser are passed on to a separator, which removes entrained droplets of liquid sulfur dioxide, and are then conveyed to a sulfuric acid plant or other plant for processing sulfur dioxide after reheating in the gas – gas heat exchanger.

Fluorinated hydrocarbons are mostly used as coolants. The coolant vapor leaving the sulfur dioxide condenser is compressed to the pressure required to restore it to the liquid state after cooling; the exact pressure therefore depends on the temperature of the available cooling water. The liquefied coolant is expanded in several stages to the pressure corresponding to the desired evaporation temperature before being recycled to the sulfur dioxide condenser.

The economics of the partial condensation process are essentially determined by the sulfur dioxide content of the gases; 16 – 18 vol % is the ideal concentration. In a partial condensation plant with 12 vol % SO_2 in the inlet gas and 6 vol % SO_2 in the residual gas, ca. 260 kWh of power and 90 m^3 of cooling water at 20 °C are consumed per tonne of sulfur dioxide liquefied [97].

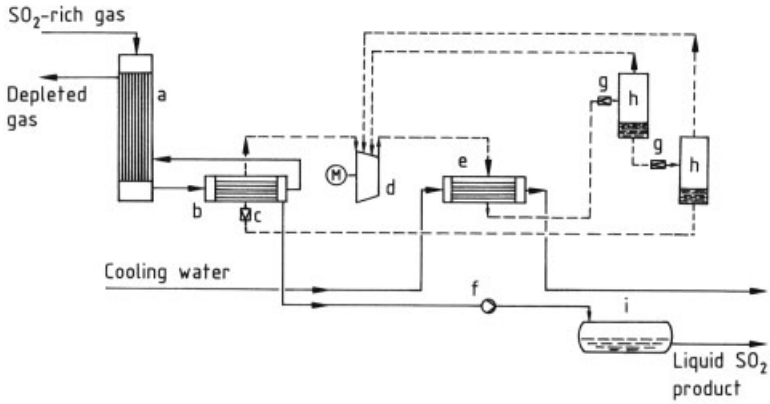

Figure 24. Liquid sulfur dioxide production by partial condensation
a) Gas heat exchanger; b) SO_2 condenser; c) Expansion valve; d) Coolant compressor; e) Coolant condenser; f) Pump; g) Expansion valve; h) Coolant expansion vessel; i) Liquid SO_2 tank

On account of its economy and operational reliability, the partial condensation process is one of the most frequently used methods of recovering pure sulfur dioxide. It is especially suited for large production capacities. The largest plant built by Lurgi so far, for example, is designed for the production of 160 t/d of liquid sulfur dioxide.

Absorption Processes. In absorption processes, sulfur dioxide is separated from the SO_2-containing gases by reversible absorption into a liquid. This may either be a solvent, in which it dissolves physically, or a scrubbing solution containing a carefully selected active component, to which it becomes loosely and reversibly chemically bound. It is then regenerated in concentrated form by subjecting the rich solution to conditions of temperature and pressure under which the sulfur dioxide is desorbed.

Absorption in Water. One of the oldest absorption processes is atmospheric water washing according to the *Hänisch – Schröder process* of 1884. It is still used today in a modified form, but usually only to recover a part of the sulfur dioxide content of the gas being processed; therefore, it must be associated with some ancillary process to treat the residual gases. For satisfactory operation the crude gas should contain at least 10 – 11 vol % SO_2.

The sulfur dioxide-containing gas is fed into the base of a packed tower and rises in countercurrent with water introduced at the top. The water leaving the bottom of the tower may contain 1 – 2 wt % of sulfur dioxide, depending on the temperature. It is passed to a desorption tower, in which the sulfur dioxide is driven off by low-pressure steam. The water at the exit of the desorber still contains ca. 0.01 wt % of sulfur dioxide. The recovered mixture of sulfur dioxide and water vapor is passed through a water-cooled condenser to remove as much water as possible, after which the sulfur dioxide is dried with sulfuric acid.

The process has relatively high utility requirements: for partial absorption, it uses about 120 m^3 of process and cooling water, 3 t of steam and 90 kWh of power per tonne of liquid sulfur dioxide [97].

A pressurized sulfur dioxide water wash (the *Grillo system*) is schematically represented in Figure 25 (see next page). Purified gas with at least 9 vol % SO_2 is compressed to ca. 5 bar in a compressor. To dissipate the heat of compression, the gas is cooled to about 50 °C in heat exchangers. It is then scrubbed in a water-irrigated packed tower, in which its sulfur dioxide content is reduced to < 0.1 vol %. The residual gas is heated against the compressed fresh gas and subsequently expanded to recover up to 35 % of the compressor energy.

The sulfur dioxide-laden wash water extracted from the absorption tower is first preheated in a liquid heat exchanger and then passed to a desorption column, in which the sulfur dioxide is stripped by injection of low-pressure

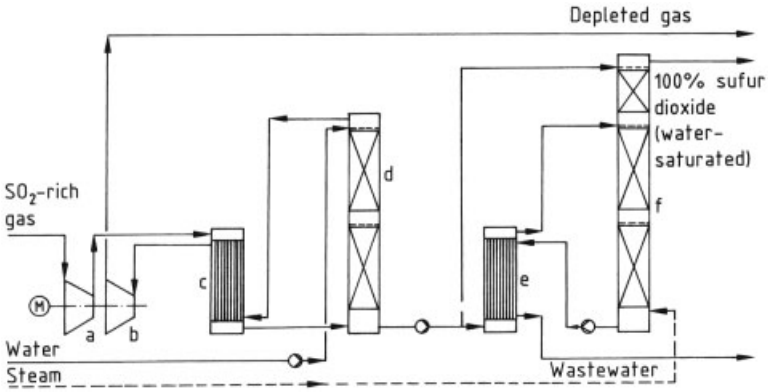

Figure 25. Production of pure sulfur dioxide by pressurized water absorption (Grillo system)
a) Compressor; b) Expansion turbine; c) Gas heat exchanger; d) Absorber; e) Liquid heat exchanger; f) SO_2 stripper

steam. The column is connected at the top with a dephlegmator in which the water vapor is disengaged from the sulfur dioxide by irrigation with a partial stream of cold, sulfur dioxide-laden water from the absorption tower. In some cases, instead of the dephlegmator, a separate condensation tower irrigated with water from an indirectly cooled circuit is used. The cooled sulfur dioxide is dried in the same way as in the atmospheric water wash. The wastewater leaving the desorption column sump is used to indirectly preheat the incoming rich scrubbing solution before being discharged.

Even though part of the compression energy is recovered in the expansion unit, the consumption of electric power is extremely high in comparison with the atmospheric water wash. Production of 1 t of liquid sulfur dioxide from a raw gas containing 10 vol % of sulfur dioxide consumes ca. 300 kWh [97]. Furthermore, ca. 1 t of low-pressure steam and ca. 50 m^3 of processed cooling water are required as well.

Both the pressurized and atmospheric water wash processes have the advantage of being able to run on almost any quality of water (except water containing volatile materials).

Absorption in Alkaline Solutions. Alkaline solutions absorb much more sulfur dioxide than water. In some instances strong bases such as caustic soda and ammonia are used, but to an increasing extent these are being replaced by alkaline salts of weak inorganic and organic acids (e.g., sodium carbonate or citrates), or (occasionally) aliphatic or aromatic amines. This is because regenerating the sulfur dioxide is easier and less energy-intensive with these weaker bases. The principle of the process corresponds essentially to that of the atmospheric water wash, except that the absorbing solution is always recirculated from the desorber to the absorber.

In the *Wellman – Lord process* [99, 100], the sulfur dioxide is extracted by means of an aqueous solution of sodium sulfite and sodium hydrogen sulfite. The absorbent is regenerated, and concentrated sulfur dioxide is released by indirect heating with steam in a separate step, which thermally decomposes the sodium bisulfide. The process is chiefly used for desulfurizing flue gases from power stations and tail gases from Claus sulfur plants (see → Sulfur, Chap. 6.), but in a few cases it has also been used for purifying the waste gas from contact (i.e., single-absorption) sulfuric acid plants.

Figure 26 shows a flow diagram of a typical Wellman – Lord plant for flue gas desulfurization. Flue gas which has already been passed through dry electrostatic precipitators or other dust control apparatus is sprayed with water in a saturator section at the bottom of the absorber. This not only cools the gas adiabatically by evaporation of some of the water but also removes any remaining particulates. A bleed stream of the scrubbing water is extracted from the circuit so as to keep the solids content in the circulating liquor down to ca. 5 % and is fed to the wastewater-treatment system. The gas next enters the main absorption zone and is washed in countercurrent with a saturated aqueous solution of sodium sulfite and sodium bisulfite, which is recirculated in several steps. At the top of the

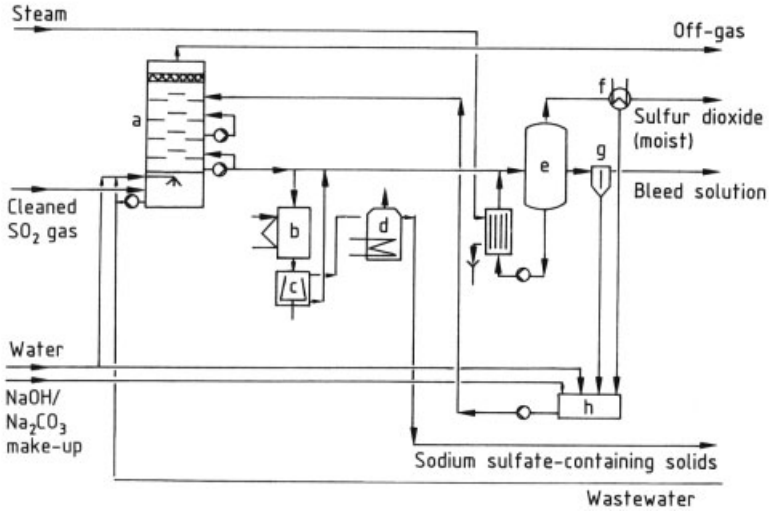

Figure 26. Wellman – Lord process for sulfur dioxide production
a) Absorber; b) Crystallizer; c) Centrifuge; d) Dryer; e) Forced- circulation evaporator/crystallizer; f) Condenser; g) Thickener; h) Dissolution tank

absorber, a droplet separator arrests entrained droplets. After reheating to about 140 °C, the moisture-laden flue gas is discharged to the stacks.

The principal mechanism by which sulfur dioxide is absorbed is its interaction with sodium sulfite to form sodium hydrogensulfite

$$SO_2 + Na_2SO_3 + H_2O \rightleftharpoons 2\,NaHSO_3 \qquad (31)$$

In a side reaction, part of the sulfite is oxidized to sulfate by residual oxygen present in the flue gas. Sulfate is also formed to a lesser extent by reaction of the scrubbing solution with traces of sulfur trioxide in the flue gas.

The hydrogensulfite-rich scrubbing liquid is regenerated in a forced-circulation evaporator, indirectly heated by steam. There, the hydrogensulfite is thermally decomposed by the reverse of the absorption reaction (Eq. 31). Some of the moisture content of the saturated sulfur dioxide liberated in the evaporator is removed by indirect cooling with water in the subsequent condenser.

Because the solution becomes oversatured in the evaporator, sodium sulfite is precipitated in crystalline form. The suspension of crystals passes to a thickener, the overflow from which is withdrawn as waste liquor. The slurry of crystals from the underflow is diluted in a tank with condensate and fresh water to extent necessary to turn it back into a saturated sulfite – bisulfite solution. To compensate for bleed losses, sodium hydroxide or carbonate is added.

The sulfates formed in the side reactions mentioned above are not decomposed in the evaporator. To prevent their concentration rising to values at which the absorptive capacity of the solution is impaired and there is a risk of clogging by accumulations of precipitated solids, part of the scrubbing solution is continuously withdrawn from the circuit [99]. The water solubility of the sodium salts in the scrubbing system decreases in the order bisulfite, sulfite, sulfate. The bleed stream is therefore taken off directly after the sulfur dioxide absorber, where the bisulfite concentration is the highest and the sulfite content is the lowest, and is passed through a water-cooled crystallizer, in which crystals consisting mainly of sodium sulfate are precipitated. The sulfate-rich solid is separated in a centrifuge from the residual solution, which is returned to the absorption circuit. The separated solid is dried by indirect steam heating and discharged from the system for further processing [99].

Other Absorption Processes. Sulfur dioxide absorption processes using aqueous amine solutions have the advantage over processes using sodium sulfite – hydrogensulfite solutions that the amine solutions can absorb more sulfur dioxide and the rich solutions can be re-

generated more easily. That is why, in former times, aliphatic and also aromatic amines were used preferably. However, amines are disadvantageous in respect of their vapor pressure, which is high enough for appreciable amounts to be discharged from the absorption system in the cleaned waste gas. Such emissions are today considered detrimental to health, quite apart from having an objectionable odor.

The *Sulfidin process*, commonly used in the past, employed an absorbent consisting of an aqueous solution of xylidine and toluidine. It is no longer used for environmental reasons.

In the United States, a process developed by *American Smelting and Refining Co.*, which uses *N,N*-dimethylaniline, is still used [101].

The *UCAP process* of Union Carbide Corporation was specifically developed for cleaning the tail gases from Claus plants, to which the pure sulfur dioxide is recycled. The absorbent used is an aqueous solution of triethanolamine [102].

During the development of flue gas and exhaust air cleaning processes physical absorption processes have been introduced. An example of such a process for the removal of SO_2 from flue gases or other SO_2-containing gas steams is the *Solinox process* introduced by Linde, in which poly(ethylene glycol) dimethyl ethers are used as absorbent. The SO_2 is recovered in form of a 80 – 90 % SO_2 gas [103].

8. Construction Materials

Dry gases containing sulfur dioxide – and even those containing water vapor as well, if at a temperature safely above the dewpoint – barely attack the steel grades used normally.

Standard steel St 37.2 (material no. 1.0114) and boiler plate HII (material no. 1.0425) are generally preferred. For extreme stress conditions, St 35.8 (material no. 1.1035) or low-alloy steel grades such as St 45.8 (material no. 1.0405) and 15 Mo 3 (material no. 1.5413) are used. Above 550 °C, heat-resistant ferritic (e.g., material no. 1.4713) or austenitic steels (e.g., material no. 1.7841 and 1.7828) are used. On account of their sensitivity to possible traces of sulfur at the furnace outlet, nickel-rich steel grades are not appropriate, especially at elevated temperatures.

For moist sulfur dioxide-bearing gases, especially in gas cleaning plants, lead was the classical material. It can be used at temperatures up to 110 °C and is also resistant to sulfuric acid at concentrations up to ca. 80 %. Lead (DIN 1719) is mainly used in the form of refined copper lead (Pb 99.9 %, Cu 0.05 – 0.08 %; material no. 2.3040) as sheet, castings, or homogeneous lead coatings on steel.

Where there is likely to be thermal cycling stress, antimony-alloyed lead (DIN 28 058) is preferable. The limited creep resistance of lead and lead alloys must be compensated by adequate support. In homogeneous lead coatings, the adhesion of the lead layer is frequently enhanced by initially tinning the steel surface before the lead coating is applied.

Nowadays plastics are normally applied for piping and ducting and for certain other equipment for wet SO_2 gas at moderate temperatures. Plastics such as polyvinyl chloride (PVC), rigid polyethylene (rigid PE) and polypropylene (PP) are chemically resistant to sulfur dioxide, dilute sulfuric acid, and other inorganic compounds which may occur as impurities in the sulfur dioxide gas cleaning system. But with plastics, factors like ageing by sunlight, cold embrittlement, and temperature cycling have to be taken into account. The upper temperature limit for practical application is ca. 60 °C for PVC, ca. 80 °C for rigid PE, and ca. 90 °C for PP, but for composites of these materials with fiberglass-reinforced plastic (FRP), the upper temperature limits are somewhat higher: for FRP/rigid PE, 100 °C, and for FRP/PP, ca. 120 °C. Fluorinated plastics (with or without reinforcement) can be used up to ca. 120 °C.

For wash acids containing sulfur dioxide, lead, rigid polyethylene, and polypropylene are used for piping, pumps, pump tanks, valves, and fittings. With lead, pitting corrosion is likely to occur at higher chloride or mercury contents.

For sulfur dioxide-saturated, dilute wash acids, plastics and silicon-containing cast iron are most often used for valves, fittings, and pumps.

Commercial acid-resistant bricks are used for lining Venturi scrubbers, scrubbing towers, and vessels. For highly fluoride-laden systems, carbon bricks are used.

Impregnated graphite, especially graphite impregnated with phenolic resin, is normally

limited to special applications in heat exchangers, valves, fittings, and pumps.

Special stainless steels are suitable for use in scrubbers, heat exchangers, etc. provided well-defined and stable conditions are maintained. This refers specifically to impurities such as HCl, HF and reducing components, which may disturb the passivating layer on the steel.

9. Industrial Output and Uses

Production for Sulfuric Acid Manufacture. The majority of the sulfur dioxide produced industrially in the world is directly used for sulfuric acid production. Sulfur dioxide requirements for other purposes are relatively small, and they are usually satisfied either by combustion of sulfur at the consumer's works or by pure sulfur dioxide supplied in liquid form. In 1979, for instance, in the Federal Republic of Germany, the consumption for sulfuric acid manufacture was about 3.3×10^6 t SO_2, whereas other uses accounted for only about 40 000 t [104]. In the same year the United States used ca. 23.6×10^6 t of sulfur dioxide for sulfuric acid production and only about 150 000 t for other purposes [105].

Miscellaneous Uses. Apart from sulfuric acid manufacture, sulfur dioxide has uses in the following industrial areas [106, 107].

Paper and pulp industry (→ Paper and Pulp, Chap. 1.3.2.)
Production of cooking solutions for the sulfite pulping process
Reduction of sodium chlorate in the production of chlorine dioxide for pulp bleaching
Pulp treatment after the first (peroxide) stage of a two-stage bleaching process to eliminate hydrogen peroxide before hydrogensulfite treatment
Chlorine scavenger, especially in sulfate pulp factories after the chlorine bleaching step

Chemical and pharmaceutical industries
Production of sulfites, hydrogensulfites, disulfites, thiosulfates, dithionites, and hydroxyalkanesulfinic acid salts
Production of alkane sulfonates by sulfoxidation of sulfochlorination
Production of gelatins and glues
Reductant or solvent
Chemical processing of manganese ores and recovery of, for instance, Co, Se, and Te from waste

Petrochemical and mineral oil industries
Solvent for separation of aromatics from hydrocarbon mixtures (Edeleanu process)
Removal of sulfur-containing impurities from mineral oils
Deodorization of lubricating greases

Textile and leather industries
Bleaching agent

Food and beverage industries
Preserving agent and disinfectant (e.g. sterilization in breweries, fermentation control, and preservation in viniculture)

Sulfur dioxide is also required for water and wastewater treatment. Liquid sulfur dioxide was formerly utilized for refrigerating machines and plants. However, in this area of application it has been replaced almost completely by fluorinated hydrocarbon refrigerants.

10. Analysis

Analytical techniques for sulfur dioxide are covered in literature reviews [108–110].

Qualitative Detection. Sulfur dioxide can be recognized even in extreme dilution by its pungent smell. Solutions of sulfur dioxide, sulfites, or hydrogensulfites decolorize iodine and permanganate on account of their reducing effect. Sulfur dioxide and sulfites are reduced by zinc in a hydrochloric acid solution to hydrogen sulfide, which is readily detected by its smell.

Quantitative Analysis. Sulfur dioxide is determined chemically by iodometry, titrimetry, gravimetry, or colorimetry. On account of the large number of quantitative legal restrictions relating to sulfur dioxide that are now in force, accuracy and reliability are a prerequisite of any analytical method, and definitive procedures and equipment are often laid down by official bodies. In Germany, for example, they are governed by VDI regulations [111–115].

Continuous Monitoring. Routine in-plant performance monitoring usually entails contin-

uous measurement of the sulfur dioxide content of, for instance, roaster gases, contact gases, and tail gases in a sulfuric acid plant by automatic recording analysers, based on a physical property such as spectroscopic absorption in the infrared [114, 116], or ultraviolet [117], or electrical conductivity [115]. Strict adherence to set sampling procedures is essential for accuracy and reliability of in-plant monitoring.

Iodometric Method. REICH's classical, manual iodometric determination is still widely used in industrial practice. A partial stream of the sulfur dioxide- containing gas is drawn through an iodine solution with starch as an indicator until it is discolored. The sulfur dioxide concentration is calculated from the gas volume and the amount of iodine originally present [108]. In another version of the method [111], an excess of iodine is used, and a predetermined volume of the gas is drawn through the solution. The unused iodine is then determined by titration with sodium thiosulfate after addition of starch. Nitrogen oxides interfere with readings from the Reich method; this disturbing influence can, however, be avoided by addition of acetate [108] or aminosulfonic acid [118].

Titrimetric and Gravimetric Methods. For titrimetric or gravimetric determination of sulfur dioxide, the gases to be analysed are passed through a hydrogen peroxide solution, which oxidizes sulfur dioxide to sulfuric acid. This sulfuric acid is either titrated with caustic soda solution against bromophenol blue [112] or is precipitated as barium sulfate and weighed [113]. The analyses are disturbed by other acid or sulfur-containing components. If sulfur trioxide or sulfuric acid are present in the gas, they are removed first by absorption in concentrated sulfuric acid in a wash bottle. On account of the solubility of sulfur dioxide in sulfuric acid, this wash acid must first be saturated with sulfur dioxide.

Colorimetric Method. In the colorimetric determination of sulfur dioxide according to WEST and GAEKE the gas is passed through a solution of sodium tetrachloromercurate [119]. After addition of formaldehyde and *p*-rosaniline, the intensely coloured *p*-rosaniline methylsulfonic acid is formed in a Schiff reaction. The sulfur dioxide concentration is determined by a comparison of the extinction at 548 nm with a calibration curve based on standard solutions.

11. Environmental Aspects and Toxicology

Environmental Aspects. A substantially larger amount of sulfur dioxide than is utilized industrially is produced by the combustion of sulfureous fossil fuels and is discharged into the atmosphere with the flue gases. The calculated total emission of sulfur dioxide from power stations, traffic, households, industry, and trade in the Federal Republic of Germany in 1975 amounted to ca. 3.6×10^6 t [120]. In contrast, only about 2.7×10^6 t were used for that year's sulfuric acid production of ca. 4.2×10^6 t H_2SO_4 [121]. In the last years the SO_2 emissions from power plants in Germany have been reduced continuously with the installation of flue gas desulfurization plant and with the utilization of low-sulfur fuels, but emissions from traffic, households, industry, etc. in 1990 still account for 620×10^3 t/a. Emissions from power stations amounted to 320×10^3 t/a. For worldwide atmospheric emissions, only rough estimates are available. For 1970, while global emissions were estimated at 157×10^6 t SO_2 [122], only about 61×10^6 t were consumed in the production of ca. 94×10^6 t of sulfuric acid [121].

Large-scale emission of sulfur dioxide close to ground level has indisputably been the cause of some spectacular environmental problems in the past. A large area around Sudbury, Ontario, was completely sterilized by sulfur dioxide from primitive ore roasting operations around the turn of the century, and much of it is still barren as a result of the ensuing soil erosion. Calamitous sulfuric acid fogs occurred in Donora, Pennsylvania, and London, England, in 1948 and 1952, when adverse climatic conditions prevented sulfur dioxide from industrial sources and domestic coal fires from dispersing. The causes of these episodes were correctly identified, and the action taken to control ground level smoke and sulfur dioxide concentrations has effectively prevented them from recurring. In the case of industrial sources, however, this has often been achieved by merely building taller chimneys to disperse the sulfur dioxide over a wider area, and as the consumption of fuels has increased over the years it has been argued that the problem has merely been shifted from the location where

the sulfur dioxide is produced to other locations downwind.

Considerable publicity has been given to the apparently increasing problem of acid rain, and there does, indeed, seem to be a correlation between the incidence of acid rainfall and the atmospheric transport of pollutants from major industrial locations [123–127]. Establishing in any degree of certainty to what extent sulfur emissions are responsible for acid rain is, however, complicated by incomplete knowledge of the magnitude of sulfur emissions from natural sources, the atmospheric chemistry of sulfur, and the importance of other potential acidulants such as nitrogen oxides. It is also, unfortunately, hampered by politics, particularly in those cases where there is a possibility that pollution from one country could be contributing to acidic rainfall in another [128, 129]. Much of the research is being conducted by or for national agencies which have the power to publish or play down the results as may be expedient. No government purporting to be dedicated to the topical causes of controlling inflation and combatting unemployment at home will lightly order a major programme of sulfur dioxide emission control, with its almost devastating economic implications, just for the purpose of ameliorating the environment in some other country if there is yet room for argument. Nevertheless, the evidence linking sulfur dioxide pollution to acid precipitation has been mounting gradually [130, 131]. A report by the U.S. National Academy of Sciences, for example, has stated that there appears to be a direct proportionate relationship between sulfur dioxide pollution and the amount of acidic sulfates in precipitation [132].

Toxicology. Ill-effects of sulfur dioxide on humans and animals are mainly related to irritation and damage of the moist mucous membranes by the formation of sulfurous acid. The odor threshold of sulfur dioxide in air is between 0.3 and 2.5 ppm. In most human beings, concentrations of 5 – 10 ppm will lead to irritation of the respiratory tract; in sensitive people they may produce spasms of the bronchi. Higher concentrations will cause heavy irritative coughing, while breathing sulfur dioxide at concentrations above 400 – 500 ppm, even for only a few minutes, is dangerous to life. The maximum workplace concentration is defined as 5 ppm (14 mg/m^3).

Long-term exposure may possibly lead to bronchopneumonia and, in extreme cases, to toxic pulmonary edema with dyspnoea, cyanosis and cardiac-circulatory failure. Short-term exposure to extremely high concentrations may cause glottal spasm, gasping or inability to breathe. Symptoms experienced after long-term exposure are loss of appetite, constipation, loss of taste, red tongue, convulsions, eye inflammation, emphysema and nasal and bronchial mucus secretion, which may be blood-stained. Contact with liquid sulfur dioxide causes extremely painful conjunctivitis, conjunctival edema and opacity of the cornea in the eye as well as frostbite on skin and mucous membranes.

Aqueous sulfur dioxide solutions (sulfurous acid) may burn eyes and skin. Swallowing, which is rather improbable on account of the intense warning smell, may cause irritation or burns in the esophagus and stomach, depending on the quantity and concentration swallowed, followed by nausea, gastric pains, vomiting, and diarrhoea. Sulfurous acid absorbed by the body is converted to sulfate and discharged in the urine, as evidenced by an increase in urine acidity.

Effects on Vegetation. Plants are impaired by sulfur dioxide at even lower concentrations than human beings. Sulfur dioxide levels of 1 – 2 ppm are enough to cause acute damage to the leaves of plants in only a few hours, followed by necrosis, probably resulting from impaired photosynthesis. For protection of the environment, the BimSchG law in Germany lays down the following emission values: 80 µg/m^3 related to a dust value of more than 150 µg/m^3, and 120 µg related to a dust value of less than 150 µg/m^3.

Further information on the toxic effects of sulfur dioxide is available in publications [107, 133–136].

12. References

1. Homer: "The Odyssey," XXII, 481 – 494 (ca. 950 BC).
2. *Gmelin,* **Suppl. 3, System no. 9.**
3. L. F. Audrieth, *Liquid Sulfur Dioxide,* Stauffer Chemical Co., New York 1969.

4. *VDI Wärmeatlas,* 3rd. ed., VDI Verlag, Düsseldorf 1977.
5. P. W. Schenk, R. Steudel: "Oxides of Sulfur," in G. Nickless (ed.): *Inorganic Sulfur Chemistry,* Elsevier, Amsterdam 1968, pp. 367 – 418.
6. *Matheson Gas Data Book,* Matheson of Canada Ltd., Whitby, Ontario 1961.
7. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin 1973.
8. R. H. Perry: *Perry's Chemical Engineers' Handbook,* 5th ed., McGraw-Hill, New York 1973.
9. C. L. Yaws, K. Y. Li, C. H. Kuo, *Chem. Eng. (N.Y.)* **81** (1974) no. 14, 85 – 92.
10. C. L. Yaws et al., *Chem. Eng. (N.Y.)* **83** (1976) no. 17, 79 – 87.
11. C. L. Yaws et al., *Chem. Eng. (N.Y.)* **83** (1976) no. 23, 127 – 135; **83** (1976) no. 25, 153 – 162.
12. *Landolt-Börnstein,* **II 2 b (Solution Equilibria I).**
13. *Gmelin,* **B 3, System no. 9.**
14. British Sulphur Corporation data (Dec 1982).
15. M. Smith, *Ind. Miner. (London)* **182** (1982) 49 – 65.
16. British Sulphur Corporation data (Apr 1983).
17. L. J. Friedman, *J. Met.* **33** (1981) no. 3, 44 – 50.
18. *Metallstatistik 1969 – 1979,* vol. 67, Metallgesellschaft AG, Frankfurt/M. 1980.
19. V. D. Terent'ev, L. S. Chernobai, S. A. Tsygankov, M. M. Kocherov, *Sov. Chem. Ind. (Engl. Transl.)* **9** (1977) no. 10, 813 – 815.
20. *Freeport Sulfur Handbook,* Freeport Sulfur Co., New York 1958, p. 22.
21. *Sulphur* **151** (1980) 35 – 39.
22. J. B. Rinckhoff: "Nature's Laws for Sulfuric Acid Plants," in: *Sulfuric/Phosphoric Acid Plant Operations,* American Institute of Chemical Engineers, New York 1982, pp. 101 – 107.
23. *Sulphur* **136** (1978) 54 – 61.
24. *Sulphur* **155** (1981) 28 – 29.
25. A. Urbanek, M. Burgiell: "Formation of Sulfur Trioxide During Sulfur Combustion," in "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp. 5th Int. Conf.* **1981,** 31 – 44.
26. G. Manocha, *Chem. Age India* **22** (1971) no. 3, 190 a – 190 d.
27. V. Krishnan, S. Bielz, Lurgi Technical Information C 131 e, Frankfurt 1979.
28. E. Wiklund, *Aufbereit. Tech.* **18** (1977) no. 6, 285 –289.
29. M. Fairlie: "Sulfuric Acid Manufacture," *ACS Monogr.* **69** (1936) 97 – 114.
30. R. B. Thompson, W. W. Jukkola: "Pyrites and Sulfide Ore Burners or Roasters," in W. W. Duecker, J. R. West (eds.): *The Manufacture of Sulfuric Acid,* Reinhold, New York 1959, pp. 61 – 84.
31. H. Kurushima, R. M. Foley, *Min. Eng.* (Oct 1958).
32. J. L. Stollerey, *Eng. Min. J.* **165** (1964) no. 10, 96 –101.
33. J. T. Graves, T. D. Heath: "Sulfur Dioxide and Sulfur from FluoSolids Systems," *AIME National Meeting,* Los Angeles, Feb. 1967.
34. W. Pfannmüller: "Die Technik des Wirbelschichtröstverfahrens," *DECHEMA Monog.* **21** (1952).
35. A. Johannsen, *Chem. Ing. Tech.* **24** (1952) no. 2, 104 –109.
36. A. Johannsen, W. Danz, *Chem. Ing. Tech.* **29** (1957) no. 9, 563 – 572.
37. Lurgi, Circulating Fluid Bed Technology, Lurgi Technical Information, Frankfurt.
38. *Sulphur* **156** (1981) 40 – 45.
39. H. Wolf, W. Gösele, *Chem. Ing. Tech.* **40** (1968) nos. 9/10, 441 – 445. *Sulphur* **78** (1968) 20 – 25.
40. F. Habashi: *Chalcopyrite – Its Chemistry and Metallurgy,* W. & J. Mackay Ltd., Chatham 1978.
41. N. B. Gray, M. R. Harvey, G. M. Willis: "Physical Chemistry of Process Metallurgy," *Richardson Conference,* London 1973.
42. A. L. Renkey, T. F. Soroka, *J. Met.* **24** (1972) no. 6, 18 – 22.
43. W. Schwartz, W. Eisert, *Erzmetall* **25** (1972) no. 10, 505 – 511.
44. B. Andersson, P. Hanniala, S. Härkki, *CIM Bull.* **75** (1982) no. 845, 172 – 177.
45. G. Melcher, E. Müller, H. Weigel, *Erzmetall* **28** (1975) nos. 7/8, 313 – 322.
46. J. Yannopoulos, Paper T III a 1, presented at the joint *MMIJ-AIME Meeting,* Tokyo 1972.
47. P. J. Lenoir, J. Thiriar, C. Coekelberg: "Use of Oxygen-enriched Air at the Metallurgie Hoboken Overpelt Smelter," *Adv. Extr. Metall. Refin,* Institution of Mining and Metallurgy, London, 1971.
48. J. L. Leroy, P. J. Lenoir, *Adv. Extr. Metall. Refin.* Institution of Mining and Metallurgy, London, 1967.
49. U. H. F. Sander, E. Weilandt, *AIME-TMS Paper Select. no. A* 1977, 77 – 105.
50. S. Dayton, *Eng. Min. J.* **175** (1974) no. 6, 85 – 96.

51. T. Suzuki, T. Nagano, Paper T IV e 4, presented at the joint MMIJ-AIME Meeting, Tokyo 1972. N. J. Themelis, G. C. McKerrow, P. Tarassoff, G. D. Hallett, *AIME Centennial Meeting,* New York 1971; J. Met. 24 (1972) no. 4, 25 – 32. L. Mills, G. Hallett, C. Newmann, Proc. Int. Symp. Extr. Metall. Copper I (1976) 458 – 487.
52. A. L. McKague, G. E. Norman, J. F. Jackson, *CIM Bull.* **73** (1980) no. 818, 132 – 141.
53. L. Renzoni, G. Machum, J. McConnel, A. Magee: Paper presented at the *Symposium Nickel der GDMB,* Wiesbaden 1970.
54. R. Conrad, W. Wuth, *Erzmetall* **23** (1970) no. 6, 282 –287.
55. N. B. Gray, G. Harvey, G. Willis: "Roasting of Sulphides in Theory and Practice," *Richardson Conference,* London 1973.
56. G. L. Bassett, *Eng. Min. J.* **182** (1982) no. 9, 102 –103.
57. A. Berg, F. Pape, *Erzmetall* **31** (1978) no. 5, 224 – 228.
58. K. Shimokawa, T. Takesue: Paper no. TIV c 2, presented at the joint *MMIJ-AIME Meeting,* Tokyo 1972.
59. R. L. Benner, H. Kenworthy, *Rep. Invest. US Bur. Mines* **6769, 6754** (1966).
60. S. Kellog, S. Basu, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **218** (1960) no. 1, 70 – 81.
61. A. Melin, H. Winterhager, *Erzmetall* **20** (1967) no. 12, 561 – 569.
62. J. L. Harrison, *Adv. Extr. Metall. Refin.* 1971, 387.
63. S. E. Woods, C. F. Harries, *Proc. Inst. Chem. Eng. Symp. Chem. Eng. Metall. Ind.* 1963, 77 – 86.
64. P. Bryk, R. Malmstrom, E. Nyholm, *J. Met.* **18** (1966) no. 12, 1298 – 1302.
65. W. Schwartz, P. Fischer, P. Queneau, R. Schuhmann: Paper presented to the *AIME Lead-Zinc-Tin World Symposium,* Las Vegas 1980. P. Fischer, *Metall (Berlin)* **36** (1982) no. 3, 296 – 299.
66. E. Stolarski: Recycling of Spent Sulfuric Acid and its Disposal, preprint Sulfur Conf., Washington, March 1994.
67. T. R. Harris, *Ind. Eng. Chem.* **50** (1958) no. 12, 81 A –82 A.
68. K. Bodenbrenner, H. von Plessen, R. Steiner, *Chem. Ing. Tech.* **50** (1978) no. 1, 30 – 36.
69. U. Sander, G. Daradimos, *Chem. Eng. Prog.* **74** (1978) no. 9, 57 – 67.
70. *VT Verfahrenstech.* **15** (1981) no. 3, 198 – 199.
71. *Sulphur* **152** (1981) 37 – 40.
72. A. Heitmann, P. Reher, *Chem. Ing. Tech.* **46** (1974) no. 14, 589 – 594.
73. J. M. Connor, *Sulphur* **131** (1977) 39 – 45; *Inf. Chim.* **204/205** (1980) 181 – 187.
74. J. M. Connor: "Energy in Spent Sulphuric Acid Recovery by Total Regeneration," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp. 5th Int. Conf.* **1981,** 375 – 391.
75. N. Lowicki: "A New Process for the Recycling of Waste Sulphuric Acids," in "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp. 5th Int. Conf.* **1981,** 401 – 409.
76. W. J. Müller, *Angew. Chem.* **38** (1925) 794 – 795; **39** (1926) 169 – 174.
77. *Phosphorus Potassium* **89** (1979) 36 – 44.
78. H. Kühne, *Chem. Ing. Tech.* **21** (1949) nos. 11/12, 227 – 229.
79. H. W. Gosch, *VT Verfahrenstech.* **7** (1973) no. 11, 319 – 321.
80. W. Q. Hull, F. Schon, H. Zirngibl, *Ind. Eng. Chem.* **49** (1957) no. 8, 1204 – 1214.
81. B. G. Mandelik, C. U. Pierson, *Chem. Eng. Prog.* **64** (1968) no. 11, 75 – 81.
82. V. Robinson, J. D. MacFarlane, *CIM Bull.* **62** (1969) no. 689, 967 – 971.
83. *Sulphur* **144** (1979) 21, 24 – 28.
84. *Ullmann,* 4th ed., **10,** 428 – 431.
85. *Winnacker-Küchler,* 4th ed., **2,** 26 – 27.
86. H. Werth, *Chem. Ind.* (Düsseldorf) 13 (1961) no. 9, 563 – 564.
87. M. Schmitt, *Chem. Ind.* (Düsseldorf) 33 (1981) no. 5, 286 – 290.
88. K. Knoblauch, *Erzmetall* **33** (1980) no. 2, 109 – 114.
89. W. Eisert, *Erzmetall* **31** (1978) no. 4, 157 – 162.
90. M. Beltran, H. Surati: "Use of Two-stage Electrostatic Precipitators in Metallurgical Roasting and Sulphuric Acid Applications," in "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp. 5th Conf.* **1981,** 509 – 523.
91. O. Sundstrom, *Sulphur* **116** (1975) 37 – 43.
92. F. Habashi, R. Dugdale, *Z. Anorg. Allg. Chem.* **380** (1971) 322 – 325.
93. IG Farbenindustrie, DE 553 817, 1931.
94. Stauffer Chemical, CH 403 728, 1966.
95. Allied Chemical & Dye., DE 1 024 935, 1958.
96. Allied Chemical Corp., DE 1 033 642, 1958.
97. G. Rowedder, Lurgi Express Information C 1235/5.77, Frankfurt 1977.
98. H. Schnell, *Kältetechnik* **4** (1952) no. 2, 33 – 35.
99. P. Leckner, R. O. Pearson, R. T. Wood, *Chem. Eng. Prog.* **78** (1982) no. 2, 65 – 70.
100. *Chem. Ing. Tech.* **50** (1978) no. 9, A 484 – A 490.

101. J. M. Henderson, J. B. Pfeiffer, *AIChE Symp. Ser.* **72** (1977) no. 156, 69 – 73.
102. C. M. Yon, G. R. Attwood, C. D. Swaim Jr., *Hydrocarbon Process.* **58** (1979) no. 7, 197 – 200.
103. *Chemieanlagen + -verfahren,* Dec. 1990. *Hoechst High Chem* no. 13, 1992.
104. Federal Statistical Office, Produktion im produzierenden Gewerbe nach Waren und Warengruppen (*Production in the manufacturing industry according to goods and goods groups)*, Wiesbaden 1980.
105. *Curr. Ind. Rep. Inorg. Chem.* **M 28 A** (1980).
106. R. Leclercq, D. W. Bixby, H. L. Fike: "Potential Uses for Sulfur Dioxide," Technical Bulletin no. 19, The Sulphur Institute, Washington, D.C., 1972.
107. BASF, Schwefeldioxid, Technisches Merkblatt M 1905 d (648), 3rd ed., Ludwigshafen/Rh. 1975.
108. *Analyse der Metalle,* 2nd ed., **vol. 2, 1st part,** Springer Verlag, Berlin 1961, pp. 666 – 689.
109. J. H. Karchmer (ed.): *The Analytical Chemistry of Sulfur and its Compounds,* **" Part 1,** Wiley Intersci., New York 1970, pp. 183 – 283.
110. F. D. Snell, L. S. Ettre (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 18,** Interscience Publ., New York 1973, pp. 360 – 440.
111. VDI Richtlinie (VDI Regulation) 2462: Sheet 1 (1974).
112. VDI Richtlinie (VDI Regulation) 2462: Sheet 2 (1974).
113. VDI Richtlinie (VDI Regulation) 2462: Sheet 3 (1974).
114. VDI Richtlinie (VDI Regulation) 2462: Sheet 4 (1975).
115. VDI Richtlinie (VDI Regulation) 2462: Sheet 5 (1979).
116. D. Schunk, *DECHEMA Mongr.* **80** (1976) no. 2, 753 – 761.
117. D. F. Barrett, J. R. Small, *Chem. Eng. Prog.* **69** (1973) no. 12, 35 – 38.
118. J. Kinnunen, B. Wennerstrand, *Fresenius Z. Anal. Chem.* **289** (1978) 110.
119. P. K. Dasgupta, K. Decessare, J. C. Ullrey, *Anal. Chem.* **52** (1980) no. 12, 1912 – 1922.
120. H. Schlachter, *Chem. Ing. Tech.* **52** (1980) no. 12, 925 – 932.
121. K.-H. Schmidt, *Chem. Ind. Tech.* **30** (1978) no. 7, 365 – 370.
122. M. Santa-Olalla, *Pet. Petrochem. Int.* **13** (1973) no. 1, 48 – 50, 52, 83.
123. E. B. Cowling, *Environ. Sci. Technol.* **16** (1982) no. 2, 110 A – 123 A.
124. *Environ. Sci. Technol.* **16** (1982) no. 6, 323 A –327 A.
125. F. Pearce, *New Sci.* (1982) Aug. 12, 419 – 424.
126. L. R. Ember, *Chem. Eng. New* (1982) Nov. 22, 25 – 26.
127. J. Franks, *Chem. Br.* (1983) Jun., 504 – 509.
128. *New Sci.* (1982) July 1, 3; (1982) July 8, 80.
129. *Chem. Eng. (N.Y.)* (1982) Dec. 13, 29 – 30.
130. J. H. Shinn, S. Lynn, *Environ. Sci. Technol.* **13** (1979) no. 9, 1062 – 1067.
131. *Chem. Eng. News* (1981) Sept. 21, 14 – 15.
132. *International Herald Tribune* (Düsseldorf) (1983) Jun. 30, 5.
133. R. Kühn, K. Birett, *Merkblätter Gefährliche Arbeitsstoffe (Loose-leaf Collection on Hazardous Materials),* 3rd ed., **Sheets 509 and 532,** Moderne Industrie, München 1977.
134. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin 1980.
135. W. Weinmann, H. P. Thomas, Arbeitsstoffverordnung, Part 2, Lfg. 5, 1980.
136. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 5th ed., Thieme Verlag, Stuttgart 1972.

Sulfur Dyes

GERT NAGL, Cassella AG, Frankfurt/Main, Federal Republic of Germany

1.	**Introduction**	1
2.	**Chemical Properties**	2
3.	**Classification and Structure**	3
3.1.	**General**	3
3.2.	**Classification**	3
3.2.1.	Sulfur Bake Dyes	3
3.2.2.	Polysulfide Bake Dyes	5
3.2.3.	Polysulfide Melt Dyes	5
3.2.4.	Water-Soluble Sulfur Dyes	7
3.2.5.	Dyes Not Produced by Sulfurization	8
4.	**Production**	8
4.1.	**Sulfur and Polysulfide Bake**	8
4.2.	**Polysulfide Melt**	8
5.	**Environmental Protection**	8
6.	**Commercial Forms**	9
7.	**Quality and Analysis**	9
8.	**Storage and Transportation**	10
9.	**Application**	10
10.	**Economic Aspects**	11
11.	**Toxicology and Occupational Health**	11
12.	**References**	11

1. Introduction

Sulfur dyes have been used for dyeing cellulose fibers since their discovery over 100 years ago. Dyeing cotton and blends thereof with synthetic fibers remains their main application. Sulfur dyes are also used for paper, but to a much lesser extent. A special form of sulfur dyes is used for the fast-dyeing of certain types of leather, but in comparatively small quantities.

From the dyeing aspect sulfur dyes are similar to vat dyes. Unlike vat dyes, which can generally be dissolved in water only with strong reducing agents such as sodium dithionite/sodium hydroxide, sulfur dyes are preferably dissolved with weak reducing agents, especially sodium sulfide or the combination glucose and sodium hydroxide.

The reduced sulfur dyes become attached to the cellulose fiber from hot aqueous solutions because of their pronounced substantivity. Like the reduced vat dyes, they are fixed on the fiber in the insoluble oxidized form by the action of oxidizing agents.

Compared with other classes of dye, the sulfur dyes mainly color cellulose fibers in muted shades. Their advantage lies in the inexpensive production of deep colorations with good fastness properties, particularly in the shades black, navy blue, blue, brown, yellow-brown, olive, and dark green. The development of various commercial forms allows sulfur dyes to be adapted well to modern high-performance dyeing processes.

The following discoveries were milestones in the development of this class of dyes:

1873: Sulfur Brown 1 [*1326-37-0*], made by heating sawdust or other lignin-containing substances with sodium polysulfide (Croissant and Bretonnière), still on the market.

1893: Sulfur Black 3 [*1326-81-4*], made by heating 4-aminophenol with sodium polysulfide. Other sulfur dyes were obtained by heating benzene and naphthalene derivatives with sodium polysulfide (Vidal).

1897: Sulfur Black 9 [*1326-97-2*], made by heating 2,4-dinitro-4′-hydroxydiphenylamine with sodium polysulfide (Kalischer). Under the name Immedial-Schwarz V, it was the first sulfur dye to compete with Aniline Black.

1899: Sulfur Black 1 [*1326-82-5*], made by heating an aqueous mixture of sodium 2,4-dinitrophenolate and sodium polysulfide (Priebs and Kaltwasser). Still irreplaceable in terms of cost.

1900: Sulfur Blue 11 [*1326-98-3*], made by purification and oxidation of Sulfur Black 9 (Bertschmann). Discovery of the possibility of producing brightly colored sulfur dyes by boiling indoaniline derivatives with sodium polysulfide solutions (von Weinberg and Herz).

1902: Sulfur Blue 7 [*85586-03-4*] from the indoaniline of 2-aminotoluene (Herz). Sul-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_613

fur Blue 13 [*1327-59-9*] from the indoaniline of diphenylamine (Herz). Sulfur Orange 1 [*1326-49-4*], made by heating 2,4-diaminotoluene with sulfur (von Weinberg).
1904: Sulfur Green 2 [*1327-74-8*], made from the indoaniline of 8-phenylamino-1-naphthalenesulfonic acid (Böniger).
1907: Sulfur Red 6 [*1327-85-1*], made from 8-amino-7-methylphenazin-2-ol (Lehmann).
1909: Vat Blue 43 [*1327-79-3*], made from the indoaniline of carbazole (Haas and Herz). This dye colors cellulose fibers best from the dithionite vat and is more resistant to chlorine than most sulfur dyes. It was sold under the name Hydron-Blau R and since then has competed with indigo. It is an important commercial product, which is described in the literature as a sulfur vat dye because of the similarity between its properties and those of vat dyes.
1911: Sulfur Black 11 [*1327-14-6*], made from 4-(2-naphthalenylamino)phenol [*93-45-8*] (Heimann). It was introduced on the market in 1926 under the name Indocarbon CL and is mainly used in combination with Vat Blue 43 because of its similar properties.
1935 – 1938: Sulfur Brown 52 [*1327-18-0*] from decacyclene. Sulfur Brown 60 [*1327-20-4*] and Sulfur Brown 51 [*1327-22-6*] from polynitro derivatives of decacyclene (Hagge and Hagen).
1940: Sulfur Green 14 [*12227-06-4*] (clear green) was the first sulfur dye to be synthesized by conversion of tetraamino copper phthalocyanine into tetrathiocyanato copper phthalocyanine (Haddock). Dyes based on copper phthalocyanine on the market are substituted with thiolate groups.
1941: Sulfur Red 14 [*81209-07-6*]: perylene pigment with structure (**1**), substituted with two thiocyanate groups, currently on the market.

1

The sulfur dyes discovered subsequently have remained relatively unimportant.

2. Chemical Properties

The insoluble conventional sulfur dyes are macromolecules cross-linked via di- and polysulfide bridges. They are amorphous and practically insoluble in all solvents. The di- and polysulfide bridges are the reactive groups, which, after reduction to thiolate groups, result in water-solubility. Addition of acid causes the precipitation of the water-insoluble thiol form.

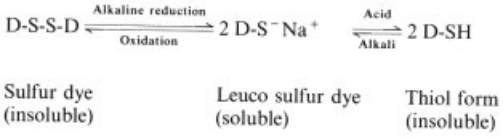

Sulfur dyes that contain the quinoneimine structural element (see Section 3.2.3), undergo a color change on reduction.

The most commonly used reducing agents are sodium sulfide; sodium hydrogen sulfide and sodium carbonate; glucose and sodium hydroxide; and sodium dithionite and sodium hydroxide.

Through the action of mild oxidizing agents (e.g., air), the insoluble sulfur dyes are regenerated on the fiber. Strong oxidizing agents can oxidize the di- and polysulfide bridges of the insoluble sulfur dyes to sulfonic acid groups [5].

Some sulfur dyes form complexes with heavy metal ions. Copper-containing commercial products were previously available, which differed considerably in shade from the copper-free dyes. Apart from the copper phthalocyanine derivatives, these copper-containing dyes are no longer made by the leading producers for ecological reasons.

3. Classification and Structure

3.1. General

Sulfur dyes with a constitution number are defined in the Colour Index by the starting materials, the type of sulfurization, and in some cases the type of work-up. These definitions naturally cannot be precise and were therefore only chosen because little is known about the structures of sulfur dyes.

Even in sulfurization of a defined compound the sulfur dye formed consists of a mixture of products, because sulfurization involves various reactions, including substitution, ring-formation, reduction, and oxidation. The mixture contains positional isomers as well as compounds with different sulfur contents and degrees of condensation. The composition and properties of a sulfur dye depend on the nature and the composition of the sulfurization agent, on the ratio of sulfurization agent to starting material, on the ratio of the starting materials to one another (for two or more starting materials), on the reaction medium, on the reaction conditions, on the method of work-up, and sometimes on the subsequent treatment. Commercial products of higher value are the result of extensive optimization.

3.2. Classification

The classification of sulfur dyes according to the type of sulfurization has become the established method.

Table 1 gives examples of the different types of sulfur dyes.

3.2.1. Sulfur Bake Dyes

Sulfur bake dyes are obtained by heating a starting material with sulfur at 200 – 350 °C. The products are yellow, yellow-brown, orange-brown, or olive. The most important starting materials are 2-(4-aminophenyl)-6-methylbenzothiazole (**2**), 2,4-diaminotoluene, 1,4-diaminobenzene, and decacyclene [*191-48-0*] (**3**).

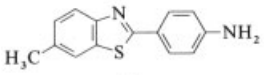

2

3

In the case of starting materials that contain amino and methyl groups the color of the dye originates from the formation of the benzothiazole ring system.

On investigating Sulfur Orange 1, produced from 2,4-diaminotoluene, Zerweck et al. found that ca. 6 – 8 benzene rings are bonded via annelated thiazole rings (**4**) [6].

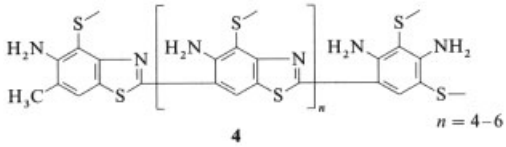

4

It is, however, not known what gives rise to the intensive orange-brown color of Sulfur Brown 52, which is obtained by baking decacyclene with sulfur.

3.2.2. Polysulfide Bake Dyes

Polysulfide bake dyes are obtained by evaporating a mixture of the starting materials with aqueous sodium polysulfide solution and subsequent heating of the thickened material to >180 °C. This method dominated sulfur dye production in the initial phases. Some of the sulfur dyes produced in this way are derived from inexpensive nitro compound starting materials, which are reduced to amino compounds in aqueous media before baking. Dye formation then takes place as with sulfur bake dyes. An example is Sulfur Brown 8 [*1326-41-6*] obtained from 2,4-dinitrotoluene.

In some cases the amino group is already present in the starting material or the starting materials are free of nitrogen. Examples are Sulfur Brown 16 [*1327-12-4*] obtained from 5-amino-1-naphthol, and Sulfur Brown 1 [*1326-37-0*] from lignin.

3.2.3. Polysulfide Melt Dyes

Polysulfide melt dyes are produced below 180 °C in a suspension which can be stirred until sulfurization is complete. Compounds with the general formulas (**5**), (**6**), or (**7**), which contain or can form the quinoneimine structural element, are mainly used as starting materials.

Table 1. Examples of sulfur dyes

Starting materials	Colour Index Generic name	Specification Constitution number	CAS registry no.
Sulfur bake dyes			
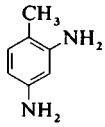	Sulfur Orange 1	53050	[*1326-49-4*]
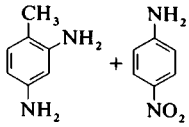	Sulfur Brown 21	53065	[*1326-54-1*]
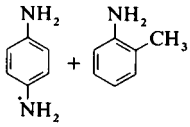	Sulfur Green 12	53045	[*1326-48-3*]
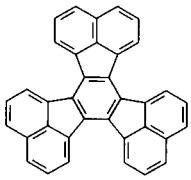	Sulfur Brown 52	53320	[*1327-18-0*]
Polysulfide bake dyes			
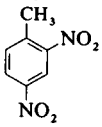	Sulfur Brown 8	53020	[*1326-41-6*]
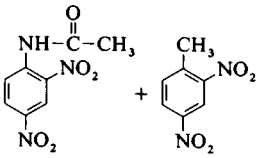	Sulfur Brown 46	53015	[*85940-08-5*]
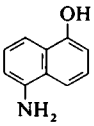	Sulfur Brown 16	53285	[*1327-12-4*]
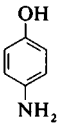	Sulfur Green 11	53165	[*12262-52-1*]
Polysulfide melt dyes			
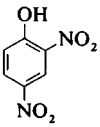	Sulfur Black 1	53185	[*1326-82-5*]
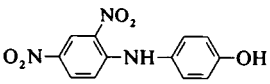	Sulfur Blue 11	53235	[*1326-98-3*]

Table 1. (Continued)

Starting materials	Colour Index Generic name	Specification Constitution number	CAS registry no.
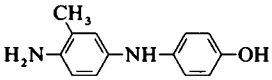	Sulfur Blue 7	53440	[*85586-03-4*]
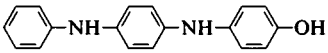	Sulfur Blue 13	53450	[*1327-59-9*]
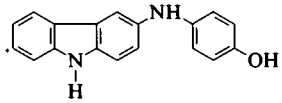	Vat Blue 43	53630	[*1327-79-3*]
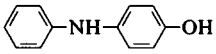	Sulfur Red 10	53228	[*1326-96-1*]
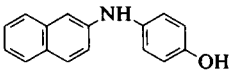	Sulfur Black 11	53290	[*1327-14-6*]
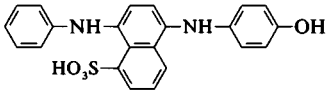	Sulfur Green 2	53571	[*1327-74-8*]
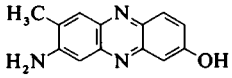	Sulfur Red 6	53720	[*1327-85-1*]

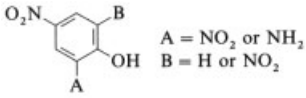

5

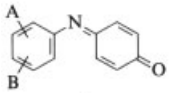

6

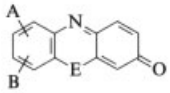

7

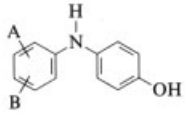

8

As an alternative to the quinoneimines the more stable aminohydroxy compounds (**8**), formed by reduction of quinoneimines, are used.

Sodium sulfide is generally used for the reduction of the quinoneimines. The indoaniline derivatives for the dyes Sulfur Blue 7, 10, and 13, and Vat Blue 43 are produced by condensation of aniline derivatives with 4-nitrosophenol (tautomer: 1,4-benzoquinone oxime) [*104-91-6*] in 70–95 % H_2SO_4 at −25° to +10 °C.

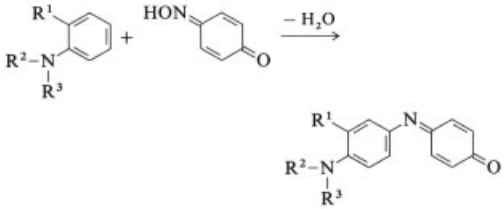

The reaction rates depend on the nature of the aniline derivative, the sulfuric acid concentration, and the temperature [7].

The indoaniline derivative for Sulfur Green 2 is produced by oxidizing a mixture of 4-aminophenol and 8-phenylamino-1-naphthalenesulfonic acid.

The polysulfide melt class encompasses the most important sulfur dyes. Sulfur Black 1 [*1326-82-5*] is formed by heating sodium 2,4-dinitrophenolate in water with sodium polysul-

fide. Most of the nitro groups are reduced. During sulfurization, which occurs on prolonged heating, ammonia is formed besides hydrogen sulfide. For every 3 mol dinitrophenolate used 1 mol ammonia is formed [8], which agrees with the N : C ratio of ca. 5 : 18 in Sulfur Black 1. The S: C ratio in Sulfur Black 1 appears to be generally independent of the sulfurization conditions and to be fairly constant if sulfurization has been carried out until maximum color yield is reached.

The suggested structures published so far do not correspond to the N : C ratio of Sulfur Black 1 [3]. Its constitution is still unknown. The speculative structural formulas **9** and **10** are in good agreement with the elemental ratios found, if oxygen is disregarded.

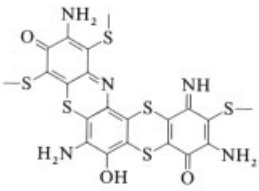

9

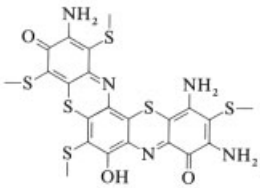

10

With regard to the constitution of the sulfur dyes from compounds with the general formula (**6**), sound ideas concerning important structural elements have been derived from investigations.

Sulfur Blue 9, obtained by sulfurization of the indoaniline derivative (**11**), has a phenothiazinone structure. Degradation of the dye with bromate and hydrobromic acid gave the brominated product (**12**), which was identical with the product formed by bromination of Methylene Violet (**13**) [9].

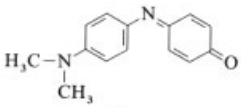

11

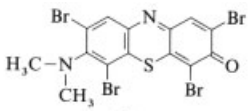

12

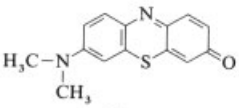

13

ZERWECK et al. prepared the thianthrene compounds (**15**) from the phenothiazone derivative (**14**), in which the three free positions of the quinoid ring are occupied by chlorine atoms, by stepwise substitution, condensation, and substitution. The compounds (**15**) are very similar in their properties to the sulfur dyes produced from the corresponding indoanilines by the polysulfide melt method [9].

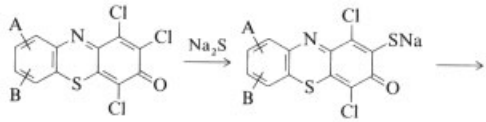

14

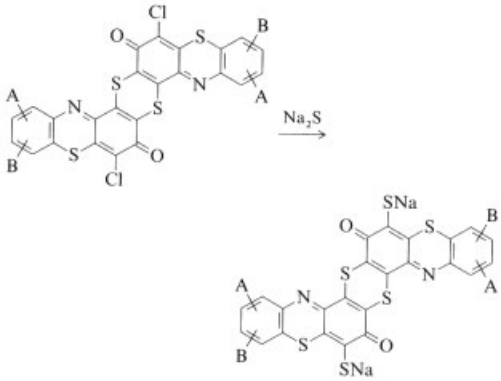

15

It is therefore assumed that the thianthrene compounds (**15**) are also formed in the polysulfide melt. The ratio of elements found in the quinoneimine sulfur dyes indicates that connection of two or more quinoid rings by sulfide or sulfoxide bridges must be involved to a considerable extent.

The Sulfur Blues 7 and 13, Vat Blue 43, Sulfur Black 11, and Sulfur Green 2 belong to the quinoneimine sulfur dyes from indoanilines of the general formula **6** (Table 1).

The quinoneimine sulfur dyes derived from compounds with the general formula **7** are no longer commercially available apart from the red-brown Sulfur Red 6 (Table 1).

3.2.4. Water-Soluble Sulfur Dyes

Sulfur and polysulfide bake dyes and polysulfide melt dyes can be converted into water-soluble thiosulfate esters by sulfite and atmospheric oxygen:

$$\text{D-S-S-D} + 2\,HSO_3^- + 1/2\,O_2 \rightarrow 2\,\text{D-S-}SO_3^- + H_2O$$

3.2.5. Dyes Not Produced by Sulfurization

The classical sulfur dyes, which do not include a red or a clear green, are complemented by two dyes with clear red and clear green shades, not produced by sulfurization.

Sulfur Red 14 with structure (**1**) [*67923-45-9*] behaves like a sulfur dye and dissolves in water in the presence of sodium sulfide with cleavage of the thiocyanate groups to give thiolate groups. Leuco Sulfur Green 36 [*68511-01-3*] is a copper phthalocyanine substituted with thiolate groups.

4. Production

4.1. Sulfur and Polysulfide Bake

In these methods of sulfurization the reaction mixtures become highly viscous to unstirrable during the course of the reaction. Special types of reactor are used to lower the temperature gradient in the melt so that the dyes achieve high color yields and acceptable quality.

Baking crucibles, heated by a heat-transfer medium and equipped with efficient stirring and comminuting equipment are in use, as are horizontally rotating steel drums heated with combustion gases. In the steel drums, steel rods or balls act as stirrers.

A twin-screw reactor has been proposed for the continuous production of the sulfur bake dye based on 2,4-diaminotoluene with addition of ammonium carbonate [10].

After sulfurization is complete, the contents of the reactor are cooled under nitrogen, comminuted, and dissolved in hot, aqueous sodium sulfide solution. If the sulfurized material contains sufficient excess sulfur, it is preferable to boil it with aqueous sodium hydroxide solution. The excess sulfur disproportionates into sodium sulfide and sodium thiosulfate, the former effecting dissolution of the dye. The solution is purified and oxidized with air, converting excess sulfide to thiosulfate. The precipitated dye is filtered off and processed to give the desired commercial form.

4.2. Polysulfide Melt

In polysulfide melt sulfurization, which is carried out with sodium polysulfide in water or other solvents, the reaction mixture remains stirrable throughout the reaction. Stainless-steel, steam-heated, stirred vessels are used as reactors. The condensate can be recycled or discharged.

Besides water, alcohols which are immiscible with water but can be steam-distilled are used, in particular *n*-butanol, or a glycol ether – water mixture. Monoethers of ethylene glycol or diethylene glycol, especially diethylene glycol monoethyl ether (carbitol), are frequently used. An alkali-stable hydrotrope (e.g., sodium xylenesulfonate) is generally added to glycol ether melts. The reaction temperature in the carbitol – water mixtures can be adjusted to a desired value in the range 105 – 140 °C by distillation and be maintained by changing to reflux. This operating procedure and the addition of a hydrotrope gives uniform sulfurization with advantages in terms of color yield and quality.

When dye formation is complete the alcohol is recovered by steam distillation. The partially dissolved dye is generally completely precipitated from the aqueous suspension by air oxidation, filtered, and washed. In some cases it is salted out with salt water. The filter cake is then further processed to the desired commercial form.

There is also the possibility of converting reacted glycol ether melts into preparations ready for use as dyes by adding of reducing agents and water.

5. Environmental Protection

Wastewater. The main salt present in the wastewater from sulfur dye production is sodium thiosulfate. It is formed as a side product of sulfurization with sodium polysulfide, and also from the oxidation of excess sodium polysulfide by atmospheric oxygen. It is also formed during alkaline work-up of sulfur and polysulfide bakes (Section 4.1). Sodium polysulfide is also present in the waste water if the corresponding dyes are isolated by salting out instead of by air oxidation. In some cases sodium sulfite is also present.

Filtrates which mainly contain sodium thiosulfate are treated with sulfuric acid under pressure at temperatures above 130 °C [11]. This disproportionates the thiosulfate to sodium sulfate and elemental sulfur:

$$3\,Na_2S_2O_3 + H_2SO_4 \rightarrow 3\,Na_2SO_4 + 4\,S + H_2O$$

The sulfur is separated and fed back into the production process.

Polysulfide-containing filtrates are mixed with thiosulfate- and sulfite-containing filtrates and then treated with sulfuric acid, so that elemental sulfur is formed [11]. The reaction is controlled so that neither hydrogen sulfide nor sulfur dioxide is liberated and the wastewater only contains sodium sulfate and small amounts of sodium thiosulfate.

Off-gas. Hydrogen sulfide is formed during sulfurization. It is completely absorbed by sodium hydroxide forming sodium hydrogensulfide, which is recycled to the production process or is used as a reducing agent in the dyeing process.

Wastewater from Dyeing Works. In the wastewater from the dyeing works the concentrations of the sulfur dyes are generally low because of their high substantivity and good bath exhaustion. Besides dye residues, it contains the excess sulfide from the dyeing process and downstream products from the oxidizing agents used in the fixing process.

To remove the sulfide, the wastewater is subjected to air oxidation, whereby sulfide is oxidized to thiosulfate. For dyeing works where it is not possible to aerate the wastewater, the use of low-sulfide sulfur dye preparations or sulfide-free finely dispersed commercial forms together with sulfide-free reducing agents solves this problem.

In the case of oxidizing agents, chromate, bromate, and Chloramin T no longer need be used, so that chromate- and AOX-free wastewater is produced. Hydrogen peroxide is the oxidizing agent of choice.

6. Commercial Forms

Powders, the oldest commercial form, consist of dried and ground press cake, standardized with sodium sulfate, sodium chloride, or sodium carbonate. Before application the powder is stirred in water, boiled with the reducing agent, diluted with water and cooled.

In the last few decades, more readily handled forms, which are better adapted to modern dyeing methods, have come on to the market and have extensively replaced the old powder forms.

To produce *low-dust granulates*, press cakes are reduced in concentrated form, mixed with a dispersing agent, and dried in special equipment. These products can be readily dissolved or dispersed in cold water.

To produce *liquid dispersions*, press cakes are slurried with water and dispersing agents and ground in bead mills.

Finely dispersed powders are produced by spray drying liquid dispersions.

Liquid preparations of reduced sulfur dyes are available in a ready-for-use form containing sodium sulfide and sodium hydrogensulfide, and in a low-sulfide form for dyeing with alternative reducing agents. They are produced by dissolving press cakes or reacted melts in water together with the reducing agents commonly used for sulfur dyes. By adding hydrotropic agents, relatively concentrated stable solutions can be obtained. If dyeing is carried out with the low-sulfide form, reducing agent must be added to the dyeing bath to achieve full color strength.

The *water-soluble sulfur dyes* are thiosulfate esters. They are sold as aqueous solutions or in dried form. They have no affinity for the cellulose fiber and are first reduced to the leuco form, which exhibits fiber affinity. This makes them very versatile and they are particularly advantageous where uniform dyeing of an article is only achieved by an excellent penetrating capacity because the packing density is high.

7. Quality and Analysis

To assess the quality of sulfur dyes the technical properties are decisive as in all classes of dyes. The comparison of the color shades and color strengths of the colorations produced under standardized conditions is carried out at differing depths of color visually and by colorimetry. With suitable toning dyes, correction of small deviations in color shades is possible.

The solubility is generally tested after the addition of reducing agents. It is only tested without reducing agents in the case of liquid preparations which are ready for use in dyeing and for thiosulfate esters. For finely dispersed types the fine distribution and the padding liquor stability are also tested. Preparations which are ready for use in dyeing are also standardized in terms of their reducing agent content. Low-sulfide preparations are tested for their total sulfide content.

In analysis for identity comparison, the sulfur dyes contained in the commercial products adjusted to type strength can be readily separated from the additives by dissolving with sodium sulfide in water, precipitation with air, filtering, and washing. Identities can be derived by comparison of the IR spectra of the isolated dyes.

Sulfur dyes cannot be removed from cellulose fibers with reducing agents because of their high substantivity. They can therefore only be detected on textiles as a dye class [12].

8. Storage and Transportation

Commercial products which do not contain any reducing agent are not subject to any specific limitations because of their low toxicity. The safety data sheets of the producer give a dust explosion class for solid commercial products. For transport the dyes are not regarded as dangerous goods. The liquid reduced preparations are subject to the storage and transport regulations of a particular country because of their sulfide content and pH value.

9. Application

Colorations with sulfur dyes predominantly have an average to high fastness level in middle to deep color shades on cellulose fiber. The lightfastness reaches values of almost 8 for black sulfur dyes. The fastness of deep shades to wet rubbing and the resistance to chlorine are problematic. Sulfur dyes are preferably used for economical colorations, e.g., for cord, leisure and outer garments, both on pure cotton and on mixed polyester – cellulose materials. Where deep color shades are to be produced in the colors black, blue, brown, and dark green at favorable prices, the sulfur dyes are without competition.

Table 2. Main producers of sulfur dyes

Producer/country	Range specification	Commercial forms (C.I. classification)
Asahi/Japan	Asathio	sulfur solubilized
	Asathiosol	sulfur
Atul Products/India	Atul Sulphur	sulfur
Cardoner/Spain	Diresul	leuco sulfur
	Pirocard	sulfur
	Pirosol	solubilized sulfur
	Sandozol	solubilized sulfur
Cassella, Hoechst/Germany	Cassulfon	leuco sulfur
	Coranil	solubilized sulfur *
	Hydron	sulfurized vat, dispersed form
	Hydrosol	solubilized sulfur
	Immedial	sulfur
	Imedial C	leuco sulfur
	Indocarbon	sulfur, dispersed form
China National Chemicals/China	Sulphur	sulfur
	Youhaodron	sulfurized vat
Mitsui Toatsu Dyes/Japan	Mitsui Sulphur	sulfur
Nippon Kayaku/Japan	Kayaku	sulfur, dispersed form
	Kayasol Homodye	solubilized sulfur
Patrick/USA	Patcosul	leuco sulfur
Robinson/England	Endurol	sulfurized vat
	Episol	solubilized sulfur *
	Sulphol	sulfur, leuco sulfur
	Sulphosol	solubilized sulfur
Sandoz Chemicals/USA	Sodyeco	sulfur leuco
	Sodyesul	sulfur
	Sodyevat	sulfurized vat, dispersed form
Shanghai Dyestuffs/China	Shangdasulfur	sulfur
Sumitomo/Japan	Nissen	sulfur

* For leather.

Sulfur dyes are used in continuous piece-dyeing works and also on the jigger. In yarn dyeing, in dyeing flocks, dyeing knitwear on winch machines or jet dyeing machines, and dyeing chain yarn for denim articles, black is predominantly used. The dyeing method and the choice of the appropriate commercial form are described in detail in the literature and in the dyeing instructions of the producer [13].

Sulfur dyes are only used in relatively small quantities in other areas. Sulfur Black 1 is used in its various commercial forms in the bulk dyeing of paper, particularly for laminates.

Thiosulfate esters are used in dyeing leather, preferably chrome leather without a reducing agent, and give good colorations throughout with excellent wet- and lightfastness. To improve the absorption of dye the use of special auxiliaries has been proposed [14]. The use of finely dispersed sulfur dyes without reducing agents has been suggested [15].

Other patents describe dyeing the elastomer component of artificial leather [16], dyeing aramid fibers [17], use in ink-jet printing inks [18–21], and the production of pigments for printing inks [22].

Table 2 lists the main producers, trade names, and commercial forms.

10. Economic Aspects

World consumption of sulfur dyes, which depends on cotton consumption and on the color shades that are in fashion, is estimated to be over 100 000 t/a. About 35 % of this is accounted for by North America and Western Europe, and 40 % by the CIS and China. The sulfur dyes account for ca. 30 % of dye consumption for cellulose fibers in terms of quantity, but in terms of value the percentage is much lower. On a long-term average, Sulfur Blacks 1 and 2 account for ca. 65 % of sulfur dye consumption.

11. Toxicology and Occupational Health

The hazard potential of sulfur dyes themselves and on the dyed fiber is regarded as very low because of their macromolecular structures and low bioavailability. Investigations on mammals gave $LD_{50} > 5000$ mg/kg and no skin or mucous membrane irritation. All sulfur dyes investigated gave negative results in the Ames test.

The solid products are now predominantly low in dust.

Reduced preparations are rendered alkaline and contain sodium sulfide, in many cases in the quantity necessary for dyeing. They must be labeled accordingly and may not be brought into contact with acidic substances.

For ecological reasons, low-sulfide reduced preparations are also available, to which additional reducing agents (e.g., glucose) must be added for dyeing.

12. References

General References

1. Colour Index, Society of Dyers and Colourists and AATCC, vol. 4, 3rd ed., 1971, pp. 4475 – 4498.
2. C. Senior, D. A. Clarke in C. Preston (ed.): *The Dyeing of Cellulosic Fibers,* Worshipful Company of Dyers, Society of Dyers and Colourists, Chap. 7, 1986, 256 – 289.
3. R. A. Guest, W. E. Wood, *Rev. Prog. Color. Relat. Top* **19** (1989) 63 – 71.
4. J. R. Aspland, *Text. Chem. Color.* **24** (1992) no. 3, 21 –24;no. 4, 27 – 31.

Specific References

5. IBM, US 4 343 653, 1981 (B. L. Beach, P. D. Gates, V. J. Schum).
6. W. Zerweck, H. Ritter, M. Schubert, *Angew. Chem.* **60** (1948) 141 – 147.
7. O. Manabe, S. Moriwaki, H. Hiyama, *Kogyo Kagaku Zasshi* **57** (1954) 523 – 525.
8. A. v. Weinberg, *Ber. Dtsch. Chem. Ges. A* **63** (1930), 117 – 130.
9. R. Gnehm, F. Kaufler, *Ber. Dtsch. Chem. Ges.* **37** (1904) 2617 – 2623, 3032 – 3033.
10. VEB Chemiekombinat Bitterfeld, DE-OS 4 007 751, 1990 (E. Lange et al.).
11. K. Holoubek, H. Aman, *Chem. Ztg.* **93** (1969) 390 – 395.
12. J. Au-Cheng, S. O. Wanogho, N. D. Watson, B. Caddy, *J. Forensic Sci. Soc.* **31** (1991) no. 1, 31 – 40. K. Venkataraman: *The Analytical Chemistry of Synthetic Dyes,* Wiley-Interscience, New York 1977, pp. 69, 78, 404. G. Dierkes, H. Brocher, *Melliand Textilber.* **44** (1963) 387 – 391.
13. H. Bernhardt, *Melliand Textilber. Int.* **65** (1984) 833 –835.
14. Sandoz, DE-OS 3 601 721, 1986 (M. Dien).
15. Cassella, EP-A 501 197, 1992 (W. Bauer et al.).
16. Asahi, JP-A 01 068 580, 1987 (K. Ooe, H. Takeuchi).
17. Seiren, JP-A 03 045 790, 1989 (M. Kurita).
18. The Mead Corp., US 4 269 627, 1979 (Ki-Sup Hwang).
19. IBM, EP-A 42 922, 1981 (K. E. Edds).
20. Canon, JP-A 63 139 960, 1986 (S. Koike, Y. Tomita).
21. Canon, JP-A 63 139 962, 1986 (S. Koike, Y. Tomita).
22. Sandoz, GB 2 254 856, 1991 (M. J. Domingo, R. Hasler, T. Robinson, V. Sanahuja).

Sulfur Halides

Hans-Dietrich Lauss, Bayer AG, Leverkusen, Federal Republic of Germany(Chaps. 1, 2)

Wilfried Steffens, Bayer AG, Leverkusen, Federal Republic of Germany(Chap. 3)

1. Sulfur – Halogen Compounds 1
1.1. Sulfur Chlorides 1
1.1.1. Disulfur Dichloride 1
1.1.2. Sulfur Dichloride 3
1.1.3. Other Sulfur Chlorides 4
1.2. Sulfur Bromides and Iodides 4
2. Sulfur – Halogen – Oxygen Compounds 5
2.1. Thionyl Halides 5
2.1.1. Thionyl Chloride 5
2.1.2. Thionyl Bromide 8
2.1.3. Other Thionyl Halides 8
2.2. Sulfuryl Halides 8
2.2.1. Sulfuryl Chloride 8
2.2.2. Other Sulfuryl Halides 9
3. Toxicology 10
4. References 11

1. Sulfur – Halogen Compounds

In sulfur halides, sulfur can exhibit the oxidation states 2, 4, or 6. The affinity of sulfur for halogens, and hence the tendency to form higher oxidation states, decreases with increasing atomic radius and decreasing electronegativity of the halogen.

Binary compounds for all typical oxidation states are only known for the sulfur fluorides (see →Fluorine Compounds, Inorganic, Chap. 6.1.→Fluorine Compounds, Inorganic, Chap. 6.2.). Sulfur forms halides of the composition SX_n ($n = 2$, 4, or 6) and S_2X_n ($n = 2$, 4, 10). The halosulfanes S_nX_2 ($n \geq 3$) contain sulfur chains. Besides the binary sulfur halides, the mixed halides SF_5Cl [*13780-57-9*], SF_5Br [*15607-89-3*], and S_2BrCl [*28077-67-0*] are also known.

1.1. Sulfur Chlorides

In the sulfur – chlorine system, the compounds disulfur dichloride (S_2Cl_2), sulfur dichloride (SCl_2), sulfur tetrachloride (SCl_4), and the chlorosulfanes (S_nCl_2, $n \geq 3$) have long been known.

The stability of the sulfur chlorides decreases with increasing oxidation state of the sulfur. Thus, at room temperature S_2Cl_2 can be stored without decomposition, whereas SCl_2 is ca. 16 % dissociated. SCl_4 is only stable at low temperature. Only S_2Cl_2 and SCl_2 have become important industrially.

1.1.1. Disulfur Dichloride

Disulfur dichloride [*10025-67-9*], S_2Cl_2, also known as sulfur monochloride, is the most stable sulfur chloride. It is readily formed on treating sulfur with chlorine at elevated temperature. S_2Cl_2 is an amber liquid with an unpleasant, pungent odor. It fumes in air due to the formation of HCl.

Physical Properties. Some physical properties of S_2Cl_2 are as follows:

Molecular mass	135.03 [6]
Melting point	−82 °C [7]
Boiling point at 0.1 MPa	138 °C
Critical temperature	392 °C [8]
Density at 20 °C	1.6773 g/cm^3 [9]
Dynamic viscosity at 20 °C	0.978 mPa · s [9]
Surface tension, 25 °C	42.5 mN/m [10]
Refractive index n_D^{20}	1.6517 [7]
Dielectric constant, 15 °C	4.79 [11]
Electrical conductivity, 20 °C	$1.3 \times 10^{-10}\ \Omega^{-1}\ cm^{-1}$ [12]
Vapor pressure, 20 °C	1.43 kPa [13]
Specific heat, liquid, 20 °C	0.921 J g^{-1} K^{-1} [14]
Molar heat capacity, gaseous, STP	72.77 J mol^{-1} K^{-1} [15]
Heat of formation	$\Delta H^0_{298} = -58.5$ kJ/mol [16]
Enthalpy of evaporation at 0.1 MPa	36.1 kJ/mol [13]
Ignition temperature	390 °C [17]
Explosion range, 0.1 MPa	4.2 – 32.5 vol % [17]

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_623

Chemical Properties. Disulfur dichloride vapors form explosive mixtures with air. On heating, the orange-yellow liquid assumes a reddish tinge due to partial dissociation. On distillation of S_2Cl_2 the vapor phase contains some SCl_2, while the liquid phase becomes enriched in sulfur. Vacuum distillation gives mostly undecomposed disulfur dichloride. S_2Cl_2 acts as a solvent for covalent inorganic and organic compounds such as $SOCl_2$, SO_2Cl_2, $POCl_3$, $AsCl_3$, $TiCl_4$, $COCl_2$, and CCl_4, but does not dissolve saltlike metal compounds. Disulfur dichloride reacts only slowly with water at room temperature. The immiscible liquids form a sulfur film at the phase interface, which inhibits further reaction. On heating, the reaction rate increases explosively. The heat of hydrolysis with excess water is − 146.5 kJ/mol [18].

On hydrolysis elemental sulfur, thiosulfate, polythionic acids $H_2S_nO_6$, HCl, and SO_2 are formed. In the presence of sulfite or SO_2, disulfur dichloride is converted into polythionic acids, elemental sulfur only being formed in trace amounts. S_2Cl_2 reacts with many elements, oxides, and sulfides, with formation of the corresponding chlorides; heating is required in most cases. Titanium reacts vigorously in the cold with incandescence. Because of the formation of protective layers of insoluble chlorides or sulfides, metals such as Ni, Co, Cr, Mn, Cd, Zn, Ag, and Cu are resistant to S_2Cl_2. Disulfur dichloride dissolves large quantities of sulfur, with formation of chlorosulfanes S_xCl_2. Dry ammonia reacts vigorously with S_2Cl_2 to form N_4S_4 [*28950-34-7*]. Depending on the reaction conditions S_7NH, $S_6(NH)_2$, sulfur, and NH_4Cl are formed as side products [19]. S_2Cl_2 is miscible with many organic solvents or is readily soluble in them. Reaction of S_2Cl_2 with organic compounds results in introduction of sulfur, chlorine, or both into the molecule [20–24]. Dehydration and condensation reactions are also known. S_2Cl_2 is an effective catalyst for many chlorination reactions.

With ethylene, S_2Cl_2 forms bis(2-chloroethyl) sulfide [*505-60-2*] (mustard gas) [25]. Reactions of S_2Cl_2 with unsaturated compounds are of industrial interest. Rubberlike products, known as factice, are formed with unsaturated vegetable oils. S_2Cl_2 acts as a cross-linking agent in unsaturated polymers. With aromatic compounds the corresponding mono- or disulfides are generally formed, and also poly(thioarylenes) under certain conditions [26], [27].

Production. Disulfur dichloride is generally produced from the elements. In the batch process, chlorine is passed into a solution of sulfur in S_2Cl_2 with cooling at ca. 50 – 70 °C until the correct chlorine content is reached. In the continuous process, gaseous chlorine is passed into molten sulfur at 220 – 260 °C, and the gaseous S_2Cl_2 leaving the reactor is then condensed [28]. Production is carried out in equipment made of steel or chromium – nickel steel. Disulfur dichloride is an intermediate in the production of CCl_4 and trichloromethylsulfenyl chloride CCl_3SCl [*594-42-3*] from CS_2 and can be obtained from the process as a byproduct. It is also formed in the chlorination of mercaptans to give the corresponding chlorides.

$$2\,RSH + 3\,Cl_2 \longrightarrow 2\,RCl + S_2Cl_2 + 2\,HCl$$

Storage and Transport. S_2Cl_2 is sold in steel or stainless steel drums, tank containers, and tank cars. To avoid corrosion the entry of moisture should be carefully excluded. For handling, see [29].

Transport classification:

IMDG Code	Class 8
UN No.	1828
Packing group	I
RID/ADR	Class 8, no. 21 a
ADNR	Class 8, no. 11 a
ICAP-TI/IATA-DGR	max. 2.5 L

Quality and Analysis. Commercial S_2Cl_2 is a clear liquid, which is light yellow at room temperature and has a chlorine content of 51.5 – 52.5 %. It can contain SCl_2, SO_2, $SOCl_2$, chlorosulfanes, and dissolved sulfur as impurities. The chlorine content is determined by volumetric analysis. The free sulfur and chlorosulfane contents can be determined as the evaporation residue in vacuo (3 hPa). An oily residue indicates the presence of chlorosulfanes. SCl_2, SO_2, $SOCl_2$, and organic impurities are determined by GLC.

Uses. Large quantities of S_2Cl_2 are used in the production of cutting oils and additives for

high-pressure lubricating oils [30]. The S_2Cl_2 is reacted with unsaturated organic compounds, such as olefins, fats, or fatty acids. Since low-chlorine lubricating oils have increased in importance, this use of S_2Cl_2 has declined considerably. By reaction with morpholine or phenols, vulcanizing agents for high-quality rubber grades are obtained [31]. Reaction with vegetable oils gives gel-like substances known as factice, which are used for erasers and as components of rubber compounds. The sulfurization of oils for varnishes is now of very minor importance. Smaller quantities of S_2Cl_2 are used in cold vulcanization of thin-walled rubber articles; for the production of organosulfur compounds, pharmaceuticals, and crop protection agents; and as a chlorination catalyst.

Herz compounds, obtained from S_2Cl_2 and aniline, are important in the production of some sulfur dyes. S_2Cl_2 can also be used as an absorbent for chlorine [32].

Economic Aspects. The consumption of S_2Cl_2 in the Western world was ca. 30 000 t/a in 1987. Of this ca. 85 % was used for the production of lubricant additives, and ca. 2000 t for the production of factice and rubber auxiliaries. Because of the decrease in the consumption of chlorine-containing lubricating agents the annual worldwide S_2Cl_2 consumption decreased to $<$ 15 000 t/a in 1992.

Producers of disulfur dichloride are as follows:

United States:	Occidental Petroleum, Miles
Western Europe:	Bayer, Germany; Elf Atochem and Produit Chimique de Loos, France; Nipa Laboratories, United Kingdom.
Asia:	Transpek, India; Sumitomo Seika, Japan.

Waste. When S_2Cl_2 is used in organic syntheses an off-gas containing HCl, SO_2, H_2S, and S_2Cl_2 is formed. The off-gas can be purified by scrubbing with water, sodium hydroxide solution, or sodium hypochlorite solution in jet washers or packed towers. Residues containing S_2Cl_2 are disposed of in special incinerators.

1.1.2. Sulfur Dichloride

Sulfur dichloride [*10545-99-0*], SCl_2, is an unstable compound which is already ca. 16 % dissociated at room temperature according to the equation:

$$2\,SCl_2 \rightleftharpoons S_2Cl_2 + Cl_2$$

The decomposition can be inhibited be compounds such as PCl_5 or $PSCl_3$.

Physical Properties. Some physical properties of sulfur dichloride are as follows:

Molecular mass	102.97 [6]
Melting point	− 122 °C [7]
Boiling point at 0.1 MPa	59.5 °C [10]
Vapor pressure, 20 °C	21.6 kPa [91]
Density at 20 °C	1.621 g/cm^3 [9]
Surface tension, 25 °C	33.42 mN/m [10]
Viscosity at 20 °C	0.548 mPa · s [10]
Dielectric constant, 15 °C	3.0 [11]
Enthalpy of formation, 25 °C	− 22.2 kJ/mol [33]
Specific heat (liquid, 0 – 140 °C)	1.01 J g^{-1} K^{-1} [10]
Molar heat capacity (gas, 25 °C, 0.1 MPa)	50.89 J mol^{-1} K^{-1} [15]
Explosion range in air, 100 °C, 0.1 MPa	16.3 – 36.5 vol % [34]

Sulfur dichloride is a red liquid which fumes in moist air. It is electrically nonconducting. Although its flash point cannot be determined, it can be ignited in closed vessels when mixed with air. Sulfur dichloride is miscible with many inorganic and organic covalent compounds such as liquid SO_2, S_2Cl_2, $SOCl_2$, SO_2Cl_2, chlorohydrocarbons, and benzene.

Chemical Properties. Sulfur dichloride decomposes rapidly and completely on treatment with water, forming HCl, thiosulfate, and polythionic acids; elemental sulfur is only formed in trace amounts. In its reactions SCl_2 resembles S_2Cl_2. It is a strong chlorinating agent which converts many elements, oxides, and sulfides into chlorides. It forms stable complexes of the type $MCl_2 \cdot SCl_2$ with many metal chlorides. SCl_2 is oxidized to thionyl chloride by oxygen, SO_2, SO_3, or SO_2Cl_2 at elevated temperature, particularly in the presence of catalysts such as activated charcoal [35]. SCl_2 reacts with NH_3 to give S_4N_4, $S_6(NH_2)_2$, and NH_4Cl [36].

Sulfur dichloride generally acts as a strong chlorinating agent towards organic compounds [20]. It readily adds to double bonds [37]. It forms mustard gas on reaction with ethylene. Reactions with organic compounds are described in detail in [20–22], [31], [37–39].

Production. Technical-grade sulfur dichloride is produced by passing gaseous chlorine into S_2Cl_2 at low temperature until the correct chlorine content is reached. Sulfur dichloride can be purified by distillation in the presence of stabilizers [10], [40], [41].

Storage and Transport. The same applies as to S_2Cl_2 (see page 2).

Quality and Analysis. Sulfur dichloride is sold in technical grade with a chlorine content of 66–70 %. The product contains ca. 72–80 % SCl_2, residual S_2Cl_2, and Cl_2. Stabilized pure sulfur dichloride should contain min. 98 % SCl_2. It can only be stored for a limited period.

The analysis is carried out as for S_2Cl_2 by determination of the chlorine content and by GLC.

Waste is treated as for S_2Cl_2 (see page 3).

Uses. As with S_2Cl_2 large quantities of sulfur dichloride are used for the production of lubricating oil additives. Smaller amounts are used in the production of antioxidants, organosulfur compounds, and rubber chemicals.

Economic Aspects. World consumption of technical-grade sulfur dichloride was ca. 8000 t/a in 1992.

Producers:

United States:	Occidental Petroleum
Western Europe:	Elf Atochem and Produit Chimique de Loos, France;
Asia:	Transpek, India; Sumitomo Seika, Japan.

1.1.3. Other Sulfur Chlorides

Sulfur tetrachloride [*13451-08-6*] is a highly unstable compound, which is formed as a fine, white powder on mixing chlorine with SCl_2 below −75 °C [42]. On warming above the melting point of −31 °C it decomposes into SCl_2 and Cl_2.

On cooling S_2Cl_2–Cl_2 mixtures containing 57.2–59.2 atom % Cl_2 to ca. −110 °C a flocculent precipitate is formed which has been referred to as *trisulfur tetrachloride* S_3Cl_4 [43].

Chlorosulfanes (S_xCl_2, $x \geq 3$) [*50864-86-3*] are oily, strongly refractive liquids at room temperature. Their color ranges from light yellow (S_3Cl_2) to orange-red for the higher homologues. The vapor pressure decreases with increasing chain length. The lower chlorosulfanes are formed by treating sulfur with S_2Cl_2. Higher homologues can be obtained by reacting polysulfanes H_2S_x with S_2Cl_2 or by the action of sulfur vapor on S_2Cl_2 at 400–900 °C. Chlorosulfanes with a chain length x of up to 100 have been reported. The analytical identification of the individual homologues can be carried out by using HPLC [44].

1.2. Sulfur Bromides and Iodides

Of the pure sulfur–bromine compounds only disulfur dibromide and the bromosulfanes (S_nBr_2) are known.

Disulfur dibromide [*13172-31-1*], S_2Br_2, is a garnet red liquid. It can be obtained from its constituent elements or by treatment of gaseous HBr with S_2Cl_2 at room temperature [45]. Some physical properties of disulfur dibromide are as follows:

Molecular mass	223.93 [6]
Melting point	−46 °C [46]
Boiling point, 24 Pa	54 °C [45]
Density, 20 °C	2.629 g/cm^3 [47]
Viscosity, 20 °C	2.41 mPa · s [47]
Surface tension, 20 °C	39.1 mN/m [47]
Molar heat capacity (gas, 25 °C)	75.28 J mol^{-1} K^{-1} [48]
Enthalpy of formation (liquid, 25 °C)	−10.5 ± 4.2 kJ/mol [48]

Bromosulfanes (S_nBr_2, $n = 3-8$) are formed quantitatively from chlorosulfanes and HBr at room temperature. They are oily liquids whose colors range from dark red (S_2Br_2) to raspberry red (S_8Br_2) and whose stability decreases as the sulfur chain length increases. Unlike the chlorosulfanes, the density decreases with increasing chain length [47].

Sulfur iodides. Of the sulfur iodides only S_2I_2 [*53280-15-2*] has been detected with certainty so far [49]. It is formed from S_2Cl_2 and I_2 in fluorohydrocarbons at −78 °C [50].

2. Sulfur – Halogen – Oxygen Compounds

Sulfur forms the following types of halogen – oxygen compounds:thionyl halides SOX_2, sulfuryl halides SO_2X_2, and complex sulfur oxyhalides.

The sulfur – halogen – oxygen compounds can be considered formally as derivatives of the sulfur oxyacids in which the OH groups have been replaced by halogens. On hydrolysis with water all S – O – X compounds are converted to the corresponding sulfur acids. Like the sulfur halides the thermal stability decreases with increasing atomic mass of the halogen. Thus SOF_2 and SO_2F_2 are very stable compounds, while $SOCl_2$ and SO_2Cl_2 already dissociate slowly just above the boiling point. $SOBr_2$ can only be distilled in vacuo without decomposing. SOI_2 is only known in solution. Hydrolysis with water shows a similar trend: SO_2F_2 is stable towards hot water, SOF_2 reacts only slowly with water, while $SOCl_2$ or SO_2Cl_2 reacts vigorously and completely with formation of HCl and SO_2, or HCl and H_2SO_4, respectively.

2.1. Thionyl Halides

Thionyl halides are derivatives of sulfurous acid $O{=}S(OH)_2$. The following thionyl halides and mixed thionyl halides are known: SOF_2, $SOCl_2$, $SOBr_2$, SOI_2, SOClF, SOBrF, and SOBrCl. Thionyl chloride has achieved major industrial importance. Thionyl bromide is used as a brominating agent and solvent in preparative organic chemistry. SOI_2 and the other thionyl halides are unsuitable for use in synthesis because of their instability.

2.1.1. Thionyl Chloride

Thionyl chloride [*7719-09-7*], $SOCl_2$, also known as sulfurous oxychloride or sulfinyl chloride, was first prepared in 1849 by treatment of SO_2 with PCl_5 [51], [52]. After it had been obtained pure, its advantages as a reagent, especially as a chlorinating agent, were recognized. From ca. 1900 onwards $SOCl_2$ was sufficiently available and reasonably priced, because of an improved production method, that it could even be used for large-scale industrial syntheses. Thionyl chloride is now one of the most important chlorinating agents in organic chemistry.

Physical Properties. Thionyl chloride is a colorless strongly refractive liquid with a pungent odor. It fumes in moist air due to HCl formation. Unlike $COCl_2$, $SOCl_2$ has a pyramidal structure. Some physical properties are listed below:

Molecular mass	118.97 [6]
Melting point	− 104.5 °C [53]
Boiling point at 0.1 MPa	76 °C [3]
Heat of evaporation at 0.1 MPa	31.3 kJ/mol [54]
Vapor pressure, 20 °C	12.9 kPa [55]
Critical temperature	296.8 °C
Density, 20 °C	1.640 g/cm^3 [56]
50 °C	1.579 g/cm^3 [56]
Surface tension, 18 °C	34.0 mN/m
Viscosity, 25 °C	0.60 mPa · s
Specific heat (liquid)	1.00 J g^{-1} K^{-1} [54]
Molar heat capacity (gas)	66.6 J mol^{-1} K^{-1} [57]
Enthalpy of formation (gas) ΔH^0_{298}	− 244.9 kJ/mol [16]
Refractive index n_D^{20}	1.517 [54]
Dielectric constant, 20 °C	9.05 [54]
Specific electrical conductivity	7 – 60 × 10^{-9} Ω^{-1} cm^{-1} [3]

Chemical Properties. Thionyl chloride is a nonflammable liquid which undergoes slow thermal decomposition on heating above the boiling point:

$$4\,SOCl_2 \longrightarrow S_2Cl_2 + 2\,SO_2 + 3\,Cl_2$$

The decomposition is measurable in the absence of catalysts above 100 °C and is rapid above 180 °C. SO and Cl_2 are probably formed initially [58].

Inorganic salts are generally insoluble in $SOCl_2$, but covalent and organic compounds mostly exhibit good solubility.

Reactions with Inorganic Compounds. Thionyl chloride is hydrolyzed rapidly and completely by water in an exothermic reaction($\Delta H = -136$ kJ/ mol with excess water).

$$SOCl_2 + H_2O \longrightarrow SO_2 + 2\,HCl$$

Thionyl chloride reacts with the water of crystallization of metal chloride hydrates to give the anhydrous metal chlorides.

With many elements $SOCl_2$ forms the corresponding chlorides, SO_2, and S_2Cl_2. The reaction rate varies widely. Some (Sb, Ti) react

vigorously at room temperature, others only on heating. Mg, Zn, Cd, Pb, Cr, and Ni are resistant up to 200 °C. The low reactivity is due to the formation of a protective layer of metal chloride that is insoluble in $SOCl_2$. In the presence of organic compounds, the rate of reaction may increase considerably. $SOCl_2$ generally only reacts with metal oxides at elevated temperature. The oxides of some transition metals (e.g., V, Cr, W) form oxychlorides with $SOCl_2$. The reaction of $SOCl_2$ with silver oxide or silver sulfide Ag_2S at reduced pressure and elevated temperature is used for the synthesis of sulfur suboxide S_2O [59].

$$Ag_2S + SOCl_2 \longrightarrow 2\,AgCl + S_2O$$

Peroxides give the chlorides, sulfates, and sulfuryl chloride. The reaction is sometimes explosive [60].

With some metal chlorides ($AlCl_3$, $ZrCl_4$, UCl_5) $SOCl_2$ forms addition compounds.

Concentrated sulfuric acid reacts with $SOCl_2$ to give chlorosulfonic acid.

$$H_2SO_4 + SOCl_2 \longrightarrow HSO_3Cl + SO_2 + HCl$$

Analogously hydrogensulfates form chlorosulfates. Sulfur trioxide forms pyrosulfuryl chloride $S_2O_5Cl_2$ on reaction with thionyl chloride.

The reaction of $SOCl_2$ with ammonia and other inorganic nitrogen compounds leads to various products, depending on the reaction conditions [19], [61]. Gaseous ammonia reacts exothermically with $SOCl_2$ vapor to give thionyl imide OSNH [*13817-04-4*] and NH_4Cl [62]. The use of thionyl chloride in preparative inorganic and coordination chemistry is described in [63].

Reactions with Organic Compounds. The most important reaction of thionyl chloride is the substitution of OH groups by chlorine in alcohols, carboxylic acids, and sulfonic acids. The reaction is particularly advantageous as the reaction products, SO_2 and HCl, are gaseous and can be removed readily. The reaction proceeds via the intermediate products sulfite and chlorosulfinate:

$$2\,ROH + SOCl_2 \longrightarrow (RO)_2SO + 2\,HCl$$

$$(RO)_2SO + SOCl_2 \longrightarrow 2\,ROSOCl$$

$$ROSOCl \longrightarrow RCl + SO_2$$

Both intermediate products can be isolated. Sulfite formation is favored by low reaction temperature, a large excess of $SOCl_2$, and use of solvents. The decomposition of the chlorosulfinates is catalyzed by tertiary amines. The chlorination of aliphatic OH groups can take place with retention or inversion of configuration, depending on the reaction conditions. In some cases intramolecular rearrangement can occur [54].

The reaction with thiols proceeds similarly:

$$RSH + 2\,SOCl_2 \longrightarrow RCl + S_2Cl_2 + SO_2 + HCl$$

Phenolic OH groups generally do not react. Carboxylic acids are generally converted rapidly by $SOCl_2$ into acid chlorides in good yield. The reaction is accelerated by catalysts such as tertiary amines, carboxylic acid or sulfonic acid amides (e.g., DMF), or inorganic salts (KCl, $ZnCl_2$). However, on using DMF as catalyst, formation of dimethylcarbamoyl chloride, a potential carcinogen, cannot be ruled out. If a large excess of $SOCl_2$ is added to the acid, then the acid anhydride is generally obtained. Thionyl chloride is also important in dehydration and condensation reactions. For reactions in organic chemistry, see [20], [54], [64–66].

Production. Thionyl chloride is now generally produced by treating SO_2 or SO_3 with SCl_2 at ca. 150 – 250 °C in the presence of activated carbon [67], [68].

$$SO_3 + SCl_2 \longrightarrow SOCl_2 + SO_2$$

$$SO_2 + SCl_2 + Cl_2 \longrightarrow 2\,SOCl_2$$

The crude product is purified by distillation after addition of sulfur. Crude thionyl chloride production is a continuous process, but the distillation is carried out either continuously or in batches. The side product SO_2 can be reacted with chlorine to give SO_2Cl_2 and thus recycled.

$$SO_2Cl_2 + SCl_2 \longrightarrow 2\,SOCl_2$$

Older processes [69], which are no longer important, are based on the chlorination of SO_2 with PCl_5 or $COCl_2$

$$SO_2 + PCl_5 \longrightarrow SOCl_2 + POCl_3$$

$$SO_2 + COCl_2 \longrightarrow SOCl_2 + CO_2$$

or the reaction of SCl_2 or S_2Cl_2 with SO_3 or HSO_3Cl.

$$SO_3 + S_2Cl_2 \longrightarrow S + SOCl_2 + SO_2$$

Technical-grade thionyl chloride can contain sulfur chlorides, SO_2, and SO_2Cl_2. Sulfur chlorides can be removed by distillation over sulfur, in some cases with the addition of $FeCl_3$ or $AlCl_3$ [70]. For laboratory-scale purification, distillation with triphenyl phosphite [71] or dipentene [72] is recommended. Modern processes give thionyl chloride in a quality which does not require any additional purification.

Storage and Transport. Thionyl chloride is sold in galvanized steel drums, and in stainless steel railway tank cars and tank containers.

Transport Classification	
IMDG Code	Class 8
UN no.	1836
Packing group	I
RID/ADR	Class 8, no. 21 a
ADNR	Class 8, no. 11 a
ICAP-TI/IATA-DGR	prohibited

Quality Specifications and Analysis. Technical-grade thionyl chloride is a clear, colorless to pale yellow liquid. The thionyl chloride content should be min. 99.6 %. The following impurities can be present: ≤ 0.3 % SO_2, ≤ 0.2 % SO_2Cl_2, ≤ 0.05 % sulfur chlorides, and ≤ 0.1 % evaporation residue.

The IR spectrum can be used for identification. The assay can also be found by determination of the chlorine content. Sulfur chlorides, SO_2, and organic impurities are determined by GLC [73], [74]. Sulfur chlorides can be determined by UV spectrometry, and SO_2 and SO_2Cl_2 by IR spectroscopy [75], [76].

Waste Disposal. In nearly all reactions with thionyl chloride an off-gas is formed which contains predominantly SO_2 and/or HCl, besides thionyl chloride. In addition, varying quantities of organic reactants and solvents can be present. The off-gas stream is generally purified by scrubbing with water or dilute alkali in packed column or jet scrubbers. The resulting wastewater contains chloride and sulfite and must be disposed of according to local regulations. The oxidation of the sulfite to sulfate can be carried out with oxidizing agents, such as chlorine, hypochlorite, or hydrogen peroxide. Oxidation with air or oxygen in the presence of catalysts is also possible. For larger quantitites of off-gas chlorine can be recovered as aqueous hydrochloric acid, and SO_2 as sodium hydrogensulfite solution.

Uses. Thionyl chloride is used in industry for the production of acid chlorides from carboxylic acids, for the production of alkyl chlorides from alcohols, for ester production, and, together with formaldehyde, for chloromethylation.

The production of aromatic sulfonic acid chlorides with chlorosulfonic acid and thionyl chloride is particularly advantageous, as higher yields are achieved compared with the use of chlorosulfonic acid alone and smaller quantities of waste acid are formed [77]. Important areas of use are the production of crop-protection agents, pharmaceuticals, dyes, paper and textile auxiliaries, and plastics.

Smaller quantitites are used as liquid cathodes in certain types of lithium battery [78–80].

Economic Aspects. The annual worldwide thionyl chloride consumption was ca. 45 000 t in 1992.

Producers

United States:	Miles
Europe:	Bayer, Germany; Säurefabrik Schweizerhall, Switzerland.
Asia:	Transpek India; Sumitomo Seika, Japan.

2.1.2. Thionyl Bromide

Thionyl bromide [*507-16-4*], $SOBr_2$, is an orange-yellow liquid, which becomes red on standing due to decomposition into SO_2, Br_2, and S_2Br_2.

Some physical properties of thionyl bromide are as follows:

Molecular mass	207.85 [6]
Melting point	− 52 °C [53]
Boiling point, 2 kPa	39 °C [81]
0.1 MPa (with dec.)	138 °C [53]
Heat of evaporation, 25 °C, 0.1 MPa	46.1 kJ/mol [81]
Density, 15 °C	2.698 g/cm^3 [53]
Heat of formation (gas) ΔH^0_{298}	− 168.3 kJ/mol [81]
Molar heat capacity (gas)	69.7 J mol^{-1} K^{-} [81]

Chemical Properties. Thionyl bromide is readily soluble in many anhydrous organic solvents, such as CS_2, C_6H_6, or CCl_4, in which it is probably associated. With water complete hydrolysis to SO_2 and HBr occurs immediately. The reactions of $SOBr_2$ with organic compounds are summarized in [54], [82]. It is sometimes used as a brominating agent in preparative chemistry. Carboxylic acids are converted into acyl bromides, and alcohols into alkyl bromides.

Production. Thionyl bromide can be produced by passing HBr into thionyl chloride with cooling and careful exclusion of moisture. For purification the reaction mixture is fractionated under vacuum [45].

2.1.3. Other Thionyl Halides

Thionyl iodide [*55032-43-4*] is formed by the reaction of $SOCl_2$ with HI at $-78\,°C$ in fluorochlorocarbons. The dark brown solid is reported to be stable below $-30\,°C$ [50].

At $0\,°C$ *thionyl chloride fluoride* [*14177-25-4*], SOClF, is a colorless liquid with a choking odor. The boiling point is $12.3\,°C$ and the solidification point $-139.5\,°C$. It is formed by reacting thionyl chloride with IF_5 and condensing the gases produced with liquid air [45]. SOClF is thermally unstable.

Thionyl bromide fluoride [*81539-20-0*], SOBrF, is formed by fluorination of $SOBr_2$ [83]. For thermodynamic data see [84].

Thionyl bromide chloride [*22650-47-1*], SOBrCl, can be detected spectroscopically but cannot be isolated as a pure substance because of disproportionation to $SOCl_2$ and $SOBr_2$. For thermodynamic data, see [84].

2.2. Sulfuryl Halides

Sulfuryl halides are formally derived from sulfuric acid by replacement of both OH groups by halogen. The sulfuryl halides SO_2F_2, SO_2Cl_2, SO_2ClF, and SO_2ClBr are known.

2.2.1. Sulfuryl Chloride

Sulfuryl chloride [*7791-25-5*], SO_2Cl_2, also known as sulfuric oxychloride, is an unstable compound, whose formation in the equilibrium reaction

$$SO_2 + Cl_2 \rightleftharpoons SO_2Cl_2$$

proceeds immeasurably slowly at room temperature in the absence of catalysts.

Physical Properties [1]. Sulfuryl chloride is a colorless liquid with a pungent odor.

Molecular mass	134.97
Melting point	$-54\,°C$
Boiling point, 0.1 MPa	$69.1\,°C$
Heat of evaporation at $69\,°C$	29.6 kJ/mol
Vapor pressure, $18\,°C$	12.7 kPa
Critical temperature	$269-276\,°C$
Density, $25\,°C$	$1.6570\,g/cm^3$
Surface tension, $23.5\,°C$	23.92 mN/m
Viscosity, $25\,°C$	$0.918\,kPa \cdot s$
Specific heat (liquid, $15-63\,°C$)	$0.976\,J\,g^{-1}\,K^{-1}$
Molar heat capacity (gas, $25\,°C$)	$77.5\,J\,mol^{-1}\,K^{-1}$
Enthalpy of formation, $18\,°C$	-389.8 kJ/mol
Refractive index n_D^{20}	1.4437
Dielectric constant, $22\,°C$	9.15
Specific electrical conductivity, $25\,°C$	$3-18 \times 10^{-8}\,\Omega^{-1}\,cm^{-1}$

Chemical Properties. The chemical properties of SO_2Cl_2 are mainly determined by its decomposition into SO_2 and Cl_2. In the absence of catalysts SO_2Cl_2 can be distilled without decomposition. The rate of decomposition is low below $200\,°C$.

SO_2Cl_2 is immiscible with water. The hydrate $SO_2Cl_2 \cdot 15\,H_2O$ is formed with ice-cold water. The strongly exothermic reaction with water, forming sulfuric acid and HCl, is slow at room temperature but becomes explosive at higher temperature or in the presence of catalysts (e.g., I_2). Sulfur dissolves in SO_2Cl_2 without reacting. In the presence of catalysts (e.g., $AlCl_3$) sulfur is chlorinated to S_2Cl_2. Metals generally only react with dry SO_2Cl_2 very slowly at room temperature because of the formation of metal chlorides which are insoluble in SO_2Cl_2. The rate of this reaction is greatly accelerated in the presence of acetonitrile [85]. With ammonia, sulfamide $SO_2(NH_2)_2$ [*7803-58-9*] and imidosulfamide are formed.

Sulfuryl chloride chlorinates, sulfonates, or dehydrates organic compounds. Depending on the reaction conditions and catalysts used the reactions proceed by a radical mechanism (in the presence of peroxides) or by an ionic one (with Lewis acids).

SO_2Cl_2 is particularly suitable for the chlorination of reactive compounds, as it behaves like chlorine of lower reactivity and only forms volatile side products. With tertiary amines SO_2Cl_2 forms stable addition compounds.

Reviews of the reactions of SO_2Cl_2 with organic compounds are given in [86–88].

Production. Sulfuryl chloride is produced almost exclusively by reacting SO_2 and Cl_2 over activated charcoal with cooling [89]. The crude product can then be freed from excess SO_2 or Cl_2 by distillation. Steel or stainless steel are used for the plant equipment, occasionally also lead.

Storage and Transport: as for thionyl chloride (see page 7).

Transport classification:	
IMDG Code	Class 8
UN no.	1834
Packing group	I
RID/ADR	Class 8, no. 21 a
ADNR	Class 8, no. 11 a
ICAP-TI/IATA-DGR	prohibited

In the United States sulfuryl chloride is classed as a material poisonous by inhalation, Zone A. Special regulations apply for packing, labeling, and transport [90].

Quality and Analysis. Technical-grade sulfuryl chloride is a clear, colorless to pale yellow liquid. The sulfuryl chloride content should be min. 99 %; Cl_2 and SO_2 can be present as impurities. The evaporation residue is < 0.01 %. The SO_2Cl_2 content can be found by determining the total chlorine and total sulfur or by GLC.

Waste. In the use of sulfuryl chloride HCl- and/or SO_2- containing off-gas is generally formed, which can be disposed of by scrubbing with alkali. For large quantities of SO_2, recovery as SO_2Cl_2 by reaction with chlorine can be feasible.

Uses. Sulfuryl chloride is mainly used as a mild chlorinating agent, e.g., in the production of chlorophenols or in the side-chain chlorination of aromatics. Other areas of application are the production of sulfonic acids and the sulfochlorination of polyethylene. The use as a liquid cathode in lithium batteries has not become established.

Producers

United States:	Occidental Petroleum
Europe:	Anglo United, United Kingdom; Bayer, Germany; Säurefabrik Schweizerhall, Switzerland.
Asia:	Sumitomo Seika, Japan.

2.2.2. Other Sulfuryl Halides

Sulfuryl chloride fluoride [*13637-84-8*], SO_2ClF, *mp* − 124.7 °C, *bp* 7.1 °C, is a colorless gas that is formed by fluorination of SO_2Cl_2 with SbF_3.

Sulfuryl bromide fluoride [*13536-61-3*], SO_2BrF, *mp* − 86 °C, *bp* 40 °C, is a colorless liquid. It is formed by reaction of SO_2 with BrF_3 and Br_2 [91]. SO_2BrF is quite stable thermally, but is rapidly decomposed by water.

Disulfuryl chloride [*72889-77-1*], pyrosulfuryl chloride, $S_2O_5Cl_2$, is formed by the action of chlorinating agents (S_2Cl_2, $SOCl_2$, PCl_5, CCl_4) on SO_3 or substances that liberate SO_3 [92]. It is a colorless, thermally stable liquid *mp* − 35 °C, *bp* 153 °C, density at 20 °C 1.837 g/cm^3, heat of formation at 18 °C − 697.1 kJ/mol. It decomposes slowly with water, and sulfonates and chlorinates organic compounds. It has not yet achieved any industrial importance.

Polysulfuryl chlorides ($S_nO_{3n-1}Cl_2$) are obtained by reaction of CCl_4 with excess SO_3 with formation of $COCl_2$ as a side product [45], [93].

3. Toxicology

The toxicity of the sulfur halides is based on their hydrolysis, which generates irritant and corrosive substances. The degree of toxicity depends, among other things, on the rate of hydrolytic decomposition. Atmospheric moisture or the moisture of the mucous membranes of the eyes and respiratory tract can be sufficient for this hydrolysis. The results range from irritation of varying degrees to corrosion of the skin, eyes, and respiratory tract. Because of the water solubility of the resulting hydrolysis products, inhalation of the strongly toxic compounds results initially in a so- called upper irritation syndrome as the substances formed dissolve in the

liquid films of the eyes and upper respiratory tract, where they cause irritation or corrosive effects. Inhalation of sufficiently large quantities can cause lung edema because the toxic substances reach the lower respiratory tract after saturation of the liquid film of the upper tract. This severe damage is seldom observed and only when escape from a contaminated atmosphere is impossible (e.g., blocked escape route, unconsciousness). Otherwise the warning action of the irritation of the eyes and upper respiratory tract protects the person from such severe harm. The specific toxic effects arising from inhalation of fluorine or bromine only occur on severe contamination.

Disulfur dichloride, S_2Cl_2, is hydrolyzed by water to give hydrochloric acid, sulfur dioxide, elemental sulfur, and polythionic acids, leading to severe irritation of the eye and respiratory tract. In mice inhalation of 150 ppm (850 mg/m^3) within 1 min is lethal [94]. The LC_{50} (vapor, 4 h) is 441 ppm (2500 mg/m^3) for rats, with an observed NOEL of 32.5 ppm (184 mg/m^3) [95]. The LD_{50} (rat) is 130 mg/kg body weight [95]. Repeated exposure (4 h per day, 25 d) to 0.1 ppm caused effects to the liver and kidneys of rats and rabbits, but not guinea pigs [96]. Corrosive action was detected on the rabbits' ears; even when washed within 30 s, reddening occurred [96]. The hydrolysis product of S_2Cl_2 gave a negative Ames test [95]. In humans 2 – 9 ppm S_2Cl_2 is only moderately irritant [97], but causes inflammation of the mucous membranes of the nose and pharynx [98]. At 10 ppm (56 mg/m^3) symptoms of illness occur after more than a short exposure, and 20 ppm (112 mg/m^3) leads to severely toxic effects within 1 min [99]. After inhalation of high concentrations for 5 h death after 1 d has been reported [94]. The MAK value is 1 ppm or 6 mg/m^3 [100].

Sulfur dichloride, SCl_2, has comparable toxicity to S_2Cl_2 and also liberates HCl. The toxicity of chlorine must also be taken into account (see → Chlorine, Chap. 16.), since SCl_2 is in equilibrium with Cl_2 and S_2Cl_2. The acute LD_{50} (rat) is 293 mg/kg body weight (not stabilized) and 343 mg/kg body weight (stabilized) [101].

Chlorosulfonic acid, $HOSO_2Cl$, is hydrolyzed to hydrochloric and sulfuric acid and therefore has a strongly corrosive and irritant effect on the mucous membranes and skin [102]. The LC_{50} (rat, 4 h) is 38.5 mg/m^3 (mouse, 2 h: 25 mg/m^3). In these animals general toxic effects occur in the range 8.1 – 10.4 mg/m^3 [103].

Thionyl chloride, SO_2Cl, hydrolyzes rapidly with water and atmospheric moisture giving HCl and SO_2. The LC_{50} (rat, 4 h) is 2700 mg/m^3 with an NOEL of 5 mg/m^3, and the LC_{50} (rat, 1 h) is 6200 mg/m^3 [104] or ca. 2500 mg/m^3 [105]. Thionyl chloride is highly irritant/corrosive to the respiratory tract, eyes, and skin [106]. It gives varying results in the Ames test [106]. In humans severely toxic effects occur on more than a short exposure to 10 ppm (50 mg/m^3), and at 20 ppm (100 mg/m^3) within 1 min [99]. Death due to lung edema after 6 min exposure to a calculated maximum of 17 000 ppm SO_2 has been reported [107].

Sulfuryl chloride, SO_2Cl_2, reacts slowly with water at room temperature and vigorously on heating, giving HCl and H_2SO_4. The latter can give rise to severe local irritant and corrosive effects. At higher temperature an equilibrium involving SO_2 and Cl_2 is established, so their toxicity must also be considered (see → Chlorine, Chap. 16.; → Sulfur Dioxide). The LC_{50} (rat, 4 h) is 159 ppm (873 mg/m^3) [108]. Repeated exposure (rat, 2 weeks, 5 d per week, 6 h per day) with 30 ppm (165 mg/m^3), reduced to 20 ppm (110 mg/m^3) after 2 days, led, after 8 d of exposure, to the death of 2 animals; with 10 ppm (55 mg/m^3) an increase in the hemoglobin and red blood cell count, the lung/body weight quotient, and fibrinous-necrotic bronchopneumonia were observed [108]. Sulfuryl chloride is corrosive to the skin [109]. One strain gave a positive result without metabolic activation in the Ames test and three strains were negative [104]. In humans 4 ppm (22 mg/m^3) can lead to symptoms of illness on more than short exposure, and 10 ppm (55 mg/m^3) causes severe toxic effects [99]. A case of lung edema has been reported for sulfuryl chloride [110].

Thionyl bromide, $SOBr_2$, also undergoes hydrolysis with an irritant effect. In humans general symptoms were observed such as somnolence, nausea, and headaches, and increased bronchovascular markings of the lung in X-ray photographs [111], [112]. Symptoms of bromine poisoning should also be taken into account (see → Bromine, Chap. 9.).

4. References

General References

1. *Gmelin,* Syst. Nr. 9, Schwefel, 1963, 1748 – 1830, 1863 – 1875.
2. *Gmelin,* Syst. Nr. 9, Schwefel, Erg. Bd. 1, 1978, Thionylhalogenide.
3. *Gmelin,* Syst. Nr. 9, Schwefel, Erg. Bd. 2, 1978, Schwefelhalogenide.
4. *Ullmann,* 4th ed., **21,** 75 – 85.
5. *Kirk-Othmer,* 3rd. ed., **22,** 122 – 163.

Specific References

6. IUPAC: "Periodic Table," *Chem. Eng. News* 1985, 27.
7. G. H. Whiting, *J. Appl. Chem.* **2** (1952) 381 – 394.
8. P. Walden, *Z. Phys. Chem. (Leipzig)* **65** (1909) 129 – 225.
9. F. Fehér, K. Naused, H. Weber, *Z. Anorg. Allg. Chem.* **290** (1957) 303 – 319.
10. R. I. Rosser, F. R. Whitt, *J. Appl. Chem.* **10** (1960) 229 – 237. C. H. G. Hands, F. R. Whitt, *J. Appl. Chem.* **10** (1960) 237 – 246.
11. T. M. Lowry, G. Jessop, *J. Chem. Soc.* 1930, 782 –792.
12. V. Gutmann, G. Schöber, *Monatsh.* **87** (1956) 792 – 793.
13. E. H. Harvey, H. A. Schuette, *J. Am. Chem. Soc.* **48** (1926) 2065 – 2068.
14. E. B. Bradley, M. S. Mathur, C. A. Frenzel, *J. Chem. Phys.* **47** (1967) 4325 – 4329.
15. S. G. Frankiss, D. J. Harrison, *Spectrochim. Acta Part A* **A 31** (1975) 161 – 167.
16. E. Neale, L. T. D. Williams, *J. Chem. Soc.* 1954, 2156 – 2158.
17. Bayer AG, DIN Safety Data Sheet Sulphur Chloride, Leverkusen 1993.
18. A. C. Batalin, I. A. Shcherbakov, *Zhurn. Obscej. Chim.* **8** (1939) 1394 – 1398; in [1].
19. M. Goehring: *Ergebnisse und Probleme der Chemie der Schwefel-Stickstoff-Verbindungen,* Springer Verlag, Berlin 1957.
20. *Houben-Weyl,* 4th ed., **5 Part 3,** 856 – 895.
21. *Houben-Weyl,* 4th ed., **E 11,** 5 – 30.
22. G. Lewandowicz, A. Zuk, *Chem. Stosow.* **31** (1987) 319 – 330.
23. *Houben-Weyl,* 4th ed., **9,** 40, 74, 90, 136.
24. L. A. Wiles, Z. S. Ariyan, *Chem. Ind. (London)* 1962, 2102 – 2105.
25. F. G. Mann, W. J. Pope, R. H. Vernon, *J. Chem. Soc.* 1921, 634 – 646.
26. K. Yamamoto et al., *J. Chem. Soc. Chem. Comm.* 1991, 596 – 597.
27. K. Miyatake et al., *Polym. Adv. Tech.* **3** (1992) 401 –405.
28. Bayer, DT 2 415 601, 1975; US 3 992 509, 1976 (B. Böhm, K. Velleman).
29. Manufacturing Chemists Assoc., Chem. Safety Data Sheet, SD77, Sulfur Chlorides, Washington 1960.
30. *Kirk-Othmer,* 3rd. ed., **22,** 255 – 266.
31. E. Kühle: *The Chemistry of the Sulfenic Acids,* Thieme Verlag, Stuttgart 1973, pp. 20 – 24, 131 –139.
32. Cabot Corp., DE-AS 1 767 625, 1978 (E. J. Holland, M. G. Fowler, R. B. Roaper).
33. G. A. Takacs, *J. Chem. Eng. Data* **23** (1978) 174 –175.
34. Bayer AG, unpublished results.
35. A. G. Evans, G. W. Meadows, *Trans. Faraday Soc.* **43** (1947) 667 – 674.
36. W. L. Jolly, M. Becke-Goehring, *Inorg. Chem.* **1** (1962) 76 – 78.
37. M. Mühlstädt, D. Martinetz, *Z. Chem.* **14** (1974) 297 – 304.
38. G. A. Tolstikov, *Sulfur Rep.* **3** (1983) 39 – 65.
39. D. Martinetz, *Z. Chem.* **20** (1980) 332 – 335.
40. Hooker, US 3 071 441, 1963 (J. H. Schmadebeck).
41. Hooker, US 3 219 413, 1965 (K. E. Kunkel, D. S. Rosenberg).
42. E. Beckmann, *Z. Phys. Chem. (Leipzig)* **65** (1909) 289 – 337.
43. T. M. Lowry, L. P. McHatton, G. G. Jones, *J. Chem. Soc.* 1927, 746 – 756.
44. R. Steudel, R. Strauss, D. Jensen, *Chem. Ztg.* **109** (1985) 349 – 350.
45. G. Brauer: *Handbuch der Präparativen Anorganischen Chemie,* 3rd ed., vol. 1, Enke Verlag, Stuttgart 1975.
46. O. Ruff, G. Winterfeld, *Ber. Dtsch. Chem. Ges.* **36** (1903) 2437 – 2446.
47. F. Fehér, G. Rempe, *Z. Anorg. Allg. Chem.* **281** (1955) 161 – 168.
48. K. C. Mills: *Thermodynamic Data for Inorganic Sulphides, Selenides and Tellurides,* London 1974.
49. K. Manzel, R. Minkwitz, *Z. Anorg. Allg. Chem.* **441** (1978) 165 – 174.
50. D. K. Padma, *Indian J. Chem.* **12** (1974) 417 – 418.
51. M. Persoz, *Compt. Rend.* **28** (1849) 86 – 88.
52. P. Kremers, *Justus Liebigs Ann. Chem.* **70** (1848) 297 – 300.
53. H. A. Mayes, J. R. Partington, *J. Chem. Soc.* (1926) 2594 – 2605.
54. M. Davis, H. Szkuta, A. J. Krubsack, *Mech. React. Sulfur Comp.* **5** (1970) 1 – 57.

55. K. Arii, *Sci. Rep. Tohoku Ser. I* **22** (1933) 182 – 199; in [1].
56. L. A. Nisel'son, K. V. Tret'yakova, V. Ya. Akhmadeev, *Zh. Neorg. Khim.* **18** (1973) 1092 – 1097; in [2].
57. G. Nagarajan, A. Müller, *Z. Naturforsch. B Anorg. Chem. Org.* **21 B** (1966) 612 – 617.
58. P. W. Schenk, H. Tribel, *Z. Anorg. Allg. Chem.* **229** (1936) 305 – 312.
59. P. W. Schenk, R. Steudel, *Z. Anorg. Allg. Chem.* **342** (1966) 253 – 262.
60. H. B. North, A. M. Hageman, *J. Am. Chem. Soc.* **35** (1913) 352 – 356.
61. M. Goehring, E. Fluck in C. B. Colburn (ed.): *Developments in Inorganic Nitrogen Chemistry,* vol. 1, Elsevier, Amsterdam 1966.
62. P. W. Schenk, *Ber. Dtsch. Chem. Ges.* **75** (1942) 94 –99.
63. V. Y. Kukushkin, *Russ. J. Inorg. Chem. (Engl. Transl.)* **35** (1990) 1273 – 1281.
64. K. Oka, *Synthesis* 1981, 661 – 681.
65. S. S. Pizey: *Synth. Reagents,* **vol. 1,** Horwood, Chichester 1974, pp. 312 – 357.
66. K. H. Bell, L. F. McCaffery, *Aust. J. Chem.* **45** (1992) 1213 – 1224.
67. Pittsburgh Plate Glass Comp., US 2 431 823, 1947 (A. Pechukas).
68. BASF, DE 842 041, 1952 (G. Koudela, K. Wintersberger).
69. Chem. Fabr. Griesheim-Electron, DE 139 455, 1902. W. Majert, DE 136 870, 1901. Farbenfabr. Bayer, DE 279 076, 1913; DE 338 851, 1919.
70. Bayer AG, EP 0 042 530, 1984 (G. Jonas).
71. L. Friedmann, W. P. Wetter, *J. Chem. Soc.* 1967, 36 –37.
72. W. Rigby, *Chem. Ind. (London)* 1969, 1508 – 1509.
73. M. Aubry, B. Gilot, *J. Chromatogr.* **32** (1968) 180 –183.
74. W. Czerwinski, W. Gromotowicz, *J. Chromatogr.* **520** (1990) 163 – 168.
75. G. R. Schlaikjer, F. Goebel, N. Marincic, *J. Electrochem. Soc.* **126** (1979) 513 – 522.
76. K. M. Abraham, R. M. Mank, *J. Electrochem. Soc.* **127** (1980) 2091 – 2096.
77. Chem. Fabr. v. Heyden, DT 752 572, 1954 (E. Haak, R. Jakob).
78. *Ullmann,* 5th ed., **A 3,** 343 – 397.
79. *Kirk-Othmer,* 4th ed., **3,** 1022 – 1028.
80. G. Eichinger, G. Semrau, *Chem. Unserer Zeit* **24** (1990) 32 – 36, 90 – 96.
81. A. Finch, P. J. Gardner, K. Radcliff, *J. Chem. Eng. Data* **13** (1968) 176 – 177.
82. G. Machell, *Chem. Prod.* **19** (1956) 307 – 309, 356 –358.
83. I. Ruppert, *J. Fluorine Chem.* **20** (1982) 75 – 78.
84. S. M. F. Payami, A. R. Prabakaran, S. Gunasekaran, *Orient. J. Chem.* **6** (1990) 233 – 237.
85. A. A. Woolf, *J. Chem. Soc. Dalton Trans.* 1991, 3325 – 3330.
86. I. Tabushi, H. Kitaguchi in S. S. Pizey (ed.): *Synth. Reagents,* **vol. 4,** Horwood, Chichester 1981, pp. 337 – 396.
87. L. Delaude, P. Laszlo, *J. Org. Chem.* 1990, 5260 – 5269.
88. H. C. Brown, *Ind. Eng. Chem.* **36** (1944) 785 – 791.
89. R. H. McKee, C. M. Salls, *Ind. Eng. Chem.* **16** (1924) 279 – 282.
90. US Fed. Regist., Winter Edition 1993, 49 CFR Parts 106 – 180.
91. H. Jonas, H. Stöhr in H. Stamm (ed.): *FIAT Review,* **vol. 23,** p. 191 ff.
92. Du Pont, US 2 530 410, 1950 (M. Sveda).
93. H.-A. Lehmann, G. Ladwig, *Z. Anorg. Allg. Chem.* **284** (1956) 1 – 9.
94. F. Flury, F. Zernik: *Schädliche Gase,* Springer Verlag, Berlin 1931.
95. *Altstoffe, Datensätze,* VCI, Frankfurt 1992.
96. E. Gross, Gewerbehygienisches Laboratorium, IG Farben, Werk Wuppertal-Elberfeld, October 10, 1944.
97. H. B. Elkins: *The Chemistry of Industrial Toxicology,* 2nd ed., Wiley & Sons, New York 1969; cited in[13].
98. S. Horiguchi, K. Shinagawa, *Sangyo Igaku* **17** (1975) 114 – 115; *Chem. Abstr.* **84** (1976) 79 058 r.
99. M. W. Goldblatt, *Br. J. Ind. Med.* **12** (1955) 1 – 20.
100. *MAK- und BAT-Werte Liste 1993,* VCH, Weinheim 1992.
101. Unpublished results, Bayer AG, T 0 011 936, Wuppertal 1982.
102. B. Sörbo: *Pharmacology of Inorganic Sulfur,* Chap. 16, 1972, pp. 143 – 169.
103. N. K. Mamleva, G. Z. Bakhtizina, *Gig. Ta. Okh. Zdorov ýa Rab. Neft. Neftekhim. Promoti* **9** (1976) 110 – 113.
104. *Altstoffe, Datensätze,* 2nd Suppl., VCI, Frankfurt 1991.
105. E. R. Kinkead, R. L. Einhaus, Air Force Aerospace Medical Research Laboratory, Natl Tech Inf Serv, AD-A148-952, Wright-Patterson Air Force Base, Ohio 1984.

106. ACGIH: *Documentation of the Threshold Limit Values and Biological Exposure Indices,* 5th ed., p. 572. ACGIH, Cincinnati 1986.
107. A. M. Ducatman, B. S. Ducatman, J. A. Barnes, *JOM J. Occup Med,* **30** (1988) 309 – 311.
108. D. P. Kelly et al., Du Pont, Haskell Laboratory for Toxicology and Industrial Medicine, Newark, in *Toxicologist* **3** (1983) 62.
109. *The Merck Index,* 11th ed., Merck & Co., Inc., Rahway, N.J., USA, 1989, pp. 1418 – 1419.
110. H. Gerbis, *Angew. Chem.* **44** (1931) 640.
111. J. Jouglard, C. Airaudo, G. Grigorian, G. Monnot, *J. Eur. Toxicol.* **5** (1972) 61 – 63.
112. T. R. Beattie, *Chem. Eng. News* **60** (1982) 5.

Sulfur Hexafluoride → Fluorine Compounds, Inorganic

Sulfuric Acid and Sulfur Trioxide

HERMANN MÜLLER, Lurgi Metallurgie GmbH, Frankfurt/Main, Federal Republic of Germany

1. Introduction 2
2. Properties 3
2.1. Physical Properties 3
2.2. Chemical Properties 7
3. Development of the Sulfuric Acid Industry 9
3.1. Early Development 9
3.2. Further Development of the Nitrogen Oxide Process 9
3.3. Ascendency of the Contact Process 9
3.4. Raw Materials Usage 10
4. Production 11
4.1. Production by Contact Processes 11
4.1.1. Reaction Kinetics and Thermodynamics 11
4.1.2. Catalysts 13
4.1.3. Process Summary 14
4.1.3.1. Gas Drying 15
4.1.3.2. Catalytic Oxidation of Sulfur Dioxide 16
4.1.3.3. Absorption of Sulfur Trioxide . . . 20
4.1.3.4. Acid Cooling 22
4.1.4. Practical Versions of the Contact Process 23
4.1.4.1. Double-Absorption Process Based on Sulfur Combustion 24
4.1.4.2. Double-Absorption Processes Based on Metallurgical Gases 31
4.1.4.3. Ordinary Single-Absorption Processes 34
4.1.4.4. Wet-Catalysis Processes 35
4.1.5. Tail-Gas Treatment 40
4.1.6. Economic Factors 44
4.2. Production by Nitrogen Oxide Processes 45
4.3. Regeneration of Spent Sulfuric Acid 48
4.3.1. Introduction 48
4.3.2. Reconcentration to 70 – 75 % H_2SO_4 50
4.3.3. Concentration to 93 – 98 % H_2SO_4 56
4.4. Production of Oleum 59
5. Construction Materials 60
5.1. Metallic Materials 61
5.2. Nonmetallic Materials 62
6. Uses of Sulfuric Acid and Economic Aspects 63
6.1. Indirect Uses 63
6.2. Direct Uses 63
6.3. Economic Aspects 64
7. Analytical Techniques 64
7.1. Concentration Measurement . . . 64
7.2. Measurement of Impurities 64
7.3. Analysis of Acid-Plant Gas Streams 65
8. Sulfur Trioxide 65
8.1. Properties 65
8.2. Manufacture 67
8.3. Handling and Uses 67
9. Toxicology 68
9.1. Sulfuric Acid 68
9.2. Sulfur Trioxide 68
10. References 68

Based on the corresponding article in Ullmann, 4th ed., written by ULRICH SANDER, ULRICH ROTHE, and ROLF KOLA (English Edition: British Sulphur Corporation, 1984)

1. Introduction

Of all the heavy industrial chemicals, sulfuric acid is perhaps the most fundamentally important, in that it has a number of large-scale uses not only within the chemical industry but in other industries as well. By far the most important user is the phosphate fertilizer industry. Other important applications of sulfuric acid are found in petroleum refining, pigment production, steel pickling, nonferrous metals extraction, and the manufacture of explosives, detergents, plastics, and man-made fibers. Many specialty areas of the chemical industry also use varying amounts of sulfuric acid including the production of dyes, pharmaceuticals, and fluorine chemicals.

The consumption of sulfuric acid has many times been cited as an indicator of the general state of a nation's economy, and although many other indicators (such as energy consumption)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_635

might today be regarded as more important, sulfuric acid consumption still follows general economic trends. For example, the recession that resulted from the "energy crisis" of 1974 was clearly reflected in the pattern of sulfuric acid consumption in the following three years. The recession a few years later was similarly accompanied by a generally declining trend in sulfuric acid consumption starting in mid-1980.

Sulfuric acid is manufactured from sulfur dioxide. The primary raw material for producing this intermediate is elemental sulfur (→ Sulfur ; → Sulfur Dioxide). A large involuntary producer of sulfur dioxide is the nonferrous metals industry, the roasting and smelting processes of which generate off-gases with a sufficiently high concentration of sulfur dioxide to permit direct processing to sulfuric acid as a by-product (→ Sulfur Dioxide, Chap. 5.5.).

A significant number of sulfuric acid facilities have been installed in metallurgical plants for the recovery of SO_2, mainly for environmental reasons. To an increasing extent the acid produced in such plants is replacing acid formerly obtained from elemental sulfur (or pyrite). An important task for the years ahead is to establish an infrastructure for distributing this "metallurgical acid" to consumers, especially in the fertilizer industry.

Pyrite still serves in several countries as a raw material for sulfuric acid production, but production rates are no longer significant compared to the other sources described.

Sulfur dioxide is also produced in the thermal decomposition stage of sulfuric acid regeneration from the heavily contaminated sulfuric acid ("spent acid") created by certain industrial processes in which sulfuric acid is introduced(→Sulfur Dioxide, Chap. 5.6.1.). A further source of sulfur dioxide for sulfuric acid production is the combustion of waste gases containing hydrogen sulfide (→ Sulfur Dioxide, Chap. 5.7.).

In theory, there is enough sulfur dioxide to supply all the world's sulfuric acid needs in exhaust gases released during the combustion of sulfur-containing coal, fuel oil, and other fossil fuels in power stations and large industrial boilers. Although it is possible to recover such SO_2 in a concentrated form by regenerative flue-gas desulfurization processes, or to convert it directly into H_2SO_4, only a minor amount of sulfuric acid is prepared by this route. Numerous flue-gas desulfurization plants installed recently as a way of complying with environmental regulations instead convert the SO_2 into gypsum, primarily for economic reasons. Unless environmental regulations require desulfurization of combustion gases, flue-gas sulfur dioxide is simply discharged into the atmosphere – where, ironically, it is then transformed by atmospheric processes into sulfuric acid, which ultimately returns to earth dissolved in rainfall.

Essentially all sulfuric acid today is manufactured by the contact process. The corresponding technology is very mature, although important alterations in the detailed arrangement of the conversion and absorption stages were introduced commercially in the 1960s to increase the sulfur dioxide conversion efficiency, primarily in the interest of environmental protection. Some plants are now designed for a sulfur dioxide conversion efficiency exceeding 99.8 %. Nevertheless, the basic principle of the process remains the same today as when it was first introduced in the 1930s.

Modern plants can be designed to be extremely efficient in terms not only of sulfur dioxide conversion but also energy recovery. It has been common practice for many years to recover 60 % or more of the total energy released in sulfuric acid production in the form of high-pressure steam. The energy efficiency of some plants has recently been very substantially increased, however, by providing for additional recovery of low-level heat from the acid system.

In the attempt to comply with requirements for existing metallurgical complexes, sulfuric acid plants have been developed and constructed for treating gases containing as little as 2 % SO_2 (in a single-catalysis plant; 5 % in a double-catalysis plant – see Section 4.1.4). A current objective is to process sulfur dioxide gas streams with the highest SO_2 concentrations possible in the interest of reducing capital and operating costs and increasing further the extent of high-temperature energy recovery. At the same time, quality requirements have become more strict for commercial acid. Given the large quantities of "metallurgical acid" now entering the market it will be necessary to develop new processes for eliminating such impurities as may still be present.

Yet another important challenge facing sulfuric acid makers is the reprocessing of acid wastes from user industries. It is a peculiarity of sulfuric acid that very little of the material actually ends up in the products it is used to make. Indeed, apart from small-volume products such as pharmaceuticals, almost the only end products with sulfur values comparable to sulfuric acid are synthetic detergents – and even they are ultimately destined for the drain. Many uses result in so-called *spent acid:* sulfuric acid in varying states of dilution as well as contamination with organic and inorganic impurities. Environmental authorities are becoming increasingly intolerant of the sometimes rather casual waste-disposal methods of the past, and the industry is under pressure to accept greater amounts of used acid for reconcentration, purification, or regeneration. Although the technology in this field is already quite extensive (see Section 4.3), further development can be expected in the use of recovered heat from sulfuric acid plants.

2. Properties

2.1. Physical Properties

Pure sulfuric acid [*7664-93-9*], H_2SO_4, M_r 98.08, is a colorless, water-white, slightly viscous liquid, *mp* 10.4 °C, *bp* 279.6 °C, and d_4^{15} 1.8356. It can be mixed with water in any ratio. Aqueous sulfuric acid solutions are defined by their H_2SO_4 content in weight-percent terms. Anhydrous (100 %) sulfuric acid is even today sometimes referred to as "monohydrate," which simply means that it is the monohydrate of sulfur trioxide. Sulfuric acid will dissolve any quantity of SO_3, forming oleum ("fuming sulfuric acid"). The concentration of oleum is expressed in weight-percent of dissolved SO_3 ("free SO_3") in 100 % H_2SO_4.

The physical properties of sulfuric acid and oleum [1–3] are dependent on the H_2SO_4 and SO_3 concentrations, the temperature, and the pressure. Figure 1 shows the *densities* of sulfuric acid and oleum as a function of temperature and concentration [4]. At constant temperature, the density of sulfuric acid increases steeply with rising H_2SO_4 concentration, reaching a maximum at about 98 %. From there up to a concentration of 100 % the density decreases slightly, but it rises again in the oleum range up to a concentration of ca. 60 % free SO_3.

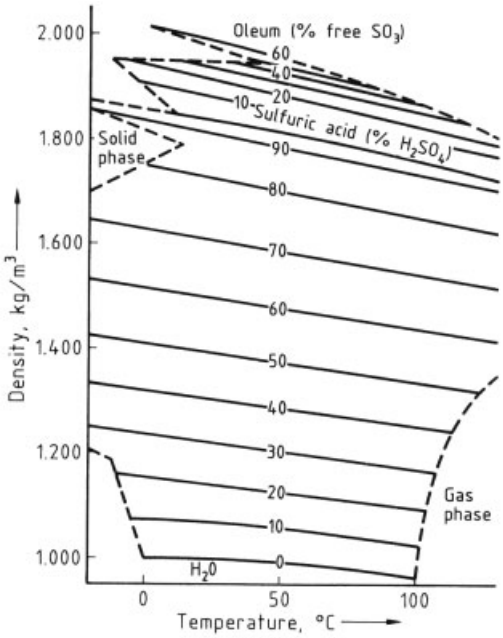

Figure 1. Oleum and sulfuric acid density as a function of temperature and concentration [4]

On account of the clear relationship between density and concentration at defined temperatures in the lower concentration range, density measurement provides a quick method for determining concentration up to about 95 % H_2SO_4.

Hydrometers used for this purpose were formerly calibrated in "degrees Baumé" (°Bé), and for that reason sulfuric acid concentration was often, and sometimes still is, expressed in °Bé. The density ρ in g/cm^3 is given by the expressions

$$r = 144.3/\left(144.3 - {}^{\circ}\mathrm{B}\right)^{[d12]}_{e} \text{ (Europe)}$$

$$r = 145/\left(145 - {}^{\circ}\mathrm{B}\right)^{[d12]}_{e} \text{ (United States)}$$

The *electrical conductivity* of sulfuric acid at 20 °C as a function of concentration is shown in Figure 2 (see next page). The peculiar shape of the curve is due to the various states of ionic dissociation present in the system $H_2O/H_2SO_4/SO_3$ at different concentrations. Conductivity measurement is thus also

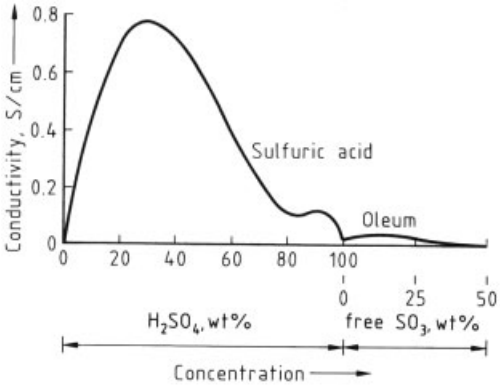

Figure 2. Electrical conductivity of sulfuric acid and oleum at 20°C

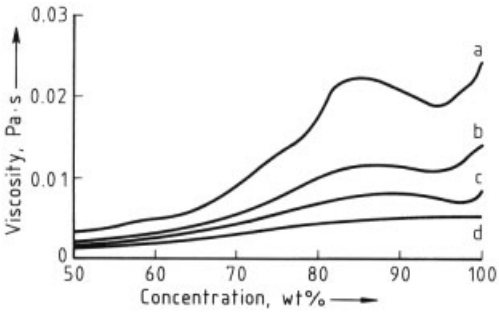

Figure 3. Dynamic viscosity of sulfuric acid [5]
a) At 25°C; b) At 45°C; c) At 60°C; d) At 80°C

useful as a method for determining the concentration of sulfuric acid.

The *dynamic viscosity* of sulfuric acid as a function of concentration at various temperatures is shown in Figure 3 [5].

The various maxima and minima observed on the *freezing point* curve of sulfuric acid and oleum, shown in Figure 4 [6], are due to the existence of different sulfuric acid hydrates at different temperatures. Whereas the 98.0 – 98.5 % acid used for sulfur trioxide absorption in sulfuric acid production (Section 4.1.3) solidifies at about − 1 to + 1 °C, commercial 96 % sulfuric acid (66°Bé) solidifies at ca. − 15 °C. Commercial 66°Bé acid in the United States contains only 93.2 wt % H_2SO_4 and has a freezing point of ca. − 29 °C. This behavior is obviously of importance if sulfuric acid is to be stored or transported under very cold conditions.

Figure 5 is a phase diagram for aqueous solutions of sulfuric acid based on boiling point measurements by Haase and Rehse [7]. The

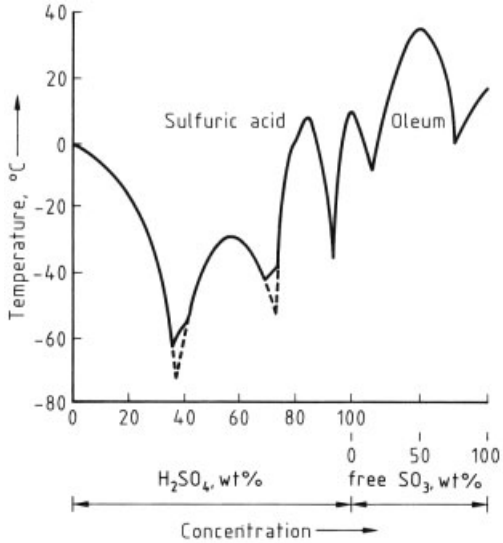

Figure 4. Freezing point curve for sulfuric acid and oleum [6]

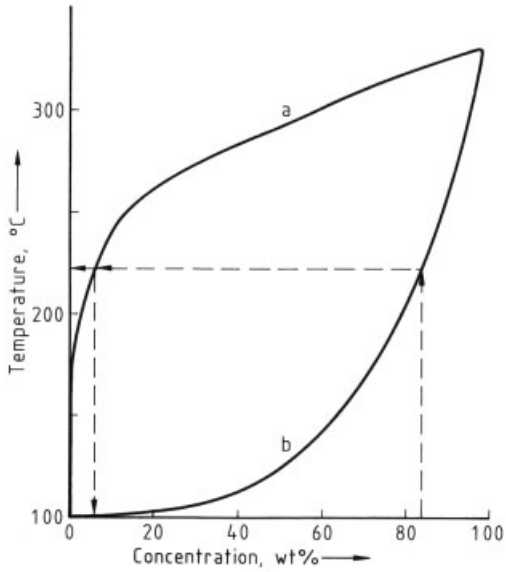

Figure 5. Boiling curves for sulfuric acid at 1013 mbar [7]
a) Vapor; b) Liquid

lower curve shows the relationship between the boiling point of sulfuric acid and its concentration, while the upper curve shows the sulfuric acid concentration in vapor evolved from acid boiling at a particular temperature. When an aqueous solution of sulfuric acid is boiled, the vapor contains more water than the boiling acid, so the concentration of the acid in-

creases and its boiling point rises. This continues until the boiling point reaches a maximum of ca. 339 °C at a sulfuric acid concentration of 98.3 wt % H_2SO_4. At this point the liquid and vapor phases are identical in composition, corresponding to an azeotrope, so the concentration of the boiling acid cannot increase further. As can be seen from Figure 5, if vapor in equilibrium with sulfuric acid of 85 wt %, boiling at about 223 °C, is completely condensed it will contain about 7 wt % H_2SO_4. At concentrations below ca. 75 wt % H_2SO_4, essentially nothing but water evaporates. The boiling behavior of sulfuric acid is especially important with respect to industrial processes for thermal concentration of dilute acid. As noted above, the azeotropic concentration (98.3 wt % H_2SO_4) represents the ultimate limiting concentration that can be reached by this method.

The vapor above concentrated sulfuric acid containing more than 98.3 wt % H_2SO_4 includes not only a greater proportion of H_2SO_4 than the liquid but also considerable quantities of SO_3. The vapor over oleum consists almost entirely of SO_3. Figure 6 shows the equilibrium vapor pressures of H_2O, H_2SO_4, and SO_3 above sulfuric acid at 60 °C in the concentration range from 85 wt % H_2SO_4 to 15 % SO_3 oleum, based on measurements made by Lucinskij [8]. These vapor-pressure curves are all-important in gas drying and SO_3 absorption, essential steps in the production of concentrated sulfuric acid by the contact process (Section 4.1.3).

Apart from the field of sulfuric acid manufacture, the system $H_2O/SO_3/H_2SO_4$ is also of immense importance in connection with the combustion of sulfur-containing fuels. It is essential to ensure that the temperature of the combustion gas does not drop below the dewpoint [9, 10] prior to discharge, since otherwise there is a danger of corrosion by condensing sulfuric acid. Various formulae have been developed [11–13] for calculating dewpoints theoretically as a function of the total gas pressure as well as the H_2O, SO_3, and H_2SO_4 partial pressures [14]. Unfortunately, the situation is further complicated by the possible formation of various types of hydrated and associated molecules in the gas phase [15].

The *specific heat* of sulfuric acid falls as concentration increases. Figure 7 shows the dependence of specific heat on concentration and temperature [4]. The *standard enthalpy of formation* for pure liquid H_2SO_4 is − 8.305 kJ/kg, and the *latent heat of evaporation* at the boiling point is ca. 605 kJ/kg [16]. Figure 8 (see next page) shows the enthalpies of liquid sulfuric acid and oleum over a wide concentration range at temperatures between 0 °C and the boiling point [17], assuming an arbitrary value of 0 kJ/kg as the enthalpy of pure water at 0 °C. This diagram

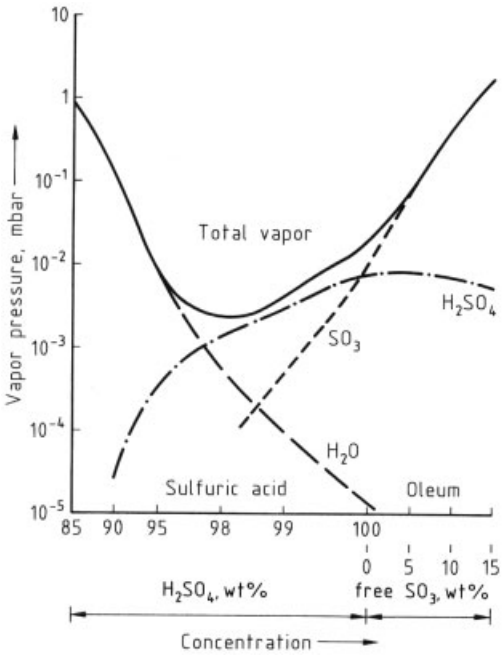

Figure 6. Equilibrium vapor pressures over sulfuric acid and oleum [8]

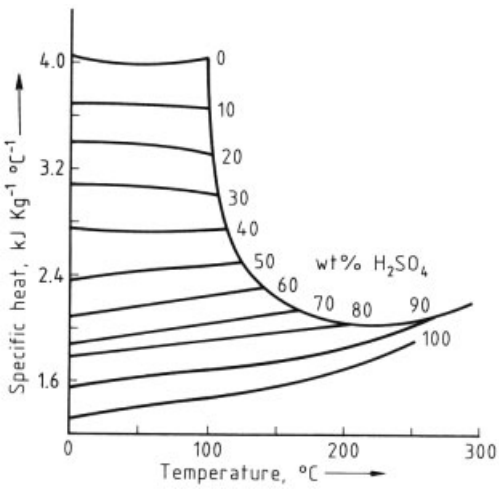

Figure 7. Specific heat of sulfuric acid [4]

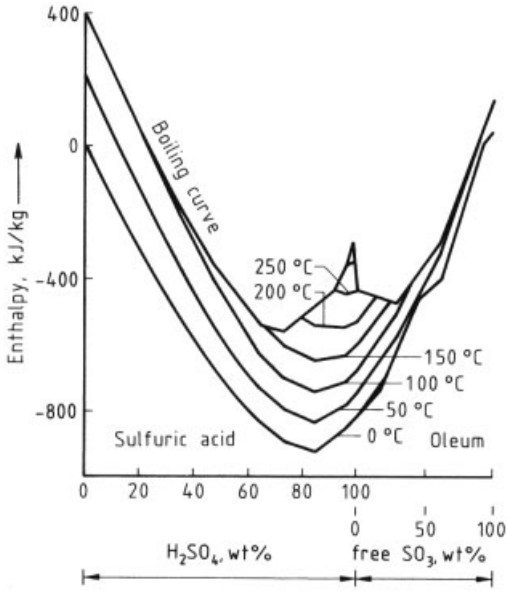

Figure 8. Enthalpy diagram for sulfuric acid and oleum [17]

provides a simple method for determining the amount of heat liberated when sulfuric acid or oleum is diluted from one concentration to another by addition of water.

The amount of heat produced by diluting concentrated sulfuric acid with water is considerable, so rapid mixing is important to ensure that local overheating and boiling are avoided. If one wishes to concentrate dilute acid, a corresponding amount of heat, the so-called *heat of dehydration*, must be supplied in addition to such heat as may be required to evaporate the water. Figure 9, which has been calculated from enthalpy values, shows the heat of hydration evolved in the dilution of 98.3 % acid to lower concentrations at 20 °C (or, conversely, the amount of heat theoretically required to dehydrate it from lower concentrations up to 98.3 %). The heat of hydration liberated by diluting between intermediate concentrations – from 75 % H_2SO_4 to 25 %, for example – is simply the difference between values read off the curve opposite the appropriate concentrations.

2.2. Chemical Properties

Sulfuric acid is a strong acid with characteristic hygroscopic and oxidizing properties. Sulfuric acid, like the sulfate ion, is chemically and

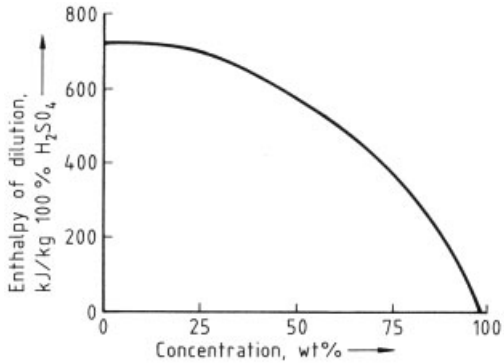

Figure 9. Heat of dilution or dehydration of sulfuric acid at 20°C

thermally very stable. The dehydrating effect of concentrated sulfuric acid is due to the formation of hydrates. Several hydrates have been identified in the solid state, and these explain the irregular variation of some of the physical properties of sulfuric acid with concentration, such as its freezing temperature (see Fig. 4). Known hydrates are $H_2SO_4 \cdot H_2O$ (corresponding to 84.5 wt % H_2SO_4); $H_2SO_4 \cdot 2\,H_2O$ (71.3 wt % H_2SO_4); $H_2SO_4 \cdot 3\,H_2O$ (64.5 wt % H_2SO_4); $H_2SO_4 \cdot 4\,H_2O$ (57.6 wt % H_2SO_4); and $H_2SO_4 \cdot 6\,H_2O$ (47.6 wt % H_2SO_4) [1].

Pure sulfuric acid is ionized to only a small extent as expressed by Equations (1) and (2) [1].

$$2\,H_2SO_4 \rightleftharpoons H_3SO_4^+ + HSO_4^- \quad (1)$$

$$2\,H_2SO_4 \rightleftharpoons H_3O^+ + HS_2O_7^- \quad (2)$$

This is the reason why the electrical conductivity of a sulfuric acid solution has its lowest value at 100 % H_2SO_4 (see Fig. 2). When pure sulfuric acid is diluted with water, dissociation occurs increasingly according to Equation (3).

$$H_2SO_4 + H_2O \rightleftharpoons H_3O^+ + HSO_4^- \quad (3)$$

The conductivity rises accordingly. Between 92 wt % and 84.5 wt % H_2SO_4, the monohydrate ($H_2SO_4 \cdot H_2O$) exists preferentially in equilibrium with the ionic species, so the conductivity decreases slightly. At lower H_2SO_4 concentrations the extent of dissociation increases, as does therefore the conductivity.

At high water content the second stage of dissociation becomes increasingly immportant (Eq. 4).

$$HSO_4^- + H_2O \rightleftharpoons H_3O^+ + SO_4^{2-} \quad (4)$$

On account of the diminishing total concentration of sulfuric acid, however, the conductivity reaches a maximum at ca. 30 wt % H_2SO_4 (the exact value depends on the temperature), and decreases steeply down to 0 wt % H_2SO_4 [1]. Dilute sulfuric acid is the preferred electrolyte for industrial metal electrowinning and electroplating plants on account of its high conductivity and the chemical stability of the sulfate ion. To take advantage of the electrical conductivity maximum, sulfuric acid of about 33 % concentration is used in lead storage batteries (→ Batteries, Chap. 4.).

Dilute sulfuric acid is a strong dibasic acid, so it will dissolve all base metals. Hydrogen is released, and the respective metal sulfates and bisulfates (hydrogensulfates) are formed. Barium and lead are exceptions, not because they do not react in the first place but because they become coated with an insoluble sulfate layer that protects them from further attack by the acid [1]. Hot, concentrated sulfuric acid has an oxidizing effect, reacting with precious metals and with carbon, phosphorus, and sulfur, by which it is reduced to sulfur dioxide.

A very important property of sulfuric acid is its ability to decompose the salts of most other acids. Examples of industrial importance include:

1) Production of sodium sulfate and hydrogen chloride from sodium chloride (→ Hydrochloric Acid, Chap. 3.1.; → Sodium Sulfates, Chap. 1.3.4.)
2) Decomposition of sulfites to sulfur dioxide
3) Decomposition of phosphate rock (natural calcium phosphates) to phosphoric acid and calcium sulfate (→ Phosphoric Acid and Phosphates, Chap. 1.2.2.).

The reactions of concentrated sulfuric acid with organic compounds are frequently dominated by its oxidizing and hygroscopic properties [1]. Carbohydrates, for example, are decomposed to the point of carbonization. Organic condensation reactions in which water is eliminated are promoted by sulfuric acid because it effectively removes the water as soon as it is formed. Sulfuric acid is therefore frequently used in industry for this purpose. It also exercises a catalytic effect on certain reactions involving organic compounds.

Sulfuric acid is thermally extremely stable. Only at very high temperatures is it partially decomposed into its anhydride, sulfur trioxide, and water vapor (Eq. 5).

$$H_2SO_4 \rightleftharpoons SO_3 + H_2O \tag{5}$$

The reverse of this reaction is the overall route by which sulfuric acid is formed in the absorption section of a contact sulfuric acid plant. However, it is not possible in practice to manu-facture sulfuric acid by absorbing sulfur trioxide directly into water, because the sulfur trioxide reacts with water vapor in equilibrium with the liquid near the surface, initially forming sulfuric acid vapor. This quickly condenses as a mist of very fine (submicron) droplets, which are practically impossible to collect. However, sulfuric acid itself reacts readily with sulfur trioxide to form disulfuric acid (Eq. 6), which can be converted back to sulfuric acid by reaction with water (Eq. 7).

$$H_2SO_4 + SO_3 \rightleftharpoons H_2S_2O_7 \tag{6}$$

$$H_2S_2O_7 + H_2O \rightleftharpoons 2\,H_2SO_4 \tag{7}$$

It is therefore quite feasible to absorb sulfur trioxide in sulfuric acid of 98 % or higher concentration, over which the partial pressure of water vapor is very low, thus avoiding the problem of mist. However, because the vapor pressures of both H_2SO_4 and SO_3 increase steeply at concentrations above 99 % H_2SO_4, the optimum sulfur trioxide absorption efficiency is achieved in 98 – 99 % acid.

If sulfur trioxide is produced in a gas stream that already contains moisture, gaseous sulfuric acid forms progressively by reaction (5), the thermodynamic equilibrium point of which is shifted towards H_2SO_4 with decreasing temperature [18]. The resulting sulfuric acid can then be condensed in a controlled manner without significant mist production. This route to sulfuric acid is also exploited in industry.

The vapor pressure of sulfur trioxide over oleum or disulfuric acid is appreciable, so when oleum is exposed to ambient air, which always contains moisture, sulfuric acid mists invariably form. It is this property that gives oleum its familiar name of fuming sulfuric acid.

Pure disulfuric acid, $H_2S_2O_7$, which corresponds theoretically to oleum with 44.9 wt % free SO_3, crystallizes at ca. 35 °C, the maximum of the freezing-point curve in the oleum range

(see Fig. 4). Disulfuric acid is partially dissociated in sulfuric acid solution (Eq. 8).

$$H_2S_2O_7 + H_2SO_4 \rightleftharpoons H_3SO_4^+ + HS_2O_7^- \quad (8)$$

It is for this reason that the electrical conductivity of oleum rises slightly as the SO_3 concentration is increased from 0 to about 10 wt % free SO_3.

Sulfuric acid is oxidized both by hydrogen peroxide and anodically to diperoxysulfuric acid, $H_2S_2O_8$, and the unstable monoperoxysulfuric acid (Caro's acid), H_2SO_5 (Eq. 9).

$$H_2O_2 + H_2SO_4 \rightleftharpoons H_2SO_5 + H_2O \quad (9)$$

Since it is a strong oxidant, Caro's acid can oxidize sulfur dioxide to sulfuric acid, a property that has been exploited in pollution control for sulfuric acid plants (the Peracidox process, Section 4.1.5).

Nitrogen oxides ($NO + NO_2$) react with sulfuric acid at concentrations above 70 wt % H_2SO_4 to give nitrosyl hydrogensulfate, $NOHSO_4$ (see Section 4.2).

3. Development of the Sulfuric Acid Industry

Detailed descriptions of the development of sulfuric acid production procedures can be found in the literature [19–23].

3.1. Early Development

In the late Middle Ages, sulfuric acid was obtained in small quantities in glass vessels in which sulfur was burned with saltpeter in a moist atmosphere. Higher rates of production first became possible with the introduction of *lead chambers* as reaction vessels by Roebuck, in Birmingham, England, in 1746. The next major step forward came in 1793, when Clement and Desormes achieved better results by introducing supplemental air into the lead chamber process. They interpreted this as meaning that the nitrous gases were acting only as facilitators of the process, and that the oxidation itself was being effected by the air (i.e., oxygen).

In 1827, Gay-Lussac introduced a method for absorbing nitrogen oxides from the lead-chamber off-gases. With the further development by Glover in 1859 of a method for recovering nitrogen oxides from the newly formed acid by stripping with incoming hot gases, it became possible to make the nitrogen oxide-catalyzed process continuous.

As early as 1831, Phillips, in Bristol, England, had patented the oxidation of sulfur dioxide to sulfur trioxide over a platinum catalyst at high temperature. Nevertheless, it was only after oleum demand for dye manufacture began to increase – from about 1872 onward – that this invention was adopted by industry, and intensive development of the *contact process* began. Better solid catalysts were then sought, and the chemistry and thermodynamics of the SO_2/SO_3 equilibrium were investigated. Systematic studies undertaken by Knietsch at BASF on the reaction equilibrium of SO_2 oxidation over a platinum catalyst, published in 1901 [24], formed an important basis for an understanding of thermodynamic principles.

3.2. Further Development of the Nitrogen Oxide Process

The growth in popularity of the contact process stimulated new competitive efforts to improve the lead chamber process. The most significant development was replacement of the lead chambers themselves with acid-irrigated towers, which substantially reduced the specific space requirements. The first so-called tower plant was built by Opl in 1907. However, widespread industrial use of the tower process had to await the development of adequate acid pumps. In 1923, Petersen introduced an improved tower process that remaine competitive with the contact process up to the 1950s. However, a fundamental disadvantage of the nitrogen oxide processes is that product concentration is limited to a maximum of 70 – 75 %, while the contact process produces concentrated (98 %) acid. With the development of relatively inexpensive vanadium catalysts for the contact process together with increasing demand for concentrated sulfuric acid, the proportionate share of world sulfuric acid output produced in nitrogen oxide process plants declined steadily. Around 1910, these accounted

for ca. 80 % of production in Western Europe and North America. By 1930 the figure had shrunk to about 73 %, by 1950 to 20 – 25 %, and by 1960 to ca. 15 % [23]; in 1980 virtually no acid was being produced in nitrogen oxide process plants in these parts of the world. Nonetheless, the nitrogen oxide process has continued to be the object of interest and a certain amount of development work, especially for the processing of gases with extremely low SO_2 content (see Section 4.2).

3.3. Ascendency of the Contact Process

Platinum remained the predominant catalyst for the contact process until the 1930s. As early as 1913, however, BASF was granted a patent [25] for a catalyst based on vanadium pentoxide, which eventually succeeded in replacing the platinum catalyst because of its insensitivity to catalyst poisons and its considerably lower cost.

In 1936/37 Lurgi introduced the wet contact process for converting moist sulfur dioxide-containing gases over a vanadium catalyst. This made it possible to process hot gases from the combustion of hydrogen sulfide in coking plants directly to sulfuric acid.

In succeeding years a number of factors affected the development of the contact process. First, the raw material basis of the industry changed progressively from mainly roaster gases to sulfur combustion gases containing higher concentrations of sulfur dioxide. Second, plant capacities increased as a result of a great rise in the consumption of sulfuric acid by the fertilizer industry. These and other factors provided a stimulus for the introduction of improvements in the individual process steps and in the design of associated equipment (e.g., the shift to tray converters from tube converters).

Double Absorption. In 1960, a patent application was filed by Bayer [26] for the so-called double-catalysis process, and the first plant using this process, built by Lurgi, started up in 1964 [27]. By incorporating a preliminary SO_3 absorption step ahead of the final catalytic stages, the improved contact process permitted a decisive increase in overall SO_2 conversion, thus reducing SO_2 emissions substantially. Because the essential difference between this process and the ordinary contact process is in the number of absorption stages, itis referred to hereafter as the "double-absorption" process.

Environment and Energy. In the 1970s the principal industrial countries introduced more stringent regulations for environmental protection, which made the use of the double-absorption process more or less mandatory in new plants. Nevertheless, the conventional contact process continues to be used in countries where environmental regulations are less exacting.

On account of the steep rise in energy costs, the main thrust of current development in the contact process is toward increasing the recovery and utilization of the very substantial amount of process heat. Indeed, a large, modern sulfuric acid plant may be looked upon not just as a chemical plant but also as a thermal power plant [28–32].

3.4. Raw Materials Usage

The principal starting material for sulfuric acid production is sulfur dioxide, which can be obtained by different methods from various raw materials (→ Sulfur Dioxide, Chap. 3.). Moderately concentrated sulfuric acid is also produced by reconcentration and purification of so-called spent or waste sulfuric acid. Reprocessing spent acids, which are generated in large quantities in many processes, and recycling the regenerated acid to the user is becoming increasingly important from an environmental protection standpoint especially in the major industrial countries [33].

Up about 1970, pyrite was the predominant raw material, maintaining a 57 – 62 % share of the continually rising total. Thus, in 1962, out of a total of ca. 3.1×10^6 t of sulfuric acid, about 1.9×10^6 t (62 %) was made from pyrite; in 1970, about 2.7×10^6 t was pyrite-based, corresponding to ca. 61 % of the total production of ca. 4.4×10^6 t [34]. Since 1970, the proportionate role of pyrite has been declining relative to the total, which has continued to rise. In 1979, pyrite accounted for only ca. 24 %.

Since the beginning of the 1970s, production of sulfuric acid from elemental sulfur in the Federal Republic of Germany has grown faster than total acid production, in step with rapid growth

in the production of recovered sulfur from the refining of crude oil or purification of natural gas from the gas fields of Northern Germany. The proportion of total sulfuric acid production derived from elemental sulfur increased from ca. 30 % in 1970 to ca. 50 % in 1979, when total production stood at about 5×10^6 t [34].

Sulfuric acid manufacture based on elemental sulfur and pyrite is, of course, relatively sensitive to market conditions, because acid produced from these materials represents a primary product. The same is not true of sulfuric acid production based on any of the other sulfur-containing raw materials. In those cases, sulfuric acid is a secondary product, manufactured as a means of disposing of waste from another process. The level of production is therefore dictated not by conditions in the sulfuric acid market, but by conditions in the market for the primary product. Typical sources of this so-called fatal acid are sulfuric acid plants associated with nonferrous metal smelters processing sulfide ores. The quantity of fatal acid produced in the Federal Republic of Germany has been rising constantly, reaching ca. 1.3×10^6 t in 1979 (ca. 26 % of the total).

The raw material basis for sulfuric acid production in other European countries has followed a pattern generally similar to that in the Federal Republic of Germany. Pyrite remained the dominant raw material until the 1950s, and it was only with the advent of large quantities of recovered sulfur from the Lacq natural gas fields in France, and later from natural gas operations in Canada, that elemental sulfur assumed its current predominance.

In contrast, the United States industry has been based since the early years of this century on elemental sulfur because of development of the Frasch industry in the Gulf states (→ Sulfur, Chap. 5.5.). In 1978, the various raw materials accounted for the following approximate proportions of the total U.S. sulfuric acid production of ca. 36×10^6 t [35] in 1963 and of ca. $280\text{x}10^6$ t in 1993:

Elemental sulfur	79 %
Nonferrous sulfide ores	9 %
Waste acid	5 %
Pyrite and hydrogen sulfide	3 %
Miscellaneous	4 %

4. Production

4.1. Production by Contact Processes

4.1.1. Reaction Kinetics and Thermodynamics

In the contact process, a gas mixture containing sulfur dioxide is passed together with oxygen over a catalyst to oxidize the sulfur dioxide to sulfur trioxide (Eq. 10) [36]:

$$SO_2 + 1/2\,O_2 \rightleftharpoons SO_3 \quad \Delta H^\circ = -99.0\,\text{kJ} \tag{10}$$

The sulfur trioxide is then absorbed in sulfuric acid where it reacts with added water to form more sulfuric acid (Eq. 11).

$$SO_3\,(g) + H_2O\,(l) \rightleftharpoons H_2SO_4\,(l) \quad \Delta H^\circ = -132.5\,\text{kJ} \tag{11}$$

The position of equilibrium in the gas-phase exothermic oxidation of sulfur dioxide to sulfur trioxide (Eq. 10) depends on the prevailing temperature, total pressure, and concentrations (partial pressures) of the reactants. Thermodynamic equilibrium is determined by the equilibrium constant K_p according to the law of mass action:4

$$K_p = \frac{p\,(SO_3)}{p\,(SO_2)\cdot p\,(O_2)^{0.5}}$$

On account of the negative reaction enthalpy of sulfur dioxide oxidation, both K_p and the SO_2 equilibrium conversion decrease with rising temperature. The classical relationship between K_p (in atm $^{-0.5}$) and temperature was developed empirically by Bodenstein and Pohl [37], although more recent data [16, 38, 39] show deviations.

$$\log K_p = \frac{5186.5}{T} + 0.611\log T - 6.75$$

where T= temperatue in K

Influences on the Conversion Equilibrium. Any increase in the overall pressure will increase the extent of conversion at equilibrium, because the reaction leads to a decrease in volume.

The maximum possible equilibrium sulfur dioxide conversion at a given temperature T and total pressure p depends upon the SO_2 and O_2 concentrations of the supplied gases. If the sulfur

dioxide concentration is $2a$ vol % and the oxygen concentration b vol %, the fraction x of sulfur dioxide oxidized to sulfur trioxide at equilibrium can be calculated from the following equation based on the law of mass action:

$$K_p = \frac{1}{\sqrt{p}} \frac{x}{1-x} \sqrt{\frac{100-ax}{b-ax}}$$

The appropriate value of K_p is determined from the equation of Bodenstein and Pohl. The dependence of SO_2 equilibrium conversion on temperature and pressure is shown in Figure 10 for the example of sulfur combustion gas containing 10 vol % SO_2.

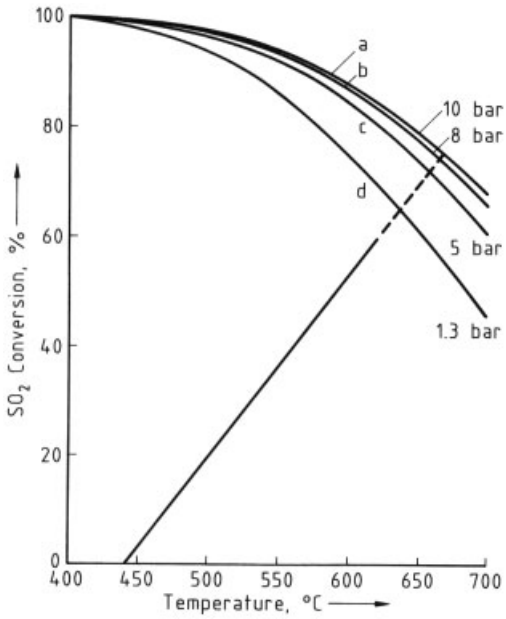

Figure 10. Theoretical conversion equilibrium in the oxidation of SO_2 to SO_3 as a function of temperature and pressure (feed gas composition: 10 vol % SO_2, 10.9 vol % O_2)
a) 10 bar; b) 8 bar; c) 5 bar; d) 1.3 bar

In accordance with the law of mass action, increasing the oxygen partial pressure will also increase the degree of conversion. However, when air alone is used as the source of oxygen in SO_2 production, as is usually the case, the oxygen and sulfur dioxide concentrations are inversely related, so the greater the oxygen concentration in the combustion gases the lower will be the sulfur dioxide content. The essential factor determining the attainable SO_2 conversion is thus the volumetric O_2/SO_2 ratio in the feed gases. Whereas sulfur dioxide oxidation requires a stoichiometric O_2/SO_2 ratio of only 0.5: 1, in industry it is normal practice to use a ratio of at least 1: 1. The presence of excess oxygen not only raises the SO_2 equilibrium conversion but is also an essential prerequisite for maintaining the activity of the vanadium catalyst. There are, however, practical limits on the amount of extra air that can be added, because nitrogen present in the air dilutes the sulfur dioxide to the point where the economics of the process are impaired. Although it would be technically feasible to avoid nitrogen dilution by using oxygen instead of air, as is sometimes done in pyrometallurgical processes that produce high-strength byproduct sulfur dioxide gas streams, it is usually difficult to justify in a sulfur-burning installation [40].

Another practical expedient for improving the conversion of sulfur dioxide to sulfur trioxide is to remove, at an intermediate stage in the process, the sulfur trioxide already formed. In a double-absorption type sulfuric acid plant this is accomplished by routing the reaction gases after two of three stages of catalytic conversion through an intermediate absorption stage and then through one or two subsequent catalytic conversion stages. Because of the large (100 %) stoichiometric oxygen excess in the original feed gas and the diminished sulfur dioxide concentration, the O_2/SO_2 ratio at this point is about six times more favorable than at the start.

Influences on Reaction Rate. In an industrial plant, actual sulfur dioxide conversion never reaches the theoretical equilibrium value. Gas-phase oxidation of sulfur dioxide is kinetically inhibited, and virtually impossible at any temperature without a catalyst. At ordinary temperatures the reaction is so slow that, in practical terms, it does not occur at all. Increasing the temperature increases the rate of reaction, but simultaneously shifts the position of equilibrium in an unfavorable way – away from sulfur trioxide and toward sulfur dioxide and oxygen. Without a catalyst, the temperature required to make the system react at a practical rate is so high that conversion is very poor. Even with present-day catalysts, a temperature of ca. 400 °C is necessary to initiate a self-sustaining reaction.

The reaction mechanism varies depending on the catalyst used. Reaction with a platinum catalyst involves heterogeneous gas – solid catalysis

[22,23,26]. In contrast, according to present understanding, oxidation over a vanadium catalyst is a homogeneous reaction that takes place in a liquid melt of active components on both the external and internal surfaces of an inert solid catalyst carrier [41,42]. The reaction mechanism and the chemical structures of the active components have not yet been clearly defined. According to the model of MARS and MAESSEN [43], reaction in the melt takes place by way of the intermediate steps shown in Equations (12) and (13).

$$SO_2 + 2\,V^{5+} + O^{2-} \rightleftharpoons 2\,V^{4+} + SO_3 \quad (12)$$

$$0.5\,O_2 + 2\,V^{4+} \longrightarrow 2\,V^{5+} + O^{2-} \quad (13)$$

The validity of kinetic equations derived from this and other reaction models is limited to certain temperature ranges [44]. However, the rate of the catalytic oxidation of sulfur dioxide depends not only on the chemical mechanism but also on mass and heat transfer at the gas – liquid interface of the catalyst [42,45].

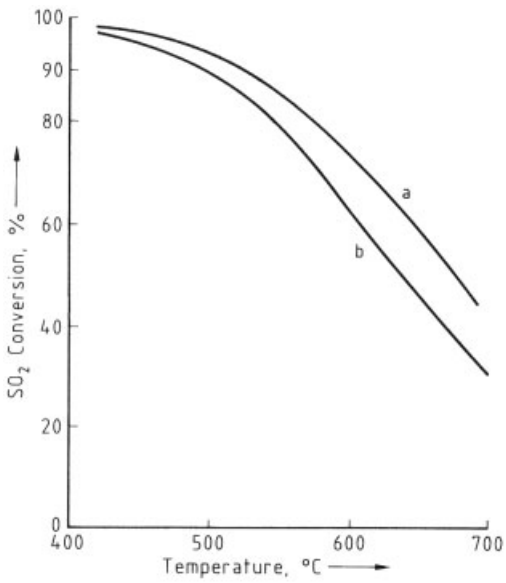

Figure 11. Comparison of (a) theoretical equilibrium SO_2 conversion for sulfur-burner gases (10 % SO_2, 10.9 % O_2) with (b) actual SO_2 conversion attained over a specific catalyst

Other purely technical parameters, such as the gas velocity, gas distribution, and residence time in the catalyst bed, help to determine how closely sulfur dioxide conversion in practice will approach the theoretical equilibrium. To estimate the sulfur dioxide conversion that can be achieved in reality in comparison with the thermodynamic equilibrium conversion, a correction function is used that takes into account the influences of the individual variables. Figure 11 is a plot against temperature of the theoretical sulfur dioxide equilibrium conversion and a typical observed sulfur dioxide conversion function. Actual conversion characteristics are significantly influenced by the specific catalyst activity, which must be determined empirically for each individual catalyst.

4.1.2. Catalysts [177]

Apart from catalyst activity, other factors including thermal stability, service life, and mechanical strength are of practical importance [20–23,42,44–47]. Of all substances tested for catalytic activity toward sulfur dioxide oxidation, only vanadium compounds, platinum, and iron oxide have proven to be technically satisfactory. Today, vanadium pentoxide is used almost exclusively.

Commercial catalysts contain 4 – 9 wt % vanadium pentoxide, V_2O_5, as the active component, together with alkali-metal sulfate promoters. Under operating conditions these form the liquid melt in which the oxidation of sulfur dioxide is thought actually to take place. Potassium sulfate is used most often in a K/V molar proportion of ca. 2.5 – 3.5. Some catalysts also contain sodium sulfate to reduce the melting point. The carrier material is silica in the form of diatomaceous earth, silica gel, or zeolites, all of which present especially large specific surface areas. Cesium-doped catalysts have also been developed and installed in various facilities. Cesium sulfate as a promoter reduces the melting point of the active components, resulting in significantly lower temperature limits for sustainable stable activity.

The catalyst components are mixed together to form a paste, which is then usually extruded into solid cylindrical pellets or rings. These are dried and baked at elevated temperature. Other catalyst forms less common in industry include spheres and tablets. Pellet-type catalysts were used almost exclusively until the mid 1980s, but plants today are usually equipped with ring-

shaped (or "star-ring") catalysts. The advantage of a ring-type catalyst is a lower pressure drop; compared to pellet-type catalysts the pressure drop is reduced by more than half. Furthermore, a ring-shaped catalyst is less sensitive to dust blockages. The relationship linking gas velocity with pressure-drop for different catalyst forms is illustrated in Figure 12.

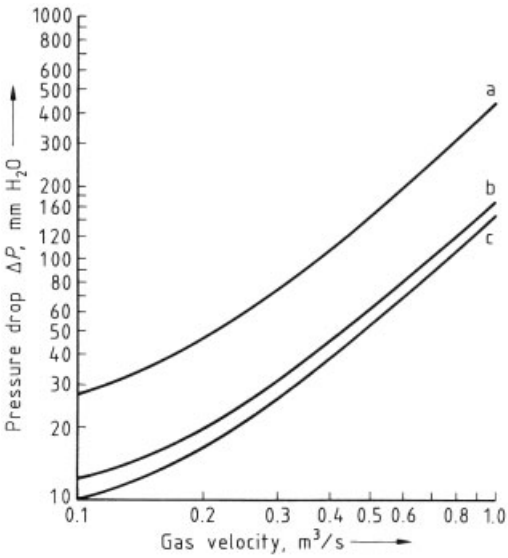

Figure 12. Pressure drops for various catalyst types per meter of catalyst depth (BASF)
a) Pellets (6 mm); b) Rings (10/5 mm); c) Star-rings (11/4 mm)

Operating Temperature Range. An important property of the vanadium catalyst is the low-temperature limit at which stable operation is possible under fixed gas conditions. This temperature is ca. 410 – 430 °C for a conventional catalyst and ca. 380 – 390 °C for a cesium-doped catalyst. Low-temperature activity depends mainly on the melting point and the chemical properties of the mixture of active constituents.

The upper operating-temperature limit is determined by the thermal stability of the catalyst. Above ca. 600 – 650 °C catalyst activity may be lost irreversibly because of damage to the structure of the carrier and reduction of its internal surface.

Service Life. The average service life quoted by most catalyst producers [46] is about ten years. Service life is generally determined not so much by progressive loss of activity as by catalyst losses incurred when filling and emptying the reactor and during routine screening. Depending on the dust load of the gas entering the converter, the size and shape of the catalyst grains, and the properties of the active melt, dust will accumulate in the catalyst bed over the course of time. This dust eventually increases the gas-pressure drop through the catalyst bed and reduces both gas throughput and SO_2 conversion efficiency. That is the reason why the catalyst must be screened from time to time to remove dust [48]. When catalyst is withdrawn, screened, and returned to the reactor, a certain amount is bound to be lost as a result of abrasion. The precise amount depends on the handling method used and the stability of the catalyst. This loss must be compensated by the addition of new catalyst.

In contrast to platinum, vanadium catalyst is largely insensitive to catalyst poisons [49]. Fluorine compounds in elevated concentrations will attack the carrier material, leading to increased abrasion loss. Chlorine compounds, especially at elevated temperatures, will volatilize the vanadium and consequently decrease the activity. Arsenic, which may be present in the feed gases that result from roasting arsenical pyrites, will accumulate in the catalyst, but it will only cause an observable decrease in activity if the As_2O_3 concentration in the catalyst mass exceeds 15 wt % [20].

Water vapor in the feed gas is not deleterious to a vanadium catalyst so long as the temperature is sufficiently high to prevent condensation of sulfuric acid. At low temperature (during plant stoppages, for example, or when the catalyst is exposed to humid air) there is a danger that water will be absorbed by the hygroscopic active constituents, and this can impair the mechanical strength of the catalyst.

4.1.3. Process Summary

There are four main process steps in the production of sulfuric acid from sulfur dioxide-containing gases by the contact process:

1) Gas drying

2) Catalytic conversion of sulfur dioxide to sulfur trioxide
3) Absorption of sulfur trioxide
4) Acid cooling

The gas-drying stage is not applicable to a plant of the wet-catalysis type. Almost without exception, contact plants operate under essentially atmospheric pressure; compression is required only for driving the gases through the plant.

4.1.3.1. Gas Drying

Gas drying is an important process step in conventional contact plants (in contrast to wet-catalysis plants). It protects cooler parts of the plant, such as heat exchangers, against corrosion by acid condensation, and it safeguards against the formation of sulfuric acid mist, which can be very difficult to absorb. It also protects the catalyst from ill effects of acid condensation when the plant is shut down for any reason. Therefore, both the operating performance (especially tail-gas purity) and the service life of the plant depend in large measure on an efficient and reliable gas-drying stage.

In sulfur-burning plants the molten sulfur used is dry from the outset, because its melting point is well above the boiling point of water, and any moisture originally present will have been driven off in the melter. The combustion air, however, must be dried. In the usual arrangement, filtered air from the atmosphere is drawn through a drying tower by the main blower.

When the feed gas is derived from smelter waste gases or pyrite roasting, cold, humid, sulfur dioxide-containing gas from the gas-cleaning system (→ Sulfur Dioxide, Chap. 6.) is mixed with such additional air as may be required to bring the O_2/SO_2 ratio to the optimum process value before the gas mix enters the drying section. Similar gas-drying equipment is used in both situations.

The gases are in most cased dried in countercurrent with fairly concentrated sulfuric acid in irrigated packed towers [20–22]. The sulfuric acid is circulated. The residual water content of the gases after drying corresponds theoretically to the partial pressure of water vapor above the drying-tower acid at the prevailing temperature and concentration. For purposes of achieving a high drying efficiency, the temperature of the irrigation acid is normally maintained at 50–60 °C.

A substantial amount of heat – not simply the heat of dilution of the sulfuric acid but also the heat of condensation of the water – is liberated in the gas-drying stage. For this reason the circulated acid is generally cooled by indirect heat exchange before returning to the dryer.

Water Balance. The sulfuric acid concentration of dryer acid is usually between 93 and 98 %, depending on the production conditions and the plant concept. This level is maintained constant by bleeding off part of the dilute acid leaving the dryer and exchanging it for a corresponding amount of concentrated acid (98.5 %) from the absorber circuit. When the desired product-acid strength is 93 – 95 % H_2SO_4, product acid can be taken directly from the drying circuit. In the case of a sulfur-burning plant, process water may even need to be added to the drying acid circuit to prevent the acid concentration from rising, and it is then not necessary to transfer dryer acid back to the absorber circuit. Occasionally, 98 % acid is used in the drying tower; this arrangement permits a common pump tank to be used for both dryer and absorption circuits instead of the usual separate pump tanks [50].

Water absorbed by the dryer acid is thus used as process water for sulfuric acid formation because of connections between the dryer and absorption systems. When processing metallurgical off-gases, the water content of the feed gas entering the dryer must be controlled by cooling in the gas-cleaning plant so that it does not exceed the stoichiometric requirement for production of H_2SO_4 based on the amount of SO_3 to be absorbed. Otherwise, the water balance in the contact plant will not be maintained, and the concentration of circulating acid will drop below the minimum level required for proper plant operation.

The permissible water content for feed gases is determined by the concentration of the product acid, the sulfur dioxide content of the contact gases, and the sulfur dioxide conversion efficiency. Figure 13 illustrates how the allowable feed-gas moisture content (represented by the gas temperature at the inlet of the drying tower, to which is is related) varies with sulfur dioxide

content for two different product-acid strengths. It can be seen that the lower the sulfur dioxide content of the gas, the cooler (i.e., the less moist) the gas must be before it enters the drying tower.

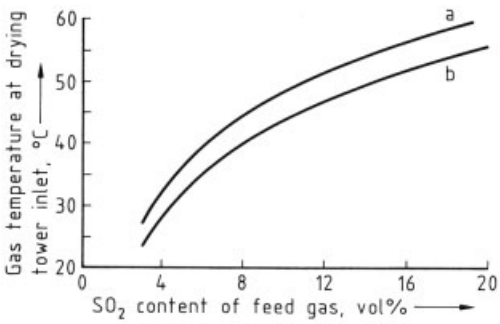

Figure 13. Maximum permissible temperature of moist gas at drying tower inlet of a roaster gas-based double-absorption sulfuric acid plant as a function of gas SO_2 content and desired product acid strength
a) 93 % H_2SO_4 product; b) 98.5 % H_2SO_4 product

In the case of metallurgical gases with extremely low SO_2 concentrations it is sometimes technically difficult and uneconomical to cool the feed gas to the low temperature required to maintain the correct water balance. For such cases Lurgi has developed a predryer-reconcentrator system that has been proven in service [51]. In the predryer, upstream from the main drying tower, surplus water is removed from the metallurgical gas by washing with sulfuric acid of medium concentration (30 – 60 %). This acid does not circulate to the main acid-plant dryer and absorber circuit. Instead, it circulates in a closed circuit between the predryer and a reconcentrator immediately downstream from the final SO_3 absorber in the main acid plant, where itgives up moisture to tail gas released to the atmosphere (see the description of the venturi process in Section 4.3.2). Alternatively, a water refrigeration plant would be required to adjust the gas temperature to the appropriate level.

Drying-Tower Design. The dryers used today are, as a general rule, vertical cyclindrical towers. Their steel shells are lined with acid-proof bricks. In the lower part of most such towers a plastic foil (polyisobutylene) is applied between the steel shell and the bricks in order to prevent acid from penetrating through the bricks to the steel shell. A grate of acid-proof material supports the packing. Normally, Raschig rings or Intalox saddles of ceramic material are used as packing. The dryer design is determined by a number of interacting parameters, such as the surface character and geometry of the packing, the packing height, the gas velocity, and the irrigation rate. The efficiency of moisture absorption by the sulfuric acid depends mainly on the diffusion resistance at the gas – acid interface. The gas flow velocity is best kept well above the region of laminar flow, because this induces turbulence that not only lowers the diffusion resistance but also improves distribution of the liquid.

A lower limit on the irrigation density is set by the minimum amount of acid required for uniform wetting of the entire packing surface. The maximum level is defined by the flooding limit, which is itself a function of the gas velocity, and must be determined empirically for each packing type.

An important factor in ensuring uniform acid distribution over the entire tower cross-section is the nature of the acid irrigation system. Preferably, this should be located in the upper part of the packing layer, and it should consist of cast iron or stainless steel nozzle tubes with lateral oblique openings directed upward. As a general rule, a wire-mesh filter of plastic or stainless steel is installed above the packing layer to separate entrained acid droplets.

Normally, a residual moisture content of 50 mg/m^3 in the dry gas is considered satisfactory.

4.1.3.2. Catalytic Oxidation of Sulfur Dioxide

The reactor in which sulfur dioxide is oxidized catalytically to sulfur trioxide is known as the *converter*. It is the heart of the sulfuric acid plant.

In the design and construction of any converter intended to assure maximum sulfur dioxide conversion, proper attention to removal of the very considerable reaction heat is of vital importance. The reaction is generally carried out under adiabatic conditions, so the temperature of the solid catalyst bed rises, thereby determining, and at the same time limiting, the attainable level of SO_2 conversion consistent with thermodynamic equilibrium (see Fig. 10). To achieve a

high final SO_2 conversion, the total catalyst mass is divided up into several catalyst beds (trays), and hot gas leaving each bed is cooled to the minimum working temperature of the catalyst before it enters the next bed. Tubular converters operating at nearly isothermal conditions have now almost completely fallen out of use in sulfuric acid plants. Isothermal fluidized bed converters are under development, however, and some have reached the point of industrial use.

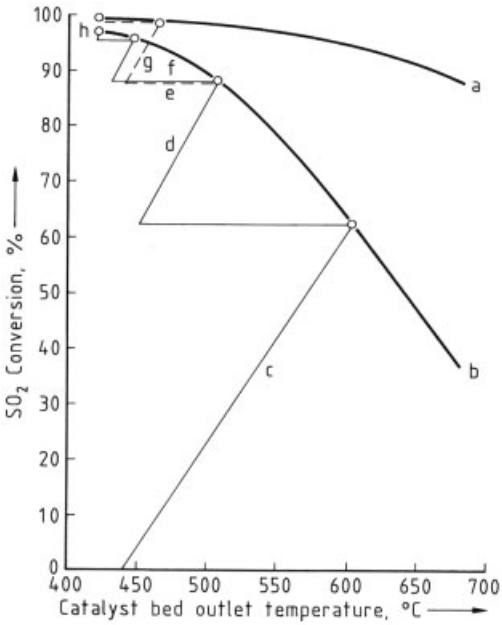

Figure 14. Comparison of reaction profiles and SO_2 conversion for 4-bed normal contact (single absorption) and (2 + 2) double-absorption processes (feed gas: 8.5 vol % SO_2)
a) Double-absorption process equilibrium curve after intermediate absorption; b) Equilibrium curve for normal contact process; c) Adiabatic reaction in bed 1; d) Adiabatic reaction in bed 2; e) Cooling and intermediate absorption; f) Cooling; g) Bed 3; h) Bed 4

Figure 14 shows the reaction profile for a converter comprising four beds and operating under adiabatic conditions, together with the SO_2 conversion attainable in each bed within the general limits of actual conversion characteristics. A profile for the normal contact process without intermediate absorption (single absorption) is compared with that for a double-absorption process with intermediate absorption after the second bed. It is evident that the overall conversion ultimately obtained in the double-absorption process is substantially greater than in a single-absorption process with the same number of catalyst beds. In the single-absorption process, the maximum achievable SO_2 conversion with a typical four-bed converter is ca. 98 %. (The exact figure depends on the feed-gas composition.) This contrasts with a final SO_2 conversion > 99.5 % in the double-absorption process.

The normal configuration for a double-absorption plant using a fixed-bed converter is either (2 + 2) or (3 + 1). That is to say, the intermediate SO_3 absorber can be placed after either the second or the third bed. If an SO_2 conversion efficiency > 99.7 % must be guaranteed, a five-bed converter in a (3 + 2) or (4 + 1) configuration may be preferred [52]. Further improvements in conversion efficiency, especially in a single-catalysis plant, can be achieved with cesium doping (see Section 4.1.2). The resulting considerable increase in low-temperature activity permits a conversion efficiency of ca. 99 % in a single-catalysis plant and > 99.7 % in a double-catalysis plans. However, the applicability of such catalysts is limited by high cost.

The double-absorption method has another significant advantage as well, in that it can process feed gases with a higher sulfur dioxide content and correspondingly lower O_2/SO_2 ratio than the single-absorption process. On account of a lower specific gas-flow rate for a given nominal H_2SO_4 production capacity, equipment for the double-absorption process can also be smaller.

Converter Design. The process design for a converter [22, 45] requires careful optimization because of the large number of interacting parameters. It has been considerably simplified by the development of appropriate computer programs [53–55]. The most important variables are the sulfur dioxide concentration and gas-flow rate, the number of beds, the specific catalyst quantity and its distribution between the individual beds, the gas-pressure drop, and the gas-inlet temperatures at the individual beds. In arriving at a final design, due consideration must be given to the relationship between equipment costs and energy costs.

The specific catalyst quantity required for production of 1 t/d of sulfuric acid is ca. 200 – 260 L for a "normal" contact (single-absorption) plant and about 150 – 200 L for a double-absorption plant. The distribution of the catalyst bet-

ween the individual beds can vary widely as a function of the gas concentration and the way the gas is routed. The catalyst bed height may vary from ca. 200 mm to 1000 mm. Different catalysts may also be used in different beds; the choice depends on the thermal stresses prevailing at each stage. The preferred grain size depends on pressure-drop considerations and the permissible dust load.

The optimum method of cooling the reaction gas between the catalyst beds is a function of the composition and initial temperature of the feed gas, and thus its origin. It is normally preferable to use indirect heat exchange. In a "cold-gas" plant running on off-gases from a metallurgical or spent-acid decomposition plant, virtually all the surplus reaction heat is transferred to the feed gas in a series of gas – gas heat exchangers to raise the gas temperature to the reaction temperature of the first catalyst bed. Modern plants processing gases with higher SO_2 loads (> 8 vol %) are equipped with boiler elements as a way of removing excess reaction heat. This improves the overall energy efficiency and maintains the gas temperature at the inlet of the absorption tower within reasonable limits. In a sulfur-burning plant, where the feed gas is already hot, surplus reaction heat from the catalytic section is recovered in steam generators and boiler feed-water preheaters (economizers). In all double-absorption plants, however, gas passing from the converter to the intermediate absorber and back is cooled and reheated in indirect gas – gas heat exchangers. Direct injection of cold feed gas or quench air may be employed to a limited extent under certain circumstances, either after the first catalyst bed for the purpose of limiting the outlet gas temperature and thus the thermal stress on the first heat exchanger or steam generator, or in later stages if the amount of available heat is insufficient to warrant provision of an indirect heat exchanger or feed-water preheater.

The idea of mounting at least some of the heat exchangers inside or around the converter shell has become attractive once again even though it was abandoned in the 1970s and 1980s. Converters fabricated entirely of stainless steel provide sufficient flexibility from a design and construction standpoint to permit such heat exchangers to be incorporated into the converter vessel. The advantage of this arrangement is of course elimination of some of the gas ducts, including expansion joints, supports, etc. However, no general rule is available to cover all types of plants; the design engineer must consider carefully the relative advantages of fewer gas ducts versus a more complicated converter vessel.

Normally, the converter is designed as a vertical cylindrical vessel, with the catalyst beds mounted above one another in separate, virtually gas-tight compartments. Gases enter and leave through lateral nozzles, passing upward and downward through the beds.

The catalyst mass is supported by a metallic bed grate. A base layer of ceramic packing prevents direct contact between the catalyst and the grate, which could give rise to corrosion. The catalyst bed is itself covered with another layer of packing. This serves the dual purpose of helping to ensure uniform gas and temperature distribution over the surface of the catalyst and preventing the catalyst from blowing around and leaving cavities in the bed, which might also adversely affect gas and temperature distribution in the main body of the catalyst bed. The converters are usually lined with ceramic material in the catalyst areas.

Three principal types of converter are in use: stainless steel (with or without heat exchangers), steel, and brick-lined [56].

The *stainless-steel converter* (Fig. 15, see next page), consisting of the shell, the tray separators, and the trays, is fabricated entirely of stainless steel or heat-resistant steel. A central tube is often used to support the trays. Heat exchangers (also made of stainless steel) can either be incorporated into the central tube or arranged around the shell.

The *steel converter* (Fig. 16, see next page) corresponds in principle to the stainless-steel converter, but boiler plate or equivalent is used as the construction material for the shell, the separations, and trays for those layers with lower operating temperatures. A steel converter is more economical compared with a stainless-steel converter and is appropriate for treatment of gases with lower SO_2 loads.

The *brick-lined converter* (Fig. 17, see following page) is completely lined internally with acid-proof bricks. The compartment separators are self-supporting domed structures made of

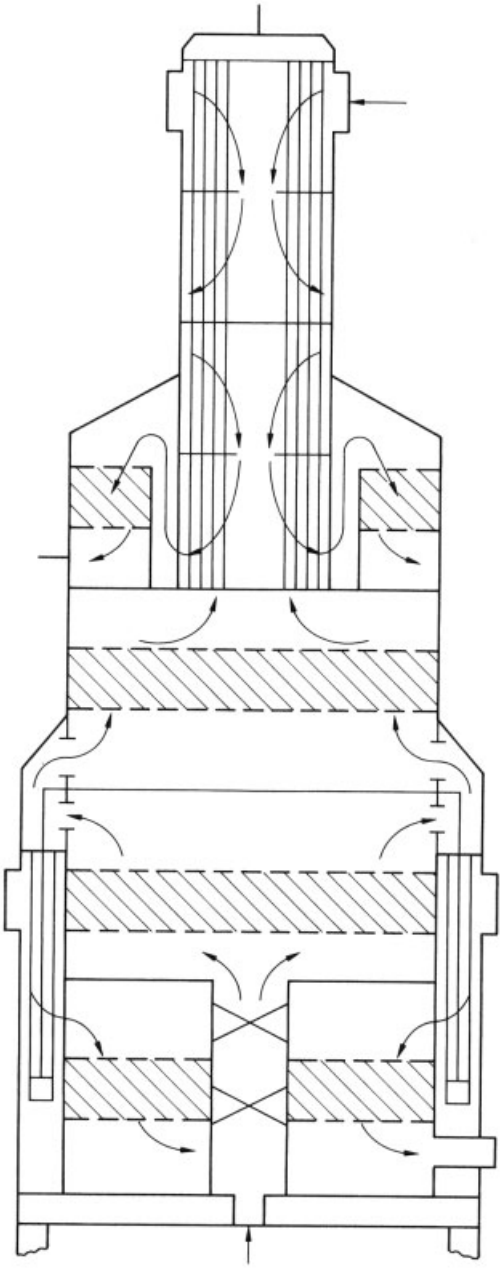

Figure 15. Lurgi converter in stainless steel with integrated heat exchangers

shaped bricks. These carry brick columns that support the catalyst-bed grates. The brick-lined converter represents a conservative design guaranteeing long lifetime together with a high "thermal inertia." This facilitates operation with fluctuating gas loads as well as start-up after idle periods.

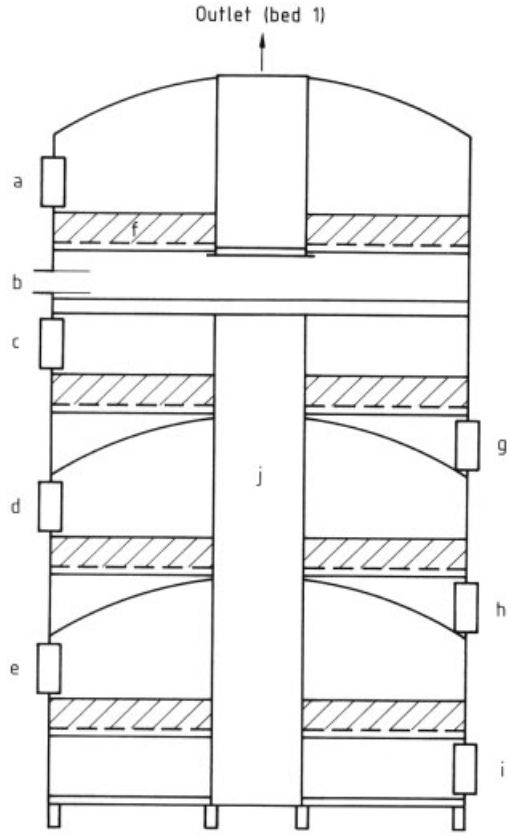

Figure 16. Four-bed all-steel converter
a) Inlet (bed 1); b) Quench inlet; c) Inlet (bed 2); d) Inlet (bed 3); e) Inlet (bed 4); f) Catalyst; g) Outlet (bed 2); h) Outlet (bed 3); i) Outlet (bed 4); j) Central bed support tube

The *cast-iron converter* that for many years constituted the standard design – specifically in the United States – has now been practically eliminated from sulfuric acid plants.

Tubular converters (→ Tubular Reactors), which were in common use in sulfuric acid plants until the 1960s, have the inherent advantage over bed converters that reaction conditions can be made almost isothermal, but they are quite unsuitable for use in today's large-capacity acid plants on account of several seri-ous disadvantages. In particular, in order to ensure that heat is dissipated efficiently from the catalyst it is necessary to use a large number of narrow-diameter tubes, even for small plant capacities. It is difficult to change the catalyst charge in such tubes, gas distribution is difficult to control, and the risk of corrosion is increased because catalyst is in direct contact with the tube material.

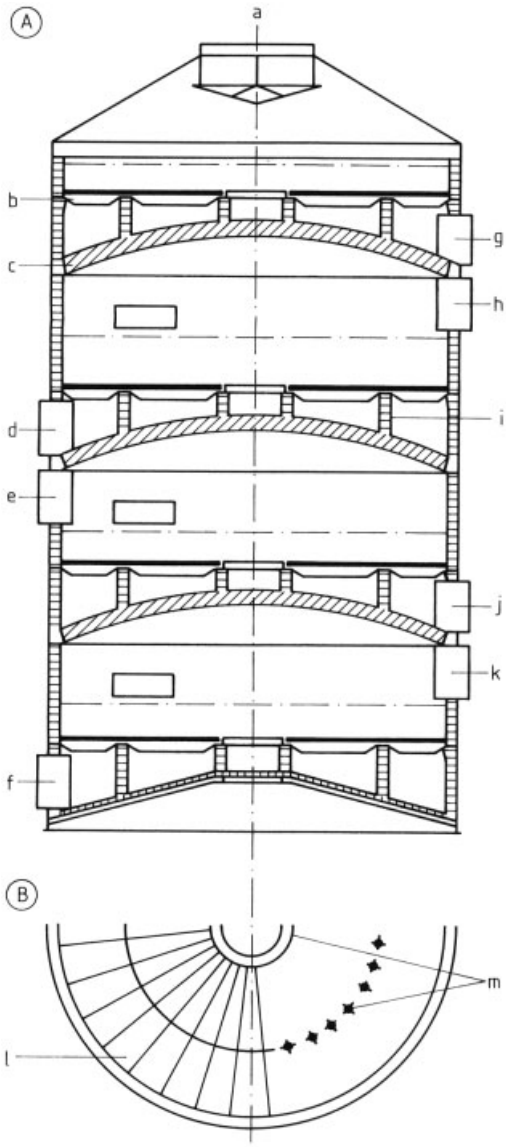

Figure 17. Four-bed brick-lined converter
A) Side elevation; B) Plan
a) Inlet (bed 1); b) Catalyst bed grate; c) Compartment separator; d) Outlet (bed 2); e) Inlet (bed 3); f) Outlet (bed 4); g) Outlet (bed 1); h) Inlet (bed 2); i) Brick supporting column; j) Outlet (bed 3); k) Inlet (bed 4); l) Catalyst grate; m) Brick columns

These shortcomings are not shared by the *fluidized-bed converter*, which also operates under virtually isothermal conditions. The good heat-transfer characteristics of a fluidized bed are well known (→ Fluidized-Bed Reactors). It is possible to conduct sulfur dioxide oxidation in this way at a constant temperature in the optimum operating range of the catalyst. The released heat of reaction is removed from the fluidized bed by immersed tube bundles that generate steam. One great advantage of the fluidized-bed converter is that cold feed gases from metallurgical processes or other sources need not be preheated to the reaction temperature; instead the conditions in the fluidized bed are simply adjusted to provide any heat necessary to the incoming gas. A further attraction of this type of converter is that it can process feed gases with much higher sulfur dioxide concentrations relative to a fixed-bed converter [57].

Because of the favorable thermodynamic conditions in a fluidized-bed converter, the conversion attainable in a single stage may equal that achieved in the first three beds of a conventional fixed-bed converter. The fluidized-bed converter is therefore ideally suited for use ahead of the intermediate SO_3-absorption stage in a double-absorption plant. However, itdoes require an extremely abrasion-resistant catalyst [58]. Bayer AG has developed a spherical catalyst for this purpose [59], and has used it in two fluidized-bed converter plants since 1971 and 1976, respectively, apparently with success [60].

4.1.3.3. Absorption of Sulfur Trioxide

Sulfur trioxide formed by the catalytic oxidation of sulfur dioxide is absorbed in sulfuric acid of at least 98 % concentration, in which it reacts with existing or added water to form more sulfuric acid [20–23]. The optimum concentration of the absorber acid corresponds to the azeotrope (see Section 2.1), where the partial pressures of SO_3, H_2SO_4, and water vapor are all at a minimum. At lower acid concentrations the water-vapor partial pressure is higher, and there is a correspondingly greater risk that sulfuric acid mist will form as a result of direct reaction of sulfur trioxide in the gas phase with water vapor above the acid. At acid concentrations above the azeotropic point the tail gas will contain increased amounts of sulfur trioxide and sulfuric acid on account of their higher partial pressures.

In the original contact (single-absorption) process, process gas passes through all the converter beds before the sulfur trioxide is absorbed

in a single absorption unit. In the double-absorption modification, now routinely practiced in both new and "revamped" plants, most of the sulfur trioxide is removed from the process gas in an extra absorption step at an intermediate stage. As explained above, this may occur after either the second or the third converter bed. In the (2 + 2) configuration, ca. 85 % of the original sulfur dioxide content of the feed gas has been converted to sulfur trioxide by the time the gas enters the intermediate absorber, and in the (3 + 1) configuration even more of the total amount of sulfur trioxide (ca. 93 – 95 %) is removed by the intermediate absorber. Sulfur trioxide formed in the last bed or beds from the small quantity of residual sulfur dioxide is absorbed in the final absorber.

Most absorbers are *packed towers*, usually operating in countercurrent. The gases pass from bottom to top through a bed of packing, which is uniformly irrigated from the top with concentrated sulfuric acid. Process gas leaving the converter system is cooled by a gas – gas heat exchanger or a steam generator, preferably in conjunction with a feed-water preheater, to a temperature of ca. 180 – 220 °C before entering the absorber. It is essential that the wall temperature in the gas coolers never drops below the acid dewpoint (ca. 110 – 160 °C, depending on the gas composition); otherwise, there is an acute danger of corrosion due to condensing acid, as well as mist formation. Gas entering the absorber is therefore not completely cold, and it releases heat to the absorber acid as it passes through the absorber; by the time it reaches the outlet it is at virtually the same temperature as the incoming absorber acid.

A substantial amount of heat is also generated in the absorber acid from absorption of sulfur trioxide and formation of sulfuric acid, and the acid temperature rises in consequence by an extent that depends on the acid-circulation rate. Efficient sulfur trioxide absorption depends not only on uniform acid and gas distribution in the absorber but also on ensuring that the temperature and concentration of the absorber acid remain at the optimum values. The acid concentration is held constant by adding process water or dryer acid to acid leaving the absorber, at a rate controlled by a device that measures the electrical conductivity or density. The optimum acid-inlet temperature depends on design conditions, but it is ca. 60 – 80 °C in most plants, maintained at that level by indirect cooling. The attainable SO_3 absorption efficiency is generally > 99.9 %.

Irrigated packed-tower absorbers are not efficient at removing *sulfuric acid mist* [1]. In spite of efficient gas drying and optimum conditions for sulfur trioxide absorption, it is often impossible to prevent mist formation completely, especially when processing high-bitumen sulfur or metallurgical feed gases with elevated hydrocarbon contents [61], or when starting up or shutting down the plant. Furthermore, in plants with an oleum tower upstream from the intermediate absorber, sulfuric acid mists may form with aerosol particle sizes in the submicron range owing to overcooling of the process gas. In such cases the mists may already have formed in the intermediate absorber. Ordinary wire-mesh spray separators have no effect on these mists, so special *mist eliminators* must be installed to avoid corrosion in the downstream heat exchangers and further mist formation in the final absorber. There are various designs for mist eliminators, and not all operate on the same principle. The most appropriate type depends on the nature of the mist, especially its particle size. For example, impingement separators are best for trapping particles above 1 – 3 μm, while submicron particles are more efficiently trapped by diffusion on Brownian-motion separators [61–65]. Glass fibers of varying degrees of fineness are preferentially used as the filter material.

The design principle underlying conventional absorption towers is similar to that for drying towers. Such towers have welded cylindrical steel shells lined on the inside with acid-proof bricks and silica-based, acid-proof mortar. Often the lower part of the tower around the acid sump is further protected by polytetrafluoroethylene (PTFE) sheeting sandwiched between the steel shell and the multilayer brick lining. As with the dryer, it is essential to ensure that gas and acid distributions are uniform over the entire tower cross-section, and that the acid flow rate is sufficiently high to wet the entire packing layer completely. Depending on gas conditions, the packing layer may have a height of 4 – 6 m.

The development of special stainless steels over the past few years has permitted the installation of steel absorption towers without any brick lining. This is a particularly great advantage when towers must be replaced during main-

tenance shut-downs. The absence of a brick lining means that a new tower can be completely prefabricated and set onto the existing foundations within a very short period of time.

Sulfur trioxide absorption systems based on *venturi scrubbers* were introduced by Lurgi in the early 1970s. These were developed specifically for the purpose of reducing heat losses during intermediate absorption so as to permit low-grade smelter off-gases to be processed autothermally in a double-absorption plant [40]. Because the gas and absorber acid flow in cocurrent in a venturi system, the gas temperature at the absorber outlet is higher than with a conventional countercurrent absorber based on acid at the same inlet temperature. Thus, less heat is transferred from the gas to the absorber acid for dissipation through the acid-cooling system. Running the absorber at a higher acid temperature reduces gas-heat losses still further. This principle of hot absorption in cocurrent with acid at a temperature of about 120 – 140 °C is useful not only for the purpose of processing low-grade feed gases but also for maximizing heat recovery generally in a double-absorption plant.

The venturi absorption system consists of vertical and horizontal venturi scrubbers arranged in series. In each case, acid is injected through a nozzle into the gas inlet of the venturi unit. Upon leaving the venturi units the gas passes through a packed tower where droplets entrained by the gases are separated.

The intensive mixing of injected acid with the turbulent gas stream in a venturi unit provides a large liquid – gas interface and, consequently, favorable mass- and heat-transfer conditions. This type of venturi system has given satisfactory service in many sulfuric acid plants, including units with H_2SO_4 capacities as great as 2000 t/d.

4.1.3.4. Acid Cooling

The acid cooling system plays a vital role in determining the efficiency and operating safety of an entire sulfuric acid plant. Appropriate choice of a particular cooling system depends not only on the acid temperature but even more on the availability and quality of the cooling water, as well as the cost of water in relation to the cost of energy.

The introduction of special stainless steels led to significant changes in the field of sulfuric acid cooling beginning in about 1980. Acid cooling in modern plants is dominated by two types of coolers:

1) Shell-and-tube coolers, either equipped with anodic protection or fabricated from special stainless steel that requires no such protection
2) Plate-type coolers

Intermediate closed-loop water circuits are also being installed more frequently both for heat recovery purposes and as a way of complying with environmental regulations (especially with respect to sea water). A further step has been the development of systems generating steam from heat released in the absorption system. A brief description of such systems is provided in Section 4.1.4.

Cast-iron cascade coolers, which were standard equipment for many decades, have been almost completely eliminated from modern acid plants; even in existing installations these coolers are in most cases being replaced. Decreasing reliability, poor quality, environmental regulations, and of course better alternatives have forced this development. The use of spiral coolers, tank coil coolers, etc., fabricated from "standard stainless" is limited to special applications. Air coolers occupy a share of the market in situations where the consumption of cooling water is restricted.

Shell-and-Tube Coolers. Shell-and-tube coolers offer many advantages, including ease of installation, compact design with correspondingly low specific space requirements, and good heat-transfer coefficients (ca. 800 – 1400 $Wm^{-2}K^{-1}$, 700 – 1200 $kcal\,m^{-2}h^{-1}K^{-1}$, depending on the design conditions and the mode of construction). Shell-and-tube (s + t) coolers made of stainless steel have been used in sulfuric acid plants since the mid-1960s when the introduction of anodic protection made this practical. At present, two types of s + t coolers are preferred:

1) Coolers fabricated from "standard" stainless steels and equipped with anodic protection systems

2) Coolers fabricated from special stainless steels (Sandvik SX, 1.4575, etc.), which do not require anodic protection

These coolers make it possible to cover the complete range of operating parameters normally encountered in drying and absorption systems. Sea water, brackish water, cooling-tower water, and closed-loop water can all be used as the cooling medium, provided appropriate materials are selected. In special cases, shell-and-tube coolers are made from other materials – Teflon or glass, for example – especially when aggressive media are being heated by the hot absorber acid, such as dilute or chemically contaminated spent sulfuric acid.

Plate Coolers. Ever since the development of adequate acid-proof gaskets, plate coolers have been increasingly called upon for sulfuric acid cooling duty. The sealing materials used today are elastomers such as Viton (a copolymer of vinyl chloride and hexafluoropropene; → Fluorine Polymers, Organic, Chap. 3.2.), which can handle acid at temperatures up to a maximum of ca. 110 °C. The special advantages of plate heat exchangers are extremely compact design, good accessibility, and easy maintenance and cleaning [66, 67]. They can be built with very thin walls, so the specific material requirement for the heat exchange area is very low, and the heat-transfer coefficient is very high (up to ca. 2300 $W\,m^{-2}\,K^{-1}$). Therefore, even expensive special materials can be used economically. Hastelloy C276 (DIN 2.4819) (→ Construction Materials in Chemical Industry, Chap. 5.3.3.), for example, has proved to be very satisfactory in plate heat exchangers designed for sulfuric acid cooling with brackish or sea water, although the water must be filtered carefully to avoid deposits in the plate cooler. For cases in which higher temperatures are required (e.g., for heat recovery purposes), an alternative type of plate cooler has been developed based on semiwelded plates. Pairs of welded plates form the sulfuric acid channels in these coolers, while plates with normal gaskets serve as the water channels.

Air Coolers. Air coolers are preferentially used in sulfuric acid plants when cooling water is not available in adequate quantity or at reasonable cost. The acid flows through tubes arranged horizontally in flat bundles, which are equipped with external fins to improve heat transfer. The cooling air is forced or sucked past the tubes by means of a fan. The tubes are usually made of stainless steel. The acid temperature is limited to ca. 80 °C unless the cooler is equipped with anodic protection, in which case higher temperatures can be tolerated [68]. Alternatively, if the cooler is equipped with acid-proof centrifugally cast tubes, acid temperatures up to about 110 °C can be handled.

Air coolers are relatively expensive to install, and a considerable amount of power is also required to drive the fans. However, in any cost comparison with water-cooled types the capital and operating costs of not only the acid-heat exchanger itself but also the requisite water-cooling system and water-treatment unit must be taken into account.

4.1.4. Practical Versions of the Contact Process

Contact sulfuric acid plants vary in a number of respects depending upon the raw material used to produce the sulfur dioxide-containing contact gases. The broadest division is into *hot-gas* and *cold-gas* plants. In hot-gas plants, which are usually based on the combustion of elemental sulfur, the hot feed gas from the sulfur furnace is cooled just to the required converter-inlet temperature. Cold-gas plants, based on metallurgical or decomposition gases, are so designated because the crude sulfur dioxide-containing gas must be cooled to low temperatures in the dedusting and cleaning systems before being introduced into the sulfuric acid plant. It is therefore necessary to reheat the cold feed gas to the requisite converter-inlet temperature with reaction heat from the converter system.

Cold-gas plants are sometimes confusingly described as "wet-gas" plants because cold feed gas leaving the gas-cleaning system is saturated with water vapor. However, this water vapor is removed by a dryer before the feed gas is preheated, so in all the remaining process steps the gas is effectively dry. By contrast, in a true wet-contact or wet-catalysis process, moist feed gas (produced, for example, by the combustion of hydrogen sulfide) is processed directly in the

converter without preliminary drying. Therefore, the term "wet" is best avoided when referring to a cold-gas plant.

These traditional contact-plant classes can be further subdivided into the double-absorption process (also referred to as the double-contact or double-catalysis process), which is the type of process now most commonly used in new plants, and the older, ordinary contact process, without intermediate absorption (also referred to as the single-catalysis or single-contact process) [69–76]. The ordinary contact process is still used where the feed gases are very low in SO_2 and where permitted by local pollution regulations. Within a given overall process concept, many variations in detailed design are possible [75–79].

4.1.4.1. Double-Absorption Process Based on Sulfur Combustion

The sulfur-burning double-absorption process is considered to be the standard sulfuric acid production process for conforming with sulfur dioxide emission limits now in force in most countries [52]. Heat is released in all steps. Whereas it has long been standard practice to utilize reaction heat from the converter for producing high-pressure steam, more recently it has become of economic interest also to recover low-level heat from the absorber and dryer-acid systems.

Process Description. Figure 18 is a flow diagram of a typical sulfur-burning double-absorption plant with a four-bed converter in the (3 + 1) configuration (intermediate absorption after the third bed), producing ca. 98.5 % sulfuric acid. The feed gas is produced by combustion of liquid elemental sulfur with dried air (→ Sulfur Dioxide, Chap. 4.). Air is drawn from the atmosphere through a filter and an air dryer (irrigated with concentrated sulfuric acid, see Section 4.1.3.1) into the main blower, which compresses it to a pressure of ca. 1.4 bar.

The combustion air flow is controlled in proportion to a predetermined sulfur feed rate so that gases of about 11 – 12 vol % SO_2 are produced by combustion. Combustion gas leaves the sulfur combustion furnace at about 1100 °C and is cooled in a waste-heat boiler to the converter-inlet temperature of ca. 420 – 450 °C. Under the adiabatic conditions prevalent in the first bed of the converter (Section 4.1.3.2), heat generated

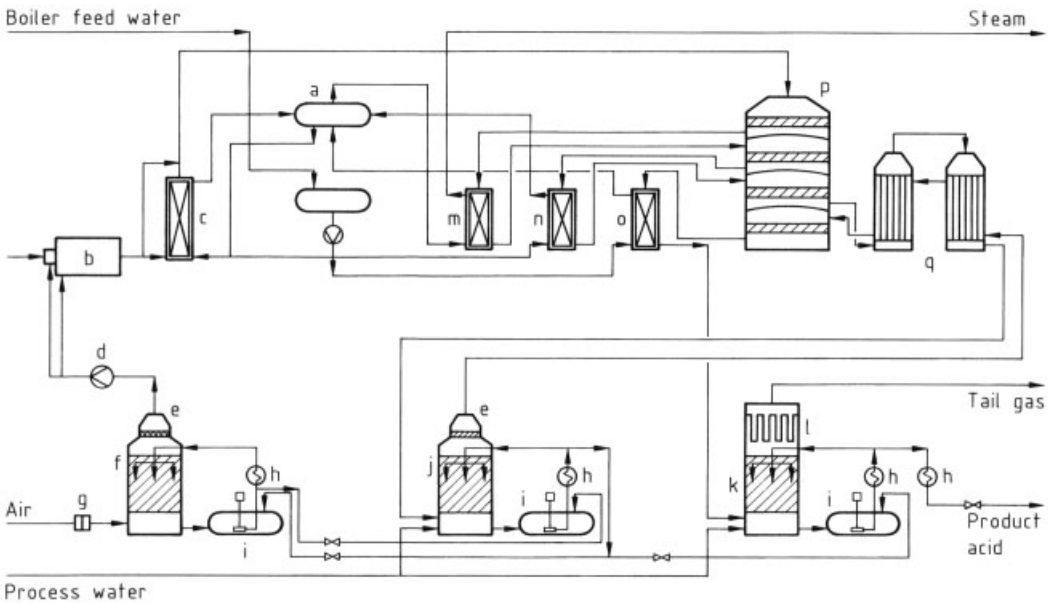

Figure 18. Sulfur-burning double-absorption sulfuric acid process (Lurgi)
a) Steam drum; b) Sulfur furnace; c) Waste heat boiler; d) Main blower; e) Mist eliminator; f) Drying tower; g) Air filter; h) Cooler; i) Acid pump tank; j) Intermediate absorber; k) Final absorber; l) Candle filters; m) Steam superheater; n) Boiler; o) Economizer; p) Converter; q) Intermediate heat exchanger

in the reaction raises the temperature of the gas mixture until it reaches a level at which the reaction is essentially in equilibrium. In practice, this corresponds to ca. 600 – 620 °C, representing a sulfur dioxide conversion of about 60 %. Passing out of the first bed of the converter, the gases are cooled in a steam superheater to the second-bed inlet temperature of 430 – 440 °C. The gas leaving the second bed is cooled in the evaporator of the steam system to the third-bed inlet temperature, ca. 420 – 440 °C. Gas leaving the third bed is cooled in two intermediate heat exchangers and passes at ca. 200 – 220 °C into the intermediate absorber, where sulfur trioxide is removed by absorption in concentrated sulfuric acid (Section 4.1.3.3). The remaining gas, which is cooled in the absorber to about 70 – 80 °C, is returned to the intermediate heat exchangers through a heated double-jacket tube to avoid the condensation of sulfuric acid. The cold gas is heated indirectly by hot gas from the third bed to the inlet temperature of the fourth bed (ca. 400 °C). The hot gases coming from the fourth bed are cooled to ca. 160 °C in an economizer (feed-water preheater) and passed into the final absorber. After final absorption, so-called tail gas is discharged to the atmosphere through a stack at ca. 70 – 80 °C. An overall sulfur dioxide conversion efficiency of ca. 99.7 % is attainable in this type of plant. Assuming the feed gas contains 10 vol % SO_2, this corresponds to a sulfur dioxide concentration in the tail gas of about 400 ppm SO_2.

Figure 18 also shows the acid circuits for the air dryer and the intermediate and final absorbers. In this example, the end product is 98 – 98.5 % sulfuric acid, which is discharged from the final absorber circuit. Water required for maintaining the correct acid concentration, apart from that absorbed in the air dryer, is supplied as process water to the absorber acid system. Water addition is controlled automatically by an acid-concentration measuring system operating within narrow limits (Section 4.1.3.3). The dryer-acid circuit is designed to discharge sulfuric acid of 93 – 97 % concentration back to the absorber circuit. If part of the production is to be discharged as oleum (see Section 4.4), and additional oleum tower is added upstream from the intermediate absorber. Plants designed for maximum oleum production preferentially reflect the (3 + 2) configuration. Five-bed converter systems with (3 + 1) configurations are used especially for achieving maximum final conversions [44, 72–74, 80].

Sulfur-burning double-absorption plants are designed for production capacities $\leq$ 3000 t/d of 100 % sulfuric acid in a single stream [81].

Energy Balance. All steps in the production of sulfuric acid from elemental sulfur are exothermic. An amount of heat corresponding to about 5.4 GJ is generated overall per ton of 100 % sulfuric acid. The thermal capacity of a sulfuric acid plant with an output of 1000 t/d H_2SO_4 is, accordingly, ca. 63 MW. The liberated heat must be dissipated under controlled conditions in such a way as to maintain optimum gas temperatures in the converter system and optimum acid temperatures in the dryer and absorber circuits, thereby ensuring that sulfur dioxide concentrations in the tail gas are minimized and sulfuric acid mist and sulfur trioxide at the stack outlet are avoided insofar as possible. The systems for gas and acid cooling are therefore essential components of a sulfuric acid plant.

Figure 19 shows the energy flow in the form of a Sankey diagram for a double-absorption plant of the type represented by Figure 18 [75]. Of the total energy input (100 %), 97 % is accounted for as energy released in the conversion of sulfur to sulfuric acid, and 3 % of the energy is consumed in driving the gas through the plant. Up to about 70 % of the total energy is normally utilized for the generation of ca. 1.35 t of high-pressure steam (40 bar, 400 °C) per tonne of sulfuric acid; the remaining 30 % is usually lost as waste heat.

The high-pressure steam is generated with high-temperature heat recovered by indirect exchange with gases from the converter system and the sulfur furnace. The waste-heat system is completely integrated into the double-absorption sulfuric acid plant (Fig. 18). In the economizer that cools the gases from the fourth bed, about 13 % of the overall process heat is utilized for feed-water preheating. The waste-heat boiler downstream from the sulfur furnace and the evaporator after the second bed use respectively 38 % and 16 % of the total heat to produce saturated steam. Another 10 % of the total energy is recovered in the steam superheater after the first bed for the generation of high-pressure steam.

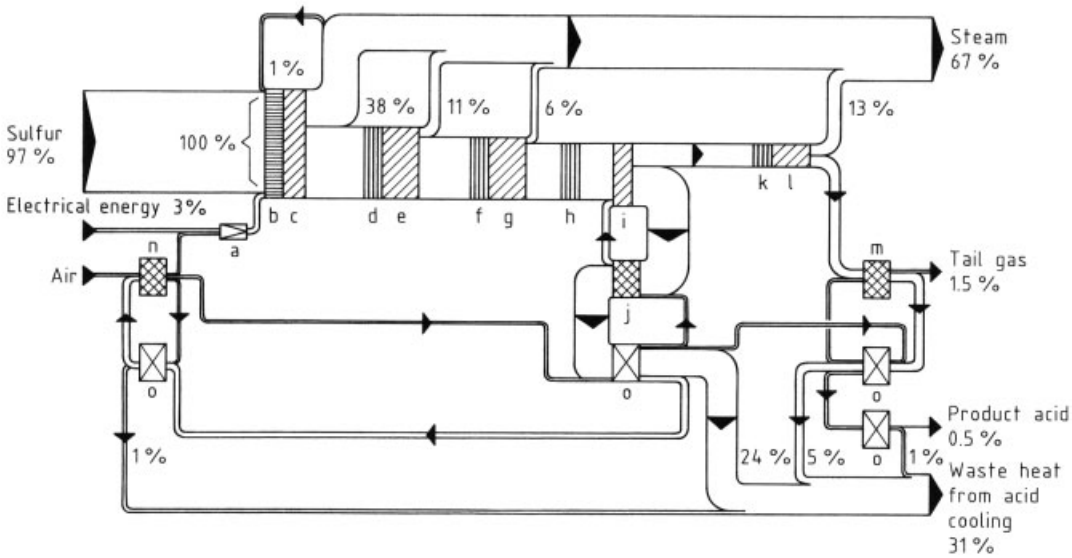

Figure 19. Sankey energy-flow diagram for a 1000-t/d sulfur-burning double-absorption sulfuric acid plant (feed gas: 11 % SO_2)
a) Blower; b) Sulfur furnace; c) Waste heat boiler; d) Catalyst bed 1; e) Steam superheater; f) Catalyst bed 2; g) Boiler; h) Catalyst bed 3; i) Intermediate heat exchangers; j) Intermediate absorber; k) Converter bed 4; l) Economizer; m) Final absorber; n) Air dryer; o) Acid coolers

Of the generally unrecovered waste heat, the majority (31 % of the total energy) is removed in the acid-cooling systems: 1 % in the dryer circuit, 24 % in the intermediate absorber circuit, 5 % in the final absorber circuit, and 1 % in the product-acid cooler. Heat in the dryer results from the condensation of water vapor and the dilution of sulfuric acid. Heat from the absorbers is made up of sensible heat transferred from the moderately hot gas to the colder acid in the absorber, the latent heat of condensation of sulfur trioxide, and the heats of formation and dilution of sulfuric acid. A small amount of the total energy is discharged from the system as sensible heat in the tail gas (2.5 %) and in the product acid (0.5 %).

Maximizing Energy Recovery. The value of energy has risen so much that systems for maximum energy recovery are becoming increasingly economical despite the associated considerable increase in capital cost. This is especially true for the extremely large-capacity double-absorption sulfuric acid plants now in use.

In the *steam system*, it is possible to improve energy utilization not only by increasing the production of high-pressure steam but also by optimizing the steam quality [76–78]. Use of a specially designed waste-heat system makes it possible because of the high excess temperature of the gas to produce steam at up to 80 bar in a sulfuric acid plant instead of the normal 40 – 50 bar. The higher the quality of the steam the more efficiently it can be converted into mechanical work or electrical energy. Production of high-grade steam is advantageous even if low-pressure steam for heating purposes is all that is required in neighboring production plants. In this case the high-pressure steam can be transformed into low-pressure steam in a back-pressure turbine driving a generator or other equipment [82, 83]. The Lurgi substoichiometric sulfur-combustion system (→ Sulfur Dioxide, Chap. 4.4. is well suited to the generation of highgrade steam, and it has been used successfully in a number of modern sulfuric acid plants [77].

Steam production can be increased by raising the acid and gas-outlet temperatures in the intermediate absorber to reduce gas-heat losses. As a general rule, hot converter gas is cooled before entering the intermediate absorber by indirect exchange with cold gases returning from the

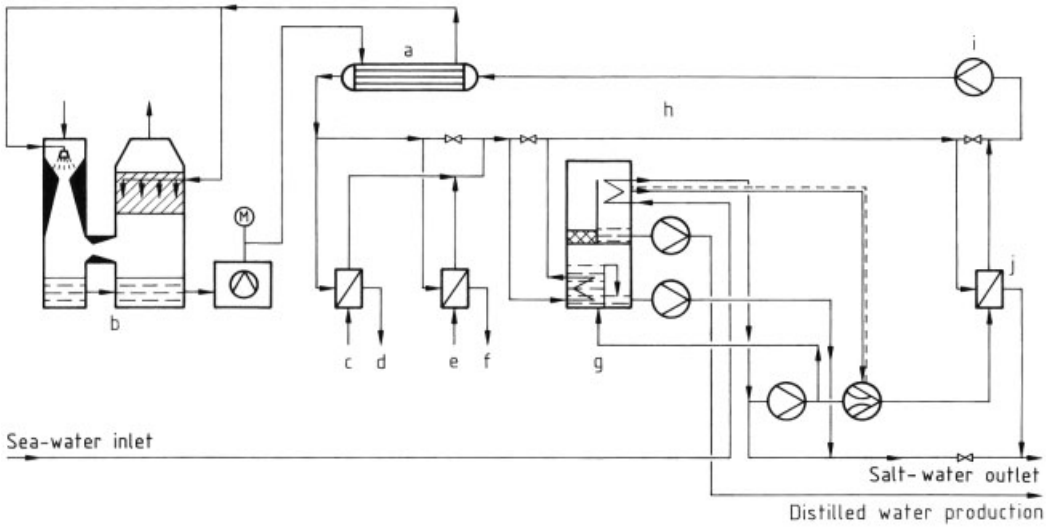

Figure 20. Absorber acid heat-recovery system with intermediate closed-loop hot water circuit
a) Absorber acid circuit; b) Venturi SO_3 absorption system; c) Boiler feed-water heater; d) Feed heater; e) Filter wash-water heater; f) Wash heater; g) Sea water desalinator; h) Closed-loop hot water circuit; i) Hot water circulation pump; j) Trim cooler

same absorber. The gas exit-temperature from the intermediate heat exchanger is usually too close to the acid dewpoint for the recovery of any more heat for steam generation. Otherwise there would be a danger of sulfuric acid condensing, which would give rise to corrosion on the heat-exchange surfaces as well as mist formation. However, use of venturi-type cocurrent absorbers and the hot absorption principle permits the gas temperature after the intermediate absorption stage to be increased substantially (see Section 4.1.3.3). Thus, the amount of heat required for heating the cold gases from the intermediate absorption stage is reduced. At the same time, an adequate amount of heat at sufficiently high temperature becomes available in the sulfur trioxide-rich gas upstream from the intermediate absorber to run an evaporator or economizer (feed-water preheater) integrated within the steam system. This can be located either before or after the intermediate heat exchangers. It is possible to boost steam production in the sulfuric acid plant by an additional 10 % by preheating the boiler feed water to about 90 °C, utilizing waste heat in acid from the absorption systems.

If recovery of acid waste heat is to be optimized, the temperatures in the absorption systems become very important. Raising the temperature creates more favorable conditions and a wide range of possibilities for heat utilization. The extent to which the acid temperature can be increased is restricted largely by the allowable corrosion rate in the heat exchangers. Considerable advances have been made in recent years in the development of special materials and techniques for controlling corrosion at higher acid temperatures. A number of double-absorption plants now incorporate acid-cooling systems modified for heat recovery and utilization. The required investment can be amortized within a short time out of energy cost savings.

Figure 20 shows one possible combination of various approaches to utilizing recovered acid waste heat [77, 84]. The heat recovery system depicted has been installed in a preexisting sulfur-burning double-absorption plant. The heat is used to produce boiler feed water by distillation of seawater, for boiler feed-water preheating, and for preheating filter wash water for a neighboring phosphoric acid plant. Heat is transferred from the anodically protected stainless-steel acid cooler to the various consumers by means of a closed-loop hot water circuit.

In other sulfuric acid plants, waste heat recovered from the acid is utilized in hot water circuits for remote heating [85]. In plants operating with absorber acid at elevated temperature (130 – 140 °C), waste heat from the acid is also used for concentrating phosphoric acid from 28 – 54 % P_2O_5 (Fig. 21). Other plants use acid

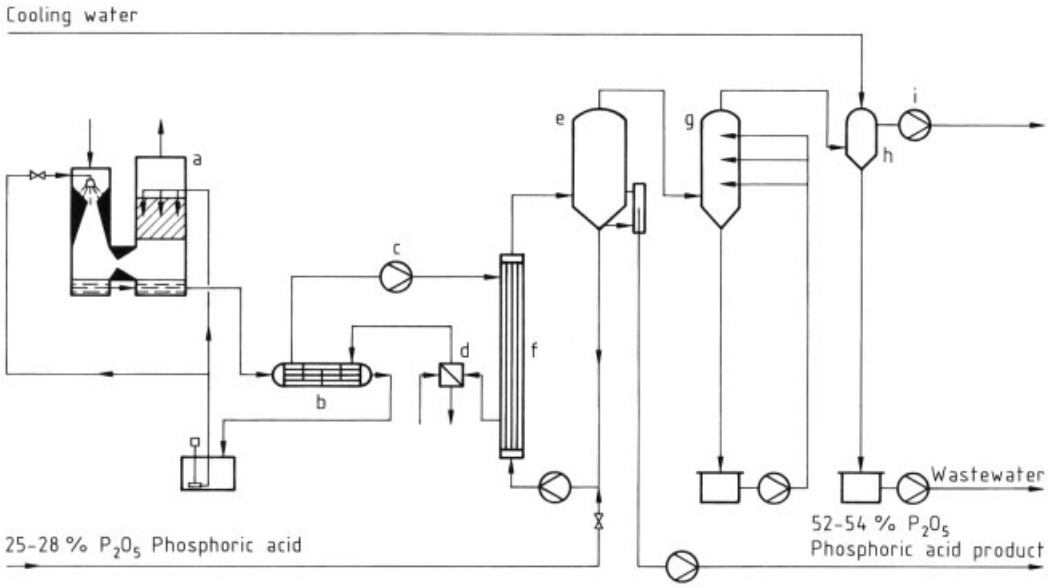

Figure 21. Phosphoric acid concentration using sulfuric acid waste heat
a) Sulfuric acid plant intermediate absorber; b) Sulfuric acid cooler; c) Hot water circuit; d) Trim cooler; e) Evaporator; f) Heat exchanger; g) Fluorine scrubber; h) Condenser; i) Vacuum pump

waste heat for reconcentration of spent sulfuric acid in a venturi system (see Section 4.3.2) [75, 84].

A significant step toward more highly energy-efficient plants was taken with the introduction of heat recovery systems generating steam from the acid circuit. Thus, part of the heat formed in the absorption circuits (and normally dissipated with the cooling water) is transferred to a higher level and used for the production of low-pressure steam at 6 – 10 bar. Depending on the nature of the plant, up to 0.5 t of steam per tonne of acid can be produced in this way in addition to the high- or medium-pressure steam obtained from a sulfur-burning plant. The overall energy efficiency is thus increased to $> 90\,\%$ based on a "standard sulfur-burning" plant [178, 179].

Such heat recovery systems are normally installed in the intermediate absorption circuits. Intermediate absorption is divided into two stages: a first stage operated at ca. 180 – 200 °C and a second stage at a "normal" absorption temperature of ca. 80 °C. About 95 % of the incoming SO_3 is absorbed in the first stage, where no sensible heat is removed from the gas. Acid is circulated at ca. 180 – 200 °C; the heat, which is a result of absorption/acid formation, is transformed into steam in a special boiler, usually of the kettle type. In the second stage, remaining SO_3 is absorbed and the gas is cooled to ca. 80 °C in order to provide appropriate gas conditions for downstream plant elements. Heat formed in the second stage absorber is normally used for boiler feed-water preheating or other heating purposes.

At present, two such systems are available on the market, representing different approaches to realizing the desired end:

1) The Monsanto HRS system
2) The Lurgi low-pressure steam recovery system

The Monsanto HRS system is based on investigations showing that "normal" stainless steels are subject to low corrosion rates when the acid concentration is $> 99\,\%$. That "window of concentration" permits the design and construction of a system consisting of 300-grade stainless steel and incorporating a first-stage absorption tower (with a countercurrent flow absorber, piping, heat exchanger, etc.). No brick lining is required. It is self-evident that the acid concentration must be carefully controlled, but this is quite feasible with modern instrumentation. The second-stage absorption circuit corresponds in principle to a standard absorption system. A typ-

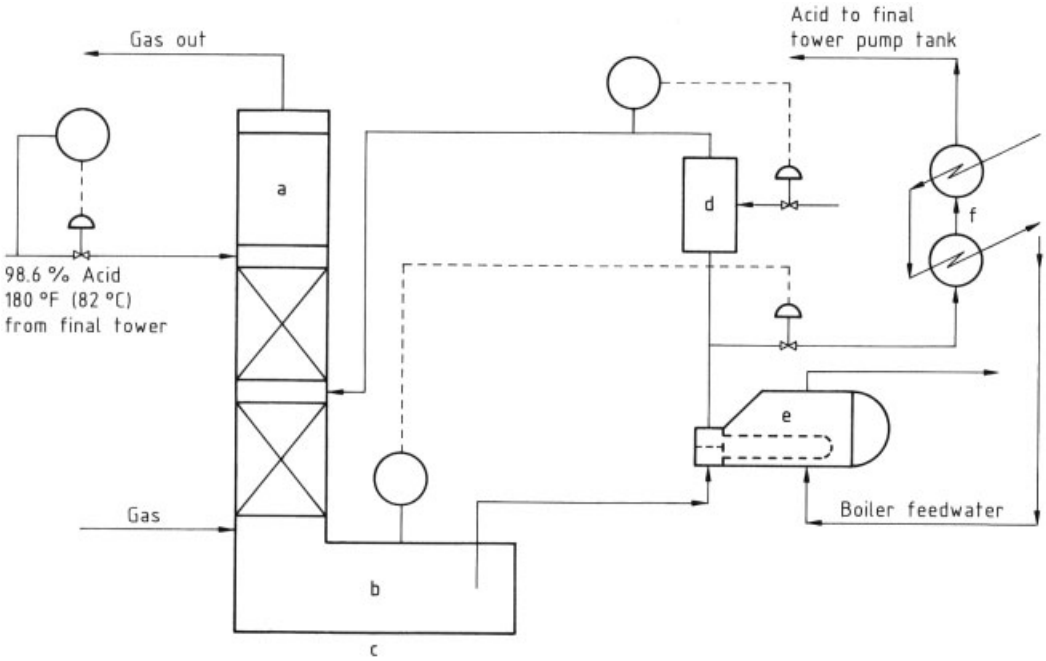

Figure 22. Monsanto heat recovery system
a) Mist eliminators; b) Pump boot; c) Heat recovery tower; d) Diluter; e) Heat recovery system boiler; f) Heat recovery system water heaters

ical flowsheet for a Monsanto HRS system is shown in Figure 22.

Lurgi has taken a different approach for its low-pressure steam recovery system. In cooperation with Thyssen and Krupp-VDM, a stainless steel was qualified for sulfuric acid service over a relatively wide range of concentrations and temperatures (see Fig. 23). The material is a ferritic stainless steel with the German material code 1.4575 and known as Superferrit (Thyssen) or Cronifer 2803 (Krupp-VDM). This material provides excellent corrosion resistance, but handling and equipment fabrication require special precautions because of brittleness. Lurgi has chosen here to again make use of its venturi absorption systems, already familiar from other sulfuric acid plant applications (see Section 4.3.2). The venturi unit constitutes the first-stage absorber, to which a separate acid circulation system is connected, including a pump tank, pumps, piping, etc. Heat generated in this system is transformed into low-pressure steam in a specially designed boiler. Second-stage absorption takes place in a small packed tower connected to the venturi unit. The venturi unit, the pump tank, the heat-exchanger tube bundle, piping, etc., are all fabricated from 1.4575 steel. Alternatively, the venturi unit and pump tank can be of a standard design: mild steel with a brick lining. A simplified flow diagram is provided in Figure 24.

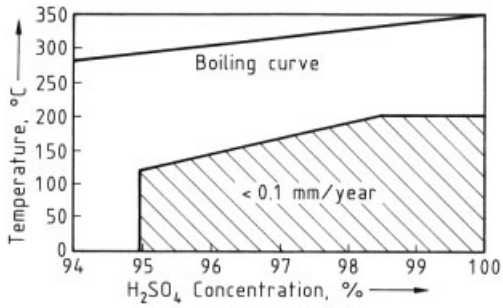

Figure 23. Range of application for Superferrit 1.4575 in concentrated sulfuric acid

With respect to the pumps, valves, etc., special cast alloys have been developed for the required high-temperature service. Additional facilities, such as continuous corrosion measurement and noise monitoring systems, are available for such heat-recovery systems as accessories to facilitate monitoring of all the relevant process data.

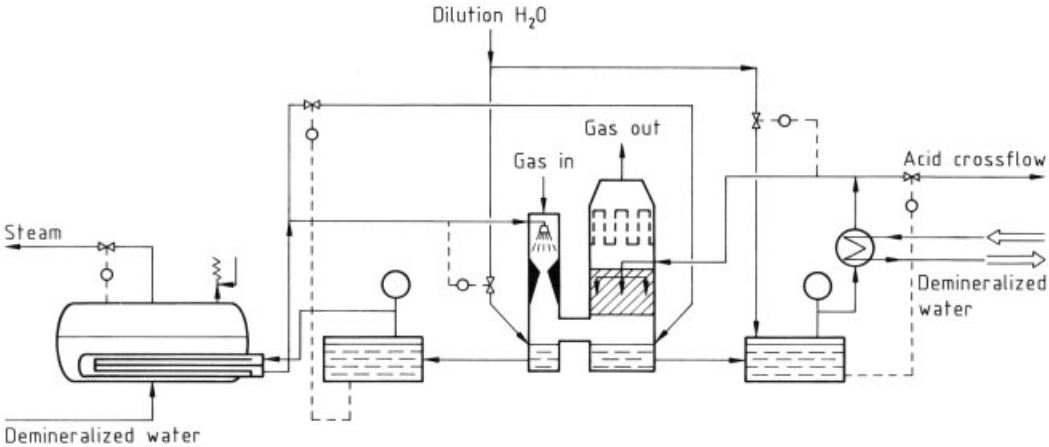

Figure 24. Low-pressure steam recovery system

Pressure-Contact Process. A specific variant of the sulfur-burning double-absorption process is the pressure-contact process, in which sulfur combustion, sulfur dioxide conversion, and sulfur trioxide absorption are all effected at elevated pressure. The idea of conducting a contact process under pressure is not new – it was first suggested over 50 years ago [86] and the notion has since been reiterated from time to time [87–91]. So far, however, is has only been embodied in one industrial double-absorption plant with a capacity of 550 – 575 t/d H_2SO_4 operated by PCUK in France [92–94]. The maximum pressure in this plant is 5 bar. It first assumed operation in 1972.

After drying in a conventional tower under normal pressure, combustion air is compressed to 5 bar for introduction into the sulfur furnace. Liquid sulfur is supplied through a special burner, and it burns at a maximum temperature of 1800 °C. The combustion gas, containing about 12 vol % SO_2 and 9 vol % O_2, is introduced at ca. 1200 °C into a fire-tube waste-heat boiler, after which it passes through a gas filter before being subjected to three stages of conversion in separate converter vessels. The process gas is then cooled and introduced into the intermediate absorption stage. Downstream from the intermediate absorber there is one further converter, from which gas passes to the final absorber. Tail gas leaving the final absorber is reheated and expanded in a turbine coupled with the air compressor. It is asserted that about two-thirds of the compression energy is recovered in this way [93].

As in an atmospheric-pressure double-absorption plant, the process gas is cooled in gas – gas heat exchangers before entering the intermediate absorber. In the pressure-contact process, however, only one exchanger is used for reheating cold gas returning from the intermediate absorber to the fourth converter stage. Another one serves to reheat the tail gas from the final absorber before it enters the turboexpander. Heat is recovered for steam generation in economizers, positioned downstream from the fourth converter and the expansion turbine, as well as in a waste-heat boiler after the sulfur furnace.

Compared with a conventional double-absorption process, two special advantages have been claimed for the pressure-contact process:

1) The position of chemical equilibrium in the sulfur dioxide oxidation reaction (Eq. 10) is more favorable, permitting a higher conversion efficiency with less catalyst. The PCUK plant is reported to have achieved 99.8 – 99.85 % conversion with a catalyst charge of only about 76 L per tonne of daily sulfuric acid production capacity [93]. The tail-gas sulfur dioxide content is reported to be reduced to about 200 – 250 ppm by volume. However, the high temperature in the sulfur furnace increases the rate of nitrogen oxide formation (→ Sulfur Dioxide, Chap. 4.4.).
2) On account of the lower operating volume with respect to the converter gases, smaller equipment can be used. This reduces the material and site-area requirements, and it raises the capacity limit on shop-fabricated equip-

ment. The resulting capital and construction cost savings are said to be ca. 10 – 17 % relative to a typical double-absorption plant [93]. It should be mentioned, however, that these savings would be nullified in many countries by the cost of conforming with requirements for extra wall thickness and higher-grade construction materials arising from safety regulations regarding pressure vessles.

The principal disadvantage of the pressure-contact process in comparison with a conventional double-absorption process is that it consumes more power and produces less steam. The magnitude of this negative effect on operating costs depends on the value of energy.

4.1.4.2. Double-Absorption Processes Based on Metallurgical Gases

Double-absorption plants based on metallurgical gases differ from hot-gas plants based on sulfur combustion in that cold feed gases must first be heated to the converter-inlet temperature using energy liberated in the oxidation of sulfur dioxide.

Process Description. Figure 25 is a flow diagram of a typical double-absorption plant based on metallurgical gases, with a four-bed converter in a (2 + 2) configuration for the production of 98 – 98.5 % sulfuric acid [40]. Sulfur dioxide-containing gases enter the sulfuric acid plant from the gas-cleaning system free of dust and mist but still saturated with moisture. To adjust the O_2/SO_2 volume ratio to the required value of ca. 1.0 – 1.1, air may need to be added to the gas before it enters the dryer. After drying (Section 4.1.3.1), the sulfur dioxide-containing feed gas, at about 50 °C, is compressed by the main blower to a pressure of ca. 1.35 – 1.45 bar, depending on the gas-pressure drop through the plant. The liberated heat of compression raises the gas temperature to ca. 90 – 100 °C.

The sulfur dioxide-containing gases are heated to the converter-inlet temperature of ca. 420 – 450 °C by indirect heat exchange with hot gases from, successively, the fourth, third, and first catalyst beds. Converted gases are simultaneously cooled to the appropriate inlet temperature for the succeeding bed and for the final absorber. In some cases, a boiler element may be installed ahead of the intermediate or final absorber. The converter system and the inter-

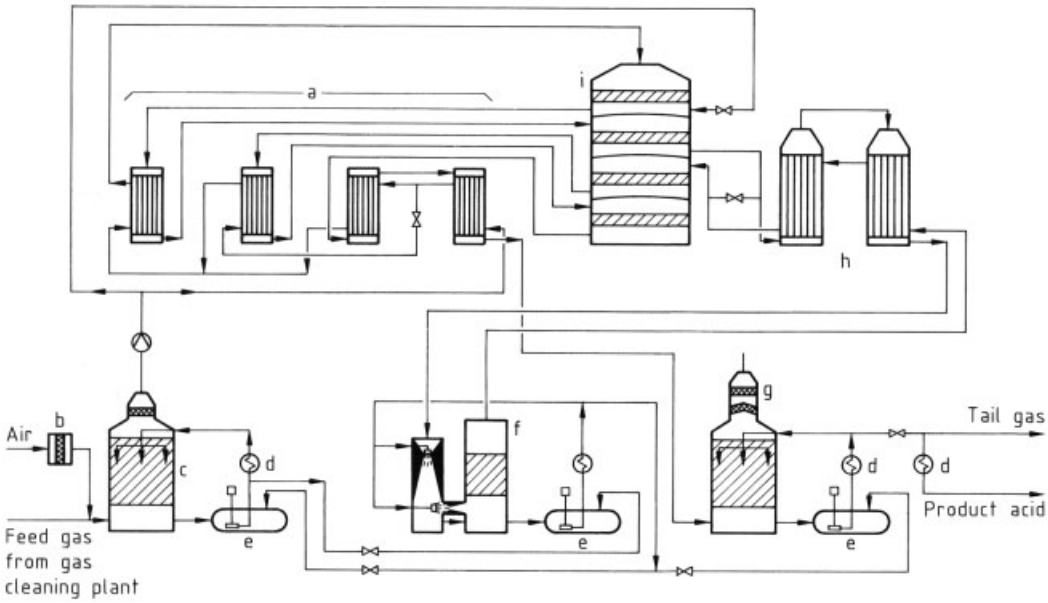

Figure 25. Double-absorption sulfuric acid process for metallurgical feed gas (Lurgi)
a) Gas – gas heat exchangers; b) Air filter; c) Drying tower; d) Cooler; e) Acid pump tanks; f) Intermediate absorber; g) Final absorber; h) Intermediate heat exchangers; i) Converter

mediate heat exchangers, as well as the intermediate and final absorption stages, are similar to those in the sulfur-based double-absorption process described previously, except that in this case the intermediate absorption stage is placed after the second catalyst bed rather than the third. Product sulfuric acid of 98 – 98.5 % concentration is withdrawn from the final absorber circuit. If 93 – 97 % sulfuric acid is to be produced, acid discharged from the dryer circuit is preferably stripped with air, which is subsequently used to adjust the O_2/SO_2 ratio of the feed gases. This removes any dissolved sulfur dioxide, reintroducing it into the plant with the feed gas. Most of the water requirement for sulfuric acid production can be satisfied by water vapor absorbed in the gas dryer.

Double-absorption plants of this type are normally designed for contact gases containing 5 – 10 vol % SO_2. Attainable final SO_2 conversions are > 99.5 %. In contrast to cold-gas plants based on pyrite roasting gases, the primary purpose of which is to produce sulfuric acid, sulfuric acid plants associated with nonferrous metal smelters and acid decomposition plants [95–97] have the primary function of cleaning the off-gases. They must therefore be adaptable to frequent large fluctuations in the output and composition of off-gases from the preceding plant. The lower limit for the sulfur dioxide concentration of off-gases subject to processing in a double-absorption plant is determined mainly by two criteria: it must be possible to maintain the water balance, and autothermal operation must be assured. The water balance can generally be maintained with a sulfur dioxide concentration of ca. 3.5 – 4 vol % SO_2 when cooling water at ca. 25 °C is available for the gas cleaning system (see Section 4.1.3.1 and Figure 13).

At the other extreme, modern nonferrous metal smelters generate gases with high SO_2 levels; SO_2 concentrations of up to 10 or 11 vol % may be achieved after mixing/dilution. In such a case the reaction heat developed in the converter system exceeds the requirements for heating the cold gas. The result is an excessively high gas temperature at the entrance to the absorption tower. Boiler elements (economizers, evaporators, superheaters, etc.) are therefore often incorporated into the converter scheme to remove excess heat. A typical flow diagram for such a plant is shown in Figure 26 [180].

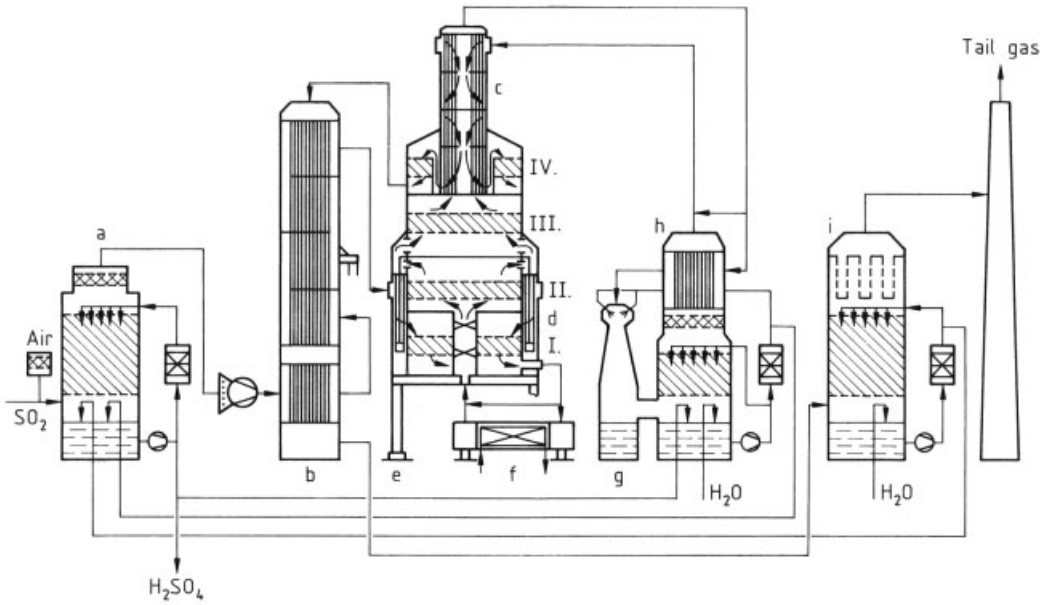

Figure 26. Contact acid plant for metallurgical gases with heat recovery
a) Drying tower; b) Heat exchanger 4; c) Heat exchanger 3; d) Heat exchanger 2; e) Converter; f) Superheater; g) Intermediate absorber; h) Heat exchanger 5; i) Final absorber

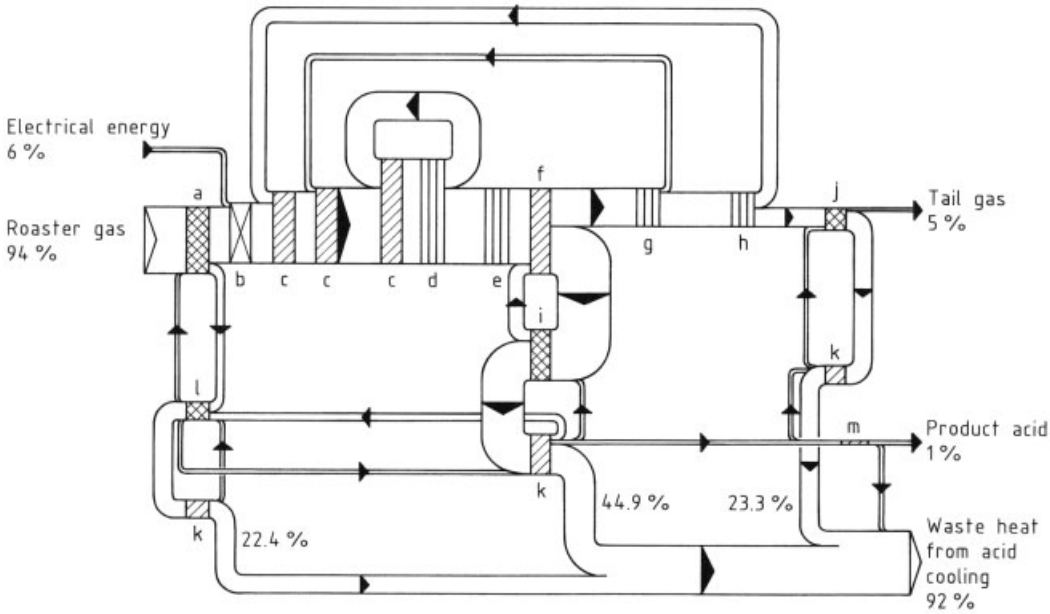

Figure 27. Sankey energy-flow diagram for a 1000-t/d metallurgical double-absorption sulfuric acid plant (feed gas: 8.5 % SO_2)
a) Drying tower; b) Blower; c) Heat exchanger; d) Catalyst bed 1; e) Catalyst bed 2; f) Intermediate heat exchangers; g) Catalyst bed 3; h) Catalyst bed 4; i) Intermediate absorber; j) Final absorber; k) Acid coolers; l) Acid tank; m) Product acid cooler

Energy Balance. Figure 27 is a Sankey diagram showing the energy balance for a double-absorption plant analogous to Figure 25 based on metallurgical gases. An essential characteristic of a conventional cold-gas plant is that almost all the energy is discharged as waste heat at low temperature. Of the total energy input, 94 % is accounted for as chemical energy released in the conversion of sulfur dioxide to sulfuric acid and in the condensation of water in the dryer. At a feed-gas concentration of 8.5 vol % SO_2 and a dryer inlet temperature of 30 – 40 °C, a total of about 2.7 GJ of thermal energy is liberated per ton of sulfuric acid. This corresponds to a thermal output of ca. 31 MW for a 1000 t/d plant.

About 45 % of the total energy is discharged through the intermediate absorber acid-cooling system, about 23 % through the final absorber acid-cooling system, and about 22 % through the dryer-acid cooling system. In order to maintain autothermal operating conditions, gas-heat losses in the absorption system must be covered by the heat of reaction released in the converter system [40]. At thermal equilibrium, the heat content initially supplied to the converter system during preheating is thus maintained.

For a particular gas throughput, the gas-heat losses remain constant, but the lower the concentration of sulfur dioxide in the feed gas, the less heat is released in the converter system. The lower limit for maintaining thermal equilibrium is about 5 vol % SO_2 for a conventional plant. Minimizing the gas-heat losses in the intermediate absorption stage through the use of venturi-type cocurrent absorbers and the hot absorption principle makes it possible to maintain autothermal operation even when processing contact gases containing as little as 4 vol % SO_2.

Proper design of the gas – gas heat exchangers in relation to the sulfur dioxide concentration is a prerequisite for maintaining thermal equilibrium. The lower the sulfur dioxide content, the smaller will be the difference between the mean inlet and outlet temperatures at each of the catalyst beds. With lower temperature differences, the specific heat-exchanger surfaces must be larger. For example, processing gases with 4 vol % SO_2 rather than 8 vol % SO_2 means that the required specific heat-exchanger surface is more than three times as large [98].

At somewhat higher sulfur dioxide concentrations (8 – 9 vol %), the required specific hea-

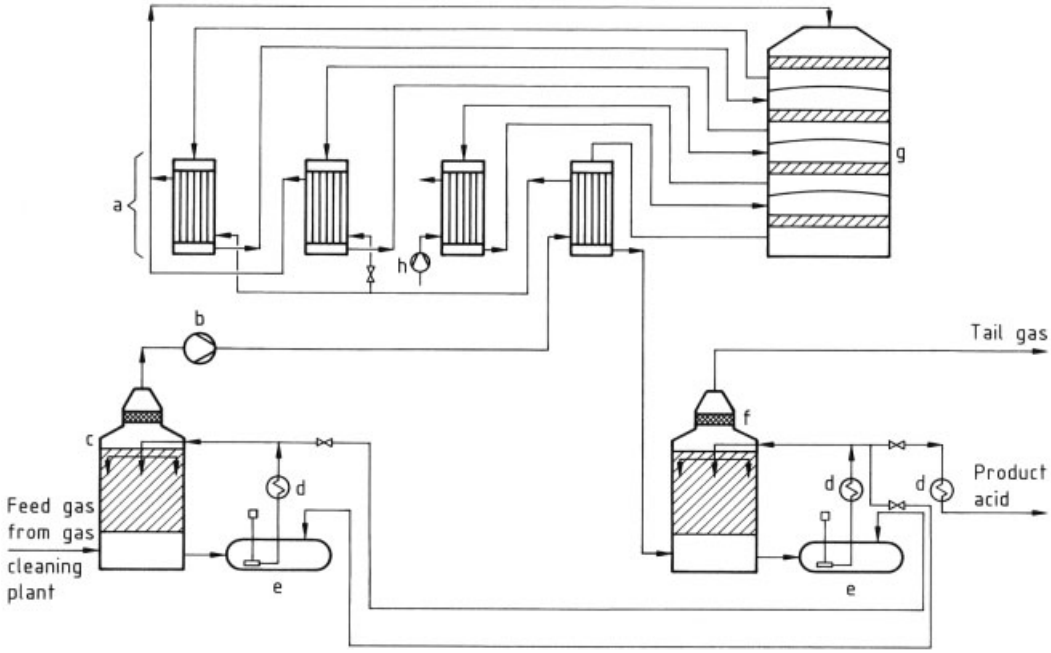

Figure 28. Ordinary (single-absorption) contact sulfuric acid plant for metallurgical roaster gas
a) Gas – gas heat exchangers; b) Main blower; c) Drying tower; d) Cooler; e) Acid pump tanks; f) Absorber; g) Converter

texchanger surface reaches a minimum. At still higher sulfur dioxide concentrations it rises again, and in this case it is advisable to install a waste-heat boiler element to convert the surplus reaction heat into steam.

Heat Recovery. In a conventional cold-gas double-absorption plant for processing relatively low-grade sulfur dioxide-containing feed gases, there is no excess high-temperature heat that can be used for generating high-pressure steam (see Fig. 27). However, if the sulfuric acid unit is associated with a fluidized-bed roaster for sulfide ores, high-pressure steam is generally available from a waste-heat boiler system utilizing at least part of the heat of the roasting reaction. In pyrite or zinc blende roasters, for instance, steam is produced at about the same rate per tonne of sulfuric acid as in a double-absorption plant based on sulfur combustion. Where the sulfuric acid plant is linked to a modern smelter (Outokumpu, QSL, Kivcet, etc.), high-level SO_2 gases are available for conversion, so the removal of surplus reaction heat in a waste-heat boiler element becomes almost obligatory. To increase the output of high-pressure steam, low-temperature heat from the absorber-acid circuits can also be utilized for preheating the boiler feed water. Recovery of surplus heat from the acid system, which accounts for ca. 92 % of the entire waste heat produced in a cold-gas plant (see Fig. 27), is of the same economic importance as in a plant based on sulfur combustion. Systems for acid waste-heat utilization have been installed in a number of roaster gas-based double-absorption plants, generating hot water for industrial purposes (e.g., P_2O_5 concentration), or for district heating schemes [99, 100].

4.1.4.3. Ordinary Single-Absorption Processes

Before the introduction of the double-absorption process, the ordinary contact or single-absorption process was the standard sulfuric acid production process [19–23]. Figure 28 is an example of a flow diagram for an ordinary contact plant based on metallurgical gases equipped with a four-bed converter. Before entering the dryer, the cleaned sulfur dioxide-containing gases are generally diluted extensively with air to an O_2/SO_2 volume ratio of at least 1.7. The dried gases are then compressed to

the pressure required to overcome the pressure drop of the overall plant. Although this is about 0.1 bar less than the pressure drop in a comparable double-absorption plant, almost the same specific blower capacity is needed on account of the substantially higher specific gas volume that results from the greater dilution of the feed gas.

The feed gas is heated stepwise in a series of indirect gas – gas heat exchangers cooling first the hot gases from the last catalyst bed and then the gases between the individual beds. The cooled sulfur trioxide-rich gases from the last bed are passed into the SO_3 absorber.

In principle, the ordinary contact process retains more gas heat than the double-absorption processes, because there is no intermediate absorption stage with its associated heat losses. In a roaster-based plant the surplus heat is simply discharged, for instance by indirect air cooling in an air cooler after the third bed (Fig. 28). The reaction heat in a plant based on sulfur combustion is recovered by steam generation in economizers and evaporators. About 10 % more steam is therefore produced per tonne of sulfuric acid in a sulfur-burning ordinary contact plant than in a double-absorption plant of the same capacity [73].

Because of the favorable energy balance, an ordinary contact plant can also process gases with a lower sulfur dioxide content [101]. The lowest sulfur dioxide concentration at which thermal equilibrium can still be maintained in continuous operation without recourse to an external heat source is ca. 2 vol %. The water balance in a sulfuric acid plant based on such a low SO_2 content can be maintained with the aid of a predryer – reconcentrator system developed by Lurgi [51] (see Section 4.1.3.1).

The maximum overall SO_2 conversion attainable in an ordinary contact plant without intermediate sulfur trioxide absorption is 98 –99 % in continuous operation under favorable conditions. Assuming a feed-gas concentration of 7 vol % SO_2, this corresponds to a sulfur dioxide content in the tail gas of about 1600 ppm by volume. Stringent environmental protection regulations introduced since the early 1970s in many industrialized countries demand a higher standard of performance than this. Most new sulfuric acid plants are therefore designed according to the double-absorption principle. Ordinary contact plants are used today only for processing off-gases of extremely low sulfur dioxide content, such as highly dilute smelter off-gases, which cannot sustain autothermal operation in a double-absorption plant. At low feed-gas concentrations (e.g., 3 vol % SO_2) and with the help of highly active (cesium-doped) catalysts, tail-gas SO_2 concentrations are attained that are equivalent to the levels associated with "standard" double absorption plants based on high SO_2 concentrations. In such cases, the ordinary contact process, which produces concentrated sulfuric acid as a marketable end product, is preferable to other off-gas purification methods such as alkaline washing. The SO_2 separation efficiency with these alternative approaches is seldom higher, and they usually lead to unmarketable end products or polluting waste materials.

When necessary, an existing ordinary contact plant can be brought into conformity with strict antipollution legislation either by retrofitting with an intermediate absorption stage, transforming it into a double-absorption plant, or by addition of a tail-gas purification system. Acid waste-heat recovery systems have also been installed in ordinary contact plants [79, 102].

4.1.4.4. Wet-Catalysis Processes

Wet-catalysis processes differ from other contact sulfuric acid processes in that the feed gas still contains moisture when it comes into contact with the catalyst. Sulfur trioxide formed by catalytic oxidation of the sulfur dioxide reacts instantly with the moisture to produce sulfuric acid in the vapor phase to an extent determined by the temperature. Liquid acid is subsequently formed by condensation of the H_2SO_4 vapor and not by the absorption of sulfur trioxide in concentrated sulfuric acid, as in a contact process based on dry gases. The concentration of the product acid depends on the H_2O/SO_3 ratio in the catalytically converted gases as well as on the condensation temperature [103]. The wet-catalysis process is especially suitable for processing the wet, dust-free gases obtained in the combustion of hydrogen sulfide-containing off-gases, which need only be cooled to the converter inlet temperature of ca. 440 °C. Processing these moisture-laden gases in a conventional cold-gas plant would necessitate cooling to an

economically unacceptable extent in order to remove the large excess of moisture, an expedient that would only be justified for a gas with a high dust content and a relatively high sulfur dioxide concentration.

Classical Process. The classical wet-catalysis process was developed by Lurgi as early as the 1930s. It is illustrated schematically in Figure 29. Dust-free hydrogen sulfide-containing gases are burned at ca. 500 – 1000 °C (depending on their composition) with a sufficient excess of air to produce a contact gas with an O_2/SO_2 volume ratio of ca. 1.1 – 1.3. The hot, wet, sulfur dioxide-containing gas is cooled to ca. 440 °C in a waste-heat boiler (or, where the volume of gas involved is very small, in a radiation cooler) and passed directly into the converter. The converter contains three or four catalyst beds, between which cold quench air is added to cool the hot reaction gases.

Gas from the last bed enters a condensation tower at ca. 420 – 430 °C. Here it is cooled to the condensation temperature of sulfuric acid by irrigation with circulated, cooled sulfuric acid. Good contact between the gas and the liquid, as well as a high irrigation density, are essential to ensure that the gas is cooled insofar as possible at the gas – liquid interface of the irrigation acid, causing the condensing sulfuric acid to coalesce with the irrigation acid. This largely avoids formation of the acid mist that would result if sulfuric acid were to condense from the gas phase due to the mixing of cold and hot gas streams. Any mist that does form is subsequently removed from the tail gases by passing them through glass-fiber candle filters. The sulfuric acid thus obtained is fed to the acid circuit of the condensation tower. The gas temperature at the condensation-tower outlet is maintained at ca. 60 °C to restrict H_2SO_4 and SO_3 emission.

Classical wet-catalysis plants are usually designed for the production of 78 % sulfuric acid. Even if the H_2O/SO_3 ratio of the gases leaving the converter were high enough for the product concentration to exceed 78 % (assuming complete condensation of the sulfur trioxide and water), it would not be advisable to operate at higher acid strength because the 80 – 90 % H_2SO_4 concentration range is quite critical with respect to the corrosion-resistance of most of the usual construction materials. The water content of moist hydrogen sulfide-containing gases is normally too high to sustain a sulfuric acid concentration > 90 %.

Acid leaving the condensation tower at ca. 80 – 85 °C is cooled to 50 – 60 °C for recirculation. If necessary, the concentration is adjusted to 78 % H_2SO_4 by addition of process water.

A sulfur dioxide conversion of about 97.5 % is achieved in the wet-catalysis process with a four-bed converter. Since wet-catalysis plants are almost exclusively used for desulfurization of hydrogen sulfide-containing off-gases, the corresponding acid production capacities are comparatively small. The largest wet-catalysis plant built by Lurgi in the United States has a capacity of 250 t/d of H_2SO_4.

The classical wet-catalysis process cannot normally be used for processing wet hydrogen sulfide-containing gases that are too dilute to sustain combustion without supplemental fuel, since the water content of the resulting sulfur dioxide-containing gases would be too high to

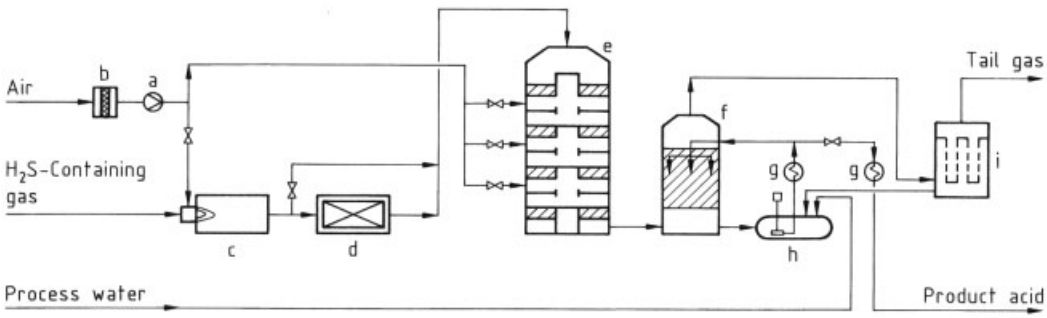

Figure 29. Wet-catalysis sulfuric acid plant processing H_2S-containing gases
a) Blower; b) Air filter; c) Combustion furnace; d) Waste heat boiler; e) Quench-cooled converter; f) Condensation tower; g) Coolers; h) Acid pump tank; i) Candle filter

permit production of 78 % sulfuric acid. In any case, the opportunities for marketing 78 % sulfuric acid are somewhat limited, and for that reason the wet-catalysis method has been developed further to permit the conversion of gases with extremely low hydrogen sulfide content to sulfuric acid more concentrated than 90 % [104–108]. Examples of such processes are the Concat process (Lurgi) and the WSA process (Haldor Topsoe). Another special development in this area is the wet – dry contact process with intermediate condensation.

Concat Process. Lurgi's Concat process, illustrated schematically in Figure 30, differs from the classical wet-catalysis process principally in the way in which the condensation is accomplished. This occurs in two stages, combining a venturi unit with an irrigated packed tower [103]. In the first stage, the so-called hot condensation, the major part of the sulfur trioxide and/or sulfuric acid condenses at ca. 180 – 230 °C from the hot gases, which leave the converter. At this high condensation temperature only a very small portion of the free water passes from the gas stream into the condensed sulfuric acid on account of the relatively high partial pressure of water vapor over sulfuric acid at this temperature. Therefore, acid with a 93 % H_2SO_4 content can be obtained in the first condensation stage. This section utilizes a venturi scrubber into which the circulating acid is injected in a cocurrent sense relative to the gas to ensure intimate mixing of the gas phase with the liquid phase and the most efficient removal possible of the heat of formation and latent heat of condensation of the sulfuric acid.

As the gas passes between the first and second condensation stages, additional air is injected both for cooling purposes and to reduce the water-vapor partial pressure through dilution. The second stage is a packed tower irrigated with relatively dilute sulfuric acid from which water evaporates as a result of contact with the still moderately hot gas. In the process, the gas is cooled further, and this promotes condensation of the residual sulfuric acid vapor. Gas leaving the condensation tower is freed from remaining acid mists in a candle filter before being discharged to the stack.

Sulfuric acid separated in the candle filter drains into the acid circuit of the condensation tower, from the sump of which excess acid flows over to the hot-condensation stage. Hot acid in the venturi circuit is cooled indirectly. To ensure that it does not exceed the maximum permissible inlet temperature for the particular material of which the cooler is made, part of the cooled acid is recycled to the pump tank. Excess acid that accumulates in the circuit of the first stage is discharged as product after additional cooling.

On account of this two-stage condensation and the cocurrent hot-condensation principle, the Concat process is able to process gas with a sulfur dioxide content of less than 1 vol % and a water content in excess of the stoichiometric requirement, resulting in 78 % or 93 % sulfuric acid. A sulfur dioxide content < 200 ppm can be attained in the tail gas with a two- or three-bed converter depending on the sulfur dioxide concentration of the incoming gases.

Topsoe Wet Sulfuric Acid (WSA) Process [181]. The Topsoe WSA process is used for

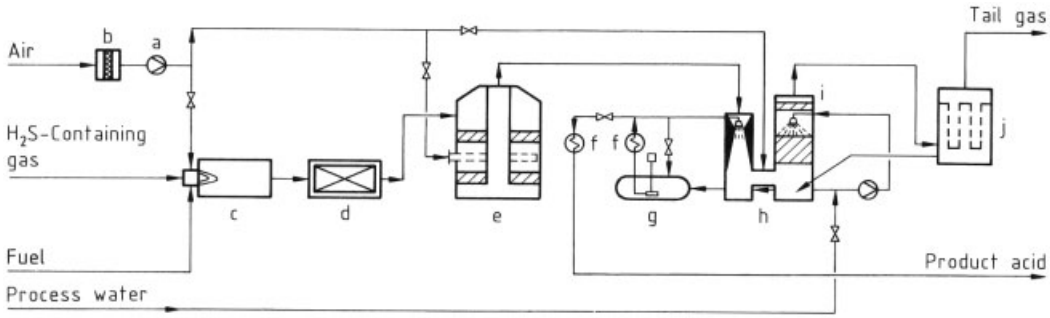

Figure 30. Lurgi Concat (hot condensation) wet-catalysis process
a) Blower; b) Air filter; c) Combustion furnace; d) Waste heat boiler; e) Converter; f) Coolers; g) Acid pump tank;h) Hot-condensation venturi; i) Condensation tower; j) Candle filter

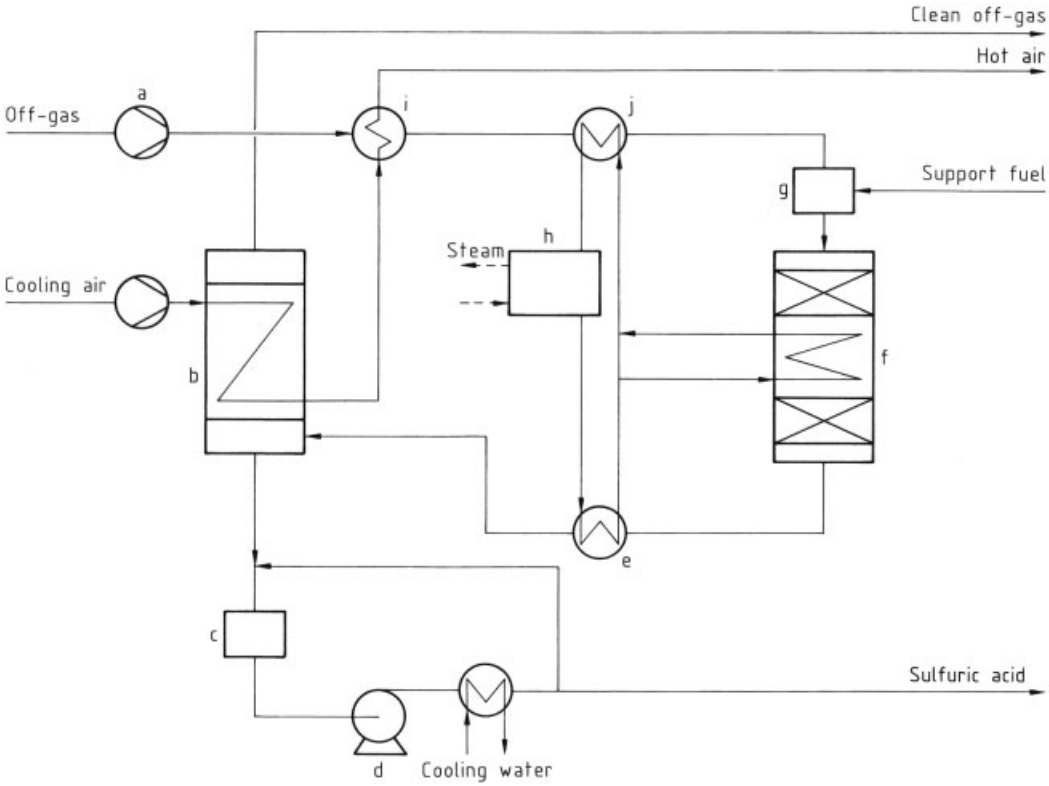

Figure 31. Haldor Topsoe WSA process
a) Fan; b) WSA condenser; c) Acid tank; d) Pump; e) Gas cooler; f) SO_2 converter; g) Burner; h) Salt system; i) Gas preheater; j) Gas heater

treating wet SO_2- containing gases – either "cold gases" (e.g., smelter off-gas) or "hot gases" (from catalytic or thermal incineration units) – without the need for prior drying. The process efficiently recovers sulfur in the form of 93 – 98 % sulfuric acid despite off-gas water contents that may exceed 20 %. Recoveries > 99 % have been achieved. The main components of the process are an SO_2 converter, a heat-exchanger system, and a WSA condenser (see Fig. 31). Upstream from the WSA process, dust and trace materials such as arsenic and hydrogen fluoride are removed by scrubbing if these are present in significant concentrations. Cold gas is heated in two steps. The first heat exchanger uses hot air from the WSA condenser, whereas the second uses the hot converted gases. An in-line burner supplies whatever additional heat is required to reach the appropriate inlet temperature for the SO_2 converter (400 – 420 °C). In cases where the SO_2 concentration exceeds ca. 2.5 %, no supplementary heat is necessary. Surplus heat is used for steam production or dispersed in a cooler.

In the converter, SO_2 is oxidized to SO_3 by means of a sulfuric acid catalyst. Conversion is accomplished either in a single-stage adiabatic converter or in a two-stage converter with intermediate cooling, depending upon the SO_2 concentration and the required level of sulfur removal.

In the next heat exchanger the temperature of the gas is reduced. Cooling causes the SO_3 to react with water in the off-gas to form sulfuric acid vapor. The heat released by this gas-phase hydration is transferred to incoming cold SO_2-containing gases.

Finally, the sulfuric acid vapor is condensed as concentrated acid in the WSA condenser/concentrator, which is an air-cooled glass-tube device. The sulfuric acid-containing gas passes on the tube side and is cooled to ca. 100 °C by ambient air passing on the shell side.

Sulfuric acid condenses selectively and flows downward on the walls of the glass tubes. The acid is concentrated in the lower parts of the tubes by countercurrent contact with hot off-gas; the acid is subsequently collected at the bottom of the condenser. Hot acid leaving the condenser is cooled in a water-cooled plate cooler before being pumped to a storage tank.

Careful control of the temperature and of sulfuric acid droplet formation makes it possible to keep the level of sulfuric acid in the cleaned off-gas low without the use of any subsequent wet electrostatic precipitator or low-velocity mechanical filter. Cleaned off-gas leaves the top of the condenser and goes to a stack. Cooling air leaves the condenser at ca. 200 °C, and is used for preheating the cold feed gas.

In power-plant boiler applications the WSA process is combined with a selective catalytic reduction (SCR) NO_x removal system. The first heat-exchanger step is omitted, because the off-gas is already hot as received after dust filtration. A rotary gas – gas heat exchanger is used to heat the off-gas ahead of the SO_2 converter. The SO_2 converter is an adiabatic dust-tolerant type with a number of vertical catalyst beds placed in parallel array. The catalyst beds can be dedusted one by one while the SO_2 converter remains online without disrupting plant operation. Hot cooling air from the WSA condenser is used as boiler combustion air.

The main construction material employed in the WSA system is either carbon steel or low-alloy steel.

Wet – Dry Contact Process. The wet – dry contact process with intermediate condensation, a joint development by Lurgi and Süd-Chemie, is a modified double-absorption process based on sulfur combustion. Instead of dried air, industrial waste air contaminated with such sulfur compounds as hydrogen sulfide, carbon disulfide, or carbonyl sulfide is used to burn the sulfur [109]. The impurities are oxidized simultaneously with combustion of the sulfur. The resultant gases, containing 9.5 – 10 vol % SO_2 and less than the stoichiometric amount of water necessary for the formation of 98.5 % sulfuric acid, are passed into the converter after prior cooling in a waste-heat boiler (Fig. 32).

In contrast to the normal double-absorption process based on sulfur combustion, the moist

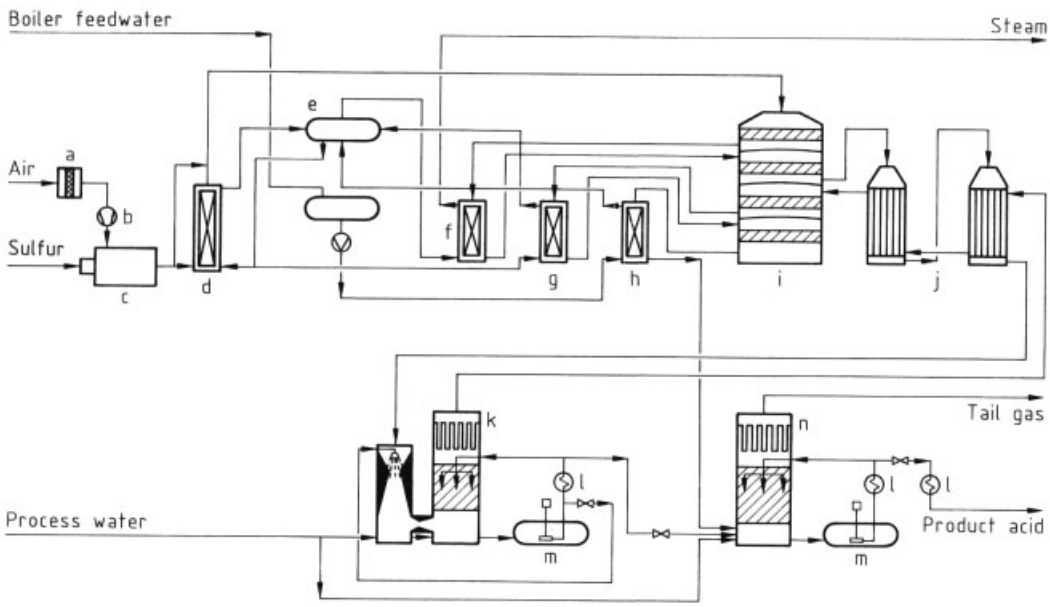

Figure 32. Wet/dry catalysis sulfuric acid plant with intermediate condensation/absorption
a) Air filter; b) Blower; c) Sulfur furnace; d) Waste-heat boiler; e) Steam drum; f) Steam superheater; g) Boiler; h) Economizer; i) Converter; j) Intermediate heat exchangers; k) Intermediate condenser/absorber; l) Coolers; m) Acid pump tanks; n) Final absorber

sulfur trioxide-containing gases formed in the first beds of a wet-catalysis system are passed on to a two-stage intermediate condenser. In the first stage, which is a vertical venturi unit, the entire water content of the gas and most of the sulfur trioxide are removed into hot sulfuric acid injected cocurrently. Remaining sulfur trioxide is absorbed in the second stage, a packed tower irrigated with cold acid. The concentration of the circulating acid is adjusted to 98.5 % by the addition of process water. Dry SO_2- containing residual gases from the intermediate condenser are processed in the same way as in a normal double-absorption plant (Section 4.1.4.1).

A plant of this type for the production of 120 t/d of sulfuric acid has been in operation at the Kelheim works of Süd-Chemie since 1980. This plant processes off-gases from a rayon staple factory containing ca. 0.5 vol % $H_2S + CS_2$ [109, 110].

4.1.5. Tail-Gas Treatment

The tail gases from a sulfuric acid contact plant consist chiefly of nitrogen and residual oxygen. They also contain sulfur dioxide at a low concentration that depends on the conversion efficiency attained in the conversion stages. The content of gaseous sulfur trioxide and sulfuric acid is essentially a function of the temperature and concentration of irrigation acid in the final absorber. Under unfavorable operating conditions (e.g., when the sulfur dioxide-containing converter feed gases are inadequately dried, or contain hydrocarbons) sulfuric acid mists may form which are not removed in the absorption system. This is true even when the concentration and temperature of the absorber acid are at their optimum values. The safest way to remove these acid mists is with a candle filter, although this is not very effective for removing excess sulfur trioxide, which may result from poor acid distribution in the absorber.

The sulfur dioxide level in the tail gases is dependent not only on the sulfur dioxide conversion achieved but also on the SO_2 content of the converter feed gases (Fig. 33). However, the specific sulfur dioxide emission rate (kilograms of SO_2 per tonne of 100 % H_2SO_4) is a function only of the sulfur dioxide conversion efficiency.

An essential prerequisite for attaining maximum sulfur dioxide conversion and correspondingly low sulfur dioxide emission is maintaining the gas conditions at the converter inlet (gas volume, gas temperature, sulfur dioxide concentration, and O_2/SO_2 ratio) as constant as possible. This is in general reasonably easy in a sulfuric acid plant based on sulfur combustion or metallurgical gases from a fluidized-bed furnace. But in sulfuric acid plants associated with metal smelters, especially copper smelters incorporating converter operations (→ Sulfur Dioxide, Chap. 5.5.1.), wide fluctuations in the gas-flow rate as well as in the off-gas sulfur dioxide concentration are quite normal [111, 112]. Consequently, the sulfur dioxide content of tail gases from the corresponding sulfuric acid plants themselves may vary quite noticeably from the optimum value. It is also almost impossible in most cases to avoid fluctuating emissions during start-up [113].

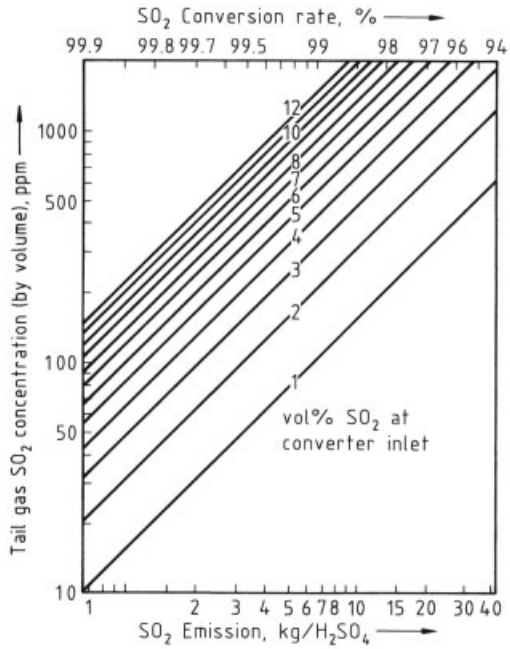

Figure 33. Sulfur dioxide emission as a function of SO_2 concentration at the converter inlet and the SO_2 conversion rate

Replacing the normal contact process by a double-absorption process results in a marked reduction in sulfur dioxide emissions [114–116]. The specific sulfur dioxide emission of a single-

absorption plant with a sulfur dioxide conversion of 97.5 % is about 17 kg of SO_2 per ton of H_2SO_4. Corresponding values for double-absorption plants with sulfur dioxide conversion efficiencies of 99.5 % and 99.7 % are ca. 3.3 kg and 2 kg of SO_2 per tonne of H_2SO_4 respectively (Fig. 33). For a plant with an H_2SO_4 capacity of 1000 t/d this translates into a reduction from 710 kg/h of SO_2 to ca. 138 kg/h or 83 kg/h. Theoretically, it would be possible to improve the conversion efficiency, and thus reduce emissions still further, by introducing an additional intermediate absorption stage, corresponding to a triple-absorption process. In practice, however, this method is not feasible, for both technical and economic reasons [80].

Environmental regulations – in Germany, for example – call for further improvement in the conversion efficiency normally achieved with a double-catalysis plant. At locations where the maximum permissible ground-level SO_2 concentration is exceeded because of emissions from other nearby sources, it may even become necessary to reduce SO_2 emissions below the value attainable with a conventional double-absorption process. Two approaches to a further reduction of the SO_2 emissions might be considered in such cases: installation of a fifth catalyst bed, or addition of a tail-gas treatment unit. Whereas the SO_2 reduction in a 5-bed converter system is achieved through an improvement in conversion efficiency, reduction of SO_2 emissions in a tail-gas treatment plant is effected by reacting the SO_2 with some type of scrubbing liquor. The 5-bed converter system will probably continue to be the most convenient solution for new plants so long as the target conversion efficiency remains at approximately 99.8 %. In the case of an existing plant, "add-on" tail-gas treatment may be feasible, although some processes of this type lead to byproducts that also require treatment. Of the known tail-gas cleaning processes, ammonia washing, the Wellman – Lord process, and the Sulfacid process are most suitable for application in normal contact plants; the Peracidox process is used only in conjunction with double-absorption plants. For a single catalysis plant, upgrading to a double-absorption operation remains an attractive alternative to a tail-gas scrubbing system.

Ammonia Washing. Sulfuric acid tail-gas purification processes based on ammonia washing are distinguished mainly by the method selected for reprocessing the sulfur dioxide-laden scrubbing liquor [111, 115–120]. Figure 34 shows a Lurgi-designed system capable of removing up to 95 % of the incoming sulfur dioxide. This high level of sulfur dioxide removal is made possible by a two-stage gas-scrubbing system.

Tail-gases from the absorber of the sulfuric acid plant are washed with circulated ammonia solution in two venturi scrubbers arranged in series. Unconverted sulfur dioxide in the tail gas dissolves as ammonium sulfite or ammoni-

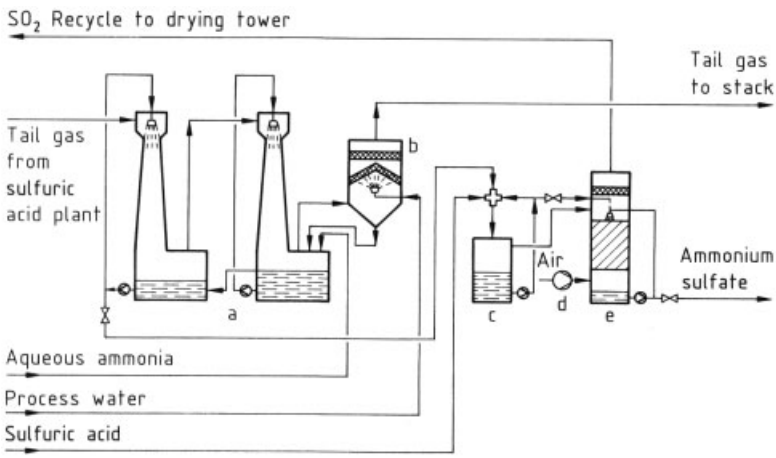

Figure 34. Ammonia scrubbing system for an ordinary (single-absorption) contact sulfuric acid plant
a) Ammonia scrubbers; b) Gas filter; c) Decomposer; d) Fan; e) Stripper

um hydrogensulfite. Any sulfur trioxide or sulfuric acid carried over from the absorber is transformed into ammonium sulfate or ammonium hydrogensulfate. Before being discharged to the stack, the cleaned gas is passed through a mist eliminator. Liquor drawn from the mist eliminator and fresh ammonia solution are then introduced into the second venturi srubber. Scrubbing solution from the second stage overflows to the first. Thus, the solution in the second stage, where the sulfur dioxide content of the tail gas is lower than in the first, has a higher pH and a lower sulfite concentration than the solution in the first stage, causing the partial pressure of sulfur dioxide to be reduced even further in this stage.

A bleed stream of sulfite-rich scrubbing solution is withdrawn from the first stage and introduced into a decomposer along with concentrated sulfuric acid. The acid decomposes the sulfites, liberating sulfur dioxide and forming an ammonium sulfate solution. This solution retains a certain amount of dissolved sulfur dioxide, which is subsequently stripped out with air. Moist sulfur dioxide gas from the decomposer is combined with the stripping air and recycled to the dryer of the sulfuric acid plant.

The economics of an ammonia tail-gas scrubbing process depend chiefly on the fate of ammonium sulfate solution discharged from the stripper [118–120]. The most favorable situation is one in which the sulfuric acid plant is part of a fertilizer complex, because in that case the solution can probably be used directly in the production of granular fertilizers, and ammonia is likely to be available at a reasonable cost. Otherwise, assuming there is an adequate market, the ammonium sulfate may be sold, so that at least part of the ammonia cost can be recovered through a credit for the ammonium sulfate. Normally, however, there is little prospect for selling the solution directly, and working it up into solid products would involve additional capital investment as well as energy, leading to high operating costs.

In a sulfuric acid plant designed to regenerate spent acid (→ Sulfur Dioxide, Chap. 5.6.1.), sulfite solution formed in an ammonia tail-gas scrubbing system can be disposed of in the spent-acid decomposition furnace. This circumvents the problem of finding an outlet for byproducts, but the resulting convenience must be paid for in lost ammonia values, since ammonia is oxidized to nitrogen. Also, decomposition of the sulfite solution and especially the evaporation of water require additional energy, which is not matched by the heating value of the ammonia that is burned.

Wellman – Lord Process. In the Wellman – Lord process, sulfur dioxide is washed from the acid-plant tail gas by a saturated aqueous solution of sodium sulfite and sodium hydrogensulfite (→ Sulfur Dioxide). Wet sulfur dioxide-containing gas expelled during thermal regeneration of the scrubbing solution is recycled to the contact plant. About 10 – 20 % of the recovered sulfur ends up as sodium sulfate owing to oxidation to sulfite in solution. This dissolved sodium sulfate cannot be decomposed, so it must be removed as a byproduct. The economics of the process are determined by the purity and marketability of the byproduct.

Sulfacid Process. The Sulfacid process is a tail-gas purification method (developed by Lurgi) that produces dilute sulfuric acid as a byproduct [111, 115, 121, 122]. Dry tail gas normally leaves the sulfur trioxide absorber of an acid plant at 75 – 80 °C. This is first humidified with steam and then passed through a reactor filled with activated carbon. Sulfur dioxide adsorbed by the wet activated carbon is oxidized to sulfuric acid at 50 – 80 °C by residual oxygen in the tail gas. This acid, together with traces of sulfur trioxide and sulfuric acid removed from the tail gas, is retained in the pores of the carbon. The purified gas is discharged to the stack at 40 – 80 °C; the attainable sulfur dioxide separation efficiency is normally > 90 % [121]. Adsorbed acid is washed out by spraying the activated carbon bed with water at regular intervals without interrupting the gas flow.

The concentration of acid discharged from the reactor is kept below 20 – 25 % in an industrial plant by the dilution effect of the flushing water. Increasing the acid concentration by decreasing the water flow-rate would reduce the solubility of sulfur dioxide and, consequently, the sulfur dioxide-removal efficiency. The resulting dilute acid is recycled to the contact plant, where it takes the place of process water.

If the amount of water introduced with the dilute acid exceeds the water requirements of

the contact plant, the dilute acid may be concentrated by means of the well-established venturi method (see also page 53). In this case acid leaving the Sulfacid reactor is passed on to a single- or double-stage venturi reconcentrator located upstream from the Sulfacid reactor. In the course of raising the acid concentration above 30 % the tail gas itself will absorb enough water vapor from the acid to render the humidification stage of the basic process unnecessary, with a consequent saving in steam. Furthermore, the tail-gas heat is utilized in concentrating the acid. Any additional heat required can be supplied by indirect heat exchange between the acid-plant absorber and the concentrator [122].

Peracidox Process. The Peracidox process is an oxidative sulfur dioxide-removal process developed by Lurgi and Süd-Chemie [121, 123] for purifying the already very dilute tail gas from a double-absorption sulfuric acid plant. It does not lead to any byproducts or waste, producing only sulfuric acid, which is recycled to the main sulfuric acid plant.

Hydrogen peroxide or electrolytically produced monoperoxysulfuric acid (H_2SO_5) is used to oxidize sulfur dioxide to sulfuric acid [55, 122]:

$$SO_2 + H_2O_2 \longrightarrow H_2SO_4 \tag{14}$$

$$SO_2 + H_2SO_5 + H_2O \longrightarrow 2\,H_2SO_4 \tag{15}$$

Figure 35 is a flow diagram of the electrolytic version of the Peracidox process. Tail gas leaving the final absorber of the sulfuric acid plant is washed in a two-stage Peracidox scrubber with ca. 30 – 40 % sulfuric acid in which the oxidant is dissolved. The attainable sulfur dioxide-removal efficiency is > 90 %. Entrained sulfuric acid vapor and residual sulfur trioxide are washed out simultaneously.

The oxidant is generated in an electrolytic unit by oxidizing H_2SO_4 anodically in a dilute sulfuric acid stream produced from concentrated acid from the product line. Diperoxysulfuric acid ($H_2S_2O_8$) initially formed in the electrolytic cell disproportionates in the dilute acid to form sulfuric acid and monoperoxysulfuric acid (Caro's acid) as the final oxidizing agent. When hydrogen peroxide is used as the oxidant, the electrolysis unit and pump tank are not, of course, required – only a storage tank for the hydrogen peroxide [55].

Oxidizing solution is introduced into the acid circuit of the second scrubbing stage at a stoichiometric rate proportional to the amount of sulfur dioxide to be removed. Any residual oxidant in solution overflowing to the first scrubbing stage reacts there with the entering tail gases. Overflow withdrawn from the circuit of the first stage thus consists only of dilute sulfuric acid, and it is introduced into the final or intermediate absorber of the main acid plant in place of process water.

Depending on the location of the plant and prevailing prices for electric power or hydrogen peroxide, the operating costs for a Peracidox process with electrolytically produced oxidant may be lower than costs incurred with the use of hydrogen peroxide [55]. If tail-gas purification is required only intermittently during short periods when sulfur dioxide emissions temporarily exceed permitted limits – during start-up, for example – it is more economical to use hydrogen peroxide on account of the lower investment costs [122].

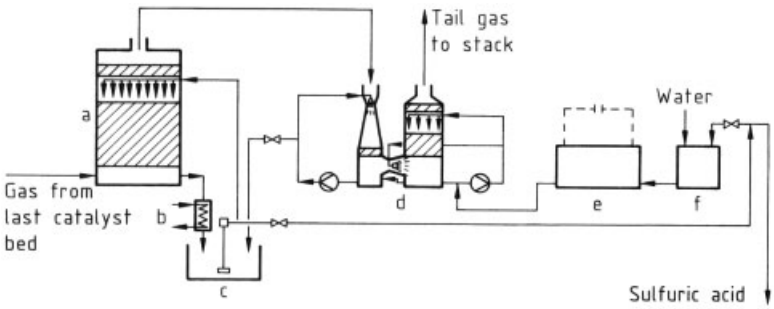

Figure 35. Lurgi Peracidox tail-gas purification system (electrolytic version)
a) H_2SO_4 plant final absorber; b) Cooler; c) Pump tank; d) Peracidox scrubber; e) Electrolytic oxidation unit; f) Electrolyte tank

A Peracidox tail-gas scrubbing unit with electrolysis has been in operation since 1972 at a 120-t/d double-absorption sulfuric acid plant in Germany [123]. Another scrubbing unit, using hydrogen peroxide and attached to a 330-t/d double-absorption sulfuric acid plant (also in Germany) has been operating since 1982 [124].

4.1.6. Economic Factors

Economies of Scale. Due to a steep increase in demand for sulfuric acid since the early 1950s, the unit size of sulfuric acid plants has been constantly increasing. For example, in 1954 the largest sulfuric acid plants in the Federal Republic of Germany were designed for production of ca. 250 t/d, and in the United States for about 450 t/d. Today, the largest double-absorption plants have capacities of ca. 1000 t/d in the Federal Republic of Germany and ca. 3000 t/d in the United States.

The incentive for steadily increasing the size of sulfuric acid production units is "economy of scale" – savings achieved in both specific capital and production costs. This is particularly important in view of the opposing effect of increasingly stringent environmental protection regulations. The measures required for conforming with these restrictions inevitably increase the complexity of the plant, introducing energy and capital cost penalties.

Economics of Voluntary Production. The relationship between specific capital cost and unit production capacity can be seen from Figure 36 for a sulfur-burning double-absorption sulfuric acid plant. The production of sulfuric acid from elemental sulfur represents the most economical and flexible option with respect not only to capital costs but also operating expenditure (utilities, manpower, absence of byproducts apart from steam). As a general rule, the sulfuric acid production costs in a large plant are determined chiefly by the raw materials, which means the price of sulfur, since the credit for steam production usually covers all other costs (such as electric power, cooling water, labor, depreciation, interest, administration, and maintenance).

In contrast, a sulfuric acid plant based on the roasting of pyrite is not only about twice as expensive to build (if the cost of the roasting and gas-cleaning systems are included) but also uses about twice as much electric power and about twice as much manpower as a sulfur-burning plant. Nonetheless, in cases involving domestic pyrite deposits that are economical to exploit for their nonferrous metals content, pyrite often represents a cheaper raw material source for sulfuric acid manufacture than expensive imported sulfur.

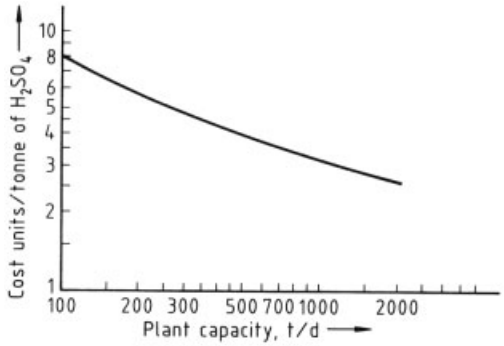

Figure 36. Relationship between specific capital cost and production capacity for a sulfur-burning double-absorption sulfuric acid plant

Without assigning a credit value for the cinder, the cost of producing sulfuric acid from pyrite is roughly comparable to that of producing it from elemental sulfur if the sulfur content of the pyrite costs ca. 25 % less than elemental sulfur. Depending on its chemical composition and prevailing market conditions, the byproduct cinder may exert either a positive or a negative influence on total costs. If the cinder can be utilized in a nearby cement fatory, for example, the credit may cover up to 15 % of the pyrite costs. In the least favorable case, the cost of removing and disposing of the cinder might constitute an additional burden on sulfuric acid production costs.

Economics of Involuntary ("Fatal Acid") Production. In any economic feasibility study, sulfuric acid obtained in the processing of metallurgical off-gases as a compulsory byproduct of a nonferrous-metal smelter operation (e.g., copper, nickel, zinc, lead, or molybdenum) must be valued differently from sulfuric acid of the same quality produced as a primary product in a plant based on sulfur combustion or pyrite roasting [125].

Sulfuric acid plants based on metallurgical off-gases are essentially pollution control plants, and they must operate in accordance with the production requirements of the accompanying smelter complex. Under such conditions it is not appropriate to expect the sulfuric acid plant to operate profitably, merely that proceeds from the sale of sulfuric acid should, if possible, cover the off-gas desulfurization costs. The capital costs of such a plant (including the wet gas-cleaning system as well as the sulfuric acid plant proper) depend heavily on the sulfur dioxide concentration of the gases to be processed (Fig. 37 A) [40].

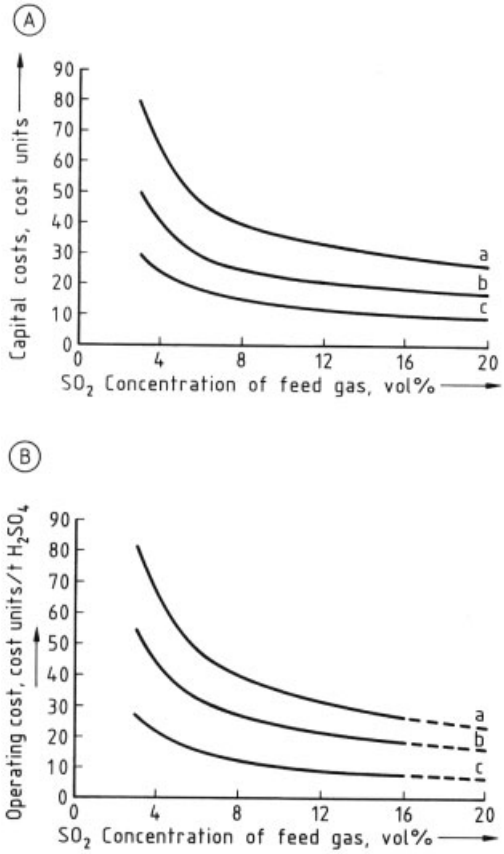

Figure 37. Capital cost (A) and operating cost (B) for a double-absorption sulfuric acid plant and wet gas - cleaning system as a function of feed gas sulfur dioxide concentration a) Total; b) Sulfuric acid plant; c) Wet gas - cleaning system

If the sulfur dioxide concentration and the gas volume of the off-gas stream are subject to frequent wide fluctuations, the capital cost for the gas-cleaning system and the sulfuric acid plant may approach that for a complete pyrite plant with a comparable acid capacity.

The current price disadvantage in the marketing of large quantities of sulfuric acid as a compulsory product (which frequently reflects a substantial freight cost owing to unfavorable location of the production facility, in relation to the market) is compensated by the fact that the sulfur dioxide content of the smelter off-gases is charged at zero cost in the cost calculation. However, in contrast to sulfur-burning and pyrite-based plants, there is no steam credit in a metallurgical sulfuric acid plant unless high-grade sulfur dioxide-containing smelter off-gases can be anticipated in the future [57].

The gas-processing costs and specific operating costs show a dependence on the feed-gas sulfur dioxide concentration similar to the capital costs, since capital-dependent costs account for most of the processing costs (Fig. 37 B).

4.2. Production by Nitrogen Oxide Processes

In a nitrogen oxide process, sulfur dioxide is oxidized to sulfuric acid in an aqueous phase, with nitrogen oxides serving as oxygen carriers. Processes of this type include the lead chamber processes (e.g., Mills – Packard and Gaillard) and the tower processes (e.g., Curtius, Petersen, Kachkaroff, Matignon, Salsasserra, and Moritz). These processes formerly constituted the basis for sulfuric acid production [19–23], but they have now been replaced almost completely by the modern contact processes (see Chap. 3). Nevertheless, nitrogen oxide processes are of interest not only from the point of view of the history of chemical theory, technology, and industrial practice, but also because they have certain characteristics that may qualify them for use even today (in an improved design adapted to specific cases), despite certain general disadvantages relative to contact processes.

Chemistry. Nitrogen oxide processes are based on oxidation of sulfur dioxide in the liquid phase. The detailed reaction mechanism is not fully understood even today, though this is one of the oldest industrially applied processes. In a simplified form, the course of reaction can be described by the following equations:

$$SO_2 + 2NOHSO_4 + 2H_2O \longrightarrow 3H_2SO_4 + 2NO \quad (16)$$
$$NO + 1/2\,O_2 \longrightarrow NO_2 \quad (17)$$
$$NO_2 + NO \rightleftharpoons N_2O_3 \quad (18)$$
$$N_2O_3 + 2H_2SO_4 \rightleftharpoons 2NOHSO_4 + H_2O \quad (19)$$
$$SO_2 + 1/2\,O_2 + H_2O \longrightarrow H_2SO_4 \quad (20)$$

Sulfur dioxide absorbed in the aqueous phase is actually oxidized by nitrosyl hydrogensulfate, $NOHSO_4$ ("nitrosylsulfuric acid"), dissolved in ca. 70 % sulfuric acid ("nitroso acid"), yielding sulfuric acid and releasing gaseous nitric oxide, NO. The colorless nitric oxide, which is virtually insoluble in aqueous systems, is oxidized by free oxygen in the gas stream to form the yellow nitrogen dioxide, NO_2. Nitric oxide and nitrogen dioxide are in equilibrium with the relatively unstable nitrogen sesquioxide, N_2O_3, which to some extent dissolves in sulfuric acid.

In sulfuric acid of greater than 70 % strength, nitrogen sesquioxide reacts to form nitrosyl hydrogensulfate, which is a relatively stable compound at higher sulfuric acid concentrations. In ca. 70 % sulfuric acid, however, nitrosyl hydrogensulfate is metastable and behaves as a very reactive oxidant toward sulfur dioxide. For this reason it is not possible to produce sulfuric acid of significantly higher concentration without accepting a higher residual nitrogen oxide content.

Application as a Pollution Control Process. The low concentration of the product acid, which (for reasons explained above) cannot economically exceed ca. 78 %, constitutes the overriding disadvantage of any nitrogen oxides process. Sulfuric acid of this concentration has relatively few applications today. It is not economical to concentrate it because of the high energy costs involved. Viewed strictly as a method of manufacturing sulfuric acid, therefore, the nitrogen oxides process is no longer competitive with contact processes. It has the further disadvantage that none of the heat released is at a sufficiently high temperature to permit its economical recovery.

On the other hand, the low working temperatures of the nitrogen oxide process offer certain advantages in specific cases. Thus, it is suitable for processing gases with extremely low sulfur dioxide contents, which cannot be converted in a contact plant without supplementary heating. The nitrogen oxide process has also been proposed by Tyco as a method for removing both sulfur and nitrogen oxides from flue gases [126].

Ciba – Geigy Process. Since 1974, Ciba – Geigy has been developing a nitrogen oxide sulfuric acid process specifically designed for the processing of gases containing ca. 0.5 – 3 vol % SO_2 and offering special advantages when the gases are laden with nitrogen oxides [127].

A flow diagram of the Ciba – Geigy nitrogen oxide process is shown in Figure 38. The

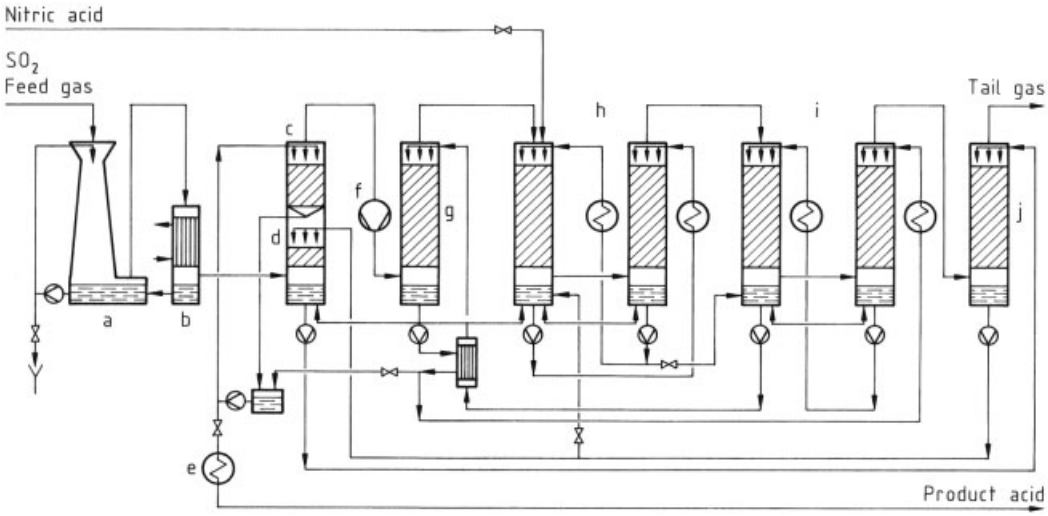

Figure 38. Ciba – Geigy nitrogen oxide sulfuric acid process
a) Venturi scrubber; b) Star cooler; c) Product acid denitrator; d) Gas dryer; e) Cooler; f) Blower; g) Denitration tower; h) SO_2 oxidation reactors; i) Nitrogen oxide absorbers; j) Concentrator

sulfur dioxide-containing gases are first passed through a wet gas-cleaning and cooling system to remove dust and reduce the water-vapor content. In a subsequent gas dryer, surplus moisture is absorbed in 50 – 60 % sulfuric acid and transferred via the outer acid loop to a concentrator, where it is evaporated into the tail gases. In principle, this is quite comparable to the Lurgi predryer – reconcentrator system used to maintain the water balance in a contact-process acid plant (see page 15).

The product *denitrator* serves as a final nitrogen oxides stripping stage for the product-acid bleed stream. Nitrogen oxides are removed from the irrigation acid partly by reaction with sulfur dioxide (reaction 16) and partly by hydrolysis (the reverse of reaction 19). Sulfuric acid leaving the tower is free of nitrogen oxides, and has a concentration of 70 – 80 %, preferably 78 % (60° Bé). Part of the sulfur dioxide feed already has been removed when the gases pass on to the denitrator.

In the denitrator up to ca. 50 % of the sulfur dioxide feed is oxidized by nitrosyl hydrogensulfate present in recycled acid from the first nitrogen oxide absorber. Nitric oxide is released into the gas stream. The incoming acid is heated in a recuperative acid heat exchanger, while acid leaving the denitrator is cooled before being recycled to the second nitrogen oxide absorber.

The two sulfur dioxide reactors serve to reoxidize nitric oxide and remove remaining sulfur dioxide. The transformation of sulfur dioxide into sulfuric acid takes place at the extensive gas – liquid interface presented in the irrigated reactor towers by packings of a special design. Reaction heat from the oxidation of sulfur dioxide is removed by cooling the circulating acid.

In the two nitrogen oxide absorbers, nitrosyl hydrogensulfate is formed by the reaction of nitrogen oxides present in the gases with sulfuric acid recycled from the denitrator via the inner acid loop. The gas and the irrigation acid flow cocurrently in the first absorption tower and in countercurrent fashion in the second. Acid discharged from the second tower is pumped to the preceding nitrogen oxide absorption tower after cooling.

Efficient absorption of the nitrogen oxides depends not only on adequate cooling of the irrigation acids but also on precise control of the NO/NO_2 ratio, which is monitored continuously at the outlet of the nitrogen oxide absorption stage by automatic analyzers. To ensure maximum nitrogen oxide absorption, the plant is designed and operated such that gases entering the absorption towers contain nitric oxide and nitrogen dioxide in almost equimolar quantities, which favors the formation of nitrogen sesquioxide. Nitric oxide in excess is scarcely absorbed at all. On the other hand, nitrogen dioxide in excess does dissolve to some extent, but it is partly converted back into insoluble nitric oxide. Some autoxidation of nitric oxide to nitrogen dioxide takes place in the absorption towers, though this is a slow reaction.

Sulfur dioxide remaining in the gas at the nitrogen oxide absorption stage will again liberate nitric oxide. A high sulfur dioxide content in the entering gases therefore causes a disproportionate rise in nitric oxide losses to the tail gases. Consequently, a high sulfur dioxide conversion efficiency in the sulfur dioxide reaction step is prerequisite to minimizing nitrogen oxide concentrations in the tail gases. The final sulfur dioxide content is actually lower than 100 ppm by volume.

Up to ca. 1 g/m^3 of nitric oxide and nitrogen dioxide may remain in the tail gas, but this is not the only cause of nitrogen oxide loss in the overall plant. Side reactions also take place in which nitrogen oxides are converted to such species as nitrous oxide, N_2O, and elemental nitrogen, while a certain amount of nitrosyl hydrogensulfate is discharged with the product acid.

Nitrogen losses must be compensated by the addition of nitric acid to the sulfur dioxide reactors in order to maintain the nitrogen oxide balance in the entire system. Another possibility is the addition of nitrogen oxides generated by burning ammonia. If the sulfur dioxide-containing feed gases already contain nitrogen oxides, the stoichiometric equivalent of nitric acid or ammonia can be saved.

The first commercial plant based on the Ciba – Geigy nitrogen oxides process has been in operation near Goslar (Federal Republic of Germany) since 1979. It is designed to desulfurize 10 000 m^3/h of off-gases containing ca. 0.8 – 1.5 % SO_2 from a molybdenum sulfide roasting unit. A characteristic of this plant is the almost exclusive use of plastics as construction materials. In comparison with earlier tower plants, the specific reaction-space requirements (in relation

to gas throughput) have been substantially reduced. An elaborate automatic measuring and control system restricts nitrogen oxide emissions to a minimum.

4.3. Regeneration of Spent Sulfuric Acid

4.3.1. Introduction

Most of the sulfuric acid produced enters not into marketable end products, but instead ends up as spent acid, impure gypsum, or other metal sulfates, all of which are waste products. Disposal of these chemical wastes represents an unsolved environmental problem that is becoming more serious as production capacities expand and antipollution regulations become more stringent.

The bulk of the spent acid is generated by the organic chemical and petrochemical industries. The nature of the acid varies widely from highly dilute to relatively concentrated, and with both low and high levels of organic and inorganic impurities. Large volumes of uniform spent acid arise only in the course of plastics production (e.g., methacrylate, caprolactam); typical data related to the quantity and composition of representative spent acids are as follows:

Composition figures for alkylation acid:	
Sulfuric acid content	ca. 88 – 90 %
Water content	ca. 5 %
Organics	ca. 3 – 4 %
Organic sulfates	ca. 2 %

Production and consumption figures for caprolactam manufacture:	
Caprolactam production rate	120 000 t/a
H_2SO_4 consumption (100 % acid)	636 kg/t of caprolactam
Oleum consumption	1 300 kg/t of caprolactam
Ammonium sulfate production	312 000 t/a (including spent acid)
Spent acid composition:	
Sulfuric acid	ca. 50 %
Salt content	ca. 0.7 %
Water	balance

Production and consumption figures for methyl methacrylate (MMA) manufacture:	
Methyl methacrylate production	50 000 t/a
H_2SO_4 consumption (100 % acid)	1.8 t per tonne of MMA
Spent acid production	2.7 t per tonne of MMA
Spent acid composition:	
Sulfuric acid	ca. 20 %
Ammonium hydrogensulfate	ca. 45 %
Water	ca. 30 %
Organics	ca. 5 %
Solids	ca. 2 %

An important source of spent acids containing inorganic impurities is the inorganic chemicals industry, especially the TiO_2 pigment industry, which generates spent acid of a relatively uniform quality containing ca. 20 – 23 % H_2SO_4 and 7 – 15 % metal sulfates. Similar in quality is the spent acid obtained from pickling operations in the metallurgical industry. Another source of waste acid is the gas-cleaning section of a roasting or other "metallurgical" acid plant.

A typical analysis of spent acid obtained from TiO_2-pigment manufacture is as follows:

Sulfuric acid	22 %
Iron sulfate	4.5 %
Calcium sulfate	0.1 %
Other metal sulfates	5.4 %
Water	68 %

In the past, typical methods of disposal or recycling for spent sulfuric acids included

1) Neutralization with lime
2) "Ocean dumping"
3) Reconcentration/blending with oleum

Stricter pollution-control regulations together with more limited disposal areas have now virtually eliminated the use of such neutralization and dumping methods in the industrialized countries. Reconcentration of weak acids by blending with oleum also has its practical limits. Indeed, the method is suitable only for small acid volumes involving high H_2SO_4 concentrations, and even then only if the specific impurity content is such as to permit recycling or use of the acid in some other process.

Regeneration of spent sulfuric acids is therefore the only method available for obtaining a product of original quality and at the same time complying with environmental regulations, although the cost of acid derived from such a process greatly exceeds the production cost in a "normal" sulfuric acid production plant.

The regeneration of spent acid normally comprises two major process steps:

1) Concentration of the spent acid to the highest feasible level, together with separation of salts to the extent necessary
2) Decomposition of the spent acid and/or sulfates

No single procedure can be considered universally suitable for concentrating all types of

spent sulfuric acid. The choice of an evaporation method also depends on the extent to which the specific application for which the regenerated sulfuric acid is ultimately destined is able to tolerate residual impurities.

The process of concentrating of dilute sulfuric acid consumes substantial amounts of heat energy, not only to evaporate water but also for dehydrating the acid (see Section 2.1). For concentration at atmospheric pressure at or near the boiling point it must also be borne in mind that the boiling point of sulfuric acid rises steeply as the H_2SO_4 concentration increases.

The combination of high acid temperature, intermediate H_2SO_4 concentrations, and certain types of impurities can give rise to extremely corrosive conditions, necessitating the use of special construction materials for the concentration equipment.

Spent sulfuric acid can be concentrated by direct or indirect heating, and in one or several stages. The methods can be classified into two large groups [128]:

1) Evaporation to an H_2SO_4 concentration $\leq$ ca. 75 wt %
2) Evaporation to high concentration (up to the azeotropic point, 98.3 % H_2SO_4 at atmospheric pressure)

In concentrating to less than 75 % H_2SO_4, water is almost the only substance evaporated. The H_2SO_4 content of the gas phase is negligible on account of the low H_2SO_4 vapor pressure (see Section 2.1), so a scrubbing system for removing sulfuric acid is seldom necessary. Evaporating to a higher concentration (up to the azeotropic point) requires a relatively high temperature. However, the specific evaporation-heat requirement in the higher concentration range is relatively low because little water remains. For example, if spent acid containing 20 % H_2SO_4 is concentrated to 98 % in two stages – from 20 % to 75 % and then from 75 % to 98 % – ca. 92 % of the total amount of water removed evaporates in the first step, leaving only 8 % to be evaporated in the second.

A more detailed description of the various approaches to reconcentration is presented in Section 4.3.2.

The second and more fundamental regeneration step is decomposition of the spent acid and/or sulfates. Such sulfates may be present in the acid itself (e.g., as ammonium hydrogensulfate) or they might be isolated in the form of free metal sulfates (e.g., $FeSO_4 \cdot H_2O$ from TiO_2-pigment spent acid). The selection of an appropriate decomposition procedure is crucial for satisfactory operation, where the composition of the acid and the solids/salts content are the determining factors. Three process concepts that have proven their worth in industrial-scale applications are:

1) Gas-phase decomposition in an empty decomposition furnace
2) Decomposition above a red-hot coke bed in a rotary kiln (Grillo process)
3) Decomposition in a fluidized bed

For more detailed descriptions of the various decomposition processes, see → Sulfur Dioxide, Chap. 5.6.2.

A decomposition plant of this type is normally found in conjunction with a gas-cleaning plant and a sulfuric acid contact plant, such that the final product is pure sulfuric acid.

The spent-acid regeneration plant realized in Sachtleben Chemie is outlined in Fig. 39 [182]. It represents an integrated regeneration process for spent acid from a TiO_2-pigment fabrication process.

Spent acid as received from the TiO_2 production facility is first concentrated from ca. 22 % H_2SO_4 to ca. 70 % in a vacuum concentration plant. The 70 % $H_2SO_4/FeSO_4$ suspension is then passed through the "maturation station," representing the 2nd stage of the process. After maturation the suspension passes through a filter plant separating the acid from the sulfates (stage 3). The sulfates recovered from the filter are prepared for decomposition, mixed with pyrite and carbon, and passed to an intermediate storage facility (stage 4). The filtrate (sulfuric acid with 70 % H_2SO_4) is further concentrated to ca. 80 % H_2SO_4 in another vacuum concentration plant (stage 5). The acid that results is ready for immediate use in the TiO_2 attack section. In stage 6, the mixture of sulfate/carbon/pyrite from stage 4 is fed into a fluidized-bed roaster for decomposition. The SO_2 gas is cooled, cleaned, and transformed into sulfuric acid in a contact acid plant.

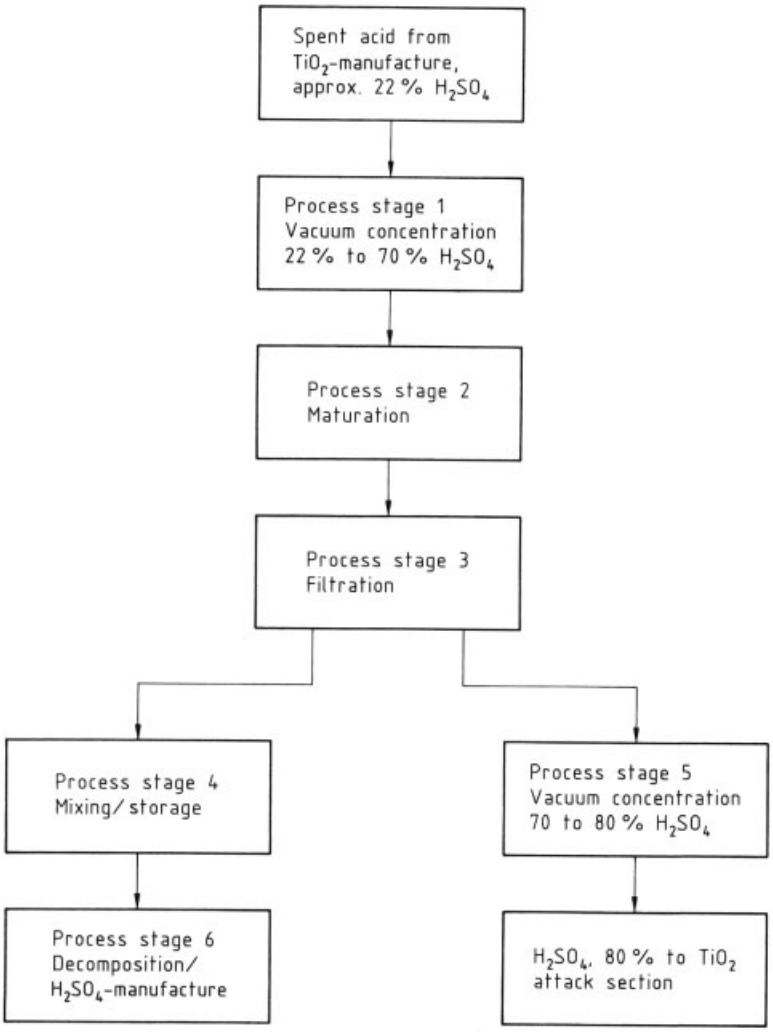

Figure 39. Integrated spent-acid regeneration process

4.3.2. Reconcentration to 70 – 75 % H_2SO_4

Vacuum Evaporation. In this process, dilute sulfuric acid is maintained at the boiling temperature under vacuum by indirect heating with steam. The vapors evolved are recondensed almost completely in a downstream cooling system so that, apart from insignificant amounts of inert gases desorbed from the acid, there is no gaseous effluent. Other dissolved volatile impurities that may be present, especially in organic-contaminated acids, can generally be condensed as well. Heavily contaminated acids containing inorganic substances, which tend to deposit solids during evaporation, are effectively concentrated in forced-circulation evaporators, whereas for comparatively pure spent acids a natural circulation system may be adequate [128]. Evaporators of this type are satisfactory for concentrating dilute sulfuric acid to 70 – 80 % [129].

Forced-Circulation Evaporator. In the forced-circulation evaporator shown in Figure 40, the heater is separate from the flash vessel. With this arrangement it is possible to prevent evaporation in the heater itself, so solids deposition on the heating surfaces is kept to a minimum. An additional advantage of forced circulation over natural circulation is that the liquid-flow velocity in the heat exchanger is

higher and more easily controlled; as a result, the heat-transfer conditions are more stable.

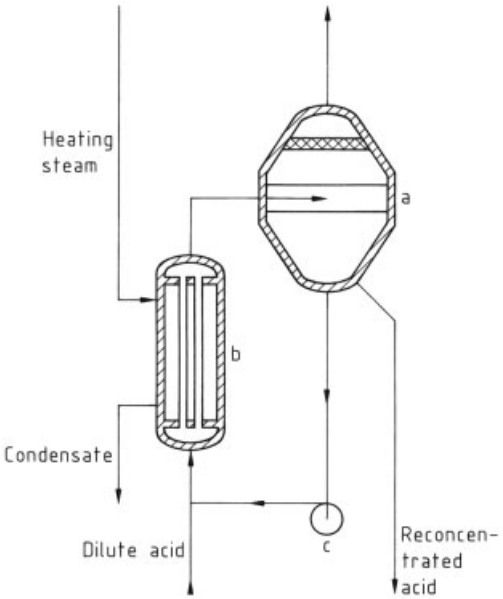

Figure 40. Forced-circulation evaporator with separate heater
a) Flash vessel; b) Heater; c) Circulating pump

A constriction at the inlet to the flash vessel keeps the acid under a slight pressure so that, although it is heated to 3–4 °C above the outlet temperature of the flash vessel, water cannot evaporate until the acid is free to expand into the evaporator. Delayed boiling in the evaporator itself can be largely prevented by efficient distribution of the acid within the evaporator space. Owing to the relatively high rate of circulation, precipitated solids remain in suspension and are discharged from the system along with the reconcentrated acid.

Multiple-Effect Evaporators. The energy required for water evaporation is normally supplied in the form of low-pressure steam. The steam requirement can be significantly reduced, especially in large-capacity evaporation systems, by coupling several evaporator units together in such a way as to reuse the heat content of the vapors. In the first evaporator unit, heat is supplied in the normal way by indirect steam heating of the acid circuit, but in each subsequent evaporator unit the acid circuit is heated indirectly by vapors produced in the preceding stage. Evaporation systems arranged in this manner are known as *multiple-effect* (or *multistage*) *evaporators* (→ 3. Evaporation, Chap. 2.2.).

The relationship between specific steam consumption and the number of thermally coupled stages in a multiple-effect evaporator is illustrated graphically in Figure 41. A single evaporator unit (i.e., a single-effect system) requires ca. 1.2 t of low-pressure steam per tonne of evaporated water. In a double-effect system this requirement is reduced to only about 0.6 t. Adding a third effect leads to a further reduction, to ca. 0.4 t of low-pressure steam per tonne of water evaporated – in other words, about one-third of the steam consumption of a single-effect system. The alternative of dividing the evaporation system into several concentration steps without such coupling (i.e., use of live-steam heating in each unit) does not reduce the specific steam requirement, though it does reduce the required total heating-surface area relative to a single-effect system.

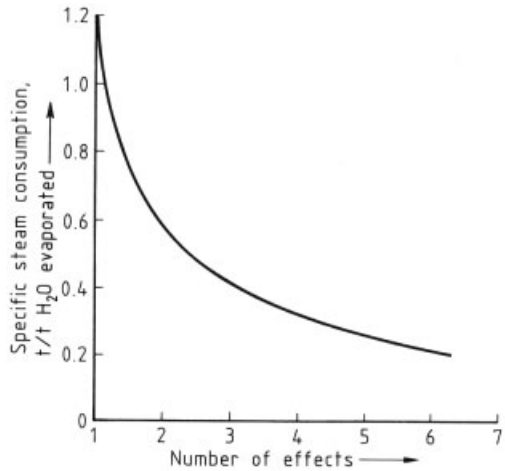

Figure 41. Theoretical specific steam consumption as a function of the number of effects

Vacuum and Condensation Systems. Multiple-effect evaporator systems are normally operated under vacuum. Depending on the relative cost and availability of electric power, steam, and cooling water, the vacuum may be produced by means of electric vacuum pumps or steam ejectors. When steam ejectors are used, several units are arranged in tandem to generate a relatively low pressure. The effect is enhanced by the installation of preceding and intermediate water-cooled vapor condensers.

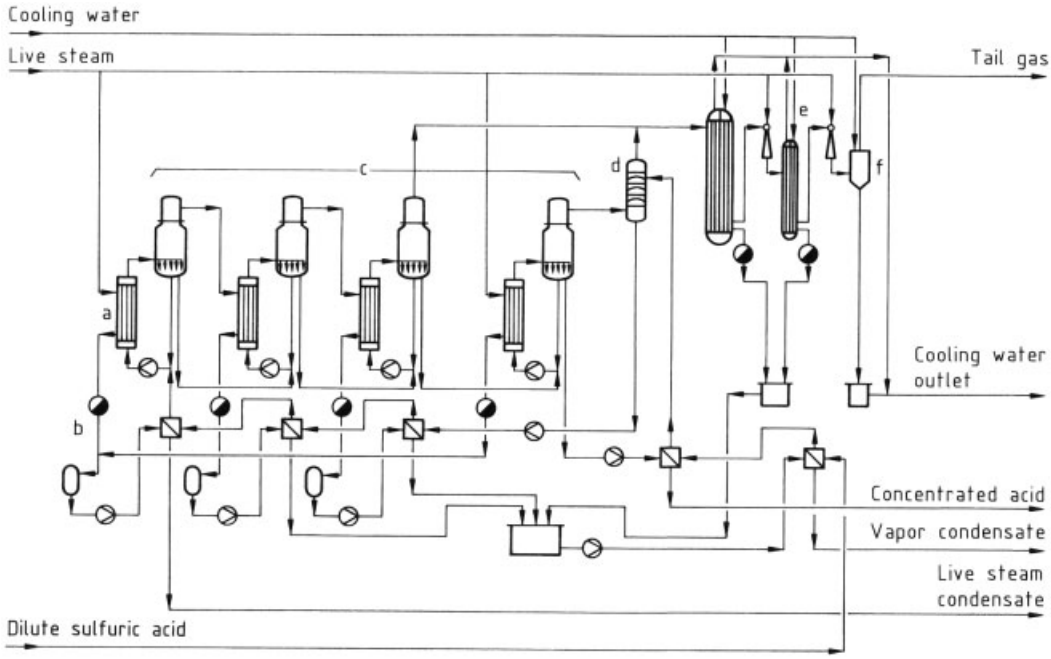

Figure 42. Forced-circulation vacuum evaporation system for concentrating sulfuric acid from 20 % to 70 – 80 % H_2SO_4
a) Evaporator heaters; b) Plate heat exchangers; c) Evaporators; d) Acid degasser; e) Ejectors; f) Contact condenser

If spent sulfuric acid containing volatile contaminants is to be concentrated it is best to use surface condensers (tube-bundle heat exchangers with cooling water flowing through the tubes), because with this type of system it is easiest to separate impurities from the vapor condensate. Where the contact (quench) type of condenser with direct water injection is used, the volume of wastewater can be reduced by recirculating part of the condensate.

Construction Materials. Evaporators designed to concentrate sulfuric acid up to 75 % can be fabricated from fiberglass-reinforced polyester or plastic-coated, rubber-lined, enamelled, or lead-coated steel. At higher acid concentrations and temperatures it is possible to use steel lined with FEP sheeting (→ Films, Chap. 6.3.3.) and protected with two layers of carbon bricks. The heat exchangers are usually made of graphite impregnated with phenol resin, in which case they can be operated at wall temperatures $\leq 165\,°C$ and acid concentrations $\leq 84\,\%$ H_2SO_4. Cast ferrosilicon pumps in common use today are capable of withstanding temperatures $\leq 140\,°C$ and a sulfuric acid concentration of ca. 75 %. Special construction material is required for high-stress conditions.

Complete Plant. Figure 42 is a flow diagram of a typical Lurgi-designed plant for vacuum concentration of spent sulfuric acid containing 20 % H_2SO_4 to an H_2SO_4 concentration of 78 % [84]. It comprises a total of four evaporator units. The acid is preconcentrated from 20 % to 54 % H_2SO_4 in the first three units, which are arranged as a triple-effect system. Final concentration to 78 % takes place in a fourth, single-effect stage with live-steam heating. A pressure gradient is maintained across the first three units (a triple-effect stage) to ensure that water continues to evaporate despite a decreasing temperature. Thus, the first stage operates virtually at atmospheric pressure, whereas the third runs at a much lower pressure – 70 mbar, for example, at which point the vapor-condensation temperature is ca. 40 °C. Live steam at 5.5 bar is used for heating the first and fourth stages. This might be derived, for instance, from 3-bar and 28-bar steam, using a steam ejector followed by a steam saturator.

Dilute feed acid is preheated in plate exchangers with the vapor condensate and concentrated production acid. It is then degassed in a bubble-tray column by stripping with vapor from the final concentration stage; this also

serves to recover any sulfuric acid that may still be present in vapors from the final concentration stage. The degassed dilute acid is preheated further in plate exchangers with vapor and steam condensate from the evaporator heaters before being introduced into the acid circuit of the first evaporator.

In the first stage, heated by live steam, water is evaporated from the acid at about 120 °C. The steam supply is adjusted to maintain an acid concentration in the first stage of ca. 25 % H_2SO_4, where the acid boiling point is used as the control parameter. Indeed, the boiling-point elevation above the boiling point of pure water at a predetermined pressure can be used as an exact measure of the H_2SO_4 concentration provided the influence of dissolved contaminants is taken into account. In the second stage, where the acid is heated indirectly by vapors from the first stage, more water is evaporated at about 95 °C despite the higher acid concentration; this is possible because of the lower pressure. Vapors from the second stage provide heat for acid circulating in the third stage, where it is concentrated to about 55 % at ca. 65 °C under an appropriate vacuum. As mentioned previously, live steam is again used for final concentration to 78 % in the fourth stage.

Because of the pressure gradient, the acid need not be pumped from one stage to the next. After the fourth stage, the concentrated acid is cooled indirectly with incoming dilute acid in a plate heat exchanger and than pumped to storage tanks. Vapors from the third-stage evaporator are mixed with scrubbed vapors from the final evaporator as they leave the degasifier column, and then passed through a surface condenser. Uncondensable gases emerging from this condenser are exhausted through a multistage vacuum system consisting of steam ejectors and auxiliary surface condensers, and are finally discharged to the atmosphere.

The steam condensate from the heating elements of the first and fourth evaporator stages is collected, expanded, and cooled through indirect heat exchange with the dilute acid before being discharged. Vapor condensate from the second and third stages is similarly utilized for preheating the spent acid in plate exchangers, after which it is mixed in a collector with condensate from the condensers of the vacuum system. The residual heat of the accumulated vapor condensate is utilized for initial preheating of the spent acid.

The process thus takes maximum advantage of the heat content of the vapors by means of a multiple-effect arrangement of the evaporators, and it also recovers the maximum possible amount of waste heat from the condensate for preheating purposes. This is of vital importance in view of the continuously rising cost of energy.

The maximum water-evaporating capacity of a single-train multistage unit with forced-circulation evaporators is presently in the range of 8 – 10 t of water per hour.

Venturi Reconcentration Process. The venturi sulfuric acid reconcentration system developed by Lurgi is preferably integrated thermally into a sulfuric acid production plant in such a way as to make use of waste heat and thereby avoid the cost of heating with steam or fossil fuels. The system is designed to utilize sensible heat from the tail gases as well as the relatively low-level heat content of acid circulating in the sulfur trioxide absorption system of an acid production plant [122, 130].

The most significant feature of the venturi reconcentration process is that – at atmospheric pressure – water is evaporated from dilute sulfuric acid at a temperature far below the boiling point of the acid solution. Gases passing through the venturi system become laden with water vapor in proportion to the water-vapor partial pressure of the acid (which is a function of the H_2SO_4 concentration and the temperature), carrying this water out of the reconcentrator system for subsequent discharge to the atmosphere. Venturi systems have long been used in gas-cleaning plants as gas scrubbers, but it was only in the early 1970s that Lurgi developed venturi systems specifically designed for acid concentration. These have meanwhile rendered satisfactory service in various industrial plants.

The concentration unit consists of a venturi scrubber into which the acid is injected cocurrent with the gas flow, followed by a tower in which dilute spent acid and gas interact in countercurrent. Despite a short gas retention time, water becomes equilibrated between the acid and the vapor because the two mix so efficiently and intimately in the venturi unit. The carrier gases acquire a higher moisture load in such a two-stage evaporation system with two separate acid

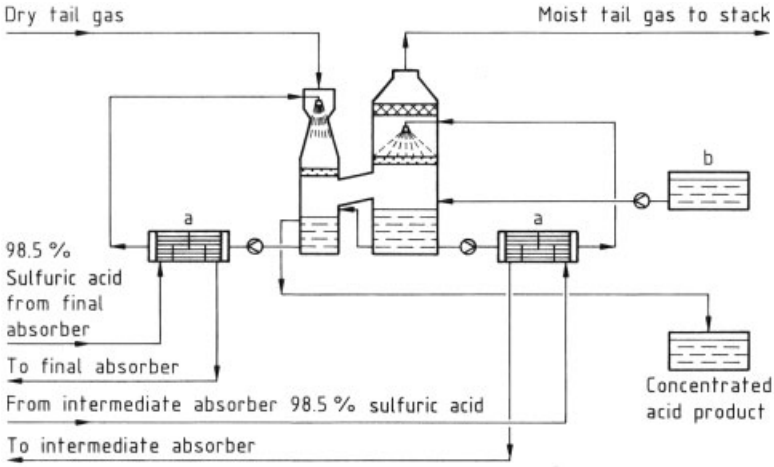

Figure 43. Two-stage venturi reconcentration system utilizing acid waste heat from a double-absorption sulfuric acid plant
a) Acid heat exchangers; b) Dilute acid tank

circuits running at different H_2SO_4 concentrations than they would in any single-stage system, so the water-evaporation capacity is greater for the same gas throughput. At the same time, the H_2SO_4 content of discharge gases remains extremely low, even when the sulfuric acid concentration is raised to more than 70 %, on account of the lower concentration of acid circulating in the second evaporation stage. It has been demonstrated in practice that venturi reconcentrators also function well as tail-gas scrubbers, substantially reducing the natural H_2SO_4 content of tail gases from a sulfuric acid plant [75]. The heat required for dehydration and water evaporation during concentration of the acid is supplied directly to the system as the sensible heat content of the carrier gases as well as by indirect heating of the concentrator acid circuits with product acid from the main acid plant. The concentrating capacity can be increased by introducing air heated with excess heat from the contact plant along with the hot, dry tail gas from the sulfuric acid plant.

The temperature demand associated with heat suitable for a venturi reconcentrator system, which has a comparatively low operating temperature, is quite compatible with the substantial amount of waste heat available in acid circulating in the sulfur trioxide absorption system of a sulfuric acid plant, which must in any case be dissipated. The venturi reconcentrator system is thus a very good way of making economical use of low-level waste heat from a sulfuric acid plant. If the concentrator unit is thoroughly integrated into the waste-heat system of the acid plant, ca. 22 – 26 t/h of water can be evaporated with the waste heat from a 1000-t/d sulfuric acid plant.

Figure 43 is a diagram of a two-stage venturi reconcentrator designed for evaporating waste sulfuric acid originally containing 30 % H_2SO_4 until it reaches a concentration of 70 %. Tail gases leaving the final sulfur trioxide absorber of the sulfuric acid plant at ca. 80 °C are passed directly into the venturi reconcentrator. Circulated acid, at a concentration of about 70 % and also heated to ca. 80 °C, is injected at the top of the venturi unit in cocurrent with the gases. Adiabatic evaporation of water cools the gases and acid to about 74 °C. A special nozzle grate installed in the outlet of the venturi section provides additional turbulence and assists in subsequent separation of gas from the liquid phase.

Gases leaving the venturi section typically have a water content of ca. 20 g/m^3. They are introduced into the bottom of a subsequent tower through which they rise in countercurrent with a spray of 55 % sulfuric acid from the second acid circuit, again heated to about 80 °C. Within the tower the water content of the gases increases to about 80 g/m^3. A droplet separator at the tower outlet removes any remaining droplets from the gases, which are subsequently discharged to the atmosphere via a stack. Acid accumulating in the tower sump becomes cooled to ca. 70 °C as a result of the water evaporation. Acid in both concentrator circuits is heated in Teflon tube-

bundle heat exchangers by 98.5 % sulfuric acid at ca. 100 °C derived from the intermediate and final absorbers of a double-absorption sulfuric acid plant.

The dilute acid subject to concentration, which is about 30 % in H_2SO_4, is fed to the tower sump of the second stage. Preconcentrated acid from the second stage simply overflows from the sump of the tower into the sump of the first-stage venturi unit. The 70 % reconcentrated acid likewise overflows from the sump of the venturi unit to product storage.

Operation of a venturi reconcentrator system is quite straightforward. The feed rate of the spent acid is controlled relative to the acid concentration in the venturi circuit so as to maintain the desired product-acid concentration; no further control measures are required.

Venturi reconcentrators are designed for a gas-pressure drop of 15 – 20 mbar. Apart from power consumed by the supplementary blower capacity and the electric drives of the acid circulation pumps, no additional energy is required.

Since they are relatively compact, venturi reconcentrators are well-suited to retrofit installation in existing sulfuric acid plants.

The first of the venturi acid-reconcentration systems has been in operation at Süd-Chemie AG at Kelheim, in the Federal Republic of Germany, since 1972. It concentrates dilute spent sulfuric acid to a level of about 70 %, and is associated with a double-absorption sulfuric acid plant designed for a capacity of 220 t/d. The largest venturi reconcentrator unit, designed for a gas throughput of 200 000 m^3/h, is thermally integrated into a 2000-t/d double-absorption sulfuric acid plant at the Inspiration Copper works in Arizona (United States). Venturi reconcentrators thermally integrated into sulfuric acid plants are also used for processing spent acids saturated with ferrous sulfate, encountered, for example, as scrubbing acids in the gas-cleaning systems of pyrite roasting plants.

Submerged-Combustion Process. Submerged-combustion processes are sometimes used for concentrating spent acids with high salt contents to 60 – 70 % H_2SO_4 [128, 131–134]. Liquid or gaseous fuels and a burner are used to generate hot flue gases at 1500 – 1600 °C that are forced through concentrated acid via an immersion tube. The flue gases are cooled by the adiabatic evaporation of water, simultaneously absorbing water vapor at a rate that depends on the water-vapor partial pressure above the acid. Depending on the acid concentration, the bath temperature remains in the range of ca. 120 – 150 °C. The flue gases, which leave at about the same temperature, contain up to ca. 60 vol % water vapor. Before being discharged to the atmosphere, these gases are cleaned and cooled in a liquid separator, a scrubbing and cooling tower, a wet precipitator and, if necessary, a special sulfur dioxide scrubbing tower.

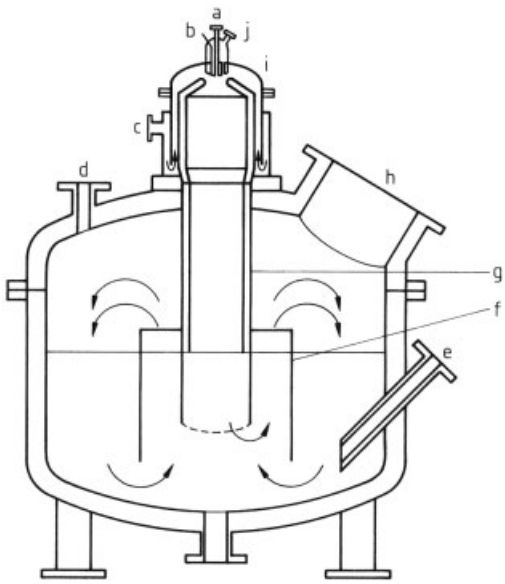

Figure 44. Submerged-combustion evaporator for gas or oil firing
a) Oil inlet; b) Igniter; c) Air inlet; d) Acid inlet; e) Acid outlet; f) Acid circulating tube; g) Immersion tube; h) Vapor outlet; i) Combustion chamber; j) Gas inlet

Figure 44 is a schematic diagram of a submerged-combustion evaporator. The evaporator vessel is constructed from lead-coated steel, and in most cases it is brick-lined as well. The combustion chamber is located above the vessel, from which the burner projects downwards beneath the surface of the acid. A circulating tube is positioned concentrically around the immersion tube, through which liquid is forced by the ascending flue gases. Apart from causing vigorous circulation of the acid, this also promotes favorable mass and heat transfer between gas and liquid.

Spent acid is fed continuously either through a separate nozzle, as shown in Figure 44, or through a dephlegmator placed on top of the gas outlet [131, 132]. A mixture of concentrated acid and precipitated salts leaves the evaporator through an overflow pipe.

This system of direct heat transfer from the flue gases to the acid largely avoids the incrustations that would normally be deposited on heating surfaces, especially during the evaporation of salt-laden spent acid. Nevertheless, the high flue-gas temperature and the consequent high wall temperature of the immersion tube impose very heavy demands on the construction materials.

The jacketed combustion chamber is cooled by incoming combustion air passing through the jacket. The immersion tube, made of cast iron, is protected in its uncooled upper region by an internal insulating lining. The lower part of the immersion tube is cooled by the acid. The critical zone with respect to corrosion and erosion is the outside of the immersion tube above the top of the circulating tube. With time, this part of the immersion tube gradually becomes eaten away to the point where the lower, submerged part actually falls off, limiting the service life of the tube to a maximum of ca. 2000 operating hours [133].

A submerged-combustion evaporator is useful, for example, for concentrating the contaminated salt-bearing sulfuric acid arising from TiO_2 pigment manufacture from 21 % to a final concentration of 65 – 68 % H_2SO_4 [133]. About 27 m^3/h of spent acid with a specific gravity of 1.3 kg/L is fed to an evaporator vessel with a capacity of ca. 3 m^3. At the designed water-evaporation rate of ca. 1.95 t/h, 3.1 – 3.2 GJ of energy in the form of fossil fuel is consumed per ton of evaporated water. However, the specific heat requirement of the submerged-combustion evaporation process can be reduced (in this case, for example, to about 1.9 GJ per tonne of water evaporated) [133] by utilizing the heat of condensation of the vapors for concentrating spent acid in a preceding stage.

4.3.3. Concentration to 93 – 98 % H_2SO_4

Vapors evolved during the concentration of spent sulfuric acid to a more highly concentrated state (93 – 98 % H_2SO_4) contain significant quantities of gaseous sulfuric acid. It is therefore absolutely obligatory that the vapors be cleaned before discharge to the atmosphere. Concentration may involve either direct or indirect heating, with the latter approach applicable both at normal pressure and under vacuum.

Chemico Direct-Fired Drum Concentrator. The Chemico drum-concentrator process [135] was developed in the 1920s. It is used especially for concentrating spent acids from the explosives industry. The principle is similar to that of the submerged-combustion method. Flue gases resulting from the combustion of gas or fuel oil are passed through an immersion tube into the acid to be concentrated, which flows countercurrent with the gases through several vessels arranged in cascade fashion (Fig. 45). The vessels are made of lead-lined steel protected with up to three layers of acid-proof bricks [136, 137]. The immersion tubes are fabricated from ferrosilicon.

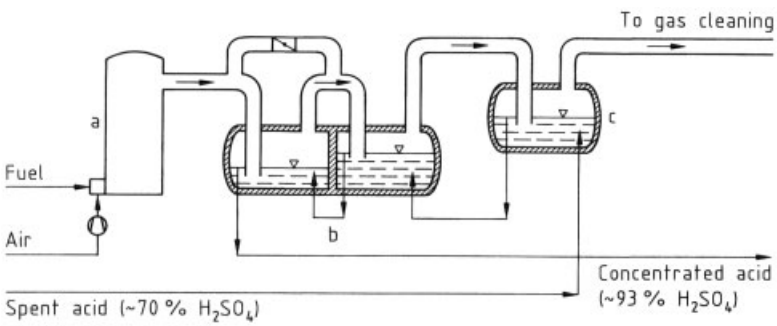

Figure 45. Chemico drum concentrator
a) Combustion chamber; b) Concentrator; c) Gas cooler/acid preheater

Flue gases enter the concentrator at ca. 600 °C [128]. Because of the relatively low flue-gas temperature, there is practically no thermal decomposition of the sulfuric acid, so almost no sulfur dioxide is formed [131].

The concentrator is subdivided into two chambers containing acid at different concentrations. In the first chamber, into which the hot flue gas enters and from which concentrated acid is discharged, the acid concentration is maintained at 93 – 94 %. The gas then passes to the second chamber at 250 – 260 °C. A portion of the hot flue-gas stream may be introduced directly into the second chamber for the purpose of controlling the acid concentration. Flue gases leaving the second chamber at 180 – 200 °C are cooled to 130 – 150 °C in a second vessel. The feed acid (usually at a concentration of 70 %) is introduced into this vessel. Discharged gases must be demisted and cleaned.

The Chemico process shares with the submerged-combustion method the advantage that it avoids heating surfaces prone to acquiring deposits during the concentration of salt-contaminated spent sulfuric acid; all the requisite heat is supplied by the flue gases. Moreover, water evaporation takes place not in the boiling range of sulfuric acid but about 50 °C below the atmospheric pressure boiling point [136]. The process is suited to the concentration of relatively large volumes of acid. Single-train plants have been designed with capacities of up to 18 t/h of 93 – 94 % acid.

A significant drawback of the Chemico process is the large amount of contaminated waste gas that is produced, the cleaning of which is very costly. A scrubbing tower and an electrical gas-cleaning system were formerly used for this purpose, but modern plants employ venturi scrubbers almost exclusively. These are combined with cyclones for separating droplets. The sulfuric acid content of the waste gas from such a plant is reported to be 75 – 200 mg/m^3 [135].

Spent acid may also contain organic impurities, in which case these and their decomposition products would be entrained by the waste gases and subject to only partial elimination in the waste-gas cleaning system. When processing spent acid containing nitrates or nitrogen oxides, NO/NO_2 liberated during concentration would similarly not be removed by the gas-cleaning system, resulting in a visible plume at the stack. In view of the increasing stringency of environmental regulations in many locations, the utilization of this process is becoming more and more problematic [137].

Concentration Processes Involving Indirect Heating.

Pauling – Plinke Process. In this process, spent sulfuric acid is concentrated to 95 – 98 % H_2SO_4 under normal pressure by indirect heating with flue gases. The process was developed by Pauling beginning in 1915, and the necessary equipment is built by Plinke [138].

Actual concentration takes place in a cast iron vessel (a "Pauling pot") suspended in a furnace. The furnace is surmounted by a distillation column (in most cases a bubble-tray column) made of silicon iron (see Fig. 46). Vessels of this type are built with a useful volume of up to 12 m^3; above about 3 m^3 an anchor mixer (also made of silicon iron) is included [138].

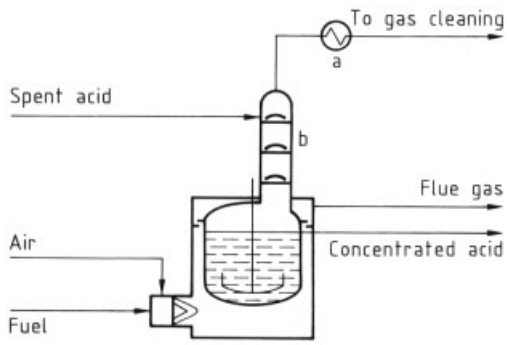

Figure 46. Pauling – Plinke pot concentrator
a) Condenser; b) Stripping column

Spent acid (ca. 70 % H_2SO_4) is fed to the column from above, and it flows downward into the vessel in countercurrent with the exhaust vapors. Incoming acid is concentrated to 82 – 85 % H_2SO_4 by heat and mass transfer with the vapors; simultaneously, sulfuric acid present in the vapors is condensed almost completely. If the spent acid has an H_2SO_4 content greater than 70 % the acid is admitted at a tray lower in the column.

The sulfuric acid concentration in the vessel itself must be kept above 80 % to avoid corrosion [136]. In most cases, the concentration of the boiling acid is maintained above 95 % (corresponding to a boiling temperature of ca. 330 °C). Under these conditions, the average corrosion

loss-rate is about 8 – 10 mm per year [138]. A service life of ca. 3 – 5 years is achieved by designing the vessel with a sufficiently thick wall.

The acid is heated indirectly by flue gases passing around the vessel. These flue gases, generated by burning fuel oil or gas, enter the furnace cavity at 800 – 1100 °C and leave it at ca. 450 °C. The mean heating-surface load of the vessel is ca. 144 000 $kJm^{-2} h^{-1} = 40 kW/m^2$ [138]. The high flue-gas temperature means that the wall material of the vessel on the furnace side is subject to heavy stress.

Flue gases leaving the furnace can be utilized for preheating the combustion air to ca. 200 °C. Part of the flue-gas stream is cooled to 200 – 250 °C and recycled to the furnace cavity. Such combustion-air preheating and flue-gas recirculation raises the furnace efficiency (ratio of heat supplied to the vessel to applied fuel energy) to 80 – 83 %.

Hot, concentrated sulfuric acid is extracted continuously from the vessel. It is cooled indirectly against water in special coolers (pot coolers) [138], where most of the dissolved inorganic impurities are precipitated. Organic impurities in the feed acid are decomposed by oxidation within the evaporator: either by the high temperature alone or – if necessary – by the addition of nitric acid.

Vapors leaving the top of the discharge column at ca. 130 °C still contain as much as ca. 1 % sulfuric acid vapor, together in some cases with sulfur dioxide, nitric oxide, and organic substances, as well as inert gases [128, 138]. The water vapor is separated in a condenser, whereas noncondensable components are passed on to a waste-gas cleaning system.

All concentration processes involving indirect heating, including the Pauling – Plinke process, have the advantage over directly heated processes that the volume of resulting waste gases to be cleaned is low. However, they suffer the drawback of heating surfaces that are prone to scaling with precipitated impurities from the spent acid. This necessitates frequent checking and entails high maintenance costs [136].

In the Pauling – Plinke process, the ratio of heating surface to vessel volume, and thus the ratio of the attainable heat-transfer rate to the rate of acid production, is relatively low. Because of the limited evaporation capacity of a single unit, high throughput requires several parallel units [136]. Typically, 0.6 – 1.2 t/h of water can be evaporated per vessel [137].

Falling-Film Evaporation Process. In this process, spent sulfuric acid descends as a thin film on the inner surface of an externally heated tube. This arrangement provides a favorable heat-transfer rate and a large surface from which water can evaporate. Evaporated vapors ascend in countercurrent with the acid, and are substantially freed from entrained sulfuric acid vapor and mist by the scrubbing action of entering spent acid in a column surmounting the evaporator, similar to that in the Pauling – Plinke process.

A falling-film evaporator process was formerly operated by Du Pont under vacuum conditions (27 – 40 mbar), with indirect heating by steam at 25 – 30 bar [139].

A more recent process – the BOSAC process, developed by Bofors – operates under normal pressure with indirect heating by flue gases [137, 140–142]. When the emerging flue gases are used for preheating the combustion air, a thermal efficiency of 82 – 85 % is achieved [141].

Bayer – Bertrams Process. This process is applicable from atmospheric pressure up to about 5 bar; indirect heating is supplied by a heat-transfer medium [137, 143, 144]. The heat carrier is a salt melt (nitrate/nitrite) that is heated indirectly by flue gases in a separate furnace.

The modern version of the process uses tubes of borosilicate or pure quartz glass with a diameter of 150 – 400 mm. Spent acid is fed to the stripping column at a concentration of 65 – 85 % H_2SO_4 and is concentrated to ca. 96 – 98 % at 300 – 330 °C. Organic impurities that may be present are decomposed almost completely, with assistance as necessary from an added oxidant.

A special advantage of the falling-film concentration process is the favorable ratio of heating surface to acid volume. By connecting a large number of tubes in parallel, an evaporation capacity comparable to that of a Pauling – Plinke unit can be achieved. A falling-film evaporator plant composed of 12 tubes is reported to have a capacity of 1 t/h of 96 – 97 % sulfuric acid starting from 65 – 70 % acid, which corresponds to a water-evaporation rate per tube of ca. 50 kg/h.

Vacuum Processes. Processes in this category transform spent sulfuric acid directly into highly concentrated material. The principle is

comparable to that used in a vacuum process for evaporation to intermediate concentration. Problems are encountered with the materials of construction, however, and it is essential that special cleaning systems be provided to remove evaporated sulfuric acid from the emitted vapors. It is not possible to make use of the multiple-effect principle because of the high temperatures required. The acid is heated indirectly with high-pressure steam at every stage.

Simonson – Mantius Process. Developed in 1921, this is one of the most well-known vacuum methods for attaining high acid concentrations. Most of the corresponding plants are multistage units operating with either natural or forced circulation, depending on their design, and with separate heat exchangers or heating tubes installed directly in the evaporator vessels [145]. The evaporator vessels themselves are made of lead-lined steel further protected by acid-proof bricks or a layer of graphite or enamel. The heat exchangers or heating tubes are made of lead, graphite, silicon iron, or tantalum.

The Simonson – Mantius process has been developed further by Chemetics International Ltd. [137, 146]. In the Chemetics version of the process, the evaporators are made of enamelled steel or, in the case of very small units, glass-lined steel, and they incorporate bayonet heaters made of tantalum. In its standard form the process leads to 93 % sulfuric acid, but it can be designed to provide concentrations of 95 – 96 % if desired [137].

The largest existing unit concentrates approximately 28 t/d of sulfuric acid (calculated as 100 % H_2SO_4) from 70 % H_2SO_4 to 93 %; this corresponds to a water evaporation capacity of ca. 10 t/h [137]. Only acids that do not contain solids, or which will not produce accretions during concentration, can be treated by this process.

Corning/EIVS Process. The Corning/EIVS process is based on horizontal evaporators [147]. Spent acid flows continuously through a horizontal cylindrical vessel made of borosilicate glass. PTFE deflectors are used to prevent backmixing. The acid is heated indirectly with 15-bar steam, which is passed through an internal tantalum tube bundle. The operating temperature of the evaporators is limited to a maximum of 180 °C to restrict corrosion.

Spent acid can be concentrated from 65 % H_2SO_4 to 88 % in a single stage operating at an absolute pressure of 93 mbar. For triple-stage concentration to 95 % H_2SO_4, pressures in the individual stages are 133, 40, and 13 mbar [147]. The steam consumption is reported to be ca. 2 t of steam per tonne of evaporated water, corresponding to ca. 4 GJ per tonne of H_2O.

4.4. Production of Oleum

Oleum ("fuming sulfuric acid") [*8014-95-7*] consists of a solution of sulfur trioxide in 100 % sulfuric acid. Normally, the oleum concentration is expressed in terms of the mass content of so-called "free" sulfur trioxide. Part of the dissolved sulfur trioxide reacts with sulfuric acid to form higher sulfuric acids such as disulfuric acid, $H_2S_2O_7$ (see Section 2.2). Pure sulfur trioxide (Chap. 8) corresponds theoretically to oleum with 100 wt % "free" sulfur trioxide.

Oleum is produced industrially in contact plants, where sulfur trioxide-containing gases are passed through a special oleum tower. In a double-absorption plant this is located upstream from the intermediate absorber; otherwise it is upstream from the final absorber.

Oleum should contain at least 20 – 22 % "free" sulfur trioxide to avoid problems with construction materials. The upper concentration limit of about 36 % is determined by the partial pressure of sulfur trioxide in the gases as well as the oleum temperature in the oleum tower, which in turn establishes the sulfur trioxide vapor pressure of the oleum (see Section 2.1 and Fig. 6). The dependence of oleum vapor pressure on temperature and concentration in the range of interest to industry is shown in Figure 47 [148–150]. Oleum formation by sulfur trioxide absorption can proceed only so long as the oleum vapor pressure remains below the partial pressure of sulfur trioxide in the gas phase. It follows from the vapor-pressure curves that, if the oleum temperature is 60 °C and the partial pressure of sulfur trioxide in the contact gases is 115 mbar, for example, the theoretical maximum attainable oleum concentration is about 37 %. This corresponds to the sulfur trioxide partial pressure in gas at a total pressure of 1.12 bar produced in the first three beds of a (3 + 1) double-absorption sulfuric acid plant operating on a feed gas initially containing 10 vol % SO_2, assuming the sulfur dioxide conversion at this point is 95 %.

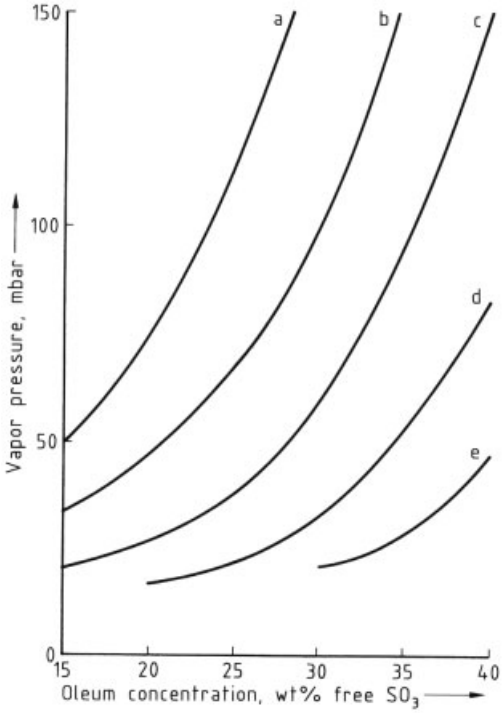

Figure 47. Vapor pressure of oleum as a function of free SO_3 content and temperature
a) 80°C; b) 70°C; c) 60°C; d) 50°C; e) 40°C

The potential rate of oleum production is limited by the water balance in the contact plant. This is determined by the rate at which water vapor enters the plant – with the combustion air in a sulfur-burning plant, or in roaster or decomposition gases in a "cold-gas" plant – and is subsequently absorbed into the acid circuit in the dryer, since sufficient sulfur trioxide must remain in the process gas after the oleum tower to maintain an acid concentration of 98.5 % in the absorbers. Therefore, the lower the rate at which water enters the acid system via the dryer, the higher the proportion of sulfur trioxide available for oleum production.

Oleum towers are similar in design to modern acid-plant main absorbers. However, because of the specific corrosion behavior of oleum, these towers can be fabricated from stainless steel without a brick lining. Sulfur trioxide-ontaining gas enters the oleum-irrigated packed tower at ca. 200 °C. Depending on the desired product concentration, the tower is irrigated with 22 % or 35 % oleum at a temperature of ca. 40 – 50 °C. Oleum of higher concentration leaves the tower at ca. 60 – 80 °C. It is adjusted to the desired final oleum concentration by addition of concentrated sulfuric acid from the intermediate or final absorber of the contact plant, and is then cooled in special coolers. Product is withdrawn from the oleum circuit downstream from the cooler. Oleum of higher concentrations (e.g., 60 – 65 %) is obtained by mixing 20 – 36 % oleum with pure sulfur trioxide.

5. Construction Materials

(→ Construction Materials in Chemical Industry)

The traditional construction material for use in a sulfuric acid plant based on the lead chamber or tower process was, of course, lead. By contrast, modern contact-process plants are constructed mainly from iron and steel. This is a striking example of the extent to which the selection of construction materials depends on the nature of a specific process and its associated operating conditions [151–153]. Conversely, new processes frequently become practical only with the development of appropriate materials. This is particularly true of acid-heat recovery systems, which are of major importance in the further development of sulfuric acid technology.

Elevated sulfuric acid temperatures are essential for the economical recovery of acid waste heat. This mandates the use in acid coolers and heat exchangers of corrosion-protected materials that are able to withstand attack not only by the acid itself (with the aid of anodic protection as appropriate) but also, given the elevated wall temperature, attack by the cooling fluid.

One general problem relates to the extent to which a particular material will display in service the same behavior that has been observed in a laboratory with respect to corrosion rates under specific conditions (as a function of oxygen concentration and temperature, for example). Many additional parameters that influence corrosion behavior are relevant in industrial practice. For example, dissolved impurities – even in minute concentrations, which may be difficult to determine analytically – as well as suspended solids and sulfuric acid flow velocities are all significant, as is the treatment and processing of a construction material prior to its installation.

Data acquired through systematic laboratory tests based on pure media, and even empirical data derived from practical experience, are therefore of only limited validity for predicting the behavior of a given material in a particular application. General statements regarding the suitability of a construction material with respect to sulfuric acid over a defined range of concentration and temperature should never be taken too literally.

A comprehensive list of important empirical values is provided in the Dechema materials tables [154].

For concentrations $\leq 90\,\%$, sulfuric acid is classified as a nonoxidizing acid, but concentrated sulfuric acid and oleum both exhibit oxidizing properties. This is the reason, for example, why hot, concentrated sulfuric acid reacts with copper and silver to form sulfur dioxide rather than hydrogen:

$$2\,H_2SO_4 + M^0 \longrightarrow M^{II}SO_4 + SO_2 + 2\,H_2O \qquad (21)$$

Carbon is also oxidized by hot, concentrated sulfuric acid and oleum, again generating sulfur dioxide. Corrosion rates for metallic materials in sulfuric acid are determined largely by the solubilities of the respective metal sulfates as a function of acid concentration and temperature.

5.1. Metallic Materials

Lead. The high corrosion resistance of lead over a wide range of sulfuric acid concentrations and temperatures is due exclusively to the formation of a passive layer of lead sulfate ($PbSO_4$) on the metal surface. The solubility of lead sulfate in sulfuric acid increases with increasing acid concentration, and the corrosion rate increases accordingly. Lead can be used with H_2SO_4 concentrations up to 78 % and temperatures $\leq 110\,°C$. Chloride ions strongly enhance the corrosion, because they penetrate the passive layer. Copper-containing lead is the most corrosion-resistant form apart from high-purity lead itself. Lead plays essentially no role whatsoever in a dry-gas contact sulfuric acid plant. Only in wet-catalysis plants ($\leq 78\,\%$ H_2SO_4) and especially in wet gas-cleaning systems is lead still relatively widely used in such equipment as scrubbers, wet-gas precipitators, tube-bundle coolers, piping, valves, and pumps. Even here it is increasingly being replaced by plastics.

The main drawbacks of lead are high density, low mechanical strength, and high installation cost.

Mild Steel. Sulfuric acid with a concentration $> 60\,\%$ produces an iron sulfate layer on mild steel surfaces. The solubility of iron sulfate decreases markedly at elevated acid concentrations, and the corrosion rate displays a similar steep decline as the H_2SO_4 concentration is increased above 78 %. At H_2SO_4 concentrations $> 100\,\%$ (i.e., in the presence of free sulfur trioxide) corrosion increases rapidly up to a sulfur trioxide content of 30 %, declining steeply at even higher sulfur trioxide concentrations. The passive layer is quickly eroded at elevated acid-flow velocities, and the corrosion rate rises accordingly. Chloride ions contribute substantially to corrosion, because they penetrate the passive layer (pitting corrosion).

Steel is used in conjunction with 68 – 100 % sulfuric acid at low temperatures for both storage and handling vessels, and it is the preferred construction material for all reaction vessels in contact plants producing 93 – 98.5 % sulfuric acid. The corrosion resistance of steel declines as the temperature rises from 50 °C to 80 °C, which is why steel surfaces in reaction vessels with turbulent flow areas are protected by a lining of acid-proof bricks. Dryers and absorbers frequently have a layer of plastic sheeting between the steel shell and the brick lining. Steel is unsuitable for acid temperatures above 80 °C. When acid piping is made of steel itis advisable to limit the flow velocity to about 0.5 m/s. Steel is also used for oleum, but predominantly in situations where the "free" sulfur trioxide content exceeds 22 – 23 %. If the "free" sulfur trioxide content is low, enamelled steel is suitable, even at elevated temperature.

Austenitic Stainless Steel [183]. Under oxidizing conditions, sulfuric acid forms a passive layer on austenitic stainless steel. This layer consists of iron oxide and chromium oxide together with incorporated sulfates, which improves its stability. At higher flow velocities, however, and under reducing conditions, the protective layer is eroded, causing the corrosion rate to increase. Reducing conditions result, for example, if sul-

furic acid is left standing in a shut-down plant for extended periods without being moved or aerated. Chloride ions also contribute to depassivation. In a modern sulfuric acid plant, austenitic stainless steel (e.g., material no. 1.4571) is used in conjunction with concentrated acids up to 85 °C, and it can withstand acid velocities up to ca. 1.0 m/s (in pipelines, for example). At higher temperatures, the corrosion resistance decreases sharply. However, the temperature resistance can be increased to about 160 °C by applying anodic protection (passivation of a metallic construction material by anodic polarization with direct current) [68]. Anodic protection not only reduces the corrosion rate for certain grades of stainless steel by two orders of magnitude, but it also prevents stress-corrosion cracking. It is increasingly being used to safeguard stainless-steel acid coolers, especially in acid waste-heat recovery systems.

Because of the limitations of standard stainless steels (types 304, 316, etc.) at higher temperatures and high H_2SO_4 concentrations, considerable effort has been directed toward the development of special forms of stainless steel. Several stainless steels with austenitic and ferritic structures are now available for sulfuric acid service at temperatures up to 140 °C (in certain cases up to 200 – 220 °C) and flow velocities up to 2.5 – 3 m/s. Anodic protection is not required for such stainless steels. The most familiar alloys for this purpose include:

Sandvik SX
Saramet
Hastelloy C 276 (DIN 2.4819)
Cronifer 2803 (DIN 1.4575)
Alloy 20 (Carpenter 20)
DIN 1.4529, 1.4539
DIN 1.4562, 1.4563

The above-mentioned materials are used mainly for the fabrication of heat exchangers, piping, irrigation systems, and to a certain extent drying/absorption towers. A similar development has occurred with cast materials. Pumps, valves, etc. are now available for sulfuric acid absorption systems operating at temperatures up to 200 °C. The best-known alloys of this type are the Lewmet family and Rheinhütte's alloys 1.4136 S/HRS. Cast iron was for decades the dominant construction material for sulfuric acid services but itis now being replaced in almost all fields of applications by these modern alloys.

Acid-Proof Cast Iron. Certain grades of cast iron alloyed with chromium or copper are very stable at H_2SO_4 concentrations of 90 – 99 % and temperatures $\leq$ 100 – 120 °C. Cast iron is used mainly for acid piping, irrigation systems in acid towers, and irrigation coolers. It is not suitable for use with oleum; although the surface-wear corrosion rate is extremely low, salts tend to form alongside the graphite lamellae, giving rise to cracks and material failure.

Silicon Iron. Silicon iron contains 14 – 18 % silicon. Its high resistance to acids is due to the formation of a protective layer of silica on the surface of the casting. It is therefore somewhat sensitive to hydrogen fluoride. Silicon iron is highly resistant to sulfuric acid at all concentrations, from extremely dilute up to highly concentrated, and over a wide temperature range. Silicon iron with at least 16 % Si is useful for pumps and agitators under the high thermal stress encountered in a sulfuric acid concentration plant. Unfortunately, however, it is highly sensitive to both thermal and mechanical shock. It also cannot be used for oleum, because its silicide content is subject to oxidation, and the corresponding salt formation gives rise to cracks and ruptures along the grain boundaries.

5.2. Nonmetallic Materials

Plastics. Plastic materials [for example, polypropylene, poly(vinyl chloride), or polyethylene] are preferred for use in systems with low acid concentrations because of their high resistance to corrosion. Their usefulness is limited, however, by high coefficients of thermal expansion and a loss of mechanical strength with increasing temperature, resulting in creep. Poly(vinyl chloride) (PVC) can be used with acid concentrations $\leq$ 78 % H_2SO_4 at temperatures $<$ 50 °C, polypropylene (PP) at concentrations $\leq$ 90 % and temperatures $<$ 80 °C. On account of its excellent corrosion resistance, however, polytetrafluoroethylene (PTFE) is applicable even at high acid concentrations ($\leq$ 100 %) and temperatures $>$ 100 °C for equipment and

pipe linings as well as in special acid heat exchangers.

Graphite. Nonporous carbon and graphite are regarded as resistant construction materials under nonoxidizing conditions (i.e., in sulfuric acid with a concentration up to 90 %), and they are used in condensation towers, coolers, storage tanks, and pumps at temperatures $\leq 170\,^{\circ}C$.

Glass. Because of its excellent corrosion resistance at all acid concentrations, glass is used in special acid heat exchangers – for heating dilute waste acid with hot absorber acid, for example. A significant disadvantage of this material is its fragility.

Enamel. Acid-proof enamel offers the same resistance as glass. Newly developed enamel grades are largely insensitive to thermal shock. Enamelled steel and cast iron are used for acid piping at elevated temperature.

6. Uses of Sulfuric Acid and Economic Aspects

Sulfuric acid is one of the most widely used of all industrial chemicals. But, as has already been noted, most of its uses can be considered as indirect, because it functions as a reagent rather than as an ingredient: surprisingly little of it appears in end products, and most ends up as spent acid or some type of sulfate waste. A number of products do incorporate the sulfur of sulfuric acid, but nearly all of them are low-volume, specialty items.

6.1. Indirect Uses

The largest single consumer of sulfuric acid by far is the *fertilizer* industry. Most goes into the production of phosphoric acid, which in turn is used to manufacture such fertilizer materials as triple superphosphate and mono- and diammonium phosphates. Lesser amounts are used for producing superphosphate and ammonium sulfate. About 60 % of the sulfuric acid produced is utilized in fertilizer manufacture [35, 155]. In the Federal Republic of Germany, however, the fertilizer industry accounts for a much less significant proportion of total sulfuric acid consumption: only about 14 % [33]. This is because of the historical predominance in that country of thermally treated phosphates and basic slag (a byproduct of steel making) relative to chemically produced phosphate fertilizers. Although thermal phosphates are generally in decline in relation to total consumption owing to their high cost, the balance is accounted for by increased use of fertilizers based on imported phosphate intermediates rather than an expansion in the domestic basic phosphate processing industry.

Elsewhere in industry, substantial quantities of sulfuric acid are used as an acidic dehydrating reaction medium in *organic chemical* and *petrochemical* processes involving such reactions as nitration, condensation, and dehydration, as well as in *oil refining*, in which it is used for refining, alkylation, and purification of crude-oil distillates.

In the *inorganic chemical* industry sulfuric acid is used notably in the production of TiO_2 pigments, hydrochloric acid, and hydrofluoric acid.

In the *metal processing* industry, sulfuric acid is used for pickling and descaling steel, for leaching copper, uranium, and vanadium ores in hydrometallurgical ore processing, and in the preparation of electrolytic baths for nonferrous-metal purification and plating.

Certain *wood pulping* processes in the paper industry require sulfuric acid, as do some *textile* and *chemical fiber* processes and *leather tanning*.

6.2. Direct Uses

Under certain conditions, sulfuric acid is occasionally used directly in *agriculture* for rehabilitating extremely alkaline soils, such as are found in the desert regions of the western United States. This is not a very important use in volume terms, however.

Probably the largest use of sulfuric acid in which the sulfur becomes incorporated in the final product is *organic sulfonation*, particularly for the production of detergents [156]. Other minor organic chemicals and pharmaceuticals are also made by sulfonation.

One of the most familiar consumer products containing sulfuric acid – the *lead-acid battery* – accounts for only at tiny fraction of total sulfuric acid consumption.

6.3. Economic Aspects

Table 1 provides a summary of world production and consumption of sulfuric acid for the years 1982 and 1992, during which world production increased by 9.9 % (despite a drop of 10.4 % between 1989 and 1992).

Table 1. Sulfuric acid production and consumption for 1982 and 1992 (10^3 t of 100 % H_2SO_4) [source: British Sulphur]

	Production		Consumption	
	1982	1992	1982	1992
World total	**130 263**	**144 646**	**132 189**	**145 505**
Western Europe	**24 094**	**20 295**	**23 806**	**20 649**
Belgium/Luxembourg	1 798	1 917	2 092	2 697
France	3 927	3 132	4 000	3 456
Germany *	4 670	3 800	4 290	2 785
Italy	2 360	1 800	2 320	1 740
Netherlands	1 692	1 125	1 600	1 530
Spain	2 690	2 490	2 623	2 190
United Kingdom	2 587	1 716	2 559	1 946
Central Europe **	**8 216**	**3 390**	**8 510**	**3 310**
former Czechoslovakia	1 252	510	1 415	510
Poland	2 682	931	2 600	831
Romania	1 669	500	1 700	540
former Yugoslavia	1 120	900	1 220	900
Former USSR	**23 801**	**17 400**	**23 600**	**17 400**
Africa	**10 141**	**15 702**	**10 265**	**15 670**
Morocco	3 072	7 279	3 072	7 279
South Africa	3 382	2 589	3 380	2 589
Tunisia	2 525	3 347	2 525	3 347
North America	**32 490**	**44 030**	**33 458**	**44 740**
Canada	3 131	3 605	3 060	2 350
United States	29 359	40 425	30 398	42 390
Latin America	**6 308**	**8 602**	**6 960**	**8 824**
Brazil	2 229	3 077	2 331	3 241
Chile	460	1 194	460	1 325
Mexico	2 920	2 825	3 370	2 849
Asia	**20 872**	**33 815**	**20 980**	**33 446**
China	8 174	14 060	8 270	14 060
India	2 232	3 970	2 323	3 970
Indonesia		1 330		1 370
Japan	6 531	7 120	5 763	5 948
South Korea	1 596	2 493	1 780	2 663
Taiwan	680	600	700	949
Oceania	**2 491**	**1 412**	**2 490**	**1 466**
Australia	1 971	950	1 920	1 030

* 1982 data include both the Federal Republic of Germany and the German Democratic Republic. ** German Democratic Republic omitted from the 1982 totals.

7. Analytical Techniques

Concentration and impurities content are the most frequently measured parameters for sulfuric acid quality-control purposes. In addition, quantitative determinations of sulfur dioxide and trioxide are essential in the production plant for monitoring the performance of the process and ensuring that the tail-gas composition conforms with environmental regulations. More detailed accounts of analytical methods are available in [157–159].

7.1. Concentration Measurement

The simplest manual method for determining sulfuric acid concentration is density measurement with a densitometer. This method is reliable for sulfuric acid with a strength of up to 95 % provided the influence of temperature is taken into account. The method can also be extended to acids of higher concentration and also to oleum by first diluting a measured sample under controlled conditions and then measuring its density.

Continuous monitoring of acid concentration in an industrial sulfuric acid plant is generally accomplished by means of automatic recording devices based either on conductivity or ultrasonic measurement. The common chemical laboratory approach to measuring sulfuric acid content is volumetric titration with sodium hydroxide solution, but in certain cases it is necessary to resort to gravimetric determination of precipitated barium sulfate.

7.2. Measurement of Impurities

Iron, arsenic, sulfur dioxide, nitrogen compounds, chloride, and fluoride in sulfuric acid can be determined by conventional means as follows:

Iron. Iron content can be established either photometrically after addition of thiocyanate or 1,10-phenanthroline, or else by iodometry. If the sample is first reduced chemically, iron can also be determined by titration with permanganate.

Arsenic. The sample is first subjected to reducing conditions to convert arsenic to arsine. The latter is then driven off by heating and absorbed in pyridine containing silver diethyldithiocarbamate. The concentration of the resulting red complex can be determined by measuring its absorbance. Other techniques include the coulometric method of GUTZEIT and a gravimetric method in which all the arsenic present is reduced to arsenic trichloride, which can be distilled off and determined as sulfide.

Sulfur Dioxide. Dissolved sulfur dioxide is determined iodometrically (→ Sulfur Dioxide, Chap. 10.).

Nitrogen Oxides and Nitric Acid. Dissolved nitrogen oxides can be measured by titration with permanganate. Alternatively, the sample may be treated with 1-naphthylamine and sulfanilic acid to convert the nitrogen oxides to an azo dye, which is then determined photometrically. Nitric acid and nitrosylsulfuric acid form a colored species upon addition of brucine, and this can also be determined photometrically.

Chlorides and Fluorides. Chloride is determined gravimetrically by precipitation with silver nitrate. Fluoride can be measured by etching according to the Spielhaczek method.

A wide range of sulfuric acid impurities is now determined on the basis of physical methods; e.g.:

- Mercury, selenium, and arsenic by flameless atomic absorption spectrometry
- Iron, nickel, chromium, zinc, cadmium, lead, copper, vanadium, aluminum, sodium, and potassium by atomic emission spectrography (with inductively coupled plasma) or atomic absorption spectrometry

7.3. Analysis of Acid-Plant Gas Streams

The sulfur dioxide content of contact and tail gases is established by one of the methods described in → Sulfur Dioxide, Chap. 10..

When determining the sulfur trioxide content of sulfuric acid-plant tail gases it is important to bear in mind that residual sulfur dioxide is easily oxidized, which can lead to unrealistically high readings. The following methods give satisfactory results.

The gas sample in question is passed through an aqueous solution of hydrogen peroxide and the resulting sulfuric acid is determined by titration or gravimetric analysis (see above). The fraction of this acid derived from sulfur dioxide is determined by measuring the sulfur dioxide concentration of a second sample and calculating the sulfuric acid equivalent. This is subtracted from the result of the first determination to leave the amount of sulfuric acid due to sulfur trioxide alone.

Alternatively, the gas can be passed through an isopropyl alcohol solution, which absorbs sulfur trioxide without oxidizing sulfur dioxide. The resulting sulfuric acid then is titrated against barium chloride solution using thorin as an indicator [160, 161].

SO_3/H_2SO_4 mists entrained in tail gas are separated in special filters and determined by measurement of the trapped acid. Particle spectra of SO_3/H_2SO_4 mists are determined by means of cascade impactors [162].

8. Sulfur Trioxide

8.1. Properties

Physical Properties. Sulfur trioxide [*7446-11-9*], SO_3, M_T80.06, is the anhydride of sulfuric acid. It is known in all three states of matter. In the gaseous and liquid phases, an equilibrium exists between the monomer, SO_3, and the annular trimer, S_3O_9. This equilibrium is shifted towards S_3O_9 at low temperature. In the presence of slight traces of moisture (ca. 100 ppm H_2O), liquid sulfur trioxide (below ca. 27 °C) and solid sulfur trioxide are transformed into solid polymers, which form asbestos-like, felted crystalline needles. These consist of β-SO_3 and α-SO_3, which are thought to correspond to long SO_3 chains with water saturation at the chain ends – the so-called polysulfuric acids, $HO(SO_2O)_xH$ [163].

Information regarding the physical properties of sulfur trioxide is tabulated in various works [16, 164–166]. The available data originate from different measurements made over the course of the past 60 years, and they are not always mutually consistent, in part because it is extremely

difficult to prepare pure, absolutely anhydrous sulfur trioxide.

Table 2 provides a selection of important physical data. Pure, solid sulfur trioxide, referred to as γ-SO_3, forms silky orthorhombic (ice-like) crystals. The melting points cited for the polymeric forms α-SO_3 and β-SO_3 indicate temperatures at which these solids depolymerize to liquid sulfur trioxide. From an industrial operation standpoint it is important to recognize that the liquid sulfur trioxide originating from α-SO_3 and β-SO_3 exerts a significant vapor pressure: 2.4 bar at 62.2 °C [163].

Table 2. Physical properties of sulfur trioxide

Property	Value	Reference
Gaseous SO_3		
Nominal density, g/L (0°C, 1013 mbar)	3.57	[164]
Specific heat C_p, kJ m^{-3} K^{-1}		
100°C	2.543	[16]
500°C	3.191	[16]
Liquid SO_3		
Density, g/cm^3 (25°C)	1.9	[163]
bp, °C (1013 mbar)	44.8	[164]
Heat of evaporation (boiling point), J/g	538	[163]
Vapor pressure, bar		
20°C	0.26	[166]
30°C	0.47	
50°C	1.32	
100°C	8	[163]
Critical temperature, °C	217.7	[163]
Critical pressure, bar	81.9	[163]
Solid SO_3		
γ-SO_3		
mp, °C	16.86	[163]
Heat of fusion, J/g	119	[163]
β-SO_3		
"Melting point", °C	30.4	[163]
α-SO_3		
"Melting point", °C	62.2	[163]

The temperature dependence of various physical properties of sulfur trioxide has been represented graphically [165]. Formulae have also been worked out from which it is possible to calculate approximate values at any temperature for a number of properties, including vapor pressure, specific heat, and density [163, 167].

Chemical Properties [163, 168]. Because of sulfur trioxide's importance as an industrial intermediate in the production of sulfuric acid, the thermodynamics and kinetics of its generation by the oxidation of sulfur dioxide (Eq. 10) have been extensively studied, as has the reverse reaction. Pure sulfur trioxide is in fact extremely resistant to thermal decomposition because of kinetic inhibition, even at elevated temperatures where thermodynamic equilibrium is shifted heavily toward $SO_2 + O_2$. However, certain catalytically active substances are able to increase the rate of equilibration substantially. In the presence of metals such as platinum [24] or of metal oxides and sulfates (e.g., of iron, copper [169], and, of course, vanadium) the decomposition approaches equilibrium at temperatures above ca. 700 °C.

The sulfur in sulfur trioxide is present in its maximum (hexavalent) oxidation state. Sulfur trioxide is therefore a strong oxidizing agent, and at the same time it is one of the stongest known Lewis acids. This bifunctional reactivity is apparent, for example, in the reaction with hydrogen halides. With hydrogen fluoride, sulfur trioxide reacts as a Lewis acid to form fluorosulfuric acid. Hydrogen bromide and hydrogen iodide, on the other hand, are oxidized under the same conditions to the respective free halogens, while the sulfur trioxide is reduced to give sulfur dioxide or even hydrogen sulfide [163].

Sulfur trioxide reacts at 50 – 150 °C with elemental sulfur to give sulfur dioxide. This reaction is used industrially for the production of pure sulfur dioxide (→ Sulfur Dioxide, Chap. 7.1.). Coke or pulverized coal at elevated temperature will reduce sulfur trioxide to sulfur dioxide, which is the basis for the regeneration step of the Bergbau-Forschung process for flue-gas desulfurization (→ Sulfur Dioxide, Chap. 5.7.2.).

Under absolutely dry conditions, sulfur trioxide is unreactive toward most metals. With metal oxides it reacts at moderately high temperature to form the corresponding metal sulfates. Sulfur trioxide undergoes an extremely violent reaction with water to form H_2SO_4 or higher sulfuric acids. Sulfur trioxide gas reacts instantaneously with wet gases to form sulfuric acid mists, which are extremely difficult to precipitate.

As a general rule, sulfur trioxide reacts very briskly with all organic compounds. In the course of the reaction, the organic compound may be sulfonated, oxidized (decomposed) with the release of water, or dehydrated.

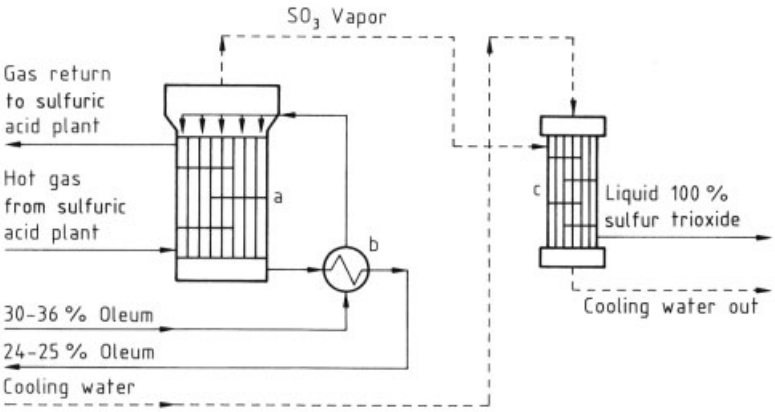

Figure 48. Production of 100 % sulfur trioxide by distillation from oleum
a) Evaporator; b) Heat exchanger; c) SO_3 condenser

8.2. Manufacture

Sulfur trioxide is of course always produced in dilute gaseous form in a contact-process sulfuric acid plant through the oxidation of sulfur dioxide-containing gases. So far, however, the only reported preparation of pure sulfur trioxide from dilute SO_3- containing gases has been a pilot-plant scale process involving cryoscopic condensation [170]. The usual procedure instead involves distillation of oleum. The heat required for oleum distillation is most conveniently supplied by hot contact gas from the associated sulfuric acid plant [171]. Figure 48 shows the method use at Hoechst.

In this approach, 30 – 36 % oleum is preheated indirectly by liquid – liquid heat exchangers and subsequently fed to the top of a falling-film evaporator for distillation. In the evaporator the oleum flows downward as a film on the inner walls of steel tubes that are heated on the outside with hot contact gas in countercurrent. The weak oleum discharged from the evaporator sump, with a residual "free" SO_3 content of 24 – 25 %, is cooled through indirect heat exchange with the entering concentrated oleum, after which it is recycled to the sulfuric acid contact plant for concentration in the oleum tower once again to 30 – 36 %. It is important that the SO_3 content of the oleum not drop below 20 % in the interest of avoiding severe corrosion of the steel tubes at the temperatures prevailing in the evaporator. Sulfur trioxide gas leaving the top of the falling-film evaporator is condensed in a tubular condenser made of steel or aluminum. The cooling water is maintained at ca. 30 °C to ensure that the sulfur trioxide temperature does not fall below 27 °C, since sulfur trioxide would otherwise solidify. The evaporation capacity of the falling-film evaporator is controlled for a particular oleum input rate by adjusting the flow rate of the hot contact gas.

The largest units of this type built so far are designed to produce ca. 3 t/h of SO_3. Apart from falling-film evaporators, forced-circulation evaporators are also used for oleum distillation. In most cases, stainless steel is utilized as the construction material.

8.3. Handling and Uses

When storing and handling liquid sulfur trioxide it is extremely important to prevent polymerization, because any solid formed is very inconvenient to remove. Sulfur trioxide reacts explosively with water, so the vessels must never be cleaned with water. Concentrated sulfuric acid is required for dissolving solid deposits of sulfur trioxide.

In order to maintain sulfur trioxide in the liquid state during storage, tanks and piping must be kept at ca. 30 °C by supplementary heating. Addition of a stabilizer also effectively inhibits the formation of solid polymers at temperatures above the natural solidification point of sulfur trioxide. A number of inorganic and organic compounds are used or have been recommended

as stabilizers, but the extensive patent literature on the subject provides no indication of their mode of action [163].

Liquid sulfur trioxide is among the most dangerous of all industrial materials, and strict safety regulations govern its handling [172, 173]. In the Federal Republic of Germany, for example, every act of transporting liquid sulfur tridioxide over public roads requires a special permit, which can be withdrawn at any time.

Pure sulfur trioxide is used in organic synthesis for sulfonation reactions, including the manufacture of chlorosulfonic acid, thionyl chloride, aminosulfonic acid, dimethyl sulfate, and sulfamide. The reaction is difficult to control if the sulfur trioxide is introduced in liquid form, so it is preferentially supplied as a gas (pure or diluted with an inert gas), as a solution in liquid sulfur dioxide or some other solvent that is inert under the reaction conditions, or as an addition compound with an organic base.

9. Toxicology

Given their high chemical reactivities, it is not surprising that both sulfuric acid and sulfur trioxide have marked toxic and ecological effects [174–176].

9.1. Sulfuric Acid

Effects on Animals. Sulfuric acid has a highly corrosive effect on the eyes, the mucous membranes, and the skin, even in low concentrations. Because it completely destroys living tissue, concentrated sulfuric acid causes burns that penetrate deeply and heal only slowly. Swallowing sulfuric acid produces extreme pain in the digestive tract, vomiting, and shock, and there is a danger of perforation. Sulfuric acid vapors or mists irritate the eyes and the muscous membranes of the nose, pharynx, and respiratory tract, causing heavy coughing and breathlessness. Chronic inflammation of the upper respiratory tract may result, and eventually impairment of the lungs (chronic bronchitis). Even concentrations well below 0.1 vol % (corresponding to ca. 4 g/m^3) will render breathing impossible. Pain and damage to the teeth have also been observed, especially the incisors. The maximum permitted workplace concentration of vapors and mists under German regulations (MAK) is 1 mg/m^3 (ca. 0.245 ppm by volume).

Environmental Effects. Sulfuric acid releases are highly deleterious with respect to ground and surface waters. The substance is toxic to both fish and algae, both directly and as a result of reaction with other materials in the water. Any concentration $\geq$ 1.2 mg/L is considered lethal to fish; 6.3 mg/L or more causes death within 24 h [176].

9.2. Sulfur Trioxide

In principle, sulfur trioxide has the same toxic effects as sulfuric acid [173]. Inhalation of the gas itself or the sulfuric acid mist formed when it comes into contact with humid air causes irritation, burning, and degeneration of the tissue of moist skin, eyes, and mucous membranes, especially in the respiratory tract. This happens relatively rapidly at concentrations $\geq$ 10 ppm by volume. Laryngal edema, chronic bronchitis, and dental damage are possible. In contrast, symptoms of resorptive poisoning have not been observed.

10. References

1. *Gmelins Handbuch der Anorganischen Chemie* "Schwefel," vol. B 2, system no. 9, Verlag Chemie, Weinheim 1960.
2. O. T. Fasullo: *Sulfuric Acid* – Use and Handling, McGraw-Hill, New York 1965.
3. R. M. Perry: *Chemical Engineers' Handbook,* McGraw-Hill, New York 1973.
4. T. R. Bump, W. L. Sibbitt, *Ind. Eng. Chem.* **47** (1955) no. 8, 1665 – 1670.
5. P. Pacquiez, *Ind. Chim. (Paris)* **45** (1958) no. 495, 292 – 298.
6. C. M. Gable, H. F. Betz, S. H. Maron, *J. Am. Chem. Soc.* **72** (1950) no. 4, 1445 – 1448.
7. R. Haase, M. Rehse, *Mitt. VGB* **62** (1959) 367 – 371.
8. G. P. Lucinskij, *Zh. Fiz. Khim.* **30** (1956) no. 6, 1207 – 1222.
9. R. Biarnes, *Plant/Oper. Prog.* **1** (1982) no. 4, 230 –235.
10. *Sulphur* **166** (1983) 34 – 37.

11. J. L. Gmitro, T. Vermeulen, *AIChE J.* **10** (1964) no. 5, 740 – 746.
12. F. H. Verhoff, J. T. Banchero, *Chem. Eng. Prog.* **70** (1974) no. 8, 71 – 72.
13. R. R. Pierce, *Chem. Eng.* **84** (1977) no. 8, 125 – 128.
14. R. J. Jaworowski, S. S. Mack, *J. Air. Poll. Control Assoc.* **29** (1979) no. 1, 43 – 46.
15. R. Schumacher, *VDI Ber.* **246** (1975) 119 – 128.
16. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin 1973.
17. W. D. Ross, *Chem. Eng. Prog.* **48** (1952) no. 6, 314 –315.
18. M. Bodenstein, M. Katayama, *Z. Elektrochem. Angew. Phys. Chem.* **15** (1909) no. 8, 244 – 249; *Z. Phys. Chem.* **69** (1909) 26 – 51.
19. M. Fairlie: "Sulfuric Acid Manufacture," *ACS Monogr.* **69** (1936).
20. *Gmelin,* "Schwefel", system no. 9, vol. A (1953).
21. P. Pacquiez: "Evolution de la fabrication de l'acide sulfurique pendant les trente dernieres annees," *Ind. Chim.* **41** (1954) no. 445, 226 – 240 to **50** (1963) no. 556, 413 – 432.
22. W. W. Duecker, J. R. West (eds.): "The Manufacture of Sulfuric Acid," *ACS Monogr.* **144** (1959).
23. B. Waeser: *Die Schwefelsäurefabrikation,* Vieweg, Braunschweig 1961.
24. R. Knietsch, *Ber. Dtsch. Chem. Ges.* **34** (1901) 4069 – 4115.
25. BASF, DE 291 792, 1913.
26. Bayer, DE 1 136 998, 1960.
27. *Sulphur* **54** (1964) 30 – 31.
28. U. Sander, E. Winkler: "Energieverbund mit Schwefelsäureanlagen," *Mehrfachnutzung industrieller Prozesswärme,* Resch, Grafelfing 1984.
29. G. B. Whyte: "Improved Energy Recovery on a Sulphuric Acid Plant," *IFA Technical Conference,* Kallithea, Oct. 1982.
30. L. J. Friedman: "Sulfuric Acid Energy Design for the 80s," in: *Sulfuric/Phosphoric Acid Plant Operations,* American Institute of Chemical Engineers, New York 1982, pp. 101 – 107.
31. *Chem. Week,* 10 Feb. 1982, 40 – 43.
32. W. A. McGill, M. J. Weinbaum: "Maximizing Energy Efficiency in Sulphuric Acid and SO2 Recovery Plants by Controlled Metallurgy," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Int. Conf.* (1981) 251 – 274.
33. U. Sander, G. Daradimos, *Chem. Eng. Prog.* **74** (1978) no. 9, 57 – 67.
34. Statistics of German Federal Bureau of Statistics, Wiesbaden.
35. J. E. Shelton: *Sulfur,* US Bureau of Mines, Washington, DC, 1979.
36. T. J. P. Pearce: "Sulphuric acid: Physicochemical Aspects of Manufacture," in G. Nickless (ed.): *Inorganic Sulphur Chemistry,* Elsevier, Amsterdam 1968, pp. 535 – 561.
37. M. Bodenstein, W. Pohl, *Z. Elektrochem. Angew. Phys. Chem.* **11** (1905) no. 24, 373 – 384.
38. *JANAF Thermochemical Tables,* 2nd ed., US National Bureau of Standards, 1971.
39. F. D. Richardson, J. H. E. Jeffes, *J. Iron Steel Inst. London* **171** (1952) 165 – 175.
40. U. Sander, *Metall (Berlin)* **31** (1977) no. 12, 1353 –1356; **32** (1978) no. 1, 95 – 98.
41. N. Neth, G. Kautz, H.-J. Schuster, U. Wagner, *Chem. Ing. Tech.* **51** (1979) no. 8, 825.
42. C. N. Kenney, *Catalysis* **3** (1980) 123 – 135.
43. P. Mars, J. G. H. Maessen, *J. Catal.* **10** (1968) 1 – 12.
44. H.-J. Becker, W. Gosele, N. Neth, J. Adlkofer, *Chem. Ing. Tech.* **51** (1979) no. 8, 789 – 795.
45. A. Urbanek, M. Trela, *Catal. Rev. Sci. Eng.* **21** (1980) no. 1, 73 – 133.
46. M. Appl, N. Neth: "Problems and Advances in the Catalytic Oxidation of Sulphur Dioxide," in: "Fertilizer Acids," Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers 1979, paper XX. *Sulphur* **145** (1979) 26 – 29, 38 – 39.
47. G. T. Ketjen: "Advances in Catalyst Quality and Conversion Equipment for Sulphuric Acid Manufacture," *Proceedings of the Fertilizer Society,* London 1978, no. 170.
48. J. R. Donovan, J. S. Palermo, *Sulphur* **138** (1978) 39 –41; *Sulphur* **131** (1977) 46 – 50.
49. S. K. May, *Sulphur* **146** (1980) 30 – 34.
50. J. B. Rinckhoff, L. J. Friedman, *Chem. Eng. Prog.* **73** (1977) no. 3, 79 – 84.
51. U. Sander, AIME-TMS paper no. A 79, Feb. 1979.
52. M. L. Brown: "The Lurgi Double Contact Sulfuric Acid Process," *Proceedings of the Fertilizer Institute, Environmental Symposium,* New Orleans 1976, pp. 471 – 486.
53. F. J. Doering, J. L. Gaddy, *Comput. Chem. Eng.* **4** (1980) 113 – 122.
54. R. Mann, I. J. Gardner, C. Morris, *Chem. Eng. Sci.* **35** (1980) 185 – 192.
55. U. Sander, *Erzmetall* **31** (1978) no. 4, 163 – 169.

56. C. V. Narayanaswamy, P. Sathyanarayana, K. Ramachandran, *Chem. Age India* **29** (1978) no. 10, 811–817.
57. U. H. F. Sander, *Sulphur* **164** (1983) 36–42.
58. I. J. Mukhlenov, V. B. Sarkich, B. N. Osipova, *Khim. Promst.* **11** (1963) 833–836.
59. Bayer AG, DE-OS 1 926 564, 1970; DE-AS 1 929 388, 1972.
60. W. Becker: "The Catalytic Oxidation of Sulphur Dioxide by Fluidized-bed Technology," in *Making the Most of Sulphuric Acid,* pp. 45–57; *Sulphur* **158** (1982) 30–33.
61. D. R. Duros, E. D. Kennedy, *Chem. Eng. Prog.* **74** (1978) no. 9, 70–77.
62. H. Geipel, *VDI-Ber.* **294** (1978) 113–117.
63. M. Eber, *Chemie-Technik* **6** (1977) no. 12, 523–527.
64. G. C. Pedersen: "New Mist Eliminator for Improved Sulfuric Acid Plant Operation," in: *Sulfuric/Phosphoric Acid Plant Operations,* American Institute of Chemical Engineers, New York 1982, pp. 161–170.
65. O. Hansen, *Sulphur* **159** (1982) 29–33.
66. P. Cross, S. Sjögren, L. Berthagen, B. Linder: "A New Design of Plate Exchanger: the Welded Heat Exchanger with Anodic Protection for Sulphuric Acid Cooling and Heat Recovery at Elevated Temperatures, in: Making the Most of Sulphuric Acid, *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 157–179.
67. P. H. Cross, *Chem. Eng.* **378** (1982) no. 3, 87–90.
68. F. Paulekat, H. Gräfen, D. Kuron, *Werkst. Korros.* **33** (1982) 254–262.
69. *Sulphur* **132** (1977) 35–40.
70. M. D. Farkas, R. R. Dukes: "Multiple Routes to Sulphuric Acid," in: *Sulphur and SO2 Developments,* American Institute of Chemical Engineers, New York 1971, pp. 8–12.
71. D. B. Burkhardt, *Chem. Eng. Prog.* **64** (1968) no. 11, 66-70.
72. T. J. Browder, *Chem. Eng. Prog.* **67** (1971) no. 5, 45–50.
73. R. W. Riedel, J. J. Knight, R. E. Warner, *Chem. Eng. Prog.* **73** (1977) no. 3, 55–60.
74. S. Lal, *CEW Chem. Eng. World* **16** (1981) no. 1, 53–58, 82.
75. U. Sander, *Chem. Eng. Prog.* **73** (1977) no. 3, 61–64.
76. W. H. Streamer, I. T. D. Knapp, *CHEMECA* **77** (1977) 63–67.
77. U. Sander, H. J. König, *Proceedings of the AiChE/IMIQ Joint Meeting,* Acapulco, Mexico, Oct. 1980, vol. 11.
78. H. C. Hollyman, A. Phillips, W. R. Stead: "Energy Conservation in Sulphuric Acid Plants," in: "Fertilizer Acids", *Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXIX.
79. D. Bhaga, *Sulphur* **147** (1980) 32,33,39–43.
80. D. J. Higgins, *Chem. N. Z.* **44** (1980) no. 2, 56–58.
81. R. B. Catlett, D. L. Hill: "Design and Equipment Considerations in the World's Largest Single-Train, Double-Absorption Sulphuric Acid Plant," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 425–437.
82. J. M. Kovacik, J. C. Franklin, *Chem. Eng. Prog.* **78** (1982) no. 2, 58–64.
83. S. A. Walsh: "Energy Recovery in a Phosphoric Acid Plant," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 111–129.
84. U. Sander, M. Beckmann, *Sulphur* **162** (1982) 33–39.
85. *Sulphur* **159** (1982) 40–42.
86. P. Askenasy, O. Merkel, DE 640 626, 1932.
87. Chemical Construction Corporation, US 3 432 263 and 3 455 652, 1966.
88. Ugine Kuhlmann, FR 1 541 087, 1967.
89. Canadian Industries, DE 2 255 206, 1972.
90. T. J. Browder, *Chem. Eng. Prog.* **73** (1977) no. 3, 70–76.
91. G. M. Cameron, P. D. Nolan, K. R. Shaw, *Chem. Eng. Prog.* **74** (1978) no. 9, 47–50.
92. B. Vidon, *Chem. Process Eng. (London)* **53** (1972) no. 7, 34–35.
93. R. A. Bauer, B. P. Vidon, *Chem. Eng. Prog.* **74** (1978) no. 9, 68–69.
94. *Sulphur* **155** (1981) 29–31.
95. G. Srinivasan, *Sulphur* **148** (1980) 32–36.
96. *Phosphorus Potassium* **89** (1977) 36–44.
97. J. R. Shafer, J. S. Bogadi, *Chem. Eng. Prog.* **76** (1980) no. 10, 70–75.
98. U. Sander, E. Weilandt, AIME-TMS paper select. A 77, paper no. 102, March 1977.
99. R. Herbert, *Chem. Ing. Tech.* **53** (1981) no. 5, 374–376.
100. E. Herrmann, *Fernwärme Int. (Frankfurt/Main)* **12** (1983) no. 1, 10–13.
101. V. K. Agrawal, *CEW Chem. Eng. World* **16** (1981) no. 11, 53–58.
102. *Sulphur* **152** (1981) 34,35.
103. *Sulphur* **123** (1976) 36–42.
104. *Sulphur* **141** (1979) 41.
105. O. B. Jorgensen, P. Schoubye, *Chem. Eng. Prog.* **78** (1982) no. 2, 44–47.
106. P. Schoubye, *Oil Gas J.* **79** (1981) no. 51, 55–59.

107. P. Schoubye, *Sulphur* **137** (1978) 38 – 42.
108. W. Kerner, P. Schmitz, *Chem. Tech. (Heidelberg)* **8** (1979) no. 2, 83 – 86.
109. R. Peichl, *Chem. Tech. (Heidelberg)* **11** (1982) 21 –24.
110. M. Schmitt, *Chem. Ind. (Düsseldorf)* **33** (1981) no. 5, 286 – 290.
111. J. M. Connor: "Sulfur removal," in B. L. Crynes (ed :): *Chemical Reactions as a Means of Separation,* Marcel Dekker, New York 1977, pp. 34 – 72.
112. G. Srinivasan, B. R. Angles: "Fertilizer Acids," *Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXII.
113. R. Mann, I. J. Gardner, C. Morris: *Symposium of the Institute of Chemical Engineers,* Series no. 57, pp. L1 – L14, 1979.
114. K. L. Sullivan, M. E. Doyle: "Modifying Contact Plants to Simultaneously Increase Sulphuric Acid Production and Reduce Emissions," in "Fertilizer Acids", *Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXVII.
115. W. G. Tucker, J. R. Burleigh: "SO2 Emission Control from Acid Plants," in "Sulphur and SO2 Developments," American Institute of Chemical Engineers, New York, 1971, pp. 103 – 109.
116. J. G. Kronseder: "Cost of Reducing Sulphur Dioxide Emissions," in "Sulphur and SO2 Developments," American Institute of Chemical Engineers, New York 1971, pp. 25 – 28.
117. L. J. Friedman, *J. Met.* **33** (1981) no. 3, 44-50.
118. L. J. Friedman in "Sulfur Dioxide Control in Pyrometallurgy," TMS-AIME (1981) pp. 189 – 203.
119. L. J. Friedman, *Environ. Prog.* **1** (1982) no. 4, 281 –284.
120. S. Villa: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp. 5th Conf.* (1981) 471 – 492.
121. *Sulphur* **117** (1975) 32 – 38.
122. U. Sander, *Dechema Monogr.* **80** (1976) no. 2, 379 –391.
123. R. Peichl, *Ber. Bunsenges. Phys. Chem.* **82** (1978) 1194 – 1198.
124. *Sulphur* **162** (1982) 14.
125. U. Sander: "Economic Aspects of Processing Very Weak SO2 Gases to Sulfuric Acid," in: *Sulfuric /Phosphoric Acid Plant Operations,* American Institute of Chemical Engineers, New York 1982, pp. 6 – 18.
126. A. Gruber, A. Walitt: Report no. APTD-1156, Contract CPA 70-59 for NAPCA, Oct. 1970.
127. V. Fattinger: "Production of Sulphur (and Nitric) Acid by a Modified Nitrogen-oxide Sulphuric Acid Process," in "Fertilizer Acids," *Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXVI.
128. K. Bodenbrenner, H. von Plessen, R. Steiner, *Chem. Ing. Tech.* **50** (1978) no. 1, 30 – 36.
129. A. Kirsch, C. K. McEwen: "Purification and Reconcentration of Waste Sulphuric Acid," in "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 275-284.
130. U. Sander, H. W. Schmidt, *Verfahrenstechnik* **10** (1976) no. 4, 197 – 201.
131. A. Heitmann, P. Reher, *Chem. Ing. Tech.* **46** (1974) no. 14, 589 – 594.
132. G. Kurz, *Verfahrenstechnik (Mainz)* **2** (1968) no. 9, 374 – 378.
133. G. Gerlach, *Chem. Ing. Tech.* **42** (1970) no. 7, 452 –456.
134. T. Kobayashi, *CEER Chem. Econ. Eng. Rev.* **8** (1976) no. 6, 42 – 48.
135. G. M. Smith, E. Mantius, *Chem. Eng. Prog.* **74** (1978) no. 9, 78 – 83.
136. J. M. Connor, *Sulphur* **131** (1977) 39 – 45.
137. I. Rodger: "Recovery and Reconcentration of Spent Sulphuric Acid," in "Fertilizer Acids", *Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXV. Sulfuric/Phosphoric Acid Plant Operations, American Institute of Chemical Engineers, New York 1982, pp. 151 – 155.
138. K. Bodenbrenner, H. von Plessen, H. Vollmüller, *Dechema Monogr.* **86** (1979) 197 – 218.
139. F. S. Chambers, R. F. Peterson, *Chem. Eng. Prog.* **43** (1947) no. 5, 219 – 224.
140. *Sulphur* **144** (1979) 39 – 40.
141. *Chem. Eng.* **86** (1979) no. 17, 109 – 110.
142. L. Douren: "The BOSAC Sulphuric Acid Recovery System," in: "*Making the Most of Sulphuric Acid,*" *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 317 – 335.
143. H. R. Kung, P. Reimann: "Cleaning and Reconcentration of Sulphuric Acid Contaminated by Organic Impurities," in: *Making the Most of Sulphuric Acid, Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) pp. 295 –316.
144. H. R. Kueng, P. Reimann, *Chem. Eng.* **89** (1982) no. 8, 72 – 73.
145. J. F. Burke, E. Mantius, *Chem. Eng. Prog.* **43** (1947) no. 5, 237 – 246.
146. J. S. Gaddas, I. M. Smith: "Reconcentration of Spent Sulphuric Acid with Preservation of Organic Values," in: *Making the Most of Sulphuric Acid, Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) pp. 337-350.

147. B. Cance, N. J. Stockley: "Sulphuric Acid Concentration," in *Fertilizer Acids, Proc. Br. Sulphur Corp. 3rd Int. Conf. Fertilizers* 1979, paper XXXII.
148. J. C. D. Brand, A. Rutherford, *J. Chem. Soc.* 1952, 3916 – 3922.
149. F. D. Miles, H. Niblock, G. L. Wilson, *Trans. Faraday Soc.* **36** (1940) 345 – 356.
150. J. W. McDavid, *J. Soc. Chem. Ind. (London)* **43** (1924) no. 12, 57 – 61.
151. G. E. McClain, W. A. Mueller, *Chem. Eng. Prog.* **78** (1982) no. 2, 48 – 50.
152. G. E. McClain: "Materials and Technology Considerations in Sulphuric Acid Plant Energy Recovery," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 207 – 226.
153. G. Berglund, R. Mellström: "Application of Stainless Steel in Sulphuric Acid Environments," in: "Making the Most of Sulphuric Acid," *Proc. Br. Sulphur Corp.* **5th** *Conf.* (1981) 227 – 250.
154. *Dechema-Werkstofftabellen,* Chemische Beständigkeit/Schwefelsäure, Dechema, Frankfurt 1969.
155. G. F. Eckert, Jr., *Eng. Min. J.* **182** (1981) no. 3, 87 –91.
156. *Sulphur* **149** (1980) 32 – 37.
157. *Analyse der Metalle,* 2nd ed., **vol. II,** part 1, Springer Verlag, Berlin 1961, pp. 666 – 689.
158. J. H. Karchmer (ed.): *The Analytical Chemistry of Sulphur and its Compounds,* part 1, Wiley-Interscience, New York 1970, pp. 183 – 283.
159. F. D. Snell, L. S. Ettre (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 18,** Wiley-Interscience, New York 1973, pp. 360 – 440.
160. VDI-Richtlinie 2462, Bl. 7 (1979).
161. W. Guse, *Staub Reinhalt. Luft* **41** (1981) no. 6, 204 –210.
162. K. G. Hebel, R. L. Protzmann, R. Davis, R. Knarr, *AIChE Symp. Ser.* **70** (1974) 299 – 306.
163. *Gmelin,* "Schwefeloxide" system no. 9, suppl. vol. 3, pp. 247 – 338.
164. *VDI-Wärmeatlas: Berechnungsblätter für den Wärmeübergang,* 3rd ed., VDI-Verlag, Düsseldorf 1977.
165. C. L. Yaws, K. Y. Li, C. H. Kuo, *Chem. Eng.* **81** (1974) no. 14, 85 – 92.
166. R. A. Hyne, P. F. Tiley, *J. Chem. Soc.* 1961, 2348 –2352.
167. C. L. Yaws et al., *Chem. Eng.* **83** (1976) no. 17, 79 –87; **83** (1976) no. 23, 127 – 135; **83** (1976) no. 25, 153 –162.
168. P. W. Schenck, R. Steudel: "Oxides of Sulphur," in: G. Nickless (ed.): *Inorganic Sulphur Chemistry,* Elsevier, Amsterdam 1968, pp. 367 – 418.
169. Produits Chimiques Ugine Kuhlmann, US 4 158 048, 1979.
170. J. Cathala, *Chem. Ing. Tech.* **25** (1953) no. 6, 285 –291.
171. IG Farbenindustrie, IT 391 569, 1941; BE 444 266, 1942.
172. Chemical Safety Data Sheet SD-101, Manufacturing Chemists' Association, Washington, DC 1976.
173. R. Kühn, K. Birett: *Merkblätter gefährliche Arbeitsstoffe,* 3rd ed., Moderne Industrie, W. Dummer & Co., München, Loseblatt-Sammlung, Blatt S13.
174. BASF, Technisches Merkblatt M 5430 d; Schwefelsäure technisch rein, Ludwigshafen 1974.
175. R. Kühn, K. Birett in [173], Blatt S12.
176. G. Hommel: *Handbuch der gefährlichen Güter,* Springer Verlag, Berlin 1980, Merkblatt 183 and 184.
177. J. Adlkofer, E. Dieckmann, E. Winkler, *British Sulphur* **229** (1993) 50 – 52.
178. *British Sulphur* **189** (1987).
179. J. Sheputis, Monsanto – Enviro-Chen, AICHE Clearwater Convention, May 1988.
180. P. Lüdtke, Lurgi Metallurgie, GDMB Seminar, Schwerin, Nov. 1993.
181. P. Morsing, Haldor Topsoe, personal communication.
182. Sachleben GmbH, *Spektrum,* April 1989.
183. *British Sulphur* **216** (1991).

Sunflower Oil → Fats and Fatty Oils
Sunscreens → Skin Cosmetics
Suntan Preparations → Skin Cosmetics

Superconductors

Hermann Rietschel, Kernforschungszentrum Karlsruhe, Karlsruhe, Federal Republic of Germany (Chaps. 1 – 3)

Hans E. Hoenig, Institut für physikalische Hochtechnologie (IPHT), Jena, Federal Republic of Germany (Chap. 4)

Günther Bogner, Siemens AG, Erlangen, Federal Republic of Germany (Chap. 5)

Denis Jérome, Université Paris-Sud, France (Chap. 6)

1. Introduction 2
2. Principles 3
2.1. Electrical Resistance and Thermal Conductivity 3
2.2. Behavior in Magnetic Fields 3
2.3. Critical Current 5
2.4. Energy Gap and Thermodynamic Properties 6
2.5. Josephson Effects 8
2.6. Theoretical Descriptions 8
3. Classes of Superconductors 9
3.1. Classical Superconductors 10
3.2. Exotic Superconductors 11
3.3. High-Temperature Superconductors 12
4. Electronic Applications of Superconductivity 14
4.1. Superconductivity Effects Important for Electronic Applications . . . 14
4.1.1. Pure Inductances 14
4.1.2. Small High-Frequency Losses 15
4.1.3. Energy-Gap Effects 15
4.1.4. Quantum Interference Effects 16
4.2. Josephson Junctions, Tunnel Junctions, and Weak Links 16
4.2.1. Junction Types and Their Significance 16
4.2.2. Josephson Circuits, Digital Circuits, Digital Signal Processing, and Voltage Standards 17
4.2.3. SQUIDs 18
4.2.4. SQUIDs and Biomagnetism 20
4.3. Applications of HT Superconductors 20
4.3.1. Materials and Techniques for HT-Superconducting Electronics . . . 21
4.3.2. Operating Temperatures of HT-Superconducting Electronics . . . 22
4.3.3. Passive Components Based on HT Superconductors 22
4.3.4. HT-Superconductor Radiation Detectors 23
4.3.5. Nonlinear HT-Superconductor Components 23
4.3.6. HT-Superconductor SQUIDs 24
4.4. Refrigerators for Cryoelectronics . 25
5. Application of Superconductivity in Magnet and Power Engineering . . 26
5.1. Introduction 26
5.2. Industrial Superconductors 26
5.2.1. Metallic Superconductors 26
5.2.2. Oxide-Ceramic Superconductors . . . 28
5.3. Potential Superconductivity for Improvements in Conventional Electrical Devices 28
5.3.1. Superconducting Magnets for High-Energy Physics 28
5.3.2. Magnetic Separation and Purification 29
5.3.3. Superconducting Levitation for High-Speed Transportation Systems 30
5.3.4. Generators and Motors with Superconducting Windings 30
5.3.5. Superconducting Transformers 31
5.3.6. Superconducting Power-Transmission Cables 32
5.4. Novel Electrical Devices for Which Superconductors Are Indispensable 33
5.4.1. Magnets for Magnetic Resonance Imaging (MRI) and Spectroscopy . . . 33
5.4.2. Magnet Systems for Magnetic-Confinement Fusion Reactors 33
5.4.3. Magnetohydrodynamic Energy Conversion 35
5.4.4. Superconducting Magnetic Energy Storage (SMES) 35
5.4.5. Superconducting Current Limiters . . 36
6. Organic Superconductors 37
6.1. Introduction 37
6.2. Electronic Structure and Superconductivity 38
6.3. Other Features of 1-D Superconductors 39
6.4. Prospects for Higher T_c and Applications 42
7. References 42

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a25_705

1. Introduction

In 1911, Heike Kamerlingh Onnes discovered that the electrical resistance of mercury abruptly drops to an unmeasurably small value at 4.2 K, and named the phenomenon superconductivity. In the following years, superconductivity was found in many other metals and in a vast number of metallic compounds and alloys. Although the highest known temperature of the transition to the superconducting state (critical temperature T_c) had reached 23 K by 1986, in the majority of cases it was < 10 K. Stable cooling to achieve this so-called conventional superconductivity is possible only with liquid helium, so that the technology is both complex and costly. As a result, the wide adoption of superconductivity in industry has been hindered until now, although a number of special applications (e.g., in magnets, sensors, and microwave electronics) do exist. A comprehensive treatment of the physical properties of conventional superconductivity and some of its industrial applications has appeared in [1].

In 1986, Bednorz and Müller discovered superconductivity at ca. 35 K in an oxide ceramic of composition $(La, Ba)_2CuO_4$ [2]. This discovery initiated a rapid development worldwide, leading to a new class of materials called high-temperature (HT) superconductors, also referred to as ceramic, oxide, or cuprate superconductors. The range of critical temperatures has now been extended up to ca. 133 K. The resulting possibility of maintaining stable superconductivity by cooling with liquid nitrogen (liquid-N_2, 77 K) has led to reconsideration of industrial applications of superconductivity. If some of the initially euphoric predictions (e.g., replacement of high-tension cables by liquid-N_2-cooled superconducting cables) have proved infeasible, many potential industrial applications of HT superconductivity can be perceived, especially in microwave technology and sensors. It will take years, however, to determine the full scope of the potential applications of HT superconductors, and the development of new processes and techniques that take into account the extreme brittleness of these materials will be a crucial step.

In physical terms, superconductivity can be understood only in the context of the quantum mechanics of multiparticle systems (quantum field theory). Furthermore, quantitative predictions require thorough and detailed knowledge of solid-state properties such as electronic band structure and lattice dynamics. Conversely, measurements of parameters in the superconducting state permit inferences about solid-state properties in the normally conducting state. Superconductivity is thus a highly important part of modern solid-state research, as is demonstrated by the awarding of four Nobel prizes in physics (to date) in this field. The 1913 prize went to H. K. Onnes for the discovery of superconductivity; the prize in 1972, to J. Bardeen, L. N. Cooper, and J. R. Schrieffer for their microscopic description of superconductivity (BCS theory); in 1973, to B. D. Josephson for his prediction of quantum coherence effects in two coupled superconductors; and the 1987 prize, to G. Bednorz and K. A. Müller for the discovery of high-temperature superconductivity.

2. Principles

2.1. Electrical Resistance and Thermal Conductivity

An ideal superconductor transports direct electrical current without loss. Although it cannot be proved experimentally that the electrical resistance of a metal in the superconducting state is exactly zero, long-term persistent current experiments with a superconducting ring have shown that the resistance drops by at least 15 orders of magnitude at the onset of superconductivity. The minuscule residual resistance, if any, is thus without detectable consequences.

In contrast, the transport of alternating current always involves losses, even in an ideal superconductor, so that additional cooling power is required. This a.c. resistance of a superconductor increases with increasing frequency ν. For industrially optimized superconductors, itis still some orders of magnitude less than the resistance of (normally conducting) copper or gold at low frequencies (e.g., 50 Hz); only above ca. 100 – 200 GHz do such superconductors exhibit higher a.c. losses.

The thermal conductivity of a superconductor also changes at the onset of superconductivity. For metals in which the contribution of conduction electrons to heat transport dominates the

lattice contribution, the thermal conductivity decreases in the superconducting state. This effect is used in thermal switches, where a bridge made of a superconductor can be switched from the superconducting state ("off") to the normally conducting state ("on") by applying a magnetic field.

2.2. Behavior in Magnetic Fields

When a superconductor is cooled in an external magnetic field, the magnetic field is completely expelled from the superconductor at the onset of superconductivity, except for a surface layer of thickness λ_L (London penetration depth, ca. 10 – 1000 nm). Persistent currents are set in motion in the surface layer, and the magnetic field due to these currents compensates the external field in the interior of the superconductor, which thus behaves as an ideal diamagnetic material.

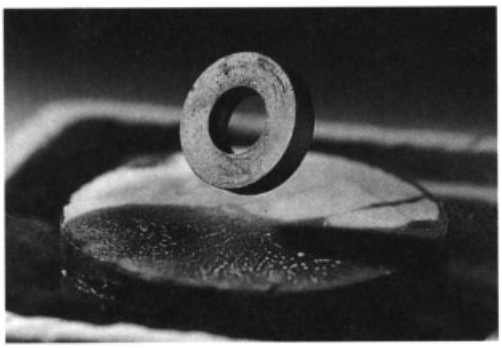

Figure 1. A ring-shaped magnet is levitated above a liquid-nitrogen-cooled high-temperature superconductor by persistent supercurrents

This Meissner – Ochsenfeld effect can be demonstrated strikingly by the levitation of a permanent magnet. If a small magnet is placed on a warm superconductor, which is then cooled (for example, by pouring LN_2 over it) the magnet is raised up at the onset of superconductivity and remains in a levitated state (see Fig. 1). Because the force levitating the magnet does not involve a time variation of the external field, this behavior cannot be accounted for by induction, which could be explained in terms of classical electrodynamics supplemented by the assumption of ideal conduction. Instead, this effect proves that the superconducting state is a new thermodynamic phase. The Meissner – Ochsenfeld effect allows unambiguous detection of the onset of superconductivity.

With regard to their response to an increasing external magnetic field, two types of superconductors can be distinguished. In type I superconductors, the shielding currents, and thus the free energy of the superconducting state, increase with increasing magnetic field strength until, at a field strength B_c (the thermodynamic critical field), the normally conducting state becomes energetically more favorable. For field strengths $B_a > B_c$, superconductivity is disrupted and, with it, the magnetization M compensating the external field. The field-free Meissner phase abruptly goes over to the normally conducting, field-filled phase (see Fig. 2).

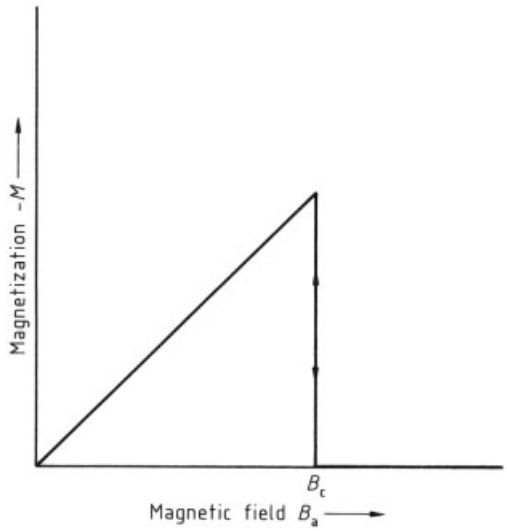

Figure 2. Magnetization curve of a type I superconductor

In a type II superconductor, the Meissner phase exists only up to a field strength B_{c1} (the lower critical field), where $B_{c1} < B_c$. If the external field exceeds B_{c1}, magnetic field in the form of flux filaments (fluxoids) penetrates the material, which remains superconducting. This is the mixed or vortex state. As the magnetic field strength B_a increases further, the surface density of fluxoids increases, so that the magnetization M compensating the external field decreases further until superconductivity collapses at $B_a > B_{c2}$ (upper critical field) and the material enters the normally conducting phase (see Fig. 3).

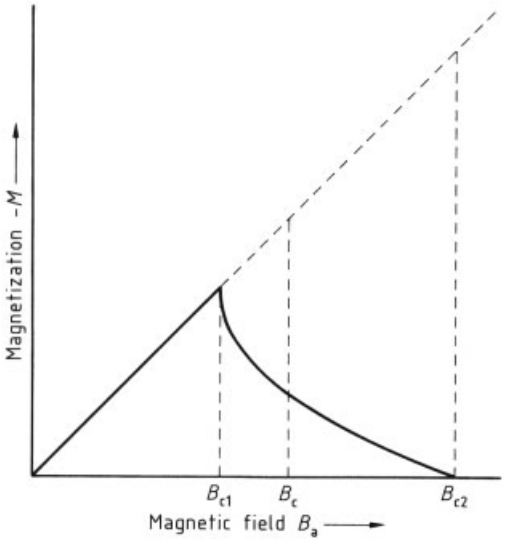

Figure 3. Magnetization curve of a type II superconductor

B_c, B_{c1}, and B_{c2} are temperature dependent and decrease steadily for $T \to T_c$. Figure 4 presents sample plots of B_c versus T. While the critical fields for type I superconductors are small, the upper critical field strengths B_{c2} in type II superconductors can take on very high values; examples are Nb_3Sn ($T_c \approx 18$ K) up to about 30 T and the high-temperature superconductor $YBa_2Cu_3O_7$ ($T_c \approx 92$ K) up to 100 T. Therefore, superconductors potentially suitable for technical applications are almost exclusively of type II.

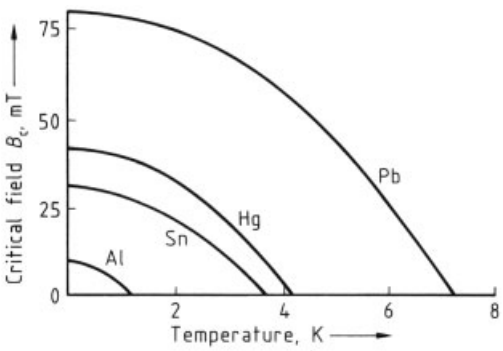

Figure 4. Thermodynamic critical field $B_c(T)$ for the elements Al, Sn, Hg, and Pb

The mean free path l of conduction electrons is a decisive parameter for the occurrence of type II superconductivity. If l is diminished in a type I superconductor (e.g., by incorporation of defects or by alloying with a second material), the transition to type II superconductivity takes place at a certain value of l. Most alloys and intermetallic compounds are therefore of type II; high-temperature superconductors represent the extreme case.

The magnetic flux penetrating a superconductor must be an integer multiple of the elementary fluxoid quantum $\Phi_0 = h/2e$, where h is Planck's constant and e is the elementary charge; Φ_0 is ca. 2×10^{-15} Wb. This flux quantization follows directly from the quantum coherence of the superconducting state; the appearance of twice the elementary charge $2e$ in Φ_0 indicates that electron pairs are the superconducting charge carriers (see Section 2.6). Each of the fluxoids in the mixed state carries exactly one fluxoid quantum.

Flux quantization can be quantitatively proved with small superconducting rings in which a persistent current has been induced.

2.3. Critical Current

The existence of a critical magnetic field, beyond which superconductivity is disrupted, directly implies the existence of a critical current (because of the intrinsic magnetic field generated by this current). Accordingly, each superconductor has a maximum transport current I_c that it can carry. This critical current is of central importance for industrial applications. It depends on the material properties and the geometry of the superconductor, the temperature, and the strength of any applied external magnetic field. Because the critical fields B_c, B_{c1}, and B_{c2} become smaller as the temperature rises and vanish for $T \geq T_c$, critical currents exhibit a corresponding temperature dependence.

Because type I superconductors have relatively low critical fields, they also have little current-carrying capacity. For example, a lead wire 1 mm in diameter has $I_c \approx 200$ A at $T = 0$ K. This value is easily derived, since the transport current flowing through the wire must generate at the surface of the wire a magnetic field with B_c (Pb) ≈ 80 mT.

Many type II superconductors have very high (upper) critical fields, which imply very high critical currents. This gives rise to a new problem that does not exist for type I superconductors.

Due to the transport current, the fluxoids penetrating the superconductor in the mixed state experience a Lorentz force, perpendicular to the plane spanned by the transport current and the fluoxid. If the fluoxids are free to move, they begin to do so. This flux movement is associated with energy dissipation and leads to a finite resistance. Only if the fluxoids are pinned at traps can their movement be prevented or reduced. Traps include lattice defects and normally conducting inclusions. Type II superconductors with good flux pinning are said to be "hard". Conventional superconductors optimized in this way, such as Nb_3Sn or NbTi, achieve critical current densities (the critical current I_c divided by the conductor cross section) J_c of $10^9 - 10^{10}$ A/m^2 at $T = 4.2$ K.

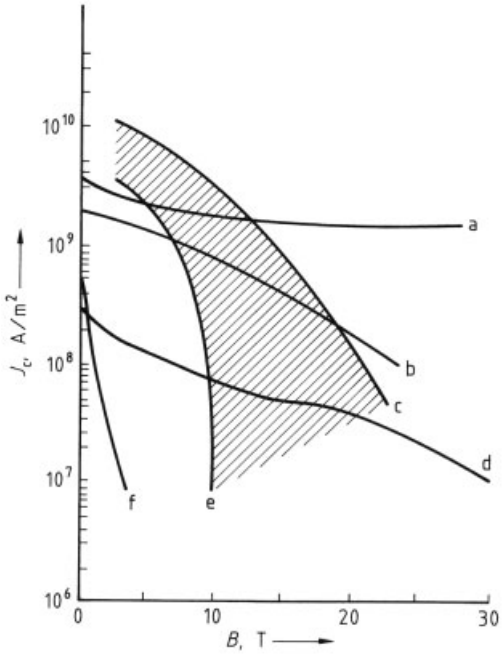

Figure 5. Critical current densities J_c (B) for selected superconductors
a) Bi-2212 tape at 4.2 K (B parallel to tape); b) $PbMo_6S_8$ wire at 4.2 K; c) Nb_3Sn at 4.2 K; d) Y123 specimen at 77 K (B parallel to CuO_2 planes) e) NbTi-50 at 4.2 K; f) Bi-2223 tape at 77 K (B parallel to tape);
All commercial helium-cooled high-current superconductors lie in the shaded area.

In liquid-N_2-cooled HT superconductors, flux movement (sometimes referred to as flux creep) is more prominent because of the higher operating temperature and the fundamentally poorer pinning properties of HT superconductors, which result in increased thermal activation of the fluxoids away from their traps. The phenomenon is especially marked for bulk HT superconductor specimens, where the J_c ($T = 77$ K) values attainable at present are only 5×10^8 to 2×10^9 A/m^2, depending on the material, despite the high B_{c2} values; a further drastic reduction occurs in an external field. In contrast, epitaxially grown HT superconductor films with thicknesses up to a few hundred nanometers exhibit J_c ($T = 77$ K) values up to almost 10^{11} A/m^2. The highly effective pinning mechanism in these thin layers has not yet been elucidated. Figure 5 shows the critical current density J_c as a function of external magnetic field for some conventional and high-temperature superconductors.

2.4. Energy Gap and Thermodynamic Properties

Absorption experiments performed with light or acoustic waves show the existence of an energy gap $2\,\Delta$ in the excitation spectrum of superconducting electrons. These electrons can absorb energy from the light or sound wave only when the quantum energy $h\nu$ associated with the frequency ν exceeds $2\,\Delta$ ($h\nu \geq 2\,\Delta$). The existence of this energy gap is due, in turn, to the quantum coherence of the superconducting state. It is roughly proportional to the critical temperature; according to the BCS theory (see Section 2.6)

$$2\,\Delta\,(T = 0) = 3.52 k_B T_c \tag{1}$$

where k_B is the Boltzmann constant.

The most precise way to measure the energy gap is to use the quantum-mechanical tunneling effect. Two superconducting electrodes are separated by a very thin (ca. 1 nm) insulating barrier, and the current – voltage characteristic is measured (Fig. 6, see next page). Because of their wave nature, single electrons can pass through the potential barrier even though the laws of classical mechanics forbid this. Accordingly, if a voltage U is imposed between the electrodes, a (tunneling) current I may in principle flow across the barrier. Provided the two electrodes are normally conducting, the result is a linear characteristic (ohmic behavior). If, however, they are in the superconducting state, the tunnel-

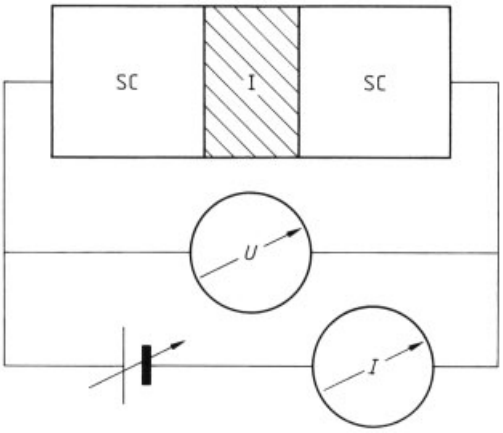

Figure 6. Circuit for measuring the I versus U characteristic of a tunnel diode
SC = Superconductor; I = Insulator

ing current can only flow when a tunneling electron extracts the minimum energy $2\,\Delta$ from the electric field (i.e., $eU \geq 2\,\Delta$). Rigorously, this holds only at $T = 0$. For $0 < T < T_c$, some normally conducting electrons (thermally dissociated Cooper pairs) are always present and contribute to the current even below the threshold voltage $U = 2\,\Delta/e$ and thus cause some smearing of the initially sharp threshold (see Fig. 7). If one of the electrodes is replaced by normally conducting material, a tunneling current can flow above half the threshold voltage ($U = \Delta/e$).

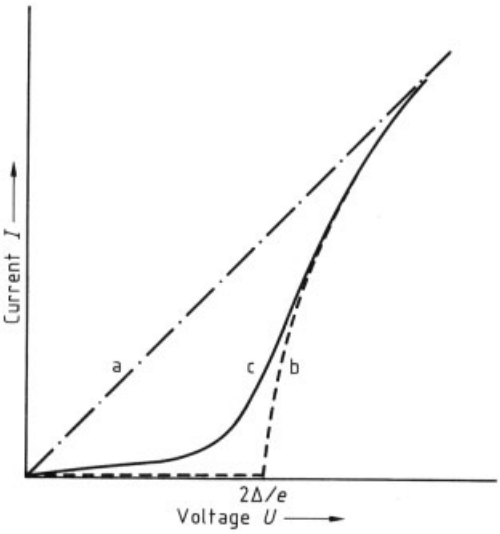

Figure 7. Qualitative curve of I versus U recorded with the circuit of Figure 6
a) Ohmic curve; b) Curve for $T = 0$; c) Curve for $0 < T < T_c$

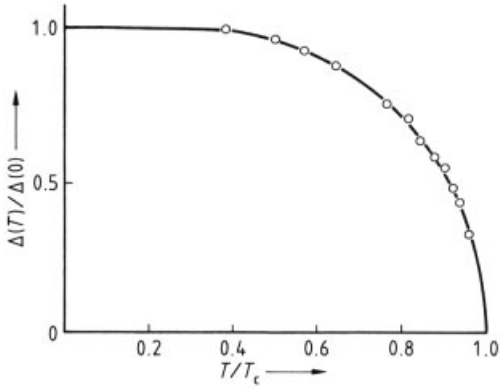

Figure 8. Temperature dependence of the energy gap for tantalum, determined by tunnel measurement

The transition of a material from the normally conducting to the superconducting state, as occurs on cooling with no applied magnetic field, is a second-order transition. No latent heat accompanies it, and the order parameter, which is the energy gap $2\,\Delta\,(T)$, vanishes steadily as $T \to T_c$. Figure 8 illustrates this behavior for the example of tantalum.

The specific heat $c_v\,(T)$, in contrast, changes abruptly at $T = T_c$. Below T_c, the electron contribution falls off exponentially as a direct consequence of the widening energy gap (Fig. 9). The jump in $c_v\,(T)$ results from the condensation of electrons into pairs at $T < T_c$ (see Section 2.6).

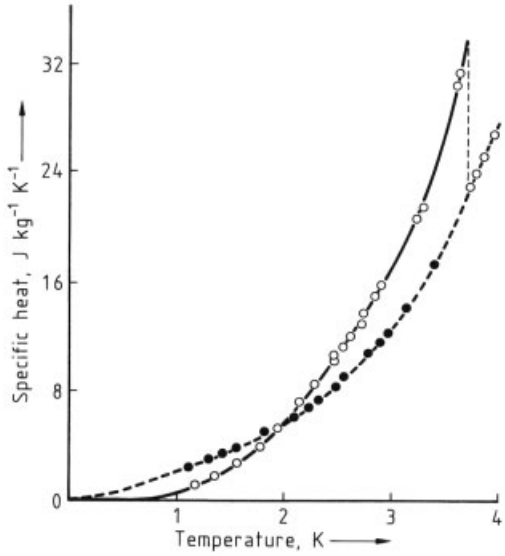

Figure 9. Specific heat $cv\,(T)$ for tin (open circles) with superconductivity suppressed by an applied magnetic field (solid circles)

The associated condensation energy, which is the difference in free energy between the normally conducting and superconducting states, is given by

$$\Delta F = F_{\mathrm{n}} - F_{\mathrm{s}} = V H_{\mathrm{c}}^2/8\pi \tag{2}$$

where B_{c} is the thermodynamic critical field defined in Section 2.2, V is the specimen volume and H_{c} is the critical magnetic flux density. For type I superconductors, ΔF can be obtained directly from B_{c}; calorimetric measurements are required for type II superconductors.

2.5. Josephson Effects

The $I(U)$ curve shown in Figure 7 describes only the current represented by the sum of tunneling currents of individual electrons. JOSEPHSON predicted in 1962 that electron pairs could tunnel through the barrier between two superconductors provided the junction is sufficiently thin. In other words, an additional supercurrent can flow across such a Josephson junction (with $U = 0$). This d.c. Josephson effect has since been confirmed experimentally.

The Josephson current is extremely sensitive to external magnetic fields, and the structure of this relationship is governed by the elementary fluxoid quantum Φ_0. Figure 10 shows the maximum Josephson current as a function of the external field for a tin – tin oxide – tin junction. The experimental curve resembles the intensity pattern for a wave diffracted at a slit. In fact, this phenomenon can be described in terms of the interference of two coherent quantum states in the two weakly coupled superconductors, with the magnetic field determining the phase shift between the two states. The strong magnetic field dependence of the Josephson current is employed in the design of extremely sensitive magnetic field sensors called SQUIDs (superconducting quantum interference devices; see Chap. 4).

If a voltage drop U is imposed across a Josephson junction, an additional alternating current occurs whose frequency is given by

$$\nu_{\mathrm{J}} = 2eU/h \tag{3}$$

For $U = 1$ mV, $\nu_{\mathrm{J}} = 4.85 \times 10^{11}\ \mathrm{s}^{-1}$. Because frequency can be measured with very high accuracy, this a.c. Josephson effect offers a way of constructing voltage standards by connecting many Josephson junctions in series.

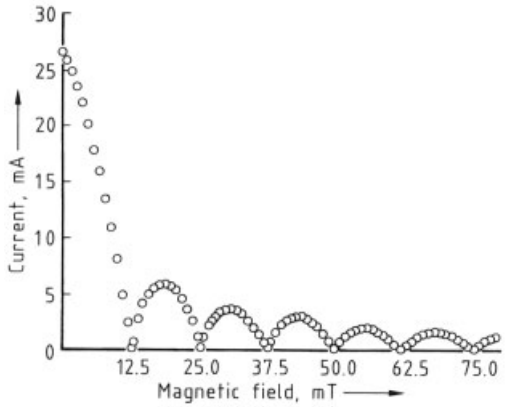

Figure 10. Critical Josephson current $I_{\mathrm{c}}(B)$ of aSn/SnO_x/Sn tunnel electrode plotted versus applied magnetic field

2.6. Theoretical Descriptions

In 1934, GORTER and CASIMIR devised a phenomenological theory that treats normally conducting and superconducting electrons as two fluids that interpenetrate without friction. This simple approach describes very well the temperature dependence of the condensation energy, and hence of the thermodynamic field. In 1935, F. and H. LONDON improved this two-fluid theory by postulating the London equation, which supplements Maxwellian electrodynamics by describing the Meissner – Ochsenfeld effect. In a nonlocal generalization of this theory, PIPPARD (1953) first introduced the notion of coherence (see below). In 1950, GINSBURG and LANDAU proposed an extension of the London theory that also permits local variation of the superconducting electron density; this concept was later stated in concrete terms by ABRIKOSOV and confirmed on a microscopic scale by GORKOV. Today the Ginsburg – Landau or GLAG theory is employed chiefly for the quantitative treatment of position-dependent phenomena in type II superconductors (e.g., calculating the spatial structure of a fluxoid).

In 1957, BARDEEN, COOPER, and SCHRIEFFER published their microscopic theory of superconductivity (BCS theory). The key assumption is pair formation: two electrons having opposite

momentum and spin form a Cooper pair. By the Pauli exclusion principle, these pairs are closely correlated with one another and condense into a coherent (macroscopic) quantum state with a single phase and a single amplitude, in which all the electrons have lost their individual degrees of freedom. This BCS state accounts for the Meissner – Ochsenfeld effect as well as the vanishing of the electrical resistance; it yields formulas for the temperature dependence of many parameters, such as energy gap $2\Delta(T)$, penetration depth $\lambda(T)$, and nuclear spin relaxation rate $\alpha(T)$, in quantitative agreement with experiment (see, for example, Fig. 8).

The prerequisite for pair condensation is the existence of an attractive interaction between the electrons. If this has strength V and the state density of electrons at the Fermi energy is given by $N(\varepsilon_F)$, the BCS theory gives the following expression for the transition temperature T_c of the superconductor:

$$T_c = 1.14\Theta \exp\left[-1/N(\varepsilon_F)\,V\right] \tag{4}$$

where $k_B\Theta = \omega_c$ is the energetic range of the interaction V. For the energy gap $2\Delta(T=0)$, Equation (1) gives a value of $3.52\,k_B T_c$.

Because electrons have like charges and therefore repel each other, the attraction required for pair condensation can arise only from a polarizable background medium. As early as 1950, Fröhlich proposed that the positively charged lattice of atomic residues could act as such a medium. Because in this model the electrons interact via lattice vibrations (phonons), the resulting coupling is referred to as electron – phonon coupling. The temperature scale Θ is then a characteristic phonon temperature, approximately equal to the Debye temperature Θ_D of the superconductor.

The BCS theory is rigorously valid only in the limiting case of weak coupling where $N(\varepsilon_F)\,V << 1$. Eliashberg formulated an extension of the BCS theory for strong electron – phonon coupling, which leads to material-specific corrections, mostly small, in the respective formulas. Numerous experiments have shown that virtually all conventional superconductors (see Chap. 3) can be described by the Eliashberg theory.

The first attempts to apply the Eliashberg theory to the high-temperature superconductors were unsuccessful. Several alternatives were proposed, involving either non-BCS states or a coupling mechanism other than conventional electron – phonon coupling. The controversy over the correct theoretical description of HT superconductivity continues, but there are clear experimental indications (flux quantization, Josephson effects) that the ground state of high-temperature superconductors is a coherent pair state [3].

The superconducting state is coherent only in a finite volume with a linear dimension, known as the coherence length ξ. For type I superconductors and $T = 0$, ξ is a few hundred nanometers; for conventional type II superconductors, a few tens of nanometers. High-temperature superconductors display strongly direction dependent ξ with values of 0.2 – 2 nm. These very small coherence lengths are a major drawback for industrial applications, because they result in high sensitivity to impurities and defects as well as strongly reduced trapping forces for fluxoids.

A thorough presentation of the theory for conventional superconductors can be found in [4], and a survey of theoretical models proposed for high-temperature superconductivity, in [5].

3. Classes of Superconductors

Many thousands of superconductors are known at present, including metallic elements, alloys, intermetallic compounds, and compounds with nonmetallic components. The rigorous classification of these materials is often difficult and to a degree arbitrary. The cuprate-based oxide superconductors discovered in 1986, however, can be delimited relatively well. Because many of them have $T_c > 77$ K, they are generally referred to as high-temperature superconductors (see Section 3.3).

Another relatively well defined class comprises organic superconductors, which were discovered in 1980 and by the early 1990s T_c values up to ca. 13 K had been attained (see Chap. 6).

By far the largest class is composed of the conventional or classical superconductors (Section 3.1). It includes all superconducting elements and alloys, as well as many intermetallic compounds and metal – nonmetal compounds such as hydrides, sulfides, oxides, and

nitrides. The maximum T_c for this group is 23 K (Nb_3Ge), but most members of the class have $T_c < 10$ K, so that helium cooling is required.

The term "exotic" superconductors (Section 3.2) includes all superconducting materials that cannot be easily assigned to the conventional, organic, and high-temperature superconductors.

While the coupling mechanisms responsible for superconductivity in most of these exotic compounds are as unclear as those involved in high-temperature superconductivity, electron–phonon coupling can be assumed for classical superconductors almost without exception.

3.1. Classical Superconductors

Elements. Twenty-nine pure metals are known to be superconductive at normal pressure (Table 1). A number of other elements exhibit superconductivity only in high-pressure phases; these in clude some metals (Ba, Ge, Sc) but also elements such as P and Se that are nonmetallic at normal pressure and enter metallic phases only at high pressure. It is possible that still further metallic elements (alkali metals or noble metals, for example) become superconductive under normal pressure if they are cooled sufficiently. Superconductivity in rhodium, for instance, was not discovered until 1982 when the critical temperature of $T_c \approx 325\,\mu$K was attained by using a complicated cryogenic technique [6].

It is, however, certain that not every pure metal becomes superconducting at a sufficiently low temperature. The main counterexamples are the ferromagnetic elements Fe, Co, and Ni. No coexistence has ever been observed between superconductivity and inherent ferromagnetism, and this is in fact a theoretical impossibility for a BCS superconductor. While the BCS ground state requires the spins in the electron pair to be

Table 1. Metals superconducting at normal pressure

Element	T_c, K	Crystal structure ∗	Melting point, °C	Debye temperature, K
Al	1.19	fcc	660	420
Be	0.026	hex	1283	1160
Cd	0.55	hex	321	300
Ga	1.09	orth	29.8	317
Hf	0.13	hex	2220	252
Hg	4.15	rhom	− 38.9	90
	(3.95)	tetr		
In	3.40	tetr	156	109
Ir	0.14	fcc	2450	420
La	4.8	hex	900	140
	(5.9)	fcc		
Mo	0.92	bcc	2620	460
Nb	9.2	bcc	2500	240
Np	0.075	orth	637	259
Os	0.65	hex	2700	500
Pa	1.3		1230	185
Pb	7.2	fcc	327	96
Re	1.7	hex	3180	430
Rh	0.00033	fcc	1966	512
Ru	0.5	hex	2500	600
Sn	3.72	tetr	231.9	195
	(5.3)	tetr		
Ta	4.39	bcc	3000	260
Tc	7.8	hex	2140	351
Th	1.37	fcc	1695	170
Ti	0.39	hex	1670	426
Tl	2.39	hex	303	88
U(α)	0.2	orth	1132	200
V	5.3	bcc	1730	340
W	0.012	bcc	3380	390
Zn	0.9	hex	419	310
Zr	0.55	hex	1855	290

∗ bcc = body-centered cubic; fcc = face-centered cubic; hex = hexagonal; orth = orthorhombic, rhom = rhombic; tetr = tetragonal.

Table 2. Properties of important classical superconductors

System	T_c, K	H_{c2} ($T=0$), T	Uses
Nb	9.2	<1	microelectronics, SQUIDS, RF cavities
NbN	ca. 16	44	microelectronics, coating of fibers
NbTi50	ca. 10	14	alround applications in multifilament wires
Nb_3Sn	ca. 18	25	high-field magnets
$PbMo_6S_8$	ca. 15	60	extremely high-field magnets

antiparallel (see Section 2.6), ferromagnetism is based on parallel spin of conduction electrons, so that ferromagnetism breaks up pairs and interferes with superconductivity. Accordingly, superconductivity is excluded for the ferromagnetic elements, Fe, Co, and Ni, even at absolute zero.

Industrial Superconductors. Only in a few cases are pure metals suitable for applications; examples are Nb in SQUID technology and Pb in heat switches. Conventional industrial superconductors are generally alloys or stoichiometrically sharply defined compounds; all are of type II. For specific technical requirements, they must be optimized in the following respects:

1) High critical temperature: $T_c \geq 10$ K for stable cooling with liquid He^4 (4.2 K)
2) High critical current density: J_c (4.2 K) $\geq 10^8 - 10^{10}$ A/m^2, especially for applications in power engineering and microelectronics
3) High critical magnetic field: B_{c2} (4.2 K) $\geq 10 - 20$ T for use in magnets
4) Good mechanical properties: ductility, elasticity, and tensile strength

Of the four points, the last is most critical. Superconducting materials display a tendency to become more brittle the higher their transition temperature, and this is also true of high-temperature superconductors. The processing of industrial superconductors thus entails complicated techniques.

Apart from niobium, the most important conventional industrial superconductors are the alloy NbTi50 and the A15 compound Nb_3Sn. In addition, the Chevrel phase $PbMo_6S_8$ and the hard material NbN have some application potential for the future. Table 2 lists important data for these materials. NbTi50 stands out by virtue of its ductility; at present it is the most widely used industrial superconductor.

3.2. Exotic Superconductors

Heavy-Fermion Superconductors. Superconductivity below $T_c \approx 0.6$ K in the compound $CeCu_2Si_2$, discovered in 1978, is due to the condensation of very heavy electrons with effective masses around 300 times the free electron mass. Similar behavior is observed in some uranium compounds, such as UBe_{13} ($T_c \approx 0.9$ K), UPt_3 ($T_c \approx 0.45$ K), and UAl_2Pd_3 ($T_c \approx 2$ K) [7]. Although technically unpromising, these systems are of major interest for modern solid-state physics. Both the symmetry of the superconducting ground state and the effective pair-formation mechanisms remain largely unclear.

Polymeric Sulfur Nitride. The first superconducting polymer, $(SN)_x$, was discovered in 1973. It consists of long chains collected into bundles. The gold-colored, lustrous material is metallically conductive along the chains and becomes superconducting at $T_c = 0.26$ K. It exhibits highly anisotropic behavior in a magnetic field. $(SN)_x$ is practically a one-dimensional conductor [8].

Fullerenes. In 1991 superconductivity was discovered in alkali-metal-doped fullerene crystals of composition A_3C_{60} (A = K, Rb, Cs) [9]. Figure 11 shows the relatively simple cubic structure of these crystals. The compounds have high transition temperatures (18 K for K_3C_{60}, 28 K for Rb_3C_{60}, 33 K for Cs_2RbC_{60}). Undoped C_{60} crystals are insulators, and A_3C_{60} exhibits metallic behavior and hence superconductivity only when the alkali metal atoms donate valence electrons to the narrow conduction band, which then becomes half-filled. The high T_c values can readily be accounted for within the BCS theory on the basis of electron – phonon coupling. They result from a high Debye temperature (due to the small size of C atoms) combined with a relatively high electron-state

density (narrow conduction band). Alkali-metal-doped fullerenes react vigorously with oxygen and moisture and so cannot be employed in industry.

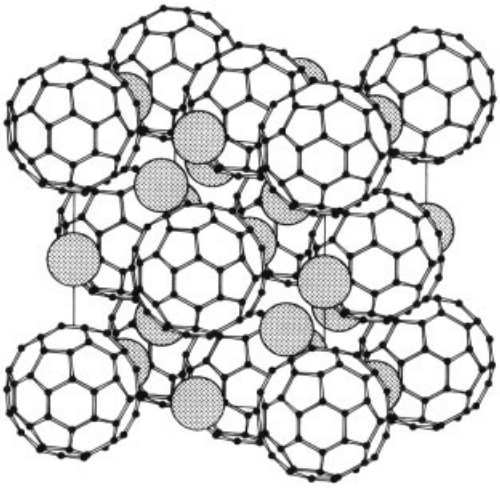

Figure 11. Crystal structure of the superconducting fullerene compound K_3C_{60} (large solid circles represent K atoms)

Because of their close relationship to alkali-metal-doped graphites, which can also become superconductive, it does not appear appropriate to classify these materials with the organic superconductors.

Doped Oxide Semiconductors. Some oxide semiconductors exhibit superconductivity after they have been rendered metallic by doping. These compounds include $SrTiO_3$ and $BaBiO_3$, both having perovskite structures, as well as $Li_{1+x}Ti_{2-x}O_4$ with a spinel structure. The highest T_c values found up to now in these three classes are ca. 0.7 K ($SrTi_{1-x}Nb_xO_3$, $x \approx 0.01$) [10], ca. 12 K ($LiTi_2O_4$) [11], and ≈ 30 K ($Ba_{0.7}K_{0.3}BiO_3$) [12]. All these superconductors have very low concentrations of free charge carriers (ca. 10^{18} to $10^{20}\,cm^{-3}$), sharing this property with high-temperature superconductors as they do a metal – insulator transition and the perovskite-like structure. Especially in view of the very high transition temperature of $BaKBiO_3$, a relationship to high-temperature superconductors is thus a plausible supposition. However, the two-dimensional structures (CuO_2 planes) typical of HT superconductors are absent, as is magnetic order in the undoped, insulating substances.

3.3. High-Temperature Superconductors

Structure. Structurally, all high-temperature superconductors are perovskite derivatives. Important building blocks are CuO_6 octahedra, which may degenerate to CuO_5 pyramids or CuO_4 squares as a result of oxygen deficiency. Linked together at the corners, these units form CuO_2 planes, which are either isolated from one another by (oxide) planes of other elements (single-layer structures) or arranged in packages of two or three layers. It is certain that these CuO_2 planes are crucial determinants of the physical properties of the materials, and hence of their superconductivity. All other components serve to stabilize the structure or provide a charge reservoir. Figure 12 illustrates the three-layer structure of Tl-2223 ($T_c \approx 125$ K).

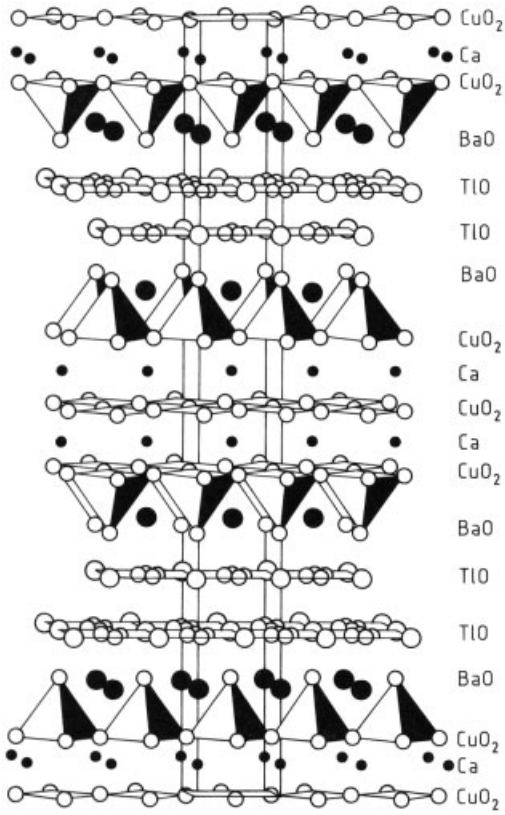

Figure 12. Crystal structure of the high-temperature superconductor Tl-2223 ($T_c \approx 125$ K)

(*x*, *T*) Phase Diagram. Each HT superconductor has a reference compound that is insulating and antiferromagnetic and from which it is derived by *n*- or *p*-type doping. Examples are $La_{1.85}Sr_{0.15}CuO_4$ and $YBa_2Cu_3O_7$, which are obtained from the compounds La_2CuO_4 and $YBa_2Cu_3O_6$ by *p*-type doping, and the *n*-doped HT superconductor $Nd_{1.85}Ce_{0.15}CuO_4$ ($T_c \approx 20$ K), which can be derived from Nd_2CuO_4. All known superconductors with $T_c > 40$ K are *p*-doped.

The effect of doping on the physical properties of a high-temperature superconductor is illustrated by the (x, T) phase diagram of $La_{2-x}Sr_xCuO_4$ (Fig. 13). Qualitatively, this plot also applies to other HT superconductors. Undoped La_2CuO_4 is an insulator and exhibits antiferromagnetism, which decays rapidly as x increases; there is an insulator – metal transition at $x \approx 0.04$, above which the system becomes paramagnetic and superconducting. T_c is a maximum for $x \approx 0.15$, falling off again from that point on and vanishing beyond $x \approx 0.3$.

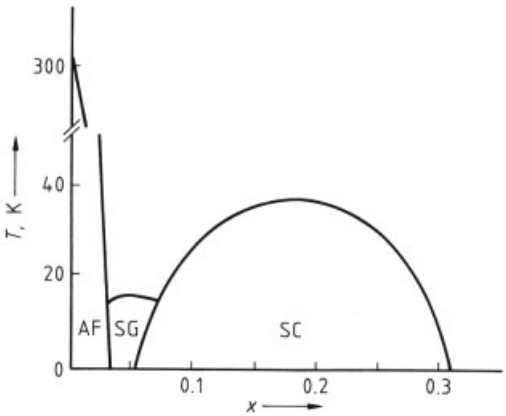

Figure 13. (x, T) Phase diagram of $La_{2-x}Sr_xCuO_4$
AF = Antiferromagnetic phase; SG = Spin-glass behavior; S = Superconductivity

High-Temperature Superconductors with $T_c > 77$ K. The most important families of superconductors with $T_c > 77$ K (nitrogen cooling) are listed in Table 3. A more comprehensive list can be found in [13].

Among these compounds, Y123, Bi-2212 and Bi-2223, and Tl-2223 are most significant for industrial applications at present.

Table 3. High-temperature superconductors with $T_c > 77$ K

Formula	Short hand	Maximum T_c
$REBa_2Cu_3O_7$*	REBCO or 123	92 K (YBCO)
$Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ (+ Pb doping)	BSCCO or Bi-22(n−1)n	90 K (Bi-2212) 122 K (Bi-2223) 90 K (Bi-2234)
$Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$	TBCCO or Tl-22(n−1)n	110 K (Tl-2212) 127 K (Tl-2223) 119 K (Tl-2234)
$Tl_1A_2Ca_{n-1}Cu_nO_{2n+3}$ (A = Sr, Ba)	Tl-12(n−1)n	90 K (Tl-1212) 122 K (Tl-1223) 122 K (Tl-1234) 110 K (Tl-1245)

* RE = Rare earth.

HT superconductor specimens prepared as ceramics (e.g., sintered from mixed oxides) have critical current densities J_c ($T = 77$ K) $< 10^7$ A/m^2 and are thus unsuitable for most applications. The low J_c values result from grain boundaries, which can be eliminated or reduced by special fabrication methods, so that J_c can be substantially increased. Two methods are generally used:

1) Epitaxial films: this method is based on the epitaxial growth of a high-temperature superconductor film on a single-crystal substrate having an appropriate crystal structure (e.g., $SrTiO_3$ or Al_2O_3). Frequently used deposition methods are laser ablation, sputtering, and MOCVD (Metal Organic Chemical Vapor Deposition). High current densities are obtained for epitaxial films with the CuO_2 planes parallel to the substrate surface. This technique is used chiefly in the fabrication of Y123 and Tl-2223 films for use in microwave electronics and sensor systems. Critical current densities up to ca. 7×10^{10} A/m^2 are observed in Y123 films at $T = 77$ K; the film thicknesses are some hundreds of nanometers [5].
2) Textured bulk materials: preferred orientation of the crystallites (texture) and the consequent reduction in grain boundaries are essential for achieving high current densities in bulk HT superconductor materials. For Y123, texture is produced by cooling the melt in a temperature gradient and/or in a strong magnetic field (melt texturing). In this way, it is possible to obtain bars or disks weighing up to hundreds of grams and offering

good flux pinning and J_c ($T = 77$ K) values up to ca. 10^9 A/m^2; the potential for these materials lies primarily in the area of passive magnetic bearings [5]. The main materials used in making wires and tapes are Bi-2212 and Bi-2223, which is cladded in silver and subjected to a complicated sequence of rolling and pressing steps as well as tempering stages. Texturing in these cases is apparently promoted by mechanical working. Bi-2212 tapes made in this way are suitable for use only with helium cooling, but they exhibit outstanding current-carrying capacities even in very strong magnetic fields: J_c ($T = 4.2$ K, $B_a = 25$ T) $\approx 2 \times 10^9$ A/m^2. Bi-2223 tapes, in contrast, are better suited to use with LN_2 cooling: J_c ($T = 77$ K) $\approx 5 \times 10^8$ A/m^2 at zero field. An applied external field does, however, lead to a drastic reduction in J_c, which can be attributed to inadequate flux pinning in this material (flux creep; see Section 2.3) [5].

4. Electronic Applications of Superconductivity

Superconducting electronics [14–17] have already become an indispensable resource for physical measurement and detection, for ultrafast "super" logic chips, for many high-frequency applications, and for certain calibration purposes. These circuits are used where very low noise, very high sensitivity, minimal losses, short switching times, and stability of voltage calibration are crucial. Advantages over competing semiconductor devices include the macroscopic quantum effects of superconductivity (Josephson effects), unparalleled low noise levels, and novel effects such as quantum interference and voltage-frequency conversion governed solely by physical constants. In such applications, superconducting components are employed in the most critical areas of physical instrumentation, ultrafast logic circuits, and high-frequency engineering, and also find use as voltage standards. Because cooling is always desirable when minimal noise is desired, superconducting electronics represents the technology of choice under extreme requirements, particularly in research and in military hardware. Helium-cooled SQUID systems have become firmly established in laboratories, where they serve as extrasensitive magnetometers and susceptometers and as instruments for measuring derived quantities.

The need for cooling to temperatures below T_c was long considered a major drawback and prevented the adoption of superconducting electronics in conventional hardware. Even where superconducting devices surpassed all competition (or where there was no competition in the first place), as in biomagnetic medical diagnostics using SQUID arrays, the cost of cooling and shielding has presented a major obstacle to the commercial success of superconducting electronics.

Important milestones in the commercialization of superconducting electronics include the transition to competitive electronics based on high-temperature superconductors and the rapid progress made recently in the development of cooling equipment for electronics, especially the pulse-tube cooler.

4.1. Superconductivity Effects Important for Electronic Applications

4.1.1. Pure Inductances

In a superconductor d.c. current transport remains virtually lossless up to frequencies in the megahertz range (see Section 2.1) [losses comparable to copper are observed only at 100 GHz in niobium at 4 K or in YBCO (yttrium barium copper oxide) at 77 K], provided two conditions are met:

1) The magnetic field amplitudes must be sufficiently small that no flux penetrates into the superconductor
2) Flux "frozen in" by the cooling process must not be able to migrate (i.e., it must be pinned)

These requirements can be satisfied in electronic applications. Two important applications of the effect are therefore possible: shielding of (cryo)electronic circuits and lossless "transformation" of magnetic flux from zero frequency up into the megahertz range. In the design of these applications, the superconducting structure can be treated as a simple equivalent circuit composed of ideal inductances and capacitances with

no resistive components. If the capacitances are negligible, the integrated Faraday induction law

$$n_1 \Delta\Phi_1 = -I \Sigma L_i$$

describes a "flux transformer" comprising a closed series circuit of inductances L_i; when the flux in L_1, having n_1 turns, changes by $\Delta\Phi_1$, a current I is induced in the circuit. This current I acts to oppose the flux in L_1, and if the L_i have been properly selected it will displace flux from L_1 to the L_i as desired. The L_i can, of course, be any extended structures, not just coils.

Within the stated constraints, this flux transformation is independent of frequency. The effect has proved extremely useful for transmitting magnetic flux from a field pickup to a magnetic flux sensor. It is also important in the construction of flux concentrators/transformers for SQUID applications ranging from simple susceptometers to gravity-wave detectors and monopole detectors [14, pp. 178 – 190], [17, pp. 388 – 404].

If there is no inductive load ($L_i = 0$ for $i > 1$), L_1 provides ideal shielding of external flux changes. This principle can be employed for creating volumes practically free of magnetic fields by "inflating" an initially folded superconducting pouch [17, pp. 354 – 359].

4.1.2. Small High-Frequency Losses

In superconducting materials, electrical losses vanish if the current densities $J < J_c$ are constant over time. The losses at small high-frequency power increase as a function of f^2 but still remain very low over a wide frequency range. For the best epitaxial YBCO films, the surface resistance R_s (77 K) at 100 GHz is always an order of magnitude smaller than in cooled copper. Figure 14 illustrates this point, comparing $R_s(f)$ for high-quality YBCO films with the surface resistance of copper. At a given frequency, losses increase rapidly above a high-frequency power threshold, corresponding to a magnetic field amplitude of ca. 0.03 T, beyond which flux penetrates into the superconductor (see Section 2.2). Signal propagation in superconductors is nearly dispersionless below $f/\text{GHz} = 73\, T_c/\text{K}$, so that this effect can be utilized for high-rate information transmission.

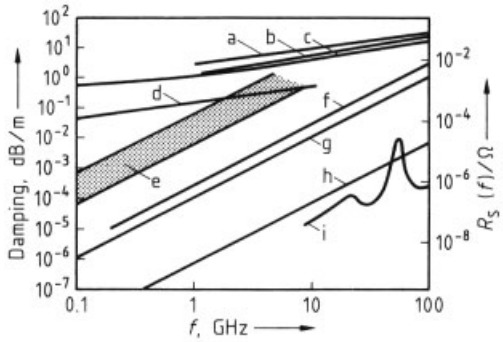

Figure 14. Measured surface resistance together with calculated and measured strip-line losses [28] for YBCO and copper at 77 K (dielectric losses not included)
a) Copper strip line at 300 K; b) Copper strip line at 70 K; c) Copper; d) RG-9 coaxial cable; e) Copper strip line at 77 K (measured); f) Epitaxial YBCO; g) YBCO strip line at 77 K (theoretical); h) YBCO strip line at 27 K (theoretical); i) Atmospheric propagation at sea level

Superconducting high-frequency resonators (cavities) and low-loss delay lines and filters for frequencies up to 100 GHz used to be important only in military hardware (special radar techniques). High-temperature superconductors are now permitting use in civilian areas: remote sensing of the Earth, mobile radio, and civilian radar (see Section 4.3.4).

4.1.3. Energy-Gap Effects

Superconductor – insulator – superconductor (SIS) tunnel junctions are used for extremely low-noise direct ("video") or heterodyne detection of millimeter radiation in radio astronomy (see Section 2.4 and Figs. 6 and 7). These devices are operated at temperatures well below T_c. Josephson effects are purposely suppressed by a static magnetic field. The devices utilize two effects: first, the extreme nonlinearity of the current – voltage characteristic for SIS junctions made with low-temperature superconductors (abrupt change in tunneling current at the total energy gap); second, the quantum effect of photon-assisted tunneling. With helium-cooled Pb/Bi/In oxide – Pb/Bi tunnel junctions, the quantum limit (cutoff) of the mixer noise temperature $T_M = \hbar\omega/2k_B$ (where $\hbar = 1.0546 \times 10^{-34}$ Js) has been reached for detection in the 100 GHz range; this makes the technique indispensable in radio astronomy [14, pp. 259 – 284]. The state of the art is an integrated submillimeter low-temperature superconductor detector

circuit like the one built at Chalmers University of Technology in Göteborg, Sweden [18]. It does not appear desirable to replace low-temperature with high-temperature superconductors in this case, since as yet no SIS tunnel junction has been devised on the basis of HT superconductors.

4.1.4. Quantum Interference Effects

The coherence property of the superconducting state is discussed in Sections 2.5 and 2.6. Because Cooper pairs "overlap" within the coherence length and in this way the coherence property is propagated, macroscopic structures of superconductors exhibit macroscopic interference phenomena (Section 2.5, Fig. 10). The magnetic flux enclosed by a macroscopic (e.g., on the scale of millimeters) superconducting ring is quantized; this can be understood as the constructive interference of Cooper-pair waves with a multiple of 2π phase difference along the ring, so that the magnetic flux enclosed by the superconducting ring is a corresponding multiple of the magnetic fluxoid quantum $\Phi_0 = h/2\varrho = 2.0679 \times 10^{-15}\,\mathrm{Tm}^2$. The Josephson junctions discussed below make it possible to introduce magnetic flux into a superconducting ring in continuous fashion and to utilize the macroscopic quantum interference in the ring for measurement purposes (the basis of SQUID devices). Magnetic-field-controlled quantum interference in a Josephson junction is discussed above (Section 2.5, Fig. 10). In accordance with the ratio of areas, external magnetic fields act much more strongly on quantum interference in the macroscopic ring (area $> 1000\,\mu\mathrm{m}^2$) than on quantum interference in a single Josephson junction enclosing less than $1\,\mu\mathrm{m}^2$ of the field. For applications with magnetic control, Josephson junctions optimized with respect to noise level and switching time are purposely made small ("point" contacts) and are linked with well-defined superconducting interferometer ring structures.

4.2. Josephson Junctions, Tunnel Junctions, and Weak Links

4.2.1. Junction Types and Their Significance

The Josephson junction is the essential nonlinear component in superconducting electronics. Figure 15 illustrates the basic current – voltage (I/U) characteristics of Josephson junctions. Such an element can be modeled by a parallel network of a Josephson junction, an ohmic resistance or SIS tunnel characteristic (Section 2.4, Fig. 7), and a capacitance, where the I and U

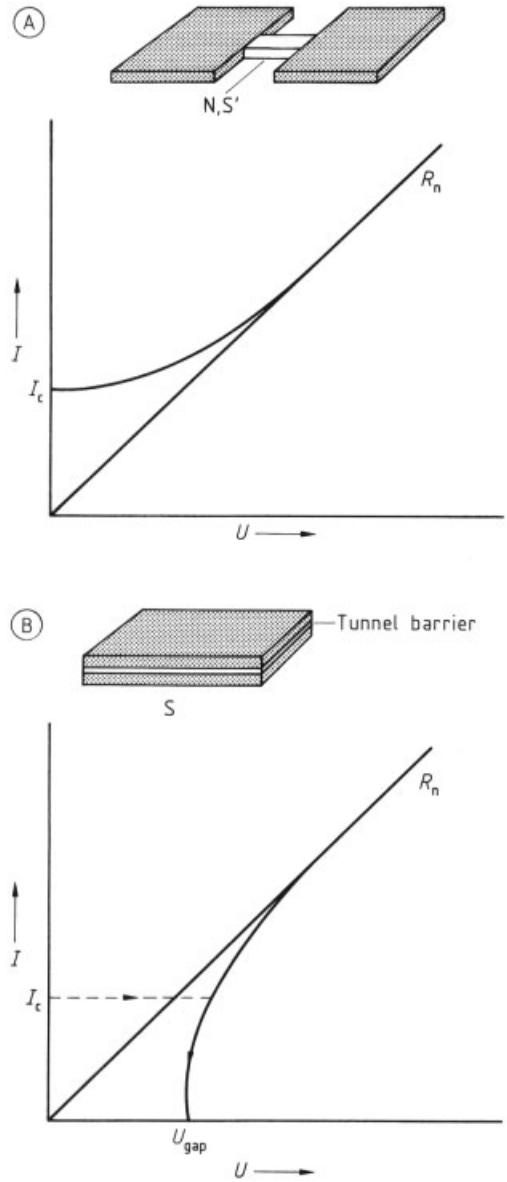

Figure 15. Ideal current – voltage characteristics of the Josephson junction
A) Steady behavior; B) Switching behavior
N = Normal conductor ; S′, S = Superconductor

values of the Josephson junction are related to each other and to time by the Josephson equation

$$I = I_c \sin(\delta_0 + 2eUt/h)$$

where I_c is the critical current, δ_0 the phase offset related to the "frozen in" flux, and h is Planck's constant.

The characteristic of Figure 15 A is a time average and is steady; that of Figure 15 B exhibits switching behavior. The transition between Josephson current and quasiparticle current, which can take place on a picosecond time scale, long formed the basis for all major logic applications of superconductivity, with the Josephson threshold-value switch as active element and with I_c as threshold value. The "weak-link" characteristic (Fig. 15 A) is well-suited to sensor (SQUID) applications and new logic circuits of RSFQ type. Characteristic 15 A can (but need not) be derived from 15 B by bridging with a resistance; 15 A simply presupposes a weak coupling between the superconductors S, not the existence of a tunnel junction. Characteristic 15 A may arise if the two superconductors S are connected by a normally conducting bridge or a bridge made up of a "weakened" superconductor. For quantum coherence to be preserved between the superconductors S, the bridges must have dimensions comparable with the superconducting coherence length in the bridge material. For well-characterized tunnel junctions, the noise properties can be theoretically described in an acceptable manner (to within a factor of two to three); for bridges, deviations of up to two orders of magnitude can occur, so that large-scale integrated circuits have not been feasible up to now [19]. So far, Josephson junctions made of high-temperature superconducting materials, which can be operated successfully at elevated temperature (ca. 70 K), have all been of type 15 A.

4.2.2. Josephson Circuits, Digital Circuits, Digital Signal Processing, and Voltage Standards

Circuits containing Josephson junctions and superconducting current paths forming closed structures are interferometers of the SQUID type and are discussed in Section 4.2.3. As current-driven switches, they have found some use in logic circuits and A/D converters as well. Important switching elements for Josephson circuits, besides the Josephson junctions proper, are thin-film resistors, microstrip lines (select lines, bias lines), and connections formed by superconducting current paths on a substrate coated with a dielectric film. Various types of latching logic gates, ranging from a relatively large-area interferometer to the very compact gates in four-junction logic, and a variety of gate families and memory cells have been devised; these have been developed to LSI (large-scale integrated) complexity, particularly in Japan [15, pp. 1 – 20], [16, pp. 101 – 134]. The crowning achievement in this line is a functioning four-bit microprocessor running at 770 MHz. The state of the art comprises superconducting structures of niobium or NbN with tunnel barriers of aluminum oxide. The mean power consumption of a logic gate in this technology is ca. 10 μW at 4 K. Despite advanced niobium technology, there is as yet no satisfactory solution for superconducting fast-access mass storage, and indeed no solution is in prospect. This defect might be remedied with a superconducting shift register as working memory, with hybrid combinations employing cooled CMOS (complementary metal oxide Silicon) semiconductor mass memory; such an approach is being pursued in Japan and the United States. A more ambitious and promising technique is being explored in the United States: the nonlatching rapid single flux quantum (RSFQ) logic family employing niobium superconductors [16, pp. 1 – 49]. Processor frequencies well over 100 GHz – surpassing the semiconductor competition by more than two orders of magnitude – appear feasible. Special processors, especially for ultrafast digital signal processing, are of interest here. In this logic family, the bit is the presence of a switching pulse lasting a few picoseconds in a time window corresponding to a 2π phase change in the Josephson relation. Josephson junctions of the type illustrated in Figure 15 A are used. The functionality of the ultrafast RSFQ basic circuits has already been demonstrated with niobium. One sensational result is the demonstration that a 32-bit RSFQ shift register at 77 K can be operated at a frequency of over 100 GHz [19]. The experiment involved HT-superconductor nanobridges, which, however, exhibit an unacceptable error rate (one wrong bit every 10 min), but the convertibility

of RSFQ circuits to the still-unperfected HT superconducting technology has now been proved feasible (at least for fault-tolerant circuits).

The basis of the Josephson voltage standard is that a Josephson junction irradiated with high-frequency radiation (frequency f) exhibits current steps at multiples of the voltage $U = h/2\,ef$. Modern Josephson voltage standards for 1 – 10 V consist of thousands of hysteretic niobium – aluminum oxide – niobium Josephson junctions connected in series, which are integrated into a superconducting microstrip conductor in such a way that all junctions are excited at the same high frequency (ca. 70 GHz) [15, pp. 21 – 36]. It appears that a new design will make it possible to lower the operating frequency to ca. 12 GHz [20]. In this way, the costs of the HF system can be cut by about an order of magnitude, which is conducive to adoption in industry.

4.2.3. SQUIDs

Superconducting quantum interference devices (SQUIDs) are extremely sensitive detectors for magnetic flux and magnetic fields. The SQUID exploits two fundamental effects of superconductivity: the Josephson effect and the single-valuedness of the quantum phase in a superconducting ring. The SQUID converts magnetic flux to electrical voltage. At very low temperatures, the intrinsic lower limit of energy resolution of a SQUID per hertz of bandwidth is given by the quantum noise, $\hbar \approx 10^{-34}$ J/Hz, as has been confirmed in laboratory specimens. For practical devices with efficient flux coupling, good helium-cooled SQUIDs achieve energy resolutions of around 30 $\hbar$ and field resolutions of ca. 10^{-15} T/Hz$^{1/2}$ at frequencies down to < 1 Hz. This high sensitivity has made SQUIDs vital for many important experiments and makes possible the measurement of extremely weak biomagnetic signals from human organs, such as the heart and brain, in the range of $10^2 - 10^5$ f T/Hz$^{1/2}$ (corresponding to 10^{-8} to 10^{-5} times the Earth's magnetic field). Measurement of biomagnetism can be regarded as the pioneering application of superconductivity in normal technology outside the physics laboratory.

The principles and technical design of SQUIDs have been presented in detail in textbooks and survey articles [21]. Here it is sufficient to illustrate the two most important versions (Fig. 16 A and B), the d.c. SQUID and the RF SQUID. Each utilizes Josephson junctions with hysteresis-free $I\,(U)$ characteristics, corresponding to Figure 15 A with parallel resistance R. In Figure 16 A, if magnetic flux Φ enters the superconducting ring (with inductance L_s) through the two identical Josephson junctions, the current – voltage characteristic of the ring will vary between two extreme positions (Fig. 16 C) corresponding to fluxes $n\Phi_0$ and $(n + 1/2)\Phi_0$, where $\Phi_0 = 2.0679 \times 10^{-15}$ Wb is the magnetic fluxoid quantum and n is an integer. If a fixed bias current $I_b > 2\,I_c$ is passed through the ring (where I_c is the critical current of the individual Josephson junction), the voltage drop U across the ring varies periodically in the flux with period Φ_0 (Fig. 16 E). The (a.c.) current – voltage characteristic of a resonant circuit made of an inductance L and a capac-

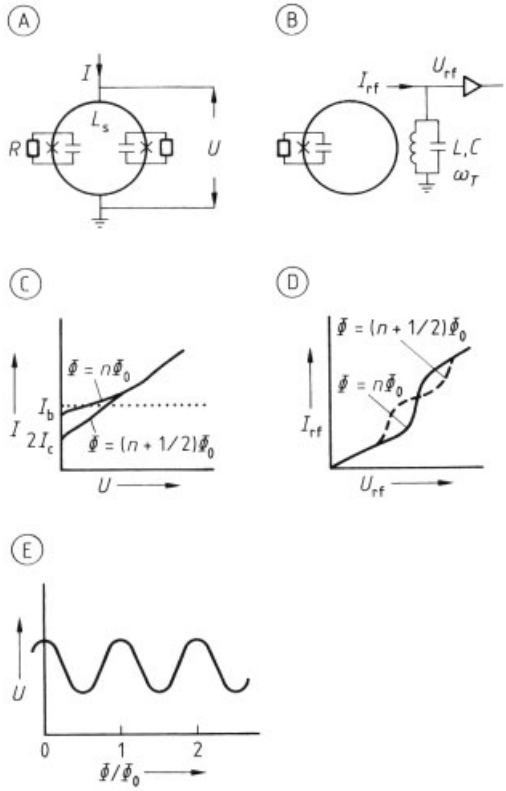

Figure 16. The two most important SQUID configurations
A) d.c. SQUID; B) RF SQUID; C), D) Extreme positions of the I versus U curve on variation of the magnetic flux; E) At a current operating point I_B, the voltage varies with period Φ_0

itance C, which is inductively coupled to a superconducting ring with a single Josephson junction, similarly varies with resonance frequency ω_T between two extreme positions, again corresponding to flux values of $n\Phi_0$ and $(n+1/2)\Phi_0$ in the ring (Fig. 16 D). If the amplitude of the high-frequency current is set equal to a current step of the I_{rf} versus U_{rf} curve and the flux Φ in the ring is varied, then U_{rf} also varies periodically with period Φ_0.

The readout electronics of a sensitive SQUID magnetometer operates as a null detector by strong reverse feedback with the aid of a compensating coil, which compensates a change in external flux Φ with a flux resolution down to ca. $10^{-6}\,\Phi_0/\mathrm{Hz}^{1/2}$. Important factors in the optimal functioning of a SQUID with characteristics 16 C and 16 D are the selection of parameters such that $2L_sI_c \approx \Phi_0$ (quantization condition) and $I_c \sim k_BT/\Phi_0$ is large enough that thermal flux noise at the SQUID operating temperature is not by itself of the order of $\Phi_0/2$ (maximum driving of SQUID). This constraint limits SQUID inductances to less than ca. 10^{-9} H at 4 K, and less than ca. 10^{-10} H at 77 K. As a result, stringent standards apply to the microfabrication of HT-superconductor SQUIDs to concentrate magnetic flux in appropriately small SQUID rings a few micrometers in diameter. The flux concentration is achieved by means of flux transformers (see Section 4.1.1), which are fabricated as thin-film structures (Fig. 17 A).

Flux concentrators without thin-film coils, in the form of slotted single-layer film structures, are less effective as transformers but are easier to manufacture.

Given the great sensitivity of SQUIDs, interference suppression is a key problem. The first thing is to maintain the functioning of the SQUID that is experiencing interference, then to isolate the signal for measurement from the noise. In the presence of external fields that vary only slightly with position, very small local field gradients can be isolated and measured if the pickup coil is made in the form of a gradiometer of first or higher order (Fig. 17 A – C). Strong interference fields are generally compensated with passive and active shielding (magnetically shielded compartments, large coils). The more open the system, the sooner it will run into dynamic limiting. If this limiting is not overcome, electronic and "software" gradiometers may be used. In these approaches, magnetometer signals are corrected in either analog or digital mode. The problem of suppressing interference is identical for SQUIDs based on low- and high-temperature superconductors. The problem with LT-superconductor SQUIDs is not so much their field sensitivity ($\approx$1 f T/Hz$^{1/2}$ at $\geq$ 1 Hz) as the performance and the cost/benefit ratio of the system. A SQUID magnetometer optimized in this respect is shown in Figure 18 [22]. This repre-

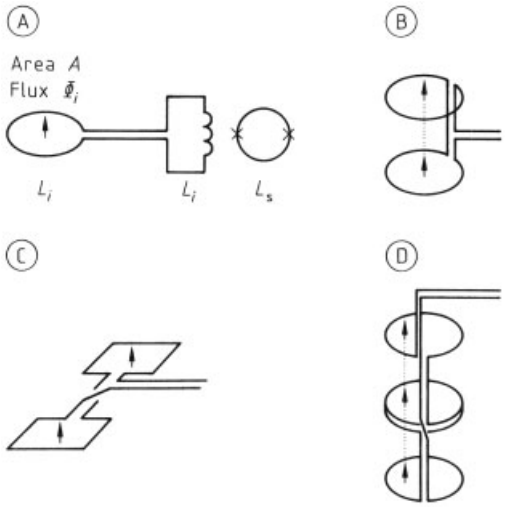

Figure 17. Flux transformer
A) Magnetometer; B) First-order axial gradiometer; C) Second-order axial gradiometer; D) First-order planar gradiometer

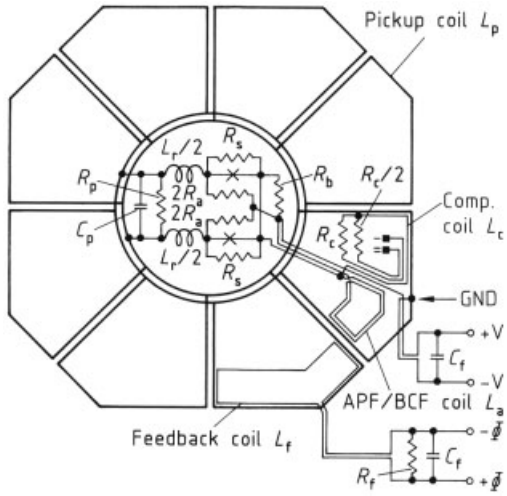

Figure 18. Layout and circuit diagram of a SQUID magnetometer designed by PTB, Berlin (chip size 7.2× 7.2 mm^2)

sents the state of the art for LT SQUIDs in large shielded systems.

4.2.4. SQUIDs and Biomagnetism

Figure 19 shows the amplitudes and frequencies of magnetic fields produced by physiological currents in some human organs. If the smallest biomagnetic signals of interest are to be resolved, the field noise contribution B_N of the SQUID should be no greater than a few tens of $fT/Hz^{1/2}$ in the frequency range from ca. 0.3 Hz to ca. 300 Hz. Also shown in Figure 19 are the interference fields in a typical urban environment, and geomagnetic activity. Measurement of biomagnetic signals, particularly from the heart and brain, is possible only if the interference fields are suppressed by several orders of magnitude. The advantage of the magnetocardiogram (MCG) and magnetoencephalogram (MEG) over ECG and EEG lies in the formation of large, parallel multichannel arrays with noninvasive localization and spatial and temporal resolution of spontaneous biological electrical current distributions. Not until the mid-1980s did reliable multichannel SQUID arrays become possible, based on Josephson junctions and niobium thin-film technology. Finally, in 1989–1991, the first commercial systems with more than 30 channels became available. These were large enough to permit spatiotemporally localized recording of biological currents in the heart or brain with a fixed array. Field resolution in each channel is 5–20 $fT/Hz^{1/2}$ for frequencies larger than 0.5 Hz. The first 37-channel system, the KRENIKON, was manufactured by Siemens [23]. Commercial systems now have up to 64 channels [24], and systems with up to 122 channels are being developed.

What the physician obtains is a series of three-dimensional images of the organism, derived from magnetic-resonance or X-ray computer tomography, onto which the space and time distributions of current are projected. Approximately ten such systems are in use worldwide. The clinical relevance of the method has been demonstrated particularly in the noninvasive localization of focal epilepsy combined with noninvasive radiotherapy (focused gamma-ray bundles), and noninvasive process monitoring (with SQUIDs) [25]. The method has also proven advantageous in the diagnosis of the heart condition Wolff–Parkinson–White syndrome. Applications in which biomagnetic measurements will probably be useful include Alzheimer's disease, schizophrenia, monitoring of recovery after trauma, and quantitative assessment of drug effects on brain function.

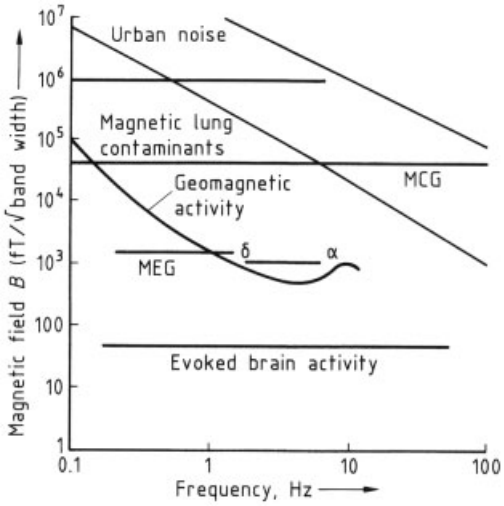

Figure 19. Frequency spectrum of biomagnetic signals compared with interference (noise) fields in an urban environment

The costs of such an apparatus – including the cooling system, magnetic shielding, interpretation system, and patient environment – are still very high (ca. € 2×10^6). This does not favor application of the device in a field important from the standpoint of patient numbers: noninvasive diagnosis of irregularities in cardiac rhythm. Nor is it contributing to the widespread adoption of this diagnostic technique. As a result, less costly methods of suppressing the interference fields are being sought, and attempts are being made to change over to the more easily cooled high-temperature superconductors.

4.3. Applications of HT Superconductors

It is now certain that the high-temperature superconductors, which have critical temperatures T_c above the temperature of liquid nitrogen (T = 77 K or − 196 °C) will find practical applications. They are already used in microelec-

tronics, where the currents and magnetic fields required are much lower than in power engineering and magnets. Power applications require conductors and cables of considerable size. HT-superconductor wires fabricated by ceramic techniques fall far short of the current-carrying capacity (at 77 K) needed in the magnetic fields encountered in industry, although advances have been made.

In contrast, microelectronic circuits can be fabricated by thin-film methods such as those known from the semiconductor industry. Superconducting properties relevant to applications, such as critical current density J_c and microwave losses (defined by the surface resistance R_s), have practically been optimized in epitaxial thin-film HT-superconductor structures. For certain components and circuit types, these properties offer marked technical advantages over established technical solutions based on normal metals at room temperature or with cooling.

4.3.1. Materials and Techniques for HT-Superconducting Electronics

The classes of HT-superconducting materials significant for microelectronic applications at 77 K are listed in Table 3 (Section 3.3). Most of the experience accumulated up to now is with YBCO. Because of the strong crystalline and electronic anisotropy of all cuprates, as well as the extremely short interaction length (coherence length) of the superconducting charge carriers, this material must be used in the form of oriented, virtually single-crystal epitaxial films which are practically free of grain boundaries. The quality attribute of such films is a critical current density $J_c > 5\times10^6$ A/cm^2 at 77 K in their own field together with the highest possible HF limiting field strength H_{rf} (the value above which high-frequency losses in the material increase rapidly). Until quite recently, films of such high quality could be made only from YBCO or other members of the REBCO (RE = rare earth) class; the best quality at present is $J_c = 5\times10^7$ A/cm^2 (50–60 K, 0 T) and $H_{rf} = 40$ kA/m (4 K) [26]. Methods have been devised for fabricating epitaxial REBCO films (*c*-axis perpendicular to film plane) on oxide substrates such as MgO, ZrO_2, $SrTiO_3$, $LaAlO_3$, and $NdGaO_3$. Substrate wafers four inches in diameter can now be obtained. Heteroepitaxial multilayer structures of YBCO and semiconducting (PrBCO) or insulating perovskites ($SrTiO_3$, $LaAlO_3$) have also been made. Accordingly, YBCO is preferred for most microelectronic application studies. High-quality TBCCO films, of interest due to the higher T_c, have also been used in high-frequency devices, but no multilayer coating technique is in prospect for this material, and the small H_{rf} (around 800 A/m) means that only small-signal applications are possible; hence it is not clear whether TBCCO will compete with YBCO. Advances have been made in the fabrication of BSCCO films and devices (single-layer films only), but film production has proved much more difficult than for YBCO.

Targets in material development are higher crystalline quality of HT-superconductor films, over larger areas, on substrates with low losses in the microwave band, at the highest possible HF field strengths, and the extremely difficult integration of HT superconductors with semiconductors such as Si and GaAs. Progress on the last problem is relatively slow, but integrated superconductor – semiconductor electronics have great potential. Buffer films are now available for YBCO on Si and GaAs, permitting deposition of high-quality YBCO films on 4″ substrates [27].

The development of processing techniques is marked by (1) major advances in the use of CVD for large-area film deposition; (2) marked reduction of the defect concentration in epitaxial films; (3) heteroepitaxy of HT superconductors and oxide-type dielectrics on prestructured surfaces and over steps (thin-film wiring, high-frequency components) and on edges (edge-type Josephson junctions); (4) epitaxy of *a*-axis films and heteroepitaxy with semiconducting perovskites yielding planar Josephson junctions; (5) epitaxy of buffer layers [e.g., ZrO_2 as Yttria stabilized ZrO_2 (YSZ)] under YBCO on "standard" large-area substrates such as 4″ sapphire wafers; and (6) the replacement of wet-etching processes (H_3PO_4 or ethylenediaminetetraacetic acid) by micromechanical process masks and by liftoff and dry-etching processes. However, in the absence of suitable selective reactive ion-etching processes, ion-beam methods must be employed.

4.3.2. Operating Temperatures of HT-Superconducting Electronics

The main applications of superconducting electronics occur where the cost of cooling is not a serious economic or technical problem. The temperature of cheap liquid nitrogen (77 K) – roughly the highest temperature at which HT-superconducting electronics can be expected to operate – is in contrast to electronics based on low-temperature (LT) superconductors, which are limited to temperatures around 4 K. Simple one-stage refrigerators can operate economically and reliably above 40 – 45 K, so that practical operating temperatures for HT-superconducting electronics are between 45 and 77 K. Thus the technical, economic, and psychological obstacle of helium cooling, which has hindered broader acceptance of LT-superconducting cryoelectronics, can be dispensed with.

4.3.3. Passive Components Based on HT Superconductors

A series of prototype high-quality (Q) coplanar, microstrip, and strip-line resonators have been made with YBCO films [28, 29]. Typical Q values may be several orders of magnitude higher than those of equivalent components made with cooled normal metal. Narrow-band multipole filters with very low insertion loss can be devised by combining a series of planar resonators. For example, a four-pole microstrip filter for the X band is made with YBCO and operates at 77 K. It has a 3 dB bandwidth of about 500 MHz (2 %), over 55 dB out-of-band suppression, dynamic range of over 20 dBm (100 mW), and maximum insertion loss under 2 dB [30].

The quality gains with YBCO films and new filter designs should increase the high-frequency rating to input power levels ≥ 50 W [26], so that usable HT-superconducting filters are possible on the transmitter end. Their compactness might make such filters useful in satellite communications and radar. Because high-frequency losses are low, planar microwave antennas can be miniaturized to dimensions much smaller than half the vacuum wavelength, while maintaining an adequate efficiency, so that the antennas can be made to operate over a wide band by appropriate folding. For the HTSSE II space experiment, the University of Wuppertal has devised a multielement antenna with radiated frequencies of 4.5 and 9.5 GHz in different directions [31]. The propagation of electrical pulses in the THz band with low dispersion and low losses has been demonstrated with YBCO strip lines [32]. Simple examples of wide-band delay lines have been realized with YBCO films of low to moderate quality [33]. The calculated and measured strip-line loss in dB/m are compared in Figure 14. For simple microstrip lines on $LaAlO_3$ wafers 50 mm in diameter, delays of 10 ns and a pass-band curve up to 20 GHz can be anticipated. A strip line 0.7 m long with modulated width, serving as a "chirp" filter, was fabricated on such a wafer; the bandwidth was 2.6 GHz, the center frequency 4 GHz, and the (dispersive) delay 8 ns, so that the time – bandwidth product was 21 [28]. Daimler-Benz has developed and reported a new approach to the implementation of such HT-superconductor chirp filters [33]. With a coplanar line configuration on a substrate measuring 3.8×3.8 inches, delays of over 100 ns were achieved for the first time. The (limited) advantages of HT superconductors as high-density conducting paths for chip-to-chip connections in computers (dispersionless, low-loss) are discussed in [34]. There is not as yet any prospect that the prerequisites for such paths, in particular the multilayer integration of HT superconductors and dielectrics with through-contacts and bump contacts.

Progress on passive HT-superconductor components in 1992 – 1993 suggests several future trends in development: low insertion losses at 77 K for a number of narrow- and wide-band components; marked extension of applications to analog signal processing, initially in satellite communications, radar, and other microwave hardware ranging down to the K band. Passive superconducting components may come into use in high-frequency technology, but only in the form of HT-superconductor components. Furthermore, a wide variety of such components must be integrated into superconductor- and/or semiconductor-based systems and subsystems, in particular so that the expense of cooling on board a satellite can be justified.

4.3.4. HT-Superconductor Radiation Detectors

Highly sensitive HT-superconductor bolometers for measuring thermal radiation can be created simply by operating a superconducting strip in the temperature range containing the very steep $R(T)$ transition to normal conduction, then measuring the voltage drop across the strip, which varies as a function of the incident radiation. Above 77 K, infrared sensitivity close to the theoretical limit of $D* = 2 \times 10^{10}$ cm Hz$^{1/2}$ W^{-1} can be achieved. Suitable design will allow response times in the microsecond range and shorter.

Prototypes of bolometers for infrared and far infrared, which can be used in spectrometry, radio astronomy, and space applications, have been described [28]. One such detector is also intended for deployment in the HTSSE experiment. The state of the art is as follows: sensitivity up to 1750 V/W, noise down to 7 ± 10^{-10} W/Hz$^{1/2}$ at 10 Hz, signal risetimes < 0.5 ms, chopper modulation at 100 Hz for 45 nm thick YBCO on a thinned Si substrate with buffer [35]. The feasibility of much faster IR quantum detectors is still being investigated, but no trend in development can yet be discerned.

4.3.5. Nonlinear HT-Superconductor Components

Figure 20 summarizes the types of Josephson junctions based on high-temperature superconductors. Up until 1990, only the "weak-link" junctions (Figs. 15 A and 20 A), formed at natural grain boundaries in polycrystalline ceramics or films, exhibited the Josephson effect at 77 K. The use of these devices for SQUIDs, microwave detectors, and mixers was demonstrated between 1987 and 1989. However, their uncontrollable characteristics made them unsuitable as a basis for an IC technology involving HT superconductors. The situation changed in 1990 when the first thin-film Josephson junctions made by a controlled process from epitaxial YBCO films and having usable weak-link $I-U$ characteristics at temperatures up to 77 K appeared [28, 29]. These Josephson junctions utilized a weak-link coupling at artificial grain boundaries [36–38] (Fig. 20 B), in YBCO – PrBCO – YBCO heterostructures [39] (Fig. 20 D), or in YBCO – Ag – YBCO junctions [40] (Fig. 20 E, produced in situ as a step junction). The use of epitaxial films also led to a reduction of the low-frequency $1/f$ noise, which originates chiefly at grain boundaries. A major improvement has been the reduction in interfacial resistance between the superconductor and the barrier material, which can be achieved by in-situ processes [41] or by wet etching [42]. Edge junctions (Fig. 20 D) with $CaRuO_3$ as the epitaxially-grown normally conducting barrier material, are now favored for future digital circuits because of the good reproducibility of I_c and R_N [43]. Microbridges etched by electron beams (Fig. 20 G)

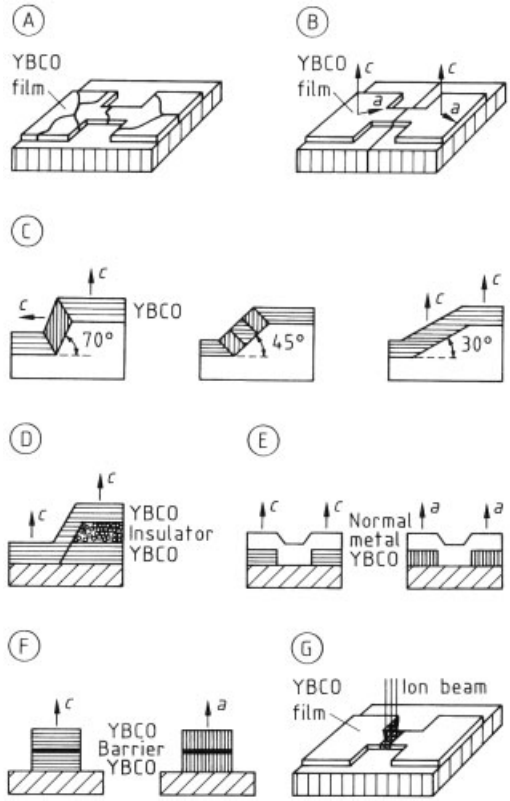

Figure 20. Types of Josephson junctions made with high-temperature superconductors [7, p. 434]
A) Grain-boundary junction (polycrystalline film); B) Grain-boundary junction (bicrystalline substrate or bicrystalline buffer); C) Step-edge junction (two different junctions connected in series); D) Edge junction with epitaxial barrier; E) Normally conducting bridge as in situ step junction; F) Planar junctions [not for 77 K operation (ca. 1990)]; G) Weakened bridge

exhibit excellent Josephson qualities but lack adequate stability [44]. Planar Josephson junctions (Fig. 20 F) for service at 77 K are currently not in prospect.

In contrast to low-temperature superconductor technology, it is still an open question whether HT-superconductor tunnel-type Josephson junctions (Fig. 15 B) are feasible for fast switches.

4.3.6. HT-Superconductor SQUIDs

Shortly after the discovery of high-temperature superconductivity in cuprates, three more discoveries came in quick succession: the Josephson effect, the SQUID function, and the fact that these effects occurred even above 77 K. However, extensive experimental activity soon showed that the sensitivity of HT SQUIDs was limited not only by thermal noise ($\sim k_B T$), but even more severely by a high level of low-frequency $1/f$ type noise. Moreover, the flux transfer from a pickup area to the SQUID, known for LT superconductors (flux transformer, usually made from niobium wire), has not been accomplished for HT superconductors using HTS wire. Such wire is still not available with the required noise performance. The outlook for HT SQUIDs was thus limited to applications where very high energy and field sensitivities were not required. SQUIDs based on HT superconductors can be used to advantage in nondestructive quality control in cases where classical flux-gate methods are insufficiently sensitive.

Improvements in the sensitivity of high-temperature SQUIDs are directly linked with reductions in $1/f$ noise and improved flux coupling. Until about 1989 – 1990, natural grain boundaries in massive ceramic HT superconductors or polycrystalline thin-film HT superconductors, chiefly of YBCO, served as Josephson junctions. The result was a large contribution to low-frequency flux noise from the many grains in the polycrystalline superconducting material. Individual grain boundaries serving as Josephson junctions in a thallium cuprate film, which is otherwise virtually free of grain boundaries, exhibited the lowest level of $1/f$ noise [45]. Since ca. 1990, alternative methods have been devised for making Josephson junctions at specified positions in the circuit. A successful method is the production of grain boundaries at specified points in epitaxial YBCO films (Fig. 20 B and C), which are otherwise of practically single-crystal character and thus low in noise, or to make the junction with a normally conducting bridge (Fig. 20 E).

Using step-edge junctions (Fig. 20 C) [38] and a SQUID inductance of 20 pH, the Siemens group obtained a SQUID energy resolution of 6×10^{-31} J/Hz (roughly $1000\,h$) above 10 Hz at 77 K [46,47]. The BTI group used highly reproducible superconductor/normal conductor/superconductor Josephson junctions, achieving 9×10^{-30} J/Hz above 1 Hz [40]. The resolution reached 3×10^{-31} J/Hz when submicrometer grain-boundary junctions were employed, but only at frequencies above 60 kHz, which is outside the range encountered in biomagnetic applications [48]. These best energy-resolution figures for HT d.c. SQUIDs thus correspond to the 10^{-30} J/Hz resolution of commercial helium-cooled d.c. SQUIDs, but are an order of magnitude lower than that of the d.c. SQUIDs in multichannel systems using LT superconductors. The HT SQUIDs under discussion did not, however, have flux coupling and so were not field-sensitive.

As a first step toward effective flux coupling in HT SQUIDs, planar high-temperature thin-film flux transformers were fabricated and combined with HT d.c. SQUIDs by the flip-chip technique, in which two chips are cemented with their active surfaces together. Such a flux transformer, with a magnetometer coupling loop combined with a grain-boundary d.c. SQUID, was first used by CONDUCTUS in the United States to record magnetic cardiac signals inside a magnetic shield; the field resolution was 600 f T/Hz$^{1/2}$ at 77 K above 10 Hz [49]. This concept was employed recently by the Sensor-Lab group in Japan, which implemented a functional SQUID array and recorded magnetocardiograms on 16 parallel channels with a field sensitivity of 370 f T/Hz$^{1/2}$ [50].

Another approach was taken by a team at the Jülich Nuclear Research Center, which created a single structure by combining HT RF SQUIDs operating at 77 K with grain-boundary Josephson junctions and large thin-film flux concentrators. The operating frequency was increased from the usual 20 – 150 MHz. In magnetic shielding, a field sensitivity of 170 f T/Hz$^{1/2}$

was achieved above 1 Hz in this way. With the addition of a ceramic flux concentrator 4.3 cm in diameter, the field sensitivity was boosted to 60 f T/Hz$^{1/2}$ above 5 Hz. The energy resolution of ca. 5×10^{-29} J/Hz was as good as that of helium-cooled RF SQUIDs. The system was able to record not only MCG cardiac signals (time signal) deemed diagnostically usable by a medical research group, but also registered signals of the human brain in the MEG for the first time [51]. Also a first-order electronic gradiometer of HT-superconductor type having field noise under 300 f T/Hz$^{1/2}$ was operated successfully above 2 Hz without magnetic shielding [52]. To derive medically relevant information from such data, however, and to perform measurements in clinics with their substantially higher ambient noise level, extensive work on noise suppression is still needed.

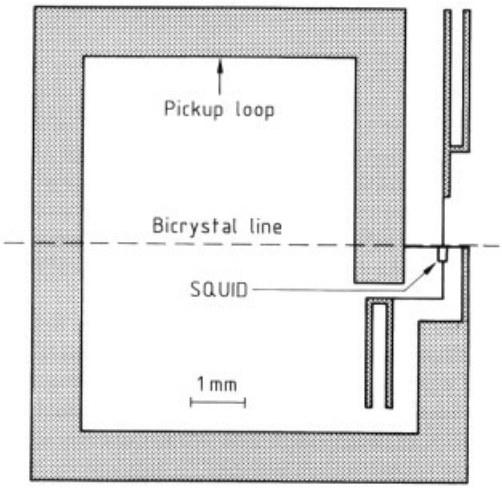

Figure 21. YBCO conducting-path layout of a direct-coupled d.c. SQUID, designed at University of California, Berkeley [49]

Progress in 1993 has created the prospect of a field sensitivity in the clinically relevant range of 20 f T/Hz$^{1/2}$. With a direct-coupled d.c. SQUID (Fig. 21), in which a magnetometer ring couples its signal into the SQUID ring via a conducting path, the Berkeley group increased the sensitivity to 39 f T/Hz$^{1/2}$ at 1 Hz [53]. This SQUID with integrated magnetometer consists of one thin-film layer. Advances in multilayer technology suggest that SQUID sensitivity can be boosted by at least a factor of two with flux transformers (which are necessarily two-layer devices). Using the flip-chip technique, the Siemens group has now achieved 65 f T/Hz$^{1/2}$ at 77 K and 5 Hz, using a thin-film flux transformer, a magnetometer with 0.4 cm^2 area, and a high-temperature d.c. SQUID. In a few optimal cases, itwas found that the flux transformer does not contribute to the noise [54].

4.4. Refrigerators for Cryoelectronics

The availability of reliable, long-lived, low-vibration, compact, economical refrigerators offering capacities of ca. 1 W at 80 K is very important for the wider adoption of cryoelectronics in industry. Although liquid nitrogen is easy to handle, there are advantages to a closed cooling cycle. Major progress has been made in Stirling and pulse-tube coolers. A Stirling cooler made by TRW offers cooling power of 250 mW at 65 K; the cooler weighs 1.4 kg, and the electronics 0.75 kg. The compressor consumes 12 W, and the electronics 5 W, part of which is used to suppress vibration. The lifetime is claimed to be 10 years. Given reasonable sales figures, prices should be in the range of \$ 10 000 – 20 000 in 1993 [55]. Pulse-tube coolers have the further advantage that there are no moving parts in the vicinity of the cooler, and the compressor can be located many meters away. Toshiba, Nihon University, and the Institute of Space and Astronautical Science [56] have developed a cooler with a 278 W linear drive, providing 10 W of cooling power at 80 K and a minimum temperature of 35 K.

The simplified cooling that can be employed with high-temperature superconductors, coupled with substantial progress in cryoelectronic HT-superconductor devices, leading to performance figures similar to those achieved with low-temperature superconductors, have extended the range of electronic applications of superconductivity to include nonlaboratory settings.

5. Application of Superconductivity in Magnet and Power Engineering

5.1. Introduction

The potential uses of superconductivity in magnet technology and power engineering can be divided into two broad categories:

1) In electrical equipment of conventional design (generators, motors, transformers, cables, etc.), the use of superconductors leads to a marked reduction in size and weight and to an appreciable gain in efficiency (energy savings).
2) The introduction of superconductors makes it possible to devise new components that may play a key role in future energy generation and distribution systems, including fusion reactors with magnetic plasma confinement, magnetohydrodynamic energy conversion, superconducting magnetic energy storage (SMES), and superconducting current limiters. On the basis of conventional technonolgy, these devices can either not be realized at all, or not at acceptable costs. The same holds for magnets with flux densities > 0.5 T for medical magnetic resonance imaging (MRI).

In the first category, the economic competitiveness of superconducting technology plays an important role. For many applications of metallic superconductors, the economic breakeven point in relation to present-day technology comes at very high unit capacities, where there may not yet be any demand. The economic difficulty results in part from expensive, inefficient cooling with liquid helium at 4.2 K (efficiency $\leq 0.4\,\%$). Technical-economic studies have shown that smaller unit capacities can become economically competitive with the new oxide-ceramic high-temperature superconductors and the far simpler and more efficient liquid-nitrogen cooling at 77 K (efficiency $\approx 10\,\%$). The use of the new superconducting technology, however, presupposes critical current densities $\geq 10^5$ A/cm^2 at 5 T and conductor prices close to those of metallic superconductors (e.g., € 5 km^{-1} A^{-1} for NbTi).

Superconductivity is essential for the second category of applications. It can be implemented with metallic superconductors, but the use of oxide-ceramic superconductors may yield a further gain in efficiency and decrease operating costs.

5.2. Industrial Superconductors

5.2.1. Metallic Superconductors

By virtue of its upper critical magnetic flux density H_{c2} (Table 2), NbTi at 4.2 K is used in magnetic fields up to 10 T, and at 1.8 K (superfluid He), up to 12 T. The predominant superconductor for the range of 12 – 20 T is Nb_3Sn. Figure 5 shows how the critical current densities of these materials depend on the magnetic field. Other candidates for use in high fields are Nb_3Al and V_3Ga.

In general, metallic superconductors are fabricated by co-extrusion (extrusion, drawing, rolling) in the shape of multifilament or multicore conductors containing a great number (up to tens of thousands) of thin superconducting filaments embedded in a matrix of normally conducting material. NbTi conductors for d.c. applications are made with a matrix of high-conductivity copper (Fig. 22 A), which can be processed directly with the ductile NbTi. In the case of the more brittle Nb_3Sn, a complicated fabrication method (bronze technique) leads to Nb_3Sn filaments in a high-resistance matrix of CuSn [57]. For d.c. applications, this is combined with a core or jacket of high-conductivity Cu, isolated from the CuSn by a tantalum diffusion barrier (Fig. 22 B). In both cases, the high-conductivity copper stabilizes the conductors in case of malfunctions (e.g., a sudden transition to the normally conducting state, known as quenching), by acting as a short-term current and heat bypass. The ratio of superconductor to stabilizing metal can vary between 1:1 and 1:20, depending on the application.

The diameters of the superconductor filaments lie in the range of 1 – 50 µm for d.c. applications. The small diameters are required to confer stability against changes in the magnetic flux, which might initiate normal conduction. The criterion for adiabatic stability is given approximately by [58]:

$$J_c\, D_{SC} < \sqrt{\gamma\, c_v\,(T)\,(T_c - T_0)\,/\mu_0} \qquad (5)$$

where J_c is the critical current density of the superconductor, D_{SC} the thickness of the super-

conductor, γ the mass density, $c_v(T)$ the specific heat, T_c the critical temperature of the superconductor, T_0 the operating temperature, and $\mu_0 = 4\pi \times 10^{-7}$ H/m the permeability of free space.

With regard to time-variant magnetic fields, the filaments must be electrically decoupled from one another, which is achieved by twisting the conductors. For power applications, conductors rated at between 100 A and 100 kA are required. These capacities are attained by fabricating several conductors (Fig. 22) into transposed stranded wires or Roebel bars; the individual conductors must be electrically decoupled for a.c. applications. Figure 23 shows a Roebel bar used for the field winding of a 400 MVA test generator in development [59].

If metallic multifilament superconductors are subjected to periodically varying (a.c.) fields at line frequency (50 or 60 Hz) and significant amplitude (e.g., in generators, transformers, and motors), the filament diameters must be lowered to ca. 0.1 µm to minimize hysteresis losses. The hysteresis loss per unit volume is proportional to the filament diameter [60]:

$$P_H = 1/2 f D_{SC} \Delta B J_c V_{SC} \quad (6)$$

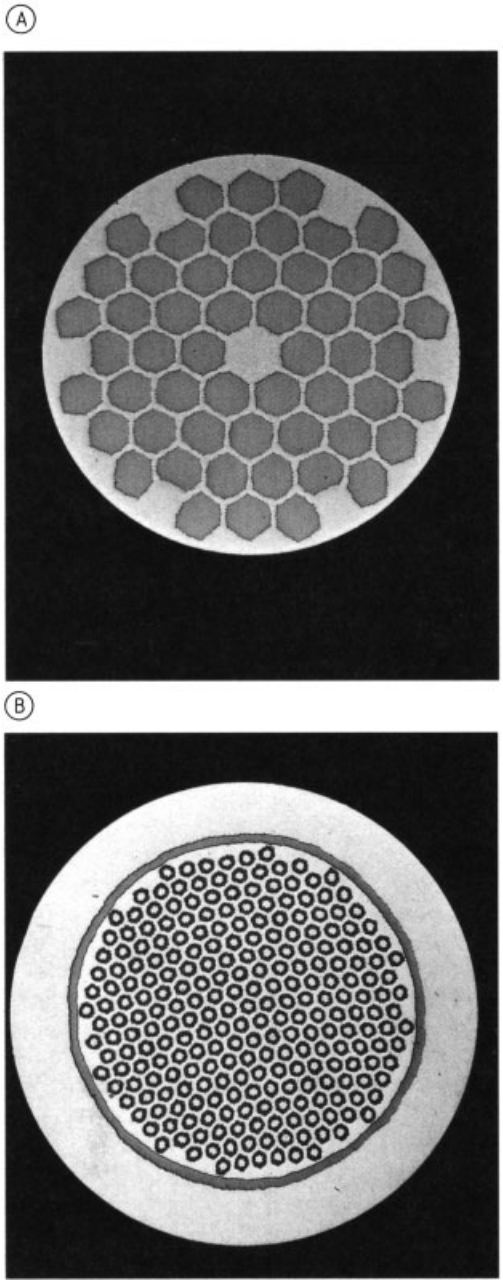

Figure 22. Metallic multifilament superconductors
A) 27 594 NbTi filaments, diameter 5 µm, in Cu matrix; B) 23 000 Nb_3Sn filaments, diameter 4 µm, in CuSn matrix
The filaments are assembled into bundles (source: Vacuumschmelze Hanau)

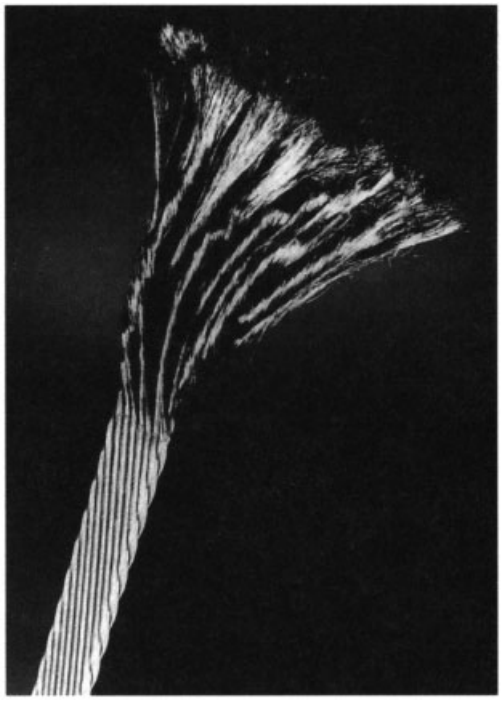

Figure 23. Transposed high-current conductor (Roebel bar) made of NbTi multifilament conductors
Current 10 k at 5 T (source: Vacuumschmelze Hanau)

where f is the frequency, D_{SC} the diameter of the superconducting filaments, ΔB is the field amplitude ($= 2\sqrt{2} \times B_{eff}$, perpendicular to filament thickness), J_c is the critical current density, and V_{SC} is the volume of the superconductor.

At lower field amplitudes, such as those in high-power transmission cables, the hysteresis losses are independent of superconductor thicknesses:

$$P_H = 2/3\mu_0^2 f \Delta B^3 A_{SC}/J_c \quad (7)$$

where A_{SC} is the area of the superconductor exposed to the field.

To minimize coupling losses (eddy-current losses) in the metal matrix, high-resistance alloys such as CuNi are employed.

5.2.2. Oxide-Ceramic Superconductors

For magnet and power applications with liquid-nitrogen cooling (77 K), the principal compounds of interest are $YBa_2Cu_3O_7$ and the lead-doped bismuth compound $(BiPb)_2Sr_2Ca_2Cu_3O_{10}$ (Bi-2223, see Section 3.3), with a high degree of texture. The state of the art is less advanced than in metallic superconductors, particularly as to the critical current density at 77 K and the control of the fabrication process for long conductor lengths. However, major progress has already been made.

The powder in silver tube process [61] is employed to make wires and thin tapes of Bi-2223 in lengths of up to 1000 m. Critical current densities are 10^4 A/cm^2 in zero magnetic field. Current densities of ca. 7×10^4 A/cm^2 have been achieved in short laboratory samples. Because its critical current density at 77 K is highly sensitive to magnetic fields, Bi-2223 at 77 K appears suitable only for low-field applications such as cables, current limiters, and possibly transformers.

The preferred configuration is also the multifilament conductor (Fig. 24). Such conductors can be stretched up to 0.3 % without degradation of the critical current density. The filament diameter can be made at least an order of magnitude larger than with metallic superconductors, chiefly because of the higher heat capacity at 77 K (adiabatic stability, see Eq. 5), and, in view of hysteresis losses at 50 – 60 Hz, the higher efficiency of liquid nitrogen cooling systems.

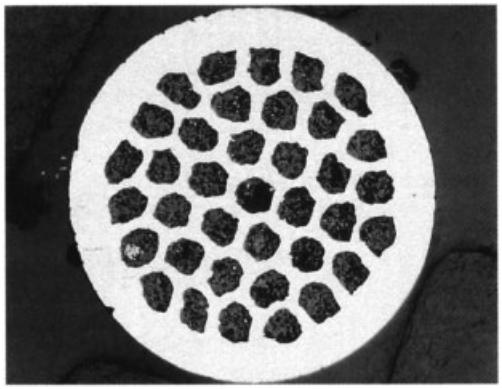

Figure 24. High-temperature multifilament superconducting wire made of Ag matrix with 37 filaments ca. 100 µm in diameter (source: Vacuumschmelze Hanau)

Two approaches have been taken in the development of highly textured conductors of $YBa_2Cu_3O_7$, which appears better suited than Bi-2223 for applications involving moderate and strong magnetic fields. The first is melt-texturing growth (MTG, see Section 3.3), while the second involves thin-film techniques, laser ablation, and chemical vapor deposition (CVD) for the production of thick layers deposited on metal tapes (Hastelloy) with intermediate buffer films (ZrO_2). The MTG process is particularly suitable for the fabrication of massive superconducting parts such as those used in passive magnetic bearings, superconducting permanent magnets, or magnetic shieldings. Critical current densities J_c at 77 K of 4×10^4 A/cm^2 have been attained in magnetic fields of 1.5 T [62]. The other approach has so far led to laboratory specimens with YBCO a few µm thick and having the following critical current densities at 77 K: 6×10^5 A/cm^2 (zero field), 8×10^4 A/cm^2 (0.4 T) [63].

5.3. Potential Superconductivity for Improvements in Conventional Electrical Devices

5.3.1. Superconducting Magnets for High-Energy Physics

High-energy physics is one of the most important applications for magnets made of metallic superconductors. Many gigantic d.c. magnets have been built for elementary particle detectors

as well as beam-guidance magnets outside the accelerator ring, and have been operated without problems for years in many laboratories worldwide. These magnets have yielded appreciable savings in electrical energy consumption and operating costs. Their higher magnetic flux densities (a factor of two to five greater than conventional iron-core magnets) permit substantial reductions in size.

These advantages have led to all modern ring-type particle accelerators being built with superconducting magnets. Examples are the 1000 GeV proton accelerator rings (Tevatron) at Fermilab in the United States (in service since 1984) and the Hera accelerator at DESY, Hamburg (operating since 1991), which have ca. 1000 and 630 dipole and quadrupole magnets, respectively, installed in their 6 km circumference rings. Larger machines of the same type are in the early phase of construction (the Superconducting Supercollider, SSC, with an 83 km ring) or in the planning stage (CERN's Large Hadron Collider, LHC, at Geneva, with a 27 km ring).

Replacing metallic superconductors with HT superconductors has little effect on installation costs, but the electrical energy costs for running these accelerators can be lowered approximately by a factor of four.

5.3.2. Magnetic Separation and Purification

In a magnetic separation process, ferromagnetic or paramagnetic particles are extracted from a stream of nonmagnetic material by the action of magnetic forces. The force acting on the magnetic particles is given by

$$F \sim \chi V \mu_0 H \nabla H \tag{8}$$

where χ is the susceptibility of the particles, V the particle volume, H the magnetizing field, and ∇H is the field gradient acting on the particle.

The use of superconducting magnets makes it possible, above all, to greatly increase the magnetic field strength and the field gradient over the values in conventional magnetic separators ($\mu_0 H \leq 1$ T). This allows magnetic separation of even weakly ferromagnetic or paramagnetic particles of small size, for which expensive and environmentally unfriendly chemical flotation methods were formerly required. Other potential applications of separators with superconducting magnets are the removal of magnetic impurities from minerals, water treatment, desulfurization, recovery of radioactive materials, and many other chemical and biological processes. This type of separator has the further advantage of a low energy consumption, roughly an order of magnitude less than for equal-capacity systems using water-cooled windings.

In the past, a number of pilot plants with liquid-helium-cooled NbTi windings were tested with success [64]. A favored separator type is the high-gradient magnetic separator (HGMS) illustrated in Figure 25. The magnetic field is generated by a superconducting solenoid; the field gradient is established by a stack of fine magnetic steel meshes (matrix) in which the magnetic particles are trapped while nonmagnetic particles can pass without hindrance. The magnetic field is then turned off, and the magnetic material is washed out with water, after which the process begins anew.

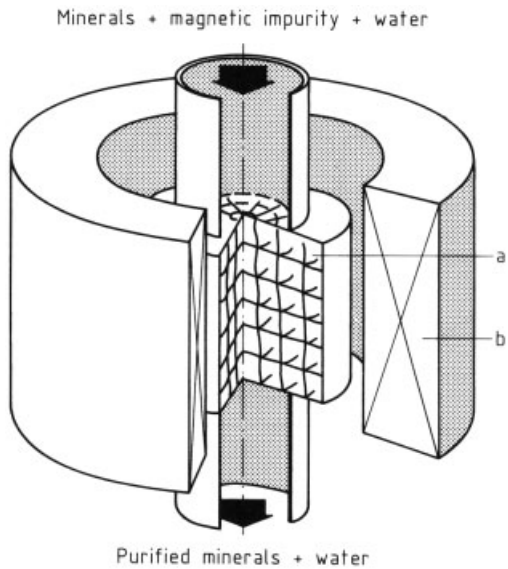

Figure 25. Schematic diagram of a high-gradient magnetic separator (HGMS)
a) Magnetic steel matrix; b) Superconducting magnetic field coil

The use of high-temperature superconductors, so that cheaper and more robust liquid nitrogen cooling can be employed, will greatly improve the market prospects for superconducting magnetic separators.

5.3.3. Superconducting Levitation for High-Speed Transportation Systems

Electromagnetic suspension, guidance, and propulsion systems are under consideration for a high-speed surface transportation system with speeds of 500 km/h or more.

In the electrodynamic system, short-circuited superconducting coils carrying a persistent current are mounted on either side of the chassis of the passenger vehicle (Fig. 26). Stationary, room-temperature copper suspension and guidance coils are located on either side of the roadway; when the vehicle moves relative to these, the magnetic flux of the superconducting coils (ca. 3 T) induces currents in the stationary coils, which exert a repulsive electromagnetic force that stably supports and guides the vehicle. An ironless linear synchronous motor with active roadway provides the propulsive force. Stationary copper drive coils embedded in the roadway are sectionally fed with a variable-frequency three-phase current to produce a traveling electromagnetic wave; the interaction of this wave with the d.c. flux of the superconducting magnets in the vehicle generates the propelling force. This system has the following advantages:

1) Wide air gap ($\geq$100 mm) for suspension, guidance, and propulsion
2) No need for air-gap regulation
3) Problem-free transmission of high power to the high-speed vehicle through contactless propulsion
4) No need for magnet power supply on board the vehicle

The performance of the system has been demonstrated in large test facilities in Europe and Japan [65, 66]. In the latter case, a 7 km test track carried a 10 t experimental vehicle at speeds up to 517 km/h as early as 1979. With the more realistic MLU 002 experimental vehicle (weight 17 t, length 22 m, capacity 44 passengers), speeds of around 400 km/h were attained on the same test track in 1989. The final introduction of the system into public transport is planned to take place by 1997, with prototype cars running on a new test track, ca. 43 km long, west of Tokyo.

The superconducting magnets use NbTi, but subsequent changeover to HT superconductors is planned. However, this will have little effect on the economics of the system since superconductivity accounts for only 1 – 2 % of costs.

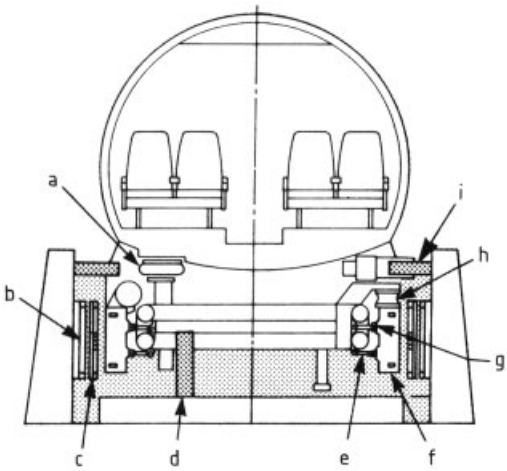

Figure 26. Electrodynamic levitation: cross section through the Japanese MLU 002 experimental vehicle (source: Japanese Railway Technical Research Institute)
a) Air-spring suspension; b) Propulsion coils; c) Suspension and guidance coils; d) Auxiliary support wheel; e) Chassis; f) Superconducting magnet; g) Magnet fixing; h) Air-spring suspension; i) Auxiliary guide wheel

5.3.4. Generators and Motors with Superconducting Windings

The power of an electrical machine with a cylindrical air gap is described by the proportionality

$$P \sim A_{\mathrm{eff}}\, B_{\mathrm{r}}\, D^2\, Ln \tag{9}$$

where D is the diameter of the air gap between the rotor and stator windings, L the length of the active part of the machine, n the rotation speed of the rotor, A_{eff} the effective linear current density at the periphery of the air gap (stator winding), and B_{r} the radial magnetic flux density in the air gap.

Superconducting windings without magnetic iron give far higher magnetic flux densities than in conventional machines ($B_{\mathrm{r}} \approx 1$ T). Omitting the iron also makes it possible to increase the effective linear current density A_{eff}. As a result, superconducting machines can be built much smaller and lighter than conventional machines with the same power rating. The low electrical losses of superconducting windings (all losses in

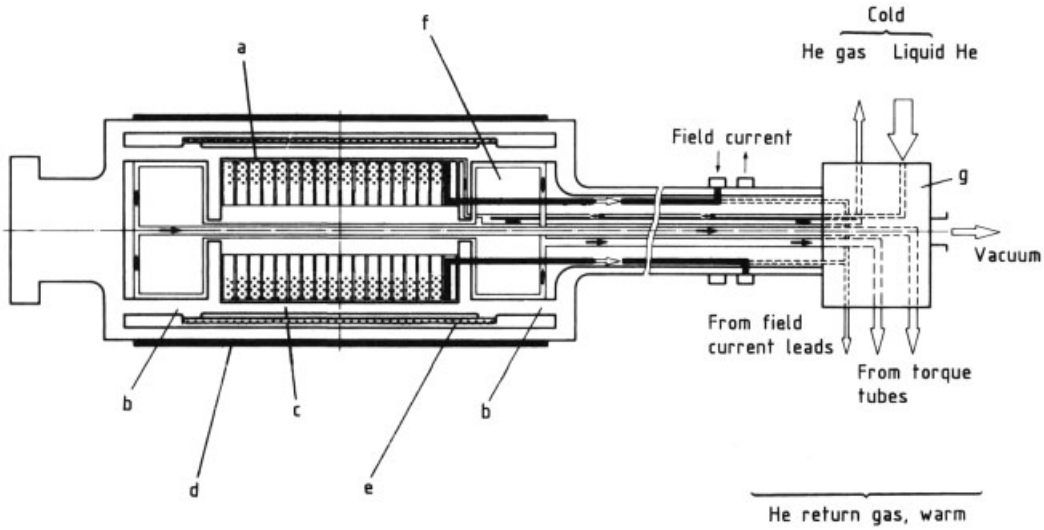

Figure 27. Schematic diagram of rotor with superconducting field winding
a) Superconducting field winding; b) Torque tube extension; c) Torque tube; d) Warm damper shield; e) Thermal radiation shield; f) Vacuum; g) Helium transfer unit

d.c. windings are thermal) and the low rotor friction losses (reduced size) lead to marked gains in efficiency, especially for high-power machines.

Preferred types of machine are d.c. and synchronous devices, because superconducting d.c. windings (without hysteresis losses) can be used here. Efforts are being made worldwide to create large superconducting synchronous generators for use in power plants. The present design has only the rotating field winding made of superconducting material, with the stationary three-phase winding built of water-cooled stranded conductors that contain fine copper wires. In comparison with conventional generators, the volume and weight are reduced by about a factor of two for large machines (500 – 1000 MVA), and the efficiency is increased from 98.85 % to 99.5 %. The economic breakeven point for NbTi conductors is in the range of 600 – 800 MVA.

Figure 27 shows the concept for a rotor with a two-pole superconducting field winding, whose conductors are laid in the grooves of a nonmagnetic steel cylinder. The design features helium off-gas cooling for the ends of the torque-transmitting cylinder and a cold radiation shield; as a result, the cold field winding (at 4.2 K) experiences minimal thermal load. Warm and cold damping shields prevent undesirable interactions between the three-phase winding and the superconducting field winding during transient operating conditions such as short circuits. The helium transfer unit delivers liquid helium to the rotating field winding and extracts warm and cold helium gas from the rotor. An important project is the Japanese Super GM program, in which Japanese industry and utility companies are developing three rotors with various superconducting field windings for a model generator of the 70 MW class [67].

A field winding with high-temperature superconductors and liquid-nitrogen cooling will shift the economic breakeven point into the 200 – 400 MVA power range. Furthermore, a completely superconducting machine (with superconducting three-phase winding) would also be conceivable with HT superconductors.

In large d.c. machines, the use of superconducting windings is limited by problems in transmitting current from the stationary to the rotating parts (commutator in heteropolar machines, very high currents in homopolar machines).

5.3.5. Superconducting Transformers

Advances in low-loss a.c. superconductors with submicron filaments in a resistive CuNi matrix have increased interest in the superconducting transformer. Figure 28 shows schematically a superconducting transformer. For more efficient removal of hysteresis losses, the iron core is held

at room temperature, cooled with a liquid such as oil. The high-voltage and low-voltage windings are placed in a thermally insulated tank that surrounds the iron core; it is made of glass-fiber-reinforced plastic in order to block eddy-current losses. The wind-ings are cooled in a bath of liquid coolant, which is also to perform part of the function of the dielectric and so must be subcooled to prevent the formation of gas bubbles, which would result in undesirable partial discharges. Other vital components are the low-voltage and high-voltage bushings [68].

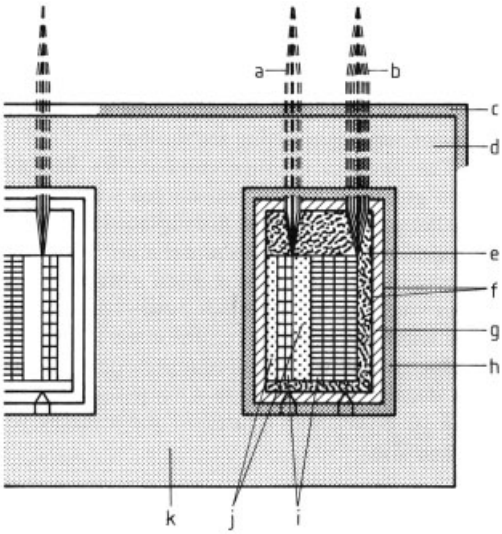

Figure 28. Schematic of a superconducting transformer (one phase)
a) Low-voltage connection; b) High-voltage connection; c) Oil tank; d) Iron yoke; e) Liquid coolant + electrical insulation; f) Coolant tank (glass-fiber-reinforced plastic); g) Thermal insulation; h) Oil; i) Superconducting low-voltage and high-voltage windings; j) Electrical insulation + coolant liquid; k) Iron core

Superconducting transformers offer the following advantages:

1) Marked reduction in 50 Hz winding losses, resulting in higher efficiency
2) Enhanced magnetic flux due to low winding losses, permitting a reduction in the cross section of flux-guiding iron and thus a substantial decrease in the weight and hysteresis losses of the iron core
3) Short-circuit current limiting through the transition of superconducting windings to the resistive normally conducting state, resulting in simpler mechanical design of the windings (milder forces) and thus leading to higher current densities and lighter-weight windings

In recent years, a number of small single-phase model transformers rated at 100 kVA have been built by using submicron NbTi filament conductors and have demonstrated the advantages of superconducting technology.

The benefits of superconductivity become much greater if the helium-cooled windings of metallic superconductors are replaced by liquid-nitrogen-cooled HT-superconductor windings. The following values have been obtained in a paper study for a 1000 MVA power transformer with windings made of a multicore HT-superconductor wire with 5 µm filaments (critical current density 10^5 A/cm^2) in a matrix of Ag – Al alloy (conventional figures in parentheses for comparison):

Weight of iron core	91 t (280 t)
Weight of windings	5 t (58 t)
Iron losses	100 kW (300 kW)
"Copper" losses (converted to 300 K)	200 kW (2000 kW)

5.3.6. Superconducting Power-Transmission Cables

High-power superconducting cables are of interest for underground power transmission in and near densely populated areas. Because transmission distances are relatively short (10 – 100 km), three-phase cable is preferred. Figure 29 (see next page) is a schematic of such a cable with HT-superconductor cores and liquid-nitrogen cooling. The flexible, hollow cable cores are of coaxial design. The actual phase conductors, made of superconductors (wires, tapes) in helical form, are enclosed in electrical insulation (wound tapes of polypropylene or polyethylene) and surrounded by a return conductor in which the a.c. current flows in the opposite direction to the phase conductors. As a result, no a.c. electromagnetic fields are produced outside the cores. A rigid or flexible double-walled thermal insulation layer surrounds the flexible cable cores; liquid coolant flows through the inner tube, while the outer one is at ambient temperature.

The practicability of superconducting cables has already been demonstrated in a number of experimental cables with helium-cooled metallic superconductors. The largest project was a

two-phase model cable of 115 m length with a power capacity of 1000 MVA at a voltage rating of 138 kV [69]. However, cables with helium cooling become economically competitive with conventional water-cooled oil cables only when carrying ca. 4000 MVA and up. Estimates indicate that the economic breakeven point for nitrogen-cooled HT-superconductor cables may be as low as 1000 MVA.

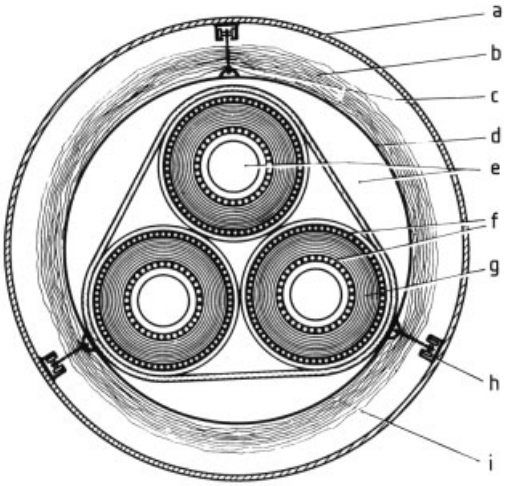

Figure 29. Liquid-nitrogen-cooled three-phase cable with high-temperature superconductors
a) Outer steel pipe; b) Superinsulation; c) Suspension; d) Nitrogen tube; e) Nitrogen coolant; f) Superconductors; g) Electrical wrapped-tape insulation; h) Suspension track; i) Vacuum

5.4. Novel Electrical Devices for Which Superconductors Are Indispensable

5.4.1. Magnets for Magnetic Resonance Imaging (MRI) and Spectroscopy

Two new medical diagnostic technologies are based on high-frequency signals emitted by protons or other nuclei (e.g., those of phosphorus or sodium) when they are aligned in a strong, highly homogeneous magnetic field and excited with high-frequency electromagnetic radiation. The detected signals are converted to cross-sectional images of the body, or used in local spectroscopy (chemical shift) to determine compounds important in metabolism.

Whole-body imaging employs magnetic fields ranging from 0.15 to ca. 1.5 T, while chemical shift spectroscopy uses fields of 1 – 2 T. Image quality improves with increasing magnetic flux density. For this reason, most installed MRI systems are equipped with superconducting magnets operating at flux densities between 0.5 and 2 T. The magnets are of cylindrical design (solenoids), 2 – 3 m long, with an internal bore about one meter in diameter in which the patient is placed (Fig. 30).

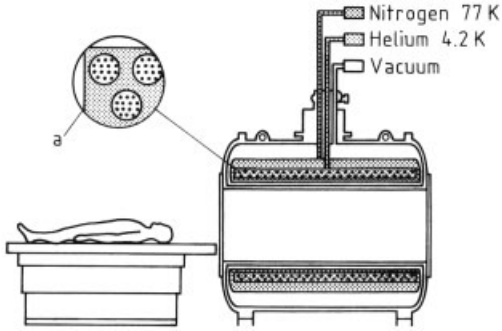

Figure 30. Superconducting magnet providing magnetic flux densities ≥ 0.5 T for magnetic resonance imaging (MRI)
a) Windings of Cu – NbTi multifilament conductors

The high spatial field homogeneity (better than ± 10 ppm in a sphere 0.5 m in diameter) required for these applications is achieved by using multicoil configurations and shim coils that surround the patient, like the high-frequency and gradient coil systems. The magnets exhibit high field stability over time ($\Delta B/B < 10^{-7}\ \mathrm{h}^{-1}$) due to the persistent-current operation that is possible with superconducting coils [70].

At present, all these superconducting magnets are constructed with NbTi metallic multicore conductors; ca. 1000 units are sold per year. A changeover to HT superconductors would lower capital costs by about 15 %, which corresponds to ca. 5 % for the entire system. In addition, cooling costs would be cut by \$ 100 000 to \$ 200 000 for a five-year service life.

5.4.2. Magnet Systems for Magnetic-Confinement Fusion Reactors

Magnetic plasma confinement is a promising route to the economical utilization of thermonuclear fusion energy. Two toroidal magnetic field configurations, the tokamak and the stellarator, appear especially suitable. Commercial reactors

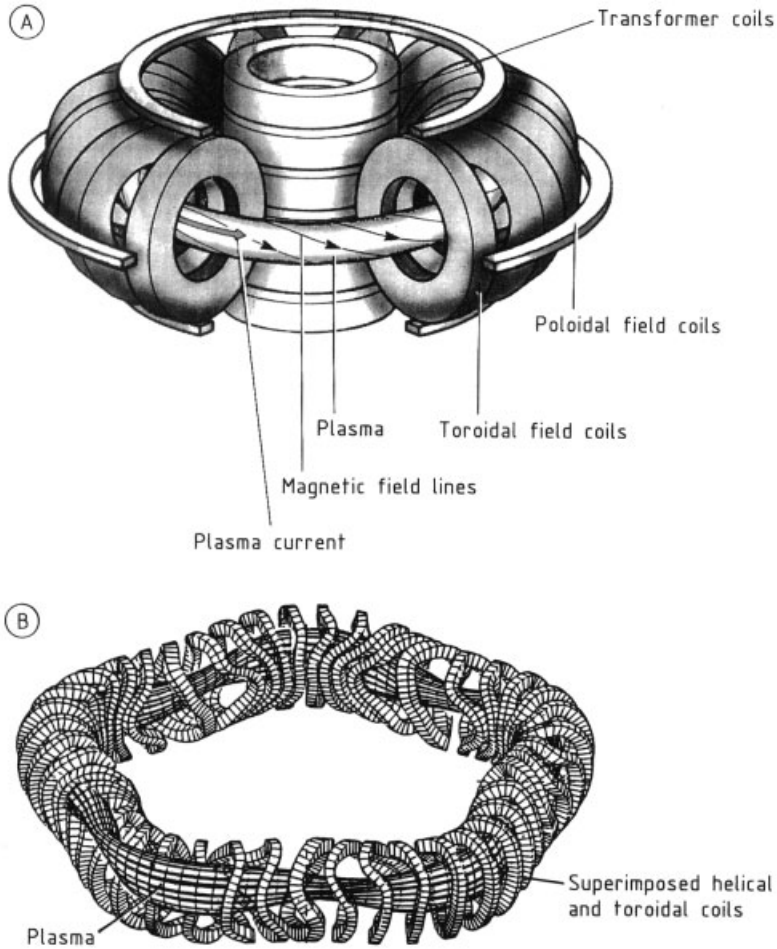

Figure 31. Principal magnet configurations for magnetic-confinement fusion reactors
A) Tokamak; B) Stellarator

will require superconducting magnet systems to achieve the required power-plant efficiency.

The coil layouts of a tokamak and a stellarator are shown in Figure 31. In the axially-symmetrical tokamak configuration, the magnetic surfaces, which as surfaces of constant pressure contain the helical field lines, are established by superposing the toroidal main field and a poloidal magnetic field. The poloidal field is generated chiefly by an electric current induced in the toroidal plasma and also by annular internal transformer coils and external poloidal field coils. In the stellarator, the magnetic surfaces are generated solely by currents in external coils, so that a purely modular design is possible for the coil system. Magnetic flux densities of 5 – 6 T are required on the ring-shaped plasma axis in the tokamak configuration, corresponding to maximal magnetic fields of ca. 12 T at the superconducting coils, which can be attained only with Nb_3Sn conductors. The stellarator makes less stringent requirements on the magnetic field, which can be satisfied with NbTi conductors.

The technical feasibility of large magnets for fusion has been demonstrated in a number of large superconducting tokamak systems. Examples are the international LCT project at Oak Ridge National Laboratory in the United States, a toroidal configuration comprising six D-shaped magnets; and the TORE-SUPRA at the Centre d'Etudes Nucleaires Cadarache in France [71]. The largest international project at this time is the development of a magnet system

for the International Thermonuclear Experimental Reactor (ITER). This will be a tokamak-type system with D-shaped toroidal field coils measuring 18×11 m and poloidal field coils up to 25 m in diameter; the maximum flux density will be around 13 T. Other magnet configurations include the stellarator project (Wendelstein W7-X) at the Max Planck Institute for Plasma Physics, Garching, Germany; and the Large Helical Device (LHD) at the National Institute for Fusion Science (NIFS) in Japan.

The state of the art is based exclusively on metallic superconductors. There would be a variety of benefits in changing over to HT superconductors. First, the overall efficiency of a fusion power plant could be increased by 2 – 4 %. Simpler cooling would result in higher reliability and availability as well. However, this application would impose the most severe requirements (magnetic fields, forces) on the high-temperature superconductors.

5.4.3. Magnetohydrodynamic Energy Conversion

Magnetohydrodynamic (MHD) generators permit the direct conversion of thermal enthalpy to d.c. electricity. Strongly heated, ionized gas (natural gas) flows at high velocity perpendicular to an imposed magnetic field between electrodes on either side of the flow channel. If such a generator is installed as the first stage, upstream of a conventional steam-turbine loop, the overall efficiency of the combination power plant can be enhanced to ca. 50 %. For this, however, at least 20 % of the enthalpy must be extracted from the hot gas, which requires magnetic field strengths of at least 6 T. Due to these high values and the huge dimensions of actual commercial systems, dipole magnets 25 m long with a 5 m free bore are required. Therefore, only superconducting magnets can be considered (on reasons of efficiency). The properties of NbTi conductors are adequate.

Although a number of smaller model systems with superconducting magnets have been tested successfully, work ceased worldwide in the early 1980s due to aging problems with the channel wall and electrodes. In 1989, a new MHD program was begun in Italy [72].

Because the electrical energy consumed by superconducting magnets is only ca. 1 % of the energy generated, a changeover to HT superconductors would not result in any major savings in operating costs or gains in efficiency for an MHD system. The same applies to installation costs.

5.4.4. Superconducting Magnetic Energy Storage (SMES)

The ability of superconductors to carry direct current without losses means that a persistent current can be maintained in a short-circuited superconducting coil with no external current supply. This makes superconducting coils very attractive as a means of storing electrical energy directly, i.e., without conversion to other forms of energy. The stored energy is proportional to the square of the magnetic flux density and to the volume of the coil. A network-controlled rectifier (rectifier/inverter) connects the storage coil with the line. The operating state of the storage device (charging, discharging, storage) is set by controlling the firing angle of the thyristors (Fig. 32). In comparison with other energy storage means, SMES has the advantage of very short response times (< 100 ms), while larger storage devices have the further advantage of high efficiency (> 90 %). In addition, they can be used for several tasks in the grid simultaneously, such as peak handling, grid stabilization, and hot standby.

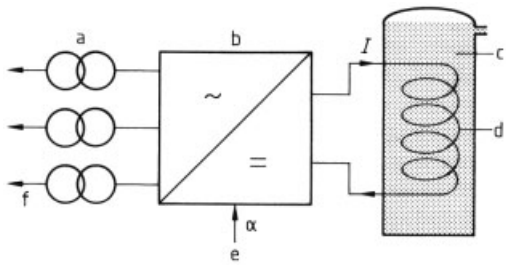

Figure 32. Operation of a superconducting magnetic energy storage (SMES) device
a) Transformer; b) Rectifier/inverter; c) Liquid nitrogen; d) Superconducting coil; e) Control; f) To three-phase grid

In the United States and Japan, where subsidiary grids are less well interconnected than in Europe, the discussion concerns the use of large storage units for peak-load storage (day/night

storage). Such devices, with power ratings in the 1000 MW range and capacity of several GWh, can compete with large pumped-storage hydro plants. Small and medium-sized storage units are being investigated as a means of improving grid stability (frequency, voltage, power fluctuations), smooth out shock loads, and serve as immediate standby sources. Their use with strongly fluctuating renewable energy sources (sun and wind) is also being studied.

Increased use of SMES in future grids may reduce the need for new plant construction and improve the utilization of existing fossil-fuel plants. This would yield an appreciable cut in fossil-fuel consumption and in the associated CO_2 emissions and other problems.

In 1991, development of a storage unit (100 kWh, 20 MW) for grid stabilization began in Japan. An Engineering Test Model (ETM) with a storage capacity of 20 MWh is under discussion in the United States. "Micro" SMES units (ca. 1 MW) are already being used to protect sensitive manufacturing equipment against voltage spikes.

The designs of large-scale storage units and experimental systems are based on metallic superconductors at 4.2 K. The use of HT superconductors at 77 K would have little effect on the capital costs of a storage unit but would improve the efficiency, even in large units (ca. +3 %). For small and medium-sized storage plants ($\leq$ 1 MWh), HT superconductors are an essential precondition for acceptable efficiency levels [73].

5.4.5. Superconducting Current Limiters

Increasing short-circuit currents in power grids impose steadily growing dynamic loads on integral components such as generators, transformers, and switchgear. Another idea is to use the fast ($\leq$ 1 ms) transition of superconductors from the superconducting to the normally conducting state (from $R \approx 0$ to very high R) to limit fault currents to acceptable values. The transition can be initiated spontaneously (when critical values are exceeded) or controlled externally (with pulsed supplemental current or supplemental magnetic field).

The three basic limiter concepts now being studied for three-phase grids are illustrated in Figure 33. In Figures 33 A and B, the superconductor is in the form of a wire or tape of well-defined length in a low-inductance configuration (e.g., a bifilar winding). In the resistive limiter (Fig. 33 A), the superconductor is connected in parallel with an ohmic resistance R_p kept at room temperature. When the superconducting winding (trigger coil) becomes normally conducting ($R_s >> R_p$), the current switches over to the parallel resistance R_p, which limits it to a predetermined value. After ca. 50 ms, the limited

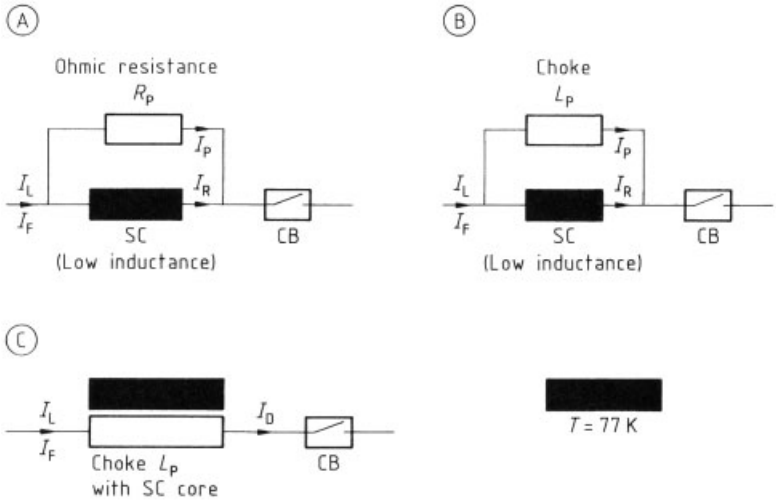

Figure 33. Main types of superconducting current limiters (one phase shown)
A) Resistive; B) Hybrid; C) Inductive, SC = Superconductor; CB = Conventional breaker

current is interrupted by a conventional breaker. The resistive limiter is also well-suited to limiting direct currents. The hybrid type (Fig. 33 B) also has a low-inductance superconducting trigger coil, but here the current is limited by a choke (limiter coil), a high-inductance copper or superconductor winding connected in parallel with the trigger.

The inductive class (Fig. 33 C) has a number of variants, such as a normally conducting choke whose windings are shielded from its iron core by a cylinder made of massive superconducting material. In normal operation, the cylinder is superconducting and the inductance of the coil is very low. If a fault current causes the cylinder to become normally conducting, it loses its shielding action and the iron core of the choke becomes fully effective; the inductance of the coil increases abruptly and limits the current.

The feasibility of superconducting current limiters at medium voltages has been demonstrated with a number of small, single-phase experimental devices using metallic superconductors. In the most advanced project, the Japanese industry is collaborating with a large electric utility to develop a three-phase, 6 kV/2 kA limiter [74]. The use of the new HT superconductors is also being tested, because the simplicity of cooling with liquid nitrogen instead of helium should yield a decrease of ca. 70 % in the operating losses and savings of ca. 60 % for the cooling plant.

6. Organic Superconductors

6.1. Introduction

Molecular conductors differ fundamentally from the regular metallic conductors in that they are composed of building blocks (the molecules) that exhibit their own specific properties (e.g., molecular orbitals with their ionization energy or electron affinity, characteristic IR or Raman active modes, and NMR chemical shifts of different atomic sites). In contrast to conventional molecular crystals made up of neutral organic molecules held together by weak van der Waals forces, organic conductors contain molecules with unpaired carriers in π molecular orbitals (open shells) which allow delocalization over all molecular sites in the crystal via strong intermolecular π overlap. The open-shell character arises from partial oxidation (reduction) of donor (acceptor) molecules with formation of a salt of an inorganic anion (cation).

These molecular conductors are therefore different from the extended conjugated polymers and graphite where the π system of the extended molecules provides the conducting pathway.

The development of molecular conductors (superconductors) has been stimulated to a large extent by the suggestion made by LITTLE in 1964 [75] that the arrangement of chain conductors in a polarizable medium could give rise to superconductivity at high temperature.

The geometry of the building blocks governs the dimensionality of the transport properties of these particular conductors (between one and three dimensions).

For one-dimensional conductors, the planarity of the molecules is responsible for their particular crystal structure (Fig. 34 A). All these molecules are modifications of the flat fulvalene skeleton. Noticeable intermolecular overlap is allowed when they form stacks. The conduction band derives from π orbitals directed along the stacks and reveals a pronounced 1-D character, i.e., a flat and open Fermi surface [76, 77].

Higher dimensionality conductors are also derived from similar planar molecules that form dimers with strong overlap between chalogen atoms both within each dimer and between neighboring dimers (Fig. 34 B). This crystal structure gives rise to a large conductivity within molecular layers and to about 10^4-times less conduction between sheets. They can be considered as 2-D conductors [78].

The discovery of the spherical C_{60} molecule has opened the way to 3-D isotropic molecular conduction. Three of the four valence electrons of each carbon atom are engaged in sp^2 bonding with the three neighbors. There remains one electron per carbon atom in a p-state which is delocalized over the whole surface of the sphere in πmolecular orbitals. The neutral C_{60} molecule possesses a closed-shell electronic

Figure 34. Crystal structure of organic conductors
A) 1-D Conductor $(TMTSF)_2PF_6$ viewed perpendicularly to the high-conductivity axis; B) 2-D Conductor $(BEDT\text{-}TTF)_2Cu(SCN)_2$ viewed along axis of least conductivity; C) 3-D Conductor K_3C_{60}

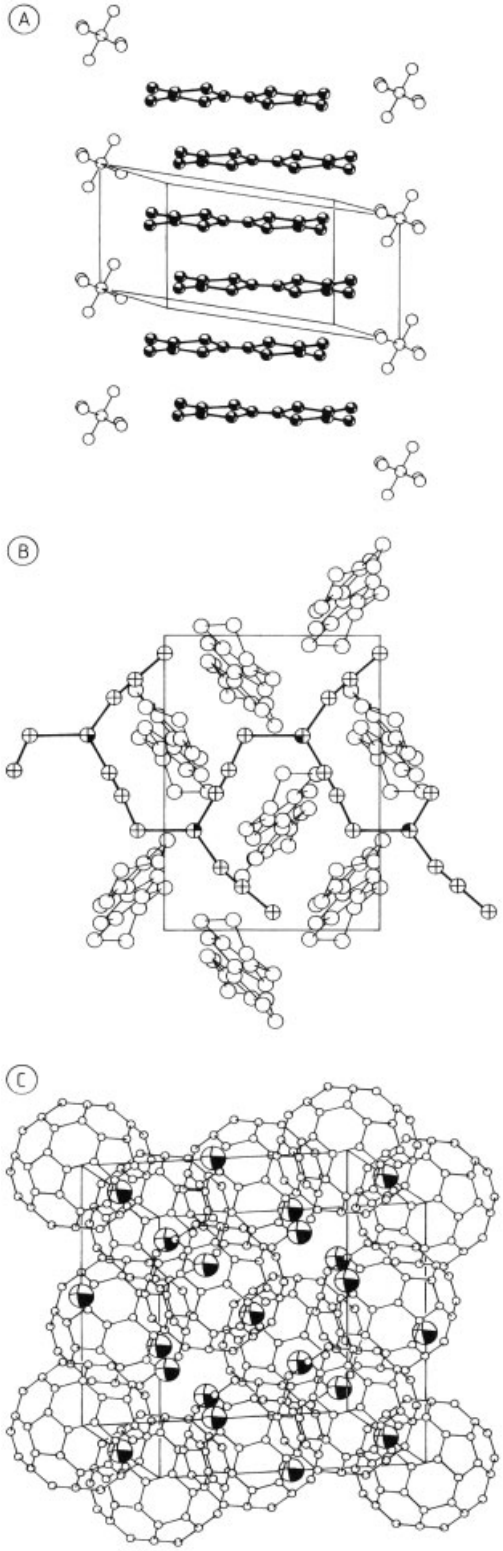

structure, giving rise to an insulating molecular solid. However, populating the highest unoccupied molecular levels (the LUMO) with carriers donated by inorganic cations such as alkali or alkaline-earth metals gives rise to a partly filled conduction band. This band is half-filled in the prototype material A_3C_{60} (A = K, Rb, Cs, etc.). The C_{60} spheres are packed in a face-centered cubic lattice, leaving cavities with octahedral and tetrahedral local symmetry, which are all occupied in A_3C_{60} (Fig. 34 C) [79].

6.2. Electronic Structure and Superconductivity

The salts of the tetramethylselenafulvalene (TMTSF) molecule are regarded as prototype materials for 1-D molecular conductors. The overlap between organic stacks along the transverse *b*-direction amounts to $t_{\perp b} \approx 20-30$ meV, which is about one tenth the value for the intrastack overlap ($t_{\parallel} \approx 200-300$ meV), making the conduction band about 1 eV wide. The overlap is negligible along the *c*-direction. Such a marked anisotropy causes the Fermi surface to be nearly planar, intersecting the $k_{\parallel}$ axis at wave vectors $\pm k_F$, where the Fermi wave vector for a 1-D conductor is related to the density of carriers per unit cell ϱ by the relation $2k_F = \varrho\pi/a$ [80]. The stabilization of superconductivity below a certain critical temperature T_c requires overcoming the strong divergence that develops at low temperature in a 1-D electron gas towards an insulating behavior (see below).

$(TMTSF)_2PF_6$, the organic compound in which superconductivity was first observed [81], requires a pressure of 9 kbar. However, $(TMTSF)_2ClO_4$ is actually the only 1-D material that undergoes a transition to superconductivity at $T_c = 1.2$ K under ambient pressure. These temperature values are rather low, but higher T_c values are reached in materials where the shape of the Fermi surface is less likely to lead to the stabilization of insulating ground states (the 1-D and 2-D molecular superconductors, Fig. 35; see next page).

In 2-D conductors, the Fermi surface is tubular with $t_{(a,b)} = 100-150$ meV, $t_{\perp c} \approx 0.5$ meV, and consequently T_c rises up to 8.1 K in $(BEDT\text{-}TTF)_2I_3$ [bis(ethylenodithio) tetrathiafulvalene

Table 4. Properties of organic superconductors

Property ∗	Compound (dimensionality)		
	TM_2X (1-D)	$\varkappa$-ET_2X (2-D)	A_3C_{60} (3-D)
T_c, K	1.2 (X=ClO_4)	9 (X=Cu(SCN)$_2$); 12 (X=Cu[N(CN)$_2$]Br)	19 (A=K); 33 (A=Rb_2Cs)
H_{c1} (0 K), 10^{-4} T	0.2 (*a*); 1 (*b*); 10 (*c*)	1 (*a, b*); 45 (*c*)	
H_{c2} (0 K), T	2.8 (*a*); 2.1 (*b*); 0.16 (*c*)	30 (*a, b*); 2 (*c*)	49
ξ (0 K), Å	700 (*a*); 335 (*b*); 20 (*c*)	174 (*a, b*); 7 (*c*)	26
λ (0 K) Å		5000 – 10 000	2400 – 4800

∗ T_c = Critical temperature; H_c^1, H_{c2} = Critical fields; ξ = Coherence length; λ = Penetration depth.

and even up to 10 – 12 K for anions such as $Cu(SCN)_2^-$ and $Cu[N(CN)_2]Cl^-$ [78, 82].

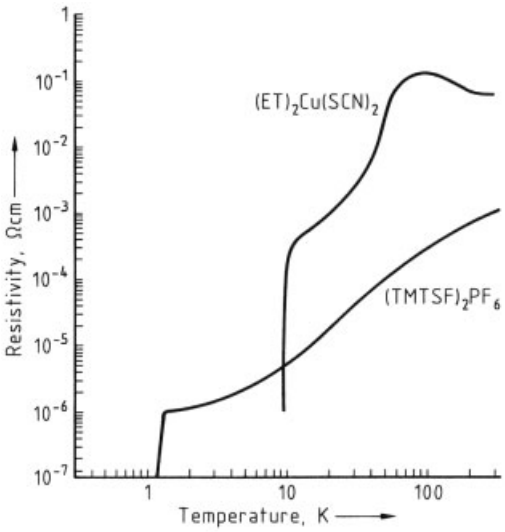

Figure 35. The superconducting transition of 1-D (under pressure) and 2-D organic conductors

Like one- or two-dimensional molecular conductors, the conductivity in fullerenes is due to π overlap between organic molecules, counterions playing essentially no role in the conduction process. The shape of the molecules and the crystal structure make A_3C_{60} compounds three-dimensional conductors with a very narrow conduction band (≤ 0.5 eV) due to the weakness of the intermolecular interactions. Superconductivity ranges from 19 K in K_3C_{60} up to 28 K in Rb_3C_{60}. The highest T_c stabilized at the moment for a molecular conductor is 33 K in Cs_2RbC_{60} [79].

All these materials display the general features which are common to superconductors, namely a state with zero resistance and the expulsion of magnetic flux below T_c. They are type II superconductors: there exists a broad field range between H_{c1} and H_{c2} in which the magnetic field is not fully expelled from the material (the vortex state). For all organic conductors but the fullerenes, parameters describing the superconducting phase are strongly anisotropic (see Table 4 where TM and ET are abbreviations of TMTSF and BEDT-TTF, respectively).

Although the search for new superconductors has been the driving force in the study of molecular conductors, many more basic physical properties related to the low-dimensional character have been discovered and are discussed briefly in Section 6.3.

6.3. Other Features of 1-D Superconductors

The TM_2X Phase Diagram. A direct consequence of the 1-D character of the TM_2X series is the existence of a wide variety of ground states which can be observed in various members of the family depending on parameters such as the chemical composition of the organic molecule or of the inorganic counterion and also on hydrostatic pressure.

Figure 36 displays the variety of ground states; it is a generic diagram for the TM_2X family.

On the left of the diagram, TM_2X compounds exhibit a metal-like behavior at high temperature, but below the temperature $T\varrho$, where the resistivity passes through a broad minimum, a loss of the charge degree of freedom is observed, leaving the spins localized and behaving like those of a chain quantum antiferromagnet. Coulombic repulsions are responsible for such localization (Wigner-like localization) [83]. In the same temperature range the electron–phonon interaction gives rise to a 1-D

lattice fluctuation with a doubled periodicity which is detected by X-ray experiments. Below ca. 20 K these 1-D lattice fluctuations order themselves three-dimensionally and the system undergoes a phase transition with loss of the spin degrees of freedom. The ground state is a spin-Peierls (SP) phase in which the lattice is tetramerized and the spins are coupled in a singet state of total spin zero.

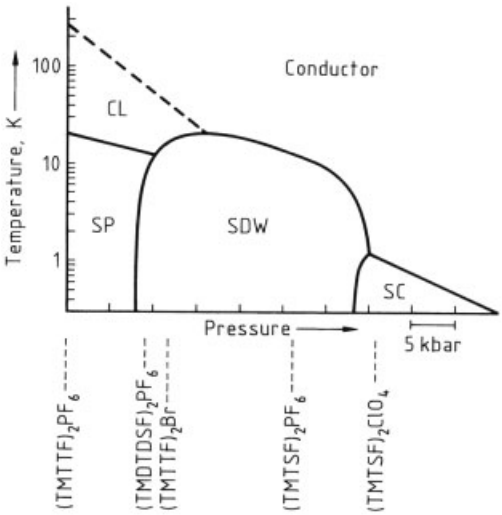

Figure 36. The generalized phase diagram for the TM_2X series
TMTTF = tetramethyltetrathiafulvalene;
TMDTDSF = tetramethyldithiadiselenafulvalene;
TMTSF = tetramethyltetraselenafulvalene
SP = Spin-Peierls; SDW = Spin density wave; SC = Superconducting ground states; CL = Region where charges are localized by Coulombic forces

As one moves toward the right-hand side of the diagram the role of the electron – phonon interaction becomes less dominant but electron – electron interactions remain important, as shown by the temperature dependence of the spin susceptibility, which is still strongly reminiscent of the 1-D antiferromagnet. In this region of the phase diagram the new order that becomes established below 19 – 12 K is characterized by a modulation of the spin density with a wave vector $\boldsymbol{Q}$ which provides the best nesting of the Q-1-D Fermi surface [80]. It is the presence of this additional periodicity in the exchange potential which is responsible for the opening of a gap $2\,\Delta$ at the Fermi level accompanied by a metal – insulator transition.

Finally, on the very right of the diagram, the spin density wave (SDW) ground state itself is suppressed by an enhanced transverse coupling and superconductivity becomes stable instead.

As far as the conducting phase is concerned, the electron gas cannot be regarded as an ordinary electron gas. Once more the one-dimensionality and the existence of strong electron – electron coulombic repulsions make this metallic state remarkable. In spite of a fairly high conductivity – reaching $10^6\,(\Omega\,\mathrm{cm})^{-1}$ at low temperature for the compounds in the $(TMTSF)_2X$ series – the susceptibility of the conducting state develops a strong tendency toward antiferromagnetism at low temperature. As shown by NMR experiments, the antiferromagnetic response becomes prominent below 150 K and displays a power law divergence $\chi_S \approx (T/E_F)^{-\gamma}$ with $\gamma \approx 0.85$ as expected from the theory of the 1-D correlated electron gas. On the other hand, the uniform spin susceptibility measured by a Faraday technique is temperature dependent, very much like that of a 1-D quantum antiferromagnet. Therefore, the superconducting instability of $(TMTSF)_2X$ conductors arises out of a background of antiferromagnetic fluctuations.

Spin Density Wave Collective Conduction. The antiferromagnetic state of Figure 36 also exhibits various novel physical properties. It is the nesting of the Fermi surface with the wave vector $\boldsymbol{Q}$ which drives the transition from a metal at $T > 12$ K in $(TMTSF)_2PF_6$ to an insulator exhibiting a magnetic modulation or spin density wave at the same wave vector $\boldsymbol{Q}$.

The broken symmetry of the SDW ground state, characterized by a magnetic modulation at wave vector $\boldsymbol{Q}$ in $(TMTSF)_2X$, $X = PF_6$, NO_3 etc., is responsible for the activated behavior of the conductivity through the single particle gap $2\,\Delta$. Moreover, two kinds of collective excitation arise in the SDW condensate. These are the magnon excitations, which manifest themselves in the antiferromagnetic resonance, and the phase mode of the condensate which is gapless provided the spin modulation is non-commensurate with the underlying lattice. In $(TMTSF)_2PF_6$, the wave-vector corresponding to the optimum nesting of the Fermi surface is incommensurate, as shown by various measurements sensitive to local magnetic fields. There-

fore, as long as $\boldsymbol{Q}$ is incommensurate with the underlying lattice the electrons can slide freely without any cost in energy, almost like in a superconductor. However, in real materials the interaction between a SDW (or the associated CDW) and impurities or crystal defects provides a finite pinning energy, and a threshold electric field E_T must be reached before the condensate can contribute to the conduction at zero frequency (Fig. 37). In all SDW phases of the $(TMTSF)_2X$ series E_T is restricted to the range of a few mV/cm [84]. Furthermore, the linear increase of E_T versus the density of lattice defects suggests that the phase of the spin modulation is locally adjusted at each defect site to provide the largest pinning energy (strong pinning limit).

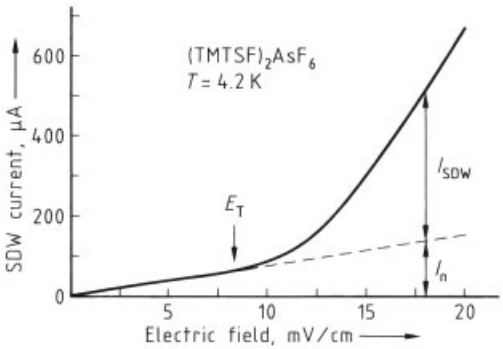

Figure 37. Current – voltage characteristics in the SDW state showing the onset of nonlinear conduction due to the collective motion of the entire electron gas above a threshold field E_T

Direct evidence for spin density wave transport is the detection of a current oscillating at a frequency proportional to the d.c. current carried collectively. The observation of such oscillations, the harmonic and sub-harmonic locking of this oscillation to an external a.c. source and a motional narrowing of the NMR spectrum in the sliding SDW state have provided firm evidence for the existence of a novel collective transport in the SDW condensate [85].

Quasi-1-D Electrons in High Magnetic Fields. The response of the Q-1-D electron gas to the application of a strong magnetic field has provided new and unexpected results, namely the transition from a Q-1-D conductor to a quasi-2-D antiferromagnetic semimetal induced by a magnetic field at low temperature. The instability of the Q-1-D electron gas at high field can be inferred from the observation of magnetoresistance oscillations above a certain threshold field in $(TMTSF)_2PF_6$ under pressure. Moreover, NMR data have shown that the ground state, which is stabilized by the magnetic field, is characterized by the existence of a spin density wave.

These phenomena are fairly well explained by a model in which, crudely speaking, the field restores enough one-dimensional character in the energy dispersion for the SDW distortion to become stable at a finite temperature [86]. This is also equivalent to the effect of negative pressure. However, the experimental data of $(TMTSF)_2PF_6$ show that the effect of the magnetic field is slightly more sophisticated. A sequence of phases must be crossed before the so-called $N = 0$ phase (which is believed to correspond to the insulating SDW ground state, which is stable at ambient pressure) can be reached above 18 T (Fig. 38 A) [87]. The integer N is related to the deviation of $\boldsymbol{Q}_x$ from $2\,k_F$, and denotes a phase containing N fully occupied Landau levels, quantized by the magnetic field. The situation $N = 0$ corresponds to zero density of states at the Fermi level, i.e., a gap which is opened over the whole Fermi surface. Phases with $N = 1, 2, 3, \ldots$, which are observed below 18 T, correspond to a sequence of semimetallic phases with a very low density of carriers (ca. 10^{-2}/unit cell). In these subphases 1, 2, or 3, ... Landau levels are completely filled. The situation where the Fermi level falls between completely filled and completely empty levels minimizes the diamagnetic energy of the 2-D carriers. This situation prevails in a finite range of magnetic fields, and the wavevector of the magnetic modulation within a given subphase can vary linearly with the field. First-order phase transitions are expected between various subphases, in agreement with the observation of hysteresis, as shown in Figure 38. The quantization of the Hall resistance at $\varrho_{xy} = h/2\,Ne^2$ (where the factor 2 comes from the spin degeneracy) is a direct consequence of the field dependence of the $\boldsymbol{Q}$ vector. The $(TMTSF)_2PF_6$ Hall data (Fig. 38 B) agree very well with the quantized nesting theory [88]

The quantization of the Hall resistance in the field-induced SDW (FISDW) phases is indeed very reminiscent of the quantum Hall ef-

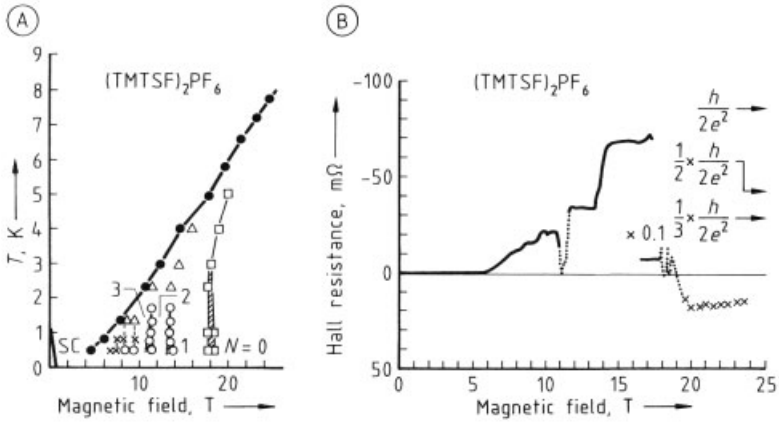

Figure 38. Temperature versus magnetic field phase diagram (A) and quantized Hall resistance (B) for $(TMTSF)_2PF_6$ under pressure

fect of the 2-D electron gas in semiconducting heterostructures. There is, however, an important difference between these two phenomena. In both cases the quantization requires a reservoir of nonconducting electronic states. This reservoir is provided by localized states in the gap between conducting Landau levels or the electron – hole (spin modulation) condensate for the 2-D electron gas and the FISDW, respectively.

6.4. Prospects for Higher T_c and Applications

Molecular superconductors are still in their infancy, but they have already demonstrated their capability to exhibit superconductivity at temperatures higher than those of any other inorganic materials apart from the high T_c cuprates. Although the mechanism for superconductivity has not been clearly identified, some guide lines have already emerged from the study of the known compounds.

As far as 1-D conductors are concerned the effectiveness of the on-site Coulombic repulsion is enhanced by the nesting properties of the Fermi surface. The resulting competition between the Cooper and Peierls channels is particularly damaging for superconductivity which cannot be stabilized above ca. 1 K in 1-D conductors derived from the TTF skeleton.

This competition is, however, less severe in 2-D materials where consequently T_c can be raised up to 12 K. Attempts to increase T_c by further decreasing the intermolecular overlap [and thus increasing $N(E_F)$] using more bulky inorganic ions have failed to provide metal-like materials, but compounds with interesting magnetic properties have been obtained.

The fullerenes, in which superconductivity is presumably mediated by high energy intraball vibrations, represent a good compromise between a weak interball interaction and the on-site Coulombic repulsion which is sufficiently small to prevent the stabilization of an insulating Wigner crystal.

The applications of the superconducting properties are still very much dependent on material preparation. However, when large areas of organic superconductors become available it may be possible to use arrays of Josephson circuits prepared by lithographic techniques to build low current devices such as high-density memories and microwave detectors.

The very narrow electronic paramagnetic resonance of some organic conductors can be used in sensitive magnetic sensors [89].

7. References

1. W. Buckel: *Superconductivity – Fundamentals and Applications,* VCH Verlagsgesellschaft, Weinheim 1991.
2. J. G. Bednorz, K. A. Müller, *Z. Phys. B: Condens. Matter* **64** (1986) 189 – 193.
3. M. N. Keene, T. J. Jackson, C. E. Gough, *Nature (London)* **340** (1989) 210 – 211.
4. J. R. Schrieffer: *Theory of Superconductivity,* W. A. Benjamin, New York 1964.

5. M. Tachiki, Y. Muto, Y. Syono (eds.): "Proceedings of the International Conference on Materials and Mechanisms of Superconductivity," *Physica C* **185 – 189** (1991).
6. C. Buchal et al., *Phys. Rev. Lett.* **50** (1983) 64 – 67.
7. N. Grewe, F. Steglich in K. A. Gschneidner, Jr., L. Eyring (eds.): *Handbook on the Physics and Chemistry of Rare Earths 14,* North Holland Publ., Amsterdam 1991, pp. 343 – 477.
8. R. L. Greene, G. B. Street, L. R. Suter, *Phys. Rev. Lett.* **34** (1975) 577 – 579.
9. A. F. Hebard et al., *Nature (London)* **350** (1991) 600 –601.
10. G. Binnig, A. Baratoff, H. E. Hoenig, J. G. Bednorz, *Phys. Rev. Lett.* **45** (1980) 1352 – 1355.
11. J. M. Heintz et al., *Z. Phys. B* **76** (1989) 303 – 309.
12. L. F. Mattheiss, E. M. Gyorgy, D. W. Johnson, *Phys. Rev. B: Condens. Matter* **37** (1988) 3745 – 3746.
13. K. Wasa et al., *J. Mater. Res.* **6** (1991) 1595 – 1604.
14. H. Weinstock, M. Nisenoff (eds.): *Superconducting Electronics,* Springer Verlag, Berlin 1988.
15. G. Costabile, S. Pagano, N. F. Pedersen, M. Russo (eds.): *Nonlinear Superconductive Electronics and Josephson Devices,* Plenum Press, New York 1990.
16. T. Ruggiero, D. A. Rudman: *Superconducting Devices,* Academic Press, Boston 1990.
17. A. Barone (ed.): *Principles and Applications of Superconducting Quantum Interference Devices,* World Scientific, Singapore 1992.
18. Y. M. Zhang, D. Winkler, T. Claeson, *4th Int. Superconductive Electronics Conf. ISEC '93,* August 11 – 14, Boulder, CO, pp. 2c-4, 9.
19. J. S. Martens et al., in [18], pp. 2d-4, 17.
20. H. G. Meyer et al., in [18], pp. 2e-9, 38.
21. J. Clarke, in [14], pp. 87 – 148.
22. D. Drung, H. Koch, in [18], pp. 3b-1, 128.
23. H. E. Hoenig et al., *Cryogenics* **29** (1989) 809 – 813.
24. J. Vrba et al., *IEEE Trans. Appl. Supercond.* **3** (1993) 1878.
25. E. Hellstrand, *Electromedica* **60** (1992) 67.
26. R. W. Ralston, D. E. Oates, A. C. Anderson, W. G. Lyons, in [18], pp. 2e-14, 48.
27. W. Prusseit, B. Utz, P. Berberich, H. Kinder, in [18], pp. 4b-1, 201.
28. *IEEE Trans. Magn.* **27** (1991) March.
29. *IEEE Trans. Microwave Theory Tech.* **39** (1991) Sept.
30. S. H. Talisa et al.: "Microwave Superconducting Filters," in [28, 29].
31. H. Chaloupka, M. Hein, G. Müller, H. Piel, Proceedings of the ESA/ESTEC Workshop on Space Applications of HTS, Nordwijk, 27 – 28 April 1993.
32. M. C. Nuss, K. W. Goosen: "Investigations of HTS with Terahertz Bandwidth Electrical Pulses," *IEEE J. Quantum Electron.* **25** (1989) 2596 – 2607.
33. G. J. Höfer, H. A. Kratz, G. Schulz, J. Söllner, V. Windte, in [18], pp. 2e-7, 34.
34. H. Föll: "Hochtemperatursupraleiter als Leitermaterial," *Mikroelektronik* **2** (1988) no. 4, 166 – 168.
35. Q. Li, D. B. Fenner, W. D. Hamblen, D. G. Hamblen, *Appl. Phys. Lett.* **62** (1993) 2428.
36. R. Gross et al.: "Characteristics of YBCO Grain Boundary Junction DC-SQUID," in [29]
37. R. W. Simon et al.: "Engineered HTS Microbridges," in [28]
38. C. L. Gia et al.: "Microstructure of Epitaxial YBCO Films of Step Edge $SrTiO_3$ Substrates," *Physica B+C* (Amsterdam) 175 (1991) 545 – 554.
39. J. Gao et al.: "Controlled Preparation of All High Tc SNS-Type Edge Junctions," *Appl. Phys. Lett.* **59** (1991) 2754 – 2756.
40. M. S. Dilorio et al., *IEEE Trans. Appl. Supercond.* **3** (1993) 2011 – 2017.
41. C. Stölzel et al., in [18], pp. 4c-25, 258.
42. M. I. Faley et al., *Appl. Phys. Lett.* **63** (1993) 2138 –2140.
43. J. S. Martens et al., in [18], pp. 2d-4, 17.
44. S. K. Tolpygo et al., *Appl. Phys. Lett.* **63** (1993) 1696.
45. R. H. Koch, W. J. Gallagher, B. Bumble, W. Y. Lee, *Appl. Phys. Lett.* **54** (1989) 951.
46. G. Friedl et al., *Appl. Phys. Lett.* **60** (1992) 3038.
47. G. Friedl et al., Proc. World Congr. Superconductivity, Sept. 14 – 18, 1992.
48. M. Kawasaki, P. Chaudhari, T. H. Newman, A. Gupta, *Appl. Phys. Lett.* **58** (1991) 2555.
49. A. H. Miklich et al., *Nature (London)* **352** (1991) 482.
50. S. Tanaka et al., *Proc. Int. Symp. Superconductivity, 6th,* 1993.
51. Y. Zhang et al., *IEEE Trans. Appl. Supercond.* **3** (1993) 2465 – 2468.
52. Y. Tavrin et al., *IEEE Trans. Appl. Supercond.* **3** (1993) 2477 – 2480.
53. D. Koelle et al., *Appl. Phys. Lett.* **68** (1993) 2271.

54. B. Roas, L. Bär, M. Kühnl, G. Daalmans, F. Bömmel, Proc. 6th Int. Symposium on Superconductivity, Hiroshima, Oct. 26 – 29, 1993.
55. C. M. Jackson, J. Raab, R. A. Davidheiser, in [18], pp. 6a-4, 341.
56. Y. Ohtani et al., CEC/ICMC '93.
57. M. Suenaga, A. F. Clark (eds.): *Filamentary A 15 Superconductors,* Plenum Press, New York 1980, pp. 17 – 34.
58. M. N. Wilson (ed.): *Superconducting Magnets,* Clarendon Press, Oxford 1983.
59. L. Intichar, D. Lambrecht: *Superconductivity in Energy Technologies,* VDI-Verlag, Düsseldorf 1990, pp. 98 – 110.
60. H. Brechna (ed.): *Superconducting Magnet Systems,* Springer Verlag, Berlin 1973.
61. K. Sato et al., *Proc. Int. Symp. Superconductivity 2nd,* 1990, 335 – 340.
62. D. F. Lee, V. Selvamanickam, K. Salama, *Physica C* **202** (1992) 83 – 96.
63. R. P. Reade, P. Berdahl, R. E. Russo, S. M. Garrison, *Appl. Phys. Lett.* **61** (1992) no. 18, 2231 – 2233.
64. P. A. Cheremnykh, *IEEE Trans. Magn.* **MAG-28** (1992) 651 – 654.
65. G. Bogner: "Superconductor Materials Science Metallurgy, Fabrication and Applications," *NATO ASI Ser. Ser. B* **68** (1981) 757 – 808.
66. E. Masada, *Proc. SPEEDAM (Symp. on Power Electronics, Electrical Devices, Advanced Electrical Motors),* Taormina, Italy, June 8 – 10, 1994, pp. 53 – 58.
67. S. Fuchino et al., *IEEE Trans. Magn.* **MAG-28** (1992) 279 – 282.
68. G. Bogner: *Superconductivity in Energy Technologies,* VDI-Verlag, Düsseldorf 1990, pp. 123 – 139.
69. R. F. Giese, Superconducting Transmission Cables, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Ill. 60439, Feb. 17, 1993.
70. H. Siebold: "Physics and Materials Science of High Temperature Superconductors," *NATO ASI Ser. Ser. E* **81** (1990) 665 – 680.
71. B. Turck, *IEEE Trans. Magn.* **MAG-25** (1989) 1473 –1480.
72. F. Negrini et al., *IEEE Trans. Magn.* **MAG-28** (1992) 390 – 393.
73. P. Komarek, *Proc. Int. Symp. Superconductivity 5th* 1992, 1199 – 1204.
74. D. Ito et al., *IEEE Trans. Magn.* **MAG-28** (1992) 438 –441.
75. W. A. Little, *Phys. Rev. A* **134** (1964) 1416.
76. K. Bechgaard, C. S. Jacobsen, K. Mortensen, H. J. Pedersen, N. Thorup, *Solid State Comm.* **33** (1980) 1119.
77. D. Jérome, *Science (Washington, D.C.)* **252** (1991) 150 924.
78. J. M. Williams et al., *Science (Washington, D.C.)* **252** (1991) 1501.
79. R. C. Haddon, *Acc. Chem. Res.* **25** (1992) 127.
80. D. Jérome, H. J. Schulz, *Adv. Phys.* **31** (1982) 299; D. Jérome in J. P. Farges (ed): *Organic Conductors,* Marcel Dekker, New York 1994, p. 405.
81. D. Jérome, A. Mazaud, M. Ribault, K. Bechgaard, *J. Phys. Lett.* **41** (1980) L-95.
82. H. Urayama et al., *Chem. Lett.* 1988 55.
83. V. Emery, R. Bruinsma, S. Barisic, *Phys. Rev. Lett.* **48** (1982) 1039.
84. S. Tomic, J. R. Cooper, K. Bechgaard, *Phys. Rev. Lett.* **62** (1989) 462.
85. G. Kriza, G. Quirion, O. Traetteberg, W. Kang, D. Jérome, *Phys. Rev. Lett.* **66** (1991) 1922.
86. L. P. Gor'kov, A. G. Lebed', *J. Phys. Lett.* **45** (1984) L 433.
87. J. R. Cooper et al., *Phys. Rev. Lett.* **63** (1989) 1984.
88. M. Héritier, G. Montambaux, P. Lederer, *J. Phys. Lett.* **45** (1984) L 943.
89. Centre National de la Recherche Scientifique, US 5 149 946 (1992).

Supercritical Extraction → Liquid-Solid Extraction
Superoxides → Peroxo Compounds, Inorganic
Superphosphate → Phosphate Fertilizers
Supraconductors → Superconductors

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Surfactants

Standard Article

Kurt Kosswig[1]

[1]Hüls AG, Marl, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a25_747
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Interfacial Phenomena
2. Overview of Surfactants
3. Properties of Aqueous Surfactant Solutions
4. Relationship between Structure and Properties of Surfactants
5. Industrial Development
6. Anionic Surfactants
6.1. Carboxylates
6.1.1. Carboxymethylated Ethoxylates
6.1.2. Amino Acid Derivatives
6.2. Sulfonates
6.2.1. Alkylbenzenesulfonates
6.2.2. Alkylnaphthalenesulfonates
6.2.3. Alkanesulfonates
6.2.4. α-Olefinsulfonates
6.2.5. α-Sulfo Fatty Acid Esters
6.2.6. Sulfosuccinates
6.2.7. Alkoxyalkane-, Acyloxyalkane-, and Acylaminoalkanesulfonates
6.3. Sulfates
6.3.1. Alkyl Sulfates
6.3.2. Ether Sulfates
6.4. Alkyl Phosphates
7. Nonionic Surfactants
7.1. General Properties
7.2. Ethoxylates
7.3. Terminally Blocked Ethoxylates
7.4. Fatty Acid Esters of Polyhydroxy Compounds
7.5. Amine Oxides
8. Cationic Surfactants
8.1. Quaternary Ammonium Compounds
8.2. Imidazoline Derivatives
9. Amphoteric Surfactants
10. Surfactants with Heteroatoms in the Hydrophobic Group
10.1. Block Copolymers of Propylene Oxide and Ethylene Oxide
10.2. Silicone-Based Surfactants
10.3. Fluorosurfactants
11. Analysis
11.1. Identification of Surfactants
11.2. Isolation and Separation
11.3. Quantitative Determination
11.4. Unspecific Additive Parameters
12. Utility Evaluation of Surfactants
13. Uses
14. Economic Aspects
15. Toxicology and Environmental Aspects
15.1. Introduction
15.2. Toxicology
15.3. Biological Degradation
15.3.1. Methods for Determining Biological Degradation
15.3.2. Biodegradation Mechanisms
15.3.3. Toxicity of Surfactants and their Metabolites to Aquatic Organisms
15.4. Preservation of Surfactants

[Top of Page]

1. Interfacial Phenomena [9-11]

Forces holding the particles of a condensed phase together (cohesive forces) become anisotropic in the phase boundary region, and their normal component becomes smaller compared to the parallel component. To simplify the description and facilitate theoretical treatment, the properties of the boundary region of a condensed phase are projected onto a two-dimensional surface (Gibbs "dividing surface"). The tensile stress resulting from the anisotropic forces in the boundary region is then termed the interfacial tension. This quantity corresponds to the reversible work required to bring particles from the volume phase to the interface during enlargement of the former, and thus corresponds to the increase in free enthalpy of the system per unit surface area:

$$\left(\frac{dG}{dF}\right)_{P,T} = \gamma$$

where *G* is the free enthalpy, *F* the surface area, and γ the interfacial tension.

The unit of interfacial tension is accordingly N · m/m^2, i.e., N/m, and is normally expressed as mN/m. The interfacial tension in most cases decreases with increasing temperature and with increasing pressure.

Interfaces exist between all contacting but immiscible phases; in current usage the term interface is understood to mean the contact region of two condensed phases, whereas the term surface refers to the boundary region between a condensed phase and a gas. Surface tensions of liquids and interfacial tensions between two liquid phases can easily be measured. The capillary rise of a liquid in a capillary tube can be used to calculate the surface tension of a liquid if the walls of the capillary tube are completely wetted by the liquid (capillary attraction).

The value of γ can be obtained directly by measurement of the force required to enlarge the surface or interface between two liquids over a specific length. A small wettable measuring plate (Wilhelmy plate) or a ring slowly withdrawn from the liquid (Lecomte du Noüy tensiometer) is used for this purpose. In the droplet-volume method the volume of a droplet released from a capillary tube, which is directly proportional to the interfacial tension, is measured. The surface tension of a liquid can also be determined by measuring the maximum pressure in air bubbles forced through a capillary tube into a liquid, this pressure being directly proportional to the surface tension. Extremely low interfacial tensions in the range from 10^{-1} to 10^{-5} mN/m can be measured with a spinning drop tensiometer, in which a droplet of the lighter phase is suspended in the heavier liquid contained in a rotating capillary [12], [13]. Also, capillary wave spectroscopy is used to measure the interfacial tension between two liquids [14].

In the equilibrium state only very small energy interactions occur between the particles at the surface of a condensed phase and those of the gaseous phase. The surface tension can therefore be regarded simply as an expression of the interactions between the particles in the surface of the condensed phase. However, at the interface between two condensed phases interactions between the particles of the two phases occur through the interface, resulting in a reduction of the interfacial tension. When examining the interactions between the particles of condensed phases a distinction has to be made between London dispersion forces and polar forces. Accordingly, the surface tension comprises components that may be attributed to pure dispersion forces (γ_d) and to polar interaction (γ_p):

$$\gamma = \gamma_d + \gamma_p$$

For example, water has a surface tension of 72.8 mN/m, composed of 21.8 mN/m for γ_d and 51.0 mN/m for γ_p.

In contrast, alkanes and also polyethylene have no surface tension attributable to polar forces, and in this case $\gamma = \gamma_d$. If one of the phases at an interface is completely nonpolar, then the interaction through the interface is restricted to dispersion forces and the interfacial tension between two condensed phases is given by Fowkes equation:

$$\gamma_{12} = \gamma_1 + \gamma_2 - 2\sqrt{\gamma_{d1}\gamma_{d2}}$$

The equation includes the polar component of the surface tension if polar interactions also act through the interface:

$$\gamma_{12} = \gamma_1 + \gamma_2 - 2\left(\sqrt{\gamma_{d1}\gamma_{d2}} + \sqrt{\gamma_{p1}\gamma_{p2}}\right)$$

If a droplet of a liquid is brought into contact with the surface of a first liquid — the two liquids being immiscible — two phenomena can occur, depending on the magnitude of γ_{12}. If γ_{12} is smaller than ($\gamma_1 - \gamma_2$), the second liquid spreads over the surface of the first liquid in the form of a thin (in the limiting case monomolecular) film. The liquid that has spread exerts a surface pressure π in the opposite direction to the surface tension γ_1 of the underlying phase, resulting in a reduced surface tension γ'_1 of this phase:

$$\gamma_1 - \pi = \gamma_1'$$

The surface tension γ'_1 can be measured by the aforementioned methods, and the surface pressure π by a Langmuir surface balance, which measures the force acting on a barrier that restricts the spreading film. If the particles of the spread liquid are sufficiently far apart, the liquid film behaves as a two-dimensional gas described by the equation of state:

$$\gamma F = RT$$

where *F* is the available surface, *R* the gas constant, and *T* the temperature. Such a surface film is elastic. If the film is compressed, a region characterized by a sharp increase in surface pressure is reached since the forces of repulsion between the particles then come into play. This region corresponds to the state of a condensed liquid. On compressing the surface film further the latter may finally break up due to furrow formation, layer stacking, etc., whereupon the surface pressure decreases.

If the interfacial tension γ_{12} between a first liquid having a surface tension γ_1 and a second liquid having a surface tension γ_2 applied as a droplet to the surface of the first liquid is greater than ($\gamma_1 - \gamma_2$), the applied droplet remains in the form of a lens in the surface of the underlying liquid (Fig. 1). The forces resulting from the two surface tensions γ_1 and γ_2, as well as from the interfacial tension γ_{12}, act at the point of contact of the two liquid phases with the gas phase. If the resultant angles α, β, and ϕ are measured and two of the tensions are known, the third can be calculated (Neumann triangle). If the underlying phase 1 is a solid and its surface is not deformed, then γ_1 and γ_2 act in opposite directions along a straight line, and the contact or wetting angle ϕ between the applied liquid and the surface of the solid is formed at the point of contact of the three phases (Fig. 2). The contact angle ϕ is given by Young's equation:

$$\cos\phi = \frac{\gamma_1 - \gamma_{12}}{\gamma_2}$$

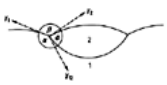

Figure 1. Lens formed by a nonspreading liquid 2 on another liquid 1 (schematic)

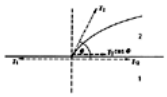

Figure 2. Formation of a contact (wetting) angle ϕ between a liquid 2 and a solid surface 1 (schematic)

The contact angle can be measured by direct observation of a droplet on a horizontally mounted plate of the solid. If the liquid is advancing the contact angle generally becomes larger than in the stationary state on account of the restricted elasticity of the liquid and the roughness of the solid surface, and if the liquid is receding the contact angle becomes smaller (Fig. 3).

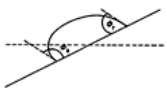

Figure 3. Contact angle ϕ of a liquid on a solid surface in the advancing state (a) and receding state (r) (diagrammatic)

The determination of the contact angle is important for evaluating the wetting of solids and for indirect determination of the surface tension γ_1 of a solid. The nonpolar and polar components of the surface tension of the solid must be determined separately by measuring the contact angles with nonpolar and polar liquids which are then added to obtain the total surface tension. The surface tension γ_1 of the solid employed in Young's equation accordingly corresponds only to the force resulting from the effective interactions (γ_{12}) between liquid and solid on the one hand, and the surface tension of the liquid (γ_2) on the other. This surface tension γ_1 therefore generally does not correspond to the true surface tension of the solid, but only to a part thereof.

The reversible work per unit area that is exchanged with the surroundings in the formation of a liquid – solid interface is the adhesion or wetting tension γ_j, which expresses the adhesion forces between liquid and solid surface:

$$\gamma_j = -\gamma_2 \cos\phi = \gamma_{12} - \gamma_1$$

If the solid is wetted by the liquid the adhesion tension is negative (ϕ < 90°), whereas if it is not wetted the tension is positive (ϕ > 90°). For example, the adhesion tension

between water and glass is negative, i.e., energy is released to the surroundings on formation of the interface, and the contact angle is < 90°. Thus water wets glass and rises in a glass capillary. In contrast, the adhesion tension between water and polyethylene is positive, i.e., work must be expended to produce a water – polyethylene interface, and the contact angle is > 90°. Hence water does not wet polyethylene completely and is displaced from a polyethylene capillary.

The wetting of a solid surface by a liquid is analogous to the spreading of a substance on a liquid surface.

The afore-described phenomena are observed for phases of a pure substance and for multicomponent phases. In the latter case, as a result of anisotropic interactions in the boundary region of a condensed phase, accumulation or depletion of components in this region may occur (positive or negative adsorption at the interface). Substances that accumulate from a liquid phase at the interface (or surface) are referred to as interface-active, usually as surface-active or, more loosely, capillary-active; substances that accumulate in the volume phase, rather than at the interface, are accordingly termed inactive.

Interfacial activity leads to a decrease of the interfacial tension compared to that of the pure substance, whereas interfacial inactivity increases the interfacial tension. Most water-soluble organic substances lower the interfacial and surface tensions of water, whereas inorganic salts increase them. For example, 20 wt % of butyric acid reduces the surface tension of water from 73 to 27 mN/m, whereas the addition of 20 wt % of water to pure butyric acid does not reduce the surface tension of butyric acid of 27 mN/m. Thus butyric acid is surface-active in water, but water is not surface-active in butyric acid. A 10 % sodium chloride solution has a surface tension that is 3 mN/m higher than that of pure water; the strongly hydrated sodium chloride ions in the volume phase are surface-inactive.

Strong accumulation of substances at the interface is termed adsorption. The adsorption of substances from a liquid phase at its interface is described by the Gibbs equation:

$$-\frac{d\gamma}{d\ \ln c} = \frac{\Gamma RT}{x}$$

where γ denotes the surface tension, c the concentration of a substance in the volume phase in mol/cm^3, Γ the excess concentration of this substance in the interface over the concentration in the volume phase in mol/cm^2, R the gas constant, T the temperature, and x is a correction factor between 1 and 2 that depends on the degree of dissociation of the adsorbed substance. For nondissociated substances $x = 1$, and for completely dissociated substances $x = 2$.

If Γ is negative, i.e., if the concentration of a dissolved substance in the interface is lower than in the volume phase, the interfacial and surface tensions increase, whereas if Γ is positive, which means a higher concentration of the dissolved substance in the interface than in the volume phase, the interfacial and surface tensions of the condensed phase decrease with increasing concentration of the dissolved substance in the volume phase. The greater the excess concentration Γ of a substance at the interface, the more strongly surface-active it is. Surfactants, also referred to as tensides, are active substances whose molecules or ions have the property that when a characteristic concentration in aqueous solution is exceeded, they associate by reversible aggregation to form larger particles, known as micelles, which impart colloidal behavior to the solution.

The reason for the large adsorption at the interface lies in the chemical structure of the adsorbed substance. Surfactants are amphiphilic substances: their molecules contain endophilic and exophilic groups. Endophilic groups are those that interact strongly with the condensed phase, whereas exophilic groups are those that interact more strongly with one another than with the volume phase, and are therefore expelled from the latter. With amphiphilic substances the exophilic group is forced out of the volume phase and the endophilic group is drawn in, an accumulation in the form of an oriented adsorption occurrs at the interface. In aqueous systems the endophilic groups are termed hydrophilic, and the exophilic groups are termed hydrophobic.

In addition to the interfacial phenomena, which can be described with the aid of thermodynamics, electrical potentials are generally produced between condensed phases in ionizable systems. By dissociation, adsorption, desorption, and exchange of ions in the interface, an electrical double layer is formed consisting of two components: a relatively rigid adsorption layer and a diffuse component that extends to a greater or lesser extent into the volume phase, depending on the ionic strength. The electrical potential of the rigid adsorption layer (Stern potential) can be determined approximately by electrophoresis or flow potential measurements (zeta potential [15]).

[Top of Page]

2. Overview of Surfactants

The highly diagrammatic tail – head model is widely used in the graphical representation of simple surfactants; the tail symbolizes the hydrophobic group, and the head the hydrophilic group.

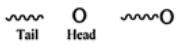

The hydrophobic groups are mainly alkyl or alkylaryl hydrocarbon groups, but fluoroalkyl, silaalkyl, thiaalkyl, oxaalkyl groups, etc., are also possible. Some important hydrophobic and hydrophilic groups are listed below:

Hydrophobic groups	Hydrophilic groups
$C_nH_{2n+1}-$	$-COO^-K^+$
$C_nH_{2n-1}-$	$-SO_3^-Na^+$
$C_nH_{2n+1}-C_6H_4-$	$-PO_3^{2-}2Na^+$
$C_nF_{2n+1}-$	$-OSO_3^-Na^+$
$CH_3-[SiO(CH_3)_2]_n-$	$-OPO_3^{2-}2Na^+$
	$-(OCH_2CH_2)_nOSO_3^-NH_4^+$
	$-(OCH_2CH_2)_n-OH$
	$-CON[(OCH_2CH_2O)_nOH]_2$
	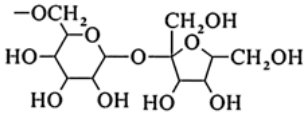
	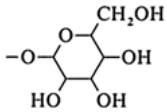
	$-N(CH_3)_2 \longrightarrow O$
	$-NR_3^+Cl^-$
	$-PR_3^+Cl^-$
	$-SR_2^+Cl^-$
	$-\overset{+}{N}(CH_3)_2CH_2COO^-$

Since surfactants mainly act in aqueous systems, it is expedient to classify surfactants according to the chemical structure of their hydrophilic groups. The hydrophilic groups may be ionic or nonionic, and their chemical structure can vary widely.

The following classification of surfactants has proved convenient (arranged in order of industrial importance).

Anionic surfactants, anionics, anion-active compounds, are amphiphilic compounds in which the hydrophobic residues carry anionic groups with small counterions such as sodium, potassium, or ammonium ions which only slightly influence the surface-active properties of the substance. Examples include soaps, alkylbenzenesulfonates, alkyl sulfates, and alkyl phosphates.

Nonionic surfactants, niosurfactants, are amphiphilic compounds that are unable to dissociate into ions in aqueous solution, for example, alkyl and alkylphenyl polyethylene

glycol ethers, fatty acid alkylolamides, sucrose fatty acid esters, alkyl polyglucosides, or trialkylamine oxides.

Cationic surfactants, cationics, cation-active compounds, are amphiphilic compounds in which the hydrophobic residues exist as cations with counterions such as chloride, sulfate, or acetate that only slightly influence the active properties of the compound. Examples include tetraalkyl ammonium chloride or *N*-alkylpyridinium chloride.

Amphoteric surfactants, amphosurfactants, (i.e., ampholytes and betaines) have zwitterionic hydrophilic groups. Examples include aminocarboxylic acids, betaines, and sulfobetaines:

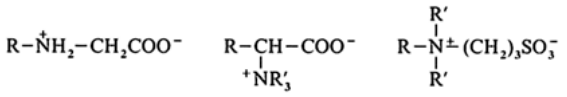

The electrical charge state of the amphosurfactants, which are often sparingly soluble at the isoelectric point, depends on the pH of the solution.

The subdivision of surfactants according to this classification is sometimes unambiguous. Nonionic surfactants may assume a cationic character in acid solution due to protonation; amine oxides are an example of this:

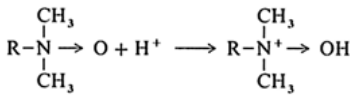

Anion – cation surfactants, in which anions and cations are amphiphilic and which are obtained by mixing anion and cation surfactants, are a special case. They are often sparingly soluble in water and in their properties resemble nonionic surfactants rather than ionic surfactants.

The structures of surfactants are not restricted to the simple tail – head model. Surfactants may contain several hydrophilic or hydrophobic groups. Examples include the sodium salts of the dialkyl sulfosuccinates:

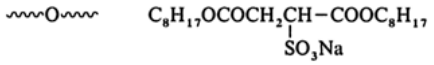

with a hydrophilic group flanked by two hydrophobic groups, or the disodium salt of 1,14-disulfatotetradecane with two hydrophilic groups at both ends of a long hydrophobic residue:

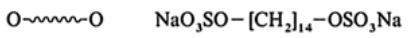

Surfactants of the latter type are also termed bola surfactants.

Industrial surfactants [8] are not pure defined substances, but mixtures of a large number of isomeric, homologous, and chemically similar substances. Low molecular mass products account for the bulk of the industrial products. Amphiphilic higher molecular mass compounds also have surface-active properties and are produced for numerous special applications. Examples include sulfated or ethoxylated condensation products of alkylphenols and formaldehyde, ethylene oxide – proylene oxide block copolymers, modified polysiloxanes, polyvinyl-*N*-alkylpyridinium salts, or copolymers of vinyl pyridine and methacrylic acid, the last-mentioned being an example of polymeric amphosurfactants. The “polysoaps” combine polymeric cations (e.g., the cation of polyvinylpyridine) with fatty acid anions.

In the theoretical treatment of surfactants the demarcation with respect to other amphiphiles that form higher ordered aggregates such as vesicles or membranes but not micelles is not always sharp [16], [17]. Substances such as polar lipids or water-soluble proteins play an extremely important role in the physiological processes of living organisms [18].

The vast majority of surfactants have hydrophobic groups derived from hydrocarbons. Only a few groups of surfactants bear hydrophobic residues containing heteroatoms. Hydrocarbon-derived surfactants are discussed here according to the nature of their hydrophilic groups (anionic, nonionic, cationic, amphoteric); the remaining surfactants are treated according to the structural features of their hydrophobic groups (polypropylene glycol derivatives, silicone-based surfactants, fluorosurfactants).

The so- called biosurfactants produced by microorganisms, are treated in [19]; for high polymer surfactants, see [20].

[Top of Page]

3. Properties of Aqueous Surfactant Solutions [21-23]

In highly dilute aqueous solutions surfactants exist in monodisperse form and are concentrated at the interfaces by hydrophilic – hydrophobic oriented adsorption. The surface concentration Γ of a surfactant can be calculated from the Gibbs equation by measuring the surface tension as a function of the volume concentration (see Chap. Interfacial Phenomena). With a rise in volume concentration the surface concentration increases until the surface is completely covered by the surfactant molecules or ions. Further surfactant particles cannot be accommodated at the surface and Gibbs' equation is no longer valid. The area requirement of a surfactant molecule or ion in the adsorption layer can be calculated from the surface concentration at maximum occupancy. Such calculations show that the surfactant molecules are adsorbed with their longitudinal axis perpendicular to the surface. The projection of the largest surface-parallel cross section of the surfactant particle onto the surface gives its area requirement.

An entropy effect is extremely important for the interfacial activity. The hydrophobic groups of the surfactant molecules or ions are hydrophobically solvated in aqueous solution. Thus the water molecules are highly ordered in the immediate vicinity of the hydrophobic species, which is associated with a decrease in entropy. If the hydrophobic group is displaced from the aqueous phase, the state of disorder of the water and thus the entropy of the system increases. This effect is also termed the hydrophobic effect. If, under total occupancy of the surface, the volume concentration of the surfactant in equilibrium with the surface concentration prevailing in this case is exceeded, the surfactant molecules or ions, which up to this concentration are mainly present in the volume phase in monodisperse form, congregate in an entropy-governed manner to form larger aggregates (micelles) in which they are oriented with the hydrophilic groups pointing towards the aqueous volume phase and the hydrophobic groups towards the interior of the micelles.

The aggregation of surfactant molecules or ions into micelles takes place in a strictly limited concentration range characteristic of each surfactant; if the surfactant concentration rises further, the number of micelles per unit volume increases, but not, however, the number of monodisperse – dissolved surfactant molecules or ions (more correctly, their activity). Since micelle formation occurs at the volume concentration of the surfactant at which the surface is largely, if not completely, covered so that the surface tension becomes independent of any increase in volume concentration, measurement of the surface tension as a function of the concentration represents a simple method of determining the so- called critical micelle concentration (CMC) at which micellization or micelle formation begins (Fig. 4).

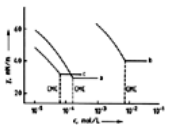

Figure 4. Surface tension Γ as a function of concentration *c*, and critical micelle concentration (CMC)

a) Nonionic surfactant (dodecylheptaglycol ether) in water; b) Ionic surfactant (dodecyl sulfate) in water; c) Dodecyl sulfate in 0.1 N sodium chloride solution

The increase in the number of particles in a surfactant solution with concentration suddenly drops at the critical micelle concentration due to micelle formation, and consequently the concentration dependence of all colligative properties such as vapor pressure and osmotic pressure, and in the case of ionic surfactants also the equivalent conductance, changes at this concentration.

Micelles are dynamic structures that are in equilibrium with the surrounding monodisperse surfactant solution; their average aggregation number fluctuates in purely aqueous solutions at the critical micelle concentration around a mean value characteristic of each surfactant. This value is between 100 and 1000 in simple nonionic surfactants, and is generally below 100 in ionic surfactants; with the latter, the electrostatic repulsion acting between the ionic head groups opposes further aggregation.

Above the CMC surfactant solutions are colloidal solutions; the surfactants form “association colloids” and behave like sols, and at higher concentrations behave like gels. This behavior is connected with the shape of the micelles that form under particular conditions. A distinction is made between spherical, rod-shaped, and disk-shaped micelles [24].

The shape of the micelle is determined primarily by the ratio of the area requirement of the head group of the surfactant under the prevailing conditions to the volume of the hydrophobic tail group. If this ratio is large, for example in the case of ionic *n*-alkyl surfactants, spherical micelles are formed. For small ratios, such as in the case of nonionic

alkyl polyglycol ethers, rod-shaped micelles are formed. However, simple ionic surfactants too can aggregate to form rod-shaped micelles if the electrostatic repulsion of the head groups is screened by adding an electrolyte and their spatial requirement on the surface of the micelle is reduced.

Surfactants with voluminous or several hydrophobic groups, as well as fluorosurfactants, form disk-shaped micelles. Some micelles as constituents of lyotropic mesophases are illustrated schematically in Figure 5.

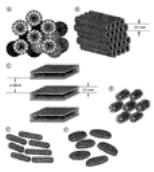

Figure 5. Structure of lyotropic mesophases formed from micellar aggregates [24] (reproduced from [6], with permission)

A) Cubic phase; B) Hexagonal phase; C) Lamellar phase; D) Inverse hexagonal phase; E) Nematic rod phase; F) Nematic disc phase

Spherical micelles can transform at higher surfactant concentrations into rod-shaped or disk-shaped micelles. Such a transition may be indicated by a second critical micelle concentration, at which the colligative properties of the solution once more change abruptly.

At higher surfactant concentrations spatially extensive, ordered structures, known as liquid crystal mesophases or lyotropic liquid crystals, are formed. In a number of nonionic surfactants of the poly(ethylene glycol) ether type such liquid crystal phases can be attributed to the formation of specific hydrates [25].

Provided they do not transform into rod-shaped or disk-shaped micelles, spherical micelles condense into cubic phases, rod-shaped micelles into hexagonal phases, and disk-shaped micelles into lamellar phases, whose basic structure resembles the bilayer lipid membranes (BLM) of living organisms. The bilayer structures present in lamellar phases transform under high shear forces (e.g., by ultrasound) into vesicles — namely spherical structures having an outer bilayer membrane (Fig. 6) [26-28]. This structure was first observed in natural fat particles (liposomes), formed from lipids.

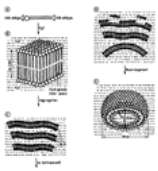

Figure 6. From bilayer lipid membranes (BLM) to vesicles

A) Amphiphilic molecule; B) BLM (section); C) Myelin structure; D) Partially broken myelin structure; E) Cross section through a vesicle (from [26], with permission)

The ordered and oriented, monolayer and multilayer adsorption of surfactants or amphiphiles leads to the formation of Langmuir – Blodgett films [29], which are important in surface treatment and coating technology.

Solutions of spherical micelles and cubic phases formed therefrom are isotropic, whereas solutions of rod-shaped and disk-shaped micelles and their liquid crystal phases are anisotropic. Due to this anisotropy the flow of surfactant sols and gels deviates from Newtonian behavior. A knowledge of this divergence from Newtonian flow is important in the preparation and formulation of surfactants, and in the design of pipelines and conveying and mixing equipment. A distinction is made between *pseudoplasticity* (shear thinning), namely decreasing in viscosity with increasing shear rate; and *dilatancy* (shear thickening), namely increasing viscosity with increasing shear.

When the mechanical stress is released pseudoplastic and dilatant liquids return to their original state.

Thixotropy is a decrease in viscosity as a result of sustained mechanical stress, and *rheopexy* an increase in viscosity as a result of sustained mechanical stress. The terms thixotropy and rheopexy imply that the liquids return to their original state, even after a time delay, when the mechanical stress is released; in practice, however, both expressions are used even if this is not the case.

Gels can also exhibit plastic behavior, acting as elastic solids below a certain critical shear stress (the yield stress or yield point) and transforming into the liquid Newtonian or non-Newtonian state only when the yield stress is exceeded.

Pseudoplasticity, thixotropy, and plasticity are often observed in surfactant solutions, whereas dilatancy and rheopexy occur less often. The flow behavior of surfactant solutions depends largely on the temperature, concentration, and on the presence of other additives, especially electrolytes. Transitions from one flow behavior to another are often produced by slight changes in the conditions, which must be taken into account when diluting or mixing surfactant solutions. The “thickening” of surfactant solutions, e.g., of alkyl polyethylene glycol ether sulfate solutions with sodium chloride is utilized in the formulation of such solutions. Conversely, the addition of a number of compounds such as urea or short- chain alkylbenzene sulfonates leads to the dissociation of the gel structure and thus decreased viscosity. Such substances are termed hydrotropics, and the effect, hydrotropy.

In addition to being viscous, many surfactant solutions at relatively high concentrations, exhibit limited elastic deformation, which is mainifested *inter alia* in vibrational phenomena. The flow behavior of such liquids is termed viscoelastic [30].

The flow behavior of surfactant solutions is measured in rotational viscometers (⟶ Rheometry). The solution is added to the gap between two concentric cylinders, one of which is at rest, while the other rotates. A shear rate gradient is generated due to the relative movement of the cylinders. The torque necessary for this movement is measured, from which the shear stress and viscosity of the solution are calculated. The cone-and-plate system is based on the same principle. A cone having an obtuse angle rotates over a plate and the liquid is subjected to shear forces in the gap between cone and plate.

Surfactant sols and gels often exhibit segregation phenomena. Coacervation is the separation of an initially uniform phase into two phases, one of which contains more of the colloidal component, and the other less. A further property of sols and gels is aging, known as deswelling or syneresis, i.e., the coagulation of gel particles and their dehydration, with the formation of a compacted gelatinous mass and a low- colloid liquid phase. To prevent such segregations, which can lead to difficulties in the storage and metering of surfactant solutions, mixing equipment must often be provided in storage vessels.

The fact that dissolved surfactants are adsorbed at interfaces and form micelles confers a number of properties on their solutions, which are described below.

Wetting. The formation of a liquid – solid interface inplace of a gas – solid interface is termed wetting. Young's equation describes the relevant relationships (see Interfacial Phenomena). A prerequisite, however, is that the surface-active substances are adsorbed at the interface so that their hydrophilic groups are directed towards the aqueous phase. If these hydrophilic groups are adsorbed on the solid surface and the hydrophobic groups project into the aqueous phase (inverse adsorption), the solid surface becomes hydrophobic, an effect that is exploited in water-repellent textiles, fabric softeners, and ore flotation, for example.

Rewetting. If a liquid, in particular a surfactant solution, displaces a liquid droplet or a solid particle insoluble in the latter from the surface of a solid, with emulsification or suspension of the displaced substrate, this is termed rewetting.

Solubilization is the apparent dissolution of substances that are insoluble or slightly soluble in water, in a surfactant solution by incorporation into the micelles. On the other hand, water can be solubilized in hydrophobic solvents by adding surfactants. Thereby, inverse micelles are formed, with the hydrophobic groups aligned outward and the hydrophilic groups inward (see Fig. 5D). Solubilization, like the decrease in surface tension and interfacial tension, is regarded as a characteristic property of surfactants in aqueous solutions.

Emulsion stabilization (see also ⟶ Emulsions) is the stabilization of disperse systems of two liquids that are insoluble or only slightly soluble in one another. This stabilization is based on covering the interfaces with surfactants that counteract coalescence of the droplets of the disperse phase. Such emulsions are metastable systems, and are also termed macroemulsions to distinguish them from the thermodynamically stable microemulsions [31-33].

Suspension stabilization is the prevention of coagulation and delay of sedimentation of solids finely dispersed in a liquid. The surfactants are solid – liquid adsorbed at the phase interface and prevent the aggregation and coagulation of the dispersed solid particles by means of steric screening and, in the case of ionic surfactants, also by electrostatic repulsion. Furthermore, under certain conditions surfactants are also able to deflocculate, without mechanical action, suspensions that have coagulated.

Protective colloid action is the suspension stabilization of hydrophobic colloids, especially against the coagulating action of inorganic salts.

Lime Soap Dispersion. A number of surfactants having relatively large or several hydrophilic groups can disperse sparingly soluble calcium and magnesium salts of fatty acids (lime soaps) in water. Examples of such surfactants include:

$$C_9H_{19} - C_6H_4 - (OCH_2CH_2)_{10} - OH$$

$C_{17}H_{35}CONH(CH_2)_3N^+(CH_3)_2(CH_2)_3SO_3^-$

$C_{10}H_{21} - C_6H_4 - SO_2NHCH_2CH_2SO_3Na$

Foams (see also ⟶ Foams and Foam Control). Surfactants accumulate at the interface of a gas bubble in aqueous solution, with their hydrophobic residues directed to the interior of the bubble. When the ascending bubble bursts through the surface of the liquid phase the film of surfactants present there becomes attached to the surface of the bubble. A lamella consisting of two films of surface-active substance is obtained, with hydrophilic groups pointing into the interior of the lamella and hydrophobic groups pointing away from the lamella. Liquid entrained from the solution is found between the films. The combination of numerous such bubbles constitutes a foam. The properties of foams are connected with the mechanical properties of the lamellae and the dynamic processes in and on the lamellae (viscosities of the surface films and interlamellar liquid, elasticity of the surface films, diffusion and spreading rate of the surface-active substances, electrical repulsion, Van der Waals attraction between the two films of the lamella, and gas permeability, leading to drying and breaking of the lamella).

The properties of surfactant solutions are the result of adsorption and aggregation of the surfactant molecules and ions; the attainment of the equilibrium state is determined by the rate of diffusion, orientation, and organization of the surfactant molecules and ions. Many properties of surfactant solutions are therefore *time-dependent.* The attainment of the equilibrium value for the surface tension of a surfactant solution is an example. Another example is the restoration force that spontaneously occurs in the expansion of a surface of a surfactant solution, as the result of a depletion in surfactant per unit area of the surface, resulting in an increase in surface tension. This increased surface tension is restored to the original value by diffusion of surfactant from the volume phase to the surface, or, in the case of spatially limited regions of increased surface tension, also by expansion of the surfactant film of the neighboring regions. The Marangoni effect, i.e., the transportation of a liquid layer adhering to the surfactant film to the site of higher surface tension, is observed in the latter case.

For the sake of convenience, the properties of the boundary region between two phases are regarded as properties of a two-dimensional surface (see Chap. Interfacial Phenomena). Although this model representation satisfactorily explains many phenomena, it does not provide a detailed understanding of the conditions in boundary regions between two phases. For example, lamellae, membranes, or multimolecular surface films should no longer be regarded as two-dimensional; rather, they form three-dimensional structures between two phases (i.e., a third phase, known as the interphase).

[Top of Page]

4. Relationship between Structure and Properties of Surfactants

The nature of the hydrophilic group of a surfactant exerts a decisive influence on the behavior of the surfactant in aqueous solution. The solubility of ionic surfactants depends on the electrolytic dissociation and solvation of the hydrophilic group, whereas the solubility of nonionic surfactants depends essentially on the hydration of the polarized but undissociated hydrophilic group. The solubility of ionic surfactants generally increases with increasing temperature, whereas that of nonionic surfactants decreases because the extent of hydration via hydrogen bond formation decreases with increasing temperature (see Figs. 7 and 8). Aqueous anionic-surfactant gels (Fig. 7) dissolve in a monodisperse fashion at low concentration, and in a micellar manner at higher concentrations. The solution temperature rises only slightly with increasing concentration, and likewise the critical micelle concentration increases only slightly with temperature. The triple point monodisperse solution – gel – micellar solution is known as the Krafft point (or temperature), and is characteristic for each ionic surfactant. It can be determined sufficiently accurately for most purposes as the point at which a surfactant solution above the critical micelle concentration becomes clear on heating. This is because above the critical micelle concentration the solution temperature (clear point) depends only slightly on the concentration.

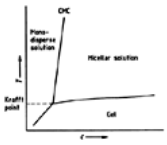

Figure 7. Phase diagram of an anionic surfactant of the sulfonate type in aqueous solution, as a function of concentration *c* and temperature *T*

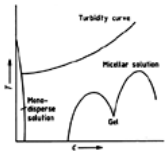

Figure 8. Phase diagram of a nonionic surfactant of the ethoxylate type in aqueous solution, as a function of concentration *c* and temperature *T*

The phase diagram of a nonionic alkyl polyethylene glycol ether is illustrated schematically in Figure 8. The critical micelle concentration decreases with increasing temperature. At elevated temperature micellar solutions separate into two phases; thus, initially clear solutions become turbid on heating. The temperature at which a solution of a nonionic surfactant becomes turbid is termed the turbidity or cloud point, and is a characteristic parameter of a nonionic surfactant.

Ionic surfactants aggregate into micelles at molar concentrations that are a factor of ten higher than in the case of nonionic surfactants. The reason for this is the electrostatic repulsion of the surfactant ions, which counteracts orientated association. In the presence of electrolytes the electrostatic repulsion of the surfactant ions is counteracted to some extent due to the increased concentration of the counterions; electrolytes thus greatly reduce the CMC of ionic surfactants and increase the aggregation number. The CMC of nonionic surfactants is only slightly influenced by electrolytes (see Fig. 4). Critical micelle concentrations of some well-defined chemical compounds are listed in Table 1, while those of some industrial products are listed in Table 2. The tables illustrate the influence of the chemical structure of the surfactants on the CMC and the surface tension:

1. With linear alkyl groups in the hydrophobic part, the amphiphilic substances require 7 – 9 carbon atoms or more in the hydrocarbon chain to exhibit surface activity and micelles formation. Only with the very highly surface-active fluorosurfactants, whose fluorocarbon residues are extremely hydrophobic, is interfacial activity observed for shorter chains, starting with perfluorobutanecarboxylic acid and perfluorobutanesulfonic acid.
2. In a series of homologues the CMC decreases with increasing size of the hydrophobic residue. For linear alkyl hydrophobic groups, Stauff's rule applies:

   $$\log(\mathrm{CMC}) = A - Bn$$

   where *A* and *B* are constants and *n* is the number of carbon atoms in the alkyl residue. Traube's rule states that in a homologous series the interfacial activity increases with increasing size of the hydrophobic group and that the surfactant concentration at which a specific interfacial activity is achieved decreases with increasing size of the hydrophobic group. For simple linear hydrocarbon chains in the hydrophobic part

   $$\log C_n = -n \log A + B$$

   where C_n is the concentration of the homologue, *n* is the number of carbon atoms in the hydrocarbon chain, and *A* and *B* are constants. For aqueous solutions of sodium salts of fatty acids, for example, the concentration required to lower the surface tension by a specific amount is three times lower when the alkyl chain is extended by one methylene group.
3. In a series of isomers the CMC and interfacial activity increase with increasing internal position of the hydrophilic group.
4. Branching of the hydrophobic hydrocarbon has the same effect on the CMC and interfacial activity as a shortening of a linear hydrocarbon residue. A central rather than terminal substitution of the alkyl chain by a phenyl group bearing the hydrophilic group has the same effect (effective chain length — see Fig. 9).
5. For nonionic surfactants of the polyethylene glycol ether type the water solubility and CMC increase with increasing length of the polyethylene glycol chain.

Table 1. CMCs of some surfactants of defined structure, and their surface tensions at the CMC

Surfactant	Temperature, °C	CMC		Surface tension,
		mmol/L	g/L	mN/m
Na 1-decyl sulfate	50	34.0	8.3	40
Na 1-dodecyl sulfate	50	8.1	2.2	38
Na 1-tetradecyl sulfate	50	2.0	0.60	37
Na 1-hexadecyl sulfate	50	0.66	0.22	38

Na 1-octadecyl sulfate	50	0.23	0.76	38
Na 2-hexadecyl sulfate	20	0.48	0.17	37
Na 4-hexadecyl sulfate	20	0.62	0.21	36
Na 6-hexadecyl sulfate	20	1.20	0.41	35
Na 8-hexadecyl sulfate	20	2.20	0.76	30
Na *p*-1-tetradecylbenzenesulfonate	75	0.67	0.25	38
Na *p*-2-tetradecylbenzenesulfonate	75	0.72	0.27	37
Na *p*-3-tetradecylbenzenesulfonate	75	0.77	0.29	36
Na *p*-5-tetradecylbenzenesulfonate	75	0.93	0.35	34
Na *p*-7-tetradecylbenzenesulfonate	75	1.73	0.65	34
Na 2-dodecene-1-sulfonate	30	13.0	3.51	36
Na 2-tetradecene-1-sulfonate	30	2.7	0.77	32
Na 2-hexadecene-1-sulfonate	30	0.61	0.81	33
Na 2-octadecene-1-sulfonate	50	0.18	0.06	31
Na 3-hydroxydodecane-1-sulfonate	30	24.8	7.15	42
Na 3-hydroxytetradecane-1-sulfonate	30	6.33	1.91	41
Na 3-hydroxyhexadecane-1-sulfonate	30	1.45	0.46	39
Na 3-hydroxyoctadecane-1-sulfonate	50	0.38	0.13	37
1-Dodecyl pentaglycol	20	0.06	0.024	31
1-Dodecyl heptaglycol	20	0.09	0.045	34
1-Dodecyl nonaglycol	20	0.12	0.070	37
1-Dodecyl dodecaglycol	20	0.17	0.121	41
Perfluoropentanecarboxylic acid	20	0.09	0.028	17
Perfluoroheptanecarboxylic acid	20	0.01	0.04	17
Perfluorononanecarboxylic acid	20	0.001	0.005	20

Table 2. CMCs of some industrial surfactants at 20 °C and their surface tensions at this concentration

Surfactant	CMC, g/L	Surface tension mN/m
Potassium oleate	0.35	25.5
Alkylbenzenesulfonate[a]	0.50	35.0
Alkanesulfonate[b]	0.44	34.5
Olefinsulfonate [c]	0.60	32.5
Fatty alcohol ether sulfate [d]	0.17	36.5
Fatty alcohol ethoxylate [e] (6 EO)[f]	0.008	30.5
Fatty alcohol ethoxylate (9 EO)	0.015	30.5
Fatty alcohol ethoxylate (12 EO)	0.020	31.5
Nonylphenol ethoxylate [g] (9 EO)	0.054	34.0
Nonylphenol ethoxylate (12 EO)	0.084	35.5
Nonylphenol ethoxylate (14 EO)	0.096	35.5
Distearyldimethylammonium chloride	2.5	35.5
Cetylpyridinium chloride	0.6	39.5

[a]Linear C_{10} – C_{13} alkylbenzenesulfonate having a mean alkyl chain length of C_{12}, as sodium salt.
[b]Linear C_{13} – C_{17} alkanesulfonate having a mean alkane chain length of C_{15}, as sodium salt.
[c]From linear C_{14}/C_{16} α-olefin, as sodium salt.
[d]From linear C_{12} – C_{14} fatty alcohol with 3 mol EO/mol, as sodium salt.
[e]From linear C_{12} – C_{14} fatty alcohol.
[f]Mol EO/mol.
[g]From alkylphenol with branched nonyl residue.

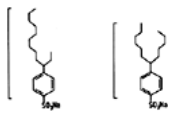

Figure 9. Effective chain length of alkylbenzenesulfonates

Micelle formation, interfacial activity, and absolute values of the interfacial tension are decisively influenced by the size and structure of the hydrophobic groups, which determine the cohesive forces between the surfactant molecules and ions as well as their spatial requirement at the interface. Cohesive forces that increase with increasing chain length exist between surfactants having unbranched alkyl residues, and are stronger than the forces between branched alkyl chains. Surfactants with linear alkyl groups accordingly associate at lower concentrations than those with branched chains, while surfactants with long chains associate at lower concentrations than those with short chains.

Fluoroalkyl groups are substantially more hydrophobic than the corresponding aliphatic alkyl groups. Fluorosurfactants therefore form micelles at much lower concentrations than their aliphatic analogues.

Surfactants with straight- chain aliphatic groups can form a tightly packed film at the interface, with parallel linear chains; their spatial requirement is small, and the strength (tension) of the film increases with increasing chain length. In contrast to branched- chain surfactants, such surfactants only reduce the surface tension slightly, though they do so already at low volume concentrations. Surfactants with branched- chain hydrophobic groups, which reach the CMC only at relatively high concentrations, lead to lower surface tensions because their surface films have a relatively low cohesion but relatively high spatial requirement. Surfactants with straight- chain aliphatic groups have been referred to as “efficient” surfactants since they are already active at low concentrations, and surfactants with branched hydrophobic residues as “effective” surfactants, since they produce a marked decrease in surface tension, albeit at higher concentrations [34] (Fig. 10).

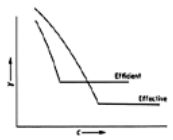

Figure 10. Surface tension γ of an efficient and an effective surfactant as a function of concentration *c*

The low cohesion between the perfluoroalkyl groups in perfluoroalkyl surfactants results in very low interfacial tensions in solution. Surfactants with methylpolysiloxanes as hydrophobic group lie between surfactants with alkyl groups, and perfluoroalkyl surfactants as regards interfacial tension, due to the moderate cohesive forces between the

dimethylsiloxane groups.

On account of the marked decrease in the interfacial and surface tension, solutions of effective surfactants generally exhibit good wetting and foaming abilities but have a low detergent power due to the high CMC; the opposite is true for efficient surfactants. This rule of thumb applies provided that the molecular masses of the surfactants are not so large that the adjustment of the equilibrium state is adversely affected by the reduced diffusion rate. Minima of the interfacial and surface tensions as a function of the length of the hydrophobic or hydrophilic (ethoxylate) residue given in the relevant literature refer to “dynamically” measured values, i.e., values that have been obtained without waiting for equilibrium to be established, which can take several hours in the case of slowly diffusing surfactants. For application technology, however, it is largely the dynamic interfacial tensions that are decisive, since in processes of wetting, washing, emulsification, etc., the instantaneous effects are of most importance.

Surfactants of varying structures form mixed films at interfaces. At low concentration surface-active substances with the higher surface activity preferentially accumulate in the interface. With increasing surfactant concentration the more strongly surface-active species may be expelled from the interface and absorbed by micelles in the volume phase, forming mixed micelles. As a result the interfacial tension, which drops with increasing surfactant concentration until the CMC is reached, may subsequently increase, reaching a constant end value (Fig. 11).

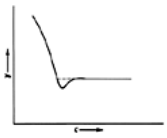

Figure 11. Surface tension γ of a contaminated surfactant as a function of concentration *c*

Such concentration-dependent interfacial tension behavior is often observed for industrial surfactants, which consist of mixtures of various chemical species. At fairly high concentrations of the second component permanent mixed films may form, whose properties deviate from the films of the individual components because of their different structure and, thus, stability. For example, lauryl alcohol in lauryl sulfate or lauric acid ethanolamide in alkylbenzene sulfonate stabilizes the foam on aqueous solutions of the corresponding surfactants because the nonionic substances are incorporated in the film and oriented in a hydrophilic – hydrophobic manner. Thus, the distance between the electrically charged hydrophilic groups of the surfactant is increased and their electrostatic repulsion is reduced.

Similar to mixed films at interfaces mixed micelles can form in solution, whose structure can be greatly altered by the mixing proportions up to the formation of “swollen” micelles, which are regarded as constituents of microemulsions.

[Top of Page]

5. Industrial Development

The industrial development of the wide range of surfactants now available received its initial and most important stimulus from the burgeoning textile industry of the mid-1800s. The various processes involved in the finishing and processing of raw cotton and wool into finished textiles required, besides the soap and natural saponins (⟶ Saponins) known from antiquity, special surface-active substances, dispersants, softeners, and sizing agents, with resistance to water hardness being one of the initial requirements. The first steps in this direction were taken by RUNGE in 1834 and by MERCER in 1846 with sulfated olive oil, followed in 1875 by turkey red oils, which were obtained by sulfating castor oil. These processes underwent many improvements, and are still of some importance.

Until 1930 industrial surfactant chemistry was almost exclusively an adjunct of the textile industry, and only in the 1930s did developments occur that opened up new markets for synthetic or semi-synthetic surfactants, including household detergents. Whereas oils and fats were formerly, with a few exceptions, the only raw materials for surfactants including soaps, during this period raw materials based on coal and petroleum also became increasingly important, although the natural materials based on oils and fats have never completely lost their importance.

In 1963 PÜSCHEL identified five periods in the development of surface-active substances [35]:

1. Up to 1925: Turkey-red oils
2. From 1925 to 1929: highly sulfonated oils
3. From 1929 to 1935: invention of the principal types of synthetic surfactants
4. From 1935 to 1945: coal and petroleum were exploited as raw material base
5. From 1945 to 1960: new, special types of surfactants were developed

This subdivision of surfactant development can be concluded by a sixth period beginning around 1960, characterized by increasing environmental awareness, which has considerable repercussions on research and technology in surfactant chemistry. Nowadays, particular attention is being paid to “mild” surfactants, i.e., those with a low irritation potential for the skin and mucous membranes.

Two developments in the 1930s in Germany have permanently affected the development of surfactant chemistry.

SCHRAUTH et al. at the Deutsche Hydrierwerke in Rodleben accomplished the catalytic high pressure hydrogenation of fatty acids to fatty alcohols. BERTSCH et al. at H. T. BÖHME in Chemnitz, who had already developed sulfated fatty acid esters (Avirol) and fatty acid amides (Humectol), obtained primary alkyl sulfates by sulfating fatty alcohols, which already appeared on the market in 1932 as a constituent of the specialty detergent FEWA, the first soap-free detergent.

In 1930, SCHÖLLER and WITTWER at I.G. Farbenindustrie in Ludwigshafen invented ethoxylation, i.e., the reaction of compounds having a dissociating hydrogen atom with ethylene oxide. This reaction leads to derivatives of poly(ethylene glycols) or also oligo(ethylene glycols), known as ethoxylates. This was the birth of a new class of surfactants, the nonionic surfactants. I.G. Farbenindustrie introduced in rapid succession the following nonionic surfactants to the market:

Fatty acid ethoxylates	Soromin SG
Fatty alcohol ethoxylates	Peregal O
Alkylphenol ethoxylates	Igepal C

Parallel to these developments, I.G. Farbenindustrie offered a range of new oleic and fatty acid derivatives:

O-acylisethionate	Igepon A
N-acyl-*N*-methyltaurine	Igepon T
N-acylsarcosine	Medialan
N-acylaminoethyl sulfate	Igepon C

The particularly dermatologically friendly condensation product formed from oleic acid and oligopeptides, Lamepon A, was developed at Chemische Fabrik Grünau.

The first cationic surfactants, *N*-acylaminoethyltrialkyl ammonium salts, Sapamines, were produced in 1927 by Ciba. In 1934 DOMAGK discovered the disinfectant action of cationic surfactants, the so- called invert soaps; I.G. Farbenindustrie developed alkyldimethylbenzyl ammonium salts under the name Zephirol. Betaines (Du Pont) and ampholytes (Th. Goldschmidt) came later, in 1937 and 1948.

In the mid-1940s Atlas Powder introduced, a new class of nonionic surfactants (Spans and Tweens) to the market with the fatty acid sorbitan esters and their ethoxylates.

The first surfactant to be based on a coal product was diisopropylnaphthalenesulfonate, which was synthesized in 1917 by GUNTHER at BASF. This compound and other short- chain dialkylnaphthalenesulfonates are still used as wetting agents under the generic name Nekal. Alkylbenzenesulfonates with longer chain alkyl groups were produced in 1933 by I.G. Farbenindustrie and, independently, in 1936 by National Aniline in the United States. The predominant position in terms of quantity that alkylbenzenesulfonates nowadays occupy was achieved, however, only after 1945. Up to then, besides soaps, alkyl sulfates and alkanesulfonates played an important role in detergents and cleansing agents. Alkyl sulfates are obtained by sulfating primary (natural) alcohols or by addition of sulfuric acid to long- chain olefins. In the latter case secondary alkyl sulfates are obtained, which were marketed as Teepol (Shell, 1934). Secondary alkanesulfonates were produced on an industrial scale from 1941 onward at I.G. Farbenindustrie in Leuna by sulfochlorination followed by saponification; they are still important nowadays as emulsifiers under the name Mersolat.

Until the expansion of refinery capacities and the conversion of chemicals production to raw materials derived from petroleum, the mainly linear hydrocarbons obtained from Fischer –Tropsch synthesis plants were used in Germany as feedstocks for the industrial production of surfactants. C_{10} – C_{14} alkenes were used for alkylating benzene to

produce alkylbenzenesulfonates, while C_{13} – C_{18} alkanes were used for the production of alkanesulfonates. The higher molecular mass hydrocarbons (Fischer – Tropsch slack wax) were oxidized to fatty acids, which were used among other things as synthetic soaps.

With the expansion of petroleum refining low molecular mass olefins became available as starting materials. Ethylene was particularly valuable as a starting material for ethylene oxide, the basic building block of nonionic surfactants, and propylene as the starting material for hydrophobic groups. Propylene trimer was used for producing isononylphenol, the starting material for nonionic surfactants, while propylene tetramer was used for producing branched- chain dodecylbenzene, the starting material for a new type of alkylbenzene sulfonates (tetrapropylbenzenesulfonate), which was primarily developed in the United States and was also widely used in Europe.

In the 1950s the alkylbenzenesulfonates, nonylphenol ethoxylates, and fatty alcohol ethoxylates were produced in such amounts that a very large proportion of surfactant demand could be met. Nevertheless, the development of special surfactants for particular purposes continued after 1950. During this period novel surfactants were synthesized, in particular by incorporating new hydrophobic groups into the surfactant molecules. Examples include the Pluronic and Tetronic types produced by Wyandotte, in which polypropylene glycol or polyoxypropylated ethylenediamine are used as hydrophobic building blocks, and also the fluorosurfactants introduced from 1950 onward by 3 M containing (per)fluoroalkyl groups, and the silicone surfactants produced by TH. GOLDSCHMIDT, containing a large variety of polydialkylsiloxanes as hydrophobic groups.

Although for a long time surfactant chemistry obtained a basic stimulus from the textile industry, nowadays the development of industrial products is largely determined by the requirements of the detergent, cleansing agent, and cosmetics industries since the bulk of surfactants is nowadays used in these sectors.

From 1960 onward the demand for biodegradability of surfactants in detergents and cleansing agents led to major changes. The most important product, tetrapropylenebenzenesulfonate, was replaced with great technical effort and expenditure by linear alkylbenzenesulfonate; in addition to nonylphenol ethoxylate, fatty alcohol ethoxylates based not only on natural raw materials but also on synthetic products appeared. A further ecological factor — the demand for a reduction or elimination of phosphate in detergents — together with the demand for a more rapid biodegradability, stimulated new technical developments: alkanesulfonates, alkenesulfonates, and α-sulfo fatty acid esters were considered as potential industrial products. Alkanesulfonates, produced by sulfoxidation of paraffins in a process developed in the 1940s by Hoechst, have in the meantime secured a firm market in Europe. The alkenesulfonates have achieved some importance in the United States and Japan, whereas the α-sulfo fatty acid esters have still not found any major use. With the expansion of oil palm plantations (e.g., in Malaysia and the Philippines), fatty alcohol sulfates based on these renewable raw materials are becoming increasingly important.

With the demand for particularly mild surfactants, besides the fatty alcohol ether sulfates and sulfosuccinic acid esters of fatty alcohol oligoglycol esters, a number of surfactants that were already developed in the 1930s have attracted renewed interest, including fatty acylisethionates and fatty acyltaurates, amino acid derivatives, ampholytes and betaines, carboxymethylated fatty alcohol ethoxylates, and so- called sugar surfactants. In parallel to the demand for mild products, there has been a more emotionally motivated wish for “natural” products, raw materials and surfactants based on renewable, preferably plant raw materials. There is thus a potential for substituting products from petrochemical raw materials by those from renewable raw materials, which will be reflected in the extent to which corresponding amounts of renewable raw materials, primarily coconut, palm, and palm kernel oils, are available at comparable prices for industrial purposes [36].

[Top of Page]

6. Anionic Surfactants

6.1. Carboxylates

Soaps, the salts of fatty acids, are still important among surfactants (⟶ Soaps). Soaps are obtained mainly by saponification of natural fats. The oxidation of paraffins, olefins, aldehydes, or alcohols (Stass reaction) is no longer of importance for the production of fatty acids. The salts of carboxylic acids occurring in petroleum (naphthenic acids), which are also surfactants, are treated in ⟶ Metallic Soaps.

6.1.1. *Carboxymethylated Ethoxylates*

The sensitivity of soaps to water hardness is a disadvantage for some applications (e.g., textile washing). In contrast, the so-called super soaps, the sodium salts of carboxymethylated ethoxylates exhibit an extreme hardness resistance combined with good water solubility.

The carboxymethylated ethoxylates or polyether carboxylates, as they have also been named, are known to be dermatologically friendly. Since they also have outstanding dispersion and emulsification properties, they are suitable for a wide range of uses, for example as surfactants in detergents and cosmetics [37], [38], and as industrial emulsifiers [39].

The carboxymethylated ethoxylates are prepared by reacting an alcohol ethoxylate or alkylphenol ethoxylate with chloroacetic acid and caustic soda:

$$\mathrm{R-O-[CH_2CH_2O]_n-H + ClCH_2COOH + 2\,NaOH}$$
$$\downarrow$$
$$\mathrm{R-O-[CH_2CH_2O]_n-CH_2COONa + NaCl + 2\,H_2O}$$

The sodium chloride byproduct is best removed by acidifying the reaction mixture, whereby the carboxylate separates as water-insoluble polyethercarboxylic acid and the sodium chloride remains in the aqueous phase. The polyether acids are clear, mobile liquids that can be transported in stainless steel or plastic containers, they are the most suitable form for handling this group of products. By neutralizing them with caustic soda colorless, highly viscous pastes are obtained that yield mobile solutions only at concentrations below 25 wt %. Typical data for some carboxymethylated products are listed in Table 3.

Table 3. Data for some commercial carboxymethylated fatty alcohol ethoxylates, R–[OCH_2CH_2]$_n$–OCH_2COOH and –OCH_2COONa, based on coconut oil alcohol (R = $C_{12}H_{25}$ / $C_{14}H_{29}$)

Property	*n* = 4		*n* = 10	
	Acid	Na salt *	Acid	Na salt *
Appearance	clear to turbid pale yellow liquid	clear to turbid pale yellow liquid	clear to slightly turbid colorless liquid	clear colorless liquid
Active substance, %	88 – 90	21 – 23	85 – 90	21 – 23
Water content, %	ca. 10	ca. 78	10 – 15	ca. 78
Degree of carboxymethylation, %	ca. 95	ca. 95	ca. 95	ca. 95
Acid number, mg KOH/g	ca. 66	0	ca. 115	0
pH (1 %)	2.5 – 3	6 – 7	2.5 – 3	6 – 7
Setting point, °C	8			
Clear point, °C	ca. 18	max. 10	ca. 11	max. 10
Density, 20 °C, g/mL	1.01	1.03	1.05	1.03
Brookfield viscosity, 20 °C, mPa · s	ca. 200	ca. 30	ca. 300	ca. 100

* CAS numbers are [33939-64-9] and [50546-32-2].

Manufacturers include Auschem, Chem-Y, and Hüls in Europe; Rhône-Poulenc Surfactant, Croda, Sandoz, Finetex, and Witco in the United States; and Nikko Chemicals in Japan.

6.1.2. *Amino Acid Derivatives*

Condensation products of fatty (oil) acids and amino acids are also less sensitive to hardness and hydrolysis than soaps. Sarcosides, for example, are obtained by reacting a fatty acid chloride with sodium sarcosinate according to the Schotten-Baumann reaction:

$$\mathrm{RCOCl + CH_3NHCH_2COONa + NaOH \rightarrow RCON(CH_3)CH_2COONa + NaCl + H_2O}$$

Sarcosides are good wetting agents and lime soap dispersants.

Industrially more important than the condensation products of amino acids are those with oligopeptides, which are obtained by partial hydrolysis of collagen, an animal skeletal

protein that has no antigenic properties. The oligopeptides have molecular masses of 200 – 600. The condensation products are prepared similarly to the sarcosides, by condensation of a fatty acid derivative (chloride or anhydride) at the terminal amino group of the oligopeptide [40], [41].

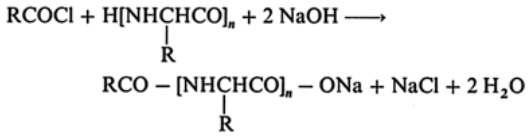

The products can be freed from byproducts by acidifying the reaction mixture; the pure condensation products separate as acids, which can be redissolved with a base.

As salts, fatty acid/protein condensation products are highly water soluble. They have a high lime soap dispersion power and foaming ability. Fatty acid/protein condensation products are preferentially used as highly skin-friendly surfactants in cosmetic preparations, bodycare agents, shampoos, and children's bathing products. They are stored and transported as 30 – 40 % solutions in steel or plastic vessels. If preserved against bacterial growth, they can be stored for prolonged periods at 20 °C and above. Typical data of two commercial products are listed in Table 4.

Table 4. Typical data for the potassium salt of a dipeptide/coconut oil acid condensate (A) and triethanolamine salt of a tetrapeptide/coconut oil acid condensation product (B)

Property	(A)	(B)
Molecular mass, g/mol	500 – 600	790 – 890
Dry content, wt %	31.5 – 32.5	39.5 – 40.5
Active substance, wt %	29.0 – 30.0	37.5 – 39.0
pH of a 10 % solution	6.2 – 7.0	6.5 – 7.0
Viscosity at 20 °C, mPa · s	1500 – 500	> 400
Density at 20 °C, g/mL	1.05 – 1.07	1.06 – 1.09
Gardner color index	max. 8	max. 10
Protein content, wt %	10 – 13	14 – 16
Ash, wt %	max. 8	max. 4

CAS Numbers. Sodium lauroyl sarcosinate [137-16-6], potassium cocoyl hydrolyzed collagen [68920-65-0], triethanolamine cocoyl hydrolyzed collagen [68952-16-9], potassium undecenoyl hydrolyzed collagen [68951-92-8], triethanolamine abietoyl hydrolyzed collagen [68918-77-4].

Producers. Sarcosides are produced by Croda, Hoechst, and ICI in Europe; Clough, Grace, and Stepan in the United States; Nikko Chemicals in Japan; and Zohar in Israel.

Fatty acid/protein condensation products are produced by Croda and Henkel in Europe; Ajinomoto, Croda, Henkel, and Inolex in the United States; and Ajinomoto and Nikko Chemicals in Japan.

6.2. Sulfonates

Sulfonates are the salts of sulfonic acids, in which a hydroxysulfonyl group is bonded via the sulfur atom directly to a carbon atom of the hydrophobic residue. This bond is thermally and chemically very stable. The strong affinity of the sulfonate group for water results in good water solubility, which decreases with increasing molecular mass. The alkaline-earth salts of sulfonic acids are less soluble in water than the alkali metal salts. The sensitivity of sulfonates to water hardness increases with increasing molecular mass. Being salts of strong acids, the alkali and alkaline-earth sulfonates give a neutral reaction in aqueous solution.

Sulfonates are nowadays the most important group among the synthetic surfactants. Besides the sulfonates discussed hereinafter, lignin sulfonates (⟶ Lignin – Commercial Lignins) and petroleum sulfonates are industrially important. Being outstanding dispersants, lignin sulfonates are used to improve the viscosity of concrete mixtures and drilling muds, while petroleum sulfonates are used mainly as oil-soluble surfactants for producing water-in-oil emulsions. Both types of sulfonates could become important as surfactants in tertiary crude oil production.

6.2.1. *Alkylbenzenesulfonates [42], [43]*

p-Alkylbenzenesulfonates having (on average) 8 – 20 carbon atoms in the alkyl chain have a wide range of uses. *p*-Alkylbenzenesulfonates with less than 6 carbon atoms in the alkyl group are not surface-active.

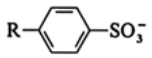

The alkylbenzenesulfonates are mainly used in the form of sodium salts, and occasionally also as potassium and ammonium salts and as salts with aliphatic amines. The octyl- to decylbenzenesulfonates have a good wetting action, but are unsuitable as emulsifiers and cleansing agents due to the shortness of the hydrophobic group. The most universally usable alkylbenzenesulfonates are those with an average of ca. 12 carbon atoms in the alkyl chain, comprising homologues with 10 – 14 carbon atoms (dodecylbenzenesulfonate). Because of their outstanding suitability as surfactants in detergents and cleansing agents (⟶ Laundry Detergents) and their low production costs compared to other surfactants, the dodecylbenzenesulfonates are nowadays the most important group of synthetic surfactants in terms of quantity. They are a constituent of powder and liquid heavy-duty household and industrial detergents. They are used less in fine detergents, shampoos, and cosmetic preparations.

Alkylbenzenesulfonates with ≥ 15 carbon atoms in the alkyl chain are sparingly soluble in water, but are readily soluble in organic media. Solutions in mineral oils play a role as drilling and cutting oils in the metal-processing industry, and as spinning, dust-binding, and batch oils in the textile industry.

Alkylbenzenesulfonates are produced by alkylation of benzene (Friedel – Crafts reaction), sulfonation of the alkylbenzene obtained, and neutralization. In the dodecylbenzenesulfonates a distinction is made between those derived from highly branched tetrapropylene (TPS = tetrapropylenebenzenesulfonate) and those with linear alkyl groups (LAS = linear alkylbenzenesulfonate). Tetrapropylenebenzenesulfonate is less soluble in water than linear dodecylbenzenesulfonate (ca. 10 % and 20 % at 20 °C), is less toxic to fish, but is a stronger skin irritant than the latter. Sodium tetrapropylenebenzenesulfonate is less hygroscopic and less plastic than linear sodium dodecylbenzenesulfonate, and accordingly can be processed more easily into powder formulations. Tetrapropylenebenzenesulfonate is, however, becoming increasingly less important because of its unsatisfactory biodegradability.

In the first stage of the alkylbenzenesulfonate synthesis, benzene is alkylated with the corresponding alkyl chlorides or olefins. Linear alkyl chlorides are obtained by chlorination of alkanes, whereby conversion is limited to at most 30 % to minimize the proportions of polychlorinated hydrocarbons. Tetrapropylene, which can be obtained by oligomerization of propylene, is already present as an olefin. Linear olefins are nowadays obtained by oligomerization of ethylene or by catalytic dehydrogenation of linear paraffin cuts.

The oligomerization of ethylene using aluminum alkyls (Alfen process) or by transition metal complexes (Shell Higher Olefin Process, SHOP) affords mixtures of homologous, even-numbered, unbranched α-olefins with a broad Schulz – Flory molecular mass distribution [44], [45] (see also ⟶ Hydrocarbons – Oligomerization of Lower Olefins). Such olefins are less suitable for producing commercial alkylbenzene mixtures, since narrow olefin cuts are required for this purpose. One possible way of narrowing the homologue distribution is to isomerize α-olefins to olefins with double bonds, as carried out in the SHOP process, followed by comproportionation or cometathesis of higher and lower olefins. Although mixtures of olefins with internal double bonds and a broad molecular mass distribution are again formed, the advantage of this process is that the desired olefin cut (e.g., C_{10} – C_{13}) can be distilled from the mixture, and the unrequired fractions can be returned to the cometathesis reaction, so that ultimately the mixture of α-olefins with a broad distribution that is obtained in the oligomerization is converted to a mixture of linear olefins of narrow distribution having internal double bonds.

The process most widely employed nowadays for producing olefins for industrial alkylbenzene synthesis is the dehydrogenation of *n*-alkanes. In conjunction with the back-integrated process for obtaining linear starting alkane from corresponding petroleum fractions (kerosene) by means of molecular sieves and the forward-integrated benzene alkylation catalyzed by hydrogen fluoride, it is nowadays carried out worldwide in numerous plants and is known as the UOP process (Universal Oil Products) [46], [47].

Nowadays, exclusively aluminum chloride and hydrogen fluoride are used as alkylation catalysts in alkylbenzene synthesis, whereby hydrogen fluoride can only be used for alkylation with olefins but not with alkyl chlorides.

A fixed-bed catalyst was described for the first time in 1990, which has advantages compared to hydrogen fluoride- catalyzed alkylation and is to be marketed by UOP under the name “Detal” [47], [48].

Alkylation always gives mixtures of isomeric (and homologous) alkylbenzenes. The isomer distribution of the alkylbenzenes obtained with the two catalysts is different. Both processes are Friedel – Crafts reactions, in which only secondary or tertiary alkyl residues occur on the benzene ring. Regardless of whether primary or secondary linear alkyl chlorides or linear olefins having terminal or internal double bonds are used as starting materials, alkylbenzenes are always obtained in which the phenyl groups is bonded internally along the alkane, but not in the terminal position.

The isomer distribution depends on the catalyst used. With aluminum chloride and linear alkyl chlorides or olefins roughly the same isomer distribution is obtained, in which 2-phenylalkanes predominate and 3-, 4-, and 5-phenylalkanes, etc., are present in lesser amounts. If benzene is alkylated in the presence of hydrogen fluoride, then 2-phenylalkanes are present in the lowest concentration in the resultant isomer mixture, and the proportion of the more internal phenylalkanes increases the closer the position of the phenyl group to the middle of the alkane chain. These relations are illustrated in Table 5 for industrial alkylbenzenes. Various isomerization reactions are responsible for these isomer distributions. Lewis acids such as aluminum chloride isomerize the phenylalkanes already formed, whereas protic acids such as hydrogen fluoride isomerize the unreacted olefins but not the phenylalkanes [49]. If branched olefins such as tetrapropylene are used as starting material, the phenyl residue is always bonded to a tertiary carbon atom, e.g.:

Table 5. Composition of industrial alkylbenzenes in wt % [50]

Component	From olefin with HF catalysis (molar mass 240)	From alkyl chloride with $AlCl_3$ catalysis (molar mass 241)	From olefin with $AlCl_3$ catalysis (molar mass 240)
Sum of *n*-alkylbenzenes	93	88	98
2-Phenylalkanes	18	29	29
Dialkyltetralins	0.5	9	0.5
Phenyldecane	14	13	14
Phenylundecane	34	30	29
Phenyldodecane	31	30	32
Phenyltridecane	20	25	24
Phenyltetradecane	1	2	1
Isomer distribution of phenyldodecanes			
1-Phenyldodecane	0	trace	0
2-Phenyldodecane	18	28	29
3-Phenyldodecane	16	19	19
4-Phenyldodecane	17	17	18
5-Phenyldodecane	24	18	18
6-Phenyldodecane	25	18	17

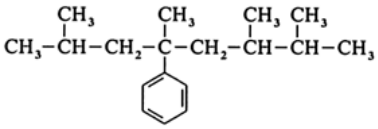

Processes for preparing linear industrial “dodecylbenzene”, a mixture of decylbenzene to tridecylbenzene, from linear alkyl chlorides and from linear olefins are described hereinafter by way of example, the preparation of the alkyl chlorides or olefins being integrated components of the processes.

Alkylation with Alkyl Chlorides (Fig. 12). The alkane, a homologous mixture of linear alkanes with average carbon number 12, is reacted with gaseous chlorine at 100 – 140 °C in a chlorination tower to a conversion of 30 mol %. To minimize backmixing and, thus, the proportion of higher chlorinated paraffins, the chlorination can also be carried out in a cascade of reactors. Suitable reactor materials are lead, silver, or enamel; iron is not suitable.

Figure 12. Preparation of a linear alkylbenzene by the aluminum chloride process

a) Storage vessel for fresh and recycled alkane; b) Chlorination tower; c) Wash towers for hydrogen chloride; d) Alkylation column; e) Separating vessel for catalyst removal; f) Water and sodium hydroxide wash; g) Dewatering column; h) Benzene column; i) Intermediate cut column; j) Paraffin column; k) Alkylbenzene column; l) Tails column

The hydrogen chloride released in the reaction $RH + Cl_2 \longrightarrow RCl + HCl$ escapes from the head of the chlorination tower and is washed in countercurrent with fresh alkane to trap unreacted chlorine. The hydrogen chloride is of high purity and can be used for further chemical syntheses. The “chlorine oil”, consisting of alkyl chlorides and unreacted paraffin, is passed to the alkylation stage. The alkyl chlorides consist of positional isomers, secondary alkyl chlorides predominating by far.

The active catalyst in the alkylation is a liquid complex compound formed in the reaction mixture and containing up to 35 wt % aluminum chloride. This complex is only slightly soluble in the reaction mixture and can, therefore, be separated as a heavier phase and recycled. Losses can be replenished by adding fresh aluminum chloride or aluminum, which reacts with hydrogen chloride to form aluminum chloride.

The alkylation is preferentially carried out in glass-lined reaction towers at 80 °C. To minimize multiple alkylation, benzene is added in a manyfold molar excess. The hydrogen chloride released in a stoichiometric amount according to the equation $RCl + C_6H_6 \longrightarrow RC_6H_5 + HCl$ maintains the reaction mixture in a turbulent state and is removed at the upper end of the reactor. It is also washed with fresh paraffin, but still contains low boiling point hydrocarbons (e.g., isobutane) and accordingly cannot be reused directly.

The reaction mixture leaving the alkylation tower is passed to a separating vessel in which the catalyst complex separates out. The organic phase is freed from residual catalyst by washing with water and sodium hydroxide and is worked up in a distillation unit. First benzene, then unreacted paraffin, and finally alkylbenzene accounts for only ca. 10 % of the alkylation mixture are separated. More highly alkylated benzenes and diphenylalkanes formed by reaction with dichloroalkanes in the “chlorine oil” can be isolated as a tail fraction.

The dichloroalkanes react also form bicyclic compounds such as 1,3-dialkyltetralins, which during distillation remain in the alkylbenzene in amounts of 5 – 10 wt %

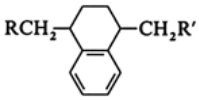

Alkylation with Olefins. The UOP process for producing alkylbenzenes, the most widely used process worldwide, is illustrated in Figure 13 [47]. The process consists of three stages: olefin production (Pacol = paraffin conversion to olefins), selective hydrogenation (DeFine), and alkylation.

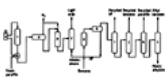

Figure 13. Preparation of a linear alkylbenzene by the UOP process

a) Heater; b) Heat exchanger; c) Dehydrogenation reactor; d) Gas – Liquid separator; e) Hydrogenation reactor; f) Stripper column; g) Alkylation reactor; h) Acid settler; i) HF stripper; j) Benzene column; k) Paraffin column; l) Rerun column

The dehydrogenation of *n*-alkanes takes place at ca. 500 °C and with a slight hydrogen excess pressure of ca. 3 bar over a fixed-bed of modified platinum catalyst on aluminum oxide.

The alkane conversion is held at 10 – 15 % to minimize further dehydrogenation to diolefins and aromatics. The hydrogen released contains small amounts of lower alkanes; some of the hydrogen is recycled to the dehydrogenation reactor, while another part is used in a connected reactor for the selective hydrogenation of diolefins, formed in small

amounts, to monoolefins (DeFine). The removal of diolefins is important to suppress the formation of byproducts,bicyclic hydrocarbons, and diphenylalkanes in the alkylation. The dehydrogenation product, containing 10 – 15 wt % monoolefins is passed to the alkylation reactor, in which it is intensively mixed with benzene and hydrofluoric acid, which is only slightly miscible with the hydrocarbons. The benzene is used in a manyfold molar excess to minimize multiple alkylation. The reaction is carried out at < 50 °C with intensive cooling. The acid catalyst is removed in a separating vessel connected to the reactor and is recycled to the alkylation reaction; a partial stream is passed through a distillation column (not shown in Fig. 13) in which the hydrofluoric acid (*bp* 19 °C) is freed from small amounts of dissolved, higher-boiling organic constituents, the so- called acid tar, a constant regeneration of the catalyst thereby being ensured. Since water is excluded from the whole alkylation section the latter can be manufactured from normal steel without risk of corrosion.

After stripping residual dissolved hydrogen fluoride in a distillation unit, the upper, organic phase from the separating vessel is fractionated into benzene, alkane, alkylbenzene, and higher-boiling compounds, the so- called heavy alkylate, which can be further fractionated. Benzene and alkane are returned to the process.

The distillation is simplified if paraffin-free olefins (e.g., tetrapropylene or linear alkenes from the SHOP process) are used for the alkylation. In the preparation of higher alkylbenzenes such as hexadecylbenzene, paraffin-free olefins are preferably used as starting materials since the distillation becomes more difficult as the boiling point rises.

Formerly, hydrofluoric acid was the exclusive catalyst for the alkylation of benzene with olefins; however, more recently a plant came on stream in which aluminum chloride is also used for this reaction [50]. The linear alkylbenzene produced in this way is reported to be particularly pure.

Sulfonation. The sulfonation of alkylbenzenes, which at low temperature mainly occurs in the *p*-position, can be performed with concentrated sulfuric acid, oleum, or sulfur trioxide. Sulfonation with sulfur trioxide is now the predominant process. It has major advantages over sulfonation with sulfuric acid or oleum, including much faster reaction rate and higher space – time yield. Also, in sulfonation with sulfur trioxide no waste sulfuric acid is formed, and accordingly there are no workup or disposal problems.

The high reaction rate, combined with strong evolution of heat (ca. 170 kJ/mol, corresponding to ca. 700 kJ/kg of dodecylbenzene), necessitates intensive mixing and efficient cooling. The sulfur trioxide used is diluted to 4 – 8 % in air or nitrogen as carrier gas. This mixture is formed in sulfur combustion combined with the contact process; it can also be obtained by passing air or nitrogen over heated oleum or liquid sulfur trioxide.

Bubble- column reactors are suitable for discontinuous sulfonation with sulfur trioxide. The mechanical momentum of the inert gas produces a high turbulence, which can be enhanced by stirrers. The heat of reaction is dissipated through the jacket, through internal cooling tubes and, in the case of rapid product circulation, by external heat exchangers. The inert gas leaving the reactor must be freed of entrained sulfur trioxide and acid droplets by washing with alkylbenzene, for example. The course of the reaction is followed by acidimetric titration, and the addition of sulfur trioxide is discontinued when the desired “titer value” is reached.

The preferentially employed continuous sulfonation with sulfur trioxide is carried out in special reactors in which backmixing of the reaction material is largely avoided and in which the reaction takes place in intensively cooled thin layers. Many modifications of suitable reactors, namely falling film, multitube, or concentric tube reactors, are available [51], [52]. Such reactors are used particularly for the sulfonation of alkenes and for sulfation of ethoxylates, since with these substrates localized overheating and excessive concentrations of sulfur trioxide can have an even more disadvantageous effect on the product quality than in the case of alkylbenzenes.

The production of alkylbenzenesulfonate is illustrated in Figure 14. The sulfur trioxide/inert gas (air or nitrogen) mixture passes via a mist eliminator, in which entrained sulfuric acid is removed, to the continuous sulfonation reactor. The sulfur trioxide is used in a slight excess (up to 10 mol %).

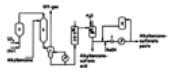

Figure 14. Continuous production of alkylbenzenesulfonate

a) Mist separator; b) Sulfonation reactor; c) Gas separator; d) Digestor; e) Hydrolyzer; f) Mixing pump; g) Intermediate vessel

The residence time of the liquid and gas in this reactor is a few minutes, and the reaction temperature is 40 – 50 °C. The reaction product passes to a separator, where the inert gas is removed. This waste gas is washed with liquid (e.g., alkylbenzene) to remove entrained acid droplets. The degassed sulfonation mixture passes through a digester where small amounts of pyrosulfonic acid present react with unreacted alkylbenzene:

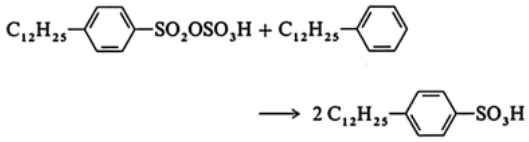

Acid anhydride, also present in small amounts, is hydrolyzed at 80 °C in a following hydrolyzer by adding 1 – 2 wt % of water:

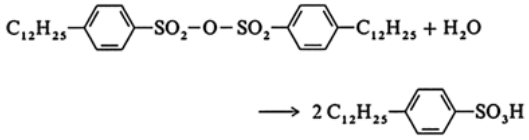

The sulfonic acid obtained in this way can then be neutralized directly, but as with sulfonic acid produced batchwise it can also be transported to the consumption site and neutralized there.

Anhydrous alkylbenzenesulfonic acids do not attack iron if water (and also atmospheric moisture) are excluded. Dodecylbenzenesulfonic acid can therefore be stored and dispatched in steel vessels. It can easily be poured, transferred, and metered as a liquid. Typical characteristic data of a high-quality dodecylbenzenesulfonic acid are given below:

Content of alkylbenzenesulfonic acid	97 – 98 %
Content of sulfuric acid	ca. 0.5 %
Content of water	0 %
Content of unsulfonated compounds	1.5 – 2 %
Titer value (mL 1 N NaOH/10 g acid)	31 – 32 mL
g Sodium salt/100 g acid	103 – 105 g
Density at 50 °C	ca. 1.03 g/mL
Viscosity at 50 °C	ca. 300 mPa · s
Color	yellow to brown

Neutralization. In the neutralization of the alkylbenzenesulfonic acid, in general only so much water is added as is necessary to form a flowable 50 % paste. The heat of neutralization is ca. 180 kJ/kg; a temperature of 60 °C is maintained by cooling. During neutralization the reaction mixture must be intensively mixed to avoid locally excessive sulfonic acid concentrations, since sulfonic acid containing 10 – 50 % water forms highly viscous gels. It has proved most convenient to add acid and sodium hydroxide continuously, while monitoring the pH, to an already neutralized mixture. In the continuous process illustrated in Figure 14 this is effected by circulating a large stream of already neutralized product and adding sulfonic acid as well as sodium hydroxide to this product stream, just in front of the mixing pump.

The brown to dark brown sulfonic acid becomes considerably lighter in color during neutralization. The light brown sulfonates obtained can be lightened by adding 1 – 3 wt % of sodium hypochlorite solution (containing 100 – 150 g of active chlorine per liter). The product then contains small amounts of chloride ions, which can cause stress corrosion cracking in steel, especially at high temperature. The apparatus for the bleaching stage must therefore be cladded with tantalum, silver, hastelloy, or earthenware.

Storage and Dispatch. Linear alkylbenzenesulfonates are water-soluble at 20 °C in an amount of up to 25 wt % to form clear solutions, and at higher concentrations pastes are formed that in the concentration range from 30 % to 60 % separate into two phases after a few hours. Such pastes must therefore be stirred during storage. The 75 % pastes do not decompose into two phases. The 65 – 75 % pastes can be dispatched only in the hot state on account of their high viscosity, and at room temperature they solidify to form rigid gels. The pastes, which can readily be pumped at 60 – 80 °C, can be stored and transported in stainless steel or glass fiber reinforced polyester vessels. They can be conveyed through heated stainless steel pipelines by using rotary piston pumps.

Anhydrous alkylbenzenesulfonates are hygroscopic and tend to agglomerate due to absorption of atmospheric moisture. Additives such as sodium sulfate or sodium toluenesulfonate counteract agglomeration; alkylbenzenesulfonates containing such additives can be worked up directly into pulverulent formulations. Anhydrous alkylbenzene

sulfonate is obtained by drying hydrated pastes, e.g., on a drum dryer; it can be packed in plastic bags and palleted. Some typical data of an industrial sodium alkylbenzene sulfonate in various concentrations are shown in Table 6.

Table 6. Typical characteristic data of an industrial sodium alkylbenzenesulfonate

Property	Concentration			
	50 %	**75 %**	**80 %**	**90 %**
Appearance at 20 °C	yellowish, flowing, inhomogeneous paste	yellowish, viscous, homogeneous paste	white to yellowish powder	white to yellowish powder
Active substance, wt %	50	75	80	90
Sodium sulfate, wt %	0.5	0.7	5	6
Sodium chloride, max. wt %	0.3	0.3	0.5	0.5
Neutral oil, max. wt %	1	1	1	1
Iron, max. ppm	5.0	5.0		
Sodium toluenesulfonate, wt %			14	
Water, wt %	remainder	remainder	1	1.5
Iodine color value (10 % solution)	1	1.5		
pH (2 % in fully deionized water)	7.5 – 8.5	7.5 – 8.5	8	8 – 9
Density, g/mL	1.06	1.06		
Bulk density, g/L			470	550

CAS Numbers. C_{10} – C_{13} alkylbenzenesulfonic acid [85536-14-7], its sodium salt [68411-30-3], its diethanolamine salt [84989-15-1], its triethanolamine salt [68411-31-4]; C_{10} – C_{16} alkylbenzenesulfonic acid [68584-22-5], its sodium salt [68081-81-2], its potassium salt [68584-27-0], its ammonium salt [68910-31-6], its calcium salt [68584-23-6], its magnesium salt [68584-26-9].

Producers. Major producers of alkylbenzenesulfonic acids and alkylbenzenesulfonates in Europe include Albright & Wilson, Berol-Nobel, Harcros-Chemicals, Hüls, ICI, Kao, Manro, Pulcra, Rhône-Poulenc, Stepan, Unger, Wibarco, Witco; in the United States, A. Harrison, Ardmore, Cromton & Knowles, Eastern Color & Chemical, Emkay Chemical, Exxon, Harcros, Henkel, ICI America, Pilot Chemicals, Rhône-Poulenc Surfactants, Stepan, Unger, Witco; in Taiwan, Taiwan Surfactant; and in Israel, Zohar.

Whereas the sulfonation of alkylbenzene and the neutralization of the sulfonic acid takes place in relatively simple plants and is accordingly even carried out by many producers for captive use (e.g., for detergents), production of the alkylbenzene is more complex. The number of producers of alkylbenzene is therefore limited. Producers of linear alkylbenzenes by hydrogen fluoride catalysis include Hüls, Enichem, and Petresa in Europa; Vista and Monsanto in North America; YPF La Plata and Deten in South America; Al Astra in Saudi Arabia; Egyptian Generalpet in Egypt; Mitsubishi Petrochemical and Nippon Petroleum in Japan; Isu Chemical in South Korea; and Formosan Union in Taiwan. Enichem and Wibarco in Europe, Vista in the United States, and Nalken in Japan also produce alkylbenzene using aluminum chloride as catalyst.

Producers of tetrapropylenebenzene include Chevron in France, Monsanto in North America, Pemex in Mexico, Mitsubishi Petrochemical and Nippon Petroleum in Japan, and Formosan Union in Taiwan.

6.2.2. *Alkylnaphthalenesulfonates*

The first synthetic surfactants based on fossil raw materials were the alkylnaphthalenesulfonates such as dipropylnaphthalene- and dibutylnaphthalenesulfonates, which still play a role as wetting agents in the textile industry. These substances are obtained in a one-pot reaction by alkylating naphthalene with isopropanol, butanol, or other alcohols in the presence of concentrated sulfuric acid as catalyst and water binder, and immediately treating the reaction product with oleum to form (di)-alkylnaphthalenesulfonic acid, in a similar manner to the sulfonation of alkylbenzene, and then separating the “waste sulfuric acid”, and finally neutralizing with sodium hydroxide.

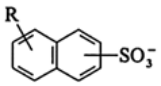

The salts of naphthalenesulfonic acid condensed with formaldehyde are chemically related to the alkylnaphthalenesulfonates, and can also be obtained in a one-pot reaction. For this purpose naphthalene is first of all sulfonated and the acid sulfonation mixture is then reacted with formaldehyde.

After the condensation the reaction product is neutralized. The molar ratio is chosen so that *n* is between 2 and 3. The salts of these condensation products are outstanding dispersants for finely divided solids, and are used as pigment dispersants and cement plasticizers.

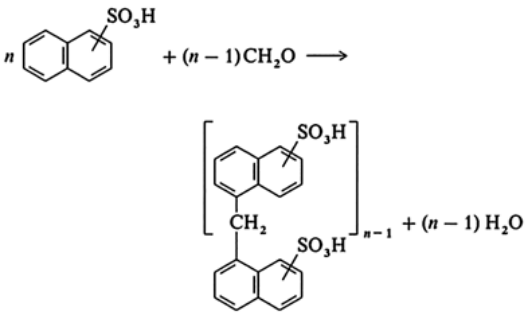

Producers of both types of products include BASF, Hoechst, Auschem, Rhône-Poulenc, and ICI in Europe; American Cyanamid, Du Pont, Emkay, Henkel, and Witco in the United States; Takemoto Oil & Fat in Japan; and Taiwan Surfactant in Taiwan.

6.2.3. *Alkanesulfonates [53-56]*

Of the large number of possible ways of synthesizing alkanesulfonates (⟶ Sulfonic Acids, Aliphatic), only sulfochlorination (the reaction of alkanes with sulfur dioxide and chlorine to form alkane sulfonyl chlorides and their saponification with sodium hydroxide) and sulfoxidation (the reaction of alkanes with sulfur dioxide and oxygen and neutralization of the sulfonic acids) are of industrial importance. The radical addition of sodium hydrogensulfite, which proceeds with satisfactory yields only in the case of terminal olefins, gives terminal alkanesulfonates, which are only slightly soluble in water and are therefore of no industrial importance [57].

In sulfochlorination and sulfoxidation, mixtures of straight- chain alkanes with 12 to 18, preferably 13 to 17 carbon atoms, are used as starting materials. Both reactions proceed appropriate by a radical- chain mechanism. In sulfoxidation a sulfoperoxy acid first of all being formed:

$$\begin{array}{lcl} R^{\bullet} + SO_2 & \longrightarrow & RSO_2^{\bullet} \\ RSO_2^{\bullet} + Cl_2 & \longrightarrow & RSO_2Cl + Cl^{\bullet} \\ Cl^{\bullet} + RH & \longrightarrow & R^{\bullet} + HCl \\ \hline RH + SO_2 + Cl_2 & \longrightarrow & RSO_2Cl + HCl \end{array}$$

$$\begin{array}{lcl} R^{\bullet} + SO_2 & \longrightarrow & RSO_2^{\bullet} \\ RSO_2^{\bullet} + O_2 & \longrightarrow & RSO_2O_2^{\bullet} \\ RSO_2O_2^{\bullet} + RH & \longrightarrow & R^{\bullet} + RSO_2O_2H \\ \hline RH + SO_2 + O_2 & \longrightarrow & RSO_2O_2H \end{array}$$

The reaction can be started by radical initiators (e.g., peroxides) or by irradiation. Ultraviolet ra-diation is used industrially. In sulfochlorination chlorine radicals are formed initially and react with an alkane:

$$RH + Cl^{\bullet} \rightarrow R^{\bullet} + HCl$$

In sulfoxidation the mechanism of the UV initiation has not been fully clarified. It is assumed that activated sulfur dioxide is formed, which reacts with the alkane:

$$RH + SO_2^{\bullet} \rightarrow R^{\bullet} + HSO_2^{\bullet}$$

Aromatics, olefins, branched alkanes, and many other substances (e.g., amines) inhibit the radical- chain reaction; the paraffins must therefore be of high purity and linearity for both processes. With high-purity alkanes quantum yields of 10 000 Einstein are achieved in sulfochlorination; termination reactions necessitate constant re-initiation (irradiation). In sulfoxidation, alkanesulfoperoxyacids are formed initially but are unstable under the reaction conditions, and decompose with the formation of additional radicals:

$$RSO_2O_2H \rightarrow RSO_3^{\bullet} + {}^{\bullet}OH$$

$$RSO_3^{\bullet} + RH \rightarrow RSO_3H + R\cdot$$

$${}^{\bullet}OH + RH \rightarrow H_2O + R^{\bullet}$$

These radicals compensate the deficit of radicals due to chain termination; being a radical- chain reaction with degenerative branching, in pure alkanes sulfoxidation accordingly takes place after a single initiation, without further radical-producing measures. The decomposition products of the peroxyacid result in a number of byproducts (esters, alcohols, discoloring substances), which adversely affect the product quality. Sulfoxidation is therefore carried out industrially with the addition of water, which together with the sulfur dioxide present in the reaction medium traps the sulfoperoxy acid before its radical-type decomposition begins:

$$RSO_2O_2H + SO_2 \rightarrow RSO_3H + H_2SO_4$$

A radical donor is lacking if the decomposition of the peroxyacid occurs by an ionic mechanism, and sulfoxidation in the presence of water therefore requires constant initiation; in the case of pure alkanes the quantum yield is ca. 10 Einstein.

In both processes substitution mainly takes place at the internal, secondary carbon atoms of the alkane. The oxysulfonyl groups are distributed randomly over the molecule. The degree of polysubstitution increases with increasing conversion. Since the interfacial activity of the sulfonates obtained after saponification decreases with increasing proportions of polysulfonates, the paraffin conversion is restricted.

Alkanesulfonates below C_8 are not surface active. C_{12} – C_{18} sodium alkanesulfonates are readily soluble in water; the solubility is increased still further by the disulfonate fraction that is always present in industrial products. On account of their proportion of species with central sulfonate groups (“effective surfactants”) the alkane sulfonates are good wetting agents. They are preferably used in liquid formulations for detergents and cleansing agents, but can also be incorporated in washing powders. Alkanesulfonates are widely used as emulsifiers, for example, in the polymerization of vinyl chloride.

Sulfochlorination Processes. Alkanesulfonates obtained from sulfochlorination processes are known as Mersolates. They are obtained in two steps:

$$RH + SO_2 + Cl_2 \rightarrow RSO_2Cl + HCl$$

$$RSO_2Cl + 2\,NaOH \rightarrow RSO_3Na + NaCl$$

In the first stage the liquid paraffin is reacted at 20 – 40 °C in chlorine- and acid-resistant reactors (earthenware, enamel, PVC) with an approximately equimolar mixture of sulfur dioxide and chlorine under irradiation from mercury vapor lamps. The waste gas consists of hydrogen chloride contaminated with sulfur dioxide; the former can be recovered as aqueous hydrochloric acid.

The paraffin conversion is adjusted to ca. 30 %; the sulfonyl chlorides then contained in the reaction mixture (Mersol 30) in amounts of ca. 30 % consist of 94 % monosulfochlorides and 6 % disulfochlorides. Mersol H, with 45 % sulfonyl chlorides in the reaction mixture already contains 16 % disulfonyl chlorides and polysulfonyl chlorides, while Mersol D, with 80 % sulfochlorides, contains ca. 40 % disulfonyl chlorides and polysulfonyl chlorides.

In the second step the degassed sulfochlorination mixture is saponified with 10 % sodium hydroxide. Elevated temperatures must be avoided since the alkane sulfonyl chloride is desulfonated to the alkyl chloride starting above ca. 80 °C, with elimination of sulfur dioxide. While hot, the neutralized product deposits part of the unreacted alkane as upper phase, and when cold deposits most of the sodium chloride derived from the saponification as an aqueous lower phase. The remaining sulfonate “glue” is freed from residual alkane in an evaporator, which together with the paraffin from the phase separation is returned to the sulfochlorination. The sulfonate is obtained as a melt flowing at 150 – 175 °C, which solidifies on cooling drums and can be removed as flakes. The flakes are hygroscopic and readily agglutinate — they can be dispersed in water and processed into pastes. The melt can, however, also be taken up directly under pressure in water. Table 7 summarizes typical data for industrial commercial products of different concentrations.

Table 7. Typical data of industrial sodium alkanesulfonates [68037-49-01] obtained by the sulfochlorination process

	Concentration		
	32 %	**62 %**	**93 %**
Appearance	solution	paste (pumpable)	colorless flakes
Active substance, wt %	32	62	93
Disulfonate and polysulfonate, wt %	5	10	14
Sodium chloride, max. wt %	2.0	3.5	4.2
Sodium sulfate	trace	trace	trace
Neutral oil, max. wt %	1.0	1.0	0.5
Iron, max. ppm	5	5	10

Half the chlorine used in the process is converted to low-grade hydrochloric acid, and half into valueless sodium chloride solution. Thus the process is economically inefficient. This disadvantage is offset by the advantage that the highly reactive alkanesulfonyl chloride intermediates can be used in the preparation of a large number of derivatives besides alkanesulfonates. For example, reaction of an alkanesulfonyl chloride with glycine yields and alkylsulfamidocarboxylic acid, which can be used to prepare effective emulsifiers and corrosion inhibitors. Esters, obtained with alcohols and phenols, are used as plasticizers for plastics.

Storage. Alkanesulfonate flakes are hygroscopic and should be processed in total. The pastes undergo segregation on standing and must be homogenized before use. The products can be stored for an unlimited time.

Sulfoxidation Processes. Figure 15 illustrates a production process for alkanesulfonates using the following reactions

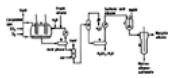

Figure 15. Production process flow sheet for alkanesulfonates using sulfoxidation

a) Photoreactor; b) Acid settler; c) Degassing column; d) Evaporation column; e) Separator for sulfuric acid; f) Neutralization vessel; g) Evaporator

$$RH + 2\,SO_2 + O_2 + H_2O \rightarrow RSO_3H + H_2SO_4$$

$$RSO_3H + NaOH \rightarrow RSO_3Na + H_2O$$

When carried out under UV irradiation with addition of water, this process is known as the light – water process.

The reactors must be resistant to acids and oxidation, and are of large volume in order to maximize the quantum yield. High-pressure 10 – 40 kW mercury vapor lamps are used. The protective tubes are made of quartz and are double-walled to absorb the thermal radiation from the lamp, with water cooling being provided between the lamp and reaction space. Since oxygen is substantially less soluble than sulfur dioxide in alkanes the reaction gas must be circulated at a high rate with a blower to maintain a sufficient supply of dissolved oxygen in the reaction mixture. This circulating gas also ensures intensive mixing of the reactor contents, which, however, can also be achieved by powerful stirrers. This is important because the aqueous and alkane phases must be constantly mixed so that the alkanesulfoperoxyacid formed initially comes into immediate contact with water and sulfur dioxide, whereby it is decomposed to the alkanesulfonic acid. Furthermore, the aqueous phase constantly extracts sulfonic acids and sulfuric acid from the alkane phase. The unusually high degree of polysubstitution — with a 1 % paraffin conversion the sulfonic acids formed already contain 10 % of disulfonic and polysulfonic acids — is due to the multiphase nature of the system. To avoid a higher proportion of disulfonic acids, the alkane conversion is limited to ca. 1 %. The mixture leaving the reactor, which consists of about 25 parts of alkane and one part of aqueous phase, passes to a separator in which the alkane forms the upper phase. The alkane is recycled through a cooler to the reactor, together with fresh alkane and process water. At 1 % conversion the amount of recycled alkane is very large, and can be used for external cooling of the reaction mixture. The reaction temperature is maintained at 20 – 40 °C.

The acid phase that separates out in the settler contains acids (sulfonic acids and sulfuric acid), water, and alkane in roughly equal amounts. This phase is freed from dissolved sulfur dioxide by degassing and is concentrated by distilling off part of the water under vacuum. The low-water- content bottom product from the evaporating column separates into two phases, the lower phase consisting of 50 to 65 % aqueous sulfuric acid, which is thereby largely removed from the sulfonic acid phase. Discolorations, which can occur in this stage due to heating (60 – 120 °C) of the sulfonic acids, can be counteracted by addition of hydrogen peroxide. The organic phase remaining after separating the sulfuric acid consists of roughly equal parts of sulfonic acids and alkane. It is neutralized with concentrated sodium hydroxide and freed from residual alkane, e.g., in a thin-layer evaporator at 200 °C in vacuo, which is recycled to the reaction. The sulfonate melt that is formed can be cooled on a rotating drum and converted to flakes, or processed with water into 60 – 65 % pastes.

As an alternative to the thermal separation of sulfuric acid, the sulfonic acids together with the alkane can be extracted from the mixture with weakly polar solvents such as alcohols, ketones, or ethers, leaving behind a 20 % aqueous sulfuric acid. The solvent must then be separated after the neutralization in an additional distillation column [58].

Some data for commercial products of various concentrations are listed in Table 8.

Table 8. Typical characteristic data of industrial sodium alkanesulfonates [85711-69-9], [7757-82-6], and [7732-18-5] obtained by the sulfoxidation process

	Concentration		
	30 %	**60 %**	**93 %**
Appearance at 25 °C	clear, almost colorless liquid	yellowish, soft paste	yellowish, waxy flakes
Active substance, wt %	30	60	93
of which disulfonate and polysulfonate, wt %	4	7	11
Sodium sulfate, max. wt %	2	4	6.5
Alkane, max. wt %	0.3	0.5	0.7
APHA color (10 % active substance in water)	40	40	50
Bulk density, g/L			500

Storage and Dispatch. The flake product can be stored for an unlimited period without undergoing any change provided moisture is excluded. The 60 % paste tends to undergo phase separation, which is prevented by stirring or circulation with a pump at 65 °C. The 30 % solution remains homogeneous and is easily pumpable.

Stainless steel, aluminum, and glass fiber reinforced polyester resin are suitable materials for storage tanks and pipelines. Normal steel is unsuitable because even very small amounts of rust seriously affect the quality.

Producers. Alkanesulfonates are produced by the sulfochlorination process by Bayer, Leuna and at Volgograd (Russia), and by the sulfoxidation process by Hoechst and Hüls.

6.2.4. α-*Olefinsulfonates [59-63]*

Olefinsulfonates are obtained by hydrolysis and neutralization of the sulfonation products of olefins. Suitable sulfonating agents include sulfur trioxide and its complexes (e.g., with dioxan), and chlorosulfonic acid; gaseous sulfur trioxide is used exclusively in industry. The fact that only α-olefins are used for the production of sulfonates is due primarily to the availability of C_{12} – C_{18} olefins of this type in the required degree of purity. Since the sulfonation takes place without double bond isomerization, α-sulfonates (α-olefinsulfonates = AOS) having good application properties are obtained from α-olefins. If olefins containing a centrally situated double bond are sulfonated, sulfonates with a centrally bonded sulfonate group are obtained, which have not yet achieved any industrial importance [64], [65].

In the sulfonation of α-olefins 1,2-sultones are probably formed initially, which then isomerize rapidly to 1,3-sultones, and more slowly to 1,4-sultones:

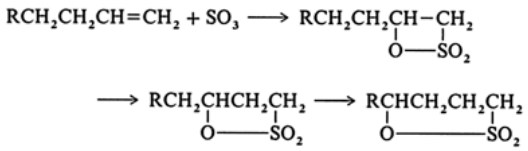

In the reaction with sulfur trioxide diluted with inert gas, at temperatures up to 40 °C, almost exclusively sultones are formed at up to 60 % olefin conversion, and at higher conversions the sultones isomerize to some extent to alkenesulfonic acids:

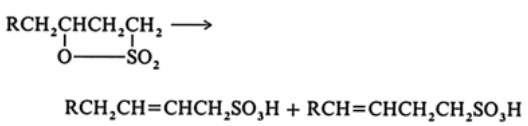

With increasing conversion side reactions such as disulfonation and oxidation also increase, leading to discoloration of the product. To convert the sultones completely to sulfonic acids or sulfonates, the sulfonation mixture must be hydrolyzed. Hydrolysis can be performed under acid or alkaline conditions. Acid hydrolysis gives predominantly alkenesulfonic acids with double bonds in the 2- and 3-positions, though compounds with more internal double bonds are also obtained; in alkaline hydrolysis about two-thirds of the reaction product consists of salts of hydroxyalkanesulfonic acids, and the remainder, salts of alkenesulfonic acids, as isomer mixtures:

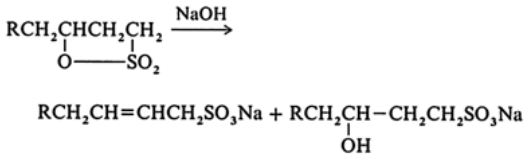

Industrial hydrolysis is carried out exclusively under alkaline conditions, mixtures of 60 to 65 % of alkenesulfonates and 40 – 35 % of hydroxyalkanesulfonates being obtained, which for the sake of simplicity are termed α-olefinsulfonates.

β-Branched α-olefins, formed by dimerizing short- chain α-olefins, are sulfonated either as such, but generally mixed with linear α-olefins. Due to the presence of a tertiary carbon atom, exclusively β-branched isomeric alkenesulfonic acids are obtained, e.g.,

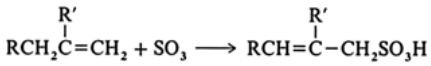

Since they have a long- chain hydrophobic residue and a terminal hydrophilic group, α-olefinsulfonates derived from linear α-olefins are efficient surfactants, with low CMCs, high detergent power, and good foaming ability in water. On account of their good water solubility sulfonates obtained from 1-dodecene to 1-hexadecene are widely used in liquid detergents and cleansing agents, in cosmetic preparations, and shampoos. Sulfonates obtained from α-hexadecene to α-octadecene can be used in powder formulations. Sulfonates obtained from β-branched α-olefins are only slightly watersoluble; they are good wetting agents, but are less suitable for detergents and cleansing agents. In industrial products they are used as a blend with linear α-olefin sulfonates in a maximum proportion of 30 %.

Production. The reaction of sulfur trioxide with olefins is rapid and strongly exothermic (ca. 180 kJ/mole; i.e., for example, 860 kJ/kg hexadecene). Local overheating, which leads to carbonization and thus brown to black products, cannot be avoided in conventional sulfonation reactors, even with intensive cooling and thorough mixing, especially if high olefin conversions are used for economic reasons.

The sulfonation of α-olefins is nowadays carried out industrially in continuous, short-time sulfonation reactors, like those used for sulfonation of alkylbenzene (see Alkylbenzenesulfonates), and for sulfating ethoxylates [51], [52]. The principle of these reactors, supplied by Allied, Ballestra, Chemithon, Lion, or Mazzoni, is to bring the substrate in the form of thin, fast-flowing and strongly cooled films or layers into contact with sulfur trioxide gas diluted with inert gas. Air or residual air from the combustion of sulfur dioxide to sulfur trioxide serves as inert gas; the concentration of sulfur trioxide is adjusted to 4 – 6 vol %.

Figure 16 illustrates the continuous production of a sodium α-olefinsulfonate. The olefin is reacted with a 10 – 20 % excess of sulfur trioxide in the reactor at as low a temperature as possible (40 °C; at lower temperature the high viscosity can cause problems). The reaction product flows into a separator, in which liquid and gas separate. The waste gas, which contains traces of sulfur trioxide and small amounts of sulfur dioxide is washed with caustic soda before being discharged to the atmosphere. The sulfonation mixture is introduced into the circulating stream of neutralized sulfonic acid and mixed with caustic soda in a pump. A partial stream is charged via a booster pump and a heater to a hydrolysis reactor where residual sultones are saponified. Hydrolysis is carried out at 150° to 250 °C with a residence time of < 1 h. If the solution still contains fairly large amounts of unsulfonated hydrocarbon, this separates as upper phase and can be removed. The hydrolysate is cooled and the pressure is released. The amount of water added in the neutralization is measured so that 30 to 40 % solutions are formed, which are bleached with hydrogen peroxide.

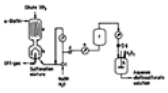

Figure 16. Continuous preparation of sodium α-olefinsulfonate

a) Sulfonation reactor; b) Gas separator; c) Mixing pump; d) Booster pump; e) Heater; f) Hydrolysis reactor; g) Pressure maintenance valve; h) Reactor for bleaching

The olefinsulfonates produced in this way contain, relative to total sulfonate, 10 – 15 % of disulfonates and small amounts of neutral oil (olefins, sulfones, oxidation products) and sodium sulfate.

Typical data of two industrial products based on C_{14}/C_{16} and C_{16}/C_{18} α-olefins are summarized in Table 9.

Table 9. Typical data of industrial sodium α-olefinsulfonates

Olefin	C_{14}/C_{16}	C_{16}/C_{18}
CAS no.	[68439-57-6]	[977067-81-4] (RD)
Appearance at 25 °C	clear, pale yellow liquid	slightly turbid, pale yellow liquid
Active substance, wt %	37	33
of which disulfonate, wt %	4	3
Sodium sulfate, wt %	1	1
Sodium chloride	trace	trace
Unsulfonated substance, wt %	1.5	1.5
Klett color index (5 % in 5 % butyl diglycol)	60	100
Viscosity, mPa · s	ca. 1000	ca. 1000

Storage and Dispatch. Olefinsulfonate solutions can be stored and dispatched in stainless steel or plastic vessels. The solutions are pumpable. Olefinsulfonates are not sensitive to heat, though overheating should be avoided. Reversible precipitation can occur below 0 °C.

Producers. α-Olefinsulfonates are produced by Akzo, Albright & Wilson, Hoechst, Rhône-Poulenc, and Witco in Europe; by Pilot Chemicals, Rhône-Poulenc Surfactant, Stepan, and Witco in the United States; and by Lion in Japan. Important producers of α-olefins, which are produced on a large scale for various purposes such as sulfonation, hydroformylation, or alkylation, are Chevron in the United States, Ethyl in Belgium and the United States, Nizhnekamsk in Russia, Shell in England and the United States, Spolana in Czechoslovakia, and Idimitsu Petrochemical in Japan.

6.2.5. α-*Sulfo Fatty Acid Esters [63], [66-70]*

The sodium salts of α-sulfo fatty acid methyl esters based on coconut, palm kernel, or tallow fatty acids are far more soluble in water than the disodium salts of α-sulfo fatty acids:

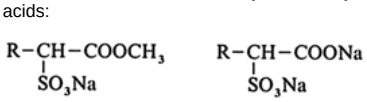

For this reason the esters are preferred. The α-sulfo fatty acid esters are stable to hydrolysis in the pH range 4 – 9 and are relatively insensitive to water hardness. They have good emulsifying and lime soap dispersing properties, and their aqueous solutions foam well and have a good cleaning ability with respect to textiles. They are accordingly used as components of soap bars and detergents.

The α-sulfo fatty acid esters are prepared by sulfonation of fatty acid methyl esters. The reaction proceeds in two stages [71]. Sulfur trioxide is rapidly absorbed by the ester at < 50 °C to give a 2 : 1 addition compound that rearranges to the α-sulfo fatty acid methyl ester at 70 – 90 °C. Sulfonation with 5 vol % of sulfur trioxide in an inert gas can be carried out in a stirred vessel cascade, in which the temperature increases from vessel to vessel. With sulfonation in a continuous short-time sulfonation reactor, a further reactor must be connected that ensures complete rearrangement at elevated temperature. Sulfonation is carried out with a 10 – 20 % excess of sulfur trioxide. The sulfonation mixtures are dark colored, and are bleached with hydrogen peroxide at 60 to 80 °C and then neutralized with sodium hydroxide. The sodium salts of the α-sulfo fatty acid esters are handled as ca. 40 % pastes or slurries.

Producers. α-Sulfo fatty acid esters are produced by Henkel and Lion.

6.2.6. *Sulfosuccinates [72], [73]*

The sodium salts of sulfosuccinate esters are obtained by reacting dialkyl maleates or salts of monoalkyl maleates with sodium hydrogensulfite in aqueous-alcoholic solution. Esterification components for dialkyl esters are mainly $C_5 – C_8$ alcohols and fatty acid ethanolamides, and for monoalkyl esters, fatty alcohols, fatty acid ethanolamides, and fatty acid ethoxylates. The sulfosuccinates thus offer a wide variety of molecular structure, which can be adapted to numerous applications. The pure sulfo esters are also termed sulfosuccinates, while the amide-group- containing products are known as sulfosuccinamates:

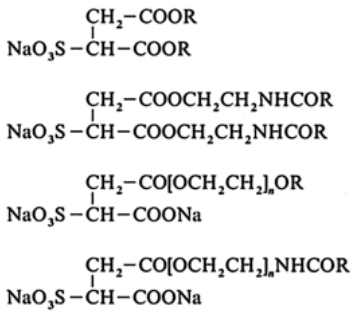

The dialkyl sulfosuccinates with short- chain alkyl groups R such as butyl, hexyl, or ethylhexyl are readily soluble in water and have outstanding wetting power (fast wetters) and dispersing properties, and are therefore used in textile processing and dyeing. They crystallize readily, like the sulfosuccinamates, and are therefore ideally suited as components of dry cleaning agents. With increasing alkyl chain length the solubility and wetting ability of sulfosuccinates decrease, while their detergent power and emulsifying power increase. The disodium salts of monosulfosuccinates are becoming increasingly important as hardness-insensitive and dermatologically compatible surfactants [74]. They are used in detergents and cleansing agents, especially for bodycare products, and as emulsifiers in cosmetics. They are also used as polymerization emulsifiers and in ore flotation. Suitable alkyl residues in these sulfosuccinates are fatty alkyl or nonylphenyl residues; the degree of ethoxylation *n* is between 3 and 12. Natural fatty acids are used as starting products for sulfosuccinamates.

A common feature of all sulfosuccinates is that they are readily saponified at the ester group. They can therefore be used only in approximately neutral media (pH 6 – 8).

Storage and Dispatch. The sodium salts of the diesterified sulfosuccinic acids can be obtained as powders by evaporating their solutions. They are generally transported as concentrated solutions in plastic drums or stainless steel vessels. Table 10 gives some typical data for three industrial products, which may contain a few percent of a solvent (isopropanol) in order to render them clear.

Table 10. Typical data of some industrial sulfosuccinates

	Sodium diethylhexyl sulfosuccinate	Disodium lauryltriglycol sulfosuccinate	Disodium lauroylaminoethyl sulfosuccinate
CAS no.	[577-11-7]	[58450-52-5]	[25882-44-4], [55101-78-5]
Apperance	clear liquid	clear liquid	clear liquid
Active substance, wt %	68 – 70	34	35
Clear point, °C	< 0	12	–2
pH (1 % active substance)	6 – 7	6 – 7	6 – 7
Klett color index (5 % active substance)	40	150	150
Density, g/cm^3	1.0	1.1	1.1
Viscosity, mPa · s	300	50	25
Flash point, °C		45	>100

Producers include Albright & Wilson, Auschem, BASF, Cyanamid, Harcros, Henkel, Hüls, Manro, Rewo, Rhône-Poulenc, Stepan, Union Carbide, Witco, and Zschimmer & Schwarz in Europe; American Cyanamide, Croda, Enkay, Harcros, Hart, Henkel, The McIntyre Group, Mona Industries, Rhône-Poulenc, Sherex, Stepan, and Union Carbide in North America; Zohar in Israel; Toho and Sanyo in Japan; and Taiwan Surfactant in Taiwan.

6.2.7. *Alkoxyalkane-, Acyloxyalkane-, and Acylaminoalkanesulfonates*

Methods of preparation of alkoxyalkanesulfonic acid salts, which would be useful for synthesizing alkylpolyglycol ether sulfonates, which in contrast to ether sulfates are resistant to hydrolysis, have not yet been carried out on an industrial scale.

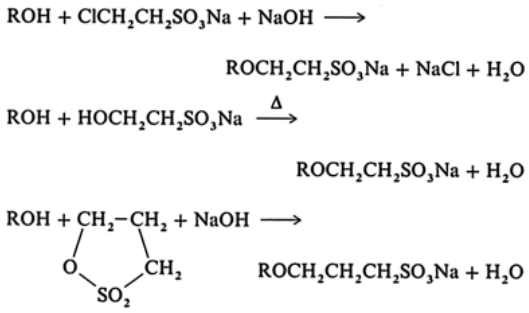

The reaction of alcohols with epichlorohydrin to form alkoxyhydroxychloropropane, which reacts with sodium sulfite to give an alkoxyhydroxypropanesulfonate, has achieved some importance:

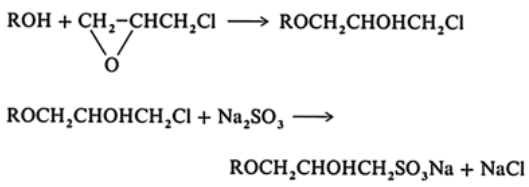

Such sulfonates are stable to water hardness and have good wetting and dispersing properties. They are dermatologically compatible and are used occasionally in the United States in body cleansing agents. Their water solubility, however, is low. Products containing $C_{16} – C_{18}$ alkyl groups are only sparingly soluble.

Due to their insensitivity to water hardness, good foaming power, soil suspending power, and dermatological compatibility, the fatty acid-derived salts of acyloxyethanesulfonic and acylaminoethanesulfonic acids are industrially fairly important (e.g., in the textile industry and in bodycare and cleansing agents). They are prepared by reacting the corresponding acid chlorides with sodium isethionate or *N*-methyltaurine.

$$RCOCl + HOCH_2CH_2SO_3Na \longrightarrow RCOOCH_2CH_2SO_3Na + HCl$$

$$RCOCl + CH_3NHCH_2CH_2SO_3Na + NaOH \longrightarrow RCON(CH_3)CH_2CH_2SO_3Na + NaCl + H_2O$$

Whereas the reaction of the isethionate with the acid chloride takes place at 100 – 120 °C without solvent and the reaction product occurs in a friable solid form after the hydrogen chloride, *N*-methyltaurine is reacted with the fatty acid chloride in aqueous solution at 20 – 30 °C. The *N*-acyl-*N*-methyltauride is obtained as a 30 –40 % solution, which can be converted into an anhydrous powder by drying.

Since the fatty acid chlorides must be synthesized from the fatty acids by reaction with phosphorus trichloride or thionyl chloride, attempts have been made to react the free acids directly with isethionate or *N*-methyltaurine at elevated temperature. Although the desired derivatives of ethanesulfonic acid are formed in 80 – 90 % yield, they are less pure than those obtained from acid chlorides.

Producers. Isethionates are produced in Europe by Hoechst, in the United States by Henkel-Emery, Hoechst-Celanese, and Rhône-Poulenc Surfactant; taurates are produced in Europe by Croda, in the United States by Finitex, Hart, Henkel-Emery, Hoechst-Celanese, Sherex, and Union Carbide, in Israel by Zohar, and in Japan by Nikko Chemicals and Nippon Oil & Fats.

6.3. Sulfates

In surfactant chemistry the term “sulfates” refers to the salts of acidic sulfuric acid esters, R–OSO_3M. In this class of compounds the sulfur atom of the hydrophilic sulfate group is bonded via an oxygen atom to the hydrophobic residue. This ester group is susceptible to hydrolysis, more strongly in acid media than in alkaline media, and is less thermally stable than the analogous sulfonates. Sulfates are prepared by sulfation of hydroxyl- containing substrates with sulfuric acid, oleum, chlorosulfonic acid, sulfuryl chloride, amidosulfonic acid, complexes of sulfur trioxide, or sulfur trioxide itself, or by addition of sulfuric acid to double bonds (⟶ Sulfonic Acids, Aliphatic). A large variety of surfaceactive sulfates is thus available, including the sulfated unsaturated fatty acids and oils that formed the starting point of surfactant chemistry and which still play a role in the textile industry [75], to the sulfates of fatty acid ethanolamides and their ethoxylates, which on account of their dermatological safety are used in cosmetics [76], and to the highly surface-active sulfated monoglycerides of fatty acids [76].

6.3.1. *Alkyl Sulfates [67], [77-79]*

The primary alkyl sulfates developed in the 1930s, based on natural alcohols obtained by hydrogenating fatty acid esters, were the first synthetic surfactants to be produced on an industrial scale. They give a neutral reaction in aqueous solution. Their physicochemical and applicational properties have been investigated intensively, especially the influence of the structure of the hydrophobic group on the above-mentioned properties has been described in detail, since individual defined substances can readily be prepared [79].

The sodium salts of alkyl sulfates prepared from primary, linear alcohols up to C_8 are scarcely surface-active. Industrially important sulfates are derived from dodecyl and tetradecyl alcohols (coconut oil and palm kernel alcohol); these sulfates are water-soluble and resistant to water hardness, and form good detergents, foaming agents, and emulsifiers. Their good dispersant properties are offset by a poor soil-suspending power. The sodium salts of the sulfates of hexadecyl and octadecyl alcohols (tallow fat alcohol) are fairly insoluble in water, especially in hard water; however, because the solubility increases markedly with increasing temperature, these sulfates are very active at elevated temperature (e.g., in textile washing).

In acidic aqueous solution (pH 3 – 4) alkyl sulfates are stable in the cold but hydrolyze on heating. They are stable in hot neutral and alkaline solution.

The primary, linear alkyl sulfates have high thermal stability, and can be spray dried at high temperature without decomposition. In contrast, the sulfates obtained from α-branched, primary and from secondary alcohols are thermally less stable, and readily decompose into sodium hydrogensulfate and alkene.

Due to the more central position of their hydrophilic groups the sulfates of highly branched alcohols obtained by the oxo or Guerbet processes or of secondary alcohols are less suitable as detergents and emulsifiers; however, in aqueous solution they have higher wetting power than the sulfates of linear primary alcohols. Secondary alkyl sulfates are no longer in use.

Production of Alkyl Sulfates. Since esterification of alcohols with sulfuric acid, followed by neutralization

$$ROH + H_2SO_4 \rightleftharpoons ROSO_3H + H_2O$$

is an equilibrium reaction the sulfuric acid must be used in excess to maximize alcohol conversion. Apart from corrosion problems caused by the water formed in the reaction, the separation of the excess sulfuric acid from the reaction mix-ture is a problem. This can be avoided if a slight excess of chlorosulfonic acid is used as sulfating agent:

$$ROH + ClSO_3H \rightarrow ROSO_3H + HCl$$

The gaseous hydrogen chloride formed is absorbed in water or sodium hydroxide. The reaction with chlorosulfonic acid must be carried out at the lowest possible temperature to avoid side reactions. Residual hydrogen chloride can be removed from the reaction mixture almost completely by air sparging.

Although sulfation with chlorosulfonic acid is still carried out in small production units, sulfation with sulfur trioxide diluted with inert gas is the method of choice for large-scale plants. In the same way as in the already discussed preparation of α-olefinsulfonates (see Section α-Olefinsulfonates), this reaction can be carried out in a continuously operating short-time sulfonation reactor [51], [52]. Sulfation is performed with equimolar amounts of sulfur trioxide; an excess promotes side reactions and discoloration of the product. Since elevated temperatures have the same effect, the reaction temperature should be below 50 °C. The reaction product leaving the reactor is degassed and neutralized as quickly as possible, preferably with sodium hydroxide:

$$ROH + SO_3 \rightarrow ROSO_3H$$

$$ROSO_3H + NaOH \rightarrow ROSO_3Na + H_2O$$

Often, the product is bleached with hydrogen peroxide.

The fatty alkyl sulfates are generally commercially available as 30 % pastes; above this concentration highly viscous pastes form, which are difficult to handle. The linear, primary alkyl sulfates crystallize well and are therefore also handled as powders that have been dried on rollers or in the spray process. Characteristic data of some industrial alkyl sulfates are listed in Table 11.

Table 11. Typical data of some industrial fatty alkyl sulfates (sodium salts)

CAS no.	[97375-27-4]	[68140-10-3]	[68955-19-1]
Appearance	paste	paste	powder
Active substance, %	30	32	93
Na_2SO_4, % max.	1.0	4.0	3.0
Nonsulfated proportion, % max.	0.5	1.8	1.0
pH	7 – 8	9 – 10	7 – 9
Color, APHA, max.	50	100	
Density, g/cm^3	1.03	0.98	0.33*

* Bulk density.

Storage and dispatch conditions correspond to those of the alkylbenzenesulfonates (see Storage and Dispatch.) or alkanesulfonates. The products must not be heated too strongly since they may be acidified further due to hydrolysis.

Producers of fatty alkyl sulfates include Akzo, Albright & Wilson, Auschem, Berol Nobel, Chem-Y, Harcros, Henkel, Hüls, Manro, Rewo, Stepan, Unger, Witco, and Zschimmer & Schwarz in Europe; Du Pont, Emkay, Henkel, Lonza, Rhône-Poulenc Surfactant, Stepan, Unger, and Witco in the United States; Daiichi, Kao, Nippon Oil & Fats in Japan; and Zohar in Israel.

Leading producers of the fatty alcohol starting materials by fatty ester hydrogenation are: Henkel, Hüls, Marchon, Oleofina, Rodleben, RWE-DEA in Europe; Procter & Gamble and Sherex in the United States; Kao and New Japan in Japan, and joint ventures of local and Western companies in the Philippines and Malaysia. Ethylene-base Ziegler alcohols are produced by RWE-DEA in Germany and by Ethyl in the United States. Important producers of oxo alcohols by hydroformylation of olefins are BASF, Enichem, Exxon, ICI, Shell in Europe, and Shell and Vista in the United States.

6.3.2. *Ether Sulfates [67], [77], [80-83]*

Ether sulfates is the short name for salts of sulfuric acid hemi-esters of alkyl or alkylaryl oligoglycol ethers of general formula

$$RO[CH_2CH_2O]_nSO_3Na$$

The oligoglycol ethers, namely ethers of oligooxyethylenes or ethoxylates, on which the ether sulfates are based are homologous mixtures where n = 0, 1, 2 . . . , etc.; the number n in the above formula specifying the mean degree of ethoxylation. Industrial ether sulfates are accordingly also homologous mixtures; with increasing degree of ethoxylation the content of alkyl sulfate (n = 0) decreases.

The oligoglycol group of the ether sulfates results in better water solubility and higher stability to water hardness, compared to sulfates containing the same alkyl or alkylaryl group. The most important surfactants of this group are the alkyl ether sulfates derived from ethoxylates of dodecanol and tetradecanol (coconut and palm kernel fatty alcohols) with a mean degree of ethoxylation of 2 – 4. Instead of natural alcohols, synthetic fatty alcohols are also used in the corresponding C- cut range. These ether sulfates have achieved considerable importance due to their good stability to hard water, dermatological compatibility, foaming and detergent power, their good emulsifying and lime soap dispersing power, their rheological behavior, and also on account of synergistic effects that they exhibit in conjunction with some other anionic surfactants (sulfonates). They are used in cosmetic preparations, foam baths, shampoos, liquid detergents, cleansing agents, especially fine detergents and rinsing agents, generally in combination with other surfactants. Being sulfuric acid hemi-esters, ether sulfates can be saponified only with difficulty in alkaline media, but are readily saponified under acid conditions. Slow saponification also occurs in neutral media, but since sodium hydrogensulfate is produced the solution becomes increasingly acidic and the saponification is thereby accelerated autocatalytically:

$$RO[CH_2CH_2O]_nSO_3Na + H_2O \rightarrow RO[CH_2CH_2O]_nH + HOSO_3Na$$

Solutions of ether sulfates are protected against this autocatalytic decomposition by adding citrate, lactate, or phosphate buffers.

A well known property of ether sulfates is the thickening effect, i.e., the increase in viscosity by several orders of magnitude of even dilute solutions on adding electrolytes. The increase in viscosity passes through a maximum, depending on the electrolyte concentration in the solution; the effect also depends on the concentration and structure of the ether sulfate (Fig. 17).

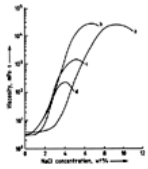

Figure 17. Thickening of 10 % aqueous solutions of fatty alcohol ether sulfates containing 3 mol EO/mol, as a function of the structure of the fatty alcohol residue

a) C_{12}/C_{14} 100 % linear; b) C_{12}/C_{15} 90 % linear; c) C_{12}/C_{15} 50 % linear; d) C_{12}/C_{15} 10 % linear

Production. Ether sulfates are obtained by sulfation of the corresponding ethoxylates and neutralization of the resulting sulfuric acid hemi-esters. Sulfuric acid is no longer used for sulfation, because it must be added in excess and remains in the product as sodium sulfate after neutralization. In smaller production units sulfation is performed with chlorosulfonic acid, as described for the preparation of primary alkyl sulfates (see Production of Alkyl Sulfates.), and the neutralized product then contains a few percent of sodium chloride. Amidosulfonic acid is occasionally used as sulfating agent, with the advantage that no byproducts or waste products are formed. In this case, however, ammonium ether sulfates are obtained:

$$RO[CH_2CH_2O]_nH + HSO_3NH_2 \rightarrow RO[CH_2CH_2O]_nSO_3NH_4$$

Amidosulfonic acid is preferably used in the sulfation of alkylphenyl ethoxylates, since in this way sulfonation of the aromatic nucleus is largely avoided.

Ether sulfates are now mainly produced by continuous sulfation of ethoxylates in short-time sulfonation reactors using gaseous sulfur trioxide, and continuous neutralization (see preparation of alkylbenzenesulfonates and olefinsulfonates, Sulfonation.). The sulfation mixture should be neutralized as quickly as possible since decomposition and disproportionation reactions occurring in the acidic reaction mixture lead to accumulation of byproducts, reduction of the content of active substance, and darkening of the color.

One of the decomposition products is 1,4-dioxane, a suspected carcinogen (⟶ Dioxane). The dioxane content can be kept below 50 ppm by precise control of the sulfation reaction (maintenance of low temperature, avoidance of excess sulfur trioxide, dilution of the sulfur trioxide gas, maintaining uniform throughput with optimum mixing and minimum residence time) as well as rapid degassing and neutralization [84], [85].

Even when the reaction is carefully controlled, the ether sulfates are not colorless. They exhibit a self-bleaching effect for several days when allowed to stand in air, presumably due to autooxidation processes. The products are normally rebleached with hydrogen peroxide or sodium hypochlorite.

Typical data of industrial sodium salts of ether sulfates are listed in Table 12.

Table 12. Typical characteristic data of some industrial ether sulfates (sodium salts)

Fatty alcohol	C_{12}/C_{14}	
Degree of ethoxylation	2	
CAS no.	[9004-82-4]	
Active substance, %	28	70
NaCl, % max.	0.3	1.2
Na_2SO_4, % max.	1.0	1.5
pH	6 – 8	6 – 8
Color, APHA, max.	75	250
Viscosity, mPa · s	300	thixotropic
Density, g/cm^3	1.05	1.1

Storage and Dispatch. Up to 30 % aqueous ether sulfate solutions are stored and transported in plastic containers or in plastic-lined vessels. Aluminum vessels are also suitable for short-term transportation. Phase separation may occur during storage in the cold, which can be prevented by stirring.

Alkyl ether sulfates in aqueous solutions of between 25 and 30 % concentration exhibit a sharp rise in viscosity from ca. 100 mPa · s to ≥ 100 000 mPa · s. This viscosity reaches a maximum value between 40 and 50 %, then decreases to 10 000 mPa · s at 60 – 70 % concentration (Fig. 18). Since pastes of this viscosity are still pumpable, 60 – 70 % products are also marketed.

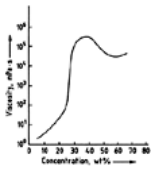

Figure 18. Viscosity of an alkyl ether sulfate as a function of its concentration [80]

The concentration-dependent viscosity behavior makes precautionary measures necessary when diluting concentrated pastes with water, so as to avoid the high-viscosity gel

state at 40 –50 % concentration. Dilute solutions are best prepared by adding the concentrated paste to water.

Producers include Akzo, Albright & Wilson, Auschem, Berol Nobel, Chem-Y, Harcros, Henkel, Hüls, Manro, Pulcra, Rewo, Rhône-Poulenc, Shell, Unger, Witco, and Zschimmer & Schwarz in Europe; Clough Chemical, Henkel, Niacet, Pilot Chemicals, Rhône-Poulenc Surfactant, Sandoz, Stepan, and Witco in North America; Kao, Lion, Marubishi Oil Chemical, Nikko Chemicals, Nippon Nyu Kazai, Nippon Oil & Fats, and Sanyo in Japan; and Zohar in Israel.

6.4. Alkyl Phosphates [86-88]

Acid phosphate esters and their salts are prepared by reacting alcohols or ethoxylates with phosphorus pentoxide:

$$3\,\mathrm{ROH} + \mathrm{P_2O_5} \rightarrow \mathrm{ROPO_3H_2} + (\mathrm{RO})_2\mathrm{PO_2H}$$

followed by neutralization, if required.

Alkyl phosphates thus generally consist of mixtures of monoalkyl and dialkyl esters of phosphoric acid. Monoalkyl, dialkyl, and trialkyl esters can specifically be prepared by stepwise reaction of a hydroxyl compound with phosphoryl chloride, followed by saponification of the resulting phosphoric ester chlorides.

Phosphoric partial esters with short- chain alkyl groups (e.g., butyl phosphoric acids) are strong acids, which have corrosion-inhibiting and bactericidal action. They are readily soluble in hard water and act as wetting and dispersing agents, and are therefore used in acid-adjusted cleansing agents, e.g., for vehicles or swimming baths.

The water solubility and acid strength of phosphoric partial esters decrease with increasing length of the alkyl chain. The sodium salts of long- chain alkyl phosphoric acids readily dissolve in water, have low sensitivity to water hardness, and are resistant to saponification, especially in alkaline media. They are good wetting agents and emulsifiers; the salts of monoalkyl phosphoric acids inhibit foam formation by other anionic or nonionic surfactants. Since the phosphates also have high dermatological compatibility, they are used both in highly alkaline cleansing agents, and in general cleansing agents, including body cleansing agents (the latter especially in Japan).

Producers of alkyl phosphates include Albright & Wilson, Auschem, BASF, Berol Nobel, Croda, Finetex, Harcros, Henkel, Hoechst, Hüls, Pulcra, Rewo, Rhône-Poulenc, Witco, and Zschimmer & Schwarz in Europe; Climax Performance, Croda, Dexter, Graden Chemical, Harcros, Hart, Henkel, Intex, Lonza, Mona Industries, Norman & Fox, Olin, PPG/Mazer, Reilly Whiteman, and Rhône-Poulenc Surfactant in North America; Kao, Nikko Chemicals, Takemoto Oil & Fat, and Yoshimura Oil Chemicals in Japan; and Taiwan Surfactant in Taiwan.

[Top of Page]

7. Nonionic Surfactants [89-91]

The polarity of covalently bonded oxygen atoms in oligoethylene glycol ethers (also termed polyglycol ethers) and oligohydroxy compounds or of oxygen atoms bonded in a semipolar manner to a heteroatom imparts water solubility and interfacial activity to hydrophobic parent structures to which such oxygen- containing hydrophilic groups are bonded. The most important nonionic surfactants are the ethoxylates, which are formally condensation products of hydrophobic alcohols, phenols, mercaptans, amines, carboxylic acids, carbonamides, etc., with oligoglycol ethers, the fatty acid esters of glycerol, diglycerol, sugars, hydrogenated sugars such as sorbitol, and alkyl(poly) glucosides. Nonionic surfactants with semipolar bonded oxygen as hydrophilic group include fatty amine oxides such as lauryldimethylamine oxide, which is used to a limited extent in liquid cleansing agents, as well as the industrially unimportant sulfoxides and phosphine oxides.

7.1. General Properties

Nonionic surfactants are colorless substances, but may be colored pale yellow to brown depending on the production process. Low molecular mass products are liquid, and with increasing molecular mass the products have a pasty to waxy consistency. In the anhydrous state, especially in the case of ethoxylates, the oligoglycol ether chains of short- chain ethoxylates are present in a stretched zig-zag form, while those of long- chain ethoxylates have a meandering structure, which is also formed in the micelles in aqueous solution.

Nonionic surfactants are less sensitive to water hardness than anionic surfactants. Their aqueous solutions foam less strongly than those of many anionic surfactants. The possible applications of nonionic surfactants are universal, given the large variability of their structure and thus of their properties. The differences between the individual types of nonionic surfactants are slight, and the choice is primarily governed having regard to the costs of special properties, e.g., effectiveness and efficiency, toxicity, dermatological compatibility and biodegradability, or permission for use in foodstuffs. The solubility of nonionic surfactants in water results from hydration of the oxygen groups by hydrogen bonding. The degree of hydration decreases with increasing temperature, and the water solubility of nonionic surfactants therefore decreases with increasing temperature. In the case of surfactants that dissolve in water to give a clear solution, i.e., ethoxylates with a fairly high degree of ethoxylation, turbidity and separation of a surfactant phase that is immiscible with water occur at a specific temperature characteristic of the surfactant (cloud point).

The cloud point can be determined for water-insoluble ethoxylates by dissolving the surfactant in aqueous butyl diglycol. For highly water soluble ethoxylates whose solutions do not become turbid even on boiling, the cloud point can be determined in aqueous sodium chloride solution. Figure 19 shows the cloud point of industrial fatty alcohol ethoxylates as a function of the degree of ethoxylation in 25 % butyl diglycol solution, in fully deionized water, and in 10 % sodium chloride solution.

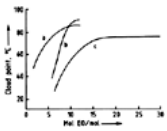

Figure 19. Cloud point of ethoxylates of a C_{12}/C_{14} fatty alcohol (linear) as a function of the degree of ethoxylation (mol EO/mol)

a) 10 % solution in 25 % butyl diglycol; b) 2 % solution in demineralized water; c) 2 % solution in 10 % NaCl

Correlations exist between the cloud points and the hydrophilic – lipophilic balance (HLB) values or phase inversion temperatures of emulsions with the corresponding surfactants as emulsifiers [92].

The concept of an HLB is of practical value [93].

The HLB system enables nonionic surfactants to be arranged on a value scale from 0 to 20. In the ideal case the HLB value of a surfactant is defined as the ratio of the molecular mass of the hydrophilic fraction in the molecule M_h to the total molecular mass M of the surfactant, multiplied by 20:

$$\mathrm{HLB} = 20 \cdot \frac{M_h}{M}$$

The HLB scale allows rough classification of nonionic surfactants according to their solubility in water and their possible areas of application. Table 13 provides a summary.

Table 13. HLB values of nonionic compounds

HLB range	Behavior in water	Example of use	Example	HLB value
0 – 3	Insoluble	defoaming agent, dispersant for solids in oil, co-emulsifier, refatting agent	Glycerol trioleate Glycerol dioleate Glycerol monooleate	0.8 1.5 2.7
3 – 6	Insoluble, dispersible	water-in-oil emulsions, co-emulsifier	Glycerol monostearate	3.8
			Propylene glycol monolaurate	4.5
			Diethylene glycol monolaurate	6.0
6 – 9	Dispersible, giving milky solution	wetting agent, water-in-oil emulsions	Sorbitan monopalmitate	6.7
			Sorbitan monomyristate	7.6
			Sorbitan monolaurate	8.6
8 – 10	Soluble, giving milky turbid to translucent solutions	wetting agent	Isotridecanol ethoxylate with 3 EO *	8.3
			Isotridecanol ethoxylate with 4 EO	10.0
			Isotridecanol ethoxylate with 5 EO	11.2
10 – 13	Soluble giving translucent to clear solutions	oil-in-water emulsions, detergents and cleansing agents	Isononylphenol ethoxylate with 5 EO	10.7
			Isononylphenol ethoxylate with 6 EO	11.5

			Isononylphenol ethoxylate with 7 EO	12.3
13 – 15	Soluble, giving clear solution	oil-in-water emulsions, detergents and cleansing agents	Octadecanol ethoxylate with 10 EO	13
			Octadecanol ethoxylate with 12 EO	14
			Octadecanol ethoxylate with 16 EO	15
> 15	Soluble, giving clear solutions	solubilizer, cleansing agent	Dodecanol/tetradecanol ethoxylate with 13 EO	15
			Dodecanol/tetradecanol ethoxylate with 17 EO	16
			Dodecanol/tetradecanol ethoxylate with 25 EO	17
			Dodecanol/tetradecanol ethoxylate with 38 EO	18

* Moles ethylene oxide per mole.

The HLB values can be determined by emulsification tests and comparison with emulsifiers of known HLB values (see also ⟶ Emulsions).

7.2. Ethoxylates

Ethoxylates are generally obtained by addition of ethylene oxide to compounds containing dissociating protons.

Substrates used for ethoxylation are primarily linear and branched, primary and secondary C_{12} – C_{18} alcohols, i.e., natural and synthetic fatty alcohols, alkylphenols with branched octyl (butene dimer), nonly(propylene trimer) or dodecyl(propylene tetramer) groups, fatty acids, fatty acid ethanolamides, fatty amines, and fatty acid esters of polyhydroxyl compounds. The degree of ethoxylation, i.e., the molar ratio of ethylene oxide added per mole of substrate, varies within wide ranges, in general between 3 and 40, and is chosen according to the intended use (HLB value, Table 13).

The addition of ethylene oxide to a substrate containing acidic hydrogen is catalyzed by bases or (Lewis) acids.

Amphoteric catalysts, prepared in situ and probably existing as finely dispersed solids having a large surface area [94], as well as heterogeneous catalysts [95], have also been described.

The reaction mechanisms of base- catalyzed and acid- catalyzed ethoxylation differ, which affects the composition of the reaction products. In base- catalyzed ethoxylation an alcoholate anion formed initially by reaction with the catalyst (alkali metal, alkali-metal oxide, carbonate, hydroxide, or alkoxide), nucleophilically attacks ethylene oxide. The resulting anion of the ethylene oxide addition product can undergo an equilibrium reaction with the alcohol starting material or ethoxylate product, or can react further with ethylene oxide:

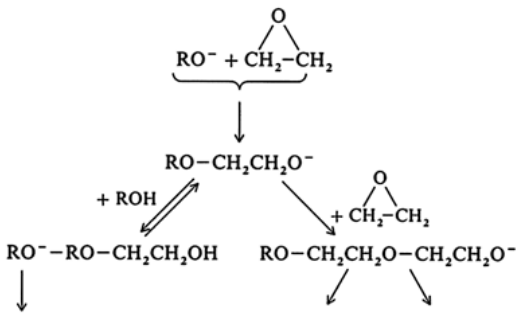

As this simplified reaction scheme illustrates, in alkaline- catalyzed ethoxylation several reactions proceed in parallel, the addition of ethylene oxide to an anion with the formation of an ether bond being irreversible.

Proton exchange between various anions, occurring as an ionic reaction, is fast; the addition reaction of ethylene oxide to an existing anion is the rate-determining step. Thus in a reaction mixture, the more acidic species preferentially react with ethylene oxide. If carboxylic acids or phenols are ethoxylated, the reaction proceeds exclusively via the left-hand path in the above scheme, and the initially formed monoethylene glycol ester or ether only reacts further when all the starting material in the reaction mixture is consumed. The fast equilibrium proton exchange reaction which precedes the addition of the ethylene oxide to the anionic species leads to a peculiarity of the reaction rate in the case of strongly acidic substrates such as carboxylic acids or phenols. On account of the lower nucleophilicity of their conjugate bases corresponding to the acidic parent substances, these bases react relatively slowly with ethylene oxide; thus, the reaction proceeds slowly until the starting material is consumed. Thereafter the reaction rate increases sharply with further supply of ethylene oxide since the ethoxylate anions now present react substantially more quickly with ethylene oxide than do the anions of the carboxylic acids or phenols [96].

The situation is different in the ethoxylation of alcohols. The ether oxygen atoms in alkyl (oligo)glycol ethers increase the acidity of the terminal primary hydroxyl group compared to the initial alcohol; glycol ethers once formed thus react preferentially with ethylene oxide and lead to the formation of a mixture of homologous oligoglycol ethers, and unreacted starting alcohol remains in the reaction mixture up to high degrees of ethoxylation. This is particularly true for the ethoxylation of secondary alcohols.

Assuming the same acidity of the starting alcohol and all (oligo)glycol ethers present in the mixture, a Poisson distribution of the individual species must be expected in the ethoxylation, with a maximum corresponding to that oligoglycol ether in which the number of added ethylene oxide units corresponds to the molar ratio of ethylene oxide to starting alcohol. However, on account of the aforementioned different acidities of the individual species in the reaction mixture the homologue distribution that is actually observed in an ethoxylation mixture differs from the Poisson distribution (Fig. 20).

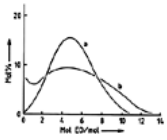

Figure 20. Homologue distribution of a lauryl ethoxylate containing 6 mol EO/mol [97]

a) Acid catalysis; b) Base catalysis

This is true for all alkaline catalysts, although the deviations in the case of alkaline-earth compounds are less strongly pronounced than in the case of sodium hydroxide or sodium methoxide [97]. The distribution pattern of the homologous polyethylene glycol ethers in an ethoxylate obtained by alkaline catalysis is independent of the temperature, pressure, and catalyst concentration [98], [99]. The occasionally observed dependence of the homologue distribution on the stirring rate can be attributed to an insufficiently complete mixing of the reactor contents.

When Lewis acids such boron trifluoride, tin tetrachloride, or antimony pentachloride are used as catalysts, homologue distributions approximating to the Poisson distribution are obtained (Fig. 20), because here it is not the proton activity but the nucleophilicity of the substrate that determines the reaction pathway. Lewis acids activate ethylene oxide and not the alcohol:

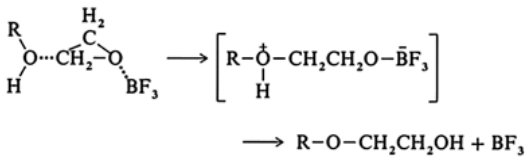

Lewis acids have not become established as catalysts since they must be laboriously removed from the reaction product and because they lead to the formation of polyethylene

glycol [the so- called polydiol, $HO(CH_2CH_2O)_nH$], methyldioxolane and dioxane, due to side reactions and decomposition reactions.

Since ethoxylates with a narrow, Poisson-like homologue distribution, so- called narrow range ethoxylates (NRE), have advantages in some applications, amphoteric catalysts have been developed that give distributions similar to those produced by Lewis acids [94].

Among these catalysts a calcined hydrotalcite of idealized empirical formula $Mg_6Al_2O_5(OH)_2$ is of interest since it can readily be handled as a pneumatically conveyable powder and can easily be separated as an insoluble solid from the reaction medium [95].

Advantages of narrow range ethoxylates, which also have a substantially lower content of starting alcohol, include: less odor, better solubility, lower volatility (reduced pluming on spray drying) and better thickening properties of both the ethoxylates and the ether sulfates derived therefrom [100-102]. However, broad distributions may also be advantageous, for example, in emulsification processes [103].

The large variability of the hydrophobic group — straight- chain and branched fatty alcohols, alkylphenols, fatty amines, fatty acids and fatty acid (alkanol)amides — as well as the option to achieve any desired degree of ethoxylation, make the ethoxylates an extremely versatile class of surfactants. The ethoxylates of fatty alcohols and alkylphenols as such or in combination with anionic, occasionally also cationic surfactants, play a major role in detergents and cleansing agents. Objections have been raised against the use of alkylphenol ethoxylates on the basis of ecotoxicological findings, which has led to their discrimination in some European countries [104]. When adjusted to the necessary HLB value by means of the degree of ethoxylation, fatty alcohol ethoxylates are used as emulsifiers in cosmetics formulations, dry cleaning, crop protection agents, metal working and processing, and textile auxiliaries, paints and coating compositions. Like the alkylphenol ethoxylates, they are also used as dispersants in petroleum production and in ores flotation. Fatty acid polyglycol esters have numerous uses as emulsifiers in bodycare products and in textile treatment. The same is also true of fatty acid alkanolamides and their ethoxylates, which moreover are used as antistatics in plastics processing and as emulsifiers and corrosion inhibitors in metal working and processing. Finally, fatty amine ethoxylates are used as finishing agents and antistatics in textile treatment and leather processing, and as emulsifiers in the petroleum industry and bitumen production.

General Production Processes. Molecularly uniform ethoxylates can be prepared by the Williamson ether synthesis. Industrial ethoxylates, obtained by ethylene oxide addition are mixtures of homologues. These mixtures generally have superior application properties to molecularly uniform substances.

Special safety precautions must be adopted in industrial ethoxylation. Ethylene oxide, *bp* 10.7 °C, is stored, transported, and metered as a liquid under pressure. The flash point is – 57 °C, the ignition temperature in air is 429 °C, and the decomposition temperature of pure ethylene oxide vapor is 571 °C. The heat of polymerization of ethylene oxide, which is released during ethoxylation, is 2090 kJ/kg; 1900 kJ/kg is released in the decomposition of ethylene oxide vapor. In ethoxylation processes the reaction can run out of control if there is overheating and a consequent rise in pressure; ethylene oxide tends to ignite or undergo self-decomposition (e.g., on overheated stirrer or pump shafts). Ethoxylation must therefore be carried out under strict pressure and temperature control. The reactors must be safe-guarded against excessive pressure by means of safety valves or rupture disks. The excess of ethylene oxide in the reaction mixture must be kept low. On account of the limited solubility of ethylene oxide in the reaction medium the presence of a second, liquid ethylene oxide phase in the reaction vessel must be avoided. The storage vessel from which the ethylene oxide is metered into the reaction vessel must be protected against backflow of reactants containing polymerization-initiating catalyst, e.g., by nonreturn valves and intermediate vessels.

Industrial ethoxylations are carried out mainly in a batchwise manner. Stainless steel pressure vessels are used as reactors. The substrate together with the catalyst, generally sodium hydroxide, is first placed in the vessel. The water introduced with the sodium hydroxide solution and the water formed in the reaction

$$ROH + NaOH \rightleftharpoons RONa + H_2O$$

is removed by heating and applying a vacuum or by passing nitrogen through the reaction mixture, to avoid formation of polyglycols ("polydiols"). Polydiols do not, in general, affect the properties of ethoxylates provided their concentration does not exceed a few percent, but not being surface-active they result in a wastage of ethylene oxide. The catalyst content is usually 0.1 – 1.0 %. The ethoxylation is carried out between 130° and 180 °C, and the substrate first must be heated up to this temperature. The reactor contents must be cooled during the addition of ethylene oxide. Heating and cooling can be performed by external jackets, internal tubes or by heat exchangers through which the reactor contents are constantly pumped.

The gaseous phase is inertized by diluting it with nitrogen. The reaction is carried out at 1 –5 bar gauge pressure. The liquefied ethylene oxide is metered into the reaction medium with intensive mixing. The temperature and pressure are monitored so that the ethylene oxide supply can be adjusted if the reaction becomes quiescent or goes out of control. The desired degree of ethoxylation is adjusted via the amounts of substrate and ethylene oxide.

The product specifications can quickly be checked by determining the cloud point.

The reaction time depends on the effectiveness of cooling and amount of ethylene oxide to be added; the production of an ethoxylate having a medium degree of ethoxylation takes a few hours in conventional stirred vessels provided with internal cooling.

The catalyst is neutralized with carbon dioxide, acetic acid, or citric acid. Phosphoric acid gives a precipitate that can be filtered off, but in most cases the neutralized catalyst remains dissolved in the product. Any slight discoloration of the products can be removed by bleaching with hydrogen peroxide.

The products are discharged hot into stainless steel storage tanks or vats, or are cooled in the reactor and discharged into plastics vessels.

Discoloration and polydiol concentration increase with increasing reaction temperature. The formation of polydiols is due to increasing elimination of water from ethoxylates as the temperature rises:

$$RO-[CH_2CH_2O]_n-OCH_2CH_2OH \longrightarrow RO-[CH_2CH_2O]_n-OCH=CH_2 + H_2O$$

$$H_2O + n\overset{O}{\overset{/\backslash}{CH_2-CH_2}} \longrightarrow HO-[CH_2CH_2O]_n-H$$

Further impurities can be formed by reactions of other compounds introduced with the ethylene oxide (e.g., formaldehyde and acetaldehyde).

Although continuous ethoxylation processes are also described in the literature [105], industrial ethoxylations, like other alkoxylations, are mainly carried out as batch operations.

The Pressindustria alkoxylation process for addition of ethylene oxide and/or propylene ox-ide operates according to a different principle than the stirred vessel method (Fig. 21). The liquid reaction mixture, circulated at high speed, is dispersed by a special device in a gas phase consisting of ethylene oxide (possibly also propylene oxide), ensuring rapid absorption of ethylene oxide [106], [107]. Catalyst and substrate are mixed in a heated receiving vessel and freed from water by heating under vacuum. This mixture is passed to the reactor and circulated by means of a pump through the special mixing unit (c) and the reactor. The mixture is heated to reaction temperature in the heat exchanger (e); residual moisture is removed by applying a vacuum or passing nitrogen through the mixture. The addition of ethylene oxide then starts, the ethylene oxide being flashed into the gas phase of the mixer (c). Here the liquid reaction mixture introduced in finely dispersed form adsorbs the ethylene oxide. The reaction proceeds in the liquid phase in the mixer and reactor, into which the liquid phase flows from the mixer. The heat of reaction is dissipated by the cooler (f). When the desired degree of ethoxylation is reached the product is passed to the neutralization vessel. The reaction is carried out at 180 – 190 °C, and a maximum pressure of 5 bar.

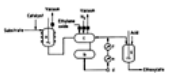

Figure 21. Pressindustria alkoxylation apparatus [106]

a) Receiving vessel; b) Reactor; c) Mixer for gas and liquid; d) Circulating pump; e) Heater; f) Cooler; g) Neutralization vessel

The process has a high reaction rate and uniform product quality. The safety risks involved with gaseous ethylene oxide are minimized by keeping the gas phase as small as possible. Since this part of the reactor contains no moving parts, ignition of the gas phase by shafts, for example, overheating, is ruled out. This process can also be operated continuously.

High reaction rates and low formation of byproducts can also be achieved in a jet nozzle reactor or jet suction reaction mixer [108].

Production of Ethoxylates of Fatty Acids, Fatty Acid Amides, and Fatty Amides. All industrial products based on primary alcohols or alkylphenols can be ethoxylated by the above processes.

In the ethoxylation of fatty acids the transesterification of the initially formed oligoglycol esters that is unavoidable with alkaline catalysis leads to the formation of diesters and free oligoglycols, which in the further course of the ethoxylation lead to undesirably high polydiol contents in the end product:

$$2\ RCOO-[CH_2CH_2O]_n-H \longrightarrow$$

$$R-COO-[CH_2CH_2]_n-COR + HO-[CH_2CH_2O]_n-H$$

As carboxylic esters, fatty acid ethoxylates are sensitive to hydrolysis.

Ethoxylates of fatty acid amides are obtained not by ethoxylation of amides but by alkaline- catalyzed ethoxylation of fatty acid monoethanolamides or diethanolamides, which are obtained from the reaction of fatty acids or their methyl esters with ethanolamine or diethanolamine, which themselves are valuable products:

$$RCO-NH-CH_2CH_2OH$$

$$RCO-N(CH_2CH_2OH)_2$$

In the ethoxylation of primary amines, in the presence of basic catalysts, only one hydrogen atom bonded to nitrogen reacts, and the initially formed *N*-alkylethanolamine reacts further with ethylene oxide at the hydroxyl group, with the retention of the secondary amino group. Amines doubly ethoxylated on the nitrogen atom are obtained if the amines are first converted to the *N*-alkyldiethanolamine under proton catalysis, for which purpose small amounts of water are generally sufficient, and this diethanolamine is then ethoxylated further with alkaline catalysis.

Properties. Ethoxylates are described according to the state of aggregation (liquid, pasty, waxy, solid) and color (iodine no. < 2), and the density (ca. 1 g/cm^3), refractive index, and viscosity are specified as physical data. Characteristic data also include the cloud point (according to DIN 53 917) and the hydroxyl value. The content of active substance is generally given as 100 %, despite the fact that ethoxylates contain amounts of polydiols that increase with increasing degree of ethoxylation. An industrial lauryl alcohol ethoxylate containing 5 mol of EO/mol contains, e.g., ca. 0.5 %, an ethoxylate with 15 mol EO/mol contains 1.0 %, and one with 30 mol EO/mol contains 3.0 % of polydiol.

Storage and Dispatch. Ethoxylates are stored and transported in stainless steel or plastic vessels. Tank wagons and storage tanks for higher ethoxylated products that are solid at ambient temperature must be thermally insulated and heatable.

When ethoxylates are dissolved or diluted with water the possibility of gel formation in specific concentration ranges must be borne in mind.

CAS Numbers of some important ethoxylates are listed below. Irrespective of the degree of ethoxylation, ethoxylates with identical hydrophobic groups have the same CAS numbers.

Hydrophobic starting material	CAS number
100 % linear C_{10}-prim. alcohol	[26183-52-8]
100 % linear C_{10}/C_{12}-prim. alcohol	[66455-15-0]
100 % linear C_{12}-prim. alcohol	[9002-92-0]
100 % linear C_{12}/C_{14}-prim. alcohol	[68439-50-9]
100 % linear $C_{12}/C_{14}/C_{16}/C_{18}$-alcohol	[68213-23-0]
100 % linear C_{16}/C_{18}-prim. alcohol	[68439-49-6]
100 % linear oleyl alcohol	[9004-98-2]
	[68920-66-1]
80 % linear C_9– C_{11} prim. alcohol, monobranched	
	[68439-45-2]
80 % linear C_{12} – C_{13} prim. alcohol, monobranched	
	[66455-14-9]
80 % linear C_{12} – C_{15} prim. alcohol, monobranched	
	[68131-39-5]
80 % linear C_{14} – C_{15} prim. alcohol, monobranched	
	[68002-97-1]
60 % linear C_{13} – C_{15} prim. alcohol, monobranched	
	[64425-86-1]
50 % linear C_{12} – C_{13} prim. alcohol, monobranched	
	[66455-14-9]
50 % linear C_{12} – C_{15} prim. alcohol, monobranched	
	[68131-39-5]
50 % linear C_{13} – C_{15} prim. alcohol, monobranched	
	[68002-97-1]
Branched C_{10}-alcohol	[61827-42-7]
Branched C_{13}-alcohol	[69011-36-5]
p-Isooctylphenol	[9004-87-9]
p-Isononylphenol	[37205-87-1]
	[68412-54-4]
Lauric acid	[9004-81-3]
Coco fatty acid	[61791-29-5]
Oleic acid	[112-80-1]
Stearic acid	[57-11-4]
Ricinoleic acid	[61791-12-6]
Tallow fatty amine	[61791-26-2]
Oleylamine	[112-90-3]
Stearylamine	[124-30-1]
Alkanolamides	
Coco fatty acid monoethanolamide	[68140-00-1]
Coco fatty acid diethanolamide	[61790-63-4]
Tallow fatty acid monoethanolamide	[68153-63-9]

Producers. Fatty alcohol ethoxylates are produced in Europe, e. g., by Akzo, Albright & Wilson, Amerchol Europe, Auschem, BASF, Berol Nobel, Chem-Y, Croda, Harcros, Henkel, Hoechst, Hüls, ICI, Kolb, Pulcra, Rewo, Rhône-Poulenc, Seppic, Shell, Stepan, Witco, and Zschimmer & Schwarz; in North America, e. g., by Amerchol, BASF, Chemax, Croda, Exxon, Hart, Henkel/Emery, Hoechst Celanese, ICI Americas, Lonza, Olin, Procter & Gamble, Rhône-Poulenc Surfactant, Sandoz, Shell, Sherex, Stepan, Texaco, Union Carbide, Vista, and Witco; in Japan by Dai-Ichi, Nikko Chemicals, Nippon Nyukazai, Nippon Oil & Fat, Sanyo Chemical, and Takemoto Oil & Fat; in Australia by ICI; and in Taiwan by Sino-Japan Chemical and Taiwan Surfactant.

Alkylphenol ethoxylates are produced in Europe by Akzo, Auschem, BASF, Berol Nobel, Chem-Y, Dow, Harcros, Hefti, Hoechst, Hüls, ICI, Kolb, Pulcra, Stepan, Witco, and Union Carbide; in North America by Chemax, Clough Chemical, Exxon, Hart, Henkel/Emery, Hoechst Celanese, Norman/Fox, Rhône-Poulenc Surfactant, Sandoz, Stepan, Texaco, Union Carbide, and Witco; in Japan by Dai-Ichi, Kao, Lion, Marubeni Oil, Nippon Nyukazai, Nippon Oil & Fat, Sanyo Chemical, and Toho Chemical Industry; in Australia by ICI; and in Taiwan by Sino-Japan Chemical and Taiwan Surfactant.

Fatty acid ethoxylates are produced in Europe by BASF, Croda, Fina Oleochemicals, Harcros, Hefti, Henkel, Hüls, ICI, Stepan, Witco, and Zschimmer & Schwarz; in North America by Chemax, Emkay, Hart, Henkel, Hodag Chemical, ICI Americas, Lipo Chemicals, Lonza, PPG/Mazer, Rhône-Poulenc Surfactant, and Reilly-Whiteman; in Japan by Dai-Ichi, Kao, Nikko Chemicals, Nippon Nyukazai, Nippon Oil & Fat, and Toho Chemical Industry; in Australia by ICI; in Taiwan by Sino-Japan Chemical and Taiwan Surfactant; and in Israel by Zohar.

Special ethoxylated fatty esters and oils are also produced by companies such as Auschem, Croda, ICI, Th. Goldschmidt and Westbrook Lanolin in Europe, and Heterene Chemical, ICI Americas, Lonza, Rhône-Poulenc Surfactants, and Sherex in the United States.

Ethoxylated amines and/or fatty acid amides are produced in Europe by Albright & Wilson, Akzo, BASF, Berol Nobel, Harcros, Henkel, Hoechst, Hüls, Rewo, Rhône-Poulenc, and Zschimmer & Schwarz; in the United States by Akzo, Chemax, Ethox Chemicals, Hart, Henkel, Heterene Chemical, Karlshanns, PPG/Mazer, Rhône-Poulenc Surfactant, Sandoz, Sherex, Union Carbide and Witco; in Japan by Dai-Ichi, Kao, Nippon Nyukazai, Nippon Oil & Fats, and Sanyo; in Australia by ICI; and in Taiwan by Taiwan Surfactant and Sino-Japan Chemical.

7.3. Terminally Blocked Ethoxylates [109], [110]

By substituting the hydrogen atom of the terminal hydroxyl group of an ethoxylate by hydrophobic residues such as benzyl, butyl, or methyl groups, terminally blocked ethoxylates are obtained that are chemically more resistant, especially in alkaline media, than the corresponding ethoxylates with a free hydroxyl group. Since blocked ethoxylates also foam less in aqueous solution than their starting ethoxylates, they have a certain value in (alkaline) cleaning processes involving strong mechanical action.

The addition of higher alkylene oxides, in particular propylene oxide, may also be regarded as a terminal blocking:

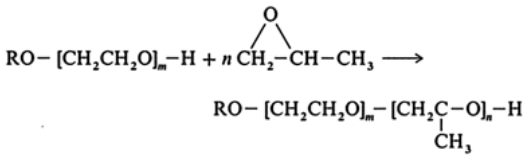

However, since such products also contain free hydroxyl groups they are no more stable than their starting ethoxylates against alkalis, but in aqueous solution foam less than the pure ethoxylates. Due to the blocking of the terminal hydroxyl groups by hydrophobic residues such ethoxylates to some extent lose their surfactant character, and become less water-soluble and less amphiphilic. This is already noticeable in propoxylated products, the propoxy group itself being hydrophobic. This property can be utilized to enlarge a hydrophobic residue before ethoxylation by addition of propylene oxide. Finally, alkoxylation can be performed with mixtures of ethylene oxide and propylene oxide and various effects can be achieved by appropriately choosing the ratio of the two epoxides, by alternating the order of addition, etc.

7.4. Fatty Acid Esters of Polyhydroxy Compounds

Polyhydroxy compounds such as glycerol, diglycerol, polyglycerol, erythritol, pentaerythritol, xylitol, sorbitol, mannitol, sucrose, and other glucosides that are partially esterified with fatty acids have surface-active properties. The hydrophilicity (HLB value) of such esters can be increased and matched to the desired application by ethoxylating the unesterified hydroxyl groups. Esters of glycerol, sorbitol, and sucrose are discussed hereinafter by way of example, as are the so- called alkyl polyglucosides, which have recently attracted increased attention.

Fatty Acid Esters of Glycerol [111]. Disproportionating transesterification of fats and oils, the fatty acid triesters of glycerol (triglycerides), with glycerol in the presence of acid, but preferably alkaline catalysts, yields, depending on the glycerol excess, equilibrium reaction mixtures containing 35 – 60 % of monoglycerides, 35 –50 % of diglycerides, 1 – 20 % of triglycerides, 1 – 10 % of fatty acids and their alkali metal salts, and 1 – 10 % of glycerol after ca. 1 h at 200 –250 °C. Such so- called mono-diglycerides, which can also be obtained by direct reaction of free fatty acids with glycerol under acid catalysis, are marketed as mixtures. They are insoluble in water, and when they contain alkali metal salts of fatty acids, they are self-emulsifying. The fatty acids that are used are predominantly from the tallow fat range (palmitic and stearic acids). The above-mentioned mixtures are generally marketed under the name “monoglycerides”, e.g., “glycerol monostearates” (GMS).

High- concentration monoglycerides are obtained from the mixtures by molecular distillation. They consist mainly of α-glycerides, and contain less than 10 % of β-glycerides:

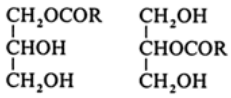

Since they crystallize in fine platelets, they impart a pearl gloss to pasty liquid detergents and cleansing agents.

Glycerides are widely used as emulsifiers in foods and cosmetics, as they are physiologically harmless. Characteristic data of some commercial products are listed in Table 14.

Table 14. Typical data of some industrial glycerides

	Glycerol monostearate	Glycerol mono/distearate	Glycerol mono/distearate, self-emulsifying	Glycerol monolaurate
CAS no.	[61789-09-1]	[68308-54-3]	[68308-54-3] [593-29-3]	[27215-38-9]
Color, consistency	white powder	white powder or flakes	whitish flakes	white, crystalline
Melting point, °C	66 – 71	56 – 61	56 – 61	56 – 60
Monoglyceride content, %	90	50	35	95
Diglyceride content, %	5	40	45	2
Glycerol content, %	1	1	3	1
Sodium stearate content, %	0	0	3	0
Acid number, mg KOH/g	< 3	< 3	< 6	< 3
Saponification number, mg KOH/g	155 – 170	160 – 175	150 – 175	195 – 205
Iodine color number, mg I_2 /100 mL	< 3	< 3	< 6	< 3

The enzymatically catalyzed synthesis of glycerides has not yet become industrially important [112].

Anionic surfactants having properties similar to those of the ether sulfates are obtained by sulfating monoglycerides and diglycerides [113].

Producers of glycerides include Akzo, Croda, Fina Oleochemicals, Gatte-Fosse, Hefti, Henkel, Hüls, ICI, Quest, Rewo, and Th. Goldschmidt in Europe; Goldschmidt Chemical, Henkel, Hodag, Hoechst-Celanese, Humko, Inolex, Karlshamns, Lonza, PPG/Mazer, and Witco in North America; Kao, Nikko Chemicals, Nippon Oil & Fats, and Toho in Japan; Taiwan Surfactant and Sino-Japan Chemical in Taiwan; and Zohar in Israel.

Fatty Acid Esters of Sorbitol [114]. Sorbitol can be converted to monoesters or polyesters by reaction with fatty acid chlorides under mild conditions in the presence of pyridine. The esterification is performed industrially with the free fatty acids or their methyl esters at elevated temperature; sorbitol readily undergoes dehydration to sorbitan, which can lose a further molecule of water to give isosorbitol, especially in the presence of acid catalysts:

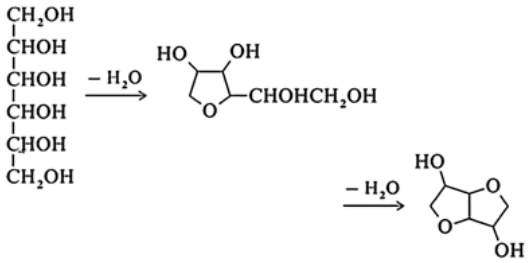

Esterification is generally carried out at 200 –250 °C after the addition of small amounts of alkali, e.g., 0.1 % sodium hydroxide; depending on the molar ratio of the starting materials, fatty acid monoesters, diesters, or triesters of sorbitan are obtained. First marketed by Atlas Powder, these are known as “Span” types. They are water-insoluble products having low HLB values that are liquid at ambient temperature or melt below 100 °C. Ethoxylation of the Span types yields products having better water solubility and higher HLB values, namely the “Tweens”, which are waxy to solid products.

Spans and Tweens are dispatched in plastic containers, plastic- coated cardboard containers, or bags.

Producers of sorbitan esters include Akzo, Auschem, Croda, Hefti, Henkel, and ICI in Europe; Chemax, Croda, Henkel/Emery, Hodag, ICI Americas, Karlshamns, Lipo Chemical, Lonza, Norman/Fox, PPG/Mazer, Stepan, and Witco in North America; Kao, Nikko Chemicals, Nippon Nyukazai, Nippon Oil & Fats, Sanyo, and Toho in Japan; and Sino-Japan Chemical and Taiwan Surfactant in Taiwan.

Fatty Acid Esters of Sucrose [115], [116]. As a plentifully available natural raw material, sucrose is a suitable parent substance for the hydrophilic group of surfactants. Surfactants based on sucrose, especially their fatty acid esters, are nontoxic, do not irritate the skin, and are fully biodegradable. However, they have not hitherto gained wide acceptance industrially due to a certain instability of the sucrose glycoside bond, and the insolubility of sucrose in conventional organic solvents, which makes reactions involving sucrose more difficult. Fatty acid esters of sucrose are prepared by alkaline- catalyzed reaction with 1 – 2 mol of fatty acid methyl ester per mole of sucrose, the methanol being distilled off under vacuum. The temperature should not exceed 100 °C. Dimethylformamide or dimethyl sulfoxide, both highly toxic substances, are used as solvents and must be removed completely from the reaction product, which is laborious since sucrose esters are temperature sensitive and must not be heated too strongly.

Producers of sucrose esters include Croda, Gattefosse, and ICI in Europe; Amerchol, Croda, and Union Carbide in the United States; and Dai-Ichi in Japan.

Alkyl Polyglucosides [117-120]. Acetal-bonded glucosides mixed with oligoglucosides are obtained as a randomly distributed mixture of homologues by acetalation of glucose with fatty alcohols; the average number of glucose units per molecule being between 1 and 3, depending on the reaction parameters. Strictly speaking the products should be called alkyl oligoglucosides, though the incorrect term alkyl polyglucosides (APG) has persisted.

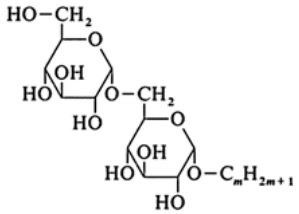

Alkyl polyglucosides are waxy, soft to glassy-solid, and colored yellow due to impurities. They are generally extremely soluble in water and are handled as 50 % aqueous solutions. The HLB values lie above 10, and can be varied in the range 11 – 15 by changing the length of the hydrophobic group and the degree of glucosidation. The surface tension of aqueous solutions is ca. 30 mN/m and the interfacial tension toward hydrocarbons 1 mN/m. Alkyl polyglucosides are moderate foaming agents and outstanding emulsifiers. They are suitable for washing textiles and hard surfaces, particularly in combination with other anionic and nonionic surfactants, with which they give synergistic effects. Alkyl polyglucosides decompose on heating; their aqueous solutions are stable in the pH range 3 –13. They are nontoxic, only slightly irritating to the skin, and biodegrade rapidly.

Alkyl polyglucosides can be prepared by direct reaction of anhydrous glucose with fatty alcohols. In another method, in which also aqueous glucose solution can be used, the latter is first reacted with butanol to form butyl glucoside, which is then reacted in a second stage with fatty alcohol to form the fatty alkyl oligoglucoside. Butanol and excess fatty alcohol are distilled off. Strong acids such as sulfuric acid are used as catalysts in both methods.

The alkyl polyglucoside thus obtained as a melt is dissolved in water and bleached with hydrogen peroxide.

CAS Numbers. C_8/C_{10}-Fatty alkyl polyglucoside [128664-36-8], C_{12}/C_{14}-fatty alkyl polyglucoside [136797-44-9].

Producers include Henkel in the United States, Kao in Japan, Seppic in France, and BASF, Henkel, and Hüls in Germany.

7.5. Amine Oxides [121], [122]

Reaction of tertiary amines such as fatty alkyldimethylamine or fatty amines doubly ethoxylated at the nitrogen atom with hydrogen peroxide yields amine oxides:

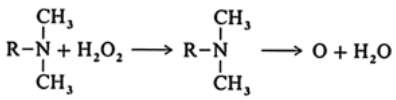

Surfactant amine oxides such as lauryldimethylamine oxide are insensitive to water hardness. They disperse lime soaps, foam satisfactorily and are mild to the skin. They are therefore widely used as constituents of dishwasher detergents, shampoos, and soaps. Amine oxides in neutral aqueous solution should be regarded as nonionic surfactants. They are protonated in acid solution and thus represent the transition to cationic surfactants.

CAS Numbers. Lauryldimethylamine oxide [1643-20-5]; Stearyldimethylamine oxide [2571-88-2]; Coco fatty alkyldimethylamine oxide [61788-90-7].

Producers of amine oxides include Akzo, Albright & Wilson, Ceca, Croda, Pulcra, Rewo, and Th. Goldschmidt in Europe; Clough, Ethyl, Goldschmidt Chemical, Henkel, McIntyre, PPG/Mazer, Rhône-Poulenc Surfactant, Sherex, Sher Chemicals, Stepan, and Witco in the United States; Nippon Oil & Fats in Japan; and Taiwan Surfactant in Taiwan.

[Top of Page]

8. Cationic Surfactants [123-125]

Surfactants whose hydrophilic group is a cation are termed cationic surfactants. The cationic structure may already be present in the surfactant molecule, as in the case of quaternary ammonium or phosphonium salts, or may be formed only in aqueous solution, as, for example, in ethoxylated fatty amines. In acidic solution, nonionic surfactants can adopt a cationic character due to protonation at a heteroatom; examples are fatty amine oxides and even ethoxylates, whose ether groups can form oxonium structures. However, strictly speaking only those surfactants that do not require protonation to become cationic are regarded as cationic surfactants.

The most important cationic surfactants are the quaternary nitrogen compounds: tetraalkylammonium salts, *N*,*N*-dialkylimidazolinium compounds, and *N*-alkylpyridinium salts. The positive electrical charge of the hydrophilic moiety confers on cationic surfactants properties that open up areas of application for which anionic and nonionic surfactants are unsuitable or less suitable. Important uses of these cationic surfactants are as microbicides, herbicides, corrosion inhibitors, oxidation inhibitors, adhesives, finishing and sizing agents, water repellents, fabric softeners, flotation aids, dispersants, levelling agents, etc.

Other cationic surfactants, e.g., the quaternary phosphonium salts or tertiary sulfonium salts are less important industrially, although they are very useful as phase-transfer catalysts in synthesis [126].

8.1. Quaternary Ammonium Compounds

Quaternary ammonium compounds are obtained by alkylating tertiary amines. At least one of the alkyl residues must be a relatively long alkyl group. In general, tertiary amines having one or two long alkyl residues are used as starting compounds (⟶ Amines, Aliphatic), to which are added shorter alkyl groups in the form of methyl chloride, ethyl chloride or bromide, or dimethyl sulfate. The corresponding ammonium chlorides, ammonium bromides, or ammonium methyl sulfates are then obtained. The addition requires a polar solvent (e.g., alcohol); in the case of alkyl chlorides, which are mainly used for the alkylation, temperatures of 50 – 100 °C and reaction times of several hours are required. The process can be simplified by carrying out the reaction under pressure in excess alkylating agent (e.g., methyl chloride as solvent). The product is isolated by releasing the pressure and evaporating unreacted alkyl chloride [127].

Benzyl chloride reacts very readily as alkylating agent; the disinfectant Zephirol is obtained by addition to dimethyldodecylamine:

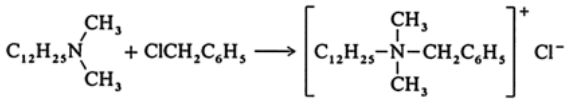

Another method of quaternizing tertiary amines, as well as ethoxylated fatty amines, is the reaction with ethylene oxide and water in the absence of a catalyst:

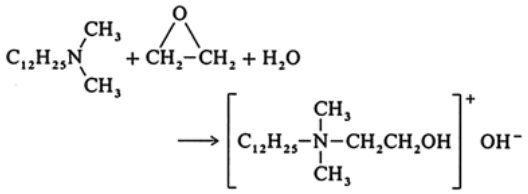

For special applications quaternary compounds can also be obtained from fatty acid derivatives such as the amides of diamines or esters of triethanolamine:

$RCONHCH_2CH_2N(CH_3)_2$

$RCOOCH_2CH_2N(CH_2CH_2OH)_2$

A further group of amines that is suitable for quaternization is obtained by hydrogenating substituted propionitriles, which are obtained by addition of fatty alcohols or fatty amines to acrylonitrile:

$ROCH_2CH_2CH_2NH_2$ and $RNHCH_2CH_2CH_2NH_2$

Distearyldimethylammonium chloride (DSDMAC) or, more correctly, since the long alkyl residues are tallow fat alkyl residues, ditallow alkyl dimethyl ammonium chloride (DTDMAC), was for a long time the most important component in fabric softeners. Since it proved to be relatively toxic to water-borne organisms [128], it has been replaced by less toxic ammonium compounds, which contain ester group as cleavage sites which lead to rapid breakdown of the compounds in water. These are the so- called esterquats obtained from fatty acid, triethanolamine or *N*-methyldiethanolamine, and dimethyl sulfate [129], [130]:

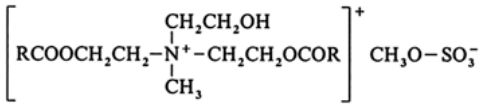

and the diesterquats obtainable from epoxypropyltrimethylammonium chloride and fatty acids [131]:

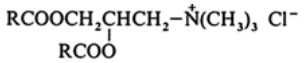

Typical data of some simple quaternary ammonium chlorides that are stored and transported in the form of aqueous or aqueous-alcoholic solution in stainless steel or plastic drums are listed in Table 15.

Table 15. Typical data of some industrial quaternary ammonium compounds

	Stearyltrimethylammonium chloride	Stearylbenzyldimethylammonium chloride	Ditallowalkyldimethylammonium chloride	Ditallowacyloxyethyl hydroxyethylmethyl ammonium methyl sulfate (“Ester quat”)
CAS no.	[112-03-8]	[122-19-0]	[61789-80-8]	[93334-15-7]
				[91995-81-2]
Consistency	liquid	liquid	liquid – solid	liquid – solid
Flow point, °C			ca. 35	ca. 35
Active substance content, %	50 – 52	50 – 52	74 – 76	ca. 90
Content of free amine and hydrochloride, %	1	1	2	
Isopropanol content, %	30	0	16	10
Density, g/mL	0.89	0.95	0.90	0.96
Color index (Gardner)	< 7	< 7	4	3
pH (1 % aqueous solution)	5 – 7	6 – 8		
Ash, %	< 0.8	1.0	< 0.35	< 0.1
Flash point, °C	15		ca. 20	26 – 32

Producers of quaternary ammonium compounds include Akzo, BASF, Bayer, Berol, Ceca, Croda, Fina Oleochemicals, Hoechst, Hüls, ICI, Pulcra, Rewo, Rhône-Poulenc, Stepan, and Witco in Europe; Akzo, Emkay, Henkel, Hoechst, Humko, ICI, Karlshamns, Rhône-Poulenc, Sherex, and Witco in North America; Kao, Marubishi Oil, Sanyo Chemical, Takemoto Oil, Tokai Seiyu, and Yoshimura Oil in Japan; and Taiwan Surfactant in Taiwan.

8.2. Imidazoline Derivatives

Condensation of fatty acids with ethylenediamine or substituted ethylenediamine yields substituted imidazolines, e.g.:

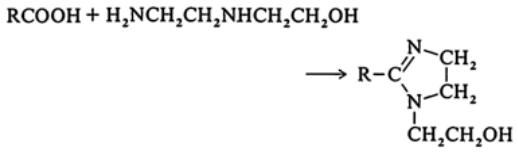

Such compounds are strongly cationic in acidic solution, but undergo slow hydrolysis [132]:

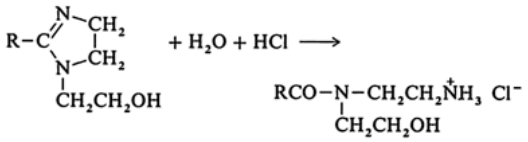

True cationic surfactants are obtained by alkylating imidazoline derivatives, mainly with chloromethane:

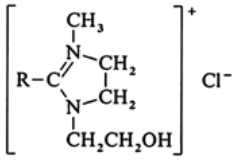

Such surfactants have properties similar to those of acyclic quaternary ammonium salts; they are dermatologically extremely compatible, are antiseptic, and are used as wetting agents, dispersants and cleansing agents, including bodycare products.

An imidazoline derivative obtained by reacting 1 mol of *N*-aminoethylethylenediamine with 2 mol of fatty acid, and quaternized with dimethyl sulfate, is an important fabric softener [86088-85-9] [133]:

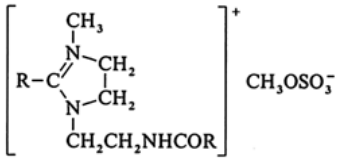

[Top of Page]

9. Amphoteric Surfactants [134-136]

Amphoteric surfactants can be classified as ampholytes and betaines. Ampholytes are compounds having at least one active proton, the best-known example being aminocarboxylic acids, which at the isoelectric point exist as inner salts, at low pH values as cationic species, and at higher pH values as anionic species:

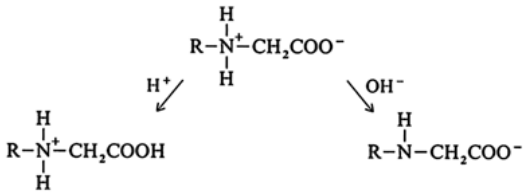

Betaines have no mobile protons and are true amphoteric ions which assume a cationic nature in strongly acid media:

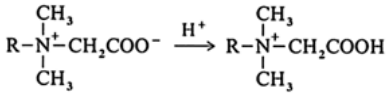

Ampholytes are of minor industrial importance, largely on account of the marked dependence of their properties on pH. Ampholytes are prepared by reacting fatty amines with chloroacetic acid and sodium hydroxide, or by alkaline- catalyzed addition of acrylic acid to fatty amines:

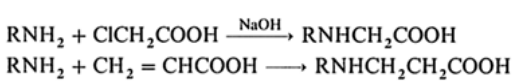

Betaines are prepared by reacting a tertiary amine with chloroacetic acid and sodium hydroxide:

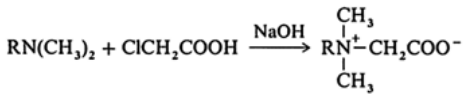

or

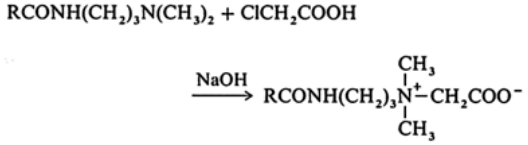

The sulfobetaines are now of only minor importance. Of the various ways of preparing sulfobetaines, the elegant synthesis from amines, alkyl chloride, and sodium hydrogensulfite is mentioned here [137]:

$$RN(CH_3)_2 + CH_2{=}CH_2CH_2Cl \longrightarrow [RN(CH_3)_2CH_2CH{=}CH_2]^+\ Cl^-$$

$$[RN(CH_3)_2CH_2CH{=}CH_2]^+\ Cl^- + NaHSO_3 \xrightarrow{-\,NaCl} RN(CH_3)_2CH_2CH_2CH_2SO_3^-$$

Nowadays only true betaines are of economic importance, especially acid amide betaines and the betaines derived from imidazolines:

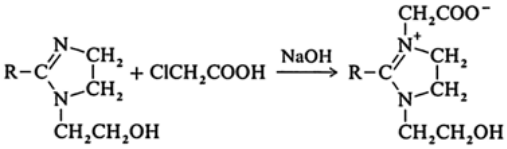

The latter, which are prepared in aqueous solution, probably exist in the open- chain form [138].

Betaines are insensitive to the water hardness and pH value of industrial water, are only slightly toxic, are compatible with the skin and mucous membranes, and have antimicrobial properties. They have good washing and foaming performance and are highly compatible with other classes of surfactants, and are therefore ideally suited for use in bodycare products.

Lecithin or phosphatidylcholine, a naturally occurring phospholipid also belongs to the class of betaines [139], [140]:

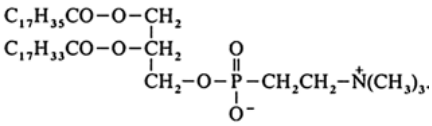

Producers of ampholytes and betaines include Akzo, Albright & Wilson, Auschem, Croda, Harcros, Henkel, Hüls, Manro, Pulcra, Rewo, Rhône-Poulenc, Seppic, Stepan, Th. Goldschmidt, Union Carbide, and Zschimmer & Schwarz in Europe; Clough, Goldschmidt Chemicals, Henkel, Inolex, McIntyre, Mona, PPG/Mazer, Rhône-Poulenc Surfactant, Sher Chemicals, Sherex, Stepan, and Witco in the United States; Lion, Nippon Oil & Fats, Sanyo, and Toho in Japan; Taiwan Surfactant in Taiwan; and Zohar in Israel.

[Top of Page]

10. Surfactants with Heteroatoms in the Hydrophobic Group

10.1. Block Copolymers of Propylene Oxide and Ethylene Oxide [141], [142]

The incorporation of propoxy groups in a surfactant increases its hydrophobicity. Hydrophobic parent substances of any desired molecular mass can be prepared by addition of propylene oxide to a low molecular mass starter molecule. Propoxylation is performed analogously to ethoxylation, with alkaline catalysts at 120 –180 °C under a slight excess pressure:

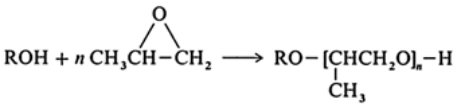

Suitable starter molecules are monofunctional or polyfunctional protic substances, for example, methanol, ethanol, propanol, or higher alcohols; diols such as butanediol; and triols such as gylcerol. Propylene glycol itself is also a suitable diol, and ethylenediamine is a suitable tetrafunctional starter molecule. Subsequent ethoxylation of hydrophobic parent substances obtained in this way yields nonionic surfactants, which can be described as block copolymers of propylene oxide and ethylene oxide. When using polyfunctional starters, branching occurs at the starter molecule. Derivatives of propylene glycol and ethylenediamine, first introduced by Wyandotte under the trade names of Pluronics and Tetronics, are now the most important surfactants of this type:

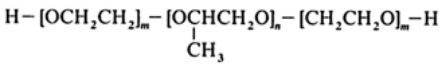

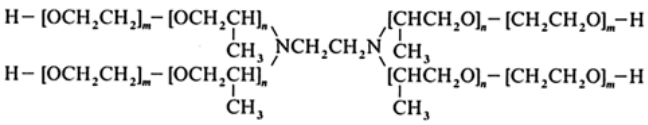

As the structural formulas indicate, a wide range of block copolymers can be synthesized. The products are not uniform, and a homologue distribution range exists for each propoxylate chain and for each ethoxylate chain. Mean molecular masses of 1000 – 5000 are preferred in the case of hydrophobic propoxylate parent substances. Ethoxylation can also be varied over a wide range; depending on the desired application, HLB value and cloud point, between 10 and 80 wt % of ethylene oxide is added.

The products are liquids, pastes, or waxy solids, depending on the molecular mass. They are water-soluble above a content of about 25 wt % of bound ethylene oxide. They are also relatively stable to alkalis and acids.

Their outstanding property is the lack of foaming ability of their aqueous solutions; the propylene oxide – ethylene oxide addition products also suppress the foam of strongly foaming solutions. The block copolymers are therefore preferably used as wetting and cleaning agents in processes involving high mechanical stress and in high-speed cleaning machines. They are also used as demulsifiers and dispersants. However, their unsatisfactory biodegradability prevents their widespread use.

Producers include BASF, Berol, Nobel, Dow, Harcros, Hoechst, Hüls, ICI, and Rhône-Poulenc in Europe; BASF, Hart, Henkel, ICI Americas, Norman/Fox, Olin, and PPG/Mazer in North America; Dai-Ichi and Nippon Oil & Fats in Japan; ICI in Australia; and Sino-Japan Chemical in Taiwan.

10.2. Silicone-Based Surfactants [143-148]

Silicone-based surfactants are mainly derived from methylsilicones (polydimethylsiloxanes, ⟶ Silicones – Polydimethylsiloxanes ⟶ Silicones ⟶ Silicones). They lower the surface tension of water to a greater degree than hydrocarbon-based surfactants, and in this respect are inferior only to fluorosurfactants. In their molecular structure silicone-based surfac-tants resemble surfactants containing polypropylene glycol as hydrophobic parent substance, since the hydrophobic silicone group also has a random molecular mass distribution. The following basic structures (linear, comb, and branched) occur, where A denotes hydrophilic groups:

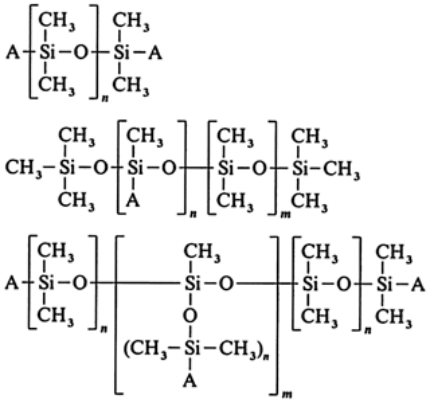

Polydimethylsiloxanes are prepared by selective hydrolysis of dimethyldichlorosilane:

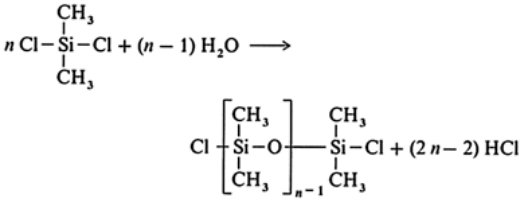

The terminal chlorine atoms may be substituted in the presence of bases by polyethylene glycol residues, resulting in nonionic surfactants of the ethoxylate type.

A special feature of silicone chemistry is the “equilibration” of the polymers, in which molecular mass distribution and distribution of the individual siloxane building blocks can be brought into a reproducible equilibrium by redistribution of the siloxane bonds under acid catalysis. The linear, comb, and branched silicones are formed in this way.

Equilibration enables substituents to be incorporated in the silicone molecule. The silicon – hydrogen bonds of silanes can add to an aliphatic double bond in the presence of hexachloroplatinic acid:

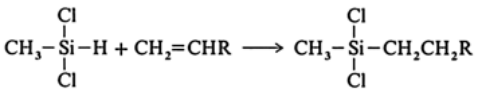

The group R can carry functional groups. By addition of allyl alcohol, for example, hydroxyl- containing dichlorosilanes can be produced, which are converted by selective hydrolysis to poly(methylhydroxypropyl)siloxane, which can be equilibrated with a polydimethylsiloxane. The hydroxyl groups introduced in this way can be reacted further, for example by sulfation to give sulfuric hemi-esters, which yield anionic surfactants.

The choice of the siloxane building blocks, selective hydrolysis, and equilibration enable a wide variety of silicone surfactants to be synthesized. The polysiloxane residues of most silicone surfactants have a polymeric character, in contrast to the previously discussed hydrocarbon-based surfactants. Polar and nonpolar groups alternate in the polymer chain, similar to the polypropylene glycol derivatives or the polypeptides. In the polysiloxanes discussed here, the alternating groups are oxygen bridges and dimethylsilyl groups. The oxygen bridges can interact weakly with water, with the result that a polydimethylsiloxane residue forms a layer on the water surface, with oxygen atoms oriented towards the aqueous phase. The same is true of pure, nonfunctionalized polydimethylsiloxanes, which have a surfactant character because they have a lower surface tension due to the weak mutual cohesive forces of the siloxane residues.

The layer structure of the polysiloxane boundary films, on the other hand, renders these films a certain strength. Polysiloxanes modified with polyethers are therefore important as integrated constituents of polyurethane foams, in which they act as foam stabilizers. The stability of films of silicone surfactants in aqueous foams makes them useful as additives to aqueous fire-extinguishing agents. Although silicone surfactants form stable foams, they are able to destroy foams of hydrocarbon-based surfactants, which they displace from the surface because of their higher surface activity.

The polysiloxane skeleton is biologically persistent, and is broken down only slowly in the environment by hydrolysis. However, polydimethylsiloxanes are physiologically harmless. Silicone-based surfactants are used in a wide range of special applications, for example, in skincare and haircare preparations [145]; in the production of fibers and textiles [146]; as wetting agents, dispersants, and flow- control agents in the production of paints and coatings [147]; as mold release agents in plastics processing; and as defoaming agents, emulsifiers, demulsifiers, and water repellents.

Producers. Silicone-based surfactants are produced in Europe by Hoechst and Th. Goldschmidt; in the United States by Dow Corning, Goldschmidt Chemical, Rhône-Poulenc Surfactant, Troy Chemical, and Union Carbide; and in Japan by Tosho and Yoshimura Oil Chemical.

10.3. Fluorosurfactants [149-151]

The fluorosurfactants used nowadays can be regarded as analogues of surfactants containing aliphatic hydrocarbon groups, in which the hydrogen atoms have been wholly or partly replaced by fluorine atoms. The extreme hydrophobicity of fluorinated hydrocarbons means, however, that amphiphilic fluorine compounds with substantially shorter hydrophobic residues than their aliphatic analogs are already surface-active. Whereas a marked interfacial activity in salts of aliphatic carboxylic acids is observed starting with 8 – 9 carbon atoms in the hydrophobic group, perfluorobutyric acid and its salts are already considerably surface-active. The very low cohesive forces between the fluorinated alkyl groups result in extremely low surface tensions of fluorosurfactants. Surface tensions ≤ 20 mN/m can be achieved only with fluorosurfactants. Fluorosurfactants have very low CMCs and are hence already active at very low concentrations, which to some extent compensates for their high cost. Fluorosurfactants have high chemical and thermal stability, and accordingly are used for special applications (high-temperature processes, electroplating techniques). They diffuse rapidly in aqueous solution, and are therefore also suitable for high-speed industrial processes.

Perfluoroalkane carboxylic or perfluoroalkanesulfonic acids are prepared by electrochemical fluorination of the corresponding carboxylic acids or sulfonic acids. Telomerization of perfluoroethylene with, for example, pentafluoroiodoethane as starter molecule yields homologous perfluoro-1-iodoalkanes, which can be converted to carboxylic acids, sulfonic acids, phosphonic acids, or amines. Perfluoroalkenes, possibly branched, are obtained by anionically catalyzed oligomerization of tetrafluoroethylene; a fluorine atom located at the double bond can be electrophilically replaced, in the presence of a base, by alcohol, phenol, ethylene glycol, or olgioethylene glycol to give nonionic surfactants of the ethoxylate type.

Fluorosurfactants are used as wetting, emulsifying, and flow- control agents; as absorption agents for imparting water-repellent and soil-repellent properties to textiles, leather, and paper; in coating photocopying paper; as mold release agents for plastics; and as additives in fire-extinguishing agents.

Producers of fluorosurfactants include Hoechst in Europe and 3 M and Du Pont in the United States.

[Top of Page]

11. Analysis [152-157]

Surface-active substances are identified qualitatively by the fact that their solutions foam and wet fibers or solid surfaces more quickly than pure water, and also by the reduction in surface tension of water, which is recognized by the rapid sinking of a small thin metal or PVC plate carefully laid on the surface of the water, which would otherwise be held on the surface in the absence of surfactants. These test methods are not unambiguous, however, not all surface-active substances should be regarded as surfactants.

11.1. Identification of Surfactants

A number of group-specific detection reactions are known for identifying different types of surfactants, which, however, have become less important with improvements in

instrumental analysis. The property of anionic and cationic surfactants of forming sparingly soluble salts in water can be used as a quick aid; an anionic surfactant, for example a sulfonate, can be precipitated with a cationic surfactant, for example a quaternary ammonium salt, and vice versa. Anionic surfactants produce a precipitate on addition of aluminum acetate, and soaps are converted to water-insoluble fatty acids on acidification. Nonionic surfactants of the ethoxylate type and cationic surfactants form precipitates sparingly soluble in water with phosphotungstic and phosphomolybdic acids, potassium hexacyanoferrate, and Dragendorff's reagent (barium bismuth iodide). Amphoteric surfactants also form precipitates with phosphotungstic and phosphomolybdic acids; generally, they are sparingly soluble at their isoelectric point.

The modern method of choice for identifying surfactants is IR spectroscopy, for which many reference spectra exist. This technique has been refined to such an extent that amounts of less than 1 µg can be measured. UV spectroscopy allows identification of surfactants containing aromatic groups.

Mass spectroscopy, which is generally employed in combination with gas chromatography and high-pressure liquid chromatography, has also become a valuable aid in surfactant analysis, especially after the introduction of soft ionization methods such as direct chemical ionization (DCI), field desorption (FD), and fast atom bombardment, which can also be applied to nonvaporizable substances. Although mass spectroscopy is complex and therefore unsuitable for routine investigations, NMR spectroscopy, especially ^{1}H NMR, has become an extremely useful tool in surfactant analysis, for example, in identifying hydrophobic hydrocarbon parent substances and in determining the number of added ethylene oxide or propylene oxide molecules in alkoxylates.

Gas chromatography is now an important method in surfactant analysis. Substances of low volatility are often modified chemically for GC analysis. Ethoxylates, for example, are determined as trimethylsilyl ethers; homologues containing up to seven ethylene oxide units per mole can be determined in this way. Fatty acids and sulfonic acids are subjected to GC in the form of their methyl esters. Gas chromatography and mass spectroscopy are suitable for identifying hydrophobic residues of surfactants after specific pyrolytic decomposition. With concentrated phosphoric acid at 215 °C the structurally unchanged alkylbenzene is obtained from the corresponding alkylbenzenesulfonate, while alkanes shortened by the hydrophilic group are obtained from alkyl sulfates, fatty acid esters, and fatty acid amides. Alkanesulfonates and α-sulfocarboxylic esters produce nonuniform fragments in this type of pyrolysis. In the pyrolysis of alkylbenzenesulfonate or alkanesulfonate in a KOH – NaOH melt at 290 °C, the parent hydrocarbons are obtained as alkylphenols or alkenes. Quaternary ammonium compounds are subjected to Hofmann degradation with potassium hydroxide or sodium hydroxide, in which alkene and tertiary amines are formed. Decomposition and GC separation can be coupled on-line (pyrolysis – gas chromatography). Decomposition is often performed as Curie point pyrolysis to ensure reproducible fragmentation under controlled conditions.

Iodoalkane, ethylene, and iodine are formed when ethoxylates are reacted with boiling 58 % hydroiodic acid:

$$\mathrm{RO-[CH_2CH_2O]_n-H} + (2n+1)\mathrm{HI} \rightarrow \mathrm{RI} + n\mathrm{CH_2{=}CH_2} + n\mathrm{I_2} + (n+1)\mathrm{H_2O}$$

The parent substances and their simple derivatives obtained from the reaction can also be identified by gas chromatography or mass spectrometry.

The decomposition of ethoxylates with hydroiodic acid is also utilized in the wet chemical determination of the degree of ethoxylation *n* of an ethoxylate; the iodine released is titrated with thiosulfate solution.

A modification of this Zeisel cleavage converts the ethoxy groups to ethyliodide and propoxy groups to propyliodide which again can be determined by GC.

11.2. Isolation and Separation

Surfactants often exist as a mixture of different types of surfactant or mixed with nonsurfactant components. A number of methods are available for separating and specifically isolating individual surfactants and for removing byproducts.

Extraction. Nonpolar constituents (neutral oil) can be extracted with petroleum ether from neutral or alkaline-adjusted aqueous or aqueous-alcoholic solutions of anionic surfactants. If the solution is acidified, fatty acids also pass into the petroleum ether phase.

Surfactant sulfonic acids are extracted from aqueous hydrochloric acid sulutions with diethyl ether. Short- chain alkylbenzenesulfonic acids, disulfonic acids, etc., remain in the aqueous phase, from which they can be extracted with butanol.

Polyglycols (polydiols) can be extracted from solutions of ethoxylates in ethyl acetate by using aqueous sodium chloride solution. Ethoxylates are extracted from aqueous solution by chloroform after saturating the solution with sodium chloride.

Blowing Out. Blowing out is a nonspecific method for isolating traces of all surface-active substances from aqueous solutions. A stream of nitrogen loaded with ethyl acetate vapor is passed through a frit into the solution. Surfactants are adsorbed on the surfaces of the nitrogen bubbles and transported upwardly, where they are absorbed in an ethyl acetate phase covering the aqueous solution. This method is used in trace analysis of surfactants, especially in effluents and wastewaters.

Adsorption. Surfactants can be separated by adsorption on aluminum oxide, activated charcoal, polymer resins, or silica gel. Using solvents of increasing polarity it is possible, by column chromatography, to separate nonionic surfactants from anionic surfactants, and nonionic surfactants of different HLB value from one another. Reverse-phase cartridges have been used to recover traces of surfactant from aqueous solution.

Ion exchangers are being used increasingly in surfactant separation and analysis, and a quantitative analysis procedure has been proposed [152]. Three ion-exchange columns are connected in series: a strongly acidic cation exchanger in the H^+ form, in which quaternary ammonium compounds, betaines, and imidazoline derivatives are retained from ethanolic solution; a strongly basic anion exchanger in the Cl^- form, which retains sulfonates and sulfates; and a strongly basic anion exchanger in the OH^- form, where fatty acids and aminocarboxylic acids are retained. The nonionic surfactants appear in the eluate. The surfactants retained in the columns are eluted with aqueous ethanolic hydrochloric acid or 2 % aqueous isopropanolic sodium hydroxide and then separated further.

A procedure for separating surfactants from heavy-duty detergents that do not contain any cationic surfactants is as follows [158]: the ethanol-soluble material is redissolved in a mixture of isopropanol and water and subjected to a double ion exchange (cation exchanger and macroporous anion exchanger). The eluate contains the nonionic surfactant. The anion exchanger is then charged with ethanol saturated with carbon dioxide, whereby the soaps are eluted. The hydrotropic compounds (e.g., cumenesulfonate) are then isolated with a solution of 1 N ammonium bicarbonate in 80 parts of water and 20 parts of isopropanol, and finally the sulfonates and sulfates are isolated with 0.3 N ammonium bicarbonate solution in 60 parts of water and 40 parts of isopropanol.

The two separation processes described above demonstrate the numerous possibilities provided by ion exchangers in surfactant analysis.

Chromatographic Methods. Besides gas chromatography (GC), thin-layer chromatography (TLC) is an important method [159], which has the advantage of being able to identify directly individual surfactants even in mixtures.

High-performance liquid chromatography (HPLC) and gradient gel-permeation chromatography (GPC) are being increasingly used. Not only can homologous and analogous compounds be separated in this way, but trace constituents can be also detected.

Supercritical fluid chromatography (SCF) has recently become important especially for higher ethoxylates [160], and capillary electrophoresis, which is suitable for separating ionic compounds, is also occasionally used.

11.3. Quantitative Determination

The afore-described methods of identifying and isolating surfactants are often also suitable for their quantitative determination. HCLP has become a powerful tool for quantitative analysis in the last few years. In chromatographic methods, quantiative determination is performed by calibration with reference substances. For extracts or eluates available in larger amounts, this is performed gravimetrically after evaporation or after precipitation with a specific reagent. Also, direct or indirect photometric determination in solution using calibration curves is often employed. The term indirect photometric determination refers to the photometric determination of a substance containing a chromophor that forms a salt, an addition compound or a complex of defined composition with the surfactant and which is quantitatively determined after reaction with the surfactant.

Anionic and cationic surfactants can also be determined by volumetric analysis if the molecular mass is known or assumed. A solution of a surfactant having an opposite electrical charge to that of the surfactant ion is used for titration. Anionic and cationic surfactants generally form sparingly soluble salts in water; the titration can be performed turbidimetrically as a precipitation titration. The determination of the endpoint can also be achieved potentiometrically with tenside-selective electrodes, e. g., by means of a graphite electrode coated with poly(vinyl chloride) [161].

Two-phase Epton titration is universally employed, especially for the volumetric determination of sulfonates and sulfates. The titration of the anionic surfactant is performed in two mutually immiscible solvents (water and trichloromethane) with a cationic surfactant such as *n*- cetylpyridinium chloride or Hyamine 1622. The water-insoluble cation and anion surfactant salt formed in the titration dissolves in the chloroform phase. The end point is indicated by dyes or dye mixtures which form salts with the surfactants and pass from one phase to the other at the end point.

Modified Epton Titration (CIA Method). The anionic surfactant is titrated in the two-phase system water –trichloromethane with Hyamine 1622, a cationic surfactant. The indicator is a mixture of the cationic dye dimidium bromide and the anionic dye Disulfin Blue VN. The anionic surfactant forms a salt with the cationic dye, which dissolves in trichloromethane to give a pinkish-red color. At the end point the Hyamine cation displaces the dimidium cation from the salt and thus from the trichloromethane, the red coloration migrating to the aqueous phase. Excess Hyamine forms a salt with the dye Disulfin Blue, which dissolves in trichloromethane to give a blue color. At the end point the color of the trichloromethane phase thus changes from red through colorless to blue.

Longwell – Maniece Methylene Blue Method. Sulfonate and sulfate surfactants form salts with the cationic dye methylene blue that are insoluble in water but soluble in trichloromethane. Since methylene blue, which is added as its hydrochloride, is practically insoluble in trichloromethane, when the aqueous and trichloromethane phases are mixed the dye passes into the trichloromethane phase only in an amount equivalent to the quantity of anionic species. The surfactant concentration in the trichloromethane extract can then be determined accurately by photometric determination at 650 nm.

The Longwell – Maniece methylene blue method can be combined with the blowing out process for trace analysis of residual surfactants in biodegradation experiments and of surfactants in effluents and wastewaters. As little as 20 µg of alkylbenzenesulfonate can be determined.

The anionic surfactant determined with this method is termed methylene blue active substance (MBAS).

Determination of Cationic Surfactants with Disulfin Blue (Kunkel Method). This method is based on the formation of a trichloromethane soluble salt of a cationic surfactant with the anionic dye Disulfin Blue. A solution of the cationic surfactant in trichloromethane is thoroughly mixed with the aqueous ethanolic solution of the dye. After separating the phases the trichloromethane phase is evaporated, and the residue dissolved in methanol. This solution is determined photometrically at 628 nm. Substance-specific calibration curves are used for evaluation. This method can be used in combination with the blowing out process for trace analysis; 10 µg of cationic surfactant can be determined. The cationic surfactant determined using this method is termed Disulfin Blue substance (DSBS).

Ethoxylates are determined not in a twophase titration, but indirectly via a precipitation reaction.

Determination of Ethoxylates with Dragendorff's Reagent (Wickbold Method). With modified Dragendorff's reagent (barium chloride/potassium tetraiodobismuthate (III) in glacial acetic acid) ethoxylates containing ≥ 6 mol of ethylene oxide per mole form sparingly soluble precipitates in water. The precipitate is filtered, washed with glacial acetic acid, dissolved in ammonium tartrate solution, and acidified with sulfuric acid. The bismuth present in solution is titrated with ethylenediamine tetraacetate solution with xylenol orange as indicator. To calculate the amount of ethoxylate, substance-specific calibration factors are required, which can also be calculated if the mean molecular mass of the surfactant is known sufficiently accurately.

In another type of titration, especially suitable for trace analysis, the precipitate is dissolved in ammonium tartrate and titrated with pyrrolidine dithiocarbamate solution; the end point is determined potentiographically. Using this method allows as little as 50 µg of nonylphenol ethoxylate (10 mol EO/mol) to be detected.

The nonionic surfactants determined by this method are termed bismuth-active substances (BiAS).

The four methods of quantitative surfactant determination outlined above presuppose a knowledge of the mean molecular mass of the investigated surfactant. As calibration or reference substance tetrapropylenebenzenesulfonate is used for anionic surfactants, nonylphenol ethoxylate containing 10 mol of ethylene oxide per mole for nonionic surfactants, and dodecyldimethylbenzylammonium chloride for cationic surfactants.

11.4. Unspecific Additive Parameters

The titrimetric methods discussed above for the analytical, in particular quantitative determination of surfactants are group-specific and are associated with the amphiphilic nature of the surfactants. The methods fail as soon as polar, hydrophilic groups are introduced into the hydrophobic residue, for example, by enzymatic attack. However, especially in investigations on the biodegradation of surfactants, not only is the loss of surface-active properties of interest but also their further degradation, i.e., enzymatic oxidation to water, carbon dioxide, sulfate, nitrate, etc. It is thus important to determine the intermediate products, the metabolites. However, due to the large number of possible metabolites, which in many cases occur only in very small amounts, the effort and cost involved in their individual determination is disproportionately large. To obtain some information about the overall biodegradability of surfactants, attempts were already made early on to determine substance-unspecific additive parameters that are generally employable for characterizing the pollution of rivers and waters by organic substances (BOD and COD values) and, more recently, the TOC and DOC values.

BOD Value. In the measurement of the biological oxygen demand (BOD) the biological decomposition (biodegradation) of a surfactant is simulated by adding it to a mineral nutrient salt solution that has been inocculated with a broad-spectrum bacterial culture. The solutions are incubated at 20 °C in closed flasks; the oxygen content in the initially air-saturated solution is determined after 5, 15, and 30 d. The oxygen consumption calculated from these values is expressed as a percentage BOD 5, BOD 15, or BOD 30 value, relative to the amount of oxygen required for total oxidation of the test substance.

COD Value. In the determination of the chemical oxygen demand (COD) the organic water-borne constituents are characterized by chemical oxidation with potassium dichromate in sulfuric acid solution in the presence of silver ions. The COD values depend largely on the chemical structure. For example, the readily biodegradable compound acetic acid is scarcely attacked by chromic acid. However, in the case of more complex compounds such as surfactants, correlations exist between COD and BOD values for groups of similar surfactants. Thus the quickly and easily performed COD determination can be used to give an indication of the BOD value. The COD is expressed as a percentage of the amount of oxygen calculated for total oxidation.

TOC Value. The total organic carbon (TOC) in an aqueous solution is the most suitable quantity for expressing the extent of mineralization of a substance in aqueous, bacterially infected solution. The TOC value is determined by combustion of the organic substances to carbon dioxide in a hot reactor through which a stream of air is passed and into which a measured water sample is injected. The amount carbon dioxide in the off-gas from the reactor is determined quantitatively by IR spectroscopy and is used to calculate the carbon content in the sample. With the development of analytical equipment that determines carbon automatically in the mg/L range, the TOC method has become an important analytical tool. The degree of biodegradation of a substance can be specified simply as a percentage by determining the TOC value before and after the degradation experiment if the investigated substance is the sole source of carbon for the microorganisms. Since inorganic carbon in the form of dissolved carbon dioxide or carbonate may also be present in the investigated solutions, the sample is acidified before combustion and the inorganic carbon is expelled as carbon dioxide. The TOC analysis can however also be carried out in two parallel determinations, one of which, performed at a high temperature, determines the total dissolved carbon, while the other, performed at a low temperature, determines only the inorganically bound carbon by acidification and sparging with air. The TOC value is then the difference between the total carbon and inorganically bound carbon.

DOC Value. If the carbon determination is carried out by the TOC method after filtering the sample, then the DOC value (dissolved organic carbon) is obtained.

[Top of Page]

12. Utility Evaluation of Surfactants

Besides the purely analytical methods a large number of utility evaluation methods, in some cases standardized, have been introduced for the characterization and quality control of surfactants. A list of the relevant ISO and DIN regulations is given below:

ISO 304:	Surface Active Agents — Determination of Surface Tension by Drawing Up Liquid Films
ISO 607:	Surface Active Agents and Detergents — Methods of Sample Division
ISO 696:	Surface Active Agents — Measurements of Foaming Power — Modified Ross – Miles Method
ISO 697:	Surface Active Agents — Washing Powders — Determination of Apparent Density — Method by Measuring the Mass of a Given Volume
ISO 1063:	Surface Active Agents — Determination of Stability in Hard Water
ISO 1064:	Surface Active Agents — Determination of Apparent Density of Pastes on Filling
ISO 1065:	Non-Ionic Surface active agents obtained from Ethylene Oxide — Determination of Cloud Temperature (cloud point)
ISO 2131:	Surface Active Agents — Simplified Classification
ISO 2174:	Surface Active Agents — Preparation of Water with Known Calcium Hardness
ISO 2267:	Surface Active Agents — Evaluation of Certain Effects of Laundering — Methods of Preparation and Use of Unsoiled Cotton Control Cloth
ISO 2456:	Surface Active Agents — Water Used as a Solvent for Tests — Specification and Test Methods
ISO 4198:	Surface Active Agents — Detergents for Hand Dishwashing — Guide for Comparative Testing of Performance
ISO 4311:	Anionic and Non-Ionic Surface Active Agents — Determination of the Critical Micellization Concentration — Method by Measuring Surface Tension with a Plate, Stirrup or Ring
ISO 4312:	Surface Active Agents — Evaluation of Certain Effects of Laundering — Methods of Analysis and Test for Unsoiled Cotton Control Cloth
ISO 4316:	Surface Active Agents — Determination of pH of Aqueous Solutions — Potentiometric Method
ISO 4317:	Surface Active Agents — Determination of Water Content — Karl Fischer Method
ISO 4319:	Surface Active Agents — Detergents for Washing Fabrics — Guide for Comparative Testing of Performance
ISO 4320:	Non-Ionic Surface Active Agents — Determination of Cloud Point Index — Volumetric Method
ISO 4324:	Surface Active Agents — Powder and Granules — Measurement of the Angle of Repose
ISO 6387:	Surface Active Agents — Determination of the Power to Disperse Calcium Soap — Acidimetric Method (Modified Schoenfeldt

	Method)
ISO 6388:	Surface Active Agents — Determination of Flow Properties Using a Rotational Viscometer
ISO 6836:	Surface Active Agents — Mercerizing Agents — Evaluation of the Activity of Wetting Products for Mercerization by Determination of the Shrinkage Rate of Cotton
ISO 6837:	Surface Active Agents — Water Dispersing Power in Dry Cleaning Solvents
ISO 6839:	Anionic Surface Active Agents — Determination of Solubility in Water
ISO 6840:	Cationic Surface Agents (Hydrochlorides and Hydrobromides) — Determination of Critical Micellization Concentration — Method by Measurement of Counter Ion Activity
ISO 6889:	Surface Active Agents — Determination of Interfacial Tension by Drawing Up Liquid Film
ISO 7535:	Surface Active Agents — Detergents for Domestic Machine Dishwashing — Guide for Comparative Testing of Performance
ISO 8022:	Surface Active Agents — Determination of Wetting Power by Immersion
ISO 8212:	Soaps and Detergents — Techniques of Sampling During Manufacture
ISO 9101:	Surface Active Agents — Determination of Interfacial Tension — Drop Volume Method
DIN 59 900:	Terms and Definitions
DIN 53 901:	Determination of the Wetting Ability of Surfactant Solutions by Measuring the Time required for a Small Piece of Cotton Cloth Added to the Solution to Sink
DIN 53 902:	Determinations of the Foaming Power of Surfactant Solutions by the Perforated Disk impact Method or the Ross – Miles Method
DIN 53 903:	Determination of the Power to Disperse Calcium Soaps for Pure Calcium Water Hardness
DIN 53 904:	Determination of the Ability to Wash Out Lubricants from Textiles
DIN 53 905:	Determination of the Hard Water Resistance of Surfactants
DIN 53 908:	Determination of the Dispersive Power of Surfactants with Respect to Dyes
DIN 53 910:	Preparation of Hard Water having Adjusted Calcium Hardness
DIN 53 911:	Sampling of Powders and Preparation of Size-Reduced Average Samples
DIN 53 912:	Determination of the Bulk Density of Powders and Granules
DIN 53 913:	Determination of the Filling Density of Pastes, Ointments, and Gelatinous Materials
DIN 53 914:	Determination of the Surface Tension
DIN 53 916:	Determination of the Flowability of Powders and Granules
DIN 53 917:	Determination of the Cloud Point of Ethoxylates
DIN 53 918:	Determination of the Krafft Point and Solubility of Ionic Surfactants
DIN 53 921:	Measurement of the Flow Properties by means of a Rotational Viscometer
DIN 53 988:	Examination of the Action of Levelling Agents in Vat Dyes
DIN 53 989:	Determination of the Turbidity Titration Number
DIN 53 990:	Principles of the Comparative Examination of Utility Properties of Detergents for Washing Textiles
DIN 53 993:	Determination of Interfacial Tension by the Stirrup or Ring Method.

Many methods used for testing the detergency and cleansing power of surfactants and surfactant formulations in their solutions have not hitherto been standardized. To evaluate the detergent power in textiles reproducibly soiled test fabrics are washed in the presence of the test substance in washing machines or washing simulators. After the test fabric has been dried and ironed its degree of whiteness is measured as remission of white light (e.g., in a Zeiss Elrepho apparatus), and compared with the remission values of the unwashed soiled test fabric, and of the unsoiled clean test fabric.

To test the cleaning effect on hard surfaces the latter are dirtied with colored greases and treated with solutions of the test substances. The degree of removal of the colorant from the surface is a measure of the cleaning effect. In the plate-rinsing test greasy plates are washed in succession in the solution of the product to be tested. The detergent power is determined on the basis of the number of plates that can be cleaned until the detergent solution no longer contains any foam and the dirt that has been washed off collects visibly on the surface of the rinsing water.

An important property of surfactants and their formulations, mainly for aesthetic reasons, is their color. Slightly colored or even colorless surfactants are preferred. Among the various methods for determining the color, those predominate in which the yellow and brown shades that mainly occur in surfactants as also in other organic substances are determined. In the iodine color value (DIN 6162, standard method C IV 4a of the Deutsche Gesellschaft für Fettwissenschaft, DGF) the color of a solution of the test substance is compared optically with the color of solutions of iodine of various concentrations in aqueous potassium iodide solution and specified as mg iodine/100 mL. In the Hazen color index (APHA, ASTM D-1209–69; A.O.C.S. Td 1b-64) solutions of platinum and cobalt salts of different concentrations are used as reference solution. In the Klett color index the extinction ($E = \log I_0/I$) of blue light of wavelength 420 –430 nm passing through a solution of the test substance is directly measured photometrically and multiplied by the factor 1000; turbidities are also expressed as color indices in this method.

The Lovibond and Gardner color indices provide a differentiated expression of the color of a solution. In the Lovibond color index (standard method C IV 4b of the DGF) the color of a beam of light passing through the solution is compared in the optical field of a colorimeter with a second beam of light passed through yellow, red, and blue filters of adjustable thickness. The color index is specified by three figures, corresponding to the thicknesses of the three filters. The measurement of the Gardner color index (DIN – ISO 4630; A.O.C.S. Td 1a-64), standard method C IV c of the DGF) is based on the same principle, though the reference colors are already standardized as mixed colors and are specified by means of progressively numbered colored glasses.

The above methods have proved their worth in practice even though they have the disadvantage that they are based on a subjective and comparative perception or, as in the case of the Klett index, are restricted to measurements within a narrow spectral range.

Although an objective determination of the color shade in the CIE color triangle is possible by measuring the overall visible spectrum and conversion to tristimulus values, this method is laborious. A simplified method based on the latter for the colorimetric characterization of transparent liquids is proposed in [162].

[Top of Page]

13. Uses [1-3], [6]

The largest proportion of surfactants is used in *detergents* and *cleansing agents* for domestic and industrial use (⟶ Dry Cleaning, ⟶ Cleansing Agents, ⟶ Laundry Detergents, ⟶ Soaps) [4], [163-166]. In universal laundry detergents mainly combinations of alkylbenzenesulfonate, fatty alcohol ethoxylate, and soaps, in addition to other components, are used. In modern washing machines the soap serves primarily not as a surfactant and cleansing agent, but as a foam regulator. Nowadays it is increasingly being replaced by, for example, the more effective silicone oils. Occasionally, tallow fat alkyl sulfate is used to partly or completely replace alkylbenzenesulfonate. Specialty detergents are similar in surfactant composition as regards the components; instead of tallow fat alkyl sulfates, coconut oil alkyl sulfates and coconut oil ether sulfates, which dissolve better in the cold are preferred.

In fabric softeners, cationic surfactants are employed. On account of their good crystallizability sulfosuccinates and olefinsulfonates are employed besides alkylbenzenesulfonate and ethoxylates in foam cleaning agents and dry cleaning agents. Besides dodecylbenzenesulfonates, alkylbenzenes with shorter alkyl chains, C_{14} – C_{17} alkanesulfonates, C_{12} – C_{14} olefinsulfonates, and fatty alcohol ether sulfates are preferably added to liquid rinsing agents and cleansing agents since they produce highly concentrated, clear solutions. Due to their better dermatological compatibility betaines are added to domestic washing up liquids. Alternatively a proportion of the anionic surfactants is replaced by nonionic surfactants such as fatty alcohol ethoxylates or alkyl polyglucosides.

In the industrial cleansing sector particular importance is placed on the degreasing and emulsifying ability, foam behavior, electrolyte compatibility, and chemical stability of the products. Besides alkylbenzenesulfonates, alkanesulfonates, fatty alcohol ethoxylates, alkylphenol ethoxylates, and fatty amine ethoxylates, phosphate esters and low-foam nonionic surfactants such as propylene oxide/ethylene oxide adducts or terminally blocked ethoxylates are also important here.

Mild surfactants that are gentle to the skin play an important role in bodycare products, and especially in cosmetic cleansing agents (⟶ Skin Cosmetics – Surfactants) [167-169]. Such surfactants include ether sulfates and ether carboxylates, sulfosuccinic esters of fatty alcohol polyglycol ethers, protein – fatty acid condensates and the related sarcosinates and taurides, isethionates, as well as ampholytes and betaines, amine oxides, fatty acid polyglycol and sorbitan esters, and alkyl polyglucosides.

A broad field of application of surfactants is the *textile industry*, historically the oldest sector, with its numerous processes such as washing, cleaning, lubricating, sizing, fulling, bleaching, mercerizing, carbonizing, and finishing (⟶ Textile Auxiliaries) [170], [171]. Almost exclusively nonionic surfactants are used for washing wool. Cotton is washed alkaline, for which purpose nonionic and anionic surfactants are suitable (alkylbenzenesulfonates, alkyl sulfates and sulfosuccinates). In lubricating, a process in which the slip of the fibers is improved, mainly fatty acid and fatty alcohol ethoxylates are used. Sulfonates, sulfates, or ethoxylates are used for desizing, in which the mutual adhesion of the fibers is overcome. Soaps or ethoxylates are used in fulling. Bleaching auxiliaries include alkyl sulfates, alkylbenzenesulfonates, and alkylphenol ethoxylates. Mercerization, the treatment of cotton with a concentrated lye, requires wetting agents such as alkanesulfonates, alkyl sulfates, or alkyl phosphates. Alkanesulfonates or alkyl sulfates are also used in carbonizing, i.e., treatment with sulfuric acid. Softening, which improves the handle of textiles, is largely carried out with cationic surfactants.

Surfactants are also required as emulsifiers, dispersants, levelling agents, and swelling agents in textile dyeing (⟶ Textile Dyeing; ⟶ Disperse Systems and Dispersants – Terminology).

The *leather industry* uses nonionic and cationic surfactants as wetting and cleansing agents. Leather oils, which are used to grease leather, contain alkane sulfonates or alkyl sulfates. Cationic surfactants are used in the aftertreatment of leather to make it supple and water-repellent.

Surfactants have a wide range of application as *emulsifiers* (⟶ Emulsions) [172-175] and *dispersants* (⟶ Disperse Systems and Dispersants). The foods industry uses mainly natural substances such as fatty acid salts, glycerides, and fatty acid esters of lactic acid, tartaric acid, and sorbitol (⟶ Foods, 3. Food Additives – Stabilizers ⟶ Foods, 3. Food Additives – Emulsifiers).

Mainly fatty alcohol ethoxylates and fatty alcohol sulfates, fatty acid ethoxylates, and fatty acid esters of glycerol and sorbitol are used in *pharmacy*, where attention must be paid to pharmacological and toxicological innocuousness (⟶ Pharmaceutical Dosage Forms – Quality Assurance) [176].

Emulsifiers are added to *crop-protection* and pest- control agents since the active substances are often only sparingly soluble in water. Emulsifiers widely used here include the sodium, calcium, and diethanolammonium salts of alkylbenzenesulfonic acids, ethoxylates of alkylphenols and condensed phenols, and fatty acid esters of polyglycols, glycerol, and sorbitol. Short- chain alkylbenzene sulfonates and alkylnaphthalene sulfonates are used as wetting agents.

Emulsifiers are required in numerous operations involved in *petroleum production.* The drilling muds used in large amounts for cooling and rinsing in drilling technology contain, for example, alkylbenzenesulfonate or lignin sulfonate to suspend the solids. Lignin sulfonate and naphthalenesulfonate are used as a cement additive in the construction of boreholes. Demulsifiers, e.g., condensation products of propylene oxide and ethylene oxide, ethoxylates of alcohols, fatty acids, alkylphenols, and condensed phenols, are indispensable in working up the oil, which is initially obtained as a water – oil emul-sion.

Emulsions of water and mineral oil are widely used in *metalworking* and machining. They act as lubricants and coolants in drilling, cutting, rolling, and drawing. Preferred emulsifiers include the heavy metal salts of fatty acids and naphtenic acids, as well as salts of phosphoric esters or carboxymethylated ethoxylates.

Quenching oils for steel hardening contain ethoxylates of fatty alcohols or alkylphenols. Emulsions prepared with ethoxylates or sulfonates are suitable as mold release agents in casting and founding. Alkaline-adjusted mineral oil emulsions, prepared using ethoxylates based on alkylphenols or fatty amines, are used as cleansing agents. Emulsions prepared with phosphates, quaternary ammonium salts or ethoxylated fatty acid ethanolamides are used as anticorrosion agents.

In the *mineral oil industry* emulsifiers play a significant role as additives for cleaning carburetors and as flow enhancers and corrosion inhibitors.

Emulsifiers used in petroleum exploration, production and processing [177], and for oil slick control are mainly nonionic dispersants or demulsifiers [178]. A potential use of ethoxylates or modified ethoxylates is for transporting heavy oils and bitumen as low-water emulsions [179], or coal as concentrated coal – water slurries[180].

Almost exclusively cationic surfactants, more specifically amine oxides, fatty amine salts and imidazoline salts, are used for bitumen application in *road construction* work.

In the *cement industry* alkylnaphthalenesulfonates are used as wetting agents, phenol – formaldehyde – sodium bisulfite adducts or lignin sulfonates are used as flow enhancers, and ethoxylates are used to form air pores.

The production of numerous plastics by *emulsion polymerization* is only possible by using emulsifiers (soaps, alkylbenzenesulfonates, alkanesulfonates, alkyl sulfates, primary and quaternary ammonium salts, pyridinium salts, fatty alcohol ethoxylates and alkylphenol ethoxylates, glycerides, sorbitol esters, etc.). Other uses of surfactants in the plastics industry include stabilization of polymer dispersions, production of foamed plastics, mold release agents, and production of microcapsules.

In the production of *paints and coatings* a broad range of emulsifiers is required to disperse pigments and emulsify oils and resins, and to act as flow- control and thickening agents. Suitable dispersants include ethoxylates of alkylphenols, fatty alcohols and fatty acids, sulfated oils and naphthenates; ethoxylates and sulfonates are used as emulsifiers.

Adhesives contain mainly nonionic surfactants, which emulsify constituents of the adhesive, and ensure rapid wetting of the surfaces to be bonded.

In *pulp and paper* production, ethoxylates, alkanesulfonates, alkylnaphthalenesulfonates are used as wetting agents to remove resinous constituents. Various ethoxylates and sulfonates as well as cationic surfactants are used in paper sizing and finishing.

Surfactants are used in *electroplating* to clean metal surfaces and also to accelerate hydrogen evolution at the cathode. Fatty alcohol ethoxylates and sulfates are often used, and in some cases also alkylbenzenesulfonates. Fluorosurfactants are particularly advantageous in electroplating due to their chemical stability.

Fire-extinguishing foams are produced by adding alkyl or alkyl ether sulfates and fatty acid esters to water. Rapid wetting of the material to be extinguished can be achieved by using fluoro surfactants.

Alkylphenol ethoxylates, sorbitol esters, glycerides, alkylbenzenesulfonates, and sulfosuccinic esters are used in the *photographic industry* to produce light-sensitive emulsions.

An increasingly important use of surfactants is *ore flotation* (⟶ Flotation) [181-183].

The applications listed above by no means exhaust the possible uses of surfactants. Further applications are described in [183-186]. A few applications of minor industrial importance are: the catalysis of chemical reactions in micelles and vesicles [17], [187]; phase-transfer catalysis [188]; extractions in multiple emulsions on liquid membranes [189]; ion-pair extraction [190] and coacervate extraction [191], which is preferably used in the separation of biological materials from aqueous solutions; and admicellar chromatography [192], in which micelles adsorbed on a solid surface act as carrier phase.

Finally, the detailed study of microemulsions has opened up a broad application potential for the latter [193-195].

[Top of Page]

14. Economic Aspects

World production of surfactants is ca. 15× 10^6 t/a. Of this, soaps account for ca. 8×10^6 t/a; in many countries they are still the most important surfactant for everyday use. After soaps, the alkylbenzenesulfonates are quantitatively the most important class of surfactants, accounting for ca. 2×10^6 t/a.

Surfactants are mainly marketed as constituents of finished products, together with nonsurfactant components. In many cases their specific consumption cannot be determined exactly on account of lack of knowledge of their content in commercial products, even when consumption data exist for products themselves. Also, most market estimates are limited to specific application sectors. There are therefore no accurate data on the total consumption. Estimated production figures in 10^3 t/a for some surfactants in 1990 were as follows:

Linear alkylbenzenesulfonates	1700
Branched alkylbenzenesulfonates	300
Lignin sulfonates	600
Fatty alcohol ether sulfates	400
Fatty alcohol sulfates	300
Petroleum sulfonates	200
Alkanesulfonates	100
Fatty alcohol ethoxylates	700
Alkylphenol ethoxylates	500
Fatty acid esters	300
Fatty acid alkanolamides	100
Quaternary ammonium compounds	300
Other amine derivatives	100

Surfactant consumption (in 10^3 t/a) according to fields of application in the United States, Japan, and Western Europe in 1982 broke down as follows [196]:

Washing and cleaning	1900
Cosmetics and pharmacy	300
Textiles and fibers	200
Leather and furs	50
Colorants, coatings, and plastics	200
Cellulose and paper	100
Mining, flotation, and oil production	300
Metalworking	130
Building and construction	50
Crop protection and pest control	100
Foods	200
Miscellaneous	400

[Top of Page]

15. Toxicology and Environmental Aspects

15.1. Introduction

Surfactants are used in many areas of human activity, in the home and in commerce, in agriculture, and in industry. Their effects on living organisms are therefore particularly important, even in cases where products, when used according to the instructions, do not come into direct contact with higher life forms. A large number of surfactants come into direct contact with human skin, as constituents of detergents and cleansing agents; their accidental oral ingestion, even if only as residues on washed dishes, therefore cannot be ruled out. Surfactants used as auxiliaries in foods and beverages are therefore subject to legal provisions that require them to be absolutely harmless. The special attention that is nowadays paid to the possible harmful effects of chemicals has meant that some surfactants intended for use in the foods sector are not permitted for use as such in all countries. A characteristic feature or surfactants permitted for use in the foodstuffs sector is that they are overwhelmingly natural products or are obtained from natural products by slight chemical modification. They are mainly derived from glycerides, fatty acids, lactic, tartaric, and citric acids, sugar, and sorbitol.

Irrespective of their intended use, product safety — including environmental protection — is of major importance for all surfactants. The fate and effect of surfactants in rivers and waters are of particular importance in environmental risk assessment, since a large proportion of surfactants is discharged after use into effluent and sewage and ultimately flows into rivers, lakes, and oceans. Here, the degradation of surfactants by microorganisms in natural waters and in sewage plants, which ultimately leads to their complete mineralization is particularly important.

15.2. Toxicology [197-199]

The majority of surfactants now in use — alkylbenzenesulfonates, alkyl ether sulfates, alkyl sulfates, alkanesulfonates, olefinsulfonates, alkyl polyglycol ethers, and nonylphenol polyglycol ethers — are nontoxic or have only a slight acute or chronic toxicity. Also, no toxic effects of the metabolites of these surfactants have been detected in mammals.

The oral LD_{50} values of both anionic and nonionic surfactants is generally in the range from several hundred to several thousand milligrams per kilogram of body weight (comparable to sodium chloride). Betaines, which are used in particular in body cleansing preparations, are nontoxic. The quaternary ammonium compounds, which are known to be bactericidal and fungicidal, exhibit a range of behavior. Whereas dialkyldimethylammonium chlorides used as fabric softeners have LD_{50} values of ca. 5 g/kg body weight and are nontoxic, alkylbenzyldimethylammonium chloride used as a disinfectant has a fairly low LD_{50} value of 0.35 g/kg.

There are few data on the toxicity of surfactants derived from polysiloxanes or fluoroalkanes. Methylpolysiloxanes are physiologically inert, but this is not necessarily the case for phenyl-substituted polysiloxanes.

Nothing is known of the carcinogenic, mutagenic, or teratogenic effects of surfactants.

There is no danger of acute lethal toxicity by resorption of surfactants through the skin or by inhalation. Surfactants can cause skin irritation and skin damage on prolonged action, since they destroy the water-lipid membrane that serves as the external protective layer of the skin, by dissolving individual constituents. The immediate results are swelling, drying, and chapping of the skin; prolonged action of surfactants can lead to eczematous changes in the skin, which recede when the source of irritation is removed.

There are a large number of standardized methods for testing irritation of the skin and mucous membranes. It is important to differentiate between tests under practically relevant conditions at realistic concentrations, and tests used to determine the potential hazard of pure substances, as required, for example, by the German Chemikaliengesetz (Chemicals Act). According to the latter most surfactants should be characterized as “irritant”, a few as “moderately toxic”, while strongly acidic or basic products, including a number of amine derivatives and ammonium salts, as “corrosive” [200]. Such effects detected at high concentration hardly ever occur if surfactants are used according to the instructions, either because such concentrations are not employed, or because the time the surfactant is in contact with the skin is too short to produce an irritant or indeed a corrosive effect.

Skin irritation generally increases in the sequence nonionic, amphoteric, anionic, cationic surfactants. Aromatic rings in the hydrophobic group increase the irritant action. Good wetting agents have a higher irritant potential than poor wetting agents. With anionic surfactants containing a linear alkyl group such as soaps, alkyl sulfates, and olefinsulfates, maximum skin irritation is found for an alkyl chain length of 12 carbon atoms. The incorporation of oligoglycol residues between hydrophobic and hydrophilic groups, as in fatty alcohol ether sulfates, fatty alcohol sulfosuccinates, or fatty alcohol carboxylates, reduces the irritant action of such anionic surfactants.

Nitrogen- containing surfactants irritate the skin less, and the action of sulfonates or sulfates is reduced by using their ammonium salts (e.g., triethanolammonium salts).

The above comments regarding skin irritation also apply to irritation of the mucous membranes, but the mucous membranes are much more sensitive than the skin.

Betaines and amine oxides are extremely weakly irritating to the skin; they reduce the irritant action of anionic surfactants.

On account of their compatibility with the skin and mucous membranes, betaines, fatty acid amidoalkylaminobetaines, amphoteric *N*-alkylamino acids, and fatty acid alkylolamidoether sulfates are playing an increasing role as constituents of body cleansing agents. Condensation products of fatty acids and protein degradation products (oligopeptides) have long been known to be particularly gentle to the skin and mucous membranes.

Allergies and sensitizations due to surfactants are uncommon. If they do occur, it should be checked whether they are caused by accompanying substances or impurities in the surfactants.

15.3. Biological Degradation [201-203]

Organic substances can be degraded aerobically and anaerobically by microorganisms in the presence of water. Since surfactants are discharged after use mainly into effluent and sewage and thus eventually into rivers and seas, their aerobic biodegradation is of primary interest, which is already reflected in Europe in a number of legislative measures

(⟶ Laundry Detergents – General Criteria for the Ecological Evaluation of Detergent Chemicals). In the determination of the biodegradability a distinction is made between the primary attack on the surfactant molecule and total degradation. Primary attack is the loss of interfacial activity due to a structural change in the surfactant molecule. Total degradation is the complete conversion of the surfactant into inorganic substances (e.g., sodium sulfate, carbon dioxide, water) and the incorporation of its constituents into the cellular material of microorganisms.

Primary attack, which has been shown to lead to total degradation, albeit after a long time, is easy to determine (in contrast to total degradation) by measuring analytically the residual surfactant in a degradation test. Intact anionic surfactants are detected as methylene blue-active substances (MBAS), nonionic surfactants as bismuth-active substance (BiAS), and cationic surfactants as DSBAS (see Section Quantitative Determination).

The investigation of the degradation pathways of a surfactant to total degradation is laborious and time- consuming, and up to now has been performed only on a few, industrially particularly important surfactants (e.g., linear alkylbenzene sulfonate; see below). To quickly obtain information on total degradation, additive parameters are nowadays preferably employed (see Section Unspecific Additive Parameters), which can readily be determined in degradation experiments.

15.3.1. *Methods for Determining Biological Degradation*

The test specifications recommended by the OECD have been adopted by the EC states as well as by some other countries. These are the static OECD screening test, which simulates the conditions of surface waters, and the continuous OECD confirmatory test in which the processes occurring in a sewage plant are approximated using activated sludge (⟶ Laundry Detergents – Environmental Exposure Assessment, ⟶ Laundry Detergents – Assessment of Environmental Effects) [204].

Whereas surfactant-specific analysis only provides information on primary attack, the introduction of the DOC method (dissolved organic carbon) provides further, but however only summary information, on the degradation of a surfactant. If the surfactant is the sole source of carbon in the test solution, then the determination of the organically bound carbon in the solution at the start and end of the experiment provides a direct measure of the degradation of the surfactant. The DOC method can be extended to the OECD confirmatory test if two units instead of one are operated, surfactant being added to the inflow of one unit but not to the other. A homogeneous composition of the bioceonosis in both batches is ensured by daily exchange of sludge between the parallel units. Entrainment of surfactant, which could influence the DOC values, is taken into account computationally in the evaluation of the results. Reliable balances and thus information on the degree of degradation of the surfactants can be obtained by this method by employing suitable technical and statistical measures [205].

Although it is complicated and costly, this so- called coupled units test has nowadays become a standard method (⟶ Laundry Detergents – Biodegradability Standard Test Methods). If the overflowing effluent in this continuously operated test is recycled up to fifty times and replenished with fresh surfactant, additional information is obtained on difficultly degradable catabolites or byproducts in the surfactant used (catabolites test) [206].

In addition to these methods a large number of other test methods are used, some of which have been adopted and numbered in the OECD Guidelines as ready, inherent, and simulation tests:

Ready Tests	
Die Away Test	301 A
Modified Sturm Test	301 B
Modified MITI (I) Test	301 C
Modified Closed Bottle Test	301 D
Modified OECD Screening Test	301 E
Inherent Tests	
Modified SCAS Test	302 A
(SCAS = Semicontinuous activated sludge)	
Modified Zahn – Wellens Test	302 B
Modified MITI (II) Test	302 C
Simulation Test	
Coupled Units Test	303 A
Inherent Biodegradability in Soil	304 A

Whereas most of the aforementioned methods determine the biodegradation via DOC analysis, in the closed bottle test the oxygen consumption is measured, while in the MITI and STURM tests the carbon dioxide evolution is measured. Information on widely used test methods is given in Table 16 [207]. In the ready tests such as the closed bottle or Sturm tests, besides information on the biodegradability of the substance information can also be obtained on the ready biodegradability, which applies if a degree of degradation of at least 70 % is achieved after a degree of degradation of 10 % is exceeded within 10 d (Fig. 22).

Table 16. Test methods for determining total degradation (mineralization)

Test method	Characteristics of the test	Analysis	Test duration, d	Degree of degradation to be regarded as degradable
Modified OECD screening test	static test (small amount of organisms)	DOC	28	> 70 %
Closed bottle test	static test (small amount of organisms)	O_2	28	> 60 %
Sturm test	static test (small amount of organisms)	CO_2	28	> 60 %
Zahn – Wellens test	static test (activated sludge)	DOC	28	> 70 %
Coupled units test	continuous test (simulation of a communal clarification plant)	DOC	≥ 30	> 70 %
Catabolites test	continuous test with recycling and replenishment of the purified effuent	DOC	≥ 30	the necessary degree of degradation depends on the carbon content of the test substance
ECETOC test	static test (digested sludge); anaerobic conditions	CH_4 and CO_2	≥ 30	not specified

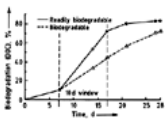

Figure 22. Curves illustrating the DOC decrease of two readily biodegradable substances [207]

All the methods mentioned here refer to the aerobic biodegradation of surfactants. Since the surfactants used nowadays in consumer products undergo rapid aerobic degradation, their anaerobic degradation plays only a minor role in ecological considerations. Accordingly, at present there are still no validated methods for determining anaerobic degradation. Occasionally, the ECETOC test is used.

A further, very informative but complex and expensive method is the use of radioactively labelled (generally ^{14}C) material in a degradation experiment and identification of the radioactive end products.

A specialist report has summarized a large amount of data on biogradation and other ecologically relevant data concerning surfactants [208]; this information is reproduced in abbreviated and supplemented form in Table 17.

Table 17. Biodegradability and ecotoxicity of important surfactants [208], supplemented

Surfactant	Confirmatory test [a], %	Coupled units test [b], %	Mod. OECD screening test [b], %	Closed bottle test[c], %	Sturm test [d], %	Fish toxicity LC_{50}, mg/L	Daphnia toxicity EC_{50}, mg/L	Algae toxicity mg/L
Linear C_{10} – C_{13} alkylbenzenesulfonate	93 – 97	92	94	55 – 65	45 – 76	3 – 9	9 – 14	50[e]
C_{14} – C_{17} alkanesulfonate	97 – 98	99	88 – 96	63 – 95	56 – 91	3 – 24	9 – 13	96
C_{12} – C_{18} fatty alcohol sulfate	98 – 99	97	88 – 96	63 – 95	64 – 96	3 – 20	5 – 70	60
Dioctyl sulfosuccinate	96	49	96	50		39	33	
C_{14} – C_{18} α-olefinsulfonate	98	70 – 78	85	85	65 – 80	2 – 20	5 – 50	10 – 100
Soap				85 – 100	80 – 99		strongly dependant on water hardness	
C_{12} – C_{18} fatty alcohol + 10 EO	93 – 98	95 – 96	94	69 – 86		5	4	3
C_{16} – C_{18} fatty alcohol + 5 EO	95	>80			65 – 75	30	2	
C_{16} – C_{18} fatty alcohol + 30 EO	98	93		27	16	>1000		
C_{13} – C_{15} oxo alcohol + 10 EO	95	>80	75		3			
Isononylphenol + 9 EO	87 – 97	77 – 90	8 – 17	5 – 10	40	5 – 11	4 – 50	20 – 50
C_{16} – C_{18} fatty amine + 10 EO	97 – 98	6 – 33						
C_{12} – C_{18} fatty acid polyglycol ester + 5/29 EO	92 – 96	71 – 92	100	60 – 80		35 – >100		
Dimethyldistearylammonium chloride	94	108	0	5		1 – 6	0.1 – 1	
1-Methyl-1-octadecylamidoethyl-2-heptadecylimidazolinium methyl sulfate		97	75 – 100	84		1.5	1	3
Ditallowacyloxyethyl hydroxyethylmethyl ammonium methyl sulfate		>90	93	>73		3.3 [f]	87[f]	3[f]

[a] MBAS, BiAS, or DSB.
[b] DOC.
[c] BOD.
[d] CO_2 evolution.
[e] EC_{10}
[f] EC_0 .

15.3.2. *Biodegradation Mechanisms*

Surfactants whose hydrophobic group are derived from hydrocarbons can be oxidized enzymatically under aerobic conditions and thus biodegraded. Surfactants derived from hydrophobic parent substances such as perfluoroalkanes, poly(propylene glycols), or polydimethylsiloxanes are not biodegradable.

The enzymatic attack leading to biodegradation occurs in most surfactants at the hydrophobic group. For ionic surfactants the nature of the hydrophilic group has only a minor influence on the rate of degradation. High biodegradation rates have been found even for bactericidal quaternary ammonium compounds such as cetyltrimethylammonium bromide, benzyldodecyldimethylammonium chloride, and distearyldimethylammonium chloride. With nonionic surfactants of the ethoxylate type the degradation rate is influenced by the length of the poly(ethylene glycol) chain (i.e., by the degree of ethoxylation); with the same hydrophobic residue, the degradation rate decreases with increasing length of the poly(ethylene glycol) chain.

The constitution of the hydrophobic hydrocarbon group of a surfactant has a decisive influence on the degradation rate.

Branchings in an aliphatic group strongly decelerates degradation. Phenylidene groups between the aliphatic and hydrophilic groups can make degradation slower. Linear alkylbenzenesulfonates are degraded more slowly than alkanesulfonates and olefinsulfonates, and the latter more slowly than linear alkyl sulfates.

Natural fatty acids and their salts as well as some of their esters with glycerol, sorbitol or sugars are readily biodegradable, although in the case of the esters a decrease in the degradation rate is already observed, which can be explained by the need for a preceding hydrolysis. The enzymatic oxidative degradation of linear fatty acids takes place mainly by β-oxidation, in which the hydrocarbon chain is shortened each time by two carbon atoms.

Aerobic biodegradation of secondary alkanesulfonates probably begins with oxidative desulfonation; the resulting oxoalkanes are readily degraded.

In the case of internal sulfonates, alkylbenzenesulfonates, and alcohol ethoxylates, microbial degradation can be initiated by ω-oxidation, i.e., by oxidative attack on both terminal methyl groups. After introduction of the terminal carboxylate groups further degradation then proceeds as β-oxidation. In the degradation of alkylbenzenesulfonates sulfophenylalkanecarboxylic acids are formed as intermediates, while in the degradation of fatty alcohol ethoxylates polyglycol ethers of ω-hydroxyalkanecarboxylic acids are formed as intermediates.

The biodegradation of the industrially most important surfactant, alkylbenzenesulfonate (also named dodecylbenzenesulfonate since the mean number of carbon atoms in the alkyl group is 12) has been largely elucidated. The degradative pathway proceeds by ω-oxidation and subsequent β-oxidation, followed by ring opening. The sulfonate group remains bound to the carbon atom after opening of the aromatic ring. The homologous and isomeric species contained in the alkylbenzenesulfonate are attacked at various rates. According to the “distance principle”, the degradation rate increases with increasing distance of the terminal methyl group from the coupling site of the sulfophenyl group, or the length of the available unsubstituted alkyl chain. For example, 2-dodecylbenzenesulfonate is degraded more rapidly than 6-dodecylbenzenesulfonate, which in turn is degraded more quickly than 6-undecylbenzenesulfonate.

That ca. 70 % of the benzene nuclei in industrial alkylbenzenesulfonate are already degraded under the conditions of the OECD confirmatory test has been confirmed by examining the degradation of an alkylbenzenesulfonate labeled in the benzene nucleus with ^{14}C [209].

Tetrapropylenebenzenesulfonate is difficultly biodegradable on account of its highly branched alkyl group.

Nonionic surfactants of the ethoxylate type are also preferentially degraded starting from the hydrophobic group by enzymatic attack on a terminal methyl group. The degradation rate decreases sharply with increasing branching of the hydrophobic alkyl group. With simple branching, which occurs for example in the case of isomeric oxoalcohols, the distance principle applies to the degradation of ethoxylates as it does for linear alkylbenzene sulfonates, whereby the surfactant is degraded more quickly the longer a linear alkyl residue starting from a branching site is:

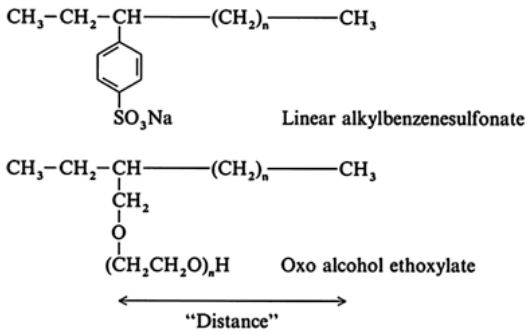

The biodegradation of linear ethoxylates proceeds by two different mechanisms that seems occur simultaneously:

1. The molecule is cleaved at the ether bridge between the alkyl group and the polyethylene glycol ether group, following which both fragments are degraded independently of one another
2. Microbial attack occurs simultaneously at both ends of the molecule, i.e., at the terminal methyl group of the alkyl chain and at the hydroxyl group of the poly(ethylene glycol) unit

The alkyl group of industrial alkylphenol ethoxylates are attacked microbially with difficulty on account of their high degree of branching. In this case the degradation proceeds from the polyglycol chain: The chain is shortened starting from the end of the free hydroxyl group, stepwise in each case by one ethylene glycol unit, glycolic acid presumably being formed as an intermediate:

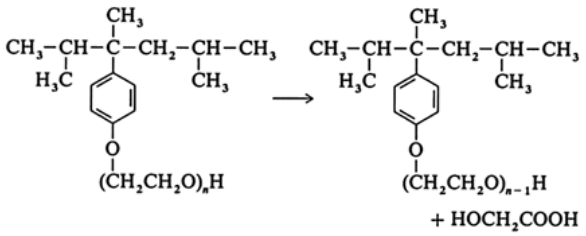

Degradation becomes slower with decreasing length of the polyglycol chain. Since the hydrocarbon skeleton is also degraded with difficulty, the total mineralization of alkylphenol ethoxylates proceeds more slowly than the fast primary degradation.

15.3.3. *Toxicity of Surfactants and their Metabolites to Aquatic Organisms [210]*

The toxicity of surfactants to aquatic organisms can only be evaluated if the rate and completeness of their biodegradation is taken into account. Thus substances with high toxicity will generally not have any harmful effect on aquatic organisms if they are degraded sufficiently quickly.

The toxicity is tested on green algae (Chlorella), water fleas (Daphnia), fish (trout, orfe, goldfish), and mixed bacterial cultures. These trout and young orfe are particularly sensitive to surfactants; therefore, lethal and nonlethal concentrations for these fish are measured to provide preliminary information (screening). In Germany orfe are used for testing; the test method, in which the fish are exposed for 48 h to graduated concentrations of the surfactant in the ambient water, has been standardized. Nonlethal concentrations (48-h LC_0 in mg/L) of some common surfactants for orfe are listed below:

Sodium tetrapropylenebenzenesulfonate	8.0
Sodium dodecylbenzenesulfonate (linear)	2.3
Sodium alkanesulfonate (C_{14} – C_{17})	3.0
Sodium olefinsulfonate (C_{14} – C_{16})	2.0
Sodium tallow fat alcohol sulfate	6.0
Sodium tallow fat alkyl ether sulfate containing 3 EO	10.0
Nonylphenol polyglycol ether with 9 EO	5 – 6
Fatty alcohol polyglycol ether with 5 EO	1 – 2
Fatty alcohol polyglycol ether with 10 EO	2 – 3
Fatty alcohol polyglycol ether with 14 EO	3 – 4

Fish toxicity is strongly dependent on the structure of the surfactant, as exemplified by the structural isomers tetrapropylenebenzenesulfonate (branched side chain) and dodecylbenzenesulfonate (linear side chain). In general, fish toxicity increases with the effective length of the hydrophobic group; in anionic surfactants, branching and an internally located hydrophilic group reduce the toxicity. This is exemplified by the fish toxicity (LC_{50}) of alkylbenzenesulfonate isomers (for *Carassius auratus* [210]):

5-decylbenzenesulfonate	76
2-decylbenzenesulfonate	36
6-undecylbenzenesulfonate	46.5
2-undecylbenzenesulfonate	14.5
6-dodecylbenzenesulfonate	20.5
2-dodecylbenzenesulfonate	4.5
7-tridecylbenzenesulfonate	8.5
2-tridecylbenzenesulfonate	2.0

These toxicity data demonstrate clearly that the toxicity of surfactants cannot be considered separately from their biodegradation in natural waters, since the toxic isomers are degraded most rapidly, (distance principle) and accordingly reach the waters in far lower concentration, if at all, than the less toxic internal isomers.

Except for one striking exception, up to now nothing is known about the aquatic toxicity of catabolites produced in the biodegradation of surfactants. The exception are the isononylphenol ethoxylates, which, because of their marked degree of branching are degraded not from the alkyl residue but from the terminal hydroxyl group of the poly (ethylene glycol) chain, whereby the degradation rate decreases with decreasing number of ethylene oxide units in the poly(ethylene glycol) chain. The diethylene glycol and monoethylene glycol ethers that are formed as intermediates and degrade only slowly and exhibit, like the parent substance isononyl(or isooctyl)phenol, a pronounced toxicity to aquatic organisms, which is much higher than that of alkylphenol ethoxylate surfactants with degrees of ethoxylation of 5 and above [104].

Anionic and nonionic surfactants are generally nontoxic to bacteria. However, reference is made here to a remarkable connection between chemical structure and bacterial toxicity in the case of nonionic surfactants, which in particular cases may result in an inhibition of the biodegradation of some nonionic surfactants. The toxicity of ethoxylates increases sharply with decreasing length of the poly(ethylene glycol) chain (see Table 17). Some toxicity threshold values (TLV in mg/L) for *Pseudomonas* are given below:

Coconut oil fatty alcohol polyglycol ether with 10 EO	1000

Coconut oil fatty alcohol polyglycol ether with 2 EO	10
Stearyl alcohol polyglycol ether with 25 EO	10 000
Stearyl alcohol polyglycol ether with 10 EO	100
Stearyl alcohol polyglycol ether with 2 EO	10
Nonylphenol polyglycol ether with 10 EO	1 000
Nonylphenol polyglycol ether with 6 EO	500
Nonylphenol polyglycol ether with 4 EO	50

A number of quaternary ammonium compounds and betaines have a substantial bactericidal activity. However, if sufficiently diluted these compounds, too, are biodegraded if the chemical structure of the hydrophobic group allows enzymatic attack to occur.

15.4. Preservation of Surfactants [211], [212]

The biodegradation of surfactants occurs not only at high dilution in waters and rivers, but in the presence of water can already start in the production process, in storage and transportation vessels, and in formulations. Bacteria cannot thrive on anhydrous surfactants or in highly concentrated solutions and pastes; however, bacterial infection must already be expected in 20 to 30 % solutions of anionic or nonionic surfactants. Also, more highly concentrated solutions or pastes that tend to coacervate with the formation of low concentration phases can act as nutrient media for bacteria and exhibit signs of putrefaction. Bacterial infection must also be expected if highly concentrated surfactant formulations are handled incorrectly; small residues of surfactants in open vessels can be diluted by absorption of atmospheric moisture, and dilute surfactant solutions can accumulate on the floor or in dead spaces of vessels that have not been thoroughly cleaned with water. Cationic surfactants and amphoteric surfactants have a bactericidal action at and below the isoelectric point; except in highly diluted solutions (ppm range), a biodegradation of these surfactants can be ruled out.

Sterile conditions must be maintained to prevent premature biodegradation of surfactants during production. A sterile environment is generally ensured by maintaing high concentrations, high temperatures, and by using chemical bleaches. Transportation vessels can be filled in the hot state and thus under sterile conditions; the vessels must then be sealed in an air-tight and germ-proof manner. To protect surfactants against putrefaction during processing and in the dilute aqueous state (e.g., washing up liquids or shampoos), preservatives are added. Formaldehyde (0.1 – 0.3 %) is an excellent preservative, but has been discredited in some countries on account of its alleged carcinogenicity. Formaldehyde-releasing substances can also be used, for example, 2-Nitro-2-bromopropane-1,3-diol. 2-Methyl-3-isothiazolone, 5- chloro-2-methyl-3-isothiazolone, and 1,3-dicyano-1,2-dibromobutane are used in ppm amounts. 2-Phenoxyethanol, methyl and ethyl *p*-hydroxybenzoates and sorbic acid are effective at higher concentrations (ca. 0.5 %).

[Top of Page]

References

General References

1. K. Lindner: *Tenside-Textilhilfsmittel-Waschrohstoffe*, Wissenschaftl. Verlagsges., vols. 1 and 2, Stuttgart 1964; vol. 3, 1971.
2. G. Gawalek: *Tenside*, Akademie-Verlag, Berlin 1975.
3. H. Stache (ed.): *Tensid-Taschenbuch*, 2nd ed., Hanser Verlag, München 1981.
4. J. Falbe (ed.): *Surfactants in Consumer Products, Theory, Technology and Application*, Springer Verlag, Heidelberg 1987.
5. M. J. Rosen: *Surfactants and Interfacial Phenomena*, 2nd ed., J. Wiley & Sons, New York 1989.
6. K. Kosswig, H. Stache (eds.): *Die Tenside, Physik, Chemie, Anwendungstechnik*, Hanser Verlag, München 1993.
7. CESIO, Proceedings, World Surfactant Congress, Munich 1984, Paris 1988, London 1992.
8. *Mc Cutcheons Detergent and Emulsifiers*, 1992, International, published annually, with index of producers and trade names.

Specific References

9. R. D. Vold, M. J. Vold: *Colloid and Interface Chemistry*, Addison-Wesley Publ. Co., London 1983.
10. P. C. Hiemenz: *Principles of Colloid and Surface Chemistry*, 2nd ed., Marcel Dekker, New York 1987.
11. S. Ross, I. D. Morrison: *Colloidal Systems and Interfaces*, J. Wiley & Sons, New York 1988.
12. J. L. Cayias, R. S. Schechter, W. H. Wade, *ACS Symp. Ser.* **8** (1975) 234.
13. J. K. Borchardt, C. W. Yates, *JAOCS J. Am. Oil Chem. Soc.* **70** (1993) 47.
14. J. A. Mann, E. Matijevic (eds.): *Dynamic Surface Tension and Capillary Waves, Surface and Colloid Science*, vol. **13**, Plenum Press, New York 1984, pp. 202 ff.
15. R. J. Hunter: *Zeta Potential in Colloid Science*, Academic Press, London 1981.
16. C. Tanford: *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, 2nd ed., J. Wiley & Sons, New York 1980.
17. J. H. Fendler: *Membrane Mimetic Chemistry, Characterizations and Applications of Micelles, Microemulsions, Monolayers, Bilayers, Vesicles, Host-Guest Systems and Polyions*, J. Wiley & Sons, New York 1982.
18. A. L. Lehninger: *Principles of Biochemistry*, Worth Publishers, New York 1982.
19. N. Kosaric, W. L. Cairns, N. C. C. Gray (eds.): *Biosurfactants and Biotechnology*, Marcel Dekker, New York 1987.
20. I. Piirma: *Polymeric Surfactants*, Marcel Dekker, New York 1992.
21. D. Atwood, A. T. Florence: *Surfactant Systems, Their Chemistry, Pharmacy and Biology*, Chapman and Hall, London 1983/1985.
22. S. E. Friberg, B. Lindmann (eds.): *Organized Solutions, Surfactants in Science and Technology*, Marcel Dekker, New York 1992.
23. J. H. Clint: *Surfactant Aggregation*, Blackie & Sons Ltd., Glasgow 1992.
24. H. Hoffmann, W. Ulbricht *Seifen Öle Fette Wachse* **111** (1985) 200 – 202, 467 – 471, 581 – 585; [6] , pp. 1 ff.
25. R. Heusch, *Tenside Surf. Deterg.* **28** (1991) 1; *Naturwissenschaften* **79** (1992) 450. Links
26. J.-H. Fuhrhop, J. Mathieu, *Angew. Chem.* **96** (1984) 124; Links J.-H. Fuhrhop, *Nachr. Chem. Techn. Lab.* **28** (1980) 124.
27. D. Arndt, I. Fischer: *Liposomen, Darstellung-Eigenschaften-Anwendung*, Akademie Verlag, Berlin 1986.
28. H. Ringsdorf, B. Schlarb, J. Venzmer, *Angew. Chem.* **100** (1988) 117.
29. J. H. Clint in [23] , pp. 59 – 81.
30. H. Hoffmann *et al.*, *Tenside Deterg.* **22** (1985) 290.
31. M. Kahlweit, R. Strey, *Angew. Chem.* **97** (1985) 655; *Tenside Surf. Deterg.* **30** (1993) 83.
32. S. E. Friberg, P. Bothorel: *Microemulsions: Structure and Dynamics*, CRC Press, Boca Raton, 1987.
33. J. H. Clint in [23] , pp. 222 – 249.
34. M. J. Rosen, *JAOCS J. Am. Oil Chem. Soc.* **51** (1974) 461.
35. F. Püschel, *Z. Chem.* **3** (1963) 338.
36. C. A. Schirber, *INFORM* **2** (1991) 1062.
37. K. Schulze, *Seifen Öle Fette Wachse* **101** (1975) 37.
38. N. A. I. van Paassen, *Seifen Öle Fette Wachse* **114** (1988) 159.
39. W. Friedrich, H. Riemer, *Chem. Ind. III* **40** (1988) no. 4, 49.
40. O. J. Muscio *et al.*, *JAOCS J. Am. Oil Chem. Soc.* **59** (1982) 217.
41. D. B. Braun, *Cosmet. Toiletries* **104** (1989) Sept., 87.
42. K. Lindner in [1] , vol. 1, pp. 768 ff.
43. F. Asinger: *Die petrolchemische Industrie*, vol. **2**, Akademie Verlag, Berlin 1975, pp. 1249 ff.
44. B. Fell, *Tenside Deterg.* **12** (1975) 3.
45. S. Warwell, *Tenside Deterg.* **19** (1982) 321.
46. B. V. Vora et al. in [7] , vol. 1 (1988) pp. 177 – 189.
47. B. V. Vora *et al.*, *Tenside Surf. Deterg.* **28** (1991) 287.
48. *Eur. Chem. News* **54** (1990) no. 1420, 20.
49. F. Asinger *et al.*, *Erdöl Kohle Erdgas Petrochem.* **20** (1967) 786, 852.
50. L. Cavalli et al. in [7] , vol. 1 (1992) pp. 105 – 114.
51. *Kirk-Othmer*, **22**, 1 – 45.
52. W. H. de Groot: *Sulfonation Technology in Detergent Industry*, Kluwer Academic Publ., Dordrecht 1991.

53. F. Asinger: *Chemie und Technologie der Paraffinkohlenwasserstoffe*, Akademie Verlag, Berlin 1956, pp. 395 ff., 534 ff.
54. H. Ramloch, G. Täuber, *Chem. Unserer Zeit* **13** (1979) 157.
55. J. M. Quack, M. Trautmann, *Tenside Deterg.* **22** (1985) 281.
56. H. G. Hauthal: *Alkansulfonate*, Deutscher Verlag für Grundstoffindustrie, Leipzig 1985.
57. R. Herke, K. Rasheed, *JAOCS J. Am. Oil Chem. Soc.* **69** (1992) 47.
58. A. Boy, R. Brard, H. Passedroit, *Chem. Eng. (N.Y.)* **82** (1975), no. 22, 84.
59. F. Püschel, *Tenside* **4** (1967) 286, 320.
60. H. Baumann, W. Stein, M. Voss, *Fette Seifen Anstrichm.* **72** (1970) 247.
61. A. Mori, M. Nagayama, *Tenside Deterg.* **10** (1973) 64.
62. A. Mori, O. Okumura in [7] , vol. 2 (1984) pp. 93 – 104.
63. I. Yamane in [7] , vol. 1 (1992) pp. 15 – 38.
64. H. Yoshimura, Y. Endo, S. Hashimoto, *JAOCS J. Am. Oil Chem. Soc.* **68** (1991) 623.
65. J. Stapersma, H. H. Deuling, R. van Ginkel, *JAOCS J. Am. Oil Chem. Soc.* **69** (1992) 39.
66. W. Stein, H. Baumann, *JAOCS, J. Am. Oil Chem. Soc.* **52** (1975) 323.
67. H. Andree, G. Jakobi, M. Schwuger in [7] , vol. 4 (1984) pp. 5 – 23.
68. F. Schambil, M. Schwuger *Tenside Surf. Deterg.* **27** (1990) 380; [7] , vol. 2 (1988) pp. 505 –524
69. K. Umehara et al. in [7] , vol. 3 (1992) pp. 234 – 242
70. T. Satsuki, K. Umehara, Y. Yoneyama, *JAOCS J. Am. Oil Chem. Soc.* **69** (1992) 672.
71. K. Schmid et al. in [7] , vol. 2 (1984) pp. 105 – 115.
72. W. M. Linfield in W. M. Linfield (ed.): *Anionic Surfactants,* part II, Marcel Dekker, New York 1976, pp.405 – 443.
73. J. A. Milne in D. R. Karsa (ed.): *Industrial Applications of Surfactants II*, The Royal Society of Chemistry, Cambridge 1990, pp. 76 – 100.
74. T. Schoenberg, *Cosmet. Toiletries* **104** (1989) Nov., 105.
75. B. A. Dombrow in W. M. Linfield (ed.): *Anionic Surfactants*, "part I", Marcel Dekker, New York 1976, pp. 233 – 251.
76. J. K. Weil, A. J. Stirton in W. M. Linfield (ed.): *Anionic Surfactants*, "part I", Marcel Dekker, New York 1976, pp. 219 – 232.
77. S. Shore, D. R. Berger in W. M. Linfield: *Anionic Surfactants*, "part I", Marcel Dekker, New York 1976, pp. 135 – 217.
78. J. Steber, P. Gode, W. Guhl, *Fat. Sci. Technol.* **90** (1988) 32; *Soap Cosmet. Chem. Spec.* **64** (1988) Nov., 44.
79. H. Kölbel, P. Kurzendörfer, *Fortschr. Chem. Forsch.* **12** (1969) 252.
80. W. E. Adam, K. Neumann, *Fette Seifen Anstrichm.* **80** (1978) 392.
81. G. Czichocki, A. Greiner, B. Koch, *Tenside Deterg.* **21** (1984) 62.
82. M. F. Cox, *JAOCS J. Am. Oil Chem. Soc.* **66** (1989) 1637.
83. H. Hensen, H.-C. Raths, W. Seipel, *Seifen Öle Fette Wachse* **117** (1991) 592.
84. G. F. Moretti, I. Adami in [7] , vol. 1 (1988) pp. 236 –261.
85. K. L. Matheson, P. A. Schwab in [7] , vol. 2 (1992) pp. 211 – 220.
86. S. G. Imokawa *et al.*, *J. Am. Oil Chem. Soc.* **55** (1978) 839.
87. H. Nielen, G. Sorbe, *Tenside Deterg.* **16** (1979) 199.
88. S. L. Paul in D. R. Karsa (ed.): *Industrial Application of Surfactants II*, The Royal Society of Chemistry, Cambridge 1990, pp. 114 – 131.
89. M. Hellsten, *Tenside Deterg.* **23** (1986) 337.
90. M. J. Schick (ed.): *Nonionic Surfactants — Physical Properties*, Marcel Dekker, New York 1987.
91. N. Schönfeldt: *Grenzflächenaktive äthylenoxid-Addukte*, Wissenschaftl. Verlagsges., Stuttgart 1976.
92. K. Shinoda, S. Friberg: *Emulsions and Solubilisation*, J. Wiley & Sons, New York 1986.
93. P. Becher (ed.): “Applications,” *Encyclopedia of Emulsion Technology*, vol. **2**, Marcel Dekker, New York 1985.
94. Vista, EP 289 159, 1987 (B. E. Leach *et al.*); Shell International, EP 344 835, 1988 (C. L. Edwards); Union Carbide, EP 361 616, 1988 (S. W. King).
95. H. C. Raths et al. in [7] , vol. 2 (1992) pp. 291 –300.
96. E. Santacesaria et al. in [7] , vol. 2 (1992) pp. 281 –290.
97. A. Behler, U. Ploog, *Fat. Sci. Technol.* **92** (1990) 109.
98. R. R. Geissler, A. E. Johnson, *JAOCS J. Am. Oil Chem. Soc.* **67** (1990) 541.
99. R. R. Geissler, A. E. Johnson, L. D. Talley, *JAOCS J. Am. Oil Chem. Soc.* **68** (1991) 544.
100. M. F. Cox, *JAOCS J. Am. Oil Chem. Soc.* **67** (1990) 599.
101. D. L. Smith, *JAOCS J. Am. Oil Chem. Soc.* **68** (1991) 629.
102. G. Crass, *Seifen Öle Fette Wachse* **118** (1992) 921.
103. Y. Saito, T. Sato, I. Anazawa, *JAOCS J. Am. Oil Chem. Soc.* **67** (1990) 145.
104. C. G. Naylor *et al.*, *JAOCS J. Am. Oil Chem. Soc.* **69** (1992) 695.
105. W. Umbach, W. Stein, *JAOCS J. Am. Oil Chem. Soc.* **48** (1971) 394.
106. P. Straneo in [7] , vol. 1 (1988) pp. 59 – 77.
107. P. Fanelli, S. Ditali, S. Senni in [7] , vol. 1 (1992) pp. 86 – 96.
108. G. M. Leuteritz, *Fat. Sci. Technol.* **93** (1991) 383.
109. R. Piorr *et al.*, *Fat. Sci. Technol.* **89** (1987) 106.
110. K. H. Schmid et al. in [7] , vol. 3 (1988) pp. 633 –645.
111. K. Hoffmann, *Seifen Öle Fette Wachse* **104** (1978) 3.
112. A. Weiss, *Fat. Sci. Technol.* **92** (1990) 392.
113. T. N. Kumar *et al.*, *JAOCS J. Am. Oil Chem. Soc.* **66** (1989) 153.
114. F. R. Benson in M. J. Schick (ed.): *Nonionic Surfactants*, Marcel Dekker, New York 1967, pp. 262 ff.
115. H. Bertsch, F. Püschel, E. Ulsperger, *Tenside* **2** (1965) 397.
116. F. R. Benson in M. J. Schick (ed.): *Nonionic Surfactants*, Marcel Dekker, New York 1967, pp. 266 ff.
117. K. Hill, W. von Rybinski, G. Stoll (eds.): *Alkyl Polyglycosides*, VCH Verlagsgesellschaft, Weinheim, Germany 1997.
118. S. Matsumura *et al.*, *JAOCS J. Am. Oil Chem. Soc.* **67** (1990) 996.
119. D. Balzer, *Tenside Surf. Deterg.* **28** (1991) 419.
120. D. Balzer, N. Ripke, *Seifen Öle Fette Wachse* **118** (1992) 894.
121. H. Rörig, R. Stephan in D. R. Karsa (ed.): *Industrial Applications of Surfactants II*, The Royal Society of Chemistry, Cambridge 1990, pp. 211 – 234.
122. K. R. Smith *et al.*, *JAOCS J. Am. Oil Chem. Soc.* **68** (1991) 619.
123. E. Jungermann (ed.): *Cationic Surfactants*, Marcel Dekker, New York 1970. J. M. Richmond (ed.): *Cationic Surfactants — Organic Chemistry*, Marcel Dekker, New York 1990. D. N. Rubingh, P. M. Holland (eds.): *Cationic Surfactants — Physical Chemistry*, Marcel Dekker, New York 1991.
124. G. Täuber, A. May, *Tenside Deterg.* **19** (1982) 151.
125. S. Billenstein, G. Blaschke, *JAOCS J. Am. Oil Chem. Soc.* **61** (1984) 353.
126. E. V. Dehmlow, S. S. Dehmlow: *Phase Transfer Catalysis*, Verlag Chemie, Weinheim, Germany 1980.
127. Henkel, GE 2 853 241, 1978 (W. Giede, H. Rützen).
128. D. J. Versteeg *et al.*, *Chemosphere* **24** (1992) 641. Links
129. L. Godefroy in [7] , vol. 3 (1988) pp. 415 – 429.
130. R. Puchta, P. Krings, P. Sandkühler *Tenside Surf. Deterg.* **30** (1993) 186; [7] , vol. 3 (1992) pp. 122 – 130.
131. J. Waters *et al.*, *Tenside Surf. Deterg.* **28** (1991) 377.
132. M. M. Watts, *JAOCS J. Am. Oil Chem. Soc.* **67** (1990) 993.
133. H. Hitz *et al.*, *Seifen Öle Fette Wachse* **109** (1983) 20, 43.
134. B. R. Bluestein, C. L. Hilton (eds.): *Amphoteric Surfactants*, Marcel Dekker, New York 1982.
135. U. Ploog, *Seifen Öle Fette Wachse* **108** (1982) 373.

136. R. E. Bilbo, R. A. Swenson, *Soap Cosmet. Chem. Spec.* **66** (1990) April, 46.
137. R. Ohme, D. Ballschuh, H. Seibt, *Tenside Surf. Deterg.* **28** (1991) 180, 235, 337.
138. Li Zongshi, Zhang Zhúangyú, *Tenside Surf. Deterg.* **31** (1994) 128.
139. H. Pardun, *Fat. Sci. Technol.* **91** (1989) 45.
140. D. E. Billek, *Seifen Öle Fette Wachse* **117** (1991) 617.
141. I. R. Schmolka, *JAOCS J. Am. Oil Chem. Soc.* **54** (1977) 110.
142. M. A. Plant in R. D. Karsa (ed.): *Industrial Applications of Surfactants*, The Royal Society of Chemistry, London 1986, pp. 318 – 332.
143. G. L. F. Schmidt in D. R. Karsa (ed.): *Applications of Surfactants*, The Royal Society of Chemistry, London 1986, pp. 24 – 32.
144. B. Grüning, G. Koerner, *Tenside Surf. Deterg.* **26** (1989) 312. I. Yilgör, E. Yilgör, B. Grüning, *Tenside Surf. Deterg.* **30** (1993) 158.
145. D. Schaefer, *Tenside Surf. Deterg.* **27** (1990) 154.
146. G. Schmidt, *Tenside Surf. Deterg.* **27** (1990) 324.
147. H. F. Fink, *Tenside Surf. Deterg.* **27** (1990) 306.
148. M. Gradzielski *et al.*, *Tenside Surf. Deterg.* **27** (1990) 366.
149. E. Kissa: *Fluorinated Surfactants, Synthesis, Properties, Applications*, Marcel Dekker, New York 1994.
150. H. C. Fielding in R. E. Banks (ed.): *Organofluorine Chemicals and their Industrial Applications*, Ellis Horwood Ltd., Chichester 1979, pp. 214 – 234.
151. J. N. Meussdoerffer, H. Niederprüm, *Chem. Ztg.* **104** (1980) 45.
152. H. König: *Neuere Methoden zur Analyse von Tensiden*, Springer Verlag, Berlin 1971.
153. E. Jungermann (ed.): *Cationic Surfactants*, Marcel Dekker, New York 1970.
154. J. Cross (ed.): *Anionic Surfactants, Chemical Analysis*, Marcel Dekker, New York 1977.
155. J. Cross (ed.): *Nonionic Surfactants, Chemical Analysis*, Marcel Dekker, New York 1987.
156. T. M. Schmitt: *Analysis of Surfactants*, Marcel Dekker, New York 1992.
157. G. Krusche in [6] , pp. 251 – 279.
158. R. Wickbold, *Tenside Deterg.* **13** (1976) 177.
159. H. Jork et al.: *Dünnschichtchromatographie,* vol. 1 a, VCH, Weinheim 1989.
160. H. J. Kalinoski, A. Jensen, *JAOCS J. Am. Oil Chem. Soc.* **66** (1989) 1171.
161. W. Selig, *J. Chem. Educ.* **62** (1984) 80.
162. F. J. Gohlke, *Tenside Deterg.* **23** (1986) 213.
163. W. G. Cutler, R. C. Davis (eds.): *Detergency, Theory and Test Methods*, **parts I (1972), II (1975), and III (1981)**, Marcel Dekker, New York.
164. W. G. Cutler, E. Kissa: *Detergency*, Marcel Dekker, New York 1987.
165. I. Geke et al. in [6] , pp. 281 – 337.
166. G. Jakobi, A. Löhr: *Detergents and Textile Washing, Principles and Practice*, VCH Verlagsgesellschaft, Weinheim, Germany 1987.
167. G. A. Nowak: *Die kosmetischen Präparate,* vol. 2, Verlag für chem. Industrie H. Ziolkowsky AG, Augsburg 1984.
168. M. M. Rieger (ed.): *Surfactants in Cosmetics*, Marcel Dekker, New York 1985.
169. A. Reng in [6] , pp. 339 – 356.
170. A. Chwala, V. Anger, C. Chwala: *Handbuch der Textilhilfsmittel*, Verlag Chemie, Weinheim, Germany 1977.
171. A. Datyner: *Surfactants in Textile Processing*, Marcel Dekker, New York 1983.
172. K. J. Lissant: *Emulsions and Emulsion Technology*, **parts I and II (1974), part III (1984)**, Marcel Dekker, New York.
173. P. Becher (ed.): *Encyclopedia of Emulsion Technology*, vol. **1**, “Basic Theory,” 1983; vol. 2, “Applications,” 1985; vol. 3, “Basic Theory/Measurements/Applications,” 1988, Marcel Dekker, New York.
174. K. J. Lissant: *Demulsification*, Marcel Dekker, New York – Basel 1983.
175. R. Heusch in K. Kosswig, H. Stache (eds.): *Die Tenside, Physik, Chemie, Anwendungstechnik*, Hanser Verlag, München 1993 , pp. 203 – 249; W. Adam, R. Heusch in K. Kosswig, H. Stache (eds.): *Die Tenside, Physik, Chemie, Anwendungstechnik*, Hanser Verlag, München 1993 , pp. 357 – 408.
176. P. N. P. Chow (ed.): *Adjuvants and Agrochemicals*, CRC Press, Boca Raton, Florida 1989.
177. P. H. Ogden (ed.): *Chemicals in the Oil Industry, Developments and Applications*, The Royal Society of Chemistry, Cambridge 1991.
178. R. Leo, R. H. Lilie: *Ölwehrhandbuch, Bekämpfung von Ölunfällen im Inland und auf See*, K. O. Storck Verlag, Hamburg 1987.
179. Petroleos de Venezuela Europe S. A.: *Imulsion, the Technology for Handling Extra Heavy Crudes and Bitumens; Orimulsion, Characteristics and Combustibility*, The Orinoco Project, London 1987.
180. H. Riemer in [7] , vol. 4 (1988) p. 15.
181. J. Leja: *Surface Chemistry of Froth Flotation*, Plenum Press, New York 1982.
182. A. N. Clarke, D. J. Wilson: *Foam Flotation, Theory and Application*, Marcel Dekker, New York 1983.
183. P. Somasundaran, B. M. Moudgil (eds.): *Reagents in Mineral Technology*, Marcel Dekker, New York 1988.
184. M. J. Rosen (ed.): *Surfactants in Emerging Technologies*, Marcel Dekker, New York 1987.
185. D. T. Wasan, M. E. Ginn, D. O. Shah (eds.): *Surfactants in Chemical/Process Engineering*, Marcel Dekker, New York 1988.
186. J. F. Scamehorn, J. H. Harwell (eds.): *Surfactant Based Separation Processes*, Marcel Dekker, New York 1989.
187. M. Grätzel, K. Kalyanasundaram (eds.): *Kinetics and Catalysis in Microheterogeneous Systems*, Marcel Dekker, New York 1991.
188. E. V. Dehmlow, S. S. Dehmlow: *Phase Transfer Catalysis*, 2nd ed., Verlag Chemie, Weinheim, Germany 1983.
189. R. D. Noble, J. D. Way: *Liquid Membranes, Theory and Application*, American Chemical Society, Washington, D.C., 1987.
190. M. Cox: *Chem. Ind. (London)* 1981, 355.
191. N. D. Gullickson, J. F. Scamehorn, J. H. Harwell in [186] , pp. 139 – 152.
192. J. H. Harwell, E. A. O'Rear in [186] , pp. 155 – 171.
193. D. O. Shah (ed.): “Macro- and Microemulsions, Theory and Applications,” *ACS Symp. Ser.* **272** (1985) .
194. G. Gillberg in [172] , part III, pp. 1 – 43.
195. R. Leung, M. J. Hou, D. O. Shah in [185] , pp. 315 –367.
196. B. W. Werdelmann in [7] vol. 1 (1984) pp. 3 – 21.
197. F. Bartnik, K. Künstler in [4] , pp. 475 ff.
198. C. Gloxhuber, K. Künstler: *Anionic Surfactants, Biochemistry, Toxicology, Dermatology*, Marcel Dekker, New York 1992.
199. K. Künstler in [6] , pp. 483 – 500.
200. Comité Européen des Agents de Surface et leurs Intermédiaires Organiques, CESIO: *Classification and Labelling of Surfactants*, Bruxelles 1990.
201. R. D. Swisher: *Surfactant Biodegradation*, 2nd ed., Marcel Dekker, New York 1987.
202. N. Nyholm, *Environ. Toxicol. Chem.* **10** (1991) 1237. Links
203. P. Schöberl in [6] , pp. 409 – 464.
204. P. Schöberl, *Tenside Surf. Deterg.* **28** (1991) 1.
205. W. K. Fischer, P. Gerike, W. Holtmann, *Water Res.* **9** (1975) 1131. Links
206. P. Gerike, W. Holtmann, W. Jasiak, *Chemosphere* **13** (1984) 121; Links *Tenside Deterg.* **22** (1985) 305.
207. P. Schöberl, Jahrestagung SEPAWA 1991, pp. 33 – 39.
208. K. J. Bock, L. Huber, P. Schöberl, *Tenside Surf. Deterg.* **25** (1988) 86.
209. J. Steber, *Tenside Deterg.* **16** (1979) 140.
210. P. Schöberl, N. Scholz in [6] , pp. 465 – 482.
211. H.-G. Schmidt, *Parfüm. Kosmet.* **68** (1987) 5.
212. S. B. Singh-Verman, *Parfüm. Kosmet.* **68** (1987) 752.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Surgical Materials

DAVID WILLIAMS, Department of Clinical Engineering, University of Liverpool, Liverpool, United Kingdom

1. Introduction 2
2. Surgical Applications of Materials . 2
2.1. General Principles 2
2.2. Orthopedic Surgery 3
2.2.1. Bone and Joint Replacement 3
2.2.2. Ligament Augmentation and Repair . 4
2.2.3. Fracture Fixation 4
2.3. Maxillofacial Surgery 5
2.3.1. Dental Implants 5
2.3.2. Maxillofacial Trauma 5
2.3.3. Reconstructive Maxillofacial and Orthognathic Surgery 5
2.4. Cardiovascular Surgery 6
2.4.1. Heart Valves 6
2.4.2. Vascular Prostheses 6
2.4.3. Extracorporeal Circuits 7
2.4.4. Cardiac Pacemakers 7
2.5. Ophthalmology 7
2.5.1. Corneal Replacement 7
2.5.2. Intraocular Lenses 7
2.5.3. Contact Lenses 8
2.5.4. Vitreoretinal Surgery 8
2.6. The Nervous System 8
2.6.1. Nerve Repair and Regeneration 8
2.6.2. Nerve Stimulators in Pain Relief . . . 8
2.6.3. Treatment of Hearing Loss 8
2.7. General Surgery 8
2.7.1. Wound Closure and Artificial Skin . . 8
2.7.2. Soft Tissue Reconstruction 9
2.7.3. Surgery of the Urinary Tract 9
2.8. Artificial Organs 9
3. Criteria for Materials Selection . . . 9
3.1. Mechanical and Physical Properties 10
3.1.1. Mechanical Performance 10
3.1.2. Electrical and Optical Properties . . . 10
3.2. Manufacturing and Sterilization . . 10
3.2.1. Constraints on Fabrication 10
3.2.2. Market Analyses 11
3.2.3. Legal and Regulatory Constraints . . . 11
3.2.4. Sterilization Techniques 11
3.3. Biocompatibility 11
3.3.1. General Principles 11
3.3.2. Corrosion and Degradation 12
3.3.3. Soft Tissue Biocompatibility 12
3.3.4. Hard Tissue Biocompatibility 13
3.3.5. Blood Compatibility 14
3.3.6. Systemic Biocompatibility 14
3.3.7. Bioactivity 15
4. Materials 15
4.1. Metals and Alloys 15
4.2. Ceramics and Glasses 16
4.3. Polymers 17
4.4. Carbon 19
4.5. Composites 19
4.6. Surface Modifications and Coatings 20
5. References 20

1. Introduction

Surgery in the late 1900s places heavy demands upon sophisticated technology, ranging from functional imaging equipment to assist in diagnosis, to lasers and robotics to improve accuracy and reliability. One of the most important areas in which technological advances have significantly enhanced surgery involves the use of synthetic materials for implantation in the body to assist in surgical reconstruction and repair. These surgical materials form an important subset of the more general group of materials referred to as "biomaterials", which are defined as materials intended to interface with biological systems to evaluate, treat, augment, or replace any tissue, organ, or function of the body [1].

Within the broad group of biomaterials, a number of subsets, including dental materials and drug delivery materials, may be distinguished (see → Dental Materials and → Pharmaceutical Dosage Forms). Surgical materials are defined in the context of this review as biomaterials used for implantation in the body to assist in surgical reconstruction and repair.

Biomaterials are rarely designed specifically for this use. Instead, they are generally materials developed for other applications but which have been found to be appropriate for medical use, often being modified or optimized for these applications. Thus the materials described in Chapter 4 are generally no different in structure and properties to materials used in chemical, aerospace, nuclear, and other industries. It

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_001

is the matching of the properties of materials to the specific and unique requirements of surgical reconstruction that is the key to biomaterials science.

2. Surgical Applications of Materials

In this section the reasons why materials are used surgically and the criteria for materials selection that can be established on the basis of the functional performance required are reviewed. General reviews of the clinical applications of materials may be found in [2–6].

2.1. General Principles

Biomaterials are implanted in the body for many different reasons but these can be broadly grouped as follows.

Congenital Deformities. Children are often born with a physical deformity which results in functional or aesthetic impairment. In some cases, the consequences may be life threatening, as with certain deformities associated with spina bifida, while in other cases they are more psychological, as in the case of facial disfigurement. The treatment and the precise usage of the materials vary from one situation to another.

Developmental Defects. A child apparently healthy and normal at birth may develop abnormally, perhaps due to impaired hormonal function, leading to defects that require correction. This usually involves the skeletal system, and procedures such as leg-lengthening and spinal curvature correction are included in this category.

Trauma. At any time in life, injuries may be sustained which require surgical intervention as part of the treatment. This may involve the use of materials and devices to assist the body to repair itself, such as sutures to close soft tissue wounds and fracture plates, nails, and screws to facilitate bone repair, or may involve structures that replace or augment damaged tissue, such as artificial skin, bone substitutes, ligament prostheses, and nerve guides.

Disease. The majority of diseases can be treated nonsurgically. However, a significant number affect tissues, leading to pain, loss of movement, or loss of vital function, such that surgery becomes essential. Specific examples involving the use of surgical materials in reconstruction include osteo- and rheumatoid arthritis, which may be treated by total joint replacement; atherosclerosis within the vascular system, which can be surgically treated with artificial arteries; and cataracts of the eye, which can be excised and replaced by intraocular lenses. Of great importance here is the order of preference in treatment: (1) nonsurgical methods (e.g., drugs); (2) minimally invasive procedures in which the defect is treated remotely or endoscopically; (3) surgical reconstruction involving the patient's own tissues, such as bone or skin grafts; and (4) reconstruction involving surgical materials. Thus, these latter procedures are only used when other treatments are ineffective or inappropriate.

Psychological Factors. There are many occasions in which adults find that their appearance causes them psychological problems. Under such circumstances, even when there is no medical or physical disability and where few others would agree that the appearance requires correcting, surgery, usually minor plastic surgery, may be carried out.

Tissue Atrophy. As part of the ageing process, tissues change their shape and consistency. In particular, some tissues lose their form and so augmentation or restoration is desirable. This may relate to loss of bone, especially in osteoporosis.

2.2. Orthopedic Surgery

Orthopedic surgery has historically been the most significant user of biomaterials. It is concerned with the musculoskeletal system and the treatment of diseases and disturbances to its components, particularly bones, cartilage, ligaments, muscles, and tendons. It is this system that produces movement and that responds to forces. It is therefore susceptible to injury and to stress-related diseases, the consequences of

which are significant since they affect the ability to move. Of equal importance is that any replacement of these tissues has to perform in the same mechanical environment, and attainment of the relevant mechanical performance is not a trivial matter.

2.2.1. Bone and Joint Replacement

Several diseases affect the bones and joints but they predominantly include osteo- and rheumatoid arthritis. These conditions affect the tissues of the joints, osteoarthritis being largely a mechanical disruption to the cartilage and underlying bone, while rheumatoid arthritis is an inflammatory disease which causes very painful structural changes to the tissues. Pain relieving drugs and anti-inflammatory agents may be reasonably effective but the diseases are progressive and in many cases, excision of the affected tissue is the only long-term solution.

Since in most cases the disease affects much of the joint, it is usual to resect a considerable amount of tissue and use a total joint replacement prosthesis. In the hip, for example, the first joint to be replaced in this way, it is usual to replace both the head of the femur and the socket within the pelvis (the acetabulum) [7]. In the knee, which is equally affected but more difficult to replace, the lower end of the femur, the upper end of the tibia, and the patella are involved [8].

The main requirements for these joint replacements are the ability to transmit forces under sliding conditions and the ability to be attached to the remaining bone, with durability measured in decades. Combinations of materials that give low friction and high wear resistance for the articulating surfaces are used, for example, a metal or ceramic bearing against ultrahigh molecular mass, high-density polyethylene, high-strength alloys for the main structural components [9], and a system to permit fixation to the bone. In the latter case, this is most often achieved by using a so-called bone cement, generally an in situ curing poly(methyl methacrylate) [10], or sometimes by means of a porous surface which allows ingrowth of bone to achieve stability [11], or bioactive coating that encourages rapid regeneration of bone at the interface [12].

Joint replacements are not without their problems but are generally able to provide pain-free movement for patients for a reasonable length of time. Whilst at one stage success rates, at least in hips, were high with an expectation of 15 – 20 years performance, as the age range in which these prostheses are used has widened, particularly as younger and more active patients have been treated, the demands on them have escalated resulting in reduced performance. An increasingly common problem is that the prostheses become loose. The principle causes of this loosening are the tissue reaction to the wear debris produced at the articulating interface [13], and the difficulty of attaining and retaining an acceptable stress distribution in the prosthesis – bone system [14]. It is for these reasons that attempts to enhance wear resistance and fixation with surface modifications and the introduction of composite materials to facilitate load transfer are being carried out (see Chap. 4).

2.2.2. Ligament Augmentation and Repair [15]

Ligaments are dense bundles of tissue, mostly composed of the structural protein collagen, that are attached to the bones within a joint and provide stability and constrain movement to acceptable limits. Damage to ligaments severely compromises this stability and therefore restricts the controlled movement of the patient. Ligaments rarely suffer disease, primarily because of their simple construction but they are very susceptible to injury, especially sports injuries. Here it is the ligaments of the knee, especially the cruciate ligaments which are frequently affected. Loss of function of these ligaments makes walking impossible.

Surgical treatment can take several forms. It is possible to attempt reconstruction by total replacement of the damaged ligament with a prosthesis; however, matching the elastic characteristics of the natural ligament and producing a reliable anchorage to the bone are difficult. The alternative is to use an augmentation device in which an implant is added to the damaged tissue and works with it to transmit the forces. Materials used in these types of device include polyester fibers, carbon fibers, and expanded polytetrafluoroethylene (PTFE).

2.2.3. Fracture Fixation

Bone possesses an extremely efficient mechanism for self-repair. Thus, when bone fractures, it heals spontaneously such that the newly formed bone is microstructurally indistinguishable from the existing bone. However, the repair only occurs with the appropriate alignment and geometrical arrangement if the fracture surfaces are held together in their correct position, without gross movement, during the healing process, which may take a few months.

With the long bones, such as the femur, tibia, and humerus, correct alignment and stability can be achieved by using a plaster cast. However, there are many occasions were this is not possible, perhaps because there are multiple fragments. In this case, surgical intervention is necessary to achieve the correct juxtaposition of the bones and to hold them in this position. This may be done through the use of various types of plate, nail, screws, or other fixtures. Although there have been a few attempts to use other materials, the vast majority of these devices are made from metals, which must combine the properties of strength, especially fatigue strength, with corrosion resistance and biocompatibility [16].

Because bone is a dynamic tissue that requires the stimulus of mechanical stress for maintenance of its internal structure, problems arise if rigid metal plates are retained within the bone after healing [17]. Ideally, all fracture plates should be removed once healing is complete; because of the slight risks involved with this additional surgical procedure, attention is now on the development of biodegradable fracture plates [18], which is difficult since the requirements for degradability and inherent strength are almost mutually exclusive.

2.3. Maxillofacial Surgery

Closely allied to the problems of the skeletal system are those concerning the structural tissues of the head, mostly involving bone but also the associated soft tissue. Here, however, the causes of the problems are a little different, as are the solutions. There are many subspecialities, including oral surgery, craniofacial surgery, orthognathic surgery, and perhaps more conveniently as an all-embracing term, maxillofacial surgery. The most important aspects follow [19].

2.3.1. Dental Implants

The area of dental implants lies at the interface between surgery and dentistry as far as biomaterials are concerned. The loss of natural teeth, through either disease or trauma, has, for many years, been compensated by the provision of artificial teeth in the form of bridges or dentures. These essentially provide an esthetic replacement of the crown of the tooth but do not replace the root and its attachment to the bones of the jaws. Whilst they are quite satisfactory in many patients, a number of disadvantages exist including the need to interfere with adjacent teeth when a bridge is used, and the problems of denture stability, and so it has long been the ambition of dentists to use implants that replace teeth in their entirety. This is difficult because the implants contact a number of different types of tissue, including an epithelial barrier, must reside in the microbiologically hostile region of the mouth, and are exposed to a complex stress system without much support. For many years, a very poor success rate reflected this difficulty, but more recently, improvements to materials, designs, and techniques have given far better performance and dental implants now represent a very respectable alternative to traditional dentures [20–22]. The materials of choice are titanium alloys (see → Dental Materials, Chap. 2.3.2.) and alumina, often coated with a porous surface or bioactive material such as calcium phosphate (see → Dental Materials, Chap. 2.5.).

2.3.2. Maxillofacial Trauma

Fractures of the facial skeleton have traditionally been treated conservatively; in contrast to fractures of the long bones, which are often compound and displaced, those of the face are usually simple without distortion of shape. The requirements for stability are more easily satisfied, for example, by interdental wiring or external halos which consist of a frame secured to the skull by means of pins. However, it is becoming increasingly recognized that better results and far more convenience to the patient

can be achieved through the use of internal plating. Small titanium or stainless steel plates for the fixation of mandibular or maxillary fractures are now widely used [23]. The requirements for degradability mentioned in Section 2.2.3 also apply here, and the less severe mechanical environment makes this a more realistic possibility.

2.3.3. Reconstructive Maxillofacial and Orthognathic Surgery

The connective tissues of the head and neck also suffer from a variety of diseases and deformities that may require treatment. These range from inflammatory disease of the periodontium, the tissues that support the teeth, to extensive hard and soft tissue tumors. The treatment of advanced periodontitis may involve the implantation of thin strips of a porous material around the root of the tooth, such as expanded PTFE, which facilitate the regeneration of new connective tissue and act as a barrier to infection, or the implantation of artificial bone. The latter approach may also be used in patients who suffer localized osteoporosis, with substantial loss of bone, within mandible or maxilla after tooth extraction.

Tumors of the head and neck present considerable problems of reconstruction [24]. This is partly because prior irradiation treatment used in an attempt to eradicate the tumor severely compromises the ability of the tissue to heal and to accept the presence of an implanted device. If the tumor affects the bone, it is necessary to remove a large amount of that bone, which makes reconstruction by a prosthesis, with attachment to the remaining part of the facial skeleton, very difficult. This is a particularly challenging area of implant surgery.

2.4. Cardiovascular Surgery

Diseases of the circulatory system occur as frequently as those of the musculoskeletal system but, by their very nature, are more life threatening than immobilizing. The diseases can affect the heart, the major circulatory system, and the microvasculature, the consequence of which is a perturbation to the flow of blood. Treatment that does not require major intervention is preferred, for example, pharmacological and endoscopic techniques.

2.4.1. Heart Valves

The valves of the heart control the flow of blood into and out of the chambers. There are four valves of interest: the mitral, aortic, pulmonary, and tricuspid, and their opening and closing mechanisms slightly differ, as do their susceptibility to disease. The mitral and aortic valves are the most critical in this respect, suffering from two major modes of malfunction, stenosis and incompetence. *Stenosis* is an obstruction of the valve orifice which disturbs the pattern of blood flow and increases the pressure necessary to achieve adequate flow. At these high pressures, turbulence becomes more significant and considerable stress is placed on the heart. *Incompetence* arises from the inability of the valve to close properly, allowing a retrograde flow back through the valve, again disturbing the flow profile.

When valvular disease has progressed to such an extent that it compromises the ability of the heart to maintain the circulation, it becomes necessary to remove the offending structure and replace it with a prosthesis [25]. This has been possible for over thirty years. The original valves consisted of a ball held within a cage which was sewn into the heart. These worked reasonably well although not without problems and the valve, both with respect to materials and design, has evolved over the years [26]. The common designs incorporate a flat disc that tilts within a mechanical constraints, or flaps which open and close on a pivot or hinge. Although mechanical failures may occur with such devices, which are usually catastrophic and fatal [27], the main problem arises from the thrombogenicity of the materials used. This means that there is a tendency for the blood to clot on the surface of the valve and, as a consequence, all patients fitted with them have to regularly take anticoagulant drugs.

Because of this, there have been attempts to produce valves that are more compatible with blood and which are able to open and close in a more physiological mode [28]. The development of the so-called bioprosthetic valve that is made of natural tissue, derived usually from the muscle wall of cows hearts or the valves of pigs

has ensured. These valves do not require anticoagulants but are susceptible to slow deterioration and calcification, necessitating periodic replacement [29].

2.4.2. Vascular Prostheses

Atherosclerosis is a condition in which deposits building up along the inner surface of blood vessels, especially the arteries [30]. The resulting stenosis interrupts blood supply and places the heart under considerable strain. Most arteries can be affected. When the coronary arteries are involved, the blood supply to the heart itself is interrupted, leading to a heart attack. If an artery in the lower leg is affected, for example, the femoropopliteal artery, the result is restricted circulation to the feet and, eventually, gangrene. The deposit that forms involves a collection of cells and proteins which can sometimes be broken up using drugs. More usually, by the time the condition is diagnosed, it is not treatable in this way. It may, however, be amenable to endoscopic techniques, which involve the insertion of a device into the vessel, its guidance to the site of the blockage, and the removal of the deposit, either mechanically, or by laser ablation [31]. If these angioplasty techniques fail, or if the problem recurs, then reconstruction may be the only option. Ideally the vascular surgeon will choose to use a vein or artery graft, i.e., transposing a vessel such as the saphenous vein or internal mammary artery to the critical, damaged site [32]. This works well under most circumstances, but it is not always possible to obtain a suitable source and very difficult to find large segments. Under these circumstances, artificial arteries, often incorrectly termed vascular grafts, may be employed [33]. Their design is crucial, as is the material and method of construction chosen for the tubular structure. Without exception, successful artificial arteries employ microporous polymers, either woven or nonwoven, such as expanded PTFE [34] or polyester textiles [35]. Their mechanical characteristics, such as compliance and kink resistance, and their compatibility with blood components are critical.

2.4.3. Extracorporeal Circuits

Although external life support systems are not included here within the definition of surgical materials, it is important to recognize that extracorporeal circulation is extremely common. The best examples are the hemodialysis machines, oxygenators, and liver perfusion equipment. In all these cases, blood is removed from the patient and treated by circulation through tubing, over membranes, or through columns (for example, to remove toxic metabolic products). All the concerns about blood flow and compatibility between blood and materials discussed in relation to implantable devices are also relevant here [36].

2.4.4. Cardiac Pacemakers

Conditions other than the mechanics of blood flow may affect the ability of the heart to function correctly, particularly if there is discontinuity in the conduction of stimuli from the nervous system. If there is such a disturbance, either permanant or transient, then an implantable cardiac pacemaker may be used to supply and deliver signals to the heart muscle to control the heartbeat. Pacemakers consist of a package of microelectronics, usually hermetically sealed inside a welded titanium casing, connected to the wall of the heart by an insulated lead, terminating in an electrode [37].

2.5. Ophthalmology

The structure of the eye is conceptually very simple, consisting of a collection of tissues and fluids that direct and focus incident light onto the retina, where the light energy is converted into electrical signals that are passed to the brain via the optic nerve. Blindness or partial loss of sight can arise from disruption to the light pathway, an inability of the retina to capture the light with appropriate fidelity, or a failure within the nervous system. Problems of the latter nature can rarely be remedied. However, there are procedures that can be used to reconstitute a light pathway that has been disturbed through disease.

2.5.1. Corneal Replacement

A common condition affecting the cornea is glaucoma, where there are inflammatory and structural changes in the tissues which prevent light passing through undisturbed. Glaucoma can be treated by drugs but also by using grafts. Corneal grafting, using corneas derived from cadavers is one of the easiest types of tissue grafting. The cornea, an avascular structure, is far more easily tolerated by the body's immune system. On the rare occasion where corneal grafts are not possible, artificial corneas, the so-called keratoplasty prostheses, typically made of a clear plastic such as poly(methyl methacrylate) (PMMA), are available [38].

2.5.2. Intraocular Lenses

Cataracts, areas of the lens which become cloudy, are extremely common. Treatment involves excision of the lens, which leaves the patient with a severe visual defect. This can be compensated by spectacles, but not efficiently, and the treatment of choice is the intraocular lens [39]. This is a poly(methyl methacrylate) lens, usually discoid in shape, which has attachments around the periphery, made from either the same material or an alternative polymer and known as haptics. The plastic lens is placed within the eye at approximately the same location as the extracted lens. Although at one stage it was difficult to achieve a high level of success with these lenses because of the damage caused to the surrounding tissue and the resulting inflammation, improvements to the surface quality and to the implantation technique have resulted in considerable success for this quite simple structure [40].

2.5.3. Contact Lenses

Although not strictly implantable materials, contact lenses are usually worn as an alternative to spectacles for correction of vision but may also be worn for medical reasons. Contact lenses must have good optical properties and excellent compatibility with the cornea [41]. The cornea is an avascular tissue which derives its oxygen by diffusion through its external surface. This should be compromised by the placement of a plastic structure over the surface. For this reason it is better that the contact lens has high oxygen permeability. It should also produce minimal mechanical irritation to the sensitive corneal surface. Several materials are used for these lenses; hard lenses are usually made of PMMA, while soft lenses may be made of silicone elastomer or a hydrogel such as poly(hydroxyethyl methacrylate) [42].

2.5.4. Vitreoretinal Surgery

A far less common cause of visual defect involves damage to the retina or associated tissues at the back of the eye, mainly detached retinas. A variety of materials, mainly silicones, may be used in attempts to physically relocate and reattach the retina [43].

2.6. The Nervous System

Intuitively, the brain and nerves appear to be tissues far too sensitive to accommodate invasion by surgical materials. However, there are a variety of circumstances in which foreign materials have been placed within the central nervous system without apparent effect. The difficulty is actually achieving anything profitable as far as performance of the system is concerned. One simple but very important application within the brain is the insertion of a catheter into the lateral ventricles to allow the drainage of cerebrospinal fluid in infants with hydrocephalus, a potentially fatal condition associated with spina bifida [44]. The tolerance of brain tissue to biomaterials is also seen from the use of a variety of substances, including in situ curing acrylics, to seal cranial defects.

2.6.1. Nerve Repair and Regeneration

Nerve tissue does have the capacity for repair, but it is limited in extent and efficiency. Repair of peripheral nerve damage can be assisted by nerve guides, small tubular structures placed around the severed ends of a nerve, which allow the nerve tissue to regenerate and reconstitute the nerve pathway. Ideally the materials used should

actively encourage this growth, although simple polymers probably achieve their goal passively [45].

2.6.2. Nerve Stimulators in Pain Relief

Chronic, severe pain is extremely difficult to treat. It is known that electrical impulses can be applied to nerves to block the impulses transmitted to the brain, that give rise to pain. The use of implantable pain relief stimulators is relatively new [46]. The same principles of construction described for pacemakers apply.

2.6.3. Treatment of Hearing Loss

Sound is conducted to the brain via the ear. The middle ear, involving the three small ossicular bones (stapes, incus, and malleus), conducts sound waves to the tympanic membrane. The disease of otosclerosis causes inflammation and eventual fusion of these bones such that their ability to vibrate is compromised. These bones may be excised and replaced by quite simple structures which can themselves vibrate within the ear to conduct sound [47]. If the inner ear is affected by disease, treatment is far more difficult; indeed restoration of hearing to the profoundly deaf is virtually impossible in the vast majority of cases. Nevertheless, some progress has been made in the conversion of sound waves into electrical signals and delivering these directly to the cochlea [48].

2.7. General Surgery

2.7.1. Wound Closure and Artificial Skin

Although there is a very strong trend toward minimally invasive surgery, in which incisions are very small, there is still a considerable need for materials and techniques for wound closure. Traditionally, tissues have been closed with sutures. These are monofilament or multifilament threads, used to sew tissues together. They may be permanent, as with polypropylene sutures, or intentionally degradable and resorbable, either natural (catgut) or synthetic (some polyesters). Materials such as silk are used where long-term retention is required but where degradation eventually takes place [49].

Clips, usually metallic, are often used instead of sutures, largely because of greater convenience. In addition, there is much interest in adhesives. The specifications for surgical adhesives are, however, very strict especially in relation to the wet environment and the need to avoid toxic effects associated with a setting adhesive. There is still much progress to be made [50].

Artificial skin is required in situations where there is considerable skin loss, usually resulting from burns. Skin grafting is preferred, but where there is substantial loss of tissue and insufficient donor tissue, a synthetic alternative is desirable. Due to the complexity of skin structure and function this is a difficult task but some progress has been made with the use of a degradable polymer scaffold, infiltrated with epithelial cells from the patient prior to application [51]. This type of structure allows skin to regenerate within the matrix, which then degrades.

2.7.2. Soft Tissue Reconstruction

Many surgical procedures used for correction or reconstruction are simple and only involve the placing of a piece of soft plastic material into the body to alter shape or contour. Since these operations are largely cosmetic, they tend to concern the face and female breast. In theory they should be the simplest of procedures and provide the best results. This is not always the case, however, and significant problems have arisen over the use of some of these materials, especially the silicones and polyurethane-coated silicones used for breast prostheses [52]. These problems have considerably reduced public demand for such prostheses.

2.7.3. Surgery of the Urinary Tract

The urinary tract, including bladder, urethra, ureter, and associated tissues, suffers a number of diseases, primarily involving incontinence and infection, and a very large number of patients are affected. However, treatment of many of these conditions remains problematic and the use of materials to assist in reconstruction or restoration of function is not well established. It

is extremely difficult to use biomaterials within the urinary system, because of the need to regulate urodynamics, to avoid infection, and to resist encrustation. There is a considerable need for prosthetic replacements for these tissues and also for devices to effectively control incontinence [53].

2.8. Artificial Organs

Clearly some organs can be transplanted, and although there remain very many logistic and ethical issues, this must be considered the preferable mode of replacement. In many circumstances, however, this is not possible and an artificial device has to be considered. A number of years ago, an implantable artificial heart was used in a few patients but this was not particularly successful. It is unlikely that any of the major organs can be replaced by a synthetic substitute, simply because of their complex and multiple functions. Progress is being made in a limited number of situations in which synthetic materials can be used in conjunction with cells or tissues, derived either from the patient or a donor, producing a hybrid artificial organ. This is undoubtedly an area of high promise for the future.

3. Criteria for Materials Selection

[54]

A wide variety of functions are required from surgical materials and no single material or group of materials can satisfy all the conditions. It is necessary, therefore, to consider, generically, the requirements of biomaterials and to identify the key elements of materials selection. These can be divided into properties that determine suitability of the device to function, those which determine practical aspects of manufacture and supply, and, most importantly, the features of biocompatibility.

3.1. Mechanical and Physical Properties

Although surgical materials are intended to replace, both in structure and function, the diseased or otherwise damaged tissues, the performance required of surgical materials is generally quite limited and largely restricted to space filling or mechanical performance, either structural or fluidic. There are some circumstances where physical properties are critical, and some situations where biological performance is required are emerging.

3.1.1. Mechanical Performance

Since the tissues to be replaced vary from the extremely hard enamel to the very soft lens, and from the rigid cortical bone to the very extensible arterial wall, a generalized mechanical performance for surgical materials cannot be specified. However, there are a number of important general points.

Strength is always important and it is crucial that stress systems within the material – tissue complex are established and performance specifications, with respect to both yield and fracture strength, determined. Finite element analyses are increasingly being used for this purpose, and the stresses generated within the human body are often surprisingly high. Because the body is a dynamic medium, both viscoelastic and fatigue behavior are crucial. Nevertheless, these mechanical considerations per se should not be serious limiting factors, since, with a few exceptions, there are sufficient materials available, off-the-shelf, to satisfy these needs.

With respect to *elasticity*, the most important aspect is the matching of the elastic constants and deformation behavior of devices to the requirements of the body. Ideally the rigidity, flexibility, compliance, or other characteristics of elastic deformation should enable the device to participate in the transfer of load and movement exactly as if it were the natural component. Failure to do so can have major consequences, both in relation to device function and to the overall effect on the body.

Because *mechanical performance* must be achieved within the chemically hostile environment of the tissue fluids, mechanical – environmental failure modes, such as stress corrosion cracking, corrosion fatigue, stress crazing, stress accelerated macromolecular degradation, and hydrolysis induced delamination of composites, may occur.

3.1.2. Electrical and Optical Properties

Electrical and optical properties are the two most significant physical properties. In some situations electrical conductivity is required, as in pacemaker or stimulator electrodes while in other cases, as in their coverings, insulation is necessary. Electrodes delivering signals to tissues are often made of noble metals, particularly platinum group alloys, because of the need for inertness under conditions in which charge has to be transferred across the interface, without the simultaneous transfer of toxic metal ions. Because of inherent difficulties here, there is much interest in other conducting materials such as conducting polymers. Insulation is not normally too much of a problem although the diffusion of body fluids, themselves electrolytes, into polymers has to be taken into account when considering long-term performance.

Optical properties are reasonably straightforward, there being a small range of optically clear amorphous polymers available. Poly(methyl methacrylate) is usually preferred.

3.2. Manufacturing and Sterilization

3.2.1. Constraints on Fabrication

The very significant advances in manufacturing technology and process engineering imply that there should not be too many problems over the fabrication of medical devices. The one major factor to consider is that no two patients are alike and that ideally many of the devices should be custom-made. This becomes difficult technically since usually prior to an operation the extent of tissue resection is not known and, therefore, the shape of the prosthesis cannot be predetermined. Nowadays, a variety of shapes and sizes have to be made available and hospitals are required to hold a large inventory so that the nearest fit can be used. In the future there will be more emphasis on the use of medical imaging techniques to determine the extent of the disease and predict the shape that will be required, and of CAD/CAM techniques for customized fabrication [55].

3.2.2. Market Analyses [56]

One of the main constraints to the supply of both materials and devices is the size and fragmented nature of the market. Whilst it is often thought that the medical device area, like the pharmaceutical industry, commands very high prices, and that markets are very large, in reality it is not quite so simple and many clinical conditions are not sufficiently common for manufacturers to take the risks associated with small production runs. Of even greater significance is the fact that raw materials suppliers usually do not benefit from the high unit prices of medical devices, but carry responsibility for quality. Few are, therefore, willing to supply small amounts of medical-grade materials to this industry.

3.2.3. Legal and Regulatory Constraints

There are quite severe regulatory constraints on the use of surgical materials in medical devices. In the United States, the industry is regulated by the Food and Drug Administration (FDA) while the European industry is now coming under the control of the European Community (EC) Medical Devices Legislation. There are stringent requirements concerning biological testing and conformity to certain requirements, especially quality control.

3.2.4. Sterilization Techniques [57]

The implantation of a medical device into the body carries with it an inherent risk of infection. Of course any operative procedure is associated with a finite risk but the use of materials placed within tissue considerably enhances that risk. It is extremely important, therefore, that the risks are minimized by appropriate sterilization of the devices. Care has to be taken since all of the available techniques can adversely affect some of the surgical materials. For example, two processes rely on heat to kill bacteria, either by dry heat or with superheated steam (autoclaving), the temperatures involved affecting many plastics. The main alternative is gamma irradiation, whereby the free radicals generated during this process can have significant and deleterious effects on many materials. The third method

involves chemical agents, principally ethylene oxide. Whilst effective, it is very difficult to remove all residues from the materials, potentially leading to toxic or irritant effects, and it is likely that this process will be prohibited in the future.

3.3. Biocompatibility [58]

The interactions between surgical materials and their host are the most influential factors determining their performance, especially in the long term. Biocompatibility can be defined as the "ability of a material to perform with an appropriate host response in a specific application" [1]. It must be emphasized that the range of materials in use today has been based on the assumption that biomaterials have to be inert and have no effect on the tissues. Whilst this may be appropriate under many circumstances, and while we never wish there to be adverse effects, it is now considered that the total lack of interaction implicit in that concept is not appropriate to the use of devices that are intended to become incorporated into the body, both anatomically and functionally, and that a distinctive and positive interaction is a far better goal. It is this concept that underpins the following discussion and which explains the developments in materials and especially their surfaces.

3.3.1. General Principles

Biocompatibility is concerned with all of the reactions and effects that take place between a material and the body [59]. It is possible to distinguish several groups or types of reaction in order to identify mechanisms, although all are potentially mutually interdependent. These categories of reaction are determined by the location of the effect.

Firstly, upon immediate exposure of the material to the tissue fluids, whether that be blood, serum, saliva, tears, or extracellular fluid, molecules are adsorbed from the fluid onto the surface. These are usually proteins, of which there are many different types in all body fluids and the characteristics of the adsorbed layer may significantly influence subsequent events. Of particular importance are the selectivity of the process, the changes to the conformational state of the proteins after adsorption, and the dynamics of adsorption and desorption as a multilayer is established. At this point, it is still difficult to identify which material characteristics control these phenomena.

Secondly, there are the effects of the tissues on the material, particularly in relation to corrosion and degradation mechanisms.

Thirdly, and usually of greatest significance, there are the effects of the material on the tissues adjacent to the device; these effects vary from one site to another and with time.

Fourthly, there is the possibility that the interfacial reactions and the systemic transport of reaction products away from the interface can give rise to remote effects, perhaps at the site of deposition or accumulation of these products.

3.3.2. Corrosion and Degradation

The environment of the human body is surprisingly hostile to synthetic materials. It is an isotonic saline solution with a variety of additional anions and cations and a range of biological macromolecules, free radicals, and cells, all of which have the potential to be biologically as well as chemically active. For metallic materials, the saline electrolyte ensures a highly corrosive environment [60], although it is known that the proteins present can also influence corrosion processes, especially if the metal concerned is able to bind to proteins [61].

Because of this aggressiveness, only the noble metals and those that are passive under physiological conditions can be used (see Section 4.1).

With ceramics, the behavior varies considerably [62]. Whereas oxide ceramics such as alumina and zirconia are extremely stable, some calcium-based ceramics are used which are degradable over a relatively short period. It is not quite clear whether this is a passive dissolution process or whether specific biological activity is implicated. It is instructive to note, however, that natural coral is used as a biomaterial; in the oceans, coral is stable over thousands of years but within the body, it may dissolve in months. Coral mainly consists of calcium carbonate.

Polymeric materials are the most difficult to judge from the degradability point of view. Polymers are normally degraded by physical

agencies such as heat, light, and ionizing radiation (which are absent in the body) so that they should be largely protected. However, some are susceptible to degradation by hydrolysis, and active molecular species within tissues, particularly free radicals and enzymes, can participate in degradation mechanisms [63], [64]. There is, therefore, a spectrum of behavior ranging from some homochain, nonhydrolysable, hydrophobic materials such as polytetrafluoroethylene, which appear to be nondegradable, to hydrolysable and rapidly degradable polymers, always heterochain and usually with carboxyl groups such as the aliphatic polyesters [65]. In the middle are a selection of polyamides, polyurethanes, polyesters, and many others which slowly degrade, perhaps over years, within the body. This degradation can affect device performance, for example, the eventual rupture of a polyester artificial artery [66], or affect the tissue response [67].

3.3.3. Soft Tissue Biocompatibility

It is still uncertain exactly what device characteristics determine the host response, but some general principles can be described. It is convenient to regard these on a temporal basis and to consider the response as a perturbation to wound healing, repair, and remodeling processes that occur in response to tissue damage [68]. If the surgical implantation of the material into an area of soft tissue is thought of as an injury, the effects of the material itself can be analysed as a superimposition onto the effects of the surgery. Inflammation results whenever tissue is damaged and has to be recognized as an inherently desirable response under these circumstances since it is designed to eliminate the source of the irritation (e.g., bacteria). Inflammation involves changes to the microvascular network and the infiltration of large numbers of cells into the area. If the source of the injury can be eliminated and any debris removed, both being brought about by the activity of these inflammatory cells, then a repair process is initiated, and the inflammatory cells slowly withdraw. In soft tissue, the repair process is normally controlled by fibroblast cells which synthesize collagen, a protein that constitutes scar tissue.

A biomaterial placed at the site of a surgical incision may interfere with either or both inflammatory or repair processes. With a totally chemically inert material, however, these opportunities are minimal and it is likely that inflammation and repair will proceed as normal, the only influence of the material being as a physical barrier to regeneration. Classically, the implantation of a sample of an inert biomaterial into soft tissue results in the formation of a thin layer of fibrous tissue, structurally analogous to scar tissue, that surrounds or encapsulates the implant [69].

If, however, the implant is not inert, which is usually the case, the reactions that take place will influence events. They may aggravate and extend the initial acute inflammatory response. More importantly, the continuation of interfacial reactions means that there remains a persistent stimulus to inflammation resulting in a chronic inflammatory response, merging with a delayed and continuous fibroblastic reaction. Their effects on the performance of the device and the health of the patient vary considerably, but if uncontrolled they are likely to lead to cell and tissue death. Moreover, these interfacial reactions usually initially involve the release of some component from the material into the tissue, either a corrosion or degradation product, or a leachable residue. The development of the inflammatory response, in which large numbers of cells are attracted to the area and activated to release large quantities of superoxide ions and oxidative enzymes, for example, rapidly increases the aggressiveness of the tissue, causing further and enhanced degradation of the material. The process is, therefore, autocatalytic [70].

The precise features of the host response are highly dependent on the individual circumstances. Material characteristics obviously influence events, particularly with respect to bulk and surface chemistry, rates of degradation, and the nature of the degradation products and physical features such as surface finish, size, and shape. Of equal importance are the patient variables, such as location of the device and age, sex, activity, general health, and pharmacological status.

3.3.4. Hard Tissue Biocompatibility

Concerning location, one type of tissue, bone, is extremely important, for the host response is a little different and the consequences of enor-

mous importance. The critical aspect is the ability to attain and maintain good apposition and preferably adhesion to bone, especially with joint replacements [71]. If the prosthesis is loose or becomes loose, then it will fail. If the generalized reaction of tissues to materials is, after inflammation and repair, a zone of soft collagenous tissue around the device, and if this were to take place around a material placed within bone, then it would be impossible to obtain the necessary attachment. It is essential for there to be a bone – material interface without an intervening layer of soft tissue. This inherent mechanism of fibrous encapsulation, therefore, has to be avoided or suppressed. This has been largely attempted through the use of materials of maximal inertness, such as titanium, alumina, poly(methyl methacrylate), or polyethylene. Of considerable current interest is the approach involving materials that are positively attractive to bone, the so-called bioactive materials, such as calcium hydroxyapatite [72], which permit preferential formation of bone rather than soft tissue at the interface.

The acquisition of a direct interface with bone is important, but so is its retention. Two factors may contrive to destroy such an interface. Firstly, as bone is a dynamic tissue it is dependent on the appropriate level of mechanical stress for it to maintain its normal structure and architecture. Without stress, bone will resorb. With the wrong type of stress, which results in relative movement between bone and prosthesis, resorption will occur. It is vital to ensure the optimal stress distribution in the bone – prosthesis system in order to maintain the bone; failure to do so, arising from poor materials selection, design, or technique, will result in the prosthesis becoming loose [73]. Secondly, if restimulation of the inflammatory response occurs, this could cause resorption through cellular activity. The main causes here are wear particles generated at the bearing interface in total joints [74]. Depending on their rate of release and the particle size, a significant chronic inflammation can arise, some of the cells involved also having the capability of destroying bone. This is probably the ultimate limitation to the performance of orthopedic prostheses.

3.3.5. Blood Compatibility

Just as the interactions between materials and bone control the performance of orthopedic prostheses, so the interactions between materials and blood control the performance of all intravascular and extracorporeal circulatory devices. These are extremely complex since they involve so many different components and mechanisms (see Section 2.4). The principle points are as follows [75].

The initial event when blood flows over a surface is that some plasma proteins are adsorbed onto that surface. The nature of the resulting *protein layer* varies from one material to another and with time. A thin layer of protein on a surface is not inherently undesirable in most circumstances; with the exception of intravascular sensors where the layer may disturb analyte transport. However, it is this protein layer which determines whether blood will clot on the surface, primarily through its effect on clotting cascade proteins and platelets [76]. The former process involves a sequence of enzyme-catalyzed reactions in which various circulating proteins are sequentially transformed into fragments, resulting in conversion of prothrombin into thrombin, which catalyzes the conversion of fibrinogen to fibrin and its subsequent polymerization. For reasons which are not entirely clear, some surfaces are far more likely to activate this sequence of events than others [77].

The second process is *platelet activation* [78]. These cells do not normally interact with the surface of blood vessels; they only do so when the vessel is damaged and then the interaction is very valuable in hemostasis. Unfortunately, the conditions which arise when a foreign surface is exposed to blood are similar to those of a damaged vessel surface, and the platelets react in the same way. A few platelets are attracted to the surface and activated. A series of chemicals are released from these activated platelets attracting vast numbers of other platelets which form an aggregate. The combination of polymerized fibrin and platelets constitutes a blood clot or thrombus, which is both damaging to device performance and compromises the health of the patient.

The selection of blood-contacting biomaterials is determined by the need to minimize this tendency. There are several possible approaches.

First, extreme inertness may yield relatively nonthrombogenic surfaces, such as PTFE, although this is no guarantee. The second approach is to use materials which inhibit or prevent these processes from taking place. This usually involves treating a surface with a substance which interferes with the clotting process (e.g., heparin [79]), or with the platelet adhesion process (e.g., prostacyclin [80]). Thirdly, the surfaces can be prepared such that they mimic the natural surface, as with the attachment of phospholipids to polymer surfaces [81], or are partially made of natural components such as surfaces seeded with endothelial cells [82].

Currenty, there is no universally blood-compatible material or surface but these approaches provide a promising future.

3.3.6. Systemic Biocompatibility [83]

If the products of an interfacial reaction can be transported away from the site, either by passive diffusion through the tissues or actively within lymphatic or vascular systems, there is the possibility that they could promote effects at some point distant from the site of application of the device. There is little evidence of this, but by the very nature of the phenomenon, evidence would be hard to obtain. The greatest fear here is that the products could accumulate in a sensitive area and cause serious effects where there was no obvious causal link to the device; the risk of biomaterial-induced cancer is at the forefront of these concerns [84].

3.3.7. Bioactivity

Inertness and lack of biological recognition is not always desirable, and a more active role for the material may be preferable. There are at least two possible mechanisms by which these so-called bioactive materials could operate, either through the inhibition of undesirable events or through the promotion of desirable events. Examples have already been given for bioactivity in the cases where bone adaptation has to be encouraged and where blood interactions have to be controlled. Further examples can be found in Chapter 4.

4. Materials

An encyclopedic review of biomaterials is given in [85]

4.1. Metals and Alloys

Metals are primarily employed in situations requiring their unique combination of mechanical properties, particularly fatigue strength, toughness, and ductility. Also when electrical conductivity is needed, metals are the materials of choice.

Because the mechanical specifications can be met by many alloy systems, considerations of biocompatibility, and especially corrosion resistance and toxicity, are decisive. It is possible to use the corrosion resistance of noble metals, and indeed, platinum group metals are used in electrical applications, gold alloys in dentistry, and silver in a variety of circumstances due to its antibacterial activity. However, these noble metals do not generally have sufficient strength, and so engineering alloys are preferred. These have to be of exceptional corrosion resistance and it is only passive alloys that can be considered. These are based on the passivity of chromium or titanium.

Steels. Stainless steels rely on chromium for their passivity and provide good corrosion resistance under many conditions. However, the physiological environment is such that their corrosion resistance is marginal, and only a small group of austenitic stainless steels are used, particularly AISI 316L [86–88]. These alloys possess a good combination of properties, but their popularity is declining in view of the enhanced performance of the alternatives discussed below.

Cobalt-Based Alloys. For over sixty years, alloys of the Stellite variety have been used in dentistry and surgery. These are alloys primarily of cobalt and chromium, in solid solution, with a variety of other elements to confer specific properties. Initially these alloys, with additions of molybdenum, nickel, and carbon gave a good balance of properties but the very restricted ductility limited the alloys to situations in which

they could be cast. Variations in alloying additions then permitted conventional forging operations to be used and now a wide variety of formulations and production technologies provide some extremely good properties [89]. The corrosion resistance is better than that of stainless steels and the excellent hardness makes these alloys the materials of choice for one of the bearing surfaces of many joint replacements. The main drawback lies with the putative toxicity of cobalt and chromium, although the significance is questionable.

Titanium and its Alloys. Titanium is the most corrosion-resistant non-noble metal in physiological fluids and is widely used in medical applications. It does not have sufficiently good mechanical properties for it to be used as the pure metal in most circumstances and so, for structural applications, alloys are employed. For many years, the $\alpha + \beta$ alloy Ti – 6 % Al – 4 % V has dominated the scene, but more recently, a few others have been introduced. The rationale for this has been a concern over adverse effects of aluminum and vanadium, but again these are speculations rather than objective decisions and there is no evidence of such adverse effects. Indeed, for many applications, the Ti – Al – V alloy possesses the best combination of mechanical, corrosion, and biological properties [90], [91]. Two points about the orthopedic uses of titanium should be noted. First, it has an elastic modulus, ca. one half that of either steel or cobalt alloys and is, therefore, more closely matched to bone with respect to rigidity. Secondly, its wear resistance is considerably inferior to these alternatives and its use as the bearing surface in joints is not recommended.

Platinum Group Metals. The inertness of platinum coupled with its excellent electrical properties suggests that it should be very appropriate for electrodes. Because implanted electrodes are subjected to forces within the body, mechanical performance is important, and certain alloys, such as Pt – Ir or Pt – Rh, may be superior [92].

Other Metals. There are few opportunities for other metals in the body. As noted above, silver is of interest because of its antibacterial activity [93]. One other possibility is the use of shape-memory alloys for intravascular strents, orthodontic wires, and orthopedic devices, where Nitinol, a nickel – titanium alloy is the most obvious candidate [94].

4.2. Ceramics and Glasses

Until a decade ago, the inherent brittleness of ceramics precluded their use in critical structural applications within the body. However, advances in ceramic processing technology have enabled significant improvements to fracture strengths and toughness to be obtained and now ceramics, and their amorphous counterparts the glasses and glass ceramics, are frequently used in these applications. There are three qualities of ceramics with underpin these uses, the inertness of certain oxide ceramics, their hardness, and within a small group, a degree of bioactivity.

Oxide Ceramics. Some of the simplest ceramics, those combining one metal with one nonmetal, provide the best properties. In particular, when the nonmetal is oxygen, a small series of extremely inert and hard ceramics may be distinguished. Of these, alumina (aluminum oxide), is the most widely used in surgery [95], [96]. Although there have been a few reports of property changes to alumina on implantation, this is not a significant problem, and the long-term biocompatibility is exceptionally good. The hardness is responsible for the selection of alumina as the bearing surface in many hip replacements. While the strength is reasonably high, its toughness is still only marginally acceptable, and there have been problems of brittle fracture in joints where alumina has been used for both femoral and acetabular components. There has, therefore, been a move toward tougher ceramics, particularly transformation-toughened, partially stabilized zirconia [97].

Calcium-Based Ceramics. Calcium compounds are not renowned for their mechanical properties nor for their hardness, but the chemical and structural similarity between some calcium compounds and the mineral phase of bones and teeth is attractive for certain medical purposes. This phase is essentially calcium hydroxyapatite, a hydrated calcium phosphate.

Calcium Phosphates. Synthetic calcium phosphates can be placed in contact with bone and stimulate bone regeneration [98]. These cannot be used in a structural capacity, but there are several forms which are of interest. Of the *chemical variations*, the two main forms are calcium hydroxyapatite and tricalcium phosphate. The former is essentially stable, while the latter is rapidly degradable. There are also *structural variations*. The materials can be used in particulate form, whereby a quantity may be placed within defects in bone, in a similar way to that employed for bone grafts. New bone forms around the particles, which effectively act as a scaffold. With the nondegradable calcium hydroxyapatite, the result is a composite structure of biomaterial and new bone; with the degradable tricalcium phosphate, new bone may continue to form as the particles degrade, resulting in total bone restoration. Bone defects in the facial skeleton and the spine may be treated in this way [99], [100].

Calcium Carbonates. Some forms of calcium carbonate are also able to demonstrate bioactivity with respect to bone. In particular, coral, a natural form of calcium carbonate, is very effective in stimulating bone ingrowth and regeneration; the coral degrades over months and is wholly replaced by bone [101].

An alternative use is as a coating on a tougher substrate. In particular, hydroxyapatite is frequently applied by plasma spraying to metallic prostheses to encourage adaptation of bone [102].

Glass Ceramics. While the oxide ceramics are completely stable and materials such as coral are totally degradable, some ceramics and glass ceramics display intermediate behavior with a degree of surface degradation. These are often referred to as bioglasses. These are complex structures, based on oxides of calcium, silicon, sodium, and phosphorus. Interaction with body fluids causes dissolution of calcium and phosphorus, which appears to permit formation of new bone at the surface, but the reaction is limited by the establishment of a silica-rich layer. These controlled surface-reaction glass ceramics are, therefore, bone-bonding [103] but their inherent brittleness and lack of toughness severely limits their applications.

4.3. Polymers

Because metals and ceramics are generally more rigid and hard than tissues, even bone, their use for tissue replacement or augmentation is counterintuitive. Polymers are preferred for these purposes because of their greater similarity to the natural structures. Furthermore, the wide variety of features, ranging from chemical reactivity to mechanical characteristics, that can be developed within the family of polymers, allows far better tailoring of these materials to meet precise specifications.

Polymers have the inherent capacity to display good biocompatibility since the macromolecules themselves are usually noncytotoxic. However, these macromolecules are very rarely used as the sole component of a material, most of which contains a variety of other substances, some desirable additives, other undesirable residues or contaminants, which tend to control biocompatibility. Plasticizers, antioxidants, residual monomers, and catalyst residues, for example, determine the properties. Polymers can be produced in a variety of forms ranging from elastomeric to rigid, cross-linked solids, from viscous resins to high molecular mass thermoplastics, by varying the molecular structure within the same chemical family.

Some of the major groups of biomedical polymers are listed below. These and other polymer families are described more fully elsewhere [104].

Polyethylene (→ Polyolefins) is, chemically, the simplest of all polymers. As a homochain polymer it is essentially stable and is suitable for long-term implantation in many situations. Moreover, it is amenable to a variety of production technologies, is relatively inexpensive, and has good mechanical properties, resulting in versatility with applications ranging from catheters to joint replacement. The latter is probably the most important and significant. After unsuccessful attempts with polymers that did not have adequate wear resistance, for example, PTFE, an ultrahigh molecular mass, high-density polyethylene became the material of choice for the acetabular component of hip replacements [105]. Although it does perform

adequately in many prostheses, wear does occur, and there is evidence now that if the wear rate exceeds a certain limit, the reaction to the wear debris will be destructive to the bone and will predispose it to failure [106].

Polypropylene (→ Polyolefins) is closely related to polyethylene, and has a similar versatility. Its medical uses range from high-strength, nondegradable sutures to the flexible integral hinges used in finger joint replacements.

Fluoropolymers (→ Fluoropolymers, Organic). The extreme inertness of some fluoropolymers is responsible for their widespread use, for example, polytetrafluoroethylene (PTFE). Unfortunately, the relatively poor mechanical properties limit their practical applications to softtissue reconstruction. There is no record of degradation of PTFE in the body, and tissue response to solid devices made of this material is minimal. Microporous expanded PTFE (Gore-Tex) is successfully used in small diameter artificial arteries [107], periodontal surgery, and other such applications.

Poly(methyl methacrylate) (PMMA) (→ Polymethacrylates) is a hard, brittle polymer that would appear to be unsuitable for most clinical applications. However, as well as being very resistant to degradation and well tolerated by tissues, it does possess two distinct advantages, which underpin many of its uses. Firstly, being an amorphous polymer it is transparent and is, therefore, the material of choice for most ophthalmological devices, especially intraocular lenses and hard contact lenses [108].

Secondly, it can be produced and fabricated under ambient conditions. It can be processed in dental laboratories which is crucial to its use in custom-made dental devices such as dentures and it may also be prepared intraoperatively, explaining its unique position as a bone cement which is made by mixing prepolymerized PMMA with monomeric methylmethacrylate, forming a dough [109]. This can be inserted into bone as a grout to hold a prosthesis firm within the bone. Although there are a number of deficiencies in this situation, including the release of heat associated with the polymerization process, the toxicity of the volatile methylmethacrylate, and the poor fracture toughness, no superior bone cement has been developed to date.

PMMA is the best known example of an acrylic polymer used in surgery but there are several derivatives of polyacids such as poly(acrylic acid) that are also employed. In particular, poly(hydroxyethyl methacrylate) is a crosslinked hydrogel that was developed specifically for use in contact lenses with the high oxygen permeability of its water-containing structure [110].

Polyesters (→ Polyesters) have a number of features which make them very attractive for medical use. The chemical structure of polyesters is intrinsically hydrolyzable so whether the polymer based on a specific polyester molecule is degraded or not will depend on the accessibility of the carboxyl groups to water.

Aliphatic polyesters based on naturally occurring substances such as lactic acid are *biodegradable*, with a combination of properties that render them highly suitable for transient functions in the body such as sutures or drug delivery systems. Since many polyesters are ideal for fabrication in the form of fibers, these materials, such as polyglycolic acid, polylactic acid, and polydioxanone, can be used as absorbable sutures. The precise molecular structure determines their rate of degradation and since the degradation products are natural and metabolizable, this process is generally well accepted [111].

The fiber-forming character of polyesters enables them to be prepared as *textiles*. Aromatic polyesters such as poly(ethylene terephthalate) (PET), are extensively used as the woven or nonwoven structures required in artificial arteries and ligament regeneration [112].

Polyurethanes (→ Polyurethanes) have had a very variable history of use in medical applications, partly arising from the wide range of materials that can be prepared with the urethane group. Bearing in mind the fact that the precursors of polyurethanes, such as toluene diisocyanate, have well-known toxicities, and the fact that the structures generally are susceptible to degradation, it is not surprising that these do display varying biocompatibility characteristics. Certain polyurethanes, however, have excellent mechanical properties, especially elas-

tomeric properties, which make them very suitable for areas of soft-tissue reconstruction. Polyurethanes also often have excellent blood compatibility. For these reasons, polyurethanes, especially polyether urethanes, have been used in critical blood-contacting applications, such as in the experimental implantable artificial heart [113]. They are also very suitable for catheters and similar devices.

However, their inherent susceptibility to degradation has resulted in a number of difficulties, for example, stress cracking of pacemaker lead encapsulation [114] and the degradation of porous coverings on breast implants [115]. The ideal characteristics, therefore, of elastic properties and blood compatibility needed for small diameter artificial arteries have yet to fully realized in devices of suitable long-term stability.

Silicones (→ Silicones, → Inorganic Polymers). Although the vast majority of synthetic polymers are based on the carbon – carbon bond, a small number are based on silicon. Linear chains of alternating silicon and oxygen atoms yield the family of siloxanes. Depending on the nature of the side groups attached to the silicon, the molecular size, and the degree of cross-linking, a range of substances may be produced, from oils to elastomers. It has long been thought that the inherent strength of the Si – O bond, coupled with the fact that this bond is not normally found in biological structures and should therefore be immune to enzyme-catalyzed degradation, should render it very stable. Thus, over several decades, substances such as polydimethylsiloxane have enjoyed widespread medical and dental use, ranging from elastomeric dental impressions to finger joint prostheses, catheters, intraocular oils, contact lenses, and a variety of soft tissue augmentation prostheses [116], [117].

However, the biocompatibility of silicones has been questioned [118], and claims have been made that silicones are capable of causing a wide variety of conditions, primarily autoimmune diseases. There is no clear evidence, but this matter is having a profound influence on the use of such materials in hitherto very common applications such as breast implants [119].

Natural Biopolymers. There is a trend toward the use of polymers derived from natural components where in some situations, the polymers themselves may be analogous to those occurring naturally.

Some tissues consist almost entirely of *collagen* and these may be used as biomaterials; one example is the tendon derived from the tails of certain animal species, for example, the kangaroo. Fortunately collagen is only a weakly antigenic protein and is unlikely to invoke an immune response when used in this way. It may be preferable, however, to prepare alternative forms of collagen, for example, by solubilization and reprecipitation. Some collagen products have quite widespread use in plastic surgery, although their stability is still questionable [120].

Hyaluronic acid is a polysaccharide that is widely distributed in tissues, especially in synovial fluid. It can be polymerized and modified in many ways resulting in a variety of degradable polymers with a number of potential applications [121].

4.4. Carbon

Carbon, as the major elemental constituent of organic molecules, would appear to be intrinsically safe to use within the body. On the basis of this putative inertness and biocompatibility, attempts have been made to use carbon, in several of its forms, as a biomaterial. Its major deficiency is that it is not a structural material and is difficult to fabricate into useful artifacts. As a bulk material it has very limited use; this can be seen by the attempts to use it in one of its few monolithic forms, as glassy carbon, in dental and other applications, the inevitable brittleness resulting in mechanical failures [122].

The interest in carbon is still intense but the forms in which it is used try to obviate these structural deficiencies. For example, it is used very successfully as a thin coating, in the form of pyrolytic graphite, deposited onto tough metallic substrates, for blood-contacting surfaces in heart valves [123]. An alternative carbon surface is the so-called diamond-like coating [124].

An alternative approach has involved the use of carbon fibers, well known for their intrinsic high strength and modulus of elasticity. Some applications have utilized tows or bundles of the fibers themselves, as with their use in artificial ligaments, but the majority are concerned with carbon compos-

ites, either carbon – carbon [125], carbon-fiber-reinforced thermosetting resins [126], or carbon-fiber-reinforced thermoplastics such as polyetheretherketone [127]. However, whilst chemically stable, the fibers themselves, by virtue of their tendency to fracture and produce small fragments of cellular dimensions, can be immensely irritating to tissues.

4.5. Composites

Attempts have been made to use composites in surgical materials, although up until now with limited success, since problems of biocompatibility have been aggravated through the increasing complexity of the structure.

There are several principles underlying the design of composites, usually mechanical in origin and commonly based upon maximizing fracture strength and optimizing elasticity. Strength is not usually markedly improved by the use of dispersed fibers or particles in a ductile plastic matrix, but it may be possible to achieve modest improvements. There is considerable scope for tailoring the elastic modulus by accurate control of components and their volume fractions. Although restorative dentistry has seen the development of a very successful range of particulate composites of quartz-like phases dispersed in dimethacrylate resin matrices [128], efforts with composite surgical materials have been far less successful. Apart from the carbon-fiber composites mentioned above, there have been unsuccessful attempts to strengthen bone cement by incorporating short fibers, and to improve the wear and creep resistance of polyethylene [129]. Especially disappointing has been the development of a carbon- or alumina-filled porous PTFE, intended as a prosthetic material to be replaced and incorporated into a variety of tissues [130].

4.6. Surface Modifications and Coatings

General Principles. Engineering materials used as structural components, with their surfaces modified to optimize biocompatibility is generally quite common. However, the varying requirements for biocompatibility and the need for long-term performance hinder their practical application as witnessed by the diverse and imaginative approaches already attempted.

Hard Coatings. Improving the hardness of a surface is thought to have a beneficial effect, not only when that surface is involved in a joint, but also when the release of any component through abrasion with tissue is possible. A variety of ionplating, ion-assisted deposition, ion-implantation, and other similar techniques may be employed [131].

Hydrophilic Coatings. Improved biocompatibility is thought to result if the surface is hydrophilic rather than hydrophobic. Such surfaces may be developed by grafting hydrogels onto polymer surfaces [132]. Better performance is likely when molecular species are grafted, in particular, when substances such as poly(ethylene glycol) are attached [133].

Pharmacological Coatings. In some circumstances it may be more appropriate to seek specific activity rather than passivity. One method would be to attach pharmaceuticals to surfaces, either by immobilizing them on the surface or incorporating them into the surface layer from which they may be released. This is most relevant in the case of blood-compatible surfaces, where anticoagulant, antiplatelet, fibrinolytic, or other agents may be attached to polymers (see Section 3.3.5).

Biomimetic Surfaces. The attachment of a layer of natural tissue-like material to engineering substrates may be difficult to imagine. One approach, however, involves biomimicry, the use of molecules that are analogous to components of tissue, which can behave as if it were that tissue. For example, phospholipids are the major molecular species of cell membranes and can be attached to a variety of substrates where, among other things, they substantially reduce the extent of protein adsorption. This is a particularly attractive property for contact lenses [134].

Cell seeding of biomaterial surfaces is of considerable interest; here it is often possible to render a surface attractive to a type of cell, for example, the endothelial cells that line blood vessels, and to incubate that material in a culture

of cells derived previously from the eventual recipient. If these cells are able to colonize the surface, then a pseudo-natural surface is generated. This technology is still under early stages of development but it is a very attractive solution to biocompatibility problems [135].

Hybrid Structures and Materials. Combinations of synthetic materials and natural components, either viable or nonviable, are now in experimental use. For example, the treatment of diabetes may be radically altered by the use of a primitive artificial pancreas, in which islets, the insulin-producing cells, are obtained from donors and encapsulated in membranes that allow the cells to respond to glucose levels and produce insulin, but protect them from the destructive proteins of the immune system [136].

5. References

1. D. F. Williams, J. Black, P. J. Doherty: in P. J. Doherty, R. L. Williams, D. F. Williams, A. J. C. Lee (eds.): "Second Consensus Conference on Definitions in Biomaterials," *Biomaterial-Tissue Interface,* Elsevier, Amsterdam 1992, p. 528.
2. D. F. Williams: *Implants in Surgery,* Saunders, London 1973.
3. D. F. Williams, *Physics in Med. Biol.* **25** (1980) 611 – 636.
4. J. Black: *Orthopaedic Biomaterials in Research and Practice,* Livingstone, New York 1988.
5. J. W. Boretos, M. Eden (eds.): *Contemporary Biomaterials,* Noyes Publications, Park Ridge, N.J. 1984.
6. M. Szycher (ed.): *High Performance Biomaterials,* Technomic Publications, Lancaster, Pa. 1991.
7. D. S. Hugerford, L. C. Jones, *Clin. Orthop. Rel. Res.* **235** (1988) 12 – 24.
8. J. A. Rand, C. D. Dorr (eds.): *Arthroplasty of the Knee,* Aspen, Rockville 1987.
9. P. Ducheyne, D. Kohn: "Materials for Bone and Joint Replacement," in D. F. Williams (ed.): *Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, chap. 2.
10. J. Charnley: *Acrylic Cement in Orthopaedic Surgery,* Williams & Wilkins, Baltimore 1970.
11. R. M. Pilliar, *Clin. Orthop. Rel. Res.* **176** (1983) 42 – 51.
12. P. Ducheyne, *J. Biomed. Mater. Res. Appl. Biomat.* **21 (A 2)** (1987) 219 – 236.
13. H. C. Amstutz, P. Campbell, N. Kossovsky, I. C. Clarke, *Clin. Orth. Rel. Res.* **276** (1992) 7 – 18.
14. R. M. Rose, A. S. Litsky in D. F. Williams (ed.): "Biomechanical Considerations in the Loosening of Hip Replacement Prostheses,"*Current Perspectives on Implantable Devices,* **vol. 1,** JAI Press, London 1989, pp. 1 – 45.
15. A. E. Goodship, P. Cooke, *CRC Crit. Rev. Biocompat.* **2** (1986) 303 – 334.
16. H. Uhthoff: *Current Concepts of the Internal Fixation of Fractures,* Springer Verlag, Berlin 1980.
17. G. Heimke, R. J. Kolbe, R. A. Latour, *Mater. in Med.* **3** (1992) 204 – 211.
18. O. Bostman et al., *Clin. Orthop. Rel. Res.* **238** (1989) 195 – 203.
19. D. F. Williams in R. W. Cahn (ed.): "Dental Implants,"*Encyclopaedia of Materials Science and Engineering,* **vol. 1, Suppl.,** Pergamon Press, Oxford 1988, pp. 115 – 121.
20. T. Albrektsson et al., *J. Prosthet. Dent.* **60** (1988) 75 –84.
21. A. Kirsch, K. Ackermann, *Dtsch. Zahnaerztl. Z.* **38** (1983) 106 – 112.
22. D. F. Williams in D. F. Williams (ed.): "Materials for Oral and Maxillofacial Surgery," *Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 259 – 283.
23. J. Breme, E. Steinhauser, G. Paulus, *Biomaterials* **9** (1988) 310 – 313.
24. B. F. Conroy, J. E. Bowerman, J. M. Harrison: "Restoration of the Mandible by Implant Prostheses," in D. F. Williams (ed.): *Biocompatibility in Clinical Practice,* vol. II, CRC Press, Boca Raton 1982, pp. 169 – 182.
25. E. Bodnar, R. Frater (eds.): *Replacement Cardiac Valves,* Pergamon Press, New York 1991.
26. D. P. Giddens, A. P. Yoganathan, F. Schoen, *Cardiovasc. Path.* **2** (1993) 167 – 177.
27. W. R. Dimitri, B. T. Williams, *J. Cardiovasc. Surg.* **31** (1990) 125 – 134.
28. Y. A. Goffin, M. M. Black, P. Lawford: "The Stability and Performance of Bioprosthetic Heart Valves," in D. F. Williams (ed.): *Current Perspectives on Implantable Devices,* vol. 2, JAI Press, London 1990, pp. 65 – 120.
29. R. J. Levy, F. J. Schoen, H. C. Anderson, *Biomaterials* **12** (1991) 707 – 714.

30. P. Libby, F. Schoen, *Cardiovasc. Path.* **2** (1993) 43–52.
31. T. Ischinger (ed.): *Practice of Coronary Angioplasty,* Springer Verlag, New York 1986.
32. R. P. Leather, A. M. Karmody, D. M. Shah, J. D. Corson, *Acta Chir. Scand. Suppl.* **529** (1985) 47–56.
33. J. C. Stanley (ed.): *Biologic and Synthetic Vascular Prostheses,* Grune and Stratton, New York 1982.
34. M. A. Golden, S. R. Hanson, T. R. Krikman, P. A. Schneider, A. W. Clowes, *J. Vasc. Surg.* **11** (1990) 838–845.
35. R. Guidoin, M. King, D. Marceau, *J. Biomed. Mater. Res.* **21** (1987) 65–87.
36. R. VanHolder, *Clin. Mater.* **10** (1992) 87–132.
37. S. S. Barold: *Modern Cardiac Pacing,* Futura Publ. Co. Mount Kisco, New York 1985.
38. P. S. Binder, E. Y. Savala, J. K. Deg, *Cornea* **4** (1983) 119–125.
39. D. F. Williams in D. F. Williams (ed.): "Materials for Ophthalmology,"*Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 416–427.
40. R. Caramazza, P. Versura (eds.): *Biomaterials in Ophthalmology,* Studio ER Congressi, Bologna 1990.
41. R. W. T. Bowers, B. J. Tighe, *Biomaterials* **8** (1987) 83–88.
42. O. Wichterle, D. Lim, *Nature* **185** (1960) 117–119.
43. B. W. McCuen, E. de Juan, M. B. Landers, R. Machemer, *Retina* **5** (1985) 189–205.
44. A. D. Hockley, P. Collins, J. R. Anderson: "Ventricular Catheters," in D. F. Wiliams (ed.): *Biocompatibility in Clinical Practice,* vol. II, CRC Press, Boca Raton 1982, pp. 151–168.
45. I. V. Yannas et al., *Trans. Soc. Biomat.* **10** (1987) 6.
46. D. Fiume, M. Palombi, V. Sciassa, M. Tamorri, *PACE* **12** (1989) 698–710.
47. J. J. Grote: *Biomaterials on Otology,* Martinus Nijhoff, Boston 1984.
48. R. S. Tyler, B. C. J. Moore, F. K. Kuk, *J. Speech and Hearing Res.* **32** (1989) 887.
49. C. C. Chu, *CRC Crit. Rev. Biocompat.* **1** (1985) 261–322.
50. B. E. Causton in D. F. Williams (ed.): "Medical and Dental Adhesives,"*Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 285–302.
51. I. V. Yannas in D. F. Williams (ed.): "Biologically Active Analogs of the Extracellular Matrix,"*Medical and Dental Materials,* chap. 12, VCH Verlagsgesellschaft, Weinheim 1991.
52. W. Kaiser, G. Biesenbach, U. Stuby, P. Grafinger, J. Zazgornik, *Ann. Rheum. Dis.* **49** (1990) 937–938.
53. P. Aebischer, M. Goddard, P. M. Galletti, M. Lysaght in D. F. Williams (ed.): "Biomaterials and Artificial Organs," *Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 133–178.
54. D. F. Williams in D. F. Williams (ed.): "Biofunctionality and Biocompatibility,"*Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1992, pp. 1–30.
55. N. Jedynakiewicz, N. Martin: *CAD-CAM in Restorative Dentistry,* Liverpool University Press, Liverpool 1993.
56. M. Szycher: "Markets for Medical Plastics," in M. Szycher (ed.): *High Performance Biomaterials,* Technomic, Lancaster, Pa. 1991, pp. 3–30.
57. I. Matthews, *Clin. Mater.* 1994, in press.
58. D. F. Williams, *Interdisc. Science Rev.* **15** (1990) 20–33.
59. D. F. Williams, *J. Biomed. Eng.* **11** (1989) 185–192.
60. D. F. Williams: "Electrochemical Aspects of Corrosion in the Physiological Environment," in *Fundamental Aspects of Biocompatibility,*" vol. 1, CRC Press, Boca Raton 1981, pp. 11–42.
61. D. F. Williams, *CRC Crit. Rev. Biocompat.* **1** (1985) 1–30.
62. S. F. Hulbert et al., *J. Biomed. Mater. Res.* **4** (1970) 433–456.
63. J. P. Santerre, R. S. Labow, G. A. Adams, *J. Biomed. Mater. Res.* **27** (1993) 97–109.
64. D. F. Williams, S. P. Zhong, *Adv. Mater.* **3** (1991) 623–626.
65. R. Smith, D. F. Williams, C. Oliver, *J. Biomed. Mater. Res.* **21** (1987) 1149–1166.
66. E. Vinard et al., *J. Biomed. Mater. Res.* **25** (1991) 499–514.
67. M. Mohanty, J. A. Hunt, P. J. Doherty, D. Annis, D. F. Williams, *Biomaterials* **13** (1992) 651–656.
68. D. F. Williams, *J. Mat. Sci.* **22** (1987) 3421–3445.
69. M. Spector, C. Cease, X. Tong-Li, *CRC Crit. Rev. Biocompat.* **5** (1989) 269–295.
70. D. F. Williams, *J. Mat. Sci.* **17** (1982) 1233–1246.
71. T. Albrektsson, *CRC Crit. Rev. Biocompat.* **1** (1984) 53–84.

72. L. L. Hench, E. C. Ethridge: *Biomaterials; An Interfacial Approach,* Academic Press, New York 1982.
73. L. Linder, L. Lindberg, A. Carlsson, *Clin. Orth. Rel. Res.* **175** (1983) 93 – 104.
74. T. P. Gross, D. W. Lennox, *J. Bone Jt. Surg.* **74 A** (1992) 1096 – 1101.
75. R. W. Coleman, *Cardiovasc. Path.* **2** (1993) suppl. 23 – 31.
76. T. A. Horbett, *Cardiovasc. Path.* **2** (1993) suppl. 137 – 148.
77. G. P. Clagett, "Artificial Devices in Clinical Practice," in R. W. Coleman, J. Hirsh, V. J. Marder, E. W. Salzman (eds.): *Hemostasis and Thrombosis,* Lippincott, Philadelphia 1987, pp. 1348 – 1366.
78. J. M. Anderson, K. Kottke-Marchant, *CRC Crit. Rev. Biocompat.* **1** (1985) 111 – 204.
79. Y. Ito, *J. Biomater. Appl.* **2** (1987) 235 – 265.
80. C. H. Bamford, I. P. Middleton, K. G. Al-Lamee, *J. Biomat. Sci.* **2** (1991) 37 – 52.
81. D. Chapman, S. Charles, *Chem. in Brit.* **28** (1992) 253 – 256.
82. S. P. Schmidt, W. V. Sharp, M. M. Evancho, S. O. Meerbaum: "Endothelial Cell Seeding of Prosthetic Vascular Grafts," in M. Szycher (ed.): *High Performance Biomaterials,* Technomic, Lancaster, Pa. 1991, pp. 483 – 496.
83. D. F. Williams, *Systemic Aspects of Biocompatibility,* **vols. I and II,** CRC Press, Boca Raton 1983.
84. R. B. Pedley, G. Meachim, D. F. Williams: "Tumour Induction by Implant Materials," in D. F. Williams (ed.): *Fundamental Aspects of Biocompatibility,* vol. II, CRC Press, Boca Raton 1981, pp. 175 – 204.
85. D. F. Williams (ed.): *The Concise Encyclopaedia of Medical and Dental Materials,* Pergamon, Oxford 1991.
86. E. J. Sutow, S. R. Pollack: "Stainless Steel," in D. F. Williams (ed.): *Biocompatibility of Implant Materials,* vol. I, CRC Press, Boca Raton 1981, pp. 45 – 98.
87. A. Cigada, G. Rondelli, B. Vicentini, M. Giacomazzi, A. Roos, *J. Biomed. Mater. Res.* **23** (1989) 1087 – 1095.
88. R. M. Pilliar, G. C. Weatherly, *CRC Crit. Rev. Biocompat.* **1** (1986) 371 – 403.
89. ASTM,"Standard Specification for Cast Co – Cr Mo alloy for Surgical Implant Applications," ASTM, Philadelphia 1984, F 75 – 82.
90. H. A. Luckey, F. Kubli, "Titanium Alloys in Surgical Implants," ASTM, Philadelphia 1983, spec. tech. publ. 796.
91. D. F. Williams: "Titanium," in D. F. Williams (ed.): *Biocompatibility of Clinical Implant Materials,* vol. I, CRC Press, Boca Raton 1982.
92. K. R. Cote, R. C. Gill, *Ann. Biomed. Eng.* **15** (1987) 419 – 426.
93. R. L. Williams, P. J. Doherty, D. G. Vince, G. J. Grashoff, D. F. Williams, *CRC Crit. Rev. Biocompat.* **5** (1989) 221 – 244.
94. L. S. Castleman, S. M. Motzkin, F. P. Alicandri, V. C. Bonawit, A. A. Johnson, *J. Biomed. Mater. Res.* **10** (1976) 695 – 710.
95. J. M. Dorlot, P. Christel, A. Meunier, *J. Biomed. Mater. Res. Appl. Biomat.* **23** (1989) 299 – 310.
96. H. Mittelmeier, *The Hip* **12** (1984) 146 – 160.
97. P. Christel, A. Meumier, M. Heller, *J. Biomed. Mater. Res.* **23** (1989) 45 – 62.
98. R. LeGeros, *Clin. Mater.* **14** (1993) 65 – 88.
99. N. M. Meenen, J. F. Osborn, M. Dallek, K. Donath, *Mater. in Med.* **3** (1992) 345 – 351.
100. H. Alexander, J. R. Parsons, J. L. Ricci, P. K. Bajpai, A. B. Weiss, *CRC Crit. Rev. Biocompat.* **4** (1988) 43 – 77.
101. G. Guillemin, J. L. Patat, J. Fournie, M. Chetail, *J. Biomed. Mater. Res.* **21** (1987) 557 – 567.
102. B. Koch, J. G. C. Wolke, K. de Groot, *J. Biomed. Mater. Res.* **24** (1990) 655 – 668.
103. L. L. Hench, J. Wilson, *Science* **226** (1984) 630 –636.
104. D. F. Williams (ed.): *Biocompatibility of Clinical Implant Materials,* CRC Press, Boca Raton 1982.
105. P. Eyerer, M. Kurth, H. A. McKellup, T. Mittlmeier, *J. Biomed. Mater. Res.* **21** (1987) 275 – 291.
106. R. M. Rose, E. L. Radin, *Clin. Orth. Rel. Res.* **170** (1982) 107 – 115.
107. R. B. Rutherford: *Vascular Surgery,* Saunders, Philadelphia 1989.
108. R. Larsson, G. Selen, H. Bjorklund, P. Fagerholm, *Biomaterials* **10** (1989) 511 – 516.
109. R. D. Mulroy, W. H. Harris, *J. Bone Jt. Surg.* **72 B** (1990) 757 – 760.
110. B. J. Tighe: "Hydrogels as Biomedical Materials," in R. W. Cahn (ed.): *Encyclopaedia of Materials Science and Engineering,* suppl. I, Pergamon Press, Oxford 1988, pp. 216 – 221.
111. M. Vert, S. M. Li, G. Spenlehauer, P. Guerin, *Mater. in Med.* **3** (1992) 432 – 446.
112. Y. Marois et al., *Biomaterials* **14** (1993) 255 – 262.
113. R. K. Jarvik, *Sci. Amer.* **244** (1981) 74 – 80.
114. K. Stokes, *Cardiovasc. Path.* **2** (1993) suppl. 111 –119.

115. H. G. Bruck, *Aesth. Plast. Surg* **14** (1990) 85 – 86.
116. J. B. Crawford, *Am. J. Ophthalmol.* **101** (1986) 680 – 683.
117. E. E. Frisch, "Polysiloxanes as Biomedical Materials," in R. W. Cahn (ed.): *Encyclopaedia of Materials Science and Engineering,* suppl. I, Pergamon Press, Oxford 1988, pp. 418 – 427.
118. N. Kossovsky, J. P. Heggers, M. C. Robson, *CRC Crit. Rev. Biocompat.* **3** (1987) 53 – 85.
119. E. E. Sahn, P. D. Garen, R. M. Silver, J. C. Maize, *Arch. Dermatol* **126** (1990) 1198 – 1202.
120. E. Sableman, "Biology, Biotechnology and Biocompatibility of Collagen," in D. F. Williams (ed.): *Biocompatibility of Tissue Analogs,* vol. I, CRC Press, Boca Raton 1985, pp. 27 – 66.
121. A. Rastrelli, M. Beccaro, F. Biviano, G. Calderini, A. Pastorello: "Hyaluronic Acid Esters, A New Class of Semisynthetic Biopolymers," in G. Heimke, U. Soltesz, A. J. C. Lee (eds.): *Clinical Implant Materials,* Elsevier, Amsterdam 1990, pp. 199 – 205.
122. A. D. Haubold, H. S. Shim, J. C. Bokros, "Carbon in Medical Devices," in D. F. Williams (ed.): *Biocompatibility of Clinical Implant Materials,* vol. II, CRC Press, Boca Raton 1982, pp. 3 – 42.
123. A. D. Haubold, R. A. Yapp, J. C. Bokros: "Carbons for Biomedical Applications," in M. B. Bever (ed.): *Encyclopaedia of Materials Science and Engineering,* vol. I, Pergamon Press, Oxford, pp. 514 – 520.
124. D. F. Williams, *Med. Dev. Technol.* 1993, in press.
125. P. S. Christel, *CRC Crit. Rev. Biocompat.* **2** (1986) 189 – 218.
126. L. Tayton, C. Johnson-Nurse, B. McKibbin, *J. Bone Jt. Surg* **64 B** (1982) 105 – 111.
127. D. F. Williams, A. McNamara, R. M. Turner, *J. Mat. Sci. Lett.* **6** (1987) 188 – 190.
128. D. C. Watts: "Dental Restorative Materials," in D. F. Williams (ed.): *Medical and Dental Materials,* VCH Verlagsgesellschaft, Weinheim 1991, chap. 6.
129. H. E. Groth, J. M. Shilling, *J. Orthop. Res.* **1** (1983) 129 – 135.
130. B. L. Florine, D. J. Gatto, M. L. Wade, D. E. Waite, *J. Oral Maxillofac. Surg.* **46** (1988) 183 – 188.
131. H. A. McKellop, T. V. Rostlund, *J. Biomed. Mater. Res.* **24** (1990) 1413 – 1426.
132. B. D. Ratner: "Biomedical Applications of Hydrogels," in D. F. Williams (ed.): *Biocompatibility in Clinical Practice,* vol. II, CRC Press, Boca Raton 1982, pp. 146 – 175.
133. E. W. Merrill, E. W. Salzman, *ASAIOJ* **6** (1983) 60 –64.
134. A. A. Durrani, J. A. Hayward, D. Chapman, *Biomaterials* **7** (1986) 121 – 125.
135. J. A. Hubbell, S. P. Massia, N. P. Desai, P. D. Drumheller, *Biotechnology* **9** (1991) 568 – 572.
136. M. F. A. Goosen, *CRC Crit. Rev. Biocompat.* **3** (1987) 1 – 24.

Suspension Polymerization → Polymerization Processes

Sweeteners

Carbohydrates is a separate keyword.

GERT-WOLFHARD VON RYMON LIPINSKI, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Introduction . . . 2
2. Sensory Properties . . . 3
2.1. Structural Requirements for Sweetness . . . 3
2.2. Sweetness Intensity . . . 4
2.3. Taste Characteristics . . . 5
2.4. Synergism . . . 5
3. Uses . . . 6
3.1. Foods and Beverages . . . 6
3.2. Table-Top Sweeteners . . . 7
3.3. Cosmetics . . . 7
3.4. Pharmaceuticals . . . 7
3.5. Feed . . . 7
3.6. Others . . . 7
4. General Toxicology and Physiology 7
4.1. Toxicology . . . 7
4.2. Metabolism . . . 7
4.3. Suitability for Diabetics . . . 8
4.4. Dental Effects . . . 8
5. Food Legislation . . . 8
5.1. International Regulations and Assessments . . . 8
5.2. National Food Legislation . . . 8
6. Substances Commonly Used as Sweeteners . . . 8
6.1. Acesulfame K . . . 8
6.1.1. General Information . . . 8
6.1.2. Physical and Chemical Properties . . . 8
6.1.3. Production . . . 9
6.1.4. Specifications and Analysis . . . 9
6.1.5. Toxicology and Legislation . . . 9
6.1.6. Uses . . . 10
6.2. Aspartame . . . 10
6.2.1. General Information . . . 10
6.2.2. Physical and Chemical Properties . . . 10
6.2.3. Production . . . 10
6.2.4. Specifications and Analysis . . . 11
6.2.5. Toxicology and Legislation . . . 11
6.2.6. Uses . . . 11
6.3. Cyclamate . . . 12
6.3.1. General Information . . . 12
6.3.2. Physical and Chemical Properties . . . 12
6.3.3. Production . . . 12
6.3.4. Specifications and Analysis . . . 12
6.3.5. Toxicology and Legislation . . . 13
6.3.6. Uses . . . 13
6.4. Saccharin . . . 13
6.4.1. General Information . . . 13
6.4.2. Physical and Chemical Properties . . . 13
6.4.3. Production . . . 13
6.4.4. Specifications and Analysis . . . 14
6.4.5. Toxicology and Legislation . . . 14
6.4.6. Uses . . . 14
7. Sweeteners of Lesser Importance . 15
7.1. Glycyrrhizin . . . 15
7.1.1. General Information and Properties . 15
7.1.2. Production . . . 15
7.1.3. Toxicology and Legislation . . . 15
7.1.4. Uses . . . 15
7.2. Neohesperidin Dihydrochalcone . . 16
7.2.1. General Information and Properties . 16
7.2.2. Production . . . 16
7.2.3. Toxicology and Legislation . . . 16
7.2.4. Uses . . . 16
7.3. Stevioside and Rebaudioside . . . 17
7.3.1. General Information and Properties . 17
7.3.2. Production . . . 17
7.3.3. Toxicology and Legislation . . . 18
7.3.4. Uses . . . 18
7.4. Sucralose . . . 18
7.4.1. General Information and Properties . 18
7.4.2. Production . . . 19
7.4.3. Toxicology and Legislation . . . 19
7.4.4. Uses . . . 19
7.5. Thaumatin . . . 19
7.5.1. General Information and Properties . 19
7.5.2. Production . . . 19
7.5.3. Toxicology and Legislation . . . 19
7.5.4. Uses . . . 20
8. New Developments . . . 20
8.1. Alitame . . . 20
8.2. Others . . . 20
9. Substances Formerly Used as Sweeteners . . . 20
9.1. Dulcin . . . 20
9.2. Others . . . 21
10. References . . . 21

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_023

1. Introduction

The term "intense sweeteners" is commonly used for substances that are much sweeter than sucrose, and which are either not metabolized in the human body or do not significantly contribute to the energy content of foods and beverages. Intense sweeteners may be produced synthetically or isolated from plants. Sucrose, related carbohydrates, and sugar alcohols should be clearly distinguished from intense sweeteners.

Plants containing sweet compounds have been used for centuries to sweeten foods and beverages, particularly in tropical countries, e.g., the thaumatin-containing plant, *Thaumatococcus daniellii* Benth [1].

The history of intense sweeteners began with the discovery of saccharin by C. Fahlberg and I. Remsen [2] in 1878. Fahlberg recognized the scientific and commercial importance of saccharin and took steps to market it; production commenced in 1884. Since then, many other sweet compounds have been discovered, mostly by chance. Only a very few of these compounds have been used as sweeteners in foods and beverages. Several substances have been withdrawn, owing to the emergence of toxicological problems or unsatisfactory properties.

Saccharin enjoyed great commercial success in periods of short sugar supply, e.g., during World Wars I and II. As a consequence, saccharin was considered a cheap and inferior sugar substitute. Taxes were imposed on saccharin to protect the sugar industry. Diabetics were the main consumers.

In the last three decades, the general perception of intense sweeteners has changed. The incidence of obesity, and increasing health-consciousness have paved the way for a more favorable attitude. The increasing importance of industrial food production, the need for systematic recipe development, and the resulting possibilities for large-scale production of products containing intense sweeteners has led to widespread availability of these products. Their main applications are no longer products especially intended for consumption by diabetics or people suffering from other metabolic disorders. Intense sweeteners are now rather used in products aimed at consumers interested in body weight management and a calorie-controlled diet.

The increasing demands on safety of food additives in general, and sweeteners in particular, resulted in a reevaluation and reassessment of the intense sweeteners commonly used until the 1960s. The results provoked an international discussion of their safety, which, in turn, stimulated an intensive search for alternative compounds with improved properties. Two of the sweeteners developed, acesulfame K and aspartame, have found widespread international approval, and others, such as thaumatin and neohesperidin dihydrochalcone are approved and used in some countries. More products are under evaluation (e.g., alitame) or about to enter the market (e.g., sucralose). There continues to be promising market opportunities for sweeteners with good taste, stability, and handling characteristics, but above all, a clean safety record.

2. Sensory Properties

2.1. Structural Requirements for Sweetness

The generally accepted theory for the phenomenon of sweetness was developed by Shallenberger and Acree [3]. According to this theory, a molecular system of a proton donor and proton acceptor is necessary, with a distance of ca. 0.3 nm between the groups. It is assumed that a component has a sweet taste whenever this so-called AH – B-system can interact with a complementary system located in the membrane of taste receptor cells.

Changes in the distance between groups, as well as changes in electronic structure influence the occurrence of the sweet taste, and may change the general taste perception, sometimes eliminating sweetness totally, or changing it to bitterness.

The AH – B-system (Fig. 1) is a necessary but not sufficient structural requirement for sweetness, and is therefore not a reliable indicator for its occurrence or absence. Accordingly, the molecular theory was extended to include a third binding site that interacts with the sweet receptor by hydrophobic contact; this model allows better correlation between molecular structure and sweetness [4].

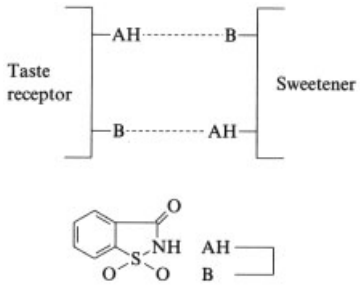

Figure 1. Interaction of sweeteners and taste receptors [3] through the AH – B-system of saccharin

On the basis of intensive research on very sweet molecules, the number of potential interaction sites of sweetness receptors with sweetening agents has been extended to eight, two of which must be in contact with sweet molecules to elicit sweetness, and four have to be in contact for extremely intense sweetness [5].

More systematic approaches have sought to reveal the molecular structure of sweetness receptor molecules. These included structure – activity investigations in homologous series of sweet substances [6,7]. Investigations on nitroanilines, sulfamates, oximes, isocumarins, dipeptides, saccharins, acesulfames, chlorosugars, tryptophans, and ureas showed that, for certain substituents, an increase in chain length results in increased sweetness, but a sharp decrease to complete absence of sweetness was observed above the optimum. Similar observations were made for amino acid amides [8].

Superposition of different sweet compounds was used to identify structural requirements for the sweetness receptor molecules [9]. Assuming a peptide-based structure, interaction with the L-aspartyl-L-phenylalanine methyl ester with a helical peptide could be demonstrated, whereas the nonsweet D – D isomer was unable to form hydrogen bonds with this peptide [10].

It is still uncertain whether there is just one sweetness receptor with different binding sites, or several types of sweetness receptor with different response characteristics. It seems probable, however, that specific binding characteristics for different sweet tasting structures have to be assumed. A sweetness receptor model with binding sites for these structures has been suggested [11].

2.2. Sweetness Intensity

The sweetness intensity of a substance can be expressed in several ways. The sweetness intensity factor, or sweetness power, is most frequently used. It is the factor by which an equisweet sucrose concentration is more concentrated than a solution of the sweetener. Commonly used values are often based on comparison with 0.1 mol/L sucrose solutions, or solutions of similar concentration. Sweetness intensity factors compared with sucrose are given below:

Sweetener	Sweetness intensity
sucrose	1
sodium cyclamate	35
glycyrrhizin	50
stevioside	160
acesulfame K	200
aspartame	200
rebaudioside A	250
neohesperidin dihydrochalcone	330
sodium saccharin	450
saccharin	550
alitame	2500
thaumatin	3500

Sweetness intensities generally depend on the sweetener concentration and the sucrose concentration used for comparison. Therefore, values other than those given above may apply at lower or higher concentrations. A demonstration of this concentration dependence is given in Figure 2 (see next page) for some commonly used sweeteners. Estimates of use levels in foods and beverages are often possible on the basis of comparison tables of sweetness intensities [12].

Much higher sweetness intensity factors are observed at the taste threshold [12]. These values are sometimes reported, but are of no practical significance.

Apparently, maximal sweetness intensities exist for all sweet substances, above which an increase in concentration does not induce a higher sweetness perception. This property is well illustrated by graphs of equisweet solutions of sweeteners and sucrose (Fig. 3, see next page).

Determination of sweetness intensities requires sensory tests; panelists have to be used according to their sensory abilities, and to be adapted to the test procedures. Owing to individual variations in sweetness perception, results

have to be evaluated on a statistical basis [13, 14].

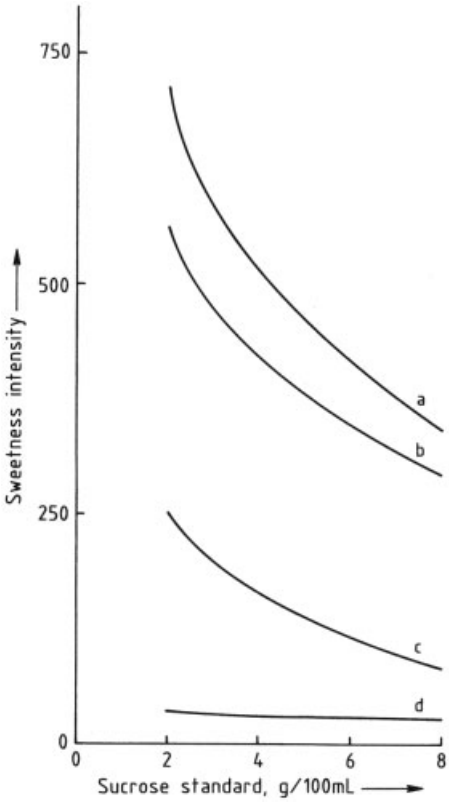

Figure 2. Sweetness intensity of saccharin (a), sodium saccharin (b), acesulfame K (c), and sodium cyclamate (d) compared with sucrose

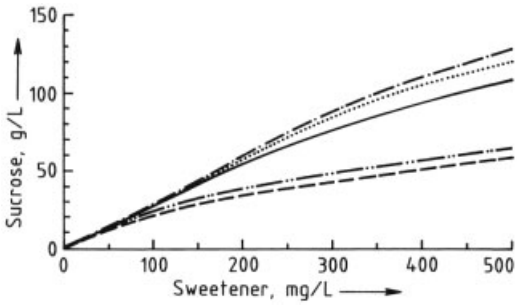

Figure 3. Equisweet concentrations of acesulfame (– – –), aspartame (– · · –), sweetener blends and sucrose
Blend ratio: (–) 2:1; (· · ·) 1:2; (– · –) 1:1

2.3. Taste Characteristics

The taste quality of sweeteners depends on several characteristics in addition to sweetness. Time – intensity profiles of sweetness perception are important for the assessment of taste quality as well as the occurrence of side-tastes and aftertastes.

Temporal sweetness perception characteristics can be studied with modern sensory techniques [15–17]. As sucrose is not only the standard for sweetness intensity assessments, but also the ideal to be matched in other sensory characteristics, the time – intensity profile of sucrose is often considered the target. Lasting sweetness or delayed sweetness onset may impair the applicability of sweeteners. Time – intensity perception of sweetness is influenced by several substrate factors, e.g., viscosity, temperature, and the presence of taste-modifying compounds. Blending of sweeteners with different time – intensity profiles often results in a mixed profile, closer to that of sucrose.

The bitter aftertaste of saccharin is well known. As sweetness and bitterness are structurally closely related [9], bitter aftertastes are observed not only in some intense sweeteners, but also in a number of sweet carbohydrates. Individual susceptibilities for bitterness perception and taste thresholds vary. Off-taste thresholds as reported in water [18], however, give general information on taste characteristics of sweeteners rather than information on limitations of practical use, as interactions with food constituents may change taste perception substantially. Other off-tastes occurring in some intense sweeteners are menthol- or licorice-like tastes.

Taste modification of intense sweeteners has been widely studied in order to improve their sensory characteristics, especially time – intensity profiles, and reduce aftertastes. A wide variety of substances is claimed to have taste-modifying properties [19].

2.4. Synergism

In sweetener blends, synergistic sweetness enhancement is often observed (Fig. 4). Commonly used synergistic sweetener blends are acesulfame – aspartame, aspartame – saccharin,

and cyclamate – saccharin. The apparent intensification of sweetness compared with single sweeteners is not fully understood. Even definitions of the term "synergism" and methods for determination of synergistic effects vary [20–22]. For a number of combinations, calculation of equisweet concentrations on the basis of sweetness intensity data for single sweeteners seems possible [12]. In these cases synergism has to be regarded as the consequence of nonlinear sweetness response functions. For certain examples, e.g., acesulfame – aspartame blends, sweetness intensities calculated on the basis of the single sweetener response curves are below the observed values.

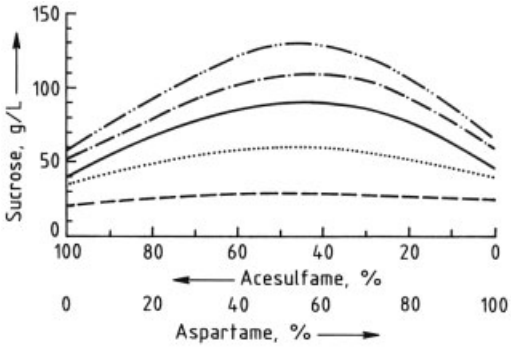

Figure 4. Equisweet concentrations of different levels of sucrose and acesulfame – aspartame blends
Sweetener blend, mg/L: (– –) 100; (· · ·) 200; (–) 300; (– · –) 400; (– · · –) 500

As time – intensity profiles of sweeteners are different, and often differ from the generally accepted standard sucrose, taste modifications can be achieved by blending; the resulting taste improvement is sometimes referred to as "qualitative synergism" [22].

When saccharin and cyclamate were generally available, blends of these two sweeteners were widely used. With the availability of acesulfame, aspartame, etc., synergism has attracted renewed interest, and blends of acesulfame – aspartame, aspartame – saccharin, and even tertiary or quaternary blends of sweeteners are increasingly used in order to exploit both types of synergism.

3. Uses

3.1. Foods and Beverages

Foods and beverages are the most important fields of application of intense sweeteners, with calorie reduction being the main goal. Single sweeteners or combinations with other sweet substances, e.g., carbohydrates, or sugar substitutes may be used. Intense sweeteners can be used in diabetic foods and beverages; depending on the type of product, either as single sweetening agents or combined with bulk sugar substitutes suitable for diabetic consumption (isomalt, lactitol, maltitol, mannitol, sorbitol, xylitol, etc.).

Whenever sugar and other sweet carbohydrates are used for their sweet taste, and not for functional reasons, replacement by intense sweeteners is generally simple. Functional sugar uses, such as viscosity enhancement, reduction of water activity, providing bulk, assistance in gelling, and others, cannot be provided by intense sweeteners owing to the minute concentrations used for normal sweetness levels. For foods requiring such functional properties, combinations of intense sweeteners and bulk sugar substitutes are an appropriate alternative whenever sugar or sweet carbohydrates are not to be used.

Beverage uses of intense sweeteners account for more than 50 % of human consumption; sugar replacement by intense sweeteners is simple, as carbohydrates do not play any important functional role in beverages, and the special taste characteristic commonly referred to as "body" can be achieved by addition of small quantities of sweet carbohydrates or viscosity-enhancing agents, such as certain polysaccharides.

Other important applications are fruit-flavored dairy products, desserts, etc.

In products such as pickles, marinated fish, and delicatessen salads, intense sweeteners not only round out the taste and balance the acidity, but also reduce the risk of spoilage (in contrast to sweet carbohydrates, they do not promote the growth of microorganisms).

In a number of applications, functional sweet carbohydrates are required for the specific product characteristics. For these applications, e.g., baked goods, confectionery products, and chewing gum, combinations of intense sweeteners

with bulk sugar substitutes or bulking agents are an alternative.

General information on applications is given in product-related brochures and some surveys [23–25].

3.2. Table-Top Sweeteners

For household use, intense sweeteners are formulated into table-top sweeteners, such as sweetener tablets, powders and spoon-by-spoon products, and liquids. Sweetener tablets may be produced as effervescent tablets containing a CO_2-releasing ingredient and an acid, such as sodium hydrogen carbonate and tartaric acid. Owing to the small amounts equivalent to a teaspoonful of sugar or a sugar cube, excipients, e.g., lactose or sugar alcohols, are often used for noneffervescent tablets.

For granular sweeteners, the same excipients or, if they are not intended for diabetics, glucose or sucrose are often used.

For spoon-by-spoon products, the bulk density of the granular product has to be reduced substantially; maltodextrins are most commonly used. To facilitate application of table-top sweeteners and to bring them close to standard sugar formulations, sugar-reduced cubes containing intense sweeteners with the same sweetness level as normal sugar cubes, and aqueous solutions of intense sweeteners, are also available.

3.3. Cosmetics

Several types of cosmetics, especially oral hygiene products, are sweetened to make them more pleasant for consumers. For oral hygiene products, noncariogenic ingredients have to be used. In dentifrices, polyols, such as glycerol or sorbitol, are used as humectants, and also provide some basic sweetness. The desired sweetness level is adjusted with an additional quantity of intense sweetener [26].

3.4. Pharmaceuticals

Intense sweeteners are used to mask undesired flavors and tastes of active pharmaceutical ingredients, e.g., bitterness, whenever the pharmaceuticals are intended for use by diabetics. Sweeteners are used in syrups, and soluble tablets and powders.

3.5. Feed

Intense sweeteners are used in animal feed to secure balanced nutrition. The main application is in starter feed for early weaning, in which intense sweeteners balance the palatability, and avoid decreased feed consumption immediately after weaning.

3.6. Others

Among other applications, tobacco and tobacco-related products are most important. In chewing tobacco, intense sweeteners are added to the flavoring systems. A further application is in cigarette tips, the palatability of which is improved by addition of intense sweeteners.

4. General Toxicology and Physiology

4.1. Toxicology

Intense sweeteners have to undergo the safety tests generally required for food additives, with long-term studies in at least two species, including a carcinogenicity study, mutagenicity studies in different test systems, studies on reproductive toxicity, pharmacokinetics, and metabolism.

Guidelines for these studies are available on an international and, in some countries, a national basis [27,28].

4.2. Metabolism

Most available intense sweeteners are not metabolized in the human body or they are excreted without metabolic utilization; they are therefore noncalorific. Some, especially peptide-based sweeteners, are digested and metabolized. In spite of this, these sweeteners are virtually noncalorific, as normal use levels are extremely low and the contribution to the nutritive value of foods and beverages is insignificant compared with nutritive bulk sweeteners.

4.3. Suitability for Diabetics

Intense sweeteners are metabolized independently of insulin, and therefore suitable for diabetics. Consumption of intense sweeteners in doses equivalent to a normal soft drink serving neither stimulates insulin secretion nor influences blood glucose levels [29].

4.4. Dental Effects

The commonly used intense sweeteners are not metabolized, or at least not into acids, by the bacteria of the oral cavity. They are therefore noncariogenic. Inhibitory effects on *Streptococcus mutans*, a caries-causing bacterium, have been demonstrated at sweetener concentrations above normal use levels. Synergistic enhancement of these effects in sweetener blends or blends of sweeteners and fluoride have also been demonstrated [30–32].

5. Food Legislation

5.1. International Regulations and Assessments

On the basis of available safety data, acceptability of intense sweeteners is assessed by several international authorities, especially the Joint Expert Committee on Food Additives of the WHO and FAO (JECFA), and the Scientific Committee for Food of the EU (SCF). JECFA bases a general evaluation on the available safety data. Acceptance for food use is demonstrated by allocation of an acceptable daily intake (ADI) value. The results of the evaluation are published as a toxicological monograph. In parallel, specifications are recommended which are normally used as guidelines for national food legislation. The SCF performs a similar evaluation, having produced a report on sweeteners in foods, and several subsequent publications [33].

Beyond national legislation, an EU directive on sweeteners harmonizes sweetener legislation in the Community, listing approved sweeteners, fields of application, and maximum use levels.

5.2. National Food Legislation

Under the food legislation of most countries, intense sweeteners are food additives, requiring approval for food use. Approvals are listed in national regulations, e.g., in the United States [34], the United Kingdom [35], and Germany [36].

6. Substances Commonly Used as Sweeteners

6.1. Acesulfame K

6.1.1. General Information

Acesulfame K was discovered by Clauss and Jensen during investigations on oxathiazinone dioxides. The sweet taste was found by chance [37, 38]. Several other oxathiazinone dioxides taste sweet, but have less favorable characteristics.

6.1.2. Physical and Chemical Properties

Acesulfame K is the potassium salt of 6-methyl-1,2,3-oxathiazin-4(3*H*)-one-2,2-dioxide (also 3,4-dihydro-6-methyl-1,2,3-oxathiazin-4-one-2,2-dioxide) [*55589-62-3*], $C_4H_4KNO_4S$, M_r 201.2.

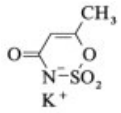

Acesulfame K

Acesulfame K does not have a well-defined melting point, and starts decomposing at ca. 225 °C. It is soluble in water (ca. 270 g/L at 20 °C and ca. 1300 g/L at 100 °C). Solubility in most organic solvents is low. Aqueous solutions of acesulfame K are neutral [39].

At pH 3 – 7, the common range for most foods and beverages, acesulfame K is virtually stable. Losses of a few percent were reported only after prolonged exposure to lower pH [39, 40]. Acesulfame K does not decompose during normal processing conditions of foods and beverages. No decomposition was observed during baking

[41], ultra-high-temperature treatment [42], and microwave heating [43]. Hydrolysis in strong acids yields acetone, CO_2, and ammonium sulfate [44].

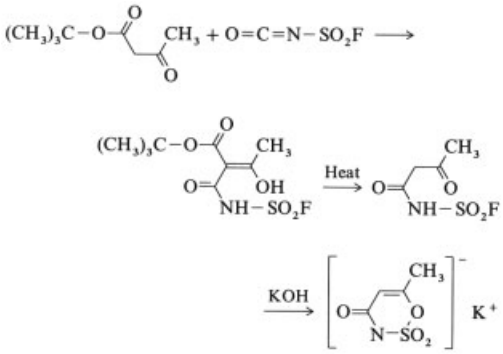

Figure 5. Synthesis of acesulfame K [37,46]

Compared with 3 % sucrose solution, the sweetness intensity of acesulfame K is approximately 200 [12]. An off-taste may become perceptible at elevated concentrations [18]. Acesulfame K is often synergistic with other sweeteners [45].

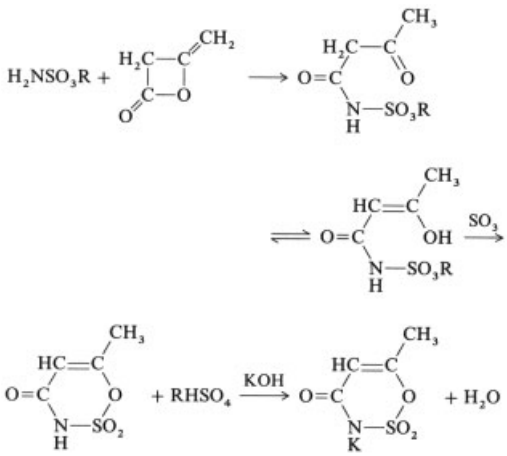

Figure 6. Synthesis of acesulfame K: alternative route [41, 48]

6.1.3. Production

Several synthesis routes for acesulfame K have been described. Potential synthesis starts from halogen sulfonyl isocyanates and keto compounds. Depending on the ketone, oxathiazinone dioxides with different substituents can be produced, some of which are also sweet, but inferior to acesulfame K in taste quality.

One production route (Fig. 5) starts from fluorosulfonyl isocyanate reacted with acetoacetic acid *tert*-butyl ester to a labile addition compound. On heating, this intermediate releases CO_2 and isobutene to form *N*-(fluorosulfonyl)acetoacetic acid amide; in the presence of bases, this compound cyclizes to the oxathiazinone dioxide ring [37,46].

An alternative synthesis route (Fig. 6) has been described in which diketene and amidosulfonic acid are reacted to acetoacetamide-*N*-sulfonic acid. In the presence of suitable dehydrating agents, acesulfame acid is formed, which can be neutralized with potassium hydroxide to acesulfame K [47,48].

6.1.4. Specifications and Analysis

For food use, acesulfame K has to conform to specifications laid down in national food legislation; these are often based on the suggestions made by JECFA [49], in which a purity > 99 % is required. Specific methods for the analysis of pure acesulfame K are given in these suggestions, and general methods are listed in a guide to the specifications [50].

Although several analytical methods for the quantitative determination of acesulfame K in foods and other products are available, HPLC analysis, allowing simultaneous determination of acesulfame K and other sweeteners, is often preferred [51,52].

6.1.5. Toxicology and Legislation

Acesulfame K has been subjected to the toxicological studies required for food additives. It is devoid of carcinogenicity, chronic toxicity, teratogenicity, and pharmacological side effects. It is absorbed quickly, and excreted at a similar rate, completely unmetabolized. Owing to the rapid excretion, accumulation in the body seems impossible. JECFA allocated an ADI of 0 – 15 mg/kg [53], as did the FDA [54].

Acesulfame K is approved for food use in most countries where intense sweeteners are used in foods and beverages, and is also approved for use in oral hygiene products and pharmaceuticals.

6.1.6. Uses

Acesulfame K is used in all fields of applications of intense sweeteners. Common applications are: table-top sweeteners; beverages; foods, such as dairy products, desserts, bakery products, confectionery, chewing gum, pickles, and marinated fish; oral hygiene products, such as tooth paste and mouthwash; and pharmaceuticals. Owing to its synergistic characteristics, acesulfame K is often used in sweetener blends, and in combination with bulk sweeteners in products requiring good stability, e.g., confectionery or bakery products [23, 40, 55].

6.2. Aspartame

6.2.1. General Information

The sweet taste of aspartame, *N*-L-α-aspartyl-L-phenylalanine methyl ester, was discovered by SCHLATTER [56, 57] in 1965, although the compound had been synthesized earlier. Systematic research on the sensory properties of dipeptide derivatives has revealed many sweet substances [58, 59].

6.2.2. Physical and Chemical Properties

Aspartame, [*22839-47-0*], $C_{14}H_{18}N_2O_5$, M_r 294.3 is a white crystalline powder virtually free from odor, with no well-defined melting point. At 196 °C itreacts to 5-benzyl-3,6-dioxo-2-piperazineacetic acid [*55102-13-1*] [60].

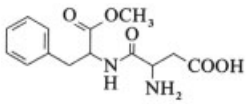

Aspartame

Solubility of aspartame in water depends on pH. At pH 5.2, the isoelectric point, solubility in water is approximately 1 g in 100 mL. With decreasing pH, solubility increases to a maximum value at pH 2.2. Water solubility also increases with temperature. In most organic solvents, the solubility is lower than in water [60].

At elevated temperature, solid aspartame slowly releases methanol to form aspartylphenylalanine and the dioxopiperazine. In aqueous solutions aspartame slowly hydrolyzes to amino acids and methanol; the dioxopiperazine can also be formed. The maximum stability in aqueous solution is at ca. pH 4. The rate of hydrolytic decomposition increases steeply with both increase and decrease of pH [60].

Aspartame stability studies have been carried out, e.g., in aqueous solutions [61, 62], beverages [63, 64], fruit preparations, yoghurt [65, 66], and under conditions favoring Maillard reactions [67]. Coating or encapsulation by cyclodextrins has been suggested to improve its stability [68].

Compared with 3 wt % aqueous sucrose solution, aspartame has a sweetness intensity of ca. 180 – 200 [12]. Its taste characteristics are generally assessed as good [18]. The sweetness perception can be slightly delayed and lasting [17].

6.2.3. Production

Production of aspartame normally starts from L-aspartic acid and L-phenylalanine or L-phenylalanine methyl ester. The production follows common routes of peptide synthesis, as the L-configuration of amino acids has to be retained (essential for sweetness).

Common production routes require protective groups for aspartic acid. The protective groups commonly used in peptide synthesis have been suggested for aspartame synthesis, but simple protection for large-scale production can be obtained by use of formylaspartic acid anhydride. It can be reacted with phenylalanine methyl ester to *N*-formylaspartame [69], from which the protective group has to be removed [70]. Especially simple production (Fig. 7, see next page) is possible from *N*-formylaspartic acid anhydride and L-phenylalanine. The resulting *N*-formylaspartylphenylalanine is treated with hydrochloric acid and methanol. The aspartame hydrochloride obtained can be crystallized for purification and transformed into aspartame after addition of a stoichiometric amount of a neutralizing agent [71, 72]. Adjustment of the solvent increases the yield by improving the ratio of α- to β-isomers [73]. Crystallization of

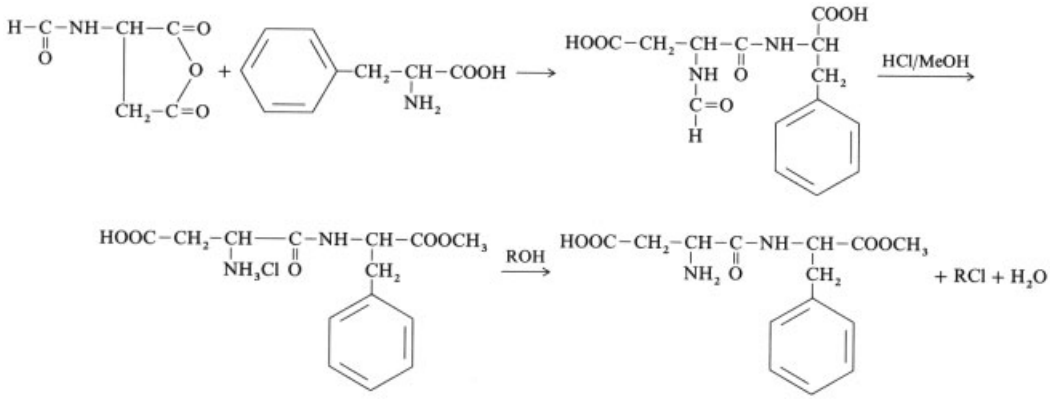

Figure 7. Synthesis of aspartame

α-aspartame hydrohalides is a suitable purification step, as β-hydrohalides are more soluble [74,75].

A commercially used alternative to chemical peptide synthesis is enzymatic formation of the peptide bond. Some proteinases, e.g., thermolysin, are able to catalyze peptide bond formation. Protected or nonprotected aspartic acid anhydride can be used. In contrast to chemical peptide bond formation, L-phenylalanine is selectively used from a racemic phenylalanine, and the remaining D-isomer can be racemized again [76].

6.2.4. Specifications and Analysis

National food legislation, for the use of aspartame in foods, is often based on the suggestions made by JECFA [77] in which a purity $> 98\,\%$ is required. Specific methods for the analysis of pure aspartame are given in these suggestions, and general methods are listed in a guide to the specifications [50]. Aspartame is also listed in the Food Chemicals Codex [78].

Although several analytical methods for the quantitative determination of aspartame in foods and other products are available, HPLC analysis, allowing simultaneous determination of aspartame and other sweeteners, is often preferred [51,52].

Methods allowing determination of aspartame and its breakdown products are available [61,63].

6.2.5. Toxicology and Legislation

Aspartame is devoid of carcinogenicity, chronic toxicity, and teratogenicity.

Being a dipeptide ester, aspartame is metabolized, with a nutritive value of 17 kJ/g (4 kcal/g); at normal use levels of several hundred milligrams per kilogram, this nutritive value has no practical importance.

JECFA allocated an ADI of 0 – 40 mg/kg [79], whereas the FDA value is 50 mg/kg [80].

Owing to its phenylalanine content, aspartame should be avoided by persons suffering from phenylketonuria, and a warning on aspartame-containing products is required in many countries.

Aspartame is approved for food use in most countries where intense sweeteners are used.

6.2.6. Uses

The sensory characteristics of aspartame allow its use in all common sweetener applications. Limitations are imposed by its susceptibility to hydrolytic decomposition, and limited temperature stability. The shelf life of aspartame-sweetened products with high water content has to be limited to periods in which no decrease in sweetness can be observed. No limitations exist for low-humidity products, unless they are exposed to high temperatures. Owing to the limited solubility of aspartame, preparations have been developed to facilitate dissolution [81].

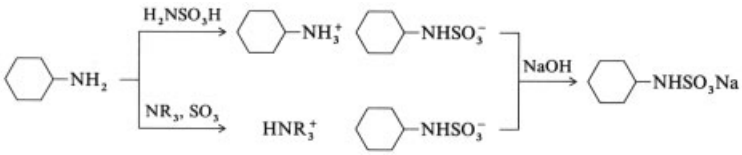

Figure 8. Synthesis of cyclamates

In addition to single-sweetener applications, the synergistic properties of aspartame are used in a variety of blends.

Applications of aspartame in foods, beverages, and other products are described in surveys [24, 82].

6.3. Cyclamate

6.3.1. General Information

The sweet taste of salts of cyclohexylaminosulfonic acid was discovered accidentally by SVEDA in 1937 [83].

6.3.2. Physical and Chemical Properties

Cyclamates are salts of cyclohexylaminosulfonic acid (cyclohexylsulfamic acid) [*100-88-9*], $C_6H_{13}NO_3S$, M_r 179.2. Cyclohexylsulfamic acid melts at 169 – 170 °C, but its salts do not have well-defined melting points and decompose at elevated temperature. Cyclohexylaminosulfonic acid is less suitable for food use but is approved, like its sodium, potassium, and calcium salts [84, 85].

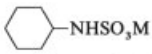

Cyclamates (M = e.g., Na, K, ½ Ca)

Sodium cyclamate, $C_6H_{12}NNaO_3S$, M_r 201.2 [*139-05-9*], is the most important cyclamate sweetener. It forms white crystals, decomposing at ca. 260 °C. Its solubility is 20 wt % in water at room temperature, whereas in most organic solvents solubility is low or extremely low [84, 85].

Calcium cyclamate [*5897-16-5*], $C_{12}H_{24}CaN_2O_6S_2 \cdot 2\,H_2O$, M_r 432.6, is of less importance. At pH 2 – 10, the stability of cyclamates to hydrolytic decomposition is high, so they are not affected by common food processing conditions [84, 85].

Compared with 0.1 mol/L sucrose solutions, the sweetness intensity of sodium cyclamate is ca. 35 [12]. Aftertastes become perceptible at elevated concentrations only, rendering cyclamate generally superior to saccharin [18].

6.3.3. Production

The starting material for cyclamate production is cyclohexylamine, which is reacted with sulfonating agents (Fig. 8) [86]. The originally used chlorosulfonic acid has been replaced by other compounds. In high-boiling-point solvents, cyclohexylamine reacts with sulfamic acid to *N*- cyclohexyl-*N*- cyclohexylammonium sulfamate. Reaction with sodium or calcium hydroxide yields cyclamates and free cyclohexylamine. In a corresponding way, equimolar amounts of cyclohexylamine and tertiary amines react to trialkylammonium-*N*- cyclohexyl sulfamates, the respective cyclamates being formed after reaction with sodium or calcium hydroxide. Further possibilities are reactions of cyclohexylamine with sodium methylthiosulfate or sulfamic acid metal salts. Reactions of cyclohexyl compounds with SO_3 or sulfuric acid, or of cyclohexylhydroxylamine and SO_2 or sulfites, producing cyclamates or cyclohexylsulfamic acid have also been described [87, 88].

6.3.4. Specifications and Analysis

The suggestions made by JECFA for cyclamates in foods [89] require a purity > 98 %. Specific methods for the analysis of pure cyclamate are given in these suggestions, and general methods are listed in a guide to the specifications [50].

Owing to the absence of a chromophore for UV detection, analytical methods for cyclamates have to be based on either color changes or

other properties. A suitable HPLC method with post-column derivatization for easy detection has been described [90].

6.3.5. Toxicology and Legislation

Reevaluations of toxicological data have concluded that cyclamates are devoid of carcinogenicity. Discussion focused on the toxicity of cyclohexylamine, a potential metabolite formed by intestinal bacteria [91]. Current ADI values are based on the toxicity data for cyclohexylamine, and the assumption of a fairly high metabolic rate in the human body. Both JECFA and SCF allocated ADI values of 0 – 11 mg per kilogram body weight [33, 92].

Whereas some countries approve cyclamate use in a variety of applications, some important sweetener markets, such as the United States, still ban them.

6.3.6. Uses

Owing to their good stability, cyclamates are suitable for all applications of intense sweeteners. They can be used in acid products, e.g., soft drinks and pickles, and withstand exposure to elevated temperature as used in the production of canned fruit [85].

The main application of cyclamates, however, is in blends with saccharin in a 10:1 ratio by weight. In these blends, quantitative and qualitative synergistic effects can be used to advantage, rendering them an important sweetener system in countries approving the use of both sweetenes [85].

6.4. Saccharin

6.4.1. General Information

Saccharin was the first intense sweetener ever introduced. It was discovered in 1878 [2]. Small-scale production commenced in 1884, and it has been available since. For decades, saccharin was the most important intense sweetener.

6.4.2. Physical and Chemical Properties

Saccharin, often referred to as *o*-benzoic acid sulfimide, or *o*-sulfobenzoic acid imide, is 1,2-benzisothiazol-3(2*H*)-one-1,1-dioxide [*128-44-9*], $C_7H_5NO_3S$, M_r 183.2. Solid saccharin is a white, crystalline, odorless powder, melting at 228 – 230 °C. Solubility in water at 20 °C is ca. 2 g/L and increases to ca. 40 g/L at 100 °C. Aqueous solutions of saccharin are acidic, pH ca. 2 for a saturated solution at room temperature. In aqueous solutions of alkali hydroxides or carbonates, saccharin is easily soluble, forming salts [93].

The sodium salt is more important than saccharin itself. Sodium saccharin [*6155-57-3*], $C_7H_4NNaO_3S \cdot 2\,H_2O$, M_r 241.2, crystallizes with water of crystallization. Commercial preparations, however, may not contain the stoichiometric water content, but a slightly lower value. In contrast to saccharin itself, sodium saccharin is soluble in water. At room temperature, ca. 1000 g can be dissolved in 1 L of water, and more than 3000 g dissolve in 1 L of water at 100 °C [93].

Calcium saccharin, $C_{14}H_8CaN_2O_6S_2 \cdot 31/2\,H_2O$, M_r 467.5 [*6381-91-5*], is of less importance than sodium saccharin. At 20 °C, ca. 370 g can be dissolved in 1 L of water, but at 100 °C solubility is almost as high as for sodium saccharin [93].

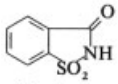

Saccharin

Compared with dilute sucrose solutions, the sweetness intensity of saccharin is ca. 550, whereas the values for its salts are ca. 450 [12]. At medium or elevated use levels, saccharin sweetness is accompanied by a metallic or bitter aftertaste [18]. Several masking agents for this side-taste have been suggested, and sweetener blends, especially a blend of 1 part sodium saccharin to 9 – 10 parts sodium cyclamate, are often used [94].

6.4.3. Production

In most countries, commercial saccharin is produced by the Remsen – Fahlberg process (Fig. 9), which basically follows the synthesis

described by its discoverers [95,96]. Toluene is used as the starting material and reacted with chlorosulfonic acid to a mixture of isomeric toluene sulfochlorides. In the presence of ammonia, 2-toluene sulfochloride forms 2-toluene sulfonic acid amide which is oxidized under appropriate conditions to saccharin. Potential oxidizing agents are potassium permanganate or chromic acid; syntheses not involving oxidizing agents have also been described [97].

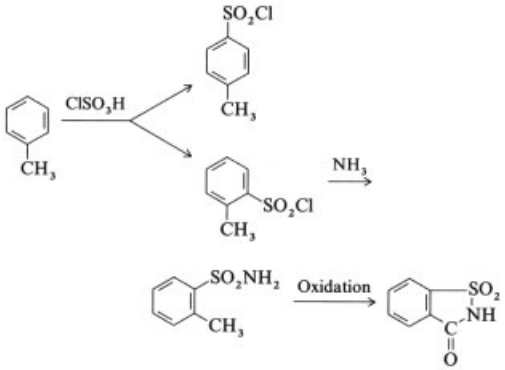

Figure 9. Remsen – Fahlberg synthesis of saccharin [96]

In the United States, the Maumee process is used (Fig. 10). This starts from either anthranilic acid or its methyl ester, which can be obtained from phthalic acid anhydride. With sodium nitrite, anthranilic acid forms a diazonium compound, which is reacted with sulfur dioxide or sulfites, and chlorinated to 2-chlorosulfonylbenzoic acid methyl ester. In the presence of ammonia, this ester yields the respective amide, from which, after de-esterification, saccharin is obtained. Reaction with sodium hydroxide or calcium hydroxide forms the respective saccharin salts [98].

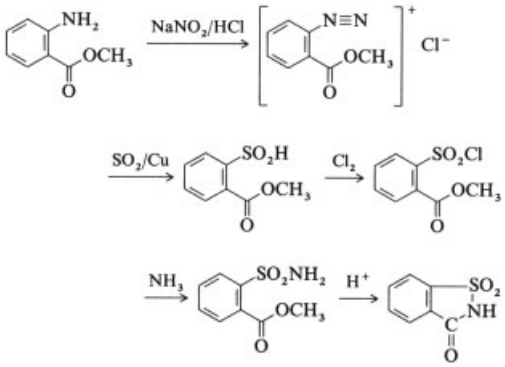

Figure 10. Maumee synthesis of saccharin [98]

6.4.4. Specifications and Analysis

The suggestions made by JECFA [99] for saccharin in foods require a purity > 99 %. Specific methods for the analysis of saccharin and saccharin salts are given in these suggestions, and general methods are listed in a guide to the specifications [50]. Saccharin and its sodium and calcium salts are also listed in the Food Chemicals Codex [98].

Although several analytical methods for the quantitative determination of saccharin and sodium saccharin in foods and other products are available, HPLC analysis, allowing simultaneous determination of saccharin, sodium saccharin, and other sweeteners, is often preferred [51,52].

6.4.5. Toxicology and Legislation

Saccharin is not metabolized in the human body. It is without influence on diabetics. Safety studies on saccharin have been repeatedly performed. In long-term feeding studies in rats an increased bladder tumor rate was observed after prolonged ingestion of high doses. This was attributed to a nongenotoxic action of saccharin, or to a co-carcinogenic effect. On the basis of further studies, the relevance of these observations for humans has been questioned. Information on saccharin toxicology has been reviewed [100]. JECFA allocated an ADI of 0 – 5 mg/kg [101].

Saccharin use is restricted in several countries, although in most countries it is still available as an intense sweetener. In the United States, a special label is required for saccharin-sweetened products [102].

6.4.6. Uses

Saccharin is most frequently used as its sodium salt. Important fields of application are soft drinks, table-top sweeteners, and desserts. For taste reasons, blends with other intense sweeteners, or combinations with reduced sugar levels are preferred wherever such blends are approved. In oral hygiene products, e.g., toothpastes and mouthwashes, saccharin masks undesired tastes of other ingredients [93].

In starter feed for livestock, saccharin is used to avoid reduced feed intake after weaning.

Beside its applications as an intense sweetener, saccharin is used in electrolytic nickel deposition. Addition of saccharin to the nickel salt solutions increases the hardness and brightness of the nickel plate. This effect is apparently specific to saccharin [93].

7. Sweeteners of Lesser Importance

7.1. Glycyrrhizin

7.1.1. General Information and Properties

Glycyrrhizin is the name for mixed calcium and potassium salts of glycyrrhizic acid which are found in licorice root, *Glycyrrhiza glabra* L., grown in Southern Europe and Asia.

Glycyrrhizic acid [*1405-86-3*], $C_{42}H_{62}O_{16}$, M_r 823.0, is a glycoside, consisting of glycyrrhetic acid, a triterpene of the β-amyrine type, and two β-1,2-glycosidic-linked glucuronic acids. Extracts containing 6 – 14 % glycyrrhizinates are available as licorice. It is used as a flavoring ingredient in confectionery and pharmaceuticals [103].

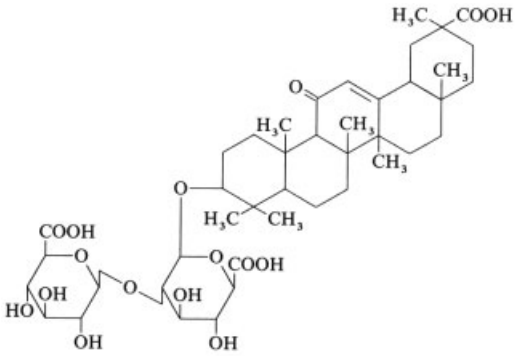

Glycyrrhizic acid

In addition to calcium and potassium glycyrrhizinates, crude ammonium glycyrrhizinate and monoammonium glycyrrhizinate are available. Ammonium glycyrrhizinate dissolves in water above pH 4.5, and in aqueous ethanol, but the solubility of monoammonium glycyrrhizinate in water is fairly low. At elevated temperature, glycyrrhinates are not fully stable to hydrolytic decomposition [103].

Glycyrrhizin is claimed to be 50 – 100 times sweeter than sucrose in dilute aqueous solution, whereas for ammonium glycyrrhizinate the sweetness intensity is ca. 50. Glycyrrhizinates are characterized by a delayed sweetness onset, a lingering sweetness, and a pronounced licorice taste; they are therefore considered to be flavorings rather than sweeteners [103].

7.1.2. Production

The starting material for the production of glycyrrhizin is dried, cut licorice root, with a glycyrrhizin content of 2 – 7wt %, depending on its origin. A crude extract is obtained by countercurrent extraction with water. After removal of polysaccharides, glycyrrhizic acid can be precipitated from the crude extract with sulfuric acid. Ammonium glycyrrhizinate is produced by dissolving crude glycyrrhizic acid in aqueous ammonia, and subsequent drying or precipitation with ethanol. Repeated recrystallization yields monoammonium glycyrrhizinate which, in contrast to the dark colored ammonium glycyrrhizinate, is almost colorless [103].

7.1.3. Toxicology and Legislation

Glycyrrhizin is characterized by a variety of physiological activities, which discourage the use of large quantities for sweetening purposes. Accordingly, it is not approved as a sweetener in many countries, but can be used as a licorice flavoring agent.

7.1.4. Uses

In addition to their flavoring properties, glycyrrhizin and ammonium glycyrrhizinate have flavor-enhancing effects in certain products. Important fields of application are confectionery, soft drinks with higher pH levels, tobacco products, and oral hygiene products [103].

Extracts of licorice root, glycyrrhizin, and glycyrrhizic acid have been suggested for treatment of a number of diseases, e.g., gastric ulcer [104].

7.2. Neohesperidin Dihydrochalcone

7.2.1. General Information and Properties

Neohesperidin dihydrochalcone, 2′,3,4′,6′-tetrahydroxy-4-methoxydihydrochalcone-4′-β-neohesperidoside [*20702-77-6*], $C_{28}H_{36}O_{15}$, M_r 612.6 is the important compound among the sweet dihydrochalcones. A number of related dihydrochalcones, e.g., naringin dihydrochalcone, are without practical importance.

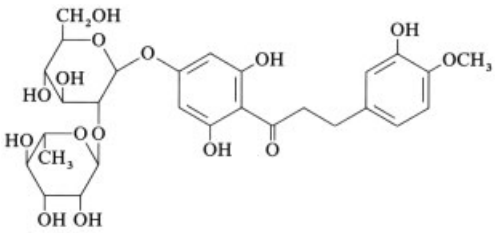

Neohesperidin dihydrochalcone

Neohesperidin dihydrochalcone has a melting range of 152 – 154 °C. In water only ca. 0.5 g/L dissolves at room temperature, but much higher quantities can be dissolved at high temperature and in ethanol (ca. 20 g/L). In alkaline solution, dissolution is favored owing to salt formation [105].

Despite the possibility of hydrolytic cleavage of the glycosidic bond, the stability of neohesperidin dihydrochalcone is acceptable under food processing and storage conditions [105].

Compared with dilute sucrose solutions, the sweetness intensity of neohesperidin dihydrochalcone is ca. 330. The much higher values sometimes reported were determined at the taste threshold. Neohesperidin dihydrochalcone is characterized by a pronounced menthol-like aftertaste which limits its applicability and allows it to make only a partial contribution to the total sweetness of most products [105].

Neohesperidin dihydrochalcone can be analyzed by HPLC [106].

7.2.2. Production

Neohesperidin dihydrochalcone is obtained by hydration of neohesperidin under alkaline conditions. This compound can be isolated from the skins of Seville oranges. A further production route starts from naringin (Fig. 11), found in the skins of several other citrus species. Treatment of naringin with alkali produces phloroacetophenone-4′-β-neohesperidoside which is condensed with isovanillin to neohesperidin chalcone. Hydrogenation under alkaline conditions yields neohesperidin dihydrochalcone [107, 108].

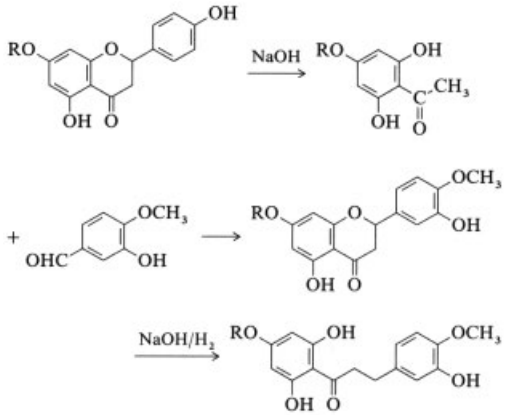

Figure 11. Synthesis of neohesperidin dihydrochalcone, starting from naringin

7.2.3. Toxicology and Legislation

Neohesperidin dihydrochalcone has been tested in a full range of safety studies, and an acceptable daily intake value of 0 – 5 mg per kilogram body weight was allocated by SCF [109]. Approval of this sweetener is limited to few countries, e.g., those in which it can be produced from the skin of citrus fruits. Extension of approval seems possible.

7.2.4. Uses

The specific taste characteristics of neohesperidin dihydrochalcone limit its use to few applications, or low use levels. It acts synergistically with a number of other sweeteners, e.g., acesulfame and aspartame, and can be used in fairly small quantities in blends with these sweeteners. The main fields of application are certain confectionery products and chewing gum, but small concentrations are suitable for use in beverages and dairy products. In addition to sweetness, neohesperidin dihydrochalcone

has flavor-enhancing properties in a variety of foods and other products [105].

7.3. Stevioside and Rebaudioside

7.3.1. General Information and Properties

Stevioside and other steviol glycosides are constituents of the South American plant *Stevia rebaudiana* Bertoni. It is claimed that the leaves of the plant have been used for centuries to sweeten foods and beverages.

Stevioside is a glycoside consisting of the aglycone steviol, *ent*-13-hydroxykaur-16-en-19-oic acid, and three glucose molecules, two of which are linked to the disaccharide sophorose. Stevioside is therefore named as 13-*O*-β-sophorosylsteviol-18-*O*-β-glucosyl ester [*57817-89-7*], $C_{38}H_{60}O_{18}$, M_r 804.9.

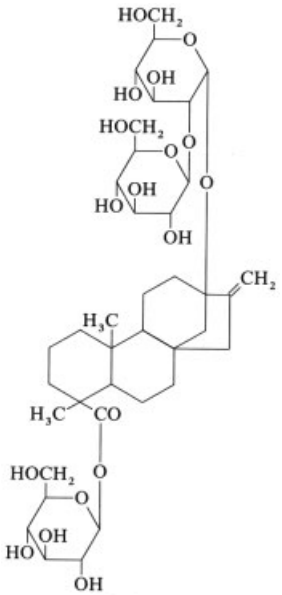

Stevioside

Beside steviol, the leaves of *Stevia rebaudiana* contain a number of other sweet glycosides, the most important of which is rebaudioside A. Rebaudioside A is 13-*O*-β-glycopyranosylsteviol-19-*O*-β-glucopyranosyl-(1,3)-sophorosyl ester [*58543-16-1*], $C_{44}H_{70}O_{23}$, M_r 967.0. Further sweet constituents of *Stevia* leaves are other steviol glycosides, such as other rebaudiosides, dulcoside, and steviolbioside.

Stevia originates in South America, but is also grown in several Asian countries. Dried *Stevia* leaves contain 2 – 22 wt % of stevioside with a mean of ca. 7 %. Rebaudioside levels are in the range 1.5 – 10 % [110].

Stevioside melts in the range 196 – 198 °C. The solubility in water is approximately 1.2 – 1.3 g/L, only slightly higher than the concentrations necessary for medium or high sweetness. Stevioside is of acceptable stability for food production and storage, e.g., for use in soft drinks. At elevated temperature, the stability decreases with decreasing pH [111].

Compared with dilute sucrose solution, stevioside is ca. 160 – 170 times sweeter. The higher sweetness intensity values often reported were determined at the taste threshold. The sweetness of stevioside is accompanied by a licorice-like aftertaste. Side-tastes and other sensory characteristics depend on the decomposition of stevioside products which are often purified extracts containing other plant constituents in addition to stevioside [112].

Rebaudioside A melts in the range 242 – 244 °C. In water, itis more soluble than stevioside. Although the stabilities of steviosides and rebaudioside A are similar, rebaudioside A has been found to decompose on exposure to light [111].

The sweetness intensity of rebaudioside A is approximately one-third higher than that of stevioside, and its taste characteristics are superior [110].

7.3.2. Production

Stevioside can be isolated from the leaves of *Stevia* by extraction with water or water – ethanol mixtures. Crude *Stevia* extracts are more or less purified, e.g., by treatment with calcium or magnesium hydroxides or carbonates. Further purification by ion-exchangers, or absorption of undesired ingredients is possible. Stevioside can be obtained from the extracts by precipitation with methanol [112].

Depending on the processing conditions, stevioside products may contain other sweet steviol glycosides, and other constituents. Accordingly,

the properties of commercially available stevioside products vary.

Rebaudioside A is isolated from *Stevia* leaves in a similar way to stevioside. The ratio of rebaudioside A to stevioside can be increased by extraction with methanol, and further purification with column chromatography yields rebaudioside-A-rich products. Stevioside can, in addition, be transformed enzymatically to rebaudioside A [110].

Commercial production routes often fail to yield pure stevioside or rebaudioside A.

7.3.3. Toxicology and Legislation

Although safety studies on stevioside and rebaudioside have been performed, they have not been accepted internationally, owing to the lack of generally accepted specifications. Accordingly, no ADI has been allocated for these compounds on an international basis. *Stevia* extracts are approved for food use in several South American and Asian countries, but lack approval in Europe and North America [110, 111].

7.3.4. Uses

Stevioside, the more important product, is used in table-top sweeteners, confectionery, certain soft drinks, fruit products, and Japanese-style vegetable products in countries approving the sweeteners, generally as stevioside-rich *Stevia* extracts.

7.4. Sucralose

7.4.1. General Information and Properties

Sucralose is the common name for 1′,6′-dichloro-1′,6-dideoxy-β-D-fructofuranosyl-4-chloro-4-deoxy-α-D-galactopyranoside, sometimes referred to as trichlorogalactosucrose.

The sweetener was discovered in 1976 [113]. Following the discovery that halogenation increases the sweetness of carbohydrates substantially, several chlorinated and brominated mono- and disaccharides of intense sweetness were found.

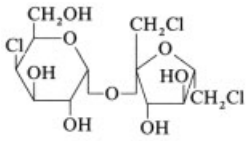

Sucralose

Sucralose [*56038-13-2*], $C_{12}H_{19}O_8Cl_3$, M_r 397.6. The solubility in water is ca. 283 g/L at 20 °C and ca. 90 g/L in ethanol [114]. Sucralose melts at 125 °C with decomposition.

Solid sucralose is not fully stable, and slowly releases HCl, with discoloration. In contrast, aqueous solutions are highly stable, and it is marketed in this form. Owing to its stability over the normal pH range of foods and beverages, it is suitable for all liquid sweetener applications [115].

Compared with dilute sucrose solution, sucralose is ca. 600 times sweeter. The sweetness is

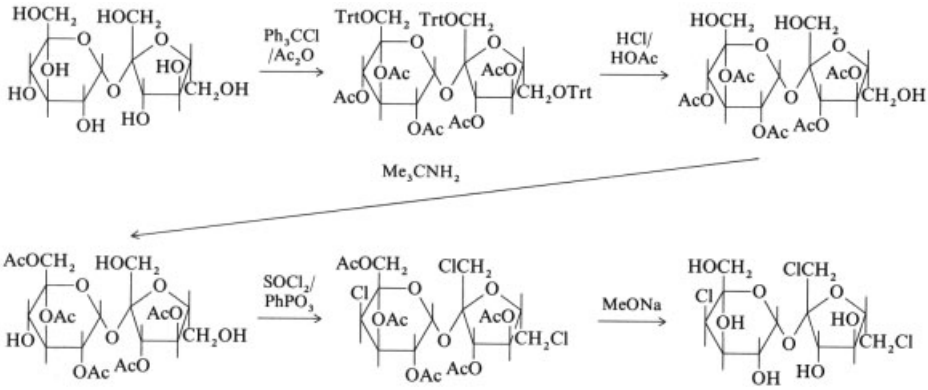

Figure 12. Synthesis of sucralose
Trt = Ph_3C = trityl group

perceived with a slight delay and a lasting effect, similar to aspartame, and generally considered of good quality.

7.4.2. Production

The starting material for the synthesis of sucralose (see Fig. 12) is sucrose. The production route commonly described blocks the three primary hydroxyl groups as trityl ethers, and transforms the secondary hydroxyl groups to acetates to yield 6,1′,6′-tritrityl-2,3,4,3′,4′-pentaacetylsucrose. By deblocking the primary hydroxyl groups, 2,3,4,3′,4′-pentaacetylsucrose is obtained. Under appropriate conditions, this compound can be transformed to 2,3,6,3′,4′-pentaacetylsucrose. Chlorination at the exposed hydroxyl groups gives 4,1′,6′-trichloro-4,1′,6′-tridesoxygalactosucrose on conversion of the glucose moiety to galactose. Removal of the acetate groups by hydrolysis yields sucralose. Other blocking agents have also been described [116].

7.4.3. Toxicology and Legislation

Sucralose has been studied in the test systems commonly used for the evaluation of food additives. JECFA allocated an ADI of 0 – 15 mg/kg [117]. Sucralose has not yet been endorsed by other international authorities, and still lacks approval in most countries.

7.4.4. Uses

Owing to the lack of approval, sucralose uses are limited. Its good stability is considered an incentive for single sweetener use, which appears feasible on the basis of its sensory characteristics, but its synergistic characteristics probably favor its use in blends.

7.5. Thaumatin

7.5.1. General Information and Properties

Thaumatin is a mixture of structurally related proteins found in the arils of the fruits of *Thaumatococcus daniellii* Benth, a plant originating in West Africa, where it is grown on plantations. The sweet taste of *Thaumatococcus* fruits, locally called Katemfe and traditionally used for sweetening purposes, was first described in the scientific literature in 1855 [1].

Thaumatin consists of several structurally related polypeptides with approximate M_r 21 000 [*53850-34-3*]. Normally, thaumatin I and II are the major components, but other similar peptides may be present in thaumatin products. The amino acid composition and tertiary structure are known [118, 119].

Thaumatin is highly soluble in water, and solutions containing more than 1000 g of thaumatin in 1 L of water can be prepared [120]. Thaumatin stability data in the literature are not consistent. Good stability below pH 5.5 has been claimed, but limited stability is also reported [120].

Compared with dilute sucrose solutions, thaumatin is approximately 3500 times sweeter. The sweetness shows a delayed onset and long persistence, accompanied by a licorice-like sidetaste. Thaumatin is often used as a flavor enhancer rather than as a sweetener [120].

7.5.2. Production

Thaumatins are extracted from the arils of *Thaumatococcus* with water; removal of solids is by centrifugation and ultrafiltration. Further purification is possible by ion-exchange chromatography [120].

As production of thaumatin in countries growing *Thaumatococcus* is difficult, attempts have been made to produce thaumatin by means of genetically engineered microorganisms. Expression of the thaumatin genes in several microorganisms has been achieved, but no biotechnological production has yet been established [121, 122].

7.5.3. Toxicology and Legislation

Thaumatin is digested like other peptide food constituents. No effects of toxicological significance have been observed in safety studies on thaumatin. It has been accepted internationally as a food additive by JECFA [123], with ADI "not specified". It is approved as a sweetener and flavor enhancer in a number of countries.

7.5.4. Uses

Although thaumatin was primarily introduced as a sweetener, the temporal sweetness characteristics do not favor its use as a single sweetener, and limit blend use to fractions of the total sweetness.

Flavor-enhancing properties, in contrast, are of some importance; thaumatin is used as a flavor enhancer in confectionery, chewing gum, and similar products [120].

8. New Developments

8.1. Alitame

Alitame, L-α-aspartyl-*N*-(2,2,4,4-tetramethyl-3-thietanyl)-D-alanine [*80863-62-3*], $C_{14}H_{24}O_4N_3S \cdot 21/2\,H_2O$, M_r 375.3, is an intensely sweet dipeptide derivative [124].

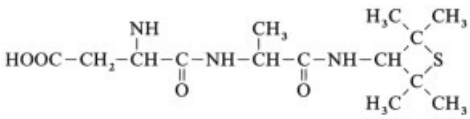

Alitame

Alitame melts at 100 °C, solidifies at 102 °C, and starts decomposing at 136 – 137 °C. At room temperature, ca. 130 g/L can be dissolved in water at pH 5.0, with increasing solubility at higher and lower pH, and higher temperature [125].

Being a dipeptide, alitame is not completely stable, but superior to aspartame at all important pH levels, especially in the neutral range.

Compared with dilute sucrose solutions, its sweetness intensity is > 2500. Sensory properties are similar to aspartame.

Alitame is under safety evaluation and has not been endorsed as a food additive by international authorities, nor yet approved in most countries [125].

8.2. Others

Systematic investigations into structure – activity relationships of sweeteners have led to a variety of products with sweetness intensities comparable to the sweetest products already known, or even higher. A number of products have been described, e.g., derivatives of *N*-phenylguanidinoacetic acid and *N*-phenylethanamidinoacetic acid [126, 127], tetrazole derivatives [128], or amide-, glycine- and β-alanine-based sweeteners with heterocyclic substituents [129]. These products are still far from being marketed, although some are apparently under safety investigation for human consumption.

Monellin is a polypeptide constituent of the fruits of *Dioscoreophyllum cumminsii* Stapf Diels, with approximate M_r 11 500. The yield from 1 kg of fruitis ca. 3 – 5 g [130, 131]. Its sweetness intensity is ca. 2000. Owing to its structure, monellin is sensitive to conformational changes caused by heat or hydrolytic decomposition, rendering it unsuitable for normal sweetener use. As monellin is intensely sweet, and has a pleasant taste, studies have been initiated to improve the stability by structural changes, without impairing the good sensory characteristics.

9. Substances Formerly Used as Sweeteners

9.1. Dulcin

Dulcin is the common name of *N*-(4-ethoxyphenyl)urea [*150-69-6*], $C_9H_{12}N_2O_2$, M_r 180.2. Dulcin melts at 173 °C and is sparingly soluble in water (ca. 1.25 g/L). Compared with dilute sucrose solutions, it is approximately 250 times sweeter. Sweetness characteristics are good.

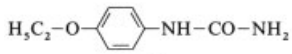

Dulcin

Dulcin was synthesized by reaction of phenetidine hydrochloride with sodium cyanate or urea, and by reaction of 4-phenetidine with phosgene and ammonia [132, 133]. Dulcin achieved some practical importance as a sweetener, despite safety concerns. It was blended with saccharin or sodium saccharin in order to mask the aftertaste of saccharin. As some dulcin metabolites were considered potentially hazardous, it is now banned in most countries.

9.2. Others

Suosan is the common name for the sodium salt of *N*-(ethyl)-*N*-(4-nitrophenyl)urea, $C_{10}H_{10}N_3NaO_5$, M_r 275.2 [134]. Pure suosan is a yellow crystalline substance ca. 300 times sweeter than sucrose in dilute solutions. It can be obtained by reacting 4-nitrophenyl isocyanate and β-alanine.

For some time, the sodium salt of *2-(4-methoxybenzoyl)benzoic acid*, $C_{15}H_{11}NaO_4$, M_r 278.3 [135], was offered as the product S 23/46. This substance is ca. 150 times sweeter than sucrose, but is bitter at elevated concentrations. Low solubility of the corresponding acid, which is released in acid foods and beverages, limited the applicability. The compound was synthesized from phthalic acid anhydride and anisole in the presence of aluminum chloride.

Several substituted nitroanilines are distinguished by high sweetness intensity. The sweetest compound in this group is *2-propoxy-5-nitroaniline*, $C_9H_{12}N_2O_3$, M_r 196.3, which is ca. 4000 times sweeter than sucrose. Attempts to market this product, as Ultrasüß or P 4000, were unsuccessful, owing to unsatisfactory sensory properties. It was synthesized by reaction of 1-chloro-2,4-dinitrobenzene and sodium propylate, with subsequent partial reduction, or by reaction of 2-nitrophenol with 1-chloropropane, reduction of the intermediate to the amino compound, and subsequent nitration [135]. *Perillatine*, the *E*-oxime of 4-(2-propenyl)-1-cyclohexene carboxaldehyde, $C_{10}H_{15}NO$, M_r 165.2, commonly called perilla aldehyde, was marketed for a time in the United States. The product is ca. 2000 times sweeter than sucrose, but has less than satisfactory sensory characteristics [136].

10. References

1. W. F. Daniell, *Pharm. J.* **14** (1855) 158 – 160.
2. C. Fahlberg, I. Remsen, *Ber. Dtsch. Chem. Ges.* **12** (1879) 469 – 473.
3. R. S. Shallenberger, T. E. Acree, *Nature (London)* **216** (1967) 480 – 482.
4. L. B. Kier, *J. Pharm. Sci.* **61** (1972) 1394 – 1397.
5. J. M. Tinti, C. Nofre in D. E. Walters, F. T. Orthoefer, G. E. DuBois (eds.): "Sweeteners, Discovery, Molecular Design, and Chemoreception," *ACS Symp. Ser.* **450** (1991) 206 – 213.
6. A. van der Heijden, H. van der Wel, H.-G. Peer, *Chem. Senses* **10** (1985) 57 – 72.
7. A. van der Heijden, H. van der Wel, H.-G. Peer, *Chem. Senses* **10** (1985) 73 – 88.
8. M. Tamura, I. Shinoda, H. Okai, C. A. Stammer, *J. Agric. Food Chem.* **37** (1989) 737 – 740.
9. H.-D. Belitz et al. in G. Charalambous (ed.): *Frontiers of Flavours,* Elsevier, Amsterdam 1988, pp. 49 –62.
10. T. Suami, L. Hough, *Food Chem.* **46** (1993) 235 – 238.
11. J. C. Culberson, D. E. Walters in D. E. Walters, F. T. Orthoefer, G. E. DuBois (eds.): "Sweeteners, Discovery, Molecular Design, and Chemoreception," *ACS Symp. Ser.* **450** (1991) 214 – 223.
12. K. Hoppe, *Lebensmittelindustrie* **38** (1991) 13 – 14.
13. A. Fricker, J. Gutschmidt, E. Prochazka, *Dtsch. Lebensm. Rundsch.* **71** (1975) 138 – 139.
14. A. Tunaley, D. M. H. Thomson, J. A. McEwan, *Int. J. Food Sci. Technol.* **22** (1987) 627 – 635.
15. D. B. Ott, C. L. Edwards, S. J. Palmer, *J. Food Sci.* **56** (1991) 535 – 542.
16. W. E. Lee, R. M. Pangborn, *Food Technol. (Chicago)* **40** (1986) no. 11, 71 – 78, 82.
17. A. C. Noble, N. L. Matysiak, S. Bonnans, *Food Technol. (Chicago)* **45** (1991) no. 11, 121 – 124, 126.
18. K. Hoppe, B. Gassmann, *Nahrung* **29** (1985) 417 –420.
19. Hoechst, DE-OS 3 331 517, 1982 (G.-W. von Rymon Lipinski).
20. N. Ayya, H. T. Lawless, *Chem. Senses* **17** (1992) 245 –259.
21. R. A. Frank, S. J. Mize, R. Carter, *Chem. Senses* **14** (1989) 621 – 632.
22. G.-W. von Rymon Lipinski, *Int. Food Marketing Technol.* **4** (1990) no. 5, 22 – 25.
23. Hoechst, SUNETT®, Product Brochure, Frankfurt 1992.
24. Nutrasweet AG, NutraSweet, Technical Bulletin, Zug 1987.
25. G.-W. von Rymon Lipinski, *Int. Food Marketing Technol.* **5** (1991) no. 2, 5, 8 – 11.
26. G.-W. von Rymon Lipinski, E. Lück, *Manuf. Chem.* **52** (1981) no. 5, 37.
27. Principles for the Safety Assessment of Food Additives and Contaminants in Food, *Environ. Health Criter.***70** (1987).

28. Toxicological Principles for the Safety Assessment of Direct Food Additives and Color Additives Used in Food, *US Food Drug Adm. Bur. Foods,* Washington 1982.
29. B. Härtel, H.-J. Graubaum, B. Schneider, *Ernaehr. Umsch.* **40** (1993) 152 – 155.
30. S. C. Ziesenitz, G. Siebert, *Caries Res.* **20** (1986) 498 – 502.
31. G. Siebert, S. C. Ziesenitz, J. Lotter, *Caries Res.* **21** (1987) 141 – 148.
32. A. T. Brown, G. M. Best, *Caries Res.* **22** (1988) 2 – 6.
33. Commission of the European Communities: "Sweeteners," *Report of the Scientific Committee for Foods,* **no. 16,** Luxembourg 1985.
34. US Code of Federal Regulations, Title 21, parts 172, 180, and 189.
35. *The Sweeteners in Food Regulations 1983,* SI 1983, no. 1211, London 1983.
36. Zusatzstoff-Zulassungsverordnung, amendment dated 13. 06. 1990, BGBl I 1990, 1053 – 1067.
37. Hoechst, DE 2 001 017, 1970 (K. Clauß, H. Jensen).
38. K. Clauß, H. Jensen, *Angew. Chem.* **85** (1973) 965 –973; *Angew. Chem. Int. Ed.* **12** (1973) 869 – 876.
39. K. Clauß, E. Lück, G.-W. von Rymon Lipinski, *Z. Lebensm. Unters. Forsch.* **162** (1976) 37 – 40.
40. G.-W. von Rymon Lipinski in D. Mayer, F. Kemper (eds.): *Acesulfame K,* Marcel Dekker, New York 1991, pp. 209 – 225.
41. C. Klug, G.-W. von Rymon Lipinski, D. Böttger, *Z. Lebensm. Unters. Forsch.* **194** (1992) 476 – 478.
42. A. Lotz, C. Klug, G.-W. von Rymon Lipinski, *Int. Z. Lebensmitteltech. Verfahrenstech.* **43** (1992) EFS 21 –EFS 23.
43. M. Korb, B. Kniel, E. Meyer, *Int. Z. Lebensmitteltech. Verfahrenstech.* **43** (1992) 494, 496, 498.
44. H.-J. Arpe in B. Guggenheim (ed.): *Health and Sugar Substitutes,* Karger, Basel 1979, pp. 178 – 183.
45. Hoechst, DE 2 628 294, 1975 (G.-W. von Rymon Lipinski, E. Lück).
46. Hoechst, DE-OS 2 264 235, 1972 (K. Clauß, H. Jensen).
47. A. Linkies, D. B. Reuschling, *Synthesis* 1990, 405 –406.
48. Hoechst, EP-A 1 55 634, 1985 (K. Clauß, A. Linkies, D. B. Reuschling).
49. *Compendium of Food Additive Specifications,* **vol. 1,** Food and Agriculture Organization of the United Nations, Rome 1992, pp. 9 – 10.
50. *Guide to Specifications for General Notices, General Methods, Identification Tests, Test Solutions and Other Reference Materials,* **FAO Food and Nutrition Paper no. 5/Rev. 2,** Food and Agriculture Organization of the United Nations, Rome 1991.
51. J. F. Lawrence, C. F. Charbonneau, *J. Assoc. Off. Anal. Chem.* **71** (1988) 934 – 937.
52. U. Hagenauer-Hener, C. Frank, U. Hener, A. Mosandl, *Dtsch. Lebensm. Rundsch.* **86** (1990) 348 –351.
53. Evaluation of Certain Food Additives and Contaminants, 27th Report of the Joint FAO/WHO Expert Committee on Food Additives, *WHO Tech. Rep. Ser.* **no. 806** (1991) 20 – 21.
54. *Fed. Reg.* **53** (1988) 28 379 – 28 387.
55. G.-W. von Rymon Lipinski, B. E. Huddart, *Chem. Ind. (London)* 1983, 427 – 433.
56. R. H. Mazur in L. D. Stegink, L. J. Filer (eds.): *Aspartame,* Marcel Dekker, New York 1984, pp. 3 – 9.
57. R. H. Mazur, J. M. Schlatter, A. H. Goldkamp, *J. Am. Chem. Soc.* **91** (1969) 2684 – 2691.
58. A. van der Heijden, L. B. P. Brussel, H. G. Peer, *Chem. Senses Flavour* **4** (1979) 141 – 152.
59. Y. Miyashita et al., *J. Med. Chem.* **29** (1986) 906 – 912.
60. J. N. Bergmann, W. Vetsch in G.-W. von Rymon Lipinski, H. Schiweck (eds.): *Handbuch Süßungsmittel,* Behr's, Hamburg 1991, pp. 425 – 444.
61. P. Langguth, R. Alder, P. Merkle, *Pharmazie* **46** (1991) 181 – 192.
62. M. N. Tsoubeli, T. P. Labuza, *J. Food Sci.* **56** (1991) 1671 – 1675.
63. W.-S. Tsang, M. A. Clarke, F. W. Parrish, *J. Agric. Food Chem.* **33** (1985) 734 – 738.
64. K. Noda, T. Iohara, Y. Hirano, H. Hayabuchi, *Kaseigaku Zasshi* **42** (1991) 691 – 695.
65. J. W. Fellows, S. W. Chang, W. H. Shazer, *J. Food Sci.* **56** (1991) 689 – 691.
66. S. E. Keller, S. S. Newberg, T. M. Krieger, W. H. Shazer, *J. Food Sci.* **56** (1991) 21 – 23.
67. L. N. Bell, T. P. Labuza, *J. Food Sci.* **56** (1991) 17 –20.
68. M. E. Brewster, T. Loftsson, J. Baldvinsdottir, N. Bodor, *Int. J. Pharm.* **75** (1991) R 5 – R 8.
69. Ajinomoto, EP-A 227 301, 1986 (T. Yukawa et al.).
70. Farmitalia, BE 900 317, 1985 (F. Dallatomasina, R. Ortica, P. Giardino, E. Oppici).

71. Monsanto, US 5 053 532, 1990 (J. B. Hill et al.).
72. Monsanto, EP-A 221 878, 1986 (J. S. Tou, B. D. Vineyard).
73. NutraSweet, WO 91 13 860, 1991 (Y. Gelman).
74. Y. Arioshi et al., *Bull. Soc. Chem. Jpn.* **46** (1973) 1893 – 1895.
75. Ajinomoto, DE-AS 2 152 111, 1970 (Y. Arioshi et al.).
76. K. Oyama, S. Irino, T. Harada, N. Hagi, *Ann. N. Y. Acad. Sci.* **434** (1984) 95 – 98.
77. *Compendium of Food Additive Specifications,* **vol. 1,** Food and Agriculture Organization of the United Nations, Rome 1992, pp. 161 – 166.
78. *Food Chemicals Codex,* 3rd ed., National Academy of Sciences, Washington 1981, pp. 28 – 29.
79. Toxicological Evalution of Certain Food Additives. 24th Report of the Joint FAO-WHO Expert Committee on Food Additives, *WHO Tech. Rep. Ser.* **653** (1980) 20 – 21.
80. *Fed. Reg.* **46** (1981) 38 284 – 38 308.
81. Toyo Soda, EP-A 255 092, 1987 (H. Wakamatsu et al.).
82. B. Homler in L. D. Stegink, L. J. Filer (eds.): *Aspartame,* Marcel Dekker, New York 1984, pp. 247 –262.
83. L. F. Audrieth, M. Sveda, *J. Org. Chem.* **9** (1944) 89 –101.
84. K. M. Beck, *Food Technol. (Chicago)* **11** (1957) 156 –158.
85. L. Kreutzig in G.-W. von Rymon Lipinski, H. Schiweck (eds.): *Handbuch Süßungsmittel,* Behr's, Hamburg 1991, pp. 413 – 424.
86. Du Pont, US 2 275 125, 1940 (L. F. Audrieth, M. Sveda).
87. Abbott, GB 662 800, 1949.
88. Abbott, GB 669 200, 1949.
89. *Compendium of Food Additive Specifications,* **vol. 1,** Food and Agriculture Organization of the United Nations, Rome 1992, pp. 283 – 285, 473 – 474, 1333 –1334.
90. M. Lehr, W. Schmidt, *Z. Lebensm. Unters. Forsch.* **192** (1991) 335 – 338.
91. B. A. Bopp, R. C. Sonders, J. W. Kesterson, *CRC Crit. Rev. Toxicol.* **16** (1986) no. 3, 213 – 306.
92. Toxicological Evaluation of Certain Food Additives, 26th Report of the Joint FAO-WHO Expert Committee on Food Additives, *WHO Tech. Rep. Ser.* **no. 683** (1982) 27 – 28.
93. L. Kreutzig in G.-W. von Rymon Lipinski, H. Schiweck (eds.): *Handbuch Süßungsmittel,* Behr's, Hamburg 1991, pp. 397 – 412.
94. H. C. Vincent et al., *J. Am. Pharm. Assoc.* **44** (1955) 442 – 446.
95. Fahlberg-List, DE 35 211, 1884.
96. J. W. Orelup, US 1 601 505, 1921.
97. Fahlberg-List, DD 237 262, 1984 (B. Müller et al.).
98. Maumee Development, US 2 667 503, 1951 (O. F. Senn).
99. *Compendium of Food Additive Specifications,* **vol. 1,** Food and Agriculture Organization of the United Nations, Rome 1992, pp. 333 – 334, 1275 – 1278, 1385 – 1387.
100. D. L. Arnold, D. Krewski, L. C. Munro, *Toxicology* **27** (1983) 179 – 256.
101. World Health Organization PSC/93.8, Summary and Conclusions of the 41st Joint FAO/WHO Expert Committee on Food Additives Meeting.
102. *Fed. Reg.* **42** (1977) 62 209 – 62 211.
103. M. K. Cook, B. H. Gominger in G. E. Inglett (ed.): *Symposium Sweeteners,* AVI, Westport 1974, pp. 211 – 215.
104. M. N. Sela, D. Steinberg in T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Appl. Sci., London 1989, pp. 71 – 96.
105. A. Bär et al., *Lebensm. Wiss. Technol.* **23** (1990) 371 –376.
106. J. F. Fisher, *J. Agric. Food Chem.* **25** (1977) 682 –683.
107. R. M. Horowitz, B. Gentili, US 3 375 242, 1968.
108. L. Krbechek et al., *J. Agric. Food Chem.* **16** (1968) 108 – 112.
109. Commission of the European Communities, *Report of the Scientific Committee for Food on Sweeteners,* Luxembourg 1987.
110. B. Crammer, R. Ikan in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Appl. Sci., London 1987, pp. 45 – 64.
111. S. S. Chang, J. M. Cook, *J. Agric. Food Chem.* **31** (1983) 409 – 412.
112. K. C. Phillips in T. H. Grenby (ed.): *Developments in Sweeteners – 3,* Elsevier Appl. Sci., London 1987, pp. 1 – 43.
113. L. Hough, P. S. Phadnis, *Nature (London)* **263** (1976) 800.
114. M. R. Jenner in T. H. Grenby (ed.): *Progress in Sweeteners,* Elsevier Appl. Sci., London 1989, pp. 121 – 141.
115. M. E. Quinlan, M. R. Jenner, *J. Food Sci.* **55** (1990) 244 – 246.
116. McNeilab, EP-A 220 907, 1986 (W. Tully, N. M. Vernon, P. A. Walsh).

117. Evaluation of Certain Food Additives and Contaminants, 37th Report of the Joint FAO/WHO Expert Committee on Food Additives, *WHO Tech. Rep. Ser.* **no. 806** (1991) 21 – 23.
118. R. B. Iyengar et al., *Eur. J. Biochem.* **96** (1979) 193 –204.
119. A. M. de Vos et al., *Proc. Natl. Acad. Sci. USA* **82** (1985) 1406 – 1409.
120. J. D. Higginbotham in C. A. M. Hough, K. J. Parker, A. J. Vlitos (eds.): *Developments in Sweeteners–1,* Applied Science, London 1979, pp. 103 – 121.
121. L. Edens et al., *Gene* **18** (1982) 1 – 12.
122. N. Overbeeke, *Biotechnol. Ser.* **13** (1989) 305 – 318.
123. Toxicological Evaluation of Certain Food Additives, 26th Report of the Joint FAO-WHO Expert Committee on Food Additives, *WHO Tech. Rep. Ser.* **no. 733** (1986) 35 – 36.
124. Pfizer, US 4 411 925, 1982 (T. M. Brennan, M. E. Hendrick).
125. Pfizer, Alitame Technical Summary 1987.
126. Université Claude Bernard, EP-A 195 731, 1986 (C. Nofre, J.-M. Tinti).
127. Université Claude Bernard, EP-B 289 430, 1988 (C. Nofre, J.-M. Tinti, F. Ouar).
128. NutraSweet, EP-A 299 533 , 1988 (R. Mazur, W. H. Owens, C. A. Klade, D. Madigan, G. W. Muller).
129. Université Claude Bernard, EP-A 321 369, 1988 (C. Nofre, J. M. Tinti, F. Ouar).
130. H. van der Wel, *FEBS Lett.* **21** (1972) 88 – 90.
131. H. van der Wel, K. Loeve, *FEBS Lett.* **29** (1973) 181 – 184.
132. DE 63 485, 1891 (J. Berlinerblau).
133. W. Kegler, *Seifen Öle Fette Wachse* **77** (1951) 606.
134. S. Petersen, E. Müller, *Chem. Ber.* **81** (1947) 31 – 38.
135. H. P. Kaufmann, D. Schweizer, *Fette Seifen* **55** (1953) 321 – 324.
136. E. M. Acton, M. A. Leaffer, S. M. Oliver, H. Stone, *J. Agric. Food Chem.* **18** (1970) 1061 – 1068.

Sympatholytics and Sympathomimetics

Sympatholytics and sympathomimetics act on the sympathetic nervous system (→ Neuropharmacology, Chap. 3.1., → Neuropharmacology, Chap. 5.3.3.). Sympatholytics are blocking (antagonist), sympathomimetics, stimulant (agonist) drugs. They are subdivided into α_1, α_2, β_1, and β_2 agents according to the receptor types that they affect.

Sympatholytics are known as adrenergic blockers since they inhibit the stimulatory (adrenergic) effects of norepinephrine (noradrenaline) and epinephrine (adrenaline) by blocking the corresponding receptors.

α-Sympatholytics are vasodilatory and they thus lead to a lowering of the blood pressure. α_1-Drugs that selectively block the α_1-adrenoceptors, such as prazosin [*19216-56-9*], indoramin [*26844-12-2*], and urapidil [*34661-75-1*], are effective in the treatment of hypertension (→ Antihypertensives, Chap. 4.2.2.).

β-Sympatholytics are a widely used class of drugs that reduce heart rate by blocking cardioaccelerator receptors, and lower blood pressure by a mechanism that is not entirely clear. The most commonly used drug of this class is propranolol [*525-66-6*], which blocks both β_1- and β_2-receptors. It is used in the treatment of coronary heart disease, hypertension, and cardiac arrhythmias and other conditions (→ Coronary Therapeutics, Chap. 4.; → Antihypertensives, Chap. 4.2.3.; → Antiarrhythmic Drugs,→ Antiarrhythmic Drugs, Chap. 2.2.). Other important β-sympatholytics are metoprolol [*37350-58-6*], nadolol [*42200-33-9*], pindolol [*13523-86-9*], timolol [*26839-75-8*]. Timolol is also used as eye drops in the treatment of glaucoma (→ Ophthalmological Preparations, Chap. 5.2.). For further information, see → β-Receptor Blocking Agents.

Sympathomimetics, also called adrenergic drugs, act on the sympathetic nerves in the same way as norepinephrine and epinephrine (→ Hormones, Chap. 2.2.). The action of α-sympathomimetics corresponds to that of norepinephrine, resulting in vasoconstriction and increased peripheral resistance and blood pressure. They include hypertensive drugs such as norfenefrine [*536-21-0*].

The stimulation of β_1-receptors leads to increased heart rate and cardiac output, whereas the stimulation of β_2-receptors causes relaxation of smooth muscle, e.g., vascular and bronchial. Sympathomimetics stimulating both α- and β- receptors are commonly employed to raise blood pressure (→ Antihypotensives, Chap. 2.) and there is limited application in the short-term treatment of heart failure (e.g., dopamine [*51-61-6*], dobutamine [*34368-04-2*]). The cardiac glycosides and synthetic cardiotonic drugs act by different mechanisms to stimulate the heart (→ Cardiac Glycosides and Synthetic Cardiotonic Drugs, Chap. 3.1.). β_2-Sympathomimetics (e.g., fenoterol [*13392-18-2*]) cause bronchodilation and are used in the treatment of asthma (→ Antiasthmatic Agents, Chap. 2.1.).

Sympathomimetics are also used as mydriatics to dilate the pupil of the eye for diagnostic purposes (→ Ophthalmological Preparations, Chap. 6.1.) and in the treatment of glaucoma (→ Ophthalmological Preparations, Chap. 5.1.).

A distinction is made between indirectly acting and centrally acting sympathomimetics. Indirectly acting sympathomimetics, e.g., stimulants such as amphetamines, ephedrine, and cocaine, release norepinephrine from sympathetic nerve endings. Centrally acting α-sympathomimetics, such as clonidine [*4205-90-7*] and methyldopa [*555-30-6*], cause stimulation of α_2-adrenoreceptors in the central nervous system which, in contrast to stimulation of reception in blood vessels, leads to a fall in blood pressure (→ Antihypertensives, Chap. 4.1.1., → Antihypertensives, Chap. 4.1.2.).

Synthesis Gas → Gas Production

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_045

Talc

WERNER TUFAR, Fachbereich Geowissenschaften der Philipps-Universität Marburg, Marburg, Federal Republic of Germany

1. **Talc** 1
1.1. **History** 1
1.2. **Structure and Mineralogy** 1
1.3. **Properties** 2
1.4. **Occurrence and Origin** 3
1.5. **Deposits** 4
1.6. **Processing and Quality Specifications** 5
1.7. **Uses** 7
1.8. **Economic Aspects** 10
1.9. **Toxicology** 10
2. **Pyrophyllite** 10
2.1. **Introduction** 10
2.2. **Structure and Mineralogy** 10
2.3. **Properties** 11
2.4. **Formation** 11
2.5. **Deposits** 12
2.6. **Processing, Quality Requirements, and Uses** 13
3. **Chlorite** 13
4. **References** 16

1. Talc

1.1. History

Talc, soapstone, and steatite have been used by humans as raw materials since prehistoric times. Molds carved from soapstone were used in the Bronze Age and early Iron Age for casting weapons and tools. In the Mediterranean cultures of the classical period, stone carvings were made from soapstone, and talc was used for treating wounds and in the production of cosmetic powder. In ancient Rome, women used large amounts of powder and rouge. The properties of talc, especially its characteristic greasy feel, were described by PLINY THE ELDER. The old Arabic word *talq*, which indicates its greasy nature, gave its name to the mineral. In 1550, CATHERINE DE MEDICI made it once again fashionable to use facial makeup in the form of powdered talc colored by the addition of pigments, a fashion that found innumerable imitators and has continued without interruption until today.

1.2. Structure and Mineralogy

Talc, an important industrial mineral, forms with pyrophyllite (Chap. 2) a separate group of silicates within the phyllosilicates, which have a mica-like sheet-silicate structure. Talc is a magnesium hydrosilicate, and pyrophyllite the corresponding aluminum hydrosilicate. The minerals are very similar crystallographically and in their properties.

Pure talc has the chemical composition $3MgO \cdot SiO_2 \cdot H_2O$, corresponding to 63.35 wt % SiO_2, 31.90 wt % MgO, and 4.75 wt % chemically bound water (hydroxide). Its mineralogical chemical formula

$${}^{2}_{\infty}Mg_3^{[6]}\ \{(OH)[Si_2O_5]\}_2 \text{ or}$$

$${}^{2}_{\infty}Mg_3^{[6]}\ [(OH)_2/Si_4O_{10}]$$

clearly represents the layer silicate structure. The talc lattice (see Fig. 1) is composed of infinite two-dimensional (${}^{2}_{\infty}$) silicate double layers ${}^{2}_{\infty}\ \{[Si_2O_5]^{2-}\}_2$ or ${}^{2}_{\infty}\ [Si_4O_{10}]^{4-}$ in which the apical O atoms of all of the SiO_4 tetrahedra within one individual layer point in the same direction, and the OH groups occupy the centers of the hexagons formed by these apical oxygen atoms. The apical O atoms of all of the SiO_4 tetrahedra within an individual layer point in the same direction, and the OH groups occupy the centers of the hexagons formed by these apical oxygen atoms. Within a double layer, the apical O atoms and OH groups of one layer are in direct contact with the corresponding atoms of the other layer, forming octahedral voids. These voids are filled by positively charged magnesium ions, which compensate for the negative charges of the silicate. One magnesium ion is coordinated octahedrally by 4 O and 2 OH, the coordination number being symbolized by $Mg^{[6]}$. The

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_047

double layer ("talc layer") so formed is therefore electrically neutral, and is only weakly bonded to the neighboring double layers by Van der Waals forces. This explains the pronounced cleavage properties of talc in the direction {001}.

Structurally, pyrophyllite is comparable to talc, which it also resembles in other respects:

$${}^{2}_{\infty}Al_2^{[6]}\ \{(OH)[Si_2O_5]\}_2 \text{ or}$$

$${}^{2}_{\infty}Al_2^{[6]}\ [(OH)_2/Si_4O_{10}]$$

Unlike talc, the two single layers in pyrophyllite are bonded by Al^{3+} instead of Mg^{2+} ("pyrophyllite layer"). All 3/3 octahedral coordination centers in talc are occupied by $Mg^{[6]}$, but in pyrophyllite only 2/3 are occupied by $Al^{[6]}$. Thus, whereas talc shows trioctahedral cation occupation, in pyrophyllite it is dioctahedral.

The pyrophyllite layer and the talc layer form the basis of the mica structure (→ Mica). In micas, within the pyrophyllite layer (muscovite series) and the talc layer (biotite series), one-fourth of the Si^{4+} ions of the SiO_4 tetrahedra are replaced by Al^{3+}. Charge balance is achieved by the 12-coordinate K^+ ($K^{[12]}$) interspersed between the double layers.

Both minerals form monoclinic crystals with space group $C_{2h}^{6}-C\,2/c$ or $C^{4}_{s}-Cc$. In the most common stacking sequence, $2M_1$, talc has the following lattice constants: $a_0 = 527$ pm, $b_0 = 912$ pm, $c_0 = 1885$ pm, $\beta = 100°\ 00'$.

The axial ratio is $a_0 : b_0 : c_0 = 0.578 : 1 : 2.067$ and $Z = 4$ (number of formula units per unit cell). The interlayer distance $d_{002} = 1/2\,c \cdot \sin\beta$ is 928 pm (9 Å layer silicate). If iron is present, these lattice constants can vary slightly.

The Mg^{2+} in talc can be partly replaced by Fe^{2+} or Ni^{2+}, giving the following talc varieties:

Minnesotaite:

$${}^{2}_{\infty}(Fe^{2+}, Mg, H_2)_3[(OH)_2/(Si, Al, Fe^{3+})_4O_{10}]$$

Willemsite:

$${}^{2}_{\infty}(Ni, Mg)_3[(OH)_2/Si_4O_{10}]$$

Idiomorphic talc formations as crystals, pseudohexagonal lamellae or platelets are rare. The mineral mainly occurs in lamellar or broad columnar, stalked, or spiky aggregates and masses. It commonly occurs in a flaky to compact, dense, tight, microcrystalline form (soapstone or steatite). Talc masses intergrown with chlorite are known as potstone.

Garnierite, $(Ni, Mg)_6[(OH)_8/Si_4O_{10}]$, which was formerly regarded as nickel serpentine or nickel antigorite, is now regarded as a mixture of 7 Å and 10 Å layer silicates, and alternately deposited mineral components. The term garnierite is now restricted to the 9 Å component, which is classed as nickel talc. Together with other nickel hydrosilicates (e.g., with the nickel chlorite schuchardtite), it occurs as an important component of the "New Caledonia" type of deposit, which commercially is a very important type of workable nickel deposit.

1.3. Properties

Talc has a greasy feel due to its very low hardness (Mohs hardness = 1) and cleaves extremely readily parallel to {001}. The density ϱ is 2.7 – 2.8 g/cm^3. The optical refraction (α, β, γ) and birefringence Δ in a thin section under the polarizing microscope are as follows:

$\alpha =$	1.539 – 1.550
$\beta =$	1.589 – 1.594
$\gamma =$	1.589 – 1.596
$\Delta =$	– (0.050 – 0.046)

The optic axial angle is $2V_\alpha = 0 - 30°$ with distinct dispersion $r > v$.

1.4. Occurrence and Origin

Talc is formed under hydrothermal conditions and is a typical mineral of weaker regional metamorphism (regional dynamothermometamorphism). It often occurs in association with chlorite, serpentine, or magnesite.

The main parent rocks that undergo metamorphic mineral reactions leading to talc formation are either (± magnesite-bearing) siliceous dolomites, or olivine- and/or pyroxene-containing ultrabasics. In ultrabasic rocks, talc is often a product of hydrothermal alteration (autometasomatism).

The formation and occurrence of talc can be illustrated by a number of metamorphic mineral reactions that can be reproduced in experimental

hydrothermal syntheses. Increasing temperature shifts the reactions towards liberation of water and CO_2.

$$\underset{\text{Dolomite}}{3\,CaMg[CO_3]_2} + \underset{\text{Quartz}}{4\,SiO_2} + H_2O \rightleftharpoons \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Calcite}}{3\,CaCO_3} + 3\,CO_2 \quad (1)$$

$$\underset{\text{Talc}}{5\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Calcite}}{6\,CaCO_3} + \underset{\text{Quartz}}{4\,SiO_2} \rightleftharpoons \underset{\text{Tremolite}}{3\,Ca_2Mg_5[(OH,F)/Si_4O_{11}]_2} + 6\,CO_2 + H_2O \quad (2)$$

$$\underset{\text{Talc}}{2\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Calcite}}{3\,CaCO_3} \rightleftharpoons \underset{\text{Tremolite}}{Ca_2Mg_5[(OH,F)/Si_4O_{11}]_2} + \underset{\text{Dolomite}}{CaMg[CO_3]_2} + CO_2 + H_2O \quad (3)$$

$$\underset{\text{Dolomite}}{2\,CaMg[CO_3]_2} + \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Quartz}}{4\,SiO_2} \rightleftharpoons \underset{\text{Tremolite}}{Ca_2Mg_5[(OH,F)/Si_4O_{11}]_2} + CO_2 \quad (4)$$

$$\underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Dolomite}}{5\,CaMg[CO_3]_2} \rightleftharpoons \underset{\text{Forsterite}}{4\,Mg_2[SiO_4]} + \underset{\text{Calcite}}{5\,CaCO_3} + 5\,CO_2 + H_2O \quad (5)$$

$$\underset{\text{Talc}}{11\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Calcite}}{10\,CaCO_3} \rightleftharpoons \underset{\text{Tremolite}}{5\,Ca_2Mg_5[(OH,F)/Si_4O_{11}]_2} + \underset{\text{Forsterite}}{4\,Mg_2[SiO_4]} + 10\,CO_2 + 6\,H_2O \quad (6)$$

$$\underset{\text{Talc}}{13\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Dolomite}}{10\,CaMg[CO_3]_2} \rightleftharpoons \underset{\text{Tremolite}}{5\,Ca_2Mg_5[(OH,F)/Si_4O_{11}]_2} + \underset{\text{Forsterite}}{12\,Mg_2[SiO_4]} + 20\,CO_2 + 8\,H_2O \quad (7)$$

$$\underset{\text{Magnesite}}{3\,MgCO_3} + \underset{\text{Quartz}}{4\,SiO_2} + H_2O \rightleftharpoons \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + 3\,CO_2 \quad (8)$$

$$\underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + H_2O \rightleftharpoons \underset{\text{Serpentine}}{Mg_3[(OH)_4/Si_2O_5]} + \underset{\text{Quartz}}{2\,SiO_2} \quad (9)$$

$$\underset{\text{Serpentine}}{2\,Mg_3[(OH)_4/Si_2O_5]} + 3\,CO_2 \rightleftharpoons \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Magnesite}}{3\,MgCO_3} + 3\,H_2O \quad (10)$$

$$\underset{\text{Serpentine}}{5\,Mg_3[(OH)_4/Si_2O_5]} \rightleftharpoons \underset{\text{Forsterite}}{6\,Mg_2[SiO_4]} + \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + 9\,H_2O \quad (11)$$

$$\underset{\text{Serpentine}}{Mg_3[(OH)_4/Si_2O_5]} + \underset{\text{Quartz}}{2\,SiO_2} \rightleftharpoons \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + H_2O \quad (12)$$

$$\underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Magnesite}}{5\,MgCO_3} \rightleftharpoons \underset{\text{Forsterite}}{4\,Mg_2[SiO_4]} + H_2O + 5\,CO_2 \quad (13)$$

$$\underset{\text{Quartz}}{4\,SiO_2} + \underset{\text{Magnesite}}{3\,MgCO_3} + H_2O \rightleftharpoons \underset{\text{Talc}}{Mg_3[(OH)_2/Si_4O_{10}]} + 3\,CO_2 \quad (14)$$

$$\underset{\text{Talc}}{9\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Forsterite}}{4\,Mg_2[SiO_4]} \rightleftharpoons \underset{\text{Anthophyllite}}{5\,(Mg,Fe)_7[(OH)/Si_4O_{11}]_2} + 4\,H_2O \quad (15)$$

$$\underset{\text{Talc}}{7\,Mg_3[(OH)_2/Si_4O_{10}]} \rightleftharpoons \underset{\text{Anthophyllite}}{3\,(Mg,Fe)_7[(OH)/Si_4O_{11}]_2} + \underset{\text{Quartz}}{4\,SiO_2} + 4\,H_2O \quad (16)$$

$$\underset{\text{Talc}}{2\,Mg_3[(OH)_2/Si_4O_{10}]} + \underset{\text{Magnesite}}{MgCO_3} \rightleftharpoons \underset{\text{Anthophyllite}}{(Mg,Fe)_7[(OH)/Si_4O_{11}]_2} + H_2O + CO_2 \quad (17)$$

Assuming that these formulas represent the chemical composition of the parent rock, talc can be formed over a wide temperature range. For example, under isobaric conditions with $X_{CO_2} = 0.3$, the equilibrium temperature for Reaction (1) is 425 °C, while that of Reaction (3) is ca. 50 °C lower (X_{CO_2} is the molar fraction of CO_2 in the gas phase, which consists of CO_2 and H_2O). The upper temperature limit for the existence of talc is set by Reaction (16), which occurs at ca. 700 °C and a water vapor pressure of 1000 bar. Reactions (4) and (7) only occur if magnesite is present with quartz and dolomite in the original rock prior to regional metamor-

phism. Reaction (15) requires the presence of CO_2.

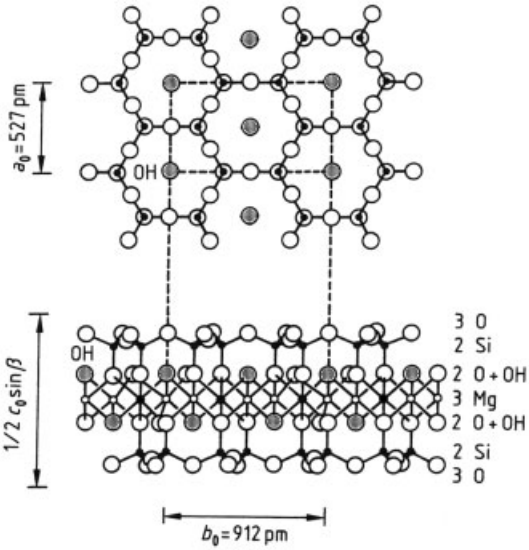

Figure 1. Structure of talc
Projection parallel to c on (001) and a plane perpendicular to the a axis; the monoclinic angle β is 100° 00′
In pyrophyllite, only the locations shown as open circles are occupied by aluminum (2 Al in place of 3 Mg)

In commercially important deposits, talc therefore occurs in association with tremolite, calcite, quartz, and dolomite, as the product of a more intense regional metamorphism of siliceous dolomites, and with forsterite and anthophyllite due to intense regional metamorphic overprint of ultrabasics. However, talc can also be observed as an authigenic new formation, e.g., in sandy sediments and salt deposits.

Reactions (1) – (17) include a number of minerals associated with talc in its deposits. Other minerals that occur in association with talc include chlorites, mica, actinolite, feldspars, rutile, pyrrhotite, pyrite, magnetite, and hematite. Limonite, a product of the weathering of iron-containing minerals, especially iron- containing ore minerals, can often be found interspersed with talc.

1.5. Deposits

Deposits of talc are widely distributed and mined worldwide in regional metamorphic complexes. The important deposits in Europe are:

Germany	Göpfersgrün – Thierstein / Upper Palatinate, Schwarzenbach on the Saale
Finland	Central Finland (Lahnaslampi) near Sotkamo, Polvijärvi and Luikonlahti near Sola), Karelia
France	Pyrenees: Département Ariège (Trimouns near Luzenac)
United Kingdom	Shetland Isles (Baltasound, Unst), Cornwall (Polyphant)
Italy	Fontane / Region Pinerolo, Val Malenco / Region Sondrio, Sardinia (Orani and Nuoro District)
Norway	Nordland (Altermark), Sognefjord (Framfjord), Oppland (Lalm, Sel, Kvam)
Austria	Eastern Styria (Rabenwald, Lassing (closed in 1998) , Weisskirchen, Kleinfeistritz, Oberfeistritz, Stubenberg, Anger)
Sweden	Handöls, Garpenberg
Spain	Provinces of Leon (Puebla de Lillo), Gerona (Pyrenees: Respina), Málaga (Fuengirola), Almeria (Purchena)

Important deposits occur in Canada (Madoc/Ontario and Van Reet/Quebec, which are mined in underground operations, and Penhorwood/Timmins, which is exploited by open-pit mining), the United States [deposits in 11 states; > 80 % of production from Texas: Allamore Talc District with Van Horn/Hudspeth County; also Montana: Yellowstone, Beaverhead, Treasure Chert; Vermont: Hammondville, Rainboro; California (trains carry talc from deposits in the south east to Nevada); New York State], Uruguay, Brazil (Paraná, Sao Paulo; Bahia: Serra das Eguas), Argentina, Australia (Three Springs, Mt. Seabrook/Western Australia; Mt. Fitton/South Australia), Japan (Hokkaido, Kyushu), the People's Republic of China (Haicheng), Taiwan, South Korea (Dongyang/Chungju region, other deposits in the Geangju, Yesan, Muju/Jinan regions), India (Rajasthan, Andhra Pradesh, Madhja Pradesh, Tamil Nadu, Bihar), and the former Soviet Union.

European talc production is ca. 10^6 t/a, of which ca. 450 000 t/a comes from Finland, 280 000 t/a from France, 130 000 t/a from Austria, 80 000 t/a from Spain, 30 000 t/a from Slovakia, and 8500 t/a from Germany.

As talc and pyrophyllite can be substituted for each other in several applications, they are usually grouped together in production and consumption statistics.

Table 1. Comparison of several grades of talc of particle size 98 % < 20 µm (data from manufacturers' information sheets)

Producer	Finnmineral	Luzenac	Norwegian Talc	Naintsch	Cyprus
Type	M 15	20 MO	A.T. 1	ST, micro	Mistron 755
Talc content, wt %	93 – 94	ca. 72	ca. 68	ca. 68	ca. 98
Main accompanying mineral	magnesite, silicate	chlorite, calcite	magnesite, magnetite	chlorite, carbonate	dolomite
SiO_2, wt %	57.0	47.8	38.9	46	62
MgO, wt %	31.3	32.0	32.9	31	31
FeO, wt %	2.3		5.9	1.5	
Fe_2O_3, wt %	0.15	1.3	0.6		1
Al_2O_3, wt %	0.7	9.4	0.9	12.5	trace
CaO, wt %	0.1		1.0	0.5	0.5
TiO_2, wt %	0.02				
CO_2, wt %	3.1		16.8	1.3	< 0.5
H_2O, chem. bound, wt %	4.85		3.0	7.2	5
Loss on ignition at 1000 °C, wt %	7 – 7.5	8.7	19.8	8.5	
Solubility in H_2O, g/100 mL	0.01		0.20		
Solubility in HCl, wt %	7 – 8	2		3	0.5
pH value	9.1	9.2	9.5	8.2	9
Specific surface area, m^2/g BET	7.5	7	5		
Oil absorption, mL/100 g	34 – 37	40	39 – 40	57	30
Brightness (Luminance A)	83 – 84	83 – 85	78	66	85 – 87
n_D	1.57	1.57	1.57		1.59
Hardness (Mohs)	1	1	1	1	1
Density, g/cm^3	2.7	2.78	2.9	2.7	2.8

1.6. Processing and Quality Specifications

Depending on the geology and minability of the deposits, talc is extracted either by open-pit mining after removal of the overburden rock layers or by underground mining. Explosives and pick axes are still used in some places, but today mining is mainly carried out mechanically with mobile cutters able to follow the talc body and produce a raw material that is already size reduced.

The processing of talc depends on the type of deposit. Commonly, the first step is selective mining, followed by flotation to separate the talc from other minerals (e.g., silicates and carbonates such as magnesite). Finally, the higher grades of talc are subjected to highly developed beneficiation processes (e.g., fine grinding, air classification).

Many common grades of talc powder are sold without grinding or sieving the mixture of minerals produced by the mine. The talc content then corresponds to the rock composition and is subject to quality variations arising both from the deposit itself and from the method of extraction (contamination with refuse rocks).

Finnish producers use flotation to treat their talc ore, which contains ca. 60 % talc and ca. 40 % associated minerals (mainly magnesite with some dolomite, and, in one deposit, a nickel ore containing ca. 0.2 % Ni). The talc concentrate contains ca. 95 % talc and 3 – 4 % carbonate, the remainder consisting of other mineral impurities.

In the mine, talc or talc-containing mineral mixtures obtained in lump form are crushed with jaw crushers. If necessary, the talc ore is further size reduced with impact or hammer crushers. Grinding is carried out with high-speed impact mills, centrifugal swing-hammer mills, roller mills, tube mills with cylinder charging (cylindrical pebbles), and jet mills.

Under optimum conditions, air classifiers can give a product with 99.7 % < 20 µm and 0.01 % > 45 µm, high-efficiency separators (so-called heavy duty air classifiers) 99 % < 10 µm,

and a jet mill with integrated screen, cyclone, and filter separation up to 98 % < 2 µm.

Grinding of talc first causes delamination of the layers, which requires only small shearing forces. However, fracture normal to the (001) plane (the plane of the layers) requires the usual energy of size reduction. The platelet-like, foliated, lamellar structure is therefore retained in the finely ground powder.

For the production of powdered talc (see Table 1) with a fineness of 98 % < 20 µm and a D_{50} value (the particle size which exceeds that of 50 % of the material and is exceeded by that of the other 50 %) of ca. 3 µm (d_{50} = 3 µm), provided that the talc contains a low level of mineral impurities and has a predominantly lamellar structure, the normal types of grinding equipment require the following amounts of electrical energy (not including the energy required for auxiliary equipment): roller mills ca. 22 kWh/t, impact mills ca. 46 kWh/t, and ball mills ca. 96 kWh/t.

Commonly used are jet mills (air and steam), which require a greater amount of energy. The technical specifications for powdered talc as a filler are laid down in ISO 3262 (see Table 2), which defines talc as "natural, platelet-shaped, lamelliform magnesium silicate hydrate". These standards apply to use in coatings, but can also be used for other areas of application. The four grades, A to D, are essentially defined by the talc content, the loss on ignition, and the content of hydrochloric acid-soluble components (carbonates, chlorites, pyrite and other sulfides, heavy metal oxides, etc.). The particle size, measured by sieve tests, is indicated by appended Arabic numerals (e.g., A1, A2, A3, B1, B2).

Table 2. Quality specifications for type A – D talcs (ISO 3262)

	A	B	C	D
Talc content, %	min. 95	min. 90	min. 70	min. 50
Loss on ingnition	4 – 6.5	4 – 9	4 – 18	4 – 27
HCl-soluble fraction, %	max. 5	max. 10	max. 30	max. 50
Sieve residue, %	1	2	3	
> 45 µm	max. 0.01	0.1	3	
> 63 µm		0.01	0.5	

The following characteristics are typical of pure talc or of talc grades low in accompanying minerals:

- Hydrophobic, organophilic boundary layers parallel to the silicate layers
- Polar and chemosorptive boundary layers (surfaces) at the edges of the platelets and on the fracture surfaces (grinding process), with unsaturated valencies
- A corrosion-inhibiting pH of ca. 8.5
- A platelet-like, lamelliform structure with correspondingly large boundary layer (surface boundary; ratio of the area of the flat surfaces of sheet silicate layers to that of the edges is 200:1)
- Barrier action against diffusion of gases and capillary water
- Morphology-dependent pseudoplasticity
- High electrical surface resistance and bulk resistance
- Low dielectric loss angle
- UV absorption
- Resistance to strong acids and alkalis

The increase in boundary layer (boundary surface) that accompanies increasing fineness of grinding accentuates the typical properties of talc. Mineral impurities lower the quality of talc or talc products, as these impurities do not generally exhibit the properties typical of talc. If the mixture of minerals has a talc content of ≤ 50 %, it is no longer regarded as talc according to DIN and ISO definitions. If the talc mixture contains ≥ 50 % chlorite, it is known as "chlorite, talc-containing" according to DIN 55 629. A product containing < 70 % talc was at one time known as asbestine.

1.7. Uses

The users of talc in the order of consumption volume are the paper industry, ceramics, plastics, paints, roofing and sealants, cosmetics and pharmaceuticals, and agrochemicals, followed by several other smaller consumers.

Some of talc's distinguishing attributes permit its wide use as a filler: smoothness, opacity, porositiy, high sheet retention, low abrasion, low yellow index score.

Paper. In the manufacturing of paper talc is utilized as a filler, for deposit control, and for coating. Paper coatings profit from talc's naturally high brightness, platy structure and its beneficial properties concerning reduced friction,

superior printability, gloss and smoothness. In resin control, the hydrophobic talc particles adhere to the resin and render it harmless. In the paper industry talc has lost some of the market share to cheaper competitors, largely to precipitated and ground calcium carbonate and, to a lesser extent, to kaolin.

Paints and Coatings. Talc is added to paints and coatings in order to produce the following effects:

- Control of the pseudoplasticity (important for processing)
- Ease of redispersion of sediments (important for coatings dispersed in water or other media)
- Improved adhesion to substrates
- Reduced diffusion through coating films (capillary water, gases)
- Synergy with inhibitor pigments (anticorrosion)
- Improved light scattering (especially in emulsion paints)
- Resistance to weathering and atmospheric influences
- Resistance to acids and alkalis
- Good dielectric properties

By combining platelet-like, flaky talc with needle-like silicates, such as amphiboles, wollastonite, xonotlite, sillimanite, and sintered mullite, interesting effects with regard to thixotrophy, close packing of particles, and elastic strengthening of the matrix can be produced.

The most important uses for talc as a filler in paints and coatings are in:

- Primer paints (based on polyvinyl butyral resin)
- Anticorrosion primer paints
- Undercoats and intermediate fillers
- Coil coatings and powder coatings
- Coatings for plastics
- Emulsion paints (especially one-coat paints), synthetic plasters, and fillers
- Epoxide – tar and polyurethane – tar coatings
- Bitumen coatings

Talc is used as a filler in a wide range of paints for domestic, industrial and protective coatings, because its particles reinforce paint, reduce sagging and give rise to a flattened coat and a smooth film. The amount of talc in paints ranges up to 30 wt %. Talc is the most common extender mineral in heavy duty applications (e.g., marine and protective coatings).

Plastics. Thermoplastics strengthened by the addition of talc to give improved tensile and transverse strength, modulus of elasticity, refractory properties, and dimensional stability are widely used as construction materials. The addition of talc has a very favorable effect on the mechanical properties and hence the serviceability of polypropylene and polyurethane (see Figs. 2 and 3).

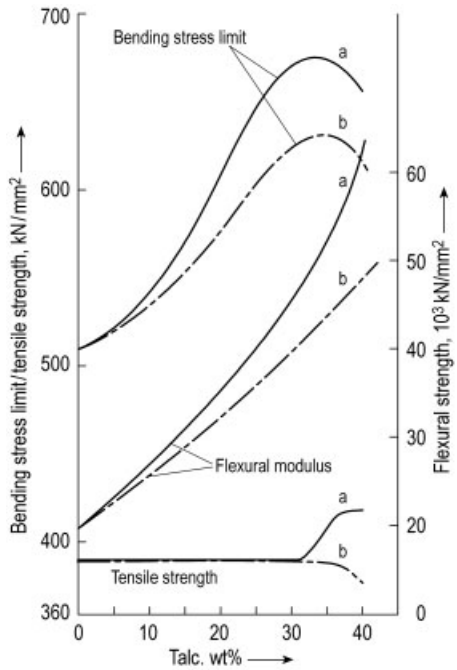

Figure 2. Mechanical properties of polypropylene reinforced with talc
a) With talc type A1 (98 % talc < 10 µm); b) With talc type C1 (68 % talc < 10 µm)

As a filler in polypropylene it improves surface properties (e.g., scratch resistance). The filler content can be up to 40 %. In the automotive industry polypropylene is used for dashboards, interior trim, bumpers, etc. Talc is also used to produce engineering thermoplastic resins (e.g., nylon compounds to replace aluminum parts in inlet manifolds).

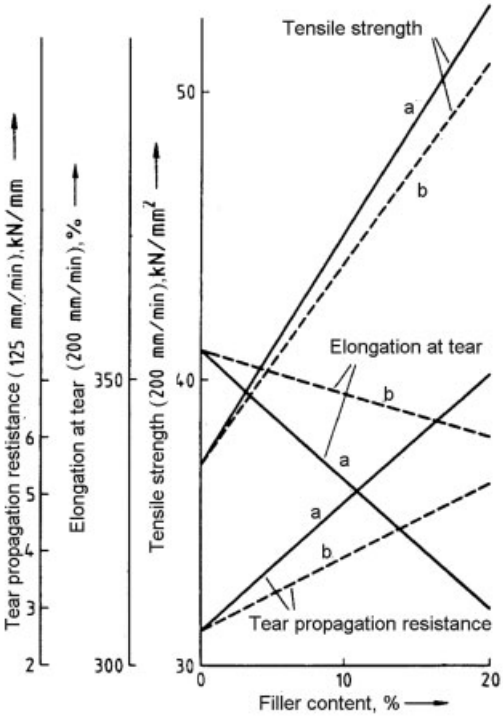

Figure 3. Mechanical properties of elastic polyurethane reinforced with talc
a) With talc type A1 (98 % talc < 10 µm); b) With talc type C1 (68 % talc < 10 µm)

Ceramics. Talc is used in traditional ceramics (e.g., 30 – 60 % in tiles, hobbyware bodies) and in technical ceramics (e.g., main silicate mineral instead of clays and feldspar in lower grade alumina products: steatite and low-expansion cordierite ceramics). In traditional ceramics, the bodies are fired for over 12 h below 1179 °C (below cone 2), and the porous bodies are then glazed and refired. For this application the talc should contain tremolite and carbonate because these minerals provide for enhanced pressing and permeability characteristics. Otherwise wollastonite must be added to purer talc to improve these properties.

Talc in the massive form known as steatite can be processed and formed directly by cutting and sawing, and in block form it can be used for thermal insulating and electroceramic products. Here, the important properties of talc are its low hardness (enabling it to be easily cut and shaped), the stability of cut products towards firing, and the very low degree of shrinkage, which means that the shaped raw material can be fired directly to porcelain.

Steatite ceramics are used in electrical insulating applications (see → Insulation, Electric, Chap. 3.1.2.2.). Steatite must be extremely pure to obtain consistency and to exclude shrinkage. For producing these low-loss bodies, ca. 10 % kaolin and 10 % barium carbonate are added to this type of talc, which is fired at 1349 – 1355 °C (cone 12 – 13).

Cordierite ceramic bodies can be produced from a mixture containing pure talc (44 %), kaolin (41 %) and alumina (15 %) or from a 1:1 mixture of kaolin and chlorite (chloritic talc). The manufacturing of these ceramics, utilized for automotive catalyst supports, is more expensive because of the higher firing temperatures.

Other Uses. For many other uses, talc is first ground. Powdered talc is used as a lubricant or polishing agent, in pastels, as a filler in the pigment industry, in cosmetics (face powder), as an additive for soaps, in paper production, as a finishing agent in the textile industry, as a filler and glazing agent in a wide range of industries, as a filler in the plastics and rubber industries, as coating for rubber products to avoid sticking and in asphalt processing.

Talc is used as an *antiblocking agent* for plastic films, for example to facilitate the opening of plastic bags such as those used for fruits and vegetables in supermarkets. For these applications a pure white talc, which is used in amounts of up to 2 kg/t, with an ultrafine size of 4 µm, is required to maintain the film color and transparency. Here talc is in competition with two cheaper rivals, namely, diatomaceous earth and synthetic silica. Sophisticated processing is required to obtain ultrafine (< 10 µm) grades of talc, and only a few producers are able to supply these.

A limited amount of talc (< 2 % of world mine production) is utilized in the *cosmetic industry*, mostly for body talcs (its pH and surface area permit it to act as a perfume carrier) but also as a filler in solid antiperspirant sticks, giving good skin slip, and in color cosmetics. Talc provides cosmetics with stability, texture, skin adhesion, slip, and water resistance. Micronized material with a grain size of < 5 µm is used in compressed products to prevent trapped air from

impairing the quality. In some applications talc is treated with hydrophobic coatings.

The amount of pure high-value talc which is consumed by the *pharmaceutical industry* is very limited. Furthermore, this talc quality requires sophisticated processing to remove all impurities (e.g., accessory minerals, carbon, iron oxide, base metal traces) and special treatments (e.g., cleaning with dilute hydrochloric acid, washing, drying, sterilization by heating to 160 °C or exposure to gamma radiation or ethylene oxide). The main applications of talc are dusting (90 – 99 % talc) and in tablets (1 – 10 % talc as a glidant or lubricant, 5 – 30 % talc as a diluent). Talc must not be used to dust surgical gloves because it can contaminate wounds.

In *agriculture*, talc satisfies all criteria (inertness, compatibility with the active ingredient, fineness, and low abrasiveness to equipment) required of a dry carrier for pesticides, fertilizers, herbicides, fungicides, and insecticides. It is used in granules and wettable powders but not in liquid formulations. In these applications, talc competes with other materials, such as fuller's earth, kaolin, diatomite, perlite, gypsum and sepiolite. The largest consumer of talc (commonly of low grade with 95 % < 325 mesh) is for pesticides.

In *animal feeds*, talc is used as an anticaking coating in feed bins and as a coating on the feed itself. Since talc is nontoxic, it is harmless and poses no danger to livestock. Its hydrophobicity prevents water loss and absorption. Furthermore, it protects the products and enables free flowing through the production and packaging plant. The amount of talc in animal feed is up to 1 – 2 % in meal and up to 0.3 – 0.5 % in pellets.

Roofing applications exploit talc (commonly < 80 mesh, passing 180 µm) as a stabilizer for the asphalt of tar paper, asphalt shingles, and roll roofing; it improves weathering and fire resistance.

A mixture of talc and chlorite with a specific size distribution, developed by Talc de Luzenac, improves the separation of waste from the water in biological *wastewater treatment* plants.

In the *rubber industry*, talc in fine grades and free of impurities is utilized mostly to lubricate molds, to prevent surfaces from sticking together, and as a mold release agent since it maintains its lubricating properties at high temperatures. As a rubber filler, talc is in carpet backing, valves, and cable insulation but is not suitable for tire manufacturing because of its low hardness and platy cleavage.

Some *minor applications* of talc are the sealant industry (e.g., joint compounds, dry wall compounds, butyl caulks, polyurethane sealants), sculpturing (e.g., steatite and soapstone for producing decorative articles), food (e.g., preparation, mostly polishing, of foodstoffs including cereals and rice, coating of chewing gum to avoid sheets sticking together) and polishing (when good color and limited abrasion is required, such as shoe, floor, and car polishing).

1.8. Economic Aspects

Worldwide, the output and consumption of talc, including pyrophyllite, exceeds 8×10^6 t/a, of which pyrophyllite accounts for ca. 2×10^6 t/a, i.e., about 25 %. In Western Europe, pyrophyllite is of minor economic importance; production and consumption of talc amount to ca. 10^6 t/a. Germany has to import most of its talc requirements. The main applications for talc in Western Europe are for paints and pigments, followed by paper, cosmetics, pharmaceuticals and rubber. The approximate percentage market shares of the major talc applications in Western Europe were:

Paper and pulp	52
Paints and varnishes	11
Roofing	7
Ceramics	6
Polymers	7
Plastics	5
Animal feedstuff	5
Fertilizers	3
Pharmaceuticals and cosmetics	3
Others	6

In Europe, the talc industry and talc production is concentrated in the hands of four companies (Talcs de Luzenac of France, Western Mining Corporation of Austria, Plüss-Staufer of Switzerland, and IMI Fabi of Italy). Finnminerals Oy of Finland, a large producer of talc, is now owned jointly by Plüss-Staufer and Western Mining Corporation.

1.9. Toxicology

Pure talc is toxicologically harmless. However, where there are high concentrations of dust in the air, face masks should be worn. If the talc contains detectable amounts of asbestos or asbestos minerals, an MAK value of 2.0 mg/m^3 applies. Talc is a nontoxic, inert substance or raw material, but it can contaminate wounds and if inhaled it can cause lung irritations.

2. Pyrophyllite

2.1. Introduction

The important industrial mineral pyrophyllite forms, with talc, a special group of phyllosilicates (layer silicates) with mica-like sheet silicate structures. Pyrophyllite is an aluminum hydrosilicate, whereas talc is the corresponding magnesium hydrosilicate. Pyrophyllite is similar to talc crystallographically and even more so in its other properties.

2.2. Structure and Mineralogy

The pyrophyllite structure (see Fig. 4) contains an infinite coherent two-dimensional silicate double layer of linked $[SiO_4]$ tetrahedra: ${}^2_\infty\{[Si_2O_5]^{2-}\}_2$ or ${}^2_\infty[Si_4O_{10}]^{4-}$, in which all apical O atoms of the SiO_4 tetrahedra within one $[SiO_5]$ layer point in the same direction, and the OH groups occupy the centers of the hexagons formed by these oxygen atoms. In the double layer, the oxygen layers formed by the O atoms and OH groups are situated directly opposite each other, and they are linked together by Al^{3+}, which is octahedrally coordinated by 4 O and 2 OH. This coordination number is symbolized by $Al^{[6]}$. The resulting silicate double layer ("pyrophyllite layer") is therefore electrically neutral and is only weakly bonded to the neighboring double layers by Van der Waals forces. This results in the laminar structure and pronounced cleavage of pyrophyllite in the direction (001). In talc, unlike pyrophyllite, the two single layers are bonded by Mg^{2+} ("talc layer") instead of Al^{3+}. Thus, in pyrophyllite only 2/3 of the octahedral coordination centers are occupied by $Al^{[6]}$, while all 3/3 in talc are occupied by $Mg^{[6]}$. Whereas pyrophyllite shows dioctahedral cation occupation, in talc it is trioctahedral, similar to the corresponding charged mica layers, e.g., in muscovite

$$K^+\{Al_2[(OH)_2/Si_3AlO_{10}]\}^-$$

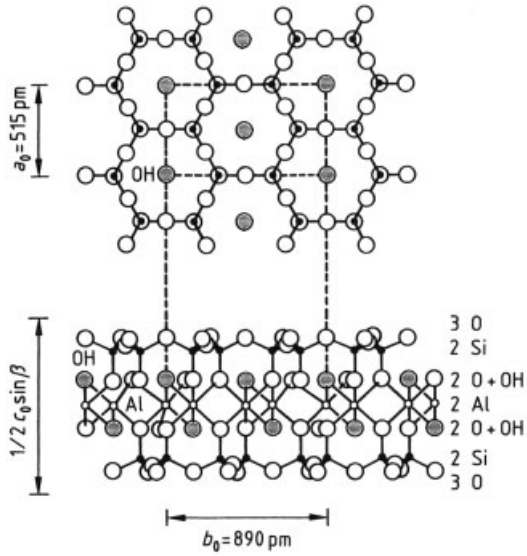

Figure 4. Structure of pyrophyllite
Projection parallel to c on (001) and a plane perpendicular to the a axis; the monoclinic angle β is 99° 55′

Pure pyrophyllite has the chemical composition $Al_2O_3 \cdot 4\,SiO_2 \cdot H_2O$, corresponding to 66.7 wt % SiO_2, 28.3 wt % Al_2O_3, and 5.0 wt % chemically bound water (hydroxide). Its mineralogical chemical formula is

$${}^2_\infty Al_2^{[6]}\{(OH)[Si_2O_5]\}_2 \text{ or}$$

$${}^2_\infty Al_2^{[6]}[(OH)_2/Si_4O_{10}]$$

Pyrophyllite forms monoclinic crystals with space group $C^6_{2h} - C\,2/c$ or $C^4{}_s - Cc$ and has the following lattice constants: $a_o = 515$ pm, $b_o = 892$ pm, $c_o = 1859$ pm, $\beta = 99°\ 55'$.

The axial ratio is $a_o : b_o : c_o = 0.577 : 1 : 2.084$ and $Z = 4$ (number of formula units per unit cell). The interlayer distance $d_{002} = 1/2\,c \cdot \sin\beta$ is 916 pm. If iron is present, these lattice constants can vary slightly.

The Al^{3+} in pyrophyllite can be partly replaced by ca. 2 atom % (average) of Fe^{3+}. The small amounts of MgO and CaO found in many

analyses can be attributed to mechanical contamination of the analytical samples. Chrome pyrophyllite, containing ca. 3 wt % Cr_2O_3, also occurs naturally.

2.3. Properties

Pseudohexagonal lamellar or tabular crystals, developed in the {001} direction, are always formed in coarse radial, random, or fan-shaped crystalline aggregates. Massive pyrophyllite is known as agalmatolite or figure stone (e.g., large deposits in Minas Gerais/Brazil). Pyrophyllite is translucent, silvery or silvery-white, white, yellowish, golden yellow, greenish, or apple green and has a nacreous luster. The lamellae are flexible and have a slightly greasy feel. Cleavage in the (001) plane is fully developed. The Mohs hardness is 11/2 and the density 2.79 g/cm^3.

Since water in pyrophyllite is in the form of hydroxide only, strong heating is needed to evolve it, leading to exfoliation and the formation of wormlike shapes. Pyrophyllite is difficult to melt. It is only slowly decomposed by sulfuric acid, but is more readily attacked by hydrofluoric acid.

In thin sections under the polarizing microscope, pyrophyllite is always colorless ($x<y\approx z$). Its optical refraction α, β, γ and birefringence Δ are

$\alpha =$	1.552 – 1.556
$\beta =$	1.558 – 1.589
$\gamma =$	1.600 – 1.601
$\Delta =$	– (0.048 – 0.045)

The optic axial angle is $2\,V_\alpha = 53-62°$ with weak dispersion $r > v$.

Depending on its origin, pyrophyllite can mainly consist either of soft, fragile aggregates or coarsely radiating lamellar crystals, very similar to lamellar talc, or of dense, coarse aggregates with an uneven, matt to lustrous fracture, similar to massive talc.

2.4. Formation

Pyrophyllite can occur in pneumatolytic formations, where it is formed by the transformation of aluminum-rich silicates (e.g., vein deposits with cassiterite or amblygonite). In hydrothermal regions, pyrophyllite can be formed from feldspar if the alkali metals can be removed by hydrothermal solutions, or from kyanite (disthene), andalusite, kaolinite, etc., provided that there is a supply of SiO_2. Suitable rocks (such as tuffs and acidic extrusive rocks, e.g., hyalorhyolite) can be (hydrothermally) altered, sometimes completely, into pyrophyllite rock or agalmatolite. The transport of pyrophyllite in hydrothermal solutions leads to its deposition in fissures, where it occurs with quartz. Pyrophyllite also occurs with quartz in nodules and lenses of regional metamorphic rocks (crystalline schists). In regional metamorphic formations, pyrophyllite can be found in the weak epizones.

Important metamorphic mineral reactions which elucidate the occurrence and distribution of pyrophyllite in certain formations include

$$\underset{\text{Lawsonite}}{CaAl_2[(OH)_2/Si_2O_7]\cdot H_2O} + \underset{\text{Quartz}}{2SiO_2} + CO_2 \rightleftharpoons \underset{\text{Calcite}}{CaCO_3} + \underset{\text{Pyrophyllite}}{Al_2[(OH)_2/Si_4O_{10}]} + H_2O$$

$$\underset{\text{Kaolinite}}{Al_4[(OH)_8/Si_4O_{10}]} + \underset{\text{Quartz}}{4SiO_2} \rightleftharpoons \underset{\text{Pyrophyllite}}{Al_2[(OH)_2/Si_4O_{10}]} + H_2O$$

The equilibrium conditions of this metamorphic reaction are (with $P_{H_2O} = P_{total}$):

1000 bar/325 ± 20 °C;
2000 bar/345 ± 10 °C;
4000 bar/375 ± 15 °C.

$$\underset{\text{Pyrophyllite}}{Al_2[(OH)_2/Si_4O_{10}]} + \underset{\text{Calcite}}{CaCO_3} \rightleftharpoons \underset{\text{Margarite}}{CaAl_2[(OH)_2/Al_2Si_2O_{10}]} + \underset{\text{Quartz}}{6SiO_2} + H_2O + CO_2$$

$$\underset{\text{Margarite}}{CaAl_2[(OH)_2/Al_2Si_2O_{10}]} + \underset{\text{Quartz}}{4SiO_2} \rightleftharpoons \underset{\text{Anorthite}}{Ca[Al_2Si_2O_8]} + \underset{\text{Pyrophyllite}}{Al_2[(OH)_2/Si_4O_{10}]}$$

$$\underset{\text{Pyrophyllite}}{2Al_2[(OH)_2/Si_4O_{10}]} + \underset{\text{Fe-rich Chlorite}}{\{Fe_2Al(OH)_6\}\cdot\{Fe_2Al[(OH)_2/Al_2Si_2O_{10}]\}} \rightleftharpoons \underset{\text{Chloritoid}}{2Fe_2AlAl_3[(OH)_4/O_2/(SiO_4)_2]} + \underset{\text{Quartz}}{6SiO_2} + 2H_2O$$

$$\underset{\text{Pyrophyllite}}{Al_2[(OH)_2/Si_4O_{10}]} \rightleftharpoons \underset{\text{Andalusite, Kyanite}}{AlAl[O/SiO_4]} + \underset{\text{Quartz}}{3SiO_2} + H_2O$$

Under certain hydrothermal conditions, pyrophyllite can be converted into diaspore (α-AlOOH).

Pyrophyllite can be synthesized from neutral or strongly acid colloidal solutions of aluminum hydroxide and SiO_2 at high temperatures (420 – 575 °C) and relatively low pressures.

2.5. Deposits

Economically important deposits are located in Hirvivaara (northern Finland), on Fukue Island off Nagasaki (Japan), and in South Korea, the United States, Canada, Brazil, Thailand, India, South Africa, Argentina, and Peru. Pyrophyllite is mainly open-pit mined. World production amounts to ca. 25 % of the total production of talc/pyrophyllite. Talc and pyrophyllite have some similar properties and are therefore grouped together in some production and consumption statistics.

Table 3. Chemical composition (in wt %), hardness, and refractive index of two grades of processed Australian pyrophyllite

	Grade 1	Grade 2
SiO_2	65.8	67.5
Al_2O_3	28.2	25.5
FeO	0.01	0.07
Fe_2O_3	0.11	0.29
TiO_2	0.18	0.18
MgO	0.01	0.03
CaO	0.28	0.26
Na_2O	0.10	0.12
K_2O	0.04	0.55
CO_2	0.05	0.10
SO_3	0.04	0.11
H_2O (chemically bound)	4.95	4.75
Hardness (Mohs)	$1-1\frac{1}{2}$	$1-1\frac{1}{2}$
Refractive index	1.58	1.58

By far the most important reserves are the Roseki deposits (“ wax stone”, the Japanese term for hydrothermally formed pyrophyllite-rich rocks with a waxy appearance) in Japan (Mitsuishi District/Okayama Prefecture, Shokozan District/Hiroshima Prefecture SW of Honshu, Goto District/Nagasaki Prefecture on Goto Island) and the Nab-Suk deposits (Korean term for deposits of similar origin to the Japanese “Roseki”) in South Korea (southeast: Wondong, Milyang, Dongnae, and in the far southwest: Wando, Kasado, Sungsan). Other deposits are found in northern Finland (Hirvivaara), in the United States (North Carolina: Robbins, Glendon, Hillsborough, and also California), in Canada (Manuels/Newfoundland), Brazil (Districts of Onca de Pitangui and Matheus Leme in Minas Gerais), India (Madhya Pradesh, Uttar Pradesh), Thailand, Turkey (Malatya Province), and Australia (Pambyula/New South Wales). In the Republic of South Africa (Ottosdal/Lichtenburg District) “ wonderstone” is extracted, which contains 90 % pyrophyllite with some epidote, chlorite, and rutile.

2.6. Processing, Quality Requirements, and Uses

Processing is similar to that for talc (see Section 1.6). As pyrophyllite has very similar properties to talc (see Table 3), its industrial uses are also similar, and the minerals can be substituted for one another. However, as pyrophyllite differs from talc of similar purity in some of its characteristic properties, it is seldom processed to a quality corresponding to Type A1 talc. Outside Europe (e.g., Japan, South Korea, United States), pyrophyllite is also used as a raw material in the refractory industry. A special application is in diamond synthesis where it is used as a lining material for the ultrahigh pressure cells. In this application it is advantageous that the melting point of pyrophyllite increases with increasing pressure.

The percentage of pyrophyllite consumed for industrial ceramics is considerably greater than that of talc.

In its use as a filler and in its properties important for this and other uses, pyrophyllite is very similar to talc. Pyrophyllite is used as a filler in coatings, plastics, adhesives, and other industrial chemical products, mainly in eastern Asia, where, in contrast to Europe and the United States, it can be produced and purchased more cheaply.

3. Chlorite

The chlorites (“14 Å chlorites”) form a very heterogeneous group and can sometimes be important rock-forming minerals. Their chemical structure (Fig. 5) shows them to be essentially hydrated Mg – Fe – Al silicates that are free of

alkali metals and calcium and are built up of two kinds of oppositely charged layer ions.

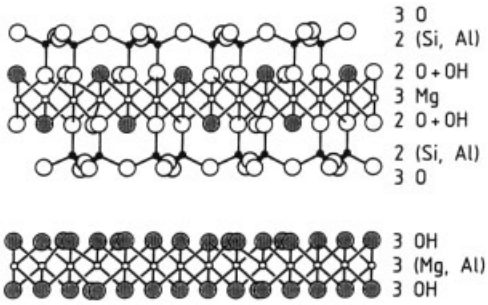

Figure 5. Idealized structure of chlorite
Projection on a plane perpendicular to the *a* axis

Formerly, low-iron orthochlorites were regarded as "true chlorites" and were distinguished from iron-rich leptochlorites. However, orthochlorites are now considered to be those in which $Fe_2O_3 < 4\,\%$, and leptochlorites those in which $Fe_2O_3 > 4\,\%$.

In the chlorites, trioctahedral phyllosilicates (layer silicates) are present, consisting of negatively charged mica layers (→ Mica; → Silicates, Chap. 1.1., → Silicates, → Silicates) alternating with positively charged brucite-like layers. All octahedral sites are occupied by cations (Mg, Al, Fe) forming trioctahedral layer silicates.

As in the other phyllosilicates, the Si^{4+} in chlorites is partly replaced by Al^{3+} in the tetrahedral positions of the mica layer. Charge balance is achieved by the replacement of divalent cations (Mg^{2+}, Fe^{2+}, Mn^{2+}, Cr^{2+}, Ni^{2+}) by trivalent cations (Al^{3+}, Fe^{3+}) in the hydroxide layer. This results in a large number of possible chemical variants, and further variations can result from the replacement of Mg^{2+}, mainly by Fe^{2+}, in the octahedral positions of both the talc layer and the brucite layer.

In contrast, the sudoite series of minerals, which have only been known since the early 1960s, are "dioctahedral chlorites". They consist of alternating muscovite and (charged) brucite layers. Charge compensation, necessary because of the partial substitution of Si by Al in the $[SiO_4]$ tetrahedra, results from the introduction of trivalent in place of divalent cations in the brucite-like layer as well as by over-occupation of the octahedral interstices.

With regard to the six octahedrally coordinated cations, the minerals of the chlorite series can also be regarded as mixed crystals of four end members, although these are hypothetical, and do not occur in nature in the pure state:

Fe_6 Chlorite $\left\{\begin{array}{l} Fe_3^{2+}[(OH)_2/Si_3AlO_{10}]^- \\ Fe_2^{2+}Fe^{3+}(OH)_6^+ \end{array}\right\}$

Mg_5Al Chlorite $\left\{\begin{array}{l} Mg_3[(OH)_2/Si_3AlO_{10}]^- \\ Mg_2Al(OH)_6^+ \end{array}\right\}$
chemical composition comparable to antigorite

$Fe_4^{2+}Al_2$ Chlorite $\left\{\begin{array}{l} Fe_2^{2+}Al[(OH)_2/Al_2Si_2O_{10}]^- \\ Fe_2^{2+}Al(OH)_6^+ \end{array}\right\}$
daphnite

Mg_4Al_2 Chlorite $\left\{\begin{array}{l} Mg_2Al[(OH)_2/Al_2Si_2O_{10}]^- \\ Mg_2Al(OH)_6^+ \end{array}\right\}$
chemical composition comparable to amesite

However, the two magnesium-rich minerals antigorite and amesite, sometimes described as end members, do belong to the serpentine minerals and not to the chlorites. The conversion from the chlorite structure to the serpentine structure only requires the apical O atoms of half of the T_2O_5 (T = Si, Al) layers to be shifted to the other side. If the transformation is only partial, the poorly crystallized 7 Å chlorites or septechlorites result.

The minerals of the chlorite series form monoclinic crystals, usually with the space group $C_{2h}{}^3 - C\,2/m$, $Z = 4$ (number of formula units per unit cell). However, the number of possible stacking variations is much greater than in the case of micas.

Much simplified, chlorite has the following lattice constants: $a_o \approx 530$ pm (5.3 Å), $b_o \approx 930$ pm (9.3 Å), $c_o \approx 2860$ pm (28.6 Å). The axial ratio is $a : b : c \approx 0.57 : 1 : 3.09$, $\beta \approx 97°$. The layer thickness is $d_{002} = 1420$ pm (14 Å layer silicate).

Of the trioctahedral chlorites (chlorites in the narrower sense), several groups ("series") can be distinguished.

The talc chlorites include:

Talcchlorite (hypothetical end member) $\left\{\begin{matrix} Mg_3[(OH)_2/Si_4O_{10}] \\ Mg_3(OH)_6 \end{matrix}\right\}$ (same chemical composition as serpentine, $Mg_3[Si_2O_5/(OH)_4]$, which is thermodynamically more stable)

Penninite $\left\{\begin{matrix} (Mg, Al)_3[(OH)_2/Al_{0.5-0.9}Si_{3.5-3.1}O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Clinochlore $\left\{\begin{matrix} (Mg, Al)_3[(OH)_2/AlSi_3O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Sheridanite $\left\{\begin{matrix} (Mg, Al)_3[(OH)_2/Al_{1.2-1.5}Si_{2.8-2.5}O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Rhipidolite $\left\{\begin{matrix} (Mg, Fe, Al)_3[(OH)_2/Al_{1.2-1.5}Si_{2.8-2.5}O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Corundophilite $\left\{\begin{matrix} (Mg, Fe, Al)_3[(OH)_2/Al_{1.5-2}Si_{2.5-2}O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

The ferro chlorites are formed from

Brunsvigite $\left\{\begin{matrix} (Fe^{2+}, Mg, Al)_3[(OH)_2/AlSi_3O_{10}] \\ (Fe, Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Daphnite $\left\{\begin{matrix} (Fe^{2+}, Al)_3[(OH)_2/Al_{1.2-1.5}Si_{2.8-2.5}O_{10}] \\ (Fe, Al)_3(OH)_6 \end{matrix}\right\}$

Pseudo-thuringite $\left\{\begin{matrix} (Fe^{2+}, Al, Mg)_3[(OH)_2/Al_{1.5-2}Si_{2.5-2}O_{10}] \\ (Fe, Mg, Al)_3(OH)_6 \end{matrix}\right\}$

The ferro-ferri chlorites include the following: (oxychlorites with partial dehydrogenation: $Fe^{2+} + OH^- \longrightarrow Fe^{3+} + O^{2-} + 1/2\,H_2$)

De-lessite $\left\{\begin{matrix} (Mg, Fe^{2+}, Fe^{3+})_3[(O, OH)_2/Al_{0-0.9}Si_{4-3.1}O_{10}] \\ (Mg, Fe^{2+})_3(OH)_6 \end{matrix}\right\}$

Chamosite $\left\{\begin{matrix} (Fe^{2+}, Fe^{3+})_3[(O, OH)_2/AlSi_3O_{10}] \\ (Fe, Mg)_3(OH)_6 \end{matrix}\right\}$

Thu-ringite $\left\{\begin{matrix} (Fe^{2+}, Fe^{3+}, Al)_3[(O, OH)_2/Al_{1.2-2}Si_{2.8-2}O_{10}] \\ (Mg, Fe, Fe^{3+})_3(OH)_6 \end{matrix}\right\}$

The following can also be mentioned here:

Pennantite $\left\{\begin{matrix} (Mn, Al, Fe^{3+})_3[(OH)_2/(Al, Si)Si_3O_{10}] \\ (Mn, Al)_3(OH)_6 \end{matrix}\right\}$

Gonyerite $\left\{\begin{matrix} (Mn, Mg, Fe^{3+})_3[(OH)_2/(Si, Fe^{3+})Si_3O_{10}] \\ (Mn, Mg, Fe^{3+})_3(O, OH)_6 \end{matrix}\right\}$

Chromium chlorites include:

Kämmererite $\left\{\begin{matrix} (Mg, Cr)_{<3}[(OH)_2/AlSi_3O_{10}] \\ (Mg, Cr)_3(OH)_6 \end{matrix}\right\}$

Kochubeite $\left\{\begin{matrix} (Mg, Al)_{<3}[(OH)_2/(Cr, Al)Si_3O_{10}] \\ (Mg, Al)_3(OH)_6 \end{matrix}\right\}$

Nickel chlorites also occur: schuchardtite is a nickel penninite which contains less nickel than magnesium, whereas in nimite, nickel predominates.

Crystals of chlorite are only occasionally found, and are usually lamellar and in flake or massive form. Well-developed crystals usually have a hexagonal, rhombohedral, or pseudohexagonal tabular crystal habit.

The Mohs hardness is ca. 2, and the density (depending on iron content) ca. 2.6 – 3.3 g/cm^3. Unlike the micas, lamellae of chlorite are not elastically flexible.

Owing to the lamellar twinning commonly present, the crystals appear optically uniaxial. There are several twinning "laws", e.g., the penninite law [twinning plane (001)] or the mica law (twinning plane perpendicular to (001) and parallel to [110]). Also, twinning occurs even in the lattice, leading to polytypes, which have different symmetries and multiple c values (variations in stacking sequence).

Depending on chemical composition, the optical character of the chlorites is either biaxially positive or biaxially negative, and is sometimes different for different wavelengths. The plane of the optical axis is usually (010), and the acute bisectrix is approximately normal to (001).

Chlorites usually have a green color, from which the name is derived. As the iron content increases, the color becomes more intense, increasing from colorless to dark green or brownish green (on oxidation of the iron). Other colors include pink to violet-pink (chromium chlorites) and orange-red (higher Mn content and lower Fe content).

The optic axial angle can vary greatly ($-20°$ to $+70°$), e.g., for clinochlore and grochauite (iron-containing sheridanite) $2V_\gamma = 0 - 70°$, and for daphnite, bavalite, and thuringite $2V_\alpha = 0 - 30°$. However, optic axial angles are usually very small, converging to 0°.

The iron content has a considerable influence on optical refraction. Increasing iron content leads to an increase in optical refraction (α,

β, γ), while the birefringence Δ is usually low. Pleochroism is usually strongly developed.

Mg chlorites:		Penninite	Clinochlore	Sheridanite
	α =	1.562	1.581	1.594
	β =	1.565	1.581	1.594
	γ =	1.565	1.586	1.606
	Δ =	−0.003	+0.005	+0.012
Chromium chlorites (Kämmererite, Kochubeite)				
	α =	1.576	1.584	1.585
	β =	1.576	1.584	1.590
	γ =	1.580	1.584	1.591
	Δ =	+0.004	+0.000	−0.006
Mg – Fe^{2+} chlorites:				
		Mg diabantite	Fe rhipidolite	Corundophilite
	α =	1.589	1.627	1.605
	β =	1.595	1.627	1.608
	γ =	1.595	1.629	1.615
	Δ =	−0.006	+0.002	+0.010
Ferro chlorites		Brunsvigite	Aphrosiderite	Daphnite
	α =	1.632	1.647	1.665
	β =	1.638	1.651	1.675
	γ =	1.638	1.651	1.675
	Δ =	−0.006	−0.004	−0.010
Ferro-ferri chlorites		Delessite	Thuringite	
	α =	1.595	1.671	
	β =	1.599	1.684	
	γ =	1.599	1.685	
	Δ =	−0.004	−0.014	

Anomalous interference colors can often occur. Optically positive chlorites with low iron content and low refractive index show typical anomalous brown (“ leather brown”) interference colors, whereas optically negative chlorites with higher iron content and higher refractive index show anomalous violet to blue (“ duck blue” or “ mallard blue”) interference colors.

If chlorites are heated, they liberate water in two steps, but their structure remains unchanged almost up to complete dehydration. With magnesium-rich chlorites, water is first liberated from the brucite layers, and olivine is formed in the second dehydration step. Further temperature increase leads to the formation of spinel and enstatite.

As a rock-forming mineral, chlorite occurs mainly in low-grade regional metamorphic rocks or epizonal crystalline schists, i.e., in the metamorphites of the greenschist facies (e.g., chlorite schists). Low-iron chlorite often occurs in association with talc. Also, chlorite is widely distributed as a hydrothermal druse mineral. Chlorite is often formed by hydrothermal decomposition and alteration (“chloritization”) of various (magnesium-rich) silicates, e.g., from biotite, garnet, hornblende, and augite; it also displaces and causes pseudomorphism of these minerals. Chlorites are also formed by the transformation of ultrabasic rocks (mainly olivine-rich rocks) into serpentinites (serpentinization). Chlorite can also be formed by contact metamorphism (e.g., of rocks of the albite – epidote – hornfels facies). Clay sediments often contain chlorite, usually of diagenetic origin.

The iron-rich chlorites, e.g., chamosite (in the Jura mountains from Chamoson in Wallis/Switzerland) and thuringite (Thuringia/Federal Republic of Germany), which practically correspond to the older leptochlorites, can occur in certain oolitic iron deposits, and can even be the predominant mineral in some places. These formations are due to sedimentary deposition, but now exist in a slightly metamorphosed form.

The nickel chlorite schuchardtite occurs near other nickel hydrosilicates (e.g., with garnierite, hitherto classified as nickel serpentine or nickel antigorite, with the mineralogical chemical formula $(Ni, Mg)_6[(OH)_8/Si_4O_{10}]$, more recently also regarded as mixtures, so that garnierite is then restricted to the 10 Å component as nickel talc) in the “New Caledonia” type nickel deposits, which are of major economic importance.

The chromium chlorites, which are characterized by their pink to peach-blossom or violet-pink colors, occur in chromite deposits. They were formed by the serpentinization of the chromite- containing ultrabasic parent rocks.

Apart from its use as an ore [an important iron ore in former times, and a nickel ore today (→ Nickel)], chlorite is at present only of minor importance as an industrial mineral (e.g., as a filler, pigment, or pigment matrix) in spite of its wide distribution and frequent occurrence in lamellar form. Its potential as a replacement for other minerals (e.g., talc, mica), which have

some similar properties is somewhat limited at present. Chlorite (penninite) is sometimes used in the wallpaper industry. Leuchtenbergite, an almost iron-free variety of clinochlore, together with sericite and quartz, is the most important component of "white schist", a metamorphic rock of which there are deposits, for example, at Aspang (Ausschlag-Zöbern in Lower Austria). Although white schist is known as "Aspang kaolin", it contains no kaolin and is mainly of interest for its sericite content.

Chlorite and talc have similar properties. Thus, chlorite can substitute talc, e.g., in some ceramic applications. Chlorite is often found in talc deposits.

More recently, amixture of talc and chlorite with a specific size distribution, developed by Talc de Lucerne, improves the separation of waste material from the process water, and is used in biological wastewater treatment where it is added to the biological pool.

4. References

1. H. U. Bambauer, F. Taborszky, H. D. Trochim, W. E. Tröger: *Optische Bestimmung der gesteinsbildenden Minerale,* part 1: Bestimmungstabellen, 4th ed., E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1971.
2. H.-R. Bosse et al.: *Untersuchungen über Angebot und Nachfrage mineralischer Rohstoffe: XIX Industrieminerale,* Bundesanstalt für Geowissenschaften und Rohstoffe Hannover – Prognos AG Europäisches Zentrum für Angewandte Wirtschaftsforschung Basel, Hannover– Basel 1986.
3. O. Braitsch, W. E. Tröger: *Optische Bestimmung der gesteinsbildenden Minerale,* part 2, E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart 1967.
4. W. A. Deer, R. A. Howie, J. Zussman: *Rock-Forming Minerals,* **vol. 3: Sheet Silicates,** Longmans, Green and Co. Ltd., London 1962.
5. W. A. Deer, R. A. Howie, J. Zussman: *An Introduction to the Rock-Forming Minerals,* 2nd ed., Longman Scientific and Technical, London 1974.
6. *Ind. Miner. (London)* (1977) no. 6, 19 – 27.
7. H. Kittel: *Lehrbuch der Lacke und Beschichtungen,* **vol. 2,** W. A. Colomb within Heenemann GmbH, Berlin 1974, p. 379.
8. O. Lückert: *Pigment + Füllstoff Tabellen,* M. and O. Lükkert, Laatzen 1980.
9. F. Machatschki: *Spezielle Mineralogie auf geochemischer Grundlage,* Springer-Verlag, Wien 1953.
10. P. Ramdohr, H. Strunz: *Klockmann's Lehrbuch der Mineralogie,* 16th ed., Enke-Verlag, Stuttgart 1978.
11. H. J. Rösler: *Lehrbuch der Mineralogie,* 3rd ed., VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1984.
12. H. Strunz: *Mineralogische Tabellen,* 7th ed., Akademische Verlagsgesellschaft Geest & Portig KG, Leipzig 1977.
13. H. G. F. Winkler: *Petrogenesis of Metamorphic Rocks,* 4th ed., Springer-Verlag, New York–Heidelberg–Berlin 1976.
14. W. Tufar: "Anthophyllit und Talk von Vorau (Oststeiermark)," *Mitteilungsbl. Abt. Mineral. Landesmus. Joanneum* **47** (1979) 37 – 52, 181 – 196.
15. W. Tufar: "Die Vererzung der Ostalpen und Vergleiche mit Typlokalitäten anderer Orogengebiete," *Mitt. österr. Geol. Ges.* **74/75** (1981/82) 265 – 306.
16. W. Tufar: "Die geologischen Grundlagen für den Bergbau in Niederösterreich," in H. Feigel (ed.): *Studien und Forschungen aus dem Niederösterreichischen Institut für Landeskunde,* vol. 10, A. Kusternig (ed.): Bergbau in Niederösterreich, Amt der Niederösterreichischen Landesregierung, Vienna 1987, pp. 1 – 53.
17. W. Tufar et al.: "Formation of Magnesite in the Radenthein (Carinthia/Austria) Type Locality," *Monogr. Ser. Miner. Deposits* **28** (1989) 135 – 171.
18. L. Weber, G. Zsak: *Welt-Bergbau-Data,* **Series A, vol. 11**, Bundesministerium für wirtschaftliche Angelegenheiten, Vienna, 1996.
19. N. C. Trivedi: "Talcs," in R. W. Hagemeyer (ed.): *Pigments for Paper* , Tappi Press, Atlanta 1997, pp. 199 – 214.
20. "Talc in Europe – Optimistic Views are Prevailing," *Ind. Miner. (London)* **1987**, no. 6, 40 – 51.
21. "Talc Review – Consolidation and Competition," *Ind. Miner. (London)* **1992**, no. 3, 23 – 37.

22. "Talc for Thermoplastics – Future Demands the Challenge," *Ind. Miner. (London)* **1995**, no. 5, 47 – 53.
23. "Talc Markets – A World of Regional Diversity," *Ind. Miner. (London)* **1997**, no. 5, 39 – 53.
24. "Minerals in European Plastics – Polypropylene in the Driving Seat," *Ind. Miner. (London)* **1993**, no. 4, 19 – 35.
25. "Fillers and Extenders – European Market Trends," *Ind. Miner. (London)* **1997**, no. 12, 29 – 35.
26. "Plastic Fillers – A Maturing Market in Europe," *Ind. Miner. (London)* **1997**, no. 5, 61 – 63.

Tall Oil

LARS-HUGO NORLIN, Bergvik Kemi, Sandarne, Sweden

1. Introduction 1
2. Crude Tall Oil as Byproduct and Raw Material 2
2.1. Composition and Properties of Crude Tall Oil 2
2.2. Recovery of Soap and Conversion to Crude Tall Oil 4
2.3. Upgrading of Soap Extraction (Crude Soap Refining Process) . . . 5
3. Upgrading of Crude Tall Oil by Distillation 5
3.1. Technical and Economic Conditions 7
3.2. Process Description 8
3.2.1. Dehydration 8
3.2.2. Depitching 8
3.2.3. Fractionation of Vaporized Tall Oil Compounds 9
4. Energy Recovery 10
5. Security and Construction Material 10
6. Analysis 10
7. Storage and Transportation 12
8. Uses 12
9. Economic Aspects 13
10. Toxicology and Environmental Protection 13
11. References 14

1. Introduction

"Tall" is the Swedish word for pine, and the term tall oil is an American adaptation of "tallolja," a Swedish product name associated with the kraft pulping process (→ Paper and Pulp, Chap. 1.3.1.1.) and with products from the fractional distillation of crude tall oil.

Pine wood contains an extractable fraction, ≈ 3 %, consisting of free and combined fatty and rosin acids and nonacidic compounds (tall oil components). In the alkaline pulping process the tall oil components are saponified into a soapy material that, when acidulated, gives a mixture of free rosin and fatty acids and neutral components – crude tall oil. Crude tall oil is refined mainly by vacuum distillation processes to separate the various compounds almost completely into rosin and fatty acid fractions.

The *rosin* part of crude tall oil is a complex mixture of rosin acids, a three-ring fused system with the empirical formula $C_{20}H_{30}O_2$ (→ Resins, Natural, Chap. 5.2.3., → Resins, Natural, Chap. 5.4.). The tall oil *fatty acids*, primarily C_{18} types, are similar in composition to those derived from soybean oil. The *neutral constituents* of tall oil – predominantly esters, sterols, and hydrocarbons – are generally regarded as troublesome "impurities" that impair its utilization and distillation.

Abbreviations of basic tall oil names and common analyses are often used in publications and reports.

Crude tall oil (CTO) [*8002-26-4*]
Tall oil fatty acids (TOFA) [*61790-12-3*]
Tall oil rosin (TOR) [*8052-10-6*]
Distilled tall oil (DTO) [*8002-26-4*]
Tall oil pitch (TOP) [*8016-81-7*]
Heads, light ends (H) [*65997-03-7*]

2. Crude Tall Oil as Byproduct and Raw Material

The kraft pulping process was developed late in the 19th century. Some years later it was discovered in Sweden that a product similar to oil or wood tar could be obtained when the soapy material was acidulated with sulfuric acid. The recovery of byproducts from the kraft pulping process is shown in Figure 1 (see next page).

Tall Oil Precursor in Pine Wood. Exploitation of the original tall oil materials in pine wood in the form of products such as tar and pitch dates from centuries B.C. Gum and wood rosin are still produced, and compete with the rosin fraction, TOR, derived from CTO.

Products such as gum rosin (residue product, → Resins, Natural, Chap. 5.2.1.), and turpentine

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_057

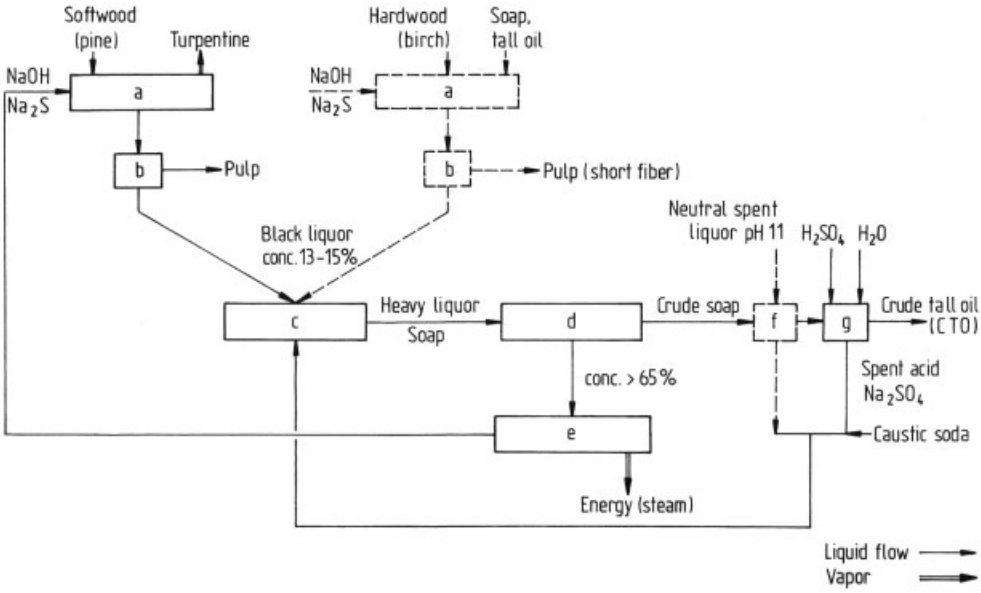

Figure 1. Recovery of byproducts from the kraft pulping process
a) Digester; b) Filter; c) Evaporator; d) Settling tank; e) Chemical recovery, energy production; f) Washing; g) Reactor

oil fractions (→ Turpentines) (distillates) obtained by distillation of exuded oleoresin from wounded living trees are still of high volume and importance. However, gum rosin production in the United States has declined drastically from ca. 300 000 t/a in 1930 to almost zero today owing to the lack of affordable labor. Despite this fact, worldwide gum rosin production in the early 1990s has been reported to be as high as 700 000 t/a. The People's Republic of China alone reports 400 000 t/a. Extraction of products such as wood rosin and turpentine fractions from stumps rich in rosin and liquid gum has declined because of the shortage of stumps (→ Resins, Natural, Chap. 5.2.2.).

Replacement of Original Products by Tall Oil. The production of CTO as a byproduct of the kraft pulping process and the subsequent refining of CTO have supplied important and renewable resources such as tall oil rosins and tall oil fatty acids as two basic chemical raw materials. The decreasing supply of wood and gum rosin in the United States is well compensated by the TOR production, which has increased proportionally to pulp production (see Fig. 2).

Tall Oil in Wood. The extractable portion of wood differs chemically, in part, from the compounds found in the CTO product. The chemical changes occurring during various processing steps are shown in Figure 3. Examples of composition of the extract from different wood species are listed in Table 1.

Table 1. Composition of the extract from different wood species

Compounds	Pine	Spruce	Birch
Esters, %	40	43	58
Free fatty acids, %	18	10	6
Free rosin acids, %	34	25	0
Total, %	92	78	64

2.1. Composition and Properties of Crude Tall Oil

The wide range of compositions and properties of crude tall oil on the market can be explained and understood from the chemistry of tall oil. Many tall oil components are sensitive to chemical reactions and changes in properties and composition. The use of wood species other than pine [e.g., spruce and hard wood (birch)] in the kraft pulping process lowers the yield of CTO and the selection of available fatty and rosin acids. It also changes the composition and properties of the CTO. The quality and yield of CTO from

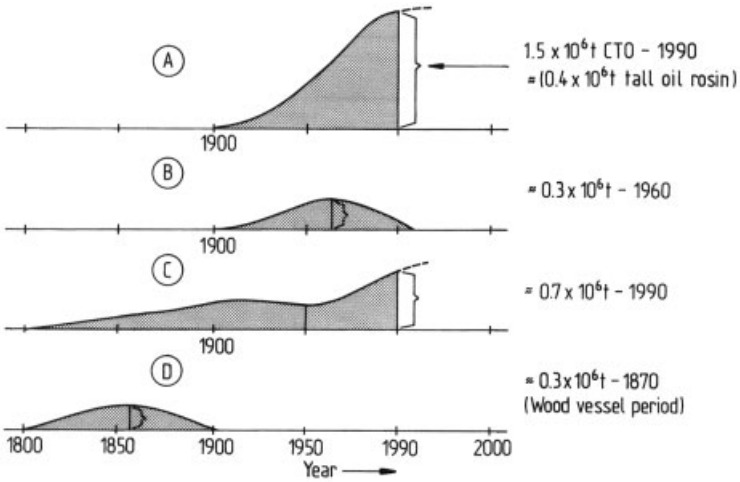

Figure 2. Historical development of global production of tall oil materials
A) CTO; B) Wood rosin; C) Gum rosin; D) Tar

the kraft pulping process consequently depend on many variables such as type and length of wood storage, season of the year, amount of wood residue used, hardwood usage, and type of pulping employed.

Table 2. Primary constituents of tall oil in different wood species

	Extractives, %	Fatty acids, kg/t wood	Rosin acids, kg/t wood	Unsap ∗, kg/t wood
Pine	≈ 3	15 – 20	5 – 10	2 – 5
Spruce	≈ 1	4 – 8	1 – 2	1 – 3
Birch	≈ 2	12 – 18		5 – 10

∗ Unsaponifiables.

Crude tall oil is a dark brown, tarlike product; its odor, influenced by sulfur components from pulp processing and soap skimming/acidulation, is unpleasant and penetrating. The three main constituent groups – fatty acids, rosin acids, and neutral materials (unsaps) – vary widely in composition, depending on the wood's origin (see Table 2). The saponification number (SN) of CTO is always higher than the acid number (AN) because of the unsplit esters in the crude oil. The difference increases for lower qualities, indicating less available rosin and fatty acids.

The crude is normally transported and stored at 50 – 80 °C to avoid crystallization and settling of rosin acids. The viscosity of the crude is variable; it is mainly a function of the rosin content. Viscosity must be measured on a clear, noncrystallized product. Impurities such as lignin, soap residue, water – oil emulsions, sodium salts, and sulfuric acid occur in the product. The amount of these undesirable materials depends on the efficiency of the soap washing and splitting process. Sour components such as sulfuric acid in the separated brine and degassed sour vapors from heated CTO in storage tanks impart the crude corrosive properties. The physical properties of tall oil (empirical values) are listed below:

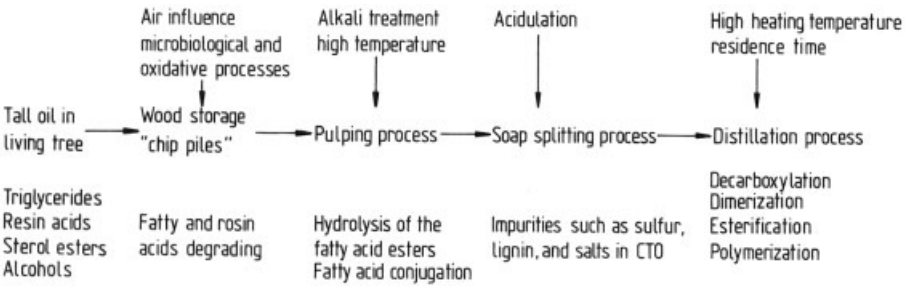

Figure 3. Chemical changes in tall oil during processing

Table 3. Typical data and composition of crude tall oil

Characteristic		Scandinavia, typical mix	United States, pine	Canada, mixed	France
Acid number		145	165	140	165
Saponification number		160	172	165	172
Fatty acids, %		45	45	42	40
Palmitic	(16:0)*	1.5	3	2	2
Stearic	(18:0)*	0.5	1	1	1
Oleic	(18:1)*	10	20	10	15
Linoleic	(18:2)*	17	13	15	11
Pinolenic	(18:3)*	5	1	3	1
Arachidic	(20:0)*	1	0.5	1	0.5
Other		10	6.5	10	9.5
Rosin acids, %		30	42	30	50
Pimaric		2	3	2	5
Palustric		4	7	5	7
Isopimaric		2	4	4	4
Abietic		12	15	10	18
Dehydroabietic		5	4	4	5
Neoabietic		4	6	3	6
Others		1	3	2	5
Neutrals, %		25	13	28	10
Moisture				$< 2\,\%$	
Ash				$< 0.2\,\%$	
Mineral acid (H_2SO_4)				$< 0.02\,\%$	
Sulfur				$< 0.3\,\%$	

Boil-up temperature (at 1.33 kPa)	180 – 270 °C
Heat of vaporization	290 – 330 kJ/kg
Specific heat	2.1 – 2.9 J/g
Heat of combustion	−(33 000 – 38 000) kJ/kg
Density (at 20 °C)	950 – 1020 kg/m^3
Liquid viscosity (at 70 °C)	25 – 40 mm^2/s

Typical data and composition of crude tall oil are presented in Table 3.

2.2. Recovery of Soap and Conversion to Crude Tall Oil

During the kraft pulping process, CTO precursors are hydrolyzed to soluble sodium soaps in the cooking liquor. The soap and spent cooking liquor (black liquor) are washed and filtered from the pulp and concentrated in multipleeffect evaporators from 13 – 15 % to > 65 % solids. Evaporation increases the solids content so much that the soap – contaminated with black liquor to a greater or lesser extent – precipitates and rises continuously to the surface of the upgraded liquors in skimming tanks connected to the evaporation process. When the soap is skimmed from the top of the black liquor tanks, contamination of the soap with some black liquor is unavoidable. The major part of this black liquor containing about 10 % lignin should be removed. Otherwise, precipitation of the lignin and gypsum (gypsum is formed because of the presence of calcium in the black liquor) will occur during subsequent acidulation. Good decantation or separation of the soap is thus important. A low black liquor content of the soap decreases acid consumption and the amount of toxic hydrogen sulfide and carbon dioxide released during acidulation. In addition, soap present in black liquor and in the heavy black liquor (i.e., concentrated black liquor) decreases the capacity of both the evaporator system and the recovery boiler.

The *splitting* or *acidulation process* is a chemical reaction between the sodium soaps and sulfuric acid. The tall oil soap is converted into free fatty and rosin acids by reaction with dilute sulfuric acid (ca. 30 %) at elevated temperature (ca. 98 °C). The reaction can be summarized as follows:

$$2\,RCOONa + H_2SO_4 \longrightarrow Na_2SO_4 + 2\,RCOOH$$

where R denotes a rosin or fatty acid alkyl group.

Three phases are formed during the reaction: liberated tall oil, a lignin – fiber phase, and the spent acid phase (Na_2SO_4, free H_2SO_4). The phases can be separated because of their different densities. The splitting process is performed

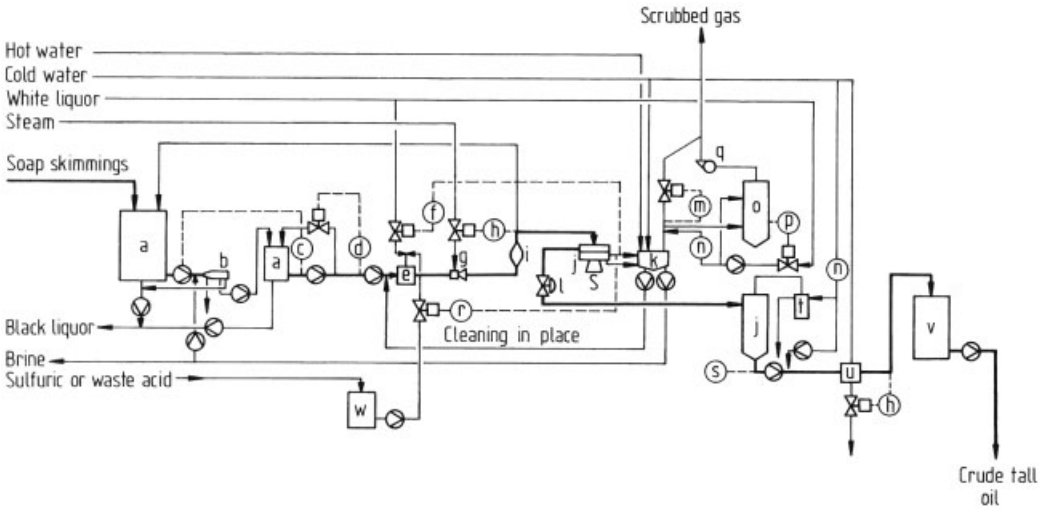

Figure 4. Hedemora – Celleco continuous tall oil system
a) Soap tank; b) Decanter – centrifuge; c) Level transmitter; d) Pressure transmitter control; e) Mixing-in piece; f) Density control; g) Direct steam heater; h) Temperature control; i) Reactor; j) Centrifugal separator; k) Brine tank; l) Constant pressure valve; m) Pressure switch; n) Flow meter; o) Gas scrubber; p) Liquid control; q) Fan; r) pH control; s) Temperature transmitter; t) Condenser; u) Tall oil cooler; v) Tall oil tank; w) Acid tank

either batchwise or continuously. In both cases the equipment must be made of acid-proof material, able to resist dilute sulfuric acid (spent acid or brine) at 90 – 100 °C. A pH of 2.5 – 3.5 is normal for complete reaction with a minimum of equipment corrosion. The Hedemora – Celleco continuous tall oil system is shown in Figure 4.

2.3. Upgrading of Soap Extraction (Crude Soap Refining Process)

On a limited scale, crude birch – pine soap is upgraded by extraction of neutrals before acidulation. The crude soap refining (CSR) process developed by Holmbom and Avela in Finland yields not only a cleaner CTO but also commercial compounds such as sitosterols, terpenes, and aliphatic alcohols from the extracted neutrals. This upgrading process has been in operation since 1977 (Kaukas, Finland).

3. Upgrading of Crude Tall Oil by Distillation

History. In 1930 – 1950 a number of relatively small distillation plants (< 10 000 t/a) came onstream in Scandinavia and the United States, producing rosin-rich distillates and distilled tall oil. The "rosin oil" fractions were crystallized and centrifuged. The rosin crystals (rosin acid content ≈ 90 %) were melted and partly saponified.

The amorphous *rosin product* replaced gum and wood rosin in Scandinavia and was used in the Delthirna process for making paper size. The *distilled tall oil fraction* (rosin acid content 20 – 30 %) – in Scandinavia named "såpolja" (soap oil) – was used to make laundry soaps, as gel and solid products.

The introduction of *Linder trays* in fractionation towers was a promising alternative to already existing "dry" distillation (randomly packed) column processes. The first Linder plant came onstream in 1947 at Mo och Domsjö, Sweden as a dry distillation process designed for a rosin fraction introduced at the bottom of the column. The tall oil flow followed a convoluted path through a series of trays, encountering vapors generated in an electrically heated reboiler at the bottom of the column. Later, several Linder plants came onstream, further developed to produce TOFA with a low rosin content.

In 1949, Arizona Chemical and E. R. Badger introduced a *"wet" distillation process* in the United States, based on the experiences and techniques of the petroleum industry. This dis-

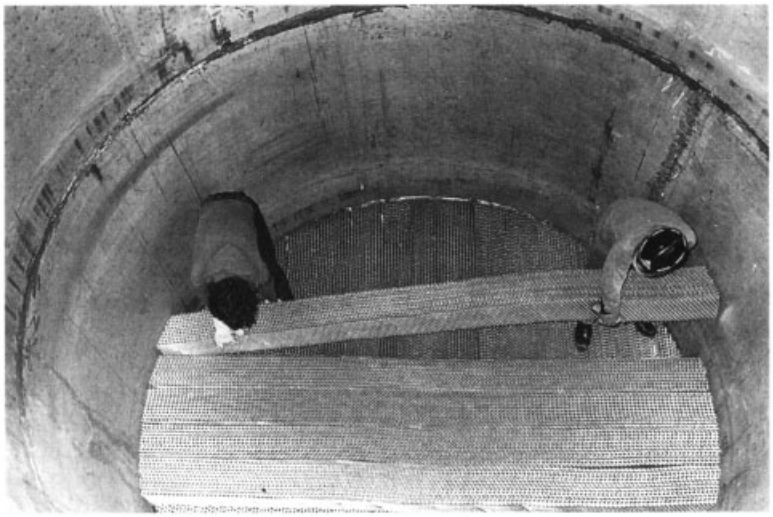

Figure 5. Sulzer's structured packing

tillation differed from the Scandinavian process. The use of bubble-cap towers for distillation made the addition of steam necessary for vaporization of the high-boiling tall oil components. To achieve sufficient separation of the fatty and rosin acids, a considerable number of trays is required. To overcome the high static pressure drop in the column, sparge steam must be used to avoid undesirable reactions of the products, caused by too high boil-up temperatures. Initially, the Arizona plant came onstream as a two-pass operation in a two-column system. The feed of the first pass is CTO, that of the second the crude fatty acid fraction from the first pass.

In 1970, Luwa, Switzerland built a plant for Krems Chemie in Austria using Luwa *thin film evaporators* for dehydration, depitching, and heat exchange for boil-up in the fractionation tower (packed columns). This approach minimized the time during which heat-sensitive compounds in tall oil were exposed to elevated temperature.

The introduction of *large-scale structured packing* (Mellapak, Sulzer Chemtech, Switzerland, see Fig. 5) at the Bergvik plant in 1979 improved the efficiency and capacity of the towers to achieve more complete fractionation of the vaporized high-boiling compounds in CTO. Worldwide, approximately 15 – 20 rosin columns and approximately 10 – 20 fatty acid columns equipped with structured packings are onstream today. A capacity comparison of two equal rosin columns, one of which is equipped with structured packing (Mellapak), is given in Table 4.

Table 4. Comparison of the capacity of two equal rosin columns

Characteristic	Pallring/minirings	Mellapak
Top pressure, kPa	0.2	0.2
Bottom pressure, kPa	1.6 – 2.0	0.8 – 1.2
Bottom temperature, °C	270 – 280	255 – 265
Capacity, t/a	40 000	80 000
Reflux ratio	4	2

3.1. Technical and Economic Conditions

Lengthy development work was required from the time CTO was discovered until a scientifically designed distillation process was achieved. Process performance may differ depending on local conditions, CTO quality, capacity demand, operation control equipment, and in-plant experiences, but the technology advances introduced have resulted in plants all over the world that are capable of producing fatty and rosin acid fractions of very high purity and quality. Special arrangements – mainly for the rosin fraction – were necessary because of variations in the quality of the CTO. However, these changes enabled the production of fractions with specific composition and properties. In practice, distillation of CTO has been achieved in two ways, both carried out in vacuum and at elevated temperature. The

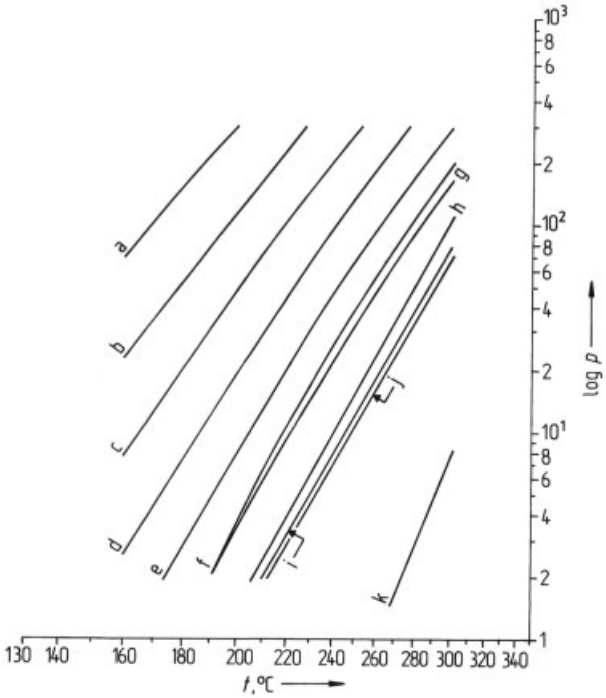

Figure 6. Vapor pressure curves of tall oil components (log p – $1/t$ curves, t in °C is depicted on $1/t$ axis for convenience)
a) Caprylic acid ($C_8H_{16}O_2$); b) Capric acid ($C_{10}H_{20}O_2$); c) Lauric acid ($C_{12}H_{24}O_2$); d) Myristic acid ($C_{14}H_{28}O_2$); e) Palmitic acid ($C_{16}H_{32}O_2$); f) Oleic acid ($C_{18}H_{34}O_2$); g) Stearic acid ($C_{18}H_{36}O_2$); h) Arachidic acid ($C_{20}H_{40}O_2$); i) Abietic acid ($C_{20}H_{30}O_2$); j) Behenic acid ($C_{22}H_{44}O_2$); k) Heavy neutrals
(Courtesy of Sulzer Chemtech AG)

high-boiling components of tall oil can be vaporized either in the presence of superheated steam (wet process) or without adding any steam at all (dry process). In both cases, low pressure is necessary to vaporize the high-boiling compounds. At actual operating temperature the compounds are sensitive to undesirable chemical reactions, depending on heating conditions. In addition, the compounds degrade rapidly in the presence of oxygen. Therefore, the entire apparatus involved in distillation must have the functional security sufficient to meet the demand of steady running.

The "dry" process is more favorable today than the "wet process"; it involves lower energy costs and less air and water pollution. The Swiss technology – rapid heating for boil-up and improved mass transfer at lower pressure drop and lower reflux ratios – for distilling CTO has been successful and used completely or partly in new and upgraded plants all over the world.

The trend in the tall oil field has been toward increased capacity, in order to gain scaleup effects, meet pollution restrictions, and produce tall oil fractions that are suitable for further processing into specific products. To obtain this, computerized control facilities are generally necessary and of great importance. Most of all, a stable operating pressure (vacuum) must be maintained to obtain fractions that meet certain values and limits. Even a small pressure change results in different process temperatures, corresponding to the new pressure. Improved techniques and resistant materials, more suited for treatment of high-boiling and heat-sensitive tall oil compounds, have led to the development of efficient distillation processes. The vapor pressure curves of tall oil components are depicted in Figure 6. The residence time of the product and the temperature differences between the product and the heating medium in different reboilers (see Table 5, next page) influence product quality and yield. Therefore, heating conditions are very important.

As heating medium Dowtherm (→ Hydrocarbons) is often used in liquid form. The desired or programmed product temperature and

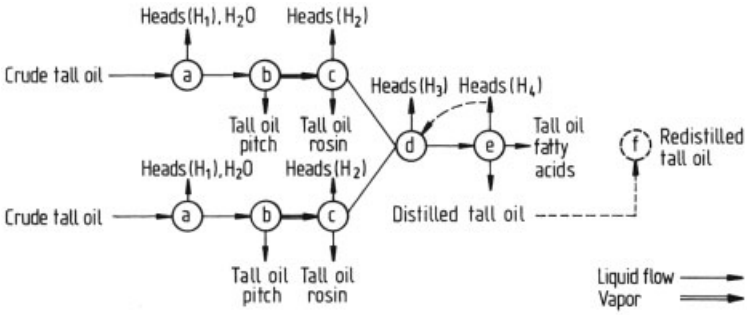

Figure 7. Simplified flow sheet for large-capacity plants (>100 000 t/a)
a) Dehydration; b) Depitcher; c) Rosin column; d) Light ends column; e) Fatty acid column; f) Distilled tall oil columns

composition can be established with great precision at each heating station by regulating the amount of Dowtherm liquid flow in the loops connected to the head coil.

Dowtherm in vapor form as heating medium is still used mainly in U.S. plants, in a way similar to conventional steam-heated systems.

The principal flow sheet for the tall oil distillation process is, in practice, quite old. A simplified flow sheet of a large-capacity plant (> 100 000 t/a) is given in Figure 7. The Bergvik Kemi tall oil plant is shown in Figure 8.

Table 5. Residence time and Δt (temperature difference between heating medium and product) for different types of reboiler

Type of reboiler	Residence time	Δt, °C
Forced circulation	≈ 60 min	≈ 5
Thin-film evaporator with fixed wipers	≈ 5 s	≈ 40
Falling-film evaporator with circulating pump	1 – 2 min	≈ 15

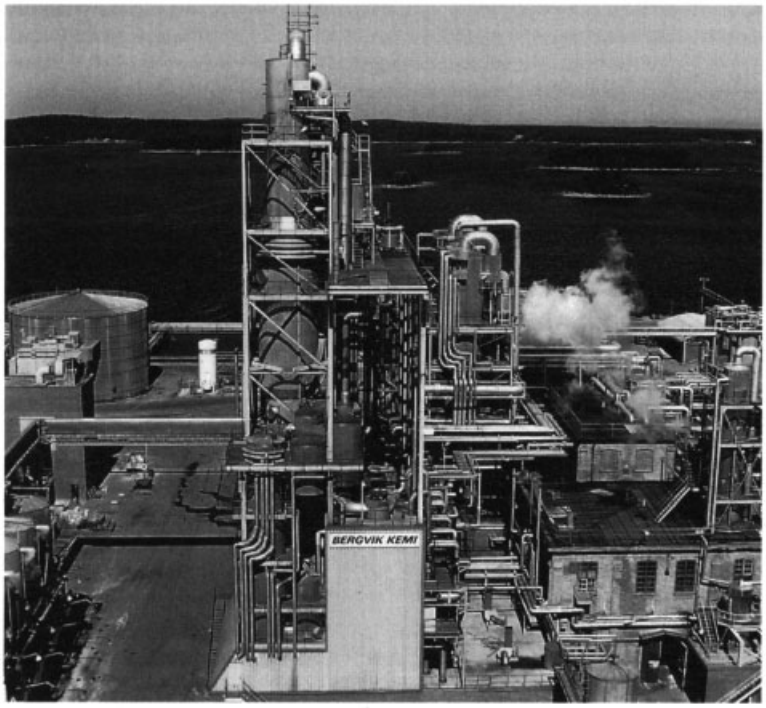

Figure 8. Bergvik Kemi tall oil plant

3.2. Process Description

3.2.1. Dehydration

CTO is fed directly from the storage tank into the thin-film evaporator (Fig. 7, a) for rapid preheating up to 200 °C under vacuum (3 – 6 kPa). Inert combustible gases, water (≈1 %), low-boiling compounds such as organic heads (H_1) and turpentine (≈1 %) are distilled during the first passage through the evaporator. This operation makes possible a stable low pressure (≈ 0.3 kPa) in the dry distillation and fractionation of the relatively high-boiling tall oil components that follow.

3.2.2. Depitching

The applicability of dry distillation not only is influenced by the CTO composition but is also highly dependent on conditions [e.g., operating pressure and temperature and residence time of the products in the depitcher (Fig. 7, b)].

The vaporization and squeezing of vaporizable products, such as separation of the fatty and rosin acids from native pitch compounds in CTO and from pitch formed by heating, require a high temperature (270 – 275 °C), even at low pressure (800 – 1300 Pa), measured at the vapor entrance into the rosin column or the first fractionation tower. The optimal depitching conditions for obtaining maximum yields and purity of rosin and fatty acid fractions are, however, of delicate nature. At elevated temperature undesirable reactions can occur such as esterification of fatty acids (by alcohols contained in the CTO), polymerization, decarboxylation of rosin acids, and isomerization of abietic-type acids. Too high a depitching temperature also leads to decomposition and vaporization of high-boiling compounds (neutrals, fatty acids) that affect the quality of rosin and fatty acid fractions negatively.

3.2.3. Fractionation of Vaporized Tall Oil Compounds

To reduce the pressure drop and achieve the lowest possible pressure in the depitching step the top section of the rosin column has a wider diameter than the lower section. Overhead vapor from the depitcher, consisting mainly of rosin and fatty acids but also of nonacidic compounds, is fed to the rosin column (Fig. 7, c) to recover rosin as a bottom fraction (TOR). The rosin acids in TOR are similar in composition to those found in gum and wood rosin. As distillate (side stream from the upper part of the column) the crude fatty acid intermediate fraction, containing about 8 % rosin acids, is taken off and fed to the light ends column (d) to remove light heads (a mixture of neutral compounds and short-chain fatty acids). The bottom fraction, containing mainly C_{18} fatty acids, a smaller portion of rosin acids, and a "minimum" of neutrals, is then fed to the fatty acids column (e) for production of high-quality tall oil fatty acids as a side stream from the top section of the column. From the top of all involved fractionation towers, smaller cuts of head fractions containing mixed neutrals and "low-boiling" ($C_8 - C_{16}$) fatty acids are distilled off ($H_1 - H_4$).

Distilled tall oil containing most of the rosin acids in the intermediate or crude fatty acid fraction, is usually distilled off as a side stream "middle-boiler" fraction as well as by other means, depending on local technical conditions and feed composition.

The fatty acids in CTO are mainly C_{18} unsaturated acids (≈ 85 %) in which a significant proportion of the linoleic acid has been isomerized to the conjugated acid during the alkaline pulping process. The rest of the fatty acids are – if compared to the C_{18} compounds – of "low-boiling" and "high-boiling" character. More than 20 different acids from C_8 to C_{24} have been identified, among which palmitic, stearic, eicosanoic, arachidic, and behenic acids are of most interest. The higher-boiling fatty acids are concentrated mainly in the DTO fraction by fractional distillation, but are also contaminants in the rosin acid fraction. Depending on the conditions during pulping and distillation, some abietic types are isomerized to dehydroabietic acid.

Both the tall oil fatty acids and the rosin acids contain two functional groups: double bonds and the carboxyl group, which allow a number of chemical reactions. Some of these occur during processing, depending on CTO composition and external physical and chemical influences. These reactions result in increased pitch yield (by esterification, polymerization, oxidation) or

Table 6. Principal composition and yields of tall oil fractions and some empirical volatility data

Fraction	Yield, %	Acid number	Composition, %		
			Rosin acids	Fatty acids	Neutrals
Head	5 – 12	70 – 120	<0.5	30 – 50	40 – 60
TOFA	35 – 45	192 – 197	<2	95 – 98	1 – 2
DTO	5 – 15	180 – 190	20 – 30	65 – 70	4 – 7
TOR	20 – 35	165 – 182	85 – 96	1 – 5	1 – 7
TOP	20 – 40	20 – 50	5 – 13	5 – 10	40 – 60

affect the molecule structure of certain sensitive materials (e.g., isomerization of linoleic acid in the fatty acid group). Levopimaric acid is converted to more stable forms, and other isomers of abietic acid are formed.

The principal composition and yields of tall oil fractions are listed in Table 6.

4. Energy Recovery

Like many other distillation processes characterized by several steps and elevated temperature (150 – 300 °C), the energy demand of tall oil distillation is large and costly. Thus, operating the distillation column at low reflux ratios and reusing the energy from the condensation system as heat (i.e., for the distillation process and for supplementary equipment, tanks, etc., belonging to a complete tall oil plant) is very important in terms of economics.

Besides the usual arrangements for heat exchange of hot and cold streams in the process, most of the thermal energy from the condensation system can be transformed into low-pressure steam (see Fig. 9). The quantity of low-pressure steam in principle meets the demands of the tall oil process. It is used mainly in vacuum jets and CTO tanks but also for other heating purposes.

5. Security and Construction Material

Acceptable construction material and process equipment that resist the corrosivity of tall oil compounds – particularly fatty acids – ensure reliable operation. Based on many years of "dry" distillation, the Co – Ni – Mo alloy, containing 2.5 % molybdenum (AISI Type 316) was regarded as an acceptable material, combating cor-

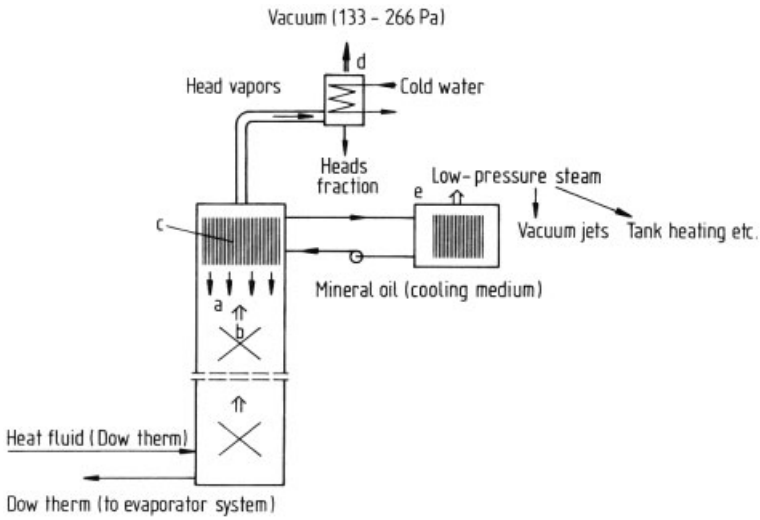

Figure 9. Heat recovery system
a) Fractionation tower; b) Flexipack packings (corrugated Al sheet); c) Deflegmator (partial condenser); d) End counter; e) Steam generator station

rosion generally in the tall oil distillation process. However, as time went on, the corrosion of replacement parts occurred, sometimes followed by unscheduled shutdowns and quite often fires. This resulted in the use of an alloy containing more than 3.5 % molybdenum (AISI Type 317).

Vapor-phase streams are more corrosive than liquid-phase streams, and corrosion increases with increasing temperature. The corrosion rate of alloys with 5 – 6 % molybdenum is negligible; thus, the use of high-molybdenum alloy in construction of the equipment is essential. Aluminum has been used for tanks, pipelines, packings, etc., and is still of interest when special strength is not required.

6. Analysis

Most of the analytical methods for tall oil are established standards and, as such, collected in ASTM and especially in SCAN (Scandinavian Pulp Paper Testing Committee) publications. About 20 standard methods are available.

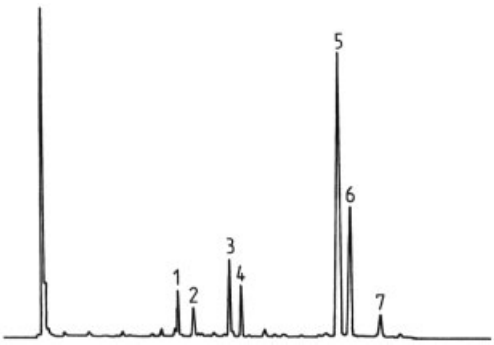

Figure 10. Gas chromatogram of tall oil rosin
1. Pimaric acid; 2. Sandaracopimaric acid; 3. Palustric acid; 4. Isopimaric acid; 5. Abietic acid; 6. Dehydroabietic acid; 7. Neoabietic acid

In addition to the standard methods of analyses, gas – liquid chromatography, mass spectroscopy, and gel permeation chromatography are used widely for better understanding of tall oil chemistry. A gas chromatogram of tall oil rosin is shown in Figure 10. These methods are essentially necessary complements to the older standard analyses for further development of the processes and of special or unique products as well. Typical data and composition of tall oil fatty acids and of tall oil rosins are listed in Tables 7 and 8.

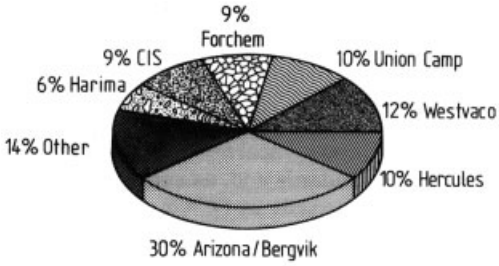

Figure 11. World tall oil fractionation capacity by company

Table 7. Typical data and composition of some tall oil fatty acids

		Scandinavia	United States
Acid number		195	197
Rosin acids, %		2	1
Unsap ∗, %		2	1.5
Iodine value		150	130
Color, Gardner		4	3
Fatty acids, %			
Saturated		2	2
Oleic	(18:1)	30	48
Linoleic	(18:2)	44	37
Linolenic	(18:3)	10	3
Conjugated	(18:2)	6	6
Other		4	2.5

∗ Unsaponifiables.

Table 8. Typical data and composition of some rosin types

	Gum rosin	Wood rosin	Tall oil rosin		
			United States	Scandinavia	
				I	II
Acid number	167	157	174	173	180
Rosin acids, %	90	85	92	90	95
Fatty acids, %			2	4	2
Softening point, °C	78	73	75	66	75
Color, USDA∗			WW	X	XA
Composition of rosin acids fraction, %					
Pimaric	7	7	2	3	3
Palustric	20	8	7	8	7
Isopimaric	15	15	8	6	6
Abietic	30	50	40	40	40
Dehydroabietic	5	8	20	22	23
Neoabietic	15	5	3	4	4

∗ For a definition of USDA color shades, see → Resins, Natural, Chap. 5.4.

Table 9. Applications of tall oil fractions

Composition, uses	Pitch	Rosin	Distilled tall oil	Fatty acids	Heads
Principal composition	esters, sitosterol, estolides, rosin dimer, free fatty and rosin acids	rosin acids (isomeric diterpene acids, complex mixture), free fatty acids, neutrals, unsaps 3 – 10 %, esters	fatty and rosin acids, neutrals, unsaps, esters $< 10\,\%$	fatty acids (oleic, linoleic types $> 90\,\%$), rosin acids (neutrals $< 10\,\%$)	fatty acids (saturated)
As intermediates in miscellaneous processes	mixtures for chemical modification	fortification, Diels – Alder reaction, dimerization (catalytic process), isomerization, esterification, disproportionation	alkyd formation, esterification, saponification, epoxidation	dimerization, isomerization, esterification, alkyd formation, saponification, separation, ethoxylation, decarboxylation, C_{21} diacids	mixtures, crystallization (palmitic acid)
Final uses	fuel, binders, coatings, rubber modifier, asphalt, sizing, printing inks, hardboard impregnation	paper size, protective coating, adhesive, ink, rubber, hot melt	soap, coating, flotation, board impregnation	protective coating, soap, ink	fuel, flotation, asphalt emulsifier

7. Storage and Transportation

The crude tall oil, melted rosin, and pitch are stored at elevated temperature to reduce viscosity. Air exposure of both fatty and rosin acid fractions results in darkening of the products. Protection with inert gases – N_2, CO_2, and steam – at elevated temperature is recommended. The fatty and rosin acid fractions are preferably stored and shipped in tanks made of stainless steel or aluminum. When rosin in the melted state is shipped in insulated tank cars, the temperature is kept at $> 160\,°C$. Rosin derivatives should be produced in direct connection with the distillation process. By storing rosin in solid form (galvanized drums), additional costs (packaging, handling, remelting, etc.) may arise. In addition, TOR in drums can crystallize easily to become nonbrittle and often difficult to remove. Today, the modified resin products are generally flaked or compressed into tablets and delivered in paper bags.

8. Uses

Uses of CTO. Most of the world production of CTO ($\approx 1.5 \times 10^6$ t/a) is utilized as raw material in continuous distillation processes, producing high-quality tall oil fatty acid fractions and tall oil rosins as primary products. If the CTO is "low grade" (e.g., the content of valuable components is too low), the product (or the corresponding soap) is used as liquid (or solid) fuel. Since CTO, irrespective of quality, can always be used as a fuel oil, its minimum price is more or less related to actual "fuel oil" prices on the market, including taxes and other associated costs.

Tall Oil Fractions. The TOFA and TOR fractions are used widely for further processing into a variety of chemical products, replacing other raw materials such as gum and wood rosin, hydrocarbon resins, and vegetable oils. Improved distillation techniques have led to upgraded fractions of TOFA and TOR that are attractive raw materials for special chemicals. The chemical structure of some fatty and rosin acids makes the primary tall oil fractions suitable as raw material for making intermediate chemicals. Examples are the production of dimeric fatty and rosin acids, polymerized rosin, and maleic and fumaric anhydride (acids) – rosin adducts including their salts and esters (→ Resins, Natural, Chap. 5.5.). The dimeric acids are utilized as raw materials to make products used in adhesives, coatings, inks and epoxy resins. Most of the intermediate chemicals are processed in

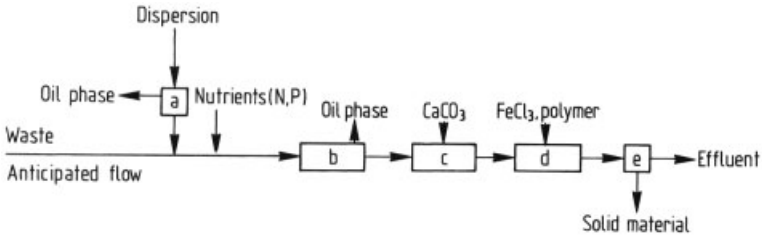

Figure 12. Principal waste treatment process
a) Phase separation equipment (Knitmesh filter); b) Oily water skimming basin; c) Aerotin basin (biological treatment); d) Chemical treatment basin; e) Filter, sludge decanter centrifuge
BOD reduction > 96 %; COD reduction > 88 %

the presence of certain catalysts. An overview of the applications of tall oil fractions is given in Table 9.

9. Economic Aspects

The total world production of CTO was in the range of 1.5×10^6 t in the early 1990s. Most of the available quantities are produced and refined in the United States and Europe. The world tall oil fractionation capacity by region is as follows:

United States	56 %
Scandinavia	21 %
Other Western Europe	7 %
Other	16 %

A breakdown of world tall oil fractionation capacity by company is given in Figure 11.

10. Toxicology and Environmental Protection

Toxicology. The tall oil compounds in living trees, stumps, CTO, fractions, etc., are nonhazardous natural substances. Modified tall oil products and solutions may, however, become hazardous because of the chemicals and solvents used. Many tall oil derivatives such as glycerol and pentaerythritol esters, alkyd resins, dimerized fatty acids, and rosin acids are generally regarded as nonhazardous products.

The toxicity of tall oil components in effluent water to fish is not a function of oxygen demand. Fish may tolerate some higher (allowed) concentration limits of rosin and fatty acids or their sodium soaps in the waste stream. These concentrations may be indicated along with the other conventional parameters (COD, BOD, etc.) of the wastewater.

Environmental Protection. During vacuum distillation, some product losses arise, resulting in polluted water and inert gases. The water streams from the various processes are of different compositions but after passing through special oil separation equipment, they are collected for the removal of organic waste constituents by biological, physical, or chemical methods before they are discharged (see Fig. 12). The noncondensed (odor) component of the gases may be disposed of as fuel gas if necessary.

The dry distillation method has the advantage of requiring fewer measures to combat pollution compared to the wet process. In distillation, most of the polluted water comes from the dehydration steps, where the "low-boiling" components (1 %) together with the water in CTO (1 %) are distilled off from the feed before depitching. Additional sources of polluted water are related to the steam jet vacuum systems, plant cleanup wastes, CTO storage tanks, pumps, etc.

11. References

1. International Tall Oil Symposium, Imatra, Finland, June 7 – 9, 1983, Turku Akademi, Turku 1983, p. 259.
2. D. F. Zinkel, J. Russel (eds.): *Naval Stores, Production Chemistry Utilization,* Pulp Chem. Assoc., New York 1989, p. 1060.
3. J. Drew, M. Propst: *Tall Oil,* Pulp Chem. Assoc., New York 1981, p. 199.
4. E. E. Sweeney, H. G. Arlt, J. Russel: *Tall Oil and Its Uses – II,* Pulp Chem. Assoc., New York 1987, p. 132.
5. W. Sandermann: *Naturharze, Terpentin, Tallöl,* Springer Verlag, Berlin 1960, p. 483.

6. P. Knoer, *South. Pulp Pap. Manuf.* **38** (1975) no. 9, 18 – 22.
7. *Ullmann,* 4th ed., **22,** 385.
8. Å. Linder: "Tall Oil Refining," *Ingenioersvetenskapsakad. Handl.* **207** (1952) 56.
9. *Analytical Procedures for Tall Oil Products,* Pulp Chem. Assoc., New York 1976.
10. K. Ukkonen, *Kem. Kemi* **18** (1991) nos. 7 – 8, 603 – 604.
11. A. Rütti, *Fette, Seifen Anstrichm.* **88** (1986) Mai, 515 – 519.

Tallow → Fats and Fatty Oils
Tanning Agents → Leather

Tantalum and Tantalum Compounds

KLAUS ANDERSSON, H. C. Starck GmbH u. Co. KG, Goslar, Federal Republic of Germany

KARLHEINZ REICHERT, H. C. Starck GmbH u. Co. KG, Goslar, Federal Republic of Germany

RÜDIGER WOLF, H. C. Starck GmbH u. Co. KG, Goslar, Federal Republic of Germany

1. Introduction 1
2. Properties 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
3. Occurrence 2
4. Extraction 3
4.1. Extraction and Processing of Ore Concentrates 3
4.2. Obtaining Ta – Nb Concentrates from Tin Slags 4
4.3. Processing Tantalum Scrap 4
5. Metallic Tantalum 5
5.1. Production 5
5.2. Semifinished Products 6
5.3. Tantalum Powder for Capacitors . . . 6
5.4. Other Uses 7
6. Tantalum Alloys 7
7. Tantalum Compounds 8
8. Analysis 10
9. Economic Aspects 11
10. Toxicology 11
11. References 12

1. Introduction

Tantalum, Ta [*7440-25-7*], atomic number 73, relative atomic mass 180.948, has the electronic structure $1s^2$, $2s^2p^6$, $3s^2p^6d^{10}$, $4s^2p^6d^{10}f^{14}$, $5s^2p^6d^3$, $6s^2$, which accounts for its principal oxidation states +2 and +5. There are two naturally occurring isotopes: ^{181}Ta and the weakly radioactive isotope ^{180}Ta (0.012 %, half-life ca. 10^{13} years). A large number of artificial radioactive isotopes are also known.

Tantalum was discovered by EKEBERG in 1802 [1,2]. The discovery and history of tantalum and niobium are closely linked (→ Niobium). The two elements usually occur together.

2. Properties

2.1. Physical Properties

The physical properties of tantalum are as follows [1–7]:

Crystal structure	Body-centered cubic
Lattice constant	$a = 0.33025$ nm
Density	16.6 g/cm^3
Melting point	2996 °C
Boiling point	5425 °C
Heat of fusion	28.5 kJ/mol
Heat of vaporization at 3273 K	78.1 kJ/mol
Specific heat capacity at 20 °C	25.41 J mol^{-1} K^{-1}
Linear coefficient of thermal expansion at 20 °C	6.5×10^{-6} K^{-1}
Thermal conductivity at 20 °C	54.4 W m^{-1} K^{-1}
Specific electrical conductivity at 20 °C	0.081 Ω^{-1} cm^{-1}
Superconductivity	4.3 K
Temperature coefficient (0 – 100 °C)	0.00383

The effect of temperature on electrical and thermal conductivity has been accurately determined [3].

The mechanical properties of the metal are strongly dependent on its purity, structure, and crystal defects, as is the case with almost all refractory metals. Even low concentrations of interstitial impurities increase the hardness and reduce the ductility. The yield strength is strongly temperature dependent, the value at 200 °C being ca. 30 % of that at 20 °C [3]. The important mechanical properties are as follows [1,3,7,8]:

Tantalum, soft annealed	
Tensile strength	240 MPa
Yield strength (0.2 % offset)	210 MPa
Breaking elongation	40 %
Vickers hardness	60 – 120 HV
Modulus of elasticity	177 – 186 MPa
Ductile – brittle transition temperature	4 K
Poisson's ratio	0.35
Tantalum, cold formed	
Tensile strength	400 – 1400 MPa
Yield strength (0.2 % offset)	300 – 900 MPa
Breaking elongation	2 – 20 %
Vickers hardness	105 – 200 HV
Recrystallization temperature	1200 – 1500 K

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_071

2.2. Chemical Properties

Like niobium, tantalum has principal oxidation states +2 and +5. Below 100 °C, tantalum metal is extremely resistant to corrosion by most organic and inorganic acids, with the exception of hydrofluoric acid. This is due to a dense adherent film of tantalum oxide, a characteristic which is utilized in the manufacture of electrolytic capacitors (Section 5.3). Above ca. 300 °C, tantalum tends to form oxides, nitrides, hydrides, and carbides. All modern tantalum capacitor powders can be handled in air at temperatures up to ca. 100 °C. However, they are classified as flammable solids as they ignite at higher temperatures, forming Ta_2O_5.

3. Occurrence

Tantalum is 54th in order of abundance of the elements in the Earth's crust (2.1 g/t). In many deposits, it occurs in association with the much more abundant niobium (24 g/t). The most important tantalum-containing minerals are tantalite, wodginite, microlite (the tantalum-rich end member of the pyrochlore series), and columbite [9]. The typical Ta_2O_5 concentration in processable pegmatitic deposits, which represent the principal raw material potential, particularly in Australia, is < ca. 0.1 %. The uranium- and thorium-containing minerals yttrotantalite, strueverite, euxenite, and samarskite are sometimes highly radioactive, which limits their workability [10]. An important potential reserve lies in a belt stretching from China to Indonesia via Thailand and Malaysia, and consists of cassiterite (SnO_2) that contains some tantalum and niobium. During tin smelting, the tantalum and niobium become concentrated in the slag (→ Niobium, Chap. 4.3.1.). The important deposits, with estimated current production capacities, are listed in Table 1, and the compositions of some tantalum-containing minerals are given in Table 2.

Table 1. Tantalum deposits [11–15]

Location	Reserves, 10^3 kg Ta_2O_5	Production (1991, estimated), 10^3 kg Ta_2O_5	Annual production capacity (estimated), 10^3 kg Ta_2O_5
Australia	17 000	235	300
Brazil	18 000	100	245
Canada	3 000	80	130
Malaysia	2 000	80	130
Thailand	16 000	150	200
Africa	10 000	25	50
Russia	> 20 000	< 50	
China	17 000	100	200
Others	10 000	< 50	< 100

4. Extraction

4.1. Extraction and Processing of Ore Concentrates

In the important Australian mines, the tantalum-containing minerals are present in finely divided form in hard pegmatitic rock. The rock is extracted by open-cast mining. Further processing, which is different from that used for the mainly niobium-containing deposits (→ Niobium and Niobium Compounds, Chap. 4.), is an initial size reduction to < 15 mm, followed by grinding in ball or rod mills to < 1 mm. Flotation is then carried out, and the soluble constituents are leached out from the concentrates. The latter are then dried, and the magnetic constituents are removed [18].

Table 2. Composition of important tantalum-containing minerals [9, 16, 17]

Mineral	Composition	Ta_2O_5 content, %	Nb_2O_5 content, %
Tantalite	$(Fe, Mn)Ta_2O_6$	42 – 84	2 – 40
Microlite	$(Na, Ca)(Ta, Nb)_2O_6F$	60 – 70	5 – 10
Columbite	$(Fe, Mn)(Nb, Ta)_2O_6$	1 – 40	40 – 75
Wodginite	$(Ta, Nb, Sn,, Fe, Mn)_{16}O_{32}$	45 – 56	3 – 15
Yttrotantalite	$(Y, U, Ca)(Ta, Fe^{3+})_2O_6$	14 – 27	41 – 56
Fergusonite	$(RE^{3+})(Nb, Ta)O_4$ *	4 – 43	14 – 46
Strueverite	$(Ti, Ta, Nb, Fe)_2O_4$	7 – 13	9 – 14
Tapiolite	$(Fe, Mn)(Nb, Ta)_2O_6$	40 – 85	8 – 15
Euxenite	$(Y, Ca, Ce, U, Th)(Nb, Ta, Ti)_2O_6$	1 – 6	22 – 30
Samarskite	$(Fe, Ca, U, Y, Ce)_2(Nb, Ta)_2O_6$	15 – 30	40 – 55

* RE = rare earth element.

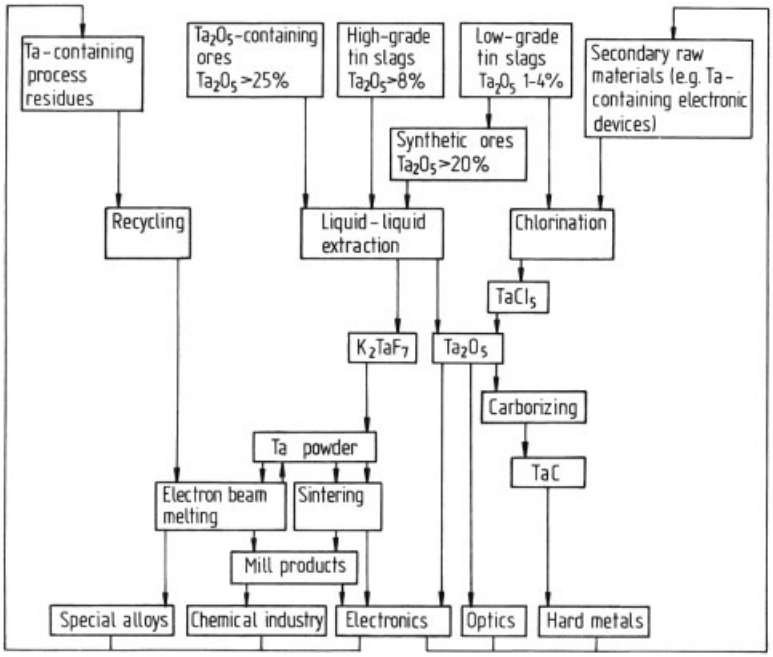

Figure 1. Major tantalum processing routes

This ore concentrate, up to 40 % Ta_2O_5, is dissolved in concentrated hydrofluoric acid (→ Niobium and Niobium Compounds, Chap. 4.3.), and is then separated from the niobium by extraction with methyl isobutyl ketone, so that the tantalum is obtained as H_2TaF_7. Potassium salts are added, and potassium heptafluorotantalate, K_2TaF_7 [*16924-00-8*], crystallizes out. This is reduced with sodium to metallic tantalum (Section 5.1).

Major tantalum processing routes are summarized in Figure 1.

4.2. Obtaining Ta – Nb Concentrates from Tin Slags

Tantalum-containing tin slags represent an important source of tantalum; e.g., in 1980, ca. 50 % of world demand was recovered from tin slags [19–21]. Owing to the low price of tin, its production is in continuous decline, so that the availability of tantalum-containing tin slags is decreasing. Moreover, the tantalum concentration in tin concentrates is also decreasing significantly. The fraction of the world demand for tantalum represented by Ta_2O_5 in tin slags is expected to fluctuate around 20 %. The price of tin is not likely to reach a level at which more tantalum-rich tin ores from Thailand and Malaysia can be processed economically.

The tantalum content of these slags fluctuates widely. Bolivian slags do not contain tantalum, but those from Thailand can have tantalum content 15 – 16 % Ta_2O_5. The Ta:Nb ratio varies from 1:0.4 (Thailand) to 1:3 (Nigeria).

The most important grades of slag have the following tantalum and niobium content [19–21]:

	Ta_2O_5, %	Nb_2O_5, %
High-grade		
Thailand	13 – 16	9 – 10
Zaire	11	10
Australia	10	5 – 6
Medium grade		
Thailand	4 – 5	4
Malaysia	3 – 4	3 – 4
Nigeria	4 – 5	12
South Africa	6 – 8	8 – 9
Low-grade		
Malaysia	0.8 – 2.5	0.8 – 2.5
Thailand	0.5 – 1	0.2 – 0.8
Brazil	1 – 2.5	1 – 2.5
Singapore	1 – 1.5	1 – 1.5

High- concentration slags are used directly in the wet chemical process (dissolution of the slag in hydrofluoric/sulfuric acid, followed by extraction and separation of tantalum and niobium).

From low- concentration slags, Ta/Nb concentrates containing 50 – 60 % Ta_2O_5/Nb_2O_5 can be produced, e.g., pyrometallurgically. The Ta/Nb content of these synthetic concentrates is > 99 % soluble in hydrofluoric acid, as is the case with natural ores.

In all the concentration processes so far developed, the slag is first melted in an electric furnace with addition of reducing agents, e.g., coke, and fluxing materials. This causes the Ta/Nb to collect in the carbide- containing ferroalloy produced, while some impurities remain in the slag. The solid ferroalloy is produced in relatively large blocks, which must be broken up and ground. The tantalum/niobium content is increased in further process steps, giving a concentrate containing 50 – 60 % oxides of tantalum and niobium.

One method of enriching the tantalum and niobium consists of oxidizing the ferroalloy, which contains the Ta/Nb carbide, with Fe_2O_3 (e.g., as hematite). This causes undesired elements to be collected in the oxidic slag, while the Ta – Nb carbide is oxidized only as far as the metals and therefore remains in the metallic phase. In a final reaction stage, this is again oxidized with Fe_2O_3 to form a slag that contains tantalum and niobium oxides [22].

The partial oxidation can also be carried out in a single smelting stage using air, oxygen-enriched air, or pure oxygen. Under these conditions, the tantalum is converted to slag much more rapidly than the niobium, enabling the tantalum to be concentrated at the expense of niobium [23]. Ta/Nb carbide- containing ferroalloy can also be treated by wet leaching processes [24–26].

4.3. Processing Tantalum Scrap

When processing tantalum- containing waste materials, physical and chemical processes are used as much as possible, avoiding processes typically used in ore treatment, i.e., dissolution in hydrofluoric acid and solvent extraction. The tantalum scrap is divided into two categories which differ in difficulty of recycling:

1) Scrap material of the first type can consist of pure unoxidized metallic tantalum, e.g., sintered tantalum powder pellets (anodes) from the manufacture of tantalum electrolytic capacitors, scrap foil, sheet, and wire, or of anodically oxidized sintered tantalum pellets containing 1.5 – 3 % oxygen from capacitor manufacture and oxidized wire scrap arising from this process.
   The following processing methods are commonly used for these scrap materials: ingot melting in an electron beam furnace, or conversion to brittle tantalum hydride, grinding, and dehydrogenation at > 600 °C in vacuo or under a protective gas to form metallic tantalum powder [1]. Alternatively, the ground tantalum hydride can be carburized with carbon black to form tantalum carbide [27].
   The oxygen content of oxidized tantalum anodes can be reduced by deoxidation with magnesium or calcium to give high-grade scrap [28].
2) Scrap of the second type consists of oxidized tantalum anodes coated with manganese dioxide or with conductive silver, or sometimes welded to nickel conductor wires. This category also includes faulty capacitors, in which the coated and welded tantalum anodes also have an encapsulation of synthetic resin or tinned brass.
   Pure oxidized tantalum anodes can be recovered from coated anodes and brass-encapsulated capacitors by successive treatments with nitric and hydrochloric acid. Resin- encapsuled capacitors require complex treatment. These can be coarsely size-reduced in grinding mills, the tantalum being recovered by density separation. Foreign metals are removed by treatment with nitric and hydrochloric acid, leaving a residue of oxidized tantalum anodes in granular form [29].
   Another method has been described for the treatment of tantalum anode scrap containing manganese dioxide. This can be directly reduced in an argon/hydrogen plasma and melted to pure tantalum [30]. If the scrap capacitor materials are resin free, the tantalum can be recovered as its pentachloride by chlorination [31].

Some types of tantalum scrap which contain other impurities are roasted to form the oxide, and treated like ore concentrates. These include condensates formed during the melting of tantalum ingots, scrapped alloy materials, impure sawings and turnings, flue dust, and sediments from the wash liquors from the sodium reduction of potassium heptafluorotantalate.

Hard metals that contain tantalum carbide can be attacked by roasting and treating with caustic soda solution [32], anodic oxidation [33], or melting with sodium nitrate [34], converting the tantalum to its oxide for further recovery treatment.

5. Metallic Tantalum

5.1. Production

Metallic tantalum products (powder, semifinished products, ingots) are produced almost exclusively by reduction of potassium heptafluorotantalate with sodium [35,36]:

$$K_2TaF_7 + 5\,Na \longrightarrow 2\,KF + 5\,NaF + Ta$$

The strongly exothermic reaction is controlled by adding inert salts (KCl, NaCl, KF, NaF). The reduction can be carried out at constant temperature by controlled addition of one of the reactants, with controlled heat removal. Tantalum powder is recovered by leaching the salts out of the reaction product. Two processes are commonly used to obtain tantalum capacitor powder from this:

1) The sodium-reduced metallic tantalum powder is purified and converted to a compactable and free-flowing product by further processing stages, e.g., high-temperature treatment in vacuo or under protective gas, deoxidation by magnesium [37], or a combination of these two processes [38]. Optimum properties for capacitor production are obtained by the addition of dopants to control sintering of porous pellets [39–41].
2) During ingot melting of compressed slugs of tantalum powder in electron beam or electric arc furnaces [42], volatile impurities are driven off, usually in two smelting operations. The ingot can be ground after hydrogenation. The powdered tantalum "hydride" is then dehydrogenated in vacuo [1] and agglomerated by high-temperature treatment.

In addition to the sodium reduction process, other processes have been described, i.e., the molten salt electrolysis of K_2TaF_7 with addition of Ta_2O_5 or $TaCl_5$, and hydrogen reduction of $TaCl_5$ [1,43], but these have not achieved economic importance. The carbothermic reduction of Ta_2O_5 in high vacuum [1,4,44] at ca. 1900 °C leads first to tantalum carbide:

$$Ta_2O_5 + 2\,C \longrightarrow 2\,TaC + 5/2\,O_2$$

This reacts above 2000 °C with residual Ta_2O_5 to metallic tantalum:

$$5\,TaC + Ta_2O_5 \longrightarrow 7\,Ta + 5\,CO$$

Because of the high quality standards for particle size and purity required by the producers of tantalum capacitors (Section 5.3), this process is of minor importance.

5.2. Semifinished Products

Tantalum metal powder for producing semifinished products is also obtained by reducing K_2TaF_7 with sodium. Semifinished products such as sheet, wire, and shaped articles are produced by two different methods:

1) Powders are compacted, and then sintered by passing an electric current at ca. 2500 °C (Coolidge process) or indirectly by resistance or induction heating in vacuo. Tantalum wire for tantalum electrolytic capacitor manufacture is produced by this process. It is advantageous to control the grain size, e.g., by adding dopants [45], to optimize the mechanical properties.
2) Bars produced by powder compression are formed into ingots by electron beam or electric arc melting, and shaped by forging, rolling, extruding, and wire drawing. Intermediate heat treatment in vacuo is usually unnecessary as tantalum does not work harden. After 95 % deformation, a tantalum product has typical hardness 180 HV, after an initial ingot hardness of 80 – 90 HV [46,47].

As well as its use for electrolytic capacitors (wire, cans, foil), semifinished tantalum is also

used in the chemical industry because of its excellent resistance to corrosion by hot, concentrated mineral acids, with the exception of hydrofluoric acid which strongly attacks tantalum. The passivation of the metal is a result of the dense, adherent oxide film, 1 – 4 nm thick, which forms spontaneously in air and in acidic media, and which is stable up to ca. 260 °C [48]. Welding of tantalum sheet is mainly carried out by the tungsten – inert gas (TIG) process, which allows complex heat exchangers and intricate heating equipment to be fabricated [49]. Impurity levels in commercial semifinished tantalum products are listed in Table 3.

Table 3. Typical chemical composition of commercial tantalum semifinished products

	Powder metallurgy grade	Electron beam melted grade
Ta	$\geq$ 99.9 %	$\geq$ 99.98 %
Impurity content, ppm		
O	250	100
C	75	30
N	75	30
H	15	10
Nb	10	20
W	10	20
Mo	5	10
Fe	75	10
Cr	15	3
Ni	20	2
Na	10	1
K	10	1

Because of the high price of tantalum, large vessels and pipework are often merely clad with the metal. Plating of steel with tantalum can be carried out by the explosion method [50], or corrosion protection can be achieved by build-up welding of tantalum [51].

5.3. Tantalum Powder for Capacitors

Since the first report that tantalum could be used for electrolytic capacitors [52], the growth of this application has been such that it now accounts for the highest consumption of the metal [53]. The capacitor consists of an anode of tantalum metal powder compressed into a pellet, typically 20 – 200 mg. This anode is sintered to give electrical continuity and density of ca. 6 – 9 g/cm^3. The electrical connection is provided by a tantalum wire, attached by sintering or welding. The whole of the pore surface of the sintered anode is coated with an amorphous dielectric layer of Ta_2O_5 by anodic oxidation, using an electrolyte of dilute H_3PO_4/H_2SO_4, sometimes with the addition of organic compounds [54, 55]. Ta_2O_5 has a resistivity of ca. $10^{15}\ \Omega \cdot$ cm [56] and dielectric constant $\varepsilon = 25$. The anode is coated with a semiconducting layer of MnO_2 by impregnation with an aqueous solution of manganese nitrate, followed by pyrolysis. This solid electrolyte provides the cathodic contact. Research has been carried out into the effect of the pyrolysis temperature on the Ta_2O_5 layer [57] and its dielectric properties [58], and also into the effect of temperature on the capacity [59] and conductivity [60–62], and the effect of these on the leakage current of the capacitor [63]. Progressive improvements to powder quality [64] have doubled the specific capacitance of the tantalum metal powder (the product of capacitance and formation voltage per unit mass) [65]. Typical qualities and properties of capacitor grade tantalum powder are given in Table 4.

5.4. Other Uses

Tantalum is used in medicine because of its lack of toxicity and very good compatibility with tissue. Reports and discussions have been published on the suitability of inhaled tantalum powder as a contrast medium in X-ray diagnostics of the throat, trachea, larynx, bronchi, esophagus, and stomach [66–68], but the material has been rarely used, mainly for cost reasons. Small tantalum spheres or tantalum wires are used in organs (e.g., the heart), in bone, or in implants as markers when controlling position and function by X-rays [69–71].

Because of its stable oxide layer, tantalum in the body is completely bioinert. It has been shown in animal experiments that, if broken bones are held in position with tantalum rods, even in septic conditions with corrosive tissue reaction, healing is better than when chromium – nickel steel or niobium are used [72]. Tantalum implants in the jaw have shown good metal – bone contact over many years with no morbid reaction [73]. Tantalum has been largely replaced as a prosthetic implant material by titanium, as this has better strength properties and is sufficiently bioinert [74]. However, tantalum has an important use in clips for rapid occlusion of ves-

Table 4. Typical properties of capacitor-grade tantalum powder

Chemical impurities, ppm	Powder from electron beam melting		Powder from sodium reduction	
	QR 7∗	QR 3∗	VFI-18 KT∗∗	STA-30 KD∗∗∗
O	800	1500	1800	2200
C	10	10	35	40
N	20	40	50	70
H	15	15	50	50
Nb	< 20	< 20	< 10	< 10
W	< 20	< 20	< 10	< 10
Fe	< 10	< 10	25	25
Cr	< 5	< 5	10	10
Ni	< 5	< 5	20	20
Na	< 1	< 1	< 2	< 2
K	< 1	< 1	5	10
Specific charge, C V/g (μF V/g)	3000	5000	15 000	30 000
Application working voltage, V	50 – 75	35 – 50	25 – 35	< 25

∗ H. C. Starck, Newton, United States.
∗∗ H. C. Starck – V-tech, Tokyo.
∗∗∗ H. C. Starck, Goslar, Germany.

sels in surgery, either temporary or permanent, [75–77]. In some cases, tantalum mesh and plate can be implanted in damaged areas of the skull [78–80].

The possible use of the very dense and ductile tantalum for the manufacture of heavy missiles for armour penetration is under investigation [81].

6. Tantalum Alloys

Tantalum-Based Alloys. The most common tantalum alloy is Ta – 2.5 W, produced by ingot melting. This has higher strength and resistance to deformation than pure tantalum, but still has good cold working and welding properties. It is used for the construction of industrial chemical equipment, especially heat exchangers, pipes, and vessels. Its resistance to hot mineral acids (except hydrofluoric acid) is even somewhat greater than that of pure tantalum [82]. The alloy Ta – 7.5 W, produced by powder metallurgy, is harder, and is therefore a preferred material for springs in chemical equipment, e.g., valves. The alloy Ta – 10 W, mainly produced by ingot melting, is relatively hard, but still ductile. This, too, is used for special applications in chemical equipment, and in aircraft construction [82, 83]. For cost reasons, tantalum – niobium alloys have been proposed for chemical plant, but these alloys, Ta – 40 Nb [83] and Nb – 25 Ta [84], are less resistant to corrosion by hot acids.

The alloys Ta – 8 W – 2 Hf (T-111) and Ta – 9.6 W – 2.4 Hf – 0.01 C (T-222), which have good machining and high-temperature properties, were developed for use in air- and spacecraft, e.g., in nozzles, and for cladding the interior of combustion chambers and the exterior of rockets [85].

Earlier research led to the development of many other binary and ternary alloys of tantalum with Nb, W, Mo, Hf, and V [86, 87], Hf, W, and Al [88], or Ti and Al [89]. Some of these alloys have improved high-temperature properties and oxidation resistance compared with pure tantalum, but have not achieved industrial importance.

Superalloys. In superalloys, tantalum contributes mainly to properties such as high-temperature strength, reduction in fatigue at high temperature, and reduction of corrosion by sea air. Superalloys (→ High-Temperature Materials, Chap. 7.3.) are metallic multicomponent materials, with nickel, cobalt, or iron as the main constituent, and additions of tantalum, aluminum, chromium, hafnium, molybdenum, niobium, rhenium, titanium, tungsten, zirconium, etc., developed for use at high operating temperatures and under high mechanical stress. Components, produced by forging, casting, or powder metallurgy, are used in aircraft engines and stationary gas and steam turbines. These conventional superalloys contain up to 9 % tantalum (e.g., TRW-NASA VI A, MAR-M 302) [90]. The

manufacture of single- crystal components, e.g., for turbine blades, and the directional solidification of superalloys specially developed for this technique (e.g., Rene 142, with 6.5 % Ta and PWA 1480 with 12 % Ta) are processes of increasing importance [91, 92].

The effect of tantalum in the nickel-based superalloys B-1900 (4 % Ta) and MAR-M 247 (3 % Ta) has been systematically investigated. The favorable high-temperature properties are mainly due to the γ'-phase of the type $Ni_3(Al, Ti, Ta)$, formed by precipitation from the γ-phase. Tantalum increases the volume fraction of the γ'-phase in the γ'/γ ratio. For the most part, it appears in this phase without changing the partitioning ratios of the other elements. A small part of the tantalum appears in the metal carbide fraction, which improves the high-temperature strength [93, 94]. As well as the composition of the material or component, a multistage heat treatment is very important for optimization of γ'-phase precipitation.

7. Tantalum Compounds

Oxides. Tantalum oxides in lower oxidation states are not industrially important. Tantalum(II) oxide, TaO [*12035-90-4*], is the only lower oxide whose existence has been confirmed. It is produced from tantalum pentoxide by reduction with carbon at 1900 °C, or with hydrogen at 1100 °C. A possible use is for the production of heat-reflecting window glass [95].

No tantalum compound analogous to niobium(IV) oxide, NbO_2, is known. However, with tantalum pentoxide, the possibility exists of introducing interstitial tantalum atoms, so that TaO_x compounds can be obtained ($x = 2 - 2.5$). These oxides have metallic conductivity, but do not form discrete phases.

Tantalum pentoxide, Ta_2O_5 [*1314-61-0*] (M_r 441.9, *mp* 1880 °C, density 8.73 g/cm^3), occurs in two thermodynamically stable modifications, α and β. The transition temperature from the orthorhombic β modification to the tetragonal α modification is ca. 1360 °C. The existence of an ε modification, produced hydrothermally from tantalic acid at 300 – 340 °C, is also known. This ε-Ta_2O_5 is isomorphous with β-Nb_2O_5, and is transformed to β-Ta_2O_5 by heating in air above 886 °C [96].

Two processes are used for the manufacture of tantalum pentoxide:

Wet Chemical Method. The raw materials (ores and/or tin slag concentrates) are dissolved in hydrofluoric acid at ca. 100 °C. Tantalum and niobium are extracted from this strongly acid aqueous solution, preferably with methyl isobutyl ketone. The impurities that are also extracted, SiF_6^{2-}, FeF_6^{3-}, AlF_6^{3-}, SbF_6^{2-}, etc., are washed out with sulfuric acid, and the tantalum and niobium are separated in a multistage operation. Ammonia gas or solution is added to the aqueous H_2TaF_7 solution, forming hydrated tantalum oxide, $Ta_2O_5 \cdot n\,H_2O$ [*75397-94-3*]. This is washed, dried, and calcined above 800 °C to form the oxide (→ Niobium) [97].

Chloride Process. The ferroalloy produced by the pyrometallurgical method is chlorinated in an iron chloride – sodium chloride melt at 550 – 600 °C. The complex salt $NaFeCl_4$ present in the melt acts as a chlorine carrier. The iron loses a chlorine atom and is reduced from the trivalent to the divalent state, and the added chlorine gas causes the iron to be continuously reoxidized to Fe(III). The $TaCl_5/NbCl_5$ mixture is separated by fractional distillation to obtain the chlorides of the two elements in a highly pure state. To obtain the oxide, the ground chlorides are hydrolyzed with steam in a fluidized bed, and then calcined (→ Niobium).

In the optical industry, tantalum pentoxide is used in lanthanum borate glasses, characterized by their high refractive index and low optical scattering; since the 1980s, niobium pentoxide has been increasingly used for this application, as tantalum pentoxide is 3 – 5 times as expensive and has twice the density [98].

Tantalum pentoxide, and especially hydrated tantalum oxide $Ta_2O_5 \cdot n\,H_2O$ [*75397-94-3*], is used as an acid catalyst. In its amorphous form, its surface acidity can be as high as $H_0 \leq -8.2$ [99]. It therefore has good catalytic activity for alkylation, esterification, Beckmann rearrangement, aldol condensation, etc. [100–103].

Tantalates. Tantalates can be produced by calcination of oxide mixtures, or by dissolving tantalum pentoxide in molten alkali metal hydroxides or carbonates. Excess alkali metal hydroxides or carbonates leads to the formation of water-soluble tantalate isopolyanions $(H_xTa_6O_{19})^{(8-x)-}$, where $x = 0, 1$, or 2 [104].

The most industrially important tantalates are those of lithium and yttrium. Single crystals of lithium tantalate, $LiTaO_3$ [*12031-66-2*], are pulled from the melt crucible by the Czochralski method [105]. The crystal growth rate for crystals 5 – 8 cm in diameter is ca. 2 – 3 mm/h. These ferroelectric single crystals in the unpoled state have domains that are statistically distributed in the crystal. They are poled on passing through the Curie temperature (665 – 618 °C) in an electric field and then have only a single domain [106, 107]. Single- crystal wafers of $LiTaO_3$ are widely used as surface acoustic wave filters (SAW) for high-frequency applications in communication systems (intermediate-frequency filters in television and video equipment). Also, lithium tantalate is used in combination with lithium niobate in the manufacture of waveguides, frequency doublers (second harmonic generation), optical modulators, and optical switches.

Yttrium tantalate powder, $YTaO_4$ [*12143-49-6*], produced in a solid state reaction, is applied as an emulsion to a film, and is used in diagnostic medicine as a phosphor for the amplification of X-rays. It converts X-rays to fluorescent light in very high yield, enabling the X-ray dose received by the patient to be significantly reduced [108].

Barium magnesium tantalate, $Ba_3MgTa_2O_9$ [*12231-81-1*], and barium zinc tantalate, $Ba_3ZnTa_2O_9$ [*12231-88-8*], are likely to play an increasingly important role in the future as the base material for microwave resonators for the stabilization of oscillators, or as frequency filters. These dielectric materials, which have the perovskite structure, have high dielectric constant and very low dielectric loss at high frequencies [109–112]. These high-quality resonators are especially necessary for the high carrier frequencies in satellite communication.

Halides. Tantalum pentafluoride, TaF_5 [*7783-71-3*], (*mp* 96.8 °C, *bp* 229.5 °C) is produced by passing fluorine over metallic tantalum, or from tantalum chloride by adding anhydrous hydrogen fluoride [113–115]. Both tantalum pentafluoride and niobium pentafluoride are used in the petrochemical industry as isomerization and alkylation catalysts. The fluorides of both elements are also useful as fluorination catalysts for the manufacture of fluorochloro- and fluorinated hydrocarbons [116].

The oxyfluorides $TaOF_3$ [*20263-47-2*] and TaO_2F [*13597-27-8*] are also known, but have no industrial importance.

Potassium heptafluorotantalate (potassium tantalum fluoride), K_2TaF_7, is an industrially important intermediate in the production of tantalum metal (Section 5.1). It crystallizes in colorless, rhombic needles when K^+ ions (e.g., KCl or KF) are added to solutions of tantalum in hydrofluoric acid (e.g., H_2TaF_7 solution) from the extraction process. If the acidity is not high enough, undesired oxyfluorides are formed in addition to K_2TaF_7.

The solubility of potassium heptafluorotantalate in hydrofluoric acid solution decreases from 60 g/100 ml at the boiling point to < 0.5 g/100 ml at room temperature. As the corresponding stable niobium salt K_2NbOF_5 is significantly more soluble, tantalum can also be separated from niobium by fractional crystallization (Marignac process) [7].

Tantalum pentachloride, $TaCl_5$ [*7721-01-9*], *mp* 216 °C, *bp* 242 °C, forms colorless, strongly hygroscopic, needle-shaped crystals. On an industrial scale, production of this chloride is exclusively by chlorination of metallic tantalum, ferrotantalum, or tantalum metal scrap. This process is mainly used in the production of high-purity oxides [117, 118].

Tantalum pentachloride is soluble in absolute alcohol, with formation of the corresponding alkoxide [119, 121]. Ultrafine tantalum oxide powder can be produced by hydrolysis of this alkoxide [121].

Tantalum pentachloride and tantalum alkoxides are suitable for use in the chemical vapor deposition of tantalum metal or tantalum oxide.

The other pentahalides, $TaBr_5$ [*13451-11-1*] and TaI_5 [*14693-81-3*], are known, but are of no industrial importance.

Carbides. Tantalum carbide, TaC [*12070-06-3*], is golden-brown, *mp* 3985 °C (one of the highest known). It is added as powder in the manufacture of hard metals. Tantalum carbide can be produced by the direct carburization of Ta metal by carbon, or preferably by reaction of Ta_2O_5 with carbon at ca. 1900 °C. In hard metal mixtures based on tungsten carbide, tantalum carbide is present at concentrations

0.5 – 10 %. It improves the high-temperature fatigue strength and thermal shock resistance of the cutting tool. For details of its properties, production, and use → Carbides, Chap. 2.3.

Nitrides. Tantalum nitride, TaN [*12033-62-4*], is a hard material, like tantalum carbide, with good wear resistance. It is produced by heating tantalum in a pure nitrogen atmosphere at ca. 1100 °C. Tantalum nitride films are used as diffusion barriers in semiconductor technology [122, 123], and as protective coatings for moisture sensors [124]. Tantalum nitride is also sometimes found in small amounts (3 – 8 %) in cermets (base material Ti(C, N)).

Borides. The tantalum – boron phase diagram [125] shows the existence of Ta_2B [*12045-26-0*], Ta_3B_2 [*12228-67-0*], TaB [*12007-07-7*], Ta_3B_4 [*12045-92-0*], and TaB_2 [*12007-35-1*]. These hard materials are of little industrial importance. The borides may be produced by heating tantalum and boron in vacuo at ca. 1800 °C, or by electrolyzing a melt of Ta_2O_5, B_2O_3, CaO, and CaF_2 (→ Boron Carbide, Boron Nitride and Metal Borides, Chap. 3.).

Silicides. The silicides $TaSi_2$ [*12039-79-1*] and Ta_5Si_3 [*12067-56-0*] [126] are produced from the elements by heating in vacuo at 1000 – 1500 °C. Tantalum silicide targets were formerly used in the semiconductor industry for production of interconnections by sputtering [127], but tantalum silicides are no longer of any industrial importance.

8. Analysis

The analytical chemistry of tantalum is described in detail in several monographs [128–131]. Analysis for tantalum as a major or minor component or trace element in raw materials, ferroniobium – tantalum, scrap materials, and alloys must be distinguished from analysis for impurities in tantalum products. The classical method of analyzing for tantalum is by gravimetry after extraction as the fluoro complex on an ion-exchange resin [132]. Low concentrations, i.e., in the μg/g region, are usually determined photometrically [131].

When tantalum is a major or minor component in ores, scrap, etc., it is usually determined by fusion with sodium or lithium borate, followed by X-ray fluorescence analysis ($\geq$ 0.01 % Ta) [133]. Analysis of solutions by emission spectrometry with plasma excitation (ICP-AES) is increasingly used. Tantalum in the ng/ml region can be determined by measurement at 296.5 nm.

Because of the high purity required for electronic components, optical glasses, and superalloys, sensitive multielement methods have been developed for the determination of trace impurities in commercial products, e.g. potassium heptafluorotantalate, metallic tantalum, and tantalum oxide. The first methods were based on OES with vaporization in a carbon d.c. arc [134]. Additionally, many elements can be determined by plasma optical emission spectrometry (DCP-AES, ICP-AES) after dissolution by treatment with hydrofluoric or hydrofluoric/nitric acid under pressure [135]. Alkalis are determined by flame atomic absorption spectrometry, and arsenic, antimony, and other hydride formers by hydride techniques. Phosphorus, as a doping element in tantalum powders, can be very reliably determined photometrically [136]. Carbon and sulfur are determined by combustion analysis, and oxygen and nitrogen by hot extraction with a carrier gas.

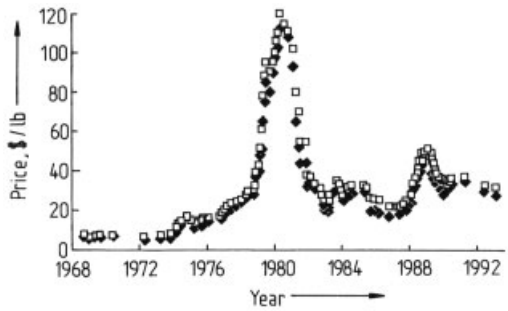

Figure 2. Tantalum ore prices
London Metal Bulletin 25/40 % Ta_2O_5, based on 30 % Ta_2O_5, cost, insurance, and freight

As the optical spectrum of the tantalum matrix contains a very large number of lines, there are many possibilities for interference; trace ele-

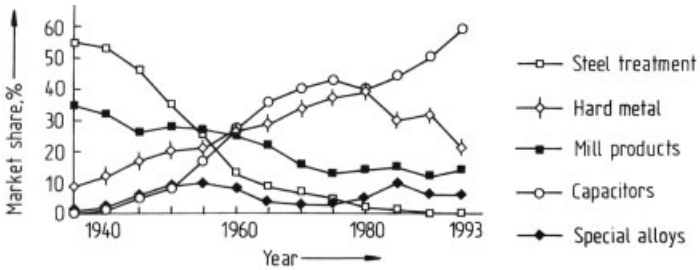

Figure 3. Market share of tantalum applications, 1935 – 1993

ment – matrix separation offers the possibility to overcome the problems [135, 137, 138]. For the final determination, ICP-AES, ICP-mass spectrometry, or neutron activation analysis are used. For trace determination in the sub-ppm region, glow discharge mass spectrometry is now the most sensitive routine method of determination [139].

9. Economic Aspects

Figure 2 shows how the price of tantalum has fluctuated. The most important tantalum processors have made long-term ore contracts with the large mining concerns [140]. World production of tantalum ores (Table 1) is well below capacity.

World demand for tantalum has changed little since 1970, remaining at ca. 1000 t/a. Forecasts based on increasing tantalum demand for electrolytic capacitors [141] have not been borne-out, as developments in tantalum powders have meant that increasingly less tantalum powder is required for a given capacitor type [142]. However, the number of tantalum capacitors produced has grown continually, from ca. 2×10^9 in 1975 to ca. 9×10^9 in 1992 [53]. The most important area of use is for electrolytic capacitors, which account for >50 % of total world production of tantalum. The use of tantalum carbide as an additive in the hard metal industry, which was still important in 1980, has decreased in recent years [17] as TaC is being replaced by chromium and vanadium carbides for cost reasons. Figure 3 illustrates the shifting market shares of the various application areas over recent decades. Applications not included, e.g., Ta_2O_5 in optical applications and in X-ray phosphors, account for considerably less than 5 % of the market.

10. Toxicology

Metallic tantalum is completely nontoxic and bioinert, and allergic reactions are unknown. This has enabled it to be used in X-ray diagnostics and surgery (Section 5.4). Tantalum carbide is not known to have acute toxicity. Tantalum oxide is classified as nontoxic, although animal experiments indicate that the LD_{50} (rat, oral) is 8000 mg/kg (RTECS). For respirable dusts of metallic tantalum, the MAK value is 5 mg/m^3. Tantalum chloride has a strongly corrosive and irritant effect, owing to the hydrolytic production of hydrogen chloride. The LD_{50} (rat, oral) is 1900 mg/kg (RTECS). MAK values have not yet been established for tantalum oxide, tantalum carbide, and tantalum chloride. Potassium heptafluorotantalate has toxic properties due to its fluoride content; LD_{50} (rat, oral) is 2500 mg/kg (RTECS), and MAK for dusts is 2.5 mg/m^3 total dust (calculated as fluoride) [143].

11. References

1. R. Kieffer, H. Braun: *Vanadin-Niob-Tantal,* Springer Verlag, Berlin – Göttingen – Heidelberg 1963.
2. *Gmelin,* **50,** part B 1, pp. 1 – 4.
3. S. M. Cardonne, P. Kumar, C. A. Michaluk, H. D. Schwartz, *Adv. Mater. Processes* 1992, no. 9, 16 – 20.
4. R. Kieffer, G. Jangg, P. Ettmayer: *Sondermetalle,* Springer Verlag, Berlin – Heidelberg 1971.
5. O. Kubaschewski: *Atomic Energy Review, Tantalum: Physico-Chemical Properties of its Compounds and Alloys,* International Atomic Energy Agency, Vienna 1972.

6. W. Köck, P. Paschen, *J. Met.* (1989) no. 10, 33 – 39.
7. G. L. Miller: *Tantalum and Niobium,* Butterworths, London 1959, pp. 91 – 93.
8. E. Erben, R. Lesser, *Metall (Berlin)* **15** (1961) 679 –686.
9. W. W. Albrecht: "Production, Properties and Application of Tantalum, Niobium and Their Compounds," in P. Möller, P. Cerný, F. Saupe (eds.): *Lanthanides, Tantalum and Niobium,* Springer Verlag, Berlin – Heidelberg 1989, pp. 345 – 358.
10. P. Borchers, G. J. Korinek: "Extractive Metallurgy of Tantalum," *AIME Annu. Meet. 110th* 1981, 97 –108.
11. TIC Tantalum-Niobium International Study Center Brussels, *Int. J. Refract. Hard Met.* **7** (1988) no. 1, 3 – 7.
12. L. D. Cunningham: *Mineral Industry Surveys, Columbium (Niobium) and Tantalum in 1993,* U.S. Bureau of Mines, Washington 1994.
13. R. Gaupp, G. Moteani, P. Möller, *Erzmetall* **36** (1983) 244 – 251.
14. H. C. Starck GmbH & Co. KG, Internal Raw Material Availability Survey, Goslar 1993.
15. TIC Tantalum-Niobium International Study Center Brussels, *TIC Bull.* **62 – 73** (1990 – 1993).
16. T. C. Pool (ed.): *Elements,* Concord Trading Corp., Denver 1992, pp. 12 – 17.
17. *The Economics of Tantalum,* 6th ed., Roskill Information Services, London 1986, p. 2.
18. C. Bird, *Met. Bull. Monthly* 1993, no. 5, 48 – 49.
19. W. Rockenbauer, *Metall (Berlin)* **35** (1981) 584 – 585.
20. W. Rockenbauer, *Metall (Berlin)* **38** (1984) 156 – 159.
21. J. Bonjer: "Proceedings of the Seminar on Production and Marketing of Associated Heaug Minerals from South-East Asian Tin Deposit in Ipoh/Malaysia and Phuket /Thailand," in W. Gocht (ed.): *Intertechnik* **26** (1985) 179 – 206.
22. Kawecki Berylko Ind., US 3 721 727, 1971 (R. A. Gustison).
23. Metallurgie-Hoboken-Overpelt, DE 2 844 914, 1978 (M. C. F. van Hecke, J. Deweck).
24. Hoboken-lez-Anvers, US 2 131 350, 1937 (J. P. Leemans).
25. Kawecki Chemical Comp., US 3 447 894, 1966 (R. A. Gustison, W. Lawn, F. Gorcyca, J. A. Cenerazzo).
26. Kawecki Berylko Ind., US 3 585 024, 1968 (J. A. Cenerazzo, C. E. Mosheim, C. E. Marvasi).
27. R. Kieffer, F. Benesovsky: *Hartstoffe,* Springer Verlag, Wien 1963.
28. Western Electric Co., US 3 697 255, 1972 (W. M. Baldwin, E. O. Fuchs, D. J. Sharp, J. H. Swisher).
29. H. C. Starck, DE 2 133 104 C 3, 1978 (H. Meyer).
30. K. Mimura, T. Takahashi, M. Nanjo, *J. Min. Mater. Proc. Inst. Jpn.* **106** (1990) no. 4, 187 – 192.
31. N. Sato, T. Takenaka, Y. Wei, M. Nanjo, *Tohoku Daigaku Senko Seiren Kenkyusho Iho* **46** (1990) no. 2, 96 – 104.
32. GTE Products Corp., US 4 255 397, 1981 (B. E. Martin, J. E. Ritsko, H. L. Acla).
33. G. P. Power, *Chem. Aust.* **46** (1979) no. 7, 303 –307.
34. VEB Kombinat Metallaufbereitung, DD 207 932, 1984 (G. Beyer, H. Bernhardt, H. Dam, H. K. Kiesling, W. Tomalik, A. Zimmermann).
35. NRC Inc., US 2 950 185, 1960 (E. D. Hellier, G. L. Martin).
36. NRC Inc., US 4 141 720, 1979 (H. Vartanian).
37. H. C. Starck, US 3 635 693, 1969 (H. J. Friedrichs, H. Meyer).
38. H. C. Starck, DE 3 130 392, 1983 (W. W. Albrecht, U. Papp).
39. Fansteel, US 4 356 028, 1982 (V. T. Bates).
40. H. C. Starck, DE 3 140 248, 1983 (W. W. Albrecht, D. Behrens).
41. NRC Inc., US 4 957 541 (T. B. Tripp).
42. W. W. Albrecht, D. P. Ingalls: "Hochtemperaturreinigung von Tantal und Niob," in Gesellschaft Deutscher Metallhütten- und Bergleute (eds.): *Raffinationsverfahren in der Metallurgie,* Verlag Chemie, Weinheim 1983, pp. 229 – 242.
43. *Gmelin* **50,** part A 2, pp. 490 – 491.
44. C. A. Hempel: *Rare Metals Handbook,* 2nd ed., Reinhold Publ. Corp., London 1961.
45. Fansteel, US 4 859 257, 1989 (V. T. Bates, C. Pokross).
46. E. Raub, E. Röschel, *Z. Metallkd.* **53** (1962) 93 – 103.
47. G. Jangg, R. Eck, *Metall (Berlin)* **31** (1977) 750.
48. G. D. Corey, *Proc. Int. Symp. Tantalum 1st* 1978, 61 – 82.
49. F. J. Hunkeler: "Properties of Tantalum for Applications in the Chemical Process Industry," in R. E. Smallwood (ed.): "Refractory Metals and Their Industrial Applications," *ASTM Spec. Tech. Publ.* **849** (1984) 28 – 49.

50. M. Hörmann, *Metall (Berlin)* **42** (1988) 400 – 406.
51. U. Draugelates, B. Bouaifi, H. Steinberg, *Werkst. Korros.* **44** (1993) 269 – 273.
52. R. L. Taylor, H. E. Haring, *J. Electrochem. Soc.* **103** (1956) 611 – 613.
53. Tantalum-Niobium International Study Center, *TIC Bull.* **73** (1993) 2 – 3.
54. M. Fernández, J. Baonza, J. M. Albella, J. M. Martínez-Duart, *Electrocomponent Sci. Technol.* **7** (1981) 205 – 210.
55. Sprague Electric Comp., US 4 781 802, 1988 (E. J. Fresia).
56. G. P. Klein, N. J. Jaeger, *J. Electrochem. Soc.* **117** (1970) 1483 – 1494.
57. T. Kudoh, M. Katoh, M. Watase, *Denki Kagaku oyobi Kogyo Butsuri Kagaku* **40** (1972) 701 – 705.
58. D. M. Smyth, G. A. Shirn, T. B. Tripp, *J. Electrochem. Soc.* **110** (1963) 1264 – 1270.
59. D. M. Smyth, T. B. Tripp, *J. Elechtrochem. Soc.* **110** (1963) 1271 – 1276.
60. D. M. Smyth, G. A. Shirn, T. B. Tripp, *J. Electrochem. Soc.* **111** (1964) 1331 – 1336.
61. D. A. Vermilyea, *J. Appl. Phys.* **36** (1965) 3663 – 3671.
62. D. A. Vermilyea, *J. Electrochem. Soc.* **112** (1965) 1232 – 1234.
63. M. Burnham, Hughs Aircraft Company Report, no. M 74-08, Culver City, Cal. 1974.
64. K. Andersson, H. Naito, T. B. Tripp, *Proc. 6th Europ. Capacitor and Resistor Tech. Symp. CARTS-Europe,* Brugge 1992, pp. 1 – 6.
65. Tantalum-Niobium International Study Center Brussels, *TIC Bull.* **69** (1992) 6 – 8.
66. D. B. Plone, *J. Am. Osteopath Assoc.* **73** (1974) no. 4, 635 – 640.
67. P. J. Friedman, G. M. Tisi, *Radiology (Easton, Pa.)* **104** (1972) Sept., 523 – 535.
68. E. Kammler, W. Weller, W. T. Ulmer, E. Bruckmann, *Pneumonologie* **146** (1972) 241 – 249.
69. G. T. Daughters et al., *J. Thorac. Cardiovasc. Surg.* **104** (1992) no. 4, 1084 – 1091.
70. T. Friden, L. Ryd, A. Lindstrand, *Acta Orthop. Scand.* **63** (1992) no. 1, 80 – 84.
71. B. Tjornstrand, G. Selvik, N. Eglund, A. Lindstrand: *Arch. Orthop. Trauma Surg.* **99** (1981) no. 2, 73 – 81.
72. L. Rabenseifner, W. Küsswetter, P. H. Wünsch, M. Schwab, *Z. Orthop.* **122** (1984) 349 – 355.
73. B. Heinrich et al., *Z. Stromatol.* **75** (1977) no. 6, 214 –224.
74. P. Tetsch: *Enossale Implantationen in der Zahnheilkunde,* 2nd ed., Hanser Verlag, München 1991, pp. 46 – 47.
75. F. Kylberg, *Acta Chir. Scand.* **141** (1975) no. 3, 242 –244.
76. F. Magistris, *Wien. Klin. Wochenschr.* **86** (1974) no. 8, 225 – 228.
77. F. Magistris, *Chirurg* **46** (1975) no. 11, 518 – 521.
78. H. Kobayashi et al., *Neurol. Res.* **8** (1986) no. 4, 221 –224.
79. S. A. Wexler, B. R. Frueh, D. C. Musch, M. A. Pachtman, *Ophthalmology (Philadelphia)* **92** (1985) no. 5, 671 – 675.
80. H. J. Gerhardt, G. Muhler, D. Szdzuy, F. Biedermann, *Zentralbl. Neurochir.* **40** (1979) no. 1, 85 – 94.
81. J. Carleone, *TIC Bull.* **71** (1992) 5 – 10.
82. R. B. Flanders, *Chem. Eng.* (1979) Dec. 17, 109 – 110.
83. M. Schussler, *J. Refr. Hardmet.* **2** (1983) no. 2, 67 –70.
84. A. Robin, M. E. de Almeida, C. A. Nunes, *Corrosion (Houston)* **47** (1991) no. 6, 443 – 448.
85. *Chem. Eng. News* **42** (1964) no. 38, 47.
86. F. F. Schmidt, E. S. Bartlett, H. R. Ogden, *Trans. Soc. Min. Eng. AIME* **227** (1963) 856 – 864.
87. *Batelle Tech. Rev.* **12** (1963) no. 1, 12 – 13.
88. H. R. Babitzke, J. G. Croeni, *Rep. Invest. US Bur. Mines* **7116** (1968) 1 – 14.
89. *Batelle Tech. Rev.* **13** (1964) no. 2, 12.
90. E. F. Bradley: *Superalloys – A Technical Guide,* ASM International, Metals Park, Ohio, 1988, pp. 10 – 14.
91. E. W. Ross, K. S. O'Hara: "René' 142: A High Strength, Oxidation Resistant DS Turbine Air-Foil Alloy," in: "Superalloys 1992," *Proc. 7th. Int. Symp. Superalloys* **1992,** 257 – 265.
92. A. D. Cetel, D. N. Duhl: "Second Generation Columnar Grain Nickel Base Superalloy," in: "Superalloys 1992," *Proc. 7th Int. Symp. Superalloys* **1992,** 287 – 296.
93. Z. Meng, G. Sun, M. Li: "The Strengthening Effect of Tantalum in Nickel-base Superalloys," in: "Superalloys 1984," *Proc. 5th Int. Symp. Superalloys* **1984,** 563 – 572.
94. G. M. Jankowski, R. W. Heckel, B. J. Pletka, *Metall. Trans. A* **17 A** (1986) 1891 – 1904.
95. Central Glass Co., JP 04 243 935, 1992 (N. Takeuchi, T. Ito, M. Takayama, K. Furuyo, H. Nakajima, H. Iida).
96. F. Izumi, H. Kodama, *J. Less-Common Met.* **63** (1979) 305 – 307.
97. H. C. Starck, DE 4 021 207 A 1, 1990 (W. Bludßus, J. Eckert).

98. T. Ichimura: "Niobiumoxide in Optical Glass Manufacture," in H. Stuart (ed.) *Niobium Proc. Int. Symp.* **1984,** 603 – 614.
99. Mitsubishi Chem. Ind., JP 63 051 945, 1986 (H. Wada, T. Ushikubo).
100. Bayer, DE 4 133 675, 1991 (O. Immel, H. J. Buysch, G. Darsow).
101. Bayer, EP 433 811, 1991 (O. Immel, H. Waldmann, R. Braden).
102. Mitsubishi Kasei Corp., JP 01 050 843, 1989 (H. Wada, T. Ushikuba).
103. Mitsubishi Kasei Corp., JP 63 225 329, 1988 (T. Maki, T. Yokoyama, Y. Sumino).
104. H. J. Lunk, S. Schönherr, *Z. Chem.* **27** (1987) 157 –170.
105. B. C. Grabmaier, F. Otto, *J. Cryst. Growth* **79** (1986) 682 – 688.
106. S. Miyazawa, H. Iwasuki, *J. Cryst. Growth* **10** (1971) 276 – 278.
107. S. Miyazawa, H. Iwasuki, *Rev. Electr. Commun. Lab.* **21** (1973) 374 – 383.
108. AGFA-Gevaert, EP 0 520 094 A 1, 1992 (D. Philip).
109. O. Renoult, J. B. Boilot, F. Chaput, *J. Am. Ceram. Soc.* **75** (1992) 3337 – 3340.
110. S. Kawashima, M. Nishida, I. Ueda, H. Ouchi, *J. Am. Ceram. Soc.* **66** (1983) 421 – 423.
111. D. A. Sagala, S. Nambu, *J. Am. Ceram. Soc.* **75** (1992) 2573 – 2575.
112. K. Tochi, *J. Am. Ceram. Soc. Jpn. Int. Ed.* **100** (1992) 1441 – 1442.
113. J. K. Gibson, *J. Fluorine Chem.* **55** (1991) no. 3, 299 – 311.
114. A. I. Papov, V. F. Sukhoverkov, N. A. Chumaevskii, *Zh. Neorg. Khim.* **35** (1990) 1111 – 1122.
115. E. I. Du Pont de Nemours, WO 92/03382, 1990 (M. Nappa).
116. Asahi Glass Co., WO 9 008 754 A 2 090 809, 1990 (S. Morikawa, S. Samejima, H. Okamoto, K. Ohnishi, S. Tatematsu, T. Tanuma, T. Ohmori).
117. Ciba, DE 1 066 194, 1957 (W. Scheller).
118. Ciba, DE 1 056 105, 1956 (F. Kern, W. Schornsein).
119. D. C. Bradley, B. N. Chakravarti, W. Wardlaw, *J. Chem. Soc.* **52** (1956) 2381 – 2384.
120. R. Gut, *Helv. Chim. Acta* **47** (1964) 2262 – 2278.
121. N. Sato, M. Nanjo, *High Temp. Mater. Proc.* **8** (1988) 39 – 46.
122. M. A. Nicolet, *Thin Solid Films* **52** (1978) 415 –443.
123. J. R. Shappirio, *Solid State Technol.* **28** (1985) 161 –166.
124. E. Jiang, C. Sun, J. Li, Y. Liu, *J. Appl. Phys.* **65** (1989) 1659 – 1663.
125. K. I. Portnoi, V. M. Romashov, *Sov. Powder Metall. Met. Ceram. (Engl. Transl.)* **11** (1972) 378 – 384.
126. S. P. Murarba, D. B. Fraser, W. S. Lindenberger, A. K. Sinha, *J. Appl. Phys.* **51** (1980) 3241 – 3245.
127. K. Hieber, F. Neppl, *Thin Solid Films* **140** (1986) 131 – 135.
128. R. Fresenius, G. Jander: *Handbuch der analytischen Chemie,* **Part 3**, "Quantitative Bestimmungs- und Trennungsmethoden," vol. 5 b, "Elemente der 5. Gruppe, Vanadin, Niob, Tantal," Springer Verlag, Berlin 1957.
129. R. W. Moshier: *Analytical Chemistry of Niobium and Tantalum,* Pergamon Press, New York 1964.
130. J. M. Gibalo: *Analytical Chemistry of Niobium and Tantalum,* Ann Arbor, London 1970.
131. O. G. Koch, G. A. Koch-Dedic: *Handbuch der Spurenanalyse,* 2nd ed., **Part 2,** Springer Verlag, Berlin – Heidelberg – New York 1974.
132. S. Kallmann, H. Obertin, R. Liu, *Anal. Chem.* **34** (1962) 609 – 613.
133. A. H. Knight, *TIC Bull.* **45** (1964) 7 – 8.
134. Autorenkollektiv: *Methoden zur Bestimmung von Spurenanalyse in Niob, Tantal und Wolfram,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1970.
135. J. Stummeyer, G. Wünsch, *Fresenius J. Anal. Chem.* **342** (1992) 203 – 206.
136. O. Hilmer, J. Peters, A. Breustedt, A. Hoppe: "Bestimmung des Phosphors in Refraktärmetallen und deren Verbindungen," in D. Hirschfeld (ed.): *Nichtmetalle in Metallen '90,* Deutsche Gesellschaft für Metallkunde (DGM), Oberursel 1990, pp. 85 – 92.
137. V. Krivan et al., *Fresenius J. Anal. Chem.* **341** (1991) 550 – 554.
138. R. Caletka, R. Hausbeck, V. Krivan, *J. Radioanal. Nucl. Chem.* **120** (1988) no. 2, 319 – 333.
139. N. Jakubowski, D. Stuewer, G. Tölg, *Int. J. Mass Spectrom. Ion Processes* **71** (1986) 183 – 197.
140. *Met. Week* **5** (1991) 21.
141. W. Gocht: *Handbuch der Metallmärkte,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1991, pp. 401 – 413.
142. I. Salisbury, *Proc. 5th Europ. Capacitor and Resistor Tech. Symp. CARTS-Europe,* Munich 1991, pp. 81 –88.

143. H. C. Starck GmbH u. Co. KG, Sicherheitsdatenblätter für Tantal-Metall, Tantalcarbid, Tantaloxid, Tantalpentoxid, Tantalchlorid und Kaliumtantalfluorid, Goslar 1994 (in press).

Tapes, Adhesive

Werner Karmann, tesa Division, Beiersdorf AG, Hamburg, Federal Republic of Germany

Andreas B. Kummer, medical Division, Beiersdorf AG, Hamburg, Federal Republic of Germany

1. Introduction 1
2. Pressure-Sensitive Adhesion 2
2.1. Mechanism of Pressure-Sensitive Adhesion 2
2.2. Dynamic Mechanical Analysis . . . 4
3. Test Methods 4
4. Raw Materials 6
4.1. Elastomers 6
4.2. Tackifier Resins 8
4.3. Plasticizers 8
4.4. Fillers 8
4.5. Stabilizers 9
4.6. Release Coatings 9
5. Manufacturing Processes 10
5.1. Compounding 10
5.2. Coating 12
5.3. Drying 12
5.4. Cross-linking / Curing 13
5.5. Slitting, Cutting 14
6. Adhesive Tape Products 14
6.1. One-Sided Adhesive Tapes 14
6.1.1. Film Tapes 14
6.1.2. Fiber-Reinforced Tapes 16
6.1.3. Paper Tapes 16
6.1.4. Textile Tapes 16
6.1.5. Other Backings 17
6.1.6. Electrical Tapes 17
6.2. Double-Sided Adhesive Tapes 17
6.2.1. Standard Double-Sided Tapes 18
6.2.2. Transfer Tapes 18
6.2.3. Adhesive Pads and Strips 18
6.2.4. High-Performance Tapes 19
6.3. Medical Tapes 19
7. References 20

1. Introduction

The first pressure sensitive adhesives were based on natural rubber and rosin. They were applied as patches and tapes for surgical use [1]. In the 19th century some early technical tapes were derived therefrom (e.g., Citoplast, P. Beiersdorf & Co., Germany), but the application range was very narrow. When spray painting was introduced in the automotive industry, such tapes were employed to mask color lines. They were imperfect and had to be improved [2]. This was the starting point for the development of industrial tapes.

Currently there are hundreds of varieties of pressure-sensitive tapes that differ in the adhesive base, the carrier material, and construction. They are used in practically all industries, in house and office, in hospitals and personal care.

Adhesive tapes generally consist of a flexible carrier or backing with a layer of a permanently tacky adhesive which sticks to most substrates by contact under pressure. There is no need for solvent or heat. The carrier often executes the function of the tape, e.g., to close a box, to protect or decorate a surface, to fill a gap, or to cover a wound.

The main types of adhesive tape are:

- *Pressure-sensitive tapes* consist of a carrier, coated, with adhesive on one side, and wound into a roll. The back side is often coated with a release agent to control the unwind force.
- *Double-sided tapes* generally consist of a carrier with adhesives on both sides and an intermediate release liner, to prevent contact of the adhesive layers when the tape is wound into a roll. There are also transfer tapes without carrier and tapes without release liner.
- *Adhesive pads* are pieces of thick layers of adhesives, or polymer foams with adhesive layers on both sides. They are often stretchable for ease of release from the surface.
- *Medical tapes* are composed of a backing coated with adhesive (porous, permeable or occlusive). Some have a wound pad.
- *Labelstock* consists of a writable or printable carrier coated with the adhesive on one side, which generally is covered by a release liner. This type is not treated here.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_085

2. Pressure-Sensitive Adhesion

Normally adhesives are applied in a liquid state to the substrate, wet the surface, and solidify to generate a bond, for example by evaporation of a solvent, cooling of a hot melt, or formation of a cross-linked polymer. Pressure-sensitive adhesives do not change their physical state, but they also must wet the surface of the substrate like a liquid and sustain loads like a solid. There have been many discussions and investigations concerning how they fulfill these conflicting demands.

The first theories were based on the observation of two phases at the surface of rubber–resin adhesives. One, having high resin content, was assumed to act as sticky spots that adhere to the surface, while the other, with a low resin content, gave cohesion to the system [3,4]. But some contradictions remained unresolved [5].

Later investigations showed that combinations of rubbers with very effective tackifiers do not show the two-phase structure. Therefore, phase separation cannot be essential to pressure-sensitive adhesion. The above-mentioned conflict rather is resolved by the dependence of mechanical strength on the rate of deformation, that is, viscoelasticity. The tackifier resins decrease the modulus at the slow rate at which wetting occurs, and increase it at the higher rate at which the adhesion tests are carried out [6]. This is demonstrated in Figure 1, in which the mechanical strength is characterized by the storage modulus G' (the explanation is given later).

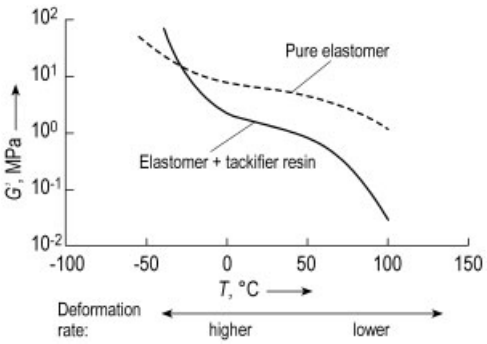

Figure 1. Effect of tackifier resin on the storage modulus of an elastomer [7]

To achieve a good compromise between wetting the surface (adhesion) and bearing a load (cohesion) the modulus should be in the range of 0.1 to 1 MPa "in the time frame of the bonding process", that is 1 – 100 s (Dahlquist criterion [9, p. 101]).

According to Williams – Landel – Ferry (WLF) principle longer time is equivalent to higher temperature. Thus at higher rates the modulus is shifted to the direction of lower temperature and vice versa [8].

2.1. Mechanism of Pressure-Sensitive Adhesion

Pressure-sensitive adhesion is a complex phenomenon determined by a number of physical processes. A comprehensive overview is given in [9].

Pressure-sensitive adhesion is determined by two key factors. First, the surface must be contacted to such a degree that molecular attraction forces come into action (wetting). The second factor is the energy of deformation of the viscoelastic adhesive until rupture or separation from the surface occurs.

The degree of wetting is controlled by thermodynamic and kinetic factors. The driving force for wetting a substrate is dependent on the surface energies of the adhesive and the substrate. The Zisman concept of critical surface tension provides the basis for estimating the surface energy of solids [10]. Typical values of critical surface tension (in mN/m) for polymer substrates follow [14]:

Polytetrafluorethylene	19.5
Poly(dimethylsiloxane)	22
Polypropylene	29
Polyethylene	34.5
Poly(methyl methacrylate)	41
Poly(vinyl chloride)	41
Polystyrene	42
Poly(ethylene terephthalate)	42.5
Polyamide 66	44

The influence of critical surface tension on peel adhesion, tack, and residues was investigated experimentally for rubber – resin [11,12] and polyacrylate [13] pressure-sensitive adhesives (PSAs).

The critical surface tension of a PSA cannot be determined directly, for when the test droplets of low polarity liquids are dissolved in the adhesive the contact angle is not defined. Therefore, this value was derived from those of chemi-

cally similar polymers, resulting in 33 mN/m for polyisobutylene and 35 – 36 mN/m for standard acrylic and rubber – rosin PSAs.

The study showed a maximum of peel force and residues on substrates having a somewhat higher surface tension than the adhesive. Hence, to have good adhesion, the adhesive should have low surface energy or the substrates should be treated to increase their critical surface tension.

Adhesion is enhanced further by contact time and contact pressure. Both increase the area of the wetted spots kinetically by flow of the adhesive. This is especially important for rough surfaces and cross-linked adhesives.

The role of the deformation energy can best be demonstrated by peeling a pressure-sensitive tape from a rigid substrate. During separation typically the adhesive becomes filamented and the "legs" are stretched to such degree that they break or separate from the surface. The former is a cohesive and the latter an adhesive failure. The energy required to achieve this is the peel force.

The correlation between peel adhesion and the viscoelastic properties of the adhesive is reviewed in [9, p. 61 ff]. A comprehensive experimental analysis can be found in [15]. The peel force is measured at various pulling rates and temperatures and superimposed on a master curve [8].

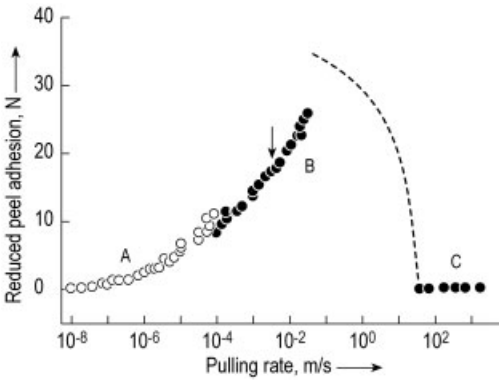

Figure 2. Master curve for the force of a tape peeled at 90 ° from a glass surface (transposed to 23 °C according to WFL principle [8])
○ Cohesive failure; ● Adhesive failure [PET film with adhesive (60 % rubber, 40 % β-pinene resin)] [15]

At low rates (i.e., at high temperatures) there is a region A of cohesive failure with a clear influence of pulling rate.

In region B at higher rate (i.e., lower temperature) the adhesive is still legging, but separation occurs at the surface. In this region the magnitude of the peel force is influenced by the area of molecular contact (i.e., by wetting). At still higher rates is region C, where the adhesive is pulled from the backing. Here the peeling force is dramaticly lower and independent of the pulling rate.

Between B and C is a transitional region, in which the peel force jumps from a higher to a lower level, so-called slip-stick. The upper limit is shown by the dashed line. It is believed that this is a periodical change from rubbery to glassy response during peeling.

Typically, pressure-sensitive tapes are applied in region B. The effective pulling rate of the standard peel test at 180° is 2.5×10^{-3} m/s, which is in the middle of region B (indicated by an arrow in Fig. 2). At very low temperature, region C may be reached. For application under freezer conditions, special tapes with lower T_g are available.

To avoid cohesive failure at higher temperature the adhesive can be cross-linked. This shifts region A to lower rates (i.e., higher temperature) or suppresses it entirely.

Viscoelastic properties are influenced essentially by the glass transition temperature T_g.

The optimal T_g is about 40 – 70 °C below the application temperature of the PSA. The T_g can be adjusted by means of the composition of the adhesive, especially by the concentration and the softening temperature of the resins.

The viscoelasticity and surface tension of adhesives are physical properties that are independent of the specific chemical composition. Therefore, the relations discussed for rubber-based adhesives are also valid for polyacrylates [16] and other elastomeric bases.

2.2. Dynamic Mechanical Analysis

The viscoelastic properties of a polymeric material can be described by its reversible and irreversible responses to deformation. These can be identified most easily by dynamic mechanical analysis (DMA). Usually, the adhesive is placed between two parallel plates, one of which is oscillating sinusoidally, and the torque is measured. From the amplitude and the phase shift

of the sinusoidal stress – strain curve, the elastic component, which is in phase, and the viscous component, which is 90° out of phase, can be derived [9, p. 158 ff].

Thus, the viscoelastic properties are characterized by the storage modulus G' (the elastic component of the modulus), the loss modulus G'' (the viscous component of the modulus), and $\tan\delta = G''/G'$ (the tangent of the phase shift between stress and strain), which are functions of the deformation rate and temperature. The method offers valuable information for the development and the application of adhesive tapes.

The maximum of $\tan\delta$ as a function of temperature is generally identified as the (dynamically defined) glass transition temperature T_g, which is dependent on the deformation rate. Usually it changes by 6 – 7 °C per decade of frequency [17, p. 253].

A mixture of two compatible polymers with different T_g values exhibits a single T_g which is influenced by the T_g values and weight fractions of the two components. In many cases this T_g can be calculated approximately by the Fox equation (Eq. 1) [17, p. 277 ff]

$$1/T_{g_m} = w_1/T_{g_1} + w_2/T_{g_2} \quad (1)$$

where T_{g_m}, T_{g_1}, T_{g_2} are the glass transition temperatures of the mixture and of components 1 and 2, and w_1, w_2 are the weight fractions of components 1 and 2.

A mixture of partially compatible polymers exhibits the T_g of at least one pure component and a T_g, which is dependent on the mixture, according to Equation (1).

3. Test Methods

Standard test methods are applied primarily for characterizing adhesive tapes and assuring product quality. To check whether a tape is suitable for a specific application generally requires more specific tests, which may be agreed with the customer.

Standard test methods for pressure-sensitive tapes are published by governmental and commercial organizations:

USA	Pressure Sensitive Tape Council (PSTC) [18]
USA	American Society for Testing Materials (ASTM) [19]
Europe	Association des Fabricants Europeens de Rubans Auto-adhesifs (A.F.E.R.A.) [20]
Europe	European Committee for Standardization (CEN) [21]

At present globally approved methods are being developed [22].

Adhesive properties of pressure-sensitive tapes are characterized primarily by three parameters: peel adhesion, shear adhesion and tack.

Peel adhesion is a highly characteristic property of a PSA. It gives the force per unit width required to remove a PSA tape by peeling from a test panel under defined separation speed and angle (0.3 m/min and generally 180° (Fig. 3), sometimes 90°). The backing can influence the results by its stiffness or stretchability. A more detailed discussion is given in [9, p. 61 ff].

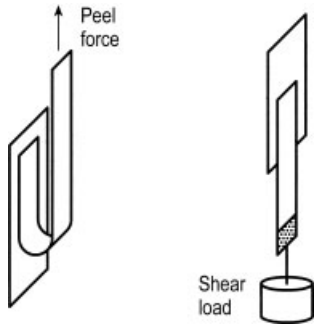

Figure 3. Standard test methods for peel and shear adhesion of pressure-sensitive adhesive tapes

Shear Adhesion (Creep). The most significant example of creep is a tape gliding down under shearing load. Therefore, most commonly creep is characterized by the time a tape holds under defined shear load (Fig. 3). This often is called holding power or shear adhesion. Further examples of creep are the telescoping of rolls of stretchable pressure-sensitive tapes and the greasy spots occurring on a paper sheet mended by a transparent adhesive tape.

Resistance to long-term creep is critical for most applications of pressure sensitive products. According to WLF principle [8] long time is equivalent to high temperature.

The creep of a PSA can be minimized by increasing the molecular mass of the adhesive and by cross-linking it. But there are limitations, because some creep is necessary to wet the surface and to relax stress concentrations when the adhesive is stressed. This often is referred to as

the balance between adhesion and cohesion. The relevance of creep to PSAs is discussed in depth in [9, p. 97 ff].

To understand the reason for a failure caused by creep it is important to identify whether it is a cohesive failure (nearly the same amount of adhesive on both sides) or an adhesive failure (residue on one side only). In the first case, the adhesive is too soft or undercured, in the second it may be too hard or overcured. Furthermore, there may be a weak boundary layer on one surface. Sometimes there are structured patterns of adhesive on the surface. This is an indication of fracture as a result of stress concentrations [9, p. 115 ff].

Tack is the ability of a PSA to create a bond spontaneously under no or minimal external con-tact pressure. There are several aspects of tack, derived from different applications. Some consider the shortest time of contact, others the lowest contact pressure; some consider smooth, others rough surfaces.

The test methods can be classified into three groups (Fig. 4) [9, p. 38 ff]:

1) Rolling ball and rotating drum tests, in which the substrate runs on the surface of the adhesive. The measured quantity is the distance traveled or the imposed force, respectively.
2) Peel tests involving low contact pressure and short application time (e.g., the loop test).
3) Probe test, in which the tip of a probe is brought into contact under low pressure for a short time. The probe is withdrawn at a defined speed, and the butt tensile strength is measured.

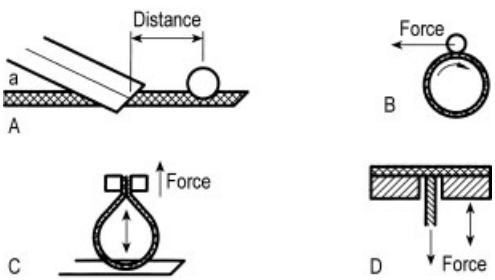

Figure 4. Standard test methods for the tack of pressure-sensitive adhesive tapes
A) Rolling-ball tack; B) Rotating-drum tack; C) Loop-tack; D) Probe-tack

Shear adhesion failure temperature (SAFT) gives information on the resistance of the PSA to stress at elevated temperature. The geometry is identical to the shear adhesion test (Fig. 4). The test is performed in a forced-air oven with temperature rising at constant rate (0.4 °C/min). The SAFT is the temperature at which the bond fails and the load falls down. As this method is not standardized yet, there are some differences in overlap and load between companies [23, 24].

Further tests on adhesive and physical properties of tapes are:

- The unwind force of pressure-sensitive tapes
- The influence of aging on the unwind force of tapes
- Stains left on a substrate after removing the tape
- Strength and elongation of tapes under load
- Water and solvent resistance

For electrical tapes there are a number of very specific tests. More information is given in Chapter 6.

4. Raw Materials

The key component of PSA is the elastomer. It has the main impact on the physical and chemical properties of the adhesive. Its glass transition temperature can be raised either by adding sufficiently compatible resins or by copolymerization with suitable monomers. The composition of the adhesive also influences its surface energy, which is important for the ability to wet a surface.

4.1. Elastomers

Polydiene Rubbers. Most important of these is natural rubber (NR). Additionally used are butadiene – styrene rubbers (SBR) and synthetic polyisoprene rubbers (IR). They have very high molecular masses (several 10^5 to some 10^6 g/mol) and low T_g (– 70 to – 40 °C).

NR is classified internationally according to its production process, purity and Mooney viscosity [25, pp. 12 – 17]. From this, information

about the molecular mass of the adhesive can be derived, which is further influenced by the mastication and mixing process.

To achieve high tack and peel adhesion the T_g of the rubber must be raised by adding tackifier resins. The ratio of resins to rubber generally covers the range of 0.75 – 1.2.

Due to their main chain double bond, polydiene rubbers are prone to oxidative degradation and can be cross-linked by reactive resins and sulfur compounds [25, pp. 233 – 261]. Cross-linking can be achieved further by an electron beam (EB) and in presence of photoinitiators by UV radiation [26].

The adhesives are applied mainly from solvents, some are applied solventless by calendering or as water-based latices. More recently, solvent-free hot-melt processes have been introduced [27].

Styrenic block copolymers (→ Thermoplastic Elastomers, Chap. 5.) are special polydiene elastomers (polyisoprene or polybutadiene) of moderate molecular mass (7×10^4 to 2.5×10^5) with end blocks of polystyrene [(7 – 25) $\times 10^3$ g/mol]. They are thermoplastic elastomers: At low temperature the polystyrene blocks form spherical hard domains that act as cross-links in the adhesive and result in very high cohesion. Above about 90 °C they soften to give a melt of comparatively low viscosity like a thermoplastic [9, pp. 317 – 373].

There are different types of block copolymers: linear ABA, AB, and branched $(AB)_n$ with various ratios of A to B (A is polystyrene, and B is polybutadiene, polyisoprene, or their hydrogenated derivatives).

They must be compounded with resins to achieve high tack and peel adhesion. All ingredients must be selected carefully so as to independently control their influence on the elastomeric and the styrenic blocks [28].

Block copolymers with hydrogenated elastomer blocks are much more stable than those with unsaturated ones but they have somewhat limited adhesive properties. New developments will overcome this limitation [29].

After coating, the hot-melt adhesives can be cross-linked to compensate for softening at higher temperatures [30, 31].

Block copolymers are especially useful for hot melt and high-solids PSAs.

Acrylics. Polyacrylates (→ Polyacrylates) are copolymerized from different monomeric acrylates, optionally with some vinyl compounds. By selecting suitable comonomers, the T_g and the surface energy can be adjusted over a wide range. The most important monomers are listed in Table 1.

Table 1. Important monomers for acrylic PSAs

Monomer	T_g *, °C	Polarity
Soft monomers		
n-Butyl acrylate	− 54	low
2-Ethylhexyl acrylate	− 70	very low
Hard monomers		
Methylacrylate	+ 10	medium
Vinylacetate	+ 30	medium
tert-Butylacrylate	+ 43	low
Styrene	+ 100	low
Polar monomers		
Acrylic acid	+ 106	very high

* Quasistatic glass transition temperature of the homopolymer [37]

Further important parameters are the average molecular mass and its distribution, as well as the sequence of the monomers in the chain. These are set by the polymerization process.

Acrylic copolymers with suitably adjusted T_g are inherent pressure-sensitive adhesives, but they can be compounded with reasonably compatible resins to adjust tack, peel, and creep properties.

For most applications acrylic copolymers must be cross-linked. This can be performed by chemical reactions of the functional groups of the comonomers (amine with acid, epoxy with acid, diisocyanate with acid or hydroxyl, metal chelate with acid [32]) or by radicals generated by thermal decomposition of peroxides or by UV or electron beam (EB) radiation [33].

Most often acrylic adhesives are applied from solvent or as water-based emulsions. For special applications, like thick double-sided tapes, a solventless process has been developed in which a partially prepolymerized mixture of acrylic monomers is coated onto a liner and subsequently polymerized by UV radiation to form the adhesive [34]. Moreover, acrylic hot-melt adhesives cross-linkable by UV and EB radiation are being developed [35, 36].

Polyisobutylene and Butyl Rubber. Polyisobutylenes (→ Polyolefins, Chap. 4.) are

polymers of low polarity and low T_g and are available in a wide range of molecular masses. For pressure-sensitive adhesives, polymers with a viscosity average molecular mass M_v of at least 10^6 are used in combination with low molecular mass types.

Butyl rubbers are copolymers of isobutylene with minor amounts of a diene that enable chemical cross-linking via unsaturated sites. The M_v is in the range of $(3-4) \times 10^5$. To improve the possibilities for cross-linking, it can be mixed with halobutyl rubber, which contains the halogen in the α-position to the double bond.

Polyisobutylene rubbers can be compounded with resins to raise the T_g and hence peel adhesion.

Ethylene – Vinyl Acetate Copolymers (EVAc). The majority of EVAc is used for hard tack-free hot-melt adhesives. Pressure-sensitive properties can be achieved with types containing 40 – 60 % vinyl acetate compounded with compatible resins. EVAc-based PSAs are used only for specialty tapes.

Silicone Elastomers. The main component of pressure-sensitive silicone adhesives is a silicone elastomer with a molecular mass of several 10^5 g/mol. There are two types: One is neat poly(dimethylsiloxane) having very low T_g (− 125 °C) [38] and critical surface tension. The other is modified by substitution of 5 – 15 % of the methyl groups by phenyl groups. This results in higher T_g (ca. − 86 °C), higher surface tension, and improved heat stability [39].

High peel adhesion is achieved by compounding with silicone resins. Since they are only partly compatible, this mixture must be stabilized by reaction of some free silanol groups in both components.

To increase cohesion, silicone elastomers can easily be cross-linked via radicals generated by thermal decomposition of peroxides or by electron beam radiation.

Silicone elastomer based adhesives cover wide range of service temperatures (− 70 to ca. 250 °C) and provide excellent aging resistance. Owing to their low surface energy, they are the only adhesives applicable to low-energy substrates like silicone release liners and PTFE.

Normally, the adhesives are applied from aromatic solvents. Newer developments use the noble metal catalyzed addition reaction of Si–H to $H_2C{=}CH-$ for high-solids and solventless silicone PSAs.

4.2. Tackifier Resins

Tackifier resins are generally brittle amorphous materials with high softening temperatures (typically 40 – 150 °C) [40,41] and low molecular masses (some 10^2 to some 10^3 g/mol). Some are liquids that soften below ambient temperature. All should exhibit reasonable compatibility with the elastomer, which can be detected by the clearness of the mixtures or by dynamic mechanical analysis (see Chap. 2). Compatibility can be predicted from the cloud point temperature in suitable mixtures of solvents [42].

The presence of high molecular mass fractions, which often impairs the solubility of a resin, can best be recognized by the Z-average molecular mass M_Z.

The resins have important impact on the aging resistance of the adhesives. Unsaturated sites are susceptible to peroxidation. Most resins reduce the efficiency of radical cross-linking considerably.

Usually the resins are delivered as solids in barrels, as broken pieces, or as pellets or flakes. An increasing number are also available dispersed in water for water-based adhesives, and molten in tank trucks for hot-melt adhesives.

Rosin and Rosin Derivatives. The main component of rosins are rosin acids with two double bonds the molecule. They can be stabilized by hydrogenation or disproportionation. The softening temperature can be varied over a wide range by esterification, from the methyl ester (liquid) to the pentaerythritol ester (ca. 110 °C). Still higher temperatures are achieved by oligomerization. Rosin-based resins have a broad range of compatibilities.

Terpene resins (→ Resins, Synthetic, Chap. 2.) are produced by cationic polymerization of terpenes from turpentine or other sources. Poly(β-pinene) resins have the best compatibility to natural rubber. They are fairly stable, in spite of the double bond in the cyclohexyl ring. Due to increasing demands for terpenes and

decreasing availability, the resins are being substituted progressively by hydrocarbon resins.

Hydrocarbon resins (→ Resins, Synthetic, Chap. 2.) are produced mainly by cationic polymerization ($AlCl_3$ and BF_3) of carefully selected fractions from mineral oil processing. There are two main types: aliphatic resins from C_5 and $(C_5)_2$ fractions, and aromatic resins from C_9 fractions.

The aliphatic resins have good compatibility with natural rubbers and polyisobutylene. The degree of unsaturation, which has an impact on aging and radical cross-linking efficiency, is characterized by the bromine number. Recently an increasing number of aromatically modified C_5 resins with improved adhesive properties in natural rubber and block copolymers are available. Some are hydrogenated to improve aging properties.

Aromatic C_9 resins have good compatibility with SBR and EVAc elastomers, and some of them to acrylate copolymers. They have limited compatibility with natural rubber. For more details, see [9, pp. 527 – 544].

4.3. Plasticizers

Plasticizers (→ Plasticizers) are generally liquid substances of low molecular mass (some 10^2 to a few 10^4) that lower the T_g of the adhesive. Thus, they often reduce peel adhesion at ambient temperature but increase low-temperature adhesion. They increase tack and creep of the adhesive.

Most important groups are:

- Mineral oils, which may be aliphatic, naphthenic or aromatic
- Oligomers with low T_g, such as polybutenes, liquid resins, polyethers, and branched polyesters
- Long-chain alkyl esters of aromatic and aliphatic acids such as phthalic, adipic, sebacic, and trimellitic acids, which are typical PVC plasticizers

4.4. Fillers

Fillers are usually powdered inorganic solids. Primary parameters are the average and distribution of particle size, as well as the particle shape. For very small particles, often the surface area per unit weight is given instead of this. The key factor is the degree of dispersion in the adhesive, which is determined by the surface energies, the mixing or kneading process, and additives.

Fillers increase creep resistance and improve removability of the adhesives. If they are used as dyes, colorimetric parameters are important. For special purposes, fibers, spheres, conductive particles, etc. may be added to PSAs.

The most important groups of fillers are:

- Carbonates, silicates, and silica are the most common fillers. Interaction with the elastomers can be modified greatly by surface treatment with fatty acids or silanes. Pyrogenic silica has an extremely high surface area.
- Zinc oxide promotes chemical cross-linking by phenolic resins and thio compounds.
- Glass and polymer fibers are used to decrease strain under the influence of unwinding stress in unsupported double sided tapes (transfer tapes).
- Glass and polymer microspheres improve the stress distribution in the adhesive layer and the adhesion of thick double-sided tapes [43]. Organic microspheres containing low-boiling liquids can be used to foam an adhesive layer [44].
- Organic and inorganic pigments are used as coloring agents. Most common is titanium dioxide for whitening. Surface treatment can prevent light induced decomposition of the elastomer.

4.5. Stabilizers

In terms of stability the polydiene elastomers are the most critical due to their main-chain double bonds. Oxidative degradation is stimulated by elevated temperatures and the presence of unsaturated resins, typically used as tackifiers.

Elastomers with saturated chains like polyacrylates and polyisobutylenes are considerably more stable.

Primary antioxidants act by trapping radicals and preventing the initiation of oxidative chain reactions. Sterically hindered phenols are widely valued for their low- or nonstaining properties. Secondary stabilizers (e.g., phosphite esters and

special thioesters) decompose peroxides, which are responsible for autocatalytic effects.

Stabilization against UV can be achieved by UV absorbers such as derivatives of benzotriazoles and benzophenones, and by hindered amine light stabilizers (HALS). For more details, see [45].

4.6. Release Coatings

Release coatings are important components of pressure sensitive tapes as they enable the formulator to adjust the release force according to the application, for example, the unwinding force of tapes and the transfer of double-sided tapes to substrates.

The surface energy of a release coating is much lower than that of the adhesive. Thus the degree of wetting and adhesion are reduced. Further factors are the viscoelastic properties of the adhesive, the surface texture, and the number of pores in the release coat. External factors such as contact pressure, contact time and temperature should have as low an influence as possible.

There are three main types of release coatings:

- Release coatings based on long-chain alkyl groups yield medium release forces (0.1 – 1 N/cm). They are primarily used for the backsize of tapes. Examples are polyvinyl – stearyl – carbamate, chromium complex of C_{14} – C_{18} fatty acids [46], and stearyl copolymers [47].
- Acrylic polymers with perfluorated alkyl groups [48] are applied from solvent or as latex [49]. Their use seems to be rather restricted.
- The most common release coatings are silicones. They are based on poly(dimethylsiloxane), yield low to medium release forces (0.01 – 0.3 N/cm), and are applied primarily on release liners.

The key process for silicones is cross-linking, since residues of low molecular mass silicone oil on the surface of the PSA will eliminate tack.

The first reactions used were tin-catalyzed condensation reactions, which could be applied from solvents only. Since ca. 1980, noble metal catalyzed addition reactions have been introduced to coating technology. The combination with retarders gives systems which cross-link efficiently at high temperature (e.g., 10 s at 150 °C) but are fairly stable at ambient temperature. They can be applied without solvents. For more details, see [9, p. 601 – 626].

The influence of high temperatures can be avoided by radiation curing (UV and EB). There are two different processes: one is initiated by radicals [25, p. 355 ff] and the other one by ionic species [52]. Both processes allow the release force to be adjusted by means of additives to achieve controlled release.

5. Manufacturing Processes

Adhesive tapes generally consist of flexible backings with layers of primer, adhesive, and release coatings. Thus, manufacturing adhesive tapes essentially comprises compounding, coating and slitting processes.

The thickness of the layers and the viscosity of the coating materials vary over a broad range. Typical values are listed in Table 2.

Table 2. Thickness and viscosity of adhesive tape coating layers

	Thickness, µm	Viscosity, ∗∗ Pa · s
PSA	10 to 150 ∗	0.5 – 100
Primer	0.5 to 5	0.01 – 1
Release layer, backsize	0.1 to 2	0.001 to 5

∗ For special double-sided tapes, up to 2 mm. ∗∗ Upper values for 100 % systems.

The tolerance in thickness often has to be kept below ± 10 %. Typical line speeds are in the range of 50 to several 100 m/min.

In the beginning of the PSA industry, most compounding and coating was effected in solution. This still is the most versatile process. With increasing demands for environmental protection, solvent recovery and incineration have become widespread. They are rather effective, but require high investment and increased energy costs. Therefore, they are being substituted increasingly by solventless processes.

Coating of fluids such as primers and release agents are well described in the literature. A guide to selection is given by Cohen [53, p. 13 ff]. Important parameters are the required thickness and precision of the layer as well as the desired line speed, the properties of the coating

Table 3. Alternative manufacturing processes for PSA coatings

Process type	Fluidization by	Solidification by	Strong points	Critical points
Solvent-based	solvation	evaporation	most versatile adhesive properties full range of thickness	high cost for energy and equipment residual solvent, especially at higher thickness influence on some backings
Waterborne dispersions	dispersion in water	evaporation and coagulation	thin to medium thickness high cohesion applicable to solvent-sensitive films	low tack on plastics and rough substrates low water resistance electrochemical corrosion
Hot melt	heating	cooling	medium to very high thickness high peel adhesion high cohesion at ambient temperature high speed	low cohesion at higher temperature, (can be overcome by cross-linking) thermal and oxidative degradation
Reactive systems	(liquid components)	curing (chemical or radiation initiated reactions)	thin to very high thickness high cohesion good reversibility	low tack on rough substrates

Table 4. Alternative equipment for different manufacturing processes

Solvent-based	Waterborne	Hot melt	Reactive system
Compounding			
granulator + kneader (solid rubbers)	mixer	kneader (discontinuous)	mixer
mixer (soluble elastomers)		mixer extruder (continuous)	
Coater			
knife over roll	Meyer bar	slot die	reverse roll
reverse roll	air brush	melt roll	slot die
slot die	reverse roll	extrusion roll	
	slot die	calender	
Drying			
dryers + solvent recovery or incineration	dryer (air cleaning often necessary)	none	none
Cross-linking/curing			
high temperature	high temperature	radiation	mixing two components
radiation	radiation		radiation

(e.g., viscosity, solvent content, shear tolerance) and of the substrate (e.g., tolerance of thickness, roughness, absorbency).

Pressure-sensitive adhesives are compounded and coated in the fluid state and used in the solid state. Key characteristics of the process are the fluidization and solidification procedures. Four alternative types of process can be applied (Table 3).

These divergent processes require very different types of equipment. The most important are summarized in Table 4. A comprehensive treatment is given in [9, pp. 355 ff].

5.1. Compounding

Bales of natural rubber can be processed in two ways [33]:

1) With low degradation: Granulating the bales and kneading them together with all other ingredients in solution kneaders with sigma blades. If high-molecular rubber types like SMR-L (Standard Malaysian Rubber, light), ADR (Air Dried Rubber), or crepe are used, the resulting PSAs can be applied for general

purpose tapes without cross-linking. Primary disadvantages are slow batch processing and low solids content or high viscosity of the coating materials.

2) With high degradation: The rubber bales are milled down together with the filler in an internal mixer with intermeshing blades within a few minutes. Subsequently this mill base is dissolved and compounded with tackifiers and further ingredients in a solution mixer. Process time is much shorter, but cross-linking is indispensable even for tapes with standard performance.

Compounding block copolymer based hot-melt adhesives can be performed batchwise in heated kneaders or continuously in mixing extruders [54, 55]. A survey of the designs and parameters of suitable extruders is given in [56].

Compounding of soluble elastomers, dispersions, and liquid components is described in detail in [53, pp. 23 – 61].

5.2. Coating

Standard equipment for solution based PSAs are knife over roll and reverse roll coaters (Fig. 5) [53, p. 13 ff]. For high-speed application, a closed reservoir or a slot die is preferred for better control of air bubbles [58]. Especially the knife over roll coater allows for an easy change of the coating material.

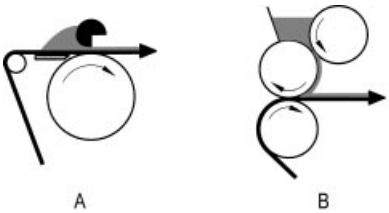

Figure 5. Standard coating equipment for PSAs
A) Knife over roll; B) Reverse roll coater

The coating can be applied directly to the web or to a release liner and subsequently transferred to the web. The latter method is often used for double-sided tapes with foam backings, for it eliminates the influence of solvent and high temperature on the substrate.

For hot melts, roll coaters and die systems are used (Fig. 6). Typical coating temperatures are in the range of 120 – 160 °C.

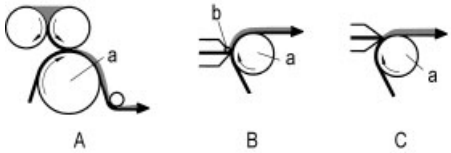

Figure 6. Coating equipment for pressure-sensitive hot-melt adhesives
A) Roll coater; B) Slot die coater; C) Extrusion die coater; a) Chill roll; b) Rotating rod

Roll coaters are largely free from interference caused by particles and variations in web thickness. The coating material can be changed quite easily. However, the influence of oxygen on the adhesive cannot be excluded completely, resulting in higher degradation of polydiene block copolymer hot-melt adhesives.

There are two types of coating dies: In the *slot die* coaters the coating is pressed onto the substrate. The thickness of the layer is controlled by the gap between the die lips and the web, which is backed by a steel or hard rubber roller. At the forward lip there may be a rotating rod to reduce stripes caused by particles. The pressure improves the contact of the adhesive to the substrate. The thickness of the coating is influenced by the web thickness. Further details are given in [59].

In the *extrusion die* coater, the coating layer is formed and calibrated in the slot and laid down on the substrate. The thickness is independent of the web thickness and can be reduced by stretching. There is no penetration through open substrates like fabrics, and there is no influence of particles on the coating. Good contact to the substrates can be provided by an air knife or a suction slit.

5.3. Drying

In solvent- and water-based systems the drying process is often the limiting factor for the line speed. Thus, it determines the economical efficiency of the production line and has to be considered carefully for all products.

Up to coating weights of about 50 g/m^2, forced air drying tunnels (Fig. 7 A) are employed. The air is pressed through holes, slots, or nozzles onto the web. Usually the dryers are segmented to cope with the different drying regimes

[53, pp. 267 – 298]. In single-side dryers, the web is backed by rolls or by a belt. Most efficient are two-sided floater dryers, in which the web hovers between specially designed air nozzles on both sides.

For higher coating weights, festoon dryers (Fig. 7 B) can be used. In this type of dryer, long loops of the web hanging on rods are moved slowly through a sealed room, which may have a volume of more than 500 m^3. The air is pressed through nozzles from the ceiling into the loops. Further details are given in [53, pp. 86 – 91].

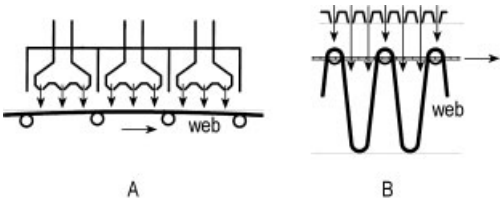

Figure 7. Drying equipment for pressure-sensitive adhesive tapes
A) Drying tunnel; B) Festoon dryer

In all air dryers, the concentration of the solvents must be kept below 50 % of the lower explosion limit (e.g., ca. 20 g/m^3 for hexane and toluene, ca. 40 g/m^3 for ethyl acetate [60]). In high-performance lines huge amounts of air have to be heated, transported and finally purified of solvent, for example, by adsorption processes.

An alternative method is incineration. As the solvent concentration is low, effective heat recovery equipment is necessary to reduce the amount of additional fuel [53, pp. 110 – 115]. Catalytic oxidation often is inhibited by widely used substances like silicones and amines.

Energy consumption can be reduced if oxygen is excluded from the dryers. This can be achieved by substituting the air by nitrogen in a closed loop. In this case the solvent concentration can be greatly increased, so that the flowing gas volumes are reduced considerably and the solvent can be condensed at about – 5 °C. The disadvantages of this method are the limited accessibility of the coaters and drying tunnels and the need to seal the inlet and outlet of the web sufficiently.

5.4. Cross-linking / Curing

Cross-linking can be effected by chemical reactions of reactive groups in the elastomer among one another or with reactive additives. The reaction must not start until after the coating process. This is effected by high temperature, which may be achieved in the last sections of the drying tunnel or in a separate high-temperature tunnel.

Radiation curing is very appropriate to cure PSAs, as the transferred energy is far higher than in thermal cross-linking. Both electron beam (EB) and ultraviolet (UV) radiation are applied. Their physical characteristics differ, as shown in Table 5 [61].

Table 5. Characteristics of EB and UV radiation

	EB radiation	UV radiation
Particle/quantum energy	150 – 300 keV	3 – 6 eV
Generation of radicals	stochastically by all matter	by the photoinitiator (PI) only
Absorption of the radiation	by all matter	by specific components
Penetration depth	dependent on density and particle energy	dependent on wavelength and concentration of components, especially the PI

These differences result in different kinetics. The energy of EB radiation is several tens of thousands times higher than the energy of chemical bonds. A primary electron transfers its energy in a cascade of secondary and tertiary electrons to several thousand molecules, generating reactive sites stochastically in the elastomers, the resins, and all other additives. The radical sites on the elastomers result in cross-linking of the adhesive. In the presence of reactive monomers, there is a high degree of grafting.

The quantum energy of UV radiation is comparable to chemical bond energies. The radiation is absorbed by the photoinitiator, which is converted to free radicals or to a highly reactive state. These sites generate cross-links at the elastomers; in the presence of monomers, they can initiate polymer chains.

The other difference concerns penetration depth. As EB radiation is attenuated by all matter, penetration depth is a function of electron energy and density, e.g., at 300 kV the useful penetration depth at a density of 1 g/cm^3 is 0.35 mm

[62]. In a clear layer of nonabsorbent polymers the penetration depth of UV radiation may increase with the absorbed irradiation energy (i.e., irradiation time) [63]. With low concentration of the photoinitiation, optimal wavelength, and sufficient irradiation time, coatings with a thickness of more than 1 mm can be cured.

Details about equipment and process parameters are given in [64,65].

5.5. Slitting, Cutting

For high efficiency, coating usually is performed over a width of more than a meter. The tape rolls are obtained by cutting the web into strips with a width of a few millimeters up to several centimeters. Four methods are applied (Fig. 8).

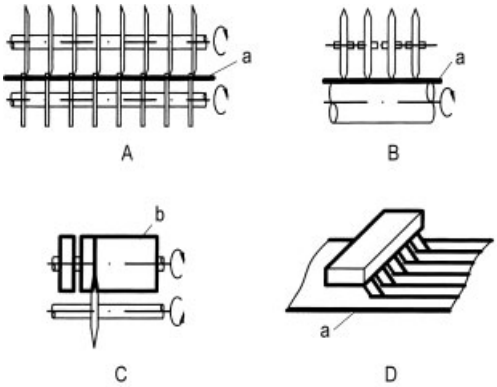

Figure 8. Slitting methods for pressure-sensitive adhesive tapes
A) Shear cutting; B) Crush cutting; C) Log cutting; D) Razor blade cutting
a) Web; b) Log

Shear cutting (Fig. 8 A) is effected between a rotating knife and a roller edge and results in tapes with smoothly cut edges. This method has the highest precision, but the width of the tape rolls cannot readily be changed. It is used for most rigid-film tapes and double-sided tapes.

Crush cutting (Fig. 8 B) is effected by pressing a rotating knife against a steel roller. The edges of the tapes are roughened by small cracks. The precision is lower than that of shear cutting, but the width can be changed more easily. This method is usually applied for paper-, fabric-, and nonwowen backed tapes. In a special case biaxially oriented polypropylene (BOPP) based films are cut by such a process to achieve hand-tearable tapes.

In log cutting (Fig. 8 C) the coated web is wound into rolls or logs in full width and in the desired length of the tape rolls. Subsequently, these logs are sliced by a rotating knife, which is moved transversally between slices to match the width of the tape rolls. As the transverse location can be adjusted deliberately by a computer, this process allows the complete customization of the width of the tape rolls.

In razor blade cutting (Fig. 8 D) the web is cut by parallel razor blades. The edges are cut very smoothly. Due to the lower stiffness of the blades the precision is lower than with the other methods. It is preferred for, e.g., carton-sealing tapes.

6. Adhesive Tape Products

Hundreds of divergent pressure-sensitive adhesive products are available to cope with numerous applications in industry, office, household, etc. They differ in material and/or in design. Most are wound into rolls; some are die-cut into separate pieces or assembled on sheets.

A detailed presentation of all the products and applications is beyond the scope of this article. Reference is made to the *Handbook of Pressure Sensitive Technology* [9].

6.1. One-Sided Adhesive Tapes

Generally such tapes are composed of four layers:

1) Pressure-sensitive adhesive
2) Primer
3) Backing (carrier)
4) Backsize (release coating)

The backsize and the primer are optional components. The former is used to adjust the unwinding characteristics and/or to seal a porous backing material. The latter provides for good bonding between adhesive and backing. In many cases, instead of using a primer the film surface is treated by corona discharge or with a flame. Sometimes the primer also contains an agent which cross-links the adhesive by diffusion into this layer.

6.1.1. Film Tapes

Film tapes have by far the largest production volume. According to the deformation properties there are two kinds of film tapes.

Rigid Film Tapes. *Cellophane* is the oldest transparent tape backing. It is fairly strong and transparent and is widely used to fasten package wrappings, to seal cartons, and for a great variety of other sealing, holding, and mending jobs. However it is not resistant to moisture and has become rather expensive due to environmental protection regulations in the production process.

Cellulose acetate film is resistant to moisture but lacks the toughness of cellophane. In combination with an age-resistant acrylic adhesive and a matte backside it is used in many countries for office tapes to repair torn books, etc.

Considerable amounts of oriented *poly(vinyl chloride)* (PVC) films are used for tapes in Europe. In combination with solvent-based rubber–resin adhesives, they yield high-performance packaging tapes with smooth unwind characteristic without the need for a backsize. *Biaxially oriented polypropylene* (BOPP) tapes with hot-melt adhesives based on styrenic block copolymers have now attained the biggest share of the market for carton-sealing tapes. They require a backsize, and on unwinding they evolve much noise.

By using waterborne acrylic dispersions, UV-resistant packaging tapes with quiet unwinding are produced. Most of them are used as carton-sealing tapes. Further applications are protection of labels and documents.

By stretching PP films predominantly in the machine direction, the tensile strength in this direction can be increased further at the expense of strength in transverse direction. Such tapes are referred to as machine direction oriented polypropylene (MOPP) and tensilized polypropylene (TPP) films and are used for strapping and bundling heavy goods, to secure pallets, and to tear the corrugated cardboard to open boxes. These tapes often are wound into special rolls with lengths of up to 70 000 m by cross-spool winding.

For more information on packaging tapes, see [9, 66].

The upper temperature limit of a film tape is given by the softening of its backing, which may be caused by the glass transition (e.g., PS, PVC films) or by the melting of crystalline regions (e.g., PE, PP, PETP films). The low temperature limit is given by embrittlement of the film, or more often by the glass transition of the adhesive, which is not considered here.

To assess roughly the temperature range of the most important film backings some values are given in Table 6, valid for adhering films at low stress. It should be noted, that for long-term service, the upper limit may be reduced considerably due to degradation processes and interaction with the adhesive. This is treated in more detail in Section 6.1.6.

Table 6. Application temperature range (short-term) of important film backings [67, 68]

Film	Application range, [a], °C	
	min.	max.
Cellophane	0	150
Low-density Polyethylene [b]	− 50	95
High-density Polyethylene [b]	− 40	80
Oriented Polypropylene	0	100
Rigid PVC	− 15	75
Plasticized PVC [b]	− 30	100
Poly(ethylene terephthalate) [c]	< − 100	160
Poly(ethylene naphthenate) [c]		200
Poly(ether ketone)		240
Polyimide	< − 100	350

[a] Temperature range over "which no significant loss in strength or stiffness occurs" [67]. [b] Conformable films. [c] Biaxially oriented.

A comprehensive survey on further film properties is given in [67].

Producers and Trade Names: Asia Chemical Corp., Beiersdorf (tesa tapes), Intertape Polymer Group, 3M (Scotch tapes), Scapa Tapes Klebetechnik (Cellux, Sellotape), Vibac.

Conformable Film Tapes. For many applications the tape should conform to a rugged or profiled surface. There are two types of films, which differ in their tensile stress elongation curve.

Type A tapes are made of semicrystalline polymers like PE. On extending such polymers beyond the yield point, the film deforms irreversibly at nearly constant stress until it finally increases to the breaking point. Type B tapes are made of amorphous plastics like plasticized PVC or rubbery EPM copolymers. Here the stress increases steadily up to the breaking point. It is a rubberlike reversible deformation. There

is always a residual stress in the adhesive bond. Therefore the adhesive needs sufficient cohesion. Typical values are shown in Fig. 9.

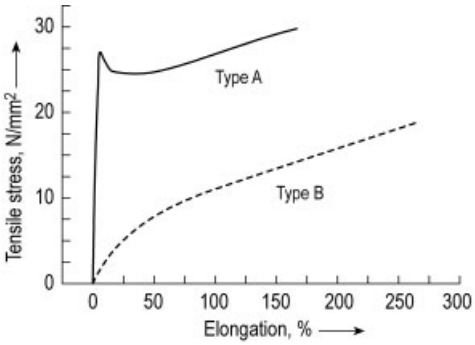

Figure 9. Deformation characteristics of conformable films

Tapes of type B are used for insulating and bundling wires and for cable harnessing. Plasticized PVC usually contains low molecular mass plasticizers, which may migrate into the substrates and adhesive. This can cause swelling of the substrate, and the tape may become brittle due to loss of the plasticizer.

Conformable films of type A are used as film-masking tapes and as thick films for corrosion protection tapes for underground transmission pipes.

Producers and Trade Names: Beiersdorf (tesa tapes), Coroplast Fritz Müller KG, 3M (Scotch tapes), Nitto Denko Corp., Raychem, The Tapecoat Company.

6.1.2. Fiber-Reinforced Tapes

To increase the mechanical strength of a film tape further, filaments can be embedded into the PSA. Most common are glass fibers; alternatives are polyester and polyamides. Such tapes have high tensile strength and low elongation under stress. Further, they are highly resistant to tear propagation, which is a weak point of biaxially oriented films. These tapes are suitable for heavy-duty applications.

Producers and Trade Names: American Tape Company, 3M (Scotch tapes).

6.1.3. Paper Tapes

Paper is a widely used backing for pressure-sensitive adhesive tapes. It is more expensive than simple plastic films but has several advantages. It is easy to tear by hand, tear strength is sufficient for many applications, it does not soften at high temperatures, and it can be upgraded by coating or saturating with polymers.

Flat Kraft Paper Backing. Usually the back side of the kraft paper is coated with PE by extrusion and additionally with a silicone release layer. The other side is coated with the PSA. Tapes with regular kraft paper have low elongation at break (ca. 4 %). This can be increased to ca. 10 % by corrugating the paper using the Clupak technology. Kraft paper tapes are quite common in Japan for carton sealing and general purpose packaging.

Creped paper backings can be divided roughly into two groups.

Thick creped papers with a thickness of more than 300 µm can be used without a backsize coat for high-elongation masking tapes for painting.

Flat creped papers saturated with rubberlike polymer latices generally have a backsize coating to control the unwind characteristic without impairing the adhesion of paints. In combination with natural-rubber-based adhesives, these tapes are used for masking in the automotive industry, for assembling electronic components, and insulation of low-voltage devices etc.

For high-temperature paint curing, the adhesive is cross-linked and thermosetting, i.e., it contains reactive resins, which produce additional cross-links during application in the high-temperature drying tunnels.

Producers and Trade Names: American Tape Company, Beiersdorf (tesa tapes), 3M (Scotch tapes), Nitto Denko Corp., Permacel, Shuford Mills, Sicad.

6.1.4. Textile Tapes

Fabric-based PSA tapes combine low stiffness with high tensile strength and low elongation. Usually they are tearable by hand. The backside can be open, coated with an acrylic finish, or extrusion coated with PE copolymers. *Duct tapes* are produced by combining a rubber-based adhesive, a wide-mesh fabric, and a PE film by calendering in a one-step process. Fabric tapes

are used for a vast number of divergent applications.

For many applications *nonwovens* or *fleeces* are used. Such tapes may be coated with a porous or impervious backsize.

Producers and Trade Names: Beiersdorf (tesa tapes), 3M (Scotch tapes), Nichiban, Scapa Tapes Klebetechnik (Fermoflex, Rotunda).

6.1.5. Other Backings

Foams. Very common are polyurethane foams. Foams based on polyesters have better heat and light stability than polyethers but have inferior stability against moisture. To increase light stability, they may contain carbon black. The PU foams are increasingly being substituted by foams based on PE and PE copolymers. For very high temperatures, silicone-based foams are available. The main applications of foam tapes are sealing gaps and damping vibrations.

Producers and Trade Names: Illbruck Industrieprodukte, Norton Company, Rubber Astic, Venture Tape Corporation.

Metal Foils. Most common is aluminum, less common are copper and lead. Such tapes are quite conformable and are used to seal against moisture and water. Tapes with aluminum foils or aluminum-coated films can shield surfaces against thermal radiation. Copper and aluminum foils and vapor-coated films also can be used to shield against electromagnetic interference.

Producers and Trade Names: Coroplast Fritz Müller, 3M (Scotch tapes).

6.1.6. Electrical Tapes

Electrical tapes comprise different types of backings. Conveniently they are treated in a separate section as they have to fulfill common specific requirements e.g. good electrical properties under dry and moist conditions, long service life at elevated temperatures, and often they need flame resistance and must be free from any substance corrosive to metal surfaces and from any emission harmful to electrical contacts.

These requirements are described by national and international standards like PSTC [18], UL [69], ASTM [19], VDE [70], and DIN [71]. The International Electrotechnical Commission (IEC) [72] has established a worldwide network to unify all national standards and test methods concerning electrical tapes.

As temperature often is the dominant aging factor in electrical insulating systems, certain basic thermal classes have been recognized internationally (Table 7).

Table 7. Thermal classes for electrical tapes for long-term service life at elevated temperature

Thermal class	Temperature, °C	Typical backings
Y	90	stabilized PVC
A	105	saturated paper, cotton fabric, acetate film
E	120	
B	130	poly(ethylene terephthalate) film, fabric, nonwoven
F	155	poly(ethylene naphthenate)
H	180	polyimide, glass fabric

Within each class certain combinations of backings and adhesives are feasible. For long-term service above 150 °C, silicone-based PSAs must be used.

For more details see Bräunling [9, p. 643 ff] and Külper [73].

Producers and Trade Names: 3M (Scotch tapes), Nitto Denko Corporation, PPI.

6.2. Double-Sided Adhesive Tapes

Double-sided tapes usually are used to attach substrates to one another; therefore, the adhesive is of primary importance. Generally, the backing material, if present, must assist the handling.

In high-performance tapes the backing consists of a relatively thick highly compliant cross-linked polymer layer, which may be foamed, and has a major impact on the strength of the bond.

Generally a double-sided PSA tape comprises five layers together with a release liner consisting of three additional layers:

- Pressure-sensitive adhesive 1
- Primer 1
- Backing (carrier)
- Primer 2
- Pressure-sensitive adhesive 2
- Release coating 1
- Release liner

– Release coating 2

Primers are optional, and in case of polyolefins can be substituted by corona, flame, or fluorine treatment.

6.2.1. Standard Double-Sided Tapes

Most double-sided tapes have cross-linked acrylic PSAs, which often are tackified with compatible resins. The adhesives on both sides can be identical or different, for example, to combine reversible and permanent bonding. Some adhesives are based on natural rubber or styrenic block copolymers.

Widely used backings are rigid and conformable films, nonwovens, fabrics, creped paper, and foams based on polyurethanes, polyolefins and their copolymers, etc. Generally the thickness of the adhesive layer has to be increased with increasing roughness of the substrates. Foams are especially useful for rugged or profiled surfaces.

An important element is the interliner, which mainly is coated with a silicone release agent, as described in Chapter 4. Glassine kraft paper is used most frequently. Creped paper, rigid and semirigid films are also used. Besides the peel force, stiffness and conformability are significant features of the interliners. More detailed information is in [9, p. 691 – 705].

Double-sided tapes are used for a vast diversity of applications in industry and household. Further important applications are laminating narrow profiles and wide substrates like foams and thick mats to endow them with pressure-sensitive adhesion. Two special applications are:

1) In the papermaking and printing industry splicing tapes are used to bind the end of one roll of paper to the beginning of another. This must be effected at high speed, and the bond often has to bear high temperature. Moreover, as the tape has to be returned to the papermill for reprocessing, the adhesive and the backing have to be repulpable (i.e., soluble or dispersible in water). Such PSAs are based on highly acidic polyacrylates and compatible water-soluble plasticizers. The dispersibility is adjusted by partial neutralization with alkali [74].
2) Double-sided tapes are used further to fix the printplates onto the printing rollers. The adhesive has to ensure that the tape is mounted easily and the plate stays down during the printing process. After finishing the plate must be easily demountable without residues.
   Very important for the print quality is the type and hardness of the backing: Film backings are mainly used to print solid areas. Foam backings are used for half-tone and combination prints. The backing has to exhibit a high recovery rate to achieve high printing speed with constant quality.

Producers and Trade Names: Beiersdorf (tesa tapes), Lohmann (Duplocoll tapes), MACtac, 3M (Scotch tapes), Nitto Denko Corporation.

6.2.2. Transfer Tapes

In transfer tapes, the PSA is coated on the release liner without additional backing. Therefore the release forces on both sides of the liner have to be precisely different and constant over storage time. The adhesives sometimes contain short-cut glass fibers to decrease the elongation during unwinding. Such tapes are used to transfer the PSA to extruded profiles or other parts to endow them with pressure-sensitive adhesion. For use in household and office, special dispensers are available.

Producers and Trade Names: Avery-Dennison Corporation (Fas tapes), Beiersdorf (tesa tapes), MACtac, 3M (Scotch tapes), Scapa Tapes Klebetechnik (Cellux).

6.2.3. Adhesive Pads and Strips

Adhesive pads and strips are die-cut pieces of relatively thick layers of highly conformable but cohesive PSA covered with release liners on both sides. Often they have a backing or core of polymeric foam. They are used to fasten objects to many surfaces.

Adhesive strips are constructed so, that in spite of their high adhesion, they can be removed easily from all surfaces just by stretching them in the bond direction. This principle first was used for medical applications [75]. More recent developments allow objects weighing up to a few

kilograms reversibly to be fastened to various surfaces [76].

Producers and Trade Names: Beiersdorf (tesa Power Strips), 3M (Scotch Command Tack).

6.2.4. High-Performance Tapes

Peel strength and adhesion under bending load can be increased by combining a highly cohesive PSA with a tough conformably balanced viscoelastic backing of adapted thickness. Such tapes also give very high bond strength on rough or profiled surfaces. Adhesion can be promoted further by application of pressure and elevated temperature (30 – 50 °C) during the bonding process and by careful exclusion of air bubbles at the interface.

The first products were based on UV-cured acrylics with a thickness of 0.3 to a few millimeters filled with glass microbubbles [77]. This has been extended now to a range of products for various interior and exterior applications, e.g., in the construction, automotive, building, and electronic industries. Such performance can be achieved also by polyurethanes [78] and other elastomers.

The performance of PSAs has been upgraded further by introducing thermosetting acrylic adhesives. This is achieved by additional epoxy groups in the PSA, which are initiated by heating after application of the tape [79].

Producers and Trade Names: Lohmann (Duplocoll tapes), 3M (Scotch VHB tapes), Norton Company.

6.3. Medical Tapes

Medical tapes are the oldest application of pressure-sensitive tapes [1]. Major progress towards skin compatibility was made by introduction of zinc oxide into the rubber-based PSAs [80]. With the development of polyisoprene, polyisobutylene, and other synthetic elastomers, natural rubber was replaced to a certain extent in medical products by synthetic rubbers. These were followed by the styrenic block copolymers [9, pp. 317 – 373].

In the late 1950s, polyacrylate-based PSAs were developed [81] and gained a dominant market share in the following decades. To a smaller extent poly(vinyl ethers) are used in medical devices. For specific applications like transdermal drug delivery systems, ostomy mounts, and electrically conductive adhesives a variety of specialty polymers such as silicones, polyurethanes, and various hydrogels are applied.

Nevertheless, both rubber- and acrylate-based PSAs are still important for medical products. Rubber-based adhesives have high tack in combination with a medium cohesive strength. The advantage of acrylics is their low allergenicity and the possibility to polymerize tailor-made products for particular applications. This is achieved by the choice of type and ratio of monomers for the polymerization and cross-linking process.

Basic characteristics of PSAs for medical applications are: Fast wetting of the skin during initial adhesion, secure adhesion on various skin types during application, and complete and pain-free removability. For resistance to moisture such as water or sweat, good permeability to water vapor and air is required.

Pressure-sensitive adhesive products are utilized in a broad variety of medical applications in very different fields: First aid products for consumers, medical tapes, island dressings, surgical drapes, wound-closure strips in hospitals and physician's practices, as well as tapes and bandages for orthopedic and athletic use. Other uses include transdermal systems for therapeutic purposes, and ostomy mounts, attaching electrocardiogram devices, and protective foot products.

First-Aid Bandages. The most widespread medical PSA products are the small bandage strips. This bandage is composed of a backing material coated with an adhesive and a small absorbent wound pad [82]. The adhesive layer is covered by a release paper or plastic film which is removed before use. For this kind of products plastic films (e.g., polyolefin), rigid or elastic fabrics, and nonwovens are applied. To attain an appropriate permeability, in many cases the backing material is perforated.

Hospital Tapes. Many hospital products are used under sterile conditions. Therefore, the raw materials have to be resistant to different sterilization methods. For several medical devices, transparency to X-rays enables examination without removing the bandage.

Depending upon the nature of the backing material, it is possible to manufacture occlusive as well as vapor-permeable products. Permeability to water vapor results in a better skin adhesion because the accumulation of moisture between skin and adhesive is avoided [83].

Medical tapes are used for all kinds of fixing purposes. The carrier of medical tapes can be woven fabric, film, perforated film, laminates, and nonwovens.

Island dressings applied after surgery are designed to absorb wound fluid either in their wound pad or in especially designed adhesive systems like hydrogels and hydrocolloids. Others are made of translucent films to enable inspection of the wound without removal of the dressing.

Self-adhering surgical drapes assure a sterile environment for surgery by isolating the operative area from resident skin bacteria. Due to the very elastic and flexible properties of the films (e.g., polyolefin or polyurethane) and their thin adhesive layer, they conform tightly to even difficult anatomical contours.

Wound-closure strips cover the field of small incisions or injuries which easily can be closed without stitches. The strips are made from a rigid fabric coated with a highly biocompatible PSA.

Orthopedic Tapes. In orthopedic applications, PSA products are most frequently used as tapes and elastic adhesive bandages. The tapes consist of a rigid backing material (e.g., cotton fabric) coated with the PSA. They are used in therapeutic medical applications to control motions and to immobilize and to stabilize joints in athletic use. Orthopedic bandages are made from coated elastic material and often are used when compression and support are to be supplied to the extremities.

Transdermal Systems. The use of transdermal drug delivery systems in pharmaceuticals has also become an important area for PSAs. This noninvasive form of drug application allows for a controlled release of bioactive substances. On applying the patch to the skin, the release of the active ingredient starts immediately and stops upon removal. During application the PSA has not only to ensure the proper adhesion of the product but also the drug delivery into the skin. This may either be achieved from a reservoir through the adhesive (membrane system) or from an adhesive that itself contains the drug (matrix system). In transdermal drug delivery systems [84] highly biocompatible adhesives with appropriate compatibility to the drug are used. These are mainly polyacrylates, synthetic rubbers, and silicone adhesives.

Electrically conductive adhesives are used in medical applications such as electrocardiogram electrodes and grounding systems for electrosurgery [9, p. 637]. They consist of electrically conductive gels, i.e., moist hydrophilic polymers containing dissolved salt.

Producers and Trade Names: Beiersdorf, Bristol-Myers Squipp, Johnson & Johnson, Lohmann, 3M, Paul Hartmann, SCA Mölnlycke, Smith & Nephew.

7. References

1. US 3965, 1845 (H. Day, W. H. Shecut).
2. C. W. Bemmels: "Pressure-Sensitive Tapes and Labels" in I. Skeist (ed.): *Handbook of Adhesives,* 2nd ed., Van Nostrand Reinhold, New York 1977, p. 724.
3. F. H. Wetzel, *Rubber Age* **82** (1957) 291 – 295.
4. C. W. Hock, A. N. Abbott, *Rubber Age* **82** (1957) 471 – 475.
5. K. Kamagata, H. Kosaka, K. Hino, M. Toyama, *J. Appl. Polymer Sci.* **15** (1971) 483 – 500.
6. M. Sherriff, R. W. Knibbs, P. G. Langley, *J. Appl. Polymer Sci.* **17** (1973) 3423 – 3438.
7. "Pressure Sensitive Adhesives" in *Encyclopedia of Polymer Science and Engineering,* **vol. 13,** John Wiley & Sons, New York 1988.
8. J. D. Ferry: *Viscoelastic Properties of Polymers,* 2nd ed., John Wiley & Sons, New York 1961.
9. D. Satas (ed.): *Handbook of Pressure Sensitive Technology,* 2nd ed., Van Nostrand Reinhold, New York 1982.
10. W. A. Zisman in: "Contact Angle, Wettability and Adhesion," *ACS Advances in Chemistry Series* (1964) no. 1.
11. M. Toyama, T. Ito, H. Morigushi, *J. Appl. Polym. Sci.* **14** (1970) 2039 – 2048.
12. M. Toyama, T. Ito, H. Morigushi, *J. Appl. Polym. Sci.* **14** (1970) 2095 – 2303.
13. M. Toyama et al., *J. Appl. Polymer Sci.* **17** (1973) 3495 – 3502.

14. A. Zosel, *Colloid Polym. Sci.* **263** (1985) 541 – 553.
15. D. W. Aubrey, M. Sherriff, *J. Polym. Sci. Polym. Chem. Ed.* **18** (1980) 2597 – 2608.
16. D. W. Aubrey: "Viscoelastic Basis of Peel Adhesion" in K. W. Allen (ed.): *Adhesion,* **vol. 3,** Applied Science Publishers Ltd., London 1978, p. 191 – 205.
17. L. H. Sperling: *An Introduction to Physical Polymer Science,* John Wiley & Sons, New York 1983.
18. *Test Methods for Pressure Sensitive Adhesive Tapes,* 11th ed., Pressure Sensitive Tape Council, 401 North Michigan Avenue, Chicago, IL 60611-4267 USA.
19. ASTM American Society for Testing & Materials, 100 Barr Harbor Drive, West Conshohocken, PA 19428 USA.
20. Association des Fabricants Europeens de Ruban Auto-adhesifs (A.F.E.R.A.), 60 rue Auber, 94408 Vitry sur Seine CEDEX, France, Manual 1983.
21. European Committee for Standardization, 36, rue de Stassart, B-1050 Bruxelles, Belgium.
22. T. B. Jensen, European Tape & Label Conference, Antwerp, April 1997, Belgium.
23. Shell Chemicals, KRATON Elastomers – Test Methods for PSAs, July 1995.
24. Exxon Chemical, Manual Number EC - AIBU/Techn-QS-04/02, Procedure TAP 5 (30.10.1997).
25. W. Hofman: *Rubber Technology Handbook,* Hanser Publishers, Munich 1996.
26. 3M Company, US 2 956 904, (J. O. Hendricks).
27. 3M Company, WO 94/11175 and WO 95/25774, (T. D. Bredahl et al.).
28. M. F. Tse, *J. Adhes. Sci. Technol.* **3** (1989) no. 7, 551 – 570.
29. G. R. Himes, L. H. Oliveri, *Adhes. Age* 1998, Feb., 41 – 44.
30. J. R. Erickson, Adhesives Age 1986, April.
31. M. DuPont, N. De Keyzer, "UV-Curing of Styrenic Block Copolymer Based Pressure Sensitive Adhesives," RadTech Europe, Maastricht 1995, p. 174 ff.
32. R. Milker, *Z. Czech Adhaesion* **3** (1985) 29 – 32.
33. W. Karmann, S. Zöllner: "Radiation Curing of Pressure Sensitive Adhesives: An Overview," RadTech Europe, Maastricht 1995, p. 155 ff.
34. 3M Comp., US 4 181 752, (J. A. Martens et al.).
35. G. Auchter, J. Barwich, G. Rehner, H. Jager, *Adhesives Age* 1994, July, 20 – 25.
36. A. B. Kummer: "Vernetzung von Acrylat-Schmelzhaftklebmassen," 23th Munich Adhesive and Finishing Seminar, 1998.
37. R. B. Fox in: *Handbook of Chemistry and Physics,* 76th ed., CRC New York, London 1996.
38. D. F. Merill: "Silicone Pressure Sensitive Adhesives," in D. Satas (ed.): *Handbook of Pressure Sensitive Technology,* 1st ed., Van Nostrand Reinhold, New York 1982.
39. S. R. Kerr: "Recent Advances in Silicone PSA Technology," PSTC Tech XVII Seminar Proceedings Hyatt Regency, Schaumburg IL, 4th – 6th May 1994, p. 187 – 201.
40. ASTM D 2398, Ring & ball softening temperatures.
41. ASTM E 28-27 "Standard Test Method for Softening Point of Resins by Ring and Ball Apparatus," pub. Sept. 1997.
42. Hercules Inc.: "The Hercules Spectrum of Hydrocarbon Resins," bulletin 900-204, Wilmington, DE.
43. 3M, EP 0 202 938, 1985 (G. F. Vesley, A. H. Paulson, E. C. Barber).
44. 3M, EP 0 257 984, (Darvell, Konsti, Klingen, Kreckel).
45. J. Murphy: *Additives for Plastics Handbook,* Elsevier Science, Amsterdam 1996, p. 116 – 118.
46. DuPont de Nemours, US 2 273 040, (R. K. Iler).
47. Beiersdorf, DE 2 845 541, 1978 (P. Gleichenhagen et al.).
48. 3M, US 3 318 852, (G. M. Dixon).
49. Permacel, US 4 513 059, (W. C. Dabroski).
50. J. D. Jones, Y. A. Peters: "Silicone Release Coatings," in [9] p. 601 – 626.
51. Th. Ebbrecht, P. Lersch, D. Wewers: "Radiation Curing of Silicone Acrylate Systems," RadTech Europe, Mediterraneo 1993, p. 322 – 328.
52. R. P. Eckberg, B. D. Shepherd: "Radiation Curable Silicone Coatings: Controlled Release Design," RadTech North America, Orlando 1994, p. 507 – 514.
53. E. D. Cohen, E. B. Gutoff (eds.): *Modern Coating and Drying Technology,* VCH Verlagsgesellschaft, Weinheim 1992.
54. R. Zimmet: "Continuous Production of Solventless Adhesives," European Tape & Label Conference, April 1997, Antwerp, Belgium.

55. P. Franz: "Kontinuierliche Aufbereitung . . . Schmelzkleber zur off-line und in-line Beschichtung von Substraten," 22. Münchener Klebstoff- und Veredlungs-Seminar, 1997, p. 6 – 11.
56. C. Rauwendaal: *Polymer Extrusion,* 2nd ed., Hanser Publishers, Munich 1990.
57. P. M. Schweizer in [53], p. 23 – 61.
58. F. Shepherd: *Modern Coating Technology Systems,* Emap McLaren Ltd., Barnet 1995, p. 13 – 22.
59. P. Herzog: "Neue Trends auf dem Sektor der Hotmelt-Düsenbeschichtung," 22. Münchener Klebstoff- und Veredlungs-Seminar, 1997, p. 1 – 5.
60. Nabert-Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* Deutscher Eich Verlag, Braunschweig 1978 (corr. 1.7.1990).
61. W. Karmann, S. Zöllner: "Radiation Curing of Pressure Sensitive Adhesives: An Overview," RadTech Europe, Maastricht 1995.
62. W. Karmann, *J. of Irradation Tech.* **1** (1983) no. 4, 305 – 323.
63. U. Köhler, G. Guse, RadTech Europe, Mediterraneo 1993.
64. G. G. Skelhorne: "Electron Beam Processing" in *Chemistry & Technology of UV & EB Formulations for Coatings, Inks & Paints,* **vol. 1,** SITA Technology, London 1991, pp. 103 – 157.
65. R. E. Knight: "UV Curing Equipment and Applications" in Chemistry and Technology of UV and EB Formulations for Coatings, Inks and Paints, vol. 1, SITA Technology, London 1991, pp. 161 – 241.
66. S. Röber: "Packaging Tapes" in D. Satas (ed.): *Handbook of Pressure Sensitive Adhesive Technology,* 3rd ed., to be published.
67. *Ullmann's,* 5th ed., **A26,** p. 97 – 100.
68. DuPont, Kaladex, Technical Data Sheet KX TD 003.
69. UL Underwriters Laboratories Inc., 333 Pfingsten Road, Northbrook IL, 60062, USA.
70. VDE Verband Deutscher Elektrotechniker e.V. VDE-Verlag GmbH, Bismarckstr. 33, D-10625 Berlin, Germany.
71. DIN Deutsches Institut für Normung e.V. Beuth Verlag GmbH, Burggrafenstr. 6, D-10787 Berlin, Germany.
72. IEC International Electrotechnical Commission, 3 Rue de Varembé, CH-1211 Geneva, Switzerland.
73. K. Külper: "Electrical Tapes" in D. Satas (ed.): *Handbook of Pressure Sensitive Adhesive Technology,* 3rd ed., in press.
74. Beiersdorf, EP 58 382, 1982 (Gleichenhagen et al.).
75. Johnson & Johnson, US 4 024 312, 1976 (R. Korpman).
76. Beiersdorf, EP 0 578 979, 1992 (B. Lühman).
77. 3M, US 4 223 067, 1977 (D. Levens).
78. Beiersdorf, DE 19 614 620, 1996 (U. Schümann).
79. G. Bennet, J. Klingen, *Adhes. Age* 1996, Sept.
80. P. G. Unna: "Weißes Kautschukheftpflaster," *Monatsh. Prakt. Dermatol. (Hamburg)* **29** (1899) 511 – 514.
81. 3M, US 2 884 126, 1958 (E. W. Ulrich).
82. E. Riedel, W. Triebsch, K. M. Sedlarik: *Verbandstoff-Fibel,* 5th ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1995.
83. I. Webster, *J. Adhes.* **17** (1997) no. 1, 69 – 73.
84. S. A. Huie, P. F. Schmit, J. Swenson Warren, *Adhes. Age* **28** (1985) no. 6., 30 – 35.
85. D. Satas, A. M. Satas, in [9], p. 637.

Tara Gum → Polysaccharides

Tar and Pitch

Gerd Collin, DECHEMA e.V., Frankfurt/Main, Federal Republic of Germany (Chaps. 1 – 6)

Hartmut Höke, Weyl GmbH, Mannheim, Federal Republic of Germany (Chap. 7)

1. Origin, Classification, and Industrial Importance of Tars and Pitches 1
2. Properties 4
3. Processing of Coke-Oven Coal Tar 9
3.1. Survey 9
3.2. Primary Distillation 9
3.3. Processing of Coal-Tar Pitch 14
3.3.1. Cooling 14
3.3.2. Production of Electrode Pitch 14
3.3.3. Production of Pitch Coke 17
3.3.4. Production of Special Pitches 19
3.3.5. Pitch – Oil Blends 21
3.4. Processing of Tar Distillates 22
3.4.1. Processing of Anthracene Oil 22
3.4.2. Processing of Wash Oil 23
3.4.3. Processing of Naphthalene Oil 25
3.4.4. Processing of Carbolic Oil 25
3.4.5. Processing of Light Oil 27
4. Processing of Low-Temperature Coal Tars 31
5. Processing of Other Tars and Tarlike Raw Materials 33
6. Uses of Tar Products and their Economic Importance 34
7. Toxicology and Ecotoxicology 35
7.1. Toxicology 35
7.2. Classification and Legislation 36
7.3. Ecotoxicology 36
8. References 37

1. Origin, Classification, and Industrial Importance of Tars and Pitches

Origin and Classification. *Tars* are liquid or semisolid products obtained by thermal decomposition of natural organic materials. *Pitches* are residues of tar distillation or meltable residues directly obtained by distillation of organic substances. Tars and pitches take their names from the source material. Tars are produced in industrial quantities by low-temperature carbonization, coking, or gasification of fossil raw materials or biomass, e.g., coal, lignite, peat, wood, and oil shale. In addition, tars result as higher boiling liquid pyrolysis products from thermal decomposition of organic secondary raw materials and waste products, e.g., pyrolysis residual oils as byproducts of steam cracking of petroleum fractions to alkenes and from pyrolysis of refuse and plastic waste.

Properties and composition of tars and pitches depend not only on the raw material but also on the temperature conditions during thermal treatment. Thus low-temperature tars are produced during low-temperature carbonization or partial gasification below 700 °C. High-temperature tars are generated by coking between 900 and 1300 °C. The differences in the properties and composition of low- and high-temperature tars are especially marked in case of coal tars. Low-temperature coal tars contain phenols, hydrogenated aromatics, aliphatic hydrocarbons, and alkenes as their main components, together with small amounts of aromatic hydrocarbons. These compounds are generated from two sources: (1) Coal bitumen, the extractable component of coal, which consists of oxygen-, sulfur-, and nitrogen-containing hydrocarbon compounds, and (2) the bitumen which arises additionally from the insoluble components by cracking of aliphatic C – C, C – O, C – N, and C – S bonds, as well as through the removal of side chains (→ Coal Pyrolysis, Chap. 1.). High-temperature coking of coal generates thermally stable aromatics, which are characteristic of high-temperature coal tars. These heat-stable aromatics arise via the following reaction schemes: Paraffins are cracked pyrolytically and dehydrogenated. The resulting alkenes are converted first into hydrogenated aromatics and finally into aromatics. Alkylphenols are dealkylated and reduced to aromatic hydrocarbons.

Lignite tars differ from low-temperature coal tars mainly in an increased content of solid paraffins and a lower amount of cyclic hydrocarbons.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_091

Table 1. Properties and composition of various tars

	Coal tars			Lignite tars	Peat tars	Wood tars	Oil-shale tars	Pyrolysis residual oils from steam cracking of naphtha and gas oil	Pyrolysis residual oils from pyrolysis of plastic waste
	Coke-oven tars	Low-temperature tars	Gasification tars *						
Density at 20 °C, g/cm^3	1.14 – 1.25	0.96 – 1.05	1.05 – 1.14	0.95 – 1.05	0.94 – 0.98	1.08 – 1.20	0.90 – 0.97	0.96 – 1.10	0.93 – 1.05
Carbon, %	90 – 93	83 – 85	84 – 86	81 – 85	78 – 82	60 – 65	78 – 86	89 – 93-,1	88 – 91
Hydrogen, %	5 – 6	8 – 9.5	6 – 8	9 – 11	8 – 10	6 – 8	9 – 12	6 – 9	8 – 10
Naphthalene, %	5 – 15	0 – 2	2 – 4	0 – 1	0 – 1	0 – 0.5		10 – 15	4 – 12
Phenols, %	0.5 – 5	10 – 45	15 – 25	5 – 30	5 – 25	20 – 40	1 – 30	0 – trace	trace
Bases, %	0.2 – 2	0.5 – 2	0.5 – 4	0.1 – 1.5	0.5 – 3	trace – 0.5	0.2 – 9	0 – trace	0 – trace
Solid paraffins, %	0 – trace	3 – 15	0 – 5	8 – 20	10 – 30	trace	0.5 – 12	0 – 5	0 – trace
Coking residue **, %	10 – 40	5 – 15	10 – 20	4 – 9	4 – 10	10 – 18	1 – 7	5 – 20	1 – 10
Toluene insolubles, %	2 – 20	0.5 – 10	3 – 10	0.2 – 1.5	2 – 8	trace – 12		trace	trace
Ash (800 °C), %	< 0.5	< 0.2	< 1.5	< 0.2	< 0.2	< 0.2		< 0.1	< 0.2
Water, %	< 5	< 5	< 5	5	1 – 10	3 – 10		0 – 1	0.2
Distillation analysis (DIN 1995)									
Light oil up to 180 °C, %	0.2 – 2	0 – 4	0 – 10	0 – 2	0 – 3	10 – 20	10 – 15	0 – 10	40 – 70
Middle oil 180 – 230 °C, %	3 – 10	5 – 25	5 – 15	0 – 4	3 – 25	10 – 14	5 – 20	20 – 30	8 – 15
Heavy oil 230 – 270 °C, %	7 – 15	10 – 20	5 – 20	2 – 18	5 – 20	7 – 12	15 – 35	20 – 30	5 – 15
270 – 300 °C, %	3 – 7	8 – 12	5 – 10	5 – 20	10 – 20	5 – 10	0 – 15	10 – 20	3 – 8
300 °C up to pitch, %	12 – 30	15 – 40	10 – 30	20 – 50	20 – 40	25 – 40	0 – 10	2 – 5	2 – 10
Pitch, %	45 – 65	25 – 40	30 – 50	20 – 60	20 – 35	20 – 30	30 – 50	25 – 30	10 – 20

* From Lurgi fixed-bed pressure gasification of coal with oxygen and steam.
** Conradson method.

The main source of paraffins is the extractable montan wax occurring in lignite. The tars are formed by thermal decomposition of initially present and newly formed "bitumen," similar to the case of coal. Tar obtained by high-temperature coking of lignite is similar to low-temperature lignite tar, as lignite bitumen, in contrast to coal bitumen, is already entirely decomposed and volatilized at lower temperatures.

Peat tar properties are determined by the structure of peat, which is a conglomerate of extractable bitumen, humic acid and its salts, and partially decomposed vegetation. In pyrolytic distillation "bitumen" is the source of the bulk of the tar, mainly paraffins. Ketones, methanol, monocarboxylic acids, and phenols arise from the plant components cellulose and lignin (→ Peat, Chap. 3.3.2.).

Wood tars are formed by thermal decomposition of the wood components cellulose, hemicelluloses, and lignin. This gives rise to acetic acid, methanol, ketones and phenols but, in contrast to tars from fossil raw materials, almost no hydrocarbons (→ Charcoal).

History. The discovery of *coal tar* goes back to J. R. Glauber, who in 1658 described the formation of " black oleum" during coal retorting. A patent for the production of pitch and tar by coking of coal with condensation of volatiles was granted in England to J. J. Becher and H. Serle in 1681. In the early 1800s coal tar was initially regarded as an annoying byproduct of town-gas and metallurgical- coke production. In 1822 the first industrial tar distilleries were built in Britain. The light oils obtained were used as solvents, the heavy oils as impregnants for building timber and railway sleepers, and the distillation residue, pitch, for the production of carbon black. In 1845 A. W. Hofmann discovered benzene in the light oil used to produce nitrobenzene as a perfume for soaps, named essence of mirbane. H. Garden (1819) had previously discovered naphthalene and J. B. A. Dumas and M. A. Laurent (1832) anthracene in coal tar. With the discovery of synthetic dyes during the second half of the 1800s, demand for these three aromatic hydrocarbons led to coal tar becoming the main source of raw materials for the flourishing organic chemical industry in central Europe. The first German tar distilleries were built in 1842 in Offenbach by E. Sell, and in 1860 in Erkner near Berlin by J. Rütgers, followed by numerous others.

Lignite tar was obtained by J. G. Krünitz in 1788 as "rock oil" when retorting central German "earth coal." In 1830 Karl Freiherr von Reichenbach isolated the main component, paraffin, by distillation. The lamp oil, obtained as a byproduct, replaced rapeseed oil, which had been used up to then for lighting. After 1858 low-temperature lignite- carbonization plants were constructed in central Germany for the production of tar-based products and town gas. During the 1930s central German lignite tar was mainly processed by hydrogenation to produce motor fuels.

Peat and wood were carbonized in early times. The Egyptians used *wood tar* for embalming corpses. Solid paraffin for candle production and liquid paraffin oil for lamp oil were obtained from *peat tar* or lignite tar during the 1800s.

Industrial Importance. Approximately 15×10^6 t/a of *coal tar* is produced worldwide, mainly in the form of high-temperature tar as a byproduct of metallurgical coke production by horizontal chamber coking (→ Coal Pyrolysis, Chap. 6.2.). Smaller amounts are produced as byproducts of low-temperature carbonization of coal to produce solid smokeless fuel (→ Coal Pyrolysis, Chap. 6.1.) and in pressure gasification of coal by the Lurgi process (→ Coal, Chap. 9.4.). Approximately 100 coal- chemical tar refineries process coal tar to give chemical feedstocks and other aromatic products. Industrial countries with major steel production account for most of the capacity: ca. 50 % in Europe and the CIS, 20 % in North America, and 20 % in South-East Asia, especially China and Japan. In the United Kingdom and India, low-temperature tars are obtained as byproducts of low-temperature carbonization of coal to produce smokeless fuels. Lurgi pressure gasification is a preliminary stage of Fischer – Tropsch synthesis in South Africa.

In contrast to coal tars, *lignite tars* are produced only in a few countries, e.g., as a byproduct of pressure gasification of lignite in the United States (North Dakota), the Czech Republic, Germany (Lausitz), Russia, and China. Lignite is carbonized to produce coke on a minor scale, e.g., in Germany (Rhineland and Lausitz) and India. Normally only the phenolic effluents

are processed, whereas the lignite tars are gasified or burnt within the process. The total amount of lignite tar arising is ca. 10^6 t/a.

Peat tars and *wood tars* arise as byproducts in the production of activated carbon and reductive carbon by carbonization of peat and wood (→ Peat, Chap. 3.3.2., → Peat; → Carbon, Chap. 5.3.). The main producing countries are Russia (peat) and Brazil (wood). The total production of these tars is at most ca. 3×10^6 t/a.

Oil-shale tars are produced by low-temperature carbonization and partial gasification of oil shale, mainly in Estonia (Kiviter process). Approximately 10^6 t/a are produced in this way (→ Oil Shale, Chap. 7.). Other processes involving pyrolysis of oil sand and oil shale are still at the pilot-plant stage.

Pyrolysis residual oils are tarlike byproducts of the steam cracking of naphtha and gas oil to produce alkenes. Approximately 10×10^6 t were produced in 1993. Pyrolysis oils with similar compositions are obtained by thermal decomposition of waste plastics. Industrial large-scale tests are being carried out to develop methods of waste-plastics pyrolysis (→ Plastics, Recycling, Chap. 4.3.).

2. Properties

In general, *tars* are dark or black viscous liquids containing small amounts of emulsified water and dispersed high- carbon solid particles (free carbon, tar sediments) entrained during formation. Organic components may crystallize from high-paraffinic tars at lower temperatures.

Pitches are generally dark or black thermoplastic materials with a broad molecular-mass spectrum. By using selective solvents it is possible to separate pitches into fractions with different properties. At temperatures below the softening range most pitches transform into glass-like solids. At higher temperatures polymerization and polycondensation reactions take place, and pitches change partly from liquid crystal intermediate stages into a nonthermoplastic state and finally into coke [5].

The chemical composition of tars and pitches varies widely and depends essentially on the type of raw materials and the temperature conditions during production. In all cases, multicomponent mixtures are produced that contain aromatic and/or aliphatic hydrocarbons as the main components, together with oxygen-, sulfur-, and nitrogen- containing heterocyclic compounds; phenols; amines; and possibly carboxylic acids, ketones, and aliphatic alcohols as minor constituents. It has been estimated that coal tar can contain more than 10 000 different constituents, most of which have boiling points above 500 °C. Multicomponent eutectics can form, especially in the case of condensed aromatics. The liquid-viscous nature of pitches and many tar oils arises from strong mutual melting-point depression of the individual components [6].

The main properties of different tars are listed in Table 1. Given approximately equal boiling range and pitch content, the density and carbon content indicate the degree of aromatization of tars. Coke-oven coal tars with average densities of ca. 1.2 g/cm^3 and carbon contents >90 % contain the highest fraction of aromatic hydrocarbons, followed by coal low-temperature gasification tars with densities of ca. 1.1 g/cm^3 and carbon contents of ca. 85 %. Smaller amounts of aromatics are found in tars produced by low-temperature carbonization of coal and lignite having densities of ca. 1 g/cm^3 and carbon contents of max. 85 %. Peat and wood tars with carbon contents of only ca. 80 % or even <70 % and varying densities, contain almost no aromatic compounds. Oil-shale tars are hardly or only partially aromatized. However, a high content of aromatic components can be found in pyrolysis residual oils from ethylene/propene plants and from the thermal decomposition of waste plastics if the pyrolysis temperature is at least ca. 700 °C. These pyrolysis residual oils have an aromatic hydrocarbon content of ca. 90 % and higher, similar to coke-oven coal tar, whereas pyrolysis oils contain almost no phenols and only small amounts of heterocyclic compounds due to the raw-material basis. The carbon content falls and the hydrogen content rises with decreasing tar aromatization. A sensitive criterion for the "aromaticity" of tars is the naphthalene content, which is ca. 10 % in coke-oven coal tars and pyrolysis residual oils, ca. 3 % in coal-gasification tars on average, less than 2 % in low-temperature coal-carbonization tars, and max. 1 % for lignite, peat, and wood tars. With increasing geological age of the raw material (i.e., wood $<$ peat $<$ lignite $<$ coal), the content of phenols and solid paraffins decrease, while the

"aromaticity" increases. Thus coke-oven coal tars contain on average ca. 2 % phenols and only traces of solid paraffinic hydrocarbons, whereas low-temperature coal-carbonization tars, lignite tars, and peat tars contain up to ca. 30 % phenols and solid paraffins. High coking values and components that are insoluble in toluene normally indicate a high degree of aromatization.

Table 2. Properties of coke-oven coal tars

	Average value	Range of analytical data
Density, g/cm^3	1.175	1.14 – 1.20
Water, %	2.5	0.2 – 5.0
Toluene insolubles, %	5.50	2 – 12
Quinoline insolubles, %	2.0	0.5 – 5
Coking value (platinum crucible), %	14.6	10 – 20
Carbon (waf) ∗, %	91.39	90 – 93
Hydrogen (waf), %	5.25	5 – 6
Nitrogen (waf), %	0.86	0.6 – 1.2
Oxygen (waf), %	1.75	1.5 – 2
Sulfur, %	0.75	0.6 – 1
Chlorine, %	0.030	trace – 0.1
Ash (800 °C), %	0.15	0.05 – 0.25
Zinc, %	0.038	trace – 0.2
Naphthalene, %	10.0	5 – 13
Distillation analysis (DIN 1995), %		
up to 180 °C		
water	2.5	0.2 – 5
light oil	0.9	0.2 – 2
180 – 230 °C	7.5	3 – 10
230 – 270 °C	9.8	7 – 15
270 – 300 °C	4.3	3 – 7
300 °C up to pitch	20.1	12 – 30
pitch (softening point 67 °C KS ∗∗)	54.5	45 – 65
distillation loss	0.5	
Crude phenols, %	1.5	0.4 – 5
Phenol, %	0.5	0.08 – 0.7
Cresols and xylenols, %	1.0	0.3 – 4.5
Low-boiling pyridine bases (in light oil), %	0.081	0.01 – 0.2
High boiling pyridine and quinoline bases (in middle and heavy oil), %	0.065	0.2 – 0.9
Unsaturated compounds (in light oil), %	0.590	0.02 – 1.0

∗ waf: water- and ash-free. ∗∗ KS: Krämer – Sarnow.

Tar properties are investigated with numerous analytical methods. Table 2 contains the analytical values of a typical coke-oven coal tar. Figure 1 shows a gas chromatogram of a coke-oven tar, and Table 3 lists the most important constituents of coke-oven coal tar. Up to now about 600 compounds have been identified in coal tars; together, they account for ca. 55 % of the tar. The remaining, unidentified compounds are mainly pitch components which decompose before they vaporize.

In general, *pitches* are multicomponent mixtures of higher molecular compounds. *Coal-tar pitch* obtained from coke-oven tar is a mixture of condensed aromatic hydrocarbons and heterocyclic compounds. Figure 2 shows a high-pressure liquid chromatogram (HPLC) of coal-tar pitch. HPLC can analyze ca. 70 % of coal-tar pitch [7]. Gel permeation chromatographic investigations indicate in case of higher molecular pitch fractions that they have molecular weights up to approximately 2500 [8]. The highest molecular components, which are partly insoluble, even in pitch, have higher molecular masses. The higher molecular fractions consist of condensed aromatic hydrocarbons and heterocyclics with insignificant fractions containing methyl- or hydroxyl-substituted compounds. The ^{13}C NMR spectrum shows that 97 % of the carbon atoms present are aromatic [9], [10]. Higher molecular pitch fractions contain monomeric aromatics and heterocyclic aromatics, as well as oligomeric systems, in some of which aromatics containing nitrogen and other heteroelements are linked by C – C single bonds or bridges such as CH_2, O, and S [11], [12]. Very large *peri*- condensed systems arise on thermal treatment of normal pitch, whereby the average molecular mass increases and the solubility of the pitch in solvents decreases. Furthermore, anisotropic liquid crystal structures known as "mesophases" are formed [13], [14].

Properties of coke-oven tar "normal" pitch are listed below:

Density at 20 °C	1.27 – 1.28 g/cm^3
Softening point (Krämer-Sarnow)	68 ± 3 °C
Softening point (Mettler)	86 ± 3 °C
Brittle point (Fraass)	30 – 40 °C
Elemental analysis:	
Carbon	92 – 94 %
Hydrogen	4 – 5 %
Nitrogen	1.0 – 1.3 %
Sulfur	0.5 – 0.8 %
Oxygen	1 – 2 %
n-Hexane solubles	40 – 60 %
Toluene insolubles	15 – 25 %
Quinoline insolubles	3 – 7 %
Coking value (ALCAN) (DIN 51 905)	43 – 48 %

Table 3. Important constituents of a typical coke-oven coal tar

Compound	Formula	*bp* at 1.013 mbar, °C	*mp*, °C	Average content, %
Hydrocarbons				
Naphthalene		217.95	80.29	10.0
Phenanthrene	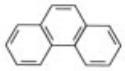	338.4	100.5	4.5
Fluoranthene	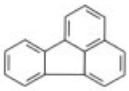	383.5	111	3.0
Acenaphthylene		270	93	2.5
Pyrene	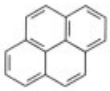	393.5	150	2.0
Fluorene	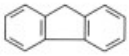	298	115	1.8
2-Methylnaphthalene	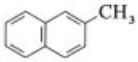	243.1	34.6	1.5
Anthracene	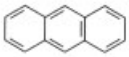	339.9	218	1.3
Chrysene	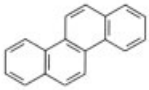	441	256	1.0
Indene		182.8	−1.8	1.0
1-Methylnaphthalene		244.4	−30.5	0.7
Benzene		80.1	5.53	0.4
Diphenyl	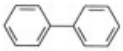	255.9	71	0.4
Acenaphthene		277.2	95.3	0.2
Heterocyclics				
Dibenzofuran	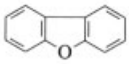	287	83	1.3
Carbazol	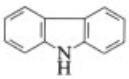	354.75	245	0.9

Table 3. (Continued)

Dibenzothiophene	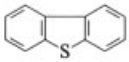	331.4	97	0.4
Benzothiophene		219.9	31.3	0.3
Quinoline		273.1	−15	0.3
7,8-Benzoquinoline	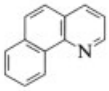	340.2	52	0.2
2,3-Benzodiphenylenoxide	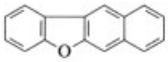	394.5	208	0.2
Indole		254.7	52.5	0.2
Acridine	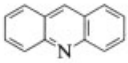	243.9	111	0.1
Isoquinoline		243.25	26.5	0.1
Quinaldine	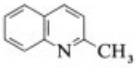	245.6	−2	0.1
Phenanthridine	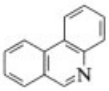	349.5	107	0.1
Pyridine		115.26	−41.8	0.03
2-Methylpyridine		129.41	−66.7	0.02
Phenols				
Phenol		181.87	40.89	0.5
m-Cresol		292.23	12.22	0.4
o-Cresol		191.00	30.99	0.2
p-Cresol	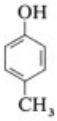	201.94	34.69	0.2

Table 3. (Continued)

3,5-Dimethylphenol	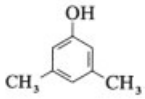	221.96	63.27	0.1
2,4-Dimethylphenol	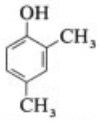	210.93	24.54	0.1

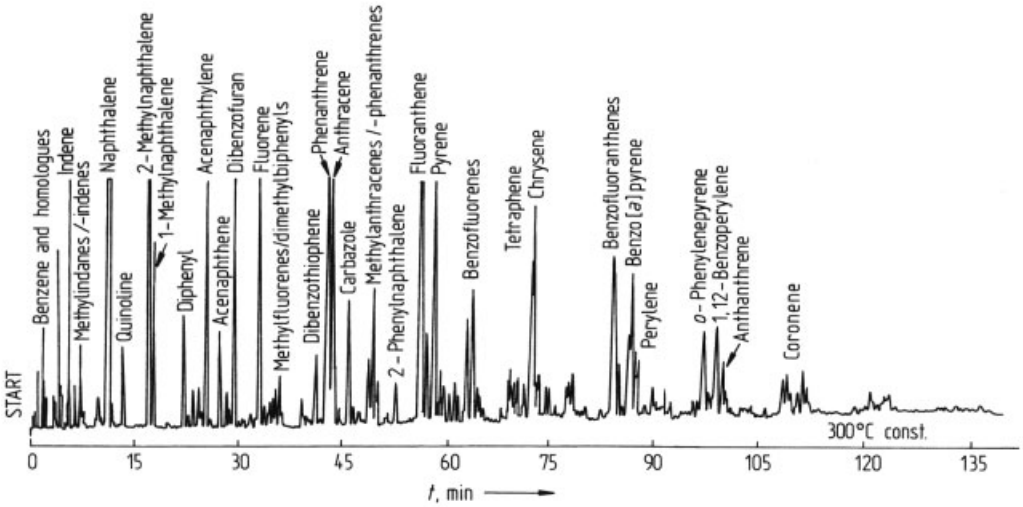

Figure 1. Gas chromatogram of a coke-oven coal tar 22 mm glass-capill. CP-sil 5 (inner diameter 0.5 mm), split, 50 °C 2 min, 2 °C/min up to 300 °C const., helium.
For reasons of clarity only selected peaks are shown with compound names. The concentrations of selected components (in wt %) are:
benzene/benzene homologues (1), indene (1), methylindanes/methylindenes (0.8), naphthalene (10), quinoline (0.3), 2-methylnaphthalene (1.5), 1-methylnaphthalene (0.5), diphenyl (0.2), acenaphthylene (2), dibenzothiophene (0.3), phenanthrene (5), anthracene (1.3), carbazole (0.7), methylanthracenes/methylphenanthrenes (1.8), 2-phenylnaphthalene (0.3), fluoranthene (3.3), pyrene (2.1), benzofluorenes (2.3).

Ash (800 °C)	< 0.3 %
Iron	200 – 300 ppm
Silicon	100 – 300 ppm
Vanadium	< 50 ppm
Specific heat capacity (140 – 210 °C)	C_p/J g^{-1} K^{-1} = 1.15 + 0.0037 T/°C
Coefficient of thermal expansion	0.00045 – 0.00056 K^{-1}
Thermal conductivity (25 – 105 °C)	0.0014 J cm^{-1} sec^{-1} K^{-1}
Surface tension (120 – 220 °C)	γ/mN cm^{-1} = 18.4 · d_t^4 (d_t = density at measuring temperature)
Dielectric constant (25 °C)	3.3 – 3.4
Magnetic susceptibility	0.69 × 10^6

3. Processing of Coke-Oven Coal Tar

3.1. Survey

The refining and processing of coke-oven coal tar preferentially isolates aromatic chemicals by a combination of six basic processes (Figure 3):

1) *Distillation* is used for primary separation of crude tar into six fractions: light oil, carbolic oil, naphthalene oil, wash oil, anthracene oil, and pitch. It is also used for further separation of primary and secondary tar fractions.
2) *Crystallization* is used to produce crystalline condensed aromatic hydrocarbons and non-

polar heterocyclics from distillate fractions (e.g., naphthalene, methylnaphthalenes, acenaphthene, dibenzofuran, fluorene, phenanthrene, anthracene, carbazole, fluoranthene, and pyrene). The aromatic oils obtained as filtrates are suitable as components for carbon-black oils, impregnating oils, fuel oils, diesel fuels, fluxing oils, and pitch – oil mixtures such as road tars, refractory pitches, anti-corrosion and protective agents for buildings.

3) *Extraction* with alkalis and acids allows the recovery of phenols and polar nitrogeneous tar bases from light oil, carbolic oil, and filtered naphthalene oil.
4) *Catalytic polymerization* with Friedel–Crafts catalysts yields thermoplastic indene–coumarone resins from light oil fractions from which the phenols and bases have been extracted. The nonpolymerizable component is crude benzole, which together with coke-oven benzol is processed to benzene and benzene homologues; the higher boiling fractions are refined to solvent naphtha and heavy naphtha.
5) *Thermal polymerization* of coal-tar normal pitch yields higher value pitch products such as hard pitch and various special pitches.
6) *Coking* of pitch leads to pitch coke as technically pure carbon which is suitable for further processing into graphite, carbon electrodes, and carburizing agents.

3.2. Primary Distillation

Crude coal tar is generally pretreated prior to primary distillation. The emulsified aqueous solution (up to 5 %) of ammonia, hydrogen sulfide, and ammonium chloride is removed as far as possible. Suspended fine solids (tar sediments) may also be removed.

The simplest procedure for a partial separation of water is to store crude tar at elevated temperatures. Storage for several days at 60 °C decreases the water content by 1 – 2 %. Due to its lower density the water collects at the top of the tank, thus enabling its periodical discharge. To reduce the tar viscosity and increase the settling rate of water, crude tar can be stored in pressure tanks at 100 – 200 °C. This pressure dehydration can decrease the water content to ca. 1 – 1.5 %. The settling rate can be further increased by centrifuging at 6000 – 7000 *g*. Disk centrifuges with discharge nozzles reduce the water content of crude tar down to ca. 0.5 – 1 %.

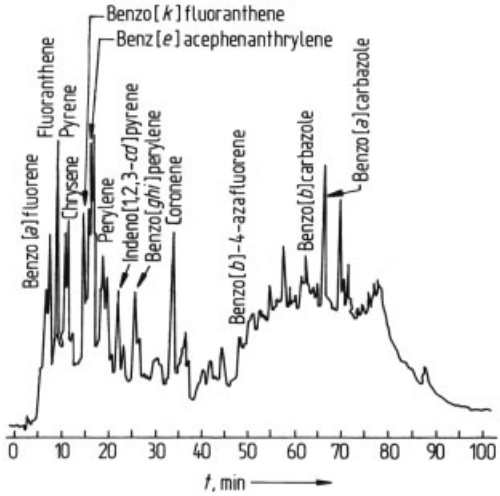

Figure 2. High-pressure liquid chromatogram of coal-tar pitch [7]
Nucleosil 5-NO_2, 5 µm, 2 columns of 200 mm length and 4 mm inner diameter and pre-column with 20 mm length and 4 mm inner diameter (switched one after the other, with gradient elution (chloroform – *n*-hexane) and UV detection at 300 nm)

The ammonium chloride contained in tar water dissociates into ammonia and hydrogen chloride above 225 °C, causing corrosion. The ammonium chloride content of crude tar can be greatly reduced by extraction with water prior to mechanical tar – water separation. Another method is to mix in an amount of sodium carbonate or sodium hydroxide solution equivalent to the ammonium chloride content. However, during tar distillation nonvolatile sodium chloride is formed and remains in the pitch, increasing its ash content.

Further impurities of crude tar are dispersed fine solids (tar sediments), which mainly consist of entrained fine coke particles. The partial removal of these sediments takes place together with mechanical water separation by sedimentation during storage. It is also possible to obtain a sediment concentrate from the crude tar by two- or three-phase centrifugal separators either with water as third or as second solid phase. Thus, erosion in pumps, pipes, and valves is avoided and the content of inorganic trace contaminants

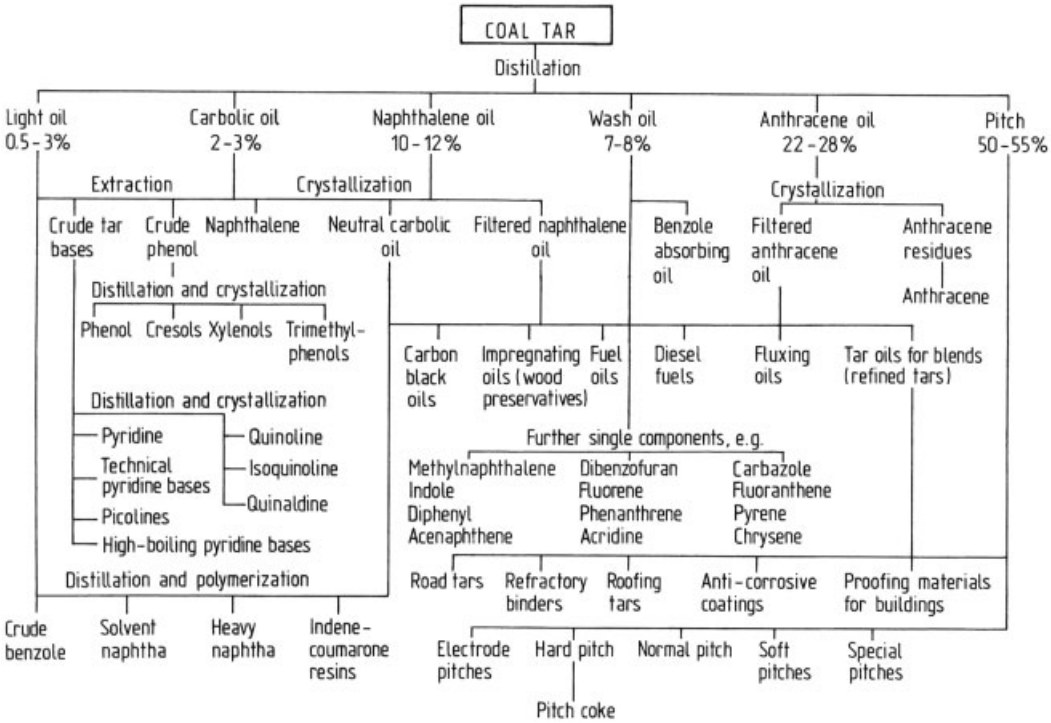

Figure 3. Processing of coal tar [1]

in the pitch primary distillation residue is decreased.

Nowadays, partly dehydrated and pretreated crude tar is distilled continuously in pipe stills and rectifying columns. Numerous distillation processes have been developed. They can be divided into flash processes, bottom-circulation processes, and processes in which these two principles are combined.

In *flash distillation* the tar is heated continuously under pressure, usually in a two-stage pipe still. It is then flashed into a system of rectification columns where the oil components are evaporated and fractionated. The first heating step dehydrates and evaporates the light oil, and in the second step the remaining oil components are evaporated, leaving pitch as the residue.

In *bottom-circulation distillation* the rectifying column bottoms are pumped continuously through pipe stills where they are provided with the thermal energy required for distillation.

These two distillation processes differ in the thermal stress on the tar fractions. In the flash process the oil fractions are subject to temperatures well above their boiling range, whereas the distillation residue pitch is not exposed to high temperatures. In the bottom-circulation process, however, the pitch is subject to high temperatures for long periods, but there is only a low thermal stress on the vaporized oil components. As several coal-tar constituents are unstable under distillation conditions, the two distillation processes differ significantly in the yield and quality of certain tar products.

Pipe stills are mostly rectangular or cylindrical, with an inner refractory lining. To maximize the heat utilization of firing gases, pipe stills are divided into radiant and convection zones and are often equipped with air preheaters. The burners of the radiant zone can be operated with gases or liquid fuels of various qualities. The tubes are made of steel; austenitic 18/8/2 chromium nickel molybdenum steels have proved to have especially good corrosion resistance.

As rectifying columns, bubble-cap columns, valve tray columns, sieve-plate columns, packed columns, and combinations thereof are used. Stoneware Raschig rings are used as packing

materials. The lower regions of the dehydration columns and flash columns mostly contain simple cascade trays. In zones with high corrosion risk the columns consist of high-grade stainless steel or cast iron, otherwise of mild steel.

To separate high-boiling tar fractions the columns normally operate under vacuum. In some processes steam is injected to lower the evaporation temperature. Heat of condensation and the thermal energy contained in the flowing bottom products are transferred via heat exchangers to the crude tar and its intermediate products to lower the total energy demand.

The *GfT/Koppers* continuous tar distillation process is shown in Figure 4 as an example of flash distillation. In the first step the crude tar is heated to ca. 160 °C in the pipe still and then flashed through an expansion valve into the middle of the atmospheric dehydration column. The pressure before the expansion valve is up to 10 bar. Light oil and water evaporate in the dehydration column, which is equipped with bubble and overflow trays, and are separated into two phases after condensation. Part of the light oil is refluxed to the top of the dehydration column. The dehydrated (topped) tar as column bottom is preheated in a heat exchanger by the hot pitch stream, then it is heated under pressure to 360 °C in the second stage of the pipe still. At this temperature it is flashed in the pitch column at a pressure of 100 – 130 mbar. The pressure before the expansion nozzle is up to 5 bar. To take off the high-boiling anthracene oil II above the flash chamber, there is a column, filled with Raschig rings, which is dephlegmated at the top with the lower boiling anthracene oil I. The column tops of the vacuum system are refluxed with a split stream of the oil fraction taken from the bottom of the next column. The bottom product of the wash oil column is directed via a side column to evaporate the remaining naphthalene. This side column is additionally heated by bottom circulation, yielding an improved rectification efficiency. Concentrated ammonium chloride is extracted with water from the bottom product of the wash oil main column before the oil is refluxed to the top of the anthracene oil column. This continuous removal of ammonium chloride prevents clogging and corrosion of the column by the salt, which condenses to form a solid in the temperature range 180 – 200 °C. The tar oil fractions anthracene oil I, wash oil, and naphthalene oil are separated as column bottoms. Carbolic oil is condensed in the final condenser as the top product of the naphthalene column. The uncondensed gases and vapors are extracted with oil and water before entering the vacuum equipment. Light oil and naphthalene vapors are extracted in the oil scrubber with cooled wash oil. Hydrogen sulfide is removed with water in the water scrubber. The heat consumption of the total plant is ca. 900 MJ/t of wet crude tar.

Figure 5 shows a diagram of the *Rütgers process* for continuous tar distillation using the bottom- circulation principle. Each of the main rectifying columns has its own pipe still for bottom circulation to achieve independent control of the column reflux. Condensing water, light oil vapors of the dehydration column, and waste heat vapor preheat the crude tar to 105 °C. Afterwards the crude tar is pumped into a dehydration column which contains cascade trays. The bottoms of the dehydration column are circulated through a pipe still and kept at 150 °C. After intermediate storage the dehydrated topped tar is passed through to heat exchangers where the heat of condensation of carbolic oil and the heat of pitch is used. The topped tar, at ca. 250 °C, enters the middle section of the carbolic oil column, which is operated at atmospheric pressure and is equipped with valve or Perform trays in the lower part and, bubble trays for example, in the upper part. The pipe still, with a discharge temperature of 380 °C, allows adjustment of the reflux ratio up to 16 : 1. A distillate fraction in the carbolic oil column is cycled to the top of a side-stream column where the naphthalene oil fraction is taken as a vapor from its base. The main-stream bottom product is led via heat exchangers to the methylnaphthalene column, which is run with a top pressure of 250 mbar and which also contains a packing or Perform trays as well as bubble trays. The reflux ratio of this column is up to 17 : 1. A methylnaphthalene fraction is obtained as top product. The top vapors are condensed in a low-pressure steam generator. A wash oil fraction is taken from a side column and a fluorene fraction is taken as a further side stream of the main column. The bottom product of the methylnaphthalene column is led to the anthracene-oil packed column with a top pressure of 90 mbar. Its top product is anthracene oil, and the bottom product is normal pitch. This column only requires a reflux ratio of 1.5 : 1. The

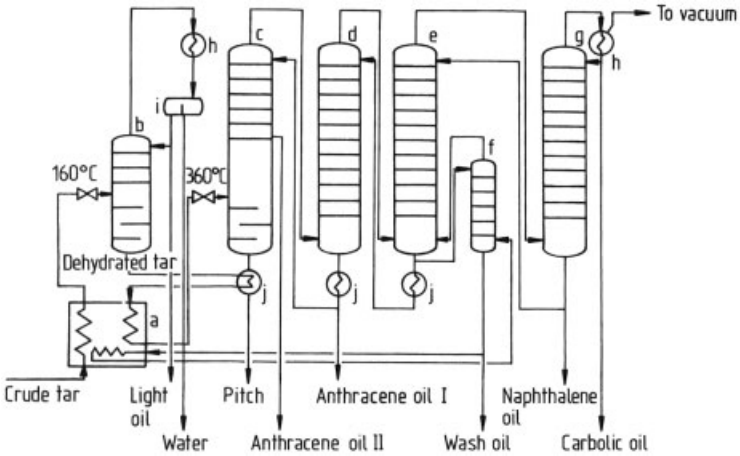

Figure 4. Continuous tar distillation with multiflash system (GfT/Koppers process)
a) Pipe still; b) Dehydration column; c) Pitch/anthracene oil II column; d) Anthracene oil I column; e) Wash oil column; f) Wash oil side-stream column; g) Naphthalene oil column; h) Condenser; i) Decanter; j) Heat exchanger

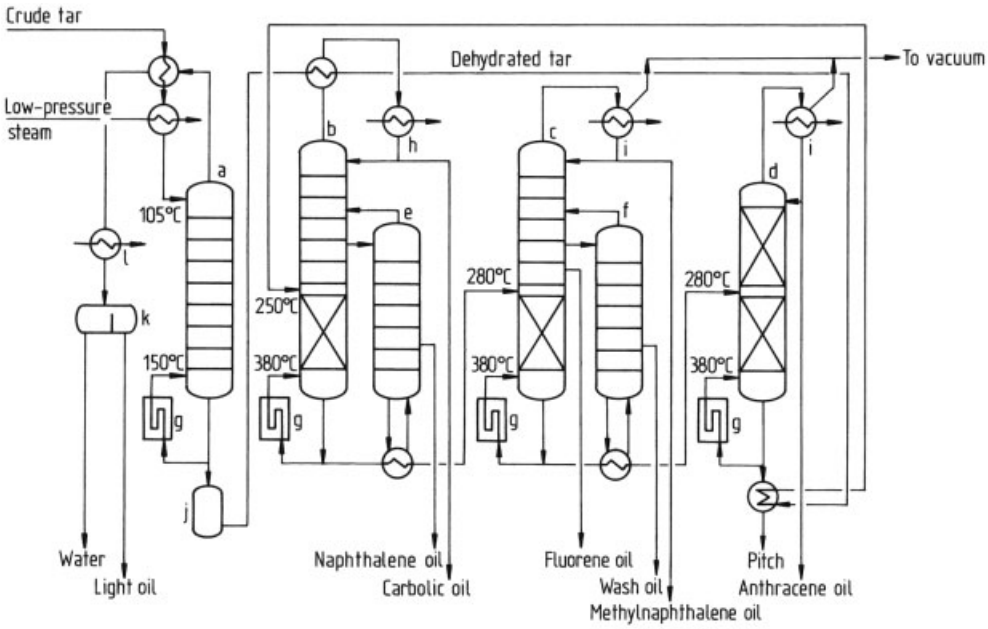

Figure 5. Continuous tar distillation (Rütgers process)
a) Dehydration column; b) Carbolic oil column; c) Methylnaphthalene oil column; d) Anthracene oil column; e) Naphthalene oil side-stream column; f) Wash oil side-stream column; g) Pipe still; h) Condenser; i) Condenser (steam generator); j) Vessel for dehydrated tar (topped tar); k) Decanter; l) Cooler

condensation heat of the anthracene oil is used for steam generation.

Besides the product flows described, other column interconnections are possible. Due to the high reflux ratios in efficient rectifying columns the tar oils are separated into many single fractions. Thus a high-yield extraction of technically pure components can be achieved. For example, the naphthalene oil fraction contains up to 90 % or more of the naphthalene occurring in crude tar at a concentration up to 90 % (crystallizing point up to 75 °C). Despite the increased energy needed for high reflux ratios, the heat requirement of the plant is only ca. 880 MJ/t crude tar due to the heat recovery system.

If the production of coal-tar pitch rather than tar aromatics and oils is the main aim then the primary distillation of crude tar should be combined with a thermal treatment at increased temperature and longer time under pressure. This is the case, for example, in the production of binders or coke for carbon and graphite electrodes.

The *Cherry-T process* of Osaka Gas [15] serves as an example of such a procedure. In this process water and light oil are separated from the crude tar as a first step, usually in a dehydration column. The dehydrated tar is heated in a pipe still to almost 400 °C and afterwards thermally polycondensed in a pressure reactor. After this reactor, a flash column is used for oil/pitch separation. If necessary, the oil can also be treated thermally in a second pressure reactor. Then, after additional flashing, a secondary oil pitch and nonpolymerized oils are obtained. The total yield of primary and secondary pitch is ca. 75 %, compared with 50 – 55 % for tar distillation without intensified thermal treatment.

Whereas in the Cherry-T process, electrode pitches are the main products of primary processing, pitch coke is the main product when tar distillation is coupled with a *delayed coker* (→ Oil Refining, Chap. 3.6.4.1.). Some Japanese plants use such a combination (see Section 3.3). Here only light, carbolic, naphthalene, and wash oils are separated by primary distillation. Soft pitch obtained as residue is pumped into the bottom of a combination tower where it is distilled together with coker oil. The bottom product of the combination tower is transferred to the pipe still of the delayed- coking plant where it is heated to almost 500 °C and pumped into the delayed- coking drum (see Section 3.3).

Nondistillation processes for the primary processing of coke-oven coal tar, such as liquid or supercritical extraction and hydrocracking, are presently not used in large-scale production.

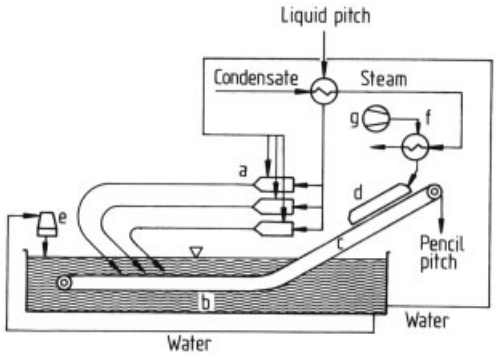

Figure 6. Direct cooling of coal-tar pitch (Rütgers pencil-pitch process)
a) Nozzles; b) Cooling water basin; c) Conveyor belt; d) Dryer; e) Cooling facility; f) Air preheating; g) Blower

3.3. Processing of Coal-Tar Pitch

3.3.1. Cooling

Approximately 50 – 55 % of the products of a coal tar primary distillation is a coal-tar normal pitch with a softening point of 68 °C (Krämer – Sarnow). This is either processed directly in the liquid state or solidified by cooling, which converts the thermoplastic pitch into its glassy state.

Modern pitch- cooling processes operate continuously and use water as cooling medium. The pitch is obtained in granulated form.

The problem of continuous pitch cooling is to obtain a rapid heat transfer despite the relatively low thermal conductivity of pitch. If the granules do not cool completely, undesired agglomeration during storage and transport can occur. For various applications the pitch granulate must be moisture-free; therefore, cooling is followed by drying. Several continuous processes for direct cooling of molten coal-tar pitch with water have been developed.

Figure 6 shows the diagram of the Rütgers pencil-pitch process [16]. Relatively thick strings are formed by pumping pitch through

heated nozzles directly into water flowing cocurrently. The water acts as cooling and transport medium. Before the liquid pitch is formed by the nozzles, it is cooled in a heat exchanger that produces low-pressure steam. The nozzles are shaped such that by adjusting the temperature and water/pitch ratio, a pitch string without internal water inclusion is obtained. The cooling water and the solidified pitch string run together through a cooling tube 20 – 25 m long. This tube may be bent to shorten the equipment. Each nozzle has its own cooling tube to prevent the warm, thermoplastic pitch strings merging. At the end of the cooling tube the outer skin of the pitch strings at least has cooled and solidified. The strings can be deposited together on a slowly moving conveyor belt running submerged in a long cooling-water basin. The cooling water is cooled and recycled. At the end of the cooling section the conveyor belt removes the cold, solidified pitch strings from the water. They pass a dryer whose airstream is heated by the low-pressure steam produced in the heat exchanger. The dried pitch strings are air-cooled and then loaded into transport and storage containers. During loading the pitch strings break into short, air- conveyable pieces. This process is also applicable to other pitches with different softening points. The throughput capacity is 1 – 2 t per nozzle per hour.

3.3.2. Production of Electrode Pitch

Electrode pitches are the most important products derived from coal-tar pitch. They are used as reactive binders, especially for the production of carbon anodes for aluminum production by electrolysis (→ Aluminum, Chap. 4.2., → Aluminum, Chap. 4.3.) and furthermore as binder and impregnating pitches for the production of graphite electrodes for electric steelmaking (→ Carbon, Chap. 4.1.; → Steel, Chap. 3.3.2.).

Electrode pitches should have good caking and binding properties towards electrode cokes (petroleum cokes, pitch cokes), the solid components of the electrode paste. On heating, electrode pitches should give a high yield of a high-strength coke. Medium and higher molecular pitch fractions are mainly responsible for the caking and binding properties. They are analyzed by solvent fractionation (e.g., quinoline insolubles, QI; toluene insolubles, TI; β-resins = TI − QI). The volatile matter of the electrode pitch should degas smoothly to prevent cracking during carbonization of the electrode paste. Moreover, electrode pitches should wet electrode coke particles well and should have a desirable viscosity – temperature behavior and their properties should not change significantly on hot storage at temperatures of, e.g., 200 – 250 °C. The content of ash-forming constituents should be low. Table 4 lists typical characteristics of various electrode pitches. Bench-scale processes have been developed that allow the quality properties of electrode pitches to be determined more precisely. It is now possible to produce test electrodes with electrode coke on a laboratory scale. These electrodes are tested according to various methods [17], [18]. Quality aspects of the test electrodes include high mechanical strength, high electrical conductivity, and low air permeability, as well as low air reactivity at, e.g., 525 °C and reactivity to CO_2 at, e.g., 1000 °C [19], [20].

The properties required for electrode pitch are generally not fulfilled by normal pitch from conventional coal-tar primary distillation. The aim of production processes for electrode pitch is to increase the pitch coking value. This is achieved either by thermal condensation and polymerization of pitch components in secondary processes or by gentle distillation of vaporizable pitch components integrated with the tar distillation. In both processes the softening point and the content of medium-molecular β-resins are increased, and in the thermal condensation and polymerization process, the amount of quinoline insolubles rises. These "secondary" quinoline insolubles, generated during thermal retreatment, mainly consist of liquid crystalline carbonaceous mesophases. These mesophases advantageously increase the coking value of total pitch, but at the same time generally decrease its wetting ability. Therefore, in many cases electrode pitches produced without mesophase formation are preferred for electrode production [21].

Figure 7 shows the flowsheet of a thermally gentle electrode-pitch production. The process is part of the coal-tar primary distillation. In this case the coal-tar primary distillation (see Fig. 5) is extended by a flash column. By regulation of

Table 4. Specifications of various electrode pitches from coal tar

	Electrode pitches for aluminum		Electrode pitches for electrographite	
	Prebaked anodes	Søderberg anodes	Binder pitches	Impregnating pitches
Softening point, °C				
Krämer – Sarnow	88 – 103	108 – 112	90 – 95	68 – 74
Mettler	106 – 121	126 – 130	108 – 113	87 – 93
Density at 20 °C, g/cm^3	$\geq$ 1.31	$\geq$ 1.31	$\geq$ 1.31	
Quinoline insolubles, %	5.5 – 13	$\leq$ 14	7 – 12	$\leq$ 5
Toluene insolubles, %	23 – 36	$\geq$ 32	27 – 34	16 – 22
β-Resins, %	18 – 26	$\geq$ 22	20 – 22	11 – 18
Coking value (Alcan), %	53 – 63	$\geq$ 57	56 – 63	$\geq$ 43
Ash (800 °C), %	0.15 – 0.35	0.15 – 0.35	0.15 – 0.35	$\leq$ 0.53
Sulfur, %	$\leq$ 0.7	$\leq$ 0.7	$\leq$ 0.7	$\leq$ 0.7
Sodium, ppm	$\leq$ 200	$\leq$ 200	$\leq$ 200	
Iron, ppm	$\leq$ 200	$\leq$ 200	$\leq$ 200	
Silicon, ppm	$\leq$ 200	$\leq$ 200	$\leq$ 200	
Distillation up to 360 °C (ASTM D 2569), %	$\leq$ 3	$\leq$ 3	$\leq$ 3	*

* Flash point $\geq$ 260 °C (open cup).

temperature and vacuum it is possible to directly control the desired increased softening points and the coking values related therewith.

The Cherry-T process of Osaka Gas (see Chap. 3.2) is a continuous thermal pitch-condensation process forming part of the coal-tar primary distillation. This process gives a high yield of electrode pitch from coal tar in two different qualities. In the first step a primary pitch yield of 55 – 60 % from coal tar is achieved. The pitch has a softening point of 80 – 100 °C, 31 – 38 % toluene insolubles, and 8 – 14 % quinoline insolubles, thus having a β-resins content of ca. 22 %. In the second step, further thermal treatment of the flash oil from the first reactor gives 40 – 50 % of a secondary pitch with a softening point of 70 – 90 °C, 23 – 31 % toluene insolubles, and a very low amount of quinoline insolubles (max. 2 %) [22].

Not only the production conditions for electrode pitches but also the choice of parent coal tars is important. The desired pitch components may already have been formed to a large extent in the coal tar. Moreover, the tars have to meet the requirements concerning the content of inorganic contamination, which very often require a separation of these microparticles by centrifugation or extraction. For use as binder pitches for carbon anodes in aluminum production, for example, a low content of zinc and iron compounds is important. Under the conditions of aluminum oxide electrolysis in the melt, elemental metals are formed together with aluminum at the cathode, thus reducing the purity of the aluminum product. Furthermore, the content of alkaline earth and alkali metal compounds should be as low as possible because they catalyze the reaction of CO_2 with the carbon anode. In the case of impregnating pitches it may also be necessary to partly remove the highest molecular mass quinoline insoluble α-resins from coal tar, as the impregnating behavior depends upon the content of α-resins and upon their particle size distribution. On the other hand, α- and β-resins considerably contribute to the stability of the pitch coke skeleton formed from electrode pitch, thus giving the mechanical and thermal stability of carbon and graphite electrodes.

For large-scale production of electrodes, bituminous binding agents other than coal-tar pitch have not gained acceptance. The reason is the good coking behavior of coal-tar electrode pitches due to the high content of condensed aromatics and fixed carbon. Coal-tar pitch has the highest carbon content of comparable bituminous binding agents. In contrast to petroleum asphalt and coal extracts, the carbon content of coal-tar pitch is considerably above 90 %, on average ca. 93 %. The atomic C/H ratio is 1.7 – 1.8 compared with, for example, 1.3 for coal extract from the solvent refined coal (SRC) process and 0.8 for high-vacuum bitumen with a comparable softening point [23]. It is only possible to produce aromatized petroleum pitches from

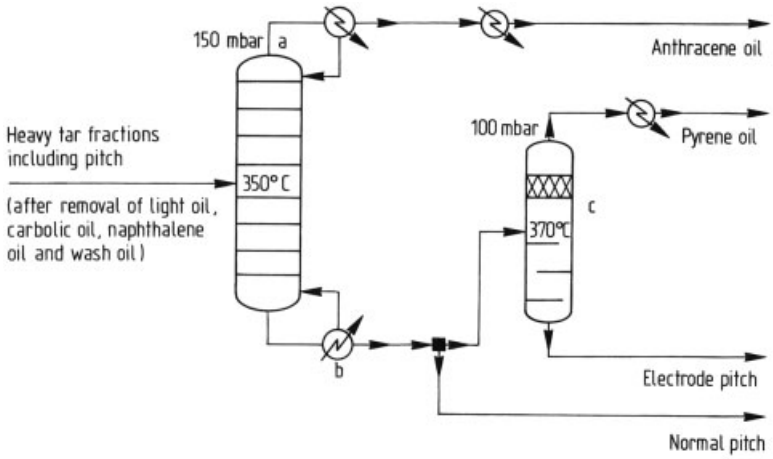

Figure 7. Flowsheet of a coal-tar primary distillation with integrated electrode-pitch fabrication (Rütgers process)
a) Pitch/anthracene oil column; b) Reboiler; c) Flash column

strongly aromatized residues of catalytic petroleum cracking by thermal treatment. These may serve as substitutes for coal-tar electrode pitches in some situations [24], [25].

Despite the advantages of coal-tar pitch as electrode binder, it does have environmental and toxicological drawbacks. Coal-tar pitch has a high content of cancerogenic polycyclic aromatics, especially benzo[*a*]pyrene. However, the uptake of harmful polycyclic aromatic hydrocarbons can be avoided if pitch fumes and dusts are not handled or inhaled. Nevertheless, attempts have been made to reduce the benzo[*a*]pyrene content of electrode pitches. The physical removal of the cancerogenic polycyclic aromatic hydrocarbons by extraction or distillation is technically possible. The softening point of the pitch increases on removal of these low-molecular compounds. Bench-scale experiments showed that it is possible to obtain carbon electrodes with improved technical properties when using such high-melting electrode pitches [26]. In an electrode pitch with an increased melting point of 170 °C, for example, the benzo[*a*]pyrene content is decreased by $> 60\,\%$, down to $0.3 - 0.4\,\%$. A further increase of the melting point up to, for instance, 260 °C by a $> 99\,\%$ reduction of the benzo[*a*]pyrene content to less than 100 or 50 ppm is possible in principle. For practical applications, however, modification of the electrode production process by using a different mixing technology for the electrode masses, or adjustment of the softening point is necessary [27].

3.3.3. Production of Pitch Coke

Like calcined petroleum coke, pitch coke is used as electrode coke, the solid component for the production of carbon and graphite electrodes. In addition, technically pure carbon pitch coke is used to produce other carbon and graphite materials, especially for high-temperature and chemically resistant equipment. It is also used as conductive filler in carbon brushes, batteries, plastics, and brake linings; as a carburizing agent in steel and cast-iron production; as adsorption coke for water purification; as an almost ash-free solid reducing agent in metallurgy and chemistry; and for the production of highly isotropic graphite for nuclear reactors. Specifications of different pitch cokes are listed in Table 5.

Table 5. Properties of pitch cokes

		Regular coke	Needle coke
Density, g/cm^3		2.04 – 2.10	$\geq$ 2.12
Microstructure		coarse mosaic or fine isotropic	coarse fibrous
Coefficient of linear thermal expansion *	10^{-6} K^{-1}	2.5 – 5.5	< 1
Ash content, %		< 0.5	< 0.2
Elemental analysis			
C, %		$\geq$ 98	> 98
S, %		0.4 – 0.5	< 0.3
N, %		0.5 – 0.7	< 0.5
H, %		0.1 – 0.2	0.1 – 0.2
Si, %		ca. 0.03	< 0.03
Fe, %		ca. 0.03	< 0.03
V, ppm		< 5	< 5

* Measured on graphite rods, 20 – 200 °C.

The most important processes for producing pitch coke are:

1) Discontinuous high-temperature coking of hard pitch in normal coke ovens with gas heating and liquid feeding, used especially in the CIS.
2) Discontinuous coking of soft pitch by the delayed- coking process in Japan and China.
3) Continuous hard-pitch coking in an indirectly heated rotary-drum reactor (still in development).

Hard pitch with a softening point > 150 °C (Krämer – Sarnow) can be obtained either from the normal pitch produced by continuous tar distillation by dehydrogenative condensation with atmospheric oxygen (air blowing) or solely by thermal treatment, similar to electrode-pitch production processes.

In the air-blowing process, air is blown through liquid normal pitch at 300 °C in a retort directly after coal-tar primary distillation [28]. There is an air distributor at the bottom of the retort, and the air is either sucked or blown through the pitch. The reaction is exothermic above 300 °C and takes place without additional heating. The temperature of the retort contents rises to ca. 400 °C. The main product is a hard pitch which can be processed to coke with a yield of up to 83 %. The byproduct pitch distillate can be converted into a secondary hard pitch by distillative thermal treatment and air blowing. By using a mixture of primary and secondary hard pitch, different pitch- coke qualities can be obtained in the subsequent coking process.

The thermal condensation of normal pitch to hard pitch is usually carried out by continuous treatment of normal pitch in pipe stills under pressure followed by flash distillation to remove vaporizable components. Compared to the air-blowing process, the yield of hard pitch is lower, but the yield of high-boiling aromatic oils is higher.

In discontinuous *pitch coking in horizontal coke ovens* hard pitch is pumped at a temperature of above 300 °C into heated pitch tanks. In the Koppers process [29] several oven blocks, each with, e.g., five pitch- coke ovens, can be fed by these tanks. The hard pitch is pumped continuously via a ring main from the pitch tanks to the oven blocks. The liquid hard pitch is pumped in portions over a period of several hours into the coke ovens from this ring main. The final coking temperature can be up to 1200 °C, and the coking time 14 – 24 h. The pitch- coke yield is ca. 70 %. The pitch- coke oil byproduct can be used for secondary-pitch production. By mixing this pitch with the primary pitch, the quality of pitch coke can be varied and the pitch- coke yield increased to 80 – 85 %. Under these conditions ca. 250 m^3/t process gas, containing ca. 80 vol % hydrogen and ca. 12 vol % methane, are generated as a byproduct. This gas can be used for the production of hydrogen, or as low- calorific gas for heating the coke ovens and pitch tanks. With heat requirements of ca. 18 MJ per tonne of hard pitch, about 80 % of the process gas is required. Analogous to metallurgical coke production, after carbonization the pitch coke is pushed mechanically out of the horizontal chamber of the coke oven and cooled in aquenching car by spraying with a calculated amount of water. It is possible to achieve a water content of 1 – 2 % in spite of the high porosity of pitch coke. Due to the high carbonizing temperature, the pitch coke produced does not require secondary calcination.

The second process for production of technical pitch- coke from coal tar is *delayed coking*, well known from petroleum refining (→ Petroleum, Coke). To produce a pitch coke with a low volatiles content a second calcination step is required. Soft pitch containing high-boiling aromatic oils, obtained by primary distillation of coal tar, is used as feed. The soft pitch is

pumped into a combination column, the bottoms of which are fed to the pipe still of the delayed coking plant.

Figure 8 shows the process flow sheet for production of pitch coke from coal-tar soft pitch by delayed coking using the Lummus process as applied in Japan and China [30]. The soft pitch coming from the coal-tar primary distillation at 150 °C is heated in a heat exchanger by the heavy oil of the combination column and then pumped into the bottom of this column. The coker oil of the delayed- coking plant is also added to the column. Light oil, middle (naphthalene) oil, and heavy (anthracene) oil fractions are taken from the combination column. The bottoms of the column are heated in a pipe still to 485 – 510 °C by gas or fuel oil. Then they are fed from below into one of two coker drums via a switching valve. In these drums the delayed coking of pitch takes place during a 24 h cycle with a residence time of 12 h at a pressure of, for example, 5 bar. When the first coker drum is full, the feed is switched to the preheated empty coker drum. After switching, the full drum is steamed for 2 h to remove residual coker oil, which is mainly present in the upper part of the coker drum. The oil obtained is recycled to the combination column. After steaming, the coke charge is cooled directly with water. Then the coke is cut out by a high-pressure water jet, which takes 5 – 7 h. The green coke obtained is dried, broken, and calcined in a hearth furnace or a rotary drum. The pipe still and the calciner can be heated with coker gas arising as the noncondensable top product of the combination column.

Normally the green- coke yield is at least 60 % of the soft pitch fed, with approximately 3 % process gas and 35 % coker oil as byproducts. The combination column yields ca. 10 % light and middle (naphthalene) oil, as well as ca. 25 % heavy (anthracene) oil. The coker gas contains ca. 50 % hydrogen and 50 % methane. The middle oil has a boiling range of 180 – 310 °C and contains ca. 30 % naphthalene. The heavy oil has a boiling range of 265 – 400 °C and a composition that differs from that of an anthracene oil from normal continuous tar distillation. Under the conditions of pressure carbonization in the delayed coker, some distillable tar aromatics are preferentially converted to coke. The green coke still has 7.5 – 9.5 % volatiles, which are removed to a residual content of < 0.5 % by calcination at up to 1300 °C.

Highly anisotropic cokes, known as *needle cokes*, are obtained via intermediate liquid-crystalline mesophases by delayed coking of pitches from which the α-resins (quinoline insolubles) have been largely removed by filtration [31] or extraction [32]. These needle cokes are used for the production of ultra-high power (UHP) graphite electrodes with high electrical conductivity and low thermal expansion coefficient.

The carbonization of hard pitch in an indirectly heated *rotary drum reactor* [33], [34] is a continuous pitch- coke production process. Hard pitch is pumped into an indirectly fired rotary drum reactor where it is carbonized with a residence time of 0.5 – 1.5 h at an inner-wall temperature of 500 – 800 °C. The gases and vapors arising are pumped in countercurrent to the carbonizing pitch, if necessary together with an inert gas. The green- coke yield is 70 – 75 % of the hard pitch feed. The grained coke has volatile matter content of ca. 3.5 wt % and relatively high strength and density. It is calcined without precooling in a second directly heated rotary drum reactor at up to 1300 °C. The yield of calcined coke is ca. 90 % of the green-coke feed, and 65 – 70 % of the hard pitch input. The advantages of this continuous process compared with the discontinuous processes of horizontal chamber coking and delayed coking are the possibility to adjust the residence times and heating periods, thus yielding a more homogeneous pitch coke.

3.3.4. Production of Special Pitches

Hard pitches, soft pitches, and extractive pitches produced by solvent extraction with considerably decreased content of α-resins (primary quinoline insolubles) are special pitches that are intermediate products of pitch carbonization. Other special pitches include plasticized pitches for protective coatings, pitches used for pitch – polymer combinations, and road-binding agents which are compatible with polymers and bitumen, as well as the newly developed and produced mesophase pitches and high-melting isotropic pitches for further processing to carbon materials.

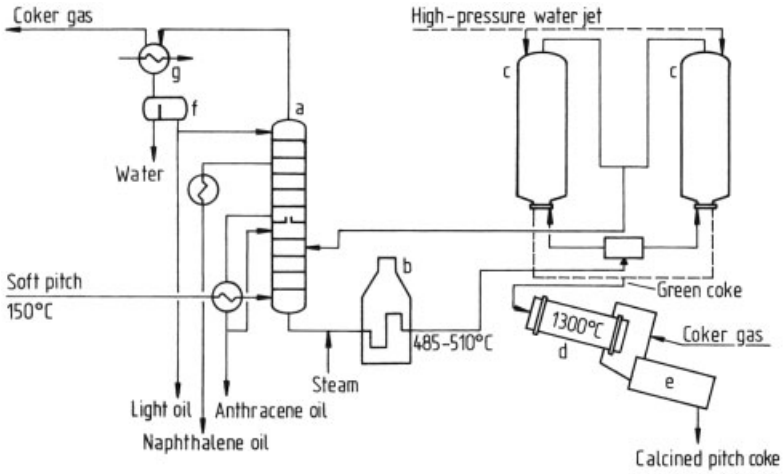

Figure 8. Production of pitch coke by delayed coking
a) Combination column; b) Pipe still; c) Coker drums; d) Rotary kiln; e) Rotary cooler; f) Decanter; g) Condenser

Soft pitches with increased coking values are used as *refractory binders*. They are reactive binding agents for the production of dolomite and magnesite bricks used especially for the lining of steel converters. During brick calcination a binding coke skeleton is formed by carbonization of the soft pitch, which considerably increases the compressive strength of the preform and counteracts corrosive attack by slags. Refractory binder pitches with increased coking values are obtained by adjusting hard pitches with high-boiling aromatic oils to give the desired softening point. Also soft-pitch mixtures with phenolic resins or their precursors are used. Special formulations contain less than 50 ppm benzo[*a*]pyrene.

For the production of *anticorrosion protective coatings* (→ Paints and Coatings, Chap. 2.14.4.) the plasticity range (difference between softening and brittle point) of 30 K of coal-tar normal pitch is too low. To increase this plasticity range to 100 K or more, hard pitches with a high content of toluene insolubles are adjusted to the desired softening point with high-boiling tar oils, preferably by adding crystalline tar aromatics. A further increase in the plasticity range is obtained by hot-mixing these pitches with extenders such as finely ground coal, minerals, diatomaceous earth, or fly ash. Such plasticized pitches are used for long-term anticorrosion pipe coatings and as sealing compounds. The advantages compared with the bitumen coatings are the excellent resistance of coal-tar pitch to moisture, bacteria, fungi, mineral oils, and gasoline. When pitch coatings are used together with cathodic pipe protection the low water absorption of coal-tar pitch is important. The water content of a 2.5 mm tar-pitch coating is only 0.3 – 0.6 % after prolonged storage in water. Plasticized pitches without extenders are used as dip- coating agents for protecting cast-iron pipes.

To meet especially high anticorrosion requirements, coal-tar pitches are combined with polymers [35]. Especially suitable for such *pitch – polymer- combinations* are epoxy resins, polyurethanes, and some elastomers (→ Paints and Coatings, Chap. 2.14.5.). Normally two-pack systems are produced for epoxy resin combinations. One component contains the nonhardened epoxy compound dissolved in a compatible special coal-tar pitch, and the other contains the hardener for the epoxide compound. After combination of the components the epoxy resin polymerizes and incorporates the pitch components. Usually low molecular-mass liquid epoxy resins are combined with soft coal-tar pitches. After chemical hardening, coatings with high-temperature stability and resistance to chemicals are obtained. Pitch – polyurethane two-pack systems are produced and processed similarly. One component contains the special soft coal-tar pitch and the polyol, and the other the polyisocyanate. In combination with polyurethanes, one-

pack systems that cure by reaction with atmospheric moisture are also feasible. Other polymers and elastomers are mixed directly with the hot liquid pitch, whereby miscibility gaps and the decomposition temperatures of the polymers must be taken into account. It is also possible to add high-boiling or unvaporizable, preferably aromatic homogenizing agents.

By using high-boiling tar aromatics as homogenizing agents, soft coal-tar pitches with a low content of α-resins can be combined in any ratio with appropriate, preferably naphthene-based bitumens. Carbobitumen, a highly deformation resistant *road binding agent*, is a blend of soft pitch and hard bitumen, containing 20 – 30 % of a special pitch. The pitch is enriched with high-boiling crystalline aromatics [36], [37]. These crystalline aromatics cause a separation of microcrystals in the pitch – bitumen matrix when cooled to ca. 100 °C. These microcrystals cross-link the thermoplastic system and harden the binding agent at lower temperatures. Thus carbobitumen is harder at temperatures below 70 – 80 °C and, therefore, more deformation resistant than the road bitumen usually used. At higher temperatures, however, it is less viscous and, therefore, can be applied at lower temperatures than road bitumen.

Normal pitches, hard pitches, and special pitches with carbon fillers and polymers are suitable as reactive additives for *cocarbonization of weak- coking coal blends* in amounts of 2 – 10 %. The mechanical strength of the metallurgical coke produced in horizontal chamber-coke ovens increases on addition of an amount of pitch ascertained by laboratory tests [38], [39]. With increasing temperature the pitch wets the coal grains in the coke-oven chamber. It partially dissolves the coal macerals in the softening range of the coal. During carbonization above 450 °C inert macerals and solid leaning agents such as coke dust or petroleum coke are also brought into the generated coke skeleton. Used plastics, for example, can be utilized as additives. These polymers should first be subjected to cothermolysis with pitch, to obtain a resistant reactive additive as liquid phase in the coal-softening range [40].

Further special pitches are fine-grained *foundry pitches* produced by spraying. They have a high softening point and high coking value. They serve as binding agents for the production of foundry molds. Another special pitch is *fiber-pipe pitch* which has a low content of α-resins. This is used as impregnating material for cellulose fibers in the production of pitch fiber pipes. *Clay pigeon pitch* is used as a brittle binding agent with increased softening point for clay pigeons used in sport shooting.

Mesophase pitches are pitches with a high content of carbonaceous mesophases (see Chap. 2). They are produced by heat treatment of pitches that have been filtered or extracted to remove primary α-resins [31]. By heat treatment, mesophase spherules are formed, which can be separated from the isotropic pitch matrix by settling and/or by solvent fractionation. Pitch pre-fractionation with solvents or supercritical extraction is often carried out before thermal mesophase formation. Two-step thermal pitch treatment is also utilized. In the first step the especially reactive pitch aromatics are separated by condensation to give secondary quinoline insolubles. The second thermal treatment step leads to mesophases with low to medium molecular mass and correspondingly low viscosity, which can be advantageous. Mesophase pitches can also be produced from special petroleum pitches, hydrogenated coal-tar pitches, and pyrolysis residual pitches. By using different processes special properties can be achieved [41]. Mesophase pitches also can be produced by catalytic polycondensation of naphthalene and its homologues and higher condensed polycyclic aromatic hydrocarbons in the presence of homogeneous Friedel – Crafts catalysts. Thus, after 2 – 17 h at 140 – 180 °C 80 – 100 % mesophase pitches with a yield of 50 – 80 % are formed from naphthalene, 1,1-binaphthylanthracene, and 1,2-benzanthracene and from a 1: 1 mixture of naphthalene and phenanthrene by reaction with aluminum chloride/copper(II)chloride [42]. From naphthalene it is possible to produce mesophase pitches by using HF/BF_3 catalysts, either by one step at 260 – 300 °C, or by two steps, the first at 80 °C and the second at 400 – 480 °C under vacuum or pressure [43]. Instead of naphthalene, 1-methylnaphthalene, 2-methylnaphthalene, or mixtures thereof can also be used. Here mesophase pitches are formed by using HF/BF_3 at 260 °C under a pressure of ca. 30 bar in 4 – 5 h with a yield of 80 – 90 % [44].

Mesophase pitches and high-melting isotropic pitches are intermediates in the production of new forms of technical carbon such as anisotropic and isotropic carbon fibers, mesocarbon microbeads, and sinterable carbon powders [41]. In their production the special pitches are normally first stabilized by oxidation after preforming by melt spinning or fine grinding and are then partly or entirely carbonized. For the production of high-modulus carbon fibers from mesophase pitches various processes have been developed (→ Fibers, 5. Synthetic Inorganic, Chap. 5.4.). Some of these processes have been used commercially. Mesophase microbeads and sinterable carbon powders can be pressed isostatically without additional binding agents and converted by thermal treatment into highly dense isotropic carbon materials [41]. In the last few years isotropic fine-grained carbon materials have found a wide range of applications, as have carbon fiber reinforced carbon materials (CFC) [45]. Widespread areas of application for these carbon materials can be predicted for the future, especially in energy and transport technology. This is due to their low density compared to metals such as steel and aluminum, their mechanical strength over a wide temperature range, and their excellent properties at temperatures up to 4000 °C if protected against oxidation.

3.3.5. Pitch – Oil Blends

Coal-tar pitches are highly soluble in various tar oils. For manufacturing of pitch – oil blends (refined tars) the liquid pitch is mixed with tar oils.

Road tars based on pure coal tars are blends of normal pitch with middle oils (boiling range 170 – 270 °C), heavy oils (boiling range 270 – 300 °C), and anthracene oils (boiling range above 300 °C). The content of normal pitch is between 60 and 80 %. These road tars are suitable for surfacing low-traffic roads. The advantages of road tars for this purpose are good adhesive strength, quick rock wetting due to their content of polar compounds, and their property of gradual surface resinification. The middle and heavy oil components evaporate during setting, leaving as the actual binder a soft pitch consisting of anthracene oil and normal pitch. The setting speed depends on the initial viscosity and the pitch/anthracene oil ratio. Normal setting road tars have a pitch/anthracene oil ratio of ca. 3.5: 1. By increasing the ratio of anthracene oil II (boiling range above 350 °C) and anthracene oil I (boiling range up to 350 °C), aging-resistant road tars can be produced. Bitumen- containing road tars include, for example, 15 % asphalt basic distillate bitumen with low oil content; bitumen-rich road tars contain 35 – 40 % bitumen, and *pitch bitumen* contains 70 – 85 % bitumen. These tar – bitumen blends with increased or predominant bitumen components combine the advantages of tar (good setting properties, resistance to mineral oil and fuel, easy processability, and good road grip) with those of bitumen (low temperature dependence of viscosity and wide plasticity range). Carbobitumen is a specially deformation resistant pitch – bitumen blend (see Section 3.3.4).

Besides bitumen, plastics such as poly(vinyl chloride), polystyrene and epoxy resins are used for the modification of road tars. Thermosetting tar – epoxy resin combinations (see Section 3.3.4) are particularly suitable for highly exposed road surfaces and airport runways due to their good wear properties and their resistance to liquid fuels.

Roofing tars used as impregnating, coating, and adhesive material for tarred felts and tarred sealing webs are usually blends of pitch and filtered anthracene oil. By using plasticized pitches or by adding extenders the plasticity and temperature stability of roofing tars is improved considerably. Tarred felts produced in this manner have the same temperature stability as bitumen felts. Moreover, they have a high resistance to water, microorganisms, and plant roots.

Pitch paints are solutions of coal-tar pitch in highly volatile aromatic solvents such as light and heavy tar-oil naphtha. Normally plasticized pitches are used or extenders are added. Like the plasticized pitches themselves, which are processed while hot, pitch paints have excellent resistance to water, bacteria, and chemicals. Anticorrosion and building-protective coatings are applied, for example, as underwater protection and as protective paints for steel, concrete, and underground brickwork [35].

Effective anticorrosion protective coatings are obtained by combining pitch with polymers. The best-known *pitch combination coatings* are two-pack systems based on soft pitches and epoxy resins (see Section 3.3.4). The coatings

are resistant both to low temperatures and high temperatures up to 100 °C. They are resistant to fresh and salt water, moderately strong mineral acids and alkalis (except ammonia), salt solutions, lubricating oils, and nonaromatic motor fuels. Two- and one-pack pitch – polyurethane coatings have a similar range of applications.

3.4. Processing of Tar Distillates

This chapter deals with the further processing of primary tar distillates to fractions and the various aromatic oils. The processing to technically pure aromatic compounds is dealt under separate keywords. Only the processing to technically pure phenols and pyridine bases is described in more detail here.

3.4.1. Processing of Anthracene Oil

In the primary distillation of coal tar, ca. 20 – 30 % of the crude tar distills as anthracene oil, which at room temperature is a semisolid, green-brown, crystalline material with a boiling range of ca. 300 – 450 °C. In most primary distillation plants two fractions are produced: the larger fraction being low-boiling anthracene oil I, and as the smaller, high-boiling anthracene oil II. Anthracene oil I contains phenanthrene, anthracene, and carbazole, whereas fluoranthene and pyrene are concentrated in anthracene oil II.

The crude anthracene oil or anthracene oil I is cooled in water- cooled discontinuous or continuous crystallizers and partially crystallized to give a filtered anthracene oil and anthracene residues. Filtered anthracene oil does not crystallize at normal temperature. The discontinuous cooling process is usually used. Anthracene oil at 80 – 100 °C is pumped from primary distillation into cylindrical vertical stirred coolers. It is cooled over 24 – 36 h to 20 – 30 °C by indirect water cooling. Early sedimentation of crystals is prevented by a slowly rotating stirrer (ca. 15 rpm). After cooling the crystal slurry is discharged via a gate through the conical lower part of the cooler and centrifuged with a peeler centrifuge. The crude anthracene oil yields 10 – 15 % anthracene residues and 85 – 90 % filtered anthracene oil. The residues are further processed to anthracene, phenanthrene, and carbazole.

Filtered anthracene oil I and high-boiling anthracene oil II are important blending components for technical *aromatic oils* and pitch – oil mixtures. Aromatic oils with anthracene oil as main component include carbon-black oils, impregnating oils (creosote), medium fuel oils, special diesel fuels, gasometer oils, fluxing oils, naphthalene wash oil and numerous others. Table 6 lists specifications for some of these aromatic tar oils.

Carbon-black oils. Due to their low sulfur content and high aromaticity, coal-tar oils are high-value raw materials for carbon black. Carbon-black oils obtained from coal tar have, due to their high content of condensed aromatics, the highest Correlation Index (Bureau of Mines). They are pyrolyzed at ca. 1500 °C by the furnace-black process to yield filler blacks, mostly for rubber, and pigment blacks, mainly for printing inks. Depending on the required quality of the carbon black the yield is 40 – 70 % (→ Carbon, Chap. 6.).

Impregnating oils (creosote) obtained from coal tar are important for long-term wood preservation (→ Wood, Preservation, → Wood, Surface Treatment). The aromatic components of anthracene oil and other tar-oil components have, when mixed, strong antifungal properties and significant pesticidal properties for wood pests such as the house long horn beetle (*Hylotrupes bajulus*), the ship worm (*Toredo navalis*), anobium beetles (*Anobium punctatum*) and termites, and furthermore inhibit dry-rot fungi. Railway sleepers, telegraph poles, pit props, timbers for use in water and other building timbers are protected against decay and insect damage over a long term by vacuum – pressure impregnation with tar oils. Coal-tar impregnating oils are not washed out by water, are noncorrosive, and have high electrical resistance. The service life of such impregnated woods is prolonged from 5 – 15 years to 30 – 50 years. The various types of impregnating oils are blends of filtered anthracene oil, wash oil, and filtered naphthalene oil.

Fuel oils based on coal tar have a short, hot flame with high radiation intensity due to their high carbon content. Because of these properties they are used in various metallurgical plants. Coal-tar fuel oils are produced in varying types

Table 6. Specifications of various coal-tar oils

	Impregnating oil for railway sleepers (Deutsche Bundesbahn)	Impregnating oil for telegraph poles (Deutsche Bundespost)	Impregnating oil AWPA (P 1 – 78) (USA)	Carbon-black oil (light quality)	Carbonblack oil (heavy quality)	Fluxing oil	Light fuel oil (DIN 51 603)	Diesel fuel	Benzole absorbing oil
Density at 20 °C, g/cm^3	1.04 – 1.15	1.04 – 1.15	at 15.5 °C ≥ 1.050	1.07 – 1.1	1.1 – 1.15	1.02 – 1.07	1.02 – 1.10	1.04 – 1.10	1.02 – 1.06
Clear point or free of sediment, °C	≤ 23	≤ 23		< 40	< 50	0	0	0	< 0
Water content, %	≤ 1	≤ 1	≤ 1.5	≤ 0.2		≤ 0.5	≤ 0.3 nonsettleable	≤ 0.3	≤ 0.1
Toluene insolubles, %	traces	traces	≤ 0.5	≤ 0.1				≤ 0.1	
Coking value [a], %				≤ 1			≤ 1	≤ 0.5	
Ash content, %							a) ≤ 0.01 [d] b) ≤ 0.04	≤ 0.01	
Phenols, %	< 3	< 2.3							
Sulfur, %				≤ 0.8	≤ 0.8		a) ≤ 0.3 [d] b) ≤ 0.8		
Correlation index [b]				≥ 140	≥ 150				
Flash point [c], °C						> 55	> 85	> 80	
Calorific value							≥ 37.7 MJ/kg	≈ 37.7 MJ/kg	
Boiling range, %	up to 235 °C: ≤ 10 up to 300 °C: 20 – 40 up to 355 °C: 55 – 75	up to 235 °C: ≤ 10 up to 300 °C: 40 – 60 up to 355 °C: 70 – 90	up to 210 °C: ≤ 2.0 up to 235 °C: ≤ 12.0 up to 270 °C: 10 – 35 up to 315 °C: 40 – 65 up to 355 °C: 60 – 77 residue: 23 – 40 density at 15.5 °C fraction 235 – 315 °C: ≥ 1.027 g/cm^3 fraction 315 – 355 °C: ≥ 1.095 g/cm^3	200 – 400 °C up to 200 °C: ≤ 20 up to 300 °C: ≥ 60 up to 350 °C: ≥ 70 up to 400 °C: ≥ 90	250 – 450 °C	180 – 360 °C		up to 250 °C: ≤ 55 up to 300 °C: ≤ 60	230 – 300 °C

[a] Conradson.
[b] Bureau of Mines.
[c] Pensky – Martens.
[d] a) and b) are two types of fuel oil.

which differ in boiling range, density, viscosity, and sedimentation behavior. Table 6 lists data for a light fuel oil. Heavy fuel oils for steel works may also contain pitch, which itself is used as extraheavy fuel oil. Coal-tar fuel oils have a minimum calorific value of ca. 38 MJ/kg.

Diesel fuels manufactured from coal tar are suitable for the operation of stationary, slowly running diesel motors. They are blends of filtered anthracene oil (< 40 %) and tar oil boiling below 300 °C (> 60 %). This diesel fuel contains no active sulfur (copper strip test) or corrosive chlorine (< 0.01 %) and has, like all coal-tar oils, a minimum calorific value of ca. 38 MJ/kg.

Fluxing oils for bitumen are, depending on their quality, blends of low- and high-boiling tar oils. They decrease the softening point and the viscosity of bitumen and, therefore, lower the processing temperature of road bitumen (cutback bitumen).

Other technical aromatic oils with coal-tar anthracene oil as main component are applied as solvents, flotation oils, quenching agents, and plasticizers.

3.4.2. Processing of Wash Oil

Wash oil, or Solvay oil, which makes up 7 – 8 % of the crude tar, is obtained by coal-tar primary distillation directly as marketable benzol wash oil. It boils between 230 and 290 °C and contains, besides slight amounts of naphthalene, mainly the naphthalene homologues, indole, diphenyl, acenaphthene, dibenzofuran, fluorene, ca. 2 % phenol homologues, and 4 – 6 % quinoline bases. If suitable rectifying conditions are used, crystals will not settle out from wash oil, even at 0 °C. This serves as a good example for the mutual melting-point depression of the tar aromatics, most of which are solids.

Wash oil can be directly used as benzol wash oil. Wash oil and anthracene oil also serve as blending components for technical aromatic oils such as impregnating oil (creosote), diesel fuels, and light fuel oils. In pitch – oil blends wash oil is the heavy oil component and offers good solubility, fungicidal activity, and slow evaporation. Furthermore, wash oil is a starting material for numerous technically pure tar aromatics such as dibenzofuran (→ Furan and Derivatives, Chap. 10.); acenaphthene, indene and fluorene (→ Hydrocarbons, Chap. 5.); indole (→ Indole) and alkylnaphthalenes (→ Naphthalene and Hydronaphthalenes, Chap. 4.).

3.4.3. Processing of Naphthalene Oil

Naphthalene oil is obtained by primary distillation of coal tar, in 10 – 12 % yield based on crude tar. Depending on the separation efficiency of the distillation process the oil fraction, boiling range ca. 210 – 220 °C, has a solidification point of 65 – 75 °C, i.e., a naphthalene content between 73 and 90 %. Besides naphthalene, the oil contains phenols, tar bases, dibenzothiophene, and small amounts of other aromatic hydrocarbons. The naphthalene oil is further processed by crystallization (→ Naphthalene and Hydronaphthalenes, Chap. 1., → Naphthalene and Hydronaphthalenes, Chap. 2.). Phenol homologues (carbolic oil II) are generally also extracted from the filtered naphthalene oil.

3.4.4. Processing of Carbolic Oil

A yield of 2 – 4 % carbolic oil based on crude tar, is obtained by primary distillation of coal tar as a fraction with a boiling range of 180 – 210 °C (carbolic oil I). It contains benzene homologues, some naphthalene, pyridine bases, and ca. 20 – 30 % phenols. The phenols are a mixture of phenol, cresols, and dimethylphenols (xylenols).

Phenol and its homologues are isolated from carbolic oil I, filtered naphthalene oil (carbolic oil II) with a crude phenol content of 15 – 25 %, light oil containing ca. 5 – 10 % phenols, and possibly the wash oil and the fractions obtained from wash oil components with a total content of high-boiling phenol homologues of ca. 2 %. Furthermore, during processing phenols extracted from the water effluents from coke-oven plants are added, either as crude phenol or sodium phenoxide solution. The coal-tar phenols are, apart from some sterically hindered dimethylphenols, readily soluble in dilute sodium hydroxide solution to form sodium phenoxides. For continuous removal of the phenols from phenol- containing oils normally a 8 – 12 % aqueous sodium hydroxide solution is used, which dissolves only slight amounts of hydrocarbons.

Figure 9 shows the processing flow sheet for the isolation of *crude phenol* from tar fractions by extraction with sodium hydroxide solution. The process has four main steps:

1) Oil extraction with dilute sodium hydroxide solution, yielding the crude sodium phenoxide solution
2) Refining the crude sodium phenoxide solution
3) Precipitation of crude phenol from the refined sodium phenoxide solution with carbon dioxide and formation of sodium carbonate solution
4) Regeneration of the sodium hydroxide solution by caustification of the sodium carbonate solution with calcium oxide

The hot carbolic oils from the primary distillation of coal tar and naphthalene crystallization are pumped with a solution of phenols in sodium phenoxide solution, recycled fom the second step, into the mixing chamber of the first extraction step. The oil – alkali mixture is continuously recirculated at ca. 50 °C and kept at a constant level. Part of the mixture is pumped continuously into a separation column from which the heavier crude sodium phenoxide solution is drawn off at the bottom. The carbolic oil is removed at the top. About 4 – 6 % phenol remains in the carbolic oil. In the second extraction step this carbolic oil is extracted analogously to the first step with fresh sodium hydroxide solution to yield a carbolic oil containing less than 0.5 % residual phenols. The solution of phenols in sodium phenoxide solution is returned to wash the fresh carbolic oil in the first step.

The crude sodium phenoxide solution contains approximately 0.5 % dissolved nonphenolic components (hydrocarbons and pyridine bases) which must be removed before the precipitation (saturation or "springing") of the crude phenols. Distillative impurity removal by steam stripping to the clear point is widely used. This takes place in continuously operating packed columns heated at the bottom with direct or indirect steam. To save energy it is favorable to connect several columns in sequence to a multistage evaporator. The vapors taken off at the top of the steam-stripping plant are condensed in a heat exchanger with the crude sodium phenoxide solution and separated into an oil and a water phase. The steam-stripped oil is fed to the light oil processing stage. The dissolved pyridine bases can be extracted from the aqueous phase with benzene.

The hot sodium phenoxide solution (ca. 100 °C) is then pumped to the top of the packed precipitation column (saturation or "springing"). Carbon dioxide is bubbled through the column. The precipitated crude phenol and the sodium carbonate solution formed flow off at the bottom and are separated.

The carbon dioxide used for precipitation of crude phenol from sodium phenoxide solution is normally produced in a lime kiln. This provides kiln gas with a carbon dioxide content of 30 – 35 % and the calcium oxide required for the caustification of the sodium carbonate solution. The sodium carbonate solution flowing from the precipitation column still contains 3 – 4 % physically dissolved phenols which can be isolated before caustification by ether extraction. Caustification takes place in stirred vessels at ca. 100 °C; the carbonate solution can be treated without cooling. The suspension of precipitated calcium carbonate in sodium hydroxide solution flowing out of the caustification plant is preseparated in a sedimentation vessel. For isolation of sodium hydroxide solution the settling calcium carbonate is filtered on a vacuum rotary filter and washed. The regenerated clear sodium hydroxide solution is recycled to the carbolic oil extraction step. The filter cake of the vacuum filter consists of precipitated microcrystalline calcium carbonate with a water content of 40 – 50 %. It can be used as neutralizing agent, for instance, in the inorganic chemical industry, for flue-gas desulfurization, or as a concrete additive.

The crude phenol from the saturation step (generally 3 parts) and the ether extract from sodium hydroxide solution refining (1 part) is subjected to extraction with water (from dehydration of the crude phenol) in a further column to demineralize the crude phenol. The crude phenol is liberated from ether and water in a continuous distillation plant prior to distillative separation of individual phenols. This process is carried out on a large scale in a plant having a capacity of ca. 30 000 t/a [3], [46], [47].

The isolation of crude phenol by this process requires ca. 550 kg of limestone and 55 kg of coke per tonne water-free crude phenol.

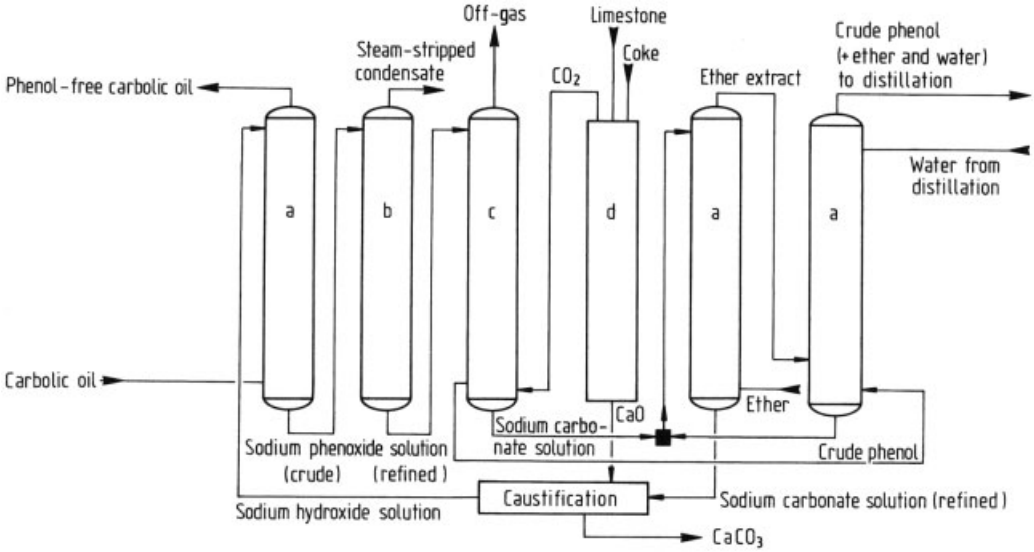

Figure 9. Isolation of crude phenol by sodium hydroxide solution extraction
a) Extraction; b) Steam stripping; c) Saturation; d) Lime kiln

The Lurgi *Phenoraffin process* is the only industrially important process in which crude phenol is extracted by selective solvents.

Figure 10 shows the processing flow sheet for crude phenol isolation by the Phenoraffin process. As selective solvent an aqueous sodium phenoxide solution is used with excellent dissolving power for phenols. The phenoxide solution can be oversaturated to 350 % of the amount of stoichiometrically bound phenol. The phenols are extracted from the sodium phenoxide solution with diisopropyl ether. The plant consists of numerous extraction and distillation columns. The carbolic oil is extracted in two steps in countercurrent flow with the phenoxide solution. Between the two steps the oil is de-based with dilute sulfuric acid. Hydrocarbons and pyridine bases are extracted from the supersaturated solution of phenols in sodium phenoxide solution with, for example, toluene. After de-basing with dilute sulfuric acid and extraction with water the toluene is regenerated by distillation. The distillation bottoms, which contain phenols and hydrocarbons, are returned to the carbolic oil. The phenoxide solution after extraction with toluene and still charged with phenols is stripped with steam to remove the remaining toluene. It is then blown with air to destroy colored components and then extracted with diisopropyl ether. After the ether extraction the ether dissolved in the phenoxide solution is removed by distillation. The regenerated phenoxide solution is recirculated to the oil extraction step. The ether phase, which contains phenoxide solution, is de-based with sulfuric acid, washed with water and separated by distillation into diisopropyl ether and crude phenol. The ether is recirculated. The crude phenol is purified by distillation.

Energy and chemical requirements for production of one tonne of crude phenol by the Phenoraffin process are 5 t steam, 165 m^3 cooling water, 60 kWh electrical energy, 20 kg diisopropyl ether, and 10 – 200 kg sulfuric acid (depending on the amount of pyridine bases).

The crude phenol extracted from coal-tar carbolic oils by sodium hydroxide solution has the following composition: 15 % water, 30 % phenol, 12 % *o*-cresol, 18 % *m*-cresol, 12 % *p*-cresol, 8 % dimethylphenols (xylenols), 5 % trimethylphenols and higher boiling phenols. The quantitative composition of the crude phenol produced in coal-tar refineries varies widely as varying amounts of sodium phenoxide solution or crude phenol from coking plant effluents are used. These mainly contain phenol and cresols but insignificant amounts of higher phenol homologues. Normally crude phenol from sodium hydroxide extraction has an ether content of ca.

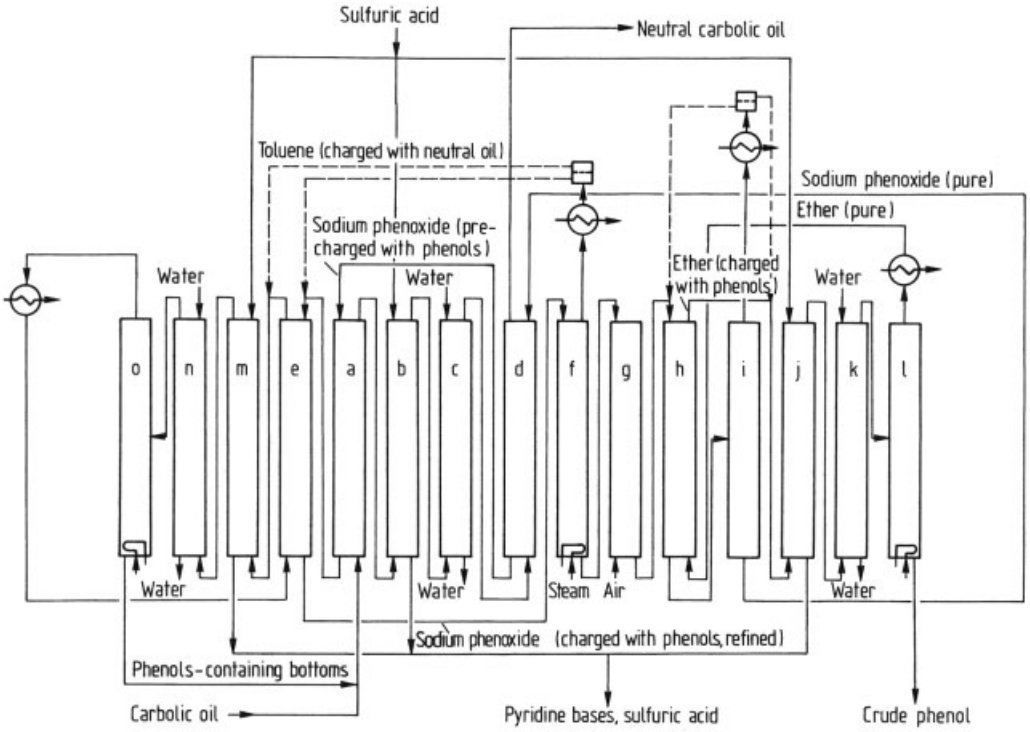

Figure 10. Crude phenol isolation (Phenoraffin process)
a) Oil extraction with sodium phenoxide (1st step); b) Oil de-basing by extraction with sulfuric acid; c) Oil extraction with water (neutralization); d) Oil extraction; e) Removal of phenols from the sodium phenoxide solution; f) Steam stripping; g) Air blowing; h) Ether extraction of the refined phenoxide solution of phenols; i) Ether stripping; j) Ether de-basing by extraction with sulfuric acid, k) Ether extraction with water (neutralization); l) Distillation column for separating ether and crude phenol; m) Toluene de-basing by extraction with sulfuric acid; n) Toluene extraction with water (neutralization); o) Toluene distillation

2 % which is decreased by water washing before distillation to ca. 0.3 %.

Figure 11 shows the processing flow sheet of a continuous crude phenol distillation. The first stage is dehydration and ether removal by distillation of the crude phenol, normally under atmospheric pressure. From the dewatered, ether-free crude phenol, a residue containing higher boiling phenol homologues and unvaporizable phenolic pitch is separated in a vacuum stripping column. The stripped vapors are rectified, for instance, in four interconnected tray columns operating under a graduated vacuum of 10 – 140 mbar and equipped with steam-heated reboilers. The tray columns are preferably made of stainless steel. For example, in the first, 80-tray column, technically pure phenol as top product is obtained with a reflux ratio of, e.g., 12: 1. From the bottom stream of this phenol column, a technically pure *o*- cresol is obtained as top product in the second, 65-tray column with a reflux ratio of, e.g., 25: 1. From the bottom stream of this *o*-cresol column, a *m/p*-cresol mixture is obtained as top product in the third 45-tray column with a reflux ratio of, e.g., 5 : 1. The bottom product of this *m/p*- cresol column together with the bottom product of the stripping column is separated in the fourth 45-tray xylenol column into a dimethylphenol fraction and phenolic pitch. The phenolic pitch contains polymerized and oxidized phenols, which can be depolymerized to phenols by distillation with tar oils.

3.4.5. Processing of Light Oil

In coal-tar primary distillation, light oil is obtained azeotropically during distillative dehydration. It accounts for 0.5 – 3 % of the crude tar and has a wide boiling range of ca. 70 – 200 °C. In addition to the main components, benzene

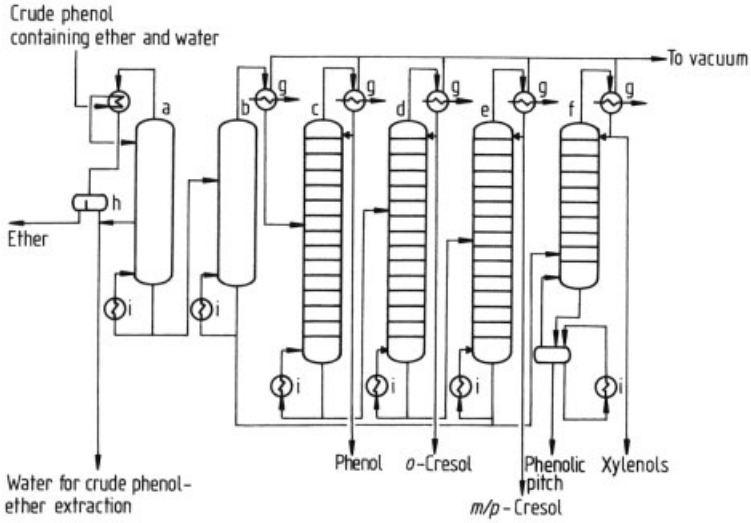

Figure 11. Distillation of crude phenol
a) Column for the separation of ether and water; b) Stripping column; c) Phenol column; d) *o*-Cresol column; e) *m*/*p*-Cresol column; f) Xylenol column; g) Condenser (low-pressure steam production); h) Separator; i) Reboiler

and its homologues, it contains 10 – 40 % unsaturated, polymerizable aromatic hydrocarbons, 3 – 10 % phenols and cresols, 2 – 7 % pyridine bases, and varying amounts of naphthalene. Coal-tar light oil, preferably together with carbolic oil I, is processed by distillation, extraction, and polymerization to obtain benzene and its homologues, pyridine bases, phenols, and indene – coumarone resins.

Pyridine bases are obtained from the dephenolized light oil and carbolic oil I, to which the strip oil from the sodium phenoxide solution refining can also be added, by discontinuous or continuous extraction with dilute sulfuric acid and precipitation with ammonia or sodium hydroxide solution.

Figure 12 shows the flow sheet for the recovery of pyridine bases. Extraction is carried out with 25 – 35 % sulfuric acid. After separation of acidic resins that may form the dissolved hydrocarbons are removed from the acidic solution of pyridine-base sulfates by steam stripping or solvent extraction with benzene. From this refined solution the crude pyridine is precipitated with gaseous or aqueous ammonia. The aqueous phase is separated and concentrated to yield ammonium sulfate.

The composition of crude pyridine bases from coal tar varies widely, depending mainly on the boiling range of the starting oil. The main components of a typical water-free tar base mixture from coal-tar light oil are listed below. The crude pyridine bases contain mainly pyridine and its methyl homologues as well as small amounts of aniline and its methyl homologues.

Pyridine	12 %
α-Picoline	10 %
2,6-Lutidine	6 %
β-Picoline	6 %
γ-Picoline	8 %
2-Ethylpyridine	2 %
2,5-Lutidine	4 %
2,4-Lutidine	11 %
2,3-Lutidine	2 %
3,5-Lutidine	$<$ 1 %
2,4,6- and 2,3,6-Collidine	9 %
Aniline	12 %
o-Toluidine	4 %
m/*p*-Toluidine	8 %
Higher boiling bases	6 %

The extracted crude pyridine is distilled (Fig. 12). The crude pyridine still contains water after precipitation with ammonia. Dehydration is preferably carried out by azeotropic distillation with benzene in a continuous two- column system. In the first column the water- containing crude pyridine is separated into a water-free bottom product and an azeotropic pyridine – water top product. After condensation and cooling,

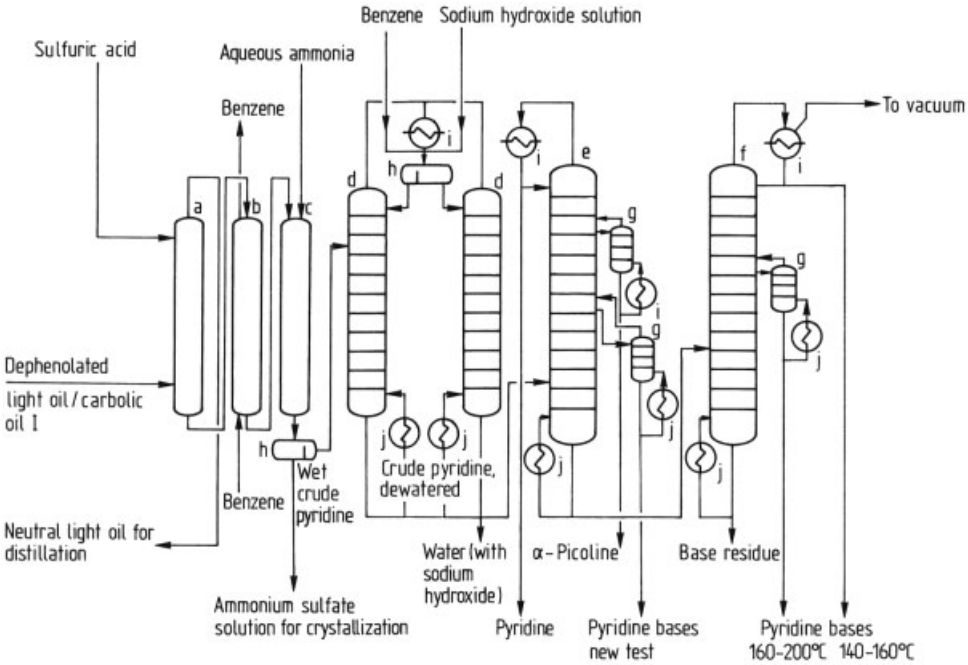

Figure 12. Recovery of pyridine bases
a) De-basing; b) Refining of the base sulfates solution; c) Base precipitation; d) Dewatering column system; e) Pyridine main column; f) Vacuum main column; g) Side column; h) Separator; i) Condenser; j) Reboiler

sodium hydroxide solution or benzene or both are added to the top product in a separator, thus obtaining a predominantly pyridine-containing and water-containing phases. The aqueous phase is pumped to the upper tray of the second column where it is separated into a pyridine-free, aqueous sodium hydroxide solution as bottom product and an azeotropic pyridine – water mixture as top product. The top products of the first and second column are both condensed, cooled, and separated into the above-mentioned phases. The phase that contains the pyridine bases from the separation vessel is fed as reflux to the first column. The dehydrated crude pyridine is separated into pyridine and various homologue fractions by, for example, continuous rectification in a column at atmospheric pressure and a vacuum column with accompanying side columns. In the main column at atmospheric pressure, technically pure pyridine is taken off as top product. An α-picoline fraction and a base fraction ("pyridine bases new test"), with a boiling range of 130 – 150 °C, consisting of picolines and lutidines, are obtained in a side column. In the vacuum main column the bottom product is split into a pyridine base fraction, boiling range 160 – 200 °C, as top product and the base residue. A fraction containing collidine and aniline bases, boiling range 160 – 200 °C, is obtained in a side column. The reboilers are generally heated by pipe stills in heat exchange. By using additional side columns or by redistillation, further splitting of the base fractions is possible. The following main fractions are obtained:

1) Technically pure pyridine, *bp* 115.3 °C
2) Technically pure α-picoline, *bp* 129.4 °C
3) A β/γ-picoline/2,6-lutidine-fraction, boiling range 144.0 – 145.4 °C
4) A 2-ethylpyridine fraction, boiling range 146 – 157 °C
5) A lutidine fraction with 2,4- and 2,5-lutidine as main components, boiling range 157 – 159 °C
6) An intermediate fraction with 2,3-lutidine, 3- and 4-ethylpyridine, and 2-methyl-6-ethylpyridine as main components, boiling range 159 – 170 °C
7) A collidine fraction with 2,4,6- and 2,3,6-collidine as main components and small

amounts of 3,5-lutidine, boiling range of 170 – 173 °C
8) An aniline fraction, *bp* ca. 184 °C
9) A toluidine fraction with the three toluidine isomers, boiling range 195 – 205 °C
10) A xylidine fraction with the six isomeric dimethylanilines, boiling range 210 – 225 °C

The fractions 1 – 7 contain pyridine and its homologues, and the fractions 8 – 10 contain aniline and its homologues.

Whereas pyridine and α-picoline are obtained technically pure by distillation alone, isolation of the other nitrogen bases from the distillate fractions is difficult. The majority of the pyridine bases, however, are not isolated in pure form, but are used as distillate fractions or as blends.

The commonly used technical pyridine-base blends, which are mainly employed as solvents, are classified according to boiling range. The "pyridine bases new test" are used to denaturize ethanol. The higher boiling base fractions are important anti-corrosion agents. The isomeric mixture of toluidines serves as absorbent for SO_2 and as a vulcanization accelerator. The pyridine homologue fractions can be demethylized catalytically in the presence of hydrogen to give pyridine and lower pyridine homologues.

After the extraction of phenols and pyridine bases, the "neutral oil" obtained from light oil and carbolic oil I is separated by distillation into different fractions. The fraction with a boiling range of 160 – 185 °C is of special interest as it contains up to 60 % polymerizable unsaturated hydrocarbons, with indene as main component and slight amounts of coumarone, dicyclopentadiene, and methylstyrenes. These compounds are polymerized cationically to indene – coumarone resins with Friedel – Crafts catalysts such as boron trifluoride and its complexes or aluminum chloride (→ Resins, Synthetic, Chap. 2.6.).

The nonpolymerizable benzene homologues which are distilled from the resins under vacuum after polymerization are combined with the remaining light-oil fractions boiling up to 185 °C, which are free of phenol and organic bases. They are processed as benzol pre-product together with the coke-oven crude benzol after refining with sulfuric acid or catalytical hydrogenation pressure refining to yield technically pure benzene, toluene, and benzene homologues (→ Benzene, Chap. 5.1.). After extraction of phenols and bases, the light-oil fractions that boil above 185 °C are used as solvents for paints, varnishes, and binders.

4. Processing of Low-Temperature Coal Tars

Only in a few locations (e.g., United Kingdom, India, China, and South Africa) are coal tars from low-temperature carbonization and gasification of coal processed by distillative and extractive refining. The aim of the recovery process is mainly the isolation of the phenols (see Table 1 for composition), rather than the isolation of aromatic hydrocarbons.

Figure 13 shows the flow chart for low-temperature tar, light condensate oil, and aqueous ammonia liquor from low-temperature carbonization of highly volatile coal to produce smokeless fuels at ca. 640 °C using the Coalite process [48–51]. In this process one tonne of coal yields 70 – 75 kg of crude coal oil and about 160 L of ammonia liquor. From the crude gas, tar and light condensate oil are obtained by two-stage condensation. Analoguous to coke-oven coal tar, they are dewatered and separated into three to five fractions by continuous primary or multi-stage distillation at atmospheric pressure and under vacuum:

1) Light oil, boiling range ca. 90 – 160 °C, about 2 % of crude tar, containing roughly 30 % aromatic hydrocarbons, 35 % alkenes, 20 % alkanes + naphthenes, 10 – 15 % phenols, and 2 % tar bases.
2) Middle oil, boiling range 200 – 300 °C, ca. 40 % of crude tar, containing roughly 45 % phenols, 25 % aromatic hydrocarbons, 10 – 15 % alkenes, and 10 – 15 % alkanes + naphthenes.
3) Upper distillate, boiling range 300 – 400 °C, ca. 30 % of crude tar, containing ca. 65 % aromatic hydrocarbons and alkanes. If required they can be separated by further vacuum distillation into a heavy oil and a wax oil with a significant content of solid paraffin hydrocarbons.
4) Low-temperature coal tar pitch as a distillation residue, ca. 25 % of crude tar.

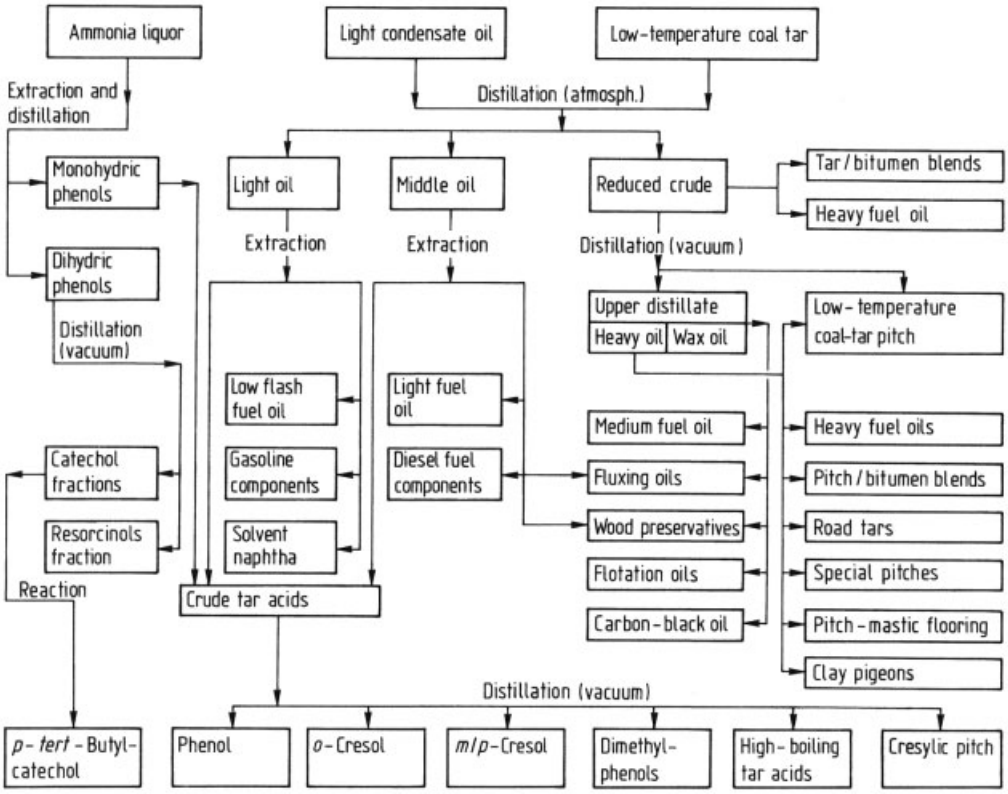

Figure 13. Processing of liquid products from low-temperature carbonization (Coalite process)

Alternatively, the residue from the base of the atmospheric fractionating column, after removal of light and middle oil (reduced crude) can be utilized, for example, as a heavy fuel oil or for making tar – bitumen blends for road construction.

Low-temperature coal-tar pitch is composed primarily of a complex mixture of hydrocarbons. It can be produced with softening points within the range of 40 – 180 °C (ring & ball). With a softening point of 55 – 60 °C it is used in production of road binders, sometimes in combination with bitumen. It is also used as an ingredient in the material prepared for coating submarine telephone cables and in the preparation of pitch – fiber pipes. The pitch that softens between 73 and 77 °C is also used for the production of pitch – fiber pipes and as a binder in briquette production. Special pitches with softening points of, for example, 82 – 88 °C and 100 – 110 °C are obtained by air blowing at elevated temperatures. These products (Florastic 85 and Florastic 100) are used for the production of colored or black pitch – mastic flooring for domestic and business premises and for the manufacture of clay pigeons. These pitch – mastic floors are resistant to oils and fats. The significantly lower content of benzo[*a*]pyrene in low-temperature pitch compared with pitch obtained from coke-oven coal tar is highly advantageous for these applications.

Heavy oil or upper distillate is processed to produce creosote oils with various boiling ranges. These are used as wood preservatives for railway sleepers, telegraph poles, electricity-transmission poles, and timber; fluxing oils for cutting bitumens and road tars; as carbon-black oil; and as medium industrial fuel oil.

Similar to the processing of coke-oven coal tar phenols are extracted from the *middle oil* continuously with sodium hydroxide solution to form a solution of water-soluble sodium salts

known as cresylate. A neutral oil fraction then separates. Neutral oil entrained in the cresylate is removed by extraction with a lighter oil, followed by steam treatment to remove the last traces of oil. The phenolic cresylate is converted to a crude mixture of free phenols (crude tar acids) by acidification with flue gases containing carbon dioxide. Further refining of the neutral middle oil affords low-viscosity wood preservatives, fluxing oils for bitumens and tar blends, as well as components for diesel fuels and light industrial fuel oils.

Crude tar acids are also separated from *light oil* by extraction with sodium hydroxide solution. Distillation and refining of the neutral oil gives low-flash fuel oils, gasoline components, and solvent naphtha. The oil is also used as a light industrial fuel oil.

The aqueous *ammonia liquor* contains ca. 1.5 % dissolved phenols, about two-thirds of which are monohydric and one-third dihydric. These phenols are extracted from the aqueous phase with, for example, isobutyl acetate or diisopropyl ether (Lurgi Phenosolvan process, → Gas Production). After distillation of the extractant, the crude phenol mixture is separated by continuous distillation into a lower boiling monohydric phenols fraction and a dihydric phenols fraction. The dihydric phenols fraction is separated by further distillation into catechol and resorcinols fractions. Further processing leads to catechol, 4-methylcatechol, a resorcinols fraction, and 2-methylresorcinol, and as a further downstream product, 4-*tert*-butylcatechol.

The monohydric phenol fraction isolated from ammonia liquor and the crude phenol obtained by extraction of light oil and middle oil is dehydrated by distillation and then separated by continuous distillation into the following fractions:

1) Technically pure phenol (8 – 10 % of the crude phenol)
2) Technically pure crude *o*- cresol (ca. 2 % of the crude phenol)
3) An *m*-/*p*- cresol fraction containing 50 % *m*-cresol (ca. 15 – 29 % of the crude phenol)
4) Various dimethylphenol fractions, for example a fraction with 90 % 2,4- and 2,5-dimethylphenol and one with 50 % *m*-/*p*-ethylphenol and ca. 50 % 3,5-dimethylphenol. These fractions make up ca. 30 – 35 % of the crude phenol.
5) Refined high-boiling tar acids (boiling range ca. 230 – 305 °C), approx. 20 % of the crude phenol.
6) Cresylic pitch (R & B softening point ca. 115 °C), approx. 18 % of the crude phenol.

The *high boiling tar acids* contain as main components C_3, C_4, and C_5 alkylphenols as well as 4- and 5-indanols. They are converted to black and white heavy duty disinfectants, soil sterilizers, horticultural fluids, sheep dips, foundry molding resins, agents for ore and coal flotation, engine degreasers and metal cleaners, specialized dye-markers, resins for food and beverage can linings, and germicides for impregnating toilet tissues.

The cresylic pitch is used as pulverized fuel and as a binding agent.

Because of the corrosive properties of dihydric phenols contained in low-temperature tar crude phenol, the distillation plants must be manufactured of high-quality stainless steel.

In the United Kingdom phenols and phenolic fractions from low-temperature tar crude phenols are extracted by the Coalite process. These phenols and phenolic fractions are then processed by chlorination and phosphorylation as well as by reaction with formaldehyde to give phenolic resins.

As an alternative to the chemical processing, crude tar from low-temperature carbonization of coal can be processed directly by catalytic hydrogenation to produce gasoline and diesel fuels [52].

Low-temperature coal tars obtained as byproducts of coal pressure gasification (e.g., by the Lurgi fixed-bed process) are processed similarly to the coal oil from low-temperature carbonization of coal. Attention has to be paid to the increased concentration of aromatic hydrocarbons which is due to higher generation temperatures (see Table 1). In contrast to the crude phenol resulting from lowtemperature carbonization of coal, the crude phenol, from water effluents and tar oils, contains almost no dihydric phenols but increased amounts of phenol, *o*- cresol, and *m*-/*p*- cresol.

Since the synthesis gas formed by coal gasification is the desired product, the crude tar

formed is preferably recycled to the gasifier, thus giving an increased gas yield.

5. Processing of Other Tars and Tarlike Raw Materials

Lignite tars and crude phenols are byproducts of lignite carbonization and gasification. Large-scale plants for lignite coking and chemical processing of lignite tars in Eastern Germany [53], were shut down for economic reasons after German reunification in 1991/1992. Lignite tars are obtained in American, Czech, Indian, and German large-scale lignite gasification plants, especially in fixed-bed gasification by the Lurgi process. These tars are either recirculated to the gasifier to increase gas production (similar to coal-gasification tars) or are utilized as heavy fuel oil for steam generation. Generally only the crude phenol extracted from wastewater is chemically processed, analogous to the refining of coal-tar crude phenol to phenol, *o*-cresol and *m*-/*p*-cresol, xylenols, and higher boiling phenols (see Section 3.4.4 and Chap. 4). The Dakota Gasification Co. in Bismarck/North Dakota runs such a plant with a phenol capacity of ca. 15 000 t/a.

Peat tars are obtained as byproducts with a yield up to 10 % during carbonization of peat to peat coke, which is used as low-ash metallurgical carbon and for production of activated carbons (→ Peat, Chap. 3.3.2.). Peat tars are used only as low-sulfur industrial fuel oils. Technically the distillation of peat tar to obtain oil is possible. This distillation product can be extracted to yield crude phenol, containing as its main components phenol, the three isomeric cresols, guaiacol (2-methoxyphenol), 2,4-, 2,3-, and 3,5-dimethylphenol, *p*-ethylphenol, and higher phenol homologues, as well as small amounts of catechol and its homologues. Peat-tar oils can be used to formulate wood preservatives.

Wood tars, methanol, and acetic acid are the liquid byproducts of wood carbonization to give activated carbons and wood charcoal for iron smelting (→ Charcoal). In these plants wood tar normally is used directly as a low-sulfur fuel oil. Wood tar can be separated into wood-tar pitch and various tar-oil fractions. The tar oils are suited to the formulation of wood preservatives and flotation oils. Extraction with sodium hydroxide solution gives wood-tar creosote, which contains phenol, creosols, guaiacol, and other phenol ethers as main components. Due to its disinfectant properties purified beechwood creosote, boiling range 200 – 220 °C, with cresols and guaiacol as main components, serves as a formulating agent for pharmaceutical preparations (e.g., cough mixtures). Furthermore, it is used as auxiliary solution in mercury maximum/minimum thermometers. Wood-tar pitch can be used as putty for the polishing of optical glasses. Crude wood tars are suitable for jute impregnation, serve as binders for briquettes and active carbon, as game deterrents, and as disinfectant formulations in veterinary medicine.

Oil-shale tars, also called shale oils, are the main liquid products of oil-shale retorting (→ Oil Shale, Chap. 7.). Depending on composition (see Table 1), oil-shale tars can be refined by pressure hydrogenation and subsequent distillation to fuel oils, gasoline and diesel fuels, or can be carbonized [54]. Estonian oil-shale tars, rich in crude phenol (cucersit oil, obtained by the Kivitter process), are separated by distillation into various oil fractions, boiling up to 360 °C. They are used for the production of wood preservatives, road binders, and plasticizers [55]. The pitch residue can be carbonized to electrode coke. Phenols are extracted from water effluents and lower boiling distillate fractions and processed to phenolic resins, injection agents for soil compaction, tanning agents, and alkyl resorcins.

Pyrolysis residual oils are tarlike byproducts of steam cracking of naphtha and gas oil to produce ethylene/propene (→ Ethylene, Chap. 5.1.; → Resins, Synthetic, Chap. 2.2.). Pyrolysis residual oils consist almost entirely of aromatic hydrocarbons (see Table 1). Due to the relatively high content of unsaturated, polymerizable aromatics they are less stable than coal tars. The main components are naphthalene and its homologues, indene and its homologues, and thermally polymerized pyrolysis residual resin (pyrolysis pitch), obtained as a dark distillation residue. By using modified coal-tar refining processes (distillation, thermal and catalytic polymerization, crystallization) the basic chemicals naphthalene, methylnaphthalenes, aromatic hydrocarbon resins and aromatic oils are obtained from pyrolysis residual oils on a large scale [56],

[57]. Pyrolysis residual oils are also used as raw materials for carbon black.

Pyrolysis residual oils from scrap plastics, tyres, etc. are obtained, similarly to petrochemical pyrolysis residual oils from ethylene/propene plants, as tarlike materials. They can be processed to chemical raw materials. If the pyrolysis is carried out at ca. 700 °C in fluidized-bed or rotary-drum reactors, up to 50 % pyrolysis oils, rich in aromatics, are obtained (→ Plastics, Recycling, Chap. 4.3.). These pyrolysis oils have a composition similar to pyrolysis residual oils from ethylene/propene plants (see Table 1). As main components they contain benzene, naphthalene and its homologues, styrene, indene and its homologues, and dark pyrolysis residual resins (pyrolysis pitch), which is obtained as distillation residue [58–60]. If modified processes of coal-tar refining are applied, pyrolysis oils yield benzene homologues, naphthalene and methylnaphthalenes as well as technical aromatic oils.

6. Uses of Tar Products and their Economic Importance

About 30×10^6 t/a of all kind of tars are recovered, of which ca. 15×10^6 t/a are processed to basic chemicals and other products. The main raw materials are high-temperature coal tars from coking plants. Low-temperature coal tars from coal carbonization and gasification; lignite tars, oil-shale tars, and pyrolysis residual oils are of lesser importance. Peat tars, wood tars, and pyrolysis oils from scrap are not currently processed on a large scale.

Tars are the basic raw material for:

1) Industrial carbon products: carbon electrodes for aluminum production, graphite for production of electrodes and high-temperature materials and carbon black as a rubber filler and black pigment (ca. 4×10^6 t/a)
2) Technically pure condensed aromatic hydrocarbons such as naphthalene and its homologues, anthracene, phenanthrene, acenaphthene and pyrene for synthesis of polymers, plasticizers, and dyes (ca. 1×10^6 t/a)
3) Technically pure aromatic heterocyclics such as carbazole, quinoline, pyridine, and indole for syntheses of dyes, pesticides, and pharmaceuticals (several thousand tonnes per annum)
4) Phenols for synthesis of thermosets, pesticides, disinfectants, and antioxidants (several hundred thousand tonnes per annum)
5) Thermoplastic indene – coumarone resins for the formulation of paints, varnishes, adhesives, and rubber products (ca. 100×10^3 t/a)
6) Aromatic oils as wood preservatives, solvents, and gas-washing oils ($>2\times10^6$ t/a)

Whereas aliphatic hydrocarbons and mononuclear aromatics presently are mainly produced from petrochemicals, polynuclear aromatics from coal tar cover a high percentage of the total requirements. In the case of naphthalene the fraction arising from coke-oven coal tar is ca. 85 % worldwide. Naphthalene homologues, the polynuclear aromatics anthracene, phenanthrene, acenaphthene and pyrene, and the heterocyclics carbazole and quinoline are obtained exclusively from coal tar. The portion of phenols derived from tar is only 3 % worldwide, whereas in case of phenol homologues it is ca. 40 %. Approximately 20 % of world aromatic hydrocarbon resin production is accounted for by indene – coumarone resins based on coal tar. In the case of binders for carbon and graphite electrodes, coal-tar electrode pitches account for over 90 %. Carbon-black oils from coal tar are a basic raw material for more than a quarter of world carbon-black production ($>2\times10^6$ t/a), and 75 % of oil-based wood preservatives for pressure-impregnation of timber ($>10^6$ t/a) are based on coal tar.

As discussed above, tars and pitches of different types are used also as industrial fuel oils, motor fuels, binders and coating materials as well as for co-processing together with coal for gasification and coking.

7. Toxicology and Ecotoxicology

7.1. Toxicology

The acute toxicity of tars and pitches is low, but high doses of polycyclic aromatic hydrocarbons and structurally related heterocyclic systems are known to result in necrosis of the adrenal gland.

Tars or their fractions may cause irritation of the skin and eyes by direct or indirect exposure. The extent of skin irritation varies, depending on the origin of the tar, the exposure time, and involvement of light. High-temperature coal tar has the strongest, and wood tar the weakest impact on skin. Contact with unprotected human skin may lead to dermatitis, inflammation of the sebaceous glands (folliculitis), and acne. Several years of exposure can cause brownish, speckled pigmentation (melanosis) and tar or pitch warts [61, pp. 130 – 132].

Certain tar constituents are activated by UV light, resulting in intensified irritant skin reactions (phototoxicity) or allergic contact dermatitis (photosensitization); this may occur also on contact with minor quantities of volatiles from tar or tar fractions.

Pharmaceutical formulations of coal tars, especially in combination with exposure to UV light, have been successful in the treatment of psoriasis [62], though after excessive usage of tar ointments, melanomas or squamous- cell carcinomas developed in isolated cases [61, p. 136]. Furthermore, such combined treatment caused marked inhibition of DNA synthesis in vivo in normal and proliferative skin of hairless mice [61, p. 128].

In general, tar constituents seem to be readily available by all exposure routes and distributed throughout the body, but are also rapidly excreted [63–66]. On application of crude coal tar (2 % solution in petrolatum) to the skin of humans for 8-h periods on two consecutive days evidence was obtained for absorption of polycyclic aromatic hydrocarbons (PAH) [65].

The potency of tumor formation can be attributed to this fraction of the mixture; PAHs are byproducts of pyrolysis of all organic materials and are generated to a major extent above 500 °C. Thus high-temperature coal tar has the highest content of highly condensed aromatic ring systems.

The experimental generation of skin carcinomas by the topical application of tars, tar fractions such as creosote, anthracene oil, and pyrolytic oil [67] of naphtha, and tar pitch is well documented [61], [68]. Case reports and epidemiological studies have provided sufficient experience about the occurrence of skin, scrotum, bladder, and other systemic cancers in exposed workers [61, pp. 121 – 127], [68, pp. 140 – 155].

Evidence has accumulated to confirm that occupational inhalation of tar or tar pitch volatiles leads to increased lung and bladder cancer mortality that is strongly associated with duration and intensity of exposure [61, pp. 134 – 136], [69, pp. 174 – 176]. There is suggestive experimental and epidemiological support for PAH causing cancer by the oral route [61], [69].

Inhalation studies with rats exposed to a condensation aerosol from tar pitch alone and in combination with carbon black [70], [71] corroborate the idea that the carcinogenic impact of PAH on the lungs is enhanced by adsorption onto respirable fine particles: After 43 weeks of exposure, the lung tumor rate was 4 % with 1.1 mg/m^3 aerosol alone (= 20 µg/m^3 benzo[*a*]pyrene), which was less than in the control with carbon black alone (8 – 18 %). With 2.6 mg/m^3 of aerosol (= 50 µg/m^3 benzo[*a*]pyrene), the tumor rate amounted to 33 %, but this concentration in combination with 2 or 6 mg/m^3 of carbon black led to 89 and 72 %, respectively. The majority of lung tumors were squamous- cell carcinomas. This potentiating effect can be explained by a PAH depot from which carcinogenic material is sustainably released, thus resulting in long-term exposure to the target cells of the lung.

Some 50 coal-tar compounds have been studied for their cancerogenic effect under experimental conditions, mainly on the skin of rats and mice. About 30 proved to be noncarcinogenic, and ca. 20 were positive to varying extent; for status reports see [72], [73]. The strongest carcinogens are found among systems with five or more condensed benzene rings. They are not tumorigenic by themselves, but must be metabolically converted into highly reactive alkylating dihydrodiol epoxides, the most potent of which have a so- called bay region.

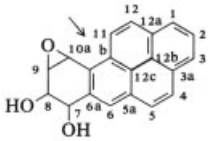

Bay-region dihydrodiol epoxide of benzo[*a*]pyrene

Benzo[*a*]pyrene, the model compound, is one of the strongest carcinogens among the PAHs,

and the carcinogenic potency of tar fractions is frequently related to its concentration.

Examples of carcinogenic high-molecular *N*- and *S*-heterocycles are benzologues of acridine, carbazole, and dibenzo[*bd*]thiophene. Apart from PAHs, 2-naphthylamine – found in tar only in very low concentrations – is well known to specifically cause bladder cancer. This was identified primarily among workers manufacturing aniline dyes where appreciable exposure to aromatic amines occurred under former working conditions [74].

7.2. Classification and Legislation

According to IARC, there is sufficient evidence for carcinogenicity of coal tars and coal tar pitches to humans (class 1) [68, pp. 174 – 176]. In the United States and Germany, the classification is A1 (human carcinogen). In Germany, the occupational limit value for benzo[*a*]pyrene in volatiles from pitch and coke processing is set at 5 µg/m^3, from other sources at 2 µg/m^3 [75]. In the United States the amount of coal-tar/pitch volatiles is determined as benzene-soluble matter recovered from sampled dust. The TLV or MAC (maximal accepted concentration) is 0.2 mg/m^3 [76]. In the European Union, complex coal-derived materials for industrial purposes must be labelled as carcinogens unless a certain limit concentration of benzo[*a*]pyrene (currently 50 mg/kg) as indicator component for PAHs is met [77]. A comparative carcinogenicity study with two high-temperature coal tar oils (benzo[*a*]pyrene < 30 and < 300 mg/kg) is in progress [78].

7.3. Ecotoxicology

The aquatic toxicity of tar and tar-borne materials depends primarily on the physical and toxicological properties and availability of their aromatic constituents from the matrix of the mixture. However, the outcome of test procedures is very much determined by the method of preparation of test solutions and the reference basis for the computation of the effective or toxic concentrations. Therefore, the evaluation for aquatic ecotoxicological properties of mixtures which consist of sparingly soluble substances is problematic.

In particular, organic solvents or detergent-like agents used as cosolubilizers, dispersants, or emulsifiers tend to bias the measurement (see below) [79]. The choice and standardization of adequate testing methods continue to be subject of controversy and debate [80].

The highest toxicity (LC_{50} values mostly < 1 mg/L) [81], [79] was obtained with acetone solutions of high-temperature coal tar creosote. In a reliable comparative study on fish and daphnia [79], the aqueous creosote extracts yielded 96h and 48h LC_{50} values between 2 and 3.5 mg/L (referring to nominal concentration of creosote); on the other hand, about 0.7 mg/L was found when the organic solvent was used. This can be explained by the finding that the composition of the acetone solution was almost identical to that of the original material, while in the aqueous extract – reflecting more realistic conditions – less toxic components with higher water solubility were enriched.

Aqueous coal tar extracts were somewhat less toxic to fish and daphnia, with LD_{50} values ranging from about 7 to 12 mg/L (nominal). This may reflect its lower content of water soluble fractions and lower availability of the toxic constituents compared with creosote. This seems to be in agreement with findings on four tar oils in the bacteria bioluminescence assay with effective 50 % doses from ca. 0.8 to 1.5 mg/L, since these data are comparable to the range of creosote toxicity mentioned above [82].

For tar pitch, the lowest LC_{50} (96 h) for fish is reported to be ca. 100 mg/L and for daphnia > 100 mg/L (nominal) [79].

8. References

General References

1. H. G. Franck, G. Collin: *Steinkohlenteer,* Springer Verlag, Berlin 1968.
2. G. Collin, E. Wolfrum: "Chemierohstoffe aus flüssigen Nebenprodukten der Verkokung, Verschwelung und Vergasung," in Winnacker-Küchler: *Chemische Technologie, Organische Technologie I,* 4th ed., Hanser Verlag, München 1981.
3. G. Collin, *Erdöl, Kohle Erdgas Petrochem.* **29** (1976) 159 – 165; **38** (1985) 489 – 496.

4. G. Collin, Proc. World Conf. Future Sources Org. Raw Mater. Chemrawn I, Toronto, July 10 – 13, 1978, Pergamon Press, London 1979, pp. 283 – 297.

Specific References

5. M. Zander, G. Collin, *Fuel* **72** (1993) 1281 – 1285.
6. H. G. Franck, *Brennst. Chem.* **36** (1955) 12 – 20.
7. G. P. Blümer, R. Thoms, M. Zander, *Erdöl, Kohle Erdgas Petrochem.* **31** (1978) 197.
8. W. Boenigk, M. W. Haenel, M. Zander, *Fuel* **69** (1990) 1226 – 1232.
9. H. D. Sauerland, J. W. Stadelhofer, R. Thoms, M. Zander, *Erdöl, Kohle Erdgas Petrochem.* **30** (1977) 215 –218.
10. P. Fischer, J. Stadelhofer, M. Zander, *Fuel* **57** (1978) 345 – 352.
11. M. Zander, *Erdöl, Erdgas, Kohle* **105** (1989) 373 – 377.
12. M. Zander, *ACS Fuel Chem. Div. Prep.* **34** (1989) 1218 – 1225.
13. J. D. Brooks, G. H. Taylor, *Nature (London)* **206** (1965) 697 – 699.
14. M. Zander, *Erdöl, Kohle Erdgas Petrochem.* **38** (1985) 496 – 503.
15. S. Takeshita, *CEER Chem. Econ. Eng. Rev.* **9** (1977) nos. 2/3, 25 – 30.
16. Rütgerswerke, DE 976 913, 1954 (H. Krieger, E. Schweym, J. Geller).
17. A. M. Odok, W. K. Fischer, *Light Met. (Warrendale, Pa.)* 1978, 269 – 286.
18. A. Alscher, W. Gemmeke, F. Alsmeier, W. Boenigk, *Light Met. (Warrendale, Pa.)* 1987, 483 – 489.
19. F. Keller, W. K. Fischer, *Light Met. (Warrendale, Pa.)* 1982, 729 – 740.
20. W. K. Fischer, F. Keller, R. Perruchoud, *Light Met. (Warrendale, Pa.)* 1991, 681 – 686.
21. W. Boenigk, A. Niehoff, R. Wildförster, *Light Met. (Warrendale, Pa.)* 1991, 615 – 619. N. I. Steward, P. H. Halley, *Light Met. (Warrendale, Pa.)* 1994, 517 – 524.
22. Osaka Gas Co., DE 2 121 458, 1971 (K. Ueda, J. Kimoto, M. Moritake).
23. G. Collin, H. Köhler, *Erdöl, Kohle Erdgas Petrochem.* **30** (1977) 257 – 263.
24. C. D. Alexander, V. L. Bullough, J. W. Pendley, *Light Met. (Warrendale, Pa.)* 1971, 365 – 376.
25. G. L. Ball, C. R. Gannon, J. W. Newman, *Light Met. (Warrendale, Pa.)* 1972, 127 – 140.
26. W. Boenigk, A. Niehoff, R. Wildförster, *Light Met. (Warrendale, Pa.)* 1992, 581 – 584.
27. W. Boenigk, J. W. Stadelhofer, Proc. Carbon '92, Essen, June 22 – 26, 1992, pp. 30 – 32.
28. Gesellschaft für Teerverwertung, DE 642 437, 1935 (F. Meyer, O. Preiss).
29. Koppers, DE 705 576, 1939 (H. Koppers).
30. A. J. Gambro, D. T. Shedd, H. W. Wang, T. Yoshida, *Chem. Eng. Prog.* **65** (1969) no. 5, 75 – 79.
31. A. Eckert, R. Marrett, J. W. Stadelhofer, *Erdöl, Kohle Erdgas Petrochem.* **38** (1985) 510 – 514.
32. I. Moshida, Y. Q. Fei, Y. Korai, K. Fujimoto, R. Yamashita, *Carbon* **27** (1989) 375 – 380.
33. Rütgerswerke, DE 3 609 348, 1986 (M. Morgenstern, C. Bertrand).
34. G. Collin, M. Zander, Prep. 19th Biennial Conf. Carbon, University Park, PA, June 25 – 30, 1989, pp. 466 –467.
35. D. Stoye (ed.): *Paints, Coatings and Solvents,* VCH Verlagsgesellschaft, Weinheim 1993, pp. 91 – 94.
36. Rütgerswerke, DE 2 623 574, 1976 (H. Louis, H. Köhler, R. Oberkobusch, G. Collin, V. Potschka).
37. G. Herion, G. von Mossen, *Strasse Autobahn* **37** (1986) 3 – 6.
38. H. Spengler, W. Weskamp, G. Collin, *Cokemaking Int.* **3** (1991) 25 – 29.
39. B. Bujnowska, G. Collin, *Cokemaking Int.* **6** (1994) 25 – 31.
40. G. Collin, B. Bujnowska, *Carbon* **32** (1994) 547 – 552; Prep. Carbon '94, Granada, July 5 – 8, 1994, pp. 80 – 81.J. Polaczek, G. Collin et al., DGMK-Tagungsber. 9401, Velen, April 18 – 20, 1994, pp. 241 – 245.
41. H. Honda, *Carbon* **26** (1988) 139 – 156.
42. I. C. Lewis, Prep. 20th Biennial Conf. Carbon, June 23 – 28, 1991, Santa Barbara, CA, pp. 156 – 157.
43. I. Mochida et al., *Carbon* **28** (1990) 311 – 319.
44. Y. Korai et al., *Carbon* **29** (1991) 561 – 567.
45. K. J. Hüttinger, *Chemiker-Ztg.* **112** (1988) 355 – 366.
46. Rütgerswerke, DE 2 919 298, 1979 (H. Spengler, D. Pachaly).
47. Rütgerswerke, DE 3 813 634, 1988 (J. Talbiersky, K. Stolzenberg, B. Wefringhaus).
48. W. A. Bristow, *J. Inst. Fuel* **20** (1947) 109 – 130.
49. G. S. Pound, *Coke Gas* **21** (1959) 395 – 401, 446 – 453, 521 – 523.
50. K. R. Payne: Yearbook Coke Oven Managers' Association 1977, pp. 154 – 164.
51. A. D. Gregory: private communication.
52. S. A. Quader, G. R. Hill, *Chem. Ing. Tech.* **43** (1971) 595 – 596.

53. *Ullmann,* 4th ed., **22,** 439 – 441.
54. R. Rammler, H.-J. Weiss, A. Bussmann, T. Simo, *Chem. Ing. Tech.* **53** (1981) 96 – 104.
55. R. E. Joonas, W. M. Jefimow, T. A. Purre, *Chem. Ind.* **34** (1982) 739 – 740.
56. Rütgerswerke, DE 1 815 568, 1968 (O. Wegener, R. Oberkobusch, G. Collin, M. Zander, H. Buffleb).
57. J. Mostecky, M. Popl, M. Kuras, *Erdöl, Kohle Erdgas Petrochem.* **22** (1969) 388 – 392.
58. H. J. Sinn, *Chem. Ing. Tech.* **46** (1974) 579 – 589.
59. H. J. Sinn, W. Kaminsky, J. Janning, *Angew. Chem.* **88** (1976) 737 – 749.
60. G. Collin, G. Grigoleit, G. P. Bracker, *Chem. Ing. Tech.* **50** (1978) 836 – 841.
61. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans* **35** (1985).
62. J. S. Comaish, *J. Invest. Dermatol.* **88** (Suppl.) (1987) 61s – 64s.
63. R. Modica et al., *J. Chromatogr.* **24** (1982) 352 – 355.
64. C. L. Sanders et al., *J. Environ. Pathol. Toxicol. Oncol.* **7** (1986) 25 – 34.
65. J. S. Storer et al., *Arch. Dermatol.* **120** (1984) 874 –877.
66. P. M. Daniel et al., *Nature (London)* **215** (1967) 1142 – 1146.
67. C. S. Weil, N. I. Condra, *Am. Ind. Hyg. Assoc. J.* **38** (1977) 730.
68. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* **34** (1984).
69. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* **updating vols. 1 – 42, Suppl. 7** (1987).
70. U. Heinrich et al. in U. Mohr (ed.) : *Toxic and Carcinogenic Effects of Solid Particles in the Respiratory Tract,* 4th Int. Inhal. Symposium, Hannover, March 1 – 5, 1993, pp. 57 – 73.
71. U. Heinrich in U. Mohr (ed.): *Assessment of Inhalation Hazards,* Springer Verlag, Berlin 1989, pp. 301 –313.
72. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans,* **32** (1983) 95 – 452.
73. J. Santodonatro, P. Howard, D. Basu, "Health and Ecological Assessment of Polynuclear Aromatic Hydrocarbons," *J. Environ. Pathol. Toxicol.* **54** (1981) 1 – 364.
74. S. Tola, *J. Toxicol. Environ. Health* **6** (1980) 1253 –1260.
75. List of MAK and BAT Values, VCH Verlagsgesellschaft, Weinheim 1993.
76. Am. Conf. of Government of Industrial Hygienists (ACGIH) : *Threshold Limit Values,* 1990/1991.
77. EEC: Draft Annex Entries for Complex Coal-Derived Substances XI/51793, July 5, 1993 as Supplement of Chemical Substance Directive 67/548/EEC, Annex 1.
78. Inter. Tar. Association, 1993 (in preparation).
79. H. Tadokoro et al., *Ecotoxicol. Environ. Saf.* **21** (1991) 57 – 67.
80. A. E. Girling et al., *Chemosphere* **24** (1992) 1469 –1472.
81. G. Sundström: "Creosote," in O. Hutzinger (ed.): *The Handbook of Environmental Chemistry,* vol. 3D, Springer Verlag, Berlin 1986, p. 158.
82. K. G. Steinhäußer et al., *Vom Wasser* **72** (1989) 93.

Tar Sands

Frank J. Mink, Energy Resources Conservation Board, Calgary, Alberta, Canada

Richard N. Houlihan, Energy Resources Conservation Board, Calgary, Alberta, Canada

1. Introduction 2
1.1. Oil Sands and Crude Bitumen Defined 2
1.2. Geological Setting 3
1.2.1. Athabasca Oil Sands Area 3
1.2.2. Peace River Oil Sands Area 6
1.2.3. Cold Lake Oil Sands Area 6
1.3. Reserves 6
1.3.1. In-Place Reserves 6
1.3.2. Minable Reserves 6
1.3.3. In Situ Reserves 8
2. Minable Oil Sands Development . . 8
2.1. Mining Operations 8
2.1.1. Waste Management 9
2.1.2. Mining 9
2.2. Bitumen Production 10
2.2.1. Conditioning and Bitumen Froth Separation 10
2.2.2. Froth Treatment 12
2.3. Tailings Disposal 13
2.4. Bitumen Upgrading 15
2.4.1. Coking Processes 16
2.4.2. Hydrogen Addition Processes 18
3. In Situ Oil Sands Development . . . 18
3.1. Cyclic Steam Stimulation: Cold Lake Project 20
3.2. Steam Drive: Peace River Project . 21
3.3. Primary Crude Bitumen Production 23
3.4. Technical Issues Affecting In Situ Recovery 23
3.4.1. Casing Design 23
3.4.2. Gathering and Distribution System . . 24
3.4.3. Production Treating 24
3.4.4. Oily Waste Disposal 25
4. Economic Development 25
4.1. Mining Projects 25
4.2. Economic Viability 26
5. Challenges Past and Present 27
5.1. Technology 27
5.1.1. Mining and Bitumen Production Technology 28
5.1.2. In Situ Recovery Technology 29
5.1.3. Bitumen Upgrading Technology . . . 30
5.2. Environmental Issues 32
5.3. Public Policy 33
6. References 33

1. Introduction

The existence of the tar sands has been known for centuries. The term "tar sands" is often used to describe the deposits of unconsolidated bituminous sands found in Canada. In that the "bitumen" is a mixture of hydrocarbons and asphaltic or tarry material "oil sands" is a more appropriate term, and is used throughout this article. The sands are exposed along the banks of the Athabasca and Clearwater Rivers in the northern part of the province of Alberta. A number of other large deposits are found in deeper formations of the province (Fig. 1). Although the oil sands were described by fur traders and explorers, it was not until 1819 that geological description was attempted and it took until ca. 1887 before the nature and extent of the resource was assessed by the Geological Survey of Canada. While the potential for bitumen production and conversion to synthetic crude oil was recognized, considerable technological development was required before commercial projects could get underway.

Developments on the recovery of bitumen started in the 1920s, and expanded rapidly during the early 1940s. It declined in the late 1940s when large amounts of conventional oil were discovered in the province. Commercial surface mining began in the late 1960s, and has expanded since then. In situ recovery testing occurred simultaneously and commercial developments got underway in the mid-1980s (see Chap. 3).

The importance of research and development to oil sands exploitation is analogous to the importance of exploration in conventional oil development. New technologies have the potential to change the oil sands industry from a relatively high-cost source of oil to one that is more

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_129

Figure 1. Oil sands deposits in western Canada

competitive. The province of Alberta participates directly in that research through the Alberta Oil Sands Technology and Research Authority (AOSTRA). To 1993, AOSTRA and its industrial partners have provided over $ 10^9 in research and development investment. Most of this is for field pilot testing of in situ production, and mining, extraction, and upgrading technology. These projects are normally carried out on an equal funding basis between AOSTRA and a company or group of companies.

In addition, industry also invests in non-AOSTRA-sponsored research and development. The Alberta Research Council is active in oil sands research, as is the Federal Government of Canada, through the Canada Centre for Mineral and Energy Technology. In total, $ 10^5 was spent on oil sands research and development in Canada in 1991.

1.1. Oil Sands and Crude Bitumen Defined

Oil sands are the combined hydrocarbon and clastic sediment host rocks containing crude bitumen. Typically, they contain 84 % solids and 16 wt % crude bitumen and water (Fig. 2). The solids are 90 % quartz, with minor amounts of feldspar, muscovite, chert, and clay minerals. Clays are mainly kaolinite, illite, and montmorillonite. The large solid particles are surrounded by a thin film of water containing fine solids; the bitumen occupies the interstitial space.

The Alberta Energy Resources Conservation Board (ERCB) defines crude bitumen as: "A naturally occurring viscous mixture, mainly of hydrocarbons heavier than pentane, which may contain sulphur compounds, and that in its naturally occurring viscous state, will not flow to a well."

Crude bitumen is a complex mixture of organic compounds, containing polymeric asphaltenes and resins of high molecular mass and high levels of nitrogen, oxygen, sulfur, and metals. Typically, it contains sulfur 4.6 – 5.6 wt %, vanadium 160 – 300 mg/kg, nickel 60 – 100 mg/kg [2], and possibly small amounts of dissolved methane and traces of hydrogen sulfide. Crude bitumen is also notable for a deficiency of hydrogen in relation to carbon.

A distinguishing characteristic of crude bitumen is its viscosity. Alberta's crude bitumens are highly viscous, to the point of pseudoplasticity, with natural viscosity $10^4 - 10^6$ MPa · s, and density 960 – 1030 kg/m^3.

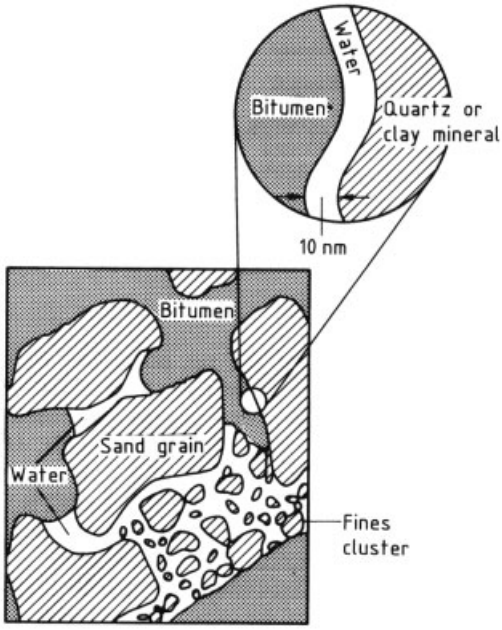

Figure 2. Structural model of the Athabasca oil sands [1]

1.2. Geological Setting

The crude bitumen deposits occur within Lower Cretaceous sands and Upper Devonian and Mississippian carbonates (Figs. 3 and 4,see next page) of the western Canadian sedimentary basin, a wedge of sedimentary strata thickening westward from a zero edge along the Canadian Precambrian Shield to the Cordilleran foreland thrust belt [3,4]. The deposits underlie an area of 86 000 km^2, about equal to that of Austria. The ERCB groups the oil sands deposits geographically into areas (Athabasca, Peace River, and Cold Lake) and separates them geologically by formation into deposits [5]; their main geological attributes are presented in Table 1 [6].

The hydrocarbons of the deposits are thought to have originated in the deeper part of the basin to the west, and to have migrated eastward updip, until a reversal in dip was reached in the north-eastern part of the province [7–9]. Overlying shales of the Mannville group form the

Table 1. Generalized geological attributes of the oil sands deposits [6]

	Depositional environment	Dominant grain minerals	Dominant clay minerals
Athabasca			
Grand Rapids			
Upper, Middle, Lower	shoreline and shallow marine	quartz, chert, feldspar, rock fragments	kaolinite, chlorite, illite } oil, sands
Wabiskaw/McMurray			
Minable, In situ	continental to marine shelf	quartz	smectite, illite, kaolinite, illite } water, sands
Cold Lake			
Grand Rapids			
Upper, Lower	continental to marine shoreline	quartz, rock fragments, feldspar	kaolinite, illite, smectite
Clearwater	marine shoreline	quartz, feldspar, rock fragments, altered grains	kaolinite, illite, smectite
Wabiskaw McMurray	marine shoreline continental	quartz, glauconite quartz	kaolinite, illite
Peace River			
Bluesky, Gething	estuarine to shallow marine	quartz, chert, rock fragments	kaolinite, illite
Carbonate deposits			
Debolt	open, marine	limestone, dolomite	not significant
Shunda	tidal flat to open marine	limestone, dolomite	not significant
Grosmont	tidal flat to open marine	dolomite	not significant

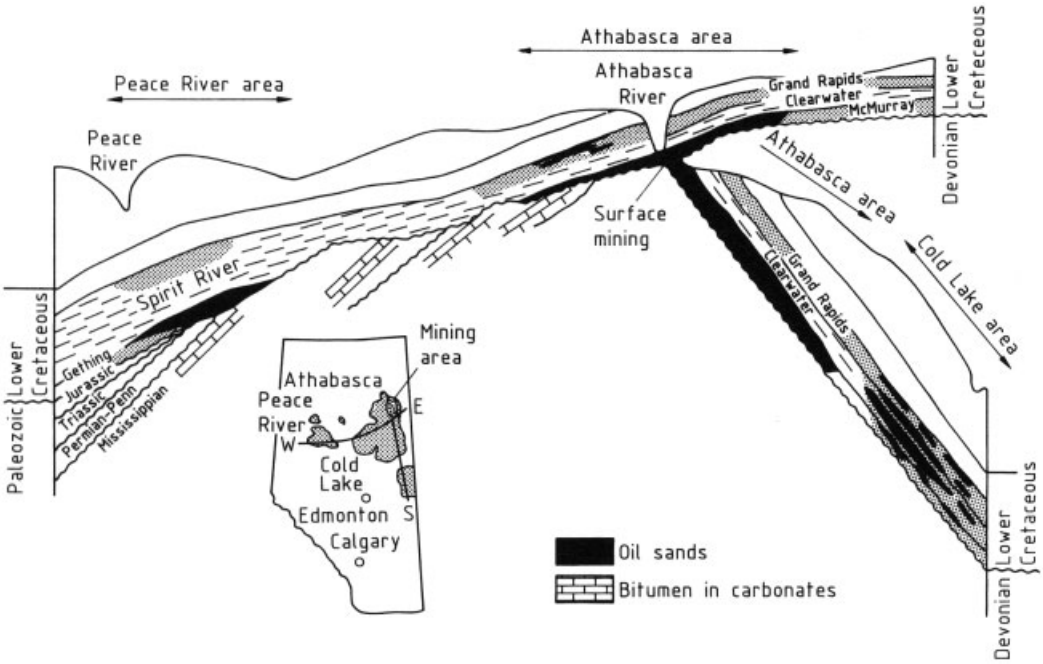

Figure 3. Stratigraphic relationships among the oil sands areas [8]

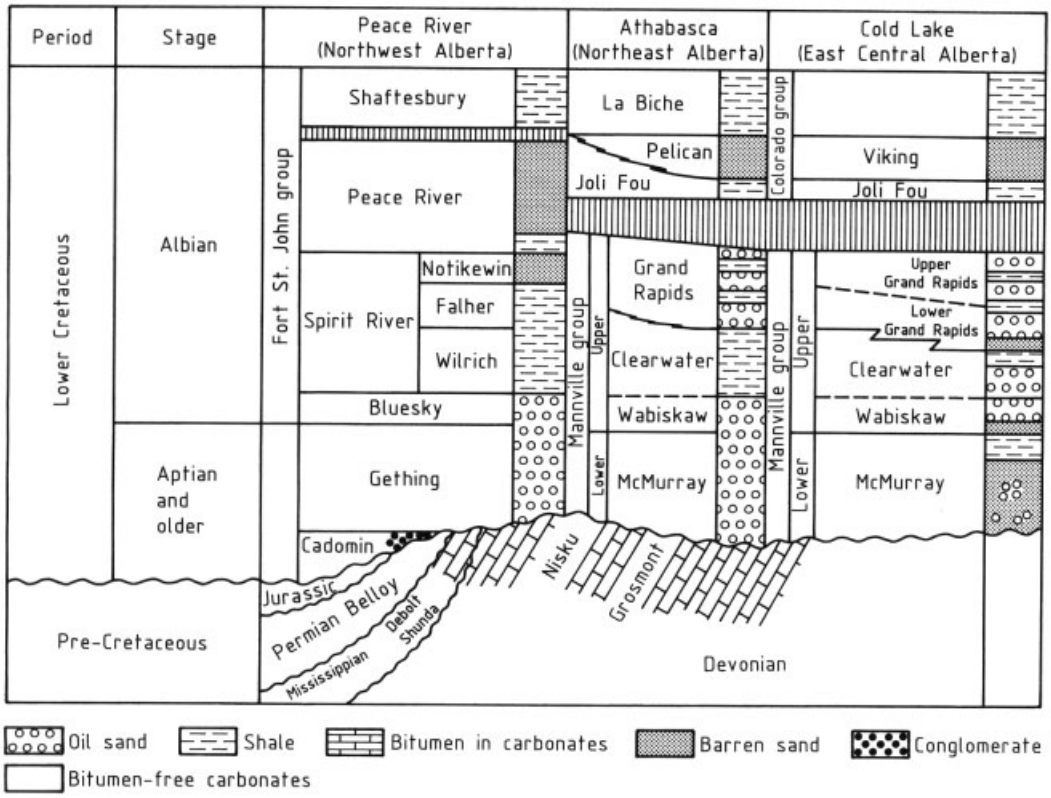

Figure 4. Stratigraphic correlation chart for the oil sands areas [8]

cap-rock for the hydrocarbons. The hydrocarbons were degraded to their present state through the action of microorganisms, water washing, and possibly inorganic oxidation [10].

1.2.1. Athabasca Oil Sands Area

Grand Rapids Deposit. The Grand Rapids sands were deposited in a high-energy, wave-dominated shoreline along a low-relief coastal plain. The reservoir sands consist of nearshore, shallow marine, and shore-attached beach complexes, with minor barrier islands and lagoons. The reservoir continuity is excellent parallel to the regional north-east to south-west depositional strike. Vertical reservoir continuity is good, but may be locally interrupted by thin, but extensive calcite or siderite cemented beds, or zones of calcareous concretions [6].

McMurray Deposit. The McMurray Formation consists of fluvial and estuarine sediments deposited in a ridge and valley system developed on the underlying Devonian carbonates [11–13]. An erosional contact separates the McMurray Formation from the overlying marine sands of the Wabiskaw member (Clearwater Formation) [6].

In the channelized McMurray Formation deposits, the sands are well sorted with little clay content, and are of excellent reservoir quality. The reservoirs are relatively narrow, but owing to vertical stacking, through the amalgamation of channel sands, thicknesses up to 70 m may occur.

The Wabiskaw member sands are marine in character, and generally consist of lower to upper shoreface and offshore sands, silts, and shales. They are generally fine to very fine grained and form laterally extensive, but relatively thin reservoirs [12, 14].

Grosmont/Nisku Deposits. Crude bitumen occurs along the eastern, updip edge of the Devonian Grosmont and Nisku Formations that subcrop below the Cretaceous sands in the Athabasca oil sands area [15]. The Grosmont Formation is a large, shallow-marine carbonate platform. Original depositional environments include relatively deep water, through patch reef accumulations within a platform interior setting, to restricted-marine shorelines and tidal flats. A vertical profile shows a series of shallowing-upward depositional cycles with excellent lateral continuity, but highly variable vertical reservoir pore fabric [6, 15]. The Nisku Formation is composed of carbonates overlying the Grosmont and separated from it by Ireton Formation shales.

1.2.2. Peace River Oil Sands Area

Bluesky/Gething Deposit. The Gething Formation consists of continental and marginal marine sands, silts, and shales overlaid by marine sandstones deposited along the trend of a paleovalley developed on the underlying Paleozoic surface. The Bluesky is a thin reworked Gething sand that represents the marine transgression that flooded the Gething coastal plain [16]. The bitumen is trapped in an updip pinchout of the sands against Paleozoic highs on the sub-Cretaceous unconformity [6].

Belloy Deposit. The Belloy Formation unconformably underlies the Bluesky/Gething. The reservoir rock was deposited as Permian, shoreline-related, basin margin sands, with sandy dolomites representing proximal, shallow marine facies [17].

Debolt/Shunda Deposits. The Debolt and Shunda Formations are Mississippian carbonate formations that unconformably underlie the Belloy. They form part of an overall shallowing-upward succession, representing restricted shelf – peritidal carbonates, siliciclastics, and evaporates [18].

1.2.3. Cold Lake Oil Sands Area

Grand Rapids Deposit. The Grand Rapids Formation consists of four distinct depositional environments, marine shelf – shore sediments, delta front sediments, bay-fill/delta plain sediments, and distributary channel-fill sediments [19]. The thickest oil sands are associated with the channel sands, while the marine sands are thinner, but form the most extensive continuous reservoirs [6].

Table 2. Reserves and characteristics of the oil sands deposits [5,6]

	In-place volume, 10^6 m^3	Area, km^2	Overburden depth, m	Pay thickness, m	Bitumen saturation		Average porosity, %	Water saturation, %
					wt %	Pore, vol %		
Athabasca								
Upper Grand Rapids	4140	3340	150 – 450+	9	6.2	55	30	45
Middle Grand Rapids	1410	1820	150 – 450+	5	7.7	68	30	32
Lower Grand Rapids	1220	1730	150 – 450+	6	5.1	45	30	55
Wabiskaw/McMurray	141 750	43 290	0 – 750+	35	9.2	62	28	30
Nisku	10 330	4990	200 – 800+	8	5.7	63	21	37
Grosmont	50 500	118 910	200 – 800	10	5.0	75	16	25
Cold Lake								
Upper Grand River	7400	8160	300 – 600	6	8.1	58	30	42
Lower Grand Rapids	12 950	7500	300 – 600	12	8.7	60	31	40
Clearwater	11 050	5890	300 – 600	15	8.9	64	30	36
Wabiskaw/McMurray	3860	6250	300 – 600	6	5.7	51	25	49
Peace River								
Bluesky/Gething	14 040	9760	300 – 700	6	6.1	60	23	40
Belloy	282	260	675 – 700	8	7.8	64	27	36
Upper Debolt	1830	1000	500 – 800	13	5.0	61	19	39
Lower Debolt	5970	2020	500 – 800	29	5.1	67	18	33
Shunda	2510	1430	500 – 800	14	5.3	52	23	48

Clearwater Formation. The sands of the Clearwater Formation are interpreted as river-dominated deltaic deposits with laterally offset shoreface sands [20]. Owing to the extensive vertical and lateral continuity of the Clearwater sands, they form the main reservoir in this area [21].

McMurray Deposit. The McMurray Formation sands in the Cold Lake area were deposited in fluvial and lacustrine environments. The Wabiskaw member sands are the transitional sediments between the underlying continental McMurray Formation and the marine Clearwater Formation [6].

1.3. Reserves

The ERCB appraises crude bitumen reserves in each of Alberta's oil sands deposits and updates its appraisal annually. Various categories of reserves are published, ranging from the volume determined to be present based on drilling information (the in-place volume) to the volume which is conservatively expected to be recovered at existing developments (the established reserve). The ERCB also provides a broad estimate of the reserve that could be recovered before activity ceases (the ultimate potential). The in-place volume, established reserves, and ultimate potential are 269×10^9 m^3, 482×10^6 m^3, and 49×10^9 m^3, respectively.

1.3.1. In-Place Reserves

The oil sands deposits in Alberta contain an estimated in-place volume of 269×10^9 m^3 crude bitumen (Table 2). Initial in-place volumes of crude bitumen in each deposit are determined with different cutoff criteria. Crude bitumen in-place volumes within the Cretaceous sands are determined with a minimum saturation of 3 wt % crude bitumen (approximately 20 % of pore volume), and a minimum saturated zone thickness of 1.5 m; in the carbonate occurrences, they are determined with a minimum bitumen saturation of 30 % of pore volume and porosity 5 %.

1.3.2. Minable Reserves

The minable area is defined as that part of the Athabasca Wabiskaw – McMurray deposit where the overburden thickness generally does not exceed 75 m. This area is ca. 3250 km^2, more than 7 % of the total area of the Athabasca oil sands. Approximately 24×10^9 m^3 crude in-place bitumen is contained in this area. The determination of minable reserves from this vol-

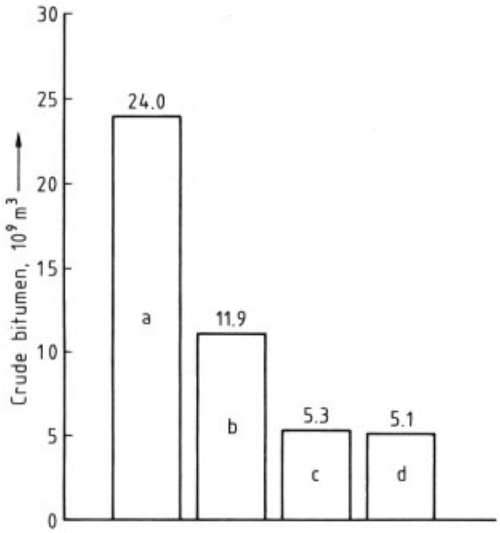

Figure 5. Crude bitumen reserves categories within the surface-minable area
a) Initial in-place volume; b) Initial minable in-place volume; c) Initial established minable reserve; d) Remaining established minable reserve

ume is discussed below and illustrated in Figure 5.

Potentially minable areas are identified by applying economic stripping ratio (ESR) criteria, a minimum saturation of 5 wt % crude bitumen, maximum shale content 45 vol %, and minimum saturated zone thickness 1.5 m. The ESR is a modified version of the more widely used cutoff stripping ratio; it defines a curve of increasing strip ratios at increasing ore grades where the area below the curve is economic and that above the curve is uneconomic (Fig. 6). The ESR considers product price, project costs, and bitumen recovery for diferent ore grades to define a pit limit within an ore body, where the net revenue is zero [22].

Within the potentially minable area, the initial, or potentially, minable in-place volume of crude bitumen is calculated as 11.9×10^9 m^3 (see Fig. 5). Allowing for surface facilities such as plant sites, tailing ponds, discard sites, environmental protection corridors along major rivers, and isolated minable areas, assuming a combined mining/extraction recovery factor of 0.82, the resulting in-place minable reserve of crude bitumen is estimated as 5.3×10^9 m^3 (initial established minable reserve). Production to date has reduced this reserve to 5.1×10^9 m^3, the remaining established minable reserve. Techno-

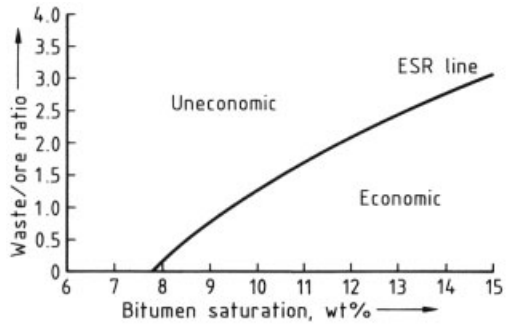

Figure 6. Economic stripping ratio relationship

logical improvements, better placement of surface facilities in future projects, and improved project economics could increase the estimates. Only a small portion of the initial minable reserve is being developed. The existing Suncor and Syncrude Canada projects have estimated established minable crude bitumen reserves at December 1992 of 82×10^6 m^3 and 352×10^6 m^3.

1.3.3. In Situ Reserves

In deeper formations, crude bitumen can be recovered through wells (in situ recovery). Since crude bitumen is not normally mobile in the reservoir, the application of thermal energy is required, and only areas that may be developed for thermal recovery are included in the established reserves. The initial in-place volume for developed areas is based on the expected drainage areas and a regard for the spacing of wells or well clusters. Established reserves are determined based on recovery mechanism, e.g., 20 % recovery is assumed for cyclic steam stimulation projects. For projects with a primary recovery component, the in-place volumes are based on the assumed full development of all project lands not currently developed for thermal recovery. The initial established primary reserves are based on an average primary recovery factor for the particular sand being developed. In the active experimental schemes, an initial established reserve figure of 17.7×10^6 m^3 is set.

The ERCB estimates the remaining established reserves of crude bitumen from the deposits under active development to be 434×10^6 m^3 for surface-minable schemes, and 48.2×10^6 m^3 for in situ schemes. The reserves

remaining to be developed are vast. In the surface-minable area the established reserve is 5.1×10^9 m^3.

2. Minable Oil Sands Development

Surface mining of oil sands and production of synthetic crude oil is well established in Alberta. The Syncrude and Suncor projects produced approximately 17 % of Canada's oil requirement in 1992. The large reserve base, proven technology, and an established infrastructure all have a positive influence on the viability of oil sands projects, but their development is handicapped by high capital and operating costs and long lead time from project initiation to production.

Figure 7 shows the Syncrude project. Table 3 provides an overview of the quantities of oil sand processed, bitumen recovered, and synthetic crude oil (SCO), sulfur, and coke produced at the Syncrude and Suncor projects in 1992.

Figure 7. Syncrude oil sands project
Dragline mining in foreground; bitumen extraction and upgrading at top right.

Table 3. Oil sands mining projects, 1992 production volumes [23]

	Syncrude	Suncor
Oil sands mined, t/d	341 650	110 140
Bitumen in oil sands, wt %	10.8	11.9
Bitumen production, m^3/d	33 610	11 620
Synthetic crude production, m^3/d	28 490	8720
Coke production, t/d	3990	2730
Sulfur production, t/d	1230	340

2.1. Mining Operations

Mining the oil sands and recovering the bitumen involves stripping the overlying overburden, mining the oil sand, and transporting it to the bitumen extraction plant, where hot water washes the bitumen out of the sand. The bitumen is upgraded and refined to produce synthetic crude oil, and the waste sand is pumped to disposal sites. For economic operation, very large volumes of material must be mined, transported, processed, and disposed of continuously. In 1992, the large Syncrude project typically handled over 700 000 t/d of in-place materials to produce 341 650 t/d of mined oil sands; additionally over 263 000 t/d of waste sand product (tailings) was disposed of.

2.1.1. Waste Management

The first steps in the mining operation are tree clearing and muskeg drainage and stripping. Muskeg is a relatively thick, water-saturated peat, composed of black organic material in varying stages of decomposition. It varies in depth from 2 to 6 m and, because of its high organic content, is stored for later use in mine reclamation. Before it can be removed, the muskeg must be drained by an extensive series of ditches.

Overburden stripping is carried out by a truck and shovel fleet. Shovels vary in size and type from 14 m^3 hydraulic diesel-powered shovels to 44 m^3 cable-electric-powered shovels. Haul fleets vary from 68 to 218 trucks.

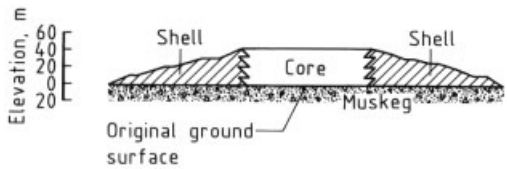

Figure 8. Cross section of a waste dump

Overburden materials are mainly composed of tills underlying glacial lacustrine soils, or a variety of meltwater or outwash alluvial deposits [24]. Most of the overburden material is used to construct waste dumps and containment dykes for the tailings produced from the extraction process. Some of the more granular deposits are used for road construction and the wetter, siltier material is dumped in-pit or within the core of the waste dumps (Fig. 8). The wet overburden soils are dumped into an inner core, formed by

constructing a shell of compacted overburden. The waste dumps are 30 – 55 m high with side slopes ca. 3.3:1 – 4.4 : 1. Most have been constructed out-of-pit and founded on thick muskeg deposits. Careful attention must be given to foundation conditions and construction rates to prevent high pore pressures developing, and to consolidate the underlying in situ muskeg.

Oversize lumps and clay or siltstone stringers are removed from the ore during the early stages of processing. This material is very wet and difficult to use as a construction material. Some success has been achieved by using it to construct a clay core within a tailings sand dyke [25]. Oversize material not needed for construction purposes is wasted into the mined-out pit areas.

2.1.2. Mining

Before mining can begin, the basal water sands in the mining area have to be depressurized. These are discontinuous silty sands 1 – 12 m thick. These aquifers pose potential stability hazards to the single-bench mining highwalls, the final pit walls, and the stability of the bucketwheel excavators operating on the final mine bench. A unique characteristic of oil sand basal aquifers is the presence of dissolved gases in the water [26].

Syncrude uses four 70 m^3 walking draglines operating from a single bench. The oil sand is spoiled on the operating bench behind the dragline and is later reclaimed by a system of bucketwheel reclaimers and conveyors. The draglines are also capable of spoiling interburden, either directly or via rehandle, into the mined-out pit. Supplementary oil sands feed is supplied via a mobile equipment fleet. It is sized and transferred to the conveyor network via feeder – breakers and double-roll crushers. A dump pocket with approximate capacities of 15 000 t live and 200 000 t dead storage provides surge capacity.

The Athabasca oil sands have been classified as locked sands [27]. Naturally occurring slopes have been observed up to 65 m high at angles of 55°. Owing to the very high strength of the oil sands and the abrasiveness of the sand, blasting is sometimes used to improve productivity and reduce wear on equipment [28]. Blasting is also done for soil stabilization. Steeply dipping clay beds are blasted at Syncrude to reduce the risk of block slide failures; blast densification of loose beach sands has been done at both Suncor and Syncrude [29, 30].

Highwall stability is a safety concern at both oil sands mines, particularly at the Syncrude mine where the dragline operates on a single-bench highwall with 50° slopes and vertical height > 60 m [31]. Four basic modes of highwall failure have been identified, largely in the Syncrude operation: slabbing failures, highwall deformations, shallow block slides, and basal failures (Fig. 9,see next page). In contrast, the present Suncor mining operation is carried out on three to five benches, each ca. 20 m high.

Both Syncrude and Suncor rely heavily on instrumentation and observation in dealing with highwall performance and design [32]. An important part of this approach is to obtain reliable geological data based on structural boreholes, highwall mapping, and past highwall performance. When a highwall has been designed, suitable instrumentation is installed to monitor the mining operation and the performance of the highwall. Instrumentation includes slope inclinometers, open standpipe piezometers, pneumatic piezometers, and electronic distance measurement systems. Visual monitoring by specially trained engineers also plays an important role.

Highwall monitoring and design at Suncor has not been required to the same extent as at Syncrude. Bucketwheel excavators, and the current truck and shovel system, operating at the base of a 20 m mining face do not require the same degree of control as that needed for a dragline operating at the top of a 60 m single bench. The first stage of recovering oil from the oil sands involves bitumen recovery.

2.2. Bitumen Production

Bitumen production operations include a feed surge system, oil sand conditioning, bitumen separation, and froth treatment facilities. The feed surge is required between the mine and extraction to decouple these operations and allow either to operate while the other is unavailable for reasons of upset or maintenance. Oil sand is

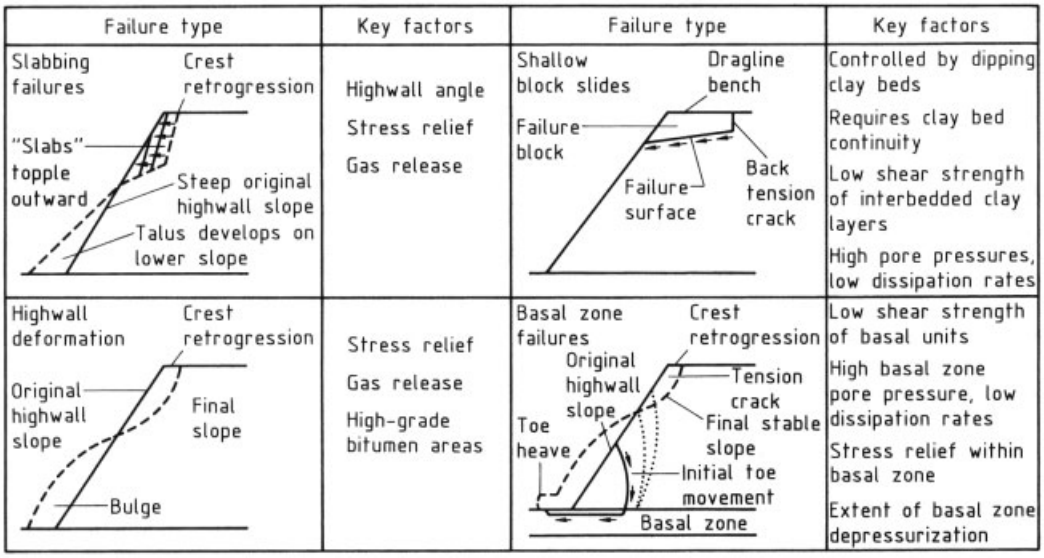

Figure 9. Highwall failure modes in oil sands

withdrawn from the surge facility through hoppers and apron feeders which feed extraction feed conveyors.

2.2.1. Conditioning and Bitumen Froth Separation

The process for bitumen froth production used at commercial operations is based on the hot water extraction process invented by Karl A. Clark in the 1920s [33, 34]; it was developed through pilot testing and first applied commercially in 1967 by Suncor (formerly Great Canadian Oil Sands). The process depends on the water-wetted nature of solids in oil sands.

The Clark hot water extraction process is illustrated in Figure 10. In the first step, oil sand is conditioned by mixing in a rotating drum with hot water, sodium hydroxide, and steam to pro-

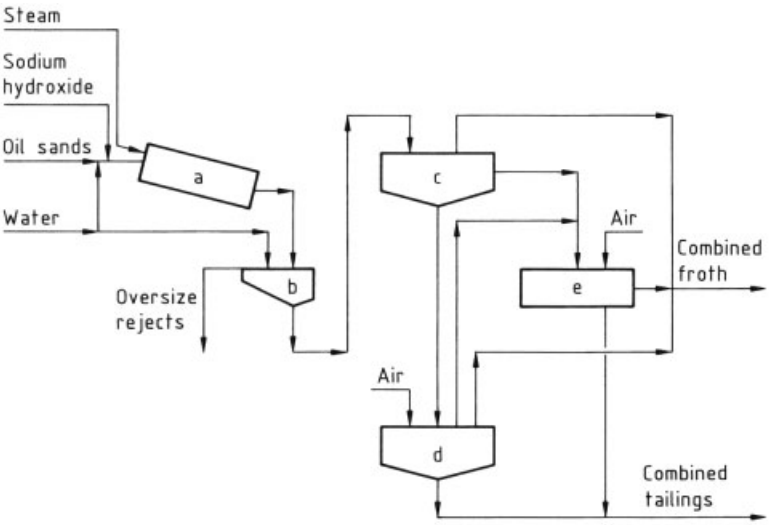

Figure 10. Bitumen froth production
a) Conditioning drum; b) Oversize screen and pump box; c) Primary separation vessel; d) Secondary separation vessel; e) Flotation vessel

duce a slurry of ca. 65 % solids at ca. 80 °C. The process involves the addition of mechanical, thermal, and chemical energy. Thermal energy is supplied by hot water and steam, mechanical energy by the rotating and tumbling action in the conditioning drum, and chemical energy results from the generation of surfactants, produced by reaction of sodium hydroxide with organic acids occurring naturally in the bitumen; occasionally surfactants are available in the oil sands without treatment. Bitumen recovery is highly dependent on surfactant concentration; sodium hydroxide is used to adjust the surfactant concentration and maximize recovery [35, 36]. The conditioning drum action causes the bitumen film which envelops the oil sand particles to rupture, allowing solids and bitumen to separate. Bitumen is also aerated in the process, and separates as a froth from the water – solid slurry in separation vessels. The residence time in the conditioning drum is set to ensure oil sand lumps are broken up and the bitumen is liberated into the slurry.

Variables affecting recovery are: process temperature, water/ore ratio, and residence time. Syncrude's warm water extraction process operates at 50 °C, and a cold water process has been developed, designed to operate at ca. 20 °C. Conditioning drum water/oil ratios in the range 0.2 – 0.4 are used, depending on the ore grade. Sodium hydroxide use depends on the ore grade and varies from 0 to 0.04 wt % of oil sands feed.

The conditioning drum slurry flows through vibrating screens to remove oversize material, consisting mainly of rocks, clay lumps, and undigested oil sands. The oversize material is washed with hot water to minimize bitumen loss, and conveyed to storage bins that are discharged to trucks and eventual disposal in mined-out pits. The conditioning drum slurry is further diluted with water and pumped to primary separation vessels.

The bitumen floats to the surface of a primary separation vessel as froth where it overflows into launders. The froth, also containing water and solids, is deaerated prior to further processing. Coarse solids settle to the bottom of the primary separation vessel and are withdrawn along with fine solids, some bitumen, and water, either to a tailings oil recovery system, or discharged to a tailings pond. The density of the separation vessel underflow is a key control variable. The tailings oil recovery system is a secondary settling vessel where, through dilution with water and other recycle streams, additional bitumen froth is produced. The underflow from this vessel is pumped to the tailings pond.

Conditions in the middle zone of the primary separation vessel are controlled by upstream variables, such as water and sodium hydroxide addition, and by withdrawing a middlings stream to ensure that the fluid density and viscosity allow aerated bitumen particles to rise to the froth layer [37]. The vessel middlings stream and the middlings stream from the tailings oil recovery circuit are treated in air flotation vessels to produce additional bitumen froth. This froth, of lower quality than primary circuit froth, is then fed to a second froth deaerator. Flotation vessel tailings flow to tailings pump boxes. Bitumen froth, following deaeration and further cleaning if necessary, is fed to storage or directly to froth treatment.

Syncrude's commercial operation has four similar extraction trains, each with oil sand throughput rate up to 3000 kg/s per train. Suncor has five trains, with throughput rate 1225 kg/s. Sufficient capacity is provided to allow for variations in oil sands feed rates, the water:ore ratios necessary to accommodate changes in oil sands properties and maintenance downtime.

The hot water extraction process is efficient. Bitumen recovery is highly dependent on oil sands feed grade (Fig. 11) [38–40]. Key factors are fines content, depositional environment, and bitumen content.

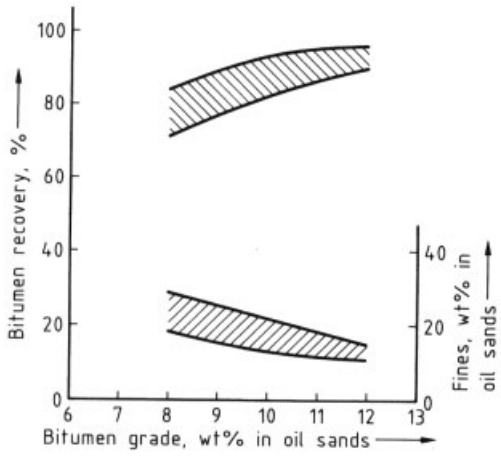

Figure 11. Bitumen recovery and relationship of fines to bitumen grade

A major concern with the existing process is the large volume of poor settling tailings produced and, as a result, the need for extensive tailings ponds and sand containment areas. Coarse solids in tailings settle rapidly in separate containment areas, or are used to form dykes to contain the remaining fluid tailings. The fluid tailings settle to ca. 30 wt % solids, and must be contained indefinitely or treated. Clear tailings water is recycled for reuse in the extraction plant. Ultimately, fine tailings which do not settle beyond 30 wt % solids in water are disposed of in mined-out pits and reclaimed (see Section 3.4).

2.2.2. Froth Treatment

Figure 12 illustrates a commercial froth treatment plant. Syncrude and Suncor processes have two-stage centrifuges of conventional design installed in parallel.

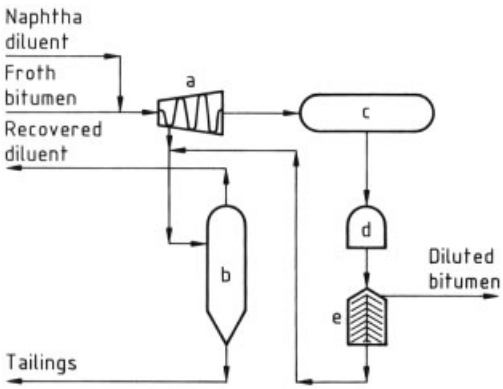

Figure 12. Froth treatment
a) Scroll centrifuge; b) Naphtha recovery unit; c) Disk centrifuge feed drum; d) Filters; e) Disk centrifuge

Deaerated bitumen froth containing ca. 8 wt % solids and 40 wt % water is treated by centrifuging to reduce solids to < 0.6 wt % and water to ca. 5 wt %, to satisfy feed requirements for downstream bitumen upgrading units. The froth is diluted with naphtha and centrifuged in two stages at ca. 76 °C. Naphtha recovered from the upgrading of bitumen is used as diluent at a naphtha/bitumen ratio of 0.8 by weight. The first stage is a scroll-type, horizontal bowl centrifuge operating at 800 rpm (ca. 250 *g*). Coarse solids (> 44 µm) are removed as a cake. The hydrocarbon product containing water and fine solids flows to a second stage disk centrifuge where most of the water and remaining solids are removed. Disk centrifuges operate at 3600 rpm (ca. 2500 *g*). A filter on the feed line to the disk centrifuge removes coarse solids. The diluted bitumen product of the disk centrifuges is pumped to a diluent recovery unit where naphtha is distilled off and recycled. Centrifuge facility tailings, mainly water and solids, contain naphtha, some of which is recovered by a vacuum flash process. Bitumen and naphtha recovery in the treatment process is ca. 99 wt %.

The Syncrude froth treatment plant throughput has been increased by the addition of an inclined plate settler that takes one-third of the feed, and operates in parallel with the centrifuge system [41]. The settler feed diluent:bitumen ratio is 0.5 by weight.

Syncrude has tested a three-stage process using inclined plate gravity settlers with the froth and diluent flowing concurrently. Other research and development has led to a process which, in two stages of mixing and gravity separation at 160 °C, results in a bitumen product of lower water and solids content.

2.3. Tailings Disposal

The tailings – sand, clay, water, and bitumen discard from the bitumen production operations – are stored in ponds where the solids are allowed to settle out and the clarified water is recycled to the extraction plant. Containment, handling, and disposal of tailings pose major economic and environmental problems.

In the initial years of operation of an oil sands mine, when recycle water is not yet available owing to the very slow fine tailings dewatering rates, tailings storage is required to hold three times the volume of oil sands for ca. 5 a. This volume increase is called a tailings bulking factor. Later, as recycle water becomes available from the tailings pond, the bulking factor reduces to 1.4. It is this bulking, and the very slow release of water from the fine tailing, that is the focus for much of the ongoing research into tailings activities and final reclamation planning.

Gradation data for oil sands tailings produced by existing oil sands operations is shown in Figure 13 [42]. While the tailings are relatively

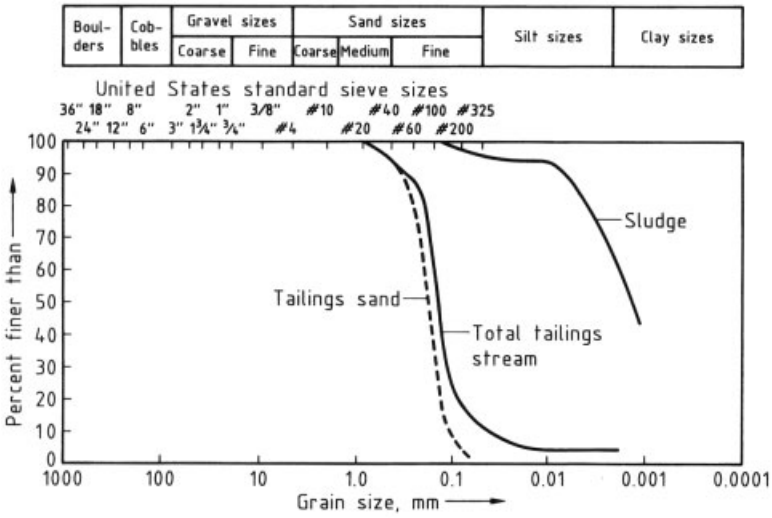

Figure 13. Grain size distribution of oil sands tailings

clean and coarse overall, a significant slime fraction is present (a mixture of silt, clay, and oil residue).

Poor sedimentation and consolidation characteristics of the slimes have complicated tailings disposal [43]. When the tailings are deposited, the coarse sand fraction settles out almost immediately, while the dispersed silts and clays remain suspended in the water to form a sludge-like material on top of the sand. During dyke construction and beaching operations, ca. 50 % of the fine tailings are captured in the voids of the coarse tailings. The remaining fine tailings enter the pond as a thin slurry (ca. 3 – 8 wt % solids). The fine tailings consolidate very slowly to a maximum concentration of ca. 35 % solids. The volume of these mature fine tailings and their toxic nature are of major public concern.

Some of the bitumen lost to the pond in the tailings stream is also released into the recycle water layer. This free bitumen either floats to the surface, and in some cases is recovered with booms and vacuums, or collects below the surface to form large suspended pools of bitumen, and is pumped to storage tanks at the shoreline for trucking back to the plant.

Containment of the tailings is also of interest. Tailings dykes have been constructed by a modified method of upstream construction (Fig. 14). The dyke typically consists of a downstream shell of compacted sand and an upstream supporting beach of sand. The compacted shell and supporting beach are placed by hydraulic construction techniques (Fig. 15, see next page).

Beaching is conducted from either a single discharge point or from multiple discharge

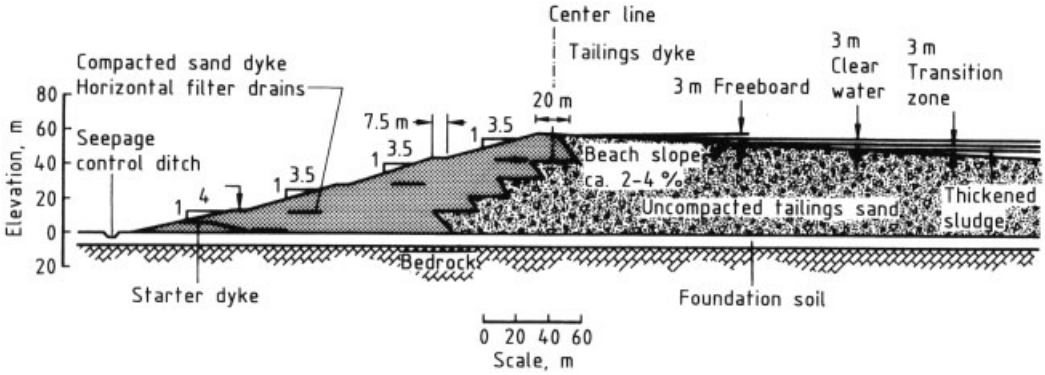

Figure 14. Typical dyke section

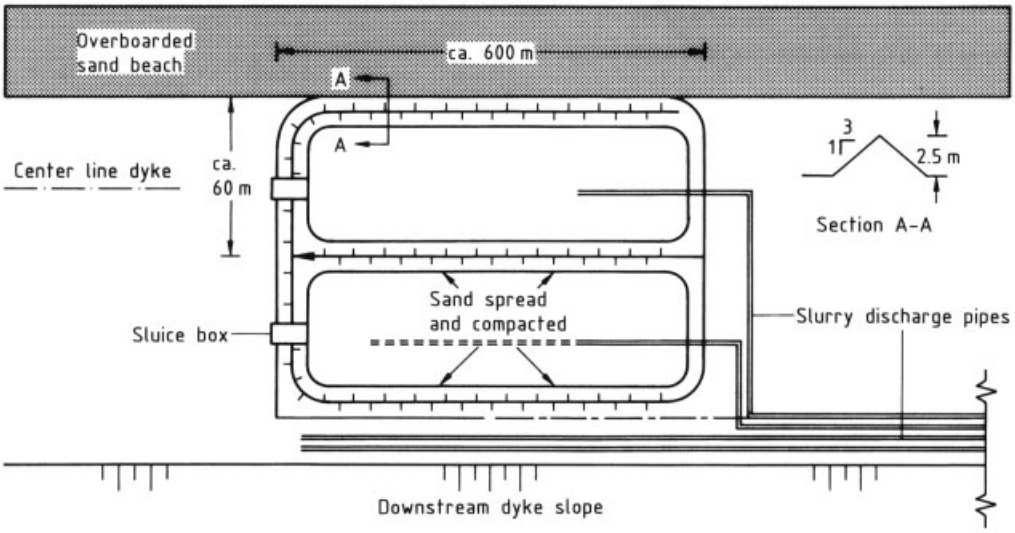

Figure 15. Construction of tailings dyke cells

points along the pipeline (spigotting). Beach slopes range from 2 to 4 %. The discharge point is moved along the dyke at appropriate time intervals to ensure adequate upstream beach support without imposing large stresses on the underwater beaches, thereby triggering a liquefaction failure. Tailings dyke design and performance monitoring are key components in a safe tailings management program. The observational approach is used as part of this program [44]. As already described in Section 2.1.2, extensive visual and instrumentation monitoring is required. Tailings dyke integrity has been reported as being a function of three major geotechnical design variables: seepage control, foundation movement, and overall dyke stability, as well as the liquefaction potential of loose sands.

Seepage control is important both to minimize groundwater contamination and to ensure structural integrity. Internal drains or filters are installed within the downstream section of the tailings dyke to control the phreatic surface and to militate against potential piping problems.

Existing operations have experienced significant foundation movements as a result of tailings dyke construction. Weak presheared clay shales are present under Syncrude's tailings pond [44], and Suncor's first tailings pond was founded on a normally consolidated alluvial silty clay [45]. Several types of remedial measures have been used to protect overall dyke integrity.

Although the Fort McMurray area is in a low seismic risk area, the potential for liquefaction was considered a possibility. Strains induced in the loose beach sands by foundation movements and from the additional loading of dyke construction, may be enough to cause liquefaction. Blast densification of loose sands and construction of toe berms has been used successfully on both projects to offset this potential risk.

Both plants are obliged to reclaim the surface of the disturbed areas to levels of productivity similar or better than those prior to development. Because of the large area and volumes, the operators of the plants are advocating an integrated reclamation strategy with a mixture of "dry" and "wet" landscapes.

2.4. Bitumen Upgrading

Bitumen produced from oil sands is a heavy viscous material which is difficult to transport and not readily processed in conventional refineries. At oil sands mining schemes in Alberta, upgrading occurs at the bitumen production site to produce "synthetic" crude oil. The characteristics of this partially refined product are described later. The need to refine bitumen depends on economics, integration of mining, bitumen production, and upgrading operations, markets for bitumen or synthetic crude oil, transportation con-

straints, and the ability of refineries to process bitumen and synthetic crude oil.

The Syncrude and Suncor projects, which commenced operation in the late 1960s and 1970s, are stand-alone, integrated operations comprising mining, bitumen extraction, bitumen upgrading, and utilities. In contrast to the minable oil sands developments, in situ recovery projects (Chap. 3) produce crude bitumen which is blended with a hydrocarbon diluent and pipelined to upgraders, primarily in the United States. Trends in the minable oil sands area are toward bitumen production operations at the mine site, and transportation of diluted bitumen to regional or remote upgraders.

The need for upgrading is illustrated by a comparison of the properties of crude bitumen and conventional oil (Table 4). Bitumen has high density, high viscosity, and high pour point. Boiling point data show that bitumen contains little or no hydrocarbon boiling below ca. 195 °C, and approximately 86 % of bitumen boils above 345 °C. The essence of upgrading bitumen is to reduce high-boiling material, to increase the hydrogen:carbon ratio, and to remove impurities.

Figure 16 illustrates the bitumen upgrading process. Diluted bitumen product is distilled for diluent recovery. Bitumen is upgraded by a primary upgrading process, and liquid products refined to produce synthetic crude oil. Hydrogen is required by the process. Byproducts are fuel gas, sulfur, and coke. Fuel gas is used in the process, sulfur is sold and, in the absence of on-site use or market, coke is stored.

Table 4. Crude properties [46]

	Bitumen	Synthetic crude	Conventional crude
°API	8	32	41
Sulfur, wt %	4.8	0.08	0.2
Nitrogen, wt %	0.45	0.03	0.04
Viscosity, 10^6 m^2/s, 40 °C	3000	3.0	2.9
Pour point, °C	18	−4.5	−6
Metals, mg/kg	450	<1	
Distillate yield, vol %			
IBP – C 5∗	0	4	3
C 5 – 195 °C	0	18	36
195 °C – 345 °C	14	47	31
345 °C – 560 °C		31	18
	86		
560 °C +		0	12

∗ IBP = Initial boiling point. C 5 = Pentane.

In the conversion of bitumen to synthetic crude oil, a two-stage upgrading sequence is used. The purpose of the primary step is to convert the high-boiling, high-molecular residue in the bitumen to distillate hydrocarbon. The secondary step involves hydrotreating distillate in desired boiling ranges (→ Oil Refining, Chap. 3.3.), and adding hydrogen to saturate hydrocarbons and remove impurities.

Primary processes to upgrade bitumen can be categorized as carbon removal or hydrogen addition. Both types of process increase the hydrogen:carbon ratio compared with crude bitumen. The carbon removal or coking processes involve heating the crude bitumen above 500 °C at near-atmospheric pressure; the high-molecular material cracks to produce distillate hydrocarbons and coke (→ Oil Refining, Chap. 3.6.4.). The coke forms by combination of reactive high-

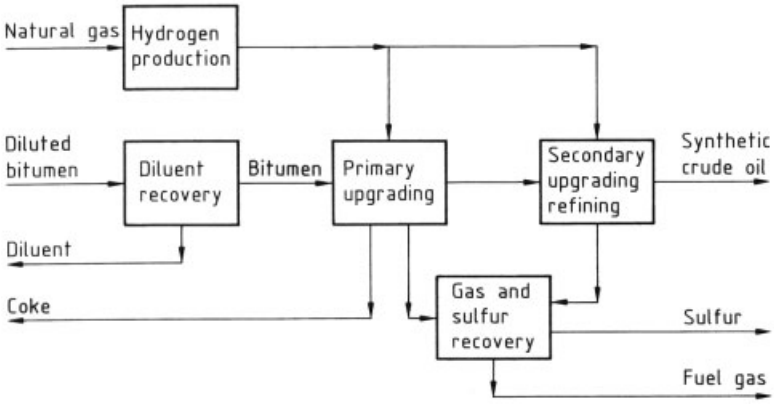

Figure 16. Bitumen upgrading

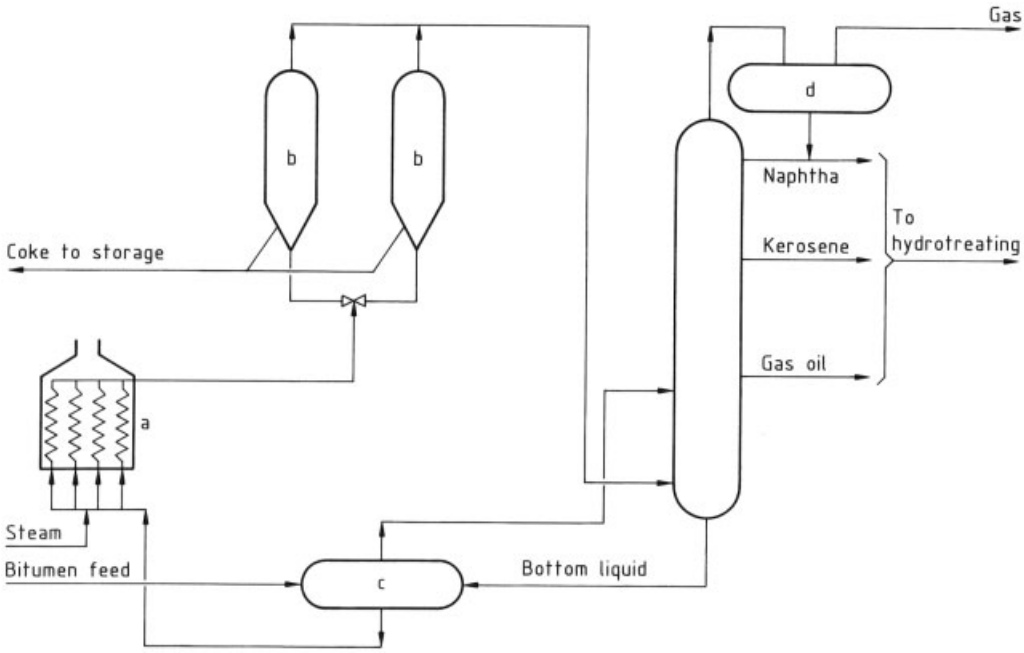

Figure 17. Suncor delayed coking
a) Furnace; b) Coke drums; c) Feed drum; d) Condenser

molecular material. Hydrogen addition processes involve reaction of bitumen with hydrogen above 450 °C and pressure > 7 MPa, sometimes in the presence of a catalyst, to produce distillate hydrocarbons and a liquid residuum. The process involves thermal cracking and hydrogen addition to the cracked product.

2.4.1. Coking Processes

A wide range of technology is available for primary upgrading, but few processes can be applied commercially for oil sands crude bitumen upgrading. Early developments relied on coking while more recent efforts have turned to hydrocracking to upgrade the bitumen. Coking processes include delayed coking and fluid coking (→ Petroleum Coke, Chap. 3.). In each case the crude bitumen charge is converted to distillate oil and coke. The coker distillate is a partially upgraded material, and is a suitable feed for hydrotreating to a sweet synthetic crude oil. The coke fraction of the product may be used for fuel.

In the delayed coking process, the bitumen is heated to 480 – 510 °C to initiate the coking reactions. The heated oil is fed to a large drum, where residue is deposited as coke, and more volatile material flows through to a fractionator. Delayed coking is a batch process, and a minimum of two coking drums are required; as coke is formed in the on-line drum, coke is being removed from the off-line drum by steam and water cutting (Fig. 17).

The Syncrude fluid coking technology (developed by Exxon) combines the principle of coking with that of fluid solids processing, as used in fluid catalytic cracking (Fig. 18).

Preheated bitumen feed at ca. 260 °C is introduced into a fluidized reactor. The reactor thermally cracks the feed at ca. 510 °C. Heat for the reaction is supplied by recycling hot coke from the coker burner. The cracked vapor is scrubbed with recycled heavy oil and fed to the fractionator. Coke from the reactor is fed to the coker burner, where a portion of the coke is combusted with air to provide heat for the reaction. Recycled coke is used as a heat carrier to the reactor. Excess coke from the burner is removed to storage. Coker burner gas is combusted with supplementary fuel and air to generate steam in a CO boiler. Flue gases are vented through an electrostatic precipitator to remove particulates, and routed to an elevated stack.

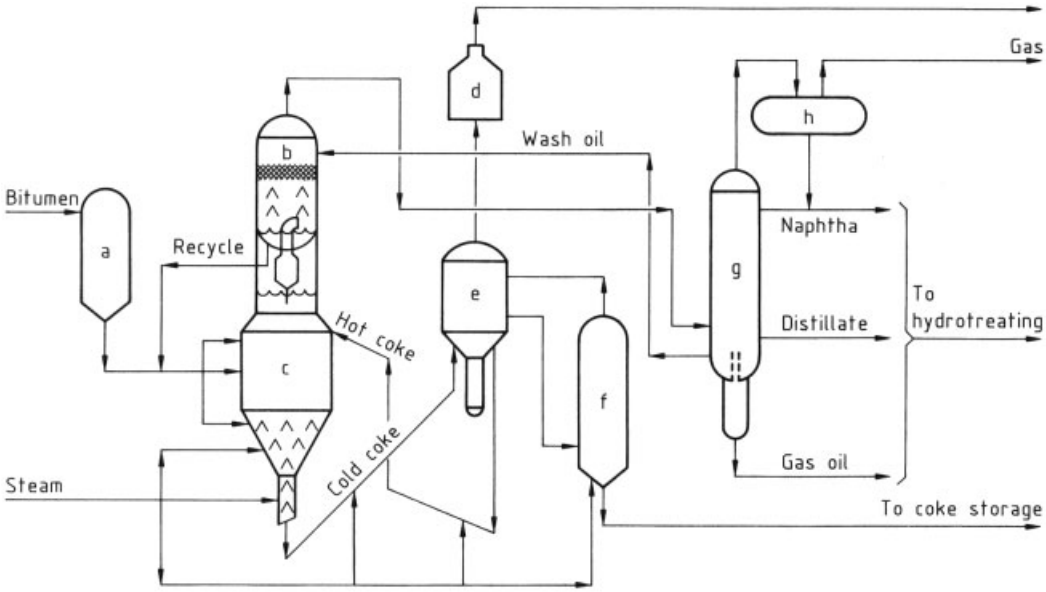

Figure 18. Syncrude fluid coking
a) Feed drum; b) Scrubber; c) Reactor; d) CO boiler; e) Coker burner; f) Elutriator; g) Fractionator; h) Condenser

The fractionator separates the vapor products from the reactor into gas, naphtha, distillate, and gas oil fractions. The gas and naphtha fractions are further fractionated in the gas recovery system. The gas is reused in the process for fuel. The naphtha and distillate are hydrotreated to reduce sulfur, nitrogen, and aromatic concentrations. The gas oil is hydrotreated to reduce sulfur, nitrogen, and metals and to add hydrogen to hydrogen-deficient aromatic compounds. Hydrotreated product is blended to meet downstream refinery specifications.

Hydrocarbon liquid yield from Syncrude's fluid coker is approximately 81 vol %, and net coke and gas make are each ca. 10 wt %. Syncrude has two fluid cokers, each operating at ca. 16 700 m^3/d. Each fluid coker is equipped with a 5700 kg/s CO boiler.

When the Syncrude project was designed in the early 1970s, fluid coking was attractive because of its low capital and operating costs, the existing commercial experience, and the availability of waste heat and fuel gas for use in the extraction and utilities areas of the integrated operation. Disadvantages include relatively low liquid yield, high SO_2 emissions, and the excess coke byproduct which, for lack of markets, must be stored. Developments to minimize or eliminate the coke production have been considered, but have not been applied for economic reasons; these include Flexicoking and coke use as a feedstock, for hydrogen production, or fuel.

2.4.2. Hydrogen Addition Processes

Hydrogen addition processes are attractive, owing to the higher liquid yields, lower sulfur dioxide emissions, and the absence of unmarketable byproducts such as coke or residuum. Hydrocracking processes are designed to convert up to 95 wt % of the heavier hydrocarbons in bitumen (i.e., material boiling above 524 °C) to lower-boiling material, and to provide product yield > 95 vol %. Bitumen hydrocracking technology is applied commercially at the Syncrude project.

At the Syncrude project, an LC-Fining hydrocracker of capacity 8000 m^3/d was added to hydrocrack 20 % of Syncrude's bitumen feed upstream of cokers. The LC-Fining technology has ebullated bed reactors designed to convert > 60 % of the 524 °C + content of bitumen. The unconverted portion of the bitumen is routed to the fluid coker for further processing (see Section 2.4.1). This addition provided a high incremental yield with no additional sulfur dioxide emissions, and is the basis for significant addi-

tional conversion capacity expansion in the future.

3. In Situ Oil Sands Development

The vast majority of crude bitumen from Alberta's oil sands can be recovered only by in situ methods. While commercial in situ recovery has been the focus of greater attention in the past 15 years, considerable research has been going on since the mid-1950s.

In situ thermal processes utilize either methods in which steam is injected into the reservoir from the surface, or methods in which heat is produced in the reservoir by hydrocarbon combustion reactions. In the mid-1990s, there were three commercial thermal in situ projects in Alberta which inject steam into the reservoir, two projects employing the cyclic steam stimulation process, and one project utilizing a steam drive process. While the combustion process (see Section 5.1.1) appears promising, it has not been applied commercially. Methods in which electrical energy is used to heat the reservoir have also been tested. Operators have also pursued the development of primary production technology in areas previously thought not to be capable of such production. In situ projects typically produce only a bitumen product that is not upgraded in Alberta.

In the Peace River oil sands area, a steam drive in situ project is operated by Shell Canada. In the Cold Lake area, Imperial Oil Resources and Amoco Canada operate projects based on cyclic steam stimulation technology. Other commercial in situ projects are producing on primary. The experimental operations are in general considerably smaller than the commercial projects. An illustration of in situ experimental activity is provided in Figure 19.

Owing to significant variations in reservoir characteristics, extensive laboratory experimentation and field testing are required prior to commercial field development of in situ schemes. Experimental pilot schemes apply recovery processes to the geological and petrophysical properties of the reservoirs to identify the operating parameters and conditions that maximize productivity and economics.

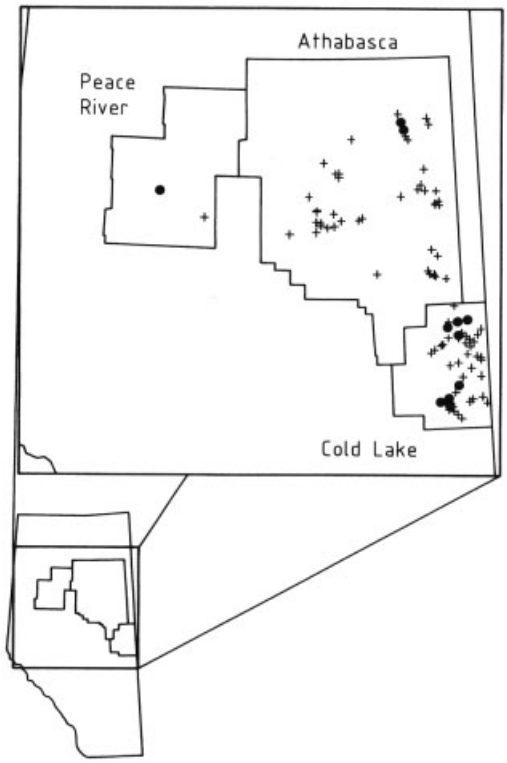

Figure 19. Alberta oil sands areas
● = Commercial; + = Experimental

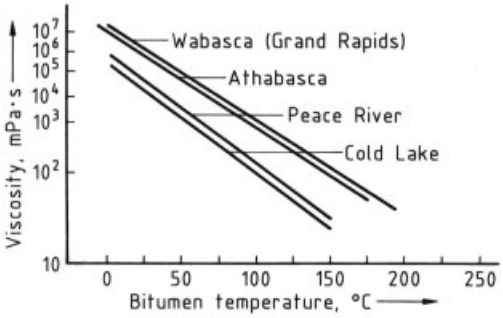

Figure 20. Bitumen viscosity – temperature relationship [47]

Most of Alberta's bitumen, in its natural state, has little or no mobility, owing to a lack of reservoir energy and high viscosity. In situ production requires heat input to the reservoir or the creation of heat in situ to reduce the viscosity of the bitumen (Fig. 20) and the reservoir flow resistance. The bitumen viscosity, and its response to elevated temperature, control the production rate in a recovery process. The application of heat also results in changes in capillary pressure, relative

permeability, and wettability. The injection of heat also results in thermal expansion of oil, increased evolution of dissolved gas, and changes in bitumen composition through thermal cracking and steam distillation.

3.1. Cyclic Steam Stimulation: Cold Lake Project

The cyclic steam stimulation (CSS) or "huff n puff" process widely used in the thermal in situ recovery of Alberta's bitumen, was discovered accidentally in 1959 in the Mene Grande field, Venezuela, when a steam injection well was backflowed to relieve reservoir pressure. The CSS process (Fig. 21) involves steam injection and fluid production in the same wellbore, conducted in a cyclic manner. Each cycle is comprised of four phases; steam injection, soaking, flowback, and pumping. Cycle steam volumes range from 4000 to 20 000 m^3 per well, and are injected over 15 – 75 d, depending on the type of operation. Soaking, for 0 – 30 d, is the period during which the well relaxes as it waits to be placed on production. The flowback period, up to 45 d, utilizes the reservoir energy to produce fluids. The pumping period, of 75 – 450 d, utilizes artificial lift to produce fluid until rates become uneconomic, at which time the cycle is repeated.

The CSS process is utilized extensively in the development of the Clearwater member in the Cold Lake oil sands area. One of the commercial projects, Imperial's Cold Lake project, is founded on a strong empirical basis with extensive piloting since 1964. The project has been commercially developed in eight phases, resulting in 1600 wells covering 26 km^2. Another two phases are planned to bring bitumen production up to 18 000 m^3/d (Fig. 22). Each rectangle represents a pad with 20 wells and area 0.32 km^2. Each pair of phases is served by a central plant, capable of daily processing 3000 m^3 of bitumen, generating 12 000 m^3 (cold water equiva-

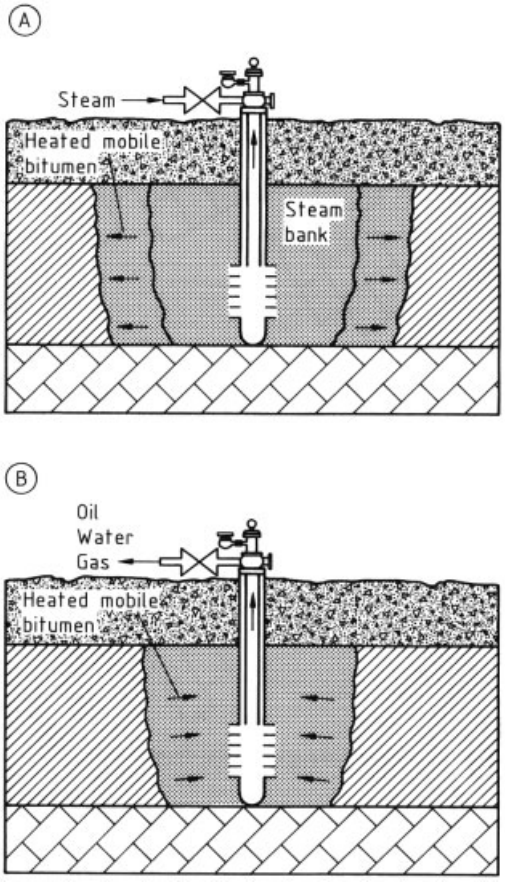

Figure 21. Cyclic steam stimulation
A) Steam injection; B) Bitumen production

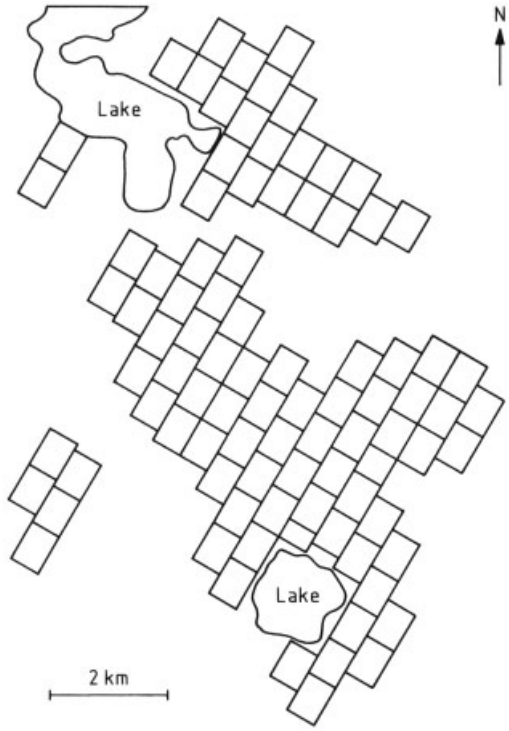

Figure 22. Cold Lake commercial development

lent) steam, and treating 9000 m^3 produced water for reuse.

To reduce surface disturbance, wells are drilled from pads (Fig. 23). The wells are directionally drilled to the reservoir. The figure shows the pattern aspect ratio of 1.7 (assumed length/width of rectangular drainage area), the well spacing, and the preferential pad orientation. Pad orientation is based on preferential flow directions observed during pilot testing, which is a function of the anisotropic permeability in the reservoir. Figure 24 shows a typical pad.

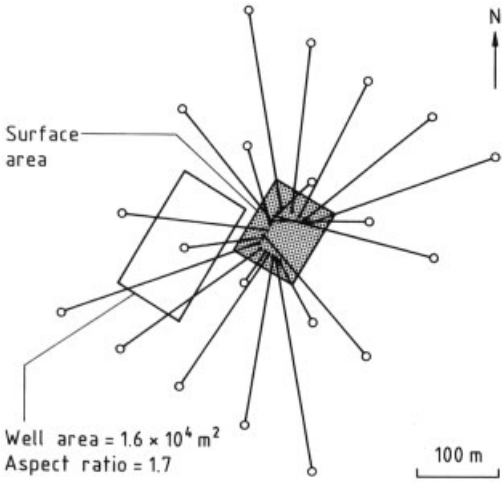

Figure 23. Well trajectories for a typical cyclic steam stimulation pad

Figure 24. Imperial Oil's Cold Lake project, typical well pad

Initially, the CSS process is managed as a single-well process, however in subsequent cycles the process must be managed as a multiwell process, requiring fieldwide synchronization of steaming with overlap steaming of adjacent rows to control unwanted communication [48]. Figure 25 illustrates the configuration of steaming, soaking, and shut-in wells at a given time. As cycle number increases, communication widens and, after about eight cycles, the process becomes uneconomic.

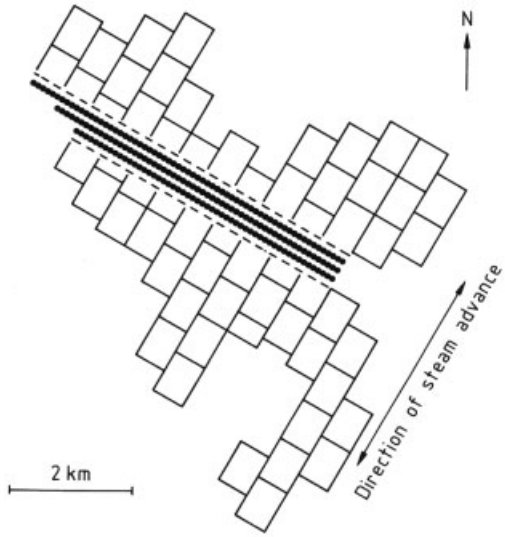

Figure 25. Overlap steaming/soaking of adjacent rows
• • • • • • • = Wells receiving steam; – – – = Wells seaking or shut-in

To allow sufficient volumes of steam injection, first-cycle injectivity is achieved by injecting above fracture pressure. In subsequent cycles, initial injectivity can be achieved below fracture pressure; however, as the depleted vapor space is compressed, formation failure is initiated again. Later cycles require increased steam volumes to initiate formation failure. Productivity decreases with increasing cycle number, and once the steam/oil ratio in a given cycle becomes uneconomic the process is discontinued. A typical well has a steam/oil ratio of 3 – 4 and overall recovery ca. 20 %.

3.2. Steam Drive: Peace River Project

Another commercial in situ site near Peace River has a steam drive that heats the reservoir via continuous injection of steam through a pattern of injection and production wells. Steam drive or steam floods require hydraulic continuity between wells, and may be preceded by a cyclic steam process to generate this continuity.

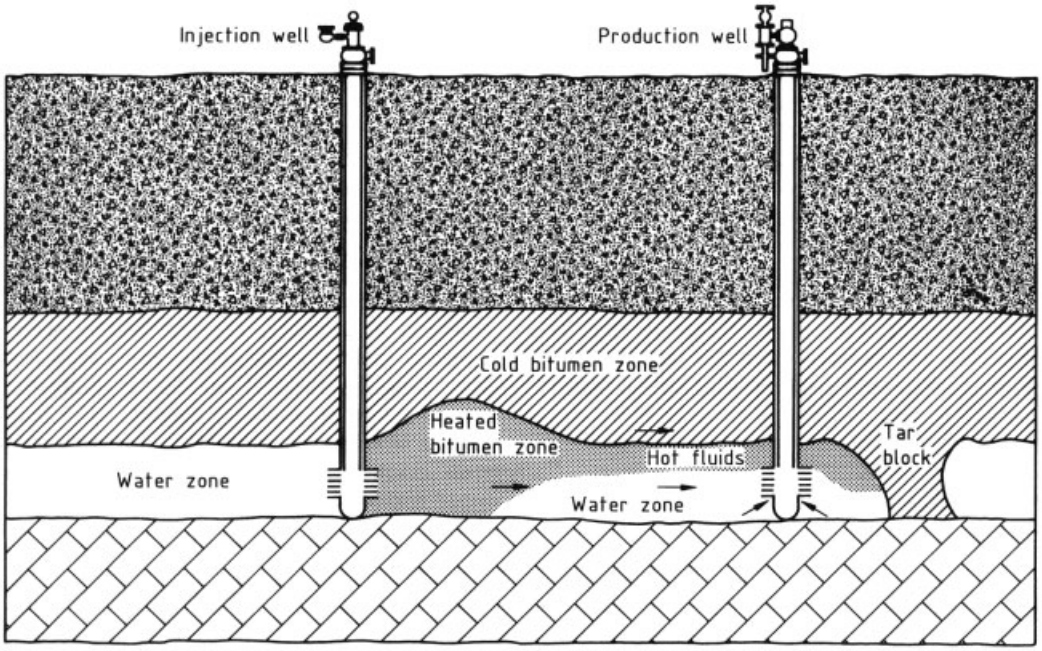

Figure 26. Shell Canada's steam drive process

Steam drive processes usually operate under less severe conditions than cyclic stimulation processes, and may be conducted in continuous or pulsed mode.

Shell Canada developed the pressure cycle steam drive (PCSD) process which is applied in the Peace River area [49,50]. In this process (Fig. 26), steam is injected into the reservoir without fracturing via an underlying water-saturated sand which is permeable to the steam. This results in a radial heated zone at the base of the reservoir, which mobilizes the overlying bitumen. The PCSD process alternates periods of high steam injection with periods of high production, which causes cycles of pressure buildup and blowdown in the reservoir.

The Shell project, which commenced in 1986, encompasses the operations at a pilot and at 213 wells directionally drilled from eight pads in four well clusters (Fig. 27). Each well cluster comprises 13 injectors and 40 producers arranged in 13 patterns; each pattern has 6 producers surrounding each injector in a hexagonal array. Operations began with 1 – 3 steam soak cycles at each well, followed by the PCSD. Operations were later modified to a continuous low-pressure steam drive process. Performance in 1992 was stabilized at 1600 m^3/d, at injection pressure 3200 kPa. The recovery efficiency to 1992 was 20 % and is anticipated to exceed 40 % with a steam/oil ratio of 3 – 4.

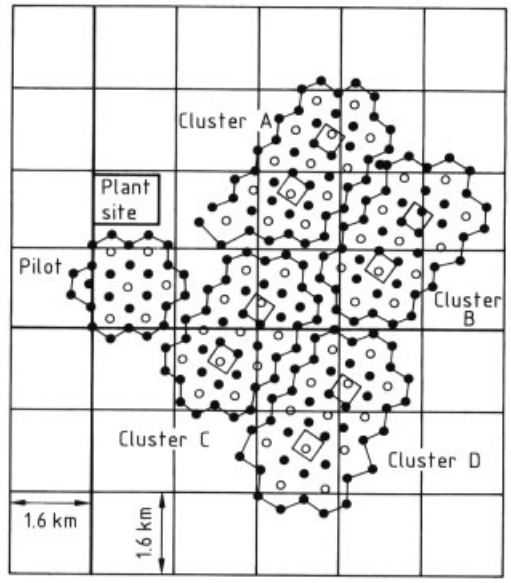

Figure 27. Shell Canada's Peace River complex
● = Production well bottom hole location; ○ = Injection well bottom hole location; □ = Pad

3.3. Primary Crude Bitumen Production

While crude bitumen recovery from oil sands generally requires thermal stimulation, primary production (pumping wells at natural reservoir temperature) is feasible in some areas. In the southern portion of the Cold Lake area there is a transition from deposits of crude bitumen to conventional oil pools. Primary production is enhanced by allowing sand to be produced. The sand production causes wormholes or channels to form in the reservoir which improve bitumen recovery through enhanced drainage radius, grain movement, gas bubble expansion, and continuous pore deblocking [51]. Although the primary reserves estimates are low (2 % for the Cummings sands and 0.1 % for the other Mannville sands) the primary production is significant. In 1992, primary production accounted for 18 % (3700 m^3/d) of the bitumen produced from Alberta's oil sands. Progressive cavity rotary pumps are more suited to pumping fluids containing significant amounts of sand (as high as 40 vol %), and in many cases are replacing the more conventional reciprocating action beam pumps.

3.4. Technical Issues Affecting In Situ Recovery

3.4.1. Casing Design

Thermal casing design must consider possible failure in tension due to thermal stresses upon heating and subsequent cooling. Well casing is designed for the maximum expected temperature and the minimum expected temperature after steaming.

Casing failures were a common problem in the early 1960s when cyclic steaming was first attempted. Most failures occurred at a coupling, usually in the upper part of the wellbore. Casing failures seldom occur now because stronger casing and better cementing practices are employed. Casing and cement integrity are important for environmental protection and production maintenance.

Figure 28 illustrates a typical thermal casing design. A thermal grade casing, along with specially prepared thermal cement, is used with oversized buttress-thread collars to withstand

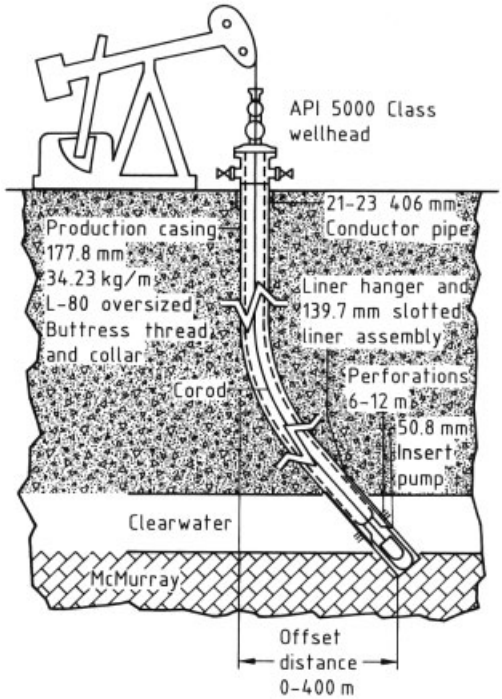

Figure 28. Basic well configuration

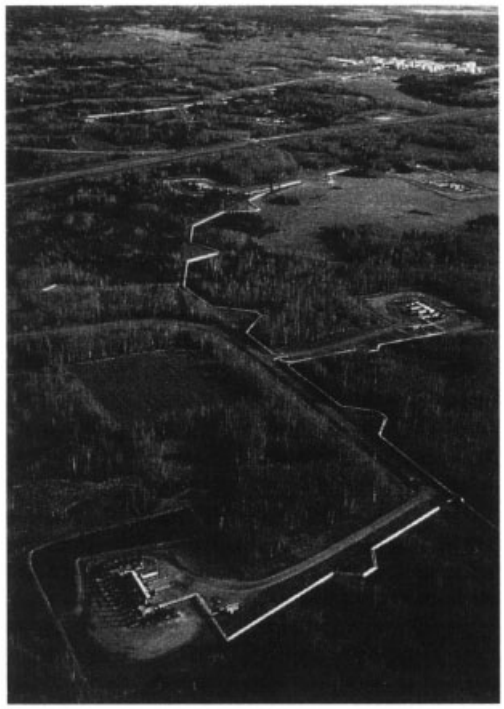

Figure 29. Imperial Oil's Cold Lake project
Well pad in foreground; pipelines (steam and produced fluids) to central plant in background

thermal stresses and to resist hydrogen sulfide stress cracking. A special spray coating and thread compound are used to ensure a tight seal at the joints.

3.4.2. Gathering and Distribution System

A pipeline gathering system collects production from the satellites for processing in the central plant. Both the production and steam gathering systems are pile-mounted, above-ground pipeline systems with expansion loops (Fig. 29, see previous page). To minimize surface disturbance, the road and power distribution grid normally follow the same route as the gathering and distribution pipeline system.

3.4.3. Production Treating

Field production enters the treating plant at the inlet flowsplitter where free gas is removed (Fig. 30). The fluid stream (bitumen and water) is routed through inlet coolers, to lower the fluid temperature, and allow treating to continue. The fluid proceeds through an inlet separator in which the free water and bitumen are separated by gravity. As the bitumen is cooled, the fluid viscosity tends to increase. At this point diluent is added to reduce viscosity and aid the separation process, and forms up to 30 vol % of the mixed diluent and bitumen stream (dilbit). The remaining water is separated from the dilbit stream in electrostatic dehydrators. The dilbit and oily produced water are then cooled to room temperature in glycol heat exchangers. The dewatered dilbit is degassed in a vertical boot degasser and sent to the dilbit sales tanks. Final product viscosity is controlled by blending diluent in the suction line to the pipeline pump station.

Cooled produced water is treated and reused for the generation of steam in once-through steam generators (Fig. 31). Oily produced water from the production treating facility enters the water treatment plant through a skim tank in which gross oil and suspended solids are removed. Further oil removal takes place in an induced gas flotation (IGF) unit. Effluent from the IGF flows to the sand filters which remove a portion of the remaining oil and suspended solids, and protect the downstream processes from slugs of oil which may result from an IGF upset. In the hot lime softener, the inlet stream is mixed with lime, magnesium oxide, and coagulant aid. These react with alkali and silica in the water to form insoluble precipitates which are purged to a sludge pit. The hot lime softener effluent then passes through filters to remove any suspended solids. Ion exchange softeners, utilizing a weak acidification resin, are used to remove the residual hardness. Effluent from the ion exchangers flows directly to the steam generators.

Oil recovered through skimming the produced water is collected and recycled through the production treatment process. Emulsions of oil and water that prove difficult to separate are handled by the recycle treater.

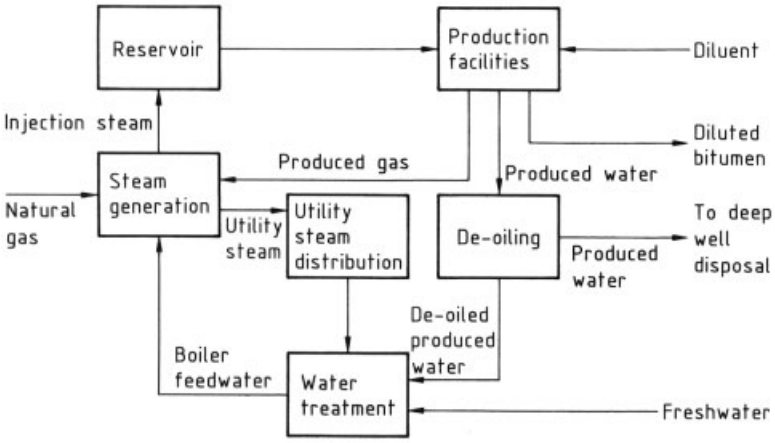

Figure 30. Bitumen production by in situ schemes

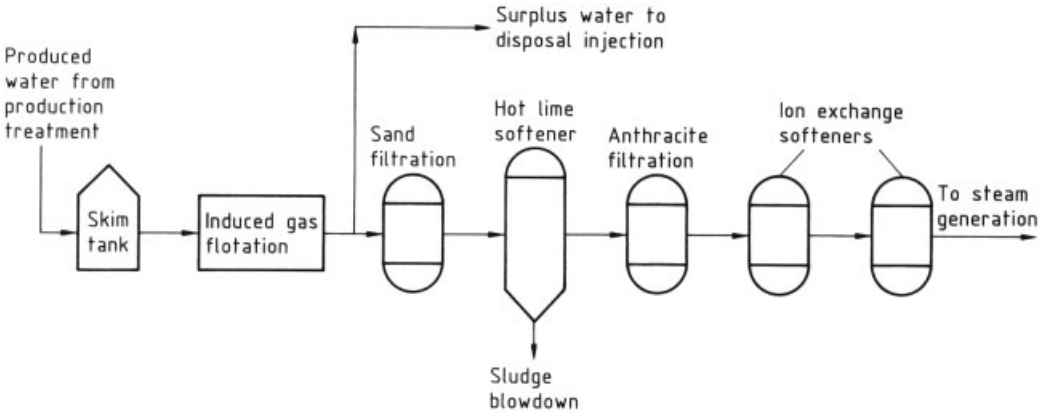

Figure 31. Produced water treatment

3.4.4. Oily Waste Disposal

Oily waste typically consists of sands, sludges, and slop oil containing bitumen, generated during oil sands production. Oily waste is disposed of by spreading an oil – sand – gravel mixture onto minor roads (as a dust supressor), by water washing and disposal of "clean sand" in an industrial landfill, landfarming, and by subsurface disposal into a salt cavern (Fig. 32). A salt cavern is created by drilling a well and washing the salt formation, using produced water only. The resulting brine is disposed into an existing water disposal well. This option is expected to have a long lifetime and to be environmentally and economically sound.

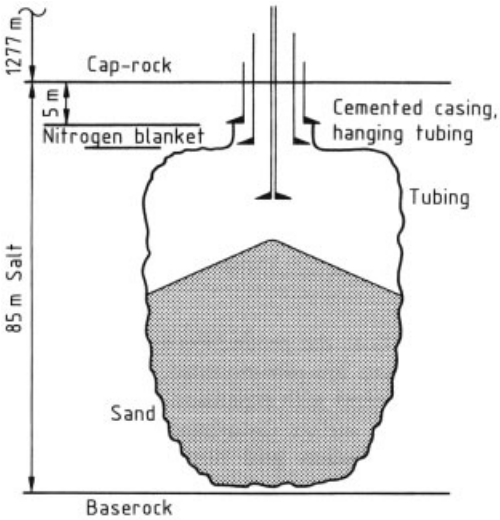

Figure 32. Sand disposal cavern

4. Economic Development

Alberta's early attempts to develop the oil sands was very sporadic with little or no commercial success. Pilot projects in the 1930s were doomed for a variety of reasons: limited capital, inadequate extraction technology, high development cost, and geographic or climatic impediments. All of these issues had to be addressed and overcome in order to exploit this vast resource; many remain today.

4.1. Mining Projects

Alberta's major oil sands deposits for mining are all located north of 56° lattitude. The area is typically forested and much of it is covered with muskeg, a permanently water-saturated soil cover. Northern Alberta is prone to extreme climatic conditions with long and extremely cold winters (average January temperature − 20 °C, sometimes − 50 °C) and short, hot, humid summers. The infrastructure necessary for industrial development of the area remained undeveloped until the first commercial project in the 1960s. Until then, the region was sparsely populated by a variety of remote indigenous settlements and a small community of 1200 at Fort McMurray.

Development of the Great Canadian Oil Sands Project (now Suncor) in the 1960s was remarkable in many respects, particularly for the long-term vision adopted by the company to scale up and commercialize unproven technology, as well as the corporate risks taken to

assure its success. At that time, it was necessary to produce a partially refined product (synthetic crude) to capture a market share. The initial investment in the integrated Suncor project was approximately $\$250\times10^6$ for synthetic crude oil production of 4930 m^3/d. Comparable crude oil typically sold for $\$16/m^3$ in 1967.

The Syncrude Canada development during the 1970s represented further advances in refining the technology and developing the infrastructure to allow the eventual exploitation of oil sands on a larger scale. It was developed when considerable conventional production was competing for a limited oil market, and crude oil prices in Canada were controlled to less than half their market value. Completed in 1978, the integrated Syncrude project promised to deliver some 20 500 m^3 synthetic crude oil per day with an initial capital cost of $\$2.3\times10^9$. To minimize the corporate risks the project was financed by a partnership of large oil companies and the Alberta and Canadian Governments. Corporate returns were shared through a negotiated fiscal package, and a favorable royalty stream in the early years of production. Alberta's oil sands remain a high-cost source of supply.

Mining of oil sands and production of synthetic crude oil is well established in Alberta. Two large commercial projects deliver ca. 17 % of Canada's oil production (38 000 m^3/d). The large reserve base, proven technology, and an established infrastructure all have a positive impact on the commercial viability of oil sands projects, and stable prices should promote further development in the future.

4.2. Economic Viability

Mining Projects. Development of reserves is hindered by high capital and operating costs and long lead time between project approval and initial production. In 1993, a typical oil sands mining and upgrading project with 13 000 m^3/d production was estimated to require investment of $\$4-5\times10^9$ over a 6 – 8 year construction period.

New technology and increasing operating efficiency have steadily reduced the cost of bitumen production at Syncrude and Suncor, and costs are close to being competitive with new conventional crude production in Alberta. The unit operating costs of commercial projects are about $\$75/m^3$, and the projects generate a modest return on investment.

The supply cost of a grassroots project has been estimated at ca. $\$150/m^3$ (1990 prices). However, significant future cost reductions are expected with the adoption of new mining and extraction technologies.

In Situ Projects. In situ recovery of oil sands was tested in the mid-1950s by Shell Canada, Pan American, Mobil Oil, and Petrofina Canada. Early work on steam stimulations and in situ combustion was largely experimental. Higher oil prices in the early 1980s prompted significantly more activity by a variety of firms and technological improvements led to commercial developments.

Although in situ developments are only marginally economic, they offer a number of environmental advantages over mining projects, e.g., limited surface disturbance and no tailings impoundment. Bitumen produced by in situ projects needs to be diluted for transportation by pipeline; it has found a ready market throughout North America.

In situ bitumen production in Alberta was ca. 21 000 m^3/d in 1992, including production from nine commercial projects employing thermal and/or primary recovery techniques, and a variety of experimental projects. In the Peace River area, a steam drive in situ project is operated; in the Cold Lake area, two projects are based on cyclic steam stimulation technology. In 1993, there were 6 other commercial projects producing on primary, while 12 experimental operations were active. Experimental projects are considerably smaller than the commercial projects, and typically operate for less than 5 years, generally producing less than 500 m^3/d.

Estimates of threshold bitumen supply costs vary, depending on the source and technology (Table 5, see next page), but are typically $\$50-75/m^3$ [52]. Additional refining costs need to be incurred in order to upgrade bitumen to higher value synthetic crude oil produced at integrated projects. Upgrading is estimated to cost $\$40-60/m^3$ [54].

Oil sands production is clearly a high-cost source of crude oil at present, and the tenuous commercial viability is likely to require ongoing royalty and fiscal concessions to make

projects commercially attractive. Nevertheless, progressive technological improvements and refinements in operating techniques have shown that Alberta's vast reserves are on the frontier of major development.

Table 5. Threshold in situ supply cost for bitumen, 1990 [53]

Area	Cost, $/m^3
Athabasca	
Steam-assisted gravity drainage	65
Cold Lake	
Cyclic steam (superior deposits)	65
Primary recovery	50
Peace River	
Pressure cycle steam drive (superior deposits)	75

5. Challenges Past and Present

5.1. Technology

Technology selection to develop the Suncor project in the 1960s was a major challenge, owing to the limited information available on mining oil sands, extracting the bitumen, and upgrading it. Although experimentation and field testing had occurred, much more work was necessary to modify and apply the available technology. When bucketwheel excavators, used successfully for nearly half a century in the large-scale German lignite mining operations, were applied to oil sands mining in Canadian winter conditions bucketwheel teeth had to be replaced every 18 h. Improved tooth design, blasting and loosening of the oil sands, and revised procedures for operating the bucketwheels during winter eventually solved the problem.

The management of tailings from the extraction of bitumen from oil sands was another huge challenge. Tailings volumes at oil sands mines are enormous compared with general mining experience. The impoundment plan initially prepared by Suncor was inadequate because of the slow fines settling rate. Solutions were found where tailings dykes used the coarse sand from the tailings, and huge impoundments were constructed; however, the challenge of increasing the fine tailings settling rate has yet to be solved.

During the 1970s, while Suncor's focus was on operational problems, Syncrude concentrated on design and construction. Although considerable testing occurred prior to start-up, new challenges arose which required innovations, such as mine depressurization and dewatering ahead of the mine highwall, slope monitoring and stabilization, and improvements to mining and extraction equipment reliability. Syncrude's work on conveyer belting, along with developments by belt manufacturers, greatly improved the operational efficiency of conveyers.

Parallel with this successful start-up at Syncrude, research and development continued on other technology. In bitumen extraction, the hot water process was improved through design modifications to the secondary recovery circuit and addition of the tailings oil recovery. In froth treatment, inclined plate settling has been found to be an effective alternative to the more complex centrifuges, and was added to increase production capability. A naphtha recovery system has also been added which significantly reduces hydrocarbon loss to tailings.

The application of huge fluid cokers at Syncrude has been a remarkable success, bearing in mind the considerable scale-up from any previous installation, and the increase in throughput from the design rate of 11 600 m^3/d to the current rate of ca. 17 500 m^3/d. The addition of a hydrocracker upstream of the fluid coker is another major innovation by Syncrude, with benefits of increased liquid yield and reduced sulfur emissions.

Challenges and innovations in the in situ recovery area have also been dramatic. Prior to commercial operations in the early 1980s at Imperial Oil's Cold Lake project, various companies conducted extensive in situ tests designed to demonstrate a wide range of in situ recovery techniques. The commercial success of Imperial and Shell is due to the development of steam stimulation processes. Other developments of note are in the area of cluster drilling, the use of slant drilling, horizontal drilling, thermal casing and coupling, and thermal cement. Along with these came a host of production, injection, metering, and control systems unique to the high-temperature, multiphase flows found at in situ operations. The large volumes of water required and produced at steam stimulation schemes led to the development of water reuse technology. The industry developed water treatment and recycle technology which meets environmental and water conservation objectives, and provides energy and cost savings.

Technology development is underway, with emphasis on cost reduction and increased reliability, service factor and product yield. For mining operations, the focus is on materials handling and transportation techniques; for bitumen production, the focus is on a lower-temperature extraction process which fits better with the mining technique. In the upgrading area, the focus is on higher yield and better quality. In the in situ area, combustion techniques have been developed and considerable success has been achieved with the application of horizontal wells which operate at lower pressures and capitalize on gravity drainage.

5.1.1. Mining and Bitumen Production Technology

Mining developments have resulted in improvements of truck and shovel capacity and oil sands crushing technology [55]. Mining shovel capacity has increased from the 11.5 m^3 hydraulic shovel used by Suncor in 1980 to a 46 m^3 shovel in 1993. Over the same period Suncor and Syncrude have increased truck capacity from 155 to 218 t. Oil sands crushing technology was tested in the early 1980s when a prototype single-roll crusher was tested. This led to the installation of a commercial-scale crusher by Syncrude in the mid-1980s and the installation of a high-speed double-roll crusher in 1991. In 1993, Suncor installed two fixed-plant double-roll crushers rated at 7500 t/h. Additionally, a low-speed mobile mineral sizer was operated at Suncor in 1990.

As a result of these developments, a truck and shovel mining system is in use at Suncor as a replacement for the early bucketwheel system.

Hydraulic Transport. Another major challenge of minable oil sands development is efficient materials handling. Besides the developments in truck and shovel mining, combined with oil sands crushing technology, operators have investigated alternatives, including dredge mining and hydraulic transport of overburden and oil sands. In the mid-1980s, Syncrude and AOSTRA conducted pilot and prototype tests of overburden and oil sands mining, but this technology has not been applied commercially.

In 1988, Syncrude launched a program to develop technology for hydraulic transport of oil sands [56]. The program resulted in the development of an oil sands slurry mixing device, a cyclofeeder, and a process to condition oil sands in a pipeline. This technology, tested at pilot scale in conjunction with a 50 °C bitumen separation process, was found to be feasible. Late in 1993, a commercial test was initiated at an oil sands feed capacity of 1250 kg/s.

The development of technology to take dry oil sand, slurry it, and pipeline that slurry so that it is conditioned and allows bitumen to be separated, is a major development. This technology, combined with truck and shovel mining, would reduce materials handling costs and energy consumption, and allow remote oil sands reserves to be exploited.

Alternative Bitumen Extraction. Commercial operation of the hot water extraction process has shown the need to reduce or eliminate tailings, to increase process energy efficiency, and to reduce water requirement. Technology development in the late 1970s and early 1980s has been assessed in a study commissioned by the ERCB [39]. While the AOSTRA/UMATAC retort process [57] was considered to offer the best long-term commercial potential, its development depended on successful scale-up and proven field demonstration. The commercial development of the Syncrude modified aqueous extraction process was attractive, owing to the relatively low risk involved in scale-up. Unfortunately, tailings disposal requirements for the alternative processes were largely unproven by 1984, and further investigations are ongoing. The technology development effort, directed by the ERCB's thrust toward improvement in energy efficiency, tailings management, and environmental acceptability, is focused largely on aqueous processes.

Minerals Recovery. The presence of minerals in oil sands and byproducts of processing has been noted, and recovery methods explored. Titanium and zirconium are found in the surface-minable oil sands, and are concentrated in centrifuge tailings streams. Preliminary analysis indicate that tailings would provide sufficient feedstock for large-scale production [52]. Tests indicate that recovery is technically feasible. Tests

on fine tailings have also indicated the potential for alumina. Minerals may be recovered from coke, fly ash, and spent catalyst; e.g., vanadium in coke, fly ash, and catalyst from the Syncrude project are estimated at 282 kg/h, 18 kg/h, and 45 kg/h, respectively [58].

5.1.2. In Situ Recovery Technology

In Situ Combustion. In situ combustion, also known as fireflooding, is an in situ thermal process that has been tested, but not developed commercially [59]. The combustion process generates heat in situ via hydrocarbon combustion reactions.

The process is initiated by igniting the reservoir bitumen. Spontaneous ignition may occur, or ignition may have to be induced, depending on the characteristics of the bitumen. Air, or air enriched with oxygen, is injected continuously to sustain and propagate the firefront away from the injection well toward the production wells. When the firefront moves in the same direction as the injected air the process is "forward" combustion. If air is injected continuously until it reaches the producer and is then ignited, the combustion front moves backward from the producer, "reverse" combustion. The efficiency of the dry combustion process can be improved by alternately injecting water and air to create the wet combustion process. The advantages of wet combustion are that water, because of its greater heat capacity, is better at scavenging heat from the burned zone than air, and that the higher viscosity of the water phase results in a more favorable distribution of the injected air, and a greater sweep efficiency of the hydrocarbons.

The most common method is forward combustion (Fig. 33) [60]. Moving from the injector to the producer, injected air first encounters the burned zone. The combustion front has passed through this zone, leaving only hot, clean sand with air-saturated pores. The air passing through this zone is preheated as it progresses to the combustion front. The heated air passes through the combustion zone itself where it oxidizes the coke left from the thermal cracking of the oil in-place in the cracking/vaporization zone, temperatures are sufficiently high to crack the oil thermally, and to vaporize any hydrocarbons and water. Just beyond the thermal cracking zone, combustion gases, steam, and vaporized light hydrocarbons (steam zone) are transported forward, where vaporized water and hydrocarbons begin to condense (altered saturation zone). The last zone is the native reservoir which has not yet been affected by the advancing fire. The residual gases flow through this zone toward the production wells.

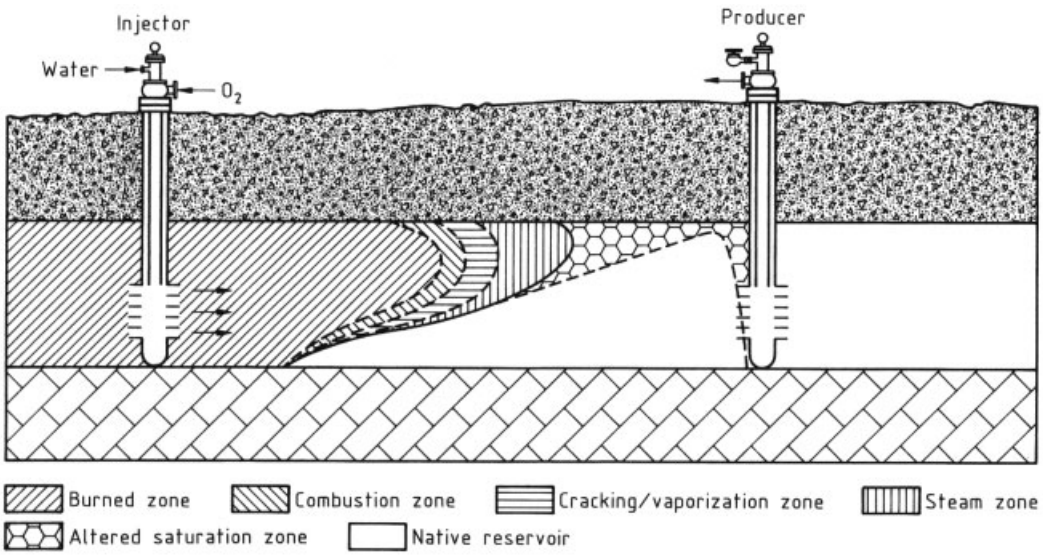

Figure 33. Forward in situ combustion

Horizontal Wells for Thermal Recovery. The in situ recovery of bitumen by horizontal wells presents new opportunities for efficient recovery and improved project economics. Commercial bitumen recovery involves close well spacing and in such cases horizontal wells can be expected to be most effective.

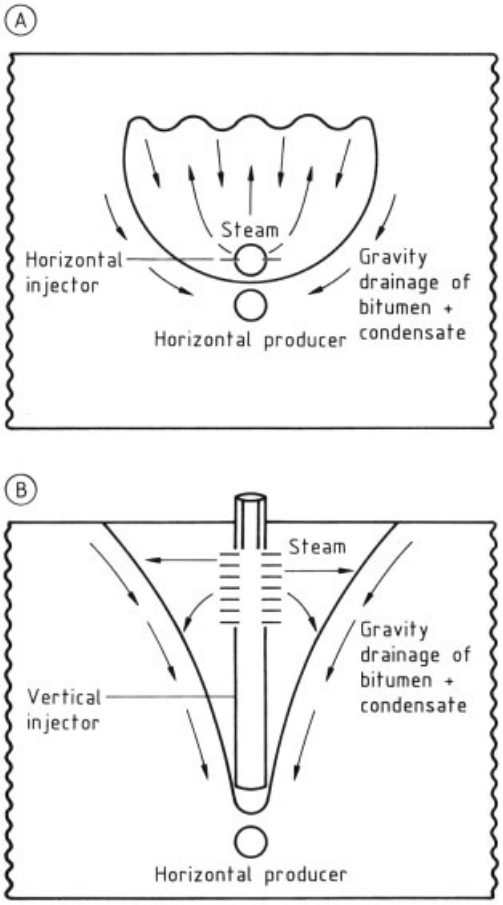

Figure 34. Steam-assisted gravity drainage
A) Upward/sideways movement of steam; B) Downward/sideways movement of steam

Additionally, horizontal wells improve productivity without coning since the pressure differential is distributed over a larger section of the reservoir [61]. Steam-assisted gravity drainage (SAGD), is a form of the steamflood recovery process that utilizes the effects of gravity and the improved reservoir contact achievable with horizontal wells. Steam is injected continuously, by a vertical or horizontal injector, above the horizontal producer, located at the base of the reservoir. The steam moves upward and sideways (Fig. 34 A) or downward and sideways (Fig. 34 B), and forms a steam chamber [62]. As the steam contacts the interface, it heats the bitumen near the interface. Owing to its reduced viscosity, the bitumen at the interface drains, by gravity, toward the horizontal producer. Although the rate of drainage at any point along the horizontal producer is relatively low, an overall high recovery rate can be achieved by utilizing a long horizontal production well.

Steam-assisted gravity drainage with underground access (SAGDUA) (Fig. 35, see next page) is also based on the SAGD process, but these horizontal wells are drilled from a mine shaft and tunnel system in the limestone underlying the bitumen-saturated zone [63]. Drilling costs are reduced and, by injecting steam directly from the mine tunnel, surface and wellbore heat losses are minimized, improving project economics. Highly promising results based on a full-scale field demonstration project indicate that commercial development of the Athabasca bitumen deposits too deeply buried for surface mining is becoming possible through the application of this process.

The heated annulus steam drive (HASD) process (Fig. 36, see next page) is a form of the steam drive process that circulates steam through a horizontal well [64]. This creates a zone of reduced bitumen viscosity, between an injector and producer. The horizontal well is not perforated so there is no fluid communication between the steam in the horizontal conduit and the formation. The zone of reduced bitumen viscosity results in increased bitumen mobility, and allows steamflooding to occur along the horizontal conduit toward the producing well. The injection of drive steam is maintained below fracture pressure so that control of the communication path is not lost.

5.1.3. Bitumen Upgrading Technology

Upgrading improvements for oil sands are of interest because ca. 50 % of the cost of producing synthetic crude oil is estimated to be due to upgrading. This, combined with the need for improved product quality to meet marketing and environmental requirements, has resulted

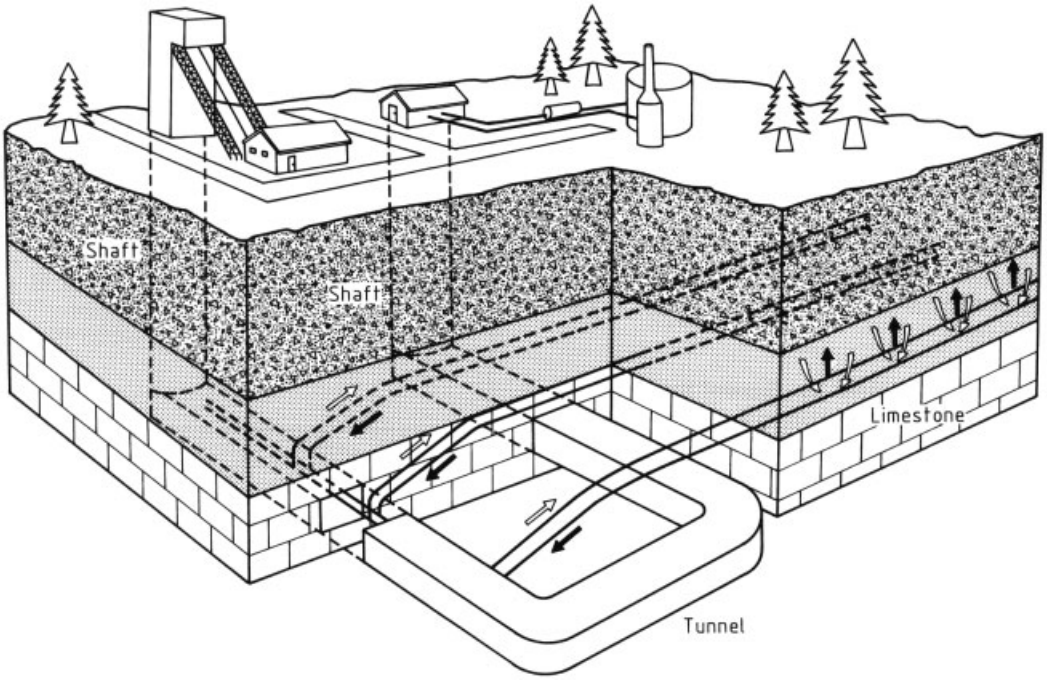

Figure 35. AOSTRA's field pilot of the SAGDUA process

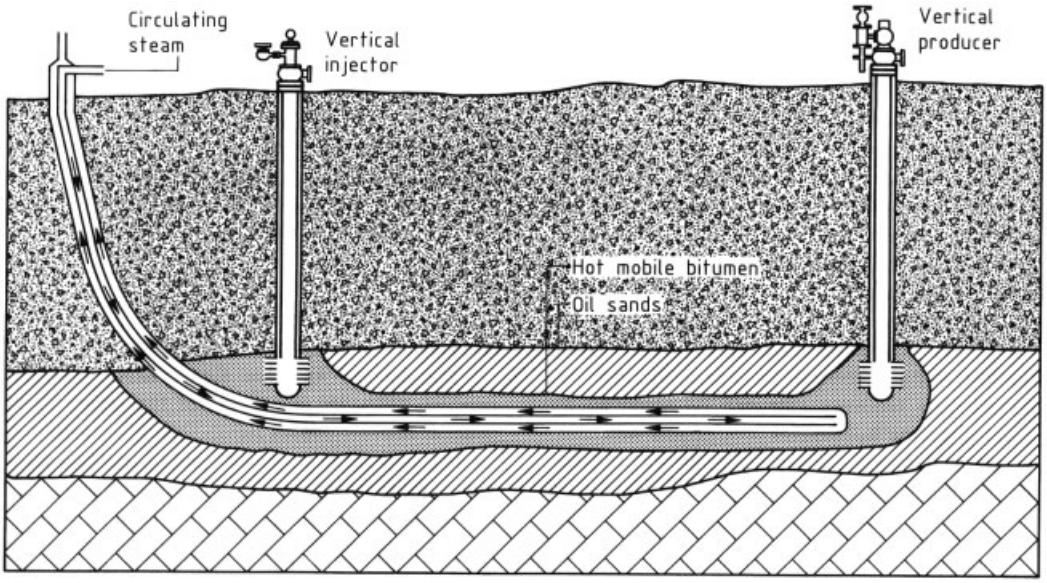

Figure 36. Horizontal configuration of HASD process

in considerable interest and activity on bitumen upgrading and distillate product hydrotreating. Technologies considered include visbreaking, coking, and catalytic hydrotreating and hydrocracking. While progress has been made in visbreaking, coking, and fixed-bed hydrotreating technologies, the major emphasis is on catalytic hydrocracking [65,66]. Although fixed-bed hydrodesulfurization technology and H-Oil and LC-Fining technologies have been available

commercially for some time in a low-conversion mode, these have not been designed and operated commercially in the high-conversion mode. Slurry phase hydrocracking processes such as Veba Combi cracking (VCC), CANMET, and MICROCAT RC, have been extensively tested on a demonstration scale.

Slurry phase hydrocracking technology is based on German coal hydrogenation technology. The process involves operation at $\geq 450\,^{\circ}C$ and 27 – 69 MPa, to crack residuum in the presence of a finely divided powder, which can be an additive or a catalyst. The powder serves both as a catalyst and as a site where coke can be deposited rather than depositing on process equipment. Such processes convert > 90 wt % of the heavier hydrocarbon, and achieve volumetric yields of ca. 105 %. The VCC process has been demonstrated on the 475 m^3/d scale in Bottrop, Germany, and the CANMET process has been demonstrated on the 715 m^3/d scale in Montreal, Quebec, Canada. MICROCAT RC is another additive residuum hydrocracking process tested on a 1.3 m^3/d pilot scale. The high conversions achieved with these technologies minimize air emissions and the production of unmarketable byproducts such as coke or pitch.

A unique feature of VCC technology is the integration of a fixed-bed hydrotreater with the high-pressure hydrocracking process. This minimizes costs in current operations associated with the pressure and temperature reduction of products of coking or hydrocracking, and reduces capital and operating costs.

5.2. Environmental Issues

Environmental issues facing oil sands development are primarily solids waste management, protection and conservation of potable water resources, reducing air emissions, and land reclamation. These issues are driven by more stringent regulatory standards and public expectations.

Fine tailings management and reclamation are the greatest challenges for the minable oil sands sector. Extensive study has identified the fundamental properties of fine tailings which give it an intractable gel-like structure and toxic character [67]. This has indicated ways to reduce fine tailings production, to treat fine tailings, and to reduce their bulk. The fine particles of size 0.1 – 1.0 µm in the presence of organic salts produced in the hot water process form the gel, which hinders water mobility. Reduced fines dispersion in a gentler oil sands conditioning process and reduced organic salt generation, by eliminating sodium hydroxide, should reduce fine tailings formation. Alternative extraction processes based on these ideas are being developed.

Reclamation techniques have been proposed for fine tailings, but have not been adopted. The principal options are to produce a dry, nontoxic, trafficable material or to cap fine tailings with water and allow then to detoxify and consolidate over an extended period. In both cases, the tailings would be contained in mined-out pits. Testing has indicated both methods have potential, but further work is required to confirm feasibility and safety. It is expected that, through the development of improved bitumen extraction and/or fine tailings treatment technology, future projects may not produce fine tailings at all.

The existing commercial projects burn the high-sulfur coke byproduct of upgrading bitumen, and consequently have significant SO_2 emissions. This and other combustion sources also produce CO_2, NO_x, and particulates; CO_2 is also a byproduct of the production of hydrogen by steam methane reforming. Volatile organic carbons (VOCs) are another source of atmospheric emission of some concern, and various measures are contemplated to reduce their effects. Particulate control in coke combustion operations is achieved by electrostatic precipitators. Studies have indicated the potential to recover CO_2 for use in enhanced oil recovery and chemical manufacture.

The need for large quantities of freshwater in oil sands recovery schemes has required maximizing water recycle and, in areas of Alberta experiencing freshwater shortages, the development of alternative sources such as brackish water and freshwater from adjacent water basins. These are feasible, but at added cost.

The disposal of process wastewater is another factor in water use at oil sands schemes. At in situ operations wastewater is disposed of to deep formations. As yet, mining projects wastewater is contained on site; in the long-term, treatment and discharge to surface is likely.

The integrity of wells, and in particular the potential to contaminate potable water aquifers, is a concern. Development of pipe material, couplings, and cement to handle the thermal application has resulted in few failures.

5.3. Public Policy

Broad policy issues involving the oil sands industry typically center on fiscal regimes to foster development by private interests, and the establishment of public infrastructure to handle the influx of population to remote or sparsely populated parts of Alberta. The bitumen/synthetic crude industry is characterized by a highly skilled workforce. Commercial development has tended to call for very large unit operations, requiring capital investment in the order of $\$ 10^9$, and thousands of employees. Economic and social factors play a major role in the pace of development. Given the personnel and technical constraints that may apply, economic issues center largely on the rate of return on invested capital, after allowances for royalties and taxes. Social issues focus largely on infrastructure development and impacts on the social fabric. It is estimated that the cost of infrastructure for a new town requires an expenditure $> \$ 800 \times 10^6$.

The economic benefits to the public and private sectors from large oil sands developments are very significant and usually dominate negotiations of the agreement for a commercial project. The direct social benefits include improvement in the scope and quality of community infrastructure (hospitals, recreation facilities, and transportation systems) although the large transient workforce required to build a project may produce significant dislocation in the community during the construction period.

Current developments have shown that, if the projects are distributed among existing and new communities, long-term community impacts are manageable. Existing institutions find it difficult to accommodate the rapid growth; issues to be considered include:

1) The need for large and advanced municipal financing of infrastructure
2) Tax-sharing arrangements
3) Rights of the indigenous population
4) Training and recruitment of the indigenous population
5) Compensation and relocation of residents affected by the project
6) The availability and affordability of houses
7) Environmental monitoring, changes in land use and land reclamation
8) Provision of adequate social facilities
9) Provision of health care, social services, education, and law enforcement
10) Transportation routes and services

These factors, if not fully resolved in a timely fashion, can constrain the pace of development.

Large bitumen/synthetic oil projects require a permanent workforce ranging from 2100 for in situ schemes to > 3000 for mining projects. Many of these positions can be filled by the construction workers involved in developing the site, but special technical skills are also required which are normally acquired through formal secondary education.

Approximately 10 years is required from the project definition phase until plant start-up. Major components of this period are: the 1 – 2 years required to obtain regulatory agency and government approvals; ca. 1 – 11/2 years for process design and preparation of a detailed estimate; and a ca. 3-year detailed engineering period. The detailed engineering period overlaps plant project construction, which normally takes ca. 5 years. Start-up operations would likely be conducted over at least a 1-year interval.

A construction workforce of 5000 is required over a 5-year period, with a peak value of 8000 – 10 000. Following the construction stage, operating staff amount to 2500 – 3000 per project, and the expected result is a total population increase of 15 000 – 20 000 per operational plant. This can place a great demand on the infrastructure, and create great need for additional government and private servicing to provide the necessary schools, hospitals, roads, and other services.

While the resource base in Alberta's oil sands is immense, the financial risks and the logistics may place limits on the rate of synthetic oil supply in the foreseeable future.

6. References

1. K. Takamura: "Microscopic Structure of Athabasca Oil Sand," *Can. J. Chem. Eng.* **60**

(1982) 538.
2. O. P. Strausz: "Bitumen and Heavy Oil Chemistry," in L. G. Hepler, C. Hsi (eds.): *AOSTRA Technical Handbook on Oil Sands, Bitumens and Heavy Oils,* AOSTRA, Edmonton, AB, 1989, p. 35.
3. B. D. Ricketts: *Western Canada Sedimentary Basin, A Case History,* Canadian Society of Petroleum Geologists, Calgary, AB, 1989, pp. 3 – 8.
4. J. W. Porter, R. A. Price, R. G. McCrossan, *Philos. Trans. R. Soc. London* **305** (1982) 169 – 192.
5. *A Description and Reserve Estimate of the Oil Sands of Alberta,* Energy Resources Conservation Board Publications, Oil and Gas Conservation Board, Calgary, AB, 1963. H. J. Webber: "The Oil Sands of Alberta," *J. Can. Pet. Technol.* **6** (1967) no. 4, 146 – 149. Energy Resources Conservation Board: "Geology and Proved In-Place Reserves of the Cold Lake Oil Sands Deposits," *ERCB Report 73-L-Geol,* Calgary, AB, 1974. Energy Resources Conservation Board: "Geology and Proved In-Place Reserves of the Peace River Oil Sands Deposits," *ERCB Report 74-R,* Calgary, AB, 1974. Energy Resources Conservation Board: "Geology and Proved In-Place Reserves of the Wabasca Oil Sands Deposits," *ERCB Report 76-A,* Calgary, AB, 1976. C. P. Outtrim, R. G. Evans: "Alberta's Oil Sands Reserves and Their Evaluation," Annu. Tech. Meet. Pet. Soc. CIM (1977).Energy Resources Conservation Board: "Crude Bitumen Reserves Atlas," *ERCB ST 91-38,* Calgary, AB, 1990. Energy Resources Conservation Board: "Alberta Reserves of Crude, Oil, Oil Sands, Gas, Natural Gas Liquids, and Sulphur," *ERCB ST 93-18,* Calgary, AB, 1993.
6. D. Wightman, B. Rottenfusser, J. Kramers, R. Harrison: "Geology of the Alberta Oil Sands Deposits," in L. G. Hepler, C. Hsi (eds.): *AOSTRA Technical Handbook on Oil Sands, Bitumens and Heavy Oils,* AOSTRA Tech. Pub. Ser. no. 6, Edmonton, AB, 1989, pp. 1 – 9.
7. J. A. Masters: "Lower Cretaceous Oil and Gas in Western Canada," in J. A. Masters (ed.): Elmworth – Case Study of a Deep Basin Gas Field, *Mem. Am. Assoc. Pet. Geol.* **38** (1984) 1 – 33.
8. S. O. Moshier, D. W. Waples: "Quantative Evaluation of Lower Cretaceous Mannville Group as Source Rock for Alberta's Oil Sands," *AAPG Bull.* **69** (1985) 161 – 172.
9. M. J. Ranger, G. Pemberton: "Trapping Mechanism and Time of Oil Migration in the Athabasca Oil Sands Deposit," *Am. Assoc. Pet. Geol. Annu. Conf. Proc.* 1992, 108.
10. G. Deroo, T. G. Powell, B. Tissot, R. G. McCrossan: "The Origin and Migration of Petroleum in the Western Canadian Sedimentary Basin, Alberta," *Bull. Geol. Surv. Can.* **262** (1977) 114.
11. R. S. Strobl, L. P. Yuan, W. K. Muwais: "Reservoir Geological Modelling of the McMurray Formation, Alberta," *Bull. Can. Pet. Geol.* **39** (1991) no. 2, 225.
12. D. A. W. Keith et al.: "Resource Characterization of the McMurray/Wabiskaw Deposit in the Athabasca Central Region of Northeastern Alberta," AOSTRA Tech. Pub. Ser. no. 7, Edmonton, Alberta, 1990.
13. G. D. Mossop, J. W. Kramers, P. D. Flach, B. A. Rottenfusser: "Geology of Alberta's Oil Sands and Heavy Oil Deposits," *1st UNITAR Conference on Future of Heavy Crude Oils and Tar Sands,* 1981, 197.
14. J. R. MacGillivray et al.: "Resource Characterization of the McMurray/Wabiskaw Deposit in the Athabasca South Region of Northeastern Alberta," AOSTRA Tech. Pub. Ser. no. 8, Edmonton, AB, 1992.
15. R. S. Harrison: "The Bitumen – Bearing Paleozoic Carbonate Trend of Northern Alberta," *AAPG Stud. Geol.* **25** (1987) 319.
16. D. Marion, B. Rottenfusser: "Geology of the Peace River Giant Oil Sands Deposit, Alberta, Canada," *Proc. of Unconventional Hydrocarbon Sources Conference,* St. Petersburg 1992.
17. J. E. Barclay, F. F. Krause, R. I. Campbell, J. Utting: "Dynamic Casting and Growth Faulting: Dawson Creek Graben Complex, Carboniferous – Permian Peace River Embayment, Western Canada," *Bull. Can. Pet. Geol.* **38 A** (1990) 115 – 145.
18. S. C. O'Connell, G. R. Dix, J. E. Barclay: "The Origin, History, and Regional Structural Development of the Peace River Arch, Western Canada," *Bull. Can. Pet. Geol.* **38 A** (1990) 4 – 24.
19. B. M. Benyon, S. G. Pemberton: "Resource Characterization of the Lower Cretaceous Grand Rapids Formation: Reservoir Geology and Hydrocarbon Distribution, Cold Lake Oil Sands Area, AB, " *AOSTRA J. Res.* **8** (1992) 108 – 130.
20. D. P. Leyden: "Geologic Reservoir Types Challenging the Future Development of the

Clearwater Formation at Cold Lake, Alberta, Canada," *Proc. of Unconventional Hydrocarbon Sources Conference,* St. Petersburg 1992.
21. D. Wightman, T. Berezniuk: "Resource Characterization and Depositional Modelling of the Clearwater Formation, Cold Lake Oil Sands Area, East Central Alberta," in *The Canadian Heavy Oil Association Handbook,* Calgary, AB, 1991, pp. 267 – 282.
22. Energy Resources Conservation Board, Alsands Fort McMurray Project, Decision Report D 79-H, Calgary, AB, 1979.
23. Energy Resources Conservation Board, Oil Sands Annual Statistics, Publication ST 92-43, Calgary, AB, 1992.
24. E. C. McRoberts: "Waste Management For Surface Mineable Oil Sands Mines," *1st District 5 Meeting, Can. Inst. of Mining and Metallurgy,* Fort McMurray, AB, 1983.
25. R. Lahaie, W. K. Chan: "Large Dykes of the Suncor Oil Sands Mine," *3rd District 5 Meeting Can. Inst. of Mining and Metallurgy,* Calgary, AB, 1987.
26. B. Purcell, A. Slawinski: "Basal Aquifer Dewatering Project at the Suncor Oil Sands Mine," *Can. Geotech. Conference,* Regina, Sask., 1987.
27. M. B. Dusseault, N. R. Morgenstern: "Locked Sands," *J. Eng. Geol.* **12** (1979) 117 – 131.
28. R. Doucet, B. Romanchuk: "Drilling and Blasting in Tarsand," *Proc. of the 11th Conference on Explosives and Blasting Techniques,* San Diego, CA, 1985.
29. A. E. Fair: "Development of Preblasting Techniques to Stabilize Mine Highwall Slopes in Areas of Steeply Dipping Clay Beds," Occupational Health & Safety Division – Mines Inspection Branch – Seminar on Slope Stability, Edmonton, AB, 1985.
30. C. Fordham, E. C. McRoberts, B. Purcell, P. McLaughlin: "Practical and Theoretical Problems Associated With Blast Densification of Loose Sands," *Can. Geotech. Conference,* Calgary, AB, 1991.
31. N. R. Morgenstern, A. E. Fair, E. C. McRoberts: "Geotechnical Engineering Beyond Soil Mechanics – A Case Study," 25th Anniversary Special Paper, *Can. Geotech. J.* **25** (1988) 637 – 661.
32. A. E. Fair, B. A. Isaac: "Highwall Stability Monitoring at Syncrude Canada Ltd. Open Pit Mining Operation," *Proceedings of the 13th Annual Mining and Metallurgy Industries Symposium and Exhibit,* Salt Lake City, UT, 1985.
33. K. A. Clark, D. S. Pasternack: "Hot Water Separation of Bitumen from Alberta Bituminous Sands," *Ind. Eng. Chem.* **24** (1932) 1410.
34. K. A. Clark: "The Hot Water Washing Method for Recovery of Oil from the Alberta Tar Sands," *Can. Oil Gas Ind.* **3** (1950) no. 6, 46.
35. E. C. Sanford, "Processability of Athabasca Oil Sand: Interrelationship between Oil Sand Fine Solids, Process Aids, Mechanical Energy and Oil Sand Age after Mining," *Can. J. Chem. Eng.* **61** (1983) 554.
36. L. L. Schramm, R. G. Smith, *AOSTRA J. Res.* **5** (1989) 87.
37. L. L. Schramm: "The Influence of Suspension Viscosity on Bitumen Rise Velocity and Potential Recovery in the Hot Water Flotation Process for Oil Sands," *JCPT J. Can. Pet. Technol.* **28** (1989) no. 3, 73.
38. R. N. Houlihan, K. H. Williams: "Recent Enhancement in Mined Oil Sands Bitumen Extraction Technology," *JCPT J. Can. Pet. Technol.* **26** (1987) no. 1, 91.
39. Dynawest Projects Ltd.: "Oil Sands Bitumen Extraction Process Evaluation," Alberta Energy Resources Conservation Board, Calgary, AB, 1984.
40. Suncor Bitumen Recovery – Ore Grade/Fines Correlation, private communication, 1993.
41. R. N. Tipman, R. C. Shaw: "Recent Advances in the Treatment of Oil Sands Froth," *Oil Sands – Our Petroleum Future Conference,* preprint, Edmonton, AB, April 1993.
42. L. T. Yano, A. E. Fair: "Overview of Hydraulic Construction Techniques Utilised at the Syncrude Canada Limited Tailings Pond," *Specialty Conference on Hydraulic Fill Structures,* Fort Collins, CO, 1988.
43. M. MacKinnon, A. Sethi: "A Comparison of the Physical and Chemical Properties of the Tailings Ponds at the Syncrude and Suncor Oil Sands Plants," *Oil Sands – Our Petroleum Future Conference,* Edmonton, AB, Apr. 4 – 7, 1993.
44. A. E. Fair, G. T. Handford: "Overview of Tailings Dyke Instrumentation Program at Syncrude Canada Limited," *International Symposium on Geotechnical Stability in Surface Mining,* Calgary, AB, 1986.
45. W. K. Chan, J. Martschuk, E. C. McRoberts: "Performance Review of Tar Island Dyke," *7th Panamerican Conference on Soil Mechanics*

and Foundation Engineering, Vancouver, B.C., 1983.
46. A. W. Hyndman, R. W. Luhning, *JCPT J. Can. Pet. Technol.* **30** (1991) no. 2, 61.
47. R. W. Luhning, "Heavy Oil, Oil Sands and Enhanced Oil Recovery; Where Will the Technology Break-through Come From," *Canadian Heavy Oil Association Conference on Heavy Oil – A New Direction,* Calgary, AB, Dec. 5, 1989.
48. P. R. Kry: "In Situ Recovery Technology at Cold Lake," *Oil Sands – Our Petroleum Future Conference,* Edmonton, AB, Apr. 4 – 7, 1993.
49. L. M. Goobie, H. L. Chang: "The Evolution of a Successful Recovery Scheme for the Peace River Oil Sands, Alberta, Canada," in [48].
50. M. J. Zatka, R. W. Kangen, J. I. Chambers: "Progress at Shell's Peace River In-Situ Recovery Operation," *AOSTRA Oil Sands 2000 Conference,* Edmonton, AB, March 26 – 28, 1990.
51. M. Dusseault: "Cold Production and Enhanced Oil Recovery," *JCPT J. Can. Pet. Technol.* **32** (1993) no. 9.
52. J. W. Kramers, R. A. S. Brown, *CIM Bull.* **69** (1976) 92.
53. S. Singh et al.: "Technical and Economic Factors for Establishing Hydrocarbon Resource Cost Thresholds," *Proc. of the 13th World Petroleum Congress,* Buenos Aires 1991.
54. National Energy Board: Canadian Energy Supply and Demand: 1990 – 2010, Calgary, June 1991.
55. A. E. Fair, W. A. Hill, F. R. Payne, B. Bruce: "Economic Comparison of Oil Sand Surface Mining Technologies," *Oil Sands – Our Petroleum Future Conference,* Edmonton, AB, Apr. 4 – 7, 1993.
56. W. Maciejewski, R. J. Oxenford: "Hydrotransport – An Enabling Technology for Future Oil Sands Development," in [55].
57. R. Koszarycz, R. Padamsey, R. M. Ritcey: "Synthetic Crude Oil from Retorted Mined Oil Sands – Economic Facts and Misconceptions," in [55].
58. Syncrude Application to ERCB, No. 921 321, Energy Resources Conservation Board, Calgary, AB, Sep. 1992.
59. R. G. Moore et al.: "A Canadian Perspective on In Situ Combustion," *Joint Canada/Romania Heavy Oil Symposium,* Sinaia, Romania, 1993.
60. R. G. Moore, D. W. Bennion, M. G. Ursenbach: "A Review of In Situ Combustion Mechanisms," *4th UNITAR/UNDP International Conference on Heavy Crude and Tar Sands,* Edmonton, AB, Aug. 7 – 12, 1988.
61. R. M. Butler: "The Potential for Horizontal Wells for Petroleum Production," *Annual Technical Meeting of the Petroleum Soceity of CIM,* Calgary, AB, 1988.
62. K. H. Chung, R. M. Butler: "A Theoretical and Experimental Study of Steam-Assisted Gravity Drainage Process," *4th UNITAR/UNDP International Conference on Heavy Crude and Tar Sands,* Edmonton, AB, Aug. 7 – 12, 1988.
63. D. P. Komery, J. C. O'Rourke, J. I. Chambers: "AOSTRA Underground Test Facility UTF Phase B Implications for Commercialization," *Oil Sands – Our Petroleum Future Conference,* Edmonton, AB, Apr. 4 – 7, 1993.
64. D. J. Anderson: "The Heated Annulus Steam Drive Process for Immobile Tar Sands," *4th UNITAR/UNDP International Conference on Heavy Crude and Tar Sands, Edmonton,* AB, Aug. 7 – 12, 1988.
65. B. L. Schulman, R. L. Dickenson: "Upgrading Heavy Crudes, A Wide Range of Excellent Technologies now Available," *5th UNITAR/INDP International Conference on Heavy Crudes and Tar Sands,* Caracas, Venezuela, 1991, p. 105.
66. H. D. Sloan, *Hydrocarbon Process.,* Nov. 1991, 99.
67. D. E. Sheeran: "An Improved Understanding of Fine Tailings Structure and Behavior," Proceedings Fine Tails Symposium, *Oil Sands – Our Petroleum Future Conference,* Edmonton, AB, Apr. 4 – 7, 1993.

Tartaric Acid

JEAN-MAURICE KASSAIAN, ETS Legré-Mante, Marseilles, France

1. **Introduction** 1
2. **Physical Properties** 2
3. **Optical Isomerism: Configuration** . . 3
4. **Chemical Properties** 3
5. **Resources and Starting Materials** . . . 3
6. **Production** 3
6.1. **Production of l(+)-Tartaric Acid** . . . 3
6.2. **Fermentation Process** 4
6.3. **Chemical Synthesis** 4
7. **Environmental Protection** 5
8. **Quality Specifications** 5
9. **Chemical Analysis** 5
10. **Derivatives of l(+)-Tartaric Acid** . . . 5
11. **Uses** 6
12. **Economic Aspects** 7
13. **Physiological and Toxicological Behavior** 7
14. **References** 7

1. Introduction

Tartaric acid is a mixed diacid and dialcohol, empirical formula $C_4H_6O_6$.

It is also known as acidum tartaricum, 2,3-dihydroxysuccinic acid and, according to the official IUPAC nomenclature, 2,3-dihydroxybutanedioic acid. The structural formula is:

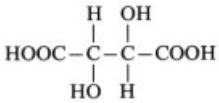

The molecule contains two asymmetric carbon atoms; there are four stereoisomeric forms.

Optically Active Forms. L(+)-Dextrotartaric acid and D(−)-levotartaric acid rotate the plane of polarization of polarized light. These two forms, which are enantiomers, have identical physical and chemical properties. The optical rotatory powers are equal, but of opposite sign.

Optically Inactive Forms. DL-Tartaric acid (racemic tartaric acid), obtained by chemical synthesis, is an equimolar mixture of the two enantiomers; *meso*-tartaric acid (unresolvable tartaric acid) is inactive owing to internal compensation. These two forms have no effect on polarized light. They have the same chemical properties (identical to the active forms), but differ in certain physical properties from each other and the optically active acids.

Tartaric acid, known from antiquity as its acid potassium salt (a sparingly soluble compound which forms a deposit in wine vats), was characterized by the Swedish chemist SCHEELE, who extracted it by decomposing its calcium salt with sulfuric acid. This extraction method is still currently used. BERZELIUS, another illustrious Swedish chemist, established the structural formula in 1830. Between 1848 and 1860, PASTEUR made one of the most fundamental discoveries in organic chemistry while studying the asymmetry and optical activity of salts of tartaric acid, establishing the relationships between racemic tartaric acid and levotartaric and dextrotartaric acids.

PASTEUR showed that, if sodium ammonium tartrate is crystallized below 28 °C, a tetrahydrate is obtained in which half the crystals have supplementary hemihedral faces oriented to the left, and the other half have hemihedral faces oriented to the right. He separated these crystals by hand, using a lens, and showed that some of them yielded a new levorotatory acid, and that the others yielded the already known dextrorotatory acid [1].

L(+)-Dextrotartaric acid is the only natural form and the main form produced commercially. It is widely distributed in plants, either in its free form or as salts. It is found in many fruits, especially in grapes, where it occurs particularly as the acid potassium salt, and to a small extent as the neutral calcium salt.

During vinification, and as alcoholic fermentation proceeds, the acid potassium tartrate becomes insoluble and crystallizes on the vat walls. This deposit, called tartar, is collected and used as starting material in the production of L(+)-tartaric acid.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_163

Table 1. Physical properties of the isomers of tartaric acid [1–6]

	L(+), D(−) Tartaric acid	Racemic tartaric acid	*meso*- Tartaric acid
Crystal form	monoclinic prisms	triclinic prisms	tetragonal plates
Density, g/cm^3	1.7598	1.788	1.666
Melting point, °C (anhydrous)	169 – 170	206	159 – 160
Solubility, g/100 g H_2O (25 °C)	147	25	167
Solubility of potassium bitartrate, g/100 g H_2O (20 °C)	0.57	0.56	13.2
Solubility of calcium salt, g/100 g H_2O (20 °C)	0.026	0.004 *	0.034
Water molecules in hydrated Ca salt	4	8	3

* 25 °C.

2. Physical Properties

The physical properties described below refer to natural L(+)-tartaric acid. They are identical to those of D(−)-tartaric acid. A comparison of the characteristic properties of the isomers are given in Table 1 [1–6].

Tartaric acid has molecular mass M_r 150.09. It crystallizes in the anhydrous state from solution above 5 °C, as large colorless monoclinic needles, density 1.7598 g/cm^3, melting point 169 – 170 °C. It has a strong acidic taste and is odorless. It is stable in air.

It decomposes on heating above 220 °C, with an odor of caramelized sugar. Tartaric acid is very soluble in water.

At various temperatures the solubility is:

t, °C	0	5	10	20
g/100 g H_2O	115	120	125	139
t, °C	30	40	50	60
g/100 g H_2O	156	176	195	218
t, °C	70	80	90	100
g/100 g H_2O	244	273	307	343

It is also soluble in alcohol: 100 g alcohol dissolves 20.4 g tartaric acid at 18 °C; it is less soluble in ether: 100 g diethyl ether dissolves 0.3 g tartaric acid at 18 °C.

Tartaric acid solutions have a rotatory power that varies according to concentration:

$$[\alpha]_D^{20} = 15.050 - 0.1535c$$

where c is the concentration of tartaric acid, in the range 20 – 50 % wt/vol. At 20 % wt/vol, $[\alpha]_D^{20} = +11.98$ (−11.98 for the D(–)-form.

Its enthalpy of combustion is 1149.9 kJ/mol. Its specific heat capacity (0 – 100 °C) is 1237 kJ kg^{-1} K^{-1}.

Tartaric acid is a diacid. Its electrolytic dissociation constants at 25 °C are $K_1 = 1.17 \cdot 10^{-3}$, $K_2 = 5.0 \cdot 10^{-5}$.

At various temperatures (°C) the relative density d_4^{15} of solutions is as follows:

1	10	20	30	40	50
1.0045	1.0469	1.0969	1.1505	1.2078	1.2696

The boiling points of solutions are: 102.2 °C for a 25 % solution, 106.7 °C for a 50 % solution.

The refractive index at the melting point n_D^{170} is 1.464.

3. Optical Isomerism: Configuration

The conventions using the symbols *l, d,* L, D to denote the enantiomers can lead to confusion, and an improved system of nomenclature for compounds containing one or more chiral centers was developed by Cahn, Ingold, and Prelog in 1960. It uses the symbols *R* (Latin, *rectus*, right) and *S* (*sinister*, left). This system enables the compounds to be classified without ambiguity [7, 8]:

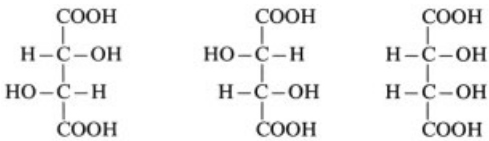

1 L(+) or (2*R*-3*R*) or [*R*-(*R**,*R**)]tartaric acid [*87-69-4*]

2 D(−) or (2*S*-3*S*) or [*S*-(*R**,*R**)]tartaric acid [*147-71-7*]

3 *meso*-tartaric acid [*147-73-9*]

DL-Racemic tartaric acid, an equimolar mixture of (**1**) and (**2**), has CAS No. [*133-37-9*].

4. Chemical Properties

Action of Heat. Tartaric acid melts between 170 and 180 °C, and isomerizes without loss of water to metatartaric acid. If heating is continued, amorphous anhydrides are obtained, which regenerate tartaric acid when boiled with water. At about 220 °C tartaric acid decomposes, swells, and ignites, leaving a carbonized residue.

Oxidation. Tartaric acid is very sensitive to oxidizing agents. The Fenton reaction involving oxidation with hydrogen peroxide in the presence of ferrous sulfate leads to the formation of dioxymaleic acid.

Reduction. Reduction with hydriodic acid leads to the formation of succinic acid.

Complexing Action. Tartaric acid acts as a complexing agent, preventing the precipitation of heavy metal salts by bases.

Boiling. Prolonged boiling of L(+)-tartaric acid solutions in the presence of alkalis (potassium and sodium hydroxides) yields the racemic and meso acids [1,3].

5. Resources and Starting Materials

L(+)-Tartaric acid is the only one of the four isomers to be produced industrially on a large scale. The starting materials are exclusively natural residues derived from vinification. The tartaric acid in these residues exists mainly in the form of potassium bitartrate and, to a lesser extent, calcium tartrate. Depending on their tartaric acid concentration and composition, these materials are classified under four headings [9–12].

Tartar. This is deposited by crystallization on the walls of wine vats as the ethyl alcohol concentration increases. Tartar contains 80 – 90 % potassium bitartrate.

Lees. These are deposited on the bottom of wine vats and contain 19 – 38 % potassium bitartrate.

Dried Lees. These are lees that have been treated, and contain 55 – 70 % potassium bitartrate.

Tartrate. This is produced by distilleries from wine lees or grape marc (the residue obtained after pressing grapes). After distillation to remove the alcohol, the lees and marc are precipitated with milk of lime, in the form of calcium tartrate.

6. Production

Two of the four isomers of tartaric acid are produced industrially, mainly L(+)-tartaric acid, but the racemic acid is also produced in very small amounts by chemical synthesis.

6.1. Production of l(+)-Tartaric Acid

Tartaric acid is produced mainly in Spain, France, and Italy. One of the oldest producers is Société Legré-Mante established in Marseilles in 1784. All processes for producing tartaric acid are based on the decomposition of calcium tartrate with sulfuric acid.

Two competing methods have been used to convert the starting materials to calcium tartrate: the Scheurer – Kestner acid process, and the Scheele – Lowitz neutral process, together with the Desfosses variant [9,13]. Nowadays, only the less costly neutral process is employed.

After grinding, the dried starting material (tartar or lees) passes to a roaster where it is heated for 2 h at 160 °C, to remove the organic material that might interfere in the filtration.

After it has been roasted, the hot product is passed to a reactor where it is diluted with water and neutralized to pH 5 with milk of lime. The reaction temperature is held at 70 °C. The addition of calcium chloride or calcium sulfate in 10 % excess with respect to the stoichiometric amount gives a more complete reaction and reduces losses.

The chemical reaction that takes place is:

$$2\,KHC_4H_4O_6 + Ca(OH)_2 + CaCl_2 \longrightarrow 2\,CaC_4H_4O_6 + 2\,KCl + 2\,H_2O$$

The calcium tartrate is separated from the mother liquor by filtration on a rotary filter and washed.

The tartrate obtained from distilleries enables this first stage to be omitted.

The second stage is the decomposition of the calcium tartrate in aqueous solution with sulfuric acid, which yields tartaric acid solution and insoluble calcium sulfate

$$CaC_4H_4O_6 + H_2SO_4 \longrightarrow H_2C_4H_4O_6 + CaSO_4$$

An excess of sulfuric acid of ca. 5 % with respect to tartaric acid is necessary in order to optimize the reaction. The pulp obtained by this decomposition is filtered and washed.

The red tartaric acid solution is recovered at a concentration of ca. 200 g/L. The wash waters are recycled to the decomposition stage and the calcium sulfate is discharged as waste.

The tartaric acid solution is concentrated in vacuum evaporators at 70 °C to a concentration of 650 g/L. This new solution is then concentrated in suitable apparatus under vacuum at 70 °C until crystals appear (at ca. 1300 g/L).

This liquid – crystal mixture is then passed to crystallizers where it is slowly cooled to provide a maximum yield of crystals. After the mixture has been cooled it is dried in centrifuges, where the crystals are separated from the mother liquor. The mother liquor is subjected to repeated evaporation, granulation, and drying cycles to recover the maximum yield of granulated tartaric acid.

The granules obtained in each operation are cooled to give a tartaric acid solution of concentration 650 g/L, which is then decolorized with activated charcoal and chemically purified if excess iron and sulfuric acid are present. This operation is carried out at 70 °C.

This solution is passed through a filter press and becomes almost colorless, and is then concentrated under vacuum until crystallization starts, following which it is passed to crystallizers or cooled to provide maximum yield of crystals.

This mixture is separated by centrifugation. The mother liquor, which still contains tartaric acid, is subjected to repeated evaporation, granulation, and drying cycles before being retreated. The purified granules are dried in an oven at 140 °C and passed to a battery of sieves for separation into different grain sizes, from 2000 µm to < 100 µm (powder). Other processes are sometimes used [14, 15].

Two other production processes have been studied: one involves production of L(+)-tartaric acid by fermentation; the other involves production of racemic DL-tartaric acid by chemical synthesis.

6.2. Fermentation Process

Studies and investigations include conversion of glucose by bacteria of the species *Acetobacter suboxydans* [16]; and conversion of *cis*-epoxysuccinic acid and its sodium derivatives by bacteria of the species *Nocardia tartaricans* [17]. Publications and patents relating to other research work are listed in [18].

6.3. Chemical Synthesis

Racemic DL-tartaric acid is produced from maleic acid or its salts [19, 20]. Maleic acid in aqueous solution is oxidized by 35 % hydrogen peroxide in the presence of potassium tungstate. Epoxysuccinic – tartaric acid is formed, which is hydrolyzed on boiling to racemic tartaric acid. The latter is recovered after cooling, centrifugation, washing, and drying.

Several variants of this process have been studied and published, and numerous patent applications have been filed.

Racemic DL-tartaric acid is produced commercially in a small production unit in South Africa.

7. Environmental Protection

In the production of L(+)-tartaric acid, the solid waste is discharged into bins before being sent to authorized dumps. The liquid waste is collected, treated with milk of lime, and filtered before being discharged into the sewage system, where it is retreated by municipal purification plants.

Daily analyses of the COD, suspended matter, and sulfates in the effluent are obligatory. Independent checks are carried out by authorized bodies.

Tartaric acid and its salts are biodegradable compounds.

8. Quality Specifications

L(+)-Tartaric acid, 2,3-dihydroxybutanedioic acid (2*R*, 3*R*) is used mainly in the food and pharmaceutical industries, and therefore has to satisfy extremely stringent quality criteria, defined in the pharmacopoeia of each country. The two principal publications are the European Pharmacopoeia (1986) and the United States Pharmacopoeia (1990), whose requirements are very similar.

Limiting values from the European Pharmacopoeia are:

Tartaric acid content	99.5 – 101 %
Loss on drying	$\leq$ 0.2 %
Calcination residue	$\leq$ 0.1 %
Rotatory power in 20 % wt/vol solution	+ (12 – 12.8°)
Sulfates	$\leq$ 150 ppm
Chlorides	$\leq$ 100 ppm
Calcium	$\leq$ 200 ppm
Heavy metals	$\leq$ 10 ppm
Oxalic acid	$\leq$ 350 ppm

9. Chemical Analysis

The official analysis of tartaric acid in starting materials is the Goldenberg – Géromont method (1898 – 1907) as modified by the VIIth Chemistry Congress, London (1909). The accuracy of this analysis is $\pm$ 0.5 %. This method enables total tartaric acid to be determined without distinction between the isomers or salts derived [9, 13, 21, 22].

The first stage consists in acidifying the starting material with 5.6 mol/L hydrochloric acid to dissolve all the tartaric acid derivatives (potassium bitartrate, calcium tartrate).

These derivatives are filtered, boiled, and converted to neutral potassium tartrate with a solution of potassium carbonate at 682 g/L. After filtration followed by evaporation on a water bath, the neutral potassium tartrate is reacted with 3.5 mL pure acetic acid to potassium bitartrate. The precipitate is rendered insoluble with 100 mL ethyl alcohol and collected by vacuum filtration, washed with alcohol, taken up and solubilized with boiling water in a porcelain evaporating dish. This solution is brought to the boil and titrated against sodium hydroxide 0.2 mol/L, using neutral litmus paper as indicator.

Other, significantly less accurate methods exist. The most widely used of these is spectrophotometry, whose accuracy is ca. $\pm$ 2 %. The reagent used is ammonium metavanadate, which forms a colored complex with tartaric acid derivatives [23, 24]. This method is most suitable for continuous analyses during production.

10. Derivatives of l(+)-Tartaric Acid

Tartaric acid forms numerous salts and esters. The commercially best-known compounds are discussed in this Chapter [1–3].

Sodium potassium tartrate is commonly known as Seignette's salt or Rochelle salt. It was discovered in 1672 by Pierre Seignette, a pharmacist working in La Rochelle.

Empirical formula $KNaC_4H_4O_6 \cdot 4\,H_2O$, [*6381-59-5*], M_r 282.23, density 1.79 g/cm^3, *mp* 70 – 80 °C, decomposes at 220 °C; solubility (g per 100 mL water) 26 at 0 °C, 66 at 26 °C.

The starting material for its production is tartar with minimum tartaric acid content 68 %. This is first diluted in water or in the mother liquor of a previous batch. It is then neutralized hot with caustic soda to pH 8, decolorized with activated charcoal, and chemically purified before being filtered. The filtrate is evaporated to 42° Bé at 100 °C, and passed to granulators in which Seignette's salt crystallizes on slow cooling. The salt is separated from the mother liquor by centrifugation, accompanied by washing of the granules, and is dried in a rotary furnace and sieved before packaging.

Commercially marketed grain sizes range from 2000 µm to < 250 µm (powder).

Acid potassium tartrate (potassium bitartrate, monopotassium tartrate) is commonly called cream of tartar. $KHC_4H_4O_6$ [*868-14-4*], density 1.96 g/cm^3; solubility (g per 100 mL water) 0.57 at 20 °C, 6.1 at 100 °C.

This salt is mainly produced from the mother liquor of Seignette's salt. These decolorized, purified, and filtered solutions are acidified to pH 3.5 with hydrochloric or sulfuric acid. Since cream of tartar is sparingly soluble, it precipitates and is recovered by centrifugation, dried, and ground before being packaged as fine powder.

Potassium antimonyl tartrate, also known as tartar emetic, was discovered in 1631 by ADRIEN DE MYNSICHT. $KSbC_4H_4O_7 \cdot 1/2\,H_2O$, [*28300-74-5*], M_r 333.93, density 2.61 g/cm^3; solubility (g per 100 mL water) 8.7 at 25 °C, 35.7 at 100 °C.

The compound is prepared by reacting 3 parts antimony oxide with 4 parts cream of tartar diluted in water. The reaction mixture is heated to 100 °C and held at this temperature for 4 h. After filtration, tartar emetic crystallizes in the cooled solution in a mixer. The granules are extracted by centrifugation. After drying, the granules are pulverized.

Metatartaric acid is obtained by heating tartaric acid to 170 – 180 °C. The conversion takes place without any loss of weight. Two or more molecules react with mutual esterification.

The compound obtained is more or less polymerized. Metatartaric acid has outstanding inhibitory properties with respect to precipitation of tartaric acid salts.

11. Uses

L(+)-Tartaric acid and its derivatives are used particularly in the food, pharmaceutical, and viniculture industries. There are also numerous applications in other fields [2, 25, 26].

l(+)-Tartaric acid (E 334) is used:

1) In the acidification of wine musts
2) As an acidifier and taste enhancer in sweets, candies, jellies, jams, fruit nectars, ice creams, gelatins, and pastes
3) In fruit, vegetable, or fish preserves, where it acts as a synergistic antioxidant, and also stabilizes the pH, color, taste, and nutritional value
4) In greases and oils, as an antioxidant
5) In the preparation of carbonated beverages
6) In the pharmaceutical industry, as an excipient or carrier for the active principal, helping to correct the basicity
7) Because of its simplicity of use, stability, and high solubility, as an acidification agent for foaming tablets and powders
8) In the cement industry, and particularly in plaster and gypsum, where its retardant action facilitates handling of these materials
9) In polishing and cleaning metals

Cream of tartar (E 336) is used:

1) To accelerate the precipitation of tartaric acid salts in wine
2) In the production of chemical yeasts
3) In the production of some drugs, by virtue of its laxative action

Seignette's salt (E 337) is used:

1) In electroplating, where it improves deposition and yield
2) In electronics and piezoelectricity
3) As a reducing agent in the silvering of mirrors
4) As a laboratory reagent; it is one of the constituents of Fehling's solution
5) In the production of cigarette paper, as a combustion regulator
6) In the pharmaceutical industry, on account of its laxative action

Potassium antimonyl tartrate is used:

1) In small amounts as an expectorant in cough syrups
2) In the treatment of some tropical diseases
3) As a mordant in the textile and leather industries

Metatartaric acid is used to inhibit tartrate crystallization in table wines.

The tartrates of fatty acid monoglycerides and diglycerides are used as emulsifiers in the bakery industry.

12. Economic Aspects

L(+)-Tartaric acid has a world market of ca. 30 000 t (ca. 1993). The main producers are France, Spain, and Italy; it is also produced in small amounts in Portugal, Argentina and Japan.

The tartaric acid market has remained static over the last decade. Production is entirely dependent on vinicultural raw materials, leading to large fluctuations in the retail price of tartaric acid.

13. Physiological and Toxicological Behavior

Tartaric acid, unlike some other organic acids in fruit, is not an intermediate of the Krebs cycle (→ Citric Acid, Chap. 1.), which is a major center of organic synthesis in the human body. It may be ingested in more or less large amounts with foodstuffs such as fruit and wine, where it occurs in the L(+)-form. After oral absorption, about 20 % is excreted in the urine, a large proportion of the remainder undergoing bacterial breakdown in the intestines. No traces are found in the feces. The absorbed fraction is rapidly eliminated from the blood, and is either excreted by the kidneys or oxidized in various tissues.

Long-term toxicity studies in rats have been carried out with L(+) monosodium tartrate. No toxic or carcinogenic effects have been detected. Moreover, no adverse effects on renal function and pathology have been detected at a daily dose of 3 g/kg.

The toxicity of the (DL)-form has been less widely studied. One study has shown that it behaves differently from the L(+)-form. The (DL)-isomer disappears more slowly from the blood and accumulates for a longer period in the kidneys, producing a weight increase in the latter. On account of this nephrotoxic effect, this isomer has not been approved for use in foodstuffs and pharmaceuticals [27–30].

14. References

1. V. Grignard: *Traité de chimie organique,* **vol. 11,** Masson et Cie, Paris 1945, pp. 285 – 331.
2. *Merck Index* 1989, pp. 1432 – 1433.
3. A. D. Wurtz: *Dictionnaire de Chimie,* **vol. 3,** Librairie Hachette et Cie, Paris 1882, pp. 198 – 246.
4. A. D. Wurtz: *Dictionnaire de Chimie,* **vol. 7,** Librairie Hachette et Cie, Paris 1908, pp. 663 – 671.
5. H. V. Ockermann: *Source Book for Food Scientists,* Avi Publishing, Westport, Conn. 1978, p. 276.
6. *Dictionary of Organic Compounds,* 4th ed., **vol. 5,** Eyre and Spottiwoode, Oxford 1965, 2943 – 2945.
7. *Vogel's Text Book of Practical Organic Chemistry,* 5th ed., J. Wiley & Sons, New York 1989, pp. 4 – 7.
8. J. March: *Advanced Organic Chemistry,* 4th ed., McGraw Hill, New York 1992, pp. 94 – 114.
9. J. Ventre: *Dérivés Tartriques de la Vendange,* Librairie Coulet et fils, Montpellier 1909.
10. C. Cantarelli: "Produits secondaires de la Vinification, Rapport Général," Office International de la Vigne et du Vin, no. 501 (1972) 947 – 967.
11. J. Mourgues: "Valorisation des sous produits des distilleries vinicoles," *C.R. Seances Acad. Agric. Fr.* **59** (1973) 475 – 480.
12. J. Mourgues, J. Mangenet, *Ind. Aliment. Agric.* **1** (1975) january.
13. U. Roux: *La Grande Industrie des Acides Organiques,* Librairie Dunod, Paris 1939.
14. Montecatini, US 3 069 230, 1962 (B. Pescarolo, V. Bianchi).
15. Orandi et Massera, US 3 114 770, 1963 (L. Dabul).
16. T. K. Yamada, T. Kodoma, T. Obata, N. Takakashi, *J. Ferment. Technol.* **49** (1971) no. 2, 85 – 92. K. Yamada, Y. Minoda, T. Kodama, U. Kotera, US 3 585 109, 1971.H. K. Bhat, G. N. Quazi, S. K. Chaturvedi, C. L. Chopra, *Res. Ind.* **31** (1986) no. 2, 148 – 152. Skowa Chem., Mitsubishi Chem. Ind., JP Kokai 93 163 193, 1993 (N. Ootomo).
17. Tokuyama Soda, JP Kokai 77 90 690, 1977 (K. Yutani, H. Takesue, K. Fujii, Y. Miura). Tokuyama Soda, JP Kokai 7 914 917, 1979 (H. Takesue, Y. Miura, M. Shibuya, M. Toyoshima). Tokuyama Soda., DE-OS 2 605 921, 1976 (Y. Miura et al.). Takeda Chemical Industries, DE-OS 2 619 311, 1976 (Y. Kamatani et al.).
18. T. Huang, Y. Qian, *Gongye Weishengwu (Ind. Microbiol.)* **20** (1990) no. 6, 14 – 17. Z. Zhang, T. Huang, *Gongye Weishengwu (Ind. Microbiol.)* **20** (1990) no. 2, 7 – 12, 24. Miles Laboratory, US 2 314 831, 1943 (I. Kamlet). Institute of Microbiology, Academy of Sciences, Moldavian S.S.R., SU 514 891, 1976.
19. J. M. Church, R. Blumberg, *Ind. Eng. Chem.* **43** (1951) no. 8, 1780 – 1786.
20. Imperial Chemical Industries, GB 1 442 748, 1976 (F. Lewis, P. A. Rodriguez). Faming Zhuanli Shenqing Gongkai Shuomingshu, Peoples Republic of China (Xintong Group), 1 050 712, 1991 (X. Wu, Z. Liu et al.). Faming Zhuanli Shenqing Gongkai Shuomingshu,

Peoples Republic of China, 85 100 455, 1986 (Z. Zhang et al.). Nippon Peroxide, JP Kokai 7 339 437, 1973. G. M. Grinberg, S. M. Gabrielyan, N. M. Morlyan, M. K. Madroyan, SU 322 043, 1975. Österreichische Chemische Werke, DE-OS 2 555 699, 1976 (K. Petritsch, P. Korl, F. Pogoriach). Nippon Peroxide, JP Kokai 7 785 119, 1977 (K. Kazutani et al.). S. Eisler et al., *Rev. Chim. (Bucarest)* **39** (1988) no. 8, 660 – 663. Kawaden Fine Chemicals, JP Kokai 6 372 647, 1988 (Nozue, Moriaki).

21. Goldenberg, Geromont et Cie, *Z. Anal. Chem.* (1898) 312.
22. Chem. Fabrik, former Goldenberg, Geromont et Cie, *Z. Anal. Chem.* **47** (1908) 57.
23. H. Rebelein, *Mitteilungsbl. GDCh Fachgruppe Lebensmittelchem. Gerichtl. Chem.* **24** (1970) 14.
24. H. Tanner, M. Sandoz, *Schweiz, Z. Obst Weinbau* **108** (1972) 10.
25. *Pharmaceutical Codex,* **vol. 11,** Pharmaceutical Press, London 1979, pp. 53, 720, 836, 916.
26. E. F. James (ed.): *The Extra Pharmacopoeia,* 30th ed., Pharmaceutical Press, London 1993, pp. 39, 900, 904, 1408.
27. F. P. Underhill et al., *J. Pharmacol. Exp. Ther.* **43** (1931) 359, 381.
28. FAO/OMS reports nos. 7, 17, 21, 27.
29. P. Dupuy, "Le Métabolisme de l'Acide Tartrique," *Ann. Technol.* **2** (1960) 139 – 184.
30. T. E. Furia: *Handbook of Food Additives.* The Chemical Rubber Co., Boca Raton, Fla. 1968, pp. 261 – 263.

Tea

ANITA CROCKER, The Tea Council Limited, London, United Kingdom

1. **Introduction** 1
2. **History** 1
3. **Major Tea Producers** 2
4. **Growth and Manufacture** 3
5. **Tea Trade** 5
6. **Tea Market and Consumption** 5
7. **Tea and Health** 5
8. **References** 6

1. Introduction

Tea is an evergreen plant belonging to the Camellia family and is known as *Camellia sinensis*. It was first discovered ca. 5000 years ago in China by the legendary emperor SHEN NUNG.

Tea is indigenous to both China and India, as the BRUCE brothers, employees of the East India Company, discovered in 1823. Hence today, the tea trade uses the term China or Indian jat to distinguish the origin of the original tea bushes on some of the older estates.

If left to grow wild, tea grows as a tree reaching some 30 m high; thus it was that monkeys used to be trained to pick the leaves and throw them down to people waiting to collect them below.

In modern cultivation, tea is kept to a height of some 1.5 m (waist height) for easy plucking. The characteristics and flavor of tea vary according to the type of soil, altitude, and climatic conditions of the area in which it grows. Other factors affecting tea's characteristics and flavor are the method by which it is processed – "made" in tea trade terms – and the blending of teas from different areas or countries. Tea offers a wide choice of varieties, more than 1500 blends. There is a tea to suit every occasion.

2. History

The early beginnings of tea are wrapped in legend and myth. Circa 2737 B.C. tea drinking was first recorded in China, when SHEN NUNG, whilst touring his empire, sat down beneath a tree as his servant boiled some water. A leaf from the tree dropped into the water and SHEN NUNG decided to try the brew. The tree was a tea tree (*Camellia sinensis*).

Two thousand years later, in the Tang dynasty (906 – 618 B.C.), tea was China's national drink, and the modern term tea derives from the early Chinese words Tchai, Cha, Tay etc., used to describe both the beverage and the tea leaf. Circa 552 A.D. Buddhist monks are reputed to have taken tea to Japan. They not only cultivated the tea drinking habit, but also planted seeds and pioneered tea cultivation in Japan. The Buddhists were also responsible for developing the Japanese tea ceremony, which is said to be based on the "Code of Practise" written by the Chinese scholar LO YU, circa 900 A.D.

Tea was first mentioned in Europe around 1559 when the Venetian author GIAMBATTISTA RAMUSIO published the stories of a Persian trader, Hajji Mahommed and his tales of tea in China.

Although, the Portuguese were the first to open up trade with the Far East, and Portugal's missionaries are reputed to have brought the tea drinking habit back to Portugal, many of the Portuguese ships were crewed by other nationalities. Thus it was a Dutch navigator, JAN HUGO VAN LINSCHOON, who is reputed to have fired the enthusiasm of Dutch merchants and sea captains to trade with the Far East, and this led to Holland trading in tea with its European neighbors, including Scandinavia and Britain.

Tea is thought to have been brought back to Britain by the East India Company as early as 1633, but the first recorded sale of tea was in 1657 when THOMAS GARWAY sold tea in Garraway's Coffee House in Exchange Alley, London. The first tea auction was held in the early 1700s, and Britain had established herself as the center of the world tea trade by 1790.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_171

3. Major Tea Producers

Tea grows in more than 32 countries around the world. Today's main producers are India, China, Sri Lanka, Indonesia, Kenya, Malawi, Tanzania, Zimbabwe, Papua New Guinea, Bangladesh, Mauritius, Rwanda, and the Cameroons. Tea is also grown in South America, Australasia, Eastern Europe, Middle and Far East, but is mostly consumed domestically.

Some teas are seasonal, others not. This depends on altitude, nearness to the equator, and climatic conditions in the producer country. The yield per hectare is also influenced by these natural factors, plus good husbandry on the part of the producer. For example, the harvest weight of plucked tea leaves on an estate can fluctuate between 1600 and 6000 kg per hectare in East Africa or Assam, but is lower in high-altitude areas such as Darjeeling.

India, one of the main tea growers, produces ca. 30 % of the world's tea, with over 400 000 hectares under tea cultivation. Teas from India are seasonal and are generally plucked from March until the end of October. The tea producing areas range from low-grown (sea level to 609 m above sea level) to high-grown tea (> 1219 m above sea level). Each growing area produces teas with unique flavor and characteristics, peculiar to a given region. Assam, Darjeeling, and Nilgiri teas are examples, and apart from their use in blended teas, they are also well-known worldwide as speciality teas.

Sri Lanka is the world's largest exporter of tea. Teas from this tropical island retain the name "Ceylon", although the traditional name of Sri Lanka was readopted in 1972 when Ceylon became a Sovereign Republic in the Commonwealth. Ceylon teas have been arriving in the London auction since 1867. Today Sri Lanka has some 240 000 hectares under tea cultivation, and because of its geographical position is able to pluck tea throughout the year.

The west and east side of the island are divided roughly by the central mountain system; thus, due to the natural climatic conditions of the area the western side's tea plucking season alternates with that of the eastern side.

Tea is one of Sri Lanka's main foreign exchange earners, and Ceylon teas are drunk as speciality teas throughout the world, as well as being used for blending. The teas span the whole spectrum of growth from low-grown to high-grown.

Indonesia's tea industry was founded by the Dutch in the 1700s. Teas from there, together with teas from India and Ceylon, dominated the black tea markets of Britain and Europe until 1939 – 40. After World War II Indonesian tea estates were in a sorry state. Wrecked factories and tea bushes reverted to their wild "tree" state were just two of the problems which faced the country.

By 1984, tea exports from Indonesia began to reappear on the tea market. Improvements in tea production, replanting of old estates, and investment in new factory machinery and manufacturing methods has enabled Indonesia to gain an estimated 12 % of world tea exports.

Teas from Kenya are probably the best known of the East African teas as Kenya sells her tea as a speciality tea as well as for blending. Kenya's tea history dates back to 1903 when C.S.L. Caine imported tea seeds from India and began to produce quality tea on a tea farm two acres in size. Other experimental farms followed and by 1928 Kenya was sending tea for sale at the London tea auction. Near to the equator, Kenya is a fertile land with a climate which allows it to grow and pick tea throughout the year.

Malawi pioneered the tea trade in East Africa. Circa 1878 tea farming commenced. However by the 1880s commercial tea production was firmly established in Mulanje. Malawi teas are used mainly for blending and have a bright color and brisk flavor. Today Malawi has a 3.8 % share of world tea exports.

Zimbabwe teas, like those from Malawi, are used mainly for blending. They have a 1.1 % share of world tea exports, which are important as a foreign exchange earner. The first fully irrigated tea estate was established in Zimbabwe, as the country has an annual rainfall of only 66 cm and tea needs some 127 cm.

The tea industry in Tanzania was originated by the Germans during the reign of Wilhelm II, and established itself between the two world wars. Teas in Tanzania are grown at low, medium, and high altitudes, have a bright color and brisk flavor, are exported as a foreign exchange earner, and are used in the West for blending purposes.

China is the second largest tea producer after India and the second largest tea exporter after Sri Lanka. The Chinese retain the majority of the tea they grow for home consumption, exporting approximately a quarter of their annual yield. World exports of tea in 1982 and 1992 are shown in Figure 1.

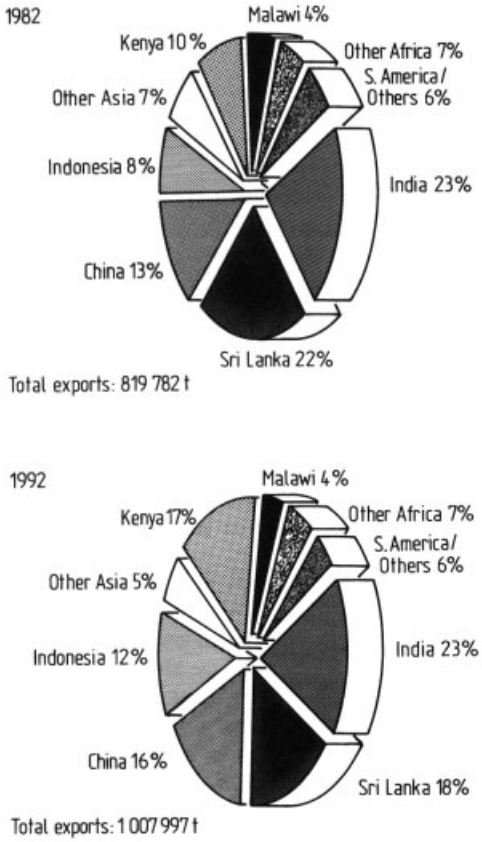

Figure 1. World tea exports (courtesy of International Tea Committee)

4. Growth and Manufacture

Tea is grown on estates or smallholdings. A smallholding can be as small as 0.5 hectares or several hectares. The owner grows tea and sells his leaves to a factory for processing. An estate is a self-contained unit, housing its own factory, schools, hospital, staff houses and gardens, shops, places of worship, reservoir, and guest house.

In modern tea cultivation, the bushes are grown from cuttings, or clones. The tender young clones are nurtured and trained into fan-shapes in nursery beds until they are ready for planting out on either a smallholding or estate. On the medium to high altitude areas the bushes are planted around the contours of the hills in rows (Fig. 2), to avoid soil erosion. The bushes are planted approximately 1.5 m apart, with a distance of some 1.0 m between each row, allowing easy access for both the gardeners and pluckers. Each bush is pruned so that it grows in a fan-shape topped by a flat "plucking plateau" roughly 1.5×1.0 m in size. Prunings from the bushes are used as mulches and some for domestic firewood. Tea bushes provide a safe haven for wild life and maintain an environmental balance. Only the top two leaves and a bud (Fig. 3) are plucked from the top of the plucking plateau, and tea is a self renewing crop, growing again within 7 – 14 d of plucking, depending on the altitude at which it is grown.

Figure 2. Tea estate

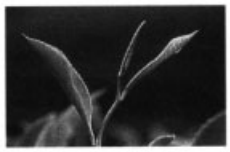

Figure 3. Two leaves and a bud

A skilled plucker can gather between 30 and 35 kg of green tea leaf in a day, sufficient to produce ca.7.5 – 9 kg of processed black tea.

Once plucked there are seven stages of processing in tea manufacture: collection, withering, breaking the leaves, fermenting, drying, sorting, and packing. For small holders the leaves are either taken to the factory or collected at a central point from the smallholders, who are paid for their crop.

On an estate the plucked leaves are collected at a central point, each plucker being credited with their own weights of leaf for subsequent payment. The collected leaf is packed into nets and taken to the factory, transported to the top floor and spread thinly on wire racks exposed to the air at 25 – 30 °C. This is known as withering; the moisture from the leaf evaporates in the warm air leaving the leaf limp. The process can take 10 – 16 h, depending on the wetness of the leaf.

The withered leaf is then broken by the "orthodox" or "unorthodox" method, so that the enzymes in the veins of the leaf are released, come into contact with air, and are oxidized (known in the tea trade as fermenting). The orthodox process (Fig. 4) rolls the leaves, whereas the unorthodox process, known as "cut, tear and curl" or "rotovane", depending on the type of machinery used, breaks the leaf into smaller particles more suited to modern market needs.

Figure 4. Orthodox rolling machine

The broken leaf is laid out on trays or in troughs in a cool humid atmosphere for 3 – 4 h and is periodically turned until all leaf particles are a golden russet color and fermentation is complete.

Next the leaf is dried by a firing process. The leaf is passed slowly through hot air chambers until all the moisture is removed and the leaf turns a dark brown, referred to as "black" in the tea-trade. Then the black tea is sorted by leaf particle size or "grades". The "grades" are Broken Pekoe, the largest particles; fannings, smaller particles which can fall under three or four classifications of decreasing size; and Pekoe Dust, the smallest leaf particles. Other descriptions such as Flowery Orange Pekoe and Golden Flowery Orange Pekoe can be used to describe the leaf appearance. The black tea is sorted by sifting through wire meshes of varying sizes into containers, before being weighed and packed into chests (Fig. 5) or sacks for loading onto pallets. Each chest or sack contains an average of 56 kg of black tea, and a loaded pallet holding 50 – 100 tea chests or sacks is known as a break.

Figure 5. Tea sorting

The above describes the manufacture or "make" of black tea. There are also two other forms of manufacture: green tea and oolong tea.

In green tea manufacture after plucking and collection the tea is left to wither, as with black tea. It is then steamed and rolled before drying. Thus the veins in the leaves are not broken and the enzymes are not oxidized. When green tea is brewed it has a yellowish liquor and the wet leaf is almost entire. Green tea is drunk mostly in China and Japan as well as some parts of South America. A little is drunk in the West as a speciality tea. Oolong tea is a semi-fermented tea. After withering and rolling, fermentation is stopped after ca. 1 h and the tea dried, then sorted. Oolong is widely drunk in the Far East and is gaining a following in the West as a speciality tea. It is a light liquoring tea with a very delicate flavor.

5. Tea Trade

After manufacture the tea is tasted in the factory and a sample airmailed to the estate's brokers in London and national local auction center. Each broker evaluates the tea for quality and price and reports back to the estate, so the tea can be sold to the best advantage. In India there are

several auction centers. Sri Lanka has Colombo; East Africa, Mombasa and Limbe; and Indonesia, Jakarta.

Tea destined for the London auction takes 12 – 16 weeks to arrive at Felixtowe, the main tea port in Britain. From there it is transferred into warehouses under the aegis of the Tea Clearing House. At this stage, samples of the tea are sent to both the selling broker and tea buyers.

The selling broker retastes the tea to ensure it has travelled well, reevalutes it, and decides the price it is expected to bring at auction. All this information is communicated immediately to the estate from which the tea came.

The tea buyers taste the tea for its quality, how much of it is needed for the company's blends and evaluate the tea pricewise from their company's viewpoint. Thus, all parties come to the auction having tasted the teas. After the teas are sold they are transported from the warehouses to the blender and packers factories.

Approximately 25 % of the tea drunk in Britain comes through the London auction, the rest being purchased from auction in the country of origin or sometimes direct from estates.

The rest of the world buys its teas from any of the auctions, on the spot market in Amsterdam, through agents or brokers in the producer countries and London, or direct from the estates. There are no future markets in tea and the price is dominated by supply and demand.

In recent years, with containerization in shipping, an off-shore auction has developed in London, whereby teas that are still on the water en route are sold. Again both sellers and buyers have tasted the teas before they begin their travels, but unlike the landed or national auction, the buyer has to buy the complete container lot, whereas with the landed auction, buyers can split lots and deal across the floor of the auction rooms.

6. Tea Market and Consumption

Black tea has the major share of the tea market in terms of production, sales, and amount drunk. Black tea can be roughly divided into popular, national or local blends, and speciality teas. The speciality teas are those which take their name from the area in which they are grown, e.g., Assam, Darjeeling, Ceylon; a blend of teas for a particular time of day such as English Breakfast, Afternoon Tea; teas blended and named after a person (e.g., Earl Grey) or tea (*Camellia sinensis*) which has been flavored with fruit or flower blossoms (e.g., Jasmine tea, Rose Pouchong, or Orange tea). Green teas, Gunpowder being the best known; oolongs; and Lapsang Souchong, with its distinctive smoky aroma and flavor, also come under the speciality tea heading.

Within the popular blend and speciality tea markets there are two main classes: loose leaf tea and teabags.

In Britain tea consumption accounts for 42 per cent of total drinks drunk, excluding tap water. Eighty per cent of the nation drink tea daily and Britain imports annually ca. 16 % of the producer countries' tea exports. In kilograms per capita per annum British consumption during the past decade has remained relatively stable at 2.56. In Eire, over the same period, per capita consumption was 3 kg. The rest of Europe and Scandinavia tend to consume ca. 1 kg of tea per capita per annum, and the United States follows this pattern. Many of the major growing countries such as India, Sri Lanka, and Kenya consume 1 kg per capita, as do Russia and Turkey. In Australasia tea consumption is around the kilogram per capita figure, whereas in Canada the American pattern seems to hold sway with consumption at 0.5 kg or less per capita.

In the Middle East consumption averages ca. 1 kg per capita per annum.

7. Tea and Health

Tea is a natural product. It contains no artificial coloring, preservatives, or flavoring and is virtually calorie free if taken without milk or sugar. Tea provides a pleasant way of taking in the fluids the body needs daily for optimum health. Tea is a rich source of two minerals essential to health: manganese and potassium.

Manganese is essential for bone growth and the body's development. Our bodies need 2 – 5 mg/d, and 45 % of this can be obtained by drinking five or six cups of tea daily.

Potassium is vital for maintaining a normal heart beat. It is one of the major constituents in living cells; helps balance sodium; enables nerves and muscles to function and regulates fluid levels within the cells. The potassium re-

quirement is 2 – 3.5 mg/d, and 25 % of this can be provided by five or six cups of tea daily.

Tea contains small amounts of carotene (a vitamin A precursor), thiamin (vitamin B1), riboflavin (vitamin B2), nicotinic acid, pantothenic acid and ascorbic acid (vitamin C).

Tea cools, calms, refreshes, and relaxes. Tea helps digestion of foods and acts as a diuretic.

A cup of tea made with a leaf tea or teabag contains 40 mg of caffeine, and with instant tea, 30 mg. Over 200 scientific papers a year are published on caffeine and there is now general consensus that caffeine has no deleterious effects on the heart, blood cholesterol, or the circulation, and there is no evidence that it gives rise to cancers. The FDA has given caffeine GRAS status, that is, it is generally recognised as safe.

Caffeine is a mild stimulant, which can increase concentration and alertness, accuracy and sensitivity of taste and smell. In high doses (> 800 mg/d) it can invoke anxiety and unpleasant gastric sensations. Tea has not been shown to have either effect, possibly due to the fact that to get this amount of caffeine from tea would involve drinking unrealistically large quantities.

Tea also contains polyphenols, termed tannins by the tea trade. The polyphenolic compounds in tea have an important sensory role. They contribute to the color, taste, and refreshing aspect of the drink.

8. References

1. *The Tea Council Annual Reports; File on Tea; Speciality Tea Leaflets; Fact File on Tea;* available from The Tea Council Limited, Sir John Lyon House, 5 High Timber Street, London EC4V 3NJ, United Kingdom.
2. International Tea Committee: *Annual Statistical Bulletin.*
3. W. Ukers: *All About Tea,* Vols I and II, Ukers, New York 1935.
4. D. Forrest *The World Tea Trade,* Woodland Faulkner, Cambridge 1985.
5. J. Weatherstone *The Pioneers,* Quiller, London 1987.
6. Flammarion *The Book of Tea,* Flammarion, Paris 1991.
7. W. Ukers *The Romance of Tea,* Ukers, New York 1935.

Technetium → Radionuclides

Tellurium and Tellurium Compounds

GUY KNOCKAERT, Union Minière, Hoboken, Belgium

1. **Introduction** 1
2. **Properties** 1
3. **Resources and Raw Materials** 2
4. **Production** 3
5. **Quality Specifications** 5
6. **Analysis** 6
7. **Tellurium Compounds** 6
7.1. **Inorganic Compounds** 6
7.2. **Organic Compounds** 7
8. **Uses** 8
9. **Economic Aspects** 10
10. **Toxicology and Occupational Health** . 10
11. **References** 11

1. Introduction

Tellurium [*13494-80-9*], Te, atomic number 52, A_r 127.60 ± 0.03, is a metallic element of group 16 in the periodic table.

Tellurium has 8 natural stable isotopes, and is known to have a total of 21 artificial unstable isotopes. The natural abundance of the stable isotopes is: ^{120}Te (0.096 %), ^{122}Te (2.60 %), ^{123}Te (0.908 %), ^{124}Te (4.816 %), ^{125}Te (7.14 %), ^{126}Te (18.95 %), ^{128}Te (31.69 %), and ^{130}Te (33.80 %) [1]. The abundance of tellurium in the Earth's crust is comparable to that of platinum, 0.01ppm.

Tellurium was discovered in 1782 by MÜLLER VON REICHENSTEIN and named by KLAPROTH in 1798 (Latin, *tellus,* Earth) [2].

2. Properties [1], [3–12]

Physical Properties. Tellurium is a crystalline, bright silver-white metal, which is rather brittle and easily crushed. Tellurium is relatively soft (2.3 Moh, nearly as hard as zinc).

The crystal structure is a three-point helical lattice with hexagonal symmetry, with the helical chains parallel to the hexagonal axis. Many physical and electrical properties, e.g., hardness, thermal expansion coefficient, Hall coefficient, resistivity, and diamagnetic susceptibility are anisotropic. The so- called amorphous phase actually contains very small hexagonal crystals.

Tellurium, a *p*-type semiconductor, demonstrates the phenomenon of piezoelectricity and becomes superconductive at 3.3 K. Some tellurium compounds have excellent thermoelectric properties which are commercially interesting. Trace additives of tellurium to steel, lead, and copper and their alloys have a positive influence on the machinability of tools.

Tellurium has a relatively low melting (723 K) and boiling point (1327 K). On melting, the specific volume increases by ca. 5 %. Tellurium vapor is yellow-gold and consists mainly of Te_2 molecules up to 2000 °C.

Some physical properties of tellurium are:

Electronic configuration	[Kr]$4d^{10}5s^2 5p^4$
A_r	127.60 ± 0.03
Thermal neutron cross section at 2200 m/s absorption	$(4.7 \pm 0.1) \times 10^{-28}$ m^2/atom
scattering	$(5 \pm 0.1) \times 10^{-28}$ m^2/atom
Crystal structure	hexagonal lattice with trigonal symmetry (A 8) $a_0 = 0.4457$ nm $c_0 = 0.5929$ nm bond angle $(103.2 \pm 0.1)°$
Atomic radius	0.1285 nm
Ionic radius	
Te^{4+}	0.089 nm
Te^{2-}	0.221 nm
Atomic volume	20.42×10^{-6} m^3/mol
Density at 300 K	6.245 g/cm^3
mp	722.6 K
bp	1327 K
Enthalpy of fusion	17.489 kJ/mol
Enthalpy of vaporization (Te_2)	107.77 kJ/mol
Specific heat at 298.15 K	25.707 J mol^{-1} K^{-1}
Entropy at 298.15 K	49.497 J mol^{-1} K^{-1}
Entropy of fusion	24.201 J mol^{-1} K^{-1}
Entropy of vaporization	36.6 kJ mol^{-1} K^{-1}
Coefficient of linear expansion, mean value (anisotropic)	16.8×10^{-6} K^{-1}
Vapour pressure *p*	
at 793 K	0.133 kPa
at 923 K	1.33 kPa
at 1111 K	13.3 kPa
Thermal conductivity at 293 K	0.060 W m^{-1} K^{-1}
Surface tension at temperature T, T_{mp} = melting point	$0.178 - 2.4 \times 10^{-5}\,(T - T_{mp})$N/m

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_177

Electronic configuration	$[Kr]4d^{10}5s^2 5p^4$
Electrical resistivity	
at 3.3 K	superconducting
at 300 K	9.9×10^{-3} Ωm
Standard electrode potentials	
$Te + 2e^- \longrightarrow Te^{2-}$	− 0.92 V
$Te^{4+} + 4e^- \longrightarrow Te$	0.63 V
Tensile strength	11.0 ± 0.25 MPa
Modulus of elasticity	4140 MPa

Chemical Properties. Metallic tellurium has many properties similar to sulfur and selenium, but it is less reactive, more basic, more metallic, and strongly develops amphoteric properties. When metallic tellurium is heated in air, it burns with a blue-green flame; tellurium powder oxidizes at room temperature.

Tellurium reacts vigorously with halogens at room temperature, cannot be combined directly with sulfur, and reacts with hydrogen at high temperature (920 K). Tellurium does not react with carbon, boron, or nitrogen; it reacts with phosphorus only in a sealed tube heated above 595 K.

Tellurium is insoluble in water, dissolves in concentrated sulfuric and nitric acid, and only slightly in dilute hydrochloric acid. The solubility in caustic alkalis depends on the temperature; tellurium is, however, insoluble in ammonium hydroxide.

When heated in an evacuated ampule, stoichiometric amounts of tellurium and one or more other elements form binary or ternary tellurides. Molten tellurium corrodes iron, copper, and stainless steel.

3. Resources and Raw Materials [13]

Tellurium, a relatively rare element with a crustal abundance of 10 µg/kg, is found in close association with sulfur as well as with selenium. Unlike the latter, tellurium does not substitute in sulfide lattices, but forms discrete minerals or microsegregations in the host sulfide mineral [14].

Native tellurium has been observed, but the element is mainly found as a gold and/or silver telluride. The concentration of tellurium minerals in nature is insufficient to allow their economic recovery as principal mining products; its recovery therefore depends on its concentration during the processing of other nonferrous metals. In copper refining, tellurium accumulates together with precious metals and selenium in the anodic slimes generated by electrolysis. Other primary metals such as zinc, gold, and lead, concentrate tellurium during refining.

Reserves of tellurium are difficult to assess because of the limited knowledge of tellurium content in the copper or other ores from which it is recovered. Estimations of refinery production of tellurium are complicated by the trade in concentrates, blister, and anode copper, as well as other nonferrous residues.

As copper ores are the main sources for tellurium, public statistics are based on the copper industry. The United States Bureau of Mines estimates tellurium resources by applying fixed recovery factors (i.e., 0.065 kg Te per tonne of Cu) and quotes the figures in tonnes [15].

Continent, country	Tellurium reserves, t
North America	
United States	6 000
Canada	2 000
Others	3 000
South America	
Chile	6 000
Peru	2 000
Others	1 000
Europe	5 000
Africa	
Zaire	2 000
Zambia	2 000
Others	1 000
Asia	
Philippines	1 000
Others	1 000
Oceania	2 000
World	34 000

The results of these estimations do not include tellurium resources included in lead, zinc, or gold reserves.

4. Production [16–22]

Tellurium can be recovered as a byproduct in the treatment of lead, copper, bismuth, precious metals, and nickel ores, and from sulfuric acid plants. The main source is copper anode slimes, which settle at the bottom of the refining tanks during the electrowinning of 99.9 % pure cop-

per. Copper anode slimes contain 0.5 – 10 % tellurium (Table 1).

Table 1. Composition of copper anode slimes: selected examples [23–25]

Element	Typical concentration, wt %		
	Canadian Copper Refiners	Nippon Mining	Inco
Cu	20.3	4.73	17
Bi	0.36	1.6	0.1
Sb	0.95	0.95	0.05
Se	10.9	15.23	7
Te	3.19	3.64	2
Pb	8.5	6.54	1
Au		0.99	0.1
Ag	21.3	20.58	6
As	1.83	1.59	0.8
Ni	0.52	0.03	26

Tellurium is collected with the precious metals in the anode slimes, and has to be separated from these together with selenium, because they interfere with the separation and refining of the precious metals. Tellurium is mostly present in the form of intermetallic compounds of silver, copper, and sometimes gold [Ag_2Te, Cu_2Te, and $(Ag, Au)Te_2$]. Pretreatment of the slimes can dissolve the copper which is harmful in the further treatment of the slimes. It is normally followed by roasting or sulfatizing of the decopperized slimes, or a combination of these and alternative methods.

Pretreatment consists of dissolving residual copper and tellurium by dilute aerated sulfuric acid or by oxidative pressure leaching with dilute sulfuric acid (normally tankhouse liquid from the copper electrolysis). The first method dissolves 70 – 80 %, the second > 90 % of the tellurium. Pressure leaching takes place in an autoclave under an oxygen pressure of 250 – 350 kPa at 80 – 160 °C. Depending on the conditions, tellurium is converted to the tetravalent or hexavalent form. Tests have been performed at oxygen pressure up to 1000 kPa whereby copper and tellurium are completely dissolved, together with part of the contained silver and selenium, which, after contact with a reducing agent (e.g., SO_2) at atmospheric pressure, reprecipitate [23]. The dissolved tellurium in the filtrate is consequently recovered as copper telluride (Cu_2Te) by cementation with copper above 80 °C. This can be performed in a rotating drum containing copper shot or in a fixed-bed reactor filled with copper chippings [26]. Other techniques use copper electrodes suspended in a bath containing the dissolved copper and tellurium sulfate, agitated by vibration or alternating current to the electrode to remove the deposited copper telluride [27], [28]. Figure 1 illustrates a possible pretreatment scheme.

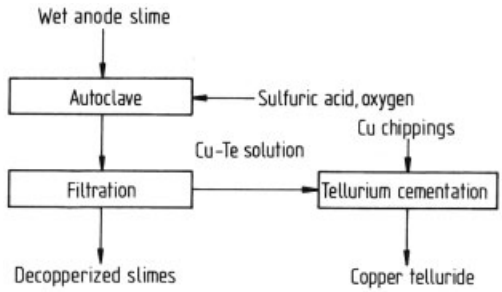

Figure 1. Anode slime pretreatment scheme

Roasting. Pretreated decopperized slimes are roasted with soda ash, whereby tellurium is converted to insoluble sodium tellurite and tellurate:

$$Ag_2Te + Na_2CO_3 + O_2 \longrightarrow 2\,Ag + Na_2TeO_3 + CO_2 \ (400 - 500\,^\circ C)$$

$$Na_2TeO_3 + 1/2\,O_2 \longrightarrow Na_2TeO_4 \text{ (insoluble) } (> 550\,^\circ C)$$

Normally, complete conversion to hexavalent tellurium is preferred with roasting temperature 550 – 650 °C. Other processes operate below 400 °C or up to 750 °C, involving different process steps. After water leaching, selenium is recovered as soluble sodium selenate. The residue containing tellurium (as sodium tellurate), lead, and precious metals is either sent to a doré furnace, or treated with sulfuric acid:

$$Na_2TeO_4 + H_2SO_4 \longrightarrow H_2TeO_4 \text{ (soluble)} + Na_2SO_4 \ (95\,^\circ C)$$

In this processing step, the residue containing lead and precious metals is separated for further treatment. The telluric acid is reduced to tellurium by treatment with hydrochloric acid and sulfur dioxide, or reduced by sodium sulfite to tellurium dioxide, or precipitated as copper telluride by adding copper. If there is no preliminary sulfuric acid treatment, the residue still containing the sodium tellurate is smelted in a

doré furnace, where the tellurium is collected in a soda slag leached with water to convert it to sodium tellurite. It is then neutralized with sulfuric acid to tellurium dioxide. Copper telluride and tellurium dioxide can be leached with dilute sodium hydroxide (with aeration in the case of copper telluride) to form a sodium tellurite solution which is further prepared for electrolysis. Figure 2 shows a soda ash roasting process combined with a doré furnace.

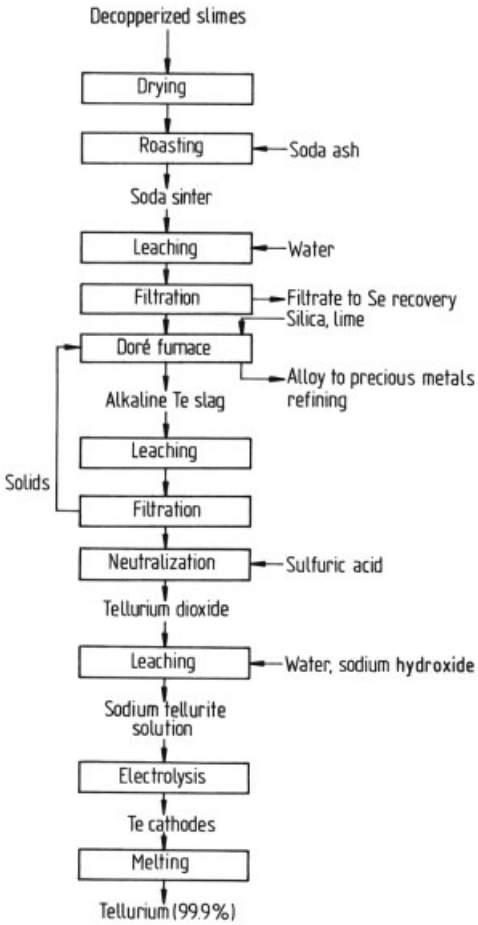

Figure 2. Soda ash roasting and doré furnace treatment

Sulfatizing takes place in two stages: below 350 °C the sulfuric acid forms metal sulfates with the proper consistency to be roasted above 400 °C in the second stage. In this stage, selenium dioxide is eliminated. For tellurium, the following reactions take place:

$$Ag_2Te + 3\,H_2SO_4 \longrightarrow Ag_2SO_4 + TeSO_3 + SO_2 + 3\,H_2O \quad (150-200\,^\circ C)$$

$$3\,TeSO_3 \longrightarrow TeO_2 \cdot SO_3 + 2\,Te + 2\,SO_2 \; (>400\,^\circ C)$$

$$TeO_2 \cdot SO_3 \longrightarrow TeO_2 + SO_3 \; (>430\,^\circ C)$$

The tellurium dioxide remains in the sulfated slimes, which are leached with water. Part of the tellurium is dissolved, and can be cemented with copper to yield copper telluride. The leach residue can be treated in the doré furnace. Alkaline leaching of the residues yields soluble sodium tellurite, but this process does not recover all the tellurium. Figure 3 shows a sulfatizing roasting process for the recovery of tellurium.

Other Techniques. The oxidation of slimes under pressure in alkaline solution can be used to give good separation of tetravalent selenium and hexavalent tellurium after oxidation. Sodium hydroxide concentration is 100 – 500 g/L, the temperature is ca. 200 °C, and the oxygen partial pressure is 170 – 1700 kPa. Pressure leaching with oxygen or hydrogen peroxide is also possible [29], [30]. Aqueous or dry high-temperature chlorination of slimes has been developed. Dry chlorination being difficult to control, wet chlorination is preferred. The slimes are sparged with chlorine gas or sodium chlorate at ca. 100 °C, and all selenium and tellurium is recovered in the soluble form.

The recovery of tellurium from secondary materials accounts for 10 – 20 % of world tellurium production. International Recoveries of the Philippines started producing tellurium dioxide and tellurium metal from copper telluride raw material in 1984, and other complex residues and waste products are now treated. Treatment of a tellurium tetrachloride residue was described in 1989[31]. Small quantities of thermoelectric scrap are recycled by different tellurium producers.

Purification. The following reactions take place during the electrolysis of a tellurite solution:

Anode:

$$4\,OH^- \longrightarrow 2\,H_2O + O_2 + 4\,e^-$$

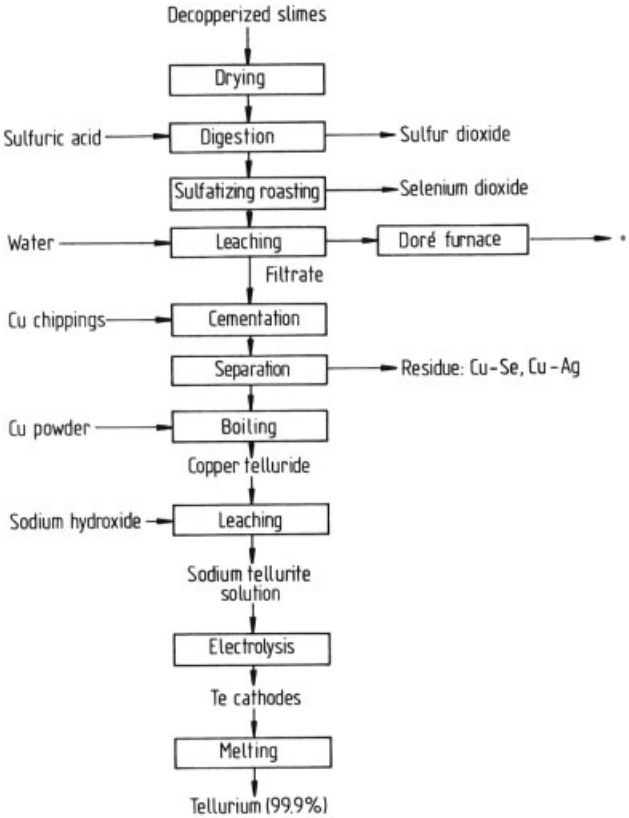

Figure 3. Sulfation roasting process
∗ As in Figure 2

Cathode:

$$TeO_3^{2-} + 3\,H_2O + 4\,e^- \longrightarrow Te + 6\,OH^-$$

Typical conditions in an electrolysis cell are: tellurium 100 – 200 g/L; free sodium hydroxide 40 g/L; current density 160 A/m^2; temperature 40 – 50 °C. After drying, melting, and casting in different shapes (or crushing to powder) the commercial tellurium grade is ready for sale. Direct precipitation with zinc is also possible instead of electrowinning.

Further refining consists of vacuum distillation and zone refining, and yields the higher-purity grades 99.99 – 99.9999 %. As zone refining is unsatisfactory for some impurities, e.g., aluminum, bismuth, and iron, another process uses chlorination with HCl to tellurium tetrachloride; distillation and hydrolysis leads to the precipitation of tellurium dioxide of > 99.999 % purity, which after reduction with hydrogen gas gives tellurium of the same purity [32].

Table 2. Detection limits obtained by GFAAS, SSMS, and ES, µg/kg

Element	GFAAS	SSMS	ES
Ag	10	1.0	100
Al	50	0.8	300
Bi	50	2.0	100
Cu	50	∗	200
Fe	50	1.0	200
Mg	20	19	100
Ni	20	1.6	300
Pb	50	1.0	200
Sb	50	1.0	300
Sn	50	4.0	100

∗ Interference by doubly charged ions.

5. Quality Specifications

Qualities ranging from 99.5 % to 99.9999 % tellurium are commercially available. Impurities

may be metallic (As, Na, Ca, Pb, Cu, Ni, Sb, Sn, Se, Fe) and nonmetallic (S, Cl, O).

Commercial grades and forms are: 99.5 – 99.9 % (2N5-3N) powder, sticks, or ingots; 99.99 % (4N) ingots or lumps; 99.999 % (5N) ingots; and 99.9999 (9) % (6N-7N) ingots.

Tellurium dioxide powder is available with purity 99.5 – 99.9 %.

6. Analysis (Table 2) [13], [33], [34]

Emission spectrometry (ES) is frequently used for the analysis of standard grade materials up to a purity of 99.99 – 99.999 % (with preconcentration if necessary). Atomic absorption spectrometry (AAS) is used for analysis down to subppm levels; this method has, however, no multielement capability. AAS can be adapted with a graphite furnace (GFAAS) and coupled with a Zeeman background correction to allow detection limits at µg/kg levels.

Techniques such as spark source mass spectrometry (SSMS) and glow discharge mass spectrometry (GDMS) offer panoramic elemental capability with low detection limits and maximum selectivity.

7. Tellurium Compounds [5], [11], [35–37]

7.1. Inorganic Compounds

Tellurium dioxide, TeO_2 [*7446-07-3*] is found in the minerals tellurite and paratellurite, which have a rhombic and a tetragonal lattice structure, respectively. It can be formed by burning tellurium in air, or by treating tellurium powder with nitric acid, followed by heating the nitrate:

$$2\,Te + 9\,HNO_3 \longrightarrow 2\,TeO_2 \cdot HNO_3 + 8\,NO_2 + 4\,H_2O$$

The melting point of TeO_2 is 1005.8 K (732.6 °C), the clear, dark red liquid oxide vaporizes between 790 and 940 °C. Tellurium dioxide is very soluble in water and is amphoteric, the minimum solubility is at pH 3.8 – 4.2, the isoelectric point. Tellurium dioxide is an important catalyst in oxidation, hydrogenation, and dehydrogenation processes; it is also used in chalcogenide glasses, and as a vulcanizing agent for rubber. Single crystals of TeO_2 are used in acousto-optic detectors and modulators. Other oxides of tellurium are tellurium monoxide, TeO [*13451-17-7*], tellurium trioxide, TeO_3 [*13451-18-8*], and tellurium pentoxide Te_2O_5 [*12036-81-6*]; TeO_3 exists in a yellow α-form and a grey β-form.

Tellurium Halides. In tellurium halides, Te has valency 2, 4, or 6. Not all the halides are formed – TeI_2 and the Te_2X_2 halides do not exist. *Tellurium tetrachloride*, $TeCl_4$ [*10026-07-0*] forms white hygroscopic needle-like crystals which are soluble in benzene, nitrobenzene, toluene, ethylacetate, and methanol. It is used as a starting material for the synthesis of many organic tellurium compounds, and is also a catalyst for the chlorination of phenol and benzene, and the hydrogenation of acids and esters to alcohols. *Tellurium tetrafluoride*, TeF_4 [*15192-26-4*] forms white hygroscopic needles. It reacts with water, glass, or silica to tellurium dioxide; metal tellurides are formed with copper, silver, gold, and nickel at 185 °C. Platinum is not attacked below 300 °C.

Hydrogen telluride, H_2Te [*7783-09-7*] is a colorless, unstable gas which decomposes above 0 °C and is prepared by electrochemical reduction of tellurium, or by the addition of hydrochloric acid to Al_2Te_3. It is used in the production of heavy metal tellurides.

Tellurous acid, H_2TeO_3 [*10049-23-7*] is obtained by reaction of alkali tellurites with nitric acid, or by hydrolysis of a tellurium tetrahalide. *Orthotelluric acid*, H_6TeO_6 [*7803-68-1*] is made by oxidizing tellurium or tellurium dioxide. It is used in some chemical processes. *Sodium tellurite*, Na_2TeO_3 [*10102-20-2*] improves the corrosion resistance of electroplated nickel layers. Solutions of *sodium tellurate*, Na_2TeO_4 [*10101-83-4*] containing Cu^{2+} ions are used for black or blue-black coatings on iron, steel, aluminum, and copper.

A large number of *metal tellurides* are known, and many are semiconductors (Table 3). The tellurides are used in infrared detectors ($Cd_xHg_{(1-x)}Te$, PbTe, and $Pb_{(1-x)}Sn_xTe$) and laser diodes (PbTe and $Pb_{(1-x)}Sn_xTe$), thermo-

Table 3. Important tellurides and their properties

Formula	Crystal type	Density, g/cm^3	Band gap at 300 K, eV	Carrier mobility at 300 K, cm^2 V^{-1} s^{-1}		Melting point, °C
				μ_n	μ_p	
CdTe	cubic	5.86	1.5	1 000	95	1092
HgTe	cubic	8.2	0.15	25 000	350	670
PbTe	cubic	8.16	0.29	1 800	900	924
ZnTe	cubic	5.7	2.26	530	900	1238
Bi_2Te_3	hexagonal	7.86	0.15	1 300	1200	588.5
Sb_2Te_3	hexagonal	6.52	0.24		360	621.5
GeTe	face-centered cubic	6.2	0.1		180	724

electric elements (PbTe, Bi_2Te_3, Sb_2Te_3, GeTe), solar cells (CdTe), and γ-ray detectors (CdTe).

7.2. Organic Compounds (Fig. 4)

The search for organic tellurium compounds and their use in chemical synthesis have developed quickly during the last few decades. In organic compounds tellurium usually has valency 2 or 4. Organotellurium compounds are thermally less stable, and are oxidized more rapidly by air (at room temperature) than organosulfur or organoselenium compounds. The names and formulas of organotellurium compounds are as follows:

Compound	Formula
Tellurol, tellane	R−Te−H
Dialkyl(diaryl)(mono,di,tri) tellurides	$R-Te_{1,2,3}-R$
Tellurenyl	R−Te−X
Alkyl(aryl)tellurium halides	R_n-TeX_n
Tellurinic acid	RTeOOH
Dialkyl(diaryl)tellurium oxide	R_2TeO
Tellurone	R_2TeO_2
Tetraalkyl(tetraaryl) telluride	R_4Te
Dialkyl(diaryl)tellurium ketone	R_2TeCR_2

These compounds have many applications in the pharmaceutical industry. Organic tellurium compounds containing the tellurium isotope ^{123m}Te are used as imaging agents for organs. Tellurium diethyldithiocarbamate (TDEC) has been used extensively in the rubber industry, and is also used in copper electroplating. Tellurofulvalenes are organic conductors [38]. Volatile organic tellurium compounds are used in organometallic vapor phase epitaxy for the manufacture of semiconductors.

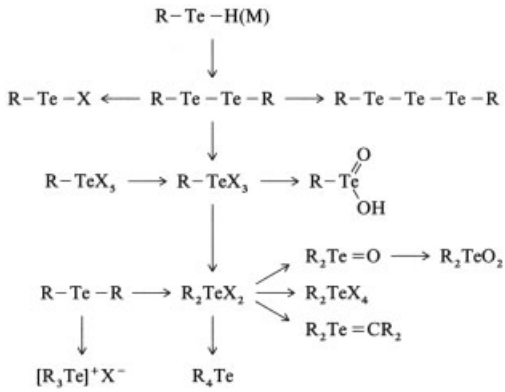

Figure 4. Known types of organic tellurium compounds with at least one carbon – tellurium bond (reproduced from [39], with permission)

In organic synthesis, organotellurium compounds are used as mild and highly selective reducing agents, e.g., sodium(hydrogen) telluride, aryltellurols, and tellurides. Other applications are tellurium-mediated formation of anionic species which then react with electrophiles (e.g., Reformatsky-type reactions with sodium telluride), and the preparation of organolithium reagents by lithium/tellurium exchange [40].

8. Uses

Consumption of tellurium (in t) according to end use in the western world in 1990 was estimated as follows [41]:

Steel additives	145
Nonferrous metals additives	55
Chemicals	45
Electronics	20
Others	5
Total	270

The machinability of steel can be improved by the addition of small quantities (up to 0.1 %) of tellurium [42–45], added during the refining of leaded and/or resulfurized, or low- carbon steel in the form of powder, sticks, or cored wire. Manganese telluride alloys are most frequently used nowadays. Additions to cast iron help control the chill depth of rapidly cooled surfaces and produce hard, wear-resistant top layers.

Tellurium steels and chilled castings are used in the automotive industry, in industrial machinery and tools, in mining, and in railroad equipment.

Many other positive effects of tellurium additions, e.g., deoxidation of liquid steel, refining of the grain size of castings, improvement of the magnetic anisotropy in electrical steels, and improvement of machinability of powder metallurgy steel have been reported.

The use of tellurium in steel could be affected because of decreased consumption of leaded steels owing to environmental constraints.

Traces of tellurium are added to many nonferrous metals [42], [44–46]. The tellurium content is mostly less than 1 %. Tellurium – copper (which also contains some phosphorus) reduces cutting resistance, tool wear, heat of cutting, and chip size. It is used in forging, automotive radiators, vacuum applications, and as electrical, motor, and switch parts. Tellurium – copper can replace lead – copper in plumbing, as the latter is under environmental pressure in some countries.

Tellurium – lead is used in the sheathing of power and marine cables and in chemical equipment piping (mostly where sulfuric acid is used). In lead acid batteries, tellurium helps strengthen the battery grids. Other nonferrous alloys are tellurium bronze, tellurium – tin, tellurium – chromium, and tellurium – magnesium. Tellurium – cobalt – titanium is used in permanent magnets.

Tellurium is used in combination with other components in many catalysts. The chemical reactions involved are oxidation, ammoxidation, hydrogenation and dehydrogenation, halogenation and dehalogenation, and phenol condensation [47]. The two major industrial applications are the Nitto process: (acrylonitrile ammoxidation) and the Mitsubishi process (tetramethylene glycol synthesis) [48]:

Nitto: acrylonitrile ammoxidation catalyst:

$$C_3H_6 + NH_3/\text{air} \xrightarrow{\text{Sb-Mo-U-Te}} C_2H_5CN$$

Mitsubishi: tetramethylene glycol synthesis (Ac = acetate):

$$C_4H_6 + \text{HOAc/air} \xrightarrow{\text{Pd-Te}} \text{AcO}-CH_2-CH=CH-CH_2\text{OAc} \xrightarrow{H_2/H_2O} \text{HO}-CH_2-CH_2-CH_2-CH_2-\text{OH} \xrightarrow{-H_2O} (CH_2)_4O \text{ (tetrahydrofuran)}$$

Tellurium is used as a secondary vulcanizing agent in natural, styrene – butadiene, and isobutylene – isoprene rubber [42], [44], [49]. The tellurium rubbers have improved aging and mechanical properties, and better resistance to heat and abrasion. They are used in portable cables, automobile tires, door and window seals, double glazing, and in conveyor belts for special applications. The use of tellurium in rubber has declined over the past few decades, and there is concern over the toxicity of tellurium handling in rubber production, which could further affect its usage.

Other uses of tellurium in chemistry are in heat- and vacuum-stable lubricants for electronics and aerospace applications, metal coatings for silverware, aluminum, and brass, in glass, pigments and fungicides, and in organic derivatives and radioactive isotopes for medicine and biology. Lately, the role of tellurium as a possible regulator in cholesterol synthesis has been studied [50].

In electronics, tellurium is a well-known additive in selenium(-arsenic) photoreceptors [51], [52]. Tellurium has the property of increasing the sensitivity and broadening the spectral response of the photoreceptor. Since organic photoconductors (OPCs) have been taking over the photocopier and laser printer market, starting with the slower machines (up to 20 – 30 pages per minute), the use of selenium – tellurium alloys is decreasing markedly, especially in the industrialized countries.

Polycrystalline cadmium telluride is used in photovoltaic solar cells [53]. CdTe has a direct band gap of 1.5 eV, which is optimal for solar energy conversion, and a high optical absorption coefficient. Research into higher efficiencies, lower cost, and better reliability has intensified, and the future of CdTe solar cells looks

very promising. By the year 2000, the use of tellurium in solar cells could be more than 50 t/a [54].

Bismuth telluride and lead telluride are thermoelectric materials [55], [56]. Thermocouples of these materials convert electricity into heat (Peltier effect) and heat into electricity (Seebeck effect). In these thermocouples, dopants such as Sb and Se are needed to make *n*- and *p*-type materials.

Bi_2Te_3 is mostly used below 200 °C, PbTe at 200 – 500 °C. The efficiency of a thermoelectric material is expressed by its figure of merit Z, a function of the thermal and electrical conductivity and the Seebeck coefficient.

Thermoelectricity has found many applications in picnic boxes, water and food coolers, fiber optic circuits, blood analyzers, and other precision temperature control instruments, remote power generators, solar energy conversion, military equipment (e.g., refrigeration systems in submarines and infrared detectors), and aerospace. Especially the low-tech applications, such as picnic boxes and car drink coolers, have seen a steady rise in sales.

Now that systems based on chlorofluorocarbons (CFCs) are increasingly banned from refrigerators and air- conditioning devices, owing to their detrimental effect on the ozone layer, thermoelectricity could offer a viable alternative, but the present efficiency of the coolers is still too low and the cost too high to suggest a rapid growth in large cooling systems [57].

Optical disks are powerful tools for storing vast quantities of digitized information (up to 1 Gbyte). Different types of write-only and erasable disks exist, and some of their active layers use tellurium alloys which are sputtered onto the disk by vapor deposition techniques from targets containing, e.g., Te–Ge–Sb or Te–Se–Sb [58].

Cadmium-, cadmium zinc-, and cadmium selenium telluride single crystals are used in several electronic applications: γ-ray detectors for medical diagnostics and dosimetry, nonlinear optics, infrared electro-optic modulators, and photorefractive devices [34], [59]. These crystals are grown by vertical or horizontal Bridgman techniques, or the traveling heater method (THM). They are also used as substrates for epitaxial cadmium mercury telluride ($Cd_xHg_{(1-x)}Te$) thin film detectors for infrared imagers [60], of which lead tin telluride ($Pb_xSn_{(1-x)}Te$) is an example. The detectors operate in the spectral ranges 3 – 5 µm and 8 – 14 µm. The thin films are grown by methods such as liquid phase epitaxy (LPE), organometallic vapor phase epitaxy (OMVPE), or molecular beam epitaxy (MBE) (→ Thin Films) [34]. The devices are normally cooled to 77 K with liquid nitrogen. The new generations of imagers use two-dimensional focal plane arrays to capture an entire image at once.

Figure 5 shows how the research and development literature focuses on the different uses of tellurium in nonferrous alloys, optical storage, and cadmium- and cadmium mercury tellurides.

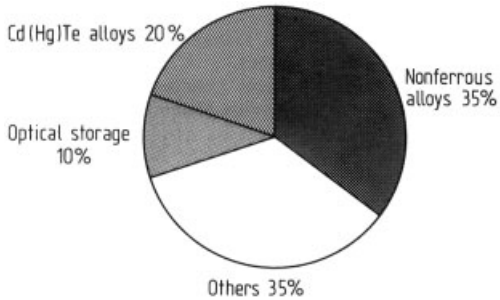

Figure 5. Tellurium in scientific publications: main areas of interest in *Chemical Abstracts* (1991)

9. Economic Aspects

World production of tellurium in 1990 has been estimated at 306 t. The estimated production (in t/a) by country and company for 1990 is listed below:

Japan	
Mitsubishi Metal	22
Mitsui Mining & Smelting	10
Nippon Mining & Smelting	12
Sumitomo Metal Mining	12
Belgium	
Union Minière	90
Canada	
Noranda Mines	15
United States	
Asarco	40
Philippines	
Pacific Rare Metal	35
Soviet Union	50
Peru	10
China	10
Total	*306*

The main producers are in Belgium, the United States, the Commonwealth of Independent States, and the Philippines [44].

Figure 6 shows the evolution of the average annual price for standard-grade tellurium (99.5 %) [44]. Tellurium 99.99 % is normally 40 – 50 % more expensive, 99.999 % is more than double the standard-grade price, and ≥ 99.9999 % ranges from 400 to 1200 $/kg.

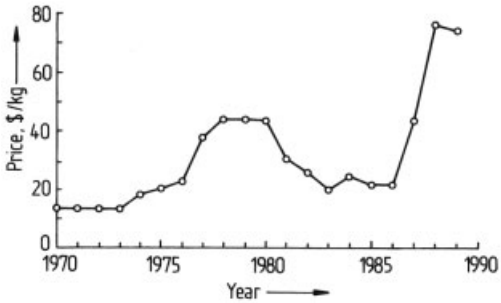

Figure 6. Average annual price for standard-grade tellurium (99.5 %)

The world economic recession of the early 1990s, particularly hitting the steel industry, and cheap material from the Commonwealth of Independent States, led to a price erosion of 50 % between 1990 and 1993.

10. Toxicology and Occupational Health [61–65]

Elemental tellurium is considered to be less toxic than selenium. Organic compounds and reactive tellurides can be a health hazard. Hydrogen telluride (H_2Te) and tellurium hexafluoride (TeF_6) are both highly toxic, colorless gases. Some LD_{50} values on ingestion of tellurium compounds are listed in Table 4. On the basis of new data for the LD_{50} value for elemental tellurium, the Environmental Protection Agency of the United States has been asked to declassify tellurium as extremely hazardous [65].

Table 4. LD_{50} values for different tellurium compounds

Compound	Animal	Administration	LD_{50}, mg/kg
Te	rat, mouse	oral	> 5000
TeO_2	unspecified	oral	> 3000
Na_2TeO_3	white mouse	oral	20
Na_2TeO_4	white mouse	oral	165

The acute inhalation toxicity value LC_{50} (4 h) for rats is > 2.42 g/m^3, the maximum attainable concentration [66].

Tellurium intoxication in the working environment can occur through inhalation and ingestion. No irritation of skin or eyes has been reported for elemental tellurium and tellurium dioxide. Exposure to tellurium fumes and dust cause a garlic odor of the breath and malaise, dryness of the mouth and metallic taste, anorexia, occasional nausea, and some other symptoms. Tellurium is accumulated in the liver, kidneys, bones, and neural cells. The elimination of tellurium in breath, sweat, urine, and feces is very slow. The typical garlic odor seems to result from the partial elimination of tellurium through the lungs as dimethyl telluride $(CH_3)_2Te$; it can be suppressed temporarily by administration of vitamin C (ascorbic acid), but the vitamin C may also enhance the toxic effects. Tellurium hexafluoride can cause a bluish-black pigmentation of the skin.

Tellurium intoxication symptoms do not appear when the tellurium concentration in the air is < 0.01 mg/m^3 and in the urine < 1 µg/L. The TLV for tellurium and its compounds is 0.1 mg/m^3 (calculated as tellurium). The TLV was established to prevent poisoning, but garlic breath may still occur. There is no specific antidote for tellurium poisoning.

There are no indications of carcinogenic, teratogenic, or mutagenic effects of tellurium or its compounds.

Unlike selenium, tellurium is not an essential trace element for humans or animals. Tellurium and its compounds cannot be used in cosmetics.

11. References

1. *Gmelin,* **Suppl. A 1,** p. 205.
2. S. C. Carapella, Jr., in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, p. 1.
3. *Ullmann,* 4th ed., **22,** 447.
4. *Handbook of Chemistry and Physics,* 66th ed., The Chemical Rubber Co., Cleveland, Ohio 1985/86.
5. J. C. Bailar et al.: *Comprehensive Inorganic Chemistry,* 1st ed., Pergamon Press, Oxford 1973, pp. 935 – 1009.
6. *Kirk-Othmer,* **22,** 659 – 661.

7. S. C. Carapella, Jr.: *Metals Handbook,* 9th ed., **vol. 2,** American Society for Metals, Metals Park, Ohio 1979, p. 806.
8. W. M. Becker, V. A. Johnson, A. Nussbaum in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 54 – 109.
9. D. M. Chizhikov, V. P. Shchastlivyi: *Tellurium and Tellurides,* Collet's Ltd., London 1970, pp. 1 – 30.
10. H. Lumbroso in P. Pascal (ed.): *Compléments au Nouveau Traité de Chimie Minérale,* **vol. 8,** Masson, Paris 1977, pp. 17 – 68.
11. W. A. Dutton in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 110 –183.
12. I. Barin: *Thermochemical Data of Pure Substances,* part II, VCH Verlagsgesellschaft, Weinheim 1989.
13. M. Caffarey, *Proc. 8th Conf. on Thermoelectric Energy Conversion,* Nancy, France 1989, pp. 40 – 44.
14. W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 1 – 13.
15. D. Edelstein in S. C. Carapella, Jr. (ed.): *Proc. 4th Int. Symp. on Uses of Selenium and Tellurium,* Banff, Canada 1989, pp. 4 – 13.
16. P. E. Skinner, *Trans. Inst. Min. Metall. Sect. C* **97** (1988) C 83 – C 87.
17. P. H. Jennings in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 14 – 53.
18. *Kirk-Othmer,* **22,** 662 – 663.
19. R. Bresee, D. Vleeschhouwer, J. Thiriar, *Proc. of the Symp. on Industrial Uses of Selenium and Tellurium,* Toronto, Canada 1980, pp. 31 – 49.
20. J. E. Hoffmann, *J. Met.* **41** (1989) no. 7, 32 – 38.
21. J. E. Hoffmann, *J. Met.* **42** (1990) no. 8, 50 – 54.
22. *Ullmann,* 4th ed., **22,** 448 – 451.
23. Noranda Inc., CA 2 049 276, 1991 (P. L. Claessens, C. W. White).
24. Nippon Mining Co., JP 60 208 431, 1985 (T. Abe, Y. Takazawa).
25. J. E. Dutrizac, T. T. Chen in W. C. Cooper, G. E. Lagos, G. Ugarte (eds.): *Copper 87,* **vol. 3,** Universidad de Chile, Santiago, Chile 1988, p. 469.
26. T. Shibasaki, K. Abe, H. Takeuchi, *Hydrometallurgy* **29** (1992) 399 – 412.
27. Mitsubishi Metal Corp., Onahama Seiren Co., JP 01 079 008, 1989 (Y. Sugawara, J. Konishi, S. Hayashi, M. Hayashi).
28. Mitsubishi Materials Corp., US 5 160 588, 1992 (T. Shibazaki, K. Abe, H. Takenchi).
29. Nippon Shin Kinzuko K.K., JP 56 084 428, 1981.
30. Interox Chemicals, EP 127 357, 1984 (A. Broome).
31. R. J. Hisshion in S. C. Carapella, Jr. (ed.), *Proc. 4th Int. Symp. on Uses of Selenium and Tellurium,* Banff, Canada 1989, pp. 25 – 30.
32. Sumitomo Metal Mining Co., JP 63 285 106, 1988, JP 63 286 531, 1988 (M. Yukinobu, J. Oowa).
33. K. Swenters, J. Verlinden, R. Gijbels, *Fresenius Z. Anal. Chem.* **335** (1989) 900 – 904.
34. D. R. Nichols, B. E. Dean, C. J. Johnson in S. C. Carapella, Jr. (ed.): *Proc. 4th. Int. Symp. on Uses of Selenium and Tellurium,* Banff, Canada 1989, pp. 282 – 299.
35. *Ullmann,* 4th ed., **22,** 451 – 452.
36. D. M. Chizhikov, V. P. Shchastlivyi: *Tellurium and Tellurides,* Collet's Ltd., London 1970, pp. 145 – 259.
37. *The Economics of Tellurium 1991,* 4th ed., Roskill Information Services, London 1991, pp. 65 – 66.
38. D. O. Cowan et al., *Phosphorus Sulfur Silicon* **67** (1992) 277 – 294.
39. K. J. Irgolic in D. Klamann (ed.): "Methods of Organic Chemistry," *Houben-Weyl,* 4th ed., **Vol. E 12 b,** Thieme Verlag Stuttgart 1990, p. XXXIX.
40. N. Petragnagni, J. V. Comasetto, *Synthesis* **10** (1991) Oct., 793 – 817, **11** (1991) Nov., 897 – 919.
41. E. Hoyne, *The Bulletin of Selenium-Tellurium Development Association, Inc.,* Sept. 1989, pp. 1 – 4.
42. *Kirk-Othmer,* **22,** 671 – 676.
43. E. S. Nachtman in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 373 –409.
44. *The Economics of Tellurium 1991,* 4th ed., Roskill Information Services, London 1991, pp. 51 – 98.
45. R. J. Raudebaugh, *Proc. of the Symp. on Industrial Uses of Selenium and Tellurium,* Toronto, Canada 1980, pp. 201 – 221.
46. P. W. Taubenblat, *Proc. of the Symp. on Industrial Uses of Selenium and Tellurium,* Toronto, Canada 1980, pp. 267 – 276.
47. W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 410 – 430.
48. T. A. Koch in S. C. Carapella, Jr. (ed.) *Proc. 4th Int. Symp. on Uses of Selenium and Tellurium,* Banff, Canada 1989, pp. 661 – 672.

49. W. R. McWhinnie, J. E. Stuckey, K. G. Karnika Da Silva, *Proc. of the 3rd Int. Symp. on Industrial Uses of Selenium and Tellurium,* Stockholm, Sweden 1984, pp. 206 – 245.
50. A. D. Toews, S. Y. Lee, B. Popko, P. Morell, *J. Neurosci. Res.* **26** (1990) 501 – 507.
51. L. Cheung, G. M. Foley, P. Fournia, B. E. Springett, *Photogr. Sci. Eng.* **26** (1982) no. 5, 245 – 249.
52. R. C. Norton, *Proc. 7th Annual Photoreceptor & Components Conf.,* Santa Barbara, United States 1991.
53. K. Zweibel: "Toward Low Cost CdTe PV, NREL/TP-413-4841," National Renewable Energy Laboratory, Golden, CO., April 1992.
54. M. C. King, *Workshop Report: Recycling of Cadmium and Selenium from Photovoltaic Modules and Manufacturing Wastes,* Golden, United States 1992, pp. 90 –101.
55. C. H. Champness in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold Co., New York 1971, pp. 322 – 372.
56. D. P. Burton, *Proc. 9th. Int. Conf. on Thermoelectrics,* Pasadena, United States 1990, pp. 109 – 123.
57. P. M. Schicklin, *Proc. 9th Int. Conf. on Thermoelectrics,* Pasadena, United States 1990, pp. 381 –395.
58. M. Hartmann, B. A. Jacobs, *Philips Tech. Rev.* **42** (1985) no. 2, 37 – 47.
59. M. Ohmori, Y. Iwase, R. Ohno, *Proc. Symposium F on New Aspects of the Growth, Characterization and Applications of CdTe and Related Cd Rich Alloys of the 1992 E-MRS Spring Conf.,* Strasbourg, France 1992, pp. 283 – 290.
60. C. Lucas, *Sens. Actuators A* **25 – 27** (1991) 147 –154.
61. J. R. Glover, V. Vouk in L. Friberg et al. (eds.): *Handbook on the Toxicity of Metals,* Elsevier/North Holland Biomedical Press, Amsterdam 1979, chap. 35, pp. 587 – 598.
62. W. C. Cooper in W. C. Cooper (ed.): *Tellurium,* Von Nostrand Reinhold, New York 1971, pp. 313 –321.
63. *Ullmann,* 4th ed., **22,** 453.
64. *Kirk-Othmer,* 3rd ed., **22,** 666.
65. *Environmental Outlook,* vol. 3, no. 1, Selenium – Tellurium Development Assoc., Grimbergen, Belgium 1992.
66. TNO Report V 92.008, TNO-Voeding, Zeist, Holland, Jan. 1992.

Telomerization → Organometallic Compounds and Homogeneous Catalysis

Teratogenic Agents

Christian Baeder, Hoechst AG, Hattersheim, Federal Republic of Germany

1. Introduction 1
2. Course of Testing 2
3. References 3

1. Introduction

Numerous agents are known to be teratogenic in humans. Examples are radiation (therapeutic radiation, nuclear contamination), infections (rubella and cytomegalovirus), maternal metabolic imbalance (alcoholism, diabetes, folic acid deficiency), drugs (androgenic hormons, alkylating agents, diethylstilbestrol, retinoic acid, thalidomide, valproic acid) and environmental chemicals (chlorobiphenyls, lithium, organic mercury) [1]. Typical representatives of teratogenic agents and effects they cause in humans are:

Androgenic hormones	adrenal hyperplasia, virilisation of the female genitalia
Alkylating agents	degeneration of the embryonic disc and intrauterine death, abortion
Diethylstilbestrol	virilisation of the female fetus
Retinoic acid	multiple CNS and cardiac malformations, thymic aplasia
Thalidomide	limb malformations, phocomelia, ear defects, facial hemangioma, atresia of the esophagus or duodenum, tetralogy of Fallot, renal agenesis, cleft palate
Valproic acid	neural tube defects (anencephaly, microcephaly, facial dysmorphology, spina bifida), heart defects, heart ventricular septal defects, postnatal growth deficiency
Lithium	cardiovascular defects, Ebstein's anomaly
Chlorobiphenyls	dark-brown staining of the skin, exophthalmos, intrauterine growth retardation

Teratogenicity is a special form of embryotoxicity [2]. It presupposes a detrimental effect on the embryo and on organogenesis. Thus, malformations induced after birth, during the lactation period are to be regarded as postnatal developmental disturbances. For the tolerance classification of an agent, not only its teratogenic effect, but the total embryotoxic potential is of significance. This includes effects on:

1) Gravidity in general
2) Developmental stages of the zygote (the period of time from fertilization of the ovum to implantation in the wall of the uterus), and
3) Intrauterine development (embryogenesis: development of the basic body form and organogenesis; fetogenesis: growth and organ differentiation after the completion of organogenesis)

Forms of embryotoxicity and fetotoxicity are death, disturbance of growth (retardation), teratogenic effects, and induction of minor anomalies. Growth disturbances can be reversible within the uterus or after birth. A distinction is to be made between structural and functional developmental defects. Structural anomalies can be detected macroscopically, microscopically, or by electron microscopy [3]. Functional anomalies are expressed as disturbed organ functions, and enzymatic, endocrine, or immunological changes. Malformations are irreversible. Minor anomalies and variations of organs and the skeleton are characterized by the fact that they do not interfere with the vital functions or the social status of the individual, and can be reversible.

Testing of the embryotoxic potential is carried out preferably on laboratory mammals. As suitable models for humans, they should fulfil the following requirements: maternal–embryonal/fetal unit; similarity in anatomy, placentation, and transplacental passage; and similarity in kinetics, toxicokinetics, metabolism, and sensitivity.

The animals used most often are rats. The advantages of using rats are that results of previous systemic toxicity studies and information on toxicokinetics and metabolism are generally available. The rabbit is the first species of experimental animal in which the teratogenic effect of thalidomide could be reproduced. Both species have always been widely used for embryotox-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_189

icological studies and have proved to be sensitive models. If differences in the toxicokinetics, metabolism, or sensitivity are found between rats or rabbits and humans or if uninterpretable results are obtained, mice and subhuman primates are often used as alternative species. For these animals, extensive knowledge of the spontaneous and agent-induced occurrence of gravidity damage and intrauterine developmental defects is available. In the assessment of test results, lineage specific differences in sensitivity that are species or genetically dependent must be taken into account.

Alternative methods and test models, such as nonmammalian systems in vivo (e.g., chicken, frog, fish or Drosophila embryos, and hydra) as well as primary embryonal cell cultures and cell cultures from mammals and nonmammals, organ cultures (e.g., limb buds), and embryo cultures (pre- and post-implantative) in vitro, concentrate on different and limited test parameters (end points) and do not cover the entire embryonal, fetal, and postnatal development. The metabolism of the mammalian organism can be adopted by introducing a metabolism activating postmitochondrial liver homogenate (S-9 mixture from supernatant after centrifugation at 9000 *g* of homogenized livers of male rats treated with an enzyme inductor, e.g., Acrolor 1254 plus NADPH) [4]. However, these systems lack above all the maternal – embryonal unit. Therefore, at their present level of development, these systems are suitable only for prescreening, especially of structurally analogous agents, and for mechanistic testing [5]. Intensive work is being done on the development of alternative methods which would lead to decreased use of experimental animals.

The studies are carried out according to national and international test guidelines [6, 7]. The following basic requirements must be complied with before planning a study:

The identification and characterization of the embryotoxic potential of an agent depend on its chemical structure, mechanism of action, and on the dosage – response relationship [8].

Agents can directly or indirectly influence the intrauterine development of the embryo/fetus. Direct actions can influence the development of cells, tissues, or organs by changing or inhibiting cell proliferation, migration, and differentiation, or can kill the cell. Indirect actions can depend on metabolic or hormonal changes, intoxication of the mother, or disturbance of placenta function. Only a few mechanisms of teratogenesis have been elucidated.

With a few exceptions, the principle of the dosage – response relationship also applies to embryotoxicology/teratology. With increasing dose or concentration of an agent, the severity, variety of effects, and the number of damaged individuals increase. In rare cases, a clear dose – response relationship is not discernible or is concealed by other toxic effects. The determination of the doses or concentrations to be tested is of particular importance. The low dose is based on the maximal or a multiple of the amount expected to be used for daily exposure. With an intermediate dose, the geometric mean of the low and the highest dose selected, the safety margin of the agent should be determined. In the highest dose, the agent should exhibit minimal toxicity for the maternal or embryonal organism. An important purpose of the testing is the determination of the no observed adverse effect level (NOAEL), i.e., the dose or concentration in which no impairment of the physical condition of the mother and embryonal development can be determined. Agents, which are not toxic for the designated species of animal in oral doses of 1000 mg/kg body weight or in concentrations of 5 mg/L of maternal respiration air, or cannot be administered in higher doses or concentrations due to their constitution, should not be tested in smaller or higher amounts because of lacking practical relevance (limit test).

As a rule, route and the frequency of administration should be similar to the exposure of humans in practice. To attain higher concentrations of the agent in the organism, other types of application can also be used.

2. Course of Testing

Adult virginal female animals in heat are mated naturally, a vaginal smear being taken to check for insemination, or artificially inseminated. The day following insemination is called day 1 of gravidity. The animals are treaded daily with the agent to be tested, starting from the implantation of the blastocyst in the uterus (e.g., rat: day 7, rabbit: day 6) and continuing through the organogenesis phase or up to the end of gra-

vidity. The physical condition of the animals is examined every day. Disturbances of gravidity can result in abortion, premature birth, or death of all the embryos.

On the day before the expected birth (e.g., rat: day 21, rabbit: day 29) the animals are killed, a cesarean section is performed, and the uterus contents are examined. The condition of the uterus, the number, appearance, size, and weight of the living and dead fetuses and placentas, as well as the size of the prematurely dead embryonal structures that are in the process of decomposition and resorption are determined. The fetuses are inspected for external anomalies, dissected, and examined for structural anomalies of the inner organs. After exenteration and clearing the animal body in potassium hydroxide solution, the bones are stained with alizarin red S and placed in glycerol. The skeleton of the fetuses is then structurally examined for the state of ossification and for anomalies.

To find functional disorders, the mothers must deliver and raise their offspring. From the duration of gravidity and the course of birth, conclusions can be drawn regarding a possible impairment of gravidity (shortening, extension, and disturbance of parturition) caused by the administration of the agent to be tested. The number, size, behavior, and physical development of the offspring after birth give information about the postnatal effects of intrauterine induced damage.

The parameters to be investigated are the number of offspring that are born alive and dead, suckling, ability to survive, weight development, and physiological development parameters, such as the time of opening of the eyelids, unfolding of the auricles from the head, eruption of the incisors through the jaw, and the beginning of fur growth. Functional tests comprise observations of the free functional behavior including sight and hearing. For behavioral studies, a large number of test methods from the area of psychopharmacology are available. The test program contains, e.g., examinations of the general behavior in the open field, activity, clinging and movement on an inclined plane or movable roll, reflexes, motor system, equilibrium sense, carrying out of given tasks in discrimination tests, as well as learning ability, memory, and ability to relearn. These examinations are effected descriptively or with measuring methods.

With due regard to the dose, sensitivity, comparative kinetics, and metabolism, all the examinations together make possible a careful estimation of the risks involved in the use on or exposure of human beings. After consideration of the criteria mentioned above, the determination of injury to the embryo in animal experiments leads to the contraindication for use in women, except where pregnancy can be definitely excluded. Uncertainty is involved in the assessment of the functional and behavioral tests. Their meaningfulness has not yet been adequately demonstrated in the absence of a cellular (histological) structural substrate.

3. References

1. J. G. Wilson: *Environment and Birth Defects,* Academic Press, New York 1973.
2. *Ullmann,* 4th ed., **22,** 517 – 518.
3. Ch. Baeder, M. Albrecht, "Embryotoxic/Teratogenic Potential of Halothane," *Int. Arch. Occup. Environ. Health* **62** (1990) 263 – 271.
4. D. M. Marion, B. N. Ames, "Revised Methods for the Salmonella Mutagenicity Test," *Mutat. Res.* **113** (1983) 173 – 215.
5. ECETOC: "Alternative Approaches for the Assessment of Reproductive Toxicity (with Emphasis on Embryotoxicity/Teratogenicity)" *Monograph* **12** (1989).
6. R. Bass et al.: "Draft Guideline on Detection to Toxicity to Reproduction for Medicinal Products," *Adverse Drug React. Toxicol. Rev.* **9** (1991) 127 – 141.
7. OECD: Guidelines for Testing of Chemicals, Section 4: Health Effects, no. 414 – 416 (1981, 1983).
8. ECETOC: "Identification and Assessment of the Effects of Chemicals on Reproduction and Development (Reproductive Toxicology)," *Monograph* **5** (1983).

Terbium → Rare Earth Elements

Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid

RICHARD J. SHEEHAN, Amoco Research Center, Amoco Chemical Company, Naperville, Illinois, United States

1. **Introduction** 1
2. **Physical Properties** 1
3. **Production** 2
3.1. **Amoco Oxidation** 3
3.2. **Amoco Purification** 4
3.3. **Multistage Oxidation** 5
3.4. **Dynamit-Nobel (Witten) Process** . . . 5
3.5. **Esterification of Terephthalic Acid** . . 7
3.6. **Hydrolysis of Dimethyl Terephthalate** 7
3.7. **Alternative and Past Technologies** . . 8
4. **Environmental Protection** 9
5. **Quality Specifications** 9
6. **Storage and Transportation** 10
7. **Uses** 10
8. **Economic Aspects** 11
9. **Toxicology and Occupational Health** . 12
10. **References** 12

1. Introduction

Terephthalic acid [*100-21-0*], and isophthalic acid [*121-91-5*], both $C_8H_6O_4$, have the IUPAC names 1,4- and 1,3-benzenedicarboxylic acid. Dimethyl terephthalate [*120-61-6*], $C_{10}H_{10}O_4$, is also known as 1,4-benzenedicarboxylic acid dimethyl ester. The acids are produced by oxidation of the methyl groups on the corresponding *p*-xylene [*106-42-3*] or *m*-xylene [*108-38-3*]. After oxidation to a carboxylic acid, reaction with methanol [*67-56-1*] gives the methyl ester, dimethyl terephthalate.

Terephthalic acid and dimethyl terephthalate are used to make saturated polyesters with aliphatic diols as the comonomer. Isophthalic acid is used as a feedstock for unsaturated polyesters as well as a comonomer in some saturated products. Structures are as follows:

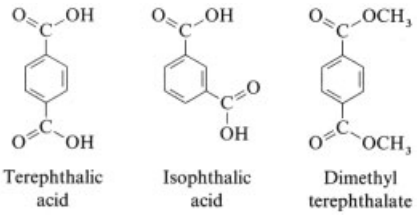

Terephthalic acid came to prominence through the work of WHINFIELD and DICKSON in Britain around 1940 [4]. Earlier work by CAROTHERS and coworkers in the United States established the feasibility of producing high molecular weight linear polyesters by reacting diacids with diols, but they used aliphatic diacids and diols. These made polyesters which were unsuitable to be spun into fibers. WHINFIELD and DICKSON found that symmetrical aromatic diacids yield high-melting, crystalline, and fiber-forming materials; poly(ethylene terephthalate) has since become the largest volume synthetic fiber. Worldwide, terephthalic acid plus dimethyl terephthalate ranked about 25th in tonnage of all chemicals produced in 1992, and about tenth in terms of organic chemicals.

2. Physical Properties

Terephthalic acid, M_r 166.13, is available acommercially as a free-flowing powder composed of rounded crystals. It forms needles if recrystallized slowly. Vapor pressure is low: 0.097 kPa at 250 °C, with sublimation at 402 °C and atmospheric pressure. Melting has been reported at 427 °C.

Dimethyl terephthalate, M_r 194.19, melts at 140.6 °C and has sufficient vapor pressure for vacuum distillation. The molten form is preferred commercially, but flakes and briquettes are available when long transport distances are required.

Isophthalic acid, M_r 166.13, melts at 348 °C. Vapor pressure is 0.61 kPa at 250 °C.

Both terephthalic and isophthalic acid are stable, intractable compounds with low solubilities in most solvents; isophthalic acid is 2–5 times more soluble than terephthalic acid in the same solvent. Terephthalic acid solubilities > 10 g per 100 g solvent at room temperature occur with ammonium, potassium, or sodium hy-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_193

Table 1. Physical properties of terephthalic acid and isophthalic acid

	Terephthalic	Isophthalic
Crystal density, g/cm^3, 25 °C	1.58	1.53
Specific heat, J g^{-1} K^{-1}, to 100 °C	1.20	1.22
Heat of formation, kJ/mol	− 816	− 803
Heat of combustion, kJ/mol, 25 °C	3189	3203
First dissociation constant	2.9×10^{-4}	2.4×10^{-4}
Second dissociation constant	3.5×10^{-5}	2.5×10^{-5}
Ignition temperature in air, °C	680	700

Table 2. Solubility of terephthalic acid and isophthalic acid (g/100 g solvent)

	25 °C	150 °C	200 °C	250 °C
Terephthalic acid solubility, g/100 g solvent				
Water	0.0017	0.24	1.7	12.6
Methanol	0.10	3.1		
Acetic acid	0.013	0.38	1.5	5.7
Isophthalic acid solubility, g/100 g solvent				
Water	0.012	2.8	25.2	
Methanol	1.06	8.1		
Acetic acid	0.23	4.3	13.8	44.5

droxide, dimethylformamide, and dimethyl sulfoxide. Tetramethylurea and pyridine each dissolve ca. 7 g per 100 g.

Dimethyl terephthalate is also stable, and more soluble in common organic solvents than the acids. Its main physical properties are a melting point below the point of degradation, and a vapor pressure which allows for purification by distillation.

Other physical properties of terephthalic acid and isophthalic acid are listed in Tables 1 and 2. Some physical properties of dimethyl terephthalate are listed below:

Solubility, g/100 g solvent	
Methanol, 25 °C	1.0
Methanol, 60 °C	5.7
Ethyl acetate, 25 °C	3.5
Ethyl acetate, 60 °C	16.0
Trichloromethane, 25 °C	10.0
Trichloromethane, 60 °C	23.0
Benzene, 25 °C	2.0
Benzene, 60 °C	14.0
Toluene, 25 °C	4.3
Toluene, 60 °C	10.4
Dioxane, 25 °C	7.5
Dioxane, 60 °C	26.5
Vapor pressure, kPa	
140 °C	1.58
160 °C	2.62
200 °C	10.88
250 °C	42.60
Density at *mp*	1.07 g/cm^3
Viscosity, 180 °C	0.0071 Pa · s
Viscosity, 200 °C	0.0060 Pa · s
Specific heat, solid	1.55 J g^{-1} K^{-1}
Specific heat, liquid	1.94 J g^{-1} K^{-1}
Heat of fusion	159.1 kJ/Kg
Heat of vaporization	355.5 kJ/Kg
Heat of formation	−740.2 kJ/mol
Heat of combustion, 25 °C	4660 kJ/mol
Flash point, DIN 51 758	141 °C

3. Production

p-Xylene is the feedstock for all terephthalic acid and dimethyl terephthalate production; *m*-xylene is used for all isophthalic acid. Oxidation catalysts and conditions have been developed which give nearly quantitative oxidation of the methyl groups, leaving the benzene ring virtually untouched. These catalysts are combinations of cobalt, manganese, and bromine, or cobalt with a co-oxidant, e.g., acetaldehyde [*75-07-0*]. Oxygen is the oxidant in all processes. Acetic acid [*64-19-7*] is the reaction solvent in all but one process. Given these constant factors, there is only one industrial oxidation process, with different variations, two separate purification processes, and one process which intermixes oxidation and esterification steps.

3.1. Amoco Oxidation

About 70 % of the terephthalate feedstock used worldwide is produced with a catalyst system discovered by Scientific Design [5,6]. Almost 100 % of new plants use this reaction. A separate company, Mid-Century Corporation, was established to market this technology, and subsequently purchased by Amoco Chemical. Amoco developed a commercial process, as did Mitsui Petrochemical, now Mitsui Sekka. Mitsui was an early licensee of Mid-Century. Both Amoco and Mitsui participate in joint-venture companies, and both have licensed the process. Licensees are distributed around the world, and some have relicensed the process to other companies.

A soluble cobalt – manganese – bromine catalyst system is the heart of the process. This yields nearly quantitative oxidation of the *p*-xylene methyl groups with small xylene losses [7]. Acetic acid is the solvent, and oxygen in compressed air is the oxidant. Various salts of cobalt and manganese can be used, and the bromine source can be HBr, NaBr, or

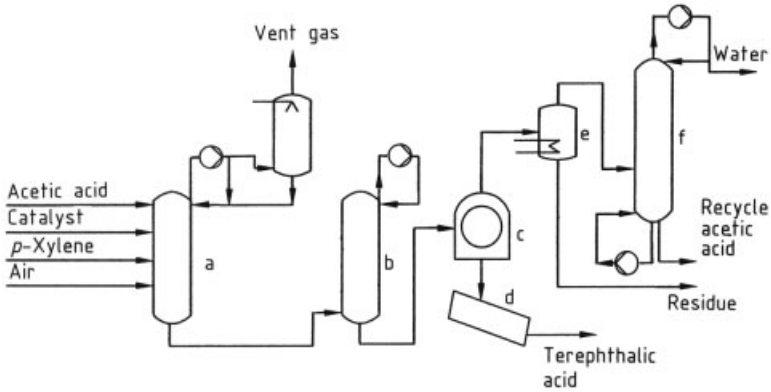

Figure 1. Catalytic, liquid-phase oxidation of *p*-xylene to terephthalic acid by the Amoco process
a) Oxidation reactor; b) Surge vessel; c) Filter; d) Dryer; e) Residue still; f) Dehydration column

tetrabromoethane [*79-27-6*] among others. The highly corrosive bromine – acetic acid environment requires the use of titanium-lined equipment in some parts of the process.

A feed mixture of *p*-xylene, acetic acid, and catalyst is continuously fed to the oxidation reactor (Fig. 1). The feed mixture also contains water, which is a byproduct of the reaction. The reactor is operated at 175 – 225 °C and 1500 – 3000 kPa. Compressed air is added to the reactor in excess of stoichiometric requirements to provide measurable oxygen partial pressure and to achieve high *p*-xylene conversion. The reaction is highly exothermic, releasing 2×10^8 J per kilogram *p*-xylene reacted. Water is also released. The reaction of 1 mol *p*-xylene with 3 mol dioxygen gives 1 mol terephthalic acid and 2 mol water. Only four hydrogen atoms, representing slightly over 2 wt % of the *p*-xylene molecule, are not incorporated in the terephthalic acid.

Owing to the low solubility of terephthalic acid in the solvent, most of it precipitates as it forms. This yields a three-phase system: solid terephthalic acid crystals; solvent with some dissolved terephthalic acid; and vapor consisting of nitrogen, acetic acid, water, and a small amount of oxygen. The heat of reaction is removed by solvent evaporation. A residence time up to 2 h is used. Over 98 % of the *p*-xylene is reacted, and the yield to terephthalic acid is > 95 mol %. Small amounts of *p*-xylene and acetic acid are lost, owing to complete oxidation to carbon oxides, and impurities such as oxidation intermediates are present in reactor effluent. The excellent yield and low solvent loss in a single reactor pass account for the near universal selection of this technology for new plants.

The oxidation of the methyl groups occurs in steps, with two intermediates, *p*-toluic acid [*99-94-5*] and 4-formylbenzoic acid [*619-66-9*]. While 4-formylbenzoic acid is the IUPAC name of the intermediate, it is customarily referred to as 4-carboxybenzaldehyde (4-CBA).

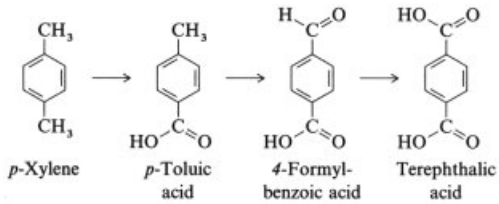

4-Formylbenzoic acid is troublesome, owing to its structural similarity to terephthalic acid. It co-crystallizes with terephthalic acid and becomes trapped and inaccessible for completion of the oxidation. Up to 5000 ppm 4-formylbenzoic acid can be present, and this necessitates a purification step to make the terephthalic acid suitable as a feedstock for polyester production.

The slurry is passed from the reactor to one or more surge vessels where the pressure is reduced. Solid terephthalic acid is then recovered by centrifugation or filtration, and the cake is dried and stored prior to purification. This is typically referred to as crude terephthalic acid, but is > 99 % pure.

Vapor from the reactor is condensed in overhead heat exchangers, and the condensate is re-

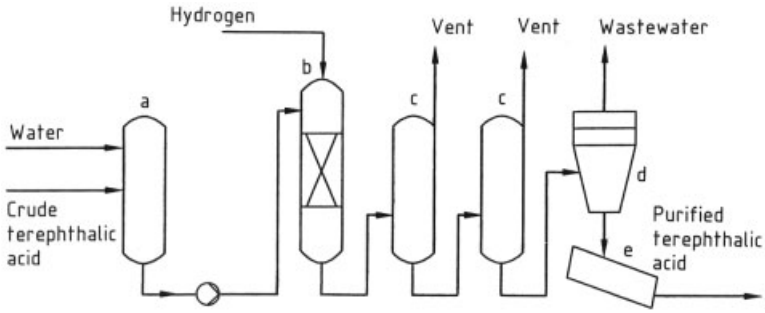

Figure 2. Purification of terephthalic acid by the Amoco process
a) Slurry drum; b) Hydrogenation reactor; c) Crystallizers; d) Centrifuge; e) Dryer

fluxed to the reactor. Steam is generated by the condensation and is used as a heating source in other parts of the process. Oxygen-depleted gas from the condensers is scrubbed to remove most uncondensed vapors. Similar to the reactor condensate, liquid from centrifuges or filters is sent to solvent recovery. Since the centrifugate or filtrate contains dissolved species, it is first sent to a residue still. Vapor from the still and other vents from throughout the oxidation process are sent to a solvent dehydration tower. The tower removes the water formed in the reaction as the overhead stream, and the acetic acid from the tower bottom is combined with fresh acetic acid to make up for process losses, and returned to the process.

Isophthalic acid is also produced by this process from *m*-xylene. Because isophthalic acid is several times more soluble than terephthalic acid, much less precipitates in the reactor. Consequently, isophthalic acid from this process contains much less 3-formylbenzoic acid, since it tends to stay in solution where complete oxidation can occur. Further purification was not carried out in the past, but a purified grade that is now being produced will become the standard.

3.2. Amoco Purification

The purification process developed by Amoco Chemical [8] and used on terephthalic acid from the Amoco oxidation process supplies over 60 % of the terephthalate feedstock for polyester production (Fig. 2).

Crude terephthalic acid is unsuitable as a feedstock for polyester, primarily owing to the 4-formylbenzoic acid impurity concentration. There are also yellow impurities and residual amounts of catalyst metals and bromine. The Amoco purification process removes 4-formylbenzoic acid to < 25 ppm, and also gives a white powder from the slightly yellow feed.

It is necessary to make all impurities accessible to reaction, so the crude terephthalic acid is slurried with water and heated until it dissolves entirely. A solution of at least 15 wt % is obtained, and this requires a temperature $\geq 260\,^{\circ}$C. The solution passes to a reactor where hydrogen is added and readily dissolves. The solution is contacted with a carbon-supported palladium catalyst. Reactor pressure is held above the vapor pressure of water to maintain a liquid phase.

The 4-formylbenzoic acid is converted to *p*-toluic acid in the reactor, and some colored impurities are hydrogenated to colorless compounds. The reaction is highly selective; loss of terephthalic acid by carboxylic acid reduction or ring hydrogenation is < 1 %.

After reaction, the solution passes to a series of crystallizers where the pressure is sequentially decreased [9]. This results in a stepped temperature reduction, and crystallization of the terephthalic acid. The more soluble *p*-toluic acid formed in the reactor, and other impurities, remain in the mother liquor. After leaving the final crystallizer, the slurry undergoes centrifugation and/or filtration to yield a wet cake, and the cake is dried to give a free-flowing terephthalic acid powder as the product. Over 98 wt % of the incoming terephthalic acid is recovered as purified product.

As with the Amoco oxidation, this purification process is also used with isophthalic acid.

3.3. Multistage Oxidation

Several companies, mostly in Japan [10, 11], have developed processes to reduce the 4-formylbenzoic acid content to 200 – 300 ppm by more intensive oxidation. A separate purification step is eliminated; the concentration of 4-formylbenzoic acid is low enough for the terephthalic acid to be suitable as a feedstock for some polyester products where high feedstock purity is not critical. The product is often called medium-purity terephthalic acid, and accounts for about 11 % of the terephthalic acid produced. Most of these processes also use the catalyst system discovered by Scientific Design.

Most medium-purity terephthalic acid is produced by Mitsubishi Kasei and its licensees. They have named this product Q-PTA, and it has a typical 4-formylbenzoic acid level of ca. 290 ppm [10]. Mitsubishi has also developed a still more intensive oxidation process where the 4-formylbenzoic acid level is further reduced. The product is called S-QTA.

The oxidation of *p*-xylene in acetic acid with a cobalt – manganese – bromine catalyst is carried out as in the Amoco oxidation. The slurry is heated to 235 – 290 °C and oxidized further in another reactor. More catalyst can be added in addition to the temperature increase [12].

Heating gives increased terephthalic acid solubility, and as crystals dissolve, some 4-formylbenzoic acid and colored impurities are released. Although the terephthalic acid is not completely soluble at the higher temperature, the crystals can digest. Digestion is a dynamic equilibrium process wherein crystals constantly dissolve and reform. This increases the release of 4-formylbenzoic acid into solution where oxidation can be completed. While the need for a separate purification process is eliminated, another reactor is needed in the oxidation process. Also, at higher temperature, acetic acid tends to be oxidized to a larger extent to carbon oxides and water [12].

The rest of the Mitsubishi process consists of solid – liquid separation and drying to obtain the powdered product. Acetic acid must be dehydrated and recycled to the process.

Eastman Chemical has also developed a medium-purity terephthalic acid product. The process does not employ manganese, only cobalt and bromine. Two oxidation stages are used, both at 175 – 230 °C [13]. Instead of heating between stages to obtain increased solubility, as performed by Mitsubishi, the contents of the first reactor are sent to hydroclones where hot, fresh acetic acid displaces the mother liquor. Samarium may be added to the catalyst. A residence time up to 2 h with air addition provides sufficient digestion of the crystals to yield a 40 – 270 ppm level of 4-formylbenzoic acid [13]. Downstream recovery of the terephthalic acid crystals by solid – liquid separation and drying must again be performed.

3.4. Dynamit-Nobel (Witten) Process

Most dimethyl terephthalate is made by a process first developed by Chemische Werke Witten, with work also being done at a division of Standard Oil of California. Modifications were made by Hercules and Dynamit-Nobel. Hüls Troisdorf licensed the process with further modifications [14]. Dimethyl terephthalate is formed in four steps. First, *p*-xylene is passed through an oxidation reactor, where *p*-toluic acid is formed. It then passes to an esterification reactor, the second step, where methanol is added to form methyl *p*-toluate [*99-75-2*]. The methyl *p*-toluate is isolated and returned to the oxidation reactor for oxidation to monomethyl terephthalate [*1679-64-7*], the third step, followed by the fourth step, esterification to dimethyl terephthalate. The sequence is as follows:

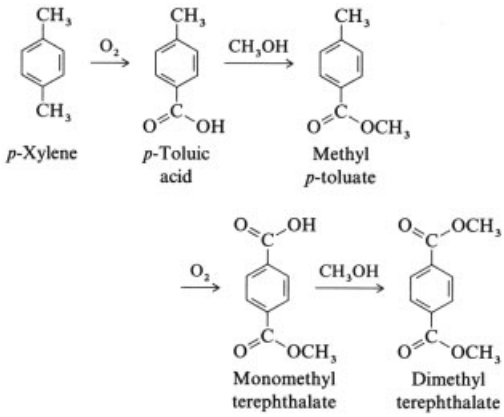

The process as licensed by Hüls Troisdorf is illustrated in Figure 3. Fresh and recovered *p*-

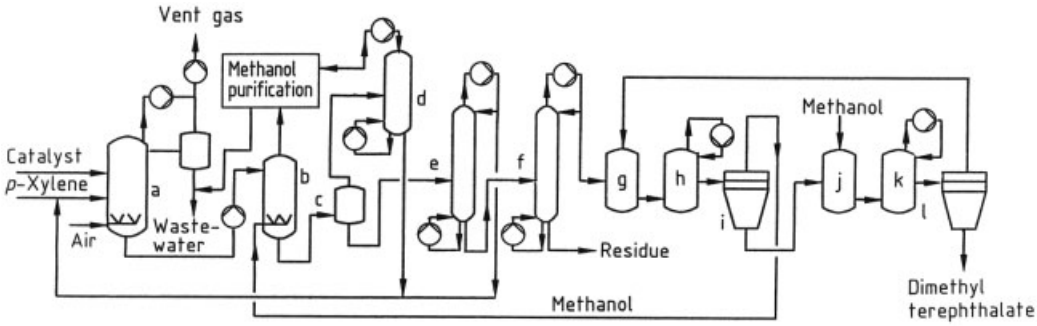

Figure 3. Production of dimethyl terephthalate by the Dynamit Nobel process
a) Oxidation reactor; b) Esterifier; c) Expansion vessel; d) Methanol recovery column; e, f) Methyl *p*-toluate and dimethyl terephthalate columns; g, j) Dissolvers; h, k) Crystallizers; i, l) Centrifuges

xylene, along with catalyst (mostly cobalt with some manganese) are combined with methyl *p*-toluate and fed to the liquid-phase oxidation reactor. Because bromine and acetic acid are not used, vessels lined with titanium or other expensive metals are not necessary. Oxygen supplied by compressed air is added at the bottom. Oxidation conditions are 140 – 180 °C and 500 – 800 kPa. The heat generated by oxidation is removed by vapors of unreacted *p*-xylene and the water of reaction. Cooling coils in the reactor are used to generate steam. The steam and reactor vapors are condensed and combined to recover *p*-xylene for recycle.

The oxidation effluent is then heated and sent to the esterification reactor, operated at 250 °C and 2500 kPa. Excess vaporized methanol is sparged into the esterifier, where the *p*-toluic acid and monomethyl terephthalate are converted noncatalytically to methyl *p*-toluate and dimethyl terephthalate, respectively. Overhead vapors from the esterification reactor are condensed and fed to a distillation system, where the water from the esterification is separated from methanol, which is recycled.

The remainder of the process separates the dimethyl terephthalate from methanol and methyl *p*-toluate, which are recycled, and residue and wastewater, which go to waste treatment.

The product from the esterifier goes to an expansion vessel. Vapor from this vessel feeds a methanol recovery column, where the methanol overhead goes to methanol recovery, and the methyl *p*-toluate bottoms are recycled to the oxidation reactor. Liquid from the expansion vessel feeds two vacuum distillation columns in series, which yield crude dimethyl terephthalate. The first column recovers more methyl *p*-toluate overhead for recycle to oxidation, and the bottoms feeds the crude dimethyl terephthalate column, where the product is taken overhead. The bottoms from the dimethyl terephthalate column, containing heavy byproducts and catalyst metals, can be mixed with water from the oxidation, which dissolves the catalyst. The resulting slurry is centrifuged; the catalyst solution is recycled, and the cake is sent to disposal.

The crude dimethyl terephthalate goes through two stages of crystallization. It is slurried with the methanol mother liquor from the second crystallization stage and dissolved by heating. Dimethyl terephthalate is crystallized from this solution on cooling, by flashing the methanol. The dimethyl terephthalate cake is separated by centrifugation, dissolved in fresh methanol, and crystallized in the same way. The wet dimethyl terephthalate cake can be melted and stored molten, or dried and flaked. There is usually a distillation column after the second centrifuge for further purification.

Mother liquor from the first centrifugation is sent to the methanol recovery system. The centrifugation yields a cleaner mother liquor which is combined with methanol recovered from the melting operation for use in dissolving the crude dimethyl terephthalate entering the crystallization section. Side-streams from throughout the process are recycled to appropriate points to maximize yield and to minimize methanol and catalyst use.

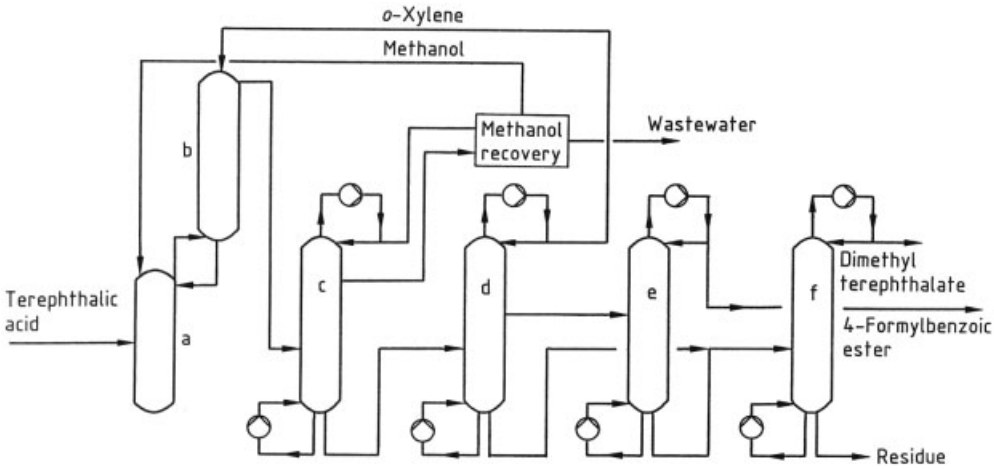

Figure 4. Esterification of terephthalic acid and separation of purified dimethyl terephthalate
a) Esterifier; b) *o*-Xylene scrubber; c) Methanol column; d) *o*-Xylene recovery column; e) 4-Formylbenzoic ester stripper; f) Purification column

3.5. Esterification of Terephthalic Acid

Terephthalic acid can be produced and, in a separate process, esterified with methanol to dimethyl terephthalate which is then purified by distillation (Fig. 4) [15, 16]. This process can be used on highly impure terephthalic acid, because of the purification achievable by distillation.

Crude terephthalic acid and excess methanol are mixed and pumped to the esterification reactor. In this example, *o*-xylene [*95-47-6*] is used to enhance the subsequent separations. The terephthalic acid is rapidly esterified by the methanol at 250 – 300 °C without catalysis, although a catalyst can be used. Methanol vapor carries dimethyl terephthalate and *o*-xylene from the reactor to a column where *o*-xylene is added to scrub out monomethyl terephthalate and return it to the reactor for completion of the esterification. The vapor contains the dimethyl terephthalate product as well as methanol, *o*-xylene, water from the esterification, and esterified impurities in the terephthalic acid feed. Several distillation steps are needed to separate out the dimethyl terephthalate and to process the separated streams for recovery of valuable components for recycle.

The reactor overhead vapor first goes to a methanol column where methanol is removed overhead, water and some methanol form a side-draw, and the bottoms contain the dimethyl terephthalate, *o*-xylene, and impurities.

Next, in the *o*-xylene recovery column, dimethyl terephthalate purification is started; it operates at 10 – 20 kPa absolute so that a temperature of 200 – 230 °C can be used. *o*-Xylene is removed overhead, the methyl esters of 4-formylbenzoic acid and *p*-toluic acid are removed in the middle, and dimethyl terephthalate forms the bottoms. A 4-formylbenzoic acid stripping column follows, where the middle stream from the previous column is sent, so that 4-formylbenzoic ester can be removed overhead. Finally, the bottoms from both the *o*-xylene recovery column and the 4-formylbenzoic acid stripper are sent to a purification column, where the dimethyl terephthalate product is taken overhead.

Methanol from the top of the methanol recovery column is sent to a purification column where the overhead contains low boilers, and the methanol from the bottoms is recycled. The methanol and water from the side-draw of the methanol recovery column are sent to a methanol dehydration column, where the water is removed and the dehydrated methanol recycled.

3.6. Hydrolysis of Dimethyl Terephthalate

High-purity terephthalic acid can be produced by hydrolysis of dimethyl terephthalate. Slightly over 2 % of terephthalate feedstock is produced by this route.

Dimethyl terephthalate and recovered monomethyl terephthalate are combined with water from the methanol recovery tower of the dimethyl terephthalate process and heated to 260 – 280 °C at 4500 – 5500 kPa in a hydrolysis reactor, where the methyl esters are hydrolyzed. The overhead methanol plus water vapor is returned to the methanol recovery tower for separation. The hydrolysis reactor liquid is sent to a series of crystallizers and cyclones. After cooling to crystallize the terephthalic acid at ca. 200 °C, washing cyclones are used to remove mother liquor which contains monomethyl terephthalate. The cyclone underflow is further cooled to 100 °C, and final crystallization occurs. This slurry is centrifuged and dried to give the final product.

Monomethyl terephthalate from the cyclone overflow is recovered, again by cooling to 100 °C and centrifuging, and the cake is recycled to the hydrolysis reactor.

3.7. Alternative and Past Technologies

Oxidation of *p*- or *m*-xylene is also possible with acetic acid solvent and a cobalt catalyst with an acetaldehyde activator. Reaction temperature is 120 – 140 °C, and as a result the oxidation requires a residence time ≥ 2 h. Titanium-lined vessels are not required because bromine is not used. The acetaldehyde is converted to acetic acid as it promotes the reaction. After reaction, the terephthalic or isophthalic acid is recovered by centrifugation or filtration and drying, much like the Amoco process, and acetic acid is recovered for recycle by distillation.

With isophthalic acid, which has a lower concentration of 3-formylbenzoic acid, considerable purification is possible by slurrying in water to ca. 20 – 25 wt %, heating to 240 – 260 °C to dissolve, and crystallizing by cooling in batch mode. Recovery of the crystals is again by centrifugation or filtration and drying.

Eastman Chemical in the United States also used an acetaldehyde activator with a cobalt catalyst to produce terephthalic acid. Bromine is now being used in place of acetaldehyde [17].

Mobil Chemical in the United States had a commercial terephthalic acid operation which has since been abandoned [18]. As with the above processes, a cobalt catalyst was used with acetic acid solvent; the activator was 2-butanone [*78-93-3*]. After reaction, the crude terephthalic acid was leached by adding pure acetic acid and heating to achieve partial solubility; this removed gross impurities. Final purification was by sublimation and catalytic treatment of the vapor. Impure terephthalic acid was vaporized in a steam carrier, and catalyst and hydrogen were added to convert undesirable organic impurities. Subsequent cooling of the stream condensed the terephthalic acid, leaving impurities in the vapor. Product crystals were removed by cyclones.

A process no longer practiced is based on Henkel technology. Starting with phthalic anhydride [*85-44-9*], the monopotassium and dipotassium *o*-phthalate salts were formed in sequence. The dipotassium salt was isolated from solution by spray drying, and isomerized to dipotassium terephthalate under carbon dioxide at 1000 – 5000 kPa and 350 – 450 °C. This salt was dissolved in water and recycled to the start of the process, where terephthalic acid crystals formed during the production of the monopotassium salt. The crystals were recovered by filtration.

Two proposed processes which were never commercialized were by Lummus, using a dinitrile route [19] and by Eastman, using the formation of 1,4-diiodobenzene [*624-38-4*] with carbonylation to aromatic acids [20].

Toluene [*108-88-3*] is a potential feedstock for manufacture of terephthalic acid and is cheaper than *p*-xylene. Mitsubishi Gas Chemical has researched a process where a complex between toluene and hydrogen fluoride – boron trifluoride is formed, that can be carbonylated with carbon monoxide to form a *p*-tolualdehyde [*104-87-0*] complex [21]. After decomposition of the complex, *p*-tolualdehyde can be oxidized in water with a manganese – bromine catalyst system to terephthalic acid. While cheaper toluene is the feedstock and acetic acid is not required, the complexities of handling hydrogen fluoride – boron trifluoride and the need for carbon monoxide add other costs. This process has not been commercialized.

4. Environmental Protection

The three chemicals discussed here consist only of carbon, hydrogen, and oxygen, so carbon

dioxide and water are the final effluents of oxidative degradation. There are traces of the cobalt – manganese – bromine catalyst in waste streams, however.

Effluents from the terephthalic acid process include water generated during oxidation and water used as the purification solvent. Gaseous emissions are mostly the oxygen-depleted air from the oxidation step. Volumes are large, but the chemicals in the streams can be effectively destroyed and removed.

Water effluents are subjected to aerobic wastewater treatment, where the dissolved species, mostly terephthalic acid, acetic acid, and impurities such as *p*-toluic acid, are oxidized to carbon dioxide and water by the action of bacteria which are acclimated to these chemicals. The bacterial growth is a sludge which can be dried and burned or spread on land. An anaerobic process has been developed to treat the wastewater [22]. Advantages include much less waste sludge production, less utility consumption, and the generation of methane, which can be burned for energy recovery.

Waste gas is scrubbed in process equipment to remove acetic acid vapors for recovery. Trace amounts of other compounds formed in the reactor can be removed by catalytic oxidation, followed by scrubbing to meet the most demanding regulations for process vents [23].

Waste streams from distillation in the dimethyl terephthalate process can be burned for energy recovery.

Another aspect of environmental protection is the recyclability of polyesters. Post-consumer containers can be ground and cleaned, and then extruded and spun into fiber to fill bedding, and quilted clothing. Textile fibers can also be made. The polyester can also be reacted to regenerate the terephthalate feedstock. High-temperature hydrolysis yields terephthalic acid. Reaction with methanol yields dimethyl terephthalate.

5. Quality Specifications

Feedstocks for polyester production must be extremely pure; they are among the purest high-volume chemicals sold by industry. If certain impurities are present in high enough concentrations, harmful effects can be measured, e.g., monofunctional compounds can cap the polyester chain and limit molecular mass buildup. Trifunctional carboxylic acids can cause chain branching which leads to undesirable rheological and spinning properties. Colored impurities can be incorporated into the polyester. In particular, 4-formylbenzoic acid limits polyester molecular mass and causes yellowness [24]. The particle size of terephthalic acid determines how the powder flows, and the viscosity of the slurry when mixed with 1,2-ethanediol [*107-21-1*].

Owing to the consistent high purity of these products and the different effects of impurities, specifications put limits on specific impurities or color, rather than overall purity (Table 3). Trends in quality specifications are away from maximum or minimum values, toward a target value with an allowable range. This has been instituted to some extent by Amoco Chemical for the *p*-toluic acid content, for example.

Table 3. Specifications and typical analyses of purified terephthalic acid

Property	Specification	Typical value
Acid number, mg KOH/g	675 ± 2	673 – 675
Ash, ppm	≤ 15	< 3
Metals,∗ ppm	≤ 9	< 2
Water, wt %	≤ 0.2	0.1
4-Formylbenzoic acid, ppm	≤ 25	15 ∗∗
p-Toluic acid, ppm	125 ± 45	125 ∗∗

∗ Co + Mn + Fe + Cr + Ni + Mo + Ti.
∗∗ Medium-purity terephthalic acid typically has 250 ppm 4-formylbenzoic acid and <50 ppm p-toluic acid.

In Table 3, acid number, determined by titration, is an overall purity measurement. A perfectly pure sample will have an acid number of 675.5 mg KOH/g, but the low impurity levels make the acid number meaningless as a quantitative indication of purity, and it is being phased out.

Impurities specified include *p*-toluic acid and 4-formylbenzoic acid, residual water, and trace metals, the ash being trace metal oxides. There are also color and particle size measurements. Polyester producers often have specific tests which they prefer for color or particle size.

Medium-purity terephthalic acid specifications address the same impurities, except that

4-formylbenzoic acid concentrations are up to 300 ppm, and residual acetic acid is present.

Typical dimethyl terephthalate specifications are shown in Table 4. Freezing point is an extremely sensitive purity indicator. Since dimethyl terephthalate is often shipped and stored in molten form, color measurements are made on the melt. A color stability test can determine the sample's resistance to degradation. This resistance is reduced by the presence of unesterified acid groups, so these residual groups are measured.

Table 4. Specifications and typical analyses of dimethyl terephthalate

Property	Specification	Typical value
Freezing point, °C	≥ 140.62	140.64
Acid number, mg KOH/g	0.03	0.01
Molten color, APHA	≤ 25	10
Color stability,* APHA	≤ 25	15
Water, wt %	≤ 0.02	0.01

* 4 h at 175 °C

Specifications for isophthalic acid are similar to those of terephthalic acid, *m*-toluic acid and 3-formylbenzoic acid being the impurities.

6. Storage and Transportation

Terephthalic acid powder is stored in silos and transported in bulk by rail hopper car, hopper truck, or bulk container with a polyethylene liner. Hopper cars and trucks should be dedicated to this service, because of the extreme purity requirements. Shipment is also in 1 t bags, or paper bags containing ca. 25 kg. Isophthalic acid, being a much lower volume chemical, has a higher proportion shipped in bags.

Dimethyl terephthalate is stored molten in insulated heated tanks, and is preferentially shipped molten in insulated rail tank cars or tank trucks. It is also solidified into briquettes or flakes, and shipped in 1 t or 25 kg bags.

The huge production volume of terephthalic acid and dimethyl terephthalate favors the use of bulk shipment with minimized packaging costs. Rail cars are preferred where possible, and they can accommodate up to 90 t. Containers with plastic liners hold 20 t terephthalic acid, and these are the preferred ocean shipping method.

7. Uses

Terephthalic acid and dimethyl terephthalate are used almost exclusively to produce saturated polyesters. Poly(ethylene terephthalate), the alternating copolymer of terephthalic acid and 1,2-ethanediol, accounts for > 90 % of this use, with a worldwide demand of 12.6×10^6 t in 1992. Textile and industrial fiber accounted for 75 % of this demand. Polyester is the largest volume synthetic fiber. Food and beverage containers accounted for 13 %, and constituted the fastest growing segment. Film for audio, video, and photography took 7 %.

Other uses are for poly(butylene terephthalate), a high-performance molding resin made by reaction with 1,4-butanediol [*74829-49-5*], and for special industrial coatings, solvent-free coatings, electrical insulating varnishes, aramid fibers, and adhesives. A small amount of bis(2-ethylhexyl) terephthalate [*6422-86-2*] is produced as a plasticizer, and some dimethyl terephthalate is ring-hydrogenated to produce the cyclohexane analog, 1,4-cyclohexanedicarboxylic acid [*1076-97-7*], for specialty polyesters and coatings. Isophthalic acid is used as a comonomer with terephthalic acid in bottle and specialty resins. It is also a feedstock for coatings and for high-performance unsaturated polyesters, being first reacted with maleic anhydride [*108-31-6*] and then this resin is cross-linked with styrene [*100-42-5*].

8. Economic Aspects

Together, terephthalic acid and dimethyl terephthalate rank about 25th in tonnage of all chemicals, and about tenth for organic chemicals. The rapid growth of polyester use has led to an above-average growth for these feedstocks, and this is projected to continue as deeper penetration into fiber and container markets is made. The versatility of terephthalate and isophthalate polyesters is projected to drive growth in the foreseeable future.

Capacity for the three types of terephthalate feedstocks since 1980 is given in Table 5. Purified terephthalic acid has grown rapidly, and is projected to supply virtually all the growth

for polyesters, about 7 % annually to the year 2000. Medium-purity terephthalic acid capacity has grown by construction of a few new plants, and by conversion of some dimethyl terephthalate capacity. Dimethyl terephthalate has shown little change in capacity since 1980, and essentially none is expected in the future.

Table 5. Capacity of terephthalate feedstocks (10^3 t/a)

Feedstock *	1980	1986	1992
PTA	2800	4000	8100
MTA	420	580	1300
DMT	3300	3600	3300

* PTA purified terephthalic acid; MTA medium-purity terephthalic acid; DMT dimethyl terephthalate.

The Far East contains most of the world's population and most of the garment manufacturing industry. As a result, it is experiencing most of the growth in manufacturing capacity for polyester fibers, and most of the growth for terephthalic acid production (Table 6). Approximately 75 % of the growth in terephthalate feedstocks in 1986 – 1992 was due to purified terephthalic acid plants built in the Far East.

Table 6. Geographical distribution of terephthalate production (10^3 t/a)

	1980	1986	1992
North America	3300	3400	3600
Europe	1500	2100	2700
Far East	1500	2300	5800
Rest of World	220	380	600

Pricing is often determined in conjunction with long-term contracts, and is most influenced by the cost of *p*-xylene. In 1992, prices were about $ 0.60/kg for terephthalic acid and $ 0.57/kg for dimethyl terephthalate [25]. Because a given mass of terephthalic acid produces 17 % more polyester than the same mass of dimethyl terephthalate, the dimethyl terephthalate price must be adjusted downward, and allowance is made for the value of the methanol generated during transesterification.

The major producer of purified terephthalic acid is Amoco Chemical, with over 17 % of the world's production in 1992. Joint ventures of Amoco, the largest being China American Petrochemical in Taiwan and Samsung Petrochemical in Korea, also produce about 17 %. Other major producers are Mitsui Sekka in Japan and Imperial Chemical Industries in England and Taiwan. China is certain to become a major producer, and large plants are being planned in Southeast Asia. Larger producers of medium-purity terephthalic acid are Mitsubishi Kasei in Japan, and Eastman and DuPont in the United States. Some polyester producers have licensed technology to supply their own feedstock. All licensors are themselves producers of terephthalic acid, or have obtained the technology from a producer.

Major dimethyl terephthalate producers are DuPont, Eastman Chemical, and Hoechst Celanese in the United States, Hoechst in Europe, and Teijin Petrochemical in Japan. There are several small, older plants, mostly in Europe and China.

Isophthalic acid production capacity is 250×10^3 t/a, ca. 2 % of the terephthalate total. Amoco Chemical is the major producer.

9. Toxicology and Occupational Health

Terephthalic acid, dimethyl terephthalate, and isophthalic acid all have low toxicity and cause only mild and reversible irritation to skin, eyes, and the respiratory system.

For oral ingestion of terephthalic acid, LD_{50} values for rats have been reported as 18.8 g/kg [26], and for mice as 6.4 g/kg [27], or > 5 g/kg [28]. Dimethyl terephthalate LD_{50} values for rats are 6.5 g/kg [29] and 4.39 g/kg [26]. Some mortality was found at 5 g/kg oral ingestion [30], and with a 28-day oral uptake of 5 % in the feed, ca. 2.5 g/kg. For isophthalic acid, LD_{50} for rats was reported as 10.4 g/kg [26].

Interperitoneal administration of terephthalic acid in rats showed LD_{50} values > 1.43 g/kg [28] and 1.9 g/kg [31], with $LD_{100} = 3.2$ g/kg [31]. Mortality was found at 0.8 g/kg in mice and 1.6 g/kg in rats [27]. For dimethyl terephthalate, an LD_{50} of 3.9 g/kg has been reported [29], and for isophthalic acid, 4.3 g/kg [26].

Ingestion of terephthalic acid or dimethyl terephthalate results in rapid distribution and excretion in unchanged form [32]. In rats, at high ingestion rates of about 3 % terephthalic acid or dimethyl terephthalate in the feed, bladder calculi are formed which are mostly calcium terephthalate; these injure the bladder wall and lead

to cancer. Calculi cannot form unless the calcium terephthalate solubility is exceeded (threshold effect) [33]. Neither compound has been found to be genotoxic (Ames test) [34].

Inhalation of terephthalic acid dust appears to pose no clear hazards. Inhalation by rats for 6 h/d, 5 d/week, for 4 weeks at 25 mg/m^3 produced no fatalities [35]. Dimethyl terephthalate dust at 16.5 and 86.4 mg/m^3 for 4 h/d caused no toxicological effects after 58 days [29]. The rats intermittently reacted to the nuisance dust levels at the higher concentration.

Irritation of the skin, eyes, and mucous membranes by any of these chemicals is mild and reversible. In rabbits, the FHSA test for terephthalic acid for eyes gave scores of 14.0/110.0 and 3.5/110.0 at 1 and 24 h after exposure. The FHSA skin test score was 0.4/8.0 [35]. For isophthalic acid, eye and skin irritancy scores were 5.3/110.0 and 0.2/8.0 [36]. Irritation due to dimethyl terephthalate has also been reported to be mild [30].

Terephthalic acid, dimethyl terephthalate, and isophthalic acid can all form dust clouds. As with any flammable substance, an explosion can occur, given proper dust and oxygen concentrations. Reported limits on the explosive region for terephthalic acid dust clouds are minimum dust content of 40 g/m^3 and minimum oxygen content 12.4 % at 20 °C. At 150 °C, minimum oxygen is 11.1 % [37]. The maximum concentration of an explosible dust cloud has been calculated as 1400 g/m^3 [38].

Dimethyl terephthalate and isophthalic acid have a more stringent oxygen limit. In tests with carbon dioxide diluent, the minimum oxygen content was 12, 14, and 15 % for dimethyl terephthalate, isophthalic acid, and terephthalic acid [39]. The oxygen content is higher for carbon dioxide diluent than for nitrogen, owing to the higher heat capacity of carbon dioxide (the 12.4 % minimum oxygen content reported above for terephthalic acid was for nitrogen diluent).

For molten dimethyl terephthalate, the flash and fire points, determined by the Cleveland open-cup method, are 146 and 155 °C [1].

10. References

General References

1. *Kirk-Othmer,* **18,** 732 – 777.
2. P. Raghavendrachar, S. Ramachandran, *Ind. Eng. Chem. Prod. Res. Dev.* **31** (1992) 453 – 462.
3. J. E. McIntyre, *Chem. Eng. Monogr.* **15** (1982) 400 – 444.

Specific References

4. Calico Printers Assoc, GB 578 079, 1941 (J. R. Whinfield, J. T. Dickson).
5. Mid-Century Corp., US 2 833 816, 1955 (R. S. Barker, S. A. Soffer).
6. Mid-Century Corp., US 3 089 906, 1958 (R. S. Barker, S. A. Soffer).
7. W. Partenheimer in D. W. Blackburn (ed.): *Catalysis of Organic Reactions,* Marcel Dekker, New York 1990, pp. 321 – 346.
8. Standard Oil Company (Indiana), US 3 584 039, 1967 (D. H. Meyer).
9. Standard Oil Company (Indiana), US 3 931 305, 1973 (J. A. Fisher).
10. K. Matsuzawa, *Chem. Econ. Eng. Rev.* **8** (1976) no. 8, 25 – 30.
11. M. Hizikata, *Chem. Econ. Eng. Rev.* **9** (1977) no. 9, 32 – 38.
12. Mitsubishi Chem. Ind. Ltd., US 4 877 900, 1988 (A. Tamaru, Y. Izumisawa).
13. Eastman Kodak Company, US 4 447 646, 1983 (G. I. Johnson, J. E. Kiefer).
14. H. J. Korte, H. Schroeder, A. Schoengen: "The PTA Process of Hüls Troisdorf AG," AIChE Summer National Meeting, Denver, Co. 1988.
15. E. I. DuPont de Nemours, US 2 491 660, 1949 (W. F. Gresham).
16. Standard Oil Company (Indiana), US 2 976 030, 1957 (D. H. Meyer).
17. *Eastman News* **43** (1988) no. 14.
18. H. S. Bryant, C. A. Duval, L. E. McMakin, J. I. Savoca, *Chem. Eng. Prog.* **67** (1971) no. 9, 69 – 75.
19. A. P. Gelbein, M. C. Sze, R. T. Whitehead, *CHEMTECH* **3** (1973) 479 – 483.
20. Eastman Kodak Company, US 4 705 890, 1987 (G. R. Steinmetz, M. Rule).
21. *Chem. Eng. (N.Y.)* **83** (1976) no. 17, 27 – 28.
22. S. Shelley, *Chem. Eng. (N.Y.)* **98** (1991) no. 12, 90 – 93.
23. T. G. Otchy, K. J. Herbert: "First Large Scale Catalytic Oxidation System for PTA Plant CO and VOC Abatement," *Ann. Air Waste Managem. Assoc. Meet. 85th* 1992.
24. W. Berger, D. Dornig, *Faserforsch. Textiltech.* **29** (1978) 256 – 262.
25. *Chem. Markt. Rep.* **242** (1992) no. 4, 42.
26. J. V. Marhold, *Sb. Vys. Toxicologickeho Vysetrien Latid A Priravku* 1972, 52.

27. A. E. Moffitt et al., *Am. Ind. Hyg. Assoc. J.* **36** (1975) 633 – 641.
28. A. Hoshi, R. Yanal, K. Kuretani, *Chem. Pharm. Bull.* **16** (1968) 1655 – 1660.
29. W. J. Krasavage, *Am. Ind. Hyg. Assoc. J.* **34** (1973) 455 – 462.
30. Haskel Laboratories, unpublished data, Wilmington, Delaware, USA, 1979.
31. E. O. Grigas, R. Ruiz, D. M. Aviado, *Toxicol. Appl. Pharmacol.* **18** (1971) 469 – 486.
32. A. Hoshi, K. Kuretani, *Chem. Pharm. Bull.* **16** (1968) 131 – 135.
33. H. d'A. Heck, *Banbury Rep.* **25** (1987) 233 – 244.
34. E. Zeiger, S. Haworth, W. Speck, K. Mortelmans, *EHP Environ. Health Perspect.* **45** (1982) 99 – 104.
35. Amoco Chemical Company, Material Safety Data Sheet, Amoco TA-33, Chicago, Ill., 1993.
36. Amoco Chemical Company, Material Safety Data Sheet, Amoco PIA, Chicago, Ill., 1993.
37. A. D. Craven, M. G. Foster, *Combust. Flame* **11** (1967) 408 – 414.
38. S. Nomur, M. Torimoto, T. Tanake, *Ind. Eng. Chem. Proc. Res. Dev.* **23** (1984) 420 – 423.
39. M. Jacobson, J. Nagy, A. R. Cooper, *Rep. Invest. U.S. Bur. Mines* **5971** (1962) 25 – 26.

Terpene Resins → Resins, Synthetic

Terpenes

MANFRED EGGERSDORFER, BASF Aktiengesellschaft, Ludwigshafen, Germany

1. General 1
1.1. Definition and Basic Structures 1
1.2. Occurrence 2
1.3. Biosynthesis 2
1.4. Biodegradation 2
1.5. Importance and Extraction 4
2. Acyclic Monoterpenes 5
2.1. Myrcene 5
2.2. Ocimene 6
2.3. 2,6-Dimethyl-2,4,6-octatriene 6
3. Monocyclic Monoterpenes 6
3.1. *p*-Menthane 6
3.2. α-Terpinene 7
3.3. Terpinolene 7
3.4. *p*-Cymene 7
3.5. 1,8-Cineole 8
4. Bicyclic Monoterpenes 8
4.1. 3-Carene 8
4.2. Pinane 9
4.3. 2-Pinane Hydroperoxide 9
4.4. 2-Pinanol 10
4.5. Camphene 10
5. Acyclic Sesquiterpenes 10
6. Monocyclic Sesquiterpenes 11
6.1. Bisabolene 11
6.2. Bisabolol 11
7. Bicyclic Sesquiterpenes 12
8. Tricyclic Sesquiterpenes 12
9. Acyclic Diterpenes 12
10. Acyclic Triterpenes 13
11. Toxicology 13
12. References 13

1. General

1.1. Definition and Basic Structures

Terpenes are natural products, whose structures are built up from isoprene units. They are classified according to the number of these units:

Monoterpenes	C_{10}
Sesquiterpenes	C_{15}
Diterpenes	C_{20}
Sesterpenes	C_{25}
Triterpenes	C_{30}
Tetraterpenes	C_{40}

Terpenes can be acyclic, or mono-, bi-, tri-, tetra-, and pentacyclic with 3-membered to 14-membered rings [1–7].

Deviations from the rule occur through rearrangement reactions or degradation of parts of the molecule during biosynthesis. The broader term terpenoids also covers natural degradation products, such as ionones, and natural and synthetic derivatives, e.g., terpene alcohols, aldehydes, ketones, acids, esters, epoxides, and hydrogenation products.

An overview of the most important basic structures demonstrates the structural variety of terpenes [8]. The compounds are normally known by their trivial names, as the systematic nomenclature is frequently less practicable for complicated structures.

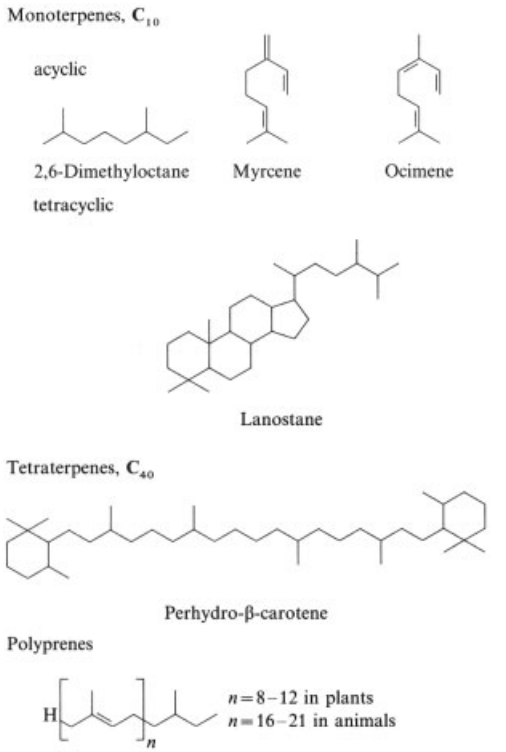

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_205

monocyclic

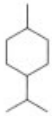

p-Menthane

Limonene

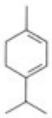

α-Terpinene

p-Cymene

Safrane

bicyclic

Pinane

α-Pinene

Carane

Camphor

Sesquiterpenes, $\mathbf{C_{15}}$

acyclic

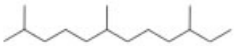

Farnesane

monocyclic

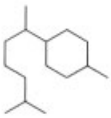

Bisabolane

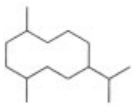

Germacrane

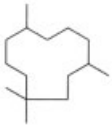

Humulane

bicyclic

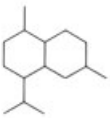

Cadinane

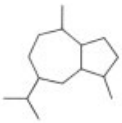

Guaiane

tricyclic

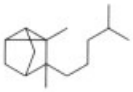

α-Santalane

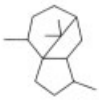

Patchoulane

Diterpenes, $\mathbf{C_{20}}$

acyclic

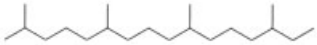

Phytane

monocyclic

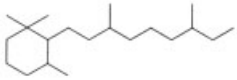

Retinane

bicyclic

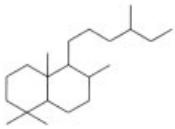

Labdane

tricyclic

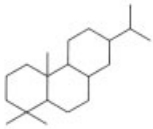

Abietane

tetracyclic

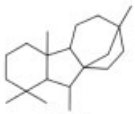

Gibbane

pentacyclic

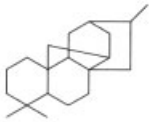

Atisane

Triterpenes, $\mathbf{C_{30}}$

acyclic

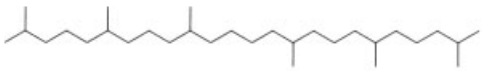

Squalane

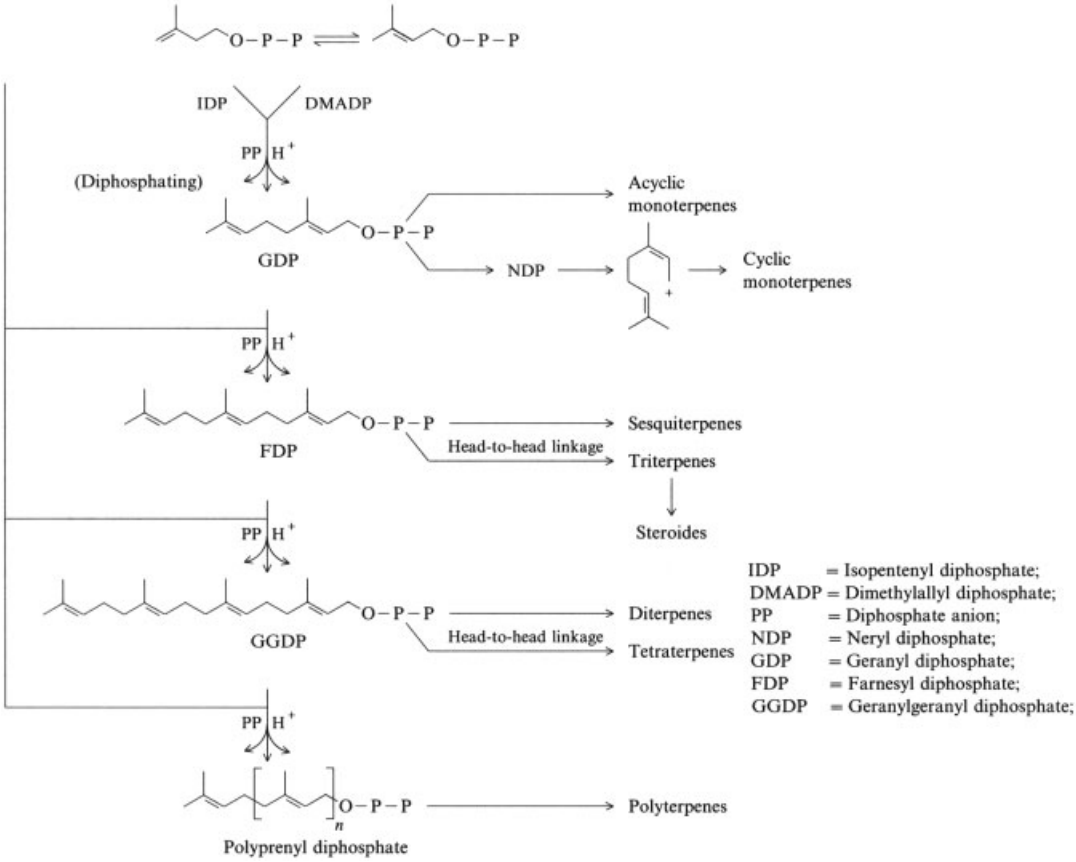

Scheme 1. Biosynthesis of terpenes

1.2. Occurrence

Terpenes occur everywhere and in all organisms, in particular in higher plants. Certain components or their combinations are often characteristic of individual types of plant, so terpenes are used for chemotaxonomy [8]. In vegetable raw materials mono- and sesquiterpenes are predominantly found in essential oils [9, 10], sesqui-, di-, and triterpenes in balsams and resins [11], tetraterpenes in pigments (carotenoids) [12], and polyterpenes in latexes [13]. Animal organisms also contain terpenes and terpenoids, which are predominantly incorporated by consumption of plants.

1.3. Biosynthesis

The biosynthesis of terpenes was explained by F. LYNEN. Starting from (3 *R*)-(+)-mevalonic acid (C_6), isopentenyl diphosphate (IDP) and dimethylallyl diphosphate (DMADP) [8, 14–16] are formed by decarboxylation and isomerization as active C_5 building blocks (Scheme 1). (3 *R*)-(+)-Mevalonic acid is formed from three C_2 units from acetyl-CoA [8]. Plants synthesize IDP in the chloroplasts, mitochondria, and microsomes; animals synthesize IDP in the liver.

Monoterpenes are formed by stereospecific condensation of IDP with DMADP, whereby a diphosphate anion (PP) is eliminated from DMADP. Geranyl diphosphate is thus formed in a head-to-tail condensation. Neryl diphosphate with the *Z*-double bond is formed when the *S*-proton in isopentenyl diphosphate is eliminated stereospecifically. Neryl diphosphate is also formed from geranyl diphosphate by double bond isomerization.

The hydrolysis of geranyl or neryl diphosphate by a prenolpyrophosphatase gives the

monoterpene alcohols geraniol and nerol. The acyclic monoterpene hydrocarbons, such as myrcene and ocimene, are formed by dehydration and isomerization of geraniol. The monoterpene aldehydes geranial and neral are formed from geraniol, and the biosynthesis of cyclic monoterpenes involves the loss of diphosphate from neryl diphosphate. The carbonium ion formed is stabilized by ring closure, sometimes after rearrangement, elimination of a proton, or addition of an anion.

The head-to-tail condensation of IDP with geranyl diphosphate gives farnesyl diphosphate, which rearranges further to give geranylgeranyl diphosphate. A large number of sesqui-, di-, and polyterpenes are formed from farnesyl and geranyl diphosphates in the same way as the monoterpenes. Triterpenes (and hence, steroids) are formed by head-to-head condensation of two C_{15} units (farnesyl diphosphate), and tetraterpenes from two molecules of geranylgeranyl diphosphate.

1.4. Biodegradation

Terpenes are degraded by microorganisms, such as Pseudomonas and Aspergillus species [17]. Generally both acyclic and cyclic terpenes are oxidized. Plants are known to be capable of degrading as well as forming terpenes.

In animals acyclic terpenes are degraded by ω-, α- and/or β-oxidation and cyclic terpenes by hydroxylation (elimination as glucuronides).

1.5. Importance and Extraction

Since the study of organic chemistry began, it has been intensively concerned with terpenes. Many reactions, such as the Wagner – Meerwein rearrangement, theories of conformational analysis, and the Woodward – Hoffmann rules have their basis in terpene chemistry. The postulation of the isoprene rule by O. Wallach (1887) and its confirmation and extension by L. Ruzicka (1921, 1953) gave this area of chemistry direction and unity.

Monoterpenes and a few sesquiterpenes are important economically as perfumes and fragrances (→ Perfumes, → Flavors and Fragrances). Depending on the properties, terpenes (mostly essential oils) are obtained from the corresponding plants by steam distillation, extraction, or enfleurage. The pure terpenes are mostly isolated from the extracts by fractional distillation.

As readily available, optically active products, some monoterpenes, such as (−)-α-pinene and (−)- and (+)-limonene are used for the synthesis of other optically active products and reagents for cleaving racemates [18].

Various essential oils and the terpenes isolated from them are still used in pharmaceutical preparations, e.g., turpentine, camomile oil, eucalyptus oil, and camphor. The activity of some plants used in medicine and flavoring is due to their terpene content [19]. Terpenes are also used in the production of synthetic resins [20–22], and as solvents or diluting agents for dyes and varnishes [23].

In the following sections only the industrially most important terpenes are dealt with, insofar as they are not part of other articles (→ Flavors and Fragrances, → Turpentines, and → Tall Oil). For natural resins, gibberellic acid (→ Plant Growth Regulators), vitamins A and E (→ Vitamins), triterpenes (→ Steroids), tetraterpenes, and the other terpenes the specialist literature must be consulted.

2. Acyclic Monoterpenes

The following acyclic terpenes and terpene derivatives are described elsewhere (→ Flavors and Fragrances):

Alcohols: geraniol, nerol, linalool, myrcenol, citronellol, dihydromyrcenol, tetrahydrogeraniol, and tetrahydrolinalool.

Aldehydes and acetals: citral, citraldimethylacetal, citronellal, hydroxydihydrocitronellal, hydroxydihydrocitronellal dimethylacetal, and methoxydihydrocitronellal

Esters and nitriles: geranyl formate, acetate, isobutyrate, isovalerate, phenylacetate, and propionate; neryl acetate; linalyl acetate, butyrate, formate, isobutyrate, and propionate; lavanduyl acetate; and citronellyl acetate, formate, isobutyrate, isovalerate, propionate, and tiglate.

2.1. Myrcene

Myrcene [*123-35-3*], 2-methyl-6-methylene-2,7-octadiene, $C_{10}H_{16}$, M_r 136.23, is a colorless liquid with a pleasant odor.

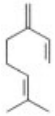

Some physical properties are listed below:

bp	166 – 168 °C (101.3 kPa)
d^{20}	0.7905
n_D^{20}	1.4697
mp	50 °C

Myrcene is very reactive. It polymerizes slowly at room temperature. For this reason it is stored in a cool place from which light and air are excluded. It can be stabilized by the addition of 0.1 % 4-(*tert*-butyl)catechol.

Although myrcene occurs naturally in many organisms, its extraction is uneconomic; it is produced industrially by pyrolysis of β-pinene [24].

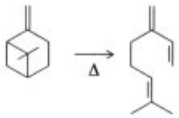

The fragmentation of linalool and linalyl acetate and the catalytic dimerization of isoprene have been described as laboratory methods [25].

Because of its functionality, myrcene is the starting material for a range of industrially important products, e.g., geraniol, nerol, linalool, and isophytol [26]. Addition of hydrogen chloride is industrially important. Depending on reaction conditions, this gives 1-chloro-3,7-dimethyl-2,6-octadiene (geranyl chloride and/or neryl chloride), 3-chloro-3,7-dimethyl-1,6-octadiene (linaloyl chloride), and 7-chloro-7-methyl-3-methylen-1-octene (myrcenyl chloride) [27]. The alcohols are produced from the chlorides by saponification.

The 1,3-diene system of myrcene reacts with a large number of dienophiles. For example, the maleic anhydride adduct is used to characterize and determine the quality of the product [28].

Isophytol is formed by transition-metal-catalyzed addition of malonic ester and subsequent acidification [29].

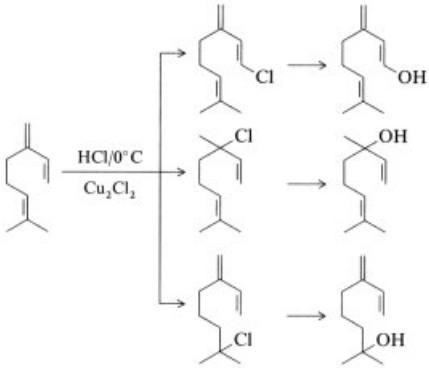

In the hydrogenation of myrcene 2,6-dimethyl-2,6-octadiene, 2,6-dimethyl-2,7-octadiene, and 2,6-dimethyloctane are formed, depending on the catalyst and the reaction conditions [30:1H 260, 1EIII 1054].

Besides its main use as an intermediate for the production of terpene alcohols, myrcene is used in the production of terpene polymers [20], terpene – phenol resins [21], and terpene – maleate resins [22]. It can also be used as a solvent or diluting agent for dyes and varnishes [23].

2.2. Ocimene

Ocimene [*673-84-7*], (*Z, E*)-2,6-dimethyl-2,5,7-octatriene, $C_{10}H_{16}$, M_r 136.23, is a colorless liquid with a pleasant odor.

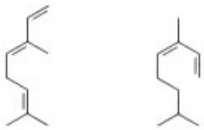

The *Z*/*E* mixture has the following properties:

bp	176 – 178 °C (101.3 kPa)
d^{20}	0.800
n_D^{20}	1.4862
mp	50 °C

Ocimene is very sensitive to oxidation and therefore can be kept for long periods only with the exclusion of air. At elevated temperatures it rearranges to alloocimene [31].

Ocimene is produced by the flash pyrolysis or photochemical isomerization of α-pinene [32].

Ocimene is mainly used as a perfume component. The derivatives ocimenol (2,6-dimethyl-5,7-octadien-2-ol), produced by acid-catalyzed

hydration, and the partially and completely hydrogenated products are also important [33, 34].

2.3. 2,6-Dimethyl-2,4,6-octatriene

2,6-Dimethyl-2,4,6-octatriene exists as two stereoisomers:

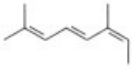

[*7216-56-0*]
(4*E*,6*Z*)-2,6-Dimethyl-2,4,6-octatriene
(Neoalloocimen)

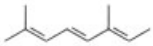

[*3016-19-1*]
(4*E*,6*E*)-2,6-Dimethyl-2,4,6-octatriene
(Alloocimen)

(4*E*, 6*Z*)-2,6-Dimethyl-2,4,6-octatriene, $C_{10}H_{16}$, M_r 136.23, is a colorless liquid with an intense odor, and is sensitive to oxidation.

bp	80 °C (1.9 kPa)
d^{20}	0.8161
n_D^{20}	1.5437

(4*E*, 6*Z*)-2,6-Dimethyl-2,4,6-octatriene is produced as a mixture with other products by the thermolysis of α-pinene at 350 °C [35], or by heating *Z*-ocimene to 190 °C [36]. The products are separated by fractional distillation.

(4*E*, 6*E*)-2,6-Dimethyl-2,4,6-octatriene, $C_{10}H_{16}$, M_r 136.23, has physical properties similar to those of the (4*E*, 6*Z*)-isomer.

bp	79 °C (1.9 kPa)
d^{20}	0.8106
n_D^{20}	1.5438

(4*E*, 6*E*)-2,6-Dimethyl-2,4,6-octatriene is formed as a mixture with the (4*E*, 6*Z*)-isomer and other products in the thermolysis of pinene at 320 – 330 °C [37], by dehydration of 3,7-dimethyl-3,6-octadien-1-ol [38], and by treating *Z*-ocimene with base [39].

2,6-Dimethyl-2,4,6-octatriene is used to a small extent in the perfume industry. It is also used as a diluting agent for varnishes and dyes [23], and as a component for terpene polymers [20, 22].

3. Monocyclic Monoterpenes

3.1. *p*-Menthane

p-Menthane [*99-82-1*], 1-isopropyl-4-methylcyclohexane, $C_{10}H_{20}$, M_r 140.25, occurs widely as a *cis*/*trans* mixture, e.g., in essential oils in the eucalyptus fruit.

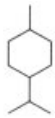

The *cis*/*trans* mixture of *p*-menthane is a colorless liquid with a fennel-like odor.

cis-1-Isopropyl-4-methylcyclohexane

bp	168.8 °C (99.2 kPa)
d^{20}	0.8086
n_D^{20}	1.4443

trans-1-Isopropyl-4-methylcyclohexane

mp	−86 °C
bp	168.1 °C (99.2 kPa)
d^{20}	0.7941
n_D^{20}	1.4369

p-Menthane is obtained as a *cis*/*trans* mixture by the catalytic hydrogenation of limonene, terpinols, and *p*-cymene [40]. Raney nickel, platinum, or copper and aluminum oxides are used as catalyst [41].

p-Menthane is predominantly used as a precursor in the production of 1-isopropyl-4-methylcyclohexane hydroperoxide, a catalyst for radical polymerization [42].

3.2. α-Terpinene

α-Terpinene [*99-85-4*], 4-isopropyl-1-methyl-1,3-cyclohexadiene, $C_{10}H_{16}$, M_r 136.23, is the main component of terpinene, the other components being the β- and γ-isomers. It occurs in various essential oils, almost always as a mixture with its isomers.

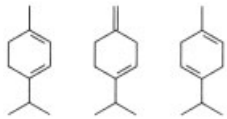

α β γ

α-Terpinene is a colorless liquid with a lemon-like odor. It resinifies on prolonged storage.

bp	172.5 – 173.5 °C
n_D^{20}	1.4780

α-Terpinene is obtained by fractional distillation of pine oil [43] and by the thermolysis of α-pinene in the presence of catalysts, such as manganese oxide [44].

α-Terpinene is used in essential oils and as an intermediate for perfumes.

3.3. Terpinolene

Terpinolene [*586-62-9*], 4-isopropylidene-1-methyl-1-cyclohexane, $C_{10}H_{16}$, M_r 136.23, occurs in many essential oils and in sulfate turpentine.

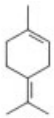

Terpinolene is a colorless liquid with an odor resembling that of turpentine.

bp	186 °C
d^{20}	0.8623
n_D^{20}	1.4861

It forms a hydroperoxide on autoxidation [45, 46].

Terpinolene used to be extracted by fractional distillation of wood turpentine [46]. It is now produced by treating α-pinene with aqueous H_3PO_4 at 75 °C [47]. It is preferable to distill terpinolene under vacuum to minimize the formation of polymeric products at elevated temperatures.

Terpinolene is mainly used to improve the odor of industrial and household products [48–50].

3.4. *p*-Cymene

p-Cymene [*99-87-6*], 4-(isopropyl)methylbenzene, $C_{10}H_{14}$, M_r 134.21, is the main component of sulfite turpentine and occurs in many essential oils [51].

p-Cymene is a colorless liquid with an odor typical of aromatic hydrocarbons.

mp	−67.7 °C
bp	177.3 °C
d^{20}	0.8573
n_D^{20}	1.4906

The enthalpy of evaporation is 44.670 J/mol.

p-Cymene undergoes catalytic hydrogenation to 1-isopropyl-4-methylcyclohexane [40]. It must be stored in the absence of light and air.

p-Cymene is formed during the sulfite leaching of wood [52]. A range of production processes starting from mono- and bicyclic terpenes have also been described. For example, *p*-cymene is obtained in good yields from α-pinene in the presence of copper catalysts [53, 54]. In this process a significant proportion of 1-isopropyl-4-methyl-cyclohexane is also formed, which is dehydrogenated in the presence of palladium catalysts at 260–280 °C to*p*-cymene [55]. The Friedel–Crafts alkylation of toluene with propene is used industrially [56]. Products containing more than one isopropyl group are converted into *p*-cymene with aluminum chloride and hydrochloric acid [57].

p-Cymene is used to improve the odor of soaps and as a masking odor for industrial products. Its use as an intermediate for musk perfumes [58], the oxidation to terephthalic acid, and its addition to antiseptic preparations have also been described [59]. *p*-Cymene is also used as a solvent for dyes and varnishes.

3.5. 1,8-Cineole

1,8-Cineole [*470-82-6*], 1,8-epoxy-*p*-menthane, eucalyptol, $C_{10}H_{18}O$, M_r 154.25, occurs widely in natural essential oils, mainly eucalyptus oil.

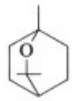

1,8-Cineole is a colorless liquid. Its odor and taste resemble those of camphor.

mp	1.55 °C
bp	176–176.4 °C
d^{20}	0.9232
n_D^{20}	1.4575

On warming, particularly in the presence of acids, 1,8-cineole forms other terpenes of the *p*-menthane type. It reacts with a range of substances, forming sparingly soluble products.

1,8-Cineole is extracted exclusively from eucalyptus oils with a high 1,8-cineole content. Various processes are used to separate it from the other terpenes, for example, treatment with H_2SO_4 in the cold [60], distillation in the pres-

ence of phenols, such as cresols or resorcinol [61], which form loose addition compounds, or by addition to β-naphthol [62]. 1,8-Cineole can also be enriched by rectification. The yield is increased by gasification in the presence of a chromium – nickel catalyst [63].

1,8-Cineole is used as a perfume and fragrance. In the past it was also used as an expectorant and an antiseptic.

4. Bicyclic Monoterpenes

The industrially most important bicyclic monoterpenes are subdivided into caranes, pinanes, and bicyclo[2.2.1]heptanes, depending on their basic structure.

4.1. 3-Carene

3-Carene [*498-15-7*], 3,7,7-trimethylbicyclo-[4.1.0]hept-3-ene, $C_{10}H_{16}$, M_r 136.23, occurs in a range of turpentines as the enantiomerically pure compound, but also as the racemate [64, 65].

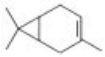

3-Carene is a colorless compound with a sweet, penetrating odor.

bp	176 – 176.4 °C
d_4^{20}	0.9232
n_D^{20}	1.4575
$[\alpha]_D^{25}$	+17.6 ° (undiluted)

It readily undergoes autoxidation and resinification in air. Addition of pyrogallol or resorcinol as stabilizers is recommended [66]. On treatment with peracetic acid in glacial acetic acid and subsequent saponification, 3-carene is oxidized to 3,4-caranediol [67].

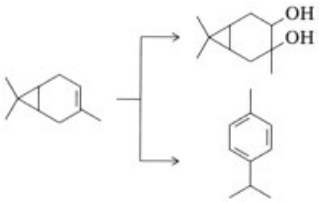

Pyrolysis at 300 – 580 °C in the presence of Fe^{III} oxide on carriers gives *p*-cymene as the main product [68].

3-Carene is obtained exclusively by fractional distillation of certain turpentines [69].

3-Carene is predominantly used in the perfume industry for producing essential oils [70]. It is also an intermediate in perfume synthesis and a precursor for insecticides of the pyrethroid type [71, 72].

4.2. Pinane

Pinane [*473-55-2*], 2,6,6-trimethylnorpinane, 2,6,6-trimethylbicyclo[3.1.1]heptane, $C_{10}H_{18}$, M_r 138.25.

cis-(1*S*) *trans*-(1*S*)

Pinane derivatives occur in wood, the leaves of many plants, algae, and insects. Pinane itself apparently does not occur naturally. However, it is formed on hydrogenation of pinane derivatives [73].

cis-Pinane (colorless liquid)

bp	167.2 – 168 °C
d_4^{20}	0.8575
n_D^{20}	1.4626
$[\alpha]_D^{25}$	−23.6 ° (undiluted)

trans-Pinane (colorless liquid)

bp	164 °C
d_4^{20}	0.854
n_D^{20}	1.4610
$[\alpha]_D^{25}$	+21.4 ° (undiluted)

The oxidation of pinane with air or oxygen gives 2-pinane hydroperoxide or 2-pinanol [74]. In industry, oxidation is carried out in the presence of catalysts [75].

The pyrolysis of (−)-(1*S*)-*cis*-pinane at 500 °C gives a mixture of (3*R*)-3,7-dimethyl-1,6-octadiene and (1*R*)-isopropyl-2,3-dimethylcyclopentane [76].

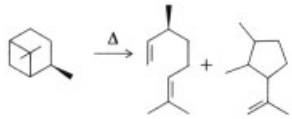

cis-Pinane is obtained, together with small quantities of *trans*-pinane, by the catalytic hydrogenation of α- and β-pinene. If (−)-α- and (−)-β-pinene are hydrogenated, the (−)-enantiomer is obtained, and from the (+)-pinenes, (+)-*cis*-pinane. In industry, α-pinene

is predominantly used for the hydrogenation. Raney nickel, Pt, or Pd (and their oxides) on carriers are used as catalysts [77,78]. The hydrogenation with Raney nickel is carried out at 2 – 10 MPa at 150 °C. In the presence of halogen compounds *cis*-pinane is obtained predominantly [79].

trans-Pinane can be obtained from β-pinene by treatment with sodium borohydride in the presence of boron trifluoride in di(ethylene glycol) and subsequent heating of the reaction mixture in propionic acid [80].

Pinane is used in industry for the production of pinane hydroperoxide [81]. It is also important as an intermediate in the production of 3,7-dimethylocta-1,6-diene, which is used to produce perfumes, such as citronellol, citronellal, and hydroxycitronellal.

4.3. 2-Pinane Hydroperoxide

2-Pinane hydroperoxide [*5405-84-5*], 2,6,6-trimethylbicyclo[3.1.1]heptane-2-hydroperoxide, $C_{10}H_{18}O_2$, M_r 170.25, is a colorless liquid, which is readily flammable and insoluble in water.

(−)-(1*R*)

bp	48 – 52 °C (1.3 Pa)
d_4^{20}	1.021
n_D^{20}	1.4898
$[\alpha]_D^{25}$	−27.4 ° (undiluted)

On heating 2-pinane hydroperoxide to > 110 °C 1-((1*R*)-*cis*-3-ethyl-2,2-dimethylcyclobutyl)ethanone is formed [82]. (1*R*, 2*S*)-2-pinanol is formed in good yields by catalytic hydrogenation.

2-Pinane hydroperoxide is produced by oxidation of pinane with air or oxygen at 95 °C [83–85]. *cis*-Pinane reacts more rapidly than *trans*-pinane [86].

2-Pinane hydroperoxide is used as a radical initiator for polymerization reactions [87], e.g., for the polymerization of diolefins and aromatic vinyl compounds, or for hardening unsaturated polyester resins. It is also an intermediate in the production of perfumes, such as pinanol, linalool, nerol, and geraniol [88].

4.4. 2-Pinanol

2-Pinanol [*473-54-1*], 2,6,6-trimethylbicyclo[3.1.1]heptan-2-ol, $C_{10}H_{18}O$, M_r 154.25, is obtained as the *cis* or *trans* isomer, depending on the production process.

cis-(−)-(1*R*,2*S*)

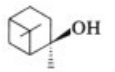

trans-(−)-(1*R*,2*R*)

cis-2-Pinanol forms colorless crystals with a camphor-like odor.

mp	78 – 79.2 °C
bp	90 – 91 °C (1.5 kPa)
$[\alpha]_D^{20}$	−24.5 ° ($CHCl_3$, C = 8 g/L)

trans-2-Pinanol forms colorless needles.

mp	58 – 59 °C
bp	88 – 89 °C (0.12 kPa)
$[\alpha]_D^{17}$	−2.3 ° (ether, 10 g/L)

The pyrolysis of 2-pinanol at 500 °C gives linalool. (−)-Linalool is formed from (+)-*cis*- and (−)-*trans*-2-pinanol, and (+)-linalool from (−)-*cis*- and (+)-*trans*-2-pinanol. The process is used industrially [89,90].

cis-2-Pinanol is produced industrially by the catalytic hydrogenation of 2-pinane hydroperoxide [91,92]. Alternatively 2-pinane hydroperoxide can be treated with sodium sulfide in aqueous sodium hydroxide [93] or with sodium methoxide [94]. *cis*-2-Pinanol can also be obtained directly from pinane by air oxidation in the presence of alkalis, such as sodium hydroxide, at 80 – 100 °C [95].

2-Pinanol is used to produce linalool by pyrolysis [96].

4.5. Camphene

Camphene [*5794-03-6*], 2,2-dimethyl-3-methylenenorbornane, $C_{10}H_{16}$, M_r 136.23, occurs in a large number of essential oils in optically active form, both as the *R* and *S* enantiomers.

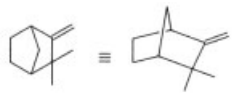

(+)-(1 *R*)

Camphene is a colorless, crumbly crystalline solid with a camphor-like odor. It has a tendency to sublime.

mp	52.5 °C
bp	157.8 °C (98.8 kPa)
d_4^{25}	0.84225
n_D^{54}	1.4564
$[\alpha]_D^{20}$	+106.8 ° ($CHCl_3$, 44 g/L)

Camphene is stable in air and light. At elevated temperatures, in the presence of oxygen, it undergoes autoxidation to camphenilone [97]. Peracids attack the double bond, giving camphene oxide [98]. Catalytic hydrogenation gives the saturated hydrocarbon isocamphane [99].

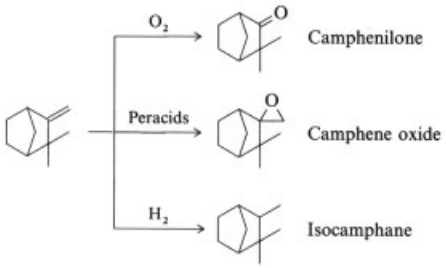

The extraction of camphene by distillation of turpentine is now hardly used in industry. Instead camphene is produced from α-pinene.

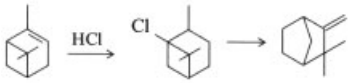

(+)-α-Pinene is converted into (+)-bornyl chloride by the action of dry hydrogen chloride in a Wagner – Meerwein rearrangement. Base-catalyzed dehydrohalogenation of the (+)-bornyl chloride gives racemic camphene [100, 101]. The reaction of α-pinene with borophosphoric acid in the gas phase [102], or on TiO_2 catalysts [103] has also been described.

Camphene is used to improve the odor of industrial products. It is an intermediate in the production of camphor, isobornyl esters, and the insecticide Toxaphen.

5. Acyclic Sesquiterpenes

Farnesene α-Farnesene, 2,6,10-trimethyl-2,6,9,11-dodecatetraene, $C_{15}H_{24}$, M_r 204.36.

α- and β-Farnesenes and their *Z*- and *E*-isomers occur in many essential oils, e.g., in apples [104], citrus fruits [105], and hop oil. α-Farnesene is a natural attractant for the larvae of the codlin moth [106]. It is also found in the gland secretions of certain ants [107].

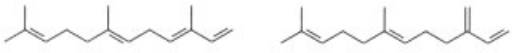

α-Farnesene [502-61-4] β-Farnesene [18794-84-8]

α- and β-Farnesenes are colorless liquids with fruity odors. They are sensitive to autoxidation and can be stored for long periods only with the exclusion of light and air.

α-Farnesene	
bp	98 – 102 °C (0.5 kPa)
n_D^{25}	1.4790
β-Farnesene	
bp	95 – 107 °C (0.5 kPa)
d^{20}	0.8363
n_D^{20}	1.4899
Mixture	
bp	129 – 132 °C (1.2 kPa)
d^{18}	0.8770
n_D^{20}	1.4995

Mixtures of α- and β-farnesenes are obtained by heating farnesol with potassium hydrogen sulfate to 160 – 170 °C [108]. Other syntheses of the mixture, involving dehydration of farnesol, nerolidol, and isoprene, have also been described [109, 110].

Farnesene is used in small quantities in the fragrance industry.

6. Monocyclic Sesquiterpenes

6.1. Bisabolene

Bisabolene, $C_{15}H_{24}$, M_r 204.33, occurs in myrrh oil and limett oil. The isomers α-, β-, and γ-bisabolene are known.

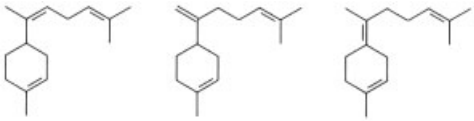

α-Bisabolene [17627-44-0] β-Bisabolene [495-61-4] γ-Bisabolene [495-62-5]

α-Bisabolene, (*E, Z*)-4-(1,5-dimethyl-1,4-hexadienyl)-1-methyl-1-cyclohexane

bp 95 °C (0.02 kPa) β-Bisabolene, 1-methyl-4-(5-methyl-1-methylene-4-hexenyl)-1-cyclohexane

bp	148 °C (2.4 kPa)
n_D^{15}	1.4893
$[\alpha]_D^{15}$	−67 ° (undiluted)

γ-Bisabolene, (*E, Z*)-4-(1,5-dimethyl-4-hexenylidene)-1-methyl-1-cyclohexene

bp	132 °C (1.4 – 1.6 kPa)
n_D^{25}	1.4928

The individual isomers and their mixtures are colorless liquids with pleasant, balsamic odors.

Isomer mixtures are produced by treating nerolidol and farnesol with acid [111]. Targeted syntheses of α- and β-bisabolene by the Wittig reaction have been described [112–114].

Bisabolene mixtures are used in the perfume and fragrance industries.

6.2. Bisabolol

Bisabolol [*515-69-5*], $C_{15}H_{26}O$, M_r 222.37, occurs in various plants.

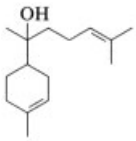

(+)-α-Bisabolol, (+)-2-methyl-6-(4-methyl-3-cyclohexenyl)-6-hepten-2-ol

bp	120 – 122 °C (0.13 kPa)
d_4^{20}	0.9213
n_D^{20}	1.4919
$[\alpha]_D^{20}$	+51.7 ° (undiluted)

(−)-α-Bisabolol, (−)-2-methyl-6-(4-methyl-3-cyclohexenyl)-6-hepten-2-ol

bp	153 °C (1.59 kPa)
d_4^{20}	0.9211
n_D^{20}	1.4936
$[\alpha]_D$	−55.7 ° (undiluted)

rac-Bisabolol, 2-methyl-6-(4-methyl-3-cyclohexenyl)-6-hepten-2-ol

bp	157 °C (1.59 kPa)
d_4^{23}	0.9223
n_D^{23}	1.4917

The bisabolols are colorless liquids with slightly flowery odors.

rac-Bisabolol is produced by acid-catalyzed cyclization of farnesol or nerolidol [115]. The individual enantiomers are obtained by extraction from the appropriate plants.

Because of their anti-inflammatory and spasmolytic properties [117], both (−)- and *rac*-bisabolol are predominantly used in the cosmetics industry and to a small extent in the pharmaceutical industry.

7. Bicyclic Sesquiterpenes

Caryophyllene Caryophyllene [*87-44-5*], 6,10,10-trimethyl-2-methylenebicyclo[7.2.0]-5-undecene, $C_{15}H_{24}$, M_r 204.36, occurs in many essential oils.

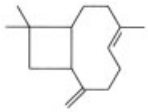

Caryophyllene is a colorless, oily liquid with an odor resembling that of cloves.

bp	254 – 256 °C
d_4^{20}	0.9019
n_D^{20}	1.4995
$[\alpha]_D^{20}$	−9.2 ° (undiluted)

Catalytic hydrogenation on platinum gives tetrahydrocaryophyllene [118], and on palladium a dihydrocaryophyllene [119]. On oxidation with potassium permanganate, a glycol (*mp* 120 – 120.5 °C) is formed [120], and with perbenzoic acid [121] or peracetic acid, a caryophyllene oxide.

Caryophyllene is best extracted from clove oil. It can also be obtained from other sources, such as certain American pine oil fractions.

Caryophyllene is used as a perfume and fragrance, for example in chewing gum (ca. 200 mg/kg). It is also used as a fixative. It can be used as an intermediate in the synthesis of other perfumes and fragrances [123].

8. Tricyclic Sesquiterpenes

Longifolene Longifolene [*475-20-7*], 3,3,7-trimethyl-8-methylene-tricyclo[5.4.0.9-*Z*9]-undecane, $C_{15}H_{24}$, M_r 204.36, occurs in the *d*- and *l*-forms in a range of essential oils. Indian turpentine contains up to 20 % longifolene.

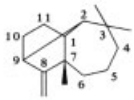

(+)-(1*R*, 2*S*, 7*S*, 9*S*)

Longifolene is a colorless, oily liquid.

bp	254 – 256 °C (93.9 kPa)
d_4^{18}	0.9319
n_D^{20}	1.5040
$[\alpha]_D$	+42.73 ° (undiluted)

Longifolene forms crystalline products with hydrogen halides [124]. On treatment with bromine followed by *N,N*-dimethylaniline, ω-bromolongifolene is formed, and with nitrogen oxides ω-nitrolongifolene [125].

Longifolene is obtained by distillation, e.g., Indian turpentine.

Longifolene is used as a solvent additive. It has been described as a starting material for the production of perfumes, e.g., by oxidation [126], or by treatment with formic acid [127].

9. Acyclic Diterpenes

Phytol Phytol [*150-86-7*], (7*R*, 11*R*)-3,7,11,15-tetramethyl-*trans*-2-hexadecen-1-ol, $C_{20}H_{40}O$, M_r 295.52, is formed by saponification of chlorophyll.

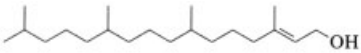

Phytol is a colorless liquid with a slightly flowery odor. The action of acids or heat effects dehydration to phytadiene.

bp	136 °C (1.3 kPa)
d_4^{21}	0.8533
n_D^{25}	1.4637

Phytol can be obtained from natural raw materials by a variety of processes [128]. The industrial synthesis of phytol starts from isophytol. The latter is first treated with formic acid to give phytyl formate, from which (*Z, E*)-phytol (isomer ratio 3:7) is obtained by saponification or transesterification [129].

Phytol is used in the perfume and cosmetics industries [130]. It can also be used, like isophytol, as the starting material for the synthesis of vitamins E and K_1 [131–133]. The hydrogenation and oxidation products are also known [134].

10. Acyclic Triterpenes

Squalene Squalene [*111-02-4*], 2,6,10,15,19,23-hexamethyltetracosa-2,*t*-6,*t*-10,*t*-14,*t*-18,*r*-22-hexaene, $C_{30}H_{50}$, M_r 410.73, occurs in liver oils of various species of shark [135], in cod-liver oil [136], in vegetable fats and oils [137], and in human skin fat.

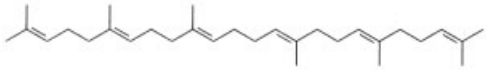

Squalene is a colorless, air-sensitive, almost odorless liquid.

bp	223 – 226 °C (0.26 kPa)
d_4^{20}	0.8577
n_D^{25}	1.4961

Squalene is extracted from shark liver [138] or synthesized from hexaphenyl-1,4-butanediyldiphosphonium dibromide and 6,10-dimethyl-5,9-undecadien-2-one (geranylacetone, industrial intermediate in the vitamin E synthesis) [139].

Squalene is hydrogenated on platinum or nickel catalysts to dodecahydrosqualene (perhydrosqualene, squalane, cosbiol) [140], which is widely used in cosmetics [141, 142].

11. Toxicology

There are excellent reviews of the toxicology of terpenes [143]. The known toxicological test results (Table 1) show that most terpenes have low acute oral toxicity and, in spite of occasionally good skin resorption, low dermal toxicity. Some are irritant to the skin of rabbits in concentrated form, but a 4 – 16 % solutions in vaseline, are generally not irritant to humans. Only 3-carene, which occurs in turpentine, has been shown to have a sensitizing potential in humans and animals.

To investigate detoxification in mammals, a metabolism study on rabbits has been carried out. On oral administration several characteristic oxidation reactions have been detected:

1) Stereoselective oxidation: formation of 3-caren-9-ol, 3-caren-9-carboxylic acid, and 3-caren-9,10-dicarboxylic acid from 3-carene

Table 1. Toxicity data for terpenes

Substance	Acute oral toxicity (rat), LD_{50}, g/kg	Acute oral toxicity (rabbit), LD_{50}, g/kg	Skin resorption	IRT∗ (rat)	Skin irritation (R = rabbit; H = human)	Irritation of rabbit eye mucous membrane	Sensitization (H = human; G = guinea pig)	Specific properties/tests	References
Myrcene	>5	>5	+		R: moderately irritant; H: 4 % not irritant		H: negative		[145]
Ocimene	>5	>5			R: moderately irritant; H: 5 % not irritant		H: negative		[145]
α-Terpinene	1.68				H: 5 % not irritant		H: negative	damage to liver and blood (forms methemoglobin)	[145]
Terpinolene	4.39 ml/kg	>5			R: not irritant; H: 20 % not irritant		H: negative		[145]
p-Cymene	4.74	>5	+	0/6 after 6 h	H: irritant; R: not irritant; H: 4 % not irritant	irritant	H:negative	affects central nervous system (narcosis)	[145, 146]
1,8-Cineole	2.48	>5			R: not irritant; H: 16 % not irritant		H: negative	cancerization study in progress	[145, 146]
3-Carene	4.8	>5			R: irritant		H: positive G: positive		[145]
Camphene	>5	2.5	+		R: slightly irritant; H: 4 % not irritant		H: negative		[145, 146]
Bisabolene	>5	>5		0/12 after 7 h	R: slightly irritant; H: 10 % not irritant	not irritant	H: negative		
Bisabolol	>5	>5		0/12 after 7 h	R: not irritant	slightly irritant			[146]
Caryophyllene	>5	>5			R: irritant; H: 4 % not irritant∗∗		H: negative		
Phytol	>10				R: moderately irritant	not irritant			[145, 146]

∗ IRT: inhalation risk test (rat, test result dependent on toxicity and volatility; 0/6 after 6 h means that after 6 h exposure in an enriched or saturated atmosphere at room temperature no animals died). ∗∗ In light petroleum jelly.

2) Regioselective oxidation: 3,10-myrcenediol, 1,2-myrcenediol, and 2-hydroxymyrcene-1-carboxylic acid from myrcene
3) Allylic oxidation: verbenol and myrtenic acid from α-pinene, and pinocarveol from β-pinene
4) Homoallylic oxidation: 6-*exo*-hydroxycamphene and 10-hydroxytricyclene from camphene
5) Hydrolysis of epoxides: camphenediol from camphene

With *p*-cymene, three different oxidation products were detected in the first step. The alcohols 2-(4-methylphenyl)-1-propanol and 2-(4-methyl-phenyl)-2-propanol are formed by the ω- and (ω-1)-oxidation of the isopropyl group, and 4-isopropylbenzyl alcohol by oxidation of the methyl group. The alcohols formed in these different oxidations are either excreted from the organism as their glucuronides or further oxidized [8, 144].

Most of the terpenes described in Table 1 are licensed as food additives by the international bodies responsible (Table 2). This licensing clearly demonstrates the low toxicity of the terpenes concerned.

Table 2. Terpenes licensed as food additives

Substance	Licensing authority
Myrcene	FDA (GRAS classification), European Council
Ocimene	FDA, European Council
α-Terpinene	FDA
Terpinolene	FDA, European Council
p-Cymene	FDA
1,8-Cineole	FDA, European Council
Camphene	FDA, European Council
Caryophyllene	FDA, European Council

FDA = Food and Drug Administration;
GRAS = generally recognized as safe.

12. References

1. J. C. Simonsen et al.: *The Terpenes,* 2nd ed., **vols. 1 – 3,** Cambridge University Press, New York 1947 – 1951.
2. A. A. Newmann: *Chemistry of Terpenes and Terpenoids,* Academic Press, London 1972. T. K. Devon, A. I Scott: *Handbook of Naturally Occurring Compounds,* **vol. II,** Terpenes, Academic Press, London 1972.
3. *Terpenoids and Steroids (A Spezialist Periodical Report)* **vols 1 – 6,** The Chemical Society, Burlington House, London 1971 – 1976.
4. W. Templeton: *An Introduction to the Chemistry of the Terpenoids and Steroids,* Butterworths, London 1969.
5. W. J. Tayler, A. R. Battersby: *Cyclopentanoid Terpene Derivatives,* Marcel Dekker, New York 1969.
6. O. Aschan: *Naphthenverbindungen, Terpene und Campherarten,* De Gruyter, Berlin 1929.
7. J. ApSimon: *The Total Synthesis of Natural Products,* vol. 2, J. Wiley, New York 1973.
8. O. W. Thiele: *Lipide, Isoprenoide mit Steroiden,* Thieme Verlag, Stuttgart 1979.
9. E. Gildemeister, F. Hoffmann: *Die ätherischen öle,* 4th ed., vols. 1 – 7, Akademie Verlag, Berlin 1960.
10. E. Guenther: *The Essential Oils,* Van Nostrand, Princeton, N. J., 1949.
11. W. Sandermann: *Naturharze Terpentinöl Tallöl,* Springer Verlag, Berlin 1960.
12. O. Isler: *Carotinoide,* Birkhäuser Verlag, Basel 1971.
13. *Beilstein,* **30** (1938) 64 – 76. S. Boström: *Kautschuk-Handbuch,* **vols. 1 – 5,** Berliner Union, Stuttgart 1958 – 1962. L. Batteman: *The Chemistry and Physics of Rubber-Like Substances,* Maclaren, London 1963.
14. T. W. Goodwin: *Aspects of Terpenoid Chemistry and Biochemistry,* Academic Press, London 1971.
15. D. V. Banthopre, B. V. Charlewood, M. J. O. Francis: "The Biosynthesis of Monoterpenes," *Chem. Rev.* **72** (1972) no. 2, 115.
16. A. Cordell: "Biosynthesis of Sequisterpenes," *Chem. Rev.* **76** (1976) no. 4, 425.
17. S. G. Cantwell, E. P. Lau, D. S. Watt, R. R. Fall, *J. Bacteriol.* **135** (1978) 324.
18. H. Siegel, W. Himmele, *Angew. Chem.* **92** (1980) 185.
19. W. Franke: *Nutzpflanzenkunde,* Thieme Verlag, Stuttgart 1976.
20. C. D. Ender, *Nav. Stores Rev.* **66** (1956) 10. W. H. Wiles, *Rubber Plast. Age* **40** (1959) 249. K. Blaetuer, *Dtsch. Farben Z.* **17** (1963) 34. H. P. Preuss, *Met. Finish* **62** (1964) no. 8, 63.
21. *Chem. Age (London)* **63** (1950) 269. *Prakt. Chem.* **10** (1959) 383. R. Werner, *Prakt. Chem.* **11** (1960) 331.
22. J. Schreiber: *Chemie und chemische Technologie der Kunstharze,* Wissenschaftl. Verlagsgesellschaft, Stuttgart 1943, p. 675. E. R. Littmann, *Ind. Eng Chem.* **28** (1936) 1150.

23. L. Désalbres, FR 10 23 339, 1950.
24. J. M. Mellor, S. Munavalli, *Q. Rev. Chem. Soc.* **18** (1964) 270.
25. F. Sorm: "Sequisterpenes with Ten-Membered Carbon Rings, A. Review," *J. Agric. Food Chem.* **19,** (1971) 1081.
26. Van Amerigen-Haebler, Inc., US 2882 323, 1957 (R. Weiss).
27. Glidden Co., US 3031442, 1958 (R. L. Webb).
28. T. Sasaki, S. Eguchi, T. Jshi, *J. Org. Chem.* **34** (1969) 3749.
29. L. A. Goldblatt, S. Palkins, *J. Am. Chem. Soc.* **63** (1941) 3517. R. L. Burwell, Jr., *J. Am. Chem. Soc.* **73** (1951) 4421. E. T. Theimer, B. M. Mitzner, *Can. J. Chem.* **42** (1964) no. 41, 959. E. L. Patton, *Am. Perfum. Essent. Oil Rev.* **56** (1950) 118.
30. *Beilstein,* 1 H 260, 1 E III, 1054.
31. J. E. Hawkins, H. G. Hunt, *J. Am. Chem. Soc.* **73** (1951) 5379. J. E. Hawkins, W. A. Burris, *J. Org. Chem.* **24** (1959) 150.
32. R. L. Burwell, *J. Am. Chem. Soc.* **73** (1951) 4461.
33. M. C. J. Enklaar, *Recl. Trav. Chim. Pays-Bas Belg.* **26** (1907) 157, 167.
34. M. C. J. Enklaar, *Recl. Trav. Chim. Pays-Bas Belg.* **26** (1907) 174; *Chem. Zentrabl.* 1907 II, 56; *Ber. Dtsch. Chem. Ges.* **41** (1908) 2084.
35. K. Alder, A. Dreike, H. Erpenbach, U. Wicker, *Justus Liebigs Ann. Chem.* **609** (1957) 1.
36. T. Saskaki, S. Eguchi, H. Yamada, *Tetrahedron Lett.* 1971, 99.
37. H. Pines, N. E. Hoffmann, V. N. Ipatieff, *J. Am. Chem. Soc.* **76** (1954) 4412.
38. G. Ohloff, *Chem. Ber.* **90** (1957) 1554.
39. V. N. Ipatieff, H. Pines, E. E. Meisinger, *J. Am. Chem. Soc.* **71** (1949) 2934.
40. V. N. Ipatieff, H. Pines, E. E. Meisinger, *J. Am. Chem. Soc.* **71** (1949) 2934.
41. H. A. Smith, J. F. Fuzek, H. T. Meriwether, *J. Am. Chem. Soc.* **71** (1949) 3765, 3766.
42. USA Secr. of Agric., US 2 775 578, 1955.
43. W. Sandermann: *Naturharze: Terpentinöl, Tallöl* Springer Verlag, Berlin 1960, pp. 269 – 310.
44. BASF, DE 960 988, 1954.
45. M. Saito, *Kogyo Kagaku Zasshi* **61** (1958) 326 –330.
46. J. N. Borgein, D. A. Lister, E. J. Lorand. J. E. Reese, *J. Am. Chem. Soc.* **72** (1950) 4591 – 4596.
47. Haarmann u. Reimer, DE 411901, 1923 (R. Moelleken).
48. Monsanto Co., US 4 097 555, 1976.
49. Mitsubishi Chem. Ind., JP 53 121 891, 1977; *Chem. Abstr.* **90** (1979) 88066 t.
50. Standard Oil Co., US 4080 592, 1977.
51. R. G. Buttery et al., *J. Agric Food Chem.* **22** (1974) 773.
52. Zellstoffabrik Waldhof, DE 727 475, 1940.
53. BASF, DE 961 979, 1953.
54. Goodyear Co., EP 77289, 1981 (J. Kuczkowski, L. Wideman).
55. Hercules Powder Co., US 2400012, 1942.
56. BASF, US 3 555 103, 1967.
57. Mobil Oil Corp., EP 12514, 1978 (G. T. Burress).
58. Sumitomo Chemical KK, DE 2910493, 1978.
59. G. Ohloff: *Riechstoffe und Geruchssinn,* Springer Verlag, Berlin 1990.
60. Rheinische Kampherfabrik. DE 499 732, 1928.
61. Newport Ind. Inc., US 2090620, 1936.
62. Takasago Perfumery Ind. Co. JP 11 72, 1950.
63. Union Camp Corp., EP 270023 (P. W. D. Mitchell, D. E. Sasser).
64. B. S. Rao, J. L. Simonsens, *J. Chem. Soc.* **127** (1925) 2494 – 2499.
65. N. T. Mirov, *J. Am. Pharm. Assoc. Sci. Ed.* **40** (1951) 410 – 413.
66. G. Widmarks, S. G. Blohm, *Acta Chem. Scand. (1947 – 1973)* **11** (1957) 392 – 394.
67. B. A. Arbusow, B. M. Michailow, *J. Prakt. Chem.* **127** (1930) 1 – 15.
68. J. Verghese, W. K. Sondhi, B. Bushan, M. L. Joshi, *Curr. Sci.* **17** (1948) 359.
69. B. S. Rao, J. L. Simonsens, *J. Chem. Soc.* **127** (1935) 2494.
70. S. Arctander: *Perfume and Flavor Chemicals,* Det Hoffensbergske, Etablissement, Kopenhagen 1969.
71. Shell Oil Co., US 4156692, 1977.
72. Shell Int. Res., EP 2850, 1977.
73. D. V. Banthorpe, D. Whittaker, *Chem. Rev.* **66** (1966) 643.
74. Monsanto US 2097 744, 1934.
75. SCM Corp., US 4254291, 1981.
76. J. Tanaka, T. Katagiri, K. Ozawa, *Bull. Chem. Soc. Jpn.* **44** (1971) 130 – 132.
77. G. S. Fisher, J. S. Stinson, R. N. Moore, L. A. Goldblatt, *Ind. Eng. Chem.* **47** (1955) 1368 – 1373.
78. G. S. Fisher, L. A. Goldblatt, J. Kniel, A. D. Snyder, *Ind. Eng. Chem.* **43** (1951) 671 – 674.
79. SCM-Corp., US 4018 842, 1977.
80. G. Zweifel, H. C. Brown, *J. Am. Chem. Soc.* **86** (1964) 393 – 397.
81. *Chem. Eng. News* **60** (1982) no. 11, 5.
82. G. A. Schmidt, G. S. Fisher, *J. Am. Chem. Soc.* **76** (1954) 5426 – 5430.
83. Tagasako, JP 61 200999, 1986.

84. G. Ohloff, E. Klelin, *Tetrahedron* **18** (1962) 37.
85. USA Secr. of Agric, US 2735870, 1950.
86. G. S. Fisher, J. S. Stinson, L. A. Goldblatt, *J. Am. Chem. Soc.* **75** (1953) 3675 – 3678.
87. Amer. Cynamid, US 306 1554, 1962.
88. *Beilstein,* **27,** 5 E IV, 277.
89. Studiengesellschaft Kohle mbH, DE 1150974, 1961.
90. G. Ohlof, E. Klein, *Tetrahedron* **18** (1962) 37 – 42.
91. Shell, NL 7900588, 1979.
92. L. A. Shutikova, V. G. Charakew, M. S. Erzhanova, A. K. Alifanova, *Maslo. Zhir. Promst.* 1973, 23 – 25.
93. Cocker Chemical Co., GB 1619649, 1963.
94. G. A. Schmidt, G. S. Fisher, *J. Am. Chem. Soc.* **81** (1959) 445 – 448.
95. Stephan Chemical Co., DE 2 305 363, 1973.
96. Studiengesellschaft Kohle mbH, FR 1328113, 1963.
97. Du Pont, *Ind. Chim. Belge Ser. 2* **11** (1940) 3.
98. W. Hückel, *Ber. Dtsch. Chem. Ges.* **80** (1947) 41 – 47.
99. P. Lipp, *Justus Liebigs Ann. Chem.* **382** (1911) 265 –305.
100. A. Meyer, DE 272562 1911.
101. V. Hilcken, DE 439 695 1924.
102. Schering-Kahlbaum, DE 578 569, 1931.
103. Du Pont, US 2551 795, 1949.
104. E. F. L. J. Anet, *Aust. J. Chem.* **23** (1970) 2101.
105. M. G. Moshonas, P. E. Shaw, *J. Agric. Food Chem.* **28** (1980) 680.
106. O. R. W. Sutherland, R. F. N. Hutchins, *Nature (London)* **239** (1972) 170.
107. G. W. K. Cavill, P. J. Williams. F. B. Withfield, *Tetrahedron Lett.* 1967, 2201.
108. C. Harries, R. Haarmann, *Chem. Ber.* **46** (1913) 1741.
109. G. Brieger, T. J. Nestrick, C. McKenna, *J. Org. Chem.* **34** (1969) 3789.
110. G. Brieger, *J. Org. Chem.* **32** (1967) 3720.
111. L. Ruzicka, E. Capato, *Helv. Chim. Acta* **8** (1925) 259.
112. F. Delay, G. Ohloff, *Helv. Chim. Acta* **62** (1979) 369.
113. G. Brieger, T. J. Nestirck, C. McKenna, *J. Org. Chem.* **34** (1969) 3789.
114. O. P. Vig, S. D. Sharma, P. Kumar, M. L. Sharma, *J. Indian Chem. Soc.* **52** (1975) 614.
115. E. V. Weber, *Justus Liebigs Ann. Chem.* **238** (1887) 101.
116. Degussa, DE-OS 2709033, 1978.
117. M. Holub, V. Heront, F. Sorm, *Cesk. Farm* **4** (1955) 129.
118. F. W. Semmler, E. W. Mayer, *Ber. Dtsch. Chem. Ges.* **45** (1912) 1384 – 1394.
119. Shell, EP 127 911, 1983 (A. J. Mulder).
120. E. Deussen, *Justus Liebigs Ann. Chem.* **388** (1912) 136 – 165.
121. H. N. Rydon, *J. Chem. Soc.* 1939, 537 – 540.
122. Soda Shangyo Co. Ltd, JP-Kokai 76 80847, 1975.
123. BBA, GB 1580184, 1976 (H. R. Ansari, R. Clark, H. R. Wagner).
124. J. L. Simonsen, *J. Chem. Soc.* **117** (1920) 564 – 578.
125. G. R. Dupont, R. Dulou, P. Naffa, G. Ourisson, *Bull. Soc. Chim. Fr.* 1954, 1075, 1078.
126. Tagasako Co., JP 04 360848, 1992.
127. Tosho Corp., JP-Kokai 01 268657, 1989.
128. E. M. Burdick, US 3248301, 1964.
129. BASF, DE 28187150, 1978.
130. H. Ianistyn, DE 1467955, 1964.
131. B. Staller-Bourdillon, *Ind. Chim. Belge* **35** (1970) no. 1, 13.
132. Eisai KK, JP-Kokai 88 108, 1976.
133. Nisshin Flour Milling, JP-Kokai 10 235, 1960.
134. E. Jellum, L. Eldjarn, K. Try, *Acta Chem. Scand. (1947 – 1973)* **20** (1966) 2535.
135. M. Tsujmoto, *Chem. Zentralbl.* 1918 I, 638, 1048.
136. J. C. Drummond, H. J. Channon, K. H. Coward, *Biochem. J.* **19/23** (1919/1923) 1055/279.
137. K. Täufel, H. Thaler, H. Schreyegg, *Fettchem. Umsch.* **43** (1936) 26.
138. M. Hamaya, JP-Kokai 00 3158, 1977.
139. D. W. Dicker, M. L. Whiting, *J. Chem. Soc.* 1958, 1994.
140. J. M. Heilbron, T. P. Hildrich, E. D. Kamm, *J. Chem. Soc.* 1926, 3135.
141. H. Janistyk: *Handbuch der Kosmetik und Riechstoffe,* 2nd ed., **vol. 1,** Hüthig Verlag, Heidelberg 1969, p. 753.
142. E. S. Lower, *Spec. Chem.* **1** (1981) no. 1, 34.
143. E. L. J. Opdyke (ed): *Monographs on Fragrance Raw Materials,* Pergamon Press, Oxford 1979.
144. T. Ishida et al.: "Biotransformations of Terpenoids in Mammals," *Chem. Abstr.* **93** (1980) 8986003.
145.
146. NIOSH: *Registry of Toxic Effects of Chemical Substances 1979,* **II,** Washington, D. C., 1979, p. 33.

Terphenyls → Hydrocarbons
Tetrachloroethanes → Chlorinated Hydrocarbons
Tetrachloroethylene → Chlorinated Hydrocarbons
Tetracycline → Antibiotics
Tetraethyllead → Lead Compounds
Tetrafluoroethylene Copolymers → Fluoropolymers, Organic

Tetrahydrofuran

HERBERT MÜLLER, BASF Aktiengesellschaft , Ludwigshafen, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 2
4.1. **Acetylene/Formaldehyde** 2
4.2. **Butadiene Acetoxylation** 2
4.3. **Propylene Oxide Process** 3
4.4. **Maleic Anhydride Hydrogenation** . . 3
4.5. ***n*-Butane – Maleic Anhydride Process** 3
4.6. **Pentosan/Furfural Processes** 3
5. **Specifications and Analysis** 3
6. **Environmental Protection and Recovery** 4
7. **Storage and Transport** 4
8. **Uses** 4
9. **Economic Aspects** 4
10. **Toxicology and Industrial Hygiene** . . 5
11. **References** 5

1. Introduction

Tetrahydrofuran (THF) [*109-99-9*] tetramethylene oxide, oxolane, is a five-membered cyclic ether with wide application in the chemical industry.

The largest proportion of the 1.4×10^5 t of THF produced worldwide in 1992 was used as a monomer in the manufacture of poly(tetramethylene oxide) (PTMO) also known as poly(tetramethylene ether glycol) (PTMEG) and polytetrahydrofuran (PTHF), important in the production of thermoplastic polyurethanes, elastic fibers, molded elastomers, and copolyesters or copolyamides. A smaller proportion finds use as a solvent for coatings, adhesives, and special varnishes based, for example, on poly(vinyl chloride) (PVC) or polyurethanes; as an extracting agent; and as the preferred medium for organometallic syntheses.

2. Physical Properties

The pure distilled tetrahydrofuran sold commercially is a colorless, clear, volatile, polar liquid with a characteristic acetone-like odor and a minimum purity of 99.9 wt %. THF is miscible in all proportions with water, alcohols, ethers, and many common solvents. Important physical properties include the following [1,2]:

M_r	72.1
bp	66 °C
mp	−108.5 °C
Critical temperature	267 °C
Critical pressure	5.19 MPa
Critical density	0.322 g/cm^3
d_4^{20}	0.886
n_D^{20}	1.4073
Specific heat	1.765 J g^{-1} K^{-1}
Heat of vaporization (66 °C/101.3 kPa)	435 J/kg
Heat of combustion	−35 141 kJ/kg
Explosive limits in air (25 °C)	
lower	1.5 vol %
upper	12.0 vol %
Flash point (Abel – Pensky)	−22 °C
Dipole moment	5.84×10^{-30} C m
Dielectric constant (20 °C)	7.6

3. Chemical Properties

THF is a valuable starting material for a wide range of important reactions. For example, cationic polymerization accompanied by ring opening leads to high molecular mass bifunctional glycol ethers with various chain lengths [3,4]. These have achieved great economic significance in the preparation of important plastics. Ring-cleavage reactions also constitute the basis for dehydration to butadiene, oxidation to succinic acid, and carboxylation to adipic acid or γ-valerolactone [5,6].

THF is the preferred reaction medium for carrying out Grignard reactions or reductions with, for example, $LiAlH_4$. It is also suitable for use as a ligand in coordination complexes, such as those sometimes employed in stereospecific polymerizations [7,8].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_221

THF reacts readily with oxygen (on contact with air, for example), the principal product being an unstable hydroperoxide [9]. Hydroquinone or 2,6-di-*tert*-butyl-*p*- cresol (BHT) (e.g., 250 mg/kg) can be added to retard peroxide formation. Distillation of peroxide-containing THF increases the peroxide concentration, resulting in a serious risk of explosion, even on a laboratory scale.

4. Production

4.1. Acetylene/Formaldehyde

A process developed by Reppe in the 1930s was for many years the preferred synthetic route to 1,4-butanediol and THF, and it is still the most common approach in Europe and the United States [10, 11]. The Reppe process involves a reaction between acetylene and formaldehyde to give 2-butyne-1,4-diol, with subsequent hydrogenation to 1,4-butanediol (→ Butanediols, Butenediol, and Butynediol, Chap. 1.3.). The saturated diol is very readily cyclized to THF with elimination of water by acid catalysis above 100 °C. Suitable catalysts include inorganic acids, acidic aluminum silicates, and earth or rare-earth oxides.

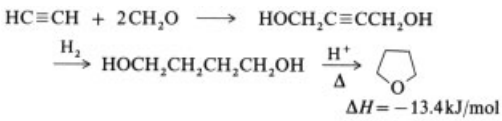

Quantitative conversion and a yield of almost 100 % can be achieved in an atmospheric-pressure, continuous process with aluminum oxide as the catalyst provided the THF is continuously distilled out of the reaction mixture as it forms and pure 1,4-butanediol is fed into the reactor at a rate equal to its rate of consumption. The amount of conversion per unit amount of catalyst is very high. Crude butanediol can be converted into THF very selectively and with little added energy by a medium-pressure process described by BASF [12]. The THF – steam mixture obtained from cyclization is first distilled in a rectification column to give the corresponding azeotrope (5.3 wt % water, *bp* 62.3 °C). Treatment with alkali hydroxide [13] and subsequent distillation provides anhydrous THF. The azeotrope problem can also be circumvented industrially by distillation under pressure.

4.2. Butadiene Acetoxylation

The Mitsubishi-Kasei Corporation in Japan produces 1,4-butanediol and THF in parallel starting from butadiene [14]. The procedure can be described by the following set of equations:

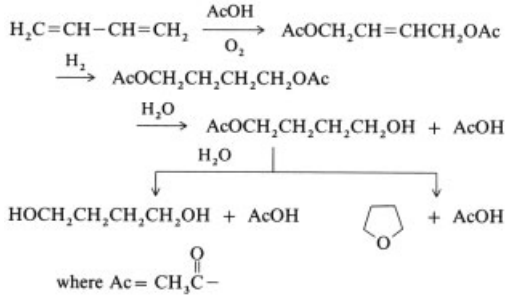

Butadiene is oxidized at 3 MPa/80 °C over a palladium – tellurium catalyst with acetic acid and a nitrogen – oxygen mixture to give 1,4-diacetoxy-2-butene. The olefin is hydrogenated to 1,4-diacetoxybutane [15], which can be hydrolyzed to butanediol or THF.

4.3. Propylene Oxide Process

Arco produces 1,4-butanediol by the following route:

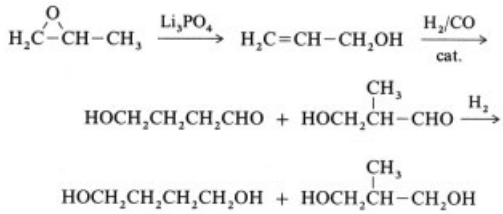

Propylene oxide is isomerized to allyl alcohol by conventional means at 250 – 300 °C/1 MPa over a trilithium orthophosphate (Li_3PO_4) catalyst. The allyl alcohol is then hydroformylated to 4-hydroxybutyraldehyde and the byproduct 3-hydroxy-2-methylpropionaldehyde by a process developed by Kuraray [16]. Subsequent hydrogenation of the aldehydes gives 1,4-butanediol as the main product together with 2-methyl-1,3-propanediol.

4.4. Maleic Anhydride Hydrogenation

Because of its structure, maleic anhydride is an attractive precursor for the preparation of butanediol, THF, and γ-butyrolactone:

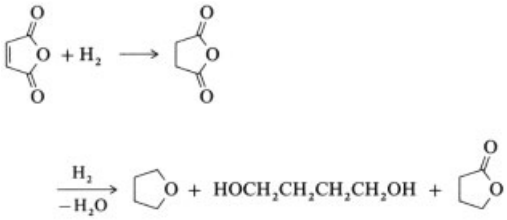

This route was originally developed and adapted to industrial scale by Mitsubishi-Kasei [17], but Mitsubishi itself has since abandoned the process and replaced it by the newer butadiene acetoxylation (Section 4.2).

Davy – McKee Ltd., BP, and Sohio have introduced a similar process for the production of butanediol and THF. Shinwha Petrochemical in South Korea has commissioned a plant based on the McKee process producing 20 000 t/a of butanediol. The first step is a vapor-phase hydrogenation of ethyl maleate [18], and the ratio of THF to butanediol in the product mix can be adjusted by altering the reaction conditions.

Increasing attention is being focused on maleic anhydride as a starting material for THF, and it appears that this approach will become more important than the Reppe process in the medium term [19].

4.5. *n*-Butane – Maleic Anhydride Process

Du Pont has developed a new two-step process for producing THF from *n*-butane. The process involves oxidation to maleic anhydride with a transported abrasion-resistant oxidation catalyst, which guarantees a high yield (70 – 75 %), with subsequent hydrogenation of aqueous maleic acid solution over a special rhenium-doped palladium catalyst [20]. This method also leads to either butanediol or THF as desired; it has been perfected on a pilot scale and is due to be implemented in 1994 in the Spanish province of Asturia with an annual production of 45 000 t. The process can be summarized as follows:

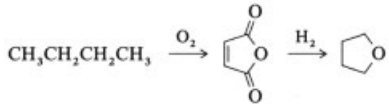

4.6. Pentosan/Furfural Processes

Because it takes advantage of a renewable raw material, THF production from furfural has attracted interest (→ Biomass Chemicals, Chap. 4.6.). Because of their widespread distribution in agricultural waste products, pentosans are especially promising as starting materials (→ Furan and Derivatives, Chap. 3.) [21]. The production of THF from furfural involves catalytic decarbonylation to furan and hydrogenation of the latter to THF [22].

5. Specifications and Analysis

Table 1 illustrates typical quality specifications quoted by a leading European producer of THF. Gas chromatographic (GC) analysis can be conducted with a 30-m WCOT (fused silica) column coated with DB-1 (FID detector, He as carrier gas) using *n*-octane as an internal standard [2].

Table 1. Quality specifications for commercial tetrahydrofuran

Characteristics	Value	Test method
THF content	min. 99.9 %	GC
Water content	max. 0.03 %	DIN 51 777, section 21
Hazen color index	max. 10	DIN/ISO 6271

6. Environmental Protection and Recovery

Based on the considerations presented in the section on toxicology (Chapter 10), it can be concluded that with proper handling THF constitutes no significant risk with respect to the environment.

THF is also quite easily recovered. Thus, THF vapors from exhaust gases can be adsorbed on activated charcoal and subsequently desorbed with steam. The efficiency of the process is said to be 98 – 99 %. In the case of high concentrations of THF in air, absorption in water is also a possibility. The THF – water azeotrope (5.3 wt % water, *bp* 62.3 °C) obtained on distillation can be converted into anhydrous THF

(< 0.05 % water) by distillation at reduced or (preferably) elevated pressure [23].

Extraction of THF from an aqueous mixture with a solvent immiscible with water (e.g., pentane), followed by phase separation and fractional distillation of the organic phase, is another drying technique practiced on an industrial scale. This results in THF containing < 0.1 % water [24]. For the separation of THF from water and such organic solvents as toluene by fractional distillation, see [24]. Various methods have been developed for drying on a smaller scale [25, 26]. Regardless which method is selected, great care is required because of the danger of peroxide formation [27].

7. Storage and Transport

The handling of THF entails only the usual precautionary measures applicable to other volatile organic compounds, and THF can be stored in the usual steel containers. For transportation through piping, welded pipes are preferable; special attention should be directed toward the seals in pumps and valves.

Like other ethers (e.g., diethyl or diisopropyl ether) THF forms peroxides in air if it has not been stabilized. At a peroxide concentration > 0.1 % THF is subject to explosive decomposition on distillation. It is stabilized with 2,6-di-*tert*-butyl-*p*-cresol (BHT). In the absence of a stabilizer THF must be stored in an inert gas atmosphere.

A safety data sheet in accordance with DIN 52 900 is available for THF.

For THF the following transport regulations are in force:

GGVE/RID	Class 3, no. 3 b RN 301
GGVS/ADR	Class 3, no. 3 b RN 2301
ADNR	Class 3, no. 3 b RN 6301
IMDG Code	Class 3.1
UN no.	2056
S phrases	16 – 29 – 33
R phrases	11 – 19 – 36/37

8. Uses

The most important area of application for THF from the standpoint of quantity and turnover is polymerization with simultaneous ring opening to give poly(tetramethylene oxide) (→ Polyoxyalkylenes). Available in various molecular masses, this is a key component in the production of important elastic construction materials, thermoplastics and molded elastomers based on polyurethanes, polyesters and polyamides, elastic Spandex fibers, and polyurethane coatings. Trade names for commercial PTMO include Terathane (DuPont), Polytetrahydrofuran (BASF), and Polymeg (Quaker Oats).

THF is also a versatile solvent for natural and synthetic resins as well as PVC. It is used extensively in the varnish and film industries and as a cosolvent for printing inks and adhesives.

9. Economic Aspects

The projected annual growth rate for tetrahydrofuran is 4 – 7 % based on anticipated increases in the consumption of polytetrahydrofuran. Major producers of THF are listed in Table 2, and regional production capacity and consumption estimates are provided in Table 3.

Table 2. Major THF producers in 1992

Company	Region	Process	Capacity, 10^3 t/a
Du Pont	USA	Reppe	70
BASF	USA, Europe, Japan	Reppe	40
ARCO	USA	PO/AA*	8
GAF	USA, Europe	Reppe	18
Quaker Oats	USA	furfural	10
Mitsubishi-Kasei	Japan	butadiene	16
Toso Corp.	Japan	Reppe	3
Shinwha	Korea	butane	10
	CIS/China	furfural/Reppe	10

* PO = propylene oxide; AA = allyl alcohol.

10. Toxicology and Industrial Hygiene

THF shows weak narcotic activity. According to very intensive investigations conducted at the Industrial Hygiene and Pharmacology Institute at BASF it is one of the least toxic solvents. This is in sharp contrast to previous reports, which were based on impure samples. Its true toxicity is evidently not greater than that of acetone.

Table 3. Worldwide production capacity and consumption data by geographical area (10^3 t/a)

	USA 1990/1995	Western Europe 1990/1995	Southeast Asia 1990/1995	China, CIS 1990
Capacity	95/133	30/50	20/40	5 – 10
Consumption	63/80	50/70	20/30	5 – 10

In the investigations at BASF, THF exposure levels of 1000 ppm for 6 h a day, 5 days a week, over a period of 1 year were tolerated by cats, rabbits, and rats with no associated symptoms of poisoning. Exposure to 3000 ppm daily for 8 h over a course of almost 2 years produced as few symptoms of poisoning in rats as did inhalation over the same length of time of 3000 ppm acetone. Concentrations of 1000 ppm or 3000 ppm are completely intolerable for humans because of severe mucous membrane irritation, and would not therefore be experienced in practice. Even higher concentrations of THF vapor produce a narcotic effect similar to that of ether or acetone. Like alcohol, THF is rapidly metabolized in the living organism.

As a consequence of its universal solvent power, THF dissolves the upper keratin-containing layers of skin and mucous membranes, penetrating very rapidly into deeper parts of the skin. Thus it gives rise to very severe skin and mucous membrane irritation, but it does not display a sensitizing effect. This irritant character is apparently reinforced by peroxides, since THF that contains peroxides due to prolonged storage is a more severe skin irritant. THF that is allowed to penetrate a wound or get under the fingernails causes very unpleasant pains.

Like methanol, THF is resorbed very rapidly by the skin. For this reason, prolonged wetting of the skin with THF must be strictly avoided, and hand protection should be provided against repeated wetting.

The maximum workplace concentration for THF has been fixed at 200 mL/m^3 (ppm). This limiting concentration is a function only of local mucous membrane irritation caused by THF vapors; much higher concentrations are required to produce a narcotic effect in humans [28].

11. References

1. A. P. Kudschadker et al.: *Key Chemicals Data Books Furan, Dihydrofuran, Tetrahydrofuran,* Thermodynamics Res. Center, Texas A & M University, College Station, Texas 1978.
2. BASF, technical report "Tetrahydrofuran," Ludwigshafen 1991.
3. H. Meerwein et al., *Angew. Chem.* **72** (1960) 927.
4. P. Dreyfuss: "Polytetrahydrofuran," in *Polymer Monographs,* vol. 1.8, Gordon and Breach Sci. Publ., New York 1982.
5. W. Reppe: *Neue Entwicklungen auf dem Gebiet der Chemie des Acetylens und Kohlenmonoxyds,* Springer Verlag, Heidelberg 1949.
6. "Furfural" 660.5020 in: *Chemical Economics Handbook,* Stanford Res. Inst., Menlo Park, Calif. 1968.
7. A. Wong et al., *J. Am. Chem. Soc.* **102** (1980) no. 13, 4529.
8. R. G. Gastinger et al., *J. Am. Chem. Soc.* **102** (1980) no. 15, 4959.
9. R. Criegee et al., *Angew. Chem.* **62** (1950) 120.
10. J. W. Copenhaver, M. H. Bigelow: *Acetylene and Carbonmonoxide Chemistry,* Reinhold Publ. Co., New York 1949; University Microfilms, Ann Arbor, Mich. 1949.
11. BASF, DE 1 043 342, 1957 (J. Schneiders); DE 2 930 144, 1979 (H. Müller et al.).
12. BASF, EP 1 536 680, 1979 (H. Müller, Ch. Palm).
13. IG Farben, DE 713 565, 1939 (W. Reppe, H.-G. Trieschmann).
14. Mitsubishi Chem., US 3 922 300, 1975 (T. Onoda et al.). T. Onoda: "1,4-Butylenglycol and Tetrahydrofuran Production from Butadien," *AIChE Summer National Meeting,* Minneapolis, Minn., Aug. 16 – 17, 1987. Kurary, US 3 872 163, 1975 (T. Shimizu, T. Yasui).
15. Mitsubishi Chem., US 4 010 197, 1977 (J. Toriya, K. Shirago).
16. Kurary, US 4 567 305, 1986 (T. Shimizu, T. Yasui). Kurary & Daicel, EP-A 129 802, 1984 (M. Milsuo, M. Shini).
17. J. Kanetaka et al., *Ind. Eng. Chem.* **62** (1970) 24. Mitsubishi Chem., GB 1 230 276, 1971. GB 1 344 557, 1974.
18. Davy McKee Ltd., EP 143 634, 1984 (M. Sherif, K. Turner). Davy McKee Ltd., GB 2 207 914, 1987 (M. Sherif, K. Turner).

19. BASF, DE 3 423 447, 1986 (R. Schnabel, H. Weitz); DE 3 726 509, 1989 (R. Fischer, W. Harder, K. Malsch). Huels, 3 539 151, 1986 (R. Nehring, W. Otte). UCB, EP 339 012, 1989 (J. L. Dalons, J. Markusel). General Electric, US 4 652 685, 1987 (W. James, E. Norman). General Aniline, US 4 797 382, 1989 (D. Waldo, D. Taylor). Standard Oil, US 4 301 077, 1980 (F. A. Pesa, A. M. Graham); US 4 810 807, 1988 (J. Budge, E. Pederson); US 4 827 001, 1988 (Th. Attig, A. Graham). Tonen Corp., EP 373 946, 1989 (S. Sadakatsu, I. Hiroyuki); EP 431 923, 1991 (S. Sadukatsu, I. Tatsumi).
20. Du Pont, EP-A 189 261, 1986 (R. M. Contractor); US 4 442 226, 1982 (T. A. Bither); US 4 371 702, 1982 (T. A. Bither); EP 147 219, 1989 (A. Marby, W. Pichard).
21. C. Godawa et al., *Inf. Chim.* **260** (1985) 171 – 177.
22. N. N. Machalaba et al., *Gidroliz. Lesokhim. Prom.-st.* **1988,** no. 5, 19 – 20; *Chem. Abstr.* 109 151883k. Quaker Oats, US 3 233 714,1963. Du Pont, US 2 846 449, 1956 (W. H. Banford, H. M. Manes); US 2 374 149, 1943 (G. M. Whitman). Quaker Oats, US 3 021 342, 1958 (H. W. Hemker); GB 855 255, 1959.
23. Du Pont, CA 546 591, 1957.
24. Du Pont, US 2 790 813, 1953 (P. F. Bente).
25. Du Pont, US 3 074 968, 1960 (J. W. Meinherz).
26. Du Pont, bulletin A-25077, recovery of THF.
27. B. G. White et al., *Org. Synth.* **46** (1966) 105.
28. BG Chemie, Merkblatt M038,588, Tetrahydrofuran, Jedermann Verlag Dr. Otto Pfeffer, Heidelberg 1988. Kühn-Birett: *Merkblätter Gefährlicher Arbeitsstoffe,* 26. Erg.Lfg., 3/85-T08-1, Verlag Moderne Industrie, München 1977. G. Hommel: *Handbuch der gefährlichen Güter,* 5th ed., Merkblatt 192, Springer Verlag, Berlin 1993.*Ullmann,* 4th ed., **12,** 21.

Tetralin → Naphthalene and Hydronaphthalenes
Tetramethyllead → Lead Compounds

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Textile Auxiliaries

Standard Article

Klaus Fischer[1], Kurt Marquardt[2], Kaspar Schlüter[3], Karlheinz Gebert[4], Eva-Marie Borschel[5], Sigismund Heimann[6], Erich Kromm[7], Volker Giesen[8], Reinhard Schneider[9], Rosser Lee Wayland Jr.[10]

[1]Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany
[2]Frankfurt, Federal Republic of Germany
[3]Henkel KGaA, Düsseldorf, Federal Republic of Germany
[4]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[5]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[6]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[7]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[8]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[9]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[10]Hickson DanChem Corporation, Danville, Virginia, United States

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a26_227
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Definitions
1.2. Historical Aspects
1.3. Economic Aspects
1.4. Environmental Aspects
2. Auxiliaries for Fiber Production and Processing
2.1. Introduction
2.2. Economic Aspects
2.3. Use of Fiber Preparations
2.3.1. Use of Fiber Preparations in Primary Spinning
2.3.2. Use of Antistatic or Lubricating Agents in Secondary Spinning
2.3.3. Use of Antistatic and Lubricating Agents in Dyeing
2.4. Physical Properties
2.5. Chemical Properties
2.6. Fiber Preparations for Special Applications
2.7. Additives for Spinning Solutions, Spinning Baths, and Spinning Melts
2.8. Environmental Aspects
2.9. Commercial Products
3. Sizing Agents
3.1. Introduction
3.2. Sizing Methods
3.3. Sizing Agents Based on Natural Raw Materials
3.3.1. Starch and Starch Derivatives
3.3.2. Sodium Carboxymethyl Cellulose

3.3.3. Sizing Agents Based on Other Natural Raw Materials
3.4. Synthetic Sizing Agents
3.4.1. Poly(Vinyl Alcohols)
3.4.2. Poly(meth)acrylates
3.4.3. Polyester Condensates
3.4.4. Other Synthetic Sizing Agents
3.5. Sizing Additives
3.6. Commercial Products
4. Pretreatment Auxiliaries
4.1. Processes and Auxiliaries for the Pretreatment of Textile Materials
4.2. Preparation of Textile Material made of Cellulose Fibers
4.2.1. Pretreatment Processes
4.2.2. Pretreatment Part Processes
4.2.2.1. Desizing
4.2.2.2. Alkaline Extraction
4.2.2.3. Bleaching with Hydrogen Peroxide
4.2.2.4. Bleaching with Sodium Hypochlorite
4.2.2.5. Bleaching with Sodium Chlorite
4.2.2.6. Mercerization and Causticizing
4.3. Preparation of Woolen Textiles
4.4. Ecological Aspects of Fabric Pretreatment
5. Textile Dyeing Auxiliaries
5.1. Dye Solubilizing and Hydrotropic Agents
5.2. Dye Protecting Agents and Boildown Protecting Agents
5.3. Wetting and Deaerating Agents
5.4. Dispersants and Protective Colloids
5.4.1. Dispersants in Dye Production
5.4.2. Dispersants for Dyeing (Batch and Continuous Processes)
5.4.3. Testing
5.5. Complexing Agents
5.6. Levelling Agents
5.6.1. Levelling Agents for Dyeing Cellulose Fibers
5.6.2. Levelling Agents for Wool Dyeing
5.6.3. Levelling Agents for Dyeing Polyamide Fibers
5.6.4. Levelling Agents for Dyeing Polyester Fibers
5.6.5. Levelling Agents for Dyeing Polyacrylonitrile Fibers
5.6.6. Levelling Agents for Dyeing Other Synthetic Fibers and Fiber Blends
5.6.7. pH Regulators (Acid- and Alkali-Releasing Agents)
5.6.8. Testing
5.7. Dyeing Accelerants (Carriers)
5.8. Anticreasing Agents
5.9. Auxiliaries for Continuous Dyeing
5.9.1. Antimigration Agents
5.9.2. Antifrosting Agents
5.9.3. Products to Increase Liquor Uptake
5.10. Aftertreatment Agents to Improve Fastness
5.11. Reducing Agents
5.12. Oxidizing Agents
5.13. Stripping Agents
5.14. Resist Agents for Dyeing and Printing
6. Textile Printing Auxiliaries
6.1. Auxiliaries for Pigment Printing
6.1.1. Thickening Agents
6.1.2. Pigment Printing Binders
6.1.3. Fixing Agents
6.1.4. Hand Modifiers
6.1.5. Emulsifiers
6.1.6. Other Additives
6.2. Auxiliaries for Dye Printing

6.2.1. Thickening Agents
6.2.2. Oxidizing Agents
6.2.3. Reducing and Discharging Agents
6.2.4. Auxiliaries for Discharge Printing
6.2.5. Fixation Auxiliaries
6.2.6. Dye Solubilizing and Dispersing Agents
6.2.7. Auxiliaries for Reactive Resist Printing
6.2.8. Auxiliaries for Burn-out Printing
6.2.9. Aftertreatment Agents and Detergents
6.3. Other Auxiliaries
6.4. Textile Printing Adhesives
7. Finishing Agents
7.1. Introduction
7.2. Easy Care Finishes
7.2.1. Cross-Linking Agents
7.2.2. Catalysts
7.2.3. Testing
7.2.4. Hand Builders
7.2.5. Softeners
7.3. Water Repellents
7.3.1. Historical Aspects
7.3.2. Reactive Quaternary Repellents
7.3.3. Organometallic Repellents
7.3.4. Wax-Based Repellents
7.3.5. Resin-Based Repellents
7.3.6. Silicone Repellents
7.3.7. Fluorochemical Repellents
7.3.8. Testing for Water Repellency
7.4. Soil-Repellent and Soil-Release Agents
7.4.1. Historical Aspects
7.4.2. Factors Influencing Soil Release; Testing
7.4.3. Soil-Release Finishes
7.5. Flame Retardants
7.5.1. Flammability of Fibers
7.5.2. Nondurable Flame Retardants
7.5.3. Semidurable Flame Retardants
7.5.4. Durable Flame Retardants
7.5.5. Flammability Testing
7.6. Wool Finishing
7.6.1. Setting
7.6.2. Scouring
7.6.3. Milling
7.6.4. Carbonizing
7.6.5. Shrink-Resistant Finishes
7.6.6. Mothproofing
8. Auxiliaries for Technical Textiles
8.1. Technical Textiles-an Overview
8.2. Auxiliaries
8.2.1. Polymers
8.2.2. Cross-Linking Agents
8.2.3. Nonslip Agents
8.2.4. Water Repellents
8.2.5. Oleophobic Agents
8.2.6. Flame Retardants
8.2.7. Antistatic Agents
8.2.8. Plasticizers
8.2.9. Coupling Agents
8.2.10. Thickeners
8.2.11. Surfactants

8.2.12. Antifoams
8.2.13. Antimicrobial Agents
8.2.14. Light Stabilizers
8.2.15. Fillers
8.2.16. Pigments
8.2.17. Producers and Trade Names

[Top of Page]

1. Introduction

Klaus Fischer

This article goes beyond the account of classical textile finishing given in the 4th edition of Ullmann's Encyclopedia, and includes nonwoven textiles, carpets, and bonded composites, which are textiles in the wider sense. Technical textiles have become increasingly important in recent decades. A chapter (Chap. Auxiliaries for Technical Textiles) has therefore been devoted to the auxiliaries for these materials.

1.1. Definitions

The TEGEWA nomenclature is used for the definition and classification of the various textile auxiliaries and finishing agents [4]. It is also the basis of the comprehensive former tpi catalogue of products [5], and is fairly well adhered to in other international product lists. This nomenclature differentiates between textile auxiliaries and textile chemicals such as acids, alkalis, salts, and solvents used in the textile industry, usually in conjunction with auxiliaries. The other chapters of this article also follow the TEGEWA nomenclature.

When textile auxiliaries and chemicals are used, they become attached to the textile. After they have carried out their auxiliary function, they may be removed at some stage of textile manufacture, or they may remain on the textile, so that the finished product then consists of the fiber, the textile structure, and the auxiliaries, which then function as finishing agents. Textile auxiliaries may be qualitatively classified as follows:

1. Nonpermanent: auxiliaries that are almost completely removed, e.g., by water, washing, or dry cleaning
2. Semipermanent: auxiliaries that are gradually removed from the textile
3. Permanent: auxiliaries that remain on the textile both in use and after treatments such as washing or dry cleaning

Generally, the only distinction made is between “permanent” and “nonpermanent.”

Environmental regulations make a further distinction, between substances and preparations [6]. According to this, substances are chemical elements or compounds of natural or synthetic origin used as auxiliaries necessary to satisfy market requirements. Preparations consist of two or more substances, and may be mixtures or solutions. Most textile auxiliaries are preparations.

The auxiliary and/or finishing functions of textile auxiliaries vary widely, and depend on the intended use of the textile. Textiles can be classified into three groups based on the demands placed on the finished textile, independent of the type of textile (i.e., whether woven, knitted, nonwoven, carpet, or bonded).

1. Classical textiles for clothing, decoration, and home
2. Technical textiles, mainly for industrial use
3. Textiles used in hygiene, care, and medicine

Textile auxiliaries include not only products that are applied to fibers after they have been made into textiles, but also those that are incorporated during fiber manufacture, e.g., optical brighteners in the spinning mass, modifiers for high-strength viscose, delustering agents, and flame-proofing agents.

1.2. Historical Aspects

Even in antiquity, auxiliaries were undoubtedly used in the production of textiles from wool, cotton, silk, and linen to facilitate working with the fiber and to give the finished product the desired appearance. This included soap for washing and fulling, oils for improved smoothness in spinning and weaving, starches to stiffen the textile, and inorganic mordants for fixing colorants. With increasing industrialization, requirements for the processibility of the raw materials and the appearance and wearing properties of the textile product became stricter. Thus, soap-washed wool had to be treated with lubricants prior to spinning, warps had to be sized, and the material had to be treated with finishing agents to give the final commercial product.

The development of modern textile auxiliaries dates back to the turn of the century and is closely associated with the development of surfactants. This was initiated by MERCER and RUNGE, who ca. 1830 observed that oils and fats can be made water soluble by treatment with sulfuric acid. In 1875 KÖCHLIN, STORCK, and WUTH, and, independently, CRUM sulfated castor oil and obtained the important Turkey red oils. By increasing the degree of sulfation, STOCKHAUSEN (1896) obtained the monopole brilliant oils, which were used as auxiliaries in dyeing and finishing due to their stability in hard water and their good dispersant and protective-colloid properties. Surfactants discovered in the following years were often first used as textile auxiliaries before becoming established in other areas.

Some other inventions that illustrate the historical development of textile auxiliaries are given in the following:

1885: M. DE CHARDONNET recommends use of additives such as fats and glycerol for the spinning baths for nitrate-silk fibers.

1904: The Stockhausen company produces solvent-containing preparation from monopole soaps as fat-dissolving soaps.

1905: KNECHT and WITT observe that felting of wool can be prevented by treatment with aqueous chlorine.

1911. LILIENFELD applies for patent for poly fatty acids (condensed castor oil) as a spinning aid for viscose fibers.

1913: GÜNTHER (BASF) develops Nekal A (a diisopropylnaphthalenesulfonate), the first high-activity wetting agent.

1921: Pott & Co. introduce the Nekal-based fat-dissolving soap Neomerpin.

1924: HERRMANN and HAEHNEL produce poly(vinyl alcohol), the first synthetic size and finishing agent. Shortly afterwards, styrene – maleic anhydride and polyacrylate sizes are introduced.

1925: BERTSCH develops the spin brighteners Avirol KM extra and Avirol AM extra by means of an improved sulfonation of oils.

1927: Stockhausen introduces Praestabitol V, a highly sulfonated oil, and recommends it as a wetting agent for wool dying and vat dying.

1928: BERTSCH and SCHRAUT develop the primary alkyl sulfates, which under the name Gardinol (Böhme Fettchemie) become widely accepted as detergents in the textile industry.

1928: Tootal Broadhurst Lee applies for patent for the use of urea – formaldehyde resins as finishing agents for textiles made of cellulose fibers. This initiated the development of permanent finishes for cotton and regenerated cellulose fibers.

1929: DAIMLER and PLATZ (Hoechst) apply for patent for the first fatty acid condensation product (Igepon A), followed in 1930 by Igepon T, (fatty acid methyl tauride), a high-quality detergent for the textile industry.

1930: SCHÖLLER, WITTWER et al. (BASF) develop the ethylene oxide adducts. Wide range of uses, depending on hydrophobic group and poly(glycol ether) chain: emulsifiers (Emulphor), detergents (Leonil, Hostapal), levelling agents (Perigal, Uniperol), softeners and preparations (Soromin, Afilan).

1930: Polyphosphates introduced into textile industry as water softeners.

1933: REED (United States) synthesizes the alkyl sulfonates. Shortly afterwards, large-scale production of these products under the name Mersolates commences in Leuna and Wolfen. Important raw materials for detergents and textile auxiliaries.

1934: Shell introduces the olefin sulfonates (Teepol). Still important detergent raw materials and textile auxiliaries.

1935: IG Farbenindustrie patents production of nitrilotriacetic acid and ethylenediaminetetraacetic acid. Widespread use of these complex formers in the textile industry begins.

1938: BENER applies for patent for a process for the wash-fast permanent shaping of cotton goods with resins.

1947: Rohm and Haas (United States) introduce dimethylolethyleneurea as easy care finish for cotton.

1949: COX (Du Pont) invents the modificator viscose spinning process in which amines are used as coagulation regulators.

1953: Sun Chemical Corp. introduces dimethyltriazinone as finishing agent.

In the 1950s and 1960s, numerous processes for the manufacture of easy care garments were developed; silicones were introduced as finishing agents and spinning auxiliaries, and fluorochemicals as oil and soil repellants.

The history of textile auxiliaries in more recent decades is linked to the growth of synthetic fiber production, as well as to the successful use of finishing techniques to optimize cellulosic chemical fibers and the most important natural fibers, cotton and wool, to meet modern consumer demands.

Discussion of environment aspects, led by Germany to a great extent, has included comparison of the properties, advantages, and disadvantages of natural and chemical fibers, including fibers such as flax and linen [8] and has led to new product development. Superfine fibers (microfibers [9]) for clothing and technical textiles required special treatment methods. The first imitation leather made of microfibers was marketed in Japan under the trade name Alcantara in 1970. Microfibers for sports clothing must compete with membranes [9].

World production of synthetic fibers increased from 70 000 t in 1950 to 17.2×10^6 t in 1992 [10], the manufacture of the most important fibers, i.e., polyester (PES), polyamide (PA), and polyacrylic (PAN) fibers having been made possible by developments that took place within the short time period 1938 – 1943. In 1968, the proportion of synthetic fibers used in the Federal Republic of Germany reached 50 % of the total [11].

In the 1960s and 1970s, interest was directed towards easy-to-keep properties, e.g., minimum ironing, resistance to soiling

and graying, ease of washing and cleaning, wearing comfort, the deliberate modification of the textile character (hand), and the fashion aspects of the finished textiles. Nonwoven materials contributed to the stabilizing function of textiles, e.g., interlining materials, and high volumes of cheap, hygienic nonwoven filling materials were used in the bedroom in place of blankets and feathers. The low density of textiles was increasingly utilized in automobile construction, and this trend will continue in the future. Shaped components made of recycled fiber from textile waste bonded with hardening resins consisting of phenol – formaldehyde (novolac) or melamine – formaldehyde precondensates, or butadiene – styrene – acrylonitrile copolymers in dispersion form provided good handling properties, elegance, practical use, comfort, safety, and ecological advantages in automobiles [12].

Textiles with technical applications have become much more important in recent times (Chap. Auxiliaries for Technical Textiles) [13]. Their importance in ancient times is shown, e.g., by the use of glass fibers by the ancient Egyptians to decorate vases [14]. In 1886, C. AUER VON WELSBACH used cotton fibers saturated with lanthanum salts and then ignited to increase the brightness of gas flames [15]. The first light bulb of T. A. EDISON contained a carbon fiber filament.

In 1921 – 1922, animal glue was marketed by the company Röhm, Darmstadt, Germany, for use as a binder for waste-fiber wadding [16], which was employed as an upholstery material for mattresses and automobile seats. In 1928, bonded fiber was used by the company Weisweiler and Kalff, Euskirchen, Germany, to make stiffening material for shoes.

The first styrene – butadiene copolymer was synthesized by Bayer in 1929. Natural rubber was used to bond nonwoven material in 1930. In 1936, the company Chicopee, United States, obtained a patent for the chemical bonding of wet and dry formed nonwoven material, and soon afterwards the company Freudenberg, Germany, used natural rubber to bind carded nonwovens [17], [18].

The trade name Appretan (from *Appretur*, the German for textile finishing) was registered on April 20th, 1934 [19]. It denoted a product group poly(vinyl acetate) used for many years to complement and replace natural materials (e.g., unmodified starch, shellac, and animal glue) to give a stiffening finish.

In 1947, the company Freudenberg introduced nonwoven interlining to the market. Malimo technology (sewing technique) was discovered in 1949. The large-scale production of nonwovens started in 1952. The development of a wide range of special polymer dispersions, mainly acrylate based and self-cross-linking, began in the 1950s connected to the name H. WILHELM, BASF [20]. The history of nonwovens in the former German Democratic Republic is described in [21].

The development work that led to spunbonded web technology was started in the late 1950s by Du Pont, United States, and Freudenberg, Germany [22]. The first spunbonded was marketed by Du Pont in 1965 under the trade name Reemay. Today, geotextiles or impregnated spun PES in roof coverings are typical technical textiles of the spunbonded type.

A patent was applied for in 1943 in the United States for the first tufted-carpet machine, and production started in the early 1950s. Production of tufted carpets began in Germany in 1956. The coating materials, mainly synthetic latex (butadiene – styrene copolymers), made it possible to manufacture these durable, hardwearing carpets, whose textile character and cheapness compared with woven carpets have now become fully accepted by the consumers.

The dates of more recent events in the history of technical textiles are given below:

1954	Lamination (foam rubber – textile)
1955/56	The first European iron-on (transfer) materials made from polymer powders
1960	Rapid growth of flocking technology in Europe (initiated in the 1930s)
1964	Durogan process: latex foam-backed carpets
1966	Discovery of paper coating (wet deposition) of clothing in the United States; production of Corfam synthetic leather (Du Pont)

New methods were necessary for the finishing of double-ribbed knitted fabrics [23].

Research and development also affect testing methods, which must meet the latest requirements. For example, as the result of increased efforts in recycling and rot-proofing biological degradation, a long used test had to be modified (DIN 53 933/1, new draft of 1990).

1.3. Economic Aspects

In 1992, 112 000 t textile auxiliaries were produced in Germany [24]. These were textile auxiliaries in the classical sense as defined by TEGEWA and 11 300 t textile dyes and 100 000 t textile chemicals. The total sales of textile auxiliaries in Germany are shown in Table 1 [25].

Table 1. Sales of textile auxiliaries in Germany, 10^6 €

	1989	1990	1991	1992
Domestic	270	277	259	246
Export	623	605	572	566

According to figures from the European Disposables and Nonwovens Association EDANA [26], ca. 40 000 t binders (calculated as 100 % solids) were used in Western Europe in nonwovens. Figures from EDANA are given in Table 2 [27]. Because not all binder consumers are members of EDANA (e.g., some wadding producers), the quantities quoted may be too low. Also, not all the needle-felt materials used as floor coverings are included.

Table 2. Binders for nonwoven fabrics in Western Europe (tonnes dry weight; EDANA statistics [26], [27]

Dispersions	1980	1985	1990	1993
Acrylates, vinyl acetate – acrylate	16 800	18 400	15 600	20 000
Vinyl acetates including vinyl acetate – ethylene	2 500	5 200	5 200	8 600
Total of above	19 300	23 600	20 800	28 600
SBR + NBR*	8 000	9 300	9 900	9 400
Halogen-containing (flame retardant)	2 300	600	700	200
Others	2 100	1 600	7 600	3 000
Grand total	31 700	35 100	39 000	41 200

* SBR=Styrene – butadiene rubber; NBR=Acrylonitrile–butadiene rubber.

According to [28] ca. 180 000 t of vinyl and acrylic dispersions, with a sales value of ca. $ 184×10^6, were produced in Western Europe in 1989. The same source quotes consumptions in the United States in 1990 for nonwoven materials as follows: 26 000 t pure acrylate dispersions, 17 000 t vinyl acetate – ethylene copolymer dispersions, 14 500 t butadiene – styrene copolymer dispersions, and 7700 t vinyl acetate – acrylic ester copolymer dispersions.

A rough estimate of latex consumption in Western Europe, which correlates mainly with carpet production, is 250 000 t, calculated as 100 % dry material [29]. In the carpet sector, by far the most important materials used are butadiene copolymer dispersions (SBR latex, carboxylated latex).

Consumption of textile auxiliaries, e.g., fiber preparation agents, can be estimated from the figures for fiber production (Table 3) [10]. Production of nonwoven material in Western Europe shows the following growth pattern [30]: 185 600 t in 1980, 450 700 t in 1990 [17], 480 000 t in 1991, and 510 000 t in 1992. In 1990, production in the United States was 650 000 t, and in Japan 200 000 t [31].

Table 3. World fiber production in 1992

Fiber	Production	
	10^3 t	%
Raw cotton	18 115	46
Raw wool	1 676	4
Man-made fibers :		
Synthetic fibers	17 212	43
Cellulosic chemical fibers	2 620	7
Total	39 623	
Glass fibers (estimated)	1 650	

Classifying fiber production by use (Table 4) [10] reveals that a high proportion of fibers are used in technical textiles: 22 % in the United States, and 18 % in Western Europe. In Japan, the proportion of fiber used for technical textiles is even higher. Technical textiles have become an important growth market [32], the highest proportion of the total (23 %) being used in motor vehicles [33]. There is increasing use of textiles for roofing and damp courses. Of the total of 228×10^6 m^2 in Germany 1992, glass-fiber nonwovens account for ca. 40×10^6 m^2. Polyester spunbonds with polymer – bitumen coatings are used in smaller, but growing, amounts. Other damp course materials include felt, cardboard, plastics, and metal [34].

Table 4. Applications of textile fibers in 1991 (10^3 t)

Use	Chemical fibers		Cotton		Wool		Total	
	USA	EC	USA	EC	USA	EC	USA	EC

Clothing	1007	1169	1470	666	65	300	2542	2135
Carpets	1334	596	9	9	12	73	1355	677
Other home textiles	402	518	582	384	5	34	988	935
Tires	128	86	2	2			130	88
Other technical textiles	1110	588	150	127	6	6	1265	721
Total	3981	2956	2213	1188	88	413	6280	4556

Table 5 shows the growth and recession that have taken place in textile finishing and printing in Germany [35]. The worldwide growth in printed textiles (ca. 18 000×10^6 m^2 in 1980, and ca. 22 500×10^6 m^2 in 1990) occurred mainly outside Europe [36].

Table 5. Textile finishing in Germany

	1985	1989	1990	1992
Textile finishing, 10^6 m^2	1721	1802	1799	1514
Outher wear, %		35		33
Lining fabrics, %		16		17
Underclothes, %		7		3
Furnishing fabrics, %		6		15
Home textiles, %		13		13
Technical fabrics, %		20		15
Other textiles, %		3		5
Textile printing, 10^6 m^2	477	527	573	502
Clothing (woven), %		44		47
Clothing (knitted), %		15		13
Furnishing fabrics, %		12		14
Home textiles, %		21		19
Technical textiles, %		4		4
Other textiles, %		4		3

1.4. Environmental Aspects

As late as the early 1970s, impressive developments in graft polymerization and finishing technologies using solvents were being made [11]. From today's viewpoint, such techniques are regarded as ecologically problematic. Potential problems include impurities in wastewater and exhaust gases, and toxic residues on the textile. Since the 1970s, producers of textile auxiliaries, the textile industry, the textile finishing industry, and producers of fibers, nonwovens, carpets, and technical textiles have been confronted by a continually growing number of regulations and restrictions relating to the environmental situation. These concern wastewater, emissions into the atmosphere, the chemical substances themselves [6], [37], waste [38], packaging [39], soil and waste disposal [40], energy recovery from residual materials [41], and the recovery and reuse of used textiles [42], unused auxiliaries, residual treatment liquor, etc.

For the year 1992, the Association of German Textile Finishers stated that the cost of environmental protection as a percentage of total finishing costs was 7.2 %. In the Netherlands, the figure was 4.1 %, in Switzerland 2.6 %, in the United Kingdom 3.25 %, and in France 1.5 % [35].

Attacks on the textile industry have been mounted by the media in Germany. Various representatives of textile companies have responded [43]. Many other publications have attempted to elucidate the environmental situation, which has become very complex [44]. Activities of this kind in the United States are described in [45], and the environmental situation with respect to polymers in [46].

Wastewater. The textile finishing industry is one of the largest consumers of water. The auxiliary agents and process steps of the textile finishing industry have been brought into question more and more by technically oriented discussions on wastewater discharge laws in the Federal Republic of Germany [47], and the introduction of a wastewater levy in 1981 [48]. Stricter threshold values, drafted in Appendix 38 (1993) to the water resources policy act indicate the further restrictions the textile finishing industry will have to deal with (Table 6).

Table 6. Wastewater from textile manufacture and finishing: specifications from the draft Appendix to 38th

Wastewater Administrative Regulation (selection as of 1993)

Parameters	Specifications new (old)	Relevance to textile auxiliaries
Chemical oxygen demand (COD)	160 mg/L (280 mg/L)	affected by all organic products, COD dependent on C content and product concentration, COD values of 300 – 2200 mg O_2/g product
Biological oxygen demand (BOD_5)	25 mg/L (40 mg/L)	wide range, e.g., 5 – 700 mg/L 700 maize starch or alkane sulfonate (60 %)
		150 melamine cross-linker (75 %)
		85 paraffinic hydrophobic agent, cationic (30 %)
		10 silicone hydrophobic agent (30 %)
Phosphorus, total	2 mg/L (–)	P-containing flame retardants, phosphates, preparation agents containing phosphoric esters, P-containing sequestrants
Ammonium nitrogen	10 mg/L (5 mg/L)	pH regulator (ammonia), thickening agents, flame retardants, ammonium phosphates, melamine cross-linking agents, amine hydrochloride catalyst, ammonium stearate
Total nitrogen in ammonium, nitrite, and nitrate	20 mg/L	pH regulator (ammonia), thickening agents, flame retardants, ammonium phosphates, melamine cross-linking agents, amine hydrochloride catalyst, ammonium stearate
Iron	3 mg/L (–)	
Aluminum	3 mg/L (–)	flame-retardant filler aluminum hydroxide, precipitating agent aluminum sulfate
Sulfite	1 mg/L (1 mg/L)	reducing agent
Total hydrocarbons	10 mg/L (10 mg/L)	certain thickening agents, defoamers in small quantities
Chromium(VI) compounds from use as oxidizing agents for sulfur dyes	none	
Chromium(VI) from other sources	0.1 mg/L	
Chlorine-organic carriers	none	
Mercury and mercury compounds	none	preservatives
Alkylphenol ethoxylates (APEO) from washing and cleaning agents	none	surfactants for dispersing, wetting, and washing
High-boiling mineral oil hydrocarbons (initial *bp* > 200 °C)	none	lubricants
Absorbable organic halogen compounds (AOX)	0.5 mg/L (–)	chlorine- and bromine-containing flame retardants, chlorinecontaining bleaching agents
Free chlorine	0.3 mg/L (0.3 mg/L)	chlorine bleaching agents
Total aromatics (benzene and homologues)	0.1 mg/L	
Hydrocarbons	15.0 mg/L	defoamers
Sulfide	1.0 mg/L (0.1 mg/L)	sulfur dyeing
Total chromium	0.5 mg/L (2 mg/L)	
Copper	0.5 mg/L (1 mg/L)	in pigments
Nickel	0.5 mg/L (–)	
Zinc	2.0 mg/L (3.0 mg/L)	Zn catalysts
Biologically difficult to degrade complexing agents such as EDTA	to be avoided	sequestering agents, stabilizing agents for coating pastes
Toxicity to fish as dilution	2.0 (4.0)	depends on product concentration

factor G_F

Some general requirements are as follows:

- Reprocessing and reuse of wash water from the printing plant.
- Segregation of dyeing and washing liquors with high dye contents, and of residues from finishing materials.
- Precleaning of vessels and hanks without generation of wastewater.
- Residues of textile chemicals, dyes, and auxiliaries to be collected separately and reprocessed to recover or dispose of the material.
- Recovery of synthetic sizing materials after desizing, followed by reuse, separate disposal, or treatment. Use of synthetic sizing agents that give a COD elimination of 80 % after 7 d.
- Avoidance of the use of organic complexing agents (such as EDTA) that are difficult to eliminate biologically.
- Surfactants must comply with the law on washing and cleaning agents (and associated regulations) concerning biological degradability.
- Separate wastewater streams for desizing, bleaching, printing, dyeing, finishing, coating, laminating, cleaning of vessels and hanks, and antifelting finishing of wool.

Equivalent Austrian regulations are given in [49].

Sizing and desizing have attracted the most attention from the outset [47], [50-52], as not only sizing agents (starches) that are easily biologically degraded were appearing in wastewater in large quantities, but also microorganismadapting synthetic sizing agents [poly(vinyl alcohol) [53]] and nonadapting sizing agents (polyacrylates, cellulose ethers).

The rate of COD elimination in the presence of activated sludge in a biological clarification plant can be determined in the laboratory by the Zahn – Wellens test [54]. If air is introduced into the system with agitation, activated sludge degrades poly(vinyl alcohol) and other sizing agents, and textile auxiliaries in general [55]. The Zahn – Wellens test has been introduced as a test method [56], [57]. The elimination rates determined by this test are applicable to wastewater treatment plants using the mechanical/ biological system [58]. The COD and BOD were at first the primary consideration to minimize pollution. Later it became necessary to minimize organic halogen compounds or to avoid them completely. These materials are mainly organochlorine carriers and hypochlorite, a bleaching agent, both of which produce high AOX values [59]. The AOX contamination caused by the bleaching agent sodium chlorite is considerably less than that caused by hypochlorites [60]. A future AOX limit of 3 mg/L for partial streams has been drafted in Germany.

Special collection of residues and excess materials from impregnation baths, from the nip rollers, spraying, coating, and sizing, and the reuse or separate waste disposal of these materials have led to considerable reduction in the pollution of natural water resources.

The decoloration of wastewater (e.g., by ozone or by adsorption) is being further developed [61].

Classification of Water Contaminants. Textile auxiliaries are divided into various classes according to the hazard to water that they represent [62]:

Class 0: Generally harmless to water (e.g., citric acid, calcium carbonate)
Class 1: Slightly harmful to water (e.g., acetic acid, cross-linking agents, polymer dispersions)
Class 2: Harmful to water (e.g., ammonia, formaldehyde, diesel fuel, surfactants)
Class 3: Very harmful to water (e.g., perchloroethylene)

Emissions to the Atmosphere. Emissions due to the use of textile auxiliaries in the textile finishing industry, including the use of the finished textile, recycling, and incineration include the following:

Application of the product	Odor from the product. MAK values apply, if available, and also TA-Luft [63].
Drying, curing, fixing	Volatile and steam-volatile substances are produced. The TA-Luft and 4. BImSchG apply.
Textile is finished	Odor from the textile. Determination of chemical constituents on the textile (e.g., formaldehyde in the chamber test). Tests for ecological labels and special trade names.
Use in interior spaces	For example, automobiles [64]: odor, fogging [65].
Recycling	Odor, soiling. TA-Luft, sometimes MAK values.
Incineration with utilization of energy	Volatile materials, combustion products. TA-Luft, etc.

Foodstuffs and Commodities. Certain textiles come into contact with foodstuffs. The laws and regulations that apply to

foodstuffs then apply to the textile auxiliaries [66]. The commodity regulations in Germany (Bedarfsgegenstände-Verordnung) [67] in which a number of EC guidelines have been incorporated, deals with plastics that come into contact with foodstuffs. The permitted monomers and other starting materials are covered by Appendix 3/Part A, while Part B covers provisionally permitted materials.

Skin Compatibility. The information given in safety data sheets [68] on skin and mucous membrane compatibility (skin and mucous membrane irritation determined on rabbits) are valid for commercial auxiliaries. However, in the case of textiles, the effects of drying and fixing always weaken these effects considerably. The material on the textile is present at low concentration or as a thin film in a solid, hardened state, with inert properties compared with liquid commercial products in ready-to-use form. In spite of the very large number of different textile finishes produced, very little skin damage can be attributed to the direct effect of chemicals on textiles. However, mechanical effects (rubbing) also play a role.

Seals of Quality and Ecological Labels. In 1991, the “Gemeinschaft umweltfreundlicher Teppichboden” (GuT) (association of environmentally-friendly carpets) [69] was founded, followed in 1992 by the “Verein für verbraucher- und umweltfreundlicher Textilien” (association of consumer-friendly and environmentally-friendly textiles) [70]. The seal of quality for environmentally-friendly carpets is granted by Western European carpet manufacturers according to established testing criteria.

The following harmful substances must be absent from floor-coverings bearing the “environmentally-friendly” label (GuT regulation of 1994):

Pentachlorophenol (PCP)
Formaldehyde
Butadiene
Vinyl chloride (VC)
Volatile chlorofluorocarbons (CFCs)
Asbestos
Dyeing accelerators (carriers)
Azo-dyes whose amino-components are in the Class III A 1 list of MAK values
Vinyl acetate
Pesticides on natural fibers

The seal of quality for textiles tested for harmful substances (MST) conforms with the criteria of German textile companies. The Austrian Textile Research Institute in Vienna and the Hohenstein Research Institute, Germany, mark products “Tested for harmful substances according to Öko-Tex-Standard 100” [71]. German and Austrian standards were unified in 1994 (Table 7). Formaldehyde, which is the substance most often discussed in the textile field, is included in all these quality seals and ecological labels.

Table 7. Qualifying criteria Öko-Tex-Standard 100 (selection as of 1995)

	Öko-Tex-Standard					
	101	**104**	**107**	**110**	**112**	**114**
Criterion	**Textiles for clothes except clothes for babies**	**Textiles for babies**	**Textile floor coverings**	**Upholstery fabrics**	**Bed linen**	**Home textiles**
Organochlorine carriers	n.d.*	n.d.	n.d.	n.d.	n.d.	n.d.
No flame-retardent finish						
No biocidal finish						
Colorfastnesses						
Water	3		3	3	3	3
Perspiration (acidic)	3 – 4			3 – 4	3 – 4	3 – 4
Perspiration (alcaline)	3 – 4			3 – 4	3 – 4	3 – 4
Abrasion wet	2 – 3	2 – 3	2 – 3	2 – 3	2 – 3	2 – 3
Abrasion dry	4	4	4	4	4	4
Saliva and perspiration		yes			yes, for babies	yes, for babies

Emissions, mg/m^3						
Toluene			0.1			
Styrene			0.005			
4-Vinyl cyclohexene			0.002			
4-Phenyl cyclohexene			0.03			
Butadiene			0.002			
Vinyl chloride			0.002			
Aromatic hydrocarbons			0.3			
Volatile organic compounds			0.5			
Odor						
SNV 195 651 (Switzerland)			3			
Others	no odor with mildew, heavy benzine, fish, aromatics or odor improver					
pH						
Wool, silk	4.0 – 7.5	4.0 – 7.5	4.0 – 7.5	4.0 – 7.5	4.0 – 7.5	4.0 – 7.5
Others	4.8 – 7.5	4.8 – 7.5	4.8 – 7.5	4.8 – 7.5	4.8 – 7.5	4.8 – 7.5
Formaldehyde, mg/kg						
		20	300	300	75	
Skin-tight	75					75
Skin-distant			300			300
For babies					20	20
Eluable heavy metals, mg/kg**						
As	1.0	0.2	1.0	1.0	1.0 (0.2)	1.0 (0.2)
Pb	1.0	0.2	1.0	1.0	1.0 (0.2)	1.0 (0.2)
Cd	0.1	0.1	0.1	0.1	0.1	0.1
Cr	2.0	1.0	2.0	2.0	2.0 (1.0)	2.0 (1.0)
Cr (VI)	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Co	4.0	1.0	4.0	4.0	4.0 (1.0)	4.0 (1.0)
Cu	50.0	25.0	50.0	50.0	50.0 (25.0)	50.0 (25.0)
Ni	4.0	1.0	4.0	4.0	4.0 (1.0)	4.0 (1.0)
Hg	0.02	0.02	0.02	0.02	0.02	0.02
Pesticides and pentachlorophenols, mg/kg**						
DDT, DDD, DDE	1.0	0.5	1.0	1.0	1.0 (0.5)	1.0 (0.5)
HCH's	1.0	0.5	1.0	1.0	1.0 (0.5)	1.0 (0.5)
Lindan	0.5	0.25	0.5	0.5	0.5 (0.25)	0.5 (0.25)
Aldrin	0.2	0.1	0.2	0.2	0.2 (0.1)	0.2 (0.1)
Dieldrin	0.2	0.1	0.2	0.2	0.2 (0.1)	0.2 (0.1)
Toxaphen	0.5	0.5	0.5	0.5	0.5	0.5
Heptachlorine, heptachlorine epoxide	0.5	0.25	0.5	0.5	0.5 (0.25)	0.5 (0.25)
2,4-D	0.1	0.1	0.1	0.1	0.1	0.1
2,4,5-T	0.05	0.05	0.05	0.05	0.05	0.05
Pentachlorphenol	0.5	0.05	0.5	0.5	0.5 (0.05)	0.5 (0.05)
Total incl. PCP	1.0	0.5	1.0	1.0	1.0 (0.5)	1.0 (0.5)
Dyestuffs						
Able to release amines of MAK classes III A1 and III A2 or dyestuffs may	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.

	be carcinogenic or allergic

* n.d. = not detectable.
** Values in parentheses: textiles for babies.

[Top of Page]

2. Auxiliaries for Fiber Production and Processing

Kurt Marquardt

2.1. Introduction

In the production and processing of fibers, products are used for which the term “auxiliaries” is hardly adequate, as they are in fact virtually indispensable for the production of man-made fibers or for spinning yarns. The products may be grouped according to their area of use, i.e., for:

1. Synthetic fibers or natural fibers
2. Primary spinning or secondary spinning

The general term “spinning auxiliaries” includes:

1. Spinning preparations (spin finishes) for fiber production, also known as primary spinning
2. Spinning oils, lubricants, and batching oils for secondary spinning
3. Antistatic and lubricating agents for aftertreatment of fibers in dyeing and finishing, as well as winding oils and lubricating (waxing) agents for yarn and threads

Products used as additives to spinning solutions and spinning melts include modifiers for the production of regenerated cellulose fibers and rayon, delusterants, optical brighteners, coloring pigments or master batches (only for spinning melts), antiaging agents, flame retardants, and incorporated antistatic agents.

In the trade there are numerous terms in use which vary with geographical areas and/or fields of application, very often without any logical coherence. Such terms include spinning preparation, avivage, spin finish, antistatic agent, antistat, fiber finish, spinning aid, winding oil, lubricating agent, lubricant, spinning oil, batching oil.

It has not yet been possible to achieve agreement on an international nomenclature that would enable the subgroups listed above under the general heading “spinning auxiliaries” to be clearly distinguished from one another.

The general term “avivage,” which is often met with, should not be used in the field of fiber production as it implies reactivation, although in this case there is no restoration of an original condition [72-76].

2.2. Economic Aspects

World production of man-made fibers, which continues to increase, reached ca. 20×10^6 t in 1992. This required an amount of fiber preparation materials (based on active substance) of ca. 150 000 t, assuming that this was used economically. Losses associated with the production process are included in this estimate [77].

Spinning preparations contribute significantly to the quality of man-made fibers and yarns, and their importance is reflected in added value. Modern products have made possible innovations in production techniques and equipment, giving higher primary and secondary production rates. They have also been important in the development of finer types of fiber.

2.3. Use of Fiber Preparations

Spinning preparations ensure the best possible results during production and all subsequent processing stages for both staple fibers and filaments. Applied soon after the product leaves the spinneret and forms the fiber, they start to act immediately and, in the case of staple fibers, accompany these through all processing stages, from secondary spinning right up to the finished yarn. They function similarly in the production of filament yarns, whether flat, draw twisted, or textured.

Spinning preparations or spin finishes, lubricants, and batching oils were developed for use in secondary spinning, and ensure optimum running properties in all processing stages of worsted, semiworsted, carded (woollen), and bast fiber spinning mills, and in the production of needle punched nonwovens.

Antistatic and lubricating agents are always used when the spinning preparation has been removed by a previous wet process (e.g., dyeing). In the case of staple fibers, antistatic and lubricating agents must restore these to a condition such that their behavior during processing is again suitable for the requirements of secondary spinning. They also support mechanical processes in the final finishing of the textile surface structure, e.g., raising a nap (teazling). In yarn dyeing, such auxiliaries are used both to facilitate rewinding and to improve processing of the yarn to form woven textiles.

In general, winding oils, lubricating and waxing agents are applied to yarns to improve especially warp and weft knitting including Raschel. With sewing threads, these auxiliaries prevent excessive heating of the needles, and help to increase the sewing speed and prevent thermally induced thread breakages and sewing defects.

Such oils are also used when winding filament yarns onto warp beams. In two-for-one twisting, these oils prevent the yarn

defects that can be caused by increased abrasion in the twisting pot.

2.3.1. *Use of Fiber Preparations in Primary Spinning*

There are various methods of applying spinning preparations in primary spinning, depending on the production technology used. These can involve one or more steps. Equipment used includes licking rolls, preparation thread guide tubes, dipping baths, and sprinkling and spraying equipment. Combinations of these methods are also used. However, in all cases, the process is one of forced application [78-83].

Viscose Rayon Production. The production of viscose rayon can be continuous or discontinuous. In the continuous process, the filaments successively pass through the various aftertreatment operations, such as washing, carbon disulfide recovery, desulfurization, bleaching, preparation application, and drying. The preparation is applied in a dipping bath [82].

In discontinuous operation, the bobbins or spinning cakes are attached to carriers similar to those of a package dyeing apparatus, and run through the various aftertreatment baths in the individual machine sections. Each section carries out a single treatment [82].

Melt-Spun Filament Yarns. When melt spinning filament yarns from polymers such as polyesters, polyamide 6, polyamide 66, and polypropylene, the spinning preparation is generally applied at the bottom of the spinning duct, whereby two processes are used. In the classical method using licking rolls, the threads are passed tangentially onto a ceramic roll rotating with its lower part in the direction of movement of the threads in a trough filled with the solution or emulsion of the preparation. A regulator ensures that the trough filling is always kept at the same level. In this way, a thin film of the dilute preparation is transferred from the roll as the filament yarn passes over it.

The amount of preparation applied is controlled by varying the rotation speed of the roll or, within limits, the degree of dilution of the preparation [78-82].

To ensure that the film thickness of the preparation remains constant on the application roll, the solutions of the products used must behave as Newtonian fluids.

In the production of yarns spun at high speed (drawing rates >2000 m/min) metering by means of preparation thread guide tubes has advantages. Here, the solution of preparation is fed by a precision toothed-wheel metering pump to a ceramic preparation tube with a fine bore. The thread then quantiatively picks up the amount delivered by the pump. Since the preparation systems are closed, addition of bactericides, germicides, or fungicides is not necessary.

Compared with application by rolls, this process gives considerably improved uniformity of application of the preparation to all yarn packages.

Staple Fiber Production. The method of application of spinning preparations to staple fibers differs from that used for filament yarns in that the process takes place in at least two stages. In the case of polyesters, the most important man-made fibers, three stages are generally used. This type of fiber is used here as an example, and the methods used and the role of the preparations applied to it are described below. The most important features also apply to other polymers such as polyamide and polypropylene.

Three processes are used:

1. Licking roll method
2. Immersion bath method
3. Spraying method

The choice of method depends on the fiber production technology and mechanical equipment used. The individual process stages (see Figs. 1 and 2) include:

1. Spinning
2. Preparation application
3. Coiling of the tow (not yet drawn) into cans
4. Withdrawal of the tow from the cans
5. Gathering and dividing of the tows to form separated fiber sheets
6. Immersion bath at the inlet of the drawing line
7. Drawing
8. Preliminary thermal annealing
9. Final preparation
10. Crimping
11. Drying and final thermal setting

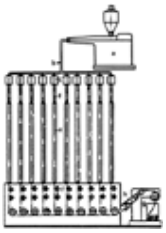

Figure 1. Schematic diagram of spinning equipment for tow production

a) Extruder; b) Melt feed; c) Spinnerets; d) Blow ducts; e) Spinning ducts; f) Spin finish application rolls; g) Tow deposited in cans; h) Spraying equipment

Figure 2. Schematic diagram of a PES tow production line

a) Can creel; b) Frame and reed; c) Infeed rollers; d) Immersion bath; e) Drawing septet; f) Heating zone for first and second drawing zones; g) Drawing and thermal setting godets; h) Cooling godets; i) Gathering tows; j) Spraying equipment; k) Dancing rolls; l) Steam chanel; m) Stuffer box; n) Dryer; o) Tow coiling; p) Dancing rolls; q) Cutter; r) Baling

Depending on the subsequent processes, the tow of man-made fiber is cut and compressed into bales, or passes as tow to the conversion stage.

The objectives and requirements of the preparation process in the individual process stages of fiber production are as follows:

The first stage of preparation application takes place after cooling of the multifilaments at the bottom of the spinning duct of each spinning station by tangential contact with licking rolls. A tow is formed by deflecting through 90°, and, before this is coiled into cans, it is again sprayed with preparation solution to increase the pickup of water.

Good wetting is necessary as the fiber must be well set at this stage to ensure that the coiling of the tow into the cans and also the withdrawal at a later stage will take place without problems.

To ensure a stable and uniform condition of each tow coming from the creel, several of them are gathered to form one or more larger sheets, and passed through an immersion bath. This compensates for irregularities caused by variation in the dwell times of the individual tows in the cans before drawing. As drawing is generally carried out wet, electrical charges play no part in this process. However, filament – filament friction is of major importance in the drawing operation. The individual filaments must slide against each other easily, as otherwise the number of capillary, breakages will increase considerably.

Other requirements include very low volatility and good thermal stability of the preparations. Before crimping, application of the final preparation is usually carried out, either in the dipping bath with subsequent squeezing off, or by spraying. As the crimping equipment has a strong squeezing action, the concentration of the bath or the spraying solution must be such that the desired final spin finish pickup is achieved [78-81].

Viscose Staple Fiber. With viscose staple fiber, the preparation is applied as the last operation of a series of aftertreatments. Directly after the drawing stage, the tow is cut to the desired staple length, placed in this form on a perforated endless belt, and passed through the aftertreatment line, which is divided into several sections, separated by pairs of squeeze rolls.

Before entering the preparation section, the wet fiber fleece is squeezed off very strongly, sprinkled with preparation solution, and again squeezed off. The preparation solution must be uniformly distributed over the staple fiber. A gentle squeezing off at the exit of the preparation section ensures good uniformity of the applied preparation. The preparation liquor squeezed off is returned to a tank, restored to a usable concentration by adding a stock solution of preparation, and returned to the preparation section for reuse [82].

2.3.2. *Use of Antistatic or Lubricating Agents in Secondary Spinning*

In secondary spinning, antistats or lubricants are applied exclusively by spraying aqueous solutions. This is carried out by using spray nozzle systems or, as is still common, e.g., in worsted spinning, with the aid of rotating disks [73].

Cotton and Rotor Spinning Mills. The mechanical equipment and functioning of these types of spinning systems require the best possible primary preparation of man-made fibers.

In exceptional cases, or in countries with unsuitable climates, the yarn feeding behavior can be improved at the spinning pretreatment stage, i.e., before carding, by spraying on extremely small amounts of antistatic spinning agents.

Cotton contains its own optimal “preparation” in the form of solid cotton wax. Moreover, its histological structure gives good spinning properties.

Worsted and Semiworsted Spinning. Lubricants and antistatic agents are a permanent feature of worsted spinning. This is true for wool, man-made fibers, and blends thereof.

The natural wool grease present in the fiber is unsuitable as a lubricant owing to its adhesive properties, and must be removed as far as possible by scouring before processing.

The scouring of the wool not only removes natural grease, grease residues, and other soil, but also creates a suitable substrate to which a lubricant appropriate for the needs of worsted spinning can be applied. Application is carried out by dripping or spraying onto the dried, scoured wool before initial carding.

After the combed top has been produced, it is again lubricated on the draw frame or during blending with man-made fibers on the melangeuse, either by the sandwich method at the infeed of the draw frame or on the delivered top.

In a semiworsted or long staple spinning, the fiber flock is sprayed before carding. Because coarser fibers must be processed here, larger quantities of spinning auxiliaries and water must be applied.

Woolen spinning mills use the largest quantities of lubricants. The lubricant is applied to the fiber flock also before carding.

2.3.3. *Use of Antistatic and Lubricating Agents in Dyeing*

The use of antistatic and lubricating agents is necessary if the original preparation is lost during a dyeing operation. The new agents must be suitable for the changed fiber parameters, since the crimping, surface condition, and stiffness of the fiber have changed during dyeing.

It is therefore not effective to apply the original preparation again. At this stage cationic active single component systems, which are suitable for exhausting processes, should be used.

In procedures for the forced application such as back washing or centrifugal methods, nonionic/anionic systems can also be used, provided that no selective adsorption on the fiber surface takes place. The product used is chosen so as to enable the frictional and antistatic properties required for the subsequent spinning process to be reestablished [84].

2.4. Physical Properties

The task of preparations is to modify the interactions of the fiber surfaces with one another and with the various surfaces of fiber guiding devices and processing machines. This requires precise knowledge of their surface behavior and hence of the resulting frictional processes.

A number of theories of surface coating, chemisorption, and intermediate valency forces are described in the literature. These explain how molecules of almost equal molecular mass and equal viscosity can show different frictional behavior. Similar phenomena can also be observed with different substrates [73], [85-97].

For example, if a prepared polymer fiber is heated above its glass transition temperature, reorientation of the molecules of the preparation on the fiber surface often takes place, leading to changes in zeta potential and frictional behavior. Moreover, frictional processes lead to partial redistribution of the preparation between the fiber material and the friction partner.

This is influenced by the surface of the solid undergoing friction, its structure, and the tendency to form a surface layer.

The object of a preparation is not to produce the lowest possible friction, but to give the optimum frictional properties that ensure steady and uniform processing.

Important parameters for frictional behavior include the consistency of the preparation (e.g., solid paste or oily liquid), viscosity, yarn speed, and reciprocal applied pressure.

These frictional phenomena are shown in the so-called Striebeck curve (Fig. 3). This shows that solid paste systems are less sensitive than oily liquid components. The latter are subject to the shear forces that develop within the liquid film between the fiber and the surface of the solid.

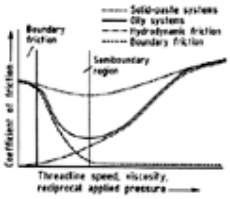

Figure 3. Striebeck curve for oily and solid-paste systems

A very important function of a preparation is conductive removal of the electrostatic charge formed by the frictional processes. This is achieved by incorporating polar groups in the preparation molecules.

Formulating a spinning preparation necessitates knowledge of the physical properties of the individual components and of the full range of requirements for the intended use. The following main properties are required for all areas of use:

1. Physiological harmlessness (e.g., dermatologically) and nontoxic by inhalation if an aerosol is formed
2. Ecological harmlessness, i.e., biologically degradable (>80 % in the OECD simulation test)
3. Antistatic action
4. Good spreading properties on the fiber surface (wetting properties)
5. Resistance to aging, i.e., no tendency towards resinification
6. Easily removable by washing
7. No detrimental effects on later finishing processes
8. No yellowing tendency
9. Good shelf life of the product and good durability after deposition on the substrate
10. Minimum migration into the interior of the fiber and no interaction with the fiber polymer (oligomer retention)

Requirements for Filament Preparations. Requirements for *flat yarns* are:

1. Low dynamic friction against solids, e.g., thread guides
2. Good fiber bond (i.e., static fiber – fiber friction for problem-free winding)
3. Low to medium dynamic fiber – fiber friction to avoid fluff formation (capillary breakage) during stretching
4. Good stability of the lubricant film

5. No spattering
6. High thermal stability and low tendency to evaporate, avoiding fume and condensate formation on hot stretching

Requirements for *textured fibers*, especially for high-speed spinning of POY (partially oriented yarn), are:

1. For friction texturing, a high dynamic dry (fiber/solid) friction coefficient is required to impart optimum twist distribution in the friction unit.
2. Little or no snow formation
3. No glazing effect on ceramic disks
4. Compatibility with polyurethane disks (does not cause swelling of the polyurethane material)
5. Residue-free thermal degradation, avoiding soiling problems on heaters
6. Consistent behavior under tension, giving uniform crimping data, extension, strength, and uniform dyeing properties with no sealed-in areas (tight spots)

Requirements for Staple Fiber Preparations. In both the production of staple fibers and in processing them in secondary spinning, the requirements of secondary spinning must be considered as well as the criteria listed above. The following are of equal importance:

1. Antistatic behavior
2. Fiber – fiber and fiber – metal friction in relation to the secondary spinning process

Thus, operating frictional conditions can be assigned to individual secondary spinning processes. The friction diagrams described in the literature indicate the operating situation for each spinning system, but vary from one author to another. These apparent discrepancies in the position of the optimum friction conditions are due to differences in measuring methods and their individual evaluation.

2.5. Chemical Properties

The basic materials for fiber preparations are oleochemical and petrochemical derivatives [83], [85], [88], [91].

Filament preparations are oily liquids formulated from several components which complement each other or have synergistic activity. They are used as aqueous solutions, stable emulsions, or neat oils.

Staple fiber preparations can be liquids, pastes, and waxlike compounds.

Depending on the area of use, antistatic and lubricating agents, as well as those applied after dyeing include both formulations of the filament preparation type and individual compounds of the staple fiber preparations.

As well as providing a purely lubricating or cohesive function, these surface-active compounds can provide all intermediate functions, depending on their chemical structure, and can also have emulsifying and antistatic effects.

The following review includes only a selection of the commonly used compounds:

1. Lubricants based on naturally-occurring materials:
   a.
      Mineral
      Oils
      Waxes

   b.
      Vegetable (triglycerides)
      Coconut oil
      Peanut oil

2. Synthetic lubricants:
   Esters of higher carboxylic acids with fatty alcohols or polyhydroxylic alcohols
   Polyethers (ethylene oxide based)
   Polyethers (ethylene oxide – propylene oxide based)
   Silicones

3. Emulsifiers and wetting agents:
   a. Anionic
      Fatty acid soaps with alkali metals or alkanolamines
      Sulfonated vegetable oils
      Sulfosuccinates
      Salts of ethoxylated alkyl phosphates

   b. Cationic
      Fatty amines and their salts
      Ethoxylated fatty amines
      Quaternary nitrogen compounds
      Quaternary phosphorus compounds

c. Nonionic
   - Glyceryl mono-oleates
   - Ethoxylated fatty alcohols
   - Ethoxylated fatty acids
   - Ethoxylated sorbitan esters

d. Amphoteric
   - Aminoacids and their salts
   - Betaines
   - Sulfobetaines

4. Antistatic agents
   a. Anionic
      - Alkyl phosphates
      - Alkyl phosphonates
      - Alkanolamine salts
      - Soaps

   b. Cationic
      - Quaternary ammonium compounds, e.g., chlorides, methylsulfates, ethylsulfates
      - Quaternary phosphonium compounds

   c. Nonionic
      - Ethoxylated fatty acid amides

   d. Amphoteric
      - Betaines
      - Sulfobetaines

5. Additives include corrosion inhibitors, antioxidants, biocides, and solubilizers

Analysis is carried out by all relevant analytical methods from, for example, oleochemistry, petrochemistry, and surfactant chemistry [74], [98].

2.6. Fiber Preparations for Special Applications

Fiber preparations for special applications include products used on fibers for hygienic and medical applications. The compounds used must comply with the U.S. Food and Drug Administration (FDA) and the European Community (EC) Regulations, and must meet strict requirements with respect to toxicological and dermatological properties.

2.7. Additives for Spinning Solutions, Spinning Baths, and Spinning Melts

These include pigments and master batches for spinning dyes, optical brighteners, delusterants to reduce the luster of man-made fibers, and flame retardants.

In wet spinning processes such as viscose fiber and polyacrylonitrile fiber production, the pigments are added to the viscose or polyacrylonitrile polymer solution. However, in melt spinning, master batches are used. These consist of pigments incorporated in high concentration in a carrier polymer. They are added to the extruder in chip form. The delusterant used is extremely finely divided titanium dioxide, which is added to the spinning solutions or polymer melts.

In the production of viscose fibers and filaments, modifiers are used, especially to improve the wet strength modulus. Without modifiers, only fibers with a pronounced core – shell structure can be produced. The core is largely amorphous and is thus responsible for the low strength. The modifiers delay the saponification of the xanthogenate, and hence lead to the formation of a uniform crystalline structure over the entire cross section of the fiber. The modifiers consist of amine derivatives such as ethoxylated fatty amines, quaternary ammonium bases, and cyclohexylamine. Nitrogen-free products include polyethylene glycols [82].

2.8. Environmental Aspects

The increased importance attached to environmental protection has led to many compounds that were formerly successfully used being dispensed with. The alkyl aryl ethoxides are an example. Although these are biologically degradable, some metabolites formed have been found to be toxic to fish.

In the case of textiles for domestic use, the products described in this chapter enter the home environment on domestic textiles and, in the case of clothing, come into direct contact with the skin.

Fiber preparations appear in the wastewater from fiber producers or, at a later stage, from textile producers. Owing to increasingly strict environmental regulations, especially in Europe, in future there will be a requirement to use only biologically degradable compounds in fiber preparations.

Some of the fiber preparation material appears in the exhaust air, especially where thermal processes are used during fiber production and textile finishing. The general legislation that regulates emissions from industrial plants also applies to the exhaust air from fiber and textile plants [99].

2.9. Commercial Products

Some producers and trade names of auxiliaries used in the production and processing of fibers are listed below [100]:

BASF	Soromin
Dr. Th. Böhme	Filapan, Synthesin
CHT	Avistat
Hansa Textilchemie	Duron
Henkel	BK-, LW-Marken, Nopco
Hoechst	Afilan, Leomin
ICI	Cirrasol
Schill & Seilacher	Limanol, Silastol, Syntex
Stockhausen	Praelanol, Tallopol
Takemoto	Delion
Zschimmer & Schwarz	Fasavin

[Top of Page]

3. Sizing Agents

Kaspar Schlüter

3.1. Introduction

Woven fabrics are two-dimensional bodies consisting of fibers in the form of threads intersecting at right angles (the warp and weft), produced using the shed-forming method. Whereas each weft thread is stressed only briefly as it is placed in position, the warp threads undergo repeated stress during each insertion of the weft and at each change of shed. The warp threads undergo stress in the form of yarn – metal abrasion when the weft thread is pushed by the reed, by yarn – yarn abrasion during the change of shed, and by cyclic stretching processes. The warp threads are normally unable to withstand these extreme stresses, and must therefore be provided with a protective coating — the sizing agent — that adheres to the fiber, forming an abrasion-resistant, elastic film.

With staple fiber yarns, the sizing agents have the task of making the yarn resistant to the frictional processes that take place during weaving. Protruding fibers are caused to adhere to the main body of the yarn, thereby preventing neighboring warp threads from catching or entangling. The overall increase in the tensile strength of the thread of ca. 20 % is of minor importance, but the increase in strength at the weakest points is crucial [101]. With filament yarns, the sizing agents cause the individual filaments to stick together, giving a compact thread, drastically reducing the risk of filament breakage. The sizing agent must adhere strongly to the fiber, and its film properties should be largely independent of the climatic conditions, especially atmospheric humidity, and be unaffected by fiber finishes and sizing additives. The elongation of the warp thread should not be reduced by presence of the sizing agent.

After the gray cloth has been woven, the task of the sizing agent is complete. As it would usually have a deleterious effect on subsequent finishing processes, it must be completely removed. Removal is simple in the case of cold-water-soluble sizing agents, but starch products that are insoluble in cold water require preliminary enzymatic or oxidative breakdown before the desizing stage (Section Desizing.). The removal of the sizing agent may pose special wastewater treatment problems in finishing plants.

A large number of classes of chemical substances are used as sizing agents. They can be divided into two main groups [102]:

1. Macromolecular natural products and their derivatives: starches and starch derivatives, carboxymethyl cellulose, galactomannan, and tamarind flour derivatives
2. Synthetic polymers: poly(vinyl alcohol), poly(meth)acrylates, polyester condensates, and polyvinyl compounds

The growth of synthetic fibers and the developments in weaving technology have accelerated the development of synthetic sizing agents. The consumption of sizing agents worldwide is ca. 700 000 t, of which ca. 70 % are starches and starch derivatives, and ca. 15 % are poly(vinyl alcohol). The remainder consists of carboxymethyl cellulose (CMC), poly(meth) acrylates, polyester condensates, and other sizing materials. The areas of use of the various sizing agents are summarized in Table 8. General information on sizing agents is given in [103], [104].

Table 8. Areas of use of sizing agents

	Starch	CMC	Galactomannan	PVAL*	Polyacrylates	Polyesters	Vinyl acetate copolymers
Staple fiber yarns							
Cotton, linen viscose staple	+	+	+	+	+		+

fiber							
Wool	+	+	+	+	+		
Polyester – cellulose	+	+	+	+	+		+
Polyester – wool		+	+	+	+		
Polyester		+	+	+	+		
Polyamide				+	+		
Polyacrylonitrile		+		+	+		
Filament yarns							
Viscose				+	+		
(Tri)acetate				+	+		+
Polyamide				+	+		
Polyester				+	+	+	

* PVAL = Poly(vinyl alcohol).

3.2. Sizing Methods

As a rule, sizing agents are applied to the warp threads from aqueous liquors (Fig. 4). The warp threads pass from the warp beams (a) through the sizing trough where they are dipped in the sizing liquor (b) and impregnated with it. The excess liquor is then squeezed off in the squeezing equipment (d), giving a liquor pickup of 50 – 200 %, depending on the warp parameters (fiber material, spinning method, thread thickness, etc.), sizing machine parameters (speed, squeezing pressure, etc.), and liquor parameters (viscosity, temperature, wetting properties, etc.) [105], [106]. The warp threads are dried in the cylinder drier, and are then separated in the dry splitting section (f). They then pass through the comb (g), which adjusts the distance between the warp threads and makes them all parallel, after which they are wound onto the weaving beam (h), which is then taken to the loom. There are two methods of applying the size: either all the warp threads are sized simultaneously or a proportion of the warp threads are sized as a warp beam (the so- called single-end sizing process). These warp beams are then assembled in a separate operation to form the complete weaving beam. This single-end sizing process is used only for filament yarns, to prevent damage to the delicate individual filaments on separation of the threads stuck together by the sizing material. Staple fiber yarns are always sized all at the same time. In state-of-the-art technology, high-pressure squeezing off is used, ensuring that the liquor pickup is small, keeping drying costs low [107], [108].

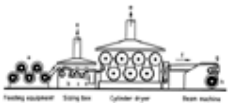

Figure 4. Schematic diagram of the sizing process

a) Warp beams; b) Sizing liquor; c) Immersion roller; d) Squeezing equipment; e) Steam outlet; f) Dry splitting section; g) Comb; h) Weaving beam

In the cold sizing process, very coarse or twisted yarns are coated with a small amount of sizing liquor by kiss roller application during warp beam preparation. The application process is carried out at room temperature, and drying of the yarn is not necessary [109], [110]. Other alternative sizing techniques, such as solvent sizing [111], [112], hot melt sizing [113-116], and foam sizing [117-122] have not yet become established.

3.3. Sizing Agents Based on Natural Raw Materials

The most important natural raw materials used as sizing agents are starch and cellulose. Other polysaccharides have so far played only a minor role.

3.3.1. *Starch and Starch Derivatives*

Starch and its derivatives are still the most important class of sizing agents with respect to total consumption. This is because of their low price, good sizing effect, and worldwide availability. The raw material basis of this class of sizing materials is naturally occurring starch, a polysaccharide based on α-D-glucopyranose. Starch is not a single chemical substance, but is composed of two structurally different polymers: amylose and amylopectin. Amylose consists of chains of glucose units linked by α-1,4-glucosidic bonds, whereas amylopectin additionally contains α-1,6-glucosidic bonds which cause branching of the polymer chain [123-126] (⟶ Starch).

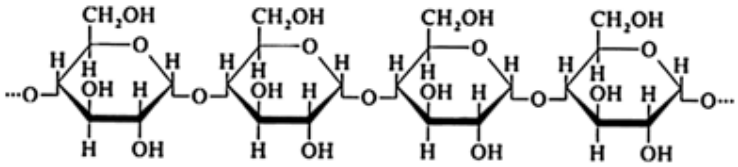

Amylose

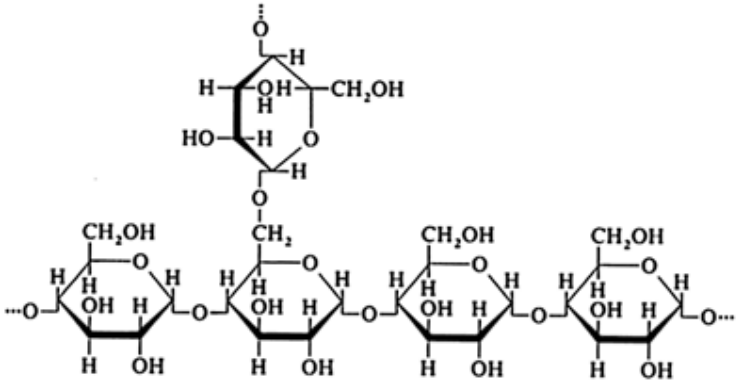

Amylopectin

The degree of polymerization of amylose is ca. 100 – 1000 glucose units, and that of amylopectin, ca. 6000 – 10^6. Amylopectin is the main constituent of starch, making up 73 – 86 % of the total, depending on the type of starch.

The most important sizing agents are potato, maize, and tapioca starches. Wheat, rice, and sago starches are also used. The characteristic properties of these starches are determined by the amylose/amylopectin ratio, the degree of polymerization of these two constituents, and the size and fine structure of the starch grain. These parameters determine the swelling and solution behavior, and also the properties of the film (Table 9).

Table 9. Composition and properties of starches

Property	Potato starch	Maize starch	Tapioca starch
Grain size, µm	30	15	20
Amylose content, %	21	28	17
Glucose units per molecule of amylose	ca. 3000	ca. 8000	ca. 3000
Starch paste temperature, °C	60 – 65	75 – 80	65 – 70
Viscosity of starch paste	very high	medium	high
Shear stability	medium to low	low	high
Retrogradation	medium to low	high	low
Strength and elasticity of film	high	low	high

Natural starch is insoluble in cold water because of the hydrogen bonds linking parallel polymer chains. The starch is brought into “solution” by heating. The starch grains first absorb water until swelling is at its maximum. Above a certain temperature, characteristic for each type of starch and known as the gelatinization temperature, the starch grains burst and form a gel. The viscosity increases to a maximum, and then decreases asymptotically to a limiting value as the solubilized polymer molecules disperse. Complete solubilization of the individual molecules of a starch grain only occurs above 100 °C. The viscosity value is important in size application, as it has a considerable effect on the amount of liquor pickup.

On storage and with decreasing temperature, starch pastes solidify to a pulpy mass. This retrogradation is caused by stretching of the molecular chains, parallel alignment of the chains, and formation of hydrogen bonds between neighboring chains, with loss of water of hydration [126]. This retrogradation has detrimental effects on the sizing agent, leading to poor storage properties, skin formation, formation of deposits on the rollers, and reduced adhesive strength. Therefore, natural starches are increasingly being replaced by starch derivatives.

All modified starches that have lost their original properties are referred to as starch derivatives. These include thin-boiling starches, dextrins, starch esters, and starch ethers (⟶ Starch).

Thin-boiling starches are produced by acid hydrolysis or oxidative degradation in aqueous suspension, and dextrins are produced by thermal depolymerization, usually in the presence of acids. They gelatinize at low temperatures, give solutions of low viscosity, and can be dissolved in high concentrations. Furthermore, it is easier to produce a liquor with a predetermined viscosity, and the tendency to retrogradation is considerably reduced.

The starch esters mainly used in sizing materials are those of phosphoric acid (the phosphate starches) and acetic acid (the acetyl starches). These starch derivatives are usually not only esterified, but also depolymerized, giving lower liquor viscosities and decreased retrogradation. In general, they give better sizing effects than thin-boiling starches.

The three most important types of starch ethers are the hydroxyethyl, hydroxypropyl, and carboxymethyl starches, produced by reaction of starch with ethylene oxide, propylene oxide, and chloracetic acid or sodium chloroacetate in the presence of caustic soda, respectively. The degree of substitution of these starch derivatives is generally ca. 0.1 or less. Because of its ionic character, sodium carboxymethyl starch is soluble in cold water and therefore does not require enzymatic desizing. Also, the starch ethers have a better sizing effect than starch derivatives that are simply depolymerized.

The mechanical properties of films cast from solutions of starch or starch derivatives depend on the degree of hydration (which depends on the relative humidity of the atmosphere in the weaving mill) and on the ratio of amylose to amylopectin and type of modified starch.

Uses. The main use of starch and starch derivatives is for sizing yarns of pure cotton and its blends with other fibers. For these yarns, a large number of sizing formulations are used which consist either exclusively or principally of starch or starch derivatives. Starch and its derivatives adhere relatively strongly to cotton [127], [128]. Cotton yarns woven on high-speed looms or yarn blends with a high proportion of synthetic fiber must be sized with sizing formulations that contain additionally carboxymethyl cellulose, poly(vinyl alcohol), or poly(meth)-acrylates to improve the sizing effect. Starches are detected by the blue coloration with iodine, and this reaction is also used for the semiquantitative determination of residual size content [129].

Ecology. Starch and starch derivatives are in general biologically degradable, and derivatives such as the hydroxypropyl, phosphate, and carboxymethyl starches are also classified as such [130], [131]. The COD of sizing agents based on starch is ca. 900 – 1000 mg O_2/g starch (or derivative). In comparison with other sizing agents, starch and its derivatives must be applied at ca. twice the amount used for carboxymethyl cellulose or synthetic sizing agents because of the poorer sizing effect, thereby leading to a higher COD in the wastewater. Toxicity towards waterborne organisms such as fish, daphnia, and bacteria is very low (LC_0 > 100 mg/L).

3.3.2. *Sodium Carboxymethyl Cellulose*

Unlike starch, cellulose contains the 1,4- β-glucosidic link and has a straight chain (⟶ Cellulose – Molecular Structure). Depending on its origin, the degree of polymerization of the macromolecule corresponds to $\bar{P}_\eta$ = 1000 – 7000 glucose units. Because of its chemical uniformity and high degree of crystallinity, cellulose hardly swells in water and is practically insoluble. This prevents the use of the natural product as a sizing agent. To obtain water-soluble cellulose derivatives, the cellulose is etherified or esterified [132], [133] (⟶ Cellulose Ethers, ⟶ Cellulose Esters).

Of the water-soluble cellulose derivatives, only sodium carboxymethyl cellulose (CMC) is used as a sizing agent. CMC is produced by reacting cellulose with sodium hydroxide and sodium chloroacetate, whereby the cellulose polymer is simultaneously depolymerized. Solubility in water is produced when the degree of substitution (DS) is greater than ca. 0.5. CMC with a DS of ca. 0.7 is usually used for sizing agents. In the production process, large amounts of salts are produced, mainly sodium chloride and sodium glycollate. Unpurified grades of CMC containing up to 40 % salt and purified grades with < 2 % salt are commercially available. The salt content only slightly affects the rheology of CMC, but because it can lead to considerable corrosion damage in sizing machines and looms, purified CMC grades are preferred. Depending on the method of production and the DS, CMC can contain small amounts of insoluble cellulose fibers. If the content of these materials is too high, deposits may form on the drying cylinders of the sizing machine, which can seriously affect the sizing process. Owing to its chemical affinity to cellulose, purified CMC adheres strongly to cotton. It also adheres to polyester and polyacrylonitrile, its performance sometimes equalling that of poly(meth)acrylates and poly(vinyl alcohol) [127], [128].

On drying CMC solutions, tough, transparent films are formed whose mechanical properties are highly dependent on water content. This in turn depends on the salt content of the film and the relative humidity of the air (24 – 31 % water content at 20 °C and 65 % R.H.).

Uses. CMC is the preferred additive for increasing the adhesion of starch sizes to cotton and cotton blends. Other important uses include the sizing of cotton yarns for weaving terry toweling and viscose staple fiber yarns. Because CMC is easily removed by washing, fabrics with warp threads sized with CMC can be dyed without a preliminary wash. Owing to the high viscosity of CMC, the concentration range for sizing liquors is limited to ca. 20 – 100 g/L. CMC is suitable for size recovery processes, whether used on its own or in combination with poly(vinyl alcohol) (see Section Poly(Vinyl Alcohols)) [134].

The methods of detecting CMC are much less selective and less practical than the iodine – starch test. CMC can be detected by precipitation with uranyl nitrate, copper (II) sulfate, or cationic surfactants [135].

Ecology. Although CMC is degradable by biological action, its rate of breakdown does not justify its classification as a readily biologically degradable material [130]. The COD is ca. 900 mg O_2/g CMC. As with starch derivatives, the toxicity of CMC towards waterborne organisms is very low (LC_0 > 100 mg/L).

3.3.3. *Sizing Agents Based on Other Natural Raw Materials*

Older sizing agents based on linseed oil [136] and bone or leather glue [137] are scarcely used today, having been largely replaced by synthetic sizing agents. However, sizing agents based on galactomannans have some importance as biologically degradable alternatives to CMC [138], [139]. Although galactomannan derivatives from guar or carob flour have long been known as thickening agents for printing, they have not been used as sizing agents. Products of this type

are produced from the polysaccharide extracted from cassia beans, which is made up of galactose and mannose monosaccharide units in the ratio 1 : 5. Solubility in water and ease of removal by washing are produced by etherification to form the hydroxypropyl or carboxymethyl derivatives.

Tamarind seed also appears to be a suitable raw material for sizing agents [140], [141]. However, like the galactomannan polysaccharides, this raw material has not yet been used in any substantial quantity.

3.4. Synthetic Sizing Agents

The growth in the use of synthetic fibers and high-speed looms (projectile and air jet looms) means that the warp yarns must receive optimum sizing treatment, and this has led to an increasing demand for high-quality synthetic sizing agents. The most important of these are poly(vinyl alcohols), poly(meth)acrylates, and polyester condensates.

3.4.1. *Poly(Vinyl Alcohols)*

Poly(vinyl alcohols) have been used as sizing agents for around 60 years, but large-scale use started only in 1960, coinciding with the appearance of the polyester – cotton yarns. They are the most important group of synthetic products used as sizing agents.

Technical-grade poly(vinyl alcohol) can be considered to be formally a copolymer of vinyl alcohol and vinyl acetate:

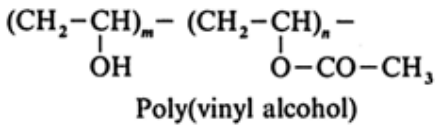

Poly(vinyl alcohol)

As monomeric vinyl alcohol does not exist, poly(vinyl alcohol) is produced by the hydrolysis or deesterification of the poly (vinyl acetate) (⟶ Polyvinyl Compounds, Others – Poly(Vinyl Alcohol)).

The parameters that determine sizing properties are the degree of polymerization (DP) and the degree of hydrolysis. Products are generally characterized by the viscosity of an aqueous 4 % solution as a measure of the DP, and by the degree of hydrolysis in mol %. These parameters are controlled via the DP of the starting material poly(vinyl acetate), and conditions during hydrolysis (deesterification). There are two grades: the partly hydrolyzed grade with a degree of hydrolysis of ca. 88 %, and the fully hydrolyzed grade with a degree of hydrolysis of ca. 98 %. Maximum solubility in water occurs for a degree of hydrolysis of ca. 88 %.

Partially hydrolyzed poly(vinyl alcohol) grades adhere more strongly to nonpolar fibers such as polyester [127], [128] and also show maximum water solubility. These are therefore preferred for use as sizing agents.

In addition to these conventional poly(vinyl alcohol) grades, copolymers with methyl methacrylate or with comonomers having carboxyl functionality are also used as sizing agents. These have the advantage of better water solubility under the alkaline conditions used for desizing and scouring, and after the thermofixing process applied to fabrics made from polyester-containing yarns [142-144].

The poly(vinyl alcohol) films are tough and elastic. Their mechanical strength, which is unmatched by other synthetic sizing agents, is less sensitive to atmospheric humidity than that of other water-soluble sizing agents. The water pickup of poly (vinyl alcohol) films is 5 – 10 % at 65 % R.H., increasing to only 20 % at 86 % R.H. [145].

Uses. Poly(vinyl alcohol) has its greatest importance in the sizing of staple fiber yarns. Because of its high efficiency on cellulose and polyester fibers compared with starch and its derivatives [127], [144], it is used as a high-quality additive in sizing formulations for cotton and especially for polyester blends. Here, as much as 50 % of the sizing agent can be saved compared with sizing exclusively with starch and its derivatives. Special areas of use include the sizing of wool where it is generally used in combination with CMC, and of polyacrylonitrile, viscose, acetate, triacetate, and polyamide filament yarns. Although poly(vinyl alcohol) adheres strongly to polyester and has an excellent sizing effect on polyester staple fiber yarns, it is normally not effective on polyester filament yarns.

Poly(vinyl alcohol) is extremely suitable for size recovery processes, whether in a partial recycling operation by means of simple washing techniques or by full recycling using ultrafiltration. Conventional partially or fully hydrolyzed poly(vinyl alcohols) and copolymers are equally suitable [144], [146-148]. Poly(vinyl alcohol) can be used alone as a sizing agent or in combination with carboxymethyl cellulose or polyacrylate [149].

Several methods for determining poly(vinyl alcohol) are available. Of these, the iodine reaction [150], intensified by addition of boric acid [151] is the most important, and can also be used for quantitative determination in wastewater [145], [152].

Ecology. Unlike most synthetic polymers, poly(vinyl alcohol) is biologically degradable, the degradability being dependent on the adaptation of the activated sludge microorganisms [130], [153], [154]. Given such adaptation, poly(vinyl alcohol) can be classified as having good biological degradability. The breakdown behavior is independent of the degree of hydrolysis and the viscosity of the poly(vinyl alcohol), which has a COD of ca. 1700 mg O_2/g. The aquatic toxicity towards fish, daphnia, and bacteria is LC_0 > 100 mg/L [130].

3.4.2. *Poly(meth)acrylates*

Poly(meth)acrylates are another important class of synthetic sizing agents. They show considerable variation in chemical structure, and hence have a very wide range of applications (⟶ Polyacrylamides and Poly(Acrylic Acids);

→ Polyacrylates; → Polymethacrylates).

The important chemical building blocks and the construction scheme for poly(meth)acrylate sizing agents are shown below [155]:

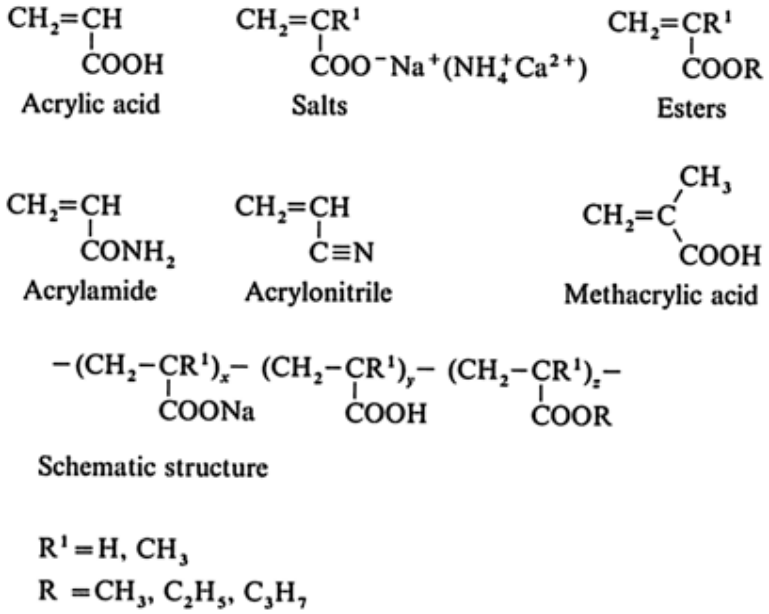

Hydrophilic monomers such as methacrylic acid and its salts and acrylamide give good adhesion to polar fibers, ease of removal by washing, and compatibility with alkalis. Hydrophobic monomers such as the acrylic esters increase the elasticity of the sizing films and improve the ease of removal from the wastewater. (Meth)acrylic acids and their esters increase the hardness of the films, while acrylonitrile improves their elasticity and abrasion resistance. The (meth)acrylic acid in the polymer is usually present as the sodium or ammonium salt. The preferred base is ammonia, because of the low hygroscopicity of the product. Complete or partial neutralization with alkaline earth bases [MgO, $Ca(OH)_2$] leads to a further reduction in the hygroscopicity, and also avoids the problem with ammonium polyacrylates that they tend to be difficult to remove by washing due to formation of cross-links during thermofixing [156].

Poly(meth)acrylate sizing agents based on methacrylic esters give good sizing effects owing to their high adhesive strength to natural and synthetic fibers and exhibit low sensitivity to atmospheric humidity [155].

An example of chemical optimization to enable extreme quality requirements to be met is provided by the sizing agents used for filament yarns for water jet looms:

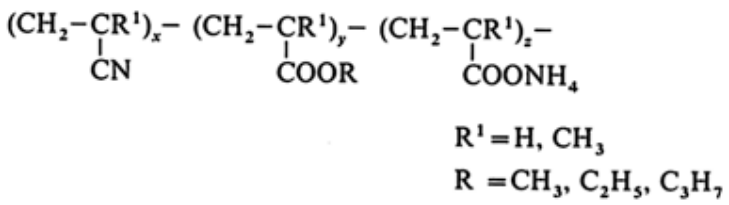

On drying, ammonia is cleaved from the water-soluble ammonium polyacrylate, converting it to the water-insoluble, almost nonswelling acid form. A polyester or polyamide yarn sized in this way can be woven on water jet looms. The sizing agent can then be washed off under alkaline conditions.

Uses. Unlike other staple fiber sizing agents, which are usually supplied as the 100 % solid product, poly(meth)acrylates are usually produced as a 25 % aqueous solution, and have a wide range of uses owing to their wide variability. Like poly (vinyl alcohol), they are used as a high-quality additive in sizing formulations for all cotton yarns, especially for blends of cotton with synthetic fibers, and also for viscose staple fiber, wool, and polyacrylonitrile. Poly(acrylic acid) is the preferred sizing agent for polyamide filament yarns. Products based on acrylic acid copolymers are suitable for viscose, acetate, and triacetate filament yarns, while the hydrophobic poly(meth)acrylates based on (meth)acrylic esters are preferred for sizing polyester filament yarns. The excellent water-solubility of the sizing agents based on (meth)acrylic acid means that the gray cloth can be treated in subsequent processes without first desizing it. Recent research has shown that hydrophobic polyacrylate sizing agents are suitable for recovery and concentration by ultrafiltration [149].

Ecology. Poly(meth)acrylates are practically biologically nondegradable. Hydrophobic methacrylic acid ester copolymers are adsorbed by the activated sludge, and can be eliminated in this way, while conventional hydrophilic products are eliminated to only a small extent [153], [155]. COD values of poly(meth)acrylate sizing agents are in the range ca. 1300 – 1700 mg O_2/g, based on solids material. Their aquatic toxicity is low, like that of the other sizing agents.

3.4.3. *Polyester Condensates*

With the introduction of the single-end sizing process [157], it became possible to apply water-dispersible sizing agents based on polyesters. Today, these are the standard sizing agents for plain polyester filament yarns.

Polyester sizing agents are, in general, polycondensates of aromatic dicarboxylic acids with diols (e.g., ethylene glycol, diethylene glycol, or cyclohexane dimethanol) and sulfonated aromatic dicarboxylic acids (e.g., 5-Na-sulfoisophthalic acid), which provides solubility or dispersibility in water [158], [159]:

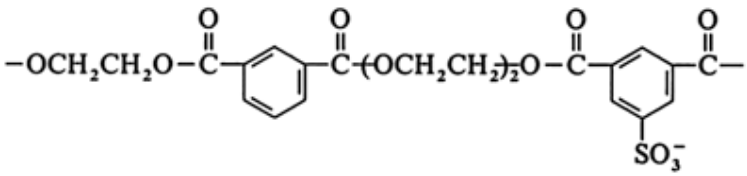

The proportion of ionic sodium sulfonate groups is such that the polyester condensates do not form true solutions in water, but dispersions of small particle size.

The properties of the polyester sizing agents can be controlled over a wide range by variation of the monomers, their relative proportions, and the molecular mass. The aim is to produce products that disperse readily in water and give films that are tough, resilient, and do not adhere to each other at high humidities. Polyester sizing agents have a chemical affinity to the polyester fiber and adhere to it strongly [127]. The possibility of interaction with the spinning preparation must be taken into account [160].

Uses. The main use of polyester sizing agents is for sizing plain polyester filament yarns. However, they are also used as a synergistic component to increase the adhesive strength when sizing staple fiber yarns made of polyester blends [161]. Note that when polyester sizing agents are washed out of the textile, there is a danger that they may be precipitated by electrolytes, especially polyvalent cations [162]. They can be determined on the textile by color reactions with basic dyes [135], [163], [164].

3.4.4. *Other Synthetic Sizing Agents*

Apart from the synthetic sizing agents described above, vinyl acetate copolymers have some practical importance. These consist of copolymers of vinyl acetate with acidic comonomers such as (meth)acrylic acid or maleic half-esters. In their almost universal applicability, they resemble the poly(meth)acrylate class of products. Owing to the tendency of vinyl acetate to hydrolyze, and the fact that the films are relatively weak and tend to stick together, these products are not so widely used as the poly(meth)acrylates. They are used as additives in starch-based sizing formulations for cotton and cotton – polyester blends to increase the adhesion, but also for acetate, triacetate, and polyester filament yarns [165].

3.5. Sizing Additives

Sizing additives are added to the sizing liquor to aid the weaving process by their softening, smoothing, or antistatic effects. Other such products are used as antifoaming agents or to increase the liquor pickup [104], [166].

Viscosity regulators are used to control the viscosity of starch-based sizing liquors. Complex formation between $Na_2B_4O_7 \cdot 10\ H_2O$ (borax) and the hydroxyl groups of starch increases the viscosity of starch paste [104], while urea reduces it [167]. Important viscosity regulators include starch-degrading agents that act by hydrolytic or oxidative cleavage of the macromolecules.

Acids are important in the industrial manufacture of starch derivatives, but are not used in the sizing process because of poor reproducibility of the effect. Amylases, which have an enzymatic action, are used in desizing, but are little used to reduce starch viscosity for sizing. The most important starch degradation agents for this application are oxidizing agents such as peroxodisulfates, peroxosulfates, and other compounds such as chloramine T (sodium *p*-toluenesulfonylchloramide, [127-65-1]). These reagents are often marketed as blends with starch, CMC, poly(vinyl alcohol), or sizing fats.

Sizing fats are used to improve the dry splitting and weaving behavior of the warp by plasticizing the sizing film, providing a smoothing and antistatic action without reducing the adhesion and abrasion resistance of the sizing agent. At the same time, the emulsifying properties must be good, and washing out of the product must be problem-free [168]. Suitable materials include self-emulsifying sizing fats such as sulfated fats and oils, and especially preparations of fatty acid esters with nonionic and anionic emulsifiers. These sizing fats often have antifoaming properties. Antistatic surfactants based on poly(glycol ethers) are added to reduce the static charge.

If it is not necessary to achieve maximum reduction of the friction coefficient, poly(ethylene glycols) with a molecular mass of ca. 4000 can be used as sizing fats. These improve the flexibility of the sizing agent film, facilitate the dry splitting process, and reduce dust formation during weaving by means of their antistatic effect. The most effective means of reducing friction coefficients is after-waxing. For this, the most suitable materials are self-emulsifying paraffin-free waxes based on fatty acids [168].

Wetting Agents. The liquor pickup during sizing depends partly on the wettability of the yarn. If the necessary size add-on can not be achieved by using high concentrations of sizing agent, a low foam wetting agent must be added. As the adhesive strength of the sizing agent can be greatly reduced by surfactants, the amount added to the sizing liquor must not exceed ca. 1 g/L. Fatty alcohol poly(glycol ethers) with a low degree of ethoxylation are usually employed.

Defoamers. If the sizing agents used tend to produce foam [e.g., poly(vinyl alcohol)] or if wetting agents are added, the addition of a defoamer is often necessary. Suitable products can be based on paraffin oils, phosphoric esters, fatty acid esters, or silicone oils. The most effective materials are silicone oils, but even small additions of these can lead to a drastic reduction in the adhesive strength of the sizing agent [168].

Preservatives. Sizing liquors that are stored for long periods and contain biologically degradable components such as starch or starch derivatives must be protected against degradation by adding fungicides and/or bactericides. Classical preservatives include formaldehyde, formaldehyde-releasing substances, phenol derivatives, and heterocyclic compounds of the isothiazolin type.

3.6. Commercial Products

A selection of sizing agents is listed in Table 10. Detailed information on these can be obtained from the various catalogues of textile auxiliaries [169] and from information leaflets provided by the manufacturers.

Table 10. Sizing agents producers

Producer	Trade name	Classification
Abco, USA	Absize, Synthesize	mixed products
	Plasticryl	poly(meth)acrylate
Air Products, USA	Airvol	poly(vinyl alcohol), mixed products
Allied Colloids, UK	Alcosize	mixed products
	Alcowax	mixed products
	Collasyn	poly(vinyl alcohol)
	Colvinal	vinyl acetate copolymer
	Sincol, Vicol	poly(meth)acrylate
	Tescol	poly(meth)acrylate, polyester
Amylum, BE	Resamyl, Supramyl	starch derivative
Avebe, NL	Kollotex	starch derivative
	Quicksolan	starch derivative, mixed products
	Solvitose, Retamyl	starch derivative
BASF, FRG	Schlichte CA/CB/CE/S	poly(meth)acrylate
	Schlichte T10	vinyl acetate copolymer
	Schlichte 3204	polyester
	Textilwachs	sizing fat
Blattmann, CH	Noredux, Noresin, Noresol	starch derivative
	Noregum	galactomannan
Bozzetto, IT	Wisacet, Wisacryl	poly(meth)acrylate, vinyl acetate copolymer
	Wisester	polyester
Du Pont, USA	Elvanol	poly(vinyl alcohol)
Eastman, USA	Eastman LB, WD	polyester
Emslandstärke, FRG	Emsize	starch derivative, mixed products
Henkel, FRG	Quellax	starch derivative
	Horsil	CMC
	Inex	poly(vinyl alcohol)
	Fibrocol, Fibropur, Fibrosint	mixed products
	Molvenin	galactomannan
	Olinor, Grünau Pillenwachs	sizing fat
Hercules, USA	Blanose	CMC
Hoechst, FRG	Arkofil	galactomannan, polyester
	Leomin	sizing fat
	Tylose	CMC
	Vinarol	poly(vinyl alcohol)
Kuraray, JP	Poval	poly(vinyl alcohol)
Lamberti, IT	Carbocel	CMC
	Lamcol	polymethacrylate, polyester, vinyl acetate copolymer, mixed products
	Overwax	sizing fat
Metsä-Serla, SF	Finnfix	CMC
Morton Intern., USA	Polyfilm	polymethacrylate, polyester, vinyl acetate copolymer

National Starch, USA	Kofilm, Lubrisize	starch derivative
Nippon Gohsei, JP	Gohsenol	poly(vinyl alcohol)
Rhône Poulenc, FR	Bevaloid	polymethacrylate, vinyl acetate copolymer
	Gerol	polyester
Roquette, FR	Tissalys	starch derivative
Seydel, USA	Aztex, Seycofilm	polymethacrylate, polyester, mixed products
	Plystran	mixed products
	Sicosize	polymethacrylate
Südstärke, FRG	Sobex, Sobotex	starch derivative

[Top of Page]

4. Pretreatment Auxiliaries

Karlheinz Gebert

4.1. Processes and Auxiliaries for the Pretreatment of Textile Materials

Pretreatment processes are used to prepare textile material (greige) for subsequent processes such as dyeing, printing, optical brightening, and finishing.

Precleansing, extraction, and bleaching processes remove foreign materials from the fibers. The reactive groups of the substrate, previously blocked by impurities, are exposed, and the whiteness of the textile is improved.

For gray goods made of natural fibers such as cotton, linen, flax, wool, and silk, the technical task is more difficult than for those made of synthetic or cellulosic chemical fibers. For example, cotton fabric can be accompanied by up to 20 % of materials that can interfere with later processing (sizes, preparations, waxes, hemicelluloses, pectins, seed husks, proteins, and alkaline earth and heavy metal compounds), whereas crude polyester textiles contain only preparations, water-soluble synthetic size, and soil, which can be removed by simple washing processes.

The processes used depend both on the form of the fiber (flocks, yarn, woven or knitted fabrics), and on the machinery available.

Pretreatment processes are substrate-specific. Because of the wide range of chemical reactions and physicochemical processes involved, the preparation of gray goods usually consists of a series of part processes. The choice and sequence of these processes mainly depend on the type of material, the quality requirements (which are often specific to the article) and the available machines or equipment (Table 11) [170].

Table 11. Pretreatment processes

Substrate	Process
Cotton, linen, polyester – cotton	singeing
	enzymatic desizing
	oxidative desizing
	demineralization
	alkaline extraction
	bleaching with hydrogen peroxide
	bleaching with sodium hypochlorite
	bleaching with sodium chlorite
	bleaching with reducing agents
	caustic soda treatment
	mercerization
Wool, silk	raw wool washing
	carbonization
	boiling-off
	bleaching with reducing agents
	bleaching with hydrogen peroxide
Synthetic fibers (PA, PES, PAN)	desizing (water-soluble sizes)
	washing
	bleaching with sodium chlorite

In the various processes, the auxiliaries have the following tasks:

1. Transport of the washing liquor into the interior of the fiber (air removal, wetting, impregnation)
2. Removal of minerals (dissolution, complexation)
3. Mobilization and removal of foreign materials and reaction products (dispersing, emulsification, complexation, colloid protecting action)
4. Fiber protection (reduction and complexation of heavy metal ions)

Modern auxiliaries are mostly multifunctional synergistic mixtures of highly active individual components; thus, a single product performs several tasks simultaneously. Other requirements include ease of handling, availability in concentrated liquid form, suitability for metering and pumping, stability in washing liquors, low foaming properties under highly turbulent washing conditions, and good storage stability. More recently, demands for minimum wastewater contamination and biological degradability have become increasingly important.

4.2. Preparation of Textile Material made of Cellulose Fibers

4.2.1. *Pretreatment Processes*

The pretreatment process depends principally on the gray goods and the object of the subsequent treatment. The nature of the material and the method of treatment determine the choice of equipment (see Table 12).

Table 12. Preparation techniques

Type of material	Type of treatment	Type of machine
Flock, cotton wool, combed sliver	discontinuous	bale apparatus
Yarn (hank, wound package)	discontinuous	apparatus for bales, hanks, cross-wound packages, warp beams
Piece (woven, knitted)	discontinuous in rope form	boiler, bale apparatus, rope apparatus, winch dyeing machine
	continuous, in rope form	place of deposit, rope storage, J-box, immersion jigger
	discontinuous, in open-width form	pad-roll, pad-batch, jigger, beam, centrifuge
	continuous, in open-width form	combination steamer, U-box, roll steamer, conveyor, sieve belt (travelling screen) steamer, roller-bed steamer, loop
		steamer, pressure HT-steamer, immersion jigger

The following methods are used:

1. “Long liquor” treatment
2. Impregnation processes
3. Open-width treatment
4. Treatment in rope form
5. Discontinuous methods
6. Continuous methods

Worldwide, the pretreatment of woven textiles is quantitatively of far greater importance than the finishing of knitted fabrics.

Discontinuous Methods. The treatment stages for open-width materials include impregnation with the chemicals containing liquor, squeezing off, batching (winding onto a roll; up to 10 000 m per roll), storage at room temperature (cold padbatch method) or in a hot steam chamber (pad-roll system), and washing in an open-width washing machine. The discontinuous treatment of fabrics in rope form is always carried out in a long bath (see above).

Continuous Methods [171-173]. When treating textiles in open-width form the material is impregnated with the treatment liquor, squeezed off, and then steamed in a reaction chamber (steamer). After steaming (2 – 45 min, depending on the type of steamer, e.g., roller steamer, combination steamer, U-box, conveyor, or open-width J-box, see Fig. 5), it is washed in an open-width washing machine. The impregnation equipment, e.g., saturator and squeezing rollers, steamer, and washing machine represent one unit of each treatment stage. In multistage preparation plants, up to three units are arranged in line (desizing, boiling-off, bleaching, see Fig. 6). In these plants, the production rate is ca. 60 – 120 m/min. When the fabric is treated in rope form, the rope is impregnated, squeezed off, and kept in a steam atmosphere in a storage device known as a J-box or boot for a treatment period of 45 – 60 min, and then washed in a rope scouring machine. Here, too, multistage plants are usual. The production rates are in the range 100 – 220 m/min.

Figure 5. Types of steamer

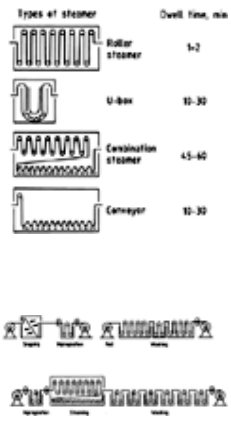

Figure 6. Continuous plant for the open-width preparation

As crease marks can be formed, especially with polyester – cotton fabrics, the open-width treatment method has become the state-of-the-art technology.

The mercerization of woven fabrics can only be carried out in the open-width state in chain or chainless mercerization plants. Circular knitted materials are usually mercerized in the tubular state.

Standard Methods for Pretreatment of Cotton and Polyester – Cotton Fabrics [170]. In the pretreatment of cotton and polyester – cotton fabrics, the three-stage process (desizing, boiling off, and bleaching) gives the best and most reproducible results. The continuous treatment of the raw textile in the open-width state ensures the best uniformity.

The mercerization of cotton and cotton blend fabrics is a standard pretreatment process in most plants for improving dye yield and dimensional stability. A continuous desizing process can be achieved by using thermally stable enzymes in a steam desizing process that can be fully integrated into the pretreatment sequence, or by simultaneous oxidative desizing and caustic scouring.

If the pretreatment equipment has only one steamer, the following two-stage processes are often used:

1. Cold bleaching, alkaline extraction
2. Cold bleaching, steam bleaching
3. Enzymatic desizing, bleaching

The “Extraction Cold Bleach” followed by the “Intensive Wash” is mainly carried out in textile finishing plants in which only pad batch equipment and an open-width washer are available. Each reduction in the amount of processing steps leads to a narrowing of the textile-quality safety margins. Special auxiliaries intended to ensure improved safety margins even using shortened methods are being developed. One of their most important tasks is to give more intensive extraction of foreign matter including heavy metal ions (catalysts).

For white articles with brightnesses of > 85 % reflectance (Elrepho, Zeiss), a second bleaching stage is always necessary (hydrogen peroxide, sodium hypochlorite, sodium chlorite).

In the European textile industry, the increasing importance attched to the limits on AOX (adsorbable organic halogens) in wastewater may lead to the complete prohibition of chlorine- containing bleaching agents.

4.2.2. *Pretreatment Part Processes*

4.2.2.1. *Desizing*

Modern dyeing and printing processes require completely desized fabrics. The most important desizing processes are enzymatic degradation of starch-based size by the impregnation-retention process, and the removal of water-soluble sizes such as carboxymethyl cellulose (CMC), poly(vinyl alcohols) and polyacrylates (see Chap. Sizing Agents). Other desizing processes depend on the machines available and the production methods. They include

1. Continuous enzymatic desizing with amylases stable under steaming conditions [high-temperature (HT) enzymes]
2. Oxidative desizing with hydrogen peroxide
3. Simultaneous oxidative desizing and boiling-off

Table 13 lists examples of desizing agents.

Table 13. Examples of desizing agents [182]

Product name	Producer	Use	Active substances
Enzymes			
Aquazym	Novo Nordisk	enzymatic desizing	enzyme

Bactosol TK	Sandoz	cold desizing	modified amylase
Beisol B 260	CHT	enzymatic desizing	amylase
Biolase TS	Hoechst	enzymatic desizing	amylase
Encylase HT	Diamalt	hot desizing	heat-resistant amylase
Rapidase fl.	International Biosynthetics	enzymatic desizing	amylase
Thermozym	Novo Nordisk	hot desizing	heat-resistant amylase
Oxidative desizing agents			
Beisol DO	CHT	oxidative desizing/boiling off	inorganic per-salts
Leonil EB	Hoechst	oxidative desizing/boiling off	mixture of peroxisulfate and surfactants
Lufibrol DO	BASF	oxidative desizing/boiling off, cold bleaching	stabilized peroxosulfate dispersion
Persulfate	Degussa	oxidative desizing/boiling off, cold bleaching	peroxosulfates

In addition to the size and the sizing auxiliaries, a number of impurities are present on the fabric, which can be partly extracted, during the desizing process. In particular, these include oils, preparations, fats, waxes that melt below 70 °C, hardening agents, and catalysts.

The result of a desizing process depends on:

1. Solvation of the polymeric sizing agent (swelling)
2. Removal of hydrophobic impurities
3. Reaction temperature
4. Reaction time
5. Washing conditions (temperature, water supply, bath turbulence)

In practice, optimum conditions hardly ever exist, so that the process only rarely gives a completely desized product. For example, cotton wax and warp wax melt during the earlier singeing process to form a film which prevents the water absorption necessary for the swelling process, and hydrophobic impurities can locally hinder liquor absorption. Furthermore, the machines and the required production rate determine or limit the reaction time. Therefore, the desizing process is usually carried out to completion in the subsequent alkaline boiling-off and bleaching stages. The initial water-insoluble starch degradation products and the residual size are broken down partly hydrolytically and partly oxidatively and removed. Likewise, the extraction of the other foreign materials usually takes place successively during the later process steps.

Enzymatic Degradation of Starches. The main component of starch, amylopectin, differs from cellulose in the linkage of the glucose units (α-glucosidic 1,4-linkage) and partial branching of the polymer chain.

Bacterial amylase [174], pancreas amylase [175], [176], and malt diastase (⟶ Enzymes – Starch-Degrading Enzymes (Amylases)) break the α-glucosidic bond of starch stereospecifically. The β-glucosidic bond of cellulose is not attacked. Fission fragments of various molecular masses (oligosaccharides, dextrins, and glucose) and various solubilities are formed.

The stability and activity of the amylases depend on pH [177], temperature, and, especially with HT amylases [178], on the presence of sodium, potassium, and calcium ions [179]. Detergents also affect enzyme activity. The effect of pH and temperature on the activity of a bacterial amylase is shown in Figure 7.

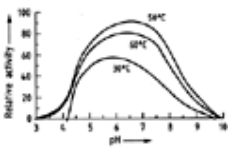

Figure 7. Activity of bacterial amylase as a function of pH and temperature

Desizing Auxiliaries. Inadequate desizing is the most frequently reported pretreatment problem [180]. Therefore, the most important auxiliaries are products that deaerate and wet the textile, so making available as much water as possible for the solvation and swelling of the hydrophilic polymer. This applies both to polysaccharides and synthetic sizes. If the wetting agent acts as an emulsifier, lipophilic impurities are removed. A dispersing action is necessary to disperse soil and water-insoluble size-degradation products. Special complexing agents act to demineralize the textile and extract the heavy metal ions (catalysts). Oxidative desizing agents, mainly peroxosulfates or peroxosulfate- containing auxiliaries are used in alkaline boiling-off and in cold bleaching with hydrogen peroxide. Wetting agents and detergents used with enzymes must not affect the enzyme activity at the concentrations used.

The effect of surfactants on the enzyme activity, measured indirectly as the glucose concentration on the textile, is shown in Figure 8.

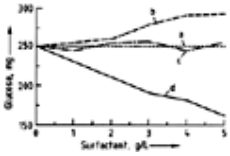

Figure 8. Effect of surfactants on the degradation of starch to glucose by enzymes

a) No surfactant; b) Octylphenol ethoxylate; c) Mixture of nonionic and anionic surfactants; d) Ester of sulfosuccinic acid

Oxidative Degradation of Starch Size. Oxidative desizing with hydrogen peroxide is carried out as the first stage of a completely continuous preparation process in some Far Eastern countries. Bleach stabilizers and wetting agents are used in addition to the bleaching agent. In Europe, simultaneous oxidative desizing and scouring is carried out, if it is desired to avoid a discontinuous batching process for enzymatic desizing, if the textile contains enzyme poisons (fungicides), or if sizes are present that are difficult to degrade (branched polysaccharides).

For oxidative desizing peroxosulfates and peroxodisulfates are generally used. Oxidative desizing with bromites discussed in the past for some time is not practiced. Oxidative desizing agents, inlcuding hydrogen peroxide, attack both starch and cellulose in the same way because of their chemical similarity.

Damage to the fibers is minimized by suitable choice of formulation and by using short reaction times, so that “mild” starch degradation occurs. The treatment should lead to dispersible rather than water-soluble degradation products. Oxidative partial desizing also takes place when a cold pad-batch bleach is carried out on undesized fabric.

Detection of Sizing Agents [181]. In practice, the detection of sizing agents on gray cloth is carried out using simple methods:

1. The iodine test for the detection of starches
2. The spot test with potassium dichromate for the detection of poly(vinyl alcohol)
3. Detection of carboxymethyl cellulose by uranyl nitrate
4. Test for soluble sizes by extraction of the textile with water

The completeness of the removal of the starch size in the pretreatment process is tested for by the very sensitive starch-iodine reaction. A scale of blue colorations is available from TEGEWA (Frankfurt/Main), which enables the residual amount of starch on the textile to be estimated.

4.2.2.2. *Alkaline Extraction*

Raw cotton contains a wide range of organic and mineral foreign materials in amounts depending on its origin:

Cellulose	88 – 96 %
Pectins	0.7 – 1.2 %
Wax	0.4 – 1.0 %
Proteins	1.1 – 1.9 %
Other organic compounds	0.5 – 1.6 %
Inorganic constituents (as ash)	0.7 – 1.6 %

The combustion residues from cotton include alkali metal salts, which do not affect the preparation process, varying amounts of calcium and magnesium phosphate, and aluminum and iron oxides. Analyses of cottons show calcium, magnesium, and iron figures that fluctuate over a surprisingly wide range. Also, modern harvesting, processing, and packing plants cause increased iron contents due to abrasion, e.g., of the often rusty steel bands used for binding the cotton bales.

Alkaline boiling-off and the subsequent washing process are aimed at the efficient extraction of the above-mentioned foreign substances, which include low molecular mass cellulose components, size degradation products, residual size, dirt, and dust. Depending on the amount of impurities and the reaction and washing conditions, a weight loss of 3 – 7 % can occur on alkaline extraction [183].

The extraction, also known as boiling-off, kier-boiling, or scouring, is carried out either batchwise or continuously by the impregnation/steaming process or by prolonged hot treatment. It is carried out at 90 – 100 °C under normal pressure. Pressurized high-temperature (HT) steamers for the open-width treatment and kier-boiling vessels for treatment in rope form are still used occasionally.

During boiling-off, intra- and intermolecular hydrogen bonds of cellulose are broken, and the polar hydroxyl groups of the polysaccharide are solvated. Swelling of the fiber takes place, which favors transport of the impurities from the inside to the outside of the fiber. During swelling, the fiber diameter increases by up to 28 %, while there is a very small increase in the length of the yarn (ca. 1 %). Uniform swelling improves the dyeing properties of the substrate.

The treatment with caustic liquor causes hydrolytic decomposition of leaves, fruit pods, and seed husk residues in the raw cotton, but these are only partially removed. Fats and waxes are also hydrolyzed, the degree of saponification depending on the temperature and treatment time. Sparingly soluble lime soaps are formed from the alkaline earths already present on the gray goods or picked up at a later stage. Pectins, which are mainly located in the cuticle of the cotton fiber and are

bound to the cellulose fiber by hydrogen bonds, can also react via their carboxyl groups with alkaline-earth metal ions and form sparingly soluble pectinates, which must also be extracted. Proteins are released from the cotton fibers during boiling-off, which is of particular significance if the cotton cloth is bleached with sodium hypochlorite, a process still operated in many countries. This leads to the formation of chloramines, which are very difficult to remove, have a yellowing effect on the material, and damage the fibers.

The alkali concentration of 40 – 60 g/L NaOH used in alkaline extraction produces sodium cellulose, which is susceptible to oxidation at high temperatures. The operations are therefore carried out in saturated steam with no air present in the equipment. Where this is not possible, e.g., in the pad-roll chamber and in older conveyors, there is a risk of fiber damage by the formation of oxycellulose and cleavage of the cellulose molecule, resulting in lower DP values (mean degree of polymerization, a measure of the fiber damage, see Section Bleaching with Hydrogen Peroxide).

Alkaline boiling-off does not completely remove all the foreign materials from the cotton. Rather, they are “chemically degraded”, and can readily be oxidatively and hydrolytically decomposed and eliminated in the subsequent bleaching stage. Depending on the initial extent of soiling of the textile materials, an amount of lightening is produced which may be adequate for dyeing in medium to dark shades. The absorbency of the treated fabric only becomes fully developed in the bleaching process.

Alternatively, the material can first be bleached (e.g., by hydrogen peroxide at room temperature in the cold pad batch process) and then boiled-off. This gives an extremely good extraction of the degradation products formed in bleaching, and hence good absorbency.

Auxiliaries for Alkaline Extraction. In alkaline extraction, the auxiliaries have the following tasks:

1. Thorough wetting of the gray goods by the boiling-off liquor
2. Emulsification of lipophilic impurities
3. Dispersion of insoluble degradation products, soil, and lime soaps
4. Complexation of hardness formers and heavy metal ions
5. Prevention of fiber damage by atmospheric oxygen

These functions of boiling-off auxiliaries are performed by alkali-stable wetting agents/detergents (e.g., alkyl sulfonates), complexing agents such as the sodium salts of nitrilotriacetic acid (NTA), ethylenediaminetetraacetic acid (EDTA), diethylenetriaminepentaacetic acid (DTPA), gluconic acid, phosphonic acids, special surfactant-free dispersing agents such as polyacrylates and phosphonates, and reducing agents such as sulfite and dithionite. Some examples of these auxiliaries are listed in Table 14.

Table 14. Some auxiliaries used in boiling-off and extraction [182]

Trade name	Manufacturer	Composition
Wetting agents/detergents		
Alcopol KBW	Allied Colloids	mixture of anionic and nonionic surfactants
DS 70	Bozzetto	sulfated ether oxide
Lanaryl NA	ICI	alkylphenol ethoxylate and solvent
Kieralon B	BASF	anionic and nonionic surfactants, APEO-free*
Solvent Scour 880	ICI	mixture of perchloroethene and surfactants
Subitol NA	CHT	modified alkanesulfonate
Surfactant-free extraction agents		
Invatex SA	Ciba – Geigy	esterified derivatives with water-soluble polymers
Ladit 1026	BK Ladenburg	polyphosphonate and hydroxycarboxylates
Lufibrol DK	BASF	mixture of polyacids
Lufibrol KB	BASF	organic polyacids and reducing agents
Sequion CM 40	Bozzetto	salt of an acrylic copolymer, phosphate-free
Surfactant- containing extraction auxiliaries		
Cottoblanc EXT	CHT	APEO-free ethoxylates, anionic surfactants, complexing agents, and dispersants
Delinol FE-NS	Dr. Th. Böhme	alkyl polyglycol ethers and complexing agents
Kappawet AT 91	Kapp-Chemie	surfactants and complexing agents
Kieralon CD	BASF	anionic and nonionic surfactants with polyacids, APEO-free
Ladit N 26 neu	BK Ladenburg	phosphonate, hydroxycarbonate, and surfactants
Sandopan SF	Sandoz	phosphoric esters, sulfonate, and polyhydroxy hydrocarbon
Sevalin 2000	TC	surfactants and sequestering agents
Tanawat BC	Sybron/Tanatex	nonionic and ionic surfactants and complexing agents

* APEO = Alkyl phenol ethoxolate.

Detailed investigations into the action of the various components of the auxiliaries are described in [184], [185]. The effects of alkali concentration, temperature, and time on the results of the treatment process are reported in [186]. The fiber-protecting caustic extraction has become known as the Lufibrol KB process (BASF) [187].

4.2.2.3. *Bleaching with Hydrogen Peroxide*

Gray goods, especially natural fibers, always contain colored foreign matter that cannot always be completely removed by washing and extraction. The whiteness of the textile materials is improved by an oxidative or reductive decomposition of these foreign materials.

Today, the most important oxidative bleaching agent is hydrogen peroxide (H_2O_2), although sodium hypochlorite (NaClO) and sodium chlorite ($NaClO_2$) are still important. The redox potentials of these substances under normal circumstances depend very much on the pH. The relatively low redox potential of H_2O_2 means that it can be used in cold and hot processes [188].

The most common reductive bleaching agent is sodium dithionite ($Na_2S_2O_4$); thiourea dioxide is used only to a small extent. Auxiliaries, including activators, stabilizers, buffer systems, and surfactants, are used to control the bleaching process to avoid damage to the treated gray goods and to develop good absorbency.

Hydrogen peroxide offers both technical and ecological advantages over sodium hypochlorite and sodium chlorite. First, wide range of bleaching processes can be used, including cold pad batch bleaching, bleaching under steaming conditions, impregnation methods, and bleaching processes in a long bath. Second, the decomposition of hydrogen peroxide which takes place during the bleaching reaction forms only water and oxygen.

Hydrogen peroxide is supplied as a 30 or 50 vol % aqueous solution. With a dissociation constant of 1.78×10^{-12} at 20 °C, H_2O_2 is a very weak acid.

A simple description of the bleaching process is as follows:

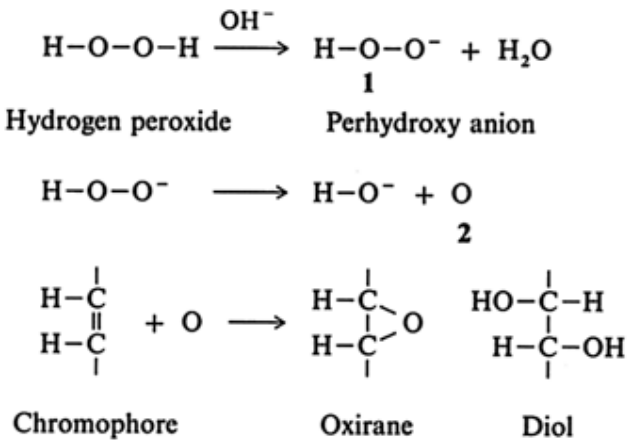

The hydrogen peroxide is activated by the addition of alkali (OH^- ions), which leads to the formation of the perhydroxy anion (**1**). This decomposes into the more stable hydroxyl ion and singlet oxygen (**2**). This “active oxygen” reacts with the double bonds of the chromophores (e.g., carotenoid pigments) that impart the characteristic brownish-yellow color to the raw cotton.

The true mechanism of the bleaching reaction is still unclear. As well as the reaction mechanism involving the perhydroxy anion and singlet oxygen atom, radical mechanisms (OH·) are under discussion [189-191].

In competition with the desired bleaching reaction, hydrogen peroxide can decompose at a comparable rate into water and oxygen with no bleaching action:

$$2\,H_2O_2 \rightarrow 2\,H_2O + O_2$$

Cleavage to form radicals requires a high activation energy (201 kJ/mol).

$$H_2O_2 \rightarrow 2\,OH\cdot$$

Heavy metal compounds and other ill-defined impurities can catalyze the decomposition reaction, which then competes with the bleaching reaction. Cleavage of the hydrogen peroxide to form radicals by the Haber – Weiss mechanism is favored especially by metals whose ions can exhibit several valencies (e.g., iron, manganese, copper, and cobalt) [192]. The attack of radicals on the cellulose fibers leads to damage, and can form “catalyst holes” in the cotton fabric [193].

Alkaline oxidation of chromophore pigments is accompanied by oxidation of cellulose (Fig. 9), with terminal hydroxyl groups being oxidized to aldehydes and carboxylic acids, and inner hydroxyl groups to ketones. Further oxidation eventually leads to cleavage of the cellulose molecule, decreasing the mean degree of polymerization (DP value).

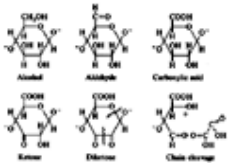

Figure 9. Oxidative degradation products of cellulose

The DP values for crude cotton, determined by viscometry, are 2600 – 3200 glucose units per cellulose molecule. In practice, DP values of 1600 – 2200 are accepted for bleached cotton. However, the measured DP values do not give any indication of the nature and number of the various oxidation stages outlined above, or of the quality of the oxycellulose formed. Eisenhut's damage factor *s* provides a relationship between the DP value of the bleached material and that of the greige material. In practice, the values for *s* are usually 0.2 – 0.7. Higher values indicate above-average damage and make it necessary to check the pretreatment process used.

$$s = \frac{1}{\log 2} \cdot \log \left[\frac{2000}{DP_{tx}} - \frac{2000}{DP_x} + 1 \right]$$

where DP_{tx} is the DP value after bleaching and DP_x the DP value before bleaching.

Effects of Hydrogen Peroxide Bleaching. In the course of the oxidative – hydrolytic bleaching reaction, the extraction processes are carried out to completion, and the following effects are achieved:

1. The seed husks are completely removed
2. The natural dyes in the raw cotton are bleached oxidatively or are extracted after chemical degradation
3. Residual size and its degradation products are hydrolyzed, oxidized, and removed
4. Whiteness is achieved with the minimum decrease in the DP value
5. The absorbency of the material is made uniform

The resulting product is suitable for dyeing and printing. To produce completely white articles, a second bleaching process is usually necessary.

The course of a peroxide bleaching under steaming conditions is shown in Figure 10, which shows the reflectance – time curves and the degradation of the cellulose as functions of the increase in brightness.

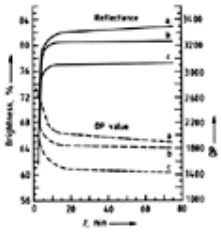

Figure 10. Peroxide bleaching under steaming conditions

a) Desizing, boiling-off, bleaching; b) Boiling-off, bleaching; c) Bleaching

The curves show that:

1. The reflectance obtained is a function of the quality of the prior extraction process.
2. The decrease in the DP values correlates with the increase in the reflectance curve, depending on the quality of the prior extraction.
3. Longer steaming times for a given formulation do not affect the bleaching. They only affect the removal of the seed husks and the oxidation of the residual sizing agent.

If the reaction time in the steamer is limited to 10 – 20 min, and if the absorption of liquor on impregnation of the fabric is only 90 – 100 %, the alkaline boiling-off stage is usually unavoidable. If the liquor pickup is 120 – 150 %, as has become technically possible, the alkali stage can usually be dispensed with.

Auxiliaries for Hydrogen Peroxide Bleaching (Table 15). When bleaching with hydrogen peroxide under alkaline conditions, bleach stabilizers must be used. These inhibit the decomposition of the bleach-active perhydroxy anions and ensure a high oxidation potential over the whole bleaching time. Residual peroxide on the fabric after bleaching in amounts of 15 – 40 % of the original hydrogen peroxide concentration indicates that the bleaching process was satisfactory, and that spontaneous catalytic decomposition has not occurred.

Table 15. Some auxiliaries used in bleaching [182]

Product	Manufacturer	Composition
Hydrogen peroxide	Peroxid-Chemie	H_2O_2

	Degussa	
Stabilizers		
Actiron ASH	Protex	mixture of complexing agents
Baystabil KF	Bayer	organic hydroxylic compounds and carboxylic acids
Candoxyl ST	Sigma	phosphonic acid derivatives
Celidon 2820	Bozzetto	organic complexing agents and silicates
Contavan SNF	CHT	organic phosphorus compounds
Cottozon 190	Rotta	phosphonates
Delinol NSR	Dr. Th. Böhme	hydroxycarboxylic acids and surfactants
Lamepon 56	Grünau	protein – fatty acid condensate
Lastabil 1020	BK Ladenburg	phosphonates and organic substances
Prestogen D	BASF	salts of organic polyacids
Stabicol A	Allied Colloids	acrylic polymers and protective colloids
Stabilol S	Henkel	combination of organic and inorganic substances
Tannex 15	Sybron/Tanatex	mixture of organic stabilizers
Thorstabilisator BF	Thor	combination of organic and inorganic compounds
Tinoclarit G	Ciba	organic metal complex compounds
Co-stabilizers		
Kieralon CD	BASF	polyacids, anionic and nonionic surfactants, APEO-free*
Securon 540	Henkel	organic phosphorus compounds
Activators for acid bleaching		
Actiron PHD	Protex	mixture of complexing agents
Prestogen SP	BASF	mixture of organic activators
Wetting agents/detergents and deaeration agents		
Arbyl BF	Grünau	alkanesulfonate with alkyl polyglycol ether
Cottoclarin OW	Henkel	combination of anionic and nonionic components
Erkantol CW	Bayer	phosphoric esters with emulsifiers
Hostapal SF	Hoechst	combination of surfactants
Kieralon JET-B konz.	BASF	defoamed nonionic surfactants, APEO-free*
Kieralon OLB konz.	BASF	combination of anionic and nonionic components, APEO-free*
Subitol L-SN	CHT	synergistic mixtures of surfactants
Tanawet PS	Sybron/Tanatex	nonionic and anionic surfactants

* APEO = Alkyl phenol ethoxolate.

Stabilizers contain magnesium ions and oxidation-stable costabilizers. The positively charged magnesium ions stabilize the perhydroxy anions, while the anions of the costabilizers (e.g., phosphonate ions) form complexes with the heavy metal ions, thereby inactivating their catalytic effect. Thus, direct stabilization is caused by the magnesium ion and indirect stabilization by the costabilizer. As sodium silicate inactivates heavy metal ions, its anion also has an indirect stabilizing action [170].

The stabilizers available commercially have various compositions. Apart from the magnesium ions responsible for the direct stabilization, there are complexing agents such as EDTA, DTPA (diethylenetriaminepentaacetic acid), gluconic acid, phosphonic acid, and poly(acrylic acid) derivatives. These so- called organic stabilizers do not contain sodium silicate. The silicate- containing stabilizers include sodium metasilicate in aqueous solution at a concentration of 38 – 40° Bé. Organic stabilizers containing surfactants are also marketed.

In modern steamers with reaction times up to 30 min, silicate-free bleaching in the presence of organic stabilizers has become established. Stabilization of bleaching liquors with sodium silicate and magnesium ions has the disadvantage that silicate encrustations (scaling) form in the bleaching equipment. These are difficult to remove, and damage the surface of the fabric. Silicate can also be deposited on the fabric, and this spoils the hand (feel) of the product and reduces its absorbency.

Stabilizers must have the following properties:

1. Good stabilizing action
2. Good resistance towards oxidizing agents
3. Prevention of silicate built-up on rollers in the steamer (scaling)
4. Inactivation of the catalysts
5. Good metering and pumping properties

Surfactant- containing stabilizers used with sodium silicate in the bleaching both must have silicate dispersing properties.

Surfactants used in bleaching, have emulsifying, dispersing, and wetting properties which promote the removal of hydrophobic impurities and soil, and assist in the transport of the reaction products formed by the bleaching process. The wetting properties are necessary to correct the absorbency of the pretreated goods and to make it uniform. To meet these requirements, the surfactants used are usually mixtures of anionic compounds such as alkyl sulfonates and alkyl aryl sulfonates with nonionic compounds such as alkylphenol ethoxylates or the biologically degradable fatty alcohol ethoxylates. The products must be stable in the bleaching bath, and must be suitable for metering equipment.

Bleaching under Weakly Alkaline to Weakly Acidic Conditions. Bleaching of cellulose fibers with hydrogen peroxide in the pH range of 6.5 – 8.0 is of some industrial importance. The process is used for goods dyed in the form of yarn when the dyes present change or bleed out in a normal alkaline bleach or are otherwise not bleachfast. The process is also used for blends of cotton with alkali-sensitive fibers such as wool. It is also carried out after bleaching with hypochlorite in order to save water in the subsequent washing process.

Whereas in alkaline bleaching the hydrogen peroxide is so strongly activated by the OH^- ions present that it must be stabilized, in the middle pH range activators are required to give bleaching activity [194].

Activators include acetylated amines, acetylated amides, mixed anhydrides of acetic acid with inorganic acids such as phosphoric [195] and phosphorous acids, and nitriles [196]. The bleaching process (oxidation) presumably does not proceed directly with the hydrogen peroxide molecule, but via an intermediate association compound. Because the pH of the bleaching bath tends to decrease, a buffer must be added.

4.2.2.4. *Bleaching with Sodium Hypochlorite*

In Europe, bleaching with chlorine compounds is in decline for ecological reasons. Outside Europe, it remains highly important as a cheap bleaching method and as a prebleach for cotton with a peroxide bleach to follow [197].

Damage to the cotton fibers is minimized if bleaching is carried out in the pH range 9.0 – 11.5 [198]. Precise control of pH and reaction time is therefore necessary. The bleaching process must be carried out at room temperature because of the high reactivity of the hypochlorite.

4.2.2.5. *Bleaching with Sodium Chlorite*

Chlorite bleaching for the pretreatment of cotton, linen, flax, jute, and other cellulosic fibers is carried out by the impregnation – steam process (e.g., pad-roll system) or long bath process. When used in conjunction with hydrogen peroxide bleaching, either before or after the chlorite bleaching, high reflectance is obtained. Textiles with a high catalyst content can be safely bleached, as heavy metal compounds are catalytically inactive in chlorite bleaching [199].

The chlorite method has declined in importance for ecological reasons and because of its severe corrosivity, but the main reason for its decline is the introduction of continuous open-width bleaching processes for cotton articles.

4.2.2.6. *Mercerization and Causticizing*

Mercerization is the treatment of cotton under tension with caustic soda solution at 15 – 25 °C for 25 – 40 s. The concentrations normally used in practice for woven materials are in the range 20 – 24 % NaOH (26 – 30° Bé), and for yarn, 24 – 26 % NaOH (30 – 32° Bé). Treatment with concentrated caustic soda destroys the spiral form of the cellulose with formation of alkali cellulose (alcoholate), which is changed to cellulose hydrate on washing out the caustic soda.

The process has the following beneficial effects on the textile [200]:

1. Improved ability to take up dyes
2. Improvement of the dyeability of dead cotton
3. Luster
4. Good dimensional stability of woven fabrics
5. Higher tear strength of yarns

In dry mercerization, the impregnation of the dry fabric can be done on the greige cloth or after any part process, e.g., desizing, alkaline boiling-off, or bleaching, followed by an intermediate drying [201].

In the Mercevic process, which is also referred to as dry mercerization, the process is carried out while drying the fabric on the stenter frame [202], [203]. In wet mercerization, intermediate drying is not carried out for cost reasons. Here, it is essential for effective mercerization that the water in the fabric should be displaced as completely as possible by the concentrated caustic solution.

For effective mercerization, complete penetration (wetting) of the cotton fibers by the concentrated caustic liquor must occur. This is hindered by the hydrophobic substances present on the gray goods, size or size residues, and by the high surface tension of the liquor. Effective mercerization therefore requires the use of wetting agents.

Hot impregnation followed by mercerization at room temperature or with cooling, in the presence of wetting agents, is also used (hot mercerization) [204-206].

In the causticizing process, the material is treated with less concentrated caustic liquor without applying tension. The material is allowed to shrink, whereby the dye absorption is improved. Cotton is treated with 13 – 16 % liquor (18 – 22° Bé), and viscose staple fiber with 3 – 5 % liquor. Causticizing of viscose staple fiber and polyester – viscose articles gives the material a wool-like hand. A combined caustic scour/causticizing process with 18 – 22° Bé NaOH under steam conditions is sometimes applied to cotton fabric. Mercerization wetting agents include C_4 – C_{12} sulfuric half-esters of

aliphatic alcohols, alkyl sulfonates from the sulfochlorination of alkanes, and ether alcohols such as butylene glycol monoethyl ether.

Alkanolamines and their mixtures with alkyl sulfonates and ether sulfates are also used. To reduce foaming phosphoric esters of C_4 – C_8 alcohols may be added. The wetting agents for treatment with 13 – 16 % caustic liquor are similar to those used for mercerization. Wetting agents used for the caustic treatment of viscose staple fiber include anionic surfactants or mixed anionic/nonionic wetting agents based on alkyl sulfonate/alkyphenol ethoxylate or alkyl sulfonate/fatty alcohol ethoxylate. These must be stable in the alkaline bath (i.e., they must not separate, forming a creamy layer at the surface). They must have a good wetting action and must not cause foaming. The first wetting agents used in the mercerization of cotton were phenol- and cresol- containing products which were mixed with amines, naphthenic acids, resin acids, and alkylnaphthyl sulfonates. The mercerization wetting agents used today are free of phenol or cresol.

The effectiveness of a mercerization wetting agent is assessed by measuring the shrinkage of a cotton yarn as a function of the wetting time.

Some auxiliaries for mercerization and causticizing are listed in Table 16.

Table 16. Some auxiliaries used in mercerization and caustic treatment [182]

Trade name	Producer	Compositon
Floranit 4028	Henkel	derivative of special alcohols
Invadin MC neu	Ciba	preparation from an alkanol sulfonate
Leonil GA	Hoechst	derivative of special alcohols
Leophen BN	BASF	alkylsulfonate with additives
Leophen MC	BASF	derivative of special alcohols
Mercerisin OR	Thor	alkyl sulfate
Mercerol QW	Sandoz	alkyl sulfate
Natron	Bozzetto	sulfated alcohols
Subitol MLF	CHT	sulfated alcohols
Sultafon MES 200	Stockhausen	combination of various sulfonates and sulfated fats

4.3. Preparation of Woolen Textiles (⟶ Wool)

The composition of raw wool varies widely, depending on the amount of impurities and foreign materials. On average, raw wool contains only ca. 50 % fiber. The approximate composition is as follows:

1. Wool wax (5 – 10 %), which is insoluble in water, consists of a mixture of fatty acid esters of higher alcohols, free fatty acids, and free higher alcohols
2. Wool grease (2 – 13 %), which is soluble in water, consists of lower carboxylic acids and aldehydes and inorganic constituents
3. Vegetable and mineral foreign materials (5 – 40 %)
4. Fibrous materials (proteins, 20 – 80 %)
5. Water (10 – 20 %)

Raw Wool Washing. The first cleaning process applied to the fleece after shearing is washing in the Leviathan wool washing machine. This continuous process removes the wool wax, wool grease, sand, soil, and other impurities from the loose fibrous material. The washing medium, which has a pH of 7 – 10, contains alkylphenol ethoxylates and fatty alcohol ethoxylates [207]. Soaps are still occasionally used.

Carbonization is the treatment of wool with acid to remove plant constituents such as burrs, straw, wood fragments, and seeds. In the treatment of spinning mill wastes, any cellulose fibers still present are also removed. The vegetable impurities are mainly decomposed by hydrolysis with sulfuric acid, and can then be easily removed by beating or brushing. If the process is correctly carried out, very little damage to the wool occurs. Carbonization with sulfuric acid consists of the following steps:

1. Impregnation of the fiber with carbonization liquor (3 – 6 % H_2SO_4)
2. Concentration of the carbonization chemicals on the wool fiber by heating to 70 – 90 °C (evaporation of some of the water)
3. “Burning” by dehydration and carbonization of the vegetable matter at 100 – 115 °C
4. Mechanical removal of the carbonized vegetable impurities
5. Neutralization and removal of the chemicals

The use of wetting agents (Table 17) [208-212] such as alkylnaphthalene sulfonates, alkylphenol ethoxylates, and alkyl polyglycol ethers leads to thorough wetting of the gray goods and reduces impregnation time. They also enable the concentration of the acid to be reduced and protect wool fibers [209].

Table 17. Some auxiliaries used in the carbonization and bleaching of wool [182]

Trade name	Producer	Composition
Carbonization auxiliaries		
Bura-Netzer	Baur & Gäbel	sulfonate mixture with nonionic surfactants
Lavotan DSU 100	CHT	fatty alcohol ethoxylates with sulfonates
Leophen RA	BASF	sulfosuccinic acid ester
Sultafon TK	Stockhausen	special ethylene oxide condensation product
Teca-Rapid S	Dr. Angele	alkane sulfonate mixture
Bleaching auxiliaries		
Blankit IN	BASF	combination of stabilized reducing agents
Blankolen K	Brüggemann	combination of stabilized reducing agents
Blankit D	BASF	combination of active reducing agents
Blankit IAN	BASF	combination of stabilized reducing agents with optical brighteners
Celidon LA	Bozzetto	phosphate derivatives
Kappazon B 84	Kapp-Chemie	alkali polyphosphate
Prestogen W fl.	BASF	salts of organic polyacids
Tubotex PCA fl.	CHT	inorganic salts, alkali buffers, and complexing agents

The treated material must be completely neutralized to avoid damage to the wool. This is achieved by having an appreciable excess of sodium carbonate in the neutralizing liquor. The surfactants present improve the wetting of the wool fiber [212].

As well as sulfuric acid, HCl gas and hydrochloric acid are used for carbonization, especially when recovering wool from rags [213]. In the United States, a mixture of aluminum chloride and hydrochloric acid at a treatment temperature of ca. 130 °C is preferred [214]. Aluminum chloride gives the best results in the carbonization of wool – synthetic fiber blends [215].

Bleaching Processes. The colors of washed wool range from cream to almost bleach (e.g., the strongly pigmented fibers of Karakul wool). All types of wool are bleached to give the whiteness necessary for dyeing with pastel tones or to enable the wool to be used for white articles. Bleaching agents are also often used during raw wool washing to brighten the color of the raw wool and to give it a more pleasing appearance.

Oxidative bleaching agents such as hydrogen peroxide and reductive bleaching agents such as sodium dithionite are used. To give a high degree of whiteness and improved lightfastness, two-stage oxidative and reductive bleaching is carried out in a conventional complete bleaching process. This bleaching method leads to high reflectance but is very time-consuming, so that ways of improving productivity are being sought. New bleaching agents can either give more rapid bleaching in existing equipment, or can enable the padding technique, a semicontinuous or continuous process, to be used. Also, several operations can be combined, leading to cost reduction. The following processes are used [216]:

1. Conventional wool bleaching in two stages: oxidative and reductive
2. Rapid acid bleaching with hydrogen peroxide in either a discontinuous or a continuous process
3. Reductive bleaching with stabilized sodium dithionite or thiourea dioxide, with or without whiteners
4. Reductive bleaching based on activated sulfinic acid derivatives, enabling fully and partly continuous bleaching with simultaneous whitening to be carried out
5. Oxidative and reductive bleaching in the dye-bath to produce clear color shades

Oxidative bleaching of wool with hydrogen peroxide is performed in the weakly alkaline pH range (8 – 9.5) below 50 °C because of the sensitivity of the wool fibers to alkali. These conditions lead to long reaction times. Sodium pyrophosphate is usually used as a stabilizer. Peracetic acid, produced in the bleaching bath from hydrogen peroxide and acetic acid, is used in the weakly acid range [217]. Pure peracetic acid is less effective in the bleaching bath than the combination H_2O_2 /CH_3COOH [218]. Activation of the hydrogen peroxide can also be achieved with formic acid or formaldehyde, which form performic acid [219], [220]. Peroxycarboxylic acids formed, e.g., by the addition of Prestogen W (BASF), have a similar effect.

Another special use of activated acid solutions of hydrogen peroxide is in the mordanting of wool and the catalytic mordanting of animal hair. Here, activators that cause decomposition of the H_2O_2 include metallic salts and modified metallic salts of iron, cobalt, nickel, manganese, and copper [221-223].

The products used in the reductive bleaching of wool are mainly based on sodium dithionite, though some have thiourea dioxide as their main component. All reductive bleaching agents contain stabilizers (Table 17) to prevent the reducing agent from rapidly decomposing, a reaction which is catalyzed by the acidic degradation products of the bleaching process.

The pH is usually kept within the range 7 – 9 at the start, and 7 – 5 at the end of the process. Phosphates are used for

buffering. Complexing agents are used to combine with calcium and iron ions and prevent the formation of precipitates (carbonate hardness). The bleaching time is determined by the consumption of reducing agent, which is affected by the amount of air dispersed in the circulating liquor [222].

When wetting agents and detergents are required, surfactants based on alkylphenol ethoxylates or fatty alcohol ethoxylates, sometimes mixed with alkyl sulfonates or alkylaryl sulfonates, are used.

4.4. Ecological Aspects of Fabric Pretreatment [224-231]

The new laws and regulations for the protection of the environment that are being applied in most industrialized countries are having a considerable impact on the textile industry.

In Germany, according to the dangerous substances regulations covering the discharge of wastewater into open sewage installations, a license is required if the amounts and concentrations exceed prescribed threshold values (Tables 18-20). According to the law on detergents, wetting agents and detergents are deemed to include sizing agents, boiling-off auxiliaries, bleaching stabilizers, complexing agents, softeners, and finishing materials. All these products must be biologically degradable or capable of being biologically eliminated. Since 1989, the TEGEWA Association of textile companies has voluntarily abandoned the use of APEO- containing auxiliaries, as the biological breakdown of these materials produces metabolites that are toxic to fish.

Table 18. Threshold values for impurities in wastewater

AOX*	3 mg/L	Cr	2 mg/L
VHC**	1 mg/L	Cr^{6+}	0.5 mg/L
Hydrocarbons	50 mg/L	Zn	10 mg/L
Cu	2 mg/L	Sn	10 mg/L
Ni	2 mg/L		

* The AOX threshold value for bleaching with chlorine and antifelting finishing of wool is 8 mg/L (until 31.12.96).
** VHC = volatile halocarbons

Table 19. Threshold values attainable by current technology

COD	160 mg/L	ammoniacal nitrogen	10 mg/L
BOD_5	25 mg/L	iron	3 mg/L
Phosphorus total	2 mg/L	aluminum	3 mg/L

Table 20. Threshold values attainable by current technology for direct and indirect dischargers

T_F (Dilution factor) 2			
AOX	0.5 mg/L	Cu	0.5 mg/L
VHC*	0.1 mg/L	Cr^{6+}	0.5 mg/L
Free chlorine	0.3 mg/L	Ni	0.5 mg/L
Hydrocarbons	15 mg/L	Pb	0.5 mg/L
Sulfide	1 mg/L	Zn	2 mg/L
Sulfite	1 mg/L	Sn	2 mg/L

* VHC = volatile halocarbons

[Top of Page]

5. Textile Dyeing Auxiliaries

Eva-Marie Borschel, Sigismund Heimann, Erich Kromm

Introduction. The dyeing of textiles requires the use not only of dyes and chemicals, but also a number of special products known as dyeing auxiliaries [237]. These materials may constitute an integral part of the dyeing process (e.g., reducing agents for dyeing with vat dyes) or can lead to technical improvements such as more (level) dyeing, better fastness, etc. [238].

The following classification and definitions of dyeing auxiliaries are taken from the TEGEWA brochures [239]. Classification is based on product use. The descriptions of the product groups indicate whether they are suitable for the exhaustion dyeing process (batchwise long-liquor dyeing) or for the padding process (⟶ Textile Dyeing), although many dyeing auxiliaries can be used in either process, including dye solubilizing and hydrotropic agents, dye wetting agents,

aftertreatment agents, and reducing agents. Some special padding auxiliaries have been placed in a category of their own (Section Auxiliaries for Continuous Dyeing). Levelling agents, carriers, and anticreasing agents are only used in the exhaustion dyeing process.

The dyeing auxiliaries form a very heterogeneous group of chemicals. Though many of them are surfactants, there are many other types, including inorganic compounds, water-soluble polymers and oligomers, polymer dispersions, and solubilizing agents. Most commercial dyeing auxiliaries are preparations containing several components. The individual chemical compounds, usually along with relevant environmental data, are described under separate keywords (e.g., ⟶ Polyacrylates, ⟶ Polyacrylamides and Poly(Acrylic Acids)), ⟶ Reduction, ⟶ Sulfites, Thiosulfates, and Dithionites, ⟶ Solvents, ⟶ Surfactants).

Consumption of dyeing auxiliaries is estimated at 60 – 70 % of the consumption of the dyes themselves. In 1992 this corresponded to ca. 280 000 t/a dyeing auxiliaries for a world consumption of 440 000 t/a dyes used for dyeing textiles [240].

5.1. Dye Solubilizing and Hydrotropic Agents

To dissolve large amounts of a dye in a small amount of water, solubilizing and hydrotropic agents are used to increase its solubility. This is mainly necessary when dye baths are used in padding processes, dye stock solutions, and printing pastes (see Section Dye Solubilizing and Dispersing Agents). The effectiveness of hydrotropic agents is due to their amphiphilic properties. Intermolecular cohesive forces lead to the formation of molecular associates with the substances to be dissolved, and these are water-soluble due to the presence of hydrophilic groups [241]. In this respect, the hydrotropic agents resemble surfactants, which also have a solubilizing effect. A difference is that surfactants can increase the solubility of sparingly soluble substances at relatively low concentrations by absorption into micelles [232].

Solvents are used in dyeing and printing to wash dye residues from equipment and apparatus. Many auxiliaries used in continuous dyeing contain solvents, hydrotropic agents, and surfactants, not only because of their ability to solubilize dyes, but also to improve the fixing process, produce of fiber swelling, and to provide wetting and foaming properties, etc. (see Section Auxiliaries for Continuous Dyeing).

The commercial products supplied for dissolving dyes often contain mixtures of solvents, dispersants, and surfactants. Solvents and hydrotropic agents are needed when dyeing with the following classes of dyes:

Developing Dyes. To dissolve naphthols by the cold dissolution process (⟶ Textile Dyeing), ethanol or certain heterocyclic bases (pyridine derivates, *N*-methylpyrrolidone) are used as solvent.

Phthalogen Dyes. To dissolve the sparingly soluble aminoisoindolines and other precursors of these dyes, special mixtures of glycols, amines, and surfactants are used (⟶ Textile Dyeing – Dyeing with Phthalogen Dyes).

Reactive Dyes. In the continuous dyeing of cellulose fibers with reactive dyes by the pad dry thermofix process (⟶ Textile Dyeing – Dyeing with Reactive Dyes), large amounts of urea (50 – 100 g/L) are added to the liquor. These increase the solubility of the dye in the pad-dyeing liquor and, during the fixing process, form a melt which allows the dye to diffuse into the fiber. In many cases, this improves the dye yield [242]. Dicyandiamide, which has a lower tendency to sublime, can be used instead of urea.

Acid dyes and metal-complex dyes often have very low solubility in water, and dye solubilizing agents must be added if they are to be used in pad liquors or printing pastes for wool or polyamide, especially with dark colors. Suitable materials include ethanol, propanol, di- and triglycol, thiodiglycol, various glycol ethers, and nonionic surfactants.

5.2. Dye Protecting Agents and Boildown Protecting Agents

Under unfavorable conditions, certain dyes can be changed or destroyed during application. In these cases, special protecting agents are added to the dye baths. Boildown effects occur when dyeing wool and wool – cellulose fiber blends, meaning that the azo dyes used are reductively decomposed by breakdown products from the wool proteins, or undergo changes in shade [243]. Dye boildown can be prevented by adding buffering agents and special auxiliaries. Azo dyes can also be reduced when dyeing cellulose fibers with direct dyes at boiling temperatures. These defects can be prevented by precise control of the pH by the addition of buffers and oxidizing agents. Similar defects can occur on dyeing with reactive dyes, which can be modified through reduction by cellulose degradation in the presence of alkali at elevated temperatures. The dye is protected by the mild oxidizing agent sodium 3-nitrobenzenesulfonate [244]. This product is also used to protect the dye from reduction when boiling off before bleaching colored piece goods.

When dyeing bleached cotton with reactive dyes, which are very sensitive to oxidation and reduction, even very small amounts of residual hydrogen peroxide must be destroyed by weak reducing agents. Vat dyes of the indanthrone type tend to give cloudy and green tints at dyeing temperatures above 60 °C owing to over-reduction. This effect can be prevented by adding sodium nitrite or glucose [245]. Some commercial products are listed in Table 21.

Table 21. Some commercial dye protecting agents [236]

Name	Producer	Uses	Composition
Buralit N	Baur & Gäbel, Germany	prevention of boildown of substantive dyes	combination of ammonium salt with aliphatic and aromatic sulfonic acids
Degal SBN	Sigma, Germany	antireduction agents	sodium nitroarylsulfonate
Dupranin	Thor-Chemie,	reducing agent for single	inorganic salt

RKM conc.	Germany	bath bleaching and dyeing with reactive dyes	
Intratex BD	Crompton & Knowles, United Kingdom	prevents reduction on boiling off, dyeing, and printing	sodium salt of an aromatic nitrosulfonic acid
Ludigol	BASF, Germany	oxidizing agent	sodium 3-nitrobenzenesulfonate
Matexil BA-PK	ICI, United Kingdom	for removal of hydrogen peroxide on dyeing with reactive dyes	mixture of inorganic salts
Matexil PA-L	ICI, United Kingdom	weak oxidizing agent	sodium 3-nitrobenzenesulfonate
Meropan SO2	CHT, Germany	prevention of boildown of direct and reactive dyes	benzenesulfonic acid derivative
Revatol S/SP Powder	Sandoz, Switzerland	prevention of reduction on boiling off, dyeing, and printing	sodium 3-nitrobenezenesulfonate
Sustilan N	Bayer, Germany	fiber protecting agent for wool and boildown protecting agents	combination of ammonium salts, aliphatic and aliphatic aromatic sulfonic acids

5.3. Wetting and Deaerating Agents

A fundamental prerequisite for successful dyeing with an aqueous dye bath is complete wetting of the textile. Only then is uniform and level dyeing possible. This is achieved by means of wetting and deaerating agents [246], whose use depends on the dyeing process and the nature and condition of the textile [232].

Exhaustion Dyeing Process (Batch Dyeing). Textiles made from natural fibers such as cotton and wool generally have good wetting properties after pretreatment such as washing, boiling-off, or bleaching. However, it may be economically advantageous to carry out pretreatment and dyeing in a single step. This can be achieved by the use of wetting agents and protective colloids (see Section Dispersants for Dyeing (Batch and Continuous Processes)). Wetting agents are not necessary when batch dyeing synthetic fibers. However, when dyeing wound packages (cross-wound bobbins or warp beams) a deaeration step is required at the beginning of the dyeing process because of the high density and thickness of the textile, irrespective of the type of fiber and its pretreatment. This prevents mechanical deformation of the wound package by the included and absorbed air when the dye liquor is introduced under pressure. The deaeration of wound packages of pretreated natural fibers and of synthetic fibers can usually be carried out by gently displacing the air in a preliminary wetting stage, accelerated by using a deaerating agent. With raw cotton, which is difficult to wet, the use of a deaerating agent is essential.

Continuous Processes (see Section Auxiliaries for Continuous Dyeing). Continuous dyeing processes are characterized by very short contact times between the pad liquor and the textile when applying the dyes and chemicals.

Inadequate, nonuniform wetting leads to reduced uptake of the dye and impairment of the fabric appearance. Wetting agents must therefore always be added to the impregnation bath when gray cloth, inadequately pretreated textile materials, or any material that is difficult to wet is dyed by the pad dyeing process [232], [234], [247].

The very short contact times in the pad dyeing process therefore require the use of products that cause very fast wetting (rapid wetting agents). Not only is the liquor caused to penetrate the interior of the textile material (capillaries, fiber crossing points) rapidly, but also a continuous film of dye liquor is formed on the surface of the textile. Such wetting agents must maintain their effectiveness not only in neutral solutions, but also in baths containing alkalis and electrolytes, and must ensure that as little foaming as possible occurs in the baths even for high throughputs [247], [248]. The stability of the pad liquor must not be affected, as even small amounts of slight creaming and dye precipitation on the cloth can lead to spots and marks on the textile. Dye wetting agents differ from pretreatment products as follows:

1. Exhaustion dyeing process: low foam formation, no dye-retention effect
2. Continuous process: no fiber affinity, good wetting action

Wetting and deaerating agents are surface-active substances. They cause spreading of the aqueous liquor over the surface of the fibers, thereby displacing the air [249]. In the usual type of surfactant wetting agents, which promote spreading by reducing the surface tension of the liquor, this is associated with foam formation [250], which causes severe problems, especially during the removal of air from wound packages. Wetting with the formation of little or no foam can be achieved in two ways. Surfactant mixtures for reducing the surface tension of the liquor can be used in combination with antifoams, or the spreading of the liquor is achieved by increasing the surface tension of the fibers by adsorption of polar, hydrophobic compounds [251], which do not form stable foam lamellae. Wetting agents for exhaustion and continuous processes usually contain the following surfactants: alkyl-or alkylaryl sulfates, alkylaryl ethoxylates, polyglycol ether sulfates. Rapid wetting agents contain a branched hydrophobic molecular component (e.g., secondary alkylsulfonates, isooctyl esters of sulfosuccinic acid, and dialkylsulfonimides). Deaerating products are emulsions of hydrophobic phoshoric triesters and mixtures of phosphoric acid partial esters, and mixtures of surfactants with antifoams. Some commercial products are listed in Table 22.

Table 22. Some commercially available wetting and deaeration agents for dyeing [236]

Name	Producer	Uses	Composition
Albegal FFA	Ciba-Geigy [a]	deaeration agent	alkylaryl ethoxylates + polysiloxane
Arbyl ASN	Grünau [b]	rapid wetting agent	diisooctyl sulfosuccinate
Erkantol CW	Bayer [b]	deaeration agent	phosphoric esters + emulsifiers
Irgapadol FFU	Ciba-Geigy [a]	pad wetting agent	alkylaryl polyglycol ether sulfate + poly (ethylene glycol)
Leophen M	BASF [b]	deaeration agent	phosphoric acid esters + emulsifiers
Leophen RBD	BASF [a]	rapid wetting agent	alkylsulfonimide
Perenin 4649	Dr. Th. Böhme [b]	pad wetting agent	phosphoric esters
Periwet SL	Dr. Petry [b]	deaertion agent	fatty alcohol ether sulfate
Primasol NF	BASF [b]	pad wetting agent	phosphoric esters
Rewopol SBD 075	Rewo [b]	rapid wetting agent	diisooctyl sulfosuccinate
Subitol SAN Neu	CHT [b]	wetting agent	phosphoric esters
Tanawet PAD	Sybron/Tanatex [c]	pad wetting agent	phosphoric esters

[a] Switzerland;
[b] Germany;
[c] The Netherlands.

5.4. Dispersants and Protective Colloids

Insoluble dyes applied in the form of aqueous dispersions are used in a large number of dyeing and printing processes. Dispersants are required to produce the dye preparations for these processes and to stabilize the finely dispersed state during application. Powdered dispersion and vat dyes contain 50 – 80 % of these products, and dyes prepared in liquid form 10 – 30 %. To maintain the stability of the dispersion throughout the dyeing or printing process, additional dispersant is added to the dye bath. Also, during the dyeing of plant fibers (e.g., raw cotton) which have not been pretreated, considerable amounts of impurities in the form of waxes, pectinates, and water hardness are introduced into the dye baths. The dispersants must also prevent these from being precipitated.

Dispersants used in textile finishing can be divided into two classes:

1. Water-soluble oligo- and polyelectrolytes (“protective colloids”) [235, p. 459]
2. Surfactants

Both have an amphiphilic structure with hydrophilic and hydrophobic components. The most important examples of these compounds are the polysulfonates, polyacrylates, polyvinylsulfonates, and anionic and nonionic surfactants. The activity of these substances is based on the formation of electrostatic and mechanical protective films around the dispersed particles which prevent precipitation and agglomeration.

Specific dispersants are available for dyes, plant fiber impurities, and precipitates due to water hardness.

Dispersants for Dyes. Condensation products of aromatic sulfonic acids with formaldehyde and lignosulfonates are widely used. Anionic and nonionic surfactants are used for special applications.

Formaldehyde Condensation Products. Condensation products of β-naphthalenesulfonic acid with formaldehyde [253] have been known since 1913 [252], and are still used in dye production and dyeing. In addition to these, condensation products of phenol with formaldehyde and sodium sulfite (sulfomethylation products) are now also used as dispersing agents. The degree of condensation depends on the conditions of the reaction with formaldehyde, and can give products with 2 – 10 aromatic nuclei. However, the structure of this, the most common dispersing agent, has not yet been fully elucidated. Conventional formaldehyde condensation products and the lignosulfonates described below are only ca. 30 % degraded in biological wastewater treatment plants (test method OECD 302 B). However, a formaldehyde condensation product which can be > 70 % eliminated from the wastewater has been commercially available since 1993 [254]. This product is used both for finishing Palanil and indanthrene dyes and as a dyeing auxiliary (see Section Dispersants for Dyeing (Batch and Continuous Processes)) under the trade name Setamol E.

Lignosulfonates. Products based on natural lignins form a further large group of disperants for finishing and as auxiliaries in the dye bath. They are produced from sulfite pulping liquor or from alkali lignins produced in the kraft pulping process [255]. The molecular mass of industrial lignosulfonates, whose structure is still not completely elucidated, lies between 2000 and 100 000 [256]. The basic building block is phenylpropane, from which a large number of derivatives can be

obtained. The suitability of the lignosulfonates for the various applications depends very much on their purity, degree of sulfonation, number of hydroxyl groups, and molecular mass distribution. The desired properties are obtained by the purification and chemical treatment of the crude products [257]. Some undesirable properties of the lignosulfonates include a reducing action on sensitive azo dyes during high temperature dyeing (130 °C) and possible soiling of the fiber due to the dark color of these products.

Other Compounds. Other natural and synthetic high molecular mass compounds are used, especially in the production of pigments [258]. These include polysaccharides, alginates, cellulose derivates, polyacrylates and polyvinyl compounds.

Dispersants for cotton impurities and precipitates from water hardness are polycarboxylates [259][260][261], which are homopolymers and copolymers of acrylic and maleic acids with relative molecular masses of 1000 – 10 000. With higher molecular mass, the protective colloid effect becomes a flocculating action [259]. These substances have wide application in many processes for dyeing cellulose fibers, as they do not cause foaming or dye retention.

5.4.1. *Dispersants in Dye Production*

Vat and disperse dyes must be finely dispersed in the dye bath (⟶ Textile Dyeing – Dyeing of Cellulose Fibers, ⟶ Textile Dyeing – Dyeing of Polyester Fibers). This is achieved by a special treatment (finishing) which follows the synthesis of the dye. Finishing consists of precipitation, filtration, grinding in the presence of dispersants, and drying. Dye finishing has a major influence on the dyeing process. Formaldehyde condensation products and lignosulfonates are widely used in dye finishing [262]. They give improved particle size reduction on grinding, good redispersibility of the finished dye, low staining of the textile fibers by the dispersant, and good solubility and temperature stability of the dispersion. To give improved compatibility with synthetic thickeners for printing, dyes for the printing process are often finished with nonionic dispersants.

5.4.2. *Dispersants for Dyeing (Batch and Continuous Processes)*

Batch Dyeing Processes with Disperse Dyes. Disperse dyes have very low solubility in water, and exist as suspensions or emulsions even at the high temperatures of the PES (polyester) dyeing process (120 – 140 °C). They provide the reservoir for a saturated dye solution from which molecularly dispersed dye is absorbed by the fiber [263]. Since fineness of this dispersion and its stability during the entire dyeing process are of crucial importance for dye deposition, that dispersants must be used. They prevent a filtering effect of agglomerated dye in package dyeing, and hence prevent unlevel results, spots, and unsatisfactory dye yield and dyefastness. Although all disperse dyes already have a high content of dispersants, 0.5 – 2 g/L dispersant is usually added to the dye liquor if the thermal and mechanical stress on the dye bath is very intense. Formaldehyde condensation products and similar compounds, nonionic surfactants, and anionic surface active compounds can be used. As the nonionic surfactants often have a levelling effect as well as a dispersing effect, auxiliaries for dyeing polyester fibers often consist of mixtures of various compounds, and are used both as dispersing agents and levelling agents.

When dyeing secondary acetate fibers and polyamide fibers with disperse dyes, the sulfoaromatic condensation products described above or nonionic surfactants are used as dispersants.

In the aftertreatment of dyed polyester fibers, dispersants help to remove the dye adhering to the surface in an additional, usually reductive washing stage, thereby improving the fastness of the dyeing. Nonionic surfactants in amounts of 0.5 – 1 g/L are usually added as dispersants and washing agents to the final washing bath.

Batch Dyeing Processes with Vat and Sulfur Dyes. Vat and sulfur dyes are converted by reduction into the water-soluble, oxidation-sensitive leuco form which is adsorbed by to the cellulose fiber.

Dispersants are therefore only necessary for process steps in which the pigment has not yet been reduced or has been re-formed by oxidation:

1. In the preimpregnation process (⟶ Textile Dyeing, ⟶ Textile Dyeing), dispersants prevent destabilization of the dye dispersion.
2. When dyeing in open equipment (jiggers, hank dyeing machines), pigment is continuously formed by atmospheric oxygen. Dispersants prevent agglomeration, and thereby enable rapid revatting to be carried out.
3. In the oxidation stage at the end of the dyeing process, a large amount of pigment is formed in the dye liquor, especially with dark colors. Dispersing agents prevent the agglomeration of these pigment particles.

Dispersing agents used include naphthalenesulfonic acid – formaldehyde condensates, lignosulfonates, sulfonated oils, and protein – fatty acid condensation products, which are used in concentrations of 0.5 – 2 g/L [264], [265].

When dyeing raw cotton, the following auxiliaries are also used in concentrations of 1 – 2 g/L to prevent the agglomeration of impurities and of precipitates due to water hardness: polycarboxylates as dispersing agents and moderately strong complexing agents, and pure complexing agents (see Section Complexing Agents) such as phosphonates, EDTA (ethylenediaminetetraacetic acid), and NTA (nitrilotriacetic acid).

Batch Dyeing Processes with Other Classes of Dye. When dyeing with developing dyes, dispersants are used during dissolution of the naphthols and preparation of the primary and development baths [266]. Suitable products for the primary baths include sulfonated oils, lignosulfonates, protein – fatty acid condensation products, and condensation products of aromatic sulfonic acids. Fatty alcohol ethoxylate is added to the development baths in amounts of 0.5 – 2 g/L to enable the excess insoluble azo pigments to be easily washed out.

For the same reason, nonionic dispersing agents are added to the development baths when dyeing with leuco esters of vat dyes.

When dyeing raw cotton with water-soluble reactive dyes, dispersants are only needed for the impurities in the raw cotton and the precipitates due to water hardness. Polycarboxylates and polyphosphates are suitable for this.

Continuous Dyeing with Disperse and Vat Dyes. In continuous dyeing with insoluble dyes, a dispersing agent is added to the pad liquor to ensure stability of the dye dispersion. This gives a uniform fabric appearance without danger of marks due to agglomerates. In chemical or development baths, dispersants facilitate the redispersion of unfixed dye during final washing. This gives improved rubbing fastness. Almost all dispersing agents can be used in pad baths. Some commercial dispersants and protective colloids are listed in Table 23.

Table 23. Some commercial dispersants and protective colloids [236]

Name	Producer	Country
Condensed aromatic sulfonates		
Anogal DA	Baur & Gäbel	Germany
Avolan IS	Bayer	Germany
Dispersogen P liq.	Hoechst	Germany
Irgasol DAM	Ciba-Geigy	Switzerland
Matexil DA-N	Zeneca	United Kingdom
Setamol E	BASF	Germany
Setamol WS	BASF	Germany
Ligninsulfonates		
Dispergal CN	Impocolor	Germany
Lamepon N	Grünau	Germany
Reax product series	Westvaco	United States
Ultrazine	Borregaards	Norway
Serlasol	Serlachius Cop.	Finland
Polycarboxylic acids and mixed products		
Alcosperse AD	Allied Colloids	United Kingdom
Dekol SN	BASF	Germany
Dispergal PS	Impocolor	Germany
Irgasol CO	Ciba-Geigy	Switzerland
Masquol A 340 N	Protex	France
Meropan VD	CHT	Germany
Sandopur RSK	Sandoz	Switzerland
Protein condensates		
Adipon TH conc.	Henkel	Germany
Lamepon BN	Grünau	Germany
Ofnapon ASN	Hoechst	Germany
Sulfonated oils		
Geipolan EB	Geissler	Germany
Monopolbrillant Oil	Stockhausen	Germany
Sandozol KB	Sandoz	Switzerland
Ethoxylated products		
Dispersogen ASN	Hoechst	Germany
Lyocol ACN	Sandoz	Switzerland
Matexil DN-VL	Zeneca	United Kingdom
Ruco Levelling Agent EN	Rudolf-Chemie	Germany
Sarabid 200 LL	CHT	Germany
Uniperol EL	BASF	Germany

5.4.3. *Testing*

Whereas improvement of the fineness of the dispersion is the main criterion when testing dispersants for use in dye production, the property that is tested when the dyes are used in a dye bath is their ability to prevent agglomeration (stabilizing action).

The effectiveness of dispersants is characterized by their effect on dye dispersions, impurities, and precipitates from hard water, whereby a large number of testing methods are used [267-273]:

1. Determination of the particle size distribution of dye dispersions
2. Determination of the filtering properties of dye dispersions
3. Determination of the increase in the solubility of dyes

4. Determination of the antiflocculation effect on accumulations of insoluble dyes and impurities (waxes, pectinates)
5. Calcium carbonate dispersing capacity (CCDC) [274]
6. Tests of performance in use, e.g., laboratory dyeing tests or flow tests

5.5. Complexing Agents

Water quality is of great importance for the success of the dyeing process. Insoluble impurities, hardness components, and heavy metal salts can cause considerable problems. The presence of alkaline earth and/or heavy metal salts lead to the following problems during dyeing [275], [276]:

1. The formation of sparingly soluble saltlike compounds with anionic dyes leads to filtering-out problems in package dyeing, levelling problems, and impairment of the resistance to rubbing and washing
2. The formation of stable complexes with dye molecules causes change of the shade, often accompanied by loss of brilliance

Purified and softened water is therefore used in textile finishing. Complexing agents are added to the dye bath to combine with multivalent cations, especially the calcium, magnesium, and iron salts which are carried into the dye liquor by the textile material (e.g., cotton [277], [278]). The use of these products is also extremely important for washing out unfixed dye after completion of the dyeing stage (see Section Aftertreatment Agents to Improve Fastness) [279], [280]. In general the products used are the same as those used in the washing and cleaning processes of textile pretreatment [232, p. 995], [281].

The use of ion exchangers is not possible as these are by definition insoluble in the treatment solution.

Complexing agents differ with respect to the stability of the metal complex and the specific effect on metal cations. Also, for all products, the stability of the complex is strongly dependent on the pH [282].

EDTA (ethylenediaminetetraacetate; ⟶ Ethylenediaminetetraacetic Acid and Related Chelating Agents), DTPA (diethylenetriaminepentaacetate), NTA (nitrilotriacetate; ⟶ Nitrilotriacetic Acid), and derived phosphonic acids are very effective on a wide range of cations including heavy metal ions and those from hard water.

Specific products to combat the effects of water hardness include mild complexing agents such as polyphosphates and various polycarboxylic acids. These also have a dispersing action on precipitates from water hardness (see Section Dispersants and Protective Colloids) [274], which the strong complexing agents referred to above do not have.

Specific mild complexing agents for the heavy metal ions of iron, copper, and manganese include various polyhydroxy compounds, e.g., sorbitol, gluconic acid, glucoheptanoic acid, and alkanolamines.

Commercial products often also contain mixtures of various compounds.

With metal complex dyes, the use of weak complexing agents is especially preferable, so that the metal is not removed from the dye molecule. Strong complexing agents are suitable for stripping these dyes, as the metal-free dye is less strongly bound to the fiber.

Detailed information on complexing agents and their use in dyeing can be found in the literature [275], [276], [282]. A list of commercial products is given in Table 24.

Table 24. Some commercial complexing agents [236]

Name	Producer	Composition
Actiron GHN 30	Protex [a]	gluconate derivative
Aquamollin BC	Hoechst [b]	EDTA
Calgon T neu	BK Ladenburg [b]	polyphosphate
Dekol SN	BASF [b]	polycarboxylate
Delinol VB	Dr. Th. Böhme [b]	polycarboxylate
Heptol NWS	CHT [b]	polycarboxylate + polyphosphate
Heptol FEW	CHT [b]	polyhydroxycarboxylic acid
Ladiquest P 97	BK Ladenburg [b]	polycarboxylate + phosphonate + polyphosphate
Lamepon KB	Grünau [b]	organic phosphorus compound
Masquol EL 40	Protex [a]	EDTA
Periquest AL	Dr. Petry [b]	polycarboxylate
Plexene UL	Sybron/Tanatex [d]	polycarboxylate
Tetralon A	Allied Colloids [a]	EDTA
Trilon FE	BASF [b]	mixed complexing agents
Trilon TA	BASF [b]	NTA

Trilon TB	BASF [b]	EDTA
Verolan NBI	Rudolf-Chemie [b]	polycarboxylate

[a] France;
[b] Germany;
[c] The Netherlands;
[d] United Kingdom.

5.6. Levelling Agents

Levelling agents promote uniform distribution of the dye in the textile in the exhaustion dyeing process, so that the dyeing is level, with a uniform shade and depth of color.

Inequalities are caused or intensified by [283]:

1. High and variable affinity of the fibers for the dye
2. Imperfect distribution of the liquor in the textile
3. Temperature differences in the textile
4. High and variable affinity of the dye for the fibers

Unsatisfactory levelling can be prevented by dyeing techniques (e.g., improvement of the diffusion of the liquor within the textile, pH control) and by levelling agents.

Levelling agents act mainly by reducing the dyeing rate, increasing the rate of migration of the dye within the textile, and improving the compatibility of dyes.

Levelling agents also have other effects which do not directly influence the dye – fiber interaction, but nevertheless have a favorable effect on dye levelling. These include improvement of the solubility or dispersion stability of the dye, and prevention of deposition of impurities. Levelling agents often exhibit several effects simultaneously, sometimes because the individual components themselves have these properties, and sometimes because mixed products with complex effects are deliberately produced. Levelling agents that directly influence the dyeing process can be divided into products with an affinity for dyes, and products with an affinity for fibers [284].

Products with an affinity for dyes form loosely bound addition compounds with the dyes whose stability is concentration dependent and usually decreases with increasing temperature. Hence, the dye distribution equilibrium between the dye liquor and the fiber is shifted towards the dye liquor. The increased dye concentration in the liquor enables areas of the textile that are dyed to different intensities to be levelled by dye migration.

Effective levelling agents have an affinity for the dye, sufficient to reduce its rate of absorption and/or to promote its migration. Also, differences in the absorption behavior of different dyes can be equalized, so that the dyes are taken up from mixtures at equal rates. However, if the interaction is too selective, this leads to variations in color shade.

There is a series of products with increasingly high dye affinity, so retaining increasing amounts of dye in the liquor. Some of these can be used as stripping agents or for brightening dyed materials.

Auxiliaries with an affinity for dyes are also used for levelling already dyed material.

Levelling agents with an affinity for fibers are absorbed onto the fibers in competition with the dye. This competitive reaction reduces the absorption rate of the dye and promotes migration.

Products with fiber affinity may also be used in some cases (polyamide, wool), to even out differences in the dye affinity of the fiber. The most important levelling agents for the most important fibers and dyes are listed in Table 25.

Table 25. Types of levelling agent

Type of fiber	Dye	Composition	Effect
Cellulose	vat dyes	nonionic surfactants	affinity to dye
		polyamide amines	affinity to dye
		polyvinylpyrrolidone	affinity to dye
	direct dyes	nonionic surfactants	affinity to dye
		polyvinylpyrrolidone	affinity to dye
Wool	acid, metal complex, and reactive dyes	nonionic surfactants	affinity to dye
		weakly cationic ethoxylated products	affinity to dye and fiber
		anionic surfactants	affinity to fiber
Polyamide	acid and metal complex dyes	nonionic surfactants	affinity to dye

		cationic surfactants	affinity to dye
		ethoxylated products	affinity to dye
		anionic surfactants	affinity to fiber
	disperse dyes	nonionic surfactants	affinity to dye
Polyester	disperse dyes	carriers, special products	affinity to fiber
		nonionic surfactants	affinity to dye
Polyacrylonitrile	cationic dyes	cationic surfactants and nonsurfactants	affinity to fiber
		anionic surfactants	affinity to dye

5.6.1. *Levelling Agents for Dyeing Cellulose Fibers*

Cellulose fibers are dyed with reactive dyes, vat dyes, leuco esters of vat dyes, direct dyes, sulfur dyes, developing dyes, and pigments (⟶ Textile Dyeing – Dyeing of Cellulose Fibers).

Levelling agents are mainly required for vat dyes, and to a lesser extent for direct dyes. Dyeing auxiliaries that improve levelling are also used for dyeing with the other types of dye, but their effectiveness depends not on directly influencing the dyeing process, but on secondary effects such as wetting and dispersing, protective colloid formation, and increase of solubility.

Levelling Agents for Vat Dyes. Because of the high affinity of the reduced leuco form of many vat dyes for the cellulose fiber, the textile is dyed very rapidly and unevenly. Evenly dyed products can be produced by:

Controlling the *concentration of the leuco form*.

The leuco form of the vat dye is not yet formed at the start of the dyeing process (exception: leuco process), but is produced:

1. In the course of the dyeing process (semipigmentation process)
2. After a phase in which the vat dye pigment (low affinity for the fiber) is distributed within the textile (prepigmentation process)
3. By controlled metering [285], [286] of the vat dye to the reducing liquor
4. By controlled metering of the reducing agent to the dispersed vat dye pigment

The dyeing equilibrium is shifted in favor of the liquor by dyeing at a high temperature (80 – 110 °C). The increased dye concentration in the liquor favors levelling.

Levelling agents with an affinity for dyes compete for the cellulose fibers with the leuco form of the vat dye. This retards the absorption process and improves levelling.

The use of levelling agents can also be combined with modification of the process.

The equilibria of the dyeing process and the effect of an auxiliary with fiber affinity are illustrated schematically in Figure 11. The interaction between the levelling agent and the leuco dye is believed to involve the formation of an adduct, which is indicated by the characteristic displacement of the bands in the absorption spectra of the dye [287].

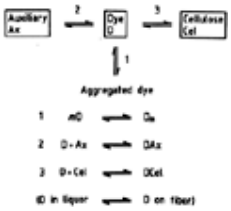

Figure 11. Equilibria involving a dissolved dye in a system with cellulose

The substances with an affinity for the dye that are most suitable for vat dyeing are stable in the strongly alkaline dye liquor. As foam formation seriously affects the dyeing process in modern dyeing equipment, the substances are classified according to their surface active properties [288]:

Surface Active Compounds

a. Fatty alcohol ethoxylates: mild levelling effect, strong bond to the dye (stripping agent)
b. Fatty amine ethoxylates: mild levelling effect, strong bond to the dye (stripping agent)
c. Alkylbenzimidazolesulfonates: moderate levelling effect, dispersing effect

Nonsurface Active Compounds

a. Betaines (*N*-pyridiniumacetic acid): levelling effect, low dye-bonding effect
b. Polyalkyleneamines

c. Oligomeric aminoamides: good levelling and absorption-retarding properties with low, unselective dye-bonding action; the most important levelling agents for vat dyes
d. Polyvinylpyrrolidone: very strong absorption-retarding effect, and very strong dye-bonding action (stripping agent).

Levelling Agents for Substantive Dyes. Mixtures of nonionic and anionic surfactants have long been used for this purpose. Nonionic components include ethoxylated fatty alcohols, fatty amines, fatty acids, alkylphenols, or propylene oxide polymers. Anionic components include fatty alcohol sulfates, alkylaryl sulfates, and Turkey red oil.

The task of these products is not so much to prevent too rapid absorption of the dye, as this can be controlled by salt addition, but more to ensure thorough wetting, and to disperse soil, preparations, dye deposits, etc.

Material dyed with substantive dyes can be levelled at a later stage by the use of polyvinylpyrrolidone, which is an auxiliary with an affinity for the dye.

Levelling Agents for Other Dyes. Other products referred to as “levelling agents” for dyeing with sulfur dyes or reactive dyes actually consist of wetting and dispersing agents.

Some commercial products are listed in Table 26.

Table 26. Some commercially available levelling agents for dyeing cellulose fibers [236]

Name	Producer	Uses	Composition
Albatex OR	Ciba-Geigy [a]	vat dyes	oligomeric amide
Albatex PON conc.	Ciba-Geigy [a]	vat and direct dyes	alkylbenzimidazole ethoxylate
Albigen A	BASF [b]	vat and direct dyes	polyvinylpyrrolidone
Anogal KF	Bauer & Gäbel [b]	vat dyes	polyglycol ether
Anogal NK	Bauer & Gäbel [b]	direct dyes	polyglycol ether
Cerafil BFA	Dr. Th. Böhme [b]	bleaching + dyeing	phosphoric ester + alkylarylsulfonate
Degal CR	Sigma [b]	vat and direct dyes	polyamide
Dupramin NFS	Thor-Chemie [b]	vat and direct dyes	polyglycol ether
Impegal SR	Impocolor [b]	vat dyes	condensation product
Levegal KNS	Bayer [b]	vat and direct dyes	polyglycol ether
Lyogen E	Sandoz [b]	vat dyes	oligomeric amide
Matexil DN-VL	ICI [c]	vat and direct dyes	polyglycol ether
Peregal P	BASF [b]	vat dyes	oligomeric amide
Ruco-Egalisierer IPM	Rudolf-Chemie [b]	vat dyes	amine condensation product
Sarabid SBF	CHT [b]	direct dyes	polyglycol ether + sulfonates
Sarabid VAT	CHT [b]	vat dyes	amine condensation product
Solidegal GL	Hoechst [b]	vat dyes	heterocyclic compound
Solidegal SR	Hoechst [b]	vat dyes	higher alkylamine
Uniperol O	BASF [b]	vat and direct dyes	polyglycol ether

[a] Switzerland;
[b] Germany;
[c] United Kingdom.

5.6.2. *Levelling Agents for Wool Dyeing*

Anionic dyes of the following types are used for dyeing wool: acid dyes, metal- complex dyes, afterchrome dyes, and reactive dyes (⟶ Textile Dyeing – Dye Classes and Dyeing Processes). Levelling agents for wool dyeing must reduce the rate of absorption of the dye, improve the migration of the dye during the dyeing phase, and improve the uniformity of color shade and color intensity of wool fibers that may have undergone damage (light, segregation, pretreatment, etc.) and may be of various origins. Nonionic, ionic, and amphoteric surfactants are used for these purposes, although their mechanism of action is still not completely understood [289].

Nonionic surfactants form a hydrophilic complex with the anionic dye, and thus are products with an affinity for the dye

[290], [291]. The molecules contain a hydrophilic group consisting of long polyglycol ether chains, and a hydrophobic portion, usually fatty alcohols, alkyl phenols, fatty acid alkylolamides, etc.

The ionic surfactants include cationic compounds, some of which are the most commonly used commercial levelling agents for wool dyeing. These have affinity both for fibers and dyes, and only have a levelling effect above the critical micelle concentration [292]. Typical compounds of this group include polyglycol ethers of fatty amines, fatty acid amide amines, and fatty alkylpolyamines.

Anionic surfactants exhibit fiber affinity. They include alkylnaphthalenesulfonic acids, fatty alkyl sulfates, alkylbenzenesulfonates, alkyl polyglycol ether sulfates, and polycarboxylates [293].

Amphoteric products combine the properties of cationic and anionic surfactants, and have an affinity for both dyes and fibers. Their effect can be varied by varying the auxiliary/dye ratio and the length of their polyether chains [294]. Typical products include ethoxylated nitrogen- containing fatty alkyl compounds with an anionic group on the nitrogen or at the end of the polyether chain.

Commercially available levelling agents for wool include many kinds of synergistic mixtures of surfactants of various ionic types. Products of universal application are available, as well as mixtures optimized for certain groups of dyes (Table 27).

Table 27. Some commercial levelling agents for dyeing wool [236]

Name	Producer	Uses*	Active substances
Albegal A	Ciba-Geigy, Switzerland	e	alkylamine polyglycol ether sulfate
Albegal B	Ciba-Geigy, Switzerland	f	ethoxylated fatty acid amide derivative
Albegal SET	Ciba-Geigy, Switzerland	e	modified nitrogen-containing fat and alcohol ethoxylates
Avolan AV 200 %	Bayer, Germany	a, c	alkylarylamine polyglycol ether
Avolan W	Bayer, Germany	e	alkyl polyglycol ether
Avolan REN	Bayer, Germany	f	alkylarylamine polyglycol ether
Avolan UL75	Bayer, Germany	a – e	alkylamine polyglycol ether sulfate
Breviol SCN	Henkel, Germany	a, d	amine condensate
Eganal SZ	Hoechst, Germany	a – e	surfactant combination
Keriolan FMK	CHT, Germany	e	polyglycol ether derivative
Lyogen WD	Sandoz, Switzerland	a – d	fatty amine polyglycol ether
Uniperol AC	BASF, Germany	a, b, d	alkylamine polyglycol ether
Uniperol KM	BASF, Germany	e	amphoteric product
Uniperol SE	BASF, Germany	a, b, d, e	amphoteric product
Uniperol W	BASF, Germany	a, c, d, e	polyglycol ether sulfate

* a = Acid dyes (strongly acid); b = Acid dyes (weakly acid); c = Afterchrome dyes; d = 1 : 1-Metal complex dyes; e = 1 : 2-Metal-complex dyes; f = Reactive dyes.

5.6.3. *Levelling Agents for Dyeing Polyamide Fibers*

Polyamide can be dyed with disperse dyes (⟶ Textile Dyeing) and with those classes of dyes that are also important for wool dyeing (acid dyes and metal- complex dyes).

Dyeing with Disperse Dyes. In long liquor dyeing of polyamide with disperse dyes, dispersants are always added to stabilize the dispersion (see Section Dispersants and Protective Colloids). With rapid acting processes, e.g., the dyeing of hosiery special mixtures of surfactants with a wetting, washing, and dispersing effect are used. These contain polyglycol ethers as dispersants as well as nonionic ethoxylation products and anionic conditioning agents. High molecular mass polyglycol ethers improve stripping of the hosiery products from the patterns.

Dyeing with Acid Dyes. When dyeing polyamide fibers with anionic dyes, anionic, nonionic, and cationic surfactants are used, as with the dyeing of wool (see Table 27).

Anionic levelling and resist agents with an affinity for fibers are used (see Section Resist Agents for Dyeing and Printing). These are mono- and polysulfonates of high molecular mass aliphatic and aromatic compounds [232], [295]. They also prevent formation of “stripes” on the dyed polyamide [296].

Cationic levelling agents (e.g., fatty amines) are similar or identical to those products used in wool dyeing (see Section Levelling Agents for Wool Dyeing).

Nonionic surfactants (ethylene oxide adducts such as ethoxylated fatty alcohols) form a complex with the anionic dye, thus having dye affinity [296-298].

When dyeing polyamide, mixtures of nonionic, ionic, and amphoteric products are also used (see Table 28).

Table 28. Some commercially available levelling agents for dyeing polyamide fibers [236]

Name	Producer	Uses	Active substances
Alviron CNF	TC, Switzerland	levelling agent, antistriping agent	sulfonates, ethylene oxide condensates
C 14	CHT, Germany	antistriping agent	araliphatic sulfonates and polyglycol ethers
Dispergal B liq.	Auschem, Italy	levelling agent	sulfonated naphthalene derivatives
Gisapon 1476	Rotta, Germany	levelling agent	modified fatty amine ethoxylate
Levegal FTS	Bayer, Germany	levelling agent	polyglycol ether derivative and sulfonates
Lyogen PN	Sandoz, Switzerland	antistriping agent	aromatic sulfonate
Perigen HTP	Dr. Petry, Germany	hosiery dyeing	fatty acid polyglycol esters, oleates, terpenes
Ruco-Egalisierer PA	Rudolf-Chemie, Germany	levelling agent, antistriping agent	alkylamine ethoxylate
Sandogen NH	Sandoz, Switzerland	levelling agent	fatty amine polyglycol ethers
Uniperol AC	BASF, Germany	levelling agent	alkylamine polyglycol ether
Univadin PA	Ciba-Geigy, Switzerland	levelling agent	combination of anionic surfactants

When choosing an auxiliary, care must be taken that it does not have a harmful effect on the compatibility of dyes properties [299] or lightfastness of dyes [300], [232].

Uniform absorption of the dye can be achieved by pH control as well as by the use of retarding agents. This is carried out by automatically metering acid into the dye bath, or by adding substances that gradually release acid (see Section pH Regulators (Acid- and Alkali-Releasing Agents)).

5.6.4. *Levelling Agents for Dyeing Polyester Fibers*

The dyeing of polyester fibers with disperse dyes by the exhaustion process is carried out either at 100 °C with dyeing accelerants (carriers) (Section Dyeing Accelerants (Carriers)) or at 120 – 140 °C under pressure (high-temperature process) (⟶ Textile Dyeing – Thermosol Process).

At boiling point and in the presence of a carrier, a levelling agent is usually unnecessary, as polyester fibers are dyed slowly and uniformly under these conditions.

To give uniform dyeing in the high-temperature process, levelling agents are necessary to prevent excessively rapid dye absorption, equalize the different absorption rates of the various dyes, and/or improve the migration of the dye in the high-temperature phase. They must not affect the stability of the dye dispersion, as otherwise filtering effects can severely disturb the process, causing unlevel dyeing and impairment of fastness.

There are three typical groups of levelling agents for the high temperature process: dyeing accelerants (carriers), ethoxylated products, and new special products.

In the high temperature dyeing process, carriers increase the rate of diffusion of the dye in the fiber and in the fiber/liquor equilibrium, thereby improving the migration of the dye [301]. They are also suitable for levelling nonuniformly dyed materials [302]. However, carriers can increase the exhaustion rate of the dye so much that unlevel dyeing takes place. They should therefore only be added to the dye bath at higher temperatures or in the migration phase. Because they cause swelling of the fiber, carriers reduce stripiness due to nonuniform dye uptake [303], [304].

The group of carriers used in high temperature dyeing includes the same compounds as are used in dyeing polyester fibers at boiling temperatures (see Section Dyeing Accelerants (Carriers)): halogenated benzenes, halogenated toluenes, 2-phenylphenol, diphenyl, diphenyl ethers, salicylic esters, methylnaphthalene, etc. [305]. However, these compounds often have disadvantages such as reduced lightfastness, environmental pollution, excessive fiber swelling, and strong odor. As alternatives to these products with their associated environmental problems, aliphatic carboxylic esters are used [306-308]. However, the carrier effect of these products is inadequate when dyeing at boiling temperatures.

Ethoxylated products have a retarding effect on dyes in the absorption phase, and are therefore used especially in rapid dyeing processes or in shading at high temperature [309-312]. The main members of this group of ethoxylated products are ethoxylated castor oil, stearic acid, alkylphenols, and the sulfuric or phoshoric esters of ethoxylated fatty alcohols or

alkylphenols.

Special levelling agents include product mixtures that produce equal rates of absorption of dye mixtures (synchronization) during the heating phase, and improve dye migration in the high-temperature phase. These products usually consist of combinations of alcohols, esters, or ketones of medium chain length ($C_6 – C_{12}$ or $C_7 – C_{16}$) with emulsifying systems [313]. The effect of cyclodextrins as levelling agents for dyeing polyester fibers under high temperature conditions is discussed in [314], [315].

Commercial levelling agents are listed in Table 29.

Table 29. Some commercial levelling agents for dyeing polyester fibers [236]

Name	Producer	Uses	Active substances
Alviron PES	Allied Colloids, United Kingdom	levelling, dispersing, migration and crease prevention agent	modified phosphoric acid esters
Anogal NTH	Baur & Gäbel, Germany	levelling and dispersing agent	ethoxylated compound with anionic part
Breviol LC	Henkel, Germany	levelling agent	ester combination
Eganal PS	Hoechst, Germany	levelling and dispersing agent	linear polycondensate
Egasol WFE	CHT, Germany	levelling agent with washing properties	aromatic esters with ethoxylates
Levegal HTC	Bayer, Germany	levelling agent	halogenated aromatic alkyl polyglycol ethers, alkylarylsulfonates
Levegal HTN	Bayer, Germany	levelling and dispersing agent	alkylphenol and fatty acid polyglycol ethers
Levegal MSF	Bayer, Germany	levelling agent, crease prevention agent	fatty acid glyceryl esters
Palegal HT	BASF, Germany	levelling agent	nonionic aliphatic compounds
Palegal SFD	BASF, Germany	levelling and migration agent	aromatic esters and emulsifiers
Perigen HTN	Dr. Petry, Germany	levelling agent	aromatic esters and emulsifiers
Ruco-Egalisierer SBB	Rudolf-Chemie, Germany	levelling agent and stripe formation prevention agent	carboxylic acid alkyl esters, oil emulsifiers, and antifoaming agents
Sandogen BM	Sandoz, Switzerland	levelling agent	carboxylic acid esters, emulsifiers
Synthapal BM	Dr. Th. Böhme, Germany	levelling and migration agent	aromatic esters and emulsifiers
Synthapal HTM	Dr. Th. Böhme, Germany	levelling, migration, and dispersing agent	polyglycol esters, aromatic sulfonate
Trevylen	Kreussler, Germany	carrier	*N*-alkylphthalimides
Univadin 3flex	Ciba-Geigy, Switzerland	levelling, migration, and dispersion agent	fatty acid esters, aryl polyglycol ethers, polyalcohol

Carrier-Free Dyeing of Modified Polyester Fibers. The modified polyester fibers known as NCD (noncarrier dyeable) fibers can be dyed with disperse dyes at boiling temperatures without the addition of carriers (⟶ Textile Dyeing – Dyeing of Modified Polyester Fibers). To prevent unlevel effects caused by excessively rapid absorption of the dye, levelling agents of the ethoxylated product type are used.

Dyeing of Polyester Fibers in Alkaline Media (⟶ Textile Dyeing). Polyester fibers are also dyed in an alkaline medium to reduce the oligomer problem and to facilitate the process of finishing (desizing and dyeing in one step). Levelling agents of the ethoxylated product type are used for this, and special products are also used to buffer the liquor in the pH range 9 – 9.5 [316].

5.6.5. *Levelling Agents for Dyeing Polyacrylonitrile Fibers*

The cationic dyes used for dyeing anionically modified polyacrylonitrile fibers usually have a very high affinity for the fiber, and diffuse very rapidly into the fiber above the glass transition temperature (⟶ Textile Dyeing – Cationic Dyes). The ability of these dyes to migrate is low, and unlevel effects due to excessively rapid absorption of the dye are difficult to rectify. The absorption rate can be moderated by controlling the temperature and adding levelling agents (cation active or anion active products), thereby improving the uniformity.

Cationic products known as retarders are of the greatest importance. These can be of the “permanent” or “temporary” type [317], the more important of these being the permanent retarders, which have a high affinity for the fiber. These retard dye

absorption by attaching themselves to the anionic sites in competition with the dye. For given characteristic parameters of the fiber, dye, and retarder, the amount of retarder required can be calculated for each dye bath formulation [318], [319]. If the amounts added are too high, the retarder can block the fiber, affecting the dye yield. It must therefore be stripped by an anionic auxiliary to enable further dyeing to be carried out. Commercially available retarders contain quaternary ammonium salts with long fatty alkyl side chains (C_{12} – C_{14} alkyltrimethyl or C_{10} – C_{14} alkyldimethylbenzyl chains) [320]. Polymeric compounds [321] or polyethoxylated amines [322] are also often used.

Cationic products with a significantly lower affinity for the fiber than for the dye are marketed as migration auxiliaries. These increase the migration of the dye during the equilibrium phase. They have only a slight effect on the absorption rate, and do not block the fiber [323]. Quaternary ammonium salts with aromatic ring systems (e.g., trimethylbenzylammonium chloride), electrolytes (e.g., NaCl, Na_2SO_4) or organic salts with larger cations [324] increase migration.

Other levelling agents used in PAN dyeing include anionic products with an affinity for the dye (e.g., naphthalenesulfonic acid – formaldehyde condensation products). The possibility of precipitation of dye – auxiliary adducts can be avoided by adding nonionic dispersing agents (ITW process, [325]). A disadvantage of the anionic products is that they reduce bath exhaustion.

A list of commercial products is given in Table 30.

Table 30. Some commercially available levelling agents for dyeing polyacrylonitrile fibers [236]

Name	Producer	Uses	Active stubstance
Astragal M	Bayer, Germany	migration agent	organic amino compound
Atragal PAN	Bayer, Germany	retarder	quaternary ammonium compound
Astragal TR	Bayer, Germany	temporary retarder	quaternary ester ammonium compound
Basacryl Salz AN	BASF, Germany	migration agent	quaternary ammonium compound
Basacryl Salz MX	BASF, Germany	retarder	quaternary ammonium compound
Basacryl Salz TR	BASF, Germany	retarder	quaternary ammonium compound
Geneucol ART	Dr. Th. Böhme, Germany	retarder	quaternary ammonium compound
Geneucol MI	Dr. Th. Böhme, Germany	migration agent, nonblocking	quaternary ammonium compound
Geneucol RTM	Dr. Th. Böhme, Germany	retarder	quaternary ammonium compound
Osimol MA	Grünau, Germany	migration agent	quaternary ammonium compound
Osimol RAC	Grünau, Germany	retarder	quaternary ammonium compound
Remartard AC	Hoechst, Germany	retarder	quaternary ammonium compound
Remartard T	Hoechst, Germany	retarder	cationic nitrogen compound
Retargal AN	Sandoz, Switzerland	retarder, nonblocking	quaternary ammonium compound
Sarabid UZ 1290	CHT, Germany	anionic retarder	ethylene oxide adducts
Tinegal B	Ciba-Geigy, Switzerland	retarder	araliphatic quaternary fatty acid amine
Tinegal MR	Ciba-Geigy, Switzerland	retarder with migration effect	*n*-tetraalkylammonium halide
Tinegal PAC	Ciba-Geigy, Switzerland	retarder	quaternary ammonium compound

5.6.6. *Levelling Agents for Dyeing Other Synthetic Fibers and Fiber Blends*

Acetate Fibers. When dyeing acetate fibers (⟶ Textile Dyeing – Dyeing Processes for Cellulose Acetate) with disperse dyes, levelling agents with a dispersing action are used (nonionic or anionic surfactants, see Section Dispersants and Protective Colloids). These products often also need to have a detergent action to remove fiber preparations during dyeing and to maintain them in emulsified form.

Triacetate Fibers. When dyeing triacetate fibers (⟶ Textile Dyeing – Dyeing Processes for Cellulose Triacetate) the same dispersants as are used for dyeing other fibers with disperse dyes (polyester, polyamide, and acetate), are used as

levelling agents. When dyeing at 100 °C, special carriers are often used (see Section Dyeing Accelerants (Carriers)).

Fiber Blends. If the various types of fiber in a fiber mixture are dyed successively (⟶ Textile Dyeing – Dyeing of Fiber Blends), the same levelling agents can generally be used as when dyeing the fibers individually. For single-bath dyeing with dyes of differing ionic character, special dispersants must be used to prevent mutual precipitation of the two types of dye, which could have a harmful effect on the levelling and rubbing fastness. These special auxiliaries are mixtures of nonionic and anionic surfactants (e.g., Uniperol KA, BASF).

If the individual components of a fiber blend are not dyed to the same color intensity by a particular dye (e.g., when dyeing wool – polyamide with anionic dyes, or cellulose – polyamide with substantive dyes), a better choice of dye, careful temperature control, and the use of an auxiliary with an affinity for the fiber can reduce the dye affinity of that component of the mixture with the greater affinity. The products used for this are the same as those used as chemical resist agents (see Section Resist Agents for Dyeing and Printing). For tone-in-tone dyeing of polyamide – polyurethane blends, cationic products with an affinity for the dye or ethoxylated fatty amines are used.

A list of commercially available products is given in Table 31.

Table 31. Some commercial levelling agents for dyeing fiber blends [236]

Name	Producer	Uses	Active substances
Dispergal WP	Auschem, Italy	WO/PA	derivatives of sulfonated mixtures
Emulsol OS	Auschem, Italy	PA/WO	ethoxylated fatty alcohol
Gelonal W	Geissler, Germany	PA/WO	alkylamine polyglycol ether
Levegal ER liq.	Bayer, Germany	CEL/PA	aromatic sulfonates
Ritandin CD	Baumheier, Germany	CEL/PAN	fatty acid condensate, polyglycol ether
Teban ERP	Dr. Th. Böhme, Germany	PA/CEL	modified arylsulfonates
Teban WPA	Dr. Th. Böhme, Germany	WO/PA	solvent-containing arylsulfonate mixture
Transferin UN	Dr. Th. Böhme, Germany	WO/PES, WO/PAN, PES/PAN	modified alkyl polyglycol ether
Uniperol KA	BASF, Germany	PAN/WO	mixture of nonionic surfactants

WO = Wool; PA = Polyamide; PAN = Polyacrylonitrile; CEL = Cellulose; PES = Polyester.

5.6.7. *pH Regulators (Acid- and Alkali-Releasing Agents)*

The pH influences the absorption of anionic dyes onto wool and/or polyamide fibers and the fixing of reactive dyes onto cellulose fibers. By controlling the pH, it is possible to improve dye levelling in the absorption phase, or, when dyeing polyester – cotton blends with disperse and reactive dyes, to control the fixing of the dye.

Dyeing of Wool and Polyamide. The level absorption of anionic dyes onto wool and/or polyamide fibers can be controlled by means of the pH of the dye liquor. The dye equilibrium in acids lies on the side of the fiber, and the dye is effectively absorbed onto the fiber. In alkalis, the equilibrium is shifted in favor of the liquor.

Dyeing is started at a pH of 8 – 8.5, and the pH is then successively reduced by adding acids or acid salts.

This method has the disadvantage that there is a sharp change in pH at the point where the acid is added, leading to a pH gradient in the dye bath.

The use of acid releasing agents based on esters of organic acids [326-331] enables the acid to be liberated continuously on heating the dye bath and hydrolyzing the ester. This causes the pH to decrease at a slow and constant rate in all parts of the dye bath, so that the anionic dye is absorbed in a controlled and uniform manner.

Dyeing of Cotton and Polyester – Cotton Blends. The dyeing of mixtures of cotton and polyester fibers by the exhaustion dyeing process in the same dye bath (single bath/two stages or single bath/single stage) is more economic than the process that includes a change of dye liquor [332].

The single-bath/single-stage process in which an increase in temperature is used to cause an acid releasing agent present in the liquor to lower the pH is especially advantageous. Fixing of the reactive dye begins in alkaline solution at pH 10 – 10.5 and 60 °C, and the dyeing conditions for the disperse dye are reached at 130 °C and pH 5 – 6 after ca. 130 min.

The acid-releasing agents used are salts of α-halocarboxylic acids which decompose on heating. Phosphates are also used as buffering agents.

For continuous dyeing with reactive dyes, separate application of the dye and the fixing alkali is used (two baths) to

prevent hydrolysis of the dye (pad steam process).

Using the alkali-releasing agent sodium hydrogencarbonate in the dye bath makes the single-bath pad steam process possible, in which the reactive dye is stabilized at room temperature and pH 7 – 7.5, and is fixed at 100 °C in the steamer. The alkaline fixing conditions are then produced by decomposition, liberating CO_2 and giving a pH of 9 – 9.5.

A list of commercially available products is given in Table 32.

Table 32. Some commercially available acid releasing agents [236]

Name	Producer	Active substances
Emenil DG	Emequimica, Spain	heterocyclic compound
Eulysin WP	BASF, Germany	organic ester
Levafix Salz PC	Bayer, Germany	aliphatic carboxylic acid, sodium salt
Meropan EF	CHT, Germany	special esters
Sandacid V	Sandoz, Switzerland	heterocyclic compound
Sandacid VS	Sandoz, Switzerland	organic formate
Verolan GBK	Rudolf-Chemie, Germany	low molecular mass ester

5.6.8. *Testing*

To test the effectiveness of levelling agents in the laboratory, the effects of the auxiliary are measured either on individual processes in the dyeing phase, or by assessing any improvement in uniformity produced by addition of the levelling agent to a system that would otherwise give unlevel results.

Testing of the Retarding Effect. To assess the retarding effect of a levelling agent, the progress of absorption of the dye from the dye bath and onto the fiber is observed with and without the auxiliary concerned. During the dyeing process, samples are removed from the dye liquor, and either the amount of dye deposited on the fiber is determined by extraction or by reflectance measurements, or the amount of dye remaining in the liquor is determined colorimetrically or by dyeing more undyed textile in this liquor [333-336]. By using dye combinations, the effect of the dyeing auxiliary on the simultaneous absorption of the dye components can also be determined from any change in the color shade [337].

Testing of the Migration Effect. The migration effect of a levelling agent leads to levelling of nonuniformly dyed material either during the course of the dyeing process or subsequently. Assessment of the migration effect is carried out in a migration, levelling, or boildown test [338-341]. Dyed material is treated with undyed material under dyeing conditions in a blind dye bath, a bath that contains all the chemicals necessary for dyeing except the dye. The effect of the levelling agent in promoting transfer of the dye from dyed material to accompanying undyed material is then observed [342], [343]. Testing of the retention effect of the levelling agent (i.e., the reduction in the bath exhaustion) is described in Part 2 of DIN 54 290 [344]. This test is important in the assessment of levelling agents.

Rectification of Unlevel Dyeing. To test the ability of levelling agents to rectify unlevel dyeing, caused by different dye absorption of fibers, special test materials and dyes are used in which the formation of stripes is especially noticeable. The levelling effect of a levelling agent is assessed visually, by reflectance measurements, or by extraction of the absorbed dye [339], [344-347].

Special Levelling Tests. To test the effectiveness of levelling agents, experimental conditions are often used which give unlevel dyeing without the use of auxiliaries. The extent to which the addition of a levelling agent improves the uniformity of dyeing is then assessed [348-354].

As well as its levelling effect, other effects of a levelling agent are often tested, e.g., wetting effect, dispersing effect, foaming tendency, etc., as these properties can also affect the uniform appearance of a dyed material.

Chemical Identification and Quantitative Determination of Levelling Agents. As most levelling agents are surfactants or similar products, the usual methods of surfactant analysis are used [232], [355].

5.7. Dyeing Accelerants (Carriers)

Dyeing accelerants (carriers) are used in the dyeing of synthetic fibers by the exhaustion process to give increased rate of absorption of the disperse dye onto the fiber, more rapid diffusion into the fiber, and better dyeing yield.

The use of carriers enables polyester fibers to be dyed with disperse dyes to a satisfactory intensity even at 100 °C. The carrier dyeing process is therefore of particular importance for blended fibers of polyester and wool, as wool must not be subjected to wet treatment at temperatures significantly above 100 °C (⟶ Textile Dyeing – Cationic Dyes). Carriers need not be used to increase dyeing yield in the high temperature dyeing of pure polyester (⟶ Textile Dyeing – Dyeing of Polyester Fibers), but they are often used as levelling agents in this process (see Section Levelling Agents for Dyeing Polyester Fibers.). They are much less used for dyeing triacetate, polyacrylonitrile, and polyamide fibers than for polyester fibers.

Research into carriers has concentrated on their effects on polyester fibers [301], [356-360]. It is believed that the activity of carriers in accelerating dyeing is mainly located in the fiber region.

Dyeing in the presence of carriers can also bring a number of problems. Most carriers are toxic, contaminate wastewater, and lead to a high COD value [361], [362]. Carrier dyeing can also cause fiber swelling sufficient to change the technological properties of the fiber. Also, carrier residues remaining on the fiber after dyeing can affect lightfastness and heat setting fastness. Apart from avoiding or minimizing these problems, an efficient carrier must also fulfil the following requirements [232]:

1. Maximum effectiveness from a minimum amount of carrier used
2. Independence of the carrier effect from the chemical constitution of the dye
3. Rapid formation of a stable emulsion of the carrier under dyeing conditions
4. Low volatility in steam
5. Ease of removal from the fiber
6. Very little or no odor contamination of the environment
7. With disperse dyes, no increase in soiling of the other fibers present in fiber blends

Commercial products are usually liquid and selfemulsifying, i.e., they contain emulsifiers and sometimes solvents as well as the active substances. Typical carrier formulations contain 60 – 80 % active substance, 10 – 30 % emulsifier, and 0 – 10 % solvent [363].

Active substances for dyeing accelerants include the following groups of compounds [305], [364], [365]:

1. Liquid halogenated benzenes
2. Aromatic hydrocarbons
3. Aromatic hydroxy compounds
4. Aromatic alcohols, ketones, carboxylic acids and their esters
5. Alkyl phthalimides
6. Substituted phenylglycols and their esters

Dyeing Accelerants for Polyester Fibers. Carriers are mainly used as dyeing accelerants in polyester – wool blends and as levelling agents in the high-temperature dyeing of pure polyester fibers. The most important active substances used in dyeing accelerants for polyester fibers are 1,2-dichlorobenzene; 1,2,4-trichlorobenzene; 2-phenylphenol; diphenyl; diphenyl ether; methyl, butyl, and benzyl benzoate; methyl salicylate; dimethyl phthalate; tetralin; α- and β-methylnaphthalene; phthalic acid *N*-butylimide; and chlorophenoxyethanol.

The composition of a dyeing accelerant is always a compromise. The volatility and toxicity of chlorobenzenes limit their use in closed equipment, and the reduction of lightfastness by naphthalene derivatives, diphenyl, or 2-phenylphenol necessitates subsequent fixing of the material at > 180 °C. Chlorobenzenes, methylnaphthalene, and carboxylic esters are particularly suitable for polyester – wool blends, as they do not increase soiling of the wool by disperse dyes.

The properties of the various types of dyeing accelerant are compared in [366-368]. A method of testing the accelerating effect and dye yield is described in DIN 54 289.

Dyeing Accelerants for Other Synthetic Fibers. With triacetate fibers, dyeing accelerants are only necessary to increase the dye yield for very deep tones or with dyes that do not diffuse well. The most commonly used compounds are esters of benzoic acid, salicylic acid, or phthalic acid.

To avoid severe shrinkage of polyacrylonitrile fibers, dyeing is occasionally carried out below the glass transition temperature T_g in the presence of dyeing accelerants. Ethylene and propylene carbonate and particularly benzyloxypropionitrile are suitable.

Certain modified polyacrylonitrile fibers become yellow and delustered or lose strength if they are dyed at 100 °C. Such fibers can be dyed at lower temperatures by using dyeing accelerants. Dyeing accelerants for the Verel (modified acrylic) fibers include triisobutyl phoshates.

Poly(vinyl chloride) fibers are normally dyed in the presence of dyeing accelerants (usually benzoic acid esters) with disperse dyes at 60 – 70 °C.

With polyamide fibers, it is often difficult to obtain a high enough dyeing temperature, especially when dyeing in a jigger. The dyeing properties are therefore improved either by means of pretreatment with a benzyl alcohol – ethanol mixture [368], or by the direct addition of benzyl alcohol to the dye bath.

When dyeing aramid fibers, which have poor dyeing properties at boiling temperatures, butyl benzoate, salicylic esters, benzaldehyde, or acetophenone are added as dyeing accelerants.

Environmental Aspects of Dyeing Accelerants. The environmental problems presented by many dyeing accelerants include toxicity to humans, fish, and sewage sludge, poor biological degradability, and odor problems [361], [369], [370]. However, as these active substances have an affinity for the fiber, and are ca. 75 – 90 % adsorbed by the textile material [305], [361], only the biologically harmless emulsifiers normally remain in the dye bath and subsequently enter the wastewater. The dyeing accelerant (carrier) that remains on the fiber is expelled during the drying or fixing process, so that the waste air cleaning processes control the emission of this material [371].

Commercially available dyeing accelerants are listed in Table 33.

Table 33. Some commercially available dyeing accelerants (carriers) [236]

Name	Producer	Uses	Active substance
Carolid 250	Sybron/Tanatex, Netherlands	PES	2-phenylphenol
Dilatin NAN	Sandoz, Switzerland	PES	aromatic hydrocarbons
Levalin ACE	Bayer, Germany	PAN	aromatic ethers
Levegal DTE	Bayer, Germany	PES, CTA	halogenated aromatics
Levegal PEW	Bayer, Germany	PES, CTA	*N*-alkylphthalimide
Levegal PTN	Bayer, Germany	PES, CTA	aromatic carboxylic acid esters
Matexil CA-MN	Zeneca, United Kingdom	PES	methylnaphthalene
Remol HT	Hoechst, Germany	PES	diphenyl
Remol LF	Hoechst, Germany	PES	diphenyl/methylnaphthalene
Sarapol DLU	CHT, Germany	PES	aromatic esters and hydrocarbons
Setavin CA	Zschimmer + Schwarz, Germany	PES	phthalic acid derivatives
Tanadel PTN	Sybron/Tanatex, Netherlands	PES	aromatic ester
Tanadel TMO	Sybron/Tanatex, Netherlands	CTA	aromatic ester

PES = Polyester; CTA = Cellulose triacetate; PAN = Polyacrylonitrile.

5.8. Anticreasing Agents

Crease formation is one of the most unpleasant faults that can occur during the finishing of piece goods. There is an extensive literature on the mechanism, prevention, and cure of creasing [372].

Because the removal of many types of crease is either impossible or extremely difficult, textile finishers always try to prevent them as far as possible [373-378].

Longitudinal creases are always formed when the textile is treated in rope form and with high lengthwise tension. If piece goods are treated in wide form, a crease-free product can be obtained, although the special character of many textile qualities (bulk, handle, fullness, etc.) can only be obtained by finishing in machines that do not produce high tension (winches, overflows, and jets). As well as measures that can be taken involving processing and mechanical aspects, (preliminary fixing of the material, low treatment temperature, slow cooling after heat treatment, finishing on special machines with low longitudinal tension), the formation of permanent creases can also be minimized by adding special auxiliaries to the treatment bath [232], [373-375], [379].

The action of anticreasing agents is not fully understood. They probably act as lubricants which improve the sliding of one part of the textile over another, so that creases formed during movement of the textile open up and disappear more easily [374], [376], [377].

The following are the main classes of compounds used:

1. Synthetic products based on fatty acids or their esters, amides, alkylol esters, and alkylolamides also fatty alcohols, usually ethoxylated. Apart from these nonionic compounds, ionic compounds with carboxyl, sulfonic, or phosphoric acid groups can also be used.
2. Products based on lecithin.
3. Products based on water-soluble alkoxylated high molecular mass polyamides.
4. High molecular mass polyethoxylates, polyacrylates, and acrylamide – acrylic acid copolymers.

A list of commercial products is given in Table 34. These are often mixed products which also contain dispersing and levelling agents as well as the anticreasing agents. The auxiliaries must not cause foaming, but should if at all possible have an antifoaming effect.

Table 34. Some commercially available anticreasing agents [236]

Name	Producer	Active substance
Breviol LFP	Henkel, Germany	modified phosphoric ester
Breviol PAM	Henkel, Germany	polyamide derivative
Cibafluid U	Ciba-Geigy, Switzerland	mixed polymerizate based on polyethers
Eganal LUB	Henkel, Germany	fatty acid polyglycol ester
Hydrocol LFV	Rudolf-Chemie, Germany	ethoxylated caprolactam derivative
Lubit LC	Sybron/Tanatex, Netherlands	modified phosphoric ester
Osimol LVP	Grünau, Germany	modified phosphoric ester
Palatex S	BASF, Germany	polyamide derivative
Persoftal L	Bayer, Germany	sulfonated fat
Pretolon L	Ciba-Geigy, Switzerland	modified lecithin

Rucolin JES	Rudolf-Chemie, Germany	polyacrylamide
Sevosoftal U	TC, Switzerland	sulfated fats and lubricating components
Tebolan MDF liq.	Dr. Th. Böhme, Germany	phosphoric ester, special emulsifiers
Viscavin S 700	CHT, Germany	modified phosphoric esters

5.9. Auxiliaries for Continuous Dyeing

Continuous and semicontinuous dyeing processes operate according to the following fundamental principle [232], (⟶ Textile Dyeing – Continuous and Semicontinuous Dyeing):

1. Padding (impregnation) of the textile with the dye liquor followed by squeezing-off to give a definite amount of absorbed liquor
2. Intermediate drying (not with all processes)
3. Developing and/or fixing of the dye by chemical, hot air, steam, or contact heat treatment
4. Removal of the unfixed dye and auxiliaries by washing

Continuous dyeing processes are used both with water-soluble dyes (substantive, reactive, acid, and cationic dyes) and water-insoluble dyes (disperse and vat dyes).

The following products are used as auxiliaries for continuous dyeing (padding auxiliaries):

1. Dispersants and protective colloids (see Section Dispersants and Protective Colloids), solubilizing agents and hydrotropic agents (see Section Dye Solubilizing and Hydrotropic Agents) are used to improve the solubility of the dye or the stability of the dye dispersion
2. Wetting agents (see Section Wetting and Deaerating Agents) and products to increase liquor pickup lead to uniform wetting and formation of a continuous adherent film of liquor on the textile
3. Antimigration agents prevent of migration of the dye during intermediate drying
4. Dye solubilizing agents and hydrotropic compounds, fiber swelling agents, and fixing accelerants (see Section Dye Solubilizing and Hydrotropic Agents) improve transfer of the dye into the fiber
5. Antifrosting agents promote dyeing of protruding parts of the fiber

An auxiliary must also have the following properties to ensure problem-free processing without harming the quality of the textile:

1. No marked affinity for the fiber, so that the composition of the pad liquor is kept constant
2. No detrimental effects on handle and fastness properties
3. Very low foaming tendency, although foam formation may sometimes be desirable (antifrosting agents)

The auxiliaries discussed below are used exclusively for continuous dyeing.

5.9.1. *Antimigration Agents*

In all continuous dyeing processes, after padding of the dye there is always the risk of dye and pigment migration due to nonuniform drying. The dye migrates with the water to the more rapidly drying parts of the textile, causing variations in color intensity and imperfect dye penetration. Antimigration agents prevent this dye movement. The action of these products, which takes place during the drying process, limits the growth of the dye particles by causing a kind of adhesion effect and viscosity increase [380], [381]. These auxiliaries are not often used when dyeing with water-soluble dyes. Highly active products include alginates or high molecular mass acrylic acid or acrylamide copolymers. Other compounds are used to a lesser extent, e.g., cellulose ethers, starch ethers, carboxymethyl celluloses, modified polysaccharides, and poly (vinyl alcohols) [382].

5.9.2. *Antifrosting Agents*

With certain textile articles and thick-pile, parts of the fiber that protrude from the yarn structure are inadequately wetted by the dye liquor, and hence are inadequately dyed. The textile then becomes covered with a gray “veil” (frosting). The effect can be prevented by two methods, in both of which the protruding fibers are held in the liquor [383], [384]:

1. Foam production during heating of the liquor in dye fixing. Mixtures of nonionic and anionic surfactants are used, sometimes with special volatile antifoaming agents that prevent foam formation during padding.
2. Increasing the viscosity and thickness of the liquor film to prevent the fibers from protruding after squeezing off. This is achieved by using high molecular mass polyacrylates and acrylamide – acrylic acid copolymers.

5.9.3. *Products to Increase Liquor Uptake*

Some continuous and semicontinuous dyeing processes rely on the textiles picking up a large amount of liquor (e.g., the wet-in-wet process for vat dyes ⟶ Textile Dyeing and the pad-jig process, ⟶ Textile Dyeing – Pad Jig Process). High liquor pickup is synonymous with maximum wetting and penetration of the textile, and is obtained by using effective pad-wetting agents (see Section Wetting and Deaerating Agents). A further increase in liquor pickup is achieved by increasing its viscosity, thereby preventing it from draining out of the interstices of the textile material. The effectiveness of this method is strongly dependent on the structure of the textile (open or compact).

Products for increasing liquor pickup include high molecular mass polyacrylates and acrylamide – acrylic acid copolymers. Some commerically available products are listed in Table 35.

Table 35. Some commercially available auxiliaries for the pad dyeing process [236]

Name	Producer	Uses	Composition
Irgapadol MP	Ciba-Geigy [b]	antimigration agent	poly(acrylic acid) derivative
Levalin KAM	Bayer [a]	antimigration agent	polyacrylate
Matexil FA-MIV	ICI [c]	antimigration agent	organic polyelectrolyte
Primasol AMK	BASF [a]	antimigration agent	acrylic acid copolymer
Solidokoll N	Hoechst [a]	antimigration agent	polyacrylamide derivative
Arbyl FRO	Grünau [a]	antifrosting agent	fatty acid condensate
Emigen DPR	Hoechst [a]	antifrosting agent	polymerization product
Lyogen CW	Sandoz [b]	antifrosting agent	alkyl polyglycol ethers
Matexil PA-S	ICI [c]	antifrosting agent	poly(ethylene glycol)
Primasol SD	BASF [a]	antifrosting agent	preparation of nonionic surfactants
Impegal EPR	Impocolor [a]	increase in liquor uptake	polyacrylamide

[a] Germany;
[b] Switzerland;
[c] United Kingdom.

5.10. Aftertreatment Agents to Improve Fastness

Many dyes of interest for textiles do not give adequate wetfastness properties (fastness to water, washing, perspiration, and ironing) or fastness to rubbing and light. Textiles dyed with direct dyes often require aftertreatment to improve fastness, though this also applies to other types of dyed textiles.

Two principal methods are used to improve fastness properties:

1. Unfixed dyes, dye aggregates, or dye degradation products are removed by washing or cleaning with complexing agents [279] or various types of surfactants with a dispersing action [385], by reductive decomposition [386], or by an organic solvent [387]. These auxiliaries are not usually classified as aftertreatment agents.
2. Bonding of poorly fixed dye to the fiber. If removal of the unfixed dye by washing does not give the required fastness properties, the poorly fixed dye is bonded to the fiber by an aftertreatment process. The auxiliaries used for this differ in chemical structure and the mechanism by which they act, depending on the type of dye and type of fiber [388].

Aftertreatment of Direct Dyes. The basic principles of the use of aftertreatment agents are illustrated by the example of dyeing with substantive dyes (⟶ Azo Dyes – Direct Dyes with Aftertreatment).

Aftertreatment with Cationic Produts [389], [390]. Cationic products improve wetfastness by forming sparingly soluble dye salts with the anionic groups of the dye [233]. Aftertreatment agents of this type include quaternary (poly)ammonium compounds with long hydrocarbon chains, polyamines, or polyethyleneimine derivatives.

Aftertreatment Agents Based on Formaldehyde Condensation Products, and Formaldehyde-Releasing Products. The use of these basic condensation products presumably leads to the formation of sparingly soluble adducts with the dye molecules [233], [391], [392]. Formaldehyde-releasing condensation products are particularly effective with dyes that have free hydroxyl or amino groups. The dye molecules probably become linked to the sparingly soluble dye aggregates by methylene bridges. Commercial products of this type include formaldehyde condensation products with amines, polynuclear aromatic phenols, cyanamide, or dicyandiamide.

Aftertreatment with Metal Salts. Aftertreatment with copper sulfate or potassium dichromate leads to complexation with certain substantive dyes. These reduce the aqueous solubility of the dye [233], [393], [394].

Aftertreatment with diazotized bases can only be used with dyes that have amino and hydroxyl groups in suitable positions, and, for example, couple with diazotized 4-nitroaniline [393].

Aftertreatment of Other Classes of Dye. Aftertreatment with cationic products can be used with all dyes that contain anionic groups (reactive, sulfur, acid, and metal-complex dyes). The following aftertreatment methods are also used:

Acid Dyes. Condensation products from aromatic sulfonic acids are used as well as cationic products, especially with polyamide fibers.

Sulfur Dyes. In addition to the use of cationic products to improve wetfastness, dark colored dyed materials must be aftertreated with sodium carbonate or sodium acetate to neutralize the sulfuric acid produced by oxidation of the sulfur dyes on storage, which can damage the fiber.

Reactive Dyes. Aftertreatment is usually carried out in two stages:

1. Removal (as thoroughly as economically possible) of the substantively bonded dye hydrolyzate by washing with very soft water. This process is very much improved by the use of complexing agents, which suppress the formation of sparingly soluble alkaline-earth salts of the dye hydrolyzate in incompletely softened water [279].
2. Fixing of the residual dye hydrolyzate by a cationic aftertreatment agent.

Some trade names used for these products are listed in Table 36.

Table 36. Trade names used for aftertreatment agents to improve wetfastness [236]

Name	Producer	Dyes whose wetfastness properties can be improved
Alcofix	Allied Colloids, United Kingdom	direct, reactive, and acid dyes
Basolan	BASF, Germany	acid and metal-complex dyes
Cibatex	Ciba-Geigy, Switzerland	acid dyes
Cyclanon	BASF, Germany	reactive and direct dyes
Fissatore	Bozzetto, Italy	direct and reactive dyes
Impofix	Impocolor, Germany	direct, reactive, and acid dyes
Levogen	Bayer, Germany	direct and reactive dyes
Matexil	Zeneca, United Kingdom	acid and reactive dyes
Mesitol	Bayer, Germany	acid dyes and cationic dyes for polyamide
Rewin	CHT, Germany	direct, reactive, sulfur, acid, and metal-complex dyes
Sandofix	Sandoz, Switzerland	direct and reactive dyes
Solidogen	Cassella, Germany	direct, reactive, and sulfur dyes
Tinofix	Ciba-Geigy, Switzerland	reactive and direct dyes

5.11. Reducing Agents

In dyeing, reducing agents are used for the following purpose:

1. Reduction of vat and sulfur dyes to form the leuco compound (which has an affinity for the fiber) when dyeing
2. Aftertreatment for improving the colorfastness of textiles dyed and printed with disperse dyes (see Section Aftertreatment Agents to Improve Fastness) by the reductive decomposition of dye adhering to the surface
3. Stripping of defectively dyed textiles (see Section Stripping Agents) by reductive decomposition of the dye
4. In dye protection agents (see Section Dye Protecting Agents and Boildown Protecting Agents), for reducing residual hydrogen peroxide in bleached cotton dyed with reactive dyes

The most important reducing agents are those used in the dyeing of cellulose fibers with vat and sulfur dyes. These reduce water-insoluble dye pigments in the alkaline pH range to form water-soluble leuco compounds with an affinity for the fiber. These are absorbed by the fiber and then fixed by oxidation.

The reducing agents used can be divided into three groups:

1. Sulfur- containing compounds (⟶ Sulfites, Thiosulfates, and Dithionites; ⟶ Sulfinic Acids and Derivatives; ⟶ Sulfides, Polysulfides, and Sulfanes). These are directly or formally derived from the following parent compounds dithionous acid ($H_2S_2O_4$) and sulfinic acid (RSO_2H): sodium dithionite ($Na_2S_2O_4$), hydroxyalkylsulfinates, thiourea dioxide (formamidinesulfinic acid); sulfurous acid (H_2SO_3): sodium sulfite, hydroxyalkylsulfonates; hydrogen sulfide (H_2S): sodium sulfide, sodium hydrogen sulfide.
2. Organic compounds. These include substances with the α-hydroxycarbonyl structure (reductones): glucose, hydroxyacetone.
3. Complex hydrides: $NaBH_4$.

Vat dyes and sulfur dyes pose different requirements on the activity of the reducing agent, and are therefore considered separately.

Reducing Agents for Vat Dyes. Vat dyes are anthraquinoid and indigoid dye pigments and are generally more difficult to reduce than sulfur dyes. They were discovered in 1901 by R. BOHN, BASF but their large-scale use only became possible in 1906 when highly reactive sodium dithionite was successfully produced in anhydrous and therefore storable form. Sodium dithionite then became the most important reducing agent in the textile industry.

The mechanism of the heterogeneous reduction reaction has been thoroughly investigated [395-397].

In the dyeing process, sodium dithionite is consumed by reduction of the dye and also by reaction with atmospheric oxygen. It must therefore either be present from the start in excess or be added continuously. A wide range of techniques have been proposed to reduce these losses [398], including the addition of stabilizing additives [399]. A number of reducing agents less sensitive to oxygen have also been proposed [397]. They include thiourea dioxide and sodium

borohydride [400], which have not been widely used, mainly for economic reasons, but also because of their low activity.

Following increasing environmental pressures, biologically degradable sulfur-free organic reducing agents have been investigated [401]. Hydroxyacetone is significantly more effective than glucose, but cannot match the range of applications of sodium dithionite.

Although many other compounds have been proposed, sodium dithionite is still the dominant reducing agent for vat dyes.

Reducing Agents for Sulfur Dyes. The most common reducing agent for sulfur dyes is sodium sulfide (⟶ Textile Dyeing). Some sulfur dyes can also be reduced with sodium hydrogensulfide. As this has a less alkaline reaction, it is used for dyeing alkali-sensitive regenerated cellulose fibers. Sodium dithionite can cause over-reduction of a number of sulfur dyes, leading to changes in color shade.

Sulfide-free reducing agents are being increasingly recommended, mainly for environmental reasons. These include low-sulfur mixtures of organic reducing agents with sodium dithionite or thiourea dioxide [402], hydroxyacetone, and glucose.

Some commercially available products are listed in Table 37.

Table 37. Some commercially available reducing agents [236]

Name	Producer*	Uses	Composition
Bayreduct PRN	Bayer	peroxide removal before reactive dyeing	inorganic sulfur compound
Cyclanon R	BASF	cleaning of dyed polyester materials	reducton + dispersing agent
Hydrosulfit conc.	BASF	vat dyes	sodium dithionite
Lavan RH	Textilcolor	cleaning of dyed polyester materials	reducton
Lorinol R	Henkel	cleaning of dyed polyester materials	thiourea dioxide
Redoxarin	Rubensperger	sulfur dyes	dextrose
Reduktionsmittel F	Degussa	cleaning of dyed polyester materials	thiourea dioxide
Rongal HT	BASF	vat dyes, high temperature dyeing	preparation of sodium dithionite
Rongal PS	BASF	vat dyes, pad steam	preparation of sodium dithionite
Zetesal UR	Zschimmer + Schwarz	cleaning of dyed polyester materials	glucose derivatives

* in Germany.

5.12. Oxidizing Agents

The oxidizing agents employed in dyeing are mostly the same as those used in textile bleaching (see Chap. Pretreatment Auxiliaries), i.e., hydrogen peroxide, sodium peroxosulfate, sodium peroxoborate, sodium hypochlorite, and sodium chlorite. For special applications, sodium chlorate, sodium nitrite, sodium dichromate, and sodium 3-nitrobenzenesulfonate are also used.

Oxidizing agents are required when dyeing cellulose fibers with vat dyes, leuco esters of vat dyes, and sulfur dyes, and also for the production of aniline black (⟶ Textile Dyeing – Oxidation, ⟶ Textile Dyeing, ⟶ Textile Dyeing – Dyeing with Leuco Esters of Vat Dyes).

When dyeing with vat dyes, the insoluble pigment is regenerated by oxidation. With some dyes and dyeing processes, this can be performed by the oxygen dissolved in the water used for washing, or by hanging the textiles in air. However, it is usually accelerated by adding an oxidizing agent [396]. Hydrogen peroxide and sodium 3-nitrobenzenesulfonate are preferred. The latter can also be used in weakly alkaline solution, enabling washing to be minimized. The color shade of the dye can vary to some extent, depending on the oxidation method used [396].

Sulfur dyes are in generally easier to reduce and hence more difficult to oxidize than vat dyes. Oxygen alone will not cause oxidation to form the insoluble dye pigment. Oxidation of the leuco compounds of the sulfur dyes can in principle be performed by all the inorganic oxidizing agents mentioned above, although color shade and handle may vary [403]. Hydrogen peroxide is preferred for environmental reasons [402].

For developing the leuco esters of vat dyes, where hydrolysis of the sulfuric dye ester in acid solution is coupled with oxidation, the usual oxidizing agent is sodium nitrite.

The oxidizing agent used to form aniline black for “single-bath black” and “oxidation black or steam black” is sodium chlorate.

The most commonly used oxidizing agents for dyeing are readily available simple chemical substances. A few such compounds are marketed as textile auxiliaries under trade names. Some of these are listed in Table 38.

Table 38. Some commercially available oxidizing agents [236]

Name	Producer	Composition
Emectol PAL	Emequimica [a]	NBS[*]
Lamesal P	Grünau [b]	NBS
Ludigol Granulat	BASF [b]	NBS
Lyoprint RG	Ciba-Geigy [c]	NBS
Perlavin OSL	Dr. Petry [b]	NBS
Revetol S	Sandoz [c]	NBS
Rewin SW	CHT [b]	inorganic oxidizing agent + nonionic additives

[*] NBS = 3-Nitrobenzenesulfonic acid, sodium salt.
[a] Spain;
[b] Germany;
[c] Switzerland.

5.13. Stripping Agents

When dyeing and printing, any defective textile products that may sometimes be produced must be rectified for economic reasons. Defects may be due to nonuniform dyeing properties of the textile, process errors (incorrect operation of machines, incorrect use of dyes and chemicals), or soiling during transportation of the material. Before the material can be correctly dyed or printed, the original dye must normally be stripped (i.e., removed or chemically destroyed). Depending on the severity of the defect, the dyed material may be simply brightened, (partial dye removal), or the dye may be chemically destroyed by reduction or oxidation, completely removing it from the fiber.

It is often possible to brighten the material while retaining the color shade and colorfastness by treating the defective product with an auxiliary with an affinity for the dye, and the effect can be intensified by repeating the treatment. The auxiliary with an affinity for the dye, which usually has the same composition as a levelling agent (see Section Levelling Agents), shifts the equilibrium between the dye on the fiber and the dye in the liquor in favor of the liquor.

Chemical decomposition of the dye is carried out with the usual reducing agents, oxidizing agents, and discharging agents used in dyeing and printing. These include common chemicals such as sodium dithionite, sodium hypochlorite, hydrogen peroxide, etc., and products marketed under special trade names (see Sections Reducing Agents, and Oxidizing Agents).

The most important methods of brightening and stripping dyed and printed materials are described below. Precise bath formulations can normally be obtained from the sample books and information sheets of the dye producers. A comprehensive treatment of the subject can be found in [404].

Stripping Dyed Cellulose Fibers. Materials dyed with direct dyes can be brightened by boiling with an auxiliary with an affinity for the dye in a solution containing sodium carbonate. In the case of dyed materials whose fastness has been improved by cationic products, these must first be removed by treatment with strong solutions of formic acid or with diluted hydrochloric acid. Stripping is performed by chemical decomposition with alkaline sodium dithionite, sodium hypochlorite, or sodium chlorite solution.

Materials dyed with reactive and naphthol dyes cannot be brightened (though in exceptional cases materials dyed with reactive dyes can be brightened in a bath containing strong alkali or acetic acid at 60 – 85 °C). The method of destroying reactive dyes is similar to that for direct dyes. Materials dyed with naphthol dyes are stripped with alkaline sodium dithionite solution with an added “reduction catalyst” (anthraquinone).

Materials dyed with vat and sulfur dyes can be brightened in a “blind vat” (alkaline sodium dithionite solution) by an auxiliary with an affinity for the dye. Only in exceptional cases is the dye destroyed with sodium hypochlorite solution.

Suitable auxiliaries with an affinity for the dye include fatty amine ethoxylates and especially polyvinylpyrrolidone (PVP).

Stripping Dyed Wool. Wool dyed with acid or metal- complex dyes can be brightened by boiling with levelling agents and sodium sulfate in neutral or weakly alkaline baths. Almost all dyed woollens can be reductively stripped with zinc formaldehyde sulfoxylate and formic acid in the presence of a levelling agent with an affinity for the fiber (e.g., fatty amine ethoxylate). The stripping effect can be improved by a neutral treatment with stabilized sodium dithionite (e.g., Blankit IN, BASF). As wool is sensitive to alkalis and oxidizing agents, it should only be treated in acidic or, if necessary, weakly alkaline baths, and only with reducing agents.

Stripping Dyed Polyamide. As polyamide is more resistant to alkalis, materials dyed with anionic dyes can be brightened in solutions of sodium carbonate and a levelling agent with affinity for the dye. Disperse dyes can be brightened with nonionic levelling and dispersing agents, or can be stripped in acid solution by reduction with zinc formaldehyde sulfoxylate or oxidation with sodium chlorite.

Stripping Dyed Polyester and Triacetate. Stripping fibers dyed with disperse dyes is very difficult and is only possible at high temperature. These textiles can be brightened by using high concentrations of nonionic levelling agents and carriers, but some fiber damage can occur. The dye can be destroyed by chlorite bleaching with addition of a carrier.

Stripping Dyed Polyacrylonitrile. Polyacrylonitrile fibers dyed with disperse dyes can be brightened by nonionic levelling agents, or stripped by sodium chlorite or zinc formaldehyde sulfoxylate. Basic dyes can be substantially brightened by anionic surfactants. Although most basic dyes can be removed by reductive or oxidative stripping by zinc formaldehyde sulfoxylate, sodium chlorite, or sodium hypochlorite, this causes permanent yellowing of the fiber.

Stripping Dyed Fiber Blends. Methods suitable for sensitive fibers should be used when stripping fiber blends. For example, a waste textile such as reclaimed wool, which is actually a blend, is treated like pure wool (i.e., with zinc formaldehyde sulfoxylate or stabilized sodium dithionite).

Some commercially available products especially suitable for stripping defectively dyed materials are listed in Table 39. For brightening dyed materials, levelling agents can generally also be used (see Section Levelling Agents).

Table 39. Some commercially available products for stripping defectively dyed materials [236]

Name	Producer	Active substances
Albigen A	BASF, Germany	polyvinylpyrrolidone
Anogal KF	Baur & Gäbel, Germany	amine ethoxylates
Decrolin	BASF, Germany	zinc formaldehyde sulfoxylate
Deflavit ZA	BASF, Germany	stabilized sulfoxylate
Emegal A	Emequimica, Spain	polymerization product
Gisapon 1430	Rotta, Germany	alkylamine polyglycol ether sulfate
Perlavin A	Dr. Petry, Germany	polyvinylpyrrolidone
Ritornal AL	Geissler, Germany	polypyrrole derivative
Tinegal W	Ciba-Geigy, Switzerland	alkylamine polyglycol ether
Uniperol AC	BASF, Germany	fatty amine polyglycol ether

5.14. Resist Agents for Dyeing and Printing

In some dyeing operations, resists are employed to prevent dyes from being absorbed on certain parts of the textile, to produce special effects. In the case of fiber blends, it is used to protect components of the fiber that give a different color shade with certain dyes, so that they can subsequently be dyed with other dyes to give the same color shade as the other components.

Resist areas can be produced mechanically (by tying or by covering selected areas with wax, e.g., batik) or chemically by using resist agents. The use of resist agents enables mixture effects to be produced, or, with fibers of different dye affinity, the tone-in-tone effect (levelling agents, see Section Levelling Agents).

In textile printing (⟶ Textile Printing), there are many ways of obtaining resist effects. Resists applied before printing are known as preprint resists, and those applied afterwards, as overprint resists. “White” resists completely prevent dyeing, half-tone resists partially prevent it, and, in colored resists, the printing paste that contains the resist also contains a dye which is itself not resisted, but causes coloration where it is applied. Discharge resists contain a discharging agent as well as a resist.

Not only can a textile be made resistant to dyeing, but it can also be made resistant to the acquisition of other surface qualities, e.g., abrasiveness, a crêpe effect (alkalis), or roughness effects. With polyester fibers, resist printing is mainly used in the production of discharge resists, and, on cellulose fibers, reactive resists.

Resist Agents for Dyeing. Resist agents to give uniform dyeing of fiber blends are mainly used in the dyeing of wool – cellulose, polyamide – cellulose, and wool – polyamide fibers with acid, metal- complex, and direct dyes.

With wool – cellulose and polyamide – cellulose blends, the cellulose fiber is usually dyed with direct dyes, which, with increasing temperatures and dyeing time, are absorbed onto both the wool and the polyamide. In the presence of anionic resist agents (e.g., polycondensation products of aryl- or alkylarylsulfonic acids with formaldehyde), which have a strong affinity for wool and polyamide and in the acid pH range can be absorbed as a colorless dye on these fibers, wool and polyamide can be wholly or partially protected before dyeing with direct dyes. In the tone-in-tone dyeing of wool – polyamide blends, anionic surfactants (e.g., fatty alcohol sulfonates, alkylarylsulfonates, sulfonated castor oil, etc.) retard absorption of the dye by the polyamide fiber, which has an affinity for the dye, and enable the two types of fibers to be dyed to an equal intensity.

To obtain mixture effects (with wool and polyamide), part of the fiber material is treated with a resist agent, mixed with untreated fiber, and then dyed. Suitable resist agents include the same products as those used in the tone-in-tone dyeing of wool – polyamide blends. In the Sando-Space-R process [405], resist effects are produced on wool and polyamide by reducing the dye affinity of fibers by a chemical reaction with a reactive resist agent (e.g., the sodium salt of dichlorotriazine-*p*-sulfanilic acid). A list of commercially available products is given in Table 40.

Table 40. Some commercially available resist agents [236]

Product	Producer	Uses*	Composition**
Alviron RM	Bayer, Germany	WO/PA, PA/CO	PCAS
Erional RF	Ciba-Geigy, Switzerland	WO/PA, PA/CO	PCAS
Levegal ER liq.	Bayer, Germany	PA/CO	PCAS
Levegal ESR	Bayer, Germany	WO/PA	PCAS
Mesitol HWS liq.	Bayer, Germany	WO/PA, WO/CO	PCAS
Perigen PAW	Dr. Petry, Germany	WO/PA	PCAS
Sandospace R	Sandoz, Switzerland	PA	chlorotriazine derivative
Sandospace S	Sandoz, Switzerland	PA	triazine derivative
Thiotan TR	Sandoz, Switzerland	PA, WO, CO	fatty amine polyglycol ether
Thiotan TRN	Sandoz, Switzerland	PA carpets	fatty amine polyglycol ether
Zetesal CRN	Zschimmer + Schwarz, Germany	PA	combination of polyglycol ethers

* CO = Cotton; PA = Polyamide; WO = Wool.
** PCAS = Polycondensates of aromatic sulfonic acids.

Resist Agents for Printing. Resist effects can be obtained by mechanical means or by printing a textile material with a concentrated paste containing a pigment or a resin, followed by dyeing or printing. This type of resist is often combined with one of the many possible chemical resists. Apart from the actual resist agents, the choice of paste plays an important part in resist technology.

In the following examples, some important methods of producing resists for various fibers and classes of dye are described. (The various resist techniques are described in ⟶ Textile Printing.)

Cellulose Fibers [406]. In the case of vat dyes, a printing paste containing heavy metal salts (Zn, Mn) and oxidizing agents (e.g., Ludigol, BASF) is first applied by printing. On overdyeing or overprinting with the vat dye, the heavy metal salts prevent dye fixing, partly mechanically (formation of hydroxides), and partly chemically (combination with the alkali). The oxidizing agent simultaneously prevents reductive development of the dye.

Reactive dyes react with the cellulose in an alkaline medium. This reaction can therefore be prevented by preprinting (or, in two-phase printing, by overprinting) acids or acid-releasing materials, producing white resists. Colored resists are produced by including dyes that are fixed in an acid medium (pigment printing, Naphtol AS dyes). Reactive dyes with a β-sulfoethylsulfonic acid reactive group (Remazol and Basilen F dyes) can be resisted by certain sulfite salts (see Section Auxiliaries for Reactive Resist Printing).

As an example of resists for developing dyes, acidic salts [e.g., aluminum sulfate or tin(II) chloride] and a reducing agent can be printed onto a textile treated with Naphtol AS, preventing subsequent coupling with the diazonium component. Colored resists are made from leuco esters of vat dyes, reactive dyes, or pigments.

The formation of phthalocyanine dyes by the condensation of isoindolenines can be prevented by preprinting with certain primary amines (Phtalotrop B), zinc salts, and organic acids.

Polyamide Fibers [407]. Resist printing of polyamide is used to a limited extent in the carpet sector. Polycondensates of aromatic sulfonic acids are used as resist agents for acid dyes containing sulfonic groups and metal complex dyes.

Polyester Fibers. With discharge resists [408], the textile is first pad dyed with dischargeable dyes and then dried. The discharge resist printing paste, which contains a reducing agent [e.g., Decrolin (BASF), or Rongalit DP (BASF)] and a white pigment or discharge-resistant dye, is overprinted. In the high-temperature process, the dyes are then discharged and/or fixed in a single step. The discharging agent decomposes the dye before it can diffuse into the fiber. This process can produce very detailed patterns with sharp contours.

Polyacrylonitrile Fibers. The discharge resist technique can be used in the printing of cationic dyes onto polyacrylonitrile fibers [409].

Some commerical products are listed in Table 40.

[Top of Page]

6. Textile Printing Auxiliaries

Volker Giesen and Reinhard Schneider

Textile printing auxiliaries are defined as the noncoloring products used at all stages of the printing process from preparation to the final wash [410]. The pronounced diversification of this product group is a consequence of the application of different dye classes [411].

In principle a distinction can be made between printing processes based on pigments, which have no affinity to the fiber, and dyes (reactive, vat, disperse, and others), which have such an affinity. The essential auxiliaries for pigment printing are necessary to adhere the pigments permanently to the fiber; they remain on the fiber and give colorfastness. In the case of dyes with an affinity for the fiber, the auxiliaries are generally removed from the fabric in a final wash.

Unlike other printing techniques pigments can be used on almost all types of textile substrates [412]. It is therefore not surprising that almost 50 % of the world's printed textiles (ca. 19 $\times 10^9$ m/a in the early 1990s) are produced by this process [413].

The advantages of the application of other dye classes in printing are especially noticeable in higher fastnesses, improved handle, and in enabling more demanding design requirements to be met.

6.1. Auxiliaries for Pigment Printing

The idea that “cheap” pigment printing is synonymous with poor quality is today completely unfounded. High-quality pigment printing has now become established as the method of choice for such articles as home textiles, sheeting, apparel, and many others. The great variety of effects from this printing technique (direct printing with pearlescents, spangles etc.; matt-white and matt- colored printing on a colored ground fabric, three-dimensional printing, etc. [414]) enables rapid response to fashion trends.

At least in industrialized countries, it is current standard practice to use solvent-free recipes, which, for environmental, safety and cost reasons, have replaced formulations containing solvents, however, these are still used in other parts of the world [412].

The recipes contain the thickening system, a binder and, if necessary, other auxiliaries (fixing agents, hand modifiers etc.), which are used either to ensure a problem-free production process or to meet the required functional properties of the finished pigment printed article.

6.1.1. *Thickening Agents*

The choice of the thickening agent determines the rheology of the print paste, and hence affects the results of the printing process in many ways. The thickening system has a considerable influence on the printing result as characterized by evenness, sharpness of outlines, color strength, and brillance, as well as on handfastnesses, consumption of print paste, and running properties.

Synthetic thickeners are widely used today in preference to white spirit- containing emulsion thickenings, which contain up to 700 g white spirit per kilogram print paste. During condensation ammonia is liberated, and the polymer is converted to the acid form, generating the slightly acidic pH necessary for cross-linking of the binder and, if used, the fixing agent.

The most commonly used pigment printing thickening agents are liquid, easily pourable preparations of synthetic polymers in a mineral oil. They can be aqueous formulations, completely neutralized with ammonia and with a solids content of ca. 25 %, or anhydrous, partially neutralized products with a solids content up to 60 %. The latter require the addition of a predetermined amount of ammonia to the print paste. In Europe and North America, completely solvent free granulated solid products are also gaining in importance. With these, pigment printing can be a practically emission-free process.

6.1.2. *Pigment Printing Binders*

The binders used in pigment printing are polymers that form clear, colorless, and relatively soft films, and have very little effect on color shades and handle. The binder film should coat the pigment and provide physical adhesion to the substrate, and so protect the printed pigments from mechanical abrasion (rubbing fastness, washing fastness, etc.). The thickness of the film is ca. 10 µm.

Commercially available binders are aqueous dispersions of polymers (mainly based on acrylic esters, butadiene, and, less commonly, vinyl acetate) with solids contents of ca. 40 – 50 %. The amount used depends on the amount of pigment and the textile substrate, and usually varies within the range 50 – 150 g per kilogram of print paste.

The surface active substances needed to produce a stable dispersion are responsible for the compatibility of the binder with the other components of the print paste.

The fastness, i.e., the resistance of the binder film to mechanical stress and its swelling tendency in water and organic solvents, must be increased to an acceptable level by the cross-linking reaction of the binder. Whereas the older types of cross-linkable binders reacted with separate fixing agents during the condensation process (hot air at 150 °C), self- cross-linking binders are now well established. These contain reactive groups, usually from copolymerization with monomers such as *N*-methylolacrylamide or similar compounds. Three-dimensional cross-linking of the binder film can then be achieved by acid catalysis under the usual condensation conditions [415].

Binder films based on butadiene can age by the action of light and oxygen, i.e., they can become yellow, and their fastness properties can deteriorate. Butadiene binders are therefore not recommended for pigment printed textiles that are continuously exposed to light (curtains, awnings, etc.) [416].

Some commercially available products are listed in Table 41.

Table 41. Some commercially available auxiliaries for pigment printing

Product	Composition	Producer

Thickeners (based on acrylic acid copolymers)		
Acraconz BN	liquid	Bayer
Alcoprint PTF	liquid	Allied Colloids
Lutexal HEF	liquid	BASF
Lutexal P	granules	BASF
Polymerconcentrate 475	liquid	Morton Intern.
Texipol 67–5028	liquid	Scott Bader
Self-cross-linking binders		
Acramin ALW	acrylate	Bayer
Acramin BA	butadiene	Bayer
Binder ACM	acrylate	Minerva
Helizarin Binder ET	acrylate	BASF
Helizarin Binder TW	acrylate	BASF
Helizarin Binder UD	acrylate	BASF
Imperon Binder MTB	vinyl acetate	Hoechst
Polybinder HS 359	acrylate	Morton Intern.

6.1.3. *Fixing Agents*

The level of the wetfastnesses (wet rubbing, home laundry, etc.) achievable simply by the use of a binder is often inadequate for many purposes, especially on smooth fibers such as polyester. A significant improvement in fastness is obtained by the use of an additional fixing agent that strengthens the three-dimensional network of the binder film [415]. However, this has a certain disadvantageous effect on the hand.

Melamine – formaldehyde condensates etherified with methanol has proven to be especially suitable, as their reactive groups take part in a cross-linking reaction under the same condensation conditions as those of the self- cross-linking binders.

However, these substances are also the main source of formaldehyde emitted by pigment-printed fabrics [417]. The various limit values imposed today, whether by official regulations, by clothing manufacturers, or by consumer-oriented eco-labels, can be complied with by using modern, modified low-formaldehyde fixing agents of similar chemical type.

Some commercially available products are Acrafix MF (Bayer), Helizarin Fixierer S, Helizarin Fixierer LF (BASF), and Tubiprint Fixierer R (CHT).

Alternative fixing agents of different chemical bases such as aziridines or isocyanates are occasionally used, but these should be treated with circumspection as experience of their use is limited, and toxicological hazards cannot be excluded.

6.1.4. *Hand Modifiers*

Hand modifiers are of two types: silicones and fatty acid esters.

The silicones, mainly poly(dimethylsiloxanes), significantly improve the dry rubbing fastness without decreasing the wetfastness. Used in conjunction with fixing agents, these substances make it possible to produce high-quality pigment printed materials. As well as improving the dry rubbing fastness, the poly(dimethylsiloxanes) give a smooth, dry hand to the fabric, which may be necessary if very soft binders with a somewhat sticky hand are applied.

“True” softeners such as dioctyl phthalate and fatty acid esters cause the binder film to be more mobile. This results in a considerably softer hand. Also, effects on the brillance and color strength (mainly positive) and on the fastness (negative, if too much is used) may be noticeable. Some commercially available products are Acramin Weichmacher SI (Bayer), Luprimol SIG, and Luprimol CW (BASF).

6.1.5. *Emulsifiers*

In high- and low-solvent pigment printing, the emulsifier stablizes the solvent (i.e. white spirit) as an oil-in-water (o/w) emulsion, which functions as a thickened paste with the appropriate rheology. Water-in-oil (w/o) emulsions are nowadays very seldom used. In solvent-free pigment printing, the main tasks of the emulsifier are to prevent agglomeration of the pigment, screen blocking, and separation of components of the print paste. At the same time, the fastness properties and the surface printing must not be affected, as sometimes happens with simple wetting agents.

The aryl- and alkyl-substituted polyglycol ethers are suitable substances, giving a reliable pigment printing process with both solvent- containing and solvent-free recipes. The ethylene oxide and propylene oxide derivatives of various alcohols are also used to overcome special problems, e.g., to improve wetting properties, cleaning properties, etc. Some commercially available products are Emulgator W (Bayer), Luprintol PE New and Luprintol MP (BASF), and Solegal W (Hoechst).

6.1.6. *Other Additives*

Antifoaming agents are discussed in Section Other Auxiliaries.

Acid donors are used in solvent-based printing to obtain the acid pH required for the fixation. They consist of ammonia salts of inorganic acids (e.g., diammonium phosphate).

Compounds are mixtures of several pigment printing additives, and are used to simplify print recipes.

6.2. Auxiliaries for Dye Printing

6.2.1. *Thickening Agents (* ⟶ Textile Printing – Thickening Agents and Thickeners in Textile Printing*)*

Natural thickening agents, both unmodified and chemically modified, are today widely used in printing with dyes, and have largeley replaced emulsion thickenings based on white spirit. Etherified or carboxymethylated starches are often used for printing with disperse dyes, and alginates for reactive printing [418]. More recently, synthetic polycarboxylates, specially developed for reactive printing, have become established. These have improved dye fixation yields [413a] and are easy to use, swelling rapidly in water, and having good resistance to bacterial attack.

Advantages of the alginates include ease of removal in the after-wash, resulting in a printed fabric with a soft hand, and also low sensitivity of the thickening effect towards electrolytes in the print paste.

6.2.2. *Oxidizing Agents*

Sodium m-Nitrobenzenesulfonate. In direct reactive printing on cellulose fiber, the addition of sodium *m*-nitrobenzenesulfonate to the print paste limits harmful effects due to chemical reduction of the reactive dyes by the natural thickening agents and by the fabric itself. This oxidizing agent thus promotes smooth and problem-free printing [419]. Sodium *m*-nitrobenzenesulfonate is also used in combination with natural thickening agents in direct disperse printing on acetate, triacetate, and polyester fibers [420]. It is not recommended for use with polyamide fibers as yellowing can occur.

In the discharge printing of vat dyes on substantive or reactive predyed fabrics, the ground colour can be reductively damaged, both in the printed and unprinted areas. Even traces of print paste containing the reducing-agent, carried from the already printed areas onto the ground color during printing, steaming, or storage of the fabric are sufficient to such damage. To prevent this undesired effect the predyed fabric can be impregnated with a solution of sodium *m*-nitrobenzenesulfonate before printing [410, p. 676]. Some commercially available products are Ludigol, Ludigol granules, and Ludigol liquid (BASF), Matexil PA-L (ICI), and Rapidoprint UL 22 (CHT).

Sodium chlorate is used only occasionally to prevent reductive dye decomposition when printing disperse dyes.

Hydrogen peroxide is used as in dyeing to reoxidize printed vat dyes [421].

6.2.3. *Reducing and Discharging Agents [422]*

In textile printing, reducing agents with a wide range of reducing power are needed. They are used for the fixation of vat dyes in direct and discharge printing, for discharge printing on natural and synthetic fibers, and for afterclearing of printed polyester fibers. Sodium dithionite is used in many applications, but for certain printing processes and equipment the reducing agent must be very stable towards atmospheric oxygen, and in these cases the alkali metal salts of the sulfinic acids are preferable. Some commercial products are listed in Table 42.

Table 42. Some commercially reducing and discharging agents

Trade name	Composition	Producer
Hydrosulfite conc.	sodium dithionite	BASF
Rongalit 2 PH-A	sulfinic acid derivative	BASF
Rongalit 2 PH-B liquid	sulfonic acid derivative	BASF
Rongalit C	sodium hydroxymethanesulfinate	BASF
Rongalit FD liquid	sodium salt of a sulfinic acid derivative	BASF
Rongalit ST liquid	sodium salt of a sulfinic acid derivative	BASF
Decrolin	zinc hydroxymethanesulfinate	BASF
Rongalit DP	stabilized sulfinic acid derivative	BASF
Reduktionsmittel F	formamidinesulfinic acid	Degussa

Sodium dithionite, $Na_2S_2O_4$ is usually known in textile finishing as “hydrosulfite”. It is widely used for the reductive clearing of polyester fibers printed with disperse dyes [423]. Hydrosulfite is a strong reducing agent with limited stability towards atmospheric oxygen. It is therefore unsuitable for direct printing with vat dyes, as severe losses of the reducing agent in the print paste and premature vatting of the dye would occur. It can, however, be used in two-phase printing [423], [424].

Rongalit 2 PH-A/Rongalit 2 PH-B liquid is a combination of reducing agents especially intended for two-phase printing with vat dyes [425]. In this process, the reducing agent is not printed directly with the dyes, but is subsequently applied to the fabric together with sodium hydroxide by an additional padding process. In the second phase the dye is then fixed by steaming (ca. 20 – 40 s) in a flash ager with exclusion of air. Cold solutions of Rongalit 2 PH-A and Rongalit 2 PH-B liquid with sodium hyroxide are less sensitive to atmospheric oxygen than hydrosulfite solutions, and are therefore significantly more stable. Unlike hydrosulfite, the pad liquor of the combination product with sodium hydroxide does not cause premature vatting after padding, and hence does not lead to running of the printed vat dye. The ease of reoxidation of the

printed fabric and the relatively low level of impurities in the used wash liquor give improved operational safety.

Sodium hydroxymethanesulfinate, sodium formaldehydesulfoxylate, $HOCH_2SO_2Na \cdot H_2O$, Rongalit C, is appreciably more stable in alkaline solution than hydrosulfite or the Rongalit 2 PH-A/2 PH-B liquid combination. It is used for printing cellulose fibers by various methods.

It can be used in direct printing or two-phase printing with vat dyes [425]. Because the steaming time is considerably longer (ca. 4 – 6 min) compared with Rongalit 2 PH-A/2 PH-B liquid (20 – 40 s), it is only used in a two-phase process if the ager does not allow rapid dye fixation. The most important application of sodium formaldehydesulfoxylate is in discharge printing [423]. Reactive dyes, and to a lesser extent naphthol and direct dyes, are used to produce the discharge ground.

In color discharge printing, the process must be carried out so that no decomposition of the reducing agent occurs before the dye is fixed by steaming. Great care must be taken to exclude moisture on storage and oxygen during steaming [426]. For white discharge printing, further auxiliaries are added to ensure a pure white discharge effect.

Rongalit FD liquid, a sodium sulfoxylate stabilized with formaldehyde and ammonia, is more stable than sodium formaldehydesulfoxylate. It is used as a reducing agent for direct printing with vat dyes [425]. The improved stability of the printed fabric before steaming is especially important for screen printing on tables. The higher stability requires a longer steaming time (ca. 15 – 20 min). Optimum color yields are obtained by using steam at a higher pressure (1.5 bar, 12 – 15 min).

Rongalit ST liquid is also a sodium sulfoxylate stabilized with formaldehyde and ammonia that is even more stable than Rongalit FD liquid. It is used as a reducing agent in pigment discharge printing. Pigment print pastes which contain Rongalit ST liquid in a slightly alkaline solution are applied to substantive or reactive predyed cellulose fabrics. The decomposition of the ground color takes place during steaming of the printed fabric. A careful choice of the dyes is necessary for this process.

Zinc hydroxymethanesulfinate, zinc formaldehydesulfoxylate, $(HOCH_2SO_2)_2Zn$ in aqueous solution has a slightly acid reaction and is stable for several days. Zinc formaldehydesulfoxylate, unlike the sodium salt, is therefore suitable for discharge printing in weakly acid media. Applications are:

1. White and colored discharge printing on acetate, triacetate, polyester, polyamide, wool, and silk [427], [428]. Only a limited number of discharge-resistant dyes are suitable for printing colored effects.
2. Color discharge printing with pigments on cellulose fibers [429], mainly on substantively dyed grounds.

Rongalit DP is a stabilized sulfoxylic acid derivative that is suitable for discharge printing on polyester [427]. It is mainly used for color discharge resist printing, whereby discharge-resistant disperse dyes are printed together with Rongalit DP. Discharge grounds are unfixed prints or pad dyeings made with dischargeable disperse dyes that are only fixed during printing. On light and medium grounds, white discharge resist printing is also possible, but on dark grounds, zinc formaldehydesulfoxylate should be used, because it results in a purer white effect.

Formamidinesulfinic acid, thiourea dioxide, THDO, is relatively stable in acid solution and unstable in alkaline solution. THDO can therefore be used for printing with vat dyes on wool, silk, acetate, and synthetic fibers. Since the application process takes place in the acid range, the dye is absorbed by the fiber in the form of the vat acid [410, p. 789]. THDO can also be used for color discharge printing with vat dyes on acetate fabrics. This product is now mainly used for the reductive afterclearing of dyed and printed polyester materials.

Tin(II) chloride has several disadvantages, in particular corrosion of the steamer by hydrochloric acid which is liberated during the discharge process. This reducing agent is therefore only used in special applications [427], e.g., for colored discharge printing on polyacrylonitrile fibers. In this process tin(II) chloride cannot be replaced by any other discharging agent [430].

6.2.4. *Auxiliaries for Discharge Printing*

In discharge printing, auxiliaries are used in addition to the discharge agents. They are not able to decompose the ground dye by themselves, but improve the discharging effect and/or facilitate the fixation of the printed dye used for the colored effet. Auxiliaries especially used in discharge printing include:

Sulfonated Dimethylphenylbenzylammonium Salt. This benzylating agent, which is sulfonated in both aromatic rings, enables white discharge prints of indigo and certain vat dyes [426]. The dye, after reduction by Rongalit C, is converted by benzylation into a water-soluble, oxidation-resistant derivative that can easily be washed out [410, pp. 532, 683, 790]. A typical commercial product is Leukotrop W (BASF).

Anthraquinone improves the discharging effect in discharge printing of vat dyes on cellulose fibers dyed with reactive, direct, or naphthol dyes [429]. It is also recommended for discharging of materials dyed with indanthrene dyes and indigo. Anthraquinone acts as a reduction catalyst and is converted to anthrahydroquinone as an intermediate. A typical commercial product is Anthraquinone Powder (BASF).

White pigments such as zinc oxide, titanium dioxide or barium sulfate can also be added to discharge print pastes. These cover the remaining undischarged ground dyes, and produce a plastic effect in patterns with small motives or fine outlines.

Optical brighteners can intensify the white effect in white discharge prints. They are mainly used with polyester fibers.

6.2.5. *Fixation Auxiliaries*

In addition to oxidizing and reducing agents, there are many other auxiliaries that improve the fixation of soluble dyes. The effect of many fixation auxiliaries is to promote solubilization of the dye during the fixation process. In other cases they cause swelling of the fibers.

Fixation Auxiliaries for Printing on Cellulose Fibers.
Hydrotopic auxiliaries. Urea has very good hydrotropic properties and is widely used in printing with reactive dyes [431]. During the fixation of the dyes by steaming, only a small amount of solubilizer, in the form of condensed steam, is present. The hydrotropic properties of urea significantly improve the fixation rate of the dye, which is especially important for regenerated cellulose fibers [410, p. 781].

When printing with vat dyes and during vat discharge printing, thiodiglycol is often added, since it increases the solubility of the leuco form of the dye and hence improves the fixation rate of certain vat dyes [410, p. 781], [429].

Alkalis such as sodium carbonate or sodium hydrogencarbonate are necessary for the fixation of reactive dyes. Alternatively, the special products Basilen Fixing Agent F-RP (BASF) and Remazol Salt FD (Hoechst), which are sodium salts of a chlorinated carboxylic acid, can be used. These salts act as alkali-releasing agents, i.e., the alkali is only liberated during the dye fixation process (steaming) [432].

Fixation auxiliaries for printing on acetate and polyester fibers are mainly fatty acid derivatives and products based on polyglycols [433], [434]. These auxiliaries not only improve the fixation rate, but also simplify the fixation process of the dye. When printing on acetate, the use of fixation auxiliaries enables the fixation time to be reduced from ca. 30 min to 8 – 12 min [435]. With triacetate fibers, suitable auxiliaries enable a dye-fixation in a continuous process at 100 °C with saturated steam instead of the batch pressure steaming process that is otherwise necessary [435].

Fixation Auxiliaries for Printing on Polyacrylonitrile Fibers. Certain fixation auxiliaries containing nitrile groups have proven to be especially suitable for this printing process. They can be used with cationic dyes and some disperse and metal complex dyes [436].

Some examples of commercial fixation auxiliaries are listed in Table 43.

Table 43. Some commercially available fixing accelerators

Trade name	Composition	Producer
For acetate		
Luprintan DCA	poly(ethylene glycol)	BASF
For polyester and triacetate		
Rapidoprint RE	mixture of organic compounds	CHT
Luprintan HDF	fatty acid derivative	BASF
Tanaprint ASD	modified fatty acid amide	Tanatex
Lamefix 680	ethylene oxide adduct to polyglycol ether	Grünau
For polyacrylonitrile		
Luprintan PFD	cyanoethylation product	BASF
Rapidoprint PAN	organic compound containing nitrile groups	CHT

6.2.6. *Dye Solubilizing and Dispersing Agents*

The dyes in print pastes must be in a completely dissolved or very finely divided form, otherwise problems could arise during the printing processes, resulting in uneven prints.

Dye solubilizing agents have the task of dissolving the water-soluble dyes during the preparation of the paste. Many polar organic solvents are recommended and used for this purpose, e.g., ethanol, ethylene glycol, ethylene diglycol, butyl glycol, butyl diglycol, glycerine, and thiodiglycol [410, p. 781].

Thiodiglycol is widely used in textile printing. It is recommended for solubilizing cationic dyes, acid dyes, metal- complex dyes, substantive dyes, naphthol dyes, and leuco esters of vat dyes. The solubilizer can be directly added to the required amount of water. With powder dyes, it is often advantageous to paste the dye with the solubilizer, and to form the solution by adding hot water to this mixture.

Dispersing agents can cause insoluble dyes to become finely dispersed during the preparation of the print paste, and stabilize this state of dispersion. The dispersing agents are usually already present in the dye preparations (e.g., of vat and disperse dyes), so that a separate addition to the print paste is not necessary.

Some commercial dye solubilizing agents are listed in Table 44.

Table 44. Some commercially available dye solubilizing agents

Trade name	Composition	Producer
Lyocol BC	aliphatic ether	Sandoz
Rapidoprint LM	aliphatic ether	CHT
Glyezin A	thiodiglycol	BASF
Tinosol G 133 %	thiodiglycol	Ciba-Geigy

6.2.7. *Auxiliaries for Reactive Resist Printing*

In reactive resist printing, fixation of the ground (vinylsulfone dyes) is prevented at the printed areas by pre- or overprinting with a resist agent (white resist). If the print paste contains monochlorotriazine dyes that are resistant to the resist agent, this is known as a colored resist print. Special sulfur- containing compounds are used as resist agents [437], [438].

Some commercial products are BASF Liquid Reactive Resist Agent (BASF) and Cleantex PWC-N (Kyoei Kagaku).

6.2.8. *Auxiliaries for Burn-out Printing [439]*

Almost the only textiles used in burn-out printing are woven and knitted goods made of polyester – cellulose and polyamide – cellulose fibers. In both fiber blends, the cellulose component is destroyed by acid hydrolysis, i.e., it is burnt out. Sodium hydrogensulfate or the less aggressive aluminum sulfate in concentrations of 80 – 150 g/kg printing paste are used for the purpose. Glycerine is added to burn-out pastes to facilitate their removal by washing. The addition of glycerine is necessary to enable a rapid rewetting of the burned-out areas and easy and complete removal of the decomposed cellulose during the washing process. If yellowing occurs or if the residues are difficult to remove, the textile must be bleached.

6.2.9. *Aftertreatment Agents and Detergents*

When printing with dyes that have affinity to the fiber, the fabric must be afterwashed. The goal is to remove the unfixed dye, the thickening agent, and the auxiliaries from the goods, giving the printed product the highest possible brillance and the required fastness properties [440].

Detergents with a good dispersing action are necessary, especially with water-insoluble dyes such as vat dyes. Detergents with a dye-solubilizing effect are used particularly for washing materials printed with disperse dyes. With printed polyester, the effect of the detergent is enhanced by treatment with the reducing agent hydrosulfite [441]. The addition of polyvinylpyrrolidone to the washing baths of prints made with reactive and direct dyes prevents reabsorption of the dissolved part of the dye. Condensation products of aromatic sulfonic acids with formaldehyde show substantivity to polyamide fibers and are mainly used for the aftertreatment of polyamide textiles printed with acid and metal- complex dyes [442]. They prevent staining of the white grounds during washing of the printed textile, and also improve the wetfastness of the printed dyes.

Some commercial products are listed in Table 45.

Table 45. Some commercially available detergents and aftertreatment agents

Trade name	Composition	Producer
Detergents		
Foryl M 8 liquid	mixture of alkyl polyglycol ethers	Henkel
Levegal AN liquid	mixture of alkylaryl polyglycol ethers	Bayer
Kieralon DB	mixture of substances that are surface active, have an affinity for the dye, or have a dye solubilizing effect	BASF
Printoblanc	mixture of fatty amine polyglycol ethers	CHT
Rucogen FWK	aliphatic ethylene oxide condensation products	Rudolf
Aftertreatment agents		
Albigen A	polyvinylpyrrolidone	BASF
Sandofix WE	methylolamide	Sandoz
Mesitol PS	condensation product of fused aromatic sulfonic acids with formaldehyde	Bayer

6.3. Other Auxiliaries

Preservatives such as formaldehyde are added to natural thickeners to prevent bacterial degradation [418].

Coagulation agents such as borax or aluminum sulfate are sometimes added to the pad liquor for flash-age printing of vat dyes. They prevent bleeding of the dye leuco compound and marking of the printed material on the padding and steaming rollers [443].

Acids such as citric acid or ammonium sulfate are used to produce mildly acid conditions in print pastes when printing with disperse and cationic dyes.

Antifoams prevent the formation of foam during the make-up of the print paste and during the printing process itself. Many such products are available based on, e.g., silicone oils, organic and inorganic esters, aliphatic hydrocarbons, and many other types of compounds.

"Print oils" (mineral oils) are sometimes added to print pastes to reduce the friction between the doctor blade and the roller of a roller printing machine.

6.4. Textile Printing Adhesives [444]

In screen printing, the textile goods are almost always secured by means of an adhesive so that it cannot move during the process.

Water-soluble adhesives are the most common and include those based on natural products such as degraded starch, starch derivatives, or vegetable gums, and synthetic adhesives such as poly(vinyl alcohol) and polyvinylcaprolactam. The latter, which does not contain hydroxyl groups, is used especially with textile goods made of synthetic fibers to which it is difficult to form an adhesive bond.

The water-insoluble adhesives include thermoplastic adhesives, which are acrylate-based water-soluble polymers and have a softening temperature of 50 – 80 °C. They are applied to the print blanket as a solution in a volatile organic solvent. In the printing process, the coated print blanket is heated by special heating equipment, and the goods are pressed onto this by a roller, causing adhesion. Thermoplastic adhesives are permanent glues, i.e., they enable printed fabrics to be bonded many times to the same adhesive film. As well as thermoplastic adhesives, permanent adhesives with a lower softening range are also used. Unlike the thermoplastic adhesives, these already form a strong bond even at room temperature, so that no heating is required to cause adhesion to the goods. They consist of dispersions of acrylic esters. Examples of commercial adhesives are listed in Table 46.

Table 46. Some commercially available adhesives [445]

Trade name	Composition	Producer
Water-soluble adhesives		
Texogum 3013	starch ether	Diamalt
Vianarol ST	poly(vinyl alcohol)	Hoechst
Lubasin RF	polyvinylcaprolactam	BASF
Permanent adhesives		
Kiwotex TDK	acrylate-based copolymerisate	Kissel & Wolf
Lubasin TP	acrylate-based copolymer	BASF
Tubigum DK 2	polymer dispersion	CHT

[Top of Page]

7. Finishing Agents

Rosser Lee Wayland Jr.

7.1. Introduction

Finishing is the name given to those final processes applied to a fabric before it is sold or made into end-use articles, such as garments or home furnishings, which impart desirable end-use properties to it. Finishing may involve both chemical and mechanical treatment. In mechanical treatment, chemical agents are often used to enhance, facilitate, or make durable the effect of the treatment.

Historically three general finishing systems have evolved, based on the processing of the three major natural fibers, cotton, wool, and silk. These general systems differ in the processes and equipment used. The synthetic fibers and their blends with natural fibers are finished in the general system that best fits the fabric type in which they are used. The cotton finishing system predominates and can be assumed in this chapter unless another system is specified.

Cotton is the most widely used textile fiber. World cotton consumption grew at an average of 3.7 % per annum in 1985 – 1990, from 15.8 $\times 10^6$ t/a to just under 19$\times 10^6$ t/a [446]. In 1988 cotton represented ca. 49 % of textile fibers used worldwide, while wool accounted for only ca. 5 %, silk only ca. 0.2 %, rayons ca. 8 % and noncellulosic synthetic fibers ca. 38 %.

In the cotton finishing system, dyeing — either in the yarn or fabric state — normally takes place prior to finishing, but some fabrics may be printed after finishing, especially when pigment prints are used. Finishing may sometimes be carried out on greige fabrics, but in most cases it is preceded by preparation processes such as singing, desizing, bleaching and mercerizing. For most chemical finishes, such as easy care, soil-release, water repellents, and fire retardants, the fabrics must be properly prepared and be free from fiber finishes and contaminants, residual sizes, surfactants, and inorganic salts if optimum results are to be obtained.

Although there have been some attempts to use solvent systems to apply chemical finishes [447], most are applied commercially from aqueous systems. The traditional method for applying finishes has been to pass dry fabric through a bath containing the finishing ingredients, then through a pair of squeeze rolls to squeeze out as much of the treating solution as possible, before passing the fabric into a drying device, such as a tenter, loop, or cage dryer. The dryer is often followed by a high-temperature curing device (e.g., an oven) if curing is needed. Wet pickup may vary from ca. 60 % to 100 %, depending upon the composition and construction of the fabric, and the composition and pressure of the squeeze rolls in the conventional pad-squeeze system. Driven initially by attempts to conserve energy in the drying step, but later by economic and environmental considerations, low wet pickup methods of finish application have been proposed [448]. These include expression methods, in which the fabric is saturated with a relatively large amount of treating solution, after which the excess is removed by special squeeze rolls, a vacuum extractor [449-451] or air or steam jets, and topical methods, in which a limited amount of the treating solution is applied to the fabric by devices such as kiss-roll applicators (e.g., Triatex MA [452]), engraved rolls (e.g., Goller, Sando, Toyoda, and Zimmer), spray applicators (e.g. Burlington, Farmer, Norton and WEKO) and open (e.g., United Merchants and Manufacturing) and closed (i.e., Gaston County – FFT [453]) foam applicators [454]. With these systems pickups as low as 12 % have been reported on 100 % synthetic fabrics. They are particularly effective on polyester – cotton blend fabrics where pickups of 25 – 35 % are usual. In addition to the savings due to reduced energy consumption for drying, the low wet pickup methods also provide savings in chemical costs, improved hand, and less migration of finish. On fabrics containing cotton or rayon, sufficient water must be applied with the finishing solution to swell the fibers and allow the finish to penetrate. In cross-linking only the cross-linking agent within the cellulosic fiber is effective. Any material on the surface or within the capillaries of the fabric or on the noncellulosic fibers present is ineffective in producing easy-care and dimensional-stability effects. It is thus essentially wasted and can contribute to migration, hand, odor, and sewability problems. However, decreasing the wet pickup increases the degree of control required for uniform pickup, thus making the control mechanisms for the low wet pickup systems critical to their success [455]. For this reason the expression methods, which are somewhat easier to control and tend to give more uniform treatment, are more widely used commercially than the topical methods. Other attempts to conserve energy in finishing have included wet-on-wet finish applications and the elimination of washing after finishing, except in cases where it is absolutely necessary, such as with certain fire retardant finishes.

In most cases agents designed to provide the various properties desired in a given fabric are all applied from a single bath rather than from separate baths. In addition to cross-linking agents and catalysis, finishing baths often also contain wetting agents, softeners, hand builders, soil-release agents, weighters, antistats, anti-crock agents, antifoams, optical brighteners, tints, odorants, and other additives. When multiple agents are applied from the same bath it is very important that all of the components be compatible and not cause precipitation or interfere with the function of other components.

7.2. Easy Care Finishes

Easy care finishes are chemical finishes applied to cellulose-containing fabrics to reduce their propensity for wrinkling in the wet and dry state and to stabilize them against progressive shrinkage during laundering. The terms easy care, wrinkle resistant, wash and wear, no-iron, durable press and permanent press have all been applied to this type of finish and although they may denote different levels of performance, they are often used interchangeably. They function by introducing cross-links between the cellulose molecules of cotton and rayon and by so doing reduce the swellability and extensibility of the fibers but increase their resilience. The cross-linking results in a fabric with a “memory” that tends to return to the state it was in when the cross-links were introduced. Thus if a fabric is smooth and flat when cross-linked it will tend to return to a smooth, flat state after wrinkling or laundering. Likewise if the fabric is creased when cross-linked it will tend to retain the crease through wear and laundering.

7.2.1. *Cross-Linking Agents*

Multifunctional methylol derivatives of urea [57-13-6], substituted ureas, or melamine [108-78-1] produced by reacting formaldehyde with these compounds have been used almost exclusively as the cross-linking agents for commercial easy care finishes. The methylol derivatives of urea itself were the first compounds used as cross-linkers and are still used to some extent, particularly on rayon fabrics and where more sophisticated products may not be readily available. They were first used by Tootal, Broadhurst and Lee around 1926 [456]. Dimethylol urea, *N,N′*-bis(hydroxymethyl)urea [140-95-4] is available as a white crystalline material with a formaldehyde to urea ratio of 2: 1. Products having lower formaldehyde to urea ratios (i.e., 1.4: 1 to 1.7: 1) are available in paste form, are more reactive than the dimethylol compound, and tend to form self-condensation polymers as well as cross-links with the cellulose. When methylol ureas are reacted with methanol under acidic conditions some of the methylol groups are converted to methoxymethyl groups accompanied by some polymerization, leading to syrup resins [68071-45-4] that are useful for stiffening fabrics, especially synthetics. All resins derived from unsubstituted urea form chloramides when fabrics containing them are bleached with hypochlorite [457]. Upon heating, as in ironing, these chloramides decompose to form hydrogen chloride, which can seriously degrade the fabric. This phenomenon is known as chlorine retention and can be tested for by AATCC (American Association of Textile Chemists and Colorists) Test Method 92 [458]. Urea – formaldehyde finishes also tend to release formaldehyde from fabrics treated with them and can evolve “fishy” odors due to the formation of methylamines if not prepared properly or if cured with ammonium salt catalysts.

Methylol derivatives of melamine, which can contain up to six methylol groups per molecule, and particularly those in which the methylol groups have been partially or completely converted into methoxymethyl groups by reaction with methanol have been used in textile finishing, either alone [459] or in combination with other cross-linkers [460]. A widely used commercial product has been the dimethyl ether of trimethylolmelamine [1852-22-8]. Melamine derivatives tend to yellow on bleaching with hypochlorite, and fabrics treated with them tend to evolve considerable formaldehyde on storage. They are not used extensively today, except for stiffening synthetic fabrics, where the methoxymethyl products are used, and in special finishes such as fire-retardant and rot-resistant fabric finishes.

The methylol derivatives of many substituted ureas have been used commercially as cross-linkers for easy care finishes. These substituted ureas include ethylene and propylene ureas, triazones, and urons. The ethylene and propylene ureas are made by reacting primary ethylene- and propylenediamines with urea at high temperatures, with the evolution of ammonia [461]. Ethylene urea reacts with formaldehyde to form 1,3-bis(hydroxymethyl)-2-imidazolidinone (DMEU) [

136-84-5]. This product was a widely used cross-linker during the 1950s and 1960s. The dimethylol derivative of propylene urea, 1,3-bis(hydroxymethyl)tetrahydro-2-(1*H*)-pyrimidinone [3270-74-4], has some advantages over DMEU, such as improved freedom from discoloration and improved durability, but it is more expensive and never gained commercial importance [462], [463].

Triazones, such as hydroxyethyltriazone, 5-hydroxy-1,3-bis(hydroxymethyl)hexahydro-*s*-triazin-2-one [1852-21-7], are made by reacting urea with formaldehyde and a primary amine [464], [465]. The triazones were used extensively as cross-linkers during the 1950s, mostly in combination with other cross-linkers such as methylol ureas where they have a “depressant” effect upon chlorine damage due to their basic character [466], [467]. They were abandoned because of their tendency to yellow fabrics, generate amine odors, and evolve formaldehyde.

The urons, i.e., tetrahydro-3,5-bis(hydroxymethyl)-4*H*-1,3,5-oxadiazin-4-one [7327-69-7] are made by reacting 1 mol of urea with 4 mol of formaldehyde and cyclicizing the tetramethylol compound. The urons are usually methylated by reaction with methanol under acidic conditions to form *N,N′*-bis(methoxymethyl)uron [7388-44-5] [468], [469] which is used alone or in combination with other products, such as melamine resins.

The early cyclic urea products greatly alleviated the chlorine retention problem but did little to reduce formaldehyde evolution from fabrics treated with them. Formaldehyde evolution became particularly critical when afterwashing was eliminated from the finishing process for cost and energy savings and because of the advent of durable press in the mid-1960s.

Methylol derivatives of carbamates, especially methyl carbamate [598-55-0], and methoxyethyl carbamate [1616-88-2], which can be produced by reacting the appropriate hydroxy compound with urea at elevated temperatures [470-472] have also been used commercially as cross-linkers and were valued for their lack of yellowing and resistance to severe laundering conditions [473]; however, they evolved considerable formaldehyde.

Many other compounds which are capable of cross-linking cellulose have been proposed as easy care finishes and some have been run commercially on a small scale, but have since been abandoned for cost, environmental, or deficiency reasons. These include polyepoxy compounds [474], polyaziridinyl compounds [475], dihydroxyethylsulfone (Ganalok) [476], polyacetals [477], formaldehyde [478], and several dialdehydes [479], [480].

Methylol derivatives of dihydroxyethylene urea [3720-97-6], produced by reacting glyoxal with urea [481], have replaced almost all of the other products formerly used as cross-linkers in easy care finishes. These products first appeared commercially in the early 1960s and were adopted as the agents of choice for permanent press when this type of finish became popular in the mid-1960s because of their lower formaldehyde evolution potential, both in the cured and uncured state, and their stability in the uncured or partially cured state. Post-cured permanent press, in which the cross-linker is left uncured [482], [483] or only partially cured in the fabric by the finisher and then fully cured in a garment oven after the garment is sewn became very popular in the 1960s, and many garment plants installed garment curing ovens for this purpose. Later it was found that precured fabrics could be reformed on hot head presses to smooth seams and install creases, and use of garment ovens declined. The original glyoxal – urea products were dimethylol dihydroxyethylene ureas (DMDHEU) [1854-26-8]. Later, buffered versions and versions with slightly less than a 2: 1 formaldehyde to DHEU ratio appeared, which offered better fabric whiteness with certain catalyst systems and slightly lower formaldehyde evolution from the finished fabric. Later, products with varying degrees of methylation appeared, such as 1,3-dimethoxymethyl DHEU [3001-61-4] and the fully methylated product [4356-60-9]. Commercial products are commonly 25 – 50 % methylated. The methylated products provide lower formaldehyde evolution from fabrics than unmethylated DMDHEU. They are prepared by reacting DMDHEU with methanol at low pH. Hydroxyl-containing compounds with low volatility, such as glycols or glycerin, when added to or reacted with either DMDHEU or methylated DMDHEU provide even lower formaldehyde evolution potential from the treated fabric. Although specific hydroxy compounds, such as the nitroalcohols [484], have been promoted from time to time, any nonvolatile hydroxy compound will work, and the efficiency of the compound in reducing formaldehyde evolution is directly related to its hydroxyl-equivalent weight. A patent has issued on DMDHEU reacted with glycols [485], but a mixture of DMDHEU with nonvolatile hydroxy compounds is equally effective [481] presumably due to reactions that take place on the fabric during curing. Various formaldehyde “scavengers” have been proposed for reducing formaldehyde evolution [486]. Products based on nitrogen-containing compounds have been proposed but not used extensively on a commercial basis. Hydroxy and active methylene compounds have also been shown to be effective formaldehyde scavengers [486], [487] and are used commercially. Methylated DMDHEU-based products in general afford about one-half the formaldehyde evolution potential of the unmethylated products. Products containing nonvolatile hydroxy compounds, such as glycols, have about one-half the formaldehyde evolution potential of the methylated products. Products that have been both methylated and contain a nonvolatile hydroxy compound provide even lower formaldehyde evolution potential. All of the various modifications of the glyoxal – urea type products are in commercial use. The products which have been both methylated and contain a hydroxy compound afford the lowest formaldehyde evolution potential of all commercially available easy care finishes of the methylol type.

In selecting a cross-linker for use in easy care finishing there are many factors to be considered, including toxicity, cost, durability to laundering, chlorine resistance, effect on hand, discoloration potential, odor, formaldehyde-evolution potential, effect on fabric strength, rate of cure, energy of activation, effect on shade of dyes, effect on lightfastness of dyes, and effect on fabric soiling and soil removal. The glyoxal – urea products have satisfactory ratings on all of these factors, which has made them the products of choice for most easy care finishes worldwide. The driving force for achieving lower formaldehyde-evolution potential from finished fabric has been the identification of formaldehyde as a suspected carcinogen and tightened threshold values for formaldehyde in the workplace. Formaldehyde-evolution potential is influenced by many factors, including the nature of the cross-linker, the amount used, the degree of cure, and the pH of the finished fabric. Many studies have been carried out on factors influencing the evolution of formaldehyde from finished fabrics [488-491]. Numerous other studies have been carried out on test methods for formaldehyde-evolution potential [492-497]. AATCC Test Method 112 [458] for formaldehyde release gives a value that includes both free formaldehyde and formaldehyde released by hydrolytic degradation of the finish and unbound reactant during the incubation step of the test. Thus it is designed to simulate the most severe conditions of high humidity and elevated temperatures likely to be

encountered during the storage of finished fabrics. Other, milder test methods are available, including the Japanese Law 112 Method.

Easy care finishes free of formaldehyde or formaldehyde precursors and thus having no formaldehyde-evolution potential on fabrics have been proposed, but have not achieved significant commercial use [498]. Dimethyldihydroxyethylene urea, 4,5-dihydroxy-1,3-dimethyl-2-imidazolidinone [3923-79-3], made by reacting dimethylurea with glyoxal [499], has been available commercially but has not been used to any significant extent because of its high cost, low effectiveness, and color and odor problems. Improvements which claim to have overcome these problems have been made [500] and there is some commercial use of the improved products.

Polycarboxylic acids such as citric acid [77-92-9] [501][502][503] and butanetetracarboxylic acid (BTCA) [1703-58-8] [504], [505] have been proposed as cross-linkers for easy care finishes, with catalysts such as hypophosphite salts being used to form ester cross-links, but these have enjoyed little commercial success because of their high cost and color problems. Zero-formaldehyde systems based on polyacrylics, polyurethane, and silicones have also been explored but have problems with cost, hand, and performance. It appears doubtful that zero-formaldehyde finishes can compete with DMDHEU-type products unless zero-formaldehyde finishes are mandated by government regulation.

7.2.2. *Catalysts*

Many catalysts have been used to cure the cross-linkers of easy care finishes, but those based on magnesium chloride are most widely used. Almost any acid or acid salt will catalyze the cross-linking reaction, but many such compounds are unsuitable because they are too strong, too weak, or present environmental problems. Zinc salts were formerly used but have been mostly abandoned for environmental reasons. Nitrate salts were used extensively during a period when there was speculation that chloride catalysts might give rise to the known carcinogen bis(chloromethyl) ether [542-88-1], when used with formaldehyde-containing products. This speculation proved unfounded, and chloride catalysts replaced the nitrates again. Magnesium chloride is often used alone, but it can be combined with organic acids such as citric acid, or more acidic salts such as aluminum chloride, to provide “hotter” catalysts which cure the cross-linkers at lower temperatures or in shorter times than pure magnesium chloride. A number of studies have demonstrated the advantages of mixed catalysts [506-509].

Curing conditions vary widely from finishing plant to finishing plant. Some plants use roller ovens while others cure on tenter frames. Curing temperatures normally range from 150 to 200°C, and curing times from 10 s to several minutes with the shorter times being at the higher temperatures vice versa.

7.2.3. *Testing*

The effectiveness of an easy care finish is measured by the washed appearance of the fabric, the washed appearance of creases and seams in a garment, and the shrinkage of the fabric on laundering (AATCC Test Method 123) [458]. Since cross-linking tends to embrittle cellulosic fibers, tests that measure the effects of this embrittelement, such as fabric strength, abrasion resistance, frosting, and dusting, are also usually measured in evaluating easy care finished fabrics. Dyefastness properties, such as lightfastness, washfastness, and crocking, may also be tested to measure the effect of the finish on these properties.

7.2.4. *Hand Builders*

Hand builders stiffen fabrics and give them a firmer, bulkier hand. They are all polymeric compounds. Starches from natural products, such as corn, potatoes, and rice, were formerly used extensively as hand builders and are still used on plain-finish fabrics which are not treated with a cross-linker. In easy care finishes on cellulose-containing fabrics, soluble and emulsion types of hand builders are used. The soluble types tend to give a crisp, “papery” hand to the fabric and include poly(vinyl alcohols) and modified starches (e.g., hydroxyethyl starches). These polymers are insolubilized by the cross-linker used in the easy care finish bath and become durable to laundering.

The emulsion polymers tend to give a thicker, “leathery” hand. Polymers that do not discolor on exposure to light, heat, or atmospheric exposure such as poly(vinyl acetate) and acrylate and methacrylate copolymers are the most widely used emulsion polymer hand builders. The stiffness of the polymer and consequently the degree of stiffness imparted to the fabric is related to the glass-transition temperature of the polymer, with polymers having a higher glass transition temperature being stiffer. Polymers with very low glass-transition temperatures that are very soft can contribute to soil pickup or redeposition during laundering. Some copolymers contain reactive groups, such as methylol amide or epoxy groups, which undergo cross-linking on curing, making the polymers more insoluble and resistant to laundering.

Polymeric methylated urea – and melamine – formaldehyde resins are used as hand builders on 100 % synthetic fabrics, such as nylon and polyester, where they adhere well and give a crisp resilient hand.

7.2.5. *Softeners*

Softeners are used to improve the hand of woven and knitted fabrics and to improve fiber, yarn, and fabric lubricity. The improved lubricity exhibits itself in improved tear strength, reduced needle cutting during sewing, improved fabric drape, and softer feel. Anionic softeners account for <10 % of the total softener market, with the remainder being almost equally divided between nonionic and cationic products. Many softeners are fatty based. For greatest softening effect the chain length of the fatty group should be about C_{18}. The fatty group is connected to a hydrophilic solubilizing group. A great variety of linkages and solubilizing groups are used.

In nonionic softeners the linkage is usually an ester or ether group, and the hydrophilic portion of the molecule may be glycerin, polyoxyethylene, sorbitan, or ethoxylated sorbitan. Nonionic softeners are especially useful where freedom from yellowing is important or where compatibility with anionic components such as optical brighteners is needed. They tend to give surface smoothness to fabrics, without excessive drape. Nonionic softeners are often used as lubricants to assist in napping and other mechanical processes which involve metal – fiber friction.

A wide variety of cationic softeners are available. Quaternary methylammonium compounds containing one or two fatty alkyl groups are used extensively as softeners in home laundering but only to a small extent in fabric finishing. Fatty amides of diamines, such as aminoethylethanolamine [111-41-1], which have been further reacted with acetic acid to form acetate salts have been used as softeners, but these compounds tend to yellow with heat or aging. More complex products formed by reacting the amino amides with ethylene oxide, followed by quaternization with chloromethane or methyl or ethyl sulfate exhibit less tendency to yellow and are widely used. The amino amides may also be cyclicized to form imidazoles before being converted into acid salts or quaternized. Cationic softeners are substantive to cellulosic fibers and can be applied by exhaustion as well as by padding. Cationics are more effective in increasing the drape and limpness of cellulose-containing fabrics than are the nonionics.

Secondary emulsions of low molecular mass oxidized polyethylenes are widely used in finish baths where they function more as lubricants than as hand-modifying agents. They improve the tear strength and abrasion resistance of fabrics and improve sewability but increase seam slippage. Polyethylene emulsions are often blended with cationic softeners.

Dimethylsilicone fluid emulsions are used as softeners and are particularly effective on synthetic fabrics. They impart a unique surface smoothness to fabrics that is different from that given by other softener types. On cellulosecontaining fabrics, organo-modified silicones containing epoxy, hydroxy, or amino groups are often used. The aminosilicones are particularly effective softeners but tend to yellow [510], [511]. They are often blended with fatty cationic softeners, with which they exhibit a synergistic effect.

Anionic softeners consisting of sulfonated oils or fatty esters, often blended with unsulfonated oils, are widely used on plain finish fabrics, such as denims, to facilitate compressive shrinkage of the fabrics. These products act as lubricants as well as wetting agents.

7.3. Water Repellents

7.3.1. *Historical Aspects*

The earliest treatments for water repellency consisted of coating fabrics with various substances which are impervious to water, such as natural fats and oils, waxes, pitch, and asphalt. Later, vulcanized natural rubber became an important waterproofing material. Even later, synthetic polymers replaced the natural products as coating materials because of their superior physical properties. The coated rainwear, often known as “slickers”, was very uncomfortable on extended use because of its impermeability. Coated fabrics are still used for tents and tarpaulins, but fabrics treated with repellents which do not reduce permeability to air and water vapor have largely replaced them for apparel rainwear. Coating materials used for waterproofing are applied as hot melts, latexes, or solvent solutions.

Durable water repellent finishes that allow air and water-vapor permeability began to be used in the mid-1930s with the introduction of reactive pyridinium type repellents. Other types were developed, and by 1990 fluorochemicals dominated the water-repellent fabric market, with ca. 90 % of the value of all products consumed. The value of all fluorochemicals used in the US textile industry in 1990 was ca. $ 100×10^6. The fluorochemicals are followed, in order of importance, by silicones, resin-based finishes, wax-metal emulsions, organometallic complexes and reactive quaternary compounds. All of these types were developed many years ago and there have been few new developments in water repellents since.

7.3.2. *Reactive Quaternary Repellents*

Reactive quaternary water repellents were introduced by ICI under the trade name Velan in 1937, and later by DuPont under the trade name Zelan [512]. These were the first truly durable water repellents. They are based on stearaminomethylpyridinium chloride [4261-72-7]. They are made by reacting stearamide with paraformaldehyde and pyridine hydrochloride. Commercial products contain not only the quaternary but other byproducts and derivatives such as methylol stearamides and methylenedistearamide. The quaternary compounds react with the hydroxyl groups of cellulose on curing, chemically linking the hydrophobe with the cellulose. The bound hydrophobe then tends to retain other unbound hydrophobic materials onto the fabric [513]. The quaternary type repellents are very durable to laundering, even at boiling temperatures. They are also durable to dry cleaning. They are dispersible in water and are applied from aqueous baths, followed by drying and curing at 135 – 205 °C. They impart a soft but sometimes waxy hand to fabrics. Reactive quaternary repellents have fallen into disuse because of the necessity of afterwashing fabrics treated with them to remove byproducts such as pyridine salts.

Quarpel, a finish developed by the US Army Natick Laboratories in 1959 [514], combines a pyridinium repellent with a fluorochemical to give outstanding water and oil repellency for military garments. It has not been used for civilian garments because of its high cost.

7.3.3. *Organometallic Repellents*

Fatty acids, such as stearic acid, form Werner-type complexes with chromium which can be used as water repellents for natural and synthetic fabrics [515]. Commercial products include Quilon C, M, and S (DuPont) and Phobotex CR (Ciba-Geigy). They are neutralized with an amine (e.g., hexamethylenetetramine) or sodium hydroxide and applied by a pad – dry – cure process. Curing at 150 – 170 °C results in maximum repellency. Chrome complexes have fair to good durability to washing and dry cleaning but are sensitive to alkali and some detergents and soaps. They are green in color and can cause a slight color change in the fabric. Aluminum complexes are colorless but are less effective than the chromium complexes. The organometallic complexes, like the pyridinium quaternaries, have fallen into disuse.

7.3.4. *Wax-Based Repellents*

Early processes involving wax emulsions used soaps as emulsifiers and precipitated the soaps as aluminum salts by treating the fabric with alum or another soluble aluminum salt in a second bath. Later, wax emulsions were combined with aluminum acetate or formate in a one-bath process. Still later, zirconium salts were substituted for the aluminum salts [516]. The wax employed is usually a paraffin wax with an emulsifier and a protective colloid such as glue, poly(vinyl alcohol) or an acrylic polymer. The wax – salt repellents are still in use, although their market share has shrunk from its

peak in the 1950s. They are inexpensive and give good water repellency, but they have poor durability to laundering or dry cleaning. The zirconium-based products are more durable than the aluminum-based products and are more widely used. Application, consisting of padding and drying without curing, is very simple.

7.3.5. *Resin-Based Repellents*

Resin-based repellents, produced by condensing fatty materials with methylolated melamines and emulsifying the resulting product, became popular in the 1950s. Many variations of the resin-based repellents have been developed and marketed under trade names such as Aerotex 96 and Permel Resin B (American Cyanamid, now Freedom Chemical), Phobotex FTC (Ciba-Geigy), Nalan W and GM (DuPont), and Norane 16 (Sun, now Sequa). Some of the manufacturing processes for these products are quite complex, as exemplified by Phobotex FTC, which is made by a threestep process involving methylated methylol melamine, stearic acid, diglycerides, triethanolamine, monochlorobenzene, and paraffin wax [517]. The resin-based repellents produce durable water repellency on a broad range of fibers and fabrics. They are less expensive than the fluorochemical, silicone, or reactive quaternary types. They are applied by the pad – dry – cure process, often together with cross-linking agents. They require an acid catalyst and curing temperatures of up to 175°C for maximum repellency and durability. They are normally applied at a rate of 1 – 4 wt % of repellent solids on the fabric. They produce a hand which is soft but firmer than that given by quaternary or silicone repellents. The resin-based repellents have been used as extenders for the more expensive fluorochemical repellents.

7.3.6. *Silicone Repellents*

Silicone repellents were introduced in the early 1950s [518]. These products are methylhydrogen- and dimethylsiloxane copolymers that react in the presence of catalysts to form cross-linked water-repellent films. The silicones are oils and must be emulsified for application. Typical catalysts are zirconium, tin, or aluminum salts, zinc octoate, tin(II) oleate, copper naphthenate, or metal-free compounds such as epoxy amines or amides. They are usually applied by a pad – dry – cure process and have good fastness on synthetic fabrics. On cellulosics they exhibit good dry-cleaning fastness but poor washfastness. Attempts to increase the washfastness of silicone repellents on cotton have been only moderately successful [519]. They are more expensive than the wax – salt repellents but less expensive than the newer fluorochemical finishes. They produce a very soft, silky hand, particularly on synthetics, where they have been found especially useful. Silicone repellents are normally applied at a level of 0.5 – 1.5 wt % solids on the fabric. Curing temperatures of ca. 150 °C are usual. Silicone repellents can be coapplied with durable press finishes. Organo-modified silicones with various reactive functional groups offer the possibility of greater durability; however, these have found more use as softeners than as water repellents.

7.3.7. *Fluorochemical Repellents*

Fluorochemicals were introduced in the late 1950s and have become the most widely used repellents. They are marketed under the trade names Scotchgard (3M), Zepel and Teflon (DuPont), Persistol (BASF), and Repellan (Henkel). They provide both oil and water repellency, due to highly fluorinated alkyl groups. For best repellency at least four fully fluorinated carbon atoms should be present, and the end group should be trifluoromethyl. The fluorochemical repellents used for textiles are mostly copolymers of fluoroalkyl acrylates and methacrylates [520]. The comonomers are esters of acrylic and methacrylic acids containing a variety of alkyl and substituted alkyl groups chosen to modify the physical properties of the polymers, improve performance, and reduce cost. Although the most expensive of the water repellents available they have become popular because of the additional oil and soil repellency which they afford. They repel both water and oil because they produce an extremely low energy surface. They are applied by a pad – dry – cure process with the cure being optional but usually used for the benefit of other finishing bath components such as cross-linkers. They firm the hand somewhat but not enough to make the fabrics unsuitable for most uses. Fluorochemical types are often used with “extenders” which may be other repellents such as the quaternary or resin-based types or even materials normally used as softeners [521]. With the extenders the amount of fluorochemical required can often be reduced by 50 % or more resulting in greatly improved economics.

7.3.8. *Testing for Water Repellency*

Probably the most widely used test for civilian rainwear is the Spray Test (AATCC Test Method 22) [458], in which 250 mL of water is allowed to fall through a spray head onto a taut fabric held at a 45° angle from a height of 6 inches (ca. 15 cm). The pattern of the wetting on the fabric is rated visually against a set of standard photographs. This test measures only the water repellency of the fabric surface and cannot predict the probable rain penetration of water through the fabric. It is, however, especially suitable for mill production control work, because of its speed and simplicity.

In the Impact Penetration Test (AATCC Test Method 42) [458] 500 mL of water is sprayed from a height of 24 inches (ca. 61 cm), and evaluation is by measuring the increase in weight of a blotter clamped under the fabric. This test measures the resistance of fabrics to penetration of water by impact, and thus can be used to predict the probable rain penetration resistance of garment fabrics. The results obtained with this test depend on the construction of the fabric as well as its water repellency.

The Rain Test (AATCC Test Method 35) [458], another impact penetration test, differs from the Impact Penetration Test in that it can be performed at different intensities. It involves directing a horizontal spray from a column of water 2 – 8 feet (ca. 60 – 240 cm) in height onto a fabric backed by a blotter from a distance of 12 inches (ca. 30 cm) for 5 min. Evaluation is by measuring the weight gain of the blotter.

The Hydrostatic Pressure Test (AATCC Test Method 127) [458] measures the pressure required to force water through a fabric and is especially suitable for evaluating fabrics which are expected to be used in contact with water such as heavy ducks and tarpaulins.

The Static Absorption Test (AATCC Test Method 21) [458] and the Dynamic Absorption Test (AATCC Test Method 70) [458] involve immersing the fabric in water under static and dynamic conditions for a specified length of time, blotting the fabric, and weighing the amount of water retained. The results of these tests depend primarily upon the resistance to

wetting of the yarns and fibers in the fabric and not upon the construction of the fabric.

The Bundesmann Test, which is used primarily in Europe, involves showering water on the fabric from 150 cm at a rate of 62 – 68 mL/min for a specified time, while the underside of the fabric is being rubbed, and weighing the amount of water which passes through the fabric or is absorbed by it.

Rain rooms are used primarily by the military for evaluating garments under severe rain conditions. In these tests water is sprayed onto subjects wearing repellent garments, simulating conditions encountered in rain storms. Such tests evaluate not only the fabrics but garment and seam construction.

7.4. Soil-Repellent and Soil-Release Agents

7.4.1. *Historical Aspects*

Soiling and soil removal have always been a problem with textiles, but with the advent of durable press, where polyester – cotton fabrics are treated with large amounts of cross-linking agents and other additives, the problem was accentuated. Durable press fabrics began to be a major factor in the market in 1964, and interest in soil-release finishes increased soon thereafter. Soiling and soil release had been studied extensively before the 1960s, and many research reports were published by workers at Harris Research, ITT, the Cotton Council, the SRRL of the USDA, and others [522], [523]. Many of the early workers realized that hydrophilic polymers such as starch or CMC improve soil removal, and these polymers were used as a laundry finish to improve soil removal but they were not durable. Deering Milliken announced a soilrelease finish based on polymers containing carboxylic acid groups in June 1966, and proposed a test for soil release in December 1966 which later was adapted into AATCC Test Method 130 [458]. A patent was issued to Deering Milliken on their process in 1968 [524]. ICI offered soil-release agents for polyester fibers under the names Cirrasol and Permalose, first in Europe and then in the United States, during the 1960s [525], [526]. During the late 1960s there was feverish activity in the laboratories of textile mills and their suppliers to develop soil-release finishes. Rohm and Haas announced SR-488, a carboxyl-containing acrylic emulsion, in July 1967. 3M introduced a fluorochemical soil-release agent, Dual Action Scotchgard, in late 1967. All through 1967 fabric producers were announcing tradenamed soil-release finishes. The trade names did not imply a particular chemical finish, and there was much changing of formulations as soil-release finishing developed. The advertising, promotion and interest in soil-release finishes probably reached a peak in late 1967 or early 1968. It has been estimated that ca. 50 % of all durable-press fabrics marketed during 1968 contained a soil-release agent [527]. Subsequently, interest in soil-release finishes began to wane, although they are still used on selected fabrics. Some of the factors which contributed to their decline were the high costs of soil-release finishes in relation to their performance, the undesirable side effects exhibited by many of them, such as poor hand, poor compatibility, and poor sewability, the misuse of the finishes, and the fact that the soil-release concept was not valued by consumers. Fabric producers found themselves promoting a concept which the consumer felt should be inherent in any fabric offered for sale. Customers wanted soil release but were unwilling to pay a premium for it. Even after interest in soil-release finishes had peaked, new and improved finishes continued to be offered as research initiated during the peak interest years came to fruition. 3M announced FC-218, an improved fluorochemical soil-release finish, in the spring of 1968 [528] and continued to make improvements in its fluorochemical products into the 1980s. When the Deering Milliken patent [524] on carboxylated acrylic emulsions was issued in April 1968 questions of infringement arose on many finishes then being used. In July 1968 Rohm and Haas announced RL-528, a water-soluble acrylic polymer soilrelease agent, with the claim that it did not infringe the Milliken patent, which included only emulsion polymers. In July 1968 ICI announced Permalose TG, an improved version of Cirrasol PT, which was more durable and required less add-on. Later they introduced Milease T. In 1968 BASF announced Perapret D [529] and promised to indemnify all users against suits for patent-infringement by Deering Milliken.

Soil-release finishes based on the caustic treatment of polyester were developed and promoted, particularly by J. P. STEVENS whose trade name was Fantessa [530]. It soon became generally recognized that any finish that imparts a hydrophilic surface to textiles would improve soil release and numerous products of various types began to be proposed, patented, and marketed. Although there are dozens of such products, few have gained appreciable commercial significance.

7.4.2. *Factors Influencing Soil Release; Testing*

Some of the factors which influence soil release are:

1. Type of soil
2. Amount of soil
3. Size and shape of fabric sample which influences work applied during laundering
4. Type and amount of detergent
5. Mechanical energy of the wash
6. Temperature of the wash

There is a conflict between the optimum conditions for durable press performance and soil release during laundering. Durable press is best at low temperatures and mild agitation, whereas soil release is most effective at high temperatures and with vigorous mechanical action.

The most widely used test method for soil release is the Oily Stain Release Method (AATCC Test Method 130) [458]. In this test a stain is produced on a test specimen by using a weight to force a given amount of the staining substance into the fabric. The stained fabric is then laundered in a prescribed manner and the residual stain rated on a scale from 5 to 1 by comparison with a graduated series of standard stain replicas.

Other tests have been devised to measure facets of the soiling problem, such as redeposition. In redeposition tests, clean samples are laundered with soil or soiled fabrics, and their reflectance is measured before and after the laundry cycle. Fabrics with a propensity for soil redeposition pick up soil in the laundry cycle and show a reduction in reflectance after laundering.

7.4.3. *Soil-Release Finishes*

The most widely used soil-release agents are:

1. Anionic polymers
2. Low molecular weight polyesters
3. Caustic treatments on polyester
4. Fluorochemicals

The anionic polymers are probably the least expensive of these finishes. They may be either emulsions or solutions. They are usually vinyl polymers containing carboxyl groups, although almost any anionic polymer which can be fixed to the fiber surface will work. A variety of acid monomers have been proposed, including acrylic, methacrylic, and maleic acids. Acrylic acid is the most important commercially [531]. A nonacidic comonomer is used to improve the flexibility and durability of the polymer. Lower alkyl acrylates such as methyl or ethyl acrylate are common comonomers. At least 15 % of the carboxylic acid monomer is needed in the polymer for practical soil-release performance. Comonomers such as *N*-methylolacrylamide, which can provide limited cross-linking of the polymer when used in small amounts (1 – 2 %), can add durability to the finish without causing excessive cross-linkage, which can inhibit swelling and decrease the effectiveness of the finish. The anionic polymers are usually supplied in concentrations of 20 – 30 % solids, and up to 15 wt % of the products as supplied is used on the fabric. They are usually applied in the finishing bath along with other chemicals in pad – dry – cure systems. Being polymers they tend to give a firmer hand to the fabric. Some method of fixing the polymer on the fabric must be incorporated into the finish to obtain durability. This can be achieved through an internal reactive monomer component or by reaction with an external cross-linking agent. The anionic soil-release agents present some compatibility problems, particularly with cationic agents such as softeners. They can sometimes interfere with the catalysis of durable press finishes. Other side effects associated with the anionic polymers are metal-ion pickup during laundering, yellowing, and the loss of fastness of pigment dyes.

The low molecular mass polyester types are used mostly on polyester fabrics, where they are claimed to “co-crystallize” with the polyester fiber component. The polyesters contain large hydrophilic polyglycol segments as recurring segments along with short blocks of poly(ethylene terephthalate) units. They are usually supplied in concentrations of ca. 15 %, and up to 15 wt % of the product as supplied is used on the fabric. The products are usually applied to the fabric by padding and are fixed by heating which causes them to be absorbed into the surface of polyester fibers. Alternatively they can be exhausted onto polyester fibers from baths at elevated temperature under conditions similar to those used for dyeing. The solubility of the products decreases with increasing temperature, which aids their exhaustion onto polyester fibers at elevated temperature. This type of finish is not effective on cellulosic fibers. Finishes of this type sometimes have an adverse effect on the fastness of dyes. Cirrasol PT, Permalose T, and Milease T (ICI) are products of this type.

When polyester fibers are treated with strong alkalis such as sodium or potassium hydroxide the surface of the fiber is hydrolyzed, liberating carboxyl and hydroxyl groups, and the resulting fibers have soil-release properties. Auxiliaries such as quaternary compounds are sometimes used with the caustic treatment to enhance the effect. There is always some weight loss associated with caustic treatment and, if carried to the extreme, severe weakening of the fiber can occur. This weakening has been proposed as a cure for pilling on polyester fabrics. Hydrolysis reduces the size of the polyester fibers and leaves the treated fabrics with a soft silky hand. The soilrelease properties of caustic-treated polyester fabrics are durable until the fabrics are heated, as in ironing, after which the soil-release properties are lost. This loss of soil-release properties on heating is thought to be due to the fact that only the surface of the fiber is modified, and that on heating the surface rearranges to bring unhydrolyzed polymer to the surface.

The fluorochemicals used for soil release differ from those used solely for water and oil repellency in that they contain hydrophilic groups as well as perfluoro alkyl groups [528]. The hydrophilic groups are usually polyoxyethylene chains. In the dry state the perfluoro alkyl groups are oriented toward the surface, giving water and oil repellency, while the polyoxyethylene chains are coiled. In the presence of water the polyoxyethylene chains are hydrated, uncoil, and become oriented toward the surface, giving good soil-release properties. These dualaction fluorochemicals were first offered by 3M as FC-216 and FC-218. Other versions have been marketed by 3M and other suppliers. Both nonionic and cationic products are available, as are low-flash-point products containing acetone and high-flash-point products without acetone. The products are supplied at concentrations from 15 to 30 %. Application of 0.3 – 0.4 wt % of active solids on the fabric is recommended. The fluorochemical agents are usually applied with other finish bath ingredients in durable press formulations by a pad – dry – cure system. The fluorochemical types are the most expensive soilrelease agents, but they are probably the most widely used types because of their performance, compatibility with durable press finishes and freedom from deleterious side effects.

7.5. Flame Retardants (see also ⟶ Flame Retardants)

7.5.1. *Flammability of Fibers*

Most fibers will burn, some more readily than others. The limiting oxygen index (LOI), which measures the lowest concentration of oxygen at which a fiber will support combustion, is a good measure of the inherent flammability of fiber types [532]. The LOIs of some common fiber types are as follows:

Acrylic	0.182
Acetate	0.186
Polypropylene	0.186
Rayon	0.197
Cotton	0.201
Nylon	0.201

Polyester	0.206
Wool	0.252
FR Cotton	0.270
Nomex	0.282

The concentration of oxygen in air at sea level is ca. 21 % or 0.210.

Judged solely on fire resistance, Nomex would be the fiber of choice, and it has become the fiber which is most often used when fire resistance is the most important property, such as in garments for firefighters, racecar drivers, steelworkers, and certain military personnel. Nomex is quite expensive, however, and it is difficult to obtain in a complete range of colors. It has not found general usage in apparel or industrial fabrics, except where fire resistance is of paramount importance. Other inherent fire resistant synthetic fibers such as Kynol, polybenzimidazole, and Durette fall in the same category as Nomex.

Cotton and rayon can present serious fire hazards, particularly in certain fabric constructions and in certain end uses. Wool, in contrast, is difficult to ignite and burns very slowly. Thermoplastic synthetic fibers, such as nylon, polyester, and polyolefin, are generally less serious fire hazards than the cellulosics because they tend to melt away from a flame, but they present the additional hazard of molten drip. When blended with cellulosics the thermoplastics represent a greater hazard because the cellulosics form a matrix which retains the molten thermoplastics and prevents them from dripping away from the flame.

Thermal decomposition of fibers, which always precedes flaming, results in gaseous, liquid, tarry, and solid products. As the flammable gases burn, the liquids and tars volatilize, and some of these fractions burn. Other fractions yield a carbonaceous material (char) which does not burn readily. The carbonaceous residue glows until essentially all organic material is consumed and only a fluffy ash remains. Thus there are two stages of burning, flaming and afterglow. The temperature of flaming cellulose is 400 – 450 °C, whereas the afterglow temperature is ca. 600 °C.

7.5.2. *Nondurable Flame Retardants*

Many water soluble salts render cellulosic fabrics flame retardant, until they are laundered or otherwise exposed to water [533]. These are Lewis acids or compounds which give rise to Lewis acids on pyrolysis. Some of the compounds which have been used include diammonium phosphate, boric acid – borax mixtures, aluminum sulfate, ammonium sulfate, sulfamic acid, ammonium sulfamate [7773-06-0], zinc chloride, various phosphate salts, various ammonium salts and combinations thereof. They are applied from aqueous solution by padding or spraying and drying, usually at concentrations of 10 – 15 %. Phosphoric and boric acids and their salts also prevent or inhibit afterglow, as well as serving as flame retardants. While effective, the nondurable types are removed by laundering and thus are suitable only for fabrics, which are seldom or never laundered and can be retreated whenever laundering is done, such as buntings and some draperies.

7.5.3. *Semidurable Flame Retardants*

Cellulose can be phosphorylated by heating it with a mixture of phosphoric acid and urea [534]. The resulting phosphate ester of cellulose is fire retardant, but it picks up anions from laundering to form the sodium, calcium, or magnesium salts, which are not flame retardant. To be effective this finish must be regenerated by reconverting it to the acid or ammonium form after each laundering.

Another semidurable finish which has been proposed is a combination of ammonium phosphate with a dicyandiamide – formaldehyde condensation product, which forms an insoluble resin within cellulosic fibers when applied to the fabric and baked. This finish has a tendency to weaken the fabric if heated above 150 °C. Various modifications of this finish have been proposed.

Cyanamide and phosphoric acid (Pyroset CP) also produce semidurable fire retardancy when applied to cellulosic fabrics, dried, and cured [535]. Optimum results are obtained when the solution contains ca. 3 mol of cyanamide per mole of phosphoric acid. The higher the solids add-on the greater the fire resistance. At 45 % add-on this finish withstands 7 – 10 home laundry cycles, but it is subject to deactivation by sodium, calcium, or magnesium ions from laundry water and must be reactivated by souring and ammonia treatment. The finish causes considerable loss of fabric strength. It has found some commercial use, especially on draperies.

7.5.4. *Durable Flame Retardants*

During the 1960s and 1970s a number of durable flame-retardant finishes were developed after which research on such finishes appeared to decline. Some of the finishes developed, such as APO (tris-aziridinyl phosphine oxide [545-55-1]) [536] which was promoted for cotton, and tris(bromoethyl) phosphate [126-72-7] (TRIS), which was promoted for polyester, are no longer in use because the reactants were found to be carcinogenic. Many of the most successful finishes have been based upon tetrakis (hydroxymethyl) phosphonium chloride [124-64-1] (THPC) or its derivatives. In the Proban process [537] THPC is reacted with urea to form a water-soluble compound that is stable in solution. This compound is applied to fabric in a pad – dry process, then the fabric is exposed to ammonia and ammonium hydroxide consecutively to produce an insoluble polymer within the fibers by reaction of the methylol groups of the THPC with urea and ammonia.

Numerous variations of THPC-derived finishes have been proposed. One of these is the THPC-amide process in which THPC is applied with urea, trimethylol melamine and sodium hydroxide by a pad – dry – cure process. A typical formulation is:

THPC	17 %
Urea	10 %
Trimethylolmelamine	10 %
Sodium hydroxide	1.5 %

This finish generally decreases breaking strength by ca. 30 %, increases stiffness, and is discolored by chlorine bleaches. It is very durable to laundering and increases wrinkle recovery somewhat.

THPC can be converted to THPOH [512-82-3] by treatment with alcoholic sodium or potassium hydroxide. If cellulosic fabrics are padded with ca. 30 % THPOH, dried to 10 – 20 % moisture, then treated with ammonia, a fire-resistant fabric results [538]. This finish produces little strength loss and little stiffness. Its durability is good, but it produces no improvement in wrinkle recovery and it discolors with chlorine bleaches.

If a methylolmelamine is combined with THPOH and an ammonia treatment, followed by curing, a more efficient finish is obtained which has the lowest cost of the THPC family [539]. A typical bath contains ca. 18.5 % THPOH and 11.5 % trimethylolmelamine. The fabric is padded with the finish, partially dried, exposed to ammonia, and finally cured for ca. 3 min at 150°C. This finish shows little strength loss, only a slight increase in stiffness and has good durability. It produces no increase in wrinkle recovery and yellows with chlorine bleaches.

The THPOH – amide finish is similar to the THPC – amide process except that THPOH is substituted for THPC [540]. A typical finishing bath contains 16 % THPOH, 10 % urea, and 10 % trimethylolmelamine. The fabric is padded, dried, and cured at ca. 150°C for 3 min. With this finish about 90 % of fabric strength is retained, with a slight increase in stiffness. Durability is very good, wrinkle recovery is greatly increased, and yellowing with chlorine bleach is very slight.

All THPC and THPOH finishes must be afterwashed to remove unreacted materials and byproducts. THPC has been replaced by the sulfate salt THPS [55566-30-8] because of possible toxicity concerns with THPC, which is no longer manufactured.

Permanent fire retardancy on polyester – cotton blends has been reported for a combination of a specially formulated halogen-free phosphonate with the THPC – urea – ammonia process [541]. The phosphonate alone is claimed to afford durable flame retardancy on 100 % polyester fabrics when ca. 5 % is applied by padding, drying and thermosoling.

Another durable fire retardant which has been used commercially is the methylol derivative of the addition product of a dialkyl phosphite with acrylamide, Pyrovatex CP [20120-33-6] (Ciba – Geigy) [542]. It is used alone or in combination with trimethylolmelamine, which improves its efficiency and durability. Add-ons of ca. 20 – 25 % are required on cotton. It is applied by a pad – dry – cure process. Treated fabrics have tear and tensile strength losses of 20 – 30 %. The finish is durable to 30 – 50 launderings and drycleanings.

Specialty resins, such as methylated thiourea – formaldehyde resins, have been used to flameproof nylon, acetate and acrylic fabrics.

7.5.5. *Flammability Testing*

There are numerous tests for the flammability of textiles. Many of these are specialized tests for specific fabric constructions or specific end uses and have been designed to simulate a particular fire hazard situation. For apparel fabrics the most widely used tests are the 45° angle test called for in US Commercial Standard 191–53, which is now a part of the US Flammable Fabrics Act, and a variety of vertical tests. The CS 191–53 test was not designed for fire-resistant textiles, but was designed to screen out those fabrics which presented an unusual flammability risk as items of commerce. In it a flame is presented to the surface of a fabric held at a 45° angle for 1 s, and if it ignites the time required for the flame to spread and burn through a drop-cord thread 5 inches (ca. 13 cm) above the point of ignition is measured. In the vertical tests a strip of fabric is held in a vertical position and a standard flame is applied to its lower edge for a standard time. If the fabric is not totally consumed the char length of the burned area is measured by applying a standard weight to one side of the fabric to tear it through the charred area, after which the length of the char from the bottom of the fabric to the end of the tear is measured. Some variation of this test is most commonly used for testing fire retardant fabrics. The shorter the char, the more fire retardant the fabric.

7.6. Wool Finishing

The finishing of wool involves both mechanical and chemical processes. Mechanical finishing to develop the hand, drape, and surface characteristics of the fabric is at least as important as chemical finishing.

Softeners, lubricants and water repellents used on wool are similar to those used on cotton.

7.6.1. *Setting*

Crabbing is normally the first process employed in wool finishing. It is a wool setting process in which latent strains in the fabric from earlier fabric forming processes are relieved to avoid distortions such as cockling, crowsfeet, and uneven shrinkage. The process can be accomplished by hot or boiling water or steam on a crabbing machine with only a wetting agent, or chemical agents can be used. Agents based on monoethanolamine sulfite, monoethanolamine formate, and ammonium sulfamate have been used [543]. Mixtures of a reducing agent and an alcohol or an ionic surfactant may also be used.

7.6.2. *Scouring*

The second operation in wool finishing is usually scouring. It is essentially a washing operation in which contaminants

applied or picked up during earlier processes are removed. It has been estimated that 70 % of all processing faults originate from inadequate scouring [544]. Nonionic surfactants, such as alkyl phenol and fatty alcohol ethoxylates are the most widely used scouring agents, although alkyl benzenesulfonates are also used, as are ethoxylated products which have been sulfated, phosphated, or carboxymethylated. Supporting electrolytes such as sodium carbonate, sodium bicarbonate, or sodium sulfate are usually added to the scouring bath.

7.6.3. *Milling*

Milling, fulling, or felting is a process in which wool fibers are progressively entangled by mechanical action to consolidate the fabric structure [545]. Wool fibers felt because of the scales on their surface. Milling is usually carried out in rotary milling machines, although other mechanical forms of milling are used. About 150 % liquor on the weight of the fabric is used in milling. Three types of liquor are used: grease, acid, or soap. In grease milling, sodium carbonate or carbonate/bicarbonate mixtures are used to form soap from the natural greases and spinning oils. Acid milling at pH 2 – 4 is used to minimize color bleeding and to increase the rate of milling. Wool felts least at its isoelectric point of pH 4.8, and felting is slow in the pH range of 4 – 7. In acid milling acid-stable surfactants such as amphoterics and amine oxides must be used. Soaps and synthetic detergents are used in milling to increase lubrication and increase the rate of felting. High-foam surfactants are often preferred for milling. Scouring and milling are sometimes combined. Alkaline milling is used on unscoured fabrics. Increasing the temperature to 40 – 50 °C increases the rate of milling.

7.6.4. *Carbonizing*

Carbonizing is used to remove vegetable matter from wool fabrics. It is based upon the difference in stability to mineral acids shown by animal fibers and vegetable matter. Sulfuric acid is the most common agent used for carbonizing, although other acid substances, such as aluminum chloride, may be used on blends containing fibers which are sensitive to sulfuric acid, such as polyester and acetate. The agent is applied in dilute solution, after which the wool is dried, and baked to degrade the vegetable matter present. After baking, the degraded contaminants are crushed and removed by mechanical action, then the wool is neutralized. Surfactants are added to the acid solutions to promote rapid and uniform wetting. The carbonizing process has been well studied [546-548].

7.6.5. *Shrink-Resistant Finishes*

A number of processes can be used to make wool shrink resistant. One of the processes most used commercially involves subjecting the fibers to a light chlorination, which confers some degree of shrink-proofing, then coating them with a resin such as Hercosett 125, which masks the surface scales [549], [550]. The chlorination – polymer treatments have been in commercial use since the 1970s and account for about 80 % of all machine-washable wool fabrics. Hercosett 125 (Hercules Chemical Co.) is a polyamide – epichlorohydrin polymer capable of self cross-linking. It is applied at 2 wt % solids on the fabric in a bath adjusted to pH 7.5 – 8.0 with sodium bicarbonate or ammonia and exhausted by raising the temperature to 30°C. After treatment, the fabrics or garments are extracted and dried. Nonchlorination processes which utilize polymers alone have been developed. One such process uses Synthappret BAP (Bayer), a reaction product of sodium bisulfite with a prepolymer formed by capping poly(propylene oxide) with an aliphatic diisocyanate, and Impranil DLN (Bayer), an aqueous anionic dispersion of a polyurethane. The polymers are exhausted from a bath containing magnesium chloride at pH 7.5 – 8.0 by raising the temperature from 30 °C to 60 °C over 30 min, after which the pH is increased to 9.0 and the temperature maintained at 60°C for 30 min to cure the polymer [551]. Other polymer-based products and processes have been proposed for shrink-proofing wool [552], [553].

7.6.6. *Mothproofing*

Mothproofing of wool is an important chemical process. One of the most widely used mothproofing agents has been polychloro-2-(chloromethylsulfonamide)diphenyl ether (PCSD) marketed as Eulan 933 and Eulan WA New (Bayer) [554]. Dieldrin, a general-purpose insecticide, has also been used [555]. Because of environmental pressures, mothproofing agents based on synthetic pyrethroid insecticides, such as permethrin, have been introduced [554].

[Top of Page]

8. Auxiliaries for Technical Textiles

Klaus Fischer

8.1. Technical Textiles — an Overview

Introduction. The typical processes of classical textile finishing — sizing, pretreatment, bleaching, dyeing, optical brightening, printing, and finishing — are supplemented by another series of finishing processes in the production of technical textiles: forming (molding), fiber bonding, heat sealing, etc. Because many of the textile auxiliaries and finishing agents [557] are the same as those used for finishing classical textiles, it is not always appropriate to refer to special auxiliaries for technical textiles. However, the emphasis is different. The important auxiliaries are not the optical brighteners, the color shades, or the softening agents that give a silky or other hand, but polymer dispersions for fiber bonding or coating, polymer powders for composite materials, finishing agents for barrier textiles, plasticizers for polymers, glass-fiber coupling agents, cross-linking agents, resins, oleophobic and hydrophobic agents, nonslip agents, antistatic agents, flame retardants, thickeners, light stabilizers, etc. Some technical textiles are treated with sizing agents, but the residual amounts are small compared with the size-free nonwovens, floor coverings, moldings, and composites.

Economic Aspects. The growth in the importance of technical textiles is described in [558]. Technical textiles were first reported separately in German textile statistics in 1988 [559]. The fraction of the total output of textile finishing production used for technical textiles was 19 % in 1992 [560]. A working group, “Technical Textiles”, was set up in 1989 by the association of German textile auxiliary-producing industries (TEGEWA, Frankfurt/Main) [561].

Many problems would be technically insoluble without technical textiles, for example, awnings tire cord fabric, tents, filters, barrier textiles, clothing for spray paint rooms and medical clean rooms, coverings for waste tips, textiles for purification processes, upholstery, fishing nets, and safety equipment for building construction [562].

Application Methods. The auxiliaries are applied by the usual processes, e.g., padding, with nip rollers, spraying, coating and strewing. Wet processes of surface finishing may involve the use of foam. Dry processes are increasingly used, both to apply powders, e.g., for sintering and melting, and in the transfer process.

The well-known conditions for drying and curing are mostly the same as those used in classical textile finishing. To these must be added molding operations and adhesion processes.

Definition. Technical textiles cannot be precisely defined. For example, a textile can be made into classical clothing, or, if it is first provided with a coating, into technical clothing. Abrasive- coated fabrics are technical textiles. Hand-made carpets are archetypal classical textiles. Tufted carpets are used in the home, but are also used as technical products in automobiles. A nonwoven padding, e.g. for an anorak lining, is grouped with the clothing textiles, but a similar product, used as a filter, is a technical textile. A coated material can be used for clothing or as a tarpaulin. This duality makes a sharp division between classical and technical textiles impossible. The roof covering or the glove tray in an automobile and the lattice textile in a facade insulating plaster must certainly be grouped among the technical textiles. Medical textiles are often included with the technical textiles, although this can be regarded as inappropriate.

The importance of technical textiles is indicated in Table 47 [563].

Table 47. Areas of use of textile fibers (%)

Use	Germany 1992		EC 1992		United States 1991	
	Total	Man-made fibers	Total	Man-made fibers	Total	Man-made fibers
Clothing textiles	39	29	47	39	40	25
Home textiles*	36	38	35	38	37	44
Technical textiles	25	33	18	23	22	31

* Including floor coverings.

The term technical textiles refers not only to loose stock, yarn, woven, warp-knitted, or knitted materials, and ropes but also includes nonwovens (⟶ Nonwoven Fabrics), carpets, and textile floor coverings, (⟶ Floor Coverings), shaped textile products, nets, layered yarn structures, bonded textiles, and composites.

Fibers. The range of fibers used in the production of technical textiles includes natural and synthetic fibers, fiber blends, organic and inorganic fibers, polyester and glass fibers, low- and high-melting fibers, high-strength and extra-high-strength fibers, chemically modified fibers, primary and recycled fibers, and normal, waste, and special fibers. Recycled fibers, which have been used in technical textiles for a long time, will become increasingly important [564].

The composite materials [565] produced from certain fibers and fiber – plastics combinations are described in ⟶ Plastics, General Survey – Polymer Composites

8.2. Auxiliaries

There are a number of technical textiles that are not given a final treatment with auxiliaries, e.g., geotextiles made from spunbonded polyester. Here, the technical textile is the end product of the production process. However, many other products only receive their desired properties on treatment with chemical products. Such auxiliaries are indispensable, as the completely finished technical textile can only be produced with their aid.

There are three basic requirement profiles for auxiliaries, which apply to the auxiliary itself, to additives used in conjunction with the auxiliary, and to the conditions under which the auxiliary is applied. Auxiliaries can be used alone or in combination with others.

The tasks of the auxiliary include:

1. To achieve and maintain stabilization of the fiber material during use (e.g., in the production of bonded fleeces), to reduce shrinkage of spunbonds, loosening of the textile structure and fraying, to prevent dust formation, to improve edge strength and resistance to cutting, and to maintain shape during heat treatment
2. Permanence of, for example, strength under various types of stress, e.g., boil washing, water, seawater, moisture, steam, climatic conditions, dry cleaning solvents, light, UV, heat, cold, and oxygen
3. To give repellency effects, e.g., hydrophobic and/or oleophobic properties, and to facilitate the removal of impurities, soil, and dust
4. Prevention of aging

A list of auxiliaries for technical textiles with TEGEWA nomenclature [566] is given below, chapters of [557] in parentheses:

Auxiliaries used alone or as polymer additive

- Polymer dispersions (3.11, 4.5, 4.5.2, 4.18, 4.19, 4.20)
- Polymer powders (4.19)
- Polymer solutions

- Resins
- Cross-linking agents (4.2)
- Catalysts (4.4)
- Nonslip agents (4.14)
- Hydrophobic agents (4.7.1)
- Oleophobic agents (4.7.22, 4.7.3)
- Flame-proofing agents (4.15)
- Antistatic agents (4.6)
- Affinity promoters
- Antimicrobial agents (4.16)

Auxiliaries mainly used as additives

- Plasticizers
- Thickeners (3.12, 4.18)
- Surfactants as emulsifiers (5.1), dispersants (5.3), and foaming agents (5.3)
- Wetting agents
- Antifoams (5.2)
- Light stabilizers
- Pigments
- Carbon black
- Inorganic fillers

Application (Fig. 12). Polymers are applied to textiles in water (hardly ever in solvents) or as dry powders. Alternatively, low-melting textile fibers can bond other textile fibers.

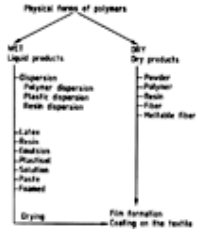

Figure 12. Physical forms of polymers suitable for application to textiles

8.2.1. *Polymers*

Polymer Dispersions. The terms dispersion and latex describe the same form of a polymer [567], i.e., finely dispersed water-insoluble polymer particles in water produced by means of dispersants (emulsifiers, protective colloids; see ⟶ Disperse Systems and Dispersants). Dispersed polymers are particularly suitable for textile applications. The terms “emulsion”, derived from emulsion polymerization (⟶ Emulsions) [568] and “resin” [569] are also used. Typical polymer emulsions include the silicone emulsions used to give textiles hydrophobic properties.

The polymer dispersions for technical textiles are a series of products with many different constituents (Fig. 13) [570].

Figure 13. Components of polymer dispersions

A typical composition of a polymer dispersion [571] is:

47.0 % Copolymer of	monomer I	60 %
	monomer II	33 %
	monomer III (dispersive effect)	1 %
	reactive monomer	6 %
2.0 % Emulsifier (ionic or nonionic)		
0.7 % Production auxiliaries: reducing/oxidizing agents, buffers, pH regulators		
0.1 % Preservative and antifoam		
0.2 % Residual monomer		
50.0 % Water		

A distinction is made between primary and secondary dispersions, the former being obtained directly by polymerization,

and the latter by dispersing solid or molten polymers or resins in a separate operation.

Environmental Aspects. According to the German chemical regulations [572], polymer dispersions are classed as preparations. They do not require special labeling, being classified as a Class 1 (low) hazard to water, and can be eliminated in wastewater purification plants, although not biologically degradable. Their LD_{50} (rat) is generally > 2000 mg/kg. There are no known significant problems associated with the drying or curing of polymers on textiles. For applications covered by the German food regulations, the dispersions recommendation XIV [573] is also supplemented by more recent EC guidelines [574] which apply to the drying of polymer dispersions on textiles to form polymers.

Plastisols [575] are polymer powders dispersed in plasticizers; they are described in ⟶ Poly(Vinyl Chloride) – Plastisols.

Polymer Solutions. Polymers used in the form of aqueous solutions include sizing agents [poly(vinyl alcohols), cellulose ethers, starches, polymers and copolymers of acrylic acid], protective colloids for the production of polymer dispersions, and some dispersants such as maleic anhydride copolymers, surfactants, ethoxylated alkylphenol, ethoxylated fatty alcohol, some finishing agents based on polyacrylamide, textile print adhesives [poly(vinyl alcohol)], and many thickening agents. Their properties include solubility in water, hydrophilicity, viscosity increase, adhesive effects, ease of removal by washing with water, and hardening of the textile (except for surfactants).

Polymer Powders. The use of polymers without water offers significant environmental advantages, although the range of problems that can be solved is more limited than when aqueous systems are used. Emissions are minimized compared with aqueous polymer systems, but control of the addition rate is more difficult.

Monomers. Polymer dispersions can be produced by radical polymerization. The ultimate properties of a textile are closely linked with the polymerizable monomers "concealed" in the polymer (see Table 48). These monomers therefore contribute to such basic properties as hardness, softness, strength, solubility in water, and oleophobicity. Some copolymers are described in [576].

Table 48. Monomers used in the production of the polymers in polymer dispersions

Monomer	Glass transition temperature of homopolymer* T_g, °C			Characteristics of polymer dispersion attributable to monomer
	A	B	C	
Methyl methacrylate	+107	+105	+105	hardness, limited elongation, strength, brittleness
Acrylonitrile	+105	+ 75	+105	thermal stability, hardness, permanence, limited elongation, brittleness
Styrene	+100**	+105	+107	thermal stability, hardness, brittleness, limited elongation
Vinyl chloride	+ 83	+ 81	+ 77	hardness, thermoplasticity, heat sealing, yellowing, flame-proofing, strength
Vinyl acetate	+ 33	+ 29	+ 42	hardness, thermoplasticity, increased adhesion, limited elongation
Dibutyl maleate	0 to –10***			filling effect
Methyl acrylate		+ 8	+ 22	slight hardening effect
Vinyl propionate	– 7		+ 8	filling effect, elasticity
Vinylidene chloride	– 17	– 19	– 18	hardness, yellowing tendency, flame-proofing, heat sealing
Ethyl acrylate	– 24	– 22	– 8	softness, permanence, tearing strength
Ethylene		– 26		softness, adhesion, elasticity
Ethylhexyl acrylate	– 50	– 85	– 58	softness, adhesion
Butyl acrylate	– 55	– 54	– 43	softness, increased adhesion, elasticity
Butadiene	– 78	–102	–102	softness, elasticity
Reactive monomers				
Acrylamide		+165	+220	hardness, ability to cross-link via *N*-methylol groups, limited elongation
N-hydroxymethylacrylamide (*N*-methylolacrylamide)				permanence, limitation of thermal adhesion, limited

			elongation, hardness
N-hydroxymethylmethacrylamide (*N*-methylolmethacrylamide)			permanence, limitation of thermal adhesion, limited elongation, hardness
Acrylic acid	+103	+130	hardness, hydrophilicity, stabilizing, poor wet strength
Methacrylic acid	+230	+162	hardness, hydrophilicity, stabilizing, poor wet strength
Itaconic acid			hardness, hydrophilicity, stabilizing, poor wet strength
Crotonic acid			hardness, hydrophilicity, stabilizing, poor wet strength
Hydroxyethyl methacrylate	+ 55		hardness, hydrophilicity
Glycidyl methacrylate	+ 46		hardness, permanence

* A: various literature sources; B: from [568a]; C: see ⟶ Polyacrylates.
** Hoechst measurement (DSC).
*** Extrapolated.

Auxiliaries in Polymer Manufacture. Polymers contain the auxiliaries used in their production (see Fig. 13). The production of polymer dispersions (latexes) in an aqueous medium normally requires a redox system consisting of peroxides (hydrogen peroxide, organic peroxides) or peroxosulfates, reducing agents (dithionite, bisulfite), polymerization initiators, and dispersing/emulsifying agents. Dispersants have a considerable influence on the ultimate properties of the treated textile, e.g., yellowing, usually due to ammonium salts, or odor from organic peroxides (see Table 49).

Table 49. Effects of dispersants on a polymer dispersion

Surfactants	
Nonionic	coarser particle size, stability with cationic additives, lower capacity for fillers, reduction of surface tension, hydrophilicity not permanent, no effect on hand
Anionic	finer particle size, stability problems with cationic additives, increase in capacity for fillers, improvement of penetration, reduction of surface tension, hydrophilicity not permanent, no effect on hand
Nonionic/anionic	hydrophilicity, increase in salt stability and penetration, reduction of surface tension, hydrophilicity not permanent, no effect on hand
Cationic (seldom used)	coarser polymer particles, thickening, yellowing, compatibility problems with anionic additives, affinity to cotton, increase of surface tension
Amphoteric (seldom used)	stability problems, yellowing
Protective colloids	
Poly(vinyl alcohol)	coarser polymer particles, thickening, reduction of penetration and capacity for fillers, hardening
Cellulose ethers	
Maleic anhydride copolymers	

Dispersants can be protective colloids or surfactants. They emulsify one or more liquid monomers, which usually have low solubility in water, and also disperse the polymer formed in the polymerization [577].

The influence of the individual components of a polymer dispersion on the ultimate properties of a textile are listed below:

Effects due to polymer

- Viscosity
- Drying properties
- Hand: hardness, softness, full, stiffness
- Elasticity, elongation, strength
- Tack, adhesion to the textile
- Swelling by water and solvents
- Resistance to water, washing liquors, and solvents
- Thermal stability, moldability
- Soiling, graying
- Welding properties

- Flame-proofing effect
- Yellowing, stability to light and UV

Effects due to dispersant and emulsifier

- Particle size
- Stability
- Ionicity
- Viscosity
- Compatibility with additives, salts, and fillers
- Spraying properties, foaming properties
- Drying behavior
- Migration
- Color intensification with pigments
- Swelling by water and solvents
- Soiling
- Adhesion/separation from substrate
- Wet strength
- Hydrophobicity, hydrophilicity, oleophobicity

Effects due to polymerization additives

- Yellowing

Effects due to water

- Limitation of stability
- Drying properties
- Volatility of constituents with water vapor

The surface tension of a polymer dispersion system under conditions of use depends on the surfactants or protective colloids used in its manufacture [578]. As a rough guide to determining whether a polymer dispersion can be regarded as hydrophilic or whether it will give poor wetting of the textile, the critical value of the surface tension is ca. 40 mN/m. Finishing agents containing polymer dispersions with a surface tension of > 40 mN/m are less hydrophilic.

A dispersing effect is also shown by carboxyl groups in the polymer derived from copolymerized acrylic, methacrylic, or itaconic acid, and from sulfonate groups in sodium vinylsulfonate. Other components with an important effect on application properties are antifoaming agents and thickeners.

Reactive Monomers (Functional Monomers). The fundamental properties of the polymers can be improved by including reactive monomers in the polymer chain such that the reactivity of the reactive groups is still present and utilizable. Polymer dispersions can provide a technical textile with many useful properties if such reactive groups are incorporated.

Permanence properties (wet strength and stability towards water, washing processes, solvents, and dry cleaning) are improved by copolymerization with the reactive monomer *N*-hydroxymethylacrylamide (NHMAM) (*N*-methylolacrylamide, NMAM) [579].

Other commonly used reactive groups include the epoxy group of glycidyl methacrylate and the hydroxyl group of hydroxyethyl acrylate.

Cross-linking of the polymers applied to the textile is necessary, especially for applications in the technical sector. This enables swelling of the fiber and shrinkage of the textile to be minimized, and high wet and dry strengths to be achieved. Cross-linking of the polymer can be produced by two methods:

1. Externally, by the addition of a cross-linking agent to the polymer, e.g., a *N*-hydroxymethylmelamine, partially or fully etherified.
2. Internally, by means of a cross-linking component copolymerized into the polymer (self- cross-linking). The polymer on the fiber cross-links in the presence of acid catalysts and heat.

For example, the effect of an increase in the percentage of an internal cross-linking agent (NHMAM) in a polymer is shown by the following comparison of the properties of two polymer dispersions (pH 3.6) based on a vinyl acetate – ethylene copolymer with a film glass transition temperature of –25 °C applied to a cellulose spunlaced nonwoven: Increasing the amount of reactive monomer in the copolymer from 3 to 4 % increases the dry tear strength from 3.7 daN to 4.7 daN, and the wet tear strength from 1.2 daN to 2.0 daN.

The 1 % increase in the percentage of reactive monomer also reduces the water swelling properties of the film and increases its resistance towards hot water and perchloroethylene.

Pre- cross-linking. Pre- cross-linked polymers can be produced from bifunctional monomers (dimethacrylic esters of ethylene glycol or of di- and polyglycols, diallyl phthalate, or divinyl benzene) and tetraallyloxyethane, a reaction product of glyoxal and allyl alcohol, which are copolymerized in small amounts during the production process. Pre- cross-linked

polymers are sparingly soluble, difficult to disperse in water, less thermoplastic, and have reduced swelling properties in comparison to chemically analogous but not pre- cross-linked polymers.

Externally and Internally Plastified and Thickened. Polymers that form brittle films as homopolymers can be softened by copolymerization with a monomer that gives a soft film when polymerized (internal plasticization).

Subsequent addition to the polymer of a plasticizer such as a dialkyl (dibutyl or dioctyl) phthalate or dioctyl sebacate, e.g., to PVC powder or by mixing 5, 10, or 20 % into a polymer dispersion, is known as external plasticization (⟶ Plasticizers). With such a product, care must be taken that the plasticizers do not migrate during use of the finished textile, as these can loosen and weaken contiguous materials or coatings.

In a similar dual approach, a monomer with a thickening effect (e.g., acrylic acid) can be incorporated into a copolymer. If this is made alkaline, thickening occurs, and the viscosity undergoes a large increase due to the formation of ammonium, sodium, or potassium salts. Alternatively, the viscosity can be increased by subsequent addition of a thickening agent.

Sensitization. Polymer dispersions are colloidal systems (see ⟶ Colloids). They are coagulated by temperature increase, the mechanical effects of shear stress, or the addition of salts, acids, alkalis, etc. (⟶ Colloids – Heterocoagulation). Polymer dispersions are normally stable up to ca. 80 °C. A dispersion stable up to this temperature can be rendered unstable by using a different dispersing agent during the production process, and/or by means of certain additives, so that controlled coagulation takes place at 50 – 60 °C. The benefit of this sensitization technique is that separation into an aqueous phase and a polymer phase occurs at lower temperatures, so that the water evaporates more readily and the coating dries more rapidly because the formation of a film that can retain water under its surface prevented. Delamination, layer formation, and migration are also prevented.

Migration [580]. Like dyes, polymers can migrate from aqueous phases to the surface of the textile. Dye and polymer migration can be detected by the formation of a layer with intensification of the color. A given amount of pigment can cause stronger or weaker coloration in polymer dispersions whose chemical compositions are similar but which are differently dispersed. In the case of color intensification, the dye migrates to the surface of the polymer, whereas in the other case it becomes distributed over the entire thickness of the coating.

Heat treatment can cause the binder to become concentrated at one side of the textile surface. Also, treatment with washing liquors or solvents can cause binders to swell and then concentrate at the surface. This surface film formation leads to loss of strength in the interior of the textile structure (e.g., a nonwoven) and to hardening at the surface. Also, soil particles that have been picked up by a solvent, e.g., during dry cleaning, can become visible at the surface.

Monomers for Direct Polymerization on the Fiber. For some decades, the possibility of directly polymerizing monomers on the fiber to form coatings and to give fiber bonding has been investigated. However, this has so far not been possible in large-scale textile production, whether by initiation with peroxides, cerium compounds, etc., or by the use of UV, or electron beam radiation, or γ-rays [581]. For later attempts to use electron beams in textile technology, see [582]. The combined use of styrene and perfluoroacrylate leads to oleophobic and hydrophobic effects [583]. The reasons why direct polymerization has so far not been used in practice are as follows:

1. High costs of equipment, shielding, and energy
2. Incomplete polymerization, leaving residual monomers which are difficult to remove from the textile
3. Difficulties in controlling the process to give reproducible results
4. Toxicological problems with monomers in the open conditions of textile production
5. Need for the replacement of cheap monomers by more expensive special monomers of lower volatility
6. Unwanted coloration and graying effects

Glass Transition Temperature. Polymers can be characterized by their glass transition temperature T_g [584]. Although this parameter was formerly of very little importance in textile applications, it has since become an aid to grading the very many different types of polymers used in textiles [585]. For polymer dispersions, T_g is measured on the film produced by drying [586], while powders, fibers, and fiber melts can be measured directly. In the case of coated textiles, an amount of sufficient coating material for a T_g determination can be scraped off. However, it is also possible to carry out a direct determination together with the associated woven or nonwoven substrate [587]. A typical T_g curve for a polymer coating is shown in Figure 14. As well as the T_g of an applied polymer, the T_g and melting point of the fiber substrate can also be determined. Polymer T_g can also be measured in the presence of organic or inorganic flame retardants or fillers provided that these do not themselves have a glass transition range.

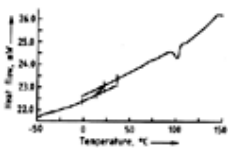

Figure 14. Glass transition temperature curve (T_g ca. 17 °C) of a methyl methacrylate – ethyl acrylate copolymer dispersion dried on polyester spunbond

Liquor bath 500 g/L, drying 120 °C, curing 5 min at 150 °C.

In polymer mixtures, the components can only be differentiated if the glass transition temperature ranges are far apart.

The T_g measurements are of interest because the results for films of polyacrylates, poly(vinyl acetates), analogous copolymers, and styrene and butadiene copolymers show good correlation with the hand of the finished textile. Since low T_g corresponds to soft hand, and vice versa, glass transition temperatures can be ordered to correlate with hand.

For primary or secondary polyurethane dispersions [588] (→ Polyurethanes – Introduction, → Polyurethanes – Two-Component Systems), this correlation between hand and T_g does not apply, although a harsher hand is noticeable for very low T_g compared to poly(vinyl acetate) (PVAC) or acrylics. Whereas a polyacrylate dispersion with T_g = – 31 °C gives a soft hand after impregnation/drying and a soft, elastic, tacky coating, a polyurethane secondary dispersion with the same T_g gives a flexible, relatively hard coating, suitable for coating vertical blinds.

Tackiness in a textile coating means that dust adheres more readily, or unrolling a roll of coated textile is more difficult. For example, a textile coated with a 70/30 styrene/butyl acrylate copolymer dispersion (T_g +50 °C) exhibits no dry soiling in a dust- containing air after 7 d open storage, while a 70/30 butyl acrylate/styrene (T_g – 8) coating leads to significant soiling of the textile in a dust- containing atmosphere after 7 d. The T_g also correlates with soiling behavior.

Glass transition temperatures and minimum film-forming temperatures of polymer dispersions are listed in Table 50 [589].

Table 50. Minimum film-forming temperature (MFT, white point temperature) according to DIN 53 787 and glass transition temperature (T_g) of polymer dispersions measured by DSC

Dispersion film	Use	MFT, °C	T_g, °C
Polystyrene	stiffening shoe toecaps	>100	98
Butadiene – styrene – acrylonitirile copolymer	moldable floor coverings	90	112; –15*
Methyl methacrylate – acrylate copolymer, pre-cross-linked	hard finish	90	103
Acrylic acid copolymer	thickener	50	65
Styrene – butyl acrylate copolymer, self-cross-linking	glass textile finish	60	50
Perfluoroacrylate polymer	oleophobizing		47; *mp* 74.5
Acrylate copolymer, self-cross-linking	nonwoven binder	29	38
Poly(vinyl acetate)	stiffening agent	15	33
Acrylate copolymer, self-cross-linking	nonwoven binder	10	24
Vinyl acetate – dibutyl maleate copolymer	handle modifier	12	19
Vinyl acetate – ethylene copolymer	carpet backing	0	5
Vinyl acetate – ethylene copolymer	carpet backing	0	–6
Butyl acrylate – styrene copolymer, self-cross-linking	glass textile finish	0	–8
Vinyl acetate – butyl acrylate copolymer, self-cross-linking	pigment binder	0	–12; *mp* 44**
Polyurethane	pigment binder	0	–30; *mp* 52

* Polymer mixture.
** Caused by surfactant; no longer present after heating twice.

Various polymers (e.g., those with perfluoro groups) [590] and certain surfactants have melting points, and are therefore at least partially crystalline. A nonylphenol ethoxylate with 30 ethylene oxide units has a melting point of ca. 50 °C, measured by DSC. This melting point can also be measured when the surfactant is used as a dispersant (Fig. 15).

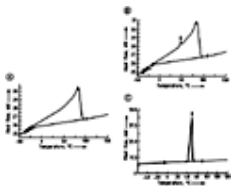

Figure 15. Glass transition temperature curve of an ethylene – vinyl acetate powder with 2 % surfactant (nonylphenol ethoxylate, 30 EO)

A) Polymer without surfactant: T_g – 29.5 °C, peak 81.7 °C; B) Polymer with surfactant on PES spunbond: T_g – 30.6 °C, peaks 47.2 °C and 83.0 °C (PES); C) Surfactant without polymer: peak 49.1 °C

No correlation is observed between the thermal resistance of a textile material and the T_g with which it is finished. Thus, while one compression molded textile treated with a polymer with a T_g of, e.g., 100 °C, may have a stable shape when the textile is later heat treated at 100 °C, another polymer of different chemical composition but with the same T_g may not give the same thermal stability when applied to the same textile material.

Uses of polymer dispersions are as follows:

Textile floor coverings	backing of woven and tufted carpets, secondary adhesion backing of tufted carpets impregnation and backing of needle punch
Nonwovens	
Nonwoven wadding	spray stiffening
Flat nonwoven	impregnation
Needle felt	coating
Spunbond	bonding
Spunlaced	lamination
Melt blown	heat sealing
	thermobonding
	flocking
	pressing
	printing
Glass textiles	
Woven and nonwoven	impregnation
lattice	coating
	finishing
Glass fibers	sizing
Woven, knitted	impregnation
fabrics, lattices	coating
	finishing
	printing
	flocking
	bonding, laminating
	hot sealing
	metallization
Ropes, nets	impregnation
Deposited materials	coating
Textiles bonded to paper,	bonding, adhesion
foil, plastics, metal	metallization
Loose stock	spray stiffening
Recycled fibers	impregnation
Mineral wool	
Flock	bonding in coating

Binders include nonwoven binders [591], pigment, printing binders [592], or dye binders, flocking binders [593], and binders for abrasives. Coating materials, either smooth or foamed, are used as carpet backings [594] or for coating textiles [595]. Adhesives [596] are used to bond textiles or coated textiles to other materials, or are termed flocking adhesives when they bind flock in the coating materials.

Technical effects achieved by auxiliaries include:

1. Surface effects, e.g., gloss, luster, roughness, oil, water, and soil repellency, surface strengthening of waddings, coatings, antipilling, and adhesive effects
2. Penetration of full bath impregnation to give through hardening of moldings, stiffness, water impermeability, and prevention of delamination, etc.
3. Incorporation of fibers or textiles into polymers in the production of fiber-reinforced materials.

As an example, the effect of using various polymer dispersions on various fiber materials can be seen in Table 51.

Table 51. Comparative effects of polymer dispersions on the technical textiles polyester spunbond and woven glass textile used for bituminized roofing material (application conditions: padding, drying at 120 °C, curing for 3 min at 170 °C)

Formulation	*T*	Amount	Glass textile	Spunbond

	g/°C		Transverse tearing strength, N	Transverse elongation, %	24-h absorption height, cm	Transverse tearing strength, N	Transverse elongation, %	24-h absorption height, cm
Vinyl acetate – dibutyl maleate copolymer dispersion, 55 %*	19	200 g/L	1230	1.7	6.5	300	50.2	0
Starch derivative		40 g/L						
Acrylic ester – styrene copolymer dispersion, 50 %**	28	300 g/L	890	1.7	10.5	326	47.5	0
Untreated			no stability	no stability	7.0	225	770	0

* Typical formulation for glass textile without hydrophobic agent.
** Typical formulation for spunbond.

Suitable Products for Various Technical Textiles. A list of selected technical textiles and suitable polymer dispersions is given in the following:

Polyester spunbond for bituminized roofing	hard thermally stable film formers based on acrylic ester – methacrylic ester or acrylic ester – styrene and/or acrylonitrile
Glass woven textile for bituminized roofing	hard film formers based on vinyl acetate or styrene, methacrylic ester/acrylic ester
Webbing belts	soft film formers based on vinyl acetate – ethylene or acrylic ester
Automobile carpets for final powder coating and molding	medium-hardness film formers based on vinyl acetate – ethylene
Backing for automobile carpets	medium-hardness film formers based on butadiene – styrene
Adhesion	vinyl acetate homopolymers, vinyl acetate copolymers with ethylene or acrylic esters, styrene – butadiene
Pressure-sensitive adhesion	vinyl acetate – ethylene or acrylic ester copolymers with very low T_g
Technical articles made by flocking	soft film formers based on acrylic esters, acrylic ester – acrylonitrile, butadiene – acrylonitrile, sometimes with added melamine cross-linking agent
Glass textile wall-coverings	tough elastic film formers based on vinyl acetate – ethylene with flame-proofing components
Glass textile structural lattice	medium hardness film formers based on styrene – acrylic ester or vinyl acetate – ethylene
Milk filters	soft film formers based on acrylic ester or vinyl acetate – ethylene
Technical nonwovens	medium hardness film formers based on vinyl acetate – ethylene, acrylic ester, or acrylic ester – styrene
Hard coating of technical woven textiles	styrene – acrylic ester copolymers, methyl methacrylate ester copolymers
Wadding, quilting	hard film formers and medium hardness binders with no surface tack based on acrylic/methacrylic ester and vinyl acetate – acrylic ester or vinyl acetate – ethylene
Nonwovens for cleaning purposes	soft film formers based on vinyl acetate – ethylene or acrylic ester
Needle felt moldable for automobiles	hard film formers based on butadiene – styrene – acrylonitrile or styrene – acrylic ester
Technical knitted goods	soft film formers based on vinyl acetate – ethylene
Back coating of technical	soft film formers based on acrylic ester, acrylic ester – styrene, vinyl

warp-knitted plush	acetate – acrylic ester, vinyl acetate – ethylene
Nonwoven filters	soft film formers with surface tack based on vinyl acetate – ethylene, acrylic ester or butadiene – styrene
Sails	combination of hard acrylic/methacrylic ester binders with melamine cross-linking agents

Some practical formulations (each case, application of the auxiliary is followed by drying and curing) are as follows:

Sizing formulation for abrasive fabrics	4 % poly(vinyl alcohol), partly hydrolyzed 4 % carboxymethylcellulose, low viscosity 0.3 % size plasticizer 0.1 % lubricant
Thermobonded polyester nonwoven, slop padded with a rapidly breaking dispersion foam in the nip roller process	100 parts polyacrylate dispersion 50 %, T_g ca. – 35 °C
Impregnation of a polyester spunbond to give a hard product for bitumen coating	500 g/L polyacrylate dispersion 50 %, T_g ca. 30 °C
Cleaning cloth with soft hand of viscose nonwoven, nip roller process	100 g/L ethylene – vinyl acetate copolymer dispersion 50 %, T_g – 18 °C
A fiber blend needle felt, nip-rolled, dried, and molded in a press at 170 °C (hard molded articles for automobiles)	500 g/L copolymer dispersion mixture based on styrene – butadiene – acrylonitrile 50 %, T_g ca. 112 °C and – 17 °C
Hardening of nonwoven filter of polyester fibers by spraying the binder (hard bonding points are formed)	250 g/L copolymer dispersion based on vinyl acetate – butyl acrylate 45 %, T_g ca. 32 °C
Spray on an high-volume nonwoven	160 g/L polyacrylate dispersion 60 %, with T_g 40 °C 40 g/L polyacrylate dispersion 60 %, with T_g 12 °C
Coating for a cotton textile, medium hard hand	75 parts polyacrylate dispersion 45 %, T_g ca. 25 °C 25 parts water 15 parts titanium dioxide 0.1 parts pigment 2.5 parts thickener 30 %, viscosity 6000 mPa · s
Soft-elastic coating on woven fabric	as above, with polyacrylate dispersion 50 %, T_g – 9 °C
Impregnation of vertical blind of Trevira CS, stiffening without reduction of flame-retardant properties	500 g/L polyurethane secondary dispersion 35 %, T_g ca. – 31 °C
Coating a fabric for later flocking	100 parts polyacrylate dispersion 45 %, T_g – 15 °C 0.2 parts antifoaming agent 8 parts *N*-methoxymethylmelamine cross-linking agent 50 % 0.8 parts amine hydrochloride catalyst 4.5 parts thickener 30 % ammonia to adjust pH to ca. 8, viscosity ca. 6000 mPa · s
Finishing of glass textile wall coverings, impregnation	300 g/L copolymer based on vinyl acetate/ethylene/vinyl chloride 50 %, T_g 15 °C 50 g/L modified starch 25 g/L flame retardant, e.g., organic bromine compound
PVC coating with plastisol on polyester nonwoven	100 parts PVC powder 65 parts plasticizer dioctyl phthalate 3 parts coplasticizer 1.5 parts stabilizer 10 parts chalk 5 parts titanium dioxide
Hot sealing [597]: sprinkling with a sifter,	100 parts copolymer powder based on

sintering, pressing for 15 s, at 5 bar and 116 ° ethylene – vinyl acetate, *mp* 87 °C
C in the sealing gap, 127 °C in the press

Polymer powders and hot melts are characterized by their melting behavior (melting point, glass transition temperature, melt index, and melt viscosity; see Table 52), adhesive properties, resistance to water and solvents, and the high-temperature strength of adhesive bonds. See ⟶ Adhesives.

Table 52. Hot melt adhesives of various chemical types

Composition	*mp*	DSC measurements		Melt flow index (MFI) [599], 160 °C/2.16 kp, g/10 min
		Peak*, °C	T_g, °C	
Polyethylene (LDPE)	107	109		22 (190 °C)
Polyethylene/vinyl acetate				
3 % VAC	102	104		1.1
		104**		
19 % VAC	87		–30	6.3
28 % VAC	70	52	–30	14.5
		62**	–34**	
Polyamide	122	103		16
		77		
		–**	+23**	
Polyamide	135	127**		16
		73**	+21**	
Polytetrafluoroethylene	370 – 375			
Tetrafluoroethylene copolymer	350 – 370			

* DSC after first heating.
** DSC after second heating.

A high melt flow index indicates that the polymer has a low melt viscosity, and melts more readily at a low temperature. The polymer powders used as hot melt adhesives [598], [600] give a strong adhesive bond which depends on the strength of the textile, the strength of the interface between the textile and the polymer adhesive (adhesion), and the cohesive properties of the film of adhesive (cohesion). It is also dependent on the method of applying the adhesive and the thickness of the textile. Adhesive bonds are also formed between a plastic film (PVC, polyurethane, polyester) and a textile.

Polymer powders are available in various particle sizes: 0 – 80, 80 – 200, 200 – 300, and 300 – 500 µm.

The polymers with the largest market share in Western Europe as hot melt adhesives are those based on ethylene – vinyl acetate.

Powders in aqueous suspension are also used [598]. The process used to disperse the large polymer particles must be carefully chosen, as surfactant-type dispersants can weaken the bond to the textile. This is also true of polymer powders that have been given into polymer dispersions [597]. If a powder and a polymer dispersion are used together, the polymer dispersion can help to give uniform distribution, and the cohesive strength of two so-bonded fabrics is increased.

Example: The cohesive strength at 120 °C of a paste coating (100 g/m^2) with polyacrylate dispersion, thickener, and alkenesulfonate wetting agent, pressed at 170 °C and 25 bar [601] is as follows:

Polymer disperion 50 %, T_g 18 °C (100 parts)	187 N/m
Polymer dispersion (100 parts) + copolyamide powder (30 parts)	377 N/m
Copolyamide powder, T_g 21 °C, *mp* 127 °C; 73 °C (100 parts)	430 N/m

Hot melt fibers for textiles are described in [602].

Trade names are given below:

Platamid, Platabond, Rilsan (Atochem, FRG), Levasint (Bayer, FRG), Abifor (Billeter, CH), Riblene (Enichem, I), Isatherm (Fuller, USA), Griltex (Ems, CH), Guttacoll-Klebefolien (Gutacoll, FRG), Hostalen, Hostalit, Hostaflon, Synthacryl (Hoechst), Coathylene (Hoechst-Celanese/Plastlabor, CH), Mybond (Mydrin, GB), Vestamelt, Vestoplast (Hüls, FRG), Permatex (Permatex, FRG), Schaettifix (Schaetti, CH), Sarpifan, Estekoll (Stockhausen, FRG)

Thermoplastics, Thermosets, and Elastomers [603]. Many polymers applied to textiles have thermoplastic properties. They range from soft elastic coatings (T_g ca. – 15 °C) up to articles with hot strength (e.g., automobile fittings) and T_g ca. 100 °C. Melamine, urea, phenol – formaldehyde precondensates, and epoxy resins are, in contrast, thermosets [604].

Phenol – formaldehyde [605] and melamine – formaldehyde precondensates are used in powder form or in an aqueous medium to produce shaped, high-strength, heat stable, press molded articles made of textile fibers for the automobile [606]. The precondensates, phenol – formaldehyde novolacs, harden at 170 – 180 °C on the textile to give resin structures. Melamine – formaldehyde precondensates can be used either alone or together with polymer dispersions, but phenol – formaldehyde precondensates are often used alone. Glass transition temperatures in the range 60 – 90 °C can be measured in the precondensates, but not after the resin has been formed.

Polytetrafluoroethylene is neither a thermoset, nor a thermoplastic, and its stability at high temperatures can be utilized, e.g., in combination with glass fibers. Some thermoplasticity can be obtained by copolymerization [607].

Starch is also suitable for technical textiles. It is a macromolecular carbohydrate with a high decomposition temperature (> 230 °C) and thermoset properties, stabilizing at high temperatures, having a hardening and thickening effect, and is biologically degradable and compatible with other auxiliaries. Starch and modified starches are therefore standard constituents, e.g., for finishes for glass textile wall coverings, to produce stiffening effects, and to reduce thermoplasticity, often in combination with polymer dispersions [608].

Pretreatment. Pretreatment processes to optimize textile properties for technical use include not only the classical processes of mercerization, scouring, bleaching, and washing, but also various methods of modifying the surface of fibers [609] such as roughening (abrasion [610]), treatment of polyester fibers with alkali, etching, treatment with affinity promoters (e.g., for tire cord [611]) for fiber-reinforced plastics and composites [565], and glass fibers and for metallization [612]. Other surface modifications include coatings or sizings for glass fibers [613], corona treatment [614], plasma treatment [615], laser treatment [616], irradiation [582], and the reaction of fibers with reactive products, e.g., 2-aminoethylsulfuric acid [617], which reacts with the hydroxyl groups of cellulose.

The penetration of the finishing agent into the textile is increased by these pretreatment processes, as is the adhesiveness of the bond between the fiber and the dried/fixed auxiliary.

Finally, appropriate pretreatment and selection of textiles is needed to give compatibility with a coating applied by spraying on the back of the textile [618], e.g., in an automobile.

The following materials can be used either as independent finishing agents and auxiliaries or as additives [619], especially to polymer dispersions [620].

8.2.2. *Cross-Linking Agents*

Melamine cross-linking agents [621] are preferably used with polymer dispersions. They are produced by hydroxymethylation of melamine in the presence of alkaline or acidic catalysts followed by etherification, e.g., with methanol. The resulting precondensates are cross-linking agents that cure by polycondensation with liberation of water in aqueous solution or in powder form on the textile in the presence of acidic catalysts at elevated temperatures. Pre-condensates in the form of aqueous solutions are used as the main components of finishing agents for sail fabrics together with a hydrophobic or oleophobic agent, or are used with poly(vinyl alcohols) for stiffening of textiles.

Melamine cross-linking agents in powder form, also water-soluble, have glass transition temperatures between 55 and 70 °C. With some products, the DSC curves show melting peaks in addition to the glass transition. After curing on the textile, neither T_gs nor melting peaks are present.

On prolonged storage in aqueous solution melamine cross-linking agents undergo continuous chemical condensation with slow reduction of pH to below the neutral point, especially if they have a low degree of etherification and are therefore very reactive. This can lead to the formation of precipitates at higher ambient temperatures.

These products give solvent-resistant bonds after curing or molding processes.

When melamine cross-linking agents are added to polymer dispersions, they give improved resistance to water, washing, perchloroethylene, and solvents, increased strength and thermal stability, and reduced shrinkage. Alone it has a flame retardant effect due to the nitrogen content.

Example: *N*-methoxymethylmelamine cross-linking agent, partially etherified, in aqueous solution, with a binder, T_g +40 °C, based on acrylic and methacrylic esters on polyester spunbond (180 g/m^2). Application by padding, dried at 120 °C, cured for 3 min at 150 °C. Dry add-on: ca. 25 % (500 g/L binder 50 %, 20 g/L melamine cross-linking agent 75 %, 5 g/L catalyst). Properties were as follows (values with cross-linking agent in parentheses):

Breaking strength at room temperature, N	585 (594)
Elongation at room temperature, %	39.5 (36.8)

Breaking strength at 180 °C, N	266 (234)
Elongation at 180 °C and 100 N load, %	14.8 (12.3)
Absorption height, cm	0.9 (3.7)

This example illustrates the reduction in hot strength and hot tensile elongation as a consequence of hardening by the cross-linking agent, and the reduction in the hydrophobic behavior of the binder.

The cross-linking agents used in resin finishes for classical textiles (see Section Cross-Linking Agents) are also used for technical textiles. Internal and external cross-linking are described in Cross-linking.

Catalysts accelerate external and internal cross-links. They are mostly acidic catalysts, i.e., salts such as amine hydrochlorides, magnesium chloride, or ammonium dihydrogenphosphate, and have good solubility in water and aqueous products.

Heavy metal salts such as zinc nitrate or zinc chloride are very effective, but are not used for environmental reasons. Acids such as phosphoric, oxalic, maleic, and citric acid are also highly effective. The low pH values of many polymer dispersions, resulting from their production, are suitable for cross-linking.

Catalysts impair the effect of thickeners.

If polymer dispersions which have been made alkaline (e.g., SBR and NBR dispersions) are used with cross-linking agents with an acid reaction and the appropriate catalysts, the extent of cross-linking, which takes place in the alkaline range, may be decreased; this can be compensated for by increasing the amount of catalyst.

8.2.3. *Nonslip Agents*

One of the main aims in the manufacture of technical textiles is to achieve high mechanical stability, which may not be provided by the fiber material alone. An important aspect of mechanical stability is the prevention of the movement of fibers, fiber systems, or yarns within the textile structure. Nonslip (antisnag) agents are also said to increase the seam strength of the material [622].

The stabilization of the textile is mainly due to the prevention of movement at points where threads meet or interlace.

Smooth fibers and open textile structures are the main causes of slipping. Knitted products made of smooth fibers or filaments, woven textiles, lattice woven materials, layered yarn structures, nets, and ropes are finished to give slip resistance.

Nonslip agents for technical textiles include:

1. Polymers: polymer dispersions, natural resins, PVC plastisols, polymer powders, poly(vinyl alcohol), starch, and starch derivatives
2. Silica and silica derivatives
3. Metal oxides

Two fundamental principles are utilized: Formation of a film on the fibers that reduces movement of individual fibers, warp threads, and yarn at points of interlacing, or production of a rough surface by silica or metal oxides.

Polymer Dispersions and Plastisols. Polymer dispersions with interior cross-linking are less effective against slipping than their chemical analogues without reactive groups. Poly(vinyl alcohol) (T_g ca. 70 °C), poly(vinyl acetate) dispersions (T_g 30 – 35 °C), or copolymer dispersions with T_g between 15 and +30 °C (hard-film formers) are suitable.

To give a softer hand, polymer dispersions with a T_g range of – 5 to – 20 °C are used. As a general rule of thumb (although there are exceptions), lower T_g values corespond to softer films and lower slip resistance. In spite of the stronger adhesion of pressure-sensitive adhesives, which form very soft films, these provide very little slip resistance. Moreover, soft films interfere with subsequent calendering processes by causing adhesion, and if many textile layers are stored under their own heavy weight, they can block. This can be avoided by using a harder film former, which improves slip resistance, but often makes the hand undesirably harsher. The addition of cross-linking agents such as *N*-methoxymethylmelamine compounds to polymer dispersions does not improve slip resistance.

PVC plastisols are used for strengthening lattice woven materials and layered yarn structures because of their softness, elasticity even at relatively low temperatures, and superior cold fracture resistance compared with the above-mentioned polymer dispersion.

Slip can also be reduced by polyurethane dispersions combined with vinyl acetate – ethylene copolymer dispersions.

Silica and its Derivatives. Products based on silica are the most typical textile nonslip agents. Silica products dry to form a brittle, powdery material. They are often combined with other textile auxiliaries, especially polymer dispersions, to improve washing stability. When using silica products, their compatibility with other auxiliaries must be tested, as, although the cationic silica nonslip agents are often very effective, many polymer dispersions have anionic character, so that precipitates are likely to be formed.

Some commercial silica-based nonslip agents follow:

Aerosil	Degussa, Germany
Burasil	Baur Gaebel, Germany
Durasol	Henkel, Germany
Durifan	Thor, Germany
Feran	Rudolf, Germany
Fixan	Rotta, Germany
Fornax	Pfersee, Germany
Hispasil	Hispaño Quimica, Spain
Lurapret	BASF, Germany
Schiebefest	Dr. Th. Böhme, Germany
Sevofast	Textilcolor, Switzerland
Silene	Auschem, Italy
Syntharesin	Bayer, Germany
Sytex	Bozetto, Italy
Tubicoat	Chem. Fabrik Tübingen, Germany

Structure Combinations. Very good slip resistance can be produced by bonding or sealing a very thin film, e.g., of polyethylene, to the textile. This allows the textile to stay soft while maximizing slip resistance. Nonwovens are made slip resistant by thermobonding in a thermo- calander.

In bonded fiber products, the slip resistance is a result of the direct bond between the fiber and the polymer produced by pressing.

Slip resistance is often required only temporarily, e.g., if a somewhat unstable material has to withstand mechanical stresses in the course of further processing. For example, a woven glass fiber for use as a wall covering can be impregnated, dried, and cured, so that it will remain stable while in contact with wallpaper paste during application to the wall.

8.2.4. *Water Repellents [623]*

Apart from the water repellents used in classical textile finishing (see Section Water Repellents), ammonium stearate is used as a hydrophobic component of polymer dispersion coating pastes. The pH must be controlled to prevent precipitation of sparingly soluble stearic acid in acid media.

8.2.5. *Oleophobic Agents*

Oil repellency is usually associated with hydrophobicity and soil repellency. Oil, water, and soil repellency are very desirable effects in technical textiles [624]. The ability to repel oils, mineral oils, naphtha, diesel fuel, and hydrocarbons, is achieved by the presence of perfluoro groups in polymers.

In principle, an “open” textile can be made oil-proof by sealing it with an oil-repellent fluorine-free polymer that does not swell in oil, e.g., a hydrophilic polymer such as poly(vinyl alcohol). Especially high oil repellency is given by polymer coating with perfluoroalkyl groups ($\geq C_6$). They are usually applied as dispersions and are generally used in combination with polymer dispersions that do not contain perfluoroalkyl groups.

Example: A thin, oil-repellent polyester nonwoven is obtained by impregnating (padding process) with a liquor of the following composition:

Vinyl acetate – ethylene copolymer dispersion, 50 % (T_g – 18 °C)	200 g/L
Oleophobic agent based on a perfluoromethacrylate, fluorine content 16.5 % (T_g 47 °C, *mp* 75 °C)	10 g/L
Melamine cross-linking agent, 50 %	25 g/L
Amine hydrochloride catalyst	5 g/L
Dried at 120 °C and cured for 3 min at 150 °C	

Unlike the silicone hydrophobic agents, there are oleophobic – hydrophobic finishes that do not affect the bonding of a hot-sealed solid polymer subsequently applied.

8.2.6. *Flame Retardants (* ⟶ Flame Retardants*)*

Low-flammability fibers such as aramid, polyester (Trevira CS), acrylonitrile copolymers, PVC, glass, and other inorganic fibers, lose some flame-resistance when treated with finishing agents such as polymer dispersions, hydrophobic agents, and softeners. Some auxiliaries do not affect the flame resistance if applied sparingly. For example, combinations of melamine cross-linking agents and polymer dispersions containing flame retardants are used for finishing glass textile wall coverings.

In addition to giving the flame-proofing effect required regulations [625], flame retardants should also exhibit stability towards water, washing, wiping, chemical cleaning, heat, and light. The following product groups are used for technical textiles:

1. Halogen- containing compounds: PVC pastes, chlorinated rubber, poly(vinyl chloride) and poly(vinylidine chloride) dispersions, vinyl acetate – ethylene – vinyl chloride copolymers, chlorinated alkanes in combination with antimony trioxide, tris(dichloropropyl) phosphate, organic bromine compounds, bromides
2. Inorganic and organic phosphorus compounds: ammonium dihydrogenphosphate
   Examples: flame-proofed viscose fiber nonwoven:

Impregnation with polymer dispersion based on vinyl acetate – ethylene, 50 % (T_g – 15 °C)	70 parts
Diammonium hydrogenphosphate	30 parts
Water	100 parts

   Guanidine phosphate and ammonium polyphosphate are also combined with polymer dispersions.
3. Nitrogen compounds: polyurethane dispersions for coatings, melamine cross-linking agents
4. Aluminum hydroxide as a filler
5. Microencapsulation of flame-proofing agents [626]

8.2.7. *Antistatic Agents*

For technical textiles, antistatic agents are mainly used in combination with polymer dispersions. Table 53 gives a brief overview of the antistatic behavior of typical materials used in a coating paste for a carpet backing. The use of special antistatic fibers is described in the literature [627].

Table 53. Effectiveness of various antielectrostatic additives to a polymer dispersion applied on a tufted carpet backing Formulation: copolymer dispersion (vinyl acetate – ethylene based, T_g 5 °C), 100 parts, polyacrylate thickener 2 parts, viscosity of coating paste ca. 5000 mPa · s, drying temperature of the carpet 120 °C; dried to form a film at room temperature.

Coating	Content of antistatic agent (100 %) in coating, %	Surface electrical resistance, Ω	
		Film	Carpet
Without antistatic agent		$\times 10^9$	2×10^{11}
With stainless steel fibers (Bekinox LT) [a]	1	$<10^7$	$<10^7$
With carbon dispersion, anionic [b]	2	6×10^8	12×10^9
With potassium formate	2	$<10^7$	3×10^9
With anionic surfactant: secondary alkyl sulfonate [c]	2	$<10^7$	3×10^8
With nonionic surfactant: fatty alcohol 13 EO [c]	2	2×10^8	4×10^{10}
With antistatic agent Tallopol [d]	2	$<10^7$	$\times 10^9$

[a] Bekaert Deutschland, Bad Homburg, FRG.
[b] Degussa, Kalscheuren, FRG.
[c] Hoechst, Frankfurt/Main, FRG.
[d] Stockhausen, Krefeld, FRG.

The electrostatic charging correlates with the hydrophobic behavior of a material. Typical auxiliaries for producing antielectrostatic finishes are hydrophilic: surfactants, salts, and carbon-black suspensions (which contain a high level of dispersing agents). This effect counteracts the often desired hydrophobicity of the textile. In some cases, metallic conductivity due to metal fibers or metal powder can be used, provided that these are not completely coated with insulating materials. The example of polypropylene spunbond in Table 54 demonstrates the difficulty of obtaining good hydrophobicity of a relatively open structure when small amounts of auxiliary are used, while maintaining the character of the textile.

Table 54. Antistatic and hydrophobic properties of polypropylene (PP) and polyamide (PA) spunbonds impregnated by padding (1 bar); polypropylene dried at 120 °C; polyamide dried at 120 °C, cured for 2 min at 150 °C

Formulation	Surface resistivity at 25 % R.H., Ω		Water column, first drop, mm		Oleophobicity (AATCC Test 118, 1966)	
	PP	PA	PP	PA	PP	PA
Untreated fabric (basis weight 35 g/m^2)	2×10^{13}	2×10^{13}	30	10	0	0
20 g/L antistatic agent (solids content 42 %)	10^8	6×10^7	10	0	0	0
20 g/L antistatic agent 20 g/L oleophobic agent (solids content 27 %), F content 16 %	2×10^8	10^9	10	30	6	5
20 g/L antistatic agent 20 g/L oleophobic agent 50 g/L polymer dispersion, styrene – acrylate, 50 %, T_g +28 °C	10^8	3×10^8	15	50	6	2
50 g/L polymer dispersion, T_g +28 °C	10^{13}	3×10^{13}	30	20	0	0
20 g/L oleophobic agent	2×10^{12}	10^{13}	55	60	0	0

8.2.8. *Plasticizers (see also Section Softeners)*

Unlike polymer plasticizers, textile softeners, typically act on the fibers. They modify the hand of a yarn or a woven or knitted material, e.g., from a rough finish to a smooth, flowing, full-bodied, or “crunch” hand.

Problems can be caused by migration of the plasticizer. The migration of plasticizers insoluble in water can loosen applied coatings and soften the material. Loss of strength can occur. In the case of a carpet backing coating that contains a plasticizer that migrates readily, this could reach a PVC underlay.

Barrier layers of softener-resistant polymers can help to prevent softeners migrating to the surface.

As plasticizers can have varying degrees of volatility, and can be emitted from the polymer, they can become deposited on car windscreens (fogging) [628].

PVC powders used for floor coverings are formulated as pastes [575]. A typical formulation (parts by weight) is as follows:

PVC powder	100
Plasticizer (e.g., dioctyl phthalate)	67
Epoxy plasticizer	3
Stabilizer	2
Chalk	10
Titanium dioxide	5

8.2.9. *Coupling Agents*

A brief overview of coupling agents is given below:

1) Applied to glass fibers to increase the affinity to organic polymers	aminopolysiloxanes
2) Applied to tire cord (viscose) to improve bonding to the rubber	resorcinol – formaldehyde condensation products together with latex or resorcinol with formaldehyde-releasing substances (hexamethylenetetramine or *N*-hydroxymethylmelamines)
3) Applied to polyester and aramid	pretreatment with isocyanate or poly-epoxide
4) Applied to nonwovens to improve the bond to resins for glass fiber nonwovens	copolymer dispersions based on acrylic esters, cationically emulsified polymers or polymer dispersions with cationic modification of the polymer

8.2.10. *Thickeners*

The application of auxiliaries and their subsequent effect on the technical textile are strongly influenced by viscosity. High

viscosities lead to increased rates of application to the surface and give sharp outlines while low viscosities give good penetration. This is true for the application of the auxiliaries both “flat” or as a foam. The auxiliaries themselves can be of low, medium, or high viscosity, pourable easily or with difficulty, or in paste form. They are often diluted with water before application. Variation of the viscosity is therefore an important method of ensuring that a material can be satisfactorily applied to technical textiles. Practical viscosity ranges (mPa · s) are as follows:

Low viscosity	20 – 600
Medium viscosity	>600
High viscosity	>6000
Application process	
Padding (foularding)	<1000
Nip roller	ca. 2000
Coating	≥ca. 6000
Printing	>10 000

The viscosity of polymer dispersions depends on:

1) The polymer/copolymer	increase in molecular mass of the polymer correlates with increased viscosity
2) Content of reactive monomers in the copolymer	*N*-hydroxymethylacrylamide has a greater tendency to increase viscosity than *N*-methoxymethylacrylamide Carboxyl groups greatly increase viscosity due to salt formation
3) Protective colloids (polyvinyl alcohol, cellulose ethers)	increase the viscosity
4) Plasticizers	increase the viscosity
5) Fillers	increase the viscosity at high concentrations
6) Incompatibility between the polymer dispersion and the additive before the onset of coagulation	different ionicity (anionic/cationic); low shear stability; low tolerance to salt; incompatibility with thickeners

In the production of coatings, thickeners based on polyacrylic acid are mainly used as low-viscosity dispersions thickened by alkali addition (usually ammonia), or are used as high-viscosity solutions of the ammonium or alkali metal salt. The so-called inverse thickeners based on polyacrylic acid have higher activity per unit solids content [629]. Acidic polymer dispersions are usually adjusted to pH > 7 by the addition of ammonia before adding the thickener. With alkaline dispersions (e.g., SBR, and NBR), pH adjustment is unnecessary.

The thickening effect of cellulose ethers is independent of pH. Carboxymethyl, hydroxyethyl, and hydroxymethyl cellulose are used, and are effective in acidic and alkaline media. Their viscosity can be controlled both by the degree of substitution and by their concentration in the liquid medium to be thickened.

The thickening effects of poly(vinyl alcohols) and starches are also exploited (see ⟶ Textile Printing – Nonionic Unmodified Polysaccharides , ⟶ Textile Printing – Nonionic Modified Polysaccharides). Thickeners based on alginate and guar gum used in textile printing are less often used in the technical textile area.

Thickeners are water-soluble polymers that harden films on drying. The special components of inverse thickeners that produce the inverse phase decrease the glass transition temperature of coating materials (Table 55).

Table 55. Polymer dispersions and thickeners: T_g of film by DSC

Styrene – acrylate copolymer dispersion 50 %	T_g, °C
Without additive	26.1
With 2 % polyacrylate thickener 30 %/ammonia	26.5
With 2 % inverse thickener 30 %	22.5
Polyacrylate thickener/ammonia (as ammonium salt, without dispersion)	none

Thickeners are effective stabilizers for auxiliaries and polymer dispersions that are processed as foams.

Salt additions and catalysts often impair the thickening effect.

8.2.11. *Surfactants (* ⟶ Surfactants***)***

Surfactants have a wide range of uses:

1. Emulsification, dispersing, dissolution
2. Increasing the mechanical stability (shear stability) of liquid products, and preventing the deposition on rollers
3. Improvement of the spraying properties of auxiliaries
4. Improvement of the stability of auxiliaries towards salts
5. Increasing the amount of filler that can be incorporated in a liquid product
6. Improvement of dyeing yield
7. Production of foamable liquid products
8. Wetting of particles or textiles and improvement of penetration
9. Increasing hydrophilicity and water absorption capacity

They may also affect the volatility of water vapor, and the reduction of the hydrophobicity and oleophobicity of coatings and the strength of adhesive bonds between the textiles and coatings. For example, nonionic surfactants in coating pastes for carpet secondary backings in concentrations of 1 % can considerably weaken the bond. If surfactants are added to polymer dispersions in a separate stage, the surfactants can form gels on mixing with water, and these are difficult or impossible to disperse homogeneously in liquid products. Certain surfactants are suitable for sensitizing polymer dispersions. The cloud point can be used as an indicator of their suitability for this use.

The various kinds of foaming behavior which the surfactants in auxiliaries can initiate can be utilized to produce low-foam polymer dispersions. These can then be treated with a surfactant with stronger foaming properties to give a stable foam for an application process.

The effect of surfactants on the surface tension (mN/m) of a polymer dispersion is illustrated by the example of a vinyl acetate – ethylene copolymer dispersion (T_g 5 °C) applied to a viscose fiber nonwoven: no additive, 39.5 mN/m; with 2 % anionic emulsifier 37 %, 35.3 mN/m; with 1 % sulfosuccinate 65 %, 28.2 mN/m. Wetting by a water drop occurred after 30 s, 10 s, and 2 s, respectively.

8.2.12. *Antifoams*

Although antifoams are useful for transporting dispersion and in preventing extreme foam formation during application, problems can arise when they are used. The products consist of oils, mineral oils, paraffin oil, high-boiling hydrocarbons, silicone fluids, silicone derivatives, high-boiling esters, and high molecular mass alcohols, and are mainly supplied as emulsions for use in aqueous products or liquors (see ⟶ Foams and Foam Control – Textile Industry).

Antifoams are used during the manufacture of polymer dispersions to prevent them foaming while they are being charged into other vessels. Excessive foaming can be caused by certain surfactants and sometimes when residual substances are evolved in gaseous form.

When antifoams are added to impregnation baths containing polymer dispersions, they can cause reduction in their wetting properties in subsequent processes, e.g., wetting of an impregnated and dried cellulose fiber nonwoven, or can cause a reduction in the bond strength of a polymer- coated nonwoven. They can also promote the pickup of dry or wet soil.

Antifoams can form agglomerates with polymer dispersions, and these can lead to the formation of deposits on rolls and textiles. The appearance of “fish eyes” in the polymer film formed by the polymer dispersion indicates the nonuniform distribution of an antifoam, and that oily particles have separated.

Antifoams also hinder the formation of foam required for textile application. However, this can be conteracted by using larger quantities of foaming agent (surfactant).

8.2.13. *Antimicrobial Agents*

The presence of microorganisms causes problems of two kinds.

Auxiliaries can contain very large numbers of microorganisms, and if they also contain sulfur- containing compounds (e.g., surfactants with sulfonate groups) hydrogen sulfide can be formed on storage. Auxiliaries generally contain < 1000 microorganisms per milliliter, and antimicrobial agents can deal with this. Microbe counts of 10^6/m L can cause problems. Product contamination is usually caused by the use of contaminated water in the manufacture of the auxiliary, and this can occur from time to time in the summer months in a central European climate. Contaminated dilution water for the storage tank in the textile plant can also cause or exacerbate the problem. The odor produced can be transferred to the textile during the treatment process. As it is almost impossible to remove the odor at a later stage, strict control of water quality is very important.

Minimization of the residual monomer content of polymer dispersions for ecological reasons can have a detrimental effect on product stability, so that special preservatives must be added. In many countries, formaldehyde/paraformaldehyde is used, which is one of the simplest and most effective methods of preservation.

The second kind of problem is the damage caused to textiles by microbial attack, mostly by mildew. This is due to storage of an auxiliary in moist, warm conditions, sometimes with pressure to the textile, which provides a breeding ground for microbial growth. The remedy is to add antimicrobial agents to treatment liquors, coating pastes, and binders [630].

8.2.14. *Light Stabilizers*

Light stabilizers prevent yellowing and embrittlement of polymers by light and ultraviolet radiation [631]. Benzotriazoles and

benzophenones are examples of such stabilizers, which bond to fibers and protect the body against sunlight [632].

The action of light stabilizers depends on the high reactivity of these substances towards the radicals formed when the polymers are subjected to UV radiation.

8.2.15. *Fillers [633]*

When inert substances are added to aqueous product systems, they produce an increase in volume, an increase in weight, a reduction in cost, and a matt rather than a transparent appearance, but seldom an improvement in strength properties. Inorganic fillers are used in considerable quantities in combination with polymer dispersions in technical textiles.

Fillers for polymer dispersions and their useful properties are:

Carbon black (aqueous dispersion)	black pigment, e.g., for a black-out coating
	antistatic agent for carpet backing
Graphite	antistatic agent
Titanium dioxide	white pigment for coatings
Calcium carbonate (chalk)	filler for flat and foamed coatings for carpet backings
Aluminum hydroxide	flame-resistant component
Kaolin	filler with a soft filled structure
Sand (fine)	for coating glass textiles
Barium sulfate	to give a highly-weighted textile

When incorporating powdered fillers into liquid products, a preliminary preparation stage in which surfactants or polyphosphates are added is advantageous to ensure that the polymer dispersions have a rapid wetting effect.

When chalk is incorporated into slightly acidic polymer dispersions, the pH must first be raised to > 7 by adding ammonia, sodium carbonate solution, or dilute sodium hydroxide solution to prevent liberation of CO_2.

A typical formulation (parts by weight) for a first backing coating with a high percentage of filler for tufted carpet is given below:

Copolymer dispersion 50 %, carboxylated butadiene – styrene, T_g ca. 5 °C	200
Dispersant	1
Chalk	700
Water	100
Thickener up to 6000 mPa · s	

For the second backing coating, the proportion of filler used is somewhat less, e.g., ca. 200 – 300 %, based on the solid polymer.

Polymer dispersions which dry to form soft films or coatings have a higher capacity for filler than hard-film formers. There is therefore a correlation with T_g, provided that these are the usual vinyl, acrylic, or butadiene copolymers.

The special properties of organic fillers such as rubber powder, recycled fibers, and polymer powders are used in a number of important technical textiles.

8.2.16. *Pigments [634] (* ⟶ Pigments, Organic*;* ⟶ Pigments, Inorganic*)*

White, black, or colored pigments are used to color coatings, flock adhesives for the flock printing process, and nonwoven binders, and for printing textile surfaces and labels.

Technical textiles are rarely dyed. The compatibility of the pigment dyestuffs with polymer dispersions and the consequent dye yield must be considered. Here, the type and quantity of the dispersing agent used in the production of the polymer dispersion are important.

8.2.17. *Producers and Trade Names*

Table 56 lists commercially available polymer dispersion auxiliaries and their producers.

Table 56. Trade names of selected polymer dispersions, etc.

Company	Polymer dispersions	Polyurethane, aqueous	Melamine – formaldehyde precondensates	Oleophobic agents	Thickeners

Air Products, USA	Airflex, Vinac, Flexbond				
Allied Colloids, UK	Alcoset, Alcoprint				Viscalex, Alcoprint
American Canamid, USA			Cymel, Aerotex		
Asahi Glass, JP	Lumiflou			Asahiguard	
Astro, USA			Astromel		
Atochem, IT	Repolem			Foraperle	
Auschem, IT	Crilat	Rolflex	Permafresh Mel		
BASF, FRG	Acronal, Butofan, Diofan, Lutofan, Perapret, Propiofan, Styrofan, Vinofan, Helizarinbinder	Emuldur	Saduren, Kaurit	Persistol	Collacral, Latekoll, Lutexal
Bayer, FRG	Acralen, Acrafloc, Perbunan, Baystal, Acramin	Desmoderm, Impranil	Acrafix	Baygard	Acraconz Verdicker
Dr. Th. Böhme, FRG	Atebin, ATB-Appretur, Elastofix, Pretavyl	Elastofix	Elastofix	Pluvion	Verdickungsmittel
Buna, FRG	Sconatex, Scopacryl, Scovinat, Präwozell, Buna-Latex, Scopadien				
Chemische Fabrik Tübingen, FRG	Arristan, Tubicoat, Tubvinyl	Tubicoat		Tubicoat	Tubivis
Chuou Rika, JP	Rikabond				
Ciba-Geigy, CH	Dicrylan	Dicrylan	Lyofix		Lyoprint, FRGicrylan
Chemische Fabrik Pfersee, FRG	Vibatex		Knittex	Oleophobol	
Daikin, JP	Polyflon			Dicguard	
Dainippon Ink, JP	Voncoat, FRGicnal	Hydran		Noxguard	Dic Thickener
Dairen, Taiwan	DA				
Daisel, JP	Sdyne				
Doverstrand, UK	Revinex				
Dow, USA	Dow-Latex				
DuPont, USA	Hypalon, Teflon			Zepel, Zonyl, Stain Master	
Enichem, IT	Ravemul				
Dai Yang Chem. Ind, KO	Polysol				
FAR, IT	Policril		Metoxil		Policril
Fuller, USA	Isatex	Isatex			
Glo-Tex, USA	Glocryl, Reacticryl				
B. F. Goodrich, USA	Geon, Hycar	Estane			Carbopol
Harco	Emultex,				

Chemical, UK	Revacryl, Viking				
Hansa Textilchemie FRG				Hansa Phob	
Henkel, FRG	Dynapress, Stabiform		Stabitex	Aversin, Repellan	
Hickson & Welch, UK				Bradfluor	
Hispaño Quimica, ES	Acronet				
Hoechst, FRG	Appretan, Imperonbinder, Hostaflon	Appretan	Cassurit	Nuva	Verdicker, Tylose
Hüls, FRG	Bunatex, Lipolan, Vestolit				
ICI, UK	Fluon			Milease	
Kanebo NSC, JP	Yodosol, Necrylic				
Kapp-Chemie, FRG	Kappatex				
Korea Polymer, KO	Koplex				
Kurare, JP	Panflex				
Lefatex, FRG	Lefasol	Lefasol		Lefasol	
3M, USA				Scotchgard	
Magma, Turkey	Magmalaminator				
Mitsui Toatsu, JP	Borron, Polylack				
Montedison, IT	Vinavil, Algoflon				
Morton, USA	Polybinder				
Mydrin, UK	Mydrin				
Nasydco, Egypt	Apomul, Poliva				
National Starch, USA	Dur-o-set, Dur-o-cryl, Dur-o-cote				
Nebopolymers, NL	Neboplast				
Nippon Carbide, JP	Nikasol				
Organic Kimia, Turkey	Orgaclear				
Para Chem., USA	Paranol				
Protex, FR	Mulron, Acrymul		Prox		Acryron
Reichhold, USA	Synthamul, Elvace, Tylac				
Revertex, UK	Revacryl, Revertex-Latex				
Rohm & Haas, USA	Primal/Rhoplex, Primax				Acrysol
Röhm, FRG	Plextol				Rohagit
Rotta, FRG	Rotta-Coating, Rotta-Finish	Rotta-Coating	Rottafix	Dipolit	
Saiden, JP	Savynol				
Sandoz, CH	Melioresin, Ratifix	Meliopromul	Finish MH	Sandofluor	
Sapkin, Turkey	Sapakril				
Schill & Seilacher, FRG	Texaran, Ukadan		Ukadan	Evoral	Verdicker
Scott Bader, FR	Texicote, Texicryl			Texipol	
Sequa, USA	Suncryl,		Permafresh	Sequapel	Sequabond

	Warcoset				
Showa High Polym., JP	Polysol				
Stahl, NL	Permuthane	Permuthane, Permutex			
Stockhausen, FRG	Estekoll, Sarpifan	Fixamin		Estarfin	Mirox
Sumitomo, JP	Sumikaflex		Sumiretz		
Synthomer, FRG	Synthomer-Latex				Verdicker
Takeda, JP	Crosron				
Teikoku, JP	Teisan Resin				
Thor, FRG	Rhenappret		Quecodur		
Toyo Ink, JP	Tocryl				
Union Carbide, USA	Ucar				Ucar
Union Chemicals, USA	RES				
Valchem, USA	Valbond				
Vinamul, UK	Vinacryl, Vinamul				
Wacker-Chemie, FRG	Vinnapas, Vinnol				
Weserland, FRG	Supron			Stralin	
Zschimmer & Schwarz, FRG	Cefacoll			Antidryn	Carbolan

[Top of Page]

References

General References

1. Gesamtverband der deutschen Textilindustrie, TEGEWA, Fachausschuß Farbstoffe und organische Pigmente im Verband der Chem. Ind. (ed.): “Wissen kleidet – Textilveredlung und was man darüber wissen sollte”, brochure, 1991.
2. TEGEWA-Rechtssammlung, Loseblatt-Sammlung, 1994. DIN/ISO 9000 – 9004 (1990), EN 29 000 – 29 004, Qualitätsmanagement- und Qualitätssicherungsnormen: 9000: Leitfaden zur Auswahl und Anwendung 9001: Modell zur Darlegung der Qualitätssicherung (QS) in Design/Entwicklung, Produktion, Montage, Kundendienst 9002: Modell zur Darlegung der QS in Produktion und Montage 9003: Modell zur Darlegung der QS bei der Endprüfung 9004: Qualitätsmanagement und Elemente eines Qualitätssicherungssystems, Leitfaden.
3.

Specific References

4. TEGEWA: *Zur Nomenklatur der Textilhilfsmittel, Leder- und Pelzhilfsmittel, Papierhilfsmittel und Tenside*, Frankfurt am Main, 1987.
5. *Textilhilfsmittel Katalog 1994/95*, Konradin Verlag, Leinfelden 1994.
6. Chemikaliengesetz, Gesetz zum Schutz vor gefährlichen Stoffen, 14. 03. 1990, BGBl I S. 521, 1. Fassung v. 16. 09. 1980.
7. *Ullmann*, 4th ed. **23**, pp. 4 – 5.
8. A. Schenek *et al.*: “Einsatzmöglichkeiten von dampfaufgeschlossenen Flachsfasern in der Spinnerei,” *Melliand Textilber.* **74** (1993) 465 – 474.
9. M. Matsui: “Microfasern gestern, heute und morgen,” *Chemiefasern/Textilind.* **43/95** (1993) 223 – 227. G. G. Feth: “Membran kontra Mikrofaser,” *Frankfurter Allgemeine Zeitung*, Dec. 28, 1993.
10. R. Bauer, H. J. Koslowski: *Chemiefaser Lexikon*, Deutscher Fachverlag, Frankfurt am Main 1993.
11. K. Fischer: “Die Entwicklung der Textilveredelung in den Siebziger Jahren,” *Dtsch. Färber Kal.* **75** (1971) 97 – 134.
12. T. Lampe, P. Neumayer, W. Roggenbach: “Alterung von Textilwerkstoffen im Kraftfahrzeug,” *Melliand Textilber.* **73** (1992) 402 – 416.
13. K. Mayer: “Kettengewirkte technische Textilien,” *Kettenwirkpraxis* 2/1986, 2/1987, 4/1989, company brochure, Obertshausen.
14. P. A. Koch, *Z. gesamte Textilind.* **56** (1954) 770 –774.
15. O. Krätz: *Faszination Chemie*, Callwey Verlag, München 1990.
16. R. Klitzsch: “Von der Polsterwatte zur Isolierwatte,” *Vliesst. Nonwoven Int.* 1993, 92 – 94.
17. A. Bührler: “Zur Geschichte der Vliesstoffe,” *Textilveredlung* **28** (1993) 219 – 223.

18. P. Böttcher: “Vliesstoffe und Freudenberg,” *Int. Nonwovens Bull.* **3** (1992) no. 3, 22 – 25.
19. K. Fischer: “Moderne Trends mit Appretan-Marken für die Ausrüstung, zum Binden und Beschichten von Textilien,” *Melliand Textilber.* **71** (1990) 290 – 303.
20. H. Wilhelm: “Reaktionsfähige Hochpolymere für die Textilveredlung,” *Textilveredlung* **23** (1988) 96 –100, 178 – 181, 225 – 230, 307 – 312.
21. P. Böttcher, Vliesstoffindustrie, in den 5 neuen Bundesländern, *Chemiefasern/Textilind.* **41/39** (1991) 273 – 274.
22. S. R. Malkan, L. C. Wadworth: “Ein Rückblick auf die Spinnvlies-Technologie,” *Int. Nonwovens Bull.* **3** (1992) no. 3, 4 – 14.
23. C. Wilkens: “Geraschelte Abstandsgewirke,” *Melliand Textilber.* **74** (1993) 993 – 997.
24. Verantwortung für die Zukunft, Enquete-Kommission des Deutschen Bundestages: *Schutz des Menschen und der Umwelt*, Economica Verlag, Bonn 1993, p. 263.
25. TEGEWA, annual reports 1990, 1991, 1992, Frankfurt am Main.
26. EDANA, European Disposables and Nonwovens Assoc., Brussels, reports.
27. K. Fischer: “Aktuelles zur adhäsiven Vliesverfestigung,” *Chemiefasern/Textilind.* **41/93** (1991) 800 –802.
28. “Der Vinyl- und Acryl-Emulsions-Markt in Europa,” *TPI Text. Prax. Int.* **48** (1993) 11.
29. “Binders: Putting it all Together,” *Nonwovens World* **3/4** (1990) 41.
30. E. Schramm: “Westeuropäischer Vliesstoffmarkt 1968 – 1988,” *Chemiefasern/Textilind.* **38/90** (1988) 596 – 597; “Vliesstoffmarkt in WEU,” *Chemiefasern/Textilind.* **42/94** (1992) 934. “Die 25 größten Erzeuger von Vliesstoffen erbringen 80 % der weltweiten Produktion im Wert von 5 Milliarden Dollar,” *Vliesst. Nonwoven Int.* **5** (1990) 177. “Westeuropa: mehr Spinnvliese,” *Chemiefasern/Textilind.* **41/93** (1991) 173. Latin America: Nonwovens Factbook and Directory, 1993, Fibers and Binders for Nonwovens, Miller Freeman Books, Gilroy, USA.
31. E. Gehmayr, W. Six: “Viskosefasern für die Nonwoven-Industrie,” *Textilveredlung* **28** (1993) 225 – 227.
32. G. Egbers: “tT – ein Wachstumsmarkt,” *Jahrbuch der TI 1986*, Gesamttextil, Frankfurt am Main 1986, pp. 27 – 32.
33. H.-J. Koslowski: “Aktuelle Markttrends für tT,” *Tech. Text.* **35** (1992) T 85.
34. Verband der Dach- und Dichtungsbahnen, Frankfurt am Main, personal communication, 1993.
35. Jahresbericht des Gesamtverbandes der deutschen Textilveredlungsindustrie, Frankfurt am Main 1992.
36. W. Kothe, G. Vogl, J. Winter: “Textildruck – quo vadis?” *Melliand Textilber.* **74** (1993) 1038 –1040.
37. Pentachlorphenol-Verbotsverordnung, Dec. 12, 1989, BGBl I, 2235. Verordnung über gefährliche Stoffe (Gefahrstoff-Verordnung) Aug. 26, 1986, BGBl I, 1470 mit änderungen 1987, 1989, 1990, 5. 6. 1991, BGBl I, 1218.
38. Gesetz über die Vermeidung und Entsorgung von Abfällen (Abfallgesetz) Aug. 27, 1986, BGBl I, 1410.
39. Verpackungsverordnung, June 12, 1991, BGBl I, 1234. EC Guideline for the Classification, Packaging, and Labelling of Hazardous Substances 92/32 EC, June 5, 1992, and Hazardous Preparations 88/379, June 7, 1988.
40. TA-Siedlungsabfall, vom 1. 6. 1993. K. Finsterwalder, U. Mann: “Emissionsabschätzung für Deponien,” *Wasser, Luft und Boden* (1992) 7 – 8, 70. “Jeden Tag acht Tonnen Müll am Krankenbett,” *Frankfurter Allgemeine Zeitung*, Feb. 14, 1980.
41. A. H. Magloth; “Umweltschutz aktiv: neue ökotechnik Sparal,” *Wirkerei-Strickerei-Tech.* **39** (1989) no. 10. F. Rüegge: “New Technology for Exhaust Air Cleaning in the Textile Industry,” *2nd Int. Conf. on Coating and Laminating*, Zürich, Nov. 9, 1992.
42. P. Knobloch: “Wertstoffe aus Restmengen von Vliesstoffen,” *Tech. Text.* **35** (1992) T 104 – T 105. “Recycling von PES-Textilien,” *Chemiefasern/Textilind.* **42/94** (1992) 924. H. Meierkord: “Recycling-Möglichkeiten für PES,” *Chemiefasern/Textilind.* **43/95** (1993) 524 – 526. H. Borufka: “Nach jeder Messe ärger mit Einweg-Teppichen,” *Frankfurter Allgemeine Zeitung*, March 6, 1993. Sandler: “Index Award für recyclingfähige Vliesmatten,” *Chemiefasern/Textilind.* **43/95** (1993) 523. R. Weber: “Recycling bei Kraftfahrzeugen,” *Konstruktion* **42** (1990) 410 – 414. J.-P. Peckstadt: “Polyolefinfaser-Abfallmanagement,” *Chemiefasern/Textilind.* **43/95** (1993) 534 – 536. E. Taeger *et al.*: “Chemische Verwertung von Polyamidabfällen,” *Chemiefasern/Textilind.* **43/95** (1993) 526 – 531. G. Menges, W. Michaeli, M. Bittner: *Recycling von Kunststoffen*, Hanser Verlag, München 1991. J. M. Marzinkowski: “Reststoffe aus Textilveredlungsprozessen,” *Int. Text. Bull.* **39** (1993) no. 2, 29 – 33.
43. W.-H. Hempel: “Baumwolle – produktökologisches Sorgenkind der TI ?,” *Textilveredlung* **28** (1993) 168 –170. C. Kummer: “Baumwolle – eine Naturfaser im Clinch zwischen Markt- und ökologieanspruch,” *Textilveredlung* **28** (1993) 170 – 176.
44. J. M. Marzinkowski: “Umweltschutz in der Textilveredlung,” *Melliand Textilber.* **73** (1992) 438 – 444. A. Weber: “Textilchemie und Umweltschutz,” *Textilveredlung* **27** (1992) 388 – 391. U. Baumann, U. Engler, W. Keller, W. Schefer: “Pragmatischer Versuch, Textilchemikalien-Sortimente umweltverträglicher zu gestalten,” *Textilveredlung* **27** (1992) 43 – 45. K. Fischer: “Technik und Umwelt – ein Bündnis für große Herausforderungen,” *Melliand Textilber.* **73** (1992) 90 – 94.
45. K. L. Kochenderfer: “The Environmental Activities of the US Textile Industry,” *2nd Int. Conf. on Coating and Laminating*, Zürich, Nov. 9, 1992.
46. R. Saffert: “PVC Plastisols for Fabric Coating, Environmental Effects,” *2nd Int. Conf. on Coating and Laminating*, Zürich, Nov. 9, 1992. J. G. Kennedy: “Carpet and the Environment,” Company brochure, Dow Chemical, Oct. 29, 1991. K. Fischer: “Binder und Binderadditive – Variationsbreite unter Umweltgesichtspunkten,” *TPI Text. Prax. Int.* **48** (1993) 709 – 714; “Polymer-Dispersionen – Einsatz als Bindemittel f. Vliesstoffe unter Umweltgesichtspunkten.” *Textilveredlung* **28** (1993) 213 – 218.
47. A. Stiebert: “Kann die TI mit dem Abwassserabgabengesetz überleben? ” *Chemiefasern/Textilind.* **24/76** (1974) 575 – 582.
48. Abwasserabgabengesetz, Sept. 13, 1976, BGBl I, 2721 – 2726. E. Fitza; “Die Bedeutung des Abwasserabgabengesetzes für die TVI”, *Taschenbuch für die Textilindustrie*, Schiele + Schön, Berlin 1977. K. Fischer: “Ab 1981, Abwasserabgabe,” *Text. Prax. Int.* **36** (1981) 1245 – 1248, 1324 – 1327.

49. Verordnung über die Begrenzung von Abwasseremissionen aus Textilveredlungs- und -behandlungsbetrieben, Sept. 24, 1992, BGBl OE, no. 612, 2813.
50. H. Leitner, W. Schenk: “Konsequenzen aus dem Einsatz und der Rückgewinnung spezieller Polyacrylatschlichten,” *Melliand Textilber.* **58** (1977) 537 – 541.
51. O. Deschler: “Probleme und Möglichkeiten der Schlichte-Rückgewinnung,” *Melliand Textilber.* **60 + 61** (1979) 1003 – 1006, (1980) 42 – 50.
52. H. Schönberger: “Schlichtemittel und Abwasserbelastung aus Sicht der Wasserbehörde,” *Melliand Textilber.* **71** (1990) 429 – 431.
53. K. Fischer: “Eliminierung von Polyvinylalkohol aus Abwässern” *Melliand Textilber.* **65** (1984) 269 – 274, 340 – 345.
54. R. Zahn, H. Wellens, *Chem. Ztg.* **98** (1974) 228 – 232.
55. K. Fischer: “Eliminierungen von Inhaltsstoffen in Abwässern der Textilveredlung,” *Melliand Textilber.* **59** (1978) 487 – 494, 582 – 588, 659 – 669; “Abwasser-Inhaltsstoffe verringern, und vermeiden,” *Wirkerei-Strickerei-Tech.* **29** (1979) 492 – 499. “Abwasser: Fällung, Bio-Abbau, Vorbehandlung,” *Chemiefasern/Textilind.* **34/86** (1984) 850 – 854.
56. DIN 38412/L 25.
57. TEGEWA: “Modifizierter Zahn-Wellens-Test,” *Melliand Textilber.* **73** (1992) 755 – 758.
58. A. Killer, H. Schönberger: “Refraktäre Stoffe im Textilabwasser,” *Textilveredlung* **28** (1993) 44 – 53.
59. G. Schulz, D. Fiebig, H. Herlinger: “Minimierung der Umweltbelastungen bei Prozessen der TVI,” *Melliand Textilber.* **74** (1993) 137 – 143.
60. R. Kleber: “Die Natriumchlorit-Bleiche und ihr ökologisches Umfeld,” *Melliand Textilber.* **74** (1993) 395 – 397.
61. G. Schulz, H. Herlinger, F. U. Gähr, T. Lehr: “Oxidativer Abbau von Farbstoffen durch Ozon,” *TPI Text. Prax. Int.* **47** (1992) 1055 – 1060, 1062. H. J. Buschmann, C. Carvalho, U. Driessen, E. Schollmeyer: “Die Entfärbung von text. Abwasser durch Bildung von Farbstoff-Einschlußverbindungen,” *Textilveredlung* **28** (1993) 176 – 182.
62. L. Roth: “Selbsteinstufung von Handelsprodukten,” *Wassergefährdende Stoffe*, vol. **2**, "chap. 2.3", ecomed Verlag, Landsberg 1989.
63. TA Luft, Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz, June 27, 1986, BGBl I 1986, 95.
64. *2. Symposium Emissionen im Fahrzeuginnenraum*, VW, Wolfsburg, Oct. 5, 1992. H. Freiberg: “Saubere Abluft an Textilmaschinen: Wohin geht der Weg?” *Melliand Textilber.* **74** (1993) 875 – 883.
65. DIN 75201, Bestimmung des Foggingverhaltens, 1992. D. Eisele: “Geruch und Fogging von Automobil-Innenausstattungsmaterialien,” *Melliand Textilber.* **68** (1987) 206 – 215. P. Ehrler, H. Schreiber, S. Haller: “Emission textiler Automobil-Innenausstattung: Ursachen und Beurteilung des Kurzzeit- und Langzeit-Foggingverhaltens,” *Textilveredlung* **29** (1994) 254 – 260. P. Hardt, T. Weihrauch: “Prognosen zum Fogging- und Abluftverhalten von Textilhilfsmitteln,” *Text. Prax. Int.* **49** (1994) 163 – 167.
66. Lebensmittel- und Bedarfsgegenständegesetz, D, Aug. 15, 1974, BGBl I, 1945, 1946 and Jan. 22, 1991, BGBl I, 121. R. Franck: *Kunststoffe im Lebensmittelverkehr*, Loseblattsammlung, 41st, Carl Heymanns Verlag, Köln 1991.
67. Bedarfsgegenstände-Verordnung v. 10. 4. 1992, BGBl I, 866 – 902 Anlage 1: verbotene Stoffe. Anlage 3: Zugelassene Monomere und sonstige Ausgangsstoffe.
68. EG-Richtlinie 91/55 v. 22. 3. 1991, Festlegung von Einzelheiten eines besonderen Informationssystems für gefährliche Zubereitungen.
69. Gemeinschaft umweltfreundlicher Teppichboden e.V., Wuppertal, Satzung v. 14. 11. 1992, Erweiterung 1994 und *Chemiefaser/Textilind.* **41/93** (1991), 262.
70. Sicherheit beim Kauf von Textilien, Informationsblatt des Vereins für verbraucher- und umweltfreundliche Textilien (seit 3. 6. 1992), Eschborn, 1993.
71. Öko-Tex Standard 100, Österreichisches Textil-Forschungsinstitut, Wien, No. 5/1992.
72. H. Rath: *Lehrbuch der Textilchemie*, 3rd ed., Springer Verlag, Berlin 1972.
73. A. Chwala, V. Anger (eds.): *Handbuch der Textilhilfsmittel*, Verlag Chemie, Weinheim, Germany 1977.
74. H. Stache: *Tensid-Taschenbuch*, Carl Hanser Verlag, München 1979.
75. *Handbuch der Werkstoffprüfung*, vol. 5, Springer Verlag, Berlin 1960.
76. B. v. Falken (ed.): *Synthesefasern*, Verlag Chemie, Weinheim, Germany 1981.
77. Chemiefasern/Textilind. (ed.): *Man-Made Fiber Year Book 1993*, p. 6.
78. K. Riggert, *Chemiefasern/Textilind.* **27/79** (1977) 1084 – 1088; *Chemiefasern/Textilind.* **28/80** (1978) 117 – 122.
79. E. Welfers: “Herstellung von Chemie-Spinnfasern,” *4th Symposium für Dozenten des Textilwesen*, Oct. 14, 1976, Karlsruhe, Chemiefasern/Textilind. 26/78 (1976) 1079 – 1096; Chemiefasern/Textilind. 27/79 (1977) 42 – 60.
80. L. Riehl: “Klassische Polyesterfaser-Technologie und Avivierung,” *Chemiefasern/Textilind.* **12/78** 1039 – 1046; *Chemiefasern/Textilind.* **1/79**, (1977) 24 – 30.
81. R. Kleber: “Avivagen und Avivierungsmethoden bei Chemie-Schnittfasern und -kabeln,” *Melliand Textilber.* **58** (1977) 187 – 194.
82. Z. A. Rogowin: *Chemiefasern*, Georg Thieme Verlag, Stuttgart 1982.
83. E. M. Verenitch, *Khim. Volokna* **6** (1991) 44 – 47.
84. K. Marquardt: “Reibungs- und griffbeeinflussende Avivagekomponenten in der Flockfärberei,” *Reutlingen – Denkendorfer Technologie-Berichte*, Konradin-Verlag Robert Kohlhammer, Leinfelden-Echterdingen 1981, pp. 19 – 22.
85. R. Kleber, *Chemiefasern/Textilind.* **27/79** (1977) 322 – 328.
86. R. Kleber, *Melliand Textilber.* **61** (1980) 831 – 832.
87. H. R. Billica, *Chemiefasern/Textilind.* **27/79** (1977) 328 – 335.

88. R. Kleber, *Melliand Textilber.* **60** (1979) 263 – 267.
89. W. Sprenkmann, *Chemiefasern/Textilind.* **6** (1984) 415 – 419; **7/8** (1984) 489 – 491.
90. H. Herlinger *et al.*, *TPI Text. Prax. Int.* **9** (1985) 970 – 972, 977 – 979; **11** (1987) 529 – 532, 537.
91. A. Speidel, *Melliand Textilber.* **62** (1981) 912 –914.
92. I. S. Olsen, *Text. Res. J.* **39** (1969) 31 – 44.
93. M. I. Schick, *Text. Res. J.*, **44** (1974) 758 – 762. Links
94. M. M. Robins *et al.*, *J. Text. Ins.* **1** (1988) 126 – 139.
95. H. I. Geursen, *Lenzinger Ber.* **47** (1979) 128 – 134.
96. A. Schulberger, Ph. D. Thesis, Universität Stuttgart 1983.
97. H. Dunken, *Freiberger Forschungsh. A*, 164 (1960), 35 – 42.
98. H. König: *Analyse von Tensiden*, Springer Verlag, Berlin 1971.
99. K. Winck, *Chemiefasern/Textilind.* **42/94** (1992) 893 – 896.
100. *Chemiefasern/Textilind.* **43/95** (1993) 893 – 897.
101. J. Trauter, R. Vialon, B. Ruess, *Chemiefasern/Textilind.* **39/91** (1989) 123.
102. H. Moroff, *Melliand Textilber.* **37** (1956) 678.
103. H. Hacking: *The Sizing of Spun Yarns*, Scientific Era Publ., Stamford 1980.
104. K. Ramaszeder: *Die chemische und mechanische Technologie des Schlichtens*, Steinkopff Verlag, Dresden 1973, p. 91.
105. J. Trauter, H. Bauer, *Melliand Textilber.* **62** (1981) 519, 621.
106. J. Trauter, H. Bauer, B. Rueß, *Melliand Textilber.* **59** (1978) 524.
107. E. Becker, *Melliand Textilber.* **64** (1983) 813.
108. J. Trauter, H. Böttle, K. Götz, *TPI Text. Prax. Int.* **35** (1980) 1200.
109. J. Trauter, W. Wunderlich, H. Böttle, *Melliand Textilber.* **73** (1992) 551, 623.
110. K. Ramaszeder, *Textilbetrieb (Würzburg)* **104** (1986) no. 5, 26.
111. M. Dawson, *TPI Text. Prax. Int.* **37** (1982) 357.
112. A. von Kannen, *Chemiefasern/Textilind.* **27 79**, (1977) 1113.
113. Burlington Ind., US 3 862 475, 1973 (W. F. Illmann, R. C. Malpass, D. M. Conklin).
114. Burlington Ind., WO 80/02035, 1980 (W. F. Illmann, D. M. Conklin, C. H. Karnes, R. I. Malpass).
115. P. Ellis, S. Galuszinsky, *Text. Inst. Ind.* **18** (1980) 268.
116. E. Sonntag, V. Hüttner, *Textiltechnik (Leipzig)* **36** (1986) 305.
117. W. S. Perkins, R. P. Walker, *Text. Chem. Color.* **16** (1984) no. 4, 89.
118. C. G. Namboodri, *Text. Res. J.* **56** (1986) 87. Links
119. W. Beck, *TPI Text. Prax. Int.* **44** (1989) 623.
120. J. Trauter, R. Vialon, *TPI Text. Prax. Int.* **44** (1989) 370.
121. J. Trauter, *ITB Flächenherstellung* **104** (1986) no. 3, 19.
122. J. Trauter, D. Scholze, *TPI Text. Prax. Int.* **48** (1993) 292.
123. J. A. Radley: *Starch and its Derivatives*, Chapmann & Hall, London 1968.
124. M. W. Rutenberg in R. L. Davidson (ed.): *Handbook of Water-Soluble Gums and Resins*, McGraw-Hill, New York 1980, chap. 22.
125. J. BeMiller in R. L. Whistler, J. N. BeMiller (eds.): *Industrial Gums*, Academic Press, San Diego 1993, p. 579.
126. G. Tegge: *Stärke und Stärkederivate*, Behr's Verlag, Hamburg 1984.
127. J. Trauter, M. Laupichler, *Melliand Textilber.* **57** (1976) 375, 443, 545, 615, 713, 797, 875, 979; **58** (1977) 23, 111. J. Trauter, H. Bauer, B. Rueß, M. Laupichler, *Textilbetrieb (Würzburg)*, **96** (1978) 46.
128. J. Trauter, *TPI Text. Prax. Int.* **44** (1989) 1297.
129. P. Wurster, G. Schmidt, *Melliand Textilber.* **68** (1987) 581.
130. K. Schlüter, *Melliand Textilber.* **71** (1990) 195.
131. M. Seekamp, *TPI Text. Prax. Int.* **48** (1993) 206.
132. R. L. Feddersen, S. T. Thorp in R. L. Whistler, J. N. BeMiller (eds.): *Industrial Gums*, Academic Press, San Diego 1993, p. 537.
133. G. I. Stelzer, E. D. Klug in R. L. Davidson (ed.): *Handbook of Water-Soluble Gums and Resins*, McGraw-Hill, New York 1980, "chap. 4".
134. CMC Warp Size Bull. VC-505, Hercules, Inc., Wilmington, 1979 **39** (1984) 252.
135. U. Denter, E. Schollmeyer, *Melliand Textilber.* **73** (1992) 267.
136. H. M. Ulrich, *Dtsch. Färber Kal.* **52** (1951) 173.
137. H. Moroff, *Melliand Textilber.* **37** (1956) 1304.
138. P. Habereder, M. Denkler, R. Cosse, *Melliand Textilber.* **64** (1983) 526.
139. K. Schlüter, *TPI Text. Prax. Int.* **48** (1993) 408.
140. H. Srivastava, S. Harshe, M. Gharia, G. Mudia, *J. Text. Assoc.* **33** (1972) 139, 148.
141. T. Gerard in R. L. Davidson (ed.): *Handbook of Water-Soluble Gums and Resins*, McGraw-Hill, New York 1980, "chap. 23".

142. M. Maruhashi in C. A. Finch (ed.): *Polyvinyl Alcohols – Developments*, Wiley Interscience, London, 1992, p. 158.
143. R. K. Tubbs in C. A. Finch (ed.): *Polyvinyl Alcohols – Developments*, Wiley Interscience, London, p. 361.
144. D. L. Nehrenberg, *Melliand Textilber.* **69** (1988) 171.
145. J. Langer, *Melliand Textilber.* **71** (1990) 666.
146. O. Deschler, *Melliand Textilber.* **64** (1983) 716.
147. J. Trauter, *Textilveredlung* **25** (1990) 27.
148. J. Trauter, *Melliand Textilber.* **71** (1990) 17.
149. H. Leitner, P. Dürrbeck, *TPI Text. Prax. Int.* **48** (1993) 330.
150. D. O. Hummel, F. Scholl: *Atlas der Kunststoffanalyse*, vol. **1**, "part 1", Hanser Verlag, München, p. 30.
151. W. T. Brown, E. S. Olson, H. J. Keegan, *Am. Dyest. Rep.* **19** (1967) 703.
152. J. M. Finley, *Anal. Chem.* **33** (1961) 1925. Links Du Pont, Quantitative Analysis of Elvanol, A 85389, Wilmington 1973.
153. H. Bauer *et al.*, *Melliand Textilber.* **73** (1992) 755.
154. Q. Wheatley, F. Baines, *Text. Chem. Color.* **8** (1976) no. 2, 28.
155. P. Dürrbeck, H. Leitner, *Melliand Textilber.* **72** (1991) 819.
156. W. Schenk, A. Würz, *Melliand Texilber.* **60** (1979) 830.
157. *Melliand Textilber.* **47** (1966) 154.
158. Eastman Kodak, US 3 546 008, 1968 (C. J. Kibler *et al.*).
159. C. Brun, Y. Girardeau, B. Pointud, D. Roberjot, *Melliand Textilber.* **68** (1987) 175.
160. J. Spijkers, H. Scheer, H. Lorenz, *Melliand Textilber.* **63** (1982) 254.
161. C. Mayfield, *Melliand Textilber.* **71** (1990) 110.
162. M. S. Kellou, C. Pividori, *Melliand Textilber.* **68** (1987) 270.
163. S. Dugal, G. Heidemann, *Melliand Textilber.* **65** (1984) 216.
164. U. Denter, S. Dugal, E. Schollmeyer, *Melliand Textilber.* **66** (1985) 142.
165. V. Heap, *Melliand Textilber.* **61** (1980) 691.
166. H. Chwala, V. Anger: *Handbuch der Textilhilfsmittel*, Verlag Chemie, Weinheim, Germany 1977.
167. L. G. Bercsenyi, I. Kovacs, *Melliand Textilber.* **61** (1980) 688.
168. J. Trauter, H. J. Schneider, *Melliand Textilber.* **56** (1975) 869.
169. *Textilhilfsmittelkatalog*, Konradin Verlag, Leinfelden 1991.
170. BASF, Techn. Information, TI/T 215, Ludwigshafen, June, 1991.
171. K. Gebert, *SVCC Symposium*, Zürich 1976, Abstracts, p. 171.
172. S. Y. Kamat, A. K. Prasad, *Colourage* **38** (1991) 33.
173. S. Fornelli, *TPI Text. Prax. Int.* **45** (1990) 1178.
174. R. Delecourt, *Melliand Textilber.* **46** (1956) 721. H. Jalke, *Z. Gesamte Textilind.* **62** (1962) 760.
175. V. Windbichler, *Melliand Textilber.* **37** (1956) 297.
176. J. Voss, *Text. Prax.* **11** (1956) 1117.
177. W. Schenk, H. Leitner, *Melliand Textilber.* **59** (1978) 148.
178. P. Habereder, *Dtsch. Färber Kal.* **81** (1977) 66.
179. F. Svoboda, *Z. Gesamte Textilind.* **66** (1964) 757.
180. P. Wurster, *Textilveredlung* **13** (1978) 341.
181. U. Denter, E. Schollmeyer, *Melliand Textilber.* **72** (1991) 267.
182. *Textilhilfsmittelkatalog*, Konradin Verlag, Leinfelden 1991.
183. J. Reicher, E. Csiszar, *Melliand Textilber.* **73** (1992), 427.
184. L. Kollmann, *Melliand Textilber.* **8** (1927) 270.
185. F. S. Gore, G. M. Nabar, *J. Soc. Ind. Res.* **8** (1949) 142.
186. R. Blankenhorn, *Melliand Textilber. Int.* **51** (1970) 1063.
187. U. Kirner, A. Würz, *Melliand Textilber.* **49** (1968) 187.
188. W. Kind: *Die Bleiche von Pflanzenfasern*, 3rd ed., Springer Verlag, Berlin 1932.
189. C. S. Foot, S. Wexler, W. Ando, R. Higgins, *J. Am. Chem. Soc.* **90** (1968) 975.
190. W. G. Steinmiller, D. M. Cates, *Text. Chem. Color.* **8** (1976) 14.
191. J. Dannacher, W. Schlenker, *Textilveredlung* **25**, (1990) 205.
192. A. F. Hollemann, E. Wiberg: *Lehrbuch anorg. Chemie*, 100th ed., De Gruyter, Berlin 1985, p. 467.
193. V. Nevelling, W. Sebb, *Textilbetrieb (Würzburg)* **93** (1975) 30; **94** (1976) 41; **95** (1977) 50.
194. E. Just, *Text. Prax.* **19** (1964) 1015. BASF, DE-OS 1 469 608, 1964 (O. Schmidt).
195. Degussa, GB 1 106 732, 1965.
196. U. Kirner, *Melliand Textilber. Int.* **51** (1970) 1069.
197. P. Wurster, F. Conzelmann, *TPI Text. Prax. Int.* **45** (1990) 1269.
198. A. Agster, *Textilveredlung* **1** (1966) 276.
199. W. Sebb, *TPI Text. Prax. Int.* **35** (1980) 1216.

200. S. K. Patel, J. Varghese, *Colourage* **30** (1983) no. 5, 3.
201. P. F. Greenwood, *J. Soc. Dyers Colour.* **103** (1987) 342.
202. K. H. Gottschalk, M. Stuhlmiller, *Melliand Textilber.* **61** (1980) 72.
203. G. Prelini, *Textilia* **58** (1982) 33.
204. G. Rösch, *TPI Text. Prax. Int. (Foreign Ed.)* **43** (1988) 847.
205. D. Bechter, A. Roth, *TPI Text. Prax. Int.* **45** (1990) 500.
206. C. Duckworth, L. M. Wrennall, *J. Soc. Dyers Colour.* **93** (1977) 407.
207. H. Rath: *Lehrbuch der Textilchemie*, 2nd ed. Springer Verlag, Heidelberg 1963, p. 242.
208. J. Knott, I. F. Polet, H. Müller, *Textilveredlung* **10** (1975) no. 277.
209. W. G. Crewther, *Proc. Int. Wool. Res. Conf. Australia* vol. **E**, (1955) p. 408. A. E. Davis, E. J. Johnson, L. R. Mizell, *Text. Res. J.* **31** (1961) 825. Links
210. H. Zahm, E. Hille, *Z. Naturforsch. B: Anorg. Chem. Org. Chem.* **133**, (1958) 824. H. Zahn, *J. Soc. Dyers Colour.* **76** (1960) 226. E. Hille, *Melliand Textilber.* **40** (1959) 893.
211. M. S. Nossar, M. Chaikin, A. Datyner, *J. Text. Inst.* **64** (1973) 490.
212. H. Zahn, W. Stein, G. Blankenburg, *Text. Res. J.* **37**, (1967) , 701. Links
213. Brändy, Brumann, *Text. Manuf.* **60** (1943) 381. I. Park, *J. Soc. Dyers Colour.* **87** (1917) 111.
214. E. Elöd, R. Rudolph, *Z. Gesamte Textilind.* **41** (1938) 295.
215. W. van Bergen, H. R. Mauersberger: *American Wool Handbook*, Textile Book Publ., New York 1938, p. 426. G. v. Hornuff, H. Hiller, *Melliand Textilber.* **42** (1961) 684.
216. P. A. Duffield: *Review of Bleaching*, IWS Development Center, Ilkley, West Yorkshire, 1986.
217. K. Prett, *Melliand Textilber.* **39** (1958) 999.
218. P. Ney, *Text. Prax. Int.* **29** (1974) , 1392, 1461, 1552, 1557, 1565.
219. BASF, DE-AS 1 272 876, 1961 (O. Schmidt, V. Steinhoff). BASF, DE-AS 1 269 991, 1964 (O. Schmidt, K. Rücker).
220. M. Giesen, K. Ziegler, *Melliand Textilber.* **62** (1981) 482.
221. K. Rücker, *Dtsch. Färber Kal.* **75** (1971) 169.
222. K. Reincke: *Taschenbuch für die Textilindustrie*, Fachverlag Schiele und Schön, Berlin 1977, p. 250.
223. S. Heimann, *Melliand Textilber.* **72** (1991) 562.
224. W.-D. Kermer, *Melliand Textilber.* **69** (1988) 586.
225. Abwasserabgabengesetz vom 05. 03. 1987 und 3. Novelle vom 21. 09. 90.
226. Klärschlammverordnung v. 25. 06. 1982 und Novellierungsentwurf v. 31. 01. 1990.
227. Bundesimmissionsschutzgesetz v. 15. 03. 1974.
228. Gesetz zum Schutz vor gefährlichen Stoffen v. 22. 03. 1990.
229. Gefahrstoffverordnung v. 26. 08. 1986.
230. Gesetz über die Umweltverträglichkeit von Wasch- und Reinigungsmitteln v. 05. 03. 1987.
231. S. Held, *Textilveredlung* **24** (1989) 394.

General References

232. H. Chwala, V. Anger: *Handbuch der Textilhilfsmittel*, Verlag Chemie, Weinheim, Germany 1977.
233. H. Rath: *Lehrbuch der Textilchemie*, 3rd ed., Springer Verlag, Heidelberg 1972.
234. K. Lindner: *Tenside-Textilhilfsmittel-Waschrohstoffe*, **3 vols.**, Wissenschaftl. Verlagsgesellschaft, Stuttgart 1964 – 1971.
235. J. Stauff: *Kolloidchemie*, Springer Verlag, Berlin 1960.
236. TPI Text. Prax. Int., TEGEWA (eds.): *Textilhilfsmittel-Katalog 1991*, Konradin Verlag, Leinfelden-Echterdingen 1991.

Specific References

237. DIN 54 295, 1986.
238. S. Heimann, *Textilveredlung* **19** (1984) 42.
239. TEGEWA: *Zur Nomenklatur der Textilhilfsmittel, Leder- und Pelzhilfsmittel, Papierhilfsmittel und Tenside*, 3rd ed., Frankfurt/Main, May, 1987.
240. Gesamttextil (Deutscher Textil-Verband): *Jahrbuch 1993*, Textilservice und Verlagsgesellschaft, Eichborn 1993.
241. H. Rath, *Tenside* **2** (1965) 1. C. Matasa, *Chem. Ztg.* **95** (1971) 507.
242. H. Herlinger, D. Fiebig, B. Kastl, *TPI Text. Prax. Int.* **45** (1990) no. 12, 1291 – 1298.
243. D. Behr: *Taschenbuch der Textilchemie*, VEB Fachbuchverlag, Leipzig 1988, p. 312. H. Zahn, I. Souren, U. Altenhofen, *Melliand Textilber.* **65** (1984) 467.
244. C. Preston: *The Dyeing of Cellulosic Fibres*, Dyer's Comp. publ. Trust, Bradford 1986, pp. 163, 203.
245. U. Baumgarte, *Melliand Textilber.* **68** (1987) 189.
246. R. H. Traber, *Textilveredlung* **27** (1992) 312.
247. G. Wecker, *Chemiefasern/Textilind.* **27** (1977) 935.
248. B. Hartmark, *Melliand Textilber.* **48** (1967) 1217.
249. H. Herlinger, *Melliand Textilber.* **67** (1986) 807.
250. H. E. Tschakert, *Tenside* **3** (1966) 317.
251. K. Reincke, *Text. Prax. Int.* **28** (1973) 461.

252. BASF, DE 292531, 1913.
253. S. Heimann, K. Dachs, *Melliand Textilber. Int.* **53** (1972) no. 5, 580 – 586.
254. P. Richter, *Melliand Textilber.* **74** (1993) no. 9, 872 –875.
255. K. v. Sarkanen, C. H. Ludwig: *Lignins*, J. Wiley, New York 1971.
256. S. Heimann, *Rev. Prog. Color Relat. Top.* **11** (1981) 1 – 8.
257. G. Prazak, *Am. Dyes. Rep.* **59** (1970) 44.
258. R. D. Athey, *Tappi* **58** (1975) 55.
259. S. Heimann: *Taschenbuch für Textilindustrie*, Verlag Schiele und Schön, Berlin 1985.
260. U. Strahm, *Textilveredlung* **19** (1984) 123.
261. BASF, DE 2 444 823, 1974 (H. Wolf, J. Gerendas, E. Wilhelm).
262. G. Prazak, *Colourage* **37** (1990) no. 19, 62 – 67.
263. J. R. Aspland, *Text. Chem. Color.* **24** (1992) no. 12, 18 – 23.
264. W. Prenzel, *Chemiefasern/Textilind.* **24/76** (1974) 293.
265. S. Heimann, *Melliand Textilber.* **63** (1982) 885.
266. Hoechst: *Naphtol AS Anwendungsvorschriften 4026*, Frankfurt 1971.
267. F. Jones, *J. Soc. Dyers Colour* **100** (1984) 66 – 72.
268. H. Brünger, A. Bossmann, E. Schollmeyer, *TPI Text. Prax. Int.* **43** (1988) no. 8, 843 – 846.
269. T. v. Chambers *et al.*, *J. Soc. Dyers. Colour* **105** (1989) 214 – 218.
270. H. Leube, H. Uhrig, *Textilveredlung* **9** (1974) 97.
271. J. Skelly, *J. Soc. Dyers Colour* **89** (1973) 349.
272. J. Prikryl, J. Ruzicka, L. Burgert, *J. Soc. Dyers Colour* **95** (1979) 349 – 351.
273. DIN 53 908, 1982.
274. F. H. Richter, E. W. Winkler, R. H. Baur, *JAOCS J. Am. Oil Chem. Soc.* **66** (1989) 1666.
275. J. Han Wood, *Am. Dyest. Rep.* **65** (1976) 32.
276. X. Kowalski, *Am. Dyest. Rep.* **68** (1979) 49.
277. P. Wurster, *TPI Text. Prax. Int.* **41** (1986) 1331.
278. U. Strahm, *Textilveredlung* **19** (1984) 123.
279. Sandoz, DE 2926098, 1980 (J. P. Chavannes, R. Fischer, S. Fornelli, F. Palacin).
280. K. H. Weible, *Melliand Textilber.* **71** (1990) , 772.
281. J. F. Leuck, *Text. Chem. Color* **10** (1978) 32 – 61.
282. B. J. J. Engbers, G. Dierkes, *TPI Text. Prax. Int.* **47** (1992) 557.
283. E. Schönpflug, *TPI Text. Prax. Int.* **30** (1975) 1554; **30** (1975) 1968; **31** (1976) 53. W. Beckmann, F. Hoffmann, *Chemiefasern/Textilind.* **27/79** (1977) 557. R. Weingarten, *Melliand Textilber.* **59** (1978) 59. A. Kretschmer, *Melliand Textilber.* **59** (1978) 823. F. Hoffmann, *Textilveredlung* **24** (1989) 340; **24** (1989) 381; **25** (1990) 49.
284. U. Baumgarte, *Textilveredlung* **15** (1980) 413.
285. U. Baumgarte, *Textilveredlung* **23** (1988) 241.
286. H. Schlüter, *Textilveredlung* **25** (1990) 218.
287. U. Baumgarte, *Melliand Textilber.* **59** (1978) 311.
288. E. P. Frieser, *SVF-Fachorgan Textilveredl.* **18** (1963) 146.
289. K. Hannemann, *Textilveredlung* **27** (1992) no. 10, 321 – 325.
290. K. Knopf, U. Denter, E. Schollmeyer, *Tenside Surf. Deterg.* **29** (1992) no. 5, 306 – 310.
291. D. M. Stevenson, D. G. Duff, D. I. Kirkwood, *J. Soc. Dyers Colour* **97** (1981) no. 13, 13.
292. E. Deniz *et al.*, *DWI Reports 1993*, Aachener Textiltagung, Nov. 1992.
293. Olin Corp., EP 479 911, 1991 (T. Hemling, H. Stitzel).
294. J. Cegarra, A. Riva, *J. Soc. Dyers Colour* **104** (1988) 227 – 233.
295. D. Fiebig, E. Schollmeyer, *TPI Text. Prax. Int.* **37** (1982) 1083 – 1085.
296. V. Wassileva, K. Stoyanov, M. Duschewa, *Melliand Textilber.* **72** (1991) no. 3, 200.
297. Y. Nemoto, H. Funahashi, *Ind. Eng. Chem. Prod. Res. Dev.* **19** (1980) 136 – 142.
298. D. Fiebig, H. Herlinger, W. Brennich, *Melliand Textilber.* **74** (1993) no. 11, 1168.
299. F. Hoffmann, *Melliand Textilber.* **59** (1978) 239.
300. I. L. Rush, E. H. Hinton, *Text. Chem. Color* **11** (1979) 26.
301. H. Herlinger, D. Fiebig, S. Koch, D. Schnaitmann, *TPI Text. Prax. Int.* **38** (1983) no. 6, 583.
302. E. S. Lower, *Int. Dyer Text. Printer* **4** (1988) 29 – 31.
303. TEGEWA, *Melliand Textilber.* **64** (1983) no. 3, 230 – 232.
304. F. Bartsch, *Dtsch. Färber Kal.* **81** (1977) 245 –253.
305. A. Murray, K. Mortimer, *Rev. Prog. Color Relat. Top.* **2** (1971) 67 – 72.
306. R. Iltscheva, *Textilveredlung* **26** (1993) no. 6, 191.
307. Ciba – Geigy, US 4 032 291, 1977 (K. A. Dellian).
308. Ciba – Geigy, US 5 009 668, 1991 (H. Berendt, R. Töpfl).
309. D. Fiebig, H. Herlinger, *TPI Text. Prax. Int.* **38** (1983) no. 8, 785 – 793.

310. K. Knopf, E. Schollmeyer, *Tenside Surf. Deterg.* **24** (1987) no. 2, 101 – 106.
311. K. Knopf, E. Schollmeyer, *Tenside Surf. Deterg.* **24** (1987) no. 3, 134 – 142.
312. D. Fiebig, D. Soltau, *TPI Text. Prax. Int.* **46** (1991) no. 6, 543 – 546.
313. BASF, DE 24 44 102, 1978 (M. Daeuble).
314. H.-J. Buschmann, D. Knittel, E. Schollmeyer, *TPI Text. Prax. Int.* **45** (1990) no. 5, 376.
315. Deutsches Textilforschungszentrum Nord-West e.V., DE 40 36 328, 1991 (D. Knittel, H. J. Buschmann, E. Schollmeyer).
316. H. Imafuku, *J. Soc. Dyers Colour* **109** (1993) no. 11 350 – 352.
317. I. Komminos et al. US 3 716 329, 1970.
318. W. Beckmann, *Z. Gesamte Textilind.* **71** (1969) 603.
319. U. Mayer, *Z. Gesamte Textilind.* **71** (1969) 792.
320. Bayer, DE-AS 1 154 788, 1961 (H. von Brachel, L. Nüßler, E. Degener).
321. S. Shukla, M. Mathur, *J. Soc. Dyers Colour* **109** (1993) 330 – 333.
322. J. Cegarra, P. Puente, J. M. Fiadeiro, J. Soc. *Dyers Colour* **102** (1986) 274 – 278.
323. J. Cegarra, P. Puente, B. Castro, *Tinctoria* **85** (1988) 67 – 74.
324. F. Feichtmayr, S. Heimann, *Z. Gesamte Textilind.* **68** (1966) 509.
325. B. A. Laddy, *Am. Dyest. Rep.* **49** (1960) 272.
326. Sandoz, CH 16126/72, 1972 (S. Fabbri, J. Frauenknecht).
327. Sandoz, DE 28 03 309, 1978 (E. Hervot, Y. Rene, A. Verdoucq).
328. BASF, DE 28 12 039, 1980 (M. Daeuble, V. Weberndoerfer, H. Schulze).
329. BP Chemicals, GB 2 132 641, 1984 (A. Currie).
330. Sandoz, DE 34 17 780, 1990 (K. H. Weible, J. Palleiro Cardona).
331. BP Chemicals, EP 0 442 749, 1991 (B. Black).
332. D. Hildebrand, J. Fiegel, *Melliand Textilber.* **64** (1983) 290.
333. J. K. Skelly, *Textilveredlung* **8** (1973) 102.
334. R. Schiffer, B. Bormeister, *Dtsch. Textiltech.* **16** (1965) 264.
335. R. Rokohl, *Tenside* **2** (1965) 76.
336. G. Siegrist, *Text. Rundsch.* **17** (1962) 143.
337. E. Schönpflug, *Melliand Textilber.* **60** (1979) 244.
338. W. Haebler, D. Hildebrand, *Text. Ind. (Moenchen-Gladbach, Ger.)* **72** (1970) 501.
339. J. A. Hughes, H. H. Sumner, B. Taylor, *J. Soc. Dyers Colour* **87** (1971) 463. S. Blackborn, T. L. Dawson, *J. Soc. Dyers Colour* **87** (1971) 473.
340. A. N. Derbyshire, W. P. Mills, J. Shore, *J. Soc. Dyers Colour* **88** (1972) 389.
341. TEGEWA, *Melliand Textilber.* **67** (1986) no. 10 750 – 751.
342. DIN 53 988, 1976.
343. DIN 54 290, 1982.
344. J. Carbonell, H. Egli, R. Hasler, *Textilveredlung* **7** (1972) 472.
345. F. Lessinsky, *Dtsch. Färber Kal.* **80** (1976) 69.
346. M. Scheller, H. Steinmüller, *Dtsch. Textiltech.* **18** (1968) 639.
347. W. Beckmann, K. Langheinrich, *Melliand Textilber. Int.* **51** (1970) 316.
348. C. L. Zimmermann, A. L. Cate, *Text. Chem. Color* **4** (1972) 150.
349. E. Schönpflug, *Melliand Textilber.* **60** (1979) 244.
350. H. Egli, *Textilveredlung* **2** (1967) 211.
351. J. Hertig, *Textilveredlung* **1** (1966) 668.
352. J. H. Brooks, *J. Soc. Dyers. Colour* **90** (1974) 158.
353. P. Richter, *Dtsch. Färber Kal.* **83** (1979) 228.
354. R. Zimmermann, *TPI Text. Prax. Int.* **33** (1978) 451.
355. R. Egginger, *Textilbetrieb (Würzburg)* **97** (1979) no. 8, 49.
356. Piedmont Section of AATCC, *Am. Dyest. Rep.* **48** (1959) no. 22; **48** (1959) no. 23, 37.
357. G. Roberts, S. Solanki, *J. Soc. Dyers Colour.* **95** (1979) 226, 427.
358. G. Schreiner, *Textiltechnik (Leipzig)* **28** (1978) 371, 780; **29** (1979) 378; **30** (1980) 192, 257; **31** (1981) 436.
359. W. Beckmann, H. Hamacher-Brieden, *Text. Chem. Color* **5** (1973) 118.
360. K. Thurner, *Text. Ind. (Moenchen-Gladbach, Ger.)* **72** (1970) 872.
361. D. Fiebig, *TPI Text. Prax. Int.* **39** (1984) no. 2, 144.
362. J. S. Gow, *Textilveredlung* **18** (1983) no. 4, 119 – 125.
363. K. Keller, *Textilveredlung* **13** (1978) 140.
364. K. Jacobs, *Text. Prax. Int.* **28** (1973) 521.
365. G. Weckler, *Text. Prax. Int.* **28** (1973) 335.
366. G. Weckler, *Text. Prax. Int.* **27** (1972); **28** (1973) 335; **28** (1973) 457.
367. K. Jacobs, *Text. Prax. Int.* **28** (1973) 521.

368. A. Bendak, H. Hanna, *Teintex* **44** (1979) 11.
369. P. Medilek, M. Quurke, P. Jablonski, *Melliand Textilber.* **57** (1976) 583.
370. P. B. Simmons, D. R. Branson, R. I. Mollenaar, R. E. Bailey, *Am. Dyes. Rep.* **66** (1977) no. 8, 21.
371. P. Richner, *Textilveredlung* **13** (1978) 134.
372. Deutsches Textilforschungszentrum Nord-West: *1. Forum für Verfahrenstechnik der Textilveredlung*, Krefeld, March 4, 1977.
373. S. Heimann, W. Tobias, *Melliand Textilber. Int.* **55** (1974) no. 10, 893 – 894.
374. G. Weckler, *Textilbetrieb (Würzburg)* **96** (1978) no. 3, 74 – 80.
375. K. Dengler, *Wirkerei Strickerei Tech.* **26** (1976) no. 1, 32 – 33.
376. W. Rüttiger, *Lenzinger Ber.* **45** (1978) 160 – 171.
377. D. Fiebig, D. Bechter, *TPI Text. Prax. Int.* **31** (1976) 1057, 1199.
378. D. C. Prevorsek, R. H. Butler, G. E. R. Lamb, *Text. Res. J.* **45** (1975) . Links
379. W. H. Hempel, *TPI Text. Prax. Int.* **34** (1979) no. 12, 1626.
380. G. Weckler, *Chemiefasern/Textilind.* **27** (1977) 935.
381. H. Gerber, *Melliand Textilber. Int.* **53** (1972) 335. H. Lehmann, F. Somm, *Text. Prax. Int.* **28** (1973) no. 1, 52. S. Heimann, *Z. Gesamte Textilind.* **66** (1964) no. 2, 129. H. Haas, E. Hilgeroth, *Melliand Textilber. Int.* **51** (1970) 699. W. Beckmann, R. Kuth, *Melliand Textilber.* **48** (1967) 1441. J. N. Etters, *Textilveredlung* **8** (1973) no. 3, 187. J. N. Etters, *Am. Dyest. Rep.* **77** (1988) no. 9, 15. P. Sarabi, *Text. Chem. Color* **23** (1991) no. 3, 21.
382. S. Heimann, K. Dachs, *Melliand Textilber. Int.* **53** (1972) 580.
383. N. D. Stewart, *J. Soc. Dyers Colour* **89** (1973) no. 7, 2.
384. Casella, DE 2 710 898, 1977 (H. Beiertz, G. Weckler, W. Weikardt).
385. C. Heinrichs, W. Becker, S. Dugal, E. Schollmeyer, *TPI Text. Prax. Int.* **40** (1985) no. 11, 1218 – 1222. G. Siegrist, *Dtsch. Färber Kal.* **74** (1970) 148 – 158. R. S. Asquith, A. K. Booth, *Text. Chem. Color.* **3** (1971) no. 3, 99 – 101. R. Vogel, *Wirkerei Stirckerei Tech.* **30** (1980) no. 6, 394.
386. S. Heimann, T. Hölken, *Chemiefasern/Textilind.* **30** (1980) no. 11, 898 – 901.
387. J. Rieker, *Chemiefasern/Textilind.* **27** (1977) no. 12, 1122. J. Rieker, W. Braun, *Melliand Textilber.* **62** (1981) no. 6, 496 – 501.
388. C. C. Cook, *Rev. Prog. Color. Relat. Top.* **12** (1982) 73 – 89.
389. H. Fischer, *Textilveredlung* **25** (1990) no. 2, 54.
390. D. M. Lewis, *J. Soc. Dyers Colour* **108** (1992) nos. 7/8, 317 – 324.
391. H. D. Pratt, *Am. Dyest. Rep.* **69** (1980) no. 11, 32.
392. C. Oschatz, *Textilveredlung* **9** (1974) no. 9, 442.
393. J. Offenbach, *Dtsch. Färber Kal.* **76** (1972) 185.
394. M. S. Aboul-Fetouh et al. *Indian Text. J.* **88** (1978) no. 8, 73; **89** (1979) no. 2, 111.
395. U. Baumgarte, *Textilveredlung* **4** (1969) 821.
396. U. Baumgarte, *Melliand Textilber.* **68** (1987) 189 –276.
397. U. Baumgarte, *Melliand Textilber. Int.* **56** (1975) 228.
398. J. N. Etters, *Am. Dyest. Rep.* **78** (1989) no. 3, 18.
399. R. C. Shah, *Text. Chem. Color* **4** (1972) 268 – 269.
400. G. L. Medding, *Am. Dyest. Rep.* **69** (1980) 30.
401. E. Marte, *TPI Text. Prax. Int.* **44** (1989) 737.
402. Hoechst: *Die Schwefelfärbung unter erhöhten ökologischen Anforderungen*, Frankfurt 1993.
403. W. Titzka, *TPI Text. Prax. Int.* **33** (1978) 1387.
404. W. H. Hempel, *Int. Text. Bull.* **1978** no. 3, 282;**1979,** no. 2, 243.
405. J. Frauenknecht, D. Schwer, *Textilveredlung* **5** (1970) 912.
406. J. Schwab, *TPI Text. Prax. Int.* **34** (1979) 282. M. Hückel, *TPI Text. Prax. Int.* **33** (1978) 1095.
407. M. Homuth, *TPI Text. Prax. Int.* **35** (1980) 55.
408. P. Kutschera, G. Vogl, *TPI Text. Prax. Int.* **34** (1979) 173.
409. W. Kühnel, *Melliand Textilber.* **57** (1976) 315.
410. H. Rath: *Lehrbuch der Textilchemie* 3rd ed., Springer Verlag, Heidelberg 1972.
411. G. Faulhaber, *Melliand Textilber.* **61** (1980) 1032 –1034.
412. a)W. Schwindt, G. Faulhaber, *Rev. Prog. Color. Relat. Top.* **14** (1984) 166 – 175. b)F. Carlier, *Ind. Text. (Paris)* **1222** (1991) 68 – 75.
413. a)W. Kothe, G. Vogl, J. Winkler, *Melliand Textilber.* **74** (1993) no. 10, 1038 – 1040. b)Stork Brabant, *Chemiefasern/Textilind.* **41** (1991) 507. c)Stork Brabant, *Int. Text. Bull.* **36** (1990) no. 4, 15 ff. d)Stork Brabant, *Melliand Textilber.* **67** (1986) 261 –264.
414. a)U. Perkuhn, *Melliand Textilber.* **67** (1986) , no. 12 E 362 – E 365. b)K. Dorfner, *TPI Text. Prax. Int.* **37** (1982) no. 10, 1048 – 1050. c)U. Perkuhn, *Int. Text. Bull. Dyeing/Printing/Finishing* **3** (1981) 229 – 236.
415. D. Bechter *et al.*, *TPI Text. Prax. Int.* **44** (1989) no. 3, 290 – 291.
416. W. Berlenbach, *Int. Dyer Text, Printer Bleacher Finish.* **140** (1968) 693 – 699.
417. F. Reinert, *Melliand Textilber.* **73** (1992) no. 4, 353 – 358, E 157 – E 160.
418. W. Tiedemann, P. Hülsberg, P. Horlacher, D. Kinast, *TPI Text. Prax. Int.* **47** (1992) , 337 – 338, 343 – 345.

419. R. Schwäbel, K. Bühler, F. Belde, *Bayer Farben Rev. Sonderh.* **12** (1970) 4 – 21.
420. G. Meyer, *Text. Prax. Int.* **29** (1974) 484 – 485.
421. R. Hofstetter, *Textilveredlung* **11** (1976) 186 – 194.
422. N. Grund, *Melliand Textilber.* **67** (1986) 896 – 902.
423. E. Feess, *Melliand Textilber.* **45** (1964) 67 – 70, 172 –181, 296 – 299, 413 – 427.
424. W. Scheuermann, W. Küppers, *Melliand Textilber.* **46** (1965) 1207 – 1212, 1333 – 1338.
425. A. Blum, *Melliand Textilber.* **41** (1960) 587 – 593.
426. G. Dillmann, *TPI Text. Prax. Int.* **38** (1983) 54 – 58.
427. P. Kutschera, G. Vogl, *TPI Text. Prax. Int.* **34** (1979) 173 – 174, 179.
428. K. Roth, *Textilbetrieb (Würzburg)* **98** (1980) no. 9, 50 – 65.
429. H. Schwab, *TPI Text. Prax. Int.* **34** (1979) 282, 295 –297.
430. W. Kühnel, *Melliand Textilber.* **57** (1976) 315 – 320.
431. H. Herlinger, D. Fiebig, B. Kastl, *TPI Text. Prax. Int.* **45** (1990) 1291 – 1298.
432. E. Feess, *Text. Prax.* **23** (1968) 335 – 341.
433. H. Ilg, H. Fischer, *Text. Prax.* **25** (1970) 484 – 487.
434. H. Ilg, W. Römpp, *Melliand Textilber. Int.* **54** (1973) 82 – 88, 661 – 667.
435. C. Sommer, *Melliand Textilber. Int.* **50** (1969) 692 –698.
436. a)G. Meyer, *Melliand Textilber. Int.* **50** (1969) 698 –702. b)G. Meyer, H. Dietz, *Text. Prax. Int.* **27** (1972) 294 – 298. c)F. Hoch, R. Bäuerle, *Ciba Geigy Rundsch.* **2** (1973) 34 – 37.
437. H. D. Optiz, *Melliand Textilber.* **71** (1990) 775 – 782.
438. R. Eisenlohr, *Melliand Textilber.* **70** (1989) 945 –947.
439. a)G. Bertolina, *Melliand Textilber. Int.* **51** (1970) 812 – 815. b)H. Schwab, *Int. Dyer Text. Printer Bleacher Finish* **146** (1971) 637 – 640.
440. a)J. Winkler, *TPI Text. Prax. Int.* **34** (1979) 302, 307 – 309. b)W. Schwindt, *Melliand Textilber.* **62** (1981) 562 –564.
441. C. Heinrichs, W. Becker, S. Dugal, E. Schollmeyer, *TPI Text. Prax. Int.* **40** (1985) 1218 – 1226.
442. R. Koch, *Melliand Textilber.* **61** (1980) 167 – 173.
443. E. Feess, *Melliand Textilber.* **45** (1964) 413 – 427.
444. K.-H. Stukenbrok, *Melliand Textilber.* **72** (1991) 46 – 49.
445. H. Dahm.: “Verdickungsmittel und Kleber für den Textildruck,” 6th ed., *Bayer Farben Rev.*, Sep. 1981 (reprint).
446. D. A. Lämmermann, *Melliand Textilber.* **73** (1992) 274.
447. E. Kurz, *Am. Dyest. Rep.* **25** (1974) Jan., 25.
448. H. B. Goldstein, H. W. Smith, *Text. Chem. Color.* **12** (1980) no. 3, 49.
449. R. A. Holser, R. J. Harper, Jr., A. H. Lambert, *Text. Chem. Color.* **18** (1986) no. 11, 29.
450. A. D. Broadbent, *Text. Chem. Color.* **22** (1990) no. 8, 13.
451. A. D. Broadbent, *Text. Chem. Color.* **22** (1990) no. 9, 53.
452. M. Schwemmer, H. Bors, A. Gotz, *Textilveredlung* **10** (1975) 15.
453. Union Carbide Corp., US 4 023 526, 1977 (D. H. Ashmus, W. W. Rankin, A. T. Walter).
454. G. R. Turner, *Text. Chem. Color.* **17** (1985) no. 10, 30.
455. R. S. Gregorian, *Text. Chem. Color.* **19** (1987) no. 4, 13.
456. Tootal Broadhurst Lee, GB 291 473, 1926; GB 291 474, 1926 (R. P. Foulds *et al.*).
457. R. L. Wayland, Jr., *et al.*, *Am. Dyest. Rep.* **49** (1960) no. 24, P843.
458. *AATCC Technical Manual* **70** (1995)
459. J. Bancroft & Sons Co., US 2 504 857, 1950 (W. R. McIntire).
460. Dan River Mills, Inc., US 2 690 404, 1954 (M. J. Spangler, R. L. Wayland, Jr.).
461. J. F. Mulvaney, R. L. Evans, *Ind. Eng. Chem.* **40** (1948) 393.
462. Dow Chemical Co., US 2 930 715, 1960 (W. N. Bakke, W. F. Tousignant).
463. Dan River Mills, Inc., US 3 158 501, 1964 (R. L. Wayland, Jr.).
464. Dan River Mills, Inc., US 3 324 062, 1967 (G. S. Y. Poon).
465. E. I. DuPont de Nemours & Co. Inc., US 2 304 624, 1942 (W. J. Burke).
466. A. M. Paquin, *Angew. Chem.* **60** (1948) 267. Links
467. R. L. Wayland, Jr., *Text. Res. J.* **29** (1959) 170. Links
468. H. Kadowaki, *Bull. Chem. Soc. Jpn.* **11** (1936) 248. Links
469. Sumitomo Chemical Co., US 3 089 859, 1963 (T. Oshima).
470. Dan River Mills, Inc., US 3 524 876, 1970 (J. E. Gregson).
471. BASF, US 3 639 455, 1972 (H. Petersen, K. Renner, H. Diem).
472. BASF, US 4 156 784, 1979 (H. Petersen, T. Dockner).
473. J. D. Reid, R. M. Reinhardt, J. S. Bruno, *Am. Dyest. Rep.* **54** (1965) 485.
474. R. Steele, *Text. Res. J.* **31** (1961) 257. Links
475. G. L. Drake, J. D. Guthrie, *Text. Res. J.* **29** (1959) 155. Links
476. G. Tesoro, P. Linden, S. B. Sello, *Text. Res. J.* **31** (1961) 283. Links

477. Quaker Chemical Products Corp., US 2 785 947, 1957 (B. H. Kress).
478. J. T. Marsh, *J. Soc. Dyers Colour.* **75** (1959) 244.
479. Shell Development Corp., US 2 826 514, 1958 (C. W. Schroeder).
480. Courtalds Ltd., US 3 312 521, 1967 (J. G. Stenner).
481. BASF, US 2 731 364, 1956 (B. v. Reibnitz, W. Ruemens, K. Beideck).
482. Korotron Co., Inc., US 2 974 432, 1961 (W. K. Warnock, F. G. Hubener).
483. H. B. Goldstein, J. M. May, *Text. Res. J.* **34** (1964) 325. Links
484. International Minerals and Chemical Co., US 4 238 545, 1980 (J. H. Hunsucker).
485. Sun Chemical Corp., US 4 396 391, 1983 (B. F. North).
486. B. F. North, *Text. Chem. Color.* **23** (1991) no. 4, 23.
487. C. Tomasino, *Text*, Chem. Color. **16** (1984) no. 12, 33.
488. R. L. Wayland, Jr., L. W. Smith, J. H. Hoffman, *Text. Res. J.* **51** (1981) no. 4, 302. Links
489. B. A. Kottes Andrews, R. M. Reinhardt, *Text. Chem. Color.* **16** (1984) no. 5, 21.
490. D. Katovic, I. Soljacic, *Text. Res. J.* **58** (1988) no. 9, 552. Links
491. Y. K. Kamath, R. U. Weber, S. B. Hornby, H.-D. Weigmann, *Text. Res. J.* **55** (1985) 519, 589.
492. E. C. Roberts, A. J. Rossano, Jr., *Text. Chem. Color.* **16** (1984) no. 3, 29.
493. D. M. Pasad, C. L. Cochran, B. A. Kottes Andrews, *Text. Chem. Color.* **21** (1989) no. 6, 13.
494. B. J. Collier, Y. Chen, J. R. Collier, *Text. Chem. Color.* **24** (1992) no. 10, 26.
495. S. H. Yoon, *Text. Chem. Color.* **17** (1985) no. 4, 33.
496. B. A. Kottes Andrews, *Text. Chem. Color.* **19** (1987) no. 10, 19.
497. B. A. Kottes Andrews, R. M. Reinhardt, B. J. Trask-Morrell, *Text. Res. J.* **58** (1988) no. 5, 255. Links
498. D. Lämmermann, *Melliand Textilber.* **73** (1992) no. 7, 274.
499. J. G. Frick, R. J. Harper, *Text. Res. J.* **51** (1981) 51.
500. BASF, US 5 246 904, 1993 (P. Hois, T. Simenc).
501. B. A. Kottes Andrews, *Text. Chem. Color.* **22** (1990) no. 9, 63.
502. B. A. Kottes Andrews, *Text. Chem. Color.* **25** (1993) no. 3, 52.
503. H. M. Choi, *Text. Chem. Color.* **25** (1993) no. 5, 19.
504. C. M. Welch, *Text. Res. J.* **58** (1988) 480. Links
505. C. M. Welch, *Text. Chem. Color.* **23** (1991) no. 3, 29.
506. H. J. Buschmann, E. Schollmeyer, *Textilveredlung* **26** (1991) no. 11, 361.
507. S. N. Pandey, C. R. Raje, *Text. Res. J.* **57** (1987) no. 4, 265. Links
508. B. A. Kottes Andrews, R. M. Reinhardt, *Text. Res. J.* **52** (1982) 123.
509. C. Chen, *Text. Res. J.* **60** (1990) 669. Links
510. H. J. Lautenschlager, J. Bindl, K. G. Huhn, *TPI Text. Prax. Int.* **47** (1992) no. 5, 460.
511. J. D. Turner, *Text. Chem. Color.* **20** (1988) no. 5, 36.
512. ICI, GB 475 170, 1937 (A. W. Baldwin, E. E. Walker).
513. H. A. Schuyten, J. W. Weaver, J. G. Frick, Jr., J. D. Reid, *Text. Res. J.* **22** (1952) 424. Links
514. U.S. Dept. of the Army, US 3 347 812, 1967 (C. G. DeMarco, G. M. Dias).
515. R. K. Iler, *Ind. Eng. Chem.* **46** (1954) 766.
516. E. B. Higgins, *Text. Inst. Ind.* **4** (1966) 255.
517. Ciba, US 2 927 090, 1960 (A. Hiestand).
518. W. Noll: *Chemistry and Technology of Silicones*, Academic Press, New York 1968.
519. F. Fortess, *Ind. Eng. Chem.* **46** (1954) 2325.
520. 3M Company, US 3 529 995, 1970 (S. Smith, P. O. Sherman).
521. H. B. Goldstein, *Text. Res. J.* **31** (1961) 377. Links
522. J. Compton, W. J. Hart, *Text. Res. J.* **24** (1954) 263. Links
523. N. F. Getchell, *Text. Res. J.* **25** (1955) 150. Links
524. Deering-Milliken Research Corp., US 3 377 249, 1968 (F. W. Marco).
525. ICI Ltd., US 3 459 590, 1969 (W. M. Corbett, D. Harrison).
526. D. A. Garrett, D. N. Hartley, *J. Soc. Dyers Colour.* **82** (1966) 252.
527. *Text. World* **118** (1968) May, 112.
528. P. O. Sherman, S. Smith, B. Johannessen, *Text. Res. J.* **39** (1969) 449. Links
529. H. E. Bille, A. Eckell, G. A. Schmidt, *Text. Chem. Color.* **1** (1969) 600.
530. J. P. Stevens, US 4 008004, 1977 (B. M. Latta, I. E. Pensa).
531. D. M. Gagarine, *Text. Chem. Color.* **10** (1978) 247.
532. G. C. Tesoro, C. H. Meiser, Jr., *Text. Res. J.* **40** (1970) 430. Links
533. W. A. Reeves, G. L. Drake, Jr., R. M. Perkins: *Fire Resistant Textiles Handbook*, Technomic Publishing Co., Westport, Conn. 1974, p. 45.
534. F. V. Davis, J. Findlay, E. Rogers, *J. Text. Inst. Trans.* **40** (1949) T839.

535. S. J. O'Brien, *Text. Res. J.* **38** (1968) 256. Links
536. G. L. Drake, J. V. Beninate, J. D. Guthrie, *Am. Dyest. Rep.* **50** (1961) 129.
537. Albright and Wilson Ltd., US 2 983 623, 1961 (H. Coates).
538. J. V. Beninate, E. K. Boylston, G. L. Drake, Jr., W. A. Reeves *Am. Dyest. Rep.* **57** (1968) no. 25, 74.
539. J. V. Beninate, R. M. Perkins, G. L. Drake, Jr., W. A. Reeves, *Text. Res. J.* **39** (1969) no. 4, 374.
540. J. V. Beninate, E. K. Boylston, G. L. Drake, Jr., W. A. Reeves, *Text. Res. J.* **38** (1968) no. 3, 267. Links
541. N. Flügel, *Melliand Textilber.* **71** (1990) 219.
542. Ciba Ltd., US 3 374 292, 1968 (A. C. Zahir).
543. M. A. White, *Text. Prog.* (1983) no. 2, 13.
544. W. H. Hempel, *TPI Text. Prax. Int.* **34** (1979) 1626.
545. B. I. J. Maxon, J. Barritt: *A Survey of Scouring and Milling in the Wool Industry*, Wira, Leeds, UK, June 1950, p. 87.
546. W. Zhao, M. T. Pailthorpe, *Text. Res. J.* **57** (1987) no. 1, 39. Links
547. W. Zhao, M. T. Pailthorpe, *Text. Res. J.* **57** (1987) no. 9, 523.
548. C. Wang, M. T. Pailthorpe, *Text. Res. J.* **57** (1987) no. 12, 728. Links
549. R. H. Earle, R. H. Saunders, L. R. Kangas, *Appl. Polym. Symp.* **18** (1971) 707.
550. A. delaMaza, J. Z. Parra, L. J. Sanchez, F. Comelles, *Text. Res. J.* **58** (1988) no. 10, 571. Links
551. K. R. F. Cockett, R. Kettlewell, D. M. Lewis, P. Smith, *J. Soc. Dyers Colour.* **96** (1980) 214.
552. A. Bourn *et al.*, *Proc. Int. Wool Text. Res. Conf. 7th* **IV** (1985) 272.
553. K. Reincke, *Melliand Textilber.* **74** (1993) 408.
554. R. J. Mayfield, *Text. Prog.* (1982) no. 4, 11.
555. M. Lipson, J. B. McPhee, *Text. Res. J.* **28** (1958) 679. Links

General References

556. B. Gnauck, P. Fründt: *Einstieg in die Kunststoffchemie*, Hanser Verlag, München 1991. Schweiz. Verband der Geotextil-Fachleute (SVG): *Das Geotextil Handbuch*, 2nd ed., Vogt-Schild AG, Solothurn 1988. K. H. Möller: *Markisen-Handbuch*, Verlag Kleffmann, Bochum 1984. *Kirk-Othmer*, 4th ed., **3**, 931 – 962; **7**, 29 – 108; **6**, 595 – 605, 606 – 635, 635 – 661, 661 – 669; **1**, 445 – 446, 459, 461; **8**, 905 – 934. *Kunststoff-Handbuch*, vol. **10**, Hanser Verlag, München 1988, pp. 490 – 508, 763 – 775, 731 – 742, 911 – 921, 908 –910. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung*, Deutscher Fachverlag, Frankfurt/Main 1989. J. Lünenschloß, G. Albrecht: *Vliesstoffe*, Thieme Verlag, Stuttgart 1982. P. Sroka: *Handbuch der textilen Fixieranlagen*, Hartung-Gorre, Konstanz 1993. M. Antonietti et a., Kolloide, Company brochure, BASF, Ludwigshafen 1994.

Specific References

557. *Textilhilfsmittel-Katalog* **1994/95**, Konradin Verlag, Leinfelden 1994.
558. G. Egbers: “Technische Textilien – ein Wachstumsmarkt,” in Gesamttextil (eds.): *Jahrbuch der Textilindustrie 1986*, Frankfurt 1986, pp. 27 – 32.
559. H. Viereck: “Die dritte Säule der Textilproduktion,” in Gesamttextil (eds.): *Jahrbuch der Textilindustrie 1988*, Frankfurt 1988, p. 42.
560. Gesamtverband der dtsch. Textilveredlungsindustrie, *Jahresbericht 1992*, Frankfurt/Main p. 55.
561. TEGEWA, *Melliand Textilber.* **71** (1990) 315.
562. K. Mayer: “Kettengewirkte technische Textilien,” *Kettenwirk Prax*. **2/1987**, company brochure, Obertshausen. P. Böttcher: “Techtextil '93, Neue Anwendungsmöglichkeiten,” *Int. Text. Bull.* **39** (1993) no. 3, 20 –24. S. Anton, Berliner Stoffdruckerei, Fahnenfabrik, “Veredlung von Fahnen,” personal communication, 1993. K. Mayer, Obertshausen: *Chemiefasern/Textilind.* **36/88** (1986) T 4. P. Ehrler, K. Maute, “Überlegungen zu einer Gliederung des Arbeitsgebiets Technische Textilien,” *Chemiefasern/Textilind.* **36/88** (1986) T 102 – T 105.
563. H.-J. Koslowski: “Aktuelle Markttrends für technische Textilien,” *Techn. Text.* **36** (1993) T 148. R. Bauer, H.-J. Koslowski: *Chemiefaser Lexikon 1993*, Deutscher Fachverlag, Frankfurt 1993, pp. 238, 239.
564. E. Otto: “Recycling von thermoplastischen Elastomeren in der Automobilindustrie,” *Kunststoffe* **83** (1993) 188 – 194. Sächs. Textilforschungsinstitut, Symposium: “Reißfaser '93,” Chemnitz, Dec. 1993. A. Watzel: “Vom Textilabfall zum Nonwoven-Produkt – Nutzen durch Recycling,” *Melliand Textilber.* **73** (1992) 397 – 401, 487 – 496, 561 – 563. T. Lampe, M. Bachor: “Anforderungen an Autotextilien am Beispiel von Sitzbezugstoffen,” *Tech. Text.* **36** (1993) T 207 – T 213. R. J. Ehrig: *Plastics Recycling, Products and Processes*, Hanser Verlag, München 1992. D. Eisele: “Nadel-Polvliesbeläge für den Automobilbau,” Fa. Borgers, company brochure, Bocholt. D. Eisele: “Nadel-Polvliesbeläge für den Automobilbau,” *TPI Text. Prax. Int.* **47** (1992) 723 – 727. E. Hufnagl, A.-M. Bartl, R. Arnold: “Möglichkeiten zur direkten Verarbeitung von Kunststoffabfällen zu technischen Texilien,” *Melliand Textilber.* **74** (1993) 913, 914.
565. J. Brandt: “Faserverbundwerkstoffe werden überwiegend dort eingesetzt, wo extremer Leichtbau gefragt ist,” *Frankfurter Allgemeine Zeitung*, **50** March 1, 1993. K. Fischer: “Technische Textilien mit eigenem Weltforum in Frankfurt,” *TPI Text. Praxis Int.* **41** (1986) 759, 760. J. C. Howson, R. J. Rymill, R. F. Pinzelli: “Aramidfaser-Laminate für den Bootsbau,” *Kunststoffe* **83** (1993) 546 – 549. P. Ehrler, M. Hottner, G. Schmeer-Lioe: “Entwicklungstechnische Fragen zu Barriere-Textilien,” *Int. Text. Bull.* **39** (1993) no. 4, 6 – 10. B. Hinz: “Aufgabe und Funktion der textilen Verstärkung in hochbelasteten Faserverbundwerkstoffen,” *Chemiefasern/Textilind.* **41/93** (1991), T 34 – T 38. R. Kleinholz, C. Lackmann: “Technische Textilien aus Glasfilamentgarnen: Märkte heute,” *Chemiefasern/Textilind.* **41/93** (1991) T 211 – T 214. J. Rellmann, H. Schenck: “Barrieremedian,” *Kunststoffe* **82** (1992) 729 – 738. *Int. Encycl. Compos.* **3** (1990) 102 – 126, 172 – 174, 394 – 420, 420 – 445, 458 – 465, 490 – 525. A. R. Bunsell: *Composite Materials*, vol. **2**, Elsevier, Amsterdam 1988. B. Clauß, H. Herlinger: “Anorganische Filamentgarne für technische Textilien und Verbundwerkstoffe,” *TPI Text. Prax. Int.* **48** (1993) 905. P. Artzt *et al.*: *Die Zukunft der Textilindustrie*, expert Verlag, Ehningen 1993, pp. 133, 149.

566. TEGEWA: *Zur Nomenklatur der Textilhilfsmittel, Leder- und Pelzhilfsmittel, Papierhilfsmittel und Tenside*, Frankfurt 1987.

567. DIN/ISO 1629, Kautschuk und Latices, 1992.

568. a)R. D. Athey: *Emulsions Polymer Technology*, Marcel Dekker, New York 1991. b)S. M. Milnera, W. Schulz, S. Sisman: "Hydrolysebeständigkeit von Baumwollem," *Melliand Textilber.* **76** (1991) 641 – 654. G. Odian: *Principles of Polymerization*, J. Wiley, Chichester 1992. H. Marschner, G. Mautner: "Abhängigkeit der Verfahrensgestaltung auf das Qualitätsbild von Emulsionspolymerisaten," *Plaste Kautsch.* **33** (1986) 85 –87. H. Warson: "Developments in Emulsion Polymerization," *Polym. Paints Colour J.* **180** (1990) 473 –475, 486, 507 – 512. G. Markert: "Zur Herstellung von Polymer-Dispersionen," *Angew. makromol. Chem.* **123/124** (1984) 285 – 306.

569. DIN 55958, Harze, 1988.

570. "Der Vinyl- und Acryl-Emulsions-Markt in Europa," *TPI Text. Prax. Int.* **48** (1993) 11.

571. K. Fischer: "Binder und Binderadditive – Variationsbreite unter Umweltgesichtspunkten," *TPI Text. Prax. Int.* **48** (1993) 709 – 714. D. Patil, R. D. Gilbert, R. E. Fornes: "Environmental Effects on Latex Paint Coatings," *J. Appl. Polym. Sci.* **41** (1990) 1641 – 1650. Links

572. BGBl I, 1945, 1946 Lebensmittel- und Bedarfsgegenständegesetz, Aug, 15, 1974; BGBl. I, 526, Neufassung v. 8. 7. 1993. Gesetz zum Schutz vor gefährlichen Stoffen (Chemikaliengesetz), 1990. 7. Änderungsrichtlinie 92/32 EWG zur Richtlinie 67/548/EWG zur Angleichung der Rechts- und Verwaltungsvorschriften für die Einstufung, Verpackung und Kennzeichnung gefährlicher Stoffe. Amtsblatt der EG No. L 154/1 – 29. v. 05. 06. 1992. 6. Änderung der EG-Richtlinie 67/548 EWG entspricht dem deutschen Chemikaliengesetz von 1980. K. Fischer: "Polymer-Dispersionen, Einsatz als Bindemittel für Vliesstoffe unter Umweltgesichtspunkten," *Textilveredlung* **28** (1993) 212 – 129; "Umweltaspekte bei Polymer-Dispersionen auf Basis Vinylacetat/Ethylen und Acrylester," *Textilveredlung* **26** (1991) 354 – 361. ISO TC 45 Rubber and Rubber Products, Working Draft II, Head Space Method, 1993. Bestimmung von leichtflüchtigen KW in Latices, Prüfverfahren TFI No. 8, Teppich-Forschungsinstitut, Aachen 1992. B. Kolb, P. Pospisil, M. Auer: "Quantitative Headspace Analysis of Solid Samples," *Chromatographie* **19** (1984) 113 – 122.

573. R. Franck: "Kunststoffe im Lebensmittelverkehr," *Loseblattsammlung*, **41st.**, XIV. Kunststoffdispersionen, 15. 04. 1991, Carl Heymanns Verlag, Köln 1991.

574. EC-Guideline 92/39 EC, Plastics Materials and Objects which come into Contact with Foodstuffs, *Off. J. EC* **L 168** (1992) 21 – 29. EC-Guideline 78/142 EC, Vinyl Chloride Monomer Containing Materials and Objects which come into Contact with Foodstuffs, *Off. J. EC* **L 44** (1978) 15 –17. Bedarfsgegenständeverordnung, Bundesgesetzblatt 1992, part I, no. 20, pp. 866 – 902, Apr. 10, 1992.

575. DIN 53773/1, Prüfung von Farbmitteln in PVC-Pasten (Plastisolen), 1988. "Beschichtung von Twaron-Aramidgarn-Geweben," *Chemiefasern/Textilind.* **39/91** (1989) T 148. W. R. Bursian: "PVC-Plastisole zur Ausrüstung synthetischer Trägermaterialien," *Plastverarbeiter* **30** (1979) 57 – 64. H. Hille: "Herstellen und Eigenschaften von PVC-Pasten," *Kunststoffe* **68** (1978) 735 – 741.

576. K. Fischer: "Polymer-Dispersionen für tT," *TPI Text. Prax. Int.* **42** (1987) 518 – 522; *Can. Text. J.* **105** (1988) 22 – 29. H. Bille: "Hochpolymere in der modernen Textilausrüstung." *Melliand Textilber.* **65** (1984) 704 –709.

577. H.-J. Adler, J. Lichtenbelt, A. Reuvers: "Moderne Methoden zur Charakterisierung von Polymerdispersionen," *Farbe + Lack* **97** (1991) 103 – 108. K. Fischer, B. Schwalenstöcker: "Zur Stabilität von Polymer-Dispersionen," *Dtsch. Färber Kal.* **93** (1988) 154 – 163.

578. K. Fischer: "Tensid-Additive für Textil-Polymerdispersionen," *TPI Text. Prax. Int.* **44** (1989) 503, 506 –508, 510, 517.

579. H. Wilhelm: "Reaktionsfähige Hochpolymere für die Textilveredlung," parts 1 – 4, *Textilveredlung* **23** (1988) 96 – 100, 178 – 181, 225 – 230, 307 – 312.

580. DIN 53775/3 Farbmittel in PVC – Ausbluten,Migration 1984.

581. F. Molenaar, P. Buijsen, C. Smit: "Haftung elektronenstrahlhärtender Beschichtungen auf Metallen," *Farbe + Lack* **98** (1992) 679 – 681. K. Fischer: "Veredlung von textilen Flächengebilden durch Bildung von Polymeren auf der Faser," *Melliand Textilber.* **53** (1972) 692 – 696, 813 –819. Z. A. Rogovin: *Chemiefasern*, Thieme Verlag, Stuttgart 1982.

582. U. Eisele: "Strahlenchemische Veredlung von Baumwolle," *Melliand Textilber.* **72** (1991) 440 – 446. B. Clauß: "Elektronenstrahlinduzierte Polymerisation wäßriger Präpolymer-Emulsionen," *Melliand Textilber.* **72** (1991) 866 – 869. P. M. Fletcher, S. V. Nablo: "Environmental Benefits of Electron beam/UV Curing," *2nd Int. Conf. on Coating and Laminating*, Zürich, Nov. 9, 1992. H. Herlinger, P. Hirt: "Strahlenvernetzte PURFasern," *Chemiefasern/Textilind.* **43/95** (1993), 810.

583. K. Fischer: "Veredlungsverfahren mit polymerisierbaren Monomeren," *Textilveredlung* **8** (1973) 397 –412.

584. DIN 51007, Differenzthermoanalyse, 1992. DIN 51005, Thermische Analyse, Begriffe 1983. S. Knappe: "Thermische Analyse in der Qualitätssicherung," *Kunststoffe* **82** (1992) 993 – 998. E. Kaisersberger, H. Möhler: "DSC an Polymerwerkstoffen," *Netsch-Jahrbücher*, vol. **1**, Netsch Gerätebau, Selb 1991.

585. K. Fischer: "Die Glastemperatur," *Textilveredlung* **19** (1984) 146 – 152.

586. K. Fischer, N. Ruiz, H.-G. Herrel: "Festigkeitsvergleich an Nonwoven und Bindemittelfilmen," *Vliesstoff Nonwoven Int.* **1** (1986) 259, 260.

587. K. Fischer: "Was leisten Polymer-Dispersionen bei der Vliesbindung," *Chemiefasern/Textilind.* **38/90** (1988) 678, 680, 682, 683. K. Fischer, G. Schwarz, H.-J. Hein: Qualitative Binderanalyse durch TG-Bestimmung, *Taschenbuch für die Textilindustrie*, Fachverlag Schiele & Schön, Berlin 1991, pp. 405 – 415. ISO TC 45, Rubber and Rubber Products, Working Draft II, Head Space Method, 1992. Bestimmung von leichtflüchtigen KW in Latices, Prüfverfahren TFI No. 8, Teppich-Forschungsinstitut, Aachen 1992. B. Kolb, P. Pospisil, M. Auer: " Quantitative Headspace Analysis of Solid Samples," *Chromatographie* **19** (1984) 113 – 122.

588. P. Hardt: "Chemie und Eigenschaften von Hydrophilausrüstungen," *TPI Text. Prax. Int.* **45** (1990) 387 – 392. R. Van der Beke, L. Dekoninck: "Neuentwicklungen auf dem Gebiet der mikroporösen Beschichtungen," *Melliand Textilber.* **67** (1986) 824 – 829. J. Hemmrich: "Polyurethanbeschichtung ohne Lösemittel," *Int. Text. Bull.* **39** (1993) no. 3, 53 – 56. D. Bechter, H. Herlinger, E. Bader, B. Wittstock: "Die Polymer-Faser-Haftung bei Beschichtungen/PUR-

Dispersionen," *Melliand Textilber.* **70** (1989), 952 –956. "Wäßrige Polyurethan-Beschichtungssysteme," *Chemiefasern/Textilind.* **39 91** (1989) T 149, T 150. W. Schröer: "Die Beschichtung von Textilien mit Polyurethanen," *Textilveredlung* **22** (1987) 459 – 467. K. H. Stukenbrock: "Ionisch koagulierbare PUR-Dispersionen," *Melliand Textilber.* **65** (1984) 756 –758.

589. DIN 53787, Prüfung von wäßrigen Kunststoff-Dispersionen; Bestimmung der Mindestfilmbildetemperatur, 1974.
590. K. Fischer, R. Schardt: "Die Glastemperatur II," *Textilveredlung* **20** (1985) 49 – 52.
591. H. Herlinger, D. Bechter, G. Kurz: "Wirkungsweise von Bindemitteln bei der Vliesverfestigung," *Chemiefasern/Textilind* **31/83** (1981) 936 – 942. K. Fischer: "Moderne Trends mit Appretan-Marken für die Ausrüstung, zum Binden und Beschichten von Textilien," *Mellaind Textilber.* **71** (1990) 290 – 303; "Aktuelles zur adhesiven Vliesverfestigung," *Chemiefasern/Textilind.* **41/93** (1991) 800 –802. K. Fischer, "Bindemittel standen zur Diskussion auf der Index '93", *Vliesstoffe Nonwoven Int.* **8** (1993) 123. EDANA, Nonwovens, brochure, Brussels 1990. K. H. Schumacher *et al.*: "Neue Generation formaldehydfreier Acrylatbinder," *Chemiefasern/Textilind.* **43/95** (1993) 621 – 624. Röhm GmbH: "Bindemittel, für Vliesstoffe," *Allg. Vliesstoff-Rep.* **8** (1980) 196 – 204. W. Stepanek: "Plextol-Bindemittel für Beschichtungsträger aus PES-Vlies," *Allg. Vliesstoff-Rep.* **18** (1990) 72 – 76. "Ethylen-Copolymer-Emulsionen als Binder für Nonwovnes," *Melliand Textilber.* **63** (1982) 894 –896. W. Loy: "Vliesstoffe – eine übersicht nach Herstellungsart und Verwendungszweck," *Textilveredlung* **28** (1993) 204 – 212. "Nonwovens: The Bonded Fabrics," *Nonwoven World* 1990 no. 1, 16 – 23.
592. U. Eisele: "Färben mit Hilfe von Bindemitteln," *Melliand Textilber.* **72** (1991) 847 – 851.
593. M. Rasche: "Gibt es einen Zusammenhang zwischen Benetzbarkeit und Haftung? " *Flock* **19** (1993) no. 71, 6, 8, 10 – 14. U. Zorll: "Einfluß von Pigmentgehalt und Polarität auf das Haftvermögen von Flock-Klebstoffen zu Substrat und Flock," *Flock* **14** (1987) no. 46, 7 –20. G. P. Bauer: "Qualitätsprüfung beflockter Materialien für die Autoindustrie," *Chemiefasern/Textilind.* **41/93** (1991) T 23 – 27. K. Fischer: "Aktuelle Aspekte zu Polymer-Dispersionen, die als Flockklebstoffe verwendet werden," *Flock* **17** (1991) no. 62, 12 – 17. V. W. Dimitriew: *Faserpräparation für die elektrostatische Beflockung*, University of Leningrad 1986.
594. A. Stei: "Die Tuftingteppichbeschichtung in der Bewährung," *Chemiefasern/Textilind.* **40/92** (1990) no. 92, 723 – 726. K. Fischer, G. Schachtner: "Aspekte der Teppichrückenbeschichtung mit Polymer-Dispersionen," *TPI Text. Prax. Int.* **40** (1985) 990 – 992, 997, 998, 1092, 1097 – 1100.
595. S. Schwenkedel: *Textilbeschichtung – Kunstleder-Syntheseleder*, Eder Verlag, brochure, Stuttgart 1970. DIN 55945, Beschichtungsstoffe, Begriffe, 1988. K. Fischer, K. Winter: "Textil-Beschichtungen," *Dtsch. Färber Kal.* **89** (1985) 150 – 170. R. N. Hildred: "Acrylic Additives for Waterborn Coatings," *Polym. Paint Colour. J.* **180** (1990) 579 –583. O. Lückert: "Bindemittelprüfung in Normung und Literatur," *Farbe + Lack* **98** (1992) 277 – 280.
596. H. Zander: "Kleben statt Weben," *Kunststoffe* **83** (1993) 100, 101. H. J. Kauderer: "Verfahrenstechniken der Kontinuierlichen Textillamenierung," *TPI Text. Prax. Int.* **48** (1993) 895 – 899. E. Bürkle, G. Rehm, K. Zweig: "Großflächige Bauteile im Urformwerkzeug kaschieren," *Kunststoffe* **82** (1992) 896 – 901.
597. C. Trebs: "Korrelationen zwischen Polymer-Dispersionen und Polymer-Pulvern," Thesis, Fachhochschule Kaiserslautern 1993.
598. K.-H. Stukenbrock: "Möglichkeiten des Druckens von Vliesen," *Melliand Textilber*, 71 (1990) 303 – 305.
599. DIN/ISO 1133, Bestimmung des Schmelzindex (MFR) 1993.
600. R. Hinterwaldner: "Schmelzklebstoffe," *Coating* **24** (1991) 110 – 114. "Eine breite Palette Hüls-Produkte für textile Anwendungen," *Allg. Vliesstoff-Rep.* **19** (1991) 68 – 70. C. Lipman: "Powder Coatings Hope for Legislation," *Eur. Chem. News.* **8** (1993) 28. "Wer liefert welches Pulver?" *Oberfläche JOT*, Market Survey (1992) 41 – 44. E. Lattke, H. Drecker: "Mit Pulver ins Jahr 2000," *Oberfläche + JOT* (1992) no. 9, 34 – 40. H. Scholten: "Hot-Melt Adhesives for Textile Laminates," *Symposium Coating*, Zürich, Nov. 9 1992. H. Jahn: "Verbunde mit Vliesstoffen unter Verwendung von Schmelzklebern," *Melliand Textilber.* **74** (1993) 297 – 299.
601. G. Schwarz, Hoechst AG, personal communication 1992.
602. V. Baron: "Herstellung, Eigenschaften und Anwendungen von Schmelzklebefasern," *Techn. Text.* **36** (1993) T 200 – 206. D. Bechter, G. Kurz, E. Maag, J. Schütz: "Der Einfluß der Kalandrierbedingungen, auf die Festigkeit von PP-Vliesstoffen," *TPI Text. Prax. Int.* **46** (1991) 1236 – 1240. "Bindefasern (Produktion Westeuropa)," *Int. Text. Bull.* **39** (1993) 34.
603. DIN 7724, Polymere Werkstoffe, 1993.
604. A. Gardiella: "Duroplaste im Automobilbau," *Kunststoffe* **83** (1993) 57 – 59.
605. R. Schäfer: "Phenolharze für schalldämpfende Verkleidungselemente im Fahrzeugbau," *Kunstharz Nachr*. (1993) no. 29, 16 – 19. DIN ISO 10082, Phenolharze, Begriffe und Prüfverfahren, 1992. A. Knop, W. Scheib: *Chemistry and Application of Phenolic Resins*, Springer Verlag, Berlin 1979. A. Gardziella, H. G. Haub: "Phenolharze," in: *Kunststoff-Handbuch*, vol. 10, Hanser Verlag, München 1988, pp. 14 – 40.
606. R. Kleinholz: "Vorgeformte textile Flächengebilde für Kunststoff-Formteile," *Chemiefasern Textilind.* **40/92** (1990) T 182. S. Anton, Berliner Stoffdruckerei, Fahnenfabrik, personal communication.
607. Hoechst, *Hostaflon*, Company brochure, Frankfurt 1984. Hoechst, *Hostafalon TF-Dispersionen*, Frankfurt 1992. Hoechst, *Fluorkunststoffe und Feuer, Fragen und Antworten*, Frankfurt 1987. P.-K. Eichhorn: "Hostaflon beschichtetes Glasgewebe in Neue Perspektiven in der Architektur," *Magazin Hoechst*, Frankfurt 1989.
608. G. Tegge: *Stärke und Stärkederivate*, Behrs Verlag, Hamburg 1984.
609. H.-J. Jacobasch, K.-H. Freitag, U. Panzer, K. Grundke: "Charakterisierung und Modifizierung der Oberflächeneigenschaften von Fasern für Verbundwerkstoffe," *Chemiefasern/Textilind.* **41/93** (1991) T 39 – T 47.
610. H. Johnen: "Schmirgeln, Rauhen, Scheren," *Textilveredlung* **27** (1992) 348 – 351.
611. B. von Falkai: *Synthesefasern*, Verlag Chemie, Weinheim, Germany 1981, pp. 370 – 373.
612. "Blei-Akkumulator mit bleibeschichteten Glasfaser-Elektroden," *Frankfurter Allgemeine Zeitung*, Nov. 24, 1993.
613. "Glass Fibers, Fiber Tables According to P. A. Koch 1993," *Techn. Text.* **43** (1993) E 89 – E 100.
614. K. W. Gerstenberg: "Sortex, Oberflächenmodifikation von Textilien durch Koronabehandlung," *Allg. Vliesstoff-Rep.*

21 (1993) 28.

615. C. Gruner, B. Rapp, H. J. Zimmermann: “Problemlösungen beim Lackieren von PP-Blends,” *Kunststoffe* **82** (1992) 802 – 806.
616. T. Bahners: “Laser zur Materialbearbeitung in der Textilveredlung,” *Textilveredlung* **26** (1991) 319 – 326. D. Knittel, E. Schollmeyer: “Haftung von UV-laserbestrahlten Hochleistungsfasern zu Gummi und Epoxidharz,” *Techn. Text.* **36** (1993) T 61 – T 66. T. Ryback, D. Knittel, E. Schollmeyer: “Pigmentdruck auf UV-laservorbehandeltem PES-Gewebe,” *Melliand Textilber.* **73** (1992) 985 – 989.
617. H. Herlinger, R. Braun, Institut f. Textilchemie, Universität Stuttgart. DE 414 2776 A1, 1991 (H. Herlinger, R. Braun); DE 421 9280 A1, 1992 (H. Herlinger, R. Braun).
618. J. Mischke, G. Bagusche: “Hinterspritzen von Textilien, Teppichen und Folien,” *Kunststoffe* **81** (1991) 199 – 203.
619. DIN 55945, Beschichtungsstoffe, Begriffe, 1988.
620. J. K. Rogers *et al.*: “Chemicals and Additives,” *Mod. Plast. Int.* 1993, Special Report, 33 – 68. K. Fischer: “Additive für Dispersionen – Dispersionen als Additive” *Taschenbuch für die Textilindustrie 1987*, Verlag Schiele & Schön, Berlin 1987, pp. 329 –343; “Additive für Polymer-Dispersionen: Tenside, Verdicker, Katalysatoren,” *Chemiefasern/Textilind.* **36/88** (1986) 920, 923, 924.
621. T. Götze, W. Adam: “Melamin-Formaldehydharze,” *Kunststoffe* **80** (1990) 1185 – 1188. W. Adam: “Melaminharze,” in: Kunststoff-Handbuch, vol. 10, Hanser Verlag, München (1988) pp. 41 – 50.
622. DIN 53868, Bestimmung des Nahtschiebewiderstandes, 1992.
623. A. J. Sabia, G. A. Policello: “Silicon finishes Enhance Nonwoven Substrates,” *Nonwoven World* 1990, no. 6, 42 – 44.
624. M. Lewin, S. B. Sello: “Chemical Processing of Fibers and Fabrics, Functional Finishes, Repellent Finishes,” *Handbook of Fiber Science and Technology*, vol. **2**, Marcel Dekker, New York 1984, pp. 144 –204. D. Lämmermann: “Fluorcarbone in der textilen Endausrüstung – Eigenschaften und Anwendungen,” *Melliand Textilber.* **72** (1991) 949 – 954. D. Lämmermann: “Moderne Fleckschutzausrüstung von Textilien,” *Melliand Textilber.* **74** (1993) 883 – 889. M. Wilhelm: “Fluorchemikalien für die Textilausrüstung,” *Int. Text. Bull.* **39** (1993) 57 – 60. R. E. Wiltgen: “Fluorchemikalien,” *Textilveredlung* **21** (1986) 384 – 389. K. Fischer: “Hygiene und Schmutz aus Sicht der Textilveredlung,” *Melliand Textilber. Int.* **54** (1973) 1239 – 1242. R. Kleber: “Moderne Aspekte der textilen Oleophob-Ausrüstung,” *TPI Text. Prax. Int.* **27** (1972) 499 – 503.
625. M. Lewin, S. B. Sello: “Chemical Processing of Fibers and Fabrics, Flame Retardancy,” *Handbook of Fiber Science and Technology*, vol. **2**, Marcel Dekker, New York, 1984 pp. 2 – 122. DIN 4102, Brandverhalten von Baustoffen und Bauteilen; Brandklassen, 1988.
626. B. Hanneken, M. Schuirer, A. Schüler, H. K. Rouette: “Moderne flammhemmende Ausrüstung für ‘Black-out'-Vorhang-Beschichtung,” *Int. Text. Bull.* **39** (1993) no. 3, 38 – 46.
627. J. Laser, J. Glaser: “Elektrisch ableitende Teppichverbundsysteme,” *Melliand Textilber.* **74** (1993) 630 – 631. S. Rolfe: “Spezielle Antistatikfaser,” *Vliesstoff Nonwoven Int.* **8** (1993), 138, 139. F. Marchini: “Metallisierte Fasern zum Schutz gegen Elektrostatik und Strahlung,” *Chemiefasern/Textilind.* **40/92** (1990) T 164 – T 170.
628. DIN 75201, Bestimmung des Fogging-Verhaltens, 1989. P. Hardt: “Prognosen zum Fogging- und Abluftverhalten von Textilhilfsmitteln,” *TPI Text Prax. Int.* **49** (1994) 163 – 167. F. Look, T. Lampe, A. M. Bahadir: “Temperaturabhängigkeit des Fogging-Phänomens,” *Kunststoffe* **83** (1993) 201 – 205. P. Ehrler, H. Schreiber, S. Haller: “Emission textiler Automobil-Innenausstattung,” *Textilveredlung* **29** (1994) 254 – 260. D. Jachowski, A. C. Poppe: “Fogging-Messungen mit variiertem Verfahren,” *Kunststoffe* **82** (1992) 818 – 820. DIN EN 665, Elastische Bodenbeläge; Bestimmung der Weichmacherabgabe, 1992.
629. H. Herlinger, U. Erzinger, G. Schulz: “Rhelogische Charakterisierung und Qualitätsbeurteilung von Druckpasten und Verdickungen,” *TPI Text. Prax. Int.* **46** (1991) 247 – 252.
630. J. Warmund-Cordelier: “Zuverlässiger Schutz gegen Mikroorganismen,” *Textilveredlung* **27** (1992) 395 –397. T. Klinkenberg, W. Stöneberg: “Symprol – umweltfreundliche Markisenstoff-Imprägnierung,” *Chemiefasern/Textilind.* **36/88** (1986) 608 – 610. K. H. Wallhäußer, K. Fischer; “Die antimikrobielle Ausrüstung von Textilien,” *Textilveredlung* **5** (1970) 3 – 14. K. Fischer, K. H. Wallhäußer: “Die antimikrobielle Ausrüstung – ein Problem unserer Zeit,” *Dtsch. Färber Kal.* **74** (1970) 324 – 344. K. H. Wallhäußer, W. Fink: “Konservierung von Polymer-Dispersionen,” *Farbe + Lack* **76** (1970) 471; **82** (1976) 108. K. Fischer: “Hygiene und Schmutz aus Sicht der Textilveredlung,” *Melliand Textilber. Int.* **54** (1973) 1239 – 1242.
631. G. Reinert, V. Misun: “Neuer UV-Absorber für PES-Faserstoffe,” *Melliand Textilber.* **74** (1993) 1007 – 1014. Hoechst, *HALS-Produkte für hochwirksamen Lichtschutz*, company brochure, Frankfurt 1992.
632. “Textilien mit hohem Sonnenschutz,” *Frankfurter Allgemeine Zeitung*, Sept. 24, 1993.
633. Degussa, Pigmente company brochure no. 47: Ruß, Frankfurt 1991. Omya-Füllstoffe in Latex-Teppichrückenbeschichtungen, Laborbericht no. 6446, Plüss-Staufer, Oftringen, Schweiz, 1988.“Keramikteilchen eingearbeitet in schützendem Vliesstoff,” *Vliesstoff Nonwoven Int.* **7** (1993) 87.
634. DIN 55943, Farbmittel, Begriffe, 1993.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Textile Dyeing

Herbert Leube, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1, 7, 8, 10, 11, Sections 4.7 – 4.9)

Wilhelm Rüttiger, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2, 3)

Gerd Kühnel, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 4.1)

Joachim Wolff, Bayer AG, Leverkusen, Federal Republic of Germany (Section 4.2)

Günther Ruppert, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Sections 4.3, 4.4)

Michael Schmitt, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Sections 4.3, 4.4)

Christian Heid, Cassella AG, Frankfurt, Federal Republic of Germany (Section 4.5)

Max Hückel, Hoechst Aktiengesellschaft, Offenbach, Federal Republic of Germany (Section 4.6)

Hans-Joachim Flath, Dresden, Federal Republic of Germany (Chaps. 5, 6)

Wolfhard Beckmann, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 9)

Rolf Brossmann, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 12)

Manfred Söll, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 13)

Ulrich Sewekow, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 14)

1. History, Economic Importance . 6
1.1. Historical Dyeing Methods 6
1.2. Economic Importance of Textile Dyeing 7
2. Dyeing Technology 8
2.1. General 8
2.1.1. History 8
2.1.2. The Field of Dyeing Technology . 8
2.1.3. Fundamental Principles of Dyeing 9
2.1.3.1. Dyeing Systems 9
2.1.3.2. Phases of Exhaustion Dyeing . . . 9
2.1.3.3. Dyeing Phase (Dyeing Kinetics) . 9
2.1.3.4. Equilibrium Phase 12
2.1.3.5. Dye Fixation, Improvement of Colorfastness 12
2.1.3.6. Sources of Further Process Data . 13
2.2. Batchwise Dyeing (Bath Dyeing) 13
2.2.1. Fundamental Principles and Equipment 13
2.2.2. Theoretical and Technical Fundamental Principles 14
2.2.3. Circulating Machines (Stationary Goods, Circulating Liquor) 16
2.2.3.1. Systems and Functions 16
2.2.3.2. Loose Stock Dyeing Machines . . 18
2.2.3.3. Package Dyeing Machines (Cross-Wound Packages) 18
2.2.3.4. Hank Dyeing Machines 19
2.2.3.5. Beam Dyeing 20
2.2.4. Circulating-Goods Machines with Textile Storage (Winch Type) . . . 20
2.2.4.1. System and Functions 20
2.2.4.2. The Winch Beck 21
2.2.4.3. Jet Dyeing Machines 22
2.2.4.4. Overflow Dyeing Machines 22
2.2.4.5. The Air Jet ("Airflow") Dyeing Machine 23
2.2.5. The Dyeing Jigger 23
2.2.5.1. Normal (Direct) Jig Dyeing 23
2.2.5.2. Pad Jig Process 25
2.2.6. Special Bath Dyeing Equipment . 26
2.2.6.1. Star-Shaped Dyeing Frames 26
2.2.6.2. Machines for Dyeing Hanks of Yarn 26
2.2.6.3. Paddle Dyeing Machine 27
2.2.6.4. Rotary Dyeing Machine 27
2.2.6.5. Cabinet Dyeing 27
2.2.6.6. Hosiery Dyeing Machines 27
2.2.7. Automatic Control of Bath Dyeing 27
2.2.7.1. Aims 27
2.2.7.2. Functions of Automatic Control . 28
2.2.7.3. Equipment Requirements 30
2.3. Continuous and Semicontinuous Dyeing 31
2.3.1. The Principal Stages of Continuous Dyeing 31
2.3.1.1. Dye Pickup 31
2.3.1.2. Intermediate Drying 33

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_351

2.3.1.3. Dye Fixation 35
2.3.1.4. Aftertreatment of the Dyed Fabric (Finishing) 38
2.3.2. Dyeing Plants 38
2.3.3. Continuous Dyeing of Yarn and Fiber 40
2.3.4. Automatic Operation of Continuous Dyeing Plants 41
2.3.4.1. Important Process Stages and their Automation 41
2.3.4.2. Technology of Automation 43
2.4. Laboratory Dyeing Techniques . 44
2.4.1. Objectives 44
2.4.2. Laboratory Dyeing 44
2.4.2.1. Typical Laboratory Equipment . . 44
2.4.2.2. Small-Scale Production Equipment 45
2.4.3. Laboratory Dyeing Technology . . 45
2.5. Techniques of Dispensing Products used in Dyeing 46
2.5.1. Dispensing of Dyes 46
2.5.2. Dispensing of Dye Auxiliaries . . 47
2.5.3. Dispensing of Chemicals 48
2.5.4. Preparation of the Initial Liquor Charge and its Replenishment . . . 48
2.5.4.1. Batch Dyeing 48
2.5.4.2. Continuous Dyeing 48
2.6. Colorimetry 49
2.6.1. Measuring Instruments 49
2.6.2. Methods of Expressing Colorimetric Results 49
2.6.3. Developments in Colorimetry . . . 50
3. Physical Properties of Textiles Important for Dyeing 50
3.1. Classification of Textile Properties 50
3.2. Fibers 51
3.3. Yarns 52
3.4. Fabrics 52
3.5. Makeup of Textiles for Dyeing . 53
4. Dyeing of Cellulose Fibers 53
4.1. Dyeing with Reactive Dyes 53
4.1.1. Fundamentals 54
4.1.2. Dyeing Techniques 56
4.1.3. Special Processes and Development Trends 58
4.2. Dyeing with Direct Dyes 59
4.2.1. Applications and Properties 59
4.2.2. Dyeing Principle 60
4.2.3. Pretreatment of Substrates 60
4.2.4. Dyeing Parameters 60
4.2.5. Dyeing Techniques 61
4.2.6. Special Processes 61
4.2.7. Aftertreatment 61
4.3. Dyeing with Anthraquinone Vat Dyes 62
4.3.1. Chemistry of Vat Dyes 62
4.3.2. Vatting 62
4.3.3. Dye Absorption in the Exhaustion Process 63
4.3.4. Oxidation 64
4.3.5. Aftertreatment (“Soaping”) 64
4.3.6. Dyeing Techniques 64
4.4. Dyeing with Indigo 65
4.4.1. Chemistry and Historical Development 65
4.4.2. Dyeing Technique 65
4.5. Dyeing with Sulfur Dyes 66
4.5.1. Types and Mode of Reaction . . . 67
4.5.2. Additives to the Dye Bath 67
4.5.3. Dyeing 68
4.5.4. Dyeing Techniques 70
4.5.5. Combination with Other Dye Groups 71
4.5.6. Wastewater 72
4.6. Dyeing with Naphtol AS Dyes . . 72
4.7. Dyeing with Pigments 74
4.8. Other Dyeing Methods 74
4.8.1. Dyeing with Leuco Esters of Vat Dyes 74
4.8.2. Dyeing with Mordant Dyes 75
4.8.3. Dyeing with Acid Dyes 75
4.8.4. Dyeing with Basic Dyes 76
4.8.5. Dyeing with Mineral Dyes 76
4.8.6. Dyeing with Oxidation Dyes . . . 76
4.8.7. Dyeing with Phthalogen Dyes . . . 76
4.8.8. Dyeing with Coupling and DiazotizationDyes 77
4.9. Other Plant Fibers 77
5. Dyeing of Wool and Silk 77
5.1. Dyeing of Wool 77
5.1.1. Principles 77
5.1.2. Dye Classes and Dyeing Processes 78
5.1.2.1. Acid Dyes 78
5.1.2.2. Chrome Dyes 79
5.1.2.3. 1:1 Metal-Complex Dyes 80
5.1.2.4. 1:2 Metal-Complex Dyes 80
5.1.2.5. Reactive Dyes (. 81
5.1.2.6. Vat Dyes, Leuco Esters of Vat Dyes 82
5.1.3. Technology of Dyeing 82
5.1.4. Properties of Dyeings 82
5.1.5. Dyeing of Wool – Polyamide Blends 83
5.2. Dyeing of Silk 83
5.2.1. Fiber Structure and Dyeing Behavior 83
5.2.2. Classes of Dyes 84
5.2.3. Technology of Dyeing 84

6. Dyeing of Polyamide Fibers . . . 84
6.1. Chemical Structure 84
6.2. Supermolecular Structure 85
6.3. Interactions Between Dye and Fiber 85
6.4. Dyes 86
6.4.1. Disperse Dyes 86
6.4.2. Acid Dyes 87
6.4.3. 1:2 Metal-Complex Dyes 88
6.4.4. Reactive Dyes 89
6.5. Technology 89
6.6. Properties of Dyed Materials . . 90
6.7. Selection of Dyes 90
6.8. Toxicological and Physiological Aspects 90
7. Dyeing of Polyester Fibers 91
7.1. General 91
7.1.1. Dyeing in Aqueous Liquor 91
7.1.2. Thermosol Process 93
7.2. Dyeing Processes for Polyester Fibers 94
7.2.1. Requirements to Be Met by Dyes and Auxiliaries 94
7.2.2. Dyeing from Aqueous Dye Baths 95
7.2.3. Special Dyeing Processes 96
7.2.4. Continuous and Semicontinuous Dyeing Processes 96
7.2.5. Dyeing of Microfibers 98
7.2.6. Dyeing of Modified Polyester Fibers 98
7.3. Aftertreatment 98
8. Dyeing of Cellulose Acetate . . . 99
8.1. Dyeing Processes for Cellulose Acetate 99
8.2. Dyeing Processes for Cellulose Triacetate 99
8.3. Aftertreatment, S-Finish 100
9. Dyeing of Acrylic (Polyacrylonitrile and Modacrylic) Fibers . 100
9.1. Economic Aspects 100
9.2. Principles 101
9.2.1. Disperse Dyes 101
9.2.2. Cationic Dyes 101
9.3. Exhaustion Process 104
9.4. Dyeing of Special Fiber Types According to the Exhaustion Process 105
9.5. Continuous Processes 106
10. Dyeing of Other Synthetic Fibers 107
10.1. Poly(Vinyl Chloride) Fibers . . . 107
10.2. Elastomeric Fibers 107
10.3. Polypropylene Fibers 108
10.4. Poly(Vinyl Alcohol) (PVA) Fibers 108
10.5. Aramid Fibers 109
11. Dyeing of Fiber Blends 109
11.1. Polyester – Cellulose Blends . . . 110
11.1.1. General Finishing Information . . 110
11.1.2. Dyeing Processes 111
11.1.2.1. Dyeing with Disperse and Vat Dyes or Appropriate Mixed Dyes 111
11.1.2.2. Dyeing with Disperse and Reactive Dyes 111
11.1.2.3. Dyeing of the Cellulose Component with Other Dyes . . . 112
11.2. Polyester – Wool Blends 113
11.2.1. Dyes 113
11.2.2. Dyeing Processes 113
11.2.3. Auxiliaries 114
11.3. Polyamide – Cellulose Blends . . 114
11.4. Polyacrylonitrile – Cellulose Blends 115
11.5. Polyacrylonitrile – Wool Blends 115
11.6. Blends of Synthetic Fibers 116
12. Characterization of Textile Dyes 117
13. Testing of Colorfastness 119
13.1. Organization 119
13.2. Standard Test Methods 119
13.2.1. Group A: General Principles . . . 119
13.2.2. Group B: Colour Fastness to Light and Weathering 120
13.2.3. Group C: Colour Fastness to Washing and Laundering 120
13.2.4. Group D: Colour Fastness to Dry Cleaning 121
13.2.5. Group E: Colour Fastness to Aqueous Agencies 121
13.2.6. Group G: Colour Fastness to Atmospheric Contaminants 123
13.2.7. Group J: Measurement of Colour and Colour Differences 123
13.2.8. Group N: Colour Fastness to Bleaching Agencies 123
13.2.9. Group P: Colour Fastness to Heat Treatments 124
13.2.10. Group S: Colour Fastness to Vulcanization 124
13.2.11. Group X: Colour Fastness to Miscellaneous Agencies 125
13.2.12. Group Z: Colourant Characteristics 126
14. Legal Aspects, Toxicology, and Ecology 126
14.1. Water Consumption 127
14.2. Legal Regulations 127
14.3. Components of Textile Wastewater 128
14.4. Reduction of Wastewater Load by Special Preventive Measures 130

14.5. Special Methods of Wastewater Treatment 131
14.6. Recycling of Water and Valuable Substances 133
14.7. Ecotoxicology of Textile Dyes . . 134
15. References 135

1. History, Economic Importance

1.1. Historical Dyeing Methods

The technique of dyeing textiles can be traced back to prehistoric times. Archaeological studies have shown that elaborate chemical – technological dyeing processes were known to all ancient highly civilized cultures. The first written documents containing detailed dyeing instructions are the papyri Leidenenses and Holmiensis dating from the 3rd century A.D. [1]. The Plictho of Rosetti is considered to be the oldest printed book on the art of dyeing [2]. In preindustrial times, textiles were dyed primarily with plant dyes [3–6]. Dyeing with inorganic pigments or the precipitation of metal salts on fibers (mineral dyes), which is still practiced today, was of only secondary importance. Also, dyeing with animal dyes was restricted to special cases. Well known are ancient Tyrian purple (6,6′-dibromoindigo), which is extracted from the Mediterranean purple sea snail, and kermes and cochineal, which contain tetrahydroxyanthraquinone carboxylic acid as the principal dye. These substances are obtained from insects and require a metallic mordant for dyeing.

Dyes isolated from plants, such as indigo from the Indian indigo plant or dyer's woad (→ Indigo and Indigo Colorants) and Turkey red, alizarin (→ Anthraquinone Dyes and Intermediates), from the root of the madder plant, were of greater significance. Dyeing black with iron salts and tannic acid from the bark of oak was important. Until recently, campeachy wood black, a pyrone dye that occurs in a special Brazilian wood, was used because of its rich hue. Yellow was frequently produced by using dyes of the flavone family, which occur in flowers and roots, e.g., weld (*Reseda luteola* L.), or carotinoids, e.g., crocin from saffron. Dyes from red berries (anthocyans) and the orchella weed (archil) gave dyeings that were less fast. Plant extracts were also used to produce the color green.

Apart from the actual dyeing process, the work of the dyer often included isolation and preparation of dyes from plants that were either imported or grown locally. As a chemical reaction with rather impure raw materials, the dyeing process was complicated and unclear. In fact, dyeing was an art. Indigo was produced by fermentation from plants and had to be applied to the fiber in the reduced form. The reduction step was performed in a fermentation vat with putrefied urine and later with iron sulfate and lime (vitriol vat). Turkey red gave especially fast dyeings as a lake (metal complex) only when the fabric was pretreated with metallic salts. Alum, tin, iron, and copper salts were most commonly used as mordants. The colors obtained depended on the skill and dexterity of the dyer. The old dyeing processes are still being used in the arts and crafts, the substrate being mainly sheep's wool [7, 8]. For ecological reasons, particularly, the old dyeing processes do not meet the requirements of modern industrial production.

In the 18th century, chemical studies dealt with the nature of color lake formation, the processes involved in vatting, the suitability of individual coloring substances for certain fibers, and making up recipes for particular shades with different dyes. Furthermore, new dyeing and printing techniques were developed in the 18th and early 19th centuries. Apart from providing new color effects, these techniques focused on producing fast dyeings and largely contributed to the upswing in textile manufacture [9]. A review of the dyeing methods commonly used at the end of the 19th century is given in [10] together with many individual recipes based on publications in *Reimanns Färbezeitung* (1872 – 1879).

1.2. Economic Importance of Textile Dyeing

The worldwide consumption of fibers by the textile industry is shown in Table 1 (see also, → Fibers, 1. Survey, Chap. 7.). The importance

of regenerated cellulose fibers has decreased and that of polyester fibers has clearly increased.

Table 1. World fiber consumption

Type of fiber	1980		1992	
	10^6 t	%	10^6 t	%
Cotton	14.1	47	18.7	47
Viscose	3.2	11	2.4	6
Wool	1.6	5	1.6	4
Polyamide	3.2	11	3.9	10
Polyester	5.1	17	8.3	21
Polyacrylonitrile	2.1	7	2.2	5
Polypropylene			1.8	4
Others (silk, linen etc.)	0.5	2	1.3	3

Developments since the mid 1970s show that Europe and North America are losing their share of the textile finishing industry and East Asia's share is becoming more significant (Table 2). The growing industrialization of China will enhance this trend.

Table 2. Regional distribution of textile processing, %

Region	1976	1992
Western Europe	17	12
Eastern Europe	19	11
East Asia	35	44
North America	19	16
South America	6	6
Africa, West Asia	4	7

The amount of fiber apportioned to dyeing is closely connected to the consumer and fashion habits of the population. It can be assumed that ca. 75 – 80 % of the world fiber consumption consists of colored materials; the rest remains undyed or is used for technical articles. Sulfur and reactive dyes for cellulose fibers and disperse dyes especially for polyester fibers are quantitatively in the lead. The estimated world consumption of synthetic textile dyes for dyeing and printing in 1992 (in 10^3 t) is given below:

Indigo	12
Vat dyes	26
Reactive dyes	108
Direct dyes	45
Naphtols	19
Sulfur dyes	100
Cationic dyes	21
Anionic dyes	74
Disperse dyes	102
Pigment preparations	40
Total	547 (for comparison 1980: 450)

More than 80 % of the total consumption of 547 000 t of dye is used for dyeing and the rest for printing.

In textile finishing in Germany (west), the quantitative importance of individual products was shown by an investigation made by the University of Münster [11] (Table 3).

The methods and processes of the textile finishing industry must adapt to these market requirements. As a result of increasingly shorter yardages, the trend is toward batchwise production and away from the more economical continuous methods. However, cotton and polyester – cotton fabrics as well as certain carpet grades are still dyed in large quantities in continuous processes.

The theoretical principles of dyeing are discussed with individual dye and fiber types.

2. Dyeing Technology

2.1. General

2.1.1. History

The very first method used to dye clothing was probably with pigments (clay or ocher as the dye and fat or glue as the binder) [12, 13]. During

Table 3. Textile dyeing and finishing in Germany (West) 1992

Application	Flock (27 000 t), %	Yarn (119 000 t), %	Fabric ($1\,800 \times 10^6$ m^2), %	Knitwear (110 000 t), %	Print (600×10^6 m^2), %
Clothing	11				60
Outer clothing		42	33	31	
Lining materials			17		
Furnishings	43	16	28		33
Underwear		4	3	34	
Industrial textiles, especially automotive	41	11	15	23	4
Others	5	27	5	12	3

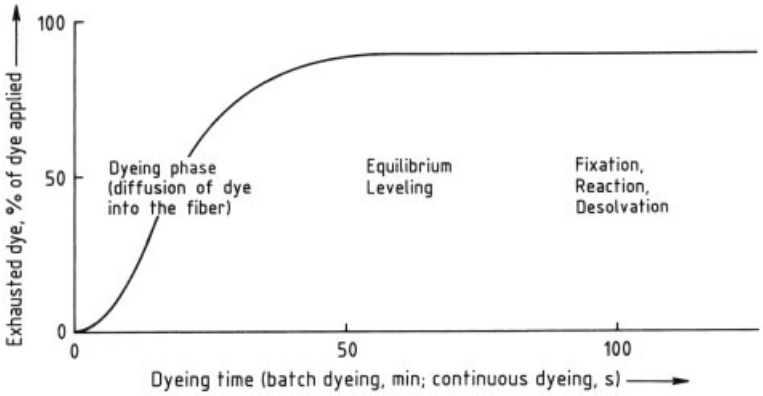

Figure 1. Phases of exhaustion dyeing

Old Testament times dyebaths were used, dyeing was performed partly in open lead vessels with fires underneath to heat the liquor [14].

For a long period, the main method of dyeing was in open stone or wood vessels in which the yarn was agitated in hank form. The use of a hand-operated rotating mechanism (reel) over the dyebath facilitated the dyeing of fabrics (piece goods).

The use of power driven equipment to move the goods and to drive liquor pumps began in the early 19th century. Also, the first continuous dyeing plants were built at this time, these being in principle identical to those still in operation today [15].

2.1.2. The Field of Dyeing Technology

Dyeing technology matches a dyeing process to the goods and the dyeing machines. Dyeing technology also includes laboratory work, necessary not only for the development of formulations (sometimes an extremely difficult task), but also contributing to the development and realization of dyeing processes.

Color measurement for research and production has become an important and quite separate part of dyeing technology. Color measurement is mainly used as an aid in the development of formulations and for the measurement of undesired color variations of the textile material. Another important aspect of dyeing technology is automatic metering. Although the metering of auxiliaries is already widely used, the considerably more difficult metering of dyes is much less widespread.

Dyeing technology is a rapidly expanding area of textile dyeing. Suitable dyeing technology can determine not only the feasibility of manufacturing a new article but also its production costs, and can be essential for the continued existence of some business enterprises.

2.1.3. Fundamental Principles of Dyeing

2.1.3.1. Dyeing Systems

In the field of dyeing technology no generally accepted system of classification exists as yet. Based on the method of introducing the dye into or onto the textile fiber three methods of dyeing can be distinguished:

Exhaustion dyeing
Diffusion of the dissolved dye into the fibers (batch and continuous dyeing) [16]
Pigment dyeing
Deposition of insoluble dye onto the fiber, and fixation with a binder [17]
Mass and gel dyeing
Incorporation of the dye during production of synthetic fibers

The market share of exhaustion dyeing exceeds that of mass dyeing, which in turn is more important than pigment dyeing.

As exhaustion dyeing is the most important process in terms of market share, the discussion of fundamental principles is limited to this process. Continuous pigment dyeing is mentioned in Section 2.3.

2.1.3.2. Phases of Exhaustion Dyeing

In exhaustion dyeing, a dye which must be at least partially dissolved migrates by diffusion through the surface of the fiber and into its interior. The dye is usually in a liquid medium (the liquor), which also contains the textile material [17].

The exhaustion dyeing process can be divided into three phases (Fig. 1). Each phase is characterized independently of the other phases as far as possible:

The dyeing phase (exhaustion and absorption)
Diffusion of the dye into the fiber (dyeing kinetics)
The equilibrium phase
The final dye concentration on the fiber (dye yield) is nearly reached; leveling and dye penetration occur at this stage
The dye fixation phase
Chemical processes, diffusion or "desolvation" of the dye, fastness improvements

2.1.3.3. Dyeing Phase (Dyeing Kinetics)

In this phase, the dye diffuses from the liquor into the textile fiber. In most cases, soft water is used as the liquid medium. Despite intensive research, organic solvents (chlorinated hydrocarbons) have not become established as dyeing media, and air – water foam systems have been even less successful [18]. However, some interesting experiments to use supercritical CO_2 as liquor have been carried out (see Section 7.2.3) [19, 20].

There are essentially two techniques available for dyeing the textile material:

1) *Batch Technique:* The liquor and the textile are placed together in a vessel, and the required amount of dye is added.
2) *Continuous Technique:* The dye is dissolved or dispersed in the liquor. A definite quantity of the dye liquor is locally applied to the textile.

In both cases, the dye subsequently diffuses into the textile fibers. Normally, the rate and uniformity of this diffusion are controlled.

The diffusion rate of the dye is normally proportional to its concentration difference between the liquor and the fiber (difference in chemical potential as the driving force for the mass transport, see also Chap. 11) [17].

During the exhaustion process, the concentration of the dye in the liquor continuously decreases, so that the concentration difference is also reduced. The dyeing rate therefore continuously slows down and reaches a final value (equilibrium).

The terms "exhaustion" and "absorption" are equivalent in many ways. However, the differences should be noted. Strictly speaking, the *exhaustion* is the decrease in the amount of dye in the bath, because of absorption of the dye by the textile and other processes such as precipitation of the dye. The *absorption* is the deposition of dye in the textile substrate. Absorption thus includes only the dye taken up by the textile material not dye residues or sedimented dye.

In the continuous technique, the term *dye fixation* or *degree of fixation* is often employed. This is the ratio of the dye deposited in the textile substrate to the total amount of dye applied.

In the following, the term exhaustion is used in a general sense, and absorption if the textile substrate is mainly considered.

The characteristic parameter that describes the dyeing properties of a dye in a dyeing system is the rate of dyeing or "strike". The dyeing rate is defined as the quantity of dye absorbed by the textile substrate per unit of time. The *absolute dyeing rate* is expressed in terms of the amount of dye in g/kg or wt % based on weight of fabric (wof). However, the end of the exhaustion process is often of more significance, and hence the rate at which dyeing equilibrium is approached, i.e., the relative dyeing rate. The amount of dye uptake in this case is expressed as percentage of the amount of dye added.

Example:
Amount of dye added: 1.8 % = 18 g/kg goods
Absolute dyeing rate = 0.4 g kg^{-1} min^{-1} (measured value)
Percentage: (0.4/18) × 100 = 2.2 % of exhausted dye per minute
Relative dyeing rate = 2.2 %/min (calculated)

If the relative dyeing rate (2.2 %/min) were kept constant, the dyeing time would be 100 % : 2.2 %/min = 45 min.

The *instantaneous dyeing rate* can be read off from an exhaustion curve. The slope of the curve

gives the value of the instantaneous (relative) dyeing rate, steeper curves indicating higher dyeing rates. Relative dyeing rates in practice fall within two very different ranges:

0.5 – 5 %/min for the batchwise process
50 – 150 %/min for the continuous process (i.e., 100 % can be absorbed in< 1 min.)

What are the causes of these well-established limits? Dwell times in a particular type of dyeing equipment are linked to the relative dyeing rates given above. They are as follows:

20 – 200 min for the batchwise process
0.7 – 2 min for the continuous process.

These are the mechanically imposed capabilities of the dyeing techniques available. (The semicontinuous process, with a dwell time of several hours, is discussed in Section 2.3.) Although the actual dyeing mechanism remains the same and also many dyes are used in both processes, significant differences between the two processes exist.

Batchwise exhaustion dyeing uses a closed system (dyebath) which contains the textile and the liquor. The dye is added in accordance with the formulation. An intense interaction between the textile and the liquor then occurs. Detailed descriptions of systems are given in Section 2.2.3.

General technical data for batchwise dyeing techniques are given below:

Mean dyeing rate	0.5 – 5 %/min
Liquor ratio ("liquor/goods")	4:1 – 25:1 (i.e., 4 – 25 L/kg goods)
Dye concentrations (2 %, i.e., 20 g dye/kg goods)	0.8 – 5g/L
Dyeing times (exhaustion)	20 – 200 min

The dyeing rate and hence the dyeing time can be controlled within wide limits: The (relative) dyeing rate increases with increase of dyeing temperature, fiber swelling, and carrier concentration, and with decreasing dye concentration, size of dye molecule (higher diffusion coefficient of smaller molecules), retarder concentration, and often decreasing liquor ratio.

Figure 2 illustrates typical exhaustion or absorption curves for batch dyeing. In general, a linear absorption curve gives better leveling.

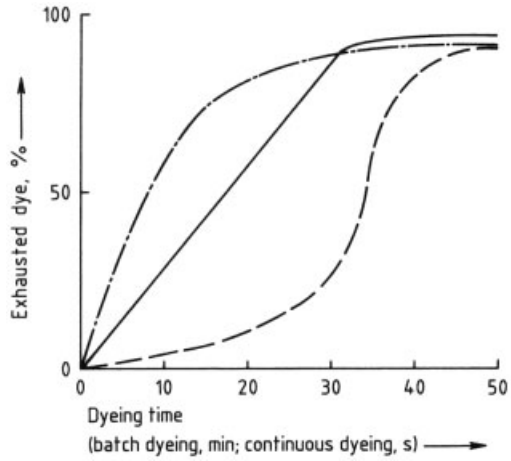

Figure 2. Kinetics of exhaustion for three different cases The relative dyeing rate is given by the slope of the curves –·–·– exponential dyeing curve; —- linear dyeing curve; ––– S-shaped dyeing curve

The batchwise dyeing method shows a tendency to unlevel results. Color correction can be carried out, giving a more reliable dyeing result. Higher dyeing rates are usually associated with poorer leveling properties [21, 22]. An example of this is the vat dyeing of cellulose fibers. This dyeing usually proceeds so rapidly that an unlevel product results even in modern machines. However, this can be very readily leveled out (see Sections 4.3.6 and 4.4.2).

The batchwise dyeing process has a very large share of the market. It is now a very modern and highly automated technique.

Continuous Exhaustion Dyeing. In continuous exhaustion dyeing the dye is first dissolved or dispersed in the liquor, and the dye liquor so produced is locally applied in definite amounts to the textile which continuously passes through. The diffusion transfer of the dye to the textile substrate is controlled by effects on the textile goods, e.g., by temperature and/or auxiliaries.

Detailed descriptions of dyeing systems are given in Section 2.3.

General technical data for the continuous dyeing technique are given below:

Mean dyeing rate	50 – 150 %/min
Liquor ratio ("liquor/textile")	0.4:1 – 1.2:1
Dye liquor loading	40 – 120 %
Dye concentration (2 %)	17 – 50 g/L
Dyeing times	0.6 – 2 min

(The liquor ratio is given only for comparison with the batchwise technique. This parameter is not normally used.)

If the fibers are hygroscopic, diffusion into the textile frequently occurs immediately after adding the fabric to the dyebath. Sometimes auxiliaries must be used to solubilize the dye and impart its substantivity. Strong heating (e.g., to 100 °C in a steamer) considerably speeds up the process. Some continuous processes also proceed in the cold (cold dwell process, semicontinuous technique).

After the dye has been applied, the textile is often first dried (intermediate drying). During intermediate drying the dye in the liquor migrates with the water into the surface zones of the textile (color intensification by impaired dye penetration, see Chap. 3). If this usually undesired migration is nonuniform, it can lead to unlevelness (at selvedges and ends).

With thermoplastic fibers, true absorption occurs at high temperature by thermosolvation of the dye in the fiber (thermosol process, see Section 7.2.4).

With hygroscopic fibers, developing agents must be applied in aqueous solution. The process then proceeds as described above (see Section 4.3.6). Here, the dyeing rate can be determined partially or entirely by the preliminary reaction, e.g., reduction.

Many proposals have been made aimed at improving the dyeing technique, but none of these has found widespread practical application even after many years. For example, ultrasound is claimed not only to improve the dyeing rate, but also to increase the dyeing yield [23].

In the continuous technique, corrections (e.g., of color discrepancies and unlevelness) are very difficult and expensive. However, the conversion of white goods to finished dyed textiles in a single operation is an elegant and hence very attractive process. It is especially important in dyeing big lots (high total yardages).

2.1.3.4. Equilibrium Phase

The equilibrium phase of dyeing follows the exhaustion phase without any sharp transition. It is reached in practice when the dye concentration in the liquor undergoes no further appreciable change [17].

Equilibrium is characterized by the terminal degree of exhaustion of the bath or the terminal dye yield as a percentage of the added dye. It is usually in the range 50 – 99.9 %. The dye still remaining in the liquor increases production and wastewater costs. An important task of dyeing technology is to reduce the amount of this dye.

The dye yield increases with decrease of liquor temperature, liquor ratio, retention effect of the auxiliaries, and dye concentration. The yield increases with increasing salt concentration, substantivity of the dye, affinity of the textile substrate, and build-up properties of the dye.

The most important relationships of the various classes of dye are well known. For example, the distribution coefficient K_D is very useful for characterizing the position of the dye equilibrium [24].

If C_F represents the concentration of the dye in the fiber material (solid phase) in g/kg and C_L represents the concentration of the dye dissolved in the liquor (liquid phase) in g/L, the distribution coefficient in L/kg is

$$K_D = \frac{C_F}{C_L}$$

The value of K_D can range between 10 L/kg for reactive dyes without additional salt to > 10 000 L/kg for cationic dyes on polyacrylonitrile.

The distribution coefficient is independent of the liquor ratio in a given dyeing process, and is therefore a suitable characteristic number for a dye (substantivity number). However, it is not an invariant, as it depends on the dye concentration (build-up properties of the dye), temperature, pH, addition of salt, and various other parameters (see Chaps. 4, 5, 6, 7, 8, 9, 10, 11 for details) (see also Fig. 3).

When dyeing equilibrium is reached, the dyeing result is often still unlevel. However, the dye that is not yet fixed can then migrate from the areas that are too dark to the lighter areas. This equalization or leveling usually occurs more rapidly with larger residual amounts of dye in the liquor and at higher temperatures.

Table 4 shows that for high distribution coefficients (very high substantivities), the dye yield differences are extremely small (< 1 %), although the concentration of residual dye in the liquor differs by a factor of two. Residual dye

concentrations of 8 – 16 mg/L are almost too low for leveling to take place.

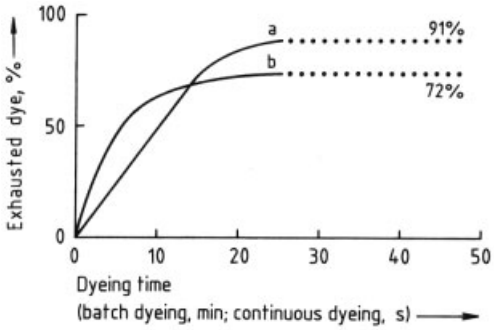

Figure 3. Dyeing yield as function of dyeing time and temperature
At equilibrium for a liquor ratio of 10:1 a distribution coefficient of 91:0.9 = 100 L/kg is obtained at 60 °C and at 80 °C a distribution coefficient of 72:2.8 = 26 L/kg.
a) Lower temperature; b) Higher temperature
· · · = Equilibrium range

Table 4. Variation in residual dye content (wastewater) for 2 % dye addition, while the difference in dye uptake (yield) is very small; liquor ratio 10:1 (10 L/kg)

Dyeing run no.	Yield	Residual dye	Distribution coefficient
A	99.2 %	0.8 % (16 mg/L)	1215 L/kg
B	99.6 %	0.4 % (8 mg/L)	2430 L/kg
Difference	0.4 %	factor of 2	

With certain leveling auxiliaries, the amount of residual dye (the so-called dye retention) can be very effectively controlled (see Chaps. 4, 5, 6 7, 8, 9, 10, 11 for details).

2.1.3.5. Dye Fixation, Improvement of Colorfastness

When the dyeing process has reached equilibrium, the dye is usually still in the outer zones of the textile fibers. It is mobile, and can bleed out again at any time (a process taking place during leveling). The dye must be fixed, a process also known as development, finishing, or aftertreatment.

The treatment required is primarily dependent on the chemistry of the dye. The following fixation methods exist:

1) Fixation of the dye by chemical reaction, e.g., oxidation of soluble dyes to give insoluble pigments, (vat dyes, see Sections 4.3.4, 4.4.2) or substantive bonding to OH groups of the cellulose (reactive dyes, see Section 4.1.1)
2) Fixation by improved deposition within the fiber substrate, e.g., by deeper penetration into the fibers (disperse dyes, see Sections 6.4.1, 9.2, 11.1.2, Chap. 7) or increased aggregation, e.g., by soaping (vat dyes) or "precipitation" of anionic dyes by cationic auxiliaries (see, e.g., Section 4.2.7)

In general, a further washing operation is also required after fixation to completely remove the residual dye and auxiliary agents to achieve satisfactory colorfastness.

2.1.3.6. Sources of Further Process Data

The computation of dyeing processes with the aid of process functions and dye data is only slowly becoming established. The most important delaying factor is the complexity of the dyeing processes. Data currently available (e.g., from sample information sheets) are usually presented in the traditional way, and parameters suitable for computer use are not provided.

The sections on individual dyeing processes give some of these data or at least the relevant literature references, e.g., for vat dyes [25, 26] or reactive dyes [27–30].

Different authors use different concepts and parameters, e.g., [27, 31]. In order to utilize the information, careful consideration of the system is required, with mathematical conversion of the data if this is at all possible.

Computer programs should be developed that can simulate the dyeing process and depict it on the computer screen, even when the information available is somewhat incomplete. Simulation on the computer screen, however, involves very complex programming [32, 33].

"Expert systems" (ES) go far beyond data and literature banks [34]. These give extremely rapid access to stored technical information (knowledge-based or rule-based systems).

2.2. Batchwise Dyeing (Bath Dyeing)

Batchwise dyeing is performed in an essentially closed system in which the textile material, the

Table 5. Forms of textile goods and appropriate dyeing equipment

Type of dyeing equipment	Loose stock (also card sliver and tow)	Yarn (bobbins, hanks)	Textiles (woven, knitted)	Finished garments
Circulating machines (Section 2.2.3)	package dyeing machines	bobbin, hank, and warp beam dyeing machines	beam dyeing machines	hosiery dyeing machines
Circulating-goods machines (Section 2.2.4)		hank dyeing machines	winch and jet dyeing machines, jiggers	paddle and drum dyeing machines

liquor, and the dye interact over a long period. The essential elements of this dyeing system are the dyebath and the storage of the textile material.

2.2.1. Fundamental Principles and Equipment

The term "bath dyeing" describes this dyeing technique more accurately than the term "batch dyeing" more commonly used today. Bath dyeing is a special type of exhaustion dyeing. In the dyeing equipment, the three phases of exhaustion dyeing – exhaustion, establishment of equilibrium, and dye fixation – usually occur in succession.

An important feature is the interaction (movement) of the textile material and the liquor to prevent unwanted concentration differences and consequently unlevelness (Section 2.2.2).

Dyeing equipment in which the textile material remains stationary and the liquor is passed through the textile material, e.g., by pumps, is called *circulating machine*. The most important types of circulating machines are described in Section 2.2.3.

However, if the textile material is moved, e.g., mechanically or hydraulically, the term "*circulating-goods machinery*" is used, even if the liquor is also moved. The most important examples of this are described in Sections 2.2.4 and 2.2.5.

Special types of equipment are described in Section 2.2.6.

Bath dyeing has the great advantage that it can be used at any stage of textile production, i.e., on the fiber, the yarn, the textile, or the finished garment. Approximately since the International Textile Machinery Exhibition (ITMA) 91 there has been a tendency to carry out the dyeing as near to the end of the production chain as possible, i.e., moving in the direction of the finished manufactured garment. A general overview of forms of textile goods with the appropriate dyeing equipment is given in Table 5.

2.2.2. Theoretical and Technical Fundamental Principles

A particular advantage of bath dyeing is that wet goods can be treated in the dyebath. For example, woven textiles can be (continuously) desized, and can then be charged, while still wet, to the dyeing equipment, avoiding drying costs.

Absorption of the Dye (Dyeing Kinetics). Here, the dyeing rate is the dominant parameter (see Section 2.1.3.3). The dye must be absorbed uniformly to give a level result [22]. Two basically different strategies can be used to achieve this. These must be optimized, also in quite different ways:

1) Accelerated absorption, giving unlevel dyeing of the goods, followed by accelerated leveling (temperature, auxiliaries, pH). Thus, the absorption phase and the leveling in the dyeing equilibrium are controlled.
2) Retarded absorption (the dyed product is always level, and longer times are required for the absorption phase).

To predict the unlevelness in the absorption phase, the significant dyeing rate can be calculated from the absorption curve [21]. The further the dyeing rate of the textile – equipment system lies above the (critical) dyeing rate, the more severe is the unlevelness. Correction of unlevelness is usually carried out by a cumbersome and expensive leveling operation, by bleaching and redyeing, or simply by dyeing black.

Unlevelness is one of the major problems of short dyeing processes, because these require the use of high dyeing rates [35].

Since the dyeing rate depends on diffusion, it is controlled by the concentration of the dye

in the liquor. A high liquor ratio (low dye concentration) leads to a low dyeing rate. For example, to produce a concentration of red dye in the goods of 2 % (20 g per kilogram fiber), the starting concentration of the dye can be 4 g/L (for a liquor ratio 5:1) or 1.33 g/L (for a liquor ratio 15:1).

However, if the system used forms a dye reservoir, the concentration in the liquor remains constant as long as the dye is supplied quickly enough (an "infinite" bath). For example, the dispersed dye particles in polyester dyeing act as a reservoir that maintains a constant concentration of dissolved dye independently of the liquor ratio [36].

In general dyeing rate can be controlled by, e.g., heating profiles (time – temperature curves), addition of leveling agents, carriers, and retarders and by pH – time and metering – time profiles.

When these control tools are used, the most favorable shape of the absorption curves shown in Figure 2 is linear, S-shaped curves being the most critical. Further details are given in Chapters 4 – 11.

Dyeing Equilibrium in Bath Dyeing. The dyes used for textile dyeing are mostly strongly substantive, i.e., they form a strong bond to the textile substrate. Thus high dye concentration in the fiber C_F is in equilibrium with a low residual dye concentration in the liquor C_L. The simplest indicator of substantivity is the distribution coefficient $K_D = C_F/C_L$ (see Section 2.1.3.4). Other substantivity parameters also exist.

This is illustrated by the following examples:

Red R 2 %, distribution coefficient = 170 L/kg		
Liquor ratio	5:1	15:1
Dye concentration in the textile (C_F)	19.44	18.38 g/kg
Dye concentration in the liquor (C_L)	0.114	0.108 g/L
Dye yield	97.2 %	91.9 %
Dye losses (wastewater)	2.8 %	8.1 %
Yellow 2G 1 %, distribution coefficient = 400 L/kg		
Dye yield	98.8 %	96.4 %
Dye losses (wastewater)	1.2 %	3.6 %

In a combination dyeing operation, a change to a higher liquor ratio (5:1 → 15:1) would cause more red dye to remain in the liquor (8.1 % compared with 3.6 %), and would shift the shade towards yellow.

The factors influencing the position of the equilibrium (distribution) (see Section 2.1.3.4) can also be used for control purposes.

The results of a bath dyeing process are easily corrected if necessary. At the end of the process, the dyeing result can still be adjusted (shading) by making further additions. Also, time can be saved by interrupting the exhaustion process early, before a close approach to equilibrium (99.0 – 99.9 %) is achieved. The following typical figures illustrate this:

Bath exhaustion, %	64	86	95	98	99
Dyeing time, min	20	40	60	80	100

In this case, the exhaustion process is preferably interrupted after 60 min at 95 % exhaustion rather than after 90 or 100 min with virtually complete exhaustion.

If the textile substrate of the same chemical composition is nonuniform, e.g., of various origins, the steps taken to prevent unlevelness must be specifically matched to the chemistry of the fiber and the dye (→ Textile Auxiliaries).

To sum up, the equilibrium phase essentially determines the dyeing result, including the ultimate levelness of a dyed product.

Dye fixation is described in Section 2.1.3.5. Specific recommendations for the various classes of dyes and methods of obtaining the required colorfastness are dealt with in Chapters 4 – 11. Tests for colorfastness are given in Chapter 13.

Bath Dyeing Technology. The most important part of the process is the interaction between the textile and the dye liquor, especially as this is influenced by the design of the equipment. It characterizes the types of dyeing equipment, which are described in more detail here.

As some dyeing processes are performed above the boiling point of water, high pressure dyeing equipment has been designed in which dyeing can be carried out up to ca. 140 °C [high temperature (HT) equipment] [37]. Since the vapor pressure of water is 2.7 bar at 130 °C and 3.6 bar at 140 °C, the safety valves are usually set at 6 – 7 bar.

To allow for thermal expansion of the dye liquor (ca. 6 % between 20 and 100 °C), either a gas cushion is used in the dyebath or an overflow with a cooler and an open collection vessel which can then also be used to hold the additives (d in Fig. 4).

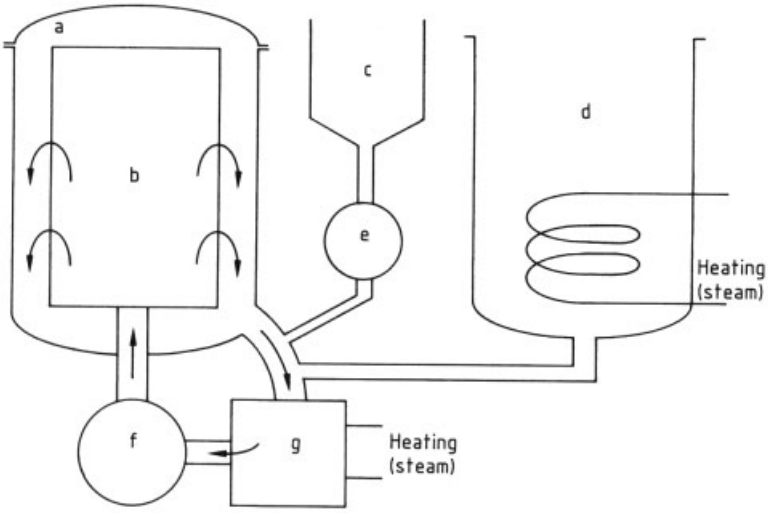

Figure 4. High temperature dyeing machine
a) Dyeing vessel, pressure-tight (water temperature: 140 °C), liquor volume: ca. 0.5 – 1 m^3 per 100 kg textile goods; b) Textile goods in package form (loose stock), on bobbins, or on a beam (woven, knitted, or warp yarn); c) Mixing vessel for additions to the dyebath (dyes, auxiliaries) (volume ca. 10 % of that of dyeing vessel); d) Auxiliary vessel for preparation of liquor (volume similar to that of dyebath); e) High pressure pump, delivery pressure: ca. 4 bar, delivery rate: ca. 2 m^3/h; f) Circulation pump showing flow through the goods (i→o), flow rate: ca. 50 – 200 m^3/h for 100 kg textile goods; g) Heat exchanger (usually steam) for the dyebath (external) and often heating coils in the dyeing vessel (internal)
Valves, device for reversing direction of circulation, and rinsing equipment are not shown.

2.2.3. Circulating Machines (Stationary Goods, Circulating Liquor)

2.2.3.1. Systems and Functions

The textile material, which can be prepared in the form of a roll or a pack, is placed in the apparatus and held there mechanically. The dye liquor is pumped through the goods. A circulating machine for HT conditions is shown in Figure 4, including the most important components [38].

An important dyeing parameter is the liquor flow rate through the textile goods maintained by the circulation pump (Fig. 4 f). In modern dyeing apparatus, the flow rate is displayed as a percentage (0 – 100 %), but this percentage usually represents the pump drive motor current and is therefore not a true flow measurement. The conversion of the percentage figure to the absolute liquor flow rate Q in m^3/h and to the specific flow rate q through the textile goods in L kg^{-1} min^{-1} is explained in Figure 5. The specific flow rate is a useful control parameter for the dyeing operation. Dyeing carried out in different types of apparatus (e.g., laboratory scale and full scale) with the same specific flow rate and the same liquor ratio usually gives similar results. This also applies within limits for dyeings in different apparatus with identical percentage figures from the pump display.

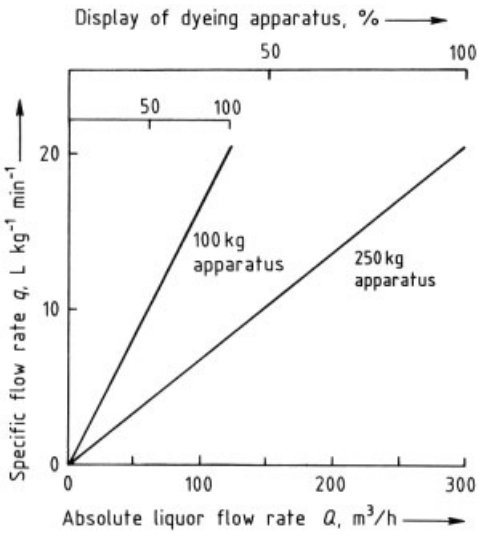

Figure 5. Characteristic figures for absolute liquor flow rate Q and specific flow rate q through the textile
The specific flow rate q is proportional to the percentage of absolute flow rate shown at the apparatus display; 100 % mostly corresponds to $q = 20 – 30$ L kg^{-1} min^{-1}.

The dye liquor should be contacted uniformly with the textile goods (no channeling and no extended flow paths, e.g., at the edges).

As the dye liquor flows through the textile goods, the dye exhausts gradually. The liquor flowing away from the textile goods therefore always has a lower dye concentration, and the textile has a lighter color at the outlet than at the dye inlet point. If this concentration difference is kept at $< 1-2\,\%$ by controlling dyeing rate and recirculation rate good leveling results can be expected.

The liquor flow usually has no effect on dyeing shades obtained (the dyeing equilibrium), especially if there is no sedimentation of the dye.

The characteristic curve of the pump is also an important characteristic of every circulating machine (see Fig. 6). It is usually nonlinear, and shows the liquor flow rate for a given pressure drop (differential pressure ΔP) across the textile goods [39]

$$Q = Q_{\max}\,(1 - \Delta P/P_{\max})\,m$$

where Q is the liquor flow rate (m^3/h); $Q_{\max}$ the liquor flow rate with no flow resistance (m^3/h); $P_{\max}$ the pressure at the textile goods when liquor flow is blocked (bar); and m the exponent of the pump characteristic curve, e.g., ca. 0.8.

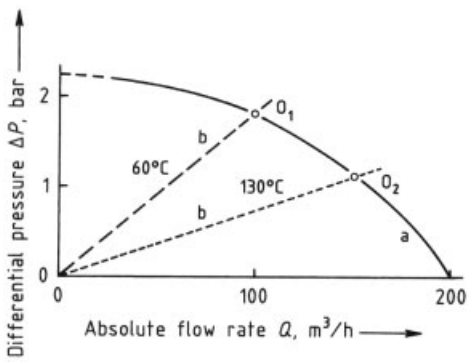

Figure 6. Pump characteristic curve and flow resistance of textile goods
a) Pump characteristic curve; b) Characteristic lines of flow resistance for two temperatures
O_1, O_2 = Operating points (O_1: $Q = 100\,m^3$/h, $\Delta P = 1.8$ bar; O_2: $Q = 150\,m^3$/h, $\Delta P = 1.1$ bar)

The characteristic curve is primarily determined by the design of the pump [39]: *Axial pumps* of the propeller type are characterized by high volumes and low pressures.

In *radial pumps* the liquor is accelerated from the center to the outer edge of the pump, high pressures can be produced and lower volumes are pumped.

Another important parameter is the ratio of diameter of the pump wheel to that of the pump housing.

Secondarily, the resulting characteristic curve also depends on the pump settings, the setting of the throttle and bypass valves, of the pump motor speed (rpm), and of the angle of the propeller blades.

Thus the performance of the pump, the most important part of circulating machines, can be largely controlled by the pump settings.

Equally characteristic of the particular dyeing system is the flow resistance of the textile goods. The absolute value R_{abs} of the flow resistance in mbar $m^{-3}\,h^{-1}$ is given by

$$R_{\text{abs}} = 1000 \times \Delta P/Q$$

(The differential pressure ΔP is sometimes incorrectly termed flow resistance.)

In normal dyeing conditions, R_{abs} is largely independent of the liquor flow rate.

In the differential pressure – flow diagram, the resistance can normally be represented by a straight line through the origin. This intersects the pump characteristic curve at the so-called operating point O.

The flow resistance often changes during a dyeing operation for reasons other than temperature, e.g., it increases on shrinkage and/or swelling of the textile or if the direction of circulation changes from i → o to o → i (inwards = i, outwards = o).

Increased flow resistance results in a decrease in flow rate and an increasing tendency towards unlevelness. Another useful measure of liquor circulation rate is the number of bath changes per minute (see Fig. 7). For normal levelness requirements, the reduction in dye concentration for each bath change should be only 2 – 3 %.

A comprehensive review of designs of circulating machines produced by the most important manufacturers (ITMA 91) is given in [40], mainly in the form of machine drawings.

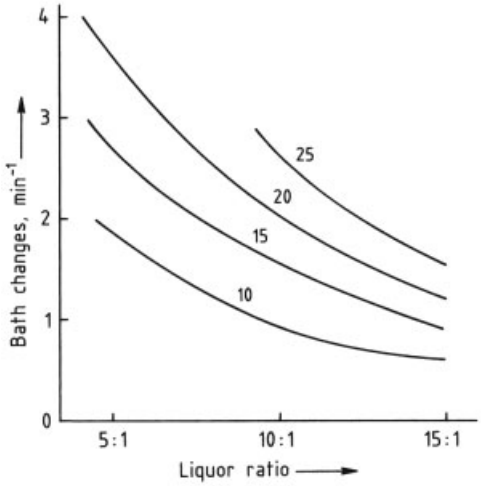

Figure 7. Frequency of bath changes as a function of liquor ratio at different specific flow rates (10, 15, 20, and 25 L kg^{-1} min^{-1})

2.2.3.2. Loose Stock Dyeing Machines

A loose stock dyeing machine is used for dyeing of flock, card sliver, and tow.

Dyeing of Flock. Flock consists of randomly oriented fiber which is packed (wet) into a cylindrical container with central tube [40].

Only the coarser and more stable types of fiber can be dyed as flock without damage. The packing density of flock is in the range of 200 – 400 g/L. It increases when water and auxiliaries are added. The liquor flow rate is in the region of 10 L kg^{-1} min^{-1}, and the liquor ratio is usually in the range 5:1 – 20:1.

It is difficult to pack and densify the loose stock uniformly, so that unlevelness can easily occur, especially because of the long flow paths in the packed material. However, subsequent spinning (followed by doubling) of the fibers often evens out unlevelness mechanically.

Loose stock dyeing apparatus is usually constructed to hold up to 1 t material, so that not too many dyeing batches are necessary for each batch of spun yarn. Loose stock dyeing comprises ca. 1 % of all bath dyeing.

Dyeing of Card Sliver. Card sliver consists of a long strip of any kind of staple fiber, the fibers being mainly parallel. Fiber dyeing is necessary in the production of melange (blended) yarns.

The slivers are placed in dyeing baskets or coiled into bundles (wool "tops" having a maximum weight of 25 kg) [40,41]. The slivers are sometimes wound onto a beam for dyeing (see Section 2.2.3.5).

Dyeing of Tow. Monofilament material produced during the manufacture of synthetic fibers (primary spinning) is formed into strips (tow), which is packed into large containers or carriers [40]. The filled containers are then placed in the dyeing apparatus.

2.2.3.3. Package Dyeing Machines (Cross-Wound Packages)

The schematic shown in Figure 4 also applies to package dyeing machines, which are usually operated at high temperature. In the (obsolete) overflow cooling system, the liquor that overflows due to expansion collects in the additives (or expansion) vessel. The advantage of this system is the ease of sampling and manual addition of products, and the disadvantage is a high energy consumption, air oxidation of the reducing agents in vat dyeing, and increased foam formation. In modern apparatus, the liquor expansion is taken up by means of a gas cushion.

Package dyeing often allows sampling of the yarn for color evaluation through a pressure lock during dyeing.

The yarns (fine to coarse) are often wound from the spinning bobbins onto dyeing bobbins (metal or plastic). For this, cross winding is preferable (weight of yarn per package: 0.5 – 2.5 kg). The packages are either wound in a conical shape (for easier unwinding of the yarn) and then stacked with separator plates, or are wound cylindrically, and then usually compressed (elastic tubes). Very high package densities (ca. 800 g/L) are obtained with polyester sewing yarns (compressed), and very low values (ca. 200 g/L) with polyacrylonitrile packages (risk of deformation because of thermoplasticity).

The cross-wound packages are held in place in the apparatus by placing them on spindles, and the columns of packages so formed are closed off

at the top by a spring-loaded plate. The spindles are metal rods (with a star section) or perforated tubes through which the dye liquor is pumped (i → o and o → i). The spindles are mounted on floor plates which provide both the support and the connection to the liquor supply.

Quantities loaded by this method into package dyeing machines can be large or small. Normal loads for production equipment are between 50 and 500 kg. For larger batches, two pieces of dyeing equipment are joined together (tandem operation).

Occasionally, package dyeing is used with parallel wound yarns (rocket packages, thermofixation packages) or with narrow ribbons instead of yarns (e.g., lace, zippers).

The liquor ratio can range from ca. 4:1 (short liquor ratio dyeing machine) to ca. 15:1 (open winding, older dyeing machine).

The short liquor dyeing machine is usually only partially flooded, and permits liquor circulation only from the inside to the outside (i → o, foaming tendency). The equipment must be specially designed to give a high enough pressure on the suction side of the circulation pump to prevent cavitation (Fig. 8).

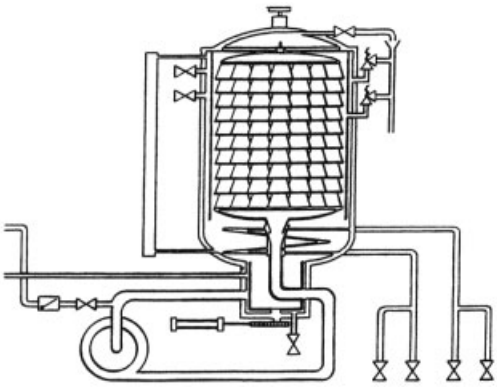

Figure 8. Short-liquor dyeing machine for cross-wound packages, operating i → o and o → i (eco bloc, courtesy of Thies)
Liquor ratio 3.5:1 – 6:1.

Higher dyeing rates can often be obtained by reversing the direction of circulation, usually every 3 – 8 min. The time required for this reversal process (dead time) is usually ca. 10 – 100 s.

The techniques used for producing the change in direction of circulation are:

1) Reversal of the direction of the pump (brief stoppage of the electric motor, expensive method)
2) Sliding valve system (pump always running in the same direction, see Fig. 9).
3) Adjustment of the pump blades (while motor is running)

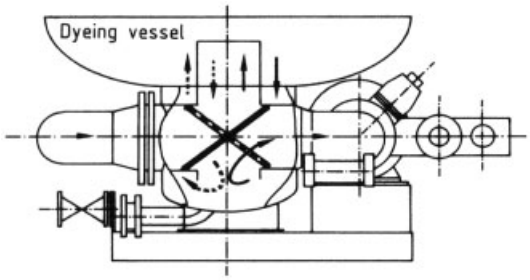

Figure 9. Special valve system for reversing the direction of flow in a dyeing machine with monodirectional circulation pump (courtesy of Thies)

Auxiliaries can be added to the additives container manually or automatically. The auxiliaries are pumped from here by the pressure pump into the dyebath.

Package dyeing equipment has a high share of the market. Most yarns for color weaving and knitting are dyed by this method.

2.2.3.4. Hank Dyeing Machines

The textile materials, which are usually delicate yarns (high bulk materials, texturized yarns, embroidering yarns) are formed into hanks, i.e., loose coils of yarn without carriers, ca. 0.5 – 1 m long. The hanks of yarn are hung fairly closely together over poles inside the apparatus, which is usually rectangular in shape (a box) [42].

The liquor flows mainly tangentially through the hanks at a high rate. This results in considerable tension in the yarn, sometimes leading to flattening of the yarn at the points of suspension (which usually cannot be changed). The liquor ratio is usually 10 : 1. A bath change occurs every 1 – 3 min. The differential pressure is only 0.1 – 0.2 bar, but the pump must deliver the liquor at a high rate (20 – 50 L kg^{-1} min^{-1}). Axial pumps (propellers) are often used.

Information on hank dyeing apparatus are given in [42], and information on ITMA 91 in [40].

Hank dyeing is now one of the older processes, and represents a small fraction of yarn dyeing. Circulating-goods machines for hanks of yarn are described in Section 2.2.6.2.

2.2.3.5. Beam Dyeing

In contrast to the apparatus described in the previous sections, the dyeing vessel in beam dyeing is normally arranged horizontally.

Piece Beams. In piece dyeing by the beam method, the woven textile (piece), which must not be affected by pressure on the surface (woven, knitted, or nonwoven material), is evenly wound onto a beam (perforated tube) in full width. The interior diameter of the beam is usually 0.3 – 0.5 m, and the thickness of material on the beam is usually 10 – 40 cm, depending on the type of fabric and size of the batch. Since the density of the roll of textile on the beam is similar to the density (g/L) measured on a single layer sample of textile, it can be predicted in this way.

Because dyeing with piece beams is so important, calculation methods have been developed based on simple measurements on a single layer sample to calculate the number of layers, the length in meters, the roll diameter, and the outside diameter. Measurement of the flow rate through a sample enables even the flow resistance of the roll of textile to be accurately predicted [34]. In practice, the calculation of the liquor – textile interaction is most advanced in the dyeing of piece goods by the beam method.

The direction of flow of the liquor is normally from the inside to the outside of the roll of textile. The ends of the beam are covered with metal sheets (sleeves) before the fabric is wound on, to prevent a short circuit of the liquor.

Although a high liquor flow rate is desirable, a limit is imposed by the resulting high differential pressure, which causes the fabric to "lift" from the beam, often at a differential pressure of only 1 bar. This disturbs the controlled flow of the liquor. A possible preventive measure is to wind the fabric under high tensile load, which, however, is problematical for certain fabrics. Batch sizes can be 500 – 3000 m. An open (permeable) fabric is usually more suitable than a dense (low permeability) fabric. Also, significant swelling of the fabric reduces its permeability to the liquor (greater pressure buildup).

The piece beam process has a market share of ca. 10 % of circulating machines.

Warp beam dyeing is similar to piece beam dyeing. Here, the warp or a part of it is wound onto the beam. The yarns are parallel to each other, and have a relatively high flow resistance as the threads support each other. If the differential pressure becomes too high, a roll of this type cannot lift, but the threads "part".

Warp beam dyeing is mainly important for the process of multicolored weaving with a plane-dyed warp, it accounts for < 10 % of all circulating machines.

2.2.4. Circulating-Goods Machines with Textile Storage (Winch Type)

2.2.4.1. System and Functions

Section 2.2.4 describes dyeing machines in which the textile is moved, i.e., it circulates through these types of dyeing machines, and is stored in folds in the dyebath to obtain additional time for dyeing. These dyeing machines have been developed from the winch beck (Section 2.2.4.2).

All the fundamental principles of exhaustion dyeing described in Section 2.1 apply to this dyeing technique, and especially bath dyeing as described in Section 2.2.1.

Dyeing machines with storage devices (rope dyeing machines) consist basically of a means of transporting the textile and a storage zone into which it is transported. Movement of the textile can be achieved

1) Mechanically by a rotating body (winch, i.e. a rotating body with an elliptical section) and electric motor
2) Hydraulically by a jet
3) By a hydraulic lifting pipe and the force of gravity in a propelling system (overflow)
4) Pneumatically (Airflow)

In the region of the drive unit, an intensive mass transfer occurs between the free liquor and the liquor in the interstitial volume of the textile, although only very briefly.

In the storage region, the dye has time to diffuse out of the interstitial spaces of the textile and into the interior of the fibers. The depleted liquor in the interstitial volume must then be replenished.

The number of cycles of the textile per minute is usually in the range 0.2 – 1, with velocities of ca. 30 – 600 m/min and batch sizes usually < 1000 m. The relationship between these process parameters is shown in Figure 10.

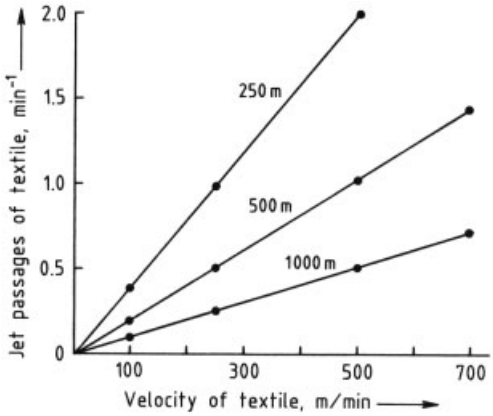

Figure 10. Jet passages per minute in dyeing machines with textile storage as a function of the velocity of textile (for batch sizes up to 1 t of goods; running speed ca. 70 – 700 m/min)

If the load in the dyeing machine is too high, the dwell times in the storage region become too long, and unlevelness and crease marks result. A rule of thumb for a level dyeing result is that only 1 – 3 % of the dye should be extracted for each cycle of the textile.

Liquor ratios range from ca. 3:1 (Airflow machine) to ca. 30:1 up to 40:1 (winch beck).

The greater the number of cycles per minute, the lower is the risk of unlevelness. At one cycle per minute, a dye exhaustion rate of ca. 2 %/min can be used, leading to exhaustion times of ca. 50 min with correct control. The liquor ratio is of greater significance for the running behavior of the fabric and the formation of crease marks than for uniformity between ends.

Loading of the dyeing machine with the fabric requires time. The fabric is loaded in rope form, usually at a speed below the operating speed of the machine.

A very comprehensive overview of dyeing machines can be found in the form of machine drawings in [40].

2.2.4.2. The Winch Beck

In the winch beck, a motor driven roller (usually elliptical, the so-called winch) is located above a dyebath. This winch draws the textile (woven, knitted, or carpet) via a guide roller out of the bath and returns it in folds into the bath. This removes excess liquor, and the liquor from the interstitial space of the textile is interchanged with "outer liquor" see Fig. 11. The velocity of the textile is 50 – 100 m/min.

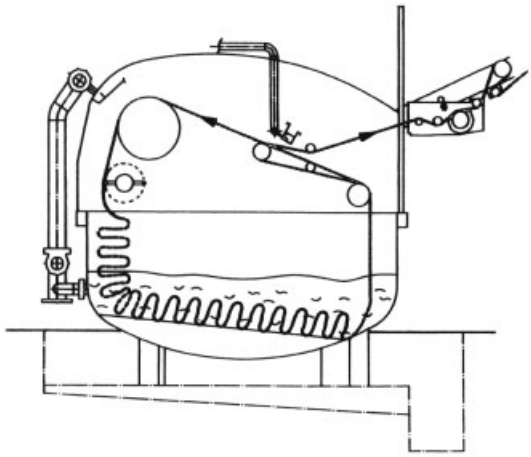

Figure 11. HT winch beck for dyeing of carpets (Supraflor, courtesy of Brückner)
Working width 4.2 or 5.2 m; capacity max. 1 t of goods; running speed of liquor circulating pump 70 – 100 m/min.

The fabric in a batch is divided into pieces of equal length whose ends are sewn together to form endless loops. Usually, there are up to 10 loops of fabric side by side on the winch, separated by rods or guiding plates. Carpets up to 5 m wide are dyed at full width or are sewn together lengthwise to form a tube (mainly in the United States). The liquor ratio is usually very high (25:1 – 50:1).

Winch dyeing is generally considered to be obsolete. However, it is a very gentle process, a very soft and voluminous fabric being produced. The fabric must not, however, be prone to form crease marks.

Mixing of the liquor in the bath is poor, so that it is difficult to maintain a constant liquor temperature in all parts of the bath and to eliminate differences in dye concentration. Pumps are often used to solve these problems by providing forced circulation of the liquor. Dyeing rates of $< 2\,\%$/min and long leveling times are often necessary to obtain level dyeing.

The high liquor ratio has a very unfavorable effect on the dye yield (see Section 2.1.3.4).

Some of the winch becks used today are high temperature machines [37]. The market share of winch becks for batchwise carpet piece dyeing is very high [44], but it is low for the dyeing of woven and knitted textiles.

2.2.4.3. Jet Dyeing Machines

In jet dyeing machines, a ring nozzle (jet) is located above the dyebath, the storage area for the fabric. The textile goods are transported through the nozzle, the required liquid flow to propel the fabric hydraulically is produced by a liquor pump. The diameter of the nozzle must match the "thickness" of the rope, i.e., it depends on the goods.

The fabric in the dyebath is moved towards the point where it is lifted up to the nozzle by sliding, a transportation device, or by floating. The liquor ratio with partly flooded jet machines is 10:1 – 20:1, and with fully flooded machines can be 30:1 – 40:1.

Jet dyeing machines are usually constructed for high temperature operation, enabling polyester textiles (knitted or woven) to be dyed satisfactorily. Because of many combinations of the various types of nozzle (including adjustable nozzles) and fabric storage devices, very different types of jet dyeing machines exist [45]. An example of an industrial design is shown in Figure 12.

The high turbulence in the nozzle as well as the entering of the fabric into the liquor lead to the entrainment of air in the dye liquor, with consequent danger of foam formation. The fabric storage system does not facilitate liquor interchange, because the fabric retains much "old" liquor (ca. 200 – 300 %), and the extent of liquor interchange in the storage zone is small. In partially flooded jet machines, a relatively large air space exists, which can cause problems especially with vat dyeing (loss of reducing agent by air oxidation). The fabric velocity is usually in the range of 200 – 400 m/min. The batch size is ca. 200 – 600 m.

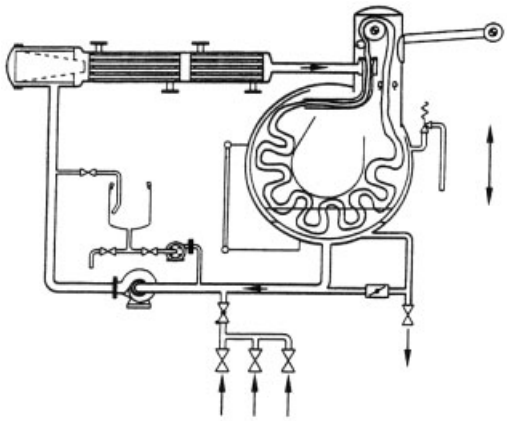

Figure 12. Schematic of a jet dyeing machine (Colorstar, courtesy of Scholl AG)
Ring nozzle (top left in dyeing vessel); fabric storage (bottom); liquor circulation pump (left); heat exchanger (top left)
Running speed ca. 60 – 500 m/min.

Although their space – time yield is not very high, jet dyeing machines have now become widely established. They represent the majority of all dyeing machines.

2.2.4.4. Overflow Dyeing Machines

In overflow dyeing machines the textile material (in rope form) is raised by the flow of liquor in a pipe system, and the excess liquor overflows. A winch (usually cylindrical and not motor driven) is situated above the fabric transport point, and the fabric "hangs" over this. A longer length of textile hangs from the exit side of the winch than from the inlet side. Gravitational forces pull the longer length of textile downwards more strongly than the shorter, and the textile is thereby transported very gently [24, p. 700, Fig. 5.]. The fabric enters into the storage zone or the pipe system into which the liquor is pumped. In the pipe system, the liquor continually flows past the fabric in cocurrent. Good renewal of the liquor in the storage zone is therefore ensured, which positively affects the liquor interchange with the fabric to some extent (see Fig. 13, next page).

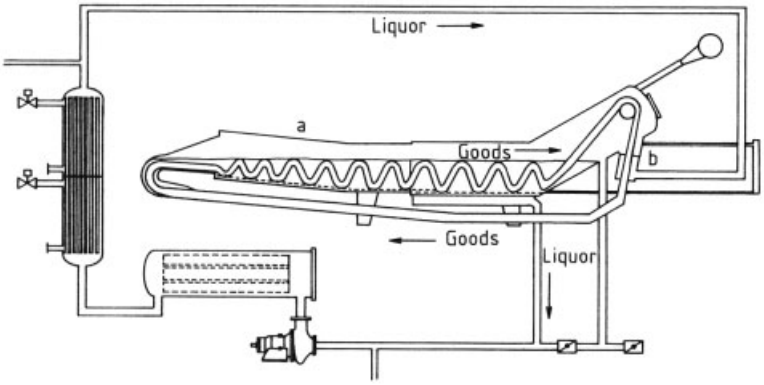

Figure 13. Schematic of an overflow dyeing machine (soft-TRD, courtesy of Thies)
a) Fabric storage area; b) Nozzle for fabric transport, with easy adjustable cross-section from ultrasoft to hard action on the fabric

The pump must provide the lifting energy to overcome the difference in height (0.1 – 0.2 bar). The flow resistance of the fabric to the liquor has almost no influence.

Overflow dyeing is suitable for woven and knitted fabrics, especially for delicate materials which justify the cost (tension-sensitive, soft, voluminous fabrics with a structured surface or high bulked yarns). The liquor ratio is ca. 15:1 – 25:1, the textile velocity is ca. 20 – 80 m/min, and the pump provides a bath circulation rate of ca. 1 – 2 circulations/min.

This special type of machine has a considerably lower market share than, e.g., jet dyeing machines.

2.2.4.5. The Air Jet ("Airflow") Dyeing Machine

The "Airflow" dyeing machine has been developed by the company Then. The fabric (in rope form) is transported pneumatically through a nozzle, by air. It then passes into a storage zone which contains a very small amount of free liquor. The dye, auxiliaries, and liberated liquor are atomized and applied to the moving fabric in the form of an aerosol (see Fig. 14) [46, 47]. A powerful fan and a suitable design of the ring nozzle enable fabric speeds of up to 900 m/min to be achieved without damaging the fabric. The air which propels the fabric is recirculated.

Because of the short liquor ratio applied (2:1 – 4:1) the dye must have adequate solubility. For a 2 % dyed product, dyeing with reactive dye requires 10 g/L dye and 40 g/L sodium chloride.

This dyeing technique is suitable for woven fabrics and stable knitted fabrics. Performance depends greatly on the particular type of textile, however, it is not possible to check in advance whether goods are suitable or not (leveling problems).

The "Airflow" machine was the first air jet dyeing machine to be established on the market. The "Airflow" now has a share of ca. 10 % of all jet dyeing machines.

2.2.5. The Dyeing Jigger

2.2.5.1. Normal (Direct) Jig Dyeing

The batch of fabric is drawn into the jigger, through a dyebath, and is taken up by a roll on the other side of the bath (Fig. 15) [24, p. 701, Fig. 6.], [47–49].

Generally woven textiles are dyed by the jig process. The textiles must not be sensitive to tension (e.g., texturized warp) or pressure (e.g., bulked material). However, fabrics susceptible to creasing can be dyed, as dyeing is exclusively performed in full width. The jigger is the most suitable dyeing machine for thin and impermeable woven textiles, e.g., lining materials. High temperature jiggers have been available since the 1980s, so that all fiber substrates, from cellulose to polyester garments, can be dyed in these machines.

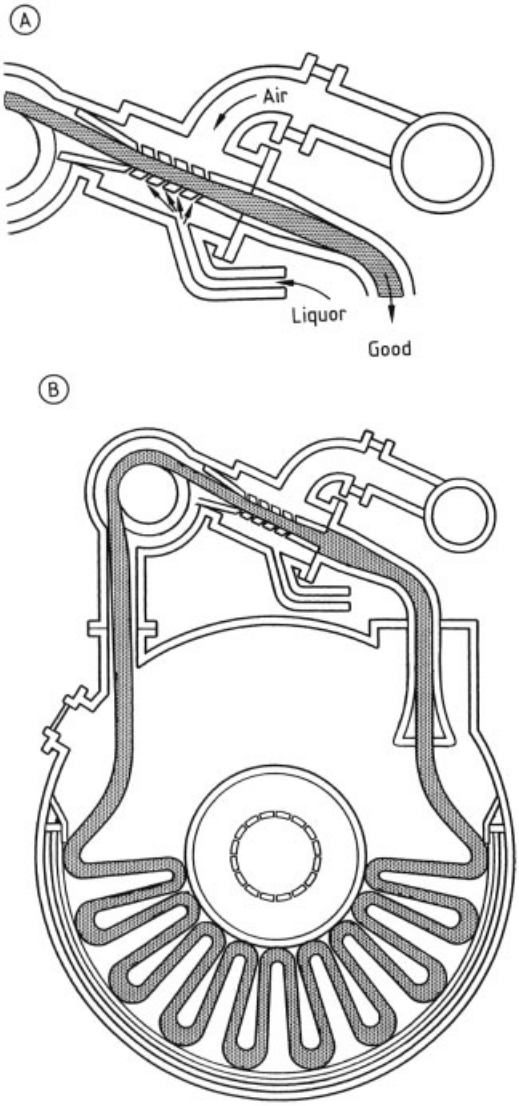

Figure 14. Air jet dyeing machine, HT type (Then-Airflow AFS, courtesy of Then)
A) Schematic of nozzle system: pressurized air strikes the fabric in an inclined angle and thus propells it; liquor is simultaneously entrained from the nozzles, sprayed and deposited on the goods; B) Schematic of the entire machine

The fabric can be loaded either dry or wet, e.g., after a continuous desizing operation. The fabric velocity is ca. 50 – 130 m/min. The liquor ratio is usually 3:1 – 5:1, i.e., very short. The size of the batch is usually 300 – 400 kg. So-called jumbo jiggers also exist in which a batch of almost 1 t textile can be dyed.

The fabric passes through the dyebath in which only a limited amount of liquor exchange occurs. A complete passage of a batch is known as a pass. During the time that the fabric is in the bath (ca. 0.5 – 1 s), the free liquor in the bath is exchanged with the depleted liquor in the textile. The exhaustion of the liquor begins in the wound up batch of fabric, where the fabric remains for a considerably longer time. This can lead to nonuniformity between the ends of the fabric.

During one pass, the dye content of the liquor becomes much depleted. The intensity of dyeing at the ends of the fabric is reduced because of short dwell times, and this must be leveled by an increased number of passes. To reduce the unlevelness of the ends, the dye is usually added during the first two passes (1:1). The dyeing process is controlled by the number of passes (exhaustion time, final concentration in the bath) and by means of temperature, pH, salt concentration, etc. (see Sections 2.1.3 and 2.2.1).

The passes are necessary to achieve intense interchange of the fabric with the liquor and to provide sufficient time for penetration of the dye. The overall dyeing time, determined by the number of passes, varies with the fabric velocity (usually constant) and the length of the batch (usually variable) (see Fig. 16, next page).

Minimum dyeing times (exhaustion times) and minimum numbers of passes required to give level dyeing are given in Figure 16. These figures are valid for a batch of 1000 m; for shorter batches (500 m), the number of passes must be doubled; and for batches 2000 m in length, only two passes may be necessary if the end unlevelness remains within limits.

A general problem with jiggers working at atmospheric pressure is to maintain the temperature constant. The covers should always be kept closed. However, even closed covers have an undesirable cooling effect on the system. The fabric in the batch certainly is at about liquor temperature (70 – 90 °C), but there is a tangential flow of air caused by the fabric until itreaches the roll, and consequent water evaporation then leads to cooling. In addition, the metal spindle at the center of the roll of fabric abstracts heat, so that the inner layers are cooler. The liquor is heated by direct steam, heating coils in the bath, or external heat exchangers, but, although the liquor is continually circulated, local temperature differences of several degrees Celsius often exist inside the bath. However, even if a

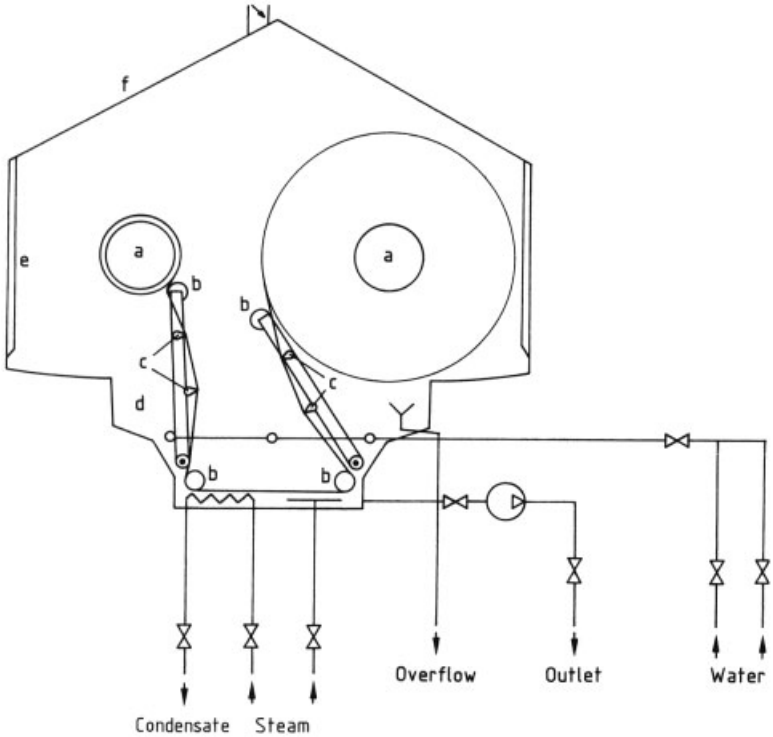

Figure 15. Schematic of a jigger (jumbo type; courtesy of Henriksen)
a) Main roller; b) Guiding devices; c) Expander (static); d) Basin; e) Window; f) Hood

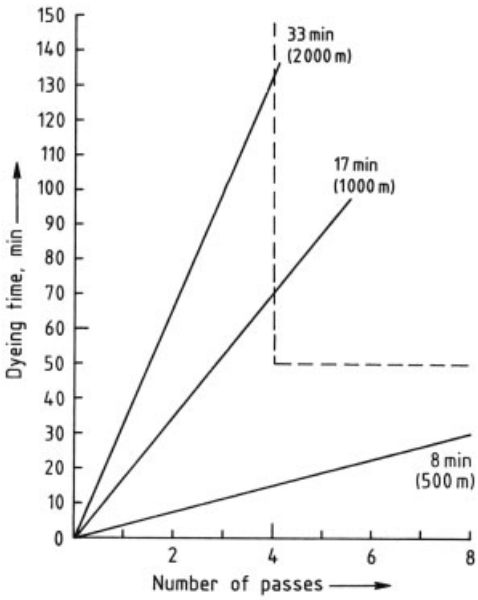

Figure 16. Dyeing time as a function of number of passes for different batch lengths and pass times
––– minimum line for good operation

pump is installed, the bath cycle times are often insufficient, > 5 min. An air flow on the sides of an open jigger often leads to edge unlevelness.

Jiggers represent a large proportion of dyeing machines used in the industry. In some sectors, they are even in the majority, e.g., for the correction of unlevelness in continuously dyed materials.

2.2.5.2. Pad Jig Process

The pad jig process is strictly speaking a semi-continuous method (Section 2.3). However, in accordance with accepted practice, it is included with the batchwise dyeing operations.

In the pad jig process, the dry goods are padded with dye liquor (see Section 2.3.1.1) and are batched up. This batch of fabric is then either hung in the jigger or drawn through in the usual way, corresponding to a single pass. The remainder of the pad liquor is usually added to the jigger (saving of dye, reduction of wastewa-

ter load, varying composition causes formulation problems).

The rest of the dyeing process is as outlined in Section 2.2.5.1. However, the minimum number of passes is considerably reduced, usually halved. This results in a considerable reduction in dyeing time for large batches (Fig. 16). The minimum number of passes is two using the pad jig process, and the dyeing time required is ca. 60 min for a 2000 m batch, whereas in the normal jig process 5 – 6 passes are needed with a dyeing time of 100 – 120 min.

Disadvantages include the problems of dye padding. The fabric to which the dye must be applied is usually sized gray cloth. Smooth fabrics with low interstitial volumes take up $< 50\,\%$ liquor, this small amount of liquor leads to problems of solubility of the dyes with dark color shades (>100 g/L). Other disadvantages are the operating costs of the dye application (smaller pieces, continuous process) and the complicated handling of the goods.

2.2.6. Special Bath Dyeing Equipment

These types of dyeing equipment can be circulating machines (e.g., star-shaped suspension equipment) or circulating-goods machines (e.g., vertical star equipment). They have only a small market share of all bath dyeing equipment. However, they play an important role in special areas (e.g., the dyeing of hosiery or velvet).

2.2.6.1. Star-Shaped Dyeing Frames

The carriers for the materials are two disk-shaped radial frameworks separated by a spacer tube at a distance equal apart to the width of the goods. On the radial metal rods, set at an angle of ca. 30°, there are small hooks (ca. 1 cm apart) on which the goods (woven or knitted) are hung. These are sometimes also pinned on one side only (usually for pretreatment, allowing free shrinkage). The star framework is an old type of dyeing equipment. The goods form a spiral cylinder and do not touch each other, i.e., dyeing is very gentle.

The *star-shaped hanging frame*, also known as a vertical star, is suspended in the bath. The goods are normally completely immersed.

The star is sometimes moved up and down (high liquor volumes and a liquor ratio of ca. 20:1 – 30:1) or it can be fixed in the dyebath and the liquor movement can be produced by a circulation pump (low liquor volumes and a liquor ratio of ca. 12:1 – 25:1). In neither case longitudinal tension occurs in the goods, and any transverse tension is usually due only to their weight.

In *horizontal star suspension systems*, the spacer tube can be rotated, and is located above the dyebath. The star is rotated by a motor. Only part of the star is immersed in the liquor along with the goods (short liquor ratio). The liquor is taken up when the goods are rotated. The spiral shape of the goods ensures that the inner parts are also fully supplied with liquor.

The star with the goods can be very easily transferred into other vessels, after dyeing in a stationary bath, e.g., into an aftertreatment bath or in a star steamer.

The space – time yields are low, so that this dyeing technique, which is now only seldom used, is expensive. However, the very gentle treatment of the goods can justify the process for textiles made from velvet and silk.

2.2.6.2. Machines for Dyeing Hanks of Yarn

Spray dyeing machines have been used for a long time for dyeing hanks of yarn, but tubular knitted goods can also be dyed by this method in some cases.

The hanks are placed on a perforated tube. The liquor is then pumped through this tube, sprays out of the perforations onto the goods and runs off. Normally, the hank does not dip into the liquor, so that there is some freedom of choice of liquor volume (liquor ratio can be between 10:1 and 40:1). A carrier arm moves the hank a short distance into a new position from time to time. This helps considerably to avoid flattening of the yarn [42].

The Colourdry dyeing machine by MCS is an interesting further development [50]. The hanks are placed loosely on triangles of tubing. The liquor is sprayed from the tubing onto the yarn (liquor ratio < 15:1). Centrifuging and drying are carried out in the same equipment without additional handling of the yarn.

The very gentle treatment is advantageous, but the space – time yield is low, so that these

machines are usually limited to special applications.

2.2.6.3. Paddle Dyeing Machine

This machine, also known as a paddle tank, is used for dyeing finished garments (sweaters, pants and trousers, etc.). The textile goods "float" in the dyebath. The goods are vigorously agitated by:

1) Motor driven mechanical paddles.
2) Injection nozzles (tangential and radial) through which the liquor is pumped. This produces the necessary relative movement of the textile and the liquor.

These dyeing machines are also constructed for high temperature operation. They have a large market share of the dyeing of ready-made goods, although this is a relatively small part of the total market [51].

2.2.6.4. Rotary Dyeing Machine

The rotary dyeing machine resembles a domestic washing machine. It can contain ca. 50 – 300 kg goods. It is mainly used for dyeing ready-made garments, and can also be constructed for high temperature operation [51, 52].

The drum dyeing machine is very suitable for producing wash-out effects. A high proportion of jeans are so treated. Various processes exist, e.g., stone washing (stones and chlorine-based bleach).

2.2.6.5. Cabinet Dyeing

The cabinet apparatus resembles a cabinet through which liquor is pumped (Fig. 17). If the apparatus is used for dyeing hanks of yarn, the cabinet contains poles that carry the hanks. Stockings and small articles of clothing can also be dyed, these being packed in perforated trays which are loaded into the cabinet. The liquor is circulated by a pump.

The cabinet dyeing machine has the advantage that it is easy to load with the textile goods. It is very important in the mentioned, highly specialized fields.

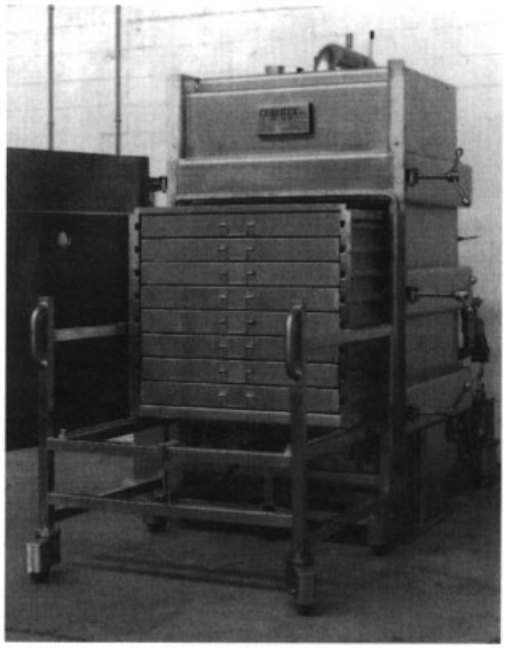

Figure 17. Cabinet dyeing machine for dyeing of hosiery (Amre P, courtesy of Cubotex)

2.2.6.6. Hosiery Dyeing Machines

In these machines, which are specially constructed for dyeing ladies' stockings, the stockings are placed over metal sheets formed into a leg shape, the sheets and stockings are briefly dipped into the vigorously agitated hot liquor, and both are then moved on automatically to the next stage. The dyeing takes place in a standing bath which is continuously replenished [51].

2.2.7. Automatic Control of Bath Dyeing

2.2.7.1. Aims

The term automatic control (process control) is used very generally here. It includes everything that can be contributed to the operation of a dyeing plant by regulation, control, and data evaluation not involving human intervention options according to the operation manual [53].

Automatic control today is expected to give good reliability and repeatability. Standard control systems should be compatible with both

types of dyeing machines in a dyehouse (use of specially adapted software).

The requirements on automatic control can be roughly divided into five stages (generations):

Simple process control
(partial automation)
Temperature – time program, alarm systems
Full automatic control
Additional requirements: filling and emptying of the liquor, liquor flow rate control, concentration control (pH, conductivity, etc.), automatic metering, data transmission, operations planning, etc.
Automatic charging and discharging
(robots)
Preparation of textile goods, charging into dyeing equipment, discharging and transportation after dyeing
Optimization and trouble shooting
Evaluation (including statistical methods) of data collected by the automatic control equipment and other measurements, special software and testing equipment [31]
Intelligent process development
Simulation of a bath dyeing operation on the computer screen including local data from equipment and the dye data (see Section 2.1.3.6)

The stage of today's technology is essentially that stage 2 is complete and stage 3 has started. On-line control in textile finishing is becoming to be part of the training program for textile engineers [54].

Automatic control has now become widely established in bath dyeing. It makes a considerable contribution to the reduction of manpower costs compared with the continuous dyeing process.

With the spread of automatic control the problem arose of precisely defining a dyeing process and representing it in an easily understood form. The tabular form in which formulations have been expressed for a long time is already well adapted for the purpose (see also Section 2.5).

The three usual forms of representation are [38,55]:

1) Description of the dyeing process in text form
2) Tables of the control functions required
3) Process diagram with symbols for the individual functions in the apparatus [38,56]

The instructions for a dyeing operation are usually rather generalized, so that the programmer has some degrees of freedom, e.g., with respect to liquor draining times during bath changes. Experienced dyers very familiar with automatic control and the dyeing process can often utilize the available degrees of freedom to achieve considerable process optimization. A dyeing diagram is shown in Figure 18.

2.2.7.2. Functions of Automatic Control

The automatic control of the charging of dyeing equipment with the textile goods (robot systems) is described in [47,57,58]. This can include automatic recognition of the batch number (stage 3 in list provided in Section 2.2.7.1) [47].

The basic classification is into (1) direct and (2) indirect effects of the control program on the process (local, decentralized) and (3) an intelligent control of the entire process, which is usually carried out centrally in a computer [32, 59]:

Direct control of the process elements by the stored program (stages 1 and 2, Section 2.2.7.1), occurs by control of valves and motors or pumps (e.g., flow direction i → o, o → i), alarm indicators for product additions, i.e., *process control*.
Indirect effect of the stored program on the process (stages 1 and 2). The set points are fed to the local control systems as a function of time. The control systems, which have local intelligence, carry out measurements (e.g., temperature, differential pressure, pH [60]), compare with the current set point, and issue the necessary control command to the actuating device, e.g., steam or metering valve, pump which is usually continuously controlled, use of feedforward and closed-loop control, i.e., *program control* (time – temperature program).
Intelligent control of the entire process (stage 4).

The data of the dyeing operation are usually fed into a central computer. The central computer produces operations planning, usage of the equipment (with or without shading), and batch control.

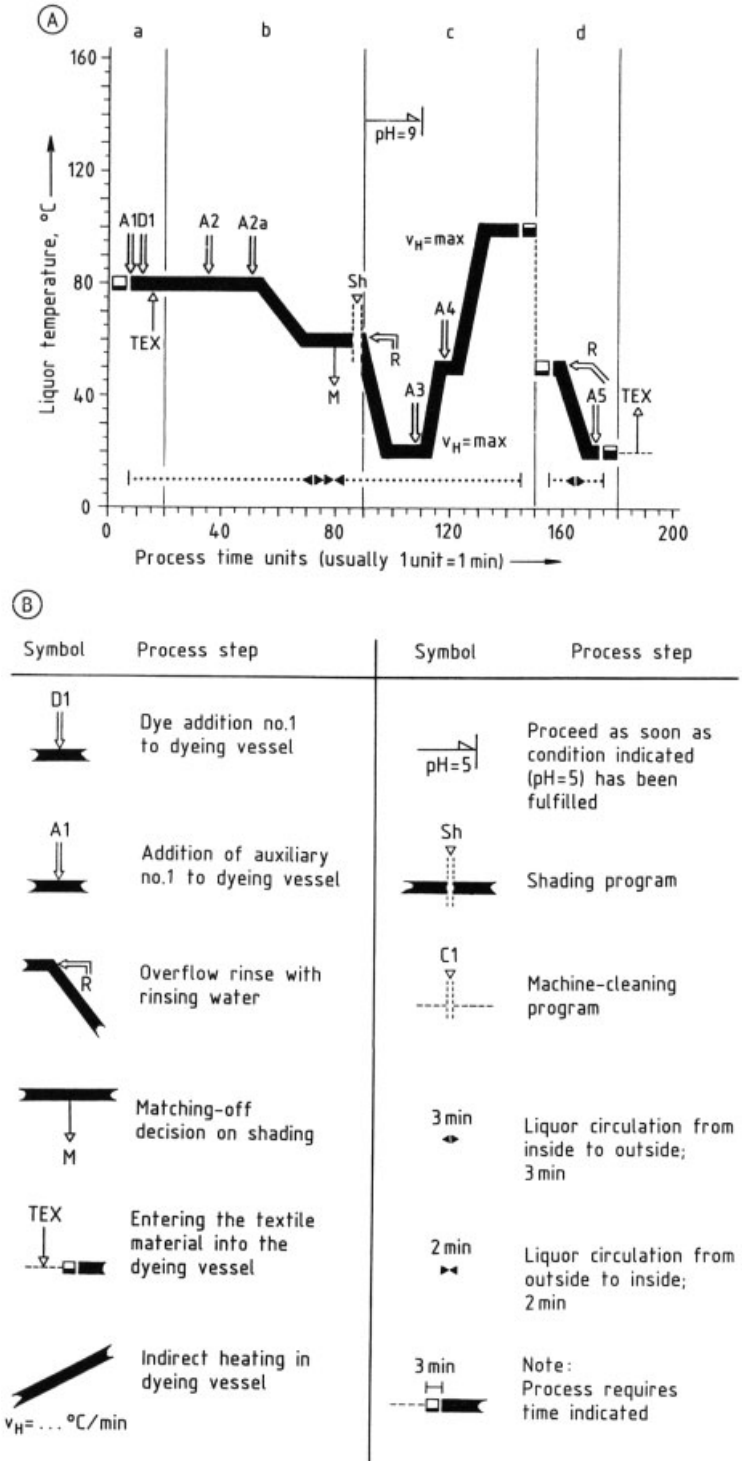

Figure 18. Process diagram for dyeing [38, 55], [77, p. 331]
A) Dyeing with vat dyes by the pre-pigmentation process; B) Table of process symbols
a) Preparing the dyeing liquor; b) Dyeing; c) Rinsing, oxidizing soaping; d) Rinsing neutralization

Data for cost calculations are sometimes also produced. The evaluation of color shading additions, dyeing rate, effects of auxiliaries, unlevelness, etc. is at an early stage of development. The preparation of dyeing programs by the dyehouse with the aid of software is described in, e.g., [26].

In all kinds of bath dyeing, i.e., in circulating – goods or circulating – liquor dyeing machines, or special equipment, the requirements are similar. Therefore the automatic control devices are essentially similar for the various tasks and are adapted to the specific tasks by additional modules. The control devices are made compatible with the software. Also, the same kind of personal supervision is applicable to all these control systems.

The requirements of an automatic control system cover a very wide spectrum today. However, a true process control system based on artificial intelligence is still awaited.

2.2.7.3. Equipment Requirements

Automatic process control usually has two components:

1) The control computer with the dyeing program and the operator's console (company-specific, Fig. 19)
2) A control system with programmable memory, a so-called stored program controller (SPC), constructed from universal low cost components (modules) which carries out individual functions

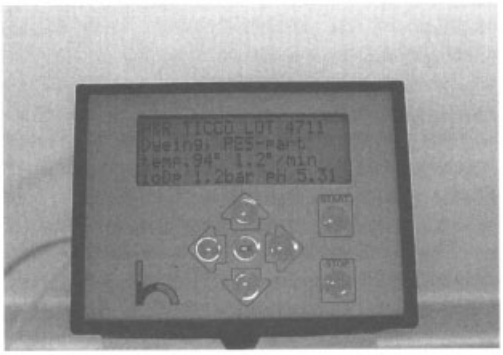

Figure 19. Local control equipment, connected to the control computer for data transfer and down loading of the dyeing programs [60, 61] (TICCO equipment, courtesy of Herrig & Rüttiger)

The individual functions are programmed into the SPC, and are called by the control computer as required. Likewise, the different closed-loop control systems are usually formed and backed up by the memory.

The programs stored in the memory of the SPC are completely adapted to the individual parts of the dyeing plant, and cannot be altered by the user.

The control computer contains the actual dyeing program. This is either entered at its console or on a PC, and is then loaded down to the control computer. It is usually specific to the dyeing equipment and not directly transferable. Only the process scheme is transferable from apparatus to apparatus in most cases.

Important values for closed-loop control
Liquor temperature (dyebath)
Liquor differential pressure (dyebath)
pH (dyebath, auxiliaries vessel)
Salt concentration

Important feedforward control functions
Filling (water, auxiliaries vessel)
Emptying (to drain or recycle)
Flow direction (i → o, o → i)
Metering (linear, nonlinear)
Time – temperature profiles
pH and redox profiles

Some devices are specifically necessary for the dyeing equipment, e.g., cycle detectors for jet and overflow dyeing machines; and passage counters to record the number of passes and devices to measure longitudinal tension for jiggers.

The control computer is often connected to a PC for transfer of process data.

In general, modern equipment for controlling dyeing machinery gives better reliability than good manual control.

Certainly the increase in production over recent years would not have been possible without automatic process control. Moreover, in view of the excellent durability of stainless steel dyeing equipment, the cost of installing or modernizing automatic control equipment can in many cases be justified by the benefits that follow (saving of manpower and gain in reliability).

2.3. Continuous and Semicontinuous Dyeing

The continuous dyeing process includes some very different dyeing techniques which can be characterized as follows:

1) The dyeing of fabrics (woven, nonwoven, knitted), which pass continuously through a line of processing units (exhaustion dyeing and pigment/binder dyeing)
2) Dyeing of flock, monofilament yarns, or warp yarns in special dyeing equipment (exhaustion dyeing)
3) Dyeing during the spinning of synthetic fibers (mass dyeing), i.e., during production

Mass dyeing is in principle carried out by the fiber producer, > 95 % of polypropylene fibers are dyed in this way. Mass dyeing is not described here, as this dyeing technique mainly lies outside the area of textile finishing.

The continuous dyeing of warp yarn is the main technique used for dyeing with indigo (→ Indigo and Indigo Colorants, Chap. 9.). Continuous dyeing of flock and monofilament yarns is described in Section 2.3.3.

The main topics of Section 2.3 are the continuous and semicontinuous dyeing of fabrics in dyeing lines, using the technique of exhaustion dyeing. Pigment dyeing, which represents a small market volume, is described in [62].

Special processing units were developed for continuous dyeing, most of them in the 19th century, and these are assembled to form the dyeing lines [15].

Some general rules and relationships apply to the fully continuous and semicontinuous dyeing technique:

1) Fabrics are treated in open width (not in rope form, because crease marks cannot be removed in a continuous system).
2) Fabrics are passed through a processing line at a constant speed of ca. 20 – 200 m/min. Possible stretching or shrinking of the fabric is compensated for by longitudinal tension sensors and speed controllers. The longitudinal tension on the fabric is in general adjustable. It usually causes a reduction in the width of the goods of several percent and leads to minimization of longitudinal creases [47].
3) The running speed of the fabric determines the (usually short) dwell time in the individual treatment units (ca. 5 – 120 s, depending on the length of fabric in the unit). In semicontinuous processes, i.e., continuous processes in the course of which the goods are batched up, dwell times of several hours are possible [63].
4) In all continuous processes, interruptions cause significant staining.
5) Any unevenness in the equipment across the width of the goods leads to unlevel dyeing (streaks and middle-to-edge unlevelness).

2.3.1. The Principal Stages of Continuous Dyeing

2.3.1.1. Dye Pickup

In this treatment stage, the dye liquor is taken up (continuously) by the fabric. Two groups of application techniques exist:

1) Direct application of the required quantity of dye liquor to the fabric, i.e., application of the liquor with a metering device. These techniques are of little industrial importance. Application techniques with direct pickup of the dye liquor include: pouring [64], kiss padding, spraying, foam application by doctor blade [18], and printing (1000-point roll).
2) Immersion of the fabric in the dye liquor. In this case an excess of dye liquor is taken up by the fabric. This is followed by removal of the excess to give the required charge of dye liquor on the fabric. This group of processes is of great industrial importance. Application techniques with immersion of the fabric and removal of the excess liquor include: squeezing off (padder) [66], squeegee (Flexnip), suction [65], and free runoff [BDA (British Dyer's Association) System].

The *padding* or *pad process* is by far the most important application technique (> 99 % of all continuous production) and is described below.

The *pouring process* is very important in the dyeing of carpets because it can handle articles up to 5 m in width. A small number of fabrics can be uniformly charged with dye liquor by a *squeegee operation*. Fashion effects can be

achieved by printing the fabric on both sides in a uniform color (→ Textile Printing). In all continuous techniques, uniform application of the dye liquor can only be achieved on a dry fabric. Therefore, in contrast to the batch dyeing process, the wet fabric must be dried before treatment which increases the costs of the process.

Padding Process. In the padding process the dry fabric passes through the pad trough, where it picks up the dye liquor. Good wetting of the textile is therefore important. With suitable auxiliaries it is often possible to pad even gray cloth (→ Textile Auxiliaries). After leaving the pad trough, the fabric is squeezed between rubber rolls. The liquor pickup, i.e., the resulting liquor charge, varies according to the fabric material (40 – 100 %).

The amount of dye liquor picked up (liquor charge or padding effect) strongly depends on the squeeze pressure produced by the two rolls (pad mangle). Figures for the liquor pickup are usually expressed as wt % based on dry fabric (at ambient atmospheric humidity). The curve of the liquor pickup plotted against squeeze pressure is nonlinear (Fig. 20).

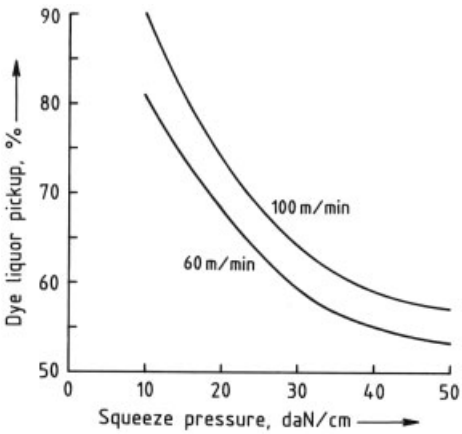

Figure 20. Effect of squeeze pressure of the pad rolls on dye liquor pickup at different fabric running speeds (60 and 100 m/min)
Values for a "swimming roll" system.

A low liquor pickup is often desirable to minimize both migration effects and the cost of an interim drying stage. A higher liquor pickup is preferred for vat dyeing of cotton (before steaming), as this gives a smoother fabric appearance.

In the pad dyeing process, the nip of the mangle ensures that the liquor flows into the fabric with high energy, i.e., the liquor is "worked" into the fabric which results in good dye penetration and very little two-sidedness.

Because the padding process is so effective, it has been improved as much as possible.

One of the problems to be solved is that the two rolls, when squeezed together by means of the end bearings, do not provide a uniform pressure over the full width of the fabric because bending takes place. A possible remedy is to fit a "bend tube" inside the body of the (tubular) roll to compensate for most of the nonuniformity (e.g., the Babcock system).

By spreading the squeezing pressure evenly over the width of the fabric, level dyeing across the full width is often ensured and middle-to-edge unlevelness and longitudinal streaking are prevented.

The most important techniques for providing a localized squeeze pressure include:

1) Swimming rolls ("S" pad rolls, Küsters see Fig. 21). Variation of the squeeze pressure is achieved by varying the oil pressure [15, 66].

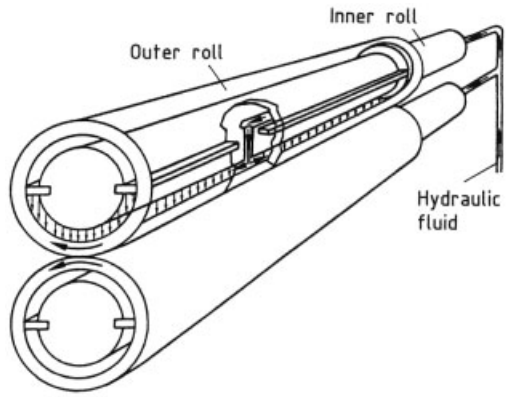

Figure 21. Swimming rolls (courtesy of Küsters)

2) The air cushion control method (the Bicoflex Foulard system of Ramisch – Kleinewefers) in which a tubular squeeze roll body slides on a tube which bears on a stator, by means of air cushions inflated to various extents. This enables the squeeze roll to be bent in various ways (see Fig. 22).
3) The wave profile of the pad rolls (Matex system, Monfort). Two squeeze rolls whose diameters vary over the width are arranged ad-

justably in opposition to one another. This enables any desired pressure distribution over the width to be produced [67].

4) Sectional rolls (Nipco rolls, Escher – Wyss). The squeeze pressure needed at particular points is obtained by varying the hydrostatic pressure of individual elements ([24, p. 702]).

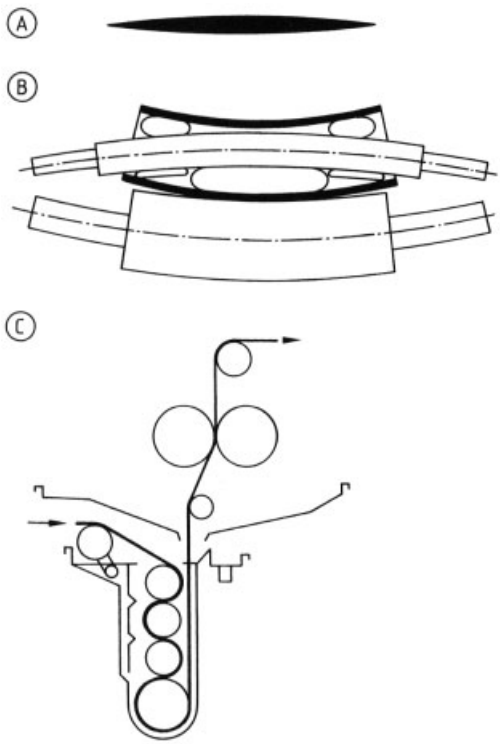

Figure 22. Pad rolls with dye liquor trough; the squeeze pressure adjustable across the width of the roll by inflated air cushions (supervision) (BICOFLEX system, courtesy of Ramisch – Kleinewefers)
A) Distribution of the squeeze pressure over the width; B) Inflated air cushions; C) Dye liquor trough (cross-section)

As a rough rule of thumb for rubber-coated rolls, increase in the softness of the rubber coating increases both the amount and the uniformity of the liquor pickup (see also Fig. 21).

Constancy of the Dye Liquor. Uniform liquor pickup is disturbed by the substantivity of the dye. Substantivity reduces color intensity at the end of the batch because of depletion of the dye content of the padding liquor. Possible countermeasures include:

1) Low immersion volumes (rapid attainment of equilibrium)
2) Low temperature, ca. 15 °C (often higher substantivity but lower absorption rate)
3) Control of liquor concentration (e.g., water addition at the start of the batch)

Problems are also caused by instability of the pad liquor which mainly results from:

1) Chemical conversions in the pad liquor, e.g., reaction of reactive dyes with alkali, inactivation of dyes by hydrolysate formation (see Sections 4.1), or reaction of vat dyes with reducing agents (see Sections 4.3, 4.4) [68].
2) Gradual precipitation of the dye if it is present at too high a concentration (higher than its solubility), mutual interaction of various dyes used for mixed fabrics (cotton – polyester dyeing), or transfer of sizing or harmful impurities from the gray cloth into the pad liquor.
3) Buildup of foam and lint (stippling due to deposition or resist effects, or printing effects due to small masses of lint).

Pickup of the dye is of great importance in continuous and semicontinuous processes, as defects that occur are extremely difficult to rectify or repair. Color intensity is determined not only by the amount of dye applied, but also by the dye penetration, which determines the "visual dye yield".

A theoretical treatment of pad dyeing is still a long way off. In particular, it cannot be predicted whether a fabric will be dyed lighter or darker at the edges even if squeezing-off is uniform. The dyeing practice is determined by the experience of the specialist, who decides on formulations, settings of the padding equipment, and on on-line measurements during production (see Section 2.3.4).

2.3.1.2. Intermediate Drying

After padding the fabric is often dried (pad dry process). In some cases this is absolutely unavoidable, e.g., when dyeing polyester of polyester mixtures by the thermosol process (see Section 7.1.2). In other cases intermediate drying can be a cost saving, especially with dark color shades. During drying the dye migrates to

the surface of the fabric, which results in a considerable increase in the "visual dye yield" (up to ca. 30 %, see Chap. 11).

In the vat dyeing (see Section 4.3.4), the application of liquor chemicals (reducing agents, alkali) on a predried fabric proceeds better, and, in particular, is less dependent on the material than a wet-in-wet process.

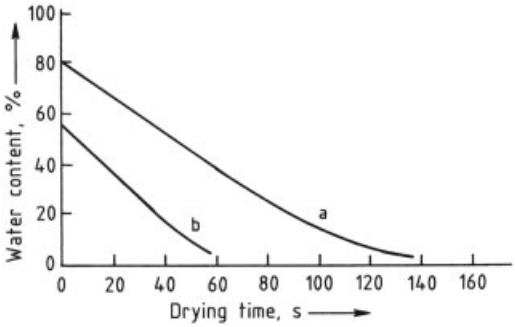

Figure 23. Drying curve of a hot flue dryer for hygroscopic fiber material
Decrease of water content with increasing drying time at 130 °C for a dryer length of 34 m. The initial linear section corresponds to Stage 1 drying.
a) Cotton, 100 %, 200 g/m^2, initial moisture content 80 %; b) PES – cotton 67:33, 156 g/m^2, initial moisture content 55 %

The drying process takes a very characteristic course (see Fig. 23) which is largely independent of the type of dryer [69], at least as long as the drying intensity is uniform along the length of the equipment. The drying process can be divided into the following stages:

Stage 1. Water evaporates from the surface of the fabric in a way largely independent of the nature of the fabric. Migration of the dye occurs, caused by flow of the dye liquor out of the interior of the fabric. The dye contained in the liquor is deposited in the surface zones of the fabric. This process ends when the liquor charge is ca. 30 %.

Stage 2. Water evaporates from the interstices in the yarns and fibers. Dye deposition occurs in the interior of the fabric where the evaporation occurs. During this stage, there is very little danger of unlevelness due to migration during drying. Dryeing conditions being equal, the drying rate is much lower than that in Stage 1.

Stage 3. This is only found with hygroscopic fiber materials [e.g., cotton, wool, poly(vinyl acetate)]. Water evaporates from the *interior* of fibers of this type. As some of the water is strongly bonded by hydrogen bonds, the energy requirement for evaporation is higher than in the other stages (further reduced drying rate).

Drying Equipment. Two different types of drying equipment are normally used for the intermediate drying of fabrics dyed by a padding process (see Fig. 24):

1) *Hot flue dryers* produce a hot air stream (jets, perforated drums), drying occurs by convection. Hot flue dryers have a higher absolute drying rate, but low drying rate per unit of surface area (low energy density).
2) The *infrared* (*IR*) *shaft* is a radiation dryer. It is usually only used as predryer before a hot flue dryer to give residual moisture levels of 30 – 25 %. Radiation dryers have a lower absolute drying rate, but higher drying rate per unit of surface area (high energy density).

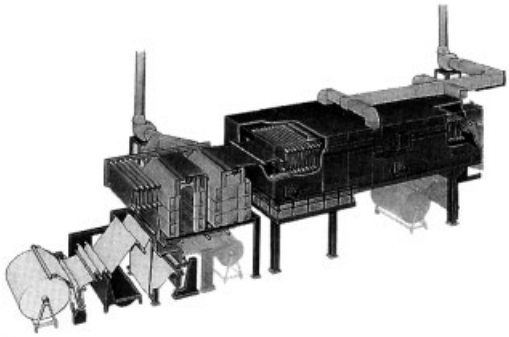

Figure 24. Intermediate drying
Fabric passage from left to right through: padder, air passage, IR shaft for predrying (VTG/3M, gas heated), and hot flue for final drying.

Predrying with a radiation dryer has become widely established, although uniform migration can be disturbed, leading to unlevelness (Stage 1). For pigment dyeing with a binder, a tentering frame is often used for drying before the fixation stage.

Drying Capacity. The drying capacity (evaporation rate) can be characterized in various ways, in which the rate can refer to either the dryer or the fabric.

The *drying capacity of a dryer* (absolute drying rate, DR_{abs} in kg/min is the amount of water evaporated per minute. For example in Stage 1 drying, $DR_{abs} = 2.3$ kg/min. An illustrative example is given below:

Fabric goods: 200 g/m^2, 1.5 m width (water content: 300 g per meter fabric); predrying: water content reduced from 65 % to 30 %.

Amount of water evaporated per meter length of fabric: 35 % of 300 g/m fabric = 105 g/m fabric. Drying capacity of the dryer: $DR = 2.3$ kg/min. Fabric running speed (Stage 1 drying): ($DR_{abs} = 2.3$ kg/min, evaporated water = 0.105 kg/m fabric) = 22 m/min.

The *relative drying rate* DR_{rel} based on the fabric indicates the intensity of drying of the fabric (energy density, fabric stress). The relative drying rate means the decrease in water content of the textile expressed as percentage units per minute. A reduction in the length of the dryer leads to an increase in the energy density for a given reduction in water content and a given fabric running speed.

Extending the above example gives:

Length of dryer:	4.2 m
Fabric running speed:	22 m/min
Drying time (dwell time):	4.2 m/22 m/min = 0.19 min = 11.4 s
Moisture reduction (65 % → 30 %):	35 %
Drying rate (relative, based on fabric):	35/0.19 = 184 %/min 35/11.4 = 3.1 %/s

Although these figures are useful when fabric changes are involved (light/heavy, sensitive/robust), they have not been widely used in practice. From the figures, it follows that:

$$\text{Fabric throughput} = \frac{200\text{g/m}^2 \times 1.5\text{m} \times 22\text{m/min}}{1000} = 6.6\text{kg/min}$$

Reduction in water content = 35 % of 6.6 kg/min = 2.3 kg/min. This is a gentle drying operation. For a soundly based assessment, figures for the "energy impact" on the fabric are also important (e.g., length of the drying zone, drying times).

Levelness. Both when drying with hot air and with radiation (steam production), differences in flow rate occur between the middle and the edges of the fabric. These can lead to differences in migration behavior (middle-to-edge unlevelness).

Energy Consumption. Dryers can be optimized with respect to energy consumption, in particular by prevention of excessive losses due to hot waste air (e.g., optimization to 150 g water vapor per m^3 air, or 7 m^3 air per kg water) [70].

2.3.1.3. Dye Fixation

The dye is fixed inside the fiber in the exhaustion dyeing process (see Section 2.1.3.5) or on the surface of the fiber in pigment binding. In the continuous exhaustion dyeing process, the penetration of the dye into the fiber is part of the fixation process. The sometimes complex processes of fixation, which are not sharply separated from each other, can be classified independently of the dyeing technique used as follows:

Chemical reactions:

1) Covalent bonding of the dye to the fiber material, e.g., reactive dyes (→ Reactive Dyes)
2) Change in the solvation of the dye molecules, e.g., oxidation of vat or sulfur dyes that have previously been reduced (see Sections 4.3.4, 4.5.3)

Incorporation into the fiber (desolvation):

1) Solvation change: dye liquor/fiber material, e.g., disperse dyes (see Chap. 7)
2) Ion pair formation and aggregation, e.g., acid dyes (see Sections 4.8.3, 5.1.2.1) or cationic dyes (see Section 4.8.4) and vat dyes at the soaping stage (see Section 4.3.5)

Incorporation into the fiber surface zone and binder film:

Surface fixation of a film of pigments, binders, and cross-linking agents, which sometimes undergo chemical reaction (see Section 4.7).

The *chemical reactions* proceed at a certain rate that determines the fixation time. In the absence of other information (see Chaps. 4 – 11), it can be assumed that the rate of reaction increases by a factor of 2 – 3 if the temperature is increased by ca. 10 °C. A reaction that requires ca. 10 h for 95 % conversion at 25 °C can take place in ca. 1 min at 85 – 95 °C.

Diffusion processes proceed slowly in solids (fibers). However, above the glass transition temperature of a synthetic fiber or with very swollen cellulose, diffusion can occur very rapidly (see Section 3.2).

In practice, the dye liquor formulations and range of products are optimized for the various requirements of each dye class, matching to the existing production technology. The most important fixation methods can be divided into two groups.

Semicontinuous Processes.

Cold Pad Batch Dyeing. The fabric is charged with dye liquor by padding wound onto rolls and these batches of cloth are stored for 4 – 10 h at room temperature. This is important in the case of reactive dyes (see Section 4.1), where the dye – cellulose reaction is brought about by the alkali in the liquor during storage. The reaction rate determines the liquor stability and fixation differences occur if the storage time is too short. The outer layer of the fabric roll is stored shorter time than the inner layer and does not reach the same degree of fixation.

Pad Roll Dyeing. The fabric, which is charged with dye liquor by padding, is fed into a chamber where it is wound onto a roll. This is most suitable for cellulose fabrics (see Chap. 4), and is sometimes also used for acetate fabrics with disperse dyes (see Chap. 8).

Continuous Processes. Here, the dyed fabric is available for further treatment after passing through the dyeing line (ca. 5 min).

Steam Treatment (Pad Steam Process). An additional developing agent is usually applied (by padding or kiss padding) to the fabric after it has been charged with dye and often predried. Examples of developing agents include reducing agents as used in vat (see Section 4.3.4) and sulfur dyeing (see Section 4.5.3) which produce a substantive form of the dye, or alkali and salt for the fixation reaction in reactive dyeing.

Saturated steam at ca. 100 °C for a dwell time of ca. 0.5 – 2 min is used to achieve fixation or at least a partial fixation reaction.

The most important machinery used for this is the continuous steamer (see Fig. 25). This is based on a fabric guide roll system, and has a fabric capacity of ca. 30 – 60 m. If atmospheric oxygen has to be excluded from the steamer during a dye reduction process, a steam blanket is provided at the inlet point and a water seal at the exit.

Other processes include the "Standfast" dyeing process in which the fabric is passed through molten metal (Wood's metal), heating the fabric in the liquor with the Williams Unit [24], and high frequency heating [71, 72].

The Thermosol Process (see Section 7.1.2). The dry fiber, loaded with dyestuff, is heated to around its glass transition temperature. The process is used with thermoplastic fibers, especially polyester, and sometimes also with polyacrylics. The fabric, whose fibers are covered with a dried film of disperse dye and auxiliary, is held at ca. 210 – 220 °C for 0.5 – 1 min. This causes the dye to diffuse into the interior of the fiber, although itis not clear whether this transfer takes place mainly by sublimation (gas phase) or whether the dye "dissolved" in the film of auxiliary exhausts from this film. Dye fixation of disperse dyes on polyester fibers as a func-

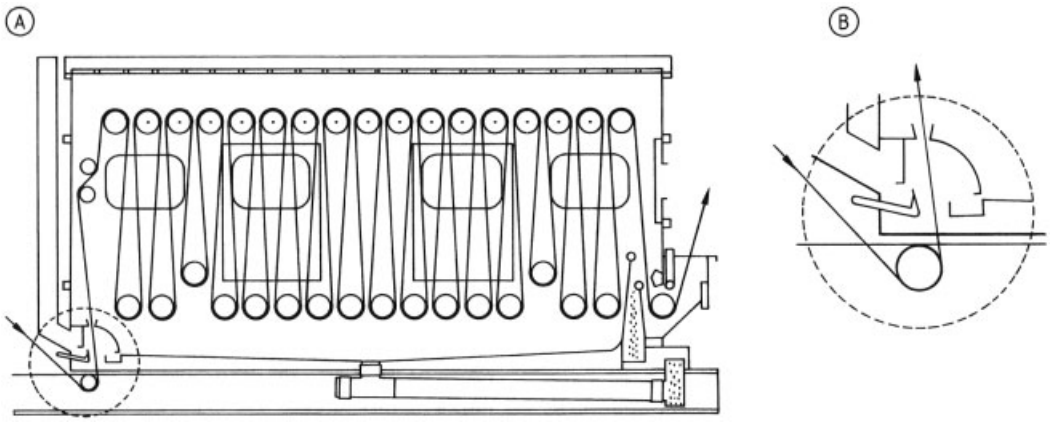

Figure 25. Continuous steamer; fabrics inlet and water seal specially designed for vat and sulfur dyes (courtesy of Babcock) A) Entire machine; B) Fabric inlet

tion of time and temperature is shown in Figure 26. The thermosol process is often used for dyeing the polyester component in polyester–cotton blends (see Section 11.1). Dark shades require subsequent washing to remove residual dye on the surface to improve fastness properties.

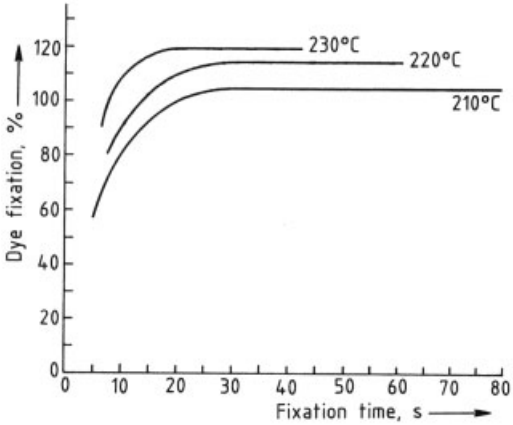

Figure 26. Fixation of disperse dyes on polyester fibers as a function of time and temperature (Rapid thermosol (RT) unit, courtesy of Fleissner)

Fixation of the Dye in Pigment Dyeing by Chemical Condensation of the Binder System. The combination of dye and (reactive) binder is first applied to the fabric by a pad dry process. The fixation stage is described in [17, 62, 73] and in Section 4.7.

Fixation by Oxidation. Dyes that have been chemically reduced before applying them to the fabric must afterwards be reoxidized to achieve fixation (see Sections 4.3.4, 4.5.3).

The oxidation is normally carried out in specially designed washing equipment (see Fig. 27). Excess alkali is first washed out of the fabric [74]. The most important oxidizing agent used today is hydrogen peroxide [61].

A typical example of the oxidizing conditions used in a vat dyeing operation is as follows: hydrogen peroxide (100 %) 1 g/L, pH 10–8.5, oxidation time 0.5–1 min, temperature 50–80 °C. If the liquor is too alkaline, this can lead to decomposition of the hydrogen peroxide [75, 76]. For deeper shades, the oxidation is sometimes completed in the subsequent soaping stage (95 °C), by entrained oxidant.

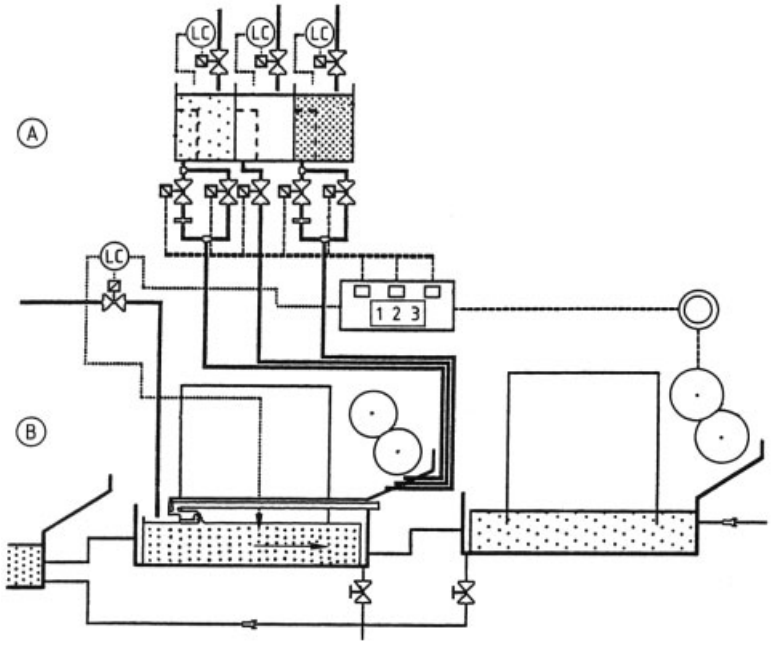

Figure 27. The oxidation stage in two sections of a washing plant with metering of hydrogen peroxide (sometimes also with acid or buffering agent) (courtesy of Babcock)
A) Metering section with buffering vessels for products and dispensing valves; B) Dyeing line with dosing controller working with running impulses and level control of the oxidation bath (movement of goods from left to right)

The dyeing produced is often dependent on the pH of the oxidation liquor.

2.3.1.4. Aftertreatment of the Dyed Fabric (Finishing)

Aftertreatment has the following main objectives:

1) To improve or complete the fixation of the dye
2) To "develop" the shade, i.e., to adjust it to the original pattern card
3) To improve fastness properties
4) To wash out residual chemicals or auxiliaries
5) To improve hand (feel)

Some process steps that have already been treated in Section 2.3.1.3, e.g., the fixation of most vat or sulfur dyes by oxidation, are often classified as aftertreatment or finishing processes.

Wet-in-wet methods are always used for aftertreatment. Aftertreatment is performed in the continuous open-width washing machine [74] (Fig. 28) or some of its units.

Application of Dye Fixation Agents. During the final wash cationic auxiliaries (to fix substantive dyes) and reactive hydrolysates (to improve wetfastness properties) are applied. The auxiliaries are metered into a section of the washing equipment or a special trough. Good squeezing-off is required before treating the fabric and aftertreatment must be carried out in a standing bath to prevent products from entering the countercurrent washing stream.

Soaping of the fabric is carried out in the soaping liquor close to boiling temperature. In the case of reactive dyes the dyeings are rinsed with pure water to remove hydrolysate residues. During vat dyeing soaping occurs in a quasi-standing bath to complete fixation and to develop the color. (It is claimed that soaping also improves rubbing resistance but this is questionable.)

The addition of auxiliaries, methods of improving colorfastness, and the problems that arise are described in Chapters 4 – 11. For control of dyeings and dyeing results see Chapter 13.

Reductive Aftertreatment. Especially in the case of dark shades produced by disperse dyes on polyester-containing fabrics (see Chap. 7), the dye that is adsorbed only on the surface of the fibers must be removed. For this, reducing agents, dispersants, and alkali are added in one or two sections of the washing equipment (standing bath method).

Neutral Washing. In this last treatment step product residues (especially alkali) are removed from the fabric [74]. During washing the liquor is often neutralized with acid, with pH control.

2.3.2. Dyeing Plants

In dyeing plants the individual stages of a dyeing operation are assembled of different machine parts (units) to form a station or production line. Generally the dyeing process can be carried out consistently and at a high production rate using the classes of dye for which the plant was designed, whereas improvisation may be necessary for other classes of dye. In some cases, tasks involving certain fabrics or classes of dye may be impossible.

A continuous dyeing plant must provide for continuous fabric supply including:

1) Synchronous throughput (start-up, production)

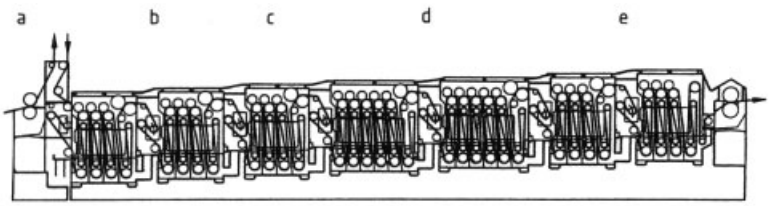

Figure 28. Continuous open-width washing machine for the aftertreatment of dyed fabrics (equipment especially for vat aftertreatment; courtesy of Benninger)
a) Padder; b) Rinsing; c) Oxidation; d) Soaping, near boil; e) Final rinsing

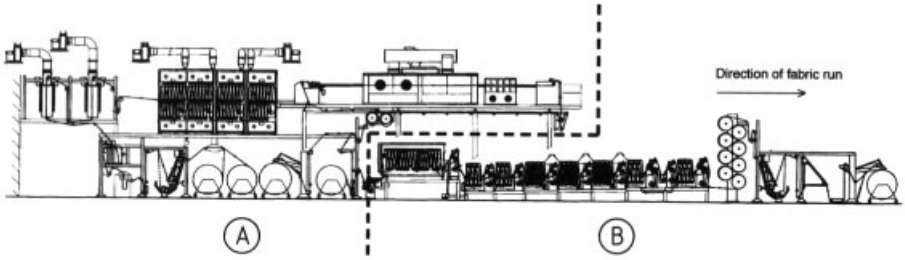

Figure 29. Thermosol pad steam line (courtesy of Babcock)
A) Disperse dye (thermosol) section: pad application of dye, dryer (IR, hot flue), thermosol unit; B) Vat dyeing section: pad application of chemicals, steamer, rinsing, oxidation, soaping and final washing (neutralization stage) sections

2) Adequate longitudinal tension (stretching of the fabric)
3) Maintenance of open width of the fabric (special mechanical devices)
4) Avoidance of longitudinal crease formation
5) Continuous product addition (metering).

Three typical plants are described as an example:

Thermosol Pad Steam Plant for Polyester – Cotton Blends (Fig. 29). In the following example a mixture of dispersion and vat dyes is first padded on the fabric. The fabric is then dried by IR shaft and hot flue dryer and subsequently treated in the thermosol unit to fix the disperse dye. This is performed in hot flue or perforated drum equipment. The reducing agent and alkali are then padded on the fabric and steamed. This causes reduction of the vat dye to the leuco form which is absorbed by the cotton, and also reduction of the unfixed disperse dye as a part of the aftertreatment of the polyester component (reductive cleaning).

Aftertreatment of the dyed cotton component is then carried out by:

1) Cooling (to ca. 30 °C, water seal on the steamer)
2) Cold rinsing (for alkali removal, in washing equipment)
3) Oxidation (at ca. 65 °C in washing equipment with metering of additives)
4) Soaping (at ca. 95 °C in special washing equipment)
5) Neutral washing (ca. 60 °C, in 2 – 3 washing sections)
6) Cooling and final squeezing-off (in squeeze rolls)

Wet-Steam Vat Dyeing of Cellulose Fabrics (Fig. 30). First the vat dye is padded on the dry fabric. The liquor that contains the reducing agent is then applied to the fabric while it is still wet. The usual intermediate drying is omitted. The dye is then fixed by steaming and oxidation (see Section 4.3.4) [77, pp. 319, 357]. The term "wet-steam" here indicates the wet-in-wet dyeing technique and not the condition of the steam.

In this technique, the fabric which is fed into the steamer is charged with a large amount of

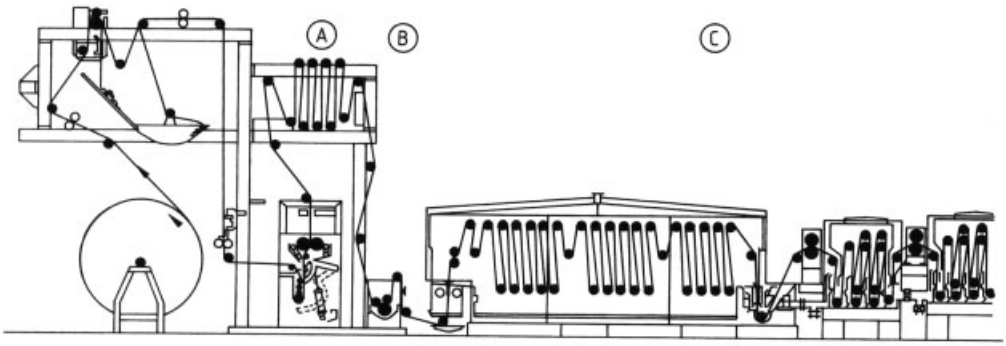

Figure 30. A wet-steam plant for dyeing with vat dyes (courtesy of Ramisch – Kleinewefers)
A) Padding of liquor (Bicoflex foulard) and air passage; B) Kiss padding of chemicals; C) Steaming and aftertreatment

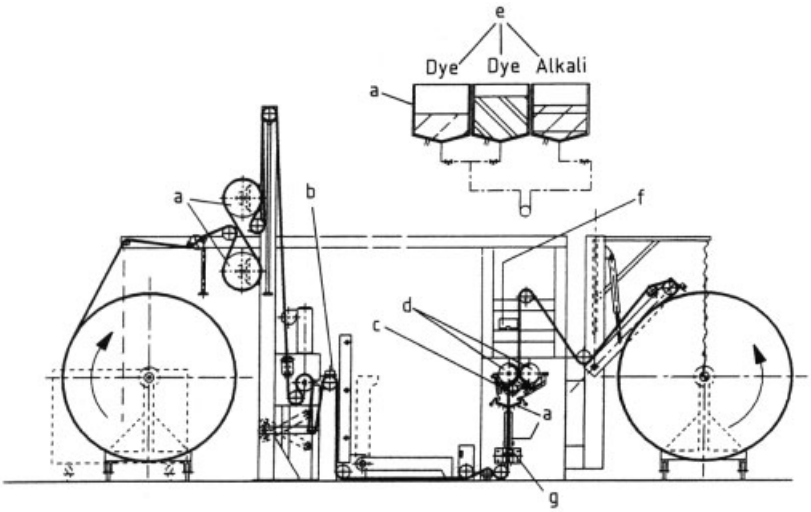

Figure 31. Cold pad-batch dyeing station (courtesy of Küsters)
Dyestuff uptake is controlled by, e.g., liquor photometric analysis, pickup, initial humidity, and fabric temperature.
a) Cooling, liquor or fabric; b) Measurement of initial moisture content; c) High-speed cleaning; d) Swimming rolls; e) Dye kitchen; f) Measurement of liquor pickup; g) Color measurement, liquor

liquor, so that there is a danger of streaking if the liquor runs back. This depends very much on the fabric. Fabrics with a high capacity for liquor often give no problems.

This dyeing technique only requires a small amount of machinery and little drying energy. The individual process steps are:

1) Dye application (absorbent fabric, padding to give ca. 60 – 80 % pickup)
2) Application of chemicals, i.e., reducing agent and alkali (padding, kiss padding, wet-in-wet, ca. 100 % total liquor pickup)
3) Steaming (vat dye steamer, saturated steam, air-free, duration ca. 1 – 2 min)
4) Cooling (water seal), rinsing, oxidation, and aftertreatment as in the thermosol pad steam plant

Cold Pad-Batch Process for Reactive Dyes on Cellulose Fabrics (see Fig. 31). In this typical semicontinuous process, the short duration steam treatment is replaced by long storage at room temperature.

The dye and alkali (reaction accelerator) are first padded on the fabric [78, 79]. The fabric is wound onto a roll and stored for ca. 5 – 10 h in the cold where reactive dye fixation occurs. The fabric is then washed and finished (see Section 4.1).

The individual process steps are:

1) Preparation of the dye liquor: addition of a weighed amount of dye followed by addition of the alkali shortly before use to prevent undesired hydrolysis of the dye in the trough
2) Application of the dye (cold padding at ca. 15 °C to give ca. 55 – 80 % liquor pickup)
3) Winding (cold, onto a roll)
4) Storage (cold, rotating batch, ca. 5 – 10 h)
5) Washing cold (ca. 15 °C) and hot (60 – 90 °C), washing machine or washing beam, preliminary rinsing and soaping
6) Aftertreatment (cold, wet-in-wet, with addition of auxiliaries)

2.3.3. Continuous Dyeing of Yarn and Fiber

By far the most important continuous dyeing process for yarn is that for dyeing with indigo [dyeing of warp yarns (slasher, rope lines) for denim material, → Indigo and Indigo Colorants, Chap. 9.]. Dyeing with other dyes (vat and reactive dyes) can be also performed on these plants, the dyes are used alone or as additions to indigo. Dyeing with dyes other than indigo is, however, rather unimportant.

The continuous dyeing of flock is occasionally carried out, e.g., in a "steamer tube" (Seracant equipment for polyacrylic fiber), or by

a padding process in high frequency fixation equipment (at ca. 80 – 95 °C) [72].

To obtain space dyeing effects (yarns with local variations in the amount of dye), the yarn can be knitted to tubes and locally printed (vigoureux process, → Textile Printing). After fixation of the dye, the knitted product is unravelled (knit-de-knit process). Yarn is also continuously dyed with local variation of the amount applied, and is then fixed in special yarn storage systems [80].

2.3.4. Automatic Operation of Continuous Dyeing Plants

The state of automation in the fabric finishing industry has overall reached a high level.

The automation of continuous dyeing plants differs from the automation of bath dyeing (see Section 2.2.7), mainly in that constant conditions are involved – requiring constant adjustments – and that no time programs are required.

Measurement and control equipment exist to a much greater extent in other industries, and the equipment used in fabric finishing represents only a part of the wide range available. There is no uniformity in the equipment used.

Computer-integrated manufacturing (CIM) is not (yet) widely used, presumably because the product information available to the dyeing industry is still too generalized and expensive [81].

The complex subject of the automation of continuous processes is here split up into two different aspects to facilitate the discussion:

1) Automation of individual items of equipment, mainly to improve their effect on the fabric
2) Automation of assemblies of equipment, mainly to improve their performance and their potential for working together as a large unit

2.3.4.1. Important Process Stages and their Automation

Selected process stages – dye application, preparation of baths, drying after padding – are described below in tabular form.

Characteristic process parameter (for both the fabric and the process), together with the *measurement and control elements* needed for determining and setting these parameters and for keeping them constant are as follows:

Dye Application (Padding of the Dye Liquor)

Process characteristic parameter	Measurement and control
Liquor pickup	control via squeeze pressure (continuously regulated)
Middle-to-edge unlevelness	control via variation of squeeze pressure over the roll width
Dye concentration in the trough	photometric measurements (transmission) of dye solutions and dispersions
Indicators of the amount of dye applied to the fabric:	
Dye concentration in the surface zones of the fabric	on-line reflection measurements on the fabric in the visible range
Amount of dye liquor on fabric	reflection or absorption of microwaves (cm range) [67] or IR [82]
Total amount of fabric fibers and liquor on the fabric	absorption of β-rays [67]
Volume of dye liquor required per running meter	measurement of liquor added to the pad trough

A complete list of characteristics of the padding process would contain still more items, e.g., signal generators and receivers to ensure uniform fabric running speed [47] and to determine residual moisture on the fabric after predrying, and the absorbency of the fabric. Control of dyeing pickup is only seldom achieved only by means of instrumentation.

Finally, the amount of dye liquor charged is of less interest than the dyeing produced on the fabric. This can easily deviate by $\pm 10\,\%$ from the figure expected from the amount of dye liquor charged. Here, the distribution of the dye in the fabric material (dye penetration and migration to the fabric surface) is presumably the important factor. Considerable research effort is needed to elucidate this effect and the often puzzling middle-to-edge unlevelness before a truly reliable automation process can be developed [83].

Control of the Treatment Bath (Continuous Washing Units). Linked to the dyeing processes, sections of open-width washing machines are widely used in important continuous process operations [47].

Process characteristic parameter	Measurement and control
Bath temperature	temperature control (heating, cooling, water mixing)
Bath renewal	control of feed rate (fresh water, countercurrent) or of amount of liquor on the fabric
Bath volume	level control (overflow flaps)
pH of the liquor	pH controllers, addition of acid, alkali, and buffering agents
Concentration of reducing or oxidizing agents	titration with redox probes as indicators or direct measurement with multiprobes to control addition rates [61,76]
Salt content	conductivity measurement (electrodeless or 2, 3, and 4 electrode systems) and control of dilution or salt addition
Concentration profile in the bath	control via partition walls, circulation pumps, or location of feed points

In practice, treatment baths have not yet reached a high level of automation. One reason for this may be the buffering capacity of the large bath volumes employed (ca. $0.5-2\,m^3$), which nearly levels out concentration changes and automatic control is therefore less important. This is even more significant because dyeing of small fabric batches (pieces) is becoming increasingly important. Simple and reliable metering devices are available which automatically adjust to the fabric running speed (e.g., Polycomat, Babcock).

The addition of acid or alkali is mainly carried out under automatic pH control. However, damage to electrodes (e.g., poisoning) and the presence of nonlinear transition functions in control characteristics still present problems.

Equipment for direct measurement of reducing and oxidizing agents is in the development stage. The most reliable method is still titration.

The main problem for a continuous plant is that a constant dwell time is necessary for a consistent dyeing operation, whether the fabric is light or heavy. Therefore, all weights of fabric pass through the plant at the same speed and differences in their dyeing characteristic must be compensated by other process parameters.

Drying after Padding (Intermediate Drying)

Process characteristic parameter	Measurement and control
Intermediate and residual moisture content of the fabric	measurement: electrical resistance of the fabric, microwaves (cm range), electrical capacitance, cooling of the drying air, fabric surface temperature
Fabric running speed, drying intensity, drying stress on the fabric	measurement: speedometer, values derived from two sources
Drying temperature (temperature profile)	measurement: air temperature, fabric surface temperature (conditions suitable for the use of antimigration agents)
Energy consumption	water content of exhaust air, e.g., "atmospheric humidity pipes" (Mahlo)

The following refers to the two-stage intermediate drying process (see Section 2.3.1.2).

It is important that all measurements should be made without contacting the fabric (to avoid streaking effects).

The *predrying stage* (IR radiation dryer) usually leads to a moisture content of 25 – 35 %. The moisture content is usually measured indirectly from the surface temperature of the fabric (thermal radiation measuring equipment). In this humidity range, nonuniform migration can occur (middle-to-edge unlevelness, streaks, two-sidedness). The drying is controlled by varying the radiation intensity of the dryer. The fabric running speed must not be changed because this would lead to marking.

Second drying by the hot flue process is considerably less intense, and therefore more uniform, with very little danger of migration. The fabric has a longer dwell time in the dryer. The residual moisture content of the dried fabric is usually 0.5 – 6 %, and this is often determined from the electrical conductivity of the fabric. The *drying controller* can adjust the temperature of the drying air (heavy fabric requires higher temperatures than light fabric), which takes some time. Alternatively, the heat transfer rate can be varied by varying the rotation rate of the fans, which produces a rapid reaction. Other devices automatically clean the idler rolls that sometimes become smeared with dye liquor (thickened) after the predrying stage.

If the drying is included in a dyeing line, the dye fixation (fixation time) determines the fabric running speed, and the drying equipment must run at a needlessly slow speed even with light fabrics.

However, if the dried fabric is batched up, the dryer can be used efficiently, as the light fabric can be run rapidly, and the heavy fabric more slowly.

2.3.4.2. Technology of Automation (see → Process Control Engineering)

General Process Control. It is first necessary to have local measurement and control equipment at or inside each machine. The measuring instruments (sensors) provide the information (measured data and signals), and the control equipment makes itpossible to influence the operation by means of control commands (actuators).

Local Measurement and Control Equipment. The measuring instruments (temperature, pressure, fabric running speed, longitudinal tension) are located very close to the point of measurement, and are designed with a direct drive of the display equipment (pointer and scale).

The closed-loop controllers are also constructed on this principle. These have a (mechanical) means of storing the set points. Their control commands mostly go to valves, which are often pneumatically actuated. The necessary controlling action (types of water, countercurrent flow, outflow, etc.) are usually carried out by lever systems or pneumatically.

The advantage of local measurement and control equipment is the low costs, but there are disadvantages:

1) Inadequate supervision and fact that only some of the functions take place automatically lead to faulty adjustment of machines.
2) The investigation and repair of faults can be difficult. Equipment is very individual, and know-how and replacement parts are difficult to obtain. This is also true for calibration.

Measurement and Control Equipment at a Central Operator's Desk. The local sensors and associated control instruments are connected to a central operation point. Amplifiers are provided to form a standard signal from the sensor output, and for transmission to the central display equipment.

A disadvantage is that determination of the weight per running meter of fabric is not carried out because this is still very expensive. Thus, quantities associated with the weight per running meter are not available, e.g., the water consumption expressed as L/kg fabric (from the water flow rate), and the moisture content of the fabric in wt % (from the amount of water per unit of area in grams per square meter) or the steam consumption per kilogram fabric.

Some control equipment (actuators) is set at the start of the lot, e.g., liquor flow control or flap valves. Also, the regulators which act on the control equipment are given their set points.

The centralized arrangement is considerably more expensive than the localized equipment, but has the advantage of security of operation.

Important Advantages:

1) Control is mainly from the operator's desk, and there is little hand operation of the machines. Change-over to operation with different dyes and fabrics is partly automatic, and uses control programs that can be called up.
2) In trouble shooting, only standard electrical and pneumatic signals are needed for control.

Process Computer as Control System. Sensors, regulators, and control equipment produce signals of a type (digital) that can be read by a process computer. Conversely, the computer sends signals (control commands, parameters) that can be understood by the equipment. Also, a PC can undertake a high proportion of the process computing tasks, if suitable software is available (on-line operation) [84].

Developments are now in the direction of local control equipment that provides the necessary digital signals or converts the analogue sensor and actuator signals to digital form (field bus operation). At present, many types of field bus are used, leading to unclear and nonstandard operating systems.

Continuous dyeing production lines with process computers sometimes afford the possibility of changing over to a kind of hand operation for controlling the plant and trouble shooting. A complete set of control instruments (usually centralized) exists for hand control of the plant in spite of the costs.

Really useful software tools for trouble shooting are usually in short supply (high programming cost, small number of possible machine combinations).

Intelligent Process Control. A computer is now expected to provide also support for the management of the dyeworks. The computer should provide information on:

1) Cost reduction (cost reporting and control)

2) Minimization of reject lots (reproducibility, buffering effect)
3) Shorter treatment time of the goods (optimization of operations planning)
4) Less production of wastewater (independent optimization, washing process)
5) Optimization of product range (amount and range of stored materials, standardization of similar products)
6) Development of new methods, e.g., by calculation of the relationships between dyeing behavior and process parameters

Personal computer programs are only available for some of these areas [31,26]. They are usually written by dyeworks staff for particular types of dye. The information required is often only available indirectly (see also Section 2.1.3.6). The simulation of a dyeing process on the computer screen to enable different procedures to be evaluated is still at an early stage of development [32,33,36].

2.4. Laboratory Dyeing Techniques

2.4.1. Objectives

An important task of the laboratory is to carry out preproduction tests, especially of new formulations, to give an indication of the results to be expected [85]. The laboratory also helps in maintaining consistency of production. It should also decide whether tolerance ranges should be wide or very narrow. Factors to be considered include dyeing rate, batch time, and cost.

Important Individual Tasks of the Dyeworks Laboratory

1) *Determination of Formulations for Production Dyeing.* Searching in the pattern collection, feasibility of producing the color shade on the given substrate, transfer to the production plant.
2) *Shading Formulations.* Determination of dye additions during dyeing.
3) *Testing of Dyed Products.* The various fastness properties are most important. Very specialized methods and equipment are sometimes required (see Chap. 13). Supporting tests of mechanical properties of textiles are also carried out (see Chap. 3).
4) *Dyeing of Samples for Color Measurement* (calibration curves). Preparation of special dyeings of new fabrics supplied by the customer.
5) *Investigation of Defects and Recommendations for their Correction.*
6) *Optimization* (cost minimization, reduction of water consumption, minimization of dye concentration in wastewater, etc., selection of dyes with improved mutual compatibility [79]; improved matching of the fastness of lighter or darker shades; statistical analyses).
7) *Reclamation Analyses.* Investigation of samples of rejected materials for quality discrepancies.
8) *New Methods and New Developments.* Development of methods of dyeing and finishing new materials with the available machinery. Ideas for new effects and transformations in the finishing process.

2.4.2. Laboratory Dyeing

2.4.2.1. Typical Laboratory Equipment

The amount of textile material available for a laboratory scale dyeing operation is often in the region of a few grams. With such small amounts, the dyeing equipment in the laboratory cannot adequately represent the full-scale plant. The main aim is to match the patterns provided and to produce dyed material for fastness tests.

Rope Dyeing Machines. The textile material (hanks of yarn, woven or knitted material) is mounted on a carrier, usually of the double-forked type, and moved up and down in the bath. The dyeing bath is placed in a large and sturdy test tube in a temperature-controlled bath at between room temperature and ca. 95 °C. Rope dyeing machines are a mechanized form of the early method of rope dyeing by hand. A typical example of this is the Ahiba apparatus.

Beaker Dyeing [High-Temperature (HT) Bomb]. The textile material and liquor are placed in a narrow, cylindrical, pressure-tight steel container with a screw top. The container is placed in a high temperature medium (oil, sand, air, glycol, aluminum block [47], or an infrared

field [86]) where it is kept in continuous motion, so giving good interaction between the fabric and the liquor. Dyeing temperatures of 140 °C can be reached in the HT bomb. This system has found wide application.

Through Flow Dyeing Equipment. A certain range of laboratory equipment is available in which the liquor flows through the textile material. This equipment requires a larger quantity of textile. These types of dyeing equipment are used for special tasks, a typical example being the "Colorstar" equipment.

2.4.2.2. Small-Scale Production Equipment

Batch Dyeing (Bath Dyeing). To provide a meaningful representation of the dyeing conditions in full-scale equipment, larger textile samples (in the range 100 g to 10 kg) are necessary, which usually make it possible to work with materials of the form used in full-scale production. Dyeing equipment taking samples up to ca. 50 kg in weight covers the size range between the laboratory and the very small scale production equipment. In general there are laboratory versions of the equipment described in Sections 2.2.3 and 2.2.4 [87].

Dyeing Equipment (Stationary Goods, Circulating Liquor). Various inserts to the dyeing vessel are available so that yarns (on spools, hanks), flock (compressed into baskets) and piece goods (beams of woven or knitted fabrics) can be dyed.

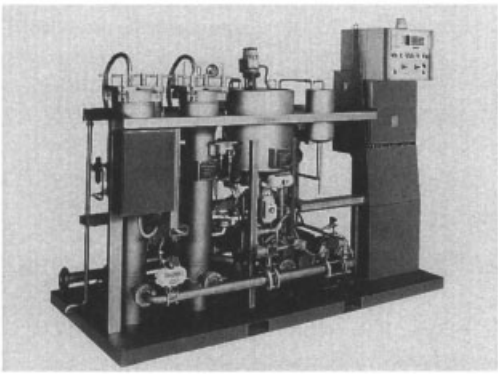

Figure 32. Laboratory Multiflex dyeing machine (courtesy of Then)

An example is the Multiflex dyeing machine (Fig. 32) which has been specially designed to provide an adaptable model of the production process.

Dyeing Machines (Circulating Goods, Stationary/Circulating Liquor). The most important dyeing machines are the *jigger*, in which the width of the fabric is ca. 40 cm, and the *jet* or *overflow* dyeing machine which can contain between 100 g and several kilograms of fabric. The extent of automation corresponds to that used in production equipment (see Section 2.2) [88].

Continuous and Semicontinuous Dyeing. The heart of the continuous laboratory dyeing machine is the padding system with a trough for applying the dyeing liquor (see Section 2.3.1.1). This determines the maximum operating width (ca. 30 – 50 cm) and the range of running speeds of the fabric (0.5 – 1.5 m/min). Many different versions of the padding process exist, ranging from the simple cold pad-batch process to a complete dyeing plant [42]. A steamer and a dryer (for intermediate drying) [89] are often available, the latter also serving as thermosol equipment. Finishing of the dyed product is often still carried out batchwise in a glass beaker or a bucket, as this does not require space and capital expenditure for operations that can easily be improvised.

In addition to universal dyeing equipment, more specialized equipment is also available, e.g., continuous yarn dyeing equipment for indigo (→ Indigo and Indigo Colorants, Chap. 9.). This is understandable, as dyeing with indigo accounts for by far the largest volume of dyed products. Here, the dye is applied to the (usually gray) yarn wet-in-wet, which does require special equipment. The Looptex laboratory equipment is of interest here.

2.4.3. Laboratory Dyeing Technology

All the principles described in Section 2.1.3 and all the recommendations that do not relate to specific equipment made in Sections 2.2 and 2.3 also apply to laboratory scale dyeing. Some dyeing methods that are special to laboratory dyeing are:

Dyeing with Exhausted Liquor. A dyeing operation is carried out on white material in the exhausted liquor. This can also be carried out in dyeing operations that have not reached equilibrium to obtain a shadow dyeing. Information about positions of equilibria, dye compatibility, and dye absorption rates can be obtained in this way. Dyeing with exhausted liquor gives results that provide more information to the dyer than plotted curves, and constitute a descriptive and reliable document.

The sample collection may contain a sample similar to the pattern provided, and this sample often has an optimized combination of dyes and reliable fastness properties, so that the relevant formulation can then easily be modified to match the pattern.

Defect Analysis and Correction. The successful correction of a lot often confirms the defect analysis. The causes of defects can be very complex. Computer aided methods for defect analysis are being developed (e.g., expert systems). However, systematic instructions have not been published (company know-how).

Statistical Evaluation. Statistical topics of interest include: frequencies of defects in returned materials (e.g., whether stains are more common than color discrepancies, effecting repair costs), "transfer factors" enabling colorimetric formulations to be converted to laboratory formulations and then production formulations, test dyeing methods [79], distribution of the additives for color shading, substrate, dyeing equipment, and, of course, "substrate calculations", e.g., viscose → cotton [90, 91].

2.5. Techniques of Dispensing Products used in Dyeing

2.5.1. Dispensing of Dyes

The dispensing of dyes forms an important part of dyeing technology [47, 92].

Amounts required for a dyeing operation range from ca. 5 g dye (small batch of 50 kg fabric and a very low dye content, 0.01 %) to ca. 50 kg dye (large batch of 500 kg fabric and dark dyeing, e.g., black, requiring 10 % dye). Hand measurement is much more common than fully automatic metering for dyes.

Hand Weighing of the Dye. Hand weighing of the dye still has advantages with which automation is only slowly catching up:

1) Wide range of products
2) Very wide range of addition rates
3) All forms of dye handleable
4) Ease of changeover between products
5) Low capital cost

Disadvantages include high staff requirements, the need for protective measures against dust from the dye, and occasional mistakes in weighing.

Computer-Aided Dye Weighing. The balance is linked to a computer. This displays the name of the dye required and the amount to be weighed.

Weighing by robots is also possible in the laboratory. Robots to some extent reproduce the range and scope of hand weighing.

Automatic Dispensing of Dye.

Continuous Dispensing of Dyes, Liquors, and Additions. For liquors with a limited service life it is advantageous to prepare stable masterbatches which are mixed shortly before use. These are almost exclusively used in reactive dyeing.

At the pad trough the dye solution and dilute alkali are delivered by a single metering pump with two heads. The finished liquor being fed into the pad trough via a mixer. In this way, the volume ratio of the two components is kept constant even if the total liquor consumption varies.

Thus, highly reactive dyes can be used in the cold pad-batch process, in which a sufficiently rapid reaction takes place on the fabric while it is stored (see Section 4.1). The very short time for which the dye liquor is stored at room temperature is not long enough to lead to any significant hydrolysis of the dye.

Batchwise Dispensing of Dyes. This involves the preparation of measured amounts of dye as required by a formulation. The most important requirements in such a system are:

1) Quantity range ca. 5 g to 50 kg

2) Measuring accuracy ca. 1 – 2 % of amount dispensed
3) Dispensing rate, e.g., 10 formulations per hour

The timing problem is solved by making all the addition times equal, which corresponds to widely different addition rates.

Metering systems are now available which operate at a rate corresponding to the quantity measured out. Two systems exist, in the first the weighing operations take place successively (rough weighing followed by accurate weighing, time-consuming), in the second large and small amounts are metered out simultaneously (using pulsing pumps in parallel, expensive).

The most important types of metering equipment are listed below:

1) *Balances* (often 2 – 3 for various ranges), which give high accuracy, and provide a practicable set of equipment for many or all dyes, whether powdered or liquid.
2) *Metering pumps* (piston or membrane pumps, both with valve problems), in which the accuracy is determined by the stroke, one stroke being the smallest metered amount. The metering rate is adjustable, product distribution is required if several dyes are metered per pump. Pumps are only suitable for liquid dyes and liquid masterbatches.
3) *Flowmeters* in which measurement is performed by direct measurement of flow rate or a counter system (water meter principle). For dispensing, valve switching is necessary between several flowmeters of various sizes to give accuracy and metering capacity. Flowmeters are only suitable for liquid dye mixtures and masterbatches.

In general, the range of dyes is selected so that the concentrated dyes from the same manufacturer can be mixed and therefore do not require preliminary dilution.

Liquid dyes and masterbatches are used for liquid metering systems. The masterbatches often require special formulation to ensure storage stability (dye sedimentation).

Powdered and granulated materials are still at present only metered by weighing systems. The dye is kept in a container (silo) from which it is fed into a feeding device (vibratory chute).

2.5.2. Dispensing of Dye Auxiliaries

Dye auxiliaries, which must be matched very specifically to the formulation, e.g., retarders for dyeing polyacrylic fibers, are usually dispensed like the dye. The products usually have an affinity to the fiber. The quantities of auxiliaries required for the formulation are usually expressed as wt % of the weight of the fabric. Their concentration in the liquor therefore varies as the liquor ratio varies. The metered volume (mL) is given by the following equation:

$$\text{Volume} = \frac{C\% \times W_{\text{Tex}} \times 10}{D}$$

where C % is the concentration of the auxiliary in wt %, W_{Tex} the weight of fabric in kg, and D the density of auxiliary in g/mL. For example, if the concentration of a retarder is 0.7 %, its density 0.93 g/mL, and the weight of fabric 100 kg,

$$\text{Volume} = \frac{0.7 \times 100\text{kg} \times 10}{0.93\text{g/mL}} = 753\text{mL}.$$

The other auxiliaries are usually dispensed with the aid of a liquid metering station. These mainly include products that modify the liquor, e.g., surfactants, dispersants, leveling agents, or complexing agents, which are thus also products with an affinity for the dye. The amounts required in the formulation are usually given in g/L.

The metered volume (mL) is given by the following equation:

$$\text{Volume} = \frac{C_{\text{g}} \times W_{\text{Tex}} \times LR}{D}$$

where C_{g} is the concentration of the auxiliary in g/L, W_{Tex} the weight of fabric in kg, LR the liquor ratio in L/kg, and D the density of auxiliary in g/mL.

Powdered products are often converted into stable masterbatches in simple dissolving stations.

The most important measuring devices are metering pumps, flowmeters, and measuring cylinders.

2.5.3. Dispensing of Chemicals

The main chemicals used in dyeing are salt, alkali, acid, reducing and buffering agents. These are often required in large quantities, and are stored separately. Salts and liquids (alkaline solution) are metered directly, salts from the salt wagons (mobile containers with feeding equipment), alkaline solutions from the solution wagons (mobile containers with delivery pumps). Stock solutions, e.g., hydrosulfite and alkali (vat dyeing, reductive washing) are prepared in a powder dissolving station. The measured amount of powder is fed into a mixer container by a screw conveyor. The solution of alkali is added, and the powder dissolved. The solution is then delivered through pipework to the point of use. Acid and alkali can be added under the control of a pH meter [60]. Alternatively, the addition can follow a predetermined program which specifies the (absolute) amount, the addition time, and the metering characteristics (linear, increasing rate, decreasing rate).

2.5.4. Preparation of the Initial Liquor Charge and its Replenishment

The required measured amounts of products must be dispersed, diluted, or mixed, and then added to the dyeing equipment. Details vary according to the dyeing technique.

2.5.4.1. Batch Dyeing

Dyes.

Initial Charge. The measured amounts of dye must be predissolved or prediluted, either at a central location or in situ on or in the dyeing equipment. The volume of the liquor and hence the concentration do not need to be exact. To achieve this, the dyes required for the dyeing operation are predissolved or dispersed in containers fitted with stirring or mixing devices, using a volume of water equal to ca. 10 % of the final volume of dye liquor, and this is added to the bath.

Transport of the product is by pumping or gravity in inert pipework, and distribution is by valves or swinging arm distributors [93].

Replenishment. When dyeing with a standing bath, the dye used must be replenished batchwise before each new batch of fabric. Replenishment is carried out with dye masterbatch solutions and dispersions. The required quantities are withdrawn with pulsed metering pumps from the silo and distributed via valves to the individual dyebaths.

Auxiliaries and Chemicals. All the auxiliaries needed for a treatment stage are added to the bath at the same time, either directly or via a central metering and mixing station [93].

The quantities and addition of products are usually called up by the dyeing program control system. Good communication between the control computer for the dyeing equipment and the central metering and control computer is necessary.

2.5.4.2. Continuous Dyeing

Dyes. In *continuous dyeing* of warp yarn with indigo standing baths and special wet-in-wet methods of application and fixation of the dye are used.

A masterbatch of replenishment liquor containing reduced indigo (indigo white, leucoindigo) is metered into the dyebath. The fabric itself in a way controls the replenishment rate via (1) the difference between the amounts of liquor on the fabric entering and fabric leaving the bath and (2) via the evaporation per pass ("skying").

The masterbatch liquor (replenishment liquor), containing ca. 50 – 70 g indigo, is prepared from indigo (100 %), hydrosulfite (excess), and sodium hydroxide solution.

Alternatively, grades of indigo are available (BASF) that already contain the reduced dye. This considerably simplifies preparation of the replenishment liquor (see Section 4.4.2).

In *pad dyeing* the total concentration of dye and auxiliary in the liquor must exactly meet the requirements of the formulation. Therefore the liquor volume must be determined as precisely as possible and the rinsing water must be included in the final volume of liquor. The predicted consumption of liquor is calculated from the weight of the fabric and the pickup. This figure is increased to allow for the liquor volume

in the trough (ca. 10 – 100 L) and a safety factor. Photometric measurements for control of the liquor preparations are under discussion.

The finished batch of liquor is transferred to a storage container near the equipment, and is transferred by pumping as required into the trough.

The dispensing and mixing of liquid additions to the liquor is described in Section 2.5.1.

Dyeing auxiliaries and chemicals (e.g., wetting agents and defoamers) are added to the dyebath along with the dyes, and are then applied to the fabric.

Liquid chemicals and stock solutions are prediluted, usually as a part of the metering process, and are added to the liquor.

Powdered products can be weighed out and made up into masterbatches which are then metered to the process. Alternatively, the powder can be dispensed directly at the point of use in the process (usually by screw feeders).

Replenishing solutions are usually made up at the point of addition to the dyeing operation.

2.6. Colorimetry (→ Dyes, General Survey, Chap. 3.)

Colorimetry means the measurement of the reflection spectrum of a sample and the interpretation of this with the aid of special software. This software calculates color coordinates (CIELAB tristimulus values) and color differences, usually for comparison purposes (patterns/samples). Alternatively, the software can calculate dye formulations based on calibration curves, and can indicate the quality of a calculated formulation. The most important uses of colorimetry are listed below:

In the *laboratory* colorimetry is used to support preparation of formulations and to evaluate the addition of dye for color shade adjustment, for the investigation of defects, and for quality control. In *production* on-line color measurement is used for the control of dye pickup, and as an aid to inspection of the fabric.

Colorimetry has virtually replaced visual assessment by the dyer. However, it cannot adequately take account of secondary effects (luster, dichromaticity, moisture in the sample, nap effects, optical color effects with samples sewn together).

2.6.1. Measuring Instruments

The most important part of the instrument is the measuring head. Here, polychromatic light falls on the multilayered sample. The reflected light is spectrally decomposed (ca. 30 wavelengths), and reflectance spectra are obtained.

The mathematical interpretation is performed by the built-in or attached computer, today usually a PC. A good review of the measuring instruments exhibited in ITMA 91 is given in [96], capabilities of the latest software are discussed in [97]. For on-line measurement, modified instruments are used which enable measurements to be carried out remotely from the fabric in full daylight [98, 99].

2.6.2. Methods of Expressing Colorimetric Results

Color coordinates and color differences are expressed using CIELAB tristimulus values (→ Dyes, General Survey, Chap. 3.2., → Dyes, General Survey, Chap. 3.3.).

Color differences (pattern – sample) are calculated from the absolute color coordinates (→ Dyes, General Survey, Chap. 3.5.). The total color difference (dE) gives information on color differences for the user.

Tolerable total pattern – sample color differences are typically within the range ca. 0.8 – 2.0 (strongly dependent on the type of fabric).

Pass/fail formulae [100] enable somewhat more reliable decisions to be made concerning the acceptability of dyed products from a given fabric.

Representation of results in tabular form is the most comprehensive system, but for some purposes is too detailed. Figures usually quoted are the color coordinates X, Y, and Z, the absolute CIELAB values (coordinates):

$a*$ = red – green (+, −)
$b*$ = yellow – blue (+, −)
$L*$ = lightness (light = 100)
$C*$ = chroma
$H*$ = hue (angle of 0° = red, 90° = yellow . . .),

and the color differences $\Delta E*$, $\Delta H*$, $\Delta C*$, $\Delta L*$, $\Delta a*$, and $\Delta b*$, together with an identification number of the sample.

The graphical representation illustrated in Figure 33 is widely used, and is almost a standard, despite some problems. The deviation of the sample from the pattern (target) is represented in a field (d*a*, d*b*) and on a scale (d*L*). This gives a very good understanding of the distribution of the samples around the target.

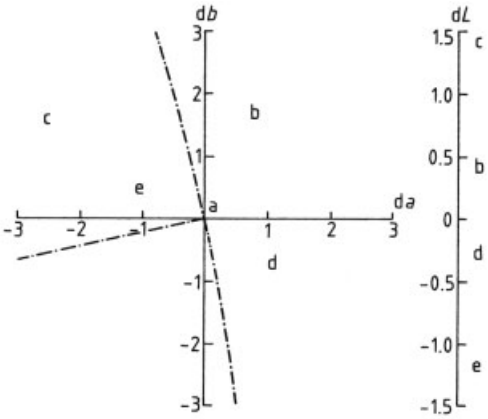

Figure 33. Graphical representation of color differences (for explanation see text; data MASTER, samples; courtesy of BPI Hohenstein)
a = Pattern; b, c, d, e = Samples

In the graphical representations, neither the color location of the pattern nor the color differences d*E* between the samples and the pattern (which are very important for many assessments) are usually shown.

The pattern *a* is in the center of the d*a*, d*b* field; sample *c* lies in the sector −d*a* (more green than the pattern), +d*b* (more yellow than the pattern); sample *d* lies in the sector +d*a* (more red than the pattern), −d*b* (more blue than the pattern).

Although results from colorimetric comparison measurements are useful for the quality control of the fabric, they are less useful to the dyer for defect analysis and prevention. For defect analysis knowledge of the concentration defects is required; these provide information about the dyeing operation (e.g., unintentional retention of a dye in the bath) [101].

Another important application is in the calculation of formulations. With a new pattern, formulations are calculated with the aid of the calibration curves. For this, *K*/*S* values (approximately proportional to concentration; *K* is the coefficient of absorption, *S* the coefficient of scattering, see Chap. 12) are calculated from the measured reflection values. The *K*/*S* curves are then superimposed until the *K*/*S* spectrum of the pattern is matched. This enables an estimate to be made of matching, fastness properties, and costs, and a series of laboratory tests are carried out until the pattern is matched with sufficient accuracy.

2.6.3. Developments in Colorimetry

The development of colorimetric measuring instruments is proceeding rapidly (e.g., hand-held instruments [47, 97] and photo-acoustic color measurement).

The development of the interpretation of measurements is remarkably stagnant. After a measurement, the software should be able to provide all possible interpretations, as with data-driven expert systems. Thus, derivation of such information as the characterization of unlevelness, the elucidation of variations in dye properties, and the assessment of migration and two-sidedness is often desirable, but is still only at the development stage.

3. Physical Properties of Textiles Important for Dyeing

3.1. Classification of Textile Properties

The *physical properties* described here include all the geometrical, mechanical, and optical properties of the textile, together with physicochemical properties such as wetting and swelling.

Information on the following physical properties can be found in the literature:

Thermal and calorific data [102–104]
Amorphous and crystalline regions [105–108]
Frictional behavior [109, 110]
Hairiness of yarns [111–113]
Conductivity and electrostatic charge [114–116]
Stiffening and strengthening [111]
Shrinkage behavior [117]
Surface appearance [118–120]

Table 6. Physical properties of various textile fibers

Type of fiber	Density, g/cm^3	Diameter, µm	Titer [a], dtex	Surface [b], m^2/kg	B [c], g/dtex	EB [d], %
Cotton	1.52	16	3	160	4	6
Fine wool	1.3	20	4	154	1.5	35
Coarse wool	1.3	60	37	50	1.2	35
Silk	1.3	10	1	310	4	22
Glass	2.54	8	1.3	200	8	2.8
Cellulose acetate (2.5)	1.32	28	8	110	1.2	27
Viscose, normal	1.52	26	8	100	2	22
Viscose, high strength	1.52	12	1.7	220	3.5	15
Polyamide, normal	1.14	27	6.5	130	5.2	28
Polyamide, high strength	1.14	30			6.9	19
Polyester filament	1.38	20	4.3	150	4.6	26
Polyester staple	1.38	21	4.8	140	3.8	42
Polyester, micro	1.38	9	0.9	322		
Polyester textile	1.38	15	2.5	190		
Polyacrylonitrile (X-51)	1.17	21	4.0	163		27
Polyacrylonitrile (Acrilan)	1.35	22	5.1	135		18

[a] 1 dtex = 1 g per 10 000 m = 1.11 denier. [b] Specific surface area. [c] Breaking strength of single fiber (dry). [d] Elongation at break (dry).

The *chemical properties* of the textile substrate and the functional groups that cause bonding of the dye are described in detail in Chapters 4 – 11.

3.2. Fibers [121–125]

(see also, → Fibers, 1. Survey, Chap. 3.; → Fibers, 6. Testing and Analysis, Chap. 2.3., → Fibers, 6. Testing and Analysis, Chap. 3.)

The *fineness* (cross section) of the fiber is an important geometrical parameter. The *titer* (mass per 10 km fiber length, unit dtex) is determined from the fineness and the density of a fiber. The specific *surface area* (ca. 50 – 300 m^2/kg) is also important for the dyeing process, indicating the area of the surface into which the dye diffuses and the extent of the dyed layer (ring zone of the fiber impregnated with dye, Fig. 34). This determines dye consumption: a thicker fiber takes up dye more slowly because the specific surface area for the entry of dye is smaller. It also consumes less dye (formation of a ring zone [126, 127]), which partially compensates the lower dyeing rate. This holds, especially for synthetics, when the orientation of the fibers is nearly the same. For a uniform dye concentration in equally thick ring layers of different fibers, the average dye concentration in the thicker fiber is lower, e.g., 75 % in the example given in Figure 34. If the transparency of the fibers is not too high the visual impression of both fibers should be the same. Some physical properties of various textile fibers are given in Table 6. (The figures listed are mean values of those cited in [126, 128]. They are reference values for comparisons and calculations.)

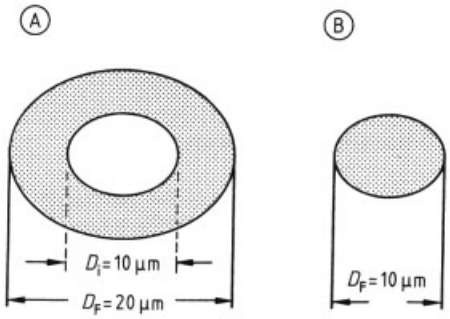

Figure 34. Cross sections of fibers of various thickness after a standard dyeing process; the shaded zone is homogeneously charged with dye
A) Specific surface area = 154 m^2/kg; the colored ring zone represents 50 % of the diameter, but 75 % of the cross-sectional area and of the fiber volume; B) Specific surface area = 310 m^2/kg

Fibers usually show *stress – strain behavior* in which an elastic (reversible) deformation is followed by permanent elongation. Internal ("frozen") stresses are produced. A partial "melting away" of these permanent stresses leads to (latent) markings. The strongly anisotropic structure of a fiber shows many kinds of reactions to stress (relaxation, retardation).

An attempt to formulate the equation of state of a fiber and to describe the transitions between states can be found in [129].

Another important parameter for dyeing is *tenacity*, even though the stresses occurring during dying are much lower than this. Tenacity is based on the cross section and the orientation of the fiber and expressed in g/dtex (Table 6). The *elongation* of fibers is equally important for certain dyeing techniques. The elongation at break is a useful first approximation (Table 6). Generally, the elongation at break is lower for higher-strength fibers, particularly among fibers of the same chemical type.

An important optical property is *light scattering*, which directly affects the depth of color (see also page 117). Light scattering increases with delustering of synthetic fibers and decreases with mercerization of cotton (corresponding to an increase in color intensity).

Hygroscopic fibers (e.g., cellulosic fibers or wool) *swell* under the influence of water or alkaline solution and sometimes also undergo structural changes [107, 130]. This swelling facilitates the diffusion of dyes into the fibers [131].

When thermoplastic fibers are below the glass transition temperature range, the dye diffuses very slowly into the fibers (which behave like a solid), but above this temperature, diffusion is very rapid (fiber behavior resembles that of a liquid). In the transition region, diffusion (dyeing rate) increases strongly with increasing temperature.

3.3. Yarns

The important properties for the dyeing of yarns are their shrinkage and stretching behavior. These are influenced strongly by the swelling and temperature (below or above the glass transition range). Other physical properties associated with the yarn are determined mainly by the fiber material (Section 3.2), the makeup used for the dyeing process (Section 3.5), and finally, the properties of the woven or knitted article (Section 3.4) that is made of the yarn.

Yarn has by far the highest marketed volume of all dyed goods, due to Indigo dyeing of warp yarn for denim fabrics made of gray cotton. Extreme ring dyeing (thin ring layer) of these yarns is preferred.

3.4. Fabrics

Fabrics [e.g., woven products (smooth and pile fabrics), knit products, and nonwovens] represent the final stage of textiles before manufacture into clothing, etc. [132]. Woven fabrics account for most of the textiles dyed by continuous dyeing. Around 50 % of fabrics are batch dyed.

The geometrical properties of fabrics are determined mainly by their structure and the type of yarn. The most important of these include mass per square meter (100 – 1000 g/m^2), width (0.8 – 5 m), and mass per running meter (80 – 5000 g/m).

A good indication of the structure is obtainable from the thread count (threads per centimeter, warp, weft) or the number of stitches [133].

The terms open or dense refer to the structure of a textile, and hence its permeability properties, which are also important. These are indicated by the thickness, interstitial volume and porosity, light transmission properties, resistance to the flow of fluids (air, water), and yarn number and twist.

Also important for good dye penetration is the accessibility of yarn crossing points to the dye and the release of air from the interstitial volume of the fabric during dyeing (prevention of white spots).

Of the mechanical properties, *stress – strain behavior* is the most important [107, 134]. Longitudinal tension during dyeing causes not only an elongation (reversible, irreversible) in the direction of stress, but also a very characteristic shrinkage in width. This must be corrected at a later stage by stretching to the specified final width [117]. Varying warp tensions over the width of the material in the weaving process lead to loose or slack edges [135, 136]. This results in variable liquor pickup (usually, the higher the warp tension, the lower the pickup).

The *running properties* of the fabric play a very important role in continuous dyeing (e.g., longitudinal creases, ribbed effects, lateral movement, weft displacement) [137]. The running behavior cannot be simulated adequately in the laboratory [138, 139]. Within limits, running properties can be improved by increasing the longitudinal tension and reducing the running speed. With certain products (especially impermeable materials), the wet material is sometimes

not in direct contact with the guide rollers and can move laterally (aquaplaning).

The *liquor pickup* after squeezing off and the *liquor-carrying capacity* of a material are other important parameters [140]. As a rough guide, dense materials have a lower liquor pickup (40 – 60 %) than open textiles (70 – 90 %). The maximum liquor-carrying capacity (with no streaking due to liquor runoff) corresponds roughly to the interstitial volume of the material (1 – 3 L/kg).

Of the *optical properties*, the luster of the material and the raising of a nap on its surface are the most important. While the luster (caused by sharp angles of reflection) usually affects only the color measurement, raising a nap often causes a change in color (brightening). Another effect in fiber mixtures is the clear visibility of one type of fiber even though the color intensity difference is small.

3.5. Makeup of Textiles for Dyeing

The textile material is usually made up in a rotational symmetrical form for the dyeing machine (rolls of yarn or fabric, packages of flock). During dyeing, the liquor flows radially through the textile material. The greatest resistance to flow and hence the greatest pressure drop occur at the inner parts of the roll. Thus, the inner parts are exposed to the highest stress on the material (ratios of the inner to the outer resistance for cross-wound packages, ca. 5: 1; for warp beams, ca. 2: 1, see Fig. 35).

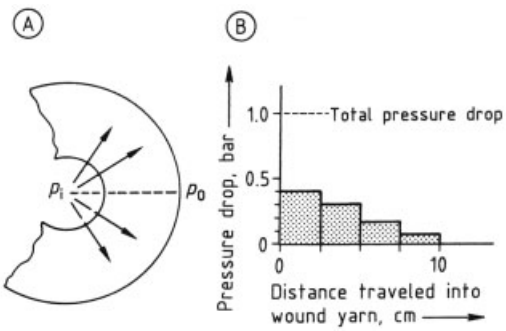

Figure 35. Local pressure drop in a cross-wound cylindrical package during radial passage of dye liquor (density of wound yarn, 300 – 400 g/L; mass, ca. 2 kg; volume of wound yarn, ca. 5.7 L; i.d., ca. 5 cm; o.d., ca. 25 cm)
A) Cross-wound package, planview; B) Pressure drop across each zone
p_i = inner pressure; p_o = outer pressure

An important factor is the uniform and rapid wetting of the material with the dye liquor [141]. For this, adhering air must be removed (deaeration) by appropriate pretreatment of the textile (dye-ready material), steaming, or the addition of auxiliaries (wetting agents; → Textile Auxiliaries).

The total resistance to the flow of dye liquor determines the pump adjustment used and the flow rate. This resistance often changes during dyeing because of alternating flow direction of the liquor (greater when flow through the textile or package o → i, smaller when i → o; o = outer part, i = inner part), swelling, temperature changes, and fiber shrinkage, to mention only the more important influences.

When using dyeing machines in which the textiles are stacked during dyeing (winch beck, jet, overflow), crease marks (from the dye) or permanent creases (mechanical) may form [142–145]. The most important methods of preventing these effects are:

1) Better presetting of the textile
2) Lower machine loading
3) Better crease control
4) Longer leveling time
5) Slower cooling

Control of the textile throughput rate or cycle time shows whether the material is being transported correctly through the jet equipment (jet diameter, liquor injection rate).

In jiggers, the running problems described in Section 3.4 can occur.

In continuous dyeing yarns [104] and woven fabrics [105] are used without a special makeup.

4. Dyeing of Cellulose Fibers

4.1. Dyeing with Reactive Dyes

Reactive dyes are the newest class of dyes for cellulose fibers (→ Reactive Dyes, Chap. 2.). In the dye molecules, a chromophore is combined with one or more functional groups, the so-called anchors, that can react with cellulose. By suitable choice of dyeing conditions, covalent bonds are formed between dye and fiber. More than 100 years after SCHOTTEN and BAUMANN first carried out fundamental investigations into the

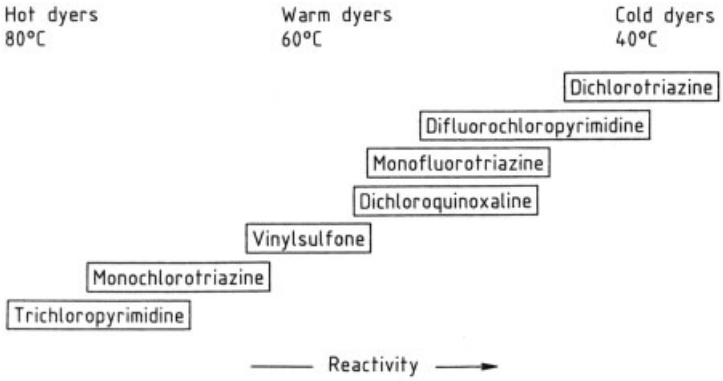

Figure 36. Reactivities of established anchor systems and optimum dyeing temperatures in exhaustion dyeing [153]

reaction between alcohols and acyl halides (acylation), ICI introduced the first group of reactive dyes for cellulose fibers, the Procion M dyes (2,4-dichloro-1,3,5-triazine anchor; → Reactive Dyes, Chap. 1.) in 1956, which initiated an intensive development effort. One-third of the dyes used for cellulose fibers today are reactive dyes. They have a better wetfastness than the less expensive direct dyes. The range of available reactive dyes is wide and enables a large number of dyeing techniques to be used. Shades ranging from brilliant to muted can be obtained. Chlorine fastness is slightly poorer than that of vat dyes, as is lightfastness under severe conditions.

4.1.1. Fundamentals [146–148]

In general, dyeing with reactive dyes is very similar to the well-known process of direct dyeing. The major difference is that the reactive dye forms a chemical bond to the fiber. Important factors determining the dyeing properties of reactive dyes include the affinity of the dye for the fiber (substantivity), its diffusibility, reactivity, and the stability of the dye – fiber bond. This bond constitutes the fundamental difference between direct and reactive dyes.

The anchor components of reactive dyes are important for both the fixing properties and the wetfastness of the dyed material. More than 300 electrophilic anchors for reactive dyes are described in the patent literature. Of this large number, fewer than ten are currently of practical importance for dyeing cellulose fibers (Fig. 36).

Hydrolysis and Reactivity. Two possible reaction mechanisms exist: addition and substitution.

Addition anchors include, for example, vinylsulfones

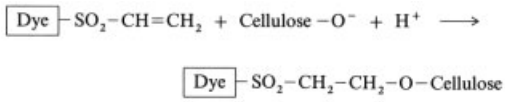

Substitution anchors include, for example, monochlorotriazines

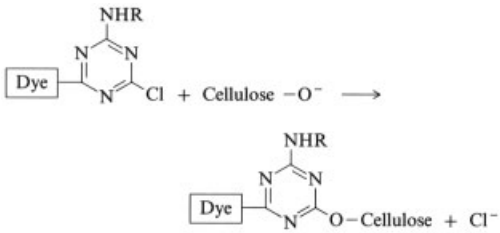

As *substitution anchors,* heterocycles with a reduced π-electron density in the heterocyclic ring are used. The rate of hydrolysis or alcoholysis (i.e., the reaction with the cellulose anion) is influenced by the nature of the heterocyclic nucleus, the substituents, and the leaving groups. In the reaction with the cellulose anion, the substitution anchors listed in Figure 36 are bonded by ester linkages. In contrast, cellulose adds to the *vinylsulfone anchor* in a nucleophilic reaction (Michael addition), and an ether link is formed.

Hydrolysis and alcoholysis are competitive processes, both being first-order reactions [149, 150]. The ratio of the rate constants of the reactions with cellulose and with hydroxide ions

(ratio cellulose-O^- : OH^-) is termed selectivity and can be determined easily in a homogeneous medium by using model substances. For triazine anchors, a decrease in selectivity depending on the leaving group can be observed in the order

$$F > Cl > DABCO > SO_3$$

(DABCO = 1,4-diazabicyclo[2,2,2]octane) [151]. However, the differences in selectivities of the established anchor systems are slight.

The preference of the anchor for the fiber is explained by the greater nucleophilicity of cellulose compared to hydroxide. At high pH, the ratio of cellulose-O^- : OH^- decreases [148, 152]. Furthermore, hydrolysis has a higher activation energy than alcoholysis (i.e., it is favored at higher temperature). As a result, the selectivity with respect to alcoholysis decreases with increasing pH and temperature.

The rate of hydrolysis increases as a rule by a factor of 10 per pH unit (first-order reaction). The rates of both reactions increase by a factor of approximately 2 – 3 per 10 °C. Therefore, for economic reasons (dyeing time) pH and temperature can be varied in practice only within narrow limits. Depending on the optimum dyeing temperature, established anchor systems are classified as cold, warm, and hot dyers (see Fig. 36).

In addition to the fastness and application properties the *fixation yield* (i.e., the percentage of dye chemically bound to the cellulose at the end of the dyeing process relative to the total amount of dye employed) is important, particularly for ecological and economic reasons. The fixation yield depends on the chemical nature of the anchor and many other factors [154].

Substantivity. Substantivity is given by the ratio of the amount of dye in the fiber to the amount in the dye liquor after chemical equilibrium has been reached (without chemical reaction). In analogy to direct and vat dyes, the adsorption properties of reactive dyes can be represented by the Freundlich relation. Substantivity is a necessary precondition for reaction with the fiber. Dye that has not become attached to the fiber is hydrolyzed when the alkali is added and lost to the dyeing process. For monoanchor dyes, high substantivity correlates with attainable fixation yields to a good approximation. Substantivity is influenced by the chemical constitution of the dye and the process parameters. In general, high substantivity is obtained with low dyeing temperature, a low liquor ratio, the addition of electrolytes (e.g., NaCl or Na_2SO_4) to the dyebath, and low dye concentration. An extended planar structure of the dye molecule is believed to favor high substantivity.

The dyer must also consider the effect of the substrate. Substantivities for mercerized cotton and regenerated celluloses are higher than those for cotton.

Limits to substantivity exist in practice. If the substantivity is too high, problems can occur with washing off of unfixed dye, wetfastness, or uneven final shade. In continuous processes, tailing can occur (see Section 4.1.2).

Diffusion Rate. The diffusion rate is the rate at which dye is transported to the interior of the fiber material. A high diffusion rate leads to rapid establishment of dyeing equilibrium and rapid leveling out of irregularities in the dyeing of the substrate. It also facilitates washing off the hydrolyzed dye. Increased dye substantivity is frequently combined with lower diffusibility. The diffusion rate doubles with a temperature increase of 10 – 20 °C.

The substrate also has an important influence on diffusion of the dye. The diffusion rate increases with caustic soda treatment or mercerization of cotton. However, with regenerated cellulose fibers, which have a marked skin – core structure, the outer parts of the fiber can act as a diffusion barrier.

Stability of the Dye – Fiber Bond. Because of the large variety of reactive dyes, generalizing about colorfastness is difficult. While wetfastness is determined mainly by the anchor system used, most other fastness properties depend on the dye as a whole or the chromophore present. Most reactive dyes are azo or anthraquinone derivatives whose standard of fastness varies greatly. Phthalocyanine, formazan, and triphenyldioxazine derivatives are also very important.

In addition, application conditions [155] and finishing processes (→ Textile Auxiliaries, Chap. 7.1.) can affect fastness. Thus, with some resin-finished textiles (dimethylolpropylenurea finish) a decrease in lightfastness is observed.

Significant differences in the stability of the bond toward hydrolysis can be observed in the anchor systems listed in Figure 36. The ether bond to cellulose formed by addition anchors (vinylsulfone dyes) is extremely stable to acid hydrolysis. Its instability in alkaline solutions is demonstrated by the poor fastness to hot sodium carbonate solution. In general, the stability of substitution anchors (fluorochloropyrimidines, monochlorotriazines, and monofluorotriazines) is significantly higher in alkaline solution. However, fastness can be reduced considerably by the presence of electron-attracting substituents in the anchor. An example is provided by dichlorotriazines, which are more reactive than monochlorotriazine dyes. The hydroxyl group formed by the second chlorine atom after reaction of the first with cellulose leads to a decreased electron density in the heterocyclic nucleus and to a significant increase in the rate of cleavage of the dye – fiber bond.

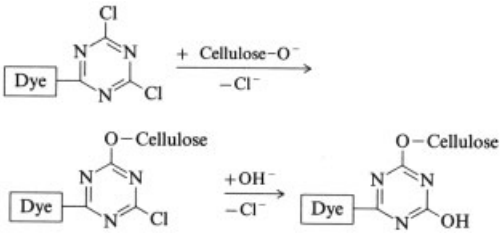

One aim of the introduction of heterobifunctional anchor systems has been to combine the good fastness profiles of different anchor systems (→ Reactive Dyes, Chap. 2.2.).

4.1.2. Dyeing Techniques [148, 157]

A wide range of reactive dyes of varying constitution is available which are suitable for many different applications. In most cases, dye is fixed to the substrate under alkaline conditions. However, dyes containing phosphonic acid groups and quaternary nicotinic acid derivatives (→ Reactive Dyes, Chap. 2.1.) are exceptions to this rule. Dyes with phosphonate groups are applied in an acid dyeing bath, quaternary nicotinic acid derivatives require neutral conditions. These dyes are applied under weakly acid or neutral dyeing conditions and are used mainly for dyeing polyester – cotton blends.

The suitability of a reactive dye for a particular application depends not only on its reactivity, but also on properties that are determined by its overall structure, such as solubility, substantivity, and diffusion. Special recommendations are therefore made by manufacturers for each type of dye.

About 20 % of reactive dyes are used in textile printing, 30 % in pad dyeing, and 50 % in exhaustion dyeing processes.

Exhaustion Dyeing. Exhaustion dyeing is a batch process. Loose stock, yarn, or piece goods can be dyed by this method. Equipment consists e.g., of package dyeing equipment, hank dyeing machines, jiggers, winch becks, or dye spraying machines (see Section 2.2.3). The general trend is toward the use of concentrated ("short") liquors (i.e., toward dyeing with the smallest possible volume of dyebath compared to the textile used). With woven textiles, liquor-to-goods ratios of 3:1 can be achieved; with knitted goods, 5:1.

In the exhaustion dyeing process, highly or moderately substantive dyes are usually employed. The two main processes are the "all-in" process and the stepwise process. In the *all-in process*, the dye, the salt, and the alkali are all added to the dyeing bath at the start of dyeing. The *stepwise process* consists of an exhaustion and a fixing phase (i.e., alkali is added only after the dye has absorbed to the fiber). The stepwise process has advantages in dyeing materials that are difficult for liquid to penetrate (e.g., dyeing wound packages in a circulating-liquor machine). A preceding exhaustion step aids in preventing an uneven final appearance. The exhaustion step is usually started at 30 °C. Dye and salts are added at the start or portionwise. The salt concentrations employed depend on the substantivity of the dye and the dye concentrations used. For <0.5 % dye based on textile weight, salt concentrations of 10 – 30 g/L are recommended; for less substantive dyes, up to 50 g/L NaCl. For deep shades with dye concentrations of >4 %, salt concentrations of ca. 50 g/L are used; with vinylsulfone dyes having low substantivity, up to 80 g/L. After the dye absorption phase, and sometimes also an additional leveling step (at temperatures sometimes higher than the fixing phase), alkali is added. With hot dyers, a uniform dye – fiber reaction can be obtained by con-

trolled heating. Heating rates are between 1 and 2 °C/min. In the case of cold dyers, fixing can be controlled smoothly by means of pH. Alkali can be added either in portions or by metering. Sodium carbonate or mixtures of sodium carbonate and sodium hydroxide are used almost exclusively today as the fixing alkali. Here, too, the amount of alkali is determined by the reactivity of the system and the depth of shade desired. Cold dyers are applied at lower pH values. The pH is normally adjusted with sodium carbonate. Warm and hot dyers are usually fixed with mixtures of sodium carbonate and sodium hydroxide. Typical concentrations are 5 g/L of sodium carbonate and 1 – 2 mL/L of sodium hydroxide solution (concentration 38° Bé or 440 g/L). The fixing time is 45 – 90 min.

After dyeing, the liquor is usually drained off; the material is rinsed and then washed off with addition of auxiliaries. The washing intensity depends on the substantivity of the dye hydrolysate. The more highly substantive hot dyers require longer washing times or more wash cycles at higher temperature. With alkali-sensitive vinylsulfone dyes soaping at the boil must be carried out in a neutral or weakly acidic liquor, to prevent cleavage of the dye – fiber bond.

Pad Dyeing Process. In this so-called impregnating process the dye, additives, and alkali are applied to the textile by dipping followed by squeezing out (see Section 2.3.2). Fixing of the dye on the textile is carried out in a subsequent step.

In these processes, dye and alkali can be applied together or separately. In the *one-bath process*, the so-called pad liquor stability is important. This term denotes the stability of the dye to hydrolysis as a function of time. With increasing reactivity of the dye (vinylsulfone < dichloroquinoxaline < monofluorotriazine < difluorochloropyrimidine < dichlorotriazine dyes), the risk of a loss in color strength with too long a dwell time in the pad box or replenishing tank increases. For this reason, metering equipment is commonly used. Dye and alkali are metered separately into the padder, which means that the dwell time in the pad box can be kept to a minimum. In addition, pad boxes are now constructed so that the liquor volume is as low as possible, and is, on average, replaced within 5 min.

When selecting dyes, the dyer must consider the possibility of so-called tailings. This refers to differences in shade and depth of shade between the beginning and end of the batch of fabric. With highly substantive dyes, the liquor running back into the padder from the squeezing rollers has a lower dye concentration than the initial liquor. This causes a gradual depletion of dye in the liquor. Also, especially with regenerated cellulose fibers, so-called negative tailing occurs. The dye uptake calculated from the liquor pickup is then greater than the actual figure.

In general, pad dyeing can be a semicontinuous or continuous process.

Semicontinuous Processes. Semicontinuous retention processes were developed especially for reactive dyes. The *cold pad-batch* process is by far the most important. After the textile has been padded with dye and alkali, the material is either plaited down in a box wagon or rolled up into batches. Fixing takes place during storage. The cold pad-batch process is of special interest from an economic point of view. The amount of alkali and the dwell times required for dye fixation depend on the reactivity and quantity of the dye. With cold dyers, adequate fixation can be achieved with sodium carbonate or bicarbonate; with dichlorotriazine and monofluorotriazine dyes, sodium carbonate concentrations of 2 – 30 g/L are used. In these systems, ca. 10 g/L of trisodium phosphate can also be used. With less reactive systems, sodium carbonate solution or mixtures of sodium carbonate and sodium hydroxide are employed. If the batches are stored in the open, carbon dioxide from the air can cause a pH decrease and hence a lighter final shade at the selvedges. A high process reliability can be achieved by utilizing the buffering action of waterglass (sodium silicate). In the past, additions of up to 100 mL/L of waterglass (concentration of waterglass solution 38° Bé = 1.35 g/mL) were recommended. For environmental reasons, the trend today is toward lower concentrations or complete avoidance of sodium silicate. By wrapping the batches in plastic film, the negative effects of carbon dioxide can be limited to the actual padding time. If necessary, the solubility of the dyes can be increased by addition of urea (up to 100 g/L). To obtain a level fabric appearance, the fabric must be uniformly absorbent. The maximum possible liquor uptake for cotton (ca. 60 – 80 %) is ca. 10 % lower than the fig-

ure for regenerated cellulose. To achieve evenness of dyeing by means of higher liquor pick-ups, auxiliaries that reduce running-down and dripping-out of the liquor can be added. Rotating batches are then used. Nowadays, highly or moderately reactive dyes are used for the cold pad-batch process. For a complete dyeing reaction, 2 – 8 h is required. However, this time is sometimes extended to 24 h for reasons of work organization. Storage at higher temperature, as in the pad roll process, or the use of accelerators (e.g., 1,4-diazabicyclo[2,2,2]octane) for slow-reacting dyes (e.g., monochlorotriazine anchor) has not widely spread. After dyeing, the material is rinsed and washed off in open-width or rope washing machines. Washing off directly on perforated beams is sometimes employed for economic reasons. The material is then dried.

Continuous Dyeing Processes. In continuous dyeing processes, padding, fixing, washing-off, and drying are carried out in one process line. Two different types of processes exist: dry-heat and steaming processes. The *dry-heat process* can be either the single-stage pad dry process or the thermofixation process. Hot dyers are preferred for use in *thermofixation*. Advantages of the thermofixation process include good color yields and covering immature cotton. The dye and alkali are padded at the same time. The preferred alkali is sodium carbonate, used at a concentration of 10 – 30 g/L. For less reactive dyes, sodium bicarbonate is recommended. In addition, 20 – 200 g/L of urea (nowadays because of ecological reasons often cyanoguanidine) is added, which in this process increases the solubility of the dye in the liquor, acts as a solvent for the dye during fixation (*mp* 133 °C), and prevents browning of the cellulose fiber. After pad dyeing, the material is dried, leaving a residual moisture content of 10 – 15 % because further drying can lead to loss of color strength. Subsequent fixation takes place at 150 – 200 °C, depending on the reactivity of the dye. In most cases, fixation times of 1 – 2 min are used.

In the *pad dry process*, fixing and drying occur simultaneously. Here, dyes of medium to high reactivity are preferred. The fixing time is 30 s to 5 min.

Of the *steaming processes*, the *pad steam process* is by far the most important. In the first stage, dye is applied by padding. Intermediate drying is then carried out at 110 – 140 °C. Alkali is applied by a subsequent padding machine. High salt concentrations are used, normally 250 g/L sodium chloride to prevent detachment of the dye and to give a high degree of fixation. The solubility of the dye can be increased by adding urea. Also here, the amount of alkali employed must be adapted to the dyeing system used. For less reactive dyes, up to 50 mL/L of NaOH solution is applied. Dye fixation is carried out with saturated steam at 103 °C for ca. 30 – 90 s. For dyes that are sensitive to reduction, a weak oxidizing agent can be added to the pad liquor.

In all heat-fixing processes, optimum fixing times must be adhered to. If the times are too short, loss of color strength occurs due to incomplete dye – fiber reactions. With too long a fixing time, cleavage of the dye – fiber bond is a danger, especially for vinylsulfone dyes. Also, in all drying operations, unevenness due to dye migration can occur. This can be prevented by the correct choice of dye combinations, auxiliaries, and drying conditions. After fixation, the material is washed off and dried.

4.1.3. Special Processes and Development Trends [158, 159]

In past years, many attempts have been made to remedy the weak points of the reactive dyeing system as much as possible. Lightfastness can be significantly improved over a wide range of shades by development of effective chromophores (e.g., introduction of the triphenyldioxazine and copper – formazan chromophores in the blue range). However, for demanding applications, especially in the red range, reactive dyes are still inferior to vat and naphthol dyes.

The salt load and discoloration of wastewater from dyeworks using reactive dyes often attract criticism on environmental grounds. With monofunctional dyes (i.e., those having only one anchor per molecule), the fixation yield in exhaustion dyeing processes is in the region of 60 %. An important step forward was achieved with the introduction of several anchor groups, giving double- or multiple-anchor dyes. These can be homo- or heterofunctional (mixed-anchor systems). When heterobifunctional dyes are provided with anchors of different reactivity, dyes

with a wide range of dyeing properties can be obtained and used in various kinds of dyeing processes (→ Reactive Dyes, Chap. 2.2.). At the same time, process reliability and dyefastness are improved. Thus far, combination of the triazine and vinylsulfone anchors has had the most important effect.

In terms of dyeing methods, a disproportionate growth in the economical padding process was seen in the 1980s. Nowadays exhaustion dyeing is regaining importance because shorter lengths of textile are being dyed to meet the changing requirements of fashion. In exhaustion dyeing processes, the trend is toward the short-liquor technique. A the same time, attempts are being made to develop new applications for reactive dyes. One example is the use of reactive dyes in continuous ranges to dye warp yarns (color denim process) [156]. These plants were originally designed for dyeing with indigo (→ Indigo and Indigo Colorants, Chap. 9.).

Special Processes. In addition, many special processes have been developed. An important aim has been the improvement of fixation yields. The two main techniques are:

1) Processes for chemically modifying cellulose fibers
2) Processes in which finishing and dyeing are combined

Introduction of nucleophilic groups (e.g., amino groups) into the cellulose fiber (item 1 above) can significantly increase the degree of exhaustion and fixation yield of reactive dyes. If reactive groups are introduced into cellulose, the possibility arises of bonding not only reactive dyes but also nucleophilic direct dyes covalently to the cellulose.

Of the combined techniques (item 2 above), the system BASF Basazol was the first to be successful commercially. After finishing the cellulose fiber with the trifunctional cross-linking agent 1,3,5-triacroylaminohexahydro-1,3,5-triazine, dyes containing sulfonamide groups were applied with high fixation yields. The Indosol system (Sandoz) is still used today. Here, copper complex dyes are used that can form mixed complexes with aminated resins. The high wetfastness is explained by covalent bonds between the resin and fiber.

However, the practical importance of this technique is still relatively low. Possible reasons for this include the additional cost of applying the dyes, changes in handling of the cellulose fiber, occasionally unsatisfactory fastness properties, and ecological problems.

4.2. Dyeing with Direct Dyes

Before the first synthetic dyes were made in 1856, natural dyes of plant or animal origin were used in the dyeing of textiles [160]. In most cases, the application of these dyes required time consuming process steps, e.g., mordanting or vatting.

Sun yellow (1883) and Congo red (1884), the first representatives of direct or substantive dyes (→ Azo Dyes, Chap. 4.), can be used on cellulose fibers "directly" (i.e., without mordanting). Because of their easy handling a variety of direct dyes has been developed.

4.2.1. Applications and Properties

Direct dyes account for 11 % (1990) of the worldwide consumption of textile dyes; 75 % of this amount is used to dye pure cotton or viscose substrates, and the rest is employed in dyeing mixed fabrics (for use in paper dyeing, see → Paper and Pulp).

Substantive dyes were not developed further as intensively as, for example, reactive dyes (→ Reactive Dyes) [161]. The main emphasis of research was on the replacement of possibly carcinogenic benzidine dyes (see Chap. 14) and on the improvement of pretreatment and aftertreatment processes.

Only in individual cases do direct dyes achieve the brilliance of reactive dyes. The lightfastnesses of substantive dyes covers the entire scale from 1 to 8, meeting the highest requirements. With increasing depth of color the wetfastnesses decrease to such an extent that direct dyeings must generally be aftertreated (see Section 4.2.7). Most direct dyes are characterized by excellent affinity to the fiber (i.e., dye liquors contain very little dye at the end of the dyeing process). This is of great importance for ecological reasons.

4.2.2. Dyeing Principle

Dyeing with direct dyes is based on the affinity of the dye for the cellulose fiber (substantivity). This affinity depends on the type of chromophore and can be influenced by the choice of dyeing parameters (see Section 4.2.4). Chromophores are azo compounds (→ Azo Dyes), stilbenes (→ Azo Dyes), oxazines (→ Azo Dyes, Chap. 3.), or phthalocyanines (→ Phthalocyanines) [162]. They always contain solubilizing sulfonic acid groups that are ionized in aqueous solution. The dye molecule is present as the anion. As a result, an electrolyte must be added to the dye liquor because cellulose fibers have a negative surface charge in water, which repels the dye anion. The cations of the electrolyte neutralize the negative charge and favor the aggregation of dye ions on the fiber (salting-out effect). Exothermic adsorption can be described with the help of Freundlich or Langmuir adsorption isotherms [163]. After adsorption dye molecules diffuse from the surface through the fiber pores to the amorphous areas of the cellulose. The rate of diffusion can be controlled by the dyeing parameters. Dyes are bound to the fiber by hydrogen bonds or van der Waals forces. Without additional measures (see Section 4.2.7), the strength of these bonds is so low that the dyes can be washed out of the fiber again. At dyeing equilibrium, the rate of absorption is equal to the rate of desorption.

4.2.3. Pretreatment of Substrates

Cellulose fibers require careful pretreatment before dyeing. Direct dyes are especially sensitive to differences in fiber affinity [161]. Pretreatment involves desizing, bleaching, possibly subsequent mercerization (→ Textile Auxiliaries, Chap. 4.2.), and increasing the wettability. Other techniques for improving the dyeability of cellulose with anionic dyes (e.g., forming of cations by fixation of quaternary ammonium compounds by reactive systems or by polymerization reactions [[161] and references cited therein]) have been investigated intensively. Although dye uptake has been successfully increased with these techniques, disadvantages such as shifts of shade, poor fiber penetration, or loss of lightfastness still hinder market acceptance.

4.2.4. Dyeing Parameters [164]

The dyeing parameters mentioned below are interdependent. The effects of these parameters on the dyeing process are described here.

The *liquor ratio* influences the dye solubility and the strength of the electrolyte effect. Efforts are made to employ lower liquor ratios for ecological, economic, and technical reasons.

The *electrolyte* (usually sodium chloride or sodium sulfate), its concentration, and the speed of addition control the adsorptive behavior of the dyes and the degree of exhaustion.

The *pH value* influences solubility, substantivity of the dyes, and their stability in the dyebath.

Temperature generally determines the position of the dyeing equilibrium, which at room temperature is shifted strongly toward adsorption on the fiber. At higher temperature (usually 80 – 95 °C) the equilibrium is attained more rapidly. The temperature dependence of the maximum degree of exhaustion is dye specific. The same applies to hydrolytic stability, especially in high-temperature processes (see below) and in the presence of carriers in dyeing polyester – cotton mixtures [165].

Dyeing time is a critical factor for the evenness (levelness) of dyeing. The longer the dyeing time, the more uniform is the result. However, dyeing times should not be prolonged indefinitely because of limited dye stability and economic factors.

Water quality (metal-ion content) may be responsible for shifts in shade. Metal-complexing sequestrants are added to avoid this kind of effects.

Other auxiliary agents support the wettability and runnability of the material and improve the evenness of dyeing.

The *leveling capacity* (i.e., ability of dyes to compensate for the unevenness of dyeing) is an important characteristic that can be determined empirically. Unevenness results if dye adsorption occurs to quickly, and can frequently be traced to mistakes made in the addition of dye or salt. The leveling of uneven dye distribution proceeds via continuous desorption of the dye from the fiber and readsorption on a position on the fiber where the dye concentration is lower (migration).

In combination dyeing dyes with similar properties are preferred. To help in the selection of combinable dyes, the Society of Dyers and Colourists has classified direct dyes according to their leveling capacity [166, 167].

4.2.5. Dyeing Techniques

Exhaustion Process. In batch dyeing, the preferable mode depends on the type of dyeing equipment, the type of material to be dyed, and the solubility and affinity of the dyes. The dye is made into a paste with small quantity of warm water, diluted with more hot water, and boiled if necessary. It is then filtered and added to the dyebath. The electrolyte is added (if necessary in portions) either during heating of the bath or after attaining the optimum dyeing temperature. The material is dyed for 30 to 45 min and redyed, if necessary, at somewhat lower temperature for a short time. After the dyebath has been drained, the fabric is quickly washed clear with cold water and generally subjected to aftertreatment (see Section 4.2.7).

High-Temperature Dyeing Process. With suitable dyes, dyeing can also be performed in a closed apparatus at up to ca. 130 °C (especially for polyester – cotton mixtures). As a result of the high rate of diffusion, high levelness is achieved in short dyeing times, even in the case of poorly penetrable materials. After the high-temperature phase, the dyebath is cooled to 80 – 90 °C and the dye continues to adsorb, resulting in the same depth of color as in normal dyeing processes.

Pad Processes. In these continuous or semicontinuous processes, the fabric is first soaked by passing it through a trough filled with dye solution. Excess liquor is subsequently removed by squeezing the fabric between rubberized rolls (padding). The substantive adsorption of dyes during padding is avoided by moving the material rapidly, preferably at ambient temperatures, and employing minimum amounts of pad liquor. Optimal distribution of the dye is later obtained by either thermal treatment or longer storage.

In *thermal treatment*, the material is passed through a steamer in 1 – 3 min without intermediate drying (pad steam process). Alternatively, it is heated to 80 – 85 °C and rolled up in a closed chamber where it is kept for 2 – 8 h (*pad roll process*).

In the *cold pad-batch process*, the padded material is rolled without heating and stored at room temperature for 8 – 24 h.

The *pad jig process* represents the easiest type of fixation of pad dyeings. The material padded with dye is passed through a salt liquor in a jigger at 80 – 90 °C, thus preventing desorption of the dye. In this process, additional dye can be used for shading (see Section 2.2.5.2).

All the processes mentioned above are followed by quick rinsing with cold water.

4.2.6. Special Processes [[161] and references cited therein]

An established process is the simultaneous bleaching and dyeing, permitting cost reduction by combining two treatment steps. However, this process can be performed only with selected direct dyes. Dyeing processes with additional energy fields (e.g., ultrasound, magnetic fields, radioactive irradiation, or high-temperature treatment) or with redox systems (persulfate or periodate) have not yet proved useful in practice.

4.2.7. Aftertreatment [161, 168]

As a rule, the wetfastness of direct dyes does not meet the demands of daily use. It can be improved by aftertreatment, giving fabrics that are fairly fast to water and perspiration and, in ideal cases, washable up to 60 °C. The improvement in fastnesses is achieved by reducing the solubility of the anionic dyes. This occurs by enlargement of the molecules via the formation of saltlike compounds with cationic aftertreatment agents. The fixation of the saltlike compound can be enhanced by cross-linking or by formation of covalent bonds to the fiber (→ Textile Auxiliaries, Chap. 5.10.). For certain azo dyes, substantial improvement in lightfastness can be achieved by metal-complex formation (→ Metal-Complex Dyes, Chap. 5.1.2.). A combination of these aftertreatments produces especially good fastnesses.

Not only formaldehyde, one of the aforementioned cross-linking agents, but also metal complexes are discussed intensively because of the

ecological aspects of these compounds. The improvement in fastness by means of diazotization and coupling has decreased in importance because of the resulting changes in shade.

4.3. Dyeing with Anthraquinone Vat Dyes

Vat dyes are water-insoluble, organic pigments that are used to dye cotton and other cellulose fibers. The principle of vat dyeing is based on chemical reduction of these dyes to the leuco compounds (which are soluble in aqueous alkali and exhibit fiber affinity), followed by reoxidation to the water-insoluble starting dye within the fiber.

Approximately 10 – 15 % of cellulose fibers or cellulose-containing fiber blends are dyed with vat dyes. The total consumption of vat dyes is about 22 000 t/a (1993) of commercial products, without indigo.

4.3.1. Chemistry of Vat Dyes

Two main groups of vat dyes are distinguished chemically:

1) *Indigoid vat dyes* are derivatives of indigo and thioindigo (→ Indigo and Indigoid Colorants). Leucoindigo compounds have a comparatively low affinity for fibers. Thus, these dyes are used mainly in textile printing rather than dyeing (see Section 4.4).
2) The leuco compounds of *anthraquinoid dyes* have a high affinity for fiber and give especially dyeings that are excellently fast in use and processing.

Vat dyes are dye preparations that consist basically of vattable colored pigment and dispersing agent. *Vattable pigments* are polycyclic quinoid compounds that contain two or more carbonyl groups in a closed system of conjugated double bonds (→ Anthraquinone Dyes and Intermediates) [169].

The pigment must be as finely dispersed as possible to achieve a satisfactory vatting rate and uniform dye application in pigmenting processes. For this reason, after the actual synthesis the dyes are wet ground in the presence of selected dispersing agents. The color strength and shade are then adjusted according to the standards of the intended commercial products (→ Dyes, General Survey, Chap. 1.3.2.).

The addition of *dispersing agent* stabilizes the fine dispersion of pigment in the aqueous liquor. Since soluble dispersing agents are discharged entirely into the wastewater, their biodegradability or eliminability in sewage treatment plants is of increasing importance.

The dyeing behavior of vat dyes, i.e., their absorptive behavior and leveling capacity, is determined basically by the substantivity and diffusion of the leuco compounds. The chemical stability and insolubility in water of the dye pigment that is reoxidized in the fiber account for the generally unsurpassed light-, wet-, and weatherfastness of vat dyeings.

The dyeing process involves the following steps:

1) Vatting (reduction) of the dye
2) Absorption by the fiber
3) Oxidation in the fiber
4) Aftertreatment of the dyeing

4.3.2. Vatting

Using anthraquinone as an example shows that the dye molecule is converted to the sodium leuco form in alkaline solution by the gain of two electrons (Fig. 37).

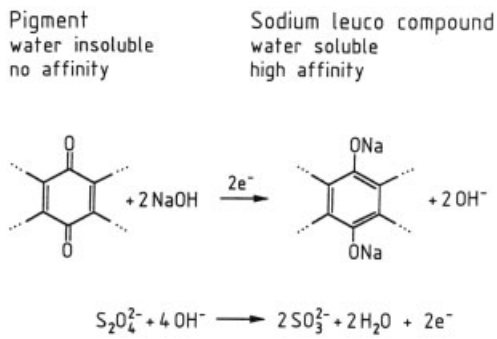

Figure 37. Conversion of an anthraquinone dye to its leuco form

For a long time, *sodium dithionite* (hydrosulfite) has been the most important reducing agent in vat dyeing. It is inexpensive and, when produced by modern methods, virtually free of

heavy metals (Zn, Hg). Its reduction potential (−970 mV) is high enough to vat all vat dyes quickly and completely. Studies of the reaction mechanism have shown that addition of the $S_2O_4^{2-}$ ion to the carbonyl group is the rate-limiting step in the overall reaction [170]. Formation of the leuco compound is achieved by the elimination of sulfite by the action of OH^-.

The rate of vatting depends not only on the concentration of dye and reducing agent but also on the crystal form, surface, and dispersion of the pigment (i.e., on its finish quality [171]). Leuco compounds are soluble in alkali. In the case of anthraquinoid vat dyes, the pH of the vat is about 13. At lower values the risk of vat acid sediments exists. Reduction is usually performed at 50–60 °C. At higher temperature, overreduction of certain dyes can occur (i.e., reductive destruction of the dye molecule).

To achieve faultless dyeings, the state of the vat should be monitored [i.e., determination of the concentration of reducing agent by using vat test paper (qualitative), by measuring the redox potential, or by titrimetric methods] because hydrosulfite is consumed by atmospheric oxygen [171]. For this reason, less reactive, more stable reducing agents are used in high-temperature dyeing [e.g., *derivatives of sulfoxylic acid* (Rongal HT, BASF)].

The oxidation products of sulfur-containing reducing agents are sulfite and sulfate. In most countries, the concentration of these substances in wastewater must not exceed maximum permissible values stipulated by law. *Thiourea dioxide* has a stronger reducing effect (−1100 mV) than hydrosulfite. Therefore, with sensitive dyes, a risk of overreduction exists. In addition, the oxidation products of thiourea dioxide contribute to the nitrogen and sulfur contamination of wastewater.

Organic reducing agents such as *hydroxyacetone* (Rongal 5242, BASF) do not contribute to the sulfur contamination of wastewater. However, the reducing effect of hydroxyacetone is much weaker (−810 mV), so it cannot replace hydrosulfite in all applications. For this reason, a combination of hydrosulfite – hydroxyacetone is recommended to lower the sulfur content in wastewater. The use of hydroxyacetone is restricted to closed systems because it forms strong-smelling condensation products in alkaline solution.

Systems for *electrochemical vatting* by means of electric current in the presence of a mediator, an *electron carrier* such as Fe^{2+}, or anthrahydroquinones have been described [172]. However, these methods have not yet been used in practice.

The vat dyeing system is also supported by auxiliary agents. *Dispersing agents* prevent the agglomeration of pigment particles or the flocculation of slightly soluble components such as vat acids or the impurities in cotton. If the presence of alkaline-earth ions is expected because of the use of hard water, the addition of *complexing agents* is required.

4.3.3. Dye Absorption in the Exhaustion Process [173, 174]

In the dissolved form, the vatted dye is present as the sodium leuco compound, either as a single molecule or as a complex of a few dye molecules [175].

In the exhaustion process, the higher the affinity, the more complete is the absorption of the dye by the fiber. The degree of absorption of dyes generally depends on dye concentration, liquor ratio, temperature, and possibly the electrolyte added [173]. Sodium salts of leuco vat dyes have a high affinity for the fiber; the exhaustion of the bath is generally between 80 and 90 % in exhaustion dyeing.

In the initial phase of dyeing, vatted dye is absorbed very quickly by the fiber. Extensive exhaustion of the bath is achieved after 10 to 15 min. The dye is then present in a high concentration in the outer regions of the fiber, but its distribution may still be entirely uneven (unlevel). In the subsequent leveling phase, the dye diffuses into the fiber and the fiber becomes dyed uniformly. The more evenly the dye absorbs to the fiber and the stronger the diffusion, the more level is the dyeing. Dyes that are less substantive absorb more slowly and evenly in the initial phase and exhibit more favorable diffusion properties. Diffusion can be improved by increasing the temperature. *Leveling agents* [176] can favorably influence absorptive behavior in the initial phase because they themselves possess a high affinity for the dye and release it in a controlled manner during the course of dyeing (→ Textile Auxiliaries, Chap. 5.6.1.) [171, 176].

However, a certain amount of dye is retained by the leveling agent, which results in slightly weaker shade.

4.3.4. Oxidation

After absorption by the fiber, the leuco dye is converted to the original pigment by oxidation and, in this way, fixed to the fiber. Oxidation can be achieved spontaneously with atmospheric oxygen or by the addition of oxidants, such as hydrogen peroxide, perborate, or 3-nitrobenzenesulfonic acid, to the liquor.

After the dyeing process, the material is usually rinsed with water to remove unfixed dye and decrease the alkalinity of the liquor. Depending on the oxidizing agent, oxidation is carried out at pH 9 – 12 and 50 – 60 °C.

The leuco dye present in wastewater is also converted to water-insoluble pigment, which can easily be separated mechanically or adsorbed to clarification sludge. Thus, colorization of the wastewater by vat dyes is negligible.

4.3.5. Aftertreatment ("Soaping")

After oxidation, all vat dyeings are subjected to heat treatment in the presence of water. As a rule, treatment is carried out in a weakly alkaline detergent liquor (soap was used in the past) at boiling temperature. The final shade is obtained by this treatment, and the more important fastness ratings (e.g., light- and chlorine-fastness) are improved substantially. Thus, this soap treatment is not restricted to the removal of pigment particles adhering to the surface but is an important individual step in vat dyeing. The processes occurring during soaping are unknown. Crystallization of amorphous dye particles or coarsening of primary particles by aggregation is assumed to occur [177]. If soaping takes too long, migration of the aggregates to the surface of the textile can cause deterioration of the rub-fastness.

4.3.6. Dyeing Techniques [178]

Vat dyeing can be carried out by using various techniques. Batch and continuous dyeing processes can be used. The dyeing properties of vat dyes enable the processes to be adapted to different machines.

Batch Processes. In batch processes, the dye must attach to the textile material from the dyebath. Therefore, the affinity and leveling-out properties of the dyes are of special importance. Dyeing is conducted in various circulatinggoods machines (e.g., jigger, winch beck, jet, and overflow) and circulating-liquor machines (yarn, piece, and package dyeing machines) (see Section 2.2.4).

Vat dyes are divided into groups according to their affinity for the fiber and the amount of alkali required for dyeing: The following abbreviations are used: IK (I = Indanthren, K = cold), IW (W = warm), and IN (N = normal). The IK dyes have a low affinity for the fiber; they are dyed at 20 – 30 °C and require little alkali. Salt is added to increase the absorption of these dyes. The IW dyes have a higher affinity than IK dyes. Dyeing is performed at 40 – 45 °C with more alkali and little or no salt. The IN dyes are highly substantive and applied at 60 °C. They require much alkali but no addition of salt.

Dyeing with vat dyes is performed according to the following processes:

Leuco Process. Auxiliaries, dye, sodium hydroxide solution, and reducing agent are added to the dyebath. The textile is introduced into the bath after a vatting time of 10 min. Depending on the depth of shade, dyeing is performed for 30 – 60 min and completed in the usual manner (rinsing, oxidation, soaping).

Preimpregnation Process. At first, the dye liquor contains only the dye as a water-insoluble dispersion. After the temperature is increased and, if necessary, sodium hydroxide solution and salt are added, the pigment partially deposits on the fiber. After the reducing agent is added, vatting occurs rapidly and the dye is absorbed by the fiber. As a result of partial deposition of the dye in the nonsubstantive form, the evenness (levelness) of dyeing is improved. Prepigmentation is conducted in a jigger and is often used in dye packages (cheese, warp, yarn, and piecedyeing beams).

Semipigmentation Process. Semipigmentation is used preferentially for dyeing lighter shades. Dye, sodium hydroxide solution, and reducing agent are added to the dyebath, and dyeing begins immediately at room temperature. Depending on the vatting rate of the dyes, only a small amount of substantive leuco dye is

present intially. For this reason, absorption proceeds more slowly and evenly. With increasing dyeing temperature, more and more dye is converted to the leuco form and can be absorbed by the fiber.

High-Temperature Process. The leveling-out properties of vat dyes can be improved by increasing dyeing temperature. In the high-temperature process, dyeing is conducted at 90–115 °C.

In bath dyeing, the substrate can be dyed in various processing forms. A distinction is made between loose material (flock, card sliver), yarn (cheese, hank), garments (socks, pullovers, T-shirts, etc.), and fabrics (wovens, knitwear) (see Chap. 3).

Continuous Processes. Vat dyes are especially suitable for continuous dyeing because of their clearly defined configuration during the dyeing process (pigment form, leuco form). Dyeing is performed according to the pad steam process or the wet steam process. The continuous process is used almost exclusively for dyeing woven fabrics and to only a small extent for knitwear.

In the *pad steam process*, which is most frequently applied, the textile is impregnated in a padder with the aqueous dye dispersion in the presence of an antimigration agent and a wetting agent, if required. The material is squeezed until it has a specified residual moisture content and then dried. Subsequently, the fabric is passed through a chemical padder, which contains the required amounts of alkali and reducing agent, and is fed immediately to a steamer. In the steamer, the dye is fixed for 45–60 s under saturated steam conditions at a wet-bulb temperature of 99 °C at sea level pressure. After steaming, the material is rinsed, oxidized, and soaped in an open-width washing machine.

To achieve maximum fixation yield and depth of shade, dyes must be very finely dispersed. However, fine dispersion also increases the migration of dye particles during intermediate drying. For this reason, antimigration agents are used (polyacrylates, alginates, etc.), which have an agglomerating effect on the pigment particles and thus restrict their mobility during the drying process [179].

The *wet-steam process* is used mainly for voluminous open fabrics with a high liquor-carrying capacity (e.g., toweling, cord, velvet). Unlike the pad steam process, this process does not require drying after dye application in the padder (energy savings).

The continuous dyeing process is especially advantageous economically in dyeing large batches.

4.4. Dyeing with Indigo (→ Indigo and Indigoid Colorants)

4.4.1. Chemistry and Historical Development

The chemical constitution of indigo was elucidated in 1883 by Adolf von Bayer. The first technically marketable syntheses were developed by Heumann and Pfleger.

As in the case of natural indigo, the dye is made by air oxidation of indoxyl (→ Indigo and Indigoid Colorants, Chap. 4.). Natural and synthetic indigo are absolutely identical chemically.

The demand for indigo depends greatly on fashion; itamounted to about 11 000 t worldwide in 1992.

The indigoid colorant thioindigo and its derivatives are also important vat dyes.

Because of the low affinity of vatted indigoid dyes for the fiber, these dyes are used mainly in textile printing. In comparison with other dyes, indigo occupies a special position because it is used almost exclusively for dyeing warp yarn in the production of blue denim. Being identical to the natural material, indigo has also been used in dyeing environmentally friendly clothing.

4.4.2. Dyeing Technique

Indigo is available as a powder, granulate, or liquid preparation. The dyeing technique has not changed over the years and, like vat dyeing, involves the following steps (see Section 4.3):

1) Dissolving of the dye by reduction in alkaline medium
2) Dyeing in the vat
3) Oxidation on exposure to air

The first step, dissolving of the dye, can now be omitted by using solutions of pre-reduced indigo

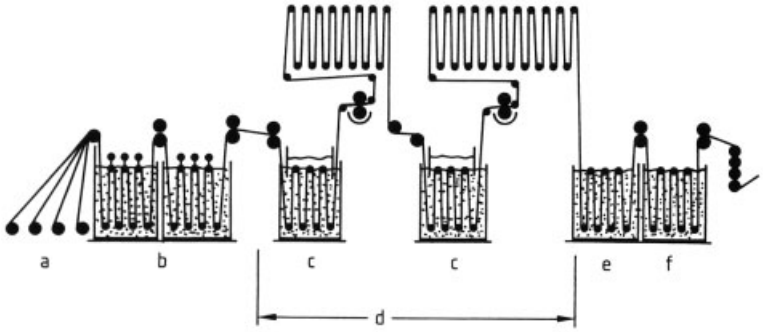

Figure 38. Schematic of an indigo continuous dyeing machine
a) Ball warps; b) Pretreatment phase; c) Dyebaths and oxidation units, up to five; d) Dyeing phase; e) Aftertreatment phase; f) Drying

(leucoindigo). In this way, about half of the reducing agent (depending on the type of machine) required for dyeing can be saved.

As a result of the relatively low affinity of indigo for the fiber, deep blue shades are attained only if dyeing and oxidation are repeated several times. Thus, in the past, indigo dyeing took several days.

In modern indigo dyeing plants for dyeing warp yarn, several dyebaths and oxidation units (usually up to five units) are arranged in a row for continuous operation [180–182] (see Fig. 38). For blue denim articles, pronounced ring dyeing of the yarn is desirable.

There are rope dyeing machines and slashers. In *rope dyeing machines*, 300 to 400 warp threads are combined into a rope and, in this form, passed through the dyeing system. *Slashers* were developed in the 1970s from sizing machines. In these machines, the warp threads are dyed side by side, parallel to each other [183]. Both types of machines have advantages and disadvantages and should be assessed according to the specific requirements (e.g., small or large batches, required fastness of dyeing).

4.5. Dyeing with Sulfur Dyes [184–186]

(→ Sulfur Dyes)

The dyeing of cellulosic fibers, especially cotton and its blends with synthetic fibers, is the main field of application of sulfur dyes. In addition, sulfur dyes such as Indocarbon are used in printing these fibers. Sulfur dyes are also used to a limited extent to dye polyamide fibers, silk, leather, paper, and wood.

Although more brilliant sulfur dyes (red and green) have been put on the market in the 1950s and 1960s, the strong point of these dyes still lies in the deeper, muted shades, such as black, dark blue, olive, brown, and green, where their favorable price has its full effect.

With respect to fastness, sulfur dyes are close to vat dyes but not equal to them. They have varying lightfastness that is generally good in the case of the blue, green, olive, and brown shades. The lightfastness in presence of perspiration of sulfur black is very good. The washfastness of sulfur dyes at 60 °C is good to very good, and their fastness to boiling water is moderate to good. Wetfastness can sometimes be improved substantially by aftertreatment of the dyeing with products of the Solidogen IH type. With a few exceptions, sulfur dyeings are not fast to chlorine. Therefore, faulty dyeing can be stripped by treatment with sodium hypochlorite bleach liquor.

A high-grade finishing operation can also improve the wetfastnesses. A decrease in lightfastness, as in the crease resist finishing of materials dyed with other dyes, does not occur; in fact, an improvement is frequently observed.

The main applications according to the textile are:

1) *Piece dyeing:* corduroy and velveteen, twill for working clothes, leisurewear, tarpaulin and rucksack fabrics, coat poplin, and lining materials. The dyeing of the cellulose part of blended fabrics containing polyester is becoming increasingly important.
2) *Yarn dyeing:* sewing thread, warp yarn for denim fabric, yarn for colored woven goods.

3) Dyeing of *flock, card sliver, roving,* etc., for blend yarns with wool and man-made fibers.

4.5.1. Types and Mode of Reaction [184, 187–190]

For technical and applicatory reasons, sulfur dyes are available in various modifications, which are classified in the Colour Index under the following general generic names:

Sulfur Dyes (C.I. Sulphur). Two types of sulfur dyes exist:

Amorphous powders, which are insoluble or partially soluble in water. They must be brought into solution by boiling with sodium sulfide and water, according to the following overall reaction scheme:

$$\underset{\text{Water-insoluble sulfur dye}}{DS-SD} + 2\,S^{2-} \rightleftharpoons \underset{\text{Reduced (leuco) dye}}{2\,DS^-} + S_2^{2-}$$

For the sake of simplicity, only one mercapto group has been assumed per dye molecule (D). In strongly caustic alkaline liquors, the following mode of reaction is possible:

$$4\,DS-SD + 2\,S^{2-} + 6\,OH^- \rightarrow 8\,DS^- + S_2O_3^{2-} + 3\,H_2O$$

According to this equation, only one-quarter the amount of reducing agent normally used is required.

With sulfur cooked dyes (i.e., quinonimine sulfur dyes, which are produced by water or solvent reflux thionation), the reducing agent can attack not only the disulfide groups but also the quinonimine group (as in the case of vat dyes).

Dispersible pigments, analogous to the Colloisol or Microdisperse grades of vat dyes, are used particularly for pad dyeing.

The dispersible, *partly reduced pigments* carrying the name "Sol" represent a special case. Apart from a dispersing agent, they contain a certain amount of reducing agent.

Leuco Sulfur Dyes (C.I. Leuco Sulphur). These ready-for-use, liquid dyes already contain the reducing agent required for dyeing. They must simply be diluted with water before application. Ecologically more favorable, low-sulfide types are now available.

Water-Soluble Sulfur Dyes (C.I. Solubilized Sulphur) [188]. Water-soluble sulfur dyes are available in the form of Bunte salts (*S*-arylthiosulfuric acid salts) of sulfur dyes and can be dissolved in hot water. The addition of alkali and reducing agent gives them an affinity for the fiber:

$$DSSO_3^- + S^{2-} \rightarrow DS^- + S_2O_3^{2-}$$

If NaSH or another reducing agent is used, alkali must be added to trap H^+ ions resulting from reduction. In the case of NaSH, H_2S formation is avoided in the following manner:

$$DSSO_3^- + SH^- + CO_3^{2-} \rightarrow DS^- + S_2O_3^{2-} + HCO_3^-$$

The leuco form DS^- is absorbed by the substrate and then oxidized to the fixed dye DS–SD. Since in reality several mercapto groups are present per dye molecule, oxidation causes not only a dimerization but also cross-linking of dye "monomers." For sulfur cooked dyes, the quinonimine group is also involved in the reaction. This is clear from the change in shade that occurs during reduction.

With respect to both application and fastness, a distinction is made between sulfur dyes and sulfur vat dyes. The latter include dyes of the type Hydron Blue (C.I. Vat Blue 43) and Indocarbon (C.I. Sulphur Black 11).

4.5.2. Additives to the Dye Bath

Reducing Agents. Ready-for-use sulfur dyes contain the reducing agent required for dyeing. Otherwise, the dye must be brought into solution by the addition of a reducing agent or the already water-soluble dye must be made to have an affinity for the fiber.

The reducing agents generally used are sodium sulfide or sodium hydrogensulfide in different forms in the following ratio (parts by weight):

100 parts of crystalline sodium sulfide ($Na_2S \cdot 9\,H_2O$; ca. 30–32 wt % Na_2S)
$\hat{=}$ 50 parts of concentrated sodium sulfide (ca. 60–62 wt % Na_2S)
$\hat{=}$ 65 parts of liquid sulfhydrate F 150 (31 wt % NaSH)

$\triangleq$ 25 parts of concentrated sodium sulfhydrate powder (95 wt % NaSH)

The amount of reducing agent required to dissolve or color the sulfur dye (Na_2S or NaSH) is given on sample cards issued by dye producers. In the dye liquor, the amount of concentrated Na_2S should be $\geq$2.5 g/L.

In aqueous solution, the sulfide ion is largely hydrolyzed according to

$$S^{2-} + H_2O \underset{10\%}{\overset{90\%}{\rightleftharpoons}} SH^- + OH^-$$

The resulting caustic alkaline product is not required at least for dyeing with water-soluble sulfur dyes.

If a combination of NaSH and soda is used, the lower alkalinity offers the following advantages:

1) The dyeing results in a better color yield and can be rinsed clear substantially faster
2) During dyeing and rinsing, the dyeing is much less sensitive to undesirable oxidation effects (bronzing, rubbing-fastness)
3) In the case of regenerated cellulose (viscose), embrittlement of the material is avoided

Sodium dithionite (hydrosulfite) is a proven reducing agent for *sulfur vat dyes*, which is used together with alkali – usually sodium hydroxide solution and occasionally soda.

The color yield of some sulfur dyes is remarkably increased by the addition of hydrosulfite. With blue grades in particular, the addition of hydrosulfite "sharpens" the dye liquor, preventing bronzing. Glucose is employed in one-phase printing with Indocarbon. As a result of possible "overreduction" or destruction of the dye chromophore, hydrosulfite and similar reducing agents can be used to only a limited extent and require a suitable selection of dye.

For the reduction of *water-soluble sulfur dyes,* the aforementioned reducing agents, glucose, and hydroxyacetone are used [191].

In the case of *water-soluble sulfur dyes* and *low-sulfide leuco sulfur dyes,* glucose-based combinations of reducing agents (binary reducing agent systems) have gained acceptance for ecological, technical, and applicatory reasons [192]. The following substances can be considered binary reducing agent components: hydrosulfite, formamidinesulfinic acid, and with limitations, hydroxyacetone. Mixtures of sodium hydroxide solution and soda are used as alkali. The aforementioned reducing agent combinations result in less pollution of wastewater. At the same time, glucose prevents or buffers possible overreduction or destruction of the dye chromophore by the reducing agent component with a higher reduction potential.

Alkali polysulfides are added to *stabilize* sulfur alkaline dye liquors against atmospheric oxygen. In jigger dyeing, dark and bronzed edges or seam marks can be avoided in this way. In the case of winch beck, jet, and continuous dyeing, undesirable oxidation during the absorption phase and rinsing is avoided, improving rubbing-fastness. Problem-free dyeing can be achieved with sulfur vat dyes of the type Hydron Blue by using alkali sulfides and alkali polysulfides, especially on the winch beck.

Wetting Agents. A wetting agent or padding auxiliary is usually added. Products based on sulfonated oleic acid amides give good results because they exert a certain leveling effect and confer softness on the material. In comparison, products based on ethoxylates are unsuitable because they have a retarding (equilibrium "dye dissolved in liquor: dye absorbed to fiber" is shifted toward "dye dissolved in liquor"), and sometimes precipitating, effect on the dye.

Water Softeners. To exclude the negative effects of alkaline-earth ions on dyeing, sequestrants based on ethylenediaminetetraacetic acid (EDTA) or polyphosphates are used. Additives of this type are required, especially in circulating-liquor dyeing with hardness-sensitive dyes, to prevent unevenness and the splitting of warp beams. Acidic treatment of the material to be dyed (1 – 5 g/L of HCl, HCOOH, or CH_3COOH) can often prevent the contamination of wastewater with sequestering agents.

4.5.3. Dyeing

Absorption of Dyes. In the exhaustion process (in a jigger, circulating-liquor machine or winch beck), the dye generally absorbs at 60 – 110 °C in 30 – 60 min and in the pad steam process, at ca. 102 – 105 °C in 30 – 60 s.

Dyeing with reducing agents is always performed according to the exhaustion process. Pad

dyeing processes, such as the pad steam and pad roll process, use a very short liquor ratio (ca. 0.7:1). The dye is basically fixed at the place where it has been padded. The fixing yield depends greatly on the liquor ratio. Optimal fixation and dye penetration are achieved with the pad roll process.

Oxidation of Dyes [184, 189, 191–193]. Sulfur colors are finally fixed on the substrate by oxidation. Before this step, the reducing agent, salts, and alkali, as well as the unfixed dye, must be removed by rinsing or washing to attain optimal rubbing-fastness. The following methods of oxidation are common.

Air Oxidation or Acid Treatment. The dyeings are subjected to cold and warm rinsing; the last rinse bath should generally be acidified with acetic acid (exception: sulfur black). In general, this process is used only for staple shades because after-oxidation usually occurs on storage, resulting in a change in shade.

Bichromate – Acetic Acid Oxidation. The material is treated with ca. 0.5 % potassium or sodium bichromate (based on the weight of the material) at pH 4 – 5 and ca. 70 – 80 °C. Until now, this process was common especially in piece dyeing. In the last years, however, it has been replaced increasingly by other types of oxidation for ecological reasons. In bichromate oxidation the material to be dyed becomes hydrophobic, evidently because of the formation of dye – chromium complexes. Although addition of copper sulfate (1 – 2 g/L $CuSO_4 \cdot 5\,H_2O$) to the oxidation bath improves the lightfastness of the colors, it is hardly ever applied. This type of oxidation cannot be used for articles to be rubberized.

Iodate Oxidation. Iodate oxidation [ca. 0.5 % potassium or sodium iodate (based on the weight of the material) at pH 4.5 – 5 (acetic acid) and >60 °C] is used only occasionally. In comparison with bichromate – acetic acid oxidation, somewhat brighter shades are obtained.

Bromate Oxidation. To attain an adequate rate of oxidation, the addition of vanadate (ca. 0.05 % $NaVO_3 \cdot 4\,H_2O$ with ca. 0.5 % $NaBrO_3$, based on the weight of material) is required. Oxidation proceeds at pH 4 – 4.5 (acetic acid) and 70 °C.

Hydrogen Peroxide Oxidation. With hydrogen peroxide oxidation, bright shades are obtained in the alkaline-to-neutral pH range. However, wetfastness is reduced somewhat compared with other oxidation processes because oxidation with hydrogen peroxide is not specific. Thus, especially if the oxidation is not performed properly [excessively high concentration of H_2O_2, excessively high temperature and pH (>50 °C, pH>7), and excessively long oxidation times], overoxidation occurs; i.e., the mercapto groups of the sulfur dyes are oxidized not only to disulfide groups, but also in part to sulfonic acid groups. Wetfastness can be greatly improved by aftertreatment of the dyeings with cationic auxiliaries, which are partly used in combination with oxidizing agents [194]. A deterioration of wetfastness can largely be avoided by oxidation in *acetic acid medium* (pH 4 – 4.5). Since oxidation then proceeds more slowly, the temperature must be increased to 70 °C. The resulting shades are much duller than those obtained by oxidation in the alkaline range. About 1 – 2 % (based on the weight of the material) of 35 % H_2O_2 is used.

Other oxidizing agents such as chlorite, bromite, chloramine T, and persulfate are not equally suitable for all dyes and consequently are seldom used. Sodium hypochlorite bleach liquor (hypochlorite) requires careful metering and can be used only with sulfur dyes that are relatively fast to chlorine.

Alkylation. With special alkylating agents, DS groups can be converted to the thioether DS – X – SD, where X is, e.g., – CHR – CHR –. In this way, the color is stabilized against oxidative and reductive effects [194]. Bi- or polyfunctional alkylating agents are especially effective because they lead to cross-linking and enlargement of the dye molecule. Solidogen IH is an appropriate product that at the same time exhibits substantivity due to its cationic character. About 1 – 2 %, based on the weight of the material, is used together with soda. In individual cases, the wetfastness of vat dyeing can be achieved by aftertreatment with alkylating agents.

In alkylation, care must be taken that only the mercapto groups and not the quinonimine groups of cooked dyes are alkylated; otherwise these compounds are fixed in their leuco form, which can lead to losses in coloring strength. Thus, in practice, dyeings are subjected to squeezing or suction before aftertreatment so

that the leuco vat form is oxidized by air before alkylation. Alternatively, the alkylating agent is used in combination with specifically acting oxidizing agents such as chlorite.

In the case of sulfur black dyeing, cationic alkylating agents can be used not only to improve wetfastness, but also to produce a buffering effect, whereby losses in tensile strength due to fiber damage can be avoided.

Rinsing or Washing. After oxidation or alkylation, the dyeing is rinsed or washed at 60–95 °C to remove excess oxidation or alkylating agent and, possibly, detached dye. A soaping process, as in the case of vat dyes, is not common because a decrease in rubbing-fastness occurs and recrystallization cannot take place with the cross-linked sulfur dye molecule.

A disadvantage of sulfur black dyeings (C.I. Sulphur Black 1) is the elimination of sulfuric acid on storage under unfavorable conditions (high temperature and higher humidity) [195]. A reaction of this type can even occur after a longer drying period. Therefore, to avoid tendering and losses in tensile strength, sulfur black dyeings must be buffered in the last rinsing bath with soda or sodium acetate. A rather washfast buffering can be achieved by using special cationic auxiliaries, such as dicyandiamide–formaldehyde condensation products and alkylating agents. An improvement can also be obtained by anticipating the elimination of sulfuric acid by means of oxidation with bichromate–acetic acid.

4.5.4. Dyeing Techniques [184, 186, 188–191, 196, 197]

Batch Processes (Exhaustion Processes). *Jigger Dyeing.* The dyebath is charged with all the chemicals, and the dissolved dye is added in two passages at the desired temperature. The salt is added after the third or fourth passage. After about six passages (or 60 min) and sampling, dyeing is completed by rinsing and oxidation.

Jigger Dyeing with Preimpregnation. In the case of water-soluble sulfur dyes, the material can be impregnated at boiling temperature with dye, salt, and wetting agent in the shortest possible liquor ratio. The volume of dyebath is then filled up with water, the reducing agent added, and the dyeing completed as usual. In comparison with the normal jigger process, better penetration is achieved.

Pad Jig Process. After the material is padded with dispersed or water-soluble dye and wetting agent (possibly drying, which is generally avoided because of the costs), it is placed in the liquor at the dyeing temperature. This liquor contains the reducing agent, salt, and some padding liquor. The dyeing is completed as usual.

Winch Beck Dyeing. The material is pretreated in a bath containing the chemicals and wetting agent at 40–50 °C. The dissolved, reduced dye is then added, the bath is brought to dyeing temperature, and salt is added. After 20–30 min, the material is rinsed clear by the influx of cold water and simultaneous overflow of the liquor. Dyeing is completed as usual.

Circulating-Liquor Dyeing. The material is often boiled or subjected to an acidic pretreatment to eliminate hardness-producing substances. The dyebath contains wetting agent, chemicals, and the dissolved, reduced dye. After 10 min of dye liquor circulation at 30–40 °C, the dyeing temperature is adjusted, and the process is terminated after about 30 min by rinsing clear and oxidation. The salt is added, in part, at the dyeing temperature.

Jet Dyeing [198]. Foam control is important in the case of semiflooded jets. After wetting the material (or predyeing the polyester part of polyester–cellulose fiber blends with disperse dyes), reducing agent (e.g., glucose–hydrosulfite) and chemicals are added at ca. 40 °C, the dissolved dye being added after about 10 min. If the jet contains a lot of air space compared to liquor, about 1.4 kg of concentrated hydrosulfite and 1.4 L of sodium hydroxide solution 38° Bé (ca. 32 wt %) must be added per cubic meter of air. After heating to dyeing temperature, more hydrosulfite is added, if necessary, and the material is dyed for 30 min. Dyeing is completed by rinsing clear, oxidation, and rerinsing.

Continuous Pad Dyeing Process. Ready-for-use liquid or water-soluble sulfur dyes and dispersible pigments are employed preferentially in continuous pad dyeing.

As a result of the (varying) substantivity of the dyes in pad processes with reducing agents,

tailing can occur. In this case a constant shade is attained only after a certain yardage. This problem can be avoided by taking into account both the feeding liquor factors of the individual dyes and the substrate factors. The recipe of the starting bath and of the feeding liquor can be calculated [199].

Pad Steam Process. The following variations in the pad steam process are possible: In the *one-bath procedure*, the material is impregnated with dye, reducing agent, and wetting agent at 20 – 30 °C and subsequently subjected to air-free steaming at 102 – 105 °C for 30 – 60 s. Rinsing is followed by oxidation and rerinsing in a washing machine. The material speed is 40 – 100 m/min. Warp for denim articles can be dyed in the same manner.

The *two-bath procedure* (for water-soluble sulfur dyes and dispersible pigments) is preferred especially for mercerized material to obtain a more even appearance. The material is padded at 20 – 30 °C with dye, wetting agent, and possibly antimigration agent. After intermediate drying, if necessary, it is impregnated with reducing agent, and the process is completed by steaming and aftertreatment, as in the one-bath process.

Pad Air-Passage Process. For light to moderate shades, the material can be impregnated with dye, reducing agent, and wetting agent at elevated temperature (70 – 80 °C). During air passage for ca. 1 min, sufficient fixation is achieved. Dyeing is completed in washing machines as usual.

The *hot pad dry process* is a variant of the pad air-passage process. After impregnation, the material is subjected to hot drying on a drying cylinder. For final fixation of the dye, the material is impregnated with a cationic alkylating agent and redried. This process is used especially in the dying of warp yarns.

Semicontinuous Padding Processes.

Pad Roll Process. After impregnation with the pad liquor, the material is batched up and stored in an air-free steam chamber (dry and wet temperature ca. 98 °C) where itis rotated slowly. After a rotation of about 1 h, dyeing is completed. By this procedure, the dye is almost completely fixed. Excellent penetration is obtained.

Special Dyeing Processes [184, 190, 200]. Dyeing with water-soluble sulfur dyes can be performed without reducing agents according to a pad thermofixing process. After padding and drying, the dye is fixed by heat treatment in the presence of acid-donating catalysts, urea, and thiourea. In this way, the water-solubilizing Bunte salt groups are split off. An excellent appearance is obtained in this process.

For denim articles, the warp yarns can be dyed with sulfur vat dyes or water-soluble sulfur dyes and sized in a continuous process. The warp is impregnated with dye, alkali, and reducing agent at 70 °C and subsequently dried and sized (mainly with potato starch). After weaving, the material is treated with a cationic alkylating agent and acetic acid in a padder and then rinsed. This process produces colored and black dyeings in which mainly the outer layer of the fiber is dyed. The stone wash effect can be imitated in this manner. A one-bath variant of this process gives poorer fastness properties, which correspond roughly to those of indigo dyeing [201].

4.5.5. Combination with Other Dye Groups

Sulfur dyes are combined with other dyes for reasons of shade, fastness, and in the case of polyamide – cellulose blends, tone-in-tone dyeing. Thus, combinations of sulfur vat dyes of the type Hydron Blue with dispersible vat dyes are generally used in the two-bath process. For stretch cord and similar articles made of polyamide – cellulose fibers, metal-complex dyes that are resistant to reducing agents are added to dye the polyamide fibers because only a few sulfur dyes stain polyamide fibers sufficiently compared to cotton fibers.

Water-soluble sulfur dyes can be combined with selected reactive dyes [202]. In this manner, more brilliant shades are obtained. Hydrosol Brilliant Red BCL, which is based on monochlorotriazine, is a reactive dye of this type.

In jigger dyeing, sulfur colors are usually brightened and shaded in a fresh bath with direct and basic dyes after oxidation.

4.5.6. Wastewater [203]

Normal wastewater from sulfur dyeing contains sulfides in addition to dye residues and salts. Since sulfides are toxic to aquatic organisms and form hydrogen sulfide with acids, they must be eliminated or their concentration reduced, at least. To avoid these problems, low-sulfide and sulfide-free dyeing methods have increasingly gained acceptance [192, 204]. Since the presence of chromium in water is restricted because it is ecologically harmful, oxidation with bichromate, which was used widely in the past, has been replaced in many cases by other types of oxidation. According to appendix 38 of the general wastewater regulations, the draft of which should come into force shortly, the presence of chromium(VI) compounds from the use of oxidizing agents for sulfur dyes is not permitted in Germany. In the United States no national standards for wastewaters from the textile industry exist. Minimum requirements are stipulated by the local authorities or the local public treatment plants. Limits for indirect discharge are generally 0.5 ppm for chromium and 5 ppm for sulfides [205]. In Japan the national wastewater standards can be adapted by the local prefectures to local requirements. Relevant values for Osaka are given as a typical example: total chromium 2 mg/L, chromium(VI) 0.5 mg/L (see Section 14.2).

Indirect Discharge. For indirect discharge, the conditions to be obeyed for components that are not dangerous depend on the municipality and the state. After appendix 38 of the general wastewater regulations comes into force, the requirements for dangerous components stipulated in these regulations must be fulfilled. This means that in textile finishing companies, sulfide values must be decreased to 1 mg/L, as in the case of direct discharge. Wastewater of this type can be treated without problems together with municipal wastewater in municipal sewage treatment plants. However, mixing with acidic wastewater and the resulting formation of hydrogen sulfide in the sewage system must be prevented. Contaminations of this type must be kept as low as possible because H_2S is oxidized microbially to sulfuric acid, which attacks concrete.

Wastewater from dyeworks can be collected in mixing and balancing tanks. Mixing with wastewater from a bleaching plant is of advantage. If possible, the wastewater should be oxidized by aeration or detention or by the addition of oxidizing agents, such as H_2O_2 or ozone.

Direct Discharge. For official requirements, toxicological, and ecological aspects, see Chapter 14.

In general, wastewater is purified in several stages. *The first stage* is a mixing and balancing tank with aeration, pH regulation, and cooling. During aeration, sulfide and polysulfide react to form thiosulfate. At the same time, the dye becomes insoluble and precipitates with polysulfide sulfur. Oxidation is accelerated by residues of sulfur dyes, which generally make the addition of catalysts such as Mn^{2+} unnecessary. By air or oxygen injection or by the action of air in trickle plants, the sulfides are oxidized in a few hours at room temperature in the presence of the dye residues.

The second stage involves chemical treatment by the addition of iron salts alone or together with $Ca(OH)_2$ and flocculation auxiliaries, if necessary. Sulfide and dye residues are eliminated by precipitation and sedimentation.

The third stage involves biological treatment, primarily using the activated sludge process with the addition of oxygen (usually air, seldom pure O_2). The activated sludge is separated in a secondary tank.

In the Katox process [204, 206], the main purification is performed in the Katox stage by means of accelerated oxidation in the presence of activated carbon. This is followed by a final purification, which involves chemical precipitation with iron sulfate, calcium hydroxide, and flocculation auxiliaries.

If no sulfides at all are allowed to be discharged into the wastewater, water-soluble sulfur dyes with sulfide-free reducing agents are an alternative [192].

4.6. Dyeing with Naphtol AS Dyes

An azo dye is formed from a Naphtol AS coupling component and a diazotized base (developing component) under suitable conditions (→ Azo Dyes, Chap. 6.1.1.2.). This coupling

can be carried out on the fiber itself. If components without solubilizing groups are used, the dye formed is an insoluble pigment, and washfast dyeing is achieved. This possibility was exploited quite early. A cotton fabric was first impregnated, for example, with a solution of sodium β-naphtholate and, after drying, was passed through a solution of diazotized *p*-nitroaniline to give an attractive deep red dyeing with marked fastness (Para Red). A large number of modified naphthols and suitable diazotizable bases (fast bases) were marketed by the then I.G. Farbenindustrie (now Hoechst and Bayer). By employing similar methods, these components can be used to produce dyeing and printing with excellent fastness in the most widely differing shades. The classic Turkey Red (Alizarin Red) has been replaced by the combinations Naphtol AS-TR/ Fast Red TR base and Naphtol AS-ITR/ Fast Red ITR base, in particular, which have good fastness. Besides bases that need to be diazotized before use in dyeings, diazotized diazonium compounds in stabilized form have also been marketed as fast color salts (→ Azo Dyes). Similar ranges later came on the market from most of the other large dye producers (→ Azo Dyes).

Nearly all cellulosic fibers can be dyed in every processing state by using Naphtol AS combinations. Other dyes can rarely give the same depth of color and shade. Handling these combinations is safe and relatively simple. Therefore, Naphtol AS dyes have been introduced into the textile industry on a wide basis. Colorations with outstanding fastness have been achieved, which in some cases are of the same standard as those using Indanthren.

Dyeing Process. In the *impregnating process* the cotton fiber is treated first with Naphtol AS, which is dissolved by boiling with caustic soda and a protective colloid (*hot solution process*) (for details of additives, see → Textile Auxiliaries). Most Naphtol AS types can also be solubilized readily with alcohol and caustic soda by using the *cold solution process.* The concentrated sodium naphtholate solution is added to the impregnating bath. For certain naphthols the addition of formaldehyde to stabilize the naphthol on the fiber (air stability) is also necessary; otherwise, coupling capacity is impaired. By addition of sodium chloride a rather substantive attachment of the naphthol to the fiber is achieved. On dyeing from a long liquor a lower temperature is used for impregnating the fabric, followed by rinsing and development. The use of a padding machine requires a higher temperature and, in general, intermediate drying.

Hank yarn is impregnated in a long liquor, for example, in a trough or on hank yarn dyeing machines. Wound yarn can be impregnated in the usual mechanical dyeing machines. Piece dyeing is carried out on the winch beck and the jigger or, preferably, continuously on the padding machine.

For *developing,* the base must first be diazotized in the cold, if necessary with the addition of ice, by using a solution of sodium nitrite in hydrochloric acid. Alternatively the base and the nitrite are made into a paste and stirred into the hydrochloric acid solution. To increase the coupling capacity, excess mineral acid must be neutralized or buffered by the addition of sodium acetate, disodium phosphate, or similar compounds. The textile impregnated with naphthol is passed through the cold developing bath containing the diazotized base so that the dye is formed rapidly. In dyeing machines the developing solution circulates through the stationary textile. Alkali carried over from the padding bath must be neutralized by the addition of an alkali-binding agent (e.g., acetic acid, aluminum sulfate) to the developing bath to prevent decomposition of the diazonium salt.

The use of naphthols and bases in solution was introduced in the 1970s. These formulations are safer and simpler to apply. Diazotization of the liquid bases generally does not require the addition of ice. After development, the textile is rinsed and subjected to careful *aftertreatment* in which excess color lake is removed from the fiber surface. Otherwise, it severely impairs the rubbing-fastness of the dyeing. In addition, through the change in pigment crystal structure occurring during aftertreatment, a stable final shade with enhanced fastness is achieved.

The possibility of producing fast, substantive, and bright dyeings on the textile from two different components has led to a large number of special processes and articles, characteristic of different geographical areas. Naphtol AS combinations that can be used at low temperature are extremely suitable for dyeing cellulose fiber materials that are partially coated with wax. Af-

ter dyeing the wax is washed out by boiling. This technique is still popular as batik, mainly in Southeast Asia.

The use of Naphtol AS dyes for the production of denim articles is of industrial and commercial interest. The cotton warps are simultaneously sized and padded with a coupling component. After weaving the impregnation is developed with a fast salt or a diazotized base and subjected to the usual aftertreatment. The process was introduced as the warp sizing padding process. Another variant is known as the warp sizing and dyeing process in which the warp is padded, dried, and developed immediately with a diazotized base.

4.7. Dyeing with Pigments

Being insoluble products, pigments can only be fixed to the surface of the fiber. For pigment coloring, very fine (particle size $\leq 0.5\,\mu m$) inorganic or organic pigments are used in a nonionic preparation (→ Pigments, Inorganic; → Pigments, Organic). They are fixed with the help of a binder. The binder is an aqueous dispersion of cross-linkable mixed polymers (copolymers or polymer blends, basis polyacrylate, polystyrene, polyurethane). A film, in which the pigment particles are embedded, is formed on the fiber. Cross-linking occurs during heating. The binder film must adhere firmly to the substrate without the material becoming too stiff or sticky. The type and amount of binder used depend on the amount of pigment and determine not only the feel, but also the rubfastness and wetfastness of the dyeing [207]. Usually, only light shades are produced by pigment dyeing. Common trade names for pigment – binder combinations are Acramin (Bayer) and Helizarin (BASF). Polyurethane-based binders (Perapret PU, BASF) form elastic films that can bind large amounts of pigment, making deep coloration possible.

Since pigments show no substantivity for textile fibers, piece goods are impregnated continuously with a liquor that contains the pigment, the binder, an antimigration agent, a cross-linking agent if necessary, an acid donor, and a softener. The pieces are dried at 90 – 120 °C and fixed at 160 – 180 °C. No afterwash is required. A resin finish can be carried out at the same time (e.g., a crease-resistant and antishrink finish, hydrophobing or a soil release finish).

Nowadays, a washout effect is produced with pigments in an exhaustion process. In this process, cotton is impregnated with a cationic product so that pigments can attach. A reduced amount of binder is applied and fixation is carried out at low temperature (120 – 130 °C). The disadvantage of the process is that a lot of pigment is lost, which contaminates the wastewater [208].

Pigment dyeing offers the advantages of universal applicability to all fibers (including glass fibers). It is a single-stage process, even for fiber mixtures (i.e., an energy-saving process). Aftertreatment is not required. Raw cotton is covered well. The resistance to light is very good; the disadvantages are limited depth of color, stiffer feel of the product, and limits to the rubfastness and wetfastness.

Pigment dyeing is used commonly for heavy textiles (e.g., canvas), light printing grounds, dress materials, shirting, bed linen, and furnishing articles.

4.8. Other Dyeing Methods

4.8.1. Dyeing with Leuco Esters of Vat Dyes

The leuco esters of vat dyes are anthraquinoid or indigoid vat dyes that have been made water soluble by reduction and esterification of the hydroxyl groups with sulfuric acid (→ Leuco Esters of Vat Dyes).

After application to the fiber, the ester is hydrolyzed, usually with sulfuric acid at room temperature or slightly elevated temperature (up to 70 °C), and the original vat dye is recovered by oxidation (e.g., with sodium nitrite).

The advantages of dyeing with the leuco esters of vat dyes are not only good levelness (evenness) and penetration but also the excellent fastness of vat dyes. The low affinity of the leuco ester for the cotton fiber and the relatively high dye costs are obstacles to wide application. These dyes are used mainly for dyeing high-quality articles of cellulose fibers in light colors and polyester – cellulose blends to moderate depths of color [209, 210], [211, p. 568].

Batch Dyeing. Because of the low substantivity of leuco esters of vat dyes for cellulose, batch processes (exhaustion processes) are rather unimportant. The usual systems are used, e.g., jigger and winch beck for piece goods and package dyeing machine for fabric and yarn. One- or two-bath processes exist. In the *two-bath mode* of dyeing, the dye is first allowed to strike (bottoming) in the presence of salt in a weakly alkaline medium (e.g., on the jigger). The liquor is then removed, and the color is developed with a liquor containing sulfuric acid and nitrite at room temperature. If high-temperature development is required (70 °C), nitrite is added to the bottoming bath. The *one-bath mode* is usually used in winch becks; i.e., the same bath is used for both bottoming and development. Packages are subjected to one-bath or two-bath dyeing, depending on the affinity of the dye.

Continuous Processes [212]. Continuous processes yield more favorable dye yields. The most important processes are:

1) Processes without intermediate drying: padding (dye, soda, nitrite), air passage, development (dipping at 70 °C in acid), air passage, and aftertreatment
2) Processes with intermediate drying: padding (dye, soda, nitrite), drying, development (dipping at 70 °C in acid), air passage, and aftertreatment (this process gives a higher fixing yield than the first method)
3) Pad steam development process without intermediate drying: padding (dye, soda, nitrite), steaming (1 min, saturated steam, 103 °C), development (dipping at 70 °C in acid), air passage, and aftertreatment

As a result of their high price, low substantivity, and toxicological problems during production, the importance of the leuco esters of vat dyes is decreasing. Remaining producers and trade names include Hoechst (Anthrasol Blue IBC), Ciech (Helasol), and Mitsui (Mikethren). Dyeing with leuco esters of vat dyes can be replaced easily by pigment coloring, by reactive or vat dyes.

4.8.2. Dyeing with Mordant Dyes [211, p. 578], [213, p. 649]

The fiber is first treated with metallic salts (i.e., mordanted). Highly adhesive, basic metal compounds are formed on the fiber. These compounds are capable of producing poorly soluble colored complexes (lakes) with certain azo and anthraquinone derivatives. Alizarin (1,2-dihydroxyanthraquinone) is the best-known anthraquinone derivative. It used to be isolated from the root of the madder plant but has now been replaced by the synthetic product. Suitable azo dyes contain, e.g., hydroxyl or carboxyl groups in the *o*-position to the azo group on one or both of the aromatic nuclei. The shade of the dyeing depends on the type of metallic mordant used. Alizarin and aluminum – calcium salts produce the well-known Turkey red.

The mode of operation in alizarin dyeing is relatively tedious. The old Turkey red dyeing process required at least 10 working steps: boiling, treatment with rancid olive oil, leaching, treatment with a tanning agent, mordanting with alumina, dyeing in hard water, reviving, acidification, and soaping. In the new process, treatment with Turkey red oil (sulforicinoleic acid) has replaced the oiling step. The light- and wetfastnesses obtained are excellent.

In the past, this method was used to dye, e.g., cambric and bunting. Today, easier methods using developing dyes and vat dyes have replaced this process.

4.8.3. Dyeing with Acid Dyes [211, p. 580]

Acid dyes (i.e., azo or anthraquinone dyes that are made water soluble by sulfonic acid groups) were occasionally used for dyeing cotton; however, they are no longer important. They do not have sufficient substantivity for cellulose. Only those dyes are usable that can form a metal complex on the fiber when applied together with a metallic salt. Excess heavy metal contaminates wastewater.

Acid dyes can attach directly to vegetable hard fibers (jute, sisal) by salt formation because the fiber companion substances contain basic groups.

4.8.4. Dyeing with Basic Dyes [211, p. 580], [213, p. 548]

Basic dyes (→ Cationic Dyes) show no substantivity for cellulose. Therefore, pretreatment of the fiber with tannic acid (which contains phenolic OH groups) is required. The tannin mordant is insolubilized with antimony salts (tartar emetic). Synthetic products (Katanol type) are also suitable. Thus, the fiber is dyeable in a weakly acidic medium, a saltlike bond being formed with the acidic phenolic hydroxyl groups. Basic dyes attach to bast fibers without a mordant. Suitable dyes are to be found in the azo, diphenylmethane, and triphenylmethane series, and among thiazine, azine, oxazine, thiazole, and quinoline derivatives. Bright colors are obtained at low cost with rhodamine, auramine, fuchsin, or methylene blue, among others. Since the fastness, especially lightfastness, is poor, this method is of little significance.

4.8.5. Dyeing with Mineral Dyes [211, p. 584], [213, p. 520], [214, p. 395]

In a pad process, inorganic metallic salts can be applied to cotton and then converted in an alkaline medium to the corresponding oxide by treatment with steam. Brown and khaki shades (mineral khaki) are obtained with chromium and iron salts. This dyeing is inexpensive, lightfast, and weather- and rotproof. The disadvantages are wastewater contamination and definite hardening of the fabric. This method is still important in some countries for dyeing tarpaulin and uniform materials.

4.8.6. Dyeing with Oxidation Dyes [213, pp. 437 – 439]

In the application of oxidation dyes, aromatic amines form insoluble polyazine derivatives in the fiber (→ Azine Dyes). Synthesis proceeds in several steps in a hydrochloric acid medium by oxidation (e.g., with dichromate). For the chief representative of this group, aniline black, the chromophore consists of dibenzopyran rings. Oxidation dyes are rapidly decreasing in importance because aniline and other aromatic amines as well as the bichromate used for oxidation are toxicologically hazardous. The colors produced by aniline black are characterized by a full bluish black shade and excellent fastness. Since they are easily reserved, they are still used occasionally for printing grounds. For individual processes, see [215].

4.8.7. Dyeing with Phthalogen Dyes

As with oxidation dyes, the actual dye in the phthalogen process (Bayer) – the insoluble phthalocyanine pigment – is formed in the fiber itself (→ Phthalocyanines, Chap. 8.2.). Precursors are involved in dyeing. A distinction is made between two processes:

Low Molecular Phthalogen Developer. Aminoiminoindolenine or its derivatives are applied together with heavy-metal donors (preferentially Cu or Ni) by padding. The pigment is formed on heating (150 °C) in the presence of weak reducing agents (e.g., glycols). This is followed by aftertreatment with hydrochloric acid, with the addition of sodium nitrite to remove secondary products. The most important aminoiminoindolenine is: C.I. Ingrain Blue 2:1, 74160 (trade names, e.g., Phthalogen Brilliant Blue IF3GM, Bayer or Sumilogen Brilliant Blue, Sumitomo).

High Molecular Phthalogen Developer. Polyisoindolenines, which are complexes of heavy metal and indolenine, are developed with reducing agents in a wet treatment. The polyisoindolenine complexes (trade names, e.g., Phthalogen Brilliant Blue IF3GK, Cu complex or Phthalogen Turquoise IFBK, Ni complex) are fiber affinitive and can be applied by using the exhaustion process. Their substantivity is, however, low but can be increased by pretreating the cellulose with anionic products (e.g., Phthalofix FN). Reductive development is performed in an alkaline medium with hydrosulfite. For details, see [215, 216].

Dyeings made by phthalogen developers are characterized by high brilliance and excellent fastness. Thus, they are suitable for washfast and weatherproof articles. No danger of photochromism exists if a finishing operation is performed with synthetic resin products.

4.8.8. Dyeing with Coupling and DiazotizationDyes [211, pp. 547 – 548], [214, p. 286]

Direct dyes (see Section 4.2) that contain aromatic amino groups can be diazotized on the fiber after dyeing and then coupled with a "developer" – a phenol, naphthol, or aromatic amine. Wetfastness, in particular, is improved by enlargement of the molecule and, naturally, the shade changes as well (→ Azo Dyes, Chap. 4.2.3.).

Conversely, water-soluble, substantive azo dyes, which carry amino or hydroxyl groups capable of coupling, can also be used followed by aftertreatment with a diazonium compound (usually diazotized 4-nitroaniline as a stabilized diazonium salt). The resulting polyazo dye shows excellent wetfastness (see Section 4.6).

4.9. Other Plant Fibers (→ Cellulose, Chap. 2.2.- 2.6.)

Hard and bast fibers, not, like cotton, isolated from seed hairs but from parts of stems and leaves, can also be colored with direct or reactive dyes. The noncellulosic portion of these fibers is frequently so large that even simple acid and metal-complex dyes attach. Basic dyes can also be used to produce colors that are brilliant but have poor fastness [214, pp. 433 – 434], [217, p. 534]. Hard fibers, such as jute and coir (the fibrous outer layer of coconut), have a strong selfcolor that is not lightfast. For this reason, they affect the quality of all types of dyeing. Bleaching does not help because the fibers become yellow again when exposed to light.

5. Dyeing of Wool and Silk

5.1. Dyeing of Wool

The dyeing properties of wool (→ Wool) vary depending on the origin, breed, age, food, season, and habitat of the sheep. Even fleece from a single sheep reveals pronounced differences in quality. In fact, differences are also found within a single fiber due to biological and environmental factors (tippy dyeing). By means of the specialized sorting of wool provenances and fleece components, batches are put together that largely correspond with each other with regard to fineness and to processing and application properties.

Apart from shorn wool, certain amounts of hide and slipe wool from slaughtered animals are available on the market. While hide wool is equal to shorn wool, slipe wool exhibits alkali damage.

5.1.1. Principles (see also, → Dyes, General Survey, Chap. 1.4.1.)

Wool fiber is composed of protein filaments and consists mainly of keratin with a very complex structure. The amino groups of keratin are very important for the dyeing process. The number of basic groups titratable with acid is 850 µmol per gram of wool fiber. In the acidic and neutral range, carboxyl groups are present largely in the undissociated state.

The number of ionic groups in the fiber depends on the pH of the medium. Keratin is most stable when the number of negative and positive ions is identical. This *isoionic region* differs somewhat from the isoelectric region of charge neutrality. This means that charges of adsorbed ions add to the ion charges of the fiber. According to Elöd, the *isoelectric point* of wool is pH 4.9 at room temperature [218].

The *morphological structure* of the fiber determines the pathway that dyes take during dyeing and is of decisive importance for the rate and extent of dye uptake. In the past, the dye was assumed to have to penetrate the scaly layer on the fiber surface. The epicuticle and the exocuticle were regarded as barriers because of their hydrophobic character. This idea was supported by the fact that oxidative or chlorinating degradation of epicuticle and exocuticle considerably facilitates dye uptake. Oxidative or chlorinating degradation is still used in the preparation of wool materials to be printed.

In contrast, Leeder [219, 220] and Leaver [221] showed that the penetration of dyes into the fiber proceeds through the intercellular regions. The cell membrane complex is regarded as a "solvent" for hydrophobic textile chemicals [222]. The strong swelling capacity of the intercellular cement is important for the dyeing

process [223–225]. Only subsequently are the sulfur-rich keratins also penetrated by the dye molecules thus, determining the position of the dyeing equilibrium.

Bonding Forces between Dye and Fiber. *Ionic Bonds.* Dye anions can participate in ionic interactions with fibers that possess cationic groups. However, the formation of ionic bonds is not sufficient to explain dye binding because compounds that can dissociate are cleaved in the presence of water.

Secondary bonds (dispersion and polar bonds, hydrogen bridges) are additionally formed between dye and fiber. Close proximity between the two is a prerequisite for bond formation. However, this is counteracted by the hydration spheres of the dye and of wool keratin. On approach, these spheres are disturbed, especially at higher temperature, and common hydration spheres are formed ("iceberg structures") [226]. The entropy of the water molecules involved is increased in this process (hydrophobic bonding).

Dye binding can be regarded as an ion exchange in which the ionic bond is supported by secondary bonds between the dye and the polymer [227]. Zollinger termed it an "Einweisungsfunktion" that is based on the far-reaching potential of ionic groups ("pilot ions").

Coordinate and covalent bonds can be superposed on secondary and ionic bonds. This aspect is discussed in connection with chrome, metal-complex, and reactive dyes.

5.1.2. Dye Classes and Dyeing Processes

5.1.2.1. Acid Dyes

Acid dyes consist of simple chromophoric systems (→ Azo Dyes, Chap. 3.2., → Anthraquinone Dyes and Intermediates, Chap. 3.2.), which are made water soluble by the introduction of sulfonic acid groups. Dissociation gives rise to dye anions that interact with the ammonium groups of the fiber. These groups are formed from amino groups in the presence of acid, which explains the name of this class of dyes.

From a coloristic standpoint, acid dyes are subdivided according to their affinity. Affinity increases from *leveling dyes* and weakly acidic absorbing (moderately leveling) *milling dyes* to neutral absorbing (poorly leveling) *super-milling dyes* (→ Azo Dyes, Chap. 3.2.). This sequence approximates the increase in size of the dye molecule. The presence of aliphatic groups in the dye molecule contributes to a substantial increase in binding to wool, converting leveling dyes to types that are fast to fulling. The sulfonic acid groups determine not only the number of possible ionic bonds to the fiber, but also hydration, which counteracts binding.

Dyes with similar absorptive behavior should be selected for combined application. The combination values for acid dyes established originally for polyamide were later transferred to wool [228–230].

In dyeing, *salt additives* exhibit a retarding and leveling effect. Higher concentrations of sulfate ions are assumed to compete with dye anions for ammonium groups of the fiber. Weakening of the electrostatic attraction between dye and fiber due to coulombic forces is more likely; this effect decreases with increasing pH. Instead, salt additives exert an aggregating effect on dye molecules, and increased attachment can occur [231].

Dyeing Processes. Dyeing instructions for different types of acid wool dyes vary, especially with regard to the pH range used (see Table 7). The greater the affinity for the fiber, the more strongly must the ionic binding component be repressed [i.e., the pH value at the beginning of dyeing must be higher (see Table 7)].

The starting temperature is 60 °C for leveling dyes, 50 °C for milling dyes, and 30 °C for super-milling dyes. Usually, the acid is allowed to react with the fiber for 10 min; then salt is added followed, after another 10 min, by the dissolved dye. After another 10 min, the system is heated for 30 – 45 min until a final temperature of 95 °C is reached and then further dyed at 95 °C for 45 – 90 min. Addition of acid toward the end of the dyeing process completes bath exhaustion. Warm and cold rinsing follows. With selected dyes dyeing can be performed at 80 °C. Milling dyes require the addition of a leveling agent (→ Textile Auxiliaries, Chap. 5.6.2.).

Trichromic dyeing can be conducted with leveling dyes. In the case of milling dyes, the

Table 7. Dyeing instructions for acid wool dyes

Addition	Leveling dyes	Milling dyes	Super-milling dyes
Na_2SO_4, cryst.	2 – 4 g/L	2 – 4 g/L	2 – 4 g/L
Acid donor	1 – 2 g/L H_2SO_4 (96 %)*	1 – 3 g/L acetic acid (30 %)	1 – 3 g/L $(NH_4)_2SO_4$
pH value	2.5 – 3	4 – 5	6 – 7
Leveling agent		0.5 g/L	0.5 g/L
Final addition		1 g/L H_2SO_4*	1 – 2 g/L acetic acid (30 %)

* Or formic acid.

desired shade is better achieved by selection of a dye with similar hue and shading.

Improvement in Fastness. The wetfastnesses of chlorinated wool that has been finished with synthetic resin (polyamide – epichlorohydrin or polyurethane) (superwash wool) can be increased by means of methylol amide compounds. Both fastness and antifelting finish can be improved through the application of a polyquaternary compound (i.e., a compound with several quaternary groups) [232] [Basolan F (BASF), Sandofix L (Sandoz)]. Anionic condensation products can form a barriere at the surface of the fiber and thus dimish the bleeding of anionic dyes (see Section 6.4.2) [Mesitol HWS (Bayer)].

5.1.2.2. Chrome Dyes

Chrome dyes (→ Azo Dyes, Chap. 3.2.1.3.) refer to selected acid dyes that form complexes with chromium ions. During complex formation a strong bathochromic shift of shade occurs. As a result of the superposition of several excited states, a marked dulling of the hue is observed.

Complex formation occurs during the dyeing process in a strongly acidic medium under the participation of electron donors (ligands) from the chromophore and the fiber [233, 234]. Chromium, the central atom of the complex, acts as a link between dye and fiber. This results in a very strong bond, which is reflected in the excellent fastnesses obtained. The binding of chromium occurs by substitution of the H in −COOH or −OH groups and via lone electron pairs from>CO, $-NH_2$ or −N=N− groups (dative bonds). Thus, the dyes must contain suitable functional groups:

1) Monofunctional: salicylic acid or alizarin type
2) Bifunctional: *o,o′*-dihydroxyazo groups

Chromium(III) with the coordination number 6 acts as the central atom. It is formed from dichromate, which is reduced by the fiber. Strong acids have an activating effect on this process [235]; the reducing effect of wool can also be enhanced by organic acids (tartaric, lactic, or formic acid) [236]. Thiosulfate also acts as a reducing agent, increasing the rate and degree of conversion in chroming [237] and decreasing fiber damage [236]. Lowering the temperature from boiling temperature to 90 °C contributes to fiber protection [238]. In the past, the amount of dichromate required was 50 %, based on the dye used, but no less than 0.25 % or more than 2.5 %, based on the fiber. According to Ciba – Geigy, the amount of potassium dichromate used is a standard $0.2 + 0.15 \times$ the amount of dye in percent. In the meantime, considerable reductions of potassium dichromate content have been found to be possible [239, 240]. Individual chrome factors are given in pattern cards.

Auxiliary agents cause the disaggregation of chrome dyes and form adducts with them. These adducts break down only at boiling temperature. In this way, the number of free amino groups available for complex formation is decreased and less acid is required. The dyeing becomes more level, and "tippy dyeing" is reduced. Ethoxylated fatty alcohols, alkylphenols, and fatty amines are suitable auxiliaries.

Application.

Afterchroming Method (*Chrome-Developing Dyes*). For dyeing the liquor is prepared with 1 – 2 g/L of formic acid (50 %) (pH 3.5 – 3.8), 0 – 2 g/L of calcined sodium sulfate, and 1 – 2 g/L of wool protectant. The process is started at 40 °C, and dissolved dye is added after 10 min. The system is heated within 30 min, and dyeing is performed at 90 °C for 30 – 45 min. If the exhaustion of the bath is too low, 0.5 – 1.0 g/L of

formic acid is added and dyeing is continued for 15 min.

For subsequent *chroming* the bath is cooled to 70 °C, followed by the addition of potassium dichromate and heating. Temperature is maintained at 90 – 100 °C for 30 – 45 min. After 15 min, 5 – 10 g/L of sodium sulfate is added to detach the saltlike bound chromate from the wool and make it accessible to complex formation (Bayer). Other variants use reducing and complexing agents to support the chemical reactions [241]. Finally, neutralization is carried out with ammonia (pH 8).

One-Bath Chroming Method (Metachrome Process). Complex formation must be preceded by dye diffusion. For this reason, the release of Cr(VI) must be delayed.

The dye liquor is prepared with 3 – 8 % of metachrome mordant (a mixture of sodium chromate and ammonium sulfate) and 2.5 – 5 g/L of crystalline sodium sulfate. After a prerun of 10 min at 40 – 50 °C, the system is heated slowly and dyed at boiling temperature for 45 – 90 min. Shortly before dyeing is completed, the bath is exhausted by the addition of 1 – 2 g/L of 30 % acetic acid.

The boiling time is shortest in the after-chroming method, and the most level and fast colors are produced. Poor shading possibilities represent a disadvantage. The one-bath chroming method is most important.

The following auxiliaries are used: Albegal W (Ciba – Geigy), Avolan AV (Bayer), Lyogen MS, WD, Lyocol CR (Sandoz), Syntegal V7 (Ostacolor), Uniperol O (BASF).

5.1.2.3. 1:1 Metal-Complex Dyes

(→ Metal-Complex Dyes, Chap. 3., → Metal-Complex Dyes, Chap. 5.1.1.)

Metal-complex dyes are chemically very similar to chrome dyes. The risk of fiber damage is reduced because the complex is formed in dye production.

Dyeing is conducted in a strong sulfuric acid medium at pH 1.9 – 2.2. Amino groups in the fiber are converted to the ammonium form, and ionic bonds to the dye anion are formed. Under these conditions, amino groups are not available as ligands. Only with increasing pH during rinsing can they be included in the complex in exchange for aquo ligands [242]. The addition of auxiliaries (alkanolethoxylates) allows a decrease in the amount of acid added (pH 2.5 – 3) because they form addition compounds with the dye and thus repress complex formation. The dyeing process is slowed down and made uniform. With synergistic amphoteric mixtures of auxiliary agents, dyeing can be performed at even higher pH (3.5 – 4) [243].

Dyeing Process. Depending on the liquor ratio, the bath is adjusted with 2 – 6 g/L of sulfuric acid (96 %) to pH 1.9 – 2.2 (pH 2.5 in the presence of 1 – 2 g/L of auxiliary agents). After addition of 5 – 10 g/L of calcined sodium sulfate, the material is treated with the liquor for 10 min at 40 – 50 °C; then the dissolved dye is added. Ten minutes later, the system is heated up in 30 – 45 min and dyeing is carried out at boiling temperature for 90 min; 1 – 2 mL/L of ammonia (25 %) or 2 – 3 g/L of sodium acetate can be added to the last rinsing bath. A lowering of the temperature to 80 °C is possible in the presence of an ethoxylated fatty amine (pH 1.9 – 2.2).

Dye Ranges. Chromolan (Ostacolor), Inochrom (Zeneca), Neolan (Ciba – Geigy), Palatinecht (BASF).

Auxiliaries. Albegal NF, Albegal Plus (Ciba – Geigy); Avolan S, SCN (Bayer); Lyogen WD (Sandoz); Syntegal V7 (Ostacolor); Uniperol O and SE (BASF).

5.1.2.4. 1:2 Metal-Complex Dyes

(→ Metal-Complex Dyes, Chap. 3., → Metal-Complex Dyes, Chap. 5.1.1.)

Metal complexes are formed in a molar ratio of 1:2 in the weakly acidic pH range [i.e., two chromophores are coordinated to one central atom (Cr or Co)] [244]. This atom is located between the two chromophores, which are usually arranged perpendicular to each other (Drew – Pfitzner complexes, → Metal-Complex Dyes, Chap. 2.2.). The high affinity for the fiber is based on the large size of the dye molecules; their compact, often almost spherical shape; and their negative charge [245].

To achieve sufficient evenness (levelness), the incorporation of sulfonic acid groups into the dye molecules was initially avoided so additional ionic interactions between fiber and dye could be excluded. Instead, water solubility was achieved by incorporation of methylsulfone ($-SO_2-CH_3$) [246] or sulfonamide ($-SO_2-NH_2$) groups.

Only after 1960 did dyes containing sulfonic acid groups also find application. They have the advantages of inexpensive production, higher yield, and cold solubility. In general, the wetfastness is somewhat higher than that of dyes containing methylsulfone and sulfonamide groups. Unfortunately, dyes containing two sulfonic acid groups, in particular, are susceptible to nonlevel dyeing. To avoid tippy dyeing, auxiliary agents, especially ethoxylated fatty amines, are added [247], which form adducts with the dye. These adducts break down at higher temperature. The leveling effect is increased by addition of Glauber's salt.

Since the 1:2 dye complexes are formed in a weakly acidic medium, they can be applied to wool under the same conditions. This enables a gentle mode of dyeing, which implies neither an oxidative attack by Cr(VI) nor a hydrolytic attack by sulfuric acid.

Dyeing Process. Ammonium sulfate (1 – 2 g/L) or ammonium acetate (2 – 4 g/L) (pH 5.5) is added to the liquor. After a prerun of 10 min at 30 – 50 °C, dissolved dye is added. The system is heated in 30 – 60 min and dyed for 30 – 60 min at boiling temperature.

In the presence of auxiliaries (1 – 2 g/L) dyeing is performed with the addition of 1 – 3 g/L of acetic acid (30 %) (pH 4.5 – 5) and 1 – 2 g/L of calcined sodium sulfate. Acid dosing using control and monitoring systems is also possible [248, 249]. After rinsing, 1 – 2 g/L of formic acid is used for acidification to improve the feel and wetfastness.

Auxiliary Agents. Albegal A, SET, SW (Ciba – Geigy); Avolan IL, IS, IWN, UL75 (Bayer); Eganal SZ (Hoechst); Lyogen FN, MS (Sandoz); Remol U (Hoechst); Uniperol SE, W (BASF); Unisol WL (Zeneca); Wofalansalz EM (Wolfen).

5.1.2.5. Reactive Dyes (→ Reactive Dyes

Reactive dyes for wool produce brilliant colors with good fastness. They are not identical to reactive dyes for cellulose fibers because the reactivity of the amino groups in the wool is considerably higher than that of the hydroxyl groups in cellulose. To achieve level dyeing on wool, the reactivity of the dye must be reduced and an auxiliary agent added.

Reactive anchor groups	Dye ranges
N-Methyltaurine ethylsulfone-	Hostalan E, Procilan E
β-Sulfatoethylsulfone-	Remalan
Acrylamide-, chloroacetyl-	Procilan
α-Bromoacrylamide-	Lanasol, selected Lanaset
2,4-Difluoro-5-chloropyrimidyl-	Drimalan F

Wool contains many reactive groups – amino, imino, and hydroxyl groups being the most important [250]. Reactions occur in a weakly acidic medium (pH 3 – 5) and include nucleophilic substitution of leaving groups (usually Cl, F, and rarely, sulfonate or ammonium groups) or addition reactions to polarized aliphatic double bonds (see Section 4.1).

The first two anchor groups mentioned (*N*-methyltaurine ethylsulfone and β-sulfatoethylsulfone) have the advantage that their functional groups are masked at the start of the process; consequently, no premature reaction occurs below boiling temperature [251]. At the same time, level dyeing is made possible by the increased solubility. Drimalan and Lanasol dyes are bifunctional and can have a cross-linking effect on wool [226, 252].

Influence of pH. In the acidic range, ionic bonds are formed between dye and fiber, with the dye still being capable of migration. At pH 5, covalent binding to the fiber predominates.

Dyeing Process. Auxiliary agent (1 – 2 g/L) is added to the liquor, and the pH adjusted to 3 – 4 with formic acid or 1 – 3 g/L of acetic acid. The process is started at 40 °C, and dissolved dye is added after 10 min. After 20 – 30 min, the pH is adjusted to 5 – 6 with sodium dihydrogenphosphate. Dyeing is conducted at boiling temperature for 1 h.

To eliminate hydrolyzed dye, an aftertreatment is performed for 15 min at 80 °C with the addition of 1.5 g/L of ammonia (pH 8.5 – 9.0). The last rinsing bath is weakly acidified.

Auxiliary Agents. Albegal B (Ciba–Geigy), Avolan REN (Bayer), Eganal GES (Hoechst), Lyogen FN (Sandoz).

5.1.2.6. Vat Dyes, Leuco Esters of Vat Dyes

In the past, vat dyes played an important role in dyeing wool. Indigo (→ Indigo and Indigo Colorants) was considered to be "king of dyes" because of its unsurpassed fastness on wool. Until the mid-1950s, navy cloth was dyed with indigo. The application was problematic with regard to the negative effect of reducing agents and alkali on wool. Today, indigo, thioindigo, and their derivatives have been replaced by classes of dyes that are easier to handle.

5.1.3. Technology of Dyeing (see Chap. 2)

Wool is dyed in equal amounts in the form of flock and slubbing. There is a trend toward piece and yarn dyeing [253]. The overflow can take the place of winch and beam dyeing machine, and jigger if the material has previously been made shrink proof [254]. High-quality, cool wool articles that are woven from fine yarns require especially gentle treatment.

Wool dyeing is usually conducted near boiling temperature. The fiber suffers clear degradation at this temperature. In some cases, the temperature can be lowered to 80–90 °C. An increase in dyeing temperature has proved necessary when poorly leveling dyes are applied. However, 108 °C is generally accepted as the upper temperature limit. A further increase in temperature can be required only in dyeing polyester–wool mixtures by the high-temperature process (see Section 11.2). In this case, the upper limit is 120 °C. Formaldehyde is added to the liquor, which leads to formation of methylene bridges in the fiber and stabilizes the keratin to hydrolytic degradation [255, 256].

The fields of application of various classes of wool dyes are based on the absorptive and leveling capacity [257]:

<1/1 standard depth	disulfonic acid dyes
1/1–3/1 standard depth	monosulfonic acid dyes
>3/1 standard depth	1:2 metal-complex dyes

Continuous processes are frequently used to dye slubbing. Here, a thickening agent based on etherified guar or locust bean gum, a special auxiliary agent, and an acid or acid donor are added to the pad liquor. Chromium trifluoride is generally used as the chroming agent. The material is steamed for 15–60 min after padding.

5.1.4. Properties of Dyeings

Among the aforementioned classes of dyes, reactive and 1:2 metal-complex dyes lead with respect to tonnage, followed by 1:1 metal-complex, chrome, and acid dyes [258]. This distribution among the different classes of dyes depends not only on the required level of fastness, but also on the necessary brilliance. Brilliant shades are produced by acid and reactive dyes, whereas chrome dyes dominate in the case of black.

Among the chrome and metal-complex dyes, 1:2 metal-complex dyes are easiest to use, and corrections of shade (shading) are possible. Superwash articles require a high level of fastness, which can be achieved with reactive dyes. The fastness of dyeings with various classes of wool dyes at 1/1 standard depth of shade is given in Table 8.

Excellent fastness to milling and potting is achieved with chrome and metal-complex dyes.

Damage of the wool during dyeing is due to action of acid, oxidizing agent (chromate), water, heat, and mechanical stress. Exposure to boiling water for a longer period results in cleavage of cystine cross-links. New bridges are formed, such as lysinoalanine, which can contribute to crease fixation. Thus, woolen fabrics should be setted before dyeing. Hydrolytic degradation increases with increasing distance from the isoelectric point [259]. The sodium sulfate present increases hydrolysis, which is closely connected to a deterioration of mechanical properties [260]. The 1:2 metal-complex dyes are considered most favorable with respect to maintenance of feel, elasticity, and strength.

Chlorinated wool exhibits a different coloristic behavior. Elimination of the scaly layer facilitates both dye uptake and release. The wet-fastness is much lower, so reactive dyes should be used. In contrast, antifelting treatments with

Table 8. Fastness of wool dyeings at 1/1 standard depth of shade

Class of dye	Washfastness (40 °C)	Fastness to perspiration (acidic)	Fastness to perspiration (alkaline)	Lightfastness
Acid dyes	4–5	2–4	2–4	3–6
Milling dyes	4–5	4–5	4–5	5–6
Super-milling dyes	4–5	4–5	2–5	4–6
Chrome dyes	4–5 *	4–5	4–5	6–8
1:1 Metal-complex dyes	4–5 *	4–5	4–5	5–7
1:2 Metal-complex dyes	4–5 *	4–5	4–5	6–7
Reactive dyes	4–5 *	4–5	3–5	4–6

* At 60 °C.

synthetic resin seldom lead to changes in coloristic behavior.

5.1.5. Dyeing of Wool – Polyamide Blends

The strength of woolen articles is increased by the addition of 5–20 % of polyamide (PA) fibers. Stretch fabrics made of wool weft and PA textured yarn warp have gained importance as ski tricot. Both types of fiber are, in principle, dyeable with the same classes of dyes. However, PA is more accessible to the dye and is, consequently, dyed more deeply than wool in the case of light colors. Only when the limited ionic binding capacity of the NH_3^+ groups in PA is exhausted does the higher binding capacity of wool become noticeable so that, for deep shades, the wool component is darker. For light shades, the different affinity can be adjusted by using fiber-affinitive retarding agents, such as aromatic sulfonates, which partly reserve the PA component. The level of auxiliary agents depends on the depth of color and is 0.5–4 %. Monosulfonated acid dyes are the most suitable.

Dyeing Process. Dyeing is performed in the presence of 1–2 g/L of acetic acid (60 %) and 2–6 g/L of Glauber's salt for 2 h at 98 °C. Fewer auxiliary agents are added for PA 66 than for PA 6 [261].

Dark colors cannot be achieved with acid dyes because of the limited dye affinity of the material. In these cases, 1:2 metal-complex dyes are used.

Auxiliary Agents. Basopal NA (BASF); Edolan A extra (Bayer); Effektan ST (Ostacolor); Endolan PAW (Hoechst); Erional NW, NWS, PW, RF (Ciba–Geigy); Intrasol CWN (Crompton); Mesitol HWS, liquid (Bayer, for 1:2 metal complex dyes).

5.2. Dyeing of Silk

Silk (→ Silk) accounts for 0.2 % (61 000 t in 1988–1989) of total fiber production. Nevertheless, because of its special properties, the importance of silk should not be underestimated, especially for ladies' wear and men's shirts, jackets, ties, and scarves.

5.2.1. Fiber Structure and Dyeing Behavior

Silk fibroin consists of 18 different amino acids. As with wool, amino groups are important for the uptake of ionic dyes. At 230 µmol per gram of fiber the number of amino groups is considerably lower than in wool.

The isoionic point of silk is at pH 5.0. The stability of silk is considerably lower than that of wool because silk has no cystine cross-links. Even under mild conditions, e.g., at pH 4.0 and 85 °C, hydrolytic degradation is observed. Therefore, acid must be added continuously during the dyeing process and not at once. For this reason, dyeing is frequently conducted at 70–80 °C, but no higher than 90 °C. In this way, damage to the surface structure and wrinkle and crease formation are minimized [262].

As a result of its fineness, light is strongly reflected from the surface of silk. Thus, considerably more dye is required to achieve the required shade than with other materials.

5.2.2. Classes of Dyes

Direct Dyes. As a result of their good fastness properties, direct dyes are frequently used on silk.

Dyeing Process. Dyeing is conducted in the neutral pH range or with the addition of 1–3 g/L of acetic acid (30 %) and 2–5 g/L of sodium sulfate. The process is started at 30–40 °C; the system is then heated up within 30–45 min and dyeing is continued at 90 °C for 30–45 min. As leveling agent, 1/3–1/5 of the degumming liquor (→ Silk, Chap. 4.1.), which has been made weakly acidic with acetic acid, can be added.

Acid Dyes. Acid dyes are the most important dyes for silk.

Dyeing Process. Dyeing is conducted with the addition of 1–4 g/L of acetic acid (30 %) or 1–2 g/L of ammonium sulfate (pH 4–5.5). After starting at 30–40 °C, the system is heated up within 30–45 min and dyeing is continued at 70–85 °C for 45–60 min. Dyeing in the soap bath used for degumming (pH 8–8.5) with the addition of sodium sulfate offers better fiber protection.

Metal-Complex Dyes. In principle, afterchroming dyes can be applied to silk [263]. They are, however, of little importance. In contrast to wool, *1:1 metal-complex dyes* can be dyed in a weakly acidic medium at 90 °C and produce excellent fastness values.

Best suited for the dyeing of silk are the *1:2 metal-complex dyes* because they can be applied in a weakly acidic medium and produce good fastness.

Dyeing Process. Dyeing is performed with the addition of 2–5 g/L of ammonium acetate or 2–3 g/L of ammonium sulfate and 0.5–1 % of leveling agent. After a prerun at 40 °C for 15 min, dissolved dye is added; the system is heated up within 30–40 min and dyeing is continued at 80–95 °C for 45–60 min.

Improvement in Fastness. Colors produced with acid, direct, and metal-complex dyes can be aftertreated with 8 % of tannic acid and 4 % of acetic acid (30 %) at 35–40 °C for 60 min. Subsequently, a fresh bath with 4 % of potassium antimony(III) oxide tartrate at 20–25 °C is used without intermediate rinsing.

Reactive Dyes. Reactive dyes are applied to silk only when brilliant shades are needed and when the colorfastness produced with acid dyes does not meet the necessary requirements [258, 264–267].

Dyeing Process. Dyeing is performed with the addition of 10–40 g/L of calcined sodium sulfate, half of which is added after 15 min and the other half after 30 min at 30 °C. The system is heated up to 50–70 °C in 30 min, and after another 15 min, 2 g/L of soda is added. Dyeing is continued for 40 min. An afterwash at 80 °C increases wetfastness.

Other Classes of Dyes. *Developing dyes* can be used for dyeing silk [268, 269]. *Vat* and *leuco esters of vat dyes*, as well as *cationic dyes*, are also used but are less important.

5.2.3. Technology of Dyeing

For the pretreatment of silk (steeping, soupling, degumming, hardening), see [267, 270].

The dyeing of yarn hanks is frequently conducted on spray or package dyeing machines. The winch, beam and overflow are preferred for fabrics (see Section 2.2.3).

Aftertreatment. To produce the typical scroopy feel of silk, weighting [269] and reviving are carried out with 1–2 g/L of formic, acetic, lactic, or citric acid.

6. Dyeing of Polyamide Fibers (see also, → Polyamides; → Fiber, 4.Synthetic Organic, Chap. 1.)

6.1. Chemical Structure

Polyamide 6 [*25038-54-4*] **and polyamide 66** [*9011-55-6*] contain, on average, 150 and 75 monomer units, respectively, linked together to give a macromolecule. One chain end consists of an amino group, which can be present in the free state or in the acylated form. Acylation results from reaction with chain transfer agents, which

can be added during polymer formation to obtain a uniform degree of polymerization (DP).

Amino groups are of special importance for dyeing because they form ammonium groups in an acidic dyebath by addition of protons. The number of amino groups is of the order of 25 µmol per gram of polymer in the case of PA 6 and 30 µmol per gram of polymer in the case of PA 66.

The reason for the limited dye uptake lies in the comparatively low number of amino groups in comparison to wool. The depth of color achieved on PA 6 is somewhat less than that on PA 66.

If the pH of the liquor falls below 2.4, more acid is taken up by the fiber because the amide groups are protonated under these conditions [271]:

$$-NH-CO- + H^+ \longrightarrow -NH_2^+-CO-$$

However, this is of no interest in dyeing because dye that is absorbed below pH 2.4 is less tightly bound, and fiber damage occurs at low pH.

Modified types of polyamide (→ Fibers, 4. Synthetic Organic, Chap. 1.4.) contain a varying number of amino groups [272]. The dyeing characteristics of PA fibers as a function of the content of amino groups (µmol per gram of polymer) are given below

Weak dyeing	15 – 20
Normal dyeing	35 – 45
Deep dyeing	60 – 70
Ultradeep dyeing	≥ 80

In addition, dyeability with cationic dyes is achieved by the incorporation of sulfonic acid groups (basic dyeable fibers). During finishing, a change in dye affinity can be attained by local reaction with fiber-reactive products, such as chlorotriazine derivatives to block part of the amino groups (Sandospace process, Sandoz).

6.2. Supermolecular Structure (see also, → Fibers, 2. Structure, Chap. 2.4.2. – 2.4.4.)

The state of order of the polymer in the fiber is of decisive importance for its dyeing behavior. The degree of crystallinity and degree of orientation are regarded as determining quantities. However, the amorphous remainder is of greater importance in dyeing because the chain segments in this part have a mobility that permits penetration of swell water and dye molecules. The system can be looked at as an uniaxial stretched dynamic network with the ordered regions as knot points [273].

The amorphous regions are present in the glassy state at lower temperature, and the mobility of chain segments begins only above the glass transition temperature T_g. In the presence of water, T_g is below the dyeing temperature for PA and, thus, has no effect on the course of dyeing.

The fiber exhibits certain structural differences in the radial direction (skin – core structure); the more compact surface layer in PA 66 obstructs dye uptake slightly [274–276]. Technological operations change the fibrous structure. Stretching leads to increased orientation and impeded dye uptake [274]. While classical drawn yarn has a monoclinic α-form, fast-spinning yarn has an unstable pseudohexagonal γ-form. However, this unstable form is converted to the more stable $\alpha - \gamma$ mixed structure during prewash [273, 277].

Thermal treatments result in changes in structure. Dyeing behavior is not changed essentially by *hydrosetting*, but the structure is loosened [278]. Thus, the size of the dye molecules has a smaller effect on dye uptake. Free shrinkage of the material results in a faster rate of dyeing [279]. In comparison, the critical range of absorption is shifted upward by 15 K on heat setting. The required penetration time is almost doubled. In *texturing*, the internal forces acting during torsion of the thread change the inner structure so greatly that dye molecules can diffuse more easily into the fiber [280]. Any unevenness in operations that changes the supermolecular structure of the polymer leads to disturbances of dye uptake ("stripiness" or "Barré effect"). In general, PA 6 dyes more evenly than PA 66 because it has a more open structure.

6.3. Interactions Between Dye and Fiber

The binding of dyes to polyamide is based on various interactions.

Secondary Bonds. The chromophoric system of the dye, with its aromatic rings (delocalized π-electron systems), easily polarizable azo groups (lone electron pairs), and other donor and acceptor groups, forms many polar bonds, dispersion interactions, and hydrogen bridges with the polymer.

These secondary bonds are superimposed on all other bonds. The larger the dye molecules and the fewer the hydrophilic groups carried by the dye, the greater is the superposition.

Ionic Bonds. Acid dyes are present in the fiber in ionic form [273]. ELÖD and coworkers were the first to detect a stoichiometric relationship between the content of amino end groups in the fiber and dye uptake [281]. This can be ascribed to electron neutrality. Dye anions are able to displace counterions of the ammonium groups in the fiber because they are supported by numerous secondary bonds (ion-exchange mechanism). In the presence of water, ionic bonds exert no binding effect but rather a "Einweisungsfunktion" (pilot function), according to ZOLLINGER [282, 283].

Additional dye uptake beyond the ionic binding capacity is possible only in the form of purely secondary adsorption ("overdyeing"). Dissociation and hydration counteract this effect. The dyeings obtained by overdyeing have poor wetfastness. The amount of these weakly bound dyes can be determined under conditions in which ionization of the amino groups does not occur ("neutral absorptive capacity" according to HOFFMANN).

The simultaneous action of ionic bonds and secondary bonds determines the degree of dye uptake in equilibrium. This is frequently reflected in a superposition of the Langmuir and Nernst isotherms [284, 285].

The amount of dye bound at equilibrium decreases with increasing number of sulfonic acid groups in the dye molecule due to increasing hydration. For this reason, mono- and polysulfonated acid dyes should not be used together in combination dyeing.

Coordinate and Covalent Bonds. Amino groups of the fiber can participate as ligands in the coordinate bonding of 1:1 metal-complex dyes. This is, however, of little practical importance. In addition, they can form covalent bonds to reactive dyes.

6.4. Dyes

6.4.1. Disperse Dyes (→ Disperse Dyes)

More than 50 % of disperse dyes are simple azo compounds, about 25 % are anthraquinones, and the rest are methine, nitro, and naphthoquinone dyes. They were originally developed for dyeing acetate fibers. With the emergence of polyamides, they also proved useful for polyamide fibers.

Disperse dyes are finely dispersed products that have a solubility in the bath of about 0.1 g/L. The solubilities of polyamide dyes are somewhat higher than those of polyester dyes. The shades of colors obtained with disperse dyes depend on the substrate. Thus, in comparison with acetate or poly(ethylene terephthalate) (PETP), many shades on PA are bathochromically shifted (orange → red, red → violet). This is attributed to interaction of the amide groups with the chromophores. The colors are generally vivid, except for the red shades.

As a result of the level dyeing of even structurally nonuniform material, disperse dyes are applied especially in dyeing lighter shades. Wetfastness deteriorates with increasing depth of color.

Dyeing Process. A dispersing agent (0.25 – 2 g/L) is added to the liquor, and the pH is adjusted to 5 with acetic acid. The process is started with the material at 40 °C and followed by addition of dye after 5 min. After another 5 min, the system is heated up within 30 min, and dyeing is conducted in 60 min near boiling temperature.

Dye Ranges. Artisil (Sandoz), Cibacet (Ciba – Geigy), Dispersol (Zeneca), Celliton (BASF), Ostacet (Ostacolor), Resolin P (Bayer), Samaron (Hoechst).

Auxiliary Agents. Eganal SZ (Hoechst), Levegal HTN (Bayer), Uniperol EL (BASF), Univadin DP (Ciba Geigy).

6.4.2. Acid Dyes

Acid dyes are derived from mono- and disazo compounds (→ Azo Dyes, Chap. 3.2.2.) (primarily yellow-red, dark blue, black) and from anthraquinones (→ Anthraquinone Dyes and Intermediates, Chap. 3.2.1.) (brilliant blue grades). Through the introduction of sulfonic acid groups, these compounds become water soluble and, at the same time, acquire the ability to form dye anions by dissociation in aqueous medium. The anions are capable of undergoing electrostatic interactions with cationic groups in the fiber. Thus, acid must be added to the dyeing process to convert amino end groups of the fiber to cationic groups, which explains the name of this class of dyes.

The binding of these dyes to the fiber is essentially supported by secondary valence forces. Affinity to the fiber can be increased by introducing additional substituents into the dye molecule that do not change the character of the chromophore [286]. The affinity increases along the series: olefinic < aliphatic < aromatic < cycloaliphatic.

Small chromophores (monoazo or anthraquinone) can be connected to one another via a binding link in such a way that no conjugation between the two can occur and – in contrast to direct dyes – the rotatability of the linked parts is maintained. So the affinity to the fiber is increased without changing the color. In larger molecules a similar effect is achieved by a steric hindrance of coplanarity within the chromophore system (benzidine dyes with sulfonic acid groups in *o,o′*-position).

The secondary binding of dye to PA is counteracted by hydratable sulfonic acid groups, which explains the large spectrum of dyeing properties. A classification according to technical application is given below:

1) *Monosulfonated Acid Dyes.* Leveling dyes and neutral absorbing dyes. These grades can easily be combined with each other.
2) *Disulfonated Acid Dyes.* These dyes produce dyeings with good wetfastness. They are used as self-colorants but are not recommended for trichrome dyeing [287].

The dyeing behavior of acid dyes can be described with the help of the combination values K (see Chap. 12), which are derived from combination dyeings at 70, 80, and 90 °C [288, 289]. In the case of matching shades from dyeing and subsequent dyeing in the same bath the combination values of the dyes involved are equated. Monosulfonated acid dyes lie in the range $K = 1 - 5$, and polysulfonated acid dyes have still higher values. The higher this number, the greater are the contrasts. The combination values of dyes from different dye manufacturers cannot be compared [287, 290, 291].

The pH of the liquor is important for level dyeing, for exhaustion of the bath, for wastewater contamination, and for the effort involved in rewashing after dyeing. With increasing dye affinity, the ionic interaction in the initial phase must be repressed to achieve uniform absorption, i.e., the liquor must be neutral at the start of the dyeing process. Only in the course of dyeing is the liquor slowly acidified by one of the following methods [292]:

1) Acid is added at the beginning of dyeing or continuously, preferably by using pH-controlling instruments: Telon ST process (Bayer), Dosacid process (Ciba – Geigy). Special care is required in the neutral range because small amounts of acid give large shifts in pH value.
2) Acid donors release acid during the dyeing process. Substances used are ammonium sulfate, which liberates NH_3 during dyeing; sodium pyrophosphate ($Na_4P_2O_7$), which is converted to the acidic disodium hydrogenphosphate [293]; esters of organic acids; and heterocyclic compounds, which slowly release organic acids when exposed to heat.

In the case of low-affinity acid dyes, addition of acid toward the end of dyeing is recommended for full exhaustion of the dyebath.

The classical temperature control is more favorable than pH control for regulation of the dyeing process [293]. In the case of pH control, absorption of ca. 60 % of the dye is quick and nonlevel, and only the last 50 – 20 % is absorbed in a controlled manner.

In the acidic range, *electrolytes* retard liquor exhaustion [294]. The competition of sulfate anions with the dye for dye-affinitive positions is responsible for this effect (see Section 5.1.2.1).

Auxiliary agents improve leveling. In the case of fiber-affinitive agents, the sum of the dye

anions and auxiliary agent anions should not exceed the binding capacity of the fiber to avoid blocking effects [295].

Aftertreatment. The wetfastness of dyeings produced with acid dyes is often unsatisfactory. In contrast to wool where the hydrophobic scaly layer hinders the escape of dye molecules from the fiber during washing, PA does not have a barrier of this kind. However, a barrier can be produced by aftertreatment.

In older processes, tannic acid (polygalloylglucose) and potassium antimony(III) oxide tartrate (tartar emetic) are used that form a 1:1 adduct on the material via hydrogen bridges [296]. This finishing process is expensive and said to be carcinogenic.

Synthanes. The synthetic tanning agents (synthanes) represent a more favorable alternative. They are high molecular mass condensation products of aromatic sulfonic acids with formaldehyde [297, 298] or condensation products of phenol, cresol, catechol, and naphthol with formaldehyde, which are made water soluble by reaction with bisulfite.

These products deposit on the fiber surface and, because of an accumulation of negatively charged sulfonate anions, form an electrostatic barrier that counteracts the release of dye anions from the fiber. In accord with this is the observation that a highly negative electrokinetic (zeta) potential is formed [299].

This treatment mainly increases the contact fastness (e.g., to water, perspiration, and seawater) of acid and metal-complex dyes and, to a lesser extent, their washfastness. The fastness to chlorinated water in swimming pools is hardly improved. Synthanes also have a stain-repelling effect on floor coverings [299].

Aftertreatment with synthanes has a number of disadvantages. Synthanes can reduce the adhesive strength of coatings or laminates [300]. In addition, they are poorly resistant to dry heat, afterfixing, and steam; treatment with synthanes results in slight changes in shade, reduced lightfastness, and a hardening of the feel. In the presence of cationic softeners, the fastness to rubbing is decreased. Synthanes are frequently applied despite their disadvantages.

The product Fadex CL (Sandoz) was developed specially to improve the fastness of bathing suits to chlorinated water. However, it cannot be used for printed articles with a white ground or bleached material because of its brown selfcolor.

Dyeing Processes.

Dyeing with Temperature Control. The pH of the dyebath is adjusted to about 4.5 – 7, and 0.2 – 2 g/L of dyeing auxiliary is added. The process is started cold with the material and allowed to prerun for 20 min. Dissolved dye is added; the system is heated up in 30 – 45 min and dyeing is continued for 20 min near boiling temperature.

Dyeing with pH Control. Dye is dissolved in the liquor, the pH is adjusted to 8.5 – 10 with ammonia, and the auxiliary agent is added as described above. The material is placed in the boiling liquor, and acid is added slowly. The final pH of 4.5 is attained after 45 – 60 min. (Alternatively, 0.25 – 1 g/L of a pH regulator can be added.)

Aftertreatment. The exhausted bath or fresh liquor is adjusted to pH 4.5 with formic acid, and 1 – 2 % (per weight of fiber) of synthane is added. Aftertreatment is started at 40 °C, the bath is heated to 70 – 80 °C, and the material is treated for 20 – 30 min; the material is then rinsed.

Auxiliary Agents for pH Control. Eulysin WP (BASF), Sandacid V, VS (Sandoz).

Synthanes. Cibatex PA (Ciba – Geigy), Mesitol NBS (Bayer), Matexil FA-SN (Zeneca), Nylofixan P (Sandoz).

6.4.3. 1:2 Metal-Complex Dyes

The classical 1:2 metal-complex dyes (→ Metal-Complex Dyes, Chap. 5.1.1.) are acid dyes that are coordinately bonded to transition metals, especially chromium and cobalt, in the molar ratio of 1 : 2. These dyes are negatively charged and can enter into ionic bonding with the fiber. The amino groups of the fiber are not included in the complex. In contrast to acid dyes, the water solubility of classical 1:2 metal-complex dyes is not based on the presence of sulfonic acid groups. Rather, it is due to sulfonamide and methylsulfone groups. Only after 1960 were dyes containing sulfonic groups put on the market (see Section 5.1.2.4). Their high affinity for the fiber is caused by secondary bonds with the large molecules that are frequently almost spherical.

As a result of their low water solubility, the traditional 1:2 metal-complex dyes are not as suitable for dyeing PA as for wool. Dyes containing sulfonic acid groups behave more favorably and, at the same time, produce better wetfastness. Their high fastness level makes them suitable for articles that are exposed to special use, such as floor coverings and car seat covers. The tendency of these dyes to mark structural differences in the material is a disadvantage. For this reason, they are recommended only for dark shades. The disulfonated acid dyes give good light- and wetfastness on PA and are frequently used in PA dyeing despite the poorer build-up properties. The high affinity of 1:2 metal-complex dyes makes the addition of amphoteric or nonionogenic leveling agents necessary.

The absorption of dyes by PA increases with decreasing pH. This tendency is especially pronounced in the case of disulfonic acid dyes. They are dyed in the presence of acetic acid. Dyeing with dyes free of sulfonic acid groups depends less on the charge of the fiber and more on the extent of formation of free complex acid, which absorbs more easily than the anion [301].

Dyeing Process. Dye and 0.5 – 1 g/L of auxiliary agent are added to the liquor. It is then made weakly acidic by the addition of 0.5 – 2 g/L of ammonium sulfate and 0.5 – 1 g/L of acetic acid (60 %), if necessary. High-affinity dyes can be used in a neutral or weakly alkaline medium. The process is started at 30 – 40 °C, followed by heating up within 30 – 60 min, and dyeing at boiling temperature for 30 – 60 min. A prerun with auxiliary agent at boiling temperature reduces the stripiness of the color.

6.4.4. Reactive Dyes

In principle, the reactive dyes used for wool are also suitable for PA (see Section 5.1.2.5). Covalent binding of dyes to the terminal amino groups produces dyeings that have excellent wet- and rubfastness (almost without exception grade 5). The lightfastness corresponds to that of analogous acidic dyes (3 – 6). The depths of color attainable are limited by the number of end groups and are higher for PA 6 than for PA 66.

Dyeing Process. The dyeing bath is made weakly acidic (pH 4.5 – 5). The process is started at 20 – 45 °C, followed by heating with a rate of 1 K/min and dyeing at near boiling temperature for 30 – 90 min. Aftertreatment (soaping) is performed with 0.5 g/L of a nonionic surfactant and 1 g/L of sodium bicarbonate or ammonia at 95 °C for 20 min.

6.5. Technology

Dyeing Material. The dye-binding capacity of a PA material can be characterized by comparison with a reference material. For this purpose, the ratio v of the dye uptake of the test (m_F) and the standard ($m_{F,\,standard}$) materials in the initial phase is determined (region of linear increase)

$$v = m_F / m_{F,standard}$$

Values between 0.2 and 5 are obtained.

A saturation value can be determined by a comparison of the dyeings at pH 4.7, the amount of dye taken up being given as the fiber sum number in S_F units (1 S_F = 25 mmol/kg) [291]. The values lie between 1 and 3. The exhausted bath can be used again, e.g., for a blue color by addition of a yellow dye and the shade compared with a chart [302].

Pretreatment (Washing, Prefixing). Before dyeing, fabrics generally must be prefixed to compensate for material-related differences in affinity and to reduce the sensitivity to creasing during the dyeing process. Prefixing can be performed as *thermofixing on the frame* [15 – 20 s at 190 °C (PA 6) or 200 – 230 °C (PA 66)] or as *hot-water fixing* (45 min at 130 °C).

Dyeing of Special Materials. The material to be dyed is available in various forms. Exceptional features in dyeing exist for the following articles.

Stockings, tights, and socks are dyed either in drum-dyeing machines or on automatic apparatus (for stockings). In these machines the stockings or tights are drawn over aluminum forms, sprayed with dye liquor, and fixed.

Belts for technical purposes (safety belts) can be dyed most efficiently according to the pad steam process in continuous systems.

Table 9. Fastness of polyamide dyeings at 1/1 standard depth of shade

Class of dye	Washfastness (40 °C)	Fastness to perspiration (acidic)	Fastness to perspiration (alkaline)	Lightfastness
Disperse dyes∗	4 – 5	4 – 5	4 – 5	4 – 5
Acid dyes	3 – 5	2 – 5		4 – 6
1:2 Metal-complex dyes	3 – 5 ∗∗	4 – 5	4 – 5	6 – 8
Reactive dyes	4 – 5	4 – 5	4 – 5	4 – 6

∗ At 1/6 standard depth of shade.
∗∗ At 60 °C.

Floor Coverings. In the piece finishing of carpets, stringent demands are made on the evenness of the shade. Dyeing on special, open-width winch becks (4 – 5-m width) is frequently performed at 90 °C according to the pH shift process, starting at pH 8. Within 30 – 45 min, the pH is lowered to 4.5 – 6.5, depending on the depth of color; pH regulators can also be used [294, 303].

In continuous dyeing of floor coverings thickening agents are added to the pad liquors to prevent dye migration. Application is carried out with the help of a minimal application process or by using foamed liquors. The liquor uptake is between 120 and 500 % [304, 305]. The next step involves steaming. Frothing auxiliaries are added to prevent dye from being washed out of the fiber tips by the formed condensate (frosting effect). Differential dyeing materials require a specific selection of dyes, auxiliaries, and pH values [306].

6.6. Properties of Dyed Materials

Dyeing with acid dyes can result in a certain degradation of the polymer [307–309]. The degree of hydrolysis depends on the concentration of the dye and the number of sulfonic acid groups per dye molecule [310].

In the case of crash and crinkle-look articles, thermal treatment under high-temperature conditions at 110 – 125 °C is frequently combined with dyeing in jet dyeing systems. In this process, loss of fiber strength can occur. This fiber damage can be largely avoided by adding 0.1 – 0.2 g/L of a hydroxylamine derivative [287].

The fastness of colors produced by various classes of dyes is shown in Table 9.

The values for PA 66 are slightly better than those for PA 6. Improvements can be obtained by aftertreatment with synthanes. The lightfastness of dyeings obtained with acid dyes on PA is slightly lower than on wool. Resistance to the chlorinated water of swimming pools plays a special role for bathing suits. PA 66 is usually superior to PA 6 in this respect.

6.7. Selection of Dyes

Very light shades are obtained on PA with *disperse dyes* because these dyes exhibit a good leveling capacity and are best able to compensate for material-dependent differences in affinity. *Acid dyes*, which also have a relatively good leveling capacity, are used for moderate depths of shade. Their fastness is usually sufficient, especially following aftertreatment with synthanes. The depth of shade has a limit of ca. 1/1 standard depth. The *1:2 metal-complex dyes*, which permit any depth of shade and give excellent fastness, are used for all dark and sober shades. They are not suitable for light shades because they have a tendency to nonlevel absorption and to mark structure-related stripiness. *Reactive dyes* are used for brilliant colors if the wetfastness of acid dyes is inadequate. They, too, have a tendency toward stripiness.

6.8. Toxicological and Physiological Aspects

Polyamide textiles often come in direct contact with the skin, especially underwear and stockings. For some time occasional allergic reactions have been observed to occur with close-fitting textiles because of their low moisture withdrawing capacity. However, only recently has the toxicological risk from the dyes employed received

Table 10. World polyester fiber production

	Amount, 10^6 t			Percentage (without polyolefin) of	
Year	Total	Staple fiber	Filament	Total fiber production	Synthetic fiber production
1970	1.6	0.9	0.7	8	33
1980	5.1	3.0	2.1	17	48
1990	8.7	4.7	3.9	23	58
1992	8.3			21	58

increased attention [311]. The amount of dye residues remaining on materials subjected to the all-in drum dyeing process, which is still occasionally used in the stocking industry, is much greater than after package dyeing with intermediate rinsing. Thus, in the first case, the potential risk of skin damage is somewhat increased but disappears after the material has been washed.

7. Dyeing of Polyester Fibers

For information on polyester (PES) fibers, see →Fibers, 1. Survey, Chap. 8. (properties); →Fibers, 2. Structure (form, structure); →Fibers, 3. General Production Technology; →Fibers, 4. Synthetic Organic, Chap. 2. (chemistry and production). For properties, see also [312].

PES fiber is quantitatively the most important synthetic fiber (see Table 10). Its inexpensive production from petrochemical raw materials and excellent textile properties alone and in combination with natural fibers guarantee PES fibers universal applicability. Starting with tire cord, technical fabrics for tarpaulin and seat belts, carpeting, and furnishing fabrics, uses also include clothing, especially mixed with wool for suiting and trouser materials and mixed with cotton for shirts, raincoats, trousers, and casual wear. Pure PES fibers, especially in textured form, are employed in the knitwear sector.

PES fibers are hydrophobic, thus water-soluble dyes do not attach. In contrast, PES fibers can be dyed easily with water-insoluble, small molecular dyes originally developed for dyeing cellulose acetate. Since the preferred dyeing medium is an aqueous liquor, the poorly water-soluble dyes must be dispersed before application (→Disperse Dyes).

7.1. General

7.1.1. Dyeing in Aqueous Liquor

Since the fundamental work of Meyer and Kartaschoff in 1925 [313] and the studies of Vickerstaff [314] in the beginning of the 1950s, a large number of studies have been published on the kinetics and thermodynamics of dyeing cellulose acetate and synthetic fibers with disperse dyes in aqueous dyebaths. For a survey, see [315, pp. 67, pp. 147]. Disperse dyes as well as acid and basic dyes must be molecularly dispersed in the dyebath; i.e., the dye must be dissolved in the aqueous bath before it can adsorb to the fiber surface and then diffuse into the fiber.

Thus, the form in which the disperse dye is present in the dye liquor is decisive for the dyeing process. The water solubility and rate of dissolution are influenced by many factors including [316–318]:

1) Temperature
2) Particle size
3) Crystal modification
4) Crystal growth
5) Melting behavior
6) Type of dispersion and dispersing agent used (fineness, stability, agglomeration, aggregation of dispersion)
7) Presence of salts
8) Behavior in dye mixtures

The water solubility of pure disperse dyes is a few milligrams per liter and increases strongly with temperature. It is also increased manifold by dispersing agents [316, 319].

The state of the dye in the fiber is often compared to a solid solution. Thus, the thermodynamic dyeing equilibrium of a disperse dye between water and fiber follows the Nernst distribution law:

$$\left(\frac{C_{\mathrm{F}}}{C_{\mathrm{L}}}\right)_{t=\infty} = K$$

At constant temperature, the distribution coefficient K [i.e., the ratio of the concentration of dye dissolved in the fiber (C_F) and in the liquor (C_L) at equilibrium] is constant.

At normal dyeing temperatures the rate of dissolution of dispersed dyes in the dye liquor is assumed to be high, so that C_L is constant as long as undissolved dye is present. Although not always valid in practice, the motion of the liquor is also assumed to be high enough so that molecularly dissolved dye is always transported to the fiber surface in sufficient amounts and the adsorption of dye proceeds rapidly. Then, diffusion in the fiber is the step that determines the rate of dyeing. The amount of dye g that diffuses in a certain time follows Fick's first law:

$$g = -\frac{D}{h} \cdot (C_L - C_F)$$

The diffusion coefficient D is 10^{-10} to 10^{-12} cm^2/s and can be increased only by an increase in temperature. The diffusion path h is predetermined by fiber geometry; the concentration gradient $C_L - C_F$, by C_L.

Many experiments were carried out to develop an equation for the kinetics of dyeing. However, only approximations result because of the complexity of the fiber state and of the processes occurring in the dye liquor. The following phenomena are of special interest:

1) The polyester fiber contains varying proportions of crystalline and amorphous regions, depending on the degree of drawing and the fixing state. Only regions that are amorphous above the glass transition temperature are accessible to dye diffusion. During dyeing, the amorphous portion can change as a function of stress and shrinkage. In addition, the dyeing medium (water) and the dye already diffused into the fiber change the fiber structure and, thus, the dyeability.
2) In the dyebath, the dye goes through a large number of stages, from solid particle to a single molecule, which interact with each other. These include:
   - Dissolving in the liquor
   - Possibly melting
   - Adsorption of dispersing agent
   - Association to dispersing agent
   - Incorporation into dispersing agent micelles
   - Diffusion through the laminar liquor layer on the fiber surface
   - Adsorption to the fiber
   - Agglomeration, aggregation, crystal growth
   - Mutual influencing of dyes, isomorphism of different dyes (formation of mixed crystals with different solution behavior than pure components)
   - Chemical changes (e.g., reductive destruction)
3) The levelness of dyeing – the basic requirement of coloring – is endangered by differences in temperature and concentration. In practice, such differences are unavoidable because of irregularities in the velocity and direction of the liquor flow at various positions in the dyeing aggregate and gradual depletion of dye in the liquor during passage through the dyebath. Hence, a constant supply of dye at the fiber surface is not guaranteed, and diffusion no longer determines the rate of dyeing.

For this reason, the practical dyeing process cannot (yet) be described by simple theoretical models. Despite many careful studies of model systems and individual processes, which permit a semiquantitative interpretation of individual steps under standardized conditions, dyeing results still cannot be mathematically predicted unambiguously. Thus, a dyeing process must be elaborated from dyeing experiments. Recipes are drawn up on the basis of trial dyeings with the help of color measurements.

When elaborating a dyeing process in practice, it is important to realize that, with regard to the rate of absorption, the system consisting of hydrophobic fiber and disperse dye occupies an intermediate position compared to other classes of dyes. The rate of diffusion of disperse dyes into hydrophobic fibers is moderately temperature dependent. Absorption can be easily regulated by controlling the temperature. Furthermore, this intermediate position between fast and slowly diffusing dyes is characterized by adequate migration capacity of the dyes, especially at high temperature. Thus, unlevel dyeings can be leveled out within manageable periods of time. Both methods (temperature control during exhaustion and high temperature leveling)

are basically available for dyeing with disperse dyes:

a) As a result of rapid absorption of the dyes, the dyeing at the end of the exhaust phase is uneven and becomes level through subsequent adequate migration at high temperature.
b) The dyeing at the end of the exhaust phase is largely level due to controlled absorption with a bath exhaustion rate that is below a previously determined limit. The subsequent high-temperature phase serves only to ensure good penetration.

Irrespective of the method selected, a series of basic rules with regard to the time requirement, safety, and quality can be deduced from knowledge of the optimization of the absorption phase (method b). These are:

1) In one formulation, dyes should be combined that, in the amounts applied, have exhaustion characteristics that are as similar as possible.
2) While the dyeing apparatus is being heated, the main exhaust range of the dyes, which covers about 30 °C, should be passed through sufficiently slowly.
3) The heating rate in the critical temperature range should be adjusted to the type and design of the textile and to liquor circulation or the circulation rate of the material [320, 321].

To follow the first of these rules, information about the combinability of dyes is obtained from dye suppliers. The fact that the temperature range in which the main amount of dye exhausts depends on the amount of dye must also be taken into account. Moreover, dye producers offer premixed disperse dyes whose components are selected in such a manner that the mixture absorbs evenly at relatively low temperature and, consequently, makes fewer demands on the dispersion stability (e.g., [322]). At the same time, the melting behavior of mixtures (usually melting point depression) should not be overlooked [323]. When using mixed dyes, differences in the fastness of the components must be accepted. Trade names of combination dyes include Foron RD (Sandoz), Resolin K (Bayer), Dianix UN-SA (Mitsubishi-Hoechst), and Sumikaron RPD (Sumitomo).

For the second and third rules, empirical values are used to elaborate apparatus-specific control programs that guarantee optimization of the dyeing process with regard to the time course and dyeing levelness. Balancing the critical dyeing rate of the apparatus system (V_{CRIT}) with the significant dyeing rate of the formulation (V_{SIG}) (see Section 2.1) is helpful. Programs of this type are marketed as fast dyeing processes [324].

The dyeing process can be also controlled with dyeing accelerants (carriers) (see Section 7.2). Carriers are aromatic compounds that exert a swelling effect on PES fibers and a dissolving effect on disperse dyes. In the presence of carriers, as in the case of special leveling agents developed for PES dyeing, theoretical elucidation of the dyeing process becomes even more complicated. Only models are available at present.

7.1.2. Thermosol Process

In dyeing PES fibers, the dyeing time can be shortened significantly by increasing dyeing temperature. The rate of dyeing increases exponentially above the glass transition temperature; a temperature increase of 10 °C doubles the rate. For this reason, at ca. 200 °C fast colored fibers can be obtained with disperse dyes within seconds. Especially for PES – cellulose mixtures, this property is used to dye PES fibers continuously in a pad-fix process, which was introduced by Du Pont as the thermosol process [315, pp. 191 – 202], [325, pp. 141 – 142] (see Section 11.1). Application of the dye from an aqueous dispersion is followed by drying and fixing.

During drying, the padded dye liquor on the fiber surface can migrate to positions with a higher rate of drying. This danger is counteracted with an antimigration aid (→ Textile Auxiliaries, Chap. 5.9.1.) [326]. To obtain a dye distribution on the fiber surface that is as homogeneous as possible, a uniform liquor film must be produced on the fiber. For this purpose, additives are used that wet the fiber and, at the same time, have a dissolving effect on dyes (e.g., ethoxylated fatty alcohols or fatty acids).

After drying, the dye is present on the fiber surface embedded in a film of dyeing aids and dispersing agents. In the case of PES – cellulose fiber blends, the dye is also present on the surface

of the cellulose fibers. At fixing temperatures of 210 °C, most disperse dyes have melted or are dissolved in the liquid film of auxiliary agents. Three processes overlap in actual dye fixation: (1) heating, (2) dissolving of the dye in the fiber surface layer, and (3) diffusion. Disperse dyes diffuse into the fiber only in the molecular form. Sorption on the fiber proceeds rapidly. Buildup of a high dye concentration on the fiber surface layer and diffusion are competing steps that depend on the prevailing temperature and determine the rate of dyeing.

In heat fixation of disperse dyes, suitable heat transfer media are hot air, superheated steam, contact heat, and radiant heat. Hot air and contact heat are used most commonly. Steam has a swelling effect on the fiber. If superheated steam is used for fixation, the temperature can be lowered in comparison with hot air, the fixing times being the same. This is an advantage especially with articles made of textured fibers.

In the case of PES – cellulose fiber blends, the largest part of the liquor is absorbed by the cellulose fiber during drying, because of its higher absorption capacity. Nevertheless, in the subsequent thermofixing step, disperse dyes are almost completely fixed to the PES fiber. Transfer from the cellulose to the PES fiber occurs both by migration via direct fiber contact points and through the gas phase. The type of transfer apparently depends on the sublimation behavior of the dyes. Dyes that are resistant to sublimation migrate by direct contact of the fibers, whereas those that are not prefer the gas phase. Transfer through the gas phase can lead to more level dyeing, but with easily subliming dyes a risk exists of contaminating dyeing equipment and incurring dye losses [327].

7.2. Dyeing Processes for Polyester Fibers

PES fibers are almost exclusively colored with disperse dyes. For some time, diazotization dyes were used for dark shades. They are coupled to the fiber with suitable components (e.g., 2-hydroxynaphthoic acid). Although fast colors are obtained, the dyeing process is time consuming and susceptible to failure, and dyeing reproducibility is poor. Vat dyes are also occasionally used for PES dyeing. Small molecular thioindigo derivatives, in particular, diffuse into PES fibers if they are applied according to the thermosol process.

7.2.1. Requirements to Be Met by Dyes and Auxiliaries

A large variety of disperse dyes are available for coloring PES fibers. For the trade names of dyes recommended for PES fibers, see → Disperse Dyes, Chap. 7. The Colour Index provides information on dyes made by various producers that correspond chemically with each other. In the case of disperse dyes, however, chemically identical dyes can exhibit marked differences not only in coloring strength and shade, but also in their preparation and finish. Especially in modern dyeing processes, exacting standards are set for fine dispersion and dispersion stability. Disperse dyes are supplied as powder and liquid products. Powdered dyes contain a high percentage of dispersing agent, which is required to maintain a very fine dispersion when drying a dye that has been ground and dispersed in an aqueous slurry. Liquid products contain much less dispersing agent. Depending on concentration, the yields obtained from liquid products in various dyeing processes can deviate from those given by powders. The advantage of liquid products is that they are easier to handle in the preparation of large amounts of dye liquor.

Among the number of trade names available, the dyer selects dyes in accordance with the required properties and with their suitability for certain dyeing techniques. A strict standard is applied to the lightfastness of articles for car seat covers, furnishings, etc., and to the wetfastness of textiles that are washed in commercial laundries. Fastness to thermofixing is important for all articles that are heat-set, pleated, or resin finished after dyeing. Thermal stability plays an important role in the selection of suitable products for thermosol dyeing. Easily subliming dyes contaminate the fixing aggregate. Many dye producers divide their dyes into groups according to thermal stability (e.g., ICI in Class A to D and Sandoz in Class S, SE, and E). Choosing the proper dye for a certain dyeing process is not easy. For this reason, dyes are also classified into groups according to their rate of dyeing and the main temperature range

in which they exhaust [328]. A combination of dyes with similar absorption characteristics quickly leads to level dyeing. To determine combinability, the American Association of Textile Chemists and Colorists (AATCC) recommends an easy-to-perform immersion test [329].

Another classification is based on covering up material differences in textured PES fibers. Dyes that have a low fastness to thermofixing often hardly mark differences in texture; they dye at relatively low temperature and are suited to carrier dyeing processes. Dyes that are very fast to thermofixing diffuse slowly into the fiber and must be dyed at high temperature. With a carrier, the color yield of these dyes is unsatisfactory at boiling temperature; texture and fixing differences in the fiber are frequently marked.

According to the recommendation of the Society of Dyers and Colourists, four tests are used to characterize the dyeing properties of disperse dyes [330]:

1) Migration test for leveling capacity
2) Building up capacity to deep shades
3) Exhaustion test to determine the temperature range in which the main amount of dye attaches
4) Diffusion test (BASF Combi Test [331]) for rate of diffusion in the fiber

Many dyeworks illustrate the coloring behavior of dyes by absorption curves (e.g., [332]).

For dyeing aids and carriers especially suited to dyeing PES fibers, → Textile Auxiliaries, Chap. 5.4., → Textile Auxiliaries, Chap. 5.5., → Textile Auxiliaries, Chap. 5.6., → Textile Auxiliaries, Chap. 5.7.

7.2.2. Dyeing from Aqueous Dye Baths

Articles made of pure PES fibers are dyed almost exclusively in exhaustion processes. Disperse dyes are most stable at a pH of 4 – 5, acetic acid being used preferentially for pH adjustment. A dispersing agent is also added to the dyebath (→ Textile Auxiliaries, Chap. 5.4.). Dyes sensitive to heavy metals (e.g., anthraquinoid red products) are dyed in the presence of a complexing agent [ethylenediaminetetraacetic acid (EDTA) type].

Dyeing under High-Temperature Conditions. Because of the slow diffusion of disperse dyes at boiling temperature, PES fibers are dyed whenever possible at 125 – 135 °C under pressure (HT). The dyeing aggregates required for this purpose are available in most dyeworks. To develop a course of dyeing that is as quick and efficient as possible, studies have been performed on optimizing the temperature – time program, using dyes together that have similar exhaustion properties and adapting the circulation of the liquor or goods to the type of material to be dyed (see Section 7.1.1). In favorable cases, the entire coloring cycle can be shortened to less than 60 min in this way. Attempts are being made to approach the same goal by using complex dye mixtures [333–335].

Efforts are being made to make HT dyeing more economical by reducing the amount of liquor substantially (see Section [**?**]).

The evenness of dye absorption is improved by dyeing in the presence of a leveling agent (→ Textile Auxiliaries, Chap. 5.6.). Carriers are also used as HT leveling agents. They are, however, technically outdated and ecologically harmful. For this reason, the carrier dyeing technique is recommended only when dyeing temperatures of 130 °C are impossible (e.g., with PES – wool blends).

Dyeing below 100 °C. The highest possible dyeing temperature is of advantage for coloring PES fibers. However, if no pressure-stable dyeing aggregates are available, dyeing can be carried out at atmospheric pressure with the addition of dyeing accelerants. In addition, all conditions must allow the theoretically attainable boiling temperature to be approached as closely as possible under local conditions (i.e., by using closed or at least covered dyeing equipment). At the same time, this also prevents inhalation problems due to the carriers, which are generally steam volatile and can be ecologically harmful (→ Textile Auxiliaries, Chap. 5.7.). Carriers must be accurately dosed; too much or too little can result in decreased yield. Since they can also affect fastness, especially lightfastness, carriers must be removed after dyeing by hot rinsing and washing or by drying at high temperature or heat setting.

7.2.3. Special Dyeing Processes

For reasons of levelness, very light shades are sometimes dyed without the use of pressure and without a carrier. Only those disperse dyes that diffuse sufficiently quickly into PES fibers even at 100 °C are suited to this purpose (e.g., C.I. Disperse Yellow 7, Orange 1, 33, Red 4, 11, 60, Blue 56, 81). The dyeing equipment must be as tightly closed as possible, and top heating in vats and jiggers is advantageous. Dyeing time is extended to several hours. Since, in a process of this type, a portion of the dye adheres to the fiber surface without being fixed, dyeing must be followed by a reductive clearing step.

Dyeing from Foam. For saving energy, attempts were made to do entirely without an aqueous dye liquor and to apply disperse dyes to PES articles from a foam medium. Problems can arise with regard to the levelness of dyeing (Aqualuft dyeing machines from the company Gaston County). When considering the energy advantages of the foam dyeing technique, it is important that although water is saved in the dyeing process, considerable amounts of water are required e.g., for pretreatment and aftercleaning. For continuous foam application, see Section 7.2.4.

Air as Transport Medium. A clear economic advantage is observed when air is used as the transport medium for the textile material (Then Air Jet, see Section 2.2.4.5). Here, efforts are made to prevent foam formation. The small amount of aqueous dye liquor present serves as the medium for dyes and auxiliaries.

Dyeing in Alkaline Medium. Oligomeric fiber components migrate out from the fiber during dyeing and can cause problems due to formation of agglomerates with dyes and deposition on the textile material or in the dyeing equipment. The cyclic trimer of ethylene terephthalate is especially harmful because of its low water solubility (< 0.5 mg/L at 100 °C). The oligomer problem is counteracted by dyeing in alkaline medium. In this process, oligomers leaving the fiber are hydrolyzed to water-soluble components before their deposition. To achieve this the pH is adjusted to ca. 10.5 with NaOH at the start of dyeing; it then decreases gradually to about 7.5. Alkali-stable dyes must be used (e.g., Dianix SPH, UPH, Hoechst – Mitsubishi). Many azo blue products are unsuitable for this process. The reductive alkaline aftertreatment can be omitted.

Dyeing from Organic Solvents. Although dyeing from organic solvents in which disperse dyes are highly soluble has been studied carefully, this method has not achieved practical importance. Chlorinated hydrocarbons, in particular, have been recommended as a medium for continuous or batch dyeing of acetate and PES fibers. For continuous dyeing, dye application from either chlorinated hydrocarbons or an aqueous liquor with fixation in solvent vapor is possible. This method is not favored for ecological and toxicological reasons.

Dyeing from Supercritical Carbon Dioxide. According to studies performed in the beginning of the 1990s, dyeing from supercritical CO_2 is possible because of its adequate solubilizing power for disperse dyes [336, 337]. Temperatures around 130 °C and pressures > 20 MPa are used. Short dyeing times, good levelness, and low disposal costs are expected. The CO_2 is recovered. The equipment is naturally very complex.

Dyeing via the Gas Phase. Dyeing with dye vapor in the vacuum or in a gas stream as the medium, has also been investigated. Although advantageous from an energy standpoint, the problems encountered on a practical scale have not yet been solved [338].

7.2.4. Continuous and Semicontinuous Dyeing Processes

Thermosol Process. The most important continuous dyeing process for PES fibers is the thermosol process. It is applied primarily to PES – cellulose fiber blends (see Section 11.1).

The thermosol process consists of four individual steps:

1) Padding of the dye liquor on the fabric
2) Drying
3) Fixing of the dyes in the fiber
4) Aftertreatment

Table 11. Fixing times and temperatures in different dyeing aggregates

Dyeing aggregate	Temperature∗, °C		Time, s
	Dyes with average rate of diffusion	Dyes with low rate of diffusion	
Hot flue, frame	200 – 215	215 – 225	60 – 30
Suction drum	200 – 215	215 – 225	45 – 20
Cylinder fixing	215	220 – 225	30 – 15

∗ Lower dyeing temperature corresponds to longer dyeing time.

The four individual steps usually follow one another in one pass. Systems also exist that include also overdyeing of the cellulose component (e.g., by using the pad steam technique; see Section 2.3.2).

The quality of thermosol dyeing is assessed according to its levelness over large yardages. For good quality, uniform pretreatment of the material and a very uniformly operating thermosol system are required.

Pretreatment ensures, above all, that the material is free of burls and wrinkles, contains no dirt or preparation stains, and is uniformly wettable. It includes careful singeing, washing, and bleaching.

In *padding*, a long dipping section is used so that the liquor penetrates the material completely and is applied evenly over the width of the material. Wetting agents are avoided if possible because they foam, and strong foam formation interferes with the process. Silicone-based foam depressants can negatively affect the stability of the dispersion. The pH of the pad liquor should not exceed 6 – 7 because disperse dyes can be sensitive to alkali. A padding auxiliary, e.g., a polyacrylate or an alginate thickener (→ Textile Auxiliaries, Chap. 5.9.), is used to prevent migration of dye liquor on the fiber surface during drying. Migration can cause two-sidedness of the material, as well as darker- or lighter-colored fiber ends (frosting). Furthermore, dye migration is counteracted by low liquor application (concentrated pad liquor, high squeezing pressure in the padder), low drying temperature, and low air circulation. *Drying* is carried out contact free up to a residual moisture of 20 – 30 % and then on any equipment available.

The *fixing time* and temperature depend on the aggregate, the dyes used, and the depth of shade (see Table 11 for guide values). For equipment requirements, see Section 2.3.1.

In one formulation, dyes with similar rates of diffusion should be used as much as possible. Very quickly diffusing dyes sublime easily, contaminate the fixing aggregate, and produce dyeings that have a low fastness to thermofixation. The color yield of very slowly diffusing dyes is sensitive to temperature variations.

After the dyes are fixed, dyes and auxiliaries that adhere superficially must be washed out, if necessary by an alkaline reductive treatment. In the case of PES – cellulose fiber blends, this washing can be combined conveniently with afterdyeing of the cellulose component, e.g., simply by overdyeing with vat dyes (see Section 11.1). For exact instructions for use of the thermosol process, see [332, pp. 122 – 131].

Pad Roll Process (Section 2.3.2). Disperse dyes are fixed only incompletely to PES fibers under the conditions of the pad roll process (padding, retention in saturated vapor atmosphere). Only light colors can be produced economically with adequate fastness. Carrier additions are not useful because very large amounts would be required. In the pad roll process, rapidly diffusing dyes must be used at any rate.

Continuous Dye Application in Foam. Dyeing from foam (Section 7.2.3) has also been studied for continuous dyeing [339]. Since a slightly inhomogeneous application of the foam becomes visible immediately in a color difference, the consistency, pore size, and disintegration time of the foam must be kept as constant as possible and the articles must be carefully precleaned [340]. Equipment for foam formation and for continuous application has been developed. The continuous foam dyeing technique offers advantages for dyeing floor coverings by greatly reducing liquor application. However, it is not an attractive method for dyeing PES piece goods.

Continuous Dye Fixation with Microwaves. The moist fabric padded with dye

liquor is irradiated with microwaves in the presence of vapor. The addition of urea or carrier is recommended for rapid diffusion of the dyes [341–343].

7.2.5. Dyeing of Microfibers

Fibers with a fineness of less than 1 dtex are called microfibers. Articles made of microfibers are characterized by a low weight, soft feel, weather resistance, and good breathing activity. The large fiber surface leads to rapid dye absorption and, thus, to the risk of unevenness. For physical reasons, the colors also appear to be lighter than on normal fibers. Therefore, greater amounts of dye are required to achieve a required shade. The large fiber surface and the great amounts of dye used also lead to lower light- and wetfastness. Preliminary tests are useful [344, 345].

PES fibers that have been subjected to a surface peeling process with alkali behave similarly to microfibers [346].

7.2.6. Dyeing of Modified Polyester Fibers

Since disperse dyes diffuse very slowly into PES fibers, efforts are made to increase the rate of dye strike by chemical or physical alteration of the fiber. The fiber is also modified to reduce the pilling tendency, to increase shrinkage and elasticity, and to reduce flammability (→ Fibers, 4. Synthetic Organic, Chap. 2.2.1.). Such modified fibers exhibit improved dye receptivity. Fibers with improved dyeability can be dyed with disperse dyes at boiling temperature without a carrier or with basic dyes when they are modified with acidic components (5-sulfoisophthalic acid). Fibers of this type are used if dyeing cannot be carried out easily above 100 °C (e.g., in the case of floor coverings, articles made of PES – wool blends, stretch materials, and cord). Strongly crimped PES bicomponent fibers are produced for special purposes. These fibers are normally also dyeable at the boil and without a carrier [347–351].

Modified PES fibers are usually more sensitive to hydrolysis than normal fibers. This must be taken into account in finishing. The lightfastness of the dyeings is often lower than on normal fibers. Thus, dyes for coloring carpeting, upholstery, and drapery must be carefully selected. When dyeing, one should take into account the fact that dyes start exhausting at low temperatures (ca. 60 °C) and the dyebath is exhausted after a short time, so problems with levelness may arise.

Anionically modified PES fibers not only exhibit affinity for cationic dyes but also are more easily dyeable with disperse dyes than normal fibers. They are dyed with basic dyes up to 115 °C in the presence of Glauber's salt to protect the fibers from hydrolysis. If the dyeing aggregates available do not allow dyeing temperatures of 110 °C, a carrier is employed to give deep shades on anionically modified fibers.

The joint use of anionically modified and normal PES types can be exploited for differential dye effects (joint application of different classes of dyes that differ in their affinity to the fiber components).

7.3. Aftertreatment

In the case of light and medium PES dyeings, as in the coloring of other fibers, the material need only be rinsed thoroughly or soaped after dyeing. In the case of dark shades, nonfixed dye components are held firmly on the fiber surface and decrease fastness. Oligoesters (oligomers) leaving the fiber can also interfere with spinning or winding because they deposit on yarn guiding devices. Therefore, loose material and yarn, in particular, must frequently be subjected to reductive alkaline clearing after dyeing. For this purpose, "hydrosulfite" (sodium dithionite) is used in the presence of an emulsifying washing agent.

In the case of loose stock, reductive aftertreatment also improves spinnability because the fiber becomes smoother. Good spinnability is especially pronounced when organic reducing agents are used for aftertreatment (Cyclanon, BASF, → Textile Auxiliaries, Chap. 5.10.).

To improve the lightfastness of dyes used to color car fittings, benzophenone or triazole compounds (e.g., Cibafast, Ciba) are often employed in the dyebath or as aftertreatment [352].

For partial and complete stripping of faulty dyeings, → Textile Auxiliaries, Chap. 5.13..

8. Dyeing of Cellulose Acetate

For properties of cellulose acetate fibers, see → Fibers, 1. Survey, Chap. 8.; for chemistry and production, see → Cellulose Esters, Chap. 2.1., → Cellulose Esters, Chap. 2.3. Cellulose 2.5-acetate (CA) and cellulose triacetate (CT) are preferentially processed as filaments. Because of their silklike feel, dull gloss, and pleasant wearing quality, they are especially popular for dress and blouse materials, scarves, and linings. In the textile sector, the importance of cellulose acetate is decreasing: world consumption in 1970 was 354 000 t, in 1980 338 000 t, and in 1992 232 000 t. Cellulose acetates are being increasingly replaced by synthetic polyamide and polyester fibers. In contrast to the other regenerated cellulose fibers, CA and CT are hydrophobic. Therefore, they have little in common with viscose or cuprammonium rayon from the viewpoint of dyeing. As hydrophobic fibers, they can be dyed with disperse dyes (→ Disperse Dyes). The dyeability of CA and CT with disperse dyes is very similar to that of PES and other synthetic fibers. The rules described for dyeing PES fibers with disperse dyes (see Chap. 7) basically apply to CA and CT as well. Tests have been drawn up by the Society of Dyers and Colourists for the selection of suitable dyes [353]. Trade names of disperse dyes available for CA and CT are, e.g., Artisil (Sandoz), Celliton (BASF), Cibacet (Ciba – Geigy), Dispersol (ICI), Dianix (Mitsubishi – Hoechst), and Serisol (Yorkshire).

In addition, the ranges of disperse dyes intended for PES fibers contain numerous products that are very well suited to CT in particular.

The choice of dye depends, on the one hand, on the desired use and production fastness and, on the other hand, on the suitability for certain dyeing techniques. Dyes that produce especially washfast dyeings require higher dyeing temperatures for level absorption. An important criterion for the selection of dyes for CA can be fastness to exhaust gases (see Section 8.3). Articles made of CT are subjected to thermal treatment in the course of finishing. The fastness to thermofixing or pleating of the disperse dyes recommended for CT is usually at least as good as on PES fibers.

8.1. Dyeing Processes for Cellulose Acetate [354, 355]

Theoretical aspects of dyeing CA and CT are described in [356, pp. 74 – 92]. Cellulose acetate is dyed by the exhaustion method with disperse dyes in the presence of a nonionic or anionic dispersing and leveling agent in a weakly acidic bath (pH 5 – 6). A series of less wetfast disperse dyes are taken up by acetate at temperatures as low as 50 – 60 °C. For this reason, the process must be started at a correspondingly low temperature to obtain level dyeings. Dyeing is normally done at 80 – 85 °C. Dyes that are fast to wet treatments require temperatures up to 90 °C. If CA is dyed at boiling temperature, the fibers may become delustered.

Piece goods made of CA are usually dyed in a jigger, less often in a star frame or winch beck with gentle material guidance. The piece *beam* is suitable for warp materials. Yarn was previously dyed in hanks. Since suitable winding machines are available, CA can also be dyed on cylindrical cheeses. Woven CA fabrics are occasionally dyed semicontinuously in a pad roll system (Artos) at 80 – 90 °C (see Section 2.2.3). Dyeing auxiliaries are subsequently washed out.

A dyeing process with developing dyes, which was used in the past for various shades, is still being employed for black. Although this process produces dyeings that are especially fast to wet treatments, it is more tedious than dyeing with direct disperse dyes. The process starts with the application of an amino-group-containing diazotizable disperse dye [e.g., Cellitazol (BASF), Cibacet-Diazo (Ciba – Geigy)], which is diazotized on the fiber and then coupled with a naphthol [e.g., 2-hydroxynaphthoic acid (Developer ON)]. Thus, a large-molecular, insoluble, wetfast dye is formed in the fiber.

8.2. Dyeing Processes for Cellulose Triacetate [355]

In terms of dyeing and finishing, CT is more similar to purely synthetic fibers than CA. It can be permanently pleated. For stress relaxation, articles made of CT, like those made of PES, are heat set (thermofixed) after dyeing [356, pp. 92 – 100]. Triacetate, like PES, can be dyed according to the thermosol process (see Section 7.1.2).

The absorption of disperse dyes from a long liquor by CT fibers is slower than CA and somewhat faster than PES fibers. With a few disperse dyes, acceptable dyeings are obtained on CT even at boiling temperature. As a rule, however, dyeing is conducted at 120 °C, especially to obtain dyeings with adequate fastness properties. If this is not possible, a dyeing accelerant (based on butyl benzoate or butyl salicylate) is used. As in the case of CA, CT is dyed in a weakly acidic liquor (pH 5 – 6) in the presence of a leveling dispersing agent. The lower the dyeing temperature, the more must leveling be supported with a suitable auxiliary agent.

Like CA, CT can also be dyed with developing dyes. Since CT is less easily accessible to ionic products than CA, the dyeing process is altered here. Dyeing is begun with the diazotizable azo dye, the developer is then applied at a higher temperature, and diazotization is performed in the cold. The diazonium salt formed couples immediately with the developer present in the fiber.

Dyeing equipment suitable for CA is also suited to dyeing CT. In accordance with the higher dyeing temperature used for CT, pressure dyeing aggregates are preferred (e.g., piece beam dyeing machines, jet dyeing machines, or HT winch becks; see Sections 2.2 and 2.3).

CT can be textured like PES. Textured yarns made of both fibers behave similarly in textile finishing.

For information on the dyeing of cellulose acetate and triacetate in mixtures with other fibers, see [355].

8.3. Aftertreatment, S-Finish

CA and CT absorb nitrogen oxides and other acidic, oxidizing exhaust and industrial waste gases from the air. These gases can destroy sensitive dyes in the fiber so that the shade changes. Products that are not resistant to gas fading are found especially among the red and blue anthraquinone dyes (→ Anthraquinone Dyes and Intermediates, Chap. 3.1.1.2.). If dyeing cannot be done with dyes that are fast to gas fading, the fiber can be protected to a certain extent by aftertreatment with cationic products that undergo saltlike bonding with the mostly acidic gases.

Articles made of CT are subjected to heat setting (thermofixing) after dyeing to make them dimensionally stable.

In the case of CT, effects similar to those produced by thermofixing are obtained by alkali treatment at higher temperature (S-finish). The S-finish results in superficial saponification of the fiber, making the fiber surface hydrophilic, which in turn corresponds to a permanent antistatic finish. The resistance to gas fading of dyeings is also improved. CT with an S-finish is dyeable with the usual direct dyes, the depth of shade depending on the degree of S-finishing.

9. Dyeing of Acrylic (Polyacrylonitrile and Modacrylic) Fibers

9.1. Economic Aspects

In 1991, worldwide production of acrylic fibers amounted to 2.31×10^6 t. Thus, these represent the third most important synthetic fibers after PES and PA. About 702 000 t were produced in Western Europe, 350 000 t in Japan, and 206 000 t in the United States. Fiber production decreased from 1988 to 1991, with an average annual rate of − 2.3 % [357].

The main advantages of acrylic fibers are their similarity to wool in feel, heat retention, and processing, as well as good dyeability. In Western Europe and the United States, ca. 60 % are used for clothes and 30 % for household textiles [357]. The proportion used for clothes worldwide is considerably larger.

Dyeing is performed mainly on yarns by exhaustion processes (batchwise). However, the tendency toward gel dyeing is increasing. The proportions of different dye application processes in 1989 and of various textile makeups are given below (in percent):

Proportion of application processes	
Exhaustion process	72
Gel dyeing	13 (increasing)
Continuous dyeing	9 (decreasing)
Printing	4
Mass dyeing	2 (decreasing)
Proportion of textile makeup for coloration	
Yarn	50
Stock, cable, top	22

Gel	13
Piece (dyeing)	9
Printing	4
Mass	2

Total dye consumption is estimated at ca. 24 000 t/a.

9.2. Principles

For methods of production and properties of acrylic fibers see, → Fibers, 4. Synthetic Organic, Chap. 5. The polymerizates used initially, which were made of 100 % acrylonitrile without suitable comonomers, gave fibers with a high glass transition temperature T_g and with insufficient dye-binding possibilities. Fibers of this type could be dyed only in light shades with *disperse dyes* or with *acid dyes* according to the complicated copper ion process [358]. Today, anionic comonomers are used, and anionic groups are incorporated into the chain ends of the fiber molecules. In this way, T_g is lowered and anionic groups are available that can act as dye sites for cationic dyes. Thus, acrylic fibers are reliably and economically dyeable.

9.2.1. Disperse Dyes

Disperse dyes (→ Disperse Dyes) can be used to produce light to medium deep shades. The dyeing mechanism and process correspond to those used on polyester fibers (see Chap. 7). However, dyeing can be performed at $< 100\,^{\circ}$C. Addition of carriers is not required. The good migration properties of disperse dyes result in problem-free level dyeing.

9.2.2. Cationic Dyes

Cationic dyes (→ Cationic Dyes) are used predominantly for acrylic fibers [359] because virtually any depth of color can be obtained and they are dyeable with high brilliancy and good fastness. In choosing the dyeing conditions, the dyeing mechanism and relevant properties of the dyes and fibers should be taken into account. This is relatively easy because despite the great variety of fibers and dyes, uniform and – in comparison with other dyeing processes – simple laws are valid for dyeing with cationic dyes. A detailed description is presented in [360], and a discussion with a literature survey of the more recent developments in certain areas (e.g., pore and surface structure, fiber modifications, thermal yellowing, effect of carriers and solvents during dyeing, and transfer print) in [361]. The physical chemistry of dyeing in general is presented in [362, 363].

The dyeing process with cationic dyes can be described by the following model. The dye rapidly absorbs to the surface of the fiber from the aqueous solution in the dyebath until saturation occurs. Above T_g, dye diffuses into the fiber. This diffusion is usually rate determining for the entire process. Because of their considerably higher affinity, dye cations displace an equivalent amount of cations (H^+, Na^+) that were previously present at the anionic groups in the fiber (ion exchange). This process determines the dye distribution in the dyeing equilibrum [364–367] and can be described by the Donnan theory [368, 369].

In acrylic dyeing, the dyeing (or absorption) rate v is especially important. Because of the high affinity of cationic dye for the fiber, their migration capacity is poor. Thus, dyeing must be sufficiently level from the start; this requires the control of v, which can be reliably achieved only if the effects of relevant parameters are known. Since v is determined by the diffusion of dye in the fiber, the following relation is valid according to the laws of diffusion [370] (see Section 2.1.3)

$$v \sim \Delta C \cdot D \tag{9.1}$$

where ΔC is the concentration gradient and D is the average, apparent diffusion coefficient of dye cations in the fiber. The parameter ΔC depends on the dye concentration on the fiber surface C_s. The amount of dye $C_F(t)$ that has diffused into the fiber after time t follows a $\sqrt{t}$ law [371]:

$$C_F(t) \sim C_s \sqrt{Dt} \tag{9.2}$$

Except when only a small number of dye cations are available (for production of very light dyeings), the fiber surface is almost saturated during dyeing, i.e., C_s is approximately constant until most of the dye is exhausted. At constant dyeing temperature, D can also be regarded as approximately constant. Thus

$$C_F(t) \sim \sqrt{t} \tag{9.2a}$$

To a good approximation, this is valid until ca. 70 % of the dye supplied is absorbed (bath exhaustion BE = 70 %). Thus, some rules exist concerning the dependence of dyeing time and rate of dyeing on the concentration p of supplied dye [372]. For a given temperature or temperature program:

1) The dyeing time required increases with p^2 (p is normally given in percent based on fiber mass).
2) The relative (based on p) rate of dyeing $v_{rel} = dp/dt$ (percentage of dye per minute) is inversely proportional to p^2.
3) The absolute rate of dyeing $v_{abs} = dC_F/dt$ (milligrams per minute) is independent of p.

Thus, the lighter the color, the faster is the absorption and the greater the danger of nonlevelness if absorption is not controlled by a suitable temperature program or retarders. Furthermore, Equation (9.2) denotes the following. Uneven surface charging with dye cations leads to differences in C_F (i.e., nonlevelness). Surface charging occurs rapidly and can hardly be influenced by temperature guidance. In the case of light colors, when the number of dye cations provided is insufficient to saturate the surface, retarders (→ Textile Auxiliaries, Chap. 5.6.5.) must be applied to achieve this purpose; only then can uniform surface charging be expected.

The *temperature dependence* of v is especially pronounced in acrylic dyeing [364, 372, 373]. Depending on the type of fiber, T_g is in the range of 70 – 80 °C; above T_g is the absorption range. In the absorption range, D and, according to Equation (9.2), v increase steeply with T in accordance with the Williams – Landel – Ferry (WLF) equation derived from the theory of chain segment motion [373–375]. The Arrhenius equation is also valid to a usable approximation and is more practical because it contains only one constant, the activation energy E [362]. The values of E for v are 250 – 300 kJ/mol [364, 372], the lower values being for wet spun fibers. This corresponds to a doubling of v for a temperature increase of only 2.5 – 3 °C. (In the case of disperse dyes on PES or acid dyes on PA a doubling of v is observed at a temperature increase of ca. 5 and 10 °C, respectively.) These values are practically independent of the choice of dye, depth of color, and mode of dyeing [372]. The steep increase of D and v with temperature is predominantly caused by changes in the fiber.

Apart from T and the amount of dye applied (in the case of v_{rel}), v also depends on the properties of individual dyes and the type of fiber. To take this into account, in practice, especially in systematic dyeing processes, these effects are separated and described in terms of dye- and fiber-specific characteristic values. For instance, the *dyeing times of dyes under standard conditions* and the *rate of dyeing V* of the fiber have been introduced for this purpose [372].

The degree of occupation by dye cations of anionic dye sites present in the fiber is decisive for the dyeing equilibrium, in practice the final exhaustion of the dyebath. At constant temperature (maximum dyeing temperature), dye distribution follows a Langmuir isotherm [376, 377].

$$C_F = \frac{s \cdot C_L}{s/a + C_L} \tag{9.3}$$

where C_L is the dye concentration in the dyebath, s is a saturation concentration, and a is essentially the affinity of the dye. The value s denotes the maximal possible dye uptake by the fiber for very high p. In this case, all dye sites are occupied, the number of dye cations n_{cat} on the fiber equals the number of anionic dye sites of the fiber n_s. [Virtually all of the anionic dye sites of the fiber are accessible to dye cations under dyeing conditions (i.e., above T_g)]. If $n_{cat} < n_s$, the supplied dye is exhausted almost completely because of the normally high affinity; if $n_{cat} > n_s$, the excess remains in the bath. This relationship is stoichiometrically valid and usually independent of dyeing conditions. Thus, the position of the dyeing equilibrium can be calculated for each individual dyeing process. For this purpose, fiber- and dye-specific values derived from molar quantities are used. The following have proved useful [372, 378]:

1) The *fiber saturation value* S_F, a measure of the concentration N (in millimoles per kilogram of fiber) of anionic dye sites in the fiber:

$$S_F = N/25 \tag{9.4}$$

2) The *saturation factor f* of a cationic dye having a molecular mass M_r and a net dye content e (grams of pure dye per gram of commercial dye):

$$f = 400e/M_r \tag{9.5}$$

The *saturation concentration* p_s (percent dye, based on the mass of textile material) of an individual dye (with $n_{cat} = n_s$) is

$$p_s = S_F/f \tag{9.6}$$

A combination of several dyes (1, 2, 3,...) exhausts almost completely if

$$(p \cdot f)_1 + (p \cdot f)_2 + (p \cdot f)_3 + ... < S_F \tag{9.7}$$

In the case of PAC types with strongly anionic (acidic) dye sites, S_F increases slightly with increasing pH (by ca. 10 % from pH 4 to 6.5). In the case of PAC types with weakly anionic dye sites (carboxyl groups), the increase is considerably greater [377].

Another important quantity is the *relative saturation* of the fiber in the dyeing equilibrium [372]

$$S_{rel} = \sum (p \cdot f)/S_F \tag{9.8}$$

The higher the S_{rel} (other parameters being equal), the better is the levelness of dyeing during absorption. If S_{rel} is about equal in different dyeings and certain other conditions are met (compatibility, use of retarder, proper temperature guidance), a similar course of the dyeing process with regard to the rate of dyeing, bath exhaustion, and levelness can be expected.

Retarders and Auxiliaries.

Cationic Retarders. With light and moderately deep shades, the amount of dye applied is often not enough for an adequate S_{rel}. Sometimes even saturating the fiber surface is not sufficient. To meet the danger of nonlevelness in such cases, cationic retarders are used (→ Textile Auxiliaries, Chap. 5.6.5.) [379, 380]. These are colorless cationic compounds that absorb like dyes and compete with dyes for acidic groups in the fiber and on the fiber surface. They are usually quaternary amines with an aromatic substituent and an aliphatic chain. The longer this chain, the greater is the affinity for the fiber [381, 382].

These compounds are used mainly to *increase* S_{rel}, thereby improving levelness. The increase can be calculated in accordance with Equations (9.5) – (9.8), the retarder being assigned a saturation factor according to Equation (9.5) and treated formally like a dye [372]. This calculation is important to guarantee the desired effect and prevent S_{rel} from becoming too large (close to 1 or higher). Dye absorption would then be incomplete (blocking of anionic groups).

Depending on the affinity and amount applied, cationic retarders partially displace the dye from the fiber surface. C_s (dye) and the concentration gradient of the dye become smaller and remain constant longer in the course of dyeing [383]. For this reason, at constant temperature T, the *rate of dyeing is lowered* by the retarders. However, this effect does not justify the use of retarders because it would be obtainable at a lower cost by dyeing at lower temperature.

Cationic Softeners. Depending on their affinity and penetrating ability, some cationic softeners – when used in the dyebath – act like cationic retarders and contribute to S_{rel}. Thus, less cationic retarder is required. However, their softening effect is lower than when applied in aftertreatment at temperatures below T_g.

Migration Aids. Colorless cationic auxiliary agents having a considerably lower affinity for the fiber than dyes are available as migrators (migration aids) (→ Textile Auxiliaries, Chap. 5.6.5.). They can be used in larger molar amounts than cationic retarders without danger of blocking. They influence S_{rel} and levelness during absorption only slightly, but they promote migration and help compensate for unevenness in the fiber [384].

Electrolytes. The addition of electrolytes has a similar effect [380]. In accordance with their much lower affinity compared to migration aids, this effect is weaker, based on the number of cations used. However, since this number is generally much higher than in the case of migrators comparable effects are obtained.

Polycationic retarders [i.e., compounds (usually polymeric) with numerous cationic groups] have been recommended as leveling agents [385]. They occupy the fiber surface without penetrating into the fiber and counteract dye adsorption, thereby reducing the rate of dyeing [386, 387]. Depending on their adsorption characteristics, they also decrease the migration capacity and the bath exhaustion in the dyeing equilibrium. However, a favorable effect on leveling has not been found. (The rate of dyeing can be influenced better through a suitable temperature program.)

Anionic Retarders. In some cases, anionic auxiliaries are also used as retarders (→ Textile Auxiliaries, Chap. 5.6.5.) [388]. These contain two or more anionic groups (sulfonic acid groups) per molecule, forming a 1:1 addition compound with the cationic dye; that is soluble in the dye liquor. To ensure this solubility, at least two anionic groups must be available from the start for each dye cation. The minimum amount of anionic retarder required can be calculated via f values according to Equation (9.5) [360]. In the course of dyeing a growing excess of auxiliary anions is formed, and the concentration of free dye cations in the liquor falls below that required for saturation of the fiber surface. Dyeing rate and bath exhaustion at dyeing equilibrium are decreased.

Anionic retarders promote the migration of cationic dyes. They reduce the danger of precipitations when anionic dyes are used in addition to cationic dyes (for fiber blends). They influence the compatibility of dyes (see below). The higher the affinity of the cationic dyes, the greater is their effect. The main disadvantages are the lower bath exhaustion and poorer reproducibility.

Compatibility of Cationic Dyes. Cationic dyes influence each other during absorption [365, 367]. If several dyes exhaust tone-in-tone from one bath (i.e., the depth of color, but not the shade, changes in the course of absorption), the dyes are referred to as compatible. In the case of cationic dyes on PAC, the property of being compatible is dye specific and also applies to dyeing under other conditions and in other shades. It depends on the product $D \cdot A$ (A = dye affinity) and can be described by the compatibility value K (see Chap. 12) [372, 389, 390].

In acrylic dyeing, combinations of dyes with the same K exhibit considerable advantages in level dyeing. In combinations with different K, dyes with a lower K absorb before those with a higher K. Cationic retarders can also be assigned K values [360, 391]. The K values change with the use of anionic retarders.

9.3. Exhaustion Process

Normal PAC textile material is dyed as cable or stock in package dyeing, as yarn in hanks or packages (cross wound, cylindrical, or conical) (320 – 350 g/L, direction of liquor inside – outside), as pieces on beam, overflow, paddel (knitwear), or drums (socks) (see Section 2.2). High-bulk (HB) yarns are applied as muffs (cylindrical bobbins without tubes, without intermediate plates) [392]. Strongly soiled textile material is prewashed, preferably in a weakly alkaline wash.

The principles of recipe and process operation are described in [383]. Because of the poor migration properties, dye absorption must be as level as possible from the start. Correspondingly, attention should be paid to uniform packaging, package density, flow or movement of textile material, and controlled machine conditions. As far as possible, dyes with the same K are selected (the largest choice of dyes and most suitable retarders are available at $K = 3$). In cases where sufficient levelness during absorption is critical, dyes with a relatively low affinity (with high K) are recommended. These dyes migrate somewhat better than normal dyes [393]. The dyeing formula is drawn up with the assistance of a color measuring system (see Chap. 12) or according to available documentation, and the amount of retarder is adjusted to give a favorable S_{rel}. The dyebath is charged with the dissolved dye and all the other additives (see Table 12) at 60 – 70 °C, and the textile material is added. After a mixing time of ca. 5 min, dyeing is conducted by controlling the temperature, as shown in Tables 13 and 14. This is followed by cooling (ca. 0.3 °C/min to 60 °C), rinsing, and aftertreatment.

Table 12. Additives to the dyeing bath, general formulation

Additive	Light hues	Medium hues	Deep hues
Nonionic dispersant, g/L	0.2 – 0.5	0.5	0.5
Acetic acid, final pH	3.6	3.6	4.5
Sodium acetate anhydrous, g/L			0.5
Sodium sulfate anhydrous, g/L	5	5	
Cationic retarder∗, %	1.5 – 3	0.3 – 1.5	

∗ Retarder with $K \approx 3$, $f \approx 0.5$, in percent based on mass of material; higher amounts are required for textiles with higher S_{F}, at critical requirements with respect to levelness, dye combintions with different K values or $K \leq 2$, and for retarders with $K > 3$ (see Section 9.2.2).

The details given in Tables 12, 13, 14 take into account in a general way (same conditions for larger groups of dyeings) the principles explained in Section 9.2.2 – especially the need to

achieve sufficiently level dyeing from the start and to have the shortest possible dyeing time. The liquor movement or circulating speed of the material can be taken into account [394–396]. As a rule, satisfactory results are obtained in this manner.

Table 13. Temperature range of controlled heating (general method) ($T_{start} - T_{end}$) and residence time (t_{max}) at maximum temperature T_{max}

	S_{rel}		
	Low	Medium	High
T_{start}, °C	75	80	85
T_{end}, °C	95	98 – 100	98 – 100
t_{max}, min			
(at T_{max} = 98 °C)	20	40	60
(at T_{max} = 102 °C)	15	30	40

Table 14. Heating rate (general method)

	Heating rate, °C/min		
Temperature range∗	Stock, top	Yarn (pack.)	Hank, piece
$< T_{start}$ (until absorption of dye starts)	quick	quick	quick
$T_{start} - T_{end}$ (usually linear)	0.5	0.3 – 0.4	0.25
$T_{end} - T_{max}$ (until absorption of dye is complete)	quick	quick	quick

∗ Numerical data for the temperature can be obtained from systematic optimization processes.

Dyeing can be performed more economically and reliably according to systematic methods for the optimization of formulations and dyeing conditions [372, 383, 397]. Here, dyes with the same K are selected and the following are determined for each dyeing:

1) The amount of retarder that leads to the most favorable relative saturation S_{rel}
2) The temperatures at which exhaustion starts or ends (the critical temperature range outside which rapid heating will not cause unlevelness, see Section 2.2.1)
3) The most favorable rate of heating in this temperature range
4) The dwelling time at the maximum (final) dyeing temperature required for penetration

In following this method, the rules mentioned in Section 9.2.2, the properties (characteristic values) of the fibers and the dyes, and the conditions of the dyeing machine and the article are systematically taken into account. This optimization can be performed with the help of tables [397] or with a computer [398].

Substantial advantages are obtained by:

1) Appropriate amounts of auxiliary agents applied
2) Shorter dyeing times
3) More protection against nonlevelness
4) Better programmability
5) Better reproducibility
6) Saving of laboratory work
7) Help in searching for errors

9.4. Dyeing of Special Fiber Types According to the Exhaustion Process

High-Bulk Material. The bulking of high-bulk material proceeds either as a pretreatment (continuous for yarn, batchwise and usually with steam for yarn and knitwear) or in the dyebath before dye absorption at ca. 80 °C for 5 – 10 min without liquor movement [392]. To avoid nonlevelness in this process, retarders are employed (high S_{rel}) to ensure that very little dye absorbs during the bulking time. The formulas given in Section 9.2.2 or systematic dyeing processes (Section 9.3) are used. With HB bicomponent fibers, a high S_{rel} can interfere with the reversible shrinkage. For this reason, anionic retarders and dyes with $K \geq 4$ are preferred [393].

Pore Fibers. For the surface and pore structure of PAC fibers, see [361]. As a result of their greater light scattering, pore fibers (type Dunova, light fibers with high water absorption capacity) require up to three times more dye than normal fibers to achieve a given shade [399]. This relation depends on the angle of observation. Navy and black shades are hardly dyeable; for dyeing other deep hues, an ammoniacal reductive aftercleaning step is required. The diffusion of dyes proceeds mainly in the pores of the fiber [400, 401], changing the K values of the dyes. With the exception of a higher softener requirement, the rules of normal PAC dyeing apply.

Microfibers. PAC microfibers (single titer below 1 dtex) behave differently in some respects from normal acrylic fibers [402]. First, more dye is required for the desired shade in each case because of the greater light scattering of

Table 15. Dyeing properties and recommended range of dyeing temperature for different types of modacrylic fiber

Type of fiber (producer)	S_F [a]	V [b]	Auxiliary	Temperature range, °C
Dynel [c] (Union Carbide)	0.9	1.0	carrier	40 – 98
Kanekalon S (Kanegafuchi)	2.2	1.2	migrator carrier (deep hues)	60 – 98
Teklan (Courtaulds)	4.3	2.4	cationic retarder	60 – 98
Velicren FRS (Snia)	4.1	2.5	cationic retarder	80 – 98
Verel (Eastman)	0.8	1.0	carrier	30 – 70

[a] S_F = Fiber saturation value (see Section 9.2.2). [b] V = Characteristic value for the rate of dyeing of a fiber (see Section 9.2.2).
[c] Several types with partially different properties.

the fiber. This requirement increases more than twofold with increasing depth of color (similar to pore fibers). For this reason, the fiber saturation S_F is usually adjusted to a higher value by the fiber producer. The characteristic dyeing rate V of the fiber is also higher. Corresponding to these increased values, more cationic retarder is required and the starting temperatures are lower. Above $S_{rel} \approx 0.75$, the rub- and wetfastness can be lower; reductive aftercleaning is sometimes necessary. Some dyes exhibit increased sensitivity to boiling down (lower dyebath stability) when the pH clearly deviates from 4.5. For this reason, dye producers recommend a special selection of dyes for microfibers. In the dyeing of stock, top, or yarn, one should take into account the fact that the liquor flow is lower and its uniformness is more problematic than in the case of normal acrylic fibers. For this reason, packaging and winding must be especially even and the rate of heating should be slower if possible. In piece dyeing, longitudinal tension should be kept as low as possible and creasing must be avoided. Cooling to < 60 °C should not proceed more than twice as fast as heating.

Modacrylic Fibers (→ Fibers, 4. Synthetic Organic). Modacrylic fibers containing more than 20 % of vinyl chloride or vinylidene chloride as comonomer are used because of their reduced flammability. Depending on the comonomers, various types of modacrylic fibers exist that differ especially in the range of dyeing temperatures. In all cases, the critical temperature range is greater (25 – 50 °C) than with normal fibers (15 °C), and absorption starts in most cases at lower temperature. Modacrylic fibers generally exhibitincreased plasticity and a strong tendency to shrink at higher dyeing temperature. The fiber types and their dyeing properties change frequently. The lightfastness on these fibers is considerably poorer than on normal PAC fiber, making a corresponding dye selection necessary [403, 404]. Dyeing is performed at pH 4.5 with acetic acid, 0.5 wt % of sodium acetate, and 1 – 2 wt % of nonionogenic dispersing agent. Depending on the type of fiber, carriers or cationic retarders and different temperatures are recommended. Examples of fiber types are given in Table 15. The relustering required in the past after dyeing is usually no longer necessary.

Modacrylic fibers can also be dyed with *disperse dyes* (similar to normal PAC fibers, but the fastness to sublimation is poor).

9.5. Continuous Processes

Stock, cable, and top can be dyed on special machines according to the *pad steam process* (see Section 2.3.2) [405–407], preferably with pressurized steam of clearly more than 100 °C to obtain short fixing times. To avoid drying and condensed water spots, the steam should not be either superheated or wet. A typical pad liquor contains acetic acid (pH 4.5) and dye solvent in addition to especially steam-resistant, readily water-soluble cationic dyes (usually liquid brands) [406, 408]. *Space dyeing* can be regarded as a variant of this mode of dyeing. A new suggestion involves simultaneous bulking and dyeing in a continuous process [409].

The *pad steam process with fixing by saturated steam* at 100 – 102 °C is used for piece goods, especially upholstery material (velour). This requires longer fixing times (10 – 20 min). Rapidly diffusing cationic dyes and dye solvents, which also exhibit a carrier effect (→ Textile Auxiliaries, Chap. 5.6.5.) [361, 410, 411], are required. If substantive dyes are used simultaneously (for cotton backs), the dye solvent should also have a good dispersing action.

Dry heat fixing (*thermosol process*, see Section 2.3.2) is unimportant for PAC dyeing be-

cause the fiber tends to yellow and harden in this process and cationic dyes are fixed slowly and incompletely [361,406]. *Transfer printing* is also not suitable without either elaborate pretreatment or special swelling agents or solvents [361]. Disperse dyes produce only very light shades.

Dyeing wet spun PAC fibers during the production process in the *gel state* is especially economical. This occurs after the solvent has been washed out and before the usual follow-up processes, especially stretching and drying [412–415]. The gel absorbs cationic dyes very rapidly (within seconds or a few minutes). There is almost no difference from normal PAC dyeing with regard to saturation, shade, and fastness. The dyes are required in large amounts. Thus, liquid brands, which are miscible without the occurrence of precipitation, are preferably used. Mixtures, even diluted, must be stable in storage (premixability).

The *mass dyeing* of PAC is important only for highly weather-resistant, special articles (awnings and tents).

10. Dyeing of Other Synthetic Fibers

10.1. Poly(Vinyl Chloride) Fibers

Poly(vinyl chloride) (PVC) fibers (→ Fibers, 4. Synthetic Organic, Chap. 6.) are characterized especially by their flame retardance. For this reason, they are used, e.g., for airplane seat covers, quilt fillings, and purely technical purposes.

PVC fibers are dyed preferably with *disperse dyes* (→ Disperse Dyes) [416, p. 404], [417, pp. 611]. As with modacrylic fibers, excessively high temperatures must not be used because of shrinkage of the PVC fiber. Hence, some fibers are dyed at 60 – 65 °C with dyeing accelerants. Other PVC fibers can be dyed at 100 °C without a carrier and a few even at 110 °C. Dyes must be selected with regard to the lightfastness desired.

Cationic dyes in combination with an anionic auxiliary can also be used to achieve average depths of shade. Here, good lightfastness is achieved with selected dyes.

Similar to certain modacrylic fibers, some PVC fibers, such as Leavil, also lose their luster on dyeing. It can be recovered by relustering (dry heat at 110 – 130 °C).

10.2. Elastomeric Fibers

Elastomeric fibers made of polyurethane (→ Fibers, 4. Synthetic Organic, Chap. 3.) are contained in most stretch articles and also in fashion materials and knitted fabrics [418]. Mixtures with polyamide are used for corsetry, orthopedic support stockings, panty hoses, and swim or sports wear. Articles made of cotton – elastomeric fibers and PES – cotton – elastometric fibers are also very common. The companion fiber is often spun around the elastomeric thread, which makes separate dyeing of the polyurethane fiber unnecessary. This is substantiated by the usually moderate fastness of most dyes on polyurethane fibers.

Polyurethane fibers can be dyed with acid dyes (→ Anthraquinone Dyes and Intermediates, Chap. 3.2.1.; → Azo Dyes, Chap. 3.2.), metal-complex dyes (→ Azo Dyes, Chap. 3.2.1.4.; → Metal-Complex Dyes), chrome dyes (→ Azo Dyes, Chap. 3.2.1.3., and disperse dyes (→ Disperse Dyes). In general, higher dyeing temperatures, especially below pH 4 and above pH 9, are not permissible because the fibers are degraded (i.e., they lose their elasticity and strength). Efforts are made to circumvent the difficulties arising due to the inadequate dye receptivity of polyurethane fibers with the help of fiber-affinitive cationic dyeing aids in combination with ethoxylated fatty alcohols or fatty amines [417, pp. 572 – 574].

Light shades can be made tone-on-tone on polyamide – polyurethane mixtures with disperse dyes at 95 – 98 °C and pH 6 – 7. However, the wetfastness of these dyeings on polyurethanes is lower than on polyamide. The mode of dyeing with acid and metal-complex dyes corresponds to that of polyamide dyeing at boiling temperature and pH 5 – 6 (see Sections 6.4.2, 6.4.3) [419]. If polyamide – polyurethane mixtures are dyed with chrome dyes, the polyurethane fibers lose elasticity and the articles acquire a brittle feel.

Because of the temperature sensitivity of polyurethane fibers, mixtures of elastomeric and polyester fibers must be dyed with small molecular, rapidly diffusing disperse dyes in 30 min at 120 °C according to the HT process [420]. Modified PES fibers that are dyeable at 100 °C

without a carrier are often used in mixtures with elastomeric fibers.

Fiber blends that contain cotton in addition to elastomeric fibers can be dyed with vat dyes, sulfur dyes, combinations of reactive and disperse dyes, and combinations of substantive and acid dyes [416, pp. 342 – 349].

In all dyeing processes for elastomeric fibers, dyeing equipment that permits low-strain guidance of the material and the lowest possible thermal stress are important.

10.3. Polypropylene Fibers

Polypropylene is an inexpensive raw material that produces fibers with a series of desirable properties such as dimensional stability, low weight, low water absorption, resistance to chemicals, and rotproofness (→ Fibers, 4. Synthetic Organic, Chap. 4.1.).

Consisting of pure aliphatic hydrocarbons, these fibers have no groups with a special dye affinity that can specifically interact with dyes normally used for textile dyeing. Lipophilic dyes, which are employed for dyeing mineral oil, have sufficient affinity for polypropylene so that they were occasionally recommended for dyeing carpet backing materials. However, they produce dyeings of such poor fastness that they are not useable for textile purposes. Hence, mass dyeing with pigment formulations is the commercial-scale method of choice for coloring unmodified polypropylene fibers (→ Fibers, 4. Synthetic Organic, Chap. 4.1.2.5.) [421].

For wide application, fibers should be dyeable in processes commonly used in textile dyeing. For this reason, numerous attempts have been made to modify polypropylene to make it dyeable from aqueous liquor. For surveys, see [422–424], [425, pp. 671 – 675], [426].

In all dyeing processes, the low melting point of the polypropylene fiber must be taken into account. The fiber becomes plastic at 130 °C, and treatment temperatures must be kept below 120 °C because of shrinkage.

Modified Polypropylene Fibers. Polypropylene, which contains nickel compounds for light stabilization, can be dyed with dyes that are capable of forming nickel complexes. Nonfading, wetfast dyeings are produced in mostly dull shades with water-insoluble, chelating dyes [e.g., Altcolene (Althouse)].

The polypropylene fibers that were on the market for some time were modified by basic components and could be dyed with the acid dyes commonly used for wool and polyamide fibers. To obtain acceptable lightfastness, pretreatment with an acid or an alkylsulfonate was required before dyeing.

In a third method, the polypropylene structure is loosened by a polyether or a polyamide, making the fiber accessible to disperse dyes. The depth of shade and fastness properties of these dyes are sufficient for many textile purposes including floor coverings.

Preliminary tests are required in any case. The type of modification can be determined easily by two test dyeings. If an acid dye is absorbed by the fiber, the modification must be basic. A test with a disperse dye capable of complex formation shows if nickel is present in the fiber. Polypropylene fibers that contain nickel are dyed brown with C.I. Disperse Red 91, and differently modified fibers are dyed pink.

10.4. Poly(Vinyl Alcohol) (PVA) Fibers

The importance of PVA fibers for the textile industry is decreasing greatly. The cotton-like feel, high strength, low moisture absorption, good rot resistance, and weatherproofness make these fibers suitable for sewing thread and for canvas and other articles used outdoors (→ Fibers, 4. Synthetic Organic, Chap. 7.). The fiber is water soluble after spinning from an aqueous solution. It is subsequently cross-linked with formaldehyde (acetalization). The higher the degree of cross-linking, the more difficult it becomes to dye the fiber because dyeability depends on the number of free hydroxyl groups.

Although PVA fibers are resistant to dry heat (exposure to 150 °C reduces the strength by only 10 %), the fibers harden. In moist heat, steam, or an aqueous dyebath, the temperature should not exceed 100 °C to prevent greater shrinkage and hardening of the fiber. In general, PVA fibers need not be bleached; optical brightening is sufficient for white materials. PVA fibers resemble cellulose and polyamide in their dyeing properties. *Disperse dyes* are suitable only for light shades because of poor light- and wetfastness.

Although the same applies to *direct dyes*, their wetfastness can be improved by the usual aftertreatment. The color yield in the case of direct dyes depends on the number of sulfonate groups in the dye molecule; it should not be too high. Deep shades can also be obtained [416, pp. 405 – 406].

Cationic, reactive, and *sulfur dyes* are taken up to PVA fibers and exhibit varying fastnesses. The most suitable are *vat dyes* (e.g., Indanthren brands) and poorly water-soluble *metal-complex dyes* (e.g., Avilon, Ciba – Geigy; Vialon, BASF). With *vat dyes*, IW or IN dyeing is carried out (see Section 4.3) depending on the individual dye. To improve color yield, the dyeing temperature should be increased to 85 °C. Nevertheless, really deep shades are difficult to achieve. In dyeing, the fact should be taken into account that although vat dyes level easily in the leuco form, they are not or are only incompletely revattable after oxidation so belated leveling becomes almost impossible. Vat dyes produce a fastness with PVA fibers that is comparable to or better than that obtained with cotton.

With *weakly acidic metal-complex dyes*, dyeing is performed at pH 4 and boiling temperature in the presence of a leveling agent, e.g., an ethoxylated fatty alcohol sulfate (Uniperol W, BASF). The higher the temperature, the better is the color yield. The leveling capacity is very good. Very fast dyeings are produced that are also suitable for weatherproof articles.

10.5. Aramid Fibers

Polyamides made from aromatic amines and dicarboxylic acids are called aramids (→ High Performance Fibers, Chap. 2.2.1.). In their textile application properties they are similar to polyamide and polyester fibers. They have the advantage of being highly heat resistant and flame retardant.

For this reason, aramid fibers are used for protective clothing. However, the relatively high cost and moderate lightfastness of the dyeings prevent further application in the textile sector.

The aramid fiber most commonly used for textile purposes is Nomex (Du Pont). Nomex must be heat set by steaming (at 120 or 90 °C, depending on the type) before it is wet finished. Steaming at excessively high temperature reduces dyeability. Thorough washing is also required for level dyeing. *Cationic dyes* are used exclusively for coloring Nomex. Suitable products are, e.g., C.I. Basic Yellow 15, 21, 23, 25, 29, 49, 53, 79, Basic Orange 22, 35, Basic Red 29, 46, Basic Blue 41, 54, and Basic Green 6. A lightfastness of 3 – 4 is achieved, with Basic Green 6 giving 5 – 6. To attain a good color yield, dyeing is carried out at pH 3 – 4 and 120 °C for 2 h. In addition, a carrier is used, preferably benzyl alcohol in amounts of 10 – 40 g/L, depending on the depth of color. The usual carriers can also be employed. Exhaustion of the dyebath is improved by the addition of 25 g/L of sodium nitrate. After dyeing, thorough afterwashing is required with the addition of sodium hydrogensulfite, if necessary, to remove superficially adhering dye [425, pp. 675 – 676].

In the case of faulty dyeing, the dye can be largely stripped by treatment at 120 °C with benzyl alcohol or with a carrier and a retarder which has a leveling effect in the presence of hydrosulfite in an alkaline liquor.

Other very strong or high-temperature-resistant fibers [e.g., poly(tetrafluoroethylene) or carbon fibers] cannot be dyed with the methods usually employed in textile dyeing.

11. Dyeing of Fiber Blends

For dyeing mixtures of wool and polyamide fibers, see Section 5.1.5.

From time immemorial the properties of textiles have been varied by mixing fiber types. Blends that are rather unimportant today (e.g., half linen, a mixture of cotton and linen, and half wool, a mixture of wool and cellulose fibers) were widely employed in the past. For the dyeing of these blends, see [427]. Of the worldwide consumption of PES fibers, 55 – 60 % is used in blends with cellulose fibers or wool. About 40 % of polyamide is used in blends, and for polyacrylonitrile fibers, 50 % is used especially in blends with wool for knitwear. Thus, the introduction of man-made fibers has resulted in rapid expansion of the mixed processing of various types of fibers. Of the large number of such blends, some have proved very worthwhile, especially PES – cellulose blends.

Since efforts are made to place the dyeing step as close as possible to the end of the finish-

ing process, the problem of coloring fiber blends arises frequently. In the joint dyeing of two components, a compromise often must be made with regard to the dyeing process, tone-on-tone dyeing or fastness of the finished product.

For this reason, not every fiber blend is advantageous from the point of view of finishing, and satisfactory dyeing methods still have to be found for some blends (e.g., PES – cellulose acetate).

11.1. Polyester – Cellulose Blends

The most important fiber blend consists of polyester and cellulose fibers. A large part of the entire production of PES, ca. 45 %, is used to make this mixture. Polyester – cellulose blends are used for all types of clothing textiles and for bed linen. A considerable amount is consumed by the knitwear sector. The cellulose component is usually cotton, but viscose staple fibers, modal fibers (→ Cellulose), and occasionally linen are also used. Modal fibers are often employed in knitwear because of their luster. The preferred mixing ratio is PES:cellulose 67:33 and, for textiles worn close to the skin, 50:50 to 20:80.

Articles produced from a mixture of cellulose and PES have more versatile application than those made of either cellulose or PES fibers alone. Cellulose with its hydrophilicity, its ability to transport moisture, and its low electrical charging contributes to wearing comfort. The PES fiber improves tear and abrasion resistance, increases dimensional stability, and makes washing and care easier.

11.1.1. General Finishing Information

For detailed information on the dyeing and finishing of PES – cellulose blends, see [428, pp. 209 – 289], [429, pp. 453 – 457], [430, pp. 592 – 600].

As in the case of cellulose articles, the dyeability of PES – cellulose mixtures is influenced greatly by the methods used for pretreatment, the boiling and bleaching of the cellulose component, and the heat setting of the PES component. Pretreatment methods depend especially on the type of cellulose fiber employed. In addition to the shades to be dyed and the scheduled dyeing method, tests must also be conducted to determine the required degrees of whiteness and hydrophilicity (→ Textile Auxiliaries, Chap. 4.2.).

In dyeing PES – cellulose mixtures, disperse dyes are used for the PES component. The cellulose portion can be dyed with practically all classes of dyes suited to cellulose (see Chap. 4). The selection of suitable dyes depends not only on the equipment available, the desired shade, and the batch size, but also on economic factors, with the required fastness taken into account.

Thus, *reactive dyes* are used preferentially for all types of clothing textiles, but *vat dyes* are employed where extreme demands are made on washfastness or weatherproofness.

Dyeing with mixed dyes that were first put on the market by BASF is especially easy. Examples of suitable mixtures of disperse and vat dyes are Cottestren (BASF) and Teracoton (Ciba – Geigy). The mixed dyes can be applied in a single bath via the exhaustion or continuous process. The mixed dyes are formulated in such a way that the same shade is obtained on both fibers, which facilitates shade matching [431].

In dyeing PES and cellulose components with two dye classes, the staining of one fiber by the dye used for the other fiber should be taken into account. *Disperse dyes* stain cellulose fibers only slightly. The soiling can be removed easily by washing or, in stubborn cases, by alkaline reductive treatment, which is required anyway in vat dyeing. Most of the dyes used for cellulose stain PES only slightly or not at all. Dyes such as selected vat dyes that are fixed to PES by a special application process (see Section 7.2) and sulfur dyes, which often dye PES deeply, exhibit good fastness in use and produce a shade that corresponds with the tint on cellulose.

In the case of light shades, a slight staining of the partner fiber can give an attractive appearance to the entire article. Therefore, for pastel shades, depending on the article and its design, only one component may be dyed, provided that the staining on the other is fast enough. *Leuco esters of vat dyes* are also suitable for the simultaneous fast dyeing of PES and cellulose in pastel shades. The mode of dyeing is simple (see Section 4.8.1). To obtain good fastness on the PES component, drying should be carried out at 170 – 180 °C. *Pigment dyeing* is also commonly used for light shades (see Section 4.7).

For matching off the two fibers, the components often must be inspected separately. Fiber blends often look better being dyed tone-on-tone if the glossy fiber, PES in this case, is kept somewhat darker. To inspect the dyeing of the components, the cellulose in PES – cellulose blends can be dissolved out by treatment with sulfuric acid (72 %, 45 min at room temperature or 5 min at 70 °C) or orthophosphoric acid (85 %, 3 – 5 min at 80 – 90 °C), which is facilitated by ultrasonic treatment. Furthermore, small samples of the individual components in a presentation similar to the material to be dyed can be dyed for inspection together with the fiber mixture.

11.1.2. Dyeing Processes

Since safe and efficient continuous dyeing methods are well known and large yardages are often dyed in the same color, PES – cellulose mixtures are frequently dyed in continuous processes. Nevertheless, the batch dyeing is of major importance for dyeing yarn as well as piece goods, especially in the knitwear sector.

11.1.2.1. Dyeing with Disperse and Vat Dyes or Appropriate Mixed Dyes

Batch Dyeing Process. In batchwise dyeing, a single bath is used in two stages. A suitable choice of dyes produces dyeings with unsurpassed fastness to washing (e.g., for work clothes) and thermofixing (e.g., for colored weaving yarn). Mixed dyes are usually prepared in such a way that PES – cellulose blends mixtures of ca. 80:20 to 50:50 can be dyed in the same shade. The PES component is dyed at ca. 130 °C without a carrier since carriers can hold back vat dyes in the dyebath (for mode of dyeing, see Section 7.2.2). During the HT step, the cellulose portion is uniformly covered with vat dye. After the PES component is dyed, the bath is cooled to 80 °C and made alkaline, the vat dye is vatted with hydrosulfite, and dyeing is completed according to the rules for vat dyeing (see Section 4.3.6). The disperse dye attached to the cellulose portion is removed by means of the vat so that a reductive clearing step to improve fastness becomes unnecessary. The mode of operation in individual dyeing aggregates is described in Section 2.2.

Continuous Dyeing Process. For PES – cellulose blends, continuous dyeing with disperse and vat dyes is especially safe, easy, and reproducible and produces very fast dyeings. Migration of the dye pigment during drying is a frequent source of faults. In PES – cellulose blends, the cellulose fiber is the absorptive component and counteracts uninhibited migration. In addition, active migration inhibitors (e.g., polyacrylates), which can be washed out easily later, are available. In continuous dyeing, the materials must be carefully pretreated to guarantee rapid wetting and uniform absorption of liquor (→ Textile Auxiliaries, Chap. 4.2.).

After dye application, the fabric is uniformly and gently dried and the disperse dye is fixed to the PES component by the thermosol process (see Section 7.2.3). Subsequently, the vat dye on the cellulose fiber portion is developed either continuously according to the pad steam process or batchwise in a jigger (see Section 4.3.6). *Jigger development* is performed according to the pad jig process for vat dyes and is advantageous with respect to penetration into nap containing or densely woven materials. The most elegant method is *pad steam development*, which frequently follows the thermosol process in one pass. In pad steam development, the fabric is padded with reducing agent and alkali, and the vat dye is fixed by steaming in saturated steam (103 – 105 °C dry thermometer). Finely dispersed vat dyes must be used in this process because of the short vatting time. The mode of operation with individual dyeing aggregates is described in Section 2.3.2.

11.1.2.2. Dyeing with Disperse and Reactive Dyes

Reactive dyes are often used to dye the cellulose portion of PES – cellulose blends because the coloring possibilities of reactive dyes are especially diverse and a large number of relatively easy application processes produce dyeings that are sufficiently fast for the clothing sector and, frequently, for household textiles.

Batch Processes. Since the worldwide tendency is toward smaller batch sizes, PES – cellulose blends are dyed mainly batchwise by the exhaustion method.

The *two-bath method* is a safe mode of dyeing that results in reproducible dyeings. The PES component is usually predyed with disperse dyes, and the cellulose portion is topped with reactive dyes after reductive intermediate clearing, if necessary.

In a *one-bath two-stage process*, the bath contains all the dyes from the start. After the PES component is dyed at 130 °C, the bath is cooled to ca. 80 °C, and alkali is added, resulting in the fixation of the reactive dyes to the cellulose component. A careful selection of dyes is required because the reactive dyes must be stable at pH 4 – 5 and 130 °C.

In a *reverse process*, the cellulose component is first predyed at alkaline pH, the pH is then lowered, and the PES component is dyed at 130 °C. In this case, alkali-stable disperse dyes must be chosen if the disperse dye is added at the start. In addition, large amounts of salt in the bath can interfere with dispersion stability. Hence, especially fast exhausting disperse dyes are used.

Alkali dosing processes with controlled addition of alkali are of great importance for dyeing the cellulose component, particularly when vinylsulfone anchor dyes are used. The control program can also allow a progressive increase in pH. Thus, the rate of absorption, which is responsible for leveling, and the rate of fixation are optimized. Occasionally, the pH is controlled by cleavable alkali-releasing chemicals.

Of little importance is a *thermosol exhaustion process* in which the PES component is predyed by the thermosol method and the cellulose portion is then topped in a bath dyeing process. Also of little significance is a *pad batch HT process* in which disperse dye, reactive dye, and chemicals are applied together in the padder. The reactive dye is then fixed in a pad batch process, and the disperse dye is fixed in an HT step. Here, too, alkali-stable disperse dyes are required.

Continuous Processes. The most important continuous dyeing methods are the thermosol pad steam process and the thermosol pad batch process. In all continuous processes, the substantivity of the reactive dyes should be taken into account (i.e., the reactive dye can exhaust from the pad liquor or accumulate to a higher or lesser extent even at room temperature as a function of the pH value and salt content).

In the *thermosol pad batch process*, a semicontinuous process, alkali-sensitive disperse dyes can be used if the alkali is applied after the thermosol process. Conversely, if the thermosol process is carried out after the pad batch step, the pH must be lowered before the thermosol process so that the disperse dye is not destroyed.

The *thermosol pad steam process* is especially important. Disperse and reactive dyes are padded together, and alkali and salt are padded after the thermosol passage. If a steamer is not available, the reactive dye can be fixed continuously by means of "alkali shock" in a roller vat with a relatively high alkali concentration.

Although the *thermosol thermofixing process* is especially easy and economical, it is less easily reproducible and, therefore, of little importance. In this process, disperse and reactive dyes are padded together with alkali and fixed by dry heat at ca. 210 °C. This is followed by washing in an open-width washing machine and soaping. Disperse dyes used for this process must be alkalistable, and the dispersing agent contained in the dye must not react with the reactive dye. Especially suitable reactive dyes are hot dyers that are stable and poorly substantive in the pad liquor.

However, none of the one-step processes with disperse and reactive dyes achieves the safety of operation of the thermosol pad steam process with disperse and vat or sulfur dyes.

11.1.2.3. Dyeing of the Cellulose Component with Other Dyes

Sulfur Dyes. The thermosol pad steam process is the primary method for application of disperse and sulfur dyes. However, the sulfur dyes that are used preferentially in the ready-for-dyeing soluble form, e.g., Cassulfon (Hoechst) and Sodyesul (Southern Dyestuff), and are specially inexpensive for muted shades, must be applied after the thermosol step because of their incompatibility with disperse dyes.

Leuco Esters of Vat Dyes. If a thermosol passage is provided under conditions that cleave the ester and the vat acid can diffuse into the PES fiber, leuco esters of vat dyes can produce up to medium shades with a high fastness level without the additional use of disperse dyes (ATE process, Hoechst).

Small Molecular Vat Dyes. Only one class of dyes is needed when small molecular vat dyes

are used. Fast dyeings are obtained in lighter shades. Padding and the thermosol process are, if necessary, followed by vatting and normal aftertreatment to improve fastness on the cellulose component.

Direct Dyes. If no great fastness demands are made on the dyed article, PES – cellulose blends can be dyed in a one-bath, one-step or two-step process with disperse and direct dyes. The direct dyes selected here must be stable under HT dyeing conditions. Dyes whose wetfastness can be improved by aftertreatment are of advantage. The subsequent resin finishing must also be taken into account in dye selection.

In a two-bath, two-step dyeing process, other dyes suited to cellulose fibers may also be used for PES – cellulose blends. For instance, in spite of its tediousness, *Naphtol AS dyeing* is still employed, especially for wetfast red shades (see Section 4.6), and the *phthalogen process* is used for turquoise (see Section 4.8.7). *Pigment dyeing* is also applied to PES – cellulose blends (see Section 4.7).

11.2. Polyester – Wool Blends

Mixtures of PES fibers and wool are used widely, especially as woven goods and knitwear in the outerwear sector. They offer the advantages of dimensional stability, easy care, and fashionable design possibilities. PES is generally used as staple fiber and frequently as a low-pilling type [432]. The ratio PES:wool 55:45 is widely employed.

In methods used for pretreating and dyeing PES – wool blends, the relatively high sensitivity of wool must be taken into account. Difficulties are circumvented by dyeing the components separately as loose stock or slubbing (see Chaps. 5 and 7). Especially fast dyeings are obtained in this manner. However, quick changes in fashion and short-term planning frequently do not permit separate dyeing.

11.2.1. Dyes

In dyeing PES – wool mixtures, disperse dyes are used for the PES component and acid or metal-complex dyes for the wool. Disperse dyes can soil wool to a great extent. Since they produce poorly fast dyeings on wool, the dyes selected must stain wool as slightly as possible or must be easily removable by a washing step, which may be reductive if necessary. Frequently used dyes are C.I. Disperse Yellow 23, 54, 64, C.I. Disperse Orange 30, 33, C.I. Disperse Red 50, 60, 73, 91, 167, 179, and C.I. Disperse Blue 56, 73, 87. Premixed dyes consisting of disperse and wool dyes are occasionally available, e.g., Forosyn (Sandoz).

Wool is stained less at high temperatures and longer dyeing times because the disperse dye that was initially on the wool component migrates to the PES fiber. Carrier-free dyeable PES fibers should preferably be employed; these fibers are easily dyeable even at boiling temperature (see Section 7.2.6). The nature of the wool also plays a role, fine wool and reclaimed wool being more strongly soiled.

11.2.2. Dyeing Processes

For practical information about the dyeing of PES – wool mixtures, see [428, pp. 291 – 328], [433]. Wool cannot be dyed at the high temperatures normal in the HT dyeing of PES fibers or PES – cellulose blends. The dyeing time should also be as short as possible so that the wool is not damaged. The following dyeing processes are usually employed:

1) At boiling temperature with a carrier
2) At 103 – 106 °C with little carrier
3) At 110 – 115 °C with the addition of formaldehyde as a wool protectant and without a carrier (HT dyeing)

The significance of HT dyeing has increased in more recent times.

The dyeing of PES – wool blends is carried out in one or two baths. In the *one-bath process* preferred in practice, PES and wool are dyed simultaneously at pH 4 – 5, which is favorable for wool. Reductive aftertreatment is possible only if the wool dyes used can withstand it.

The *two-bath process* is applied for deep shades and for stringent fastness requirements. Disperse dyes, which soil wool to a somewhat greater extent, can also be used in this case. The first step involves the use of disperse dyes for predyeing the PES part, followed by reductive

intermediate clearing and wool dyeing. However, while dyeing the wool component, dye must always be expected to detach from PES and retint the wool.

Prebleached wool becomes somewhat yellow on dyeing. To obtain clear pastel shades bleach and dye are used preferably in the same bath. Mildly bleaching reducing agents (especially stabilized bisulfite or hydroxylamine salts) can be employed because they do not affect the dyes used.

Continuous or semicontinuous dyeing processes are uncommon because dry heat treatment (thermosol) hardens wool. Besides, application of the wool dye requires dyeing in aqueous liquor or dye fixation by steaming.

11.2.3. Auxiliaries

At dyeing temperatures > 100 °C, formaldehyde (5 % on weight of fabric) is added to protect the wool. Complexing agents should be used with care in the dyebath if metalcomplex dyes are present. Ethoxylated fatty amines or ethoxylated fatty alkyl sulfates are suitable leveling dyeing aids. Condensation products of naphthalenesulfonic acid and formaldehyde, which hardly foam, have proved especially useful dispersing agents. A leveling agent is required primarily if tippy-dyeing grades of wool are present.

In the selection of a carrier, technical as well as ecological factors must be considered (→ Textile Auxiliaries, Chap. 5.5.). As in the dyeing of pure PES fibers, the amount of carrier to be applied is determined by the type of carrier, the dyes, the type of fiber, the dyeing temperature, the depth of shade, the liquor ratio, and the dyeing aggregate.

If possible, yarn is dyed at 106 °C or still higher temperatures because, in this way, the production fastness (e.g., fastness to setting) required for further processing is easier to guarantee. This is possible because thermally stable disperse dyes that exhaust only at higher temperatures are used.

To separately inspect the PES component of PES – wool mixtures, the wool can be dissolved out with sodium hydroxide (5 % solution, boil for 5 min) or a chlorine bleach (50 g/L active chlorine, 5 min at 50 °C).

Aftertreatment. After dyeing, an afterwash is always indicated to remove the disperse dye attached to the wool. The materials are treated with an ethoxylated fatty amine in a weakly acidic liquor at 60 °C. Reductive aftertreatment is possible only in the case of appropriately stable wool dyes.

11.3. Polyamide – Cellulose Blends

Polyamide (PA) fibers are included in articles made of cellulose fibers to improve dimensional stability, ease of care, and durability. Examples of articles include sportswear and knitwear, corduroy fabrics, or plush articles in which a PA pile is often anchored to a cellulose fabric base. Different dyeing methods are described in [429, pp. 433 – 437], [430, p. 570].

Dyeing Processes. In dyeing PA – cellulose mixtures, consideration should be given to the fact that PA fibers are more or less strongly dyed by all dyes normally used for cellulose. The shades on the PA and cellulose components often do not correspond in color strength and in shade. Their fastness can also vary. Since good color matching between the PA and cellulose components is usually required, a careful dye selection must be made and the dyeing process adapted appropriately.

If dyeing is carried out in the presence of 1 % sodium dihydrogenphosphate without salt at boiling temperature, solid shades from light to moderate can be obtained with high-leveling *direct dyes* (e.g., C.I. Direct Yellow 44, Direct Red 81, and Direct Black 51). The choice of suitable dyes is greater if dyeing is performed in a single bath with *direct and disperse dyes* at pH 8 in the presence of a leveling agent and a resist for the PA fiber (→ Textile Auxiliaries, Chap. 5.6.3.) to prevent the direct dyes to exhaust mainly on the PA component.

If, instead of disperse dyes, wetfast *acid or metal-complex dyes* are used for the PA component, union shades having good lightfastness and moderate wetfastness can be obtained. Dyeing is carried out at pH 5 – 8, depending on the depth of shade and the type of acid dye used. The dyeing temperature is raised to boiling to

guarantee good leveling and firm fixation of the wetfast acid dyes in the PA fiber.

Dyeings of the highest wetfastness are obtained on PA – cellulose blends with *vat dyes*. The coloring strength on the PA component depends on the type of PA fiber and on dyeing conditions (the depth of shade on PA 6 is deeper than on PA 66). The lightfastness of some vat dyes can be lower on PA than on cellulose fibers. The advantage of using vat dyes is that in a single bath with selected products, dyeings of the same shade are obtained on both components. For instance, C.I. Vat. O. 26, V.B. 4 and 14, V. Br. 55, V. Bl. 9 and 25 are widely applicable (in the exhaust and pad steam process). The dye yield on the PA component increases with increasing temperature. With a dyeing time of 60 min, the most favorable temperature for the dyes mentioned above is 80 – 90 °C. The additives to the dyebath and the dyeing processes are the same as those used for dyeing cellulose fibers with vat dyes (see Section 4.3).

Clear shades with very good fastness are obtained on PA – cellulose blends with *reactive dyes*. Like vat dyes, the depth of shade of reactive dyes depends relatively strongly on the type of PA and structural differences. Dyeing is carried out in a three-step process with appropriately selected products. First, the reactive dyes in a weakly acidic liquor are allowed to absorb to the PA component. Salt is then added to improve the yield on the cellulose component. Finally, the liquor is made alkaline for reaction with the cellulose fiber. Dyes (e.g., with monochlorotriazine anchor) that dye PA from a neutral liquor in the presence of salt are applied in a two-step process, as in the case of cellulose. In the reversal of this dyeing process, the cellulose component is dyed first at alkaline pH, followed by neutralization with acid, and the PA component is then covered at elevated temperature.

Dyeing processes with reactive dyes are described in Section 4.1.

11.4. Polyacrylonitrile – Cellulose Blends

Polyacrylonitrile (PAC) – cellulose blends are used for household textiles and imitation fur. In these plush and fur materials, the pile consists of PAC fibers while the back is made of cotton or cellulose regenerated fibers. The wool-like character, favorable price, and easy care of PAC fibers combine with the high water-absorbing capacity of cellulose fibers. Furthermore, PAC – cellulose mixtures are used for leisure wear, sport stockings, drapery, and table linen. The percentage of PAC in the mixtures varies between 30 and 80 %. For a survey of dyeing methods, see [429, pp. 474 – 476] and [430, pp. 608 – 610].

Continuous dyeing of PAC – cotton plush with *cationic* and *direct dyes* according to the pad steam process plays an important role. The choice of dyes must take into account liquor stability, reservation of PAC or cellulose fiber, and solubility. Precipitation of cationic and anionic dyes present in the pad liquor at relatively high concentrations cannot be avoided solely by dye selection. Suitable auxiliary systems have been developed. Differently charged dyes are kept in solution separated from each other in two phases by the combination of anionic and nonionogenic surfactants. With the help of fixing accelerators, good penetration of PAC fibers can be achieved in 10 – 15 min with saturated steam at 98 – 100 °C.

With *cationic and vat dyes*, dyeing can be conducted batchwise according to the pigment process in one bath and two steps. Here, the PAC component is dyed with cationic dyes and the cellulose portion is simultaneously pigmented. After being cooled to 70 °C (maximum) and vatted, the Indanthren dye is fixed (see also Section 4.3). Another possibility is to complete the dyeing of the PAC component first with cationic dyes and then add the vat dye to the same bath after cooling.

PAC – cellulose mixtures can also be dyed in a one-bath, two-step process with *reactive and cationic dyes*. Here, too, the PAC component is predyed at pH 5 with cationic dyes, and after the pH is raised, the cellulose component is covered with reactive dyes. Process safety is obviously increased when a two-bath process is employed. For the dyeing of yarn, see [434].

11.5. Polyacrylonitrile – Wool Blends

No other synthetic fiber achieves the wool-like character of PAC fibers. For this reason, PAC fibers are especially suited to the production

of blends with wool for applications in which the woolly character is indispensable. The high water-absorbing capacity of wool and the high strength of PAC fibers complement each other and confer good wearability on articles made of this mixture. Woven and knitted fabrics possess a greater dimensional stability and mechanical stressability than corresponding wool textiles. The very low water-absorbing capacity of the PAC fiber does not have such an unpleasant effect as in textiles made of PAC alone, and the felting tendency is lower than that of wool. The blending ratio of PAC to wool varies from about 20:80 to 80 : 20. The blends are preferred for outer clothing, especially knitwear (e.g., pullovers and socks), and for household textiles.

Dyeing Processes. In principle, three possibilities exist for dyeing PAC – wool mixtures [429, pp. 473 – 474], [430, p. 610]:

1) The two-bath method, i.e., dyeing of wool and PAC in succession in two baths
2) The one-bath, two-step method, i.e., dyeing of wool and PAC in succession in the same bath
3) The one-bath, one-step method, i.e., dyeing of wool and PAC simultaneously

The one-bath, one-step method is preferred for economic reasons. However, problems arise due to the simultaneous presence of cationic and anionic dyes in the dyebath, with the danger of mutual precipitation. Precipitation can be prevented by dispersing auxiliaries and precise dye selection. Very deep shades (e.g., black) are normally dyed in a one-bath, two-step process.

Although only cationic dyes may be used for dyeing the PAC component, metal-complex dyes, chrome dyes, or reactive dyes can be used for wool.

In the case of 1:1 metal-complex dyes, almost no danger of precipitations exists because of intramolecular compensation of the negative charge by the positively charged metal atom. The pH is adjusted to 2.5 – 3 (see also Section 5.1.4). Most cationic dyes can be applied at this low pH value. Only the auxiliaries and the acid are added while the dyebath is being heated. After the initial dyeing temperature (ca. 70 °C) is reached, the cationic dyes are added first. After balancing concentration differences (ca. 5 to 10 min), the anionic dyes are then added. A pH of 4.5 – 5.5 is maintained for acid, chrome, reactive, and 1:2 metal-complex dyes.

In the choice of dyes, attention must be paid not only to the prevention of precipitation but also to the reserve of PAC fibers and wool. Cationic dyes attach first to wool, producing colors of very poor fastness, and then boil over onto the PAC fiber. Even if well-reserving dyes are selected, dyeing must be conducted for a sufficiently long time (60 – 90 min) until the boilover is complete, in order to obtain the absolutely essential good wool reserve.

Apart from chemical problems (precipitation, reserve), the mechanical properties of the materials must also be taken into account. Knitwear made of PAC – wool must not be exposed to excessively high tensile stress during dyeing.

11.6. Blends of Synthetic Fibers

No fiber exists that has not at least on a trial basis been combined with every other fiber to attain new fashionable or technical effects (see survey in [429, pp. 420 – 421]). From a dyers' standpoint, these mixtures do not always make sense. Practically all synthetic fibers can be dyed with disperse dyes. Thus, the affinities of these dyes for the fiber components and their fastness properties should always be taken into account in dyeing blends of synthetic fibers [435]. In the selection of a dyeing process, the individual components must be considered. With blends of *PA and PES fibers*, for example, one should note that although disperse dyes have a higher affinity for PES than for PA fibers, the rate of dyeing on PES at boiling temperature in the absence of dyeing accelerants is so low that the PA component is usually dyed to a deeper and, frequently, in a different shade. If the same shades are to be produced on a blend of this type, a small selection of disperse dyes is used at the highest possible temperature under exactly stipulated dyeing conditions. To obtain better fastness on the PA component, the disperse dye can then be reductively eliminated to a large extent and the material redyed with acid or metal-complex dyes [428, pp. 359 – 363].

A frequently processed mixture (e.g., for outerwear or furnishing fabrics), consists of *PES and PAC fibers*. No basic difficulties are involved in dyeing, both union shades (tone-on-tone dyeing) and bicolor dyeing being possible. Selected disperse and cationic dyes are used for dyeing. Although disperse dyes color PAC more weakly than PES, the shade produced is the same. In the case of bicolor dyeing, the PAC component must always be kept darker. A dyeing temperature of 106 °C should not be exceeded because of the PAC component. Thus, carriers must be used as a rule. In one-bath dyeing, attention must be paid to the compatibility of the various dyebath additives with each other. Cationic retarders are not used in one-bath dyeing [428, pp. 346 – 355], [430, pp. 601 – 602].

Problems with the compatibility of the dyes used can also arise in dyeing mixtures of *PA and PAC fibers*. Mixtures of this type are occasionally produced for skiing articles and for wool – PA imitations. The cationic dyes used for the PAC component do not dye normal PA fibers. Acid and metal-complex dyes leave PAC fibers practically white. Sulfo group containing 1:1 metal-complex dyes possess a zwitterionic character and, therefore, can be applied in low concentration in the same bath with cationic dyes, without the occurrence of precipitation. In this manner, one-bath dyeing can be carried out. The stability of the bath can be improved by addition of a nonionic auxiliary (e.g., an ethoxylated fatty alcohol). Cationic retarders should be excluded in all cases. Dyeing is accomplished most safely in two steps: first with cationic dyes, because they exhaust almost quantitatively, and then with 1:1 or 1:2 metal-complex dyes for the PA component. Disperse dyes can naturally be used as well. However, a lower wetfastness level must be expected, and a distinct amount will absorb to the PAC component [429, pp. 471 – 473]. For the dyeing of mixtures containing elastomeric fibers, see Section 10.2.

12. Characterization of Textile Dyes

This chapter gives a description of those properties that are of general importance for dyes or groups of dyes. Such a description is based on experimentally determined quality data.

Color Strength, Hue, Chroma (→ Dyes, General Survey, Chap. 3.2., → Dyes, General Survey, Chap. 3.3., → Dyes, General Survey, Chap. 3.4., → Dyes, General Survey, Chap. 3.5., → Dyes, General Survey, Chap. 3.6.). The dominating characteristics of a dye are its coloring properties. These are assessed by preparing a dyed test sample whose color is evaluated. This must, in principle, always be done by the human eye because color perception, being a subjective sense impression, is not accessible to direct measurement. However, with the aid of colorimetry, which can be regarded as a simulation of the visual perception of color, this visual perception can be represented by measurable quantities. Since colorimetry is an objective method and is therefore more accurate and reproducible than subjective visual assessment, it is very widely used today. Color is a three-dimensional quantity and must therefore be expressed by a set of three numbers (color coordinates). In practice, these are typically the values of color strength, hue, and chroma (→ Dyes, General Survey, Chap. 3.4.) [436].

Color can be described with any accuracy only in a comparative way (i.e., in terms of deviations from a standard). Although the three-dimensional color coordinates can be determined by colorimetry (as absolute values), this is in practice not sufficient to provide an exact specification. In the case of a dye, the reference is a standard dye sample provided by the producer against which all further supplies are controlled and standardized. To be used as a basis for assessment the standard sample must be extremely stable.

To test a dye, dyed test specimen are prepared from both the sample dye and the standard under conditions that are as identical as possible and then assessed visually or colorimetrically. In practice, color strength is calculated from various color strength values agreed upon by the producer and the customer. A widely used method is a determination based on the weighted sum of the K/S values (K is the coefficient of absorption, S the coefficient of scatter), which is used in standard specifications [437] (→ Dyes, General Survey, Chap. 3.6.). The relative color strength is calculated as the ratio of the color strength values for the sample dye and the standard expressed as a percentage.

Dyers can in practice compensate for differences in color strength by adjusting the concentration. Therefore, the hue and chroma of a dyeing are always judged against a standard dyeing of equivalent color strength. This is carried out by visual assessment in which the colorist estimates the remaining differences in hue and chroma for a strength-corrected dyeing. In a colorimetric evaluation, the residual color difference is determined by a mathematical correction to give equivalence in color strength. The result is expressed in CIELAB color difference units (DIN 6174) and is usually separated into components for hue ($\Delta H*$) and chroma ($\Delta C*$) [437, 438] (→ Dyes, General Survey, Chap. 3.5.).

In practice, color strength is often determined by the simpler and more easily reproducible measurement of the extinction of the dye solution. However, this does not in all cases agree with the values determined by dyeing because the dye is not always absorbed quantitatively by the fiber [439].

Solution and Dispersion Behavior. For the dyeing process in aqueous liquor, the dye must have adequate solubility or dispersibility. In general, good solubility is necessary for good application properties. If the solubility is poor (i.e., if any of the dye is present in the dye liquor in the form of undissolved particles), local coloration (specks), spots, uneven effects, and poor fastness can be produced, leading to serious defects and costly complaints.

In addition to solubility properties, the stability of the liquor during preparation and use (i.e., the *solution stability*) is important. For reactive dyes applied to cellulose, the solution stability in the presence of electrolytes is important (*electrolyte stability*). To test for this, the dye is dissolved under defined conditions of temperature, time, and concentration, and is filtered after a storage time or after addition of electrolytes. The filter should then be free of residues. To give consistent and comparable results, the values of various parameters (temperature, etc.) should be kept constant and should be specified in standards. For example, the temperatures for hot and cold solubility are set at 90 °C and 25 °C, respectively (ISO/DIS 105-Z07, ISO/DIS 105-Z08, and ISO/DIS 105-Z09) [440].

For *disperse dyes*, the dispersion properties are of similar importance, but they are not easy to describe because of the complex processes and unstable dispersion states. Therefore, a large number of test methods exist, the results of which are usually limited in their application. A critical review can be found in [441].

Dust. The dust produced from dyes is of great importance for many reasons, including possible health hazards and intense coloration, leading to severe soiling, that can be caused by very small amounts of dust particles. The content of fine dust (1 – 10 μm) must be reduced by adhesive bonding or prevented by the use of special formulations (granulation) of the dye. Fine dust is especially troublesome because of the stability of its suspensions in air, and it can constitute a health hazard when inhaled.

The amount of dust depends not only on the dust behavior of the dye, but also on the nature and intensity of handling operations that generate dust [442]. Therefore, the dusting behavior of a product is significant only for comparison with other products and cannot be used to predict dust exposure. In a widely used test method, the dust produced when a defined amount of dye is allowed to fall is determined quantitatively or qualitatively [443]. The dust is first collected on a filter, and assessment occurs by visual comparison of the dust-laden filter with a gray scale divided into five levels (ISO 105-A03) or by quantitative gravimetric or photometric determination of the quantity of dust.

Compatibility of Basic Dyes for Acrylic Fibers. Basic dyes usually have limited migration ability. Therefore, with combination (trichromic) dyeing processes, the dyes are chosen so that they are absorbed at as uniform a rate as possible to prevent local color differences that would otherwise occur readily.

To characterize combination behavior, dyes are divided into five groups numbered 1 to 5 (compatibility values). Higher numbers correspond to a lower bath exhaustion rate. The compatibility of a dye is determined by producing samples of dyed materials in which the test dye is combined with five standard dyes whose compatibility values are known (AATCC Test Method 141 – 1987; ISO/DIS 105-Z03) [444].

13. Testing of Colorfastness [445]

In the 1980s most of the standard methods for testing colorfastness were revised, and some new methods were introduced.

13.1. Organization

Questions of standardization of test methods and of the availability of apparatus and reference materials used in testing colorfastness are treated in Germany by the Working group 511 of the Normenausschuß Materialprüfung at DIN. Working group 511 is called the "Deutsche Echtheitskommission" (DEK) and was founded in 1911. About 70 DIN standards exist describing test methods for colorfastness.

In the European Colourfastness Establishment (ECE), Germany is represented by DEK. Proposals of new test methods to the ISO are prepared by ECE.

In the ISO, Technical committee 38 (Textiles), Subcommittee 1 (TC 38/SC1) is engaged with questions of colorfastness testing, color measurement, and dye characterization. Until now 16 meetings of ISO/TC 38/SC1 have taken place.

Since 1978 the test methods have been published as *International Standard ISO 105*; each separate part is characterized by code numbers, for instance, ISO 105-B02 for testing the "Colour fastness to artificial light: Xenon arc lamp fading test." Since 1978, International Standard 105, which contains more than 70 different test procedures, has been revised, and the different parts have been given code numbers. These standards are published in English, French, and Russian.

The latest, very important development is the edition of *European Standards* (*EN*), which are published in English, French, and German. For testing colorfastness, these European standards correspond completely with ISO 105, which became EN 20 105 with the same code numbers (i.e., ISO 105-B02 becomes EN 20 105-B02). Since the editorial work takes considerable time, publishing the European standards will take several years. At the moment, very few EN denominations exist. When the European standards are published the national standards (e.g., DIN standards) are withdrawn.

13.2. Standard Test Methods

The following summary for each test method gives the denomination of the international standard and the title and denomination of the German DIN standard given.

13.2.1. Group A: General Principles

ISO 105-A01: General Principles of Testing (= DIN 54 000-A01: Grundlagen für die Festlegung und Durchführung der Prüfungen und für die Bewertung der Prüfergebnisse). This part of ISO 105 provides general information about methods for testing the colorfastness of textiles for the guidance of users. The uses and limitations of the methods are pointed out, several terms are defined, an outline of the form of the methods is given, and the contents of the clauses constituting the methods are discussed briefly.

"By colour fastness is meant the resistance of the colour of textiles to the different agencies to which these materials may be exposed during manufacture and their subsequent use. The change of colour and staining of undyed adjacent fabrics are assessed as fastness ratings."

The methods may be used not only for assessing the colorfastness of textiles but also for assessing the colorfastness of dyes. Conditions in the tests have been chosen to correspond closely to the treatments usually employed in manufacture and to the conditions of ordinary use. At the same time, they have been kept as simple and reproducible as possible.

ISO 105-A02: Grey scale for assessing change in colour (= DIN 54 001-A02: Herstellung und Handhabung des Graumaßstabes zur Bewertung der änderung der Farbe) describes the gray scale for determining changes in the color of textiles in colorfastness tests, and its use. A precise colorimetric specification of the scale is given.

ISO 105-A03: Grey scale for assessing staining (= DIN 54 002-A03: Herstellung und Handhabung des Graumaßstabes zur Bewertung des Anblutens) describes the gray scale for determining staining of adjacent fabrics in color-

fastness tests, and its use. A precise colorimetric specification of the scale is given.

ISO 105-A04: Method for the Instrumental Assessment of the Degree of Staining of Adjacent Fabrics (DIN standard in preparation). This part of ISO 105 specifies an instrumental method for assessing the degree of staining of adjacent fabrics in any fastness test, as an alternative to the visual method.

13.2.2. Group B: Colour Fastness to Light and Weathering

ISO 105-B01: Colour Fastness to Light: Daylight (= DIN 54 003-B01: Bestimmung der Lichtechtheit von Färbungen und Drucken mit Tageslicht). A specimen of the textile is exposed to daylight under prescribed conditions, including protection from rain, along with eight dyed blue wool references. Colorfastness is assessed by comparing the change in color of the specimen with that of the references. Five different methods are explained. In an annex, general information on colorfastness to light is given.

ISO 105-B02: Colour Fastness to Artificial Light: Xenon Arc Fading Lamp Test (= DIN 54 004-B02: Bestimmung der Lichtechtheit von Färbungen und Drucken mit Xenonbogenlicht). A specimen of the textile is exposed to artificial light under prescribed conditions, along with blue wool references. Two different sets of blue wool references exist, which are not interchangeable. Colorfastness is assessed by comparing the change in color of the specimen with that of the references used. For white (bleached or optically brightened) textiles, the fastness is assessed by comparing the change of whiteness with that of the references used. Five different methods are described.

ISO 105-B03: Colour Fastness to Weathering: Outdoor Exposure (= DIN 54 070-B03: Bestimmung der Wetterechtheit von Färbungen und Drucken in freier Atmosphäre). Specimens of the textile are exposed under specified conditions in the open air without any protection from weathering. At the same time and place, eight dyed wool references are exposed to daylight, but they are protected from rain, snow, etc., by a sheet of glass. Fastness is assessed by comparing the change in color of the specimen with that of the references. Two different methods are described.

The wide variations in conditions under which outdoor exposures are usually carried out make replicate exposures at different times of the year desirable. The most reliable indication of weathering fastness is obtained by taking the mean of the assessment of several exposures.

ISO 105-B04: Colour Fastness to Weathering: Xenon Arc (DIN 54 071-B03: Bestimmung der Wetterechtheit von Färbungen und Drucken durch künstliche Bewetterung in Xenonbogenlicht). Specimens of the textile are exposed under prescribed conditions to light from a xenon arc lamp and to water spray. At the same time, eight dyed wool references are exposed to the same light source but are protected from water spray by a sheet of window glass. Fastness is assessed by comparing the change in color of the specimen with that of the references. Two different methods are described. This test procedure can also be used for detection of wet light-sensitive textiles.

ISO 105-B05: Detection and Assessment of Photochromism (= DIN 54 069-B05: Verfahren zur Erkennung und Bestimmung der Photochromie von Färbungen und Drucken). A specimen of the textile is exposed to light of high intensity for a time much shorter than that necessary to cause a permanent change. The change in color of the specimen is assessed immediately after exposure, by using the gray scale. The specimen is then stored in the dark and assessed again.

13.2.3. Group C: Colour Fastness to Washing and Laundering

ISO 105-C01: Colour Fastness to Washing: Test 1 [= DIN/EN 20 105-C01: Bestimmung der Waschechtheit von Färbungen und Drucken Wäsche 40 °C (Waschprüfung ISO 105-C01)]. This part of ISO 105 specifies Test no. 1 of a series of five washing tests that have been established to investigate the fastness to washing of colored textiles. The five tests cover

Table 16. Conditions of colorfastness to washing tests according to ISO

ISO Designation	Test no.	Corresponding DIN/EN designation	Time, min	Temperature, °C
ISO 105-C01	test 1	(DIN/EN 20 105-C01)	30	40
ISO 105-C02	test 2	(DIN/EN 20 105-C02)	45	50
ISO 105-C03	test 3	(DIN/EN 20 105-C03)	30	60
ISO 105-C04	test 4	(DIN/EN 20 105-C04)	30	95
ISO 105-C05	test 5	(DIN/EN 20 105-C05)	240	95

the ranges of washing procedures from mild to severe.

The five tests are designed to determine the effect of washing only on the colorfastness of the textile. The method is not intended to reflect the result of the comprehensive laundering procedure. A specimen of the textile in contact with one or two specified adjacent fabrics is agitated mechanically under specified conditions of time and temperature in a soap solution, then rinsed, and dried. The change in color and the staining of the adjacent fabric(s) are assessed with the gray scales.

The description of the test procedure for test 1 applies to the four other tests; all five are distinguished by the specified conditions of time and temperature as stated in Table 16.

ISO 105-C06: Colour Fastness to Domestic and Commercial Laundering (= DIN 54 017-C06: Prüfung der Waschechtheit von Färbungen und Drucken bei der Haushaltswäsche und bei dem gewerblichen Waschen). The test methods in this part of ISO 105 are intended to reflect the effect of comprehensive laundering by either domestic or commercial procedures, as distinct from the washing test methods given in ISO 105-C01 to C05.

A specimen of the textile in contact with specified fabric or fabrics is laundered, rinsed, and dried. Specimens are laundered under appropriate conditions of temperature (40, 50, 60, 70, and 95 °C), duration (30 or 45 min), alkalinity (adjustment of pH to pH 10.5 or no adjustment), bleaching (none or 1 g/L sodium perborate, none or 15 mg/L available chlorine), and abrasive action such that the result is obtained in a conveniently short time. The abrasive action is accomplished by the use of a low liquor ratio and an appropriate number of steel balls (10, 25, 50, or 100). The change in color of the specimen and the staining of the adjacent fabric(s) are assessed with the gray scales.

13.2.4. Group D: Colour Fastness to Dry Cleaning

ISO 105-D01: Colour Fastness to Dry Cleaning (= DIN 54 024-D01: Bestimmung der Trockenreinigungsechtheit von Färbungen und Drucken). A specimen of the textile in contact with a cotton fabric bag together with noncorrodible steel disks is agitated in perchloroethylene, then squeezed or centrifuged, and dried in hot air. The change in color of the specimen is assessed with the gray scale. Subsequently, the coloration of the solvent is assessed by comparing the filtered solvent with unused solvent by transmitted light, via the gray scale for assessing staining.

ISO 105-D02: Colour Fastness to Rubbing: Organic Solvents (DIN standard in preparation). A specimen of the textile is rubbed with cotton cloth impregnated with solvent. The change in color of the specimen and the staining of the cotton rubbing cloth are assessed with the gray scales.

13.2.5. Group E: Colour Fastness to Aqueous Agencies

ISO 105-E01: Colour Fastness to Water (= DIN 54 006-E01: Bestimmung der Wasserechtheit von Färbungen und Drucken). A specimen of the textile in contact with one or two specified adjacent fabrics is immersed in water, drained, and placed between two plates under a specified pressure in a test device for 4 h at 37 °C. The specimen and the adjacent fabric(s) are dried separately. The change in color of the specimen and the staining of the adjacent fabric(s) are assessed with the gray scales.

ISO 105-E02: Colour Fastness to Sea Water (= DIN 54 007-E02: Bestimmung der Meerwasserechtheit von Färbungen und Drucken).

This test is performed in the same manner as colorfastness to water but a 30-g/L aqueous sodium chloride solution is used instead of water.

ISO 105-E03: Colour Fastness to Chlorinated Water (Swimming-Bath Water) (= DIN 54 019-E03: Bestimmung der Farbechtheit von Färbungen und Drucken gegenüber gechlortem Wasser). A specimen of the textile is treated with a weak aqueous chlorine solution of a given concentration and dried. The change in color of the specimen is assessed with the gray scale.

Three alternative test conditions are specified: Active chlorine concentrations of 50 mg/L and 100 mg/L are intended for swimwear. An active chlorine concentration of 20 mg/L is intended for accessories such as beach robes and towels.

ISO 105-E04: Colour Fastness to Perspiration (= DIN 54 020-E04: Bestimmung der Schweißechtheit von Färbungen und Drucken). Specimens of the textile in contact with adjacent fabrics are treated in two different solutions containing histidine, drained, and placed between two plates under a specified pressure in a test device for 4 h at 37 °C. The specimens and the adjacent fabrics are dried separately. The change in color of each specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-E05: Colour Fastness to Spotting: Acid (= DIN 54 028-E05: Bestimmung der Säureechtheit von Färbungen und Drucken). Drops of a solution of acid (acetic acid, sulfuric acid, tartaric acid) are placed on the specimen, whose surface is rubbed gently with a glass rod to ensure penetration. The changes in color of the textile, while it is still wet and after drying, are assessed with the gray scale.

ISO 105-E06: Colour Fastness to Spotting: Alkali (= DIN 54 030-E06: Bestimmung der Alkaliechtheit von Färbungen und Drucken). Drops of a solution of sodium carbonate are placed on the specimen, whose surface is rubbed gently with a glass rod to ensure penetration. The change in color of the textile is assessed with the gray scale.

ISO 105-E07: Colour Fastness to Spotting: Water (= DIN 54 008-E07: Bestimmung der Wassertropfenechtheit von Färbungen und Drucken). Drops of water are worked into the specimen with a glass rod, so that a spot approximately 20 mm in diameter is formed. The change of color of the textile is assessed with the gray scale after 2 min and after drying.

ISO 105-E08: Colour Fastness to Water: Hot Water (= DIN 54 047-E08: Bestimmung der Heißwasserechtheit von Färbungen und Drucken). A specimen of the textile in contact with adjacent fabrics is rolled around a glass rod, treated with slightly acidified hot water (pH 6 at 70 °C), and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales. The method is applicable mainly to wool and to textiles containing wool.

ISO 105-E09: Colour Fastness to Potting (= DIN 54 048-E09: Bestimmung der Pottingechtheit von Färbungen und Drucken). A specimen of the textile between adjacent fabrics is rolled around a glass rod and treated with boiling water. The specimen and adjacent fabrics are dried separately. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-E10: Colour Fastness to Decatizing (= DIN 54 054-E10: Bestimmung der Dekaturechtheit von Färbungen und Drucken). A specimen of the textile is wrapped around a perforated cylinder, and steam is passed through it for 15 min. The change in color of the dried specimen is assessed with the gray scale. Correct handling of the method is controlled by use of a test-control specimen tested under identical conditions. Two tests, mild and severe, are given.

ISO 105-E11: Colour Fastness to Steaming (= DIN 54 058-E11: Bestimmung der Dämpfechtheit unter atmosphärischem Druck von Färbungen und Drucken). A specimen of the textile in contact with specified adjacent fabrics is rolled into a cylinder and placed for 30 min in the neck of a flask containing boiling water. The staining of the adjacent fabrics is assessed with the gray scale.

ISO 105-E12: Colour Fastness to Milling: Alkaline Milling [= DIN 54 041-E12: Bestimmung der alkalischen Walkechtheit (schwere Beanspruchung)]. A specimen of the textile in contact with one or two specified adjacent fabrics is milled at 40 °C in a jar containing 50 steel balls and a solution of soap and sodium carbonate. The severity of action is controlled by means of a test-control dyeing milled separately in the same way. After rinsing and drying, the change in color and the staining of the adjacent fabrics are assessed with the gray scale.

ISO 105-E13: Colour Fastness to Acid-Felting: Severe [DIN 54 043-E13: Bestimmung der sauren Walkechtheit von Färbungen und Drucken (schwere Beanspruchung)]. A specimen of the textile in contact with adjacent fabrics is milled in solutions of acetic or sulfuric acid as used under severe conditions in the acid-felting process, rinsed, and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-E14: Colour Fastness to Acid-Felting: Mild [= DIN 54 042-E14: Bestimmung der sauren Walkechtheit (leichte Beanspruchung)]. A specimen of the textile in contact with adjacent fabrics is milled in mineral acid solutions at 60 °C, as used under mild felting conditions in the hatmaking and felt industries, rinsed, and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

13.2.6. Group G: Colour Fastness to Atmospheric Contaminants

ISO 105-G01: Colour Fastness to Nitrogen Oxides (= DIN 54 025-G01: Bestimmung der Stickoxidechtheit von Färbungen und Drucken). Specimens of textiles are exposed to nitrogen oxides in a closed container until either one (dyes with poor color fastness to nitrogen) or three (dyes with good color fastness to nitrogen) test-control specimens exposed simultaneously with the test specimens have changed color to a predetermined extent. The change in color of the textile is assessed with the gray scale.

ISO 105-G02: Colour Fastness to Burnt Gas Fumes (no DIN standard issued up to now). A specimen of the textile and the test-control fabric are exposed simultaneously to oxides of nitrogen from burnt gas fumes until the test-control shows a change in color corresponding to that of the standard of fading. The change in color of the specimen is assessed with the gray scale. If no color change is observed in the specimen after one exposure period or cycle, exposure may be continued either for a specified number of periods or for the number of periods required to produce a specified amount of color change in the specimen.

ISO 105-G03: Colour Fastness to Ozone in the Atmosphere (no DIN standard issued up to now). A specimen and a swatch of test-control sample are exposed simultaneously to ozone in an atmosphere at ambient temperature and relative humidities not exceeding 65 %, until the control sample shows a color change corresponding to that of a standard of fading. This exposure period constitutes one cycle. The cycles are repeated either until the specimen shows a definite color change or for a prescribed number of cycles.

13.2.7. Group J: Measurement of Colour and Colour Differences

ISO 105-J01: Measurement of Colour and Colour Differences (no DIN standard). This part of ISO 105 selects from several options published by the CIE those best suited to the needs of the textile industry whenever the difference in color between two specimens has to be quantified.

ISO 105-J02: Method for the Instrumental Assessment of Whiteness (no DIN standard). This part of ISO 105 specifies a method intended for quantifying the whiteness of textiles, including fluorescent materials.

13.2.8. Group N: Colour Fastness to Bleaching Agencies

ISO 105-N01: Colour Fastness to Bleaching: Hypochlorite [= DIN 54 035-N01: Bestimmung der Hypchlorit-Bleichechtheit

(schwere Beanspruchung)]. A specimen of the textile is agitated in a solution of 2 g/L sodium hypochlorite or lithium hypochlorite, rinsed in water, agitated in a hydrogen peroxide solution or sodium hydrogensulfite solution, rinsed, and dried. The change in color of the specimen is assessed with the gray scale.

ISO 105-N02: Colour Fastness to Bleaching: Peroxide (= DIN 54 033-N02: Bestimmung der Peroxid-Bleichechtheit von Färbungen und Drucken). A specimen of the textile in contact with adjacent fabrics is immersed in the bleaching solution, containing sodium peroxide in concentrations commonly used in textile bleaching processes, rinsed, and dried. Four different bleaching baths are described. The change in color and the staining are assessed with the gray scales.

ISO 105-N03: Colour Fastness to Bleaching: Sodium Chlorite: Mild [= DIN 54 036-N03: Bestimmung der Chlorit-Bleichechtheit von Färbungen und Drucken (leichte Beanspruchung)]. A specimen of the textile in contact with specified adjacent fabrics is treated in a sodium chlorite solution (1 g/L, pH 3.5 at 80 °C for 1 h), rinsed, and dried. The change in color and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-N04: Colour Fastness to Bleaching: Sodium Chlorite: Severe [= DIN 54 037-N04: Bestimmung der Chlorit-Bleichechtheit von Färbungen und Drucken (schwere Beanspruchung)]. A specimen of the textile in contact with specified adjacent fabrics is treated in a sodium chlorite solution (2.5 g/L, pH 3.5 at 80 °C for 1 h), rinsed, and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-N05: Colour Fastness to Stoving (= DIN 54 038-N05: Bestimmung der Schwefelechtheit von Färbungen und Drucken). A specimen of the textile, containing its own mass of soap solution, and a test-control specimen are exposed in an atmosphere containing sulfur dioxide. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

13.2.9. Group P: Colour Fastness to Heat Treatments

ISO 105-P01: Colour Fastness to Dry Heat (Excluding Pressing) (= DIN 54 060-P01: Bestimmung der Trockenhitzefixier- und Trockenhitzeplissierechtheit von Färbungen und Drucken). A specimen of the textile in contact with specified adjacent fabrics is heated by close contact with a medium that is brought to the required temperature (30 s at 150, 180, or 210 °C). The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-P02: Colour Fastness to Pleating: Steam Pleating (= DIN 54 059-P02: Bestimmung der Dämpfechtheit mit überdruck von Färbungen und Drucken). A specimen of the textile in contact with specified adjacent fabrics is steamed under pressure (135, 170, 270 kPa, for 5, 10, or 20 min) and dried. The change in color and the staining of the adjacent fabrics are assessed with the gray scales.

13.2.10. Group S: Colour Fastness to Vulcanization

ISO 105-S01: Colour Fastness to Vulcanization: Hot Air (= DIN 54 065-S01: Bestimmung der Heißluftvulkanisierechtheit von Färbungen und Drucken). A specimen of the textile in air is in direct contact with an (initially) unvulcanized rubber compound. The change in color of the specimen is assessed with the gray scale.

ISO 105-S02: Colour Fastness to Vulcanization: Sulfur Monochloride (= DIN 54 064-S02: Bestimmung der Kaltvulkanisierechtheit von Färbungen und Drucken). A specimen of the textile is exposed to sulfur monochloride vapor under conditions usually occurring during the cold vulcanization of rubber. The change in color of the specimen is assessed with the gray scale before and after neutralizing with aqueous ammonia.

ISO 105-S03: Colour Fastness to Vulcanization: Open Steam (= DIN 54 066-S03: Bestimmung der Heißdampfvulkanisierechtheit von

Färbungen und Drucken). A specimen of the textile is heated in live steam in direct contact with an (initially) unvulcanized rubber compound – the textile material being wrapped in either

1) Sheeting impermeable to steam and water (Method A) or
2) Undyed bleached cotton cloth, but ensuring that live steam is not prevented from infiltration into the specimen (Method B)

The change in color of the specimen and the staining of the adjacent fabric are assessed with the gray scale.

13.2.11. Group X: Colour Fastness to Miscellaneous Agencies

ISO 105-X01: Colour Fastness to Carbonizing: Aluminium Chloride (= DIN 54 044-X01: Bestimmung der Karbonisierechtheit von Färbungen und Drucken mit Aluminiumchlorid). A specimen impregnated with aluminum chloride solution is dried, baked (15 min at 115 °C), rinsed, and neutralized. The method is applicable mainly to wool and textiles containing wool. The changes in color after rinsing, neutralizing, and drying are assessed with the gray scale.

ISO 105-X02: Colour Fastness to Carbonizing: Sulphuric Acid (= DIN 54 045-X02: Bestimmung der Karbonisierechtheit von Färbungen und Drucken mit Schwefelsäure). A specimen impregnated with sulfuric acid solution is dried, baked (15 min at 105 °C), rinsed, and neutralized. The method is applicable mainly to wool and textiles containing wool. The changes in color after rinsing, neutralizing, and drying are assessed with the gray scale.

ISO 105-X04: Colour Fastness to Mercerizing (= DIN 54 039-X04: Bestimmung der Mercerisierechtheit von Färbungen und Drucken). A specimen of the textile in contact with a specified cotton adjacent fabric is treated with a 300 g/L sodium hydroxide solution (5 min at 20 °C), rinsed with 70 °C water and subsequently with cold water, acidified, rinsed again, and dried. The change in color of the specimen and the staining of the cotton adjacent fabric are assessed with the gray scales. Completely resistant specimens may show an apparent increase in depth of color. In such cases an assessment of 5 cannot be given; the fastness number is marked by an asterisk, which must be explained in a footnote.

ISO 105-X05: Colour Fastness to Organic Solvents (= DIN 54 023-X05: Bestimmung der Lösemittelechtheit von Färbungen und Drucken). A specimen of the textile in contact with adjacent fabrics is agitated in the solvent and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-X06: Colour Fastness to Soda Boiling (= DIN 54 031-X06: Bestimmung der Sodakochechtheit von Färbungen und Drucken). A specimen of textile between specified undyed cloths is rolled around a glass rod and treated with boiling sodium carbonate solution for 1 h with and without the addition of a reduction inhibitor. The composite specimen is rinsed and dried. The change in color and the staining of the undyed cloths are assessed with the gray scales.

ISO 105-X07: Colour Fastness to Cross-Dyeing: Wool (= DIN 54 049-X07: Bestimmung der Überfärbeechtheit von Färbungen und Drucken). Specimens of the textile in contact with adjacent fabrics are treated in different types of wool dyebaths, but without any dye. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-X08: Colour Fastness to Degumming (= DIN 54 055-X08: Bestimmung der Entbastungsechtheit von Färbungen und Drucken). A specimen of the textile in contact with adjacent fabrics is treated with a soap solution, similar to those used in degumming raw silk, then rinsed, and dried. The change in color of the specimen and the staining of the adjacent fabrics are assessed with the gray scales.

ISO-105-X09: Colour Fastness to Formaldehyde (= DIN 54 061-X09: Bestimmung der Formaldehydechtheit). A specimen of the textile is exposed in a closed container to the action of gaseous formaldehyde (24 h at 20 °C). The

change in color of the specimen is assessed with the gray scale.

ISO 105-X10: Assessment of Migration of Textile Colours into Polyvinyl Chloride Coatings (= DIN 54 072-X10: Bestimmung der Echtheit von Färbungen und Drucken gegenüber PVC-Weichmachern). A specimen of a textile impregnated with plasticizer is brought into contact with a white-pigmented poly(vinyl chloride) film and kept under pressure for 3.5 h at 80 °C. Then the specimen and excess plasticizer are removed from the film, and the staining of the film is assessed with the gray scale.

ISO 105-X11: Colour Fastness to Hot Pressing (= DIN 54 022-X11: Bestimmung der Bügelechtheit von Färbungen und Drucken). Three methods are described; in each method, hot pressing is performed for 15 s at 110, 150, or 200 °C:

1) *Dry Pressing.* A dry specimen is pressed with a heating device at a specified temperature and pressure.
2) *Damp Pressing.* A dry specimen is covered with a wet cotton adjacent fabric and pressed with a heating device at a specified temperature and pressure.
3) *Wet Pressing.* The upper surface of a wet specimen is covered with a wet cotton adjacent fabric and pressed with a heating device at a specified temperature and pressure.

After hot pressing, the change in color of the specimen and the staining of the adjacent fabric are assessed with the gray scales immediately and again after a period of exposure to air.

ISO 105-X12: Colour Fastness to Rubbing (= DIN 54 021-X12: Bestimmung der Reibechtheit von Färbungen und Drucken). Specimens of the textile are rubbed with a dry rubbing cloth and with a wet rubbing cloth in a suitable testing device. Two alternative sizes of rubbing finger are specified. The staining of the rubbing cloths is assessed with the gray scale.

ISO 105-X13: Colour Fastness of Wool Dyes to Processes using Chemical Means for Creasing, Pleating, and Setting (no DIN standard). A specimen of the textile, treated with the chemical solution, is placed in contact with specified adjacent fabrics and subjected to steam pressing. A comparison specimen, not treated with the chemical solution, is steam pressed simultaneously. The specimens are dried, and any differences between the colors of the two specimens and the staining of the adjacent fabrics are assessed with the gray scales.

ISO 105-X14: Colour Fastness to Acid Chlorination of Wool: Sodium Dichloroisocyanurate (no DIN standard). A specimen of the textile is treated in a formic acid buffer solution to which solutions of sodium dichloroisocyanurate and sodium hydrogensulfite are added successively, and is then rinsed and dried. The change in color of the specimen is assessed with the gray scale.

13.2.12. Group Z: Colourant Characteristics

ISO 105-Z01: Colour Fastness to Metals in the Dyebath: Chromium Salts (= DIN 54 052-Z01): Bestimmung der Farbechtheit gegenüber Chrom-VI-Verbindungen im Färbebad). The difference in color between dyeings made with and without potassium dichromate is assessed with the gray scale.

ISO 105-Z02: Colour Fastness to Metals in the Dyebath: Iron and Copper (= DIN 54 053-Z02: Bestimmung der Farbechtheit gegenüber Eisen(III)- und Kupfer(II)-Ionen im Färbebad). The difference in color between dyeings made in the presence and in the absence of salts of the metal is assessed with the gray scale.

14. Legal Aspects, Toxicology, and Ecology

The textile finishing industry is one of the branches of industry with the highest specific water consumption. According to the German Association of Textile Finishers (Deutscher Textilveredlungsverband, TVI), the German textile finishing industry consumed ca. 65×10^6 m^3 of water in 1992, which is 0.6 % of the total ca. 11×10^9 m^3 water requirement of the mining and processing industries combined. Of this quantity, 88.4 % is discharged as wastewater, and

11.6 % represents evaporation losses. The average water consumption was 146 m^3 per tonne of fiber material. About 90 % of the raw water is self-produced and costs, on average, 0.26 €/m^3; wastewater costs are 1.20 (range 0.32 – 2.75) €/m^3. Together, the fresh water and wastewater costs average 5.2 % of the turnover (only finishing work) [446].

14.1. Water Consumption

The water consumption of the textile finishing industry depends on the finishing process and method, type of fiber, form, and the machines and equipment used. A survey of the amounts of water required in terms of textiles and dyeing equipment is presented in Tables 17 and 18 [447].

Table 17. Water consumption (including pretreatment)

Textiles	Consumption, m^3 per tonne fiber material
Cotton fabric	80 – 240
Cotton woven goods	70 – 180
Woolen fabric	100 – 250
Polyacrylic fabric	10 – 70

Table 18. Liquor ratio of various dyeing equipments

Dyeing equipment	Liquor ratio
Star dyeing apparatus	100:1
Winch beck	10:1 – 30:1
Jigger	3:1 – 5:1
Beam dyeing machine	8:1 – 10:1
Overflow	4:1 – 10:1
Padder	0.6:1 – 0.8:1

In the past, water consumption has been reduced perceptibly by the use of water-saving equipment and continuous processes.

14.2. Legal Regulations

As a result of the intensive use of water in the textile finishing industry, the regulations for wastewater discharge are most important. A distinction must be made between wastewater that is discharged into a municipal sewage system (indirect discharge) and wastewater that is discharged into surface waters (after purification, if necessary). The demands to be met in the second case are greater.

Germany and Austria. The regulations for wastewater that is discharged into surface waters are stipulated in the German Water Resources Act (Wasserhaushaltsgesetz, WHG) [448] and particularly in the decrees and administrative directions given in § 7a. The requirements for discharge are regulated in a branch-specific manner through the ordinance on the source of wastewater [449] and the general administrative regulations [450] with the corresponding appendices. For wastewater that comes largely from the textile finishing industry, the 38th administrative regulation from September 5, 1984, is still valid (end of 1994) (Table 19), but it should be replaced shortly by Appendix 38 of the general administrative regulations [452].

Apart from the minimum requirements for wastewater discharge in keeping with the state of the art and generally accepted technological conventions (Table 19), general demands are also made on the use of certain chemicals, treatment of partial streams, concentrates, and residual amounts of materials used. About 95 % of the German textile finishers are indirect dischargers. Their wastewater discharge is controlled not only by state water laws, but also by local regulations of the concerned sewage treatment plants. The amended § 7a of the WHG represents a reform because the same minimum requirements must be met for dangerous substances by direct and indirect dischargers [453]. (This refers to substances that are assessed as dangerous because of the fear of toxicity, longevity, capability of accumulation, or a carcinogenic, teratogenic, or mutagenic effect.) A similar decree exists in Austria for direct and indirect dischargers [454].

England and Wales. In England and Wales, the Water Law of 1980 [455] gave rise to the National River Authorities (NRA) that establish and enforce quality standards for waters [456, 457].

United States. In the United States, no uniform national or state wastewater standards exist. As of now, industry-specific guidelines established by the EPA do not exist for the textile industry. The minimum requirements are determined by public-owned treatment plants or local officials.

Table 19. Minimum requirements for wastewater direct discharge in Germany

Parameter	38. VwV [a] from 1984 [b]	Appendix 38, generally accepted technological conventions	Appendix 38, state of the art [c]
Filterable substances, mg/L	40		
Settleable substances, mg/L	0.3		
COD, mg/L	280	160	
BOD_5, mg/L	40	25	
Phosphorus, total, mg/L		2	
Ammonium nitrogen, mg/L	5	10	
Nitrogen from nitrate, nitrite and ammonium, mg/L		20	
Iron, mg/L		3	
Aluminum, mg/L		3	
Sulfite, mg/L	1	1	
Hydrocarbons, mg/L	10	10	
Fish toxicity as dilution factor	4		2
Absorbable organic halogens (AOX), mg/L			0.5
Free chlorine, mg/L			0.3
Aromatics (benzene and homologues), mg/L			0.1
Sulfide, mg/L	0.1		1
Chromium(VI), mg/L			0.1
Chromium, total, mg/L	2		0.5
Copper, mg/L	1		0.5
Nickel, mg/L			0.5
Zinc, mg/L	3		2
Color (translucent color value)			
436 nm (yellow)			$7\,m^{-1}$
525 nm (red)			$5\,m^{-1}$
620 nm (blue)			$3\,m^{-1}$

[a] VwV = Administrative regulation.
[b] 2-h mixed sample.
[c] Qualified random sample or 2-h mixed sample.

Japan. In Japan, the environmental authorities determine the national wastewater standards, which can, however, be raised according to the special situation of local prefectures. A selection of Japanese wastewater standards for direct discharge is given below [458]

pH (inland waters)	5.8 – 8.6
BOD, COD	160 mg/L
Settleable substances	200 mg/L
Copper	3 mg/L
Chromium, total	2 mg/L
Chromium(VI)	0.5 mg/L
Zinc	5 mg/L
Iron	10 mg/L
Fluoride	15 mg/L
Nitrogen	120 mg/L
Phosphorus	16 mg/L
Tetrachloroethylene	0.1 mg/L

A compilation of the requirements to be met for the discharge of some important substances into public-owned treatment plants in various countries is given in Table 20 [457,459].

14.3. Components of Textile Wastewater

The composition of textile wastewater depends to a large extent on the dyed substrates, dyes, textile auxiliaries, and chemicals used, which makes the statement of average values impossible. As a rule, untreated textile wastewater has an alkaline pH, a higher temperature, high conductivity, and a poorer BOD_5 : COD ratio (i.e., lower degradability) compared to domestic wastewater. Typical BOD_5 : COD values lie between 1:2.5 and 1:5 [455]. The organic load of the wastewater originates largely from pretreatment of the materials – BOD_5 values of up to 140 kg O_2 per tonne of material being obtained for desizing and 210 kg/t in the case of alkaline boiling out of polyester – cotton blend fabrics [461, 462]. Wastewater from dyehouses, which work primarily with sulfur or vat dyes, have high COD values because of the organic or inorganic reducing agents (glucose, hydroxyacetone, sodium

Table 20. Requirements for indirect discharge (selection)

Parameter	Belgium	Denmark	Switzerland	Austria	United States∗	Japan (Osaka)∗∗
pH	5.5 – 10	5.5 – 10	5.5 – 10	6.5 – 9.5	6 – 9	5 – 8
Temperature, °C	45	30 – 35	60	40	50	40
AOX, mg/L			0.1	0.5		
Sulfate, mg/L	2000		300	200		
Chlorine, mg/L		0.5	3	1		
Chromium(VI), mg/L	0.3	2	0.5	0.1		0.5
Total chromium, mg/L	4		2	1	0.5	2
Copper, mg/L	0.5	1	1	0.5	0.5	3
Nickel, mg/L	1	1	2			
Zinc, mg/L	5	1 – 2	2	2	1	

∗ Typical example [459].
∗∗ [460].

Table 21. Typical wastewater of a cotton finisher [446]

Parameter	Average	Minimum	Maximum
pH		8.5	10.3
Conductivity, mS/m	650	420	1400
Temperature, °C	27	25	38
COD, mg/L	650	420	1400
BOD_5, mg/L	180	80	500
Phosphate (as P), mg/L	50	26	80
Sulfate, mg/L	810	750	1050
Ammonium (as N), mg/L	0.7	0.6	1.0
Hydrocarbons, mg/L	5	3	15
Chloride, mg/L	800	400	1500
AOX, mg/L	0.8	0.5	1.2
Settleable substances, mg/L	10	1	150
Copper, mg/L	< 0.1	< 0.1	0.6
Chromium, mg/L	< 0.01	0.015	0.034

dithionite) used. An analysis of the wastewater from a cotton finisher is shown in Table 21.

The contaminants, which are controlled on discharge into the public sewage system or into surface water, can have many different sources (Table 22) [459,461,463–465]. Apart from metal complexes, the heavy-metal content in dyes is usually in the range of a few micrograms per gram and is due partially to the catalysts used in production [459].

Table 22. Important sources of contaminants in wastewater from dyehouses

Parameter, contaminant	Source
Alkaline pH value	dyeing with reactive, vat, and sulfur dyes; mercerization, boiling-off, bleaching
Acidic pH value	dyeing with basic, acidic, and disperse dyes; easy care finish
Color	reactive and sulfur dyes
Heavy metals	
Chromium	metal-complex dyes, oxidant
Cobalt	metal-complex dyes
Copper	metal-complex dyes, pigments, improvers of lightfastness
Nickel	metal-complex dyes
Zinc	cationic dyes, resist and reducing agents, biocides, catalysts, water piping
Antimony	flame retardant
Tin	biocides
Halogenated hydrocarbons	detergents, degreasers, carrier, chlorine bleaching
AOX	dyes, carrier, chlorine bleaching, preservatives, mothproofers, wool chlorination
Mineral oils	yarn preparations, spinning oils, defoamers, dedusting agents, print thickening
Phosphorus, phosphates	buffers, sequestrants, flame retardants
Sulfide	sulfur dyes
Neutral salts	reactive and substantive dyeing

Depending on the coloristic class, dyes are fixed in various quantities (see below) [447]:

Direct dyes	70 – 95 %
Vat dyes	80 – 95 %
Sulfur dyes	60 – 70 %
Reactive dyes	50 – 95 %
Disperse dyes	80 – 92 %
Acid dyes	80 – 93 %
Afterchroming dyes	98 – 99 %
Cationic (basic) dyes	97 – 98 %

Nevertheless, their share (ca. 1 – 2 %) of the organic wastewater load is very low [466].

14.4. Reduction of Wastewater Load by Special Preventive Measures

If certain requirements are not fulfilled, the wastewater must be treated or the discharge of the contaminants in question must be lowered or completely eliminated. As a rule, preventive measures are more economical than subsequent contaminant elimination. The discharge of a number of the contaminants listed in Table 22 can be reduced or eliminated by process changes or exchange of products. Residual color from reactive dyes can be reduced by using dyes with a higher fixing yield (i.e., with more affinitive, bifunctional, or bireactive dyes [467]). Higher yields are also achieved by a shorter liquor ratio or by using continuous or semicontinuous processes instead of exhaustion processes [468]. Another possibility is the use of higher salt concentrations, which, however, stand in the way of the demand for lower electrolyte concentrations in wastewater. From an economic viewpoint, optimization of the formulation can be achieved by means of suitable computer programs [469]. In the case of copper-containing metal-complex dyes, these methods result in the lowering of an increased copper concentration. The complete avoidance of metal complexes, which are employed for some violet, blue, navy blue, and black shades in direct and reactive dyes, results in marked losses in lightfastness. Turquoise and brilliant green shades can be produced only with copper or nickel phthalocyanine complex dyes.

In padding processes, a general problem is the *unused dye liquor* in the pad-trough, piping, and dissolving tank. Better utilization of dye liqour can be achieved by using padders with a smaller content [470] or by reusing the remaining liquor in other dyeing processes, which requires colorimetric determination of the correct formulation in the laboratory [471]. With reactive dyes, however, reuse is limited because the liquors are not stable for long in the presence of fixing alkali.

In the wastewater from wool dyeworks, increased *chromium concentrations* are frequently encountered. They are caused by chromium metal-complex dyes and/or by the chroming of chrome dyes. To lower chromium concentration, optimized chroming methods have been developed. In fact, chromium concentrations in residual liquors have been reduced from 20 – 150 mg/L in the case of traditional dyeing to 1 – 6 mg/L. Thus, the concentrations lie below those obtained on black dyeing with 1:2 metal-complex dyes, which can have up to 13 mg/L of chromium [472,473]. Optimized chroming methods include the International Wool Secretariat (IWS) low-temperature chrome dyeing method at 90 °C with the addition of thiosulfate, the Bayer Glauber's salt technique with dye-specific chrome factors, and the Sandoz method with minimum chrome factors and a special auxiliary agent.

The *bleaching of cotton* with hypochlorite is an important source of absorbable organic halogen (AOX). In laboratory experiments with raw cotton, the AOX detected in the bleaching liquor was up to 30 mg/L and, in the case of desized materials, up to 80 mg/L. In addition, up to 11 mg/L of chloroform can be formed by a haloform reaction during hypochlorite bleaching under standard conditions. In contrast, bleaching with sodium chlorite or hydrogen peroxide gives ca. 10 or less than 0.5 mg/L AOX, respectively [474,475]. The transition to bleaching with hydrogen peroxide can reduce both AOX content and free chlorine in the wastewater.

Phosphorus enters wastewater largely via buffer substances or sequestering agents. Although some countries have already installed stages for phosphate elimination in sewage treatment plants, a reduction is desirable because of the possibility of eutrophication of surface waters. Apart from the use of phosphorus-free substances, an alternative to phosphorus compounds as buffers is dosing with acids or bases via automatic dosing equipment. As sequestering agents and peroxide stabilizers in peroxide bleaching, polyhydroxy carbonates, aminocarboxylates [e.g., ethylenediaminetetraacetate (EDTA) or nitrilotriacetate (NTA)], phosphonates or modified polyacrylates are used instead of phosphorus compounds [476]. However, EDTA should not be used because it is poorly biodegradable and can remobilize heavy metals from the sediment. The use of sequestering agents can be almost avoided by performing an extraction step with formic or

hydrochloric acid before bleaching, eliminating hardness and heavy metals from the material.

In reactive dyeing, the use of *salt* can be avoided when continuous or semicontinuous processes are used instead of exhaustion processes. In the case of reactive dyes that have a low salt sensitivity, the amount of salt can be lowered without yield losses [469]. *Sulfites and sulfates* are also formed in the oxidation of sodium dithionite (hydrosulfite), which is used both as a stripping agent for faulty dyeing and for reduction of vat dyes. Thiourea dioxide or reductones such as hydroxyacetone are alternatives to hydrosulfite. The disadvantages of both chemicals are overreduction in the case of vat dyes of the indanthrone type and changes in shade [477]. Hydroxyacetone, which is easily degradable, is being used increasingly in indigo dyeing. *Sulfide* formed when dyeing with sulfur dyes must be oxidized before discharge. The use of low-sulfur, prereduced dyes, which are reduced by glucose, lowers the electrolyte load and avoids the required sulfide oxidation [471].

Multiple use of exhausted dye liquors is usually restricted to dyes with high exhaustion yield and exhibit low interactions with the auxiliaries remaining in the liquor. Examples are the dyeing of cotton with indigo or sulfur dyes and the dyeing of cotton velvet with direct dyes. Polyamide can be dyed with acid dyes, acrylic fibers with cationic dyes, and polyester with disperse dyes in a standing bath [466, 477]. Multiple use of the residual liquors from finishing processes should also be aimed at. However, the possible staining of the padding liquors with dyes, the correct analysis of present composition, and the storage stability can cause problems [478]. The advantages of reuse are reduced organic load of the wastewater and savings of water, heating energy, and, in part, auxiliary agents and chemicals. Water consumption can also be decreased by using dyeing equipment with shorter liquor ratios [479] (see Section 14.1), optimized process operation (e.g., washing in reactive dyeing [480]), and a series of other measures [466].

14.5. Special Methods of Wastewater Treatment

The usual methods of wastewater treatment can also be applied to the textile finishing industry. A special case is, however, the removal of dyes. Apart from paper dyeing, the presence of dyes in wastewater to this extent is specific to the textile industry.

Biological Treatment. As mentioned in Section 14.3, textile wastewater is frequently poorly biodegradable. This is of minor importance if textile wastewater represents a relatively small part of the total wastewater of a municipal biological sewage treatment plant. Nevertheless, textile wastewater can be subjected satisfactorily to biological treatment even if it is not mixed with municipal wastewater. Because of the different textile processing program, textile wastewater has a complex and rapidly changing composition. For this reason, preceding equalization of composition in adequately dimensioned balancing tanks is required for optimal purification. In addition, a certain temperature compensation is achieved, extreme pH values are avoided by internal neutralization, etc. The balancing tanks also eliminate the need for design of the treatment plant based on maximum wastewater values, serve charge of the biological treatment plant on work-free weekends, and represent safety stages in case of breakdown. The COD value of wastewater has been reported to be reduced almost 50 %, and the mineral oil concentration by 70 %, by appropriate design of the balancing tank, introduction of adapted microorganisms, and additional aeration [481]. The pilot plant and the operation of a one-stage aerated biological sewage treatment plant run by the finisher of cotton and blended fabrics are described in [482]. Adaptation of microorganisms in the one-stage plant was shown to considerably improve treatment efficiency in comparison with the pilot plant and the results at the start of operation. Biological treatment plants can be improved by combination with preceding or subsequent precipitation – flocculation [483, 484] and addition of activated carbon or cheaper brown-coal cokes in the activated sludge part [485–487]. Pilot experiments on the decolorization of the outflow from a biological sewage treatment plant using an activated carbon fixed-bed reactor show a bacterial degradation of the dye, which occurs in addition to purely adsorptive processes [488]. Highly loaded wastewater (e.g., partial streams from pretreatment and desizing), is very suitable for anaerobic treatment. Partial

streams from dyeworks are not easily degradable anaerobically, especially in the presence of better degradable substrates [489,490].

Precipitation – Flocculation. Chemical precipitation – flocculation belongs to the generally accepted technological conventions and is used for extensive wastewater treatment in many branches of industry. The advantages are increased elimination of contaminants, easy combinability with other processes, and relative insensitivity of these processes to substances that are toxic for bacteria and to fluctuations in wastewater composition [491]. The usual inorganic precipitating agents are lime, aluminum and iron salts [e.g., iron chloride sulfate, iron(II) and iron(III) chloride]. Instead of direct dosing into the wastewater, electrochemical production of the metal cations or their hydroxides by corresponding sacrificial electrodes is possible. Apart from reducing the contaminant load (COD, BOD, heavy metals, and mineral oils), dyes are also eliminated, at least partially. These dyes include vat, disperse, sulfur, direct, and acid dyes, as well as pigments [492–495]. Reactive dyes are not adequately removed by the usual inorganic precipitants. For this purpose, polycationic precipitating agents are used. With dyes containing sulfonic acid groups, these agents form compounds that are poorly soluble in water and flocculate [461,468,496–498]. In both cases, the sludge can be separated by flotation or sedimentation.

Oxidation. The disadvantage of precipitation and flocculation is the production of large amounts of sludge, which must be treated and disposed of. Chemical processes, especially oxidation and reduction, usually do not have this disadvantage. The oxidizing agents used in practice or in research projects for the treatment of textile wastewater (especially for the removal of residual color) are chlorine [487] or hydrogen peroxide, the latter usually together with iron(II) sulfate (Fenton's reagent) [499–501], ozone [502–504], or UV irradiation [505,506]. A combination of ozone and UV irradiation also exists. The disadvantage of chlorination is the formation of organic chloro compounds. Fenton's reagent effectively destroys water-soluble dyes. However, it also nonspecifically oxidizes organic pollutants expressed as COD, leading to a high consumption of peroxide and correspondingly high costs. Ozone reacts specifically with dyes, especially at neutral and acidic pH. Thus, other water components such as surfactants and thickeners are hardly attacked. Wastewater treated with oxidizing agents shows better degradability and a lower toxicity for bacteria. The reaction products of ozone treatment include saturated and unsaturated carboxylic acids (e.g., oxalic, malonic, maleic, and phthalic acids) [507]. For the special treatment of wastewater from sulfur dyeing, see Section 4.5.6.

Reduction. Sodium dithionite is the agent of choice for reductive decolorization. Under suitable reaction conditions, not only azo dyes but also dyes based on other chemical structures (e.g., anthraquinone or formazane dyes) [461,508] can be permanently reduced. Treatment with iron(II) sulfate – lime, which uses the reduction potential of divalent iron, represents a combination of reduction and precipitation. This stage precedes the biological wastewater treatment plant [509]. Divalent iron ions can be produced electrochemically on activated iron electrodes, which can also reduce hexavalent chromium from wool dyeing [473]. Both direct and indirect cathodic dye reduction [i.e., via an iron(II) – iron(III) – triethanolamine redox system] is possible by electrolysis with higher current strengths [510]. In purely chemical methods of treatment, which proceed without the use of precipitants, no elimination of heavy metals (e.g., from metal-complex dyes) occurs.

Adsorption. The most important adsorbent in wastewater treatment is activated carbon. A distinction is made between the stirring process with powdered carbon and the percolation process with granulated carbon [455,466,487, 511]. The former has already been mentioned in connection with biological treatment. Besides the actual adsorption, biological processes (i.e., combinations of aerobic and anaerobic degradation) also play an important role. Apart from water-soluble dyes, polar substances are preferentially adsorbed to activated carbon. As a result of the limited adsorption capacity and high cost, activated carbon should be used preferably for the treatment of less loaded wastewater (e.g., in fine purification for reuse or as safety filters). Cheaper adsorbents under investigation are

coking coal, peat, kieselguhr, activated alumina, clay, chitin, and other natural substances. Synthetic polymers and ion exchangers are suited to the separation of polar and nonpolar water components. As with activated carbon, periodic regeneration is required [487]. For the selective separation of heavy-metal cations, AOX, and dyes by polymeric diaza crown ethers, cyclodextrins, calixarenes, and curcubituril, see [512,513].

Membrane Separation. The separation of contaminants by using membrane techniques (e.g., reverse osmosis, nano- or ultrafiltration) results in no final decrease in contamination. Concentrates with the separated contaminants must be treated further by using known techniques, while the clean permeates can be discharged or reused in the plant [514,515]. Membranes that are applied include those made of cellulose acetate, polyamides, and polysulfones, and the robust, dynamic membranes made of zirconium(IV) hydroxide – polyacrylates on porous ceramic or steel carriers [487].

Neutralization. Textile wastewater must be neutralized since the pH is frequently more alkaline than permitted for discharge. Apart from strong acids such as sulfuric or hydrochloric acid, CO_2 is also well suited to neutralization and can be used in the pure state [516] or as a component of flue gas [517]. In comparison with mineral acids, the advantages of neutralization with CO_2 are lower chemical costs, less material corrosion, no salting, and more safety in dosing.

14.6. Recycling of Water and Valuable Substances

Recycling of Water. Because of high costs for treatment or discharge of industrial and wastewater, reuse of water occasionally appears reasonable. The multiple use of dye liquors was mentioned in Section 14.4. Similarly, slightly loaded rinsing baths can be reused in many cases without previous treatment. For universal applicability, however, purification, usually a combination of several processes, is required. Some of these are given below:

1) *Weakly loaded wastewater*
   Powdered activated carbon
   Polyaluminum chloride precipitation and sedimentation
   Multilayer filter [518]
2) *Total wastewater*
   Charcoal-supported biology in stirred loop reactor
   Adsorption in stirred loop reactor on pulverized brown coal
   Polyaluminum chloride precipitation and sedimentation
   Fixed-bed filtration with granulated brown coal [486,519]
3) *Total wastewater, mainly from reactive dyeing*
   Precipitation with cationic precipitating agents and flotation
   Fixed-bed filter with sand and brown coal [520]
4) *Wastewater from polyester dyeing*
   Electrocoagulation
   Filtration [494]
5) *Total wastewater*
   Precipitation – flocculation – filtration
   Nanofiltration [521]
6) *Total wastewater*
   Ozonization
   Precipitation with calcium chloride and sedimentation
   Electrolysis
   Treatment with powdered activated carbon
   Fixed-bed filtration [522]
7) *Total wastewater*
   Distillation or evaporation [523,524]

Recovery of Valuable Substances. Unused dyes, textile auxiliaries, and chemicals are valuable materials that can be recovered and reused, saving considerable costs. The energy content of hot baths can be used to heat raw water by means of heat exchangers. A lowering of temperature of dyebaths is often required anyway to maintain discharge limits [525]. The recovery of synthetic size by ultrafiltration is becoming increasingly important. Not only the value of the size, but also the reduction of wastewater loading with poorly degradable agents should be taken into account [526]. The ultrafiltration plant amortizes in less than two years. Diluted sodium hydroxide solution, which is produced in the washing of mercerized fabrics, is purified by

ultrafiltration before it is concentrated by evaporation [527]. The recovery of indigo from rinsing baths is also successfully carried out by ultrafiltration [526,527]. Dyes and auxiliaries can be eliminated from the rinsing baths and concentrated with hot water- and pH-resistant dynamic membranes. Reuse is then possible after combination with residual dye liquors and correction of the chemical concentrations [528]. Salts, which have a concentration of up to 100 g/L in exhaustion dyeing with reactive dyes, can be largely recovered if the salt-containing residual liquor and the first rinsing bath, if necessary, are subjected to ultrafiltration. (The less loaded rinsing baths are discharged into the sewage system without treatment.) The amortization time for the investment of these plants is less than three years because of the recovery of the salt [514, 529].

14.7. Ecotoxicology of Textile Dyes

Biological Degradability. In view of their high stability to a number of physical and chemical influences, dyes have a similarly high stability to microorganisms. Although indications of mineralization have been found in some investigations [530], dyes are regarded as nondegradable during their relatively short detention time in normal *aerobic biological wastewater treatment plants* [531]. For some dyes (e.g., C.I. Basic Violet 1 and 3 [532], the azo dyes C.I. Acid Orange 12 and 20, and the triphenylmethane dye C.I. Acid Red 88 [533]), degradation by adapted bacteria has been detected. Dyes can also be biodegraded on biofilters or biofilms (i.e., packing with a bacterial layer) [534,535]. The most important route of elimination in wastewater treatment plants is adsorption on clarification sludge, the extent depending on the constitution of the dye [530,531,536,537]. Microbial degradation is possible under *anaerobic conditions* (e.g., in digestion tanks, soil, or sediments). This applies not only to azo dyes, in which reductive cleavage of the azo bond occurs, but also to other classes of dyes [538]. The aromatic amines expected from the anaerobic degradation of azo dyes have been detected. Although the amines are poorly degradable anaerobically, aerobic degradation occurs readily [539]. From a model landfill with sludge from sewage treatment plants containing mostly textile wastewater, neither dyes nor the corresponding amine components could be eluted [540].

Toxicity for Aquatic Organisms. Dyes have a low toxicity for aquatic organisms. Of 200 compounds tested, only 18 had a harmful effect on wastewater bacteria at concentrations of < 100 mg/L. These were mostly cationic dyes [541]. However, these dyes exhaust very well in the dyebath, are easily adsorbed by biosludge, and thus present no problems in wastewater treatment plant effluents. With respect to algae and fish toxicity, cationic dyes are usually the culprits [455,542]. Of more than 3000 tested commercial dyes, 59 % were nontoxic to fish at a concentration of > 100 mg/L, 27 % had an LC_{50} between 10 and 100 mg/L, and only 3 % were toxic at concentrations < 10 mg/L [543]. Dye concentrations in surface waters are estimated at the most at 1 – 10 µg/L (i.e., three to five orders of magnitude lower than the acute fish toxicity) [544]. Bioaccumulation in fish does not occur with water-soluble dyes and is practically negligible for disperse dyes [545].

Toxicity for Warm-Blooded Animals and Occupational Health. In general, dyes have a low toxicity for *warm-blooded animals*. Of 4500 tested commercial products, 82 % had an LD_{50} (rat, oral) of > 5000 mg/kg and < 1 % had a value of < 250 mg/kg [546]. According to the ChemG, only 13 known textile dyes are considered toxic ($LD_{50} < 200$ mg/kg oral and $LC_{50} < 25$ mg/m^3 inhalation), several of which are no longer produced [547]. Studies on rats to determine the subchronic toxicity of eight commercial products with an $LD_{50} > 2000$ mg/kg showed no measurable effects in the case of 22 daily doses during 28 days of 1000 mg/kg each [548]. A carcinogenic effect of 13 textile dyes, mainly azo dyes, has been detected in animal experiments [548–550]. In metabolism, aromatic amines may be formed by the enzymatic reductive cleavage of azo dyes. According to a recommendation of the German MAK Commission (1988), azo dyes that release carcinogenic arylamines of the MAK groups III A 1 and III A 2 under reductive conditions should be treated like the released amines themselves. In the GefStoffV [551], this recommendation led to consequences

with regard to classification, labeling, packaging, and contact in the factory. The important representatives involved are dyes based on benzidine and its congeners. The members of the Ecological and Toxicological Association of the Dyestuff and Pigment Industry (ETAD) have agreed voluntarily to neither produce nor market benzidine dyes. Up to now the use of benzidine dyes and the importation of textiles dyed with them have been banned only in Sweden. In India, the production of benzidine dyes has been banned since February 1993. In Germany, the sale of textiles dyed with dyes that are able to cleave under formation of carcinogenic amines and can come into contact with skin will be banned from January 1996 on (2nd and 3rd Ordinance for Amendment of Consumer Goods Ordinance from July 15, 1994, BGBl. 1670 – 1671, and from December 16, 1994, BGBl. 3836). Benzidine and its metabolites could be detected in the urine of workers who had contact with benzidine dyes and had inhaled their dust. In plants with a high standard of industrial hygiene, the results were negative [552].

In the case of fast dyeing – since the migration of dyes on and through the skin is estimated at most at a few micrograms – the risk of endangering health is very low. An exception is the risk of allergies, which exists for a few dyes. In comparison with worldwide dye consumption, the number of skin allergies due to dyes is negligibly small [553]. The dyes with increased allergic potential are essentially those listed below [554–556].

Colour Index	*Dye class*
Disperse Yellow 1, 9	nitro dye
Disperse Yellow 39, 49	methine dye
Disperse Yellow 54, 64	quinoline dye
Disperse Orange 1, 3, 76	azo dye
Disperse Red 11, 15	anthraquinone dye
Disperse Red 1, 17	azo dye
Disperse Blue 1, 3, 7, 26, 35	anthraquinone dye
Disperse Blue 102, 106, 124	azo dye
Acid Violet 17	azo dye

In textile dyehouses, especially in dyeing kitchens, additional exposure to inhalable dust should be considered. Apart from skin irritation, respiratory or nasal problems and itching of the eyes can occur, especially due to reactive dyes. Remedial action can be taken by means of good occupational hygiene and dust-free working conditions [557–559].

15. References

1. H. Vogler, *Textilveredlung* **27** (1992) 352 – 358.
2. S. M. Edelstein, H. C. Borghetty: *The Plictho of Gioanventura Rosetti,* The MIT Press, Cambridge – London 1969.
3. *Ullmann,* 4th ed. **11,** 99 – 134.
4. H. Schweppe: *Handbuch der Naturfarbstoffe, Vorkommen, Verwendung, Nachweis,* Ecomed Verlag, Landsberg 1993.
5. *Kirk-Othmer,* 3rd ed., **8,** 351 – 373.
6. F. Brunello: *The Art of Dyeing in the History of Mankind,* Neri Pozza, Vicenza 1973, English Translation by Phoenix Dyeworks, Cleveland, OH.
7. G. Fieler: *Farben aus der Natur, Eine Sammlung alter und neuer Färberezepte für das Färben auf Wolle, Seide, Baumwolle und Leinen,* M. H. Schaper, Hannover 1978.
8. R. J. Adrosko: *Natural Dyes and Home Dyeing, a Practical Guide with over 150 Recipes,* Dover Publ. Inc., New York 1968.
9. E. Bemiss: *The Dyers Companion,* **vol. 2,** New London 1806, 3rd. enlarged ed. New York 1973. C. G. Gilroy: *A Practical Treatise on Dyeing and Calico-printing,* 2nd ed., New York 1846.
10. P. Richter, *Wirkerei- Strickereitech.* **40** (1990) no. 3 to 41 (1991) no. 10.
11. Gesamtverband der deutschen Textilveredlungsindustrie, TVI-Verband e.V.: Annual Report 1992.
12. Arbeitgeberkreis Gesamttextil: *Textilveredelung, Färben – Aus der Geschichte der Färberei,* Frankfurt 1990, pp. 1.01 – 1.02.
13. R. Germer: *Die Textilfärberei und die Verwendung gefärbter Textilien im alten ägypten,* Verlag Otto Harrasowitz, Wiesbaden 1993, p. 150.
14. C. Hütz: "Textiltechnik zur Zeit des alten Testaments," *Textilveredlung* **28** (1993) 59 – 67.
15. A. Kretschmer: "Textilveredlungsmaschinen (gestern), heute und morgen," *Melliand Textilber.* **64** (1983) 676 – 680.
16. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung, Handbuch der Technologie, Verfahren, Maschinen,* 13th ed., Deutscher Fachverlag, Frankfurt 1989.
17. H. Zollinger: *Color Chemistry,* Verlag Chemie, Weinheim 1987.

18. J. Skoufis: "Low-Liquor Ratio Foam Processing," *Am. Dyest. Rep.* **68** (1979) 20 – 23.
19. W. Saus, D. Knittel, E. Schollmeyer: "Färben aus überkritischem Kohlendioxid – eine Alternative zur HT-Färbung von Polyester," *Text. Prax. Int.* **47** (1992) 1052 – 1054.
20. W. Saus, S. Hoger, D. Knittel, E. Schollmeyer: "Färben aus überkritischem Kohlendioxid, Dispersionsfarbstoffe und Baumwollgewebe," *Textilveredlung* **28** (1993) 38 – 40.
21. W. Rüttiger, J. Ehlert: "Die kritische Färbegeschwindigkeit, eine verfahrenstechnische Kenngröße für die zeitsparende und egale Färbung auf dem Apparat," *Text. Prax.* **27** (1972) 609 – 616.
22. H. Leube, W. Rüttiger: "Egalität und Färbesicherheit bei unterschiedlich geführten Badfärbungen von Polyesterfasern," *Melliand Textilber.* **59** (1978) 836 –842.
23. B. Smith, G. Mcintosh, S. Shanding: "Ultrasound – A Novel Dying Accelerant," *Am. Dyest. Rep.* **77** (1988) no. 10, 15 – 20.
24. *Ullmann,* 4th ed., **22,** 635 – 716.
25. U. Baumgarte: "Das Aufziehverhalten der Indanthrenfarbstoffe als wichtige Grundlage für die Kurzzeitfärbetechnik," *Melliand Textilber.* **55** (1974) 953 – 962.
26. U. Nahr, W. von Bistram: "Grundlagen für die EDV-gestützte Chemikalienberechnung zum Färben mit Indanthren-Farbstoffen in teil- oder vollgefluteten geschlossenen Apparaten," *Text. Prax. Int.* **46** (1991) 978 – 983.
27. H. Gerber: "Kennwerte zur Charakterisierung des färberischen Verhaltens von Reaktivfarbstoffen," *Textilveredlung* **28** (1993) 405 – 411.
28. P. S. Collishaw, D. A. S. Philips, M. J. Bradbury: "Controlled Coloration: A Success Strategy for the Dyeing of Cellulosic Fibers with Reactive Dyes," *J. Soc. Dyers Colour.* **109** (1993) 284 – 292.
29. W. Schrott, K. Bacher: "Kreativität führt zu Spitzenleistung. Wo stehen wir im Heißausziehsegment mit Reaktivfarbstoffen?" *Text. Prax. Int.* **47** (1992) 1145 – 1150.
30. F. Hoffmann, H. Schubert, J. Fiegel: "Einsatz von Rechnern zur Optimierung von Färberezepten und Verfahren," *Text. Prax. Int.* **47** (1992) 233 – 237.
31. U. Baumgarte: "Zustand und Verhalten von Farbstoffen beim Färben am Beispiel der Leukoküpenfarbstoffe," *Textilveredlung* **15** (1980) 413 – 425.
32. W. Rüttiger: "Prozeßgrößen und Kostenfaktoren in der Veredlung am Bildschirm verfolgen," *Text. Prax.* **38** (1983) 1216 – 1217.
33. W. Rüttiger: "Process Simulation and On-Line Data – A Step Towards Microprocessor Control of Textile Processing," *1985 AATCC Technical Programm Book,* 1985, pp. 234 – 252.
34. W. Rüttiger: "Expertensysteme – das künftige Instrument technischer Führungskräfte," *Textilveredelung* **23** (1988) 127 – 134, 159 – 162, 199 – 203.
35. H. Wolf, W. Rüttiger: "Möglichkeiten und Grenzen der Kurzzeit-Färbetechniken," *Melliand Textilber.* **55** (1974) 876 – 879.
36. W. Rüttiger: "Prozess-Simulation am Beispiel der Polyesterfärbung," *Chemiefasern/Textilind.* **40/99** (1990) 446.
37. E. Saas, W. Franke: "Welche Vorteile bietet der Einsatz einer Hochtemperatur-Haspelkufe," *Chemiefasern* **17** (1967) 218 – 219.
38. P. Toldrian, W. Rüttiger: "Verfahrenssymbole als Voraussetzung für die automatisch gesteuerte Färbung mit Apparaten," *Melliand Textilber.* **55** (1974) 973 – 978.
39. A. Kretschmer: "Welchen Pumpentyp für den Färbeapparat von morgen," *Text. Prax.* **25** (1970) 39.
40. H. U. von der Eltz: "ITMA 1991 Färbeapparate und -maschinen für das Ausziehfärben," *Text. Prax. Int.* **47** (1992) 154 – 162, 246 – 251.
41. G. Rordorf: *Neues großes Handbuch der Textilkunde,* Fachbuchverlag Dr. Pfanneberg, Giessen 1956, pp. 271 – 272.
42. J. Ehlert: "Maschinen und Apparate für die Textilfärberei," *Deutscher Färberkalender 1970,* Deutscher Fachverlag, Frankfurt 1969, pp. 89 – 147.
43. A. Kretschmer: "Die Durchströmungsrechnung als Optimierungshilfe in der Apparatefärberei," *Textilveredlung* **15** (1980) 50 – 55.
44. H. Beiertz: "Färben und Drucken textiler Bodenbeläge aus Trevira CF- und CS-Fasern," *Chemiefasern/Textilind.* **42/94** (1992) 64 – 67.
45. MCS: "Färbemaschinen der Baureihe Lang-Horn erweitert," *Text. Prax. Int.* **48** (1993) 995.
46. THEN: "Airflow noch wirtschaftlicher," *Text. Prax. Int.* **47** (1992) 1171 – 1172.
47. H. K. Rouette: "ITMA 91: Highlights, Trends und Zukunft – Perspektiven in der technischen Entwicklung von Färberei, Druckerei und

Veredlung," *Chemiefasern/Textilind.* **42/94** (1992) 57 – 61.
48. Küsters: "Universal-Jigger für Vorbehandlung, Färberei und Spezialbehandlungen," *Text. Prax. Int.* **48** (1993) 248.
49. Henriksen: "Hochgeschwindigkeitssteuerung für Jigger bringt 30 % Zeitersparnis, soft-set-Trockner jetzt auch kontinuierlich arbeitend," *Text. Prax. Int.* **48** (1993) 131 – 132.
50. MCS: "Neue Strangfärbemaschine," *Melliand Textilber.* **74** (1993) 566.
51. "ADR's Annual Buyers' Guide to Garment Dyeing/Washing Machinery and Stone/Chemical Suppliers," *Am. Dyest. Rep.* **82** (1993) no. 5, 38 – 45.
52. Thies: "Trommelfärbeapparat für Fertigkonfektion," *Melliand Textilber.* **67** (1986) 723.
53. J. Park: "Developments in Dyeing Process Control," *Rev. Prog. Color. Relat. Top.* **15** (1985) 1 – 5.
54. *On-Line-Steuerungen in der Textilveredlung,* lecture held at Fachhochschule Reutlingen.
55. W. Rüttiger, P. Toldrian: "Einheitliche Verfahrensdarstellung zur Automatisierung und Optimierung der chargenweisen Färbung," *Text. Prax.* **27** (1972)674 – 678, 719 – 722.
56. Sepa: "Offene Prozessleittechnik für die Textilveredlung," *Text. Prax. Int.* **47** (1992) 674.
57. J. H. Heetjans: "Handhabungs- und Transportsysteme in der Garnfärberei," *Chemiefasern/Textilindustrie* **39/91** (1989).
58. G. Hofmann: "Automatisches Beschicken von Färbeapparaten und Garnveredlung," *Chemiefasern/Textilind.* **42/94**(1992) 61 – 63.
59. J. Ehlert: "Rationalisierung in der Textilfärberei," *Deutscher Färberkalender 1971,* Deutscher Fachverlag, Frankfurt 1970.
60. E. Voss: "Erfahrungen mit pH-Automaten für die Färbung von Stück- und Konfektionsware, sehr hohe Ausziehgrade für Polyamid und Wolle in der Badfärbung," *Melliand Textilber.* **74** (1993) 564 –566.
61. K. Tiedemann, G. Grüninger, W. Rüttiger: "Methoden zur On-Line-überwachung der Wasserstoffperoxidkonzentration für die Bleiche und Küpenfärbung," *Deutscher Färberkalender 1993,* Deutscher Fachverlag, Frankfurt 1992, pp. 140 – 172.
62. H. Fuchs: "Erkenntnisse beim einbadigen Pigmentfärben und Ausrüsten," *Text. Prax. Int.* **49** (1994) 507 – 510.
63. W. Kambach: "Kontinuierliches Aufwickeln von Textilbahnen zu Großkaulen," *Text. Prax. Int.* **48** (1993) 36 – 38.
64. A. Günther: "Geringere Abwasserbelastung durch neue Konstruktionen bei Teppich-, Färbe- und Druckanlagen," *Text. Prax. Int.* **48** (1992) 848 –852.
65. Farmer Norton: "Neues Vakuum-Flottenauftragssystem," *Text. Prax. Int.* **48** (1993) 129 – 130.
66. J. Polster: "Foulardtechnik – tema con variazioni," *Melliand Textilber.* **74** (1993) 1173 – 1174.
67. K. van Wersch: "Kontrollierte Applikation von Färbeflotten," *Textilveredlung* **28** (1993) 218 – 223.
68. P. Salzmann, W. Schindler: "Die gemeinsame Applikation von Küpenfarbstoff und Reduktionsmittel beim Kontinuefärben ohne Zwischentrocknung," *Melliand Textilber.* **74** (1993) 525 – 529.
69. W. Rüttiger, W. Rümens, G. Burkhardt, H. Petersen: "übersicht der industriellen Verfahrenstechniken der Hochveredlung von Zellulose," *Melliand Textilber.* **11** (1972) 1278 – 1289.
70. H. Paulsen: "Energiebilanzen in der Textilveredlung," *Melliand Textilber.* **74** (1993) 550 – 556.
71. S. M. Burkinshaw, W. J. Marshall: "Continuous Dyeing of Piece Goods Using Radiofrequency Heating," part I, *J. Soc. Dyers Colour.* **102** (1986) 263 –268.
72. P. Lennox-Kerr: *Int. Dyer Text. Printer Bleacher Finish.* **171** (1986) 20 – 21.
73. U. Einsele: "Neuere Methoden zum Färben von Polypropylen – 1. Mitteilung: Färben mit Hilfe von Bindemitteln," *Melliand Textilber.* **72** (1991) 847 –851.
74. W. Rüttiger: "Das Ausnutzen des Waschwirkungsgrades – eine Einführung in das Beherrschen von Waschprozessen," *Text. Prax.* **34** (1979) 1380 – 1387, 1544 – 1551, 1629 – 1643.
75. W. Rüttiger, U. Kirner: "Die Geschwindigkeit des Wasserstoffperoxide-Eigenzerfalls als wichtige verfahrenstechnische Kenngröße in der Textilveredlung," *Text. Prax.* **25** (1970) 21 – 30.
76. D. W. Ravensberger, R. B. M. Holweg: "Monitoring and Control of Hydrogen Peroxide Dosing," *J. Soc. Dyers Colour.* **109** (1993) 72 – 75.

77. BASF: *Ratgeber Cellulosefasern, Schlichten, Vorbehandeln, Färben,* Verlag BASF AG, Ludwigshafen 1977, p. 319, 357.
78. K. van Wersch: "Anlagen für das Färben nach dem Pad Batch-Verfahren," *Melliand Textilber.* **73** (1992) 431 – 432.
79. U. Denter, E. Schollmeyer: "Verfahrensanalyse und Entwicklung einer Testfärbemethode für die Rezeptierung der KKV-Reaktivfärbung von Baumwollgeweben (AIF 7997)," *Melliand Textilber.* **74** (1993) 251.
80. Suderba: "Space-dyeing-Maschine für sechs Farben," *Text. Prax. Int.* **48** (1993) 1006.
81. Carbonell: "Praxiserfahrungen beim Versuch, automatische Prozeß-Steuerungen auf die Erfordernisse einer computerintegrierten Fertigung umzugestalten," *Text. Prax. Int.* **47** (1992) 468 – 472.
82. Qualico: "Schnellste kontinuierliche Feuchtekontrollen," *Text. Prax.* **48** (1993) 247.
83. J. Rieker, T. Guschlbauer: "Farbabläufe in der Kontinue-Färberei und Materialdichteunterschiede in Geweben – Kann die On-Line-Farbmessung helfen?" *Melliand Textilber.* **74** (1993) 1256 –1260.
84. W. Rüttiger: "Meßwerte im Computer für die Veredlungspraxis," *Text. Prax.* **41** (1986) 555 – 562, 776 –781.
85. O. Annen, H. Gerber: "Probleme bei der übertragung von Laborergebnissen in die Praxis," *Melliand Textilber.* **75** (1994) 406 – 411.
86. Roaches: "Neue infrarotbeheizte Laborfärbemaschinen entwickelt," *Text. Prax. Int.* **48** (1993) 131.
87. "ITMA 91 – Labor- und Musteranlagen für die Textilveredlung," *Melliand Textilber.* **73** (1992) 180 –183.
88. J. Oegarra: "On-Line-Qualitätskontrolle bei der Ausziehfärbung," *Melliand Textilber.* **74** (1993) 538 – 549.
89. K. Stommel: "Einsatz von Laborgeräten in der Textilveredlungsindustrie," *Melliand Textilber.* **45** (1964) 283 – 286.
90. O. Annen, H. Gerber, B. Seuthe: "Das Färbeverhalten von Viskose- und Modalfasern im Vergleich zu Baumwolle," *J. Soc. Dyers Colour.* **108** (1992) 215 – 261.
91. E. Floegel: *Statistik in BASIC,* Hofacker-Verlag, Holzkirchen 1986.
92. M. Warmes: "The Central Dye Preparation Station – A Necessity for an Efficient Process Cycle in the Dyehouse," *International Textile-Bulletin* **2/82** (1982).
93. Then: "Das Farbküchensystem AMC-CKM," *Text. Prax. Int.* **48** (1993) 1009.
94. H. P. Locher, M. Firmann: "Verbesserte Produktionssicherheit in der Foulardfärberei durch photometrische Kontrolle der Farbansätze," *Textilveredlung* **26** (1991) 393 – 398.
95. J. Rieker, U. Scherschel-Gaedke: "Zur Problematik der photometrischen Bestimmung von Farbstoffkonzentrationen in Färbebändern," *Text. Prax. Int.* **48** (1993) 980 – 982.
96. J. Rieker: "Labor- und Farbmessungen," *Melliand Textilber.* **73** (1992) 258 – 261.
97. Datacolor: "Integrierte Farbdatenverarbeitung," *Chemiefasern/Textilind.* **43/95** (1993) 70.
98. K. van Wersch: "On-Line-Farbmessung in der Kontinuefärberei aus der Sicht des Maschinenbauers," *ITB Veredlung* **36** (1990) no. 2, 21 – 42.
99. W. Pape: "Farbmessung in der Produktion," *Melliand Textilber.* **73** (1992) 152 – 155.
100. J. Rieker, E. Furthmüller: "Untersuchung der farbmetrischen Abmusterung mit Pass/Fail-Formeln," *Text. Prax. Int.* **42** (1987) 422 – 423.
101. J. Rieker, K. Hilscher: "Über die visuelle Relevanz von Farbstoffkonzentrationsunterschieden auf gefärbten Textilien," *Text. Prax. Int.* **49** (1994) 499 –501.
102. B. Wulfhorst, K. Meier: "Untersuchungen an einem kurzen Hochtemperaturheizer," *Chemiefasern/Textilind.* **43** (1993) no. 95, 40 – 45.
103. *Melliand Textilber.* **74** (1993) 204.
104. Y. G. Bryant: "Neue Fasern als temperaturgesteuerte Wärmefluß-Sperre," *Tech. Text.* **35** (1992) T 88 – T 89.
105. J. Gacén, D. Cayuela, J. Maillo, T. Rodriguez: "Einfluß der thermischen Vorbehandlung auf die Feinstruktur von thermofixiertem Polyester," *Melliand Textilber.* **74** (1993) 797 – 801.
106. H.-J. Flath, C. Scholz: "Charakterisierung zugänglicher Strukturanteile in Cellulosefasern," *Melliand Textilber.* **74** (1993) 219 – 221.
107. W. Rüttiger: "Mechanische Zustandsänderungen von Baumwolltextilien bei der Alkalikaltbehandlung, part 1: Grunderscheinungen der mechanischen Vorgänge und ihre Deutungsmöglichkeiten

über Faserstrukturmodelle," *Melliand Textilber. Int.* **51** (1970) 1449 – 1465.
108. R. V. Meyer, E. Haug, G. Spilgies: "Elastane – Chemie, Eigenschaften, Einsatzgebiete," *Melliand Textilber.* **74** (1993) 194 – 198.
109. H. G. Dejong: "Yarn-to-yarn Friction in Relation to some Properties of Fiber Materials," *Text. Res. J.* **63** (1993) 14 – 18.
110. G. Bühler: "Garnreibwerte. Ein altes, jedoch immer noch nicht ganz bewältigtes Problem in der Praxis der Maschenwarenherstellung," *TPI Text. Prax. Int.* **48** (1993) 108 – 113.
111. P. Offermann, G. Putzger: "Qualitätsanalyse im Garnveredlungs- und textilen Fertigungsprozeß," *Melliand Textilber.* **74** (1993) 237 – 242.
112. U. Scholze, J. Trauter: "Auswirkungen der Haarigkeit von Rohgarnen und der Fasereigenschaften auf das Webverhalten geschlichteter Kettgarne (Part 1)," *TPI Text. Prax. Int.* **48** (1993) 773 – 779.
113. J. P. Rust, S. Peykamian: "Yarn Hairiness and the Process of Winding," *Text. Res. J.* **62** (1992) 685 –689.
114. *TPI Text. Prax. Int.* **48** (1993) 122.
115. W. Löbel: "Weiterentwicklungen der Restfeuchtemessung von Textilien nach dem konduktometrischen Verfahren," *Melliand Textilber.* **73** (1992) 771.
116. W. Rüttiger: "Regel- und Steuerprobleme bei der kontinuierlichen Textilausrüstung," *Melliand Textilber.* **47** (1966) 552 – 559, 680 – 690.
117. W. Rüttiger, G. Schmidt: "Krumpf und Verformung, zwei wesentliche Faktoren zur Gütesicherung von Maschenware," *Melliand Textilber.* **61** (1980) 526 – 532.
118. A. Eickmeier, U. Beller, D. Knittel, E. Schollmeyer: "Untersuchungen von Polymeren im oberflächennahen Bereich mit dem Verfahren der Akustomikroskopie," *TPI Text. Prax. Int.* **48** (1993) 534 –536.
119. L. Andrews: "Charakterisierung von Texturen geschmirgelter Textilgewebe mit Hilfe der Statistik dritter Ordnung," *Melliand Textilber.* **74** (1993) 231 – 233.
120. A. Lehnen: "Untersuchung der Zusammenhänge zwischen Ergebnissen der subjektiven Beurteilung von Teppichoberseitenstrukturen sowie deren Ver-änderungszustände und optoelektronisch meßbaren Kennwerten," *TPI Text. Prax. Int.* **48** (1993) 104.
121. R. H. Peters: *Textile Chemistry,* **vol. III, The Physical Chemistry of Dyeing,** Chapt. 2, Importance of Physical Properties of Fibers, Elsevier, Amsterdam –Oxford – New York 1975, pp. 23 – 26.
122. M. Peter, H.-K. Rouette: *Grundlagen der Textilveredlung, Handbuch der Technologie, Verfahren, Maschinen,* 13th ed., Deutscher Fachverlag, Frankfurt 1989, Chap. 5.1.
123. W. Albrecht: "Welche Bedeutung haben die physikalischen Eigenschaften von Cellulosefasern für ihr Färbeverhalten," *Melliand Textilber. Int.* **51** (1970) 1437 – 1448.
124. H. Rath: *Lehrbuch der Textilchemie,* 3rd ed., Chap. I. A. Chemie und Physik der Cellulose und ihrer Begleitstoffe, Springer Verlag, Berlin 1972.
125. W. Bubeth (ed.): *Textile Faserstoffe – Beschaffenheit und Eigenschaften,* Springer Verlag, Heidelberg 1993.
126. P. Richter: "Färben von alkalisierten Polyesterfasern und Polyester-Microfasern," *Chemiefasern/Textilind.* **41/93** (1991) 1118 – 1125.
127. H. Gerber: "Kennwerte zur Charakterisierung des färberischen Verhaltens von Reaktivfarbstoffen," *Textilveredlung* **28** (1993) 405 – 411.
128. M. Harris: *Handbook of Textile Fibers,* Textile Book Publ., New York 1954.
129. W. Rüttiger: "Eine Zustandsgleichung für Cellulosefasern als Erfassungsmöglichkeit für Längung, Krumpf, Knitterbildung und -erholung," *Textilveredlung* **2** (1967) 428 – 435.
130. D. Fiebig, H. Herlinger, B. Kastl: "Strukturveränderungen bei Cellulosematerialien durch Netz-/Trockenprozesse," *TPI Text. Prax. Int.* **48** (1993) 789 – 794.
131. P. R. Brady: "Diffusion of Dyes in Natural Fibers," *Rev. Prog. Color. Relat. Top.* **22** (1992) 58 – 78.
132. B. Reichstädter, O. Pajgrt: "Struktur und Eigenschaften von Chemiefasern und Eigenschaften der fertigen Textilien," *Chemiefasern Text. Anwendungstech.* **20** (1970) 877 – 884.
133. P. Ehrler: "Präzisionsgewebe: Notwendigkeit verfahrenstechnischer Modifikationen," *Tech. Text.* **35** (1992) T 94 – T 97.
134. L. Vangheuwe: "Methode zur Bestimmung des dynamischen Moduls von Garnen," *Melliand Textilber.* **74** (1993) 717 – 718.
135. H. Grams: "Neues Websystem," *Melliand Textilber.* **74** (1993) 198 – 201.

136. J. Rieker, W. Braun: "Kontinuierliche Messung von Kante/Mitte/Kante-Unterschieden an Webware nach dem Durchströmungsstaudruck-Prinzip," *Melliand Textilber.* **68** (1987) 749 – 752.
137. D. H. Abrahams: "Reducing Cracks Creases Chafes," *Am. Dyest. Rep.* **63** (1974) no. 9, 32 – 36.
138. G. Heidemann, H.-J. Berndt: "Effektivtemperatur und Effektivspannung, zwei Meßgrößen zur absoluten Bestimmung des Fixierzustandes von Synthesefasern," *Melliand Textilber.* **57** (1976) 485.
139. G. Heidemann, H.-J. Berndt: "Die substrat- und prozeßspezifischen Größen der Fixierung von Synthesefasern, eine Analyse am Beispiel der Thermofixierung von Polyesterfasern," *Chemiefasern/Textilind.* **24/76** (1974) 46.
140. A. Kretschmer: "Grenzwerte der Flottenbeladung von flächigem Textilgut beim Foulardieren," *Text. Prax.* **25** (1970) 217 – 218.
141. E. D. Dresden: "Transparenzmessung – eine Möglichkeit zur Kontrolle von Veredlungsvorgängen," *TPI Text. Prax. Int.* **48** (1993) 230 – 231.
142. W. Rüttiger: "Ein plausibler Mechanismus für das Entstehen von Falten in Ablage-Färbeaggregaten," *Lenzinger Ber.* **45** (1978) 160 – 171.
143. Deutsches Textilforschungszentrum Nord-West e. V.: "Lauf- und Liegefalten," Krefeld 1977.
144. *Z. Gesamte Textilind.* **58** (1956) 504 – 505.
145. Bayer-Faser-Institut, company brochure, Bildung und Fixierung von Falten – Dralon, VII f 1, Leverkusen, Feb. 1967.
146. K. Venkataraman: *The Chemistry of Synthetic Dyes,* vol. 6: "Reactive Dyes," Academic Press, New York 1972.
147. D. Hildebrand, *Text. Prax. Int.* **25** (1970) 292 – 296, 351 – 354, 428 – 430, 492 – 495, 621 – 624; **26** (1971) 45 – 47, 431 – 433, 564 – 567.
148. M. R. Fox, H. H. Sumner in C. Preston (ed.): *The Dyeing of Cellulose Fibers,* Dyers' Company Publications Trust, Bradford 1986, pp. 142 – 195.
149. P. Rys, H. Zollinger, *Helv. Chim. Acta* **49** (1966) 749 – 754.
150. J. P. Luttringer, P. Dussy, *Melliand Textilber.* **62** (1981) 84 – 94.
151. A. H. M. Renfrew, J. A. Taylor, *J. Soc. Dyers Colour.* **105** (1989) 441 – 445.
152. H. H. Sumner, T. Vickerstaff, *Melliand Textilber.* **42** (1961) 1161 – 1166.
153. U. Ruf, W. B. Egger, *Textilveredlung* **13** (1978) 304 – 308.
154. S. Abeta, K. Akahori, U. Meyer, H. Zollinger, *J. Soc. Dyers Colour.* **107** (1991) 12 – 19.
155. M. Matsui, U. Meyer, H. Zollinger, *J. Soc. Dyers Colour.* **104** (1988) 425 – 431.
156. BASF, DE-OS 4 216 591, 1992. (J. Schulze et al.).
157. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung,* 13th ed., Deutscher Fachverlag, Frankfurt a. M. 1989, pp. 510 – 516.
158. A. H. M. Renfrew, J. A. Taylor, *Rev. Prog. Color. Relat. Top.* **20** (1990) 1 – 9.
159. D. M. Lewis, *J. Soc. Dyers Colour.* **109** (1993) 357 – 364.
160. N. S. Knaggs, *Am. Dyest. Rep.* **81** (1992) no. 11, 109 – 111.
161. J. Shore, *Rev. Prog. Color. Relat. Top.* **21** (1991) 23 – 42.
162. J. R. Aspland, *Text. Chem. Color.* **23** (1991) no. 12, 30 – 32.
163. J. J. Porter, *Text. Chem. Color.* **25** (1993) no. 2, 27 – 37.
164. J. R. Aspland, *Text. Chem. Color.* **23** (1991) no. 11, 41 – 45.
165. M. I. Khalil, A. Bendak, S. S. Aggour, *Am. Dyest. Rep.* **70** (1981) no. 10, 31.
166. D. R. Lemin, E. J. Vickers, T. Vickerstaff, *J. Soc. Dyers Colour.* **62** (1946) 132.
167. Report of the Commitee on Direct Dyes, *J. Soc. Dyers Colour.* **62** (1946) 280 – 285; **64** (1948) 145 – 146.
168. J. Varghese, N. Bhattacharyya, A. S. Sahasrabudhe, *Colourage* **36** (1989) no. 3, 16 – 31.
169. F. Weiss: *Die Küpenfarbstoffe,* Springer Verlag, Wien 1953. M. R. Vox: *Vat Dyes and Vat Dyeing,* Chapman and Hall, London 1948.
170. U. Baumgarte, *Textilveredlung* **4** (1969) 821 – 832.
171. U. Baumgarte, *Textilveredlung* **2** (1967) 896 – 907.
172. T. Bechtold, E. Burtscher, D. Gmeiner, O. Bobleter, *Melliand Textilber.* **72** (1991) 50 – 54. T. Bechtold, E. Burtscher, A. Amann, O. Bobleter, *Angew. Chem.* **104** (1992) 1046 – 1047.
173. U. Baumgarte, *Melliand Textilber. Int.* **55** (1974) 953 – 962.
174. T. Vickerstaff: *The Physical Chemistry of Dyeing,* Oliver & Bayd, London 1954.
175. U. Baumgarte, *Textilveredlung* **15** (1980) 413 – 424.

176. U. Baumgarte, *Melliand Textilber.* **59** (1978) 311–319.
177. E. I. Valko, *J. Am. Chem. Soc.* **59** (1941) 1433–1437.
178. BASF-Ratgeber Cellulosefasern: Schlichten, Vorbehandeln, Färben, Ludwigshafen 1977.
179. U. Baumgarte, H. Schlüter, *Melliand Textilber.* **62** (1981) 555–561.
180. P. Richter: "Kontinuefärben von Baumwollkettgarn mit Indigo," *Textilveredlung* **10** (1975) no. 8, 312–316.
181. BASF, Continuous Dyeing of Cotton Warp Yarn with Indigo, Technical Information, Ludwigshafen 1985.
182. J. A. Greer, G. R. Turner: "Indigo Denims: the Practical Side," *Text. Chem. Color.* **15** (1983) no. 6, 101–107.
183. L. Haas: Indigo-Färberei: Verfahrens- und maschinentechnische Lösungen, *Int. Text. Bull. Veredlung* **2**(1990) 45–50.
184. C. Heid, K. Holoubek, R. Klein: "100 Jahre Schwefelfarbstoffe," *Melliand Textilber.* **54** (1973) 1314–1327.
185. E. Krusche: "Die Bedeutung der Schwefelfarbstoffe heute und in der Zukunft," *Text. Prax.* **23** (1968) 342–344.
186. D. G. Orton: "Sulfur Dyes," in O. K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. VII, Academic Press, New York 1974, pp. 26–33.
187. *Colour Index,* **vol. 3,** pp. 3649–3704 (1971), **vol. 4,** pp. 4475–4501 (1971), **vol. 6 (Revised Third Edition),** pp. 6375–6385 (1975).
188. C. Heid: "Chemismus und Anwendung der wasserlöslichen Schwefelfarbstoffe," *Melliand Textilber.* **45** (1964) 648–652, 833.
189. W. Prenzel: "Der heutige Stand der Schwefelfärberei," *Chemiefasern Textind.* **24/76** (1974) 293–302.
190. C. Heid: "Neue Anwendungsmöglichkeiten von Hydrosol- und Hydrosol-Licht-Farbstoffen beim kontinuierlichen und halbkontinuierlichen Färben cellulosehaltiger Stückware," *Z. Gesamte Textilind.* **70** (1968) 626–630.
191. W. Titzka: "Das Färben mit Schwefelfarbstoffen unter besonderer Berücksichtigung der Reduktions- und Oxidationsmittel," *TPI Text. Prax. Int.* **33** (1978) 1387–1390.
192. M. Hähnke: "Die Schwefelfärbung in ökologisch anspruchsvoller Umwelt, Internat. Veredler-Jahrbuch," *Dtsch. Färber-Kal.* **96** (1992) 154–156.
193. L. Tigler: "Sulfur Dye Oxidation," *Am. Dyest. Rep.* **61** (1972) no. 9, 46, 107.
194. C. Heid: "Die Verbesserung der Wasch- und Peroxidwaschechtheit von Färbungen mit Schwefel- und Schwefelküpen-Farbstoffen durch Solidurit IH," *Melliand Textilber. Int.* **51** (1970) 322–325. C. Heid: "Theorie und Praxis der Verbesserung der Wasch- und Peroxidwaschechtheit von Färbungen mit Schwefel- und Schwefel-Küpenfarbstoffen," *Melliand Textilber.* **59** (1978) 247–254.
195. C. Heid: "Die Anwendung von Fibradurit FS bei Schwefelschwarzfärbungen zur Vermeidung von Faserschädigungen während des Trocknens und Lagerns." *Z. Gesamte Textilind.* **67** (1965) 364–367.
196. C. Heid: "Hydron-Blau, der Indigo des 20. Jahrhunderts," *TPI Text. Prax. Int.* **33** (1978) 285– 287. C. Klopfstock: "Färben von Frottiermaterialien mit Immedial-, Hydrosol-, Cassulfon-, Hydron-Blau- und Indanthren-Farbstoffen," *TPI Text. Prax. Int.* **32** (1977) 1119–1121, 1230–1232. W. Titzka: "Einsatz von Schwefelfarbstoffen in der Cord-Färberei," *Melliand Textilber.* **60** (1979) 254–256.
197. G. Krauzpaul: "Die Schwefel- und Schwefelküpenfarbstoffe in der Ausziehfärberei aus der Sicht des Praktikers," *TPI Text. Prax. Int.* **42** (1987) 140–142.
198. H. M. Tobin: "Jet Dyeing with Sulfur and Vat Colours," *Am. Dyest. Rep.* **68** (1979) no. 9, 26, 28.
199. H. Bernhardt: "Welche Problematik wird durch die Substantivität bei Klotzprozessen mit Schwefelfarbstoffen ausgelöst? Gibt es dafür Problemlösungen?" *Int. Text. Bull.* **33** (1987) 5–6, 8–9, 12, 16, 19. H. Bernhardt: "Der Einsatz des Taschencomputers bei der Rezepterstellung in der Kontinuefärberei mit Schwefelfarbstoffen," *Int. Text. Bull.* **33** (1987) 46, 50, 54, 57.
200. C. Heid: "Neue Fixiermöglichkeiten von wasserlöslichen Schwefelfarbstoffen," *Z. Gesamte Textilind.* **69** (1967) 106–108. Cassella Farbwerke, DE 1 619 403, 1967 (C. Heid, W. Gunzert).
201. R. Klein: "Phänomen Denim," *Textilveredlung* **10** (1975) 112–117.
202. Cassella Farbwerke, DE 1 276 596, 1963 (M. Stuhlmiller, G. Böttiger, C. Heid).
203. "Probleme und Problemlösungen textiler Abwässer," *Chemiefasern Textilind.* **28/80** (1978) no. 1, 85–89. J. M. Marzinkowski, W. Keiler: "Konzepte zur Vermeidung von

Abwasserbelastungen in einem Textilveredlungsbetrieb," *Chemiefasern Textilind.* **41/93** (1991) 895 – 901.
204. B. Kerres, G. Valk: "Verfahren zur Beseitigung von Geruchsbelästigungen aus dem öffentlichen Kanalnetz bei Einleitung von Färbereiabwässern mit Schwefelfarbstoffen," *Melliand Textilber.* **66** (1985) 588 – 590.
205. B. Smith, J. Rucker: "Water and Textile Wet Processing – Part II," *Am. Dyest. Rep.* **76** (1987) 68 – 78.
206. C. Oehme: "Abwasserklärung nach dem Katox-F-System," *TPI Text. Prax. Int.* **33** (1978) 1076 – 1077. B. Höke: "Die Reinigungsleistung des Katox F-Verfahrens," *Textilbetrieb (Würzburg)* **96** (1978) no. 6, 92 – 94;nos. 7, 8, 60 – 62. G. Wysocki: "Katox-Fällungsverfahren für die Reinigung von Abwässern der Textilveredlungsindustrie: Technologie und Wirtschaftlichkeit," *Chemiefasern Textilind.* **26/78** (1976) 353 – 356.
207. W. Schwindt, G. Faulhaber, A. J. Moore, *Rev. Prog. Color. Relat. Top.* **2** (1971) 33.
208. T. Lever, *J. Soc. Dyers Colour.* **108** (1992) 477 – 478.
209. H. U. von der Eltz, H. Walbrecht, *Dtsch. Färber-Kal.* **77** (1973) 79 – 129.
210. H. U. von der Eltz, *Textilveredlung* **7** (1972) 692 – 697.
211. H. Rath: *Lehrbuch der Textilchemie,* 3rd ed., Springer Verlag, Berlin 1972.
212. U. Baumgarte, *Rev. Prog. Color. Relat. Top.* **5** (1974) 17 – 32.
213. R. H. Peters: *Textile Chemistry,* vol. 3, Elsevier, Amsterdam 1975.
214. A. Schäffer: *Handbuch der Färberei,* vol. 1, Konradin-Verlag R. Kohlhammer, Stuttgart 1949.
215. *Ullmann,* 4th ed., **22,** p. 656.
216. H. Vollmann: "Phthalogen Dyestuffs," in K. Venkataraman (ed.): *The Chemistry of Synthetic Dyes,* vol. 5, Academic Press, New York 1971, p. 283.
217. M. Peter, H. K. Rouette, *Grundlagen der Textilveredlung,* 13th ed., Deutscher Fachverlag, Frankfurt 1989.
218. E. Elöd, H. Reutter, *Melliand Textilber.* **19** (1938) 67 – 72.
219. J. D. Leeder, J. A. Rippon, F. E. Rothery, I. W. Stapleton, *Proc. Int. Wool Text. Res. Conf. 7th* **vol. V,** Tokyo, 1985,vol. V, 99 – 108.
220. J. D. Leeder: "The Cell Membrane Complex and its Influence on the Properties of the Wool Fibre," *Wool Sci. Rev.* **3** (1987) 3 – 35.
221. V. Sideris, L. A. Holt, I. H. Leaver, *J. Soc. Dyers Colour.* **106** (1990) 131 – 135.
222. H. Zahn, *Proc. Int. Wool Text. Res. Conf. 6th* Pretoria 1980,suppl. p. 1.
223. M.-L. Klotz, U. Altenhofen, H. Zahn, *Textilveredlung* **21** (1986) 119 – 121.
224. H. Zahn, *Lenzinger Ber.* **42** (1977) 19 – 34.
225. Forschungsprojekt des DWI: Interne Lipide der Wolle: Zusammensetzung, Veränderung und Bedeutung bei Vergilbung und bei Veredlungsprozessen von Wolle (AIF-Nr. 7160).
226. D. Mäusezahl, *Textilveredlung* **5** (1970) 839 – 845.
227. H. Zollinger, G. Back, B. Milicevic, A. N. Roseira, *Melliand Textilber.* **42** (1961) 73 – 80.
228. F. Hoffmann, *Rev. Prog. Color. Relat. Top.* **18** (1988) 56 – 64.
229. F. Hoffmann, W. Langmann, H. Elberzhager, M. Schnee, *Int. Text. Bull. Färberei/Druckerei/-Ausrüstung* **no. 3** (1978), 205, 206, 211, 212, 217, 218.
230. H. Flensberg, W. Mosimann, H. Salathé, *Melliand Textilber.* **65** (1984) 472 – 477.
231. D. Hildebrand, H. Zahn, *SVF Fachorgan Textilveredl.* **13** (1958) 376 – 385.
232. K. Reincke, *Melliand Textilber.* **74** (1993) 408 – 417.
233. W. Ender, A. Müller, *Melliand Textilber.* **19** (1938) 182 – 183.
234. F. R. Hartley, *J. Soc. Dyers Colour.* **85** (1969) 66 – 71.
235. O. K. Dobozy, *Am. Dyest. Rep.* **62** (1973) no. 3, 36 – 48, 76 – 77.
236. D. M. Lewis, G. Yan, *J. Soc. Dyers Colour.* **109** (1993) 193 – 197.
237. P. Spinacci, N. C. Gaccio, 12th IFVTCC Congr., Budapest 1981, pp. 10 – 13.
238. P. A. Duffield, R. D. D. Holt, *Textilveredlung* **24** (1989) 40 – 45.
239. L. Benisek, *Textilveredlung* **12** (1977) 406 – 415; *J. Soc. Dyers Colour.* **94** (1978) 101 – 105.
240. G. Meier, *TPI Text. Prax. Int.* **31** (1976) 898 – 901.
241. A. C. Welham, *J. Soc. Dyers Colour.* **102** (1986) 126 – 131.
242. I. D. Rattee, *J. Soc. Dyers Colour.* **69** (1953) 288 – 295.
243. D. de Meulemeester, I. Hammers, W. Mosimann, Aachener Textiltagung 1989; *Melliand Textilber.* **71** (1990) 69.
244. F. Beffa, G. Back, *Rev. Prog. Color. Relat. Top.* **14** (1984) 33 – 42.

245. G. Schetty, W. Kuster, *Helv. Chim. Acta* **44** (1961) 2193. H. Pfitzner, *Melliand Textilber.* **35** (1954) 649 – 651.
246. G. Schetty, *J. Soc. Dyers Colour.* **71** (1955) 705.
247. J. A. Bone, *Schriftenr. Dtsch. Wollforschungsinst. Tech. Hochsch. Aachen* **93** (1984) 170 – 178.
248. W. Mosimann, *Textilveredlung* **17** (1982) 289 – 295.
249. C. Mancheno, H.-W. Hensen, I. Souren, H.-K. Rouette, *Melliand Textilber.* **71** (1990) 875 – 882.
250. G. v. Hornuff, H. J. Flath, *Faserforsch. Textiltech.* **12** (1961) 559 – 567.
251. D. M. Lewis, *Melliand Textilber.* **67** (1986) 717 –723.
252. D. M. Lewis, *J. Soc. Dyers Colour.* **98** (1982) 165 –175.
253. K. Hannemann, F. Grüner, *Int. Text. Bull. Veredlung* **38** (1992) no. IV, 22 – 31.
254. H. Angelmahr, F. Bartsch, *TPI Text. Prax. Int.* **47** (1992) 455 – 456.
255. A. Würz, *Melliand Textilber.* **42** (1961) 439 – 444.
256. H. Baumann, *Textilveredlung* **14** (1979) 515 – 522.
257. D. Schwer, H. Ritter, K. Zesiger, *Textilveredlung* **16** (1981) 479 – 484.
258. H. Flensberg, W. Mosimann, H. Salathé, *Schriftenr. Dtsch. Wollforschungsinst. Tech. Hochsch. Aachen* **93** (1984) 150 – 169.
259. D. M. Lewis, *Rev. Prog. Color. Relat. Top.* **19** (1989) 49 – 56.
260. W. Mosimann, *Am. Dyest. Rep.* **80** (1991) no. 3, 26, 28, 30, 32, 34, 62.
261. F. Blum, *Bayer Farben Rev.* **7** (1964) 1 – 9.
262. R. Hofstetter, *Melliand Textilber.* **72** (1991) 366 –373.
263. F. Shimizu, K. Jyoko, I. Sakaguchi, *Sen'i Gakkaishi* **32** (1976) 388 – 393.
264. H. Putze, *TPI Text. Prax. Int.* **39** (1984) 1051 – 1054.
265. A. Liddiard, *J. Soc. Dyers Colour.* **99** (1983) 56 – 59.
266. R. Rohrer, *Textilveredlung* **20** (1985) 85 – 87.
267. K. Y. Chu, J. R. Provost, *Rev. Prog. Color. Relat. Top.* **17** (1987) 22 – 28.
268. A. Kremer, *TPI Text. Prax. Int.* **47** (1987) 746 – 757.
269. A. Kremer, H. Kremer, *Melliand Textilber.* **73** (1992) 580 – 582.
270. M. I. Guljarani, *Rev. Prog. Color. Relat. Top.* **22** (1992) 79 – 89.
271. R. H. Peters, *J. Soc. Dyers Colour.* **61** (1945) 95 –100.
272. A. Anton, *Text. Chem. Color.* **13** (1981) 46 – 50.
273. G. Kühnel, H. J. Flath, R. Gärtner, *Melliand Textilber.* **72** (1991) 288 – 290;360 – 362.
274. J. A. Coates, V. Ellard, I. D. Rattee, *J. Soc. Dyers Colour.* **96** (1980) 14 – 18.
275. I. D. Rattee, S. So, *Mater. Sci. Monogr.* **21** (1984) 575 – 584.
276. B. Catiore, R. Hagege, *Bull. Sci. Inst. Text. Fr.* **2** (1973) 209.
277. G. Kühnel, Dissertation, TU Dresden 1989.
278. F. Hoffmann, *Chemiefasern/Textilind.* **31/83** (1981) 762 – 768.
279. B. von Falkai, H. Wilsing, *TPI Text. Prax. Int.* **33** (1978) 1324 – 1330.
280. D. Dornig, Dissertation, TU Dresden 1974.
281. E. Elöd, T. Schachowsky, *Melliand Textilber.* **23** (1942) 437 – 440.
282. T. Iijima, H. Zollinger, *Teintex* **32** (1967) 245 – 263.
283. H. Zollinger, G. Back, B. Milicevic, A. N. Roseira, *Melliand Textilber.* **42** (1961) 73 – 80.
284. E. Sada, H. Kumazawa, T. Ando, *J. Soc. Dyers Colour.* **98** (1982) 121 – 125.
285. V. S. Praptowidodo, K. Hamada, T. Iijima, *Angew. Makromol. Chem.* **144** (1986) 159 – 165.
286. J. Meybeck, P. Galafassi, *Appl. Polym. Symp.* **18** (1971) 463 – 472.
287. D. Schwer, *Textilveredlung* **23** (1988) 296 – 301.
288. W. Beckmann, F. Hoffmann, H.-G. Otten, *Melliand Textilber. Int.* **54** (1973) 641 – 646; *J. Soc. Dyers Colour.* **88** (1972) 354 – 360.
289. W. Langmann, *Melliand Textilber. Int.* **54** (1973) 654 – 659.
290. H. Beiertz, *Am. Dyest. Rep.* **68** (1979) no. 6, 22 – 26.
291. F. Hoffmann, *Rev. Prog. Color. Relat. Top.* **18** (1988) 56 – 64.
292. B. C. Burdell, *Rev. Prog. Color. Relat. Top.* **13** (1983) 41 – 49.
293. O. Annen, J. Carbonell, *Textilveredlung* **15** (1980) 296 – 302.
294. H. Scheidegger, H. Flensberg, R. Bauhofer, *Textilveredlung* **13** (1978) 302 – 304.
295. J. R. Aspland, *Text. Chem. Color.* **25** (1993) 19 – 23.
296. J. Shore, *J. Soc. Dyers Colour.* **87** (1971) 37 – 48.
297. J. Guthrie, *J. Appl. Polym. Sci.* **27** (1982) 2567 – 2575.
298. C. C. Cook, *Rev. Prog. Color. Relat. Top.* **12** (1982) 73 – 89.

299. T. F. Cooke, H. D. Weigmann, *Rev. Prog. Color. Relat. Top.* **20** (1990) 10 – 18.
300. O. Annen, *Textilveredlung* **19** (1984) 8 – 11.
301. H. Egli, *Textilveredlung* **2** (1967) 856 – 864.
302. Bayer, Musterkarte Telon-S-Verfahren Sp 502, Leverkusen.
303. R. Weber, *Melliand Textilber.* **58** (1977) 48 – 51.
304. W. Hartmann, *Am. Dyest. Rep.* **69** (1980) no. 6, 21 –22.
305. M. Mitter, *Chemiefasern/Text. Anwendungstech.* **31** (1981) 55 – 56, 558.
306. I. Strelecky, *Ciba Rundsch.* 1968, no. 2, 40 – 44.
307. W. R. Remington, E. K. Gladding, *J. Am. Chem. Soc.* **72** (1953) 2553.
308. M. Greenhalgh, A. Johnson, R. H. Peters, *J. Soc. Dyers Colour.* **78** (1962) 315 – 321.
309. N. J. Bhatt, E. H. Daruwalla, *Textile Res. J.* **34** (1964) 435 – 444.
310. G. Valk, G. Stein, A. El Bendak, *Melliand Textilber.* **57** (1976) 166; **58** (1977) 580 – 582. G. Valk, G. Stein, *Melliand Textilber.* **59** (1978) 593 – 594.
311. D. Dornig, H. Sehm, *Melliand Textilber.* **74** (1993) 212 – 215.
312. Faserstoff-Tabellen according to P. A. Koch, *Chemiefasern/Textilind.* **43** (1993) no. 6, 508 – 522.
313. K. H. Meyer, *Melliand Textilber.* **6** (1925) 737 – 739; V. Kartaschoff, *Helv. Chim. Acta* **8** (1925) 928 –942; **9** (1926) 152 – 173.
314. T. Vickerstaff: *The Physical Chemistry of Dyeing,* Oliver & Boyd, London 1954.
315. R. H. Peters: *Textile Chemistry,* vol. 3: The Physical Chemistry of Dyeing, Elsevier, Amsterdam 1975.
316. H. Braun: "Particle Size and Solubility of Disperse Dyes," *Rev. Prog. Color. Relat. Top.* **13** (1983) 62 – 72.
317. H. Leube, *Text. Chem. Color.* **10** (1978) no. 2, 32 –39, 39 – 46.
318. F. Jones, *J. Soc. Dyers Colour.* **100** (1984) no. 2, 66 –72.
319. G. L. Baughman, T. A. Perenich, *Text. Chem. Color.* **21** (1989) no. 1, 33 – 37.
320. H. Leube, W. Rüttiger, *Melliand Textilber.* **59** (1978) 836 – 842.
321. T. M. Baldwinson, *Rev. Prog. Color. Relat. Top.* **15** (1985) 6 – 14.
322. H. Gerber, *Melliand Textilber.* **69** (1988) 195 –200.
323. A. Keil, H. Noack, H. J. Flath, *Textiltechnik* **40** (1990) 264 – 267, 318 – 320, 382 – 384, 432 – 437.
324. R. Turner, *Text. Chem. Color.* **15** (1983) no. 9, 180/61 – 188/69.
325. D. M. Nunn (ed.): *The Dyeing of Synthetic Polymer and Acetate Fibres,* Dyers Comp. Publ. Trust, Bradford 1979.
326. F. Somm, R. Buser, *Int. Dyer Printer* **170** (1985) no. 12, 6 – 7.
327. H. Gerber, *Textilveredlung* **8** (1973) 449 – 456.
328. A. Maier, AATCC Int. Dyeing Symp., Washington 1977, pp. 66 – 69.
329. R. A. Walsh, *Text. Chem. Color.* **7** (1975) no. 10, 184/35 – 187/38.
330. Disperse Dye Comm., *J. Soc. Dyers Colour.* **93** (1977) 228 – 237.
331. M. Hammoudeh, E. Schönpflug, *Melliand Textilber.* **52** (1971) 1063 – 1068.
332. BASF Manual: Dyeing and Finishing of Polyester Fibers B 363 e, pp. 384 – 501, 1975.
333. J. Park, *Rev. Prog. Color. Relat. Top.* **15** (1985) 26 –27; T. M. Baldwinson, 6 – 14.
334. G. R. Turner, *Text. Chem. Color.* **15** (1983) no. 9, 61 – 69.
335. E. Podogrocka, *Melliand Textilber.* **70** (1989) 198 –199.
336. K. Poulaikis et al., *Chemiefasern/Textilind.* **41** (1991) 142 – 147.
337. W. Saus, D. Knittel, E. Schollmeyer, *TPI Text. Prax. Int.* **48** (1993) 32 – 36; *Text. Res. J.* **63** (1993) no. 3, 135 – 142.
338. F. Jones, J. Kraska, *J. Soc. Dyers Colour.* **82** (1960) 333 – 338.
339. Ch. Guth, *Textilveredlung* **14** (1979) 270 – 274.
340. AATCC Palmetto Section, *Text. Chem. Color.* **11** (1979) no. 12, 270/23 – 282/35.
341. Nippon Kayaku, JP 59 066 585, 1982 (N. Mitsuro).
342. J.-H. Chiao-Cheng, B. M. Reagan, *Text. Chem. Color.* **15** (1983) 12 – 19.
343. C. A. Li, *Chem. Abstr.* **100** (1984) no. 24, 193 415.
344. W. Griesser, H. Tiefenbacher, *Textilveredlung* **28** (1993) no. 4, 88 – 96.
345. M. Laufer, *Textilveredlung* **28** (1993) no. 4, 96 –101.
346. P. Richter, *Chemiefasern/Textilind.* **41/93** (1991) no. 9, 1118 – 1125, E 129 – 133.
347. W. Beckmann, H. Hamacher, *Chemiefasern + Text. Anwendungstech. Text. Ind.* **23** (1973) no. 5, 436 –439.
348. C. Renard, *Melliand Textilber.* **54** (1973) 1328 –1335.
349. J. Cegarra, F. J. Carrion, P. Puente, *Melliand Textilber.* **65** (1984) 405 – 409.

350. C. Grardel, *Ind. Text. (Paris)* **1155** (1985) no. 5, 519 – 520.
351. H. Beiertz, *Chemiefasern/Textilind.* **42/94** (1992)no. 1, 64 – 67, E 1 – 3; H. Beiertz, K. H. Röstermundt, *TPI Text. Prax. Int.* **48** (1993) no. 2, 135 –137.
352. G. Reinert, V. Misun, *Melliand Textilber. Int.* **74** (1993) 1007 – 1014.
353. Report of Committee: "Dyeing Properties of Disperse Dyes, I Cellulose Acetate," *J. Soc. Dyers Colour.* **80** (1964) 237 – 242; "II Cellulose Triacetate," *J. Soc. Dyers Colour.* **81** (1965) 209 – 210.
354. H. Leube, *Z. Gesamte Textilind.* **70** (1968) no. 4 to **71** (1969) no. 2; English version see: Dyeing and Finishing of Acetate and Triacetate and their Blends with Other Fibres, BASF Manual S 388 e.
355. D. Blackburn in D. M. Nunn (ed.): *The Dyeing of Synthetic Polymer and Acetate Fibres,* Comp. Publ. Trust, Bradford 1979, pp. 76 – 128.
356. R. H. Peters: "The Physical Chemistry of Dyeing," *Textile Chemistry,* vol. 3, Elsevier, Amsterdam 1975.
357. Stanford Research Inst.: Chemical Economics Handbook, Fibers 5 433 000 D, Menlo Park, Sept. 1993.Comité International de la Rayonne et des Fibres Synthetiques, Bruxelles, Brochures 1993.
358. T. A. Field, *Am. Dyest. Rep.* **41** (1952) 475.
359. R. Raue, *Rev. Prog. Color. Relat. Top.* **14** (1984) 187.
360. W. Beckmann in D. M. Nunn (ed.): *The Dyeing of Synthetic Polymer and Acetate Fibres,* The Dyer's Company, Bradford 1979, p. 359 ff.
361. I. Holme, *Rev. Prog. Color. Relat. Top.* **13** (1983) 10.
362. F. Jones, *Rev. Prog. Color. Relat. Top.* **4** (1973) 64.
363. H. Sumner in A. Johnson (ed.): *The Theory of Colouration of Textiles,* chap. 2, Society of Dyers and Colourists, Bradford 1989.
364. O. Glenz, W. Beckmann, *Melliand Textilber.* **38** (1957) 296, 783, 1152.
365. W. Beckmann, *J. Soc. Dyers Colour.* **77** (1961) 616.
366. S. Rosenbaum, *Text. Res. J.* **33** (1963) 899.
367. U. Mayer, W. Ender, A. Würz, *Melliand Textilber.* **47** (1966) 653, 772.
368. G. Alberghina, M. Amato, S. Fisichella, *J. Soc. Dyers Colour.* **105** (1989) 163.
369. H. Zollinger, *Text. Res. J.* **59** (1989) 264.
370. P. R. Brady, *Rev. Prog. Color. Relat. Top.* **22** (1992) 58.
371. A. V. Hill, *Proc. R. Soc. London B* **104** (1928) 39, 65.
372. W. Beckmann, *11th Int. Textile Seminar,* Kingston, Ontario, 1968, Book of Papers, p. 97; *Z. Gesamte Textilind.* **71** (1969) 603.
373. S. Rosenbaum, *J. Polym. Sci. Part A* **3** (1965) 1949.
374. M. Williams, R. Landel, J. Ferry, *J. Am. Chem. Soc.* **77** (1955) 3701.
375. R. Asquith, H. Blair, N. Spence, *Text. Res. J.* **47** (1977) 446; *J. Soc. Dyers Colour.* **94** (1978) 49.
376. S. Rosenbaum, *Text. Res. J.* **34** (1964) 159.
377. D. Balmforth, C. Bowers, T. Guion, *J. Soc. Dyers Colour.* **80** (1964) 577.
378. D. G. Evans, G. J. Bent: "Basic Dyes on Acrylic Fibres" *J. Soc. Dyers Colour.* **89** (1973) 292.
379. J. Neary, R. Thomas, *Am. Dyest. Rep.* **46** (1957) 625.
380. I. Holme, *Rev. Prog. Color. Relat. Top.* **1** (1970) 31.
381. S. Cohen, A. Endler, *Am. Dyest. Rep.* **46** (1957) 625; **47** (1958) 325.
382. S. Heimann, F. Feichtmayr, *Z. Gesamte Textilind.* **68** (1966) 509.
383. F. Hoffmann, *Melliand Textilber.* **59** (1978) 239.
384. C. Zimmermann, A. Cate, *Am. Dyest. Rep.* **83** (1972) 150.
385. Bayer, CA 793 184, 1968 (W. Beckmann, J. Nentwig, H. Rudolf, J. Schneider).
386. M. Dullaghan, A. Ultee, *Text. Chem. Color.* **8** (1976) 22.
387. S. Shukla, M. Mathur, *J. Soc. Dyers Colour.* **109** (1993) 330.
388. J. Leddy, *Am. Dyest. Rep.* **49** (1960) 272.
389. D. G. Evans, C. J. Bent: "Basic Dyes on Acrylic Fibres Comm." *J. Soc. Dyers Colour.* **88** (1972) 220.
390. F. Hoffmann, *Rev. Prog. Color. Relat. Top.* **18** (1988) 56.
391. AATCC Metropolitan Sect., *Text. Chem. Color.* **8** (1976) 165.
392. J. Park, J. Shore, *J. Soc. Dyers Colour.* **97** (1981) 223. J. Park: *A Practical Introduction to Yarn Dyeing,* Society of Dyers and Colourists, Bradford 1980.
393. W. Biedermann, *Rev. Prog. Color. Relat. Top.* **10** (1979) 1.
394. F. Hoffmann, M. Schnee, H. Schubert, *Melliand Textilber.* **65** (1984) 478.
395. F. Hoffmann, *Text. Chem. Color.* **22** (1990) 11.
396. A. Kretschmer, *TPI Text. Prax. Int.* **44** (1989) 1098, 1316; **45** (1990) 225; **46** (1991) 1220.

397. W. Beckmann, K. Jakobs, *Bayer Farben Rev.* **16** (1968) 1.
398. F. Hoffmann, H. Schubert, J. Fiegel, *TPI Text. Prax. Int.* **47** (1992) 233.
399. P. Dorsch, H. Wilsing, K.-H. Peters, *Melliand Textilber.* **62** (1981) 188.
400. R. Rohner, H. Zollinger, *Text. Res. J.* **56** (1986) 1.
401. U. Meyer, J. Zhang, H. Zollinger, *Textilveredlung* **19** (1984) 39.
402. U. Reinehr, G. Häfner, A. Nogaj, *Chemiefasern/Textilind.* **35** (1985) no. 87, 588.
403. G. Müller, *Textilveredlung* **10** (1975) 438.
404. L. Witt, P. Toldrian, *Chemiefasern/Textilind.* **28** (1978) no. 80, 469.
405. J. Shore, *Rev. Prog. Color. Relat. Top.* **10** (1979) 33.
406. W. Beckmann, W. Langmann, *Dtsch. Textiltech.* **18** (1968) 243.
407. H. Gerber, H. Lehmann, F. Somm, *Melliand Textilber. Int.* **54** (1973) 77.
408. L. Kostova, R. Iltscheva, R. Detscheva, *Textilveredlung* **27** (1993) 398.
409. I. Hardalov, J. Mikhailova, *J. Soc. Dyers Colour.* **109** (1993) 369.
410. R. Weber, *Text. Prax.* **21** (1966) 670.
411. W. Ingamells, *J. Soc. Dyers Colour.* **96** (1980) 332.
412. P. Ackroyd, *Rev. Prog. Color. Relat. Top.* **5** (1974) 86.
413. R. Detscheva, R. Iltscheva, K. Dimov, *Textilveredlung* **9** (1974) 312; **10** (1975) 318.
414. A. Läpple, A. Schneider, *Textilveredlung* **10** (1975) 63.
415. I. Holme, *Rev. Prog. Color. Relat. Top.* **7** (1976) 1.
416. D. M. Nunn (ed.): *The Dyeing of Synthetic-polymer and Acetate Fibres,* Dyers Comp. Publ. Trust, Bradford 1979.
417. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung,* 13th ed., Deutscher Fachverlag, Frankfurt a. M. 1989.
418. H. W. Partridge, *Rev. Prog. Color. Relat. Top.* **6** (1975) 56.
419. BASF: technical information TI/T 138 e, Ludwigshafen 1986.
420. BASF: technical information TI/T 151 e, Ludwigshafen 1985.
421. G. Biehler, *Chemiefasern/Textilind.* **29/81** (1979) 848 – 853.
422. H. Leube, *Melliand Textilber.* **46** (1965) 743 – 749.
423. J. Shore, *Rev. Prog. Color. Relat. Top.* **6** (1975) 7 –12.
424. J. Shore in [416], pp. 393 – 403.
425. R. H. Peters: *Textile Chemistry,* vol. 3: The Physical Chemistry of Dyeing, Elsevier, Amsterdam 1975.
426. B. D. Gupta, A. K. Mukherjee, *Rev. Prog. Color. Relat. Top.* **19** (1989) 7 – 19.
427. *Ullmann,* 3rd ed. 7, 46 – 54.
428. BASF Manual Dyeing and Finishing of Polyester Fibres B 363 e, 1975.
429. J. Shore in D. M. Nunn (ed.): *The Dyeing of Synthetic-polymer and Acetate Fibres,* Dyers Comp. Publ. Trust, Bradford 1979.
430. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung,* 13th ed., Deutscher Fachverlag, Frankfurt a. M. 1989.
431. BASF Manual Dyeing and Finishing of Cellulose Fibres B 375 e, pp. 383 – 414, 1975.
432. K. H. Rostermundt, *TPI Text. Prax. Int.* **47** (1992) no. 7, 649 – 654; and XIX – XXIII.
433. S. M. Doughty, *Rev. Prog. Color. Relat. Top.* **16** (1986) 25 – 38.
434. W. Haertl, *Textilveredlung* **24** (1989) no. 6, 214 – 218.
435. D. Blackburn, *Int. Dyer Text. Printer, Bleacher Finish.* **171** (1986) no. 1, 13 – 14.
436. A. Berger-Schunn: *Praktische Farbmessung,* Muster Schmidt, Göttingen 1991.
437. ISO/CD 787-26 (Draft).
438. W. Baumann et al., *Melliand Textilber.* **67** (1986) 562 – 566; *J. Soc. Dyers Colour.* **103** (1987) 100 –105.
439. R. Broßmann et al., *Melliand Textilber.* **67** (1986) 499 – 502; *J. Soc. Dyers Colour.* **103** (1987) 38 – 42.
440. A. Berger-Schunn et al., *Melliand Textilber.* **67** (1986) 638 – 639, 716 – 717, 812 – 813; *J. Soc. Dyers Colour.* **103** (1987) 138 – 139, 140 – 141, 272 – 274.
441. T. v. Chambers et al., *Melliand Textilber.* **69** (1988) 755 – 758; *J. Soc. Dyers Colour.* **105** (1989) 214 – 218.
442. Brit. Occup. Hyg. Soc., Technical Guide no. 4, part 1 and 2 (1985).
443. A. Berger-Schunn et al., *Melliand Textilber.* **70** (1989) 690 – 693; *J. Soc. Dyers Colour.* **107** (1991) 270 – 273.
444. *J. Soc. Dyers Colour.* **88** (1972) 220 – 222.
445. P. J. Smith, *Rev. Prog. Color. Relat. Top.* **24** (1994) 31 – 40.
446. TVI-Verband: Water/Wastewater Investigation 1992 (Association of the Textile Finishing Industry), Frankfurt a. M. 1993.
447. Arbeitsbericht der ATV-Arbeitsgruppe Textilveredlungsindustrie (Working Report of the ATV Working Group Textile Finishing Industry), *Korresp. Abwasser* **36** (1989) 1074 – 1084.

448. Gesetz zur Ordnung des Wasserhaushaltes, Wasserhaushaltsgesetz, Neufassung vom 12.2.1990, BGBl. I (Water Resources Act, revised February 12, 1990, Federal Law Gazette), p. 205
449. Abwasserherkunftsverordnung vom 3.7.1987, geändert am 27.5.1991, BGBl. I. (Wastewater Origin Decree, July 3, 1987, revised May 27, 1991, Federal Law Gazette), p. 1197.
450. Allgemeine Rahmenabwasserverwaltungsvorschrift über Mindestanforderungen an das Einleiten von Abwasser in Gewässer, Neufassung vom 29.10.1992, GMBl 1992 (General Legal Framework on Minimum Requirements on Discharge of Wastewater into Surface Waters, revised October 29, 1992, Joint Ministerial Law Gazette 1992), pp. 1265 – 1266.
451. 38. Abwasserverwaltungsvorschrift über Mindestanforderungen an das Einleiten von Abwasser in Gewässer (Textilherstellung) vom 5.9.1984, GMBl. [38th Wastewater Ordinance on Minimum Requirements for the Discharge of Wastewater into Surface Waters (Textile Processing), September 5, 1984, Joint Ministerial Law Gazette] 1984, p. 348.
452. Bundesratsdrucksache 181/93 (Printed Matter of the German Bundesrat), pp. 13 – 21.
453. 612. Verordnung des Bundesministers für Land- und Forstwirtschaft über die Begrenzung von Abwasseremissionen aus Textilveredlungs- und Behandlungsbetrieben, BGBl. 207. Stück v. 24.9.1992 (612. Ordinance of the Federal Minister for Agriculture and Forestry on Limitations of Discharge from Textile Finishing and Processing Plants, Federal Law Gazette, September 24, 1992), pp. 2828 –2835.
454. The Water Act (London. HMSO).
455. I. G. Laing, *Rev. Prog. Color. Relat. Top.* **21** (1991) 56 – 71.
456. P. Cooper, *J. Soc. Dyers Colour.* **108** (1992) 176 –182.
457. H. Reetz, *Melliand Textilber.* **72** (1991) 932 –933.
458. The Government of Japan, Environment Agency: Annual Book, Quality of the Environment in Japan in 1992.
459. B. Smith, J. Rucker, *Am. Dyest. Rep.* **76** (1987) 68 –78.
460. Sewerage Bureau of Osaka Municipal Government: The Sewerage Law and Regulations for Control of Water Quality, Osaka 1990.
461. U. Sewekow, *Melliand Textilber.* **70** (1989) 589 –596.
462. U. Rössner, *Textiltechnik (Leipzig)* **38** (1988) 661 –665.
463. B. Smith, *Am. Dyest. Rep.* **78** (1989) 26 – 32.
464. W.-D. Kermer, *Melliand Textilber.* **69** (1988) 586 –590.
465. B. M. Müller, *Rev. Prog. Color. Relat. Top.* **22** (1992) 14 – 21.
466. J. Park, J. Shore, *J. Soc. Dyers Colour.* **100** (1984) 383 – 399.
467. J. Wolff, H. Henk, *Textilveredlung* **25** (1990) 213 –218.
468. W. Beckmann, U. Sewekow, *TPI Text. Prax. Int.* **46** (1991) 346 – 348, 445 – 449.
469. F. Hoffmann, H. Schubert, J. Fiegel, *TPI Text. Prax. Int.* **47** (1992) 233 – 237.
470. R. Puk, D. Sedlak, *TPI Text. Prax. Int.* **47** (1992) 238 – 245.
471. J. M. Marzinkowski, *Chemiefasern/Textilind.* **41/93** (1991) 895 – 901.
472. P. A. Duffield, K.-H. Hoppen, *Melliand Textilber.* **68** (1987) 195 – 202.
473. P. A. Duffield, R. R. D. Holt, J. R. Smith, *Melliand Textilber.* **72** (1991) 938 – 942.
474. W. Sebb, *TPI Text. Prax. Int.* **44** (1989) 841 – 843.
475. F. Conzelmann, P. Wurster, A. Zahn, *TPI Text. Prax. Int.* **44** (1989) 644 – 649.
476. G. Rösch, *TPI Text. Prax. Int.* **45** (1990) 495 – 499.
477. W. Beckmann, J. Pflug, *TPI Text. Prax. Int.* **38** (1983) 160, 165 – 168.
478. R. Teichmann, *Melliand Textilber.* **74** (1993) 900 –902.
479. H.-U. von der Eltz, *TPI Text. Prax. Int.* **45** (1990) 727 – 731.
480. D. Fiebig, D. Soltau, *TPI Text. Prax. Int.* **43** (1988) 644, 649 – 650; **44** (1989) 533 – 535; **44** (1989) 1124 –1126.
481. D. Frahne, S. Koscielski, *Melliand Textilber.* **68** (1987) 594 – 596.
482. N. Athanasopoulos, *Melliand Textilber.* **71** (1990) 619 – 628.
483. U. Sewekow, G. Diesterweg, *Textilveredlung* **26** (1991) 142 – 146.
484. A. Wilking, D. Frahne, *Melliand Textilber.* **74** (1993) 897 – 900.
485. C. Oehme, *Melliand Textilber.* **67** (1986) 582 – 588.
486. J. Janitza, S. Koscielski, H. Schnabel, *TPI Text. Prax. Int.* **46** (1991) 1216 – 1219.
487. P. J. Halliday, S. Beszedits, *Can. Text. J.* **103** (1986) 78 – 84.

488. E. Thomanetz, D. Bardtke, E. Köhler, *GWF, Gas-Wasserfach: Wasser/Abwasser* **128** (1987) 432 – 441, 475 – 481.
489. M. Seekamp, *Textilveredlung* **25** (1990) 125 – 129.
490. W. Marte, W. Keller, *Textilveredlung* **26** (1991) 224 – 230.
491. G. von Hagel, *Korresp. Abwasser* **33** (1986) 908 –915.
492. T. R. Demmin, K. D. Uhrich, *Am. Dyest. Rep.* **77** (1988) 13 – 14, 17 – 18, 32.
493. W. C. Tincher, *Text. Chem. Color.* **21** (1989) 33 – 35.
494. A. E. Wilcock, S. P. Hay, *Can. Text. J.* **108** (1991) 37 – 44.
495. A. E. Wilcock, M. Brewster, W. C. Tincher, *Am. Dyest. Rep.* **81** (1992) 15 – 22.
496. G. Schulz, D. Fiebig, H. Herlinger, *Textilveredlung* **23** (1988) 1 – 4.
497. G. Schulz, H. Herlinger, G. Mayer, *Textilveredlung* **25** (1990) 23 – 27.
498. A. Hövelmann, S. C. Bidinger, A. Linder, *TPI Text. Prax. Int.* **48** (1993) 507 – 509.
499. K. H. Gregor, *Melliand Textilber.* **71** (1990) 976 –979.
500. U. Sewekow, *Melliand Textilber.* **74** (1993) 153 –157.
501. D. L. Michelsen, W. W. Powell, R. M. Woodby, L. L. Fulk, AATCC Technical Conference, Book of Papers, Atlanta 1992, pp. 135 – 150.
502. J. M. Green, C. Sokol, *Am. Dyest. Rep.* **74** (1985) 50 – 51, 67.
503. M. Kolb, B. Müller, B. Funke, *Vom Wasser* **69** (1987) 217 – 223.
504. J. Carriére, J. P. Jones, A. D. Broadbent, AATCC Technical Conference, Book of Papers, Atlanta 1992, pp. 231 – 236.
505. M. Pitroff, K. H. Gregor, *Melliand Textilber.* **73** (1992) 526 – 529.
506. H. Reissig, T. Jentsch, R. Fischer, *Gewässerschutz, Wasser, Abwasser* **125** (1991) 275 – 304.
507. G. Schulz, H. Herlinger, F. U. Gähr, T. Lehr, *TPI Text. Prax. Int.* **47** (1992) 1055 – 1062.
508. M. Kolb, P. Korger, B. Funke, *Melliand Textilber.* **69** (1988) 286 – 287.
509. M. Kolb, B. Funke, A. Baur, N. Peschen, *Korresp. Abwasser* **32** (1985) 986 – 987.
510. E. Burtscher et al., *Melliand Textilber.* **74** (1993) 903 – 907.
511. J. J. Porter, *Am. Dyest. Rep.* **61** (1972) 24 – 27.
512. H.-J. Buschmann, E. Schollmeyer, *Melliand Textilber.* **72** (1991) 543 – 544.
513. H.-J. Buschmann, *Chemiefasern/Textilind.* **42/94** (1992) 408 – 411.
514. D. Tegtmeyer, *Melliand Textilber.* **74** (1993) 148 –151.
515. U. Sewekow, AATCC Conference, Book of Papers, Montreal 1993, pp. 235 – 246.
516. W. Peukert, *Seifen, öle, Fette, Wachse* **114** (1988) 474 – 477.
517. *Text. Horizons* 1987, 58 – 59.
518. K. Roennefahrt, *Melliand Textilber.* **70** (1989) 203 –206.
519. J. Küßner, J. Janitza, S. Koscielski, *TPI Text. Prax. Int.* **47** (1992) 736 – 741.
520. P. Wragg, *J. Soc. Dyers Colour.* **109** (1993) 280 –282.
521. S. N. Gaeta, U. Fedele, *Desalination* **83** (1991) 183 – 194.
522. G. M. Elgal, *Text. Chem. Color.* **18** (1986) 15 – 20.
523. F. Richartz, *TPI Text. Prax. Int.* **46** (1991) 567 –572.
524. W. Hess, *Bild Wiss.* **29** (1992) 84 – 88.
525. E. Spachmann, E. Tillig, *TPI Text. Prax. Int.* **45** (1990) 129 – 130.
526. J. Trauter, *Melliand Textilber.* **74** (1993) 559 – 562.
527. J. J. Porter, *Text. Chem. Color.* **22** (1990) 21 – 25.
528. D. Kuiper, C. A. Brandon, H. G. Spencer, G. L. Gaddys, *TPI Text. Prax. Int.* **40** (1985) 1124 – 1126.
529. A. Ersvell, C. J. Brouckaert, C. A. Buckley, *Desalination* **70** (1988) 157 – 167.
530. G. M. Shaul, T. J. Holdsworth, C. R. Dempsey, K. A. Dostal, *Chemosphere* **22** (1991) 107 – 119.
531. U. Pagga, D. Brown, *Chemosphere* **15** (1986) 479 –491.
532. G. B. Michaels, D. L. Lewis, *Environ. Toxicol. Chem.* **5** (1986) 161 – 166.
533. T. Ogawa, C. Yatome, E. Idaka, H. Kamiya, *J. Soc. Dyers Colour.* **102** (1986) 12 – 14.
534. K. Ackermann, D. Frahne, *Melliand Textilber.* **63** (1982) 66 – 69.
535. C. Harmer, P. Bishop, E. Holder, P. V. Scarpino, *Proc. Ind. Waste Conf.* **46th,** Lewis Publishers, Chelsea, Michigan 1992 pp. 217 – 228.
536. G. M. Shaul, C. R. Dempsey, K. A. Dostal, R. J. Liebermann, *Proc. Ind. Waste Conf.* **41st,** publ. 1987, 603 – 611.
537. H. R. Hitz, W. Huber, R. H. Reed, *J. Soc. Dyers Colour.* **94** (1978) 71 – 76.
538. D. Brown, P. Laboureur, *Chemosphere* **12** (1983) 397 – 404.
539. D. Brown, B. Hamburger, *Chemosphere* **16** (1987) 1539 – 1559.

540. W. C. Tincher: *Dyes in the Environment: Dyeing Wastes in Landfill,* Final Report, Georgia Institute of Technology, Atlanta 1988.
541. D. Brown, H. R. Hitz, L. Schäfer, *Chemosphere* **10** (1981) 245 – 261.
542. R. Anliker: "Organic Colorants" in M. Richardson (ed.): *Toxic Hazard Assessment of Chemicals,* The Royal Society of Chemists, London 1986, pp. 167 – 187.
543. E. A. Clarke, R. Anliker, *Rev. Prog. Color. Relat. Top.* **14** (1984) 84 – 89.
544. D. Brown, *Ecotoxicol. Environ. Saf.* **13** (1987) 139 – 147.
545. R. Anliker, E. A. Clarke, P. Moser, *Chemosphere* **10** (1981) 263 – 274.
546. R. Anliker, *Textilveredlung* **16** (1981) 431 – 438.
547. K. H. Leist, *Ecotoxicol. Environ. Saf.* **6** (1982) 457 – 463.
548. R. A. Moll, *Melliand Textilber.* **72** (1991) 836 –840.
549. *IARC Monographs on the Evaluation on the Carcinogenic Risks to Humans* **48** (1990) 139 –147.
550. U. S. Dep. of Health and Human Serv., National Toxicology Program, Technical Report Series no. 285 (1992) 11 – 13.
551. GefStoffV vom 26.8.1986, in der Fassung vom 26.10.1993, BGBl. I (GefStoffV, August 26, 1986, version October 26, 1993, Federal Law Gazette), 1782 – 1810.
552. R. Anliker, D. Steinle, *Textilveredlung* **25** (1990) 42 – 49.
553. O. P. Hornstein, *Melliand Textilber.* **70** (1989) 222 – 227.
554. K. L. Hatch, *Text. Res. J.* **54** (1984) 644 – 682, 721 – 732.
555. B. M. Hausen, K. H. Schulz, *DMW Dtsch. Med. Wochenschr.* **109** (1984) 1469 – 1475.
556. B. M. Hausen, A. Kleinheinz, H. Mensing, *Allergo J.* **2** (1993) 13 – 16.
557. J. M. Wattie, *J. Soc. Dyers Colour.* **103** (1987) 304 – 307.
558. G. A. Heath, W. Dyson, *Text. Chem. Color.* **22** (1990) 25 – 31.
559. P. Rosenthal, *Int. Text. Bull.* **36** (1990) 58 – 66.

Textile Printing

Robert Koch, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 1 – 4, 6 – 8, Sections 5.1 – 5.9)

Johann Heinrich Nordmeyer, Bayer AG, Leverkusen, Federal Republic of Germany (Section 5.10)

1. History . . . 3
2. Introduction . . . 5
3. Textile Printing Technology . . . 7
3.1. Classification . . . 7
3.2. Preparation of a Master Pattern for the Printing Form . . . 8
3.3. Relief Printing . . . 8
3.4. Gravure Printing . . . 9
3.4.1. Roller Printing . . . 9
3.4.2. Screen Printing (Stencil Printing) . 11
3.4.3. Economic Developments in Gravure Printing . . . 13
3.5. Transfer Printing . . . 14
3.5.1. Wet-Transfer Printing . . . 14
3.5.2. Heat-Transfer Printing . . . 14
3.6. Special Printing Methods . . . 16
3.6.1. Carpet Printing . . . 16
3.6.2. Other Processes . . . 17
3.7. Two-Phase Printing . . . 19
3.8. Drying . . . 19
3.9. Steaming and Other Fixation Processes . . . 20
3.10. Washing Operations . . . 21
4. Thickening Agents and Thickeners in Textile Printing . . . 22
4.1. Introduction . . . 22
4.1.1. Purpose and Effect of a Thickener 22
4.1.2. Demands Imposed upon Thickening Agents and Thickeners . . . 23
4.2. Types of Thickening Agents . . . 24
4.2.1. Emulsions . . . 25
4.2.2. Nonionic Unmodified Polysaccharides . . . 25
4.2.3. Nonionic Modified Polysaccharides 25
4.2.4. Anionic Polysaccharides . . . 26
4.2.5. Cationic Polysaccharides . . . 27
4.2.6. Synthetic Polymeric Thickening Agents . . . 27
5. Printing on Cellulose . . . 28
5.1. Pretreatment . . . 28
5.2. Printing Methods: Classification and Significance . . . 29
5.3. Substantive (Direct) and Acid Dyes . . . 30
5.4. Cationic and Mordant Dyes . . . 31
5.5. Vat Dyes . . . 31
5.5.1. All-In (Sulfoxylate) Printing Processes . . . 32
5.5.2. Two-Phase Printing Processes . . . 34
5.5.3. Discharge Printing Processes . . . 35
5.5.4. Resist Printing Processes . . . 36
5.6. Sulfur Dyes . . . 36
5.7. Leuco Esters of Vat Dyes . . . 37
5.8. Azoic Dyes Produced by Coupling 37
5.9. Phthalocyanine Developing Dyes 38
5.10. Reactive Dyes . . . 39
5.10.1. Introduction . . . 39
5.10.2. Reaction Mechanism . . . 39
5.10.3. Important Reactive Dyes . . . 40
5.10.4. Pretreatment of the Fabric . . . 40
5.10.5. Printing Pastes . . . 40
5.10.6. Fixation (One-Phase Method) . . . 42
5.10.7. Two-Phase Application . . . 42
5.10.8. Washing of the Prints . . . 43
5.10.9. Fastness . . . 43
5.10.10. Special Printing Processes . . . 43
6. Printing on Synthetic Fibers . . . 44
6.1. 2,5-Acetate . . . 44
6.2. Triacetate . . . 45
6.3. Polyamide . . . 47
6.3.1. Pretreatment . . . 47
6.3.2. Printing Processes . . . 48
6.3.3. Dyes and Printing Pastes . . . 48
6.3.4. Fixation . . . 49
6.3.5. Aftertreatment . . . 49
6.3.6. Discharges and Resists . . . 50
6.3.7. Reactive Dyes . . . 50
6.3.8. Polyamide Blends . . . 50
6.4. Polyacrylonitrile . . . 51
6.4.1. Pretreatment . . . 51
6.4.2. Printing Processes . . . 52
6.4.3. Dyes and Printing Pastes . . . 52
6.4.4. Fixation . . . 53
6.4.5. Aftertreatment . . . 53
6.4.6. Discharges and Resists . . . 54
6.4.7. Other Classes of Dyes . . . 54
6.4.8. Blends with Other Fibers . . . 54
6.5. Polyester . . . 54
6.5.1. Pretreatment . . . 54
6.5.2. Printing Processes . . . 55
6.5.3. Dyes and Printing Pastes . . . 56
6.5.4. Fixation . . . 57
6.5.5. Aftertreatment . . . 58

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_489

6.5.6. Discharges and Resists 59
6.5.7. Direct Printing on Polyesters with an Affinity for Basic Dyes 60
6.5.8. Blends of Polyester with Other Fibers 60
6.6. Other Synthetics 61
7. Printing on Protein Fibers 62
7.1. Pretreatment 62
7.2. Direct Printing 63
7.2.1. Printing with Acid Dyes 63
7.2.2. Printing with Metal-Complex Dyes 64
7.2.3. Printing with Reactive Dyes 64
7.2.4. Printing with Chrome Dyes 65
7.3. Discharge Printing 65
7.4. Printing on Blends Containing Protein Fibers 66
8. Pigment Printing 67
9. References 67

This contribution is based on the article Textildruck in Ullmann's, 4th ed., written by Joachim Zahn, Johannes Eibl, Werner Kühnel, Helmut Dahm, Robert Koch, Richard Schwaebel, and Wilhelm Berlenbach.

1. History

Early testaments to printed textiles have survived only because of the dry climate of Egypt. These include a cylindrical printing block 4.4 cm in height with embossed patterns on both flat surfaces; a linen tunic for a child of ca. 3 years, printed by the resist technique with a white diamond pattern on a light-blue ground dyed with indigo (woad); and fragments of directly printed woven textiles, including a two-color printed fabric in red and black. These items, discovered by Forrer in 1894 in the necropolis of Achmim (the Panopolis of antiquity), are Coptic remains from the period between the 5th and 10th centuries A.D. The gap between these particular items and the true beginnings of the localized dyeing of textiles, which lie much farther back in time, is unlikely ever to be closed. Substantial amounts of material have simply rotted away. There are also no known written reports with the exception of a passage in Pliny the Elder (23 or 24 to 79 A.D.), where he describes a process used in Egypt that later came to Europe via India. Printing with solutions of metallic salts (mordanting) was followed by dyeing of the fabric in a hot bath of madder (alizarin). Apart from the vat dyes Tyrian purple and indigo, almost all natural dyes are of the mordant type, and they produce truly dyed textiles only when combined with metallic salts, forming the so-called colored "lakes." Some dyes produce very different colors with different metals.

In the Middle Ages, most of the textile printing in Europe occurred in monasteries. The few records surviving from this epoch also specify the dyes used. Almost all were the same inorganic pigments employed by painters; e.g., black: soot from carbonized vines; white: white lead (basic lead carbonate); yellow to brown: various types of ocher; green: verdigris (mixtures of basic copper(II) acetates); blue: azurite (basic copper carbonate) and powdered indigo; red: red ocher, red lead (lead oxide), and cinnabar (mercury sulfide). All of these were bonded with linseed oil in the printing process. Printing with such oil colors was practiced for many centuries, and it is rightly regarded as the precursor of today's *pigment-printing* process.

After the 15th century, textile printing also developed outside the walls of the monasteries. It gradually became an independent trade, and reached industrial scale in certain countries as early as the 18th century. A powerful impulse came from India, with its rich textile tradition. This increased interest in printing brought remarkable achievements, and not only in the use of color. The first process that did not use oil colors was *indigo blue printing*, established in Holland in the 16th century. In principle, this was the same technique that had been used for the child's tunic in Achmim/Panopolis: printing a pattern with a paste (a liquid wax resist in the case of the Copts) by means of a printing block, dyeing the result in a cold indigo vat, and washing out the printed resist to reveal a pattern that had stayed white during dyeing. In the 17th century, Georg Neuhofer brought indigo blue printing from Holland to Augsburg, where the first German calico printing works was built by his brother, Jeremias Neuhofer, in 1689.

In 1759, Johann Heinrich Schüle began a manufacturing operation that at its peak em-

ployed 3500 workers. In addition to indigo blue printing, *direct printing* and hand coloring with supplementary dyes such as alizarin soon were of major importance as well. White and color *discharges* also took their place in company with the conventional printing process. Printing was even initiated on the basis of engraved copper plates instead of wooden patterns.

CHRISTOPH PHILIPP OBERKAMPF, who was born in Weissenburg near Ansbach in 1738, moved in 1758 to Paris and soon thereafter (1760) founded the most important textile printing works of the time at Jouy, near Versailles. "Toiles de Jouy" gained a worldwide reputation, and became the absolute embodiment of artistic textile printing. Indeed, LOUIS XVI elevated OBERKAMPF to the nobility and designated his operation a "Manufacture Royale."

The technique of *chlorine bleaching* was practiced at an early date in Jouy, and wooden patterns were replaced by firmly bonded copper plates of arbitrary size, permitting the printing of large designs. Printing with copper plates required use of a *press* for the first time, because the printing paste was to be transferred to the fabric not from a raised area but rather from the recesses in the plate. A simple printing machine with a roller bearing a raised pattern was also part of the technical arsenal at Jouy.

First reports of a continuous *rotary press* appeared at the turn of the 18th century. This wooden device was operated by a hand crank and included two rollers, one above the other. The top roller bore the raised design, to which colored printing paste had to be applied manually. The lower roller pressed the moving fabric against the upper roller.

ALOIS SENEFELDER (1771 – 1834), the inventor of lithography, failed to achieve significant success with this particular process, so he turned toward calico printing, but his invention of an improved rotary printing machine came too late. By 1783, the Scotsman THOMAS BELL, a printer of copper engravings, had already patented a machine in which engraved copper rollers continuously printed a multicolored pattern onto a fabric. He was thus the true inventor of *cylinder printing*. Present-day roller printing machines developed on the basis of his design. By contrast, rotary printing with a roller containing a raised pattern was not destined for a great future.

In 1834, the French machinist PERROT constructed a flatbed printing machine that quickly came into widespread use, and was in fact called the *perrotine* after its inventor. A rather complicated mechanism coated the printing surface with paste and pressed it against the fabric, subsequently causing the fabric to advance to the next printing position (i.e., by one pattern-repeat unit). The printing blocks extended across the entire width of the fabric. The significant advantages of the perrotine over hand printing included great accuracy, production of a more uniform print thanks to the mechanical transport feature, multicolor printing during a single pass of the fabric, and higher production rates.

An ancient, artistically very ambitious, and still much used Japanese printing technique, the Yuzen process, is named for its inventor, the painter YUZENSAI MIYASAKI (1654 – 1736). It was this that later developed into the *screen printing* process. A Yuzen screen stencil consisted of several layers of paper, glued together, into which a pattern was cut. The protective paper regions were joined together with hairs (later: silk threads) to keep the edges of the stencil from bending up. Eventually the threads were replaced by silk gauze, then by woven screens made of chemical and metallic fibers. The function of the paper was ultimately assumed by a coating of paint. Starting in 1926, the screen-printing process also became established industrially in Europe. Further development led to the modern continuous rotary-screen printing process based on hollow, cylindrical patterns made of thin nickel foil, which has become a more economical way of producing prints than roller printing.

Since the mid-1960s the *transfer process* has become increasingly important. Here, the pattern is first printed onto an intermediate carrier (usually paper), after which it is transferred to the fabric.

Total world production of printed textiles is now estimated to be ca. 24×10^9 m^2/year.

2. Introduction

Methods of Textile Printing. Ornamental patterns consisting of one or several colors can be reproduced on fabrics by the local application of sharply defined areas of dye preparations

(printing pastes) either to undyed or previously dyed cloth.

If in the fixation process that follows application of such a printing paste there is destruction of a dye applied previously, the procedure is known as *discharge printing*. Use of a discharge printing paste that also contains its own discharge-resistant and fixable dye is known as *color discharge printing*.

Resist printing is actually the oldest of the printing processes. In this case a special printing paste is applied to certain areas of the fabric to prevent dye fixation. In the case of a physical resist, penetration of the dye liquor is inhibited by a wax-induced hydrophobicity, or by application of a difficultly wettable thickener with a high solids content. With a chemical resist, dye fixation is prevented by a chemical reaction. Resists are most easily applied in a *preprinting* step; i.e., the resist paste is first printed and dried, after which the entire fabric is exposed to dye in a continuous dyeing device, for example by dipping the fabric in the dye liquor, squeezing it between rollers, and drying it in a continuous dryer (the Foulard or *padding* process). Overprinting resists can also be applied, but only if dye already present in the previously dyed and dried fabric is still in its unfixed state, as in the case of oxidation dyes (aniline black goods) and other developing dyes. A resist appropriately described as an "intermediate printing resist" can be prepared with naphthol dyes that are subject to two-stage application. Here, a fabric previously treated with a naphtholate is printed with a resist that prevents the normal coupling with a stabilized diazonium salt to form an azo dye.

Color resists contain a dye that is compatible with the resist agent. *Discharge resists* constitute a special case. The difference between this process and normal discharge printing is that here a dyed fabric is printed in a dry but still unfixed state. Synthetic fibers are especially well-suited to this process, since after padding and drying the deposited dye resides mainly on the surface of the fibers, leaving it much more readily accessible to the discharge chemicals than if fixation has occurred in the interior of the fibers. Preprinted resists are often based on the combination of a physical resist with a chemical resist.

A more recent approach is the "wet-in-wet" printing of a resist, leading to what are known as "African prints." In practice, the resist is overprinted with a surface- covering dye in a continuous "wet-in-wet" process. The contours produced by this type of resist are not as sharp as those obtained either with dried, preprinted resists or with discharge printing.

Direct printing, resist printing, and discharge printing differ with respect to the mechanism of action of the various chemicals, as well as in the dyes used, but all can be conducted by any of the standard printing methods (printing blocks, rollers, etc.), and either by hand or mechanically.

Dye Application and Fixation. There are significant differences between absorption of a printed dye and absorption of a dye applied from a dye bath. In the latter case, dye is often present in the dye bath at low concentration, and the process is allowed to proceed to the point of true equilibrium between dissolved and absorbed dye. The position of this equilibrium depends on the substrate (fiber), the dye, the temperature and pressure, and sometimes also the presence of additives (e.g., salts). However, the end of the process is associated with a well-defined final state.

Dye availability from a printing paste is characterized by high dye concentration in the medium and a printing thickener that inhibits dye diffusion. The extent of dye transfer is determined in part by the printing method used (e.g., the depth of engraving). The printing thickener is responsible for limiting capillary flow between the threads of the fabric, as a result of which dye penetration is much reduced into cellulosic fibers, which are subject to swelling. Except in the case of carpet printing, smearing of the dye as a result of transporting the print over guide rollers during drying is a potential problem. Evaporation of the water leads to an increase in the dye concentration and ultimate restoration of the dye to a solid state. As a result, soluble dyes diffuse to some extent into fibers that exhibit swelling properties under the influence of a concentration gradient.

The actual fixation step is responsible for transporting into the fibers as much as possible of the dye that has been retained in the dried thickener. This is especially important with dyes that do not dissolve in the printing paste but are converted into a soluble form only under the influence of chemicals introduced during the fixation process. In the great majority of cases, fixa-

tion is completed by treatment with water vapor (steaming), in the course of which steam condenses on the printed material, flows into the thickener, heats the fabric, and provides a dye-diffusion transport medium in the form of the water that swells the thickener.

The distribution of dye between fiber and thickener is determined in part by the "retaining power" of the thickener. The thickener is often composed of polysaccharides, and thus competes with cellulose in the fibers. For this reason the extent to which a thickener is subject to removal by washing is especially important. This depends not only on the swelling properties and cold solubilities of the various modified starches, celluloses, or mucilages present, but also on the nature of the drying and fixation processes employed. Direct fixation with hot air or high-temperature steam may make the removal of a thickener very difficult.

The color intensity of a print depends very much on the thickener used. The dye-distribution ratio between fiber and thickener is important, but so is the possibility of chemical reactions between the thickener and the dye (reactive dyes, metal- containing complex dyes, developing dyes of the Phtalogen type) and between the thickener and various auxiliaries.

The behavior of thickeners toward dyes and printing auxiliaries depends in turn on the textile substrate: synthetic fibers often act quite differently from cellulosic fibers in this respect.

Printing Auxiliaries. As a general rule, the auxiliaries used for printing are the same as those used in dyeing with a dye bath (→ Textile Auxiliaries). They can be grouped into various categories as follows:

Humectants; e.g., urea, glycerine, glycols
Solvents; alcohols, ethers, esters, polyglycols
Fiber swelling agents; cyanoethylformamide
Wetting agents; nonionic, cationic, anionic
Reducing agents; e.g., sodium oxymethanesulfinate (Rongalit C, BASF), sodium dithionite (Hydrosulfit, BASF), formamidinesulfinic acid, tin(II) chloride ($SnCl_2$), boranates, glucose
Oxidizing agents; e.g., hydrogen peroxide, sodium *m*-nitrobenzenesulfonate (Ludigol, BASF)
Carriers; e.g., cresotinic acid methyl ester (Levegal PN, Bayer), trichlorobenzene, *n*-butylphthalimide in combination with other phthalimides (Levegal PEW, Bayer), methylnaphthalene (Remol TRM, Hoechst)
Retarders; e.g., derivatives of quaternary amines (with basic dyes), but also levelling agents
Resist agents; e.g., zinc oxide, alkalis, amines, complexing agents (Trilon A and Trilon B, BASF)
Metal complexes; copper or nickel salts of sarcosine or hydroxyethylsarcosine
Acids; acetic acid, tartaric acid, citric acid, lactic acid, glycolic acid, formic acid

Thickening agents, discussed in detail in Chapter 4, constitute a special group of printing auxiliaries. There also exist various special auxiliaries, which frequently consist of mixtures; examples include emulsifiers, amides, glycol ethers, and aminoalcohols (Levasol, Bayer). The use of certain once-common auxiliaries (e.g., petroleum ether or naphtha for emulsion printing) is severely restricted as a consequence of environmental legislation.

Dyes. Regulations related to workplace safety have led to the almost complete disappearance of certain dyes and processes, including aniline black and other black oxidation dyes based on aromatic amines. As a result, dyes applicable to cellulose fibers are now limited to a few main groups. The commercially most important categories are the *vat dyes* (for color printing and color discharge printing), *reactive dyes*, and *pigments* together with their associated binder systems.

More detailed information on dyes is provided in Chapters 5, 6, and 7.

Pretreatment. The appropriate pretreatment of textiles prior to printing is crucial to the success of the process. This includes singeing, washing, bleaching, etc. (→ Textile Auxiliaries ; see also Chapters 5, 6, and 7 of the present article). Pretreatment is usually carried out on the full-width fabric, usually with continuously operating equipment ("continuous lines"). Cellulosic fibers in woven form should be bleached and sometimes also mercerized before printing to increase the brilliance and intensity of the colors. Slivers, yarns, and knitted goods also require

pretreatment prior to printing, typically in the form of cleaning and processing in such a way as to increase wettability. It is essential that spinning preparations and sizes be removed if one hopes to achieve an adequate degree of whiteness. Removal of fiber components is sometimes necessary as well, such as lignin in the case of bast, stalk, and leaf fibers, and "silk gum" (sericin) in the case of silk. The dye-absorption properties of wool are considerably improved by chlorination (see Sec. 7.1).

Particularly with synthetic fibers, but also cellulose fibers, it is necessary that a fabric's shape be stabilized, for example by preshrinking, removal of tension introduced during weaving, stabilization of the weave structure, and straightening of warps and wefts in the "thread direction." Tentering frames of various designs are used for this purpose.

Pretreatment also includes the application of certain preparations designed to improve dye absorption (e.g., urea and wetting agents), or weak oxidizing agents may be applied to a "ground color" (dyed fabric) to prevent damage to the dye by "tarnishing" (carryover of a discharge printing paste to the unprinted surface), especially with light-colored fabrics. Such preparations are applied by padding or "slop padding" (one-sided application with a roller that rotates half-immersed in the pretreatment solution while at the same time wiping against the fabric), after which the treated fabric is dried.

The only case in which an unprepared fabric is printed is when a nap resist is to be applied. This type of resist prevents the nap from being raised by a napping machine, producing a unique type of pattern.

3. Textile Printing Technology

3.1. Classification

The following list summarizes the various techniques available for printing on textiles:

1) *Manual Methods* (∗ indicates a historical method)
   Raised printing forms (relief printing): hand or block printing∗, yarn printing∗, warp printing∗
   Recessed printing forms (gravure printing): copperplate printing (of calico)∗
   Stencils: spray printing, screen printing
   Special processes like those often associated with tie dyeing, wax batik, or the Golgas process∗ (mechanical resists produced by squeezing between plates)

2) *Automatic and Semiautomatic Printing Processes*
   Raised printing forms: perrotines∗, relief roller printing∗, hank yarn printing, vigoureux (melange) printing (especially for carpets), Stalwart – Pickering, space-dyeing system (yarn and warp printing), Stalwart system (warp printing)
   Recessed printing forms: roller printing, space dyeing (knit – deknit process)
   Stencils: flat-screen printing, rotary-screen printing
   Carpet printing: BDA process (flat-screen printing with vacuum), Aljaba carpet printing∗
   Special printing processes: carpet jet-spray printing, deep-dye process (special printing equipment), transfer printing, STAR printing,
   Orbis printing

The emphasis in the discussion that follows will be on automatic and semiautomatic methods.

3.2. Preparation of a Master Pattern for the Printing Form

A master pattern is produced originally as a drawing, which should be as accurate as possible consistent with the capabilities of the proposed printing process. Single- color or multicolor patterns must be incorporated several times into the actual printing form (roller or screen) to permit single-pass printing onto a surface. The repeating portion of the pattern is known as the *pattern-repeat*. The length of the pattern-repeat in the warp direction defines the pattern-repeat height, that in the weft direction the pattern-repeat width. The pattern repeat may involve no offset (1/1) or it may be offset longitudinally to

the extent of 1/4, 1/3, or 1/2. Each color incorporated into a pattern drawing requires a separate printing form. (Normally, the pattern designer is careful to ensure that the number of colors required for color separation – and therefore the number of printing forms – is kept to a minimum.) Proper placement of the colors as each is printed in sequence by the printing forms is referred to as the *registration* [17].

Several methods exist for transferring a pattern drawing to a printing form. The basic procedure is the same as that used to make printing forms for book production (→ Imaging Technology), where an intermediate step involves reproducing the color-separated design on a transparent film or photographic substrate (→ Photography) of a size appropriate to the printing form. Transfer to the printing form itself occurs by a copying process adapted to the nature of the form (roller or screen). This is described in conjunction with printing processes themselves in Section 3.3.

Color separations can be prepared manually, by a photographic method based on color filters, or with electronically controlled color-separation equipment (textile engraving machines such as those produced by Hell in Kiel, Germany, or by Scitex in Israel, which operate on a scanning principle). Scanners of this type, utilized in conjunction with a computer and a data base, open the way to color signal processing, addition, pattern reflection and offset, color reversal, scale modification, contour tracing, electronic rastering, etc. This makes it possible to prepare a color separation with a repeating pattern of exactly the desired size. Transfer to a printing form is usually accomplished with the aid of photographic film, although a process involving direct control of an engraving machine (e.g., of the type manufactured by Hell) has also become well established. The resulting engraving is very shallow (0.03 – 0.05 mm), but it can be used for printing plastic films (e.g., woodgrain patterns for furniture) as well as transfer papers. Digitalized patterns have also been transferred by means of a laser beam to undercoated templates. In the process, the laser beam burns its way through the undercoat to expose selected portions of the screen constituting the basis of the nickel template.

3.3. Relief Printing

Block Printing. The pattern form in this case consists of several layers of wood glued together to form blocks ca. 7 – 8 cm thick such that the grain in the various layers runs in alternating directions. Areas where printing should not occur are carved out on the basis of a traced drawing. Strips cut from a sheet of brass are used to reinforce the contours, and any required dots are fashioned from brass pegs. Large printing areas may be covered with a felt made from wool fleece to ensure greater uniformity of dye application.

Manual printing is carried out on long felt-covered tables over which rubber sheets have been stretched. For difficult (multicolored) patterns, the fabric is held in place by an adhesive. The printing block is wetted with dye from a trough (chassis). This trough is a box filled with thickener (false color) in which floats a wooden frame with a rubber sheet stretched over it. The frame in turn contains a felt cloth saturated with printing dye. This cloth constitutes the "stamping cushion" from which dye is transferred to the raised portions of a printing block placed upon it. Printing is accomplished by laying the prepared block on the fabric and striking it with a wooden mallet (at the corners as well as in the center). This manual process still finds use today, although mainly in the fine arts field.

Attempted mechanization of this particular printing process led to development of the *perrotine*, named after its inventor Perrot. A perrotine operates in a repetitive mode: a printing block that has already been applied over the entire width of the fabric is lifted mechanically, fresh dye is applied to the block from a roller covered with felt and impregnated with dye, the fabric is moved by the equivalent of one pattern repeat and then secured again to the padded area, and the printing block is lowered once more. Machines of this type have been designed for printing up to 6 colors, although they are rarely encountered today.

Relief Printing. The once widely distributed relief printing machines have suffered a similar fate, although they were still used in carpet printing until a few years ago, and modified versions play a role in this application even today (see Sec. 3.6.1).

3.4. Gravure Printing

A feature shared by all recessed-form (gravure) printing processes is application of dye to the entire surface of the printing roller. A doctor blade is used to remove excess dye from the surface, leaving it only in the depressions (engraved areas). The textile subject to printing is supported on a somewhat deformable surface and the roller is then applied to the fabric at relatively high pressure, causing the cloth surface to be pressed into the engraved areas where it picks up dye.

3.4.1. Roller Printing

Roller printing is a process (Fig. 1) that is declining steadily in importance.

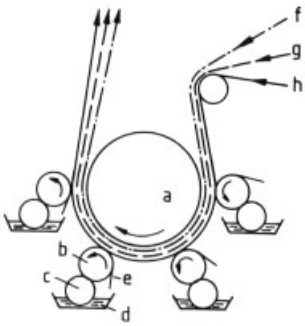

Figure 1. Four-color roller printing machine
a) Impression cylinder; b) Printing roller; c) Dye roller; d) Dye box; e) Doctor blades; f) Undercloth; g) Back gray; h) Textile

Whereas most manual printing techniques originated in Asia and the Far East, roller printing is a European invention (BELL, 1785, in Lancashire). With only minor improvements, such printing machines continued to operate until the 1950s with virtually no new design features. Subsequent changes have included automated positioning of the pattern repeat, separate control of the doctor blades independent of the rate of fabric transport, somewhat later the introduction of split bearings for the printing rollers, and, finally, the use of hydropneumatic pressure in place of weights or springs. Use of the undercloth (an endless blanket) for driving the drying system (*mansard*) was abandoned in favor of an independent drive mechanism for the dryers and folders. Furthermore, washing systems were introduced for the undercloth and the backing cloth ("back gray"), simplifying the separation especially of those textiles printed by the pigment process from the undercloth and the back gray.

The older type of machinery, used mainly for printing staple-cotton fabrics with a maximum width of 90 cm, has been almost completely displaced by mechanized screen-printing systems (see Sec. 3.4.2), although modern (Saueressig) roller-printing machines have been installed for printing knitted fabrics ca. 160 cm in width. The classical design with a central press has in this case been abandoned in favor of the so-called portal printing machine based on a frame construction in which each printing roller is directly opposed by its own impression cylinder. The frame (portal) consists of two segments joined along one side by hinges. The front part can be opened like a door, facilitating dye changes and minimizing expensive down time.

Printing Rollers. Hollow copper cylinders constitute the heart of the roller-printing process. Such a cylinder may have a weight of 40 – 70 kg depending on its size. A slightly conical spindle is pressed into the hollow center of the roller, which has a matching conical shape. Also mounted on the spindle are a pattern repeat wheel, a drive wheel for the dye-application roller, and the necessary bearings.

In newer machines, the continuous spindle has been replaced by a pair of cones pressed pneumatically into the hollow cylinder from opposite ends. These shortened half-axles, each with a cone that can be disengaged pneumatically, remain permanently set in the bearings. Continuous spindles had the disadvantage of being very heavy (80 kg or more), which made manipulation of the printing roller and installation and removal of the spindle both difficult and time-consuming.

The crude roller barrel, already corresponding approximately to the appropriate dimensions, is ground and polished prior to engraving to produce the smoothest possible surface. This is an essential prerequisite for gravure printing, because the doctor blade must be capable of wiping the roller completely free of excess dye. Small defects, such as scratches, inevitably lead to marks on the finished textile.

Engraving. The printing roller can be engraved with various types of cutting instruments (graving chisels, hooks), either by hand or mechanically. Embossing or etching processes are applicable as well.

The latest development is the optoelectronically or magnetoelectronically controlled "helioklischograph" engraving machine developed by Hell, (→ Imaging Technology, Chap. 3.4.3.), although the very shallow engraving that results with this technique is so far used mainly for printing "transfer papers," and only rarely for textile printing.

Photogravure. Color intensities captured as gray levels on a transparent film can be transferred to a copper roller coated with a photosensitive resist (formerly chrome gelatine) as one way of reproducing color shades, but the only technically feasible approach involves the use of a half-tone *screen* (→ Imaging Technology). To produce such an autotype, the color-separation film (negative) of interest is placed directly over an appropriate screen film, and both films are then wrapped together around the coated roller. Light scattered at the edges of the dots constituting the half-tone screen causes these dots to be reproduced on the roller, but with varying sizes. Relatively dark areas of the negative produce dots of a size equal to or even somewhat larger than those of the screen, as there is only a small amount of scattering, while a bright region leads to small dots on the photosensitive-resist coating. After removal by washing of all photosensitive resist that has not been fixed by exposure to light, an engraving is prepared by an etching process, which leads ultimately to the desired autotype.

When an engraved roller is to be prepared for gravure printing, a screen (e.g., a cross or hexagonal screen) is first placed over the coated roller and this is exposed, after which the color separation is set in place. Hardening of the photosensitive resist within the screen apertures takes place to an extent that depends on the local "color intensity" of the separation. A selective etchant is then used, one which penetrates only slowly or not at all into those parts of the coating that have been more strongly illuminated. "Development" with iron chloride leads to an engraving in which the various screen dots, all of which are the same size, have been etched to different depths.

This type of engraving produces what are known as "true halftones." Obviously such a photogravure process could also be used to create printing surfaces with a constant color depth over the full width of the fabric, but other methods are normally employed when the screen pattern itself is the only thing to be copied (e.g., the Szigeti process, or use of an Agfa contour film) [18, 19].

An engraving suitable for textile printing is always deeper and coarser than one intended for paper printing. When printing onto a textile, interlacing points formed by the threads themselves constitute a second raster pattern, so that higher pressures and larger dye volumes are required to achieve a printed result similar to that obtained with paper. In fact, the two printing techniques are complementary, which is also the case, for example, with the transfer-printing process.

Aftertreatment and Regeneration of the Rollers. For all practical applications today the engraved copper rollers are plated with chromium, because the copper of the rollers is much too susceptible to damage (e.g., by the doctor blades). A thin chromium layer (maximum depth 0.04 mm) is applied by electroplating after polishing and degreasing of the roller, and the cleaned rollers are then polished a second time. This hard chromium coating considerably prolongs the useful life of a roller.

It is self-evident that roller printing involves considerable cost, not only in the form of the printing machine, but also through preparation of the printing rollers. These costs can only be justified if the output will be large and color and design changes are limited. For this reason, mechanized screen printing and rotary-screen printing have become the established methods for short production runs. Indeed, rotary-screen printing has become popular for long runs as well, in part because of the width of the textiles currently being printed.

3.4.2. Screen Printing (Stencil Printing)

As noted previously (Chap. 1), the first stencils were made in Japan of reinforced paper. These devices permitted dye to be applied to the textile underneath through openings that corresponded

to the proposed pattern. At a later date a woven screen was used instead, with those parts of the screen where no printing was to occur being simply covered over. Metallic and plastic stencils were also developed for dye application with a spray pistol, through which appropriate patterns were cut, machined, or, in some cases, etched. Thus, a sheet of zinc might be coated with paint wherever printing was to be prevented, permitting the remainder to be etched away. Shading effects were achieved manually in the course of spraying by the use of low-viscosity printing inks. Such techniques are seldom practiced today except occasionally for disguising the joins between large repeating patterns.

A number of different screen-printing methods of ever-increasing importance have evolved from these relatively primitive approaches. The most widely practiced today include:

1) **Flat-screen printing on tables**, where the screens are laid along rails to facilitate alignment of the repeat patterns. The tables are sometimes subject to heating, although in other respects the process is similar to that described in Section 3.3. The dye is applied by hand.
2) **Screen printing with a wheeled stencil carriage** that can be moved on rails from one repeat position to the next, where it stops and is lowered so dye can be applied by hand with the aid of a weighted double squeegee.
3) **Screen printing with a fully automatic screen carriage** that moves the screen and stops it at the pattern-repeat point, where the squeegee is activated by an electric motor. Rails that support the equipment on both sides also carry hand-operated or electrically driven carriages for applying adhesive to the textile and anchoring it in place, as well as a washing carriage to wash off and wipe away excess glue. Some hand-printing establishments still dry the fabric passively by hanging it over the tables. Mechanical drying devices are described in Section 3.8.
4) **Mechanized screen printing with stationary screens mounted in a frame.** A screen frame performs the function of lifting the screens while the textile, which is glued to a moving endless rubber belt, is advanced to the pattern-repeat point, after which the screens can again be lowered. The screens are wiped with a squeegee while stationary. Squeegees of many different types are used, including wipers, rollers, and magnetic rollers. With a magnetic roller system a magnetic bar located beneath the undercloth is used to move the roller over the screen. The roller pressure and the amount of dye applied are functions of the thickness of the "magnetic squeegee" (a cylindrical iron bar) and the magnitude of the magnetic force, which can be adjusted stepwise. Double rollers and box or slit squeegees are used for printing carpets. These are especially well-suited to printing many colors and pressing the dye deep into the pile.

Printing machines for this purpose consist basically of a facility for moving the textile, a gluing device that causes liquid glue to be applied by a squeegee to the undercloth, and the printing equipment itself. Undercloths precoated with thermoplastic glues are sometimes used instead, in which case the heated textile is squeezed by a roller when it first meets the rubber- coated transport belt, causing glue in the immediate vicinity of the textile to soften and instantly adhere. With permanent adhesives the textile can be fixed simply by placing it in position and pressing it down. Printing occurs with all the stencils on the conveyor belt at once, although the printed design is complete only after it has passed the last screen. The rubber- coated undercloth is then directed downward over a guide roller, washed with water and rotating brushes, wiped to remove excess rinse water, and sent back to the gluing device.

The textile, which has been pulled away from the undercloth, is fed directly to a drying chamber. One type of printing machine manufactured by the Buser company incorporates an ingenious design that permits transport through the dryer to be continuous even though the printing process itself remains discontinuous. This prevents potential damage in the drying chambers due to heat that might accumulate in stationary carrier rollers. Alternatively, drying in a chamber must be preceded by hot-air jet drying.

Automatic screen printing involves the least set-up time, the lowest screen costs, and the smallest loss of printing dye, which in turn minimizes contamination of the waste water that results from washing of the screens.

Rotary-Screen Printing. This is the most efficient textile-printing process available. Its immediate precursor was a Portuguese printing machine (Aljaba) that for the first time incorporated cylindrical screens. Dye was applied from the interior of the screens by a dye tube and a squeegee that pressed the dye through the stencil openings onto the textile.

The true breakthrough was achieved with development of a rotary-screen printing machine in which the textile was fixed by adhesive to an endless rubber transport belt, permitting cloth to pass continuously under a set of rotary screens arranged in relatively close proximity to one another. The devices used for feeding the fabric, applying adhesive, and removing and drying the printed goods correspond essentially to those used with automatic flat-screen printing, but less space is required for a rotary-screen printing machine (Fig. 2).

The production rate achieved in rotary-screen printing equals or even exceeds that in roller printing. The short distance between the screens and the very brief dwell time experienced by the textile under the screens eliminates adhesion problems even with fabric made of synthetic fibers, such as polyamide, whereas flat-screen printing can be complicated by partial detachment of the fabric from the undercloth, leading to bubble formation.

The availability of automatic equipment for both flat-screen and rotary-screen printing makes it possible to use one or the other technique selectively depending on the circumstances. Thus, rotary-screen devices permit trouble-free printing of longitudinal stripes in the warp direction, for example, whereas with flat-screen printing all the places where it was necessary to lift the frames tend to remain visible. On the other hand, flat-screen printing machines are in many respects more versatile (e.g., squeegee operation can be repeated as often as necessary for application of more dye and increased dye penetration).

Stencil Preparation. Flat screens consist of frames that have been covered with a gauze. The frame material may be wood, but today it is more likely to be welded metal tube. Adhesive is used to attach to the frame a stretched prestressed gauze made of monofilament or multifilament artificial fiber.

While the desired pattern might be created by manual application of paint to those areas of the gauze where printing is to be prevented, the process is today accomplished almost exclusively by first coating the entire screen with a light-hardening photosensitive resist. Color separations (film diapositives) are then placed over the coated screens and illuminated; in the case of small repeating patterns a "copy-adding machine" may be used to expose the same screen several times as an alternative way of reproducing a complete design. Because of the high degree of mechanical stress to which it will be subjected, a hardened photo-resist is strengthened by relacquering and fixation after washing out of any unhardened screen coating and drying. Rotary screens for "printing cylinders" can be prepared by any of the following methods:

1) Fabrication of a thin-walled nickel cylinder and subsequent etching of the pattern.
2) Preparation of a cylindrical screen, perforated over its entire surface, which can be processed by the photo-resist method, just like a flat screen; alternatively, a screen might be coated over its entire surface and a pattern then produced by partially burning out the paint with a laser (under computer control based on digitalized design data).
3) The direct electrolytic method. Here the color separation, together with a half-tone screen, is illuminated on a cylinder that has been coated with photoresist. After dissolution of the unfixed photoresist and drying, the cylinder is connected up as the cathode in an electrolytic bath. A nickel coating is then applied, varying in thickness from a few hundredths to a few tenths of a millimeter depending on the size of the screen. This results in a thin-walled cylinder bearing the pattern. Unlike coated screens, rotary screens of this type cannot later be given a new pattern. On the other hand, there is also a limit to the number of times a fairly resistant paint coating can be removed from a conventional screen without damaging it.

It should be noted that the electrolytic method is capable of producing smaller apertures separated by less metal (e.g., 78 – 100 apertures per cm), opening the way to subtle color gradations and reduced dye consumption (for transfer printing onto paper, for example).

Figure 2. Rotary screen-printing machine (Johannes Zimmer Co., Kufstein, Austria)

3.4.3. Economic Developments in Gravure Printing

Penetration figures for the textile-printing machinery market in Germany and their evolution with time are presented in Table 1 based on data developed by the Research Center for the General and Textile Market Economy at the University of Münster. Current estimated mean capacities (1000 m^2/a) for these facilities excluding carpet printing are as follows:

Roller printing	1233
Rotary-screen printing	3031
Flat-screen printing	620

Table 1. Number of textile printing machines in the Federal Republic of Germany

	Roller printing	Rotary-screen printing	Flat-screen printing (mechanized)	Flat-screen printing (manual) ∗
1960	320			
1968	202	22	112	13 000
1972	135	73	114	7 500
1980	81	117	94	3 265
1988	29	124	59	

∗ Data in meters of overall table length

3.5. Transfer Printing [20,21]

In transfer printing, the pattern is first created on an intermediate carrier and then transferred from there to the textile. The dye may be fixed either in the course of the transfer step or in a subsequent fixation and finishing process.

Transfer processes were of relatively little importance until the mid-1960s, but this situation changed rapidly with development of dry-transfer printing processes, especially the Sublistatic process (heat transfer on polyester; see Sec. 3.5.2), and transfer printing is today involved in the production of ca. 700–800 $\times 10^6$ m^2 of textile per year.

3.5.1. Wet-Transfer Printing

The oldest of the transfer-printing processes is the *Sark process* [22], also known as the "Orbis" process. A pattern is first generated on a metal roller using a printing-dye mix in which the dyes and auxiliaries are present as a soapy, plastic mass. The pattern is then transferred by rolling it onto a textile previously prepared with a solvent/water mixture. The transfer step is followed by steaming and washing. Only a limited length of fabric can be printed with a single roller.

This process is capable of reproducing patterns with a unique range of colors and elegance, but interest in it has been limited to high-fashion applications. Although the process can in principle be used on all types of fibers, it has so far seen service only for silk and artificial silk.

In the *Fastran* and *Dew-Print* processes [23], and more recently in the Danish "*Cotton-Art*" process [24], paper is used as an intermediate carrier onto which textile dyes have been printed by paper-printing methods. Moist fabric that has been impregnated with auxiliaries is then pressed onto the paper and heated to 100 – 105 °C. The resulting steam atmosphere effects transfer of the dye. This is followed by a washing process that may be less intensive than would be required in conventional textile printing because a large part of the unfixed dye and thickener remains on the paper. This technique is again applicable to all types of fibers, including wool and cotton. Since the patterns are printed by paper-printing machines, they are usually limited to four colors.

Neither of the wet-transfer processes described has yet become significant from the standpoint of production volume.

3.5.2. Heat-Transfer Printing

STAR Printing [25]. The *STAR* process was developed in 1950 in Italy. Paper is again printed with a solvent and a paste containing the dyes, but in contrast to the Fastran and Dew-print processes, special thermoplastic binders on specially prepared papers are used to create what amounts to a type of transfer picture.

Dye is transferred to the textile by squeezing the fabric and the paper together between rollers heated to 120 – 150 °C, causing the film to soften and the entire pattern to be transferred to the fabric. The dyes are then fixed by steaming. The thermoplastic film has of course been transferred as well, so it must be removed by chemical cleaning as part of the finishing process. The STAR method is applicable to all classes of dyes with the exception of developing, oxidation, and pigment dyes. It is so far practiced only by a single company.

Thermachrome Process. The *Thermachrome* process is a variation of the STAR process, but one based on pigment printing. The thermoplastic binder in this case is so formulated that it acts as a satisfactory binder on the textile as well. No cleaning step is required subsequent to transfer. The Thermachrome process is of some importance in label printing, and for printing emblems containing photographically precise patterns on finished items of clothing (e.g., T-shirts).

Sublistatic Process [26, 27]. Of all the transfer-printing processes for textiles, the *Sublistatic* process is the only one that is important from a quantitative standpoint, accounting for well over 90 % of the $700 - 800 \times 10^6$ m^2/year of transfer-printed textiles. As with the STAR process, the requisite papers are prepared mainly by four-color printing with paper-printing machines, but thermoplastic binders are avoided. The only coloring materials used are disperse dyes that sublime unchanged in the range 180 – 230 °C.

The paper and fabric are brought into contact for 60 – 20 s at 200 – 220 °C with the aid of presses or rollers, causing only the dye to be transferred from paper to textile. No aftertreatment is required.

The process is restricted almost exclusively to polyesters and polyester blends, because the fibers must be capable of effectively absorbing a disperse dye, and they must not be overly thermoplastic in the range 200 – 220 °C.

Transfer Papers. The stock used in preparing the transfer papers should have a weight of 60 – 80 g/m^2, and it must be coated on the printing side with a thin layer of starch (ca. 3 g/m^2). Paper that is to be printed on paper-printing presses that accept pastes based on alcohol or solvents need not be as absorbent as the paper used with aqueous pastes on roller-printing or rotary screen-printing equipment. In the case of aqueous pastes, the paper should accept 60 – 80 g of water per square meter in the course of 60 s, and it must display "good transparency" when held up to the light. Cloudy papers cause uneven printing.

The papers are subject to processing by all the various paper-printing techniques (→ Imaging Technology).

Gravure printing is well-suited to the production of very fine half-tones, shadings, and color sequences. Printing is from roll to roll.

Flexographic printing, another gravure process, is appropriate for less demanding designs characterized by large, uniform surfaces. Here again, printing is from roll to roll.

Offset printing is capable of very fine half-tone reproduction, but only from one sheet to another.

Screen printing is of particular interest because of the low cost of preparing the corresponding stencils. Printing can be either from roll to roll or from sheet to sheet.

Transfer papers can also be printed with textile-printing equipment. Thus roller printing (Sec. 3.4.1) opens the way to very fine half-tone images and shadings, just like gravure printing. Rotary-screen printing (Sec. 3.4.2) is a less expensive approach to reproducing simpler designs. Printing in both cases is from roll to roll with aqueous pastes.

Dyes. The dyes employed are selected disperse dyes with fairly low relative molecular masses. These are supplied by the dye manufacturers in various forms suitable for the preparation either of paper-printing inks based on solvents (alcohol or toluene) or water-based printing pastes.

Sublimation and solubility characteristics of the various dyes must be well-adapted to the substrate in question in order to ensure that color shades will remain constant despite variations in the transfer temperature. The depth of penetration into the substrate can also be controlled by suitable choice of the dye.

Substrates. Polyesters are the most important substrates because of their good colorfastness characteristics and low thermoplasticity. Over 90 % of thermal printing occurs on woven and knitted fabrics made of this fiber. The remaining 10 % represents mainly polyester – wool mixtures or mixtures of polyester with other synthetics. Other types of synthetic fiber are of minor importance.

Triacetate and acetate fibers tend to be deformed rather severely during the transfer process. They also become somewhat harsh to the touch, and their fastness properties are only moderate. Polyamide fibers can be imprinted to give intense and brilliant colors, but the resulting wetfastness is so poor as to be adequate only in exceptional cases.

There has been no lack of attempts to extend dry-transfer processes based on sublimable disperse dyes to natural fibers, especially cotton. One suggestion involved partial esterification or etherification of the cotton, and many preparations of this type have been described. However, the fundamental problem has still not been solved, because optimization of the dye yield is not accompanied by an improvement in fastness toward moisture and light. It has therefore not proven possible to establish the Sublistatic process in the broad area of polyester – cotton blends, though some of the resulting products might be acceptable for reasons of fashion.

Nontextile Substrates. Apart from textiles, plastic films and other flat materials are also subject to printing by the dry-transfer process. Prolonged storage of such products may be accompanied by migration phenomena, however, since disperse dyes are subject to progressive diffusion. With fibers, on the other hand, such diffusion is practically invisible due to the limits imposed by small fiber cross-sections.

Transfer Machines. Presses are used for printing single items by the transfer process, whereas rollers are more suitable for long runs of fabric.

Apart from the usual equipment, which results in contact times of 30 s to 4 min at temperatures of 180 – 230 °C, special machines have been designed that take advantage of a vacuum [28] to reduce the transfer temperature and time, which also helps to preserve the structure of the textile.

3.6. Special Printing Methods

3.6.1. Carpet Printing (→ Floor Coverings and [20])

Jet Printing of Carpets. Many equipment manufacturers have attempted to perfect processes for the spraying of printing dyes through independently controllable jets [29–33]. The potential advantages are:

1) A printing technique that avoids direct contact with the product
2) Good pile penetration due to the high kinetic energy of the air-borne dye, an effect that cannot be achieved with other printing processes
3) Use of low viscosity printing pastes that are easy to fix and wash out

4) The cost of producing printing forms is eliminated, costs that are not inconsiderable in the case of carpet printing due to the size of the stencils

The technical and electronic challenge is formidable, however. There are at present two ways of carrying out a jet-printing process.

In the first approach, a separate fixed jet holder is assigned to each color. A "spray line" is thus created on the moving textile that advances in the warp direction. The jets must be mounted so close to each other that opening all the jets results in complete coverage. By controlling each individual jet, lines of color of various lengths are produced, permitting the reproduction of any desired pattern (Millitron, Deering Milliken, USA).

In the second method, the jets are mounted on a sliding frame that can itself be moved in the direction of the warp while the fabric remains stationary during the spraying process (discontinuous process). The number of jets employed can thereby be considerably reduced, but with a correspondingly lower rate of production. It would be possible in principle to operate with only one jet assigned to each color, and to mount all the jets on a single sliding frame. In practice, jet-printing machines have been constructed in which each color has its own sliding frame, whereas with others all the colors (eight, for example) are sprayed from a single frame. The number of available fabric colors is of necessity limited, as with any printing processes not based on the trichromic system (Chromotronic or Chromojet, J. Zimmer, Klagenfurt).

Jet-printing machines differ from one another in another respect as well: the method used for controlling the dye jets. The first approach entails jets producing a continuous stream of dye that can either be directed onto the textile or diverted into a channel for return to the dye-storage tank. In the second (Chromotronic) method the dye stream is switched on and off by a magnetically operated needle valve.

The application of jet printing is limited by distortion that occurs as a result of contact between the moving textile and the jet, or, in the Chromotronic process, the moving jets and the stationary textile. A further limitation (though more in the sense of cost) arises from the number of jets that must be installed close to each other. For textile-printing purposes it is easy to incorporate 5 – 7 jets per centimeter. This arrangement in turn limits the minimum size of the individual points of color to ca. $2 \times 2 = 4\ mm^2$ based on the minimum width of the corresponding lines and their minimum length as established by the speed of the on/off mechanism. This is roughly equivalent to the pattern potential provided by the tufts of a woven carpet. Nevertheless, it is to be anticipated that there will be further advances in the spray technique, which is already highly developed for such purposes as printing addresses for the delivery of newspapers. Further progress is especially welcome since, except for flat screens, the production of stencils – especially electrolytically produced metal screens – is very costly and consumes a great deal of energy. "Ink-jet devices" that avoid the use of stencils were demonstrated successfully at the 1991 Hannover International Exhibition of Textile Machinery for producing patterns on both paper and textiles (Stork, Brabant), which is to say that the process is by no means restricted to carpets.

Deep-Dye Process (Bigelow, USA). The printing element in this case is a shallow plastic tank that permits all the colors required for a design to be applied simultaneously. Individual dyes within the tank are separated by divisions, and each is fed by a separate injector. The carpet is led through the system with the pile side down. Printing requires that the dye container be raised and pressed against the pile (discontinuous operation).

BDA System (Bradford Dyers Association, UK). In this screen-printing process based on a double squeegee system, dye penetration is enhanced by a vacuum applied to a perforated plate that acts as an undercloth beneath the carpet (discontinuous process).

Stalwart – Pickering System (UK). This is a relief printing process involving rollers to which pieces of foam corresponding to the pattern shape are attached by means of an adhesive. A low-viscosity dye solution is picked up from a tank (continuous process).

All three of the above printing processes are normally carried out before the carpet backing is applied.

Yarn-Printing Machines. Special yarn-printing machines have been developed for yarns intended for tufted carpets. An attempt is made whenever possible to apply the dye in an irregular way to prevent the formation of "rivers" (undesirable repeats) on the resulting tufted carpet. With the Laing Space-Dyeing machine (Pickering, UK) the yarn is passed between a pair of rollers, each with a relief printing profile (usually a bar pattern advancing in the longitudinal direction with respect to the rollers). The lower roller runs in a trough of dye, and the upper roller has the larger diameter of the two. Yarn is imprinted only when bars on the two rolls meet or partially overlap. This produces a printed stripe of irregular width on the yarn bundle.

Stalwart System. This system operates with two smooth rollers, with an upper roller that can be moved vertically. Dyed bands of varying width are produced by electrically raising and lowering the upper cylinder at irregular intervals.

Knit – Deknit Process. The knit – deknit process begins with previously knitted tubes made of a texturable material. These are then imprinted on both sides with deeply engraved rollers and subsequently fixed and washed. Once the yarn has been again unravelled it is ready to be fed into a tufting machine.

Other techniques exist as well for dyeing yarns in a partial and irregular way, but these do not involve printing.

Warp Printing. Warp threads arranged parallel to each other and containing "auxiliary wefts" introduced at various intervals are subject to printing with rollers or rotary-screen printing machines. Here, too, a problem is the width of the textile: warp threads cannot be placed parallel to each other in sufficiently close contact to prevent excessive spread in the case of a carpet pile printed with standard equipment. Such carpet can of course also be printed intact by the flat- or rotary-screen process.

3.6.2. Other Processes [34]

Brush Printing. Dye is applied in this case with brushes rather than squeegees (e.g., in the printing of flags). This results in extremely good penetration, corresponding essentially to a completely dyed product. Brush printing can also be used in making imitation fur and other "long-pile materials." However, most challenges of this type are today met by mechanical means (e.g., spray printing) or through use of special printing dyes.

Melange Printing of Slivers. Slivers are printed with relief-type printing rollers equipped with diagonally arranged channels (S- or Z-shaped). The width of the printed stripes determines the color shade of the blend (melange) ultimately produced when the fibers are later combed. Printed slivers are not dried, but are instead placed in wire baskets and wet steamed in a round steamer for ca. 1 – 2 h in the case of wool. Polyester slivers are steamed under pressure.

Yarn Printing. This technique is known as either *pearl* or *flame* printing depending on the size of the printed pattern. The yarns are imprinted by two relief rollers bearing axial stripes or "discs" running in precise opposition to each other, providing both pressure and back-pressure. Each roller has its own device for dye application. The hank, which is stretched between two beams, is passed through the pair of rollers, causing perhaps one-third of the total length to be printed. An equivalent amount of hank is then transferred from one beam to the other and printed once again, the process being repeated until the entire hank of yarn has been printed. The printed hanks are hung up, dried, and fixed – usually by steaming, but sometimes by hot air, as in the case of pigment printing.

Flock Printing [35–37]. Here the textile is first coated with an adhesive and immediately "sprinkled" with "flocks" 0.3 – 3 mm in length in the presence of an electrostatic field (20 000 – 60 000 V). These flocks consist of fibers cut to a constant length and usually made of pretreated viscose. In such an electrostatic field the flocks tend to orient themselves in a single direction, and they are impelled like small spears into the adhesive. The potential acceleration of the flocks is limited by the maximum achievable voltage. The distance between the poles must also not be

too great, since air exerts a retarding effect such that further increases in the voltage eventually result in no further acceleration, and may in fact cause spark discharge.

Loose flock dust on parts not subject to printing can be removed by reversing the electrostatic field, or by heating the textile after fixation with concurrent suction.

A flocking apparatus of this type can be constructed as a flock-printing carriage (Wirth), which permits its use in conjunction with screen-printing tables. Special devices also exist for combined printing and flocking (Menzel). "Multicolor" flocking (Hofer, Octen/Switzerland) takes advantage of several discrete printing and flocking units. Intermediate drying of the adhesive and vacuum extraction of excess dyed flocks occur between the individual units.

Flock printing is highly dependent on current fashion. Whereas flock covering of complete surfaces has become firmly established, flock patterns are less popular today than formerly.

The Printing of T-Shirts. In recent years there has been a great demand for printed T-shirts, prepared, for example, on a modern version of a carousel printing machine, one in which both the carrier and the garments rotate in a circular fashion underneath the stencils. The process is discontinuous, and individual T-shirts are positioned and removed by hand (Variprint KSD II system from Maag and Schenk).

Batik Printing. Batik in its original form was not in fact a printing process, but today there are several ways of achieving this resist effect. For example, the penetration of dye liquor can be prevented by tying (*tie dyeing*) or by sewing such that a resist effect is achieved. A related technique is the *bandanna* process, which emphasizes dyeing that occurs under cold or only slightly warm conditions over a short period of time. A similar mechanical resist system, known as the *Golgas* process, involves several layers of textile pressed together between metal plates bearing a relief pattern that are immersed in the dye liquor.

In the classical batik process, wax resists are poured onto a fabric according to a predetermined pattern. It is also common to imprint the molten wax with the aid of stamps made of brass or wood. Unprinted areas are then dyed under conditions that will not cause the wax to melt. The principal dye is often indigo. This process of resist application and dyeing can be repeated many times, whereby the wax resist is "boiled off" after each step.

Batik imitations are produced with the aid of various resist-printing pastes, and on the basis of several printing processes. The one most closely resembling the original batik process entails application of a liquid wax, for example by heated printing rollers. Veined designs result from breaks in the wax layer. Resists are also often printed wet-in-wet, and batik imitations can be produced by printing appropriate resists and bases onto naphtholized fabric (see also Chap. 2 and Sections 5.3 and 5.5.3).

3.7. Two-Phase Printing

Two-phase printing refers not to a specific printing technique, but to the separate application of dye and fixing agent. The dye is printed and dried, and the fixing agent is applied later, after which the textile is then steamed. This separation was an essential one with *vat dyes* (see Sec. 5.5.2) due to instability of the associated reducing agents during storage prior to steaming. Production rate also plays a role. Printing machines are capable of operating much more rapidly than the steamers used in the Rongalit – potash process, which entails 8 – 10 min of steaming time. Separation of the process into two phases makes it possible to use rapid fixing auxiliaries such as sodium dithionite ("Hydrosulfit," BASF), thereby reducing the steaming time to 30 s – 1 min at higher steaming temperatures and at the same time opening the way to greater consistency and efficiency. Two-phase printing also provides the solution to a special problem: the reuse of residual dye. In the Rongalit – potash process, printing-dye residues from the body of the roller machines usually contain excess dye from the doctor blade together with metal ions resulting from chemical or electrochemical reaction. These ions act as oxidation catalysts, and can accelerate decomposition of the Rongalit C to such an extent that no fixing occurs on steaming. It can therefore be dangerous to mix such dye residues into fresh dye.

Analogous two-phase fixation with reduced steaming time is employed with *reactive dyes* (see Sec. 5.10.7). An energy-saving variant of two-phase dyeing is the slop-padding process, also used in conjunction with reactive dyes, which helps avoid the risk of staining by highly concentrated alkali or sodium silicate solutions (osmotic effect; see Sec. 5.10).

Other related processes include the application of dye in two stages (e.g., base application and naphtholate printing; see Sec. 5.8), the resist process with and without intermediate drying, and conversion printing.

Conversion Printing. This process, which is used only with cellulose fibers, is a combination process in which overprinted patterns are created by the selective application of a fixation auxiliary (e.g., Rongalit C) that fixes one dye and decomposes another, while in those places where no fixing agent is applied all that occurs is fixation of the second dye.

The most prevalent combination is that of a vat, leucoester, or coupling dye with an oxidation, basic, or reactive dye. The prerequisite for the success of this dyeing process is a difference in behavior on the part of the two dyes in question with respect to the chemicals applied (fixing agents, discharge agents). For environmental reasons, such processes are now mainly of historical interest.

3.8. Drying

Drying Prior to Fixation. To prevent smearing of the print as it passes over the guide rollers, drying is commenced in such a way that the rollers come into contact only with the back of the textile, or in some cases the material is kept in suspension by heated air emerging from jets on both sides. Drying equipment is usually coupled with the associated printing machine. Various drying methods are practiced, including ones based on hot air or the combustion gases from burners, direct contact with the fabric (cylinder drying), and radiative heating, but the best method found so far utilizes hot air drying in a chamber ("mansard"), preferably one fitted with air jets (Fig. 3). Care must be taken with the use of exhaust gases, because many dyes are sensitive to sulfur dioxide and nitrous gases.

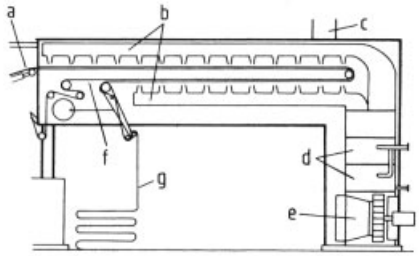

Figure 3. Jet-drying chamber
a) Textile inlet; b) Hot-air pipes with jets; c) Exhaust-air extraction; d) Heat exchangers; e) Air recirculation fan; f) Transport belt; g) Textile exit

The introduction of dye categories and substrates whose color tones are easily altered has required the development of special drying procedures. For example, residual moisture has become an important control characteristic. Also, there must be provisions for uniform air velocity, and the ratio of inlet air to exit air must be precisely controlled.

The drying of textiles made of synthetic fibers brings with it additional problems, since the corresponding fibers are almost incapable of absorbing water except in capillary spaces that the fibers create. For these and other reasons, cylinder drying machines are rarely used in conjunction with modern printing machines.

Fixation by Dry Heating. This technique was introduced for use with certain types of dyes and for pigment printing on cellulosic textiles. For most sublimable disperse dyes the high temperatures of the thermosol process (Sec. 6.5.4) lead to better fixation yields with polyester fabrics than can be achieved through the steaming method. Sublimation fixing also works well with the transfer-printing process in the temperature range 190 – 205 °C.

A suitable facility for dry fixation consists of tentering frames, jet dryers, various condensing devices, and high-temperature cylinder arrays heated by gas, oil, or electricity rather than steam. Economical operation is possible because of the much shorter fixation time.

3.9. Steaming and Other Fixation Processes

The standard fixation method for dyes on cellulosic fibers is steaming, and in some cases (e.g.,

with vat dyes) this is the only feasible approach (see Sec. 5.5). The required steamer capacity depends on the rate at which prints are produced, the type of dye, the printing method (flat-screen printing, rotary-screen printing, or roller printing), and the type of fabric.

Steaming is carried out with saturated or even "wet" steam (i.e., steam containing condensate). With certain dyes (vat dyes) and certain substrates (e.g., wool), the use of saturated steam is preferred. For this purpose, the steam is usually "moistened" by passing it into a tank of water (indirect steam). After heating the water in the tank, a stream of saturated steam passes directly through the vessel, picking up water droplets as it goes, and then enters the steamer from below via a sieve tray. Additional "direct" steam may be introduced through jets. Once all the air has been displaced, the steamer is ready for operation. In general, air inside a continuously operated steamer is vented by way of the same slots through which fabric enters and leaves. The absence of air can only be ensured if excess steam is always present. Steam is lighter than air, so the air should be displaced from above in a downward direction. This is the principle underlying the Arioli-type steamer, which is completely open on the underside and can be operated with either saturated or superheated steam. Fabric is transported through the steamer by motordriven rollers.

Continuous Steamers. The *rapid steamer* system of Mather and Platt includes two rows of rollers between which the textile is moved backward and forward, with only the upper rollers normally being driven. The lid and the inlet box may be heated to prevent drops of water from forming, since these could produce "water stains" on the fabric.

In the *spiral steamer* (Krostewitz) system the cloth is transported by guide rollers that come into contact only with the back of the fabric, and which have a "broken carrying surface" to facilitate keeping the cloth at full width (Fig. 4). The rollers are so arranged that the cloth traverses a spiral path inward toward the center of the chamber where it is deflected by a guide roller, the so-called "sword," set at a 45° angle to the direction of motion. This directs the cloth out through the middle of the side wall.

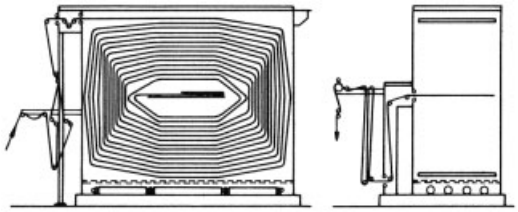

Figure 4. Krostewitz steamer

Hanging or *loop* dryers have only a single upper row of driven rollers. The cloth is first formed into loops by a mechanical looper, and the resulting loops are advanced by a chain of rotating bars. The rollers turn in a synchronized way, but the rate of transport is kept low to prevent the formation of marks at the support surfaces. This type of steamer is exceptionally versatile compared with those described earlier, and is becoming more and more firmly established. Increasing demands with respect to reproducibility of color shades have made it necessary also to incorporate measurement and control devices (from various sources) to ensure constant moisture contents for both cloth and steam.

Two-phase steamers (for use in the two-phase printing process, Sec. 3.7) are designed for rapid textile throughput ($\leq$120 m/min) and short fixation times (arched steamers, portal steamers, lightning steamers).

Batch Steamers. With a *star steamer*, the cloth, sometimes together with a back gray, is hung in spiral fashion on a star-shaped stand, the cloth being "pinned" in place by outward-facing needle bars. This star, complete with the suspended cloth, is then introduced into the steamer. Loading may be through an open lid, through a "door" (cabinet steamers), or from below into a raised "bell" (bell steamers). The steamer itself is a cylinder with a double casing. The bell steamer is a cylinder closed only at the top which is latched to a floor plate after the bell has been lowered.

Steam enters the steamer in most cases via the double casing through openings near the top, passing out through flaps in the floor. However, with a bell steamer (Dedeko, Dupuis), the steam flows through the support columns of the star to the underside of the heated lid, with a steam outlet in the floor.

Most steamers are capable of operating either with or without a sump. Bell steamers, which operate above atmospheric pressure (1 – 2.5 bar), are exceptionally well-suited to fixing dyes on synthetic fibers, especially polyester. A further advantage is that the steaming time can be varied between wide limits. Nevertheless, the number of such steamers in use is decreasing steadily due to their low productivity.

Other Fixation Processes. *Wet developing* is possible only with a small number of dyes. The process depends on the action of chemicals in an aqueous bath, usually heated, and it is conducted in the first compartment of a full-width washing machine. Dyes that can be fixed in this way include the leuco vat dye esters (Sec. 5.7), the coupling dyes (Sec. 5.8), and the Phtalogen complex dyes (Sec. 5.9).

Aftertreatments for dyed textiles depend on the dyes in question, and may consist of oxidative or sometimes reductive steps as well as passage through acids or alkalis.

3.10. Washing Operations

The washing process is even more important with respect to a print than it is for a conventionally dyed textile, because dyes in different concentrations – sometimes also of different dye types with differing fastness properties – may be present together on a white background. Often, the final colors and fastness properties are established only after the washing process.

Washing devices and washing machines are designed to suit the character of particular fabrics. Woven textiles are washed whenever possible in full-width washing machines, though sometimes rope or winch-back machines may be used. Washing of a printed textile is intended to accomplish removal of thickeners, unfixed dye, and auxiliaries. Unfixed dyes may be either soluble or insoluble, but both cases present the danger of staining or bleeding.

Washing begins with a rinsing process to loosen or swell the thickener. The duration of the swelling effect can be improved by "skying" or laying the fabric over chutes, which is often combined with spraying of the cloth. The overall effectiveness of the wash can be greatly increased by inducing transverse motion of the fabric with beater rollers, turbinators, or vibrators as the material moves progressively forward in the longitudinal direction.

Full-width washing machines are divided into several sections. At the end of each section, squeezing equipment is incorporated to prevent dirty wash liquor from being carried forward. Such squeezing equipment greatly increases the effectiveness of the washing process. Washing is accomplished mainly in countercurrent flow with addition of a calculated amount of feed liquor.

The use of perforated drums with eccentrically mounted bearings results in longitudinal and transverse motion of the wash liquor together with good washing action (Vibrotex, E. Küsters, Krefeld, FRG). A perforated drum with wash liquor flowing through the holes is useful generally for both washing and drying purposes (Fleissner).

The design of newer washing machines is directed increasingly toward conservation of water and energy. Examples include the Mäander full-width washing machine and the Rotojet, both manufactured by E. Küsters and both operating on the countercurrent principle, where the cloth moves upward from below and is steeped and squeezed off many times, while the wash liquor flows downward from above.

Before a washed textile is dried it is often dewatered by centrifugation, squeezing off, application of a vacuum (through a slit), or a blow-through with air or steam. Drying with the aid of equipment like that described in Section 3.8, preferably on tentering frames, is almost always combined with the finishing process.

4. Thickening Agents and Thickeners in Textile Printing

4.1. Introduction

The composition of a textile printing paste is determined by the printing method (direct, discharge, or resist printing), the substrate, the application method, and the fixation method, not by the type of dye used. Moreover, all printing pastes contain not only dyes and chemicals, but also thickeners.

Thickeners are those components of a printing paste that impart a printable, pasty consistency to an otherwise aqueous liquor, and they consist mainly of thickening agents. Thickeners supplied in concentrated form for storage purposes are known as *thickener solutions*.

Thickening agents in the narrower sense are substances in powder or granular form (sometimes supplied as low- or high-viscosity "concentrates" or "clears") that display marked swelling properties in water. Such materials form colloidal systems with water, thereby increasing to a greater or lesser extent the viscosity of a dye or chemical solution [38, 39].

Printing paste normally contains 40 – 70 % thickener solution. Depending on the solids content, this typically corresponds to 2.5 – 10 % (16 % maximum) thickening agent [40].

Thickening systems also include emulsions composed of emulsifying agents, water, and heavy naphtha [39, 41]. The addition of such an emulsion to a swollen colloidal thickening agent produces what is known as a "semiemulsion." However, the use of heavy naphtha emulsions is now decreasing rapidly, mainly for environmental reasons.

4.1.1. Purpose and Effect of a Thickener

A thickener confers the required viscosity on an aqueous liquor containing dyes or chemicals to ensure that it can be printed and pumped, producing on the one hand good flow and levelling properties, but at the same time keeping the liquor "short" (i.e., immobile) so that the dye will stay where it should based on the design, permitting the production of sharp contours. In case of an interruption of the mechanical printing operation it is also important that the dye mixture not flow out from engraved areas or screens.

Most thickeners used in textile printing have structural viscous properties; i. e., the apparent viscosity of the mixture changes with changes in mechanical stress (shear stress). Some thickeners also have a flow limit, which is to say that their viscosity in a stationary state is infinitely high, and flow will commence only above some critical value of shear stress. Viscosity requirements differ greatly for the various operations associated with the printing process (e.g., pumping, doctoring) as well as for different printing rates and printing equipment, so it is essential that a printer be familiar with the rheological properties of the various thickeners. The manufacturers of thickening agents provide this information in the form of flow or viscosity curves [39, 42].

However, the effect of the thickener in a printing paste is not limited to providing some desired set of viscosity or rheological properties, thereby permitting particular effects to be achieved, such as sharp contours, surface leveling, or through-printing of a textile substrate. The thickener must also serve as a protective colloid for the various dyes and chemicals dissolved in the paste, and it must ensure the most complete dye transfer possible from the paste to the substrate. Thickeners are also responsible for regulating the moisture balance during the fixation process, which affects the dye yield [43]. Furthermore, dried thickener provides a protective film that prevents the loss of unfixed dye by frictional effects after the textile leaves the drying chamber and before it undergoes dye fixation, a process that does not always follow immediately. This film-forming property sometimes proves inadequate with extremely low-solids thickeners, such as those prepared from the so- called synthetic thickening agents, in which case the further addition of a film former is recommended [44]. On the other hand, the film of thickener must not be so brittle that it flakes off prior to dye fixation, a risk that is always present when printing on hydrophobic, smooth fabrics. The choice of a thickening agent is also influenced by the substrate [45, 46]. It is seldom possible to achieve all the desired effects with a single thickening agent. In practice, several thickening agents are often therefore mixed together, or a ready-mixed combination of thickening agents is utilized.

4.1.2. Demands Imposed upon Thickening Agents and Thickeners

The purposes of thickening agents and thickeners as listed in Section 4.1.1 require that they display a number of specific characteristics:

1) Their rheological properties must match the intended use.

2) They must be available in a standardized form at reasonable cost. The most commonly used thickening agents are modified natural products, the availability of which depends upon weather conditions, cultivation potential, and labor content. Unlike finishing agents, thickeners add no value to the textile, but are simply washed out after use. The cost must therefore be as low as possible.
3) Thickening agents and the thickeners prepared from them must have good storage characteristics. Most "natural" thickening agents as they are supplied contain 5 – 10 % water, and during prolonged storage under "dry" conditions or in a prepared composition the effects of hydrolysis or microbial attack may lead to a gradual loss of thickening power. The addition of preservatives by manufacturers or users can only delay such decomposition, not prevent it. Furthermore, a thickener solution with a high solids content tends to dry out and acquire a surface crust. Residues of such a crust remaining in a finished dye- containing printing paste might lead to defective "skittery" printing.
4) Thickening agents must be easy to use. Thickening agents are now supplied almost exclusively in cold-soluble form [47]. Nevertheless, users often consider it advantageous to mix many thickeners with warm water, or to heat the mixture later [48].
   Mixing methods commonly employed today (i.e., dispersion in water, perhaps by injection pumping, followed by agitation with highspeed stirrers) require free-flowing, nondusting, nonlumping powders or granules. Therefore, some thickening agents are modified chemically so that they only begin to swell slowly some time after wetting ("delayed swelling," "dispersible"), or only after the addition of acid or alkali [49].
5) The finished, fully swollen thickener must be clean and free of large, undissolved swollen particles. Solid impurities could cause the doctor blade of a roller machine to produce marks, and swollen particles might also block the openings of flat or rotary screens. In either case the result is the same: defective printing. The thickener solutions – or better: the printing dyes prepared from them – should therefore be filtered on a vacuum-operated passing machine prior to use. The diameter of the apertures of the corresponding filter sieve should be ca. 15 % smaller than the diameter of the printing-screen apertures.
6) The thickeners must be compatible with all dyes and chemicals that are to be used in the printing pastes, must not react with them, and should themselves have no color or be uniformly colored to prevent problems with spotty or unlevel printing. An important role is also played by ionic properties of a thickener (e.g., possible interaction between an anionic thickener and cationic dyes), as well as by the pH and the electrolyte content of thickeners or the dyes. On the other hand, the incompatibility of some thickeners with alkali, salts of polyvalent metals, and sodium borate (borax) is used to advantage in certain processes (e.g., thickener coagulation in the two-phase printing process).
7) During dye fixation by steam or dry heating, or during wet development of a dye, a good thickening agent should be displaced as completely as possible by the substrate to ensure maximum dye yield, and it should itself adhere or "crab" only slightly to the textile. A thickener that can easily be washed out permits aftertreatment processes to be kept brief, resulting in savings of time and money. Stained white backgrounds attributable to inefficient removal of residual dye and thickener, and printed areas that are harsh to the touch, result in "second-rate" products [50].
8) Thickeners should be environmentally friendly. Quite apart from the continually increasing cost of heavy naphtha, environmental considerations and legal restrictions in some countries place severe limits on the use of hydrocarbons in emulsion thickeners. Other types of thickening agents pass completely into the wastewater when a print is washed, so problems associated with wastewater contamination are becoming increasingly important [47,51]. This applies both to "natural" polysaccharide thickeners and "synthetic" polymer thickeners [52].

4.2. Types of Thickening Agents

Especially in manufacturers' information sheets, thickening agents are classified according to the raw materials used in their preparation.

The standard categories include alginates, seed derivatives (e.g., guar gum), starch degradation products, and the so-called natural, semisynthetic, and fully synthetic thickening agents. However, this classification system says nothing about the most important characteristics of the products. The characteristics of the starting materials, once considered unchangeable, have over the course of time been dramatically altered by improved processing. The guar flour available ten years ago has nothing in common with a modern highly depolymerized guar ether other than the starting material.

It would therefore be better to classify thickening agents according to chemical criteria; i.e., on the basis of the chemical constitution and ionic charge of the various dissolved long-chain polymers [53]. However, this information is not always available, especially for combinations of thickening agents, and it provides little insight into technical printing characteristics of the products. The practitioner therefore seldom looks for a thickening agent of a particular type (whereby exceptions only prove the rule!), but rather follows the recommendations provided by the producer or a product information sheet, determining the true suitability of the product for a particular application through preliminary tests.

Inorganic products are almost never used as thickening agents, but they are added as *fillers* that provide a thickening and water-absorbing effect in conjunction with true thickeners or thickening agents. Examples include zinc oxide, organically modified silicates (Bentone EW/Kronos Titan; → Clays, Chap. 5.2., → Clays, Chap. 5.3.), and hydrophilic silicic acid (Aerosil/Degussa) [54] (→ Silica, Chap. 6.).

Proteins also have no importance as thickening agents. *Albumen* was formerly used as an adhesive with a thickening effect in printing pastes for white pigments in direct and discharge printing, but it is now almost never encountered. However, proteins are present in certain polysaccharides, especially the galactomannans (see below), which are processed into carob-flour thickeners. Seedling residues in galactomannan also contain protein, which can have a harmful effect in carob thickeners [55].

4.2.1. Emulsions

Emulsion thickeners consist of water, nonionic or anionic emulsifiers, and heavy naphtha from a particular boiling range. Their use is very much on the decline worldwide for both cost and environmental reasons. The "water-in-oil" emulsion system was very popular until the late 1980s for pigment printing in the United States primarily because of the high printing rates it permitted, but "oil-in-water" and so-called "semiemulsion" thickeners for printing with reactive, direct, and disperse dyes are now the most widely used systems.

Semiemulsions are prepared by mixing emulsions with thickeners based on swollen polysaccharides, or heavy naphtha may be emulsified into a thickener that already contains emulsifier. The advantages of emulsion printing are good color brilliance, ease of washing out, and minimal influence on the feel of the textile, and these cannot easily be obtained using other thickening systems.

4.2.2. Nonionic Unmodified Polysaccharides (→ Polysaccharides)

Wheat, corn (maize), and rice starch, once popular constituents of thickeners, are now rarely used because of their lack of cold-solubility. The same applies to gum arabic and gum tragacanth, which once were standard products in the printing industry, but today are no longer used because of their high cost and the long preparation times involved.

Unmodified finely ground guar flour is still used in carpet printing processes if it can be obtained at an unusually low price.

4.2.3. Nonionic Modified Polysaccharides

All polysaccharides can be hydrolyzed by acids, alkalis, oxidizing agents, enzymes, and the influence of high temperature. This degradation process significantly alters the properties of the raw material, especially its flow characteristics.

Starch Products. Examples include the hot-water-soluble heat-treated starches ("British gum"; trade names: Diatex, Diamalt) and the cold-water-soluble heat-treated swelling starches (trade names: Diatex SL, Diamalt). These thickeners are introduced at concentrations of ca. 30 – 50 % and are very resistant to alkalis. They are used, for instance, in alkali-based pastes for crepe production, and are also added as color-intensifying components, such as in acrylic printing [56].

Nonionic Starch Ethers and Starch Esters (→ Starch, Chap. 2.2.4., → Starch, Chap. 2.2.5.). These are seldom used alone, but sometimes serve as thickeners in combination formulations. They are more often used as adhesives for woven cellulose (trade names: Solvitose H and H4, Avebe; Texogum 3013, Diamalt).

Gum Resin. The sparingly soluble gum resins are degraded by boiling under pressure. After purification and drying they are marketed in flake or granular form as cold-water soluble *crystal gum* or "industrial gum" (e.g., Cordofan). Solutions (ca. 30 %) of these products display low viscosity, almost Newtonian flow characteristics, and good film-forming properties, but for reasons of cost they are used only as combination thickeners or in high-quality applications, such as outline printing (trade names: Karagum Super, Diamalt; Lamegum, Chem. Fabrik Grünau; Nafka Kristallgummi, Avebe).

Galactomannans (→ Polysaccharides, Chap. 13.). The galactomannan group includes carob flour, which was formerly used extensively in textile printing but is now avoided for reasons of cost. Because of the better availability, good swelling characteristics, and cold-water-solubility of guar flour, this material is now modified in such a way that it satisfies the demands of textile printing [57].

The cold-water-solubilities and hydration rates of nonionic guar derivatives have been improved by hydrolysis or reaction with ethylene oxide or propylene oxide. In contrast to anionic guar ethers (Sec. 4.2.4), the fully etherified products do not react with heavy-metal salts.

Guar derivatives, including the anionic varieties, are marketed under the following trade names, where various grades are distinguished only by added letters: Ago-Gum, Ago Chemicals; Diagum, Diamalt; Guaranate, SFC; Indalca, Cesalpinia; Lameprint, Chem. Fabrik Grünau; Lamberti Print, Fratelli Lamberti; Meypro Gum, Meyhall Chemical; Polygum, Polymer Industries; Polyprint, Polygal; Prisulon, Chem. Fabrik Tübingen; Solvitose, Avebe.

Cellulose Derivatives. Methyl, ethyl, hydroxyethyl, and hydroxypropyl derivatives of cellulose are increasingly recommended as thickener components [58], but they are seldom used alone. They once played a very important role as stabilizing components for emulsion thickeners containing naphtha.

Trade names include Cellosize, Union Carbide; Fine Gum, Dai-Ichi; Natrosol MR, Hercules; Tylose H and MH, Hoechst; Walocel HT, Wolff Walsrode.

4.2.4. Anionic Polysaccharides

Alginates (→ Polysaccharides, Chap. 4.). The alkali-metal salts of alginic acid, especially sodium alginates obtained from species of brown algae, have in the last 35 years become extremely important as thickening agents, especially since the introduction of reactive dyes [59, 60]. Advantages of the sodium alginate thickeners include good water-solubility (facilitating preparation of the thickeners) and ease of removal by washing after fixation of a dye, even at high temperatures. Magnesium and ammonium alginates are also used, the latter for reactive dyes fixed under acidic conditions.

The viscosity of an alginate thickener depends on the origin and species of the algae as well as the corresponding climatic and growth conditions. Degradation reactions occurring during use may lead to reduced viscosity. High-viscosity, medium-viscosity, and low-viscosity alginates are used as thickener solutions in the concentration range 3.5 – 10 %. Calcium salts may alter the flow characteristics of alginate thickeners, but this effect can be prevented by the addition of sequestering agents.

Alginates form elastic films on drying, and it is this property that is exploited when they are used as components in thickening combinations, as for printing on polyesters. Low-

viscosity sodium alginates and high-viscosity magnesium alginates are also applicable alone in this case. Unfortunately the cost of alginates has risen considerably because of cost-intensive methods involved in their production, so attempts are now being made everywhere to find suitable substitutes. Trade names include Alginate, SFC; Cecalgum, Sanofi; Dialgin, Diamalt; Keltex, Kelco; Lamitex, Protan; Manutex, Kelco.

Carboxymethylated Polysaccharides. These anionic products, utilized as their sodium salts, have become extremely important in textile printing. They are obtained by reaction of the alcoholic hydroxyl groups of starch, galactomannans, or cellulose with chloroacetic acid, followed by neutralization [43]. The products contain a variable excess of sodium chloride.

Anionic starch ethers are usually introduced at a concentration of 10 % or as combination thickeners in conjunction with certain classes of dyes in African prints. They display alkaline properties, and therefore have only limited application with acidic printing pastes. Such thickeners coagulate in the presence of salts of trivalent metals. Trade names include Emprint CE, Emslandstärke; Monagum W, Diamalt; Solvitose C5, Avebe.

Anionic etherified seed flours, mainly based on guar, are supplied in various viscosity grades, and have many uses as thickeners, both alone and in mixtures. They too usually have an alkaline reaction, and the pH should therefore be adjusted so that the mixture is neutral or acidic after swelling is complete. Very highly substituted grades are recommended for printing with reactive dyes. The trade names are typically identical to those assigned to nonionic etherified seed flours (see above), but with different letters added as distinguishing marks.

Carboxymethylcelluloses and their sodium salts are thickening agents with good film forming characteristics. They are seldom used alone, but are advantageous as components in thickener combinations.

Trade names include CMC Cellulose Gum, Hercules; Courlose, British Celanese; Tylose C600, Hoechst; Verdickung V extra Super 2, Chem. Fabrik Grünau; Walocel CRT, Wolff Walsrode.

Xanthan. This is a so- called biopolysaccharide (→ Polysaccharides, Chap. 11.) produced by the controlled fermentation of glucose. Its relative molecular mass is reported to be 2×10^6. Aqueous xanthan solutions have unusual plastic flow properties. When subjected to shear stress their viscosity decreases much more rapidly than is the case with other polysaccharide thickeners once the flow limit has been exceeded. These characteristics are exploited in the printing of thick-pile carpet products.

Incompatibility with cationic dyes can be reduced by special additives introduced into the printing paste [61]. Such thickeners are relatively stable toward acids, alkalis, and salts, and combining them with seed-flour products leads to a synergistic increase in viscosity.

Trade names include Kelzan, Kelco; Rhodopol, Rhône-Poulenc.

4.2.5. Cationic Polysaccharides

Products in this category are of little practical importance. The reaction of guar with 2,3-epoxypropyltrimethylammonium chloride has been described [62]. Such thickeners have been recommended for achieving special effects.

4.2.6. Synthetic Polymeric Thickening Agents

The term "synthetic" thickening agents was formerly interpreted to include compounds like acetylcellulose, nitrocellulose, alkylcellulose, hydroxyethylcellulose, polyvinyl alcohol, chlorinated rubber, and the poly(acrylic acids). Of these polymers, only the poly(acrylic acids) are important as fully synthetic thickeners today [→ Polyacrylamides and Poly(Acrylic Acids), Chap. 5.1.].

Other products are obtained by the copolymerization of maleic anhydride and, for example, ethylene, styrene, or vinylmethyl ether with polyfunctional monomers (e.g., divinylbenzene). Aqueous emulsion copolymers of 20 – 40 % (meth)acrylic acid, olefinic compounds (e.g., ethylacrylate), and up to ca. 2 % polyfunctional monomers are of some importance as well [63].

All these thickening agents are long-chain polymers bearing carboxy groups subject to some measure of cross-linking. They swell considerably in water upon neutralization (thickening by swelling), and form high-viscosity gels.

These thickeners are most effective in neutral or slightly alkaline solutions. The commercial products are either already neutralized so that they swell immediately upon addition to water, or else swelling is induced after mixing with water by the introduction of a precisely determined amount of alkali (ammonia, sodium hydroxide, or an amine, depending on the application). The commercial products are supplied either in powdered form or as special liquid preparations, the so-called "concentrates" or "clears" [64].

The preparation time for a thickener of this type (i.e., the hydration time for a polymer thickener) is significantly shorter than the swelling time for a polysaccharide thickener [65].

The structural viscosity of a prepared thickener is usually higher than that of a polysaccharide thickener. A large decrease in viscosity under shear permits the use of pastes with relatively high static viscosities. Process interruptions thus lead to little or no "running" (drop formation), although at the moment of printing-paste application a corresponding high shear stress causes the mixture to become so thin that it passes without difficulty through very fine screen apertures, whereupon it again becomes highly viscous on the textile. It is therefore possible to use these thickeners for printing very fine outlines.

Polymeric thickeners were first introduced into pigment printing starting about 1967. It thereby became possible to reduce or completely eliminate the use of heavy naphtha in printing emulsions. In later years, the unique rheological properties of synthetic thickeners were exploited for printing with acid and metal-complex dyes, resulting in improved penetration with such voluminous materials as upholstery and carpet fabric [66]. However, another feature of polymer thickeners must also be taken into account, namely their sensitivity toward electrolytes. Negative charges on the thickener molecules cause them to become extended and react strongly with added electrolytes present in the form of water hardness, anionic dyes, or setting-up salts. The observed viscosity decreases more or less drastically depending on the type and quantity of electrolyte, and may collapse completely. This has led to the release of special "nonionic" grades of dyes with a low electrolyte content, but so far only in the area of disperse dyes. Completely synthetic polymer thickening agents for printing on polyester textiles or transfer papers have to date been used only to a very limited extent [63, 67, 68].

Trade names include Acraconz, Bayer; Acramin Concentrate, Miles; Aqua Hue Clear, Blackman Uhler; Clear, Minerva; Hostatherm Verdicker, Hoechst; Lutexal, BASF; Lyoprint, Ciba-Geigy; Tanaclear, Charles S. Tanner.

5. Printing on Cellulose (For pigment printing see Chapter 8)

5.1. Pretreatment (See also → Textile Auxiliaries, Chap. 4.)

The quality of a printed product depends not only on the properties of the printing paste and the dye but also on properties associated with the substrate, so careful pretreatment of textiles is always essential. This applies especially to weaves and knits made of cellulose fibers, because these are heavily contaminated on the one hand with natural substances (e.g., seed-shell residues and waxy substances in the case of cotton), but also with sizing agents and preparations (both cotton and regenerated cellulose). The surfaces of some articles may also have been modified by mechanical treatment. For example, textiles with plain surfaces may be singed (with gas, or occasionally an electrically heated plate) in an attempt to remove protruding fibers and loose fiber dust that might stand in the way of a smooth surface print.

Alternatively, textile materials intended for the production of printed fabrics with a "nap," as well as flannels and similar articles, are "raised" on one or both sides with the rotating wire cards of a raising or napping machine. This may be undertaken either before or after printing. For example, if blurring of the pattern is to be avoided the material is usually raised before printing and again briefly at the conclusion of the various finishing steps. The "nap" of surface prints can be leveled by a cropping machine. Similar effects can be achieved by abrading the material. Fiber dust must in any case be completely removed by

air extraction to prevent complications during the printing process.

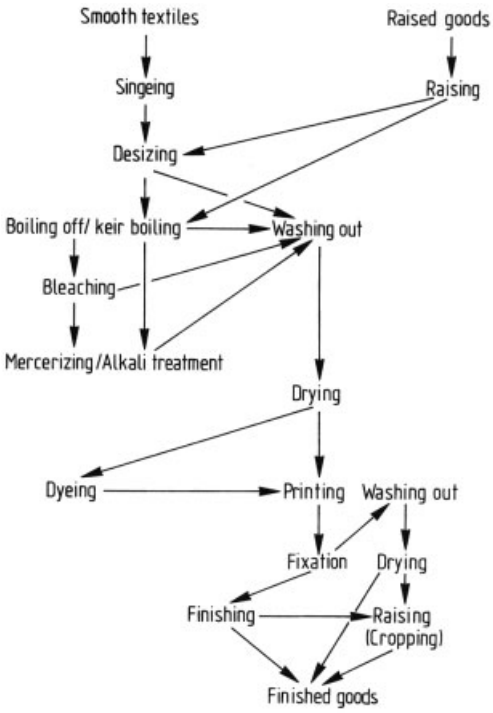

Figure 5. Overall process for the pretreatment and printing of cellulose textiles as it applies to both smooth and raised goods

All soil is removed during the first wet treatment (desizing and/or washing). In the case of cotton this was formerly accomplished with a boiling-off or keir-boiling process in special pressure vessels, but the process has now been considerably simplified through the use of modern apparatus (coiling machines, machines for storage and reaction, and perforated-drum washing units) together with special textile auxiliaries. The next step is usually alkali treatment or mercerization (→ Textile Auxiliaries, Chap. 4.2.2.6.) to improve the material's dye absorption characteristics and luster. Major improvements in the feel (smooth, silky) are achievable especially with fine knitted materials.

Many cellulose textiles in their natural state or subsequent to manufacture display a yellowish to brownish color (linen), so a bleaching step is usually added, involving hydrogen peroxide, sodium hypochlorite, sodium chlorite, or a combination process (→ Textile Auxiliaries, Chap. 4.2.2.3., → Textile Auxiliaries, Chap. 4.2.2.4., → Textile Auxiliaries, Chap. 4.2.2.5.). A pure white color is essential not only for optimum brilliance of the color tones, but also for attractive white-discharge and white-resist effects. Optical brighteners (fluorescent whitening agents) are therefore very often incorporated into the formulations for bleaches, white-discharge printing pastes, and white resists.

A flow diagram outlining the entire operation from gray cloth to finished material is provided in Figure 5.

Viscose fibers (filament and staple), including the high-modulus fibers (→ Cellulose, Chap. 3.1.), are in principle treated like cotton, but always under milder conditions to allow for their greater sensitivity to mechanical and chemical effects. After desizing and/or washing they may be bleached, but only if necessary, and then nearly always exclusively with hydrogen peroxide (rarely with sodium chlorite). Their dye-absorption properties can be improved by treatment with sodium hydroxide solution ($< 8°$ Bé $= 5.2\,\%$, $d = 1.06$). After careful rinsing and sometimes acidification they are dried, but without the use of high temperature (max. 150 °C) to avoid keratinization of the fibers and any diminution in swelling properties. The greater sensitivity of such regenerated celluloses is also relevant to all subsequent processes as well. For example, with strongly acid or alkaline printing pastes the risk of damage in this case is greater than with cotton. Brittle, dried films of thickening agent can lead to creases and folds, especially with fabrics made of fine or highly twisted yarns, such as crepe articles. The thickening agent must therefore be selected very carefully. Especially delicate, fine-quality fabrics of this sort are often wound into rolls after printing (preferably along with the back gray if the printed paste has a tendency to rub off!), and they are folded and washed in small batches after fixation.

Knitted goods are manufactured in both warp and tubular form, and they must be processed with special equipment even during pretreatment that precedes printing. Tubular goods may be processed as tubes on the production line or they may first be cut. Despite edge gumming, an unusable edge remains on both sides in the

latter case, and this must be cut off after drying immediately subsequent to passage through the tentering frame.

In the treatment of *cellulose blends*, especially cotton – viscose staple blends or, to a much smaller extent, blends containing linen, the differing natures of the various fiber components necessitates compromises. If, for example, a linen component must be very white, as in the case of table linens, increased damage of fibers derived from the cotton component is unavoidable. On the other hand, if maximum strength is to be maintained with a cotton – rayon blend it may be necessary to accept the presence of seed-shell residues in the fabric, since boiling-off under pressure would necessarily be ruled out [69–72].

5.2. Printing Methods: Classification and Significance

Specific fiber materials and dye types are the subject of mutual interactions, and it is these interactions that determine the optimum composition of a printing paste, as well as the nature and sequence of fixation and aftertreatment steps. These matters are discussed here not in the order of their importance, but rather in the order of their historical development, although the most important aspects are consistently emphasized. Each class of dye is treated separately under the classical headings of direct, discharge, and resist printing.

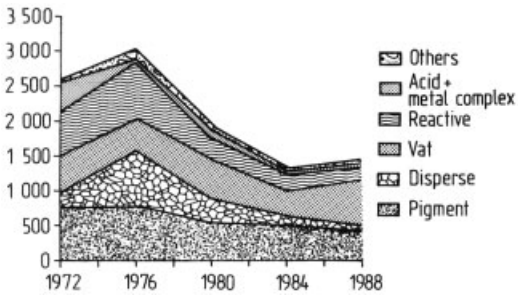

Figure 6. Consumption (t/a) of various classes of dyes by the textile-printing industry in the Federal Republic of Germany, 1972 – 1988 [73]

An impression of the relative importance of the various classes of dyes can be gained from Figure 6, in which the overlap of economic factors and style trends is noteworthy [73]. Figure 6 also shows that pigment dyes are by far the most important, and that a small number of dye categories account for most of the total demand. Substantive and acid dyes represent barely 1 % of total dye consumption, and cationic and discharge dyes have almost completely disappeared from cellulose printing.

5.3. Substantive (Direct) and Acid Dyes

These two classes of dyes differ less in their chemistry (both consist mainly of azoic dyes carrying acidic groups that confer solubility in water) than in their differing affinities for the various fibers (→ Dyes, General Survey). Here they are discussed together because they are usually used together in printing on cellulose, and their subtle differences complement each other. Although their fastness is in most cases only moderate, those dyes are still occasionally used for inexpensive goods because of their favorable price. Also, further savings are possible through suitable choice of particular dyes and simplification of the process.

Direct Printing. A printing paste is prepared by dissolving the dyes in hot water to which is added urea and a solvent (ethylene glycol, thioethylene glycol; sometimes glycerine or a similar substance), and this solution is stirred into a thickener that is easily removed by washing. Small amounts of oxidizing agents are added to prevent reduction effects due to the fiber and/or the thickening agent.

Although many substantive and acid dyes reach maximum dye yield and fastness only after fixation with saturated steam for 30 – 60 min, the steaming time can be reduced to 15 min without incurring significant disadvantages by careful choice of the dye.

If the maximum possible fastness is required it is desirable not only to use a long steaming time, but also to treat the printed textile with auxiliaries that improve wetfastness (cationic quaternary polyammonium compounds) in the course of the final wash. The function of these agents can be interpreted as the linking together of several dye molecules in the fiber so that they become less soluble and also acquire such an

awkward shape that they are much less able to diffuse out of the fiber. One negative side-effect must also be tolerated, however: a slight reduction in the lightfastness of prints so treated.

Discharge Printing. As fashions change in textile printing, there often occur periods in which discharge printing plays an important role. Substantive dyes with azo bridges subject to facile reductive cleavage are attractive candidates for use as dischargeable ground dyes, whereas acid dyes are rarely used for this purpose on cellulose. Sodium hydroxymethanesulfinate serves as the reducing agent (e.g., Rongalit C, BASF, → Textile Auxiliaries), either alone or in combination with discharge auxiliaries such as anthraquinone (a reduction catalyst) or quaternary ammonium compounds (Leukotrop W), depending on the intensity of the color and the ease of discharging the dye. Whiteness effects are improved only in exceptional cases by the addition of alkali, but humectants and wetting agents are advantageous, as are white pigments (preferably zinc oxide or titanium dioxide) or optical brighteners added to improve appearance. Less satisfactory discharge dyeing results are usually obtained with regenerated celluloses, and the fact that there is a risk of fiber damage means that storage of the material (e.g., overnight, either before or after steaming) should be avoided if at all possible. The storage properties of the printed material depend very much on its moisture content and the relative humidity of the atmosphere, and less on ambient temperature.

Special care is required with substantively dyed grounds, whose discharging properties are subject to harm by treatments designed to improve wetfastness. The extent to which such effects are observed is very much a function of the auxiliary used, and it may lead to inferior results (pale or muddy colors) even with colored vat discharge dyeing, especially in the case of bright colors. The cause is not only reduced solubility of the cleavage products of the dyes, which are therefore more difficult to wash out, but also a change in wetting behavior and, in the case of colored vat discharge dyes, the formation of adducts between dyes and auxiliaries that in turn display poor diffusion characteristics under alkaline conditions. In such cases it is preferable to delay the fastness-improvement process until the final wash bath subsequent to printing. Previous washing temperatures should therefore not exceed 30 – 40 °C at most. Also, the wash-bath liquor should be changed sufficiently frequently to prevent accumulation of dissolved ground dye in the wash liquor and thus avoid any danger of bleeding onto the discharge pattern.

Another difficulty is the possibility of "overdischarging" the ground. In roller printing, rough roller surfaces or damage to the rollers or doctor blades can sometimes cause discharge paste – albeit in very small amounts – to appear outside the intended pattern on the textile, leading to a slight overdischarge of the overall dyed product, or fine lines appearing in the wrong places. One possible countermeasure is treatment of the fabric with oxidizing and/or acidic chemicals such as hydrogen peroxide or sodium *m*-nitrobenzenesulfonate (e.g., Ludigol, BASF), alone or in combination with nonvolatile organic acids, preferably glycolic or lactic acid. This "Ludigolization" can be accomplished by dipping or slop padding either before or after printing. The appropriate concentration depends on the amount of liquor applied, and is typically 5 – 15 g/L of oxidizing agent and 3 – 8 g/L of acid (based on 100 % acid).

Resist Printing. This technique is today restricted almost exclusively to batik articles, and involves substantive dyes containing groups subject to coupling (i.e., capable of reacting with diazotized bases or diazonium salts, known as "fast color salts"). Cotton fabric to which a wax pattern has been applied is introduced at room temperature into the cold dye solution, subsequently dried in air, coupled with the diazonium salt, and then rinsed. The wax is finally removed by boiling-off in water [74–76].

5.4. Cationic and Mordant Dyes

These two classes of dyes (→ Dyes, General Survey; → Cationic Dyes) which are characterized by a very wide variety of chemical groups, are now almost never used in cellulose printing for the following reasons. Cationic dyes alone have insufficient affinity for cellulose, and therefore require fixing with a mordant (a tannin; i.e., a synthetic tanning agent such as Katanol ON or

Mesitol) as well as aftertreatment with an antimony salt. Even then the observed colorfastness properties are only moderate, with lightfastness being especially poor. A further disadvantage is that the corresponding printing pastes can be stored only for a very limited time.

Mordant dyes lead once again to environmental problems. The unavoidable use of heavy metal salts, especially chromium salts, is a powerful argument against their application. Also, the colorfastness of mordant dyes (especially wetfastness) is usually only moderate.

Finally, both of these dye categories mandate special thickening agents and dye-specific formulations. All these disadvantages outweigh the favorable economics of the intensely colored and even brilliant representatives of these two dye groups.

5.5. Vat Dyes

Vat dyes (→ Dyes, General Survey) are not only the oldest and most versatile dyes used for printing on cellulose, but also among the most colorfast of all dyes. They are still widely used today in various direct and discharge printing processes, though seldom in resist printing. They are often extremely fast toward moisture, light, and weathering, and are used to great advantage for decorative fabrics, upholstery, and window shades. These characteristics apply with respect to all types of cellulose fibers, including cotton, regenerated cellulose, and linen [73].

The dyes are supplied as water-insoluble pigments, and are also printed in this form (→ Indigo and Indigo Colorants; Anthraquinone Dyes and Intermediates; → Pigments, Organic). They are then converted into the corresponding alkali-leuco compounds ("vats") by the action of alkali and a reducing agent in an atmosphere saturated with water vapor (→ Leuco Esters of Vat Dyes). This provides substantivity toward cellulose, permitting the dyes to be absorbed by the fiber. There, treatment with oxidizing agents (air in the simplest case) in the course of wet treatment (washing) in weakly acidic medium causes them to be reprecipitated as finely dispersed pigments. Dye pigment deposited in the fibers in this way exhibits extremely good fastness properties.

Nevertheless, the reoxidation process must be carried to completion so that no leuco dye will redissolve in the alkaline soap-containing bath and/or bleed onto the white fabric. Also, the subsequent boiling with alkali must remove not only pigment dye adhering to the surface (to improve rubbing resistance), but also ensure that the vat pigment is converted into its final crystalline (crystallite) form. In very many cases it is only in this way that optimum fastness and brilliance and the correct color shade on printing can be achieved with vat dyes [77].

Not all vat dyes that produce good results in normal dyeing are equally well-suited to textile printing. This depends partly on the substantivity of the leuco compounds, which must diffuse as completely as possible into the cellulose fibers during the limited steaming period, but partly on the redox potentials of both dye and reducing agent. Finally, optimum vatting and fixation conditions should be adhered to for each dye with respect to quantities of alkali and reducing agent, time, and temperature.

However, since several very different color shades (and hence different dyes) must ordinarily be fixed simultaneously in a given printed design, the conditions selected necessarily represent a compromise. Under circumstances where, for example, a high alkali content or a longer steaming time might otherwise be desirable, these conditions might in fact lead to the hydrolysis of other dyes present, resulting in changes in shade or a reduction in color intensity. Similar considerations apply in the reverse case as well: use of too little alkali and insufficient steaming can cause rearrangements to dye forms that are more difficult to reoxidize [78].

Apart from careful selection of the dyes, this problem can be combated in two ways (→ Dyes, General Survey):

1) Use of fine or extremely fine pigment-dye powders with increased surface area, permitting considerably faster vatting and hence more rapid dissolution of the dye.
2) Additives introduced by the dye manufacturer to increase the extent of dispersion and provide stabilization and homogenization so that cellulose fibers can be dyed over a wider range of tolerance. With liquid products (pastes), other components are added that confer the maximum possible protection

against sedimentation, drying, and frost damage. Especially printing pastes intended for the sulfoxylate printing process (see below) often contain hydrotropic and hygroscopic chemicals along with reduction catalysts. A single dye from a given manufacturer is thus frequently supplied in three or more forms (powder, granules, pastes).

5.5.1. All-In (Sulfoxylate) Printing Processes

The term "sulfoxylate" refers to sodium formaldehydesulfoxylate = sodium hydroxymethanesulfinate [*149-44-0*], C.I. Reducing Agent 2, marketed, for example, as Rongalit C (→ Textile Auxiliaries).

Of the various sulfoxylate printing processes, the "Rongalit C – potash process" developed by W. Sieber in 1908 has become the most important. It is equally suitable for direct printing or for producing color-discharge effects. It is still widely used today for printing cellulose with vat dyes, although many precautionary measures are necessary, a circumstance that has led to an increase in the importance of the "two-phase printing process" (Sec. 3.7).

In the Rongalit C – potash process, also known as the "all-in process," the printing paste contains not only the dye but also all the chemicals necessary for fixation. For storage of the paste the dye is often converted to a reduced state, resulting in better dye yields for dyes with medium or dark colors. With light colors, however, decomposition by the reducing agent may lead to an unacceptable loss of color intensity. In the interest of good reproducibility it is therefore advisable to carry out printing tests prior to large-scale use of such a paste.

Manufacturers' recommendations should be followed when selecting a particular *dye*. Products with very soluble leuco forms are preferable, as these will be rapidly and completely reduced and will provide good substantivity with respect to cellulose fibers. Both of these properties depend on molecular constitution, since dyes with relatively low molecular mass are the most soluble, and indigoid dyes are more readily reduced than anthraquinoid dyes, although the latter usually have better fastness properties.

At one time, only paste grades were normally available commercially, but today many users take advantage of the availability of very fine granules with a low dust content. With care, these can be mixed directly into stock pastes using high-speed stirrers to produce speck-free products.

The *reducing agent*, almost always sodium hydroxymethanesulfinate, reduces dye in the printing paste rather slowly at pH 10 and temperatures up to 55 °C. At higher temperatures the reaction proceeds more rapidly, reaching a maximum at 100 °C, which corresponds to the steaming conditions during fixation. If a print is stored before steaming, however, Rongalit C in the printed surface oxidizes relatively rapidly under the influence of high humidity and high temperatures (though hot, dry air is relatively harmless). Complete reduction and fixation of the printed dye is then no longer possible. Storage prior to steaming should therefore be kept to a minimum (< 6 h).

Disodium iminobismethanesulfinate, first marketed in 1959 under the name Rongalit FD [*8065-69-8*] (BASF), is a more stable sulfinate. However, not all dyes useful for printing can be completely fixed with this substance. Also the steaming time must be extended for maximum dye yield, or else the process must be carried out slightly above atmospheric pressure in a star steamer. Rongalit ST liquid (BASF) is a product with many applications that has more recently been introduced for this purpose (→ Textile Auxiliaries) [79–81].

The *alkali-metal carbonate* in printing pastes or thickener stock is usually potassium carbonate (potash), because this is more readily soluble and also more hygroscopic than sodium carbonate. Maximum fixation yield and brilliance are achieved by fixing fabrics printed with vat dyes under saturated-steam conditions at 100 – 102 °C. As a result of the properties noted above, potash increases the tolerance range for steam superheating. On the other hand, sodium carbonate is not only cheaper, but also effective at considerably lower concentration. However, it has the disadvantage of crystallizing out during prolonged paste storage, increasing the risk of damage to printing rollers and screens (especially rotary screens). Only mixtures of the two carbonates are therefore recommended. Sodium hydroxide is applicable as the source of alkali only in special cases (prereduction processes),

since it causes hydrolysis of many of the dyes before the required fixation times have elapsed. Sometimes, especially with blue dye varieties, aliphatic bases are used as the alkali (e.g., triethanolamine) as a way of improving color intensity and purity.

The *thickening agents* (Sec. 4.2) must be stable toward all the chemicals used, especially the alkali, and toward electrolytes in general. Starch/British gum/tragacanth combinations were once very popular, but these have today been replaced by cold-soluble, frequently premixed formulations of starch ethers with seed-flour derivatives and/or special alginates or gums.

Other *auxiliaries* used for printing vat dyes include hygroscopic and hydrotropic substances (glycerine, glycols, glycol ethers, thioglycols, and urea), and sometimes printing oils and defoamers. Whereas oils act in a purely mechanical way to improve printing properties, doctor-blade operation, and the levelness of printed surfaces, the effect of the first-named agents is delayed until the steam-fixation stage, at which point they promote dispersion, dissolution, color intensification, and stabilization, also reducing the negative effect of (slightly) superheated steam by encouraging condensation of the steam on the printed surfaces. Excessively superheated steam (> ca. 106 °C) leads inevitably to the deposition of weak, muddy colors.

Several blue anthraquinoid dyes print at their full brilliance only if smaller amounts of reducing agent are used and triethanolamine is added. The same effect can be achieved by stirring 3 – 5 g/kg of sodium nitrite or sodium chlorate into the standard mix.

Steaming. After printing and drying (whereby the latter must be rapid and above all complete), the fabric is often cooled by air blowing or by passing it over a cooling cylinder to improve its storage properties prior to steaming. Possible effects of unfavorable weather conditions can be excluded by air conditioning (e.g., silica-gel drying).

Steaming with continuous or star steamers using air-free saturated steam at 100 – 102 °C requires 5 – 15 min. Regenerated celluloses usually demand longer steaming times. Temperatures in the steamer must be carefully controlled to prevent damage from overheating due to the effect of exothermic processes, including liberation of heat of swelling of the fabric (depending on the initial moisture content), heat of solution for certain chemicals in the condensate, heat of condensation, and the decomposition of reducing agents. It is safer (though more expensive) to achieve the necessary control by maintaining a high flow rate of steam rather than by spraying water into the steamer, as this expedient leads to frequent droplet formation. With water spraying there is also the danger of air leakage, best prevented by a slight overpressure in the steamer.

Aftertreatment. After steaming, the printed fabric must not be stored for too long prior to washing because reducing agent residues may continue to decompose exothermically, leading to heat build-up in the stacked material and defective dyeing or even browning of the cellulose. If a delay of several hours is anticipated before the wet aftertreatment the fabric should be cooled with air ("skying") to oxidize at least some of the excess reducing agent [82].

Washing of the printed material, whether at full width or in rope form, begins with a thorough rinsing in cold water. After this, reoxidation is carried out with hydrogen peroxide in the presence of a small amount of acetic acid at 50 – 60 °C. A soap treatment with sodium carbonate at the boiling point should be initiated only after this process is complete.

5.5.2. Two-Phase Printing Processes (See also Section 3.7)

The storage stability of unsteamed fabric from the all-in process is poor, and there are sometimes problems with color reproducibility (prereduction). Also, fiber damage is possible if drying is inadequate, and the results are sensitive to fluctuations in the steaming conditions. This led at an early date to the development of alternative processes (E. Pfeffer, 1926, "Colloresin process"), which became well-established only after about 1954 in conjunction with the construction of special steamers and the development of suitable reducing agents.

In the two-phase processes of today (mainly "flash-aging" methods), dye and chemicals are applied in two separate steps. In the first step only the dye and thickener (the "solid phases")

are applied, thereby avoiding all stability problems. The chemicals (reducing agent and alkali, sometimes together with other auxiliaries) are added in a second step in the form of an aqueous solution (the "liquid phase") by padding. The fabric is then immediately steamed; i.e., the dye is fixed by exposure to an air-free atmosphere of steam for 15 – 60 s. Provided the fabric is wet, no damage is possible during fixation due to heat arising from condensation, enthalpy of reaction, or superheated steam, because the surface of the textile is under saturated steam conditions. For the same reason the steaming temperature also plays a relatively unimportant role. However, it is necessary to ensure that the relationship between the weight of the textile, the amount of water applied during padding, the heat caused by condensation, and the extent of the condensate is such that it excludes possible running of the printed pattern. Considerable help in this respect is provided by addition to the padding liquor of substances for coagulating the thickener. In some cases the added alkali is sufficient, but borax is often added when seed-flour derivatives are used, and aluminum sulfate is introduced with alginates. This measure has the further advantage that it helps prevent marking of the printed product by the guide rollers, a problem that can also be overcome mechanically by only allowing the guide rollers to contact the back of the cloth.

Unfortunately, thickener coagulants may hinder diffusion of leuco compounds into the fiber and thereby decrease fastness toward washing. For this reason it is common practice today to employ mixtures of coagulable products with thickening agents that cannot be coagulated, such as starch or, more recently, starch ethers [83].

Dye Fixation. Fixation is accomplished by a combination of padding and steaming. Many variations on the process are possible depending on the equipment available. Cloth is either run continuously through the developing bath or it is slop-padded with the liquor on one side only prior to being squeezed off by passage between a pair of rubber rollers or a combination of rubber and steel rollers. The fabric is then fed into the steam chamber through which it is guided by rollers contacting the reverse side only. The steam chamber can be of the tower, channel, vault, spiral, or hanging type, with the latter being used for knitted goods. Depending on the fabric load and the production rate, long- or short-duration steaming methods may be used, in each case with the appropriate formulation and choice of reducing agent (see Table 2).

Table 2. Typical formulations for padded developing liquors

	Long-duration steaming	Short-duration steaming	
Water∗	700 mL	700 mL	700 mL
Sodium hydroxide	40 g	60 g	60 g
Borax	0 – 10 g	0 – 10 g	0 – 10 g
Rongalit C	80 – 100 g		
Hydrosulfite (conc.)		80 – 100 g	
Rongalit 2PH			80 – 100 g

∗ Additional water is added for a total batch weight of 1000 g

"Hydrosulfite" (sodium dithionite [*7631-94-9*], sodium hyposulfite, C.I. Reducing Agent 1) is quite sensitive to oxygen, so air leakage must be prevented between the squeezing-off step and the steam inlet (40 cm maximum) when this agent is employed, and steaming times must be kept very short (20 – 40 s). Sodium formaldehyde sulfoxylate (sodium hydroxymethanesulfinate, C.I. Reducing Agent 2) is used instead in the long-duration steaming method (3 – 15 min). A substance superior to hydrosulfite for the short-duration process is Rongalit 2 PH (BASF), which is also a stabilized derivative of hydroxymethanesulfinic acid (C.I. Reducing Agent 13). This substance displays good liquor stability and does not reduce dissolved vat dyes in the temperature range 20 – 30 °C, thus eliminating the possibility of color changes in the ground. It also permits more thorough wetting of dense fabrics through inclusion of an intermediate air-blowing stage.

The alkali used is normally sodium hydroxide, sometimes mixed with sodium carbonate or potassium carbonate. The most likely further additives apart from coagulation agents are alkali-resistant wetting agents.

If aluminum sulfate is used instead of borax for the coagulation of alginates it is important that it be predissolved in water and then reacted with sodium hydroxide solution to form sodium aluminate, where aluminum hydroxide represents an intermediate stage.

Immediately after padding with the developing liquor (70 – 80 % uptake of liquor with a concentration equivalent to that listed in Table 2), the

wet cloth is steamed. A slight overpressure in the steamer produces an outward flow of steam through the cloth inlet and outlet slots, ensuring the absence of air in the steamer.

The steaming temperature is usually in the range 120 – 130 °C, but may fluctuate within wide limits depending on the cloth. It is important that the cloth not be allowed to dry out at high temperature, because the leuco compound of the dye would then diffuse back, and sensitive dyes might partially hydrolyze.

Washing and Reoxidation after Printing. This is usually carried out continuously in a single production line to prevent marking of the moist cloth during storage. The method is in principle analogous to that used in the all-in process. Thus, the fabric is first thoroughly sprayed in a chute or J-box (although perforated drums with spray tubes and/or water recirculation are often used today), preferably with water from the last rinsing bath of the washing machine. It is then reoxidized by air or added hydrogen peroxide, and finally boiled with soap and sodium carbonate solution [84].

5.5.3. Discharge Printing Processes

Vat dyes are very seldom used as discharge grounds, and then only for white discharging of light to medium color shades, because the discharge pastes required are so aggressive that color-discharge printing is practically ruled out because of instability. However, vat dyes have long been used for the illumination of other ground colors with good discharging properties, since the same reducing agents that cleave the ground color (substantive dyes, insoluble azo dyes, reactive dyes) in washable products are also suitable for fixing a vat dye by the sulfoxylate process (see Sec. 5.5.1). However, a successful application requires careful choice of the dyes, and this is true of both the ground dye and the color-discharge dye. Otherwise, although the color might well be discharged, it is possible that the vat dye applied simultaneously would not be fixed, or that fixation would be achieved on printing, but the available reducing power might no longer be sufficient to discharge the underlying ground color. This effect depends on the redox potentials of the dyes employed.

Auxiliaries for increasing discharge effects include anthraquinone, Leukotrop W (BASF), the calcium salt of a disulfonic acid of dimethylphenylbenzylammonium chloride, alkalis (sodium carbonate, potassium carbonate, sodium hydroxide solution), and hydrotropic agents or dispersants. However, the tendency of these substances to attack the color-discharge dye and/or the cloth represents a disadvantage, because it results in poor paste stability, poor storage stability, unsatisfactory reproducibility, and the risk of fiber damage. For example, augmented discharge pastes of this type are not at all suitable for regenerated cellulose [85].

In discharge printing, aqueous solutions of the weak oxidizing agent Ludigol (BASF, → Textile Auxiliaries, Chap. 6.2.2.) are applied to the fabric by padding either alone or in combination with nonvolatile organic acids, especially glycolic or lactic acid, using a Foulard or special padding machine. The fabric is subsequently dried. This precaution, sometimes delayed until after the printing step, affects the decomposition of any Rongalit introduced through unintentional application of white or color discharges caused by defective rollers, screens, or doctor blades, as well as discharge halos (aureoles) that might be present.

Table 3. Typical formulations for white discharge pastes

Ground dye	Vat dye		Substantive dye	Insoluble azoic dye
Thickener*	450 g	350 g	450 g	450 g
Zinc oxide 1 : 1	150 g	100 g	100 g	100 g
Rongalit C	240 g	240 g	150 g	200 g
Leukotrop W (conc.)	60 g	160 g		80 g
Ludigol	30 g	30 g		
Sodium hydroxide (solution, 38 ° Bé)		180 g		
Potassium carbonate			0 – 90 g	60 g
Anthraquinone				20 g
Water or thickener to 1000 g				

* Gum resin/British gum mixtures or alkali-resistant seed-flour ether thickeners

The risk of defective printing is especially severe if errors are not immediately visible, appearing only after steaming. The extra precaution is therefore taken of treating with Ludigol

or adding optical brighteners (5 – 20 g per kilogram of printing paste) that fluoresce under UV light, requiring that special lamps be mounted on the printing machines [82].

Examples of typical formulations for white discharge pastes are provided in Table 3.

The formulations for color-discharge pastes are identical to those used in the all-in process (Sec. 5.5.1), although depending on the ease with which the ground is discharged, 50 – 150 g of Rongalit C may be added per kilogram of printing paste.

After printing and drying, the cloth must be steamed as quickly as possible, because the dual role of the sodium formaldehyde sulfoxylate doubles the harmful effects anticipated during storage. The steaming time, other steaming conditions, and the conditions applicable to the post-printing wash are comparable to those in direct printing (see Sec. 5.5.1). It is essential that the colorfastness of the ground color at the proposed washing temperatures be thoroughly tested.

5.5.4. Resist Printing Processes

Neither the mechanical (paste resist) nor the chemical resist process is now used with vat dyes. This is not only a consequence of the relative complexity and uncertainty of the operation, but also the fact that the chemicals required (heavy-metal salts) should be avoided for environmental reasons. This leaves only the use of wax resists with indigo as a continuation of the Indonesian folk tradition of batik printing, as well as certain techniques associated with African printing [76, 86].

5.6. Sulfur Dyes (→ Sulfur Dyes)

Apart from black grades, sulfur dyes play almost no part in textile printing because of their dull colors and only modest fastness. They behave in the dyeing process like vat dyes, except that they can be converted to the vat form with such weak reducing agents as sodium sulfide or glucose. However, the black versions (Indocarbon, Hoechst, → Sulfur Dyes, Chap. 1.) can be used in the two-phase printing process alongside vat dyes without any change in formulation.

5.7. Leuco Esters of Vat Dyes (→ Leuco Esters of Vat Dyes)

Leuco esters of vat dyes are the water-soluble sodium or potassium salts of sulfuric acid esters of the leuco vat dyes. Their importance in printing has diminished greatly, and they are now used only in special cases, if for no other reason than cost. For example, color shades that could not be obtained by coupling reactions of diazotized bases or stabilized diazonium salts were formerly reproduced by direct printing onto Naphtol AS grounds. Thus red naphthols (AS, AS-OL, AS-D, AS-PH, AS-BO, etc.) were in this way supplemented by yellow, green, and brilliant blue. Today reactive and Phtalogen dyes are used for this purpose in preference to leuco vat-dye esters.

5.8. Azoic Dyes Produced by Coupling (→ Azo Dyes)

The production of insoluble azoic dyes directly on a piece of fabric is one of the oldest dyeing techniques. The coupling component 2-hydroxy-3-naphthoic acid anilide [*92-77-3*], Naphtol AS, was introduced in 1911 (→ Azo Dyes), and it is still used as such or in the form of one of its many derivatives. The *two-stage* process requires that a "naphtholated" fabric be printed with a diazotized base, usually a diazotized aromatic amine. For this reason the process is also referred to (incorrectly) as "base printing." A naphthol preparation includes the naphthol itself, sodium hydroxide solution, alcohol, and a dispersing agent that prevents the precipitation of naphthol when the concentrated solution is diluted. The naphthol preparation is applied by padding and drying. It is somewhat sensitive to atmospheric moisture, light, and carbon dioxide.

The diazonium salts obtained by diazotizing a base are rather unstable, and must be printed immediately. Apart from the diazotizing solution, the corresponding printing pastes contain sodium acetate, acetic acid, and a suitable printing-paste thickener. The commercial process is simplified today by extensive use of stabilized diazonium salts. These are complex double salts of diazonium ions with, for example,

zinc chloride, 1,5-naphthalenedisulfonic acid, or tetrafluoroboric acid (→ Diazo Compounds and Diazo Reactions).

Coupling to form the dye, starting with either a diazotized solution or stabilized diazonium salts, is carried out immediately after printing so that after drying the printed article can be washed to remove unreacted naphthol; it may be, however, that other dyes added to produce specific color shades will require an additional fixing step. Supplementation in this sense is frequently necessary, because not all shades can be obtained with naphthol preparations alone. Missing colors include yellow (except for Naphtol AS-G and other yellow naphthols), turquoise, and a brilliant blue with the clarity of a phthalocyanine dye. Printing on Naphtol AS and AS-G, supplemented with selected reactive dyes, Phtalogen dyes, and pigment dyes, and with the aid of sophisticated resist techniques, is also still extremely important in African printing. Otherwise, however, the technique has declined today to the point of insignificance. The same applies to the single-step application of mixtures of diazo and coupling components (e.g., the Rapidecht, Rapidazol, and Rapidogen dyes).

This displacement in favor of reactive and pigment dyes is somewhat regretable from the standpoint of fastness properties. Mixtures of the type described were of considerable importance in direct and reserve printing with aniline black and Phtalogen dyes. Following disappearance of aniline black articles, with which the best color resists were always obtained (e.g., using Rapidogen dyes), one of the last areas of application for this type of dye has probably been lost. Apart from ecological and toxicological considerations, the following causes can be identified:

1) Interference from so-called cross-coupling in the event of overprinting. Thus, when two dyes consisting of different naphthols as well as different diazo compounds are combined, instead of the desired two dyes, four dye combinations result, with different colors and sometimes differing fastness properties. A predictable result is possible only if a match is present for at least one of the components.
2) The stability of the printing paste depends on stabilization of the diazonium salt, and is in most cases only modest.

5.9. Phthalocyanine Developing Dyes [87, 88]

Insoluble phthalocyanine–metal complexes (→ Phthalocyanines) can be caused to form on the fibers starting with an isoindolenine and a heavy-metal compound, both of which must be present in the printing paste:

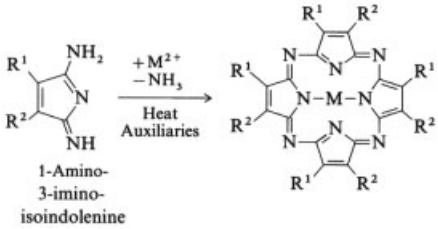

The substituents R^1 and R^2 may also be joined in the form of a ring.

Such a printing paste also contains other components designed to improve the solubility of the isoindolenines, suppress hydrolysis, and ensure the liberation of ammonia on heating. The only metals used in practice are copper, cobalt, and nickel. A range of phthalocyanine developing dyes is available from Bayer under the trade name Phtalogen.

Direct Printing. Phtalogen dyes are mixed into a paste with solvents, usually consisting of glycols, glycol ethers, aminoalcohols, and amides. To this paste is added an ammoniacal solution of a copper or nickel complex, and the resulting solution or fine dispersion is mixed into a previously prepared thickener. Thickeners for this application must be compatible with the solvents, and must also remain unaffected by the metal ions. Practical advice is available from manufacturers of the thickening agents.

The print must be dried immediately and completely in a batch process, where the endpoint is recognizable by a color change in the direction of the final color. The printed pattern is at first almost colorless, but this changes upon drying at $> 80\,^{\circ}C$ with appropriate air circulation as a consequence of complex formation, even though this initial color is not yet that of the final pigment. Fixation can be accomplished by

steaming for 5 – 8 min at 100 – 102 °C, or by hot-air treatment for 5 – 3 min at 135 – 145 °C. Inadequately dried prints lose color through hydrolysis in the course of steaming. If the drying-chamber capacity is insufficient to produce consistent drying results, the cloth can be further dried with drying cylinders.

Excess heavy metal should then be removed in an acid treatment (10 g/L HCl, 70 – 73 °C, 30 s) in order to prevent subsequent damage by peroxide or chlorine from the washing or bleaching agents. The thickener and any dye pigment deposited loosely on the surface are removed by rinsing and boiling-off. Ionic copper can also be removed by a Trilon B bath (1 – 2 g/L or more, depending on water hardness) if auxiliary dyes happen not to be compatible with a strong-acid treatment.

Combinations with Other Dyes. Phtalogen dyes can be printed together with reactive dyes, azoic coupling dyes, leuco-vat esters (Anthrasol and Indigosol), and pigment dyes. With other classes of dyes there is often a problem of incompatibility with the aftertreatment required for Phtalogen dyes, or else the fastness properties do not match.

When printing with reactive dyes, acidification is carried out with 5 – 10 g of formic acid instead of the usual hydrochloric acid. The boiling-off and rinsing processes must also be adjusted to suit the unique demands of reactive dyes.

Combination of Phtalogen dyes with vat dyes is possible in the two-phase printing process (Sec. 3.7) provided the fabric is dried sufficiently vigorously after printing to ensure proper development of the Phtalogen dye.

Phtalogen dyes are especially important for the printing of cellulosic fibers. In fiber blends (e.g., polyester and cotton), a 50/50 mix usually represents the limit for use of these dyes alone. Mixing Phtalogen and disperse dyes in a paste can open the way to printing other fiber blends. The Phtalogen dyes are so exceptionally colorfast that it is possible in this case to implement washing procedures capable of removing the disperse dye from the cotton constituent.

Printing with Complex Dyes (→ Metal-Complex Dyes). The initially formed metal-containing complex is a precursor to the actual phthalocyanine. Early examples include Phtalogen Brilliant Blue IF3GK and Phtalogen Turquoise IFBK. These display low solubility in water, and are therefore more resistant to hydrolysis than other Phtalogen developing dyes. Their solubility in mixtures of diglycol ethers and tertiary aminoalcohols (triethanolamine) can be described as adequate.

The complex dyes are developed in the usual way by steaming, hot air treatment, or wet development with reducing agents. Development with reducing agents can be used for direct or two-phase printing in conjunction with vat dyes, because the dye color remains unaffected.

Because of the large size of the molecules (or association compounds) constituting the Phtalogen complex dyes, together with low solubility in water, their use is limited largely to suitably pretreated cotton.

The high proportion of solvent mandates a careful choice of thickening agents. Apart from the classical starch – tragacanth thickeners the only option available is starch ethers etherified to a specific degree.

Phtalogen Blue IBN is the most important complex dye from a commercial standpoint. It is a cobalt complex that has been rendered water-soluble by the presence of a basic group. Fixation causes this group to be cleaved, and the resulting cobalt phthalocyanine is deposited on and in the fibers. In addition to thermal fixation methods, reductive treatment can also be used to produce the cobalt phthalocyanine pigment, which has a clear, dark blue color.

The solubility of Phtalogen Blue IBN in dilute acetic acid is high, so no additional auxiliary is required. The importance of the product lies in direct printing and printing on naphtholized textiles, where the acidic printing mixture also permits admixture with stabilized diazonium salts or even diazotized bases. The dark green, brown, or black colors so obtained are both intense and fast.

A typical formulation for 1000 g of printing paste combines 30 g of Phtalogen Blue IBN, 10 g of 60 % acetic acid (up to 30 g for an alkaline Naphtol AS preparation), ca. 500 g of thickener, and 460 g of water or additional thickener. Storage properties are improved by the use of lactic acid in place of acetic acid.

Polyacrylonitrile – cotton blends can be printed with combinations of basic dyes and Phtalogen Blue IBN, and polyester – cotton blends

with combinations of disperse dyes and Phtalogen Blue IBN.

A chemical resist procedure is very difficult to conduct with Phtalogen Blue IBN, but it is possible to work with either a preprinted paste resist (with intermediate drying) or a wet-in-wet resist (Chap. 2). Resists with reactive dyes can also be obtained with acidic printing pastes.

5.10. Reactive Dyes [89–97] (→ Reactive Dyes)

5.10.1. Introduction

There are two basic reasons for the steadily increasing importance of reactive dyes in textile printing:

1) The brilliance and wide range of available colors
2) The ease with which the dyes are applied

Other advantages include stability of the fiber – dye bond, facile removal of unfixed hydrolyzate by washing (i.e., low substantivity), a high fixation yield, and good build-up properties, as well as good stability with respect to the printing paste.

Based on this profile of characteristics the monochlorotriazine (MCT) reactive dyes have established themselves extremely well in textile printing applications. Sulfatoethylsulfone (SES) dyes are used for special two-phase applications. All the other types of reactive dyes, developed mainly for dyebath application, are of minor importance in direct printing.

5.10.2. Reaction Mechanism

Reaction occurs primarily in the amorphous region of the cellulose. Depending on the type of reactive group present in the dye, the result is a more ether- or ester-like bond to the fibers. Aliphatic reactive groups substitute preferentially at the C-2 hydroxyl group, forming ether links, whereas triazine derivatives form esters at C-6 hydroxyl groups of the cellulose.

The basic equation is as follows:

$$\text{Chr}-\text{R}-\text{X} + \text{HO}-\text{Cell} \longrightarrow \text{Chr}-\text{R}-\text{O}-\text{Cell} + \text{HX}$$

where Chr = chromophore, Cell = cellulose, R = supporting (bridging) group, and X = leaving group

Thus,

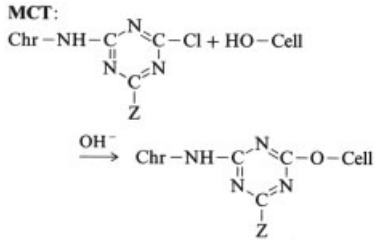

where Z = $-NH_2$, $-CH_3$, or other groups

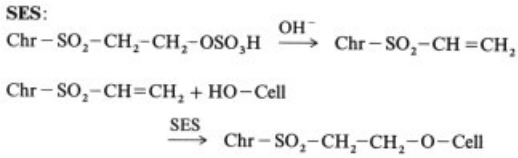

SES dyes must first be converted to the vinylsulfone form in alkaline medium before they can participate in an addition reaction with cellulose.

5.10.3. Important Reactive Dyes

The reactive dyes most important in textile printing include:

MCT Dyes: Basilen P, Cibacron P, Drimaren P, Kayacion P, Levafix PN, Procion P

SES Dyes: Remazol, Sumifix

Reactive dyes for textile printing are chosen on the basis of reaction rate, fixation properties, and stability of the corresponding pastes. They react with cellulose in the presence of an alkali such as sodium carbonate or sodium hydrogencarbonate. Application conditions depend upon the alkali used, the reactivity of the dye, and the fixation temperature. Reactive dyes react with hydroxyl compounds generally, so hydrolysis by water competes with the fiber – dye reaction, and a certain amount of hydrolysis is unavoidable. For this reason, moderately reactive dyes are the only ones used in the one-phase fixation processes for textile printing. More highly reactive ranges are in turn applica-

ble only in the two-phase processes (see Section 5.10.7).

5.10.4. Pretreatment of the Fabric

Achieving a fiber – dye bond and intense, brilliant, clear colors requires that the fabric first undergo a thorough pretreatment. It should be completely desized, bleached, and given good wetting properties. Mercerization is recommended with cotton, and regenerated cellulose should be treated with alkali. Viscose staple fiber should be treated with 3 % sodium hydroxide at room temperature under tension-free conditions. The mercerized or alkali-treated material should be well rinsed, but not subjected to acid treatment. Further pretreatment with urea and/or alkali should be avoided. Although such a treatment might sometimes lead to an increase in color yield, serious problems can develop in other directions. Fabrics so treated may yellow in the course of steaming, and the moisture – temperature equilibrium of a steamer can be upset by the presence of large quantities of urea. Urea removes large amounts of moisture from the steamer, causing the steam to become superheated. The temperature then rises, and after only a few meters of production there is a noticeable decrease in color (or fixation) yield. This can be remedied only by the introduction of excess steam. Large quantities of urea also present an environmental problem. (For details of the conditioning process see Sec. 5.10.6.)

5.10.5. Printing Pastes

Thickening Agents. Reactive dyes may also react with thickening agents if these are similar in structure to cellulose. Starches, starch derivatives, tragacanth, and similar materials are therefore unsuitable as thickeners, causing the fabric to become harsh to the touch and producing poor fastnesses. Commercial grades of alginate are used instead, specifically sodium salts of alginic acid obtained from seaweed. High-viscosity sodium alginates are preferred for reasons of cost. Low- and medium-viscosity sodium alginates offer the best dye leveling and dye yields, especially with regenerated cellulose. Since sodium alginates are natural products they are subject to biological degradation, with a consequent loss of viscosity on prolonged storage even in the powdered state. In the colloidal state (i.e., in aqueous medium) a viscosity decrease due to bacterial attack occurs within a few days, which makes it necessary that preservatives be added to the printing paste. Sodium alginate swells in water, but forms precipitates with the substances responsible for water hardness. For this reason the swelling process (often incorrectly referred to as "dissolving") should be permitted to proceed only in soft water. For added safety, a sequestering agent (polyphosphate) is added to the system to combine with and mask any alkaline-earth or heavy-metal ions.

Apart from sodium alginate, so-called *half-emulsions* of the oil-in-water type are often used, but their importance is diminishing for environmental reasons. Half-emulsions are widely employed in markets where cost and environmental considerations are secondary because they result in high degrees of brilliance, color intensity, and dye leveling, and they are easily removed from the print by washing.

In the Asiatic world, carboxymethylcellulose (CMC) has also been used successfully. Suitable grades of CMC must have a very high degree of substitution (> 2); i.e., all the reactive hydroxyl groups should be inactivated by substitution.

Synthetic thickeners have been utilized for some time as well. These are sodium salts of poly(acrylic acids), and they have the advantages of producing very brilliant and intensely colored printed results, being virtually immune to bacterial attack, and existing as standardized industrial products available in ample quantity. Synthetic thickeners are sensitive to the presence of electrolytes, however, which cause the viscosity of the paste to decrease. The concentration of thickener must therefore be varied in accordance with the dye concentration. Also, synthetics may confer a harsh "hand" (feel) because they are difficult to wash out. Synthetic thickeners are not soluble in water, but merely swell, so a relatively high degree of agitation is required in the washing equipment to remove them without leaving a residue. It is therefore preferable to use synthetic thickeners in combination with sodium alginates.

Urea. Reaction between a reactive dye and cellulose occurs in a liquid phase, which permits

the dye to diffuse from the printing paste into the fibers. The function of urea can be regarded as that of "binding" the water; i.e., it causes a certain amount of steam to condense on the print due to urea's negative heat of solution. The resulting aqueous urea phase, in which the dye is dissolved, functions as a dyeing medium. Steam fixation requires the use of 50 – 80 g/kg of urea for cotton and 100 – 150 g/kg for viscose rayon. Hot-air fixation is suitable only for cotton, and it requires larger amounts of urea. However, the high nitrogen content of urea makes it a threat to the environment. Urea passes through a sewage treatment plant unchanged, and may contribute to hypertrophy in surface waters. For this reason urea must be introduced at the lowest concentration possible (see also Section 5.10.6.).

Alkali. The reaction of a reactive dye with fibers occurs only in an alkaline medium at a rate that is a function of both pH and temperature. Roughly speaking, the reaction rate increases by a factor of ca. 10 with an increase of 1 pH unit or 20 °C. As the reaction rate increases, the rate of hydrolysis by water (the competitive reaction) also increases, which is equivalent to a reduction in the stability of the printing paste. With highly reactive dyes (SES, DCC, FCP, DFT, etc.) applied by the one-phase method, only weak bases (sodium hydrogencarbonate, sodium trichloroacetate, etc.) are used. For dyes of medium reactivity (e.g., MCT) sodium carbonate is appropriate, offering a printing-paste stability of 3 – 4 weeks at ca. 20 °C. If MCT dyes are used with sodium hydrogencarbonate a yield reduction of ca. 10 % must be anticipated with intense colors. Application of pastes free of alkali is described in Section 5.10.7.

Oxidizing Agents. Reactive dyes are more or less sensitive to reduction during steaming depending on the nature of the chromophoric group. The effects of reduction can be countered by adding weak oxidizing agents, such as derivatives of benzenesulfonic acid (Ludigol, BASF). Light colors are especially sensitive to steam fixation. The risk of reduction is generally lower with hot-air fixation.

Preparation of the Printing Pastes. Most reactive dyes are quite soluble, so reliable and rapid incorporation into a printing paste can be accomplished by either the premixing or the sprinkling method. In the latter approach the powdered dye is fed slowly with rapid stirring into a mixture containing all the chemicals and thickeners necessary for fixation. Several manufacturers also supply reactive dyes as liquid formulations that are very easy to handle and are suitable for processing in direct dispensing systems known as automated dye "kitchens."

Typical formulations for 1000 g of paste are as follows:

1) Sodium alginate system: water, 50 – 150 g of urea, 10 g of Ludigol, 15 g of sodium carbonate (25 g of sodium hydrogencarbonate), 450 g of sodium alginate (4 %), and dye
2) Semiemulsion: water and chemicals as above, 150 g of sodium alginate (4 %), 10 g of emulsifier, and 400 g of kerosene
3) Synthetic thickener: water and chemicals as above, 150 g of sodium alginate (4 %), and 40 – 70 g of synthetic thickener (dispersion)

All these formulations apply to single-phase fixation, the so- called "all-in method."

5.10.6. Fixation (One-Phase Method)

The fixation rate is determined principally by the reactive group. Interaction also takes place between the supporting group and the chromophore, however, and the diffusion rate of the individual dye plays a part as well, so reactive dyes may display different reaction rates with the same reactive system. It is well known that yellow (azoic) dyes of low molecular mass are fixed more rapidly than the bulky turquoise (phthalocyanine) types. The fixation time must therefore be adapted to the slowest type of dye used. With MCT reactive dyes in a sodium carbonate medium, exposure to saturated steam at 102 °C for 8 – 10 min or hot air at 160 °C for 3 min is sufficient. Highly reactive systems could be fixed in less time, but these are seldom utilized in all-in pastes because the storage characteristics of such pastes are unsatisfactory. In practice, reactive-dye prints are usually treated with saturated steam in short-loop steamers for 8 min. Hot air is less often used because it is suitable only for cotton and less satisfactory for regenerated fibers (low yield). Regenerated cellulose requires for optimum dye fixation a relatively

high moisture content, and this must be achieved during the fixation process. The problems associated with addition of large amounts of urea have been described in Sec. 5.10.4. In practice, printed goods are often conditioned by external application of moisture immediately prior to steaming. This can be accomplished by spraying according to the Weko rotor principle, foam application by the Stork system, or by contacting the back of the fabric with a wet cloth, the water content of which is controlled (e.g., with equipment of the type supplied by Zimmer). In order to ensure the greatest possible reproducibility, moisturizing equipment of this sort should be subject to hygrometric monitoring and control.

5.10.7. Two-Phase Application

Reactive dyes with high reactivity are applied in alkali-free printing pastes because of their low stability. Such pastes consist only of sodium alginate, oxidizing agent, and dye; urea is not required. After printing and drying, the alkali is applied as a separate second step (second phase). Application may occur by one-side "blotch" or two-sided "pad" application of the alkaline liquor. The print is then fixed without intermediate drying in a single operation, through a brief steaming process (20 – 40 s). Shock fixation by passage through a heated ($< 90\,°C$) alkaline bath is also possible, as is cold-pad batch application with swelling over the course of several hours at room temperature. In all cases, there must be a high electrolyte concentration in the alkaline liquor so that the dye will not bleed, but rather be "pressed" onto the fibers. Suitable additives for this purpose include sodium sulfate or sodium silicate (which also acts as an alkali).

Typical alkali baths are based on 180 g/L of sodium sulfate, 150 g/L of sodium carbonate, 50 g/L of potassium carbonate, and 100 mL/L of NaOH, $d = 1.36$ (38° Bé); the alternative is 800 mL/L of sodium silicate, $d = 1.53$ (50° Bé) and 200 mL/L of NaOH, $d = 1.36$ (38° Bé).

MCT reactive dyes are less suitable for two-phase application. Good color intensities are obtained with regenerated cellulose, but the touch of the fabric may suffer. A steaming step prior to alkali application may prove beneficial in this respect, although there is no theoretical explanation for the phenomenon. The two-phase printing process is questionable from an environmental standpoint because of the large amount of water required for the final wash.

5.10.8. Washing of the Prints

The washing of textiles printed with reactive dyes is less problematic the lower the substantivity of the dyes, as this reduces the likelihood that dye will be absorbed from the wash liquor onto the white ground. If the dyes are chosen properly, thorough cold rinsing is enough to remove most of the thickener and dye. This is followed by a boiling step. Detergents are unnecessary with a sufficiently high bath temperature ($> 90\,°C$). Because of the incompatibility of alginate thickeners with hard water, soft water must be used for rinsing, and a complexing agent must be added in the boiling step. Prints treated with cationic aftertreatment agents are subject to the risk of mordanting effects, in which dye bleeds or is absorbed out of the wash liquor onto the white ground, which is covered with cationic substances. This is especially true of prints sewn onto monocolor articles and subjected to wet storage.

5.10.9. Fastness

Apart from the colorfastness associated with a given article under normal use, there is one unusual fastness phenomenon associated specifically with prints based on reactive dyes: under unfavorable conditions, the fiber – dye bond may undergo "splitting" (cleavage), permitting dye to bleed onto the white ground. This splitting may be thermally induced (thermal hydrolysis, thermal cracking), or it may occur in an acidic or alkaline medium (acid or alkaline hydrolysis). MCT dyes are sensitive in the acidic range, SES dyes in the alkaline range. Dye splitting can be initiated by excessive drying, excessive acidification during washing, the use of inappropriate catalysts at the finishing stage, condensation at temperatures that are too high, or storage in an acidic atmosphere.

5.10.10. Special Printing Processes

Reactive Discharge Processes. Discharge printing on a reactive dye ground is of great im-

portance worldwide. Vat dyes are used to produce the printed colors because the reducing agent (sodium formaldehyde sulfoxylate) and alkali (potassium carbonate) required for their development act as the discharge medium for the reactive dye ground. Discharge paste, with or without vat dye, is first printed onto the reactive dye ground, dried and then steamed immediately in the absence of air.

Typical discharge paste: 100 g/kg of TiO_2 (white) or vat dye, 150 g/kg of Rongalit C, 90 g/kg of potassium carbonate, and thickener (seed-flour derivative).

The ground is usually dyed with SES dye, which is especially sensitive to alkali and is easily and completely discharged without any reaction residue being left on the fabric. Dischargeable reactive dyes with a fiber – dye bond that is more stable under alkaline conditions may leave behind decomposition products that lead to subsequent oxidative yellowing. Such dyes are usually split off reductively at the azo bridge.

Resist Printing.

"Pigment-under-reactive" resists with the usual organic acids (tartaric and citric) have never become well established because of poor application properties associated with the pigment paste (leading to premature cross-linking of the binder) and unsatisfactory resist effect. However, the use of an organic acid water-soluble only in the form of its ammonium salt (e.g., a pyrazolonecarboxylic acid derivative) permits optimal effects to be achieved. A typical system consists of a normal pigment paste to which is added ca. 20 – 40 g/kg of seed-flour derivative, 30 – 40 g/kg of 25 % ammonia, and 40 – 60 g/kg of Levalin 46040 (Bayer) as the resist agent. This paste is preprinted and then overprinted without intermediate drying ("wet-in-wet") using a normal alkaline reactive paste (MCT or SES), after which the fabric is dried, fixed with hot air for 3 min at 160 °C, and finished in the usual way.

Reactive/reactive resists can be obtained by blocking SES reactive dyes with stabilized sodium hydrogensulfite compounds (BASF resists) under MCT reactive dyes. A 30 – 40 g/kg of BASF resist is added to the MCT printing paste (standard formulation), which contains sodium carbonate. This paste is preprinted, and, without intermediate drying, wet-in-wet overprinted with an SES printing paste (sodium hydrogencarbonate, standard formulation). It is then dried, fixed in saturated steam, and finished in the usual way.

Crêpe Effects. Crêpe articles can be prepared by overprinting cotton – which has not been subjected to mercerization or alkali treatment – with a solution of sodium hydroxide. The fabric shrinks in the areas printed with sodium hydroxide, thus producing the desired crepe effect. The selected dye must be characterized by a strong fiber – dye bond in alkaline medium (unlike SES, for example).

6. Printing on Synthetic Fibers

Pigment printing is discussed in Chapter 8. Thickeners mentioned in this chapter are discussed in Chapter 4; for other relevant substances → Textile Auxiliaries.

Fabrics based on regenerated cellulose are printed with the same dyes used for cotton (see Chap. 5). However, dyes must be carefully selected for use with fibers derived from cellulose esters (→ Cellulose Esters) and synthetic polymers (→ Fibers, 4. Synthetic Organic), and it is sometimes necessary that dyes be developed specifically for a particular situation.

Compared with natural fibers, man-made fibers based on cellulose esters and synthetic polymers create special dyeing problems associated mainly with the constitution of the fiber material. Other factors are important as well, however, including delustering of the spinning material, fiber cross-section, the spinning processes themselves, and processing to produce filament or staple fiber yarn. The thermoplastic character of synthetic fibers in this group also limits the maximum treatment temperature in processes such as fixation.

Unlike natural fibers, man-made fibers (with few exceptions) are relatively hydrophobic, so dyeing is generally more difficult.

The nature of the textile surface to be printed (woven, knitted, or tufted) also has a significant effect on properties. For example, a closely woven fabric made of polyamide filament behaves like "umbrella silk" and is very hydrophobic. On the other hand, the same polyamide raw material can be used to produce carpets that pick up

so much moisture during the printing process that an intermediate drying stage is unnecessary prior to dye fixation by steaming, and no running is observed with pattern contours. This effect is less unusual than it may appear. After all, a rubber sheet is extremely hydrophobic, whereas a rubber sponge can absorb many times its own weight of water.

6.1. 2,5-Acetate

Developed about 1925, 2,5-acetate was the first of the fibers to differ significantly in its physicochemical behavior from the natural fibers used exclusively until that time. Nevertheless, this pioneering material has now been largely replaced by other man-made substances.

Acetate in both woven and knitted form is used for dresses, blouses, linings, and scarves. Its current share of the total volume of printed products is significantly less than 1 %.

Pretreatment. It is important to note that acetate fiber is sensitive to the action of alkali at temperatures above 85 °C. Partial fiber saponification occurs under these conditions, leading to loss of luster and strength.

Acetate fibers are today sized only with water-soluble preparations, which means that they need only be prewashed at 60 °C with a solution of an anionic detergent (ca. 1 g/L), such as Levapon TH (alkylsulfonate in combination with polyglycol ethers).

The fiber is usually sufficiently white from the outset, so it is seldom bleached. If brightening is necessary (→ Optical Brighteners), the usual choice is a dispersion brightener, such as Blankophor DCR (Bayer) or Hostalux AR (Hoechst). This is applied after the tentering process or after steam padding. The final white shade first becomes apparent following steaming of the printed product.

Printing Processes. Printing can be carried out either on tables or with roller or rotary-screen machines. Depending on its quality, the fabric may be glued directly to the undercloth or provided with a back gray.

Dyes [98] **and Printing Pastes.** Woven and knitted acetate fabrics are today printed almost exclusively with select *disperse dyes* (→ Disperse Dyes). Liquid preparations are preferred, since these are easy to incorporate into printing pastes. Predispersion is unnecessary.

The dyes are selected according to their color shade, lightfastness, wetfastness, sublimationfastness, and fabricationfastness. The relative importance of the various criteria differs from one finished article and fiber to another.

Some dye manufacturers have applied special designations to disperse dyes particularly well-suited to acetate. Examples include the Celliton (BASF) and Cibacet dyes (Ciba-Geigy).

Fixation. Disperse dyes (and sometimes cationic dye as well) are fixed by treatment with saturated steam in a star or hanging-loop steamer at 102 °C for 20 – 40 min. The steam should not be too moist, since the pattern might otherwise run.

Aftertreatment, Fastness Properties. The fixed prints are thoroughly rinsed, washed with an anionic detergent at 40 – 50 °C, and again rinsed until the water is clear.

The fastness of disperse and cationic dyes is only modest. What is known as "fumefastness" can be a problem with some anthraquinoid dyes, but it is subject to improvement by addition of GFN Inhibitor (BASF) – an organic base [99] – to the printing paste (see Sec. 6.2).

Discharge Printing. Textiles dyed with selected disperse and cationic dyes can be used as a ground for discharge printing. The best white discharge results are obtained with Decroline (zinc formaldehydesulfoxylate [*24887-06-7*], → Textile Auxiliaries). With color discharge dyeing, tin(II) chloride is used as the discharging agent because of the wider range of dyes resistant to its effects. Processes are also known based on a tautomer of thiourea dioxide (formamidine sulfinic acid [*1758-73-2*], Manofast).

Other Classes of Dyes. *Acid dyes* and *metal-complex dyes* are subject to fixation without steaming. Instead, advantage is taken of the swelling of 2,5-acetate fibers in solvents. A typical 1000-g printing paste formulation consists of 10 – 60 g of dye, 30 g of urea, 280 – 330 g of water, 130 g of a solvent mixture, and 500 g

of thickener solution. The solvent mixture is made up of 30 g of Glyezin A (thiodiethylene glycol, [*111-48-8*]), 50 g of benzyl alcohol, and 50 g of ethyl alcohol, whereas the thickener solution (1000 g) contains 40 – 50 g of 2-hydroxyethyl methyl cellulose [*9032-42-2*] (Tylose MH), 600 g of ethyl alcohol, 280 – 290 g of water, and 70 g of ammonium rhodanide [*1762-95-4*].

Fixation occurs during drying at room temperature or with warm air. Because of its high alcohol content, the printing paste must be kept away from open flames, and the prints may not be dried with infrared radiation. In the final washing process, alcohol previously introduced as a swelling agent is diluted and removed. Dye remaining on the fiber displays a relatively high degree of fastness. White grounds are also not subject to bleeding. However, the brilliance of the resulting prints is somewhat inferior to that obtained with disperse dyes.

Blends with Other Fibers. Acetate blends are sometimes produced for reasons of fashion, including the blends with polyamide known as Rhodialon, which are printed with disperse and cationic dyes just like pure acetate. In the case of dark colors (navy blue, wine-red, black) the final wash is complicated by the presence of the polyamide, and the white ground is very easily stained. Nevertheless, very fashionable discharge prints have been achieved with precisely this fiber blend. The ground is dyed with selected disperse dyes, whereas cationic dyes are used for the color-discharge dye system, with a tin(II) salt as the discharge agent.

6.2. Triacetate

Triacetate is more resistant to alkalis relative to 2,5-acetate, and it also has less tendency to crease. It occupied an important position in the 1960s in the production of printed fabrics for fashionable clothing. At that time up to 10 % of the prints produced in certain industrialized countries were made of triacetate. Today, it has been largely replaced by polyester, though a small amount is still used in blends with polyamide (e.g., under the name Tricelon).

Triacetate fiber resembles acetate fiber in several of its mechanical and chemical properties, but it is more like polyester in its absorption capacity with respect to disperse dyes.

Pretreament. *Woven* fabrics are first washed with ca. 1 g/L Levapon AN or TH liquid at ca. 60 – 80 °C, usually in mildly alkaline medium, and subsequently rinsed, all under tension-free conditions if possible. The fabric is then given an "S-finish" whereby a surface soaping produces a delustering effect, also serving to improve the feel, increase fastness toward exhaust gas for subsequently applied dyes, and provide a wash-resistant antielectrostatic finish. However, not all fabrics are subjected in practice to an S-finish because it can reduce their capacity to absorb disperse dyes. This problem does not arise if the dye is fixed in a pressurized steamer (saturated steam at 2.5 bar), but use of a rapid or high-temperature steamer results in a significant loss in color intensity.

Triacetate knits processed on a large scale are washed in full-width low-tension facilities. Temperatures close to the boiling point are required for stabilization. On a smaller scale, book-form or high-temperature equipment is appropriate. The optimum fixation temperature is 130 °C.

Knitted fabrics do not normally receive an S-finish.

Printing Processes. Printing is carried out on tables or with roller, rotary-screen, or flat-screen printing machines (see Sections 3.3 and 3.4.2). Especially with discontinuous table and flat-screen printing, the textile must be held firmly to the table by adhesives applied directly or adhesives applied to a previously attached back gray.

Dyes [100, 101] **and Printing Pastes.** Nearly all printing on triacetate is accomplished with specially selected disperse dyes, available from all the major dye producers.

Fixation Accelerators (→ Textile Auxiliaries, Chap. 6.2.5.). The range of available color-intensifying fixation accelerators (auxiliaries) is greater with triacetate than polyester. Even thiourea is effective with some dyes. Various manufacturers offer fixation auxiliaries, but their effectiveness must always be tested as a function of fabric, pretreatment, dye, and fixation process. It should also be noted that many

such products fail to dry completely on the fabric, so a film of printing paste may smear if too much fixation accelerator is used. Examples of fixation auxiliaries include Luprintan HDF and PFD (BASF), and Sandotherm ACS liquid (Sandoz) [99] (→ Textile Auxiliaries).

Fastness toward Nitrogen Oxides ("Fume Fading"). This can be an important consideration with both triacetate and 2,5-acetate. A number of disperse dyes on these fibers display more or less marked color change on exposure to nitrous gases like those present in the off-gas from industry and coal combustion plants. Fastness toward nitrogen oxides can be tested by the method prescribed in DIN 54 025 and ISO 105 601, and it can be improved if necessary by addition to the printing paste of inhibitors, such as GFN Inhibitor.

Fixation. Disperse dyes printed on triacetate can be fixed by various methods:

1) Continuous steaming with saturated steam at 100 – 102 °C for 30 – 45 min. Under these "mild" conditions, addition of appropriate fixation auxiliaries considerably improves the color intensity of many dyes.
2) Steaming under pressure in a bell or kettle steamer for 20 – 30 min with saturated steam at 2.5 bar (ca. 127 °C). Optimum yields are obtained without a fixation accelerator, and the presence or absence of an S-finish has no effect on the results. The disadvantage is the need for batch operation.
3) Continuous fixation with superheated steam (HT-steam fixation) and a steaming time of 6 – 8 min at 185 – 165 °C. Experience shows that temperatures must be somewhat higher in this case than for the same dyes on polyester. Only moderate dye yields are obtained in the absence of fixation accelerators. Disperse dyes on fabric to which an S-finish has been applied cannot be adequately fixed by this method. Also, the high temperatures involved require use of dyes that are stable to sublimation.
4) Continuous thermosol fixing with hot air at 190 – 200 °C for 60 – 45 s. The comments above concerning HT-steam fixation apply here as well.

Aftertreatment. Care is necessary with the final wash after steaming to discourage dye from bleeding onto the white ground. Unfortunately, no way has yet been found to prevent this completely. It is therefore recommended that the fabric first be rinsed cold, then warm, and finally washed at ca. 60 °C. As soon as the washing bath becomes contaminated with dye it must be changed immediately. In order to help keep the ground white it is recommended that small amounts of hydrosulfite (1 – 2 g/L) and sodium hydroxide solution (1 – 2 cm^3/L) be added to the wash baths. The bath temperature should not exceed 50 °C to prevent attack on dye already fixed in the triacetate fiber, which is more strongly hydrophilic than polyester.

After washing, the fabric is rinsed thoroughly both warm and cold, perhaps neutralized, and then dried.

Discharges and Resists. At the time that triacetate was important in textile printing, discharge and resist printing were rarely employed (for reasons of fashion), so no appropriate process of this type was ever developed. Nevertheless, the processes described for polyester should in principle be applicable.

Blends Containing Other Fibers. The only triacetate blends available are combinations with polyamide. Marketed under the trade name Tricelon, these can be printed satisfactorily with select disperse dyes. Although such a print can be fixed with high-pressure steam, continuous steamers based on saturated steam are usually employed instead. The use of HT-steamers is not recommended for these articles, especially if the fabric has been subjected to an S-finish.

The same chemicals are used as fixation agents here as with triacetate itself. Excess fixing agent leads to running of the pattern. Optimum amounts depend on the construction of the fabric, and should be ascertained on the basis of preliminary tests. Fixation in high-pressure steam (20 min, 2.5 bar) avoids the need for a fixing agent.

Special care is required when washing the steamed fabric. Especially when large areas have been printed in dark colors there is a danger that the polyamide fraction of the unprinted surface will be stained with excess disperse dye. One should therefore use a long liquor ratio in the first rinse, and the water should be changed

rapidly. Alternatively, the fabric might be thoroughly sprayed beforehand.

6.3. Polyamide [102, 103]

Dyeing characteristics of polyamide fibers are described elsewhere (→ Textile Dyeing, Chap. 6.).

The printing of fabrics made of polyamide first became important with the development of dimensionally stable warp-knitted fabrics made from filament yarns. Woven and knitted fabrics based on polyamide 6 (made from ε- caprolactam) and polyamide 66 (from hexamethylenediamine and adipic acid) are especially suitable for printing, whereas other polyamides (e.g., polyamide 11, from ω-aminoundecanoic acid) play no role whatsoever because of their poor dye-absorption properties.

For some years, so- called nonelastic knitted goods (porous Perlon or Nyltest) attracted the greatest interest. These were followed by articles with a velour effect. Fabrics of this type are of little importance today, at least in the textile-printing industry of Central Europe. Other products have retained their importance, however, including polyamide locknit fabrics, polyamide taffeta in the form of "umbrella silk," and tubular knitted goods for texturized swimwear. The largest area of current application is the carpet sector.

6.3.1. Pretreatment

Woven and knitted fabrics made of polyamide fiber must be subjected to several pretreatment steps prior to printing.

Two processes are common with knitted shirt materials and locknit fabrics. Thus, if tentering frames are to be used in the thermofixation process (hot air fixation), a tension-free jigger is required for prewashing and optical brightening. If an HT (high temperature) facility is available, prewashing, thermofixation, and (if necessary) optical brightening are carried out as a single-bath process in the appropriate HT equipment (hot water or hydrofixation).

In the first process, which is also well-suited to woven polyamides, the raw goods are washed in the jigger at 40 °C and thermofixed on the tentering frame (polyamide 6 at 192 °C, polyamide 66 at 210 °C). Grades that have not been brightened during spinning are brightened in the jigger and dried on the tentering frame.

In the case of hydrofixation in an HT facility, the fabric is prefixed on a tentering frame at 160 – 170 °C, wound onto the perforated beam, washed in the HT apparatus, brightened if necessary, thermofixed (nonelastic knitted goods made of polyamide 6 at 120 °C, Nyltest at 130 °C), and then dried on the tentering frame and pressed.

Thermofixation in an HT facility is especially recommended for knitted shirt fabrics made of polyamide. This leads to more stable fixation, but may also contribute to the appearance of dreaded moiré (optical) effects. The latter are most reliably countered by careful winding onto a beam as a precursor to thorough prefixing. In the absence of prewashing, however, a prefixing can easily lead to graying as a result of diffusion into the fibers of spinning lubricant and other impurities. The best defense against this problem is a full-width wash on the tentering frame prior to prefixing.

A similar pretreatment is appropriate for polyamide velours. The velour effect is produced by a raising machine, preferably after washing and perhaps optical brightening, but before thermofixing on a tentering frame and printing. The process causes the fabric to decrease in width by as much as 50 %, although the length remains virtually unchanged. Printing in advance is impossible due to subsequent pattern distortion.

Optical brighteners may impair the lightfastness of a printed product. For this reason polyamides that have been brightened at the spinning stage are always preferred. If a polyamide must be brightened later in the process the only auxiliaries that can be recommended are those that display the greatest lightfastness (e.g., Blankophor CL or CLE).

Washing Agents for Precleaning. Both nonionic and anionic detergents are used for washing the various polyamide products to remove pretreatment materials and soil. Alkylarylpolyglycol ethers are the most widely used of the nonionic detergents. Phosphates are the preferred alkalis for washing, since these help to

preserve a soft feel, unlike sodium carbonate, which makes the fabric harsher.

6.3.2. Printing Processes

Pretreated woven and knitted polyamide fabrics are printed by roller and screen-printing machines, as well as on tables. Polyamide velour knits are preferentially printed on roller and rotary-screen printing machines. Texturized polyamide tubular knitted goods, polyamide warp-knitted velour, and polyamide locknits, which display high elasticity and poor dimensional stability, are printed almost exclusively on screen-printing tables and machines. They must be passed through the machines by the shortest possible route with as little tension as possible to prevent distortion. In many cases it is useful to attach the fabric to a cotton back gray.

Nonelastic warp-knitted products such as Nyltest or porous Perlon are less problematic, since they acquire a special "fringe" that confers dimensional stability.

If polyamide knits are glued directly onto a rubber undercloth, wet expansion of the polyamide fibers can become a problem. These fibers may expand by as much as 4 % when moist, so a heavy first coating with dye can very easily lead to the formation of bubbles and wrinkles, preventing the immediate pursuit of subsequent printing steps. Knitted goods behave more satisfactorily if their width is extended by ca. 4 % during the last pass through the tentering frames, and if they are glued to the screen-printing table at this width. Such fabrics remain firmly fixed to the table even when printing multicolored patterns that cover the surface completely, and printing proceeds without problems.

6.3.3. Dyes and Printing Pastes

Direct printing is accomplished with selected metal-complex and disperse dyes, as well as reactive (Sec. 6.3.7) and occasionally substantive dyes. Apart from the required high degree of fastness, dyes are selected according to their color, solubility, and stability.

The solvents used with all types of dyes are thiodiethylene glycol [*111-48-8*] (Glyezin A) and urea. Urea, which can be replaced by thiourea, increases both the dye solubility and the color intensity of the print. Maximum concentrations of thiodiethylene glycol and thiourea quoted in pattern books must not be greatly exceeded, since otherwise the printed patterns may easily run. Other solubilizing agents include diethyleneglycol monobutyl ether [*112-34-5*] (Glyezin BC) and, in special cases, cyclohexanol.

Ammonium sulfate is added to stabilize the pH between 5 and 3, the region in which polyamide fibers display optimum dye-absorption characteristics. This is especially important with acid dyes, whereas metal-complex dyes are sometimes printed without such an addition.

Because polyamide fibers absorb moisture to a relatively small extent compared with cellulose, the situation is different with respect to thickening agents. In the case of polyamides, water soluble dyes together with a larger amount of solvent must be firmly retained on a less absorbent material during the steaming process, but without any running taking place. A satisfactory thickener must therefore have a high solids content.

Medium- to low-viscosity thickeners, or mixtures thereof, are preferred. High-viscosity products are unsuitable. It is particularly difficult to produce a uniformly printed coating on a closely woven fabric like umbrella silk.

The thickeners with the most satisfactory properties are etherified seed-flour and guar products in combination with degraded starches. The solids content should not be less than 12 – 16 %.

6.3.4. Fixation

After drying, the prints are fixed in steam. For woven and nonelastic knitted products, a continuous or star steamer is appropriate. Printed Perlon warp velours, Nylon velours, polyamide locknits, and polyamide texturized tubular knitted goods are fixed in a star or kettle steamer or a hanging loop steamer. In continuous steamers of the Mather & Platt or Krostewitz type they tend to become curled and stained. Minimum steaming conditions for maximum dye yield are 20 – 30 min steaming time at ca. 102 °C. In all cases, shorter steaming times lead to lighter color shades.

The best fixing results are obtained in a star or kettle steamer at 1.5 bar. With approximately the same steaming times, the resulting colors are ca. 10 – 20 % more intense.

6.3.5. Aftertreatment [104]

Aftertreatment processes for polyamide include washing and measures designed to improve fastness. Washing removes excess dye, dye-paste thickener, and chemical additives from the fabric. Constituents of the printing pastes are sometimes difficult to eliminate by washing, and any residues left behind can easily lead to bleeding onto the white ground. The extent of unfixed dye can vary greatly, and depends upon the extent of the coated area, the amount of dye applied, the weight of the fabric per square meter, and the nature of the paste thickener.

Fabric made of polyamide 6 is capable of taking up more dye than polyamide 66 under equivalent conditions. The dye substantivity of both polyamides can be improved considerably by increasing the steam pressure and temperature during the dye-fixation process, which means that the extent of unfixed dye is also a function of conditions.

Unfixed dye can be removed by cold rinsing, washing at 40 – 45 °C, and rinsing again.

Bleeding onto a white ground can be prevented by addition to the rinse of certain auxiliaries (2 – 3 %), usually condensation products of formaldehyde with aromatic sulfonic acids (Mesitol PS or NBS, Erional NWS; → Textile Auxiliaries, Chap. 6.2.9.), and adjustment to pH 3 with formic acid. Such additives block the amino groups of the polyamide, preventing any further absorption of dye. Under these conditions the temperature can be increased to 40 – 45 °C.

Very heavy fabrics, such as upholstery or velour articles containing spandex, can be washed in a bath containing sodium carbonate and Levapon MR (alkylammoniumpolyglycol ether).

Improvement of Fastness. A considerable increase in fastness to moisture can be obtained by further aftertreatment with Mesitol PS or NBS (6 % based on the weight of textile) or Erional NWS at 70 – 80 °C and a pH of 3, obtained by adding formic acid. The maximum increase in wetfastness requires a treatment time of 15 – 20 min. While fastness to light and rubbing are not adversely affected, the recommended procedure may lead to a harsher feel, although this can be compensated for with softeners.

6.3.6. Discharges and Resists [105]

White Discharging with Decrolin. The most suitable ground dyes are selected acid and Isolan (1 : 2 metal- complex) dyes. Some disperse dyes are also used. Although a number of substantive dyes are also very suitable for white discharging, the dyeing of piece goods has been found to be difficult with these materials.

The white effects obtained with Decrolin tend to yellow to a varying extent over the course of time.

Dyed textiles are sometimes given an aftertreatment with Mesitol PS or NBS to improve wetfastness, preferably subsequent to printing and discharging.

Color-Discharge Printing with Vat Dyes. For color discharging with selected vat dyes on grounds dyed with acid and metal- complex dyes, Rongalit C and Rongalit DS (→ Textile Auxiliaries, Chap. 6.2.3.) are the preferred reducing agents. Vat dyes with good fixing properties are suitable. In addition to a few anthraquinoids, indigoid and thioindigoid dyes are particularly worthy of note.

It is advantageous to carry out reoxidation of the color-discharge prints at the pH of acetic acid. Reoxidation with hydrogen peroxide and ammonia generally leads to more or less severe changes in the color of the ground.

Discharging with tin (II) chloride is the most important of the processes used in conjunction with polyamides. It is applicable to both white and color discharge printing, although the color intensity of the ground must be limited to ensure good white effects.

Prior to treatment with Mesitol PS/NBS or Erional NWS (see above), the print must always be thoroughly rinsed to remove any tin chloride residues. Certain concentrations of tin chloride and Mesitol or similar substances result in precipitates that may have a detrimental effect on fabric feel and the clarity of color shades.

The resist process is rarely applied to polyamides.

6.3.7. Reactive Dyes [106]

Brilliant colors stable to a boiling wash can be obtained with selected reactive dyes. Reactive dyes are absorbed by polyamide fibers even from mildly acidic printing pastes, subsequently undergoing reaction at pH 5 – 7. The resulting covalent bonds are very stable toward both alkali and acid. Because of this stable dye – fiber bond, textiles printed with suitable reactive dyes display very good fastness properties.

Suitable thickening agents include high- and medium-viscosity alginates and guar ethers, but not starch- containing thickeners.

Fixation. The dye can be fixed by steaming in a continuous steamer for 10 min; a star steamer is of course also appropriate, but with a maximum overpressure of 1.5 bar.

Final Wash. Prints are first rinsed thoroughly with a cold solution containing 2 g/L of sodium carbonate (to give a pH of 9) and, if necessary, a complexing agent to remove water hardness. Alternatively, however, the fabric can be passed immediately through a hot alkaline rinse bath and then washed with a liquor containing 0.5 – 1 g/L of Levapon TH at ca. 80 – 90 °C. After rinsing both hot and cold, a little acetic acid is added to the final cold rinse to effect neutralization. The washing process must completely remove any unfixed residual dye, since this might otherwise discolor the ground in the acidic medium.

Unlike cotton, polyamide is always washed at a temperature near the boiling point, because unfixed reactive dye on the polyamide surface is retained by a salt-like bond analogous to that of an acid dye, whereas residual reactive dye on cotton is held only absorptively to the cellulose fibers.

6.3.8. Polyamide Blends

Polyamide – Cotton. Blends of polyamide with cotton or viscose staple fiber are so far of minor importance in textile printing because of serious problems – except in the case of pigment printing – of printing in such a way as to produce adequate wetfastness. Moreover, pigment printing is subject to coloristic limits with respect to color intensity and overprint potential.

Blends with Other Synthetic Fibers [107]. Mixtures with spandex (elasthan) marketed under the trade names polyamide/ Lycra and polyamide/ Dorlastan are important in the production of swimwear. Because these fabrics are elastic, and also because of the dye-uptake potential of the polyurethane-based spandex, fibers, special procedures must be followed during pretreatment. For example, the fabric is usually treated with Mesitol PS (0.5 – 1 % based on the weight of the fabric) even before printing (→ Textile Auxiliaries, Chap. 6.2.9.) in order to keep the ground white during the wash that follows printing.

Obtaining good tone-in-tone dyeing results requires careful selection of the dye. Acid and metal-complex dyes are used in a formulation analogous to that employed with polyamide printing. Fixing and final washing are also carried out in the same way as with polyamide.

6.4. Polyacrylonitrile

Polyacrylonitriles have a character very similar to that of wool. Their market share with respect to total fiber consumption is small: in textile printing, only ca. 1 – 2 %. Japan and Italy are the two countries where this fiber is printed extensively, especially for women's outer garments. In Germany it is in steady demand for decorative articles and upholstery, whereas the market for women's outer garments is very much governed by fashion.

The polyacrylonitrile fibers from individual manufacturers vary in their coloristic and technological characteristics. This is partly due to a varying content of acidic groups in the polymer (sulfonic, sulfato, and carboxyl groups), which are essential for fixing cationic dyes. The spinning process (dry or wet spinning) also has an influence on properties. It is even possible to induce an affinity for anionic dyes by incorporating cationic compounds into the polymer.

The flame resistance of polyacrylonitrile is increased by the incorporation of a high proportion (> 15 %) of copolymerized vinyl chloride (Modacrylic fibers; → Fibers, 4. Synthetic Organic).

6.4.1. Pretreatment

It is important to note with respect to pretreatment that there is always a need to remove the

sizing agents (most of which are water soluble) from woven goods, as well as water-insoluble paraffinic compounds from knits. Pretreatment of high-bulked (HB) grades of yarn should lead to an increase in volume, whereas filament and normal yarns are more likely to be stretched, thus requiring that different procedures be followed.

Woven Goods [108, 109]. In general, fabrics with a warp containing polyacrylonitrile filament are sized with water-soluble products, but those in which the warp consists of polyacrylonitrile staple fiber are sized with starch-containing preparations. Starch-containing sizes must be eliminated with appropriate enzymatic agents during pretreatment because unhydrolyzed starches cause hardening of the fabric.

A subsequent wash to remove soil, residual preparations, and other foreign substances is carried out with the aid of anionic or nonionic detergents.

The combined washing – shrinking process for all polyacrylonitrile piece goods is effectively carried out in stainless-steel Mezzera full-width washing and alkali-treatment equipment. Continuous-rope washing machines are suitable only for grades that are not overly susceptible to stretching. Stretching by the reeling machine can be prevented by winding the fabric onto the reel cold. The reel is then stopped, and the bath is heated almost to boiling. To desize the fabric the reel is run briefly a few more times during heating, after which the fabric is allowed to cool for 20 – 30 min with the reel stationary in the machine. It is then run over the reel for a further 10 min for straightening purposes, followed by acid treatment and rinsing.

Polyacrylonitrile piece goods pretreated by one of the above methods are cold rinsed, dewatered by spinning or suction, dried on tentering frames at 110 – 140 °C, and smoothed. The fabric is provided with an oppertunity to shrink on the tentering frames by stretching it to slightly more than its wet width and overfeeding by ca. 5 %.

Knitted Goods [110]. In the 1970s, polyacrylonitrile plain knit (jersey) was produced on a large scale. Pretreatment in this case differs in several respects from that described for woven textiles. Removal of the preparation (paraffinic) and soil requires that the fabric be washed. Suitable detergents include anionic (Levapon TH) and nonionic (Diadavin, polyglycol ether) agents, most of which also contain fat solvents. Cationic auxiliaries should be avoided, since these adhere readily to the fiber surface, impairing fixation of the dye. It is recommended that the last rinse be adjusted to pH 5 – 6 with acetic acid, because alkaline media may lead to yellowing of the fiber on steaming, as well as to decomposition of the dye.

Fabrics with severely crimped edges are pretreated in the tubular state, and are cut after spin drying. In order to prevent excessive longitudinal tension and the formation of longitudinal creases, both washing and shrinking are carried out on low-tension wool-reeling machines using a long liquor. Creasing and folds can be avoided only by cooling extremely slowly.

After rinsing and spinning briefly the fabric is dried on a tentering frame at 130 – 150 °C using maximum possible overfeed, after which it is wound onto a reel. The edges should be glued and cut. Minimum ventilation should be ensured in the tentering frames.

The running behavior of the fabric on a rotary-screen printing machine and in continuous hanging-loop steamers can be improved by padding the knits with an alginate thickener at low concentration after washing and intermediate drying, and then leveling them on the tentering frame.

Whether stretched or placed on a printing table, the fabric must correspond precisely to the intended length, width, weight per square meter, mesh number, and wale count as anticipated for the finished product. After printing, steaming, washing, and brightening these parameters associated with the final stretched product must remain absolutely unchanged, since the printed pattern would otherwise be distorted.

6.4.2. Printing Processes

The pretreated fabric can be printed on roller printing machines or flat- or rotary-screen presses.

Especially with roller and flat-screen printing it is useful to glue the knitted fabric to a cotton back gray. Only gluing machines with drying cylinders are suitable for this purpose, because the strength of the adhesive bond between the

fabric and the back gray will be inadequate unless it is dried immediately. Suitable adhesives include gum – Cordofan 1 : 1 (see *gum resins* in Sec. 4.2.3), dextrin – water 1 : 1, and 15 – 20 % poly(vinyl alcohol). Partial replacement of water in the solvent by alcohol leads to more rapid drying of the adhesive film on the undercloth. Heavy fabrics can also be glued directly onto the rubber printing blanket or undercloth without use of a back gray, although the rubber blanket must be dried completely before the glue is applied. With a screen-printing machine the adhesive-bond strength can be improved considerably by pressing a dummy screen onto the fabric and then drying in a drying device.

6.4.3. Dyes and Printing Pastes

Cationic dyes are of greatest importance for printing on polyacrylonitriles because of their high brilliance, spreading power, and considerably better lightfastness relative to other kinds of fibers [111–113].

The dyes in question (Astrazon, Basacryl, Maxilon, and Remacryl) include the classical triphenylmethane and oxazine dyes along with almost all the other known categories; e.g., anthraquinone, azo, naphthalimide, methine, and cyclammonium dyes (→ Cationic Dyes).

In contrast to the localized charges on sulfonate groups in anionic dyes, the positive charges of cationic dyes are often not localized, instead moving around between several basic groups. Moreover, the relative basicities of these groups may differ greatly. This leads to differences in migration, fixation rate, and fixation yield from one dye to another, as well as differences in the effect of pH on these properties and the formation of more or less sparingly soluble salts with anionic constituents from the printing pastes. For example, dyes with a localized cationic group are fixed more slowly than those in which the charge is not localized.

Polyacrylonitrile fibers also differ with respect to their dye-absorption properties and dyeing rates, so reproducible results can be expected only if the type of fiber, the nature of the fabric, the pretreatment, the paste formulation, and the fixing conditions are all held constant [114].

Suitable *thickeners* are based on materials with high solid contents that give clear pattern outlines and optimum color intensities, in conjunction with substances that form an elastic film after drying of the printed product [115], thereby suppressing flaking when the textile is plaited down and transported.

If large areas are to be printed and exceptionally uniform coloring is desirable, leveling can be achieved by the addition of carboxymethylcellulose, which acts on the cationic dye like an anionic thickener.

Similar leveling effects can be achieved by adding other anionically active materials to the printing pastes. Suitable additives include anionic dispersing agents, such as the condensation products of naphthalenesulfonic acids with formaldehyde. Alginate thickeners behave similarly. In all cases, however, changes in the shade and intensity of the colors must be anticipated.

6.4.4. Fixation

Fixation with High-Pressure Steam. Over the entire range of basic dyes, the best dye yields are obtained by fixation with pressurized steam at 1.2 – 1.5 bar for 30 min. Certain dyes of the azo type are subject to reductive decomposition at temperatures above 115 – 120 °C, so steam temperatures corresponding to the pressures indicated (104 °C and 110 °C, respectively) should not be significantly exceeded. In special cases involving sensitivity of the printing dyes to steaming, a 30 % solution of sodium chlorate should be added at the rate of 5 – 10 g/kg shortly before printing. A discontinuous mode of operation is not recommended.

Continuous Steam Fixation at 102 °C for 30 min. Rapid steamers of the Mather & Platt or Krostewitz type are inappropriate for fixing light woven fabrics. The corresponding steaming temperatures lie within the softening range of polyacrylonitrile fibers, and tensile stress produces distortions of the fabric that are very difficult to correct. Hardening of the goods is also very common.

Continuous steaming at 102 °C in saturated steam became possible with the introduction of hanging-loop steamers. However, selection of the dyes must take these fixing conditions into account. With deep shades, the addition of Luprintan PFD [*69071-73-4*] is necessary.

Considerations Related to Steaming. It is difficult to overstate the importance of constant steaming conditions in fixing water-soluble dyes on synthetic fibers. Unlike cotton and wool, synthetic fibers have only a very low moisture content after drying, so it is almost impossible to compensate for fluctuations in the moisture content of the steam. If the steam is too wet, colors begin to run. If it is too dry, extreme cases may result in a dewatering effect rather than the desired extensive swelling of the thickener. Ion exchange between a dye and the fibers requires the presence of water so that the dye can dissociate. The temperature in the steamer must therefore be monitored carefully throughout the fixing process.

Hot-Steam and Hot-Air Fixation [116]. Fixation processes of this type have been proposed for cationic dyes on polyacrylonitrile, but they have not been widely used.

6.4.5. Aftertreatment

Proper aftertreatment is very important for assuring good rubbingfastness, brilliant colors, and a satisfactory feel. The fabric is first rinsed with cold water containing ammonia or sodium carbonate, washed at 40 – 60 °C with an anionic detergent and hydrosulfite, and then rinsed and washed in fresh liquor at ca. 60 – 70 °C. If the thickeners used contain starch that has not yet been hydrolyzed, it is wise to include in the first wash bath an enzymic desizing agent.

After washing, the textile is dewatered by spinning or suction and then dried. The final finish is advantageously applied by padding, followed by tentering. The padding liquor might, for example, contain 5 – 10 g/L of softening agent at pH 5 (citric or acetic acid). The fabric would then be dried at 100 – 110 °C on a tentering frame and subsequently pressed.

The feel of polyacrylonitrile textiles can be influenced in any desired direction by appropriate softening agents.

6.4.6. Discharges and Resists [105, 117]

Discharge printing methods are of some importance for polyacrylonitrile fabric intended for use in clothing. Polyacrylonitrile is usually printed in the form of a knit. This means that resist methods are inapplicable because they entail a preliminary padding process. In practice, especially in Italy, color-discharge printing is carried out mainly on dyed or printed grounds with the aid of tin(II) salts.

White discharge effects can be produced with either Decroline or tin(II) chloride. The resulting effects are subject to brown discoloration on storage, however, hence these methods are of no practical significance. The alternative is over-printing with a white pigment.

Color-discharge printing also takes advantage of tin(II) chloride. The fabric is in this case dyed with selected cationic dyes at the lowest possible temperature (< 100 °C). Printing pastes for color-discharge printing contain selected discharge-resistant cationic dyes, sometimes in conjunction with metal- complex dyes, as well as 40 – 80 g of tin(II) chloride per kilogram of printing paste. The amount of tin salt used depends on the color intensity of the ground that is to be discharged.

6.4.7. Other Classes of Dyes

Apart from cationic dyes, selected disperse dyes are also used for printing polyacrylonitrile. Thus, sharp pattern outlines can be obtained by mixing an anionically dispersed disperse black with a cationic black. It is also sometimes preferable to use select disperse dyes for bright colors, since these are subject to better leveling. Metal-complex dyes can be used if there is a need for especially light, pale colors with very good lightfastness.

Some types of polyacrylonitrile are modified to make them suitable specifically for dyeing with acid and metal- complex dyes.

6.4.8. Blends with Other Fibers

Polyacrylonitrile – Cellulose. Blends of polyacrylonitrile fibers with cellulosic fibers have been the subject of experimental printing tests with combinations of cationic and reactive dyes. However, fixation by steaming under acidic conditions followed by an “alkali-shock” treatment for the reactive dyes leads to a risk of yellowing of the polyacrylonitrile.

Polyacrylonitrile – Wool. Wool blends are printed with selected metal-complex and cationic dyes. The fabric must afterward be very carefully washed to ensure adequate wetfastness. For this reason, interest has developed in mixtures containing grades of polyacrylonitrile that can be acid dyed, since it is easier to print when only one class of dye is involved.

Polyacrylonitrile Blended with Other Synthetic Fibers. Apart from blends with polyester (Sec. 6.5.8), only the blends with PVC and acrylonitrile – vinyl chloride copolymers (Modacryl fibers) are of interest, specifically because of their flame-resistant properties. These are printed with selected disperse and cationic dyes.

6.5. Polyester

The introduction of HT steamers first made it possible for disperse dyes to be fixed to polyester fibers in a continuous process. As a result, polyester rapidly became the most important synthetic fiber for textile printing. Between 1975 and 1990, ca. 2×10^9 m^2 (300 000 t) were printed annually, ca. 25 – 35 % by transfer printing and the remainder by the classical direct, discharge, and resist techniques. Nearly all these prints go to the clothing sector, with minor amounts for curtains, upholstery, and decorative fabrics.

Polyethylene terephthalate accounts for well over 90 % of the market; other printed polyester fabrics include polydimethylolcyclohexane terephthalate and fibers modified by the addition of anionic compounds to permit their use with cationic dyes.

6.5.1. Pretreatment

Woven Fabrics [118, 119]. These are produced from staple-fiber or continuous-filament yarns. The sizing agents employed are water-soluble, so enzymes are not required for desizing. Precleaning therefore consists only of washing with nonionic or (better) anionic detergents (Diadavin or Levapon TH) at 40 – 80 °C in the presence of sodium carbonate or trisodium phosphate. Higher temperatures may lead to creasing. If nonionic detergents are used, it is important that their cloud point not be exceeded; otherwise, detergent becomes concentrated at the fiber surface, leading to serious printing problems. Colorfastness under conditions of use may also be impaired.

Knits. Unlike woven materials, knits are usually produced from continuous-filament yarn, which might be either smooth or texturized. To facilitate processing on the knitting machines such yarns are treated with preparations based on paraffinic substances melting in the range 40 – 80 °C. These must be removed completely prior to printing. The residual moisture content should not exceed 0.2 % on fabric ready for printing. The precleaning technique resembles that used for woven fabrics, but with the addition of fat solvents, usually chlorinated hydrocarbons.

Care must be exercised in all high-temperature wash processes to avoid excessive longitudinal stress on the fabric, since this can encourage crease formation. The problem can be combated with the aid of softening agents. These comments apply as well to polyester microfilaments (i.e., fibers with a fineness < 1 dtex or 0.1 g/1000 m) and articles made from them.

Heat Setting of Woven and Knitted Goods. Achieving dimensional stability and crease-resistance requires that woven or knitted goods be set by hot air on a tentering frame subsequent to precleaning. Drying and setting can be combined, but the fabric must be completely dry before entering the setting zone. Setting should be accomplished at a temperature ca. 10 – 20 °C above the highest temperature to which the fabric will be exposed during dye fixation or other aftertreatments. The usual setting conditions for woven fabrics are 20 – 30 s at 190 – 210 °C. For articles made of texturized yarn the appropriate setting temperature is somewhat lower: ca. 150 – 180 °C. Because of anticipated shrinkage – usually 5 – 7 % in both the warp and weft directions – the fabric is stretched to ca. 5 – 7 % less than its wet width, and a corresponding amount of overfeed is used. Once it leaves the hot-air zone the fabric must be thoroughly cooled.

The dye-absorption properties of polyester are influenced by this hot-air treatment, with dye uptake by nontexturized materials increasing as setting temperature increases. This effect is not observed with goods made of texturized yarn,

however, because they are exposed to high temperature during texturizing.

If the fabric is later to be embossed it is preferable to use a setting temperature ca. 20 – 30 °C lower.

Alkali Treatment. Better fabric feel and increased elasticity can be obtained by an alkali treatment with a 2 – 3 % solution of sodium carbonate for 15 – 20 min at the boiling point. Depending on the delicacy of the fabric, which is usually made from filament yarn, the process is conducted on a winch-back machine, a jigger, or a star-dyeing device. This treatment causes polyester molecules on the fiber surface to be hydrolytically cleaved and partially eliminated. The resulting loss of material loosens the weave structure and produces a soft, silky feel.

The optimal duration and temperature for alkali treatment depend on the desired effect. The weight loss is normally 3 – 6 %. After alkali treatment the fabric is acidified with acetic acid, rinsed, and dried.

6.5.2. Printing Processes

Woven or knitted fabrics are printed by screen processes on tables or with flat or rotary machines, as well as by the roller-printing method.

The fabric can be glued either to a back gray or directly to the rubber blanket depending upon the fineness of the material and the intricacy of the proposed design.

As with cellulose fibers, use of a back gray is uncommon with roller printing machines.

In the screen-printing process, heavy-grade textiles (e.g., decorative materials and jersey) are also glued directly to the rubber blanket without a back gray, but knitted goods sometimes cause major problems, especially with flat-film printing. Experience shows that thermoplastic adhesives give the best results, consisting for example of poly(vinyl acetate), poly(vinyl chloride), and an organic solvent. They are used principally with screen-printing machines (flat or rotary). In every case it is important that the adhesive be applied as uniformly and as thinly as possible to prevent mechanical resist effects with the printing paste.

If a back gray seems unavoidable, as is often true with fine textiles made of polyester filament yarn, the fabric is bonded to the cotton back cloth with gum – Cordofan (1 : 1) and/or dextrin – water 1 : 1, or with 20 % poly(vinyl alcohol). Only machines equipped with a drying cylinder are suitable for gluing, because immediate drying is essential to ensure adequate strength in the adhesive bond between the polyester and the cotton back gray.

6.5.3. Dyes and Printing Pastes

Only disperse dyes are used for direct printing on unmodified polyester, although cationic dyes are applicable to anionically modified polyester (Sec. 6.5.7).

Disperse Dyes [120–123]. A choice based on several criteria is necessary from among the large number of available dyes when printing on polyester.

Lightfastness. The tolerance range is wider for printing than it is for solution dyeing, because colorations based on deep navy blues, clarets, and greens always include lighter nuances as well. For decorative articles in pale colors a light-fastness value of 4 – 5 on the German fastness scale is permitted, and for women's outer garments a value of 3.

Wetfastness. Good wet-fastness properties can be obtained with disperse dyes on polyester if recommended printing and aftertreatment instructions are observed. The strongly hydrophobic character of the fiber prevents penetration of wash liquors and perspiration into the fibers, thereby preventing removal of dye. A significant deterioration in wetfastness is observed especially toward perspiration and water if excessive amounts of nonionic products are utilized in the form of detergents, carriers, fixing accelerators, softening agents, or antistatic agents, because they tend to remain on the fiber surface. Heat treatment > 140 °C then causes dye to migrate to the fiber surfaces on which these water-soluble nonionic products have become concentrated (thermomigration). Subsequent testing of fastness toward water or perspiration often shows a decrease of ca. 2 – 3 fastness units. An especially critical – but altogether justified – test is a comparison against polyamide present as a companion fiber, because polyamide accepts disperse

dye considerably more readily than polyester at low temperatures.

Fastness toward Processing. Whereas dyes are today usually fixed by HT steam at 165–180 °C with a steaming time of 8–6 min, it was once common to use high-pressure steam (2.5 bar, 125 °C, 20–30 min). In such a saturated steam atmosphere, azoic dyes subject to reduction may undergo color changes. Addition of oxidizing agents does not always ensure protection from this effect.

In order to prevent discoloration of a white ground it is important that there be no dye sublimation during dye fixation with HT steam or thermosol treatment (40–60 s, 190–210 °C). This is one of the important criteria by which the dyes should be selected.

Another important criterion relates to the behavior of a dye in the final wash. The only suitable dyes are those that show no affinity for the fiber under washing conditions, or are destroyed in the wash bath by the reductive washing process.

Applicable dyes are marketed under the trade names Dispersol, Foron, Palanil, Resolin, Terasil, and Samaron. Fabrics made from polyester microfilament do not require increased dye concentrations (in contrast to the situation in vat dyeing), because the unique characteristics of the fabric result in a higher concentration of printed dye at the fabric surface. However, the achievable light-fastness and wet-fastness properties are in this case inferior by ca. 1 fastness unit.

Printing Pastes. The requirements with respect to stability, adhesive properties, elasticity, and ease of wash removal cannot be met by one thickening agent alone. Obtaining printed products with optimal pattern sharpness requires the use of high-solids thickeners such as starch ethers and dextrin (British gum), but thickeners of this type lead to a brittle film that can easily break and flake. On the other hand, pastes based on low-solids thickeners, such as low-viscosity alginates or seed-flour derivatives, produce an elastic film that is easily removed by washing. Unfortunately, pastes of this type tend to run, especially with high-pressure steam fixing on a finely woven fabric. In spite of this, they are the most commonly used pastes for polyester printing, sometimes in combination with starch and cellulose ethers.

In addition to the thickening agent, the thickener solution may contain such auxiliaries as defoamers and/or printing oils to improve the performance of doctor blades or squeegees. It is more important, however, to add a nonvolatile acid or acid donor. Many disperse dyes are susceptible to chemical change in the presence of alkaline thickeners, especially in the course of high-pressure fixing. This can be countered by adding monosodium phosphate, which has the advantage over a number of organic acids that it has little tendency to attack nickel screens. It is also more compatible with alginates than is an organic acid. In the case of other types of dyes the reductive effect of steam or thickener can lead to color variation, especially with light shades. Here the addition to each kilogram of printing paste of ca. 5 g of an oxidizing agent such as sodium chlorate or sodium nitrobenzenesulfonate is indicated.

Following the successful use of *synthetic thickening agents* in pigment printing based on ethylene–maleic anhydride polymers and poly(acrylic acid), their use in other sectors of textile printing has been investigated. Promising formulations are so far known only for printing disperse dyes, apart from special applications in carpet printing involving acid dyes [124–129]. It has since been shown that their use is limited to exceptional situations, however, especially because solvent aftertreatment has proven to be a failure.

6.5.4. Fixation [130]

Because polyester fibers are strongly hydrophobic they have considerably less swelling capacity than, for example, cellulose. Therefore, a certain amount of effort is necessary to open the fibers to such an extent that dye is able to diffuse into them sufficiently. Appropriate measures include steaming at elevated pressure, steaming with superheated steam, setting with dry heat at ca. 200 °C, and the use of suitable fixation auxiliaries.

The dyeing and fixation of disperse dyes on polyester are purely physical solution processes, where the fiber represents a "solid" solvent for the dye. The capacity of polyester fibers to ab-

sorb dye is highly temperature dependent, and is subject to influence by added carriers, also designated in this case as *fixation accelerators* (→ Textile Auxiliaries, Chap. 6.2.5.). However, such chemicals may have a negative influence on the fastness properties of certain dyes in the finished print. The specific auxiliaries required are a function of the fixation process.

Fixation with Saturated Steam (30 min, 102 °C). Under normal steaming conditions in the absence of carriers, only edge-zone dyeing is obtained, with light to medium color intensities and maximum dye yields of 20 – 50 %.

Because of the deficiencies associated with this fixation process it is of virtually no significance, even with the use of specially selected dyes and large amounts of carrier. For polyester printing it can therefore be regarded as a solution only in an emergency.

Fixation with High-Pressure Saturated Steam (ca. 30 min, 2.5 bar). The dye-absorption capacity of polyester is significantly enhanced by fixation in a kettle steamer under pressure. The increase is proportional to the steam pressure, reaching a color intensity corresponding to essentially 100 % dye fixation at 3.5 – 4.0 bar, although by this point the dyes begin to sublime. The high dye yields obtainable at high steam pressures cannot be duplicated by longer steaming times at lower pressures. Very good results are achieved at 2.5 – 3.0 bar with a steaming time of 20 – 30 min. At 2.5 bar without addition of a carrier the dye yield is 60 – 90 % depending on the dye used. Addition of a carrier increases the yield only slightly, and is therefore not justified.

Fixation by HT Steam [131–134] (7 min, 175 °C). This is becoming an increasingly important method because of the growing use of hanging-loop steamers, and it can be regarded as the best fixation process for printing on polyester.

The process is conducted with continuously operating high-temperature hanging-loop steamers and superheated steam at 170 – 180 °C. These steamers differ from thermofixing facilities (e.g., tentering frames) in their high capacity (200 – 400 m), permitting exposure of the fabric to the requisite high temperatures for 5 – 10 min at maximum throughput.

Because steam is used as the heating medium instead of air, the film or thickener swells at the start of the heating phase (20 – 100 °C), with the water evaporating again on further heating to 170 – 180 °C. As a result, thickeners are less likely to be burned in, the fabric remains softer, and dyes can diffuse more easily from the film into the fibers.

Such a high fixation temperature means that only dyes very resistant to sublimation are applicable. In addition to sublimation or thermofixation stabilities (as established by the DIN method and quoted in pattern books), the sublimation resistance of a print in its unfixed state in steam must also be taken into account. Therefore, when changing to HT steamers, dyes previously used with a high-pressure steam process must be tested once again for fastness in order to avoid the production of prints that are defective due to sublimation onto a white ground. The risk of staining the ground depends on the fixation temperature, the fixation time, and the circulation pattern of the heat-transfer medium.

Thickeners with a high solids content are unsatisfactory for this process because they burn in and harden (albeit to a somewhat lesser extent than in the thermofixation process), and they are therefore difficult to remove by subsequent washing. Good results are obtained with mixtures of alginates and starch ethers in the ratio 4 : 1 to 3 : 1. Seed-flour products are also occasionally utilized, either alone or in combination with starch ethers. Pastes containing alginate thickeners should also contain 1 – 3 g/kg of a polyphosphate to facilitate removal by washing and to improve compatibility with acid donors and carriers.

With medium and dark colors the dye yield can be increased significantly by a fixation auxiliary (usually an ethoxylated substance), such as Levegal PEW or Samaron HT Fixer (see also → Textile Auxiliaries, Chap. 6.2.9.) [122, 133].

Thermosol Fixation (dry heat, 1 min at 200 °C). In addition to the discontinuous high-pressure steam-fixing process and the continuous HT process, hot-air fixation, which is also continuous, is of some importance. The print is in this case passed through a high-performance tentering frame or a condensation apparatus. The best results are obtained at 200 – 220 °C, but above 200 °C the disperse dyes involved must

be exceptionally resistant to sublimation, which greatly limits one's choice.

A corresponding dwell time must be determined for each hot-air temperature. With a high-performance tentering frame this time is typically 40 – 50 s. At 200 °C, the mean dye yield without a carrier is 50 – 70 %. Color intensity can be increased considerably by adding Levegal PEW. In general, fixation by dry heat, because of the high temperature, is suitable only for woven fabrics made of untexturized yarn.

6.5.5. Aftertreatment

Washing [135]. After printing with disperse dyes and fixing by one of the processes described above, the fabric is thoroughly rinsed (warm or cold) and then treated with an alkaline reducing agent (2 g/L of NaOH, 2 g/L of hydrosulfite, and 1 g/L of Levegal HTN). At a bath temperature of 40 – 50 °C, most of the thickener and unfixed dye is either dissolved or reductively decomposed. Because of the low temperature, neither dye nor decomposition products are absorbed onto a white ground.

The first wash bath is followed, sometimes after an intermediate rinse, by a second and third wash at 70 – 80 °C with lower concentrations of chemicals. These are in turn followed by warm and cold rinses under acidic conditions, and drying at 110 – 130 °C under very low tension.

Finishing with Fabric Conditioners. These agents further improve the feel, volume, and softness of what is essentially already an "easy-care" and crease-resistant fabric. Treatment with the appropriate substances is possible either in the last rinsing bath or by padding prior to final treatment on a tentering frame.

Antielectrostatic Finishing. A certain amount of antielectrostatic effect is conferred by fabric conditioners, although it is not permanent. To achieve a lasting antielectrostatic finish that does not produce a softening effect one can apply 5 – 15 g/L of Statexan PAN (an ester of phosphoric acid) to the fabric by padding. Statexan PAN and Persoftal SWA (which contains silicone) are often applied in combination. For more information on antielectrostatic agents.

6.5.6. Discharges and Resists [105, 117, 136, 137]

Polyester is increasingly being printed by the discharge and resist methods in cases where the patterns have very fine designs or registration problems arise due to short pattern repeats. Another motivation is the desire for fashionable coloristic effects that would be difficult to achieve by direct printing. Three methods are available:

1) True discharge printing on a dyed ground
2) Discharge resist on an unfixed ground
3) Ordinary resist on an unfixed ground.

In general, discharge and resist processes always require much stricter production control and more process development than direct printing. In each case, the optimum formulation and operating procedure must be established through preliminary tests carried out by the printer. It is especially important to determine the proper quantities of reducing or resist agents relative to the ground color intensity and the amount of dye applied.

True Discharge Printing on a Dyed Ground. If Decrolin or tin(II) chloride is used with a hydroxydiphenyl carrier as the discharge intensifier (Tanalid HP), the result is a white effect that at first appears acceptable. However, because of the hydrophobic nature of polyester, the discharge process usually attacks only the surface of the dyed fabric. Light falling on dye or chemical residues remaining in the fibers and not removable by washing leads to yellowing over the course of time. Moreover, dye diffuses from the interior of the fibers to the surface, resulting in a color change, especially with dark ground colors. For these reasons, neither white discharging nor color discharging can be recommended.

Discharge Resists with Tin(II) Chloride. In the discharge-resist process, the ground that is to be discharged is applied by padding or printing, and then simply dried. The temperature of the fabric must not be allowed to rise above 90 °C to prevent premature fixation of the dye.

In a second step, paste and a discharge agent are printed onto the dried ground, after which

the ground and printed pattern are fixed simultaneously in a third step.

The discharge resist/tin(II) chloride combination is used mainly for coloristic reasons. On one hand it provides access to a wide range of dischargeable dyes, but at the same time there is also a broad choice of printing dyes that display good resistance to tin(II) chloride.

The amount of reducing agent required is a function of color intensity and the fixing process. In the two-stage fixing method (presteaming at 100 °C followed by poststeaming for 6 – 7 min at 170 °C), 60 – 100 g of tin salt – water 1 : 1 is usually sufficient, whereas the single-stage process (6 – 10 min of HT steam at 170 °C) requires not only the presence of a polyglycol in the discharge paste but also approximately twice as much tin salt.

One disadvantage of tin(II) chloride is its corrosive effect, attributable to hydrochloric acid liberated during fixing but also present in the paste.

A high-viscosity thickener should be used for printing the dischargeable ground so that the printed surface will remain as elastic as possible after drying.

Fixing processes are described below.

Discharge Resists with Decrolin or Rongalit DS. The same process steps are involved with these products as with tin(II) chloride. They offer the advantage of no risk of corrosion. However, one disadvantage is modest stability and reproducibility of the printing dyes. Especially in the red range, light shades of disperse dyes are less stable toward Decrolin and Rongalit DS than toward tin(II) chloride.

Dye in the dischargeable ground is not yet fixed to the fiber, so a relatively small amount of Decrolin (30 – 50 g/kg) is sufficient to cause its decomposition. On the other hand, there are a number of disperse dyes that are stable toward such quantities of Decrolin if their concentration is sufficiently high. Light- colored dilute dyes are not particularly stable even toward small amounts of Decrolin. Since the amount of reducing agent required depends on the intensity of the ground, it may under some circumstances happen that the concentration of Decrolin normally recommended (50 g/kg) will prove insufficient, depending on the fabric and the dye that is to be discharged. However, increasing this concentration severely limits the number of Decrolin-resistant disperse dyes, which are required for color printing. Therefore, in the interest of greater security, Rongalit DS or even tin(II) chloride is preferred in practice.

Methods of fixing discharge reserves with tin(II) chloride, Decrolin, or Rongalit DS include the following:

1) 10 min in a star steamer at 102 °C followed by 20 min at 2.5 bar
2) 20 min in a star steamer at 2.5 bar
3) 8 min in a hanging-loop steamer at 170 – 175 °C with HT steam
4) 8 min in a hanging-loop steamer with saturated steam at 102 °C, followed by 8 min in a hanging-loop steamer at 175 °C with HT steam.

Subsequent to the fixation process the fabric is rinsed and then washed at 60 – 80 °C in the presence of sodium hydroxide solution, hydrosulfite, and Levegal HTN.

Alkali Resists. The ground, consisting in this case of special disperse dyes that can be hydrolyzed by alkali, is first applied and then gently dried (maximum temperature 100 °C).

Onto this ground are printed the pastes, which contain sodium carbonate, sodium hydroxide or sodium silicate solution, and disperse dyes that offer good resistance toward alkali.

This process offers environmental advantages, and costs for both the dye and the auxiliaries are low. Disadvantages include sensitivity of the ground to unduly high drying and steaming temperatures, lack of sharpness of the printed pattern, and fiber damage resulting from the application of alkali to thin monofilament fabrics that may already have undergone an earlier alkali treatment.

Crimped Effects. These can be achieved with polyester by applying a printing paste containing ca. 400 g/kg of Tumescal PH (*p*-phenylphenol) and then fixing the printed product.

6.5.7. Direct Printing on Polyesters with an Affinity for Basic Dyes

Polyester fibers modified by the incorporation of anionic (acidic) groups have not so far achieved

great importance in textile printing. They can be dyed with basic (Astrazon) dyes.

Pretreatment before printing is analogous to that for ordinary polyesters, and the procedure for printing with basic dyes is no different from that for printing on polyacrylonitrile (Sec. 6.4.2).

Fixation. The dye yield depends on steaming conditions. Steaming is best conducted in a star steamer for 30 min at a maximum pressure of 1.5 bar, because higher pressures can have harmful effects on the dye. However, it may prove necessary to print disperse and basic dyes together in a single design. The disperse dye must then be steamed at ca. 2.0 bar to produce a better dye yield. In this case, dyes sensitive to reduction must absolutely be avoided. Addition to the printing paste of the fixing agent Luprintan PFD can be omitted.

HT fixation for 10 – 6 min at 150 – 175 °C gives unsatisfactory results.

Selected Astrazon dyes can also be fixed very successfully in unpressurized saturated steam (hanging-loop steamers).

Appropriate *thickeners* include seed-flour and guar products, either alone or in combination with such cold-soluble starch derivatives as Diatex SL or Printex S, or proprietary mixtures with this type of formulation (e.g., Diaprint 1024 P).

Aftertreatment of the fixed print is usually similar to that applicable to polyacrylonitrile; i.e., cold rinsing, washing at 60 °C with 1 g/L of Levapon TH and 1 g/L of hydrosulfite, and rinsing under acidic conditions.

6.5.8. Blends of Polyester with Other Fibers

Polyester – Cellulose. These are widespread in the form of mixed cotton blends, which are printed at a rate of 4×10^9 m^2/a. Blends with viscose staple fiber are produced in much smaller quantity, and are printed mainly with pigment dyes (see Chap. 8). Only if pigment-printed fabrics are unsatisfactory with respect to feel and rubbingfastness are alternative – and much more complicated – processes used, as described below [101, 138–142].

Mixtures of disperse and reactive dyes can be prepared by the printer himself, but it is better to use mixtures supplied by the dye manufacturers. Apart from dyes, the corresponding printing pastes contain urea, Ludigol (sodium nitrobenzenesulfonate), and a mild alkali (e.g., sodium hydrogencarbonate). Alginates must be used as the thickening agents.

The print is fixed by the thermosol process (1 min, 200 °C) or better with hot steam in a hanging-loop steamer (6 min, 175 °C).

Careful attention must be directed toward the final wash, especially if the design contains white effects next to large areas of dark ground, and for this reason special washing procedures and test methods have been developed. For example, the *acetone test* can quickly show whether unfixed dye has been completely removed from the textile. In this test a small piece of the printed and washed fabric is shaken with acetone in a test tube. Problem-free washing is indicated if the acetone remains virtually colorless.

Selected *disperse dyes* can be fixed on the cotton in the presence of swelling agents [143]. The dyes in this case must be soluble in the swelling agents (usually polyglycols or their esters and ethers), at least at elevated temperature. The swelling agents are added to the printing pastes at the rate of 10 – 15 %, and their role is to displace water from the micelles of the cotton during the drying and fixing process. If the swelling agents are also solvents for the dye, solubilized dye is transported simultaneously into the interior of the cotton fibers. If the solvent partially volatilizes at the fixing temperature (> 200 °C), the dye is precipitated. Precipitation in the cotton increases during cooling after fixation, and is essentially complete after dilution of the swelling agent and solvent in the final washing process.

Because disperse dyes also dye the polyester fibers, both types of fiber can be dyed with a single dye group using this method. This approach was introduced by DuPont in 1970 as the Dybln process. The Cellestren dyes of BASF work on the same principle, but both have in the meantime been superceded.

Burn-out Textiles. Whereas discharges and resists on polyester – cotton are still in the laboratory stage despite much research, the so- called burn-out process has already achieved considerable importance. The process is based on the fact that cellulose fibers can be destroyed by heating with strong acids or their salts.

Core-spun yarns (i.e., polyester filament yarns that have been surrounded during the spin-

ning process by cotton or viscose staple fiber) are printed with a printing paste that contains strongly acidic metal salts and disperse dyes. They are then subjected to a heat treatment (carbonization) that destroys the cellulose fibers and at the same time fixes the disperse dyes on the polyester component.

The carbonized cellulose is removed by a combination aftertreatment consisting of mechanical action followed by washing, or simply an extremely intense wash.

The dye yield of the selected disperse dyes from this strongly acidic paste is always lower than from pastes free of sulfuric acid salts, and the loss when the salt is aluminum sulfate is greater than with sodium hydrogensulfate.

An important variant of this technique is used in the production of Indian *saris*. Blended yarns of polyester and cotton are woven into fabric and dyed or printed with disperse dyes. The cloth is then treated with 70 % sulfuric acid to remove the cotton component, leaving a typical sheer sari fabric.

Polyester – Wool. These blends are printed with combinations of disperse dyes with selected acid and metal- complex dyes [144]. Fixation is carried out with HT steamers (8 min, 165 °C), although higher dye yields are possible through two-stage fixation with saturated steam (30 min, 100 °C) followed by HT steam (8 min, 165 °C).

Polyester Blended with Other Synthetic Fibers. Blends of this type put in an appearance from time to time, although they have so far not achieved commercial importance. Blends of polyester and polyacrylonitrile fibers are perhaps worth special mention. These are printed with select disperse dyes, either alone or in combination with cationic dyes.

6.6. Other Synthetics

This group accounts for < 1 % of printed textiles.

Polypropylene Fiber. Base-modified polypropylene, which is printed with selected acid and metal- complex dyes, and nickel- containing polypropylene have so far attracted only passing interest in the carpet-printing sector. In the case of the nickel-modified fibers, the dye forms a complex with nickel located both on and in the fibers, thus producing a true dyeing effect [145, 146].

Other proposed methods of printing polypropylene (e.g., through prior chlorination of the fiber) have not led to practical applications [147].

Polyurethane fibers are processed together with polyamide in the production of swimwear material. Printing methods are described in Section 6.3.8.

Special polyamide fibers with high strength and good flame resistance have been introduced to the market under the name "Nomex" by DuPont. These have hitherto proven difficult to dye, but new carrier-free grades that are subject to dyeing are now of interest for military clothing and tents (camouflage).

7. Printing on Protein Fibers

The only protein fibers of importance from the standpoint of printing are wool and silk, termed "animal" fibers because of their origin, (→ Silk, → Wool). These two fibers, together with cotton, formed almost the sole basis for clothing in the past. Today, competition from the comparatively inexpensive synthetic fibers (which unlike natural fibers are also continuously available, and with a consistent level of quality) is the main reason for the steadily decreasing importance of natural products as a fraction of total fiber consumption. Nevertheless, the quality of the natural materials cannot be equalled in many respects, and they are assured of a secure future, especially for high-quality articles.

In the case of wool, its polypeptide backbones bear long side chains, which are very tangled, whereas with silk the polypeptide chains are extended and have only short side chains. Another important characteristic of wool is the fact that the polypeptide chains are linked by the sulfur bridges of cystine.

The simultaneous presence of amino, hydroxyl, and carboxyl groups makes it possible for wool and silk to be dyed by acid, basic, reactive, and the usual substantive dyes. Metal-complex formation is also possible (preferably with chromium, but also copper) in which both principal and secondary valencies of the fibers and dyes play a role.

Several other characteristics are equally important with respect to dyeing and printing of protein fibers. Wool is very sensitive to alkali, but very stable toward dilute acid. On the other hand, silk tends to be more sensitive to acid and less subject to damage by dilute alkali. Both protein fibers are susceptible to abrasive damage, and both display a yellowing tendency above 120 – 130 °C. High temperature leads to deformation, embrittlement, and irreversible impairment of swelling, and hence of dyeing potential. The felting tendency of wool necessitates special attention before and after printing, or the provision of a special antifelting finish.

7.1. Pretreatment

For information on the washing, desizing, crabbing ("burning in"), and bleaching of wool → Textile Auxiliaries, Chap. 4.3.; → Wool. *Wool chlorination* and the *degumming of silk* are especially important processes with respect to printing.

Chlorination of Wool. The acid chlorination of wool, formerly carried out only with dilute hydrochloric or sulfuric acid (first bath), followed by treatment with chlorinated lime or sodium hypochlorite (second bath), alters the scaly surface of the wool fiber such that the material is more easily wetted and exhibits better swelling properties so that it will in turn absorb more dye. In the process, the scales are caused to lie flat on the hard outer layer. The wool also tends not to felt so severely, and it acquires a slight luster. It is extremely difficult to achieve completely uniform chlorination, however, and there is always the risk that "chlorination marks" will form. For this reason two improved processes were developed. In the first, nitrogen- containing auxiliaries are added to the chlorination bath. These combine with the liberated chlorine to form chloramines that react with the wool slowly and gradually (e.g., the Melafix process of Ciba-Geigy). In the second process the chlorinating agent is the less reactive dichloroisocyanuric acid [*2782-57-2*] or its potassium [*2244-21-5*] or sodium [*2893-78-9*] salt (e.g., Basolan DC, BASF) (→ Chloroamines).

All the processes described, correctly implemented, produce such a dramatic improvement in the color intensity and brilliance of printed wool – irrespective of the dye – that almost all printed woolen piece goods or knits are subjected to chlorination.

Antifelting Finish. The creation of a nonfelting finish (e.g., by the IWS-Hercosett process) so that wool can be washed in a washing machine is almost always carried out on the combed sliver. This produces on the chlorinated fibers a very thin film of resin that has no detrimental effect on typical wool properties but ensures simultaneously both good dye-absorption capacity and nonfelting behavior. Yarn, woven fabrics, and knitted goods produced from wool of this type are printed preferentially with very colorfast reactive dyes to satisfy the demand for good washing properties along with colorfastness.

Degumming of Natural Silk. The double fiber emerging from a silk spinneret consists of fibroin covered with a second secretion known as *sericin* (*silk gum*). Woven fabrics produced from silk of this type are matt, rough, and harsh to the touch. They are therefore usually subjected to degumming; i.e., they are carefully boiled with a 4 % solution of hard soap either as piece goods in a star-dyeing machine or, more often, in a tension-free full-width washing machine or at an earlier stage in rope form (→ Silk, Chap. 4.1.). The resulting colloidal solution of sericin in the wash liquor, known as "degumming liquor," is used in silk dyeing as an excellent protective colloid and levelling agent. The following alternative degumming method is also effective: a solution consisting of 5 g/L of a nonylphenolpolyglycol ether (e.g., Levapon TH, Bayer), 3 g/L of hexametaphosphate, and 2 g/L of sodium carbonate (soda ash) is used to prepare both a degumming bath and a second processing bath. The fabric is treated for 1 h at 90 – 95 °C in each of these baths. After rinsing and drying, the material retains its silky luster, displays good dye-absorption properties, and can in principle be printed exactly like wool.

7.2. Direct Printing

7.2.1. Printing with Acid Dyes

In principle, both basic and substantive dyes can be used for printing on wool and silk, but

in practice, acid dyes are mainly used, as the others, though often giving excellent brilliance, have unsatisfactory fastness properties. Even with acid dyes it is necessary to restrict the selection to substances that combine brilliant results with acceptable lightfastness and wetfastness properties with respect to the articles in question.

Printing Paste. Dissolving the dye requires the use of urea or thiourea, solubilizing agents such as thiodiethylene glycol, and hot water. In rare cases, other solvents (e.g., polyglycol ethers) or dispersing agents are added to prevent thickening of the printing paste on prolonged storage and the possibility of stippled printed effects. Glycerine, which is hygroscopic, is recommended only for wool printing, where it is able to compensate for the harmful effects of slightly superheated steam introduced during fixing, whereas with silk, significant prolongation of drying brings the risk of smearing.

Thickening agents used today include seed-flour derivatives (mainly guar products) either alone or in combination with cold-soluble British gum, as opposed to the formerly widely used tragacanth and its mixtures with British gum and/or gum arabic. High-solids thickeners are preferred for fine effects, sharp pattern contours, etc., whereas low-solids products with their better leveling properties and reduced risk of fold and crease formation are preferred for larger printed areas. One of the most important prerequisites for the viability of a thickening agent is ease of removal by washing. The printing pastes also contain an acid donor to promote dye fixation (e.g., ammonium sulfate, tartrate, or oxalate), sometimes also acetic or glycolic acid, and small amounts of sodium chlorate to counteract the reductive action of wool and certain thickening agents in the course of steaming (occasionally essential with "steam sensitive" dyes!). Defoamers and printing oils are usually also necessary for producing smooth, sharp prints.

Wool that has been poorly or inadequately chlorinated or is extremely voluminous may profit from the use of a coacervate former (e.g., Levalin VKUN, Bayer, an alkylarylpolyglycol ether) to prevent "frosting effects." This auxiliary causes dye to be absorbed uniformly during steaming, even at the outer tips of the fibers. If silicone defoamers are introduced simultaneously this effect is negated, therefore high-boiling alcohols are used as defoamers instead (ca. 5 g/kg).

Fixation. Fixation of the dye on wool or silk usually requires relatively prolonged steaming (i.e., 30 – 60 min), and the most brilliant and colorfast prints are obtained only with saturated-steam fixation at 100 – 102 °C. Excessive drying of the fabric after printing, especially in the case of wool, must therefore be strictly avoided. For this reason wool is sometimes sprayed prior to steaming to ensure that the steam is saturated.

Aftertreatment. The final washing process is carried out on winch-back equipment, although one of the various types of tricot washing machine can also be used, depending on the batch size. Just like silk, wool should be treated as carefully as possible – i.e., with minimum mechanical stress and at a maximum temperature of 40 °C (0.5 – 1 g/L of an anionic synthetic detergent). To prevent bleeding into unprinted areas and the formation of marks, products based on the condensation of high molecular mass aromatic sulfonic acids with formaldehyde may also be introduced (Erional NWS, Ciba-Geigy; Mesitol NWS, Bayer, etc.). These lead to significant improvement in the wetfastness properties of wool if applied for ca. 20 min at the level of 5 – 6 % of the textile weight at 60 °C in a bath acidified with acetic or formic acid.

Similar improvements in fastness properties are produced by certain polyammonium compounds (e.g., Levogen, Bayer; Tinofix, Sandoz) if all anionic auxiliaries are thoroughly rinsed out after the final wash (to prevent the formation of spots) and the fabric is then treated in a bath containing 2 – 3 g/L of one of the auxiliaries for ca. 20 min at 35 – 40 °C.

These inexpensive and very effective processes unfortunately have the disadvantage that cationic aftertreatment may lead to reduced lightfastness, especially with pale colors, and color changes (dulling effects) may be detectable in rare cases.

7.2.2. Printing with Metal-Complex Dyes

Both 1 : 1 and 2 : 1 metal- complex dyes give a higher degree of colorfastness – especially lightfastness – than acid dyes, but they have the disadvantage of producing rather subdued, dull colors. There are almost no differences in printing applications between metal- complex and acid dyes, except that it is better to print the former in the absence of acids or acidic compounds. The acid pH range decreases the stability of the paste and encourages a tendency toward dye agglomeration (stippling effects); in addition, one also frequently obtains prints in which leveling of the printed dye is inferior. Leveling problems are in any case common with this type of dye in fullcovered design applications.

The dissolution process, the choice of paste-thickening agent, fixation, and the final wash are also identical for both types of dye. This leads to a great many combination possibilities, although printing must be accomplished without acids or acid donors. The use of auxiliaries in aftertreatment processes is also analogous to the situation with acid dyes.

7.2.3. Printing with Reactive Dyes

Unlike the dyes described previously, reactive dyes (→ Reactive Dyes) do not form salt-like bonds to protein fibers; instead, true chemical bonds are formed in an acidic medium (pH 3 – 5) at 80 – 100 °C with $-SH$, $=NH$, and $-NH_2$ groups from the polypeptide. This results not only in dyes with good fastness properties, but also very brilliant products, thereby significantly enriching the market. The high demand for goods that carry the "woolmark," together with the high serviceability associated with articles produced from these fibers, justify the additional costs entailed in use of these dyes.

From the point of view of printing techniques, there is again no difference between reactive dyes and acid dyes with respect to printing-paste formulation or thickening agents, except that formic acid is often included in the reactive-dye pastes to ensure acidity. In general, the steaming times for fixation can even be reduced – to 10 – 20 min (saturated steam at 100 – 102 °C) – but it should be noted in terms of the final wash that a high degree of fastness will be achieved only if dye that has not reacted with the fiber is removed completely. To achieve this end it is necessary after the rinsing bath (to which a small amount of a synthetic tanning agent such as Erional NWS or Mesitol is also normally added) that the fabric be washed in several baths with progressively increasing temperatures (40 °C, 60 °C, 80 °C) using 2 g/L of disodium phosphate (with ammonia added to a pH of 8) and an anionic detergent.

This has two consequences. The energetic washing process does not permit the concurrent use of any but reactive dyes in the color scheme, because these would be subject to serious bleeding. Also, this type of wet aftertreatment is possible only with wool that has received a good antifelting finish; i.e., it is essentially restricted to chlorinated or Hercosett-treated wool.

With silk, which is more resistant to alkali, an alternative procedure is available at alkaline pH with sodium hydrogencarbonate and alginate thickeners. The advantage of this approach is that it often gives a more brilliant result with a better dye yield. The disadvantage is a sometimes inferior print with designs that overlap. No differences are apparent in fastness. (Very detailed patterns can in any case be obtained with pigment dyes without affecting the feel of the fabric.)

7.2.4. Printing with Chrome Dyes

The conversion of chrome dyes (→ Azo Dyes, Chap. 3.2.1.3.) into chromium complexes during steaming leads to very colorfast prints. However, this technique is restricted almost exclusively to melange printing (Sec. 3.6.2), because the range of usable dyes is limited, and the printed colors on clothing materials are not very attractive. Chromium(III) salts (usually chromium fluoride, more rarely the acetate) are added to the printing paste together with formic acid (occasionally oxalic acid), which has a delaying effect on complex formation, thereby preventing premature formation of the chromium pigment. This technique is being used less and less frequently, however, largely for environmental reasons.

After printing, the sliver is steamed for 60 – 90 min without intermediate drying, after which it is rinsed, washed, dried, and combed.

7.3. Discharge Printing

Although the discharge-printing technique on cellulosic materials was almost completely forgotten for almost 15 years, wool and silk for high-fashion articles have always been dyed (albeit in small quantity) with dischargeable substantive or acid dyes and then printed with discharge-resistant acid or substantive dyes. The usual discharging agents are sodium or zinc formaldehyde sulfoxylate (Rongalit C or Decrolin, BASF). A wide range of additives is used in the interest of achieving more or less perfect results. The production of these discharge-dyed articles is difficult, because success depends on many conditions, which are often subject to change. The fabric, its structure, the pretreatment, the degree of whiteness, the color intensity of the ground, the particular combination of dyes, the discharge formulation, the type and amount of paste applied, and the steaming and final wash conditions – all these contribute to the result (as measured, for example, by fastness properties, brilliance, and yellowing tendency). For this reason, only general information can be provided with respect to formulations, which must be adjusted from case to case to suit local conditions.

Special attention must be directed toward the risk of fiber damage through attack on the protein fibers by decomposition products from the reducing agents. Thiourea dioxide (formamidinesulfonic acid) is a mild reducing agent, but its use as a discharging agent on protein fibers has the disadvantage that its poor solubility limits the choice of dischargeable dyes. It is therefore applicable only in special cases.

To ensure good fastness properties, especially with reactive dyes, the dyed fabric must be thoroughly washed before printing. This is also beneficial with respect to the desired discharge effect, and it facilitates the final washing stage. Dischargeable grounds on silk should display a neutral reaction; dye effects achieved with the aid of added acid require that the fabric be neutralized with ammonia and rinsed very carefully.

Color-Discharge Printing. The demand for efficiency has led to a reduction in the choice of dyes offered by manufacturers, and the number of dyes available for special applications is decreasing steadily. For this reason it has become increasingly difficult to put together a satisfactory palette of discharge-resistant dyes for wool and silk.

In principle, the following approaches are possible:

1) Color-discharge printing with vat dyes according to the sulfoxylate procedure, using dyes that can be fixed with a small amount of alkali, which means accepting a certain amount of unavoidable fiber damage. This path is only very rarely taken.
2) Color-discharge printing with basic dyes, using tin(II) chloride as the reducing agent. In spite of a high degree of brilliance and an adequate selection of dischargeable grounds, this method is seldom if ever used because of poor fastness to light and moisture.
3) Color-discharge printing with selected discharge-resistant pigment dyes using soft binders. It is recommended that suitable softening agents be added, with zinc formaldehyde sulfoxylate as the reducing agent. Otherwise, fixation may be unsatisfactory, and the differing swelling properties of protein fibers and binder film may lead to extremely poor rubbingfastness.
4) Color-discharge printing with selected discharge-resistant substantive and acid dyes, where sodium formaldehyde sulfoxylate or zinc formaldehyde sulfoxylate is the discharging agent. Although the choice of dyes in this case is relatively limited, it has proven to be the best method.

7.4. Printing on Blends Containing Protein Fibers

Printing of blends of wool with silk, or of one of the two with cellulose, has become a peripheral activity since the emergence of blends with synthetic fibers, although even these are not of great importance in textile printing (see Chap. 6).

Wool – Silk. The wool fraction is pretreated like wool (Sec. 7.1), though without chlorination, and the silk fraction is degummed (if necessary) at 90 °C for 1 h in one or two mildly alkaline baths containing hard soap (5 g/L) as well as hexametaphosphate (2 g/L) and disodium phosphate (3 g/L). Silk is sensitive to chlorine, so any

bleaching must be accomplished with hydrogen peroxide. The choice of dyes and the formulations are exactly as with pure woven wool products. This is also true for aftertreatment of the print and the production of discharge-printed articles.

Wool – Cellulose. Blends of wool with cellulose ("half-wool") are pretreated according to the nature and ratio of the blend, always bearing in mind the sensitivity of wool to alkali. Chlorination can normally be a consideration only if wool is the predominant fiber. Here, too, hydrogen peroxide is the preferred bleaching agent.

Most of the dyes used are substantive, supplemented by certain neutral-dyeing acid and metal-complex dyes. The formulation again corresponds to that recommended for printing on wool alone; additives intended to improve fastness (e.g., cationic quaternary polyammonium compounds) must not be omitted from the aftertreatment process. In the case of half-wool products various special aftertreatment agents have been developed.

Textiles printed with selected reactive dyes display a higher degree of fastness. In most cases these can also be fixed using shorter steaming times. It is essential, however, that the formulations contain only ca. 10 g/kg of sodium hydrogencarbonate, and the two fibers must be intimately mixed in the fabric structure; otherwise, fiber damage and crimping effects must be anticipated due to localized shrinkage.

White- and color-discharge printing can be carried out as described in Section 7.3, and if there is a high proportion of cellulose in the fabric it is also possible to use vat dyes that can be fixed with small amounts of alkali.

Silk – Cellulose. Grades known as "half-silk" are pretreated like pure silk and printed like half-wool. However, the lower sensitivity of silk to alkali means that direct or discharge printing with easily fixed vat dyes is more feasible. Otherwise, the same recommendations apply as those given for half-wool.

8. Pigment Printing

"Pigment printing" refers to producing printed textiles by the application of colored pigments with the aid of pigment binders, softening agents, flow modifiers, defoamers, and special thickeners.

The colored pigments consist mainly of organic products of the azoquinacridone, dioxazine, and phthalocyanine types, to mention only a few examples. These are supplemented by carbon black and inorganic white pigments, mainly titanium dioxide, and sometimes with iron oxides for brown shades.

All these pigments are used in textile printing in the form of 25 – 50 % dispersions reduced in size mechanically (by grinding) to the optimum particle-size distribution and mixed with various additives (emulsifiers, dispersing agents, evaporation inhibitors, and preservatives).

The resulting viscosity properties should be as stable as possible, and it is important to strive for a consistency that facilitates pumping and metering so that the preparations can be processed by automatic metering equipment. Special effects sometimes call for the use of copper and aluminum alloys in powdered form ("bronze printing"), basic dyes modified with synthetic resins ("luminescent pigments"), or mica that has been coated with titanium dioxide ("nacreous effects"). For further details → Pigments, Inorganic; → Pigments, Organic.

9. References

1. W. H. v. Kurrer: *Geschichte der Zeugdruckerei,* Nürnberg 1844.
2. R. Forrer: *Die Kunst des Zeugdrucks vom Mittelalter bis zur Empirezeit,* Straßburg 1898.
3. H. Züblin: *40 Jahre Kattundruck,* v. Berchtold & Gommeringer, Singen 1927.
4. A. Bollinger: *Ein Beitrag zur Entwicklung des europäischen Textildrucks,* Springer Verlag, Wien 1950.
5. E. Bindewald, K. Kasper: *Bunter Traum auf gewebtem Grund,* G. Westermann, Braunschweig 1950.
6. J. Zahn: *Das Abenteuer Farbe – Eine Kulturgeschichte,* Hoffmann und Campe, Hamburg 1983.
7. "Uralter Zeugdruck," *Bayer Farben Rev.* **1 – 8** (1962 – 1964).
8. L. W. C. Miles: *Textile Printing,* Dyers Company Publications Trust, London 1981.

General References

9. G. Georgievics, R. Haller, L. Lichtenstein: *Handbuch des Zeugdrucks,* **vols. I – III,** Akad. Verlagsgesellschaft, Leipzig 1930.
10. L. Diserens: *Fortschritte in der Anwendung der Farbstoffe,* Birkhäuser, Basel 1949.
11. H. Rath: *Lehrbuch der Textilchemie einschl. der textilchemischen Technologie,* 3rd ed., Springer Verlag, Berlin Göttingen Heidelberg New York 1972.
12. W. Bernard: *Appretur der Textilien (mit Waschmaschinen, Trocknungseinrichtungen, Spannrahmen usw.),* 2nd ed., Springer Verlag, Berlin 1967.
13. W. Bernard: *Praxis des Bleichens und Färbens,* Springer Verlag, Berlin 1966.
14. W. Bernard: *Druck von Textilien,* Springer Verlag, Berlin Göttingen Heidelberg New York 1969.
15. K. Schmidt: *Der Textildruck,* Dr. Spohr Verlag, Wuppertal 1973.
16. K. Spitzner: *Textildruck,* 3rd revised ed., VEB Fachbuchverlag, Leipzig 1980.

Specific References

17. H. G. Weissgerber, F. Weissgerber: *Walzengravur und Schablonenherstellung,* Melliand Verlag, Heidelberg 1960.
18. *Bayer Farben Rev.* **14** (1968) 28.
19. J. Eibl, *Melliand Textilber.* **57** (1976) 407.
20. E. Bobec, *Melliand Textilber. Int.* **56** (1975) nos. 10 – 12; *Melliand Textilber.* **59** (1978) no. 9; **61** (1980) no. 7 (literature references).
21. N. L. Moore, *J. Soc. Dyers Colour.* **90** (1974) 318 – 325 (literature references).
22. Sark, DRP 614 786, 1933.
23. Fastran, GB 1 284 824, 1971.
24. E. Küsters: "Reactiv Transferdruckverfahren für Cellulosefasern," *Int. Textil Service* **1** (1992) 63.
25. Star, DE 35 509 D/8 n, 1949.
26. N. de Plasse, FR 1 223 330, 1958. Ciba, GB 1 221 126, 1968.
27. Sublistatic, DE-OS 2 458 660, 1973.
28. M. Fox, *J. Soc. Dyers Colour.* **89** (1973) no. 12, 474 – 485.
29. A. Kershaw, *Textilveredlung* **11** (1976) 262.
30. T. L. Dawson, B. P. Roberts, *J. Soc. Dyers Colour.* **93** (1977) 439.
31. J. Eibl, *Chemiefasern/Textilind.* **27** (1977) 636 – 645.
32. J. Eibl, *Melliand Textilber.* **62** (1981) 565 – 567.
33. T. L. Dawson: "Jet Printing," *Rev. Prog. Color. Relat. Top.* **22** (1992) no. 12, 22.
34. K. H. Fluss: "Space Dyeing," *Bayer Farben Rev.* **26** (1975) 19. J. Wirtz: "Teppichdruck," *Bayer Farben Rev.* **30** (1978) 47.
35. *Bayer Farben Rev. (special issue)* **2 (Beflockung)** (1963).
36. *Melliand Textilber.* **38** (1957) 669.
37. P. Weber, *Text. Faserstofftech.* 1956, no. 3, 129, 130.
38. R. Müller, W. Schwindt, *Melliand Textilber.* **49** (1968) 1325 – 1330; *Melliand Textilber. Int.* **53** (1972) 201 – 203.
39. O. Schlösser, *Bayer Farben Rev.* **17/18** (1969).
40. H. Dahm, *Bayer Farben Rev.* **6** (1981) Sept., reprint Verdickungsmittel und Kleber.
41. H. Barth, *Bayer Farben Rev.* **15/16** (1968).
42. W. Berlenbach, *Melliand Textilber. Int.* **53** (1977) 207 – 210.
43. P. Habereder, F. Bayerlein in A. Chwala, V. Anger (eds.): *Handbuch der Textilhilfsmittel,* Verlag Chemie, Weinheim 1977, p. 621.
44. W. Schwindt, *Melliand Textilber.* **58** (1977) 1009 – 13.
45. U. Karsunky, M. Thissen, *Melliand Textilber. Int.* **53** (1972) 577 – 79.
46. F. R. Alsberg, P. R. Dawson, *Textilveredlung* **8** (1973) 365 – 370.
47. F. Bayerlein, *Melliand Textilber.* **58** (1977) 1017 – 1020.
48. E. Heym, *Melliand Textilber. Int.* **53** (1972) 213 – 214.
49. G. Barnhardt, *Text. Chem. Color.* **11** (1979) no. 10, 32 – 45.
50. M. Thissen, *Melliand Textilber. Int.* **53** (1972) no. 5, 573 – 576.
51. F. Bayerlein, P. Habereder, *Text. Prax. Int.* **29** (1974) 1411 – 1413.
52. P. Habereder, *Melliand Textilber.* **61** (1980) 165 – 166.
53. F. Bayerlein, L. K. Schwörzer, *Melliand Textilber. Int.* **53** (1972) 204 – 206.
54. K. Roth, *Textilbetrieb (Würzburg)* **98** (1980) no. 9, 50 – 65.
55. Chem. Fabrik Grünau, Druckverdickungsmittel, company brochure.
56. U. Karsunky, *Chemiefasern/Textilind.* **25** (1975) no. 2, 159 – 161.
57. J. Rosenbalm, J. Shelso, *Text. Chem. Color.* **11** (1979) no. 10, 28 – 31.
58. N. S. Volkonskaja et al., *Faserforsch. Textiltech.* **12** (1975) 321 – 323.
59. H. Schulzen, *Z. Gesamte Textilind.* 1958, nos. 6, 8, 11.
60. *Ciba-Rundsch.* **1** (1969) 19 – 34.
61. J. S. Racciato, *Text. Chem. Color.* **11** (1979) no. 2, 31 – 35.

62. Meyhall, Guar- und Guar-Derivate, bulletin.
63. W. Kühnel, *Bayer Farben Rev. (special issue)* **18** (1979).
64. E. G. Hochberg, *Text. Chem. Color.* **11** (1979) no. 5, 27 – 28.
65. L. T. Holst Jr., *Text. Chem. Color.* **11** (1979) no. 3, 17 – 19.
66. T. G. Perkins, *Text. Chem. Color.* **11** (1979) no. 10, 21 – 25, 44.
67. R. Hofstetter, G. Robert, *Textilveredlung* **14** (1979) no. 2, 51 – 56.
68. W. Badertscher, *Textilveredlung* **14** (1979) no. 7, 279 – 82.
69. W. Kothe: "Aktuelle Möglichkeiten der Druckwarenvorbehandlung," *TPI Text. Prax. Int.* **34** (1979) 52.
70. D. Bassing, W. Küppers: "Was bieten die BASF-Vorbehandlungssysteme dem Textildruck?" *TPI Text. Prax. Int.* **34** (1979) 58.
71. BASF-Ratgeber, Zellulosefasern, B 375 d, Ludwigshafen 1977.
72. W. Guth, A. Brüggemann: "Kontinuierliche Verfahren zum Bleichen und Weißtönen von Geweben und Gewirken in Verweilgeräten,"*Bayer Farben Rev. (special issue)* **15** (1975) 97.
73. T. Mandt, B. Brüssow: Der Stoffdruck in der Bundesrepublik Deutschland. (Forschungsstelle f. allg. Textil-Marktwirtschaft), Universität Münster 1980.
74. K. Schmidt: "Textildruck," in O. Spohr, E. Wagner (eds.): *Handbuch für Textilingenieure und Textilpraktiker,* 3rd ed., Dr. Spohr Verlag, Wuppertal-Barmen 1951.
75. H. E. Fiertz-David, E. Merian: *Abriß der chemischen Technologie der Textilfasern,* Birkhäuser, Basel 1948, p. 240.
76. IG-Farbenindustrie, Ratgeber für das Bedrucken von Baumwolle und andere Fasern pflanzlichen Ursprungs, 1934.
77. A. Schaeffer: "Die Oxydation der Leukoverbindung zum Küpenfarbstoff," *Melliand Textilber.* **38** (1957) 428.
78. A. Schaeffer: "Vorgänge beim Verküpen von Küpenfarbstoffen," *Melliand Textilber.* **36** (1955) 1033.
79. U. Baumgarte: "über Reaktionen von Reduktionsmitteln bei der Küpenfärberei," *Textilveredlung* **2** (1967) 896.
80. N. Grund: "Redoxprozesse im Ätzdruck auf Textilien," *Melliand Textilber.* **67** (1986) no. 12, 896.
81. BASF, Rongalit ST flüssig, technical report no. 10, Ludwigshafen 1989.
82. E. Fees: "Der Ätzdruck auf Cellulose," *TPI Text. Prax. Int.* **31** (1976) 161.
83. H. Ilg, V. Krentz: "Untersuchungen zur Ermittlung der günstigsten Einsatzbedingungen für das Zweiphasendruckverfahren." *Text. Prax.* **20** (1965) 324.
84. E. Fees: "Der Zweiphasendruck mit Küpenfarbstoffen," *Melliand Textilber.* **45** (1964) 45.
85. U. Blum: "ätz- und Reservedruck," *Textilveredlung* **15** (1980) 323.
86. U. Altmann: "Der Afrikadruck – Vita einer interessanten Entwicklung," *Bayer Farben Rev.* **14** (1968) 1.
87. F. Baumann et al., *Angew. Chem.* **68** (1956) 133 –150.
88. J. Eibl, *Melliand Textilber.* **39** (1958) 522 – 527, 660 –663, 772 – 773; **45** (1964) 789 – 796; *Bayer Farben Rev.* **24** (1974) 47 – 57, *Bayer Farben Rev. (special issue)* **no. 5 (Rapidogen-Farbstoffe auf neuen Wegen)**; *Bayer Color.* **10** (1960) 21.
89. D. Hildebrand, R. Schwaebel: "LEVAFIX P – Neue faserreaktive Farbstoffe für Cellulose," *Text. Prax.* **22** (1967) 796 – 808.
90. D. Hildebrand: "Zum Mechanismus des Färbens mit Reaktivfarbstoffen," *Text. Prax.* **25, 26** (1970) H 5 – H 8.
91. L. Schmidt: "Probleme mit Reaktivfarbstoffen aus der Sicht des Druckereipraktikers," *Textilveredlung* **13** (1978) no. 8, 293 – 298.
92. H. U. von der Eltz: "Die Chemie der Remazol-Farbstoffe," *Melliand Textilber.* **63** (1982) no. 11, 798 – 801.
93. H. Herlinger et al.: "Rheolog. Charakterisierung und Qualitätsbeurteilung von Druckpasten und Verdickungen," *TPI Text. Prax. Int.* **42** (1987) no. 10, 1231 – 1241.
94. H. Beckstein et al.: "Verringerung der Umweltbelastung in der Textildruckerei/Feuchtigkeitsauftrag v.d. HS-Daempfer," ITB Veredlung 1/92.
95. K. H. Blank, V. Wuertz: "Remazol-Farbstoffe, die Voraussetzung für anspruchsvolle, vielseitige Bunteffekte modischer Drucke," *Melliand Textilber.* **64** (1983) no. 10, 752 – 757.
96. K. H. Blank, "Remazol-Farbstoffe drucken und kalt fixieren," *TPI Text. Prax. Int.* **39** (1984) 249 –251.
97. H. Gutjahr, F. v. Vlissingen, EP 0 167 711, 1989.

98. C. Sommer, *Melliand Textilber.* **50** (1969) 692–698.
99. *Textilhilfsmittel-Katalog 1981/82,* Konradin Verlag, Leinfelden-Echterdingen.
100. L. T. Holst, *Text. Chem. Color.* **7** (1975) 154–156.
101. H. J. Manderla, *Chemiefasern* **15** (1965) no. 2, 96–107.
102. H. J. Manderla, *Melliand Textilber.* **43** (1962) 1310–1316.
103. H. J. Manderla, *Bayer Farben Rev.* **8** (1964) 32–38.
104. R. Koch, *Bayer Farben Rev.* **13** (1967) 23–29.
105. W. Kühnel, *Melliand Textilber.* **57** (1976) 315–320.
106. W. Kühnel, *Bayer Farben Rev.* **20** (1971) 27–32.
107. H. Dahm, *Bayer Farben Rev.* **12** (1967) 26–30.
108. H. J. Manderla, *Bayer Farben Rev.* **4** (1963) 24–31.
109. H. J. Manderla, *Bayer Farben Rev.* **5** (1963) 38–51.
110. H. Gutjahr, H. Dahm, *Melliand Textilber.* **10** (1968) 1207–1210.
111. G. Meyer, *Melliand Textilber.* **50** (1969) 698–702.
112. W. Kühnel, *Melliand Textilber. Int.* **51** (1970) 325–330.
113. R. Hofstetter, *Teintex* **44** (1979) 691–696.
114. F. G. Rebske, *Melliand Textilber. Int.* **54** (1973) 519–520.
115. U. Karsunsky, *Chemiefasern/Textilind.* **25** (1975) 728–731.
116. G. Meyer, *Melliand Textilber. Int.* **51** (1970) 1341–1344.
117. Mitsubishi, *JTN* **295** (1979) 77–86.
118. J. Winkler, *Melliand Textilber. Int.* **56** (1975) 728–731.
119. H. J. Manderla, *Bayer Farben Rev.* **7** (1964) 16–30.
120. C. Sommer, G. Vogl, *Melliand Textilber. Int.* **53** (1972) 440–445.
121. F. R. Alsberg, *Chemiefasern/Textilind.* **23** (1973) 439–444.
122. R. Rafael, *Melliand Textilber. Int.* **55** (1974) 361–364.
123. J. F. Dawson: "The Structure and Properties of Disperse Dyes in Polyester Coloration," *J. Soc. Dyers Colour.* **99** (1983) nos. 7/8, 183.
124. R. Hofstetter, G. Robert, *Textilveredlung* **14** (1979) 51–56.
125. W. Schwindt, *TPI Text. Prax. Int.* **34** (1979) 278–281.
126. J. Skoufis, *Text. Chem. Color.* **11** (1979) 279–282.
127. W. Badertscher, *Textilveredlung* **14** (1979) 106.
128. L. T. Holst, *Text. Chem. Color.* **11** (1979) 53–55.
129. W. Kühnel, *Bayer Farben Rev. (special issue)* **18** (1979) 21–27.
130. H. P. Schoepflin, *Text. Chem. Color.* **10** (1978) 225–229.
131. R. Hofstetter, *Text. Chem. Color.* **5** (1973) 172–180.
132. R. Hofstetter, *Textilveredlung* **11** (1976) 186–194.
133. J. Eibl, *Melliand Textilber.* **57** (1976) 663–667.
134. L. T. Holst, *Am. Dyest. Rep.* **65** (1976) 47.
135. J. Winkler, *TPI Text. Prax. Int.* **34** (1979) 302–309.
136. D. Brierley, J. R. Provost: "The Use of Disperse Dyes Containing Diester Groups to Produce Discharge Effects," *J. Soc. Dyers Colour.* **99** (1983) no. 12, 358.
137. D. Knittel, W. Günther, E. Schollmeyer: "Ätzdruck auf laservorbehandelten Polyestergeweben," *TPI Text. Prax. Int.* **45** (1990) no. 1, 46.
138. W. Kühnel, *Bayer Farben Rev.* **13** (1967) 16–22.
139. R. Schwaebel, *Bayer Farben Rev.* **24** (1974) 38–46.
140. G. Gabry, *Teintex* **40** (1975) 267–274.
141. B. Glover, *Am. Dyest. Rep.* **67** (1978) 47.
142. E. Fees, *Melliand Textilber.* **60** (1979) 595–605.
143. F. Miksovsky, A. Blum, *TPI Text. Prax. Int.* **34** (1979) 152–154, 167–168.
144. P. R. Brady, *J. Soc. Dyers Colour.* **95** (1979) 302–344.
145. R. R. Hyness, *Text. Chem. Color.* **2** (1970) 25–28.
146. G. Buzzoni, *Rev. Gen. Caoutch. Plast.* **52** (1975) 459–462.
147. A. Agster, *Melliand Textilber. Int.* **56** (1975) 470–472.
148. J. Lösch: *Fachwörterbuch Textil,* Verlag J. Lösch, Frankfurt/M. 1975.
149. Knecht-Fothergill: *The Principles and Practice of Textile Printing,* 4th ed., Griffin, London 1952.
150. H. E. Fiertz-David, E. Merian in [75].
151. D. Hildebrand: "Strukturchemie der natürlichen Polyamidfasern, *Z. Gesamte Textilind.* **71** (1969) 274–280, 311–317.
152. K. Reincke: Vorbehandlung von Wolle für den Druck, *Text. Prax.* **25** (1970) 419 ff.

153. C. Frommelt: "Bleichen und Weißtönen von Wolle," *Bayer Farben Rev.* **25** (1975) 80 ff.
154. J. Lewis: "Neuere Entwicklung bei Superwaschwolle," *Textilveredlung* **11** (1976) 214 ff.
155. H. Heiz: "Vorbehandlung von Wolle und Wollmischgeweben," *Textilveredlung* **13** (1978) 205 ff.
156. W. Schefer: "Nachweis und Beurteilung chemischer Schäden an Wolle," *Textilveredlung* **13** (1978) 55 –57.
157. W. Steger: "Die Naturseide, ihre heutige Veredlung und Ausrüstung," *Text. Prax.* **24** (1969) 379 – 383.
158. R. Koch: "Bedrucken von Halbwolle," *Bayer Farben Rev.* **33** (1981) 32 – 41.
159. B. Kramer: "Am seidenen Faden," *Bayer Farben Rev.* **29** (1977) 3 – 10.
160. P. Bakker, A. Johnson: "The Application of Monochlorotriazinyl Reactive Dyes to Silk," *J. Soc. Dyers Colour.* **89** (1973) 203 – 208.
161. V. Bell: "Recent Developments in Wool Printing," *J. Soc. Dyers Colour.* **104** (1988) no. 4, 159.
162. A. R. Czerny: "Textildruck auf Naturseide," *Melliand Textilber.* **69** (1988) no. 6, 438.
163. K. Y. Chu, J. R. Provost: "The Dyeing and Printing of Silk Fabrics," *Rev. Prog. Color. Relat. Top.* **17** (1987) 23.
164. H. Thomas et al.: "Anfärbetest zur Überprüfung der Reproduzierbarkeit in der Vorbehandlung von Wollgeweben für den Textildruck," *Melliand Textilber.* **70** (1989) no. 5, 352.

General References

165. P. Wengraf: "Sammelbericht Pigmentdruck in den Kriegsjahren," *Text. Rundsch.* **2** (1947) 125 – 131.
166. K. Reinartz: "Das ACRAMIN F-Verfahren," *Melliand Textilber.* **33** (1952) 626 – 628.
167. N. S. Cassel: "New Horizons for Pigment Printing," *Am. Dyest. Rep.* **49** (1960) 184 – 189.
168. W. Schwind: "Grundlagen und Anwendung der Pigmentfärbe- und Drucktechnik," *Melliand Textilber.* **45** (1964) 529 – 533, 668 – 672.
169. M. Glastra, F. van Lamoen: "Zusammenhänge zwischen rheologischen Eigenschaften und Verhalten von Druckpasten beim Rouleauxdruck," *Melliand Textilber.* **46** (1965) 1339 – 1346.
170. W. Berlenbach: "New Pigment Printing Techniques," *Int. Dyer Text. Printer Bleacher Finish.* **140** (1968) 693 – 699.
171. W. Schwindt: "Der Pigmentdruck – Bedeutung und Entwicklungstendenzen," *Melliand Textilber.* **50** (1969) 670 – 672.
172. W. Berlenbach: "Dispersionsverdickungen im Textildruck," *Melliand Textilber. Int.* **53** (1972) 207 –210.
173. P. Rasche: "Benzinfreier Druck mit Pigment," *Textilveredlung* **10** (1973) 31 – 34.
174. U. Perkuhn: "Bronzepigmente im Pigmentdruck," *Int. Text. Bull.* 1981, 229 – 236.

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Textile Technology

Standard Article

Burkhard Wulfhorst[1], Oliver Maetschke[2], Markus Osterloh[3], Alexander Büsgen[4], Klaus-Peter Weber[5]
[1]Institut für Textiltechnik, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany
[2]Institut für Textiltechnik, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany
[3]Institut für Textiltechnik, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany
[4]Institut für Textiltechnik, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany
[5]Institut für Textiltechnik, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page
BROWSE THIS TITLE
Article Titles A–Z
Topics
SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a26_553
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. General Introduction
2. Spinning
2.1. Introduction
2.2. Steps in the Cotton-Spinning Process
2.3. Harvesting and Ginning
2.4. Preparation for Spinning
2.5. Spinning Methods
2.5.1. Drawing
2.5.2. Combing
2.5.3. Conventional Spinning
2.5.4. Unconventional Spinning
2.6. Automation
2.7. Recycling
3. Weaving
3.1. Introduction
3.2. Steps in the Weaving Process
3.3. Weaving Preparation
3.3.1. Winding
3.3.2. Twisting
3.3.3. Preparation of the Warp Beam
3.3.4. Sizing
3.4. Weaving Methods
3.4.1. Fabric Construction
3.4.2. Loom Technology
3.4.2.1. Shedding Systems
3.4.2.2. Warp Let-Off Systems
3.4.2.3. Weft-Insertion Principles
3.4.2.4. Selvage Formation
3.4.3. Fabric Defects
3.4.4. Special Weaving Methods
4. Braiding
4.1. Introduction
4.2. Definition of a Braid
4.3. Production of Braids
4.3.1. Principles of the Mechanical Braiding Technique
4.3.2. Functional Elements of a Braider
4.3.3. Mechanical Systems
4.4. Interlacing Techniques
4.5. Materials Used in Braiding
4.6. Types of Braids and Their Applications
4.6.1. Ribbon Braids
4.6.2. Round Braids
4.6.3. Packing Braids
4.6.4. Three-Dimensional Braids
5. Knitting
5.1. Introduction
5.2. Classification of Knitted Fabrics
5.3. Technology of Stitch Formation
5.4. Structural Elements and Structures of Knitted Fabrics
5.5. Classification of Stitch-Forming Machines
5.5.1. Flat-Bed Knitting Machines
5.5.2. Circular Knitting Machines
5.5.3. Weft Knitting Machines
5.5.4. Warp Knitting Machines
6. Acknowledgement

[Top of Page]

1. General Introduction

Burkhard Wulfhorst

The inclusive term *textiles* encompasses items in the following categories:

1. Textile fibers
2. Textile fabrics
3. Goods manufactured from the above

Textile fibers (⟶) also fall into three categories:

Natural Fibers. Natural fibers of vegetable origin include cotton (⟶ Cellulose – Cotton) and bast fibers (i.e., jute, flax, and hemp; ⟶ Cellulose – Jute, ⟶ Cellulose – Flax, ⟶ Cellulose –

Hemp). The important natural animal fibers are silk (⟶ Silk) and wool (⟶ Wool). An example of a natural mineral fiber is asbestos (⟶).

Man-Made Fibers Produced from Natural and Synthetic Polymers. Among the fibers made from natural polymers are viscose (⟶ Cellulose – Viscose Fibers), cupro (⟶ Cellulose – Cuprammonium Fibers), acetate (⟶ Cellulose Esters – Cellulose Acetate), alginate, and diene. Man-made fibers from synthetic polymers include polyamide (⟶ Fibers, 4. Synthetic Organic – Polyamide Fibers; ⟶ Polyamides), polyester (⟶ Fibers, 4. Synthetic Organic – Polyester Fibers; ⟶ Polyesters), acrylic (⟶ Fibers, 4. Synthetic Organic – Polyacrylonitrile Fibers), and elastane (polyurethane, also known as spandex; ⟶ Fibers, 4. Synthetic Organic – Polyurethane Fibers).

Other Man-Made Fibers. These are based on inorganic materials, and include carbon fibers (⟶ Fibers, 5. Synthetic Inorganic – Carbon Fibers), metal fibers (⟶ Fibers, 5. Synthetic Inorganic – Metal Fibers), glass fibers, and spinning paper.

Worldwide production of fibers in 1992 is estimated at 40×10^6 t/a, of which 50 % represents cotton.

Fibrous raw materials can be divided into spinnable fibers of limited length (cotton, wool, and man-made fibers that have been cut to a particular length) and so- called endless fibers with essentially unlimited lengths (filaments and filament yarns). Such fibers lend themselves to direct processing into nonwoven fabrics and felts. Spun yarns can be produced from the appropriate fibers by means of various spinning processes. These and the filament yarns are linear in structure, and are used primarily for the manufacture of woven fabrics, braids, knits, and hosiery fabrics. Materials based on other types of surfaces have been developed for special purposes. Thus, bobbinet ("bobbinet lace") presents an open surface comprised of two distinct warp-thread systems, one weft-thread system, and a bobbinet-thread system running diagonally. Fabrics of this type are used for making curtains. Nets are used primarily in the fishing industry. Laces provide surfaces that are both open and transparent. Other materials useful especially in industrial applications include *warp-knitted multiaxial layers* (WIMAG) and *stitch-bonded fabrics* (NVG).

In the case of tufted fabrics, multiple needles are normally used to introduce new threads into a preexisting fabric backing. This results in a series of loops that may subsequently be cut. Fabrics of this type are often used as carpeting material.

Once manufactured, textile fabrics are subject to various types of finishing, including dyeing (⟶ Textile Dyeing), printing (⟶ Textile Printing), and both wet and dry processing (⟶ Textile Auxiliaries). Final make-up leads to products in three major marketing categories: apparel, domestic fabrics and linens, and industrial products. Roughly 50 % of total textile output is directed toward the apparel sector, with each of the others garnering ca. 25 %. The use of industrial fabrics is increasing rapidly, especially in the industrialized nations of the world.

This article is intended to provide an overview of major processes used in converting essentially one-dimensional fibers and yarns into various types of two- and three-dimensional surfaces for further utilization on the basis of the four fundamental technologies spinning (Chap. Spinning), weaving (Chap. Weaving), braiding (Chap. Braiding), and knitting (Chap. Knitting).

[Top of Page]

2. Spinning

Oliver Maetschke

2.1. Introduction

The technique of forming staple-fiber yarn includes all the process steps involved in production of staple-fiber yarn from a fibrous raw material. The fibrous raw materials themselves can be either natural products, such as cotton, wool, hemp, and flax, or synthetic materials, such as viscose, polyester, polyamide, or polyacrylic. The work of a spinning mill begins with preparation of the fibers for the actual spinning process, their transformation into a cohesive continuous structure, and the creation of fiber units suitable for further processing.

The basic principle of yarn formation has remained essentially unchanged since the early days of spinning. Hand spinning with simple whorls made of bone, wood, or clay was practiced in the New Stone Age as early as ca. 4000 B.C. By the 13th century A.D. simple hand spinning wheels were already in use in Europe. In 1764, JAMES HARGREAVES constructed the first fully mechanized spinning machine with several spinning positions, and it was this that formed the basis for introduction of industrial spinning into Europe.

Many different spinning processes are practiced depending on the field of application (Table 1), with the cotton-spinning process being the most common. Total annual world production of fiber amounts to about 40×10^6 t (Table 2), where cotton accounts for 20.7×10^6 t (i.e., ca. 50 % of total fiber production). Given that many man-made fibers consist of a combination of filament yarns and staple fibers, and that the corresponding staple fibers are processed in part using cotton-spinning technology, then it is clearly this process that predominates. For this reason the cotton-spinning process is the only one that will be described here.

Table 1. Spinning processes

	Cotton spinning mill (see also Table 2)		Worsted yarn spinning mill	Carded yarn spinning mill	Semi-worsted yarn spinning mill
	Ring spinning	OE* Rotor spinning			
Raw materials	cotton, man-made fibers		wool, man-made fibers	wool, man-made fibers, cotton	synthetic fibers (wool)
Fiber length	≤ 50 mm		≤ 80–200 mm	≤ 120 mm	80–200 mm
Spinning-machine set-up	blowroom machines cards drawing frames flyers ring-spinning frames spooling frames	rotor spinning	preparatory machines card sets combs drawing machines gill boxes flyers/apron frames ring-spinning frames	preparatory machines card sets tape divider/ rubber condensor carded wool ring-spinning frames mules carded wool rotor spinning machines	preparatory machines card sets drawing frames flyers/rubbing frame ring-spinning frames covering machines OE* rotor spinning machines
Yarn gauge	1000–20 tex (carded) 5 tex (combed)	1000–12 tex	84–10 tex	1000–50 tex	10000–25 tex
Yarn type	firm, hairy (carded) smooth (combed)	less firm, smooth with wrapped fibers	smooth, shiny	less firm, bulky, voluminous, high stretch	
Applications	underwear, shirts, blouses, handkerchiefs, sport clothes, home furnishing fabrics, table linens, domestic items		outerwear	outerwear (cloth), home furnishing fabrics, industrial goods (filters)	Floor coverings, home furnishing fabrics (tapestry, upholstery), technical textiles

	Cotton spinning mill (see also Table 2)		Worsted yarn spinning mill	Carded yarn spinning mill	Semi-worsted yarn spinning mill
	Ring spinning	OE* Rotor spinning			
Raw materials	flax		hemp	jute	
Fiber length	ultimate fibers 25–30 mm bast fibers 200–1500 mm		ultimate fibers 15–25 mm bast fibers 1000–2500 mm	ultimate fibers 2–3 mm bast fibers 1000–2500 mm	
Spinning-machine set-up	heckling machines (industrial fibers 100–600 mm) gill boxes direct spinning flyers flyer-/ring-spinning frames		heckling machines gill boxes/ drawing frames fine gill spinning frames flyer-/ring-spinning frames	heckling machines (bast fibers 200–250 mm) breakers jute-spinning frames	
Yarn gauge	680–7.2 tex	spinning limit 8.4 tex	1700–50 tex	dependent upon fiber type; 84 tex to > 840 tex	
Yarn type	irregular, smooth, shiny, firm		irregular	high strength	
Applications	woven fabrics knitted goods (blended) nets women's/men's outerwear household fabrics (upholstery, linens) technical textiles (tarpaulins, sails) twisted yarns		technical textiles (straps, drive belts) carpet backing twisted yarns	technical textiles (straps, hose linings) upholstery fabric, carpets garments (only in blends)	

* OE = Open end.

* See also Table 2.
** OE = Open end.

Table 2. Fiber production in 1000 t (1991)

	Staple fiber	Filament yarn	Total
Cotton	20 793		20 793
Country of origin			
USA	3 835		3 835
Russia	2 491		2 491
China	5 675		5 675
Pakistan	2 676		2 176
India	2 023		2 023
Wool	3 010		3 010
Silk (raw silk)			76
Man-made fibers	9 658	8 014	17 673
Natural origin	1 682	833	2 515
Synthetic	7 976	7 181	15 158

The cotton-spinning process — also known as the short-staple or three- cylinder process — is suited to all types of fibers with lengths ranging up to ca. 40 mm, and it is very flexible with regard to properties and fields of application for the resulting yarns. Staple-fiber yarns made according to the cotton-spinning process are further processed into fabrics, knits, and braided materials for use as apparel, domestic fabrics and linens, and industrial products.

2.2. Steps in the Cotton-Spinning Process

The processing of fiber into yarn is divided into several stages, as indicated in Scheme 1. Depending on the desired yarn properties and the nature of the fibrous raw materials, various mechanical arrangements and spinning methods are employed. In each case it is necessary to establish an optimum compromise between ideal yarn characteristics and economical production.

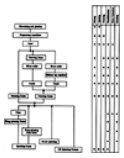

Scheme 1. Processing of fiber into yarn

The matrix in Scheme 1 underscores the main functions associated with the various process steps. The process step *mixing* covers both mixing of a nonhomogeneous fiber substrate and the dosed blending of fibers of differing origin. *Opening* refers to the breaking up of compressed fiber packets to yield individual fibers. *Cleaning* refers to the removal of interfering particles, such as wood or leaf components and fiber knots ("naps") that cannot be opened. *Parallelization* refers to alignment (insofar as possible) of all the fibers in a single direction. *Drawing* refers to the extension ("drafting") of an oriented collection of fibers. The purpose of drawing is to promote thorough mixing and blending of multiple raw materials as well as further homogenization of sliver weight. This is facilitated by the doubling of several slivers prior to actual drawing.

2.3. Harvesting and Ginning

Almost all cotton is now harvested mechanically in accordance with either the "stripper" or the "picker" principle. In *stripping*, all plant components are removed simultaneously by the action of rotating brush rolls. A *picker* incorporates rotating spindle fingers that separate cotton fibers selectively. The raw cotton is then freed of coarse impurities and subsequently separated from seeds in a process known as ginning. Finally, the fiber material is compacted into bales weighing ca. 250 kg each for trade or direct transport to a spinning mill.

The most important cotton-producing countries are listed in Table 2. Cotton bales are traded throughout the world at a price (ca. 2 – 6 DM/kg) that depends on quality as well as market demand. The most important quality characteristics of cotton include fiber length, fiber fineness, fiber strength, dirt content, and color.

2.4. Preparation for Spinning

Two examples of fiber preparation systems are illustrated in Figure 1. The fiber is transported pneumatically from one machine to the next, with a throughput in the range 500 – 800 kg/h depending on mechanical design.

Figure 1. Examples of fiber preparation systems offered by two different manufacturers

A) Trützschler; B) Rieter Spinning Systems

a) Bale opener; b) Particle separator; c) Twin-roll cleaner; d) Six-fold mixer; e) Single-roll intensive cleaner; f) Four-fold mixer; g) Quadruple-roll intensive cleaner; h) Dedusting device; i) Card; j) Free-beating cleaner; k) Intensive cleaner

The specific tasks associated with spinning preparation include:

1. Opening of the bales
2. Separation of interfering particles (refuse)
3. Dedusting

4. Opening of fiber flocks
5. Homogenization of the raw material
6. Mixing of various raw materials

Bale Opening. Bale opening refers to breaking up of the bales into individual flocks (see Fig. 1). The nature and intensity of the opening process has a decisive influence on all subsequent processing steps. The more finely the cotton is opened, the more interfering particles will be found on the surface, from which they can more readily be eliminated. In modern bale-opening systems, up to 80 bales are placed side-by-side on the floor and then stripped into layers by a programmable bale opener. Opening into fiber flocks is accomplished mechanically with the aid of milling rolls, after which the flocks are transported further in a stream of air with concurrent dedusting.

Opening, Cleaning. The main functions of opening and cleaning machines are additional breaking up of the fiber flocks, elimination of interfering particles (refuse), and dedusting of the cotton. Two basic principles are utilized in accomplishing this task.

With a *free-beating* system, flocks are caught in free flight and accelerated by the working components (beaters) of the cleaner. Interfering particles are separated on grids through the combined effects of centrifugal force and gravity. Such machines may incorporate one or more rolls for this purpose, as illustrated by the example in Figure 2 (see also Fig. 1).

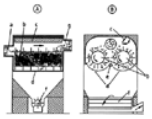

Figure 2. Schematic diagram of a twin-roll cleaner

A) Cross section; B) End view

a) Fiber inlet with guide plate; b) Cleaning rolls; c) Flashing; d) Grid; e) Guide; f) Rotary vane sluice; g) Fiber discharge port with fresh-air inlet

In the case of *constrained beating* the flocks are held between two feed rolls, or between a feed plate and a feed roll, while the beating elements perform their function. Depending on the degree of opening, nose-beaters, spikes, or saw teeth are used as the beating elements (Fig. 3). As a result of the restraining effect, the extent of the opening achieved with constrained beating is greater, but the process is also more aggressive than free beating. Cleaning is accomplished with grids, knives, and exhaust hoods.

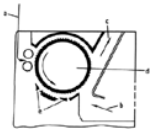

Figure 3. Schematic diagram of a sawtooth cleaner

a) Fiber inlet; b) Air inlet; c) Fiber discharge in a stream of air; d) Beating roll; e) Knives (grids) for particle separation

Mixing. Mixing comprises two functions: blending and homogenization. *Blending* refers to the establishment of a definite quantitative relationship between various raw-material partners. This is accomplished through one or more of the following processes:

1. Manual deposition of the appropriate components (imprecise)
2. Automatic bale opening
3. Use of weighing-hopper feeders (batch) and continuous dosing devices inside the preparatory machines

The purpose of *homogenization* is to achieve a distribution of the various components as homogeneous as possible in the final product. With the *mixing- chamber* principle, the maximum available space is filled in horizontal layers and subsequently processed vertically from one side. Multiple mixers (Fig. 4) consist of several independent chambers arranged one behind the other. In this case a distinction can be made between two basic working principles: flocks may be fed in all at once and processed at different times, or fibers that are fed in at different times may be processed simultaneously.

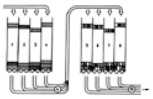

Figure 4. Multiple mixer

Carding. The last machine required in spinning preparation is the card [Fig. 5; see also Fig. 1 (i)]. The functions of a card are

1. Elimination of interfering particles and short fibers
2. Opening of the flocks into individual fibers
3. Parallelization of the fibers
4. Mixing
5. Drawing
6. Slivering
7. Sliver coiling

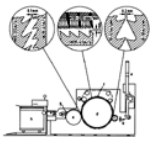

Figure 5. Schematic diagram of a card

a) Pneumatic feeder; b) Feed device; c) Liker-in; d) Tambour or drum; e) Doffer; f) Card flats; g) Stripper rolls; h) Can coiler

The various working elements inside a card are equipped with steel “clothings” (for details see Fig. 5). The fibers themselves are brought between sets of card tips arranged both opposite and adjacent to one another.

Carding (i.e., the separation and parallelization of individual fibers) is a result of relative motion occurring between clothings on the several working elements. If two working elements subject to motion relative to each other are outfitted with card tips inclined in opposite directions, the system is arranged in what is called the *carding position* (Fig. 6). The intensity of the actual carding action is a function of the speed differential between the carding elements and the angles of inclination of the corresponding clothings. If the working elements are instead inclined in the same direction, fibers pass from one fitting to the other, and the system is said to be arranged in the receiving or transfer position.

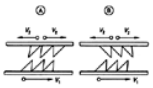

Figure 6. Operating principle of a card (schematic)

A) Carding zone; B) Transfer zone

v_1, v_2 = velocities of the working elements, where $v_1 >> v_2$

Homogenized fibrous material is introduced into the card (Fig. 5) via a pneumatic flock feeder (a). The material is then fed to the liker-in (c), a roll covered with saw teeth, via the feed device (b). Here the flocks are also opened as a consequence of a difference in peripheral speeds. Simultaneously, 70 – 75 % of all interfering particles accumulating at the card are eliminated at the liker-in with the aid of knives. The principal operating component of a card is the tambour (d). The surface of a tambour is covered with more than 4×10^6 card tips, and the tambour moves with a peripheral speed of ca. 26 m/s. Between the liker-in and the tambour is a carding zone, which ensures further separation of the fibers. However, most of the carding effect takes place between the tambour and the card flats (f), which are arranged in a carding position relative to each other. It is here that separation into individual fibers is achieved, along with fiber parallelization and the simultaneous elimination of particles and short fibers. Processing of man-made fibers is accomplished with fixed flat elements. Modern cards often incorporate additional fixed carding bars and dirt separators in front of and behind the card flats. Even though the region between the tambour and the doffer (e) constitutes another carding zone, fibers are nevertheless transferred from the tambour to the more slowly moving doffer as a result of the angle of inclination of the fittings. The observed transfer behavior is described by a *transfer factor*, which specifies the fraction of the fibers on the tambour that is expected to pass to the doffer with each revolution of the tambour. Interfering particles and dust are eliminated from fibrous material remaining on the tambour by means of grids and knives located beneath the tambour itself.

The condensed fiber web (fleece) is removed from the doffer by means of stripper rolls (g). Remaining interfering particles are pulverized by two squeezing rolls before the fleece is formed into a sliver. A coiler is then used to deposit the sliver in cycloidal form into a revolving can (h). Cards are provided with short- and long-term control devices to compensate for fluctuations in the mass of the sliver.

2.5. Spinning Methods

2.5.1. *Drawing*

After carding, the fiber slivers must be doubled and drawn. *Doubling*, a process in which several slivers are combined and subsequently drawn, results in a homogenization of the sliver. A regulating device is often used to reduce fluctuations in the sliver mass. This usually involves mechanical scanning of the incoming slivers, with the feed rate controlled as a function of the measured sliver mass. In the case of slivers of differing origin, doubling can also be used to achieve blending.

A drawing frame (Fig. 7) generally consists of a can creel (a), a framework for the drive and control units, the *drafting* (drawing) *unit* (b), and a can coiler with an automatic changer (c). The drafting unit comprises two or three pairs of rolls arranged one behind the other. A continuous fiber band is drafted between the "nip lines" because the peripheral velocities of successive cylinder pairs increase in the direction of transport. Modern drawing frames are equipped with three pairs of cylinders, resulting in two drafting zones. A light preliminary draft (ca. 1.05-fold) in the pre-drafting zone is generally followed by ca. 6 – 8-fold principal draft. The drafts in the individual zones must be multiplied in order to compute an overall draft. The stretched band (sliver) is deposited in cans in cycloidal form. Most spinning processes combine two drawing passages in sequence to achieve effective homogenization, parallelization, and mixing. A typical sliver exit speed is 500 – 1000 m/min.

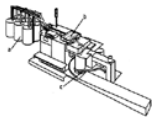

Figure 7. Schematic representation of a drawing (drafting) frame

a) Can creel; b) Drafting unit; c) Can changer

2.5.2. *Combing*

The combing process is designed to remove short fibers, impurities, and naps, permitting the spinning of fine yarns and an improvement in the yarn quality. The percentage by weight of combed-out material ("noil") is subject to adjustment. Depending on the raw material, noil percentages between 5 and 30 are common.

The fibrous material must be specially prepared before being fed to the combing machine (see Scheme 1). On a *comber lap machine*, several parallel slivers are combined into a single lap, which is then introduced into the combing machine. Successful removal of fiber "hooks" requires taking the orientation of the hooks into account, because the working principle of the combing machine is such that it is most effective with "guide hooks" — those in which the hook is located at the forward end of the fiber relative to the direction of travel. A fiber hook at the rear of a fiber is called a "trailing hook." The ratio of guide to trailing hooks in a card sliver is typically ca. 15: 65. For effective hook removal there must be an even number of machine passages between the card and the combing machine.

Combing is a batch process. Inside the combing machine (Fig. 8) the fiber lap (a) is fed to the feed rolls (b) via a lap trough. These rolls transport the fleece between the top nipper (c) and the bottom nipper (d) into the vicinity of the combing cylinder (e). The two sets of nippers hold the fibers while the combing cylinder combs the fiber tuft. The nippers then open and move toward the detaching rolls (f). The detaching rolls move the slubbing (h) back so that the combed-out fiber tuft can be returned to the slubbing. Subsequently, the detaching rolls rotate in the direction of transport and pull the fibers of the combed-out tuft out of the fleece. A top comb (g) then pierces the material to comb the back end of the fiber tuft, after which the nippers return to their original position and a new combing cycle begins. A rate of 300 – 400 combing cycles per minute is normal.

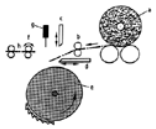

Figure 8. Operating principle of a combing machine

a) Lap; b) Feed rolls; c) Top nipper; d) Bottom nipper; e) Combing cylinder; f) Detaching rolls; g) Top comb; h) Slubbing (combed sliver)

2.5.3. *Conventional Spinning*

Conventional spinning refers to the combination of a flyer, a ring-spinning frame, and a spooling frame (see Scheme 1). About 80 % of all staple-fiber yarn is spun according to this process. In an unconventional spinning process the three operations are combined in a single machine.

Flyer. The functions of the flyer (or "roving machine") are further drafting of the sliver, coupled with the introduction of protection against faulty drafting in the form of a protective twist (Fig. 9). The twist must be adequate to prevent faulty drafting, but it must also be sufficiently limited to permit a slubbing to be drafted up to the desired final yarn fineness ("gauge") in the drafting unit of the ring-spinning frame.

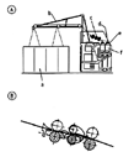

Figure 9. Typical arrangement of a flyer

A) Overview; B) Detailed sketch of the drafting unit

a) Can creel; b) Framework; c) Drafting unit; d) Opening in the top of the flyer leg; e) Flyer leg; f) Flyer bobbin

Slivers are first drafted in the flyer drafting unit (c), which is usually a 3-roll, 2-apron system (see Fig. 9). Aprons are used to guide the fibers into the drafting zone. Drafts as high as ca. 50-fold can be achieved with flyer drafting units. From the drafting unit the drafted sliver ("flyer roving") is transported to the flyer leg (e) via an opening in the top of the flyer leg (d). The sliver exits through an opening at the bottom end of the flyer leg and is led around a spring finger to the bobbin (f). With each revolution of the flyer the roving acquires one twist. Winding occurs owing to a leading or lagging of the bobbin with respect to the flyer leg. The vertical displacement required for smooth winding is attended to by the bobbin. Since the flyer operates at a constant feed rate, the vertical displacement and rotational speed of the bobbin are subject to constant adjustment as a function of the spool diameter, thereby ensuring uniform twisting and winding. The current velocity limit for flyers is 1300 – 1500 rpm.

As a result of technical limitations and the fact that it cannot easily be automated, efforts have long been underway to find a replacement for the flyer. Unconventional spinning processes have already made flyers partially obsolete. Certain ring-spinning frames also permit the spinning of slivers directly, although their share of the market is still very small.

Ring-Spinning Frames. In order for yarn to be produced from slubbing, further drafting is required up to the desired yarn gauge, and the appropriate degree of twist must be introduced. All of this occurs on the ring-spinning frame (Fig. 10), after which the finished yarn must be wound.

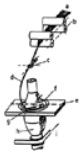

Figure 10. Principle of the ring-spinning process

a) Slubbings; b) Drafting unit; c) Thread guide; d) Rotating thread balloon; e) Ring rail; f) Ring; g) Traveler; h) Spindle; i) Spindle drive

As shown in Figure 10, slubbings (a) are fed from a creel to a double-apron drafting unit (b) where they are drafted to the desired yarn gauge. The thread passes through a thread guide (c) mounted concentric to the spinning axis and then through the traveler (g) to the spindle (h). The traveler is dragged along by rotation of the spindle, causing it to rotate on the ring (f), whereby the yarn is twisted once with each rotation of the traveler. This twist is propagated up to the spinning triangle. The traveler lags behind the spindle, facilitating winding of the yarn. The extent of the twist *T* applied to the yarn can be adjusted by altering the ratio of delivery speed to rotational speed of the spindle according to the relationship:

$$T = n_T / f$$
$$n_T = n_{Spi} - f/(d \cdot \pi)$$

where *f* is the feed rate at the drafting-unit exit, *d* is the current spindle diameter, n_T is the traveler speed, and n_{Spi} the spindle speed.

The stroke displacement required for satisfactory winding is achieved by means of the ring rail (e). The performance limit for ring spinning is determined by the maximum traveler speed, which is ca. 40 m/s, producing local temperatures as high as 450 °C. With higher traveler speeds it is impossible to provide adequate dissipation of the heat generated, resulting in potential damage to the traveler. In the case of man-made fibers, such heat build-up may also lead to spots of melted fiber in the yarn.

Thread forces in the thread balloon (d) can be reduced with the aid of *balloon- constricting rings*, which diminish the magnitude of centrifugal forces acting on the yarn. Automatic bobbin changing has also become a standard feature of modern ring-spinning frames. Although automatic devices have been developed for repairing breaks in the yarn, this task is still largely performed manually.

2.5.4. *Unconventional Spinning*

Open-End (OE) Rotor Spinning. Carded sliver, drawn sliver, and combed sliver can all serve as starting materials for OE rotor spinning. In the spinning unit itself (Fig. 11) the sliver (a) is fed to an opening roller (b) via a feed roll and a feed plate. The surface of the opening roller rotates at 20 – 30 m/s, and leads to very intensive combing and separation of the fibers. Refuse particles are eliminated below the opening roller, whereas the fibers themselves are removed in a stream of air and accelerated further in a conical fiber-guiding channel (c). The fiber-guiding channel ends in a rotor (d) that revolves with a velocity of up to 130 000 rpm. Centrifugal force causes the fibers to slide along the rotor wall and collect as a ring in the rotor groove. The OE rotor-spinning process is called an "open-end" process because the sliver is broken up into individual fibers immediately prior to yarn formation. Since the fibers are accelerated throughout the process, nearly all fibers collecting in the groove assume an extended configuration. Spinning is accomplished with the help of a previously finished piece of yarn that is fed into the rotor through the doff (discharge) tube (f) and the navel (e). As this piece of yarn rotates in the rotor groove, fibers lying in the groove become bound to its end. Subsequent successive withdrawal of the yarn through the navel and the doff tube results in a continuous spinning process. Yarns can be spun by this method with a fineness ("count") down to 12 tex (where "tex" = g/1000 m).

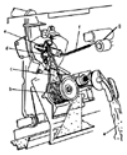

Figure 11. Open-end (OE) rotor spinning unit

a) Sliver; b) Opening roller; c) Fiber-guiding channel; d) Rotor; e) Navel; f) Doff tube; g) Yarn

One technological disadvantage of the OE rotor-spinning process is the presence of “wrapped fibers.” Fibers fed onto the previously twisted end of yarn in the rotor groove wrap themselves in such a way that one end of each fiber is oriented in the direction of the yarn twist, whereas the other end wraps around the yarn in the opposite direction (Fig. 12). Wrapped fibers make rotor yarns unsuitable for use in cutpile materials (e.g., carpets and velvet) because they interfere with opening of the pile threads.

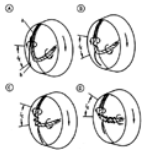

Figure 12. Steps (A – D) in the formation of a wrapped fiber

a) Fiber head; b) Fiber end

l_E = length of yarn-forming zone

As a result of their structure, OE rotor yarns are not as strong as comparable ring yarns. However, the OE rotor-spinning process is more economical, because it avoids the use of both flyers and spooling machines, and an OE rotor-spinning machine can be fully automated.

OE Friction Spinning. Another spinning process based on the open-end principle is OE friction spinning (Fig. 13). This process starts with several slivers (a) that are broken up into individual fibers by an opening unit (b). The fibers are subsequently transported in a stream of air through a guide channel (c) into the yarn-formation zone (d). The frictional forces required for conferring a twist are generated by friction elements (e) whose surfaces move perpendicular to the direction of yarn withdrawal. In most cases this entails two perforated, cylindrical friction drums that rotate in a common direction. Fibers are fed into the wedge-shaped gap between the two perforated rolls. Reduced pressure in the region of the wedge ensures that the fibers will be sucked through the perforations and onto the rolls themselves. Normal and frictional forces arising in the process cause the fibers to rotate and bind to the open end of the yarn.

Figure 13. Friction spinning unit

a) Slivers; b) Opening unit; c) Guide channel; d) Formation zone; e) Friction elements; f) Winding unit

OE friction yarns offer relatively little strength, so the process has gained only limited acceptance, and then only for yarns in the coarse range. However, it is suitable for the production of hybrid yarns for industrial applications.

Air-Jet Spinning. Air-jet spinning (Fig. 14) is a process that is becoming more and more important. Here the take-off sliver is first drafted to the desired yarn count by means of a drafting unit (a). Two air-jets (b1, b2) are located immediately behind the exit of the drafting unit. Air flowing tangentially through the twist jet (b2) confers upon the thread a “false twist” that is propagated as far as the exit of the drafting unit. This temporary twist is confined to a definite region, and it dissipates as the yarn leaves the false-twist zone. The other jet (b1) has the task of spreading out individual fibers that may protrude from the thread created by the false twist. This is accomplished with the help of air flowing through tangential holes arranged diagonally relative to the direction of yarn transport. Once the false twist has been eliminated through yarn withdrawal, the resulting yarn core is completely free of twist. Fiber ends that have been spread apart by the second jet wrap themselves around the core as the false twist dissipates. Axial force exerted in the course of this wrapping around the yarn core causes yarn compression as a result of frictional contact.

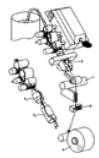

Figure 14. Schematic diagram of an air-jet spinning device

a) Double-apron drafting unit; b) Air jets; c) Discharge rolls; d) Yarn cleaner; e) Spool

Very fine yarns (down to ca. 10 tex) displaying good uniformity can be spun with the air-jet process. Good yarn characteristics depend upon a fibrous raw material that is fine, strong, and as uniform in length as possible. For this reason the process is best suited to man-made fibers or mixed yarns based on a combination of cotton and man-made fibers. Fine yarns can be produced very economically in this way.

Other spinning processes are known in addition to those described, but none has yet achieved commercial significance.

2.6. Automation

All the machines used in preparation for spinning, including the cards, are fully automated, and fibers are transported pneumatically from one machine to the next. The card produces a carded sliver, which is deposited in cans, and these cans then become available for subsequent transport.

Slivers are transported from the card via the drawing frame or the combing machine to the flyer either in cans or as laps. Different forms of transport have been adopted in different mills, but much of the work is still accomplished by hand.

In the case of a conventional spinning line (flyer, ring-spinning frame, winding machine) there exists a fully automated approach that is increasingly being adopted. Unconventional spinning systems are already fully automated. Online quality assurance is usually introduced simultaneously with automation.

2.7. Recycling

The recycling of textile raw materials has a long and exemplary history in both the industrial and private sectors. For example, one company in Bocholt (Federal Republic of Germany) has been recycling textiles for more than 130 years. The Italian city of Prato near Florence is also famous in this respect. Initially, a shortage of raw materials, especially natural materials, always constituted a major incentive.

The total quantity of new textiles introduced into the German market in 1992 is estimated at 2.1×10^6 t. It must be assumed that roughly the same amount of material is subject annually to disposal. Recycling in this context may take various forms, including *material recycling, chemical recycling* (recovery of chemical starting materials), and *energy recycling* (recovery of energy consumed in production).

Until now, material recycling has been the most common type of recycling in the textile industry, but in the future chemical and energy recycling will necessarily increase in importance. In the case of material recycling, various methods are available, depending in part on the intended purpose of the textiles with respect to the three categories apparel, domestic fabrics and linens, and industrial products. In the case of old textiles there exists the possibility of reuse in the form of second-hand materials. Old- clothes collection represents a second way of disposing of waste textiles. Clothes accumulated in this way are sorted in special facilities into items that are still usable, or from which useful fabric can be recovered, and waste. A third possibility is direct disposal as trash. In this case, fibers can sometimes be recovered through tearing processes, providing materials that can be used as secondary raw materials.

The incorporation of secondary raw materials into carded yarn has a long tradition. Recycling in a cotton-spinning operation can be accomplished either offline or online. In the offline approach, production waste is collected, prepared for use, and returned to the process as a secondary raw material. In the online process, prepared fibers are reintroduced into the same spinning procedure from which they were originally removed in the form of production waste.

[Top of Page]

3. Weaving

Markus Osterloh

3.1. Introduction

The technique of weaving has been in use for about 5000 years as a way of producing textile fabrics. Weaving is recognized worldwide as the dominant technology for this purpose. World output of woven fabrics amounts to ca. 30×10^6 t/a [21], [22]. Fields of application for woven fabrics include apparel (especially men's and ladies' garments), table and bed linens, domestic fabrics (carpeting, upholstery, drapery), and industrial products.

The economic importance of industrial textiles is increasing. These now account for about 25 % of all woven fabrics, and include reinforced composites commonly used in light construction as well as geotextiles for road, tunnel, and dam construction. Industrial textiles are also used as protective or filtering materials, and in medical applications.

A woven fabric consists of two or more thread systems that cross each other at right angles, known as the *warp* and the *weft* (see Fig. 15). The warp threads (a) run in the direction of production, and are wound onto a warp beam (c) in the desired width and spacing (ends per inch). Alternate raising and lowering of the heald frames (d) causes warp threads that are supported by these frames to form a passageway (*shed*), through which the weft thread (b) can be inserted. After weft insertion the weft thread is beaten firmly against the edge of the fabric by a reed (e). The relative positions of warp threads are then changed, binding the newly inserted weft thread in its place.

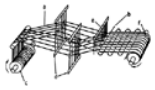

Figure 15. Schematic representation of a loom

a) Warp threads; b) Weft thread; c) Warp beam; d) Heald frames; e) Reed; f) Cloth beam

3.2. Steps in the Weaving Process

A summary of the process steps from ring-spun yarn to loom-state fabric is provided in Figure 16. Here a distinction is made between preparation for weaving and the weaving process itself.

Figure 16. Steps in the weaving process

A) Winding; B) Warping; C) Sectional warping; D) Sizing; E) Drawing-in; F) Weaving; G) Inspection

a) Cops; b) Crosswound package; c) Warp beam; d) Unsized yarn; e) Sized yarn; f) Drawn-in warp; g) Woven cloth

Preparation for weaving consists of several operations: winding (A), warping (B) or sectional warping (C), sizing (D), and drawing-in (E). Yarns produced on a ring-spinning frame are supplied as spindles (cops), which must always be rewound in the form of crosswound bobbins. OE rotor-spinning machines, OE friction-spinning machines, and air-jet spinning machines all deliver crosswound bobbins directly. Yarn supplies for the warp and the weft are prepared in the *winding* process. Warp threads are then processed into a warp beam either by *warping* or by *sectional warping*. In the subsequent sizing operation the outer surface of the yarn is smoothed by application of a sizing agent. This is intended to reduce friction as well as thread damage during the succeeding weaving process. The fact that the warp is drawn in at several places — by the warp-stop motion, the heald frames, and the reed — minimizes the time during which the weaving machine (*loom*) is in disuse in the course of resetting. If twisted yarn happens to be required, twisting would also be included as one of the steps in weaving preparation.

In the actual weaving process (F in Fig. 16), the two thread systems (warp and weft) are linked together by interlacing (see Fig. 15). The nature of the crossing is called the *weave* or *fabric construction*, and it is this that determines the mechanical and technological properties of a fabric. In the fabric inspection step (G), any flaws in the fabric are detected and recorded. This is followed by textile finishing and make-up.

3.3. Weaving Preparation

3.3.1. *Winding*

The purpose of winding (Fig. 17) is to maximize efficiency in the subsequent processes. Winding ensures both the availability of an adequate supply of thread and optimal unwinding characteristics. An integral feature of the winding process is an improvement in yarn quality through the removal of defects (e.g., thick or thin regions).

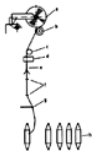

Figure 17. Schematic representation of winding

a) Crosswound package; b) Grooved drum; c) Broken-thread stop-motion; d) Thread clearer; e) Thread brake; f) Thread guides; g) Balloon restrictor; h) Cops

With respect to spools, a distinction is made between those with parallel winding and crosswound packages (Fig. 18). Crosswound packages are used for weaving because of their superior unwinding properties. A spool with parallel winding has a higher package density, and is usually unwound tangentially.

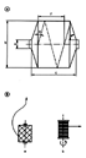

Figure 18. Factors associated with yarn spools

A) Spool characteristics; a) Tube diameter; b) Spool diameter; c) Maximum thread-guide traverse; d) Helix angle α; e) Slope gradient; f) Pitch

B) Unwinding behavior; a) Crosswound package with axial unwinding; b) Parallel-wound spool with tangential unwinding

Random Winding. The characteristic feature of random winding (Fig. 19 A) is peripheral drive of the tube with the aid of a grooved drum or a smooth friction roll and a moving thread guide (Fig. 17). Thread is thereby wound onto the spool at a constant helix angle α (Fig. 19; see also Fig. 18 A). At the same time, the number of windings per thread-guide pass, known as the winding ratio *W*, decreases with increasing spool diameter. If the winding ratio is an integer, individual layers of thread are wound closely next to each other or on top of each other. This is called patterned or ribbon winding. It may lead to uneven dyeing as a result of variations in the density, as well as thread-layer slippage during thread removal because of insufficient crossing of the layers. Winding machines are therefore equipped with special devices to counteract such “ribboning” by introducing periodic slippage between the drive and the spool.

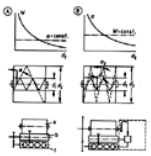

Figure 19. Comparison between different types of windings

A) Random winding; B) Precision winding

a) Spool; b) Peripheral drive; c) Grooved drum

d_1 = tube diameter; d_2 = spool diameter; α = helix angle; *W* = winding ratio

The advantages of random winding include spool stability, large packages, and mechanical simplicity. The disadvantages are ribboning and difficult unwinding.

Precision Winding. In precision winding (Fig. 19 B), the drive impulse is applied to the axis of the spool, resulting in a constant winding ratio. In this case the helix angle α decreases with increasing diameter (see Fig. 19 B). Another characteristic of this type of winding is increasing yarn-winding speed at constant rotational velocity of the spool. Stepwise precision winding is an attempt to combine the advantages of both random and precision winding. Here the winding ratio is changed in a stepwise manner.

The advantages of precision winding are an absence of ribboning and favorable unwinding characteristics. The disadvantages are spool fragility, small spool size, varying thread density, and expensive machinery.

3.3.2. *Twisting*

Twist is a general term applicable to all textiles made by the twisting together of simple yarns or ply yarns. In each twisting step, several starting threads are twisted together to form a single ply yarn (Fig. 20). Distinctions are made between one-stage and multistage twisting (Fig. 20 C, D, E), as well as between Z and S twisting (Fig. 20 A, B). Multistage twist yarns are especially important in the production of cord twist for automotive tires and conveyor belts, including yarns made of polyester, polyamide, or aramide, as well as rayon or steel cord.

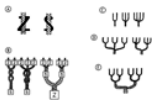

Figure 20. Examples of various types of twists

A) Twist direction (Z or S); B) Twist-direction sequences; C) One-stage twist; D) Two-stage twist; E) Three-stage twist;

A new twist is usually applied in the direction opposite to that of the preceding operation. Successive twisting in the same direction leads to very hard twisted yarns characterized by low extensibility.

The process of twisting serves as a convenient way of increasing strength, improving stretch behavior, reducing unevenness, and achieving special surface and color effects. Ply yarns with special surface effects are widely used in the clothing sector and for home-furnishing fabrics, serving to enliven the appearance of the product. Such materials are made by combining several different take-off yarns. It is also possible to vary periodically the individual yarn-delivery speeds, permitting the introduction of such effects as knops, loops, snarls, or knots.

Twist yarns are made on ring-twisting frames, cablers, double-twisting frames, or Tritec twisters (Fig. 21). These machines differ according to the number of twists applied in a single operation. The

importance of the ring-twisting process (Fig. 21 A) is declining, and the double-twist process is now standard [23]. Two spools placed on top of each other are fed simultaneously to the twisting frame. Alternatively, a “doubling bobbin” might be utilized, where the term *doubling* refers to assembly of the threads prior to twisting. In the case of a cabler (Fig. 21 B), the pre-twisted yarn need not be subjected to additional twist. With the double-twist process, two twists are obtained for every spindle rotation (Fig. 21 C), whereas three twists are achieved in the Tritec twister (Fig. 21 D).

Figure 21. Twisting frames

A) Ring-twisting frame; B) Cabler; C) Double-twisting frame; D) Tritec-twister

3.3.3. *Preparation of the Warp Beam*

A loom requires that there be available a warp system with a defined number of threads, all of which have the same length and thread tension. A distinction can be made between four different production processes with respect to the warp:

1. Weaving from the creel
2. Direct beaming
3. Warping
4. Sectional warping

In the first two processes the number of warp threads is limited by the capacity of the creel, a disadvantage that is avoided with warping or sectional warping.

Warping. In warping, a portion of the total number of required threads is first wound onto a sectional beam [(a) in Fig. 22 A) across the entire width of the warp. As many as about 16 sectional beams are then assembled into a warp beam [(c) in Fig. 22 B]. This step is often combined with the sizing process. The process just described offers only limited patterning possibilities, so it is best suited to large lots with simple patterns.

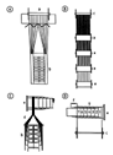

Figure 22. Preparation of a warp beam

A) Warping; B) Assembly; C) Sectional warping; D) Rebeaming

a) Sectional warp beam; b) Creels; c) Warp beam; d) Wraithe; e) Conical warping drum; f) Warp layer pitch; g) Sections of warped threads

Sectional Warping. In the sectional warping process (Fig. 22 C, D), narrow bands of warp threads (g) with the required end spacing and the desired length are wound side by side onto a conical warping drum (e). The conical design prevents slippage of the individual warp sections. As many sections are laid down as required by the desired width. Once sectional warping is complete the warp is rebeamed from the drum to a warp beam (c). The presence of individual sections means that patterning possibilities are virtually unlimited. Therefore this process is appropriate for smaller lots and more complicated warp patterns.

3.3.4. *Sizing*

Sizing refers to a special pretreatment of the warp [24]. A size is a substance that has the function of improving the mechanical properties and loadability of the thread without at the same time reducing its elasticity (⟶ Textile Auxiliaries – Sizing Agents). Demands placed on a sizing material relate to its behavior during weaving as well as the potential for desizing and recycling. Sizing processes can be classified according to the method of application [25]. The most important categories are:

1. Hot-melt sizing
2. Cold sizing
3. Dry sizing
4. Wet sizing
   a. Conventional wet sizing, which involves the application of a water-soluble size followed by a drying process
   b. Solvent sizing, where dissolved size is deposited on the warp by evaporation of a non-aqueous solvent
5. Warp waxing

Sizing materials themselves also fall into several groups:

1. Starch sizes
2. Carboxymethylcellulose (CMC) sizes based on cellulose ethers
3. Protein sizes
4. Synthetic sizes

Starch sizes are used most commonly because they are both inexpensive and environmentally safe. Sizing processes can be further subdivided according to the nature of the size-application procedure (Fig. 23 A – D).

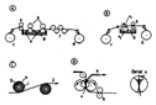

Figure 23. Processes for size application

A) Trough sizing; B) Tangential size application; C) Sinter-roll sizing; D) Foam sizing

a) Feed rolls; b) Squeeze rolls; c) Unsized goods; d) Warp beam; e) Immersion rolls; f) Cylindrical dryer; g) Sizing agent; h) Porous steel cylinder; i) Size film; j) Spun yarn; k) Foam introduction

v_1 = fabric velocity; v_2 = roll velocity

A sized fabric must be subjected to a *desizing* process prior to the finishing stage. Desizing has a decisive effect on the wastewater load in textile production. Two possibilities exist for reducing this load: use of biodegradable sizing materials, or size recovery by ultrafiltration or evaporation.

3.4. Weaving Methods

3.4.1. *Fabric Construction*

As noted previously, a woven fabric is formed by the interlacing of two thread systems, the warp and the weft, and the mode and manner of thread crossing is called the weave or fabric construction. Fabric design can be influenced in many ways by choice of a particular type of weave.

Rapport. Ordinarily, a weave repeats itself in both the transverse and longitudinal directions. The smallest unit which, when duplicated, gives rise to the overall fabric pattern is called the weave rapport. This can in turn be subdivided into a warp rapport and a weft rapport.

Floating. “Floating” refers to the extent of free, unbound thread between two crossing points.

Weave Design. The weave design is a graphic representation of a weave pattern, reproduced on squared design paper. A particular square on the paper is filled if at this crossing point the warp thread is raised and the weft thread is lowered (see Fig. 24).

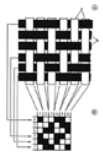

Figure 24. Diagrammatic representation of a particular weave

A) Fabric sample; B) Weave design

a) Warp threads; b) Weft threads

Pattern Drafting. A fabric's weave can also be represented by a pattern drafting (see Fig. 25). This is a code based on the type of weave, information related to warp raising and lowering, the number of adjacent comparable threads, and the racking number (see below). An international standard for pattern drafting is stipulated in ISO 9354. The first number in a pattern drafting specifies the base weave (1 for plain weave, 2 for twill weave, or 3 for satin weave; see below).

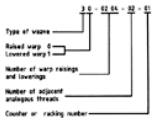

Figure 25. Example of a pattern drafting

Plain Weave. Plain weave represents the simplest interlacing pattern for warp and weft threads, but also the tightest. Plain weave and its simple modifications can be achieved with the use of only two heald frames. Four, six, or more heald frames are used for very closely woven fabrics with a high fiber density. This weave and its simple modifications are illustrated in Figure 26.

Figure 26. Plain weave and its simple modifications

A) Plain weave; B) Weft-rip fabric; C) Hopsack (panama) weave

Twill Weave. Twill weaves are characterized by ribs that run diagonally across the fabric. A distinction is made between Z- and S-rib twill depending on the orientation of the rib. A Z-rib results from a consistent displacement of the first warp thread in the "upper right" direction. The height of the shift is indicated by the racking number. Examples of basic twill weaves are shown in Figure 27.

Figure 27. Twill weave

Satin/Sateen Weave. Satin/sateen fabrics display a closed, smooth, and dense appearance. A feature of this weave is the fact that none of the uniformly distributed weave points (see Fig. 28) touch each other. The distance from one weave point to the next on the following warp thread is called the *counter* (or move number). Permitted move numbers *i, j* may not have a common integral divisor, and they must also meet the following supplementary conditions:

$$n = i + j;\ i \neq j;\ i, j \gg 1;\ i, j \in IN$$

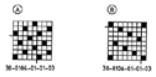

Figure 28. Satin/sateen weave, illustrated with a counter of three

A) 5-End (3); B) 7-End (3)

In a satin weave the warp raisings predominate, whereas in a sateen weave it is the warp lowerings that predominate.

3.4.2. *Loom Technology*

Looms are distinguished according to the nature of the weft-insertion process (Section Weft-Insertion Principles) or the shedding system (Section Shedding Systems) used. The *shedding system* is what makes possible a selective raising and lowering of the warp threads for insertion of the weft thread. *Weft insertion* refers to the process by which a weft thread is transported through the opened shed. The basic structure of a modern loom is illustrated in Figure 29.

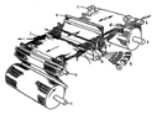

Figure 29. Schematic illustration of a modern power loom

a) Warp beam; b) Back rest; c) Right-hand thread spools; d) Warp stop-motion; e) Shedding; f) Selvage formation (in this case, leno weave); g) Weft insertion; h) Reed; i) Temple; j) Selvage cutter; k) Fabric take-up roller; l) Fabric roller

3.4.2.1. *Shedding Systems*

The task of the shedding elements is to separate the warp threads into an upper shed and a lower shed (Fig. 30, c and d). The weft thread is inserted between these two warp positions. Subsequent changes in warp positions result in binding of the weft threads. The threads are carried and guided in heddles. Various types of shedding mechanisms are described below.

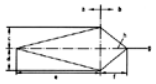

Figure 30. Shed geometry

a) Back shed; b) Front shed; c) Height of the upper shed; d) Height of the lower shed; e) Length of the back shed; f) Length of the front shed; g) Fabric; h) Closed-shed position

Eccentric Dobbies. This shedding system is characterized by a direct, rigid connection between the heald frames and a cam shaft. A particular series of movements repeats itself with each rotation of the cam shaft (Fig. 31 A). Eccentric dobbies are compatible with very high weft-insertion frequencies, but they have the disadvantage of a small weft rapport, and a great deal of effort is involved in resetting the loom for a different weave. Eccentric machines are used mainly for mass-produced fabrics.

Figure 31. Shedding systems

A) Eccentric dobby

B) Dobby: a) Control unit; b) Knives; c) Drawbars; d) Balance arm; e) Spring

C) Jacquard loom: a) Knife box; b) Knives; c) Hook; d) Prism batten; e) Pattern cylinder; f) Needles; g) Compression spring; h) Bottom board; i) Cable; j) Harness cord

Dobbies. With a dobby, the position to which each heald frame should be moved is ascertained from an information carrier (punch card, computer software) prior to every weft insertion (Fig. 31 B). Theoretically, therefore, the weft repeat pattern is unlimited in variety. A change in weave is achieved simply by introducing or activating a new information carrier. A *double-lift* dobby inserts two weft threads per rotation of the main drive shaft. Double-lift dobbies may be based on the Hattersley, positive, or rotary principle.

Jacquard Looms. Jacquard looms (Fig. 31 C) permit individual control over each warp thread via a set of harness cords (j). Any desired type of pattern can be produced with this process. A selection mechanism establishes whether or not the knife box (a) takes with it the hook (c) attached to a particular heddle. A selection occurs via needles (f) or electromagnets.

3.4.2.2. *Warp Let-Off Systems*

The warp let-off (Fig. 32) has the functions of ensuring the availability of an adequate supply of warp materials and keeping the warp-thread tension as constant as possible. Together with the movable loom back rest [(g); see also Fig. 29], the warp let-off is expected to help compensate for any thread elongation caused by shedding [26].

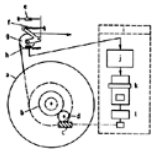

Figure 32. Schematic diagram of a warp let-off control system

a) Warp beam; b) Warp-beam gear; c) Worm; d) Worm gear; e) Damper; f) Tension spring; g) Back rest; h) Position sensor; i) Let-off device; j) Controller; k) Motor; l) Drive

The principle of *passive* warp let-off is equivalent to applying a brake on the warp beam. *Active* warp let-off implies that the warp beam itself is driven. The system can be so adjusted as to maintain constant fabric withdrawal as well as either a constant distance between adjacent weft threads (negative fabric withdrawal) or a constant center-to- center weft-thread distance.

3.4.2.3. *Weft-Insertion Principles*

A distinction is made between conventional weft insertion with a shuttle, unconventional or shuttleless systems (missile, rapier, and jet looms), and multiphase weft-insertion systems. Unconventional looms always insert an end of the thread. The inserted weft yarn is subsequently tightened and separated from large, fixed packages, usually cones or cheeses. With a conventional shuttle loom, however, one long continuous thread is inserted repeatedly after being bent around as necessary at the edge of the fabric. A traditional bound edge (selvage) is formed only by a shuttle loom.

The weft-insertion efficiency of a loom is expressed in m/min, obtained by multiplying the rotational velocity by the fabric width. The progressive development in weft-insertion efficiencies as a

function of time is shown in Figure 33 for various insertion systems. As a result of the very high and variable unwinding rates from crosswound packages, weft storage systems are used as prewinding devices.

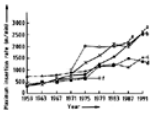

Figure 33. Development of weft-insertion efficiencies as a function of time and insertion mechanism, based on presentations at the annual International Textile Machinery Exhibit (Internationale Textilmaschinen-Ausstellung, ITMA) [27]

a) Water; b) Air; c) Missile; d) Rapier; e) Multiphase; f) Shuttle

Shuttle Looms. In this type of loom (Fig. 34 A), a shuttle, which contains a spool wound with yarn (the weft spool), is pushed through the shed by a beating device. With each weft insertion, the entire mass of the shuttle (b) and the weft spool must be accelerated and subsequently brought to rest.

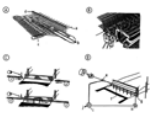

Figure 34. Weft-thread insertion techniques

A) Shuttle; B) Missile; C) Rapier; D) Air jet

a) Warp; b) Shuttle and shuttle spool (filling bobbin); c) Traditional selvage; d) Reed; e) Main jet; f) Echelon jets; g) Weft stop-motion device; h) Microprocessor; i) Pressure-regulating valve; j) Compressed air

Missile Looms. With a missile loom (Fig. 34 B) the weft thread is clamped to the end of a projectile (the “missile”) that is shot through the shed. The weft is then tightened and separated from the external weft supply. The missile itself must later be restored to its original position externally, which means that there are always several missiles in circulation.

Rapier Looms. Looms of the rapier type (Fig. 34 C) are representative of form- closed weft-insertion systems. The head of one rapier (the giver) takes a thread end from the weft bobbin and carries it into the middle of the shed where the thread is then passed along to a second, opposing rapier (the taker). Weft insertion of this type is subject to control and regulation throughout the insertion process. The rapier insertion system is very flexible, and suitable for sensitive materials as well as both fine and coarse fabrics.

Jet Looms. In the case of a jet loom (Fig. 34 D), a distinction must be made on the basis of the insertion medium between air-jet and water-jet devices. In other words, the weft thread may be transported through the shed by means of either air pressure or a water jet. A main jet (e) is supported by echelon jets (f) located at the reed. Jet-weaving machines offer the highest weft-insertion efficiencies, but they are limited in flexibility with respect to both material and fabric width.

Multiphase Weaving Machines. In a multiphase weft insertion system, several weft threads are inserted simultaneously. A series-shed loom may have several sheds arranged sequentially in the warp direction, each with its own weft-insertion device. By contrast, with a wave-shed weaving machine the warp is shedded with a wave-like motion to permit passage of a succession of “shuttles,” each of which lays in its own length of filling. The beat-up must also be accomplished in a wave-like manner, realized by a rotary slay. Multiphase weaving has not gained broad acceptance because of poor fabric quality and lack of flexibility.

3.4.2.4. *Selvage Formation*

A firm, non-fraying selvage is required to permit such further processing as textile finishing. With all unconventional weft-insertion processes it is necessary that the weft threads be specially tied at the edge. If possible, the selvage should retain the same thickness and length as the crude fabric. A satisfactory selvage can be achieved in several ways, including:

1. Insertion of an auxiliary thread (*inserted* selvage)
2. Binding of the weft-thread ends (*tuck-in* selvage)
3. Twisting of the warp threads (*leno* selvage)
4. Gluing or welding of the edges

3.4.3. *Fabric Defects*

Fabric Defects in the Weft Direction. Restarting a loom can lead to variations in the distance between individual weft threads. Depending on the origin of the problem, a distinction is made between flaws of the *starting*, *beating*, *regulating*, and *shed* type. The extent of observable defects is a critical function of the behavior of the warp let-off device. Defects attributable to improper weft tension or weave errors are less common.

Fabric Defects in the Warp Direction. Fabric defects in the warp direction can be a consequence of weave errors, incorrect drawing in the heald frames or the reed, or varying tension on the warp threads. Flaws in the selvage also fall in this category.

Fabric Defects Unrelated to Weave Direction. Such defects include soiling, interwoven foreign matter, holes, and tears. Defects in this class are usually a result of operator carelessness or incorrect machine settings.

3.4.4. *Special Weaving Methods*

Special weaving methods include the multiphase weft-insertion looms described in Section Weft-Insertion Principles and systems adapted to the production of special fabrics.

Circular Looms. The circular loom is a particular type of multiphase loom. Here the shed moves in a circle, and the loom is in fact similar in appearance to a circular knitting machine (see Section Circular Knitting Machines). Such a shed is formed by circulating cams and needles.

Terry Looms. A terry loom (Fig. 35) processes two warps simultaneously, a base warp (a) and a pile warp (b), into a single fabric. The base warp is stretched tightly, while the pile warp is very loosely stretched. A group of three or more wefts is inserted one after the other and then beaten as a group. During beating, the weft threads slip over the tightly stretched base warp, thereby pushing up the pile warp and forming loops. The resulting loops are sometimes cut, producing a cut pile.

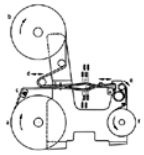

Figure 35. Cross-section of a terry loom

a) Base warp; b) Pile warp; c) Tension beam for the base warp; d) Tension beam for the pile warp; e) Fabric take-off; f) Fabric roller

Velour and Plush Looms. With machines of this type, loops or naps in the pile warp are introduced with the aid of pile wires. As the pile wires are withdrawn, all the loops are cut with a knife, again producing a cut pile.

Carpet Looms. Double-velvet or double-plush carpet fabric is produced by machines in this category. The pile is achieved by attaching the upper fabric only loosely to the lower fabric via pile threads. These pile threads are subsequently cut with a knife after weft beating.

Large amounts of material are also produced on the basis of the *tufting process* (especially wall-to-wall carpeting). In this case each pile thread is drawn in a needle of a special type of sewing machine, and the pile loops are stitched together into a base fabric. The resulting loops can either be left closed (looped carpeting), or they can be cut (cut-pile carpeting). The back of the carpet is coated in order to fix the loops.

Shaping Looms. Almost any type of surface can be achieved by varying the distance between weave points in both the weft and warp directions (Fig. 36). A special shaping loom for this purpose has been developed at the Institute for Textile Technology of the Aachen University of Technology (Federal Republic of Germany).

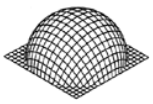

Figure 36. Three-dimensionally shaped fabric

[Top of Page]

4. Braiding

Alexander Büsgen

4.1. Introduction

Machine-made braids exist in many different forms, including cords, tubes, band-like flat products, and three-dimensional objects. All have in common a diagonal thread- crossing and the principle of independent thread motion upon which their production is based.

The first report of a mechanical braider dates back to 1748 when an Englishman, THOMAS WALFORD, received a patent (No. 638) for a device designed for the creation of braids. The first braider made of iron instead of wood was built in 1767 by a mechanic from Barmen (Germany) by the name of BOCKMÜHL [31]. Most of the early braided products were shoelaces and shoe straps, but mechanical braiding was gradually extended to other fields of application. These include articles of clothing, household fabrics, and increasingly, industrial braids, such as sheathing or sealing

cords. In the future, braided three-dimensional structural elements may be useful in light construction.

4.2. Definition of a Braid

The term “braiding” refers to the continuous twisting of a set of threads or equivalent elements in the direction of propagation. The resulting textile product consists of a single system of threads twisted diagonally [32]. Uncrossed threads called “cores” or “standing ends” may also be included in the braid. According to DIN 60 000, braids (Fig. 37) are defined as “flat or three-dimensional products with a regular thread density and a closed appearance whose braid (lace) threads cross over in a direction diagonal to the edges of the product.”

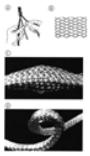

Figure 37. The technique of braiding

A) Hand braiding (“right rib”); B) General form of a braid; C), D) Examples of braided material

4.3. Production of Braids

4.3.1. *Principles of the Mechanical Braiding Technique*

Machine braiding has its origin in the hand-braiding technique, which requires that the material to be braided consist of at least three thread packets. In the course of braiding, the left hand repeatedly moves strands from the upper left to the lower right position. The resulting strand arrangement is known as a “left rib.” If the right hand is used to move strands from the upper right to the lower left, what results is a “right rib” (Fig. 37 A).

During plaiting of the thread strands each hand executes a rotational motion about the central thread. In creating a right rib the left hand moves counterclockwise, while the right hand moves clockwise. Mechanical braiding is accomplished similarly by means of *horngears* rotating in counterclockwise and clockwise fashion (Fig. 38), simultaneously carrying the thread strands around each other just as would the hands. Notches are distributed along the periphery of the horngear. Each notch is capable of accepting a spool carrier, the *braiding carrier*, and transporting it as the horngear revolves. Beneath the horngears is a surface into which have been cut guide grooves known as *races*. It is through these races that the lace butts are led. The races also establish paths for the braiding carriers during rotation of the horngears.

Figure 38. Horngears

A) Gliding system; B) Carried system

4.3.2. *Functional Elements of a Braider*

The basic design of a braider is illustrated in Figure 39 (see also Fig. 37 B).

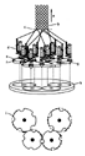

Figure 39. Schematic representation of a braiding machine

a) Direction of propagation; b) Braided strand; c) Braiding point; d) Braiding carrier; e) Horngear; f) Notch; g) Sliding shoe; h) Race; i) Path followed by the braiding carrier

Horngears. With few exceptions, braids are always produced with the aid of rotating horngears. In most braiders these horngears are arranged in a circular pattern. The differences between braiders for ribbon and for round products are clarified in Figure 40. In a ribbon machine the braiding carriers are not passed along to the next horngear at the point of reversal, but are instead returned, and in a direction opposite to that of the previous motion.

Figure 40. Horngear arrangements for (A) a ribbon braider and (B) a round braider

a) Point of reversal; b) Forward course; c) Reverse course; d) First braiding- carrier series; e) Second braiding- carrier series

From a design standpoint, a distinction must be made between two horngear systems (Fig. 41) [33]. The *gliding* (Barmer) system relies upon a guide table mounted above the horngears. The upper flange at the base of the braiding- carrier foot (the lace butt) secures the carrier to the guide table as the entire assembly glides along the race. The lower part of the lace butt projects into the notches. Continuous friction associated with the gliding system requires intensive lubrication, and severely limits the operating speed of the machine.

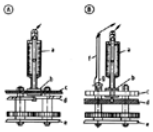

Figure 41. Schematic illustrations of two types of braiding systems

A) Gliding (Barmer) system: a) Braiding carrier with braid spool (bobbin); b) Lace butt; c) Upper plate; d) Impeller; e) Lower plate

B) Carried system: a) Braiding carrier with braid spool (bobbin); b) Lace butt; c) Impeller; d) Upper plate; e) Lower plate; f) Standing-end spring; g) Standing end

The *carried* system uses horngears located above the guide table. Here, the lace butt surrounds the notch, which means that the braiding carrier is carried along by the horngear. Only the lower portion of the braiding carrier, the sliding shoe, is subject to friction as the braiding carrier passes along through the race. Considerably higher horngear rotation rates can therefore be achieved relative to a machine based on the gliding system.

The size of the horngears is frequently specified as one characteristic of a braider. The diameter of the horngear usually corresponds to the distance between the axes of adjacent horngears, and is called the “pitch.” Braiders typically have pitches in the range of a few centimeters to 1.5 m.

Braiding Carriers. As a rule, the yarn spools themselves are not guided along the notches, being confined instead to the braiding carriers, a measure intended to ensure proper uninterrupted unspooling. The braiding carrier plays an important role in the braiding process. Its most important functions are:

1. Mating with an interchangeable spool
2. Controlling the spooling-off process as the braiding carrier moves
3. Maintaining an adjustable level of thread tension
4. Interrupting the machine in the event of thread breakage

Braiding carriers come in many different sizes and shapes. The thread spool is usually joined such that its axis is aligned vertically, although a horizontal spool alignment is preferred in the case of a material that is extremely sensitive to bending.

During the braiding process the braiding carrier follows a wave-like course. The changing distance between the carrier and the *braiding point*, where the threads are compacted to form a braid, must be compensated for by a special thread-tightening device, consisting normally of thread-tightening weights or spring-loaded compensator levers. Tension on the thread holds an interrupter located on the braiding carrier in an upright position. If the spool becomes empty, or if the thread breaks, this interrupter falls in such a way as to activate a mechanical or electromagnetic control that shuts down the machine.

A special device, usually a pin or a catch rod, prevents unintentional unwinding of the thread (as a consequence of the rotary motion of the braiding carrier) by locking the grooved edge of the spool. Spool release to permit renewed unwinding of the thread is usually controlled by tension. A conventional braiding carrier with the functions described is illustrated in Figure 42. The lever carrier (c) is lifted by tension of the thread acting against the force of a spring. This pushes down on a pin that pulls a catch rod (d) out of the spool. The spool can rotate to release thread only when this catch rod is completely extracted.

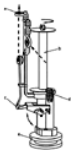

Figure 42. Schematic representation of a braiding carrier with faller arm and catch rod

a) Thread eyelet; b) Spool; c) Lever carrier; d) Catch rod serving as a spool brake; e) Lace butt

Races and Track Switches. It is the race that controls the path of the braiding carrier. Most often this consists of a set of grooves cut in a guide table into which the sliding shoes of the braiding carrier are in turn inserted. When a braiding carrier is pushed from one horngear into the operating zone of an adjacent horngear, the sliding shoe must pass over a race- crossing between the two gears. One alternative to such a crossing is a track switch that blocks the race. Track switches can be controlled by means of cogged repeat rolls or by a Jacquard device. A more recent development is the electronically controlled pneumatic track switch, used especially in braiders for the production of three-dimensional braids [36].

Material Take-Off. Above the center of the machine, threads are compacted at the so- called braiding point to form a braid (see Fig. 39). The rate of take-off must be adjusted to match the braiding rate of the braiding carriers. This rate influences both the thread angle in the braid and the extent of thread compacting. Rapid take-off results in a sharp crossing angle and a relatively loose braid. A slow rate of take-off leads to an obtuse crossing angle and a firm, compacted thread structure.

A three-roll take-off unit is frequently employed for ribbon and lacing. The band-like braid is wound around three pressure rolls, the speeds of which can be regulated either with change gears or electronically. Braids that are flexurally rigid or sensitive to bending are withdrawn via a disk with a special adhesive coating.

Standing Ends. It is not essential that all the threads of a braid participate in the braiding process on braiding carriers. Individual threads are often introduced in unbraided form, typically as a way of filling hollow spaces in the braid or adding elasticity. Such unbraided threads may constitute either standing ends or cores.

Standing ends are threads that are fed through the bored centers of the horngears and incorporated individually into the braid as the braiding carriers pass around the horngears. This technique is often utilized to make elastic tape, where prestressed rubber threads are introduced in the form of standing ends to add elasticity. Inelastic standing ends are used to provide additional support to the interlacing, or for patterning.

Standing ends introduced through the center of the braiding machine itself are called *cores*. A sheath is created around a single core by all the circulating braiding carriers simultaneously. Filling materials, filling threads, insulating materials, wood and cork forms, electrical cables, lead wires, lead balls, and tubes have all been used as cores in braided products.

4.3.3. *Mechanical Systems*

Horngear Systems. The most common type of machine for making braids uses horngears to pick up the braiding carriers, push them through a race, and pass them on to a neighboring horngear (see Figures 39 and 40). The number, arrangement, and notch patterns of the horngears determine the types of braid that can be produced. Braiders for flat and round braided products usually entail a circular horngear arrangement, whereas packing braiders use square horngear arrays. There are a great many variants as well, combining, for example, circular and flat braiders with individual track switches for shifting the races [34], [35].

Three-dimensional rotational braiding can be used to produce industrial three-dimensional braids in the form of components [36], [37]. Track switches direct the braiding carriers on individual tracks through a complex array of horngears. This technique permits the introduction not only of cores and standing ends, but also shorter *zero-degree threads*, and for any desired period of time. For this purpose, individual horngears are selectively disengaged at various junctures. Computerized simulation and operational control programs are required to ensure collision-free braiding and to control the activities of the machine. This techniques makes it possible to produce braids with various cross-sectional shapes.

Horn's System. Braiders operating according to the Horn system avoid the use of horngears. The general principle is illustrated in Figure 43. Threads are run from the inner spool circle directly into the braid. Threads from an outer spool circle are led over and under the inner circle with the aid of planetary wheels and guide levers. These machines are characterized by a very high production rate, but they also take up a great deal of space. The spools remain the same distance from the braiding point throughout their rotation, which eliminates the thread compensation required with horngear machines. Horn machines are especially suitable for braiding around a core made of a material that is sensitive to bending, such as wire.

Figure 43. Schematic representation of a braider based on the Horn system

a) Planetary wheel or guide lever; b) Core

Cartesian 3-D Machines. To date, Cartesian braiders have been investigated only on a laboratory scale for the production of industrial three-dimensional braids. The braiding carriers in this case are pushed back and forth on linear tracks, for example by pneumatic cylinders, as a way of producing braid profiles with a constant cross-section. The nature of a Cartesian braider is illustrated in Figure 44.

Figure 44. A Cartesian braider

a) Framework; b) Table; c) Material take-off; d) Braid eyelet; e) Braiding carrier; f) Pneumatic cylinder

4.4. Interlacing Techniques

A braid interlacing describes the course of a particular thread and its crossing path with respect to threads coming from the opposite direction. The nature of the interlacing depends on the number of notches in the corresponding horngear and the distribution of the braiding carriers among these notches. Depending on the number of notches and the braiding- carrier distribution, a given thread will cross over one or more threads from the opposite direction during each half-rotation of the horngear.

The most common types of braid interlacing are single braid (each thread crosses over and under one opposing thread), double braid (thread crosses over and under two opposing threads), and triple braid (thread crosses over and under three opposing threads).

The interlacing can be altered by changing the notch and braiding- carrier occupation (or fitting). The *notch fitting* establishes which notches are occupied by braiding carriers and which remain empty. The *braiding- carrier fitting* indicates which braiding carriers are to be provided with spools. A braiding carrier lacking thread is called a “missing” or “empty” spool.

4.5. Materials Used in Braiding

The choice of materials for braiding depends on the intended application. Demands are placed on the braiding yarn on one hand by the tensile force of the thread, which can lead to frequent thread breaks in larger machines, but also by deflection and friction points in the braiding carrier. Very brittle industrial fibers are most subject to damage by small radii and frequent deflections.

Cotton is often used for making shoelaces and also tapes and ribbons because of its good wet-tear resistance. Viscose and wool are used for trim materials and borders because of the surface effects they are capable of producing. The tear resistance of dyed wool is very low, so wool yarns intended for braiding are mixed with polyamide. Polyamide and polyester are both found in a wide range of applications.

Linen can be regarded as a “rediscovered” natural raw material, and it is being increasingly braided into trimmings. The processing of the corresponding yarns is problematic, however, and it requires air humidification. Polypropylene is braided into ropes and cords. Glass, carbon, and aramide fibers are processed into packing braids and braided three-dimensional structures.

4.6. Types of Braids and Their Applications

Braids are classified on the basis of a numerical system and, in most cases, a more detailed description of the character of the product. The numbering begins with the number of braiding carriers required for manufacture, followed by an *interlacing index* that indicates the “braidedness” of the product. For example, a ribbon made from nine braiding carriers with a quadruple-braided interlacing is described as “9/4.”

Since this system provides no clear description of the braid itself, a descriptor is often added to clarify the material's character, use, form, or pattern. Examples are [31]:

9/4 ribbon
32/2 herringbone
44/2 shoelace
33/3 curved ribbon

4.6.1. *Ribbon Braids*

The terms “heald” refers to a band-like, flat braid. Examples of this type of braid are bobbinet, tailor's braid, elastic ribbon, and high-profile ribbon. Braiding in this category most often serves as a trim or border for outerwear. Compared with a woven band, braided ribbon is especially soft, and it can be sewn in tight curves without creasing. Elastic ribbons are made by including elastomers in the braid. As a rule, double-braided interlacing is preferred for tight enclosure of the rubber thread. Ric-rac braid and bow braids are used as trimmings for folk dress and children's clothes.

4.6.2. *Round Braids*

Round braids are also referred to as round or hollow *cords.* They include all braided articles with a circular or approximately circular cross-section, in contrast to flat braid. The appropriate interlacing pattern can usually be obtained without special control over the braiding carriers. Examples and applications of round braids include cords and laces for trimmings and for the shoe industry; drawstrings for curtains and underwear, rubber cords for the haberdashery, clothing, and sportswear industries; and cables, tubes, and ropes.

4.6.3. *Packing Braids*

Originally, packing braids were classified along with round braids. However, the corresponding cross-sections tend no longer to be round, but rather square (see Fig. 37 C). The horngears in this case are arranged not in a circular pattern, but as a surface array below the braiding point, so the threads are also braided through the center of the cross-section. Packing braids require 2×2, 3×3, or 4×4 horngear arrangements, and the races and braiding- carrier tracks are not adjustable. The square cross-section of such a braid is constant, with edges ≤ 30 mm in length.

The main application of packing braids is as stuffing-box sealings [34]. Because of their great flexibility packing braids are especially effective for narrow-radius seals against high pressures. Pumps and propelling shafts are examples of devices that often entail the use of such seals.

4.6.4. *Three-Dimensional Braids*

Braided three-dimensional (3-D) structures are particularly valuable in light construction (see Fig. 44) [36], [37]. In combination with appropriate synthetic materials, 3-D braids can be effectively processed into ribs, stringers, or patterns. The diagonal fiber alignment in the braid is advantageous because it provides stability against torsional loads. Development of the corresponding mechanical technology and subsequent processing methods is still in its early stages. Progress to date coupled with the potential for great flexibility in the production of braids similar to construction parts suggest a promising future for the use of braids as structural components in many areas of technology.

[Top of Page]

5. Knitting

Klaus-Peter Weber

5.1. Introduction

Knitted material has been produced manually for thousands of years. The Egyptians, Incas, and other ancient advanced civilizations accomplished their knitting with rods. Indeed, it was not until 1589 that the Reverend WILLIAM LEE, a theologian from Calverton, England, invented a hand frame, thereby making it possible for the first time to mechanize the knitting process in the form of *weft knitting*. This device was capable of producing about 16 times the output of a hand knitter. In 1775 a knitter by the name of JOSIAH CRANE from Nottingham, England, patented the first *warp* knitting frame, and in 1863 another theologian, ISAAC W. LAMB, an American Baptist, invented the first *flat* knitting machine.

The production of knitted goods has increased tremendously in the final decades of the twentieth century with the increasing use of synthetic fibers, because these materials are ideally suited to processing on knitting and hosiery machines. CAD-supported devices for manufacturing knitted fabrics provide extremely high output rates and the potential for a wide variety of patterns. Current equipment offers 500 000 times the output capacity of a hand knitter. Knitted materials are used for all types of clothing, household fabrics, and industrial textiles.

5.2. Classification of Knitted Fabrics

Knitted stitches consist of thread loops that hang inside one another. Each stitch has upper and lower interlacing points, and consists of a *head*, *shanks*, and *feet*. A distinction is made between stitch loops (which have only lower interlacing points) and thread loops (with no interlacing points at all), as shown in Figure 45 A. A row of stitches aligned side-by-side forms a stitch course (SC), whereas stitches arranged one above the other (i.e., in the direction of production) constitute a wale (W), as illustrated in Figure 45 B.

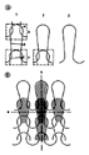

Figure 45. The basic characteristics and terminology of knitted stitches

A) Stitches and loops: 1) Knitted stitch; 2) Stitch loop; 3) Thread loop

a) Lower interlacing point; b) Upper interlacing point; c) Feet; d) Shanks; e) Head

B) Stitch combinations: a) Stitch course; b) Wale

The quality of a knitted fabric can be characterized in terms of two parameters:

Density of the stitch course = (no. of SC)/(unit length)
Density of the wale = (no. of W)/(unit length)

In principle, all knitted materials can also be classified according to the nature of the thread take-off as either single-thread (one thread crosswise) or warp-thread (with warp threads lengthwise) knits (Fig. 46).

Figure 46. The two fundamental types of thread path in knitted work

A) Single-thread knitware; B) Warp-thread knitware

5.3. Technology of Stitch Formation

Most modern knitting and hosiery machines utilize bearded, latch, or compound needles.

The *bearded needle* (Fig. 47 A), invented in 1589 by LEE, consists of an extended hook (d) with a tip (e), a shaft (b) containing a groove (f), and a foot (c). At the start of stitch formation (Fig. 47 B), with the stitch loop already in the hook (1), the needle is pushed forward ("enclosing the loop"), and a thread is placed in the hook (2). The needle is subsequently retracted (3), and a press (or presser bar) pushes the tip into the groove ("pressing") so that a stitch loop can be cast onto the tip ("casting on"). The thread is then pulled through to form a stitch loop (4), and the stitch loop is knocked over the hook, producing a new stitch ("knock-over" or "casting off"). Bearded needles are used in both weft knitting and warp knitting machines.

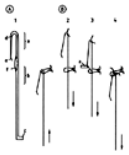

Figure 47. Knitting with a bearded (or spring) needle

A) Bearded needle: a) Head; b) Shaft; c) Foot; d) Hook; e) Tip; f) Groove

B) Principle of stitch formation with a bearded needle: 1) Enclosing; 2) Thread laying; 3) Pressing and casting on; 4) Knocking over; a) Press

The *latch needle* (Fig. 48 A), invented by TOWNSEND in 1847, has a short hook (d), a hinged latch (e) in the shaft (b), and a foot (c). Stitch formation (Fig. 48 B) starts with expulsion of the needle (1). The stitch loop slides out of the hook, over the latch, and onto the shaft of the needle (enclosing the loop). A thread is then laid into the hook (2) and the needle is pulled back with the foot. During this process, the stitch loop closes the hook by rotating the latch (3). It then slides onto the latch ("casting on") and is subsequently cast off over the hook (4). The stitch loop thereby gives rise to a new stitch, and the thread produces a new stitch loop. The latch needle is used in both standard knitting machines and warp knitting machines.

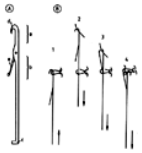

Figure 48. Knitting with a latch needle

A) Latch needle: a) Head; b) Shaft; c) Foot; d) Hook; e) Latch

B) Principle of stitch formation with a latch needle: 1) Enclosing; 2) Thread laying; 3) Casting on; 4) Knocking over

The *compound* or *slider needle* (Fig. 49 A), invented by WILCOMB in 1924, consists of a hook with shaft and groove and, in addition a separate slider (d). The slider is subject to independent motion in the shaft in such a way that it opens and closes the hook. In forming a stitch (Fig. 49 B), the needle is pushed forward with the slider retracted (1), and the stitch loop is moved off the hook onto the shaft (enclosing the loop). After thread laying (2), the needle and the slider are moved independently (3) in such a way that the hook is closed and the stitch loop is able to slip onto the slider (casting on). Finally, the stitch loop is pulled over the hook (knocking over), a new stitch is formed, and the thread produces a new loop (4). Slider needles are used in warp knitting machines as well as standard knitting machines (although not extensively at the present time).

Figure 49. Knitting with a compound needle

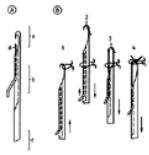

A) Compound needle: a) Head; b) Shaft; c) Foot; d) Slider

B) Principle of stitch formation with a compound needle: 1) Enclosing; 2) Thread laying; 3) Casting on; 4) Knocking over

To further ensure secure and satisfactory stitch formation, small shaped-steel rods (sinkers) are used in knitting and hosiery machines to improve needle performance, especially in the clearing, casting-on, and casting-off operations (Fig. 50). With respect to the criteria needle mobility and thread take-off, a distinction is made in the production of knitted materials between standard knitting, weft knitting, and warp knitting processes.

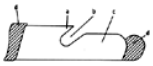

Figure 50. The characteristics of a sinker

a) Nose; b) Throat (enclosing the loop); c) Shaft (casting on and knocking over); d) Unit

The *standard knitting process* (LAMB, 1863), illustrated in Figure 51 A, is characterized by independently moving needles (usually latch needles) that accomplish the various phases of stitch formation successively, and are served with one stretched thread (single-thread take-off). Stitches are also formed from the thread sequentially. Depending on the arrangement of the needles, a distinction can also be made between *flat-bed* and *circular* standard knitting machines.

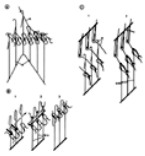

Figure 51. Schematic illustrations of the principles underlying various knitting processes

A) The standard knitting process: a) Independently driven latch needles; b) Single-thread feed

B) The weft knitting process: 1) Thread laying and loop sinking; 2) Pressing and casting on; 3) Knocking over; a) Collectively driven bearded needles

C) The warp knitting process: 1) Laying of the warp thread; 2) Stitch formation; a) Warp-thread placement in front of and behind the needles; b) Latch needles moved simultaneously; c) Guide bar with thread guides (guide needles)

In the *weft knitting process* (LEE, 1589), shown in Figure 51 B, all the needles (bearded needles) move in concert (the principle of weft knitting), again with a single thread presented to the hook side of the needles. Unless it is highly elastic, a single stretched thread cannot be simultaneously transformed into stitches by several needles all moving at the same time, so the thread must first be preformed into a series of loops ("sinking the loops"). This is followed by simultaneous stitch formation, with all the needles moving together. In contrast to a flat weft knitting machine, stitch formation on a circular weft knitting machine (with needles arranged in a circle and moved all at once) occurs in a sequential fashion. This process is currently of little importance.

The *warp knitting process* (CRANE, 1775), illustrated in Figure 51 C, also uses bearded, latch, or compound needles in a planar arrangement and all moving together. The threads in the thread chains are usually processed in the longitudinal direction (warp-thread take-off). Threads from the thread chains are transferred in front of and behind the needles with the help of guide bars equipped with thread guides. At the same time, the threads are formed into stitches by needles all moving in concert.

5.4. Structural Elements and Structures of Knitted Fabrics

Knitted materials consist of various structural elements: stitches, loops, wefts, floatings, and filler threads. The primary structural element of any knitted material is the stitch, although other structural elements may be present as well depending on the technology in question, the structural group (see below), and the pattern.

A stitch (Fig. 52 A) can be either open or closed. The head and the shank of the stitch constitute the needle loop, whereas the foot forms the sinker loop. A distinction is also made between the two sides of a stitch: the "face loop" (or *face stitch*, R) and the "reverse loop" (*reverse stitch*, L), recognizable by the form of the lower interlace point (Fig. 52 B). On the face side, the feet are underneath the head of the preceding stitch, whereas on the reverse side they are above it. This distinctive feature allows one to divide knitted materials into three structural groups:

1. Single-face knits (RL)
2. Double-face knits (RR)
3. Purl knits (LL)

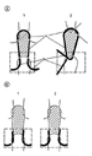

Figure 52. Stitches and the notion of "sidedness"

A) Two basic stitch types: 1) Open stitch (no foot crossing); 2) Closed stitch (foot crossing); a) Needle loop (head and shank); b) Sinker loop (foot bonding)

B) "Sidedness" with respect to a stitch: 1) Face loop ("R"); 2) Reverse loop ("L")

The basic structure of a *single-face* fabric (Fig. 53 A) presents on one face of the fabric the face-loop sides of the stitches, and on the other only the reverse-loop sides. A characteristic feature is a common direction of passage for all the stitches, possible only with a single-needle system (RL technique).

Figure 53. Different types of knits

A) 1) Single-face knitted fabric (face loop); 2) Single-face warp-knitted fabric (reverse loop)

B) Double-face knitted fabric: a) Face wale: b) Reverse wale

C) Purl-stitch knitted fabric: a) Course of face stitches; b) Course of reverse stitches; c) Double-latch needle

In *double-face* fabrics with the RR structure (Fig. 53 B), the face and reverse sides of stitches alternate along a given stitch course. Nevertheless, only the face sides of the stitches are visible on either face of the fabric because the perpendicular elasticity of the material draws the wales together, concealing the reverse wales. Warp knits of the RR type also display only face stitches on both sides of the fabric. Two needle systems with two directions of passage are required for producing RR knits (the RR technique). Turning off one of the needle systems on an RR machine permits the production of RL materials, and LL materials (described below) can also be produced provided stitch transfer is possible (see Fig. 58).

The basic structure of an LL knit (Fig. 53 C) exhibits alternating face and reverse stitches in the wale. The longitudinal elastic behavior of these materials results in the appearance of mainly reverse stitches on both faces of the fabric. Double-latch needles are used in the production of such knits. These needles, activated by selectors, work alternately in different needle beds, causing the direction of passage to be repeatedly reversed (LL technique). Properly sorting the needles in advance in one or two needles beds makes it possible to use an LL device for making RL and RR knits as well. Technology of this type exists, but it is of little commercial significance.

The following structural elements are incorporated or tied into stitches or other structural elements to establish desired patterning or structural properties of the knitted fabric:

1. A loop (Fig. 54 A) consists of a thread that has been placed in the needle hook along with the stitch-forming thread, and which is bound in place by the stitches.
2. A floating (Fig. 54 B) is a length of thread in a single-thread knit that skips at least one wale, and may skip successive stitch courses. In a warp knit this length of thread must skip at least one stitch course, and may extend over the wales.
3. A weft (Fig. 54 C) is a thread laid in the direction of the course. In warp knits, a distinction is made between a partial weft and a full weft.
4. A filler thread (Fig. 54 D) is a thread laid in the direction of the wale. Single-face knits exist with standing threads, but they are rare.

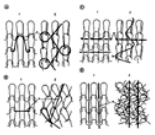

Figure 54. Loops, floatings, wefts, and filler threads

A) Loops: 1) RL Single-face knitted fabric (reverse loop); 2) RL Warp-knitted fabric (reverse loop); a) Loops

B) Floatings: 1) RL Single-face knitted fabric (reverse loop); 2) RL Warp-knitted fabric (reverse loop); a) Floatings

C) Wefts: 1) RR Single-face knitted fabric; 2) RL Warp-knitted fabric; a) Full weft; b) Partial weft

D) Filler threads: 1) RL Single-face knitted fabric (reverse loop); 2) RL Warp-knitted fabric (reverse loop); a) Filler thread

5.5. Classification of Stitch-Forming Machines

Scheme 2 illustrates classification of knitting machines based on needle mobility, thread take-off, needle arrangement, and structural group. These include:

1. RR and LL flat-bed knitting machines
2. RL, RR, and LL circular knitting machines
3. RL flat weft knitting machines (Cotton system)
4. RL and RR warp knitting machines

Scheme 2. Classification of the machines capable of producing knitted goods

5.5.1. *Flat-Bed Knitting Machines*

Flat-bed knitting machines are used mainly for outerwear — sweaters, vests, and dresses — with colored or textured patterns. Flat-bed knitting technology has developed tremendously in the last two decades with the help of electronic controls and computer-aided design (the CAD technique). Electronic control and design programs have made it possible to go beyond the limits previously imposed by mechanical control and patterning, and the potential for diversity is unlikely ever to be exhausted by knitting designers.

The knitting zone of an RR flat-bed knitting machine like that shown in Figure 55 contains two flat needle beds mounted in a roof-like arrangement (Fig. 56 A). Latch needles are independently driven out of channels in these beds and subsequently retracted by the action of moving cams. For this purpose, each cam (Fig. 56 B) consists of driving and retracting segments that constitute a channel for guiding the needle foot during stitch formation.

Figure 55. Electronically controlled flat-bed knitting machine

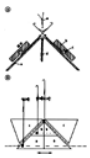

Figure 56. Characteristic features of an RR flat knitting machine

A) Knitting zone: a) Needle bed; b) Cam; c) Needle butt in the cam track; d) RR knit; e) Thread guide

B) Knitting cam: a) Lower driving segments; b) Upper driving segments; c) Retracting segments;

– · – · – · – · Needle path for knitting (stitch)

····· Needle path for capturing (loop)

- - - - - - - - - Needle path when not knitting (RL, floating)

⊗ ⊙ (here and elsewhere) symbolizes motion perpendicular to the plane of the paper

In the knitting zone of an LL flat-bed knitting machine (Fig. 57) are two flat needle beds mounted horizontally with their channels aligned opposite to each other, facilitating stitch transfer from one double-latch needle to another. Push rods (selectors) are required to move the needles, and these are in turn controlled by cams. Thus, as a consequence of the sliding motion, knitting may either be confined to one of the two needle beds or the growing strand can be transferred from one needle bed to the other. Depending on the direction of passage of a particular needle, a face stitch or a reverse stitch will appear in the knitted material.

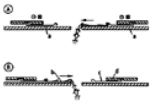

Figure 57. Schematic diagram of the knitting zone of an LL flat-bed knitting machine with needle transfer

A) Stitch formation in the right needle bed; B) Stitch formation in the left needle bed

a) Needle beds; b) Double-latch needle; c) Selector (clavette); d) Clavette cam

With the help of the *Jacquard technique* (individual needle selection), a particular needle can be caused independently to knit (stitch), tuck (loop), or not knit (floating or RL; three-way technique) in each knitting process, using either the RR or the LL bonding technique (Fig. 56 B).

With the RR technique it is possible to transfer any stitch from a needle on one needle bed to one on the other. With this stitch-transfer method (Fig. 58) one is able not only to create fully fashioned parts by adding to or subtracting from the edge of the knit, but also to develop textural designs and patterns similar to those obtained with the LL technique. RR technology is compatible as well with structures of the RL and LL structural groups, either through the stitch-transfer method or by selectively disengaging one needle system.

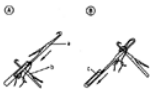

Figure 58. The stitch-transfer method

A) Presentation of a loop by needle (a); B) Capture of the loop by needle (b)

a) Donor latch needle; b) Acceptor latch needle; c) Loop-spreader spring for stitch transfer

Implementation of the Jacquard technique in the LL structural group facilitates needle transfers of every type, and, depending on the direction of transfer, formation of either face or reverse stitches in the knitted material. Patterning based on the “RL – RR – LL” technique, the structural elements “stitch-loop-floating”, a stitch-and-needle transfer technique, or some process involving thread changing, needle-bed shifting, etc., is accomplished with modern, electronically controlled flat-bed knitting machines equipped for integral CAD programming (Fig. 59). Pattern input may involve a scanner, a video camera, a manual drawing, or keyboard commands. Pattern development or the modification and generation of an appropriate control program can be achieved quite quickly with most CAD systems. Prior to knitting, the complete knitting program for a particular product is transferred either online or via a punched tape or cassette to the flat-bed knitting machine's own computer.

Figure 59. Example of a CAD system for a flat knitting machine

The fineness of a knitted fabric depends on both the yarn strength and the machine gauge, which indicates the number of needles per inch in the needle bed (1 inch = 25.4 mm), where the gauge “E 10” means ten needles per inch. “Spacing” refers to the distance between two needles in the same needle bed. Thus, a machine might also be described as having “12-gauge spacing,” meaning twelve needles to the inch (equivalent to “E 12”).

5.5.2. *Circular Knitting Machines*

Circular knitting machines are characterized by a very high fabric output ($\leq 10^6$ stitches/min), and they are commonly used in the production of apparel, domestic fabrics, and industrial products. A basic distinction can be made between small (diameter < 165 mm; Fig. 60 A) and large (diameter > 165 mm; Fig. 60 B) circular knitting machines.

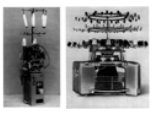

Figure 60. Circular knitting machines

A) Small device; B) Large, electronically controlled machine

RL circular knitting machines (Fig. 61 A) utilize a rotating needle cylinder mounted in the knitting zone, in the channels of which latch needles (cylinder needles) also move vertically. The needle feet slide in the channels of fixed cylinder cams (Fig. 61 B), and can produce several stitches during the course of a single revolution. Each knitting point is known as a system. Up to ca. 150 such systems may be arranged along the periphery of a single circular knitting machine.

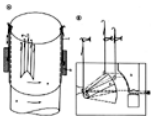

Figure 61. Schematic illustration of the RL circular knitting technique

A) General overview: a) Cylinder; b) Cylinder cam; c) Cylinder needle; d) Knitting point; e) RL knitted fabric

B) Detailed view of a cylinder cam: a) Driving segment; b) Retracting segment

—·—·— Needle path (cam track) for knitting (stitch)

· · · · · · · · Needle path for capturing (loop)

– – – – – Needle path when not knitting (RL, floating)

In addition to a cylinder and the cylinder cams, an RR circular knitting machine also possesses a dial (Fig. 62 A), in the channels of which are radially movable latch needles (dial needles) driven by dial cams. The dial cams are located above the rotating dial, and serve to apply the appropriate knitting motions to the dial needles. If the dial and cylinder needles are staggered, all the needles can be caused to knit simultaneously (the *dial technique*). If they are directly opposite each other in a head-to-head arrangement the needles in a given system must work together in a criss- cross fashion. Thus, only after two cycles will all the needles have knitted, leading to two RR knits that are interlocked (*interlock technique*; Fig. 62 B).

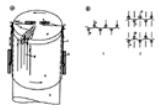

Figure 62. Schematic illustration of the RR circular knitting process

A) General overview: a) Cylinder; b) Cylinder cam; c) Cylinder needle; d) Dial; e) Dial cam; f) Dial needles; g) Knitting point; h) RR knitted fabric

B) Path of the thread, alternative systems: 1) Rib technique; 2) Interlock technique; a) Knitting by a cylinder needle; b) Knitting by a dial needle

LL circular knitting machines (Fig. 63) have two rotating needle cylinders mounted one above the other. The needle channels are so aligned that double-latch needles activated by selectors (clavettes) can not only knit but also transfer thread from cylinder to cylinder. These machines can produce RL, RR, or LL structures depending on the needle arrangement.

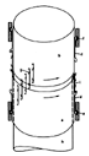

Figure 63. Schematic illustration of the LL circular knitting technique

a) Lower cylinder; b) Upper cylinder; c) Double-latch needle; d) Selector; e) Lower selector cam; f) Upper selector cam; g) Needle transfer; h) LL knitted fabric

With circular knitting, selection of the needles plays an important role in pattern development. Needles with feet that differ (high- or low-butt needles, long- or short-butt needles) make it possible to activate certain needle groups selectively, opening the way to a limited amount of patterning. Only with the Jacquard technique and fully independent needle control, however, is there complete freedom of patterning. Jacquard needle selection in the cylinder can occur mechanically based on such devices as pattern wheels or pattern drums (peg drums), or the process may be driven electronically via electromag-nets. This permits any desired cylinder needle to be designated as “knitting” (stitching), “tucking” (forming a loop), or “nonknitting” (floating).

All the design and process-engineering advantages of the CAD technique can be exploited with electronically controlled circular knitting machines. The gauge designations applied to circular knitting machines correspond to those of flat-bed knitting machines.

5.5.3. *Weft Knitting Machines*

Weft knitting machines are constructed with linear or circular needle arrangements, and they almost always employ the RL technique (single-needle system). Circular RL weft knitting machines are of almost no significance, whereas the flat RL weft knitting machine (Fig. 64), known as the Cotton machine after its inventor, W. COTTON (1864), is of some importance for fully fashioned fitted garments like sweaters and dresses. Such fashion forming is achieved by means of stitch transfer within the needle system (narrowing attachment).

Figure 64. Electronically controlled flat weft knitting machine

The knitting elements (Fig. 65) include bearded needles (a) arranged linearly in a needle bar subject to both horizontal and vertical motion. Loop-forming sinkers (b) are also utilized to form thread delivered by the thread guide (d) into successive loops, requiring an independent drive system based on a slurcock or culier cam (j) and jacks (i). The sinkers are activated simultaneously and intermittently by a sinker catchbar (h) as a way of releasing previously formed loops for subsequent stitch formation. Other knitting elements are the hold-down/knock-over sinkers or web holders (c) that facilitate needle operations during stitch formation (holding down the loop, casting-on, casting-off) as well as narrowing needles (e) arranged above the bearded needles. These narrowing needles transfer the stitches to the side for forming and patterning purposes.

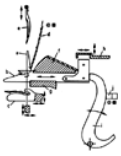

Figure 65. Schematic illustration of the key components in a flat RL weft knitting machine

a) Bearded needle; b) Loop-forming sinker; c) Web holder; d) Thread guide; e) Narrowing needle; f) Sinker head, upper part; g) Sinker head, lower part (press); h) Sinker catchbar; i) Jack; j) Slurcock

Stitch formation (Fig. 66 A – C) begins with laying of the thread. Loop-forming sinkers are pushed one after another into the needle lanes, forming the thread into a row of loops (loop sinking). Subsequently, all the needles simultaneously move downward and in front of the presser bars (pressing), thereby closing them in such a way that the stitch loops are pushed onto the tips of the needles (casting-on) and can be thrown over the hooks (casting-off). The sinkers grasp the stitch loops in their grooves (enclosing) and the needles rise, initiating a new stitch-forming process. For forming and patterning, stitches must be transferred sideways with the help of a narrowing machine. For this purpose, the narrowing needles are lowered onto the bearded needles (Fig. 67) where they capture the stitch loops, move them sideways, and deposit them on the adjacent needles. Stitch transfer can be used both for narrowing purposes and to produce an eyelet pattern in the material (Fig. 68). Color patterning (ringlet, intarsia, and plated patterns) can also be achieved by changing or repositioning the thread guide.

Figure 66. Loop formation with a weft knitting machine

A) Thread laying and loop forming: a) Laying of the thread; b) Sinking the loop; c) RL knitted fabric

B) Pressing and enclosing: a) Thread loop; b) Stitch loop

C) Casting off and clearing: a) Thread loop; b) Stitch loop; c) Stitch

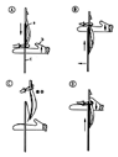

Figure 67. Stitch transfer with the aid of a narrowing needle

A), B) Capture of a stitch loop by the narrowing needle; C) Sideways motion of the narrowing needle; D) Transfer of the stitch loop to the adjacent needle

a) Narrowing needle; b) Sinker; c) Bearded (spring) needle

Figure 68. The use of “transferred stitches” for forming and patterning

a) Fully fashioned forming (narrowing); b) Patterning (eyelet pattern)

The fineness of a Cotton machine is expressed by a *gauge* (gg), which specifies the number of needles in a span of 1.5 inches (e.g., 21 gg = 21 needles/1.5 inch).

5.5.4. *Warp Knitting Machines*

Warp knitting machines use needles that move together and are arranged in a planar array. These machines are equipped with guide bars, each with its own thread guides (guide needles) and thus able to move many warp threads simultaneously in front of and behind the needles (warp-thread take-off; see Fig. 70). Sinkers facilitate the needle operation in these processes as well. Warp

knitting machines offer extremely high output rates of up to ca. 50×10^6 stitches per minute.

A distinction is made between automatic warp knitting machines (Fig. 69 A), Raschel machines (Fig. 69 B), and crochet galloon machines (Fig. 69 C). The latter are used primarily in the manufacture of ribbons. Their mode of operation is very similar to that of the Raschel machines, and for this reason only the automatic warp and Raschel machines will be described in detail.

Figure 69. Examples of warp knitting machines

A) Automatic warp knitting machine; B) Raschel knitting machine; C) Crochet galloon knitting machine

Automatic warp knitting machines are usually equipped with compound needles, whereas Raschel machines have latch needles. The angle between the thread chain and the fabric is ca. 90° in an automatic warp knitting machine and ca. 170° in a Raschel machine. The force applied to the needles is thus considerably higher with automatic warp knitting machines than with Raschel machines. Since stitch size in the knitted material is a function of this force, automatic machines are used for fine articles, and Raschel machines are used for coarser products (apparel, domestic fabrics, industrial products).

The process of stitch formation with an automatic warp knitting machine (Fig. 70 A – D) starts with ascent of the needle (enclosing the loop). The guide bars have already completed their lateral motion (thread laying) on the back side of the needle (underlap). They first swing with their guide needles into the needle lanes, move sideways on the hook side (overlap), and then swing back once again. The needles next drop, and in the subsequent casting-on and casting-off processes create a stitch. The knitted material is guided by sinkers in all phases of stitch formation.

Figure 70. Stitch formation with an automatic warp knitting machine

A) Enclosing the loop: a) Compound needle; b) Enclosing/casting-off sinker; c) Warp thread; d) Guide needle; e) Guide bar G_1; f) Guide bar G_2; g) Warp knitted fabric

B) Overlap: a) Underlap position; b) Overlap position

C) Casting on: a) Stitch loop; b) Thread loop

D) Knocking-over: a) Stitch; b) Stitch loop

The course of the thread ("thread laying") can be seen from a *loop-structure diagram* (Fig. 71). The sinker loops are formed by underlappings, and the needle stitches result from "swinging in," "overlapping," and "swinging out." Depending on the nature of the lappings, various basic structures can be formed, which can be represented by *lapping diagrams* (Fig. 72) showing a top view of the course of the needles in successive stitch-forming processes as seen by the knitter. The illustrated course of the thread shows the resulting structure in a simplified form (pattern notation). In commercial knitted materials, two or more basic structures are combined with each other and with other structural elements as required (Fig. 73).

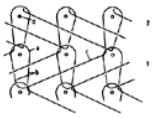

Figure 71. Loop-structure diagram for the product of a warp knitting machine

Process of forming (1) the first course and (2) the second course

a) Stitch head (overlap); b) Stitch shank (swinging in and out); c) Stitch foot (underlap); d) Needle

Note: (a) and (b) needle loop; (c) sinker loop

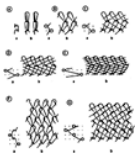

Figure 72. Lapping diagrams (stitch bindings)

A) Open-lap pillar stitch; B) Closed one-and-one lapping movement (tricot lap); C) Closed two-and-one lapping movement; D) Closed three-and-one lapping movement; E) Closed four-and-one lapping movement; F) Two-row open atlas lapping movements; G) Two-row floating closed atlas lapping movements with closed reverse

a) Lapping diagrams; b) Loop structures

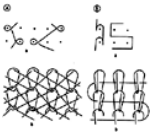

Figure 73. Combined stitch formations

A) Combination of closed two-and-one and closed one-and-one lapping movements; B) Combination of open-lap pillar stitch with inlay

a) Lapping diagrams; b) Loop structures

Lateral movement of the guide bars (i.e., the overlap and the underlap) is accomplished by pattern gears operated either mechanically via chain-link drums or pattern wheels, or else electromechanically.

The stitch-forming process with an RL Raschel machine (Fig. 74 A – D) also begins with enclosing the loop. Web holders (or "hold-down sinkers") (c) support the stitch loops while the latch needles (a) ascend. The underlaps are completed, the guide bars (g, h) swing into the needle lanes, move around the needle hooks (overlap), and then swing back. The needles subsequently descend to form stitches (casting-on and casting-off).

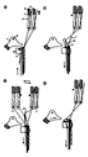

Figure 74. The stitch-formation process with an RL Raschel machine

A) Enclosing the loop: a) Latch needle; b) Knock-over bit; c) Web holder; d) Latch-stop wire; e) Guide needle; f) Warp thread; g) Guide bar G_2; h) Guide bar G_1

B) Overlap: a) Underlap position (e.g., for weft in guide bar G_2); b) Overlap position (e.g., for pillar in guide bar G_1)

C) Casting on: a) Stitch loop (open pillar stitch); b) Thread loop; c) Inlay lap

D) Casting off (knocking over)

One technique typical of Raschel machines, "pillar stitch with inlay," is illustrated in Fig. 73 B. The weft requires only underlaps, and the pillar stitch only overlaps.

Many patterns can be produced by appropriate laying of the thread: through use of various basic structures, structural elements, independent guide-needle displacements within a guide bar on the basis of the Jacquard technique, colored thread draws in the guide bars, weft insertions, etc.

RR Raschel machines (Fig. 75) operate with two vertically mounted needle bars. Stitch-forming processes are carried out in succession at the two needle bars. This mode of operation permits the production of tube-like warp-knitted fabric (packaging materials, clothing) or RR warp-knitted goods, which when cut open result in two plush or pile fabrics.

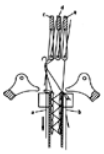

Figure 75. Schematic diagram of a double-needle-bar Raschel machine (RR Raschel machine)

a) Needle bar N_1; b) Needle bar N_2; c) Guide bar for N_1; d) Guide bar for N_1 and N_2; e) Guide bar for N_2

The fineness of automatic warp knitting machines is expressed on the basis of one English inch, that of Raschel machines on two English inches. Thus, with an automatic warp knitting machine a gauge of "E 28" means 28 needles per inch; for a Raschel machine the gauge "ER 36" means 36 needles per two inches.

[Top of Page]

6. Acknowledgement

The entire topic was coordinated by Burkhard Wulfhorst

[Top of Page]

References

General References

1. I. Hormes, M. Weber, B. Wulfhorst: *Faserstoff-Tabelle nach P.-A. Koch; Verarbeitung von Baumwolle*, RWTH Aachen 1991.
2. B. J. Naarding: "Neue Konzepte bei Baumwollanbau-, Ernte- und Entkörnungsverfahren und ihre Bedeutung für die Spinnereien," *Int. Text. Bull. ITS Spinnerei* **2** (1982) 187, 188, 193, 194, 199, 200, 215.
3. H. H. Perkins Jr., J. D. Bargeron III: "Einflußfaktoren auf den Kurzfasergehalt von Baumwolle," *Textilbetrieb (Würzburg)* **100** (1982) no. 9, 21 – 28.
4. W. Klein: *Manual of Textile Technology, Short-staple Spinning Series*, vol. 2: A Practical Guide to Opening and Carding, The Textile Institute, Manchester 1987.
5. K. Gilhaus: "Ein Beitrag zum Mechanismus der Baumwollreinigung," Dissertation, RWTH Aachen 1981.
6. B. Wolf: "Die Karde im Zeichen der modernen Spinnerei," *Int. Text. Bull. ITS Garnherstellung* **34** (1988) no. 3, 37, 40, 41, 44.
7. "Research on Tandem Carding," parts I – IV, *Text. Top.* **17** (1989) no. 7, 1 – 3; no. 8, 1 – 4; no. 9, 1 – 3; no. 10, 1 – 2.
8. W. Klein: *Manual of Textile Technology, Short-staple Spinning Series*, vol. 1: The Technology of Short-staple Spinning, The Textile Institute, Manchester 1987.
9. M. Frey: "Steigerung von Qualität und Produktivität durch den Einsatz einer modernen Baumwollkämmerei," *Melliand Textilber.* **68** (1987) 157 – 163.
10. W. Klein: *Manual of Textile Technology, Short-staple Spinning Series*, vol. 3: A Practical Guide to Combing and Drawing, The Textile Institute, Manchester 1987.
11. W. Klein: *Manual of Textile Technology, Short-staple Spinning Series*, vol. 4: A Practical Guide to Ring Spinning, The Textile Institute, Manchester 1987.
12. E. Dyson: *Rotor Spinning – Technical and Economic Aspects*, The Textile Trade Press, Stockport 1975.
13. R. Nield: *Open-end Spinning*, The Textile Institute, Manchester 1975.
14. D. E. Henshaw: *Self-twist Yarn*, Merrow Publ. Co., Watford 1971.
15. W. Klein: *Manual of Textile Technology, Short-staple Spinning Series*, vol. 5: New Spinning Systems, The Textile Institute, Manchester 1993.
16. L. Hunter: "The Production and Properties of Staple-fiber Yarns Made by Recently Developed Techniques," *Text. Prog.* **10** (1978) nos. 1/2, 1 – 163.
17. H. Deussen: *Rotor Spinning Technology*, Schlafhorst Inc., Charlotte, N.C. 1993.
18. J. Möller: "Transportautomation in der Baumwollkämmerei," *4. Reutlinger Ringspinnkolloquium*, Reutlingen 1989.

General References

19. P. R. Lord, M. H. Mohamed: *Weaving: Conversion of Yarn to Fabric*, Merrow Publishing Co., Watford 1963.
20. J. J. Vincent: *Shuttleless Looms*, The Textile Institute, Manchester 1980.

Specific References

21. R. Cocchi: "La Tessitura a Negli Anni 1990," *Tecnol. Tessile* 1985, no. 4, 24 – 36.
22. W. Loy (ed.): *Taschenbuch für die Textilindustrie 1993*, Fachverlag Schiele & Schön, Berlin 1993.
23. DIN ISO 9947, 11.92: Textilmaschinen und Zubehör: Doppeldraht-Zwirnmaschinen. Begriffe.
24. J. Trauter, W. Wunderlich, H. Böttle: "Untersuchungen zum Kaltschlichten von Spinnfasergarnen," part I: "Grundlagen: Der Einfluß von Parametern auf den Beschlichtungsgrad," *Melliand Textilber.* **73** (1992) 551 – 554, 623 – 626.
25. L. Simon, M. Hübner: *Vorbereitungstechnik für die Weberei, Wirkerei und Strickerei*, Springer-Verlag, Berlin 1983.
26. S. Schlichter: "Der Einfluß der einzelnen Maschinenelemente auf die Bewegungs- und Kraftverläufe in Kette und Schuß an Hochleistungswebmaschinen," Dissertation, RWTH Aachen, Aachen 1987.
27. H. Weinsdörfer: "ITMA 1991: Weberei und Webereivorwerk," *Chemiefasern/Textilind.* **42/94** (1992) 110, 112 – 118.

General References

28. H. Engels: *Handbuch der Schmaltextilien. Teil 1: Maschinen und Verfahren zur Erzeugung konventioneller Geflechte*, Verlag der Fachhochschule Niederrhein, Institut für Textil- und Bekleidungswesen, Mönchengladbach 1994.
29. H. Engels: *Handbuch der Schmaltextilien. Teil 2: Maschinen und Verfahren zur Erzeugung von Flechtprodukten mit speziellen physikalischen und chemischen Anforderungen*, Verlag der Fachhochschule Niederrhein, Institut für Textil- und Bekleidungswesen, Mönchengladbach 1994.
30. F. K. Ko, C. M. Pastore, A. A. Head: *Handbook of Industrial Braiding*, Atkins & Pearce, Covington, Kentucky 1989.

Specific References

31. Arbeitgeberkreis Gesamttextil (eds.): *Ausbildungsmittel Unterrichtshilfen: Textiltechnik Maschinengeflechte*, Frankfurt 1981.
32. F. Goseberg: Meisterkursus Flechterei, Lehrgangsunterlagen der Textilen Abendschule, Wuppertal, Oct. 1, 1967 – Jan. 15, 1970.
33. B. Lepperhoff: *Die Flechterei*, Eugen G. Leuze Verlag, Saalgau/Wttbg. 1953.
34. H.-A. Lange: "Die moderne Flechttechnik und ihre Anwendung in zukunftsorientierten Industriezweigen," *Band. Flechtind.* **24** (1987) 20 – 29.
35. H. Kruse: "Schneller höhere Leistung — 125 Jahre Flechtmaschinenbau — vom Barmer System zur Flechtmaschine HP," *Band Flechtind.* **24** (1987) 72 –76.
36. B. Wulfhorst, A. Büsgen, J. Brandt, H.-F. Siegling: "Anwendung der Flechttechnik für 3-D-verstärkte Verbundwerkstoffe," *TPI Text. Prax. Int.* **48** (1993) 394 – 400.
37. A. Büsgen, B. Wulfhorst, J. Brandt, H.-F. Siegling: "Struktur und Eigenschaften von dreidimensionalen Geflechten zur Herstellung von FVW," *Tech. Text.* **36** (1993) T153 – T154, T156.

General References

38. C. Iyer, B. Mammel, W. Schäch: *Rundstricken*, Meisenbach GmbH, Bamberg 1991.
39. E. Lesykova: *Fachwörterbuch der Maschenwarenproduktion*, Meisenbach GmbH, Bamberg 1991.
40. S. Raz: "Warp Knitting Production," *Melliand Textilber.* **68** (1987) .
41. S. Raz: *Flat Knitting*, Meisenbach GmbH, Bamberg 1991.
42. D. Markert: *Maschen-Abc*, Deutscher Fachverlag, Frankfurt 1990.
43. K. P. Weber: *Die Maschenbindungen der Kettenwirkerei*, Verlag Werkgemeinschaft Karl Mayer, Obertshausen 1966.
44. K. P. Weber: *Wirkerei und Strickerei*, Deutscher Fachverlag, Frankfurt 1987.
45. H. Wignall: *Hosiery Technology*, National Knitted Outerwear Association, New York 1968.
46. A. Reisfeld: *Warp Knit Engineering*, National Knitted Outerwear Association, New York 1966.
47. D. S. Paling: *Warp Knitting Technology*, Columbine Press, Manchester and London 1965.
48. E. Lesykova: *Technical Dictionary for Knitwear and Hosiery Production*, Meisenbach GmbH, Bamberg 1991.
49. K. P. Weber: *An Introduction to the Stitch Formations in Warp Knitting*, Employees Association, Karl Mayer e. V., Obertshausen 1966.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007 John Wiley & Sons, Inc. All Rights Reserved.

Thallium and Thallium Compounds

HEINRICH MICKE, Sachtleben Chemie, Duisburg, Federal Republic of Germany (Chaps. 1 – 5)

HANS UWE WOLF, Universität Ulm, Ulm, Federal Republic of Germany (Chap. 6)

1. Thallium Metal 1
1.1. Historical 1
1.2. Properties 1
1.3. Occurrence 2
1.4. Extraction 3
1.5. Uses 5
2. Thallium Alloys 6
3. Thallium Compounds 6
3.1. Individual Compounds 7
3.2. Glasses and Single Crystals 10
3.3. Uses 10
3.4. Environmental Aspects 10
4. Analysis 10
5. Economic Aspects 11
6. Toxicology and Occupational Health . 11
7. References 13

1. Thallium Metal [1–18]

Thallium [*7440-28-0*], Tl, A_r204.39, atomic number 81, is in group 13 of the Periodic Table, and has electronic structure $1s^2$, $2s^2p^6$, $3s^2p^6d^{10}$, $4s^2p^6d^{10}f^{14}$, $5s^2p^6d^{10}$, $6s^2p$ [19]. It is a dense, fairly reactive metal, the fresh surface having a bluish-white luster. Thallium is softer than lead, but harder than indium, with a low *mp* 303 °C.

There are 20 known isotopes of thallium, of which two, atomic numbers 203 and 205, are relatively stable. The isotope ^{204}Tl is used for materials testing (Section 1.5).

The only thallium-containing substances of commercial importance are the metal itself and thallium sulfate. As thallium compounds are prohibited for insect pest control in the United States and parts of Europe, methods of extraction have not been improved during the last 10 years. Several European producers have stopped production of thallium and its compounds.

1.1. Historical

Thallium was discovered in 1861 by the Englishman CROOKES and in 1862 by the Frenchman LAMY, probably independently. During spectrographic examination of a lead chamber slime, CROOKES found a green line that could not be ascribed to any known element. He was convinced that he had discovered a new element, and gave it the name thallium after the Greek word $\theta\alpha\lambda\lambda o\varsigma$, a young green branch. LAMY must have been the first person to isolate the element and establish its metallic character. A heated dispute over priority developed between the two investigators [20].

1.2. Properties

In many of its physical properties, thallium resembles its neighbor, lead. Its chemical behavior combines the properties of the alkali metals with those of silver, mercury, lead, and aluminum. Thallium is highly toxic (see Chap. 6).

Chemical Properties. Thallium is a very reactive metal. It oxidizes slowly in air, even at ca. 20 °C (more rapidly on heating), to thallium(I) oxide and thallium(III) oxide. Hydrogen peroxide and ozone oxidize the metal to thallium(III) oxide. In the presence of water, a hydroxide is formed. Thallium is therefore stored under petroleum spirit, glycerine, or deaerated water. On prolonged storage under water with free access of air, metallic thallium becomes covered with large crystals of thallium(I) carbonate, and the water becomes saturated with this salt. To protect against oxidation, rods of thallium are coated with paraffin.

Thallium combines with fluorine, chlorine, and bromine at room temperature, and with iodine, sulfur, phosphorus, selenium, and tellurium on heating. Nitrogen, carbon, and molecular hydrogen do not react with the metal, which does not dissolve in liquid ammonia. The metal

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_607

is dissolved slowly by hydrochloric and dilute sulfuric acids, but more rapidly by nitric and concentrated sulfuric acids. Its slow reaction with hydrofluoric acid is noteworthy.

In the series of electrode potentials of cations, thallium has a similar normal potential to that of indium (Tl/Tl^{+} -0.335 V, In/In^{3+} -0.34 V), lying between cadmium (Cd/Cd^{2+} -0.402 V) and tin (Sn/Sn^{2+} -0.140 V). Solutions of thallium(III) salts are not stable toward metallic thallium, reacting almost quantitatively:

$$Tl^{3+} + 2\,Tl \longrightarrow 3\,Tl^{+}$$

The electrode potential of Tl^{+}/Tl^{3+} is +1.21 V.

Physical Properties. Thallium is a soft metal (Brinell hardness 29.4 N/mm^2, the value for lead being 39.2 N/mm^2). Thallium exists in two allotropic crystal forms: α-thallium, hexagonal close-packed, stable at room temperature; and β-thallium, body-centered cubic, stable above 226 °C. A volume increase of 3.23 % takes place on solidification. On bending, the metal emits a creaking sound similar to the "cry" of tin. It has good cold forming properties and low strength. It can be melted and worked like lead. Physical properties are as follows [21, 22]:

Atomic number	81
Isotope abundance	
^{203}Tl	29.52 %
^{205}Tl	70.48 %
Density at 20 °C	11.85 g/cm^3
Atomic volume	17.26 cm^3/mol
Melting point	303.6 °C
Boiling point	1457 °C
Latent heat of fusion	30 J/g
Latent heat of vaporization at normal boiling point	800 J/g
Vapor pressure of melt, Torr	$\log p = \frac{-9300}{T} - 0.892 log$
Thermal expansion of melt	3.1 – 4.3 %/°C
Electrical resistivity	
2.38 K	superconducting
0 °C	$18 \times 10^{-6}\ \Omega \cdot$ cm
Specific magnetic susceptibility	
Solid thallium	-0.165×10^{-6} cm^3/g
β-Thallium	-0.158×10^{-6} cm^3/g
Molten thallium	-0.22×10^{-6} cm^3/g
Surface tension	4.01 mN/cm
Tensile strength	9.8 kg/mm^2
Relative elongation	135 %
Brinell hardness	3 kg/mm^2
Mohs hardness	1.2

Like mercury, lead, indium, etc., thallium can become superconducting, the average upper temperature limit being ca. 2.38 K [23]. Some compounds, e.g., Tl_3Bi_5, Tl_2Hg_5, and Tl_7Sb can also become superconducting [24]. It is especially valuable as a superconductor in ceramic materials at higher temperatures than are possible with other metals [25].

Vessels for containing liquid thallium can be made of Fe, W, Ta, Mo, Nb, and Co, in decreasing order of suitability.

1.3. Occurrence

Very variable figures are quoted for the concentration of thallium in the Earth's crust (1 – 3 ppm). It is said to be the 58th most abundant element [26]. In its occurrence, thallium shows a dual character. As a chalcophilic element, it occurs typically, though in low concentration, with the heavy metals that occur in sulfidic ores, usually associated with cadmium, mercury, indium, and germanium. Much more frequently, thallium accompanies the alkali metals potassium, cesium, and rubidium, where its concentration is also low, e.g., in carnallite, sylvine, leucite, lepidolite, feldspar, mica, etc. It is found in seawater, some mineral waters, and in native sulfur [1]. It also occurs in true thallium minerals, of which the following are the most common:

Lorandite	$TlAsS_2$
Pirotpaulite	$TlFe_2S_3$
Chalcothallite	Cu_3TlS_2
Vrbaite	$Hg_3Tl_4As_8Sb_2S_{20}$
Hutchinsonite	$(Pb,Tl)_2(Cu,Ag)As_5S_{10}$
Bukovite	$Cu_{3+x}Tl_2FeS_{4-x}$
Wallisite	$PbTlCuAs_2S_5$
Hatchite	$PbTlAgAs_2S_5$
Crookesite	$(Cu,Tl,Ag)_2Se$
Avicennite	$7\,Tl_2O_3 \cdot Fe_2O_3$

The thallium content of these minerals is 16 – 85 %. For many years, only five true thallium minerals were known. Because of their extreme rarity, these are of no importance for the extraction of the metal [27, 28].

1.4. Extraction

Although the amounts of thallium in ores, including those of lead and zinc, are small compared with the amounts in salts, rock, etc., only the former play a significant role in the winning of the metal. They can be economically worked

only if the thallium becomes concentrated in the side products in the course of processing the ore to obtain the main metal, and if other metals are present in this side product to an extent that justifies their extraction. Only a small part of the thallium that occurs naturally is therefore available for extraction.

Thallium is present only as a trace element in nonferrous metal concentrates. Only in a few smelting works does the thallium become sufficiently concentrated in particular process steps for extraction to be practical.

The following companies are thallium producers: Vieille Montagne, Belgium; Cimkent, Russia; Toho Zinc, Japan; Cerro de Pasco, Peru; Preussag, Germany.

Thallium produced as a byproduct of the extraction of nonferrous metals is of only minor importance, both quantitatively and economically.

The following concentration processes can be used for the treatment of solid raw materials:

leaching and precipitation
volatilization
special processes

If the raw materials are in solution, the thallium is precipitated by cementation, and is then processed like a solid raw material.

Leaching and Precipitation. The simplest process is leaching with water without pretreatment. With airborne dusts from lead or zinc smelting works, or cementation slimes, a yield of 85 – 90 % of the thallium can be obtained. The thallium-containing liquor, which usually also contains cadmium and zinc, is filtered, precipitated with zinc amalgam, and the relatively pure metal is obtained by an electrolytic process (Fig. 1) [31].

There are many methods of pretreating the starting material before leaching. These include:

oxidative roasting at 500 °C
sulfatization at 175 – 250 °C
alkaline pretreatment

Leaching of the raw material after pretreatment can extract 50 – 60 % of the thallium. Flue dusts that contain bismuth, copper, zinc, iron, arsenic, cadmium, thallium, selenium, etc., are given an alkaline pretreatment (fusion with caustic soda and lead sulfide). The alkaline slag

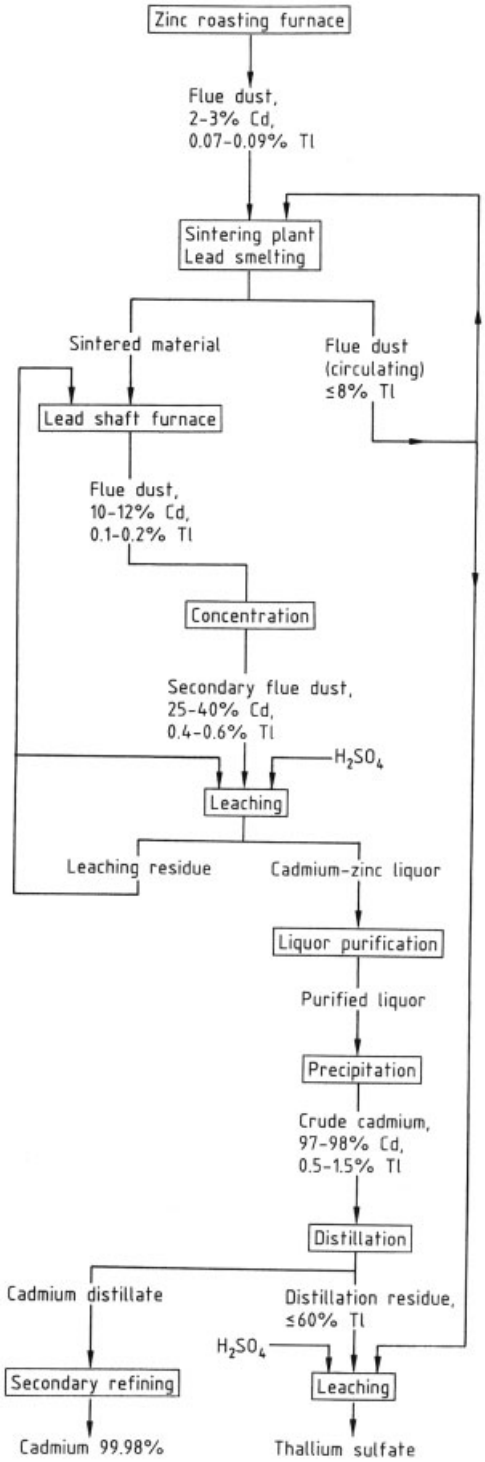

Figure 1. Simplified flow diagram of a plant for the combined recovery of cadmium and thallium sulfate from lead and zinc raw materials [29, 30]

so formed is granulated with hot water. On cooling, sodium and thallium arsenates are precipitated, and processed to obtain the thallium [32].

The thallium is usually precipitated and recovered as a sparingly soluble salt (cementation thallium and thallium amalgam) from liquors produced in a variety of concentration methods, with and without pretreatment, and also from liquors that are themselves the raw materials.

Processing and refining of the precipitates obtained by cementation at the Duisburg copper refinery [33,34] was discontinued in 1983.

Volatilization. In this process, crude arsenic containing 96 % As_2O_3 and 0.2 % Tl is treated by adding 5 % H_2SO_4 and 5 % lime and heating to 430 °C, causing 98 % of the arsenic trioxide to be volatilized while 99 % of the thallium remains in the residue. Sodium chloride is added, and the thallium is volatilized at 800 °C [33]. The volatilization process also gives good results with materials of complex composition from lead smelting. The secondary flue dust can be reprocessed. The most well-known process was that of the former company Zinkhütte Magdeburg [34].

Electrorefining. Electrorefining is operated in one or several stages, or as amalgam electrolysis. Depending on the purity of the starting metal and the operating conditions, thallium cathodes with variable thallium content (ca. 99.9 – 99.99 % and above) are produced [35–37]. Metal of the highest purity can be obtained by multistage electrorefining, in which a perchlorate electrolyte containing 40 – 70 g/L $TlClO_4$ and 60 – 120 g/L $NaClO_4$ with small amounts of organic additives is used in the final stage. The current density is 0.3 – 0.6 A/dm^2. The sum of all the impurities in the metal is < 1 g/t, i.e., the thallium produced has a purity of 99.9999 % [38].

Wet metallurgical refining is normally combined with the concentration process. The high solubility of thallium sulfate in water is used when lead and other metals that form insoluble sulfates are precipitated. The insolubility of thallium sulfide in alkaline solvents is used to separate it from the group 1 metals, and its solubility in dilute acids is used to separate it from silver, mercury, bismuth, copper, arsenic, antimony, cadmium, etc. The solution is treated with hydrogen sulfide, or preferably with freshly precipitated thallium sulfide, so that an easily handled thallium sulfate precipitate is produced at the end of all operations [39]. Iron and arsenic can be removed after oxidation with air followed by neutralization, e.g., by adding zinc oxide. The high solubility of thallium carbonate in water is used to separate it from zinc, iron, cadmium, and other impurities, usually in combination with a concentration process. The low solubility of thallium chromate is then used for both concentration and refining.

Very pure metal can also be obtained by fractional dissolution of the impure metal or by fractional precipitation, mainly by fractional cementation of the impurities and the thallium from impure solutions. Impure thallium metal or thallium amalgam can be used as the starting material. When thallium amalgam is heated with 70 – 100 % sulfuric acid to boiling, the thallium dissolves first. If dissolution is interrupted after almost all the thallium, but none of the impurities, has gone into solution, the solution contains pure thallium sulfate which can be the raw material for metal winning [40].

To obtain compounds of the highest purity for use as intermediates for producing very high-purity metal, several unusual reactions have been proposed. These include precipitation of the thallium as cyclopentadienyl thallium [41,42], or extraction of thallium(III) chloride, $TlCl_3$, from solution in hydrochloric acid (6 – 7 mol/L) with ether. First, however, the iron must be converted to the divalent form, and the heavy metals removed. Almost complete removal of all impurities takes place by taking the impure starting substance into solution in hydrochloric, hydrobromic, or hydroiodic acid, and extracting the thallium complex with isopropyl ether at acid concentration 0.1 – 4 mol/L. The metal obtained by electrolysis of the purified compound has purity 99.9999 – 99.99999 % [43].

During production of an ultrapure product from the already relatively pure metal by crystallization (especially zone melting and crystal pulling from the melt), the various impurities behave very differently. As well as the distribution coefficients, the operating conditions (mainly the migration rate of the zones and the crystallization rate) have a great influence on the refining. However, many other factors have an influence, e.g., the rotation rate during zone melting,

and the rotation rate of the seed crystal during crystal pulling from the melt. Secondary processes are also important, e.g., volatilization and oxidation of certain impurities. The behavior of some of the more common potential impurities for a given combination of operating conditions can make it difficult to obtain very pure end products.

Zone melting is generally carried out in graphite boats. The ingots used are 150 – 180 mm long, with a ca. 15 mm molten zone, the process being carried out in an atmosphere of purified nitrogen. With a zone migration rate of 0.5 – 1.0 mm/min and five passes, the copper content in most of the ingot can be reduced from 50 g/t in the starting metal to 2 – 7 g/t, and, under the same conditions, the silver content in approximately half the length of the ingot is reduced from 3 to < 0.3 g/t. By pulling crystals from the melt in vacuo (ca. 10^{-2} Pa, 10^{-4} mbar), crystals 100 – 220 mm long and 8 – 10 mm diameter can be obtained. With a crystallization rate of 0.4 – 0.5 mm/min, the manganese content is 0.2 g/t in the lower part and 0.1 g/t in the middle part of the crystal. After a second recrystallization, 0.05 g/t manganese is found in the middle and upper part.

Not all impurities can be removed by crystallization, e.g., lead and iron, and there is a group of impurities that can only be removed by secondary processes. These include mercury, tin, and sulfur [44].

In Germany, thallium is produced solely by Preussag; annual production capacity is 4 – 5 t of 99.4 – 99.64 % thallium [45]. Thallium is supplied as rods and granules in "4 nines" to "5 nines" quality (i.e., 99.99 – 99.999 %). The catalogue also includes "pure" monovalent thallium compounds, including sulfate, bromide, chloride, and iodide. Other compounds can be produced on request.

1.5. Uses

No large application potential for thallium has so far been found. The unalloyed metal is unsuitable for direct use, owing to its unfavorable mechanical properties and marked tendency to oxidize. The radioactive isotope ^{204}Tl is a β-emitter with half-life ca. 3.5 years, and is used as a radiation source in materials testing, principally for thickness measurements on metals and nonmetals. It has been discovered that ^{205}Tl has the most constant atomic vibrations so far observed. This property has led to its use in atomic clocks.

These special applications require extremely pure grades. The very pure metal supplied by Unterharzer Berg- und Hüttenwerke contains > 99.999 % Tl, i.e., the sum of all the impurities does not exceed 10 g/t [46]. American Smelting and Refining supply "high-purity thallium", with the same purity. The following figures are quoted for the impurities [47]: Cu < 1 g/t; Pb 1 g/t; Fe < 1 g/t; Ag, Sn, Mg, and Si spectrographically undetectable, but < 1 g/t; other elements spectrographically undetectable.

2. Thallium Alloys

Thallium readily forms alloys with most other metals. There is complete mutual insolubility with iron, and limited solubility in the liquid state with copper, aluminum, zinc, arsenic, manganese, and nickel. The melting points of thallium – lead alloys are higher than those of the pure metals, the maximum, 380 °C, being reached at 37.8 % Pb. There is a resemblance to the two-component systems of lead, especially in alloys of thallium with copper, aluminum, zinc, and nickel. Gold, silver, cadmium, and tin form simple eutectics with thallium. Thallium also forms binary alloys with antimony, barium, calcium, cerium, cobalt, germanium, indium, lanthanum, lithium, magnesium, strontium, tellurium, and bismuth.

Alkali metals, alkaline earth metals, some metals of the rare earth group (e.g., lanthanum, cerium, praseodymium), and mercury form intermetallic compounds with thallium, some of which have unusually high melting points, e.g., CeTl (1240 °C) and PrTl (ca. 1150 °C). The compound with mercury, Hg_5Tl_2, has *mp* 14.5 °C and forms eutectics with thallium and mercury: Tl – Hg_5Tl_2 with 40.45 % Tl melts at 0.6 °C and Hg – Hg_5Tl_2 with 8.5 % Tl at – 58.4 °C.

Of the systems with three or more components, the combinations with bismuth, lead, and cadmium have been much investigated. These have very low melting points over wide ranges, and in most cases form ternary or quaternary eutectics. For example, the melting point of the

Bi – Pb – Tl eutectic with 11.5 % Tl, 55.2 % Bi, and 33.3 % Pb is 90.8 °C, and the melting point of the quaternary Bi – Cd – Pb – Tl eutectic with 8.9 % Tl, 44.3 % Bi, 35.8 % Pb, and 11.0 % Cd is 81 °C [9].

The ternary alloys Tl – Pb – Bi, Tl – Al – Ag, In – Hg – Tl, Sn – Cd – Tl, Bi – Sn – Tl, and Bi – Cd – Tl are used as semiconductors or in ceramic compounds [48, 17].

The quaternary alloys Tl – Sn – Cd – Bi, Pb – Sn – Bi – Tl, and Pb – Cd – Bi – Tl are known [17].

Thallium, like tin, cadmium, lead, etc., has good wear resistance when used in bearings for steel shafts. As additions of thallium also increase the deformation resistance, breaking strength, yield, strength, and hardness of lead and its alloys, thallium-containing alloys are frequently recommended for bearings [49]. The alloy 72 % Pb, 15 % Sb, 5 % Sn, and 8 % Tl is superior to the best tin-based bearing metals with respect to yield strength and breaking strength. Bearing metals based on Pb – Tl – Cd with high cadmium content can have compositions in the range 0.01 – 5 % Tl, 5 – 40 % Cd, the rest being Pb [47]. However, thallium has since been replaced by indium as an alloy constituent [45].

Addition of thallium to gold, silver, and copper contacts in the electronics industry (e.g., 0.05 – 20 % Tl added to copper alloys) reduces the tendency to sticking [45].

In solid-state rectifiers, especially selenium rectifiers, addition of thallium increases the blocking resistance [50].

Addition of thallium can reduce the melting point of mercury from − 39 °C to ca. − 59 °C. One use for this low-melting Hg – Tl alloy is for low-temperature thermometers [51].

The increase in the melting point of lead by addition of thallium is used in special types of electrical fuses and in solder (thallium content ≤ 20 %) where a higher melting point is required [5, 6].

3. Thallium Compounds

Thallium can be monovalent or trivalent, the stability of the 6*s* electron pair in the electronic structure making thallium preferentially monovalent in inorganic compounds. Thallium(I) compounds have a certain similarity to alkali metal compounds, e.g., high solubility of the sulfate, carbonate, and hydroxide in water. However, they also show similarities to compounds of silver, mercury, and lead, e.g., low aqueous solubility of the chloride, bromide, iodide, and sulfide.

3.1. Individual Compounds

Thallium(I) oxide [*1314-12-1*], Tl_2O, is formed by heating thallium in air at < 100 °C, or by dehydration of TlOH. It is reddish-black to black, volatile in air at higher temperatures, being converted to Tl_2O_3, and is very hygroscopic, forming a colorless, alkaline solution of TlOH with water. The enthalpy of formation is given by:

$$2\,Tl + 1/2\,O_2 \longrightarrow Tl_2O$$

$$\Delta H_f = -\,180.8 \pm 0.8\,\text{kJ/mol}\ (-425.5 \pm 2.1\,\text{kJ/kg})$$

With B_2O_3, borates, or mixtures of borates and fluorides of alkali and alkaline earth metals, Tl_2O forms stable, clear, homogeneous melts which can be used for the electrolytic production of metallic thallium.

Thallium(III) oxide [*1314-32-5*], Tl_2O_3, is more stable, and can be formed by heating Tl_2O to > 100 °C, treating thallium(III) salts with KOH or NH_3, or treating thallium(I) salts with oxidizing agents, e.g., TlCl with NaOCl. It is brown to black, forms cubic crystals with density (4 – 21 °C) 9.65 g/cm^3, *mp* 717 ± 5 °C, and begins to sublime above 500 °C with partial decomposition:

$$Tl_2O_3 \rightleftharpoons Tl_2O + O_2$$

It can be reduced to Tl_2O, and even to the metal, by strongly heating with hydrogen, carbon, or carbon dioxide. It reacts very vigorously with sulfur and hydrogen sulfide, even at room temperature, doubtless in accordance with the equations:

$$2\,Tl_2O_3 + 5\,S \longrightarrow 2\,Tl_2S + 3\,SO_2$$

$$3\,Tl_2O_3 + 5\,H_2S \longrightarrow 3\,Tl_2S + 5\,H_2O + 2\,SO_2$$

Treatment of an aqueous suspension with sulfur dioxide leads to Tl_2SO_4:

$$Tl_2O_3 + 2\,SO_2 \longrightarrow Tl_2SO_4 + SO_3$$

It is sparingly soluble in water, but does not form a hydroxide. It dissolves in hydrochloric, sulfuric, and nitric acids with formation of hygroscopic, unstable thallium(III) salts, from which it is precipitated unchanged by alkali. It is also dissolved by acetic and tartaric acids. With oxalic acid, the sparingly soluble oxalate is formed, which decomposes on heating:

$$Tl_2C_2O_4 \longrightarrow 2\,Tl + 2\,CO_2$$

Thallium(I) hydroxide [*12026-06-1*], TlOH (92.32 % Tl), is formed on heating metallic thallium with water in the presence of air, or dissolving Tl_2O in water. It is white, becoming dark gray on exposure to light. Its enthalpy of formation is given by:

$$Tl + 1/2\,O_2 + 1/2\,H_2 \longrightarrow TlOH$$
$$\Delta H_f = -240.4\,\text{kJ/mol}\ (-1085.8\,\text{kJ/kg})$$

It dissociates to Tl_2O and H_2O on heating above 139 °C.

It is soluble in water, forming a colorless markedly alkaline solution which rapidly absorbs O_2 from the air, with formation of Tl_2CO_3, and rapidly oxidizes, forming Tl_2O_3. The aqueous solution is a stronger base than NaOH, attacking glass very strongly on heating. Sodium sulfide precipitates Tl_2S:

$$2\,TlOH + Na_2S \longrightarrow Tl_2S + 2\,NaOH$$

Thallium(I) chloride [*7791-12-0*], TlCl, is formed when HCl or alkali metal chlorides are added to neutral thallium(I) salt solutions. The precipitated product is sparingly soluble in water, and is of some metallurgical importance. It is white, becoming gray-brown to blackish-brown on exposure to light. It forms cubic crystals, density 7.00 – 7.05 g/cm^3, *mp* 427 °C, *bp* 806 °C, initial sublimation temperature ca. 360 °C, specific heat capacity c_p (2.8 – 45.3 °C) = 0.221 J g^{-1} K^{-1}. The enthalpy of formation is given by:

$$Tl + 1/2\,Cl_2 \longrightarrow TlCl$$
$$\Delta H_f = -203.7\,\text{kJ/mol}\ (-847.3\,\text{kJ/kg})$$

No thermal decomposition takes place at the melting point. On heating in a stream of chlorine, it is first converted to $3\,TlCl \cdot TlCl_3$, and then to $TlCl_3$. On melting with $NaHSO_4$, it is converted to Tl_2SO_4:

$$2\,TlCl + NaHSO_4 \longrightarrow Tl_2SO_4 + NaCl + HCl$$

Metallic thallium is liberated on melting with alkali metal cyanides, or zinc:

$$2\,TlCl + Zn \rightleftharpoons ZnCl_2 + 2\,Tl$$

It is very sparingly soluble in water, 0.29 g dissolving in 100 g H_2O at 15.6 °C, and 2.41 g at 99.35 °C. It reacts with a dilute solution of Na_2SO_4 to form Tl_2SO_4:

$$2\,TlCl + Na_2SO_4 \longrightarrow Tl_2SO_4 + 2\,NaCl$$

On heating with H_2SO_4, Tl_2SO_4 or $TlHSO_4$ are formed (see above). HNO_3 reacts in the cold to form $2\,TlCl \cdot TlCl_3$, and, on heating, the very soluble $TlCl_3$ (with some $TlNO_3$), in accordance with the approximate equation:

$$9\,TlCl + 8\,HNO_3 \longrightarrow 3\,TlCl_3 + 6\,TlNO_3 + 2\,NO + 4\,H_2O$$

On treatment with a 20 % solution of Na_2CO_3, Tl_2CO_3 is formed, which is converted to TlCl on addition of NaCl:

$$2\,TlCl + Na_2CO_3 \rightleftharpoons Tl_2CO_3 + 2\,NaCl$$

It reacts with ammonium sulfide to form Tl_2S:

$$2\,TlCl + (NH_4)_2S \longrightarrow Tl_2S + 2\,NH_4Cl$$

and with sodium hypochlorite to form Tl_2O_3:

$$2\,TlCl + 3\,NaOCl \longrightarrow Tl_2O_3 + 3\,NaCl + Cl_2$$

With $CdCl_2$, TlCl forms a sparingly soluble double salt, $TlCl \cdot CdCl_2$, which decomposes again on boiling with water.

Thallium(I) iodide [*7790-30-9*], TlI, is formed by direct combination of the elements or by precipitation from thallium(I) salt solutions by addition of alkali metal iodide. There are two modifications: a yellow rhombic form stable at room temperature, and a red cubic form stable above ca. 165 °C. The yellow form has density 7.1 g/cm^3 and the red form 7.45 g/cm^3, *mp* 440 °C, *bp* 824 °C. Sublimation starts at 355 °C. The specific heat capacity in J g^{-1} K^{-1} is given by $c_p = 0.152\,T + 3.43$. The enthalpy of formation is given by

$$Tl + I \longrightarrow TlI$$
$$\Delta H_f = -126.28\ \text{J/mol}\ (-381.2\ \text{J/kg})$$

Metallic thallium is liberated on melting with alkali metal cyanides. The following reaction takes place with sodium sulfide:

$$2\,TlI + Na_2S \longrightarrow TlS + 2\,NaI$$

Its solubility in water is considerably less than that of TlCl, but is increased by addition of KOH. It is not attacked by sulfuric and hydrochloric acids, but nitric acid (dilute or concentrated) reacts with liberation of iodine.

Thallium(I) fluoride [*7789-27-7*], TlF, is very soluble in water; unlike the other thallium(I) halides, the solution is alkaline.

Thallium(I) sulfide [*1314-97-2*], Tl_2S, is formed on melting together thallium and sulfur (along with Tl_2S_5, depending on the mixing ratio). It is also precipitated from thallium(I) salt solutions on adding $(NH_4)_2S$, or an amorphous precipitate is formed when H_2S is added to very weakly acid solutions. This becomes crystalline on heating to 150 – 200 °C with a solution of $(NH_4)_2S$. From thallium(III) salt solutions, a mixture of Tl_2S and sulfur is precipitated. Tl_2S is also formed when a solution of Tl_2SO_4 is heated with $Na_2S_2O_3$. It is black, and the crystals have a blue metallic luster. Its density is 8.4 g/cm^3, and its melting point 448 °C. It vaporizes on heating above 300 °C, at first unchanged, but decomposing at higher temperatures (e.g., 900 °C). The enthalpy of formation is given by:

$$2\,Tl + S \longrightarrow Tl_2S$$
$$\Delta H_f = -90.6\ \text{kJ/mol}\ (-205.4\ \text{kJ/kg})$$

Precipitated Tl_2S is readily oxidized in air to Tl_2SO_4. It is sparingly soluble in water and almost insoluble in alkaline solvents, but dissolves in boiling mineral acids (dilute H_2SO_4 and HNO_3). Reaction with fuming nitric acid leads to the sulfate:

$$3\,Tl_2S + 8\,HNO_3 \longrightarrow 3\,Tl_2SO_4 + 8\,NO + 4\,H_2O$$

The other sulfides (Tl_2S_3, Tl_2S_5) become increasingly insoluble as the sulfur content increases, and are converted to Tl_2S with liberation of sulfur on heating.

Thallium(I) sulfate [*7446-18-6*], Tl_2SO_4 (80.97 % Tl), is the most important starting material for thallium metal winning. Depending on its method of manufacture, it forms colorless, very lustrous rhombic crystals or a snow-white crystalline powder, density 6.765 g/cm^3, *mp* 632 °C. It vaporizes unchanged at white heat. Above 900 °C, its vapor pressure increases rapidly. During the roasting of zinc concentrates (900 °C) and the sintering of lead concentrates (1000 °C), a considerable amount of thallium can volatilize as its sulfate. The effect of temperature on the vapor pressure can be approximately represented by:

$$\log p = -\frac{9300}{T} - 0.892 \log T + 11.1$$

where p is in Torr (1 Torr = 133 Pa); T is in K.

The enthalpy of formation is given by:

$$2\,Tl + 2\,O_2 + S\ (\text{rhombic}) \longrightarrow Tl_2SO_4$$
$$\Delta H_f = -933.9\ \text{kJ/mol}\ (-1849\ \text{kJ/kg})$$

It is very soluble in water, e.g., 100 g boiling water dissolves 18.45 g Tl_2SO_4. The solution is hardly hydrolyzed, and is therefore neutral. Addition of hydrochloric acid or alkali metal chlorides to the solution precipitates sparingly soluble TlCl, though never completely:

$$Tl_2SO_4 + 2\,HCl \rightleftharpoons 2\,TlCl + H_2SO_4$$

Potassium iodide addition can give complete precipitation:

$$Tl_2SO_4 + 2\,KI \longrightarrow 2\,TlI + K_2SO_4$$

Addition of sodium thiosulfate at the boiling point produces Tl_2S:

$$3\,Tl_2SO_4 + 4\,Na_2S_2O_3 \longrightarrow 3\,Tl_2S + 4\,Na_2SO_4 + 4\,SO_2$$

With excess sodium carbonate, partial hydrolysis of the Tl_2CO_3 formed, and oxidation of the TlOH, produces a black precipitate. Addition of $Na_2Cr_2O_7$ or Na_2CrO_4 produces sparingly soluble Tl_2CrO_4, which can be redissolved to Tl_2SO_4 and $CrSO_4$ by treatment with dilute H_2SO_4 and $NaHSO_4$. Treatment of Tl_2SO_4 with H_2SO_4 or treatment of TlCl with excess H_2SO_4 gives thallium(I) hydrogen sulfate, $TlHSO_4$, or $Tl_2SO_4 \cdot H_2SO_4$, which melts at 115 – 120 °C and is converted to neutral Tl_2SO_4 with liberation of SO_3 by strong heating.

Thallium(I) nitrate [*10102-45-1*], $TlNO_3$, is formed when metallic thallium or Tl_2CO_3 is dissolved in HNO_3. It is white, and forms rhombic crystals (a cubic modification is also known), density 5.55 g/cm^3, *mp* 206 °C. It melts without decomposition, but decomposition begins at ca. 300 °C, and is complete at 450 °C:

$$2\,TlNO_3 \rightleftharpoons Tl_2O_3 + N_2O_3$$

The enthalpy of formation is given by:

$$Tl + 1/2\,N_2 + 11/2\,O_2 \longrightarrow TlNO_3$$
$$\Delta H_f = -243.3\ \text{kJ/mol}\ (-913.4\ \text{kJ/kg})$$

It is very soluble in water, the solubility increasing very rapidly with temperature, 100 g H_2O dissolving 7.93 g $TlNO_3$ at 15.4 °C and 593.93 g at the boiling point. Its solutions are neutral. H_2SO_4 reacts with $TlNO_3$ to form Tl_2SO_4.

Thallium(I) carbonate [*6533-73-9*], is formed by the action of CO_2 on TlOH, or by reacting thallium(II) salt solutions with an alkali or alkaline earth hydroxide, e.g., $Ba(OH)_2$, while passing CO_2 into the solution.

It is white, and forms monoclinic crystals, density 7.164 g/cm^3, *mp* 272 °C, decomposing at elevated temperatures. The Tl_2O thus formed is soluble in molten Tl_2CO_3, and is converted to Tl_2O_3 on prolonged heating.

The enthalpy of formation is given by:

$$2\,Tl + C + 11/2\,O_2 \longrightarrow Tl_2CO_3$$
$$\Delta H_f = -699.45\ \text{kJ/mol}\ (-1492.1\ \text{kJ/kg})$$

It is fairly soluble in hot water (27.78 g Tl_2CO_3 dissolves in 100 g H_2O at 100 °C). On dilution, rapid hydrolysis takes place with formation of TlOH, so that the solution is alkaline. It is rapidly decomposed by acids.

Thallium metasilicate [*29209-99-2*], Tl_2SiO_3, is obtained by treatment of thallium(I) hydroxide with sodium metasilicate, Na_2SiO_3, dissolved in a solution of $NaNO_3$. The orthosilicate, Tl_4SiO_4, is prepared by reacting thallium(I) hydroxide with sodium orthosilicate, Na_4SiO_4. Both silicates are water soluble. Solid vitreous thallium silicate has a high refractive index.

The total quantity of thallium compounds in Germany does not exceed 500 kg (based on thallium metal), and is distributed in laboratories, universities, and research institutes.

3.2. Glasses and Single Crystals

Glasses composed of the three-component systems As – Tl – S and As – Tl – Se may be prepared by melting the elements together. They are very mobile liquids, even below 400 °C. Viscosity as low as 3 Pa · s has been observed at 250 °C for a composition of, e.g., 25 % As, 45 % Tl, and 30 % S. The glasses are chemically very stable, and insoluble in dilute acids. They are very slightly attacked by alkalis. The electrical resistivity is ca. $10^6 - 10^{18}\ \Omega \cdot \text{cm}$. The glasses have good wetting properties toward many metals, and are extremely impervious. They are therefore suitable for hermetically sealing sensitive electrical equipment (capacitors, resistors, semiconductors, etc.) to protect them from moisture and atmospheric effects. The glasses can be vaporized in vacuo and directly condensed onto the article to be coated [52].

The properties of photochromic glass can be improved by the addition of 0.2 – 4 % thallium (as TlCl added to AgCl).

Single crystals of three-component systems, e.g., Tl_3VS_4, Tl_3NbS_4, and Tl_3PSe_4, are used in acoustical – optical measuring equipment, e.g., laser modulators [53].

In the manufacture of quartz single crystals, Tl compounds can provide wavelength selection (optical windows) [54].

3.3. Uses

Since the 1970s, it has not been permitted to use thallium compounds for insect pest control in Europe and the United States, so that they are now sold only for special applications. Thallium compounds have also been proposed for impregnating wood and leather to kill fungal spores and bacteria, and for the protection of textiles from attack by moths. A solution of 0.2 g thallium(III) sulfate in 1 L water is suitable both for wood impregnation and seed treatment.

In the manufacture of ethylene oxide, catalysts containing 1 mmol Tl per kilogram catalyst have certain advantages [55].

When recycling polyurethanes at 120 – 200 °C, Tl_2O_3 0.01 mol/L and catalysts are added to the solutions [56].

Not only colloidal thallium metal, but also its compounds, especially the benzoate, oleate, and amyl alcoholate have been proposed as antiknock agents for internal combustion engines. These compounds are added to the fuel as insoluble suspensions or soluble compounds, or are introduced into the cylinders in other ways [57].

Binary mixed crystals of thallium halides, especially TlBr – TlI and TlCl – TlBr, are highly transparent in the infrared. They are used for the manufacture of plates, prisms, and lenses for infrared instruments [8].

Addition of thallium to the filling materials of high-pressure mercury vapor lamps improves light emission considerably. If thallium iodide is used in place of the metal, provided that certain operating conditions are observed, the spectrum can be modified, i.e., the shade can be corrected [58]. Thallium salts are also added to fluorescent materials as activators.

Thallium(I) sulfide is one of the most important photosemiconductors, and is used in the manufacture of photocells. These are mainly constructed as barrier layer photoconductive cells, i.e., the resistance of the cell is reduced by absorption of light. The cells are especially sensitive in the long wavelength visible and near infrared, and are superior to selenium cells at low radiation intensities [59].

3.4. Environmental Aspects

As the natural concentration of thallium in the Earth's crust is 1 – 3 ppm, it does not constitute a danger to groundwater or rivers (Rhine water, 2 ppb). The most important sources of air pollution are smelting and roasting plants for iron and nonferrous metals, and lignite power stations. Over 95 % of the Tl is retained in cyclones and hot electrostatic filters [18]. The total thallium content of all the lignite used in Germany is 1.7 t/a. Of this, 95 % is retained in the ash, and only 5 % is released to the atmosphere (at an exhaust gas concentration $< 1\ \mu g/m^3$).

4. Analysis

The quantitative determination of thallium is carried out with the aid of the grass-green coloration of the flame, and the spectral line at 535.1 nm, which is easily distinguished from the barium band at 534.7 nm by its sharpness and intensity. There are several wet chemical methods: precipitation of the monovalent ion by hydrochloric acid; the very sensitive reaction with potassium iodide to bright yellow thallium iodide (minimum detectable Tl concentration 1 in 3×10^5); precipitation from neutral solution as the yellow chromate, Tl_2CrO_4; or, the most sensitive detection method, precipitation with thionalide as a lemon-yellow complex (concentration limit 1 in 10^7).

Quantitative gravimetric determination is by precipitation as thallium(I) iodide or chromate.

The reaction:

$$Tl^{3+} + 3\,KI \longrightarrow TlI + I_2 + 3\,K^+$$

is used for volumetric determination, the liberated iodine being titrated against thiosulfate after addition of starch.

The Blumenthal permanganate precipitation is used to concentrate traces of thallium, ether being used to extract Tl^{3+} (Tl^+ cannot be extracted quantitatively with ether).

For the determination of thallium in the microgram and nanogram regions, atomic absorption spectrometry is used, by flame atomic absorption or flameless techniques. Optimal limits of detection are:

solid samples (ores, earth samples, dusts, etc.), 0.5 ppm
aqueous samples, 0.001 mg/L
biological materials, 0.004 ppm [60]

Field desorption mass spectrometry (FD-MS) enables detection and quantitative determination of metallic cations from biological samples and environmental samples in the picogram range, without a separate conversion to ash. With this new technique for the determination of trace and ultratrace quantities, the limits of detection for thallium are ca. 10^{-11} g. The extreme specificity of FD-MS makes elaborate sample preparation unnecessary, except for the tissue and cell components of samples, which must be homogenized and centrifuged; the time required per analysis is only ca. 20 min. The sample size need be only ca. 1 mg [61].

5. Economic Aspects

The limited demand can be fully met by suppliers, as the total world demand in 1988 was only ca. 6 – 8 t. No increase is expected for either the metal or its compounds. Annual production in 1975 was 15 t. The amount of thallium available annually from raw materials used in metal refineries is estimated as > 600 t/a.

The price of the metal has increased tenfold from 1975 to 1990, and in 1990 was ca. 170 US $/kg in 1990. The price of the 99.999 % metal was ca. 350 US $/kg [62].

6. Toxicology and Occupational Health [63, 65, 66]

Thallium is a well-known, highly effective poison. Intoxications occurred in former times due to its use as a medicinal agent for many diseases such as venereal disease, dysentery, and tuberculosis. Another main source of accidental, suicidal, and homicidal exposures was its use as a rodenticide which, however, has been banned during the last decades in many countries. Thallium may be accumulated in mushrooms.

The toxicological significance of thallium is mainly restricted to some inorganic and organic salts of Tl^+ such as TlCl, Tl_2SO_4 (used as a rodenticide [67]), and thallium acetate. Intoxications by elemental thallium are comparatively rare. There is some evidence that, analogously to mercury, Tl^{3+} can be converted in the environment to methylated compounds by microbial action [68], and thus represents a potential cumulative hazard.

Toxicokinetics. Tl^+ ions are readily absorbed from the gastrointestinal tract and distributed into various tissues, resulting in high concentrations in kidneys, myocardium, testes, salivary glands, intestine, skeletal muscle, thyroid and adrenal glands. Uptake of thallium-containing dusts via the respiratory tract results in increased thallium toxicity. Transdermal absorption is also possible, especially after the use of thallium-containing ointments. Thallium ions are able to cross the placental barrier and can produce nail abnormalities and alopecia in fetuses exposed during the last trimester [69]. Thallium ions do not undergo metabolic reactions comparable to those of organic xenobiotics (e.g., hydroxylation and conjugation reactions).

Excretion of Tl^+ ions after strong acute intoxication mainly occurs via the feces. Due to pronounced enterohepatic circulation, the excretion half-life varies widely in the range 1.7 – 30 d. Only 3 % of the entire body burden is excreted per day in the urine [64, 70].

The mechanism of toxicity is not completely understood. There is evidence for interaction of Tl^+ ions with thiol groups, resulting in reactions with a variety of proteins and in inhibition of enzymes such as succinic dehydrogenase. The exchange of the activating K^+ ion by the non-activating Tl^+ ion leads to inhibition of Na^+, K^+-ATPase. Furthermore, Tl^+ is stored in axonal mitochondria, resulting in their destruction [71].

Acute and Chronic Toxicity. The most prominent and characteristic features of acute, subchronic, and chronic thallotoxicosis are seen in the nervous system, skin, and cardiovascular system. Concerning the clinical presentation, the time course for the development of symptoms can be divided into four stages [65, 72]:

1) *During 3 – 4* h onset of predominantly gastrointestinal symptoms such as nausea, vomiting, and diarrhea occurs. In severe cases, hematemesis may occur.
2) *Within hours to days* various severe symptoms of the CNS (e.g., disorientation, coma, convulsions, psychosis, cerebral edema with central respiratory failure), PNS (combined motor and sensory neuropathy, including severe hyperesthesia of palms and soles), and ANS (tachycardia, hypertension, fever, salivation, and sweating) develop. Symptoms of the cardiorespiratory system, which may predominate in severe cases, include myocardial necrosis with dysrhythmia and pump failure, hypotension, and bradycardia. Skin symptoms include blue-gray lines on gums, dark pigmentation around the hair roots, and acne. The ophthalmologic symptoms are optic neuritis and ophthalmoplegias.
3) *During 2 – 4 weeks* the skin becomes dry and scaly, and white stripes across the nails (Mees's lines), and scalp and facial hair loss appear as characteristic signs.

4) *During several months* various CNS and PNS abnormalities such as ataxia, tremor, foot drop, and memory loss may persist.

Genotoxicity. Chromosome aberrations and an increased frequency of DNA breaks were found in embryonic cell cultures. After chronic thallium exposure, precancerous lesions in the female genital tract occurred in mice [64, 73].

Reproductive Toxicity. The reproductive system is highly susceptible to thallium toxicity; in humans some of the highest concentrations have been found in the testes after Tl poisoning [64]. Thallium has been shown to produce teratogenic effects in chick embryos, including achondroplasia, leg bone curvature, parrot beak deformity, microcephaly, and reduced fetal size; the results of teratological investigations in rats, mice, and cats are conflicting [73].

Toxicological Data and Analysis. The average lethal dose of thallium salts is ca. 1 g for adults (ca. 10 – 15 mg/kg body weight). Low blood thallium levels are only an indication for an exposure, but do not allow conclusions on the degree of an intoxication to be drawn. However, levels above 300 µg/L blood or 100 µg/L urine indicate severe intoxication (0.3 µg/L urine is the average level in unexposed persons). Moreover, the incorporation of thallium into scalp hair is of analytical significance, 7 – 15 ng/g hair being the normal range.

Treatment. Decontamination of the gastrointestinal tract such as by use of syrup of ipecac, lavage, serial doses of charcoal, and use of cathartics are useful during the first few hours after ingestion. Since thallium occurs as a monovalent cation, chelating agents used for the enhancement of excretion such as dithiocarb, $CaNa_2EDTA$, BAL (British anti-Lewisite, dimercaprol, 2,3-dimercaptopropanole), and penicillamine are not effective. The combination of hemodialysis (protecting the kidneys) and potassium diuresis has been recommended in the case of severe poisoning. However, hemoperfusion clearance seems to be superior to hemodialysis clearance.

Prussian blue, potassium ferricyanoferrate, in daily doses up to 24 g (combined with a mild cathartic) may be used to interrupt the marked enterohepatic circulation by exchanging Tl^+ for K^+. The resulting compound cannot be absorbed and therefore causes only minor side effects [71]. Moreover, it does not liberate substantial amounts of cyanide.

Occupational Health. The MAK value is 0.1 mg/m^3 soluble Tl compounds measured as total dust. In the United States and the United Kingdom the TLV-TWA is 0.1 mg/m^3.

Occupational diseases caused by thallium are included in the list of the German Berufskrankheitenverordnung (Occupation Disease Regulations) [73].

Since thallium tends to accumulate, persons exposed at the workplace should be monitored adequately even in cases of low chronic exposure.

7. References

General References

1. *Gmelin,* Thallium, System no. 38. *Ullmann,* 3rd ed., **17,** 299.
2. J. De Ment, H. C. Dake: *Rarer Metals,* Temple Press, London 1949.
3. N. Lowitzki: "Neueres Schrifttum über Thallium aus den Jahren" 1938 – 1948, *Z. Erzbergbau Metallhüttenwes.* **3** (1950) 201.
4. H. E. Howe, A. A. Smith, Jr.: "Properties and Uses of Thallium," *J. Electrochem. Soc.* **97** (1950) 167 C.
5. L. Sunderson: "Thallium," *Can. Min. J.* **65** (1944) 624.
6. W. H. Waggaman, G. G. Heffner, E. A. Gee: "Thallium," US Dept. of the Interior, Bureau of Mines Information Circular 7553, 1950.
7. V. Tafel: *Lehrbuch der Metallhüttenkunde,* vol. 2, Hirzel, Leipzig 1953.
8. D. E. Eilertsen: "Thallium, Mineral Facts and Problems," US Dept. of the Interior, Bureau of Mines, Bulletin 585, 1960.
9. C. A. Hampel: *Rare Metals Handbook,* Reinhold Publ., New York 1961.
10. W. Schreiter: *Seltene Metalle,* vol. 3, VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1962.
11. R. Kleinert: "Thallium, ein seltenes Metall, ein Begleitmetall," *Erzmetall* **16** (1963) 67.
12. R. Kleinert: "Die Verwendung von Thallium," *Erzmetall* **18** (1965) 363.

13. A. G. Lee: *The Chemistry of Thallium,* International Idens. Inc., Philadelphia 1974.
14. H. R. Babitzke: "Thallium, Mineral Facts and Problems," US Dept. of the Interior, Bureau of Mines, Bulletin 667, 1975.
15. W. W. Stanzo, M. B. Tschernenko: *Die chemischen Elemente Antimon – Wismut, Bausteine der Erde,* vol. 3, Verlag MIR, Moskau, Urania Verlag, Leipzig 1976.
16. R. J. De Filippo: "Thallium, Commodity Data Summeries," US Dept. of the Interior, Bureau of Mines, 1977.
17. J. C. Smith, B. L. Carson: *Trace Metals in the Environment,* vol. 1, Ann Arbor Sci. Publ., Ann Arbor, Mich. 1977.
18. Abfallwirtschaft Forschungsbericht 10301327 des Bundesministers des Innern, Federal Republic of Germany.

Specific References

19. F. A. Cotton, G. Wilkinson: *Anorganische Chemie,* 2nd ed., Verlag Chemie, Weinheim 1970.
20. E. Pilgrim: *Entdeckung der Elemente,* Mondus, Stuttgart 1950.
21. W. Schreiter: *Seltene Metalle,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1960.
22. J. Fischer: "Die Dampfdruckkurve des Thalliums," Festschrift der TH Breslau, Verlag Korn, Breslau 1935.
23. W. Meissner, H. Franz, H. Westerhoff, *Ann. Phys. (Leipzig)* **13** (1932) 555.
24. C. J. Smithells: *Metals Reference Book,* Butterworth, London 1962.
25. *Am. Ceram. Soc. Bull.* **68** (1989) no. 5, 1070.
26. B. Mason: *Principles of Geochemistry,* Wiley, New York 1958.
27. P. Ramdohr: *Die Erzmineralien und ihre Verwachsungen,* Akademie Verlag, Berlin 1960.
28. K. H. Wedepohl: *Handbuch der Geochemie,* **Element 81, vol. 2, no. 3,** Springer Verlag, Berlin 1972.
29. J. Feiser, *Erzmetall* **15** (1962) 578.
30. J. Feiser, personal communication, 1963.
31. M. T. Koslovskij et al., *Tr. Inst. Khim. Nauk Akad. Nauk Kaz. SSR* **3** (1958) 5.
32. Z. A. Serikov, S. M. Anisimov, *Izv. Vyss. Uchebn. Zaved. Chern. Metall.* **3** (1960) no. 6, 65.
33. I. D. Prater, D. Schlain, S. F. Ravitz, US Dept. of the Interior, Bureau of Mines, report of investigations no. 4900, 1952.
34. W. Langner, A. Göbel, *Erzmetall* **3** (1950) 370.
35. W. Kangro, F. Weingärtner, *Erzmetall* **11** (1958) 70.
36. L. G. Plechanov, *Izv. Akad. Nauk Kaz. SSR Razdel. Metallurgii* **4** (1957) 38.
37. A. Gäumann, *Schweiz. Arch.* **21** (1955) 337.
38. A. A. Sokol, L. F. Kozin, *Ukr. Khim. Zh. Kiev* **25** (1959) no. 2, 249.
39. A. R. Powell, The Institution of Mining and Metallurgy, London 1950.
40. Duisburger Kupferhütte, DE 954 236, 1954.
41. H. Meister, *Angew. Chem.* **69** (1957) 533.
42. Chem. Werke Hüls, DE 942 989, 1954.
43. Licentia Patentverwaltungs-GmbH, DE 1 045 999, 1957.
44. T. J. Darvojd, V. N. Vigdorovic, N. A. Iordanskaja, *Izv. Akad. Nauk SSSR Otd. Tekh. Nauk Metall. Topl.* **3** (1961) 55.
45. P. R. Mallory & Co., US 2 180 845, 1939.
46. F. Ensslin, *Erzmetall* **15** (1962) 419.
47. Unterharzer Berg- und Hüttenwerke, DE 740 780, 1942.
48. *Chem. Abstr.* **68** (1968) 98239m.
49. A. E. Roach, C. L. Goodzeit, P. A. Totta, *Nature (London)* **172** (1953) 301.
50. A. Hoffman, F. Rose, *Z. Phys.* **136** (1953/54) 152.
51. H. Moser, *Phys. Z.* **37** (1936) 885.
52. S. S. Flaschen, *J. Am. Ceram. Soc.* **42** (1959) 450; **43** (1960) 168, 247.
53. Westinghouse Electric Corp., US 3 929 970, 1975; 3 929 976, 1975.
54. Schafer, *Proc. SPIE Int. Soc. Opt. Eng.* **285** (1981) 183.
55. US 4 267 073, 1981.
56. JP 8 199 244, 1981.
57. Ethyl Corp., US 3 328 440, 1967.
58. D. A. Larson et al., *Illum. Eng. (N.Y.)* Preprint no. 29 (1962).
59. W. Brugel: *Physik und Technik der Ultrarotstrahlung,* Vincentz, Hannover 1961.
60. H. R. Schulten, C. Achenbach, U. Bahr, F. Kochler, R. Ziskoven, *Angew. Chem.* **91** (1979) no. 11, 944. C. Achenbach et al., *J. Toxicol. Environ. Health* **6** (1980) 519 – 528.
61. M. Sager, G. Tölg: "Spurenanalytik des Thalliums," in: *Analytiker-Taschenbuch,* vol. 4, Springer Verlag, Berlin 1984, pp. 443 – 466.
62. *Am. Ceram. Soc. Bull.* **69** (1990) no. 5, 885.
63. L. Manzo, E. Sabbioni: "Thallium," in H. G. Seiler, H. Sigel, A. Sigel (eds.): *Handbook on Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988.
64. M. J. Ellenhorn, D. G. Barceloux: *Medical Toxicology, Diagnosis and Treatment of Human Poisoning,* Elsevier, New York 1988.

65. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 7. ed., Georg Thieme, Stuttgart 1986.
66. H. U. Wolf: "Thallium" in H. U. Wolf (ed.): *HAGERs Handbuch der Pharmazeutischen Praxis,* 5. ed., **vol. 3: Gifte,** Springer-Verlag, Heidelberg 1992.
67. W. P. Ridley, L. J. Dizikes, J. M. Wood, *Science* **197** (1977) 329.
68. S. Moeschlin, *Clin. Toxicol.* **17** (1980) 133 – 146.
69. W. Stevens et al., *Int. J. Clin. Pharmacol. Ther. Toxicol.* **10** (1974) 1 – 22.
70. P. S. Spencer et al., *J. Cell. Biol.* **58** (1973) 79 – 95.
71. F. H. Lovejoy, *Clin. Toxicol. Rev.* **4** (1982) 1 – 2.
72. E. Sabbioni, L. Manzo in L. Manzo (ed.): *Advances in Neurotoxicology,* Pergamon Press, Oxford 1980.
73. Berufskrankheitenverordnung, BK Nr. 1106: Erkrankungen durch Thallium oder seine Verbindungen.

Thermoelectricity

VOLKER TEGEDER, Siemens AG, Erlangen, Federal Republic of Germany

CHARLOTTE TEGEDER, Universität Erlangen, Federal Republic of Germany

1. Introduction 1
2. Seebeck Effect 2
2.1. Seebeck Coefficient 2
2.2. Thermoelectric Power 3
2.3. Determination of Absolute Thermoelectric Power 4
2.4. Source of Thermo-emf 4
2.4.1. Interaction with the Lattice 5
2.4.2. Effect of Charge Carriers 5
2.5. Examples 5
2.5.1. Thermoelectric Magnet 5
2.5.2. Thermocouples 6
2.5.3. Determination of Conduction Type . . 6
2.5.4. Radiation Measurement 6
2.5.5. Thermoelectric Generator (Seebeck Element) 6
3. Peltier Effect 7
3.1. Peltier Coefficient 7
3.2. Demonstration of the Peltier Effect 8
3.3. Peltier Module 8
3.4. Optimization of the Peltier Element 8
3.5. Applications 9
3.6. Other Effects 9
4. Thomson Effect 9
4.1. Demonstration of the Thomson Effect 10
4.2. Other Homogeneous Thermoelectric Effects 11
4.3. Kelvin Relations 11
5. References 11

1. Introduction

Thermoelectric effects result from interactions between thermal and electrical transport processes in electrical conductors in either the solid or the liquid state. For example, in an open circuit made up of two unlike electrically conducting materials that form two junctions, a thermoelectric emf (voltage) can be measured when the junctions are at different temperatures (Fig. 1). This behavior is known as the Seebeck effect.

Thermoelectric phenomena fall into two categories, homogeneous and inhomogeneous. *Inhomogeneous thermoelectric effects* include the Seebeck effect (or the thermoelectric effect in the narrow sense, see Chap. 2) and the Peltier effect (see Chap. 3). *Homogeneous thermoelectric effects* (see Chap. 4) include the Thomson, Benedicks, Bridgman (or internal Peltier effect), and phonon-drag effects (Peltier effect of the lattice). The two classes differ in that inhomogeneous effects occur only when unlike materials are combined.

One well-known application of the Seebeck effect is the thermocouple, which transforms thermal to electrical energy and can thus be used to measure temperature (see Section 2.5.2).

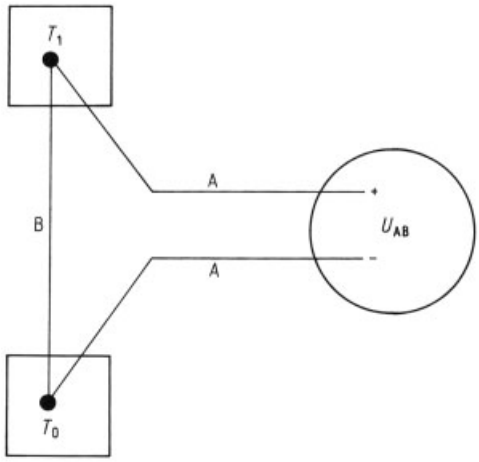

Figure 1. A thermoelectric element consisting of two unlike materials A and B; if $S_{AB} > 0$, the potential U_{AB} is present when the junctions are at temperatures T_1 and T_0 ($T_1 > T_0$)

The *Peltier effect* is practically the inverse of the Seebeck effect. If a current flows through the circuit already described, heat is reversibly absorbed or liberated at the junctions. The *Thomson effect* is the reversible absorption or liberation of heat when a current flows through a single, homogeneous conductor over whose length a temperature gradient is maintained.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_621

To a good approximation, these relatively weak thermoelectric effects can be described by linear equations. Three coefficients occur in these equations: the absolute thermoelectric power (thermopower) S, the Peltier coefficient π, and the Thomson coefficient μ. Each of these has a well-defined value for a homogeneous conductor at a constant temperature. The three coefficients are linked by the Kelvin (Thomson) relations so that if one is known at all temperatures, the other two can be calculated. For ease of measurement, the thermopower is usually determined experimentally. The electrical resistivity, thermal conductivity, and thermopower sufficiently describe the electron-transport properties of a conductor.

The principal applications of thermoelectricity are in the study of interactions between electrons and lattice vibrations, the detection of impurities in solids, the measurement of temperature, and cooling and heating.

2. Seebeck Effect

In 1822, Thomas Johann Seebeck, working with a system of copper and bismuth as diagrammed in Figure 1, discovered that a voltage is generated in a circuit comprising two unlike conductors between whose junctions a temperature gradient is present [1–4].

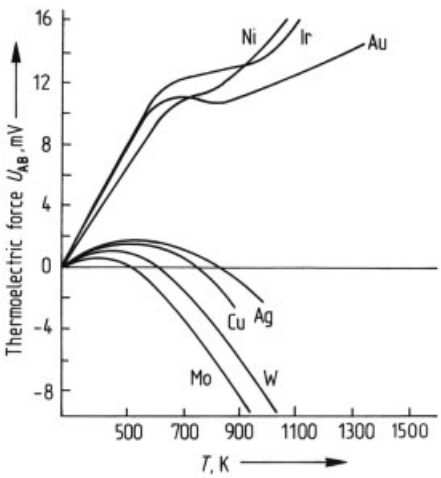

Figure 2. Thermoelectric force U_{AB} in couples consisting of various metals (B) and an iron reference, plotted versus the temperature T of one junction; the other junction is at $T = 273$ K [5]

Figure 2 shows the thermoelectric force U_{AB}, plotted versus temperature for several metals (conductor B in Fig. 1). Iron is the reference material (conductor A) in each case.

2.1. Seebeck Coefficient

The thermo-emf is a function of the materials in contact and the temperature difference ΔT between the junctions. If the temperature difference ΔT is small, so that $S(T)$ is virtually constant, the voltage is described by

$$U_{AB} = S_{AB}\,\Delta T \tag{1}$$

The constant of proportionality S_{AB} is called the differential thermoelectric force or the relative Seebeck coefficient [5–12]. S_{AB} is given by the Seebeck coefficients S'_x of the two materials:

$$S_{AB} = S'_A - S'_B$$

The Seebeck coefficient is thus a bulk property of the material, with units μV/K; its sign depends on the type of charge carrier and the way in which the carriers interact with the lattice. The thermo-emf is independent of the shape of the materials used and the detailed temperature distribution between the junctions. The thermo-emf thus has its source in the materials themselves rather than in the contact, but it can be measured only when unlike materials are in contact, for example, at the junctions of a thermocouple.

The Seebeck effect must be distinguished from the Volta effect. The junction voltage seen in the latter is due to differences in the electron work functions of the two materials and is present even when there is no temperature difference. The thermo-emf is caused by redistributions of charge carriers due to the temperature gradient.

Phenomenologically, all thermoelectric materials (metals, alloys, semiconductors) can be ranked in a thermoelectric series by the value of the thermo-emf against a reference such as copper. One (soldered or brazed) junction of the thermocouple is held at 273 K. In a thermocouple consisting of two elements from this thermoelectric series (Table 1), if the junction is heated, the element standing lower in the series takes on a positive potential, while the element standing higher takes on a negative potential.

Table 1. Thermoelectric series for 100 K temperature difference (thermo-emf values in mV) [13]

Metal x at 273 K	Platinum (373 K) vs.	Bismuth (373 K) vs.	Copper (373 K) vs.
Bi	−7.0	0.0	−8
Constantan*	−3.4	−4.1	−4.1
Ni	−1.5	+5.5	−2.2
Pd	−0.3	+6.7	−1.0
Pt	0.0	+7.0	−0.8
Hg	0.0	+7.0	−0.8
C	+0.2	+7.2	−0.6
Al	+0.4	+7.4	−0.4
Pb	+0.4	+7.5	−0.4
Sn	+0.4	+7.5	−0.3
Manganin **	+0.6	+7.6	−0.1
Ir	+0.7	+7.7	−0.1
Rh	+0.7	+7.7	−0.1
Zn	+0.7	+7.7	0.0
Ag	+0.7	+7.7	0.0
Au	+0.7	+7.7	0.0
W	+0.8	+7.8	+0.1
Fe	+1.8	+8.8	+1.0
Si	+45	+52	+44
Te	+50	+57	+49
Se		+85	+98

* Alloy of 54 % Cu, 45 % Ni, 1 % Mn. ** Alloy of 86 % Cu, 12 % Mn, 2 % Ni.

2.2. Thermoelectric Power

When larger temperature differences are involved, the differential thermoelectric force must be regarded as a function of temperature given by

$$S_{\mathrm{AB}}(T) = \mathrm{d}U_{\mathrm{AB}}/\mathrm{d}T \tag{2}$$

(Eq. 1 holds for small ΔT). The function $S_{\mathrm{AB}}(T)$ is commonly referred to as the thermoelectric power. Figure 3 shows how the thermoelectric power varies with temperature for several materials against copper. The curves are simpler in form than the ones for the thermo-emf (Fig. 2). The thermoelectric power for metals at higher temperatures is generally a linear function of temperature; transition metals exhibit more complicated behavior. The temperature dependence is weaker for semiconductors at higher temperatures.

Accordingly, the x component of the electric field in a homogeneous conductor is, in general, given by

$$E_x = S_{\mathrm{AB}}(T)\ \mathrm{d}T/\mathrm{d}x$$

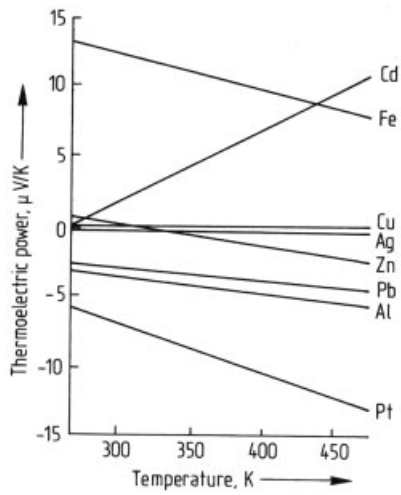

Figure 3. Thermoelectric power S_{AB} as a function of temperature for some metals against copper [5]

If the function $S_{\mathrm{AB}}(T)$ is known, the voltage U_{AB} can be obtained by integrating this equation

$$U_{\mathrm{AB}}(T_0,T_1) = \int_{T_0}^{T_l} S_{\mathrm{AB}}(T)\ \mathrm{d}T$$

2.3. Determination of Absolute Thermoelectric Power

To determine the absolute thermoelectric power (thermopower) S_{A} of a material, the material can be formed into a thermocouple with a superconductor. The thermo-emf of the superconductor below its transition temperature is 0 V, so that the measured thermo-emf is entirely due to the normally conducting part of the couple, since $S_{\mathrm{AB}} = S_{\mathrm{A}} - S_{\mathrm{B}}$. At present, however, this method is restricted to temperatures of ≤ 24 K. At higher temperature (higher than the transition temperature of superconductors), lead (because of its low thermo-emf) or platinum (because it can be prepared in high purity) is used as the thermoelectric reference.

The thermopower $S_{\mathrm{A}}(T) = \mathrm{d}U_{\mathrm{A}}/\mathrm{d}T$ (in μV/K) for a given conductor with one end at 0 K and the other at the test temperature must be distinguished from the Seebeck coefficient (also in μV/K) S_{A}, which is the thermo-emf of a thermocouple in which a platinum reference forms the

Table 2. Absolute thermoelectric powers (μV/K) for some metals [14, 15]

T, K	Pb	Cu	Ag	Au	Pt	Pd	W	Mo
100	−0.58	1.19	0.73	0.82	4.29	2.00		
200	−0.83	1.29	1.05	1.34	−1.27	−4.85		
273		1.70	1.38	1.79	−4.45	−9.00	0.13	4.71
300	−1.05	1.84	1.51	1.94	−5.28	−9.99	1.07	5.57
400		2.34	2.08	2.46	−7.83	−13.00	4.44	8.52
500		2.83	2.82	2.86	−9.89	−16.03	7.53	11.12
600		3.33	3.72	3.18	−11.66	−19.06	10.29	13.27
700		3.83	4.72	3.43	−13.31	−22.09	12.66	14.94
800		4.34	5.77	3.63	−14.88	−25.12	14.65	16.13
900		4.85	6.85	3.77	−16.39	−28.15	16.28	16.86
1000		5.36	7.95	3.85	−17.86	−31.18	17.57	17.16
1100		5.88	9.06	3.88	−19.29	−34.21	18.53	17.08
1200		6.40	10.15	3.86	−20.69	−37.24	19.18	16.65
1300		6.91		3.78	−22.06	−40.27	19.53	15.92
1400					−23.41	−43.30	19.60	14.94
1600					−26.06	−49.36	18.97	12.42
1800					−28.66	−55.42	17.41	9.52
2000					−31.23	−61.48	15.05	6.67
2200							12.01	4.30
2400							8.39	2.87

second leg and one junction is held at a constant 273 K. The thermopower (Table 2) depends on just one temperature. The sign convention states that the polarity is as shown in Figure 1 when $S_{AB} = S_A - S_B > 0$ ($T_1 > T_0$).

2.4. Source of Thermo-emf

The experimentally determinable thermo-emf is present because the heat flux due to a temperature difference in a metal or semiconductor is carried partly by phonons (quantized lattice vibrations) and partly by charge carriers, which transport energy along with their charge. In metals, only electrons (charge carriers) with energies near the Fermi level E_0 (Fig. 4) contribute to the conductivity. The electron kinetic energy distribution depends on temperature. High-energy electrons from the hot end and low-energy electrons from the cold end diffuse along the conductor at different velocities because of different interactions with the lattice, so that a net current flows until the voltage resulting from this nonuniform distribution stops any further net current flow. The Fermi level also varies with temperature.

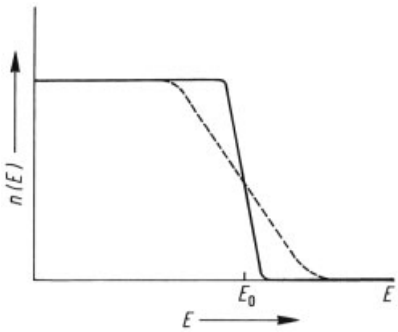

Figure 4. Electron-density distribution $n(E)$ versus electron energy E near the Fermi level E_0 for two temperatures
The dashed curve represents the distribution at the higher temperature

2.4.1. Interaction with the Lattice

The strength of interactions with the lattice and phonons is important in determining the thermo-emf generated. Especially at very low temperature, the phonon-drag effect (Gurevich effect) increases the thermopower, which is given by

$$S = S_d + S_g$$

where S_d is the electron diffusion component and S_g is the phonon-drag component. This additional thermopower contribution occurs when electrons are entrained by phonons [16] moving from the hot to the cold portion of the conductor because of the temperature difference. The interaction of electrons with these phonons decreases as the temperature increases, since phonon–phonon interactions come to dominate at higher temperature. For this reason, the phonon-drag

effect cannot be employed in practical situations (i.e., at higher temperature; see Fig. 5).

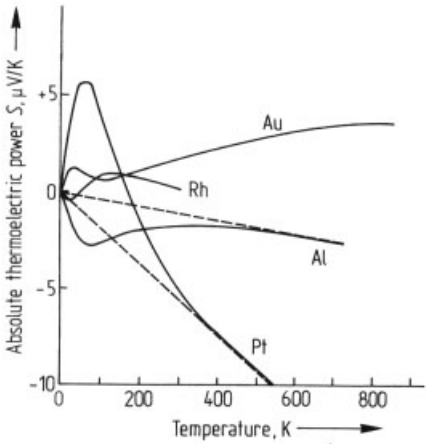

Figure 5. Increase in absolute thermoelectric power at low temperature due to phonon drag [16]

2.4.2. Effect of Charge Carriers

Thermopower is a function of the carrier concentration. In semiconducting materials, the thermopower increases as the electron or hole density declines; it is thus inversely proportional to the dopant concentration, because the mean carrier energy as a function of temperature increases faster, as carrier concentration decreases. Semiconductors, with their much lower carrier density, thus have much higher thermopowers than metals. The thermopower S can range up to 1 mV/K for semiconductors; it is commonly a few μV/K for metals.

2.5. Examples

2.5.1. Thermoelectric Magnet

Even though thermo-emf values are very small, relatively large currents occur in some thermocouples if the resistance of the system is low enough. The following experiment demonstrates these high thermoelectric currents [13]: A copper wire A (diameter 10 mm) is bent into a loop; one end is placed in a beaker of ice water, while the other end is heated over a candle flame. The circuit is completed by two sections B of Constantan brazed to the copper. A thermo-emf of ca. 4.2×10^{-3} V is produced. The thermoelectric current of some 40 A can be detected by its magnetic action on two pieces of iron C_1 and C_2 above and below the loop; C_2 can be loaded additionally by a weight of several kilograms.

2.5.2. Thermocouples

Thermocouples used for temperature measurement (Fig. 6) are the most important application of the Seebeck effect. A thermocouple is a conducting circuit made up of two or more unlike materials with the junctions held at unequal temperatures [17–24]. One of the junctions is kept at a reference temperature (usually melting ice in water, 273 K), while the other takes on the temperature to be measured.

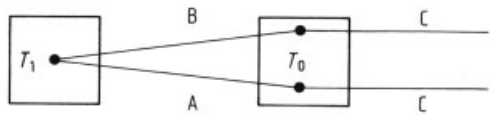

Figure 6. Block diagram of a thermocouple made of metals A, B, and C and used for temperature measurement

As a consequence of the Seebeck effect, a thermo-emf is generated in the circuit, causing a thermoelectric current to flow. This current is measured with appropriate instruments; the result can be converted to a temperature through use of a table. Many combinations of materials are used in thermocouples in order to cover various temperature ranges. Material pairs for thermocouples and ranges of temperature measured are listed below:

Copper – Constantan	73 – 673 K
Iron – Constantan	up to 1073 K
Platinum vs. platinum – 10 % rhodium	up to 1573 K
Nickel vs. chromium-nickel	up to 1373 K
Constantan vs. chromium – nickel	up to 1073 K
Chromel (alloy of ca. 90 % Ni, ca. 9 % Cr, and minor amounts of Fe, C, Mn, Cu, Si) vs. Alumel (Ni – Al – Mn)	up to 1473 K
Bi_2Te_3	273 – 573 K
$Bi_2Te_{2.4}Se_{0.6}$	273 – 573 K
$Pb_{1-x}Sn_xTe$ ($x = 0 - 0.25$)	423 – 873 K
$Si_{1-x}Ge_x$ ($x = 0.15 - 0.3$)	873 – 1473 K

Because the sign of the thermopower in semiconductors depends on whether the charge is

carried by electrons (−) or holes (+), a thermocouple made up of semiconductors will have the highest efficiency if both legs consist of different conducting materials, i.e., of *p*-type and *n*-type conductors. Despite the higher thermoemf obtained with semiconductors, most thermocouples are fabricated from metals because these can be produced with reproducible results at lower cost and can more easily be formed into a variety of shapes.

The producer of a thermocouple must take account of the fact that thermoelectric effects occur even in like materials when the slightest variation in chemical composition exists or when mechanical stresses are imposed (i.e., the Fermi energy changes). For example, a thermocouple made of annealed brass wire and hard drawn brass wire will behave like a thermocouple made of unlike metals; a thermoelectric effect will also be seen if a wire under elastic stress is paired with an unstressed wire or if crystals of the same metal are joined with differently oriented faces coming together at the junction.

2.5.3. Determination of Conduction Type

The conduction type of a semiconducting material can be determined by contacting the material with two metal tips, one of which is heated. The polarity of the cold tip will indicate the predominant type of charge carrier.

2.5.4. Radiation Measurement

Thermocouples can also be employed for the sensitive measurement of microwave pulse power, as thermoelectric transducers, and as thermoelectric radiation detectors (theoretical detection threshold 3×10^{-11} W, practical threshold 3×10^{-10} W).

2.5.5. Thermoelectric Generator (Seebeck Element)

A thermoelectric generator or heat engine utilizing the Seebeck effect can be made in a configuration similar to a thermocouple (Fig. 7). Here, however, the two legs composed of unlike materials are joined by a plate made of a third material. Each leg has a contact. The plate joining the legs is heated relative to the contacts, and electrical energy can be removed at the contacts [24, 25]. A thermoelectric generator combined with a heat source (e.g., the decay of radioactive materials, solar radiation, or combustion) can generate energy for isolated stations such as lighthouses, navigation buoys, remote weather stations, offshore drilling platforms, and man-made satellites.

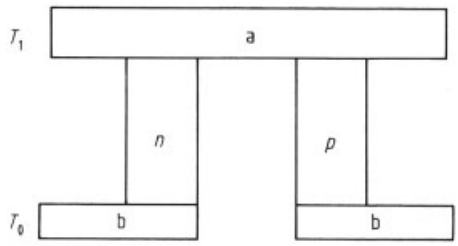

Figure 7. Block diagram of a thermoelectric generator with heavily doped connecting plate (a), metallic contacts (b), and *n*- and *p*-type semiconductors

An alloy of 70 % silicon and 30 % germanium has proved suitable for the legs of a thermoelectric generator.

Generators of this type connected in cascade (thermally parallel, electrically series) offer markedly higher yield, especially if generators of different materials are used in which each material is held at its most efficient temperature. For multistage cascade circuits, an efficiency of 22 % ($\Delta T = 1000$ K) appears possible, and 10–15 % efficiencies have been attained; the corresponding value for single-stage circuits is 7 % ($\Delta T = 800$ K). An output of 1.7 W has been obtained at $T_1 = 1273$ K (T_2 = room temperature) with a generator having arm measurements of 6 mm × 6 mm × 6 mm (Table 3).

Table 3. Parameters of important semiconductors for thermoelectric generators [26, 27]

Material	Temperature range, K	Conduction type	$(z*T)_{max}$
Bi_2Te_3	273–573	*p*	0.6
$Bi_2Te_{2.4}Se_{0.6}$	273–573	*n*	0.7
$Pb_{1-x}Sn_xTe$ (x = 0–0.25)	423–873	*n*	1.1
		p	1.0
$Si_{1-x}Ge_x$ (x = 0.15–0.3)	873–1473	*n*	1.0
		p	0.7

$*\,z$ = Figure of merit (see Section 3.4).

3. Peltier Effect

When a current flows through a circuit made up of two unlike conductors, one junction becomes warm and the other cold. This phenomenon was discovered by the clockmaker and physicist JEAN CHARLES ATHANASE PELTIER in 1834. Working with an antimony – bismuth couple, Peltier was able to freeze a drop of water [7–9, 12, 28, 29]. If the direction of the current is reversed, the effect is also reversed so that the first junction is cooled and the second heated. The heat released (or absorbed) is directly proportional to the current I.

3.1. Peltier Coefficient

The Peltier coefficient π_{AB} is defined as the quantity of heat generated per unit current and second at the junction of materials A and B. It is measured in volts and is by convention positive if that junction becomes warm at which a positive current flows from A to B. The Peltier coefficient does not depend on the current I or the shape and size of the conductors, and the effect is reversible. The coefficient π depends solely on the materials in contact and the junction temperature; it is a measure of the mean energy of the charge carriers in the material. Clearly if

$$\pi_{AB}(T) = \pi_A(T) - \pi_B(T)$$

where π_A and π_B are the Peltier coefficients of materials A and B, the coefficient is a bulk property and can be defined for a single conductor.

The cause of the Peltier effect lies in the energy transported by charge carriers when a current flows. The Peltier heat Q_π (i.e., the heat liberated or absorbed at both junctions when a current flows through the circuit) is given by

$$Q_\pi = \pi_{AB} I t$$

where the product of current I and time t is the quantity of charge transported through the junction.

The Peltier effect is superimposed by simultaneous Joule heating and Thomson heating or cooling, so the Peltier coefficient can be difficultly measured with accuracy. Accordingly, it is commonly determined by applying the Kelvin relations (see Section 4.3) to the thermopower. Table 4 lists some Peltier coefficients determined by calorimetry.

Table 4. Peltier coefficients π_{AB} for various material pairs [30]

Fe – Constantan		Cu – Ni		Pb – Constantan	
T, K	π_{AB}, mV	T, K	π_{AB}, mV	T, K	π_{AB}, mV
273	13	292	8.0	293	8.7
293	15	368	9.0	383	11.8
403	19	478	10.3	508	16.0
513	26	563	8.6	578	18.7
593	34	613	8.0	633	20.6
833	52	718	10.0	713	23.4

3.2. Demonstration of the Peltier Effect

Despite the superimposed Joule effect, the Peltier effect can be demonstrated experimentally. One setup (Fig. 8) has a piece of antimony (a) between two pieces of bismuth (b), with the two junctions sealed into two vessels. The vessels are joined by a U-shaped glass tube with a drop of liquid in it. When a current flows, the liquid drop moves as the junctions are heated and cooled [13].

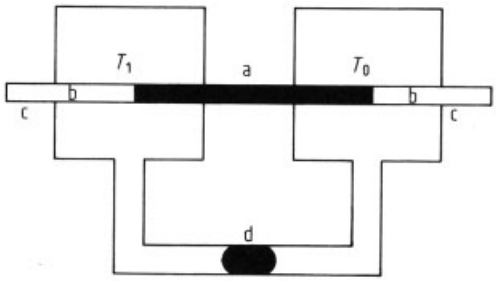

Figure 8. EDLUND's experimental demonstration of the Peltier effect with two unlike metals [31]
a) Antimony; b) Bismuth; c) Electrical contacts; d) Liquid drop
T_1, T_0 = temperatures (unequal) in the chambers

This effect can be demonstrated at high temperature by brazing together an iron and a Constantan wire (each ca. 12 cm long and 0.3 – 0.5 mm in diameter) in such a way that the junction is no bigger than the wires. The wires are placed in an evacuated glass tube, and a current is passed through them until they become incandescent. Depending on the direction of the current, the junction will glow more or less brightly than the wires.

3.3. Peltier Module

The Peltier module (Fig. 9), a specially designed semiconducting thermocouple utilizing the Peltier effect, can be employed for either cooling or heating [32, 33]. The two arms of one element are joined by a piece of metal, which gives off or absorbs heat when a current is passed through the element.

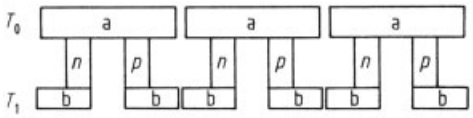

Figure 9. Block diagram of a Peltier module
a) Metal bridge; b) Contact
n, *p* = Semiconductors; T_0 = Lower temperature; T_1 = Higher temperature

3.4. Optimization of the Peltier Element

The selection of materials must take into account the facts that thermal conduction opposes the Peltier effect and that Joule heat is produced because the materials used always have finite electrical resistances. The semiconducting materials must therefore have the best possible electrical conductivity σ and poor thermal conductivity $\varkappa$. Suitable materials are Bi – Sb – Te mixed crystals for low temperature and Si – Ge mixed crystals (oxidation resistant) for high temperature. Commonly, materials differing in conduction type are combined. For good cooling performance, the figure of merit $z = S^2\sigma/\varkappa$ must be made as large as possible; z is the ratio of effective power (proportional to $S^2\sigma$) to thermal conductivity $\varkappa$.

Figure 10 shows how to locate the expected optimum point. Because the thermal conductivity $\varkappa$ is the sum of thermal conduction by electrical charge carriers (electrons or holes) $\varkappa_C$ and by the lattice $\varkappa_L$, and because $\varkappa_C$ (like the conductivity σ) increases with carrier density while thermopower S decreases, an optimum must exist. An optimum for both parameters is located at a free carrier concentration of ca. 10^{19} cm^{-3}. In semiconductors, however, the lattice conductivity $\varkappa_L$ is usually the crucial parameter for thermal conductivity and varies over three orders of magnitude. The best thermoelectric materials are therefore selected in accordance with this parameter $\varkappa_L$.

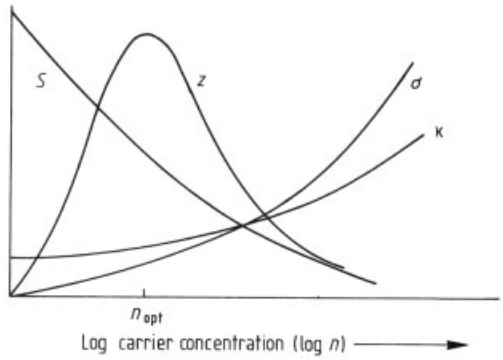

Figure 10. Optimal charge-carrier concentration for a Peltier element as a function of S, σ, and $\varkappa$ [25]

The best efficiency $[z(T) \cdot T]_{max} \approx 1$ is obtained for a few highly doped semiconducting alloys with narrow band gaps. The maximum possible temperature difference ΔT_{max} between the hot and cold junctions of a single stage is $\Delta T_{max} = (1/2)\, z T^2_{cold}$; the efficiency for such a stage approaches zero. Accordingly, little heat is transported, but for some applications, heat transport is not needed. This temperature difference can be increased by a suitable cascade arrangement, but no appreciable gain in efficiency results. The best thermoelectric materials, Pb – Te and Bi – Te, exhibit z values of 2 – 4×10^{-3} K^{-1} at their best temperatures.

Temperature drops as great as 150 °C can be achieved through the choice of materials and cascading of such elements. Single elements normally have maximum temperature differences of 50 – 70 K. Table 5 gives the thermoelectric properties of some semiconductors.

3.5. Applications

The Peltier element is used to cool sensitive solid-state lasers, the specimen stage in microscopy, infrared detectors, and other sensitive electronic (semiconductor) components that must operate at higher ambient temperatures (e.g., in computers). Other uses are cooling of oil-vapor traps in vacuum pumps, of isothermal microcalorimeters, and of medical devices for cryosurgery and cryotherapy. Small, battery-

Table 5. Thermoelectric properties of some semiconductors for Peltier elements [25,34]

Material	E_g, eV	Type	$\varkappa_L$, W m^{-1} K^{-1}	z_{max}, 10^{-3} K^{-1}	T for z_{max}, K	Maximum temperature, K
Bi_2Te_3	0.15	*n*	1.6	2.8	300	450
Bi_2Te_3	0.15	*p*	1.6	2.2	300	450
$BiSb_4Te_{7.5}$		*p*	1.0	3.3	300	450
Bi_2Te_2Se	0.3	*n*	1.3	2.3	300	600
PbTe	0.3	*n,p*	2	1.2	300	900
GeTe (Bi doped)		*p*	1.6	1.6	800	900
ZnSb	0.6	*p*	17	1.2	500	600
$AgSbTe_2$	0.6	*p*	0.6	1.6	700	900
InAs (P doped)	0.45	*n*	6.5		900	1100
Cu_3Te_2S			1.2	1.5	1100	
70 % Si – 30 % Ge		*n*	5	0.9	1100	1200
70 % Si – 30 % Ge		*p*	5	0.7	100	1200
88 % Bi – 12 % Sb		*n*		6.0	80	
88 % Bi – 12 % Sb (magnetic field)		*n*		8.6	100	

powered units can be automatically held within a narrow band around 0 °C (reference temperature) for long periods of time. Multistage devices find use as cooling batteries (electrically parallel/thermally serial configuration or cooling cascade in thermal series).

The advantages of Peltier cooling are higher efficiencies at low cooling power and small ΔT, no rotating parts (hence virtually no wear), small size, ease of control, and low probability of malfunction.

3.6. Other Effects

Phenomena related to the Peltier effect (electrothermal inhomogeneity effect) include the Bridgman effect (or internal Peltier effect), the electron-drag effect (Gurevich effect or lattice Peltier effect), and the transverse Peltier effect.

The *Bridgman effect* occurs when current flows in anisotropic materials; heat is produced when the current changes direction relative to the crystal axes [35–38]. In the *electron-drag effect*, an electron current generates an additional flow of heat through the electron – phonon interaction, which is observed chiefly at very low temperature. The *transverse Peltier effect* occurs in noncubic conducting crystals. The formation of a transverse temperature gradient and the associated longitudinal thermo-emf cause the d.c. resistance to be greater than the a.c. resistance (Lenard effct) since no temperature gradient is present if the frequency is sufficiently high [39].

4. Thomson Effect

In 1854, William Thomson (later Lord Kelvin) postulated a reversible process of heat production when a current flows in a homogeneous conductor along which a temperature gradient is imposed [40–43]. This phenomenon, characterized by the Thomson coefficient μ, is of little practical value but had great theoretical significance, because it was essential for a complete thermodynamic theory of thermoelectricity.

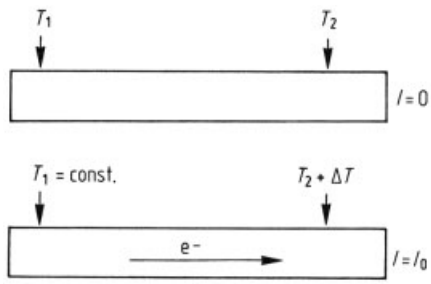

Figure 11. Additional heat flux produced when current flows in a conductor with a temperature gradient

If a temperature gradient dT/dx is present in a homogeneous conductor carrying a current I (Fig. 11), a quantity of heat $dQ_T = (\mu I\, dT/dx)\, dx$ per second is developed over an element of length dx (in addition to the Joule heat). This heat is absorbed or liberated, depending on the sign of μ and the direction of the current relative to the temperature gradient. If the positive

current flows in the direction of rising temperature and $\mu > 0$, heat is released. In Figure 11, therefore, if T_1 is held constant, the temperature T_2 is changed by ΔT as a result of the transported heat.

The Thomson coefficient is positive for silver and gold, and negative for iron, platinum, zinc, nickel, and Constantan (Table 6). The reason for the effect is that charge carriers flowing through the temperature gradient must exchange energy with the lattice in order to maintain local equilibrium, so that heat is created or absorbed.

Table 6. Thomson coefficients for some metals [30]

Pt		Cu		Pb	
T, K	μ, μV/K	T, K	μ, μV/K	T, K	μ, μV/K
203	−9.6	70	−0.26	120	−0.19
233	−9.3	90	−0.49	150	−0.30
273	−9.1	120	−0.26	200	−0.45
313	−9.0	170	+0.47	300	−0.68
353	−9.0	220	+0.96	400	−0.85
393	−9.2	300	+1.52		

In contrast to the inhomogeneous thermoelectric effects (Seebeck and Peltier effects), the Thomson effect is homogeneous. In other words, itcan be observed in a circuit composed of a single (uniform) material; junctions between unlike conductors are not required.

4.1. Demonstration of the Thomson Effect

The Thomson effect can be demonstrated in the following way [13]: Wires of equal length and cross section, connected to a current source, are arranged vertically with their free ends dipping into a bath of mercury. The mercury not only completes the circuit but also acts as a cold reservoir. If a current is passed through the wires so as to make them incandescent, the Thomson heat released or absorbed is visualized as a difference in maximum brightness in different regions of the wires, because the current flows with the temperature gradient in one case and against it in the other.

An inverse Thomson effect, in which a varying temperature gradient produces a thermoemf, must be postulated on grounds of symmetry [44]. This effect has not, however, been observed. The Thomson effect and the inverse Thomson effect are second-order thermoelectric effects.

4.2. Other Homogeneous Thermoelectric Effects

Other homogeneous thermoelectric effects include the two *Benedicks effects* [4,45–48]: (1) A thermo-emf is produced in a homogeneous conductor if a very strong temperature gradient exists along the conductor. (2) A very slight rise in temperature occurs at a restriction in a homogeneous conductor when a current flows through the material.

4.3. Kelvin Relations

The Kelvin (or Kelvin – Onsager) relations link the thermopower S, the Peltier coefficient π, and the Thomson coefficient μ [40–43]:

First Kelvin relation: $\pi = ST$

Second Kelvin relation: $\mu = T\,\mathrm{d}S/\mathrm{d}T$

Thus, if one thermoelectric coefficient is completely known, the other two can be calculated with these equations. The second relation implies that the integral

$$S_{\mathrm{A}} = \int_0^T \frac{\mathrm{d}\mu_{\mathrm{A}}}{\mathrm{d}T}\,\mathrm{d}T$$

makes possible the determination of absolute S_{A} on a single conductor, provided the Thomson coefficient is known. Here $S\,(T = 0)$ has been set equal to zero in accordance with the third law of thermodynamics. Because these measurements are very difficult, they have been performed for only a few substances, such as lead and platinum, which can thus serve as references [49].

5. References

1. T. J. Seebeck, *Abh. Kgl. Akad. Wiss. Berlin* **23** (1822) 265.
2. T. J. Seebeck, *Ann. Phys. (Leipzig)* **73** (1823) no. 1, 430.
3. T. J. Seebeck, *Ann. Phys. (Leipzig)* **6** (1826) no. 2, 1 – 20, 133 – 160.

4. C. A. F. Benedicks, *C.R. Hebd. Seances Acad. Sci.* **167** (1918) 296.
5. F. J. Blatt, P. A. Schroeder (eds.): *Thermoelectricity in Metallic Conductors,* Plenum Press, New York – London 1978.
6. D. D. Pollock: *Thermoelectricity: Theory, Thermometry, Tool,* ASTM STP 852, Am. Soc. Test. Mater., Philadelphia, PA, 1985.
7. D. D. Pollock: *Physical Properties of Materials for Engineers,* vols. I, II, III, CRC Press, Boca Raton, FL, 1982.
8. R. R. Heikes, R. W. Ure, Jr.: *Thermoelectricity; Science and Engineering,* Interscience, New York 1961.
9. R. D. Barnard: *Thermoelectricity in Metals and Alloys,* Taylor & Francis, London 1972.
10. F. J. Blatt et al.: *Thermoelectric Power of Metals,* Plenum Press, New York – London 1976.
11. T. C. Harman, H. M. Honig: *Thermoelectric and Thermomagnetic Effects and Applications,* McGraw – Hill, New York 1967.
12. D. K. C. MacDonald: *Thermoelectricity: An Introduction to the Principles,* J. Wiley & Sons, New York 1962.
13. L. Bergmann, C. Schaefer: *Lehrbuch der Experimentalphysik,* **vol. 4,** De Gruyter, Berlin 1981.
14. R. B. Roberts, *Philos. Mag.* **36** (1977) 91.
15. N. Cusack, P. Kendall, *Proc. Phys. Soc. London* **72** (1958) 89.
16. C. Herring: "The Role of Low-frequency Phonons in Thermoelectricity and Thermal Conductivity," *Proc. Int. Coll.,* Garmisch-Partenkirchen 1956, Vieweg, Braunschweig 1958, p. 184.
17. A. F. Joffé: *Semiconductor Thermoelements and Thermoelectric Cooling,* Infosearch, London 1957.
18. A. F. Joffé: *Halbleiter-Thermoelemente,* Akademie-Verlag Berlin 1957.
19. R. P. Benedict (ed.): *Manual on the Use of Thermocouples in Temperature Measurement,* **ASTM STP 740 B,** Am. Soc. Test. Mater., Philadelphia, PA, 1981.
20. R. P. Benedict: *Fundamentals of Temperature, Pressure and Flow Measurements,* 3rd ed., Wiley, New York 1960.
21. N. A. Burley et al.: "The Nicrosil Versus Nisil Thermocouple: Properties and Thermoelectric Reference Data," *NBS Monogr. (U.S.)* **161** (1978).
22. R. L. Powell et al.: "Thermocouple Reference Tables Based on IPTS-68," *NBS Monogr. (U.S.)* **125** (1974).
23. American Institute of Physics: *Temperature, its Measurement and Control in Science and Industry,* **vol. 1/ vol. 2,**Reinhold Publ. Co., New York, 1941/ 1955.
24. P. Gerthsen: "Thermoelektrische Anwendung von Halbleitern," *Z. Angew. Phys.* **13** (1961) no. 9, 435 –442.
25. D. M. Rowe: "Thermoelectric Power Generation," *Proc. Inst. Electr. Eng.* **125** (1978) (11R), 1113 – 1136.
26. H. Preier et al., *Appl. Phys.* **12** (1977) 277.
27. A. S. Ochotin, A. S. Puskarskij, R. P. Borovikova, V. A. Simonov: "Metody izmerenija charakteristik termoelektriceskich materialov i preobrazovatelej," Moskau: Izd. Nauka 1974.
28. J. C. A. Peltier, *Ann. Chim. Phys.* **56** (1834) 371.
29. P. W. Bridgman: *Thermodynamics of Electrical Phenomena in Metals,* Dover, New York 1960.
30. *Landolt-Börnstein* **II/6,** 59.
31. P. O. Gehlhoff, E. Justi, M. Kohler, *Abh. Braunschw. Wiss. Ges.* **2** (1950) 149.
32. H. J. Goldsmid: *Thermoelectric Refrigeration,* Plenum Press, New York 1964.
33. R. W. Ure: "Thermoelectric Effects in III – V Compounds," in R. K. Willardson, A. C. Beer (eds.), *Semiconductors and Semimetals,* vol. 8, Academic Press, New York – London 1972, pp. 67 – 102.
34. R. Wolfe: "Magnetothermoelectricity," *Sci. Am.* **210** (1964) no. 6, 70 – 82.
35. P. W. Bridgman, *Proc. Natl. Acad. Sci. USA* **11** (1925) 608.
36. P. W. Bridgman, *Proc. Am. Acad. Arts Sci.* **61** (1925) 101.
37. P. W. Bridgman, *Phys. Rev.* **39** (1932) 702.
38. P. W. Bridgman, *Phys. Rev.* **42** (1932) 858.
39. P. E. A. Lenard, *Wiedem. Ann. Phys. (Leipzig)* **39** (1890) 28.
40. W. Thomson, *Math. Phys. Papers* **2** (1882) 192.
41. W. Thomson, *Philos. Trans. R. Soc. London* **146** (1856) 663.
42. W. Thomson, *Math. Phys. Papers* **2** (1883) 207.
43. W. Thomson, *Math. Phys. Papers* **1** (1882) 232, 266.
44. E. Justi: *Leitfähigkeit und Leitungsmechanismen fester Stoffe,* Vandenhoek & Ruprecht, Göttingen 1948.
45. C. A. F. Benedicks, *Ann. Phys. (Leipzig)* **55** (1918) no. 1, 103.
46. C. A. F. Benedicks, *Ann. Phys. (Leipzig)* **62** (1920) 185.

47. C. A. F. Benedicks, *Math. Astron. Phys.* **23A** (1933) no. 27, 1.
48. C. A. F. Benedicks, *Math. Astron. Phys.* **24A** (1933) no. 1, 7.
49. J. W. Christian et al.: "Thermoelectricity at Low Temperatures: VI. A Redetermination of the Absolute Scale of Thermo-electric Power of Lead," *Proc. R. Soc. London A* **245** (1958) 213 – 221.

Thermography → Nondistructive Testing

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Thermoplastic Elastomers

Standard Article

Traz Ouhadi[1], Sabet Abdou-Sabet[2], Hans-Georg Wussow[3], Larry M. Ryan[4], Lawrence Plummer[5], Franz Erich Baumann[6], Jörg Lohmar[6], Hans F. Vermeire[7]

[1]Advanced Elastomer Systems, Brussels, Belgium
[2]Advanced Elastomer Systems, Akron, Ohio 44334, United States
[3]Bayer AG, Dormagen, Federal Republic of Germany
[4]E. I. Du Pont de Nemours and Co., Inc., Engineering Polymers, Wilmington, Delaware 19880, United States
[5]E. I. Du Pont de Nemours and Co., Inc., Engineering Polymers, Wilmington, Delaware 19880, United States
[6]Degussa AG - High Performance Polymers, Marl, Germany
[7]Shell Research SA, Louvain-La-Neuve, Belgium

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a26_633.pub3
Article Online Posting Date: January 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Thermoplastic Polyolefin Elastomers
1.1. Introduction and Definition
1.2. Thermoplastic Polyolefin Blends
1.2.1. Morphology
1.2.2. Elastomer Component, Soft Domain
1.2.3. Plastic Component, Hard Domain
1.2.4. Compounds, Trade Names
1.2.5. Processing
1.3. Thermoplastic Vulcanizates
1.3.1. Dynamic Vulcanization
1.3.2. Production and Morphology
1.3.3. Types
1.3.4. Processing
1.3.5. Performance and Product Positioning
1.3.6. Uses
1.3.7. Commercial Products and Trade Names
2. Thermoplastic Polyurethane Elastomers
2.1. Introduction
2.2. Raw Materials
2.3. Production
2.4. Properties
2.5. Processing and Use
2.6. Economic Aspects
2.7. Toxicology and Occupational Health
3. Thermoplastic Copolyester Elastomers
3.1. Introduction
3.2. Raw Materials
3.3. Production
3.4. Microstructure and Composition of Segments
3.4.1. Short-Chain Ester Units (Hard Segments)

3.4.2. Long-Chain Polyether Soft Segments
3.4.3. Hydrocarbon Soft Segments
3.4.4. Polyester Soft Segments
3.5. Properties
3.6. Blends
3.7. Uses
3.8. Producers, Trade Names
4. Thermoplastic Polyamide Elastomers
4.1. Introduction
4.2. Raw Materials
4.3. Manufacture of Block Polyetheramides
4.3.1. Polyetheresteramides
4.3.2. Synthesis of Polyetheramides
4.4. Physical and Chemical Properties
4.4.1. Block Variations
4.4.2. Morphology
4.4.3. Outstanding Properties
4.4.4. Chemical Resistance
4.4.5. Processing
4.5. Uses
4.6. Trade Names
5. Styrenic Block Copolymers
5.1. Introduction
5.2. Synthesis
5.3. Properties
5.4. Uses
5.4.1. Adhesives, Sealants, and Coatings
5.4.2. Modification of Bitumen
5.4.3. Compounded Products
5.4.4. Modification of Plastics

[Top of Page]

1. Thermoplastic Polyolefin Elastomers

Traz Ouhadi and Sabet Abdou-Sabet

1.1. Introduction and Definition

The significant growth of thermoplastic elastomers in the marketplace since the 1970s can be attributed to the family of thermoplastic polyolefin elastomers because of the wide range of products available, their favorable price – performance relationship, and their acceptance by the automotive industry. Historically, Uniroyal Chemical Company was the first company in 1972 to introduce a thermoplastic polyolefin elastomer to the market.

In 1971, W. K. FISHER [1], [2] filed patent applications on thermoplastic elastomer (TPE) blends based on rubbers that are copolymers of monolefinic monomers — such as ethylene – propene copolymer (EPM) or ethylene – propene – diene terpolymer (EPDM) (⟶ Rubber, 3. Synthetic) — with crystalline polyolefin resin, e.g., polypropylene (PP). FISHER employed partial vulcanization of the rubber component to maintain the thermoplastic processibility of the blend by limiting the amount of peroxide. Cross-linking was performed while mixing under dynamic shear, a process known as dynamic vulcanization. However, the dynamic vulcanization process and the first EPM – PP blend were discovered independently by GESSLER et al. [3] (Exxon) and HOLZER et al. [4] (Hercules) in 1958 and 1961, respectively. Two types of thermoplastic rubber (TPR) were produced. One type consisted of simple blends of EPDM and PP; the other contained *partially vulcanized rubber* in the blend. In 1972 – 1976, most major EPDM producers commercialized products based on simple blends. In general, the TPR products were elastomeric according to DIN 7724 and ISO 1382–1982 definitions, yet these products were not seriously considered for true rubber replacement applications.

Significant improvement in the properties of these blends was achieved in 1975 by CORAN et al. [5] by *fully vulcanizing the rubber phase* under dynamic shear without affecting the thermoplasticity of the blend. This discovery was advanced by ABDOU-SABET et al. [6] in 1977 by the use of phenolic curatives to achieve further improvement in rubber-like properties and processing characteristics.

Therefore, the field of thermoplastic elastomers based on polyolefin rubber – plastic blends grew along two distinctly different product lines or classes: one class consists of a simple blend of α-olefin rubber in a crystalline polyolefin resin and classically meets the definition of thermoplastic polyolefin (TPO). In the production of the other class, the rubber component in the TPO is subjected to dynamic vulcanization to generate a TPE with true rubber-like properties. This class is called thermoplastic vulcanizates (TPVs) or dynamically vulcanized alloys (DVAs).

These two classes of materials share a common basic composition. The plastic phase is a semicrystalline polyolefin such as isotactic and syndiotactic polypropylene, propene copolymer, high-density polyethylene, low-density polyethylene, linear low-density polyethylene, ethylene – vinyl acetate copolymer, ethylene – methacrylate copolymer, or ethylene – ethyl acrylate copolymer. The rubber phase may consist of EPM, EPDM, isobutene – isoprene copolymer (butyl rubber), halobutyl rubber, isobutene – *p*-methylstyrene copolymers, halogenated isobutene – *p*-methylstyrene copolymers, natural rubber, styrene – butadiene copolymers, nitrile – butadiene copolymers, styrene – isoprene – styrene block copolymers and their hydrogenated version, etc. (For a description of component polymers, see ⟶ Polyolefins; see ⟶ Rubber, 3. Synthetic).

1.2. Thermoplastic Polyolefin Blends

Thermoplastic polyolefin compositions based on the blends of a polyolefin rubber and a semicrystalline polyolefin have increased dramatically in popularity in recent years. The first thermoplastic polyolefins blends commonly referred to as TPOs were *mechanical blends* of polyolefin polymers. The most common types of TPO are composed of polypropylene and ethylene – propene rubber (EPR). In principle there are a wide variety of starting materials, an almost unlimited number of formulations for this type of TPOs in major part custom-tailored to meet the varied requirements of the most demanding low temperature high impact applications.

In addition to these mechanical blends, advances in Ziegler – Natta catalysis and reactor design have allowed the preparation of such blends in the polymerization reactor from ethylene and propene monomers. These are commonly called *reactor thermoplastic polyolefins* (RTPO). The amount of EPR content in RTPO can be varied from less than 15 wt %, which is most common, to around 70 wt %. Commercial demand for TPO and RTPO has been particularly strong in the automobile bumper application, especially in Europe and Japan. Such dominance has overshadowed the opportunities for TPO in other market sectors. Advanced compounding technology is being used to generate TPO compounds specific to the application; such technology can also be applied to RTPO.

Mechanically prepared blends are produced by using either batch or continuous mixing techniques. The most common batch mixers are internal intensive mixers, such as the Banbury. Continuous mixing techniques can be performed in either single- or twin-screw extruders. Mixing procedures can be extremely varied because of the wide range of TPO recipes in use, as well as the mechanical characteristics of individual process lines.

Reactor thermoplastic polyolefins are the newest variety of TPO. The amount of rubber (generally EPR) that may be included in a reactor-generated TPO is strongly dependent on the catalyst and process used. The development of a new generation of Ziegler – Natta catalysts (4th generation) with long-lived activity and the capacity of providing PP particles with very high porosity permit the polymerization of a high-rubber component (up to 70 wt %) (soft component) inside a shell or skin of crystalline PP. This has been accomplished by the *Cataloy process* [7], which is based on a three-stage reactor. The first stage consists of a tubular loop reactor that produces semicrystalline polypropylene. In the second and third reactors, ethylene and other olefinic monomers are introduced, and ethylene copolymerizes with propene to form the ethylene – propene copolymer rubber inside the semicrystalline homopolymer PP produced in the first reactor; this minimizes reactor fouling. The formation of a continually growing polymer skin and porous particle provides a reaction bed within which the copolymer is produced. This advanced technology in Ziegler – Natta catalysis has been referred to as *Reactor Granules Technology* [8] because the growing polymer particle itself has become the polymerization reactor.

1.2.1. *Morphology*

Thermoplastic polyolefins possess primarily a two-phase morphology. At least one phase — the thermoplastic component of the blend — is continuous. The second, or rubber, phase can be either co- continuous or dispersed, depending on the composition of the blend. The mixing conditions, surface energy, viscosity, and molecular masses of the two polymers determine at what polymer ratio the two phases are co- continuous [9], [10], since at extreme concentrations (i.e., a 10 : 90 or 90 : 10 ratio) the minor component is expected to be the dispersed phase.

ONOGI and coworkers [11], [12] reported phase inversion in the blends of crystalline polypropylene with EPR or EPDM elastomers when the polypropylene content reaches 50 to 60 %.

LOHSE [13] describes the morphology of the simple blend as unstable because the polymers are not miscible, especially at a 50 : 50 polymer ratio. The rubber phase, particularly when present as the dispersed phase, tends to agglomerate or coalesce. Such phase growth is undesirable and would lead to changes in properties and performance as a function of mixing and processing. To stabilize against phase growth, additives or *compatibilizers* are usually added. The preferred compatibilizer is a graft or block copolymer of isotactic PP and EPR. One approach has been described in [14]. The compatibilizer is generated from EPDM containing vinylnorbornene as the diene; the pendant double bond is then used to graft to the PP. The addition of 3 % of a compatibilizer to the blend improves adhesion between the phases and leads to finer dispersion of the phases. This is demonstrated by the smaller size of the rubber phase and the considerably decreased phase growth (Fig. 1).

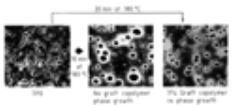

Figure 1. Morphology of EPDM – PP thermoplastic polyolefins containing 0 and 3 mol % compatibilizer

1.2.2. *Elastomer Component, Soft Domain*

Either ethylene – propene or ethylene – propene – diene rubber is suitable for these blends. The double bonds in the EPDM are not used as chemically active sites for curing. However, they provide considerably wider choices in raw material selection and undoubtedly affect the green strength properties of the elastomer by changing the number of branch points in the rubber molecule. The diene most commonly used in the EPDM is ethylidene norbornene (ENB), but 1-4 hexadiene

(1,4 HD), dicyclopentadiene (DCPD), or vinylnorbornene (VNB) may be used.

EPR rubbers contain 45 – 85 mol % ethylene, 55 – 15 mol % propene, and 0 – 10 mol % diene. EPDM with an ethylene content > 75 mol % exhibits crystallinity that can be examined by differential scanning calorimetry or ^{13}C NMR by measuring the ethylene sequence index [15]. The net effect of ethylene crystallinity is that it greatly strengthens the rubber and, consequently, the blend.

1.2.3. *Plastic Component, Hard Domain*

In most commercial TPO, the hard domain is isotactic polypropylene homopolymer or a copolymer (random or block) with a small amount of ethylene as comonomer for random PP and until 25 mol % ethylene for block copolymer PP. These resins are semicrystalline. The homopolymer has an *mp* of 155 – 165 °C, while the random copolymer has an *mp* of 140 – 155 °C. Ethylene decreases the rigidity of the polymer and improves the low-temperature properties at a sacrifice of the upper performance temperature.

1.2.4. *Compounds, Trade Names*

The two main components of a TPO blend are ethylene – propene rubber and polypropylene. Additional materials, such as other elastomers, processing oils, antioxidants, reinforcing and nonreinforcing fillers, colorants, and other additives to enhance the adhesion or the paintability of the product, are often used. Typical TPO compounds and their properties are listed in Table 1, and the suppliers and trade names of such materials are listed in Table 2.

Table 1. Typical TPO compositions and properties

Components, properties	TPO type		
	1	2	3
EPDM, parts by weight	20	57	85
PP, parts by weight	80	43	15
$CaCO_3$, parts by weight	60	8	8
Irganox 1010, parts by weight	0.2	0.2	0.2
Tinuvin 327, parts by weight		0.15	0.15
N-110 black, parts by weight	3.0		
Calcium stearate, parts by weight	0.10		0.10
Process oil, parts by weight		7.0	64.00
T.S.,* MPa	16.55	13.50	5.50
U.E., ** %	200	740	750
Flexural modulus B, MPa	1240	415	
Shore hardness	62 D	54 D	65 A
Gardner impact (at –29 °C), J	20	36+	

* T.S. = Tensile strength.
** U.E. = Ultimate elongation.

Table 2. Suppliers of TPO

Trade Name	Producer	Location
Vistaflex	AES	Europe
Dexflex, Sequel, Depro, Ontex, Fortilene	Solvay	North America, Europe
Hifax, Adflex	Basell	North America, Europe
Flexathene	Equistar	North America
Kelburon	DSM	Europe
Keltan TP	DSM	Europe
Polytrope	A. Schulman	North America
Telcar	Teknor Apex	North America
Tefabloc	Cousin Tessier	Europe
Multiflex	Multibase	Europe
Vestolen EM	Pawtucket, RI, United States	North America
Vestoprene	Chem. Werke Hüls	Europe
Exxtral	ExxonMobil	North America
Plastomer	Yukong	Asia
Garaflex O	Alpha Gary	North America

Dong Il	Deayoung	Asia
Zeonex, Zeonor TPO	Zeon	Asia
Silac	Lucent	North America
WPP TPO	Washington Penn	North America
Vyflex	Lavergne	North America
Novalast	Nova Polymers Inc.	North America
Elastamax	PolyOne	North America
Salflex	ABC Group	North America
J-Prene	J-Von	North America
Tamfer	Mitsui	Asia
Thermoran	Mitsubishi/JSR	Asia
P.E.R.	Tokoyama Soda	Asia

Automotive and electrical applications are the major markets for TPO. Automotive constitutes by far the largest market because of the excellent weatherability and low cost of TPOs. Some examples are bumper covers, interior trim, fender extensions, etc.

TPOs have good electrical properties in combination with good ozone resistance, which makes these materials a good choice for low voltage wire and cable applications where hot tension set at 200 °C is not a requirement. Such applications can be found in automotive booster cable, flexible cord, and welding cable insulation.

1.2.5. *Processing*

TPOs can be processed on most common thermoplastic equipment. Products can be formed by injection molding, extrusion, vacuum forming, injection blow molding, extrusion blow molding and so on. The versatility of the compounding TPO products permits designing of compounds to fit the available process and equipment.

1.3. Thermoplastic Vulcanizates

Thermoplastic vulcanizates made via a dynamic vulcanization (TPV) are rubbery engineered polymers blends with fabrication characteristics of a conventional thermoplastic and perform the properties of a conventional thermoset rubber.

TPVs offer a variety of practical advantages over conventional thermoset rubbers. These advantages are:

1. Little or no compounding or blending. Most TPVs are fully formulated and ready for use.
2. Simpler processing with a fewer steps. TPVs have the processing simplicity of thermoplastic.
3. Shorter fabrication cycle time and consequently higher productivity for fabrication parts.
4. Recycling of scrap material (regrind). After each step, the scrap from thermoset rubber should be discarded while with TPVs all the eventual scrap is reground and used as virgin material.
5. Lower energy consumption
6. Better quality control and closer tolerances on finished articles
7. In most cases a density lower than a comparable thermoset rubber.

1.3.1. *Dynamic Vulcanization*

Dynamic vulcanization was discovered by GESSLER et al. [3]. In an attempt to improve the impact properties of PP they partially vulcanized halobutyl rubber in a polypropylene matrix under continuous mixing. Dynamic curing generates the same necessary cross-links or three-dimensional polymer structures as static curing. In dynamic curing, however, these structures are generated in small rubber particles dispersed in the uncross-linked thermoplastic polymer matrix as a microgel. The formation of such dispersed-phase morphology depends on establishing a high degree of cure (CORAN et al. [5]). The rubber phase is fully vulcanized in a plastic matrix (thermoplastic vulcanizate). The rubber particles are in the order of 1 – 2 µm in size. Fully curing the rubber phase has led to significant improvement in the material properties including

1. Higher ultimate tensile strength
2. Lower tension and compression set
3. Better property retention at elevated temperature
4. Greater resistance to attack and swelling by fluids
5. Greater flexural fatigue resistance
6. More consistent processibility

Partially cross-linked blends give modest improvement [1], [2].

The effect of the size of the rubber particles and the degree of cure on material properties can be seen in Figures 2 and 3. By proper choice of the curatives, S. ABDOU-SABET et al. [6] demonstrated that rubber-like properties can be improved beyond those achieved by CORAN. Improvements in compression set, oil resistance, and the processing characteristics of the material were achieved by applying phenolic curatives (e.g., dimethyloloctylphenol). These improvements are beyond those expected from the state of cure. They probably are due to in situ formation of a graft copolymer between EPDM and PP that optimizes the interfacial adhesion between the rubber particle and the polypropylene matrix, thus allowing for better elastomeric properties. The graft copolymer could have been generated through the functionalization of PP with dimethyloloctylphenol, which then reacts with EPDM [16].

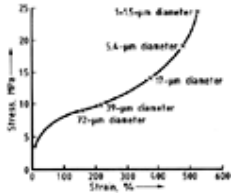

Figure 2. Effect of particle size on stress – strain relationship

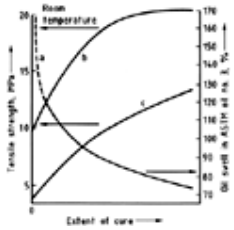

Figure 3. Effect of the degree of cure on mechanical properties

a) Percentage of swell in ASTM oil no. 3; b) Tensile strength at 100 °C; c) Tensile strength at room temperature

More importantly, significant improvement in processing was achieved over that obtained with the sulfur cure, as shown in Table 3. The use of dimethyloloctylphenol as curative allowed the generation of very soft TPV (as soft as 35 Shore A, with compression set approaching that of a thermoset rubber) while maintaining excellent thermoplastic processing. The poor processing characteristics of sulfur- cured TPV are due to phase growth of the dispersed rubber particles. The polysulfidic bonds generated during vulcanization undergo sulfur exchange reactions leading to significant coalescing of the rubber particles.

Table 3. Effect of cross-linking agent on properties of EPDM rubber – PP TPV

	Cross-linking agent			
Property	**None**	**Sulfur**	**Dimethylolalkylphenol**	**Peroxide**
Hardness Shore D	36	43	44	39
T.S.,[a] MPa	4.96	24.3	25.6	15.9
M_{100},[b] MPa	4.83	8.00	9.72	8.07
U.E.,[c] %	190	530	350	450
ASTM oil no. 3 swell, %	disintegrated	194	109	225
Comp. set,[d] % (22 h at 100 °C)	91	43	24	32
Processing characteristics[e]	+	–	+++	+

[a] T.S. = tensile strength.
[b] M_{100} = modulus at 100 % elongation.
[c] U.E. = ultimate elongation.
[d] Comp. set = compression set.
[e] + = good; – = poor; +++ = excellent.

Other vulcanization systems are used to cure the rubber phase of TPV. Peroxide with an adequate coagent is the most common curing system after phenolic curatives. The role of coagent is to enhance the cross-linking efficiency for the rubber phase. Apart from cross-linking efficiency, the use of a coagent can also affect the molecular structure of the cross-links. The most common type of coagents are polyfunctional organic molecules.

Another effective vulcanization system reported in several patents [17], [18] uses a hydrosilylation agent with platinum containing catalysts. This cross-linking system works with rubber containing diene groups, preferably with sterically unhindered carbon-carbon double bond.

A new cross-linking system to make TPV based on hydrolyzable organosilane grafted rubber molecules is reported in [19-21]. The organosilane function is hydrolyzed during the dynamic mixing of rubber with the thermoplastic phase (PP) followed by condensation reaction with help of an organotin derivative catalyst such as dibutyltin dilaurate.

1.3.2. *Production and Morphology*

Melt mixing of the two polymers prior to dynamic vulcanization is the first requirement for a good TPV. The phase morphology of the uncross-linked blend will be influenced by the same variables discussed for TPO. S. ABDOU-SABET et al. [10] demonstrated for an 80 : 20 EPDM – PP blend that PP is the dispersed phase in an EPDM matrix. During dynamic vulcanization of such a blend, EPDM and PP have to undergo a phase inversion to maintain the thermoplasticity of the blend. In the initial stages of dynamic curing, two co- continuous phases are generated, and as the degree of cross-linking

increases during mixing, the continuous rubber phase becomes increasingly elongated and then breaks up into polymer droplets. As these droplets form, PP becomes the continuous phase. This can be seen in the SEM electron micrograph (Fig. 4), in which the EPDM phase appears white and the PP phase is black. Photograph 4 B shows that both phases are co- continuous. Figure 5. shows a TEM electron micrograph of a very soft TPV where EPDM phase appears in black and PP in white color.

Figure 4. SEM-Electron micrograph of an 80 : 20 EPDM – PP thermoplastic vulcanizate

A) Unvulcanized blend; B) Partially vulcanized; C) Fully vulcanized; D) Fully vulcanized and diluted to 20 : 80 EPDM – PP

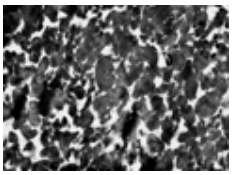

Figure 5. TEM-Electron micrograph of a very soft TPV based on EPDM/PP

Reactor TPO can also be used as a raw material to make a TPV [22]. Reactor TPO, however, is not yet commercially produced with a diene function or other functionality in the rubber molecules. Therefore, cross-linking of the rubber phase can be performed only with peroxide. Peroxide curing is carried out while the necessary precautions are taken to protect polypropylene from chain scission and unzipping by the peroxy radical. Protection can be achieved via the incorporation of compounds in the blend that preferentially undergo reaction with the peroxides without affecting PP [23]. Such compounds include polyisobutylene, polybutadiene, and polybutene. Even then, achieving full cure is quite difficult, and only partial curing of the rubber phase is usually accomplished.

To achieve the best improvement in properties of TPV, certain conditions must be satisfied:

1. The higher the crystallinity of the plastic component, the better are the mechanical and elastomeric properties
2. The smaller the difference in critical surface tension between the rubber and the plastic, the smaller is the size of the dispersed rubber phase
3. The lower the entanglement spacing of the rubber chain, the better are the properties of TPV

1.3.3. *Types*

Theoretically it should be possible to make a large number of different classes of TPV by combination of at least one type of rubber with one type of thermoplastic. But practically, there is a limitation in the number of commercial TPV families due to limited number of compatible rubber-thermoplastic pairs. Obviously by use of adequate compatibilizing agent between a given rubber and a given thermoplastic, it is possible to expand the number of TPV families.

The discussion in this section is limited to those containing at least one polyolefin polymer.

EP/EPDM rubber -polypropylene based TPVs are the most common types of TPV [24]. The EP/EPDM-PP pair is the most used rubber (except for tire application) and thermoplastic material due to the excellent rubbery properties, heat and weathering resistance of EP/EPDM rubber and the excellent versatility, chemical resistance and inertness, and mechanical properties of PP. A broad range of vulcanization systems can be used as mentioned earlier to cross-link EPDM while only peroxide system can be used as cross-linking agent for EP rubber.

Natural rubber (NR) - polyolefin based TPVs [25] have been thoroughly investigated in an attempt to develop commercial products that take advantage of the attributes of natural rubber. Most of the research work was focused on *polypropylene-based material*. These NR – PP TPV thermoplastic natural rubbers (TPNRs) exhibited significant disadvantages compared to their EPDM counterparts and were not acceptable. Drawbacks stemmed from the poor thermal stability and the generation of excessive odor from the natural rubber component on processing at higher temperature. Again, partially cross-linked products exhibited co- continuous morphology, while those cross-linked to 95 % or higher of the natural rubber phase showed dispersed-phase morphology. Properties of this type of TPV are listed in Table 4.

Table 4. Typical properties of thermoplastic natural rubbers

Property	Rubber type				
	1	2	3	4	5
Shore A hardness	50	60	70	80	90
	3.1	4.1	5.3	6.3	8.2

M_{100},[a] MPa					
T.S.,[b] MPa	6.5	8.8	11.2	13.2	15.3
U.E. at break,[c] %	285	315	340	370	405
Tear strength, die C (kN/m)	22	31	38	43	51
Tension set, %	9	10	14	18	23
Comp. set,[d], %					
72 h, 23 °C	26	28	31	34	39
24 h, 70 °C	30	35	36	41	55
22 h, 100 °C	38	42	43	47	57
Volume swell after 7 d in					
ASTM oil no. 1, 23 °C	14	9	9	7	
ASTM oil no. 2, 23 °C	19	13	12	9	
ASTM oil no. 3, 23 °C	71	53	47	35	
ASTM oil no. 1, 100 °C	101	80	67	61	
ASTM oil no. 2, 100 °C	151	123	108	82	
ASTM oil no. 3, 100 °C	190	164	139	116	

[a] M_{100} = modulus at 100 % elongation.
[b] T.S. = tensile strength.
[c] U.E. = ultimate elongation.
[d] Comp. set = compression set.

Natural rubber – polyethylene-based compositions [26] have also been studied extensively. These TPVs should possess good low-temperature flexibility and elastic recovery for low-temperature applications.

Butyl rubber-polypropylene based TPVs are another family of TPVs where a butyl or halobutyl rubber is used as elastomer componen. In 1988, R. C. PUYDAK et al. [27] introduced *butyl rubber-based TPV* for applications requiring impermeability to oxygen and water vapor transmission. Such products are useful in sports bladders and blood vial stoppers.

Polymers with significantly different critical surface tension [28] can still be mixed together with the help of a compatibilizer to improve wetting characteristics or interfacial adhesion [29] between the two phases, thus allowing for the generation of preferred morphology prior to dynamic vulcanization. The application of compatibilization technology allowed the mixing of polar rubbers (e.g., *nitrile rubber* [30], [31] and *acrylate rubber* [32]) with polypropylene, followed by dynamic vulcanization to improve the properties as much as possible. Figure 6 illustrates the effect of the compatibilizer (maleated polypropylene) on polypropylene – ethylene – acrylate rubber TPV.

Figure 6. Effect of compatibilizer (maleated polypropylene) on ultimate tensile strength of polypropylene – ethylene – acrylate rubber TPV

Only a small amount of compatibilizer is needed for compatibilization. The function of the compatibilizer is analogous to that of a surfactant that emulsifies and stabilizes a water – oil mixture. Effective compatibilizers are generated by forming a graft or block copolymer; the structures of the blocks must be similar to the phases to be compatibilized. In that case, one block of the copolymer is soluble in the polyolefin resin, while the other is soluble in the polar rubber phase. Graft formation must be controlled to achieve the desired interfacial adhesion while preserving a dispersed phase morphology. Generation of primary chemical bonds between the two phases could lead to an interpenetrating network morphology. Such networks could result in thermoset characteristics and the loss of the thermoplastic processing advantage.

1.3.4. *Processing*

Thermoplastic vulcanizates or block copolymers that are made of soft and hard blocks are processed or fabricated by the same techniques normally used for rigid plastics. This enables the fabrication of elastomeric articles with improved efficiency and economy over those of a conventional thermoset rubber. Scrap from TPV can be recycled readily, whereas thermoset rubber scrap is generally discarded, resulting in a sizeable cost penalty. The thermoplastic nature of TPV enables fabrication by methods unavailable to thermoset rubbers. These include blow molding, thermoforming, and heat welding. Thermoplastic vulcanizates have extremely non-Newtonian flow properties, and their viscosity is highly sensitive to shear [33]. At low shear rates, they flow very poorly; at high shear rates, however, they are quite fluid, enabling rapid injection molding cycles.

1.3.5. *Performance and Product Positioning*

ASTM D-2000 attempts to classify rubber products according to their performances in dry heat (dry heat service temperature) and ASTM oil no. 3 (see also ⟶ Rubber, 3. Synthetic). Conventional thermoset rubber occupies certain regions on a performance chart. Such classification is, however, insufficient because other characteristics such as

dynamic properties, stress relaxation, barrier properties, UV resistance, adhesion, and colorability, also play an important role in material selection. Nevertheless, according to the ASTM D-2000 classification, EPDM – PP TPVs have performance characteristics superior to chloroprene (Neoprene) rubber. The properties of TPO, however, are considerably less elastomeric, and TPOs occupy a lower-performance category (Fig. 7).

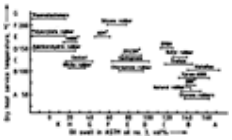

Figure 7. Classification of elastomers according to heat and oil performance

*AEM = ethyl acrylate – ethylene copolymer; CM = chlorinated polyethylene; CSM = chlorosulfonated polyethylene; HNBR = hydrogenated acrylonitrile – butadiene rubber; SBR = styrene – butadiene rubber; Geolast, Santoprene, Trefsin, Vistaflex, and Viram-6000 are registered trade names of commercial TPVs (see Table 5)

1.3.6. *Uses*

Since their introduction in 1981, TPVs have found thousands of commercial uses worldwide. Generally speaking, these materials have been used in most rubber applications with the exception of tires. Thermoplastic vulcanizates are employed, e.g., in automotive and construction applications, in foods, hardware, appliances, business machines, mechanical rubber goods, and sporting goods, as well as electrical and medical applications. The uses can vary from the syringe stoppers in medical syringes to the rack-and-pinion boots on the steering mechanism of a front-wheel-drive automobile.

1.3.7. *Commercial Products and Trade Names*

Advanced Elastomer Systems who commercialized the first TPV (Santoprene) in 1981, is leading the market for this family of products. Global market demand for TPVS is forecast to increase 8 to 10 % per year. The strong gain anticipated for TPVs results from an intense product development activity led by Advanced Elastomer Systems. Currently, a number of products are on the market (see Table 5). Only fully vulcanized materials achieve the full benefit of dynamic vulcanization.

Table 5. Suppliers of thermoplastic vulcanizates

Trade name	Supplier	Polymer type	Degree of cure
Santoprene [a]	AES	EPDM – PP	full
Trefsin [b]	AES	butyl-PP	full
Dytron [c]	AES	EPDM – PE	full
Sarlink 3000	DSM	EPDM – PP	partial
Sarlink 2000	DSM	butyl – PP	partial
Hifax-XL	Himont	EPR – PP	partial
Vyram [d] 9000	AES	EPDM – PP	partial
Vyram [d] 6000	AES	natural rubber – PP	full
Milastomer	Mitsui	EPDM – PP	partial
Sumitomo TPE	Sumitomo	EPDM – PP	partial

[a] Santoprene is a registered trademark of Monsanto Company, licensed to Advanced Elastomer Systems, L.P.
[b] Trefsin is a registered trademark of Advanced Elastomer Systems, L.P.
[c] Dytron is a registered trademark of Advanced Elastomer Systems, L.P.
[d] Vyram is a registered trademark of Advanced Elastomer Systems, L.P.

[Top of Page]

2. Thermoplastic Polyurethane Elastomers (Polyurethane Rubbers, see ⟶ Rubber, 3. Synthetic – Polyurethane Rubber ; see also ⟶ Polyurethanes)

Hans-Georg Wussow

2.1. Introduction

Thermoplastic polyurethanes (TPUs) are a group of elastic materials whose common structural feature is the urethane group –O–CO–NH– arising from the reaction between diisocyanates and long- and short- chain diols. The long- chain polyester or polyether diols form flexible soft segments in the temperature range of use, while the short- chain diols and diisocyanates form rigid or hard segments. Since soft and hard segments are arranged alternately along the molecular chain, TPUs belong to the $[AB]_n$-type block copolymers. Intermolecular interactions between the hard segments lead to the formation of glassy or paracrystalline hard domains, which, as physical cross-linking points, can be reversibly melted during thermoplastic processing. Variation in the ratios of soft and hard segments and monomers permits the production of materials in the modulus range of ca. 5 – 700 MPa, corresponding to ca. 60 Shore A to 70 Shore D. Because of their very good mechanical properties, especially extraordinary resistance to abrasion, the great choice of monomers available for the production of “tailor-made materials”, and efficient thermoplastic processibility, TPUs have found a wide range of applications since their introduction to the market at the beginning of the 1960s. Further information on the production and properties of TPUs can be found in [34-37].

2.2. Raw Materials (⟶ Isocyanates, Organic; ⟶ Polyesters; ⟶ Polyoxyalkylenes)

The raw materials used in TPUs are diisocyanates, short- chain diols with molecular masses of 61 – ca. 600, and long- chain polyester and polyether diols with average molecular masses between 600 and 4000, so- called polyols.

Diisocyanates. 4,4′-Diisocyanatodiphenylmethane (4,4′-methylenediphenyl diisocyanate, MDI) [101-68-8] is industrially by far the most important diisocyanate. It can be used to produce materials that offer the most favorable compromise between price, thermoplastic processibility, and mechanical properties. Other *aromatic diisocyanates* used are 3,3′-dimethyl-4,4′-biphenyl diisocyanate (3,3′-tolidene-4,4′-diisocyanate, TODI) [38] and 1,4-phenylene-diisocyanate (PPDI) [104-49-4] [39]. Among the *aliphatic diisocyanates* the following are important: 4,4′-dicyclohexylmethane diisocyanate (H_{12}-MDI) [5124-30-1] (as the isomeric mixture) [40] and 1,6-diisocyanatohexane (hexamethylene diisocyanate, HDI) [822-06-0]. The aliphatic diisocyanates are much less reactive toward alcohols than MDI [41].

Short-Chain Diols. The most important short- chain diol used as a chain extender is 1,4-butanediol. For TPUs with improved dimensional stability at elevated temperature, 1,4-bis(2-hydroxyethoxy)benzene [hydroquinone bis(2-hydroxyethyl) ether] [104-38-1] is also used. The use of diol mixtures (e.g., with ethanediol or 1,6-hexanediol) gives products with a broader processing range [42].

Polyols.
Polyester Polyols. The most important polyols are polyadipates of short- chain diols with two to ten carbon atoms. Butanediol, hexanediol, ethanediol, and mixtures of these diols are most commonly used. The resistance to hydrolysis of the polyol and TPU increases as the molecular mass of the diol increases. Polycaprolactones and polycarbonate diols with their good to excellent resistance to hydrolysis are also employed. The most widely used products have average molecular masses of 1000 – 2000. As polycondensation products, polyadipates have a broad molecular mass distribution [43]. They frequently contain unreactive ring esters formed as byproducts. For TPU production the acid number of the ester polyols should be as low as possible, preferably < 1 mg KOH/g, because carboxyl groups react with isocyanates and impair the resistance to hydrolysis.

Polyether Polyols (⟶ Polyoxyalkylenes). Poly(tetramethylene oxides) are the most important polyether polyols for the production of TPUs. Poly(propylene oxides) [poly(propylene glycols)] and copolymers with ethylene oxide give products with poorer properties; normally, they are used together with polyester polyols. The molecular mass distribution of the polyether polyols follows the Poisson distribution and is thus considerably narrower than that of the corresponding polyesters. Poly(propylene oxides) have predominantly secondary hydroxyl groups as terminal units and are therefore considerably less reactive than other polyols.

Additives. Internal release agents, such as esters of montanic acid (octacosanoic acid), silicones, or amide waxes in quantities up to 2 %, are used as *processing aids*. TPUs based on polyesters can be protected effectively against *hydrolytic degradation* by the addition of carbodiimides [44]. Protection against *microbial degradation,* for example on prolonged contact with soil, is also possible. Sterically hindered phenols and amines are used as *antioxidants,* mainly to protect against discoloration and degradation, particularly in the case of polyether TPUs at the high production and processing temperatures often required. Yellowing because of the effect of light can be slowed considerably through the use of *light stabilizers* such as benzophenones sterically hindered amines (HALS) or benzotriazoles [45].

Plasticizers are used only to achieve hardness levels below ca. 70 Shore A [46-48]. Short glass fibers are used mainly to *reinforce* TPUs. Many additives, such as colorants, antiblocking agents, antistatic agents, and flame retardants, are supplied as concentrates.

2.3. Production

General. The industrial production of TPUs is carried out virtually exclusively in continuous, solvent-free processes [49]. According to Flory statistics [43] the required high molecular masses are achieved only at high conversions and an essentially stoichiometric ratio of NCO and OH groups. At an NCO : OH ratio < ca. 0.97, products with the required properties are generally not obtained [50]. Too high an excess of NCO can lead to cross-linking through allophanate and biuret formation even after long periods [51]. A prerequisite for maintaining exact stoichiometric ratios is the use of raw materials that are as pure — especially anhydrous — as possible and of exact metering devices. Prepolymer and one-shot processes can be used to carry out the reaction (⟶ Polyurethanes, ⟶ Polyurethanes – Reactive One-Pack Systems). In the one-shot process, OH- and NCO-functional reactants are mixed with each other in one step. In the prepolymer process, a prepolymer with terminal NCO groups is first formed from a long- chain polyol and diisocyanate. This prepolymer is then mixed with the chain extender. Reactivity differences (e.g., between polyether polyols and butanediol) can be overcome with this process. The process used affects the segment length distribution and thus the morphology of the final product [52], [53].

Belt Process [54]. Liquefied raw materials are mixed continuously in a mixing head and poured onto a continuous belt, which passes through an oven heated to ca. 70 – 130 °C, where the reaction proceeds further. The congealed product is ground to irregular pellets. It can be melted in extruders, possibly along with additives, and formed into regular pellets. According to a modification of this process [55] the reaction is carried out to a desired degree of conversion, generally > 80 %, on the belt, and the congealed mixture is then fed immediately into an extruder where the reaction is continued and the mixture is homogenized. Finally, the polymer is pelletized. In the belt process the temperature and residence times in the heated zone can be adapted to the requirements of the particular product.

Reaction Extruder Process. A twin-screw extruder serves as the reactor [56-59]. The screws are conrotating and can be set to the required shearing ratios by a combination of kneading and transporting parts. Temperature is controlled by using barrel sections that can be heated separately. The reactants and additives can be charged at different locations. In this way a continuous prepolymer or one-shot process is possible without additional equipment. The bulk temperatures are generally > 200 °C. The product is extruded continuously and pelletized immediately after the head or after passage through a cooling area. As in the belt process, subsequent sieving and drying processes occur. In the reaction extruder process, finished pellets can be obtained from starting materials with the addition of all additives (including glass fiber or

other thermoplastics) in one process step [60].

Comparison of Products. Because of different shearing and temperature conditions, the type of process used significantly affects the morphology and thus the processing properties of the products. Starting from identical raw materials more opaque products are formed in the belt process with larger hard segment domains that melt at higher temperature, whereas the reaction extruder process gives more transparent, lower-melting products.

2.4. Properties

Morphology [37], [61], [62]. The physical properties of TPUs are determined by their polyphase structure consisting of soft segments with very low glass transition temperatures, hard segments with high melting ranges, and mixed phases. This structure arises from the incompatibility of the hard and soft segments, which, on cooling from the melt during production and processing, leads to partial or complete separation of the phases that are bonded covalently with each other. The extent of separation and the phase structure are determined mainly by the stoichiometric composition of the polymer; the statistical distribution of the hard and soft segments [52]; the structure of the polyol [63], the diisocyanate and the chain extender, and the thermal history. The latter results from the kinetics of the separation and crystallization processes, and from the cleavage and transurethanization during production, processing, and possible tempering of the product [64-67]. In polyethers, phase segregation is more pronounced than in polyesters and increases as the molecular mass of the polyol increases. Many investigations of the structure [65], [68], [69] phase interactions [70], [71], [69] and structure – property relationships of TPUs have been carried out, mainly on model substances [72-74]. Macroscopically, the soft segments determine the elastic ductility, the low-temperature properties, and the basic chemical properties, while the hard segments are important for dimensional stability at high temperature and the processing properties.

Technical Properties. By varying the proportion of the hard segment the modulus of elasticity can be varied over a range of ca. 5 – 700 MPa without the use of plasticizers or reinforcing agents. Typical contents of the long chain diol range from 30–75 wt%. Products containing no [41] or only very small proportions of soft segments can no longer be regarded as thermoplastic elastomers [75]. Table 6 gives an overview of some properties of different types of product.

Table 6. Typical mechanical properties of TPUs with different hardness based on polyester polyols (example Desmopan, Bayer)

Property	Test standard (ISO)	Desmopan			
		385	345	359	372
Density, mg/m^3	1183	1.2	1.21	1.23	1.24
Hardness, Shore A/D	868	86 A	47 D	59 D	73 D
Tensile stress at break, MPa	527-1, -3	52	52	60	70
Elongation at break, %	527-1, -3	550	450	400	330
Tear resistance, kN/m	34-1	70	100	160	220
Flexural modulus, MPa	178		70	180	650
Abrasion loss, mm^3	4649	30	30	35	30
Compression set, %					
70 h/22 °C	815	30	25	30	
24 h/70 °C	815	55	42	60	

Figure 24 shows the dependence of tensile modulus on temperature for adipate-based TPU with different hardness; typical stress-strain diagrams are shown in Figure 25.

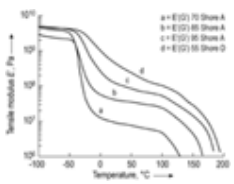

Figure 24. Tensile modulus of elasticity vs. temperature for polyadipate based TPU (example Desmopan, Bayer)

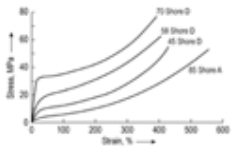

Figure 25. Stress-strain curves of polyadipate-based TPU with different hardness (example Desmopan, Bayer)

High extension with very low residual extension, high tensile and tear strength, high moduli of elasticity at a given Shore hardness, and excellent abrasion resistance are common to all TPUs. Polyester TPUs are superior to polyether products with regard to strength and resistance to oxidation or swelling by oils, fats, and water, while they are inferior with regard to

resistance to hydrolysis or microbial attack and impact resistance at low temperature. Hybrid types from polyester and polyether polyols combine the advantages of both types to a certain extent [76], [77].

Unstabilized TPUs based on MDI turn yellow through the action of light, initially without the impairment of mechanical properties. TPUs based on aliphatic diisocyanates do not show this discoloration. The lower temperatures of use are determined by the glass transition temperature of the soft phase (–20 to – 60 °C), and the upper ones by the onset of softening of the hard domains. Depending on the hardness, softening starts at ca. 80 – 100 °C (see ⟶ Polyurethanes). Hard products can be kept at up to ca. 120 °C for short periods. TPUs are soluble in polar, aprotic solvents such as *N,N*-dimethylformamide, methyl ethyl ketone, or *N*-methylpyrrolidone.

2.5. Processing and Use [34], [35], [78], [37], [81]

TPUs can be processed by all methods available for thermoplastics, such as injection molding, extrusion, blow molding, sintering, and calendering, although not all types are suited to every processing method. Melt temperatures range between 180 and 250°C, depending on hardness and other parameters. Typically, the melt viscosity of TPU is strongly temperature-sensitive ▪in which way? ▪. Generally only products with hardnesses up to ca. 50 Shore D can be extruded. Generally, extrusion becomes more and more difficult with increasing hardness above ca. 50 Shore D, due to formation of gel particles resulting from crystallization effects or side reactions.

TPUs are used in all those areas where their high strength and excellent resistance to abrasion, oil, and many other solvents are important. The large number of uses includes bases for shoe heels, soles and studs, or sports shoes, screens for the classification of minerals, rollers, tool parts, toothed belts, gaskets, electrical plugs, coupling elements, ear tags for animals, parts for car bodies, instrument panels, bearing shells, ski boots, seals, cable sheathings, tubing of all types, blown films, profiles, coatings, and medical equipment such as tubing and catheters. TPUs are also used for melt spinning of elastic fibers and modification of the impact resistance of polyacetals and as blend components for polycarbonates and acrylonitrile – butadiene – styrene (ABS) polymers [79], [80].

2.6. Economic Aspects

TPUs are produced in Europe, the United States, and Japan by the three large producers BASF, Bayer, and Noveon (quantities of > 20 000 t/a) and a number of medium and small producers (Table 7).

Table 7. Important producers of TPU and trade names

Region	Producer	Trade Name
Europe	Bayer	Desmopan
	Elastogran (BASF)	Elastollan
	Noveon	Estane
	Huntsman	Avalon, Irogran
	COIM	Laripur
	Mercquinsa	Pearlthane
USA	Bayer	Texin, Desmopan
	Noveon	Estane
	Huntsman	Krystalflex
	Dow	Pellethane
	BASF	Elastollan
Asia	Dainichi Seika	Resamine
	Nippon Miractran	Miractran
	DIC Bayer Polymers	Pandex, Desmopan
	BASF	Elastollan
	Hosung	Neothane
	SK Chemicals	Skythane

Almost all producers offer materials of differing hardness based on different polyols within several product ranges. In 2000, the estimated market volume was ca. 150 000 t/a. The regional distribution has a peak in Europe (40%), followed by North America (30%) and Eastern Asia (27%). The chances of growth are estimated to be positive in view of the possibilities for replacing rubber and plasticized thermoplastics by TPUs and the versatility of this class of materials.

2.7. Toxicology and Occupational Health

Since the production process involves polyaddition in a melt, very small quantities of byproducts are formed. Correctly dimensioned and efficiently functioning air extraction allows the permitted workplace concentrations for diisocyanates to be adhered to with certainty by producers and processors. Generally neither aromatic amines nor diisocyanates can be extracted from properly produced TPU articles. If only the allowable raw materials are used, TPUs comply with health authority guidelines for contact with foods (FDA: CFR 175.300, 175.105, 177.1680; European Commission Directive 2002/72/EC). Because of their good compatibility with blood and tissue, some products also fulfill the requirements of the ISO standard 10993/1 [82], [83].

[Top of Page]

3. Thermoplastic Copolyester Elastomers

Larry M. Ryan and Lawrence Plummer

3.1. Introduction

Thermoplastic copolyester elastomers are segmented block copolymers, consisting of repeating, high-melting, rigid “hard” blocks and amorphous, flexible “soft” blocks, which have a very low glass transition temperature. These polymers exhibit a remarkable combination of properties as a result of microphase separation of the hard block (in the form of crystalline domains) from the matrix of the amorphous soft blocks. The crystalline domains impart excellent mechanical properties (tensile strength, toughness, resistance to creep, good retention of properties at elevated temperature, etc.) and outstanding chemical resistance to the polymer. Concurrently, the amorphous domains provide exceptional flexibility and fatigue resistance over a broad temperature range. By varying the ratio of hard to soft blocks, products ranging from relatively rigid plastics to very flexible elastomers can be produced.

Copolyester elastomers are typically produced by condensation of an aromatic dicarboxylic acid or ester with a low molecular mass aliphatic diol and a polyalkylene ether glycol (molecular mass generally between 1000 and 4000) [84-86]. Reaction of the aromatic dicarboxylic acid with the diol leads to the crystalline hard segment phase, while the soft segment is the product of the diacid (or diester) and the long- chain glycol. Other combinations are possible, however (see Sections Long-Chain Polyether Soft Segments, Hydrocarbon Soft Segments).

3.2. Raw Materials

Polyester elastomers are produced from aromatic dicarboxylic acids or diesters, polyoxyalkylenes, and an excess of short- chain diols [84]. The most common raw materials follow:

Dicarboxylic Acids. *Terephthalic Acid* [100-21-0]. For production and properties, see ⟶ Teratogenic Agents.

2,6-Naphthalenedicarboxylic Acid [1141-38-4]. For production, see ⟶ Carboxylic Acids, Aromatic.

Polyoxyalkylenes. *Poly*(*tetramethylene oxide*); *poly*(*propylene glycol*) [25322-69-4]; *poly*-(*ethylene glycol*) [25322-68-3]; for production and properties, see ⟶ Polyoxyalkylenes.

Short-Chain Diols. *1,4-Butanediol* [513-85-9]. For production and properties, see ⟶ Butanediols, Butenediol, and Butynediol – 1,4-Butanediol.

Ethylene glycol [107-21-1]. For production and properties, see ⟶ .

3.3. Production

(see also ⟶ Polyesters – General Principles, ⟶ Polyesters – Production of Poly(Ethylene Terephthalate) from Dimethyl Terephthalate, ⟶ Polyesters – Production of Poly(Ethylene Terephthalate) from Terephthalic Acid, ⟶ Polyesters – Production of Poly(Butylene Terephthalate))

Polyester elastomers are manufactured by a *multistage melt polymerization process*. Soluble titanium alkoxides are generally used as catalyst. In the first stage of the process, all reactants are mixed and heated at atmospheric pressure to 190 – 220 °C to form a low molecular mass prepolymer. In the latter stages the prepolymer is converted to high molecular mass polymer by increasing the temperature to 250 – 260 °C while lowering the pressure to < 133 Pa to distill off the excess short- chain diol. The degree of polymerization is monitored by the viscosity change of the molten reaction mass by measuring the torque on the reactor agitator. To protect the polyether ester from oxidative degradation during synthesis and in end use applications, polymerization is usually carried out in the presence of an antioxidant. Hindered phenols such as Irganox 1098 (Ciba – Geigy) and Ethanox 330 (Ethyl Corp) or an aromatic secondary amine [i.e., Naugard 445 (Uniroyal)] are examples of antioxidants that are added during the preparation stage of the prepolymer. At the end of polymerization, the polymer is extruded, quenched, and cut into pellets. Predominantly hydroxyl and carboxylic acid end groups are produced by this process. The acid end groups are formed by the thermal cleavage of polymer chains or by degradation of a butanol end group with simultaneous elimination of tetrahydrofuran.

Both batch and continuous processes produce materials of sufficient viscosity to be fabricated by either extrusion or injection molding. However, a polymer with ultrahigh viscosity is required for blow molding operations and in extrusion processes where very precise dimensions are required. Generally, the polymer viscosities required for these methods of processing cannot be obtained by a melt condensation process. One technique used to manufacture ultrahigh-viscosity copolyester elastomers is *solid-phase polymerization* [87-90]. In this process the solid polymer is heated to just below its melting point under reduced pressure. This leads to the reaction of acid and hydroxyl end groups or of two hydroxyl end groups with elimination of water or short- chain glycols. They diffuse out of the polymer, and the molecular mass increases. A second method of increasing viscosity is by polymer chain extension in a postcompounding step. In numerous examples [91-93], diisocyanates have been used as chain extenders to increase the viscosity of low and moderate molecular mass polyesters through the reaction of the polyester hydroxyl end group to form a urethane.

Thermoplastic copolyester elastomers are stable for an indefinite period of time. However, they will absorb moisture, reaching equilibrium with existing atmospheric conditions very rapidly. Since the ester group hydrolyzes at processing temperature, these products must be dried thoroughly prior to fabrication. The moisture content should not exceed 0.1 wt % when molding or extruding [94].

3.4. Microstructure and Composition of Segments

3.4.1. *Short-Chain Ester Units (Hard Segments)*

The hard segment employed in the polyester elastomer strongly influences the load-bearing capabilities and processing characteristics of the polymer. Because of its high melting point, rapid crystallization, and low cost, the predominant hard

segment in commercial use is tetramethylene terephthalate (4 GT), an ester of 1,4-butanediol (4 G) and terephthalic acid (T). Hard segments based on ethylene glycol (2 G) and terephthalic acid (2 GT) are high melting and yield polymers with excellent physical properties. However, crystallization of this hard segment is very slow, requiring 36 h to completely develop the physical properties of the polymer. This limitation greatly reduces the usefulness of 2 GT thermoplastic elastomers in injection molding and extrusion applications [95]. WOLFE [96] has prepared a series of copolyester elastomers with a wide range of short- chain diols and compared their properties (Table 8). The properties of the elastomers resulting from 2 G, 3 G, (1,3-propanediol), and 4 G are very similar except for tensile and tear strength. Hard segments based on diols with five, six, and ten carbon atoms do not form high-melting crystals, and as a result, copolyesters made with these diols generally have inferior properties.

Table 8. Properties of 50 % alkylene terephthalate – PTMO terephthalate copolymer as a function of diol structure [96]

	Diol					
Properties	**2 G**	**3 G**	**4 G**	**5 G**	**6 G**	**10 G**
Melting point, °C	224	198	189	106	122	106
Shore D hardness	46	48	48	32	33	35
Stress at 100 % elongation, MPa	11.4	11.7	11.7	4.5	5.2	6.1
Tensile strength, MPa	45.5	22.8	48.4	15.4	13.5	15.5
Elongation, %	675	660	755	880	750	640
Tear strength, kN/m	42	15	48	25	18	7.8

Substituting a cycloaliphatic glycol, 1,4- cyclohexanedimethanol, for 1,4-butanediol produces a high-melting polyether ester elastomer that crystallizes somewhat slower than one containing tetramethylene terephthalate but rapidly enough to be blow or injection molded. Because of their slower crystallization characteristics, finished parts made of these elastomers are clear [97], [98].

Naphthalenedicarboxylic acid, in place of terephthalic acid, in combination with 1,4-butanediol provides products with improved thermal properties compared to their 4 GT counterparts [99]. Some mechanical properties such as tear strength are also improved. A major limitation in using naphthalenedicarboxylic acid in commercial thermoplastic copolyester elastomers, however, is its high price.

3.4.2. *Long-Chain Polyether Soft Segments*

The long- chain polyether soft segments are formed by reaction of poly(ether glycols) and diacids. The predominant poly (ether glycols) used in commercial thermoplastic copolyester elastomers are poly(tetramethylene oxide) (PTMO) and poly (propylene glycol) end- capped with ethylene oxide (EOPPG). In general, the chemical resistance and certain physical properties of thermoplastic elastomers, such as tear strength, are influenced by the poly(ether glycol) segment. For example, products based on PTMO have much higher tensile and tear strengths and significantly better resistance toward oxidative degradation than polymers made with EOPPG. Melting point, modulus, and hardness are generally independent of the type of polyether employed.

Because of their polarity, polymers made using poly(ethylene glycol) (PEO) show the greatest water swell, yet the least swell in nonpolar oils [100]. These products have high moisture vapor transmission rates.

A comparison of typical properties of thermoplastic polyester elastomers containing soft segments based on PTMO, EOPPG, or PEO is shown in Table 9 [101].

Table 9. Properties of 57 % 4 GT polyether elastomers as functions of polyoxyalkylene structure [101]

Properties	**Polyoxyalkylene**		
	PTMO	**EOPPG**	**PEO**
Nominal molecular mass	1000	1000	1000
Stress at 100 % elongation, MPa	13.4	11.7	13.2
Tensile strength, MPa	47.2	27.2	42.6
Elongation, %	660	675	705
Tear strength, kN/m	62	23	31
Volume swell, % [97]			
Water (3 d/24 °C)	0.5	1.0	17.0
	13.0	12.0	4.0

ASTM oil no. 3 (7 d/100 °C)

Through a novel set of reactions, polyethers end- capped with amino groups [102] also have been utilized to manufacture copolyester elastomers. These polyoxyalkylenamines are first converted to acid-terminated polyethers by reaction with benzene-1,2,4-tricarboxylic acid, followed by copolymerization with terephthalic acid and a short- chain diol. Since the backbone of the polyoxyalkylenamines is based on ethylene and propylene oxides, the resulting thermoplastic elastomers have properties similar to EOPPG-based products [103].

3.4.3. *Hydrocarbon Soft Segments*

Unsaturated straight- chain aliphatic carboxylic acids can be dimerized to form high molecular mass dicarboxylic acids, generally referred to as dimer acids [104]. When purified, these dibasic hydrocarbons have been used in place of poly (ether glycols) as soft segments in the preparation of thermoplastic copolyester elastomers. HOESCHELE [105] prepared several of these polymers and demonstrated the striking similarities with their polyether counterparts.

One major difference is that products based on dimer acid exhibit poor low-temperature properties because of their high glass transition temperatures. However, due to the hydrocarbon nature of the backbone, these polymers have much better heat aging characteristics and resistance to UV degradation. MCCREADY [106] has noted that the heat aging of thermoplastic polyether esters can be improved by using a mixed soft segment of dimer acid and poly(ether glycol).

3.4.4. *Polyester Soft Segments*

The use of low molecular mass polyester glycols as soft segments gives thermoplastic copolyester elastomers with improved heat aging characteristics and stability toward UV degradation compared to their polyether counterparts. However, the preparation of these products presents special problems because of the randomization that can occur between the hard and soft ester segments during synthesis. This negatively affects formation of blocks, and the resulting polymers are low melting with poor mechanical properties [107].

One approach to solving this problem is to react preformed hard and soft polyester seg-ments in the presence of a titanate catalyst until the reaction mass becomes clear. At that point the catalyst is deactivated by addition of a phosphorus compound [108]. On further processing into end use articles, very little ester exchange occurs between the hard and soft blocks, and as a result, parts with good mechanical properties are obtained. Generally, better results are obtained by chemical coupling of the blocks with diisocyanates [109] or acylating agents such as diacyl-bis-(*N*-lactams). A detailed description of diisocyanate coupling between poly(tetramethylene terephthalate) and various aliphatic polyester diols has been published by VAN BERKEL [107]. After deactivating the catalyst used for formation of 4 GT, coupling is carried out by using either toluene diisocyanate or methylene diisocyanate. A slight excess of isocyanate is used to obtain optimum mechanical properties. These thermoplastic polyester ester products show improved heat aging and UV resistance, but also have poor hydrolytic stability and stiffen much faster than polyether ester elastomers at low temperature.

3.5. Properties

Physical and Chemical Properties. Polymers prepared by the method described in Section Production have the molecular mass distribution expected from a condensation polymerization. A number-average molecular mass in the range of 20 000 – 25 000 can be obtained. The two-phase hard – soft segment structure of these segmented polyesters has been demonstrated clearly by differential scanning calorimetry [110]. By varying the ratio of these two segments, thermoplastic polyesters ranging from soft, transparent rubber-like elastomers to semirigid plastics can be prepared. Melting points ranging from 140 to 220 °C have been reported for polymers containing 20 – 80 wt % 4 GT [111]. As expected, the glass transition temperatures of this same series of polymers increase from – 70 to 25 °C.

Thermoplastic copolyester elastomers have good to excellent resistance to hydrocarbons. Products containing > 50 % 4 GT are particularly well-suited for contact with oil, grease, or hydraulic fluids. These polymers offer better resistance to hot hydrocarbons than nitrile rubber.

Resistant to

- Nonpolar fluids
  - Oil
  - Fuels
  - Hydraulic fluids
- Polar fluids at room temperature
  - Glycols
  - High molecular mass alcohols
  - Weak acids and bases

Not resistant to

- Hot glycols
- Hot alcohols
- Strong acids and bases
- Amines

Resistance to polar fluids, including water, acids, bases, amines, and glycols, is a function of both the composition of the polymer and the temperature of use. At > 70 °C, polar fluids attack the polymer.

Mechanical Properties. The predominant hard segment in thermoplastic copolyester elastomers is poly(tetramethylene terephthalate). WOLFE [96] has shown that a hard segment content of 30 to 80 wt % produces a wide range of useful products (Table 10). A complete summary of mechanical properties for these elastomers has been published [112].

Table 10. Properties of tetramethylene terephthalate – PTMO terephthalate copolymer as a function of tetramethylene terephthalate content [96]

Properties	Percentage 4 GT				
	30	**40**	**50**	**57.4**	**80**
Melting point, °C	152	172	189	197	212
Shore D hardness	34	44	48	54	70
Stress at 100 % elongation, MPa	5.9	8.3	11.7	13.4	24.4
Tensile strength, MPa	24.6	31.0	48.4	46.7	54.5
Elongation, %	900	830	755	660	535
Tear strength, kN/m	13	26	48	63	108

3.6. Blends

The low melt viscosity and good processing characteristics of thermoplastic elastomers facilitate their use in polymer blends to give products that are otherwise not easily obtainable. In many instances the properties of these blends are an average of the properties of the individual components, for example, the combination of poly(vinyl chloride) (PVC) and a polyester based on 4 GT and PTMO [113], [114]. The addition of 10 – 30 % of thermoplastic polyester elastomer provides PVC with improved heat resistance, low-temperature flexibility, and resistance to oils and fuels. Another important advantage is that the polyester elastomer cannot be extracted from the blend, which extends the service life of PVC molded parts.

In other instances, blends offer improved physical properties that are not obtainable by direct synthesis. Blending poly (butylene terephthalate) with a thermoplastic copolyester elastomer containing 58 wt % hard segment provides a polymer with improved low-temperature impact resistance compared to a directly synthesized polymer of identical composition [113]. The improvement in low-temperature properties is said to result from the incompatibility of the components of the resultant blend.

High-melting, crystalline thermoplastic copolyester elastomers can be softened (thus improving their low-temperature properties) by blending with an amorphous or very low melting polyester made from PTMO, 4 G, and a diacid [115]. Since the melting point of the resultant blend is similar to that of the unblended crystalline thermoplastic, evidence that the two components of the blend are not miscible is good. Soft thermoplastic copolyester elastomers can also be prepared by blending the polyester with ethylene copolymers. SHIH [116] reports blends with EPDM that retain the melting point of the polyester yet exhibit reduced stress – strain properties. In an extension of this work, blends of polyesters have been prepared with ethylene – carboxylic acid copolymers that had been partially neutralized [117]. This novel combination of polymers provides blends with enhanced melt viscosity, which greatly improves the processibility by blow molding and film extrusion of the thermoplastic copolyester elastomers.

Very efficient mixing is required to achieve good disperson of polyester homopolymers and thermoplastic copolyester elastomers because of their incompatibility [113]. If the materials being blended differ greatly in polarity, homogeneous blends can sometimes not be achieved unless *compatibilizers* are employed. These compatibilizers generally contain reactive functional groups to assist in dispersing the components of the blend. In two separate examples, polycarbonate resins having improved flexibility and impact strength were obtained by blending with copolyester elastomers. To obtain optimum properties, polystyrene grafted with organosiloxanes or styrene – butadiene copolymerized with acrylic acid derivatives has been used to compatibilize the blend components [118], [119].

3.7. Uses

Thermoplastic copolyester elastomers are available in grades ranging from soft, rubber-like to semirigid plastics. This provides customers with products covering a wide spectrum of load-bearing characteristics, flexibility, and low-temperature properties. These materials are simple to process and can be made into high-performance products by a variety of polymer-forming processes, including injection molding, extrusion, blow molding, rotational molding, and melt casting. All thermoplastic copolyester elastomers have sharp, well-defined melting points, crystallize rapidly, and have good melt stability.

Grades having a Shore D hardness of 50 – 80 have excellent chemical resistance, which makes them suitable replacements for specialty rubber such as nitrile and epichlorohydrin rubbers in hose and tubing applications. They can be extruded into long, continuous lengths with relatively low wall thickness, thereby providing lightweight constructions. Compared to competing flexible polyamide compounds, no plasticizer is required, so the material remains flexible over long periods of time. Overall, hose construction costs using thermoplastic copolyester elastomers are comparable to those using conventional rubber. This is due to the less expensive plastics processing techniques involved and the fact that thermoplastic copolyester elastomers require no postcuring [112].

The good impact strength and toughness, dimensional stability, and good paintability of thermoplastic copolyester

elastomers have been utilized in automotive applications to replace metal in exterior automotive bumpers and front and rear end spoilers. The use of thermoplastic copolyester elastomers is not limited, however, to the exterior of automobiles. Overmolded interior parts such as door handles having good “feel” properties also have been commercialized [120].

In a very demanding application requiring excellent flexural fatigue resistance, thermoplastic polyesters have replaced polychloroprene rubber in constant-velocity joint boot bellows for front-wheel-drive automobiles. In addition to the necessary flexibility, this application also requires a product with good thermal properties, high tear resistance, very good low-temperature properties, and resistance to lubricating greases [121].

Thermoplastic copolyester elastomers containing either EOPPG or PEO soft segments absorb significant quantities of water. Because of this characteristic, thin films produced from these materials have high moisture vapor transmission rates. In addition to their use in laminated films for moisture-proof sports and rain wear [122], [123], these materials have also found application in the medical field [124]. By varying the EOPPG or PEO content, a wide range of moisture vapor transmission rates can be obtained. These products are suitable for medical wound care treatment, transdermal patches, and surgical drapes [124].

3.8. Producers, Trade Names

Thermoplastic copolyester elastomers were originally introduced by DuPont in 1972 under the trade name Hytrel and are manufactured by DuPont in the United States and Luxembourg, and by DuPont – Toray in Japan. Other U.S. producers include General Electric (trade name Lomod), Hoechst – Celanese (trade name Riteflex), and Eastman Chemical Products (trade name Ecdel). Both copolyether esters and copolyester esters are manufactured by DSM in the Netherlands, under the trade name Arnitel, and by Toboyo in Japan, under the trade name Pelprene. Pibiflex is manufactured in Europe by Montedison. More recent manufacturers entering this field include two South Korean producers, Kolon Industries (Kopel) and Cheil Synthetics (Esrel).

[Top of Page]

4. Thermoplastic Polyamide Elastomers

Franz Erich Baumann and Jörg Lohmar

4.1. Introduction

Polyamides have been highly regarded thermoplastics for more than 50 years because of their high heat deflection temperatures, rigidity, and excellent resistance against solvents and aggressive chemicals. Their low melt viscosity permits easy processing, e.g., in injection molding. These and other positive properties have been transferred successfully to a greater or lesser extent to intrinsically softened polyamide elastomers.

In the early 1980s Chemische Werke Hüls (today Degussa) and Ato-Chimie (today Atofina) independently developed processes for the synthesis of multiblock copolymers containing sequences of different polyamides and various kinds of polyethers linked by either ester or amide bonds. These polyamide elastomers (PAE) – sometimes referred to as **P**oly **E**ther **B**lock **A**mides (PEBA) – behave like thermoplastic elastomers, i.e., they combine elastomeric properties over a wide range of temperature with an easy thermoplastic processibility. Because their properties are closely related to the basic polyamides PAE can be assigned to the high-performance class of thermoplastic elastomers.

4.2. Raw Materials

Polyamide elastomers consist formally of difunctional crystallizable oligoamides (hard segments) and difunctional soft segments having a low glass transition temperature. In many cases this sequential cocondensation is also a viable preparative procedure. Various embodiments of this “internal plastification” concept have been described such as (co-) polyolefin derivatives [125-131] and polysiloxane soft segments [132], in 2002 polycarbonate diols as soft segments have been described [133], however only polyether derivatives have become technologically important on a commercial scale so far [134-138].

Polyether diols and diamines undergo condensation with dicarboxylic acid terminated oligoamides to form sequential multiblock copolymers via polyesterification or polyamidation, respectively [137], [138]. The typical structural patterns of these copolymers can be simply visualized as shown below.

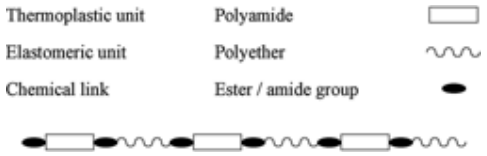

Typical PAE building blocks are polyamides with carboxylic acid terminal groups, e.g., PA 12 (polylaurolactam, number average molar mass, M_n = 600–4000 g/mol), polyethers with hydroxyl terminal groups or polyethers with amino terminal groups, e.g., polytetrahydrofurandiol (M_n = 600–3000 g/mol).

Oligoamide Unit. Oligoamides having carboxylic terminal groups are obtained by reacting excess dicarboxylic acid with a lactam, an aminocarboxylic acid or a nylon salt [138]. The dicarboxylic acid excess is tantamount to adjust a number average molar mass M_n of usually 600 to 4000 g/mol.

Polyether Units. *Polyether diols* applied in commercial PAE grades are based on polyethylene glycol (PEG), polypropylene glycol (PPG), and most frequently on polytetramethylene glycol (polytetrahydrofuran, PTHF). Polytetramethylene glycol diols have been commercialized in Europe and America, e.g., by Du Pont (TETRATHANE),

BASF (PTHF), or Great Lakes (POLYMEG). In Japan PTHF diol grades are manufactured, e.g., by Asahi Chemical or by Hodogaya.

Polyether diamines have been commercialized by Huntsman - formerly by Texaco – as JEFFAMINE D-grades (polypropylene glycol diamine) and as JEFFAMINE ED-grades (polyethylene glycol diamines).

4.3. Manufacture of Block Polyetheramides

The most important group of polyetheresteramides is derived from polyamide 12. The monomer basis is laurolactam which is converted to polyether block amides (PEBA) usually in a two-step process [134], [138]. The intermediate is an oligoamide dicarboxylic acid which is then condensed to a PEBA by forming ester or amide links (Fig. 8).

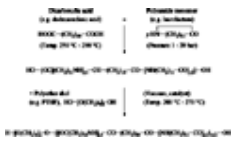

Figure 8. Two step formation of polyetheresteramide type PAE

4.3.1. *Polyetheresteramides*

The first step - ring opening polymerization of laurolactam - requires either acidolytic conditions, i.e., sufficiently high amounts of dicarboxylic acids [141] or the presence of water at high temperatures [142]. It may be advantageous to enhance the miscibility of oligoamide and oligoether by pressurizing the reaction mixture with water vapor [143].

The second polyesterification stage usually requires addition of a catalyst. Recommended classes of catalysts are dialkyltin(IV) compounds [138], phosphoric acid [144], tetraalkyl titanates [144], [145], tetraalkyl zirconates [146], tin(II)) salts [147], or antimony oxide [148], the latter also being applied as a mixture with tin and phosphorus compounds [149].

A *single-step batch synthesis* has been described [144], in which both lactam polymerization and polyesterification are performed in the same reactor: Usual polyamidation equipment can be employed, such as stirred autoclaves, provided that sufficient water vapor pressure up to 25 bar for the lactam cleavage can be applied and the reaction vessel can be evacuated for the second stage.

The one-pot synthesis, i.e., charging the reaction vessel with both polyamide forming monomers, including excess dicarboxylic acid, and polyether diols or diamines is always the method of choice when polyamide monomers more reactive than laurolactam are used: Drastic reaction conditions necessary for the ring opening of laurolactam – high temperatures over an extended period of time may cause thermal degradation of polyethers – can be avoided in these cases, in particular for caprolactam [140], aminocarboxylic acids [150], nylon salts, or equimolar mixtures of diamines and dicarboxylic acids [151] as polyamide forming monomers.

A continuous process applying *thin layer evaporation* has been described by Atofina [152-154], it requires, however, a separate synthesis of oligoamide dicarboxylic acid. This dicarboxylic acid terminated oligoamide is fed with a stoichiometrically equivalent amount of polyether diol via a static mixer into one or several thin-layer reactors equipped with rotating film-forming wipers. These elements are inclined to promote a reverse flow in the initial phase of the polyesterification, whereas in the final stages they are positioned in a way to promote downstream flow of the higher viscous melt [152]. The published procedure makes use of three sequential thin-layer reactors [152].

4.3.2. *Synthesis of Polyetheramides*

Condensation of oligoamide dicarboxylic acids with polyetherdiamines can be performed without the necessity of applying vacuum in the final phase of polycondensation [155], [158].

Polyamide elastomers based on this structural pattern have been developed by Ube, Ems (GRILAMID, ELY), and former Rhône Poulenc (DYNYL) [155], [159]. An additional benefit of the amide link instead of the ester group is the increased hydrolytic stability. The synthesis is usually performed batchwise in pressurized polyamide reactors. Both a single-step process [160] and a two-step process comprising a separate preparation of oligoamide dicarboxylic acids [161] have been described by Ube. Cocondensation of polyetherdiamines with low molar mass aliphatic diamines has also been described [162]. In the Ube synthesis a dimeric fatty acid is preferred both as a nylon chain terminator and as a link between polyetherdiamine segments.

Hybrid elastomers based on polyetheramides terminated with both amino and hydroxyl groups have been described [163], [164]: Reaction with isocyanates converts these polyetheramides into polyurethanes or polyureas, respectively.

Polyetheramides with very high melting points derived from nylon 46 segments have been prepared [165], [166] employing polyethers terminated by two primary amine moieties.

4.4. Physical and Chemical Properties

4.4.1. *Block Variations*

The properties of PAE can be varied and determined by the selection of both hard and soft block types (Table 11). The choice of hard blocks defines the upper service temperature, corresponding with the melting temperature of the base polyamide. Also water absorption, chemical stability, and density are determined by the chosen polyamide. The length of the hard blocks is controlled by the amount of chain limiting dicarboxylic acid and affects the melting range, the crystallinity and the transparency of the PAE.

Table 11. Variation of block types and correlations with PAE properties

Variable	Influenced property
Type of polyamide block:	melting range (upper service temperature)
	water absorption
PA 12, PA 11, PA 6, PA 66, PA 612	chemical stability
	density
Length of polyamide block	melting range (upper service temperature)
	crystallinity
	transparency
Type of polyether block:	glass transition temperature (lower service temperature)
water absorption, water permeability	
PTHF, PEG, PPG	electrostatic properties
	chemical and thermal stability
Length of polyether block	transparency
Block composition	rigidity, hardness, flexibility
Polyamide / polyether	

The choice of the soft segment type determines the glass transition temperature and, hence, the lower service temperature. Using hydrophilic soft segments such as polyethylene glycol or polypropylene glycol the moisture absorption and permeability is enhanced and permanent antielectrostatic properties induced. Soft segments with higher molar mass tend to phase separation causing lower transparency. Using polyethers with amino end groups (i.e., JEFFAMINE) the resulting polyetheretheramides have higher thermal and hydrolytic stability.

Finally, the block composition of polyamide and polyether blocks can be varied in a wide range and the resulting products encompass an outstanding range of rigidity. This is demonstrated in Figure 9 in terms of Shore hardness. PAE with high polyamide block contents have a Shore hardness D up to 70, i.e., close to polyamides themselves. With lowest polyamide block contents PAE can be very soft with Shore hardness A down to 60, hence, in the domain of soft rubbers. Such soft grades cannot be achieved using other TPEs.

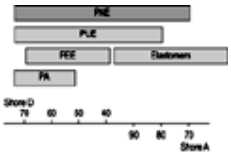

Figure 9. Range of hardness for PAE

PAE = Polyamide elastomer; PUE = Polyurethane elastomer; PEE = Polyester elastomer; PA = Polyamide

Figure 10 shows a series of tensile stress-strain curves of PA 12-PAE with varied PA 12 block content. PAEs rich in PA 12 exhibit the typical behavior of thermoplastics with yielding. With increasing polyether content there is a transition to more and more elastomeric characteristics.

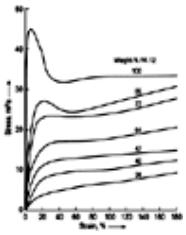

Figure 10. Stress-strain diagrams of PA 12-PAE from tensile tests according to ISO 527

4.4.2. *Morphology*

The specific mechanical properties of PAE are strongly determined by the phase morphology of the principally incompatible polyamide and polyether blocks. Four different phases had been identified by thermoanalytical methods and electron microscopy [167-169]. Differential scanning calorimetry revealed two melting signals, the dominant one corresponding with the melting of PA 12 blocks, the other one with a little content of crystallized PTHF blocks (Fig. 11). With decreasing level of PA 12 hard blocks the melting temperature drops indicating that the size of crystallites is becoming smaller. Amorphous phases have been detected by torsional vibration analysis. Figure 12 shows the loss modulus plots of the investigated PA 12 – PAE series. The maximum values of the plots correspond with the glass transition temperatures of amorphous phases. In a range of higher PA 12 block content, glass transition temperatures of two mixed amorphous phases with different block compositions are observed. At higher PTHF block content, the loss modulus plots are dominated by the signal of a PTHF-rich mixed phase.

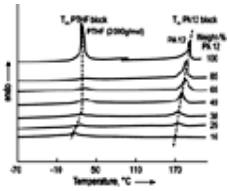

Figure 11. Melting point diagrams (DSC) indicating two crystalline phases in PA 12-PTHF-PAE (T_m = melting temperature)

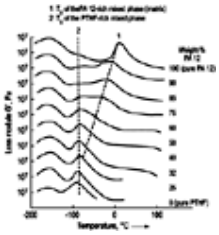

Figure 12. Loss modulus curves from torsional vibration analyses indicating two amorphous mixed phases in PA 12-PTHF-PAE (T_g = glass transition temperature)

With the aid of highly sophisticated contrast enhancing techniques, transmission electron microscopy (TEM) images were obtained demonstrating the unique morphology in PAE. Crystallized lamellae of PA 12 blocks form the same spherolitic superstructures as known from polyamides (Fig. 13). These crystalline superstructures form an interpenetrating reinforcing “physical network” for the amorphous mixed phase of PA 12 and PTHF blocks between the lamellae. At lower PA 12 content, only fragments of spherolites or single lamellae can be formed with a weak reinforcing effect on the soft amorphous PTHF-rich mixed phase (Fig. 14). For completeness it should be mentioned that PTHF-rich PEBA molecules build up a separate disperse phase due to the strong incompatibility between both blocks. This is the second amorphous mixed phase with the low glass transition temperature observed in torsional vibration analysis.

Figure 13. Transmission electron micrographs showing spherolitic superstructures of crystallized lamella in PA 12-PTHF-PAE (PA 12-rich PAE)

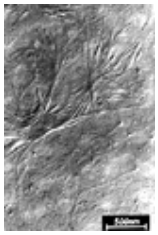

Figure 14. Transmission electron micrographs showing spherolitic superstructures of crystallized lamella in PA 12-PTHF-PAE (PTHF-rich PAE)

A number of the unique mechanical properties which distinguish PAE from all other TPE can be deduced from the interpenetrating network of crystallized polyamide blocks. This is valid for example for the lower temperature dependence of stiffness and elasticity compared to other TPEs.

4.4.3. *Outstanding Properties*

The specific advantages of PAE (PA 12-PTHF) compared with other TPE can be summarized as follows:

- Low density: 1.01 – 1.03 g/cm^3 (PUE, PEE: 1.14 – 1.25 g/cm^3),
- High stability against chemicals and solvents,
- Easy processing, coloring, and overmolding,
- Excellent ability to be decorated by thermodiffusion printing,
- Excellent cold impact strength,
- Wide range of hardness and flexibility,
- High elasticity and resilience,
- Low temperature dependence of mechanical properties,
- Excellent abrasion resistance (only surpassed by TPU),
- Highest service temperatures (especially PAE with PA 6 blocks).

Some of the outstanding properties of PAE are listed in Table 12 for a series of commercial products (VESTAMID E, Degussa AG) based on PA 12 and PTHF: The density (close to 1 g/cm^3) is up to 20 % lower than that of competing TPE. The flexural modulus of PAE comprises a full decade of magnitude, from 45 to 500 MPa. The cold impact strength of PEBA is excellent, as demonstrated by Izod notched impact data at –30 °C in Table 12. Very soft and medium soft PAE do

not break at all, stiffer grades with Shore hardness D >50 and tensile moduli above 200 MPa have still excellent toughness data.

Table 12. Major properties of PA 12-PAE compared with PA 12

Properties	Test-method	Unit	PA 12-PAE (VESTAMID E) with Shore hardness D (ISO 868)						PA 12
			30*	40	47	55	62	68*	72
Density at 23 °C	ISO 1183	g/cm^3	1.01	1.01	1.02	1.03	1.03	1.03	1.01
Tensile modulus	ISO 527 –1/–2	MPa	45	85	115	240	360	500	1500
Izod notched impact strength at –30 °C	ISO 180/1A	kJ/mm^2	no break	no break	no break	21	8	7	6
Heat deflection temperature									
Method A 10N	ISO 75-1	°C		45	45	45		50	
Method B 50N	ISO 75-2	°C		55	65	90	100		110
Vicat softening temperature Method A/10 N	ISO 306	°C	90	125	140	160	165	170	174

*Presently not commercially available.

The uptake of water is a crucial property for many PAE applications. In this regard PAE based on PA 12 and PTHF blocks do not differ significantly from PA 12 which has the lowest water absorption of all polyamides. With increasing polarity of polyamide and polyether building blocks the water uptake of PAE increases dramatically (Table 13). This creates a strong influence on the electrical behavior of the PAE. PAE with PA12 and PTHF-blocks are good insulators, while, e.g., PAEs based on PA 6 and polyethylene glycol blocks are permanent antielectrostatic due to their fast and high moisture uptake from the ambient air. High water uptake of PAE is not tolerable for applications where dimensional stability of parts is crucial. On the other hand PAE with higher absorption and permeability of moisture have found applications as antistatic modifier and, e.g., for moisture breathing film.

Table 13. Density and water absorption of various PAE in comparison

TPE-type	Density ISO 1183 g/cm^3	Water absorption (saturated) 20 °C / 65 % rh, wt %	20 °C / 24 h water storage, wt %
PA 12-PTHF	1.01 – 1.03	0.5	1.2
PA 6-PPG	1.10	2.4–2.8	6.1 – 6.4
PA 6-PEG	1.14	4.5	120
PUE*	1.14 – 1.25	0.1 – 5	0.5 – 15
PEE*	1.17 – 1.25	0.2 – 0.4	0.6 – 0.8
Rubber*	>1.20	0.6 – 0.8	0.1 – 15

*Strongly varying for different types.

4.4.4. *Chemical Resistance*

Due to their close relationship to polyamides PAE exhibit high stability against many chemicals and solvents compared with other TPE. The stability of PAE against diluted hydrochloric and sulfuric acid as well as bases should be especially noted. Rigid PAE grades show only low swelling in ASTM oil, even at 100 °C. This is also valid for hydraulic liquids, aliphatic solvents, and alcohols. With softer PAE grades the swelling in aromatic solvents and fuels is rather high, however, the mechanical properties are not affected dramatically.

4.4.5. *Processing*

PAE can be processed on standard equipment and with processing conditions usual for polyamides. For injection molding of PA 12-PAE melt temperatures of 190–230 °C are recommended for rigid grades and 170–210 °C for softer grades. Mold temperatures of 20–40 °C are suitable for all grades. Because of the very low melt viscosity and rapid crystallization, short injection cycle times are possible.

At extrusion temperatures of 210–230 °C, thin sheets, hoses, tubes, profiles, plates, and monofilaments can be processed. Other processing techniques applied for PAE are blow molding, rotational molding, thermoforming, and powder coating.

4.5. Uses

Since the polyamide block building monomers are rather expensive because of their multistep synthesis, PAE are ranking at the high end of the scale of prices of TPE together with PUE and PEE. As a consequence they are only used where one or more of their outstanding product and processing properties are required.

Polyamide elastomers are especially acknowledged for the construction of lightweight soles for specialists sport shoes. PAE are used primarily because of their low density, but also because of their good resilience, cold impact strength, and low temperature dependence of flexibility. The excellent processing and overmolding properties allow high sophisticated sole constructions. This has been applied for designs with sole parts of different stiffness and for decorative effects in overmolding of parts with different colors (Fig. 15). Even PA 12 based composites and PA 12 coated metal sheets have been incorporated in sole constructions. Recently the combination of overmolding and decorative printing techniques have been applied for new sports shoe soles constructions.

Figure 15. Sport shoe soles with overmolded PAE parts and PA 12 composites

Another main application field for PAE are winter sports goods such as lightweight ski boots or PAE films for decorative ski and snowboard surfaces. By means of thermodiffusion printing processes brilliant decorations are applied often on the bottom side of a PAE film. The colorants penetrate only about a 100 µm into the film of typical 500 µm total thickness. The film itself then acts as a protective layer for the decoration and the print motifs are almost indestructible (Fig. 16).

Figure 16. Design of skies with decorative PA12-PAE films, sublimation printed from the bottom side

Soft PEBA grades, with Shore hardness D below 40, have been used in such applications as straps of watches. Beside flexural properties, the dermatological compatibility, UV stability, and translucence of PEBA is striking. For frames of sunglasses, dermatological compatibility, easy processing, and coating stability of these elastomers is important. Medical uses are, e.g., catheters, or controlled drug release carriers. Hydrophilic grades are applied for moisture breathable film in surgical drapes and garments.

The excellent processing and demolding properties and the dimensional stability of PAE make them well suited for manufacturing precise moldings. PAE gear wheels (Fig. 17), for example, contribute to the noiseless operation of tape recorders or CD players.

Figure 17. Precisely molded PAE wheels for noiseless gears

For permanent antistatic housings PAE are compounded with special conductive carbon blacks which have still good impact resistance in spite of the necessary high loadings. Applications are permanently antielectrostatic parts, for example, castor wheels (Fig. 18), conveyor belts, rollers, and paint spray hoses.

Figure 18. Tough antielectrostatic PAE parts highly loaded with conductive carbon blacks

4.6. Trade Names

With a consumption of some 5000 t/a PAE play only a niche role in the TPE market. Suppliers for PAE are Degussa AG/Germany (VESTAMID E-grades, Shore hardness D 30–70, PA 12 base), Atofina/France (PEBAX, Shore hardness D 25–70 , PA 12 and PA 6 base, including hydrophilic grades), EMS Chemie/Switzerland (GRILAMID ELY, GRILON ELX, Shore hardness D 50–63, PA 12 and PA 6 base), and Ube/Japan (UBESTA, selected grades).

[Top of Page]

5. Styrenic Block Copolymers

Hans F. Vermeire

5.1. Introduction

Thermoplastic elastomers known as styrenic block copolymers (SBC) have emerged from developments and research in the organometallic-initiated polymerization of diene monomers in hydrocarbon solvents (⟶ Rubber, 3. Synthetic). Commercial block copolymers are essentially based on styrene, butadiene, and isoprene monomers. The first commercial quantities were produced by Shell in 1964 in its production plant in Torrance, Ca., and marketed under the trade name Kraton [170]. Worldwide capacity for styrenic block copolymers in 1993 was estimated at 500 000 t/a.

Styrenic block copolymers combine rubber performance with thermoplastic processibility on the basis of their specific structure. Anionic polymerization allows the selective polymerization of butadiene or isoprene and styrene in such a way that the resulting polymer is essentially a polybutadiene or polyisoprene molecule tipped at both ends with a polystyrene molecule [171]. Such a structure is commonly identified as ABA, where A represents the plastic and B the elastomer. More specific notations used are SBS or SIS polymers. Styrenic block copolymers with saturated rubber blocks are produced from SIS or SBS precursors by selective hydrogenation. This process converts the polyisoprene into poly(ethylene – propene) and the polybutadiene into poly(ethylene – butene). To provide rubber performance, the volume fraction of plastic (polystyrene) in these structures must be < 50 %, and commercial hydrogenated grades usually contain around 30 %. Newer variants of the styrene – ethylene/butene – styrene S – E/B – S or styrene – ethylene/propene – styrene S – E/P – S polymers are functionalized with maleic anhydride or other reactive comonomers. Functionalization levels are between 0.2 and 5 % and the reactive group is grafted onto the rubber mid block [172].

5.2. Synthesis

The first step in the production of SBCs is activation of purified styrene in a hydrocarbon solvent with butyllithium and subsequent anionic polymerization. The molecular mass of the first polystyrene block is determined by the butyllithium : styrene ratio (see Fig. 19). The formed living polymer, PS – Li, is then reacted with purified butadiene to form the second block. When all the diene has been converted, two routes can be followed to complete the formation of SBC:

1. Addition of a coupling agent to produce linear or branched copolymers
2. Addition of a second amount of styrene monomer to form the third block

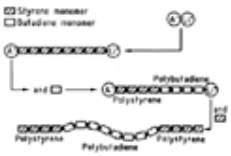

Figure 19. Formation of block copolymers by anionic polymerization (A^- = butyl anion)

Technically feasible coupling agents include di-, tri-, tetra-, or multifunctional compounds that allow formation of a multitude of different polymer structures [173].

The procedure for production of the precursor for *saturated block polymers* is the same. To ensure a rubbery mid block after hydrogenation in the S – E/B – S polymer, a modifier is added during polymerization of the polybutadiene mid block to increase its vinyl content (1,2- content) (⟶ Rubber, 3. Synthetic) to approximately 40 % which will ensure optimum elastic properties after hydrogenation. This allows optimization of the ratio of ethylene and butene units in the hydrogenated polymer. After preparation of the precursor, polymer solution is transferred to the hydrogenation section where the double bonds of the polydiene are reacted with hydrogen in the presence of a catalyst.

Functionalized saturated block copolymers are made by extruder grafting of an S – E/B – S or S – E/P – S molecule. The polymer is contacted in an extruder with a peroxide to create free radical sites along the rubber mid block. These react in the extrusion step with a compound containing the functional group and a group able to react with the free radical sites (e.g., maleic anhydride). The process requires careful optimization of the amount of peroxide and reaction conditions to minimize polymer degradation.

The result of the polymerization process is a solution of polymer in hydrocarbon solvent. At this stage, additives such as stabilizers, antioxidants, or plasticizers can be added to the solution which is subsequently converted into solid polymer during finishing. The following systems are used industrially:

Direct Desolventizing. The polymer solution (cement) is introduced directly into a special vacuum extruder; the solvent is evaporated into a recovery system and the extruded polymer melt is face- cut and cooled at the die head of the extruder to form compact transparent pellets.

Coagulation with Steam. Cement is pumped to a vessel where steam is passed into the solution. The resulting coagulate separates from the solvent, and the polymer – water mixture is dried to remove water. The usual form resulting from the coagulation – drying route is referred to as crumb.

5.3. Properties

Thermoplastic rubbers based on styrenic block copolymer can be considered polybutadiene [or polyisoprene, poly (ethylene – butene), or poly(ethylene – propene)] tipped with polystyrene end blocks. Polybutadiene (butadiene rubber, BR) and polystyrene of sufficiently high molecular mass are incompatible, which forms the basis for the performance of these thermoplastic rubbers. At temperatures above the softening point of polystyrene the SBC molecule is quite mobile and can be melt processed by, e.g., extrusion or injection molding. On cooling, the PS end blocks separate from the BR

mid blocks, and two distinct phases are formed. The separate, discrete phase of the PS domains acts both as a cross-linking material and as a reinforcing filler analogous to the sulfur and carbon black in vulcanized rubber. Figure 20 shows an SBC where the PS domains hold together a rubbery BR network. The tensile strength (Table 14) and resilience data (Table 15) illustrate the truly rubbery behavior of a linear SBS polymer with a polystyrene content of 30 %.

Table 14. Comparison of tensile strengths of BR and SBS

Polymer	Tensile strength, MPa
Unvulcanized BR	0.2 – 0.5
Vulcanized BR	4.0 – 5.0
Vulcanized + carbon black reinforced BR	5.0 – 20.0
Unvulcanized pure SBS	30.0 – 35.0

Table 15. Resilience (Lüpke-rebound) of different polymers

Polymer	Resilience, %
Unvulcanized SBS	65
Vulcanized BR	65
Unvulcanized SIS	60
Vulcanized NR/IR	60
Vulcanized SBR	50
Vulcanized EPDM	45
Unvulcanized EVA (28 % VA)	40

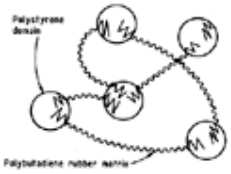

Figure 20. Schematic phase structure of an SBS block copolymer

The *molecular structure* of an SBC can be varied by changing the proportion of PS in the base polymer. Furthermore the total molecular mass can be varied, but practical limits exist to the processibility in the melt. Figure 21 shows the effect of altering polystyrene content in a series of linear polymers. Polymers can be made that range from very low strength at low (10 %) PS content, through rubbery materials of about 30 % PS, to those with 80 % PS content that behave much like toughened polystyrene. SBC polymers can be linear or branched; branched types are usually higher in molecular mass than linear ones and have a higher melt viscosity at equal strength levels. Furthermore, a minimum molecular mass of the end block exists below which no phase separation occurs, no domain structure forms, and hence unsatisfactory properties are obtained [174]. A comparison of properties of polymers differing in the molecular mass of the PS blocks and in total molecular mass is given in Table 16.

Table 16. Properties of styrenic block copolymer structures with different PS block and total molecular mass

Polymer structure	Molecular mass		Viscosity	Strength
	PS block	Total		
Linear	12 000	80 000	low	high
Branched	12 000	160 000	medium	high
Branched	6000	80 000	very low	low

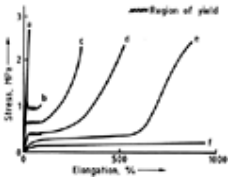

Figure 21. Stress – strain curves for various SBS copolymers

a) 80 % styrene; b) 65 % styrene; c) 50 % styrene; d) 40 % styrene; e) 30 % styrene; f) 10 % styrene

The *processing properties* of SBCs can be derived from their basic chemical structure. The molecular mass of commercially available grades varies from about 75 000 to 400 000. Their behavior in the molten state is similar to that of polyethylene and polystyrene, for example. However, even in the molten state the tendency to form PS domains persists, especially for those types with a saturated rubber mid block. *Melt processing SBCs* at higher shear rates leads to lower apparent viscosity; this effect is relatively independent of temperature (see Fig. 22). Reprocessing of SBCs at < 200 °C has little effect on their properties; saturated grades can be processed and reprocessed at up to 280 °C.

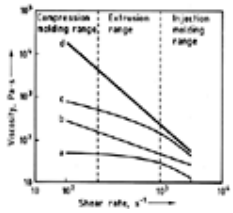

Figure 22. Viscosity – shear rate relationship at 200 °C

a) Clear polystyrene; b) SBS; c) polypropylene melt flow rate (MFR) = 5; d) S – E/B – S

Solution processing of SBCs is possible, provided that solvents of sufficiently high solubility are used to dissolve the PS end blocks. Solvents with solubility parameters between 8 and 10 are usually employed. Because of the easy solubility of SBCs, adhesive, sealant, and coating systems can be designed that have low viscosity and high solids content. A consequence of their good solubility in organic solvents is that the oil and solvent resistance of parts based on SBC is limited. However, their resistance to dilute aqueous solutions of the most commonly encountered chemicals is generally satisfactory.

The *service performance* of SBCs and the products derived from them is also related to their basic chemical structure. Because of their moderate molecular mass and their physical and reversible cross-linking, they flow under pressure and exhibit creep. Block copolymers with *BR or IR mid blocks* are unsaturated and sensitive to aging. Exposure to elevated temperature can result in polymer degradation with simultaneous decrease in viscosity and tensile strength. With the exception of black compounds, exposure to UV light for any length of time will result in discoloration and surface cracking, which becomes more pronounced on prolonged exposure. These block copolymers are also susceptible to ozone attack, and in shaped products at points with local stresses, microcracks develop on exposure to atmosphere containing ozone. The use of antioxidants, UV stabilizers, and antiozonants reduces these degradation effects to workable limits in end uses.

The *hydrogenated S – E/B – S or S – E/P – S types*, in contrast, have a saturated rubber mid block and are therefore highly resistant to degradation by oxidation, ozone, and UV light. They can be processed over a much wider processing temperature range.

5.4. Uses

Because of their rubbery characteristics (especially at low temperature), high strength, compoundability, and thermoplastic nature, SBCs are used in a variety of applications as highlighted in Figure 23.

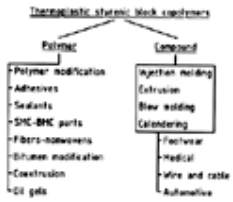

Figure 23. Applications of styrenic block copolymers

SMC = sheet molding compounds; BMC = bulk molding compounds

On the left-hand side of Figure 23 the list includes outlets that use SBC as an additive or modifier for other materials. The polymers are delivered to the end user who will incorporate them in his existing process. The right-hand column lists applications that utilize a compounded product with performance tailored to the specific end use and plastic conversion technology. A comprehensive review of the applications of thermoplastic rubber is given in [175], [176].

5.4.1. *Adhesives, Sealants, and Coatings*

Adhesive properties are determined primarily by the styrenic block copolymer type chosen for a typical application. SBS polymers with high cohesive strength (high polystyrene content) are commonly used in contact adhesives, whereas SIS polymers with low polystyrene content that can be tackified easily are preferred for pressure-sensitive adhesives. For applications in hot melt form, low molecular mass polymers with a high melt flow rate are selected. When high UV resistance is required, hydrogenated block copolymers should be considered as base polymer. Furthermore, adhesive properties are also determined by the nature and the concentration of the various resins and plasticizers used and their degree of compatibility with the respective mid or end block phase of the block copolymer [177].

Pressure-sensitive adhesives are materials that remain permanently tacky and adhere instantaneously to solid surfaces on the application of light pressure. Major applications are tapes and labels. Tapes require high shear strength and holding power, while labels require less shear strength but good convertibility. For general-purpose solvent-based pressure-sensitive adhesives, SIS block copolymers are combined with resins that are compatible with the mid block, such as wood rosins, polyterpenes, or synthetic hydrocarbon resins. A typical composition is shown in Table 17, formulation no. 1. The formulation for an adhesive used for permanent labels is given in Table 17, formulation no. 2. Pressure-sensitive adhesives for peelable labels have a relatively high cohesive strength and a low peel strength (Table 17, no. 3). Labels that are applied under cold conditions must have a low T_g, and formulations usually contain low amounts of high-melting resin but large proportions of plasticizer (Table 17, no. 4).

Table 17. Pressure-sensitive adhesive formulations and properties

Composition, property	Formulation number			
	1	2	3	4
Parts by weight				
SIS polymer *	100	100	100	100
Piccotac 95	90		25	
Escorez 1310		120		90
Hercures C 10	10	50	125	
Wingtack 10				40
Irganox 1010	1			3
ZDBC		3	3	
Properties				
Melt viscosity (175 °C), Pa · s		20		60
Rolling ball tack, cm	2	3	3	1
Probe tack, N	10	>10		10
Loop tack, N/25 mm	25	30	8	27
180° peel adhesion, N/25 mm	17	15	5	16
SAFT, °C	115	100	80	100
T_g (Fox), °C	–20	–19	–33	–29

* Kraton D-1107 from Shell Chemicals.

In outdoor applications where the adhesive must withstand UV light, saturated block copolymers are employed. Pressure-sensitive adhesives with very high temperature resistance can be obtained by UV or electron beam curing (⟶ Radiation Chemistry). A special polymer developed by Shell for this application is Kraton D-1320X. The shear adhesion failure temperature (SAFT) of a pressure-sensitive adhesive based on this polymer is increased from 80 °C to ca. 140 °C by either UV or electron beam curing [178].

Contact adhesives are applied as liquid onto one or two substrates; after a certain time to allow partial evaporation of the solvent (open time) the coated substrates are pressed together to form a strong bond. SBCs are ideal for these types of adhesives because they offer high cohesive strength without vulcanization and are readily soluble in many solvents. For some applications, the contact adhesive is allowed to dry completely to a tack-free nonblocking adhesive film. Exposure to heat reactivates the adhesive, and a firm bond forms under pressure. Large surfaces are commonly coated by spraying with a uniform adhesive layer (e.g., on flexible foams used in the furniture and automotive industries). Because of their molecular structures, SBC-based adhesives combine high strength with low viscosity and ease of spraying. A variety of melt adhesives are used in the construction of hygienic products such as disposable diapers, sanitary napkins, etc. These adhesives can be applied by spray or slot-die coating techniques. Selection of the styrenic block copolymer depends on the properties required and the application technique used.

Sealants. Styrenic block copolymers with *saturated rubber mid block* are ideally suited as base polymers for both hot-melt and solvent-based sealants that require resistance to long-term degradation and weathering [179], [180].

Coatings. In several applications, SBCs are used to formulate *protective coatings* based on solvent systems, but the trend is toward solvent-free systems, for example, in chemical milling (⟶ Imaging Technology – Photochemical Machining).

5.4.2. *Modification of Bitumen*

The addition of SBCs to bitumen increases its upper and decreases its lower operating service temperature, which results in substantial performance improvement in applications such as roofing felts, waterproofing membranes, and asphalt pavement. Selective development of SBS polymers has enabled bitumen to maintain its strength and integrity at elevated temperature without flowing and to remain flexible and ductile at temperatures well below the freezing point [181]. Moreover, SBCs greatly enhance the lifetime of bitumen-based products (e.g., roofing felts). This is achieved by development of a three-dimensional network. To optimize the effect of SBS addition, bitumen with the right degree of compatibility must be selected to ensure effective formation of the elastomer network. For roofing felts, SBS addition levels are commonly between 8 and 12 %. Blending into the bitumen is carried out at 170 – 190 °C. For fast dissolution, especially when low-shear paddle mixers are employed, SBS polymer is ground to a powder with particle size of ca. 1000 µm. For most applications, high-shear mixers are used with regular polymer crumb [182], [183].

The use of polymer-modified bitumen for asphalt pavements is growing rapidly and offers large potentials for SBC in the future. The SBS addition levels for road applications are lower than for roofing: typically 3 – 7 %. In addition to improved deformation and flexibility, laboratory data clearly demonstrate that the durability of asphalt roads is increased by addition of SBS polymer [184]. Many test roads have been installed that are being monitored to gain experience and evidence for performance under practical conditions. For documentation of the most important practical experiences so far, see [185-191].

5.4.3. *Compounded Products*

As mentioned earlier, end use parts that are made via established plastic processing techniques do not employ pure SBC polymers but rather — like traditional vulcanized rubber — compounded and end use-tailored compositions. Depending on their requirements, such compounds have hardnesses in the range from 35 Shore A to as high as 55 Shore D. The actual SBC content can be as low as 15 wt % but is usually between 25 and 40 wt %. SIS block copolymers are rarely used as base polymer for compounds. The main classes of compounding ingredients include the following:

Mineral oil is used as plasticizer to increase flow and reduce cost. Paraffinic or naphthenic oils are used that compatibilize with the rubber mid block. Aromatic oils or polar plasticizers such as dioctyl phthalate (DOP) are avoided because they strongly plasticize the end blocks and reduce strength.

Polystyrenes are important compounding ingredients to increase hardness, strength, and wear resistance, particularly in SBS polymers. Both clear and high-impact polystyrenes are used.

Polyolefins, particularly polypropylene, are important ingredients in S – E/B – S and S – E/P – S block copolymers.

Fillers, mainly cheapeners and nonreinforcing types, are used to decrease cost and increase hardness. They also increase density, which is exploited in formulations for sound-deadening applications.

Additives such as antioxidants, UV stabilizers, and antiozonants, as well as additives that induce special effects, can be used. Examples are pigments, surface-active slip agents, blowing agents, and antistatics.

Compounding is carried out at elevated temperature and at shear conditions that result in a fine, stable dispersion of the various ingredients. Compounding equipment originates from both the rubber industry with two-roll mills, Banbury-type internal mixers, and Farrel continuous mixers, and the plastic processing industry with mainly single- and double-screw compounding extruders. Modern compounding plants are fully instrumented robots in which a twin-screw corotating compounder with a length : diameter ratio (L : D) up to 42 is the key unit. Screw types with feeding, kneading, homogenization, and venting sections are designed for the specific compounding operation. Ingredients are added either as preblended mixture or via separate weigh feeders and oil injection pumps. The granulating section at the end of the extruder is important because it is responsible for formation of the homogeneous, free-flowing pellets required in subsequent plastics conversion equipment. Depending on line capacity and product characteristics (i.e., soft or hard), die plates with knife cutters, water-ring pelletizers, underwater pelletizers, or a strand die with water cooling bath and rotating knife granulator, can be used.

Detailed compositions and selection of base polymers have been published by SBC suppliers in the form of technical bulletins or formulating guides. Compounding know-how and formulations based on SEBS or SEPS polymers is rather specific [192]. Table 18 lists properties of typical SBC-based compounds tailored for selected end uses. The largest volume of SBCs in compounded form is consumed by the footwear industry. Here, SBS-based compounds are established mainly as high-quality soling materials. An important aspect is the multitude of colors that can be combined in shoe design. Compounding aims to optimize the end products with respect to melt flow rate (for ease of processing), high abrasion resistance (for durability), and good resistance to repeated flexing. Hardness for most typical soling materials falls within the Shore A 55 to 75 range.

Table 18. Properties of compounded SBCs

Property	Reference no.*, base SBC polymer				
	1 SEBS	2 SBS	3 SEBS	4 SEBS	5 SEBS
MFR (E or G), dg/min	50 G	35 E		3 G	17 G
Density, g/mL	0.91	1.0	1.94	0.91	1.01
Hardness, Shore A/D	45 A	65 A	36 D	60 A	82 A
Tensile strength, MPa	5.0	7.6	4.8	7.5	17.3
Elongation, %	610	650	250	480	670
DIN abrasion, mm^3		180			
Haze, %	7				
UL 62**, Class 34 rating, °C					105
Compression set 22 h at 100 °C, %				42	

* 1 = Transparent medical molding; 2 = soling material; 3 = automotive sound deadening; 4 = high-temperature performance; 5 = electrical wire insulation.
** UL = Underwriters Laboratories code.

Compounds designed for use in medical devices and articles intended for food contact are also important outlets and take advantage of the high purity of polymers made by anionic polymerization. SBS- and SEBS-based compounds which meet the requirements set in the United States by the FDA and in Germany by the BGA, are used for syringes, tubing, bottle nipples, stoppers, i.v. bag systems, etc. Contact of SBS and SEBS with fatty food is restricted.

Thermoplastic rubber products can be recycled, and scrap, trimmings, and off-specification parts can to a certain extent be reintroduced in the manufacturing stream.

5.4.4. *Modification of Plastics*

Most plastics are brittle and lack impact resistance, particularly below 0 °C. Rubber is added to achieve

1. Increased impact strength (Izod, Charpy, FWIS, Dart, etc.) especially at low temperature
2. Increased environmental stress corrosion resistance

The addition of a rubber to a thermoplastic also changes properties such as stiffness, tensile strength, yield modulus, surface friction, softening point, clarity, electrical properties, vapor transmission, and gloss.

With many plastics, an impact resistance improving rubber fraction can be incorporated in the polymerization reaction. Examples are propene – ethylene copolymers, high-impact polystyrene (HIPS), and ABS. Nevertheless, styrenic block copolymers are used as melt-blended component in plastics that do not allow reactor modification or to obtain impact performance in excess of that achievable in the reactor. Polystyrenes are often compounded with flame-retarding additives and then require the addition of SBS polymer to compensate for the loss in impact strength due to addition of the flame retardant [193]. Engineering plastics require processing temperatures beyond the operating range of SBS. Here, saturated SEBS and SEPS block polymers are used, especially as modifiers in polycarbonate (PC) and poly(phenylene ether) (PPE). The more polar engineering plastics such as polyamides and polyesters can be adequately impact modified by SEBS or SEPS functionalized with maleic anhydride [194].

Styrenic block copolymers are also capable of compatibilizing different and dissimilar plastics into useful and workable blends. Addition of SBS or SEBS to a blend of polystyrene and polypropylene or polyethylene, for example, creates a ductile material that can be extruded or molded without the delamination or phase separation ecountered in the absence of block polymer addition [195]. Similar results are obtained when SEBS is used in blends with polycarbonate and polypropylene, as well as polyethylene – poly(ethylene terephthalate) (PETP) blends. Functionalized block polymers are good compatibilizers for, e.g., polyamides and polypropylenes. SBCs can also be used in thermoset resin systems, particularly as additives for sheet molding or bulk molding processes with unsaturated polyesters. Moldings modified with specifically developed SBS or SEBS polymers to control shrinkage show high-quality surface appearance [196], [197].

The use of SBCs as compatibilizers for mixed plastic streams is becoming more and more important with respect to waste management and upgrading and compatibilization of used plastics [198].

[Top of Page]

References

1. Uniroyal, US 3 758 643, 1979 (W. K. Fisher).
2. Uniroyal, US 3 806 558, 1974 (W. K. Fisher).
3. Exxon, US 3 037 954, 1962 (A. M. Gessler, W. H. Haslett).
4. Hercules, US 3 262 992, 1966 (R. Holzer, O. Taunus, K. Mehnert).
5. Monsanto, US 4 130 535, 1978 (A. Y. Coran, B. Das, R. P. Patel).
6. Monsanto, US 4 311 628, 1982 (S. Abdou-Sabet, M. A. Fath).
7. SRI Int. Report no. 104 A, Thermoplastic Elastomers, June 1993.
8. P. Galli, J. C. Haylock: “Polyolefins VII International Conference,” Society of Plastics Engineers, Houston, Texas, Feb. 24 – 27, 1991.
9. J. Yin *et al.*, *Sci. Sin. Ser. B (Engl. Ed.)* **29** (1986) no. 12, 1233.
10. S. Abdou-Sabet, R. P. Patel, *Rubber Chem. Technol.* **64** (1991) 769. Links
11. S. Onogi, T. Asada, A. Tanaka, *J. Pol. Sci. Part A-2* **7** (1969) 171.
12. S. Onogi, T. Asada, *Prog. Polym. Sci. Jpn.*, **2** (1971) 261.
13. D. M. Lohse, *Annu. Tech. Conf. Soc. Plast. Eng.* **43rd** (1985) 301.
14. D. M. Lohse, S. Datta, E. N. Kresge, *Macromolecules* **24** (1991) 561. Links
15. B. F. Goodrich, US 4 036 912, 1977 (J. Sticharczuk).
16. A. Y. Coran, R. P. Patel, D. Williams-Headd, *Rubber Chem. Technol.* **58** (1985) 1014. Links
17. Union Carbide; US 4,803,244, 1989 (J. Umpleby).
18. Advanced Elastomer Systems; US 5,672,660, 1997 (R. E. Medsker, R. Patel).
19. Continental; EP 0 510 559 B1, 1992 (S. Lüpfert, R. Anderlik, H.-G. Fritz, F. Röthemeyer).
20. Bernhard Rustige; EP 0 913 427 A1, 1997 (C.-P. Kirchner, J. Pac, H.-G. Fritz).

21. Chemplast Marketing Services; WO 00/68287, 1999 (C.-P. Kirchner, Q. Cai, H.-G. Fritz).
22. Himont, US 5 196 462, 1993 (D. A. Berta).
23. Mitsui, US 4 212 787, 1980 (A. Matsuda, S. Shimizu, S. Abe).
24. C. P. Rader, S. Abdou-Sabet in S. K. De, A. K. Bhowmick (eds.): *Thermoplastic Elastomers from Rubber-Plastic Blends*, Ellis Horwood, Chichester, England, 1990, p. 159.
25. D. J. Elliott in S. K. De, A. K. Bhowmick (eds.): *Thermoplastic Elastomers from Rubber-Plastic Blends*, Ellis Horwood, Chichester, England, 1990, p. 102.
26. N. R. Choudhury, P. P. De, A. K. Bhowmick in S. K. De, A. K. Bhowmick (eds.): *Thermoplastic Elastomers from Rubber-Plastic Blends*, Ellis Horwood, Chichester, England, 1990, p. 71.
27. R. C. Puydak, D. R. Hazelton, *Plast. Eng.* **44** (1988) no. 9, 38.
28. W. A. Zigman, *Adv. Chem. Ser.* **43** (1964) 1.
29. S. Wu: *Polymer Interface Adhesion*, Marcel Dekker, New York 1982.
30. A. Y. Coran, R. P. Patel, *Rubber Chem. Technol.* **56** (1983) 1045. Links
31. S. Abdou-Sabet, Y. L. Wang, E. F. Chu, *Rubber Plast. News* 1985, Nov. 4, 20.
32. Monsanto, US 4 654 402, 1987 (R. P. Patel).
33. L. A. Goettler, J. R. Richwine, F. J. Wille, *Rubber Chem. Technol.* **55** (1982) 1448. Links
34. G. Oertel (ed.): *Polyurethane Handbook*, 2nd ed., Hanser Publ., München – Wien – New York 1993.
35. C. S. Schollenberger in A. K. Bhowmick, H. L. Stephens (eds.): *Handbook of Elastomers*, Marcel Dekker, New York – Basel 1988.
36. E. C. Ma in B. M. Walker, C. P. Rader (eds.): *Handbook of Thermoplastic Elastomers*, 2nd ed., Van Nostrand Reinhold, New York 1988.
37. W. Meckel, W. Goyert, W. Wieder in G. Holden, N. R. Legge, R. P. Quirk, H. E. Schroeder (eds.): *Thermoplastic Elastomers*, 2nd edition, Hanser Verlag, München 1996..
38. Upjohn, US 3 899 467, 1974 (H. W. Bonk, T. M. Shah).
39. A. Singh, *Advances in Urethane Science and Technology* **13** (1995) 112 – 139.
40. H. Hespe *et al.*, *J. Appl. Polym. Sci.* **44** (1992) 2029 –2035. Links
41. J. H. Saunders, K. C. Frisch: *Polyurethanes*, part I: “Chemistry,” Interscience, New York 1962.
42. Bayer, EP 4393, 1978 (B. Quiring, H. G. Niederdellmann, W. Goyert, H. Wagner).
43. P. J. Flory: *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, N.Y., 1953.
44. W. Neumann, P. Fischer, *Angew. Chem.* **24** (1962) 806.
45. F. Gugumus in H. Zwiefel (ed.): *Plastics Additives Handbook*, 5th ed., Hanser Verlag, München 2001.
46. BASF, EP 0 134 455, 1984 (G. Zeitler, G. Bittner, K. Faehndrich, H. M. Rombrecht).
47. BASF Schwarzheide, DD 300 900, 1987 – 1992 (R. Krech *et al.*).
48. Elastogran, DE 41 12 805, 1991 (A. Chlosta, F. Lehrich, J. Sadlowski, L. Thil).
49. U. Barth in Verein Dt. Ingenieure (eds.): *Polymerreaktionen und reaktives Aufbereiten in kontinuierlichen Maschinen*, VDI-Verlag, Düsseldorf 1988.
50. C. S. Schollenberger, K. Dinbergs, *J. Elastoplast.* **5** (1973) 222; **7** (1975) 65.
51. H. J. Maaß *et al.*, *Plaste Kautsch.* **34** (1987) 251 –254.
52. J. A. Miller *et al.*, *Macromolecules* **18** (1985) 32 – 44. Links
53. S. Abouzaar, G. L. Wilkes, *J. Appl. Polym. Sci.* **29** (1984) 2695 – 2711. Links
54. ICI, GB 1 057 018, 1964 (J. P. Brown, G. Trappe).
55. BASF, EP 922552, 1998 (P. Bartholomaus, H. Dauns, T. Friedl, G. Scholz, H. Loock, W. Lukat).
56. Mobay, US 3 233 025, 1966 (B. F. Frye, K. A. Piggot, J. H. Saunders).
57. A. Bouilloux, C. W. Macosko, T. Kotnour, *Ind. Eng. Chem. Prod. Res. Dev.* **30** (1991) 2431 – 2436.
58. Bayer, DE 4217367, 1993 (F. Mueller, W. Braeuer, K. Ott, H. G. Hoppe).
59. Bayer, DE 4406948, 1995 (W. Braeuer, F. Mueller, H. Heidingsfeld, B. Schulte, J. Winkler).
60. Bayer, DE 2 854 406, 1978 (W. Goyert *et al.*).
61. C. Li, J. G. Homan, R. A. Phillips, S. L. Cooper in K. C. Frisch, Jr. (ed.): *Recent Developments in Polyurethanes and Interpenetrating Polymer Networks*, Technomic, Lancaster – Basel 1988.
62. Z. Petrovic, J. Ferguson, *Prog. Polym. Sci.* **16** (1991) 695 – 836. Links
63. C. G. Seefried, Jr., J. V. Koleske, F. E. Critchfield, *J. Appl. Polym. Sci.* **19** (1975) 2493 – 2502, 2503 – 2513, 3185 – 3191. Links
64. W. P. Yang, C. W. Macosko, S. T. Wellinghoff, *Polymer* **27** (1986) 1235 – 1240. Links
65. L. Born, H. Hespe, J. Crone, K. H. Wolf, *Colloid. Polym. Sci.* **260** (1982) 819 – 828. Links
66. G. Pohl, D. Joel, H. Goering, H. E. Carius, *Plaste Kautsch.* **40** (1993) 357 – 362.
67. *Houben-Weyl*, E 20, 534.
68. J. Foks, I. Naumann, G. H. Michler, *Angew. Makromol. Chem.* **189** (1991) 63 – 76.
69. E. Tocha, H. Janik, M. Debowski, G.J. Vancso, *J. Macromol. Sci. B* **41** (2002) 1291 – 1304.
70. S. Hwang, D. J. Hemker, S. L. Cooper, *Macromolecules* **17** (1984) 307 – 315. Links
71. L. M. Leung, J. T. Koberstein, *Macromolecules* **19** (1986) 706 – 713. Links
72. J. W. C. Van Bogart, P. E. Gibson, S. L. Cooper, *J. Polym. Sci. Polym. Phys. Ed.* **21** (1983) 65 – 95.

73. C. P. Christenson *et al.*, *J. Polym. Sci. Polym. Phys. Ed.* **24** (1986) 1401 – 1439.
74. C. D. Eisenbach, T. Heinemann, A. Ribbe, E. Stadler, *Angew. Makromol. Chem.* **202/203** (1992) 221 –241.
75. Dow Chemical, EP 80 031, 1983 (D. J. Goldwasser, K. Onder).
76. B. F. Goodrich, DE 1 720 843, 1967 (E. G. Kolycheck).
77. Bayer, DE 1 940 181, 1969 (E. Meisert, A. Awater, C. Muehlhausen, U. J. Doebereiner).
78. W. Goyert, H. Hespe, *Kunststoffe* **68** (1978) 2 – 8.
79. E. C. Ma, *Rubber World* **199** (1989) 30 – 35.
80. A. T. Chen *et al.*, *Elastomerics* **122** (1990) no. 9, 19 –22.
81. H. G. Hoppe, *Plastverarbeiter* **44** (1993) no. 9, 75 –80;no. 10, 40 – 44.
82. S. Gogolewski, *Colloid Polym. Sci.* **267** (1989) 757 –785. Links
83. R. Zdrahala, I. Zdrahala, *J. Biomat. Appl.* **14** (1999) no. 1, 67 – 90.
84. DuPont, US 3 651 014, 1972 (W. K. Witsiepe).
85. DuPont, US 3 763 109, 1973 (W. K. Witsiepe).
86. DuPont, US 3 766 146, 1973 (W. K. Witsiepe).
87. L. H. Buxbaum, *J. Appl. Polym. Sci.* **35** (1979) 59 –66.
88. DuPont, US 3 801 547, 1974 (G. K. Hoeschele).
89. BASF, US 4 056 514, 1977 (H. Strehler).
90. General Electric, US 4 732 948, 1988 (R. J. McCready, J. A. Tyrell).
91. Kuraray, JP 61 81 419, 1985 (M. Ishiguro, H. Hira).
92. DuPont, US 5 004 748, 1991 (G. Huynh-Ba).
93. DuPont US 3 726 569, 1973 (G. K. Hoeschele).
94. DuPont: “Hytrel,” Product Bulletin H-33430–1, 1994, p. 6.
95. G. K. Hoeschele, W. K. Witsiepe, *Angew. Makromol. Chem.* **29/30** (1973) 267.
96. J. R. Wolfe, Jr., *ACS Adv. Chem. Ser.* **176** (1979) 129.
97. Eastman Kodak, US 4 256 860, 1981 (B. Davis, R. B. Barbee, H. R. Musser).
98. General Electric, US 4 711 947, 1987 (R. J. McCready, J. A. Tyrell).
99. DuPont, US 3 775 374, 1973 (J. R. Wolfe Jr.,
100. DuPont, US 3 784 540, 1974 (G. K. Hoeschele).
101. J. R. Wolfe, Jr., *Rubber Chem. Technol.* **50** (1977) 688. Links
102. Texaco Chemical Co.: Polyoxyalkyleneamines, product bulletin, 1991.
103. General Electric, US 4 544 734, 1985; US 4 556 705, 1985 (R. J. McCready).
104. E. C. Leonard, *Dimer Acids*, 1975, 1.
105. DuPont, US 3 954 689, 1976 (G. K. Hoeschele).
106. General Electric, US 4 594 377, 1986 (R. J. McCready).
107. R. W. M. van Berkel, S. A. G. de Graaf, F. J. Huntjens, C. M. F. Vrouenraets: Development Series, Developments in Block Copolymers-1, p. 261.
108. Teijin, BE 834 004, 1975.
109. Goodyear, US 3 446 778, 1969 (R. Ca. Waller, M. H. Keck).
110. R. J. Cella, *J. Polym. Sci. Symp.* **42** (1973) no. 2, 727.
111. R. J. Cella, Polymer Symposia No. 42, Helsinki 1972.
112. S. C. Wells in B. Walker (ed.): *Handbook of Thermoplastic Elastomers*, "Chap. 4".
113. M. Brown, *Rubber Ind. (London)* **9** (1975) 102.
114. DuPont, US 3 718 715, 1973 (R. W. Crawford, W. K. Witsiepe).
115. DuPont, US 3 917 743, 1975 (H. E. Schroeder, J. R. Wolf, Jr.
116. DuPont, US 3 963 802, 1976 (C. Shih).
117. DuPont, US 4 010 222, 1977 (C. Shih).
118. General Electric, US 4 939 205, 1989 (Nan-I Liu).
119. General Electric, US 4 814 380, 1990 (Nan-I Liu, R. J. McCready).
120. M. Veldstra: *Thermoplastic Elastomers-II*, Rapra Technology Ltd., London 1989.
121. H. Reinhardt, M. Negri, *Thermoplastic Elastomers-II*, Rapra Technology Ltd., London 1989.
122. Teijin, JP 50–35 623, 1976.
123. AKZO, US 4 493 870, 1985 (C. M. F. Vrouenraets).
124. M. H. Horn, K. P. Schodt, G. J. Ostapchenko, *Annu.. Tech. Conf. Soc. Plast. Eng.* 1992.
125. M. Tessier, E. Marechal, *J. Polym. Sci. A1*, **27** (1989) 539.
126. M. Tessier, E. Marechal, *Eur. Polym. J.* **20** (1984) 281. Links
127. M. Tessier, E. Marechal, *Eur. Polym. J.* **22** (1986) 877. Links
128. Bayer, EP 83 434, 1983 (U. Grigo, K.-H. Köhler, R. Binsack, L. Morbitzer, J. Merten, L. Trabert, W. Heitz).
129. Daicel-Degussa, EP-PS 73 838, 1983 (K. Arita, H. Suzuki).
130. Bayer, EP-A 244 601, 1987 (K.-H. Köhler, K. Reinking, G. Weber, R. Prinz).

131. Bayer, DE-OS 3105365, 1982 (U. Grigo, K.-H. Köhler, R. Binsack, L. Morbitzer, J. Merten, L. Bottenbruch, W. Heitz).
132. A. Koichi, *J. Polym. Sci. A* **25** (1987) 1591.
133. Ube, EP-A 1235165, 2002 (H. Okushita, T. Muramatsu).
134. G. Deleens, P. Foy, E. Marechal, *Eur. Polym. J.* **13** (1977) 337. Links
135. G. Deleens, P. Foy, E. Marechal, *Eur. Polym. J.* **13** (1977) 343. Links
136. G. Deleens, P. Foy, E. Marechal, *Eur. Polym. J.* **13** (1977) 353. Links
137. G. Deleens, *Ph.D. Thesis*, University Rouen 1975.
138. S. Mumcu, K. Burzin, R. Feldmann, R. Feinauer, *Angew. Makromol. Chem.* **74** (1978) 49.
139. Hodogaya, JP 63 105 029-A, 1988 (K. Ogabe, S. Takeuchi).
140. Asahi, EP-A 221 188, 1987 (Y. Suzuki, M. Nakamura, A. Aochiama).
141. Degussa, DE-OS 2 936 976, 1981 (S. Mumcu).
142. Degussa, DE-PS 2 936 977, 1981 (S. Mumcu).
143. Degussa, DE-OS 2 932 234, 1979 (H.-J. Panoch, S. Mumcu).
144. Degussa, DE-PS 2 712 987, 1978 (K. Burzin, S. Mumcu, R. Feldmann, H. Jadamus, R. Feinauer).
145. Atofina, DE-PS 2 523 991, 1975 (P. Foy, C. Jungblut, G. Deleens).
146. Atofina, DE-PS 2 837 687, 1979 (G. Deleens, J. Ferlampin, M. Gonnet).
147. Ems-Chemie, DE-PS 3 428 404, 1985 (H. J. Liedlof).
148. Atofina, DE-OS 2 632 120, 1977 (P. Foy, R. Kern).
149. Toray, EP-AS 163 902, 1985 (C. Tanaka, M. Kondou, Y. Yamamoto).
150. Toray, DE-OS 3 145 998, 1982 (C. Tanaka, A. Chita, S. Nakashima, N. Shimobu).
151. Toray, EP 95 893, 1983 (C. Tanaka, S. Nakashima, M. Kon Dou).
152. Atofina, DE-OS 2 856 787, 1979 (G. Deleens, J. Ferlampin, C. Prulain).
153. F. Widmer, *Adv. Chem. Ser.* **128** (1973) 51 – 67.
154. B. Loretan, E. Heimgartner, *Chem. Ing. Tech. (1979) MS* **467/77**.
155. Ems-Chemie, DE-PS 3 006 961, 1980 (W. Isler, E. Schmid).
156. Rhone-Poulenc, EP-PS 130 927, 1986 (I. Coquard; J. Goletto).
157. Rhone-Poulenc, EP-PS 187 607, 1986 (I. Coquard; J. Goletto).
158. Rhone-Poulenc, EP-PS 201 434, 1986 (I. Coquard; J. Goletto).
159. H. G. Elias, R. Vohwinkel: *Neue Polymere Werkstoffe für die industrielle Anwendung II*, C. Hanser Verlag, München-Wien 1983, p.125 – 134.
160. Ube, JP 59-133224, (1983) (H. Okamoto, Y. Okushita).
161. Ube, JP 59-131628, (1983) (H. Okamoto, Y. Okushita).
162. Ube, JP 59-193923, (1983) (H. Okamoto, Y. Okushita).
163. Huntsman, EP-PS 432 943, 1991 (G. P. Speranza, W.-Y. Su).
164. Huntsman, EP-A 449 419, 1991 (G. P. Speranza, W.-Y. Su).
165. R. J. Gaymans, P. Schwering, J. L. de Haan, *Polymer* **30** (1989) 974. Links
166. DSM, EP-A 360 311, 1990 (R. J. Gaymans, R. Roerdink, P. J. Schwering ▪ Walch).
167. E. Bornschlegl, G. Goldbach, K. Meyer, *Progr. Colloid Polym. Sci.* **71** (1985) 119 – 124.
168. J. Lohmar, K. Meyer, G. Goldbach, *Makromol. Chem.* **189** (1988) 2053 – 2065.
169. G. Goldbach, M. Kita, K. Meyer, K. P. Richter, *Progr. Colloid Polym. Sci.* **72** (1986) 83 – 96.
170. N. R. Legge: "Thermoplastic Elastomers Based on Three-Block Polymers," *Chemtech.* **13** (1983) 630.
171. Shell Oil Co., US 3 149 182, 1964 (L. M. Porter).
172. R. J. Ceresa: *Block and Graft Copolymers*, Butterworth, Washington, D.C., 1962.
173. P. Dreyfuss, L. J. Fetters, D. R. Hansen, *Rubber Chem. Technol.* **53** (1980) 728. Links
174. E. V. Gouinlock, R. S. Porter, *Polym. Eng. Sci.* **17** (1977) 535.
175. B. M. Walker, C. P. Rader (eds.): Handbook of Thermoplastic Elastomers, 2nd ed., Van Nostrand Reinhold, New York 1987.
176. A. D. Thorn: *Thermoplastic Rubbers – A Review of Current Information*, RAPRA, Shawsbury, UK, 1980.
177. D. J. St. Clair, *Rubber Chem. Technol.* **55** (1982) 208.
178. J. R. Erickson, *Rubber Plast. News* 1985, Sept. 9, 16.
179. G. Holden, S. S. Chin, Paper to the Adhesives and Sealants Conference, Washington, D.C., March 1986.
180. G. Holden, Paper to the Adhesives and Sealants Council Seminar, Chicago, Il., Oct. 1982.
181. E. J. Van Beem, P. Brasser, *J. Inst. Pet.* **59** (1973) 91.
182. Shell Int. Chem. Co., Bulletin TR-18, London 1984.
183. *The Shell Bitumen Handbook*, Shell Bitumen UK Ltd., London 1990.
184. J. H. Collins, W. J. Mikols, *60th Meeting of Asphalt Paving Technologists*, San Antonio, Tx., Feb. 1985.
185. D. M. Colwill, M. E. Daines: "Progress in the Trials of Pervious Macadam," *Highways*, Jan. 1989, p. 15.
186. A. Dinnen: "Bitumen Thermoplastic Rubber Blends in Road Applications," 3rd Eurobitumen Symposium, The Hague, Sep. 1985.
187. M. J. W. Downes, R. C. Koole, E. A. Mulder, W. E. Graham: "Some Proven New Binders and Their Cost

Effectiveness," *7th AAPA Int. Asphalt Conf.*, Brisbane, Aug. 1988.
188. K. H. Kolb, T. Pallay, J. P. Serfass, O. Ruud: "Asphalt for Better Roads," Symposium, E.A.P.A., Copenhagen 1986.
189. L. Laitinen: "Rubberised Bitumen in Road Surfacing Applications," *Asfaltti* **43** (1988) 18.
190. J. P. Marchand: "15 Years of Experience in Thick Bitumen Pavements," Petrol. Inf. **1987**, Nov., 62.
191. S. H. Carpenter, T. Van Dam: "Modified and Unmodified Asphalt Concrete Mixtures," *Transp. Res. Rec.* **1115** (1987) p. 62 – 74.
192. H. F. Vermeire, *Kautsch. Gummi Kunstst.* **43** (1990) 11.
193. A. L. Bull, G. Holden, *J. Elastomers Plast.* **9** (1981) 281.
194. Shell Int. Res. Co., EP 0 215 501, 1982 (W. P. Gergen, R. G. Lutz, M. K. Martin).
195. D. W. Bartlett, D. R. Paul, J. W. Barlow, *Mod. Plast.* **58** (1981) no. 12, 60. Links
196. C. L. Willis, W. M. Halper, D. L. Handlin, Jr., *Polym. Plast. Technol. Eng.* **23** (1984) no. 2, 207.
197. Shell Int. Chem. Co., Technical Bulletin TR.3.6, London 1991.
198. J. Schneider, Paper presented at Recycle '92, Davos, Apr. 1992.

[Top of Page]

Abbreviations for polymers and monomers:

2G: ethylene glycol

2GT: dimethylene terephthalate, ethylene terephthalate

3G: 1,3-propanediol

4G: 1,4-butanediol

4GT: tetramethylene terephthalate, butylene terephthalate

5G: 1,5-pentanediol

6G: 1,6-hexanediol

10G: 1,10-decanediol

ABS: acrylonitrile – butadiene – styrene copolymer

BR: butadiene rubber

DOP: dioctyl phthalate

DVA: dynamically vulcanized alloys

EOPPG: poly(propylene glycol) capped with ethylene oxide

EPDM: ethylene – propene – diene terpolymer

EPM: ethylene – propene copolymer

EPR: ethylene – propene rubber

EVA: ethylene – vinyl acetate copolymer

H_{12}-MDI: 4,4′-dicyclohexylmethane diisocyanate

HDI: hexamethylene diisocyanate

HIPS: high-impact polystyrene

IR: isoprene rubber

MDI: 4,4′-methylene diphenyl diisocyanate

NDI: 1,5-naphthalene diisocyanate

NR: natural rubber

PAE: thermoplastic polyamide elastomer

PB: polybutadiene

PC: polycarbonate

PEO: poly(ethylene glycol); poly(ethylene oxide)

PETP: poly(ethylene terephthalate)

PP: polypropylene

PPE: poly(phenylene ether)

PS: polystyrene

PTMO: poly(tetramethylene oxide)

PVC: poly(vinyl chloride)

RTPO: reactor thermoplastic polyolefins

S – E/B – S, SEBS: styrene – ethylene/butene – styrene copolymer

S – E/P – S, SEPS: styrene – ethylene/propene – styrene copolymer

SBC: styrenic block copolymers

SBS: styrene – butadiene – styrene block copolymer

SIS: styrene – isoprene – styrene block copolymer

T: terephthalic acid

TODI: 3,3′-tolidine-4,4′-diisocyanate

TPE: thermoplastic elastomer

TPNR: thermoplastic natural rubber

TPO: thermoplastic polyolefin elastomer

TPR: thermoplastic rubber

TPU: thermoplastic polyurethane elastomer

TPV: thermoplastic vulcanizates

VA: vinyl acetate

Other abbreviations used
BMC: bulk molding compound

MFR: melt flow rate

SAFT: shear adhesion failure temperature

SMC: sheet molding compound

T. S.: tensile strength

U. E.: ultimate elongation

UL: Underwriters Laboratories code

[Top of Page]

Related Articles

A general introduction into Thermoplastic Elastomers is given in ⟶ Rubber, 3. Synthetic – Thermoplastic Elastomers.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Thermosets

WALDEMAR SCHÖNTHALER, Bakelite Gesellschaft mbH, Iserlohn-Letmathe, Federal Republic of Germany

1. **Introduction** 1
2. **Manufacture** 3
2.1. **Fusion Process** 3
2.2. **Liquid Resin Process** 5
2.3. **Other Processes** 5
3. **Commercial Forms and Manufacturers** 5
4. **Storage and Transport** 6
5. **Properties** 7
5.1. **Uncured Molding Compounds** 7
5.2. **Cured Moldings** 10
6. **Processing and Aftertreatment** 13
7. **Uses** 13
8. **Economic Aspects** 13
9. **Environmental Protection and Recycling** 14
10. **References** 15

1. Introduction

Typical of thermosetting plastics (thermosets) is that they are formed in the processing mold from moderately large molecular building blocks (reactive precursors) by chemical reaction. This contrasts with thermoplastics, where the finished high-molecular material is merely physically remelted in the mold. That also explains the main difference between these two kinds of plastics: thermosets no longer melt and are processed at temperatures far below their application temperatures; thermoplastics can be remelted and must therefore always be processed at temperatures far above their temperatures of use.

In German, the thermosets were formerly known as *Pressmassen* (compression-molding compounds) since they were processed by compression molding. Thermoplastics were known as *Spritzgussmassen* (injection molding compounds) since they were processed by injection molding. Since the mid-1960s thermosets have also been processible on injection molding machines, so the term *Formmasse* (molding compound) has been introduced whereby a distinction is made between curable and noncurable molding compounds (thermosets and thermoplastics, respectively).

The classical thermosets, which mainly are discussed here, are based on phenolic, urea, or melamine resins. These were followed by others based on unsaturated polyester, diallyl phthalate, epoxy resin, or silicone resin. Polyurethanes (→ Polyurethanes) and polyimides also can be thermosets. Strictly speaking, even cross-linked thermoplastics can be regarded as thermosets.

A comprehensive survey of the wide field of the thermosets is provided in [1]. Also discussed there are future developments based on nadimide- (polyimide), maleimide- and acetylene-terminated thermosetting polymers.

History [2]. Compression-molding compounds were the first fully synthetic plastics. Resins were manufactured from phenol and formaldehyde according to the patents of L. H. BAEKELAND as early as 1910 and, after mixing with fillers, cured to plastics by applying heat and pressure. Phenolic resin compression molding materials were initially based on phenolic resols (one-step resins), and later (in the 1920s) on novolacs ("two-step resins") (see also → Phenolic Resins). The use of novolacs with curing by hexamethylenetetramine was an important development (fast-curing molding compounds), since it also improved the manufacture and the storability.

In the late 1920s, the urea resin compression molding compounds came onto the market and made white and colored light-resistant compression moldings possible. At the end of the 1930s melamine resin compression molding compounds were introduced, which made good water resistance possible.

Compression-molding compounds based on unsaturated polyester resins (see → Polyester Resins, Unsaturated) and epoxy resins (see → Epoxy Resins) became widespread after World War II, the polyester resin compression-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_665

molding compounds, at first being only wet-compounded. Today, dry, free-flowing polyester resin compression molding compounds also have gained major importance.

Molding compounds are generally unshaped products which can be processed by pressure and temperature to form moldings.

In the thermosetting plastics, because of the chemical curing process, the composition of the molded material differs from that of the molding compound. In the case of thermoplastics, the finished plastic is processed, and only remelting takes place in the mold.

Classification. From the beginning, thermosetting molding compounds were compounded from binders and fillers. This was done to remove the brittleness from the moldings, to economize on resin, and to broaden the range of applications of the materials. An extremely wide variety of materials are used as fillers. The term filler can refer to nonreinforcing and reinforcing fillers.

Fillers can be classified into inorganic and organic, into powder, fibrous, and lamellar types (see Table 1) [3].

These fillers can be combined with various binders. It is also possible to compound combinations of binders with one filler or combinations of fillers with a single binder. Finally, there is also the possibility of combining mixtures of several binders with the mixture of several fillers.

The binder (resin) can be manufactured from diverse molecular building blocks and by various reaction mechanisms (see Table 2) [4]. Today, phenolic resins, urea resins, melamine resins, unsaturated polyester resins, and epoxy resins are widely used. Resins based on diallyl phthalate or silicones are less important.

The curing or cross-linking process in the mold can proceed by other chemical reactions than the manufacture of the resin itself. Polycondensation, polymerization, or polyaddition can be involved. In case of polycondensation, low-molecular substances (e.g., water) are split off, whose vapor pressure at processing temperature must be compensated for.

The properties of the cured molded material are the sum of the properties of the resins and of the fillers. Thus for a given filler, the typical binder properties, and for a given binder typical filler properties are observed. Therefore, resin-determined and filler-determined properties must be differentiated.

Standardization. For the standardization and monitoring of thermosetting molding compounds in Germany, the Technical Association of Manufacturers and Processors of Plastics

Table 1. Fillers for thermoset molding compounds

	Particle form	Example	Advantages			Disadvantages
Inorganic	powder short fibers long fibers platelets	stone powder mineral fiber glass fiber mica	high thermal stability	uniform distribution, high mechanical values	high density	low mechanical values, orientation
Organic	powder short fibers long fibers platelets	wood flour cellulose textile threads macerated fabric	low density	uniform distribution high mechanical values	low thermal stability	low mechanical values orientation

Table 2. Resins for thermoset molding compounds

	Symbol	Resin structural units	Resin-forming reaction	Cross-linking reaction	Reaction product split off on cross-linking
Phenoplastic	PF	phenol and formaldehyde	polycondensation	polycondensation	water
Aminoplastic	UF	urea and formaldehyde	polycondensation	polycondensation	water
	MF	melamine and formaldehyde	polycondensation	polycondensation	water
Unsaturated polyester	UP	unsaturated dibasic acid and diol	polycondensation	polymerization	none
Epoxide	EP	bisphenol and epichlorohydrin	polycondensation	polyaddition with diamine or dibasic acid	none
Diallyl phthalate	DAP	monomeric diallyl phthalate	polymerization	polymerization	none

Molding Compounds (Technische Vereinigung der Hersteller und Verarbeiter von Kunststoff-Formmassen) has been established to promote cooperation between raw material manufacturers and processors. Type tables with minimum requirements for several important properties of the molded materials have been drawn up. These properties are tested on defined test specimens according to DIN specifications. The tables, also specify the binder and filler for each type. The most important DIN standards are DIN 7708 (phenolics and aminoplastics) and DIN 16 911 (polyester resin molding compounds). For epoxy resin molding compounds, DIN 16 912 has been withdrawn owing to the rapid development in this field.

Beyond the type values listed in the DIN specification tables, a further series of properties of the molding compounds is defined. In the case of the phenolics, this is information on the base resin, resin content, and color. For example, a molding material of Type 31 might carry the designation 31-1449, where the first two digits indicate the type according to DIN 7708. The third digit indicates phenolics as base resin, the fourth defines a minimum resin content and the last two the color:

3 = 35 % resin content
4 = 40 % resin content
5 = 45 % resin content
6 = 50 % resin content
7 = 55 % resin content
8 = 60 % resin content
00 – 09 = white, yellow, natural
10 – 19 = light brown to dark brown
20 – 29 = red to mahogany
30 – 35 = green
35 – 39 = blue
40 – 49 = grey to black

An additional definition is sometimes appended to the type designation (e.g., 51.5):

.5 = electrically high-quality
.7 = suitable for tableware for eating and drinking
.8 = content of volatile acids < 0.18 %
.9 = ammonia free

Attempts to develop a system analogous to DIN 7708 for the quality assurance of thermoplastics failed. This was partly due to the difficulty of producing standard test specimens by injection molding. To offer designers the possibilities of comparing the diverse plastics, in Germany the four largest plastics manufacturers, BASF, Bayer, Hoechst, and Hüls, have created the CAMPUS System (Computer-Aided Material Preselection by Uniform Standards), which has been joined by more than 20 international chemical companies. Since then thermoset manufacturers have also decided to use this system on grounds of rationalization, especially since the single-point values originally used for quality assurance are now supplemented by curve diagrams.

The CAMPUS System is offered in the form of diskettes free of charge by the participating firms and can be run on IBM-compatible PCs.

In the course of internationalization, DIN standards have already in many cases been replaced or supplemented by ISO or EN standards [5].

2. Manufacture

Despite the great variety of possible thermosetting molding compounds, only a few processes are used in their manufacture. The binder (resin) must be converted to the liquid state so that good impregnation of the fillers is possible. This is referred to as the fusion process if liquefaction is carried out by melting the resin, and as the liquid resin process if the resin is already liquid at room temperature or is used in solution in water or organic solvents [3].

2.1. Fusion Process

In the fusion process (Fig. 1) a premix is produced from finely ground solid resin, fillers, colorants, lubricants, and release agents, etc. Examples of recipes (in wt %) follow:

Type 31 (PF)	
Phenol novolac	43
Hexamethylenetetramine	7
Wood flour	45
Magnesium oxide	1
Zinc stearate	2
Dyes	2
Type 150 (MF)	
Solid melamine resin	50
Accelerator	0.5
Wood flour	45
Zinc stearate	1.5
Dyes	3

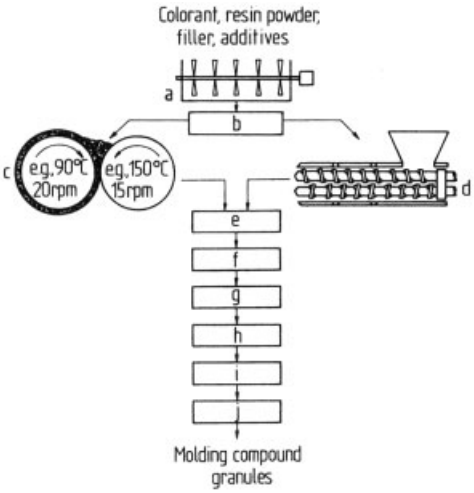

Figure 1. Compounding of thermoset molding compounds by the fusion process
a) Premixer; b) Premix; c) Rolls; d) Screw kneader; e) Cooling; f) Precomminution; g) Grinding; h) Sieving; i) Classifying; j) Mixing

The premix is then compounded discontinuously or continuously on roll mills or extruders to form the thermoset molding compound. The roll mills have two contrarotating steel rolls, up to ca. 3 m long, heated by steam, high-pressure water, oil, or electric (for smaller rolls), and rotating at different speeds. Their roll gap, a few millimeters, is variable. The rolls differ in temperature (e.g., 90 °C for the front and 150 °C for the back roll), so that on feeding the premix a layer (sheet) forms on the (cooler) rolls covering itcompletely. On passage of the sheet through the narrow gap between the rolls, the material is kneaded and mixed as a result of the different roll speeds. In addition, a bank forms on the roll gap rotating contrary to the front roll. On the roll mill, the fillers are impregnated with the binder, the material is homogenized, and in the case of polycondensation compounds (phenolics and aminoplastics) also precross-linked.

The precross-linking is continued until a thermoset molding compound has been formed that the processor can still process in his specific mold with the forming pressure available to him. This enables increased savings of energy and time to be made at the processor's works.

In addition, extensive precondensation also improves the properties of the molded material, such as surface gloss, electrical properties, shrinkage and post-shrinkage behavior. Compounding on roll mills can be carried out discontinuously or continuously. In the discontinuous process, when the sheets have reached the required degree of precondensation they are cut open along the roll and pulled off. In the continuous process, the premix is fed continuously in the middle or on one side of the roll, and the sheet formed is cut off in narrow continuous strips at the roll edge.

In the extrusion process, single- or twin-screw machines or planetary screw extruders are used. In this case the premix is usually processed in large batch premixers.

Processing on the extruders is continuous. In both the roll-mill and extruder processes, the sheet or lumps formed are cooled, giving a grindable product. From this, a granular molding compound of defined grain size is obtained by breaking, milling, and sieving.

For the fusion process, powder or short-fiber fillers such as wood flour, stone powders, and glass fibers, are best suited. Long-fiber fillers are broken down by the high shear forces in the roll gap. Materials manufactured by the hot melt process include phenolic resin molding compounds (PF) of Types 11 (stone powder), 31 (wood flour) according to DIN 7708; the melamine – formaldehyde (MF) and melamine phenol (MP) Types 150 (wood flour), 180 (wood flour), and 155 (stone powder); and the polyester resin Types (UP) 802 and 804 (short glass fibers) according to DIN 16 911. The UP compounds are not precured since hardening by polymerization takes place only in the molding tools.

2.2. Liquid Resin Process

The liquid resin process (Fig. 2) is used mainly for compounding thermosetting molding materials that contain long-fiber or macerate fillers (coarse-structured materials). Kneaders (e.g., sigma kneaders, which bakeries also use) or edge mills are used as mixing units. The compounded materials can be removed in the dough state after mixing and impregnation. Examples are Types 801 and 803 (DIN 16 911), which are compounded with liquid, styrene-containing unsaturated polyester resins and contain long glass fibers.

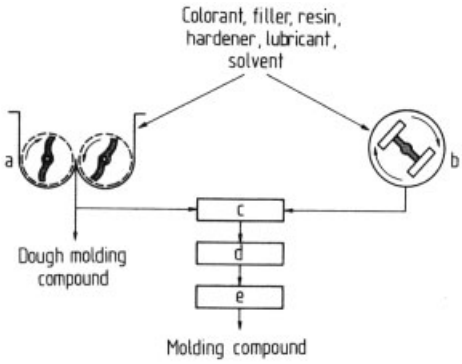

Figure 2. Compounding of thermoset molding compounds by the liquid resin process
a) Kneader; b) Edge mill; c) Drying; d) Cooling; e) Coarse grinding

The compounds can also be dryed and further condensed, as in the case of the phenolic of Type 71 (textile fibers) or 74 (textile macerate) or the aminoplastic Types 153 (textile fibers) and 154 (textile macerate) (DIN 7708).

Discontinuous tray driers or continuous-flow driers can be used for drying and further condensation. By suitable choice of temperature and drying time, the flow – cure behavior of phenolics and aminoplastics can be adapted to the requirements of the processor. When using solvent-containing resins, however, care must be taken that prior to further condensation the solvent can escape from the molding compound particles. If solvent residues are trapped because drying process has not been carried out properly, surface defects can occur on molding.

The liquid resin process is mainly used for the mild impregnation of long-fiber fillers, but it also allows low-resin molding compounds to be produced if the liquid resin is highly diluted by solvent.

2.3. Other Processes

A number of special methods exist. For example, a turbomixer (Fig. 3), which permits very short mixing times, can be used in the liquid resin process. Turbomixers are also suitable for the hot melt process since the large amount of mechanical energy introduced can lead to fusion of the finely ground resin. In addition to impregnation of the filler, the mixture is agglomerated to dust-free granules. As a result, subsequent grinding can be dispensed with, and particularly free-flowing granules are obtained. In addition, recycling of dust and oversize grain to the process, which is otherwise necessary for granule production, is unnecessary.

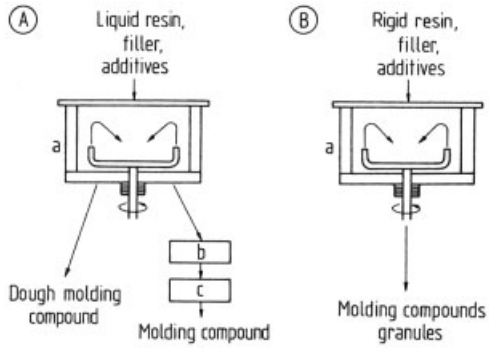

Figure 3. Compounding of thermoset molding compounds in a turbomixer
A) Wet mixing and impregnation; B) Dry mixing and impregnation
a) Turbomixer; b) Drying; c) Coarse grinding

3. Commercial Forms and Manufacturers

Thermoset molding compunds are supplied in diverse forms, from dust-fine material through various grain sizes up to briquetted grades, depending on the mechanical demands made on the molding to be manufactured. Thus, a material compounded with coarse-structured filler cannot be supplied in a fineground form without loss of strength. Fine grinds are therefore useful mainly in recipes containing powdered filler or with short-fiber fillers.

Conventional size reduction by crushers and mills (e.g., hammer mills, pinned disk mills, cutting mills etc.) of the brittle sheets or clods obtained in the production process gives fairly wide grain-size distributions. By sieving out the dust and oversize grain, narrow grain-size ranges can be obtained fixed, but with simultaneous reduction of the production capacity. Oversize grain must be returned to the mills, and the dust component is usually added to the premix and therefore passes once more through the production plant. A reasonable grain size therefore usually comprises the range from 0.2 to 1.0 mm grain diameter. Such a range of grain sizes has a good

flowability and does not tend to demix, even during long-distance transport. Removing the very fine fractions on vibrating sieves is sometimes problematic because the very fine dust adheres strongly to the grains of molding compound, and because in the subsequent mixing processes the brittle granules rapidly form dust by abrasion. By careful control of the process, however, almost dust-free granules can be produced. Although these modern low-pollution granules are more convenient for the processor, the older commercial forms with a high dust content (up to 30 %) gave moldings with better mechanical properties and surface quality.

With coarser fillers, the upper limit of grain size is extended to ca. 8 mm, whereby the flowability deteriorates. If high-strength materials with good flowability are to be produced, special granulation processes are necessary, such as pelletizing or compression granulation. For thermoset molding compounds with high fiber content the latter is generally carried out on extruders, giving "sausage-like" granulates with diameters of 1 – 6 mm, depending on the filler.

Molding compounds with particularly coarse fillers, such as those produced by wet compounding of macerates (e.g., Type 74), are also supplied as briquettes, weighing up to ca. 1 kg. A special granulation method is used for the aminoplastic Types 131 and 152. These cellulose-filled molding compounds are usually produced in undyed form and are subsequently dyed in ball mills. This procedure is useful because of the variety of shades of color required for these molding materials. Very fine dust is formed in the ball mills and must subsequently be recompacted in, e.g., tabletting machines, pelletizers, extruders, or pan granulators. Depending on the method used, compaction may be followed by grinding and sieving.

Manufacturers. An extensive compilation of the manufacturers and processors of thermoset molding compounds appears in the journal *Kunststoffe*, usually in the February issue.

Some important manufacturers of thermoset molding materials in Europe are:

- Bakelite GmbH, Iserlohn Lemathe, Germany
- Phönix Gummiwerke AG, Hamburg, Germany
- Dr. F. Raschig GmbH, Ludwigshafen, Germany
- Resart Ihm AG, Mainz, Germany
- Süd-West Chemie GmbH, Neu-Ulm, Germany
- Perstorp, Sweden
- Vynckier, Belgium

4. Storage and Transport

Since the molding compounds discussed here cure on warming they should preferably be stored in cool rooms. The storage temperature depends on the type of material.

Phenolics and aminoplastics should preferably not be stored above room temperature. Phenolics based on novolac – hexamethylenetetramine (the majority) can be kept in this way for years. The storability of phenolics based on resols, and of aminoplastics is lower. Since these polycondensation molding compounds readily absorb or lose water, depending on the ambient humidity, the relative humidity of the store rooms also plays an important role. A relative humidity of ca. 50 % is recommended.

Packaging of molding compounds in multilayer paper sacks with polyethylene liners protects against the uptake or release of moisture for several months. Changes in humidity in phenolics and aminoplastics affect the flow – cure behavior of the molding compounds, in particular.

Polyester resin molding compounds also should be stored below room temperature. The storage life is then at least 6 months. Modern styrene-free, free-flowing polyester compounds achieve storabilities comparable to those of phenolics.

In earlier times epoxy resin molding compounds were stored in refrigerated rooms, and sometimes shipped in cooled vessels. However, for modern epoxy resin molding compounds this is no longer necessary.

Thermoset molding compounds are increasingly transported and stored in containers. This guarantees the exclusion of moisture and permits processing in closed systems on modern injection molding machines. The molding compound can thus be conveyed via suction-pneumatic or mechanical conveyor systems, without dust problems, into the feed hopper of the injection molding machine.

The containers range in capacity from 500 kg to ca. 2000 kg and can be made of metal, cardboard, or plastic. Cardboard containers are used as disposable packaging. They are less moisture-proof and more easily damaged. Containers made of steel sheet or aluminum have considerable weight when empty, and therefore cause additional costs for return freight. However, they are undoubtedly the safest packaging and protect the contents even under highly unfavorable weather conditions.

5. Properties

5.1. Uncured Molding Compounds

For the uncured molding compound the most important property is the processing behavior. The following properties are tested [4, 6]:

1) Flowability
2) Grain size (or grain-size distribution)
3) Bulk density
4) Filling factor
5) Flow – cure behavior

The *flowability* (DIN 53 492) of granular plastic is usually assessed in funnels of defined shape or with automatic metering devices. A simple method is determination of the angle of repose.

The *grain size* (DIN 53 477) is measured on sieving machines, the residues on the test sieves being weighed. The measured values can be evaluated either directly or by the Rosin – Rammler method, in which the measured values are plotted on a double-logarithmic scale. A straight line is obtained, whose slope allows the uniformity of the grain distribution to be inferred. An advantage of this method is that deviation of the curve from the straight line indicates a fault in the sieving plant.

The *bulk density* (DIN 53 468) is of importance for volumetric metering. First the weight of loose material that can be accommodated in 1000 mL is determined and then divided by the volume to give the bulk density. Since volume metering is often employed in the processing of thermoset molding compounds, the maintenance of the bulk density from delivery to delivery is very important.

The *filling factor* is the ratio of the relative density of the molded material to the bulk density of a molding compound. It plays a role when the transfer chamber of a mold is limited in size. Then, a certain minimum bulk density must be used.

The *flow – cure behavior* is the most important processing property of a thermosetting molding compound. Contrary to thermoplastics, not only does the viscosity of thermosets decrease on heating, it also rises again with the onset of curing (Fig. 4). The viscosity curve is thus U-shaped, the falling branch being determined by the physical process of heating, and the rising leg by the chemical curing reaction. The more steeply the left-hand branch falls, the more readily the molding compound softens. For compression molding, the softening behavior is the most important property of the thermoset molding compound. The middle part of the U-shaped curve is a measure of the lowest possible viscosity and determines the lowest possible processing pressure. The width of the curve and the steepness of the rising leg are measures of the curing rate.

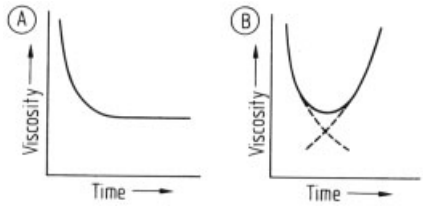

Figure 4. Flow-cure behavior of thermoplastics (A) and of thermosets (B) on heating to processing temperature

There is no test method that covers the whole curve. Generally two complementary tests are employed.

In the plates method (Fig. 5) a defined amount of molding compound is pressed flat between two heated plates at a given pressure. The width or height of the "cake" is then used as a measure of the flowability. The drawback of this process is that the specific pressure changes continually during the deformation.

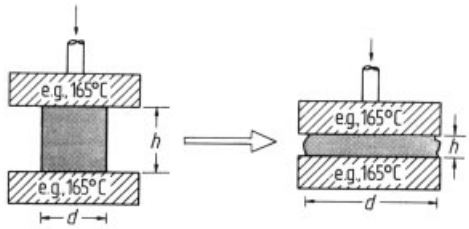

Figure 5. Testing of flow behavior by the plates method

In the beaker method (DIN 53 465, Fig. 6) a molding in the shape of a beaker is produced at a given pressure and a given temperature. The time from the lowering of the ram onto the material to complete sealing of the compression mold is measured. This "beaker time" is a measure of the softening and flow behavior of the material. In addition, the stiffness on demolding is tested. The beaker, still hot, is loaded on its top edge with a defined force (e.g., 10 N) and compressed.It undergoes a deformation that depends on the degree of cure. At a given curing time and curing temperature, the compression in millimeters is a measure of the stiffness on demolding. If several beakers with increasing curing time are tested, the curing rate of a molding material can be determined. In the American version of the beaker method (ASTM D-731) minimum pressure for the outflow of the beaker is measured. Here, therefore, several pressings are for time deter.

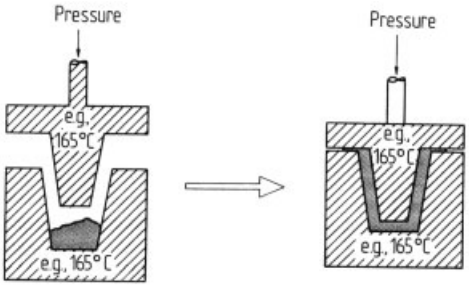

Figure 6. Testing of flow behavior by the beaker method

In the rod method (Fig. 7) a tablet of the material is pressed from a heated chamber under defined pressure into a heated flow channel. The length of the rod formed is a measure of the flowability. Recording the growth of the rod with time gives softening curves. If the tablet is left in the heated chamber for various lengths of time before it is pressed into the flow channel, and the length of the rod obtained is plotted against the time, curing curves are obtained. This Rossi – Peakes flow test is only suitable for short-fiber molding compounds. For coarse-structured materials, rods of larger diameter are used.

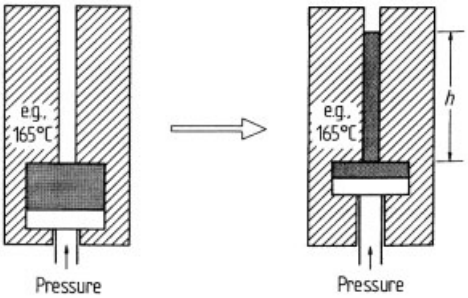

Figure 7. Testing of flow behavior by the rod method

A variation of the rod method is the spiral test, in which the rod is spirally wound, allowing long rods to be accommodated in a relatively small mold. These long rods are required for testing thermoset molding compounds that flow particularly readily. Thermosets of this kind, also known as low-pressure materials, are used in the electronics industry for encapsulating sensitive components. Epoxy resin molding materials in particular are tested by this method in EMMI spirals (EMMI: Epoxide Molding Material Institute).

In the kneader method (Fig. 8) the forces developed at the kneading blades when the molding compound is kneaded are measured (e.g., with a Brabender plastograph). If a defined amount of the molding material is introduced into the small heated kneader, the expenditure of energy at first falls on melting and then, as

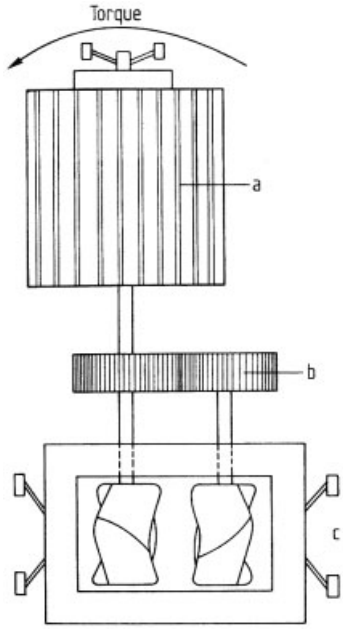

Figure 8. Testing of flow behavior with a Brabender plastograph
a) Motor; b) Gearing; c) Motor kneader

curing begins, rises again. Finally the material becomes crumbly, and the expenditure of energy falls again. The quantity measured is the torque developed by the driving motor, which gives a U-shaped curve. However, this test does not measure the entire flow – cure behavior, since the material crumbles before it is completely cured.

The rotary viscometer or Kanavec method (Fig. 9) is the most precise test of flow – cure behavior. A test specimen resembling a bicycle hand-grip is molded in a special mold, and during the curing process the male mold is rotated slowly (e.g., 1 revolution in 20 min) relative to the mold cavity. The quantity measured is the torque required for uniform rotation or the shearing stress generated. This method records very accurately the lower part and the rising branch of the U-shaped viscosity curve of thermoset molding compounds.

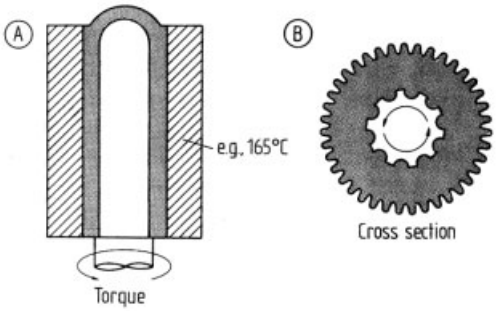

Figure 9. Testing of flow behavior by the Kanavec Method

A special form of the test is the Philips method, in which a mushroom-shaped ram is pressed into a cylindrical mold cavity. The ram has slits, through which the molding compound can escape upwards during the deformation process. The quantity measured is the loss of material or the thickness of the cured residual material.

5.2. Cured Moldings

The properties of cured moldings are always a combination of the binder properties and the filler properties. A survey is provided in DIN 7708 for phenolics and aminoplastics, and in DIN 16 911 for unsaturated polyester resins.

Materials Based on Phenolic Resins. Phenolic resin molding compounds yield moldings with small susceptibility to cracking and small aftershrinkage (Table 3). They are mostly of low flammability, but generally have a low tracking resistance. They are not light stable, and slowly discolor under the influence of light. They are therefore manufactured only in black, brown, or muted red and green shades. Aminoplastics are fully light stable and can be supplied in light and colored shades.

Materials Based on Urea, Melamine, and Melamine – Phenolic Resins. These molded materials have particularly low flammability and, in combination with suitable fillers, exhibit very high tracking resistance (Table 4, see next page), especially in the case of melamine resins. However, both the urea – formaldehyde (UF) and melamine – formaldehyde (MF) aminoplastics have relatively high aftershrinkage values. In unfavorable cases this can lead to suscepti-

Table 3. Properties of phenolic moldings (DIN 7708) ∗

Type	Filler	Density, g/cm^3	Bending strength, N/mm^2	Impact strength, kJ/m^2	Notch impact strength, kJ/m^2	Martens temperature, °C	Surface resistance, comparative figure	Water absorption, mg
11.5	stone powder	1.8	50	3.5	1.3	150	10	45
12	mineral fiber	1.8	50	3.5	2	150	8	60
13	mica	1.9	50	–	2	150	10	20
31	wood flour	1.4	70	6	1.5	125	8	150
85	wood flour + cellulose	1.4	70	5	2.5	125	7	200
83	wood flour + textile fibers	1.4	60	5	3.5	125	8	180
84	chopped cotton cloth + cellulose	1.4	60	6	6	125	8	150
51	cellulose	1.4	60	5	3.5	125	7	300
71	textile fibers	1.4	60	6	6	125	7	250
74	chopped cotton cloth	1.4	60	12	12	125	7	300

∗ Minimum values, except for water absorption (maximum value).

Table 4. Properties of aminoplastic and aminoplastic – phenolic moldings (DIN 7708, minimum values)

Type	Filler	Base resin	Density, g/cm^3	Bending strength, N/mm^2	Impact strength, kJ/m^2	Notch impact strength, kJ/m^2	Martens temperature, °C	Surface resistance, comparative figure	Water absorption∗, mg	Tracking resistance
131	cellulose	urea	1.5	80	6.5	1.5	100	10	300	600
150	wood flour	melamine	1.5	70	6	1.5	120	10	250	600
152	cellulose	melamine	1.5	80	7	1.5	120	10	200	600
153	textile fibers	melamine	1.5	60	5	3.5	125	9	300	600
154	chopped cotton cloth	melamine	1.5	60	6	6	125	8	300	600
155	stone powder	melamine	2.0	40	2.5	1	130	8	200	600
156	mineral filler (or fibers)	melamine	1.8	50	3.5	2	140	8	200	600
157	mineral filler (or fibers) + wood flour	melamine	1.7	60	4.5	1.5	140	9	200	600
180	wood flour	melamine/phenol	1.5	80	6	1.5	120	10	180	175
181.5	cellulose	melamine/phenol	1.6	80	7	1.5	120	10	150	600
182	stone powder + wood flour	melamine/phenol	1.6	70	4	1.2	120	10	120	600
183	stone powder + cellulose	melamine/phenol	1.6	70	5	1.5	120	10	120	600

∗ Maximum values.

bility to cracking and to distortion at elevated temperatures.

The aminoplastic – phenophlastic (MP) molded materials (Types 180 – 183), as a result of the small proportion of plasticizing phenolic resin, behave more advantageously. With pale shades, however, only a very small proportion of phenolic resin is possible.

Materials Based on Polyester Resins. Polyester resin molding compounds (UP) have extremely small aftershrinkage (Table 5). They can be pale or colored and are light stable, resistant to tracking, and dimensionally stable. Thus they exhibit a combination of the advantageous properties of phenolics and aminoplastics, without their drawbacks. At first, UP materials were produced from styrene-containing unsaturated polyester resins with a relatively high proportion of glass fibers. Owing to their moist, fibrous structure, they were also known as "sauerkraut compounds". They yielded mold-

Table 5. Properties of polyester resin molding (DIN 16 911, minimum values)

Type	Filler	Form supplied	Bending strength, N/mm^2	Impact strength, kJ/m^2	Notch impact strength, kJ/m^2	Martens temperature, °C	Surface resistance, comparative figure	Water absorption∗, mg
801	glass fiber long + stone powder	moist, semidry and dry	60	22	22	125	10	100
802	glass fiber short + stone powder	dry, free-flowing	55	4.5	3	140	12	45
803	glass fiber long + stone powder	moist, semidry and dry	60	22	22	125	10	100
804	glass fiber short + stone powder	dry, free-flowing	55	4.5	3	140	12	45

∗ Maximum values.

ings of relatively uneven structure and were therefore mainly used for engineering components.

Improved surface quality was achieved by so-called low-profile additives (thermoplastic additives), which made possible the use of UP molding materials for high-quality visible components.

The use of monomer-free solid UP resins was a further decisive breakthrough. With them, perfect surfaces without sink marks, such as are required for domestic appliances, can be made. Such UP molding materials are dry, free-flowing, and low-polluting and can be processed on conventional automatic machines like phenolics and aminoplastics. However, unlike the latter, the polyester resin molding materials do not cure by a polycondensation process but by polymerization. Since with this curing process no low-molecular condensation products such as water are split off, polyester moldings can be removed from the mold sooner and molded at lower pressures. The curing process is initiated by the addition of peroxides. The mold temperature necessary for the curing is adjusted to suit the peroxide used. An advantage of the polyester resins is that their synthesis from diols and diacids as well as the chain length between cross-links can be varied within a wide range. A further interesting property of styrene-containing polyester resins is that their viscosity is increased greatly by adding magnesium oxide. This enables the production of UP molding compounds that are resin-rich but still highly viscous. The production of prepregs (DIN 16 913) makes use of this effect (→ Reinforced Plastics, Chap. 3.9.).

Dry, free-flowing polyester resin molding compounds containing organic fibers (e.g., cellulose) instead of glass fibers have gained acceptance particularly in the domestic appliance sector.

Materials Based on Epoxy Resins. Epoxy resin molding (EP) were classified under DIN 16 912 (Types 870 – 872). However, this field is undergoing changes in the binders used, so that some older types have been withdrawn. However, DIN 16 912 enables a survey to be made of the basic properties of EP materials (Table 6). With these resins the number of possible combinations is even greater than with polyesters. Epoxy resins can be cured by three different methods: basically with diamines, acidically with diacids or acid anhydrides, and catalytically with Lewis acids. Epoxy resin molding compounds of the older Types 870 – 872 tend to yellow; therefore, only muted, dark colorations are possible. Hence they are used only for engineering components. Formerly, the storability of EP materials was poor, and cooled containers and cold rooms were necessary. Meanwhile, however, it has become possible to produce types with good storability and considerably improved light stability [7], although the light stability does not reach that of the UP molding compounds. The advantage of the epoxy resin molding compounds lies in their combination of good physical properties, freedom from overcuring, even at very high curing temperatures, and the very low processing shrinkage and aftershrinkage. Epoxy resin moldings therefore have a particularly low tendency to thermal stress cracking. Owing to their chemical composition they

Table 6. Properties of epoxy resin moldings (DIN 16 912) ∗

Type	Filler	Base resin	Density, g/cm^3, ca.	Bending strength, N/mm^2	Impact strength, kJ/m^2	Notch impact strength, kJ/m^2	Martens temperature, °C	Surface resistance, comparative figure	Water absorption, mg
870	inorganic granular stone powder	epoxy	1.9	50	5	1.5	110	12	30
871	inorganic short-fibrous glass fibers	epoxy	1.9	80	8	3	120	12	30
872	inorganic long-fibrous glass fibers	epoxy	1.9	90	15	15	125	12	30

∗ Minimum values, except for water absorption (maximum value).

have a very good resistance to water and chemicals.

In addition, epoxy resin moldings have the advantage of retaining their electrical properties even under unfavorable climatic conditions. They permit formulations that are processible at low pressure (low-pressure compounds). The adhesion to metallic inserts is very good.

Miscellaneous Binders. For particularly demanding electrical requirements, diallyl phthalate molding compounds are used, with a prepolymer from diallyl phthalate as binder. The fillers can be impregnated with the powdery prepolymer by the melt fusion process. Peroxides are suitable for curing, as with the UP resins. Molded materials on this basis have outstanding electrical values, especially in moist atmosphere, and meet U.S. Military Specifications, above all in aircraft construction.

Silicone resin molding compounds satisfy still higher electrical and thermal requirements. These materials play a certain part in electronics for sheathing semiconductors. They have a very high temperature stability, even over long periods of time, but are very expensive.

Cold molding compounds based on special phenolic resins have only limited storability. They have gained some commercial importance because they permit the manufacture tracking-resistant PF moldings. Cold molding compounds (DIN 7708, part 14) are compounded wet in kneaders and molded cold. The moldings are subsequently cured pressureless with stepwise increase of temperature over a fairly long period of time (up to several days). The surface quality and dimensional accuracy of such moldings do not, however, approach the high standards of conventional moldings.

Test Methods for Molded Materials. Most of the test methods used for testing molded materials are defined in DIN specifications. For thermoset plastics, the test specimens are produced by compression molding. This eliminates the filler-orientation effects which invariably occur in injection molding. Studies are available that allow the directional dependence of the properties of injection molded test specimens, to be estimated, especially the shrinkage behavior [8, 5].

The most important DIN test methods are:

Bending strength	DIN 53 452
Impact and notch impact strength	DIN 53 453
Compressive strength	DIN 53 454
Modulus of elasticity	DIN 53 457
Ball indentation hardness	DIN 53 456
Martens heat resistance	DIN 53 458
Linear expansion coefficient	DIN 53 752
Thermal conductivity	DIN 52 612
Incandescence resistance	DIN 53 459
Surface resistance	DIN 53 482
Electrical volume resistivity	DIN 53 482
Dielectric loss factor	DIN 53 483
Relative permittivity	DIN 53 483
Dielectric strength	DIN 53 481
Tracking resistance	DIN 53 480
Water absorption	DIN 53 495

In addition there is also a series of special test methods laid down by firms or associations, and test specifications for thermoset plastics from ISO, European Standards EN, and the UL (Underwriter's Laboratories).

Users' requirements for comparability of very different plastics have been generally accepted, which led to revision of DIN 7708. This began with a harmonization of the dimensions and production of the test specimen and was completed with the new edition of DIN 7708, part 8 (February 1992).

The new test specimen is of a multipurpose type, originally proposed for thermoplastics, and adapted to the special requirements for the processing of thermosets. Dyes for producing the test specimens by injection molding and compression molding have been developed, and comparisons made between the test values obtained and the previous values.

6. Processing and Aftertreatment

Thermoset molding compounds are hot molded and cured under pressure, sometimes with the addition of hardeners. The usual processes are compression molding, injection molding, transfer molding, injection – compression molding, and extrusion. For a detailed description of the processing of molding compounds to give cured moldings, see [4].

Moldings must generally be deflashed. This can be performed manually by filling, grinding, punching, sawing, abrasive blasting, or brushing. For small moldings or large numbers, mechanical processes are used, such as tumble deflashing, abrasive blasting, or brushing. Saw-

ing, milling, turning, and drilling are also possible [9]. Tools with carbide cutting edges are usually used. Detailed information on steel quality, cutting speed, and tool quality can be found in [9].

If necessary, the moldings can be polished with polishing wheels or in drums. Thermoset moldings can be joined by detachable joints (e.g., screw or clamp joints) or nondetachable joints such as rivets and adhesive bonds [10].

Thermoset moldings are sometimes finished by painting, flocking, hot-embossing, printing, or metallizing by electrodeposition or vacuum deposition.

7. Uses

The main applications for moldings made from thermoset molding compounds are electrical engineering, domestic appliances, mechanical engineering, the automotive industry, and the sanitary sector. Thermoset molding materials also have versatile uses as bottle seals, in precision engineering, and in office machine manufacture. Typical moldings in the electrical engineering sector are collectors for electric motors, slip rings, contacts, strip terminals, distributor box cases, switches, plugs, and other connectors. Collectors also play a major role in the automotive industry, as well as ignition distributors, spark plug sockets, control pistons for power brakes, bushes, and axial face seals. In the household sector, thermoset moldings are used for high-quality camping ware, ashtrays, handles for irons, pot and pan handles, grill and toaster components, lamps, and cutlery. In the sanitary sector, high-quality toilet fittings are manufactured in many colors from thermoset molding compounds. They are used for a great variety of cosmetic containers (including metallized ones). In mechanical engineering, they are suitable for highly stressed bearings, including self-lubricating bearings, and for coverings, rollers, reels and operating controls.

8. Economic Aspects

The "thermosets" discussed here (thermoset molding compounds) constitute only a narrow area of the thermosetting plastics. They contain 20 – 60 % of thermosetting binders, but such binders are also constituents of a wide variety of other products. Production of curable free-flowing molding compounds is summarized in Table 7.

Table 7. Production of curable, free-flowing molding compounds

Region	Production, 10^3 t			
	1979	1982	1985	1988
Federal Republic of Germany *	58	55	59	60
Western Europe	210	190	215	207
United States	189	115	135	133
Japan	120	115	114	103

* Without urea molding materials, which are no longer produced in Germany.

For the phenolic resins, which make up ca. 2 % of foundry molding compounds and ca. 60 % of electrical laminates, a total production volume of material for the Federal Republic of Germany in 1988 of almost 3×10^6 t can be estimated [11]. The production of thermoset molding compounds based on phenolic resin, on the other hand, is much lower, Table 8 [12].

Table 8. Production of phenolic molding compounds

Region	Production, 10^3 t				
	1987	1988	1989	1990	1991
Japan	59.2	64.9	67.3	65.2	64.8
Korea	5.0	5.4	5.8	5.8	6.0
South-East Asia	16.2	17.8	25.0	26.0	28.0
United States	89.5	89.7	79.9	76.3	70.8
Western Europe	75.5	74.0	72.0	73.0	67.0
Germany *	29.0	31.0	32.0	32.0	31.5

* Included in data of Western Europe.

Market growth for thermoset molding compounds is estimated at 3 – 4 % for the next few years. In automobile manufacture even 6 – 7 % may be reached. Here the thermosets compete mainly with metals. Quantitatively, the largest fields of application are electrical engineering, followed by domestic appliances and tableware, and sanitary products. These considerations do not include the large field of polyester dough molding compounds (bulk molding compounds, BMC) and prepregs (sheet molding compounds, SMC).

9. Environmental Protection and Recycling [13]

Safety data sheets (DIN 52 900) are available for all thermoset molding compounds. They give information on storage and handling, transport, fire protection, toxicology, ecology, waste disposal, etc.

Phenolic moldings that contain $> 1\,\%$ free phenol must be labelled as harmful (X_n). However, by the 1970s manufacturers had already undercut this 1 % limit by using low-phenol resins. Most phenolic molding compounds therefore do not require such labelling. According to all transport regulations, molding compounds are not deemed to be hazardous materials.

In connection with processing, further questions arise such as which gases and vapors are formed, and in what amounts? For phenolic molding compounds, extensive studies on this subject are available. During processing, phenol, formaldehyde, and ammonia are released, in addition to water vapor. The quantities have been determined by using an enclosed mold, extracting the gases, and absorbing them in appropriate solutions. The following approximate values were obtained:

- Phenol, ca. 5 mg per kg molding compound
- Formaldehyde, ca. 2 – 4 mg per kg molding compound
- Ammonia, ca. 20 – 40 mg per kg molding compound

The MAK values for these substances :

- Phenol 5 ppm (19 mg/m^3)
- Formaldehyde, 0.5 ppm (0.6 mg/m^3)
- Ammonia, 50 ppm (35 mg/m^3)

can be maintained by using extraction hoods.

During processing of aminoplastics, formaldehyde quantities of up to ca. 25 mg/m^3 have been observed, and demands on the extraction equipment are therefore greater.

Thermoset molding compounds based on phenolic resin, melamine resin, and urea resins contain as binder polycondensation products produced using formaldehyde. However, owing to polycondensation formaldehyde is not present in the free form. Curing converts the resin molecules via reactive centers derived from formaldehyde into an almost inert network, which gives the molded materials their resistance to temperature and aggressive media. With correct curing, the regeneration of formaldehyde in the molding is possible only to a minor extent in the case of the aminoplastic molding compounds (UF, MF and MP). However, it is observed only under extreme thermal or hydrolytic conditions. In moldings based on phenolic resin, such a release of formaldehyde is practically immeasurable.

No formaldehyde is used in the synthesis of unsaturated polyester resins. For special applications, polyester molding compounds are modified with up to 5 % melamine resin. In these cases small quantities of formaldehyde can escape during the processing.

During the processing of styrene-containing unsaturated polyester dough molding compounds, styrene may be released. Meeting the MAK value for styrene of 20 ppm (85 mg/m^3) requires elaborate ventilation systems. Producers of such UP molding compounds are therefore attempting to reduce the styrene content of their products.

A few years ago asbestos was still a popular reinforcing fiber for thermosets. Although the fibers are enveloped by resin in the process, in compounding and aftertreatment of the moldings exposure of the fibers cannot be avoided. Such asbestos fibers are classified as carcinogenic in certain geometries (particles with a length of $> 5\,\mu$m, and a diameter of $< 3\,\mu$m with a length-to-diameter ratio $> 3:1$). Substitutes have since been found and new asbestos-containing molding materials are no longer marketed. The Bakelite company, for example, ceased production of asbestos-containing molding compound at the beginning of 1991.

Recycling of thermoset molding compounds was carried out as early as the 1930s, but was not economic at that time, since the wastes had to be very finely ground to avoid surface defects in the new moldings. There was also a risk of entraining impurities. For controlled, quality-assured moldings in Germany the addition of reground material is only allowed if the processor undertakes the inspection necessary to safeguard quality.

As early as the 1970s, studies were carried out worldwide to facilitate the addition of reground material. The conclusion was that the addition of

10 – 20 % of reground material is possible without significantly impairing the test values.

A similar particle recycling process is used for sheet molding compounds. Raw material manufacturers and processors have established a company for this purpose (Ercom, Karlsruhe).

The problem of damage to thermoplastics on recycling does not occur with thermosets, since the reground material is used as a filler. In the finely ground state, this reground material is a particularly high-quality filler since it is impregnated.

The regenerate can be added either batchwise in mixing vessels or continuously in injection molding machines. The particle size of the regenerate should be ca. 0.1 – 0.4 mm. Too coarsely ground material leads to surface and textural defects.

Incineration. Since most thermoset molding compounds are free of halogens, disposal by incineration is also possible. The calorific values of such wastes of 12500 – 21000 kJ/kg are relatively high. However, thermoset wastes are not accepted by all refuse incineration plants. It is occasionally possible to use them as additives in cement manufacture to economize on oil and coal.

Dumping. Because of their insolubility, cured molded materials can be dumped on landfills for domestic waste. Uncured wastes are regarded as pollutants.

In Germany an active environmental study group for thermosets has existed for several years [14].

A comprehensive survey of the problems of plastics recycling is given in [15].

Thiamine → Vitamins

10. References

1. S. H. Goodman: *Handbook of Thermosets,* Noyes Publications, Park Ridge, N.J., 1986.
2. "Duroplaste," *Kunstoff-Handbuch,* **vol. X,** Hanser, München 1988.
3. W. Schönthaler, *Maschinenmarkt* **77** (1971) 2227 –2231.
4. W. Schönthaler: *Verarbeiten Härtbarer Kunststoffe,* Taschenbuch T 40, VDI-Verlag, Düsseldorf 1973.
5. Vergleich von Prüfnormen nach ISO, DIN, ASTM, JIS und BS, Bakelite company brochure, Iserlohn 1996, p. 66.
6. W. Bauer, W. Woebcken: "Verarbeitung duroplastischer Formmassen," *Kunststoffverarbeitung,* **vol. 3,** Hanser, München 1973.
7. P. Tschanz, *Kunstst. Plast.* **25** (1978) 1.
8. W. Schönthaler, K. Niemann, *Maschinenmarkt* **82** (1976) 1638 – 1641 and **83** (1977)229 – 230.
9. H. Zickel: "Das spanabhebende Bearbeiten der Kunststoffe," *Kunststoffverarbeitung,* **vol. 4,** Hanser, München 1963.
10. W. Schönthaler, J. Tetz, *Maschinenmarkt* **81** (1975) 257 – 259.
11. W. Schönthaler, U. Braun, *Kunststoffe* **80** (1990) 1165 ff.
12. M. Kraemer, *Plast. Eur. Kunstst.* Oct. 1992, p. 493.
13. W. Schönthaler: "Duroplaste Zukunft von Anfang an," Script, Würzburg 1993.
14. Technische Vereinigung der Hersteller u. Verarbeiter duroplastischer Kunststoff-Formmassen, Würzburg, brochure.
15. H. Breuer, E. Dolfen, *Kunststoff-Recycling Handbuch 1992,* A. Hüthig Verlag, Heidelberg.

Thin Films

Georg Wahl, Technische Universität Braunschweig, Institut für Oberflächentechnik, Braunschweig, Federal Republic of Germany (Section 1.1)

Paul B. Davies, University of Cambridge, Department of Chemistry, Cambridge, United Kingdom (Sections 1.2 and 2.6.2)

Rointan F. Bunshah, University of California, Department of Materials Science and Engineering, Los Angeles, United States (Section 1.3)

Bruce A. Joyce, University of London, Imperial College of Science, Technology and Medicine, London, United Kingdom (Section 1.4)

Colin D. Bain, University of Oxford, Physical Chemistry Laboratory, Oxford, United Kingdom (Section 1.5)

Gerhard Wegner, Max-Planck-Institut für Polymerforschung, Mainz, Federal Republic of Germany (Sections 1.6 and 2.6.1)

Markus Remmers, Max-Planck-Institut für Polymerforschung, Mainz, Federal Republic of Germany (Sections 1.6 and 2.6.1)

Francis C. Walsh, University of Portsmouth, Department of Chemistry, Portsmouth, United Kingdom (Sections 1.7 and 2.4)

Konrad Hieber, Fraunhofer Gesellschaft, Institut für Festkörpertechnologie, München, Federal Republic of Germany (Section 2.1)

Jan-Eric Sundgren, Linköping University, Department of Physics and Measurement Technology, Linköping, Sweden (Section 2.2)

Peter K. Bachmann, Philips GmbH, Research Laboratories, Aachen, Federal Republic of Germany (Section 2.3)

Shintaro Miyazawa, NTT LSI Laboratories, Kanagawa, Japan (Section 2.5)

Alfred Thelen, Ingenieurbüro Thelen für optische Interferenzschichten, Schmitten-Seelenberg, Federal Republic of Germany (Section 2.7)

Heiner Strathmann, University of Twente, Faculty of Chemical Technology, Enschede, The Netherlands (Section 2.8)

1. Preparation of Thin Films 3
1.1. Chemical Vapor Deposition Processes 3
1.2. Plasma Polymerization 6
1.3. Physical Vapor Deposition (PVD) Processes 7
1.3.1. Sputter Deposition 8
1.3.2. Evaporation Deposition 10
1.3.3. Deposition of Metals, Alloys, and Compounds 11
1.4. Thin Film Growth by Molecular Beam Epitaxy (MBE) 12
1.4.1. Introduction 12
1.4.2. The Configuration of MBE Growth Systems 12
1.4.3. Monitoring Facilities 16
1.5. Thin Films by Adsorption Processes 17
1.5.1. Self-Assembled Monolayers 18
1.5.2. Self-Assembled Multilayers 19
1.6. Langmuir – Blodgett and Related Techniques 19
1.6.1. Low Molecular Mass Substances . . . 22
1.6.2. Polymeric Amphiphilic LB Layers . . 22
1.6.3. Hairy Rod Polymers 25
1.7. Electrochemical Deposition Processes 28
1.7.1. Scope 28
1.7.2. Electrochemical Deposition of Metals and Alloys 28
1.7.3. Electrophoretic Coatings 34
1.7.4. Trends in Electrodeposition 35
2. Thin Films as Materials 37
2.1. Coatings for Electronic Applications 37
2.1.1. Introduction 37

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_681

2.1.2. Hybrid Technology 37
2.1.3. Microsystems 37
2.1.4. Data Storage 37
2.1.5. Microelectronics 37
2.1.6. Semiconducting Films 39
2.1.7. Ferroelectrics 39
2.2. Hard and Decorative Nitride and Carbide Coatings 39
2.2.1. Introduction 39
2.2.2. Growth Techniques 39
2.2.3. Coating Microstructures 40
2.2.4. Adhesion 41
2.2.5. Hard Coatings 41
2.3. Carbon Films 43
2.3.1. Carbon Phases: Composition and Stability 43
2.3.2. Properties and Applications of Diamond Films 43
2.3.3. Properties and Applications of a-C:H, a-C, and ta-C Films 47
2.4. Metallic Films 48
2.4.1. Introduction 48
2.4.2. Electronics Applications 48
2.4.3. Engineering Coatings 50
2.4.4. Dispersed Catalytic Coatings 50
2.4.5. Developments and Trends 51
2.5. Superconductive Coatings 51
2.6. Polymer Films 53
2.6.1. Langmuir – Blodgett Layers of Polymer Films as Materials 53
2.6.2. Films by Plasma Polymerization . . . 58
2.7. Optical Coatings 60
2.7.1. Introduction 60
2.7.2. Theory 61
2.7.3. Deposition Methods 61
2.7.4. Coating Materials 61
2.7.5. Classical Coatings 62
2.7.6. Nonclassical Coatings 63
2.7.7. Applications 63
2.8. Membranes 63
2.8.1. Introduction 63
2.8.2. Fundamentals 64
2.8.3. Membrane Structures and their Preparation 64
3. References 65

1. Preparation of Thin Films

1.1. Chemical Vapor Deposition Processes

Chemical vapor deposition (CVD) produces layers from the gas phase with thicknesses in the range ca 0.1 – 10 µm, by means of a chemical reaction. Typical pressure ranges are different from those used in physical vapor deposition (< 1 hPa); ca 1 – 50 hPa for low-pressure CVD (LPCVD), 50 to < 1000 hPa for medium-pressure CVD, and atmospheric pressure for atmospheric CVD (ACVD) [1–8]. Accordingly, the Knudsen number λ/d (the ratio of the mean free path of molecules in the gas phase to a typical reactor dimension) is small, and diffusion and convection processes are the main transport mechanisms. Reynolds numbers are in the range 1 – 1000, so laminar gas flow is the normal flow mode and buoyancy effects are important, especially at higher pressure [7].

In contrast to chemical transport processes [9], CVD is an open system with a gas inlet and outlet. The equipment (Fig. 1) has the following main parts: a gasification system (a, b, c, j), a chemical reactor with volume of the order of cm^3 or m^3, which contains the samples or workpieces to be coated (d, e, f), and an exhaust system for the reacted gas (g, h, i). Precursor compounds can be inorganic (e.g., halides, hydrides, carbonyls) or organic (alkyls, β-diketonates, alkoxides, or organometallics [10, 11]). Almost all compounds can be deposited by CVD, provided the material can be dissolved in the gas phase. Usually, the possibility of depositing material from a gas mixture and the expected phase composition can be calculated by thermodynamic equilibrium calculations, minimizing the Gibbs free energy. These estimates are especially suitable for CVD processes at high deposition temperatures (e.g., SiC deposition at ca. 1000 °C [12]). At low temperature, kinetic effects are often important, and the deposition of metastable phases (e.g., amorphous SiO_2) is possible.

The structure of the coating is the result of several steps: adsorption – desorption processes, surface reactions, surface diffusion, nucleation, growth of the critical nuclei, layer formation, and aging processes in the layer during deposition. The result ist often the formation of preferred orientations of the deposited layer on crystalline and amorphous substrates, depending on the deposition conditions. On monocrystalline sub-

strates, epitaxy is possible if the lattice misfit is not too large [13].

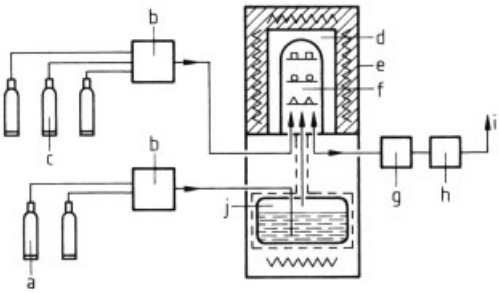

Figure 1. CVD system for deposition on tools (e.g., TiN, TiC)
a) Gases; b) Gas control unit; c) Precursor gases; d) Interspace to remove reactor from the heating system for cooling; e) Heating system; f) Deposition reactor (hot-wall reactor); g) Filter; h) Pump; i) Ventilation, j) Evaporator (liquid or solid)

There are a number of methods to activate the chemical reactions necessary for deposition. These methods are described in the following.

Thermal CVD. The surface or homogeneous reactions necessary for layer formation can be thermally activated (thermal or conventional CVD) by heating the reactor, including the samples (hot-wall reactor, Fig. 1) or by heating the samples only (cold-wall processes) to the deposition temperature T_{d} (Table 1). These processes have the "ideal" temperature dependence shown in Figure 2. At medium temperature, in the so-called transport-controlled regime, transport in the gas phase determines the deposition rate, and the temperature dependence is very smooth, proportional to T^{ν} with $\nu = 0.7 - 1.0$. This range can be used only for deposition on simple geometries, because the deposition depends critically on gas flow conditions. Deposition rates are determined by the concentration of precursor molecules in the gas phase and by the gas flow. In stagnation flow reactors, deposition rates up to 100 µm/min can be reached, as shown for tungsten deposition with a precursor content in the hydrogen gas phase of ca. 30 % WF_6 [15].

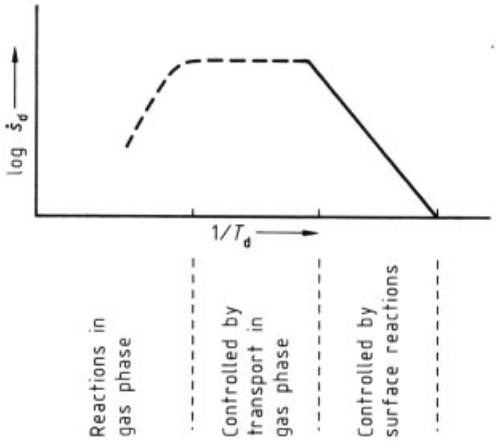

Figure 2. Temperature dependence of an "ideal" CVD process

At low temperature, the deposition rate is given by an Arrhenius law:

$$\dot{s} = k \exp(-E/RT) \tag{1}$$

where k is a constant, E is the activation energy, and R the gas constant. In surface reactions, T is equal to the deposition temperature T_{d}. If the rate-determining step is in the gas phase, it is dif-

Table 1. Examples of thermal CVD processes

Deposited material	Type of reaction	Reaction	Deposition temperature, °C	Activation energy, kJ/mol	Ref.
W	pyrolysis	$WCl_6 \rightarrow W + \ldots$	>1000		
W	reduction	$WCl_6 + 3\,H_2 \rightarrow W + \ldots$	>800	627	[14]
W	reduction	$WF_6 + 3\,H_2 \rightarrow W + \ldots$	>400	60 ± 10	[15]
Si	pyrolysis	$SiH_4 \rightarrow Si + \ldots$	>500	150 ± 10	[16]
Si	reduction	$SiCl_4 + 2\,H_2 \rightarrow Si + \ldots$	>1000	150 ± 10	[17]
SiO_2	oxidation	$SiH_4 + \ldots\ O_2 \rightarrow SiO_2 + \ldots$	>400		
SiC	displacement	$SiCl_4 + \ldots\ CH_4 \rightarrow SiC + \ldots$	>900		
Ge	disproportionation	$2\,GeI_2 \rightarrow Ge + GeI_4$	>600		
TiC	disproportionation	$TiCl_4 + \ldots\ CH_4 \rightarrow TiC + \ldots$	>900		
Ni	pyrolysis	$Ni(CO)_4) \rightarrow Ni + \ldots$	>150		
$Y_1Ba_2Cu_3O_{7-\delta}$	oxidation	$1\,Y(dpm)_3 + 2\,Ba(dpm)_2 + 3\,Cu(dpm*)_2 + \ldots\ O_2 \rightarrow Y_1Ba_2Cu_3O_{7-\delta}$	>600	82 ± 5	[18]

* dpm = Dipivaloyl methanoat

ficult to determine the reaction temperature if the reactor is not isothermal. Typical activation energies are given in Table 1. At high temperature, the deposition rate often decreases because homogeneous reactions and nucleation in the gas phase reduce transport of precursor molecules to the surface. Very often the deposition characteristics is much more complicated than shown in Figure 2. A more detailed description of the dependence of deposition on pressure, temperature, gas composition, and gas flow is then possible by the methods of chemical kinetics and the theory of catalysis including fluid dynamics. The kinetically controlled range shown in Figure 2 can also be used for the deposition inside porous materials. This process is called chemical vapor infiltration (CVI) [19,20]. Thus the temperature dependence is often more complicated than shown in Figure 2.

The activation of the chemical reaction at the surface can also be achieved by laser irradiation [21]. In this case, localized deposition with lateral dimensions down to a few micrometers is possible. This laser CVD process can be mainly thermal or mainly photochemical (photo CVD), and allows patterning of surfaces by laser-direct write. The fabrication of three-dimensional microobjects by laser CVD has been demonstrated [22].

In recent years there has been interest in the development of efficient lamps to replace the cost-intensive lasers. The intense ultraviolet (UV) light is generated on these devices using dielectric barrier dischargers to form excimer molecules of rare gases and rare gase halides [23]. This has provided the oportunity for inexpensive large area photo CVD [24].

Practical deposition rates (0.1 – 1 μm/min) are mostly a compromise between the coating structure and the throwing power (the ability to produce layers of constant thickness on the surface of the workpiece). Thermal CVD processes are used industrially for the production of hard coatings (TiC, TiN, Al_2O_3), and in the electronics industry (Si, SiO_2, Si_3N_4, III – V compounds).

Plasma CVD. Very often the deposition temperature of conventional CVD is too high; deposition methods have been developed to decrease the deposition temperature, e.g., plasma-activated deposition processes (PACVD), also called plasma-enhanced CVD (PECVD). A plasma is an electrically neutral gas in which neutral species, ions, and electrons are all present. Plasmas are ignited by d.c. or a.c. in the frequency range MHz to GHz. ln d.c. (continuous or pulsed) discharges, the sample is the cathode and the discharge is ignited between the cathode and the anode. The cathode surface is coated and at the same time bombarded with positive ions with energy in the range of a few hundred volts, and current density 1 – 10 mA/cm^2. If the sample to be coated is insulating, an a.c. discharge is necessary to avoid charging effects. The different mobilities of ions and electrons produce a space-charge layer on the electrodes, with a potential drop similar to that in a d.c. plasma. The voltage can be further increased by changing the ratio of the electrode areas, smaller areas being associated with higher voltages (approximately proportional to $(A_1/A_2)^\nu$, where A_1 and A_2 are the areas of the the electrodes, and $\nu = 1-4$ [6], [25]. In plasmas used for CVD, the ratio of charged to neutral species can be as low as $10^{-6} - 10^{-5}$ [6,26] (weakly ionized plasmas). The total pressure is $10 - 10^3$ Pa. This low pressure leads to few collisions in the gas phase before species reach the walls or recombine. Therefore, no complete temperature equilibration of neutral species, ions, and electrons is formed. Neutral species and ions have temperatures in the range $kT = 0.1$ eV (where k is the Boltzmann constant, 0.1 eV = 1160 K); electron temperatures are in the range 1 – 10 eV. The role of the plasma can be twofold:

1) To change the gas phase composition and thus change the reaction system by formation of intermediates, radicals, excited molecules, and ions. Plasmas for this application can also be produced remote from the sample position (Fig. 3 A).
2) To modify the gas phase and deposition conditions. This is achieved by ignition of the gas discharge directly above the deposition surface (Fig. 3 B). Ion bombardment through the space-charge layer changes the surface processes and hence the structure of the coating, usually improving adhesion to the substrate.

Plasma CVD processes are used for the production of hard coatings on steels [26] and in the electronics industry (Si_3N_4).

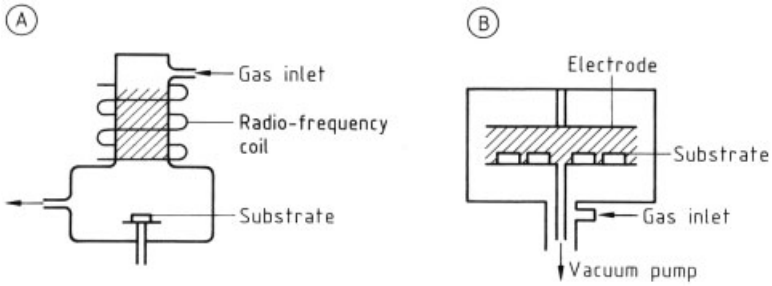

Figure 3. Basic configurations of plasma reactors for CVD
A) Remote plasma reactor (inductively coupled); B) Parallel-plate reactor (capacitively coupled)

1.2. Plasma Polymerization

Plasma polymerization refers to the formation of polymeric films on surfaces in proximity to a low-pressure glow discharge in an organic or organometallic vapor. Among the unique properties of these films are: controlled thickness up to ca. 2 µm, good adhesion to a variety of substrates, and variable physical properties such as hardness. The useful characteristics of such films has led to their widespread use in applications as diverse as protective coatings and osmotic membranes. Some of the features of plasma polymerization are similar to those found in plasma-enhanced chemical vapor deposition (PECVD). Several plasma parameters are known to influence the polymerization process, including pressure, flow rate, power and frequency of the discharge, and reactor geometry.

A variety of reactor geometries have been used, including those with internal or external electrodes. Reactors with internal electrodes usually have a parallel plate design in which one electrode is grounded (onto which the substrate is placed) and the other is fed with capacitively coupled radiofrequency (r.f.) power. Tubular reactors are often inductively coupled.

A typical bell jar reactor with horizontal, parallel plate electrodes is shown in Figure 4. It has been used [27] to prepare protective films of fluorinated polymers from tetrafluoroethylene (TFE) and chlorotrifluoroethylene (CTFE) monomers. Plasma reactors must have good vacuum integrity; in this example it was better than 1 Pa. The pressure in the operating reactor was ca. 10 Pa. Typical deposition rates over a 5-cm diameter substrate were $7.9\times10^{-6}\,\mathrm{g\,cm^{-2}\,min^{-1}}$ ($48\,\mathrm{nm\,min^{-1}}$) with 20 W r.f. power, 13 Pa TFE, and 13 Pa argon. Deposition time for a 1-µm coating on NaCl substrate was ca. 40 min.

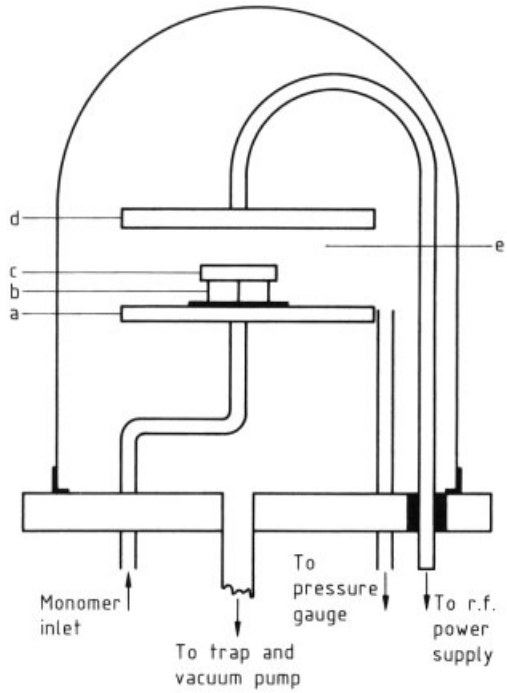

Figure 4. A parallel plate capacitively coupled plasma reactor [27]
a) Ground electrode; b) Support pedestal; c) Optical crystal; d) Hot electrode, e) Plasma

The control and measurement of pressure in a reactive plasma presents several problems. The pressure in the reactor can change significantly when the discharge is turned on, even at constant flow rate, owing to changes in the concentration and nature of the plasma species. The ideal device for measuring the pressure in the plasma must be immune from the electrical effects of the

plasma, and not disturb the plasma chemistry. A pressure transducer which is independent of the nature of the plasma species is usually selected. At constant flow rate and current density, the deposition rate increases linearly up to some limiting pressure. Maximum reactor pressures which have been used are ca. 500 Pa, but pressures below 250 Pa are usually employed.

Closely allied to the gas pressure is the flow rate F. The dependence of deposition rate on flow rate at constant pressure is as follows for ethane plasma polymerization. At high flow rates deposition tends to decrease with increasing flow rate, owing to decreasing residence time. Conversely, the longer residence times at low flow rates enable more of the monomer to react, and under these conditions deposition rate increases with flow rate. While this general behavior is repeated at different pressures, an increase in pressure leads to a decrease in deposition rate, owing to a lower in electron density [28].

Increasing the power density increases the deposition rate, as might be expected from the increased electron density. However, the plasma chemistry also changes which directly affects the rate of polymerization. Power dissipation depends uniquely on the reactor geometry and on the monomer under study.

To facilitate comparisons of one plasma with another Yasuda [29] has used a composite parameter W/FM, where M is the molecular mass of the monomer. The relevant dimensions are $(W/FM)\times 1.34\times 10^9$ J/kg, where W is the discharge power in Watts and F is in cm^3 (STP)/min. The most useful form of this parameter is at the critical power value W_c, defined as the power required to fill the reactor with the glow discharge. Above this power level, polymerization is independent of power and the pressure remains constant as the power is increased. The slope of plots of W_c/FM versus flow rate are approximately the same for different monomers containing the same functional group(s). The slope is highest for saturated hydrocarbons and W_c/FM is almost independent of flow rate for compounds with triple bonds or aromatic structure. The slope correlates with the quantity of H_2 produced in the plasma and is highest in the degradation of noncyclic saturated hydrocarbons. This has led to general classification [29] of hydrocarbons into three categories: compounds containing triple bonds in addition to aromatic structure; compounds containing double bonds and/or cyclic structure; noncyclic saturated hydrocarbons. This classification assumes unchanged reactor geometry.

Plasma polymerization discharges have been operated from d.c. to microwave frequencies (2.45 GHz). Probably the most commonly used frequency is 13.56 MHz. The disadvantage of d.c. plasmas is that extensive polymerization at the electrodes changes the discharge conditions. At high frequencies (13.56 MHz) the energetic charged species in the plasma are contained within the plasma itself. Polymer films which might be sensitive to highly energetic species are therefore best formed at high radiofrequencies.

The mechanism of plasma polymerization is not fully understood, but includes both homogeneous and heterogeneous reactions. A characteristic of plasma polymers is the presence of large concentrations of free radicals. This implies that free radicals are important in the polymerization process. More conventional polymerizations do not involve free radicals, which are normally quenched in the termination step. A free radical mechanism which explains the preponderance of these species is the rapid step growth principle. In RSGP, mono- or biradicals generated in the plasma can add to further monomer or other plasma molecules to form a new radical species, which can itself add further monomer or, if inactive, can be reactivated in the plasma and reenter the cycle [29]. In this scheme no distinction is made between gas phase or surface reactions.

Unlike conventional condensed-phase polymerization, which requires the monomer to contain one or more unsaturated functional groups, no such restriction applies to plasma polymerization and almost any organic molecule with a vapor pressure of a few hundred Pascals is amenable to this treatment. Monomers used include hydrocarbons, fluorocarbons, silane and other silicon compounds (siloxanes and silazines), and organometallic species.

1.3. Physical Vapor Deposition (PVD) Processes

The basic PVD processes fall into two general categories: sputtering and evaporation. The application of PVD techniques ranges from

Table 2. Classification of PVD processes

	Metals/alloys		Compounds	
Process	Basic	Hybrid	Basic	Hybrid
Conventional	evaporation deposition sputter deposition	ion plating sputter ion plating	direct evaporation direct sputtering	
Plasma-assisted (PAPVD)			activated reactive evaporation (ARE) reactive sputtering (RS)	reactive ion plating (RIP) biased RS (BRS)

decorative to high-temperature superconducting films. The thickness of the deposits can vary from 10^{-10} to 10^{-3} m. Very high deposition rates (25 µm/s) have been achieved with the advent of electron-beam-heated sources. A very large number of inorganic materials – metals, alloys, compounds, and mixtures as well as some organic materials – can be deposited by PVD technologies. "Ion plating" is a hybrid PVD process, i.e., an atomic film deposition process in which the substrate surface and/or the depositing film is subjected to a flux of high-energy particles sufficient to cause changes in the interfacial region between the film and the substrate, as well as in the properties of the deposited film compared with a nonbombarded film. These changes may be in the adhesion of the film to the substrate, film morphology, density, or stress. The source of the depositing species can be evaporation, sputtering, gases, or vapors (Table 2).

1.3.1. Sputter Deposition

Sputtering (Fig. 5) is momentum transfer from an indirect energetic projectile, such as an ion, to a solid or liquid target, resulting in the ejection of surface atoms or molecules. The target and the substrate are placed in the vacuum chamber and evacuated to a pressure of 10^{-2} – 10^{-5} Pa. The target (cathode) is connected to a negative voltage supply and the substrate generally faces the target. A glow discharge is initiated after an inert gas (usually argon) is introduced into the evacuated chamber. Typical working pressure is 2.5 – 20 Pa.

Planar Diode Glow Discharge Sputter Deposition. This is the simplest sputtering system, consisting of the cathode (target) and anode facing each other. The substrates are placed on the anode. The target, which is usually water cooled, performs two functions: as the source of coating material and as the electrode sustaining the glow discharge. The distance between cathode and anode is usually ca. 5 – 10 cm.

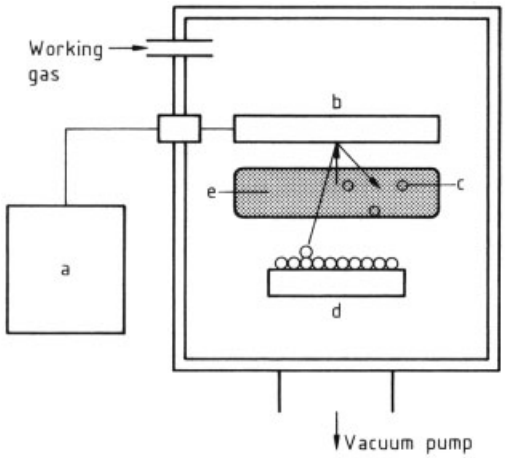

Figure 5. The sputter deposition process
a) Power supply; b) Cathode (target); c) Ion; d) Substrate; e) Plasma

The deposition rate is mainly determined by the power density at the target surface, size of the erosion area, source – substrate distance, source material, and working gas pressure. Some of these factors are interrelated, e.g., pressure and power density; optimal operation is obtained by controlling the parameters to get the maximum power flux which can be applied to the target without causing cracking, sublimation, or melting. The maximum power limit can be increased if the cooling rate of the target is increased by designing the coolant flow channels properly and improving the thermal conductance between the target and the target backing plate.

Even though planar diode glow discharge sputter deposition techniques are widely used, owing to their simplicity and the relative ease of fabrication of targets for a wide range of mate-

rials, they have several disadvantages: low deposition rate, substrate heating due to the bombardment of high-energy particles, and relatively small deposition surface areas.

Magnetron Sputter Deposition. By appyling magnetic fields to the diode sputtering process, the ionization efficiency can be greatly increased. In the magnetron sputter deposition process, an applied magnetic field parallel to the cathode surface forms electron traps and restricts the primary electron motion to the vicinity of the cathode. The magnetic field strength is of the order of 10^{-2} T and therefore itcan influence the plasma electrons, but not the ions. The electrons trapped on a given field line can advance across the magnetic field to the anode or walls via collisions (mostly with gas atoms). Therefore, their chances of being lost to the walls or anode without collisions are very small. Because of the higher efficiency of this ionization mechanism, the process can be operated at ca. 0.15 Pa with high current densities at low voltages, thus providing high sputtering rates.

There are several configurations of the magnetron sputter deposition technology. Figures 6, 7, 8 show the cylindrical magnetron, the planar magnetron, and the S-gun magnetron. The cylindrical magnetron is very useful to prepare uniform coatings over large areas, since long cathodes are employed. Furthermore, the cylindrical hollow magnetron technique is effective for coating complex-shaped objects. The cylindrical post magnetron can be used to avoid substrate bombardment by energetic particles, thus preventing heating of the substrate. Metallic and dielectric films can be deposited at higher deposition rates by planar magnetron sputtering than by diode sputtering. To deposit films on thermally sensitive substrates, such as electronic devices, the S-gun magnetron can be used, since this technique allows good isolation of the substrate from the plasma.

The balanced magnetron was developed primarily for metallization in microelectronic applications where electron/ion bombardment of the growing film is to be avoided. However, for metallurgical applications, ion bombardment of the film is desirable for the changes it produces in its structure and properties. Often, a separate ion gun is incorporated into the system to ion bombard the growing film. The development of the unbalanced magnetron has resulted in a practical way to carry out ion bombardment without the extra ion gun. Figure 9 shows the balanced and unbalanced magnetrons. The latter incorporates a controlled leak into the magnetic trap, permitting magnetic field lines as well as elec-

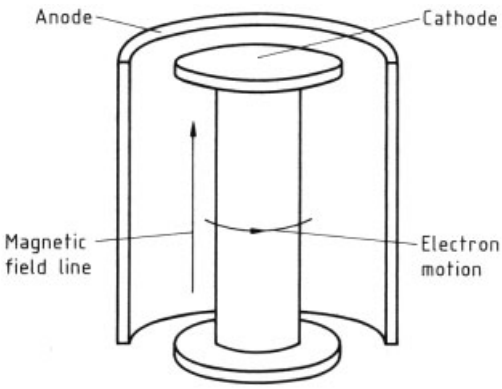

Figure 6. A cylindrical magnetron sputtering source

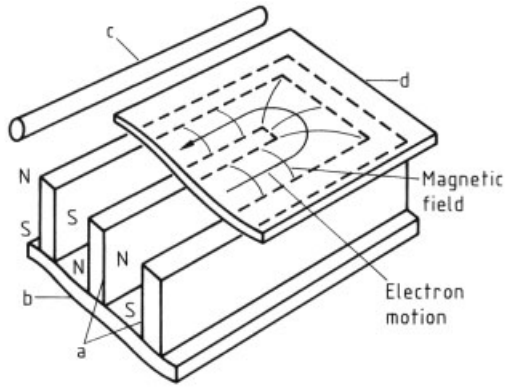

Figure 7. Planar magnetron sputtering sources
a) Permanent magnets; b) Pole piece; c) Anode; d) Cathode

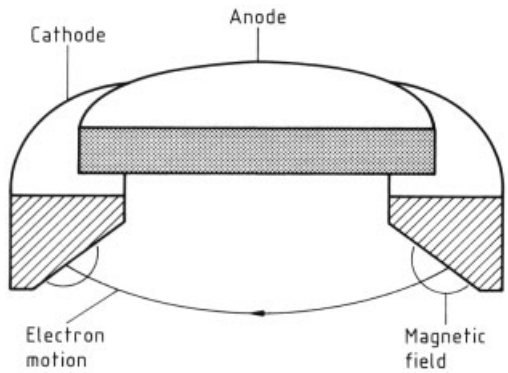

Figure 8. S-gun magnetron sputtering source

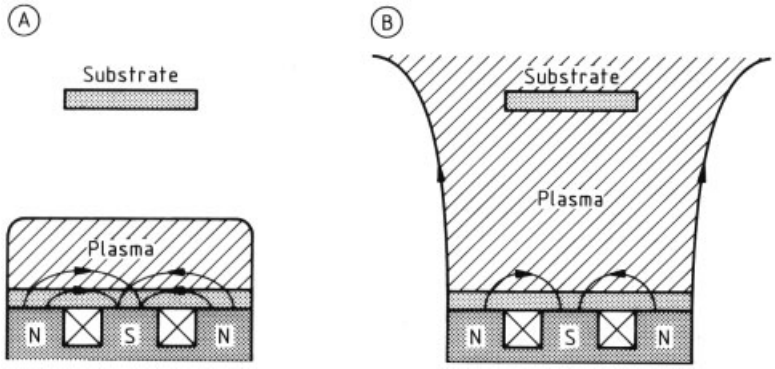

Figure 9. A) Balanced magnetron; B) Unbalanced magnetron

trons and ions to stream toward the substrate. A plasma is thus present near the negatively biased substrate from which positive ions can bombard the growing film.

Another variation is ion beam sputtering where ions of controlled energy from an ion source impinge on a target. The advantages are excellent adhesion; a resulting high-purity deposit because of the low operational pressure (ca. 0.015 Pa); and very low substrate heating effects since the substrate is not in contact with the plasma, as is the case for diode sputtering. If a second ion source is used to ion bombard the film, the process is called dual ion beam sputtering or ion-beam-assisted deposition. Optionally, an evaporation source can be used instead of the sputtering source to generate the vapor species.

Radiofrequency Sputter Deposition. Deposition can be performed at considerably lower pressures, e.g., 0.5 – 2 Pa, in r.f. sputtering compared with the planar d.c. discharge, since electrons oscillating at high frequency can obtain sufficient energy to cause ionizing collisions, and the number of electrons lost (without collisions) can be reduced.

Radiofrequency sputtering is widely used to deposit various kinds of conducting, semiconducting, or insulating coatings despite the complexity of the r.f. power source. This technique can also be applied to magnetron sputtering sources.

1.3.2. Evaporation Deposition

In the evaporation process, vapors are produced from a material located in a source which is heated by various methods (Fig. 10). The system consists of an electron-beam-heated evaporation source to vaporize the desired material. The substrates are located at an appropriate distance, facing the evaporation source. Resistance, induction, cathodic and anodic arc, electron beam, or lasers are the possible heart sources. The substrate can be heated and/or biased to the desired potential with a d.c./r.f. power supply.

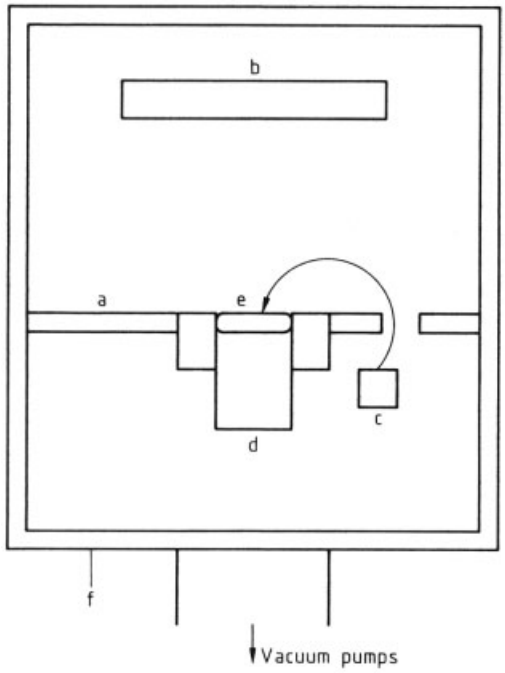

Figure 10. Evaporation/deposition process with an electron beam source
a) Separator plate; b) Substrate(s); c) Electron beam gun; d) Ingot rod; e) Molten pool; f) Vacuum chamber

Evaporation is carried out in vacuum at 10^{-3} – 10^{-8} Pa. In this pressure range, the mean free path is very large ($5\times10^2 - 10^7$ cm) compared with the source – substrate distance. Hence the evaporated atoms essentially undergo collisionless, line-of-sight transport prior to condensation on the substrate, thus leading to thickness buildup directly above the source and decreasing steeply away from it. At a source – substrate distance of 20 cm, the vapor pressure should be ca. 1.5 Pa in order to get reasonable deposition rates (ca. 0.1 μm/min). The source temperature should be adjusted to give this vapor pressure. Specialized evaporation techniques include flash evaporation, laser or electron beam ablation, cathodic arc evaporation (including filtered arc), and anodic arc evaporation.

In the flash evaporation process, pellets of alloy are dropped onto a very hot strip and are vaporized completely, thus maintaining the composition of the alloy in the deposit. It works very well for elements with high vapor pressures. A newer version of flash evaporation is pulsed laser or electron beam ablation where a high energy pulse vaporizes a small chunk of the target material, be it alloy or compound, thus retaining target composition in the deposit.

Cathodic arc evaporation which evaporates small volumes of the cathode (target) may also be used. However, the problem of droplet ejection from the target causing major flaws in the deposited film precludes the application of this process for thin film deposition. One way around the droplet problem is the use of a filtered arc where the ion stream from the cathodic arc plume is extracted by an electromagnetic field and the ions are condensed to deposit a film.

The anodic arc is similar to the cathodic arc except that the anode material evaporates. There is no particulate emission with the anodic arc process. It is as yet a laboratory-scale process.

1.3.3. Deposition of Metals, Alloys, and Compounds

The great versatility of the PVD processes is their ability to deposit a very large number of materials, including metals, alloys, semiconductors, superconductors, polymers, and to fabricate composites of various types (particulate, fibrous, or laminate).

Single-Element Deposition. This can be carried out by evaporation or sputter deposition processes. The deposition rate depends on the process and process parameters.

Alloy Deposition. Alloys consist of two or more components, which have different vapor pressures and hence different evaporation rates. As a result, the composition of the vapor phase and the deposit are constantly varying. The following methods have been used to obtain alloy films with stoichiometry close to the source composition by evaporation-based techniques.

1) Co-evaporation or co-sputtering using multiple sources: this technique involves simultaneous co-evaporation of the constitutive elements of the alloy. The composition of the deposited film is controlled by adjusting the evaporation/sputtering rate of the respective elements. In elaborate systems, separate deposition rate monitors are used with feedback networks to control the deposition rate from each source independently.
2) Evaporation from a single source: this technique involves evaporation of an alloy using a rod-fed electron beam source. Evaporation operates under steady-state conditions where the composition and volume of the liquid pool on the top of a solid rod are kept constant [30].
3) Flash evaporation: in this process, pellets of the alloy are dropped onto a very hot strip and are vaporized completely, thus maintaining the composition of the alloy in the deposit. It works very well for elements with high vapor pressure.
4) Sputter deposition from an alloy target.
5) Sputter deposition from a segmented target where the segments consist of each of the two components of the alloy and the ratio in the target of each element is inversely proportional to their sputtering yields.
6) Laser ablation from an alloy target.

Deposition of Compounds. Deposition of compounds can be performed in two ways: direct evaporation/sputtering of a compound target, where the composition of the target is the same as that of the compound to be deposited; and reactive evaporation, where the elements of the compound are evaporated/sputtered and re-

act with the reactive gas to form the compound [30,31].

1.4. Thin Film Growth by Molecular Beam Epitaxy (MBE)

1.4.1. Introduction [32–35]

Film growth by molecular beam epitaxy (MBE) is based on the formation of thermal, neutral, collision-free molecular (or atomic) beams which interact under ultra-high-vacuum (UHV) conditions with a heated single crystal substrate to form a thin epitaxial film. The beams are formed by effusion from Knudsen sources and the condition for molecular flow is that the mean free path of the atoms or molecules of the beam material contained in the source is greater than the diameter of the effusion orifice. Beams can be formed from elements or compounds, which may be solid, liquid, or gaseous at room temperature, and the design of the Knudsen source is critically dependent on the state of matter involved. It will also influence the pumping system used to achieve UHV.

MBE has been used to prepare films of a wide range of materials, including metals and alloys, semiconductors, superconductors (especially high critical temperature oxides), alkali halides, and other insulators, but by far the most important application is to semiconductor films, with some emphasis on III – V compounds and alloys.

The major advantage of MBE over other methods of thin film formation is that the growth processes can be controlled at the atomic level. This is achieved by the application of a range of in situ diagnostic techniques which can operate in the UHV environment.

1.4.2. The Configuration of MBE Growth Systems

The essential elements of an MBE system are illustrated in Figure 11. It comprises a group of Knudsen cells whose detailed design depends on the nature and state of the source material, a heated stage on which the substrate can be mounted, and a mechanical shutter interposed between each source and the substrate, whose function is to intercept the beam and effectively provide a switching capability. All components are contained within a UHV chamber.

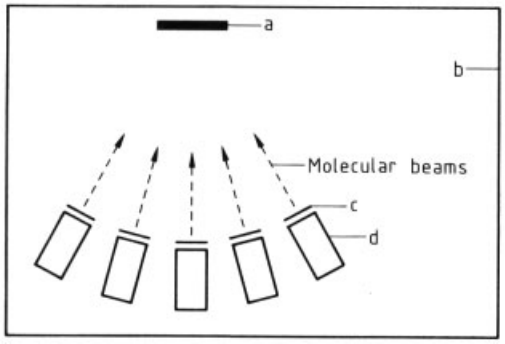

Figure 11. The basic arrangement of an MBE system
a) Monocrystal substrate; b) Ultrahigh vacuum chamber; c) Shutters; d) Knudsen sources (effusion cells)

Despite this apparent simplicity, implementation of the process requires a complex system, as illustrated in Figure 12. The important components are:

1) The UHV growth chamber, with associated pumping facilities
2) A vacuum interlock arrangement which enables substrates to be inserted and removed from the growth chamber without breaking the vacuum
3) A heated, rotatable substrate holder
4) A set of effusion cells and shutters, the operation of which is computer controlled

Process Environment. In general, systems are constructed from stainless steel, but all components should be capable of withstanding bake-out to 200 °C for outgassing. To achieve a base pressure in the growth chamber $< 10^{-8}$ Pa, the primary pumping for solid source systems is usually provided by a combination of sorption and ion pumps, perhaps with additional Ti sublimation and closed-cycle He cryopumps. Where a high gas throughput is involved, which is the case with gaseous or volatile liquid sources, trapped oil diffusion pumps, or more usually turbomolecular pumps are used instead of ion pumping. Secondary pumping takes the form of liquid-nitrogen-cooled cryopanels surrounding as much of the growth environment as possible and is used in all systems except those in which a toxic or pyrophoric condensible (at > 77 K) source gas is involved (e.g., SiH_4).

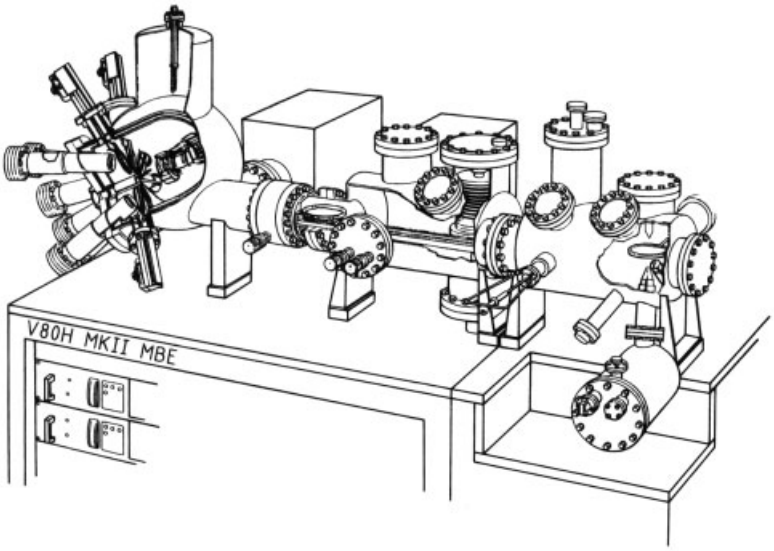

Figure 12. A solid source MBE system (VG Semicon, Fisons Instruments)

Vacuum Interlock and Substrate Preparation Chamber. It is critical that UHV conditions are maintained in the growth chamber as substrates are introduced and removed. A vacuum interlock system is therefore used in which the substrates are mounted on a movable holder, transferred from an entry lock into a preparation chamber, and finally into the growth chamber via a series of gate valves. The preparation chamber is used for the initial cleaning of the substrate surface, the final stage being carried out in the growth chamber. It is important that the substrate surface is clean at the atomic level and this is usually achieved by ex situ etching and in situ high-temperature treatment to desorb impurities, especially oxides.

Substrate Holder. Semiconductor substrates are typically circular wafers ca. 75–100 mm diameter and 0.25 mm thick. The function of the substrate holder is to provide a carriage for the wafers and to heat them uniformly. Growth temperatures are very dependent on materials, but are typically in the range 200–700 °C, most usually 500–700 °C. The holder is usually constructed from a refractory material block (most frequently Mo) heated by resistance or radiation from behind. In addition it must be capable of rotation within the UHV growth chamber at a speed comparable to the rate of growth of a monolayer of material, typically 1 s, so the rotation speed is ca. 1 Hz. The incident beams are neutral, and even with collimation they are necessarily divergent, so uniformity of thickness and composition over a reasonable area ($> 1\,\mathrm{cm}^2$) can be achieved only with substrate rotation. Substrate temperatures are usually measured with a calibrated optical pyrometer, but a thermocouple may also be incorporated in the holder block.

Beam Sources. Beam formation is at the heart of the MBE technique and the material from which the beam is to be formed determines the source design and also influences the choice of pumping system. In fact, MBE has become the generic term for a number of closely related techniques which are defined by the nature of the source material (Table 3).

Solid Sources. A typical design is illustrated in Figure 13. For low-vapor-pressure materials, such as Ga and Al, the operating temperature may reach 1200 °C. The crucible must therefore be refractory, of high purity, and chemically inert; the preferred material is pyrolytic boron nitride (PBN). The heater is a refractrory metal (usually W or Ta) wound noninductively and supported on PBN insulators; the cell is surrounded by Ta foil radiation shields. The flux is determined and regulated by the cell temperature, so measurement and control are crucial. Measurement is based on an optimally po-

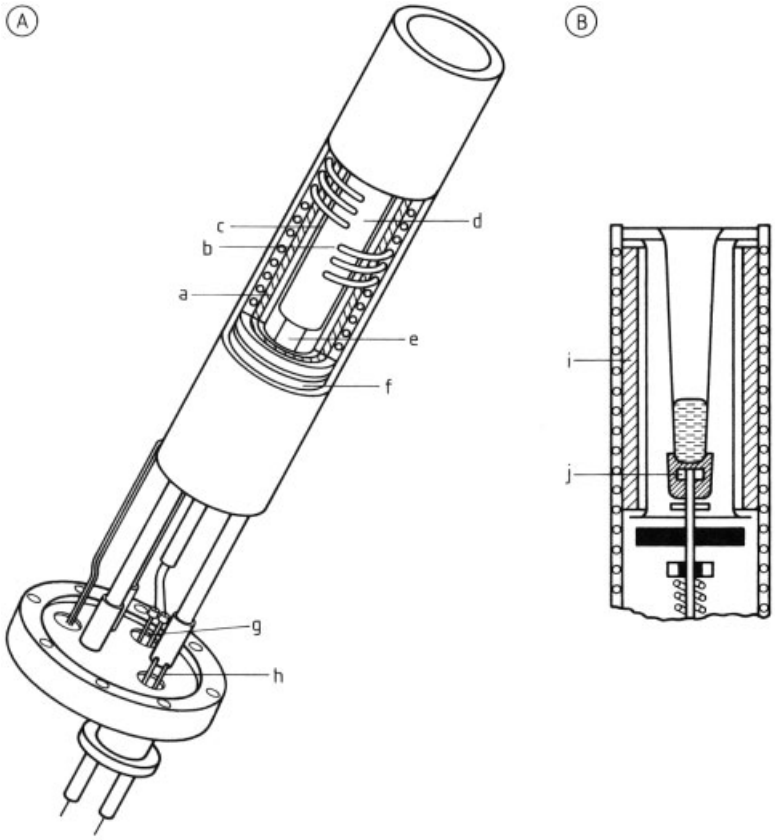

Figure 13. A) A typical Knudsen source for group 13 elements; B) Section through the tapered crucible
a) Foil tube; b) Water cooling; c) Heater foil; d) Crucible; e) Thermocouple; f) Multilayer radiation shield; g) Thermocouple feedthroughs; h) Power feedthroughs; i) Radiation shielding; j) Spring-loaded W – Re thermocouple

sitioned thermocouple, usually W – Re (5 % and 26 % Re) and control is established by proportional integral derivative (PID) controllers and thermocouple feedback, with computer control of the deposition sequence so that temperature – time profiles for each cell can be programmed, together with shutter sequencing and substrate temperature.

The flux J at a distance d (cm) from the cell is given by:

$$J = \left(1.118 \times 10^{22}\right) \frac{pA}{d^2\,(MT)^{1/2}} \quad \text{atoms(molecules) cm}^{-2}\text{s}^{-1}$$

where p is the pressure in Torr (1 Torr $\simeq$ 133 Pa) of material of molecular mass M in the cell, A is the area of the orifice, and T the cell temperature in K. Typical fluxes at the substrate surface are $10^{14} - 10^{15}$ atoms (molecules) cm^{-2} s^{-1}.

Group 15 elements (group numbering according to IUPAC recommendations) evaporate as tetramers (P_4, As_4, Sb_4), and molecular beams from a conventional Knudsen source are consequently also tetrameric. There is evidence, however, that certain electrical properties of films of III – V compounds may be improved by the use of dimeric group 15 species and these can be produced with a special source known as a cracker cell. In essence this is two cells in tandem; a beam of tetramers is formed in the first by normal evaporation from the element and this beam is transmitted into a second cell, referred to as a "cracker," which is maintained at a higher temperature and fitted with baffles to maximize the number of collisions of tetramer molecules with heated surfaces. Under these conditions, the tetramers dissociate to form dimers and the resultant beam can contain up to

Table 3. Processes covered by the generic term MBE, the source materials involved, and some applications

Process	Source materials	Applications
Molecular beam epitaxy (MBE)	generic name, but now used for solid source systems, e.g., metals, Si, Ge, and elemental group 13 and 15 sources	magnetic films, silicon, III – V compounds, e.g., GaAs, (Al, Ga)As
Metal-organic molecular beam epitaxy (MOMBE)	group 13 organometallic compounds, e.g., triethyl gallium, trimethyl indium, (mostly volatile liquids); elemental group 15 sources	III – V compounds, especially GaAs and $(In_xGa_{1-x})As$
Chemical beam epitaxy (CBE)	group 13 organometallic compounds; group 15 hydrides (PH_3 and AsH_3)	III – V compounds and alloys, especially P-containing materials, e.g., $(In_xGa_{1-x})P$
Solid source molecular beam epitaxy (SSMBE)	usually applied only to Si and Ge elemental sources with electron beam evaporation	Si and SiGe alloys
Gas source molecular beam epitaxy (GSMBE)	applied mainly to Si and Ge growth from their gaseous hydrides (SiH_4, Si_2H_6, GeH_4)	Si and SiGe alloys
Atomic layer epitaxy (ALE); migration-enhanced epitaxy (MEE); growth interruption (GI)	a range of methods characterized by pulsed sources, usually elemental, and applied to III – V compounds and alloys	most III – V compounds and alloys

100 % of the dimerized species, depending on the temperature of the cracker region.

Although it employs a solid source, silicon deposition from elemental silicon is an exception, since it is referred to as Si MBE (or SSMBE), but is not strictly a molecular beam technique. It relies on the evaporation of Si from a Si boule, the central region of which is heated to ca. 2000 °C by a focused electron beam. This source temperature is necessary to obtain an adequate vapor pressure, but molten Si (*mp* ca. 1420 °C) is extremely corrosive and can only realistically be contained by solid Si. The arrangement is to mount the Si boule in a water-cooled copper hearth, which allows a solid outer region to be maintained, even though the central part is molten. The rest of the system is the same as for conventional MBE.

Gas and Volatile Liquid Sources. Although solid sources have provided the basis for MBE technology, volatile liquid and gaseous sources have important areas of application. In compound semiconductor technology, organometallic precursors such as trimethyl gallium (TMGa) or triisopropyl gallium (TiPGa) which in general are volatile liquids, are frequently used as the cation source while the gaseous group 15 hydrides (AsH_3, PH_3) can provide an anion source. They have the advantage of continuous replenishment from outside the UHV system, and operate at or just above room temperature, so potentially generate less contamination than a high-temperature solid source. This low-temperature concept is even more important for Si technology, where the very high-temperature electron beam evaporation process is being superseded by the use of gaseous hydrides (SiH_4, Si_2H_6). With both volatile liquids and gases, fluxes are normally controlled by mass flow meters, even though it is ultimately the pressure behind the cell orifice which determines the magnitude of the flux.

An additional feature in the growth of III – V compounds from group 15 hydrides is the necessity to dissociate (crack) them before they impinge on the substrate. The technology is similar to that used to produce dimer beams from group 15 elemental solid sources, but in this case the reason for predissociation is simply that the hydrides are too stable thermally to allow growth to occur directly.

The use of nonsolid sources has generated a terminology appropriate to each type and for reference to the literature it is important to define them, as indicated in Table 3.

Pulsed (Noncontinuous) Sources. Several modifications to the process of continuous deposition, particularly of III – V compounds and alloys, have been introduced, all based on interruption of growth. The modifications were introduced either to produce more perfect interfaces between layers of different composition or to grow material with superior electrical and optical properties at lower temperature. In the simplest version of growth interruption, the shutter of the group 13 element source, but not that of the group 15, is closed briefly (up to 120 s) to allow the surface to smoothen before deposition of the next layer is started.

Atomic layer epitaxy (ALE) and migration-enhanced epitaxy (MEE) involve periodic interruptions of both fluxes during growth; ALE

is based on an alternate supply of constituent elements to the substrate surface, regulated so that a complete monolayer is provided with each pulse. The adsorption of group 15 molecules is self-regulated by the group 13 atoms in the surface, but the latter can accumulate as the free element if the amount per pulse is too large. The MEE process also involves the alternating supply of cation and anion fluxes, and it is claimed that by reducing the surface concentration of the group 15 element during the impingement of the group 13 atoms, their surface migration rate is enhanced, resulting in smoother, higher quality growth at lower temperatures.

1.4.3. Monitoring Facilities

The vacuum environment of MBE growth allows a number of powerful diagnostic techniques to be used to monitor various stages of the process. The vacuum quality can be checked with a mass spectrometer to obtain quantitative information on the partial pressures of the background gases; characteristic patterns are observed which may indicate the presence of small leaks, or other contamination, e.g., an oxidized As_4 source. The total pressure in the system can be measured with a Bayert – Alpert ion gauge, and a second such gauge placed in the substrate position can be used to quantify the flux of atoms from the growth cells, manifest as a rise and fall in pressure when the shutter in front of each cell is opened and closed. The flux of dopant atoms is usually too small to be detected in this way, and calibration of these cells is performed indirectly, e.g., by measuring the electrical properties of doped layers.

A technique which can provide valuable information on the MBE growth process is reflection high-energy electron diffraction (RHEED). As shown in Figure 14 a beam of electrons of en-

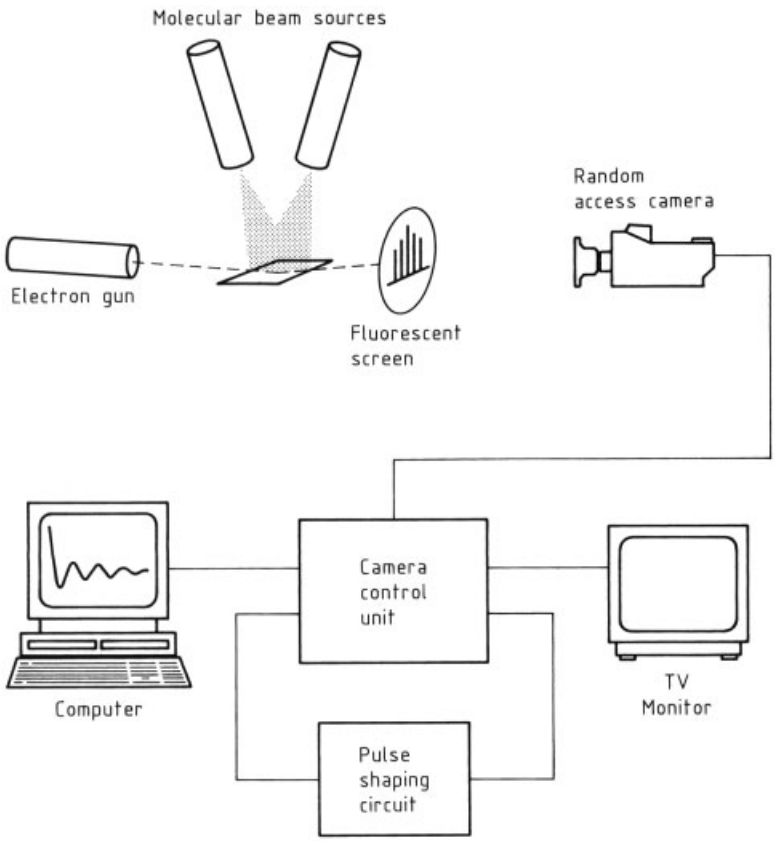

Figure 14. A RHEED system used for monitoring MBE growth

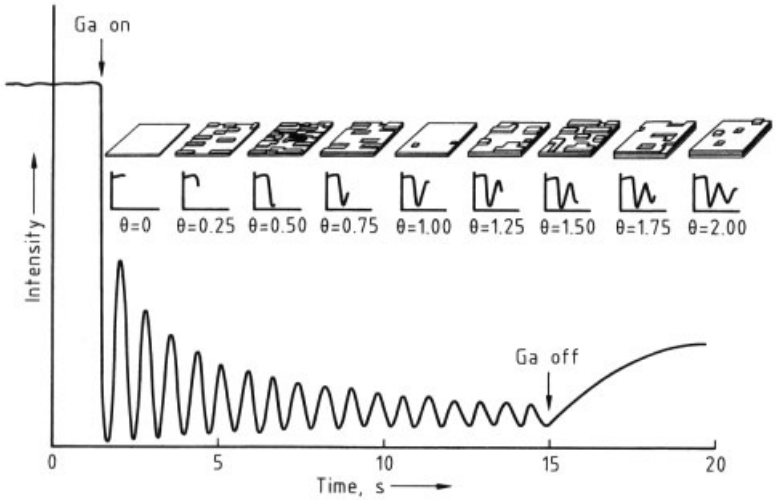

Figure 15. Two-dimensional, monolayer-by-monolayer growth and the corresponding RHEED intensity modulation Θ is the coverage in monolayers.

ergy ca. 10 – 25 keV is incident at glancing angle (ca. 0.5 – 4°) on the substrate, and the diffracted beams strike a phosphor screen on the opposite side of the growth chamber. The diffraction patterns produced are related to the surface structure of the substrate (or deposited layer) and can be used to detect the structure of the surface during substrate cleaning, the presence of three-dimensional growth modes (rather than the desired monolayer-by-monolayer deposition), and the occurrence of modifications to the configuration of the surface atoms, the so-called surface reconstructions. In addition, monitoring the intensity of the specularly reflected beam allows a very accurate determination of the growth rate. This is possible because the deposition of each monolayer occurs by a "nucleation plus coalescence" process which results in an alternating roughening and smoothing of the surface (Fig. 15) which produces a corresponding intensity modulation of the specularly reflected beam at various stages in the growth. Since electrons are scattered out of the specular beam as the surface roughens, the intensity oscillates with a period equal to the time taken to deposit one monolayer of material; thus, not only can the number of individual monolayers deposited be counted, to give a very precise measure of thickness, but by timing the oscillations, an accurate measure of the growth rate is obtained.

1.5. Thin Films by Adsorption Processes

The outermost nanometer of a solid governs many interfacial poperties, including wettability, friction, and electrochemical reactivity. Very thin films of foreign materials on the surface may alter these properties. Consequently, a widely used method of controlling interfacial chemistry is to adsorb a monolayer of a surface-active agent (surfactant) onto the surface. Surfactants are organic molecules, normally adsorbed from solution, although vapor-phase adsorption is also possible.

Adsorbed monolayers are classed as physisorbed or chemisorbed, according to the nature of the interactions attaching the molecules to the surface. Although this division is useful, the distinction between chemisorption and physisorption is not sharp. Physisorbed films are stabilized only by van der Waals or dipolar interactions. Since the interaction with the surface is weak, the monolayers do not normally retain their integrity when the solid is removed from the adsorption solution. A practical example is the use of detergents to enhance the wettability of hydrophobic surfaces. In polymeric adsorbates, the relative weakness of the interaction with the surface is compensated by multiple interactions involving different parts of the polymer chain. The suppression of the growth of wax crystals

in diesel fuel is one important use of physisorbed polymers.

In chemisorbed films, a covalent or ionic interaction exists between the surfactant and the solid. This specific interaction with the solid increases the stability of the film and helps prevent multilayer formation, both of which lower the amount of additive required. Surfactants used in aqueous solution frequently also contain one or more long hydrocarbon chains. Hydrophobic interactions between the alkyl chains and water favor aggregation of the surfactant in solution and at a surface; van der Waals interactions between the hydrocarbon chains in the monolayer contribute additional stability to the adsorbed film.

In commercial applications, mixtures of surfactants are invariably used and the optimum adsorption conditions are developed empirically. The factors governing the choice of surfactant may be illustrated by the practical problem of controlling wettability. In the recovery of ore by flotation, surfactants are required that adsorb selectively to ore particles and make their surfaces hydrophobic. Aqueous solutions of sodium ethylxanthate ($C_2H_5OCS_2^-Na^+$) are commonly used for sulfide ores. It is believed that the monolayer is stabilized by a polar covalent bond between the mineral surface and the sulfur atoms of the surfactant. Metal oxides are more ionic than sulfides. In the flotation of oxide minerals, electrostatic interactions between the surface and an ionic surfactant are exploited. However, ionic interactions alone are insufficient for the formation of stable monolayers, and long hydrocarbon chains are also required. For example, dodecylammonium acetate is used in the flotation of negatively charged oxides, such as goethite and quartz. Conversely, anionic surfactants are used at low pH when the mineral surface is positively charged. Similarities in the chemical structure of the surfactant head group and the solid can be exploited, e.g., sodium stearate is used to make calcite hydrophobic. Typical surfactant concentrations are $10^{-5}-10^{-3}$ mol/L.

Surfactants can be deposited from dispersions as well as solutions. Double-chained cationic surfactants, such as $[CH_3(CH_2)_{17}]_2(CH_3)_2N^+Cl^-$, are widely used as fabric conditioners, even though they are virtually insoluble in water. These molecules are present in solution as lamellar mesophases and are adsorbed onto the surface as a result of electrostatic interactions.

Adsorption of monolayers from nonaqueous media requires surfactants that are soluble in organic solvents. A classical example is the adsorption of stearic acid from oil onto iron. The resulting densely packed monolayer acts as an efficient lubricant. Oil-soluble corrosion inhibitors, such as long-chain imidazolines, are highly efficient, even when present in monolayer quantities on the surface of steel.

The silanes are another important class of surface modifiers. Silanes, such as $NH_2(CH_2)_3Si(OCH_3)_3$ (APS), are used as coupling agents to promote adhesion of polymeric resins to mineral surfaces [36]. Hydrolysis of the methoxy groups yields free silanols which condense with hydroxyl groups on the surface to form strong covalent Si – O – Si bonds. The terminal functional group is free to couple to the resin: e.g., the amino group of APS can react with polyamide resins. Reactions between silanols strengthen the film by cross-linking, but also lead to polymerization, so the resulting thin films are rarely monomolecular – ten monolayers is a more typical thickness. Silane coupling agents can be applied, either from dilute aqueous solutions or as a minor component of a resin.

1.5.1. Self-Assembled Monolayers

Self-assembled monolayers (SAM) are densely packed monolayers formed by adsorption; the term is usually restricted to structurally well-defined monolayers that can be removed from the adsorption medium and characterized by a range of analytical techniques. The molecular architecture and physical properties of SAMs can be varied in a controlled way, in contrast to most of the systems used in commercial processes. Two types are most important: alkanethiols on metals and alkyltrichlorosilanes on oxides.

Long-chain alkanethiols of the generic structure $X(CH_2)_nSH$ adsorb on gold, silver, platinum, and copper (as well as on gallium arsenide and probably other materials) through the formation of a metal thiolate [37]. The chain length n is typically 10 – 21 and the tail group X can be any functional group chemically compatible with a thiol. The monolayers are stabilized by a strong metal – sulfur bond and by van der Waals inter-

actions between all-*trans* hydrocarbon chains. The tilt of the chains from the surface normal depends on the lattice parameters of the solid and varies from substrate to substrate: 30° on Au (111), 12° on Ag (111), and 57° on GaAs (100).

A SAM on gold can be prepared by immersing a freshly prepared gold mirror in a 10^{-3} mol/L solution of a thiol in degassed ethanol (Fig. 16). Adsorption is largely complete within the first few seconds, but a slow period of consolidation follows and many hours of adsorption may be needed before the limiting structure is reached. Multiple functional groups can be incorporated into SAMs by coadsorption of two or more thiols. The composition of a mixed monolayer does not in general reflect the composition of the adsorption solution and must be determined separately. In the study of interfacial processes such as wettability, surface reactivity, and electron transfer, SAMs of thiols on gold have been extensively exploited as model systems. Practical applications as electrochemical sensors and monolayer resists in microlithography are also being developed.

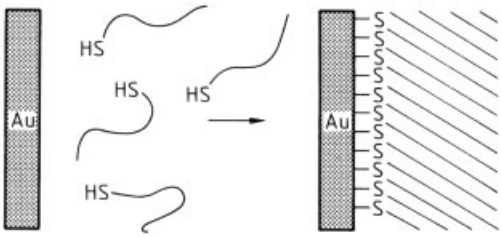

Figure 16. Formation of a self-assembled monolayer of an alkanethiol on gold

Alkyltrichlorosilanes ($X(CH_2)_nSiCl_3$) are similar to the silane coupling agents, but less prone to multilayer formation [38]. Elimination of HCl leads to the formation of siloxane bonds to oxide surfaces, in particular to silica. A substrate is typically immersed in a 0.1 % solution of the trichlorosilane in a 3:1 hexadecane/CCl_4 mixture at room temperature for 2 min, although the adsorption time depends critically on the amount of water present. SAMs of alkyltrichlorosilanes are densely packed and the hydrocarbon chains are oriented approximately perpendicular to the surface. One particularly versatile molecule is octadecyltrichlorosilane (OTS). An OTS monolayer on glass is 2.5 nm thick, chemically and thermally stable, highly hydrophobic, and resistant to chemical contamination.

1.5.2. Self-Assembled Multilayers

Multilayers can be built up in a controlled fashion by sequential adsorption of self-assembled monolayers. These multilayers are being developed principally for applications in nonlinear optics and optoelectronics, which require oriented thin films of molecules with large nonlinear susceptibilites. More than 25 layers have been built up by the following route at Eastman Kodak (Fig. 17).

An alternative approach, developed at the University of Texas and at AT & T Bell Laboratories, uses layers of alkyldiphosphonates ($^{-2}O_3P(CH_2)_nPO_3^{2-}$) held together by sheets of zirconium ions (Zr^{4+}) [39].

1.6. Langmuir – Blodgett and Related Techniques

Monomolecular films have been a subject of scientific interest for more then a hundred years, but it was not until 1917 that LANGMUIR introduced the film balance for the study of molecular films on water surfaces [40], the basis of today's experimental and theoretical concepts. He demonstrated that fatty acids of different chain lengths form insoluble monolayers at the air – water interface and that they occupy the same cross-sectional area. Therefore they are oriented perpendicular to the water surface. The transfer of the monomolecular film from the water surface onto a solid substrate provides a method of preparation which is now called the Langmuir – Blodgett (LB) technique, first published by BLODGETT in 1935 [41].

The LB technique is commonly used to produce highly ordered ultrathin layered films of monomeric or polymeric materials. The term "LB multilayers" is frequently used to describe the result of the assembly technique. The LB technique offers a number of advantages:

1) Film thickness can be controlled directly by the number of monolayers transferred.

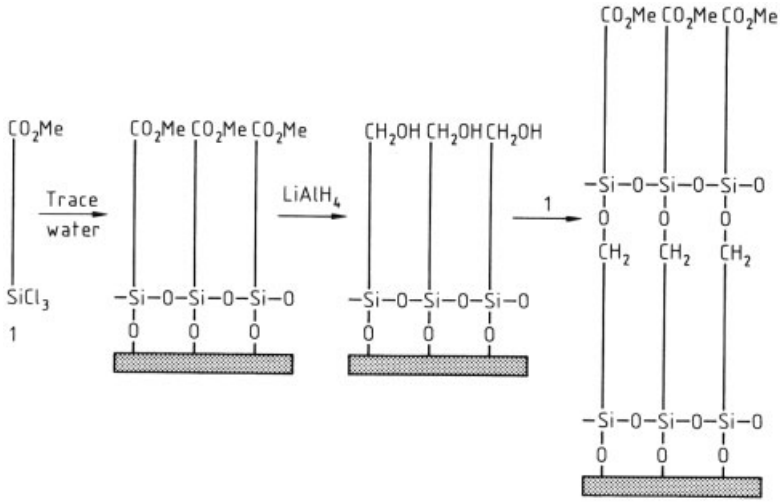

Figure 17. Synthesis of organized multilayers of alkyltrichlorosilanes [38]

2) The orientation of the molecules in each layer can be precisely controlled.
3) The molecular species in each layer can be varied at will. This is of particular significance with regard to the bottom and the top layers of a multilayered assembly. The bottom layer is responsible for the interaction between film and substrate whereas the top layer determines the surface characteristics, e.g. wetting.
4) Control of the molecular interactions within the film allows optimization of properties such as conductivity or transparency.

Classical substances for the LB technique are amphiphilic molecules consisting of a long alkyl chain with a hydrophilic head group. Films are prepared by first depositing the amphiphilic material, dissolved in a volatile solvent immiscible with water, on the surface of water (subphase) contained in a trough. The subphase temperature is generally controlled by a thermostat. When the solvent has evaporated, the organic molecules may be compressed by a floating barrier to form a floating "two-dimensional" solid. Owing to their hydrophobic head groups, individual molecules align themselves during this process. Once the monolayer is compressed at a suitable constant surface pressure, the molecules can be transferred onto a hydrophilic substrate by dipping the substrate slowly into the subphase [42, 38]. Deionized water or dilute aqueous solutions of electrolytes have mainly served as subphase, although mercury, hydrocarbons, or glycerol have also been studied [43].

Depending on the substance and the substrate, transfer may occur during either the downstroke or the upstroke, or during both (Fig. 18). Three arrangements of amphiphiles in the multilayers (X-, Y- and Z-films) are possible (Fig. 19, see next page).

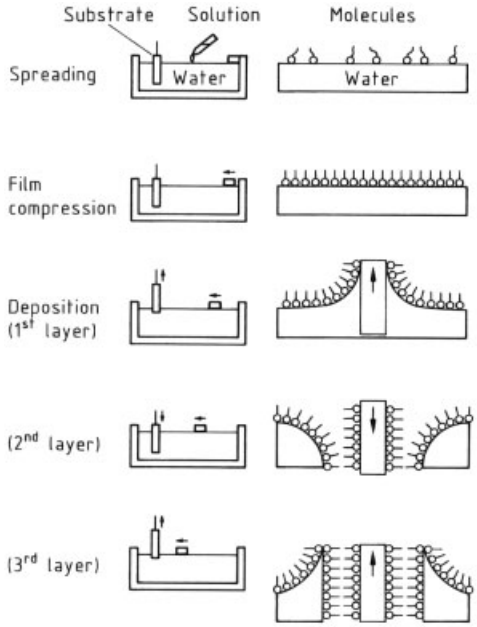

Figure 18. Spreading, compression, and deposition of amphiphilic molecules on the subphase

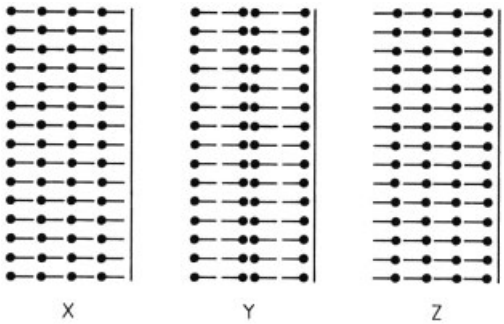

Figure 19. Structure of X-, Y-, Z-type amphiphilic multilayers

During compression, the monolayer undergoes a number of phase transformations. The different phases are in some respects two-dimensional analogs of gases, liquids, and solids, and can be identified by measuring the surface pressure of the monolayer as a function of the area per molecule. Figure 20 shows isotherms of an amphiphilic diacetylene monocarbonic acid at three different temperatures [44].

For the 5 °C isotherm, at an area of 0.33 nm^2 per molecule the first phase transition occurs to form the "liquid expanded" state. In this phase, the monolayer is coherent, except that the molecules occupy a larger area than in the "liquid condensed" state, which is formed at ca. 0.25 nm^2 per molecule. In the condensed phase, the molecules are closely packed and uniformly oriented. If further pressure is applied, the monolayer collapses and a sharp decrease in the pressure is observed. Frequently, amphiphilic substances show different condensed phases which differ by the mode of packing in two dimensions, depending on the applied surface pressure and temperature. Phase diagrams have been proposed which show the influence of temperature and chain length on the existence range of certain phases [45]. A specific state of packing of

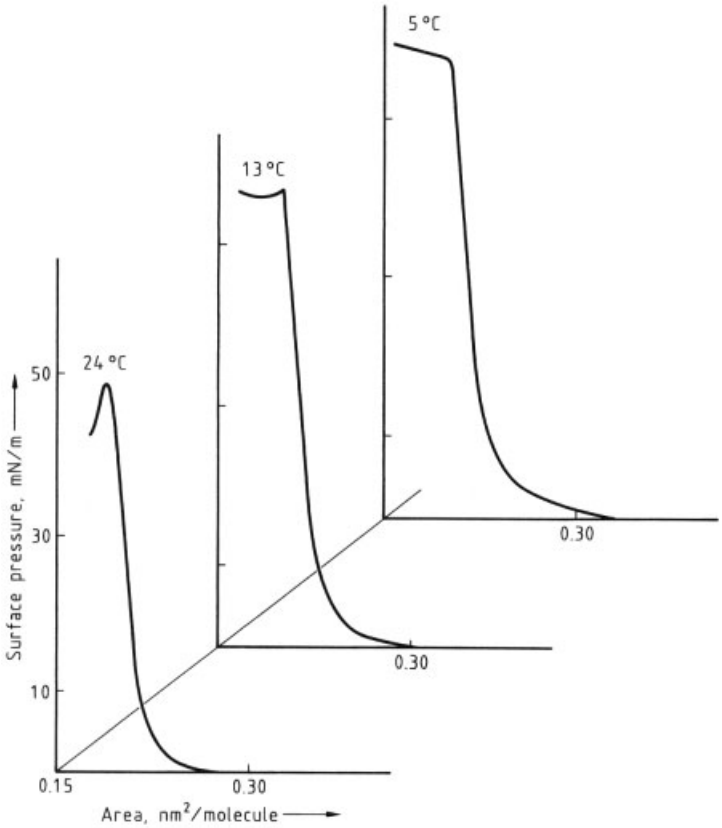

Figure 20. Surface pressure – area isotherms of 12,8-diacetylene monocarboxylic acid, $C_{12}H_{25}-C{\equiv}C-C{\equiv}C-(CH_2)_8-COOH$, at the air – water interface on 10^{-3} mol/L $CdCl_2$ subphase, pH 6.1 at three temperatures

the molecules on the substrate can be obtained by choosing suitable transfer conditions (temperature, surface pressure, etc.).

Molecular Design. Several factors affect the suitability of a substance for the formation of a monolayer:

Solubility
Volatility
Interaction between molecules
Interaction with the subphase
Chemical stability under spreading and transfer conditions

The deposition of the floating monolayer can be hindered if the affinity of the film to the substrate is too weak. The formation of multilayers is impossible if the substance lacks affinity to itself. Molecules which are able to form monolayers at the air – water interface can be divided into three classes:

1) Low molecular mass substances
2) Amphiphilic polymers
3) Polymers with a rod-like backbone and flexible side chains (hairy rods)

1.6.1. Low Molecular Mass Substances

Most of these substances with high affinity to the water subphase consist of a hydrophilic head group and at least one long alkyl chain. The classical film-forming substances are fatty acids [46], their derivates and salts, but a number of other substances have been found to form LB layers (Table 4).

Hydrophilic head groups can be formed by nearly all kinds of polar functional groups. The amphiphiles often behave like the corresponding fatty acids.

Aromatic compounds, some of which contain heteroatoms, as well as anthracene derivatives, have been modified by polar groups to form amphiphile head groups. Liquid crystalline (mesogenic) compounds inherently form layered structures; their incorporation increases the thermal and mechanical stability of a two-dimensional fluid phase. Properties of LB films such as selective binding to particular surfaces or conductivity (Section 2.6.1) can be achieved by attaching long alkyl chains to the suitable functional moieties; LB layers containing dyes and ionophores have been prepared. On the other hand, ferrocene is an example of a compound which forms LB layers, if polar substituents are attached.

Lipids are the matrix-forming ingredients of biological membranes and are therefore an important class of LB layer materials. These molecules often show complicated behavior at the air – water interface, due to the presence of two hydrophobic tails.

Porphyrins and phthalocyanins are examples of more complex LB layer molecules. They can carry a variety of functional groups, some of which deviate from the normal requirement of having well-defined head and tail groups. This deviation is reflected in their monolayer properties; the molecules form multilayer stacks in which the structure and stability of the film is governed by molecular packing rather then by hydrophilic – hydrophobic interactions.

For some amphiphiles it may be necessary to add certain ions to the subphase in order to ensure good deposition behavior by salt formation. For example, bivalent ions, e.g., Cd^{2+}, in the subphase are sometimes needed to achieve transfer of fatty acids from the water surface onto solid substrates. They are then deposited as Cd salts.

Monolayers can also be built up by a mixture of different molecules; LB films containing molecules which do not themselves form stable films can be prepared from a mixture with a suitable partner.

1.6.2. Polymeric Amphiphilic LB Layers

Thermal and structural stability is limited by the melting temperature of the compounds. Thermal desorption occurs readily above the melting temperature. The stability can be enhanced by using polymers. Such polymers can be obtained from amphiphilic monomers by in situ polymerization of monolayers or by spreading preformed amphiphilic polymers. The functional groups active toward polymerization of amphiphilic monomers are typically diacetylenes [47], acrylates, methacrylates, or cinnamic acid derivatives (Table 5).

Polymerization can be performed on the water surface or in the deposited film. It is gener-

Table 4. Amphiphiles forming Langmuir – Blodgett films

	Hydrophobic	Hydrophilic
Fatty acids and long-chain derivatives	C_nH_{2n+1}, $n = 12-22$	$-COOH$ $-CH_2OH$ $-CN$ $-CONH_2$ $-NHCOCH_3$ $-CH=NOH$
Aromatic and long-chain compounds	$C_{16}H_{33}-C_4H_6$ — OH $C_{18}H_{37}$ — $O-C_4H_6-NH_2$ C_4H_9 — CH_2OH $C_{12}H_{25}$ — $N(CH_3)$ — $N=N$ — $COOH$, NO_2	
Others	$C_{18}H_{37}$ — OCH_2 — (18 crown 6) Ferrocene — $C_{18}H_{36}$ — $CONH_2$	
Lipids	$C_{15}H_{31}$ — $C(=O)$ — $O-CH_2$ $C_{15}H_{31}$ — $C(=O)$ — $O-CH$ $CH_2-O-P(=O)(O)-OR$	
Porphyrins and phthalocyanins	$HOOC-HC(C_{18}H_{37})$, $CH(C_{18}H_{37})-COOH$ $HOOC-HC(C_{18}H_{37})$, $CH(C_{18}H_{37})-COOH$	$C(CH_3)_3$ $(CH_3)_3C$ $C(CH_3)_3$ $C(CH_3)_3$

ally initiated by irradiation or by the addition of radical-forming substances which can be excited directly or via energy transfer from sensitizers.

The topochemical polymerization of diacetylenes proceeds within LB layers with retention of the crystalline order and the layer structure. The active group to be polymerized can be located on either end of the monomer: near the hydrophilic end or in the hydrophobic part. Oxiranes undergo very little structural change during their ring-opening polymerization.

In polycondensation, the constituent molecules join together with the elimination of another molecule. One of the few examples in the present context is the reaction of a monolayer of octadecanal $C_{17}H_{35}CHO$ with a preformed polymer (poly-L-lysine) which is initially dissolved in the subphase.

Preformed amphiphilic polymers have alternating hydrophilic and hydrophobic segments in different parts of the polymer. Spacer groups can be used to decouple the motions of the main

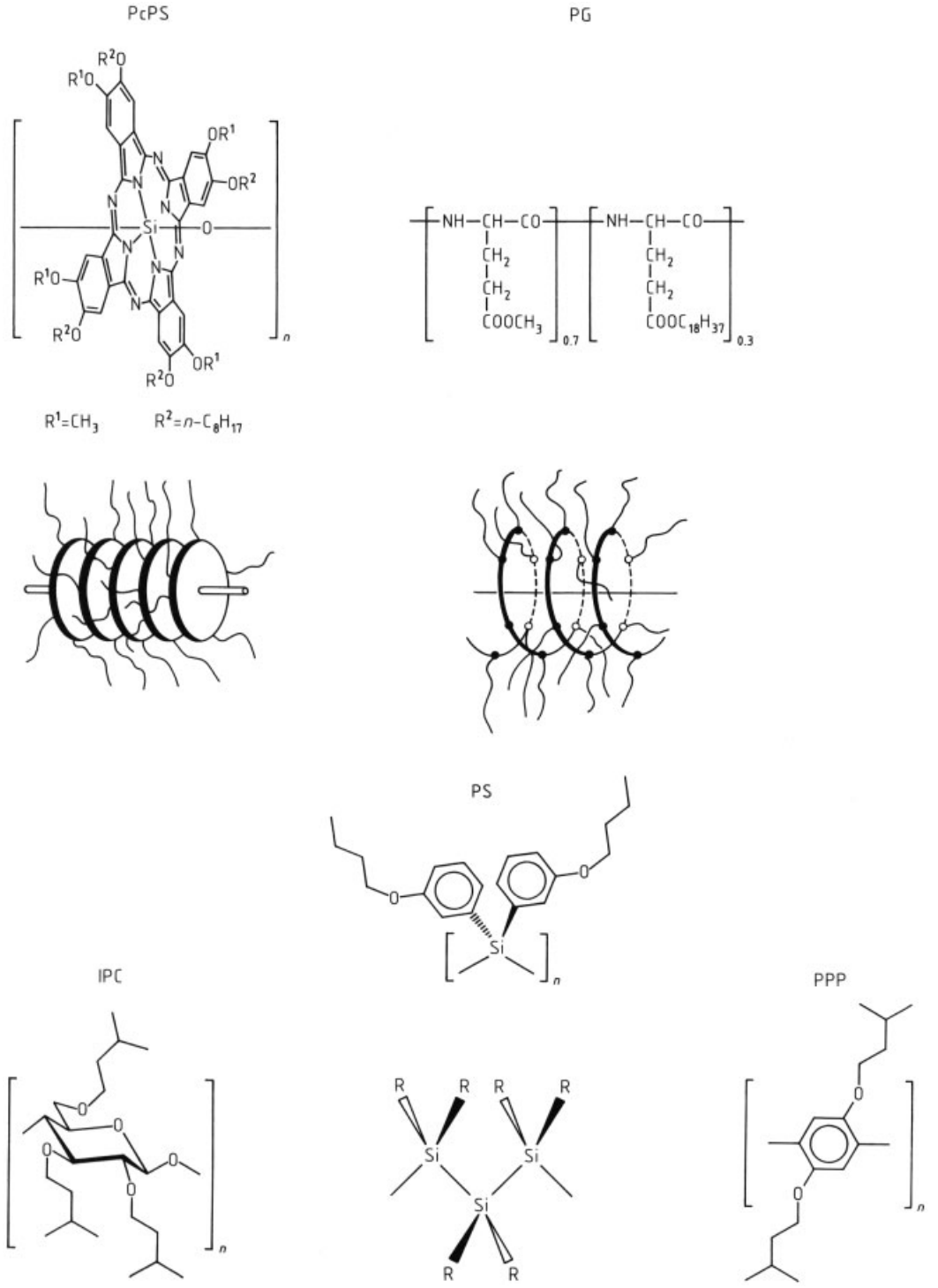

Figure 21. Typical hairy rod molecules
PcPS = phthalocyaninato-(poly)siloxanes; PG = polyglutamate; PS = polysilane; IPC = Isopentyl cellulose; PPP = poly(*p*-phenylene)

chain from that of the amphiphilic side groups. Polymer analogous reactions, if carried out in the deposited multilayer system, may be used to prepare thin films of insoluble polymers, e.g., poly(imides) and poly(*p*-phenylenevinylenes), on solid substrates.

Table 5. Polymerizable amphiphilic monomers forming Langmuir – Blodgett films

	Hydrophobic	Hydrophilic
Amphiphiles containing double bonds	C_nH_{2n+1}, $n = 10-22$	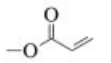
		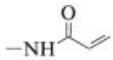
		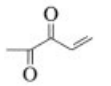
		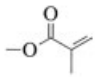
		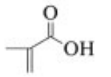
	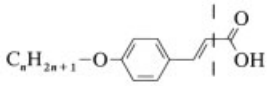	
Diacetylenic amphiphiles	$CH_3(CH_2)_n-C\equiv C-C\equiv C-(CH_2)_m$	
Oxiranes	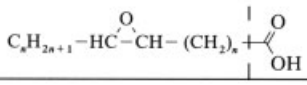	

1.6.3. Hairy Rod Polymers

Rod-like polymers can not be classified according to the rules developed for the behavior of amphiphilic compounds at the air – water interface. They exhibit cylindrical symmetry, with low-polarity segments distributed equally over the whole molecule (Fig. 21, see previous page). These polymers lie flat on the surface. It is essential to surround the rod-like polymers with conformationally mobile side chains. This provides fluidity of the monolayer at the air – water interface and screens direct backbone interactions. The alkyl chains act as chemically attached solvent molecules. Transferring monolayers of hairy rods onto solid substrates aligns the stiff polymer backbones parallel to the dipping direction, producing highly ordered multilayers (Fig. 22) [48].

Typical examples of such polymers are phthalocyaninato-(poly)siloxanes (PcPS), several polyglutamates in the α-helical conformation, cellulose alkylethers, specifically substituted polysilanes, and poly(*p*-phenylenes).

Film Deposition. Figure 23 is a schematic diagram of a typical LB trough. The bath (a) is usually made of PTFE or PTFE-coated material. A moving barrier (b) allows control of the pressure applied to the monolayer. The pressure-control device receives information from a pressure sensor (e) on the water surface. A motor (d) is used to lower and raise the solid substrate (e). The trough is temperature controlled.

Two methods can be employed to measure surface pressure. In the Wilhelmy technique, the surface tension is measured by weighing a hydrophilic plate which is dipped through the monolayer into the aqueous subphase. In the Langmuir method, the difference in surface pressure is measured between the pure air – water surface and the surface covered with the film.

The transfer of monolayers requires a constant surface pressure. Thus, during transfer, the barrier moves continuously toward the substrate.

Other Types of LB Trough. The circular trough (Fig. 24) [49] has a simple reliable motor drive for the barriers (linear troughs require special provision to convert the circular motion of the motor to a linear one). The multicompartment circular trough has two independent drive barriers enclosing the monomolecular film, so that molecules can be compressed as well as being transported from one compartment to another. This allows, for example, a change of subphase while the trough is in use. The texture of the films transferred from the trough reflect the state of order of the monolayer at the air – water interface. The texture is strongly influenced by the flow conditions of the monolayer in compression and transfer. Since the flow processes depend on the trough design, different textures may be obtained by using different designs.

In the constant-perimeter barrier trough [50] some of the difficulties encountered with conventional compression systems can be avoided

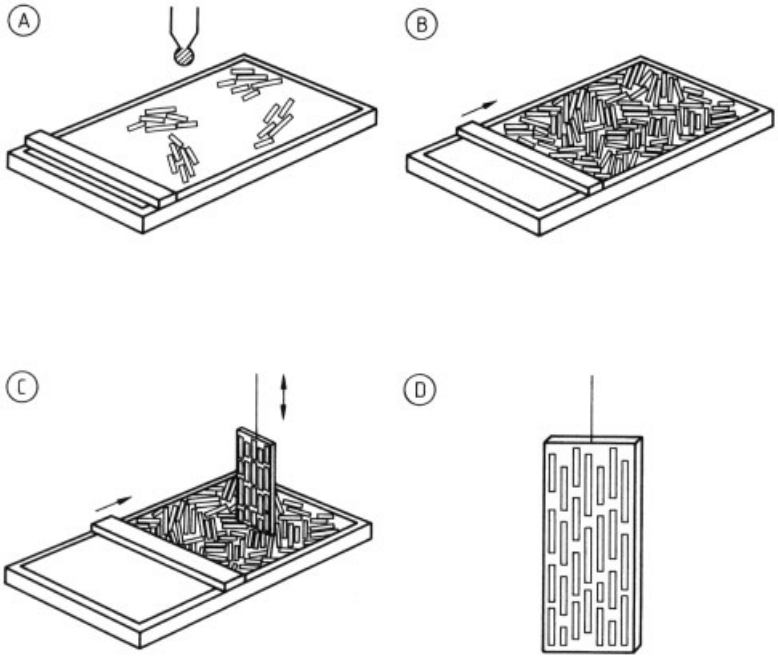

Figure 22. Deposition of hairy rod molecules
A) Spreading; B) Compression; C) Transfer; D) Monolayer structure

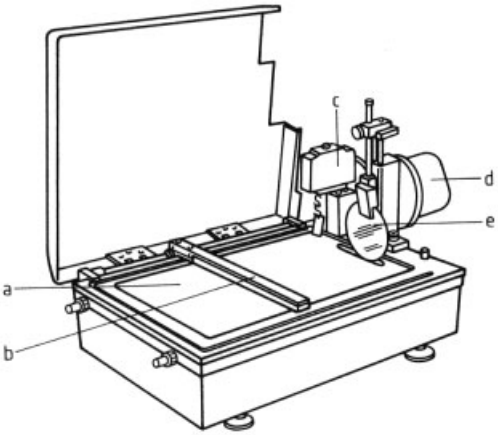

Figure 23. A linear LB (Langmuir – Blodgett) trough (Nima Technology, Coventry, United Kingdom)
a) Water basin; b) Moving barrier; c) Pressure sensor; d) Motor; e) Substrate

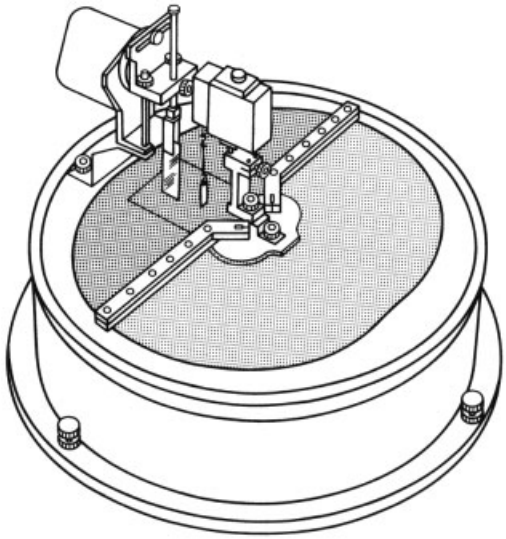

Figure 24. Circular LB trough (Nima Technology, Coventry, United Kingdom)

by enclosing the monolayer within a continuous, flexible plastic barrier. Usually, a PTFE band is stretched around four vertical glass spools and a rod. The rod can be rotated to take up the slack as the area is changed.

Double linear troughs [42] facilitate deposition of alternating layers. The substrate can be handed over automatically from one trough to another above or below the water level, which allows preparation of a variety of periodically or aperiodically stacked multilayer systems.

All troughs described so far are specifically designed for small-scale production of LB films (typical areas $1-20\,cm^2$). The number of lay-

ers on a substrate is limited by the subphase area and substrate size. For deposition of larger numbers of layers, continuous troughs have been constructed. In principle, the Langmuir time sequence (spreading, compressing, deposition) is replaced by a sequence distributed in space; all the elementary operations take place simultaneously in different compartments. Several different troughs have been developed.

In a trough with a steadily flowing subphase (Fig. 25) [51], the horizontal trough is replaced by a flow channel. Water as subphase is circulated to generate a steady laminar flow film on the ramp and thereby compresses the molecules on the surface to form a monolayer. The amphiphile can be continuously spread and deposited.

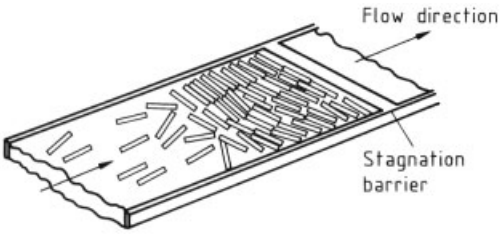

Figure 25. Monolayer compression of rod-type molecules in a continuously working LB trough with a steadily flowing subphase

A roller trough (Fig. 26) [52] is divided into four compartments connected by rollers which transport amphiphilic molecules according to the direction of rotation. In the first compartment, the amphiphile is continuously spread by a microelectrovalve and slowly transported into the second compartment where a constant surface pressure of a few mN/m is maintained. The speed of the second roller transporting the molecules into the third compartment, where deposition takes place, is adjusted according to the amount of film removed by a dipping substrate. If the surface pressure exceeds the desired value, an auxiliary roller is actuated to take the excess film to a fourth compartment, where it is removed via a capillary tube.

Another possibility of building LB multilayer structures is the horizontal lifting method (Schäfer's method) [53], introduced by Langmuir and Schäfer in 1938. Schäfer's method is useful for deposition of very rigid films, such as those located in the two-dimensional solid region of the surface – pressure – area diagram (see Fig. 20). A flat substrate is placed horizontally on the compressed monolayer. When this substrate is lifted and separated from the water surface, the monolayer is transferred onto the substrate. In theory, the molecules should maintain the same orientation; in practice, this is not always the case. The molecules near the outer rim of the substrate tend to lose their orientation.

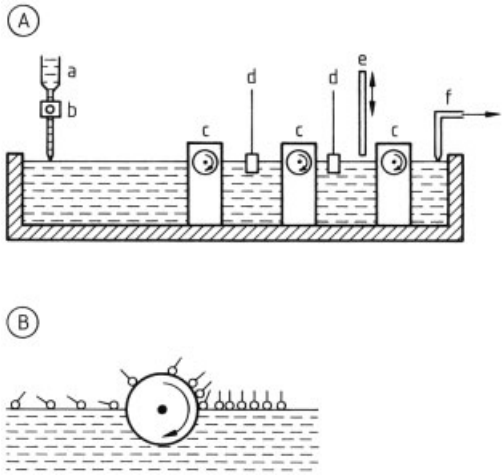

Figure 26. A) Continuously working LB trough with rollers
a) Reservoir; b) Microelectrovalve; c) Rollers; d) Pressure sensor; e) Substrate; f) Capillary tube
B) Transport of amphiphilic molecules

A large number of experimental conditions have to be taken into account for the deposition of LB layers: control of atmosphere, temperature, and humidity around the trough is recommended. The LB trough is usually installed in a laminar flow box to ensure a dust-free environment. The trough should stand absolutely shock free. Since the spread substance has an extremely large surface, special care has to be taken to avoid chemical reactions with the atmosphere or photodegradation.

The water is usually deionized through a Millipore system that contains a final filter to remove bacteria. It is particularly important to keep the subphase free of all surface-active substances. Ionic content and pH must be carefully controlled.

A variety of solids can be used as substrates: glass, quartz, or silicon wafers are commonly used, but in principle most metals or semicon-

ductors can serve as substrates, depending on the desired application of the film. Hydrophobization may be necessary. Cleaning is essential; even trace impurities may adversely affect the deposition process.

1.7. Electrochemical Deposition Processes [54–79]

1.7.1. Scope

Electrodeposition is the process of producing a thin layer on a substrate, under controlled conditions, via an electrochemical reaction. The choise of substrate and coating may encompass a large number of single metals and alloys. The substrate may be polymeric, ceramic, or a composite. The coating may be a single metal, an alloy, a polymer, a ceramic, or a metal – polymer or metal – ceramic composite. In practice, the choice is severely restricted by technical and economic constraints and there is a trend toward relatively cheap substrates (ferrous, zinc, aluminum, or copper-based alloys) as bulk constructional materials. The most common case of electrodeposition is the d.c. electroplating of a metallic coating onto a metal substrate, the workpiece being made the cathode of the electrolytic cell. However, the field of electrodeposition is much wider, including the production of anodic films (e.g., metal oxides or polymers) and electrophoretic or electropolymerized deposits.

The objective of an electrodeposition process is to produce a deposit which adheres well to the substrate and which has the required mechanical, chemical, and physical properties. It is essential that the process is predictable and reproducible. On the other hand, many metals may (by modification of the bath and plating conditions) be deposited with different properties. For this reason it is not possible to define a single set of conditions for electroplating each metal. Electrolyte composition, current density, temperature, etc. depend on the deposit properties required. Electrochemical techniques are widely used in surface finishing to produce thin layers (typically 1 – 100 µm) with particular engineering, chemical, corrosion resistance, decorative, or electronics properties.

1.7.2. Electrochemical Deposition of Metals and Alloys

The deposition of metal onto a substrate can occur by three electrochemical routes: electrodeposition, electroless plating, and immersion plating. These all involve reduction of a dissolved metal ion, M^{z+}, via electron gain to form the metal M:

$$M^{z+} + ze^- \longrightarrow M$$

The three deposition processes are differentiated by the source of the reducing electrons.

The most commonly used technique is electrodeposition or electroplating. An externally applied current provides the electrons. An electroplating bath consists of an anode and a cathode in an electrolyte, usually an aqueous solution of a salt of the metal to be plated. The item to be plated is made the cathode and reduction of ions to metal occurs here. Anodes are made either of an inert conductor, e.g., platinum or graphite, or of the metal which is being plated. Oxidation occurs at the anode. Inert anodes evolve oxygen as current flows. Anodes of the plated metal are dissolved. The thickness is mainly determined by the deposition time. Electrodeposits can be up to 100 µm thick, but thinner layers are much more common.

Electroless plating does not utilize an external power supply; instead, metal ions are reduced by chemical agents in solution and deposition is controlled so that plating occurs only on the substrate. This is achieved by an activation pretreatment of the workpiece which acts to catalyze the reduction or as a nucleation site for film growth. Metal is deposited on both conducting and nonconducting surfaces. The rate of electroless plating is considerably slower than electroplating. It is only used to deposit thin layers, typically 0.25 – 1 µm.

Immersion plating involves plating of a base metal substrate by reaction of that metal with the salts of a more noble metal, via reduction and deposition of the dissolved ions, and oxidation of the surface of the substrate. This technique has limited application, owing to restrictions on the types of metal that can be plated, and the limited thickness of deposits.

Electroplating. Alloys and composites can also be deposited by electrolysis. The uses of electroplated coatings include engineering, electronic, and decorative applications. The properties and applications of commonly electrodeposited metallic coatings are summarized in Table 6. The essential components of the electroplating process are shown in Figure 27.

Table 6. Properties and applications of electrodeposited metals

Property	Metals and alloys used	Example of use
Corrosion resistance	Cr, Ni, Sn, Au, Zn, Rh, Cd	corrosion protection, food cans
Buildup of material on surface	Cr, Fe, Ni	reclamation of worn parts
Decorative appeal	Cr, Au, Ag, Pt, Ni, 70:30 Cu – Zn	jewelry, cutlery, general decoration
Wear resistance	Cr, Ni, Fe, Sn, Ru, Pd	rollers, pistons, bearings, contacts, switches
Hardness	Cr, Ru, Os	molds, presses
Reflectivity (optical or thermal)	Cr, Rh, Au	lamps, projectors, aerospace screens and visors
Electrical conductivity	Cu, Ag, Au	printed circuit boards, antennae, bus bars, cables
Oil retention	Cu, 65:35 Sn – Ni	hydraulic systems, lubrication
Solderability	Ni, Sn, Cd, 60:40 Sn – Pb	printed circuit boards, electrical contacts
Low contact resistance	Ag, Au, Rh, Pd, Sn, 80:20 Pd – Ni	electrical contacts

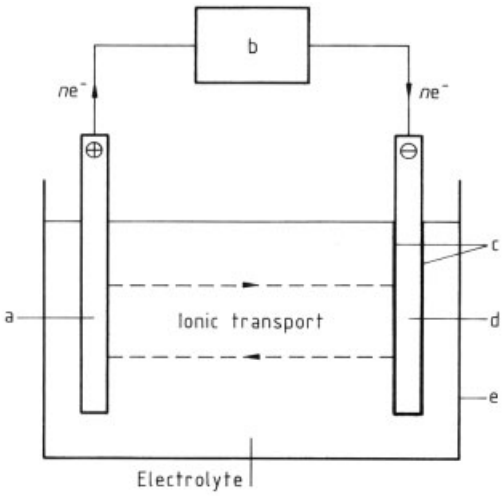

Figure 27. Principle of electroplating
a) Anode; b) Power supply, d.c.; c) Electroplated metal layer; d) Cathode (workpiece); e) Electroplating tank
The electrolyte in the plating bath consists of: water, source of metal ions, conducting salts, buffer, and additives.

The source of metal ions for the cathode reduction reaction may be a complex cation or an anion, rather than the simple metal cation. Examples include $Zn(CN)_4^{2-}$ and $Sn(OH)_4^{2-}$ in cyanide – zinc and alkaline – tin plating baths. The mass of metal deposited is directly related to the charge transferred:

$$w = \frac{\varphi M Q}{zF}$$

where w is the mass of electroplated metal, M the molar mass of metal, Q the electrical charge, φ the cathode current efficiency for metal deposition (< 1), z the cation charge, and F is the Faraday constant, 96 485 C/mol.

Most electroplating processes are carried out at constant current. If uniform deposition is assumed, the average rate of deposition per unit area is:

$$\frac{w}{tA} = \frac{\varphi M I}{zF}$$

where A is the electrode area, t the time for deposition, and I the current. The rate of thickness development can be expressed as:

$$\frac{x}{t} = \frac{\varphi M I}{\varrho z F A} = \frac{\varphi M i}{\varrho z F}$$

where x is the deposit thickness, A the electrode area, and ϱ the density of the deposited metal. The current efficiency depends on the process used. Electroplating baths are usually designed to have a high efficiency, $0.9 < \varphi < 1$, but some baths, e.g., chromic acid – chromium, are operated at much lower cathode efficiencies, typically $0.1 < \varphi < 0.2$. The most common secondary reaction at the cathode is the evolution of hydrogen which creates safety/environmental problems and the possibility of hydrogen damage to the plated deposit via embrittlement or blinding of the surface.

The throwing power of a bath is a measure of how evenly the workpiece is plated. It is assessed by comparing the plating thickness of areas at different distances from the anode. A high throwing power leads to even coverage all over the workpiece. This is important when coating items with recesses or other surface features. The main factors which affect throwing power are bath composition, temperature, and current density. A conductive electrolyte aids current flow to all parts of the cathode, minimizing the effects of increased resistance with increased distance.

If there is a high concentration of metal ions in the bath, mass transport considerations should not affect the throwing power. An even current density leads to even plating across the workpiece, so the metal plating is affected by the conductivity of the substrate surface. How evenly a metal is plated can be enhanced by auxiliary anodes (usually inert), to help distribute current to inaccessible areas.

Where possible the anode is made of the same metal as that being plated. This dissolves as plating occurs. To maintain a constant metal ion concentration, the rate of dissolution at the anode must be the same as the rate of deposition at the cathode. The current efficiency of the two electrodes must therefore be similar. It is important to minimize production of insoluble sludge as the anode dissolves (from impurities in the metal or particles left over from fabrication). Incorporation of impurities in the plate can affect the smoothness, conductivity, and contact resistance of the metal. The structure of the anode can also be important. A coarse-grain anode can dissolve unevenly, leading to fluctuations in metal ion concentration. The best anodes have a nonporous rolled structure and dissolve steadily at steady current.

Before electroplating, the surface must be cleaned to remove dirt, grease, and surface oxide. This is essential to give good adhesion between the metal plate and the substrate.

Mechanical cleaning methods include abrasion, shot blasting, and polishing. These remove scale and some dirt, but further treatment is usually necessary.

Solvent cleaning is used to remove residual oils and greases from the workpiece. This allows the surface to be better wetted by water-based electrolytes, but will not remove oxide layers.

Alkaline cleaners remove grease, soils, and scale from the metal surface. They consist of a mixture of alkali salts to saponify oil and grease to a water-soluble form, detergents to aid wetting by lowering surface tension at the metal – soil – solution interfaces, sequestering agents such as sodium gluconate to prevent precipitation of metal ions as insoluble hydroxides or carbonates, and a water softener to prevent hard water forming scale or scum on the surface. Alkaline cleaning is often carried out electrolytically. Cleaning is enhanced because the evolution of gas bubbles at the metal – solution interface dislodges soiling and transports fresh cleaner to the site.

Acid pickling is used to remove impurities by chemical attack and to stabilize impurities by ionic dissolution. An acid pickle removes oxides and scale from the surface of a metal and is often used after alkaline cleaning. A number of acids are used, either individually or as a mixture. These include nitric, sulfuric, hydrochloric, hydrofluoric, and phosphoric acids at concentration 2 – 85 %, depending on the metal and finish needed. Acid pickles are usually heated to 50 – 80 °C. As in the case of alkaline cleaners, electrolytic cleaning is often used to improve pickling.

When the plating process starts, the metal is deposited onto a surface that has a different structure from the plate. This is due to it being a different type of metal, to surface impurities, or to the surface having been machined. This can lead to poor adhesion between the base and the plated layer. Surface preparation such as the techniques outlined above improves adhesion by removing impurities and oxides, but there may still be some difference in structure between the substrate and the plate. Adhesion between these layers is often enhanced by deposition of a thin "strike" plate of metal before the main deposit is applied.

A strike plate is electrodeposited from a low-concentration solution. Such solutions have relatively low throwing power and efficiency. At low current density, a slow plating rate leads to good adhesion of the metal onto the substrate. The strike solution is used to build a good basis upon which to plate from the high-concentration main tank.

Plating baths are a mixture of metal salts, electrolytes, and trace additives which modify the plating characteristics. The precise makeup of commercial baths and chemicals is usually secret, but general functions of each component are well documented.

The metal ion to be plated is kept at high concentration, usually 1 – 3 mol/L. It may be either the simple hydrated ion or a complex. If the deposition becomes mass transport controlled,

owing to a lack of metal ion at the cathode surface, plating quality is impaired because plating is now favored at sites on the workpiece that are easily accessible to the solution, and the lack of metal ions available for reduction promotes electrochemical side reactions. For this reason the concentration of metal ion is kept high.

Items of relatively simple shape are usually coated with solutions of the simple hydrated ion, derived from a salt. This is because the higher throwing power of complex ions is not needed. There are exceptions to this, e.g., when the metal being deposited is difficult to dissolve. To electroplate complicated shapes higher throwing power is needed, utilizing a complexed metal ion electrolyte or auxiliary anodes.

To maximize the conductivity of the bath, high concentrations of electrolyte are added; acidic, alkaline, and neutral electrolytes are used, e.g., H_2SO_4, KCN, NaOH, NaF, KF, NaCl, and NH_4Cl.

Complexing agents are added to make the deposition potential more negative, if it is necessary to prevent spontaneous reaction between the plating ion and the cathode material. For example, plating copper onto iron, the copper ions in solution attack the metal surface, dissolving iron:

$$Cu^{2+} + Fe \longrightarrow Cu + Fe^{2+}$$

This chemical reaction will lead to a very poor deposit, both porous and badly adherent.

Complexing agents can modify the redox potential of the Cu^{2+}/Cu couple so that it is more negative than the Fe^{2+}/Fe couple and the above reaction is no longer favorable thermodynamically. Common complexants are cyanide (CN^-), pyrophosphate ($P_2O_7^{4-}$), and hydroxide (OH^-). Complexing agents can also prevent passivation of dissolving anodes which maintains current efficiency.

The structure, morphology, and properties of the electroplate can be modified by low concentrations of organic molecules. The exact identity and concentration of additives for specific baths is kept secret by commercial suppliers. There are four general types of additive.

Brighteners produce a visually brighter or more shiny surface finish, by reducing the microscopic roughness of the deposit. The electrical properties are unaffected and brighteners are used when decorative appearance is important. Relatively high concentrations (g/dm^3) are used.

Levelers act by adsorption at points on the plating surface where there would be rapid deposition of metal. The adsorbed molecules reduce the rate of electron transfer, slowing the plating rate at these points and making plating across the whole workpiece more even. Many additives act as both brighteners and levelers.

Structure modifiers act by changing the electrodeposit structure to optimize particular properties of the plate. They are used when the plating is required to have certain mechanical properties, e.g., wear resistance for electrical contacts.

Wetting agents facilitate the release of gas from the cathode surface. Inclusion of hydrogen can cause the deposit to become brittle and prone to cracking.

Some tyical electroplating conditions are summarized in Table 7.

Metals are usually electroplated to a thickness of 1 – 100 µm. Deposits below 1 µm are generally too porous, and a plating thickness greater than 100 µm is usually ruled out by considerations of cost and time, although chromium plating may give thickness up to 500 µm, in special cases. Plating thickness for a particular metal is determined by the current, time, area, and electrode efficiency. For any electroplate, the thickness can be controlled by adjusting the current and time of plating. Most plating baths are current controlled and thus operated at known current densities.

The quality of the plating is affected by the rate of growth of the metal. If the plate is build up slowly, the microstructure is well defined, with layered, ridged, or block formation. If the plating rate is increased, a less dense, polycrystalline solid results. At high plating rates, nodules and dendrites of metal form. At the maximum deposition rate the metal goes down as a powder which does not bond to the surface. The change in plating characteristics with rate is mainly due to mass transport considerations. When the metal is deposited slowly, there is time for metal ions to diffuse to the cathode surface and there is always a high concentration of metal

Table 7. Typical process conditions for electroplating single metals [58]

Metal	Electrolyte composition, g/dm^3	Temperature, °C	Current, mA/cm^2	Current efficiency, %	Additives	Anode	Characteristics
Cu	$CuSO_4$ (200–250) H_2SO_4 (25–50)	20–40	20–50	95–99	dextrin, gelatin, S-containing brighteners, sulfonic acids	P-containing rolled Cu	not suitable for iron and iron alloy substrates; high current, but low throwing power; additives essential
	CuCN (40–50) KCN (20–30) K_2CO_3 (10)	40–70	10–40	60–90	Na_2SO_3	O_2-free, high-conductivity Cu	good throwing power; deposits adhere well and are bright; used as undercoat; cyanide decomposes at anode
Ni	$NiSO_4$ (250)	40–70	20–50	95	coumarin, saccharin, benzenesulfonamide acetylene derivatives	Ni pellets or pieces	H_2 evolved with deposition; additives depend on purpose of deposit; medium throwing power
	$NiCl_2$ (45) H_3BO_3 (30) pH 4–5						
	Ni sulfamate (600)	50–60	50–400	98	not essential; naphthalene-1,3,6-trisulfonic acid	Ni pellets or pieces	good throwing power; some H_2 with additives, mirror finish at 400 mA/cm^2
	$NiCl_2$ (5) H_3BO_3 (40) pH 4						
Ag	$KAg(CN)_2$ (40–60)	20–30	3–10	99	S-containing brighteners	Ag	good throwing power; functional or decorative
	KCN (80–100) K_2CO_3 (10)						
Sn	$SnSO_4$ (40–60)	20–30	10–30	90–95	phenol, cresol	Sn	poor throwing power; additives essential
	Na_2SnO_3 (50–100) NaOH (8–18)	60–80	10–20	70		Sn (control current)	good throwing power, but uses twice charge (Sn^{4+})
	$SnCl_2$ (75)	65	50–200	90	SCN-disulfononaphthalinic acid	Sn	reasonable throwing power; more stable than $SnSO_4$; rarely used
	NaF (37) KF (50) NaCl (45) pH 2.5						
Zn	NH_4ZnCl_3 (200–300) NH_4Cl (60–120) pH 5	25	10–40	98	dextrin, organic additives	Zn	poor throwing power; additives essential

ions available for reduction. Since deposition is slow, the metal atoms can form organized, well-ordered structures. When the rate is increased, the concentration of metal ions at the cathode is lower. Despite the short-range order in the structure the faster growth rate does not allow long-range lattices to form (polycrystalline). At high deposition rates, the metal ions arrive quickly at the cathode, and do not form ordered structures. The metal ion concentration at the surface becomes much lower than that of the bulk electrolyte, promoting the growth of nodules and dendrites since plating is favored further from the plane of the workpiece surface where there is a higher concentration of metal ions. The formation of dendrites is further favored since they cause an increase in surface area which reduces the overall current density.

The limiting factor in deposition rate is diffusion-controlled mass transport of the metal ion. At a particular temperature and electrolyte concentration, this has a maximum value (maximum rate of deposition). Deposits laid down this fast have very little structure and form powders which are easily removed by physical agitation. For a typical bath, the limiting current density is ca. 70 mA/cm^2. To ensure good adhesion, the current is kept at a much lower level. The main way of increasing plating thickness is to lengthen the plating time.

Electroless Deposition. In electroless deposition, the substrate acts as a catalyst for a spontaneous redox reaction. No external current is required and the process takes place by immersion of the workpiece in the plating bath. It is essential, however, that the substrate is pretreated so that deposition of metal takes place over the entire surface. Under constant process conditions, deposition takes place at a constant rate and the deposits tend to be microcrystalline, with the alloying element (e.g., P or B in nickel deposits) in solid solution. Despite the relatively high cost of electroless plating (compared with electrolytic deposition), the technique offers the advantages of a very uniform deposit thickness, even on workpieces with complex shapes. In Ni – P deposits, heat treatment (say 400 °C for 1 h) after deposition can increase the hardness (typically from 500 to ca. 1000 VHN) owing to crystallization of an Ni_3P dispersion throughout the coating. In addition to a metal salt and a suitable reducing agent, electroless plating solutions are formulated with additives such as complexants, stabilizers, pH buffers, accelerators, and brighteners to provide stable electrolyte solutions and the required deposit characteristics.

The most common electroless deposits are copper (to produce conductive tracking on printed circuit boards) and nickel (as a corrosion-resistant and engineering coating). Copper deposition is commonly carried out from an alkaline bath containing formaldehyde as the reducing agent while nickel baths commonly contain hypophosphite (Ni – P deposits) or less commonly borohydride (Ni – B). Typically, Ni – P deposits contain 3 – 15 wt % Ni.

Oxidation of the reducing agent must have an electrode potential which is more negative than that of the metal deposition process (Table 8). In electroless copper deposition from a bath containing formaldehyde, the simplified reactions are deposition of copper by reduction of its ions at cathodic sites:

$$Cu^{2+} + 2e^- \longrightarrow Cu$$

Table 8. Standard electrode potentials of reactions in electroless and immersion deposition

Reversible electrode process	Electrode potential∗, V
$Ag^+ + e^- = Ag$	+0.799
$Cu^{2+} + 2e^- = Cu$	+0.337
$HCOOH + 2H^+ + 2e^- = HCHO + H_2O$	+0.056∗∗
$Ni^{2+} + 2e^- = Ni$	−0.250
$Fe^{2+} + 2e^- = Fe$	−0.440
$H_2PO_3^- + 2H^+ + 2e^- = H_2PO_2^- + H_2O$	−0.500
$Zn^{2+} + 2e^- = Zn$	−0.763
$HCOOH + 2H^+ + 2e^- = HCHO + H_2O$	−1.07∗∗∗

∗ Relative to standard hydrogen electrode at 25 °C. ∗∗ At pH 0. ∗∗∗ At pH 14.

accompanied by reduction of formaldehyde to formic acid at anodic sites:

$$HCHO + H_2O \longrightarrow 2H^+ + 2e^- + HCOOH$$

The overall redox process on the metal surface is:

$$Cu^{2+} + HCHO + H_2O \longrightarrow Cu + 2H^+ + HCOOH$$

In electroless deposition from baths containing hypophosphite, the simplified reactions are as follows. The anodic reaction is oxidation of hypophosphite to orthophosphite:

$$H_2PO_2^- + H_2O \longrightarrow H_2PO_3^- + 2\,H^+ + 2e^-$$

and the cathodic reaction is deposition of nickel:

$$Ni^{2+} + 2\,e^- \longrightarrow Ni$$

The overall redox process on the metal surface is:

$$H_2PO_2^- + H_2O + Ni^{2+} \longrightarrow H_2PO_3^- + 2\,H^+ + Ni$$

It is important to note that:

1) The substrate metal and the deposited metal must support the electrode process in a catalytic manner
2) In the case of noncatalytic surfaces, the electroless process may be initiated by preliminary deposition of a catalytic metal (as in the case of plastics) or brief electrolytic deposition of the coating metal by passing a current for a short period
3) The process must be operated so as to avoid spontaneous decomposition of the electrolyte in the bath or on the tank surfaces
4) A pH decrease accompanies the overall process
5) The reducing agent is depleted and its oxidation product accumulates
6) The source of metal, i.e., Ni^{2+}, declines in concentration as the bath is used
7) In practice, the deposit is an alloy with a significant phosphorous level ($< 15\,\%$), showing that the above reactions are greatly oversimplified.

Points (4), (5) and (6) are addressed by planned makeup additions to the electrolyte.

By suitable formulation of the bath, electroless deposition can produce single metals, or metals containing a nonmetal component (e.g., P or B). In addition, special baths are available for the deposition of composite "inclusion" coatings, e.g., Ni – PTFE (as a nonstick coating) or Ni – SiC (as a wear-resistant coating), the inert particles being introduced to the electrolyte by agitation and/or surfactant dispersion techniques.

Immersion Deposition. If steel is dipped in a solution of copper sulfate, a deposit of copper is spontaneously formed on the surface of the steel. This reaction can be explained with reference to the electrochemical series (Table 8). The more positive (noble) redox system is reduced (deposition) while the more electronegative (base) is oxidized (dissolution).

The surface of the steel consists of a mixed electrode comprised of a mosaic of anodic and cathodic sites; the following reactions take place as the standard electrode potential for copper is more positive than that of iron. Anodic site:

$$Fe \longrightarrow Fe^{2+} + 2\,e^- \;+0.44\,V$$

Cathodic site:

$$Cu^{2+} + 2\,e^- \longrightarrow Cu \;+0.34\,V$$

Overall mixed-electrode reaction:

$$Fe + Cu^{2+} \longrightarrow Fe^{2+} + Cu \;+0.78\,V$$

The deposition of copper continues until almost the entire surface of the steel is coated. At this point there are insufficient anode sites left to dissolve (oxidize) and supply electrons for the cathodic reaction, and deposition ceases. Copper is usually deposited to a thickness of ca. 1 µm. Deposits tend to be very porous and adhere poorly to the substrate. The copper solution becomes contaminated in a relatively short time with ions of the base metal. These considerations limit the applications of immersion plating. It is normally used to produce deposits of silver, gold, or palladium on base metals. Examples include the deposition of gold onto nickel-clad printed circuit boards and the metallization of printed circuit plastics with palladium by deposition onto a tin layer.

1.7.3. Electrophoretic Coatings

Electrophoresis is the migration of charged organic molecules to electrodes of the opposite charge. At the electrode, the particles accumulate and deposition follows rapidly. In the organic coating of automobile bodies, prior to the finishing coats of paint, electrophoretic coating is also known as electropriming, electropainting, or electrocoating. It is used to deposit films 1 – 20 µm thick, the majority of applications being in the range 4 – 10 µm. These films have high corrosion resistance and an even coverage over the whole workpiece.

There are two types of electropriming: anodic and cathodic. Both use a water-based primer solution as the electrolyte. Water is favored as the solvent because it is cheap, harmless, and effluent treatment is straightforward.

In anodic coating, the article to be coated is given a positive charge. A water-insoluble, organic acid resin is dispersed in water by the action of a base to form colloidal-size anions:

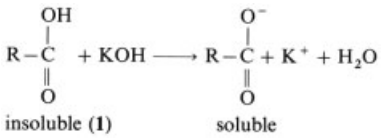

The anions are attracted to the anodic workpiece where water is electrolyzed to H^+ and O_2 by removal of electrons:

$$2\,H_2O \longrightarrow 4\,H^+ + O_2 + 4\,e^-$$

The protons produced neutralize the resin, which is deposited as an insoluble, resistive film of (**1**).

The chemistry of the more commonly used cathodic electrocoating is similar in many ways. Acid is used to form colloidal cations from insoluble aminated resins:

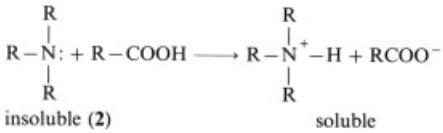

As these ions collect at the cathode, electrolysis of water produces hydroxyl ions and hydrogen:

$$4\,H_2O + 4\,e^- \longrightarrow 2\,H_2 + 4\,OH^-$$

The resin reacts with the hydroxyl ions to deposit on the surface as an insoluble, resistive layer of (**2**).

Additives can be mixed with the resins to modify their properties. Examples are metals, metal oxides, carbon, sulfur, and cellulose. These are added as colloidal-sized particles which gain a positive charge by adsorption of hydroxonium ions in the case of cathodic electropainting. Large particles can migrate under the influence of a few charged ions. Electropriming is usually performed onto a metal substrate. The quality of the coverage is affected by the preparation of the metal surface. Conventional cleaning and degreasing techniques are often used, but the adhesion of some primers is enhanced if the base metal has been phosphated prior to electrocoating.

The high throwing power of electropainting is due to the resistive character of the paints. Once a layer has formed, this electrically insulates the surface, preventing further deposition. Coverage on less heavily coated areas is favored.

Following electrocoating, the work is rinsed and cured. Curing removes solvents and irreversibly alters some resins, improving coating finish. The technique is efficient in terms of energy and materials. This has led to a number of applications in large-scale processes, notably the priming of large metal parts such as car bodies prior to further spray coating. Electropainting is also used to coat coil stock, pipes, sheet steel constructions, and cans.

Commercial electropriming tanks use refinements such as ion-selective membranes around the counter-electrode to prevent diffusion of unwanted ions, and thus control paint pH (Fig. 28). The anodes in such systems are usually made of stainless steel, typically with a surface area about one-quarter that of a car body. The body is immersed in the paint, large tanks are required. Direct current at 50 – 500 V is commonly used to electrodeposit the primer layer. The requirement of large-scale processing equipment leads to electrocoating being used where a large number of parts need to be coated quickly and cheaply.

Electrophoretic techniques may also be used to codepositpolymer or ceramic particles within an electrodeposited metallic film. The inert particles (e.g., PTFE, alumina, silicon carbide, or diamond) can be dispersed in the electrolyte by high-stability (usually fluorocarbon-based) surfactants. The choice of surfactants and their concentration are critical in determining deposit properties and composition. Examples of the resultant "inclusion coatings" are Ni – PTFE, Co – SiC, and Ni – SiC.

1.7.4. Trends in Electrodeposition

Electroplating is vital in almost every branch of industry where the technique can provide a versatile and economic method of applying a thin

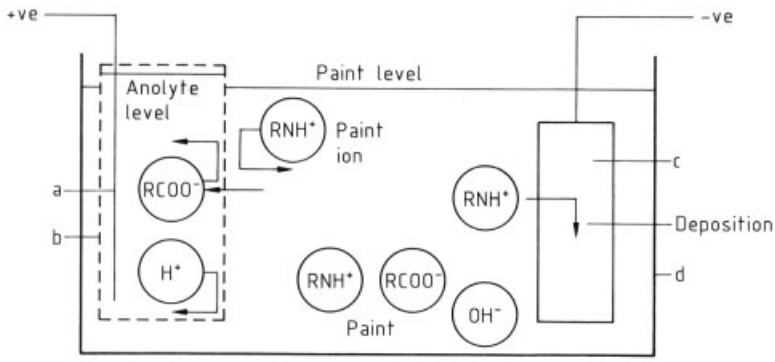

Figure 28. Cathodic priming of car bodies
a) Anode plate; b) Ion-selective membrane; c) Car body; d) Paint tank
An ion-selective membrane prevents diffusion of H^+ and controls pH.

(1 – 75 µm) metallic coating. However, itis essential to consider the effect of the shape, texture, and material composition of the workpiece at the design stage. Of the many modern developments in electrodeposition technology, several trends may be highlighted:

1) Electroplating of alloys of controlled composition, e.g., Ni – Fe or Pd – Ni as cheaper alternatives to single-metal coatings
2) Codeposition of metals with either polymers or ceramics to provide unique wear-resistant or "self-lubricating" coatings, e.g. Ni – PTFE, Ni – diamond, Co – WC, Zn – alumina
3) Electroless (zero-current) plating of an increasingly wide range of single metals, alloys, and composites, e.g., electroless Ni – P – PTFE
4) Low-temperature nonaqueous electrolytes for deposition of refractory metals, e.g., Ti, Ta, and Nb
5) Selective use of special, programmed current waveforms, e.g., periodic current reversal (rather than the normal steady direct current) in order to produce thick, dense, pore-free coatings, particularly for demanding electronic applications and severe wear-resistant applications
6) Selective, high-speed, automated electroplating processes for reel-to-reel plating of electronics materials
7) Local enhancement of deposition rates by use of laser irradiation, jetted electrolyte flow, as well as by acoustic, mechanical, or ultrasonic vibration
8) Development of polymer coatings based on electrochemical deposition from a monomer in the electrolyte, e.g., polypyrroles and polythiophenes whose conductivity can vary over many orders of magnitude via variation in monomer, electrolyte, and deposition conditions
9) Continued developments in electrophoretic deposition of polymers and ceramics, particularly in the finishing of car bodies

Journals and Periodicals. The electrochemistry of electrochemical deposition reactions is considered in the following journals:

Journal of Applied Electrochemistry
Journal of Electroanalytical Electrochemistry
Journal of the Electrochemical Society
Electrochimica Acta
Soviet Electrochemistry
Bulletin of Electrochemistry

Journals devoted to aspects of the surface finishing of metals include:

Plating and Surface Finishing
Metal Finishing
Transactions of the Institute of Metal Finishing
Galvanotechnik
Oberfläche-Surface
Metal Finishing (India)

Other journals relevant to the electrochemical deposition of thin films include:

Surface Technology
Thin Solid Films
Synthetic Metals

2. Thin Films as Materials

For a discussion of corrosion-resistant coatings → Steel, Chap. 5.3., → Steel, Chap. 5.4., → Corrosion, Chap. 5.1.2.3..

2.1. Coatings for Electronic Applications

2.1.1. Introduction

The world market of semiconductor-related products increased from $\$45\times10^9$ in 1988 to $\$74\times10^9$ in 1993. The corresponding market for deposition equipment increased from $\$0.9\times10^9$ in 1989 to $\$1.5\times10^9$ in 1992. The worldwide hybrid market was estimated at $\$5.5\times10^9$ in 1990.

As microelectronics pushes the deposition processes toward shorter cycle time, higher reproducibility, uniformity/conformality, and lower cost, deposition processes have become highly sophisticated. For integrated circuit (IC) assembly, thin film multichip modules are of increasing importance [80], but are not discussed here.

2.1.2. Hybrid Technology

One of the oldest applications of thin films in electronics are carbon resistors deposited on cyclindrical ceramic substrates by pyrolysis of a mixture of hydrocarbons at 400 – 600 °C. Because of the high negative temperature coefficient of resistivity (TCR) and the high tolerance of the resistance (± 10 – 20 %) metal resistors (base materials: NiCr, TaN, or CrSi) are now dominant [81]. In R – C networks the negative TCR of the resistor is compensated by the positive TCR of the capacitor [82]. Thin film hybrids are at present only a small portion of the hybrid market, but their share is likely to grow because large-scale integration for multipole ICs requires thin film memories, microprocessors, logic devices, semicustom, linear and discrete components, and also because higher pin-in/pin-out counts for dense modules are now achievable.

2.1.3. Microsystems

Combining micromachining technology of Si [83] with data collection (sensors) and microelectronic data processing has created a new class of products called microsystems (intelligent sensors).

Infrared sensors are made of HgCdTe or GaAs. For HgCdTe deposition, a lot of effort has been put into vapor-phase deposition, because of the performance-limiting defects which occur with epilayers. The range of detectable wavelengths can reach 8 – 12 µm.

Optical sensors/detectors consist of Si deposited by evaporation, sputtering or CVD. Applications include scanning, colorimetry, position sensing, fiber optic sensing, multiplexing, and bar-code reading.

Gas sensors are made of doped Si or polymers. Applications include smoke detectors, detection of trace quantities of toxic gases, and combustion control for car engines.

Pressure sensors employ NiCr, Si, piezoceramics (e.g., ZnO), GaP, or heterostructures. They are used in strain gauges, pressure sensing, acceleration sensing, and microphones.

2.1.4. Data Storage

Thermal print heads [84] are fabricated from highly resistive, highly stable compounds with low TCR, e.g., CrSiO, TaN, or HfN as heating elements; SiO_2 or Si_3N_4 for wear protection. They are used in facsimile machines, barcode printers, industrial printers and dye-transfer printers. For bubble-jet print heads and magnetic record heads see [85, 86].

Magneto-optic (MO) storage. In MO storage [87] the Kerr effect is used for data storage and reading. The materials used are generally amorphous alloys of rare earths and transition metals (Fe, Co) deposited by sputtering. These amorphous materials do not suffer from "grain noise".

2.1.5. Microelectronics

The two most important physical processes in microelectronics materials preparation do not

involve deposition. Diffusion/ion implantation is used for the local, electrical activation of the bulk Si, and thermal oxidation produces the high-performance gate oxides [88]. Nevertheless, for a 16 megabit dynamic random access memory, about 20 deposition processes are needed (two-level metallization with W contact hole-filling). Before depositing the first metallization layer [sputtered Al – Si (1 wt %) or Al – Si (1 wt %) – Cu (0.5 – 2 wt %)] deposition temperatures above 450 °C may be applied.

High-Temperature processes. Applications include insulation (capacitor dielectrics):

$$SiH_2Cl_2 + 2N_2O \xrightarrow{910\,^\circ C} SiO_2 + 2HCl + 2N_2$$

oxidation masking (capacitor dielectrics):

$$3SiH_2Cl_2 + 7NH_3 \xrightarrow[40\ Pa]{750\,^\circ C} Si_3N_4 + 3HCl + 6H_2 + 3NH_4Cl$$

and formation of conformal insulation layers, spacers, etch-stops, and protection against dopant out-diffusion:

$$\underset{\text{(TEOS)}}{Si(C_2H_5O)_4 \text{ at } 30-70\,^\circ C} \xrightarrow{700\,^\circ C} SiO_2 + 2H_2O + 4C_2H_4$$

Medium-Temperature Processes. Applications include the use of BPSG (Boron – phosphorus – silicate glass) for insulation, smoothing of edges, and planarization of topography by a 850 – 900 °C anneal/flow step:

$$Si(C_2H_5O)_4 + B(OCH_3)_3 + PH_3 + (O_2) \xrightarrow{630-680\,^\circ C} \text{BPSG + decomposition products}$$

This process is hard to control under production conditions and new processes have been developed [89].

Another application is the production of conduction lines:

$$SiH_4 \xrightarrow[40\ Pa]{630\,^\circ C} Si + 2H_2$$

This CVD process is important in semiconductor technology, and has influenced the development of equipment (tube furnaces, low-pressure CVD). The process gives high-purity deposited polycrystalline Si and high throughput (ca. 100 150-mm wafers per hour) and gives outstanding uniformity and conformality of the film thickness.

The resistivity of doped poly-Si is too high for some applications, and a new class of materials has been introduced (refractory disilicides: $TiSi_2$, $MoSi_2$, $TaSi_2$, WSi_2) which are in equilibrium with Si at higher temperatures. The films are sputtered from compound or mosaic targets.

Metallization Processes [90, 91]. For the electrical connections between activated areas, an alloy of Al – Si (0.5 – 2 %) is sputter deposited. The low-cost Al reduces the native silicon oxide at ca. 450 °C, leading to low ohmic contact resistances. The Si content in the alloy avoids pronounced out-diffusion of the bulk Si into the Al conduction line. After cooling, the Si alloy component forms lightly doped precipitates, preferably on the bulk Si, with average diameter 0.3 – 1.5 µm.

For sub-micrometer technology, TiN or Ti – W – N layers are applied to reduce Si precipitates in the contact holes. Since these barrier layers have excessive contact resistance to Si, an additional contact layer (e.g., 50 nm sputtered Ti) is necessary. A further problem is that the step coverage of the sputtered films is no longer sufficient. Therefore CVD processes have been developed for TiN deposition [92]:

$$Ti[N(CH_3)_2]_4 + 2H^* \xrightarrow[40\ Pa]{430\,^\circ C} TiN + 3C_2H_6 + 2CH_4 + {}^3\!/_2 N_2$$

(where H∗ is produced by microwave irradiation) and for contact hole filling with W [93]:

$$WF_6 + 3H_2 \xrightarrow[8\ kPa]{430\,^\circ C} W + 6HF$$

Passivation. *Low-temperature plasma CVD processes* include:

$$3SiH_4 + 4NH_3 \xrightarrow[30\ Pa]{300\,^\circ C} Si_3N_4 + 12H_2$$

for compressive stress $10^8 - 10^9$ Pa, and:

$$3SiH_4 + 2N_2 \xrightarrow[0.5\ kPa]{390\,^\circ C} Si_3N_4 + 6H_2$$

for compressive stress 5×10^7 Pa. This is used for protection against alkali and H_2O diffusion.

In *multilevel metallization*, the Al – Si – Cu conduction line levels have to be separated by an intermetal insulating layer deposited by PECVD below 430 °C.

$$Si(C_2H_5O)_4 \xrightarrow[1.3\ kPa]{400\ °C} SiO_2 + \text{decomposition products}$$

For further developments see [94].

2.1.6. Semiconducting Films

The basic process for *thin film transistors* (TFTs) is the PECVD deposition of amorphous Si (a-Si) at 300 – 400 °C, starting from SiH_4. Many substrates are in use, e.g., Si, GaAs, glass, quartz, and alumina. The breakthrough was reached when doping of the a-Si by a simple codecomposition of *p*- and *n*-type dopant gas species (e.g., PH_3, B_2H_6) became possible. Applications include television, logic circuits for image sensors (CCDs), printer/copiers, and active matrix liquid crystal displays.

Polycrystalline Si TFTs seem likely to become an important competitor with a-Si for displays, because the drive electronics can be fabricated directly on the glass substrate during the formation of the TFT array.

In the field of *photovoltaics*, a-Si solar cells and GaAs, InP, CdTe, and $CuInSe_2$ cells are fabricated. GaAs and InP cells are attractive for use in space because of their high efficiency and radiation resistance. The efficiency of solar cells is limited by recombination of electrons at the transparent conductive oxide – *p*-layer interface; thin film materials such as ZnO, SnO_x and indium oxide – tin oxide (ITO) are under investigation [95].

Optoelectronic devices [96] are made from $LiNbO_3$ or polymers [97]. Applications include phase modulators, amplitude modulators, splitters, polarizers, and switches.

GaAs and GaAlAs [98] can also be used for detectors, waveguide components, and high-speed electronic circuitry on the same substrate. Thin film deposition processes promise to play a major part in integrated optics, e.g., semiconductor planar waveguides could be used to integrate several optical components with associated waveguides on a single substrate.

For the use of semiconductor superlattices grown from III – V, II – VI, and IV – VI compounds for quantum wells see [99].

2.1.7. Ferroelectrics [100]

Ferroelectrics can be made from: perovskite compounds, e.g., $BaTiO_3$, $KNbO_3$, $K(Ta, Nb)O_3$, $PbTiO_3$, $Pb(Zr, Ti)O_3$; or non-perovskite compounds, e.g., $LiNbO_3$, $Bi_4Ti_3O_{12}$, $(Sr, Ba)Nb_2O_6$.

The breakthrough for this multifunctional class of materials was an improvement of the deposition process [101]. For optical applications of ferroelectric films, the system technologies are not completely mature. Several alternative materials and device technologies exist. For electronic applications, ferroelectrics have commercial potential, e.g., in nonvolatile memories and dynamic random access memories, microactuators, thin film capacitors, and surface acoustic wave devices.

2.2. Hard and Decorative Nitride and Carbide Coatings

2.2.1. Introduction

Transition metal nitrides and carbides exhibit characteristics of covalent, ionic, and metallic bonding. They have high thermal and electrical conductivities while the covalent nature gives rise to high mechanical microhardnesses, typically 20 – 30 GPa [102, 103]. They exhibit a metallic luster, and for some compounds (e.g., TiN, VN, ZrN, TaC) interband transitions locate the reflectance edge in the middle of the visible region, giving a golden color. Most of the compounds are also completely soluble in each other, making it possible to tailor the properties by depositing carbonitrides, alloy nitrides, or alloy carbides [102, 104]. The combination of high thermal and electrical conductivities, hardness, and attractive color have resulted in extensive use as both wear-protective and decorative coatings. Nitride thin films are also used as diffusion barriers in integrated circuits.

2.2.2. Growth Techniques

Nitride and carbide coatings are grown by chemical vapor deposition (CVD) or ion-assisted physical vapor deposition (PVD). In the for-

mer, substrate temperatures > 900 °C are typical; in the latter, growth temperatures of 300–550 °C are common. For acceptable quality of coatings grown by PVD at 300–550 °C, ion bombardment is necessary during growth, usually by plasma-assisted techniques, e.g., magnetron sputtering, arc evaporation, or ion plating [104]. Owing to the nonequilibrium nature of PVD techniques, film microstructure and morphology, stress levels, hardness, adhesion, etc. depend strongly on the deposition conditions, and a large spread in properties is common for coatings prepared in different equipment [103]. The most systematic correlations between deposition parameters, microstructures, and properties are found for TiN coatings.

2.2.3. Coating Microstructures

The detailed microstructure of the coatings depends strongly on growth parameters, but under most conditions columnar microstructures are formed. Films grown by PVD at < 500 °C are low-density coatings with a pronounced columnar microstructure [105]. The columns extend from the substrate surface to the top surface (Fig. 29). The surface is typically rough with faceted columns. For dense films, temperatures significantly higher than 500 °C are required, but even in this case the microstructure is columnar. By using ion irradiation during growth, dense films can be grown, even below 500 °C. Ion bombardment enhances adatom mobility, forward sputtering, and recoil implantation, leading to densification [106]. The energy required to obtain dense films depends on growth conditions, but for typical growth rates of the order of μm/h and ion current densities in the mA/cm^2 range, ion energies of 50–100 eV are required [104–107].

As the ion energy is increased above that required for creating dense films, ion-induced defects (point defect clusters, small dislocation loops, sputtering gas atoms) become trapped within the film. This can lead to continuous renucleation during growth and interruption of the columnar microstructure (Fig. 30). A relatively smooth surface is obtained. As the microstructure changes from a porous columnar structure to a dense noncolumnar structure, the stress in the coatings changes from tensile to compressive. The compressive stress usually increases to a saturation value as the ion energy is increased [107]. Compressive stress levels as high as 10 GPa can be obtained. Thermal stresses due to different thermal expansion coefficients of substrate and coating are usually smaller than growth-induced stresses, unless temperatures significantly above 500 °C are used. The different microstructures are reflected in the physical properties of the coatings. For TiN hardness values of 10–40 GPa are reported [103] (the single-crystal bulk value for TiN is 20 GPa [102]). The lowest values are obtained for low-density films, the high-

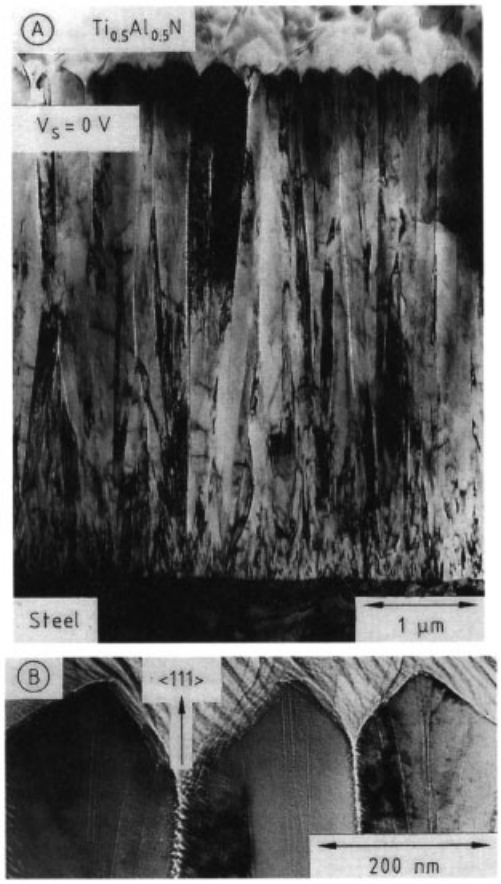

Figure 29. Cross-sectional transmission electron micrograph of a $Ti_{0.5}Al_{0.5}N$ film grown by reactive magnetron sputtering on a stainless steel substrate
Grown with the substrate grounded, i.e., no intentional ion bombardment.
A) Overview micrograph; B) Higher magnification micrograph of the film surface showing faceted top surface and intercolumnar porosity

est values only for coatings with high compressive stress levels. For optical or decorative purposes, low-density films with rough surfaces are not acceptable since their properties change with time and they are very sensitive to fingerprints. For most applications, dense, moderately compressively stressed coatings are required.

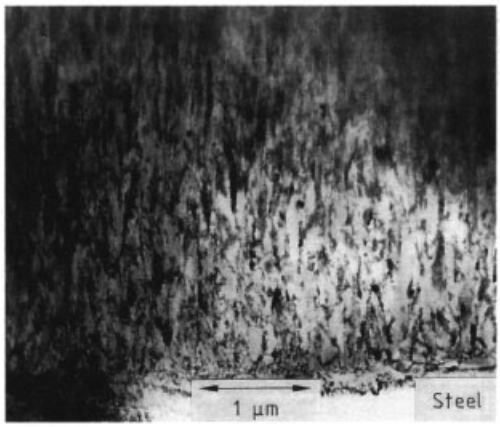

Figure 30. Cross-sectional transmission electron micrograph of a $Ti_{0.5}Al_{0.5}N$ film grown by reactive magnetron sputtering on a stainless steel substrate
Grown with substrate at a negative potential of 150 V, i.e., intentional ion bombardment.

2.2.4. Adhesion

Since coated tools are subjected to very large forces, the adhesion strength between the coating and the substrate has to be very high. Although the available empirical results do not dictate a unique mechanism for adhesion-enhancing processes, some general requirements for good adhesion can be noted. The most important is to eliminate contamination of the substrate surface prior to growth. During high-temperature growth (e.g., CVD) organic contaminants are readily desorbed from the surface. High temperature also produces interdiffused graded zones between film and substrate, with strong metallurgical bonding, and thus a high adhesion strength (provided no brittle phase is formed). For low-temperature growth, in situ sputter etching in a noble gas (most commonly Ar) plasma is necessary [108]. In addition to removing surface contaminants, sputter etching avoids low nucleation density, since lateral growth may result in interfacial voids. During low-temperature PVD growth, graded interfacial zones can be formed by depositing interfacial layers. For TiN coatings, it is common to use Ti interlayers to promote adhesion [108, 109]. Collisional mixing by ion bombardment can also be used to create graded interfaces, and thus higher adhesion strength.

2.2.5. Hard Coatings

Nitrides. The most used and studied material for hard and decorative coatings is TiN [103, 110]. Stoichiometric bulk TiN has a hardness of 20 GPa and a golden color. However, owing to the wide stability range ($0.6 < N/Ti < 1.2$) its properties vary with composition [102]. The color can range from that of metallic Ti to almost brown. The different microstructures obtainable give coatings with a hardness range of 10 – 40 GPa. The majority of commercially used TiN films have compositions close to stoichiometric, hardness 20 – 27 GPa, and compressive stresses of a few GPa. For N/Ti ratios below 0.6, films consisting of both the NaCl-structure δ-TiN phase and the α-Ti phase are reported. Only in a few cases has the tetragonal ε-Ti_2N phase been reported in films grown from the vapor phase [111]. These usually have high compressive stress values and they have not yet found any commercial use. The exact phase composition is given by the kinetic conditions during growth.

Other common nitrides are the group 4 nitrides ZrN and HfN, both of which have properties similar to those of TiN. Bulk HfN has been reported to have a higher hot-hardness than TiN, and has been claimed to be beneficial in high-temperature applications, but so far HfN films have not been used in commercial wear-protective or decorative applications.

Except for CrN, transition metal nitrides from groups 5 and 6 are not used for wear-protective or decorative purposes.

Alloy Nitrides. Most transition metal nitrides are completely soluble in each other, making it possible to vary the properties of the coatings over a large range. The most studied alloy nitride is $Ti_{1-x}Al_xN$ [112]. Even though the NaCl-structure TiN and the

wurtzite-structure AlN are not mutually soluble, metastable $Ti_{1-x}Al_xN$ films with NaCl structure can be grown for $0.4 < x < 0.6$, depending on the growth conditions [113]. For growth of such metastable phases it is necessary to use strong kinetic limitations, through low growth temperatures, high growth rates, or ion-induced collisional mixing during growth. The metastable $Ti_{1-x}Al_xN$ phase decomposes above 900 °C to TiN and AlN [114].

$Ti_{0.5}Al_{0.5}N$ has a much higher oxidation resistance than TiN, owing to the formation of a protective Al-rich oxide layer, beneficial for high-temperature applications [114]. The oxidation temperature for $Ti_{0.5}Al_{0.5}N$ and a few other hard coatings are as follows [104]:

Coating	Oxidation temperature °C
TiC	400
TiN	550
CrN	700
$Ti_{0.5}Al_{0.5}N$	750
$Cr_{0.5}Al_{0.5}N$	910

The possibilities of using alloy nitrides to increase the oxidation temperature is clearly demonstrated. The structural and mechanical properties of $Ti_{1-x}Al_xN$ are similar to those of TiN. The electrical and optical properties depend on the Al content (the color changes from golden to dark violet as x is increased from 0 to 0.5).

Carbides. The hard carbide coatings used resemble those of nitride hard coatings, i.e., carbides from group 4 (TiC, ZrC, and HfC) dominate. Also, B and Si carbides have been studied for wear-protective applications. The structures of the transition metal carbides resemble those of the corresponding nitrides; all monocarbides in groups 4 and 5 have a B 1-NaCl structure. The hardness of the carbide is in general higher than that of the corresponding nitride (more pronounced covalent bonding) [102, 103]. Compared with nitrides the growth process is more complex (possibility of C precipitation as the C/metal ratio approaches 1, caused by the complex cracking patterns and chemisorption behavior of the hydrocarbon precursors). As free C precipitates, the film structure changes drastically to porous films with very small grain sizes [103].

Carbonitrides. Since most transition metal carbides and nitrides are completely soluble in each other, carbonitrides are sometimes used. As in the case of alloy nitrides, the properties can be varied continuously from those of the carbide to those of the nitride. Only TiCN coatings have been studied to any significant degree [103].

Multilayers. By using multilayers of different materials it is possible to take advantage of the characteristic features of each material and to obtain synergetic effects [115, 116]. The main reasons for using multilayers of nitrides and carbides are to facilitate strong adhesion, to obtain wear-protective films with low chemical reactivity and low friction, and to increase strength and hardness. Multilayers of the type $TiN/Al_2O_3/TiC$ deposited on cemented carbide substrates by CVD are used routinely in order to optimize adhesion and wear properties on cutting tool inserts [110].

By depositing nitride multilayer films with individual layer thicknesses of a few nanometers, significantly increased hardness values have been reported for films with total thicknesses of a few micrometers [115–122]. Examples of systems with hardness more than twice that of the individual materials include TiN/VN [117], TiN/NbN [118], and $TiN/V_{0.6}Nb_{0.4}N$ [119]. Figure 31 shows the variation in hardness for TiN/VN and TiN/NbN multilayers as a function of the multilayer wavelength Λ (the sum of the individual layer thicknesses). The enhanced

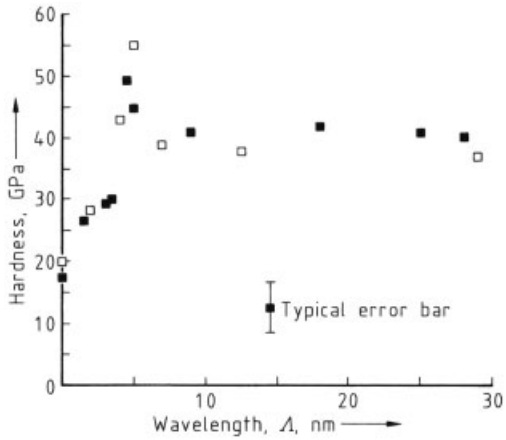

Figure 31. Vickers hardness of TiN/VN (open squares) and TiN/NbN (full squares) multilayers as a function of modulation wavelength [122]
Total film thicknesses 2.5 – 3 µm in both cases; hardness values for $\Lambda = 0$ nm correspond to the TiVN and TiNbN alloy values.

hardness can be explained by hindrance of dislocation motion across the layer interfaces (different dislocation line energies in the two materials) and by coherency strain [119]. Metal/nitride multilayers such as Ti/TiN, Hf/HfN, W/WN [121], and Ti/BN [122] have very high hardness values.

2.3. Carbon Films (General references [123–128])

2.3.1. Carbon Phases: Composition and Stability

Diamond and graphite are well-defined, long-range-ordered crystalline allotropes of carbon, that occur naturally. Structure, properties, and quality of vapor-deposited thin film variants are best judged by comparison with their natural equivalents. Other carbon phases are less well defined. They do not occur naturally, but are grown as thin films by chemical and physical vapor deposition (CVD and PVD) methods. Composition and properties of these phases depend strongly on the preparation conditions. Resulting differences are manifested by a confusingly long list of names: amorphous hydrogenated carbon (a-C:H), amorphous carbon (a-C), ion-beam-deposited carbon (i-C), mass-selected ion-beam-deposited amorphous carbon (MISB a-C), tetrahedral amorphous carbon (ta-C), and diamond-like carbon (DLC). DLC is also frequently used for the whole class of amorphous carbon phases. They may contain considerable quantities of hydrogen, varying amounts of sp^3- and sp^2-hybridized carbon atoms. The ternary phase diagram (Fig. 32) is a convenient way to display the compositional differences [129, 130], to correlate them with different properties, and to outline the relation between the various carbon phases. The sp^2-vertex represents graphite, a hydrogen-free crystalline material that consists of aromatic carbon layers built from threefold-coordinated sp^2- carbon atoms. The sp^3-vertex corresponds to diamond, hydrogen-free, cubic crystals formed by exclusively covalently bound, fourfold-coordinated sp^3- carbon tetrahedra. Noncrystalline, hydrogen-free carbon phases, called either a-C or, with more than 80 % sp^3- carbon, ta-C, are located on the sp^2 – sp^3 side line. Films with a C – C skeleton are feasible only up to a certain level of hydrogen in the solid. At higher H concentrations films do not form (Fig. 32, dashed line [130]). The compositional coordinates of the various carbon thin films in Figure 32 determine whether their physical properties are closer to diamond (high sp^3-C, low H content), to graphite (high sp^2-C, low H content), or more like polymers (high H content).

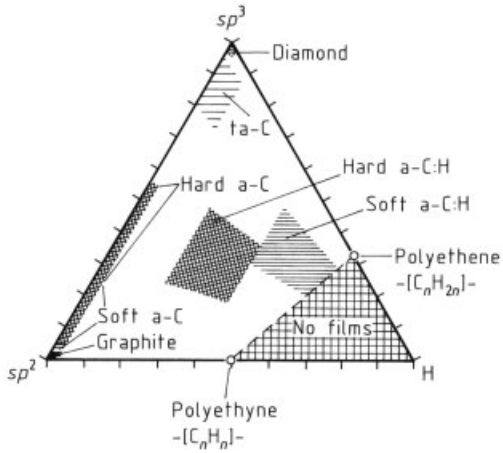

Figure 32. Ternary sp^3-C/sp^2-C/H diagram for carbon films ([129, 130], modified)

Of all the carbon phases mentioned, only graphite is thermodynamically stable. All other carbon phases are metastable, and, although reasonably stable for many applications at room temperature, convert to graphitic material on heating in an inert atmosphere. While a-C:H and a-C change their structure at 300 – 600 °C [131], ta-C seems to tolerate more than 1000 °C [132]. Diamond graphitizes rapidly at temperatures > 1500 °C [124, 128]. Composition, structure, and stability define the physical and chemical properties of the films and, along with the deposition temperature, define their application areas.

2.3.2. Properties and Applications of Diamond Films

There is no other material with such a combination of properties as diamond. It is the hardest of all materials. At room temperature, it has the highest thermal conductivity. Diamond is transparent from the UV to the microwave region of

Table 9. Properties and applications of diamond films

Property		Comparison with competitors	Possible applications
Indentation hardness, kg/cm^2	10 000	β-BN < 4500; SiC < 4000; WC = 2190; steel = 800	drill bits, polishing material, cutting tools
Coefficient of friction	0.05 – 0.1 (only in air)	PTFE ≈0.05	ultra-low-wear coatings
Young's modulus, GPa	1050	2× the value of alumina; 5× that of steel	stiff membranes for lithography masks, lightweight diaphragms for audio devices
Sound propagation velocity, km/s	18.2	1.6× the value of alumina	
Chemical inertness	oxidation at T > 600 °C	at room temperature, resistant to all acids, bases and solvents	coatings for reactor vessels
Range of high transmittance, µm	0.22 – 3 and>6	in the IR orders of magnitude lower than most other materials	UV – vis – IR windows and coatings, microwave windows, X-ray windows (low Z)
Refractive index	2.41	1.6× the value of silica	
Band gap, eV	5.45 (negative electron affinity)	1.1 for Si; 1.43 for GaAs; 3.0 for β-SiC	high-power/high-temperature electronics, e.g., thermistors, transistors, particle detectors, cold cathodes for displays, vacuum microelectronics
Electron/hole mobility, cm^2 V^{-1} s^{-1}	1900/1600	1500/600 for Si; 8500/400 for GaAs	
Dielectric constant	5.5	11 for Si; 12.5 for GaAs	
Thermal conductivity, W m^{-1} K^{-1}	1300 – 2200	highest at room temperature; 4× the value of Cu or Ag	heat sinks, heat spreaders
Thermal expansion coefficient, K^{-1}	0.8×10^{-6}	at room temperature, similar to silica (0.57×10^{-6})	thermally stable substrates, e.g., for X-ray lithography masks

the spectrum (except for a small gap between 3 and 6 µm). It is an excellent insulator, but can be doped and converted into a low-resistivity semiconductor. Diamond is chemically inert at room temperature and, owing to its high atom density, is extremely rigid. The properties of the best vapor-grown diamond films are identical with those of natural diamond. Significant differences between natural and vapor-grown material result from the polycrystalline structure of the latter. Polycrystalline diamond films usually exhibit a columnar growth morphology with crystal grains of 0.1 to> 100 µm. The grain boundaries tend to be decorated with contaminants and, depending on the deposition conditions and process, anything from black films with considerable amounts of non-diamond carbon phases, other impurities (Si, O, etc.) and several percent hydrogen, to transparent material with less contamination than natural diamond, is feasible. Table 9 lists the physical properties of diamond, in comparison with some competing materials and possible applications of vapor-grown diamond films.

Mechanical Properties and Applications. The hardness and wear resistance of diamond and its resulting application as a material for tools and abrasives is common knowledge. Diamond CVD permits growth of polycrystalline films of at least comparable hardness and wear properties, but free of the cobalt binder that is required when compacting high-pressure diamond powder into tool bits. Binder-free, polycrystalline films have an advantage over single crystals. Cleavage planes are absent, because of the random orientation of crystallites. Grain boundaries prevent crack propagation and the material is tougher and mechanically isotropic. The Young's modulus of CVD diamond is identical to that of single crystals: 1050 GPa. The rupture strength of 10 – 20 mm diameter CVD diamond plates with thicknesses of 150 – 300 µm is ca. 1000 ± 200 MPa, a factor of two lower than for single crystals. The fracture toughness is 6 MPa m$^{0.5}$, higher than the 3.5 – 5 MPa m$^{0.5}$ value reported for single crystals. The test performance of thick CVD diamond plates is hardly affected by the surface morphology of top-quality material, i.e., unpolished disks show the same values as polished material (for mechanical data see [133, 134]).

For films with thicknesses< 50 µm, the phase purity and hence all film properties depend strongly on the thickness. Vapor-grown, free-standing (substrate-free) diamond plates can be used as single-point tools or, if brazed onto a tool insert, can replace high-pressure synthetic poly-

crystalline diamond compacts, i.e., in turning, milling, drawing, or bonding tools [135]. Diamond can also be directly deposited onto various tool materials, e.g., tungsten carbide or SiAlON inserts. With appropriate surface treatment, the adhesion of such films is sufficient for practical applications. Coated tools require a minimum of diamond. A $< 10\,\mu m$ thick film is usually sufficient. They perform as well or better than conventional diamond tools, without their size limitations and often at a lower price [135]. Owing to the complex shapes required, directly deposited diamond layers are specially suitable for drilling tools [135]. The excellent mechanical properties are also required for other diamond applications, e.g., as high-pressure or vacuum windows.

Optical Properties and Applications. Type IIa natural diamond is optically transparent from its band edge at $0.220\,\mu m$ to ca. $3\,\mu m$. The lattice vibrations of diamond are excited between 3 and $6\,\mu m$, hence diamond is weakly absorbing in this region. At higher wavelengths it transmits electromagnetic radiation up to the microwave region of the spectrum [124, 128]. Figure 33 sketches the optical transmission spectra of diamond and other common optical window materials, e.g., LiF (UV windows), silica (visible), and CsI (IR) [136]. Transmission losses of 31 % for diamond are due to reflections at two diamond – air interfaces (refractive index ca. 2.41). The best quality CVD diamond films are now approaching the characteristics of type IIa natural diamond [137]. Problems are related to:

1) Excess absorption between 220 and 350 nm, owing to nitrogen incorporation in the solid
2) Excess absorption between 220 nm and $3\,\mu m$, owing to non-diamond carbon phases
3) Scattering losses due to the surface roughness of the polycrystalline structure (affected wavelength region depends on grain size)

While excess absorption can be avoided by contamination-free deposition procedures, scattering at the rough growth surface is inherent to polycrystalline material. Either by polishing (for thick diamond plates) or by deposition of very fine-grain material (suitable for thin films and membranes), scattering in the visible region can be minimized. For top-quality material, the grain boundaries inside thick polished plates hardly affect the transmission properties.

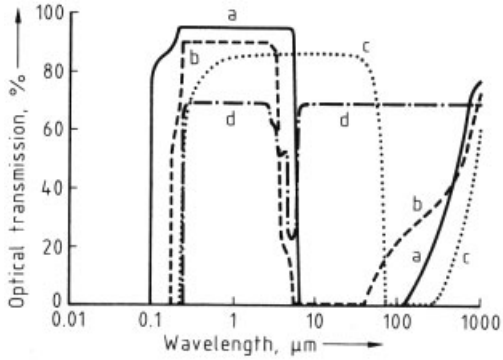

Figure 33. Optical transmission spectra for diamond and other optical window materials [136]
a) LiF; b) Silica; c) CsI; d) Diamond

Vapor-grown diamond films and plates can be used as windows for electromagnetic radiation of any kind. With its unique combination of mechanical strength, chemical inertness, high transparency, and low atomic number, diamond is particularly useful in harsh environments and as a protective optical coating for sensitive materials such as IR optics.

Thermal Properties and Applications. At room temperature, diamond is the best thermal conductor available; at 20 °C, values of 2000 – 2200 $W\,m^{-1}\,K^{-1}$ were measured for type IIa natural single crystals – four times the value of copper or silver. The thermal conductivity of diamond peaks at ca. 65 K with values around 17 000 $W\,m^{-1}\,K^{-1}$. Between room temperature and 500 °C, the thermal conductivity κ decreases, following the empirical formula [138]:

$$\kappa = 27\,350\,T^{-1.26}$$

Top-quality material has conductivity ca. 700 $W\,m^{-1}\,K^{-1}$ at 500 °C, still twice the value of copper or silver (Fig. 34). The transport of thermal energy in diamond is a phonon-coupled phenomenon. Despite the polycrystalline nature and the numerous grain boundaries that could disturb phonon propagation, CVD diamond films with the thermal properties of the best natural single crystals have been produced [137–140]. Film thicknesses $> 100\,\mu m$ with grains larger than $20\,\mu m$ are required to obtain values above 1800 $W\,m^{-1}\,K^{-1}$. Thin films suffer from contamination by excess non-diamond carbon phases and from phonon scattering at the

grain boundaries of numerous small ($< 5\,\mu m$) crystals. For $4\,\mu m$ thick films, values of ca. $700\,W\,m^{-1}\,K^{-1}$ are feasible [140]. Isotopically enriched ^{12}C-diamond single crystals as well as films exhibit a thermal conductivity 20 – 50 % higher than that of natural (99 % ^{12}C) material [141].

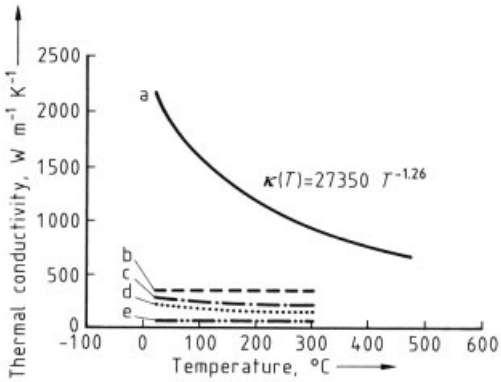

Figure 34. Temperature dependence of the thermal conductivity of diamond and other commonly used heat spreader materials
a) Type IIa natural and best quality CVD diamond; b) Cu; c) BeO; d) AlN; e) Al_2O_3

In addition to its extremely high thermal conductivity, diamond has a very low thermal expansion coefficient, $0.8\times10^{-6}\,K^{-1}$ at room temperature. However, this value applies only to single crystals and top-quality thick CVD diamond plates. Non-diamond carbon contamination in thin films may increase the values to $> 2\times10^{-6}\,K^{-1}$.

The application of high-thermal-conductivity diamond as a highly efficient heat spreader for high-power electronic devices and high-thermal-load laser windows is well established. CVD films make it possible to overcome the cost and size limitations imposed by the use of natural single crystals and may compete in applications currently served by poisonous BeO or AlN.

Electrical Properties and Applications. Pure diamond is electrically insulating with resistivity $> 10^{11}\,\Omega\,cm$ and breakdown electrical field 10^7 V/cm, which is similar to silica [124, 128]. The values for polycrystalline films are lower [136] and depend on the thickness. Again, grain boundary decoration with non-diamond carbon phases poses a problem and only ultra-pure CVD diamond approaches the breakdown voltage values of single crystals. The resistivity of diamond can be modified either by UV illumination or by doping with boron. For both natural boron-doped crystals and vapor-grown films, resistivities < 0.1 to $100\,\Omega\,cm$ are feasible. Boron-doped diamond is a *p*-type semiconductor with activation energy ca. 0.38 eV [136, 142]. The feasibility of true *n*-type doping of diamond with significantly low resistivity is still under discussion. Phosphorous, nitrogen, or lithium are possible *n*-type dopants. Implantation of C^+ may also lead to *n*-type conductivity. In all cases, the material is still highly resistive and $p-n$ junctions are not yet feasible. The carrier mobilities in undoped bulk diamond are $2200\,cm^2\,V^{-1}\,s^{-1}$ for electrons and $1600\,cm^2\,V^{-1}\,s^{-1}$ for holes [124, 128]. The polycrystalline nature of CVD diamond severely affects the mobilities. For randomly oriented boron-doped material, carrier mobilities $< 50\,cm^2\,V^{-1}\,s^{-1}$ are common. At room temperature, highly oriented (100)-textured boron-doped films exhibit values of $300-400\,cm^2\,V^{-1}\,s^{-1}$ and for homoepitaxially grown boron-doped films 800 – $1000\,cm^2\,V^{-1}\,s^{-1}$ has been reported [142]. Device operation fieldeffect transistor, FET at 300 – 600 °C is feasible. For (100)-oriented boron-doped films, the carrier concentration increases from $10^9\,cm^{-3}$ at room temperature to $10^{16}\,cm^{-3}$ at 500 °C. Unfortunately, the hole mobility was found to drop from $350\,cm^2\,V^{-1}\,s^{-1}$ at room temperature to $120\,cm^2\,V^{-1}\,s^{-1}$ at 500 °C [142]. The properties of CVD diamond now allow the construction of the first high-power, high-temperature electronic devices (e.g., transistors, diodes, and sensors). Tremendous progress has been made over the past decade: prototype devices operating at 500 °C are available and the first logic circuits [143] have been demonstrated. However, to implement a diamond-based semiconductor technology still needs a lot of work.

Diamond is not only interesting as an active device and sensor material, but has unusual electronic surface properties which might be exploited in cold cathodes. The (111) surface of natural boron-doped diamond was found to exhibit "negative electron affinity" (NEA) [144]. This means that the energy level of the conduction band of diamond is slightly higher than

that of electrons in vacuo. NEA materials emit electrons once they are excited across the band gap (5.45 eV for diamond) by, e.g., UV light or high electric fields into the conduction band. Diamond might, therefore, serve as a cold electron source for various applications, e.g., flat panel displays. The NEA properties depend strongly on the surface state of the diamond; H- and HO-terminated surfaces emit electrons while reconstructed diamond surfaces show no emission. CVD diamond films display NEA properties and large-area electron-emitting surfaces seem feasible.

2.3.3. Properties and Applications of a-C:H, a-C, and ta-C Films

Amorphous carbon phases are also called diamond-like carbon (DLC) films, referring to the hardness of some of these materials. Hardness, and the fact that the films are carbon-based, are about the only relationship between diamond and some members of the DLC family. Figure 32 illustrates that DLC usually consists of a mixture of sp^2- and sp^3-hybridized carbon and, depending on the preparation conditions, may contain up to 65 atom % hydrogen. The properties of the different DLC types depend strongly on the composition. They range from polymeric, soft, transparent, and insulating, to hard, dark, and conducting. In all cases the films are amorphous and the properties are isotropic.

Figure 35 correlates the hydrogen content, density, optical band gap, and hardness of DLC films grown under different deposition conditions, i.e., different bias voltages.

Mechanical and Chemical Properties and Applications. The hardness, low friction coefficient, and wear resistance are the most useful properties of some members of the DLC family. If deposited properly, a-C:H films can reach a density of 2.0 g/cm^3 and a hardness of 40 GPa (diamond: 100 GPa). The values for soft, hydrogen-rich a-C:H films and soft, sp^2-C-rich a-C are ca. 2 – 5 GPa. With 90 – 100 GPa and densities of 2.5 – 3.1 g/cm^3, mass-selected, ion-beam-deposited hard a-C are closest to diamond (see also Fig. 32). Unfortunately, the thickness of these films is limited by the high internal compressive stress [132].

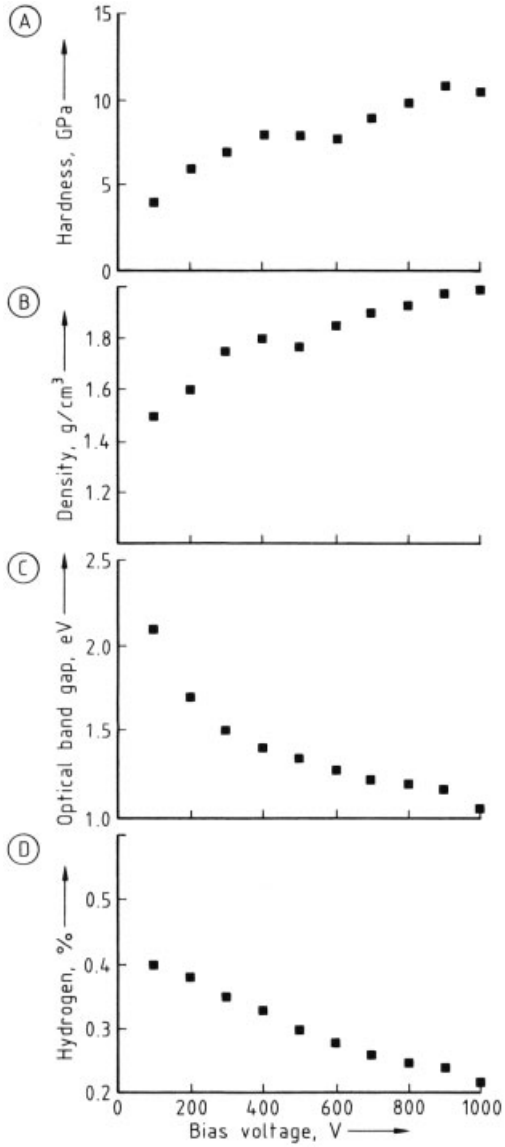

Figure 35. Correlation of A) Hardness; B) Density; C) Optical band gap; and D) Hydrogen content of DLC films grown from benzene by r.f. plasma CVD at different bias voltages [130]

Friction coefficients < 0.1 are common for DLC. Like that of diamond, the coefficient of friction is very sensitive to humidity, but unlike diamond, it is reduced in dry air or vacuum.

Chemically, DLC films are inert to most acids and bases at room temperature. At higher temperatures they are attacked. Moreover (see the discussion of stability in Section 2.3.1) they

change their structure and ultimately turn into graphite films [130, 132, 145, 146].

Because of their excellent mechanical and chemical properties and the fact that it is possible to deposit DLC films at room temperature, they are widely used as protective, low-wear, low-friction, or antireflection coatings on, e.g., IR optics, sunglasses, magnetic recording disks, or metal film recording tapes. Although the properties are inferior to diamond, the low-temperature growth option is a big advantage of DLC films over diamond for such applications [146, 147].

Biomedical Properties and Applications. Carbon has been used for many biomedical applications. Tissue can adhere to it and a protein layer that forms on its surface prevents the formation of blood clots. The ideal solution for biomedical implants would be a combination of the mechanical strength of metals and the biocompatibility of carbon. DLC films deposited onto metal implants seem to provide this combination [146, 147].

Optical and Electrical Properties and Applications. The optical properties of DLC films are also highly dependent on the sp^2/sp^3 ratio and the hydrogen content. Films with a high hydrogen content are transparent in the visible with optical band gaps of up to 4 eV. Hard a-C and a-C:H films are less transparent with gaps of 1 – 2.5 eV. With a gap of 0.4 – 0.8 eV, soft, graphite-like a-C films are hardly transparent. The optical properties of DLC are often exploited in conjuction with the mechanical properties, e.g., in protective coatings for optical components [130, 145, 147].

Electronically, all DLC films are best described as composite materials, consisting of conductive, graphitic islands embedded in an insulating matrix. Like diamond, ta-C films show photoconductivity, however, white light is sufficient to change its conductance by two orders of magnitude at 125 K [132]. Doping with nitrogen changes this effect, and also effects the electrical conductivity of the material. Pure ta-C is a good insulator ($10^{-9}\ \Omega\,\mathrm{cm}$), N-doped ta-C is more conductive [132]. Doping of DLC films has been investigated, but device applications of DLC films, if any, are still in their infancy.

2.4. Metallic Films [148–160]

2.4.1. Introduction

Techniques for the preparation of thin films include: electrodeposition, hot dipping, cementation, diffusion, flame spraying, plasma spraying, laser coating, roll bonding, powder rolling, mechanical impingement, explosive bonding, chemical vapor deposition, physical vapor deposition, sputtering, and ion implantation. The properties of metallic coatings depend on the technique used and on the processing conditions; the last decade has seen the increasing use of dry techniques, such as vapor-phase deposition, which are becoming competitive with wet processing techniques in some areas of surface finishing.

Thin metallic films find extensive use in the electronics and engineering industries, as coatings on a substrate or as detached films (e.g., electroforming of foil or mesh). By selection of deposition technique and process conditions, it is possible to produce films in the range 0.1 – 20 μm of materials which include compact single metals and alloys, composite metal – polymer and metal – ceramic layers, conductive polymers, and solid-state ceramics and dispersed coatings of catalysts on a porous substrate (Table 10).

2.4.2. Electronics Applications

Thin metallic films may be used to alter the electrical properties of polymer, ceramic, and metal substrates, or the substrate may be viewed as a support for the production of a specific film. Examples include the metallization of laminates, such as glass-reinforced polyester and epoxy materials to produce printed circuit board tracks, switch contacts, and connectors. Following catalytic activation of the surface by a thin palladium deposit, copper may be deposited on a plastic printed circuit board laminate by electroless plating or by screen printing of copper-containing inks. The film may be thickened by electroplating of copper to produce conductive tracks. For edge connectors, further electrodeposition of nickel, then gold, is common, while solderability can be provided by the electrodeposition of tin – lead alloys. Porosity and adhesion problems require care in the formulation

Table 10. Thin film materials and their applications

Material	Typical film thickness, μm	Deposition technique	Applications
Zinc	75 – 100, 7 – 20, 20 – 40	hot dipping diffusion electroplating	corrosion protection of steel
Chromium	10 – 120	electroplating	corrosion- and wear-resistant coatings on steel
Copper	2 – 10	electroless plating, electroplating	conductive tracking
Gold	1 – 5	electroplating	electronic contacts
Copper, nickel, iron, Ni – Co	10 – 1500	electroforming	metal foils and flat meshes; audio and video disk stampers; waveguides; electric razor "foils" and liquid food filter meshes; molds for plastics; magnetic display boards (Ni and Fe)
Nickel – phosphorus	5 – 30	electroless plating	bearing and valve surfaces
Nickel – boron	5 – 30	electroless plating	high-temperature wear-resistant coatings
Dispersed platinum	(1 – 10 g/m^2)	electroplating, chemical deposition	fuel cell electrodes, electrolytic conductivity probes
Tin – lead	5 – 10	electroplating	solderable tracks on printed circuit boards
Cobalt – nickel – gold	2 – 10	electroplating	edge connectors on printed circuit boards
Nickel – silicon carbide	5 – 30	electroplating, electroless deposition	abrasive wheels and cutting tools
Nickel – PTFE	5 – 30	electroplating, electroless deposition	self-lubricating, non-stick coatings
Titanium nitride, titanium carbide	4 – 8	chemical vapor deposition, plasma deposition	wear-resistant cutting tools
Alumina	15 – 25	anodizing	corrosion-resistant coatings on aluminum
Tantalum oxide	1 – 8	anodizing	high-stability capacitors
Molybdenum disulfide	4 – 10	chemical vapor deposition	lubrication of internal combustion engine surfaces

of pretreatment sequences and rinses between coating stages. Conductive inks are cost effective and versatile for the formation of conductive tracks on printed circuit boards and for other applications (Table 11). They consist of a conductive pigment (metal flakes or particles), a binder (polymer resin), and a solvent or dispersant. The pattern is formed by forcing the liquid ink through a masked mesh or screen. The ink is then cured (usually by heat) to remove the solvent. The resin contracts, forcing the pigment particles together to produce a conductive coating. After curing, the coating thickness is usually 3 – 12 μm, although thicker coatings are possible. Conductive inks are printed onto a range of substrates, principally ceramics, printed circuitboards, and flexible plastics (e.g., polyester or polycarbonate). It is possible to electroplate metals onto some cured inks to increase conductivity and durability. By blending metal pigments it is possible to combine the properties of the constituent metals.

Table 11. Conductive inks and their applications

Substrate	Metal(s)	Applications
Ceramics	gold, silver, copper, silver – palladium blend	conductive tracking; capacitors
Printed circuit boards	silver, copper	conductive support for tracking
Flexible polymers	silver	conductive tracking; keyboards; switches; RFI∗ shielding; medical sensors; communications antennae
	copper	flexible printed circuits
	nickel	RFI shielding; magnetic shielding; environmental protection
	silver – carbon	environmental protection; chemical sensors; resistors; keyboards
	carbon – rubber	force-sensitive resistors
	silver – nickel	conductive tracking; magnetic conductors; phone cards
	iron oxides	magnetic cards, "smart" cards

∗ RFI = radiofrequency interference.

A variety of gas-phase techniques can be used to produce special materials. Reactive sputtering can produce compound semiconductor coatings of CdS by sputtering from a Cd target in an $Ar + H_2S$ atmosphere. The ability of sputtering to control coating composition makes it a useful technique in the electronics industry (metallization of aluminum alloy microcircuits, refractory metal microcircuits, optically transparent conducting electrodes, piezoelectric transducers, optically addressed memory devices, thin film resistors and capacitors, video disks, and thin film lasers). Sputtering can also produce

multilayer stacks of dielectric materials such as TiO_2, Ta_2O_5, Al_2O_3, and SiO_2, which are harder and denser than evaporated layers.

2.4.3. Engineering Coatings

As seen in Section 1.7.1, a wide range of single metals, alloys, and composites can be electrodeposited or electrolessly plated from aqueous solution. Examples include electroless nickel–phosphorous coatings for wear resistance on valves and corrosion protection, and chromium coatings for wear resistance under arduous conditions (hydraulic cylinders, gun barrels). Special applications include the use of Ni–PTFE inclusion deposits for carburetor butterfly valve spindles and the use of Co–WC as a wear-resistant coating for cutting tools.

In dry processing techniques, flame and plasma spraying can produce wear-resistant layers of materials such as stainless steel alloys, WC, Si_3N_4, SiC, and Al_2O_3. Sputtering, in combination with heating of the substrate (or the use of a reactive atmosphere) can produce hard, wear-resistant coatings such as titanium nitride. A 3-mm TiN film with hardness > 2000 Knoop/mm^2 can be produced for high-temperature hobs. Solid lubricants can greatly decrease wear, e.g., sputtered films of MoS_2 can have a coefficient of friction of ca. 0.04 against steel, with lifetime more than 100 times that of brushed-on MoS_2 films.

For some metals, a number of coating technologies are available and the choice depends on the environment and service lifetime, as well as cost. In corrosion protection, sacrificial coatings of zinc on steel may be produced by electroplating, hot-dip galvanizing, or high-temperature diffusion (Sheradising). Hot-dipped coatings have thickness 20–200 μm, the usual range being 70–100 μm. A series of Fe–Zn alloy layers are produced, the top layer consisting of pure zinc if a low-temperature process is used. The result is excellent adhesion and reasonable resistance to abrasion; this coating process may be chosen when arduous conditions are involved, a long lifetime is sought, or if no further coatings (e.g., paints) are to be applied. Zinc diffusion coatings are usually restricted to smaller workpieces, e.g., fasteners, metal brackets. The surface of the steel is converted to a film of iron–zinc alloys, in a batch process involving zinc powder at elevated temperature, with typical coating thickness 7–25 μm. The coating has reasonable abrasion resistance and uniformity, and is used when no further anticorrosion coatings are applied. For large-scale or continuous coating (e.g., steel coil or sheet) or for substrates that are deformed by high-temperature processes, electroplating of zinc (or an alloy with higher corrosion resistance, e.g., Zn–Ni, Zn–Fe, Zn–Co) is often used, as in car body panels. In electroplating, the films tend to have less toughness and resistance to abrasion, owing to the absence of intermetallic, alloy layers.

Zinc, aluminum, magnesium, and iron alloy surfaces can be chemically modified by a conversion coating to produce, e.g., oxides (anodizing of aluminum), chromates, and phosphates. Such treatments can significantly improve corrosion resistance and, in the case of phosphating, lubrication conditions can also be improved.

For noble coatings, e.g., electroless nickel, electroplated nickel, or chromium, the coating must have low porosity, as pitting of the underlying steel can occur at coating defects. The stress distribution and cracking pattern in chromium-plated layers is important in long-term performance, both for corrosion protection and tribology.

Gas-phase techniques have specialist uses in corrosion protection, e.g., ion plating to coat aircraft and spacecraft steel and titanium fasteners with aluminum. The aluminum layers have good uniformity, adhesion, strength, and mechanical properties, and have, in some cases, replaced cadmium electroplated layers (which tended to suffer from hydrogen embrittlement and stress cracking). Ion plating of aluminum is also used to coat thin-walled stainless steel heat exchanger tubes and interconnecting sections. Other applications include Cr, Al_2O_3, Si_3N_4, Cd, Ti, C, and Ta on steel for medical, biochemical, and industrial chemical applications.

2.4.4. Dispersed Catalytic Coatings

The majority of metallic films are compact, uniform-thickness layers on a nonporous substrate, but an important class of coatings are produced by dispersing a precious metal onto

a porous substrate, to produce a catalytically active, high surface area coating which uses a minimum of the costly coating material. Examples include:

1) Deposition of platinum or platinum alloys onto porous carbon paper, carbon cloth, or ionically conductive polymer membrane to produce porous gas diffusion electrodes. These are used in aqueous electrolyte fuel cells, speciality batteries using gaseous reactant (e.g., the aluminum – air cell), in electrochemical gas sensors, and as a hydrogen anode. The electrode surface must remain electrically conductive and act as a three-phase boundary at the interface of gases, liquid electrolyte, and solid catalyst, and it is common to control the wetting properties of the electrode to avoid flooding of the structure, by the use of PTFE which also acts as a binder.
2) The deposition of platinum-group metals and alloys onto porous metal and ceramic substrates to produce catalytic converters for car exhausts. The ability to recycle the expensive catalyst is an important consideration.

In both of these applications, there is a trend toward lower loadings of the precious metals. In fuel cell electrodes for phosphoric acid H_2/O_2 cells, the loading of platinum metal on carbon has decreased toward ≤ 1 mg/cm^2. Techniques to produce dispersed coatings include combinations of chemical impregnation, electrodeposition, and thermal treatment. More recently, sputtered and other vapor-phase-technology coatings have been considered, although they are much more costly than chemical impregnation followed by heat treatment.

2.4.5. Developments and Trends

The range of thin metallic films continues to increase as more advanced techniques are implemented and new alloy, polymer, and composite coatings are explored:

1) Increasing use of vapor-deposition techniques, particularly chemical vapor deposition and ion implantation, to produce very thin, nonporous layers for the electronics and engineering industries
2) Introduction of multilayer coatings by a number of different techniques or the alteration of process conditions during the application of a single technique
3) Increasing use of polymer, alloy, and composite coatings
4) Development of conductive polymer films with electrical conductivity ranging from that of an insulator to about one-tenth the value of some metals
5) Increasing need to use thin films of precious metal catalysts on a less costly, engineered substrate (fuel cell electrodes, specialized gas sensors, and car catalytic converters)
6) Development of cost-effective, environmentally attractive techniques for selective deposition of thin films (ion plating, laser or jet electroplating, and screen printing by jet techniques)
7) Formation of thin diffusion coatings by ion implantation

2.5. Superconductive Coatings

Since the discovery of high-temperature oxide superconductors [161], coatings (deposition and/or thin film growth) of several cuprates with a superconducting transition temperature T_c above liquid nitrogen temperature, 77.3 K, have been developed. Among the many high-T_c oxide materials synthesized to date, representative thin film materials for practical use are $LnBa_2Cu_3O_{7-\delta}$ (where Ln is a rare-earth element other than Pr, Ce, and Tb) with $T_c \geq 90$ K, and $Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ with T_c ca. 20 K for $n = 1$, ca. 80 K for $n = 2$, and ca. 110 K for $n = 3$. Deposition of $LnBa_2Cu_3O_{7-\delta}$ thin films is well established by different deposition techniques, while that of the $Bi_2Sr_2Ca_{n-1}Cu_nO_{2n+4}$ system for $n = 2$ and 3 is being investigated. Thin film deposition of $Tl_2Ba_2Ca_{n-1}Cu_nO_{2n+4}$ ($n = 1, 2, 3$) is limited because of the toxicity of Tl. A more recent development is thin film deposition of $HgBa_2CaCu_2O_{6+\delta}$ by reactive r.f. magnetron sputtering, followed by oxygenation annealing. The film shows a relatively sharp transition up to ca. 120 K [162].

Typical methods for in situ deposition of high-T_c cuprate thin films are physical deposition, e.g., r.f. sputtering, pulsed excimer laser

ablation or pulsed laser deposition (PLD), and reactive coevaporation, sometimes referred to as molecular beam epitaxy (MBE). Radiofrequency magnetron sputtering was first used for deposition or epitaxial growth of $YBa_2Cu_3O_{7-\delta}$ thin films on MgO and $SrTiO_3$ substrates; PLD with ArF ($\lambda = 193$ nm) and KrF ($\lambda = 248$ nm) excimer lasers is a promising method for $LnBa_2Cu_3O_{7-\delta}$ thin films with $T_c > 90$ K [163]. It can now produce in situ deposited multilayered films with different targets installed in a deposition chamber. Figure 36 shows a typical transmission electron micrograph of a $YBa_2Cu_3O_{7-\delta}/PrGaO_3/YBa_2Cu_3O_{7-\delta}/PrGaO_3$ quadrilayered film grown in situ on a $SrTiO_3$ substrate. The great advantage of MBE is its ability to control deposition sequences by observing oscillations of the reflection high-energy electron diffraction (RHEED) intensity [164], which enables "layer-by-layer" deposition of constituents and fabrication of superlattices, e.g., $PrBa_2Cu_3O_x/YBa_2Cu_3O_x$.

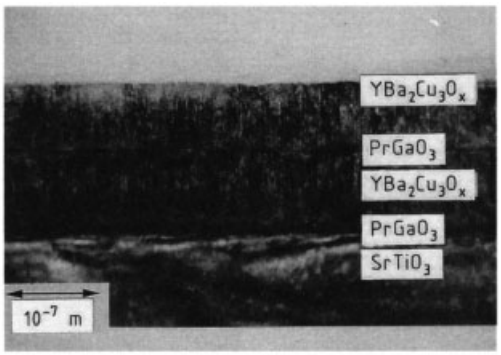

Figure 36. Transmission electron micrograph of a $YBa_2Cu_3O_{7-\delta}/PrGaO_3/YBa_2Cu_3O_{7-\delta}/PrGaO_3$ quadrilayered structure on a $SrTiO_3$ substrate deposited in situ by pulsed laser deposition

Most $YBa_2Cu_3O_{7-\delta}$ films grown by physical deposition contain spiral growth islands [165] originating from screw dislocations caused by relatively high supersaturation (Fig. 37). Surface roughness is of the order of 0.01 µm.

Physical depositions require low supersaturation of cuprate-generating species, oxygen pressure control, and choice of epitaxial substrate materials. For supersaturation, the most advanced techniques are metallorganic chemical vapor deposition (MOCVD) for Bi systems, and liquid-phase epitaxy (LPE) for $LnBa_2Cu_3O_{7-\delta}$ systems. Current MOCVD research has obtained high-quality *c*-axis-oriented $YBa_2Cu_3O_{7-\delta}$ thin films with T_c of ca. 90 K and $Bi_2Sr_2CaCu_2O_6$ ($n = 1$, 2212 phase) thin films with a T_c ca. 80 K.

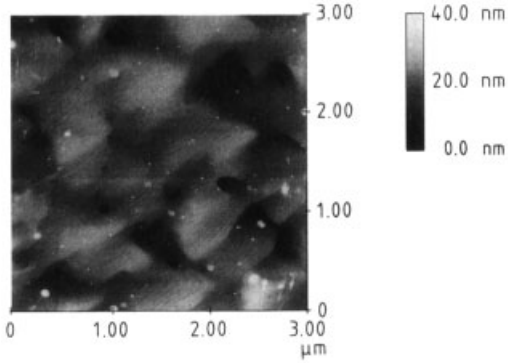

Figure 37. Spiral growth islands of *c*-axis-oriented $YBa_2Cu_3O_{7-\delta}$ thin film revealed by atomic force microscopy
Terraces with step height ca. 12×10^{-10} m.

Liquid-phase epitaxy is the most advanced technology for $LnBa_2Cu_3O_{7-\delta}$ films [166]. The complex phase diagram of $LnBa_2Cu_3O_x$ gives a relatively high epigrowth temperature, 900–970 °C, and as-grown films do not show a superconducting transition, owing to oxygen deficiency; post-growth annealing in oxygen is necessary. Growth is conducted at minimal supersaturation.

Deposition of $LnBa_2Cu_3O_{7-\delta}$ thin films, specifically $YBa_2Cu_3O_{7-\delta}$, has been thoroughly investigated for different deposition techniques. To obtain superconducting $YBa_2Cu_3O_{7-\delta}$ thin films in an as-deposited state, the relationship between oxygen partial pressure during deposition and substrate temperature has been established for different physical deposition techniques (Fig. 38). Oxygen partial pressure during deposition has to be kept close to the critical oxygen pressure at the decomposition of bulk $YBa_2Cu_3O_6$ (nonsuperconducting tetragonal phase). Activation processes to enhance oxidization involve ozone, oxygen plasma, and atomic oxygen radicals. Electron beam deposition involves post-deposition annealing in oxygen.

For epitaxial thin films of $LnBa_2Cu_3O_{7-\delta}$, the choice of substrate materials becomes more important, especially for active device applica-

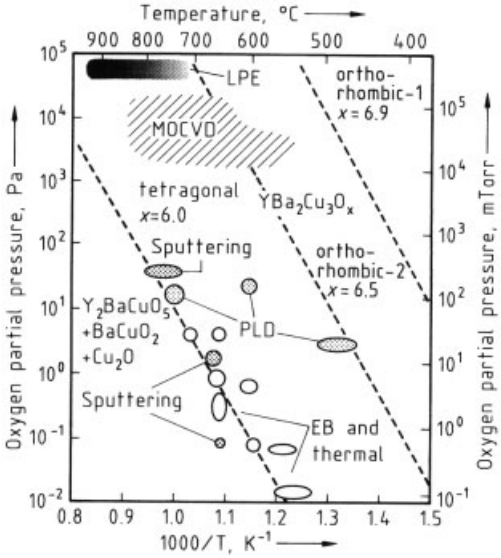

Figure 38. Equilibrium oxygen partial pressure and deposition temperature of $YBa_2Cu_3O_{7-\delta}$ thin films [167]
LPE = Liquid-phase epitaxy; MOCVD = Metal-organic chemical vapor deposition; PLD = Pulsed laser deposition; EB = Electron beam

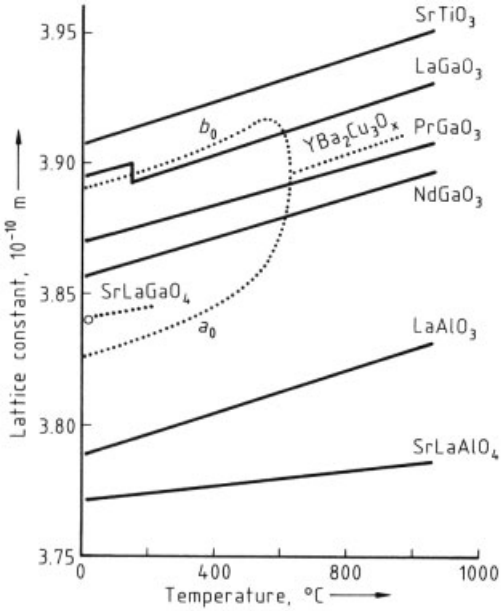

Figure 39. Lattice constants and thermal expansivities of new substrate materials
Data from $YBa_2Cu_3O_{7-\delta}$ for comparison [168].

tions. Figure 39 [168] shows thermal expansivities of different substrate materials. $LaAlO_3$ seems practical for microwave applications because of its relatively low dielectric constant [169], while $NdGaO_3$ is suitable for *c*-axis-oriented epitaxial thin films because of its good lattice matching [170].

Of current interest is the growth of *a*-axis-oriented $YBa_2Cu_3O_x$ thin films, because of the anisotropy in coherent length ($\xi_{a/b} = 30-40\times10^{-10}$ m and $\xi_c = 5-10\times10^{-10}$ m). This takes advantage of a Josephson coupling junction across an *a*-axis SIS (superconductor – insulator – superconductor) structure. Many workers have reported *a*-axis-oriented $YBa_2Cu_3O_{7-\delta}$ film growth in which the deposition temperature is lowered by ca. 100 °C from the usual *c*-axis film deposition temperature. However, this gives *a*-axis films with 90° domains, T_c ca. 20 K. Meanwhile, *c*-axis in-plane-aligned, *a*-axis-oriented $YBa_2Cu_3O_x$ thin films have been grown on (100)-oriented tetragonal $SrLaGaO_4$ and Nd_2CuO_4 substrates with a K_2NiF_4-type crystal structure. A growth model called "atomic graphoepitaxy" has been proposed [171], where the atomic geometry of a (100)-oriented face of tetragonal K_2NiF_4-type materials is considered. Films grow epitaxially with perfectly in-plane-aligned *c*- and *b*-axes, along both of which $T_c = 90$ K.

When $YBa_2Cu_3O_x$ films are deposited on semiconducting Si substrates, Si ions diffuse easily into the films, and T_c degrades to ca. 70 K or less. As a result, a variety of oxide materials are being investigated as a buffer layer between the film and Si substrate in order to suppress interdiffusion. One such buffer material is yttrium-stabilized ZrO_2 (YSZ).

2.6. Polymer Films

2.6.1. Langmuir – Blodgett Layers of Polymer Films as Materials

The Langmuir – Blodgett (LB) technique (see Section 1.6) is a powerful method for the preparation of ultrathin organic films of defined thickness and molecularly controlled internal structure. It consists of the following essential steps:

1) Spreading molecules on a water surface
2) Compressing them to create a two-dimensional condensed layer
3) Transfer of the layer onto a solid substrate to yield mono- or multilayers

The thickness of the films can be varied, in multiples of the monolayer thickness, between a few nanometers (one monolayer) to a few micrometers (several thousand layers). Polymeric LB layers [172] can be obtained by: in situ polymerization of multilayers of amphiphilic monomers (Fig. 40 A); spreading and transfer of preformed amphiphilic polymers (Fig. 40 B); spreading and transfer of hairy rod polymers (Fig. 40 C).

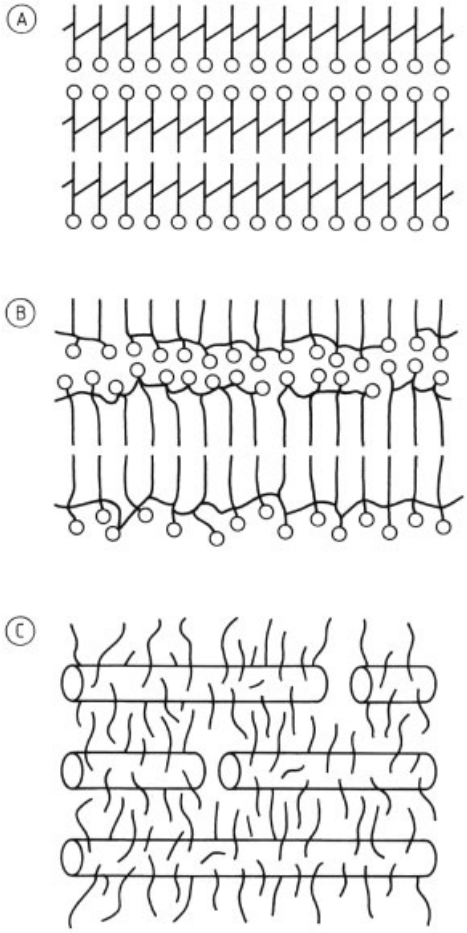

Figure 40. Architecture of polymeric LB layers
A) Layers prepared from solid-state polymerization of amphiphilic monomers (the polymerization can be carried out either within the monolayer prior to transfer, or within the multilayer assembly); B) Layers prepared from preformed polymers; C) Layers composed of rod-like main chains

Individual layers of amphiphilic monomers or polymers are generally composed of two-dimensional crystallites (Fig. 41). The intercrystallite grain boundaries contribute significantly to the properties of LB films. This can result in losses by scattering in such applications as waveguides or cladding layers, and can play a major role in defining transport properties, e.g., in diffusion barriers or electrical insulators.

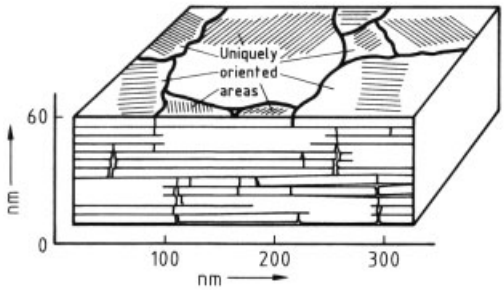

Figure 41. Domain structure in multilayer systems

In multilayers of rod-like macromolecules with flexible side chains (hairy rods), the polymer backbones are oriented parallel to one another and are maintained at a fixed separation by the side chains, which form a continuous matrix (Fig. 42). LB layers of hairy rod polymers exhibit good mechanical stability and high optical transparency, and may be considered as molecular composites. The stability of hairy rod polymer films can be increased by cross-linking functional groups attached to the side chains [173].

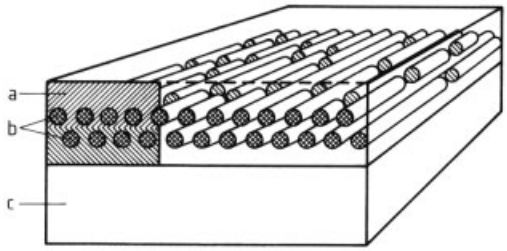

Figure 42. Arrangement of hairy rod molecules in a multilayer system
a) Hydrocarbon; b) Rigid rod polymer backbone; c) Hydrophobic substrate

Characterization Methods (Table 12) [174–176].

Film Thickness. If the thickness of an individual monolayer is known, counting the number of deposition cycles is the simplest method to as-

Table 12. Techniques for the investigation of LB films

	Radiation used	Contrast by	Typical information
Optical interference microscopy	light	optical phase shift	optical thickness, topography of surface
Ellipsometry	light	refractive index difference	optical thickness
Surface plasmon resonance spectroscopy	light	optical phase shift	optical thickness (metric thickness)
X-ray reflectometry	X-rays	electron density difference	electron density profile along layer normal, metric thickness, roughness
Neutron reflectometry	neutrons	selective deuteration	scattering length profile along layer normal, metric thickness, roughness
Contact angle		surface tension	wetting behavior
Adhesion test			adhesion forces
Infrared (IR)/Raman spectroscopies	light	IR/Raman absorption	functional group composition and orientation
Transmission electron microscopy (TEM)	electrons	electron density difference in plane	crystal structure, domain structure and texture
Low-energy electron diffraction (LEED)	electrons	electron density difference	crystal structure of surface layers
Atomic force microscopy (AFM)		height profile or surface tribology	surface topography
Scanning tunneling microscopy (STM)	charge carriers	conductivity differences	chemical composition periodicity of groups
X-ray photoelectron spectroscopy (XPS)	X-rays		elemental surface composition

sess the thickness of the film. The thickness can be determined by a number of experimental techniques. Optical interference techniques yield the "optical thickness," defined as the product of the metric thickness d and the refractive index n; knowledge of the refractive index enables evaluation of the metric thickness. The birefringent character of most LB films is a complicating factor.

Phase-shift interferometry [177] provides a three-dimensional profile of the film. Study of the profile edge enables the film thickness to be determined. This technique requires that a thin metal film be evaporated on top of the LB layers in order to obtain a reflecting surface. The thickness profile can also be determined by mechanical contact methods and, in the case of only a few layers, by atomic force microscopy (AFM).

In ellipsometry, the change in polarization of light reflected from a thin film on top of a metal surface is analyzed to yield accurate information on the optical thickness. Ellipsometry is also a sensitive technique for detecting minor changes during the lateral organization of LB films on the subphase, and has been used to investigate the properties of monomolecular films of fatty acids on both water and mercury subphases.

Surface plasmon spectroscopy [178] is a further method for optical thickness analysis. Surface plasmons are electromagnetic surface waves propagating from a thin metal film into a dielectric medium. Plasma oscillations in the metal film are excited by polarized light from a laser, which couples into the film at a particular angle. The resonance condition for a given system depends on the refractive index and the film thickness. When the LB film is thick enough to be used as a waveguide, the surface plasmon and waveguide modes can be detected simultaneously. Analysis of the waveguide modes gives exact values of the film refractive index and thickness.

The most detailed and accurate information on the thickness of multilayer systems comes from grazing incidence X-ray or neutron reflectometry. The absolute layer thickness can be derived from analysis of the characteristic interference patterns (Kiessig fringes). The thickness of the individual monolayers and the surface roughness of the film can also be obtained.

Surface Properties [174]. Contact angle measurements of various liquids placed on the LB layer can be used to determine surface energy and wetting properties. The adhesion of the first monolayer to the underlying substrate, and of subsequent monolayers to each other, can be determined only qualitatively. A standard adhesive tape is applied to the surface of the film and then detached. Comparison of the film before and after this procedure gives a measure of the amount of film removed.

Grazing incidence infrared (IR) reflection by mono- and multilayers on a metal substrate and polarized IR transmission are powerful tools for

Table 13. Properties and applications of LB films

	Properties	LB films	Applications
Mechanical properties	stability	low	
Optical properties	birefringence	often observed	optical filters
	nonlinear optics (χ^2/χ^3)	observed in films of special composition	optical switches, frequency doubling
Electrical properties	conductivity	up to 10^{-2} – 10 S/cm of special components	interconnection in microelectronic circuits
	resistance	good insulating behavior, low electric loss, e.g., in arachidic acid	capacitors
	photoconductivity	up to 1 % quantum efficiency in special cases	
	electroluminescence	blue light emitted from devices based on aromatics	displays
	pyroelectricity	observed in noncentrosymmetric structures	infrared TV cameras
	piezoelectricity	observed in noncentrosymmetric structures	
Surface properties	lubrication	drastic reduction of static friction by a few monolayers of fatty acids	magnetic storage
	wettability	polarity can be tailored	surface modification of substrates
	surface anisotropy	order parameter high, monodomain structure possible	liquid crystal alignment in displays
	biocompatibility	possible tailoring of surface	artificial implants, sensors
Miscellaneous	permeability	selective permeability often observed	membranes for gas separation, reversed osmosis, ultrafiltration
	chemical reactivity	composition dependent	microlithography, positive and negative resists

the investigation of the orientation of molecular moieties in the LB film. These techniques are sensitive enough to investigate individual monolayers [177].

Raman spectroscopy gives similar information about the structure, orientation, and conformation of the components of thin organic layers. Because of the weak signals, an enhancement mechanism is usually required for the direct observation of LB films. A dramatically increased signal can be produced by waveguide Raman spectroscopy (WRS) and surface plasmon enhanced Raman spectroscopy.

Transmission electron microscopy (TEM) as well as low-energy electron diffraction (LEED) have been used to investigate LB layers. Under optimum conditions single polymer molecules can be observed.

Various scanning near-field microscopies (AFM, STM, SEM, etc.) have been used to gain insight into the surface properties of mono- and multilayer systems.

In X-ray photoelectron spectroscopy (XPS) or electron spectroscopy for chemical analysis (ESCA), the surface is exposed to monochromatic X-rays and the energy distribution of electrons emitted gives information on the chemical composition, environment, and binding of atoms at the surface.

Elastic constants of LB assemblies can be determined by Brillouin spectroscopy [179] or by a quartz crystal resonator [180].

Properties of LB Multilayers and Applications (Table 13) [174, 175, 181]. Polymeric LB films are much more stable than layers composed of monomers, owing to covalent bonding in the film plane. However, the fragility of these systems remains a problem in many applications. The mechanical properties of mono- and multilayers are generally anisotropic, and may be strongly influenced by structural defects.

Optical and Optoelectronic Properties [182, 183]. Birefringence and dichroism are frequently observed in LB layers. Thus, optical filters and devices for optical applications can be prepared. Optical anisotropy can be induced through the introduction of mesogenic side groups. The optical properties of the film may be varied by thermal treatment or the application of an external electrical field. LB assemblies may be used as waveguides; the best attenuation of guided light reported has been ca. 2 dB/cm.

Nonlinear optical effects have received particular attention. To produce χ^2 effects (frequency doubling, electrooptical switching, etc.) noncentrosymmetric LB films of hemicyanines,

phenylhydrazones, or similar polarizable dyes are necessary. Values of the χ^2 parameter ca. 10 – 100 times the value of the commonly used material $LiNbO_3$ have been reported. For χ^3 effects (frequency tripling, phase conjugation, etc.) highly polarizable conjugated π-electron systems with donor – acceptor moieties can be used.

As discussed previously (Fig. 41), the presence of cracks and crystalline domains of sizes in the range of the optical wavelengths must be avoided, and the thermal stability must be enhanced, in order to produce LB films of suitable optic and stability.

Electric and Dielectric Properties [184]. Electrical measurements on LB layers frequently prove difficult, owing to the inherently fragile nature of the specimens. Various effects have been reported which depend on the chemical nature of the constituents and the state of order of the LB assemblies. For example, tunneling barriers have been studied where an LB film was sandwiched between two metal electrodes. LB layers composed of "organic metals" or semiconducting materials have also been reported. The behavior of LB films as conductors and insulators can be optimized by tailoring the molecular components of the film.

Ultrathin organic films between metal contacts have an enormous charge storage capability, which suggests an application as the dielectric layer in simple capacitors.

LB films on semiconductors can be used for various applications, including metal – insulator – semiconductor diodes, Schottky barriers, or field effect transistors. The properties of LB layers in terms of photoelectric transport and as the active layer in electroluminescent devices have been investigated.

Pyroelectricity and piezoelectricity are observed in noncentrosymmetric LB multilayer systems. Owing to the small heat capacity of the film, the pyroelectric response is very fast. Pyroelectric detector systems for thermal radiation based on LB films could be used in fast infrared TV cameras.

Surface Properties [181]. The presence of a single monolayer on any substrate can drastically reduce static friction. This effect can be used to reduce frictional forces between different surfaces in mechanical contact. For example, magnetic mass storage media require very thin, but very stable lubricant layers to avoid wear during transit of the tape across the magnetic head. Good results have been achieved with a few layers of fatty acids on the tape, which also serve as an encapsulant to protect the magnetic media.

The wettability of a specific surface can be profoundly influenced by coverage with only one or a few monolayers, e.g., turning a hydrophobic surface into a hydrophilic one. This offers the possiblity of tailoring the surface properties of different substrates. In the field of liquid crystal displays, orientation layers are commonly used to align liquid crystals parallel to the surface. LB films with uniaxial or biaxial orientation, prepared by carefully controlled deposition procedures, are contenders for liquid crystal alignment layers. Excellent results have been reported for ω-tricosenic acid films, polymerized by an electron beam. With azobenzol-dye-containing monolayers, a change of orientation of the liquid crystals from parallel to homeotropic can be induced by photoinduced *cis – trans* isomerization.

The LB technique allows control of surface charge densities, surface potential, degree of hydrophobicity, and type and concentration of functional groups on the surface. It has therefore been proposed to use this to produce coatings on medical devices and, in particular, on implants. To date, only model systems have been developed.

Permeability [174]. The architecture of LB multilayers of amphiphilic compounds in many ways mimics biological membranes, suggesting possible applications in specific separation tasks, such as ultrafiltration, reverse osmosis, and gas separation. In these situations LB films with high selectivity are deposited on a porous, mechanically stable support. For example, eighteen monolayers of hexacosa-10, 12-diyonic acid on a porous polypropylene support reduce the gas flow of methane by a factor of 30 compared with the uncoated substrate. Structural defects in LB films deposited directly onto the porous support often generate leakage problems. These can be resolved by a composite technique. A very thin, nonporous, but highly permeable layer is installed between the LB film and the porous membrane, which dramatically reduces the number of defects in the film.

Chemical Variation of LB Multilayers [185]. The polymerization of an LB layer composed of amphiphilic monomers can be initiated locally by UV or high-energy radiation. Resist structures may thus be prepared with a resolution better than 100 nm. This is of importance for integrated circuit technology, where submicrometer resolution lithography is a crucial element of process technology. Both positive and negative resists have been successfully used; simple fatty acid salts show high resolution as positive resists, subliming under irradiation. Enhanced sensitivities are more easily achieved with negative resists, in which case polymerization is initiated by an electron beam. Several types of polymerizable materials have been investigated. The topochemical polymerization of diacetylenic amphiphiles and of ω-tricosenoic acid show promising results. Transfer of these ultrafine structures to solid substrates, however, requires improvement of the resistance of the LB film to plasma etching. Polysiloxanes and polysilanes are good candidates for this task.

LB Films as Sensors (Table 14) [175]. The properties of LB films as described above suggest many possible applications to sensory devices. The extremely high ratio of surface area to bulk volume of LB films prevents bulk diffusion, and yields fast response and recovery times. The ordered surface allows a maximum of reactive species per unit surface area, and offers the potential of high selectivities via appropriate molecular design. In particular, exploitation of the intrinsic selectivity of biological systems (e.g., enzyme – substrate, antibody – antigen, and receptor – transmitter) can be incorporated into the design.

Table 14. LB films as sensors

	Sensitive absorbing LB layer on top of	Physical parameter changes by absorption of analyte
Field-effect transistor devices	semiconductor	dielectric properties
Optical sensors	optical waveguide	optical spectrum
Surface plasmon resonance devices	sandwiched between metal surface and prism	refractive index, film thickness
Ellipsometric sensors	metal surface	refractive index, film thickness
Acoustoelectric sensors	oscillating quartz	mass of film

A wide range of sensors based upon LB films have been investigated. Chemically sensitive field-effect transistor (ChemFET) sensors detect chemical changes in a medium which is in contact with a metal oxide – semiconductor FET. Changes in this medium cause an alternation in the surface potential within an LB layer placed on top of the FET device. Profound effects on the measurable parameters of a device are noted with even minute changes in the dielectric properties of the interfacial region.

Changes in surface conductivity induced by the interaction of specific gases with semiconducting LB films have formed the basis for the construction of gas sensors.

Optical sensors detect changes in the optical spectrum of an LB film resulting from the adsorption of an analyte. As an example, the fluorescence of porphyrin LB layers can be quantitatively quenched through the absorption of gases such as NO_2 and HCl. Other optical methods include plasmon resonance or ellipsometric detection of changes in the refractive index or surface thickness when a substrate interacts with the surface layer.

Acoustoelectric sensors are based on the principle that the resonant frequency of a vibrating body is altered when the surface coverage is changed. Thus, the resonance frequency of a piezoelectric quartz crystal coated with LB layers of ω-tricosenoic acid is changed on adsorption of hydrogen sulfide.

2.6.2. Films by Plasma Polymerization

Films formed by plasma polymerization differ from normal polymer material in several respects. A wide range of structure can be fabricated by modifying the plasma conditions (pressure, flow rate, and power). A map delineating different material regions may be constructed using these parameters. In ethylene, for example, it is possible to produce soft and hard films, powder and oil, as well as mixtures of these [186]. It is believed that powder is formed when the reaction occurs primarily in the gas phase, and is facilitated by long residence times, high power, and reactive monomers. Both IR and NMR have been used to investigate polymer film structure. Models of the structure of the oil and film products from plasma-polymerized

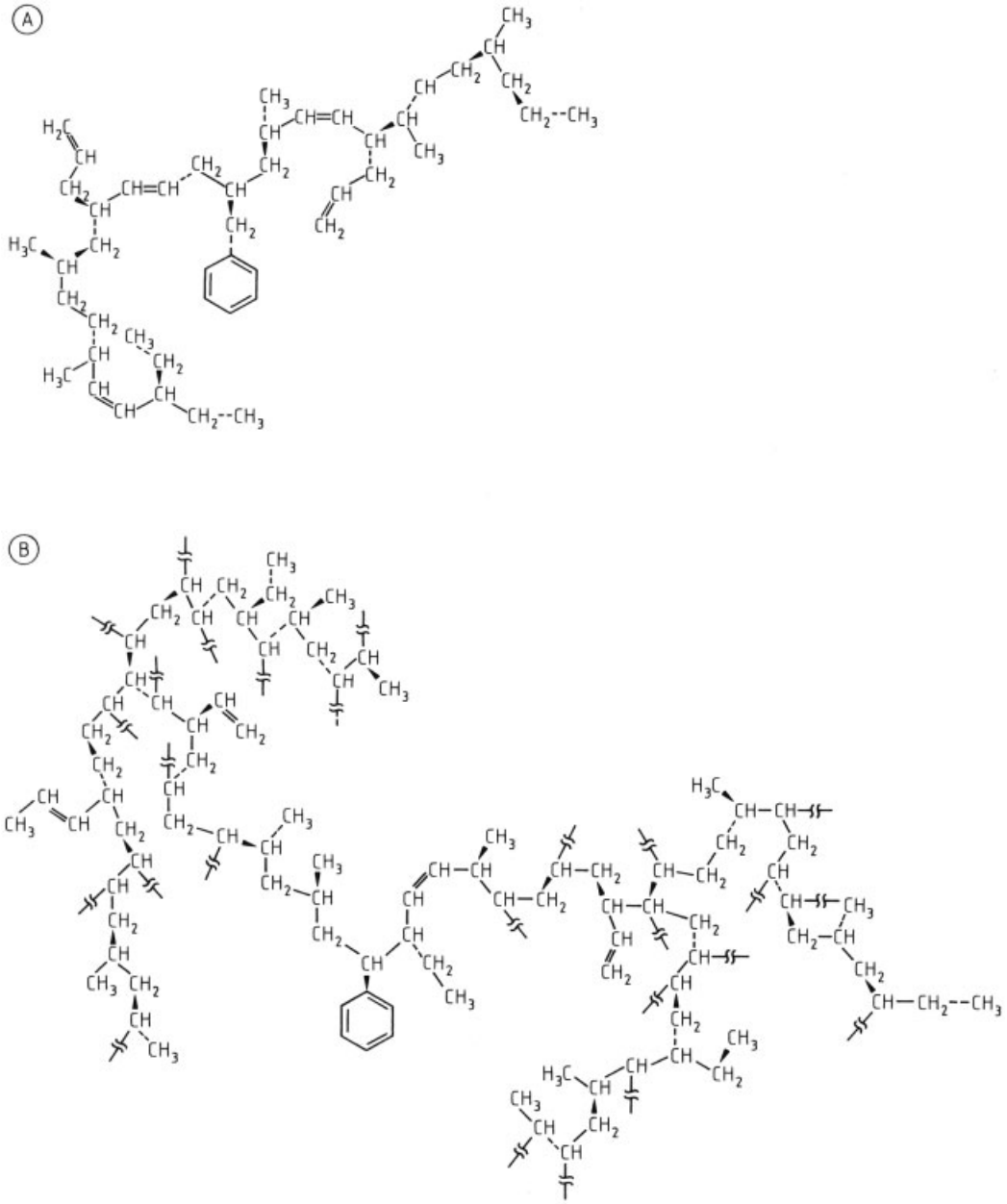

Figure 43. Plasma-polymerized material from ethylene monomer [187]
A) Oil model; B) Film model

ethylene are shown in Figure 43 [187]. A second feature of plasma polymerization is that almost any monomer can be used: organic, organometallic, or containing silicon. The monomer need not contain unsaturated groups. In plasma-polymerized material, very few monomer units remain in a regular repeating pattern as in conventional polymerization. The morphology of plasma films has been investigated by X-ray scattering und electron microscopy (e.g., powder particles in plasma-polymerized ethylene on a chromium substrate observed by transmission electron microscopy [186]).

Plasma films contain significant quantities of long-lived free radicals. In plasma-polymerized styrene, one in twenty polymer molecules is a radical site. The radicals are formed either directly in the energetic plasma, and then incorporated in the growing film, or by reactive plasma species or plasma radiation impinging on surfaces. Because the polymer molecules are strongly cross-linked (Fig. 43), the radicals are immobile and long lived. They are shorter lived if the film is exposed to the atmosphere, when they react with water vapor or O_2 (appearance of the characteristic IR bands of OH and CO in exposed film). If a high free radical density is undesirable, it can be reduced by heating or by subsequent exposure to a hydrogen plasma.

Another important characteristic of plasma-polymerized film is its impermeability. This is due in part to cross-linking. The mechanism of permeation in these films has been considered by Yasuda [188] who classified plasma films as intermediate between two ideal cases: solution – diffusion or molecular sieve mechanisms. Plasma films often show extremely low permeability to water, e.g., a methane polymerized film is two orders of magnitude less permeable than conventional polymeric material [188].

An important feature of polymerized plasma films is the presence of compressive intrinsic stress, the molecular origin of which is believed to be the high nonequilibrium density of the films due to their highly cross-linked structure. Intrinsic stress has been widely studied because of its implications for film stability. Methods of alleviating it include annealing at 100 – 300 °C. Intrinsic stress is not observed for all polymer films, and films formed from closely related monomers often exhibit quite different stress characteristics. Internal stress is generally larger under slower deposition conditions and increases with film thickness, eventually leading to cracking. Better barrier and adhesion properties are to be found in thinner films. The amount of stress also depends on the substrate and its thermal properties relative to the plasma film deposited on it [188].

The electrical properties of plasma films have received much attention. Initially it was believed that they had great potential for forming thin dielectric coatings, but there are some serious drawbacks. Their dielectric properties are susceptible to aging, primarily due to the reactive nature of the film when exposed to air. This behavior is often irreproducible from one film to another, even from the same batch. A second problem is high dielectric loss. Both features are believed to originate from the free radicals in the films, and attempts have been made to eliminate them by heat treatment in vacuo. The effects of O_2 and humidity on conduction properties have also been recognised. The effect of O_2 on plasma polymer from styrene monomer is an irreversible decrease in conductivity by three orders of magnitude. Photoconduction and bistability have also been investigated.

Potential applications of plasma films which seek to exploit their special properties have been reported [189]. The synthesis of osmotic membranes from plasma polymer deposited on porous substrate has produced materials which can efficiently separate salt solutions. Salt rejection as high as 98 % has been repeated with nitrogen-containing polymers at fluxes of $320\,\mathrm{L\,m^{-2}\,d^{-1}}$. Permselective membranes for separating gas mixtures have been applied to H_2/CH_4, H_2/CO_2, and He/CO_2 mixtures.

The adhesion properties of the films on a number of substrates and their pinhole-free, cross-linked structure have led to their use as barrier coatings. They have been used to protect metals and moisture-sensitive optical components such as alkali halide crystals. For the latter, fluorinated ethylene films have been used which also give the coated components antireflection properties, which can be varied by changing the monomer and its thickness. Abrasion-resistant coatings can be made from hard films and several silicon-containing monomers have been used.

2.7. Optical Coatings

2.7.1. Introduction

The characteristic colors of most substances have their origin in selective molecular absorption (chemical colors). Other colors are a result of physical structure. These physical colors have the advantage over chemical colors that they do not fade.

Some physical colors found in nature are iridescent, interference colors, e.g., peacock feathers. “Interference colors are the purest and most brilliant colors known to man. They cannot be

matched by even the brightest pigment colors in depth and intensity" [190].

Optical coatings are artificial thin films. Although they can be used to color objects, they are normally deposited on optical elements to modify spectral reflection and transmission. Their uses range from the UV to the far IR. The number of individual films can be up to several hundred. Since each additional thin film adds another interface (the origin of an additional partial reflection, contributing to the interference) very high reflections, very low reflections, and steep transitions can be accomplished.

2.7.2. Theory

The primary elements of thin film interference are the reflections and the spacings of the interfaces. For nonabsorbing films, this can be related to the indices of refraction n and the optical thicknesses nd of the individual films. The reflectance R of a nonabsorbing optical coating consisting of m individual layers is given by

$$R = 1 - \frac{4n_0 n_s}{2 + (n_0/n_s)\, M_{11}^2 + n_0 n_s M_{12}^2 + M_{21}^2/n_0 n_s + (n_s/n_0)\, M_{22}^2}$$

where n_0 is the refractive index of the surrounding medium (air), n_s that of the substrate, and M_{11}, M_{12}, M_{21}, M_{22} are the elements of the characteristic matrix of the multilayer:

$$M = \begin{bmatrix} M_{11} & iM_{12} \\ iM_{21} & M_{22} \end{bmatrix} = \prod_{\nu=1}^{m} \begin{bmatrix} \cos\Phi_\nu & \frac{i\sin\Phi_\nu}{n_\nu} \\ in_\nu \sin\Phi_\nu & \cos\Phi_\nu \end{bmatrix}$$

where n_v is the refractive index, $\Phi_\nu = \frac{2\pi}{\lambda} n_\nu d_\nu \cos\alpha_\nu$ the phase thickness, λ the wavelength, d_ν the physical thickness of the νth layer, and α_ν the angle of incidence at the $(\nu+1)$th interface.

The "design" process of fitting the spectral reflectance of a multilayer to a prescribed spectral curve generally follows one of two methods:

1) The designer is in charge and the computer is used merely as a calculator (split-filter synthesis [191], equivalent layers [192], and Chebyshev prototypes [193]).
2) The computer is in charge and the decisions of the designer are reduced to a minimum (refinement [194] and computer synthesis: comprehensive search [195], flip-flop [196], and inverse Fourier transforms [197]). Several excellent programs are commercially available [198].

2.7.3. Deposition Methods

Four deposition methods are used commercially:

1) High-vacuum evaporation from boats or with electron guns [199]
2) Dipping [200]
3) Energetic deposition processes [201] (d.c. magnetron sputtering, ion plating, and ion-assisted evaporation)
4) Chemical vapor deposition (low-pressure CVD [202], plasma-enhanced CVD [203], and plasma-impulse CVD [204])

Figure 44 estimates the share and trend of each of these methods in world manufacture of optical coatings (total market \$ 1.1×10^9 in 1992 [205]).

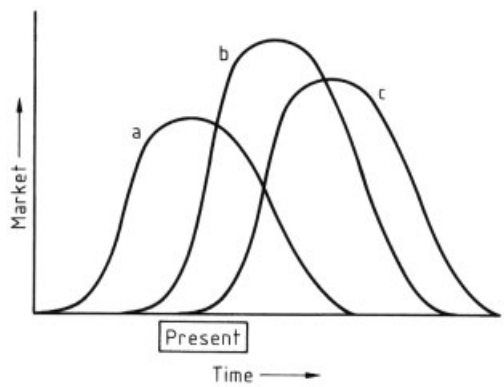

Figure 44. Past and projected market share of deposition methods for optical coatings
a) Evaporation/dipping; b) Sputtering/ion plating/ion-assisted evaporation; c) Plasma CVD, PICVD/PECVD

2.7.4. Coating Materials

Table 15 lists typical coating materials. Mixed oxides are also feasible, but the proportions must allow congruent evaporation, with energetic deposition processes, silicon and aluminum nitrides are also possible.

Table 15. Refractive indices and ranges of transparency for typical coating materials [206]

Material	Wavelength, µm					Range of transparency, µm
	0.35	0.55	1	1 – 8	10	
Ge				4.20	4.20	1.7 – 25
Si			3.90	3.42		1 – 9
TiO_2		2.32	2.20			0.4 – 6
ZnS		2.36	2.27			0.4 – 14
ZrO_2	2.15	2.05	2.00			0.3
Ta_2O_5	2.31	2.16	2.09			0.3
SiO		2.00	1.90	1.85		0.5 – 8
Al_2O_3	1.66	1.63	1.60			0.15 – 9
Si_2O_3		1.57	1.55			0.3 – 8
SiO_2	1.48	1.46	1.45			0.16 – 8
MgF_2	1.39	1.38		1.36		0.13 – 10
Cryolite			1.30 – 1.35			0.13 – 9

The quality of optical coatings depends on the preparation of the coating materials. Most materials come in 1 – 2 mm granules and are vacuum sintered. Often, small additions of foreign material reduce crystallization and increase homogeneity.

2.7.5. Classical Coatings

The characteristics of Figure 45 can be generated from the UV (0.2 µm) to the IR (50 µm) range of the spectrum. Since different coating materials are necessary in different spectral regions (Table 15), the quality of match to the ideal curve depends on wavelength.

In the visible (0.4 – 0.7 µm) the following performance figures are possible:

1) Reflectances as low as 0.01 % for narrow regions and < 0.3 % over the whole visible spectrum
2) Reflectances as high as 99.99 % for narrow regions and 99.8 % over the whole visible spectrum
3) Transition from low (< 5 %) to high reflectance (> 95 %) within an interval $< 0.01\ \lambda$
4) Narrow bandpass filters with bandwith as small as 10^{-10} m
5) Minus filters with bandwidth ca. $0.01\ \lambda$

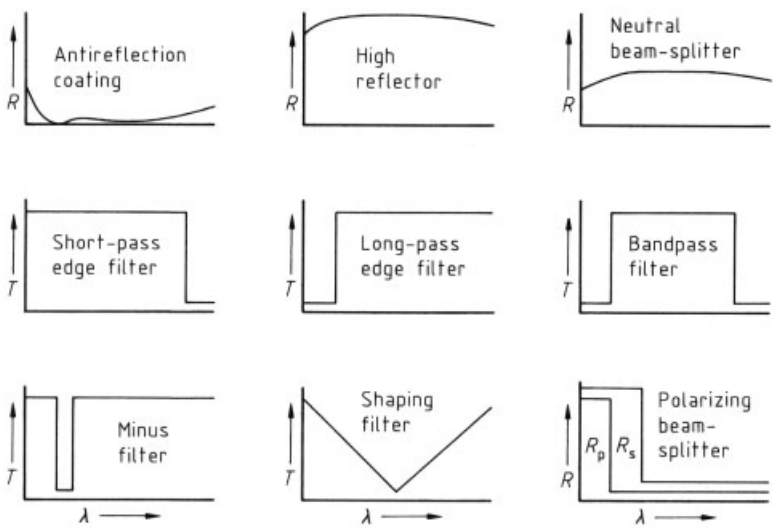

Figure 45. Characteristics of classical optical interference coatings

2.7.6. Nonclassical Coatings

Of increasing importance are optical coatings which, in addition to being part of the interference system, serve an electrical or magnetic function:

1) Transparent conductive coatings: indium–tin oxide conducts electricity well, but is optically transparent [207]. Surface resistance of a few Ω/m^2 at 90 % transmittance can be accomplished.
2) Electrochromic coatings: colorless amorphous WO_3, on injection of electrons and charge-compensating positive ions, transforms to blue hydrogen tungsten bronze [208].
3) Electroluminescent coatings: Mn-doped ZnS films, deposited at ca. 225 °C, emitlight in an electrical field [209].
4) Magneto-optical films: sputtered alloys of Tb, Fe, and Co show a large Kerr rotation, and can be used for high-efficiency magneto-optical recording [210].
5) Diamond-like films: under certain conditions, carbon films can assume a diamond or diamond-like structure [211].

2.7.7. Applications

Optical coatings have the following four effects:

1) Improved performance of optical components
2) Adaptability to most needs
3) Protection (often harder than the substrate)
4) Solution of otherwise intractable problems

They play key roles in the following areas:

1) Displays: cathode ray tube contrast enhancement, r.f. interference shielding, transparent electrodes for liquid crystal displays, liquid crystal light valves for high-definition television, head-up displays, shop window antireflection coatings
2) Photography/video: antireflection coatings for lenses, mirrors, filters, color film processors, enlarging heads, professional and amateur television cameras
3) Eyeglasses: antireflection coatings, sunglasses, photochromic glasses, high-efficiency sunglasses, frames, laser goggles, welding goggles, electrochromic glasses
4) Lighting: regenerative lamps; low-height, wide-angle street lamps; shopping display lights; traffic/railroad lights; movie projectors; disco lights; stage lighting; dental/surgical lighting
5) Instruments/lasers: antireflection coatings, beam dividers, mirrors, filters, variable filters, Raman spectroscopy filters, fluorescence microscopy filters, laser end mirrors, laser windows, high-power polarizers, laser gyro coatings
6) Photocopiers/data storage: high reflectors for photocopiers, compact disk players, optical data disks, magneto-optical data disks, optical read/write heads
7) Security: burglar alarms; anticounterfeiting for documents, licenses, and currency
8) Buildings: summer films, winter films, antireflection coatings, "smart" windows
9) Automobiles: rear-view mirrors, scratch-resistant coatings for plastic windows, heat-reflecting windows, plastic headlights
10) Optical communications: antireflection coatings on laser ends, narrow-band filters, wavelength-division multiplexers
11) Integrated optics: waveguide coatings, antireflection coatings, integrated filters
12) Home appliances: oven windows, oven wall heater reflectors, optical pot heaters
13) Applied art: jewelry, iridescent color effects, optical art
14) Flexible substrates: architectural window films, cathode ray tube contrast enhancement films, touchscreens, inexpensive solar reflectors, iridescent packaging, interference ink, lightweight mirrors, optically enhanced packaging

2.8. Membranes [212–214] (→ Membranes and Membrane Separation Processes)

2.8.1. Introduction

Membranes consist of thin solid or liquid films that act as barriers for the transport of molecular components. Synthetic membranes are widely used in applications such as water desalination

and gas separation or the concentration and purification of food and chemical products, and notably in medical devices, such as artificial kidneys or drug delivery systems. In energy conversion and storage systems, e.g., batteries, fuel cells, and electrolyzers, membranes also play an important role.

2.8.2. Fundamentals

Membranes used in various applications differ in their functions, structures, and materials. In general terms, a membrane is defined as a solid or liquid, homogeneous or heterogeneous barrier that separates two phases and restricts the transport of different components from one phase to the other in a specific way. To separate molecular mixtures, membranes must have different permeabilities for different components. For high mass transport rates per unit area, membranes should be as thin as possible; they are usually thin films or hollow fibers. To obtain a transport through a membrane, driving forces are required, i.e., differences in the electrochemical potentials of the various components such as differences in concentration, hydrostatic pressure, or electrical potential between the two phases separated by the membrane. Transport of a component is determined by its concentration and mobility in the membrane. Mobility is determined by the hydrodynamic resistance in pressure-driven processes, and by diffusivity of the component in the membrane matrix in concentration gradient- or electrical potential-driven processes. Concentration in the membrane is determined by solubility in homogeneous membranes, by exclusion due to size in porous membranes, or due to electrical charge in ionexchange membranes. Concentration and mobility of certain components in a membrane may be facilitated by functional groups incorporated in the membrane matrix (facilitated transport).

2.8.3. Membrane Structures and their Preparation

Classification. Although synthetic membranes vary widely in structure, function, and material, they can be classified into typical groups (Table 16). This classification is rather arbitrary and many structures used today may fit none or more than one of these classes.

Table 16. Synthetic membranes

Structure	Material	Function
Porous films, hollow fibers	polymers, ceramics, metals	gas separation micro- and ultrafiltration, dialysis
Homogeneous films, hollow fibers	polymers, metals	reverse osmosis, gas separation, pervaporation
Homogeneous or porous films with fixed ions	polymers, ceramics	electrolysis, electrodialysis, dialysis
Asymmetric porous or homogeneous films	polymers, ceramics	pressure-driven membrane processes
Thin film composite membranes	polymers, ceramics	pressure-driven membrane processes
Liquid films with selective carriers	organic liquids	selective removal of ions, gases, etc.

Preparation. The techniques used for manufacturing the various membrane structures are very different. For the preparation of microporous membranes from ceramics or polymers, pressing and sintering of a fine powder is used. Irradiation of a thin film followed by acid or alkaline etching or the precipitation of polymer from a solution by a non-solvent are also used to prepare porous membranes. Homogeneous membranes are prepared by extrusion of a metal or a polymer into a thin film or by casting polymer solution as a thin film and evaporating the solvent. Asymmetric structures are generally made by phase inversion techniques and composite membranes by dip-coating or interfacial polymerization. Techniques for making liquid membranes include the formation of emulsions for the preparation of unsupported liquid membranes and the filling of porous structures with liquid for the preparation of supported liquid membranes.

Processes. Membrane separation processes can differ significantly with regard to the membranes, the applied driving forces, the areas of application, and the industrial relevance (Table 17).

Applications. Synthetic membranes are used not only in separation processes, but also in medical devices as barriers for the controlled

Table 17. Technical relevant membrane separation processes

Separation process	Membrane type	Driving force	Application
Microfiltration	symmetric porous membranes	hydrostatic pressure 10 – 500 kPa	sterile filtration, clarification
Ultrafiltration	asymmetric porous membranes	hydrostatic pressure 0.1 – 1 MPa	filtration of macromolecular solutions
Reverse osmosis	asymmetric homogeneous membranes	hydrostatic pressure 2 – 10 MPa	desalination of saline solutions
Dialysis	symmetric porous membranes	concentration gradient	separation of microsolutes from macromolecular mixtures
Electrodialysis	ion-exchange membranes	electrical potential gradient	desalination of ion-containing solutions
Electrolysis	ion-exchange membranes	electrical potential gradient	chlor-alkaline production
Gas and vapor separation	asymmetric homogeneous membranes	vapor pressure gradient	separation of gases and vapors
Pervaporation	asymmetric homogeneous membranes	vapor pressure gradient	separation of azeotropic mixtures

Table 18. Technically relevant applications of synthetic membranes

Separation process	Controlled-release systems	Membrane reactors, immunoisolation	Energy conversion and storage
Microfiltration	therapeutic drug delivery devices	enzyme and catalytic reactors	battery separators fuel cell and electrolyzer separators
Ultrafiltration	artificial organs	biosensors	
Reverse osmosis	sustained-release of pesticides	immunoprotection of cell implants	solid electrolytes
Dialysis			
Electrodialysis			
Gas Separation			
Pervaporation			

release of drugs and in energy-conversion systems as ion-permeable separators (Table 18). For large-scale applications in the chemical and food industry or in environmental protection, separation processes, such as micro- and ultrafiltration or gas separation are most important, but commercially the medical uses of membranes, e.g., in artificial kidneys, are equally interesting.

3. References

1. K. Jensen, G. W. Cullen,: XII Int. CVD Conference, Int. CVD Conferences, Electrochemical Soc., Pennington 1993.
2. T. Kodas, M. Hampden-Smith: *The Chemistry of Metal CVD,* VCH Verlagsgesellschaft, Weinheim 1997.
3. A. C. Jones, P. O'Brien: *CVD of Compound Semiconductors,* VCH Verlagsgesellschaft, Weinheim 1997.
4. W. S. Rees: *The Materials Science of Thin Films,* Academic Press, San Diego 1991.
5. M. Ohring: *CVD of Nonmetals,* VCH Verlagsgesellschaft, Weinheim 1997.
6. M. L. Hitchman, K. Jensen (eds): *Chemical Vapor Deposition, Principles and Applications,* Academic Press, London 1993.
7. C. E. Morosanu: *Thin Films by CVD,* Elsevier, Amsterdam 1990.
8. C. F. Powell, J. H. Oxley, J. M. Blocher (eds): *Vapor Deposition,* Wiley, New York 1966.
9. H. Schäfer: *Chemische Transaktionen,* Verlag Chemie, Weinheim 1962.
10. D. M. Schleich: "CVD: A Chemical Approach to Ceramic Materials," *J. Phys. Colloq.* **C5** (1989) Suppl. to no. 5, vol. 50, 961.
11. T. F. Kuech, *Mater Sci. Rep.* **2** (1987) 1.
12. Y. Wang, M. Sasaki, T. Goto, T. Hirai, *J. Mater Sci.* **25** (1990) 4607.
13. E. A. Fitzgerald, *Mater Sci. Rep.* **7** (1991) 91.
14. G. Wahl, P. Batzies in G. F. Wakefield, J. M. Blocher (eds): *Proc. IV CVD Conference,* Princeton, N.J. 1973, p. 425.
15. G. Wahl, P. Batzies in G. F. Wakefield, J. M. Blocher (eds): *Proc. IV Int. CVD Conference,* Electrochemical Society, Princeton, N.J. 1973, p. 363.
16. W. A. P. Claassen, J. M. Bloem, W. G. J. N. Valkenberg, C. H. J. v. d. Brekel, *J. Cryst. Growth* **57** (1982) 259.
17. J. M. Bloem, L. J. Giling in E. Kaldis (ed): *Current Topics in Mat. Sci,* **vol. 1,** North Holland, 1978, p. 147.

18. F. Schmaderer, R. Huber, H. Oetzmann, G. Wahl, *Appl. Surf. Sci.* **46** (1990) 53.
19. J. O. Ofori, S. V. Sotichos, *J. Electrochem. Soc.* **143** (1998) 1962.
20. St. K. Griffiths, R. H. Nilson, *J. Electrochem. Soc.* **145** (1998) 1263.
21. D. Bäuerle: *Laser Processing and Chemistry,* Springer Verlag, Heidelberg 1996.
22. O. Lehmann, M. Stuke, *Materials Letter* **21** (1994) 131 – 136. O. Lehmann, M. Stuke, *Appl. Phys.* **A53** (1991) 343 – 345.
23. B. Eliasson, U. Kogelschatz, *Appl Phys. B* (1996) 633.
24. J. Y. Chang, L. J. Bie, I. W. Boyd, *Jpn. J. Appl Phys.* **37** (1998) L27.
25. B. Chapman; *Glow Discharge Processes,* Wiley, New York 1980.
26. B. Rother, J. Vetter: *Plasmabeschichtungsverfahren,* Verlag Grundstoffindustrie, Leipzig 1992.
27. J. R. Hollahan, T. Wydeven, C. C. Johnson, *Appl. Opt.* **13** (1974) 1844 – 1849.
28. R. J. Jensen, A. T. Bell, D. S. Soong, *Plasma Chem. Plasma Process.* **3** (1983) 163 – 192.
29. H. Yasuda: *Plasma Polymerization,* Academic Press, Orlando 1985.
30. R. F. Bunshah (ed.): *Deposition Technologies for Films and Coatings,* 2nd ed., Noyes Publications, 1994.
31. J. L. Vossen, W. Kern (eds.): *Thin Film Processes,* Academic Press, London 1991.
32. E. H. C. Parker (ed.): *The Technology and Physics of Molecular Beam Epitaxy,* Plenum Publ., New York 1985.
33. E. Kasper, J. C. Bean (eds.): *Silicon Molecular Beam Epitaxy,* CRC Press, Boca Raton, FL 1988.
34. M. A. Herman, H. Sitter: "Molecular Beam Epitaxy, Fundamental and Current Status," *Springer Ser. Mater. Sci.* **7** (1989).
35. B. A. Joyce, D. D. Vvedensky, C. T. Foxon: "Growth Mechanisms in MBE and CBE of III – V Compounds," in S. Mahajan (ed.): *Handbook on Semiconductors,* **vol. 3 a,** North Holland, 1994, p. 275.
36. E. P. Plueddemann: *Silane Coupling Agents,* 2nd ed., Plenum Publ., New York 1991.
37. L. H. Dubois, R. G. Nuzzo, *Annu. Rev. Phys. Chem.* **43** (1992) 437 – 463.
38. A. Ulman: *An Introduction to Ultrathin Organic Films,* Academic Press, San Diego 1991.
39. M. L. Shilling et al., *Langmuir* **9** (1993) 2156 – 2160.
40. I. Langmuir, *J. Am. Chem. Soc.* **39** (1917) 1848.
41. K. B. Blodgett, *J. Am. Chem. Soc.* **57** (1935) 1007.
42. G. Roberts: *Langmuir-Blodgett Films,* Plenum Press, New York 1990.
43. A. H. Ellison, *J. Phys. Chem.* **66** (1962) 1867.
44. B. Tieke, G. Lieser, G. Wegner, *J. Polym. Sci. Polym. Chem. Ed.* **17** (1979) 1631.
45. A. M. Bibo, I. R. Peterson, *Adv. Mater.* **2** (1988) 309.
46. H. Ringsdorf, B. Schlarb, J. Venzmer, *Angew. Chem.* **100** (1988) 117.
47. G. L. Gaines: *Insoluble Monolayers at Liquid-Gas Interfaces,* Interscience, New York 1966.
48. G. Wegner, *Thin Solid Films* **216** (1992) 105.
49. J. Boyle III, A. J. Mantone, *Colloids Surf.* **4** (1982) 77 – 85.
50. D. B. Ziversmit, *J. Colloid Sci.* **18** (1963) 794.
51. C. Kurthen, W. Nitsch, *Adv. Mater.* **3** (1991) 445.
52. A. Barraud, M. Vandevyver, *Thin Solid Films* **99** (1983) 221.
53. I. Langmuir, J. V. Schäfer, *J. Am. Chem. Soc.* **57** (1938) 1007.
54. F. A. Lowenheim (ed.): *Modern Electroplating,* 3rd ed., J. Wiley & Sons, New York 1974.
55. W. H. Safranek (ed.): *The Properties of Electrodeposited Metals and Alloys: A Handbook,* 2nd ed., American Electroplaters and Surface Finishing Society, Orlando, FL, 1986.
56. J. J. Durney (ed.): *Electroplating Engineering Handbook,* 4th ed., Van Nostrand Reinhold, New York 1984.
57. *The Canning Handbook: Surface Finishing Technology,* 23rd ed., W. Canning, Birmingham 1982.
58. D. Pletcher, F. C. Walsh: *Industrial Electrochemistry,* 2nd ed., Chapman & Hall, London 1991, chap. 8, p. 358.
59. D. R. Gabe: *Principles of Metal Surface Treatment and Protection,* 3rd ed., Merlin Books, London 1993.
60. H. Silman, G. Isserlis, A. V. Averill: *Protective and Decorative Coatings for Metals,* Finishing Publications, Teddington, UK, 1978.
61. R. Suchentrunk: *Metallising of Plastics – A Handbook of Theory and Practice,* Stevenage, Hertfordshire, UK, 1993.
62. R. Weiner: *Electroplating of Plastics,* Finishing Publications, Teddington, UK, 1977.
63. G. Lorin: *Phosphating of Metals,* Finishing Publications, Teddington, UK, 1974.

64. W. Rausch: *The Phosphating of Metals,* Finishing Publications, Teddington, UK, 1990.
65. W. Riedel: *Electroless Nickel Plating,* Finishing Publications, Teddington, UK, 1991.
66. R. Weiner, A. Walmsley: *Chromium Plating,* Finishing Publications, Teddington, UK, 1980.
67. H. Geduld: *Zinc Plating,* Finishing Publications, Stevenage, UK, 1988.
68. S. Wernick, R. Pinner, P. Sheasby: *The Surface Treatment and Finishing of Aluminium and its Alloys,* 5th ed., ASM International, Metals Park, OH, 1987.
69. A. T. Kuhn (ed.): *Techniques in Electrochemistry, Corrosion and Metal Finishing – A Handbook,* J. Wiley & Sons, Chichester, UK, 1987.
70. R. L. Yeates: *Electropainting,* 2nd ed., Robert Draper, Teddington, UK, 1970.
71. C. F. Simpson, M. Whittaker (eds.): *Electrophoretic Techniques,* Academic Press, London 1983.
72. W. Machu: *Handbook of Electropainting Technology,* Electrochemical Publications, Hatch End, UK, 1978.
73. A. Brenner: *Electrodeposition of Alloys: Principles and Practice,* Academic Press, New York 1963.
74. J. K. Dennis, T. E. Such: *Nickel and Chromium Plating,* Butterworths, London 1993.
75. J. W. Price: *Tin and Tin-Alloy Plating,* Electrochemical Publications, Ayr, Scotland, 1983.
76. E. Raub, K. Muller: *Fundamentals of Metal Deposition,* Elsevier, Amsterdam 1967.
77. D. R. Gabe (ed.): *Coatings for Corrosion Protection: A Guide for Production Engineers,* Institute of Production Engineers, London 1983.
78. V. E. Carter (ed.): *Corrosion Testing for Metal Finishing,* Butterworths, London 1982.
79. K. N. Stafford, P. K. Datta, C. G. Coogan (eds.): *Coatings and Surface Treatment for Corrosion and Wear Resistance,* Ellis Horwood, Chichester, UK, 1984.
80. P. Lall, S. Bhagath: "An Overview of Multichip Moduls," *Solid State Technol.* **36** (1993) Sept, 65–67, 70, 72, 74, 76.
81. P. L. Kirby: "Applications of Resistive Thin Films in Electronics," *Thin Solid Films* **50** (1978) 211 – 221.
82. W. Worobey: "Thin Film Technology Used in Bell System Telecommunication Circuits," *Electrocomponent Sci. Technol.* **8** (1981) 3 – 8.
83. F. Krull, H.-E. Endres: "Spürnasen," *Tech. Rundsch.* **75** (1993) no. 18, 28 – 34.
84. F.-W. Drees et al.: "Model of The Thermal Transfer Process," *Proc. 4th Int. Congress on Advances in Non-Impact-Printing Tech,* SPSE, New Orleans, 1988, pp. 442 – 447.
85. E. Unger: "Die Erzeugung dünner Schichten," *Chem. unserer Zeit* **25** (1991) no. 3, 149.
86. J. Bond: "The Incredible Shrinking Disk Drive," *Solid State Technol.* **36** (1993) Sept., 39 – 45.
87. R. J. Gambino: "Optical Storage Disk Technology," *MRS Bull.,*April (1990) 20 – 23.
88. K. Hieber, H. Körner, H. Treichel: "Chemical Vapour Deposition of Oxide and Metal Films for VLSI Applications," *Thin Solid Films* **181** (1989) 77.
89. H. Treichel, Th. Kruck: "Molecular Engineering in Semiconductor Technology: Borphosphorous-Silicatglass by Decomposition of a Monomolecular Precursor," *J. Phys. Colloq. C 5* **50** (1989) no. 6, 747 – 756.
90. D. B. Fraser in S. M. Sze (ed.): *VLSI Technology,* Mc Graw Hill, New York 1984, pp. 347 – 383.
91. H. Joswig, W. Pamler: "Stoichiometry Effects in TiN Diffusion Barriers," *Thin Solid Films* **221** (1992) 228 – 232.
92. A. Intemann, H. Körner, F. Koch: "Film Properties of CVD TiN Deposited with Organometallic Precursors," *J. Electrochem. Soc.* **140** (1993) 3215 –3222.
93. H. Körner et al.: "A Low Cost, High Throughput Tungsten Plug Contact Metallization with Improved Al Electromigration," *Proc. of the 10th Int. VLSI Multilevel Interconnection Conf. (VMIC) 1993,* VIMIC Catalogue No: 93 IS MIC-102.
94. O. Spindler, B. Neureither: "In Situ Planarization of Intermetal Dielectrics: Process Steps, Degree of Planarization and Film Properties," *Thin Solid Films* **175** (1989) 67 – 72.
95. H. C. Weller, R. H. Mauch, G. H. Bauer: "Studies on the Interfaces Between Novel Type ZnO and p-a-SiC:H in 1.5 eV a-SiGe:H Pin Diodes in Comparison to SnOx and ITO," *Solar Energy Mater.* **27** (1992) 217 – 231.
96. M. Kawachi: "Silica Waveguides on Silicon and their Applications to Integrated-Optic Components," *Opt. Quantum Electron.* **22** (1990) 391 – 416.
97. R. T. Chen: "Polymer-based Passive and Active Guided Wave Devices and Their Applications," *SPIE, Critical Review on*

Integrated Optics and Optoelectronic, **vol. CR** (1993) 298 – 335.
98. H. Heinecke, E. Veuhoff: "Evaluation of III – V Growth Technologies for Optoelectronic Applications," *Mater. Sci. Eng.* **B 21** (1993) 120 – 129.
99. L. Esaki: "A Bird's-Eye View on the Evolution of Semiconductor Superlattices and Quantum Wells," *IEEE J. Quantum Electron.* **QE-22** (1986) no. 9, 1611 – 1624.
100. S. L. Schwartz, V. E. Wood: "Ferroelectric Thin Films," *Condens. Matter News* **1** (1992) no. 5, 4 – 13.
101. R. Bruchhaus: "Thin Films for Ferroelectric Devices," *Ferroelectrics* **133** (1992) 73 – 78.
102. L. E. Toth: *Transition Metal Carbides and Nitrides,* Academic Press, New York 1967.
103. J.-E. Sundgren, H. T. G. Hentzell, *J. Vac. Sci. Technol. A* **4** (1986) 2259 – 2279.
104. J. Musil, J. Vyskocil, S. Kadlec in M. H. Francombe, J. L. Vossen (eds.): *Physics of Thin Films Mechanic and Dielectric Properties,* vol. 17, Academic Press, San Diego 1993.
105. L. Hultman et al., *Thin Solid Films* **205** (1993) 153 –164.
106. J. E. Greene, S. A. Barnett, J.-E. Sundgren, A. Rockett in A. Itoh (ed.): *Ion Assisted Deposition,* Elsevier, Amsterdam 1988, chap. 5.
107. S. J. Bull, A. M. Jones, A. R. McCabe, *Surf. Coat. Technol.* **54/55** (1992) 173 – 179.
108. K. A. Pischow et al., *Surf. Coat. Technol.* **58** (1993) 163 – 172.
109. U. Helmersson et al., *J. Vac. Sci. Technol. A* **3** (1985) 308 – 315.
110. H. E. Rebenne, D. G. Bhat, *Surf. Coat. Technol.* **63** (1994) 1 – 13.
111. M. K. Hibbs, J.-E. Sundgren, B. E. Jacobsson, B.-O. Johansson, *Acta Metall.* **33** (1985) 797 – 803.
112. W.-D. Münz, *J. Vac. Sci. Technol. A* **4** (1986) 2717.
113. U. Wahlström et al., *Thin Solid Films* **235** (1993) 62 – 70.
114. D. McIntyre et al., *J. Appl. Phys.* **67** (1990) 1542 –1553.
115. J.-E. Sundgren et al., *Thin Solid Films* **193/194** (1990) 818 – 831.
116. S. A. Barnett in M. H. Francombe, J. L. Vossen (eds.): *Physics of Thin Films Mechanic and Dielectric Properties,* vol. 17, Academic Press, San Diego 1993.
117. U. Helmersson et al., *J. Appl. Phys.* **62** (1987) 481 –484.
118. M. Shinn, L. Hultman, S. A. Barnett, *J. Mater. Res.* **7** (1992) 901 – 911.
119. P. B. Mirkarimi et al., *J. Mater. Res.* **9** (1994) 1456 –1467.
120. X. Chu et al., *J. Vac. Sci. Technol. A* **10** (1992) 1604 –1609.
121. K. K. Shih, D. B. Dove, *Appl. Phys. Lett.* **61** (1992) 654 – 656.
122. T. Friesen et al., *Surf. Coat. Technol.* **48** (1991) 169 –174.
123. A. Lettington, J. W. Steeds (eds.): *Thin Film Diamond,* Chapman & Hall, London 1994, p. 117.
124. J. E. Field (ed.): *The Properties of Natural and Synthetic Diamond,* Academic Press, London 1992.
125. J. Wilks, E. Wilks (eds.): *Properties and Applications of Diamond,* Butterworth-Heinemann, Oxford 1991.
126. *Diamond and Related Materials,* Journal of the Science and Technology of Diamond and Related Materials, Elsevier Science, Lausanne.
127. *Proc. IEEE* **79** (1991), special issue on "Large Band Gap Electronic Materials and Components."
128. C. A. Davies (ed.): "Properties of Diamond," *EMIS Data Rev. Ser.,* no. 9, INSPEC, Institution of Electrical Engineers, London 1994.
129. P. Reinke, W. Jacob, W. Möller, *J. Appl. Phys.* **74** (1993) 1354.
130. J. Robertson, *Diamond Related Mat.* **3** (1994) 361.
131. B. Dischler, A. Bubenzer, P. Koidl, *Solid State Commun.* **48** (1983) 105.
132. D. R. McKenzie et al., *Diamond Related Mat.* **3** (1994) 353.
133. R. S. Sussmann et al., *Diamond Related Mat.* **3** (1994) 303.
134. T. J. Valentine et al., *Diamond Related Mat.* **3** (1994) 1168.
135. B. Lux, R. Haubner in A. Lettington, J. W. Steeds (eds.): *Thin Film Diamond,* Chapman & Hall, London 1994, p. 127.
136. M. W. Geis, A. A. Tamor in G. L. Trigg (ed.): *Encyclopedia of Applied Physics,* vol. 5, VCH Verlagsgesellschaft, Weinheim 1993.
137. C. H. Wort, C. G. Sweeney, M. A. Cooper, G. A. Scarsbrook, R. S. Sussmann, *Diamond Related Mat.* **3** (1994) 1158.
138. M. Seal in A. Lettington, J. W. Steeds (eds.): *Thin Film Diamond,* Chapman & Hall, London 1994, p. 143.
139. J. E. Graebner et al., *Diamond Related Mat.* **2** (1993) 1059.

140. P. K. Bachmann et al., *Diamond Related Mat.* **4** (1995) 820.
141. T. Anthony, W. F. Banholzer, *Diamond Related Mat.* **1** (1992) 717.
142. P. Bade: "The Business and Technical Outlook for CVD Diamond and Diamond-Like Carbon," *Proc. 2nd Int. Gorham Conference,* Monterey, CA, Jan. 31 – Feb. 2, 1994.
143. B. Fox, presented at "Diamond Films 94," Il Ciocco, Italy, accepted for publication in Diamond and Related Mat. 1995.
144. F. J. Himpsel, J. A. Knapp, J. A. Van Vechten, D. E. Eastman, *Phys. Rev. B* **20** (1979) 624.
145. J. Robertson in A. Lettington, J. W. Steeds (eds.): *Thin Film Diamond,* Chapman & Hall, London 1994, p. 107.
146. A. Lettington in A. Lettington, J. W. Steeds (eds.): *Thin Film Diamond,* Chapman & Hall, London 1994, p. 117.
147. A. Matthews, S. S. Esklidsen, *Diamond Related Mat.* **3** (1994) 902.
148. W. G. Wood (ed.): *Metals & Materials Handbook,* vol. 5, American Society for Metals, Metals Park, OH, 1992.
149. C. Bocking, B. Cameron: "The Use of High Speed Selective Jet Electrodeposition of Gold for the Plating of Connectors," *Trans. Inst. Met. Finish.* **72** (1994) 33.
150. W. Riedel: *Electroless Nickel Plating,* Finishing Publications, Teddington, UK, 1991.
151. J. D. Greenwood: *Hard Chromium Plating,* 3rd ed., Portcullis Press, Redhill, UK, 1981.
152. C. F. Coombs, Jr. (ed.): *Printed Circuit Handbook,* 2nd ed., McGraw-Hill, New York 1979.
153. J. A. Scarlett: *The Multilayer Printed Circuit Handbook,* Electrochemical Publications, Ayr, Scotland, 1985.
154. A. Kempster: "The Principles and Applications of Chemical Vapour Deposition," *Trans. Inst. Met. Finish.* **70** (1992) 68 – 75.
155. Y. Sun, T. Bell: "Combined Plasma Nitriding and PVD Treatments," *Trans. Inst. Met. Finish.* **70** (1992) 38 – 44.
156. M. S. Wilson, S. Gottesfeld: "Thin-Film Catalyst Layers for Polymer Electrolyte Fuel Cell Electrodes," *J. Appl. Electrochem.* **22** (1992) 1 – 7.
157. A. K. Shukla, K. V. Ramesh, A. M. Kannan: "Fuel Cells: Problems and Prospects," *Proc. Indian Acad. Sci.* (Chem. Sci.) 97 (1986) 513 – 528.
158. A. J. Appleby, F. R. Foulkes (eds.): *Fuel Cell Handbook,* Van Nostrand Reinhold, New York 1989.
159. D. Pletcher, F. C. Walsh: *Industrial Electrochemistry,* 2nd ed., Blackies, Glasgow 1993, chap. 8, p. 358.
160. D. R. Gabe: *Principles of Metal Surface Treatment and Protection,* 3rd ed., Merlin Books, London 1993.
161. J. G. Bednorz, K. A. Muller, *Z. Phys. B: Condens. Matter* **64** (1986) 189 – 193.
162. Y. Q. Wang et al., *Appl. Phys. Lett.* **63** (1993) 3084 –3086.
163. T. Venkatesan et al., *Appl. Phys. Lett.* **53** (1988) 243 –245.
164. T. Terashima et al., *Jpn. J. Appl. Phys. Part I* **27** (1988) 91 – 93.
165. D. G. Schlom et al., *Z. Phys. B: Condens. Matter* **86** (1992) 163 – 175.
166. C. Klemenz, H. J. Scheel, *J. Cryst. Growth* **129** (1993) 421 – 428.
167. R. H. Hammond, R. Bormann, *Physica B+C* **162/164 B+C** (1989) 703 – 704.
168. J. D. Jorgensen et al., *Phys. Rev. B: Condens. Matter* **36** (1987) 3608 – 3616.
169. R. W. Simon et al., *Appl. Phys. Lett.* **53** (1988) 2677 –2679.
170. M. Sasaura, S. Miyazawa, M. Mukaida, *J. Appl. Phys.* **68** (1990) 3643 – 3644.
171. S. Miyazawa, M. Mukaida, *Appl. Phys. Lett.* **64** (1994) 2160.
172. G. Wegner, *Thin Solid Films* **216** (1992) 105.
173. G. Wegner, *Mol. Cryst. Liq. Cryst.* **235** (1993) 1.
174. G. Roberts: *Langmuir-Blodgett Films,* Plenum Press, New York 1990.
175. A. Ulman: *An Introduction to Ultrathin Organic Films,* Academic Press, Boston 1991.
176. M. Stamm in: *Macromolecules: Synthesis, Order and Advanced Properties; Advances in Polymer Science,* vol. 100, Springer Verlag, Berlin 1992, p. 357.
177. C. Bubeck, D. Holtkamp, *Adv. Mater.* **3** (1991) 32.
178. E. F. Aust, S. Ito, M. Sawodny, W. Knoll, *Trends Polym. Sci.* **2** (1994) 313.
179. M. Nieto-Vesperinas, J. C. Dainty: *Brillouin Scattering from Thin Films,* Elsevier Publ., Amsterdam 1990, p. 305.
180. D. Johannsmann, K. Mathauer, G. Wegner, W. Knoll, *Phys. Rev. B: Condens. Matter* **46** (1992) 1088.
181. H. Fuchs, H. Ohst, W. Prass, *Adv. Mater.* **3** (1991) 10.
182. J. D. Swalen, *J. Mol. Elec.* **2** (1986) 155.
183. G. Khanarian, *Thin Solid Films* **152** (1987) 265.

184. P. S. Vincett, G. Roberts, *Thin Solid Films* **68** (1980) 135.
185. K. Ogawa, H. Tamura, M. Sasago, T. Ishihara, *Proc. SPIE Int. Soc. Opt. Eng.* **771** (1987) 77.
186. M. Shen, A. T. Bell: "Plasma Polymerization," *ACS Symp. Ser.* **108** (1979) 1 – 33.
187. J. M. Tibbitt, M. Shen, A. T. Bell, *J. Macromol. Sci. Chem.* **A 10** (1976) 1623 – 1648.
188. H. Yasuda, *Plasma Polymerization,* Academic Press, Orlando 1985.
189. M. N. Morosoff in R. d'Agostino (ed.): *Plasma Deposition, Treatment and Etching of Polymers,* Academic Press, Boston 1990.
190. H. Simon: *The Splendor of Iridescence,* Dodd, Mead & Company, New York 1971.
191. S. D. Smith: "Design of Multilayer Filters by Considering Two Effective Interfaces," *J. Opt. Soc. Am.* **48** (1958) 43 – 50.
192. L. I. Epstein: "The Design of Optical Filters," *J. Opt. Soc. Am.* **42** (1952) 806 – 810.
193. L. Young: "Synthesis of Multiple Antireflection Films Over a Prescribed Frequency Band," *J. Opt. Soc. Am.* **51** (1961) 967 – 974.
194. P. Baumeister: "Design of Multilayer Filters by Successive Approximations," *J. Opt. Soc. Am.* **48** (1958) 955 – 958.
195. J. A. Dobrowolski: "Completely Automatic Synthesis of Optical Thin Film Systems," *Appl. Opt.* **4** (1965) 937 – 946.
196. W. H. Southwell: "Coating Design Using Very Thin High- and Low-index Layers," *Appl. Opt.* **24** (1985) 457 – 460.
197. J. A. Dobrowolski, D. Lowe: "Optical Thin Film Synthesis Program Based on the Use of Fourier Transforms," *Appl. Opt.* **17** (1978) 3039 – 3050.
198. P. Baumeister: "Computer Software for Optical Coatings," *Photonics Spectra* **22** (1988) no. 9, 143 –148.
199. H. A. Macleod: *Thin-film Optical Filters,* 2nd ed., Adam Hilger Ltd., Bristol 1986, pp. 357 – 445.
200. H. Schroeder: "Oxide Layers Deposited from Organic Solutions," in G. Hass, R. E. Thun (eds.): *Physics of Thin Films,* vol. 5, Academic Press, New York 1969, pp. 87 – 141.
201. P. J. Martin, R. P. Netterfield: "Optical Films Produced by Ionbased Techniques," in E. Wolf (ed.), *Prog. Opt.* **13** (1986) 113 – 182.
202. D. Z. Rogers: "Manufacture of Optical Interference Coatings by Low Pressure Chemical Vapor Deposition," *Proc. SPIE Int. Soc. Opt. Eng.* **1168** (1989) 19 – 24.
203. W. Klug, R. Schneider, A. Zöller: "Plasma Enhanced CVD Hard Coating for Opthalmic Lenses," *SPIE* **1323** Optical Thin Films III, New Developments (1990) 88 – 97.
204. J. Segner: "Plasma Impulse Chemical Vapour Deposition – a Novel Technique for the Production of High Power Laser Mirrors," *Mat. Sci. Eng. A* **140** (1991) 733 – 740.
205. R. W. Bryant: "Coating Growth Accelerates Through the Decade and Beyond," *Laser Focus World* **28** (1992) Oct., 117 – 118.
206. A. Thelen: *Design of Optical Interference Coatings,* McGraw-Hill, New York 1989, p. 243.
207. P. Nath, R. F. Bunshah: "Preparation of In_2O_3 and Tin-doped In_2O_3 Films by a Novel Activated Reactive Evaporation Technique," *Thin Solid Films* **69** (1980) 63 – 68.
208. F. G. K. Baucke, B. Metz, J. Zauner: "Elektrochrome Schichtsysteme mit variierbaren optischen Eigenschaften," *Phys. Unserer Zeit* **18** (1987) no. 1, 21 – 28.
209. C. King: "Thin Film Electroluminescent Displays," *Society of Automotive Engineers, West Coast International Meeting and Exposition,* Portland, OR, Aug. 4 – 8, 1985, Paper 851 461.
210. E. Schultheiss et al.: "Production Technology for Magneto-optic Data Storage Media," *Solid State Technol.* **31** (1988) March, 107 – 112.
211. J. M. Pinneo, L. C. Conner: "CVD Diamond: New Material, New Applications," *Photonics Spectra* **23** (1989) Oct., 123 – 126.
212. W. S. Wiston Ho, K. K. Sirkar: *Membrane Handbook,* Van Nostrand Reinhold, New York 1992.
213. S. T. Hwang, K. Kammermeyer: "Membranes in Separations," *Techniques in Chemistry,* **vol. VII,** J. Wiley & Sons, New York, 1965.
214. R. E. Kesting: *Synthetic Polymeric Membranes,* J. Wiley & Sons, New York, 1985.

Thiocyanates and Isothiocyanates, Organic

Frank Romanowski, Degussa AG, Frankfurt, Federal Republic of Germany

Herbert Klenk, Degussa AG, Frankfurt, Federal Republic of Germany

1. Organic Thiocyanates 1
1.1. Introduction 1
1.2. Physical Properties 1
1.3. Chemical Properties 1
1.4. Production 2
1.5. Use and Commercial Names 3
1.6. Toxicology 3
2. Organic Isothiocyanates 4
2.1. Introduction 4
2.2. Physical Properties 4
2.3. Chemical Properties 4
2.4. Production 6
2.5. Quality Specifications and Analysis 7
2.6. Storage and Transport 7
2.7. Commercial Names 7
2.8. Uses 7
2.8.1. Methyl Isothiocyanate 7
2.8.2. Other Organic Isothiocyanates 8
2.9. Toxicology and Industrial Hygiene . 8
2.10. Safety Measures and Processing Advice 9
2.11. Ecological Aspects 9
3. References 9

1. Organic Thiocyanates

1.1. Introduction

Organic thiocyanates are esters of thiocyanic acid. In the general formula R−S−C≡N, the group R can be aliphatic, alicyclic, aromatic, or heterocyclic. So far, no organic thiocyanates have been detected in plants or animals.

Special applications have been discovered and developed for only a few synthetic thiocyanates. Certain organic thiocyanates are used as vulcanization accelerators, others have been used as preservatives because of their biocidal and fungicidal properties. Small quantities are produced as fine chemicals.

1.2. Physical Properties [1]

The low molecular mass alkyl thiocyanates are colorless oils with an odor reminiscent of leeks. They are insoluble in water, but soluble in ethanol and ether, for example. Some physical properties are listed in Table 1.

1.3. Chemical Properties

Thiocyanates isomerize on warming to isothiocyanates. Allylic and tertiary thiocyanates isomerize rapidly, so some thiocyanates can be converted quantitatively to isothiocyanates [2]. In the case of methyl thiocyanate, it is possible to produce methyl isothiocyanate on an industrial scale by isomerization (Section 2.4).

Reducing agents cleave thiocyanates to thiols and hydrogen cyanide.

$$\mathrm{RSCN + [H] \longrightarrow RSH + HCN}$$

The reduction can be carried out with Sn/HCl, $LiAlH_4$, or sodium in liquid ammonia.

Organic thiocyanates are converted to sulfonic acids by strong oxidizing agents:

$$\mathrm{R-SCN + [O] \xrightarrow{H_2O} RSO_3H}$$

The most commonly used oxidizing agent is nitric acid, but alkaline hypochlorite or hydrogen peroxide can also be used. Acid hydrolysis of organic thiocyanates leads to thiols via thiocarbamates:

$$\mathrm{R-SCN \xrightarrow[(H^+)]{H_2O} R-S-C(=O)NH_2 \xrightarrow[(H^+)]{H_2O} RSH + CO_2 + NH_3}$$

Alkaline hydrolysis gives the corresponding disulfides in good yields [3]:

$$\mathrm{2\,RSCN + 2\,OH^- \longrightarrow RSSR + OCN^- + CN^- + H_2O}$$

Because of its polarity, the C≡N bond can undergo addition reactions. Intramolecular reac-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_749

Table 1. Physical properties of some thiocyanates

Name, formula	M_r	CAS	Freezing point, °C	Boiling point, °C (kPa)	Density, g/cm^3
Methyl thiocyanate, CH_3SCN	73.11	[*556-64-9*]	– 51	132 (101.3)	1.074
Ethyl thiocyanate, C_2H_5SCN	87.14	[*542-90-5*]	– 85.5	145.5 (101.9)	1.011
3-Thiocyanatopropyltriethoxysilane, $(C_2H_5O)_3Si(CH_2)_3SCN$	263.42	[*34708-08-2*]	– 80	95 (0.013)	1.000
2-(2-Butoxyethoxy)ethyl thiocyanate, $C_4H_9OC_2H_5OC_2H_5SCN$	205.28	[*112-56-1*]		120 – 122 (0.266)	1.016
Methylenebis(thiocyanate), $NCSCH_2SCN$	130.18	[*6317-18-6*]	103 – 104		

tions for the synthesis of heterocycles are particularly important [4], [5].

1.4. Production

An important method for the production of alkyl thiocyanates is substitution of a halogen in an alkyl halide by the thiocyanate anion. The most frequently used solvents are ethanol, water, and DMF [4], [6]. As the thiocyanate anion is ambidentate, the reaction can give mixtures of thiocyanates and isothiocyanates, or can proceed selectively to either isomer [7]:

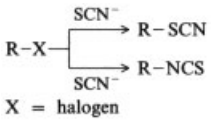

The selectivity of the reaction depends on the solvent, temperature, and thiocyanate concentration. Methylenebis(thiocyanate) [8–10] and methyl thiocyanate [11], [12], for example, can be produced by this method.

Organic thiocyanates isomerize partially or completely to the corresponding isothiocyanates during the production process, depending on the type of compound. For example, only isothiocyanates can be produced from β,γ-unsaturated alkyl halides.

Aromatic thiocyanates can be produced by reaction of arenes with dicyanodisulfane. Addition of a Lewis acid increases the electrophilicity of the dicyanodisulfane so much that even unreactive aromatics react:

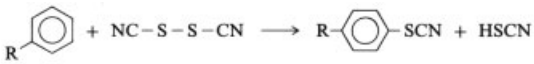

A further route for the production of aromatic thiocyanates involves reaction of diazonium compounds with salts of thiocyanic acid in the presence of copper salts:

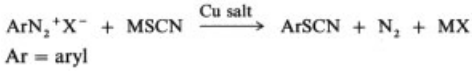

A valuable extension of the possibilities for producing organic thiocyanates is the reaction of thiols with cyanogen chloride [13], particularly as no isomeric isothiocyanate can be formed.

$$RSH + ClCN \longrightarrow RSCN + HCl$$

Other methods exist for the production of organic thiocyanates, but these are not of industrial interest.

Thiocyanic acid adds to epoxides, forming 2-hydroxyalkyl thiocyanates [14], [15]:

Addition of dicyanodisulfane to olefins results in quantitative formation of dithiocyanatoalkanes; the reaction is used to detect double bonds analytically in fat chemistry [16]. Radical initiators or sunlight accelerate the reaction, so a radical addition mechanism is probable.

Reaction of alkylating agents, e.g., dialkyl sulfates or alkyl sulfonates, with salts of thiocyanic acid gives organic thiocyanates [17], [18]:

$$R^1\text{–}SO_2OR^2 + MSCN \longrightarrow R^2SCN + R^1\text{–}SO_2OM$$

An overview of production methods for thiocyanates can be found in reference [6].

1.5. Use and Commercial Names

Industrial applications of organic thiocyanates have been developed for only a few compounds.

Methyl thiocyanate is isomerized to methyl isothiocyanate on an industrial scale (Section 2.4). A small number of compounds are used as insecticides. The compounds known as Lethanes [19] are used as contact insecticides for flies, mosquitoes, and household pests. Lethane 384 [*112-56-1*] has the formula $C_4H_9OC_2H_4OC_2H_4SCN$; Lethane 60 [*301-11-1*] has the formula $C_{11}H_{23}COOC_2H_4SCN$.

Lethane 60 (2-thiocyanatoethyl dodecanoate) is tolerated better by plants than Lethane 384 (2-(2-butoxyethoxy)ethyl thiocyanate) and can thus be used for combatting greenfly.

Methylenebis(thiocyanate) (MBT) and 2-(thiocyanatomethylthio)benzothiazole [*21564-17-0*] [20] exhibit biocidal properties. Both compounds are supplied ready for use in widely differing formulations under the name Busan. The quantities required for use as biocides are predicted to rise [21].

MBT hinders the propagation of slime-forming bacteria and fungi. As Busan 110, it is preferably used in refrigerator circulation systems and in papermills as a 10 % solution in inert solvents [22]. To lower the COD in the circulation water in papermills MTB is used as a stable aqueous dispersion [23]. MTB is employed as a fumigant for counteracting fungal growth [24]. Other applications of methylenebis(thiocyanate) are described in the literature [25–28].

2-(Thiocyanatomethylthio)benzothiazole has been known as a pesticide for a long time [29]. It is sold as a solution in an inert solvent under the name Busan. It is used as a broad-spectrum bactericide und fungicide in various areas. It is thus possible to preserve water-based or organic-solvent-based dyes with Busan 30 [30]. All types of wood are protected from fungi and insects by a coating of the substance. Attack of leather by mold during processing can be prevented by treatment with 2-(thiocyanatomethylthio)benzothiazole [31], [32]. An advantage of the compound over pentachlorophenol is its better environmental compatibility.

3-Thiocyanatopropyltriethoxysilane (Si 264), $(C_2H_5O)_3SiCH_2CH_2SCN$, is used as a reinforcing additive in molding compounds which can be vulcanized by heating with peroxides. The molding compounds are more readily processed, and the polymer obtained after vulcanization has markedly improved properties [33].

1.6. Toxicology

In the case of alkyl thiocyanates, the toxicity in rodents after oral and subcutaneous administration decreases with increasing molecular mass. The higher homologues have an irritant, sometimes corrosive effect on the skin and mucous membranes. They cause irreversible damage to the outer epithelium, death of tissue (necrosis), and defect formation (ulceration), with subsequent hyperplastic reconstruction of the outer epithelium. The systemic toxicity is high (Table 2). Acute poisoning leads to a high degree of circulatory disturbance in the inner organs, with increases in circulation, bleeding, and possibly death.

2. Organic Isothiocyanates

2.1. Introduction

Organic isothiocyanates are esters of isothiocyanic acid. In the general formula R−N=C=S, the group R can be aliphatic, alicyclic, aromatic, acyl, or heterocyclic.

Organic isothiocyanates do not occur naturally in the free form. They are liberated from glucosinolates in plant tissues by enzymatic degradation. Their pungent odor and sharp, mustard-like taste gave these compounds the name mustard oils, by analogy with the ethereal oils of black mustard [34]. Allyl isothiocyanates are liberated from the glucosinolate sinigrin of black mustard by myrosinase [35], [36]. Methyl isothiocyanate is formed by the enzymatic degradation of glucocapparin, which is contained in the caper plant.

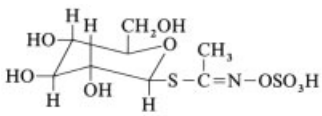

Glucocapparin

Table 2. Toxicological data for important organic thiocyanates

Name	RTECS	Administration route	Species	LD_{50}, mg/kg
Methyl thiocyanate	XL 1575000	oral	rat	60
		intravenous	mouse	18
Methylenebis(thiocyanate)	XL 1560000	oral	rat	161
		intravenous	mouse	3.6
2-(2-Butoxyethoxy)ethyl thiocyanate	XK 8400000	oral	rat	90
		intravenous	mouse	56
2-(Thiocyanatomethylthio)benzothiazole	XK 8150000	oral	rat	1590
		dermal	rabbit	642

Many other organic isothiocyanates with, e.g., sulfur-containing or aromatic groups have become known as natural products through the degradation of the corresponding glucosinolates.

Aliphatic and aromatic isothiocyanates are important starting materials for the production of thiourea derivatives, substituted thiosemicarbazides, heterocycles, and other organic intermediates. Some isothiocyanates have nematocidal, bactericidal, and fungicidal properties. Allyl (→ Flavors and Fragrances, Chap. 2.1.6.), methoxycarbonyl, and methyl isothiocyanate are produced on an industrial scale.

2.2. Physical Properties

Most aliphatic and aromatic isothiocyanates are colorless oils with a pungent odor. Methyl isothiocyanate (→ Nematocides) is a solid at room temperature. Many isothiocyanates are not only relatively stable to water in the cold, but can also be steam distilled. Aliphatic isothiocyanates with the general formula R−N=C=S, in which R is an unbranched or branched alkyl group with 2 – 6 carbon atoms, distil azeotropically with water at 80 – 120°C. The boiling points of the pure compounds lie in the range 120 – 200 °C.

Organic isothiocyanates are readily soluble in most common organic solvents.

Table 3 shows the most important physical properties of some isothiocyanates (for further data on isothiocyanates see reference [1]).

2.3. Chemical Properties

The chemical properties of organic isothiocyanates are similar to those of organic isocyanates, but the latter are generally more reactive [37]. Most reactions of isothiocyanates involve addition of a nucleophile to the −N=C=S group. Some examples of nucleophilic addition are given below [1], [38]:

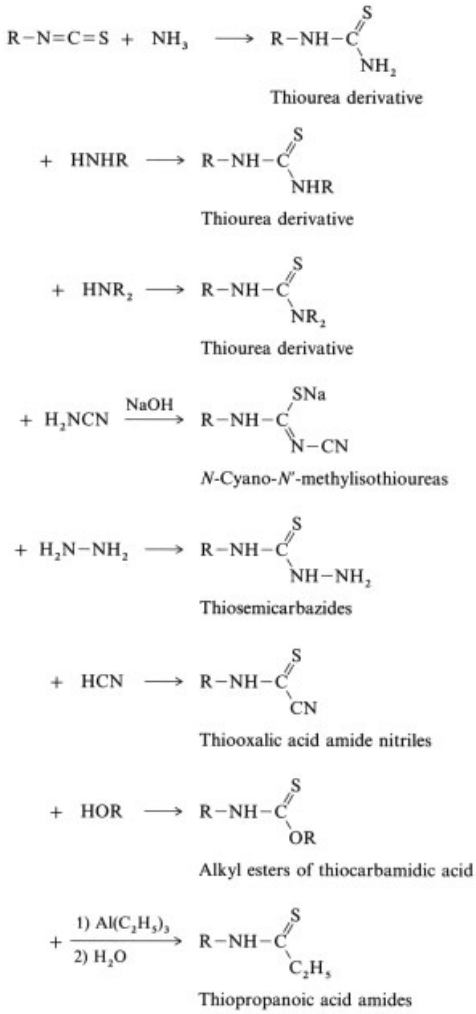

Table 3. Physical data for some isothiocyanates

Name	M_r	CAS	Freezing point, °C	Boiling point, °C (kPa)	Density, g/cm^3
Methyl isothiocyanate, CH_3NCS	73.11	[*556-61-6*]	35.9	118.7 (100.9)	1.0691 (37°C)
Allyl isothiocyanate, $CH_2{=}CHCH_2NCS$	99.15	[*57-06-7*]	−102.5	151.8 (100.7)	1.015 (20°C)
Phenyl isothiocyanate, C_6H_5NCS	135.18	[*103-72-0*]	−21	221 (100.5)	1.138 (15.5°C)
Vinyl isothiocyanate, $CH_2{=}CHNCS$	85.12	[*1520-22-5*]		46 (10.0)	1.018 (15°C)
Methoxycarbonyl isothiocyanate, $CH_3OCONCS$	117.10	[*35266-49-0*]		30 (0.12)	1.152 (15°C)

Isothiocyanates have a very important function in the synthesis of heterocycles. The large number of reactions which have been discovered is notable, and with them access to new compounds has become possible [39–42].

Interest in the development of corresponding synthetic processes was directed towards five- and six-ring heterocycles, but four- and seven-ring systems can also be synthesized, such as benzimidazoles, benzothiazoles, and other bicyclic heterocycles.

Amino acids can be converted to thiohydantoin derivatives [43]:

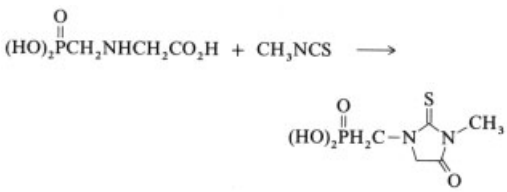

Acyl isothiocyanates react with 1,3-diketones to 5-acyl-4*H*-1,3-oxazine-4-thiones [44]:

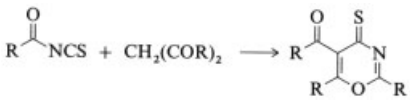

Anthranilic acid reacts with methyl isothiocyanate to 3-*N*-methyl-2-thioxotetrahydroquinoxalin-4-one [45]:

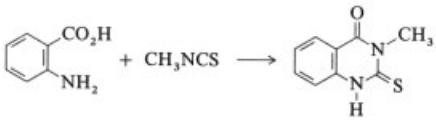

The isothiocyanate group −N=C=S can undergo cycloadditions with suitable reaction partners. Methyl isothiocyanate reacts with sodium azide to 5-mercapto-4-methyl-1,2,3,4-tetrazole [46]:

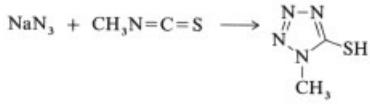

2.4. Production

A large number of processes are known for the production of isothiocyanates [39], [41], [47], [48]. The structure desired and sometimes the other functional groups on the organic skeleton determine the choice of production process.

The most important production methods are:

1) Reaction of alkyl and acyl halides with salts of thiocyanic acid [49], [50]
2) Isomerization of esters of thiocyanic acid [51], [52]
3) Reaction of dithiocarbamates (→ Dithiocarbamic Acid and Derivatives, Chap. 4.6.) with: hydrogen peroxide [53], [54], cyanogen chloride [55], cyanuric chloride [56], chloroformates [57], alkali hypochlorite or alkali chlorite [58], phosgene [59], heavy metal salts [60], atmospheric oxygen [61], [62], phosphorus oxychloride [63]
4) Reaction of primary amines with thiophosgene [64]
5) Reaction of thiourea derivatives with mineral acids, or acetic anhydride, or by heating [39]
6) Addition of thiocyanic acid to unsaturated compounds [65]

For vinyl isothiocyanates, special production methods are usually necessary [48].

Methyl isothiocyanate, which is industrially important, is produced by two different processes:

1) Rearrangement of methyl thiocyanate

2) Oxidation of sodium *N*-methyldithiocarbamate with hydrogen peroxide

The isomerization of methyl thiocyanate begins above 100 °C. The reaction is catalyzed by salts, e.g., $ZnCl_2$, alkali-metal thiocyanate [2].

$$CH_3SCN \longrightarrow CH_3N{=}C{=}S$$

A patent specification by Morton Chemical [52] describes a multistep continuous process for the isomerization. The temperature of the reaction, which requires a long induction period, is 130 – 160 °C. In the course of reaction, a mixture is formed, which promotes the isomerization autocatalytically and favors the shift in the equilibrium. The methyl isothiocyanate formed is continuously removed from the vapor phase together with methyl thiocyanate. This mixture is subjected to fractional distillation and the methyl thiocyanate is recycled into the isomerization. The methyl isothiocyanate removed is replaced by fresh methyl thiocyanate. The side products formed (sulfides, disulfides) are removed periodically from the liquid phase. It is possible to convert up to 90 % of the methyl thiocyanate into isothiocyanate by this process.

A patent specification by Stauffer [51], describes the isomerization of methyl thiocyanate in the gas phase at 300 – 400°C on a fixed-bed catalyst. The carrier (silica gel, activated charcoal) is impregnated with catalytically active salts, e.g., alkali thiocyanate, $ZnCl_2$. The vapor phase at the head of the reactor contains up to 55 % methyl isothiocyanate. It is purified by fractional distillation and the recovered methyl thiocyanate is recycled.

The production of methyl isothiocyanate from sodium methyldithiocarbamate and hydrogen peroxide has been developed on a large industrial scale [53], [54]. The formation of methyl isothiocyanate takes place via several steps:

$$2CH_3NH{-}\overset{\overset{\displaystyle S}{\|}}{C}{-}SNa + H_2O_2 \longrightarrow \left[CH_3NH\overset{\overset{\displaystyle S}{\|}}{C}{-}S{-}S{-}\overset{\overset{\displaystyle S}{\|}}{C}NHCH_3\right] + 2NaOH$$

$$CH_3HN{-}\overset{\overset{\displaystyle S}{\|}}{C}{-}S{-}S{-}\overset{\overset{\displaystyle S}{\|}}{C}{-}NHCH_3 \longrightarrow 2CH_3{-}N{=}C{=}S + S + H_2S$$

$$H_2S + 4H_2O_2 \longrightarrow H_2SO_4 + 4H_2O$$

The reaction can be represented by the overall equation:

$$2CH_3NH\overset{\overset{\displaystyle S}{\|}}{C}{-}SNa + 5H_2O_2 \longrightarrow 2CH_3{-}N{=}C{=}S + S + Na_2SO_4 + 6H_2O$$

With smaller quantities of hydrogen peroxide, sodium thiosulfate is predominantly formed as the side-product:

$$2CH_3NH\overset{\overset{\displaystyle S}{\|}}{C}{-}SNa + 4H_2O_2 \longrightarrow 2CH_3{-}N{=}C{=}S + Na_2S_2O_3 + 5H_2O$$

Industrial-scale production is as follows: ca. 40 % aqueous sodium methyldithiocarbamate solution (pH 10 – 12) is reacted with 30 – 40 % hydrogen peroxide at ca. 100°C. The pH during the reaction is 3 – 4. The reaction mixture is heated so that the methyl isothiocyanate immediately distils as an azeotrope with water. Methyl isothiocyanate (ca. 5 – 7 %) is dissolved in the water that distils over. This can easily be separated by a short distillation. A very pure product is formed in ca. 85 % yield. The sulfur precipitated during the reaction can be separated readily from the wastewater, which is readily biodegradable. The small quantity of waste-gas formed is incinerated.

Allyl isothiocyanate is produced by reaction of allyl chloride with alkali rhodanides in the two-phase system water/1,2-dichloroethane [66].

All commercial methoxycarbonyl isothiocyanate is produced in situ from methyl chloroformate and potassium thiocyanate in acetone. It is then converted immediately to thiophanate methyl (Section 2.8.2) [67].

Vinyl isothiocyanate can be produced in 63 % yield in a two-step reaction. 2-Aminobromoethane is reacted with thiophosgene in ether. The 2-bromoethyl isothiocyanate obtained after work-up and distillation is converted into vinyl isothiocyanate by elimination of hydrogen bromide [68].

2.5. Quality Specifications and Analysis

The processing of industrially produced organic isothiocyanates to higher-value products and their use in pharmaceuticals requires high purity. The quality of the product is indicated by the −N=C=S content. To determine this, the product is treated with excess dibutylamine in dioxan or chlorobenzene/methanol, and the excess amine is titrated [69], [70]. Gas – liquid chromatography gives a detailed overview of the side-product spectrum.

2.6. Storage and Transport

Methyl isothiocyanate is generally stored in tanks or in 200-kg steel drums. The product is stable to storage, provided it is free of water. The containers should be stored and handled tightly closed at normal temperature in a dry, well-ventilated area, because the product has a severely irritant effect on the skin and mucous membranes and is toxic [71].

Methyl thiocyanate: UN no. 2477, EINECS 209-132-5, IMDG 6.1 PG II, RID/ADR 6.1 no. 20b.

2.7. Commercial Names

Methyl isothiocyanate is sold worldwide by Degussa (Germany) and Morton Chemical (United States).

In formulations, it is sold under the names Trapex and Di-Trapex by Schering. Di-Trapex is sold by Nor-Am under the name Vorlex in the United States. Osmose sells methyl isothiocyanate as MITC-FUME in glass and aluminum ampules. Allyl isothiocyanate is sold in Germany by Haarman & Reimer and in the United Kingdom by Rhône-Poulenc.

2.8. Uses

2.8.1. Methyl Isothiocyanate

Consumption of methyl isothiocyanate is <2000 t/a (1993); 80 % is used in agrochemicals, ca. 15 % in pharmaceuticals, and ca. 5 % in other applications.

Agrochemicals. Methyl isothiocyanate is used directly as a 20 % formulation in an inert solvent (Vorlex, Trapex, Di-Trapex) for protecting soils against fungi and nematodes.

1,3,4-Thiadiazole derivatives are known for their herbicidal properties. Methyl isothiocyanate is used for their synthesis. The most important examples are Spike (Eli Lilly, United States) [72], Ustilan (Bayer, Germany) [73], and Erbotan (Ciba-Geigy, Switzerland) [74].

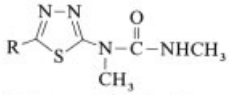

Spike	R = *t*-butyl
Ustilan	R = ethylsulfonyl
Erbotan	R = trifluormethyl

Pharmaceuticals. Methyl isothiocyanate is an important basic chemical for the synthesis of the H_2-blockers ranitidine (Zantac, Glaxo) [75], [76] and cimetidine (Tagamet, Smith Kline & French) [77].

5-Mercapto-1-methyltetrazole is obtained by treating methyl isothiocyanate with sodium azide. It is a side-group in the cefalosporin antibiotics Latamoxef [78] and Cefamandol [79].

Other Uses. Methyl isothiocyanate is very effective in the protection of wood against rotting [80] and premature decay.

A special application of decay prevention for telegraph poles in the United States has been developed by Osmose Wood Preserving. Four holes, on average, are bored at an angle of 45° in the poles. Ampules of methyl isothiocyanate (MITC-FUME) are inserted into the holes and the latter are closed with wood plugs. The methyl isothiocyanate diffuses slowly into the wood. After ca. 30 d the substance can be detected at 60 – 80 cm from the opening of the ampule in dry Douglas pine [81]. The moisture content of the wood determines the rate of migration, the effectiveness, and also the time at which further treatment is required. The decision as to whether a secondary impregnation with MITC-FUME is necessary is made on inspection of the poles after 6 years [82]. Effectiveness is estimated to last 12 years, so that the second inspection cycle is within the expected period of activity. The lifetime of the telegraph poles is expected to be

doubled by subsequent treatment with MITC-FUME [83].

2.8.2. Other Organic Isothiocyanates

Allyl isothiocyanate is mainly used in the sugar industry because of its fungicidal properties. It protects sugar beet from fungi during storage [84]. Allyl isothiocyanate is also known as an insecticide and an antimicrobial agent, and is used in perfumes. It is used for combatting the cabbage maggot fly *H. brassicae* .

2-(Hydroxyethyl)allylthiourea, produced from allyl isothiocyanate, has been used for producing light-sensitive paper. Methoxycarbonyl isothiocyanate [67], [85] is treated in situ with *o*-phenylenediamine to give thiophanate methyl (→ Fungicides, Agricultural). Thiophanate methyl [*23564-05-8*] is a systemic fungicide, sold under the names Topsin M (Nippon Soda, Japan), Mildothane (May & Baker, United Kingdom), and Cycosin (American Cyanamid, United States). In 1980, United States consumption of thiophanate methyl was 150 t.

2.9. Toxicology and Industrial Hygiene

As vapors, organic isothiocyanates have a severely irritant effect on the skin, eyes, and respiratory tract. Direct contact with the body leads to severe damage to the areas affected. Skin resorption is possible. Oral administration leads to general disturbance of function, cramp, and unconsciousness. Acute toxicity after oral administration in animals is generally high (Table 4).

Methyl Isothiocyanate [86–88]. The vapor has a severely irritant effect on the skin and mucous membranes of the eyes and respiratory tract, and gives rise to a burning pain. On dermal contact, first- or second-degree acid burns are possible. Without immediate treatment, corneal opacity of the eyes is also possible. After swallowing, burning, pain, vomiting, cramp, unconsciousness, complete loss of reflexes, reduction in blood pressure, and death can occur. Methyl isothiocyanate can damage the liver and kidneys and cause lung edema.

Allyl Isothiocyanate [89]. Local and systemic toxic effects are very similar to those of methyl isothiocyanate.

2.10. Safety Measures and Processing Advice

In all work with organic isothiocyanates a protective mask with combination filter B, thick protective clothing, and safety glasses must be worn. Work areas must be well ventilated and, because sight can be impaired in the event of an accident, straight-line, free escape routes must be provided. Apparatus and opened vessels must be extracted while filling and emptying. Vapors form explosive mixtures with air, so ignition sources must be excluded.

Methyl isothiocyanate which has leaked must be covered with cold water, and the crusts formed added in portions to 10 % aqueous ammonia, and disposed of separately. Spilled allyl isothiocyanate should be covered with activated charcoal and the latter disposed of carefully.

2.11. Ecological Aspects

Organic isothiocyanates, particularly methyl and allyl isothiocyanates, represent a hazard to bodies of water. Measures should be taken

Table 4. Toxicological data for organic isothiocyanates

Name	RTECS	Administration route	Species	LD_{50}, mg/kg
Methyl isothiocyanate	PA 9625000	oral	rat	67
		dermal	rabbit	174
		inhalation	rat	29.6
Allyl isothiocyanate	NX 8225000	oral	rat	112
		dermal	rabbit	88
Phenyl isothiocyanate	NX 9275000	oral	mouse	87

to avoid contamination of lakes, streams, and groundwater, under all circumstances.

3. References

1. E. E. Reid: *Organic Chemistry of Bivalent Sulfur,* vol. VI, Chem. Publ., New York 1966.
2. A. Fava in N. Kharasch, Cal. Y. Meyers (eds.): *The Chemistry of Organic Sulfur Compounds,* **vol. 2,** Pergamon Press, New York 1966, pp. 73 – 91.
3. E. Hoggart, W. A. Sexton, *J. Chem. Soc.* 1947, 815.
4. R. G. Guy in S. Patai (ed.): *The Chemistry of Cyanates and their Thio Derivatives,* part 2, J. Wiley & Sons, Chichester 1977, pp. 819 – 878.
5. B. Schulze, M. Mühlstädt, *Z. Chem.* **19** (1979) 41 – 48.
6. *Houben/Weyl,* **E4,** 941 – 969.
7. D. Knoke, K. Kottke, R. Pohloudek-Fabini, *Pharmazie* **28** (1973) 574, 617; **32** (1977) 195 – 215.
8. Sumitomo Chem., DE 1 232 574, 1967 (T. Kawanami, G. Suzukamo).
9. Nalco Chem., DE 1 668 494, 1967 (J. Matt, E. W. Hunter, L. A. Goretta).
10. Kinkai Kagaku, JP-Kokai 6 007 2858, 1985 (K. Shinya, H. Seuda, S. Kojima).
11. Morton, DE 1 295 549, 1966 (J. T. Venerable, J. Miyashiro, A. W. Seiling).
12. Bayer, DE 1 183 903, 1965 (K. Goliasch, E. Grigat, R. Pütter).
13. Degussa, DE 1 270 553, 1966 (Ch. Kosel).
14. H. D. Vogelsang, T. Wagner-Janregg, R. Rebling, *Justus Liebigs Ann. Chem.* **569** (1950) 183.
15. F. G. Weber, H. Ließert, R. Radeglia, *Pharmazie* **35** (1980) 330.
16. H. P. Kaufmann: *Analyse der Fette und Fettprodukte,* **vol. I,** Springer Verlag, Berlin 1958, p. 570.
17. Dow Chemical, US 2 939 875, 1958 (U. D. Floria).
18. E. J. Corey, R. B. Mitra, *J. Am. Chem. Soc.* **84** (1962) 2938.
19. K. H. Büchel in R. Wegler (ed.): *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 1,** Springer Verlag, Berlin 1970, pp. 457 – 459.
20. Buckman Laboratories, US 1 129 575, 1966.
21. Kline & Company, Biocides-Industry Report, Investext, March 26, 1990, pp. 1 – 4.
22. Buckman Laboratories, Product Data Sheet, Memphis, Tenn., July 1981.
23. W. D. Henkels, A. Schenker, J. F. Schuetz, *Wochenbl. Papierfabr.* **112** (1984) no. 23/24, 869.
24. Yamato Chem. Ind., JP-Kokai 59 010 530, 1984 (A. W. Chester, Chu Yang Feng).
25. Ichikawa Gosei Kagaku, JP-Kokai 03 007 205, 1991 (Y. Igarashi, T. Tsunoda, K. Keisake, R. Imai).
26. Lester Technologies Corp., US 4 975 109, 1990 (L. A. Friedman, R. F. McFarlin).
27. T. C. Presuell, D. D. Nicholas, *For. Prod. J.* **40** (1990) 57 – 61.
28. W. R. Grace and Co., EP 282 203, 1988 (D. Clarkson, R. P. Cliffort).
29. T. Egli, E. Sturm in R. Wegler (ed.): *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* **vol. 6,** Springer Verlag, Berlin 1981, p. 380.
30. Buckman Laboratories, product information "BUSAN 30L".
31. H. Gattner, W. Lindner, H.-U. Neuber, *Leder* **39** (1988) 66.
32. W. M. Fowler, A. E. Russel, C. G. Whiteley, *J. Am. Leather Chem. Assoc.* **85** (1990) 243.
33. Degussa, DE 4 100 217, 1992 (U. Görl, S. Wolff).
34. A. W. Hofmann, *Ber. Dtsch. Chem. Ges.* **1** (1868) 25.
35. F. Hoffmann, *Chem. Unserer Zeit* **12** (1978) 182.
36. A. Kjaer in N. Kharasch (ed.): *Organic Sulfur Compounds,* **vol. 1,** Pergamon Press, New York 1961, p. 409.
37. *Houben/Weyl,* **IX,** 768.
38. S. J. Assony in N. Kharasch (ed.) : *Organic Sulfur Compounds,* **vol. 1,** Pergamon Press, New York 1961, p. 326.
39. S. Sharma, *Sulfur Rep.* **8** (1989) 327.
40. A. K. Mukerjee, R. Ashare, *Chem. Rev.* **91** (1991) 1.
41. M. Avalos et al., *Heterocycles* **33** (1992) 973 – 1010.
42. G. Giesselmann, K. Huthmacher, H. Klenk, F. Romanowski, *Chem. Ztg.* **114** (1990) 215 – 224.
43. Schering, DE 3 427 794, 1986 (F. Blume et al.).
44. H. Dehne, P. Krey, *J. prakt. Chem.* **324** (1982) 915.
45. Ch. H. Chan, *Heterocycles* **26** (1987) 3193.
46. H. W. Altland, *J. Org. Chem.* **41** (1976) 3395.
47. *Houben/Weyl,* **E4,** 834 – 883.
48. K. Schulze, B. Schulze, Ch. Richter, *Z. Chem.* **29** (1989) 41.

49. Ciba-Geigy, DE 3 504 016, 1985 (M. Böger, J. Drabek).
50. B. S. Drach, E. P. Sviridov, A. V. Kirsanov, *Zh. Org. Khim.* **8** (1972) 1872.
51. Stauffer Chem., US 2 954 393, 1960 (J. N. Haimsohn, G. E. Lukes).
52. Morton Chem., DE 1 253 265, 1962 (J. T. Venerable, J. Miyashiro, P. L. Weyna).
53. Degussa, DE 2 105 473, 1971 (G. Giesselmann, K. Günther).
54. Degussa, DE 4 001 020, 1990 (G. Giesselmann, W. Schwarze).
55. Degussa, DE 2 603 508, 1976 (G. Giesselmann, R. Vanheertum, G. Schreyer).
56. Degussa, DE 1 935 302, 1969 (W. Schwarze, W. Weigert).
57. Bayer, DE 1 178 423, 1963 (R. Heusch).
58. Bayer, DE 952 084, 1956 (E. Schmidt, R. Schnegg, F. Zaller, F. Moosmüller).
59. Schering, DE 1 068 250, 1957 (H. Werres).
60. G. Losse, H. Weddige, *Justus Liebigs Ann. Chem.* **636** (1960) 144.
61. Stauffer Chem., US 4 713 467, 1987 (J. E. Telschow, D. A. Bright).
62. Story Chemical, US 3 923 852, 1975 (A. G. Zeiler, H. Babad).
63. DD 43 996, 1965 (D. Martin, E. Beyer, H. Gross).
64. Bayer, DE 1 172 257, 1963 (E. Mühlbauer, E. Meisert).
65. M. S. Kharasch, E. M. May, F. R. Mayo, *J. Am. Chem. Soc.* **59** (1937) 1580.
66. IG Farbenindustrie, DE 705 561, 1941 (W. Flemming, H. Buchholz).
67. May & Baker, DE 2 129 960, 1982 (D. A. Eichler et al.).
68. Dow Chemical, US 2 757 190, 1956 (G. D. Jones, R. L. Zimmerman).
69. A. G. Williamson, *Analyst (London)* **77** (1952) 372.
70. M. Ottnad, N. A. Jenny, C.-H. Röder, *Anal. Methods Pestic. Plant Growth Regul.* **10** (1976) 563 – 573.
71. Degussa, DIN Sicherheitsdatenblatt "Methylisothiocyanat," Hanau, 1991.
72. Air Products and Chemicals, DE 2 100 057, 1971 (T. Cebalo, J. F. Alderman).
73. Bayer, DE 1 816 568, 1968 (C. Metzger, H. Hack).
74. Bayer, DE 1 770 467, 1968 (D. Rücker, C. Metzger, L. Eue, H. Hack).
75. Glaxo Group, GB 2 160 204, 1985 (J. F. Seager, R. Dansey).
76. Union Quimico Farmazeutika, ES 2 003 781, 1988 (I. Lopez Molina, A. D. Coto).
77. Smith Kline & French, US 3 894 151, 1975 (J. W. Black, M. E. Parsons).
78. Eli Lilly, US 4 452 778, 1984 (L. G. Brier).
79. Eli Lilly, US 3 641 021, 1972 (Ch. W. Ryan).
80. J. J. Morrell, M. E. Corden, *For. Prod. J.* **36** (1986) 26.
81. J. N. R. Ruddick: "Fumigant Movement in Canadian Wood Species," The International Research Group on Wood Preservation, *15th Annual Meeting,* Stockholm, Sweden 1984.
82. J. J. Morrell: "The Use of Fumigants for Controlling Decay of Wood: A Review of Their Efficacy and Safety," International Research Group on Wood Preservation, *20th Annual Meeting,* Lappeenranta, Finland, 22 – 26 May 1989.
83. J. J. Morrell: *Market Potential for Safer Fumigants for Wood Preservation,* Oregon State University, Corvallis 1986.
84. I. Leifertova, M. Lisa, J. Stanek, N. Hejtmankova, *Folia Pharm. (Prague)* **5** (1983) 43.
85. Nippon Soda, DE 1 930 540, 1969 (T. Noguchi et al.).
86. Industrieverband Agrar: *Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmittel: physikalisch-chemische und toxikologische Daten,* 2nd ed., BLV Verlag, München 1990, p. 284.
87. Kühn/Birett: Methylisothiocyanat M 45, *Merkblätter gefährliche Arbeitsstoffe,* ecomed Verlag, Landsberg 1992.
88. Neue Datenblätter für gefährliche Arbeitsstoffe nach der Gefahrstoffverordnung, Methylisothiocyanat Datenblatt 1032, Weka Verlag, Augsburg, 1992.
89. Kühn-Birett: Allylisothiocyanat A 106, *Merkblätter gefährliche Arbeitsstoffe,* ecomed Verlag, Landsberg, 1992.

Thiocyanates, Inorganic

THEO H. J. VAN HOEK, Akzo Nobel Chemicals, Köln, Federal Republic of Germany

1. **Introduction** 1
2. **Thiocyanates** 1
2.1. **Ammonium Thiocyanate** 1
2.2. **Sodium Thiocyanate** 4
2.3. **Potassium Thiocyanate** 6
2.4. **Alkaline Earth Thiocyanates** 6
2.5. **Other Thiocyanates** 7
3. **Thiocyanic Acid** 7
4. **Economic Aspects** 8
5. **Toxicology** 8
6. **References** 8

1. Introduction

Thiocyanates are the salts of thiocyanic acid, H−S−C≡N. The older name, rhodanides, is still used in many countries. Although HSCN, which is unimportant industrially, is readily decomposed, thiocyanates are stable both in solution and in the crystalline state. Some chemical reactions of thiocyanates are similar to those of halides. One of their characteristic reactions is the blood red coloration they give with the iron(III) ion. Pure thiocyanate solutions are titrated using the Volhard method in nitric acid against 0.1 N silver nitrate with iron alum as indicator.

The older production methods for thiocyanates were based on the use of coke oven gas, which used to be in plentiful supply. These very economical methods have become less important over recent decades because of higher quality requirements for the final product and the availability of suitable alternative processes.

Thiocyanates, in particular the sodium and ammonium salts, are mainly used as raw materials or auxiliaries in fiber production, for agricultural products, in photography, in the chemical industry, and in building. The realization that thiocyanates play an important role in many biochemical processes in animals and humans [1] has been increasingly exploited in the production of personal hygiene products and in the foodstuffs and pharmaceutical industries.

2. Thiocyanates

2.1. Ammonium Thiocyanate

Properties. Ammonium thiocyanate [*1762-95-4*], NH_4SCN, M_r 76.12, ϱ 1.305 g/cm^3, is a white, hygroscopic salt which forms stable monoclinic crystals below 90 – 92 °C [2], [3]. At higher temperatures it is converted to the rhombic form. The *mp* of the pure salt is 149 °C. If it is heated to above 100 °C it partially rearranges to the isomeric thiourea.

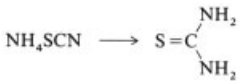

The degree of rearrangement depends on the temperature and the heating time. Lower temperatures and longer heating times give the highest proportion of thiourea [4]:

200 h at 122 °C	31.7 mol % $SC(NH_2)_2$
10 – 15 h at 150 °C	28.0 mol % $SC(NH_2)_2$

If the melt is further heated to 180 – 200 °C, guanidinium thiocyanate is formed [5].

$$2\,NH_4SCN \longrightarrow [H_2N{=}C(NH_2)_2]SCN + H_2S$$

Here, CS_2, NH_3, and – via the intermediate melamine – a compound with the formula $C_6H_3N_9$ are formed as byproducts.

Ammonium thiocyanate is readily soluble in many polar solvents. The solubilities in 100 g solvent are for liquid ammonia (25 °C) 312 g, for liquid sulfur dioxide (0 °C) 46.8 g, and for acetonitrile (18 °C) 7.5 g. Ammonium thiocyanate is readily soluble in ethanol and acetone.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_759

The density of aqueous NH_4SCN solutions (in kg/L) is as follows:

15 % NH_4SCN	1.033
30 % NH_4SCN	1.068
45 % NH_4SCN	1.103
60 % NH_4SCN	1.140

and the solubility at various temperatures (in wt %):

0 °C	55.0
30 °C	66.8
60 °C	77.8
90 °C	87.5

The boiling point of a 30 % solution in water is 105 °C, of a 50 % solution 110 °C, and of an 80 % solution 140 °C.

The enthalpy of solution in water is 23.86 kJ/mol. By mixing 100 g water with 130 g ammonium thiocyanate a temperature decrease of 18 °C can be achieved (cooling baths). As the salt of a weak base ammonium thiocyanate is weakly acidic in aqueous solution. The pH value of a 5 % solution is between 4.8 and 5.0.

Since small quantities of iron in ammonium thiocyanate which has been exposed to light can cause the typical red coloration, attention must be paid to the chemical resistance of vessels used for storage or processing. Aqueous ammonium thiocyanate solutions and crystals severely corrode mild and stainless steels up to and including the grade 1.4301. The stainless steels 1.4571 (UNS S31635) and 1.4404 (UNS S31603) have limited resistance. Only austenitic steels with a molybdenum content of over 4 % or a high nickel content, such as 1.4449/1.4439 (UNS S31726), 1.4539 (UNS N08904), 2.4858 (UNS N08825) and 1.4563 (UNS N08028), have corrosion rates of < 10 µm per year up to 80 °C in concentrated ammonium thiocyanate solution.

Glass, titanium, and rubber or baked enamel linings have also proved to be suitable materials. Plastics, such as polypropylene, poly(vinyl chloride), polyvinylidene fluoride, and PTFE, and epoxy, polyester and vinyl ester resins also show good chemical resistance towards ammonium thiocyanate. Tests have shown that small changes in the chemical compositions of the solutions arising, for example, from differences in origin or pH cause changes in the corrosion behavior with metals.

Production. Ammonium thiocyanate is obtained from CS_2 and NH_3 or from the reaction of hydrogen cyanide or cyanides with sulfur and/or polysulfides.

Production from Ammonia and Carbon Disulfide. This continuous, pressure-free process is carried out in two steps [6]. In the first step CS_2 and ammonia gas are reacted in the presence of water over activated charcoal at ca. 50 °C in an interface reactor to give ammonium dithiocarbamate.

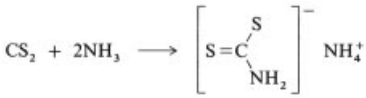

The latter is decomposed in the second step at 95 – 100 °C on activated charcoal.

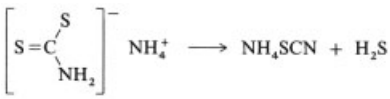

If the NH_3 and CS_2 used are of high purity, the 35 – 45 % ammonium thiocyanate solution formed can be marketed after concentration without further purification. The crystals, which are produced by vacuum evaporation and centrifugation, are of higher purity. In this process, apart from H_2S and excess CS_2, no other waste substances are formed. After CS_2 and H_2S have been separated, the former can be recycled and the latter can be used, for example, for sulfur production in the Claus process. There is also a one-step variant of this process in which the reaction of NH_3 with CS_2 and decomposition take place at ca. 115 °C and 5 bar [7].

Production from Hydrogen Cyanide contained in Coke Oven Gas. Cooled coke oven gas, which has been freed from tar, contains HCN, NH_3, and H_2S. NH_3 and H_2S are reacted with an aqueous slurry of sulfur in a circulating gas scrubber forming $(NH_4)_2S$, which dissolves the elemental sulfur to form ammonium polysulfide. At the same time the HCN combines with the NH_3 dissolving in the washing liquid to give NH_4CN, which then reacts with the polysulfide to give ammonium thiocyanate and ammonium sulfide.

$$(NH_4)_2S_2 + NH_4CN \longrightarrow NH_4SCN + (NH_4)_2S$$

The ammonium sulfide is distilled off. A 25 – 50 % ammonium thiocyanate solution remains, which, after purification, is further processed as in the other processes [5], [8], [9].

Production from Hydrogen Cyanide, Aqueous Ammonia, and Sulfur/Ammonium Polysulfide. Hydrogen cyanide is neutralized with aqueous ammonia, and the ammonium cyanide formed is treated with elemental sulfur and/or ammonium polysulfide with warming to give ammonium thiocyanate [10].

Commercial Forms and Specifications. Ammonium thiocyanate is generally packed in 25 kg (or 50 lb) sacks with polyethylene liners. It is also transported as a 50 or 55 % solution in tank cars or containers of 1000 kg capacity or plastic drums of 200 – 250 kg capacity. The bulk density of the crystalline product is600 – 700 g/L. Ammonium thiocyanate is not a hazardous substance under the terms of the transport regulations and, because the corrosion rates measured on aluminum and steel (St 37) are less than 6.25 mm/a, aqueous ammonium thiocyanate solutions also need not be declared as hazardous substances for air and sea transport. Specifications for ammonium thiocyanate solutions and crystals are listed in Table 1.

Table 1. Typical specifications for NH_4SCN

Property	Solution (e.g., 50 % in water)	Crystals
Appearance	colorless, clear liquid	crystalline, white
NH_4SCN content	49 – 51 %	min. 99.0 % in dry mass
Water content		max. 1.0 %
pH	4.5 – 6.5	4.0 – 6.0 (5 % solution)
Iron	max. 10 mg/kg	max. 2.0 mg/kg
Iodine consumption	max. 20 mg I/100 g	max. 10 mg I/100 g
Sulfides	max. 1500 mg/kg	max. 800 mg/kg
Chloride		max. 50 mg/kg
Sulfate		max. 200 mg/kg
Density (20 °C)	1.113 – 1.118 g/ml	
APHA number	max. 80	

Uses. Most of the ammonium thiocyanate is produced as an intermediate in the production of other inorganic thiocyanates. It is also used as the starting material for compounds such as aminothiazoles [11].

Large quantities of ammonium thiocyanate are used in agriculture. It is used to combat worms in cattle and, as a mixture with aminotriazole, reinforces the activity of the latter as a broad-spectrum herbicide. Ammonium thiocyanate is also used as a raw material for the production of other herbicides, such as diallates, triallates and tricyclazoles. It is also used as a raw material for fungicides, such as methyl isothiocyanate.

Ammonium thiocyanate is extensively used in photography as an accelerator with sodium thiosulfate in fixing baths, where it removes undeveloped silver (Ag^+). Addition of ammonium thiocyanate to electrolytic cells in the production of ammonium peroxodisulfate, $(NH_4)_2S_2O_8$considerably increases the yield.

Ammonium thiocyanate is used as a solvent for cellulose [12–14], which can also be spun [15]. In the textile industry ammonium thiocyanate is widely used to improve desired properties:

1) In dyeing [16] and improving the quality [17], [18] of cotton
2) For improving the resistance of silk to creasing, light, and wear and tear [19], [20]
3) For lowering the absorption capacity of marks on nylon floor coverings [21]
4) For softening jute fibers [22]
5) For the heat-stabilization of polyester fibers [23]
6) In printing [24] or dyeing [25] fabrics

In the steel industry ammonium thiocyanate is used for coating [26], [27] and as a corrosion inhibitor against acidic gases [28]. It is used as an additive for silver [29] and tin [30] plating baths and for etching aluminum [31].

Because of the neutron-absorbing properties of hafnium, only hafnium-free zirconium can be used in nuclear technology. Following their extraction, Zr and Hf are dissolved in aqueous ammonium thiocyanate so that the $Hf(SCN)_4$ complex can subsequently be extracted with an organic solvent.

Ammonium thiocyanate can also be used as a component in the following applications:

1) For energy storage in combination with heat pumps [32–35]
2) In vulcanization [36], [37]
3) For the production of fire-extinguishing agents [38]

4) In absorbents for ammonia [39] or steam [40]

2.2. Sodium Thiocyanate

Sodium thiocyanate [*540-72-7*], NaSCN, M_r 81.08, and its dihydrate [*1702-40-5*], $NaSCN \cdot 2\,H_2O$, M_r 117.11, form white crystals. The hygroscopic NaSCN crystal is orthorhombic [41]. An α- and a β-form of monoclinic $NaSCN \cdot 2\,H_2O$ are known [42]. The density of $NaSCN \cdot 2\,H_2O$ is 1.56 g/cm^3. On recrystallization the crystals have a severe tendency to agglomerate if residual moisture content is above ca. 0.3 – 0.5 %, particularly in the case of high-purity NaSCN. Anhydrous NaSCN melts at 287 °C and the dihydrate dissolves in its own water of crystallization at 30.4 °C.

The solubility of NaSCN in water (in wt %) at various temperatures is:

30.4 °C	63.2
60 °C	65.2
90 °C	67.3

and the density (in kg/L):

40 wt % NaSCN	1.228
50 wt % NaSCN	1.298
55 wt % NaSCN	1.333

The solubility increases sharply up to the triple point (30.4 °C) and thereafter only gradually. The *bp* of a 20 % solution in water is 107 °C, of a 40 % solution 115 °C, and of an 80 % solution 140 °C. The solubility of sodium thiocyanate in methanol increases from 26 – 35 % over the temperature range 16 – 52.5 °C, that in ethanol from 15.5 – 19.6 % at 19 – 71 °C, and that in acetone from 6.4 – 17.5 % at 18.8 – 56 °C. Sodium thiocyanate is less corrosive than ammonium thiocyanate. Besides the materials named under ammonium thiocyanate (see Section 2.1), aluminum and 1.4571 steel (UNS S31635) are also resistant.

Production. Sodium thiocyanate can be obtained by using the process described for ammonium thiocyanate from CS_2 and NH_3, but substituting 1 mol NH_3 with 1 mol NaOH (50 % solution) [43].

$$CS_2 + 3\,NH_3 + NaOH \longrightarrow NaSCN + (NH_4)_2S + H_2O$$

If ammonium thiocyanate is available and it is possible to use the ammonia formed, sodium thiocyanate can be produced by treating ammonium thiocyanate with sodium hydroxide. This process can also be carried out continuously in a column.

$$NH_4SCN + NaOH \longrightarrow NaSCN + NH_3 + H_2O$$

In another process sodium cyanide formed by neutralizing HCN with NaOH, is treated with rhombic sulfur and/or sodium polysulfide to give sodium thiocyanate [44], [45]. The sulfur can also be added as a suspension in ethanol [46].

Sodium thiocyanate is also formed when wastewater from desulfurization plants for coke-oven gas is treated with NH_3, NaOH, and $Ca(OH)_2$ [47].

Like ammonium thiocyanate, sodium thiocyanate is formed as an aqueous solution which may be evaporated after purification to higher concentrations or to give crystals. If the evaporation to form crystals takes place above 30.4 °C the anhydrous salt is formed, and below this temperature the dihydrate.

Table 2. Typical specifications for NaSCN

Property	Solution (e.g., 50 % in water)	anhydrous crystals
Appearance	colorless, clear liquid	crystalline, white
NaSCN content	49 – 51 %	min. 99.0 % in dry mass
Water content		max. 3.0 %
pH	7.5 – 9.5	5.5 – 7.5 (5 % solution)
Iron	max. 1.0 mg/kg	max. 2.0 mg/kg
Iodine consumption		max. 20 mg I/100 g
Sulfides	max. 80 mg/kg	max. 15 mg/kg
Chloride		max. 100 mg/kg
Sulfate		max. 200 mg/kg
Ammonium (as NH_3)	max. 300 mg/kg	max. 200 mg/kg
Density (20 °C)	1.291 – 1.305 g/ml	
APHA number	max. 40	

Commercial Forms and Specifications. Sodium thiocyanate is packed as the anhydrous salt and as the dihydrate in 25 (or 50) kg sacks with polyethylene liners. The 50 or 55 % solutions are dispatched in tank cars or containers of 1000 kg capacity or in 250 kg plastic drums. Sodium thiocyanate is not classified as a hazardous substance in transport regulations, and

because the corrosion rates measured on aluminum and steel (St 37) are less than 6.25 mm/a, aqueous sodium thiocyanate solutions need not be declared as hazardous substances for air and sea transport. For the specifications for sodium thiocyanate solutions and crystal form, see Table 2.

Uses. Much of the sodium thiocyanate produced is used in polyacrylonitrile fiber production. Since the decomposition of polyacrylonitrile begins below its melting point, spinning the melt is not possible. However, polyacrylonitrile can be processed to give fibers when dissolved in aqueous sodium thiocyanate solution in the presence of additives. In a similar process polyacrylonitrile is dissolved in an organic solvent.

Sodium thiocyanate is used in the chemical industry as raw material for the production of organic isothiocyanates [48], such as methyl isothiocyanate, other monoalkyl isothiocyanates, or thiosemicarbazides. Some of these products are used as pesticides. Fungicides, such as methylene bisthiocyanate (MBT) and 2-thiocyanomethylthiobenzothiazole (TCMTB) are also produced from sodium thiocyanate. Like ammonium thiocyanate, sodium thiocyanate is used in the production of sodium peroxodisulfate ($Na_2S_2O_8$). In many countries sodium thiocyanate has replaced calcium chloride as the additive for reducing the hardening time of concrete. However, as the corrosive action of thiocyanate on the iron in reinforced concrete cannot be fully explained, there are restrictions in the case of concrete with high-quality requirements and prestressed concrete.

Alkali metal thiocyanates are used in the production of solid cell batteries [49], [50], and in the production of dyes for liquid crystal indicators [51].

Because of complex formation with Fe^{3+} ions, thiocyanates are used as laboratory reagents. The blood red color of $Fe(SCN)_3$ in water is used in oil fields to detect the underground connections between different drillholes. Sodium thiocyanate is also mixed with oil field explosives as a stabilizer. Furthermore, thiocyanates are used in the production of ink for ink-jet printers [52], [53].

In the pharmaceutical industry thiocyanate has been used as an antihypertensive, but the high concentrations required gave rise to side effects. Thiocyanates are used as raw materials for antithrombolytics [54] and antitumor drugs [55], and as components of hair growth preparations. In hair dyeing, thiocyanates reduce the duration of the treatment [56], [57].

In recent decades sodium thiocyanate has increased in importance as a preservative. In the lacto-peroxidase system thiocyanate, together with peroxides, forms hypothiocyanite ($OSCN^-$), which exhibits pronounced antibacterial activity. Fresh cows' milk is protected from immediate degradation by the natural presence of thiocyanate (4 – 6 ppm) and peroxides. In countries where cooling the milk before processing cannot be guaranteed, a further ca. 10 ppm sodium thiocyanate can be added as a preservative together with the oxidizing agent sodium carbonate peroxohydrate ($2\,Na_2CO_3 \cdot 3\,H_2O_2$) [58], [59]. As this method seems to be effective and free of side effects, it is being increasingly used as a means of preservation in the cosmetics industry.

2.3. Potassium Thiocyanate

Potassium thiocyanate [*333-20-0*], KSCN, M_r 97.18, forms colorless, prismatic crystals with *mp* 175 °C [60], [61]. At 20 °C it has a density of 1.89 g/cm^3. The solubility in water at 20 °C is 20.8 g/100 g. It is readily soluble in ethanol.

Potassium thiocyanate is produced by the same processes as sodium thiocyanate, but using KOH instead of NaOH. Potassium thiocyanate is used for some of the same purposes as sodium thiocyanate, but to a lesser extent because of its higher price. Potassium thiocyanate is particularly important as an intermediate and a final product in the pharmaceuticals industry. It is added to toothpastes and mouthwashes to approximately double the quantity of thiocyanate which is naturally present in saliva. In the lacto-peroxidase system it increases the antibacterial activity in combination with peroxides [62–64]. Halides behave similarly, but, unlike thiocyanates, they also cause undesired side effects.

Although the properties of potassium and sodium thiocyanates are comparable, the potassium salt is more corrosive. At the boiling point potassium thiocyanate solutions cause pitting in aluminum and attack 1.4571 stainless steel

(UNS S31635) to the same extent as ammonium thiocyanate. In laboratory tests only 1.4577 and 1.4438 steels (UNS S31703) have corrosion rates of < 10 µm per year. With mild steels potassium and sodium thiocyanates exhibit comparable corrosion rates, which are clearly exceeded by that of ammonium thiocyanate.

2.4. Alkaline Earth Thiocyanates

Of the known alkaline earth thiocyanates only calcium and barium thiocyanates are important industrially. Some properties of these two compounds are given in Table 3.

Table 3. Properties of calcium thiocyanate tetrahydrate and barium thiocyanate trihydrate

Property	$Ca(CSN)_2 \cdot 4H_2O$	$Ba(SCN)_2 \cdot 3H_2O$
M_r	228.31	307.58
mp, °C	ca. 59	ca. 70
Solubility, g/100 g H_2O	ca. 1000 g	ca. 240 g

Besides $Ca(SCN)_2 \cdot 4H_2O$, di- and trihydrates are known.

Calcium and barium thiocyanates are obtained by treating aqueous ammonium thiocyanate solutions with the corresponding alkaline earth hydroxides.

$$M(OH)_2 + 2\,NH_4SCN \longrightarrow M(SCN)_2 + 2\,NH_4OH$$

After the ammonia has been distilled off, the salt solutions are evaporated under reduced pressure, and the salt crystals are obtained by centrifugation. They then have an assay of 97 – 98 % based on the hydrate. The yield is almost quantitative.

Calcium and barium thiocyanates [and also strontium thiocyanate $Sr(SCN)_2$] can also be produced from NH_4SCN, NH_3, and CS_2 with the corresponding metal hydroxide [43]. Calcium thiocyanate can also be obtained from the wastewater produced in the desulfurization of coke-oven gas, together with NH_3 and $Ca(OH)_2$ [65].

Calcium and barium thiocyanates are used in the chemical and pharmaceutical industries. Calcium thiocyanate is used for the production of special materials in the building industry. Barium thiocyanate is used in printing and dyeing textiles.

2.5. Other Thiocyanates

Of the large number of other known thiocyanates, only copper(I), lead(II), and aluminum thiocyanates are important industrially.

Copper(I) thiocyanate [*18223-42-2*], CuSCN, M_r 121.62, is a yellowish-white powder with *mp* 1084 °C and ϱ 2.843 g/cm^3 at 20 °C. It is sparingly soluble in water (5 mg/L). It is soluble in alkali thiocyanate solutions and ether, but not in ethanol. It forms a deep blue copper – ammonia complex in aqueous ammonia.

The industrial production of CuSCN starts from ammonium or sodium thiocyanate. Aqueous solutions of the thiocyanate are warmed with copper sulfate solution. SO_2 or alkali metal sulfites and hydrogensulfites can be used to reduce the copper(II) thiocyanate formed initially.

$$2\,NaSCN + 2\,CuSO_4 + Na_2SO_3 + H_2O \longrightarrow 2\,CuSCN + 2\,NaHSO_4 + Na_2SO_4$$

The CuSCN, which is precipitated quantitatively, is separated from the dissolved sodium salts by decantation and centrifugation, washed until sulfate-free, and dried [66], [67].

CuSCN is used in the production of antifouling agents for protecting the underwater surfaces of ships against vegetation. Anti-fouling paints based on tributyltin (TBT) are only allowed to be used for ships more than 25 m long.

Lead(II) thiocyanate [*592-87-0*], $Pb(SCN)_2$, M_r 323.38, forms white to pale yellowish, monoclinic crystals with *mp* 190 – 195 °C (decomp.) and ϱ 3.82 g/cm^3 at 20 °C. The solubility in water at 20 °C is ca. 0.05 g/100 g, and increases on warming. Lead(II) thiocyanate dissolves in nitric acid, aqueous ammonia, and alkali metal thiocyanate solutions.

Lead(II) thiocyanate is produced by treating water-soluble lead salts (e.g., lead nitrate) with alkali metal thiocyanate solutions [68].

Lead(II) hydroxythiocyanate, Pb(OH)(SCN), can be obtained by boiling a slurry of $Pb(SCN)_2$ in water or by the action of basic lead(II) acetate hydroxide on $Pb(SCN)_2$ [69].

Lead(II) thiocyanate is used for producing fuses in the explosives industry and, together with antimony sulfide and potassium chlorate, for producing phosphorus-free matches.

Aluminum thiocyanate, $Al(SCN)_3$, M_r 201.21, can be produced as a ca. 30 % aqueous solution by treating an aluminum sulfate solution with barium or calcium thiocyanate. After separation of the alkaline earth sulfate, the $Al(SCN)_3$ solution can only be evaporated under reduced pressure to a concentration of ca. 30 %. Above this concentration the $Al(SCN)_3$ begins to decompose. $Al(SCN)_3$ solutions are used as mordants in alizarin printing in the textile industry.

3. Thiocyanic Acid

Properties. Thiocyanic acid, formerly also known as hydrogen rhodanide, HSCN, M_r 59.09, *mp* 5 °C, is a colorless, very volatile liquid. It is readily soluble in water and behaves as a strong mineral acid. In very dilute (max. 5 %) aqueous solutions thiocyanic acid is relatively stable up to 20 °C. At higher temperatures or at higher concentrations it polymerizes with loss of hydrogen cyanide, forming 5-imino-1,2,4-dithiazolidine-2-thione.

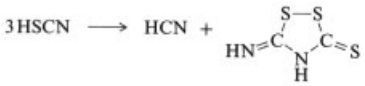

Production. Anhydrous thiocyanic acid is obtained as white crystals by passing dry H_2S into dry $Hg(CN)_2$. Pure, gaseous thiocyanic acid can be produced by heating a mixture of KSCN and $KHSO_4$ in H_2 under reduced pressure (ca. 60 mbar) [69]. Aqueous solutions of thiocyanic acid are obtained by decomposing a 25 % ammonium thiocyanate solution with dilute sulfuric acid.

$$NH_4SCN + H_2SO_4 \longrightarrow HSCN + NH_4HSO_4$$

The yield is almost quantitative when stoichiometric quantities of the starting materials are used. The thiocyanic acid liberated is distilled at ca. 30 mbar and collected in water. The industrial and economic importance of free thiocyanic acid is low.

4. Economic Aspects

The annual worldwide demand for thiocyanates is ca. 15 000 t, based on 100 % crystalline product. As industry mainly uses sodium and ammonium thiocyanates, these account for by far the greatest part of the overall requirement. About 20 % of the thiocyanate is used in North America and 35 % in Western Europe.

Around 30 – 40 % of the thiocyanate is used in fiber production. The second largest consumer (ca. 20 %) is the production of agricultural auxiliaries. About 15 % is used in photography. Of the many other uses, application as a reagent for extracting metals such as copper, molybdenum, nickel, lead, zinc, and zirconium from ores is important.

The world market for thiocyanates appears to be relatively stable to slightly recessive. Decreases in the use in agricultural products in the early 1990s were only partially compensated for by a slight increase in the requirement for the photography industry.

The largest producers in order of importance are Western Europe (Germany and the United Kingdom), South East Asia (Japan and China), and the United States.

5. Toxicology

Thiocyanates are only slightly toxic. Most inorganic thiocyanates have been assigned to toxicity classes 3 – 5 according to the Swiss toxicity law. In animal experiments the lethal dose (LD_{50}) has been found to be 50 – 15 000 mg/kg [70]. Ammonium thiocyanate with an LD_{50} of 500 mg/kg was assigned to toxicity class 3, and sodium thiocyanate with LD_{50} 764 mg/kg to class 4. The NOEL for rainbow trout for NH_4SCN is 100 mg/L, and the EC_{10} value for bacteria (*Ps. putia*) is 8 g/L. Although resorbed thiocyanate is biologically degradable, degradation can become problematic at high concentrations ($>$ 36 mg SCN^-/L H_2O).

On handling thiocyanates skin contact should be avoided, as prolonged contact can lead to skin irritation and rashes, and thiocyanates are readily resorbed.

The action of strong acids on thiocyanates can liberate toxic hydrogen cyanide. Ammoni-

um and sodium thiocyanates have been assigned to water hazard class 1 (WGK 1) in Germany (September 1994).

6. References

1. W. Weuffen: *Medizinische und biologische Bedeutung der Thiocyanate (Rhodanide),* VEB Verlag Volk und Gesundheit, Berlin 1982.
2. J. W. Bats, P. Coppens, *Acta Crystallogr. Sect. B: Struct. Crystallogr. Cryst. Chem.* **B 33** (1977) 1542–1548.
3. V. E. Zavodnik, Z. V. Zvonkova, G. S. Zhdanov, E. G. Mirevich, *Kristallografiya* **17** (1972) 107–110.
4. W. Klempt, *Chem. Tech. (Leipzig)* **15** (1963) 1.
5. W. Klempt, *Brennst. Chem.* **30** (1949) 149, 155.
6. Enka Glanzstoff, DE 1 297 088, 1967 (H. Maegerlein, H. Rupp, G. Meyer).
7. Akzo, DE 3 508 404, 1986 (G. Meyer, G. Sudheimer, H. Zengel, H. Grothaus).
8. Sumikin Coke and Chemicals, JP 87-113 753, 1987 (T. Sato, K. Takeda).
9. Scientific Research Institute of Coal Chemistry, SU 87-4 171 977, 1987 (V. Markov, T. Batyeva, V. Melikentsova).
10. Sumikinkako, JP 79-164 355, 1979.
11. BASF, DE 3 227 329, 1984 (G. Seybold).
12. J. Cho, S. Hudson, J. Cuculo, *J. Polym. Sci., Polym. Lett. Ed.* **27** (1989) no. 8, 1699–1719.
13. B. Lukanoff, H. Schleicher, B. Philipp, *Cellul. Chem. Technol.* **17** (1983) no. 6, 593–600.
14. K. Yang, M. Theil, J. Cuculo, *ACS Symp. Ser.* **384** (1989) 156–183.
15. North Carolina State University, US 4 750 939, 1988 (J. Cuculo, H. Theil, K. Yang, Y. Chen).
16. US Dept. of Agriculture, US 171 626, 1981 (E. Blanchard).
17. T. Calamari, D. Thibodeaux, *Text. Chem. Color.* **20** (1988) no. 12, 13–16.
18. H. Bober, J. Cuculo, P. Tucker, *J. Polym. Sci. Polym. Chem. Ed.* **25** (1987) no. 8, 2025–2032.
19. H. Ban, JP 01 020 379, 1989.
20. Kanebo, DE 3 814 450, 1988 (T. Fuse, A. Yamamoto, J. Sano).
21. Allied Signal, WO 8 902 949, 1989 (D. Hangey et al.).
22. Indian Jute Ind. Res., IN 152 988, 1984 (P. Bhattacharjee, U. Ghosh, M. Masumdar, A. Ray).
23. Allied, US 4 348 314, 1982 (S. Lazarus, R. Lofquist).
24. BASF, DE 3 401 500, 1985 (A. Blum, N. Grund, S. Schreiner, G. Treiber, N. Zimmermann).
25. Osrodek Badawczo, PL 116 165, 1982 (M. Gralinski et al.).
26. Nippon Steel, DE 3 902 457, 1989 (K. Saito, Y. Miyauchi, T. Murata, Y. Shindo).
27. Wiener Brückenbau- und Eisenkonstr., AT 364 218, 1981 (H. Willmitzer).
28. Dow Chem., US 4 446 119, 1984 (M. DuPart, B. Oakes, D. Cringle).
29. Schering, US 3 915 718, 1975 (R. Ludwig, K. Fuchs).
30. Kemiska, SE 428 805, 1983 (I. Fjell).
31. Hunt, NL 7 303 505, 1974 (A. Philip).
32. J. Ruiter, Rep. ETN-88-91614, Order no. N88-17108, NTIS Fr (1987) 156.
33. Y. Liu, *Proc. Int. Soc. Energy Convers. 23rd* **4** (1988) 249–252.
34. M. Jeday, P. Le Goff, *Int. J. Refrig.* **11** (1988) no. 3, 164–172.
35. I. Fujiwara, M. Ebe, M. Sato, *Kagaku Kogaku* **51** (1987) no. 4, 302–304.
36. VEB Transportgummi, DD 216 944, 1985 (J. Hopfe, A. Nechwatal).
37. Degussa, US 3 947 436, 1976 (G. Rocktaeschel et al.).
38. Amalgamated Chem., US 4 816 186, 1989 (F. Acitelli).
39. D. Erickson, US 4 784 783, 1988.
40. D. Erickson, WO 8 810 290, 1988.
41. J. W. Bats, P. Coppens, Å. Kvick, *Acta Crystallogr. Sect B: Struct. Crystallogr. Cryst. Chem.* **B 33** (1977) 1534–1542.
42. K. Mereitner, A. Preisinger, *Z. Kristallogr.* **169** (1984) 95–107.
43. Enka Glanzstoff, DE 67-1 592 339, 1967 (H. Rupp, H. Maegerlein, G. Meyer).
44. Bergwerksverband, DE 73-2 329 894, 1973 (G. Huck, N. Schaefer).
45. Mitsui Toatsu Chem., JP 76-15 771, 1976 (N. Koike, Y. Hatazaki, M. Omura).
46. Degussa, DE 85-3 514 408, 1985 (E. Bilger).
47. Sumikin Coke and Chem., JP 87-113 752, 1987 (T. Sato, K. Takeda).
48. Am. Cyanamid, US 4 778 921, 1988 (M. Lewellyn, S. Wang, P. Strydom).
49. Exxon Research and Engineering, US 4 190 706, 1980 (B. Rao, B. Silbernagel).
50. Inst. of Phys. and Chem. Research, JP 61 256 573, 1986 (A. Yamada, J. Shigehara, Y. Kurata, K. Yonahara).

51. Casio Computer, JP 01 146 960, 1989 (H. Aoki, K. Kodera, M. Hikasa, T. Okugawa).
52. A. Dick, GB 1 543 703, 1979 (K. Hwang, D. Zabiak).
53. Dainippon Toryo, Matsushita Electric, JP 59 053 564, 1984.
54. Wellcome Foundation, EP 304 311, 1989 (H. Berger, S. Pizzo).
55. Engelhard, EP 98 133, 1984 (A. Amundsen, E. Stern).
56. Kintex, NL 7 409 228, 1976.
57. Bucaria, US 3 823 231, 1974.
58. L. Björck, O. Claesson, W. Schulthess, *Milchwissenschaft* **34** (1979) 726.
59. H. Korhonen, *World Anim. Rev.* **35** (1980) 23.
60. S. Yamamoto, M. Sakuno, Y. Shinnaka, *J. Phys. Soc. Jpn.* **56** (1987) 4393 – 4399.
61. D. Cookson, M. Elcombe, T. Finlayson, *Mat. Sci. Forum* **27** (1988) 113 – 116.
62. O. Poulson, WO 880 2600, 1988.
63. State University of New York, US 461 848, 1986 (J. Pollock, T. McNamara).
64. Laclede Professional Prod., CA 1 167 381, 1984 (M. Pellico, R. Montgomery).
65. Sumikon Coke and Chem., JP 87-113 751, 1987 (T. Sato, K. Takeda).
66. Hokko Chem., JP 58 036 921, 1983.
67. I. Krut'ko, *Koks Khim.* 1988, no. 3, 31 – 32.
68. Du Pont, US 1 977 440, 1935.
69. H. Williams: *Cyanogen Compounds,* 2nd ed., Arnold, London 1948, pp. 257 – 316.
70. L. Roth, M. Daunderer: *Giftliste,* Verlag Moderne Industrie, München 1979.

Thioethers → Thiols and Organic Sulfides
Thioglycolic Acid → Mercaptoacetic Acid and Derivatives
Thioindigo → Indigo and Indigo Colorants

Thiols and Organic Sulfides

Kathrin-Maria Roy, Langenfeld, Federal Republic of Germany

1. Aliphatic Thiols, Sulfides, Disulfides, and Polysulfides . . . 2
1.1. Introduction . . . 2
1.2. Properties of Aliphatic Thiols . . . 2
1.2.1. Physical Properties . . . 2
1.2.2. Chemical Properties . . . 3
1.3. Production of Aliphatic Thiols . . . 4
1.3.1. From Alcohols . . . 4
1.3.2. From Alkenes . . . 5
1.3.3. Other Production Methods . . . 6
1.4. Construction Materials, Storage, and Transport . . . 6
1.5. Uses of Aliphatic Thiols . . . 7
1.6. Alkyl Sulfides, Disulfides, and Polysulfides . . . 7
1.6.1. Physical and Chemical Properties . . . 7
1.6.2. Production of Alkyl Sulfides . . . 8
1.6.3. Production of Alkyl Disulfides . . . 9
1.6.4. Production of Alkyl Polysulfides . . . 10
1.6.5. Uses of Alkyl Sulfides, Disulfides, and Polysulfides . . . 10
2. Aromatic Thiols, Sulfides, Disulfides, and Polysulfides . . . 11
2.1. Physical and Chemical Properties . . . 11
2.2. Production of Aromatic Thiols . . . 11
2.2.1. By Reduction . . . 11
2.2.2. By Thiolation . . . 12
2.3. Uses of Aromatic Thiols . . . 13
2.4. Production of Aryl Sulfides, Disulfides, and Polysulfides . . . 13
2.5. Uses of Aryl Sulfides, Disulfides, and Polysulfides . . . 13
3. Heterocyclic Thiols, Sulfides, and Disulfides . . . 13
3.1. Introduction . . . 13
3.2. Physical and Chemical Properties . . . 14
3.3. Production . . . 15
3.3.1. Heterocyclic Thiols by Substitution . . . 16
3.3.2. Heterocyclic Thiols by Cyclization . . . 16
3.3.3. Heterocyclic Sulfides and Disulfides . . . 17
3.4. Uses of Heterocyclic Thiols, Sulfides, and Disulfides . . . 18
3.4.1. Pharmaceuticals . . . 18
3.4.2. Agricultural Chemicals . . . 19
3.4.3. Other Uses . . . 19
4. Toxicology . . . 19
5. References . . . 21

The author would like to thank Bayer AG for undertaking the literature research for this article

1. Aliphatic Thiols, Sulfides, Disulfides, and Polysulfides

1.1. Introduction

Aliphatic thiols, R – SH, were formerly known as mercaptans because of their affinity for mercury compounds (Latin: *corpus mercurium aptans*). They can be regarded as organic derivatives of hydrogen sulfide or thio analogs of alcohols.

According to IUPAC nomenclature the names of aliphatic thiols are constructed by adding the ending "thiol" to the name of the corresponding alkane (e.g., ethanethiol). If substituents with higher priorities are present, the prefix "mercapto" is used instead (e.g., mercaptoacetic acid).

Thiols are found in a wide range of natural materials. Low molecular mass thiols are responsible for the aromas of various foods, such as cheese, milk, coffee, cabbage, and bread. They are probably liberated during processing through enzymatic cleavage of sulfides, disulfides, and other sulfur-containing substances [12]. Thiols can also be detected in a number of plants [13]. Methanethiol occurs in radishes and various alliums, such as onions, leeks, and garlic. The characteristic odor of fresh grapefruit juice is due to the presence of 1-*p*-menthene-8-thiol in high dilution [14]. Asparagus contains 3-mercapto- and 3,3′-dimercaptoisobutyric acids [15]. Butanethiol and 2-methylbutanethiol have been found in the defense secretions of skunks [16]. Low molecular mass alkanethiols are formed in the degradation of biological material during putrefaction [17], and consequently

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_767

are found in varying quantities in most crude petroleum, depending on its origin [18].

L-Cysteine, one of the natural amino acids found in proteins and a substance that participates in an important reversible redox system with the disulfide cystine, is of considerable biochemical interest, as is coenzyme A (CoA−SH), whose acetyl derivative (CoA−S−Ac) is a key participant in the metabolism of carbohydrates, fats, and proteins.

In 1834 W. C. Zeise achieved the first thiol synthesis by preparing ethanethiol from barium hydrogensulfide and calcium ethyl sulfate [19].

1.2. Properties of Aliphatic Thiols

1.2.1. Physical Properties

With the exception of methanethiol, which is gaseous at room temperature, aliphatic thiols with up to 16 carbon atoms are colorless or yellowish liquids. They are characterized by extremely offensive odors, perceptible even at very high dilution. The odor threshold for ethanethiol is less than 0.001 ppm, and that for the *S*-enantiomer of 1-*p*-menthene-8-thiol is only 2×10^{-8} ppm [13]. The odor sensitivity is highly concentration dependent, however, particularly for the lower thiols. At very low concentrations the C_1 to C_4 thiols have an odor similar to that of coal gas. That of other low molecular mass thiols has been described as pleasantly fresh or fruity, but at higher concentrations these compounds have extremely nauseous odors.

The bond energy of the S−H bond (339 kJ/mol) is lower than that of the O−H bond (462 kJ/mol), so hydrogen bonding to sulfur is also much weaker than to oxygen. This in turn means that thiols are less highly associated than the corresponding alcohols, as manifested in higher volatilities and lower boiling points for simple aliphatic thiols compared with their oxygen counterparts. This boiling point difference diminishes as the chain length increases, however, and is eventually reversed. Thus, from 1-octanethiol onward the straight-chain thiols have higher boiling points than the corresponding alcohols (Fig. 1).

The physical properties of important alkanethiols are summarized in Table 1. The high flammability of thiols containing up to six carbon atoms is particularly noteworthy. All the corresponding flash points lie below room temperature.

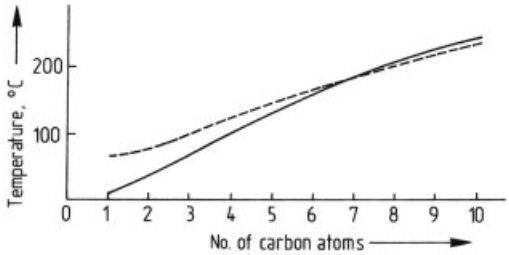

Figure 1. Comparison of boiling points for *n*-alcohols and *n*-thiols containing 1 – 10 carbon atoms
—- RSH – – – – ROH

Unlike their oxygen analogs, aliphatic thiols are only slightly soluble in water, but they are readily soluble in hydrocarbons, ethers, and alcohols. Ethane-, propane-, and butanethiols form azeotropes with aliphatic hydrocarbons, but not with benzene or toluene.

Aliphatic thiols are only weakly acidic, but they are significantly more acidic than the corresponding alcohols, which explains the ready solubility of low molecular mass alkanethiols in dilute base. However, the solubility decreases sharply with increasing molecular weight.

$$R-SH + Na^+OH^- \longrightarrow R-S^-Na^+ + H_2O$$

The resulting thiolate salts, RS^-Na^+, are less basic but more strongly nucleophilic than the equivalent alkoxide salts.

1.2.2. Chemical Properties

Salt Formation. Aliphatic thiols form thiolates (formerly known as mercaptides) with numerous metal cations. Many of these salts are sparingly soluble, so their formation can be used for the qualitative and quantitative determination of thiols. Many mercury thiolates, $Hg(SR)_2$, have been used for the identification of thiols because of their defined melting points. Silver thiolates have also been used for the qualitative and quantitative detection of thiols. The formation of lead thiolate constitutes the basis for the "doctor test," a method of determining the thiol content of crude petroleum. Heavy metals, in particular mercury and cadmium, can be removed from

Table 1. Physical data for selected aliphatic thiols

Thiol	CAS registry no.	M_r	*bp*, °C/kPa	d_4^{20}	n_D^{20}	Flash point, °C
CH_3-SH	[*74-93-1*]	48.11	6.2	0.8665		
CH_3CH_2-SH	[*75-08-1*]	62.13	35.0	0.8391	1.43105	−17
$CH_3-(CH_2)_2-SH$	[*107-03-9*]	76.16	67.8	0.8415	1.43832	−20
$(CH_3)_3C-SH$	[*75-66-1*]	90.19	64.2	0.8002	1.4232	−24
$CH_3-(CH_2)_3-SH$	[*109-79-5*]	90.19	98.4	0.8416	1.4430	12
$CH_3-(CH_2)_5-SH$	[*111-31-9*]	118.24	152.6	0.838	1.4497	20
	[*1569-69-3*]	116.23	158	0.950	1.4921	43
$CH_3-(CH_2)_7-SH$	[*111-88-6*]	146.30	199.1	0.8433	1.4540	68
$(CH_3)_3C-CH_2-C(CH_3)_2-SH$	[*141-59-3*]	146.30	155.5	0.845	1.456	
$CH_3-(CH_2)_{11}-SH$	[*112-55-0*]	202.41	142–5/2	0.8450	1.4589	
tert.-$C_{12}H_{25}-SH$ (isomeric mixture)	[*25103-58-6*]	202.41	ca. 233	0.86	1.467	

wastewater with SH-group-containing resins or ion exchangers [20].

Thiolate formation is also utilized in the treatment of heavy-metal poisoning (copper, lead, mercury, zinc, cobalt, and gold). The thiol groups of essential enzymes that have been blocked by such metal ions are liberated again with the aid of the more strongly chelating thiols. The water-soluble substances 2,3-dimercaptopropane-1-sulfonic acid [*4076-02-2*] and dimercaptosuccinic acid [*304-55-2*] are being increasingly used for this complexation reaction, alongside 2,3-dimercapto-1-propanol [*59-52-9*], also known as dimercaprol or British Anti-Lewisite (BAL), and penicillamine [*52-67-5*], which have been known for a long time [21]. Mercaptoacrylic acids are also said to be suitable antidotes for lead and nickel poisoning [22]. Zinc and cadmium salts of cylcohexanethiol can be used as catalysts, crop-protection agents, and polymer stabilizers [23]. Certain gold thiolates are used in the treatment of arthritis. Organic antimony thiolates can be used as passivators for catalysts employed in the cracking of hydrocarbons [24]. Thiolate-coordinated transition metals play an important role in many physiological processes [25].

Oxidation. Thiols are readily converted into the corresponding disulfides by mild oxidizing agents.

$$2\,R-SH \underset{\text{Red.}}{\overset{\text{Ox.}}{\rightleftharpoons}} R-S-S-R$$

This facile oxidizability is the basis of quantitative thiol determination, and it is used industrially for sweetening crude-petroleum distillates. In this process the thiol-containing acidic fractions are oxidized with atmospheric oxygen under basic conditions in the presence of such catalysts as cobalt phthalocyanine monosulfonate [26] or EDTA complexes of copper(II) [27] or cobalt(II) [28]. Many other reagents can also be used for oxidizing thiols to disulfides (see Section 1.6.3). The thiol–disulfide redox system also plays an important role in a range of biological processes (glutathione, cysteine–cystine) [29].

Thiols and disulfides are oxidized to sulfonic acids by stronger oxidizing agents via a series of isolable intermediates. These oxidizing agents include nitric acid, potassium permanganate, chlorine gas [30], a mixture of hydrogen peroxide and aqueous hydrochloric acid [31], and atmospheric oxygen in the presence of catalytic amounts of dimethylsulfoxide and hydrogen halide [32].

Addition Reactions. Being strong nucleophiles, aliphatic thiols add to multiple bonds to give thioethers. This reaction, which is also used for the synthesis of unsymmetrical sulfides (see Section 1.6.2), produces a normal Markovnikov adduct (→ Alcohols, Aliphatic,

Chap. 2.3.7.) under ionic reaction conditions and an anti-Markovnikov adduct in the presence of radical-forming agents.

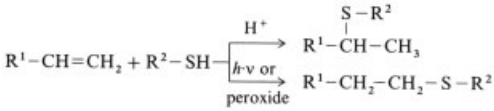

Thiols add to carbonyl compounds in the same way as alcohols. With aldehydes or ketones the products are thioacetals or thioketals [33], and with carboxylic acids and their derivatives they are *S*-esters of thiocarboxylic acids [34]. Other transformations of thiols are described in [35].

1.3. Production of Aliphatic Thiols

1.3.1. From Alcohols

The reaction of hydrogen sulfide with alcohols is used mainly for industrial syntheses of low molecular mass thiols.

$$ROH + H_2S \longrightarrow R\text{–}SH + H_2O$$

Such reactions are carried out in the gas phase> 300 °C, and generally require a catalyst. The thorium oxide [36] once used for this purpose has now largely been replaced by basic catalysts on a carrier of basic aluminum oxide. These catalysts favor thiol formation, whereas disulfides are the main products in acidic medium [37]. Promoters for thiol formation include a range of metal oxides, potassium and sodium tungstates, and heteropolyacids and their salts [37–40]. Thioethers and ethers may form as byproducts, along with olefins (except in the case of methanol).

The method described is used mainly for the industrial production of *methanethiol* [38,41], although ethanethiol is also produced by this route [42]. The method is not suitable for primary thiols with more than three carbon atoms because of the small boiling-point differences between the products and the starting alcohols (see Fig. 1).

Tertiary thiols can be produced in the liquid phase by reacting alcohols with hydrogen sulfide in the presence of sulfuric acid. This method leads to *triphenylmethanethiol* [*3695-77-0*] in 75 % yield [43], and is used industrially to produce *2-methyl-2-propanethiol* (*tert*-butanethiol) [44].

The reaction of alcohols with phosphorus pentasulfide is frequently employed for this purpose as well, and has been patented for alkanethiols [45]. C_4 to C_{16} alcohols are converted in this way into dialkyl dithiophosphates, which give the corresponding thiols on acid hydrolysis.

A more recent method uses Lawesson's reagent [*19172-47-5*] for the conversion of alcohols into thiols [46].

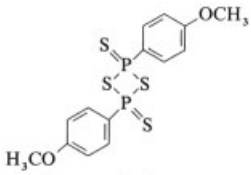

Lawesson's Reagent

1.3.2. From Alkenes

A second industrially important process for aliphatic thiol production is the addition of hydrogen sulfide to alkenes. In principle, both Markovnikov (**i**) and anti-Markovnikov (**ii**) adducts can result from this reaction, but formation of one isomer or the other can be favored by the use of suitable catalysts.

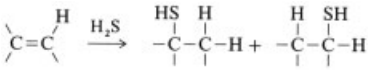

The Markovnikov product (**i**) is formed preferentially under ionic conditions in the presence of oxides, sulfides, acids, or elemental sulfur. Branched alkenes react more rapidly than straight-chain or terminal alkenes. Secondary thiols are prepared by the addition of hydrogen sulfide to linear alkenes in the presence of acid catalysts [47]. The analogous reaction of branched alkenes to give tertiary thiols is catalyzed by aluminum chloride, Al/Si oxides [48], zeolites [49], or cation exchange resins [50].

This method is exploited for the industrial production of a variety of thiols. Reaction is carried out in either the gas or liquid phase depending on the catalyst used. The yield of thiol increases with excess hydrogen sulfide (3 –10 mol

per mole of alkene), high pressure (≤ 7 MPa), and relatively low reaction temperature. Dialkyl sulfides are formed as byproducts. *Ethanethiol* is prepared from ethylene and hydrogen sulfide in the gas phase over aluminum oxide catalysts containing potassium tungstate [51] or nickel and tungsten oxides [52]. *Cyclohexanethiol* is obtained with cyclohexene and hydrogen sulfide in the gas phase over nickel sulfide on silica gel [53]. The reaction of cyclohexene with hydrogen sulfide over cobalt molybdate has also been described [54]. The industrial production of "*tert*-dodecyl mercaptan," a mixture of isomeric thiols, starts from oligomers of isobutylene or propylene. In the presence of hydrogen sulfide and a boron trifluoride catalyst, "*tert*-dodecyl mercaptan" can be prepared from triisobutylene at $-40\,^{\circ}$C [41] and from propylene tetramer at $60\,^{\circ}$C [55]. Aluminum chloride [41], ethyl aluminum dichloride [56], a mixture of boron trifluoride with phosphoric acid and a low molecular weight alcohol [57], and sodium oxide on zeolite [58] have also been described as catalysts. The industrial production of *2,4,4-trimethyl-2-pentanethiol* is similar to that of *tert*-dodecyl mercaptan. Terpenes are also used as industrial starting materials in this reaction. Thus, 1-*p*-menthene-8-thiol [*71159-90-5*], a component of the grapefruit aroma, is obtained selectively and in high yield by the addition of hydrogen sulfide to limonene in the presence of inorganic Lewis acid catalysts [59].

Electron-withdrawing groups, such as the carbonyl, cyano, nitro, carboxyl, or ester functions, activate alkenes toward nucleophilic addition of hydrogen sulfide. Basic catalysts are particularly suitable for these reactions, including amines [60], anion exchange resins [61], and magnesium oxide [61].

$$CH_2{=}CH{-}X + H_2S \xrightarrow{\text{base}} HS{-}CH_2{-}CH_2{-}X$$

$$X = -CO{-}CH_3,\ -CN,\ -NO_2,\ -COOH,\ -COOR$$

Formation of the anti-Markovnikov product (**ii**) results from a radical addition of hydrogen sulfide to alkenes (→ Hydrogen Sulfide, Chap. 3.3.2.). In industry this reaction is usually initiated with UV light ($\lambda = 253.7$ nm) [62, 63]. Longer wavelength radiation ($\lambda = 300 - 400$ nm) can also be used in the presence of photosensitizers [62–64]. Among the chemical radical initiators, azobisisobutyronitrile is particularly suitable [65]. This can also be used in the presence of accelerators, such as organic phosphines or phosphites [66]. Peroxides, on the other hand, can oxidize the hydrogen sulfide to elemental sulfur, which then acts as a catalyst for the ionic reaction pathway (see above) [67].

The method described is used mainly for the industrial production of primary thiols. The formation of dialkyl sulfides as byproducts is minimized by ensuring that excess hydrogen sulfide is present. Thus, *ethanethiol* is obtained from ethylene and hydrogen sulfide in the liquid phase in the presence of azobisisobutyronitrile and tributylphosphine [68]. *1-Dodecanethiol* is prepared by photochemically catalzyed addition of hydrogen sulfide to 1-dodecene. The olefin is reacted with a five-fold excess of hydrogen sulfide in the presence of α,α-diethoxyacetophenone as initiator at 45 °C and 3 MPa [63]. The industrial preparation of *1-octanethiol* is similar to that of 1-dodecanethiol.

1.3.3. Other Production Methods

Other methods for the industrial-scale preparation of thiols include the reductive thiolation of carbonyl compounds, epoxide ring-opening with hydrogen sulfide, and nucleophilic substitution of alkyl halides or tosylates.

Aldehydes and ketones are reduced to thiols by hydrogen sulfide and hydrogen in the presence of metal sulfides [69]. This process is used for the industrial production of cyclohexanethiol [70].

$$\overset{\backslash}{\underset{/}{C}}{=}O + H_2 + H_2S \xrightarrow{\text{cat.}} \overset{\backslash}{\underset{/}{C}}H{-}SH + H_2O$$

Base-catalyzed cleavage of epoxides with hydrogen sulfide gives mercaptoalcohols. Symmetrical bis(1-hydroxyalkyl) sulfides are formed as byproducts. Inorganic bases, amines and anion exchange resins, zeolites [71], or guanidine [72] can be used as catalysts. In industry, preparation of *mercaptoethanol* [*60-24-2*] by addition of H_2S to ethylene oxide is catalyzed by either a cation exchange resin [73] or the byproduct thiodiglycol [74].

$$H_2C{-}\underset{O}{CH}{-}R + H_2S \xrightarrow{\text{base}} H_2\underset{HS}{C}{-}\underset{OH}{CH}{-}R$$

The reaction of alkyl halides or tosylates with metal hydrogensulfides is frequently used to prepare thiols on a laboratory scale [75].

$$R-X + M-SH \xrightarrow[-MX]{} R-SH$$

X = Cl, Br, I, –O-tosyl

Alcohols are the usual solvents. The competing formation of sulfide can be suppressed by the presence of hydrogen sulfide or sulfuric acid. Sulfide formation is also strongly inhibited in the reaction of alkyl halides with a mixture of hydrogen sulfide and ammonia or alkylamines [76]. C_2 to C_{12} alkanethiols form in essentially quantitative yield from the corresponding bromides and sodium hydrogensulfide in dimethylsulfoxide [77]. The nucleophilic substitution reaction is used industrially to produce *1-dodecanethiol* from 1-chloro- or 1-bromododecane [78].

Substitution of – SH for chlorine occurs particularly readily if the process is activated by the presence of electron-withdrawing groups (e.g., C=O, C≡N) or heteroatoms (O, S, Si) in the α-, β-, or γ-positions [79]. This method is used to prepare mercaptocarboxylic acids and thiols containing additional functional groups [79, 80].

$$Cl-(CH_2)_n-COOH \xrightarrow[2)\ HCl]{1)\ NaSH} HS-(CH_2)_n-COOH$$

$$Cl-CH_2CH_2-O-R \xrightarrow{NaSH} HS-CH_2CH_2-O-R$$

Besides the processes commonly used in industry for thiol production, a number of generally applicable laboratory methods are also available, such as the reaction of an alkyl halide with thiourea, thiosulfate, or dithio- or trithiocarbonate. Details of these methods can be found in the general literature.

1.4. Construction Materials, Storage, and Transport

Manufacturing plant and transport facilities for thiols are preferentially made from (coated) normal steel, stainless steel, aluminum, or copper-free alloys. Thiols in containers made of carbon steel should be handled dry and under a blanket of protective gas to prevent the formation of pyrophoric iron – sulfur complexes. Copper and copper-containing alloys are not suitable for use because they are rapidly attacked by thiols.

Stringent demands apply to the impermeability of thiol piping systems. Seals are normally based on asbestos or fluoroplastics.

Regulations in force with respect to the transport of industrially important thiols are summarized in Table 2.

Table 2. Transport restrictions applicable to aliphatic thiols

Thiol	ADR/RID		IDMG Code	
	Class	Number		
CH_3-SH	2	3bt	2	UN 1064
C_2H_5-SH	3	18b	3.1/II	UN 2363
n-C_4H_9-SH	3	3b	3.2/II	UN 2347
t-C_4H_9-SH	3	3b	3.2/II	UN 1228
cyclo-$C_6H_{11}-SH$	3	31c	3.3/III	UN 3054
n-$C_8H_{17}-SH$	3	32c		
n-$C_{12}H_{25}-SH$	3	32c		
	6.1	21c		
t-$C_{12}H_{25}-SH$	6.1	21c		

1.5. Uses of Aliphatic Thiols

Many aliphatic thiols are important starting materials for the synthesis of crop-protection agents (especially thiophosphoric acid esters) [81], pharmaceuticals, and polysulfides (see Section 1.6.4). They are also widely used as polymerization regulators in rubber and plastics manufacture. *Methanethiol* is important as a starting material for the industrial production of the amino acid DL-methionine, used as an animal-feed additive. Methanethiol is also used as a gas odorizer, as are other low molecular mass thiols (in particular, ethanethiol). For applications of thiolates and chelate-forming thiols see Section 1.2.2.

Pharmaceuticals. Mercaptoamino acids and their derivatives display a broad spectrum of activity. *N*-Isobutyryl- and *N*-pivaloylcysteine are said to exhibit improved bioavailability compared with *N*-acetylcysteine [*616-91-1*], which is used as a mucolytic [82].

Vanadyl – cysteine complexes can be used for oral treatment of insulin-dependent diabetes [83], and substituted cysteinamides for the localized treatment of inflammatory skin diseases [84]. Certain mercaptoamino acids, such as the 2*S*-enantiomer of 1-(3-mercapto-2-methylpropionyl)-L-proline (Captopril [*62571-86-2*]), act as antihypertensives [85].

Other mercaptoamino acid derivatives have been described as analgetics [86] and antiarthritic agents [87].

Diaminodithiols are chelating agents, and thus useful in radioactive diagnosis [88]. Thiol-substituted dioxolanes [89] and cyclohexenes [90] are mucolytics. Mercapto-substituted propanamides exhibit psychotropic [91] and analgesic activity [91,92].

Agricultural Chemicals. Aminodithiols have been described as effective low-toxicity insecticides [93]. Bismuth dithiolates can be used to combat coccidiosis in poultry and as insecticides [94].

Thiol-substituted 2-propanols are fungicides with low toxicity [95]. Bis(dithiobenzyl)nickel complexes absorb IR radiation and can be used as plant-growth regulators [96].

Other Uses. *N*-Substituted1-amino-2-propylmercaptans are used as antioxidants in lubricating oils [97]. 4-Mercaptobutyramides can act as reducing agents in the permanent waving of hair. They are said to be less toxic and accompanied by less noxious odors than mercaptoacetic acid [98]. Mercaptoalkyl esters of acrylic and methacrylic acids are used as binders for metallic fillings for teeth [99]. Many thiol-substituted terpenes [100], esters [101], ketones [102], and mercaptoalkyloxathiolanes and -oxathianes [103] have been patented as fragrances and aroma substances (e.g., for foods, cosmetics, or pharmaceuticals).

1.6. Alkyl Sulfides, Disulfides, and Polysulfides

1.6.1. Physical and Chemical Properties

Sulfides (R^1-S-R^2) are the sulfur analogs of ethers, and are thus derivatives of hydrogen sulfide. They were referred to previously as thioethers or sulfanes. Alkyl *disulfides* ($R^1-S-S-R^2$) and *polysulfides* ($R^1-S_x-R^2$) are derived from di- and polysulfanes. Symmetrical sulfides contain R groups that are identical ($R^1 = R^2$), whereas unsymmetrical ones have different R groups ($R^1 \neq R^2$). Sulfides are named in the same way as ethers (e.g., diethyl sulfide, methyl phenyl sulfide). Di- and polysulfides are named either as sulfides or as sulfane derivatives (e.g., dimethyl disulfide, diethyltrisulfane). Most low molecular mass alkyl sulfides have intense, unpleasant odors. Organic sulfides and disulfides are widely distributed in natural materials. For example, they occur (together with thiols) in onions and garlic [104], and have been found to be the source of various well-known aromas, including those of asparagus and coffee. In polecats, mink, and skunks such compounds act as pheromones [105]. The cyclic polysulfide lenthionine confers a characteristic odor on cooked mutton and the shiitake mushroom [106]. Many disulfides are components of biologically active substances (e.g., lipoic acid, cystine, holomycin) and physiologically important peptides (e.g., insulin, oxytocin). Here the activity is associated directly with the presence of the disulfide group. Aliphatic sulfides are also present in crude petroleum, accounting for up to 45 % of the total sulfur, but disulfides are virtually absent [107].

Aliphatic sulfides, disulfides, and polysulfides are colorless to yellow or orange liquids. Most are insoluble in water, but readily soluble in ether, acetone, chloroform, and ethanol. Physical data for a few representative compounds are provided in Table 3 (see next page).

Alkyl sulfides, like their oxygen analogs, are very weak bases. Because of nucleophilicity attributable to the free electron pair on sulfur, they form stable addition compounds with halogens and salts of heavy metals. Aliphatic sulfides also form crystalline trialkylsulfonium salts with alkyl halides or dialkyl sulfates.

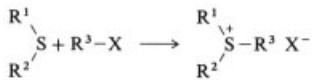

X = halogen, $O-SO_2-O-R^3$

Depending on the reaction conditions, oxidation of sulfides leads to sulfones or sulfoxides [107]. Sulfonic acids are formed under forcing conditions, and these are then further oxidized to sulfuric acid [108].

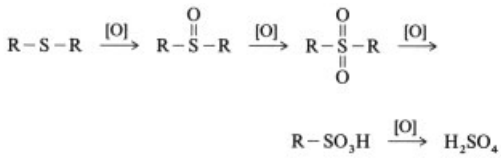

Table 3. Physical data for selected alkyl sulfides, disulfides, and polysulfides

Sulfide	CAS registry no.	M_r	*bp*, °C/kPa	d_4^{20}	n_D^{20}	Flash point, °C
$(CH_3)_2S$	[*75-18-3*]	62.13	37.3	0.846	1.4355	−36
$(CH_3CH_2)_2S$	[*352-93-2*]	90.19	92.1	0.8362	1.4430	−9
$(CH_2{=}CH{-}CH_2)_2S$	[*592-88-1*]	114.21	138	0.887	1.4889	46
$[CH_3(CH_2)_2]_2S$	[*111-47-7*]	118.24	142	0.838	1.4487	28
$[CH_3(CH_2)_3]_2S$	[*544-40-1*]	146.30	188	0.838	1.4530	76
$[CH_3CH_2CH(CH_3)]_2S$	[*626-26-6*]	146.30	165	0.839	1.4500	39
$[(CH_3)_3C]_2S$	[*107-47-1*]	146.30	149	0.815	1.4506	48
$[CH_3(CH_2)_5]_2S$	[*6294-31-1*]	202.40	230	0.849	1.4587	>110
$(CH_3)_2S_2$	[*624-92-0*]	94.20	109	1.046	1.5253	24
$(CH_3CH_2)_2S_2$	[*110-81-6*]	122.25	152	0.993	1.5073	40
$(CH_2{=}CH{-}CH_2)_2S_2$	[*2179-57-9*]	146.28	187	1.008	1.5410	62
$[CH_3(CH_2)_3]_2S_2$	[*629-45-8*]	178.36	226	0.9382	1.4926	93
$(CH_3)_2S_3$	[*3658-80-8*]	126.27	42.5/0.4	1.2013	1.6012	
$(CH_3)_2S_4$	[*5756-24-1*]	158.33	59/0.1	1.3065	1.6612	
$(CH_3)_2S_5$	[*7330-31-6*]	190.39	70/0.01			

Sulfides can be reduced to hydrocarbons or, depending on the reducing agent and the substrate, thiols or hydrogen sulfide.

$$R{-}S{-}R \xrightarrow{\text{Red.}} R{-}H + R{-}SH \xrightarrow{\text{Red.}} 2R{-}H + H_2S$$

The sulfide contraction reaction, a desulfurization with simultaneous C−C bond formation, is important in organic synthesis [109].

Reaction conditions determine whether disulfides are oxidized to thiosulfinic or thiosulfonic acid esters or to sulfonic acids.

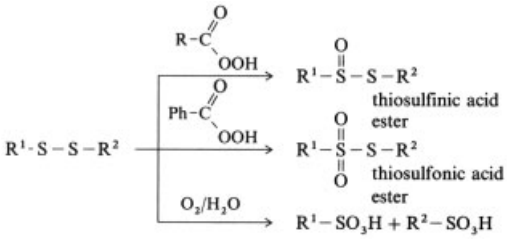

The reduction of disulfides to thiols is a reversible reaction of great physiological importance. Hydrogen, zinc in the presence of acid, complex hydrides, or sodium are generally selected as the reducing agents [110]. Disulfides can be reduced selectively with sodium hydrogentelluride in good yield under very mild conditions in the presence of other functional groups [111]. Organic polysulfides can also be reduced to thiols.

1.6.2. Production of Alkyl Sulfides

Many of the procedures for the synthesis of dialkyl sulfides correspond to those for aliphatic thiols, including the alkylation of hydrogen sulfide or alkali metal sulfides with alcohols, alkenes, or alkyl halides to give symmetrical sulfides (Section 1.3). *Dimethyl sulfide*, the industrially most important sulfide, is produced by treating hydrogen sulfide with excess methanol over an aluminum oxide catalyst. Unsymmetrical sulfides are formed in a similar way from thiols and their salts [112, 113], but they can also be obtained by alkylation of thiols with acidic C−H compounds [114]. The alkylation of thiols with alkyl halides in the presence of K_2CO_3 in DMF occurs particularly rapidly, and in almost quantitative yield [115]. Unsymmetrical sulfides can also be made by exchange reactions involving symmetrical sulfides and thiols, catalyzed by phosphotungstic or phosphomolybdic acids [116].

$$R^1{-}S{-}R^1 + R^2{-}SH \xrightarrow{\text{cat.}} R^1{-}S{-}R^2 + R^1{-}SH$$

Cleavage of an epoxide ring with a thiol proceeds analogously to the cleavage with hydrogen sulfide (Section 1.3.3). Bases (such as sodium hydroxide or triethylamine [117]), metal salts (e.g., lithium perchlorate [118]), and acids [119] have all been used as catalysts. The resulting β-hydroxythioethers are important intermediates for a large number of other products.

Alkylthio-substituted carboxylic acids and their thioesters can be prepared from lactones

by treatment with thiols in the presence of an acid catalyst at 100 – 350 °C [120].

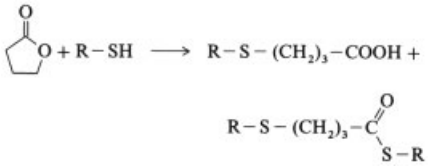

The reductive thiolation of carbonyl compounds (Section 1.3.3) can also be carried out with thiols instead of hydrogen sulfide. The borane – pyridine complex in trifluoracetic acid [121] or treatment with boron trifluoride monohydrate followed by triethylsilane [122] can be used to accomplish the reduction.

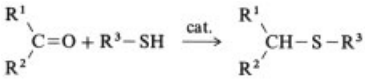

α-Substituted dialkyl sulfides can be prepared from sulfoxides by means of the Pummerer rearrangement (→ Acetic Anhydride and Mixed Fatty Acid Anhydrides). The deoxygenation of sulfoxides to sulfides is a particularly important approach with respect to asymmetric syntheses. Several methods are known for achieving this reduction [123]. The systems sodium iodide – titanium(IV) chloride [124] and sodium iodide – sulfonic acid [125] are described as especially efficient and mild reagents. The selective reduction of various functionalized sulfoxides can be effected using Lawesson's reagent [126] (see Section 1.3.1).

1.6.3. Production of Alkyl Disulfides

Dialkyl disulfides are often formed as byproducts in the synthesis of sulfides. However, the most important synthetic method is the mild oxidation of thiols. In industry this oxidation is carried out in aqueous alkaline solution using air or oxygen. The reaction is catalyzed by radical initiators (copper, cobalt, nickel, or iron salts, or UV light) [127]. Even high molecular mass disulfides can be prepared in this way [128]. Since in this case water-insolubility of the products can be a disadvantage, the reaction is sometimes carried out in a two-phase mixture in the presence of phase-transfer catalysts [129]. For the oxidation of tertiary thiols a catalyst system consisting of alkali-metal and/or alkaline-earth hydroxides and cobalt molybdate is particularly suitable [130]. The oxidation of thiols with atmospheric oxygen is also exploited industrially for sweetening crude petroleum distillates (see Section 1.2.2).

Other potential oxidizing agents include hydrogen peroxide, halogens (in particular, iodine) [131], and potassium hexacyanoferrate(III) [132]. The fire hazard is lowered and the yield of dialkyl disulfide increased using carbon tetrachloride in aqueous alkali as the oxidizing agent together with a phase-transfer catalyst [133].

Interest has been shown in the oxidation of thiols with nitrogen oxides. Whereas symmetrical sulfides are obtained in good yields with nitric oxide and nitrogen dioxide [134], dinitrogen tetroxide leads first to thionitrites [135], which may then react with an additional equivalent of thiol to give symmetrical or unsymmetrical disulfides [136].

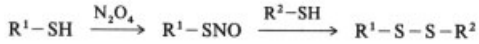

The analogous coupling of thiols with iron(III) nitrate on a bentonite carrier to give symmetrical disulfides takes place under very mild conditions [137].

An important method for the production of unsymmetrical disulfides is the thioalkylation of thiols with sulfenyl halides [138].

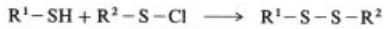

1.6.4. Production of Alkyl Polysulfides

Alkyl polysulfides are produced industrially by treating thiols with sulfur in the presence of bases (alkali-metal hydroxides or tertiary amines) [139]. Mixtures of various polysulfides are frequently formed in this reaction, so the isolation, purification, and characterization of defined compounds can be difficult. Higher molecular mass dialkyl trisulfides can be obtained in very good yields from the corresponding thiols and sulfur with magnesium oxide as the base at 30 – 100 °C [140]. The reaction of thiols with electrolytically formed $S^{2+}{}_x$ gives trisulfides in 53 – 74 % yields [141]. Alkyl polysulfides containing an average of five sulfur atoms can be prepared at room temperature in the presence of thioxanthates as catalysts [142].

The reaction of dihaloalkanes with alkali polysulfides gives polymeric polysulfides with the general formula $-[R-S_x-]_n$. These polysulfide rubbers (thioplastics) are widely used in industry because of their resistance toward chemicals and light and their impermeability toward gases (→ Rubber, 3. Synthetic, Chap. 5.3.; → Sealing Materials, Chap. 3.).

1.6.5. Uses of Alkyl Sulfides, Disulfides, and Polysulfides

The most important industrial application of aliphatic sulfides and disulfides is as intermediates in the production of pharmaceuticals, crop-protection agents, and a wide range of other substances [143]. Alkyl sulfide groups themselves are also present in the active components of many pharmaceuticals and crop-protection agents, especially insecticides [144]. *Dimethyl sulfide*, the industrially most important sulfide, is widely used in the form of the borane complex $(CH_3)_2S \cdot BH_3$ [*13292-87-0*] for the reduction of a variety of functional groups [145]. Dimethyl sulfide is also the starting material for production of the solvent dimethyl sulfoxide (DMSO; → Sulfones and Sulfoxides, Chap. 2.2.1.).

Organic disulfides have a variety of applications as well. L-Cystine [*56-89-3*] and lipoic acid [*1077-28-7*] are therapeutic agents for the liver. Holomycin and the pharmaceutically active component in garlic, di(1-propenyl) disulfide [146], exhibit antibacterial activity. Dialkyl disulfides with 2 – 8 carbon atoms per alkyl residue can be used as insecticides and acaricides (→ Acaricides, Chap. 3.5.), or to combat algae [133]. Higher molecular mass disulfides are used as additives in the plastics industry for chain transfer during polymerization, and as antioxidants (→ Oxidation, Chap. 4.4.3.). Di- and polysulfides are used as high-pressure additives for lubricating oils.

2. Aromatic Thiols, Sulfides, Disulfides, and Polysulfides

2.1. Physical and Chemical Properties

Aromatic thiols, Ar–SH, are sulfur analogs of phenols. This relationship is also manifested in many names, where the prefix thio is added to the trivial name of the corresponding phenol (e.g., thiosalicylic acid, thiocresol). According to IUPAC rules, aromatic thiols are supposed to be referred to as mercaptoarenes (e.g., mercaptoaniline) or arenethiols (e.g., 2-naphthalenethiol).

Aromatic thiols differ only slightly from alkanethiols with respect to acidity, salt formation, odor, and solubility. Many thiophenols are volatile in steam. Table 4 provides an overview of physical properties for several important arenethiols.

Unlike aliphatic thiols, arenethiols occur only to a very limited extent in natural materials. One reason for this may be the more facile oxidation of aromatic thiols, which are converted into disulfides simply on standing in air. The powerful nucleophilicity of arenethiols is exemplified primarily by substitution and addition reactions, most of which lead to thioethers. Such reactions correspond largely to those of aliphatic thiols (see Section 1.2.2). Unlike phenol derivatives, derivatives of aromatic thiols undergo a number of rearrangements. A review of the reactions of aromatic thiols can be found in [11].

Table 4. Physical data for several thiophenols

R	CAS registry no.	M_r	*bp*, °C/kPa	*mp*, °C	Flash point, °C
H	[*108-98-5*]	110.18	169	−15	50
4-CH_3	[*106-45-6*]	124.21	195	43	68
4-Cl	[*106-54-7*]	144.62	206	50	>110
2,5-Cl_2	[*5858-18-4*]	179.07	132/3.9	28	>110
2,6-Cl_2	[*24966-39-0*]	179.07		49	>110
2-OH	[*637-89-8*]	126.18	150/3.3	34	>110
2-O-Me	[*7217-59-6*]	140.20	99/1.1		95
2-NH_2	[*137-07-5*]	125.19	71/0.03	20	79
2-COOH	[*147-93-3*]	154.19		167	

2.2. Production of Aromatic Thiols

2.2.1. By Reduction

The reduction of sulfonyl chlorides, which are readily accessible from the corresponding aromatic systems by sulfochlorination, is one of the most important methods for producing arenethiols.

$$Ar-H + HO-SO_2Cl \xrightarrow{-H_2O} Ar-SO_2Cl \xrightarrow{Red.} Ar-SH$$

Many different reducing agents can be employed. The use of zinc or tin in dilute acid is still important in industry [147]. For example, *thiophenol* [148], *halothiophenols* [149], or *aromatic dithiols* [150] continue to be produced by this method. The resulting levels of ecologically harmful metal salts can be considerably reduced by the use of such catalysts as lead [151] or various harmless metal salts [152].

Sulfonyl chlorides can be hydrogenated directly in the presence of metal sulfide catalysts to the corresponding thiophenols at 150 – 275 °C and a pressure of 3 – 17 MPa [153]. This process is used to produce *p-thiocresol*. The preparation of *2,5-dichlorothiophenol* by pressurized hydrogenation of the corresponding sulfonyl chloride over a platinum catalyst is similar [154].

Another commonly used reducing agent is red phosphorus in the presence of hydrogen iodide in aqueous [155] or glacial acetic acid solution [156]. Two-step reduction of a sulfonyl chloride is also practiced on an industrial scale, as in the preparation of thiophenol. Depending on the reaction conditions, sodium arenesulfinates or disulfides are formed initially by sulfite reduction, and these are subsequently converted into thiophenols by pressurized hydrogenation in the presence of metal sulfides [157].

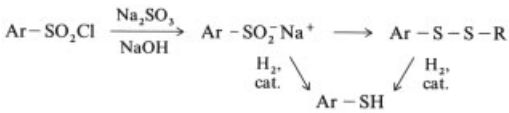

The reduction of aromatic sulfonic acids with triphenylphosphine in the presence of a reaction accelerator has also been described [158].

The tendency of aromatic thiols to undergo oxidation means that the actual products are often disulfides, so reduction of the latter and the corresponding polysulfides is frequently important. This indirect route to arenethiols often leads to better yields than the direct synthesis despite the additional reaction step. Many reducing agents are suitable for the reductive cleavage of disulfides. Pressurized hydrogenation in the presence of a catalyst is used in industry [159]. For halosubstituted compounds the preferred catalysts are Raney cobalt or palladium on charcoal [160]. Other reducing agents include metallic aluminum in liquid ammonia with inorganic halogen as the catalyst [161], zinc and acid in the presence of another metal [162], lithium aluminum hydride [163], hydrogen sulfide [164], and alkali-metal sulfides [165]. Functionalized aromatic disulfides are reduced selectively and under mild conditions using lithium (tri-*tert*-butoxy)aluminum hydride, where halogen, nitro, and carboxyl groups are spared from attack [166]. High yields of arenethiols are also obtained in the presence of a wide range of functional groups with a phosphine reagent bonded to a polymeric carrier [167]. For other reducing agents, particularly those used in the laboratory, see [2].

2.2.2. By Thiolation

A nucleophilic exchange between aromatically bonded halogen and the thiol group occurs under mild conditions only in the presence of other strongly electron withdrawing substituents, such as nitro or sulfonyl groups [168]. Either sodium hydrogensulfide or sodium disulfide is used as the thiolating agent. Poor yields and selectivities are often observed with sodium sulfide, particularly with weakly activated haloaromatics. Chloronitrobenzenes can be converted into the corresponding thiophenols either with retention [165, 169] or simultaneous reduction of the nitro group [170], depending on the reaction conditions.

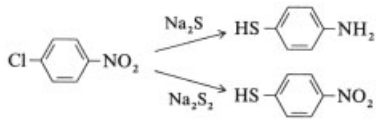

Polychlorinated benzenes also react with sodium hydrogensulfide to give the corresponding thiophenols [171].

The thiolation of haloaromatics is often used in the industrial preparation of arenethiols. *Pentachlorothiophenol* is obtained

from hexachlorobenzene by treatment with sodium sulfide and sulfur in methanol, or with sodium hydrogensulfide [172, 173]. Various halothiophenols can be produced by a similar method in such polar solvents as DMF, sulfolane, dimethylsulfoxide (DMSO), and *N*-methylpyrrolidone [174]. Mercaptoanilines are formed from halonitrobenzenes with simultaneous reduction of the nitro group [175]. Aromatic dithiols can be obtained in a particularly pure state and high yield from polyhalogenated aromatics by reaction with hydrogen sulfide and sulfur in the presence of iron [176]. Sodium alkanethiolates can be used for the thiolation of nonactivated aryl halides. The aryl alkyl sulfide formed initially is converted into the desired aromatic thiol by acid hydrolysis [177]. Nucleophilic substitution of an aryl iodide with thiourea in the presence of transition metal complexes leads to *S*-arylisothiuronium salts, which can by hydrolyzed to arenethiols [178].

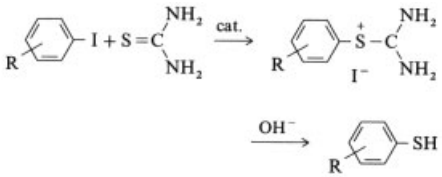

Another thiolating agent, one that is also suitable for the production of alkoxythiophenols, is disulfur dichloride. Disulfides are formed initially, and these are then reduced as described above (see Section 2.2.1). Various metal salts can serve as catalysts for this thiolation [179]. Similarly, monohalogenated benzenes can be converted into halothiols via polysulfides [180].

Other preparative methods applicable to arenethiols include the catalytic reaction of phenols with hydrogen sulfide in the gas phase [181] and reductive cleavage of aryl alkyl sulfides [182].

2.3. Uses of Aromatic Thiols

The most important industrial applications of arenethiols are as intermediates in the production of pharmaceuticals, agrochemicals [183], dyes, and pigments, and as chemicals for the electronics industry. Thiophenols are used in the rubber industry as polymerization regulators, plasticizers, and stabilizers [184], as well as for rubber reclamation. Aromatic dithiols and mercaptoanilines are used in the plastics industry as monomers and modifiers [185].

Thiophenolamides act as collagenase inhibitors, and can be used for treating arthritic conditions [186]. Thiophenolate-containing preparations can dissolve seals and coatings based on polysulfides [187].

Methyl and ethyl 2-mercaptobenzoates have been patented as scents [188]. Substituted mercaptoanilines are flotation agents for lead and zinc ores [189]. Polythiols can be used as optical materials in plastics [190], and for the production of lightweight reversible electrodes with high theoretical energy density [191]. The use of mercaptophenyl-substituted pyrrolinones as herbicides and growth regulators for crop-protection purposes has been described [192].

2.4. Production of Aryl Sulfides, Disulfides, and Polysulfides

The production processes for aromatic sulfides, disulfides, and polysulfides correspond largely to those for the related aliphatic compounds. Symmetrical aryl sulfides can be obtained by arylation of hydrogen sulfide or its salts with alkyl halides [193]. Similarly, unsymmetrical or mixed (aliphatic – aromatic) sulfides are formed from thiols or thiolates on treatment with alkyl or aryl halides [194]. The addition of thiols to multiple bonds [195] and the reaction of sulfenyl chlorides with arenes or phenols [196] also lead to unsymmetrical sulfides. Sulfides are formed as well in the cleavage of disulfides with phenols [197] or aromatic amines [198].

2.5. Uses of Aryl Sulfides, Disulfides, and Polysulfides

Aryl sulfides are used mainly as intermediates in the production of biologically active substances and dyes, and as monomers in the plastics industry.

Tetrasul and Mercaptodimethur are examples of commercial insecticides. Other aryl sulfides have been introduced as fungicides [199], herbicides [200], microbial growth inhibitors [201], acaricides and nematocides [202], miticides

[203], and gametocides [204]. Aryl thioethers are components of color developers in recording materials [205] and of the photosensitive elements in photocopiers and printers [206]. Certain monomeric aryl sulfides can be polymerized to materials with interesting optical properties, such as transparency or a high refractive index [207]. In the plastics industry arylthio compounds can be used as hardeners [208], antioxidants [209], or heat and light stabilizers [210]. Other aryl sulfides function as lubricating-agent additives [211], hair dyes [212], and flotation agents [213]. Aromatic sulfides are used in the pharmaceutcial sector for treating immunological diseases [214], skin diseases [215], and asthma [216], and also as muscle relaxants [217].

Aromatic disulfides have been patented as stabilizers for polycarbonate plastics [218], corrosion inhibitors in transport systems for crude petroleum and natural gas [219] and antioxidants for use in lubricating agents [220].

Aryl polysulfides in rubber act simultaneously as antioxidants and cross-linking agents [221].

Poly(*p*-arylenesulfides) are industrially important as thermoplastic polycondensates with good mechanical and thermal characteristics [222].

3. Heterocyclic Thiols, Sulfides, and Disulfides

3.1. Introduction

Heterocyclic thiols and their derivatives are widely distributed in nature. For example, the eggs of such marine invertebrates as sea urchins, starfish, and squid contain relatively high concentrations of mercaptohistidine and the corresponding methyl derivatives, known as "ovothiols" [223]. These compounds function as biological antioxidants in the fertilized eggs [224].

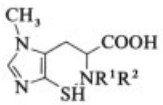

R^1	R^2	Ovothiol
H	H	A
CH_3	H	B
CH_3	CH_3	C

2-Methyl-3-furanthiol and its disulfide and *S*-methyl derivatives have all been identified as aroma components of cooked meat [225], and they have also been incorporated into synthetic flavorings [226]. 2-Methyl-3-furanthiol has been shown to be present in the aromas of roasted coffee [227] and canned tuna [228]. *S*-Substituted furans, thiophenes, thiazoles, and 3-thiazolines contribute to the odor of yeast extracts [229]. Heterocyclic sulfides occur in cytokinins of the bacteria *Escherichia coli* [230] and in the phallotoxins, toxic substances from the "death cup" mushroom *Amanita phalloides* [231].

3.2. Physical and Chemical Properties

Heterocyclic thiols are generally colorless or pale yellow solids with relatively faint characteristic odors. Table 5 provides an overview of the physical properties of several heterocyclic thiols.

Table 5. Physical data for selected heterocyclic thiols

Formula, name	CAS registry no.	M_r	*mp*, °C
Imidazole-2-thiol	[*872-35-5*]	100.14	228–231
2-Thiazoline-2-thiol	[*96-53-7*]	119.21	105–107
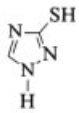 [1*H*]-1,2,4-Triazole-3-thiol	[*3179-31-5*]	101.13	221–224
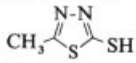 5-Methyl-1,3,4-thiadiazole-2-thiol	[*29490-19-5*]	132.21	188–189
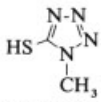 1-Methyl-[1*H*]-tetrazole-5-thiol	[*13183-79-4*]	116.15	125–128
Pyridine-2-thiol	[*2637-34-5*]	111.17	128–130
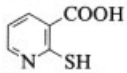 2-Mercaptonicotinic acid	[*38521-46-9*]	155.18	270 (decomp.)

Table 5. Continued.

Formula, name	CAS registry no.	M_r	*mp*, °C
Pyrimidine-2-thiol	[*1450-85-7*]	112.15	230 (decomp.)
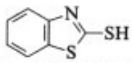 Benzothiazole-2-thiol	[*149-30-4*]	167.25	182
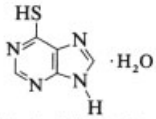 6-Purinethiol hydrate	[*6112-76-1*]	170.19	>300
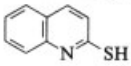 2-Quinolinethiol	[*2637-37-8*]	161.23	174–176
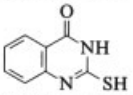 2-Mercapto-4[3*H*]-quinazolinone	[*13906-09-7*]	178.21	>300

Just like the corresponding aliphatic or aromatic compounds, heterocyclic thiols form salts with many metals, some of which have therapeutic applications (see Section 3.4).

Nitrogen-containing heterocyclic thiols are subject to the same tautomeric equilibrium between thiol and thione forms exhibited by the analogous oxygen species.

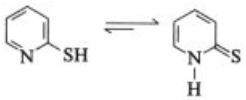

Spectroscopic studies [232] have shown that this equilibrium is displaced in the direction of the thione form. Reactions of heterocyclic thiols can take place from either the thiol or the thione form.

The mild oxidation of heterocyclic thiols leads to disulfides, just as with the corresponding aliphatic and aromatic systems. Catalytic oxidation of 2-mercapto-1,3-benzothiazol with atmospheric air is not practical, so the corresponding disulfide is prepared industrially by oxidation with sodium chlorate [233].

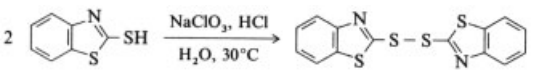

Stronger oxidizing agents like potassium permanganate or potassium peroxide transform heterocyclic thiols into sulfonic acids [234]. Treatment of 3-mercapto-1,2,4-triazole or 2-mercaptopyridinoimidazole with nitric acid leads to oxidative desulfurization [4].

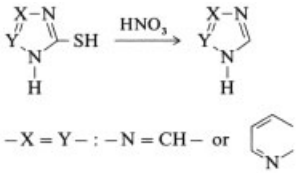

However, a more widespread desulfurization method for heterocyclic thiols is the reductive approach, in which Raney nickel [235] or sodium borohydride [236] is usually the reagent of choice.

Oxidation of heterocyclic thiols with hypochlorite in the presence of excess amine results in the corresponding sulfenamides [237], and disulfides are subject to similar treatment [238].

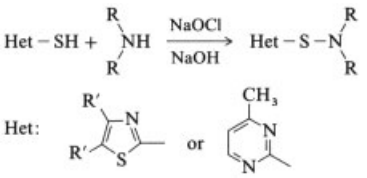

Heterocyclic thiols or sulfides are easily transformed into amino or hydrazo derivatives by substitution [239].

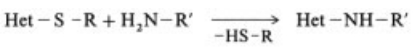

Moreover, heterocyclic thiols are subject just like their aliphatic and aromatic counterparts to such transformations as etherification, esterification, or rearrangement.

3.3. Production

A wide variety of methods is available for the preparation of heterocyclic thiols. Thus, a thiol function can either be introduced by substitution into an existing heterocyclic system, or else the heterocycle itself may be constructed by cyclization of a sulfur-containing precursor.

3.3.1. Heterocyclic Thiols by Substitution

Generally speaking, heterocycles undergo substitution to thiols under substitution conditions similar to those applicable to activated benzene derivatives. One frequently applied method is the thiolation of a halo-substituted heterocycle, taking advantage of a wide range of thiolation reagents. Particularly suitable in this context are sodium or potassium hydrogensulfide [240] and thiourea [241]. Other possibilities include elemental sulfur [242], thiosulfate [243], alkali thiocyanates, and thiocarboxylic acid derivatives.

Heteroaromatic organomagnesium or organolithium compounds react with sulfur to give thiolates, hydrolysis of which leads to the corresponding thiols. This method permits the preparation of mercaptothiophenes in yields up to 70 % [244].

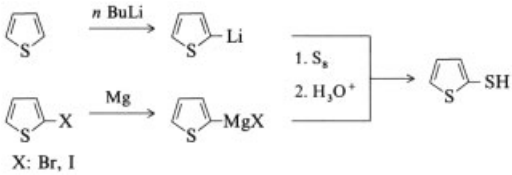

Similarly, the mercapto function can be introduced into *N*-protected 2-aminopyridines with dibenzylsulfide [245].

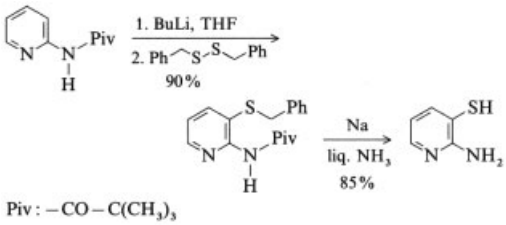

Just like aryl amines, heteroaromatic amines are subject to conversion by way of diazonium salts into the corresponding thiols [246].

Heterocycles bearing oxygen substituents can also be transformed into thiols. Suitable reagents include especially phosphorus pentasulfide [247], hydrogen sulfide [248], and the Lawesson reagent [249], just as in the case of alkanethiols (see Section 1.3.1).

Another reaction that has been used to prepare heteroaromatic thiols is the reduction of a sulfonic acid with iodine and triphenylphosphine [250].

3.3.2. Heterocyclic Thiols by Cyclization

Cyclocondensations with such sulfur-containing species as carbon disulfide, thiourea, or thiocyanates are especially important reactions for the synthesis of heterocyclic thiols.

Carbon Disulfide. 2-Mercapto-1,3-benzothiazoles are prepared on an industrial scale from anilines by cyclocondensation with carbon disulfide and sulfur. One reaction of this type exploited industrially, that of *o*-aminothiophenols, occurs even at low temperature [251].

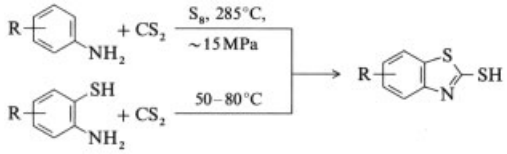

o-Phenylenediamines react analogously to 2-mercaptobenzimidazoles, whereby potassium xanthogenate can be utilized in place of carbon disulfide if so desired [252].

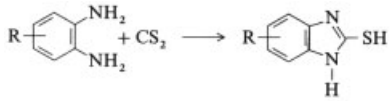

1,2-Diaminoalkanes are transformed by carbon disulfide into 2-mercaptoimidazolines [253].

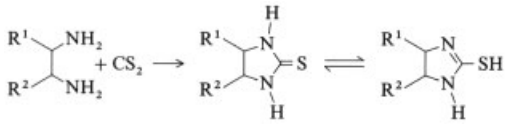

The cyclization of 2-aminobenzonitriles with carbon disulfide or potassium xanthogenate leads to mercapto-substituted quinazolines [254].

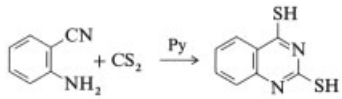

Cycloaddition of carbon disulfide to sodium azide produces the sodium salt of 5-mercapto-1,2,3,4-thiatriazole [255].

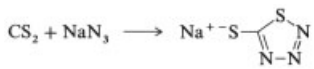

Thiourea. Condensation of carbonyl compounds with thiourea derivatives represents another important approach to the synthesis of heterocyclic thiols. Thus, cyclocondensation of α-hydroxyketones with thiourea results in 2-mercaptoimidazoles [256].

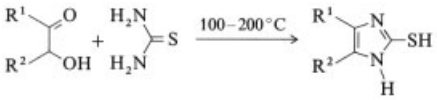

The analogous conversion of β-dicarbonyl compounds to 2-mercaptopyrimidines has been subject to especially widespread utilization [257].

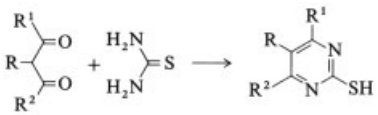

The usual starting materials are β-diketones or β-ketoaldehydes along with the corresponding acetals and acetoacetic or malonic ester derivatives. The latter lead to thiobarbiturates, which are used as short-duration narcotics.

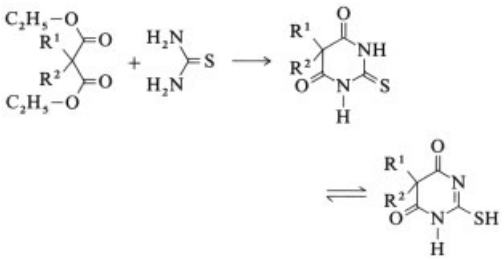

Thiourea reacts with two equivalents of an aldehyde and one equivalent of amine to provide mercapto-substituted 1,3,5-triazines [258].

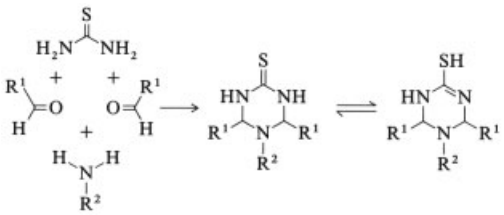

Thiosemicarbazide undergoes a reaction similar to that of thiourea, and can be transformed with aliphatic or aromatic carboxylic acids into 3-mercapto-1,2,4-triazoles [259].

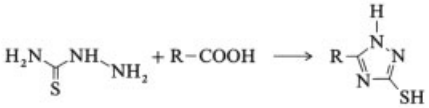

Thiocyanates. Potassium thiocyanate is a frequently employed reagent that lends itself to cyclization with aminocarbonyl compounds to 2-mercaptoimidazoles [260].

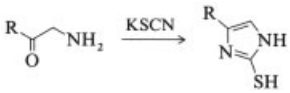

Isothiocyanates react with acylhydrazides in the presence of sodium alcoholate to give 3-mercapto-1,2,4-triazoles [261].

Other cyclization processes are described in [262].

3.3.3. Heterocyclic Sulfides and Disulfides

The synthetic methods applicable to heterocyclic sulfides and disulfides correspond largely to those for aliphatic and aromatic compounds (see Sections 1.6 and 2.4). Thus, heteroaromatic sulfides can be prepared just like their aromatic analogues by heteroarylation of such inorganic compounds as sodium sulfide or dichlorosulfane [263].

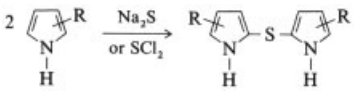

Unsymmetrical sulfides can be obtained from heteroaryl halides by treatment with sodium alkanethiolates or thiols [264]. Good yields are also observed in the palladium-catalyzed reaction of heteroaromatic halogen compounds with organotin sulfides [265].

Sulfenyl chlorides can be transformed into unsymmetrical sulfides with alkenes [266] or arenes [267].

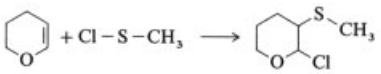

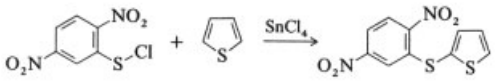

4,5-Dinitroimidazoles are converted in high yield to thioethers through substitution of a nitro group [268].

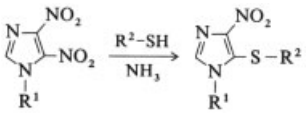

For the preparation of heterocyclic disulfides see the methods described in Section 1.6.3.

3.4. Uses of Heterocyclic Thiols, Sulfides, and Disulfides

Numerous heterocyclic thiols and their sulfur derivatives have found application as intermediates in the synthesis of pharmaceuticals, crop-protection agents, and dyes. Additional applications are described in the sections that follow.

3.4.1. Pharmaceuticals

Many heterocyclic thiols and sulfides are among the active ingredients in commercially available drugs. Table 6 presents a few examples.

A wide range of such substances with antibiotic activity has been developed in the last few years, including systems with skeletons that differ from those of penicillin and cephalosporin. A review of the patent literature is incorporated into [274]. The description that follows provides a few examples of the broad spectrum of pharmaceutical activity associated with heterocyclic thiols and their derivatives.

Mercapto- (or alkylthio-) triazoles [275], imidazoles [276], imidazolines [277], and pyridines [278] inhibit the activity of dopamine-β-hydroxylases. 7-Thiosubstituted oxoquinolinecarboxylic acids [279], 2-amino-3-mercaptopyridines [280], and thiopyridine derivatives of 5-nitro-2-vinylfuran [281] display antibacterial activity, and the latter are also fungicides. Analgesic, antipyretic, and anti-inflammatory activity has been reported for certain thiopyridine derivatives [282], 2-mercaptoimidazoles [283], 3-amino--4-mercapto-6-methylpyridazines [284], and 3-mercaptotriazoles [285]. Mercapto-substituted imidazoles [286] and triazoles [287] have been patented as medicinal fungicides. 2-Mercaptobenzoxazoles can be used for both prophylaxis and treatment in the case of liver diseases [288]. Various heterocyclic thiols and sulfides inhibit the secretion of stomach acid, and can be utilized in ulcer treatment [289]. Mercaptopyrrolidine derivatives have been described as cough-prevention agents [290]. Mercaptopyrazolo-pyrimidines and -triazines represent potential schistosomicides [291]. Furanyl and pyranyl sulfides display immune-controlling characteristics [292]. Thio-substituted pyridazinones [293], imidazoles [294], and other heterocycles [295] display antiallergic and antiasthmatic activity. The gold salt of thioglucose shows in vitro antiviral activity toward the immune-deficiency virus HIV-1 [296].

Table 6. Heterocyclic thiols and sulfides used as pharmaceuticals

Thiol/Sulfide	Application	Ref.
2-Thiobarbituric acid 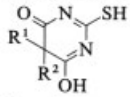	anaesthetic narcotic hypnotic sedative	[269]
6-Thiopurine 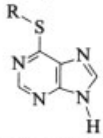	cytostatic agent antileukemic	[270]
2-Thiouracil 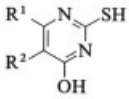	thyreostatic	[271]
Cephalosporin 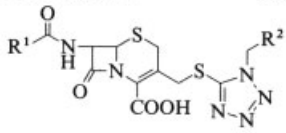	antibiotic	[272]
2-Pyridinethiol-*N*-oxide Na salt Zn salt disulfide	antimycotic antiseborrhoeic fungicide, bactericide	[273]

3.4.2. Agricultural Chemicals

Heterocyclic thiols and sulfides are included as crop-protection agents in numerous commercial products. For example, several herbicides contain methylthio-1,3,5-triazines [297].

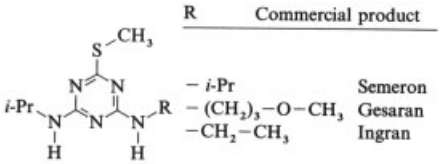

Other effective herbicides of considerable economic significance are the 3-methylthio-1,2,4-triazine-5(4*H*)-ones, which inhibit photosynthesis [298–300].

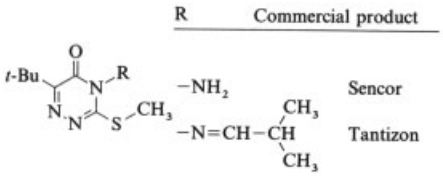

Mercapto-substituted pyrazoles [301] and imidazole-5-carboxylic acids [302] as well as 5-mercaptopyridine-3-carboxylic acid derivatives [303] have also been reported to show herbicidal activity.

A broad spectrum of activity has been reported in the extensive patent literature with respect to 1,2-pyridazinones containing thio substituents. For example, they can be used not only as pesticides in veterinary-medicine and crop-protection applications [304], but also as plant fungicides [305]. Other compounds effective as pesticides are derivatives of 1,2,5-thiadiazol-3-thiol [306] and heteroaryl vinyl sulfides [307]. 1-Oxopyridyl disulfides can be utilized as feed supplements in the raising of poultry [308].

3.4.3. Other Uses

Heterocyclic thiols and their salts are employed in the photographic sector as stabilizers and as protection against fogging. Patents have been issued in this context covering pyrazoles and triazoles [309], thiadiazoles [310], pyrrolidines [311], and benzimidazoles [312] with free mercapto groups, but also heterocyclic sulfides [313]. Other photographic auxiliaries contain mercaptopyridines [314]. 2-Mercaptobenzothiazole and its derivatives have been used to protect copper and copper alloys against corrosion. Corrosion-inhibiting properties have also been reported for 2-mercaptopyridines [315], 4,5-bismercapto-1,2-dithiol-3-one and its salts [316], and certain pyridine thioethers [317].

A few heterocyclic thiols are effective as stabilizers for microbicidal 3-isothiazolones [318].

Mercaptobenzimidazole and its derivatives have proven valuable in preventing the aging of rubber [319]. 2-Mercaptobenzothiazole is an important vulcanization catalyst in the rubber industry. Other compounds suggested for this purpose include mercapto-substituted triazine [320] and pyrimidine [321] derivatives. Derivatives of 2,5-dimercaptothiadiazole [322] as well as aminobis- and aminotris(alkylthio)triazines have been suggested as stabilizers for natural and synthetic rubber [323].

Heterocyclic anthraquinone thioethers have been described as dyes for liquid crystals [324]. Thioethers [325] and disulfides [326] of furans and thiophenes are cited in patent literature as flavoring food additives.

2-Mercaptobenzothiazole plays a role in analysis as a reagent for cadmium as well as for the determination of copper, lead, bismuth, silver, mercury, thallium, gold, platinum, and iridium. 4-Amino-3-hydrazino-5-mercapto-1,2,4-triazole (Purpald, Aldrich) is a reagent specific for aldehydes [327]. Other analytical applications of heterocyclic thiols and sulfides are discussed in [328].

Cosmetic applications include the use of 2-mercaptopyridine to prevent the formation of melanin in the skin [329], and of certain heterocyclic thiols as active ingredients in deodorants [330].

4. Toxicology [331–335]

Most thiols are classed as moderately toxic compounds. Since many thiols and sulfides are readily detectable on the basis of their odors at levels far below the danger point, these odors serve to provide a warning. Exceptions include 2,4,4-trimethyl-2-pentanethiol (*tert*-octanethiol), cyclohexanethiol, and thiophenol, which are significantly more toxic than short-chain thiols. Toxicity decreases with longer-chain thiols.

The SH group plays a significant role in biological metabolism (e.g., in metabolite transfer), and for this reason thiols may exhibit either inhibitory or accelerating effects on metabolic processes. A comprehensive treatment of the metabolic, biochemical, and toxicological aspects of xenobiotics containing thiol or sulfide functions is presented in [333].

The discussion that follows constitutes a brief description of the toxic properties of a few representative thiols, sulfides, and disulfides.

Methanethiol (Methyl Mercaptan). The greatest risk with respect to this compound involves inhalation, and the corresponding toxic properties are comparable to those of hydrogen sulfide. Low concentrations lead to irritation of the eyes and mucous membranes, headache, nausea, and vomiting. Irritation of the respiratory tract may lead to edema in the lungs. Higher concentrations result in increased blood pressure and pulse rate, coma, severe anemia, and death. Inflammation of the skin (dermatitis) is observed to accompany chronic exposure to low concentrations. MAK: 0.5 ppm (1 mg/m^3); TLV: 0.5 ppm.

Ethanethiol (ethyl mercaptan) has shown moderately toxic effects in animal experiments in conjunction with inhalation, oral ingestion, or intraperitoneal application. Low concentrations (4 ppm) lead in the case of humans to irritation of the skin and eyes, headache, and nausea, and respiratory rate also decreases at an exposure level of 50 ppm. Extended exposure to 4 ppm of ethanethiol results in a decrease in odor sensitivity. MAK: 0.5 ppm (1 mg/m^3); TLV: 0.5 ppm.

n-Propanethiol (propyl mercaptan) has been shown in animal studies to display little toxicity upon oral administration, and it is essentially nontoxic with respect to inhalation. Its effect as an irritant to mammalian skin and eyes is also low. On the other hand, human subjects residing adjacent to a field treated with the pesticide Ethioprop (Mocap) displayed health impairments that persisted for as long as six weeks as a result of the release of propanethiol. The symptoms described included headache, sore throat, diarrhea, and irritation of the eyes, as well as hay fever and asthma attacks [336].

Butanethiols. The inhalation of 50–500 ppm of *n-butanethiol* over a period of one hour led in the case of seven human subjects to modest to severe signs of poisoning, including such symptoms as weakness, increased respiratory rate, pain in the neck, nausea, and depression of the central nervous system (diminished reaction capability). MAK: 0.5 ppm (1.5 mg/m^3); TLV: 0.5 ppm. *tert-Butanethiol* has been shown in animal experiments to be somewhat less toxic than the straight-chain isomer. Both butyl thiols are eye irritants. MAK and TLV values for these compounds have not yet been established.

Cyclohexanethiol causes eye irritation and severe skin damage. Animal experiments (rats) involving oral ingestion have revealed a low LD_{50} of 558 mg/kg.

tert-Octanethiol has been found in animal studies to be toxic by oral ingestion, inhalation, or intraperitoneal application, and moderately toxic upon resorption through the skin. The lowest lethal dose (intraperitoneal) is reported for rats at 11 mg/kg.

Dodecanethiols. The toxicity of *n-dodecanethiol* is low. Chronic exposure has been observed to produce skin irritation and sensitization. In the case of *tert-dodecanethiol* (isomeric mixture) animal studies based on oral administration produced a very low LD_{50} of 309 mg/kg. Skin and eye irritation was also noted.

Thiophenol (phenyl mercaptan) has a severe irritating effect on the skin and eyes, and can lead to acute dermatitis. Headache and dizziness may accompany working with the compound. Animal studies have shown that chronic inhalation of high concentrations may cause damage to the liver, lungs, and kidneys. No MAK value has yet been established; TLV: 0.5 ppm. Thiophenol as well as *o-* and *p-aminothiophenols* cause the transformation of oxyhemoglobin into methemoglobin in human erythrocytes [337]. *p-Chlorothiophenol* and *pentachlorothiophenol* cause severe eye irritation, and the former may be carcinogenic.

Alkyl sulfides are generally described as moderately toxic compounds that may lead to hemolytic anemia and allergic dermatitis. *Dimethyl sulfide* is responsible for skin irritation and severe eye irritation. Certain diaryl sulfides [e.g., 2,2′-thiobis(4,6-dichlorophenol) and 4,4′-thiodianiline] are suspected of being carcinogenic.

2-Mercaptobenzothiazole has a sensitizing effect, and with chronic exposure (e.g., through

the use of rubber gloves) it may induce skin reactions [338]. *2-Mercapto-1-methylimidazole* displays teratogenic effects in humans, and it is suspected of being a carcinogen. *6-Mercaptopurine derivatives* influence cell division, and have been employed as cytostatic agents. They have been found to be teratogenic in animal studies. *Methylthio-substituted triazines*, which are used as herbicides, are only slightly toxic. The corresponding LD_{50} or LC_{50} values derived from animal studies are above the level of 1000 mg/kg [339]. Similar values have been established for a methylthiotriazinone that is the active ingredient in the herbicide Sencor [340].

The professional literature should be consulted regarding the diverse toxic side effects of the many pharmacologically active heterocyclic thiols and sulfides.

5. References

General References

1. *Ullmann,* 4th ed., **23,** 175.
2. *Houben-Weyl,* **E11,** 32, 129; **9,** 55.
3. N. Kharasch (ed.): *Organic Sulfur Compounds,* Pergamon Press, New York 1961 – 1966.
4. G. C. Barrett in: *Comprehensive Organic Chemistry,* vol. 3, Pergamon Press, Oxford 1979, p. 3.
5. J. L. Wardell in S. Patai (ed.): *The Chemistry of the Thiol Group,* Wiley-Interscience, London 1974.
6. *Kirk-Othmer,* **22,** 946.
7. H. Goldwhite in: *Rodd's Chemistry of Carbon Compounds,* 2nd ed., vol. IB, Elsevier Publ., Amsterdam 1965, p. 73.
8. D. W. Brown in: *Rodd's Chemistry of Carbon Compounds,* 2nd ed., vol. IB, 2nd Suppl., Elsevier Publ., Amsterdam 1991, p. 344.
9. A. Ohno in S. Oae (ed.): *Organic Chemistry of Sulfur,* Plenum Press, New York 1977, p. 119.
10. S. R. Sandler, W. Karo in H. H. Wasserman (ed.): *Organic Functional Group Preparations,* 2nd ed., vol. 1, Academic Press, New York 1983.
11. *Rodd's Chemistry of Carbon Compounds,* Elsevier Publ., Amsterdam, 2nd ed., vol. 3A, 1971, p. 421; Suppl. 3A, 1983, p. 241.

Specific References

12. A. J. Virtanen, *Angew. Chem.* **74** (1962) 374; *Angew. Chem. Int. Ed. Engl.* **1** (1962) 199.
13. A. E. Johnson, H. E. Nursten, A. A. Williams, *Chem. Ind. (London)* 1971, 556.
14. E. Demole, P. Enggist, G. Ohloff, *Helv. Chim. Acta* **65** (1982) 1785.
15. H. Yanagawa et al., *Tetrahedron Lett.* **13** (1972) 2549 –2552.
16. E. E. Beach, A. White, *J. Biol. Chem.* **127** (1939) 87.
17. T. Janowski, J. Gawlik, S. Zimnal, *VDI Ber.* **226** (1975) 123.
18. E. H. Phelps, *Proc. World Pet. Congr. 7th.,* **9** (1967) 201.
19. W. C. Zeise, *J. Prakt. Chem.* **1** (1834) 257.
20. C.-D. S. Lee, W. H. Daly, *Adv. Polym. Sci.* **15** (1974) 123.
21. H. V. Aposhion, *Annu. Rev. Pharmacol. Toxicol.* **23** (1983) 193. L. Magos, *Br. J. Pharmacol.* **56** (1976) 479.
22. B. L. Sharma et al., *Jpn. J. Pharmacol.* **45** (1987) 295 –302.
23. Pädagogische Hochschule, N. K. Krupskaja, DD 276 481, 1990 (K. Andras, H. R. Hoppe, R. Sziburies, H. Selzer).
24. Sakai Chem. Ind., US 4 873 351, 1989 (K. Fujita, T. Wachi, Y. Ikeda).
25. B. Krebs, G. Henkel, *Angew. Chem.* **103** (1991) 785 –804. J. H. Dawson, M. Sono, *Chem. Rev.* **87** (1987) 1255 – 1276.
26. UOP, US 4 490 246, 1983 (T. A. Verachtert); US 4 498 977, 1983 (R. R. Frame).
27. Co. Française de Raffinage, EP 91 845, 1983 (C. Marty, P. Engelhard).
28. Inst. Française du Petrole, FR 2 560 889, 1984 (P. H. Bigeard, J. P. Franck, P. Ganne, H. Mimoun).
29. Y. M. Torchensky: *Sulfur in Proteins,* Pergamon Press, Oxford 1981. H. Sies, *Naturwissenschaften* **76** (1989) 57.
30. Monsanto, US 4 966 731, 1990 (Y. Chou).
31. Pennwalt, EP 313 939, 1988 (A. Husain, G. A. Wheaton).
32. Atochem North America, EP 424 616, 1990 (A. Husain, G. A. Wheaton).
33. E. Campaigne in N. Kharasch (ed.): *Organic Sulfur Compounds,* **vol. 1,** Pergamon Press, New York 1961, chap. 14.
34. S. Ahmad, J. Iqbal, *Tetrahedron Lett.* **27** (1986) 3791. T. Imamoto, M. Kodera, M. Yokoyama, *Synthesis* 1982, 134. T. Kunieda, Y. Abe, M. Hirobe, *Chem. Lett.* 1981, 1427.
35. *Kirk-Othmer,* **22,** 953. D. J. R. Massy, *Synthesis* 1987, 589.
36. P. Sabatier, A. Mailhe, *C. R. Acad. Sci.* **150** (1910) 1217. R. C. Kramer, E. E. Reid, *J. Am. Chem. Soc.* **43** (1921) 880.

37. H. O. Folkins, E. L. Miller, *Proc. Am. Pet. Inst. Sect. 3* **42** (1962) 188.
38. H. O. Folkins, E. L. Miller, *Ind. Eng. Chem. Prod. Res. Dev.* **1** (1962) 271. A. V. Mahkina et al., *React. Kinet. Catal. Lett.* **36** (1988) 159.
39. Pennwalt, US 3 035 097, 1962 (T. E. Deger, B. Buchholz, R. H. Goshorn).
40. O. Weisser, S. Landa: *Sulphide Catalysts, their Properties and Applications,* Pergamon Press, Oxford 1973, p. 268.
41. P. Bapsères, *Chim. Ind. (Paris)* **90** (1963) 358.
42. Toyo Chem. JP 50 040 511, 1973 (S. Ohno, R. Orita, S. Akita).
43. M. P. Balfe, J. Kenyon, C. E. Scarle, *J. Chem. Soc.* 1950, 3309.
44. Mitsui Toatsu, JA 51 048 606, 1974 (T. Hayakawa, T. Ichikawa).
45. Albright & Wilson, GB 917 921, 1963 (H. Coates, P. A. T. Hoye).
46. T. Nishio, *J. Chem. Soc., Chem. Commun.* 1989, 205.
47. Societé Nationale Elf Aquitaine (Production), Demande brevet français 88 018 79, 1988 (E. Arretz).
48. Pennwalt, US 2 951 875, 1960 (B. Loev, R. H. Goshorn).
49. Pennwalt, US 4 102 931, 1978 (B. Buchholz).
50. Societé Nationale Elf Aquitaine (Production), EP 101 356, 1982 (E. Arretz, A. Mirassou, C. Landoussy, P. Augé).
51. Pennsalt Chem., NL 142 406, 1963.
52. Societé Nationale Elf Aquitaine (Production), FR 2 330 674, 1975 (E. Arretz, A. Pfister).
53. Seitetsu, JP 79 132 551, 1978 (M. Kawamura et al.).
54. Phillips Petroleum, US 3 963 785, 1975 (D. H. Kubicek).
55. J. F. Frantz, K. I. Glass, *Chem. Eng. Prog.* **59** (1963) 68.
56. M. A. Korshunov, SU 518 489, 1974; *Chem. Abstr.* **85** (1977) 159410.
57. Phillips Petroleum, US 3 214 386, 1965 (R. D. Franz, P. F. Warner).
58. Pennwalt, BE 866 426, 1977 (B. Buchholz).
59. Ogawa, JP-Kokai 63 201 162, 1987 (H. Masuda, H. Kikuiri, S. Mihara).
60. R. Dahlbom, *Acta Chem. Scand.* **5** (1951) 690. Soda Aromatic Co., JP 1 052 711, 1987 (T. Shirakawa, K. Kinoshita, K. Suzuki, T. Eto).
61. Phillips, EP 208 323, 1985 (J. S. Roberts).
62. Phillips, US 3 050 452, 1962 (R. P. Louthan).
63. Pennwalt, US 4 140 604, 1978 (D. A. Dimmig).
64. Shell, US 2 411 983, 1946 (W. E. Vaughan, F. F. Rust). Societé Nationale Elf Aquitaine (Production), EP 5 400, 1979 (J. Ollivier, G. Souloumiac, J. Suberlucq).
65. Du Pont, US 2 551 813, 1949 (P. S. Pinkney).
66. Stauffer, US 3 689 568, 1969 (R. J. Eletto).
67. F. W. Stacey, J. F. Harris, *Org. React. (N.Y.)* **13** (1963) 150.
68. Nippon Oils & Fats, US 3 459 809, 1969 (S. Ishida, T. Yoshida). Stauffer, BE 750 555, 1969 (D. J. Martin, E. D. Weil).
69. M. W. Farlow, W. A. Lazier, F. K. Signaigo, *Ind. Eng. Chem.* **42** (1950) 2547. Phillips Petroleum, US 3 994 980, 1976 (D. H. Kubicek).
70. Seitetsu, JP 79 135 747, 1979 (M. Kawamura et al.).
71. Pennwalt, US 4 281 202, 1980 (B. Buchholz, C. B. Welsh, H. C. Henry).
72. Societé Nationale Elf Aquitaine (Production), EP 274 934, 1988 (E. Arretz, P. Augé, A. Mirassou, C. Landoussy).
73. Uniroyal, CA 882 159, 1971 (G. S. Pande).
74. Societé Nationale Elf Aquitaine (Production), US 4 507 505, 1985 (E. Arretz).
75. L. M. Ellis, E. E. Reid, *J. Am. Chem. Soc.* **54** (1932) 1674.
76. J. E. Bittell, J. L. Speier, *J. Org. Chem.* **43** (1978) 1687.
77. A. M. Vasil'tov, B. A. Trofimov, S. V. Amosova, *Zh. Org. Khim.* **19** (1983) 1339; *Chem. Abstr.* **99** (1983) 70162x.
78. Dow Corning, BE 834 487, 1974 (J. L. Speier). A. G. Koscova et al., *Zh. Prikl. Khim. (Leningrad)* **38** (1965) 170.
79. Akzo, US 3 927 085, 1974 (A. G. Zenkel, A. Toth, H. Nagerlein, G. Meyer).
80. Stauffer, US 4 740 623, 1988 (J. B. Heather). Societé Nationale Elf Aquitaine (Production). Demande brevet français 8 804 693, 1988 (Y. Labat).
81. Crown Zellerbach, US 3 326 980, 1964 (D. W. Goheen).
82. Draco, EP 317 540, 1987 (A. R. Hallberg, P. A. S. Tunek).
83. Panmedica, EP 305 264, 1987 (R. Lazaro, G. Cros, J. H. McNeil, J. J. Serrano).
84. L. R. Morgan, US 4 827 016, 1985.
85. Schering, EP 355 784, 1988 (M. F. Czarniecki, L. S. Lehmann). Societé Civile Bioprojet, EP 419 327, 1989 (J. C. Plaquevent et al.). Squibb and Sons, EP 361 365, 1988 (N. G. Delaney).
86. Ono, EP 341 081, 1988 (M. Kawamura, Y. Arai, H. Aishita).
87. Santen, EP 326 326, 1988 (T. K. Morita, S. Mita, T. Iso, Y. Kawashima). Beecham, EP 322 184, 1987 (R. E. Markwell, S. A. Smith, L. M. Gaster).

88. Nihon Medi-Physics, EP 322 876, 1987 (H. Yamauchi et al.). NeoRx, EP 344 724, 1988 (A. R. Fritzberg, A. Srinivasan, D. S. Jones, D. W. Wilkening). Du Pont, EP 279 417, 1987 (P. L. Bergstein, E. H. Cheeseman, A. D. Watson).
89. Proter, EP 288 871, 1987 (P. Braga).
90. Camillo Corvi, EP 343 367, 1988 (P. Braga).
91. Roussel-Uclaf, EP 280 627, 1987 (J. P. Vevert, J. C. Gasc, F. Delevallee, F. Petit).
92. Dainippon, JP 1 149 763, 1987 (T. Mimura et al.).
93. Takeda, JP 64 000 063, 1986 (H. Uneme, H. Mitsuders, T. Uekado).
94. Nitto, JP 63 225 391, 1986 (H. Imazaki et al.).
95. Takeda, EP 421 210, 1989 (K. Itoh, K. Okonogi).
96. Midori, JP 1 061 492, 1987 (H. Nakasumi, T. Kitao, M. Oizumi).
97. M. A. Allakhverdiev, *Neftekhimiya* **30** (1990) 570; *Chem. Abstr.* **114** (1990) 26893r.
98. L'Oreal, EP 368 763, 1988 (J. Maignan, M. Colin, G. Lang).
99. Kurakay, JP 3 246 360, 1987 (M. Kawashima, I. Komura, J. Yamauchi).
100. Firmenich, EP 54 847, 1982 (E. P. Demole, P. Enggist). Givaudan, US 3 896 175, 1972 (D. Helmlinger et al.). Fritzsche Dodge, GB 1 546 283, 1979 (B. J. Willis, F. Fischetti, D. I. Lerner). International Flavors and Fragrances, EP 369 668, 1988 (R. M. Boden, J. A. McGhie). BASF K & F, EP 418 680, 1989 (P. A. Christenson, R. G. Eilerman, P. J. Riker, B. J. Drake).
101. Soda Aromatic, JP 1 052 711, 1987 (T. Shirakawa, K. Kinoshita, K. Suzuki, T. Eto).
102. Ogawa, JP 1 102 019, 1987 (H. Masuda, H. Kikuiri, Y. Shishido, S. Mihara).
103. General Foods, US 4 496 600, 1985 (T. H. Parliment).
104. E. Block, *Spektrum Wissensch.* 1985, 66.
105. H. Schildknecht et al., *Angew. Chem.* **88** (1976) 228.
106. *Römpp Chemie Lexikon,* 9th ed., Thieme Verlag, Stuttgart 1992, p. 2484.
107. E. B. Krein in S. Patai, Z. Rappoport (eds.): *The Chemistry of Sulphur-Containing Functional Groups,* Suppl. S, Wiley & Sons, New York 1993, p. 984.
108. E. Block in S. Patai (ed.): *The Chemistry of Ethers, Crown Ethers, Hydroxyl Groups and Their Sulphur Analogues,* part 1, Suppl. E, Wiley & Sons, New York 1980, p. 539.
109. S. Blechert, *Nachr. Chem. Tech. Lab.* **28** (1980) 577.
110. S. Oae, H. Togo, *Kagaku (Kyoto)* **38** (1983) 48; *Chem. Abstr.* **98** (1983) 197543g.
111. F. Kong, X. Zhou, *Synth. Commun.* **19** (1989) 3143.
112. Kuraray, JP 61 155 365, 1986 (T. Onishi, S. Suzuki, M. Shiono, Y. Fujita).
113. AS Urals Bashkir, SU 1 574 595, 1988 (U. M. Dzhemilev, A. G. Ibragimov, A. B. Morozov, E. G. Galkin); *Chem. Abstr.* **114** (1991) 5808b.
114. Humboldt-Universität Berlin, DD 266 350, 1987 (E. Wenschuh et al.).
115. J. M. Khurana, P. K. Sahoo, *Synth. Commun.* **22** (1992) 1691.
116. Phillips Petroleum, US 4 837 192, 1989 (J. S. Roberts).
117. M. E. Peach in S. Patai (ed.): *The Chemistry of the Thiol Group,* Wiley, New York 1974, p. 771. E. J. Corey et al., *J. Am. Chem. Soc.* **102** (1980) 1436, 3663.
118. P. Crotti et al., *Synlett* 1992, 303.
119. Nippon Mining, JP 4 352 764, 1992 (T. Katagiri, K. Furuhashi).
120. Atochem, EP 446 661, 1991 (B. Buchholz, R. B. Hager, M. J. Lindstrom).
121. Y. Kikugawa, *Chem. Lett.* **1981,** 1157. Y. Kikugawa, JP 5 824 557, 1983; *Chem. Abstr.* **99** (1983) 22095.
122. G. A. Olah, Q. Wang, N. J. Trivedi, G. K. S. Prakash, *Synthesis* 1992, 465.
123. Reviews: J. Drabowicz, T. Numata, S. Oae, *Org. Prep. Proced. Int.* **9** (1977) 63. M. Madesclaire, *Tetrahedron* **44** (1988) 6537.
124. R. Balicki, *Synthesis* 1991, 155.
125. J. Drabowicz, B. Dudzinski, M. Mikolajczyk, *Synlett* 1992, 252.
126. H. Bartsch, T. Erker, *Tetrahedron Lett.* **33** (1992) 199.
127. T. J. Wallace et al., *Ind. Eng. Chem. Process Des. Dev.* **3** (1964) 237.
128. Phillips Petroleum, EP 316 942, 1989 (S. H. Presnall).
129. Shell, EP 288 104, 1988 (E. M. Van Kruchten, C. M. J. Leenaars).
130. Phillips Petroleum, US 4 277 623, 1981 (D. H. Kubicek).
131. M. Goodrow et al., *Tetrahedron Lett.* **23** (1982) 3231. J. Nakayama et al. *Sulfur Lett.* **1** (1982) 25. J. Drabowicz, M. Mikolajczyk, *Synthesis* 1980, 32.
132. F. S. Jargensen, J. P. Snyder, *J. Org. Chem.* **45** (1980) 1015.
133. Organic Chem. Tech. Inst., SU 1 712 358, 1992 (O. E. Orlov, L. V. Kaabak, B. A. Kusnetsov, Y. I. Baranov).

134. A. Pryor, D. F. Church, C. K. Govindan, G. Crank, *J. Org. Chem.* **47** (1982) 156.
135. Waku Pure Chem. Ind., JP 59 172 456, 1984 (S. Daikiyo, Y. Kin, D. Fukushima).
136. Waku Pure Chem. Ind., JP 59 172 457, 1984 (S. Daikiyo, Y. Kin, D. Fukushima). S. Oae et al., *J. Chem. Soc., Perkin Trans. 1* 1978, 913.
137. A. Cornélis, N. Depaye, A. Gerstmans, P. Laszlo, *Tetrahedron Lett.* **24** (1983) 3103.
138. L. Field, W. S. Hanley, I. McVeigh, *J. Org. Chem.* **36** (1971) 2735. Bayer, DE 3 014 717, 1981 (R. Schubart, U. Eholzer, T. Kempermann, E. Roos).
139. Rhein Chemie, EP 25 944, 1979 (K. Morche, K. Nuetzel, M. Sauerbier, K. Schilling).
140. Nippon Soda, WO 83/2 773, 1983 (Y. Koyama, H. Nakazawa, T. Yamashita, H. Moriyama).
141. Electricité de France, FR 2 669 042, 1992 (G. Le Guillanton, T. Do Quang, J. Simonet).
142. Societé Nationale Elf Aquitaine (Production), EP 356 318, 1990 (Y. Vallee, Y. Labat).
143. Nippon Soda, WO 90/543, 1990 (S. Tazaki et al.).
144. Squibb, US 4 888 424, 1989 (J. E. Sundeen, D. Floyd, V. G. Lee). Troponwerke, DE 3 012 000, 1981 (H. Jacobi, W. Opitz).
145. H. C. Brown, S. U. Kulkarni, *J. Org. Chem.* **44** (1979) 2417, 2422.
146. Nippon Miktron, JP 1 117 856, 1989 (T. Hiramitsu).
147. R. Adams, C. S. Marvel, *Org. Synth. Coll.* **1** (1964) 504.
148. Nippon Soda, JA 3 075 165, 1961.
149. Sumitomo Seika, JP 5 186 418, 1991 (M. Suzuki, N. Kitagishi, A. Fujisawa, J. Ookawa).
150. Yaroslav Polytechnic Inst., SU 1 421 736, 1987 (E. M. Alov et al.); *Chem. Abstr.* **110** (1988) 192428.
151. Jenapharm, DD 251 343, 1986 (G. Langbein, H. J. Siemann, S. Berbig, A. Bocker). Sumitomo Seika, JP 3 153 664, 1989 (T. Morishita, M. Sato, T. Yagi, K. Kaneda).
152. Sumitomo Seika, JP 3 170 456, 1989 (T. Morishita, M. Sato, N. Kitagishi, A. Onoe).
153. K. Itabashi, *Yuki Gosei Kagaku Kyokaishi* **19** (1961) 601; *Chem. Abstr.* **55** (1961) 23412.
154. Ciba-Geigy, EP 2 755, 1978 (H. Thies, F. von Kaenel).
155. F. Vögtle, R. G. Lichtenthaler, M. Zuber, *Chem. Ber.* **106** (1973) 719. Bayer, DE 1 643 385, 1967 (H. Leuchs, H. W. Hammen).
156. A. W. Wagner, *Chem. Ber.* **99** (1951) 375. J. Morgenstern, R. Bayer, *Z. Chem.* **8** (1968) 106.
157. Uniroyal, NL 6 611 430, 1967.
158. Waku Pure Chem. Ind., JP 63 054 355, 1989 (S. Daikyo, K. Fujimori, H. Togo).
159. O. Weisser, S. Landa: *Sulphide Catalysts, their Properties and Applications,* Pergamon Press, Oxford 1973, p. 193.Seitetsu, EP 474 259, 1991 (M. Kawamura et al.).
160. Bayer, FR 2 008 330, 1970.
161. Nippon Kasei, JP 62 209 057, 1986 (M. Saito, R. Sato).
162. Konica, JP 62 226 957, 1986 (S. Sugita, K. Masuda, S. Nakagawa).
163. W. Rundel, *Chem. Ber.* **101** (1968) 2956.
164. Phillips Petroleum, US 4 005 149, 1977 (D. H. Kubicek).
165. C. C. Price, G. W. Stacy, *J. Am. Chem. Soc.* **68** (1946) 498.
166. S. Krishnamurthy, D. Aimino, *J. Org. Chem.* **54** (1989) 4458.
167. R. A. Amos, S. M. Fawcett, *J. Org. Chem.* **49** (1984) 2637.
168. Bayer, DE 2 903 505, 1979 (H. Hagemann, E. Klauke, G. M. Petruck). F. G. Bordwell, H. M. Anderson, *J. Am. Chem. Soc.* **75** (1953) 6019.
169. Y. Takikawa, S. Takizawa, *Nippon Kagaku Kaishi* 1972, 756; *Chem. Abstr.* **77** (1972) 5081.
170. S. K. Jain, D. Chandra, R. L. Mital, *Chem. Ind. (London)* 1969, 989.
171. K. R. Langgille, M. E. Peach, *J. Fluorine Chem.* **1** (1972) 407.
172. Bayer, US 3 560 573, 1971 (M. Blazejak, J. Haydn).
173. ICI, DE 2 347 462, 1972 (D. M. Dawson, D. J. Harper).
174. Sumitomo, JP 4 182 463, 1990 (M. Suzuki, H. Hata, M. Yoshikawa, S. Oe).
175. Cassella, EP 409 009, 1990 (W. Bauer, W. Steckelber). Hoechst, DE 2 127 898, 1971 (S. Planker, K. Baessler). Bayer, DE 1 670 760, 1966.
176. Mitsui Toatsu, JP 5 117 225, 1991 (T. Kobayashi et al.). Ricoh, JP 4 074 160, 1990 (R. Momiyama, K. Nagai).
177. Phillips Petroleum, US 4 978 796, 1990 (J. E. Shaw). J. E. Shaw, *J. Org. Chem.* **56** (1991) 3728.
178. K. Takagi, *Chem. Lett.* **1985,** 1307. Nippon Chem. Ind., JP 62 201 862, 1986 (K. Takagi).
179. Konishiroku Photo, JP 62 036 354, 1985 (S. Sugita, K. Masadu, S. Nakagawa). Konishiroku Photo, JP 62 067 063, 1985 (S. Sugita, K. Masuda, S. Nakagawa).
180. Sumitomo, JP 5 140 086, 1992 (M. Suzuki, N. Kitagishi, S. Kimura, A. Fujisawa).

181. Monsanto, US 4 088 698, 1978 (N. A. Fishel, D. A. Gross). Mitsubishi Petrochem., JP 8 036 409, 1980 (A. Sakurada, N. Hirowatari).
182. W. E. Truce, J. J. Breiter, *J. Am. Chem. Soc.* **84** (1962) 1621. A. Ferretti, *Org. Synth. Coll.* **5** (1973) 419.
183. Sumitomo, AU 8 814 024, 1988 (T. Haga, S. Hashimoto, E. Nagano, R. Yoshida).
184. Soc. France Hoechst, EP 368 697, 1989 (Y. Christidis).
185. Mitsubishi Gas Chem., JP 2 075 656, 1988 (K. Okabe, S. Ametani, S. Murabayashi).
186. Beecham, EP 358 305, 1990 (I. Hughes).
187. Commonwealth of Australia, WO 87/01724, 1986 (W. Mazurek).
188. Naarden, EP 219 901, 1986 (P. J. Devalois, H. J. Wille, H. L. Vandenheuv).
189. Consiglio Nat. Ricerche, DE 3 601 286, 1986 (G. Bornengo, F. Carlini, A. Marabini, V. Alesse).
190. Showa Denko, JP 3 212 430, 1990 (T. Arakawa, N. Minorikawa, S. Maruyama, H. Takoshi). Mitsui Toatsu, JP 63 165 392, 1986 (K. Sasagawa, M. Imai).
191. Matsushita, EP 569 037, 1993 (Y. Sato, T. Sotomura, K. Takeyama, H. Uemachi).
192. Bayer, EP 408 891, 1989 (H. L. Elbe et al.).
193. Bayer, EP 302 321, 1988 (H. Hagemann, K. Sasse, R. Fischer).
194. Ciba-Geigy, EP 258 190, 1987 (C. Frey, B. Dill). BASF, DE 1 593 769, 1967 (H. Eilingsfeld, H. Scheuermann).
195. Hoechst, DE 4 131 355, 1991 (J. Dannheim). Chisso, JP 3 287 573, 1990 (T. Onishi).
196. Nisso Petrochem Ind., JP 62 077 361, 1985 (M. Hoyko, K. Kokoma, T. Hida). Fuji Photo Film, JP 2 286 655, 1989 (T. Kawagishi, Y. Shimada, H. Tamai).
197. DuPont, US 4 792 633, 1986 (P. W. Wojtkowski). Societé Nationale Elf Aquitaine, EP 318 394, 1988 (C. Vottero, Y. Labat, J. M. Poirier).
198. Ethyl Corp., US 4 670 597, 1985 (P. F. Ranken, R. L. Favis). Ethyl Corp., WO 92/15560, 1992 (D. E. Ballhoff, J. F. Ballhoff).
199. Shell, EP 411 720, 1990 (A. C. G. Gray, T. W. Naisby, S. J. Turner, T. M. Machin). Sumitomo, JP 5 039 257, 1991 (S. Tsuyunashi, K. Kumagai, N. Nabeshima, M. Kato). BASF, EP 379 098, 1990 (H. Wingert et al.).
200. Stauffer, ICI, EP 249 150, 1987 (C. G. Knudsen). Nihon Bayer, JP 5 117 109, 1991 (T. Goshima et al.). BASF, EP 402 751, 1990 (U. J. Vogelbacher et al.). ICI, GB 2 225 014, 1989 (D. Cartwright). Rhône-Poulenc, EP 351 332, 1989 (P. Desbordes, M. Euvrard).
201. Dow Chem., WO 91/11503, 1991 (C. J. Deford et al.).
202. Hoechst, EP 336 199, 1989 (M. Kern et al.). Kureha, EP 255 935, 1987 (S. Kato, T. Hayaoka).
203. Sumitomo, JP 5 229 905, 1992 (T. Takagaki, H. Kishida, K. Umeda, T. Ishiwatari).
204. Plant Science Res. Inst., SU 1 632 393, 1980 (M. A. Fedin, T. K. Alsing, T. A. Kuznetsova).
205. Ricoh, JP 62 011 676, 1985 (K. Taniguchi, H. Furuya). Fuji Photo Film, DE 3 704 627, 1987 (M. Takashima, M. Satomura, K. Ichikawa).
206. ICI, EP 427 404, 1990 (P. Mistry, P. Patel).
207. Mitsubishi, EP 424 144, 1990 (Y. Tamura, F. Watari). Toray, Green Cross, JP 2 289 622 (K. Kato, C. Ohama).
208. Bayer, EP 570 798, 1993 (K. Hentschel, U. Walter). Ethyl Corp., US 4 757 119, 1987 (P. L. Wiggins, G. G. Knapp).
209. Sumitomo, JP 4 031 489, 1990 (H. Hayami). Bayer, DE 3 938 422, 1989 (B. Kohler, K. Stahlke, A. Sommer, K. Sommer).
210. Ciba-Geigy, EP 307 358, 1988 (H. O. Wirth, H. H. Friedrich).
211. Ciba-Geigy, EP 224 442, 1986 (H. R. Meier, G. Knobloch). Ethyl Petroleum, US 4 946 610, 1989 (W. Y. Lam, G. P. Liesen). Lubrizol, WO 88/09804, 1988 (M. F. Salomon).
212. Clairol, Bristol Myers, US 4 799 934, 1987 (M. I. Lim, J. S. Anderson).
213. Phillips Petroleum, US 5 132 008, 1991 (M. D. Bishop, J. E. Shaw).
214. SmithKline Beecham, EP 327 306, 1989 (J. G. Gleason et al.). Searle, EP 293 895, 1988 (R. A. Mueller, R. A. Partis). Ciba-Geigy, EP 228 045, 1986 (A. Vonspreche et al.). Lilly, GB 2 184 121, 1986 (S. R. Baker, A. Todd). Kanegafuchi, EP 211 363, 1986 (T. Shiraishi et al.).
215. BASF, EP 386 452, 1990 (B. Janssen, H. H. Wuest).
216. Green Cross, EP 120 426, 1988 (Y. Naito et al.).
217. Taisho, JP 4 145 062, 1990 (K. Tomizawa et al.).
218. Bayer, Miles, US 5 214 078, 1992 (G. Fennhoft et al.).
219. Kennelley, CA 2 058 266, 1991 (K. J. Kennelley et al.).
220. Lubrizol, US 4 894 174, 1988 (M. F. Salomon).
221. Hüls, DE 3 941 002, 1989 (G. Horpel, K. H. Nordsiek).

222. Bayer, EP 23 313, 1979 (K. Idel, J. Merten). Hay, US 5 239 052, 1990 (A. S. Hay, Z. Y. Wang).
223. F. Rossi, G. Nardi, A. Palumbo, G. Prota, *Comp. Biochem. Physiol. B Comp. Biochem.* **80B** (1985) 843. E. Turner, R. Klevit, L. J. Hager, B. M. Shapiro, *Biochemistry* **26** (1987) 4028.
224. E. Turner, L. J. Hager, J. Lisa, B. M. Shapiro, *Science (Washington D.C.)* **242** (1988) 939. T. P. Holler, P. B. Hopkins, *J. Am. Chem. Soc.* **110** (1988) 4837.
225. U. Gasser, W. Grosch, *Z. Lebensm. Unters. Forsch.* **190** (1990) 3. G. MacLeod, J. Ames, *Chem. Ind. (London)* 1986, 175.
226. U. Gasser, W. Grosch, *Z. Lebensm. Unters. Forsch.* **190** (1990) 511.
227. J. Blank, A. Sen, W. Grosch, *Z. Lebensm. Unters. Forsch.* **195** (1992) 239.
228. D. A. Withycombe, C. J. Mussinan, *J. Food Sci.* **53** (1988) 658, 660.
229. P. Werkhoff et al., *Chem. Mikrobiol. Technol. Lebensm.* **13** (1991) 30.
230. W. J. Burrows et al., *J. Biochem.* **9** (1970) 1867.
231. T. Wieland et al., *Liebigs Ann. Chem.* **12** (1981) 2318.
232. A. R. Katritzky, J. M. Lagowski in A. R. Katritzky (ed.): *Advances in Heterocyclic Chemistry,* **vol. 1,** Academic Press, New York 1963, p. 396. A. R. Katritzky, *Chimia* **24** (1970) 134.
233. Bayer, DE 1 108 692, 1958 (E. Kracht, F. Eichler, J. Farana).
234. Y. H. Kim, G. H. You, *Bull. Korean Chem. Soc.* **2** (1981) 122.
235. G. Pettit, E. E. van Tamelen, *Org. React. (N.Y.)* **12** (1962) 356. Y. Nakayama et al., *J. Heterocycl. Chem.* **18** (1981) 631.
236. Y. Sanemitsu et al., *J. Heterocycl. Chem.* **18** (1981)1053. S. D. Aster, S. S. Yang, G. D. Berger, *Org. Prep. Proced. Int.* **23** (1991) 658.
237. Monsanto, FR 1 549 002, 1969 (J. J. d'Amico). D. I. Banks, P. Wiseman, *J. Appl. Chem.* **18** (1968) 202. E. Kühle: *The Chemistry of Sulfenic Acids,* G. Thieme Verlag, Stuttgart 1972. Monsanto, DE 2 201 989, 1972 (J. J. d'Amico).
238. M. D. Bentley et al., *J. Chem. Soc., Chem. Commun.* 1971, 1625.
239. E. C. Taylor, W. A. Ehrhart, *J. Am. Chem. Soc.* **82** (1960) 3138. H. Hartmann, S. Scheithauer, *J. Prakt. Chem.* **311** (1969) 827.
240. Nissan Chem. Ind., JP 62 004 272, 1987 (T. Sato, K. Morimoto, S. Yamamoto). Nissan Chem. Ind., JP 61 109 777, 1986 (T. Ogura, Y. Kawamura, H. Suzuki). Ricoh, JP 3 048 661, 1989 (K. Nagai). V. Konecny, S. Kovac, *Chem. Pap.* **41** (1987) 827.
241. Ishihara Sangyo Kaisha, JP 63 203 632, 1987 (Y. Tsujii, H. Uenishi, T. Kimura). Ishihara Sangyo Kaisha, JP 1 275 562, 1988 (R. Nasu, S. Kimura, A. Mori, R. Koto).
242. T. Kamiyama, S. Enomoto, M. Inoue, *Chem. Pharm. Bull.* **33** (1985) 5184.
243. Ricoh, JP 4 049 282, 1990 (K. Nagai, R. Momyama).
244. E. Jones, I. M. Moodie, *Org. Synth.* **50** (1970) 104. J. Cymerman-Craig, J. W. Loder, *Org. Synth. coll. vol.* **V** (1973) 667.
245. Bayer, DE 3 545 124, 1985 (A. Klausener, G. Fengler).
246. C. Corral, J. Lissavetzky, A. S. Alvarez-Insua, A. M. Valdeolmillos, *Org. Prep. Proced. Int.* **17** (1985) 163. Takeda Chem. Ind., EP 220 695, 1987 (I. Aoki, T. Kuragano, N. Okajima , Y. Okada).
247. R. Gompper, W. Elser, *Org. Synth. coll. vol.* **V** (1973) 780. R. F. Langler et al., *Can. J. Chem.* **49** (1971) 481. SmithKline Beckman, EP 308 159, 1987 (L. I. Kruse, S. T. Ross).
248. S. D. Aster, S. S. Yang, G. D. Berger, *Org. Prep. Proced. Int.* **23** (1991) 658.
249. S.-O. Lawesson, *Nouv. J. Chim.* **4** (1980) 47. K. Clausen, S.-O. Lawesson, *Bull. Soc. Chim. Belg.* **88** (1979) 305 Poli Ind. Chim., EP 515 995, 1992 (S. Poli, G. Coppi , G. Signorelli).
250. Waku Pure Chem. Ind., JP 1 230 557, 1988 (S. Daikyo, K. Fujimori, H. Togo).
251. C. Ferri: *Reaktionen der organischen Synthese,* G. Thieme Verlag, Stuttgart 1978.
252. J. A. van Allan, B. D. Deacon, *Org. Synth. coll. vol.* **IV** (1963) 569.
253. C. F. H. Allen, C. O. Edens, J. A. van Allan, *Org. Synth. coll. vol.* **III** (1955) 394. SmithKline Beckman, WO 89/6126, 1989 (L. I. Kruse, T. B. Leonard, S. T. Ross).
254. E. C. Taylor, R. N. Warrener, A. McKillop, *Angew. Chem.* **78** (1966) 333. H. J. Kabbe, *Synthesis* 1972, 268.
255. E. Lieber, E. Ofterdahl, C. N. R. Rao, *J. Org. Chem.* **28** (1963) 194. C. Cristophersen, *Acta Chem. Scand.* **25** (1971) 1162.
256. Taisho Pharmaceut. Co., JP 2 083 372, 1988 (Y. Yoshikawa et al.).
257. H.-C. Chiang, J.-R. Ko, *Heterocycles* **19** (1982) 529. Dynamit Nobel, EP 46 856, 1980 (H. Peeters, W. Vogt). H. Petersen, *Synthesis*

1973, 258. D. G. Crosby, R. B. Berthold, H. E. Johnson, *Org. Synth. coll. vol.* **V** (1973) 703. H. M. Foster, H. R. Snyder, *Org. Synth. coll. vol.* **IV** (1963) 638.
258. H. Petersen, *Synthesis* 1973, 251.
259. C. Ainsworth, *Org. Synth. coll. vol.* **V** (1973) 1070.
260. SmithKline Beckman, US 4 798 843, 1987 (L. I. Kruse). Ciba-Geigy, Janssen Pharmaceutica, EP 277 387, 1986 (G. R. E. Van Lommen et al.). Janssen Pharmaceutica, EP 275 603, 1986 (W. R. Lutz, W. G. Verschueren, H. Fischer, G. R. E. Van Lommen).
261. SmithKline Beckman, EP 359 505, 1988 (L. I. Kruse, J. A. Finkelstein). SmithKline Beckman, EP 323 737, 1978 (L. I. Kruse, J. A. Finkelstein).
262. Akad. Wissenschaft DDR, DD 252 372, 1986 (A. Rumler et al.). N. Okajima, Y. Okada, *Synthesis* 1989, 398. Technische Universität Dresden, DD 224 320, 1984 (L. Jakisch, H. Viola, R. Mayer). Karl-Marx-Universität Leipzig, DD 276 283, 1988 (D. Briel et al.).
263. M. J. Broadhurst, R. Grigg, A. W. Johnson, *J. Chem. Soc., Perkin Trans. 1* 1972, 1124.
264. R. A. Rossi, S. M. Palacios, *J. Org. Chem.* **46** (1981) 5300. Taiho Yakuhin, JP 56 100 777, 1980.G. M. Kulkarni, V. D. Patil, *J. Indian Chem. Soc.* **67** (1990) 693. Ube Ind., EP 283 271, 1988 (H. Yoshioka et al.).J. Matsumoto et al., *Chem. Pharm. Bull.* **38** (1990) 2190.
265. C. Jixiang, G. T. Crisp, *Synth. Commun.* **22** (1992) 683.
266. S. Z. Ivin, V. K. Promonenkov, *Zh. Obshch. Khim.* **37** (1967) 489; *Chem. Abstr.* **68** (1968) 114701.
267. R. D. Schütz, W. L. Fredericks, *J. Org. Chem.* **27** (1962) 1301.
268. BASF, EP 293 742, 1988 (H. Köhler, T. Dockner).
269. Abbott, US 2 876 225, 1956 (G. H. Donnison). E. Miller et al., *J. Am. Chem. Soc.* **58** (1936) 1090. G. Renzoni, *Fac. Farm.* **36** (1965) 42. M. Steinmann et al., *J. Med. Chem.* **16** (1973) 1354. Lilly, US 2 561 689, 1949.
270. Burroughs Wellcome, US 3 056 785, 1960 (G. B. Elion, G. H. Hitching). P. Nuhn et al., *Pharmazie* **34** (1979) 285.
271. A. Kleemann, J. Engel: *Pharmazeutische Wirkstoffe,* **vol. 5,** G. Thieme Verlag, Stuttgart 1982. D. A. McGimty, W. G. Bywater, *J. Pharmacol.* **84** (1945) 342. Parke & Davis, US 2 153 711, 1934. Ciba, US 2 666 764, 1951 (A. E. Lanzilotti, A. C. Shabica, J. B. Ziegler). G. W. Anderson et al., *J. Am. Chem. Soc.* **67** (1945) 2197.
272. P. Actor et al., *J. Antibiot.* **28** (1975) 471. Eli Lilly, US 3 928 592, 1974 (J. M. Greene, J. M. Indelicato).Bristol-Myers, US 3 796 717, 1972 (R. Smith, L. Cheney).Takeda Chem. Ind., DE 2 715 385, 1977 (M. Ochiai, A. Morimoto, Y. H. Matsushita).Warner-Lambert, EP 15 771, 1979 (T. H. Haskell, D. Schweiss, T. F. Mich, T. P. Culbertson).Sankyo, DE 2 455 884, 1974 (H. Nakao et al.).H. Nakao, *J. Antibiot.* **32** (1979) 320. Eli Lilly, US 3 943 126, 1973 (C. W. Ryan, W. B. Blanchard).
273. Olin Mathieson, US 2 745 826, 1953 (M. A. Dolliver, S. Semnoff). Olin Mathieson, GB 761 171, 1956.
274. Fujisawa, EP 272 456, 1987 (M. Murata et al.). Sanraku-Ocean, JP 60 163 882, 1984 (T. Yoshioka et al.). Rhône-Poulenc, EP 248 703, 1987 (J. C. Barriere, C. Cotrel, J. M. Paris). Schering, EP 238 285, 1987 (M. P. Kirkup, A. S. Boland). Fujisawa, EP 111 281, 1983 (T. Takao et al.). Meiji Seika Kaisha, EP 209 751, 1986 (T. Tsuruoka et al.). Banyu, JP 58 154 588, 1982. Bristol-Myers, DE 3 312 533, 1983 (C. U. Kim, P. F. Misco, Jr.). Hoffmann-LaRoche, US 4 348 518, 1978 (M. Montavon, R. Reiner). Asahi, EP 68 403, 1982 (J. Nishikido, E. Kodama, M. Shibukawa).
275. SmithKline Beckman, EP 323 737, 1987 (L. I. Kruse, J. A. Finkelstein). SmithKline Beckman, EP 359 505, 1988 (L. I. Kruse, J. A. Finkelstein).
276. SmithKline Beckman, US 4 798 843, 1987 (L. I. Kruse). SmithKline Beckman, EP 371 730, 1988 (L. I. Kruse, S. T. Ross). SmithKline Beckman, US 4 749 717, 1987 (L. I. Kruse).
277. SmithKline Beckman, WO 89/6126, 1987 (L. I. Kruse, T. B. Leonard, S. T. Ross).
278. SmithKline Beckman, EP 308 159, 1987 (L. I. Kruse, S. T. Ross).
279. J. Matsumoto et al, *Chem. Pharm. Bull.* **38** (1990) 2190.
280. Bayer, DE 3 545 124, 1985 (A. Klausener, G. Fengler).
281. D. Vegh, J. Kovac, M. Kriz, CS 223 450, 1982; *Chem. Abstr.* **105** (1986) 226324g.
282. Degussa, EP 149 088, 1984 (G. Scheffler et al.).
283. Taisho, JP 2 083 372, 1988 (Y. Yoshikawa et al.).
284. R. Bluth, *Pharmazie* **37** (1982) 285, 441. R. Bluth, DD 150 150, 1980.

285. Warner-Lambert, US 4 962 119, 1989 (D. H. Boschelli et al.).
286. Sumitomo, EP 54 974, 1981 (I. Saji).
287. Pfizer, EP 91 309, 1983 (K. Richardson, P. J. Whittle).
288. Yamanouchi, JP 62 039 583, 1985 (H. Iwamoto, A. Tanaka, K. Nishikiori, H. Ikadai).
289. Wyeth, John & Brothers, US 4 374 992 (R. Crossley).Sankyo, JP 3 014 566, 1989 (A. Yoshida, K. Oda, K. Tabada).Ohta, JP 60 218 375, 1984 (Y. Hasegawa et al.).
290. Proter, EP 184 861, 1985 (L. Dall'Asta, G. Coppi, M. E. Scevola). Poli Ind. Chim., EP 515 995, 1992 (S. Poli, G. Coppi, G. Signorelli).
291. M. R. H. Elmoghayar, *Arch. Pharm. (Weinheim, Ger.)* **316** (1983) 697.
292. Taiho, JP 56 100 777, 1981.
293. Nissan Chem. Ind., JP 61 260 018, 1985 (M. Mutsuto, K. Tanigawa, K. Shikada, R. Sakota).
294. SmithKline Beckman, EP 168 950, 1986 (W. E. Bondinell, D. T. Hill, B. M. Weichman).
295. Yamanouchi, EP 181 779, 1985 (M. Kiyoshi, M. Toshiyasu, H. Hiromu, T. Kenichi).
296. T. Okada, B. K. Patterson, S. Q. Ye, M. E. Gurney, *Virology* **192** (1993) 631.
297. Ciba-Geigy, CH 337 019, 1955. Geigy, DE-OS 1 186 070, 1963. H. Gysin, J. R. Geigy, *Chem. Ind. (London)* 1962, 1393. Geigy, US 3 207 756, 1963 (E. Knusli, W. Stammbach).
298. W. Draber, K. Dichoré, K. H. Büchel, *Naturwissenschaften* **55** (1968) 446.
299. Bayer, US 3 671 523, 1970 (K. Westphal, W. Meiser, L. Eue, H. Hack). Bayer, US 4 058 526, 1976 (W. Merz, G. Schümmer).
300. Bayer, DE-OS 2 407 144, 1974 (R. R. Schmidt, L. Eue, C. Metzger, K. Dichoré).
301. Sankyo, AU 515 639, 1981 (T. Konotsune, K. Kawakubo).
302. Janssen Pharmaceutica, Ciba-Geigy, EP 273 531, 1986 (W. R. Lutz, G. R. E. Van Lommen, V. Sipido, W. G. Verschueren).Janssen Pharmaceutica, Ciba-Geigy, EP 289 066, 1987 (M. F. L. De Bruyn, G. R. E. Van Lommen, W. R. Lutz).
303. Monsanto, JP 62 273 955, 1987 (M. G. Dolson, K. F. Lee, K. L. Spear).
304. Sumitomo, JP 1 180 882, 1988 (H. Tomioka, H. Fujimoto, I. Fujimoto). Ube, EP 283 271, 1988 (H. Yoshioka et al.). Nissan Chem. Ind., JP 1 211 569, 1988 (Y. Nakajima, K. Hirata, M. Kudo). Nissan Chem. Ind., EP 183 212, 1985 (T. Ogura et al.). Nissan Chem. Ind., EP 199 281, 1986 (T. Ogura et al.). Nissan Chem. Ind., EP 232 825, 1987 (Y. Nakajima et al.). Hoechst, DE 3 733 220, 1987 (G. Salbeck et al.). BASF, DE 3 742 266, 1987 (J. Leyendecker et al.). Sumitomo, JP 1 197 479, 1988 (H. Tomioka, H. Fujimoto).
305. Sankyo, EP 89 650, 1983 (H. Takeshiba, T. Kinoto, T. Jojima). Nissan Chem. Ind., WO 9208715, 1991 (Y. Nakajima, J. Watanabe, Y. Hirohara, T. Mita).
306. Schering, DE 3 822 371, 1988 (D. Huebl, E. A. Pieroh, H. Joppien, D. Baumert).
307. Bayer, DE 3 216 416, 1982 (U. Heinemann, P. E. Frohberger, W. Brandes).
308. SmithKline Beckman, EP 104 836, 1983 (G. K. Menon, W. J. Sanders).
309. Konica, Nissan Chem. Ind., JP 63 223 635, 1987 (H. Sakamoto, M. Fujiwara, I. Sakata, S. Yamamoto).
310. Olin, US 4 252 962, 1980 (E. F. Rothgery).
311. Chugai Shashin Yakuhin, JP 62 178 247, 1986 (M. Ito, N. Watanabe, H. Mukai).
312. Fuji Photo Film, GB 2 083 243, 1980 (K. Shiba, S. Nakao, T. Toyama).
313. Agfa-Gevaert, EP 40 771, 1980 (E. Ranz, H.-D. Schütz, J. W. Lohmann). Fuji Photo Film, JP 57 026 848, 1980 (I. Itoh, T. Toyoda, M. Yamada). Eastman Kodak, EP 54 414, 1980 (H. W. Altland, D. D. F. Shiao).
314. Fuji Photo Film, JP 1 292 339, 1988 (M. Goto, K. Morimoto). Mitsubishi Paper Mills, JP 4 012 349, 1990 (T. Miura, H. Seiyama, Y. Tsubakii).
315. F. Zucchi et al., *Werkst. Korros.* **44** (1993) 264.
316. Technische Universität Dresden, DD 246 322, 1986 (G. Reinhard et al.).
317. Ciba-Geigy, EP 330 613, 1989 (E. Phillips, R. C. Wasson).
318. Rohm & Haas, US 5 210 094, 1991 (P. F. D. Reeve).
319. IG Farben, DE 557 138, 1931 (M. Bögemann). DuPont, US 3 895 025, 1973 (A. L. Goodman). Bayer, DE 2 703 313, 1977 (R. Schubart, K. Wagner, K. F. Zenner). Bayer, DE 3 008 159, 1980 (R. Schubart, G. P. Langner).
320. Sankyo, JP 53 027 641, 1976 (Y. Nakamura, K. Mori, A. Umehara). Degussa, DE 2 850 338, 1978 (K. Hentschel, F. Bittner). Bayer, Degussa, DE 2 109 244, 1971 (U. Eholzer, T. Kempermann). H. Ahne et al., *Kautsch. Gummi Kunstst.* **28** (1975) 135.
321. T. Kempermann, *Bayer-Mitt. Gummi-Ind.* **50** (1978) 29; **51** (1979) 25.
322. Toyo Linoleum, JP 56 026 943, 1979 (K. Miyahara, T. Matsukura, I. Niwa). Vanderbilt, US 4 704 426, 1986 (L. A. Doe, Jr.).

323. Geigy, DE 1 470 832, 1961; *Chem. Abstr.* **62** (1965) 11834. Mitsubishi Chem. Ind., JP 56 129 249, 1980 (A. Sugio et al.).
324. Bayer, DE 3 314 467, 1983 (M. Blunck, U. Claussen, F. W. Kroeck, R. Neeff).
325. Haarman und Reimer, DE 3 831 980, 1988 (R. Emberger et al.). Haarman und Reimer, DE 4 016 536, 1990 (R. Emberger et al.).
326. Givaudan-Roure, WO 93/07134, 1992 (U. Huber).
327. R. G. Dickinson, N. W. Jacobson, *J. Chem. Soc., Chem. Commun.* 1970, 1719. H. Brandl, *Prax. Naturwiss. Chem.* **40** (1991) 25.
328. U. Schmidt, G. Pfleiderer, F. Bartkowiak, *Anal. Biochem.* **138** (1984) 217. B. Narayana, M. R. Gajendragad, *Asian J. Chem.* **5** (1993) 121.
329. Sansei Pharmaceut. Co., JP 1 228 908, 1988 (Y. Oyama).
330. Procter & Gamble, EP 483 426, 1990 (R. Chatterjee).
331. R. J. Lewis, Sr. (ed.): *Sax's Dangerous Properties of Industrial Materials,* 8th ed., vol. III, Van Nostrand Reinhold, New York 1992.
332. E. E. Sandmeyer in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2A,** Wiley-Interscience, New York 1981, p. 2061.
333. L. A. Damani (ed.): *Sulphur-Containing Drugs and Related Organic Compounds,* Ellis Horwood Limited, Chichester 1989.
334. *Threshold Limit Values for Chemical Substances and Physical Agents in the Workroom Environment,* American Conference of Governmental Industrial Hygienists (ACGIH), 1978.
335. E. J. Fairchild, H. F. Stokinger, *Am. Ind. Hyg. Assoc. J.* **19** (1958) 171.
336. R. G. Ames, J. W. Stratton, *Arch. Environ. Health* **46** (1991) 213.
337. P. Amrolia, S. G. Sullivan, A. Stern, R. Munday, *J. Appl. Toxicol.* **9** (1989) 113.
338. H. T. H. Wilson, *Br. J. Dermatol.* **81** (1969) 175. A. A. Fischer, *J. Dermatol. Surg.* **1** (1975) 63.
339. Industrieverband Pflanzenschutz e. V. (ed.): *Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmitteln,* Pressehaus Bintz-Verlag, Offenbach 1982.
340. E. Löser, G. Kimmerle, *Pflanzenschutz-Nachr. Bayer* **25** (1972) 186.

Thionyl Chloride → Sulfur Halides

Thiophene

JONATHAN SWANSTON, Degussa Knottingley Limited, Knottingley, Yorkshire, UK

1. **Introduction** 1
2. **Physical and Chemical Properties** . . 1
3. **Resources and Raw Materials** 3
4. **Production** 4
5. **Specifications, Transportation, Health and Safety** 4
6. **Uses** 4
6.1. **Acylated Thiophenes** 5
6.2. **Halogenated Thiophenes** 8
6.3. **Alkylthiophenes** 9
6.4. **Amino- and Hydroxylthiophene Derivatives** 10
6.5. **Reduced Thiophenes** 11
6.6. **Miscellaneous Thiophenes** 11
7. **Toxicology** 12
8. **References** 12

1. Introduction

The simple derivatives of thiophene are stable liquids closely resembling the corresponding benzene derivatives in boiling points and even in odor. Thiophene and its simple alkyl derivatives occur in neutral coal tar distillates, which led to the discovery of thiophene by VICTOR MEYER in 1882. Prior to the discovery a classical test for benzene was the development of a blue color on heating with isatin and concentrated sulfuric acid. During a lecture demonstration by MEYER in 1882 this test failed. Upon investigation it became apparent that the lecture assistant had used benzene made by decarboxylation of benzenecarboxylic acid rather than commercial benzene containing the impurity which was discovered to be a new ring system. The new system was subsequently named thiophene from *theion*, the Greek word for sulfur, and *phaino*, another Greek word meaning shining. The word *phaino* had already been used in naming phenol because of its occurrence in coal tar, a byproduct of the manufacture of gas used in lighting in the 19th century.

Thiophene compounds occur rarely in plants and do not play any part in animal metabolism. However, they have found extensive use in pharmaceuticals and agrochemicals and play a minor role in commerce as constituents of reactive dyes, conductive polymers, ligands for catalyst systems, and in flavor and fragrances. The extensive literature on thiophenes is indicative of the commercial and research interest in this five-membered sulfur heterocycle. Some hundreds of patents appear each year, many using thiophene derivatives as alternatives to benzenoid products, in many cases the thiophene containing molecule exhibiting higher activity than the benzene analogue.

The process to manufacture thiophene (and its homologues) by the gas-phase interaction of aliphatic compounds with sulfur derivatives is capable of unlimited expansion, but thiophene production remains a fine-chemical business in the scale of around 1000 t/a.

This article concentrates on the products of current commerce. Chapter 6. Uses covers over 90 % of the present uses; the remainder consists of minor products, mainly pharmaceuticals, together with several which are still in the very early stages of development.

2. Physical and Chemical Properties

The physical characteristics of thiophene (thiofuran) are closer to benzene than any other five-membered heterocyclic. This enables thiophene to replace a benzenoid moiety in the whole spectrum of aromatic chemicals [1] and, as illustrated in Chapter Uses, sometimes to introduce worthwhile advantages over the analogous benzenoid compounds.

Some physical and physicochemical properties of thiophene are listed below:

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_793.pub2

Boiling point, 100 kPa	84.16 °C
Melting point	−38.30 °C
Density	
d_4^4	1.0873
d_4^{20}	1.06485
d_4^{25}	1.05887
Refractive index	
n_D^{20}	1.52890
n_D^{25}	1.52572
Viscosity	
at 20 °C	0.662 mPa·s
at 25 °C	0.621 mPa·s
Vapor pressure	
at 20 °C	8.4 kPa
at 25 °C	10.61 kPa
at 60 °C	45.32 kPa
at 110 °C	143.24 kPa
at 130 °C	270.07 kPa
Enthalpy of vaporization	32.483 kJ/mol
Critical temperature	317 °C
Critical pressure	4.53 MPa
Dipole moment	0.53 – 0.63 D
Dielectric constant	2.76
Flash point	−6.7 °C
Enthalpy of combustion	−2807 kJ/mol
Enthalpy of formation	82.13 kJ/mol

The numbering of the thiophene ring is orthodox:

Nomenclature can be summarized as

2-Thenyl

2-Thienyl

2-Theonyl

The usefulness of thiophene and its derivatives lies in their reactivity to substitution by electrophilic reagents. Thiophene is a π-electron-rich aromatic with six π-electron distributed over five ring atoms (Fig. 1); this facilitates the attack by electrophiles at ring carbons. The literature illustrates the ease of electrophilic substitution in the 2- and 5-positions (Fig. 2). As in other five-membered heterocyclics, the S atom of the thiophene ring plays its part in the formation of cationic reaction intermediates. Thus, the reactivity of thiophene is more akin to that of phenol and aniline than to the comparatively unreactive benzene ring.

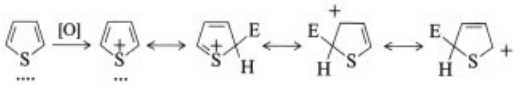

Figure 1. Delocalization of -electrons in the thiophene ring

With the 2-position occupied by an *o*/*p* -directing group, a second group is substituted in the 5-position of the thiophene ring. With a *m*-directing group in the 2-position a second group is substituted in the 4-position.

As can be seen from Figure 2, many 2-substituted derivatives can be produced by orthodox and comparatively mild organic chemical reactions. However, amino, hydroxyl, and thiol substituted thiophenes are unstable unless an electron attracting group is also attached to the thiophene ring, e.g., reduction of 2-nitrothiophene gives 2-aminothiophene which decomposes very rapidly in air but is stable in an inert atmosphere or as a salt.

While 2-substitution is practically 100 %, single-percentage amounts of the 3-isomer may be formed. Routes and reaction conditions to limit this effect need to be optimized.

The order of reactivity toward electrophilic substitution is:

Pyrrole >> thiophene > furan >> benzene e.g., bromination of thiophene is 1.25×10^7 faster than that of benzene; chlorination 1.7×10^7 faster.

Thiophene and alkylthiophenes do not react with nucleophilic reagents by either substitution or addition. Similar to other five-membered heterocycles they only react with strong bases and then by deprotonation of the 2-position carbon. Alkyllithium bases such as butyllithium give a synthetically useful thienyllithium. Because the factors which determine the deprotonation are different to those affecting electrophilic substitution such reactions can be used to produce isomeric products.

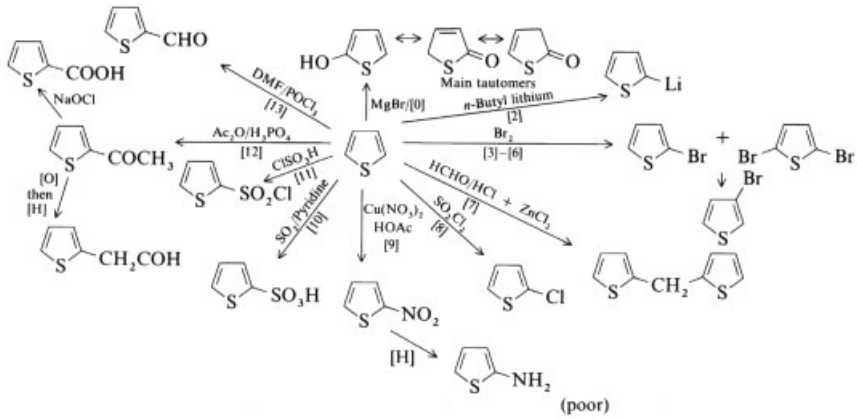

Figure 2. Electrophilic substitution reactions of thiophene

Given the chemistry and availability of commercial thiophene derivatives most 2-substituted thiophenes are obtained by electrophilic substitution of thiophene and subsequent modification of the substituent, whilst 3-substituted thiophenes are made by modification of the two commercially available 3-thiophenes, namely 3-methylthiophene and 3-bromothiophene.

Irradiation causes thiophene to react with H_2O [2]:

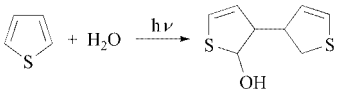

Heating thiophene does not yield C_2H_2, but bithenyl [2]:

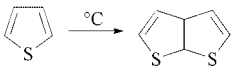

3. Resources and Raw Materials

Thiophene occurs in the light hydrocarbon fraction distilled from coal tar. The proportion of thiophene depends on the type of coal, the temperature of carbonization, and the type of retort. Typical quantities in a tar-derived benzol stream are 0.4 – 1.4 wt % [3].

Traditionally, thiophene is removed from coal tar benzol by washing with H_2SO_4, because the thiophene is more easily sulfonated than the aromatic materials present in the mixture, more rarely thiophene can be extracted by solvent extraction.

The very large volumes of light hydrocarbons produced from coal carbonization plants represent a considerable quantity of associated thiophene, but the quality demands of the principal markets for thiophene require a purity which cannot be obtained economically from coal tar due to the difficulty in separating it from benzene as demonstrated by Meyer.

Commercialization of thiophene for the fine chemicals intermediate market required a synthetic high-purity production method. This was initiated in the 1960s in the United States by Pennsalt/Pennwalt, now Arkema. Midland-Yorkshire Tar Distillers/Synthetic Chemicals, now Degussa, first produced synthetic thiophene on a commercial scale in Europe in 1970.

A key factor in the commercial production of pure thiophene and its homologues is the specificity and cleanliness of the processes used, because often the impurities and byproducts of the reaction are difficult to separate from the main product. Vapor phase (VP) cyclization methods are the processes of choice.

The original VP process, using furan and H_2S as the sulfur source [4], is no longer preferred. Degussa operates its own VP process at its site in Knottingley, UK [5], using CS_2 plus a selected C_4 skeleton. A similar VP process is patented by Arkema [6].

Other, non-VP processes may be described as laboratory or semiscale methods and include: (**1**) syntheses from C_4 unsaturates, using H_2S, Li_2S, SO_2, and S as sulfur sources; (**2**) succinic salts and γ-diketones with, e.g., P_4S_7 or P_4S_{10}; (**3**) thiodiacetates and 1,2-dicarbonyl compounds

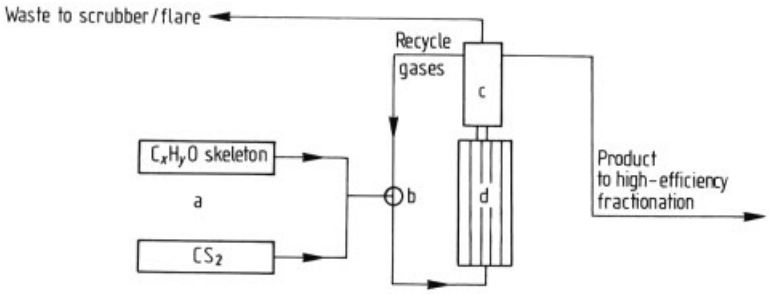

Figure 3. Thiophene production
a) Raw materials tank farm; b) Mixer/flow control; c) Condenser; d) Multitube, fixed-bed VP (vapor phase) reactor

(the Hinsberg synthesis); (**4**) mercaptoacetates and 1,3-dicarbonyl compounds. Such processes often result in very complex reaction products and large amounts of waste making them unsuitable for large-scale production.

4. Production

Since the current international market for thiophene is <1000 t/a, the type of VP installation required for its manufacture is much smaller than a typical petrochemical unit. However, the process is capable of expansion to match any required demand.

Tight operating temperatures of 500–530 °C, catalyst performance, and regeneration plus overall control, particularly with respect to environmental requirements, are key factors in the successful operation of such a unit (Fig. 3).

A typical feedstock portfolio is shown in Figure 4.

The advent of a range of commercially available alkylthiophenes, including those substituted in the difficult 3-, (4-) position, resulted in an alternative raw material to synthesize a range of 3-, (4-) side-chain thiophenes, other than from, e.g., 3-bromothiophene [7].

5. Specifications, Transportation, Health and Safety

Data for thiophene, together with its 2-methyl, 3-methyl, 2-acetyl, and 2-bromo derivatives are given in Table 1.

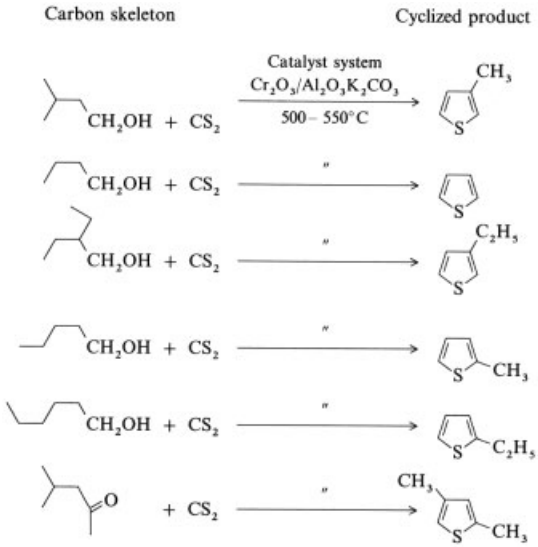

Figure 4. Feedstocks for thiophene and its derivatives

6. Uses

Successful commercial derivatization is governed by the physical and chemical characteristics of the basic thiophene nucleus outlined in Chapter 2.

By far the greatest volume of simple thiophenes are produced by means of VP cyclizations (see Chapter 4). Derivatives of these homologues account for most of this business, in areas as diverse as pharmaceuticals, veterinary drugs, agrochemicals, to dyes, photography, and polymers.

Table 1. Specifications, transportation, health and safety of thiophene and derivatives

	Thiophene	2-Methylthiophene	3-Methylthiophene	2-Acetylthiophene	2-Bromothiophene
Specification					
CAS No.	[*110-02-1*]	[*554-14-3*]	[*616-44-4*]	[*88-15-3*]	[*1003-09-4*]
Appearance	clear liquid	clear liquid	clear liquid	yellow liquid (≥9°C)	clear/darkening
Odor	aromatic	aromatic		aromatic	mild pungent
Purity (GLC), wt %	99	98	98	99	98
Boiling point, °C	83 – 85 (95 %)	113	115	213 – 214 (95 %)	153 – 154
Flash point, °C	-7	15.5	15.5	96	
Miscibility with water	almost immiscible	almost immiscible	almost immiscible	slightly miscible	immiscible
Transportation					
Tariff code	29340 100	293490 900	293490 900	293490 900	293490 900
Classification	class 3.2 highly flammable	class 3.2 highly flammable/irritant	class 3.2 highly flammable/irritant	class 6.1 toxic	class 6.1 toxic
Packaging group	II	II	II	II	II
IMDG page No.	3284	3230	3230	6231	6231
UN (identification)	2414	1993 NOS	1993 NOS	2810 NOS	2810 NOS
ADR	item. 3°(b)	item. 3°(b)	item. 3°(b)	item. 21°	item. 20°(b)
Health and Safety *					
R phrases	R11 R21/22/20 R38/45	R11 R36/37/38 R52/53	R11 R36/37/38	R23 R24	R23/24/25 R36/37/38
S phrases	S9/16/33/43/53/61	S9/16/33/43	S9/16/26/33/43	S18/24/30/41/44	S28/36/37/39/44

* Risk and Safety phrases as legislated by the European Commission and stated fully in United Kingdom. Statutory Instruments 1982 No. 1496, Health and Safety Executive, Sudbury, Suffolk, United Kingdom.

6.1. Acylated Thiophenes

The Friedel-Crafts acylation of thiophene is a much used reaction producing more than 50% of the world market of thiophene derivatives. Use of standard Friedel-Crafts catalysts such as anhydrous aluminum chloride directly with thiophene gives rise to intractable tars and very low yields and as a consequence the catalyst has to be added gradually to the mixture of thiophene and acylating agent. Acylation with acid anhydrides can be catalyzed by strong acids such as phosphoric acid and this method can be used to manufacture 2-acetylthiophene. Vilsmeier formylation proceeds readily to make thiophene-2-aldehyde.

Thiophene 2-aldehyde (T2A), C_5H_4OS M_r 112.15, *bp* 198 °C, is prepared as shown in Figure 2 [8]. It was the first oxygenated derivative to reach a market of several hundred tons per year in the 1970s.

It was originally used to produce a range of antihistamines, including methapyrilene (**1**) [9], methaphenilene (**2**) [10], and thenalidine (**3**) [11]. However, this usage has practically disappeared.

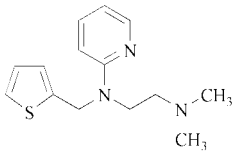

1

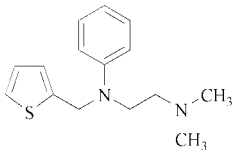

2

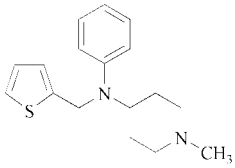

3

The anthelmintic pyrantel (**4**) [12] (→ Anthelmintics) is an important outlet for T2A, enhanced by new formulations and the development of the medicinal use of pyrantel beyond the original veterinary market.

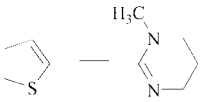

4

An important T2A derivative is the antihypertensive eprosartan (2-thiophenepropionic acid methyl ester) (**5**) [13], which acts as a selective angiotensin II receptor antagonist.

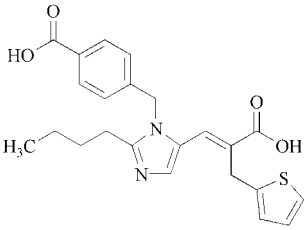

5

Other pharmaceuticals containing T2A are azosemide (**6**) [14] a diuretic, and teniposide (**7**) [15] an antineoplastic.

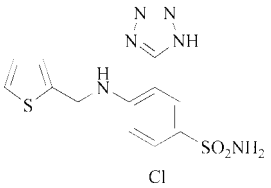

6

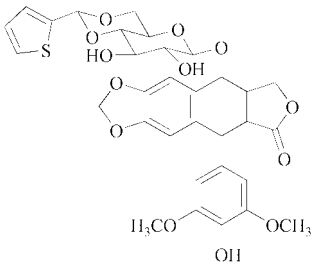

7

2-Acetylthiophene (2AT) C_6H_6OS, *bp* 214 °C, *mp* 9 °C, is prepared as shown in Figure 2. 2AT is used as an important precursor to thiophene-2-acetic acid (T2AA) or its acid chloride (Cl) and thiophene-2-carboxylic acid (T2CA) (Cl). It is important that the process used to synthesize 2AT yields the lowest amount of 3-acetylthiophene isomer.

Thiophene 2-carboxylic acid, T2CA, $C_5H_4O_2S$, M_r 128.15, *bp* 260 °C, *mp* −130 °C, is prepared via oxidation of 2-acetylthiophene (see Fig. 2) using hypochlorite, because oxidation of alkylthiophenes leads to general breakdown of the thiophene ring.

The sodium salt of T2CA is used to treat influenza (trophires) [16]. Other salts, e.g., the Li salt, have minor medicinal uses and the free acid is a mucolytic. A US Patent [17] describes the use of the Na salt as a lubricating oil thickener.

Suprofen (**8**) [18], a nonsteroidal anti-inflammatory (NSAI), and tienilic acid (ticrynafen) (**9**) [19], a potent diuretic, were both very large prospective markets for T2CA or its acid chloride, but were almost completely withdrawn in the 1980s. Sanofi Aventis' NSAI tiaprofenic acid (**10**) was comparatively successful in this market, but although it is still marketed it has subsequently been superceded by alternative products.

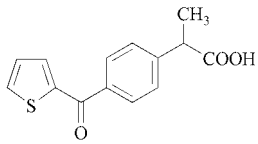

8

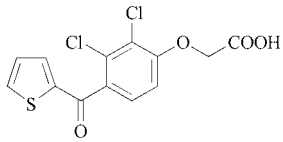

9

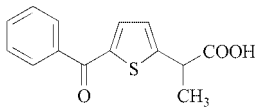

10

The 5-nitro derivative of T2CA is used to produce nifurzide, an intestinal antibacterial.

Pfizer's anti-inflammatory tenidap (**11**) [20] is a developing drug which uses the acid chloride of T2CA as the starting intermediate.

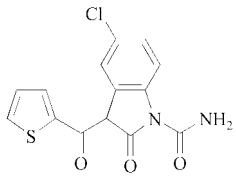

11

Thiophene 2-acetic acid, T2AA, $C_6H_6O_2S$, M_r 142.18. The use of 7-aminocephalosporanic acid (7ACA), following the classical work by the National Research and Development Council in the 1950s [21] to produce a family of semisynthetic cephalosporins, resulted in important first-generation antibiotics (→ Antibiotics, Chap. 3.1.4). Cephaloridine (**12**) [22] and cephalothin [23] were the major products licensed from this work. The second-generation cephalosporin cefoxitin (**13**) [24] came a few years later.

T2AA, or its acid chloride, is used to form the 7-side-chain of these antibiotics. Many processes have been designed to synthesize T2AA (Cl), including chloromethylation of thiophene until the late 1980s in the United States, and starting from T2A in the FAMSO process in Japan.

However, these processes are now superseded, and the oxidation and reduction of 2-acetylthiophene is now the process of choice.

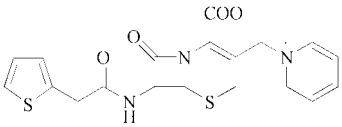

12

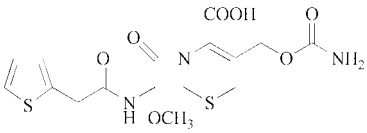

13

Ketotifen fumarate (**14**) [25] is an antiasthma drug (→ Antiallergic Agents, Chap. 2.1.7), which is thought to possess prophylactic properties, and this market is growing.

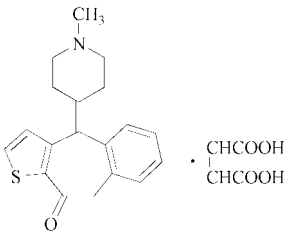

14

Pizotyline (**15**) [26], an antimigraine agent, is synthesized from T2AA (Cl).

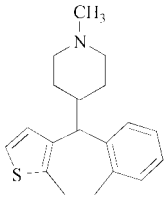

15

Thiophene aldehydes and ketones undergo all the carbonyl reactions of benzenoid aldehydes and ketones. For example 2-propionoylthiophene undergoes Wolff- Kishner reduction to prepare 2-propylthiophene used in the antiparkinson and antiabuse drug adrogolide [27]. Acetylthiophene will undergo a classical Mannich reaction to form the important intermediate 3-(dimethylamino)-1-(2-thienyl)-1-propanone for the antidepressant and antiurinary incontinence drug duloxetine (**16**) [28]. Similar chemistry is used to convert 2AT to the sedative indiplon [29]. Thiophene-2-aldehyde reacts with nitromethane to form the 2-(2-nitroethenyl)thiophene used in the manufacture of the ACE inhibitor temocapril (**17**) [30].

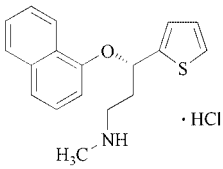

16

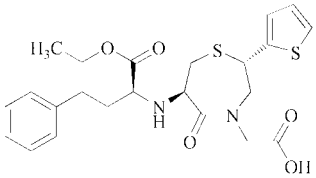

17

Thiophene Dicarboxylic Acids. Adipic acid, or its dichlorinated derivative, provides a means of forming thiophene 2,5-dicarboxylic acid [31]. Thiophene aldehydes and other substituents exhibit pronounced fluorescence [32]:

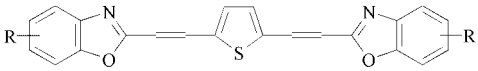

More complicated thiophene polybenzoxazoles are used as optical brighteners, and may be synthesized via the 2,5-carboxylic acid groups.

6.2. Halogenated Thiophenes

The halothiophenes are used widely in the synthesis of thiophene derivatives as intermediates for the preparation of thienyllithium or thienyl Grignard reagents.

2-Bromothiophene (2BT), C_4H_3Br, M_r78.17, bp 153.5 °C, is easily produced to a good purity (see Fig. 2) [33].

Tipepidine (**18**) [34], is an example of an antitussive produced from 2BT via a Grignard reaction. Other examples include tiotropium bromide, an antiasthma drug [35], tiquizium bromide [36], tiemonium iodide [37], and timepidium bromide [38], all antispasmodic drugs.

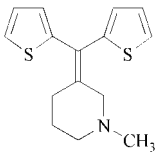

18

Reaction of either 2-thienyllithium or thiophene-2-magnesium Grignard with ethylene oxide generates thiophene-2-ethanol an important intermediate in the manufacture of ticlopidine (**19**) [39] and clopidogrel (**20**). Both are potent pharmaceuticals, used to prevent blood platelet aggregation in the treatment of strokes.

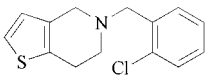

19

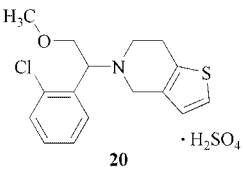

20

Reaction of 2-thienyllithium with sulfur generates the 2-mercaptothiophene, which, whilst not stable, can be used in situ and this is employed to make the antiglaucoma drug dorzolamide [40] and antineoplactic drug batismastat [41].

Dichlorination and acetylation is utilized in the synthesis of the antiglaucoma drug brinzolamide (**21**) [42]

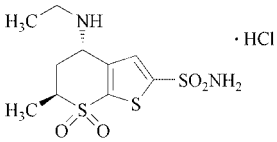

21

3-Bromothiophene (3BT), C_4H_3Br, *bp* 160 °C. In order to obtain 3-(4-)substituted thiophenes, more elaborate chemistry is required than straightforward electrophilic substitution. Successive bromination to 2,3,5-tribromothiophene can be followed by removal of the 2- and 5-Br atoms (see Fig. 2), but, a newer, simpler process has been patented [43] which gives 3-bromothiophene essentially free of 2-Br isomer.

3-Bromothiophene may be used to produce thiophene 3-malonic acid (TMA), the side-chain intermediate for the production of ticarcillin and temocillin from 6-aminopenicillinic acid (6APA). Timentin (**22**) is a powerful antibiotic, produced from ticarcillin plus clavulanic acid [44]. An alternative route to TMA is via 3-methylthiophene (see Section 6.3).

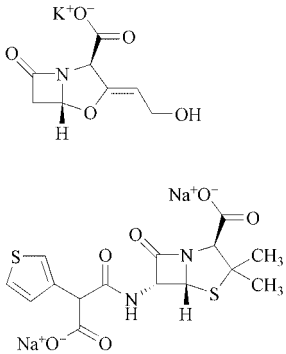

22

3-Thienyllithium derived from 3-bromothiophene is used in the preparation of the vasodilator cetiedil [45].

6.3. Alkylthiophenes

Alkylation of thiophene occurs readily but is not of any utility given the poor selectivity and as a consequence resinification of reaction mixtures. Generally the methylthiophenes are available by VP cyclization (see Fig. 4) or by modification of bromo or acyl groups on the thiophene ring.

The presence of an alkyl group on the thiophene ring does not hinder the reactions outlined elsewhere but also allows derivatization via the halogenation of the alkyl group.

2-Methylthiophene (2MT), C_5H_6S, *bp* 112.5 °C, *mp* 63.4 °C, is commercially produced to 98 wt % purity by VP cyclization (see Fig. 4).

It is a feasible raw material for the production of several of the important 2-substituted thiophenes. It may be used to manufacture 2-chloromethylthiophene (2-thenyl chloride), but this and further syntheses must be completed in situ, because of its instability.

3-Methylthiophene (3MT), *bp* 115 °C, is commercially produced by an analogous process to 2MT, and to the same purity.

Thenyldiamine (**23**) [46], an antihistamine, has been produced since the late 1940s. The synthesis requires a source of 3MT to prepare the 3-thenylbromide precursor.

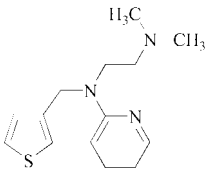

23

Morantel (**24**) [47] is a potent anthelmintic, differing from pyrantel (→ Anthelmintics) only in the additional 3-methyl group on the thiophene ring. Production of morantel is an important market for 3MT.

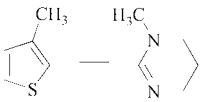

24

As stated above, the semisynthetic penicillins ticarcillin (**25**) [48] and timentin possess a 3-thienyl side-chain; 3MT, via free radical bromination of the methyl side-chain, is an alternative precursor to TMA.

The antifungal tioconazole (**26**) [49] is synthesized from 2-chloro-3-methylthiophene.

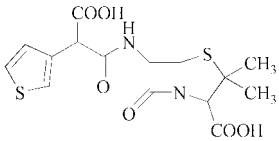

25

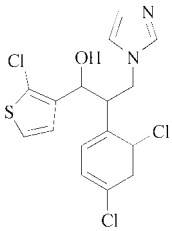

26

Cetiedil (**27**) [50] is a peripheral vasodilator which incorporates a thiophene ring in the 3-position; 3MT provides a means of synthesis.

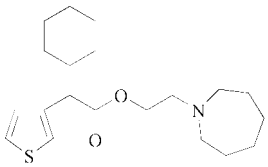

27

Tiagabine (**28**) is an antiepileptic [51] which incorporates two 3-methylthiophene rings that are linked to each other via organometallic intermediates.

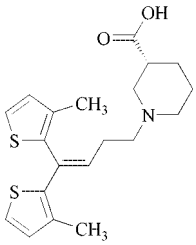

28

There has been much growing interest in conducting polymers based on polythiophenes and poly-3-alkylthiophenes which has been extensively reported [48, 52]. In order to produce a workable polythiophene, a fairly long side-chain

may be introduced at the 3-position of the thiophene ring. Choice of the structure of this side-chain is a compromise between the workability and the maintenance of required conductivity. The chain length of the attached n-alkyl group generally lies between C_4 and C_{12}. A commercial polythiophene has been promoted based on polymerization of 3,4-ethylenedioxythiophene (PDOT) manufactured by the Hinsberg synthesis – two consecutive aldol condensations between diethyl oxalate and diethyl thiodiacetate. Polymerization is via the 2- and 5-positions.

Removal of the 3-Br atom and replacement by an alkyl group is a means of obtaining the desired monomer.

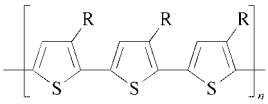

R = $C_5 - C_{10}$ alkyl
Conducting polythiophene

6.4. Amino- and Hydroxylthiophene Derivatives

The commercial availability of a range of thiophenes substituted with an amino group, together with other ring substituents, is based on a modification of the Gewald reaction [53], for 2-NH_2 derivatives. An example of a 2-aminothiophene of synthetic utility is the fungicide sillthiopham [54] made from 2-amino-3,4-dimethyl-3-thiophenecarboxylic acid ester via a Gewald synthesis. Substitution of an amino group in the 3-position requires custom-designed chemistry [55, 56]. Aminothiophenes containing electron attracting groups such as ester groups can be easily diazotized and coupled.

Methyl 3-amino-4-methylthiophene 2-carboxylate $C_7H_{10}NO_2S$. The 3-amino group is used to attach a side-chain to the thiophene ring for manufacture of the local anesthetic carticaine (**29**) [57].

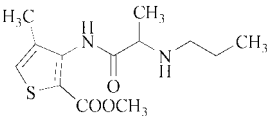

29

Methyl 3-aminothiophene 2-carboxylate $C_6H_7NO_2S$. This derivative is used to produce a sulfonylurea post-emergent herbicide, Harmony (M6316 DuPont) (**30**), which has extremely rapid biodegradability [58]. This characteristic, together with the very low dose requirements, are important environmental safeguards.

The NSAI tenoxicam (**31**) [59] is one of a range of oxicams which contain a thiophene moiety and may be synthesized from this aminothiophene carboxylate. Chlorination of tenoxicam yields lornoxicam (**32**) [60] a NSAI with higher activity.

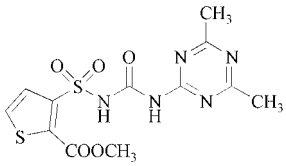

30

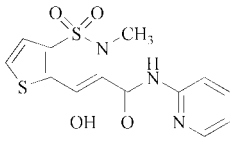

31

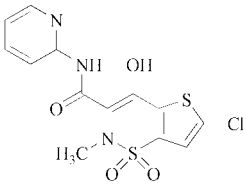

32

Similar chemistry —cyclization to amino- and hydroxythiophenes —is used to manufacture the antihypertensive sitaxsentan [61], sedative brotizolam [62], peristaltic stimulant pumosetrag [63] and rice herbicide thenychlor [64].

The antipsychotic drug olanzapine (**33**) is made from the 5-methyl-2-aminothiophene-3-carbonitrile

Table 2. Toxicological data for thiophene and derivatives

	Thiophene	2-Methylthiophene	3-Methylthiophene	2-Acetylthiophene	2-Bromothiophene
Acute Toxicity					
Oral LD_{50}	1400 mg/kg (rat)	3200 mg/kg (mouse)	2300 mg/kg (rat)	50 mg/kg (mouse)	200 – 250 mg/kg (rat)
Dermal LD_{50}	830 mg/kg (rabbit)	no data	2000 mg/kg (guinea pig)	370 mg/kg (rat)	134 mg/kg (rat)
Inhalation LC_{50}	9500 mg/kg (mouse)	no data	10 g/m3, 2 h (rat)	1460 mg/m3, 1 h (rat)	1.04 mg/L, 4 h (rat)
Irritation					
Skin	yes	expected	yes	not reported	yes
Eye	yes	expected	yes	predicted	yes
Skin sensitization	no data	no data	not found in tests	no data	no data
Occupational exposure limits	none specified	none specified	none specified	2 ppm (provisional)	1 ppm (provisional)
Ecotoxicity	possibly only slight harm to aquatic organisms	possibly only slight harm to aquatic organisms	possible only slight harm to aquatic organisms	harmful to aquatic organisms	expected only slight harm to aquatic organisms
Persistence	expect ultimate biodegradability	expect ultimate biodegradability	expect ultimate biodegradability	expect ultimate biodegradability	expect very slow biodegradability
Stability					
Materials to avoid	strong oxidizers	strong oxidizers	strong oxidizers	strong oxidizers	strong oxidizers
Conditions	stable under normal conditions	stable under normal conditions	stable under normal conditions	stable under normal conditions	slow decomposition unless cool/dark
Combustion products	toxic oxides of S+CO, CO_2	toxic oxides of S+CO, CO_2	toxic oxides of S+CO, CO_2	toxic oxides of S+CO, CO_2	HBr during storage (prolonged), CO, CO_2

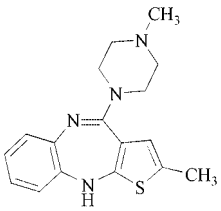

33

6.5. Reduced Thiophenes

The thiophene ring is not normally reduced by hydride reagents and the tetrahydrothiophene ring is normally made via other routes such as cyclization or vapor phase routes. Treatment of thiophenes with Raney nickel can give rise to desulfuration of the thiophene.

Tetrahydrothiophene (THT), C_4H_8S, M_r 88.18, *bp* 121.1 °C, *mp* −96 °C, is easily confused with thiophene itself, particularly since it may also be called thiophane. THT was certainly produced in commercial quantities before thiophene, and it remains a commercial product.

Tetrahydrothiophene is not manufactured from thiophene, but by VP replacement of the oxygen atom in tetrahydrofuran by sulfur from H_2S. The plant and catalyst used are very similar to a thiophene unit. However, THT is used for its very strong odor, either on its own or, more commonly, within a cocktail, e.g., alkylsulfides, as a powerful gas odorant. The saturation of comparatively sweet-smelling thiophene thus produces vastly different odor characteristics.

The herbicide dimethamid-P [65] is made via the tetrahydro-3-imino-2,4-dimethylthiophene which is made by cyclization of a thioester with methacrylate.

6.6. Miscellaneous Thiophenes

The use of *thiophene* mixed with other constituents to produce a denaturant has been developed [66]. Since thiophene does not occur in fermented alcohol, the addition of a marker containing thiophene facilitates the identification of alcohol denatured in this way. The thiophene is very difficult to remove.

Uses of thiophene itself have otherwise been restricted to analytical techniques, where a standard mass of sulfur is required for measurement control. Thiophene may also be used for quantitative deactivation of cracker catalysts.

The amino group, when in the 2-position of a substituted thiophene derivative, may be diazotized to a range of monoazo disperse dyes [67]. Coupling occurs with a pyrazolone or, e.g., phenol, and colors toward the yellow/green part of the spectrum are produced.

Substituted thiophene molecules are found in various components of animal and vegetable flavorings. The combination of H_2S with unsaturated aldehydes may be the source of such thiophene derivatives in, e.g., cooked beef, pork liver, onion, and coffee [68, 69].

As well as the fused ring structures illustrated above, which are valuable commercial products, others are worth mentioning, because of their potential markets.

Many methods are available to synthesize fused thieno rings to a nitrogen heterocyclic [70, 71].

There are several benzothiophenes used as pharmaceuticals including the anticancer drug *raloxifene* (**34**), these are generally manufactured by the cyclization of a thiophenol derivative.

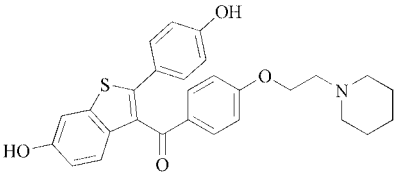

34

7. Toxicology

Toxicological data for thiophene and its derivatives are given in Table 2.

8. References

1. L. S. Fuller, J. W. Pratt, F. S. Yates, *Manuf. Chem. Aerosol News* (1978) no. 5; *Inf. Chim.* (1979) no. 10, 19.
2. Private Communication.
3. *Ullmann*, 4th ed., **22**, 416.
4. Pennsalt Chem. Corp., GB 963 385, 1964 (T. E. Deger, B. Bucholtz, R. H. Goshorn).
5. Midland York Tar Dist., GB 1 345 203, 1974 (N. R. Clark, W. E. Webster).
6. Soc. Nat. Elf Aquitaine, GB 1 585 647, 1981.
7. J. A. Clarke, O. Meth-Cohn, *Tetrahedron Lett.* **52** (1975) 4705 – 4708.
8. E. Campaigne, W. L. Archer, *J. Am. Chem. Soc.* **75** (1953) 989.
9. Monsanto, US 2 581 868, 1952.
10. Lenard, Solmssen, *J. Am. Chem. Soc.* **70** (1948) 2066.
11. Sandoz, US 2 717 251, 1955; US 2 757 175, 1956.
12. Pfizer, US 3 502 661, 1968/70.
13. Smithkline Beecham, EP 403159, 1990.
14. Boehringer Mannheim, DE 1 815 922.
15. Sandoz, US 3 524 824, 1970.
16. Syde, Malleray, *Bull. Soc. Chim. Fr.* 1963, 1276.
17. Standard Oil, US 2 576 031, 1951.
18. Janssen, GB 1 446 239, 1975.
19. CERPHA, DE 2 048 372, 1971.
20. Pfizer, US 4 556 672, 1971.
21. Nat. Res. Dev. Corp., GB 810 196, 1959; US 3 082 155, 1963.
22. Glaxo, FR 1 384 197, 1965.
23. Lilly, BE 618 663, 1962.
24. Merck, DE 2 129 675, 191; DE 2 203 653, 1971.
25. Sandoz, DE 2 111 071, 1971.
26. Sandoz, US 3 272 826, 1966; BE 636 717, 1964.
27. Abbott, WO 94 22858, 1994.
28. Eli Lilly, EP 0273658, 1988.
29. American Cyanamid, US 4521422, 1985.
30. Sankyo, EP 161801, 1985.
31. *Tetrahedron Lett.* **42** (1970) 3719 – 3722.
32. R. E. Atkinson, F. E. Hardy, *J. Chem. Soc. B* 1971, 357 – 361.
33. I. Hirao, *Yakugaku Zasshi* **73** (1953) 1023; *Chem. Abstr.* **48** (1954) 10723.
34. Tanabe, JP 17 988, 1962.
35. Boehringer Ingelheim, EP 418716, 1991.
36. Hokuriku, BE 866988, 1978.
37. C.E.R.M., GB 953386, 1964.
38. Tanabe, FR 2100750. 1972.
39. Parcor (Sanofi), US 4 127 580, 1978 (E. Braye).
40. Merck and Co., EP 457586, 1991.
41. British Biotech, US 5240958, 1993.
42. Alcon, WO 9115486, 1991.
43. Shell Int. Res., EP 029 586, 1989 (P. Grosvenor, L. Fuller).
44. Beecham, US 3 282 926, 1966.
45. Innothera, FR 1460571, 1967.
46. E. Campaigne, W. M. Le Suer, *J. Am. Chem. Soc.* **71** (1949) 333.
47. Pfizer, GB 1 120 587, 1968.

48. M. L. Blohm, P. C. Van Dort, J. E. Pickett, *CHEMTEC* 1992, no. 2, 105 – 107.
K.-H. Kocher, *Hoechst High Chem. Mag.* 1992, no. 12, 39 – 42.
49. Pfizer, US 4 062 966, 1977 (Gymer).
50. Innothera, FR 1 460 571, 1966 (Pons, Robba).
51. Novo, WO 87 00,171, 1987,;
J. Med. Chem. **36** (1993) no. 12, 1716 – 1725.
52. *Eur. Chem. News*, April 9, 1991.
R. D. McCullough, R. D. Lowe, *J. Chem. Soc., Chem. Commun.* 1992, no. 1, 70 – 72.
53. K. Gewald, *Chem. Ber.* **98** (1965) 3571.
K. Gewald, *Chem. Ber.* **99** (1966) 94.
54. Monsanto, WO 96/18631, 1996.
55. H. Fiesselmann, *Chem. Abstr.* **55** (1961) 6497c/17651d.
56. Shell Int. Res., EP 298 542, 1989.
57. Hoechst, US 3 855 243, 1974.
58. Meeting of Weed Soc. of America, Seattle **25** (1980) 10.
59. Hoffmann-LaRoche, DE 2 537 070, 1976.
60. CL Pharma, EP 340472, 1989.
61. Texas Biotechnology, EP 1048657, 2000.
62. Boehringer Ingelheim, DE 2410030, 1974.
63. Mitsubishi, EP 0560348, 1993.
64. Tokuyama, JP 04279505, 1992.
65. Sandoz (BASF), US 5457085, 1995.
66. ACNA, EP-Appl. 424 886, 1991.
67. ICI, GB 1 394 365;
GB 1 394 367;
GB 1 394 368.
68. M. Shankaranarayana, *CRC Crit. Rev. Food Technol.* **4** (1974) 395.
69. R. Wilson, *J. Agric. Food. Chem.* **21** (1973) 873.
70. Schneller, *Int. J. Sulfur Chem.* (1974) 309 – 316.
J. M. Barker, *Adv. Heterocycl. Chem.* **21** (1977) 65.
71. Croda Synthetic Chem., EP 0 003 645, 1982 (O. Meth-Cohn).

Thiourea and Thiourea Derivatives

Bernd Mertschenk, SKW Trostberg AG, Trostberg, Federal Republic of Germany (Chaps. 1, 2)

Ferdinand Beck, SKW Trostberg AG, Trostberg, Federal Republic of Germany (Chaps. 1, 2)

Wolfgang Bauer, Cassella AG, Frankfurt/Main, Federal Republic of Germany (Chap. 3)

1. Introduction 1
2. Thiourea 1
2.1. Physical Properties 1
2.2. Chemical Properties 2
2.3. Production 3
2.4. Environmental Protection 3
2.5. Quality Requirements 3
2.6. Chemical Analysis 4
2.7. Storage and Transportation 4
2.8. Uses 4
2.9. Legal Aspects 4
2.10. Economic Aspects 4
2.11. Toxicology 5
3. Thiourea Derivatives 10
3.1. Physical Properties 10
3.2. Chemical Properties 10
3.3. Production 10
3.4. Uses 12
3.5. Toxicology 12
4. References 12

1. Introduction

Thiourea is a chemically interesting compound, which has three functional groups: amino, imino, and thiol. This results from tautomerism between thiourea and isothiourea:

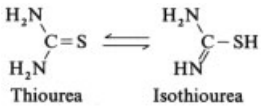

Because of this polyfunctionality and also because of its complex-forming properties, thiourea has been widely used for more than 30 years [1–5]. Its most important uses are as a starting material for nitrogen- and sulfur- containing heterocycles and formamidinesulfinic acid (thiourea dioxide), as a reaction partner for aldehydes, and as a component of addition compounds and complexes.

Thiourea derivatives are thermally stable, mostly crystalline compounds. They can be produced in good yields from readily available starting materials by simple syntheses. This class of compounds, described in detail in several review articles [6–11], has gained importance in the pharmaceutical sector, in plant protection, in various technical applications, and in the synthesis of heterocycles.

2. Thiourea

2.1. Physical Properties

Thiourea [*62-56-6*], 2-thiourea, CH_4N_2S, M_r76.12, d_4^{20} 1.405, is a white, odorless solid, and crystallizes in a rhombic bipyramidal structure with four molecules per unit cell. It is the only known molecular crystal which is ferroelectric within a certain temperature range. Space group $D_2^{16}{}_h - P_{\mathrm{nma}}$.

Thiourea is soluble in protic and aprotic polar solvents (Table 1) and insoluble in nonpolar ones.

Table 1. Solubility of thiourea

In H_2O						
Temperature, °C	0	20	40	60	80	100
Solubility, g/100 g	4.9	13.7	30.7	70.9	138	233
In MeOH						
Temperature, °C			25	40.7	53.7	61.9
Solubility, g/100 g			11.9	16.4	22	24.6
In EtOH						
Temperature, °C		20	31.9	45	58	64.7
Solubility, g/100 g		3.6	4.7	6.3	8.5	9.8

Thiourea has no sharp melting point, as it rearranges relatively rapidly to ammonium rhodanide (NH_4SCN) from 135 °C. On rapid heating, a melting point of 180 °C has been determined. The heat of combustion at constant vol-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a26_803

ume is 1482.2 kJ/mol (19.47 kJ/g), and at constant pressure 1485.6 kJ/mol (19.52 kJ/g).

2.2. Chemical Properties

Although it is known in only one form, thiourea can also react as the tautomeric iminothiol (see Section 3.2). Salts such as $(NH_2)_2CS \cdot HNO_3$ or $(NH_2)_2CS \cdot H_2SO_4$ are formed with strong acids [3]. Addition compounds and complexes can be synthesized with metal oxides or inorganic salts in acid solution, e.g., $[Cr\{(NH_2)_2CS\}_3]Cl_3$.

In alkaline solution, a large number of heavy metal ions can be precipitated quantitatively as sulfides with thiourea. These precipitates are readily filtered, so that the method can be used for analytical determination of metal salts [12]. The production of light-sensitive PbS layers [13] and the purification of mercury- containing wastewater [14] are based on this reaction.

Thiourea forms clathrates (inclusion compounds) with branched aliphatic and cyclic hydrocarbons [15]. This property can be used, e.g., for separating branched and straight- chain aliphatic hydrocarbons [16].

Thiourea can be alkylated and acylated, and reacts with aldehydes and ketones. With formaldehyde, for example, it forms *N*-(hydroxymethyl)thiourea ($H_2N{-}CS{-}NH{-}CH_2{-}OH$), which condenses to resin-like products on heating [17]. By condensation of thiourea with chloroketones, diketones, hydroxyketones, oxocarboxylic acids, and similar ketone derivatives, a large number of heterocycles can be produced. From 2-haloketones or 2-haloaldehydes (in the enol form) 2-aminothiazole derivatives are formed:

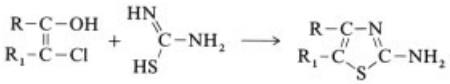

1,1-Dihaloalkanes condense with thiourea to 1,3,5-dithiazines:

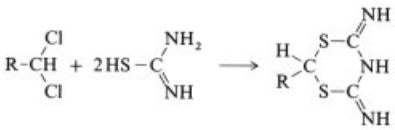

2-Hydroxyketones give 4-imidazoline-2-thiones:

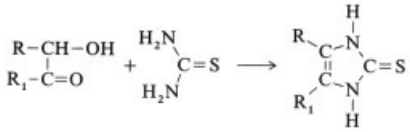

With 2-oxocarboxylic acids (e.g., pyruvic acid) 2,3-dihydro-2-thioxo-4*H*-imidazole-4-one derivatives are formed:

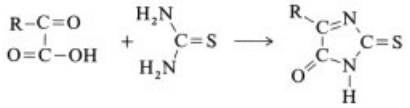

Malonic esters react with thiourea to thiobarbituric acid:

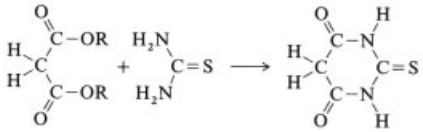

2-Oxocarboxylic acid esters condense with thiourea to thiouracils, and 1,3-diketones to mercaptopyrimidines:

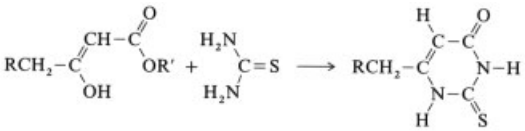

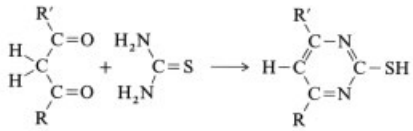

2.3. Production

Thiourea was first prepared by REYNOLDS by thermal rearrangement of ammonium rhodanide at ca. 150 °C:

$$NH_4SCN \rightleftarrows (NH_2)_2CS$$

As this is an equilibrium reaction and separation of the two substances is quite difficult, better methods of preparation have been investigated.

The reaction between carbon disulfide and ammonia or ammonium carbonate under pressure at ca. 140 °C has not achieved industrial importance [18].

Production of thiourea is now carried out by treating technical-grade calcium cyanamide ($CaCN_2$) with hydrogen sulfide or one of its precursors, e.g., ammonium sulfide or calcium hydrogen sulfide [19]. The calcium cyanamide must contain at least 23 % nitrogen and must be free from calcium carbide (CaC_2), as otherwise explosive acetylene is liberated by water or hydrogen sulfide.

In Germany, thiourea is produced in a closed system by reaction of calcium cyanamide with hydrogen sulfide:

$$CaCN_2 + 3H_2S \longrightarrow Ca(SH)_2 + (NH_2)_2CS$$

$$2CaCN_2 + Ca(SH)_2 + 6H_2O \longrightarrow 2(NH_2)_2CS + 3Ca(OH)_2$$

$$Ca(OH)_2 + CO_2 \longrightarrow CaCO_3 + H_2O$$

A reaction vessel is charged continuously with an aqueous suspension of calcium cyanamide and a mixture of hydrogen sulfide and carbon dioxide, such that the above reactions go to completion. The converted reaction mixture is removed continuously from the reactor and excess H_2S is removed by stripping. Wastegas from the reactor contains hydrogen sulfide, which is purified in a scrubber connected to the reactor.

The stripped thiourea solution is filtered to remove the precipitated calcium carbonate formed as one of the reaction products. The filtrate is evaporated in vacuum and thiourea is crystallized by cooling. It is filtered, washed, and dried in an air stream.

2.4. Environmental Protection

In Germany, the scrubbed waste gases resulting from the production and filtration of thiourea are treated in an incinerator. All installations where thiourea dust is handled are equipped with air extraction. The extracted air is collected and, together with the waste air from the product drier, is purified in a scrubber.

Mother liquors containing thiourea are removed batchwise from the production process and are used in high-temperature incineration processes to remove NO_x. Side products of low solubility separated from the mother liquor are incinerated externally in a licensed plant.

The precipitated calcium carbonate arising in the production of thiourea is reprocessed (e.g., in cement factories or brick works). This is a contribution to the conservation of natural lime resources.

2.5. Quality Requirements

The German producer gives the following specification for the standard product:

Thiourea	$\geq$ 99.0 %
Water	$\leq$ 0.30 %
Ash	$\leq$ 0.10 %
Rhodanide (SCN^-)	$\leq$ 0.15 %
Other N-containing compounds	$\leq$ 0.5 %
Iron	$\leq$ 10 mg/kg

Special qualities with higher purity are also produced.

2.6. Chemical Analysis

Tests of the content (purity) of thiourea can be carried out titrimetrically with Ce(IV) and potassium iodate solutions [20]. Since the commercial product is of relatively high quality, it is more efficient to determine the impurities as a means of quality control. The main impurity, dicyandiamide, is best determined by TLC. The incineration residue (ash) and the water content are also measured. A quantitative determination of sulfate, heavy metals, and rhodanide can also be carried out.

Traces of thiourea (e.g., in water) can be determined by HPLC (C_{18} reverse-phase with water as the mobile phase and subsequent UV detection at 245 nm) [21].

Reaction of thiourea with sodium nitrite in acetic acid solution to give thiocyanate, which gives a red coloration with Fe^{3+} ions, is also suitable for the determination of trace amounts. The red coloration is used to determine thiourea photometrically at 450 nm [22].

2.7. Storage and Transportation

Thiourea is packed in polyethylene bags ($\leq$ 50 kg) or in folding containers ($\leq$ 1500 kg). Bags and containers must be kept tightly closed, and in a cool, dry place. The contents must be protected from the action of light. Thiourea is not regarded as dangerous goods in transport regulations (RID/ADR, etc.).

2.8. Uses

Thiourea has a wide range of uses, e.g., for producing and modifying textile and dyeing auxiliaries [23], [24], in the production and modification of synthetic resins [25], in repro technology [26], in the production of pharmaceuticals (sulfathiazoles, tetramisole [27], and cephalosporins [28], in the production of industrial cleaning agents (e.g., for photographic tanks [29] and metal surfaces in general [30], [31]), for engraving metal surfaces [32], as an isomerization catalyst in the conversion of maleic to fumaric acid [33], in copper refining electrolysis [34], in electroplating (e.g., of copper) [35], and as an antioxidant (e.g., in biochemistry) [36].

Other uses are as an additive for slurry explosives [37], as a viscosity stabilizer for polymer solutions (e.g., in drilling muds [38]), and as a mobility buffer in petroleum extraction [39]. The removal of mercury from wastewater from chlorine – alkali electrolysis [14] and gold and silver extraction from minerals [40], [41] are also of economic importance.

2.9. Legal Aspects

Thiourea is included in international inventories of existing chemicals, e.g., EINECS (European Union), TSCA Inventory (United States), DSL (Canada), ACOIN and AICS (Australia), and MITI (Japan). In Germany and in the European Union, it is classified as a suspected carcinogen (MAK III B [42], [43]; EU cancer category 3). Similar evaluations have been made by the International Agency for Research on Cancer (Lyon, France) [44], and within the framework of the National Toxicology Program in the United States [45]. Within the European Union, thiourea must be labeled as a hazardous substance with hazard symbol Xn (St Andrew's cross – harmful), and R 22 (harmful if swallowed), R 40 (possible risk of irreversible effects), S 22 (do not breathe dust), S 24 (avoid contact with skin), and S 36/37 (wear suitable protective clothing and gloves). In the European Union Council Regulation on Existing Chemicals, thiourea is named in Annex I [46]. In Germany, the substance is classified in water hazard class 2 (hazardous for waters).

2.10. Economic Aspects

The quantity of thiourea produced and sold worldwide is ca. 10 000 t/a. The only producer in Western Europe is SKW Trostberg, Germany. Thiourea is also produced by Sakai Chemicals in Japan and at least seven Chinese companies.

According to estimations of SKW Trostberg, the largest producer, the quantity of thiourea produced worldwide is divided into the following main areas of use:

Intermediate in the production of thiourea dioxide	ca. 30 %
Ore leaching	25 %
Diazo papers	ca. 15 %
Catalyst in fumaric acid synthesis	ca. 10 %

The remainder is divided between smaller areas of use (Section 2.8).

The United States EPA carried out a survey of the uses of thiourea in the United States. It is used, for example, in animal glue liquefiers, in silver and metal cleaners, in the production of flame-retardant resins, and as a vulcanization accelerator [45].

2.11. Toxicology [42–44], [47]

Acute Toxicity and Local Tolerance. The oral LD_{50} for mice has been reported to be 1000 mg/kg [48], for rabbits 10 000 mg/kg [49], and for rats 1750 mg/kg [50] and 6200 – 6300 mg/kg [51]. After intraperitoneal administration LD_{50} values varying between 4 and 1340 mg/kg were found for rats [52], [53]. Lung edema and pleural effusion have been described as symptoms. The dermal LD_{50} is $>$ 2800 mg/kg for rabbits [50] and $>$ 6810 mg/kg for rats [51]. After inhalation of dust for 4 h (particle size 0.8 – 4.7 µm) LC_{50}

for rats was $> 170\,mg/m^3$ [50], after aerosol inhalation (10 % solution) $> 195\,mg/m^3$ [50]. An alcoholic solution (droplet size $< 10\,\mu m$) gave $LC_{50} > 900\,mg/m^3$ [51].

In skin tolerance tests (normal and abraded rabbit skin, 24 h application), slight to distinct erythema and slight edema were observed [50]. A 4 h application of 0.5 g gave no reaction [51]. Application of a 10 % solution into the conjunctival sac of a rabbit's eye gave no reaction [50], but 100 mg of the solid substance gave a slight reddening and swelling [51]. Thiourea does not have a sensitizing effect in the guinea pig [51].

Subacute and Chronic Toxicity. In long-term application of thiourea, the effect on the thyroid gland stands out [54]. The substance disturbs incorporation of iodine in the synthesis of thyroxine and thus leads to a release of thyrotropic hormone from the pituitary gland, followed by severe stimulation of the thyroid gland, resulting in hyperplasia, adenomas, and, under certain circumstances, formation of thyroid carcinomas. Administration of thyroxine extensively prevents these effects.

Nonspecific effects of thiourea include hemosiderosis in the spleen, lymph nodes, and intestinal villi [55], weight loss of the liver, kidneys and spleen [54], anemia, leukopenia, increased erythropoiesis, and reduced bone growth.

Some examples of long-term tests are given below: rats received 0.625 % thiourea in the diet for 14 days. This resulted in 52 % growth inhibition [56]; 2.5 g per kilogram food (over 13 weeks) did not affect the body weight of mice [57]. Administration of 2.5 ppm in the drinking water of rats over 13 weeks gave no reaction [58]; after $27.5\,mg\,kg^{-1}\,d^{-1}$ for 53 weeks in the drinking water normal growth was observed in rats [59]. Subcutaneous application in the mouse of $500\,mg\,kg^{-1}\,d^{-1}$ for 10 d led to a slight decrease in the colloid content of the thyroid gland [60].

Mutagenicity [42], [43], [47]. In the Ames test with *Salmonella typhimurium* (TA 98, 100, 1535, 1537, 1538; 10 – 5000 µg per plate) thiourea was shown to be nonmutagenic, with and without activation (S9 mix) [51], [61–63]. A DNA repair test with isolated rat hepatocytes (0.3 – 30 mmol thiourea) also gave negative results [51]. A further test of this type gave a slightly positive result, as did a test with Chinese hamster cells V79 [64]. A CHO/HGPRT cell mutation test [65], a test for mitotic recombination (*Saccharomyces cerevisiae* D3) [66], and a micronucleus test (rat, 2×350 mg/kg, oral) [50] were also negative. A host-mediated assay (mouse) with TA 1530, TA 1538 (125 mg/kg thiourea, i.p.) showed a slight increase in the mutation rate, but the same test with TA 1535 and *Saccharomyces cerevisiae* (1000 mg/kg thiourea, i.p.) did not [48].

Carcinogenicity. A cell transformation test with Syrian golden hamster embryo cells (0 – 100 g/mL) was negative [67].

Experiments on mice with different test periods (4 – 10 months) and dose levels (0.1 – 0.5 % in drinking water) showed a slight enlargement of the thyroid gland and slight thyroid hyperplasia [68]. Application of 2 % thiourea in the diet over 1.5 years led to hyperplasia and cystic changes in the thyroid gland, but not to carcinomas [69]. In rats receiving 0.25 % in drinking water for 5 – 23.5 months, adenomas and carcinomas in the thyroid gland were observed [70]; in other tests (0.2 % in drinking water) tumors in the orbita, nose, and ear regions [71] were observed as well as in the liver (0.01 – 0.25 % thiourea in the diet) [72].

Tumor formation is associated with excessive stimulation of the thyroid gland by thyrotropic hormone from the pituitary gland. The fact that thiourea has only a weak mutagenic potential also indicates an epigenetic mechanism for tumor formation. Since tumors are also found occasionally in organs other than the thyroid gland, thiourea is classified as a suspected carcinogen [42], [43].

Embryotoxicity. In mice and rats, maternally toxic oral doses of 1000 mg/kg body weight were found to be embryotoxic [73]. In rat fetuses, whose dams had received 2000 ppm in drinking water, maturation defects were apparent in the central nervous system, skeleton, and eyes [74]. These effects are attributed to the thyreostatic activity of thiourea. Therefore, at dosages which are not sufficient to inhibit thyroid function effects are not expected [42]. A single oral dose of 480 mg/kg was found to be

Table 2. Uses of thiourea derivatives as vulcanization accelerators [102]

Compound	CAS No.	Polymer	Trade name	Producer
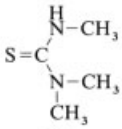	[*2489-77-2*]	polychloroprene	Thiate EF-2	Vanderbilt
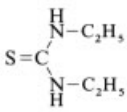	[*105-55-5*]	polychloroprene EPDM *	Pennzone Robac DETU Sanceller EUR Thiate H	Pennwalt Robinson Sanshin Vanderbilt
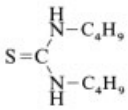	[*109-46-6*]	polychloroprene EPDM *	Robac DBTU Thiate U	Robinson Vanderbilt
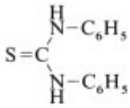	[*102-08-9*]	polychloroprene EPDM *	A-1 Thiocarbanilid Stabilizer C	Monsanto Mobay
	[*96-45-7*]	polychloroprene	Akrochem ETU-22 Robac 22 Sanceller 22 Valcazit NPV/C	Akron Chemical Robinson Sanshin Mobay
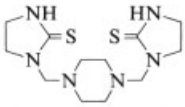	[*14764-02-4*]	polychloroprene	Vulafor 322	Vulnax

* EPDM = ethylene–propylene–diene terpolymers.

neither maternally toxic nor teratogenic in rats [75].

Absorption and Metabolism. In humans, thiourea is rapidly and completely absorbed from the gastrointestinal tract following oral administration. Most is excreted unchanged in the urine [76]. Small amounts appear as sulfate and ether sulfate, as shown by tests with radioactive material in rats [77]. A whole-body autoradiography with ^{14}C-thiourea showed enrichment and longer residence time in the thyroid gland compared with other tissues [78]. Following cutaneous application of 2 g/kg as an aqueous solution in rabbits, ca. 4 % of the dose was detected in the urine, but only 0.1 % after application of the solid [50].

Effects on Humans. Thiourea has been used temporarily in therapy because of its thyroid-depressant activity. After initial doses of 2 – 3 g/d, dosages of 100 – 200 mg/d were administered, and finally 25 – 70 mg was recommended as the daily dose. Doses of 10 – 15 mg/d were ineffective in 5 of 8 people [79]. Its use in therapy has been frequently associated with side effects, e.g., gastrointestinal disorders, pain in the muscles and joints, leukopenia, fever, and reddening of the skin. There are only a few reports on observations in the workplace [80], [81]. Apart from hypothyroidism, symptoms appeared which were similar to side effects known from therapeutic use.

In the use and processing of thiourea, contact dermatitis has occasionally been reported, e.g., on use in silver polish and diazo (blueprint) paper, sometimes after UV irradiation [82–86].

A Dutch expert committee recommended an occupational exposure limit of 0.5 mg/m^3 [47]. This value is also cited in the documentation of the German Research Association with reference to the thyroid gland [42], [43].

Table 3. Pharmacologically active thiourea derivatives

Compound	CAS No.	Use	Generic name	Trade name	Producer
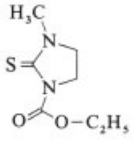	[*22232-54-8*]	thyrotherapeutic agent (thyreostatic)	carbimazole	Atirozidina Basolest Neo-Carbimazole Neo-Mercazole Neo-Morphazole Neo-Thyreostat Neo-Tomizol Neral Tyrazol	Confas Pharmachemie Landerlan Schering Nicholas Herbrand Novag Conradson Neofarma
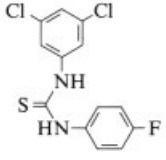	[*790-69-2*]	fungicide, dermatological agent	loflucarban	Fluonilid	Continental Pharma
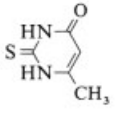	[*56-04-2*]	thyrotherapeutic agent (thyreostatic)	methylthiouracil	Methiacil Methiocil MTU Thimecil Thyreostat Thyrostabil Tiouracil	Schwarz Helvepharm Philopharm Physicians' Drug Herbrand Streuli Lefa
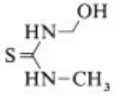	[*15599-39-0*]	antiseptic	noxytiolin	Gynaflex Noxyflex-S	Geistlich Geistlich
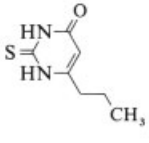	[*51-52-5*]	thyrotherapeutic agent (thyreostatic)	propylthiouracil	Propycil Propyl-Thiocil Propyl-Thyracil Prothiucil Thyreostat II Tiotil Tirostaz	Kali-Chemie Teva Merck-Frosst Donau-Pharmazie Herbrand Pharmacia Yurtoglu
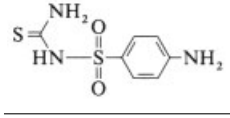	[*515-49-1*]	antiinfective agent	sulfathiourea	Badional Fontamide Salvoseptyl	Bayer Specia Alkaloida

Table 3. (Continued)

	[*77-27-0*] [*337-47-3*] Na-salt	ultrashort narcotic, anesthetic	thiamylal	Bio-Tal Bio-Tel Surital Thioseconal	Philips Roxane Philips Roxane Parke Davis Lilly
	[*6028-35-9*]	thyreostatic	thiabenzazolin	Thyreocordon	Donau-Pharmazie
	[*92-97-7*]	tuberculostatic	thiocarbanidin	Thioban	Parke Davis
	[*76-75-5*] [*71-73-8*] Na-salt	narcotic	thiopental	Farmotal Intraval Leopental Nesdonal Pentothal Thio-Barbityral Thionembutal Tiobarbital Trapanal	Farmitalia Carlo Erba May & Baker Leo Specia Abbott Amino Abbott Miro Byk Gulden

Table 3. (Continued)

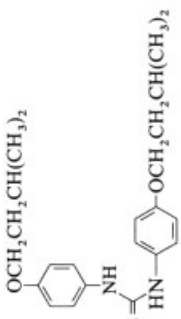	[*910-86-1*]	tuberculostatic	thiocarlide	Amixyl Disoxyl Isoxyl	Inibsa Ferrosan Continental Pharma, Inibsa
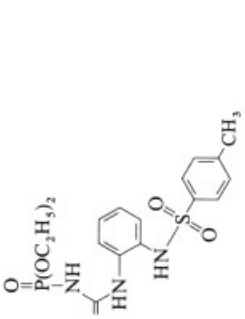	[*52406-01-6*]	anthelmintic (veterinary)	uredofos	Sansalid	Röhm & Haas

3. Thiourea Derivatives

3.1. Physical Properties

Melting points and solubilities of thiourea derivatives are given in [7], [9], [11].

The free energy of activation for rotation about the C – N bond determined by dynamic NMR spectroscopy is in the range $\Delta G^{+} = 29 - 50$ kJ/mol [87].

Various physical methods, e.g., neutron diffraction [88] and IR spectroscopy [89], confirm a strong contribution of the polar, planar structure I to the ground state:

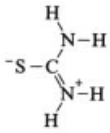

I

Monoalkyl-substituted thioureas exist as mixtures of *E* and *Z* isomers . In monoaryl substitution products, the *E* isomer is preferred [90].

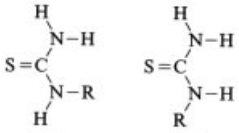

(*E*) (*Z*)

3.2. Chemical Properties

The chemistry of thiourea and its derivatives is dominated by two structural features:

1) The nucleophilicity of sulfur gives isothiuronium salts in good yields on protonation [91], alkylation, and arylation [9].
2) The polyfunctional character of the thiourea system can be used to synthesize 4-, 5-, 6-, and 7-membered heterocycles [9], [92].

Halogenation and oxidation reactions of monoaryl-substituted thioureas give 2-aminobenzothiazoles [93], important in the production of, e.g., polyester dyes [94].

3.3. Production

Methods for the production of thioureas can be divided into seven groups [11]:

1) Reaction of alkyl, aryl, and heteroaryl isothiocyanates with ammonia, or primary or secondary amines can be used universally, is simple to carry out on a preparative scale, and gives mono-, di-, and trisubstituted thioureas in good yields:

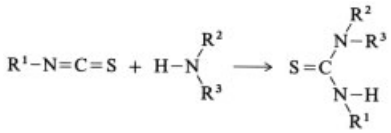

2) Reaction of primary amines with carbon disulfide is of fundamental importance for the synthesis of *N,N*-diaryl-substituted thioureas and results in quantitative yield with aromatic amines. Dithiocarbamates are formed as isolable intermediates (→ Dithiocarbamic Acid and Derivatives, Chap. 3.), which are converted to thioureas with loss of hydrogen sulfide:

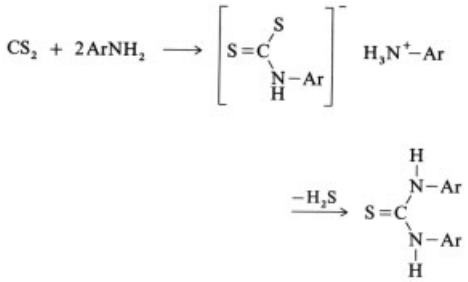

Nitroarenes can also be used as starting materials if the reaction is carried out in an alkaline medium [95].

3) Primary and secondary amines react with thiophosgene, forming thiocarbamoyl chlorides as intermediates, which are then converted to di- and tetrasubstituted thioureas:

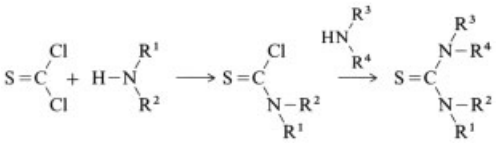

However, other thiocarbonic acid derivatives, e.g., thiocarbonylbisazoles or diesters of trithiocarbonic acid, react only with primary amines.

4) By analogy with the thiourea synthesis from ammonium thiocyanate, primary alkyl- and arylammonium thiocyanates are converted by thermolysis into monosubstituted thioureas in good yields:

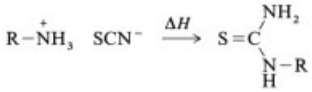

5) Addition of hydrogen sulfide to alkyl- and aryl-substituted cyanamides offers an effective route to *N*-mono- and *N,N*-disubstituted thiourea derivatives by analogy with the industrial thiourea synthesis from calcium cyanamide:

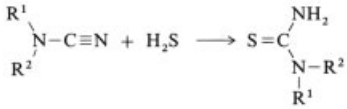

6) For the production of sterically hindered, tetrasubstituted thioureas, sulfurization of the corresponding urea derivatives with phosphorus pentasulfide is suitable:

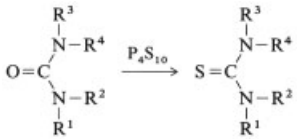

7) Cyclic thioureas can be obtained by the methods known for open-chain thioureas, and by the reaction of diamines with thiourea [96]:

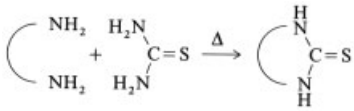

3.4. Uses

Analysis. For the quantitative determination of thiourea compounds, titration with phenyl iodosoacetate is suitable [97]. Chromatographic methods have been developed for separation [98].

Table 4. Thiourea derivatives as plant protection agents and pesticides [107]

Thiourea derivative	CAS No.	Use	Generic name	Trade name	Producer
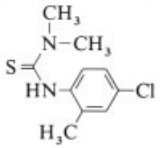	[*28217-97-2*]	acaricide, anti-tick agent	chloromethiuron	Dipofene	Ciba-Geigy
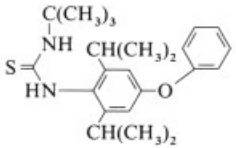	[*80060-09-9*]	insecticide, acaricide	diafenthiuron	Pegasus Polo	Ciba-Geigy Ciba-Geigy
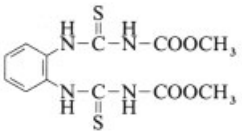	[*23564-05-8*]	fungicide	thiophanate methyl	Topsin M Cercobin M Mildothane Cycosin	Nippon Soda Nippon Soda Rhone-Poulenc American Cyanamid
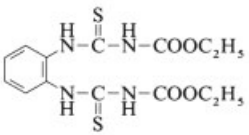	[*23564-06-9*]	fungicide, anthelmintic	thiophanate	Topsin Nemafax Verdamax	Nippon Soda May & Baker May & Baker

Because of their complex-forming properties [99], thiourea derivatives can be used as analytical reagents [100] and for the extraction of noble metals [101].

Industrial Applications. The thiourea derivatives listed in Table 2 are used as accelerators for the vulcanization of polychloroprene and ethylene – propylene – diene terpolymers (EPDM) [102].

N-Substituted thiourea derivatives are used as corrosion inhibitors, as antioxidants in the rubber industry, in electroplating processes [8], and in the production of light-sensitive materials [103].

Biological Activity. Some thioureas are used as active substances in human and veterinary medicine [104–106], as insecticides [107–109], and as rodenticides [110]. The structures of thiourea derivatives which are important as pharmaceuticals and as plant protection agents and pesticides are given in Tables 3 and 4.

3.5. Toxicology

The toxicological properties of a selection of alkyl- and aryl-substituted thiourea derivatives are described in [111].

More recent studies have dealt with metabolism and the effect of steric factors on the toxicity of *N*-phenylthiourea derivatives [112], and with the mutagenic, teratogenic, and carcinogenic properties of imidazolidine-2-thione (ethylenethiourea) [*96-45-7*], a metabolite of ethylene-bisthiocarbamate fungicides [113].

4. References

General References

1. *Beilstein,* 4th ed., **E IV 3, 342 ff.** (1977).
2. *Houben-Weyl,* **9,** 799 – 803, 884 – 899.
3. *Gmelin,* 8th ed., **(14),** part D6, Kohlenstoff, 1978.
4. SKW Trostberg AG, Produktstudie Thioharnstoff, Trostberg 1980.
5. BUA, Stoffdossier Thioharnstoff, in press.
6. D. C. Schröder, *Chem. Rev.* **55** (1955) 181 – 221.
7. E. E. Reid: *Organic Chemistry of Bivalent Sulfur,* Chem. Publishing Co., New York 1963, pp. 11 – 65.
8. V. V. Mozolis, S. P. Iokubaitite, *Russ. Chem. Rev. (Engl. Transl.)* **42** (1973) 587 – 595.
9. F. Duus in D. Barton, W. D. Ollis (eds.): *Comprehensive Organic Chemistry,* **vol. 3,** Pergamon Press, Oxford 1979, pp. 452 – 486.
10. D. H. Reid: *Organic Compounds of Sulphur, Selenium, and Tellurium,* The Chemical Society, London, vol. 1, 1970, pp. 209 – 216; vol. 2, 1973, pp. 236 – 243; vol. 3, 1975, pp. 267 – 279; vol. 4, 1977, pp. 142 – 153; vol. 5, 1979, pp. 139 – 151; vol. 6, 1981, pp. 165 – 174.
11. *Houben-Weyl,* **E 4,** 484 – 521.

Specific References

12. R. Bauer, I. Wehling, *Fresenius* ' Z. Anal. Chem. 199 (1964) 171 – 183.
13. F. Kicinsky, *Chem. Ind. (London)* **1** (1948) 54 – 57.
14. D. Napoli, G. Ratti, G. Tubiello, *Quad. Ist. Ric. Acque* **45** (1979) 347 – 361.
15. K. Gawalek: *Einschlußverbindungen,* VEB Deutscher Vlg. der Wissenschaften, Berlin 1969, pp. 226 – 236.
16. W. Schlenk, *Justus Liebigs Ann. Chem.* **573** (1952) 142 – 162.
17. BASF, DE-OS 1 241 612, 1962 (H. Scheuermann et al.).
18. G. Inghilleri, *Gazz. Chim. Ital.* **39** (1909) 634 – 639.
19. S. A. Miller, B. Bann, *J. Appl. Chem.* **6** (1956) 89 – 93.
20. R. P. Agarwal, S. K. Ghosh, *Chim. Anal. (Paris)* **47** (1965) 407 – 488.
21. SKW Trostberg AG, unpublished.
22. K. Hayashi et al., *Bunseki Kagaku* **26** (1977) 514 –515.
23. Farbenfabriken Bayer, FR 1 576 094, 1969.
24. Ciba Geigy, DE-OS 2 059 373, 1971 (J. Wegmann).
25. SKW Trostberg, DE-OS 2 027 085, 1971 (H. Michaud et al.).
26. Ricoh Co., Ltd., GB 1 218 933, 1971 (T. Iwaoka et al.).
27. S. A. Janssen, US 3 274 209, 1960 (A. Raeymaekers).
28. Lonza, EP 0 045 005, 1984 (A. Huwiler et al.).
29. Degussa, DE-OS 2 511 433, 1975 (H. Knorre et al.).
30. Trio Chemical Works, US 2 907 716, 1959 (R. J. Wassermann).
31. Geigy, FR 1 577 582, 1968 (J. Casanova).
32. IBM, US 4 588 471, 1986 (J. Griffith et al.).

33. W. Schliesser, *Angew. Chem.* **74** (1962) 429 – 430.
34. F. J. Kaltenböck, H. Wöbking, *Metall (Berlin)* **40** (1986) 1144 – 1146.
35. A. Hung, *J. Electrochem. Soc.* **132** (1985) 1047 – 1049.
36. E. Van Driesche et al., *Anal. Biochem.* **141** (1984) 184 – 188.
37. ICI, DE-OS 2 553 055, 1976 (G. H. McCallum et al.).
38. Philips Petroleum, US 4 124 073, 1978 (D. R. Wier).
39. Shell Int. Res. Maatschappij B.V., DE-OS 2 715 026, 1977 (S. L. Wellington).
40. T. Groenewald, *J. South Afr. Inst. Min. Metall.* 1977, 217 – 223.
41. SKW Trostberg AG, company brochure, Laugung und Gewinnung von Edelmetallen mit Thioharnstoff, Trostberg 1984.
42. D. Henschler: *Gesundheitsschädliche Arbeitsstoffe. Toxikologisch-arbeitsmedizinische Begründung von MAK-Werten, Thioharnstoff,* Verlag Chemie, Weinheim 1988.
43. DFG: *Occupational Toxicants. Critical Data Evaluation for MAK Values and Classification of Carcinogens,* **vol. 1,** Verlag Chemie, Weinheim 1991, p. 301.
44. *IARC Monogr. Eval. Carcinog. Risk Chem. Man* **7** (1974) 95; **Suppl. 4** (1982) 270.
45. U.S. Department of Health and Human Services: *National Toxicology Program. Sixth Annual Report on Carcinogens,* National Institute of Environmental Health Services, Research Triangle Park, NC, 1991.
46. Council of the European Communities, Council Regulation (EU) of March 23, 1993 on the Evaluation and Control of the Environmental Risks of Existing Substances, Official Journal of the European Communities, L 84, vol. 36, 5th April 1993.
47. Dutch Expert Committee for Occupational Standards: *Health-Based Recommended Occupational Exposure Limits for Thiourea,* Directorate-General of Labour of Ministery of Social Affairs and Employment, Monograph RA II/90, 1990.
48. V. F. Simmon, H. S. Rosenkranz, E. Zeiger, L. A. Poirier, *JNCI J. Natl. Cancer Inst.* **62** (1979) 911 –918.
49. F. B. Flinn, J. M. Geary, *Contr. Boyce Thompson Inst.* **11** (1940) 241 – 247.
50. A. P. de Groot et al., Berichte von CIVO (Central Institute for Nutrition and Food Research), TNO, in Zeist (Holland) an SKW Trostberg AG 1975 – 1983.
51. F. Korte, H. Greim: Überprüfung der Durchführbarkeit von Prüfungsvorschriften und der Aussagekraft der Grundprüfung des E. Chem. G. UFOPLAN-Nr. 10704006/01, 1981.
52. S. H. Dieke, C. P. Richter, *J. Pharmacol. Exp. Ther.* **83** (1945) 195 – 202.
53. G. S. Wiberg, H. C. Grice, *Food. Cosmet. Toxicol.* **3** (1965) 597 – 603.
54. E. Huf, F. Auffarth, *Naunyn-Schmiedebergs Arch. Exp. Pathol. Pharmakol.* **206** (1949) 394 – 415.
55. L. Arvy, M. Gabe, *Compt. Rend. Soc. Biol.* **144** (1950) 487 – 488.
56. C. C. Smith, *J. Pharmacol. Exp. Ther.* **100** (1950) 408 – 420.
57. H. P. Morris, C. S. Dubnik, A. J. Dalton, *J. Natl. Cancer Inst.* **7** (1946) 159 – 169.
58. M. R. Osheroff, Hazleton Laboratories America Inc., 13-Week Drinking Water Study in Rats with Thiourea, HLA Study No. 2319-119, Bericht an SKW Trostberg AG 1987.
59. A. Hartzell, *Contr. Boyce Thompson Inst.* **12** (1942) 471 – 480.
60. R. P. Jones, *J. Pathol. Bacteriol.* **58** (1946) 483 – 493.
61. J. McCann, E. Choi, E. Yamasaki, B. N. Ames, *Proc. Natl. Acad. Sci. USA* **72** (1975) 5135 – 5139.
62. H. S. Rosenkranz, L. A. Poirier, *JNCI J. Natl. Cancer Inst.* **62** (1979) 873 – 891.
63. V. F. Simmon, *JNCI J. Natl. Cancer Inst.* **62** (1979) 893 – 899.
64. K. Ziegler-Skylakakis, S. Rossberger, U. Andrae, *Arch. Toxicol.* **58** (1985) 5 – 9.
65. M. L. Augustine, N. K. Poulsen, J. E. Heinze, *Environ. Mutagen.* **4** (1982) 389 – 390.
66. V. F. Simmon, *JNCI J. Natl. Cancer Inst.* **62** (1979) 901 – 909.
67. R. J. Pienta, J. A. Poiley, W. B. Lebberz, *Int. J. Cancer* **19** (1977) 642 – 655.
68. E. Vazquez-Lopez, *Br. J. Cancer Res.* **3** (1949) 401 –414.
69. A. Gorbman, *Cancer Res.* **7** (1947) 746 – 758.
70. H. D. Purves, W. E. Griesbach, *Br. J. Exp. Pathol.* **28** (1947) 46 – 53.
71. A. Rosin, H. Ungar, *Cancer Res.* **17** (1957) 302 – 305.
72. O. G. Fitzhugh, A. A. Nelson, *Science (Washington D.C., 1883)* **108** (1948) 626 – 628.
73. S. Teramoto, M. Kaneda, H. Aoyama, Y. Shirasu, *Teratology* **23** (1981) 335 – 342.

74. M. Kern, Z. Tatàr-Kiss, P. Kertai, I. Földes, *Acta Morphol. Acad. Sci. Hung.* **28** (1980) 259 – 267.
75. J. A. Ruddick, W. H. Newsome, L. Nash, *Teratology* **13** (1976) 263 – 266.
76. R. H. Williams, G. A. Kay, *Am. J. Physiol.* **143** (1945) 715 – 722.
77. J. Schulman, R. P. Keating, *J. Biol. Chem.* **183** (1950) 215 – 221.
78. P. Slanina, S. Ullberg, L. Hammarström, *Acta Pharmacol. Toxicol.* **32** (1973) 358 – 368.
79. A. W. Winkler, E. B. Man, T. S. Danowski, *J. Clin. Invest.* **26** (1947) 446 – 452.
80. A. G. Zaslavskaja, *Klin. Med. (Moscow)* **42** (1964) 129 – 132.
81. Y. N. Talakin et al., *Gig. Tr. Prof. Zabol.* **11** (1990) 18 – 20.
82. A. Dooms-Goossens et al., *Br. J. Dermatol.* **116** (1987) 573 – 579.
83. A. Dooms-Goossens et al., *Contact Dermatitis* **19** (1988) 133 – 135.
84. D. S. Nurse, *Contact Dermatitis* **6** (1980) 153 – 154.
85. J. C. Van der Leun, *Arch. Dermatol.* **113** (1977) 1611.
86. J. K. Kellett, M. H. Beck, G. Auckland, *Contact Dermatitis* **11** (1984) 124.
87. D. R. Eaton, K. Zaw, *J. Inorg. Nucl. Chem.* **38** (1976) 1007.
88. D. Mullen, G. Heger, W. Treutmann, *Z. Kristallogr.* **148** (1978) 95.
89. Y. Mido, T. Yamanaka, R. Awata, *Bull. Chem. Soc. Jpn.* **50** (1977) 27.
90. W. Walter, K. P. Rueß, *Justus Liebigs Ann. Chem.* **743** (1971) 167.
91. J. T. Edward, I. Lantos, G. D. Derdall, S. C. Wong, *Can. J. Chem.* **55** (1977) 812, 2331.
92. T. S. Griffin, T. S. Woods, D. L. Klayman in A. R. Kartritzky, A. J. Boulton (eds.): *Advances in Heterocyclic Chemistry,* **vol. 18,** Academic Press, New York 1975, pp. 100 – 145.
93. Hoechst, EP 207 416, 1986 (H. Rentel, T. Papenfuhs); EP 529 600, 1992 (S. Dapperheld, H. Volk, K. Peter).
94. M. A. Weaver, L. Shuttleworth, *Dyes Pigm.* **3** (1982) 81.
95. Olin Mathieson Chem., US 3 636 104, 1972 (E. H. Kober, G. F. Ottmann).
96. A. V. Bogatskii, N. G. Luk'yanenko, T. I. Kirichenko, *Chem. Heterocycl. Compd. (Engl. Transl.)* **19** (1983) 577.
97. K. K. Verma, *Fresenius' Z. Anal. Chem.* **275** (1979) 287.
98. H. Nakamura, Z. Tamura, *J. Chromatogr.* **96** (1974) 195.
99. L. Beyer, J. Hartung, R. Köhler, *J. Prakt. Chem.* **333** (1991) 373.
100. S. Singh, S. P. Mathur, R. S. Thakur, M. Katyal, *Acta Cien. Indica Chem.* **13** (1987) 40.
101. S. Singh, S. P. Mathur, R. S. Thakur, L. Keemti, *Orient. J. Chem.* **3** (1987) 203.
102. *Kirk-Othmer,* **20,** 337.
103. Fuji Photo Film, JP 0 504 449, 1991 (H. Kawakami, K. Iwakura); *Chem. Abstr.* **118** (1993) 263906u.
104. S. N. Pandeya, P. Ram, V. Shankar, *J. Sci. Ind. Res.* **40** (1981) 458.
105. M. Negwer: *Organic-Chemical Drugs and their Synonyms,* 6th ed., Akademie Verlag, Berlin 1987.
106. Swiss Pharmaceutical Society (ed.): *Index Nominum, International Drug Directory 1990/1991,* Medpharm Scientific Publ., Stuttgart 1990.
107. C. R. Worthing, R. J. Hance (eds.): *The Pesticide Manual,* 9th ed., The British Crop Protection Council, Farnham 1991.
108. K. H. Büchel: *Chemistry of Pesticides,* J. Wiley & Sons, New York 1983, p. 315.
109. J. Drabek et al., *Spec. Publ. R. Soc. Chem.* **79** (1990) 170.
110. R. Wegler: *Chemie der Pflanzenschutz- und Schädlingsbekämpfungsmittel,* Springer Verlag, Berlin 1970, pp. 605 – 606.
111. *Ullmann,* 4th ed., **23,** pp. 172 – 173.
112. A. E. Miller, J. J. Bischoff, K. Pae, *Chem. Res. Toxicol.* 1988, 169.
113. R. A. Frakes, *Regul. Toxicol. Pharmacol.* **8** (1988) 207.

Thomas Phosphate → Phosphate Fertilizers

Thorium and Thorium Compounds

WOLFGANG STOLL, Hanau, Federal Republic of Germany

1. Introduction 2
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 6
3. Occurrence and Raw Materials . 8
4. Production 8
4.1. Concentration 8
4.1.1. Ore Preparation 8
4.1.2. Ore Digestion and Thorium Recovery 10
4.1.2.1. Acid Digestion 10
4.1.2.2. Alkaline Digestion 11
4.2. Fine Purification by Solvent Extraction 12
5. Important Compounds 16
5.1. Thorium Oxides 16
5.2. Thorium Halides 19
5.2.1. Thorium Tetrafluoride 19
5.2.2. Thorium Tetrachloride 19
5.2.3. Thorium Tetraiodide 21
5.3. Other Compounds 21
6. Production of Pure Thorium Metal 22
7. Metalworking 26
8. Intermetallic Compounds and Alloys 28
8.1. Thorium as the Main Component 28
8.1.1. Thorium and Carbon 28
8.1.2. Thorium and Nitrogen 28
8.1.3. Thorium and Aluminum 29
8.1.4. Corrosion-Resistant Thorium Alloys 30
8.1.5. Thorium Alloys Used in Thorium Production and Nuclear Applications 30
8.2. Alloys Containing Low Concentrations of Thorium 31
9. Uses 31
9.1. Nuclear Technology 31
9.2. Conventional Applications 33
9.2.1. Illuminants 33
9.2.2. Electron Emitters 33
9.2.3. Ceramics and Glass 34
9.2.4. Catalysts 34
9.2.5. Medicine 34
9.2.6. The Demand for Thorium 34
10. Product Quality and Analytical Methods 35
11. Toxicology 36
12. Industrial Safety 36
13. Environmental Protection 38
14. Legal Aspects 38
15. References 38

1. Introduction

Thorium has atomic number 90 and relative atomic mass 232.038. It was originally placed in group 4 of the periodic table under titanium, zirconium, and hafnium, following $_{89}$Ac. However, the $5f$ electron shell is filled in the series of elements from $_{90}$Th to $_{103}$Lr. In analogy to the lanthanides, in which the $4f$ shell is filled, these elements are therefore referred to as the actinides, thorium being the first in the series [1, 2]. Its similarity to cerium and its mineralogical occurrence in association with the rare earths and uranium confirm this placing. The number of thorium isotopes known thus far, all radioactive, is 25 [3]. The most common of these has the relative atomic mass 232, and five others occur naturally in the decay chains of uranium and thorium. First isolated by BERZELIUS in 1828 from the silicate mineral thorite, thorium attained industrial importance in the late 1800s as a component of gas mantles. By World War II, total consumption worldwide had amounted to ca. 8000 t, mainly for the production of gas mantles, reaching 250 t in some years. Thorium occurs in only a few regions of the earth's crust in minable amounts, although its average abundance is three to four times that of uranium [4]. After absorbing a neutron, thorium is transformed in several steps into fissile (i.e., fissionable by thermal neutron) ^{233}U. Thorium thus represents quantitatively by far the greatest energy reserve based on nuclear fission. Therefore, in the early years of nuclear technology, up to 900 t/a thorium was extracted and stockpiled as a strategically important raw material, and some hundreds of tonnes were irradiated in reactors, although less than 3 t of ^{233}U was obtained from this.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_001

However, for partly technical and partly economic reasons this route to energy production is now of little importance except in India. Therefore, since thorium is little utilized in the nonnuclear field, consumption since 1990 has declined to a few tens of tonnes per year.

2. Properties

2.1. Physical Properties

Formation and Decay. The most abundant thorium isotope (exceeding the abundance of the other isotopes by a factor of several millions) is ^{232}Th. Its concentration in the earth's crust obeys Harkins' rule, which states that even-numbered isotopes are more abundant than odd-numbered ones [5]. However, its concentration overall is more than twice that expected from the decrease in abundance of nuclides with increasing atomic number. The half-life of ^{232}Th (α decay) is 1.4×10^{10} a, the longest of all actinides, and a calculation based on the half-lifes of higher actinides shows that this concentration excess must be a result of the decay of primordial actinides such as plutonium and curium in the initial phases of the earth's history. Mineralogical evidence is provided by the fact that the very old thorium mineral bastnaesite, a thorium fluorocarbonate, is so far the only mineral in which traces of the longest-lived transuranic, the isotope ^{244}Pu with a half-life of 8.26×10^{7} a, have been detected [6].

As early as 1922, the isotopic ratio ^{208}Pb : ^{206}Pb and the age-proportional Th:U ratio in old ores were thought to indicate that a parent nuclide to thorium must have existed, whose half-life was estimated to be $(5.9–6.3)\times10^{7}$ a [7]. This closely matches that of ^{244}Pu.

Alpha decay of ^{232}Th leads to ^{228}Th, while in the uranium decay chain, ^{238}U gives the short-lived ^{234}Th and ^{230}Th, and decay of ^{235}U gives ^{231}Th and ^{227}Th. Because of the low concentration of ^{235}U (0.711 %) in natural uranium and the short half-lifes of the last two thorium isotopes, these are of little quantitative significance.

The thorium decay chain from ^{232}Th to ^{208}Pb is shown in Figure 1 [8].

The concentration ratio of ^{232}Th to ^{228}Th at equilibrium corresponds to the ratio of their half-lifes. In ^{232}Th irradiated in a reactor, an even higher proportion of ^{228}Th and of the follow-on daughter nuclides from ^{232}U decay chain is present, which is indicated, e.g., by the intensified 2.6-MeV γ-radiation from ^{208}Tl.

Neutron absorption leads to the formation of ^{233}Th from ^{232}Th. By two subsequent β decays ^{233}U is formed which decays to form the long-lived ^{229}Th (half-life 7340 a). The other thorium isotopes with mass numbers 212 – 225 are very short-lived α-emitters formed by decay of the corresponding uranium isotopes, and the isotopes ^{235}Th and ^{236}Th are short-lived β-emitters; all are produced by artificial transmutation and are of no industrial importance.

Like all other actinides, ^{232}Th also undergoes spontaneous fission with neutron emission, although, being the first member of the series, it has the longest half-life (i.e.,$>10^{21}$ a). The xenon isotopes expected as fission products can be found in old thorium ores, which also contain 1 – 10 cm^3 of helium per gram from α-decay. The helium can be boiled out with water.

Where tetravalent uranium is present in primary ores, it is usually associated with thorium, whose principal valence is also 4. Therefore, the thorium isotopes present come from all three decay chains (^{235}U, ^{238}U, ^{232}Th). The isotope ^{230}Th (ionium), a decay product of ^{234}U, is important because of its relatively long half-life (75 400 a) and, at equilibrium, is present at a level of 1.76×10^{-3} % of the proportion of uranium in the thorium.

The activity of freshly separated ^{232}Th passes through two activity maxima due to the formation of decay products. The first of these, which gives approximately a threefold increase in activity, is due to the six short-lived decay products of ^{228}Th (half-life 1.9 a) and occurs after ca. 5 weeks. The second increase begins after ca. 4 a when ^{228}Th is formed from ^{228}Ra (half-life 5.75 a), and the original radiation level increases almost fivefold after 40 a (see Fig. 2).

Pure thorium is a silvery-white, ductile metal, with low strength properties and chemical resistance [9]. It becomes gray on exposure to air. Finely divided thorium is pyrophoric. These disadvantages can be mitigated slightly by adding alloying elements. Solid thorium has the highest atomic density of all thorium compounds, which favors its use in nuclear reactors.

The physical properties and thermodynamic data of thorium are listed below. Thorium exists

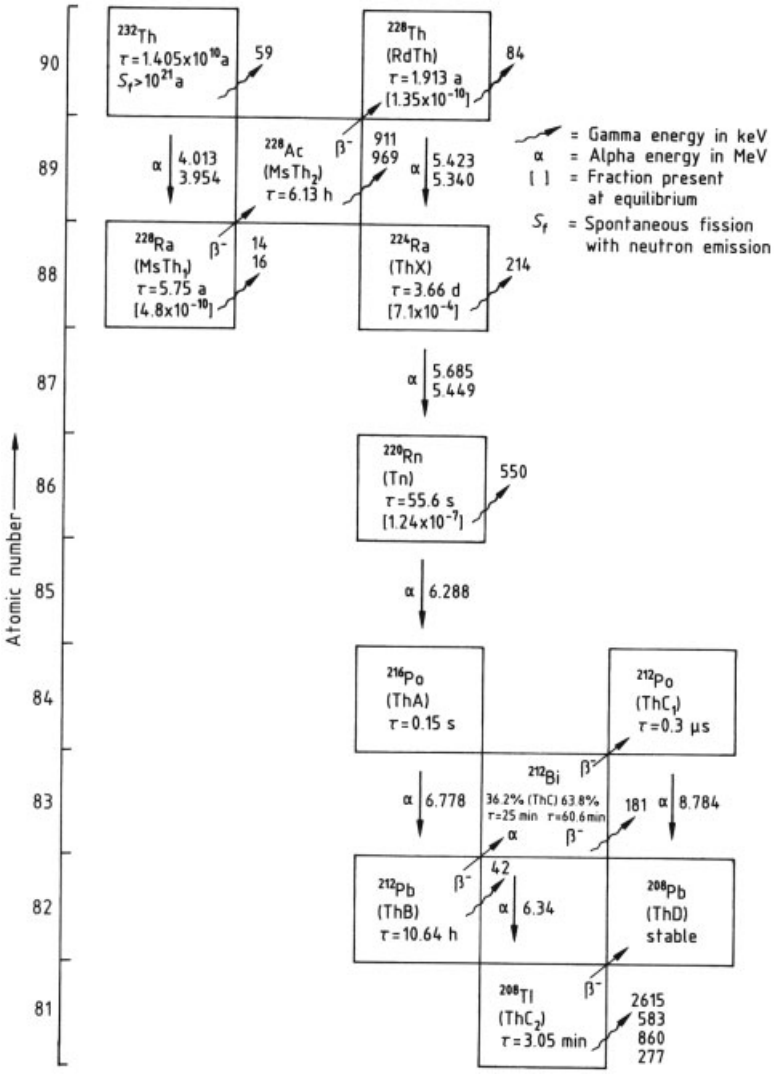

Figure 1. Thorium decay chain

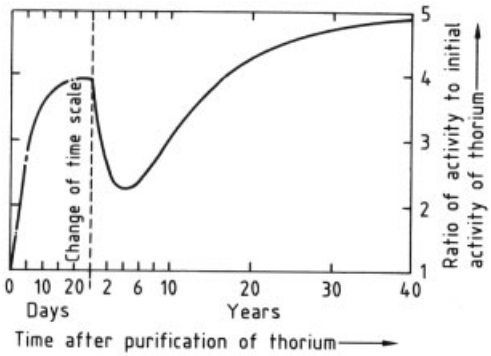

Figure 2. Change in reactivity with time of natural ^{230}Th plus ^{228}Th after separation from their daughters at time zero

in two solid phases, the α-phase (lattice constant 0.5068 nm) is stable at room temperature and is transformed to the β-phase (lattice constant 0.411 nm) at $1360 \pm 10\,°C$ (heat of transformation 3600 ± 125 J/mol). The phase transformation β-phase – liquid occurs at $1750 \pm 10\,°C$ (heat of fusion $13\,817 \pm 1256$ J/mol). Thorium has a *bp* of 4702 °C. The molar heat capacities are as follows:

α-phase	a:	25.5 J/mol
	$b \times 10^3$:	4.48 J/mol
	$c \times 10^{-5}$:	−0.51 J/mol

Table 1. Effect of production process on strength properties

Production process	Ultimate tensile strength, MPa	Tensile strength (0.2 % offset), MPa	Elongation, %
From iodide	125	42	40
By Ca reduction	175	105	60
Compressed and sintered thorium	245	175	35

where a denotes the linear, b the quadratic, and c the cubic term in the equation for the heat of transformation ($Q = a\,t + b\,t^2 + c\,t^3$)

β-phase	a:	15.70 J/mol
	b:	11.97 J/mol
liquid		46.06 J/mol

The vapor pressure p (atm) is given by

$$\log p = a - \frac{b}{T}$$

where T is the temperature in kelvin, and a and b are 27.84 and 129.67 J/mol for the α-phase and 25.10 and 124.02 J/mol for the liquid.

Thorium crystals have a density of 11.72 g/cm^3 at 25 °C; the apparent density of the cast metal is only 11.5 – 11.6 g/cm^3.

Mechanical, thermal, and electrical properties of thorium metal are very sensitive to the level of impurities, especially the carbon content. The level of impurities typically varies according to the method of production. The purest and hence most ductile metal, prepared from the iodide by the van Arkel process (see Chap. 6), is used only in exceptional cases.

The effect of the thorium production process on strength properties is shown in Table 1. The dependence of Brinell hardness (N/mm^2) on the metal treatment process is given below:

Annealing or sintering	400 – 600 N/mm^2
Moderate deformation (50 – 80 %)	600 – 800 N/mm^2
Strong deformation (10 × elongation)	1100 – 1500 N/mm^2

Figure 3 shows the temperature dependence of the mechanical impact strength as a function of carbon content. The linear coefficient of expansion of Westinghouse and Ames thorium is listed in Table 2. As in the case of carbon-containing steels, marked cold embrittlement occurs with carbon-containing thorium over a certain range of carbon content, whereas with pure thorium, embrittlement increases continuously with decreasing temperature.

Table 2. Linear thermal coefficient of expansion (α) of thorium

Temperature range, K	$\alpha \times 10^6$, K^{-1}	
	Westinghouse Th	Ames Th
298 – 473	11.2	10.8
473 – 948	12.3	11.8
948 – 1273	13.8	13.4
298 – 1273 *	12.6	12.4

* Mean value.

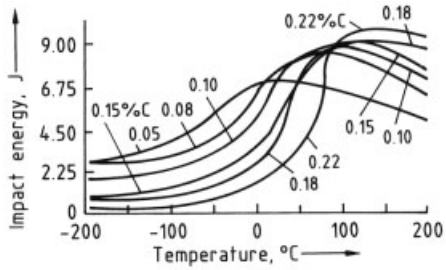

Figure 3. Effect of carbon content of thorium on impact energy [10]

The variation of thermal conductivity (W cm^{-1} K^{-1}) with temperature is given below [11]

100 °C	0.377
200 °C	0.389
300 °C	0.402
400 °C	0.419
500 °C	0.427
600 °C	0.444
650 °C	0.452

This property is affected only slightly by the content of impurities, whereas the electrical conductivity is much reduced by impurities. At room temperature, pure thorium has an electrical resistance of $14 \pm 1 \times 10^{-6}$ Ω/cm. With 0.03 % C, the figure is 18×10^6 Ω/cm, and with 0.2 % C, 37×10^{-6} Ω/cm – the temperature coefficient of resistivity being $(3.6 - 4) \times 10^{-3}$ K^{-1}.

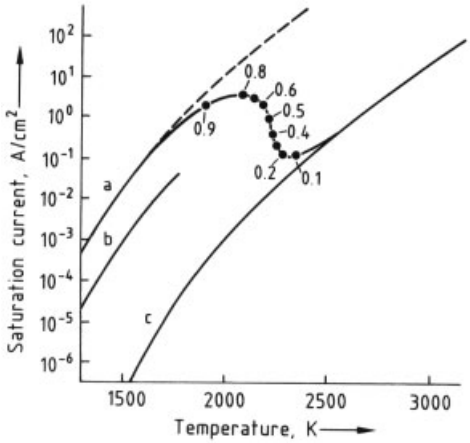

Figure 4. Effect of temperature on state of activation
a) Thoriated tungsten (monoatomic layer); b) Thorium; c) Pure tungsten

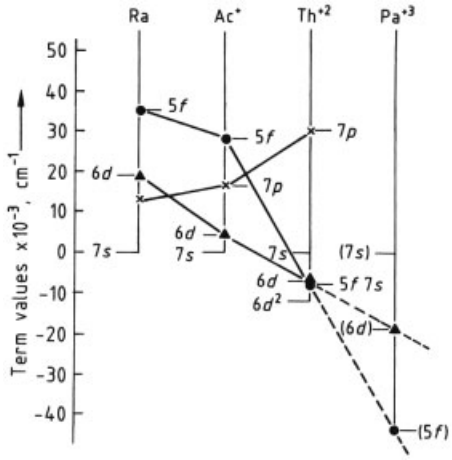

Figure 5. Relative binding energies in the radium isoelectronic sequence

Thorium metal is superconducting below 1.368 K.

Thorium has a very low work function of thermal electrons [9] (Fig. 4). The emission of electrons from "thoriated" tungsten wires is up to four orders of magnitude higher than that from pure tungsten in the lower temperature region. (For further details, see Section 9.2.)

This low work function is in fact more likely to be a property of the oxidized surface of thorium, possibly formed by a suboxide such as "ThO," rather than of the pure metal. This is suggested by the fact that in Th^{2+} the electronic transition from $5f$ to $6d$ is facilitated by the extremely small difference in bonding energies (i.e., only 570 cm^{-1}), which is better than an order of magnitude less than the $5f-6d$ transition in neighboring elements and is a result of the interaction of a valence electron with the empty $5f$ shell, which can occur only with thorium (Fig. 5) [1, p. 144].

The standard electrode potential of thorium is

$Th \rightarrow Th^{4+}$ $E_o = -1.90$ V

It is therefore somewhat more negative than zirconium or aluminum. The overvoltage of H_2 on thorium is 0.38 ± 0.01 V. Thorium cannot be deposited electrolytically from aqueous solutions, and its deposition at a mercury electrode is also doubtful. Oxide-aquates are formed instead. The production of thorium metal by fused salt electrolysis is described in Chapter 6.

2.2. Chemical Properties

Thorium is nearly always tetravalent but, exceptionally, can have a valence of two (thorium hydride) or three.

In moist air, thorium metal oxidizes slowly at its surface even at room temperature. Oxidation proceeds in two stages. First, a surface layer of the dark gray hydride is formed by reaction with atmospheric moisture, the hydride is then oxidized, with formation of a strongly bonded oxide layer. Above 130 °C the rate of oxidation increases together with ablation of the oxide film, and at 450 °C, rapid oxidation begins with the evolution of heat. Thorium burns like magnesium, with an intense flame – pulverized thorium ignites spontaneously. Above 670 °C, thorium reacts with nitrogen to form Th_2N_3. Thorium reacts with hydrogen at 300 – 400 °C, with the evolution of heat, to form Th_4H_{15}, which itself decomposes above 1100 °C, probably forming ThH_2, a gray-black powder that ignites spontaneously and is isomorphous with ZrH_2. Thorium can therefore be melted only in an inert-gas atmosphere or in a vacuum.

Thorium reacts violently with the halogens and with sulfur on heating to 450 °C, producing a flame, and it reacts with fluorine even in the cold. The oxidation state of thorium is 4 in all the compounds formed. Thorium is hardly attacked at all by water, even when hot, but reacts at 800 –900 °C with water vapor to form ThO_2.

On heating with phosphorus, finely divided boron, or carbon, tetravalent compounds are formed immediately, as in the reaction with nitrogen. Alloys are formed with all actinides and lanthanides, iron- and platinum-group metals, alkali metals (except Li), Mo, W, Ta, Zn, Zr, Hf, Be, Al, and Si.

Thorium is stable toward dissolved or molten alkali and is attacked only slowly by dilute mineral acids. It is passivated in HNO_3, but the effect disappears in the presence of small concentrations of HF or H_2SiF_6 (0.05 mol/L), and thorium is then dissolved rapidly. Thorium dissolves in aqua regia and concentrated HCl.

Thorium salts generally contain tetravalent thorium. They have only a slight tendency to hydrolyze. A pH of 3.7 is required to remove the first hydrogen ion from the water molecule according to

$$Th^{4+} + H_2O \rightarrow Th(OH)^{3+} + H^+$$

The thorium ion forms anionic complexes with a large number of ions in acidic or neutral solution (e.g., CO_3^{2-}, $C_2O_4^{2-}$, SO_4^{2-}, HSO_4^-, PO_4^{3-}, citrate, and tartrate). Thorium forms positively charged complexes with F^- if the molar ratio of F^- to Th^{4+} is<2. Complexes are also formed with an excess of thiocyanide and nitrate. Thorium has a strong tendency to form double salts such as $K_2ThF_6 \cdot 4\,H_2O$, $NaTh_2(PO_4)_3$, $MnTh(SO_4)_3 \cdot 7\,H_2O$, and $ThSO_4HPO_4 \cdot 4\,H_2O$; it also forms basic salts such as $ThOCO_3 \cdot 2\,H_2O$. The strong adsorptive forces of the Th^{4+} ion, which are exemplified by the formation of stable colloidal solutions in salt solutions with full gel – sol reversibility, especially in the system Th^{4+} – NO_3 – OH^-, is used in industrial production of ThO_2-particles of high density.

Some of the salts have very low solubility in acids (e.g., the iodates, peroxides, oxalates, fluorides, phosphates, complex sulfophosphates, ferrocyanides, and the double salt potassium thorium sulfate); these are used industrially to separate thorium. For more recent separation processes, the fact that thorium nitrate forms complexes with tributyl phosphate (TBP) and other organophosphates that are soluble in organic solvents is of great importance. These are useful for separating thorium from the rare earths (RE).

Although the readiness of the valence electrons to occupy the empty inner $5f$ shell is too small for any oxidation states other than 4 to be stable, it is deemed to be sufficient for covalent bonding. This gives thorium also a remarkable catalytic activity.

3. Occurrence and Raw Materials

Primary thorium deposits occur in acidic granitic magmas and pegmatites. They are locally concentrated, but of a small size. Secondary deposits are found in placer deposits at the mouths of rivers in granitic mountain regions (e.g., in India, Indonesia, and Brazil). In these deposits, thorium is enriched along with other heavy minerals.

Thorium, together with the related rare earths, is believed to have remained in the silicate layers during the phase separation of the earth because of their affinity for oxygen. Wherever these elements have not become bonded in the apatites formed in the initial crystallization, thorium became concentrated in the acidic residual magmas, especially in the nepheline – syenite pegmatites, where it can be found at concentrations of ca. 80 mg/kg (Hungary) compared with the average level in granites (5 – 15 mg/kg Th). If these minerals have been formed at high temperature, uranium is also present because of the isomorphism of UO_2 and ThO_2, although at lower temperature in the presence of oxygen and water, uranium can migrate away as a hexavalent ion.

Since thorium minerals are attacked only slightly by reactive components of the atmosphere (e.g., oxygen, nitrogen) and by aqueous solutions, they have been deposited in a sedimentary sequence (monazite has a density of 4.3 –5.6) along with other heavy minerals such as magnetite, rutile, cassiterite, garnet, ilmenite, and zircon in placer deposits, which sometimes also contain gold. If any thorium was present in the rivers in solution, it was precipitated together with dissolved iron as carbonate and fluoride on entering the sea.

Because of its low water solubility, thorium is scarcely present in water, soil, plants, or animals, even if the sand under a biotope contains thorium.

Table 3. Thorium minerals containing >3 % Th (mean value)

Name	Formula	Th content, %	Principal deposits
Oxides			
Thorianite	ThO_2	45.3 – 87.9	Sri Lanka
Uranothorianite	$(U,Th)O_2$	15 – 48	Canada, Madagascar
Phosphates/silicates, carbonates			
Thorite	$ThSiO_4$	25.2 – 62.7	Langesundfjord (Norway)
Orangite	$ThSiO_4$		Madagascar, Idaho
Huttonite	$ThSiO_4$		New Zealand
Uranothorite	$(U,Th)SiO_4$	20 – 42	Mid-Norway, Blind River (Canada)
Cheralite	$(Ce,La,Pr,Nd,Th,Ca,U)(PO_4,SiO_4)$	25.9 – 27.7	Travancore (India)
Monazite	$(Ca,La,Y,Th)PO_4$	3 – 26	India, Brazil, and most other deposits
Thorogummite	$Th(SiO_4)_{1-x}(OH)_x$	18.2 – 50.8	Brazil, Jisaka (Japan)
Tscheffkinite	silicates of RE, Fe, Mn, Mg, Ca, Al, Ti, Th, U	$\leq$18.4	Madagascar
Pilbarite	$PbO \cdot UO_3 \cdot ThO_2 \cdot 2\,SiO_2 \cdot 4\,H_2O$	$\leq$27.4	Western Australia
Allanite	$(Ca,Ce,Th)_2(Al,Fe,Mn,Mg)_3(SiO_4)_3OH$	$\leq$3.2	Montana, Sri Lanka
Niobates/tantalates			
Pyrochlore	$(Na_4,Ca_2,U,Th)(Nb,Ta)_4O_{12}$	$\leq$10	Colorado
Euxenite	$(Y,Ca,Ce,U,Th)(Nb,Ta,Ti)_2O_6$	$\leq$4.5	Western Australia, Nigeria

Table 4. Composition of monazite concentrates

Constituent	India	Brazil	Florida beach sand *	South Africa Monazite Rock	Malagasy Republic
ThO_2, wt %	8.88	6.5	3.1	5.9	8.75
U_3O_8, wt %	0.35	0.17	0.47	0.12	0.41
$(RE)_2O_3$ **, wt %	59.37	59.2	40.7	46.41	46.2
Ce_2O_3, wt %	(28.46)	(26.8)		(24.9)	(23.2)
P_2O_5, wt %	27.03	26.0	19.3	27.0	20.0
Fe_2O_3, wt %	0.32	0.51	4.47	4.5	
TiO_2, wt %	0.36	1.75		0.42	2.2
SiO_2, wt %	1.00	2.2	8.3	3.3	6.7

* Florida beach sand contains about 70 % monazite.
** Rare earth oxides, including Ce_2O_3.

More than 100 different ores contain thorium minerals. The most important are listed in Table 3. No deposits exist that are worth extracting for thorium alone, so it is almost always extracted as a coupled product with the rare earths, which themselves can be a side product of the production of other heavy minerals when they are in the form of monazite from a gravity concentration process.

The most widely occurring and important mineral is monazite, a phosphate of rare earths and thorium, in which the content of individual metals varies with the origin of the mineral (→ Rare Earth Elements, Chap. 2.). Typical compositions of monazite concentrates from five important deposits are given in Table 4 [12]. The currently known and extracted deposits amounting to ca. 1.3×10^6 t are in India and Malaysia (30 %); Egypt (20 %); Canada and the United States (15 %); and Brazil, Russia, and Australia (5 %); the remainder is in Indonesia, Nigeria, Nyasaland, and the Republic of South Africa. These deposits can meet the demand for the foreseeable future, so new deposits are not being sought.

The largest mining areas are in India in the province of Travancore, where the coastal sands over a distance of >70 km contain monazite in the form of honey-yellow grains up to several millimeters in size.

The sands are usually subjected first to simple gravity concentration, which yields a concentrate containing 65 – 80 % ilmenite ($FeTiO_3$), 3 – 6 % rutile (TiO_2), zircon ($ZrSiO_4$), 1 – 5 % sillimanite (Al_2SiO_5), and garnet [$(Fe, Ca, Mg)_3Al_2SiO_4$], as well as some cassiterite

(SnO_2) and a small amount of gold, together with 0.5 – 1 % monazite.

4. Production [13]

The principal production stages are the concentration of thorium minerals, extraction of thorium from them, purification, and conversion to the metal or desired compounds, usually ThO_2. Whereas up to 0.2 % impurities, including rare earths, have no detrimental effects on conventional applications, nuclear technology demands a much lower level of impurities (by one or more orders of magnitude), especially of neutron-absorbing elements such as rare earths and boron, for which additional purification processes have been developed.

4.1. Concentration

4.1.1. Ore Preparation

The methods used for ore concentration are different for primary and secondary deposits. The process used by Molybdenum Corp. in the United States is described here as a typical example of beneficiation of a *primary deposit*. The pegmatites, which are usually obtained by mining, are first coarsely and then finely ground. The ores are then subjected to a flotation process at 60 °C. This gives a 63 % concentrate of rare earths and thorium. The alkaline-earth carbonates are dissolved out of this with HCl, further concentrating the ore to 73 %. Thickening, filtration, and calcination yield a concentrate containing almost 90 % Th and RE suitable for further processing.

For ore preparation from Indian coastal sands (*secondary deposit*), the sand is subjected to gravity separation on dredges just off the coast; i.e., the sand that is collected from the seabed is wet sorted in batteries of separators. Alternatively, after removal of the coarse fraction, the sand is sorted in a dry state on the coast. The second step is magnetic sorting using a series of magnets of increasing strength. Magnetite and ilmenite are first removed by the weakest magnets; garnet is removed by magnets of the next higher strength; and coarse and fine monazite is successively removed in the next two stages, the paramagnetism of monazite being due to its rare-earth content. The monazite concentrate obtained has a purity of up to 98 %. Nonmagnetic residues are then usually treated by flotation to recover rutile, zircon, and gold, which are valuable byproducts.

4.1.2. Ore Digestion and Thorium Recovery

Monazite is fairly inert chemically. The industrial chemical ore digestion process, which was started in the last century, was first based on the simplest possible methods (i.e., treatment with hot, concentrated sulfuric acid in cast iron vessels, followed by selective precipitation by dilution with water). Because of the low rate of solution and the simultaneous presence of rare earths, thorium, and the complexing ions phosphate and sulfate in the solution, only slight variation in operation was possible, and success depended on following the formulations precisely and also on the composition and grain size of the monazite. A number of other processes were therefore developed to avoid these problems. For economic reasons, the only successful alternative method was to use alkaline digestion with hot sodium hydroxide solution. Although the process itself is a little more expensive than acid digestion, it removes phosphate and is therefore more suitable as a first stage before extractive separation to produce high-purity thorium. After some process development, it has been used mainly in the United States, India, and Brazil, where one of the target products is (or was) thorium of nuclear purity [14].

4.1.2.1. Acid Digestion

Digestion with sulfuric acid is usually carried out in two stages with 93 % acid at 210 – 230 °C (see Fig. 6). In the first stage, in which an acid excess of ca. 60 % is used, the reaction mixture thickens as reaction products are formed – first to a slurry and then to a solid gray mass. Then, fuming sulfuric acid is added, and treatment is continued for additional 5 h at the same temperature. The acid concentration chosen is the result of a compromise between the reaction rate (which increases with concentration) and the viscosity (which also increases with concentration,

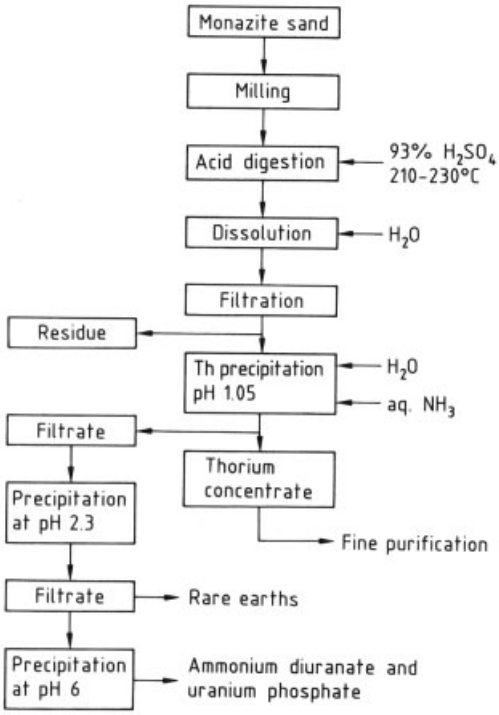

Figure 6. Digestion with sulfuric acid and thorium concentration

thereby retarding the reaction). The completeness of the reaction depends on the grain size of the monazite sand, sand:acid ratio, temperature, and reaction time. Unfortunately, increasing the reaction temperature to 300 °C brings a risk of formation of insoluble thorium pyrophosphate and must therefore be avoided. Since the dissolution reaction is strongly exothermic, an upper limit exists to the rate of addition of monazite to the acid. In contrast, below 200 °C the reaction proceeds too slowly. If stoichiometric amounts of sand and acid could be used, the required amount of acid added would be only 60 % of the sand mass. However, to keep the salts formed in solution and prevent blockage of the reactive surface of sand by the formation of precipitates, the mass of acid used must be twice the mass of sand (see Fig. 7). To reduce acid consumption, reaction times of >5 h are used to minimize the volume of solution after dilution. After the pasty mixture is cooled to 70 °C, it is diluted with ten times its volume of cold water. The salts of rare earths and thorium remain in solution, while undissolved monazite rapidly settles out because of its high density. This can be recycled, and the

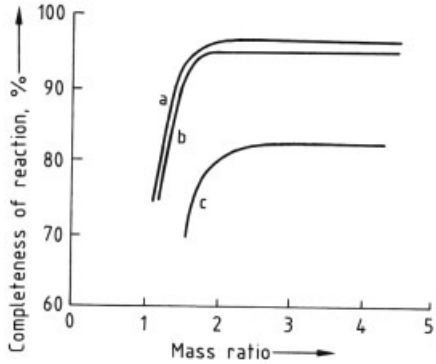

Figure 7. Effect of mass ratio of H_2SO_4: monazite sand at 300 °C and 30-min digestion on completeness of reaction
a) Completeness of reaction of thorium fraction; b) Completeness of reaction of sand (overall); c) Completeness of reaction of uranium fraction

finer and lower-density gangue materials are filtered off.

The subsequent separation of thorium and rare earths is based on the fact that thorium is almost completely precipitated as phosphate at pH values of ca. 1.3, while the main quantity of rare earths is precipitated later at a pH of ca. 2. Neutralizing agents such as sodium carbonate and ammonia must be added slowly and distributed homogeneously so that the pH limit is never exceeded, thus preventing local precipitation of rare earths, which do not redissolve. Furthermore, the slow rate of precipitation and the occlusive properties of the finely divided thorium precipitate must be kept in mind. The most uniform results would be obtained by diluting with pure water, but this is ruled out for practical reasons (i.e., excessive volume increase).

If the solution contains much uranium, simultaneous precipitation of thorium and rare earths by oxalic acid can be used – a selective, but expensive method of separation. Attempts have also been made to precipitate the rare earths first as double sulfates by adding sodium sulfate; to selectively precipitate thorium along with uranium from the solution as peroxides; or to precipitate rare earths from the concentrated solution by phosphoric acid and then evaporate the solvent to achieve precipitation of thorium as a double salt (sulfate – phosphate) [15].

An overview of the principal proven process variations is given in Figure 8. None of the processes is sufficiently selective to be used without subsequent removal of the rare earths from

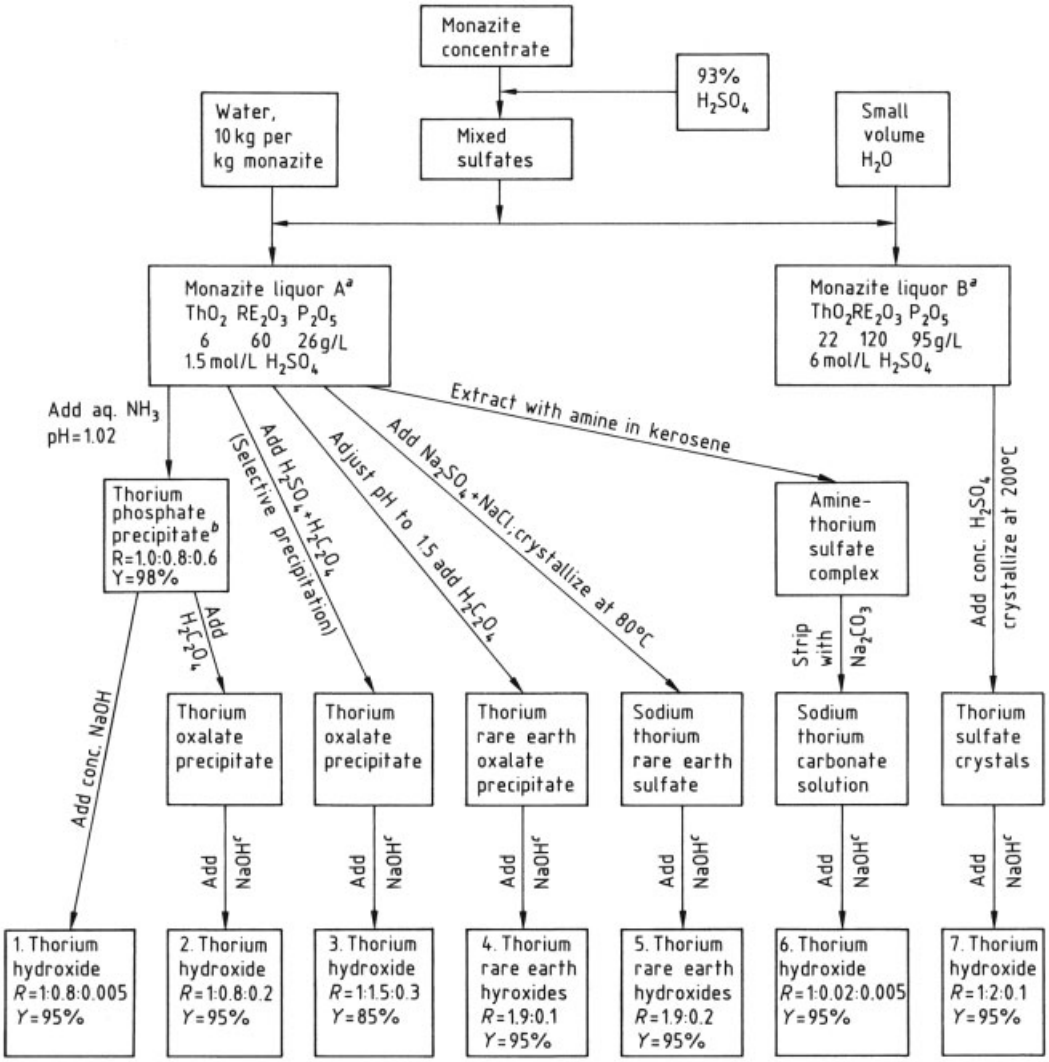

Figure 8. Principal processes for extracting thorium from monazite acid leach liquor
R = Mass ratio $ThO_2 : RE_2O_3 : P_2O_5$; Y = Approximate overall ThO_2 yield in concentrate
[a] Filtered. [b] Washed. [c] 10 % excess.

thorium and vice versa. Also, sometimes one or more rare earths are the most marketable products and sometimes thorium is, the purity requirements being different in each case.

In the early 1950s, interest in thorium of nuclear purity suddenly increased in the United States. This led to the development of the Iowa process in the Ames Laboratories of the USAEC [16] in which H_2SO_4 digestion was linked with purification by solvent extraction of the Th and RE nitrates with TBP. After acid digestion, phosphate and sulfate were removed by oxalate precipitation; the oxalates were digested with alkali to form hydroxide precipitates, which were subsequently dissolved in HNO_3. The process was functionally reliable, but expensive, and produced a number of waste streams (variation 4 in Fig. 8, with details in Fig. 9).

4.1.2.2. Alkaline Digestion

As with acid digestion, treatment with sodium hydroxide solution (Fig. 10) is limited by a number of factors. If the temperature is too high,

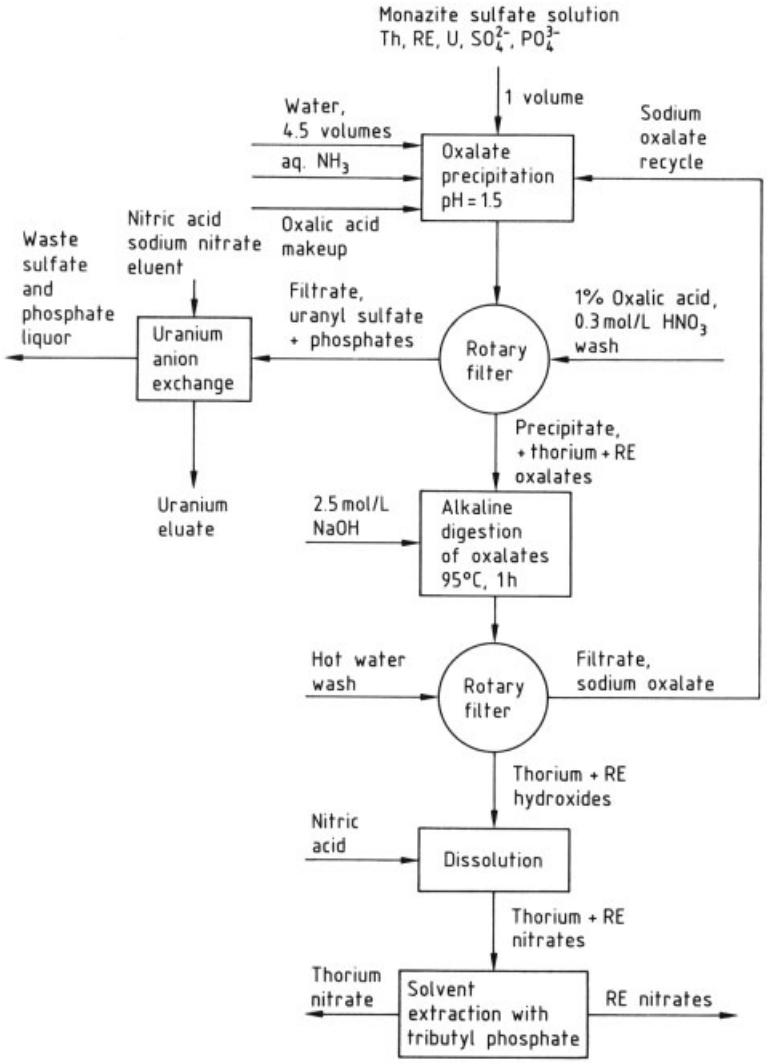

Figure 9. Iowa process for separating thorium, rare earths, and uranium from monazite sulfate

a difficultly soluble thorium oxide is formed and too much uranium appears in the filtrate. If alkali concentrations are too low, the dissolution process takes too long. The optimum values appear to be 30 – 45 % sodium hydroxide solution at ca. 140 °C and a reaction time of 3 h. However, these comparatively mild reaction conditions require finely ground monazite sand (particle size<45 µm) to ensure adequate reaction. After filtration the filter cake contains rare earths and thorium as hydroxides and uranium as sodium diuranate; all of the phosphate is present in solution in the filtrate as trisodium phosphate. On cooling this below 60 °C, $Na_3PO_4 \cdot 10\,H_2O$ crystallizes. The uranium content of trisodium phosphate increases with the amount of dissolved SiO_2, and it must be recrystallized before it becomes a salable byproduct.

The hydroxides are dissolved in 37 % HCl at 80 °C. After filtering off the undissolved material, 47 % NaOH is added, and thorium along with uranium almost completely precipitates at pH 5.8. However, the alkaline filter cake must not be completely dried; otherwise, some Ce^{4+} is formed by contact with air, and this liberates chlorine from HCl. The rare earths are precipitated by further increasing the pH. Neutralization can be carried out with NaOH residues from

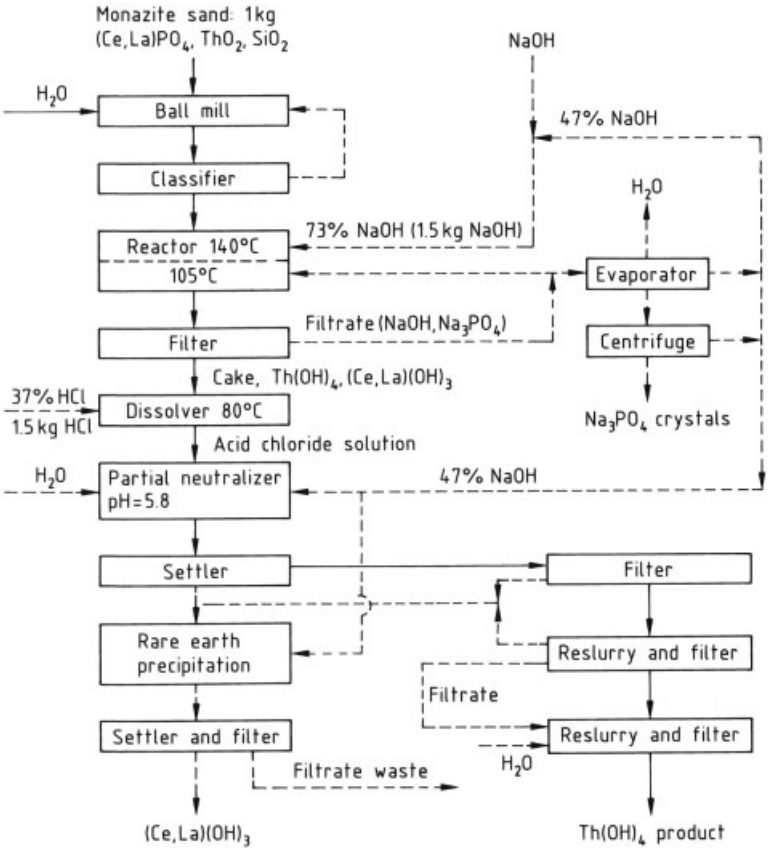

Figure 10. Caustic soda process for recovery of thorium from monazite

the filtrate from alkaline digestion, provided concentrations of phosphate are low enough as not to precipitate insoluble RE phosphates.

If thorium and uranium are to be separated from rare earths by solvent extraction, the filter cake from alkaline digestion is preferably dissolved immediately in HNO_3. However, complete dissolution of thorium cannot always be guaranteed, especially in the presence of titanium hydroxides, which bind large amounts of thorium. The good solubility of ThO_2 at 10 bar and 230 °C in HNO_3 was only discovered in 1986 [17], and has hitherto not yet been utilized on an industrial scale.

4.2. Fine Purification by Solvent Extraction

In recent decades, purification of chemically similar substances by techniques involving repeated transitions between solid and liquid phases has been replaced by technologically simpler liquid – liquid extractions. At the same time, purity requirements for thorium and its compounds have become more strict. Especially in nuclear technology, concentrations of elements with a high parasitic neutron capture cross section must be low. (For example, the maximum permitted concentrations equivalent to 4 mg/kg boron are only 8 mg/kg Sm, 2 mg/kg Er, or 1 mg/kg Gd.) Thus, the process of multiple solution and precipitation or recrystallization to separate rare earths from each other or from thorium and uranium has been replaced by multistage countercurrent solvent extraction methods. However, to give suitable distribution coefficients between the aqueous and organic systems, the chemical bonding of the ions that take part in the exchange must not be too different in the two systems. Thus, stable complexes of thorium and uranium that are soluble only in water,

especially those with sulfate and phosphate ions, are unsuitable.

Hence, attempts to extract thorium and uranium directly from monazite sulfate solutions with TBP in organic solution were unsuccessful. Only the use of extraction media with strongly cationic character enabled successful extraction from these solutions to be carried out. Whereas the anionic character of UO_2^{2+} is strong enough for it to bond to quaternary nitrogen (in amines dissolved in organic solvents), the situation is less favorable in the case of Th^{4+}. Some distribution coefficients are given in Table 5 (see next page), which also shows the clear difference between thorium and uranium. This difference favors branched primary amines.

Amex Process. The use of a combination of different amines in three consecutive cycles (see Fig. 11) is the basis of the prototype scale Amex process [18], which enables pure products (U, Th, RE) to be obtained directly from diluted solutions of monazite in sulfuric acid. However, in the absence of any large demand, this flow scheme has not yet been tested on a full scale.

Solvent Extraction with TBP. In the absence of phosphate ions (e.g., in the case of oxidic and carbonate-containing thorium ores) or if phosphate ions have been previously removed in the alkaline treatment method, the U, Th, and RE nitrato complexes can be separated and purified by extraction with TBP in an organic solvent, preferably in kerosene. The overall process is simple and easily controlled. The effect of the molarity of nitric acid on the distribution coefficients of rare earths, uranium, and thorium is shown in Figure 12. With a mean acid concentration of 3 mol/L, the distribution coefficient (organic/aqueous) for rare earths is 0.02, for thorium 1, and for uranium 20, so that a multistage separation is possible [19]. Trivalent metallic

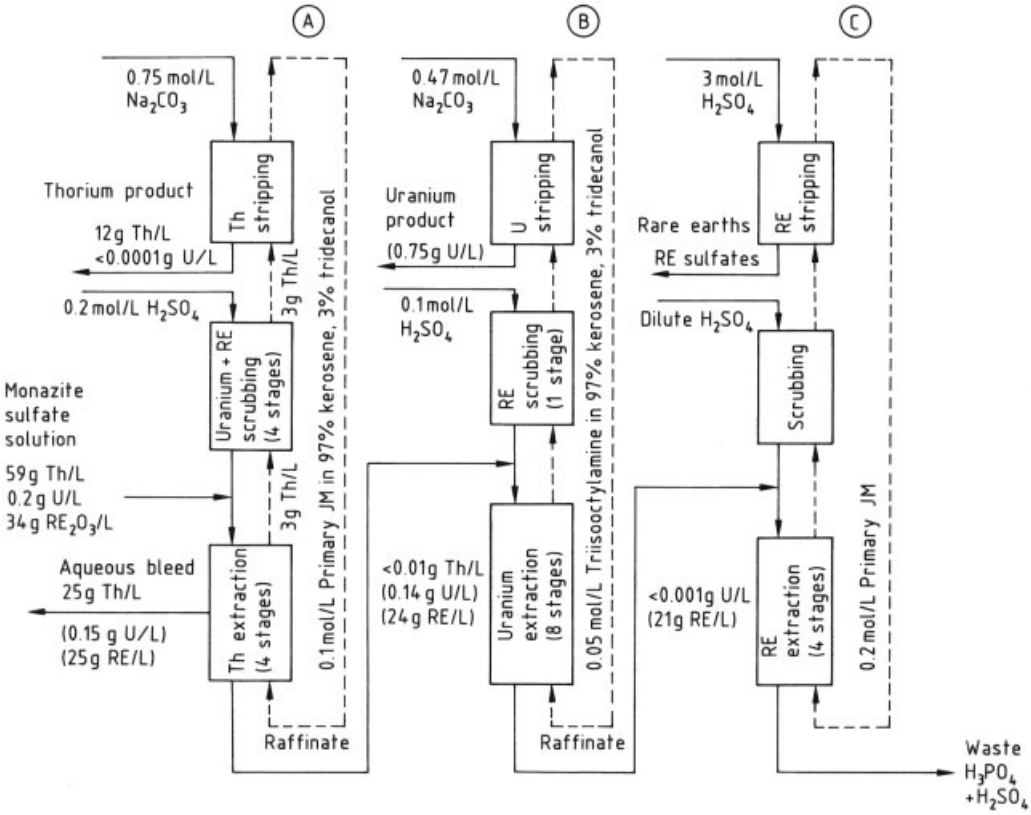

Figure 11. Separation of thorium, uranium, and rare earths from monazite by solvent extraction in Amex process
A) Thorium separation; B) Uranium separation; C) Rare earth separation

Table 5. Distribution coefficients for uranium, thorium, and cerium between organic amines and aqueous sulfate solution [a] [18]

Amine type	Examples of amines	Distribution coefficient		
		U (VI)	Th (IV)	Ce (III)
Branched primary	primary JM [b] and 1-(3-ethyl-pentyl)-4-ethyloctylamine	5 – 30	>20 000	10 – 20
Secondary with alkyl branching distant from the nitrogen	di(tridecyl)amine [c]	80	>500	<0.1
Secondary with alkyl branching on the first C	amberlite LA-1 [d] and bis(1-isobutyl-3,5-dimethylhexyl)amine	80 – 120	5 – 15	<0.05
Tertiary with no branching or branching no closer than the third C	alamine 336 [e,f] and triisooctylamine [f,g]	140	<0.03	<0.01

[a] 0.1 mol/L SO_4^{2-}; pH = 1; ≈1 g metal per liter; 0.1 mol/L amine in kerosene; 1:1 phase ratio.
[b] Trialkylmethylamine, homologous mixture, 18 – 24 carbons.
[c] Mixed C_{13} alkyls from tetrapropylene by oxo process.
[d] Dodecenyltrialkylmethylamine, homologous mixture, 24 – 27 carbons.
[e] Trialkylamine with mixed *n*-octyl and *n*-decyl radicals.
[f] Kerosene diluent modified with 3 vol % tridecanol.
[g] Mixed C_8 alkyls from oxo process.

ions such as Fe^{3+} can mask residues of phosphate. A suitable flow scheme is shown in Figure 13. Solvent extraction, which is optional in the purification of natural thorium compounds, is the only feasible method of separating thorium, ^{233}U, and fission products after irradiation in a reactor.

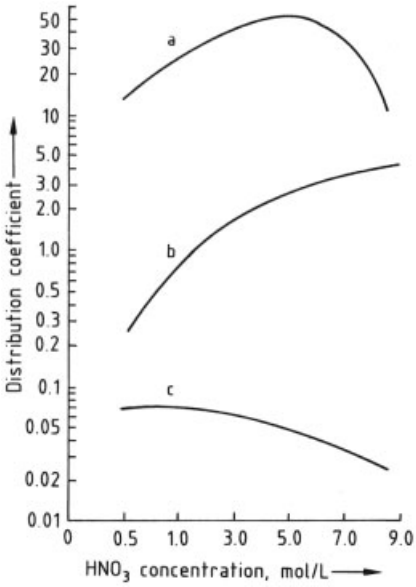

Figure 12. Effect of acid molarity on distribution coefficients (organic – aqueous phase) of nitrato complexes in TBP – kerosene extraction
a) Uranium; b) Thorium; c) Rare earths

In an intermediate route that can be used in the presence of high concentrations of sulfate but in the absence of phosphate (e.g., in residual solutions from uranium extraction in the Canadian Blind River district), thorium is extracted with di(2-ethylhexyl)phosphoric acid.

5. Important Compounds

5.1. Thorium Oxides [20]

Two compounds, ThO and ThO_2, exist in the oxygen – thorium system. The black compound ThO, which crystallizes with the sodium chloride lattice, is known only from coatings on the metal. It is probably an interstitial compound with an electron deficit structure. When ThO_2 is heated in an electric arc, ThO band structures appear in the spectrum.

The compound ThO_2 is the stable, well-known oxide, being formed if either a salt that contains a volatile anion or the oxidaquate is strongly heated. It is obtained as a very fine powder by precipitating and then strongly heating thorium oxalate, according to

$$Th(NO_3)_4 + 2\,H_2C_2O_4 \cdot 2\,H_2O \rightarrow Th(C_2O_4)_2 \cdot 2\,H_2O + 4\,HNO_3 + 2\,H_2O \text{ (at 60 °C)}$$

and

$$Th(C_2O_4)_2 \cdot 2\,H_2O + O_2 \rightarrow ThO_2 + 4\,CO_2 + 2\,H_2O \text{ (at 650 °C)}$$

At pH 1 and with careful temperature control, the reaction with oxalic acid leads to a crystalline precipitate with good filtering properties. Precipitation can be carried out selectively also in the presence of trivalent metals, uranium, and

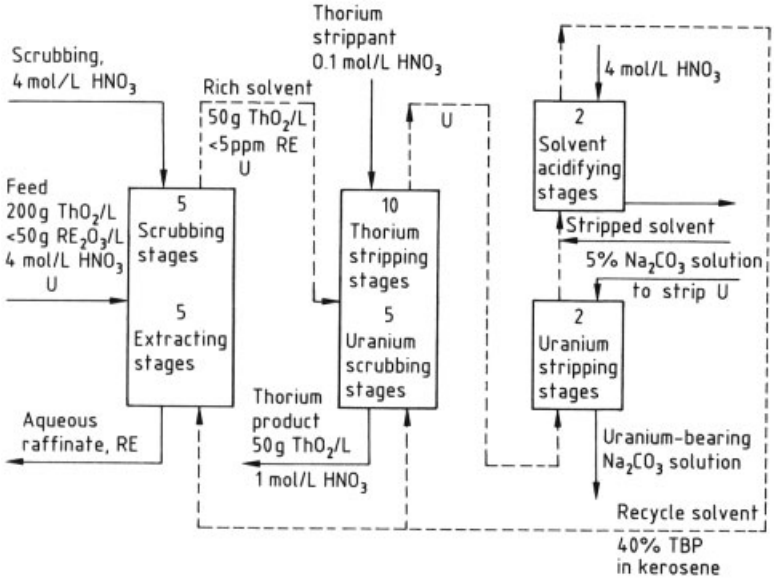

Figure 13. Thorium purification by solvent extraction with TBP
– Aqueous phase; - - - 40 vol % TBP in kerosene

titanium. Several other 1,2-dicarboxylic acids have been tested as alternatives, but only tetrachlorophthalic acid gave significantly improved selectivity against the rare earths [21].

The structure, decomposition temperature used, and calcination temperature can affect the rate of reaction of ThO_2 in other reactions. For example, ThO_2 freshly prepared by low-temperature thermal decomposition of thorium oxalate (*meta*-ThO_2) is catalytically active and fairly soluble in acids, whereas thorium oxide calcined at high temperature is very slowly attacked by acids. It can be dissolved only in hot, concentrated H_2SO_4 or in concentrated HNO_3 in the presence of F^- ions. Physical properties of ThO_2 are listed below [22]:

Color	white
Crystal system	face-centered cubic
Density (X-ray, 25 °C)	10.00 g/cm^3
Linear expansion from 25 °C	9×10^{-3} at 1000 °C; 20×10^{-3} at 2000 °C
Thermal conductivity	
(100 °C)	0.10 W cm^{-1} K^{-1}
(600 °C)	0.04 W cm^{-1} K^{-1}
Melting point	3370 ± 30 °C
Entropy (25 °C)	65.3 J K^{-1} mol^{-1}
Heat of formation from elements (25 °C)	−1227.6 kJ/mol
Free energy of formation from elements (25 °C)	−1170.0 kJ/mol

The vapor pressure (atm) between 2180 and 2865 K can be calculated according to:

$$\log p = 8.00 - 3517/T$$

(ThO_2 dissociates partially into ThO and O; at low oxygen pressure the vapor pressure is somewhat higher.) The temperature dependence of the molar heat capacity (J/mol) between 298 and 3000 K is

$$C_p = 69.3 + 9.71 \times 10^{-3}T - 9.19 \times 10^{-5}T^{-2}$$

Two crystal forms are known, i.e., a tetragonal modification similar to rutile that crystallizes from borate melts and a cubic form that crystallizes from phosphate melts. The lattice constants and densities are very similar to those of the metal since the oxygen occupies interstitial locations. However, these densities are not reached by sintered oxides (maximum values: ca. 10.2 g/cm^3). Thorium dioxide melts at 3370 °C and thus has the highest melting point of all ceramic materials. The tendency of *meta*-ThO_2 to form colloidal solutions (thorium oxidaquate) in the presence of nitrate ions is important, because these can be dried to form stable gels that can be sintered to give high-density ceramic bodies. Another important property of ThO_2 is the emission of long-wave red light on

heating (a historical test of purity), and the shifting of the emitted wavelength further into the visible region by the addition of rare earths, especially cerium, which is an essential feature of gas mantle technology (see Section 9.2 for further details).

Coatings of ThO_2 on tungsten are used to promote the emission of electrons from incandescent cathodes. These coatings must be thin and uniform; they are produced by a simple process of dipping the cathodes into suspensions of thorium oxidaquate and various finely dispersed thorium oxide slurries or by electrophoretic deposition of the oxide from alcoholic suspensions. They must be "formed" by heating the cathodes to 2900 – 3000 °C.

Thorium oxide forms mixed crystals with a number of oxides of the actinides (e.g., U, Np, and Pu), with no miscibility gap. No phase transformation occurs up to the melting point.

When ThO_2 is used to form ceramics, for dimensional accuracy of the preformed bodies (crucibles, nuclear fuel pellets) the powders used must shrink by a small but definite amount during sintering/densification. Choice of the powder involves a compromise between less good sintering properties and minimized shrinkage. The *meta*-ThO_2 is preferably precalcined in an atmosphere of steam at 800 – 1000 °C before itis formed by pressing or slip casting. The effect of precipitation temperature and calcination temperature on the specific surface area of the powder is shown in Figure 14 [23].

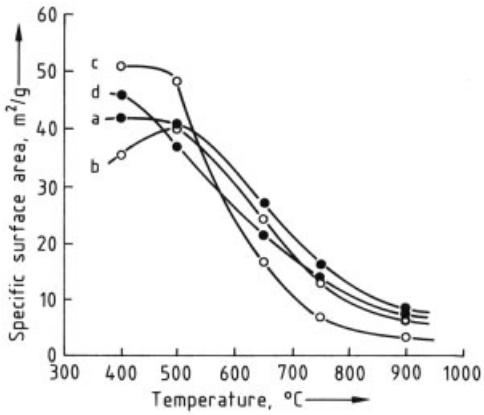

Figure 14. Effect of temperature on the specific surface of thorium oxide from thermal decomposition of thorium oxalate
a) 40 % Precipitation; b) 10 % Precipitation; c) 100 % Precipitation; d) 70 % Precipitation

5.2. Thorium Halides

An overview of the properties of thorium halides is given in Table 6.

5.2.1. Thorium Tetrafluoride

Precipitation of thorium salt solutions with HF leads to the formation of gelatinous, white precipitates of ThF_4 hydrates with 8, 4, or 2 molecules of water of crystallization, but these hydrolyze to form mixtures of ThO_2 and $ThOF_2$ even on careful drying. Anhydrous ThF_4 can be obtained only by fuming ThF_4 hydrates many times with NH_4F or by dehydration in a stream of dry HF at 300 – 400 °C. Thorium tetrafluoride is therefore produced industrially by reaction of ThO_2 and HF in the gas phase:

$$ThO_2 + 4\,HF \rightarrow ThF_4 + 2\,H_2O$$

At 566 °C the reaction is exothermic and at first proceeds rapidly, but thermodynamic equilibrium is attained very quickly, leaving an unreacted excess of HF. At lower temperature, the reaction goes to completion but proceeds much more slowly. A schematic of the production of thorium fluoride from the nitrate via the oxalate, as carried out by the National Lead Company (Fernald, Ohio), is shown in Figure 15 [11, p. 153]. The plant consists of four stainless steel tubular reactors arranged in series, each provided with exterior heating and interior screw conveyors made of Inconel. A countercurrent flow system with graded temperature increases utilizes HF to give complete conversion of ThO_2. However, to obtain ThF_4 that does not contain oxide – fluoride, it was necessary to use excess anhydrous HF and produce and sell byproduct aqueous HF (70 %). New plants will probably be operated as fluidized-bed reactors following experience with uranium.

Thorium tetrafluoride is formed as a solid residue when enriched $^{235}UF_6$ is stored. The ^{234}U contained in uranium, which is also enriched, decays to form $^{230}ThF_4$. This remains behind after steaming out the UF_6 container and must be scrupulously removed because of its high α-toxicity.

In the electrolytic production of thorium from a fluoride melt, the double fluoride $KThF_5$ can be used as the starting material. The $KThF_5$ can

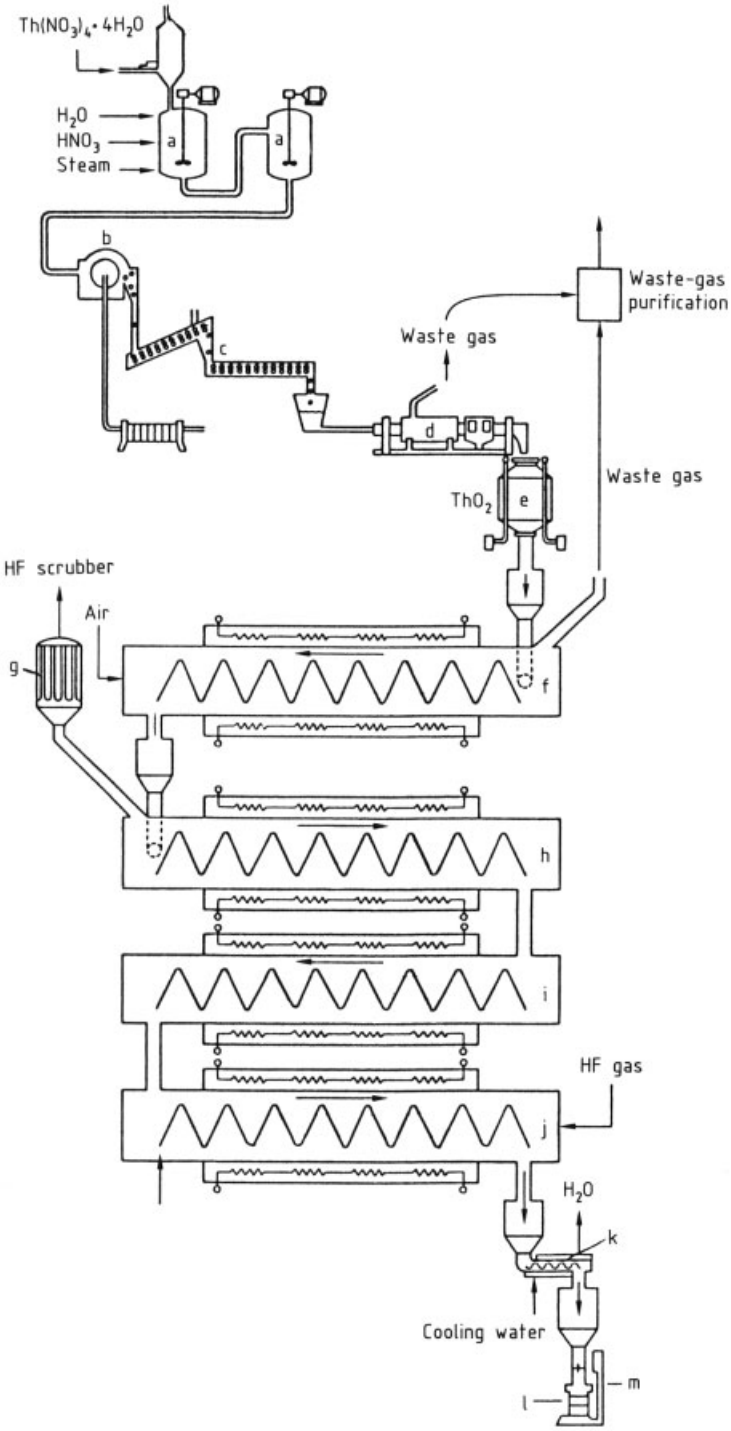

Figure 15. Production of ThF_4 by calcination and fluorination
a) Thorium oxalate precipitation; b) Drum filter to filter off $Th(C_2O_4)_2$; c) Band dryer; d) Calcination; e) Weighing; f) Calcination (650 – 675 °C); g) HF scrubber; h) Hydrofluorination (260 °C); i) Hydrofluorination (370 °C); j) Hydrofluorination (566 °C); k) Cooling trough; l) ThF_4 collector; m) Weighing

Table 6. Physical properties of thorium tetrahalides

Property	ThF_4	$ThCl_4$	$ThBr_4$	ThI_4
Color	white	white	white	yellow
Density at 25 °C, g/cm^3	6.12	4.62	5.72	6.00
Crystal system				
Low-temperature form	monoclinic	tetragonal	orthorhombic	monoclinic
High-temperature form		orthorhombic	tetragonal	orthorhombic
Transition temperature, °C		(406)	≈420	?
Melting temperature, °C	1110	770	679	570
Normal boiling point, °C	1782	942	905	853
Vapor pressure, atm				
$\log p = A - B/T$				
Solid, A	9.345	9.426	9.498	9.747
B, K	17 089	10 630	10 151	9894
Liquid, A	6.395	5.229	5.260	5.714
B, K	13 080	6346	6187	6425
Heat capacity,				
$C_p = A + 10^{-3}\,BT - 10^5\,C/T^2$				
Solid, A, J mol^{-1} K^{-1}	112.0	120.4	127.7	129.8
B, J mol^{-1} K^{-1}	24.51	23.84	15.1	13.0
C, J mol^{-1} K^{-1}	7.56	6.15	6.15	6.15
Liquid, J mol^{-1} K^{-1}	152.8	167.5	171.7	175.9
Heat of transition, J/mol		5.02	4.2	?
Heat of fusion, kJ/mol	44.0	61.5	(54.4)	48.1
Heat of vaporization at normal boiling point, kJ/mol	233.2	?	111.0	115.1
Heat of formation at 25 °C, kJ/mol	−2112.8	−1187.4	−966.3	−664.9
Free energy of formation,				
$\Delta G = A + BT$				
Solid, A, kJ/mol	−2102.5	−1177.0	−1019.3	−783.1
Solid, B, J mol^{-1} K^{-1}	293.6		283.5	290.8
Liquid, A, kJ/mol	−2031.0	−1093.3	−942.9	−716.0
B, J mol^{-1} K^{-1}	240.3	201.4	211.7	210.4

be precipitated by adding KF to a solution of thorium nitrate according to

$$Th(NO_3)_4 \cdot 4\,H_2O + 5\,KF \cdot 2\,H_2O \rightarrow KThF_5 + 4\,KNO_3 + 6\,H_2O$$

and is dried at 120 °C to give a product free of water of crystallization.

5.2.2. Thorium Tetrachloride

Depending on conditions, thorium tetrachloride crystallizes from aqueous solution with 11, 9, 8, 7, 4, or 2 molecules of water of crystallization, but oxide – chlorides are formed on drying. Anhydrous $ThCl_4$, like anhydrous ThF_4, can be produced only by reaction in the gas phase. All reactions that simultaneously reduce and chlorinate ThO_2 are suitable (see overview in Table 7), for example, chlorination of a mixture of ThO_2 and C above 600 °C:

Table 7. Reaction of chlorine and ThO_2 with reducing agents

ThO_2 type	Reducing agent	Temp., °C	Time, h	Yield, %
$Th(C_2O_4)_2$	carbon	700	2	99
	CCl_4	800	4	10
↓ >675 °C	sulfur	800	2	98
	C_3Cl_8	400	5	25
ThO_2	NH_4Cl	425	4	0
$Th(C_2O_4)_2$	carbon	600	2.5	99
	CCl_4	300	5.5	92
↓ 445 to 525 °C	sulfur	500	3	99
	C_3Cl_8	400	4	92
ThO_2	NH_4Cl	500	1.5	99

$$ThO_2 + 2\,C + 2\,Cl_2 \rightarrow ThCl_4 + 2\,CO$$

Because the $ThCl_4$ formed contains carbon as an impurity, it must be distilled at 942 °C. This step is complex due to the very hygroscopic and aggressive nature of the chloride. Distillation is not necessary if thorium oxalate is chlorinated with CCl_4 with a small addition of chlorine as a catalyst according to

$$Th(C_2O_4)_2 + CCl_4 \rightarrow ThCl_4 + 2\,CO + 3\,CO_2$$

The reaction is carried out at temperatures up to 525 °C in several stages in a graphite reactor.

5.2.3. Thorium Tetraiodide

Thorium tetraiodide has an intense yellow color, an *mp* of 566 °C, and a *bp* of 837 °C, and can be produced by reaction of the elements. Thorium tetraiodide reacts with excess thorium to give subiodides of trivalent and divalent thorium, depending on the relative amounts of reactants and the temperatures and pressures used. Like the other halides, it can in principle be produced by reaction of intimate mixtures of carbon and ThO_2 with elementary iodine. In the production of pure metallic thorium, the main use of ThI_4 , itis not isolated as a separate product but is used only as an intermediate for gas-phase matter transfer between crude and pure thorium in the van Arkel – de Boer process (see Chap. 6) [24].

5.3. Other Compounds

Thorium Nitrate. The compound from which almost all other thorium compounds are produced is thorium nitrate, $Th(NO_3)_4 \cdot 6\,H_2O$, which is highly soluble in water at room temperature (65 wt %). Evaporation of the aqueous solution yields the tetrahydrate (one of seven hydrates) in the form of a finely crystalline white powder [25].

Thorium Hydrides. The hydrides ThH_2 and ThH_4 are known. On reduction of ThO_2 with excess magnesium in the presence of H_2, a somewhat substoichiometric, highly flammable powder is formed, which is isomorphic with ZrH_2 and pseudoisomorphic with ZrC_2 and ThC_2. The other compound, which is also substoichiometric, $ThH_{3.75}$, has a cubic crystal structure that does not resemble that of the metal. Both compounds can be used to produce finely divided thorium metal since they release all their hydrogen at 900 °C. Anhydrous thorium halides can be obtained by reaction of the hydrides with hydrogen halides.

Thorium hydroxide, $Th(OH)_4$, is formed by precipitation from solutions of salts by adding alkali. This compound is not a true hydroxide, but an oxidaquate ($ThO_2 \cdot x\,H_2O$), a stable colloid that remains amorphous up to 340 °C.

Thorium peroxide, $Th_2O_7 \cdot 4\,H_2O$, is formed as a sparingly soluble precipitate when H_2O_2 is added to thorium salt solutions. Only the actinides react in this way. The reaction can be used to separate thorium from the rare earths.

Thorium Nitrides. The compounds Th_2N_3 and ThN are known. The first, Th_2N_3, is a hexagonal, dark red compound, which is formed from the elements, and decomposes in the presence of water to form NH_3 and ThO_2.

ThN has the face-centered cubic structure like the nitrides of uranium, lanthanum, and cerium, in which nitrogen has a quasimetallic character similar to that of carbon.

Thorium Phosphide and Borides. The corresponding phosphide ThP_4 is stable toward all the usual solvents. The two borides ThB_4 and ThB_6 are also very stable, hard materials, the latter forming dark red crystals.

Thorium Carbides. The mono- and dicarbide exist, although both compounds have a substoichiometric composition (i.e., 0.97 and 1.90). The monocarbide crystallizes with a face-centered cubic structure. The dicarbide is monoclinic up to 1410 °C and tetragonal above this temperature. The densities of ThC and ThC_2 are 9.6 and 8.76, respectively, and the melting points 2500 °C and 2640 °C. Depending on the availability of carbon, the carbides are formed readily from the elements or by reaction of an intimate mixture of graphite and thorium oxide at 1400 °C:

$$ThO_2 + 4\,C \rightarrow ThC_2 + 2\,CO$$

ThC has a metallic appearance and a thermal conductivity comparable to that of thorium metal. The dicarbide is readily oxidized to ThC at high temperature in the presence of oxidic substances. Partial mutual substitution of O_2, N_2, and carbon is possible, leading to relatively thermally stable compounds of the type $Th(O_{1-(x+y)}C_xN_y)$ when the sum of x and y is <0.6. The formation of solutions of carbon in thorium is described in Section 8.1.1.

Thorium sulfides are not readily formed from aqueous solution, but are formed by reaction of the elements. The compound ThS, which has a metallic appearance, is formed first and then converted in several stages to ThS_2 via the orthorhombic Th_4S_7 and other intermediates.

Thorium sulfate is important mainly because it is formed when monazite is digested with H_2SO_4. Thorium sulfate forms five different hydrates, which contain up to nine molecules of water of crystallization. Also, sparingly soluble double salts are formed with alkali-metal sulfates.

Thorium Carbonate. When thorium salt solutions react with alkali-metal carbonate, a poorly defined basic carbonate of the type $ThOCO_3 \cdot 8\,H_2O$ precipitates. It dissolves in an excess of alkali-metal carbonate (similarly to the corresponding uranium) to form relatively stable complex double carbonates of the type $Na_6Th(CO_3)_5$ or $(NH_4)_2Th(CO_3)_3 \cdot 6\,H_2O$. Thorium hydroxide does not precipitate from solutions of these substances.

6. Production of Pure Thorium Metal

Because of thorium's strongly negative standard electrode potential, its compounds cannot be reduced to the metal with carbon or hydrogen. The route to thorium metal is therefore reaction of the oxide or a halide with metals, electrolysis in molten salts, or thermal decomposition of a halide. The principal reactions that have led to production processes are summarized below:

1) Electrolysis of fused salts:
   $KThF_5$ in NaCl
   ThF_4 in NaCl – KCl
   $ThCl_4$ in NaCl – KCl
2) Reduction with reactive metals:
   ThO_2 with Ca
   $ThCl_4$ with Mg
   ThF_4 with Ca
3) Thermal dissociation of ThI_4

The results of the processes are listed in Tables 8, 9, 10, 11 [11, pp. 170 – 171].

Table 8. Production of thorium metal by reduction of halides or double halides

Thorium halide	Reducing agent	Product
$ThCl_4$	Na or K	Th powder
$KThCl_5$	Na or K under H_2 atmosphere	Th powder
NH_4ThCl_5	Al	alloy
$ThCl_4$ (70 %) in $ThOCl_2$ (30 %)	Ca	powder
$ThCl_4$	Mg	powder
$ThCl_4$	$Na(Hg)_x$	Th – Hg alloy
ThF_4	Ca with $ZnCl_2$ booster	Zn – Th alloy biscuit(sponge)

Table 9. Production of thorium metal by reduction of the oxide

Thorium source	Reducing agent	Product
ThO_2	Mg	impure alloy
ThO_2	Al	Th beads
ThO_2	Ca	Th powder
ThO_2	Ca and $CaCl_2$	Th powder

Table 10. Production of thorium metal by thermal decomposition of halides

Thorium source	Reducing agent	Product
ThI_4	hot-wire technique $ThI_4 + \text{heat} \longrightarrow Th + 2\,I_2$	crystal bar of thorium

Table 11. Production of thorium metal by electrolytic methods

Thorium source	Electrolyte	Product
Aqueous solution		
$Th(OH)_4$	HBF_4	?
Thorium salts	$HSO_3 \cdot NH_2$	?
Organic solution		
Thorium salts	C_2H_5OH, CH_3OH, and other alcohols	thorium organic compounds
Complex Th salts	organic acids	?
$Th(NO_3)_4$	C_2H_5O with 15 – 20 %	
	Me_2CO_3	ThO_2
Fused salt solution		
$ThCl_4$	KCl – NaCl	Th with ThO_2 and C
$KThF_5$	NaCl – KCl	Th powder
$KThF_5$	NaCl	Th powder
$ThCl_4$	NaCl and KCl – NaCl	Th powder
ThF_4	NaCl – KCl	Th powder or Th in liquid Zn cathode

The tolerable impurity level of intermediate products to be met when thorium metal is used in the nuclear industry are given in Table 12.

Table 12. Typical analyses of pure thorium products

Element	$Th(NO_3)_4$	ThO_2	ThF_4
Th, % *	40.3	86	75.4
U, mg/kg *	8	2.5	16.3
Traces, mg/kg **			
Al	13	<10	<10
B	5.5	0.4	1.2
Ba	<20		<20
Be	<1	<1	<1
Bi	<2	<2	<2
Ca	<200	<200	
Cd	<0.2	<0.2	<0.2
Co	<1	<1	
Cr	<4	8.1	38
Cu	2	46	81
Fe	26	46	161
Mg	25	8	24
Mn	<1	2	6
Mo	<1	<1	5
Ni	3	15.6	338
P	<40	<40	<40
Pb	2	27	24
SiO_2	109	207	33
Sn	<1	6	
V	<10	<10	
Zn	<10	<10	29

* Chemical analysis.
** Spectrochemical analysis.

In the production of thorium metal, problems arose because of its highly reactive nature and sensitivity to both air and moisture, as well as its high melting point (1750 °C). The following four production methods have been tested on a full scale:

(1) reduction of fluorides and chlorides; (2) reduction of ThO_2; (3) thermal decomposition of ThI_4; and (4) fused salt electrolysis.

Reduction of ThO_2 with calcium metal was first carried out in the United Kingdom at 1200 °C on a pilot plant scale in a reactor lined with CaO. In the United States, the reaction was carried out in molten $CaCl_2$ in an argon atmosphere to provide a more coarse-grained product, the heat of reaction being insufficient to melt the thorium. Both processes give thorium sponge, which requires many further processing stages. These production processes were abandoned because of the high temperatures required and the cost of the apparatus.

Fused Salt Electrolysis. In fused salt electrolysis, $KThF_5$ is used as a 15–20 % mixture with NaCl to give a eutectic. Thorium is produced in an argon atmosphere as a nonadherent finely crystalline powder deposited on molybdenum cathodes in graphite cells at 900 °C, using a current density of 20 A/dm^2. In fused salt electrolysis, only Cl_2 is formed at the anode, and the buildup of thorium powder in the electrolysis vessel causes short circuits. Thus, batchwise operation must be used. A continuous method of operation is possible if $ThCl_4$ is used in a mixture of NaCl and KCl as electrolyte at 800 °C with strict exclusion of moisture. A very pure, coarse-grained thorium metal is obtained. Plants of this type were operated briefly in the United Kingdom, the United States, and the Soviet Union.

Thermal decomposition of ThI_4 yields an extremely pure, ductile thorium metal. Thorium iodide is usually produced from impure metal pieces, which are first heated in a muffle furnace to>500 °C for 24 h in a vacuum, and then reacted with iodine vapor, which is vaporized at 260 °C. Thorium iodide is formed between 450 and 480 °C, and purified by distillation. Thorium metal is deposited from the iodide on incandescent filaments of tungsten or thorium at 900–1700 °C. The temperature range between 500 and 900 °C must be avoided because the tetraiodide decomposes to form thorium and nonvolatile ThI_3 at a little over 600 °C. The total reaction (formation of ThI_4 and decomposition) is carried out in long, stationary, heated tubes filled with iodine vapor. Concentric perforated sheet metal tubes retain unreacted pieces of crude thorium at the walls of the tube, and pure thorium is deposited on heated filaments at the center. The thickness of the thorium rod that forms in the center is limited only by the electric current used, which eventually reaches several hundred amperes. Both decomposition of the iodide and electrolysis give very pure metal, but these processes do not give high throughputs or yields in the present state of the technology.

Reduction of Halides. The larger quantities of thorium metal used in the nuclear industry are produced on an industrial scale mainly by the reduction of halides. The fluoride and chloride are not equivalent, as can be seen from the reaction enthalpies (Table 13). Thorium fluoride can be reduced only by calcium, whereas a choice of sodium, calcium, and magnesium exists for the reduction of the chloride. Two processes were developed in the United States almost simultane-

Table 13. Free-energy change in reduction of thorium compounds to metallic thorium

	Oxide		ThF_4		$ThCl_4$
Reaction temperature T, K	1223		2023		1150
Free energy of formation from elements at T, kJ/mol					
	ThO_2 (s)	−998.4	ThF_4 (l)	−1545.7	$ThCl_4$ (l)	−861.8
	$2\,Na_2O$ (s)	−487.7	4 NaF (l)	−1303.0	4 NaCl (l)	−1219.7
	$2\,MgO$ (s)	−934.4	$2\,MgF_2$ (l)	−1472.1	$2\,MgCl_2$ (l)	−931.9
	$2\,CaO$ (s)	−1015.7	$2\,CaF_2$ (l)	−1774.5	$2\,CaCl_2$ (l)	−1241.2
Free-energy change in reaction, kJ/mol						
Th reduced with						
4 mol Na		+510.6		+243.4		−357.9
2 mol Mg		+64.0		+73.6		−70.1
2 mol Ca		−17.3		−255.5		−379.4

ously: the reduction of $ThCl_4$ with sodium amalgam (Metallex process, Oak Ridge, 1954 [26]) and reduction with magnesium in the Kroll process, which was favored because of the price of magnesium and its lower sensitivity to oxygen (1956).

Metallex Process. The flow diagram of the Metallex process, which requires relatively simple equipment, is shown in Figure 16. Originally, $ThCl_4$ was dissolved in propylenediamine and reacted with liquid sodium amalgam. However, this step was then replaced for cost reasons by

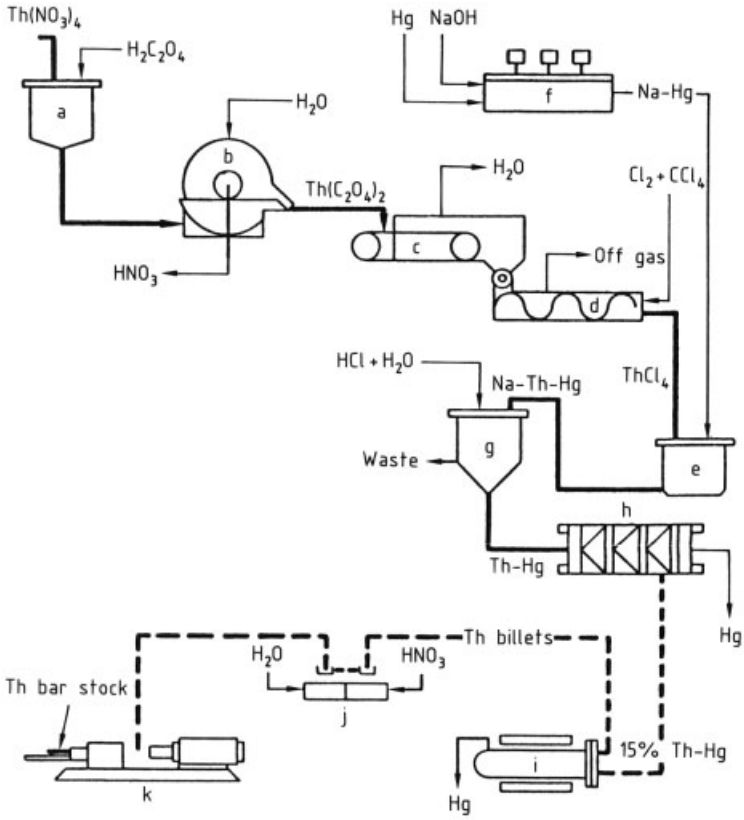

Figure 16. Schematic of Metallex process
a) Precipitator; b) Filter; c) Dryer; d) Chlorinator; e) Contractor; f) Electrolyzer; g) Scrubber; h) Hydraulic filter press; i) Retort; j) Pickler and washer; k) Extrusion

direct reaction of solid thorium chloride with the amalgam at 180 – 300 °C in a kneader mixer, a conversion of 97 % being obtained after 30 min. Excess sodium and sodium chloride were removed by washing with dilute HCl, and solid $ThHg_3$ was removed from excess Hg (in which thorium metal is soluble to<0.1 % at room temperature) by means of a filter press. Thorium amalgam was then decomposed in a steel retort at 1100 °C in vacuum, to remove mercury and produce a sintered solid. Since mercury has a high parasitic neutron capture cross section and can also attack the aluminum cladding on thorium rods in the reactor, it must be rigorously excluded. The process is consequently not without risk.

Kroll Process. The Kroll process, in which thorium chloride is reduced with magnesium [27], was developed at about the same time by the National Lead Company of Ohio. Although this process was originally expected to have a wide application for the treatment of all air-sensitive transition metals, it has in fact become widely used only for titanium, zirconium, hafnium, and thorium. In further development of the process by Ames Laboratories, thorium chloride is completely reacted at 850 – 950 °C in a furnace provided with exterior heating, by using twice the stoichiometric quantity of magnesium to give a Th – Mg alloy (Fig. 17). This alloy melts and forms a compact regulus. Magnesium chloride and excess magnesium are then distilled off at 950 °C in a high-vacuum furnace (Fig. 18) and collected in collection vessels.

Reduction of ThF_4 with Calcium. Quantitatively, the most important method for producing thorium of nuclear purity in the United States was reduction of ThF_4 with calcium metal. The process developed by Ames Laboratories and Iowa State University [28] was analogous to the

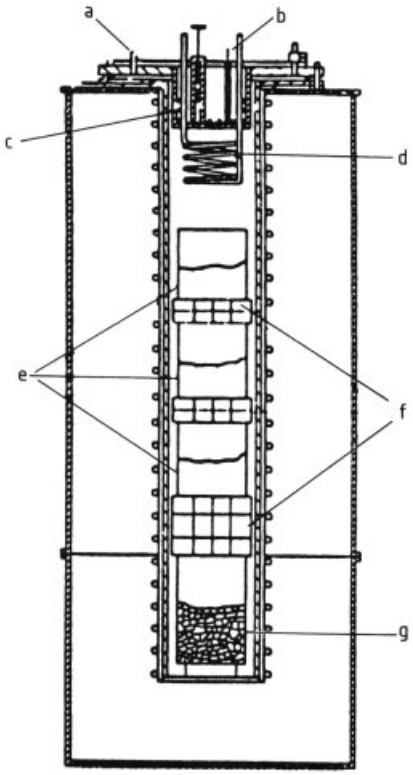

Figure 17. Furnace for $ThCl_4$ reduction (Kroll furnace)
a) Vacuum connection; b) Thermocouple; c) Lift heater; d) Air-cooled condenser; e) $ThCl_4$ container; f) Distance disks; g) Melting crucible containing magnesium

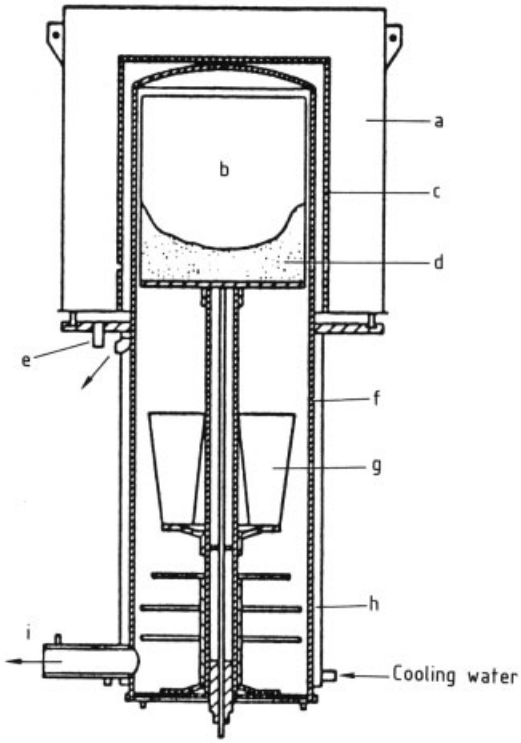

Figure 18. High-vacuum distillation furnace
a) Furnace; b) Distillation crucible; c) Heating element; d) Thorium sponge; e) Furnace vacuum connection; f) Inner wall of reactor; g) Collecting vessel; h) Cooling jacket; i) Reactor vacuum connection

production of uranium metal. However, when calcium metal reacts with ThF_4, the heat of reaction is insufficient to melt the thorium metal. Additional heat is therefore generated by reacting $ZnCl_2$ with calcium metal. This also produces a lower-melting thorium – zinc alloy and a lower melting point eutectic of CaF_2 and $CaCl_2$. The process requires careful preparation of the reactants since most of the metallic impurities in the reactants end up in thorium. Calcium metal must be sieved to remove fine-grained material after preliminary size reduction because this material contains most of the impurities. Zinc chloride must also be of high purity; it must be anhydrous and not too fine grained, otherwise the reaction with calcium can take place prema-turely. The reduction furnace is illustrated in Figure 19. The furnace pot (e), which is made of nonscaling steel, is placed inside the muffle (g). This is heated by three heating elements (d, f, h), outside which is a lining of calcined dolomite. The reaction mixture consists of 75.3 kg of ThF_4, 27.22 kg of granulated calcium metal, and 7.26 kg of anhydrous $ZnCl_2$. A layer of dolomite powder is placed above and below the reactants to protect the walls. The furnace is heated to 660 °C, and when the contents reach 475 °C a strongly exothermic reaction begins, producing a Zn – Th alloy containing 18 mol % Zn. The furnace is operated behind safety walls because deflagrations occasionally occur. After cooling, the slag is removed, and 90 % of the zinc can be distilled off in vacuum (<27 Pa). The thorium sponge is cooled in a protective gas and then inductively melted in a BeO crucible in vacuum. The remaining zinc evaporates, and the thorium is cast in graphite molds (yield 94 – 86 %).

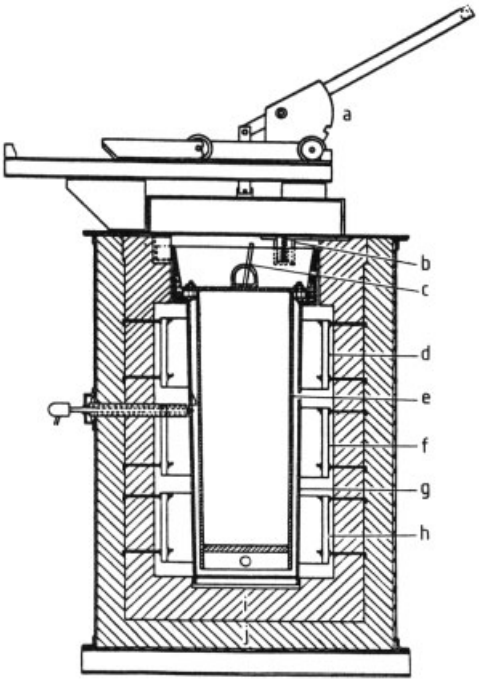

Figure 19. Section view of reduction furnace and furnace pot
a) Cover lifting mechanism; b) Muffle hanger; c) Pot alignment bar; d) Top element; e) Furnace pot (steel); f) Center element; g) Muffle; h) Bottom element; i) 2600 °C insulation brick; j) 1600 °C insulation brick

7. Metalworking [11, pp. 222 ff], [29, 30]

Most production processes lead to thorium powder or sponge. This must therefore be compacted by sintering or melting before it can be worked in a metal-forming process. In the literature, densification by powder metallurgy is described in less detail than melting and casting.

Melting and Casting. To cast thorium metal, it is first melted at 1800 °C by inductive heating in vacuum to avoid the danger of picking up O_2 and N_2. Also, zinc left behind from the production process (reduction of ThF_4 with Ca, see section page 25pp. 22 –23) must be evaporated, and pickup of impurities from the crucible material must be as low as possible. For this, either zinc is removed first in graphite crucibles at a temperature at which thorium hardly picks up any carbon from graphite (1300 °C) and the actual melting is carried out later in BeO crucibles, or both processes are carried out together in a stabilized ZrO_2 crucible. Thorium metal has a very low heat of fusion and, when fused, attacks all known crucible materials, so that excessive heating must be avoided. Both the service life of the BeO crucible and the limitation of pickup of impurities by thorium are only moderately satisfactory.

A two-stage electrode-consuming arc melting process is a better alternative for producing the cylindrical ingots generally used in nuclear technology. The molten thorium metal is subsequently solidified in water-cooled copper

Table 14. Average analyses of thorium ingots

Type of analysis	Impurity	Induction-melted;	Arc melted	
		charge: 40 % scrap, 60 % virgin metal	Charge: virgin metal	Charge: virgin and scrap metal
Chemical	ThO_2, %	2.8	1.34	2.42
	ZrO_2, %	0.14		
	N, mg/kg	290	142	211
	C, mg/kg	465	211	193
Spectrochemical	Al, mg/kg	370	20	13
	B, mg/kg	3.9	0.4	0.5
	Fe, mg/kg	600	243	441
	Ni, mg/kg	520	324	486
	SiO_2, mg/kg	500	34	30
	Zn, mg/kg	10	128	43
	U, mg/kg	20	10	6
Hardness, BHN ∗, 500-kg load		64	54	56
Density, g/mL		11.52	11.57	11.64

∗ BHN = Brinell hardness number.

Table 15. Effect of carbon content on mechanical properties of iodide-source thorium (cold rolled to 85 % reduction and annealed 30 min at 650 °C)

Carbon content, wt %	Yield strength (0.2 % offset), MPa	Tensile strength, MPa	Elongation, %	Reduction of area, %	Hardness, VHN ∗
0.015	52.4	124.1	27.0	60	40
0.022	44.8	121.3	50.0	62	34
0.047	113.1	179.3		43	60
0.070	185.5	237.2	46.0	51	83
0.130	242.0	302.7	41.0	46	109
0.187	346.8	369.6	38.0	39	125
0.22	346.8	418.5	30.0	39	140

∗ VHN = Vickers hardness number.

crucibles. The pieces of thorium sponge obtained from the first dezincification are sawed into strips in chambers filled with inert gas and then welded together to form electrodes. However, metal is lost in this laborious operation, and attempts have therefore been made to replace it by hot plastic forming of the sponge (casting or compaction). The properties of the ingots obtained by induction and arc melting are listed in Table 14.

In subsequent metalworking, thorium resembles soft iron in its ductile behavior. Even the presence of several percent oxygen does not cause problems, because oxygen is virtually insoluble in thorium and therefore only contained in inclusions of ThO_2. To prevent surface oxidation – especially of the small pieces used during hot working by, e.g., rolling, hammering, or extruding – the pieces are encased in soft copper or iron cladding, which is later dissolved in acid. Welding of thorium is possible only in an inert-gas atmosphere. Thorium can be soft-annealed after cold working (e.g., rolling) by heating to 500 – 700 °C. No problems occur due to phase transition since these occur only above 1400 °C. Impurities picked up during the various production processes affect workability. Small amounts of oxygen and beryllium have little effect, but carbon causes hardening. As with steel, considerable precipitation hardening occurs on quenching from the soft-annealed state at 600 – 800 °C (see Table 15). Elements that form carbides (e.g., titanium, niobium, and tantalum) considerably counteract this effect.

The dimensionally accurate tubes – required for fuel elements used in breeder reactors – were usually extruded at 600 – 700 °C, using a protective coating of copper or soft iron. Both of these metals form reaction zones with thorium that must subsequently be machined off. To prevent this, oxidic coatings on thorium were investigated. Alternatively thorium can also be pre-

heated in a molten salt bath consisting of nitrites and cyanides of alkali metals. The adherent salt film acts as a protective and lubricating film. The method eventually favored was to coextrude thorium with an inner and outer cladding of aluminum. Zirconium of similar plasticity obtained by preheating zirconium ingots was used as cladding at a later stage. Process details are unknown for reasons of military secrecy.

8. Intermetallic Compounds and Alloys

8.1. Thorium as the Main Component

Thorium has limited solubility in carbon and nitrogen, and can form compounds of metallic character with both elements. These are described below and in Chapter 5.

Alloys and compounds in which the alloying element vaporizes below the melting point of thorium cannot normally be produced. When the alloying elements have especially high melting points, impurities can be picked up from the crucible or electrode materials (e.g., W, Be, Zr, Cu, or C).

This discussion is limited to alloys and compounds that produce improvements in the properties of thorium, especially its corrosion resistance and strength and hardness properties.

8.1.1. Thorium and Carbon

Carbon is usually present as an impurity in thorium. Thorium produced by decomposition of thorium iodide contains 0.01 – 0.05 % C, and from halide reduction 0.03 – 0.12 % C. As mentioned in Chapter 7, carbon is one of the few elements that increases hardness and age hardening properties of thorium. Of all the elements that improve steel properties, only chromium can by itself (and even more in combination with carbon) produce comparable effects in thorium. Only indium seems to be more effective but is excluded from nuclear applications due to its neutron absorption. The other metals investigated (Ti, V, Mo) have a minor hardening effect, probably due to the formation of carbides with the carbon in thorium. However, the effect is partially negated by the hardness reduction caused by removal of carbon. The Th – C constitutional diagram (Fig. 20) shows the two compounds ThC and ThC_2 (see Chap. 5). The effect of temperature on the solubility of carbon in thorium is given below:

20 °C	0.35 wt %
800 °C	0.43 wt %
1018 °C	0.57 wt %
1215 °C	0.91 wt %

The hardening effect of carbon in the presence of chromium is shown in Table 16. (Chromium and thorium form a simple eutectic system without intermetallic compounds.) The hot forming property (>700 °C) of thorium is not negatively affected at a carbon content ≤0.4 wt %.

8.1.2. Thorium and Nitrogen

The compounds ThN and Th_2N_3 are described in Chapter 5. The solubility of nitrogen in thorium increases linearly with temperature, being 0.05 % at 850 °C and 0.35 % at 1500 °C.

Table 16. Mechanical properties of thorium – chromium – carbon alloys (arc melted, cold rolled about 85 % and annealed 30 min at 650 °C)

Analyzed composition, wt %		Proportional limit,	Yield strength (0.2 % offset),	Tensile strength,	Elongation (in 2 in.),	Reduction of area,	Hardness
Cr	C	MPa	MPa	MPa	%	%	
0.22	0.041	159.3	194.4	232.4	43.0	55.0	106 R_H *
0.28	0.112	182.7	222.7	272.3	42.0	35.0	55 R_B **
0.78	0.104	195.8	248.9	317.8	35.5	56.0	58 R_B **
1.55	0.101	199.3	268.2	355.1	29.0	37.0	64 R_B **
2.42	0.001	221.3	285.4	386.1	32.0	36.0	68 R_B **
3.60	0.075	223.4	317.2	428.9	20.0	16.5	68 R_B **

* R_H = Rockwell hardness.
** R_B = Rockwell hardness B.

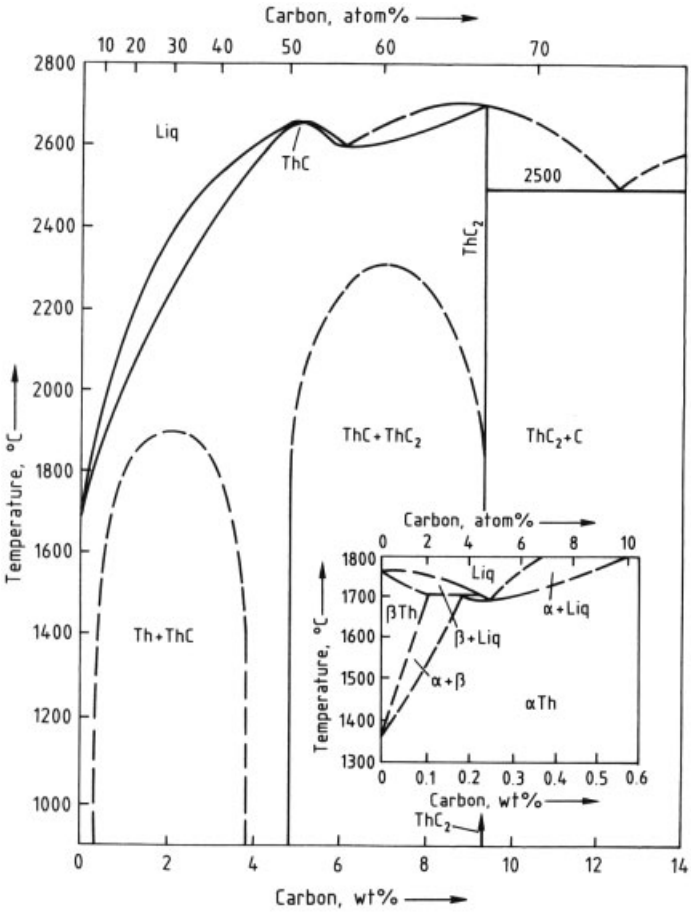

Figure 20. Phase diagram of the thorium – carbon system

Table 17. Mechanical properties of thorium – nitrogen alloys (hot rolled, cold rolled 50 % and annealed 2 h at 850 °C)

Nitrogen content, wt %	Yield strength (0.2 % offset), MPa	Tensile strength, MPa	Elongation (in 1 in.), %	Reduction of area, %	Hardness, VHN
Unalloyed	77.2	135.8	44	60	45
0.17	113.1	205.5	32	54	69
0.18	122.7	207.5	33	54	68
0.41	130.3	233.0	32	48	74

∗ Analyzed 0.012 wt % carbon and 0.08 wt % oxygen.
∗∗ VHN = Vickers hardness number.

In thorium formed by halide reduction, nitrogen is usually present only at levels of 0.02 wt %. The strength of thorium – nitrogen alloys increases moderately with increasing nitrogen content (see Table 17). Precipitation hardening (age hardening) of the alloy after annealing at 850 °C occurs only with quenching temperatures of $<$300 °C.

8.1.3. Thorium and Aluminum

In the thorium – aluminum system, six definite compounds are thought to exist. Of these, only $ThAl_3$ is well described. The solubility of aluminum in thorium is very low (0.6 – 0.7 wt %), so that alloys containing small amounts of aluminum are very brittle. The solubility of thorium

in aluminum is even lower. Alloys containing >10 % aluminum decompose rapidly in moist air, so that, unlike corresponding compounds of uranium and the higher actinides, they will probably not be used in nuclear technology.

8.1.4. Corrosion-Resistant Thorium Alloys

Small additions of alloying elements (to the extent permissible when thorium is used in nuclear technology) can give moderate improvements in resistance to corrosion by water. Improvements are obtained with carbon and beryllium at 100 °C and 200 °C, silicon at 100 °C, and titanium and zirconium at 200 °C. All other alloying element tested thus far have shown either no effect or harmful effects. The carbide formers are the most important of these.

8.1.5. Thorium Alloys Used in Thorium Production and Nuclear Applications

Thorium and Zinc. The thorium – zinc constitutional diagram (Fig. 21) shows four compounds with intermediate eutectics. With 30.5 atom % zinc, an alloy melting at only 1055 °C is formed. This is utilized in the thorium production process (see Chap. 6). However, high temperatures are required to drive off residual zinc.

Thorium and Beryllium. In casting operations with BeO crucibles, thorium picks up 0.02 –0.06 % Be as an impurity. However, beryllium, which has a high vapor pressure, can be driven off from the melt. The brittle compound $ThBe_{13}$ is formed in analogy to beryllium compounds of other actinides, and the eutectic containing 1.75 % Be melts at 1215 °C. Corrosion properties of thorium are improved only with low concentrations of beryllium.

Thorium and zirconium are mutually soluble in all proportions in the β-thorium region (≥1380 °C), and 1.5 wt % zirconium dissolves in thorium at 600 °C (Fig. 22). Zirconium concentrations of >80 wt % give good resistance to corrosion by water (hafnium behaves similarly to zirconium).

Thorium and Uranium. The constitutional diagram for the thorium – uranium system is shown in Figure 23. At the melting point of the eutectic (1270 °C), the solubility of uranium in α-thorium is only 2.5 atom %, while the solubility of thorium in uranium at 900 °C is only 0.3 atom %. Whereas all authors confirm the existence of two eutectics, they are in agreement on the size of the miscibility gap only between 6 and 49 atom % Th. High uranium contents harden thorium, and the addition of small amounts of carbon leads to work hardening. The solubility of uranium in thorium is adequate for the production of ^{233}U in a breeding process, but this is of little significance in the presence of fission products.

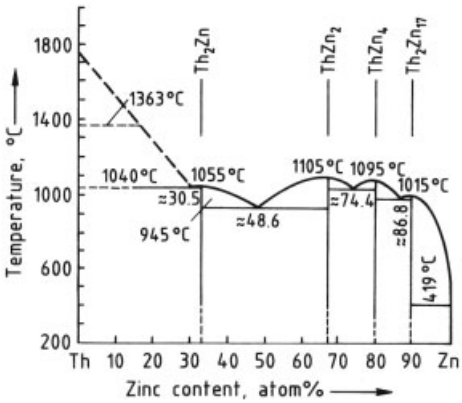

Figure 21. Phase diagram of the thorium – zinc system

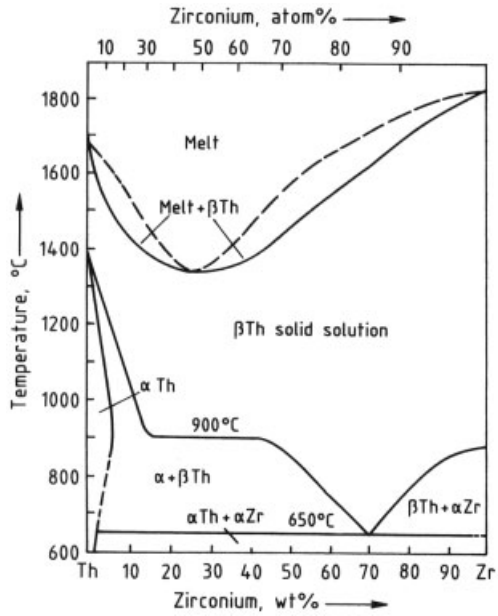

Figure 22. Phase diagram of the thorium – zirconium system

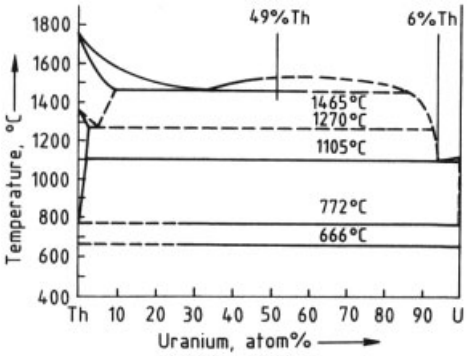

Figure 23. Phase diagram of the thorium – uranium system

Thorium and Plutonium. The thorium – plutonium system has been investigated only at the plutonium-rich end, as shown in Figure 24. The region with high thorium content that is relevant to nuclear physics is so far unknown. Nothing is known about the effect of plutonium on the β-phase of thorium (>1400 °C). The compound Pu_2Th (orthorhombic, unit cell dimensions: $a = 0.982$ nm, $b = 0.8164$ nm, $c = 0.6681$ nm) forms a eutectic with plutonium at 7 atom % thorium, with a melting point of 615 °C.

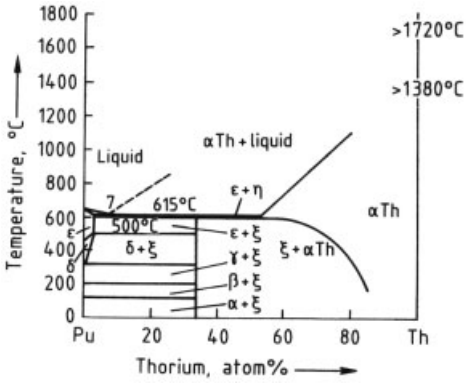

Figure 24. Phase diagram of the thorium – plutonium system

8.2. Alloys Containing Low Concentrations of Thorium

Thorium in low concentrations can be used similarly to cerium mischmetal as a hardness and hot-strength improver in magnesium and Mg – Zn alloys [31] and in aluminum alloys, whose seawater resistance is also improved. The added thorium is precipitated at the grain boundaries in the form of intermetallic compounds. These cause grain growth, although this can be counteracted by small additions of zirconium.

Thorium was widely employed in light magnesium alloys used in automobile crankcases produced by pressure die casting, and provided work-hardening properties and increased hot strength and creep resistance. The principal alloys used were of the type Mg – Zn – Th – Zr, which contained 0.15 – 3 % thorium and up to 3 % zinc. More than 100 t of thorium was used for this application. Although thorium gives the required properties at lower levels than those required with mischmetal, the latter is preferred for cost reasons.

In general, small additions of thorium, like rare earth additions, increase the hot strength and antiscaling properties of many alloys. Such additions are therefore used in jet engines and in resistance heaters. The effect is due partly to deoxidation and partly to the formation of dense, adherent coatings of scale.

Copper and silver become much harder with small additions of thorium. These alloys are used to produce "nonstick" electrical contacts with long lifetimes. Copper – thorium alloys are also used as welding electrodes. An intermetallic compound of thorium with silver (Ag_5Th_3) is pyrophoric. Alloys of nickel with 3 % thorium form a good bond to ceramic bodies and are used for posts in dentistry [32].

Other intermetallic compounds with a promising future include $ThMn_{12}$ as a material for permanent magnets [33]. Compounds of the type $ThCr_2Si_2$ have been tested as superconductors [34].

9. Uses

9.1. Nuclear Technology [35]

Production of ^{233}U. Use of thorium in the nuclear industry is based on its ability to behave as a breeder. It absorbs neutrons and forms ^{233}U by successive release of two β-particles according to

$$^{232}_{90}\mathrm{Th} + {}^{1}_{0}\mathrm{n} \longrightarrow {}^{233}_{90}\mathrm{Th} \xrightarrow{\beta^-,\ 22.3\ \mathrm{min}}$$

$$^{233}_{91}\mathrm{Pa} \xrightarrow{\beta^-,\ 27.0\ \mathrm{d}} {}^{233}_{92}\mathrm{U}$$

The isotope ^{233}U is notable for the fact that the ratio of neutrons liberated on fission to those absorbed is more favorable than the corresponding ratio of ^{235}U and ^{239}Pu. This has aroused interest in this nuclide – which is complicated to produce – in both the military and the civil areas. The ^{233}U formed from thorium can either be separated from the breeder material by chemical methods or undergo fission in situ by absorption of further neutrons, thereby releasing energy. Both methods are subject to a number of limitations.

Chemical Separation of 233 U. In separating ^{233}U, it must be kept in mind that ^{232}U can be formed in an *n*,2*n* reaction with neutrons with an energy of >6.27 MeV from the ^{233}U already formed, and mainly from ^{232}Th via ^{231}Th and ^{231}Pa by further neutron capture. To avoid this reaction, thorium must be allowed to interact only with low-energy thermal neutrons. The same isotope ^{232}U is also formed with a cross section 14 times as large by low-energy neutron capture by ^{230}Th. Although ^{230}Th is present in thorium at a level of only a few parts per million, its concentration can be higher if thorium is a byproduct of uranium production, in which case ^{230}Th is present as an α-decay daughter of ^{234}U.

Whereas ^{232}U is not itself harmful, a serious handling problem arises from the fact that it decays relatively rapidly to its daughter product ^{208}Tl, which emits penetrating γ-radiation with an energy of 2.62 MeV.

In practice, this means that thorium must not be used in a thermal neutron flux in molecular proximity to neutron sources (e.g., as a U – Th alloy, mixed oxide, or mixed carbide) – which would give optimum neutron capture – but should be kept separate from the neutron source ^{235}U (or ^{239}Pu) in breeder rods covered with a layer of moderator.

To give an idea of quantities, if 1 t of thorium is irradiated for some weeks in the carefully thermalized neutron flux and then cooled for 150 d to enable the fission product iodine, which interferes with the separation process, to decay, this forms ca. 1 kg ^{233}U containing ca. 8 mg/kg ^{232}U. This then remains in a handleable state for some weeks without complicated gamma shielding. The overall complexity of this operation can be reduced by replacing the thorium metal rods with ceramic ThO_2 tablets, which are easier to produce but more difficult to dissolve later since $HF + HNO_3$ must be used. Even under these conditions, with a handling concentration limit of 10 – 20 mg/kg ^{232}U, an optimistic price estimate for ^{233}U is still twice that of highly enriched uranium [36]. According to published data, no more than 3 t ^{233}U has been produced worldwide, of which ≤ 800 kg in the United States. New fuel rods have been produced from this in only one full-scale experiment (Kilorod Facility, Oak Ridge, Tennessee [37]).

Attempts to irradiate thorium in the form of a fluid medium or of ThF_4 dissolved in fused salts [38] were finally not successful.

"In Situ" Fission. In an alternative method, ^{233}U is caused to undergo fission with generation of heat at the place where it is formed. Here, fewer possibilities exist for controlling the nuclear reaction, and thorium is by no means completely consumed, but the system is technically easier to realize. Also, a particular property of thorium can be utilized – i.e., its ability to form a series of solid solutions with other actinide oxides (UO_2, PuO_2) – thereby stabilizing them at high temperature. This is mainly a property of ThO_2, to a lesser extent of the carbide, and to a small extent of the metal. Also, since the formation of ^{232}U does not present a problem, the higher neutron flux (by at least an order of magnitude) of the fast breeder reactor is advantageous.

Thorium oxide forms a series of mixed crystals with a cubic fluorite structure with both UO_2 and PuO_2, so that fuel elements can be produced in which thorium is the main component and which are stable when heated in air to >1200 °C. This facilitates both production and safety analysis. The mixed crystals show no phase transformation up to their melting point of >2600 °C. However, used fuel elements of this type can be dissolved only in mixtures of HNO_3 and HF, which attack most available containment materials. Except in India where much thorium but comparatively little uranium is available, all fast breeder programs currently in operation are based on the uranium – plutonium cycle, and no thorium is used.

Thorium could be used in thermal reactors to extend the use of ^{239}Pu, which is produced as a fuel by transformation of ^{238}U. However, continuous hydrogenation of the cladding material (Zircaloy) limits the burnup in water reactors to

ca. 50 000 MW · d/t, so that there is little advantage in using thorium, which is therefore limited to *graphite-moderated thorium high-temperature reactors* (THTRs), which allow ten times the burnup. Here, the fuel elements consist of small particles of uranium and thorium oxide or uranium and thorium carbide, embedded in a graphite matrix. For more details see → Nuclear Technology, Chap. 3.2., → Nuclear Technology, Chap. 5.3.2.; [39].

Owing to parasitic neutron capture in the thorium cycle, Th – U fuel required highly enriched uranium (93 % ^{235}U), and the possible consequent proliferation of weapons-grade material led to concern on the part of source countries (United States). Therefore, new fuel variations are no longer based on thorium. Slightly enriched uranium (10 % ^{235}U), which is not suitable for weapons, is used instead.

So far, high-temperature reactors have not reached the threshold of economic viability achieved by other methods of generating electricity. The safety advantage has not compensated for this disadvantage, and no power stations of this type are currently under construction or being operated.

Because of the wide variations in targeted burnups and the question of whether – and to what extent – thorium is recycled to the reactor, figures for thorium consumption fluctuate not only with reactor capacity but also according to the mode of operation. In using the minimum consumption figures that correspond to maximum recycling, the THTR reactor requires 7.3 t/a thorium for each 1000-MW_e capacity, 340 kg ^{233}U can be recovered from the used fuel elements, and 6.8 t Th (containing isotopes that interfere with a second burnup) must be disposed of as nuclear waste. The initial inventory of the reactor is ca. 60 t of thorium.

9.2. Conventional Applications

9.2.1. Illuminants

The most important application of thorium followed the discovery by Auer Von Welsbach in 1892 that some rare earth oxides and thorium oxide, when introduced in the form of woven yarn into a gas flame, considerably increased the luminous efficacy of the flame. This emission of light is excited by cathode rays, heat, and UV radiation. Although pure cerium oxide converts the UV fraction of an oxidizing gas flame into visible light more efficiently than ThO_2, it is not used in practice because so much light is emitted in the IR region that "cooling" of the gas flame occurs. Excitation of cerium oxide gives a gas flame temperature of only 1500 °C, whereas thorium oxide gives 1930 °C. Maximum luminous efficacy occurs with 1 % cerium oxide in ThO_2 [40].

The requirements of maximum light-emitting surface area together with stability in a gas flame in which high flow rates occur at the zone of maximum temperature led to the development of the gas mantle, whose shape matches that of the combustion zone of the flame. Today, commercially available gas mantles are produced by soaking artificial silk (viscose or cellulose ester fibers) in a solution of thorium and cerium nitrates. The hydroxides are then formed by treatment with alkali and fixed to the fiber; the organic carrier material is burned off, and the remaining ash structure is fixed with collodion lacquer.

Modern gas mantles give a luminous efficacy of 3 lm/W with propane, which is even better than that produced by electric incandescent lamps. Gas mantles are expressly excluded from the restrictions on radioactive substances, although the inhalation of ash from used gas mantles, which can easily occur on careless handling, can lead to harmful deposits in the lungs.

Thorium oxide also had limited application as a light converter (UV into visible light) in fluorescent lamps before modern oxide mixtures were developed.

9.2.2. Electron Emitters

In the production of incandescent filaments, the rate of recrystallization of tungsten is considerably reduced by adding 1 % ThO_2 to the tungsten sintering powder before drawing the filaments.

A side effect of this addition was noticed: i.e., that a small addition of thorium oxide to the tungsten used in heated electrodes considerably reduces the work function of electrons (see Fig. 4) [41]. This is probably due to the formation of a ThO dipole layer, which is constantly renewed from the interior of the elec-

trode. Therefore, since 1922, thoriated tungsten wires have been used in electronic tubes and in the cathodes and anticathodes of X-ray tubes and rectifiers. Owing to the reactivity of thorium with O_2 and N_2, italso acts as a getter for impurities in the evacuated tubes.

These applications decreased very much following the introduction of transistors in the 1950s, but reduction of the electron work function is still a useful phenomenon, having a beneficial effect on the process of welding in an inert gas (argon, helium), for which thoriated tungsten electrodes are used.

9.2.3. Ceramics and Glass

Thorium oxide crucibles represent the most refractory material known and can be used for melting pure, carbon-sensitive metals at 2300 °C. The crucibles are produced by sintering fused and ground thorium oxide – small additions of ZrO_2, P_2O_5, and cryolite being used as sintering aids. Although its application is very limited, ThO_2 is also used to increase the density of ZrO_2 crucibles.

Like ZrO_2, ThO_2 increases the hardness of dispersion-hardened alloys, e.g., in an Ni – 16 Cr – (3 – 5)Al alloy for gas turbine blades.

The addition of ThO_2 to glass leads to extremely high refractive indices with low dispersions and densities lower than those of lead glass. However, this application is also limited because rare-earth oxides have a similar effect and are not α-emitters.

9.2.4. Catalysts [42]

The catalytic effect of ThO_2 seems to be due to an increase in the mobility of hydroxyl ions and hydrogen. Thus, a large number of hydroxylations, hydrogen additions, hydrogen eliminations, oxidations, and reductions have been documented in which thorium catalysts are used. The earliest use of thorium as a reactant and catalyst at the same time was the reaction of aliphatic dicarboxylic acids to give ring ketones (civetone, synthetic musk) by dry distillation of thorium salts [43].

Thorium oxide was used as a fixed-bed catalyst with nickel and cobalt on a kieselguhr carrier, first in ammonia synthesis and then in oil refining and the Fischer – Tropsch process. It was also employed in dehydrogenation, isomerization, and a number of other reactions.

An alloy of platinum with 10 % thorium is a very good catalyst for the oxidation of ammonia to nitrogen oxides. However, a 5 % rhodium alloy with platinum is now preferred because of its superior mechanical properties and longer lifetime in the form of a gauze.

In almost all cases, ThO_2 has been replaced by cheaper catalysts (e.g., based on rare earths) and is now hardly used.

9.2.5. Medicine

Thorium dioxide was used as a contrast medium (Thorotrast) when making X-ray photographs of hollow organs but has been replaced by nonirradiating substances.

^{225}Ra has been used experimentally for therapeutic irradiation, as this is the only isotope of radium that does not have a gaseous decay product. It is obtained either as a short-lived daughter nuclide of ^{229}Th, which is the most stable daughter nuclide (half-life 7340 a) in the ^{233}U decay chain, or can be produced by neutron irradiation of ^{226}Ra.

9.2.6. The Demand for Thorium

In the first phase of utilization from 1893 to 1940, mostly for the production of gas mantles, worldwide consumption was ca. 8000 t ThO_2, of which 3300 t came from India, 4000 t from Brazil, 370 t from the United States, 100 t from Indonesia, 60 t from Sri Lanka, and the rest from Norway, Australia, and a few other countries, together with a small amount of recycled material.

India, which produced 280 ± 40 t/a between 1940 and 1954, has not published any figures since development of the nuclear application.

Nonnuclear use has in general decreased since 1950 [44].

In Germany in the 1990s, thorium consumption decreased to ca. 1.5 t/a, having been 3 – 5 t/a a few years before. Four companies produce thoriated tungsten electrodes, and one company produces electric discharge lamps. Gas mantles are mainly imported from France. Large quantities of thorium oxide, a coupled product from

rare earth production, are stored in various parts of the former German Democratic Republic. This material will probably not be utilized but will be deposited as radioactive waste in a waste disposal site.

10. Product Quality and Analytical Methods

Product Quality. Thorium is commonly marketed in the form of solid thorium nitrate. Here, the thorium content and the level of certain impurities are important. Rare earths, magnesium, calcium, and heavy metals that can be precipitated by sulfide (mainly Fe) are determined.

Thorium nitrate used in the production of gas mantles is highly purified by multiple recrystallization or oxalate precipitation. Rare earths (except in concentrations $<1\,\%$) are undesirable because they stimulate the emission of IR radiation, thereby reducing the emission of visible light. Very small proportions ($>0.02\,\%$) of SiO_2 lead to embrittlement of the gas mantle [45].

Sulfate in ThO_2 has a detrimental effect on crucible production and is removed by multiple dissolutions and precipitations of thorium as oxalate.

The rare earths are also undesirable in nuclear technology because some of them (e.g., erbium and gadolinium) are strong neutron absorbers. Therefore, these elements must be removed. If thorium is used in metallic form, the oxygen and carbon contents are important, the latter from the point of view of hardness properties. When processing irradiated thorium, uranium and thorium must be separated.

The determination of low concentrations of thorium is important in the quality control of alloys. Very low concentrations must be determined in geochemistry and environmental analysis, and when removing ^{230}Th from uranium compounds.

Analytical methods are based on the selective separation of thorium. The natural accompanying elements and compounds in thorium include La and lanthanides, Y, Sc,Ti, Zr, Hf, U, Fe, Al, Ca, SiO_2, and P_2O_5. Only a few reactions exist that are specific for thorium; moreover, it has a stable valence of four so redox separations are impossible. The most difficult precipitation separation is that of scandium, although large amounts of this element are seldom associated with thorium.

Precipitation processes usually start with solutions in relatively strong mineral acids. Precipitation of thorium as the oxalate, fluoride, peroxide, iodate, and hypophosphate gives the best selectivity, the last of these even against rare earths. Some organic dicarboxylic acids, especially tetrachlorophthalic acid, are more selective than oxalic acid. Dimethyl oxalate, which slowly hydrolyzes in warm solutions of mineral acids, can also be used to give slow formation of a precipitate that is free of inclusions.

Most of the compounds soluble in organic solvents that are used for *liquid–liquid extractions* in nuclear technology have been tested for the selective separation of thorium, mainly from solutions in nitric acid, for analytical purposes. They include 1-(2-thenoyl-3,3,3-trifluoroacetone (TTA), mesityl oxide, dibutoxytetraethylene glycol ("penta ether"), long-chain amines (trilaurylamine), TBP [46], and a series of organic phosphonium compounds, especially bis(dialkylphosphinyl)methane [47]. The latter extractants, in particular, separate uranium and thorium selectively from the rare earths. Uranium is most effectively separated from thorium by means of the appropriate choice of ligands (branched or straight-chain) and of the concentrations of the acid and salt in aqueous stripping. Bis(dialkylphosphinyl)-methanes are more selective extractants than trialkylphosphine oxides. The branched-chain, sterically hindered phosphine oxides extract U(VI) more effectively, whereas the straight-chain compounds extract thorium more effectively.

Good selectivity against the rare earths and selectivity against uranium can also be obtained by extracting thorium with TBP from nitric acid solution if the TBP concentration is 40 % in the organic solution (instead of 10 % as used with uranium). In contrast to uranium, stripping is best carried out by salting out with nitrates in the presence of moderately concentrated HNO_3.

Few selective *analytical reagents* for thorium in solution are known. Their selectivity is generally considered to depend on the bond formed with inner electrons of thorium, as in the case of precipitation with dicarboxylic acids. The

compounds formed are generally more sparingly soluble than the corresponding rare earth compounds.

Reagents include a series of weak organic acids with acid groups on adjacent carbon atoms, and other compounds that can form chelate-like complexes. The pink coloration produced by the reaction of ammoniacal solutions of pyrogallol with Th^{4+} has been known for some time, although this is only moderately stable and selective. The formation of star-shaped groups of red needles with Orange II (*p*-(2-hydroxy-1-naphthyl)azo]benzenesulfonic acid), is used for microchemical detection. The most selective reaction known is the formation of a red coloration with thorin, 4-(2-arsonophenylazo)-3-hydroxy-2,7-naphthalenedisulfonic acid [*3688-92-4*] (in the absence of oxalate and fluoride). This coloration is measured at 54.5 nm in sulfuric acid solution, although a number of other compounds that react in a similar way (dibenzoylmethane, aurintricarboxylic acid, alizarin S) also give color reactions or form color lakes.

When impurities originating from production and processing of thorium metal and its compounds are to be analyzed most of the thorium must be separated first. The impurities can then be determined by the usual methods. This applies also to spectrographic analysis, for which most of the thorium should first be removed because of its complex line spectrum.

11. Toxicology

Thorium, its natural decay products, and its compounds are toxic both as heavy metals and as radioactive substances. However, their heavy-metal toxicity is low because of their low solubility in body fluids and organs, and has been compared with that of aluminum.

12. Industrial Safety

In addition to toxicological risks, the spontaneous flammability of thorium metal constitutes a handling hazard.

Thorium has been extracted and processed on an industrial scale since 1893, mostly as the oxide. However, the incidence of harmful effects associated with the handling of thorium is very low. In the absence of strong complex formers (e.g., citrates), water-soluble thorium compounds are immobilized close to the point of incorporation into the body in the form of insoluble hydroxides. Animal experiments first gave an LD_{50} of 1 – 2 g/kg. The low water solubility also prolongs the excretion process, mainly via the feces, and to a lesser extent the kidneys. Exposure to thorium occurs mainly by inhalation of dust into the tracheobronchial tract and the lungs. No toxic chemical effect has thus far been observed in these areas, but radiation damage can occur in the lungs and from the small amounts of thorium transported to the bones.

Thorium is often compared with uranium, which has been well investigated radiologically. Thorium differs from uranium in the more rapid buildup of its daughter nuclides, which also have chemically different properties (Fig. 1). Important daughter nuclides are ^{228}Ra, ^{228}Ac, ^{228}Th, ^{224}Ra, and ^{220}Rn. Other decay products are of little radiological importance. As shown in Figure 2, freshly separated thorium passes through two activity maxima, one after ca. 5 weeks and a second after ca. 60 a, with an intermediate activity decrease after 3 a. In addition to the relatively low β- and γ-radiation, α-radiation has a significant effect on the living organism due to the high ionization density that it produces. During chemical isolation of thorium from ores or residues that have been stored for a long time, ^{228}Ra and ^{224}Ra become concentrated in the filtrate (to>70 % in oxalate precipitation). These nuclides are precipitated with sulfate to concentrate them in sediment and prevent them from passing into receiving waters. Furthermore, phase transformations and chemical changes of thorium and its compounds cause the release of all the ^{220}Rn formed in the solid up to that point, which can lead to the buildup of significant α-activity in poorly ventilated areas. Table 18 gives the maximum permissible concentrations of thorium and its decay products according to the USAEC Regulations of 1955 (still in curies). The German legal requirements for thorium according to the Strahlenschutzverordnung are listed in Table 19. The recommended annual limits for incorporation (ALI) in the International Commission on Radiological Protection (ICRP) 30 Regulations are based on a metabolic inhalation model. This has been used to test the relationship between the uptake of tho-

Table 18. Maximum permissible concentrations of thorium and the thorium daughter products radium-228 and radium-224

Element	Body limit, μC	Water MPC $*$, μC/cm^3	Air MPC $*$, μC/cm^3
^{232}Th (soluble)	0.01	5×10^{-7}	3×10^{-13}
^{232}Th (insoluble)	0.002 $**$		3×10^{-13}
^{228}Ra (soluble)	0.04	8×10^{-8}	1×10^{-11}
^{228}Ra (insoluble)	0.002 $**$		2×10^{-12}
^{224}Ra (soluble)	0.06	5×10^{-5}	8×10^{-9}
^{224}Ra (insoluble)	0.003 $**$		3×10^{-10}

$*$ MPC = Maximum permissible concentration.
$**$ Lung only.

Table 19. Exposure and annual limits for inhalation and digestion of radon, radium, actinium, and thorium

Atomic no.	Element	Radionuclide	Exposure limit, Bq	Annual limit: Inhalation, Bq	Annual limit: Ingestion, Bq
86	radon	^{220}Rn	5×10^6	3×10^8	
		^{220}Rn [c]		5×10^5	
		^{222}Rn	5×10^5	2×10^8	
		^{222}Rn [c]		4×10^6	
88	radium	^{223}Ra	5×10^3	9×10^3	1×10^5
		^{224}Ra	5×10^4	2×10^4	2×10^5
		^{226}Ra	5×10^3	9×10^3	4×10^4
		^{228}Ra	5×10^3	2×10^4	5×10^4
89	actinium	^{227}Ac	5×10^3	9 [a]	4×10^3
				40 [b]	
				100 [c]	
		^{228}Ac	5×10^4	2×10^5 [a,c]	6×10^7
				9×10^5 [b]	
90	thorium	^{227}Th	5×10^3	4×10^3 [c]	2×10^6
				6×10^3	
		^{228}Th	5×10^3	200	1×10^5
				200 [c]	
		^{230}Th	5×10^3	100	9×10^4
				400 [c]	
		^{232}Th	5×10^4	30	2×10^4
				30 [c]	
		^{234}Th	5×10^5	2×10^6	3×10^6
				3×10^6 [c]	
		Th-nat	5×10^4	50	3×10^4
				100 [c]	

[a] All, except nitrates.
[b] Nitrates.
[c] In equilibrium with decay products.

rium and biological detection in a large group of workers in a monazite processing plant [48]. It showed that the analysis of urine and feces correlated only moderately well with the short-term uptake of thorium. After inhalation of insoluble thorium compounds urine analyses showed a maximum thorium uptake fraction of 1.4×10^{-3}, up to 12 % were found in feces, both at the second day after incorporation. Almost all measurements yielded higher figures than the ICRP recommendation, although there was no detectable damage to health. Cytogenetic studies in parallel with this showed no correlation between the frequency of chromosomal aberrations and the duration of contact. Clearance from the lungs has a biological half-life of>500 d, and from bones ca. 200 a.

The workplace should therefore be kept as dust free as possible.

13. Environmental Protection

Thorium occurs almost ubiquitously in nature in low concentrations. Because of its poor solubility in water, it occurs in extremely low concentrations in the biological cycle, and because its use in industry is not increasing, special environmental protection measures are unnecessary. Only in a few locations where thorium was used or processed in the past are soil remediation operations occasionally required.

14. Legal Aspects

According to German atomic law (AtG), thorium that has not been irradiated is a "miscellaneous radioactive material" but is not a "nuclear fuel" so that according to §3/§9 AtG it must be safely disposed of as a radioactive material. According to the nuclear nonproliferation treaty, thorium is classed as a controlled substance. Quantities of 20 t (source material) must be checked at least once per year, as also prescribed for Europe by Euratom Regulation no. 3227/76. Persons constructing and operating plants for the production, separation, utilization, and storage of thorium must report kilogram or higher quantities of thorium.

Exceptions to this duty to report include gas mantles, to which the limits imposed by the radiation protection regulations do not apply, although the amount of freshly purified thorium in a gas mantle represents the maximum permitted intake of activity for 100 person-years. Aged thorium contained in gas mantles could represent ten times this amount.

15. References

1. G. T. Seaborg: *The Transuranium Elements,* Yale University Press, Washington 1958, p. 113.
2. J. R. Naegele: "Surface Analysis of Actinide Materials," in U. Benedikt (ed.): *Transuranium Elements Today and Tomorrow,* North Holland Publ., Amsterdam 1989, p. 66.
3. W. Seelmann-Eggebert, G. Pfennig, H. Münzel, H. Klewe-Nebenius: *Nuklidkarte,* Kernforschungszentrum, Karlsruhe 1981.
4. S. Tilson: "The Earth's Crust," *Int. Sci. Technol.* **X** (1962) p. 24, Fig. 3.
5. H. Remy: *Lehrbuch der Anorganischen Chemie II,* Akad. Verlagsgesellschaft, Leipzig 1942, p. 544.
6. D. F. Peppard et al., *J. Am. Chem. Soc.* **73** (1951), 2529.
7. G. Kirsch, *Ber. Wien. Akad.* **131, II a** (1922) 551 – 568.
8. G. T. Seaborg, *Chem. Eng. News* **26** (1948) 1902 – 1906.
9. H. J. Vink, *Chem. Weekbl.* **44** (1948) 685 – 688.
10. H. A. Wilhelm: *The Metal Thorium, Am. Soc. for Metals,* Cleveland, Ohio, 1956.
11. J. R. Murray: "The Preparation, Properties and Alloying Behavior of Thorium," Report AERE-M/ TN-12, 1952.
12. R. K. Garg et al.: "Status of Thorium Technology," in: *Nuclear Power and its Fuel Cycle,* vol. 2, International Atomic Energy Agency, Vienna 1977, p. 457.
13. F. L. Cuthbert: *Thorium Production Technology,* Addison-Wesley-Publ. Comp., Reading, Mass., 1958, p. 62 ff.
14. P. Krummholz, F. Gottdenker: "The Extraction of Thorium from Monazite," *Proc. Int. Conf. At. Energy* **8** (1956) 126 – 133.
15. F. Soddy, GB 572 411, 1949; *Chem. Abstr.* **42** (1948) 731.
16. J. Barghusen, Jr., M. Smutz: "The Ames Process," *Ind. Eng. Chem.* **50** (1958) 1754.
17. M. Steinhauser: "Löslichkeit von PuO2 and ThO2 in HNO3," Ph. D. Thesis, Universität München 1987.
18. D. J. Crause, K. B. Brown: "Recovery of Th, U, and SE from Monazite Sulfate Leach Liquors by the Amine Extraction (= Ames) Process," *Rep. Oak Ridge National Lab.,* **no. 2720,** July 16, 1959.
19. R. A. Ewing et al.: "Purification of Thorium Nitrate by Solvent Extraction with TBP," *Fin. Rep. USAEC-BMI* **946** (1952) p. 3 ff.
20. F. Sachs: "Preparation of ThO_2," *Nucl. Sci. Abstr.* **6** (1952) 647 ff.
21. L. Gordon, C. H. Vanselow: "Precipitation of Th from Homogeneous Solution," *Anal. Chem.* **21** (1949) 1323 – 1325.
22. IAEA: "Thorium, Physicochemical Properties of its Compounds and Alloys," *At. Energy Rev. Spec. Iss.* Vienna (1975) Nov. 5.
23. V. D. Allred, S. R. Buxton, J. P. McBride, *J. Phys. Chem.* **61** (1957) 117.
24. N. D. Veigel et al.: "The Preparation of High Purity Thorium by the Iodide Process," *USAEC-Rep.* AECD-3586, BMI 1953.

25. R. J. Meyer, A. Gumperts: "Reinheitskriterien von Thoriumnitrat," *Ber. Dtsch. Chem. Ges.* **38** (1905) 817 – 825.
26. R. E. Blanco: "Status of the Metallex Process," *USAEC-Rep.* CF-55-1-53, ORNL 1955.
27. T. C. Runion et al.: "Production of Thorium by Kroll Technique," *USAEC-Rep.* NLCO-616, National Lead Comp. of Ohio, 1956.
28. F. H. Spedding: "Progress-Report in Metallurgy," *USAEC-Rep.* ISC-6, Iowa State College, 1947.
29. H. A. Wilhelm: *The Metal Thorium,* Am. Soc. for Metals, Cleveland/Ohio, 1956.
30. H. A. Wilhelm, B. A. Rogers: "The Physical Metallurgy of Thorium," *AIME-IMD Spec. Rep. Ser.* 1955, Oct. 17, 40.
31. G. S. Foerster et al.: "Investigation of Alloys of Mg and their Properties," Report WADC-FR-56-88, Dow Chemical Corp., US-Air-Force, 1956.
32. *Winnacker-Küchler,* 2nd ed., **5,** 544.
33. A. P. Goncalves et al. "Preparation and Characterisation of Actinide Metals and Compounds, " in *Annual Rep. 1992,* Institute for Transuranium Elements, Karlsruhe 1992, p. 131.
34. M. Boge et al.: "Magnetic and Electrical Properties of $NpRu_2Si_2$," *J. Nucl. Mater.* **166** (1959) 77 – 82.
35. M. Benedict, T. Pigford, H. W. Levi: *Nuclear Chemical Engineering,* McGraw-Hill, New York 1981, pp. 283 –317.
36. E. D. Arnold: "Radiation Limitations on Recycle of Power Reactor Fuels," *Proc. 2nd Int. ThoriumFuel Cycle Symp.* , Gatlinburg/Tenn., May 3 – 6, 1966.
37. R. E. Brooksbank, J. P. Nichols, A. L. Lotts: "The Impact of Kilorod Facility Operational Experience on the Design of Fabrication Plants for ^{233}U-Th-Fuels," *Proc. 2nd Int. Thorium Fuel Cycle Symp.* , Gatlinburg/Tenn., Feb. 1968, pp. 321 – 339.
38. W. R. Grimes, D. R. Cuneo: "Molten Salt Reactor Fuels," chap. 17, in C. R. Tipton Jr. (ed.): *Reactor Handbook,* 2nd ed., Interscience Publ., New York 1960.
39. E. R. Merz, G. Kaiser, E. Zimmer: "Progress in Th-^{233}U-Recycle Technology," *Am. Nucl. Soc. Topical Meeting,* Gatlinburg/Tenn., May 1974.
40. F. Möglich, *Angew. Chem.* **53** (1940) 405 – 409.
41. I. Langmuir, *J. Am. Chem. Soc.* **38** (1916) 2221 – 2295.
42. R. Bastasz, C. Colmenares: *Catalytic Uses of Thorium and Uranium,* UCID 17499, Lawrence Livermore Laboratories, Livermore, June 1977.
43. L. Rusicka in P. Karrer (ed.): *Lehrbuch der Organi-schen Chemie,* Thieme Verlag, Stuttgart 1948, p. 789.
44. J. E. Crawford: "Thorium, Minerals, Facts and Problems," *Bull. U.S.-Bur. Mines* **556** (1956) p. 1 ff.
45. H. V. Hodgson, *J.-Soc. Chem. Ind. Trans.* **41** (1922) 284.
46. K. Pan, Y. Mao Chen, S. Ti Lee: "A Study of Extraction of Thoriumnitrate with TBP," *J. Chin. Chem. Soc. Taiwan* **7** (1960) p. 36 ff.
47. J. R. Parker, C. V. Banks: "Some Bis(Dialkyl-Phosphinyl)Methanes as Solvent Extractants," Contract W-7405-eng-82, Ames-Lab., Iowa, June 1964, OTS.
48. J. L. Lipstein et al.: "Bioassay Monitoring Studies for Thorium," *Radiat. Prot. Dosim. Ser.* **26** (1989) nos. 1 – 4, 57 – 60.

Thyrotherapeutic Agents

Simone M. Dahlmanns, Forschungszentrum Jülich GmbH, Jülich, Federal Republic of Germany

Hans-Wilhelm Müller-Gärtner, Forschungszentrum Jülich GmbH, Jülich, Federal Republic of Germany

1. Introduction 1
2. Thyroid Hormones as Therapeutic Agents 2
2.1. Pharmacology of the Thyroid Hormones 2
2.2. Chemical and Physical Properties of Thyroid Hormones 3
3. Thyroid Depressants as Therapeutic Agents 4
3.1. Iodination Inhibitors 4
3.2. Iodization Inhibitors 5
3.3. Inhibitors of Hormone Incretion . . . 6
3.4. Inhibitors of the Conversion of T_4 to T_3 6
3.5. Radioiodine 6
3.6. Chemical and Physical Properties of Thyroid Depressants 6
4. References 9

Abbreviations:

BTU	benzylthiouracil
CBZ	carbimazole
DIT	diiodotyrosine
INN	International nonproprietary name
MIT	monoiodotyrosine
MMI	methimazole, thiamazole
MTU	methylthiouracil
PTU	propylthiouracil
rT_3	reverse L-triiodothyronine
T_3	L-3,5,3′-triiodothyronine, liothyronine
T_4	L-thyroxine, L-3,5,3′,5′-tetraiodothyronine
TBA	thyroxine binding albumin
TBG	thyroxine binding globulin
TBP	thyroxine binding protein
TBPA	thyroxine binding prealbumin
Tg	thyroglobulin
TRH	thyrotropin releasing hormone
TRIAC	triiodothyroacetic acid
TSH	thryroid-stimulating hormone
TSI	thyroid-stimulating immunoglobulins

1. Introduction

Thyrotherapeutic agents are drugs that affect the function and growth of the thyroid gland. The interest in this class of substances is due to the commonness of diseases of the thyroid. For example, 15 % of the German population suffer from enlargement of the thyroid gland (goiter) [1].

The thyroid is an endocrine gland which provides the organism with the metabolic hormones L-thyroxine (T_4) and L-triiodothyronine (T_3). In terms of weight, the largest component of the thyroid hormones T_4 and T_3 is the element iodine. Iodine is taken in as iodide in food, resorbed, and actively transported into the thyroid gland, where it is used for the synthesis of the thyroid hormones. The adult human organism requires an iodine supply of 150 – 300 µg/d for sufficient synthesis of thyroid hormones [1].

The synthesis of thyroid hormones involves several steps: iodide entry into the thyroid gland (iodination) and oxidation to I_2, incorporation of iodine into tyrosine with the formation of MIT and DIT (iodization), and the coupling of iodotyrosines to form iodothyronines. Iodization and coupling occur in thyroglobulin (Tg). This is followed by the storage phase in Tg and the incretion phase (secretion of the thyroid hormones into the bloodstream). Depending on the requirements of the organism, the thyroid gland releases its hormones into the bloodstream, regulated by hypothalamic TRH and hypophysial TSH.

About ten times more T_4 than T_3 is produced and released into the bloodstream. Outside the thyroid gland, T_4 is mostly deiodinated in the peripheral tissue to T_3, which is 100 times more active, and partly to rT_3, which is inactive. Serum levels of the thyroid hormones that are too high result in suppression of TSH secretion in the hypophysis, while serum T_4 and T_3 levels that are too low lead to stimulation of TSH secretion.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_039

This feedback mechanism guarantees the maintenance of a constant level of thyroid hormone and, consequently, normal metabolism.

Diseases of the thyroid are classified according to functional and morphological criteria. From a functional standpoint, a distinction is made between hypofunction (hypothyreosis) and hyperfunction (hyperthyreosis). Morphological classification includes diffuse struma, multinodular struma, thyroiditis, Morbus Basedow, thyroid cysts, adenoma, and malignant tumors.

Hyperthyreosis is characterized by an increased peripheral blood level of thyroid hormones. The resulting complaints include insomnia, weight loss, diarrhoea, tachycardia and, in severe cases, central nervous disorders.

Diseases of the thyroid that can be connected to hyperfunction are: Morbus Basedow, functional autonomy, thyroiditis, and secondary hyperthyreosis in the case of pathologically increased hypophysial TSH incretion.

Hypothyreosis is characterized by decreased hormone levels in the blood. The resulting complaints include increase in weight, obstipation, bradycardia, central nervous disorders, depression, sensitivity to cold, myxedema, and cretinism. Hypothyreosis can be congenital due to organ dysgenesis and disturbances of hormone synthesis and can appear in the course of a thyropathy or as a result of medication, radiation therapy, or surgical treatment. The most frequent cause of acquired hypothyreosis is autoimmune atrophy.

Thyroid hyper- and hypofunction can be rectified by the administration of the appropriate preparations.

2. Thyroid Hormones as Therapeutic Agents

2.1. Pharmacology of the Thyroid Hormones (→ Hormones, Chap. 2.1.1.).

Thyroid hormone preparations find therapeutic application for:

1) Substitution in the case of hypothyreosis of any genesis
2) Preventive treatment against recurrence of struma after strumectomy or radioiodine treatment
3) Suppressive treatment of euthyroid struma
4) Substitution therapy after thyroidectomy and radioiodine treatment for thyroid malignancies
5) Treatment accompanying thyroid depressant therapy for hyperthyreosis

Four types of preparations are available for medicamentous therapy [2]:

1) Levothyroxine preparations (T_4)
2) Liothyronine preparations (T_3)
3) Preparations that combine T_3 and T_4
4) Preparations that combine T_4 and iodide

T_4 and T_3 differ fundamentally in their pharmacokinetic properties, as shown in Table 1. Today, monotherapy with T_4 is preferred to treatment with T_3 or combined treatment with T_4 and T_3. In this way a demand-orientated supply

Table 1. Pharmacokinetic data of the hormones T_4 and T_3 (active hormone)

	Levothyroxine T_4	Liothyronine T_3
Incretion rate [25]	94 – 110 µg/d (120 – 140 nmol/d)	16 – 22 µg/d (24.6 – 33.8 nmol/d)
Utilization rate [25]	80 µg/d (102 nmol/d)	60 µg/d (92 nmol/d)
Blood level [26]	60 – 80 µg/L (75 – 105 nmol/L)	1.0 – 1.5 ng/L (1.5 – 2.3 nmol/L)
Unbound physiologically active form [26]	0.03 – 0.05 %	0.3 – 0.5 %
Plasma protein binding [25]	99.9 % [a]	99.7 % [b]
Dosage	50 – 300 µg	10 – 100 µg
Relative activity [c]	1	4
Enteral resorption [d]	75 – 85 %	78 – 95 %
Bioavailability	75 – 85 %	80 – 95 %
Distribution volume [25]	12 L	46 L
Total metabolic clearance rate [25]	1.17 L/d	26.1 L/d
Plasma half-life	6 – 7 d	1 d
Onset of action	3 – 5 d (oral), 6 h (i.v.)	12 – 48 h
Maximal effect after	ca. 9 d (single dose)	ca. 2 – 3 h (single dose)
Duration of effect	7 – 10 d	3 – 5 d
Plasma transfer	slow	slow
Excretion in breast milk	in small amounts	in small amounts
Biliary excretion	yes	yes
Enterohepatic circulation	yes	yes
Renal unchanged	0 %	0 %

[a] Up to 80 % on TBG, up to 15 % on TBPA, up to 10 % on TBA.
[b] Up to 99.7 % on TBG, up to 0.3 % on TBPA and TBA; binding is ten times weaker than T_4.
[c] For enteral administration.
[d] In fasting state; ingestion of food lowers resorption to 35 %.

of the metabolically active T_3 is guaranteed to a certain extent. The biological half-life of T_4 of 6 – 7 d is much longer than that of T_3 (ca. 19 h). T_4 is converted extrathyroidally to metabolically active T_3. Thus, even a single dose of T_4 results in essentially constant hormone levels in the serum. Monotherapy with T_3 can easily result in unphysiological T_3 peak values in the serum and overdosage symptoms. Therefore, treatment with T_3 offers no advantages over T_4 therapy. An exception is the short-term application of T_3 if thyroid hormone therapy must be interrupted prior to diagnostic and therapeutic measures.

Combined treatment with both thyroid hormones attempts to imitate physiological conditions. The balanced combination of both thyroid hormones (T_4 : T_3 = 5:1 or 10:1) should guarantee physiologically adequate substitution and suppression.

Combination preparations are indicated when conversion to T_3 does not occur in the case of monotherapy with T_4. These preparations are used to treat struma caused by lack of conversion. Preparations that combine levothyroxine and iodide are given preference in the treatment of iodine-deficiency struma.

An adequate supply of iodide is required for protection against iodine-deficiency struma. Since in many areas of Germany, the alimentary supply of iodine is less than the iodine requirement and endemic iodine deficiency affects both the function and morphology of the thyroid gland [2] (formation of struma, degeneration, autonomy), adequate prevention of iodine deficiency by administration of iodide preparations is recommended (100 – 200 µg I^-/d).

As a rule, substitution therapy with thyroid hormones starts very slowly. The dosage is established individually, depending on the age, condition, and indication, the regular determination of the hormone values in the serum being imperative. If the dosage is correct, no undesirable effects occur. If the individual tolerance limit or the proper dose is exceeded, the clinical symptoms typical of hyperthyreosis can occur, especially if the dose is increased too fast at the start of treatment.

During pregnancy, substitution treatment for hypothyreosis of the mother should be rigorously carried out because otherwise the insufficient hormone production in the mother can result in defective development of the child. However, overdosage must be carefully avoided. Even in the case of high-dose therapy, the amount of thyroid hormones secreted into the breast milk during lactation is not sufficient for the development of hyperthyreosis or suppression of TSH secretion in the infant.-,10

2.2. Chemical and Physical Properties of Thyroid Hormones

Levothyroxine (INN), T_4, L-thyroxine, L-3,5,3′,5′-tetraiodothyronine, L-3-[4-(4-hydroxy-3,5-diiodophenoxy)-3,5-diiodophenyl]-alanine (IUPAC) [*51-48-9*], $C_{15}H_{11}I_4NO_4$, M_r 776.93, decomp. 235 – 236 °C, $[\alpha]_D^{20}$ +4.4° (3 % solution in 0.13 N NaOH in 70 % ethanol). Colorless, light-sensitive crystals, insoluble in water, ether, ethanol, and chloroform.

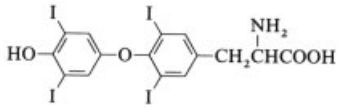

Levothyroxine sodium (INN), L-3-[4-(4-hydroxy-3,5-diiodophenoxy)-3,5-diiodophenyl]alanine sodium (IUPAC) [*55-03-8*], $C_{15}H_{10}I_4NNaO_4$, M_r 798.85.

Levothyroxine sodium *x*-water (INN) [*25416-65-3*], $C_{15}H_{10}I_4NNaO_4 \cdot x\,H_2O$, pH 8.9 (saturated, aqueous solution). White to lightly colored, light-sensitive, crystalline powder. Solubility: very sligthly soluble in water (solubility decreases with pH), slightly soluble in ethanol, insoluble in acetone, chloroform, ether, and soluble in alkali metal hydroxide solution. The analytical profile and synthesis of levothyroxine compounds are described in [3].

Liothyronine (INN), T_3, 3,5,3′-triiodothyronine, L-3-[4-(4-hydroxy-3-iodophenoxy)-3,5-diiodophenyl]alanine (IUPAC), [*6893-02-3*], $C_{15}H_{12}I_3NO_4$, M_r 651.01, *mp* 233 – 234 °C, $[\alpha]_D^{29,5}$ +21° (4.75 % solution in 1:2 1 M HCl/ethanol, pK_a 8.5, insoluble in water and ethanol, soluble in dilute alkali.

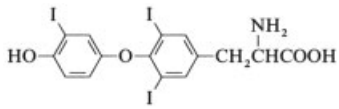

Liothyronine sodium (INN) [*55-06-1*], $C_{15}H_{11}I_3NNaO_4$, M_r 672.99, UV_{max} 319 nm. White to slightly colored powder, insoluble in water and ether, slightly soluble in ethanol, soluble in dilute alkali metal hydroxide solution. For synthesis, see → Hormones, Chap. 2.1.3.

Tiratricol (INN), TRIAC, TA3, 3,5,3′-triiodothyroacetic acid, 4-(4-hydroxy-3-iodophenoxy)-3,5-diiodophenylacetic acid (IUPAC) [*51-24-1*], $C_{14}H_9I_3O_4$, M_r 621.95, *mp* 65 °C from methanol and water; after recrystallization at 110 °C: 180 – 183 °C. Synthesis is described in [4].

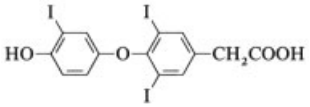

TRIAC [5,6] is a natural metabolite of T_3. In spite of its high affinity for the T_3 receptor, TRIAC evidently has no physiological role. In vivo, it exerts a strong, dose-dependent and reversible inhibitory effect on the hypophysial secretion of TSH. The effects on peripheral tissue are negligibly small. Consequently, TRIAC has been used as an alternative to T_4 for hypophysial TSH suppression in patients suffering from thyroid carcinoma. More recent studies show that TRIAC results in increased L-thyroxine activity, even in peripheral tissue.

l-3,5-Diiodothyronine (INN), T_2, L-3-[4-(4-hydroxyphenoxy)-3,5-diiodophenyl]alanine (IUPAC) [*1041-01-6*], $C_{15}H_{13}I_2NO_4$, M_r 525.10, *mp* 256 – 257 °C (decomp.), $[\alpha]_D^{25}$ +25.2° [5 % solution in 1 M HCl:95 % ethanol (1:2)], colorless crystals. Thyroid hormone with a L-thyroxine-like activity. For synthesis, see [7].

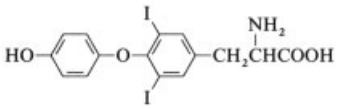

3. Thyroid Depressants as Therapeutic Agents

Thyroid depressants are divided into five groups, based on their mechanism of action.

1) Iodination inhibitors, which inhibit iodide transport into the thyroid gland
2) Iodization inhibitors, which directly inhibit the synthesis of thyroid hormones
3) Inhibitors of hormone incretion
4) Inhibitors of T_4 and T_3 conversion
5) Radioiodine, which can destroy thyroid tissue by means of ionizing radiation

3.1. Iodination Inhibitors

Iodination inhibitors act as thyroid depressants by competitively inhibiting the uptake of iodine by the thyroid cell (iodination). They release accumulated iodide from the thyroid gland provided it is not incorporated into Tg molecules. In this manner, these agents inhibit the synthesis of the thyroid hormones, resulting in iodine depletion.

Iodination inhibitors are administered to block the thyroid gland prior to scintigraphic investigations of other organs with iodine isotopes, during the unavoidable application of iodine-containing contrast medium if an autonomic thyropathy is known to exist, and in the treatment of hyperthyreosis if other thyroid depressants must be discontinued due to unjustifiable side effects. These agents should not be used if treatment with radioiodine or iodine is to be performed. Perchlorate finds therapeutic application in the form of the sodium or potassium salt in a dosage of up to 1500 mg/d. The undesirable effects observed are gastrointestinal disturbances; serious but more seldom side effects are agranulocytosis, aplastic anemia, and nephrotic syndrome.

3.2. Iodization Inhibitors

Iodization inhibitors are derivatives of thiourea. Agents that are therapeutically used are: benzylthiouracil (BTU), methylthiouracil (MTU), propylthiouracil (PTU), carbimazole (CBZ), methimazole (MMI), and thibenzazoline. Today, CBZ and MMI have gained acceptance in the clinical practice [8].

These substances inhibit the iodine oxidation reaction catalyzed by a thyroid peroxidase [9], the incorporation of iodine into tyrosine, and the

coupling of tyrosine molecules to form iodothyronines. Thus, they inhibit the synthesis of thyroid hormones. This makes possible the symptomatic treatment of thyroid hyperfunction irrespective of its etiology.

In addition, thiourea derivatives are attributed with a favorable effect on the immunopathogenesis of hyperthyreosis [10]. At present, it cannot be definitely established to what extent the underlying immunopathogenetic process can be suppressed.

Processes that are not influenced are the transport of iodide into the thyroid gland, the incretion of the thyroid hormones, and hyperthyreosis due to hormone liberation after destruction of the thyroid cells (e.g., after radioiodine therapy or thyroiditis). This mechanism of action explains why the clinical effect starts only after the hormone depots in the gland have been emptied and the peripheral hormone concentrations decrease. An additional mechanism of action being discussed for PTU is the blocking of the peripheral deiodination of T_4 to T_3.

Indications for the administration of thiourea thyroid depressants include:

1) Initial and long-term treatment of Basedow hyperthyreosis
2) Initial treatment of thyroid autonomy
3) Treatment of transitory hyperthyreosis caused by inflammation
4) Intravenous treatment of thyrotoxic crisis (preferably PTU)
5) Intermediate therapy after radioiodine treatment until an euthyroid metabolic condition is reached
6) Preoperative and prior to radioiodine therapy in all hyperthyreosis conditions because treatment must always be given in the euthyroid state
7) Not for the treatment of hyperthyreosis caused by inflammation

Due to their goitrogenic action when given over an extended time period, antithyroid substances can be administered in combination with synthetic thyroid hormones.

In principle, all thyroid depressants give rise to the same undesirable effects. The total rate of side effects is about 14.3 % [11]. The most frequent are skin reactions (5.6 %) and ailments of the joints (1.6 %). The most dangerous side effect, which occurs with a frequency of up to 1 %, is toxic bone marrow damage (agranulocytosis, leukopenia, thrombopenia). Other rare side effects include gastrointestinal complaints, swelling of the joints, fever, headache, and rash.

The frequency of side effects varies greatly from agent to agent. CBZ, MMI, and PTU are favorably assessed. As a result of the lower dosage and toxicity, thioimidazoles are preferred to PTU, and MMI as an active metabolite to CBZ.

The highest rate of side effects is shown by MTU (14 %; agranulocytosis in 1 – 3 % of all cases). Hence, the use of MTU in ready-for-use drugs is no longer recommended [12].

Care must be taken during pregnancy because the drugs pass the placental barrier and can result in the development of struma or cretinism in the child. For this reason, during pregnancy, the lowest possible doses of thyroid depressants must be administered without the addition of thyroid hormones, which pass the placental barrier. In the case of combination therapy with a thyroid depressant and T_4, in spite of the euthyroid metabolic state of the mother, there is a risk of the development of hypothyreosis in the child due to thyroid suppression.

The agent of choice during pregnancy and lactation is PTU in the United States, and MMI in Germany.

A survey of the pharmacokinetic data of antithyroid agents is presented in Table 2 (see next page).

3.3. Inhibitors of Hormone Incretion

Higher concentrations (10 – 500 mg/d) of iodide ions influence all the important steps of iodine metabolism and indirectly inhibit the secretion of thyrotropin. The active uptake of iodide by the thyroid gland is inhibited by high concentrations of iodide in the blood. At the same time, the organic binding of iodine is inhibited (Wolff – Chaikoff effect), resulting in the inhibition of hormone synthesis.

The inhibition of hormone release from Tg is clinically most significant and is the cause of the immediate (after ca. 2 h) action of iodide. This inhibition is due to the blocking of the proteases involved. However, the thyroid-depressing effect decreases with continuous use. For this reason, therapy with iodides (with the

Table 2. Pharmacokinetic data of antithyroid substances

	Methylthiouracil	Propylthiouracil	Carbimazole	Methimazole, Thiamazole
Notes	not commercially available in Germany since July 92		prodrug (inactive) enzymatic conversion to methimazole during or directly after resorption	active metabolite
Initial dose	400 – 600 mg/d	300 – 600 mg/d	30 – 60 mg/d	20 – 40 mg/d
Maintenance dose	100 – 300 mg/d in single doses	50 – 200 mg/d in single doses every 8 h [27]	5 – 15 mg/d in single doses every 8 h [27]	5 – 10 mg/d in single doses every 8 h [27]
Relative activity		10 mg ≈ 1 mg Methimazole [28]	15 mg ≈ 10 mg Thiamazole [29]	
Enteral resorption	almost complete	fast and almost complete	fast and almost complete	fast and almost complete [7,8,30]
Bioavailability (oral)	80 – 90 % [30]	80 – 95 % [30]	80 – 95 % [30]	80 – 95 % [30]
Max. plasma level after	1 – 2 h	1 – 2 h [30]	3 h	3 – 6 h [31]
Plasma level		2 – 3 µg/mL (dose: 200 mg)		
Distribution volume		20 – 30 L [28,30]		40 L [28]
Plasma clearance		120 mL/min [30]		200 mL/min [28]
Plasma half-life	1 – 1.65 h	75 min [28]		4 – 6 h [28]
		1 – 2 h [30]		2 – 6 h [30]
Plasma protein binding		75 – 80 % [28,30]	high	none [8]
Onset of action		after 1 – 2 weeks	after 1 – 3 weeks	after 1 – 3 weeks
Period of effect		2 – 3 h	24 h	24 – 40 h [28]
Placental transfer		low	higher than PTU [28]	higher than PTU [28]
Excretion in breast milk		1/10 of the serum concentration of the mother	higher than PTU [28]	higher than PTU [28]
Distribution		intrathyroidal accumulation	intrathyroidal accumulation	intrathyroidal accumulation
Metabolism				
intrathyroidal		propylthiouracil sulfinic acid	3-methylthiohydantoin [8, 31]	3-methylthiohydantoin [8, 31]
		PTU – SO_2H [8]	methimazole,	
		SO_4^{2-} [7]	SO_4^{2-} [7]	
		6-propyluracil, SO_4^{2-} [7,16]		
		S-methyl-6-propylthiouracil [16]		
extrathyroidal	65 % glucuronide	65 % glucuronide [7,16]		
		S-methyl-PTU		
Excretion				
biliary		yes	yes	yes
enterohepatic circulation		yes [7,8]		yes [7,8]
renal (unchanged)	ca. 30 %	<10 % [30]		<10 % [30]
Rate of side effects	7 – 14 %	5 %	2 – 3 %	5 %
	agranulocytosis in 1 – 3 % of all cases			agranulocytosis in 0.3 – 0.6 % of all cases

exception of iodine prophylaxis) in preparation for thyroidectomy in a thyrotoxic crisis or hyperthyroidal struma (Plummering) is possible only for a duration of 2 – 6 weeks. The agents used are KI, Lugol's solution, and organic iodine preparations.

Polonium iodide, which was administered as an i. v. preparation until now, is no longer produced for economic reasons.

Lithium salts, which have attained considerable importance in the treatment of mental diseases, inhibit the proteolytic release of the thyroid hormones [1]. As a result of their side effects and small therapeutic range, lithium salts are reserved for special indications such as thyrotoxic crisis, especially when caused by the i. v. application of contrast media [13].

In preoperative thyroid depressant therapy for hyperthyreosis, the thyroid depressants mentioned in Section 3.2 and symptomatic treatment with β-blockers are preferred [14].

3.4. Inhibitors of the Conversion of T_4 to T_3

Blockers are used in the clinical symptomatic treatment of hyperthyreosis. The agent of choice, propranolol, inhibits the conversion of T_4 to T_3. In the case of PTU (iodization inhibitor), the blockage of peripheral deiodination of T_4 to T_3 is being discussed as an additional mechanism of action. Glucocorticoids, which are intravenously administered in the treatment of thyrotoxic crisis, are also believed to inhibit the peripheral conversion of T_4 to T_3. The principle of action of the conversion inhibitors is preferentially employed in the treatment of thyrotoxic crisis [1].

3.5. Radioiodine

A radioactive form of iodine, ^{131}I [*7790-26-3*], $T_{1/2}$ 8 d, γ energy 364 keV, is used in the treatment of hyperthyreosis in Morbus Basedow or thyroid autonomy, in the treatment of differentiated thyroid carcinomas after thyroidectomy, in struma therapy, in recurrences after strumectomy, and in metastasizing thyroid carcinomas that store iodine [15, 16]. Like the stable isotope ^{127}I, it enters thyroid metabolism and is incorporated into the thyroid hormones. The radiation comprises 90 % of β-particles and 10 % of γ-rays. As a result of the low range of β radiation (1 mm), thyroid tissue is largely destroyed selectively. The γ radiation can be used for diagnostic purposes. According to recommendations made by the German Federal Board of Health (BGA), itshould be used only in the case of thyroid malignancies (whole body scintigraphy) or to calculate the dosage for radioiodine therapy.

3.6. Chemical and Physical Properties of Thyroid Depressants

Aminothiazole (INN), 2-aminothiazole (IUPAC) [*96-50-4*], $C_3H_4N_2S$, M_r 100.14, *mp* 93 °C from benzene and petroleum ether, *bp* 140 °C (1.47 kPa). Crystalline substance, soluble in hot water, slightly soluble in cold water, ethanol, and ether, readily soluble in dilute hydrochloric acid and in 20 % sulfuric acid.

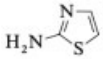

This agent is synthesized according to the King method by the conversion of paraldehyde with thiourea in the presence of sulfuryl chloride [17].

Benzylthiouracil (INN), BTU, 6-benzyl-2,3-dihydro-2-thioxo-4(1*H*)-pyrimidinone (IUPAC) [*33086-27-0*], $C_{11}H_{10}N_2OS$, M_r 218.30, *mp* 223 –224 °C. The synthesis is described in [18].

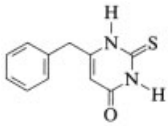

Methylthiouracil (INN), MTU, 2,3-dihydro-6-methyl-2-thioxo-4(1*H*)-pyrimidinone (IUPAC) [*56-04-2*], $C_5H_6N_2OS$, M_r 142.18, *mp* 326 –331 °C with decomposition; sublimes. White, finely crystalline, odorless powder with a slightly bitter taste, light-sensitive.

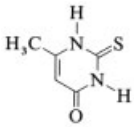

Solubility: 1:2000 – 2500 in water, 1:130 – 150 in boiling water, 1:800 in ethanol, and 1:1300 in 85 % glycerol. Readily soluble in alkali metal hydroxide and ammonia solutions, and methanol, slightly soluble in acetone, chloroform, and ether. Sparingly soluble in dilute mineral acids, and practically insoluble in benzene. The synthesis is conducted by reacting thiourea with ethyl 3-oxobutyrate [18].

Propylthiouracil (INN), PTU, 2,3-dihydro-6-propyl-2-thioxo-4(1*H*)-pyrimidone (IUPAC) [*51-52-5*], $C_7H_{10}N_2OS$, M_r 170.23, *mp* 219 – 221 °C. White, finely crystalline, bitter tasting powder.

Solubility: 1:900 in water at 20 °C, 1:100 in boiling water, 1:60 in 90 % ethanol, and 1:60 in acetone. Very slightly soluble in ether, chloroform, and benzene. In an aqueous medium, PTU is a monobasic acid with a pK_a of 8.3 (20 °C). It dissolves in aqueous ammonia and aqueous alkali metal hydroxide with formation of salts.

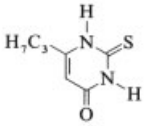

PTU is made by the condensation of thiourea with ethyl 3-oxohexanoate [18]. A summary of the analytical data of PTU is given in [19].

Carbimazole (INN), CBZ, ethyl-3-methyl-2-thioxo-4-imidazoline-1-carboxylate (IUPAC) [*22232-54-8*], $C_7H_{10}N_2O_2S$, M_r 186.23, *mp* 122–125 °C. White or creamy crystalline powder with a characteristic odor.

Solubility: 1:5000 in water at 20 °C, 1:50 in ethanol, 1:330 in ether, 1:3 in chloroform, and 1:17 in acetone.

It is produced by the reaction of methimazole with ethyl chloroformate [20].

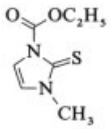

Methimazole, thiamazole (INN), MMI, 1-methyl-2-imidazole thiol (IUPAC) [*60-56-0*], $C_4H_6N_2S$, M_r 114.17, *mp* 146–148 °C from ethanol, *bp* 280 °C with slight decomposition. White, crystalline powder with a weak characteristic odor and a bitter taste, light-sensitive.

Solubility: readily soluble in water (1:5), soluble in ethanol (1:5), and chloroform (1:5), slightly soluble in ether (1:125) and petroleum ether.

It is produced by the reaction of aminoacetaldehyde diethylacetal with methylisothiocyanate or of thiocyanic acid with N-methylaminoacetals [21]. A collection of analytical data is given in [22].

Thibenzazolin (INN), 1-3-bis(hydroxymethyl)-2-benzimidazolinethione (IUPAC) [*6028-35-9*], $C_9H_{10}N_2O_2S$, M_r 210.26, *mp* 160–162 °C. Colorless crystals having a very bitter taste, soluble in dilute alkali metal hydroxide solutions.

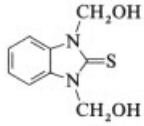

It is synthesized as described in [23], and is a thyroid depressant with a sedative effect. A selection of the common trade names of the substances mentioned above is given in [24].

4. References

1. P. Pfannenstiel: *Schilddrüsenkrankheiten, Diagnose und Therapie,* Berliner Medizinische Verlagsanstalt GmbH, Berlin 1991.
2. F. A. Horster: *Schilddrüsenkrankheiten, Diagnose und Therapie in der Praxis,* Deutscher Ärzte Verlag, Köln 1989.
3. A. Post, R. J. Warren in K. Florey (ed.): *Analytical Profiles of Drug Substances,* **vol. 5,** Academic Press, New York 1976, pp. 225–281.
4. R. I. Meltzer et al., *J. Org. Chem.* **26** (1961) 1418–1428.
5. C. Mechelany et al., *Clin. Endocrinol. (Oxford)* **35** (1991) 123–128.
6. H.-W. Müller-Gärtner, C. Schneider, *Clin. Endocrinol.* **28** (1988) 345–351.
7. J. Elks, G. J. Waller, *J. Chem. Soc.* 1952, 2366–2370.
8. G. Benker, T. Olbricht, *Krankenhausarzt* **55** (1982) 591–601.
9. G. Benker, D. Reinwein, *Klin. Wochenschr.* **60** (1982) 531–539.
10. A. Pinchera et al., *J. Clin. Endocrinol. Metab.* **29** (1969) 231.
11. M. Meyer-Geßner et al., *DMW Dtsch. Med. Wochenschr.* **114** (1989) 166–171.
12. R. Thesen, M. Schulz, R. Braun, *Pharm. Ztg.* **25** (1989) 1590.
13. H. P. Wolff, T. R. Weihrauch: *Internistische Therapie,* 7th ed., Urban & Schwarzenberg, München 1988, pp. 868–890.
14. F. Raue, *Dtsch. Ärzteblatt* **87** (1990) no. 42, 31.
15. H. J. Hermann: *Nuklearmedizin,* 2nd ed., Urban & Schwarzenberg, München 1989.
16. B. Marchant et al., *Pharmacol. Ther.* **Part B 3** (1978) 305.
17. H. Erlenmeyer, L. Herzfeld, B. Prijs, *Helv. Chim. Acta* **38** (1955) 1291–1294.
18. G. W. Anderson et al., *J. Am. Chem. Soc.* **67** (1945) 2197–2200.

19. H. Y. Aboul- Enein in K. Florey (ed.): *Analytical Profiles of Drug Substances,* **vol. 6,** Academic Press, New York 1977, pp. 457 – 486.
20. J. A. Baker, *J. Chem. Soc.* 1958, 2387 – 2390.
21. R. G. Jones et al., *J. Am. Chem. Soc.* **71** (1949) 4000 –4002.
22. H. Y. Aboul-Enein, A. A. Al-Badr in K. Florey (ed.): *Analytical Profiles in Drug Substances,* **vol. 8,** Academic Press, New York 1979, pp. 351 – 370.
23. L. Monti, G. Venturi, *Gazz. Chim. Ital.* **76** (1946) 365.
24. Arzneibüro der Bundesvereinigung deutscher Apothekerverbände (ABDA): *Pharmazeutische Stoffliste,* Werbe- und Vertriebsgesellschaft deutscher Apotheker mbH, Frankfurt 1993.
25. K. Oberdisse, E. Klein, D. Reinwein: *Die Krankheiten der Schilddrüse,* 2nd ed., Thieme Verlag, Stuttgart –New York 1980.
26. W. Forth, D. Henschler, W. Rummel: *Allgemeine und spezielle Pharmakologie und Toxikologie,* Bibliographisches Institut, Mannheim 1988.
27. J. E. F. Reynolds: *Martindale, The Extra Pharmacopoeia,* Pharmaceutical Press, London 1989, pp. 682 –689, 1487 – 1491.
28. D. S. Cooper, *N. Engl. J. Med.* **311** (1984) 1353 – 1362.
29. R. Jansson et al, *J. Clin. Endocrinol. Metab.* **57** (1983) 129 – 132.
30. J. P. Kampmann, J. M. Hansen, *Clin. Pharmacokinet.* **6** (1981) 401 – 428.
31. G. Skellern, *Br. J. Clin. Pharmacol.* **9** (1980) 137 –143.

Tiles (Roof) → **Construction Ceramics**
Tiles (Wall) → **Whitewares**

Tin, Tin Alloys, and Tin Compounds

GÜNTER G. GRAF, Freiberg, Federal Republic of Germany

1. History . . . 2
2. Properties . . . 2
3. Occurrence; Ore Extraction and Beneficiation . . . 5
3.1. Minerals . . . 5
3.2. Deposits . . . 5
3.3. Mining . . . 6
3.4. Ore Beneficiation . . . 7
4. Smelting . . . 8
4.1. Fundamental Theory of Smelting 8
4.2. Special Aspects of the Winning of Tin from its Ores . . . 10
4.3. Production of Crude Tin . . . 11
4.3.1. General Aspects . . . 11
4.3.2. Ore Preparation prior to Reduction 11
4.3.2.1. Pyrometallurgical Enrichment of Low-Grade Concentrates . . . 11
4.3.2.2. Roasting . . . 12
4.3.2.3. Leaching . . . 13
4.3.3. Reduction . . . 14
4.3.3.1. Reduction in a Shaft Kiln . . . 14
4.3.3.2. Reduction in a Reverberatory Furnace . . . 15
4.3.3.3. Reduction in Rotary Kilns . . . 16
4.3.3.4. Reduction in an Electric Furnace . . 17
4.3.3.5. Other Reduction Processes . . . 18
4.3.4. Slag Processing . . . 18
5. Refining . . . 20
5.1. Pyrometallurgical Refining . . . 20
5.1.1. Removal of Iron . . . 20
5.1.2. Removal of Copper . . . 21
5.1.3. Removal of Arsenic . . . 21
5.1.4. Removal of Lead . . . 22
5.1.5. Removal of Bismuth . . . 22
5.2. Electrorefining . . . 22
5.2.1. Electrorefining in Acid Medium . . 22
5.2.2. Electrorefining in an Alkaline Medium . . . 23
5.2.3. Other Methods of Electrorefining . 23
6. Recovery of Tin from Scrap Materials and Residues . . . 23
7. Analysis . . . 24
7.1. Analysis of Ores and Concentrates 24
7.2. Analysis of Metallic Tin . . . 25
8. Economic Aspects . . . 25
9. Tin Alloys and Coatings . . . 26
10. Inorganic Tin Compounds . . . 28
10.1. Tin(II) Compounds . . . 28
10.2. Tin(IV) Compounds . . . 29
11. Organic Compounds of Tin . . . 30
11.1. Properties of Organotin Compounds . . . 31
11.2. Production of Organotin Compounds . . . 31
11.3. Industrially Important Compounds . . . 32
11.4. Analysis of Organotin Compounds . . . 33
11.5. Storage and Shipping of Organotin Compounds . . . 33
11.6. Pattern of Production and Consumption . . . 33
12. Toxicology . . . 33
13. References . . . 34

1. History [1–4], [6], [11], [15]

Because of its luster and softness, tin was usually assigned to the planet Jupiter, more rarely to Venus. The name of the element is derived from the Old High German *zin* and the Norse *tin*. The symbol Sn from the Latin *stannum* was proposed by BERZELIUS. Historically tin is of major cultural importance, being an essential component of the copper alloy bronze which gave its name to the Bronze Age. The first bronze objects appeared in Egyptian tombs dating from the end of the 4th millennium B.C.

Pure tin was first produced in China and Japan around 1800 B.C. Around 600 B.C., the ancient Egyptians occasionally placed pure tin artifacts in mummies' tombs. Tin is not only an essential constituent of tin bronze, but is also a constituent of lead alloys for solders and tin plating. Tin and especially its alloys have shaped the development of many geographical regions, e.g., China, Indochina, Indonesia, India, the Near East, North Africa, and Europe.

The cultural and historical importance of tin from the Middle Ages to early modern times lay

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_049

in its use for sacred objects, articles of daily use, and jewelry.

There is no historical evidence concerning the oldest methods of tin extraction. It is fairly certain that in 100 B.C. in Cornwall, England, tin was smelted from very pure ore over wood fires in pits and later in small furnaces. Up to the 1200s, Cornwall provided most of Europe's tin. Today, these deposits are virtually exhausted. Tin was probably produced in Bohemia around 1150. Also, the first tin mines were opened in Saxony at this time, and these supplied European requirements until they were destroyed in the Thirty Years War. Then, as these various deposits gradually became exhausted and as ocean transport developed, tin from overseas became dominant.

The largest tin mines are in Asia, the most important ore-supplying countries in the world being Malaysia and Indonesia, followed by China. The second largest tin-producing region includes Brazil and Bolivia. The countries exporting the largest quantities of tin ores also produce the most tin metal.

World annual production has developed as follows:

ca. 1800	9 100 t
ca. 1850	19 000 t
ca. 1900	91 900 t
ca. 1950	172 100 t
1980	243 600 t
1990	225 600 t

The principal consumer countries are the United States, Japan, China, and Russia.

2. Properties

Physical Properties [1], [2], [4], [16–19], [20]. Tin, Sn, exists in two crystalline modifications, the α- and β-forms. A third modification may also exist. Some physical properties of α- and β-tin are listed in the following. (see right column)

In the periodic table, tin lies on the boundary between metals and nonmetals. The transformation of the α- to the β-modification is accompanied by a complete change of lattice structure, affecting the physical, chemical, and mechanical properties. Also, at 170 °C there is a secondorder transformation accompanied by a discontinuous change in the lattice parameters and thermomechanical properties. A tetragonal highpressure modification of tin, stable between 3500 and 11 000 MPa, is described in the literature, the lattice constants being $a = 381$ pm and $b = 348$ pm.

Natural isotopes	10
Relative atomic mass	118.69
Crystal structure	
α-Sn (gray tin)	fcc (A_4) diamond type
β-Sn (white tin)	tetragonal (A_5)
Transformation temperature α-Sn $\leftrightarrow$ β-Sn	286.2 K
Enthalpy of transformation	1966 J/mol
Lattice constants at 25 °C	
α-Sn	$a = 648.92$ pm
β-Sn	$a = 583.16$ pm
	$c = 318.13$ pm
Density of β-tin	
20 °C	7.286 g/cm^3
100 °C	7.32 g/cm^3
230 °C	7.40 g/cm^3
Density of α-tin	5.765 g/cm^3
Density of liquid tin	
240 °C	6.992 g/cm^3
400 °C	6.879 g/cm^3
800 °C	6.611 g/cm^3
1000 °C	6.484 g/cm^3
Molar heat capacity of β-tin	
25 °C	27.0 J mol^{-1} K^{-1}
230 °C	30.7 J mol^{-1} K^{-1}
Liquid	28.5 J mol^{-1} K^{-1}
Melting point	505.06 K
Enthalpy of fusion	7029 J/mol
Boiling point	2876 K
Enthalpy of vaporization	295 763 J/mol
Vapor pressure	
1000 K	9.8×10^{-4} Pa
1800 K	750 Pa
2100 K	8390 Pa
2400 K	51 200 Pa
Cubic coefficient of expansion	
α-tin at −130 °C to +10 °C	14.1×10^{-6} K^{-1} to 4.7×10^{-6} K^{-1}
β-tin at 0 °C	59.8×10^{-6} K^{-1}
β-tin at 50 °C	69.2×10^{-6} K^{-1}
β-tin at 100 °C	71.4×10^{-6} K^{-1}
β-tin at 150 °C	80.2×10^{-6} K^{-1}
Molten tin at 700 °C	105.0×10^{-6} K^{-1}
Coefficient of thermal conductivity of β-tin at 0 °C	0.63 W cm^{-1} K^{-1}
Surface tension	
232 °C	0.53 – 0.62 N/m
400 °C	0.52 – 0.59 N/m
800 °C	0.51 – 0.52 N/m
1000 °C	0.49 N/m
Dynamic viscosity	
232 °C	2.71×10^{-3} Pa · s
400 °C	1.32×10^{-3} Pa · s
1000 °C	0.80×10^{-3} Pa · s
Specific electrical resistivity	
α-tin at 0 °C	5×10^{-6} Ωm
β-tin at 25 °C	11.15×10^{-6} Ωm
Transition temperature for superconductivity	3.70 K
Magnetic susceptibility of β-Sn	2.6×10^{-11} m^3/kg

The transformation of β-tin (white tin) into α-tin (gray tin) is of practical importance, as it involves a volume increase of 21 %. The transformation process requires a high energy of activation, and can be very strongly hindered. White β-tin can therefore exist for many years at $-30\,^{\circ}C$. The presence of α-tin seed crystals is important for the transformation process, and these are formed by repeated phase transitions. Foreign "elements" also affect the transformation temperature and rate. These can consist of impurities and deformations. The effect of impurities on transformation behavior is described in [21]. Tin vapor consists of Sn_2 molecules.

Mechanical Properties [1], [2], [4], [20–23]. Mechanical properties are not of great relevance to most applications of pure tin. The most important are listed in the following:

Yield strength at 25 °C	2.55 N/mm^2
Ultimate tensile strength	
−120 °C	87.6 N/mm^2
15 °C	14.5 N/mm^2
200 °C	4.5 N/mm^2
Brinell hardness (10 mm, 3000 N, 10 s)	
0 °C	4.12
100 °C	2.26
200 °C	0.88
Modulus of elasticity E	
−170 °C	65 000 N/mm^2
−20 °C	50 000 N/mm^2
0 °C	52 000 N/mm^2
40 °C	49 300 N/mm^2
100 °C	44 700 N/mm^2
200 C	26 000 N/mm^2

Chemical Properties [1], [2], [4], [20]. Tin has the atomic number 50 and is a member of group 14 of the periodic table. The electronic configuration is $1s^2 2s^2 p^6 3s^2 p^6 d^{10} 4s^2 p^4 d^{10} 5s^2 p^2$. Tin can be di- or tetravalent. It is stable in dry air, but is considerably more rapidly oxidized at a relative humidity of 80 %. Bright metallic tin becomes dull within 100 d even in indoor atmospheres. Oxygen is rapidly and irreversibly chemisorbed, and the oxide layer formed grows at an exponentially increasing rate. Typical impurities present after metallurgical production (e.g., Sb, Tl, Bi, and Fe) promote oxidation. Treatment with carbonate or chromate solutions leads to passivation.

Molten tin at temperatures up to ca. 500 °C picks up oxygen from the air at a rate that obeys a parabolic law, a result of the compact layer of oxide formed.

Gaseous water and nitrogen do not dissolve in solid tin. Dissolution in molten tin only occurs at high temperatures (ca. $> 1000\,^{\circ}C$). Under the conditions of electrochemical reduction in hydrochloric acid solutions, atomic hydrogen forms SnH_4, and elemental nitrogen forms Sn_3N_4.

Tin is stable towards fluorine at room temperature, but SnF_2 or SnF_4 are formed at higher temperatures.

Rapid and vigorous reactions occur with chlorine, bromine, and iodine, these reactions being accelerated by moisture and elevated temperatures. The reaction products are $SnCl_4$, $SnBr_4$, SnI_2, and SnI_4.

Sulfur reacts rapidly with molten tin at $> 600\,^{\circ}C$ to form the sulfides SnS, Sn_2S_3, and SnS_2. The reaction rate is lower above 900 °C, and only SnS is formed.

Reaction with hydrogen sulfide is slow and only occurs in the presence of oxygen and moisture. Sulfur dioxide reacts with molten tin to form SnO_2 and S, and a molten solution of tin in copper reacts with SO_2 to form SnO_2 and Cu_2S (an important reaction in pyrometallurgy). Tin is stable towards pure hot water, steam, and dry ammonia. Nitrogen oxides only react with molten tin.

Tin is amphoteric, reacting with both strong bases and strong acids with evolution of hydrogen. Having a normal electrode potential of −0.136 V, tin lies between nickel and lead in the electrochemical series.

With sodium hydroxide solution, tin forms $Na_2[Sn(OH)_6]$, and with potassium hydroxide solution $K_2[Sn(OH)_6]$.

Tin reacts slowly with acids in the absence of oxygen. The high hydrogen overvoltage is caused by a layer of atomic hydrogen at the metal surface preventing further attack. Vigorous reactions occur with nitric acid, the rate depending on the acid concentration. The reactions are very vigorous with 35 % acid, but complete passivation can occur at concentrations > 80 %. Tin is stable towards fuming nitric acid. While hydrogen fluoride does not attack tin, hydrochloric acid reacts even at a concentration of 0.05 % and temperatures below 0 °C. Tin is not attacked by sulfurous acid or by < 80 % sulfuric acid.

The most important use of tin and tin-plated materials is in the preserved food industry. For this reason, the possibility of reactions of tin with certain organic acids is important. Lactic, malic, citric, tartaric, and acetic acids either do not react at all at normal temperatures or do so to a negligible extent, especially in the absence of atmospheric oxygen. This is also true of alcohols and hydrocarbons.

3. Occurrence; Ore Extraction and Beneficiation

The average concentration of tin in the earth's crust is estimated to be 2 – 3 ppm, comparable to cerium and yttrium. Owing to the high atomic mass of tin and the high density of its important minerals, its volume concentration is very low. However, it occurs in only a small number of locations, where consequently its relative abundance is high. In general, 1000-fold enrichment is necessary to give workable tin deposits, i.e., with a tin content of at least 0.2 %. The question of whether a deposit can be economically extracted, for a given world market price level, depends on the mining conditions. For example, there are deposits in Bolivia containing 1 % Sn which cannot be economically extracted, whereas in South East Asia placer deposits containing 0.02 % Sn are successfully mined.

3.1. Minerals

Native tin occurs only very rarely and has only been identified with certainty in Canada.

Cassiterite, SnO_2, is the most economically important tin mineral. It forms tetragonal crystals, has a Mohs hardness of 6 – 7, a density of 6.8 – 7.1 g/cm^3, and a tin content of up to 79 %. The color is usually brown to brownish black. The presence of Ti, Fe, Nb, Ta, or Mn can lead to colors varying from gray to white. Contact deposits of cassiterite can be combined with, e.g., magnetite, arsenical iron pyrites, or zinc blende. Placer deposits of cassiterite are of major importance. "Wood tin" consists of gel-like or very fine grained aggregates of cassiterite. Cassiterite is an oxidic mineral and is chemically very resistant, in particular towards weathering.

Stannite (bell metal ore), $Cu_2(Fe,Zn)SnS_4$, forms tetragonal crystals, has a Mohs hardness of 4, a density of 4.4 g/cm^3, and a tin content of up to 27.6 %. Its color is steel gray with an olive green tinge. It is chemically less resistant than cassiterite, seldom occurs in hydrothermal deposits, and is of little economic importance.

Hydrocassiterite (varlamoffite), H_2SnO_3, is a tetragonal gel-like stannic acid. It occurs in Bolivia, usually accompanying cassiterite.

Other tin minerals include teallite, (Sn,Sb)S; herzenbergite, SnS; franckeite, $Pb_5Sn_3Sb_2S_{14}$; cylindrite, $Pb_3Sn_4Sb_2S_{14}$; thoreaulite, $SnTa_2O_7$; hulsite, (iron tin borate); and stokesite, $CaSn(Si_3O_9) \cdot 2\,H_2O$. None are of economic importance.

3.2. Deposits

The economically important tin deposits are closely associated with acidic to intermediate magmatic rocks which were formed in the orogenic phases of the earth's history. Tin, a volatile metal, was deposited primarily during the pegmatitic, pneumatolytic, or hydrothermal phase in the region of the exo- or endocontact of the intrusive bodies. The economically much more important secondary tin deposits exist as eluvial, alluvial, or marine placer deposits which can be near to or remote from these acidic magmatic rock complexes. An overview of the ore reserves and of the amounts extracted in 1990 is given in Table 1.

Table 1. Tin ore reserves and mine outputs for various countries in 1990

Country	Ore reserves (t metal content)	Output (t metal content)
Malaysia	1 200 000	28 500
Thailand	1 200 000	14 600
Indonesia	1 550 000	31 700
Bolivia	980 000	17 300
Russia	1 000 000	13 000
China	1 500 000	35 800
Australia	330 000	7 400
Brazil	400 000	35 100
Zaire	200 000	1 600
United Kingdom	260 000	3 200
South Africa	50 000	1 100
Nigeria	280 000	200
Others	765 000	17 000
World total	9 715 000	210 700

Primary deposits originate from the pegmatitic formation of cassiterite in contact with granites or their secondary rocks. Cassiterite occurs there in idiomorphic pyramidal crystals of > 2 mm diameter. These tin-bearing granites and pegmatites typically include the minerals quartz, albite, potassium feldspar, muscovite, and cassiterite. Columbite is an important accompanying mineral. Deposits of this type are found in central and southern Africa, Brazil, and Russia (Siberia). They account for < 5 % of world production.

Cassiterite quartzes of the pneumatolytic catathermal phase were formed in vein fissures of granites and their secondary rocks. Individual veins or lodes can have a tin content of up to 3 %. They can be 0.2 – 1 m thick and up to 200 m deep.

There are various types of paragenesis of the granites, leading to the greisen type (mica – feldspar – quartz formed by pneumatolysis with fluorspar, lepidolite, and tourmaline), the topaz – quartz type, the feldspar – quartz type, and the quartz type. Such deposits contain ca. 20 % of the world's tin reserves, and are found in Malaysia, Russia (Siberia), and Germany (Erzgebirge in Saxony).

Cassiterite – sulfide deposits of the hydrothermal phase are formed as vein ores associated with intrusions of granodioritic rocks, or are formed by remigration of the metal content of older pegmatitic tin deposits caused by younger acidic subvolcanoes. Characteristic of these types of deposit is the paragenesis of cassiterite with stannite, iron pyrites, arsenical iron pyrites, galena, zinc blende, magnetic iron pyrites, and copper sulfides. These deposits, which can be very large, are mined in Bolivia, Russia, southern China, Thailand, Burma, Australia, South Africa, and also England. They constitute ca. 15 % of the world's tin reserves.

Because cassiterite is resistant to weathering and is hard and dense, the weathering of the primary tin-bearing rocks enables it to become concentrated as it is transported to form secondary deposits, i.e., eluvial, deluvial, or marine placers.

Eluvial placers are formed by intense weathering and breakdown of cassiterite-containing granites or granodiorites, especially under tropical climate conditions. The lighter minerals are washed out or carried away by wind, while the cassiterite and other heavy minerals remain behind and can become concentrated in deposits of considerable thickness, as in Malaysia and Zaire. Very coarse and deluvial placers are also formed by gravitational enrichment due to landslips and eluviation at the bottom of mountainsides.

Alluvial or fluviatile placers are formed by the transport of weathered tin-bearing rocks by flowing water derived from atmospheric precipitation. The softer and lighter components of the rock are more extensively size reduced and therefore transported further than the hard, resistant cassiterite minerals, which sink due to their high density and are deposited at points where flow rates are low. The most important alluvial placer deposits are in central Africa (Zaire, Ruanda), western Africa (Nigeria, Niger), and Brazil.

Marine placers are formed where primary tin-bearing rock complexes have been directly transported by surf, or where rivers have carried the cassiterite-containing sediment into the sea, where it is then deposited in coastal strips. These are the most important deposits, and represent ca. 60 % of the world's workable reserves. The largest of these deposits are in southeast Asia, the coast of Thailand, the Thai Island of Phuket,in Malaysia, and on the Indonesian Islands of Bangka and Billiton.

Secondary deposits usually have tin contents of 0.05 – 0.5 %, reaching 3 % in some cases. Marine placers have tin contents of 0.01 – 0.03 %. At present, deposits containing 0.1 % Sn are workable by the open-pit method. For underground mining, deposits should contain 0.3 % Sn.

3.3. Mining

Primary tin ores are extracted by underground mining. Depths can reach 1000 m in exceptional cases. Most of the technologies used in nonferrous metal mining are used, the method in a given situation being determined by the thickness, shape, and orientation of the ore body, and geological factors.

In secondary deposits, the loosely packed weathered hard mineral rock which contains the cassiterite together with associated deposits of sand and gravel is extracted by high-production loading techniques which also perform prelimi-

nary classification. General local conditions, including the state of economic development in the region, have a great influence on the mining conditions. For example, in Zaire, surface eluvial deposits (weathered pegmatites) with tin contents of up to 0.15 % are extracted by conventional open-pit methods. In Thailand, Malaysia, and Indonesia, loose alluvial and marine deposits in river valleys and in the undersea regions just off the coast are extracted by dredging shovels, dragline excavators, chain and bucket excavators, and similar equipment specially designed for local conditions. The initial separation of gangue and other foreign materials (e.g., wood) is performed by this equipment. Cassiterite in thick deposits of loose sediments and coarse detritus in Southern China and Thailand is treated with powerful water jets operating at pressures of up to 1.5 MPa. These generate a mixture of water and heavy sand which is then fed to the treatment plant.

Off the coasts of Indonesia, Thailand, and Malaysia, chain and bucket excavators are used to extract cassiterite from alluvial deposits under water at depths of up to 40 m. This also gives a preliminary beneficiation.

3.4. Ore Beneficiation [11], [24–28]

The beneficiation of primary tin ores is difficult. The principal mineral, cassiterite, is nonmagnetic and is not suitable for flotation, so that mainly gravimetric sorting processes must be used. Furthermore, cassiterite is often strongly intergrown, and the accompanying minerals behave similarly to cassiterite during processing.

Current technologies are characterized by controlled multistage size reduction of the ores and separation of the cassiterite released after each size reduction stage using sorting methods based on density. Screen jigs and shaking tables of various designs are used. However, very small particles ($< 30\,\mu m$) cannot be processed to give satisfactory yields and production rates. If the degree of intergrowth of the ores requires finer grinding, as is increasingly the case with ores from Russia, the United Kingdom, Bolivia, South Africa, and Portugal, the flotation method for sorting particles $< 100\,\mu m$ is sometimes used. This technique was also used for the tin ore from the Altenberg region of Saxony, whose mines were closed in 1990. The following stages of beneficiation of primary tin ores are used:

Ores with an average degree of intergrowth are concentrated mainly by processes based on density. Flotation is increasingly used to sort fine-grained material and ground middlings obtained by the density-based sorting process, and has now become the preferred method for treating the most finely intergrown, complex tin ores. Flotation of cassiterite with particle sizes between 40 and 10 µm is mainly carried out with arsonic acids.

The flow diagram (Fig. 1) shows the flotation of primary tin concentrates to remove sulfides of similar paragenesis, followed by flotation of cassiterite from the preconcentrate, and magnetic separation of paramagnetic minerals from the flotation product.

Composition ranges for complex tin concentrates are:

Sn	5.6 – 60 wt %
S	1.0 – 15 wt %
As	0.1 – 3 wt %
Sb	0.1 – 2 wt %
Bi	0.1 – 0.5 wt %
Cu	0.1 – 0.7 wt %
Pb	0.1 – 3 wt %
Zn	0.1 – 4 wt %
Ag	up to 500 ppm
W	0.1 – 5 wt % WO_3
Nb/Ta	1 – 3 wt % $Nb_2O_3 + Ta_2O_3$

The ores from placer deposits are thoroughly broken down by natural weathering processes which separate the material roughly according to the rate at which it settles out of suspension. The fine-grained cassiterite is mixed with coarser sand or gravel. On board the floating dredges, which operate in artificial dredging ponds or natural surface waters, there is ore beneficiation equipment which produces a preconcentrate for further processing on shore (Fig. 2). The ore passes through a drum screen which removes coarse gravel (10 – 20 mm), wood, and other foreign bodies. The material passing through the screen is desludged in a hydrocyclone, and treatment on a three-stage screen jig then produces heavy metal concentrate for further processing on shore.

The ore obtained from placer deposits on land using water cannons or from the seabed using special ships with suction pumps is processed using conventional ore beneficiation methods such as screen jigs or screen troughs.

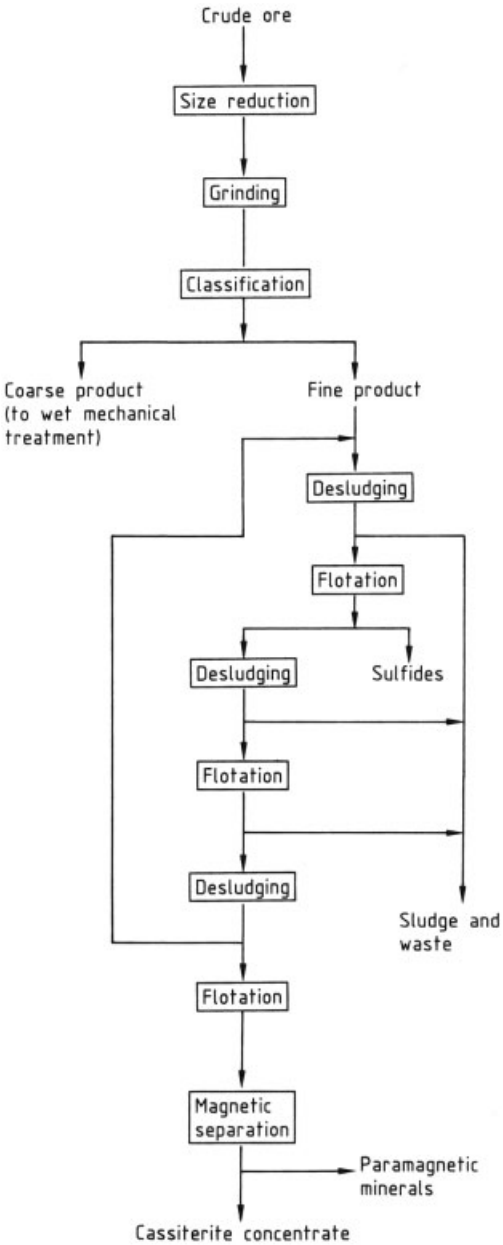

Figure 1. Flotation of primary tin ores

Concentrates from placer deposits are relatively pure; a typical analysis follows:

Sn	70 wt %	Ni	0.01 wt %
As	0.1 wt %	Ta_2O_5	0.2 wt %
Sb	0.05 wt %	Nb_2O_5	0.1 wt %
Pb	0.008 wt %	WO_3	0.05 wt %
Cu	0.005 wt %	SiO_2	5.0 wt %
Zn	0.01 wt %	CaO	0.1 wt %
Bi	0.015 wt %	TiO_2	0.1 wt %
Fe	0.3 wt %	Al_2O_3	1.0 wt %

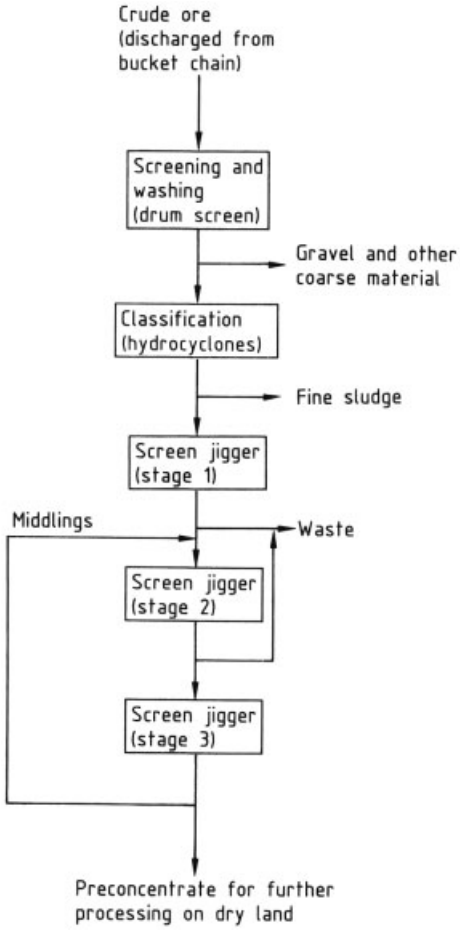

Figure 2. Preparation of secondary tin ores (placer deposits) on floating dredgers

4. Smelting

4.1. Fundamental Theory of Smelting

[1–3], [5], [29–34]

Because the most important tin-bearing mineral in tin ore is cassiterite (SnO_2), the carbothermic reaction

$$SnO_2 + 2\,CO \rightleftharpoons Sn + 2\,CO_2$$

is of fundamental importance.

A theoretical consideration of tin smelting must include the temperature dependence of this equilibrium and the behavior of the important reactants (Sn, O, and C) and accompanying elements and impurities in the concentrate, e.g., Fe, Cu, Sb, Bi, Pb, Ag, Si, Ca, Al, Mg, Nb, Ta, etc., of which Fe is the most important. The equilibrium diagrams for Sn – O – C and Fe – O – C are of crucial importance in the reduction of tin. At temperatures above 1100 °C, not only the Sn but also the Fe present in the oxidic precursors is reduced, so that selective winning from industrial tin-containing raw materials, which always also contain iron, is impossible. Furthermore, in this temperature range, up to 20 % tin dissolves in iron, and iron – tin compounds (known as hard head) are formed at lower temperatures, so that iron-free tin cannot be obtained under these conditions. This restricts the advantage of using a high process temperature to give a faster reaction rate in carbothermic cassiterite reduction. Also, the technique of stably binding iron in fayalite, to enable tin reduction to be carried out selectively, though realized on a laboratory scale by WRIGHT [29], could not be scaled up to production conditions.

The elements involved in tin reduction can be divided into the following groups:

1) Elements more noble than tin are reduced at lower temperatures than tin, and dissolve in molten tin (e.g., Cu, Pb, and Sb)
2) Elements that are much less noble than tin and which are not reduced under the reduction conditions, but which act as important slag formers, such as Ca, Al, and Si in the form of their oxides
3) Iron, the most important accompanying element, which behaves similarly to tin
4) Gaseous compounds produced in the reduction process
5) Sulfur, which has an important role in the reduction and volatilization process of pyrometallurgical tin production

Both the iron and the slag formers must be removed in a liquid slag; this determines the minimum process temperature. The iron content of the metal product depends on the Fe/Sn ratio in the slag. This relationship, represented in Figure 3, illustrates the principal problem of tin production from oxidic raw materials. If a high tin yield is to be obtained, i.e., small losses of tin in the slag (e.g., < 10 %), high reduction temperatures must be used, giving an iron content in the tin of > 8 %, so that subsequent refining is more difficult. If a purer metal is desired (e.g., 0.5 % iron in the tin), there will be high losses of tin in the slag, whose tin content can be 10 – 25 %. These slags are starting materials for a second process stage.

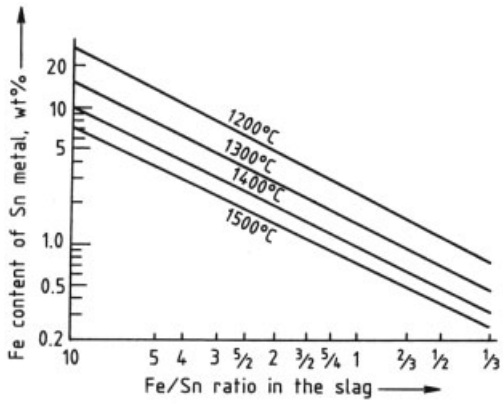

Figure 3. Calculated Fe/Sn ratios in the metal and slag at equilibrium as a function of temperature

The reactions in the slag phase are of major importance for selective reduction, whereby SnO is an important component. Experiments on pure substances have shown that although SnO melts at 980 °C it is unstable below 1100 °C, decomposing into SnO_2 and Sn. The activity of SnO in $SnO - SiO_2$ melts obeys Raoult's Law between 1000 and 1250 °C. The negative free energies of mixing SiO_2 with FeO, ZnO, PbO, MnO, and MgO increase in this order, and increase the SnO activities in the silicate melts. The position of the metal – slag equilibrium in the reaction

$$Fe + SnO \rightleftharpoons Sn + FeO$$

is expressed by the distribution coefficient

$$K = \left(\frac{Sn}{Fe}\right)_{metal} \cdot \left(\frac{Fe}{Sn}\right)_{slag}$$

K should be as large as possible, and at the usual reaction temperature of 1000 – 1100 °C used in

tin metallurgy, should be ca. 300 to give an iron content of ca. 1 % in the tin.

Binary and ternary slag systems containing tin oxides have been thoroughly investigated [33] and give useful guidance for carrying out the tin reduction. However, practical results depend very much on the viscosity of the molten products, the density differences, the surface tension, and colloidal dispersion and chemisorption of the slags. Thus, under production conditions stronger bases displace SnO bound in silicate, and FeO increases the fluidity of the slags.

The difference between the kinetics of tin oxide reduction and iron oxide reduction affects the selectivity of the reduction process and hence the iron content of the tin and the performance parameters of the furnace systems. Other important parameters are the pore structure and particle size of the raw materials, the partial pressures of the reduction gases, removal of the reaction gases, formation of seed crystals and coatings, and heat transfer. Thus, from the point of view of reaction rate, the reverberatory furnace is not the best equipment for carrying out reduction as it contains a large slow moving mass of material with a large bed thickness where heat is supplied only from above. The poor heat transfer leads to an extremely low smelting capacity, i.e., < 0.7 t metal per m^3 furnace volume per day. Because slag formation is very slow, some of the tin reduced at the start of the process can be tapped off as a relatively pure, low-iron product before the entire charge is smelted.

In contrast, highly turbulent reaction systems lead to process rates orders of magnitude higher. Under practical reaction conditions above 900 °C, the reaction

$$SnO_2 + 2\,CO \rightleftharpoons Sn + 2\,CO_2$$

is rapid, and the reaction

$$C + CO_2 \rightleftharpoons 2\,CO$$

becomes rate determining. This is why oxygen has to be added to the mixture of solid charge and reducing carbon in the furnace.

World production of tin is two-thirds from oxidic and one-third from sulfidic raw materials. The main problem in treating sulfidic tin concentrates is their complex composition. As many impurities as possible are vaporized in an initial roasting stage.

Arsenic requires special treatment, as it is oxidized to As_2O_3 and As_2O_5, which combines with Fe_2O_3, formed by roasting, to give nonvolatile iron(III) arsenate. In practical operation, a somewhat reducing atmosphere is therefore produced by adding charcoal to the charge. Another possibility is to vaporize heavy metal chlorides by adding NaCl.

4.2. Special Aspects of the Winning of Tin from its Ores

Problems in ore beneficiation often lead to concentrates with low tin contents, as unacceptably high losses of material would occur if ore concentrates with higher tin contents were produced. Therefore, a pyrometallurgical "thermal ore beneficiation" stage is necessary prior to the actual reduction process. This is a volatilization process that exploits the fact that the iron compounds and other slag formers have low vapor pressures at 1000 – 1500 °C, while SnO and SnS volatilize very readily. This technique can also be used to treat tin-containing slags.

The vapor pressure of SnS is considerably higher than that of SnO. Therefore, in practice, SnS is vaporized and then oxidized to SnO and SO_2. Pyrites (FeS_2) added as sulfur source causes problems in later stages of the process, as the FeO formed must be slagged, and the SO_2 evolved makes waste-gas cleaning necessary. However, this is outweighed by the advantages of a high yield of tin at relatively low process temperatures.

Economic advantages can be achieved by using cheap sulfur-containing grades of heating oil as fuel in the roasting process.

The naturally occurring ore beneficiation process that takes place in cassiterite deposits in surface waters and the increasing use of ore concentration processes with finely intergrown ores lead to extremely fine grained materials. Processes of agglomeration or compaction would be very costly. For this reason, these types of raw material are usually treated in an ore reverberatory furnace.

The main problem in pyrometallurgical tin production processes is separating tin from iron. Under production conditions, simultaneous reduction of SnO and FeO cannot be prevented. Molten tin can dissolve large amounts of iron,

and intermetallic compounds, which are very difficult to separate, can be formed on solidification. To minimize this problem, tin production is carried out in two stages. In the first stage, under mild reduction conditions, a relatively pure tin and a rich slag are produced. The latter is treated under strongly reducing conditions in the second stage, giving a discardable slag and a very impure tin – iron compound. The metallic phase is returned to the first stage, where the iron is reoxidized. The two-stage process must be carried out such that the iron initially in the concentrate is eventually removed from the process in the waste slags.

4.3. Production of Crude Tin

4.3.1. General Aspects

The choice of a crude-tin production process involves consideration of factors associated with both raw materials and location, and the following questions must be posed:

1) Whether the raw materials are highly enriched concentrates with low levels of impurities
2) Whether the raw materials used are low-grade concentrates whose principal impurities are slag components
3) Whether complex raw materials containing at least one other valuable element (e.g., W, Nb, or Ta) are used

Other important factors include the availability of raw materials, energy costs, environmental considerations, and personnel costs.

4.3.2. Ore Preparation prior to Reduction

As a result of the problems described in Section 3.4 and also of the efforts to recover as much of the tin from the ore as possible, the tin content of the ore concentrate can range from 8 to 60 %. Hence in most cases, pretreatment is necessary, e.g., pyrometallurgical enrichment of low-grade raw materials, a roasting stage (sometimes with addition of fluxes), or a leaching operation.

4.3.2.1. Pyrometallurgical Enrichment of Low-Grade Concentrates [1–5], [28–30], [34–41]

Low-grade tin concentrates are subjected to pyrometallurgical enrichment. The tin is vaporized as sulfide, and then oxidized in the gas phase by atmospheric oxygen to form SnO_2. The nonvolatile components of the raw material are selectively removed, and with careful process control, 90 – 95 % of the tin can be recovered as an oxidized product containing 40 – 60 % Sn.

Optimum results are only obtained if the reaction of the sulfide with oxygen takes place exclusively in the gas phase; therefore, reducing conditions must be maintained in the furnace. Although gaseous sulfur is the best sulfiding agent from a thermodynamic viewpoint, pyrites is used almost exclusively under production conditions. If calcium sulfate is used, the disadvantage that energy is required for its dissociation must be balanced against the advantage that the CaO produced is a useful slag former.

Although the vaporization of tin as SnS is used in all production plants, there is no generally accepted model of the reactions that take place. However, it is probable that in view of the S/Sn ratios required, the available polysulfidic sulfur in the pyrites does not take part in the formation of tin sulfide. Furthermore, it can be deduced from the importance of the added carbon to the reaction product obtained and from the CO content or oxygen demand of the reaction gases that SnO_2 is reduced first to SnO in an intermediate stage and SnS is then formed by the reaction of SnO with FeS. This is supported by the fact that the pyrites or the FeS formed therefrom forms SnS more readily than does sulfur vapor, contrary to what would be expected from thermodynamic considerations.

These results, especially when supported by practical experience, show that the overall equation and the individual reactions are probably as follows [28]:

$$SnO_2 + FeS_2 + 1/2\,SiO_2 + C + 3/2\,O_2 \rightleftharpoons SnS + 1/2\,(2\,FeO \cdot SiO_2) + SO_2 + CO_2$$

$$FeS_2 \rightleftharpoons FeS + S$$

$$S + O_2 \rightleftharpoons SO_2$$

$$C + 1/2\,O_2 \rightleftharpoons CO$$

$$SnO_2 + CO \rightleftharpoons SnO + CO_2$$

$$SnO + FeS \rightleftharpoons SnS + FeO$$

$$2\,FeO + SiO_2 \rightleftharpoons 2\,FeO \cdot SiO_2$$

Pyrometallurgical enrichment of low-grade tin concentrates can be carried out in various types of furnace.

Initially, rotary and shaft kilns were used for the vaporization process. Although the operation of both types of equipment was technically sophisticated, there were considerable disadvantages, which eventually led to their abandonment. Apart from the high energy and fuel requirements typical of both systems, the shaft kiln process led to a relatively low direct yield of tin in the flue dust, mainly due to the production of matte, which required separate processing. The rotary kiln method led to a low tin concentration in the flue dust, and, for raw material of high iron content, to the formation of matte, and hence to a drum clinker, which had then to be processed in a shaft kiln.

High-capacity thermal concentration processes, already established on a large scale in the nonferrous metal industry, were therefore adapted to tin enrichment. The fluidized bed and the cyclone smelting processes were not used. Also, the flash smelting (levitation smelting) process has severe limitations because of the raw materials used. Good results could be obtained with concentrates in which most of the material had a particle size of 200 – 300 µm. Using such concentrates containing 10 – 12 % Sn, discardable slags containing 0.2 – 0.4 % Sn and flue dusts containing 50 – 60 % Sn could be obtained. On using low-grade concentrates that had not been desludged and which contained significant amounts of material with a particle size of 50 –60 µm, the amount of primary flue dust increases significantly, and the tin concentration in the flue dust sometimes decreases to < 40 %. Also, the tin content of the slags is unacceptably high (1 – 2 %).

The main disadvantage of the flash smelting process is that it imposes strict requirements on the physical form of the concentrate.

The slag blowing process, originally used for detinning the slags from the reduction process, has been increasingly used for enrichment of tin in low-grade concentrates and exploitable gangue from ore processing. These products were at first added to the initial smelting process, but they were later added in solid form directly to the blowing furnace. Smelting and blowing can be carried out in a single furnace, so that it is not necessary to build a special smelting plant. This enables capital and operating costs to be reduced.

An important part of the practical operation is the maintenance of the correct fuel – air mixture with uniform distribution of the fuel and air to the individual jets. Heating oil and natural gas are preferred because they are more easily metered than solid fuels. If pyrites is used as the sulfiding agent, an S/Sn ratio of 0.8 is necessary, which places high demands on the equipment for removing sulfur dioxide. When pyrites is used, only FeS is effective in the slag phase, so that only 50 % of the sulfur is used for sulfiding. If pyrrhotite (magnetic pyrites) or calcium sulfide is used, the S/Sn ratio need only be 0.4, and the emission of sulfur dioxide decreases by 50 %.

The good metallurgical results of the slag blowing process, in which flue dusts containing 65 – 70 % Sn and final slags containing 0.1 % Sn are obtained, must be balanced against the disadvantage of a batch process. This leads to nonuniform loading of the downstream plant, e.g., the waste heat recovery and gas cleaning equipment.

4.3.2.2. Roasting [1], [2], [4], [5], [11], [17], [32], [42–50]

The roasting process, not only converts sulfides to oxides, but also volatilizes major oxidic impurities (e.g., arsenic). Roasting can be an independent process, or a pretreatment prior to hydrometallurgical leaching. It is significant that the level of impurities As, Sb, Pb, and Bi in tin concentrates has increased in spite of great efforts to improve ore beneficiation technology. If the levels of As and Pb are > 0.1 %, and of Bi and Sb > 0.03 %, a roasting process, sometimes with the addition of a leaching stage, is both useful and necessary for the benefit of the final tin reduction and refining processes.

The important reactions in the roasting process are as follows (M = metallic impurity):

1) Dissociation
   $FeS_2 \rightleftharpoons FeS + 1/2\,S_2$
   $4\,FeAsS \rightleftharpoons 4\,FeS + As_4$
2) Roasting reactions
   $MS + 3/2\,O_2 \rightleftharpoons MO + SO_2$
3) Oxidation
   $MO + 1/2\,O_2 \rightleftharpoons MO_2$

Although the roasting reactions and especially the oxidation reactions are exothermic, addition of fuel is necessary in industrial-scale roasting processes. The pore structure of the material must be maintained to enable the gaseous metallic and nonmetallic impurities to be vaporized. The upper temperature limit for the roasting process is imposed by the melting point of the low-melting sulfide eutectic. However, the temperature should be kept as high as possible to prevent the formation of sulfates, e.g., lead and calcium sulfate.

A mildly reducing atmosphere is necessary to suppress sulfate formation and also prevent formation of higher nonvolatile oxides of the impurities (e.g., As_2O_5).

The roasting processes are carried out in multideck or rotary kilns. The use of fluidized-bed furnaces with carefully controlled operating parameters has been reported [48].

Chloridizing roasting is also suitable for the pretreatment of tin concentrates, owing to the high affinity of the main impurities for chlorine. However, problems can be caused by chloridation of tin, which itself then volatilizes as $SnCl_4$ and $SnCl_2$.

Chloridizing roasting of tin concentrates in rotary kilns is carried out in Thailand. The contents of lead and bismuth are lowered from 2.0 % to 0.04 %, and from 0.1 % to 0.02 %, respectively. The flue dust contains 10 % As, 3 % Sn, 20 % Pb, and 4 % Bi. Treatment of this product presents serious problems [2].

The processes for removing impurities by chloridation differ fundamentally from those for the volatilization of tin from concentrates, being based on the fact that a large phase field exists in the phase diagrams $Sn-HCl-H_2$ and $Fe-HCl-H_2$ between 900 and 1000 °C in which selective chloridation and volatilization of tin is possible without chloridizing iron.

The overall reactions are as follows (M = Mg, Ca, or Na):

$$SnO_2 + MCl_2 + CO \rightleftharpoons SnCl_2 + MO + CO_2$$

$$SnO_2 + Cl_2 + C \rightleftharpoons SnCl_2 + CO_2$$

$$SnO_2 + 2\,HCl + CO \rightleftharpoons SnCl_2 + H_2O + CO_2$$

The presence of iron oxide increases yield and reaction rate, which can only be explained by intermediate formation of $FeCl_2$:

$$4\,SnO_2 + 6\,FeCl_2 \rightleftharpoons 2\,Fe_3O_4 + 2\,SnCl_2 + 2\,SnCl_4$$

Here, the instability of iron chloride in the presence of tin oxide is the reason for the good separation of tin from iron. In a reducing atmosphere, i.e., in the presence of C or CO, only $SnCl_2$ is formed.

In the Warren Spring process [50], $CaCl_2$ is used as chloridizing agent, and the CaO formed is used in the reductive smelting by adding limestone in accordance with the following equations:

$$SnO_2 + CaCl_2 + C \rightleftharpoons SnCl_2 + CaO + CO$$

$$SnCl_2 + CaCO_3 + C \rightleftharpoons Sn + CaCl_2 + CO + CO_2$$

The industrial-scale reaction is difficult to carry out, as moisture resulting from the hygroscopic properties of the tin chloride and the quicklime must be avoided. The corrosive properties of the reaction gases make very high demands on the construction materials used.

A special form of roasting is the heating of high-tungsten tin concentrates with NaOH or Na_2CO_3 to give soluble Na_2WO_4, which can be leached out.

4.3.2.3. Leaching [1–4], [32], [47], [51–55]

In tin metallurgy, the main use of leaching processes is to remove typical impurities from the concentrate. Tin, which is nearly always present in the raw materials as SnO_2, can only be brought into aqueous solution if the SnO_2 is reduced to SnO in a precisely controlled CO/CO_2 atmosphere. It can then be extracted in acid or alkaline media. This procedure has not yet been operated on an industrial scale.

Only hydrochloric acid is used on an industrial scale to remove impurities by leaching, the typical impurities in the concentrate, Fe, Pb, Cu, Sb, or As, going into solution in the form of their

chlorides. Best results are obtained using > 20 % hydrochloric acid at 100 – 110 °C. Suitable reaction vessels are high-pressure, acid-resistant spherical boilers with a capacity of 20 t. This batch process must sometimes be repeated several times. The solids are removed by thickeners and vacuum filters, and the dissolved impurities are precipitated from the liquor by cementation on scrap iron.

The high costs of these special reactors, the batch mode of operation, and the expense of the process for recovering the hydrochloric acid have restricted the use of hydrochloric acid leaching to some special cases.

It is preferable to carry out chloridizing roasting before leaching out the impurities. This then only requires a dilute acid solution.

Leaching of tungsten-containing tin concentrates to recover tungsten is important. After digesting the ore with sodium carbonate in spherical boilers, the tungsten is converted to its hexavalent form, which is soluble in hot water. It is then precipitated from the neutral solution with $CaCl_2$:

$$Na_2WO_4 + CaCl_2 \rightleftharpoons CaWO_4 + 2\,NaCl_2$$

An easily filtered precipitate containing up to 60 wt % WO_3 (on dry basis) is obtained.

Tungsten can also be extracted from tin concentrates by leaching with an aqueous solution of ammonia:

$$H_2WO_4 + 2\,NH_4OH \rightleftharpoons (NH_4)_2WO_4 + 2\,H_2O$$

The concentration of WO_3 in the filtrate can reach 50 g/L, and this can be precipitated as artificial scheelite ($Ca[WO_4]$). The tungsten content of the tin concentrate can be reduced to 0.5 %.

Most of the tin can be leached from a tin concentrate containing 2 wt % bismuth with 5 % hydrochloric acid at 80 °C [46].

Treatment with sulfuric acid is used to remove iron present as carbonate in Australian tin concentrates. By removing gangue material in this way, the tin content of the concentrate is increased from 37.2 % to 47.1 %.

A process for the removal of arsenic from tin concentrates by bacterial leaching with *Thiobacillus ferrooxidans* has been tested in Russian research institutes [54].

The tin can be leached out of ores that are difficult to treat if these are first reductively smelted with $CuCl_2$ and HCl, in accordance with the following series of reactions [42]:

$$Sn + CuCl_2 \rightleftharpoons Cu + SnCl_2$$

$$Cu + CuCl_2 \rightleftharpoons Cu_2Cl_2$$

$$Cu_2Cl_2 + HCl \rightleftharpoons 2\,CuCl_2 + H_2$$

Tin is precipitated from solution by adding zinc, and copper by adding iron.

4.3.3. Reduction

As explained in Section 4.1, it is not possible to obtain high yield and high metal purity at the same time. Reduction is therefore carried out in two stages, the first stage giving a relatively pure metal (up to 97 % Sn) and a rich slag (8 – 35 % Sn). This slag is treated in a second stage, and sometimes in a third. Slag treatment processes are described in Section 4.3.4.

Various types of furnace are used for reduction. Very low grade ore in lump form can be treated in a shaft kiln. However, as most concentrates obtained in ore beneficiation are very finely divided, and agglomeration, e.g., by sintering, is impossible, other types of furnace must be used in this case. Reverberatory or rotary air furnaces are often used for reduction, and electric furnaces are also employed. Each furnace type has its own advantages and disadvantages.

4.3.3.1. Reduction in a Shaft Kiln [1–3], [5], [30], [56], [57]

Shaft kilns are historically the oldest type used for tin reduction. They have their origin in the old Chinese natural draught furnaces made of rammed clay held together with wooden posts and operated on mountainsides. The hearth was usually sloped so that the molten product ran off continuously. Modern shaft kilns are water-jacketed, and have a melting capacity of 5.5 – 8 t m^{-2} d^{-1}. The good heat transfer and the continuous method of operation give a higher melting capacity than is obtained in comparable reverberatory furnaces. The air flow rate must be low to prevent volatilization. Shaft kilns require raw material in lump form. The iron content must be as low as possible in order to limit the formation of hard head in the reducing atmosphere.

4.3.3.2. Reduction in a Reverberatory Furnace [1–3], [5], [30], [32], [58–61]

The necessity for treating fine-grained concentrates from ore beneficiation has led to the replacement of the shaft furnace by the reverberatory furnace, and this is today the most important type of reduction equipment used in tin metallurgy. Modern reverberatory furnaces have internal dimensions of 3 – 4 m width, 10 – 13 m length, and 1 – 1.5 m height. In a freshly lined hearth the height available for the molten charge is no more than 0.5 m. This increases with increasing wear of the furnace bottom. The bath volume of modern tin ore reverberatory furnaces is 20 – 50 m^3. As the tin is very fluid at the reaction temperatures (up to 1400 °C), high pressures can be produced at the bottom of the hearth. To prevent this, either holes must be provided in the steel sheet bottom to allow the tin to drip into a lower chamber where it can solidify in stalactite form, or the reverberatory furnace must have a water-cooled bottom so that the tin solidifies in the gaps between the bricks.

The refractory bricks and mortar used to line the furnace must be made of high-grade chrome-magnesite. This must be very pure because of possible reactions of SnO or Sn with iron oxides or silica. Chamotte can only be used above the slag zone. To avoid disturbance of the brickwork in the melting zone, charging is carried out through the roof.

The burners, which use heavy fuel oil, are situated on the narrow sides of the furnace. The furnaces operate discontinuously using the regenerative principle, the duration of a heat being 16 – 20 h. Specific melting capacities are in the range 1.2 – 2.0 $t\,m^{-2}\,d^{-1}$. Charge batches consisting of concentrate, carbon, and fluxes weigh between 40 and 70 t. Optimum results can only be achieved by extremely careful operation.

Thus, at the start of a new heat, the amount of material charged to the furnace should be limited, e.g., by adding the charge in two portions, to prevent a large drop in temperature. The amount of carbon added in the form of coke, coal, petroleum coke, or charcoal is determined by the tin and iron contents and also by the incidence of hard head. The carbon addition is adjusted so that coreduction of the iron is prevented as far as possible.

The furnace draught must be kept low to prevent entry of cold air and to maintain the oxygen content in the waste gas below 5 vol %, thereby minimizing oxidation of the tin.

The first tin can be tapped off after 4 – 5 h of a 24 h cycle. As this is comparatively pure, it is preferable to treat it separately. The material tapped off later separates into metal and slag in a settler outside the smelting furnace. Some furnaces have facilities for tapping off the metal and the slag separately.

Up to 6 % of the weight of charge material is converted to flue dust, which consists of entrained charge, SnO, and some SnS. Its tin content is therefore considerably higher than that in the charge material.

Compositions of products obtained when a relatively pure concentrate (> 70 % Sn) is treated in a reverberatory furnace are listed in Table 2. The amount of slag is relatively small (14 wt % of the charge), but the slag must undergo a further stage of processing because of its high tin content.

Table 2. Typical compositions of products from a reverberatory furnace (in wt %)

Component	Crude tin	Slag	Flue dust
Sn	97 – 99	8 – 25	40 – 70
Fe	0.02 – 2.0	15 – 40	0.8 – 4.0
As	0.01 – 2.0		0.1 – 0.7
Pb	0.01 – 0.1	0.1 – 0.5	0.2 – 1.6
Bi	0.003 – 0.02		0.01 – 0.1
Sb	0.005 – 0.2	0.01 – 0.02	0.1 – 0.4
Cu	0.001 – 0.1		0.02 – 0.05
Al_2O_3		7 – 12	≤1.0
SiO_2		10 – 30	≤2.0
CaO		4 – 14	≤0.6
MgO		1 – 4	≤0.6
S	0.01 – 0.05		≤1.0

A simplified flow diagram for the treatment of high-grade concentrates is shown in Figure 4.

The process flow diagram for low-grade, impure, and complex concentrates is considerably more complex. With high-grade concentrates, a tin yield, including that obtained from slag treatment, of 99.6 % can be achieved, as actual losses of tin can only occur in the waste slag and by loss of airborne dust. In the treatment of low-grade concentrates, the yields of tin may only be 95 %, and in exceptional cases 92 %, because of the numerous tin-containing waste products.

Control of the CO partial pressure, which determines the reduction potential, is more diffi-

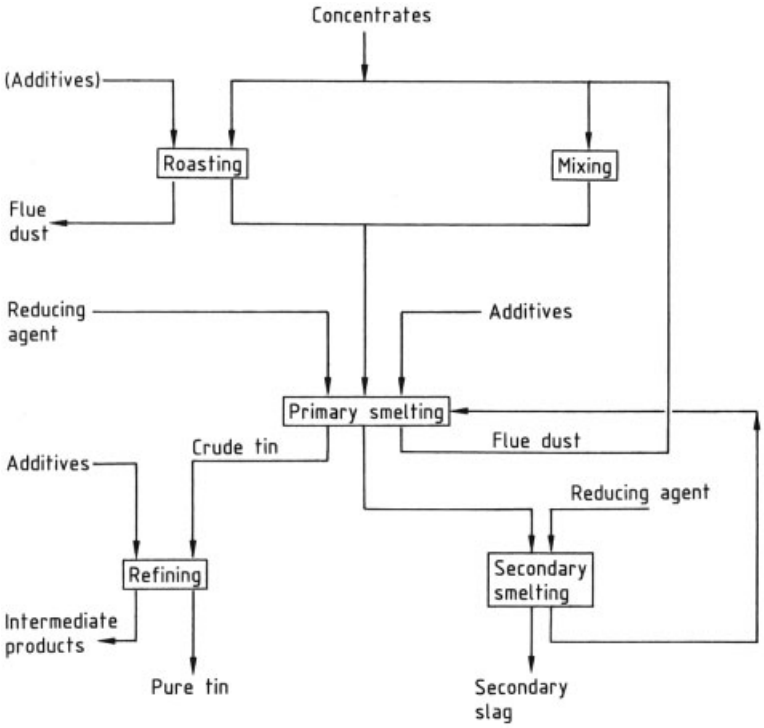

Figure 4. Process for treating high grade concentrates (simplified)

cult in a stationary reverberatory furnace, which contains a slow-moving mass of charge, than in a shaft furnace, in which the coke and air move in countercurrent flow. This is also the reason for the poor heat transfer and hence the low rates of reduction and melting in a reverberatory furnace.

Under operating conditions, care must be taken that the slag formers do not melt too rapidly, as this impairs the contact between the furnace atmosphere and the unmelted charge. At the same time, the melting temperature of the slag must be as low as possible, so that the use of such slag formers as CaO and SiO_2 should be limited. Added lime is of special significance, because it displaces tin from silicate, can lead to calcium stannate formation if present in excess, and increases the melting point of the slag.

The effect of the slag constituents, especially on melting point and viscosity, is described in detail in [3]. In a process optimization/cost minimization exercise, it is always necessary to include the effect on costs of the amount of tin tied up in materials being recycled [59].

4.3.3.3. Reduction in Rotary Kilns [1–3], [5], [30], [32], [57–62]

Rotary kilns (rotary air kilns) are horizontal smelting units that operate batchwise. They have a higher melting capacity than stationary reverberatory furnaces, but lead to considerably more severe stress on the refractory lining. Operating procedures in two tin smelting works in Indonesia and Bolivia in which the reduction process is based on the rotary kiln principle are reported in the specialist literature [63]. The furnaces have a length of 8 m, a diameter of 3.6 m, a surface area of reacting material of ca. 22 m^2, and a specific melting capacity of $1.36 - 1.5\ t\,m^{-2}\,d^{-1}$.

The furnace availability (300 d/a) is superior to that of reverberatory furnaces (260 d/a). Also, there is less requirement for mixing operations or agitation of the melt by stirring (rabbling). The metallurgical results of both types of furnace are very similar. However, the considerably poorer stability of the refractory lining, the higher energy requirement, and the significantly larger quantities of flue dust are all disadvan-

tages. Separation of the tin from the slag has to be carried out outside the furnace in a settler.

Tin concentrate can also be reduced in short drum kilns in which the ratio of the length to the cross-section is 1. The metallurgical function is basically similar to that of the rotary kilns.

The compositions of the smelting products in an Indonesian tin smelting works in which high-grade concentrates are reduced in a rotary kiln are given in Table 3.

Table 3. Typical compositions of products from a rotary kiln at Mentok, Bangka (in wt %)

Component	Crude tin	Slag	Flue dust
Sn	99.78 – 99.83	14 – 25	60 – 72
Fe	0.089 – 0.144	15 – 26	1 – 4
Pb	0.010 – 0.031		
As	0.010 – 0.188		
Sb	0.005 – 0.010		
Bi	0.0025 – 0.003		
Cu	0.002 – 0.025		
SiO_2		8 – 24	0.2 – 2.0
CaO		2 – 10	1 – 1.2
MgO		2 – 4	
S			0.2

4.3.3.4. Reduction in an Electric Furnace
[1–3], [5], [30], [32], [58], [64–68]

Electric resistance and arc furnaces used in metallurgy are characterized by high reaction temperatures and low waste-gas volumes. Disadvantages are the necessity for thorough premixing of the raw materials and the batch operation. Tin smelting is often carried out in regions where electrical energy is less readily available than energy from other sources such as gas, coal, or oil. Wherever electric furnaces are used in tin metallurgy, the object is to utilize their advantages of high reaction temperature and the production of heat by the Joule effect directly in the smelting bath.

Because their reducing action is so effective, electric furnaces are particularly suitable for extracting tin from slag (see Section 4.3.4). In concentrates whose iron content is significantly less than 5 %, it is even possible to produce crude tin in a single stage, with tin levels in the slag of < 0.7 %.

Electric furnaces are used in many tin smelting works for primary tin production from concentrates. These concentrates are often imported, e.g., in Germany, France, Italy, Canada, and Japan, although tin concentrate producing countries, e.g., Brazil, Zaire, South Africa, Russia, Thailand, and China also use electric furnaces. The raw materials for the electric furnace process must be intensively mixed. Very fine-grained materials such as flue dust are pelletized. A moisture content of ca. 2 % must not be exceeded.

Typically, circular furnace vessels with an outside diameter of up to 4.5 m and a height of 1.5 – 3 m are used, or oval furnace vessels with dimensions $2.5 \times 1.8 \times 1.8$ m. Heating is carried out with three-phase electric arcs using graphite electrodes. In single-phase furnaces, the furnace bottom acts as the counterelectrode for the immersed graphite electrodes.

Both stationary and tilting furnaces are used. Linings of carbon bricks give a service life of up to three years. Electric currents of 6 – 20 kA at 50 – 150 V are used. Depending on the characteristics of the raw materials, energy consumption is 750 – 1400 kW · h/t concentrate (1300 – 1860 kW · h/t tin). Precise control of the electrode immersion depth is essential for good control of the process.

Electric furnace technology enables a wide range of process parameters to be used. For example, when low-iron concentrates are treated, a tin quality suitable for normal refining can be produced. The high-tin slags produced (up to 30 % tin) are treated in a second stage to recover the tin. If the iron contents are much greater than 3 %, the tin obtained has a high iron content (3 – 10 %).

It is also possible to operate at ca. 1400 °C by using strongly reducing conditions. A tin-containing slag is then obtained together with hard head containing ca. 40 % tin and 50 % iron. The iron can be removed by smelting with ferrosilicon in a second pass, and a crude tin containing ca. 1 % iron is obtained.

Great efforts have been made to overcome the disadvantage of batch operation of the electric furnace. For example, a continuously operated lengthened double chamber electric furnace has been reported in Russia. Each chamber has a hearth area of 1 m^2 and two electrodes. The reduced tin flows out of the first chamber over an air-cooled overflow, and the tin is extracted from the high-tin slag in a second chamber [68].

The process is still at the pilot stage. The furnace capacity is reported to be less than 10 t/d.

4.3.3.5. Other Reduction Processes [1–3], [5], [30], [58], [60], [69]

All the processes described above have disadvantages, and many other methods and types of equipment have therefore been suggested for the reductive treatment of tin concentrates, but few of these proposals have led to an industrial plant.

One of the few methods tested at full scale is the top blown rotary converter (TBRC) developed in the United States and based on the "Kaldo" converter used in ferrous metallurgy. Oxygen is blown onto the top of charge as the converter rotates about its axis, which is set at an angle. The volume of waste gas is very low as there is no ballast nitrogen. The favorable heat transfer to the charge leads to a high reaction rate. Disadvantages include batch operation, a high rate of wear of the refractory lining, and the complexity of the system for controlling the oxygen lance.

After completion of the reduction process, the tin is tapped off. The tin in the remaining slag is vaporized as chloride by adding calcium chloride, and a discardable slag is thereby produced. The tin chloride is scrubbed out of the waste gases and then precipitated as SnO_2 by adding CaO, regenerating $CaCl_2$.

The potential for transferring proven processes used in nonferrous metallurgy to tin metallurgy is discussed in detail in [69].

Apparently no completely new processes for reducing tin concentrates have become established in the industry because of the high capital investment required and the hidden risks involved.

4.3.4. Slag Processing [1], [2], [5], [32], [38], [47], [62], [63], [70–79]

The slags produced during the reduction of high-tin concentrates can contain 5 – 10 % of the tin. This can increase to 20 % in the case of low-grade and complex ores. One- or two-stage treatment of the slag is then necessary. It is in principle possible to use strongly reductive smelting (e.g., in a reverberatory or rotary furnace) or blowing processes to volatilize the zinc from the slag.

On reductive smelting, the tin and iron form an alloy, the so-called hard head, which is recycled to the primary tin production process. The secondary slag has such a low tin content that it can be removed from the process.

In the blowing process, the tin is converted into a flue dust, which is recycled to the primary smelting process. The slag, which usually has a high iron content, can be discarded.

Under production conditions, the intermediate products, e.g., the tin-containing secondary slag, the hard head, and the flue dusts, contain considerable quantities of tin. These are important for the economic operation of the process because of the amount of capital tied up if they are not immediately treated. The tin and iron balances in the treatment of a concentrate with a very high tin content and a low iron content are shown in simplified form in Table 4.

Table 4. Tin and iron balance for two-stage smelting of a concentrate containing 73 % Sn and 0.7 % Fe based on 1000 kg concentrate treated

	kg Sn	kg Fe
Smelting of concentrate		
Charge :		
1000 kg concentrate	730	7
95 kg hard head	38	42
220 kg coke/coal		2
40 kg flue dust	18	
11 kg recycled material	6	3
Total	792	54
Product :		
733 kg crude tin	730	2
200 kg primary slag	47	51
30 kg flue dust	15	1
Total	792	54
Smelting of slag		
Charge :		
200 kg primary slag	47	51
40 kg coke/coal		
Total	47	51
Product :		
100 kg secondary slag	1	8
95 kg hard head	38	42
10 kg flue dust	5	
2 kg Fe alloy		1
Total	44	51

The balance shows that the iron introduced into the concentrate smelting process with the concentrate and reducing agent (7 + 2 kg) must be removed from the slag smelting process in the secondary slag and iron – tin alloy (8 + 1 kg). Of

the tin in the concentrate, 6 % is in circulation in the primary slag (47 kg out of 730 kg).

However, the concentrates treated usually contain considerably more iron, so that a larger amount of primary slag and hard head are produced, and the amount of tin contained therein can rise to 20 % of the raw material used.

In practical slag smelting, the SnO in silica-rich slags is mainly present as $2\,SnO \cdot SiO_2$, and the high activity of the SnO in silica- and lime-containing slags decreases with increasing FeO contents, so that in practice simultaneous reduction of tin and iron occurs. Theoretically, it is only possible to produce crude tin and low-tin slag if the FeO activities are extremely low. However, the processes would proceed at high temperatures with large additions of reducing carbon.

The binary Fe – Sn phase diagram shows a miscibility gap at $> 1100\,°C$ between 20 and 50 wt % Fe. On cooling to room temperature, separation of α-iron first takes place, followed by formation of FeSn and $FeSn_2$. At room temperature, the region of the composition of the hard head is always in the α-iron/FeSn two-phase region.

Wright has found by calculation of the distribution constant K at equilibrium

$$K = \left(\frac{Fe}{Sn}\right)_{slag} \cdot \left(\frac{Sn}{Fe}\right)_{metal}$$

that the balance of the process is improved as the ratio in the first quotient increases, i.e., as the amount of iron removed in the slag increases. However, there must then be a higher iron concentration in the tin. If the tin in the primary slag is present as $2\,SnO \cdot SiO_2$, the reaction mechanism is as follows:

$$C + SnO \rightleftharpoons CO + Sn$$

$$CO + SnO \rightleftharpoons CO_2 + Sn$$

$$CO + FeO \rightleftharpoons CO_2 + Fe$$

$$CO + Fe_2O_3 \rightleftharpoons CO_2 + 2\,FeO$$

$$3\,FeO \rightleftharpoons Fe + Fe_2O_3$$

$$Fe + SnO \rightleftharpoons FeO + Sn$$

$$2\,FeO + SnO \rightleftharpoons Fe_2O_3 + Sn$$

Thus the reaction first occurs only at the surface of the carbon, and then via the iron oxides as intermediate phases. A high relative rate of reaction of carbon, slag, and reaction gas is therefore important for the reaction kinetics.

The reactions that occur in the detinning of slag are described in Section 4.3.2.1 (pyrometallurgical enrichment of low-grade concentrates), but here they start from SnO. As a matte phase must be present, the reaction will be:

$$SnO + FeS \rightleftharpoons SnS + FeO$$

As the slag is saturated with FeS, the amount of SnS formed is proportional to the activity of SnO in the slag, but indirectly proportional to the activity of FeO.

Reverberatory or electric furnaces are usually used in the reductive smelting process for the extraction of tin from primary reduction slags, though shaft kilns are occasionally used. The necessary intense reducing effect is achieved by adding 10 – 20 % reduction carbon and by operating at temperatures up to 1500 °C. In the electric furnace process, energy consumptions are 500 – 1000 kW · h/t, and 1 – 10 kg electrode is consumed per tonne of slag. The reaction products, i.e., the final slag and the hard head alloy, are separated in settlers and then granulated in water.

As highly turbulent conditions are favorable to the process, reductive detinning can be carried out with a lance ("submerged combustion"), which produces a high reaction rate by agitation of the bath. Methane or natural gas have been used as reducing gas, and experiments have also been carried out using hydrogen, carbon monoxide, and powdered solid fuels. Heat is produced by partial combustion of the injected gases and transferred to the molten slag. However, because SnO is volatile, 20 % of the tin goes into airborne dust, a further undesired tin-containing reaction product in addition to the hard head.

An effective method of detinning primary tin reduction slags is by the blowing process, which can be carried out in reverberatory furnaces or true blowing furnaces. Gypsum or pyrites can be used as the sulfur source. The gypsum is first reduced to CaS, and then reacts with SnO to form SnS and CaO. If pyrites is used, impurities such as Pb, Zn, or As are introduced into the metallurgical process, adding to the difficulties of

treating the tin-containing flue dusts. Blasting technology is also often used as the third stage of slag detinning. In the second stage, a thermal reduction, Sn contents of 2 – 5 % and FeO contents of ca. 30 % are obtained. The tin content of the molten slag in the furnace is reduced to ca. 0.5 % by addition of sulfur sources. The addition of pyrites can be up to 400 % of the theoretical amount calculated for the formation of SnS. Since smaller additions are possible in the case of acid slags, CaS is probably formed in basic slags. The treatment of primary slags in true blowing furnaces consists of blowing the pyrites into the molten slag. However, the process is discontinuous. The SnS vapor is oxidized in the furnace atmosphere, and recovered as SnO dust by filtration.

The slag blowing process can be carried out in furnaces of various designs, e.g., in a type of shaft kiln with a floor area of 0.5 – 6 m^2 and a water-cooled shaft 7 m in height. The charge of primary slag can be 5 – 20 t. Cyclone and short drum kilns are also used. Process parameters, e.g., ratios of sulfur source to oxygen, fuel to atmospheric oxygen, and tin to sulfur must be carefully controlled to obtain optimum results.

Other proposed methods of detinning primary tin slags, such as vacuum-assisted volatilization of SnO (which has a considerably lower vapor pressure than SnS), have not been used on an industrial scale [29].

5. Refining [1–3], [5], [32], [46], [47], [80–84]

The crude tin obtained by the reduction process is insufficiently pure for most applications.

Most national standards specify maximum contents of typical impurities. However, nonmetallic impurities such as oxygen and sulfur and less common impurities such as noble metals are neglected, and the tin content is simply determined by subtracting the total amount of analytically determined impurities from 100.

The following three standard grades are accepted internationally: Sn 99.0 %, Sn 99.75 %, and Sn 99.9 %.

There are some variations in these standards from country to country; the German standard is DIN 1704.

The level of impurities in the crude tin determines the extent of the refining operation. Treatment of very pure concentrates can give a tin content of up to 99.0 % in the crude metal. The main impurity is iron (0.8 %), the sum of all the other impurities being only 0.2 %. In the case of low-grade and complex concentrates, the situation is very different, the tin content of the crude tin obtained sometimes being only 92 %.

The impurities Fe, As, Sb, Cu, Ni, Pb, Bi, and the noble metals affect the amount of work involved in the refining process. The metals Zn, Cd, Mg, Si, Ca, Te, Se, and also sulfur and oxygen do not require special treatment, as they are present in the crude metal in only small concentrations, and are removed together with the other impurities during the various stages of purification.

The phase diagrams for tin and its typical impurities lead to the following conclusions:

In the temperature range between 1000 and 1300 °C used in pyrometallurgical reduction, the impurities are completely soluble with the exception of Fe and Cr.

Only Sb, Cd, Bi, Zn, and Pb are significantly soluble in tin at room temperature. This is the basis for the removal of insoluble impurity elements by liquation. However, the liquation products have high tin contents, making costly recovery processes necessary.

In pyrometallurgical tin refining, the individual impurities are removed stepwise in batch processes. The use of the time-consuming operations is justified by their high selectivity.

Proposed continuous processes have not been operated on an industrial scale.

5.1. Pyrometallurgical Refining

5.1.1. Removal of Iron

The process for removing iron is based on the temperature dependence of the solubility of iron or Sn – Fe mixed crystals in tin. Accurate experiments have shown that the solubility of iron in tin at 250 °C is 0.0058 wt % [81]. In industrial practice, even lower figures are achieved, which can only be explained by other impurities, such as Cu, As, or Sb, causing deviations from ideal solubility behavior. On cooling molten crude tin,

α-Fe, γ-Fe, FeSn, and $FeSn_2$ precipitate successively. The density of the precipitated compounds is almost the same as that of the molten tin. In practice, "poling", i.e., passing steam or air into the melt, is used to coagulate the precipitated particles, which rise to the surface of the bath and are removed from the molten tin by filtration through graphite, slag wool, or a slab made of silica and limestone chippings. This should be carried out just above the melting point of tin, or in practice at a temperature not less than 260 °C. The process is sometimes carried out in two stages. Iron contents of 0.003 –0.01 wt % can be achieved.

As Ni, Co, Cu, As, and Sb form intermetallic compounds with each other as well as with Fe, these impurities are also removed to some extent.

The treatment of the recovered intermetallic compounds is very complex, as these are in the form of a slurry with large amounts of adhering molten tin. The iron content is only a few per cent, i.e., considerably less than that of the intermetallic compound $FeSn_2$.

The metal slurry is treated in small liquation furnaces. A controlled temperature increase over the range 230 – 300 °C enables pure tin to be removed, and a residue containing ca. 15 % Fe suitable for use in the primary smelting stage to be obtained.

In high-capacity tin smelting works, high-temperature centrifuges are used. These enable a solid residue containing up to 25 % Fe to be obtained [83].

5.1.2. Removal of Copper

After iron has been removed by liquation, the copper content is up to 0.01 %. Elemental sulfur (2 – 5 kg/t) is stirred in at 250 – 300 °C, enabling copper contents as low as 0.001 % to be attained. The resulting copper dross can be removed from the process after several stages of enrichment.

5.1.3. Removal of Arsenic

After removal of iron by liquation, the arsenic content is ca. 0.1 %, significantly higher than permitted levels. For example, the commonly used Sn 99.75 grade should contain < 0.05 % As according to DIN 1704.

Arsenic can be removed from the melts along with some Cu, Ni, and residual Fe by forming intermetallic compounds with aluminum. For this, the aluminum must be present as an ideal solution in the tin. For this reason, the melt must be heated to a temperature close to the melting point of aluminum before the aluminum is added. Special precautions are necessary, e.g., operation in a closed vessel so that the aluminum does not burn on the surface of the bath. The amount added must be approximately three times the stoichiometric amount. The use of Al – Sn master alloys enables the operation to be carried out at a lower temperature. Intense agitation if followed by a settling process with cooling to 350 – 400 °C, and the Al – As mixed crystals can then be removed. Separation of the Al mixed crystals is assisted by poling.

With correct operation, As contents of < 0.02 % can be achieved, i.e., below the permitted level for Sn 99.90 %. This process enables Sb contents of 0.005 %, Cu contents of 0.02 %, and Ni contents of 0.005 % to be attained, and any remaining Fe to be removed.

Any aluminum remaining in the molten tin can be removed by adding sodium, sodium hydroxide, chlorine, or steam, and residual sodium by adding sulfur.

The storage, transport, and treatment of the Al – As product presents problems. Contact with water must be avoided as this leads to the formation of highly toxic arsine and stibine. The material is stored in closed vessels and is converted into a safe product as soon as possible by oxidative roasting or by treating with alkali solution and collecting and burning the liberated arsine to form As_2O_3 and H_2O. The residues obtained both from the roasting and the leaching processes can be recycled to the process.

The large differences in vapor pressure between the impurities (arsenic, antimony, bismuth, and lead) and tin enable selective evaporation at reduced pressure to be used. However, numerous proposed processes have resulted in only two industrial applications.

In a system tested in Russia, the impure tin flows from the top of a vertical reactor under a vacuum of 1 Pa into heated evaporating dishes. The evaporated impurities are collected in a separate chamber. Barometric valves are used to remove the purified tin and the condensate.

In the Bergsöe – Redlac system, a cylinder, cooled on the inside, rotates in a horizontal vacuum chamber above the melt, and the vaporized impurities are deposited on this in solid form. In the next stage of the process, they are scraped off and removed. The results of the vacuum distillation process depend on the reaction temperature and time. The high energy requirement for heating and for producing the vacuum, 400 – 700 kW · h/t tin, is a disadvantage. Moreover, significant quantities of tin are vaporized. The use of selective evaporation of typical impurities in crude tin always involves a compromise between the purity and tin yield. Practical experiments are described in [85].

5.1.4. Removal of Lead

If the lead content is still not low enough after the first stages of the refining process, the lead can be converted into its dichloride by treatment with chlorine, tin dichloride, or tin tetrachloride:

$$SnCl_2 + Pb \rightleftharpoons PbCl_2 + Sn$$

The equilibrium is shifted to the right at low temperatures, so that the process must be carried out at a temperature only a little above the melting point of tin. The process also removes any remaining zinc and aluminum.

The best results are obtained by using two-stage operation, i.e., the product of the second stage, with a reduced lead content, is returned to the first stage. Precise control of the operation can lead to final lead contents of 0.008 wt %.

5.1.5. Removal of Bismuth

In analogy to the thermal refining of lead, bismuth can be precipitated as an intermetallic compound by adding calcium or magnesium. The molar ratio Ca/Mg should be ca. 2 to obtain the best results [46]. A ternary compound is probably formed at this ratio. Under production conditions, scrap magnesium is used, as this is the most economic material.

Final bismuth contents of 0.06 – 0.003 wt % have been reported for full-scale plant. In analogy to a technique used when removing arsenic from aluminum, the calcium and magnesium remaining in the tin can be converted to their chlorides and removed, e.g., by treatment with ammonium chloride.

5.2. Electrorefining

The theoretical conditions for the electrorefining of tin are favorable. The position of tin in the electrochemical series of the elements in aqueous solution show that Au, Ag, Cu, Bi, As, and Sb do not go into solution under electrorefining conditions, but will appear in the anode slime. The elements Ni, Fe, Zn, and Al can be largely removed by a preliminary pyrometallurgical refining operation. Only lead lies close to tin in the electrochemical series. The high electrochemical equivalent weight of tin also favors the use of an electrometallurgical refining process. However, there are considerable problems in the practical realization of the process.

Simple and cheap electrolytes such as solutions of sulfate or chloride lead to spongy or needle-like deposits, and these effects are only slightly moderated by extremely large additions of colloidal materials. The process can only be operated at low current densities, leading to low process rates and inefficient utilization of energy (e.g., low current yields). Also, the presence of large amounts of expensive metal tied up in the process is undesirable economically.

These negative aspects mean that electrorefining is only worthwhile if the tin contains high concentrations of noble metals.

Electrorefining can be carried out in acid or alkaline medium.

5.2.1. Electrorefining in Acid Medium

When sulfate electrolytes are used, additions of chloride, fluoride, crude cresol, glue, nicotine sulfide, α- and β-naphthol, diphenylamine, phenol, and/or cresolsulfonic acid are made. The sulfate ions cause the anodically dissolved lead to go into the anode slime in the form of lead sulfate. Also, sulfides such as nicotine sulfide can lead to the formation of lead sulfide, which is deposited in the anode slime. The organic sulfonic acids prevent the formation of basic tin salts on the anodes.

In spite of these precautions, the formation of coatings on the anode is the main problem in the electrolytic refining of tin in an acid medium. The main cause of coating formation is the lead content of the anodes, which must be removed mechanically when the bath voltage increases. The following operational data are quoted:

Anode mass:	100 – 200 kg
Anode thickness:	30 mm
Cell dimensions:	3.0 – 4.5 m long, 1.0 – 1.2 m wide, 1.0 – 1.5 m deep
Cell construction:	wooden cells with lead cladding or concrete
Cathode replacement:	after 6 d
Anode replacement:	after 10 – 12 d
Anode composition:	94 – 96 % Sn
	0.01 – 0.03 % Fe
	0.3 – 1.3 % Pb
	0.1 – 0.6 % Cu
	0.1 – 3.5 % Bi
	0.02 – 0.35 % As
	0.1 – 0.25 % Sb
	100 – 300 g/t Ag
	0.3 – 0.7 g/t Au
Cathodes:	starter sheets of pure tin

The current yield is largely determined by the rate of removal of anode passivation. The energy consumption is 150 – 200 kW · h/t tin. Because iron accumulates in the electrolyte, regeneration of the electrolyte is necessary.

5.2.2. Electrorefining in an Alkaline Medium

In alkaline medium, i.e., in NaOH or Na_2S electrolytes, less pure anodes (75 % Sn) can be used than are used in an acid medium. A smooth deposit can be obtained without addition of colloids. However, current densities are very low, and the process must be carried out at 90 °C.

Detailed information about the possibilities and limitations of electrorefining, deposition behavior, deposition mechanism, and effects of additives to the electrolyte, pH, and impurities are given in [4].

5.2.3. Other Methods of Electrorefining

Many attempts have been made to use molten salt electrorefining to overcome the disadvantages of electrorefining in aqueous solutions. The electrolyte was molten $CaCl_2$ – KCl –NaCl. Various grades of crude tin were used. The electrodes consisted of graphite crucibles and graphite rods. The operating temperature was ca. 650 °C, and the current density 50 –200 A/dm^2. Arsenic was effectively removed (reduced from 1.5 wt % to 0.005 wt %), and the antimony content was reduced from 0.32 wt % to 0.01 wt %. It proved impossible to scale up from pilot scale to full scale operation, mainly because of problems in the control of the high-temperature process [77].

6. Recovery of Tin from Scrap Materials and Residues [1–3], [5], [11], [86], [87]

Scrap materials and residues which are produced during the processing of metals to give semifinished and finished products are usually known as "new scrap," while the returned old material from industry, trade, building construction, factories, and consumption is known as "old scrap." The use of old and new scrap supplements primary production. It is collected by scrap merchants who work directly with the smelters. Also, in the neighborhood of metal smelting works there are often small scrap metal operations which sort, separate, refine and blend with primary metals to produce a primary metal that corresponds to the standard specification of an original metal [87].

The recovery of tin from tinplate is becoming increasingly difficult, as the change to electrolytic methods of tinning is leading to very thin coatings which sometimes diffuse into the steel sheet. The recycling of this material will continue to be a technically and economically difficult task. Two processes are used for recovering tin from tinplate: the alkaline electrolytic method and the alkaline chemical method.

In the alkaline electrolytic detinning of tinplate, baskets of cleaned scrap are immersed in hot 5 – 10 % sodium hydroxide solution. The baskets form the anodes, and the tinned steel sheet forms the cathodes. The tin is deposited in the form of a sponge. As contact with atmospheric carbon dioxide cannot be prevented, sodium carbonate is formed, and the electrolyte bath must be frequently regenerated. The lacquer coating on the scrap tinplate is removed by adding solvents to the bath or by a special

pretreatment process. The bath is operated at a temperature of 65 – 75 °C, a voltage of 1.5 V, and a cathode current density of 300 A/m^2.

In the alkaline chemical detinning of tinplate, the scrap material in perforated containers is immersed in sodium hydroxide solution. Hydrated sodium stannate is formed according to the equation

$$Sn + 2\,NaOH + 4\,H_2O \rightarrow Na_2SnO_3 \cdot 3\,H_2O + 2\,H_2$$

The hydrogen liberated must be reacted with an oxidizing agent; sodium nitrate is suitable. The dissolution process is accelerated by motion of the container in the liquor. The detinning time is 2 – 4 h, depending on the concentration and the temperature. If sodium nitrite is used instead of sodium nitrate, the tin goes into solution, the foreign metals, e.g., lead, iron, and antimony, can be precipitated, e.g., by hydrogen sulfide, and the tin can then be recovered by electrowinning.

Processing tin-containing alloys is easier than recovering tin from scrap tinplate. However, the tin content of many alloys has decreased over recent years, and alloys have in some cases been replaced by cheaper materials. Both these developments have tended to limit the potential for recovering secondary tin, and also explain why the amount of recovered tin in Europe has decreased from 15 800 t in 1980 to 13 300 t in 1990. In the United States, the amounts of recovered tin have remained almost constant at 16 900 t in 1980 and 17 100 t in 1990 [9].

The quantification of tin recovery from secondary raw materials is difficult because most of it is obtained from scrap alloy.

In addition to the tin concentrates treated in the primary smelting process, there are also other materials that must be treated, including the slags (see Section 4.3.4), oxidic flue dusts, ash, and sweepings containing very variable amounts of tin. Tin-containing processing scrap is also produced during casting, metal forming and cutting, tinning, and alloying. Considerable amounts of recycled scrap also come from the manufacture of cans, tinplate, tubes, foil, pure tin articles, and alloys such as solders, antifriction and bearing metals, type metal, etc.

In recycling it is essential to sort tin-containing materials into standard grades with exact analytical specifications so that the metallurgical process can be optimized.

Short drum furnaces are suitable for processing oxidic materials, although shaft kilns and reverberatory furnaces are also used. The reduction is performed by coke/coal, with added sodium carbonate or fluorspar as flux. The process is operated in one or two stages, depending on the material. Scrap alloys containing high proportions of lead, antimony, or copper are remelted to form alloys. Scrap babbitt (bearing metal) contains zinc, which is either removed in the slag or selectively volatilized. Scrap solder alloys are refined like crude tin.

7. Analysis [88]

7.1. Analysis of Ores and Concentrates

Determination of Sn. In the determination of tin in ore concentrates, the choice of method depends on the presence or absence of typical impurities.

Tungsten-Free Ores and Concentrates. After fusion with sodium peroxide and dissolution in water, the solution obtained is acidified with hydrochloric acid and partially evaporated to drive off arsenic. The antimony is then precipitated with iron powder (cementation). The tin in the filtrate can then be "cemented" by adding aluminum powder, and determined by iodometric titration.

Tungsten-Containing Ores and Concentrates. The tungsten can be precipitated from the acidified solution by adding cinchonine. The excess of this reagent in the filtrate is decomposed by fuming with sulfuric acid. The solution is then taken up in hydrochloric acid and treated as in the determination of tin in tungsten-free ores.

Silicate-Containing Ores and Concentrates. The material is boiled to dryness with nitric acid, and the residue is then strongly heated, fumed with HF/H_2SO_4, and fused with sodium peroxide. The tin determination can then be carried out as for tungsten-free materials.

Determination of Other Elements. The determination of tungsten is carried out by precipitating with cinchonine after the fusion stage. The precipitate is strongly heated (ca. 750 °C) to form WO_3, and tungsten is then determined gravimetrically.

Arsenic is determined by distilling it out of the acidified solution of fused product and then carrying out a bromatometric titration on the distillate.

Antimony is determined by adding iron powder to the acidified fusion product to precipitate antimony sponge. This is dissolved in Br_2/HCl, and any arsenic still present is driven off by evaporating with HCl. The antimony is purified by reprecipitation, and determined by bromatometric titration.

For the determination of other elements such as cadmium, iron, nickel, copper, bismuth, lead, zinc, and the noble metals, it is best to remove zinc, arsenic, and antimony from the dissolved product of a sodium peroxide fusion (acidified with HCl) by evaporating with Br_2/HBr. The elements can be determined in the solution by atomic absorption spectrometry.

To determine sulfur, the sample is heated in a stream of oxygen, the SO_2 formed is collected in a dilute solution of H_2O_2, and the H_2SO_4 formed is determined by titration. The sulfur content of the tin is calculated from this.

7.2. Analysis of Metallic Tin

The preferred methods of analysis of pure and commercial-grade tin are detailed in DIN 1704 [88].

Arsenic is determined by dissolving the metal in a strongly acidic solution of $FeCl_3$. The $AsCl_3$ formed is distilled off, and can be determined photometrically by the molybdenum blue method.

Antimony is determined by dissolving the metal in concentrated sulfuric acid. After adding hydrochloric acid, the antimony is oxidized with cerium(IV) to antimony(V). This is then extracted with isopropyl ether and can be determined photometrically as yellow potassium tetraiodoantimonate.

Lead is determined by dissolving the metal in a mixture of Br_2, HBr, and $HClO_4$, fuming this to volatilize the zinc, removing any Tl that might still be present by extraction with isopropyl ether, and determining the lead by polarography.

In the determination of copper, the initial dissolution is similar to that used in the determination of lead. The copper in the solution then is determined photometrically as the diethyldithiocarbamate complex after extraction with CCl_4.

The determination of zinc also requires a dissolution process similar to that for lead. The solution is acidified with hydrochloric acid, and heavy metals still present are precipitated as sulfides. The zinc in the filtrate obtained can be determined by polarography.

Iron is determined by first carrying out the dissolution and precipitation of heavy metals as for zinc determination. The iron in the filtrate can be determined photometrically as the sulfosalicylic acid complex.

The determination of aluminum also requires dissolution and removal of heavy metals by precipitation of their sulfides as for iron determination. The aluminum in the filtrate can be determined photometrically as the eriochrome cyanine complex.

As well as the analytical methods recommended in DIN 1704, rapid and efficient methods such as atomic absorption spectrometry (AAS) and atomic emission spectrometry with inductively coupled plasma excitation (ICP-AES) are also being used to an increasing extent [89]. These methods are specified in the German and international standards for the analysis of water, wastewater, and sewage sludge.

After dissolution of the tin in a mixture of Br_2, HBr, and $HClO_4$, the accompanying antimony, arsenic, bismuth, copper, iron, lead, aluminum, cadmium, and zinc can be determined with a high degree of accuracy in the presence of each other, without first separating them, by means of AAS and ICP-AES. In the determination of arsenic and antimony, other dissolution processes must first be performed. The analysis of metallic tin and tin alloys is often carried out by spectral analysis with spark or arc excitation, and the analysis of tin slags and concentrates by X-ray fluorescence analysis.

8. Economic Aspects [2], [9], [11]

The amounts of tin mined, smelted, and consumed in the years 1980 – 1990 are given in Table 5.

Even in 1938, smelting production amounted to 171 200 t [3], and this had increased by only 66 500 t or 38.8 % in 1990. This is an average increase of 9.75 % per annum.

Table 5. World production and consumption of tin (in 10^3 t)

Year	Mining output	Smelting output	Consumption
1980	235.5	243.6	221.4
1985	185.3	216.2	214.3
1990	210.7	225.6	232.7

Typically in the production and consumption of tin, the production processes of mining and smelting are geographically remote from the places where the metal is consumed.

The main producers, i.e., Brazil, Indonesia, Malaysia, Bolivia, and to some extent China, consume very little tin themselves (see Table 6). On the other hand, the United States, Japan, Germany, and the United Kingdom produce little or no primary tin themselves and are the main tin consumers. In the CIS, production and consumption are approximately equal.

Table 6. Mining output and consumption of tin in various countries 1990 [9]

Country	Mining output		Consumption	
	in 10^3 t	as %	in 10^3 t	as %
Brazil	39.1	18.6	6.1	2.6
Indonesia	37.1	15.1	1.4	0.6
Malaysia	28.5	14.3	3.3	1.4
Bolivia	17.3	8.2	0.4	0.2
China	35.8	17.0	18.0	7.7
Russia (CIS)	13.0	6.1	20.0	8.5
United States	0.1	0.04	37.3	16.0
Japan			33.8	14.5
Germany			19.3	8.2
United Kingdom	3.4	1.6	10.4	4.3

The continuous fall in the price of tin (> 30 000 DM/t in 1985, ca. 10 000 DM/t in 1990) has led to many casualties worldwide. The worst effects have been in Malaysia, where over 80 suppliers have been closed down. In 1980, Malaysia was the main producer of tin, with a mine output of 61 000 t. This fell to 28 500 t in 1990, and Malaysia now produces less tin than Brazil, China, or Indonesia. Even the well-known tin smelter Capper Pass in the United Kingdom has ceased production, as have the tin mines in the Altenberg region of Saxony.

9. Tin Alloys and Coatings [1], [2], [11]

Tin is one of the most important constituents of low-melting nonferrous alloys. The following important properties of the metal are exploited

1) Low melting point
2) Low hardness
3) Good wetting properties
4) Effective incorporation of foreign particles
5) Good compatibility with foodstuffs

Tin utensils have been produced for over 5000 years, and well-preserved examples exist from the various epochs.

Tin utensils are always made of tin alloys, as the pure metal is too soft. The most important alloying elements are antimony, copper, and lead. The tin is usually melted first, and it then readily forms alloys with pure metals, which have good solubility properties. To minimize burning, especially of antimony, master alloys are usually used.

As tin has good flow and casting properties, casting is the most important method of producing tin articles. All the processes used in a modern foundry are suitable. Those most often used are sand, shell, centrifugal, and pressure die casting. The casting methods are sometimes automated, whereby the low casting temperatures make small demands on the material used for the shells and molds.

As very thin-walled cast tin cannot always be produced, pressure forming methods are now also widely used. For example, a circular sheet can be spun over a former. This method is especially suitable for simple rotationally symmetrical items without asymmetrical surface effects.

Owing to the good forming properties of tin, other methods can also be used. For example, stamping and extrusion present no special problems, and individual items can be produced by forging and hammering processes.

The composition of the metal used to produce tin articles is specified in DIN 17 810 and the materials regulation RAL-RG 683. The DIN 17 810 specification is as follows:

Grade	Sn 90-10
Material no.	2.3710
Tin	min. 90 %
Antimony	$\leq$ 7 %
Copper	$\leq$ 3 %
Silver	$\leq$ 4 %
Lead	max. 0.5 %
Sum of others	0.3 %

The lead limit of 0.5 % is imposed merely because this is technically feasible. It has been

shown that a lead content of 2 % leads to insignificant releases of lead even after utensils have been kept at unusually high temperatures and for unusually long periods of time.

Solders. Most solders are based on the tin – lead binary system (→ Lead Alloys, Chap. 5.), which has a eutectic at ca. 63 % tin and 183 °C. The solid solubility of 1 – 2 % lead in tin and 13 % tin in lead is not relevant to production conditions.

In solder applications, it is of great importance to know what percentage of impurities can cause problems, and, conversely, whether alloying elements can have a detrimental effect on soldered joints under certain conditions. This question is extremely important in the electronics industry because of the small amounts of solder used in a soldered joint, and the small distances between the soldered joints. The state of knowledge is as follows [11, Chap. 5.8]:

Zinc:	Visible impairment of the surface of the solder by oxide formation at 0.005 %. Recommended limit: 0.001 %.
Aluminum:	Impairment of adhesive bond, hot brittleness, and dull appearance at 0.005 %. Recommended limit: 0.001 %.
Phosphorus:	Increased oxidation at 0.001 %. Lower concentrations reduce oxidation in unstirred baths.

It is known that other elements, such as arsenic and sulfur, can have detrimental effects, but precise quantitative experimental results are not available.

Low friction materials and bronzes have the following useful properties:

1) High mechanical strength with good electrical conductivity
2) Good soldering properties
3) Extremely good properties as a bearing (antifriction) metal
4) Good machinability at room temperature
5) Good general corrosion resistance towards the atmosphere and seawater, and, in the case of zinc-free alloys, towards stress corrosion

Apart from cast articles, the most important forms are wire, rolled profile, sheet, and strip. The material is also used in bearings and in the chemical industry. Its use in domestic items, e.g., fittings and mountings, is also considerable. The alloys are classified as casting alloys and wrought alloys, the latter having lower tin contents. Some important copper – tin alloys, as specified in DIN 1705/1716, are listed in the following:

Gunmetal	
G-Cu Sn 12	88 % Cu, 12 % Sn
G-Cu Sn 12 Ni	86 % Cu, 12 % Sn, 2 % Ni
G-Cu Sn 12 Pb	86 % Cu, 12 % Sn, 2 % Pb
Red bronze	
G-Cu Sn 10 Zn	88 % Cu, 10 % Sn, 2 % Zn
G-Cu Sn 7 Zn Pb	83 % Cu, 7 % Sn, 4 % Zn, 6 % Pb
Leaded bronze	
G-Cu Pb 5 Sn	85 % Cu, 5 % Pb, 10 % Sn
G-Cu Pb 15 Sn	77 % Cu, 15 % Pb, 8 % Sn
G-Cu Pb 22 Sn	76 % Cu, 22 % Pb, 2 % Sn
Wrought alloys	
Cu Sn 2	98 % Cu, 2 % Sn
Cu Sn 6	94 % Cu, 6 % Sn
Cu Sn 6 Zn	88 % Cu, 6 % Sn, 6 % Zn

Although the copper – tin alloys are some of the oldest materials used by humans, their development is not yet exhausted even today.

The most important development aims are improvement of mechanical properties and corrosion resistance, and reduction of the tin content.

Sintering Metallurgy of Bronzes. An interesting new use for tin is as an addition in powder form when sintering bronze. Especially when this is for use as a bearing metal, economic advantages are obtained by the addition of 4 % tin to the copper powder with or without lead addition.

Low-melting alloys are of great importance in several technical applications. Their melting points usually lie significantly below 150 °C. Bismuth is always an essential alloy constituent.

Their most important application areas are in mold making, safety systems for the prevention of fire and overheating, and stepwise soldering.

Melting points, compositions, and areas of use of typical low-melting alloys are listed in Table 7.

Amalgams. Tin has been used for dental fillings since the Middle Ages, and amalgams since the 1800s. Subsequent developments have led to the silver – tin amalgams used today (→ Dental Materials, Chap. 2.3.2.). A typical amalgam has

Table 7. Melting points, compositions, and applications of typical low-melting alloys

Melting point, °C	Typical composition, wt %					Applications
	Bi	Sn	Pb	In	Cd	
47	44.7	8.3	22.6	19.1	5.3	test castings, fixing lenses
58	49.0	12.0	18.0	21.0		fixing components
70	49.4	12.9	27.7		10.0	supporting components
70 – 73	50.0	13.3	26.7		10.0	tube bending, radiation protection, fixing components
96	52.0	16.0	32.0			temperature fuses
138 – 170	40.0	60.0				accurate molds, cores, gravity die casting

the following composition: 52 % Hg, 33 % Ag, 12.5 % Sn, 2 % Cu, 0.5 % Zn.

When the mixture of metals is ground together, the following hardening reaction takes place:

$$AgSn_3 + 4\,Hg \longrightarrow Ag_2Hg_3 + Sn_{7-8}Hg + Ag_3Sn$$

The metallographic structure of a dental amalgam consists of islands of solid undissolved particles of alloy (Ag_3Sn) in a soft matrix of silver and tin amalgams.

Alloy Coatings. Tin alloys are important in the production of coatings by electroplating and hot tinning. The most important of these are tin – zinc, tin – nickel, tin – cobalt, and tin – copper.

Tin – lead coatings are mainly used for corrosion protection and as a preparation for soldering. Electrolytically applied coatings must be treated with hot palm oil or by infrared heating. This melts the coating and can prevent the formation of whiskers.

Tin – zinc coatings are increasingly replacing the toxic cadmium coatings.

Tin – nickel and tin – cobalt coatings are mainly used in electrical installations, e.g., to produce electrical connectors.

Additions to Alloys. Tin is increasingly used as an alloy addition in the steel industry. The addition of 0.1 – 0.5 % tin causes cast iron to solidify with a pure pearlitic structure, making it uniformly hard and wear-resistant.

Sinter Metallurgy. In the sinter metallurgy of iron, addition of 2.5 – 5 % of a tin – copper alloy (2 : 3) gives reduced sintering temperatures and times, and in particular improves the dimensional accuracy of the sintered components.

10. Inorganic Tin Compounds

Consumption of inorganic tin compounds is lower than that of organic tin compounds, but they are often the starting materials used to produce the organic tin compounds.

In minerals, tin is nearly always tetravalent and bonded to oxygen or sulfur. The only exception is herzenbergite, SnS, in which it is divalent.

In industry, tin(II) and tin(IV) compounds are produced from metallic tin. Many tin(II) compounds which are sufficiently stable for practical purposes have a strong tendency to change into tin(IV) compounds and are therefore strongly reducing. For example, $SnCl_2$ precipitates gold and silver in metallic form from their salt solutions.

The salts in both valency states hydrolyze in aqueous solution to form insoluble salts. In alkaline media, stannites (divalent) and stannates (tetravalent) are formed.

10.1. Tin(II) Compounds

Tin(II) chloride, $SnCl_2$, is the most important inorganic tin(II) compound. It is produced on an industrial scale by reducing tin(IV) chloride with molten tin, or by direct chlorination of tin.

Solutions of tin(II) chloride are obtained by dissolving metallic tin in hydrochloric acid, or by reducing a solution of $SnCl_4$ with metallic tin.

The anhydrous substance is white, has a greasy luster, and dissolves readily in water, alcohol, ethyl acetate, acetone, and ether. The clear, nondeliquescent, monoclinic dihydrate, $SnCl_2 \cdot 2\,H_2O$, crystallizes from aqueous solution and is the commercial product.

On dilution, the aqueous solution becomes cloudy as hydrolysis causes precipitation of the basic salt:

$$SnCl_2 + H_2O \rightleftharpoons Sn(OH)Cl + HCl$$

The cloudiness can be prevented by small additions of hydrochloric acid, tartaric acid, or ammonium chloride. Because of its strong tendency to hydrolyze, the dihydrate can only be dehydrated over concentrated sulfuric acid or by heating in a stream of hydrogen chloride.

Tin(II) chloride is an important industrial reducing agent, being used to reduce aromatic nitro compounds to amines, aliphatic nitro compounds to oximes and hydroxylamines, and nitriles to aldehydes.

Tin electroplating can be carried out in a fused eutectic salt mixture of 20 % $SnCl_2$ and 80 % KCl at 200 – 400 °C [90], [91].

Tin(II) Oxide Hydrate and Tin(II) Oxide. If aqueous solutions of $SnCl_2$ or other tin(II) salts are reacted with alkali metal carbonate or ammonia, an amorphous white precipitate of tin(II) oxide hydrate, $5\,SnO \cdot 2\,H_2O$, is obtained [92]. $Sn(OH)_2$ does not exist. Tin(II) oxide hydrate is amphoteric. Dehydration in a stream of carbon dioxide gives tin(II) oxide. Tin(II) oxide hydrate and tin(II) oxide are starting materials for the production of other tin(II) compounds.

Other Tin(II) Compounds.

Tin(II) fluoride, SnF_2, is formed from tin(II) oxide hydrate and hydrofluoric acid, and is added to toothpastes as an anticaries agent.

Tin(II) fluoroborate hydrate, $Sn(BF_4)_2 \cdot n\,H_2O$, is formed by dissolving the oxide hydrate or the oxide in aqueous fluoroboric acid. Sulfuric acid reacts with the oxide hydrate or the oxide to form tin(II) sulfate. Both tin(II) sulfate and tin(II) fluoroborate are important in the production of metallic tin coatings by electroplating.

Tin(II) bromide, $SnBr_2$, and *tin(II) iodide*, SnI_2, are produced by reacting metallic tin with the appropriate hydrogen halide.

Tin(II) cyanide, which is produced from bis-(cyclopentadienyl)tin(II) and hydrogen cyanide, is the only known compound of tin with inorganic carbon:

$$(C_2H_5)_2Sn + 2\,HCN \longrightarrow Sn(CN)_2 + 2\,C_2H_6$$

The tin(II) salt of ethylhexanoic acid is an effective catalyst in polyurethane production.

10.2. Tin(IV) Compounds

Tin(IV) Hydride. The toxic, colorless, flammable gas, tin(IV) hydride, is formed by the reduction of tin(IV) chloride by $LiAlH_4$ in diethyl ether at −30 °C. It is stable for several days at room temperature, and decomposes into its elements at 150 °C in the absence of air, forming a tin mirror.

Tin(IV) Halides and Halostannates(IV). Anhydrous tin(IV) chloride is a colorless liquid which fumes strongly in air. It is a good solvent for sulfur, phosphorus, and iodine, and is miscible in all proportions with carbon disulfide, alcohol, benzene, and other organic solvents. It hydrolyzes in water, evolving much heat and forming colloidal tin(IV) oxide and hydrochloric acid:

$$SnCl_4 + 2\,H_2O \longrightarrow SnO_2 + 4\,HCl$$

In moist air, the pentahydrate, $SnCl_4 \cdot 5\,H_2O$, is formed, the so-called butter of tin, a white deliquescent crystalline mass with a melting point of 60 °C. In industry, $SnCl_4$ is produced by the reaction of chlorine with tin. The anhydrous product is obtained if the metal is covered with $SnCl_4$. Anhydrous tin(IV) chloride is an important starting material for the production of organic tin compounds (see Sections 11.2 and 11.3).

Tin(IV) bromide, $SnBr_4$, and tin(IV) iodide, SnI_4, are also obtained by the reaction of metallic tin with the halogens. Tin(IV) fluoride is produced by the reaction of tin(IV) chloride with anhydrous hydrogen fluoride:

$$SnCl_4 + 4\,HF \longrightarrow SnF_4 + 4\,HCl$$

Tin(IV) halides react readily with metal halides to form the halostannates(IV), the coordination number of the tin increasing from four to six. The reaction proceeds as follows (X = halogen):

$$SnX_4 + 2\,MX \rightleftharpoons M_2SnX_6$$

One of the best known compounds of this type is ammonium hexachlorostannate, $(NH_4)_2SnCl_6$, the so-called pink salt.

Hexachlorostannic acid, $H_2SnCl_6 \cdot 6\,H_2O$, is formed by passing HCl into a concentrated solution of $SnCl_4$.

Tin(IV) Oxide, Tin(IV) Oxide Hydrate, and Stannates(IV). Tin(IV) oxide decomposes above 1500 °C without melting to form tin(II) oxide:

$$2\,SnO_2 \longrightarrow 2\,SnO + O_2$$

Pure tin(IV) oxide can be obtained by the combustion of powdered tin or sprayed molten tin in a hot stream of air. It is insoluble in acids and alkalis. Specially prepared tin(IV) oxide has many uses, total world consumption of this material being > 4000 t/a. It is used in combination with other pigments to produce ceramic colorants. It has a rutile structure, and hence can absorb the colored ions of transition metals. The products obtained form the basis of ceramic colors, and include tin vanadium yellow, tin antimony gray, and chrome tin pink. The thermal stability of the tin colors enables them to be used both in and under the glaze.

Electrodes made of SnO_2 are used in the production of lead glass [93]. They are resistant to corrosion by molten glass, and have good electrical conductivity when hot.

Coatings of tin(IV) oxide treated with indium oxide (<100 µm thick) give good transparency properties to aircraft window systems, increase their strength, and give protection from icing.

If tin(IV) oxide is reacted with a solution of alkali, or a solution of stannate is reacted with acid, a white, gel-like precipitate of tin(IV) oxide hydrates is formed which are very soluble in alkalis and acids. This precipitate was formerly known as "α-stannic acid", and the aged product, which is sparingly soluble, as "β-stannic acid" (metastannic acid). Today, this product is regarded as tin(IV) oxide hydrate with the formula $SnO_2 \cdot n\,H_2O$, where n decreases with aging.

The reaction of powdered or granulated tin with concentrated nitric acid leads to the formation of the reactive β-tin(IV) oxide hydrate. This can be used as a catalyst for aromatics.

The β-tin oxide hydrate gels precipitated from $SnCl_4$ by ammonia and then dried are stable towards nuclear radiation, and can be used in chromatographic columns for separating radioactive isotopes [94].

Fusion of tin(IV) oxide with alkali metal hydroxides leads to formation of alkali metal hexahydroxystannates according to the following reaction scheme (M = alkali metal):

$$SnO_2 + 2\,MOH + 2\,H_2 \longrightarrow M_2\,[Sn(OH)_6]$$

The sodium and potassium complexes are used as alkaline electrolytes in electrolytic tin plating.

Tin(IV) sulfide, SnS_2, is formed as golden yellow flakes with a hexagonal crystal structure by passing hydrogen sulfide through a weakly acidic solution of a tin(IV) salt. On heating, the crystals become dark red to almost black, reverting to yellow on cooling.

The tin disulfide known as "mosaic gold" is produced industrially by heating tin amalgam with sulfur and ammonium chloride. It is used for gilding picture frames, and in painting to produce deep yellow to bronze color shades.

11. Organic Compounds of Tin [3], [95], [96]

The organic chemistry of tin has attracted major interest since 1945. The first organotin compounds were prepared in 1849 by Frankland and in 1852 by Löwig [97]. In the first technical application in 1936, the discovery of the stabilizing effect of these compounds on poly(vinyl chloride) was utilized. The biocidal properties of other organotin compounds have been known since ca. 1950.

Only the organic compounds of tetravalent tin are used in industry.

Most organotin molecules contain a single tin atom with 4 substituents. They are classified according to the number of direct tin-carbon bonds: R_4Sn, R_3SnX, R_2SnX_2, $RSnX_3$.

Here, R denotes any hydrocarbon group and X denotes a group such as a halide, OH, OR, SR, acid group, etc. Most of these compounds are colorless liquids at room temperature or slightly above. They are very soluble in organic solvents but sparingly soluble in water.

Organotin compounds have a very wide range of applications, depending on their type, including the stabilization of PVC, catalysis, crop protection, and wood preservation. Also, for some time, various organotin compounds have been increasingly used as laboratory chemicals, especially organotin alkoxides and hydrides. These are used as synthesis auxiliaries and as mild and selective reducing agents.

Organotin compounds do not occur in nature.

11.1. Properties of Organotin Compounds

Under the influence of light, atmospheric oxygen, or certain microorganisms, organotin compounds are degraded in a relatively short time, the hydrocarbon groups being split off to leave behind nontoxic inorganic products. Although both the tin – carbon bond (average dissociation energy 209 kJ/mol) and the tin –oxygen bond (average dissociation energy 318 kJ/mol) are reactive, they are sufficiently stable for general handling purposes.

The symmetrical tetraalkyltin compounds have a very slight odor. They are colorless, form monomolecular solutions, are fairly stable towards water and air, and can be distilled without decomposition at $< 200\,^{\circ}C$. Their solubilities and boiling points are similar to those of the branched chain paraffins with similar molecular mass; the higher homologues are waxy substances.

The symmetrical tetraaryltin compounds are stable towards air and water and are also colorless. They melt at temperatures above 150 °C.

The organotin hydrides, with the exception of some aryl tin hydrides which are solid at room temperature, are colorless, nonassociated liquids which are rapidly attacked by oxygen and therefore can only be prepared and stored under inert gas. They are important reducing agents.

The organotin fluorides, the diorganotin dihalides, and the aromatic organotin monohalides are solids at room temperature, while the aliphatic organotin monohalides and trihalides are liquids. The higher triorganotin derivatives have a broad biocidal effect on microorganisms such as fungi, bacteria, and harmful waterborne organisms such as algae, tube worms, shellfish, etc. The most active compounds are the tributyl-, trichlorohexyl-, and triphenyltin compounds.

The di- and monoorganotin derivatives in which methyl, butyl, and octyl groups are bonded to the tin stabilize polymers sensitive to light and temperature such as PVC if the tin is bonded via oxygen or sulfur to certain other groups.

11.2. Production of Organotin Compounds

Tetraorganotin Compounds from Tin Tetrachloride and Organometallic Compounds. Tin tetrachloride is the key substance for the production of organotin(IV) compounds. In industry, the tin tetrachloride is first alkylated with organic compounds of magnesium, aluminum, or sodium to form tetraorganotin compounds. The process is usually continuous.

Organotin Chlorides from Tetraorganotin Compounds and Tin Tetrachloride. If tetraorganotin compounds are reacted with stoichiometric amounts of tin tetrachloride, the corresponding organotin chlorides are obtained:

$$SnR_4 + 1/3\,SnCl_4 \rightleftharpoons 4/3\,R_3SnCl$$

$$SnR_4 + SnCl_4 \rightleftharpoons 2\,R_2SnCl_2$$

$$SnR_4 + 3\,SnCl_4 \rightleftharpoons 4\,RSnCl_3$$

The production of triorganotin chlorides and diorganotin dichlorides proceeds smoothly according to this reaction. Monoorganotin trichlorides can only be obtained in a few cases where R = acryl or vinyl and in special solvents or with catalysts.

Direct Synthesis of Organotin Chlorides from Tin. Organotin chlorides can be obtained by direct reaction of tin with unsaturated organic compounds and hydrogen halides (ester tin process) or from organic halides (catalytic direct process).

Ester Tin Process for the Production of Tin Carboxylic Acid Derivatives from Tin. Unsaturated organic compounds, such as esters of substituted or unsubstituted acrylic acid, acrylonitrile, or vinylphosphoric diesters, react with metallic tin and hydrogen chloride in a polar medium (ethanol, concentrated hydrochloric acid, or diethyl ether) to form the so-called ester tin compounds, e.g., bis(2-methoxycarbonylethyl)tin dichloride [98].

Catalytic Production of Organotin Halides from Tin. Organotin halides can be produced discontinuously or continuously from metallic tin and organic halides with the aid of catalysts at elevated temperatures. The most effective of these catalysts are the tetraalkylammonium halides, tetraalkylphosphonium halides, and other derivatives of N, P, As, or Sb.

Production of Organotin Trichlorides by Addition of Tin Dichloride and Hydrogen Chloride. The unsaturated starting compounds suitable for the ester tin process, when reacted with tin(II) chloride and hydrogen chloride, form the corresponding organotin trichlorides, e.g., (2-methoxycarbonylethyl)tin trichloride [99].

Production of Unsymmetrical Organotin Compounds by Addition of Organotin Hydrides (Hydrostannation). Unsymmetrical tetraorganotin compounds that may contain functional or unsaturated groups in the alkyl groups can be synthesized by addition of organotin hydrides to alkenes and alkynes. The reaction is favored by radical formers and UV light.

Conversion of Organotin Chlorides to other Halides or Derivatives with Heteroatoms via Organotin Oxides. The organotin chlorides are quantitatively converted into organotin oxides in alkaline medium:

$$2\,R_3SnCl \longrightarrow (R_3Sn)_2O$$

$$R_2SnCl_2 \longrightarrow 1/n\,(R_2SnO)_n$$

$$RSnCl_3 \longrightarrow 1/n\,(RSnO_{1.5})_n$$

In acid media, many organotin compounds, e.g., fluorides, bromides, iodides, pseudohalides, carboxylates, thiolates, etc., are obtained from the organotin compounds:

$$(R_3Sn)_2 \longrightarrow 2\,R_3SnX$$

$$1/n\,(R_2SnO)_n \longrightarrow R_2SnX_2$$

$$1/n\,(RSnO_{1.5})_n \longrightarrow RSnX_3$$

X = halogen or other functional group

These transformation reactions are of industrial importance, as they enable the sensitive organotin compounds to be converted into organotin oxides that are insoluble in water, stable to air, easily transported, and storable for long periods. These can then be used to produce either the original organotin oxides or other organotin compounds when required.

11.3. Industrially Important Compounds

A few typical examples of the many applications of organotin compounds are described here.

Methyl Compounds. A legally permitted stabilizer for PVC used with foodstuffs is produced from dimethyltin dichloride and methyltin trichloride. It consists of:

$(CH_3)_2Sn[SCH_2COOC_5H_{10}CH(CH_3)_2]_2$	75 %
$CH_3Sn[SCH_2COOC_5H_{10}CH(CH_3)_2]_3$	25 %

Butyl Compounds. Tetrabutyltin, $(C_4H_9)_4Sn$, can be converted to tributyltin chloride, $(C_4H_9)_3SnCl$, and this can be converted to bis(tributyltin) oxide, $(C_4H_9)_3SnOSn(C_4H_9)_3$, by treatment with NaOH. This product, which is insoluble in water and miscible with industrial solvents, is moderately toxic. It is an active biocide with many uses. For example, it is used as an antifouling paint for ships, for the prevention of slimes in industrial recirculating water systems, for combating freshwater snails that cause bilharzia, as a wood and textile preservative, and as a disinfectant.

The readily interconvertible compounds dibutyltin chloride $(C_4H_9)_2SnCl_2$ and dibutyltin oxide $[(C_4H_9)_2SnO]_n$ (a polymeric powder) are the starting substances for the production of the most common PVC stabilizers, i.e., the liquid dibutyltin dilaurate $(C_4H_9)_2Sn(OCOC_{11}H_{23})_2$, the polymeric solid dibutyltin maleate $[(C_4H_9)_2SnOCOCH{=}CHCOO]_n$, the liquid dibutyltin bis(isooctylmaleate) $(C_4H_9)_2Sn[OCOCH{=}CHCOOC_5H_{10}CH(CH)_3]_2$, and the liquid dibutyltin bis(thioacetic acid isooctyl ester) $(C_4H_9)_2Sn[SCH_2COOC_5H_{10}CH(CH_3)_2]_2$.

Octyl Compounds. The stabilizers used in the production of PVC film for the foodstuffs industry, i.e., the liquid octyltin(thio)acetic acid isooctyl esters with the formulas $(n\text{-}C_8H_{17})_2Sn[SCH_2COOC_5H_{10}CH(CH_3)_2]_2$ and $n\text{-}C_8H_{17}Sn[SCH_2COOC_5H_{10}CH(CH_3)_2]_3$ are produced from isooctyl mercaptoacetate and dioctyltin dichloride $(C_8H_{17})_2SnCl_2$ or octyltin trichloride $C_8H_{17}SnCl_3$, respectively.

Cyclohexyl Compounds. Tricyclohexyltin hydroxide $(cyclo\text{-}C_6H_{11})_3SnOH$ is obtained by alkaline hydrolysis of tricyclohexyltin chloride $(cyclo\text{-}C_6H_{11})_3SnCl$. It is a colorless, crystalline substance, insoluble in water, and has a very marked acaricidal action, attacking many mites and acarides, thereby protecting plants and useful insects. It is the main component of the

commercial product Plictran (Dow Chemical), which is used in fruit growing, viticulture, and greenhouses.

Phenyl Compounds. The phenyl derivatives of tin are used as fungicides, e.g., for the treatment of potato rot and leaf spot in root tubers. Typical compounds are triphenyltin hydroxide, $(C_6H_5)_3SnOH$ (Du Ter, Philips-Duphar), and triphenyltin acetate, $(C_6H_5)_3SnOCOCH_3$ (Brestan, Hoechst).

11.4. Analysis of Organotin Compounds

The analysis of organotin compounds is a complex field. Methods used are described in detail in [100].

In the determination of individual organotin compounds by gas chromatography, it is first necessary to convert the organotin halides or oxides into unsymmetrical tetraorganotin compounds by methylation or butylation with Grignard reagents [100].

11.5. Storage and Shipping of Organotin Compounds

Organotin compounds are stored and transported in drums and vessels of steel sheet coated on the inside with a special paint. In the case of solids, the emission of hazardous dust can be prevented by incorporating these products in paste formulations.

11.6. Pattern of Production and Consumption

World production of organotin compounds was ca. 50 t/a in 1950, 35 000 t/a in 1981, and 40 000 t/a in the mid-1990s. The tin content of these materials is ca. 25 %.

In the main producing and consuming areas – United States, Western Europe, and Japan – 76 % of the organotin compounds are used as stabilizers for PVC, 10 % as antifouling biocides, 8 % as agricultural biocides, and 5 % as catalysts for the production of polyurethanes and silicones.

The main producers of organotin compounds are:

United States:	M & T Chemicals, Thiokol-Charstal, and Interstab (Akzo)
Japan	Hokko Chemical Industries, Yoshitomi Pharmaceutical Industries, Nitto Kasei, and Somkyo Organic Chemicals
Europe	Schering, Akzo Chemie, and Ciba-Geigy

12. Toxicology

Metallic tin is generally considered to be nontoxic. As early as the Middle Ages, it was used in the form of plates, jugs, and drinking vessels. Even large amounts of tin salts in the digestive system cause negligible harm. Apparently, tin can migrate only very slowly through the intestinal walls into the blood. Orally ingested tin is poorly resorbed by animals and humans. The half-life in the kidneys and liver is 10 – 20 d, and in the bones 40 – 100 d. Cases of poisoning are almost unknown. Massive inhalation of tin or tin oxide dust by exposed industrial workers can lead to irritation of the respiratory tract. In extreme cases, metal fume fever with similar symptoms to those of "zinc fever" or "brass fever" also occurs. As the peroral ingestion of tin and its inorganic compounds is comparatively harmless because of the relatively low resorption, the limit value for the prevention of sickness and diarrhea for fruit conserves is ca. 250 – 500 mg/kg, and for fruit juices 500 – 1000 mg/kg. In current industrial practice, the tin cans are additionally lacquered, and the tin concentrations in preserved fruit and vegetables are < 250 ppm. The MAK value, calculated as tin and measured in total dust, has been 2 mg/m^3 for many years [101], [102].

Hydrochloric acid formed by the hydrolysis of tin chloride can cause acid burns. Tin hydride is very toxic, having a similar effect on the human organism to arsine [103].

Organotin compounds are very toxic; the MAK value is 0.1 mg/m^3 [102]. These compounds can differ widely in their effects, and can also be slowly converted to other compounds in the organism, so that toxic symptoms can change during the induction time. The toxicity of alkyl and aryl tin compounds decreases in the se-

ries: trialkyl > dialkyl > tetraalkyl > monoalkyl compounds. Whereas tributyl- and triphenyltin compounds are almost as toxic as HCN, the monoalkyl compounds have a toxicity similar to that of the inorganic tin compounds.

The resorption of tin alkyl and tin aryl compounds can have harmful effects on the CNS, such as edema of the brain and spinal cord and damage to the respiratory center. The volatile organotin compounds cause persistent headaches, epileptiform convulsions, narcosis, and respiratory paralysis.

Dibutyltin dichloride, tributyltin chloride, and analogous alkyl tin halides, after a latent period, cause irritation and burning of the skin and especially the mucous membranes, and are intensely sternutatory and lachrymatory [104], [105].

In cases of poisoning by organotin compounds, long-term observation of the functioning of the liver and CNS is necessary.

13. References

General References

1. *Ullmann,* 3rd ed., **19,** 145 – 170.
2. *Ullmann,* 4th ed., **24,** 641 – 679.
3. V. Tafel: *Lehrbuch der Metallhüttenkunde,* 2nd ed., **vol. 2,** Hirzel-Verlag, Leipzig 1953, pp. 221 – 305.
4. *Gmelin,* Teil B, pp. 205 – 214, 347 – 355.
5. F. Pawlek: *Metallhüttenkunde,* De Gruyter, Berlin –New York 1983, pp. 482 – 500.
6. *Meyers Enzyklopädisches Lexikon,* 9th ed., **vol. 25,** Bibliographisches Institut, Mannheim 1979, p. 730.
7. F. Ahlfeld: *Zinn und Wolfram. Die metallischen Rohstoffe,* vol. 11, Enke Verlag, Stuttgart 1958.
8. *Landolt-Börnstein,* New Series, Springer Verlag, Berlin 1971.
9. Metallgesellschaft AG, *Metallstatistik,* **vol. 78,** Frankfurt, (1980 – 1990).
10. *Brockhaus Taschenbuch der Geologie,* Brockhaus Verlag, Leipzig 1955.
11. *Zinn-Taschenbuch,* 2nd ed., Metall-Verlag, Berlin 1981.
12. A. Leube: "Zinn," in: *Lehrbuch der angewandten Geologie,* vol. 2, part 1, Enke Verlag, Stuttgart 1968.
13. W. Gocht: *Handbuch der Metallmärkte,* Springer Verlag, Berlin – Heidelberg – New York 1974.
14. S. Jankovic: *Wirtschaftsgeologie der Erze,* Springer Verlag, Wien – New York 1974.

Specific References

15. P. Klemm: *Der Weg aus der Wildnis,* vol. 3, KB Verlag, Berlin 1962.
16. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* 3rd ed., Springer Verlag, Berlin 1964.
17. I. Barin et al.: *Thermochemical Properties of Inorganic Substances,* Suppl., Springer Verlag, Berlin – Heidelberg – New York 1977.
18. R. Zimmermann, K. Günther: *Metallurgie und Werkstofftechnik,* **vol. 1,** Deutscher Verlag für Grundstoffindustrie, Leipzig 1977, p. 635.
19. A. Eiling, J. S. Schilling, *J. Phys. F* **11** (1981) 623.
20. Chemical Rubber Comp.: *Handbook of Chemistry and Physics,* 62nd ed., CRC Press, Boca Raton 1981.
21. G. V. Raynow, R. W. Smith, *Proc. R. Soc. London Ser.* **A244** (1958), Nov. 27, 101.
22. W. Köster, *Z. Metallk.* **39** (1948) 1.
23. E. G. Shidkovskii, A. A. Durgartan, *Nauchn. Dokl. Vyssh. Shk. Fiz. Mat. Nauki* 1958, no. 2, 211.
24. H. Gundlach, W. Thormann, *Z. Dtsch. Geol. Ges.* **112** (1960) no. 1, 1.
25. H. Breuer, *Glückauf Forschungsh.* **42** (1981) 213.
26. H. Breuer, A. Guzman, *Erzmetall* **32** (1979) 379.
27. *World Min.* **31** (1978) no. 5, 50.
28. E. Müller, *Erzmetall* **30** (1977) 54.
29. P. A. Wright: *Extractive Metallurgy of Tin,* Elsevier, Amsterdam 1966.
30. J. M. Floyd, *Proc. Symp. Lead-Zinc-Tin '80,* AIME, Materials Park, Ohio 1980.
31. I. J. Bear, R. J. T. Canay, *Trans. Inst. Min. Metall. Sect. C* **85** (1976) no. 9, 139.
32. N. N. Murach et al.: *Metallurgy of Tin,* vol. 2, National Lending Library for Science and Technology, Boston, Spa, England 1967.
33. E. M. Llevin et al.: "Phase Diagrams for Ceramists," Am. Ceram. Soc., 1974.
34. R. Zimmerman, K. Günther: *Metallurgie und Werkstofftechnik,* **vol. 2,** Deutscher Verlag für Grundstoffindustrie, Leipzig 1977, p. 116.
35. P. Paschen, *Erzmetall* **27** (1974) 28.
36. W. van Rijswijk de Jong, *Erzmetall* **27** (1974) 22.
37. J. Barthel, *Freiberg. Forschungsh. B* **B 112** (1971) 13.
38. K. Leipner, *Neue Hütte* **16** (1971) 395.

39. L. Sivila, *Neue Hütte* **18** (1973) 395.
40. W. T. Denholm, *Proc. Symp. Extr. Met.,* 1981.
41. K. A. Foo, J. M. Floyd, *Proc. Symp. Lead-Zinc-Tin '80,* AIME, Materials Park, Ohio 1980.
42. V. I. Evdokimov, *Freiberg. Forschungsh. B* **B113** (1976) 6.
43. E. B. Kingl, L. W. Pommier, *5th Int. Conf. Tin,* La Paz 1977.
44. G. Severin et al., *Neue Hütte* **20** (1975) 752.
45. A. I. Evdokimov, I. S. Morosov, *Chim. Ind. (Paris)* **88** (1962) no. 2, 115.
46. A. Lange, *Erzmetall* **15** (1962) 113.
47. H. Weigel, D. Zetsche, *Erzmetall* **27** (1974) 237.
48. Y. P. Zvonkov, A. A. Rozlovskii, *Tsvetn. Met. (N.Y.)* **10** (1969) no. 7, 35.
49. G. Daradimos, U. Kuxmann, *Erzmetall* **24** (1971) 171.
50. A. W. Fletscher et al., *Trans. Inst. Min. Metall. Sect. C* **76** (1967) 145.
51. V. E. Dyakov, N. S. Saturin, *Tsvetn. Met. (N.Y.)* **9** (1968) no. 9, 55.
52. L. W. Pommier, S. J. Escalera, *J. Met.* **31** (1979) no. 4, 10.
53. H. Winterhager, U. Kleinert, *Erzmetall* **22** (1969) 310.
54. G. I. Karaivko et al., *Orig. Izd. Nauka*, Moskau, engl. trans., Technikopy Ltd. London 1977.
55. G. Holt, D. Pearson, *Trans. Inst. Min. Metall. Sect. C* **86** (1977) 77.
56. T. R. A. Davey, *5th Int. Tin Conf.,* La Paz 1977.
57. T. R. A. Davey, *Aust. Min.* **61** (1969) Aug., 62.
58. T. R. A. Davey, *Proc. Symp. Lead-Zinc-Tin '80,* AIME, 1980.
59. P. A. Wright: "Advances in Extractive Metallurgy and Refining," *Trans. Inst. Min. Metall.* **24** (1972) 467.
60. T. R. A. Davey, G. M. Willis, *J. Met.* **29** (1977) 24.
61. P. Paschen, *Erzmetall* **29** (1976) 14.
62. T. R. A. Davey, F. J. Flossbach, *J. Met.* **24** (1972) 101.
63. K. Batubara, *5th Int. Tin Conf.,* La Paz 1977.
64. W. Gocht: "Der metallische Rohstoff Zinn," *Volkswirtsch. Schriften,* no. 131, Dunker & Humbold, Berlin 1969.
65. F. K. Oberbeckmann, M. Porten, *Proc. Symp. Lead-Zinc-Tin '80,* AIME, Febr. 1980.
66. C. Ferrante in: *Handbuch der technischen Elektrochemie,* vol. 4: "Die Anwendung des elektrischen Ofens in der metallurgischen Industrie," Akad. Verlagsgesellschaft, Leipzig 1956.
67. O. M. Katkov, *Tsvetn. Met. (N.Y.)* **17** (1976) 38.
68. O. M. Katkov, *Tsvetn. Met. (N.Y.)* **12** (1971) 28.
69. J. E. Joffre, *5th Int. Tin Conf.,* La Paz 1977.
70. M. F. Barett et al., *Trans. Inst. Min. Metall Sect. C* **84** (1975) 231.
71. J. Lema Patino, *4th Int. Tin Conf.,* Kuala Lumpur 1974.
72. B. Möller, *Neue Hütte* **16** (1971) 400.
73. D. N. Klushin, O. V. Nadinskaya, *J. Appl. Chem.* **29** (1956) no. 10, 1593.
74. J. M. Floyd et al., *Aust. Min.* **64** (1972) Aug., 72.
75. R. Kammel, H. Mirafzall, *Erzmetall* **30** (1977) 437.
76. P. A. Wright, *Trans. Inst. Min. Metall Sect. C* **80** (1971) 112.
77. A. P. Samodelov, Y. K. Delimarskii, *Tsvetn. Met. (N.Y.)* **9** (1968) 41.
78. K. Orlich, W. Dittrich, *Freiberg. Forschungsh. B* **B156** (1970) 39.
79. R. P. Elliott: *Constitution of Binary Alloys,* 1st Suppl., Mc Graw-Hill, New York 1965.
80. J. D. Esdaile, *5th Int. Tin Conf.,* La Paz 1977.
81. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* Mc Graw-Hill, New York 1958.
82. S. Y. Shiratshi, H. B. Bell, *Trans. Inst. Min. Metall. Sect. C* **79** (1970) 120.
83. E. Müller, P. Paschen: "Complex Metallurgy '78," *Proc. Inst. Min. Met.* **1978,** 82; *Erzmetall* **31** (1978) 266.
84. P. Paschen, *Metall (Berlin)* **33** (1979) 137.
85. L. Marone, G. Lanfranco, *Metall Ital.* **63** (1971) no. 3, 121.
86. *Winnacker-Küchler,* 3rd ed., **6.**
87. L. Müller-Ohlsen in: *Die Weltwirtschaft im industriellen Entwicklungsprozeß,* J. C. B. Mohr, Tübingen 1981.
88. O. Proske, G. Blumenthal: *Analyse der Metalle,* 3rd ed., **vol. 1,** Springer Verlag, Berlin – Heidelberg – New York 1986.
89. K. Slickers: *Die automatische Atom-Emissions-Spektralanalyse,* 2nd ed., Buchvertrieb K. A. Slickers, Gießen 1992.
90. A. I. Vitkin, K. Delimarskii, *Prot. Met. (Engl. Transl.)* **11** (1975) 245 – 246.
91. T. Williamson, *Zinn Verw.* **124** (1980) 7 – 8.
92. J. D. Donaldson, *Prog. Inorg. Chem.* **8** (1967) 287.
93. C. J. Evans, *Zinn Verw.* **132** (1982) 5 – 8.
94. W. B. Hampshire, C. J. Evans, *Zinn. Verw.* **118** (1979) 3 – 5.
95. *Gmelin,* Tl 1 (Erg.Bd. 26), 1975; Tl 2 (Erg.Bd. 29), 1975; Tl 3 (Erg.Bd. 30), 1976; Tl 4 (Erg.Bd. 35), 1976; Tl 5, 1978; Tl 6, 1979; Part 7, 1980; Part 8, 1981; Part 9, 1982.

96. *Houben-Weyl,* **13/6,** 181 – 521.
97. E. Frankland, *Ann. Chem. Pharm.* **71** (1849) 171, 212.
98. Akzo, DE-OS 2 607 178, 1976.
99. Akzo, DE-OS 2 540 210, 1976.
100. *Houben-Weyl,* **13/6,** 490 – 518.
101. MAK-Werte 1982, Maximale Arbeitsplatzkonzentrationen, DFG, Verlag Chemie, Weinheim.
102. *Gefahrstoffe 1992,* Verlag Universum, Wiesbaden 1992.
103. S. Moschlin: *Klinik und Therapie der Vergiftungen,* 5th ed., Thieme Verlag, Stuttgart 1972.
104. R. Ludewig, K. Lohs: *Akute Vergiftungen,* G. Fischer Verlag, Jena 1988.
105. *Merkblätter für den Umgang mit gefährlichen Stoffen,* Verlag Tribüne, Berlin 1985.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Tires

Standard Article

Dietrich Overhoff[1]

[1]Groß-Umstadt, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.a27_083.pub2
Article Online Posting Date: April 30, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Tire and Vehicle
2. Tire and Road
3. Tire Characteristics
4. Construction
5. Size Designation
6. Load Capacity
7. Standardization
8. Products
8.1. Tires for Two-Wheel Vehicles
8.2. Passenger-Car Tires
8.3. Tires for Commercial Vehicles
9. Economic Aspects
10. Environmental Aspects

[Top of Page]

Tires belong probably to the most complex rubber products. There is hardly any technical rubber article which is so difficult to describe comprehensively in mathematical and physical terms and has such a wide-ranging profile of requirements including:

- Load carrying capacity
- Damping of road irregularities
- Adhesion on the road
- Transfer of longitudinal forces
- Transfer of lateral forces
- Immediate steering response
- Behavior as damped spring
- Dimensional stability
- Structural resistance
- Good mileage
- Low energy consumption
- Low noise generation

A tire functions mainly according to three physical principles [1], [2]:

1. *The spring behavior of the enclosed air:* The air enclosed in the tire acts as a spring and in a first approximation supports any kind of vehicle fitted with pneumatic tires.
2. *The rubber elasticity of the rubber components*: The action of tires is due to their rubber components that are dynamically deformed. In a vehicle, which can be modeled as a system of many components that are able to vibrate, tires act as dampened vibrators across a wide frequency spectrum.
3. *The rubber friction of the tread compound:* Due to the laws of rubber friction, tires are able to build up longitudinal and lateral forces in the contact areas with the road. These forces are necessary to move and to hold a vehicle on a road.

[Top of Page]

1. Tire and Vehicle [3], [4]

A large proportion of the properties of a pneumatic tire can be evaluated only if the tire is considered in combination with a vehicle. An overview on different vehicles is shown in Figure 1.

Figure 1. Applications of tires in different vehicles

The tire is an integrating component of the chassis and determines the quality of a vehicle to a large extent. Since the vehicle dynamics are differently designed from manufacturer to manufacturer, but also due to differences in riding comfort adjustment, the effect of one and the same tire in different types of vehicle – but above all in vehicles from different manufacturers – will be quite different. Furthermore, the evaluation of a tire on a vehicle is relevant only if the second acting partner of the tire, the road, is also taken into consideration.

[Top of Page]

2. Tire and Road [5-9]

Apart from the driver the most important reference element of the vehicle environment is the road surface. Some vehicles have to be used across a wide range of surfaces, for example military vehicles, agricultural or quarry vehicles, and mining industry vehicles. However, the majority of vehicles is intended to be used more or less exclusively on road surfaces specially designed for them. The road surface has a defined structure and is supported by a permanent foundation layer, that is, e.g., in Europe, made of asphalt macadam or concrete.

The kinematic design of a road depends on its intended use. In order to forecast and describe the vehicle movement on a road, the road design must be described physically.

For the description of the mutual interaction between road and tire it has proven to be advantageous to define the spectral distribution of the surface irregularities (Table 1.).

Table 1. Road spectral analysis

Characteristic		Wavelength	Influence on:
Slopes		50 m< λ	Pseudo-static
Undulations		1 m< λ<50 m	Dynamic (ride)
Roughness		10 mm< λ<1 m	Dynamic (NVH)
Texture	Macro	1 mm< λ<10 mm	Friction and noise
	Micro	10 µm< λ<1 mm	Friction
Material	Molecular	Friction	

Surface Irregularities. For wavelengths greater 50 m (ascents, hills) and vehicle speeds of approx. 20 m/s (70 km/h), the acting period of such irregularities is > 2.5 s; this is sufficiently slow for a vehicle to remain in the static balance of its suspension. For describing the tire behavior, these irregularities are not relevant.

Wavelengths of approx. 1–50 m, i.e., acting periods of 50 ms – 2.5 s, are designated as *undulations*. Undulations effect essentially a vertical dynamic response of the vehicle. This includes the excitation of the vehicle body with a resonance frequency of approx.1 Hz and the excitation of the resonance wheel frequency of approx. 10 Hz. These excitations are controlled or reduced by the spring/damper system (shock absorber). In coping with such vibrations, tires play only a minor role (influencing the vertical dynamics).

Wave lengths of 10 mm – 1 m, with acting periods of 0.5 – 50 ms, are called *roughness*. Roughnesses are essentially mastered by the tire deformation, and, in consequence, lead to a filtered dynamic response behavior of the vehicle (influencing the ride comfort).

Wavelengths under 10 mm influence the dynamics of a vehicle only to a minor extent. The respective acting period is < 0.5 ms. But within this wavelength range, the *road holding behavior* of the tire is mainly influenced (influencing the horizontal dynamics of the vehicle).

The subdivision of the roughness into macrotexture and microtexture (Fig. 2.) is the result of an analytic “dissection” of real roads in relation to their influence on tire friction.

Figure 2. Macro- and microtexture of road surfaces

Normally, the macrotexture is classified as “rough” or “smooth” whilst the microtexture is described as “harsh”. The macrotexture describes, e.g., the type of gravel used in an asphalt surface (1–10 mm) and thus the extent to which it protrudes from the matrix surface or has been rolled into the same.

The microtexture on the other hand describes the surface properties of the gravel elements. The influence of both characteristics can be seen on dry but even more pronounced on wet road surfaces. Here, the macrotexture assists in the actual water displacement in the contact area of the tire and road and thus supports the effect of the tire tread pattern, whilst the harsh microtexture – similar to the lamellas of a tire tread pattern – provides for local ground pressure and thus ensures the removal of the water film produced.

Under wet road surface conditions, surface porosity has an additional influence on water drainage (aquaplaning). In the case of dense nondivisible surfaces (e.g. concrete), this water displacement task must be performed by the tire tread pattern alone. By the way, tires with tread patterns are not required if surfaces are dry, smooth and consistent (frictional connection only!). If the “roughness” is of molecular magnitude, then the road surface material influences strongly the actual frictional connection of a tire (“adhesion”–related).

As far as the tire properties on a vehicle are concerned, above all with regard to handling, the road characteristics must always be taken into consideration. Thus, tire properties are tire/road surface properties.

[Top of Page]

3. Tire Characteristics [10-12]

As mentioned previously, the motion of a vehicle is practically determined totally by the forces which the road surface exerts on the tires. Thus, the tire characteristics (materials, structure, geometry) have a considerable influence on the handling of a vehicle. Within the tires, the forces must be generated, that “carry” the vehicle load (*load capacity*) and “move” the same (acceleration, braking, cornering).

However, the tire equipment must also serve as an intermediate medium to disengage the vehicle from road surface irregularities (*ride comfort*) and to provide a maximum number of kilometers travelling distance (*mileage*), and all this with maximum reliability for its entire service life (structural integrity).

Handling (Horizontal Dynamics). The contact areas of the four tires (for two- track vehicles, in the case of two-wheel vehicles there are only two tires), that is, the interacting tire/road areas, are the locations where the necessary forces for the required kinematic behavior of a vehicle are generated (Figs. 3. and 4.). The pneumatic tire made from rubber and equipped with a reinforcing material (carcass) has been the ideal technical product for this purpose without any alternative right from the start, i.e. for more than a hundred years now. The dominance of this construction principle is essentially the result of the rubber properties of the tire construction elements. Rubber elasticity and rubber friction render the tire suitable for load bearing and moving the same under very different ambient conditions (Fig. 5.). For a given tire/road surface pair, the lateral force – built up in the contact area by a specific vertical load and a specific internal tire pressure – mainly is a function of the slip angle (angle position of the tire relative to the direction of travel) and the tire camber (Fig. 4.). It is furthermore a rule that the longitudinal forces (acceleration forces, braking forces) can only be generated by slip (relative movement of the contact area to the road surface).

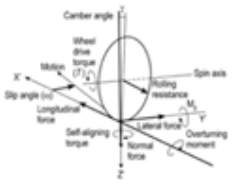

Figure 3. Tire forces

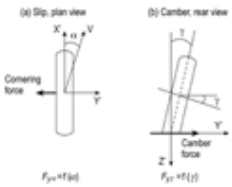

Figure 4. Force exerted by the ground by side slip and camber

Figure 5. Tire friction in longitudinal direction

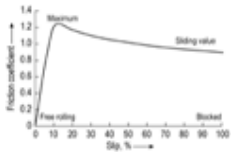

Due to the many dependencies it is practically impossible to predict the magnitude of the resulting lateral force on a tire under real conditions of handling and even less so the resulting lateral movement of a vehicle with four tires: This is due to the fact that the tire characteristics vary strongly from manufacturer to manufacturer, the vehicle concepts can be very different; and the road surfaces can vary greatly, which results in a strong variation of grip and time behavior, running performance (caused by the frictional wear behavior of the tire), etc.

Ride Comfort Behavior. To describe its behavior on the vehicle, a pneumatic tire can be represented as a spring-type action-damper combination (Fig. 6.). Its speed- and frequency-dependent spring-type action or damper behavior is determined by the air pressure inside the tire, geometry factors of the tire-rim combination and by the material components of the tire itself (elastic properties of the reinforcement carriers, rubber elastic properties of the rubber matrix materials). Depending on their actual frequency, the vibrations induced by the road surface during the rolling process (or by tire irregularities, tire nonuniformity) are transmitted to the vehicle via the axles; in this process, the vibrations are dampened by the tire to a greater or lesser degree. Therefore, the ride comfort of a vehicle is determined largely by the tires, acoustically (acoustic comfort) as well as on a tactile basis (plastic comfort).

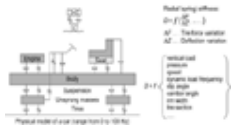

Figure 6. The tire as a part of the vibrating system vehicle

[Top of Page]

4. Construction [13]

Tire Reinforcements. The essential construction element of a pneumatic tire is the *carcass*, a rubber compound matrix, reinforced by one or several layers of textile or metal cord fabric. The carcass “contacts” the wheel rim on both sides (beads). The tire beads are used to fix the cord reinforcement, wrapping it around the bead cable (carcass wrapping), and on its external surface to “attach” the tire to the rim. The metal bead cables are rubberized and shaped such that they compress the carcass firmly onto the rim seats (incline from 5–15°) under the internal air pressure of the tire.

For passenger car tires, depending on load capability or speed category, the internal air pressure is 1.5–3.5 bar; in the case of commercial vehicle tires (trucks) 7.0–9.0 bar (higher loads).

Depending on the actual use of the tire, the reinforcing materials (*tire cords*) are normally made of rayon, nylon, polyester, aramid or “brassed” steel. The tire cords have a higher elastic modulus than the rubber matrix. Also, their creep behavior is significantly lower. In this way, in the case of pneumatic tires, they take up the forces built up by the internal tire pressure and kinematic vehicle movements, and transmit them to the road. Thus, very significant tensions are applied in the tire cord. The tensions influence all tire properties such as handling, comfort, wear etc.

In the case of a *radial tire* the carcass cords (one layer) run from bead to bead, 90° to the running direction of the tire (see Fig. 7.). In *diagonal* (*bias*) *tires*, the tension direction runs at approx. 35–40° to the running direction (diagonally), with two or more cord layers with a respective mirror-symmetric angle position of the cords being necessary here (Fig. 8.). The rubberization of the cords is relatively soft (low filled compounds) in order to ensure a good fatigue behavior at dynamic tire loads.

Figure 7. Construction of a radial tire

Figure 8. Construction of a diagonal tire

In a radial tire, the carcass is fixed in its position — and thus its cross-section contour — by a *belt*, mostly made up of steel cord “bandages” of a defined width crossed at an angle of 20–30°. This principle makes it theoretically possible to produce any cross-sectional tire contour (low-section tires); this is one of the essential benefits of this construction as compared to the diagonal principle.

Tire Compounds. In a radial tire, the belt is circumferentially surrounded by the *tread*. The tread is the contact medium to the road surface and is thus used to build up the driving forces and also for the transmission of forces “into the tire”. The tread is equipped with a *tread profile* featuring longitudinal and transverse grooves as well as lamellas with a depth of approx. 8 mm in the case of tires for passenger cars, and 12–20 mm (depending on the area of use) in the case of commercial vehicles. The profile is essentially used for the drainage of water on a compact road surface or for the generation of acceleration and deceleration of the vehicle on loose road foundations by means of form fit (traction). The tread compound must be relatively hard (highly filled) in order to ensure a good running performance; otherwise the grip to the road surface must be assured by means of the composition and deformability of the profile elements (block stiffness). Typical passenger car tread compounds feature blends of SBR and BR whilst commercial vehicle tread compounds – depending on their area of use – consist of 100% NR or blends of NR with SBR and/or BR.

Modern passenger car and commercial vehicle tires are *tubeless*, i.e., as their innermost rubber layer the tires contain a membrane (liner) which is highly non-permeable for air. As a result of the combination of tubeless tires with air-tight wheel rims (tubeless rims) the tubes used as standard formerly have become obsolete; this increases safety, is an advantage for mounting capability and reduces the weight of the tire system.

In total, a tire is composed of up to about ten different compounds which in themselves are again built up from different rubber blends (natural rubber, SBR, BR), fillers (carbon black, silica), softeners (oils, resins), antiaging agents and vulcanization systems (sulfur, accelerators), that are designed for the locally required mechanical and thermal properties of the tire components. A car tire contains 50% rubber (natural and synthetic), 26% carbon black and other fillers, 14% textile and steel cord, 5% bead wire and 5% vulcanization accelerating and antiaging system chemical (Fig. 9.). The required properties of tire compounds are listed in Table 2.

Figure 9. Tire materials

Table 2. Required properties of fire compounds

Tire Component	Requirements	Rubber used within	
		Passenger Car	Truck
Tread	low wear	SBR/BR/(75/25 to 55/45)	NR
	high friction		
	low hysteresis	NR/SBR/BR(25/35/40)	NR/SBR/BR (50/30/20)
	low tearing		
	Ozone resistance		
Casing	high elasticity	NR/BR (50/50)	NR/BR
	good adhesion to textile cord		
	low hysteresis		
Belt	good flow	NR	NR
	good adhesion to textile cord		
	low hysteresis		
	fatigue/tear resistance		

Apex	high modules good adhesion	NR	NR
Linear	bending strength low permeability curing properties	Halobutyl NR/SBR (60/40) NR/Halobutyl	Halobutyl

The actual tire geometry and the selection of the materials for the various components (compounds, reinforcements) of a tire must be carefully balanced by the tire development engineer. Even if an individual tire property can be optimized, this is in most cases connected with significant detrimental effects on other tire properties (interdependent variables).

[Top of Page]

5. Size Designation

A tire size refers to two base dimensions of a wheel rim, the rim diameter and the associated rim width. On the rim, the tire is “built” up, with the dimensional measures tire width and cross-sectional height. The external tire diameter results from the rim diameter and cross-sectional height (Fig. 10.).

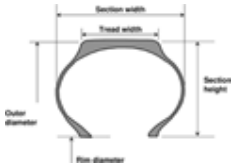

Figure 10. Tire cross-section

Since the 1960s there has been a trend leading – on a wide scale – to tires with steadily decreasing cross-sectional heights. This is because on the one hand lower cross-sectional heights provide significant advantages especially with respect to the handling of any type of vehicle, and because it was possible on the other hand to compensate their poorer comfort on the vehicle. Modern motorcycle and passenger vehicle tires range from a cross-sectional ratio of 80 over 70, 65, 60, 55, 50, 45, 40, 35 to 30. Even for commercial vehicle tires cross-sectional ratios of 100, 85, 80, 75, 70, 65, 60, 50 and even 45 for super single tires are found today (Fig. 11.).

Figure 11. Development of the tire cross-sectional ratio

[Top of Page]

6. Load Capacity

In a first approximation, the tire load capability is determined by the air volume of the tire and the internal tire pressure. When loaded, the tire experiences a vertical deflection. During the rolling process, this leads to a cyclic deformation of the tire in the running surface, belt and side wall areas; and due to the rubber elastic, hysteresis-type properties of the rubber components involved, this also leads to energy losses (rolling resistance, heat build up).The permitted tire deflection and thus the permitted load depends on the load-time duration of the tire application (e.g. aircraft tires: short term maximum loads with high deflection; commercial vehicle tires: long-term high load with low deflection). Non-permitted high loads, excessive speed, insufficient tire air pressure and unfavorable chassis geometries (e.g. a high camber) lead to nonpermitted high deflections and thus a major heat build up in the rubber materials. As a consequence this leads to tire failures (bubble formation, chunking of block elements, belt separations, and finally tire bursts).

[Top of Page]

7. Standardization [14]

The purpose of tire standardization is the worldwide interchangeability of products from different manufacturers. Standardization encompasses the tire designation and the tire dimensions. In addition, it comprises the correlation between tire load carrying capability and tire pressure as well as the influence of vehicle parameters such as maximum speed and wheel camber. Also, the assignment of the recommended and permitted rims is controlled by the standards.

The ISO, in Europe ETRTO (European Tire and Rim Technical Association), in the USA T&RA (Tire and Rim Association), and in Japan JATMA (Japan Automobile Tire Manufacturers Association) are responsible for standardization. At a national level, the following are also active: e.g. AFNOR in France, BSI in the UK, in Germany DIN as well as WdK (Wirtschaftsverband der Deutschen Kautschukindustrie).

In addition, in some countries there are tire-approval regulations for tires as vehicle components. These serve to provide road traffic safety and differ from country to country. International safety legislation and safety codes are listed in Table 3.

Table 3. Safety legislation and safety codes

Tire attribute	Safety legislation and regulations				Quality legislation	
	EEC directive 92/93[a]	ECE Rule (Europe)[b]	FMVSS 109/110 (Canada) FMVSS 119/120 (United States)[c]	ADR 23B (Australia)[d]	UTQG (United States)	
Tire marking	PC	PC	PCM	PCM	P	P
Rims	PC	PC	PCM	PCM	P	
Dimensions	PC	PC	PCM	PCM	P	
Load capacity	PC	PC	PCM	PCM	P	
Maximum speed rating	PC	PC				
Inflation pressures	PC[f]	PC[f]	PCM[g]	PCM[g]	P[g]	P[f]
Structural durability			PCM	PCM	P	
High-speed ability	PC	PC	PCM	PCM	P	
Impact resistance			PCM	PCM	P	
Tire seat on rim (Bead unseating resistance)			P	P	P	
Traction (braking)						P
Treat wear						P
Noise (tire-road)	PC[h]	PC[h]				
Rolling resistance						P[i]

P passenger; C commercial; M motorcycle
[a]European Economic Community;
[b]UN Economic Commission for Europe;
[c]Federal Motor Vehicle Safety Standard;
[d]Australian Design Rule;
[e]Uniform Tire Quality Grading;
[f]For testing;
[g]maximum allowable value;
[h]In preparation;
[i]planned.

.

The side wall of a tire provides information on some of the standards or type approval parameters (see Fig. 12.).

Figure 12. Tire marking

[Top of Page]

8. Products

8.1. Tires for Two-Wheel Vehicles

For their own stabilization – which is based on the gyroscopic principle – two-wheel vehicles with tilt instability always require a lateral force component to assure lateral guidance. The lateral forces, that the tire has to build up, are mainly compensated by the inclined position of the vehicle, i. e., by its camber.

The significantly lower running performance of motorcycle tires as compared to passenger car tires is caused by the lower

contact area and, in the case of motorcycles with powerful engines by the greater forces to be transmitted to the road (greater slip).

Up to the early 1980s only diagonal tires were used for two-wheel vehicles. With the increase in vehicle engine power and speed — above all in the case of motorcycles — the diagonal construction principle reached the limits of its performance capability (tire growth, insufficient contour reliability, with the consequences high wear, insufficient grip in curves, and poor suitability for high speed). Radialization provided the technical solution to the problem (twice bent shell geometry).

Depending on the vehicle (type, construction), a range of different tire constructions are today on the market (see Figs. 13. and 14.).

Figure 13. Segmentation of motorcycle tires

Figure 14. Construction of motorcycle tires

8.2. Passenger-Car Tires

In the highly industrialized countries with good road systems and high vehicle speeds, passenger-car tires are built practically exclusively according to the *radial principle*. The tires are classified according to their use (summer tires, winter tires, see Fig. 15.).

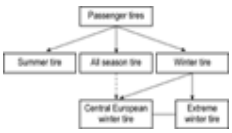

Figure 15. Segmentation of passenger-car tires

The typical requirements for European *summer tires* are:

- Best performance on good, solid road surfaces
- No loss of performance even at temperatures >50 °C
- High speed capability
- Maximum dry adherence
- Maximum wet adherence
- Aquaplaning resistance
- High steering precision and stability
- High comfort
- Minimum noise generation on all road surfaces
- High mileage
- No loss of air

The following requirements must be met by a Central European *winter tire*:

- Use from –25 to +7 °C
- Use on dry and wet surfaces, temporarily covered with snow, ice and mud
- High dry adherence
- High wet adherence
- Traction on ice and snow
- Traction on mud
- Aquaplaning resistance
- Steering precision and stability
- Good comfort
- Low noise generation
- High mileage
- High speed capability (“H”)

As winter conditions are very different in the mountains or in arctic countries, the requirements for extreme ice and snow tires differ quite strongly from the Central European type:

- Use from –40 to +7 °C
- Use on irregular surfaces, covered for months with ice and snow
- Traction on ice (also with studs!)
- Traction on cold, loose but also hardened snow
- Good mud performance

In the USA, all-season tires have become dominant. These tires constitute a compromise between summer tires and

Central European winter tires. A comparison between all-season and European winter tires is given in Table 4 and Fig 16.

Table 4. Requirements of all seasons tires in comparsion with European summer tires

Criteria	Demands in reference to base
Dry handling, lateral dynamics	losses acceptable
Straight ahead, longitudinal grooves	at least equal
Wet handling performance	losses acceptable
Aquaplaning	equal
Comfort	noticeable better
Noise	noticeable better
Mileage	significant better
Irregular Wear	low
Winter Performance	minimum mobility
Vibrations and Flat Spot	equal
Weight, Rolling Resistance	significant better
Brand acceptance	preprequisite

Figure 16. Performance of all season tires vs. European summer tires

Depending on the vehicle speed, tires are offered with different speed symbols (S ≤ 180 km/h, T ≤ 190 km/h, H ≤ 210 km/h, V ≤ 240 km/h, W ≤ 270 km/h, Y ≤ 300 km/h, Z ≥ 300 km/h), see Fig. 17.). Depending on the manufacturer, tires feature different tread patterns (technical properties, optical appearance).

Figure 17. Different Pirelli car tires

In the case of passenger cars, one aim of the vehicle industry is to eliminate the spare wheel to this end, so-called *run -flat systems* have been developed. In the event of a loss of air because of a tire puncture, these systems provide for the “onward movement” of the vehicle (> 200 km) without having to change the tire until a service point is reached (Fig. 18.).

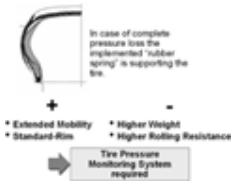

Figure 18. Self supporting tires

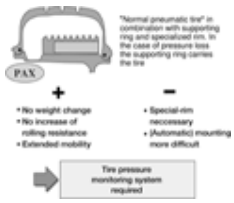

Figure 19. Run flat system

8.3. Tires for Commercial Vehicles

In principle, the same requirements as for passenger cars apply to commercial vehicles; however, economic efficiency is of prime importance here. Apart from all security arguments the predominant economic factors are:

- Good mileage
- Regular wear (between tire “center” and “shoulder”)
- Low rolling resistance (for low fuel consumption!)

- High reliability
- Low tire weight
- Regroovability
- Retreadability of the casing (more than three times)
- Low cost of the tire

In addition, depending on the kind of use, the following may be important:

- Traction
- Comfort (tactile)
- Easy mounting of snow chains

In comparison with passenger-car tires, tires for commercial vehicles have to cover a wider range of total vehicle weight, vehicle performance, and areas of use. This leads to widely differing tire requirements. A classification of commercial tires is given in Figure 20.

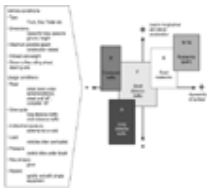

Figure 20. Classification of commercial tires

In Europe mainly *radial tires* are used for commercial vehicles. In countries with lower vehicle speeds and poorer road conditions, however, *diagonal tires* are predominantly in use. Also, *tubeless tires* are standard in Europe. Different types of Pirelli truck tires are presented in Figure 21.

Figure 21. Pirelli truck tires

[Top of Page]

9. Economic Aspects [14]

In 2001 the world tire production was 1141 × 10^6 units, 822 × 10^6 of which were tires for passenger cars, 319 × 10^6 units were produced for commercial vehicles. The world tire production for passenger cars and commercial vehicles by geographical area is shown in Table 5. Production broken down by leading manufacturing countries is listed in Table 6

Table 5. World tire production by geographic area (unit×10^6) Source: JATMA

		1997	1998	1999	2000	2001	2001 Share (%)
North America	Passenger car tires	245	249	248	257	237	28.9
	Commercial vehicle tires	52	56	56	57	50	15.8
	Total	297	305	304	315	288	25.2
South and Central America	Passenger car tires	42	43	43	44	42	5.2
	Commercial vehicle tires	24	25	25	26	26	8.2
	Total	66	68	68	71	68	6.0
Europe	Passenger car tires	218	231	229	231	227	27.6
	Commercial vehicle tires	37	41	44	45	43	13.5
	Total	255	272	273	276	270	23.7
Africa	Passenger car tires	9	10	11	12	12	1.5
	Commercial vehicle tires	4	4	5	6	6	1.8
	Total	13	14	16	18	18	1.6
Asia/Oceania	Passenger car tires	200	206	218	227	226	27.4
	Commercial vehicle tires	103	105	105	107	106	33.1
	Total	303	311	323	334	331	29.0

Others	Passenger car tires	53	59	61	65	78	9.4
	Commercial vehicle tires	80	79	84	87	88	27.6
	Total	133	138	145	152	166	14.5
Total	Passenger car tires	767	798	810	737	822	100.0
	Commercial vehicle tires	300	310	319	328	319	100.0
	Total	1,067	1,108	1,129	1,165	1,141	100.0

N.B.: 1. “Commercial vehicle tires” includes truck, bus and light truck tires.
2. “Others” includes the People/s Republic of China, as well as several former Communist regions, including the Commonwealth of Independent States and some countries in Eastern Europe.
3. Including some estimates.
4. Table values for unit volume are million-tire approximations. Consequently, these figures may not correspond exactly to the total.

Table 6. Tire production by leading manufacturing countries (unit×10^6) Source: JATMA

		1999	2000	2001	2001 Share (%)	01/00
United States	Passenger car tires	215	223	203	24.7	91.0
	Commercial vehicle tires	53	54	44	13.7	81.3
	Total	268	277	247	21.6	89.1
Italy	Passenger car tires	29	28	29	3.5	103.1
	Commercial vehicle tires	4	3	3	1.0	92.6
	Total	33	32	32	2.8	102.0
France	Passenger car tires	59	60	57	6.9	93.8
	Commercial vehicle tires	8	7	7	2.3	97.6
	Total	67	68	64	5.6	94.2
Germany	Passenger car tires	51	53	56	6.8	105.2
	Commercial vehicle tires	9	10	10	3.0	94.6
	Total	59	63	65	5.7	103.5
Korea	Passenger car tires	49	53	51	6.1	95.7
	Commercial vehicle tires	17	17	17	5.3	97.3
	Total	67	70	68	5.9	96.1
Japan	Passenger car tires	119	122	122	14.9	100.6
	Commercial vehicle tires	43	43	40	12.7	94.7
	Total	161	164	163	14.3	99.1

N.B.: 1. “Commerical vehicle tires” includes truck, bus and light truck tires.
4. Table values for unit volume are million-tire approximations. Consequently, these figures may not correspond exactly to the year-on-year comparison.

.

[Top of Page]

10. Environmental Aspects [15]

Due to tire wear on the road surfaces caused by rubber friction, tire wear material in the form of minute rubber particles (30–70 µm) is generated. On average, approximately 60 000 t rubber wear material is produced in this way per year in Germany (for example). The wear material contains mainly degraded rubber and carbon black, and in addition approximately 1% zinc in the form of its oxide, the oxide being contaminated by approximately 1 % lead and cadmium. Tire wear material does not represent any health hazard as it is naturally degraded chemically as well as bacteriologically.

[Top of Page]

References

1. S. K. Clark (ed.): *Mechanics of Pneumatic Tires*, US Dep. of Transportation, NHTSA, Washington DC 1981.

2. H. B. Pacejka (ed.): *Tires and Vehicle Dynamics*, SAE, N.Y. 2002.
3. J. C. Dixon (ed.): *Tires, Suspension and Handling*, 2nd ed., SAE, N.Y. 1996.
4. M. Klüppel, Th. Alshuth, R. H. Schuster: "Theorie der Gummireibung and Traktionseigenschaften von Reifen," *Fortschritts-Berichte VDI, Reihe 12, No. 362*, 1998.
5. P. Bösch, D. Ammon, F. Klempau: "Reifenmodelle – Wunsch und Wirklichkeit aus Sicht der Fahrzeugentwicklung," *Fortschritts-Berichte VDI, Reihe 12, No. 511*, 2002.
6. H. Huinik: "Reifen" in H. Braess, U. Seiffert (eds.): *Handbuch Kraftfahrzeugtechnik*, Vieweg Verlag, Braunschweig 2000.
7. R. Weber: "Reifenführungskräfte bei schnellen Änderungen von Schräglauf und Schlupf", Habilitations-Schrift, Universität Karlsruhe, 1980.
8. G. Heinrich, J. Schramm, A. Müller, M. Klüppel, N. Kendziorra, S. Kelbch "Zum Einfluß der Straßenoberfläche auf das Bremsverhalten von PKW-Reifen beim ABS-nass und ABS-trocken-Bremsvorgang", *Fortschritts-Berichte VDI, Reihe 12, No. 511*, 2002.
9. H. Braess: "Der Reifen –Schlüsselkomponente für Fahrdynamik und Lenkverhalten von Kraftfahrzeugen", *Fortschritts-Berichte VDI, Reihe 12, No. 437*, 2000.
10. G. Leister, in G. Runtsch: "Ermittlung objektiver Reifeneigenschaften im Entwicklungsprozess mit einem Reifenmessbus," *Fortschritts-Berichte VDI, Reihe 12, No. 362*, 1998.
11. J. Reimpell, P. Sponagel (eds.): *Fahrwerktechnik: Reifen und Räder*, Vogel Buchverlag, Würzburg 1988.
12. G. Breuer: "*Vorlesung Fahrzeugtechnik I+II", TU-Darmstadt*, 2000.
13. D. Overhoff: "*Vorlesung Reifentechnologie I+II" TU-Darmstadt*, 2003.
14. JATMA-Report: "*Tire industry of Japan*" JATMA 2003.
15. R. Stark: Environmental impact of a tire, *Kautsch. Gummi Kunstst*. **48** (1995), 448.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Titanium, Titanium Alloys, and Titanium Compounds

HEINZ SIBUM, Deutsche Titan GmbH, Essen, Federal Republic of Germany (Chaps. 2 – 10, Section 11.1, Section 11.2 in part, Chap. 12)

VOLKER GÜTHER, Gesellschaft für Elektrometallurgie, Nürnberg, Federal Republic of Germany (Section 11.2 in part, Chap. 13 in part)

OSKAR ROIDL, Gesellschaft für Elektrometallurgie, Nürnberg, Federal Republic of Germany (Chap. 13 in part)

FATHI HABASHI, Université Laval, Department of Mines and Metallurgy, Québec, Canada (Chap. 14)

HANS UWE WOLF, Universität Ulm, Federal Republic of Germany (Chap. 15)

1. Introduction 2
2. Physical Properties 2
3. Mechanical Properties 3
4. Corrosion Behavior 4
5. Occurrence 6
6. Production of Titanium Tetrachloride 7
7. Production of Titanium Sponge . . 8
7.1. Reduction of Titanium Dioxide . . 9
7.2. Reduction of Titanium Halides . . 9
7.3. Thermal Decomposition of Titanium Halides 11
7.4. Electrowinning of Titanium 11
8. Processing and Reuse of Scrap Metal 13
9. Processing of Titanium Sponge . . 13
10. Production and Processing of Semifinished Products 14
11. Titanium Alloys 16
11.1. α-, ($\alpha + \beta$)-, and β-Alloys . . . 16
11.2. Highly Alloyed Titanium 18
12. Uses and Economic Aspects 19
13. Titanium Master Alloys 20
13.1. Ferrotitanium 20
13.1.1. Composition and Uses 20
13.1.2. Production 22
13.1.3. Economic Aspects 23
13.2. Titanium – Aluminum 23
13.3. Other Master Alloys 23
14. Titanium Compounds 23
14.1. Titanium(II) Compounds 24
14.2. Titanium(III) Compounds 24
14.3. Titanium(IV) Compounds 25
14.4. Interstitial Compounds 27
15. Toxicology and Occupational Health 27
16. References 29

1. Introduction

Titanium, Ti, atomic number 22, relative atomic mass 47.88, is a member of group 4 of the periodic table. It has two 4 *s* and two 3 *d* valence electrons and is a transition element. The electronic configuration is $1s^2\,2s^2p^6\,3s^2p^6d^{\,2}\,4s^2$. Natural titanium is a mixture of five stable isotopes [7]:

Mass number:	46	47	48	49	50
Abundance, %:	7.95	7.75	73.43	5.51	5.34

The following unstable isotopes also exist:

Mass number:	43	44	45	51
Half-life:	0.58 ± 0.004 s [8]	1000 a	3.08 ± 0.06 h [9], 21 d [10]	6 min [11], 72 ± 2 d [12]

History. In 1791, GREGOR isolated a metal oxide from the black, magnetic iron-containing sand found at Manaccan in Cornwall, and conjectured that it contained a hitherto unknown metal, which he named manaccanite. In 1795, KLAPROTH investigated the red schorl rock from Boinik in Hungary, and he also decided that this mineral must contain an as yet unknown metal, which he named titanium after the Titans, the sons of the primeval Earth Mother. After detecting the new metal, also in oxidic form, in other minerals such as titanite and ilmenite, KLAPROTH showed that it was identical with the metal in manaccanite.

Although the fundamental reactions of modern titanium production (digestion with sulfuric acid, hydrolysis of the resulting liquor, and re-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_095

duction in acidic solution) were already known before 1800, the first preparation of pure titanium oxide was carried out only in 1908, while full-scale production began in 1916 in Norway.

In 1910, HUNTER prepared 98 – 99 % pure titanium metal by reduction with sodium. The strong affinity of titanium for oxygen, nitrogen, and carbon made preparation of the pure, ductile metal difficult, and this was first achieved by VAN ARKEL in 1922 by reducing potassium hexafluorotitanate with sodium in an inert atmosphere. The foundations for large-scale industrial production were laid by KROLL's discovery of the reduction of titanium tetrachloride by magnesium. WARTMAN of the Bureau of Mines then developed a method of treating the titanium sponge so obtained and melting it in an electric arc furnace.

2. Physical Properties

Pure titanium is a silvery white, ductile metal, *mp* 1668 ± 50 °C [13, 14], *bp* 3500 °C [15]. The atomic radius is 0.145 nm for coordination number six in the crystal lattice [16].

α-Ti has hexagonal structure of the Mg type, almost close-packed and somewhat compressed along the *c*-axis. The lattice constants of α-Ti at room temperature are $c = 0.4679$ nm, $a = 0.2951$ nm, $c/a = 1.585$ for twice-refined iodide titanium [17].

At 882.5 °C [18], titanium transforms into the body-centered cubic β-phase, the heat of transformation being 3.685 kJ/mol [19]. Extrapolation of the lattice constants of titanium alloys whose β-phase can be quenched to room temperature, gives a lattice constant $a = 0.3269$ nm at room temperature [20] and 0.328 nm above 620 °C. Increasing contents of oxygen, nitrogen, and carbon, which are incorporated interstitially, slightly elongate the *a*-axis of α-Ti and considerably elongate its *c*-axis. These effects are most marked with carbon, and are very slight with hydrogen. The effect on the lattice constants of titanium of elements that form substitution mixed crystals is variable.

Some important physical data of titanium are as follows:

Density at 25 °C	4.5 g/cm^3 (high purity)
	4.51 g/cm^3 (commercial purity)
at 870 °C (α-Ti)	4.35 g/cm^3
at 900 °C (β-Ti)	4.33 g/cm^3
Coefficient of linear expansion at 25 °C	8.5×10^{-6} K^{-1}
Mean coefficient of linear expansion, 20 – 700 °C	$(9.0 - 10.1) \times 10^{-6}$ K^{-1}
Latent heat of fusion	20.9 kJ mol^{-1}
Latent heat of sublimation	464.7 kJ mol^{-1}
Latent heat of vaporization	397.8 kJ mol^{-1}
Specific heat capacity	
at 25 °C	0.523 J g^{-1} K^{-1}
at 200 °C	0.569 J g^{-1} K^{-1}
at 400 °C	0.628 J g^{-1} K^{-1}
Thermal conductivity at 20 – 25 °C	0.221 W cm^{-1} K^{-1} (high purity)
	0.226 – 0.201 W cm^{-1} K^{-1} (commercial purity)
Surface tension at 1600 °C	1.7 N/m
Diffusion coefficient (self diffusion)	
at 750 °C (α-Ti)	4×10^{-13} cm^2/s
(β-Ti)	2.4×10^{-9} cm^2/s
Modulus of elasticity at 25 °C	100 – 110 GPa
Temperature coefficient	− 0.687 GPa/K
Modulus of rigidity at 25 °C	411.8 – 431.5 GPa
Temperature coefficient	0.265 GPa/K
Bulk modulus at 25 °C	122.6 GPa
Transverse contraction number at 25 °C	0.30
Electrical resistivity	
at 25 °C	42 μΩ · cm
at 600 °C	140 – 150 μΩ · cm
Superconductivity (iodide titanium), transition temperature	0.40 ± 0.04 K
Magnetic susceptibility of α-Ti at 25 °C	3.2×10^{-6} cm^3/g
Temperature coefficient for − 200 to 800 °C	0.0012×10^{-6} cm^3 g^{-1} K^{-1}
Magnetic permeability at 25 °C	1.00005

Further details can be found in [13–15, 21–37]. Many properties of titanium at a given temperature show variations that depend on the composition and condition of the metal (purity, alloying elements, thermal and mechanical pretreatment).

Variation in properties such as electrical and thermal conductivity and plastic behavior are caused by lattice defects. Cold working increases hardness and strength by producing dislocations, and reduces modulus of elasticity and electrical conductivity, e.g., by producing holes in the lattice. Annealing in the recovery range (300 – 500 °C) and in the recrystallization range (500 – 800 °C) restores the original values. The moduli of elasticity and shear elasticity also depend on heat treatment. The presence of foreign elements, even at low concentrations, often has

a marked effect. In general, electrical resistivity, hardness, and strength increase with decreasing purity.

3. Mechanical Properties

High-purity titanium (iodide and electrolytic titanium) still contains traces of other elements due to the high affinity of the metal for atmospheric gases. Above room temperature, the strength decreases, by 50 % at 200 °C, whereas the fracture strain remains virtually unchanged. Increasing the grain size decreases tensile strength and yield strength but increases elongation and reduction in area at fracture. Compositions and mechanical properties of high-purity titanium grades are listed in Table 1.

Titanium of Commercial Purity. As oxygen, nitrogen, and hydrogen contents increase, strength increases and toughness decreases. Whereas oxygen is the only element deliberately added to give increased strength, the others, together with iron and carbon, are impurities introduced during production.

The chemical compositions and strength properties of commercial grades of titanium in the recrystallized-annealed state are listed in Table 2. The Grade 1 material is the softest grade of titanium, and has excellent cold working properties. It is used for deep drawing, cladding of steel reactors, as a coating applied by explosive cladding, for components that must have good corrosion resistance but need not have high strength, and in electrical technology. Titanium Grade 2 is the most widely used commercially pure titanium. It has good cold working properties, and may also be used for load-bearing components at temperatures up to ca. 250 °C. Titanium Grade 3 is used almost exclusively for pressure vessels. This material has adequate cold-working properties and can be used at lower wall thicknesses than softer grades of titanium. Titanium Grade 4 has the highest strength and is used for the manufacture of fittings and in aircraft construction. Complex components made of this grade of titanium have to be formed at ca. 300 °C.

As shown in Figure 1, the high-temperature tensile strength of all industrial grades of titanium decreases rapidly with increasing temperature. As titanium creeps under prolonged stress even at room temperature, the characteristic high-temperature strength values used in calculations based on 100 000 h periods, as required for chemical pressure vessels, are considerably lower than those for short periods [40]. The use of titanium Grade 1 for pressure vessels is not recommended.

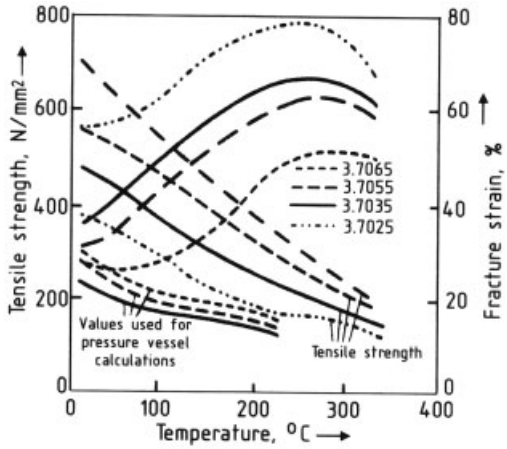

Figure 1. High-temperature tensile strength and fracture strain of titanium and values used for titanium pressure vessel calculations (10^5 h)
Titanium grade numbers are explained in Table 2

In addition to the fundamental strength properties determined by destructive testing and the creep strength values determined by long-term static load tests, fatigue strength (resistance to oscillatory stress) is also important. In the case of titanium, it is between 50 and 75 % of the ultimate tensile strength.

Cold working (80 – 90 % deformation) doubles the tensile strength of titanium, and both the elongation and reduction in area at fracture decrease by ca. 50 %.

4. Corrosion Behavior (→ Construction Materials in Chemical Industry, Chap. 5.9.)

The standard electrode potential of the reaction $Ti \rightleftharpoons Ti^{2+} + 2e$ is -1.75 V. The corrosion resistance of titanium metal is due to the formation of a thin, dense, stable, adherent surface film of oxide, which immediately reforms after mechanical damage if oxygen is present in the

Table 1. Percentage chemical compositions and strength properties of high-purity grades of titanium

	O	N	C	H	Fe	Al	Tensile strength, MPa	Yield strength, MPa	Fracture strain∗, %	Reduction in area, %
Iodide titanium [38]	0.021	0.004	0.015		0.005	0.04	230	121.5	55	60
Electrolytic titanium∗∗ [39]	0.020	0.003	0.009	0.003	0.001		229	103.0	55	

∗ Measured length 25.4 mm (1 inch).
∗∗ Typical figures for Brinell hardness 60 – 65.

Table 2. Chemical compositions (maximum contents, wt %) and strength properties (minimum values) of commercially pure grades of titanium

Grade number∗	O	N	C	H	Tensile strength, MPa	0.2 % Yield strength, MPa	1 % Yield strength, MPa	Fracture strain, %
3.7025 (Ti 1)	0.12	0.05	0.06	0.013	290 – 410	180	200	24
3.7035 (Ti 2)	0.18	0.05	0.06	0.013	390 – 540	250	270	22
3.7055 (Ti 3)	0.25	0.05	0.06	0.013	460 – 590	320	350	18
3.7065 (Ti 4)	0.35	0.05	0.06	0.013	540 – 740	390	410	16

∗ DIN 17 850, Nov. 1990 and DIN 17 860, Nov. 1990.

surrounding medium. Titanium is less resistant to corrosion in strongly reducing media.

The corrosion-resistant oxide layer is destroyed in completely water-free environments such as dry chlorine, dry oxygen, and red-fuming nitric acid, and in reducing corrosive media. In neutral aqueous solutions, especially in the presence of chloride ions, localized penetration of the passive layer can lead to pitting and crevice corrosion.

Passivation. In a corrosive medium, the electrode potential of a titanium surface decreases with time [41]. The passivation potential usually lies between -0.2 and -0.5 V. With increasing concentration and temperature, the passivation current density increases in the passive range, while the passivation potential remains constant [42]. The corrosion loss in the passive range therefore increases [43]. Small additions of oxidizing metal ions such as Cu^{2+}, Fe^{3+}, Cr^{4+}, Au^{3+}, and Pt^{4+} cause passivation of titanium in reducing media. With increasing temperature and acid concentration, the concentration of inhibitor necessary to give passivation increases. Anodic treatment strengthens the coatings and can delay their degradation under reducing conditions.

Additions. The corrosion behavior of titanium-based materials can often be improved by the addition of at least 0.15 % palladium to the alloy, without affecting its strength properties [44, 45].

Likewise, small additions (< 1 %) of other noble metals improve the corrosion behavior of titanium [46]. Improvement of corrosion resistance as well as strength properties is achieved by addition of, e.g., 0.8 % nickel with 0.3 % molybdenum [47].

Crevice Corrosion, Stress Corrosion Cracking, and Pitting. Titanium is not very susceptible to crevice corrosion (→ Corrosion, Chap. 3.2.2.) except at elevated temperatures in the presence of high concentrations of halide and sulfate ions or in moist chlorine gas. Additions of palladium to the alloy have a beneficial effect. Titanium is very resistant to pitting in the presence of chlorides, except in high concentrations of chlorides of calcium, aluminum, or zinc at elevated temperatures [48]. Titanium is not susceptible to stress corrosion cracking, except by red-fuming nitric acid containing < 1.5 % water, by uranyl sulfate, or by anhydrous methanol.

Reaction with Hydrogen, Oxygen, Nitrogen, and Air. The presence of hydrogen (produced in a chemical process or as a corrosion product) or of hydrogen-containing compounds leads to local hydrogen pickup by titanium, with consequent embrittlement and impairment of corrosion resistance. The rate of hydrogen pickup depends on temperature, time, and the

condition of the metal surface [49]. Oxide coatings reduce or prevent hydrogen pickup [50].

The oxide film that forms on the surface of activated titanium metal at room temperature reaches a thickness of 1.7 nm in 2 h, 3.5 nm in 40 d, and 25 nm in 4 years [51]. The color of the oxide film varies with its thickness and temperature of formation, and also with the alloy content of the parent metal. As the temperature increases from 200 °C to ca. 500 °C, the film thickness increases, and its color changes from golden yellow to brown, dark blue, violet, and light blue, showing iridescent effects, while at higher temperatures a gray haze forms followed by a silvery-white and then a dirty gray scale layer, after which a yellowish white to brown scale layer of variable thickness appears, depending on the alloy composition [52]. With increasing heating time and temperature, the depth of penetration of oxygen, and hence the hardness, increases [53].

Table 3. Corrosion behavior of commercially pure titanium

Resistant	Limited resistance	Nonresistant
Nitric acid	Sulfuric acid	Fluorine
Chromic acid	Hydrochloric acid	Dry chlorine
Sulfurous acid	Phosphoric acid	gas
Alkali lyes	Oxalic acid	Red-fuming
Ammonia	Formic acid	nitric acid
Aqueous chlorides		
Brine		
Seawater		
Wet chlorine gas		
Acetic acid		
Maleic acid		
Acetaldehyde		
Carbamate		
Dimethylhydrazine		
Liquid hydrogen		

Up to ca. 500 °C, the weight increase of titanium due to oxidation is small. Higher temperatures lead to a rapid weight increase associated with dissolution of oxygen in the metal.

In a nitrogen atmosphere, a titanium nitride layer with metallic character is formed. Above 700 °C, the titanium nitride increasingly reacts with oxygen to form titanium oxide. Titanium reacts with nitrogen considerably more slowly than with oxygen, so that the reaction with nitrogen makes little contribution to scale formation in air. In the presence of water vapor, hydrogen is also taken up.

Behavior in Salts, Bases, Acids, and Molten Materials. The corrosion behavior of commercially pure titanium is summarized in Table 3. Oxidizing conditions and the presence of oxidizing agents under reducing conditions improve corrosion resistance, whereas strongly reducing conditions or increases in concentration and temperature impair it [52].

Titanium is stable in molten alkali metals up to 600 °C, in magnesium up to 650 °C, in gallium, tin, lead, and lead – bismuth – tin alloys up to 300 °C, and in mercury up to 150 °C [54]. Titanium is corroded by molten mixtures of chlorides and fluorides.

5. Occurrence

Titanium is relatively widely distributed and abundant. With an estimated average concentration in the earth's crust of 0.6 %, it is ninth in order of abundance after oxygen, silicon, aluminum, iron, magnesium, calcium, sodium, and potassium. Its concentration is ca. 1/20 that of aluminum, and 1/10 that of iron. It is 5 – 10 times as abundant as chlorine, sulfur, or phosphorus, and more abundant than the rest of the metals put together. Thus, for example it is 60 times as abundant as copper or nickel and 300 times as abundant as molybdenum.

As the ionic radius of titanium is similar to that of some of the most common elements (Al^{3+}, Fe^{3+}, Mg^{2+}), most minerals, rocks, and soils contain small amounts of titanium, although true titanium minerals, containing > 1 wt % titanium, of which there are at least 87, are found in only a few locations.

Primary titanium occurs in igneous rocks, where it forms the acidic component of basic magmas and the basic component of acidic magmas. In the first case, titanates are present, the most important of which are ilmenite ($FeTiO_3$) and perovskite ($CaTiO_3$). In the second case, oxidic compounds of titanium are formed. Intermediate forms also exist, e.g., in silicates, in which titanium is present mainly as a basic element (e.g., in zircon minerals and aluminosilicates), but also as a replacement for silicon. These transitional forms include titanite (sphene), $CaTi(SiO_4)O$, a common component of magmatic rocks. All these forms also occur

in metamorphic titanium-containing rocks. The two TiO_2 modifications anatase and brookite occur exclusively in metamorphic deposits.

Secondary deposits of titanium include the widely distributed ilmenite placer deposits and ilmenite sands, which mostly occur in coastal regions, and the TiO_2 deposits in clays. Ilmenite is the main starting material for titanium dioxide pigments.

The most important titanium minerals are anatase (TiO_2); ilmenite ($FeTiO_3$), which contains up to 53 % TiO_2, and its low-iron weathering product leucoxene; perovskite ($CaTiO_3$); rutile (TiO_2); and sphene [$CaTi(SiO_4)O$]. Of these, only ilmenite, leucoxene, and rutile are of economic importance due to the ease with which they can be processed. The anatase deposits in Brazil and the perovskite deposits of Colorado (United States) may achieve economic importance in the future.

The most useful mineral for the extraction of titanium and titanium compounds is rutile (TiO_2). Although it is rarer than ilmenite, its TiO_2 content is higher. Naturally occurring enriched rutile is brown to black, and contains 90–97 % TiO_2 together with impurities including up to 10 % silica, iron oxides, vanadium, niobium, and tantalum, and traces of tin, chromium, and molybdenum compounds.

Primary rutile deposits occur at Kragero in south east Norway (albite with 25 % rutile) and in Virginia (United States). However, more important are the secondary placer deposits, e.g., in Brazil, Cameroon, and Arkansas (United States), and the beach sands, of which the most important are on the east coasts of Australia and Florida (United States) and in northern Transvaal (South Africa). All rutile and two-thirds of ilmenite are extracted from placer deposits and sands, the rest of the ilmenite being obtained from titanomagnetites and titanohematites.

Rutile-bearing beach and dune deposits can be mined if TiO_2 contents exceed 0.3 %. The TiO_2 content of the ilmenite sands depends on the degree of weathering. In the South African deposits it is 10–48 %, in Australian sands 54 %, and in ilmenite from Kerala (India) and the east coast of Sri Lanka up to 80 %. The ilmenite-bearing titanomagnetites and titanohematites contain 35 % TiO_2 in Canada, ca. 20 % in the United States, 18 % in Norway, and 13 % in Finland [55].

Titanium occurs in many stars. It is less abundant in meteorites than in the earth's crust, and is mainly associated with the silicate phase [56].

Reserves. Proven world reserves of rutile and ilmenite, calculated as TiO_2, content have been estimated at 423×10^6 t to 600×10^6 t. The largest reserves of ilmenite are in South Africa, India, the United States, Canada, Norway, Australia, Ukraine, Russia, and Kazakhstan, and of rutile in Brazil [57]. The existence of other reserves, e.g., in Bangladesh, Chile, Italy, Mexico, and New Zealand, makes it likely that the actual mineable world reserves considerably exceed the accepted figures.

6. Production of Titanium Tetrachloride

Titanium metal is produced exclusively by reduction of titanium tetrachloride, which is manufactured from natural rutile or from the so-called synthetic rutile obtained from ilmenite or from the TiO_2-rich slag produced by metallurgical treatment of ilmenite (ore preparation, production of TiO_2-rich slags, and production of synthetic rutile are described in → Pigments, Inorganic, Chap. 2.1.2.2.).

The TiO_2 produced for the pigment industry by treatment of ilmenite with sulfuric acid (→ Pigments, Inorganic, Chap. 2.1.3.) is unsuitable as a starting material for production of the metal because of the impurities present. Also, titanium white (TiO_2) is increasingly produced from $TiCl_4$ (→ Pigments, Inorganic, Chap. 2.1.3.3., → Pigments, Inorganic, Chap. 2.1.3.2.). At the present time, approximately one-third of TiO_2 world production (2.1×10^6 t/a) is produced by the chloride process. Over 95 % of the total quantity of titanium minerals extracted worldwide is used to produce TiO_2 pigment, and only ca. 4 % for titanium metal.

In spite of the wide distribution of titanium minerals, the mining of rutile-containing ores is mainly concentrated on Australia, from which 90 % of world production comes. Other important producing countries are Sierra Leone

and South Africa. The most important ilmenite-producing countries are Australia (one-third of total world production) and Norway, the United States, and the Ukraine, while the leading producers of titanium slag from ilmenite are Canada and South Africa.

Chlorination of titanium dioxide is now carried out almost exclusively by the fluidized-bed process.

In the discontinuous fixed-bed process, which is now hardly used, rutile concentrate (> 96 %) is mixed with 20 – 25 % petroleum coke and a binder (wood tar, asphalt, etc.), sometimes with added catalysts (e.g., MnO_2), and briquetted [58]. The briquettes are stacked in the lower part of a brick-lined chlorination tower over a layer of carbon which acts as an electrode, and are reacted with chlorine at 500 – 850 °C. The chlorine gas is heated to ca. 1000 °C by the resistance-heated carbon bed and then reacts exothermically with the rutile and carbon in the briquettes.

The process can also be carried out in two stages, in the first of which rutile is reduced by carbon at 1200-1400 °C to give titanium carbide and titanium monoxide. The latter reacts with chlorine more readily than rutile.

Since ca. 1950, chlorination has been carried out almost exclusively by the fluidized-bed process (→ Pigments, Inorganic, Chap. 2.1.3.2.) because of its higher reaction rate and improved heat transfer, and also because of the shortage of briquetting facilities. Also, higher temperatures (ca. 1000 °C) can be achieved. After the reaction has started at ca. 600 °C no further external heat supply is necessary. Apart from the benefit of continuous operation, fluidized-bed chlorination has the further advantage of being less sensitive to impurities in the rutile or the carbon. The conversion of chlorine is 95 – 100 %, of rutile titanium 90 – 95 %, and of carbon 95 %.

Ilmenite is now less often used as starting material, because of excessive chlorine consumption due to formation of iron(III) chloride for which there is little demand and from which chlorine can only be recovered at high cost. Furthermore, in spite of its higher boiling point, $FeCl_3$ is entrained by the $TiCl_4$ vapor and is deposited in the coolers as a powder, its solubility in liquid $TiCl_4$ being only 0.03 % [59].

In contrast to the ilmenite-treatment process, the treatment of TiO_2-rich slags containing < 10 % iron is increasing in importance: As these materials contain lower oxides of titanium, they react with chlorine at lower temperatures and require less carbon than rutile [60]. However, compared with rutile, problems are caused by the higher iron content of these slags and their alkaline earth metal content. Calcium and magnesium are converted to their chlorides, which are molten at the reaction temperature and coat the surface of the briquettes in the fixed-bed process, preventing chlorine from penetrating. In the fluidized-bed process, the continuous addition of an inert bed (sand) can give good gas distribution and dilution of the chlorides, so that with continuous removal of residues, slags containing up to 6 % calcium or magnesium can be treated.

The use of other reducing or chlorinating agents (CO, $COCl_2$, CCl_4, sulfur chlorides) and other methods of producing $TiCl_4$ has not achieved industrial importance. For example, the electrolysis of a chloride melt using titanium carbide anodes has been proposed [61]. The most promising process seems to be to digest ilmenite with sulfuric acid and precipitate potassium hexachlorotitanate from the sulfate solution after removal of iron. On heating, this decomposes to form $TiCl_4$ vapor and KCl, which is returned to the process [62].

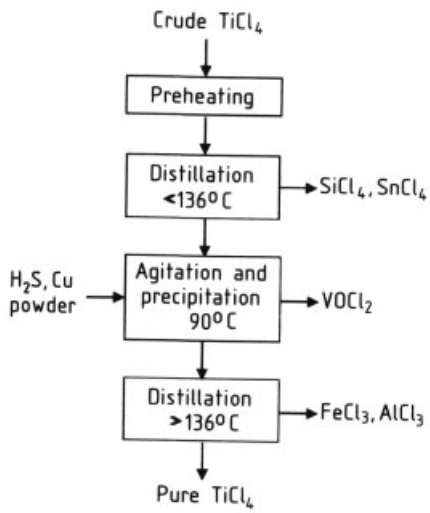

Figure 2. Purification of crude titanium tetrachloride

Separation and Purification of Titanium Tetrachloride. In the chlorination of rutile, the reaction products are gaseous, consisting of $TiCl_4$ and CO, with small amounts of CO_2, phosgene ($COCl_2$), and other metal chlorides (Fig. 2).

The gases leaving the chlorination reactor are cooled by heat exchangers and by spraying with titanium tetrachloride, and the iron(III) chloride that precipitates at 150 °C is scrubbed out by the $TiCl_4$ [63]. Further cooling causes the titanium tetrachloride to condense. The yellow filtered product contains 94 % $TiCl_4$, ca. 4 % solid constituents such as rutile, carbon, sulfur, and insoluble metallic chlorides, and 2 % soluble metal oxide chlorides. The chlorides include $SiCl_4$ and $SnCl_4$ (low-boiling substances), $VOCl_3$ which has a similar boiling point (127 °C) to $TiCl_4$ (136 °C), and $FeCl_3$ and $AlCl_3$, which have higher boiling points. Other high-boiling chlorides, e.g., of calcium, manganese, magnesium, and sodium, produce unwanted deposits in the equipment.

To purify $TiCl_4$ (→ Pigments, Inorganic, Chap. 2.1.3.3.), the solid constituents are first allowed to settle out. Small amounts of water are added to precipitate aluminum as its oxide chloride. $SiCl_4$ and $SnCl_4$ are removed by distillation at < 136 °C. H_2S is passed in and copper powder is added at 90 °C to reduce $VOCl_3$ to $VOCl_2$, which precipitates. $FeCl_3$ and $AlCl_3$ distil at > 136 °C. Unsaturated organic compounds, especially oleic acid, promote good separation of chromium and vanadium oxychlorides [64].

Dissolved chlorine can be removed by simply heating or by heating with metal powders (iron, copper, or tin). Purification with H_2S can also be carried out continuously in a fluidized bed of silica sand or iron sulfide at 140 – 300 °C. The impurities then form solid products which can be continuously removed from the fluidized bed.

The purified titanium tetrachloride contains only 0.002 % V_2O_5, and is fractionally distilled to remove residual phosgene and $SiCl_4$. The purity of the $TiCl_4$ is 99.9 % min.

7. Production of Titanium Sponge

The large heat of formation of titanium dioxide (945.4 kJ/mol) combined with the high solubility of oxygen in titanium at high temperatures has so far made it impossible to develop an economic process for the direct reduction of titanium dioxide to low-oxygen titanium metal. The standard industrial processes for the production of titanium metal are therefore based on titanium halides.

7.1. Reduction of Titanium Dioxide

The reduction of titanium dioxide by carbon is only possible above 6000 °C. Using carbon-containing reducing agents, some titanium carbide is formed, even in high vacuum [65, 66]. Reduction by hydrogen in the presence of inert gases leads to mixtures of lower oxides [67]. A more complete reduction of titanium dioxide is only possible with alkaline earth metals, whereby reduction with calcium gives the lowest oxygen content [68]. Using reactions at temperatures of 600 – 1200 °C in a vacuum followed by dissolution of excess calcium and calcium oxide in hydrochloric acid, an oxygen content of 0.1 – 0.3 % in the titanium is obtained [69]. Reduction with calcium hydride at 600 – 700 °C gives titanium hydride, which decomposes at 900 °C into titanium (containing 0.2 % oxygen) and hydrogen [70].

7.2. Reduction of Titanium Halides

When $TiCl_4$ is reduced with hydrogen in an electric arc, subchlorides are formed in side reactions, and this reaction is consequently uneconomic on an industrial scale [71]. The reduction of $TiCl_4$ with calcium is strongly exothermic [72], but this reaction also has not been used commercially.

Reduction of Titanium Tetrachloride with Sodium. In 1910, Hunter succeeded in producing larger quantities of pure titanium by reacting titanium tetrachloride with sodium in an evacuated steel bomb [73].

The Degussa process uses a mixture of oxide-free sodium and potassium at 700 – 800 °C. To prevent overheating (sodium chloride *mp* 797 °C, sodium *bp* 877 °C), molten sodium chloride is first placed in the reactor [74]. Finely powdered sodium chloride at 200 – 620 °C can also be used. Molten sodium is then fed into the

reactor from above, and $TiCl_4$ is blown in from below, together with the inert gas [75].

In industrial plant molten sodium and $TiCl_4$ are fed simultaneously into a steel reactor filled with argon, heated to 650 °C. After the reaction has begun, the temperature can rise to 900 °C. When reaction is complete, more sodium is added and the temperature is raised to 950 °C.

In the two-stage process, $TiCl_4$ is first converted at 235 °C into low-melting sodium titanium chloride compounds and titanium(II) chloride, and is then reduced to titanium and sodium chloride in a second reactor after further addition of sodium. This technique distributes the heat of reaction, and the process is therefore more easily controlled. After size reduction of the reaction product, the NaCl is dissolved in water, and the titanium sponge left behind is centrifuged off and dried.

The Hunter process has been almost completely superseded by the Kroll process. Industrial scale plants are currently only operated in China [76].

Reduction of Titanium Tetrachloride with Magnesium (Kroll process). The reduction of titanium tetrachloride with magnesium

$$TiCl_4 + 2\,Mg \longrightarrow Ti + 2\,MgCl_2$$

was discovered by Kroll [72, 77], and was developed into an industrial process by the Bureau of Mines in Boulder City, Nevada [78]. Magnesium boils at 1120 °C, and magnesium chloride melts at 711 °C. The resulting temperature range and the high purity of magnesium are advantageous for the industrial production of titanium sponge from $TiCl_4$.

The reactor is constructed of plain carbon steel, chromium – nickel steel, and the interior walls are cleaned by brushing, or are titanium coated. The reactor is charged with oxide-free lumps of magnesium and filled with argon, and the magnesium is melted at 651 °C (see flow diagram in Figure 3). When the temperature reaches 700 °C, purified $TiCl_4$ is run in slowly from above or blown in as a vapor, such that a reaction temperature of 850 – 950 °C is established. The titanium sponge is deposited on the reactor walls and forms a solid cake above the molten magnesium. The molten $MgCl_2$ collects beneath the magnesium and is drawn off. The magnesium rises through the porous cake to its surface by capillary action, and reacts there with the gaseous $TiCl_4$.

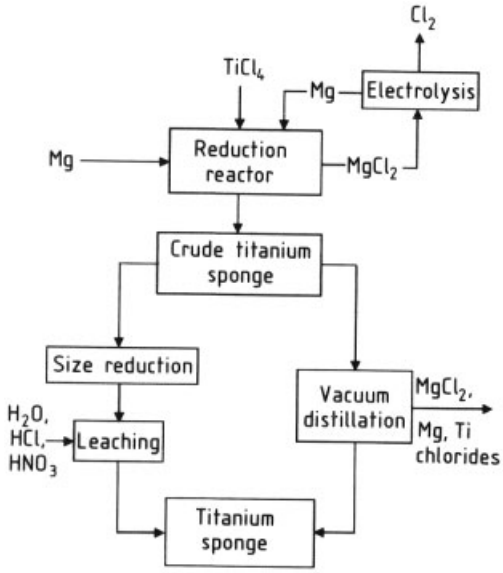

Figure 3. Reduction of titanium tetrachloride to titanium sponge by the Kroll process (reduction with magnesium)

The temperature must not exceed 1025 °C, to prevent reaction between the titanium and the iron of the reactor, although titanium can pick up iron even below 950 °C. The amounts of $TiCl_4$ reacted are 10 – 15 % substoichiometric amount, as some of magnesium and $MgCl_2$ remain in the titanium sponge. However, an excess of $TiCl_4$ leads to the formation of lower titanium chlorides and iron chlorides, which increases the iron content of the titanium sponge.

The temperature range that can be used in practice is 850 – 950 °C, the lower temperatures giving longer reaction times but purer titanium sponge. After the molten $MgCl_2$ has been removed and the temperature of the reactor contents has fallen below 200 °C, the reaction vessel is opened in a dry room ($MgCl_2$ and titanium chloride are hygroscopic).

The crude titanium sponge is generally purified by vacuum distillation; leaching is now rarely used. Around 98 % of titanium sponge produced is vacuum distilled. Immediately after completion of the reaction a vacuum hood usually is attached to the reduction reactor. At ca. 0.1 Pa (10^{-3} mbar) and 900 – 1000 °C, the metal chloride impurities, mainly magnesium and ti-

tanium chlorides, are distilled off together with magnesium metal while the titanium sponge remains behind.

In the leaching process, the size-reduced sponge is treated with a hot mixture of hydrochloric and nitric acids (8 % HCl, 3 % HNO_3) in tanks or in an inclined rotating titanium drum containing an Archimedean screw. Because of the heat of solution of the $MgCl_2$, cooling must be provided to prevent the temperature from exceeding 25 °C. During leaching, the titanium sponge picks up hydrogen, which can be removed by melting in a vacuum electric-arc furnace with a high-performance vacuum pump.

The process of drilling out the crude sponge from the reactor can be avoided by using a perforated crucible if the level of the molten magnesium is maintained above the perforated base [79]. Removal of the sponge from the perforated crucible can be facilitated by facing it with stainless steel or high-carbon steel sheet.

The reduction reactors, usually gas heated, have titanium sponge capacities between 1.5 and 10 t. Production of 1 kg titanium sponge theoretically requires 3.96 kg $TiCl_4$ and 1.015 kg magnesium, 3.975 kg $MgCl_2$ being formed. However, in practice, only 65 – 70 % of the magnesium takes part in the reduction process. Most of the remainder is recovered in vacuum distillation, whereas it is lost in the leaching process. The magnesium is recovered from $MgCl_2$ electrolytically. The proportion of the titanium sponge that can be used for the production of titanium metal is 75 – 85 %, and the total titanium yield is ca. 98 %. The energy consumption is 30 kWh/kg titanium sponge. The reduction process takes ca. 95 h, and the distillation ca. 85 h.

Variants of the Kroll process, so far not used industrially, are aimed at giving better separation of the sponge from the $MgCl_2$ and at achieving continuous operation. For example, in a ram reactor the sponge is compressed by a ram and is ejected as compact metal from the magnesium chloride bath [80].

The reduction can be carried out under vacuum instead of in an argon atmosphere, and the magnesium chloride can be distilled off immediately after completion of the reaction [81].

The simultaneous injection of $TiCl_4$ vapor and magnesium vapor into a reactor packed with titanium netting at 800 – 1100 °C and ca. 10^{-2} Pa produces pure titanium, which grows as crystals on the netting, while the byproducts (Mg vapor and $MgCl_2$ vapor) distil off [82].

In Japan, a quasi-continuous process is being developed in which the distillation of $TiCl_4$, Mg, and $MgCl_2$ and the melting of a titanium block are performed simultaneously [83]. A production of $\geq$ 500 t titanium sponge per furnace per month is expected.

Comparison of the Sodium and Magnesium Reduction Processes. Both methods are used industrially and produce titanium sponge of sufficiently high purity. Their advantages and disadvantages are as follows:

Sodium has a lower melting point than magnesium and can therefore be transported in the molten state. Reduction with sodium requires only the theoretical quantity, whereas ca. 130 % of the stoichiometric quantity of magnesium is required. With sodium reduction, the titanium sponge is more easily removed from the reactor, as NaCl collects on the reactor walls, while in the Kroll process the sponge builds up on the reactor walls. Sodium chloride is soluble in water and is therefore easily removed from the titanium sponge. This is a more economic method than vacuum purification of the titanium sponge produced by magnesium reduction. The density of magnesium-reduced sponge is 1.0 – 1.2 g/cm^3, (sodium-reduced sponge 0.7 – 0.9 g/cm^3) and it has larger particles than sodium-reduced sponge, which contains a high proportion of fines. Typical figures for the chemical composition of various types of sponge produced by both processes are listed in Table 4. This titanium from the Kroll process has a lower average oxygen content and that from the sodium reduction process, a lower average iron content. The highest purity achieved in the Kroll process so far is reported to be 99.999 % [83].

The costs of reduction with sodium are usually ca. 10 % lower than reduction with magnesium, although the economics of the Kroll process can be improved by using larger reduction reactors, improved vacuum distillation, and higher capacity electrolysis cells for recovery of magnesium. In both cases, batch operation is a disadvantage.

The sponges produced by different processes of chemical reduction, size reduction, and purification differ in hardness, lump size, density, and compressibility, but all are equally suitable

Table 4. Chemical composition (in wt %) of titanium sponge

Reduction process	Maximum hardness (Brinell)	C	N	O	H	Fe	Mg	Na	Cl
Magnesium	95	0.006	0.003	0.045	0.002	0.02	0.03		0.10
	100	0.008	0.005	0.05	0.002	0.04			0.09
	110	0.008	0.005	0.06	0.002	0.05	0.03		0.08
	120	0.008	0.005	0.07	0.002	0.06	0.03		0.09
Sodium	120	0.009	0.004	0.07	0.01	0.016		0.08	0.13
Electrolysis	60	0.003	0.002	0.01	0.003	0.001			0.08
	75	0.013	0.004	0.03	0.004	0.004			0.09
	90	0.018	0.004	0.04	0.004	0.020			0.16

for melting into ingots. Therefore, the reasons for selecting a particular reduction process are mainly economic, e.g., availability and reprocessing of reduction metals, energy costs, and capital costs.

In 1995, ca. 98 % of world titanium-sponge production was by the Kroll process.

Reduction of Titanium Tetrafluoride. Titanium tetrafluoride is easily produced from titanium dioxide and hydrogen fluoride, can be readily purified by sublimation, and, in contrast to titanium tetrachloride, titanium tetrafluoride can also be reduced by silicon. The high-purity titanium so produced contains only 0.009 % silicon [84].

7.3. Thermal Decomposition of Titanium Halides

The oldest process of this type is that of VAN ARKEL and DE BOER in which titanium tetraiodide is decomposed on electrically heated tungsten wires at 1000 °C. If wires or single crystal fibers of titanium are used, tungsten-free titanium of the highest purity can be obtained (as a single crystal in the latter case). The iodine liberated is reacted with crude titanium to regenerate titanium tetraiodide in the same vessel below 200 °C or above 500 °C. Above 200 °C lower titanium iodides (TiI_2, TiI_3) are formed, which only act as titanium carriers above 500 °C because of their low volatility [85]. The process enables titanium of the highest purity to be produced from crude titanium or titanium scrap.

Very high purity titanium can also be obtained by thermal disproportionation of titanium(II) chloride or bromide. By passing $TiCl_4$ vapor over titanium-containing material (titanium scrap, crude titanium, or titanium alloys) at 950–1500 °C, $TiCl_2$ vapor is formed, which condenses and is decomposed at ca. 1000 °C

$$2\,TiCl_2 \longrightarrow Ti + TiCl_4$$

The $TiCl_4$ is fed back to the process [86].

7.4. Electrowinning of Titanium

The high affinity of titanium for oxygen and hydrogen together with its electrode potential ($Ti/Ti^{2+} = -1.75$ V) prevents its deposition from aqueous solution [87]. Hence only molten salt electrolysis can be used.

Titanium dioxide is almost insoluble (0.07–0.02 %) in molten chlorides of alkali metals and alkaline earth metals. It is soluble (ca. 8 %) in sodium and potassium fluorides and in potassium hexafluorotitanate (ca. 14 %) [88]. However, attempts to electrolyze the oxide in halide melts (in analogy to aluminum production) only produced lower titanium oxides [89]. Thus only titanium halides are suitable starting materials for molten salt electrolysis.

Titanium tetrafluoride combines with alkali metal and alkaline earth metal fluorides to form complex hexafluorotitanates. These compounds decompose below their melting points with vaporization of titanium tetrafluoride. They can only be used in molten salt electrolysis as components of a low-melting molten salt bath containing other diluent salts [90]. Special problems arise in molten salt electrolysis with graphite electrodes due to the anode effect. Although this can be counteracted by controlled addition of oxides to the salt bath, oxygen pickup by the titanium metal produced must be prevented by using low temperatures, by separating the anode space from the cathode space, and by avoiding turbulence in the electrolyte [91].

Although *titanium tetrachloride* is insoluble in molten alkali and alkaline earth metal chlorides, stable molten electrolytes can be obtained by using lower titanium chlorides. These can be produced directly from titanium tetrachloride in the electrolytic cell, but they react with the chlorine liberated at the anode to reform titanium tetrachloride, which then evaporates. This can be avoided by using a porous, basket-shaped cathode into which the titanium tetrachloride is fed, as the electrolyte (e.g., an $NaCl - SrCl_2$ eutectic) can then be kept free of lower titanium chlorides [92, 93].

The problems of electrolyte composition and the anode effect can be completely solved by using soluble anodes, as in this case no changes to the electrolyte occur and no gas is emitted at the anode, at least at low current densities [94]. Soluble anodes can be made of crude titanium, titanium scrap, and titanium compounds such as titanium carbide, sometimes mixed with titanium monoxide [95, 96] and titanium nitride [97], which can be produced directly from titanium ores.

Semitechnical plants using processes analogous to the production of magnesium from magnesium chloride can operate at 900 °C under argon, using $TiCl_4$ and NaCl as starting materials. The NaCl fed into the cathode space is decomposed, and the sodium liberated reduces the $TiCl_4$ (which is fed in later) stepwise via $TiCl_3$, $TiCl_2$, and TiCl to give titanium. The energy consumption is only 40 % of that required for the reduction of $TiCl_4$ with magnesium (ca. 43 kW h/kg) [98].

All electrolytic processes yield a very pure metal (> 99 % titanium), but have the disadvantage that the titanium is produced at the cathode in the form of poorly adhering dendritic crystals. The cathodes must be lifted out of the molten electrolyte and cooled in an inert atmosphere. The titanium metal is then scraped off and removed from the electrolytic cell through an air lock.

A production plant, now shut down, to produce titanium sponge electrolytically operated at 520 – 650 °C with an electrolyte consisting of a mixture of LiCl and KCl of approximately eutectic composition and with $TiCl_4$ as the raw material. The titanium concentration in the electrolyte is ca. 1.5 %. The $TiCl_4$ is fed through a steel tube into the molten electrolyte in a metal crucible. The graphite anode is surrounded by a wire mesh coated with cobalt or nickel. Metallic titanium is deposited on steel cathodes. After size reduction, dissolution of the salt in 0.5 % hydrochloric acid, washing, and drying, the titanium yield is ca. 98 % [99].

A number of attempts have been made to produce titanium by electrolytic processes on an industrial scale but none have led to permanent application [76].

8. Processing and Reuse of Scrap Metal

With increasing titanium production, the use of titanium scrap as a raw material instead of titanium sponge has become increasingly important. Titanium scrap is produced during the production of semifinished products, their processing to finished products, and the scrapping of used equipment. Before remelting, titanium scrap must be pretreated. Adhering scale and other residues must be removed from the surface by abrasive blasting, pickling, or surface grinding. The various alloys must be sorted so that the alloy added to the melt is of consistent quality. Depending on the melting process, the scrap may be mixed with the sponge and compressed, or may be welded together for addition in the form of a consumable electrode. Alternatively, it can be directly added to the melt in small pieces. In another process, ground, hydrogenated titanium scrap is used instead of titanium sponge for melting to obtain titanium.

Untreated titanium scrap is used as an additive to steel, nickel, aluminum, copper, and zinc alloys, and in the production of master alloys such as ferrotitanium.

In the United States, ca. 40 % of titanium consumption is from processed titanium scrap. In Europe, due to the low price of titanium sponge only 10 – 30 % is accounted for by scrap [100, 101].

9. Processing of Titanium Sponge

Melting Processes. Titanium reacts in the molten state with air and with all the conventional crucible materials. Melting is therefore

only possible under vacuum or in an inert gas atmosphere in cooled metal crucibles of high thermal conductivity, usually copper. The crucible walls are cooled with water or molten sodium–potassium alloy. The energy source is chiefly an electric arc, although to an increasing extent the metal is melted by a plasma, by electron beams, or, on the laboratory scale, by induction.

On the industrial scale, vacuum arc furnaces are generally used, with consumable electrodes made of titanium sponge, titanium scrap, and alloy components. Vacuum arc furnaces with rotating water-cooled copper electrodes are only used for first stage melting.

To produce consumable electrodes, titanium sponge, sometimes mixed with scrap and alloying components, is compressed to form electrode pieces, which are welded into electrodes, by plasma or MIG welding. Large pieces of scrap can also be welded together to form electrodes.

The melting of the electrode (negative pole) takes place when the arc is formed with the bottom of the crucible under argon or a vacuum of ca. 1 Pa (10^{-2} mbar), with currents of 15 – 40 kA and an average melting capacity of 1 kg/kWh. The solidified titanium is remelted in a similar manner, as this gives improved homogeneity. For special applications involving critical service conditions, the titanium is melted three times.

Vacuum arc with solid negative electrodes [102] and cold-hearth melting furnaces with electron-beam (EBCHM) [103] or plasma heating (PCHM) [104] do not require consumable electrodes. The sponge and scrap in lump form are melted by feeding them continuously into the molten metal. These furnaces have so far only been used for the first melt. In cold-hearth melting, sponge and processed scrap are melted in a pre-crucible. The melt is fed into the mold through a channel, whereby impurities can be retained in the skull – the solidified surface forming on the melt in the channel. This method is expected to give better retention of impurities from the scrap. The mold can be designed for producing ingots or slabs. Multiple burners are employed to obtain a continuous ingot or slab. The crucible is equipped with a withdrawable bottom, and the crude ingot is melted to form a strand which is lowered at a rate corresponding to the melting process.

The commercial crude ingots produced by melting can have a mass of 1 – 13 t, but are usually in the range 3 – 8 t. Diameters are in the range 500 – 1100 mm.

Casting. When arc furnaces with consumable or solid electrodes are used [105], the titanium drains into water-cooled tilting copper crucibles, or the melt is superheated by electric current under an inert gas. The metal is cast in graphite, ceramic, or metallic molds, or by centrifugal casting. Electron-beam furnaces can also be used [106]. Depending on the process, the mass of a casting can range from 750 kg to a few grams in the case of high precision castings with small wall thicknesses produced by the lost wax process. The output of titanium castings is barely 1 % of the total output of semifinished products. In a new melting process a combination of cold-wall crucible and plasma burner is used.

Powder Metallurgy. The production of components from titanium and titanium alloys by powder metallurgy has so far not achieved industrial importance. Production of powder by sieving the fine material from titanium sponge or by supersaturating the sponge with hydrogen, milling under inert gas, and dehydrogenation [107] are not widely used because of the poor flow properties of the powder obtained. These properties can be improved by partial melting of the powder to give spheroidal particles.

Spheroidal powder with good flow properties and good densification behavior can also be produced by an atomization process in which titanium rods are rotated at high speed in an electric arc or plasma [108].

Advances in cold-wall crucible technology have now made possible inert gas atomization with argon or helium in a ceramic-free process [109]. Fine, spheroidal titanium powder can be produced by the induction drip-melting atomization process and by the electrode-induction-melting, gas-atomization process.

In the calciothermic production of titanium alloys in powder form, the starting materials are the metal oxides, CaO, Ca, and $KClO_4$. Fine powders with good densification properties (60 %) and with oxygen contents of 0.14 – 0.18 % and carbon contents of 0.03 – 0.06 % [110] are obtained.

Table 5. Composition and strength properties of commercial titanium alloys at room temperature (minimum values)

Alloy∗	Alloy type	Tensile strength, MPa	0.2 % Yield strength, MPa	Density, g/cm^3	Modulus of elasticity, GPa	Main requirement
Ti 5 Al 2.5 Sn	α	830	780	4.48	110	high strength
Ti 6 Al 2 Sn 4 Zr 2 MoSi	near α	900	830	4.54	114	high temperature strength
Ti 6 Al 5 Zr 0.5 MoSi	near α	950	880	4.45	125	high temperature strength
Ti 5.8 Al 4 Sn 3.5 Zr 0.7 Nb 0.5 Mo 0.2 Si 0.05 C	near α	1030	910	4.55	120	high temperature strength
Ti 6 Al 4 V	$\alpha+\beta$	900	830	4.43	114	high strength
Ti 4 Al 4 Mo 2 Sn	$\alpha+\beta$	1100	960	4.60	114	high strength
Ti 6 Al 6 V 2 Sn	$\alpha+\beta$	1030	970	4.54	116	high strength
Ti 10 V 2 Fe 3 Al	near β	1250	1100	4.65	103	high strength
Ti 15 V 3 Cr 3 Sn 3 Al	β	1000	965	4.76	103	high strength
Ti 3 Al 8 V 6 Cr 4 Zr 4 Mo	β	1170	1100	4.82	103	high corrosion resistance
Ti 15 Mo 3Nb 3 AlSi	β	1030	965	4.94	96	high corrosion resistance

∗ Figure before chemical symbol denotes wt %.

10. Production and Processing of Semifinished Products

Semifinished products are chiefly produced by pressing, forging, rolling, or drawing.

Hot Forming. In hot forming, the temperature of the metal is above its recrystallization temperature. A slightly oxidizing atmosphere and a uniform temperature are maintained in the reheating furnace to prevent hydrogen pickup and consequent embrittlement. In hot rolling, temperatures are usually 20 – 100 °C below those used for forging.

For pressing or forging heavy section components, a preforming stage is carried out in the β-range, and the final is then carried out, usually in the $(\alpha+\beta)$-range sufficiently below the $(\alpha+\beta)/\beta$-interphase, with deformations of at least 60 – 75 % to produce the desired structure and properties of the material.

Cold Forming. In cold forming to produce wire, thin sheet, and foil, the high yield strength of titanium, which increases with increasing tensile strength, and the low elongation without necking must be taken into consideration. Commercially pure titanium should be work annealed after 30 – 40 % deformation.

Heat Treatment after Deformation. Low-strength titanium (Grade 1, DIN designation 3.7025) generally does not require stress-relief annealing, but annealing is necessary for higher strength and especially complex components.

Recrystallization annealing to give a ductile product is required after hot forming of semifinished and drop forged products, and for work annealing after forming of sheet.

Descaling and Pickling. Titanium is descaled by abrasive blasting followed by pickling in aqueous mixtures of 20 – 40 % HNO_3 and 1 – 3 % HF, in sodium hydroxide-based salt baths with added oxidizing agents such as sodium nitrate and borax at 450 – 510 °C, or in sodium hydride baths at 380 °C. Hydrogen pickup can occur if the pickling time is too long.

Processing of Semifinished Products. Of the noncutting forming processes, deep drawing can be used with titanium-based materials, but stretch-forming is less suitable. In cutting, stamping, pressing, and hole punching, titanium behaves like steel. To give crack-free bending, folding, and edging, radii should be large and deformation rates small. The use of load-bearing lubricants and plastic coatings gives improved drawing behavior [111].

In machining, the low thermal conductivity, low specific heat capacity, low modulus of elasticity, and high toughness lead to considerably higher temperatures in the contact zone than with steel under similar cutting conditions. Also, because of the small extent of chip compression, the contact zone is restricted, leading to additional thermal stress on the tip of the tool. For

low heat generation at the tool a high cutting speed, a low chip removed rate, and sufficient cooling should be provided.

Chemical and electrochemical machining of mechanically machined surfaces are widely used processes, and give high accuracy. Hydrofluoric acid, usually with additions of nitric or chromic acid, is used for chemical machining, while electrochemical machining is carried out with direct current in a solution of NaCl.

Welding. In principle, titanium of all strength grades is weldable. However, it can only be welded with exclusion of air or in an inert gas atmosphere because hot solid or molten titanium reacts readily with the atmosphere, reducing toughness and causing embrittlement, even with low levels of pickup. Electron-beam welding is limited to special applications, particularly the aerospace industries, but MIG, TIG, plasma, and laser-beam welding are widely used (→ Welding and Cutting). Friction and diffusion welding can also be employed with titanium.

Soldering is only possible if a solderable coating is first applied to titanium. For brazing, the dense passive layer on the titanium must first be removed by a flux, which usually consists of alkali metal halides with addition of AgCl. Suitable solders include silver, titanium – silver, and aluminum – silver alloys. It is preferable to solder in a vacuum or in an inert gas atmosphere to minimize diffusion of atmospheric gases into the metal [112]. Soldered joints can have a tensile strength of $> 800\,\mathrm{N/mm^2}$ and a shear strength of ca. $150\,\mathrm{N/mm^2}$.

Adhesives. Adhesive bonds with shear strengths of $> 200\,\mathrm{N/mm^2}$ can be produced if the titanium surface is pretreated and anodically oxidized or coated with metals by spraying or vapor deposition [113, 114].

Explosive Cladding. Industrial low-strength grades of titanium can be applied to metallic substrates, usually steel, by explosive cladding at ca. 5000 MPa, giving shear strengths of ca. $300\,\mathrm{N/mm^2}$. To prevent diffusion between the parent metal and the coating leading to embrittlement of the bonding zone, the upper limit of the usage temperature should be 550 °C [115]. Other substrate metals such as aluminum, copper, and zirconium can also be explosively cladded.

Super Plastic Forming (SPF). Alloyed titanium with a two-phase fine structure often exhibits ductility properties that allow ca. 1000 % elongation. In combination with diffusion welding, highly complex structural components can be produced in a single operation [116].

Tribology. In general, the surface of titanium components is not very resistant against wear. Therefore, wear resistant films are formed by ion implantation or by PVD or CVD. In-situ formation of titanium nitride, carbide, or boride films of varying thickness is achieved by annealing under the corresponding atmosphere or by laser melting [117–121].

11. Titanium Alloys

The main aim of adding alloying elements to titanium is to improve its mechanical properties. The most important of these alloys are listed in Table 5.

The increase in corrosion resistance caused by addition of 0.15 % Pd and traces of Ni and Mo has been mentioned in Chapter 4. Titanium master alloys are discussed in Chapter 13.

11.1. α-, $(\alpha+\beta)$-, and β-Alloys

The transformation temperature of titanium from the α- to the β-phase can be raised or lowered by adding alloying elements which can stabilize the α- or β-phase. Oxygen, nitrogen, and carbon, which form interstitial mixed crystals, and aluminum, which dissolves as a substitution mixed crystal, stabilize the α-phase. Elements that stabilize the β-phase include hydrogen, which forms an interstitial mixed crystal, the alloying elements vanadium and molybdenum, and iron, chromium, copper, palladium, and silicon, while zirconium and tin are highly soluble in both phases.

The alloys are classified as α-, $(\alpha+\beta)$-, and β-titanium alloys according to the phases present in the annealed state at room temperature.

As the cubic structure of β-titanium contains considerably more slip planes than the hexagonal α-form, β-titanium is more easily deformed. The alloys in the β- and $(\alpha+\beta)$-regions are therefore hot formed. β-Alloys and some $(\alpha+\beta)$-alloys contain, in addition to molybdenum or vanadium, additives whose β-phase at temperatures in the range between $< 550\,°C$ and up to a maximum of $860\,°C$ undergo eutectoid decomposition into the α-phase and an intermetallic compound. This process is usually associated with embrittlement and reduces thermal stability and must be avoided during production and processing.

The critical alloying content of β-stabilizing elements in two-component alloys in which the β-phase is preserved on quenching to room temperature and the eutectoid data for alloys showing eutectoid decomposition are listed in Table 6. Also, many two- and multi-component phase diagrams can be found in [122].

Table 6. Eutectoid temperature and critical alloy content of titanium binary alloys (concentrations in wt %)

	Eutectoid composition	Eutectoid temperature, °C	Critical alloy content
Manganese	17	550	6.4
Iron	17	595	3.5
Chromium	15	670	6.3
Cobalt	9.5	685	7.0
Nickel	6.2	765	9.0
Copper	7	790	13.0
Silicon	0.65	865	
Vanadium			14.9
Molybdenum			11.0
Niobium			36.7
Tantalum			45.0

The equilibrium state is established very slowly for many alloy systems at lower temperatures. For systems showing eutectoid decomposition of the β-phase, the process proceeds more slowly at lower temperatures and with higher alloy contents of the eutectoid. This is especially noticeable in multicomponent alloys [123].

Hardening of titanium alloys can be carried out in various ways: An intermetallic compound can be precipitated from the α-phase by quenching and annealing (e.g., with Ti – Cu alloys). In alloys with β-stabilizing alloying elements, the supercooled metastable β-phase can be formed in addition to the α-phase by solution annealing, quenching, and aging. On aging, the former is transformed via an intermediate ω-phase into the α-phase and to β-phase stabilized by enrichment of the alloying elements. Both these processes are used in industry to improve the mechanical properties of titanium alloys.

Mechanical Properties. The strength properties of the most important titanium alloys are listed in Table 5.

The α-alloys, which have a strength range of 830 – 1030 MPa, are used as forging alloys, and also in the production of sheet metal because of their good welding properties. Ti 5 Al 2.5 Sn has good high-temperature tensile strength. The near α-alloys are members of the group of high-temperature titanium alloys, used in aircraft engines.

The $(\alpha+\beta)$-alloys include the most common titanium alloy Ti 6 Al 4 V. Its strength of 900 MPa in the annealed state can be in-creased by ca. 200 MPa by hardening. Strengths exceeding 1000 MPa with better through hardenability can be obtained with the alloys Ti 6 Al 6 V 2 Sn and Ti 4 Al 4 Mo 2 Sn.

β-Alloys with a strength of 1000 – 1200 MPa in the annealed state include Ti 15 V 3 Cr 3 Sn 3 Al, Ti 3 Al 8 V 6 Cr 4 Zr 4 Mo and Ti 15 Mo 3 Nb 3 AlSi. These are mainly used as sheet alloys in aircraft construction [124].

The through hardening properties of the alloys Ti 6 Al 6 V 2 Sn and Ti 4 Al 4 Mo 2 Sn are superior to those of the usual commercial titanium alloys. These alloys are also preferable to Ti 13 V 11 Cr 3 Al, as they have better figures at > 100 mm, the latter alloy tending to have a coarse-grained structure and poor toughness properties.

The fatigue strength of titanium alloys depends on the section of the semifinished product and its structure. In the unnotched state, it is 30 – 50 % of the tensile strength, and in the notched state 30 %. The fracture toughness of metal with a Widmannstätten structure reaches higher values than that of metal with a polygonal structure [125].

The rate of decrease of high-temperature tensile strength of alloys with increasing temperature up to ca. 250 °C is lower than that of commercially pure titanium, and the creep strength figures are higher.

The ratio yield point is > 90 %. As the elongation without necking is small, these materi-

als are usually hot formed. Solution annealing and stabilization give improved properties, e.g., increased thermal stability and improved creep properties.

In Germany, materials standards for titanium and titanium alloys are given in a number of DIN standards, materials performance sheets (Werkstoffleistungsblätter, WL), and AECMA/EN standards. In France, this information is provided by AFNOR (Association Française de Normalisation) and NORME AIR of the Ministère de la Défense, in the United Kingdom by the British Standards Institution (BSI) and the DTD of the Ministry of Defence, and in the United States by ASTM, ASME, AMS, and MIL of the Air Force Materials Laboratory.

Physical Properties. In the commercial α- and $(\alpha+\beta)$-titanium alloys, the density at 25 °C is in the range 4.37 – 4.56 g/cm^3. In β-titanium alloys, values of 4.94 g/cm^3 are reached. Moduli of elasticity and rigidity increase with increasing content of interstitial elements and aluminum, and with annealing temperature [126]. With additions of β-stabilizing additives, they decrease in comparison with pure titanium, as hardening increases [25]. With both pure titanium and commercial titanium alloys they decrease with increasing temperature [23]. At 25 °C titanium alloys usually have electrical resistivities of ca. $150\times10^{-6}\,\Omega$ cm. The linear coefficient of expansion at 25 °C for α- and $(\alpha+\beta)$-titanium alloys is ca. $9.5\times10^{-6}\,K^{-1}$. The specific heat capacity is similar to that of titanium, and is somewhat higher for $(\alpha+\beta)$-titanium alloys. The thermal conductivity is only half that of titanium.

Chemical Resistance. Additions of noble metals and other metals such as molybdenum, zirconium, hafnium, nickel, tantalum, or niobium improve the anticorrosion properties of titanium, while additions of iron, chromium, and aluminum, like high levels of oxygen, nitrogen, and hydrogen, reduce corrosion resistance [51].

The corrosion behavior of commercial titanium alloys is very similar to that of the commercially pure metal. Unlike titanium containing 0.15 % Pd, alloys containing 2 % Ni or 15 – 30 % Mo [127] have not achieved industrial importance, in spite of their superior corrosion resistance under reducing conditions.

With increasing aluminum content, the susceptibility of titanium alloys to stress corrosion cracking at >200 °C increases [128].

11.2. Highly Alloyed Titanium

Intermetallic Ti – Al Alloys. Three intermetallic titanium aluminides are known: Ti_3Al, TiAl, and $TiAl_3$ (Fig. 4). In particular, the phases α_2-Ti_3Al (hexagonal) and γ-TiAl (tetragonal) exhibit low density and good high-temperature strength and were therefore chosen for development of construction materials. The main aims are increasing the room-temperature ductility of these brittle phases and improving the corrosion resistance in the temperature range 500 – 900 °C [130, 131].

The α_2-Ti_3Al alloys were developed in the United States and first used commercially in the 1990s. They mainly replace conventional titanium alloys, allowing applications above 500 °C. The ductility is increased by alloying with 8 –18 wt % Nb. Other possible alloying elements are Mo, V, Ta, and Ni. In addition to Ti_3Al, the phases β-NiAl (cubic) and Ti_2AlNb can also be present [132]. Characteristic data of α_2-Ti_3Al can be found in [133]. α_2-Ti_3Al is produced by melting in vacuum arc furnaces.

γ-TiAl alloys have been under development for around ten years and are now on the brink of commercial application. They can be used up to 850 °C, and with their density of 3.9 g/cm^3, they could potentially replace nickel superalloys in aircraft and turbines [134, 135]. Another potential application is as a valve material for combustion engines [136]. γ-TiAl alloys consist of two phases: γ-TiAl (tetragonal face-centered $L1_0$ structure) and a small amount of α_2-Ti_3Al (hexagonal DO_{19} structure) [137]. The aluminum content is 45 – 48 % Cr. H, Mn, and V are added to increase the room-temperature ductility, while Nb, Ta, Mo, and W improve the high-temperature creep resistance and corrosion resistance. Si, B, and C can be used as grain refiners. The total content of alloying elements is usually in the range 2 – 7 % [137].

Two-phase γ-TiAl materials can be produced by powder metallurgy [138] or by melting in vacuum arc furnaces.

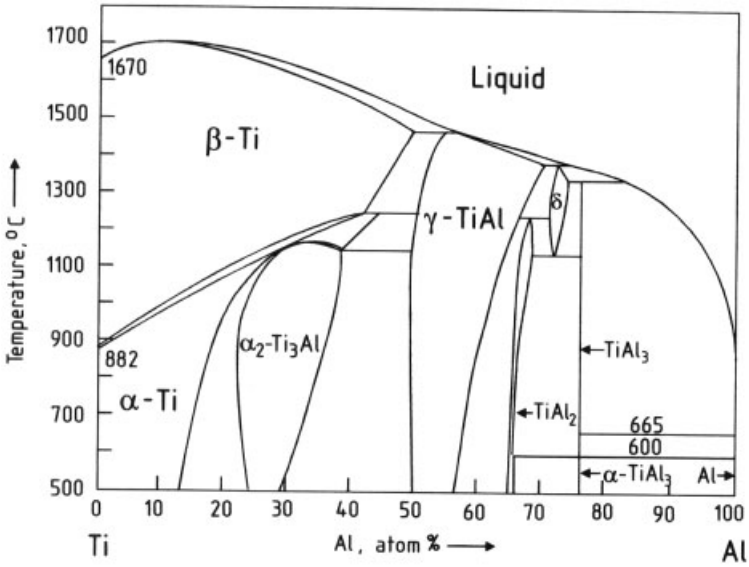

Figure 4. Ti – Al phase diagram [129]

Table 7. Properties of titanium aluminides, titanium-base conventional alloys, and nickel-base superalloys ∗

Property	Conventional titanium alloys	Ti_3Al	TiAl	Nickel-base superalloys
Density, g/cm^3	4.5	4.1 – 4.7	3.7 – 3.9	8.3
Modulus, GPa	96 – 117	100 – 145	160 – 176	206
Yield strength, MPa	380 – 1150	700 – 990	400 – 450	
Tensile strength, MPa	480 – 1200	800 – 1140	450 – 800	
Creep limit, °C	600	760	1000	1090
Oxidation limit, °C	600	650	900	1090
Ductility, %				
at room temperature	20	2 – 10	1 – 4	3 – 5
at high temperature	high	10 – 20	10 – 60	10 – 20

∗ From *ASM Handbook*, 10th. ed., vol. 2, ASM International, Metals Park, OH, 1990, p. 226.

Table 7 compares the properties of titanium aluminides with those of conventional titanium alloys and nickel superalloys [139].

Shape-Memory Alloys. Ti – Ni alloys containing ca. 50 atom % Ni exhibit an austenite – martensite transformation at a definite temperature. Depending on the type of activation used (working at a certain temperature), either a shape-memory effect is obtained, or superelasticity [140].

Superconductors. Ti – Nb alloys containing ca. 50 atom % Nb are low-temperature superconductors with good ductility, high critical field strength, and high critical current (→ Superconductors, Chap. 5.2.1.). Wires of ca. 5 µm diameter can readily be made into coils [141].

Metal-Matrix Composites. The tensile strength, especially at high temperatures, of titanium alloys can be increased by strengthening the metal matrix with fibers such as SiC [142].

Hydrogen Storage. Titanium alloys with high contents of Fe, Cr, Mn, and V can take up large amounts of hydrogen, which is released under certain conditions of temperature and pressure. They can therefore be used for hydrogen storage, as in gas purification, and for vacuum generation [143].

12. Uses and Economic Aspects

Chemical Industry. In chemical plant, commercially pure and palladium-containing titanium lead to increased lifetime, reduced maintenance, and reduced iron contamination of products. Important applications include petrochemicals and the production and handling of, e.g.,

acetic acid, propionic acid, nitric acid, and chlorine gas. Titanium is used in all types of chemical equipment, including pressurized reactors. The temperature limit for equipment constructed of pure titanium is usually 250 °C, for titanium-clad equipment 200 °C, and for titanium-plated equipment ca. 500 °C.

In the food industry, cooking equipment, stirring devices, and containers are constructed of titanium because of its resistance to a wide range of organic acids. In the cellulose, paper, and textile industries, titanium is important because of its resistance to hypochlorite, chlorite, and chlorine dioxide. In deep-sea research, seawater desalination, and closed-loop cooling water systems in power stations, the good resistance toward salt water even at high flow rates and at temperatures up to 165 °C (e.g., in tube and plate heat exchangers) is utilized. In electrolysis technology and surface finishing, commercially pure grades of titanium, especially the softer grades 3.7025 (Grade 1) and 3.7035 (Grade 2), are used for a wide range of components including anodes (sometimes platinum-coated and cathodes [52, 144]. Offshore and onshore drilling represent a growing demand for titanium.

Aerospace. The largest consumer of titanium alloys is the aircraft industry, in which the nonmilitary share continues to increase. In aircraft engines, the current limit for weight saving by the use of titanium lies at Mach 3.5 and ca. 530 °C. In subsonic engines, the titanium content is 20 – 35 %, and the weight saving achieved is up to 22 % [145].

In airframes, temperatures of 300 °C are seldom exceeded. With increasing flying speed, the weight saving achievable by the use of titanium can reach ca. 17 %. In civilian wide-body jets, the proportion of titanium-based materials in the body of the aircraft can reach 8 %, and in military supersonic aircraft 25 % or more. Unalloyed grades of titanium are used for lightly stressed components, and alloys for highly stressed components.

Other Applications. Applications in the automotive industry include components of high-performance engines and undercarriage made of titanium alloys [146]. Pure titanium is used as a getter in electronic tubes used in the measurement, control, and electrical technologies, X-ray amplifiers, television cameras and transmitter tubes, and in sonar detectors and gyrocompasses. In medicine, endoprostheses, jaw implants, pins for fixing broken bones, plates, heart pacemaker capsules, and surgical instruments are manufactured from titanium and the alloy Ti 6 Al 4 V because of their compatibility with human tissue.

Applications in precision engineering and optics include sonotrodes, rotors for ultracentrifuges, spectacle frames, watches, golf clubs, and film-developing equipment. Titanium has so far not been widely used in steam turbines, although it is suitable. The same applies to its use in underwater exploration and submarine construction.

Economic Aspects. At present, titanium sponge is produced in the United States, Japan, Russia, Kazakhstan, China, and India. World production capacity for titanium sponge in 1994 was 113 300 t. An overview of the producing countries and existing capacities for sponge and ingot production is given in Table 8. Ingot production capacity in Western Europe, the United States, and Japan reached ca. 85 000 t in 1994.

Table 8. Production capacities for titanium sponge and melting capacities in 1994

Country	Capacity, t/a	
	Sponge	Ingots
United States	19 500	57 700
Japan	25 000	22 400
Europe		9 500
CIS	65 000	110 000
China	3 000	3 000
Total	113 300	202 600

Worldwide consumption of titanium semifinished products in 1994 was ca. 45 000 t. European demand in 1994 was ca. 10 000 t, expected to reach 13 300 t in the year 2000.

13. Titanium Master Alloys

13.1. Ferrotitanium

13.1.1. Composition and Uses

Ferrotitanium is described in DIN 17 566 as a master alloy containing at least 28 wt % Ti, obtained by reduction of the corresponding raw

Table 9. Composition of commercial ferrotitanium alloys from DIN 17 566

Alloy grade	Alloy number[a]	Chemical composition, wt %						
		Ti[b]	Total Al max.	Si max.	Mn max.	C max.	P max.	S max.
FeTi 30	0.4530	28 – 32	4.5	4.0	1.5	0.10	0.050	0.060
FeTi 40	0.4540	36 – 40	6.0	4.5	1.5	0.10	0.10	0.060
FeTi 50	0.4550	46 – 50	7.5	4.0	1.0	0.10	0.10	0.060
FeTi 70	0.4570	65 – 75	2.0	0.20	1.0	0.20	0.040	0.030
FeTi 70 VB (vacuum treated)	0.4571	65 – 75	0.50	0.10	0.20	0.20	0.030	0.030

[a] In accordance with DIN 17 007, Sheet 3 (draft);
[b] Variation within a batch must not exceed ± 2 wt %.

materials or their concentrates. The International Standard for ferrotitanium is ISO 5454-1980 (E), which specifies a Ti content of at least 20 wt % and allows greater variation in the Al content (up to 10 wt % Al) than DIN 17 566. The Standards specify not only the chemical composition (see Table 9) but the condition, testing methods (sampling and analysis), shipping, and storage. Ferrotitanium in "normal lump form" usually means lumps up to 25 kg in weight. Ferrotitanium can also be supplied on demand as crushed material and/or sieved as agreed. It is supplied in sealed steel drums.

Melting points and densities of some commercial ferrotitanium alloys are listed in Table 10.

Table 10. Melting points and densities of some commercial ferrotitanium alloys

Composition, wt %			*mp*, °C	ϱ, g/cm^3
Ti	Al	Si		
30.6	4.0	3.7	1325 – 1500	6.2
39.1	5.8	4.0	1325 – 1480	5.9
46.6	7.6	3.2	1305 – 1480	5.5
70.0	0.1 – 0.5		1070 – 1135	5.4

The binary titanium – iron phase diagram is shown in Figure 5 [147].

Applications in the Iron and Steel Industry. Compared with pure titanium, ferrotitanium has the advantages of better solubility (lower melting point and higher density) and lower price. Titanium is useful mainly because of its ability to form a carbide. However, as the affinity of titanium for oxygen is considerably higher than for carbon, the steel must be effectively deoxidized before titanium is added. Titanium is only used for this deoxidation in special cases. Titanium is also used for nitrogen removal and for bonding with sulfur [148].

In the production of stainless austenitic chromium – nickel steels, titanium is added to combine with the small amount of carbon present to form TiC ("stabilization"), thereby preventing intergranular corrosion (e.g., adjacent to welds).

Deoxidation of killed steel with titanium improves its quality by minimizing oxidic inclusions. Depending on the carbon content, it also reduces cracking tendency (grain refinement of austenite) and increases yield strength (titanium-containing ferrite). Titanium addition also minimizes segregation in the upper part of the ingot on casting, thus improving the yield.

Addition of titanium to tool steels gives a finer grain and reduces the depth of hardening and hence the cracking tendency. In sulfur-containing free-cutting steels, especially highly alloyed stainless chrome and chromium – nickel steels, the sulfur is bonded to the titanium. This prevents red shortness and improves ductility.

Some complex master alloys with boron (→ Boron and Boron Alloys, Chap. 2.4.2., → Boron and Boron Alloys, Chap. 2.5.) contain titanium and other elements (BATS = boron – aluminum – titanium – silicon). The boron is added in small amounts to the steel to improve the hardening properties and grain refinement, and the titanium is added to protect the boron from nitrogen (by forming TiN).

High-temperature hardenable alloys (superalloys, → High-Temperature Materials) based on iron and especially nickel (e.g., nimonic alloys) contain up to 4.5 % Ti [149]. Here, the titanium reacts with the aluminum and nickel to form coherent particles with the composition $Ni_3(Ti, Al)$, which cause precipitation hardening. Titanium is also used as a microalloying el-

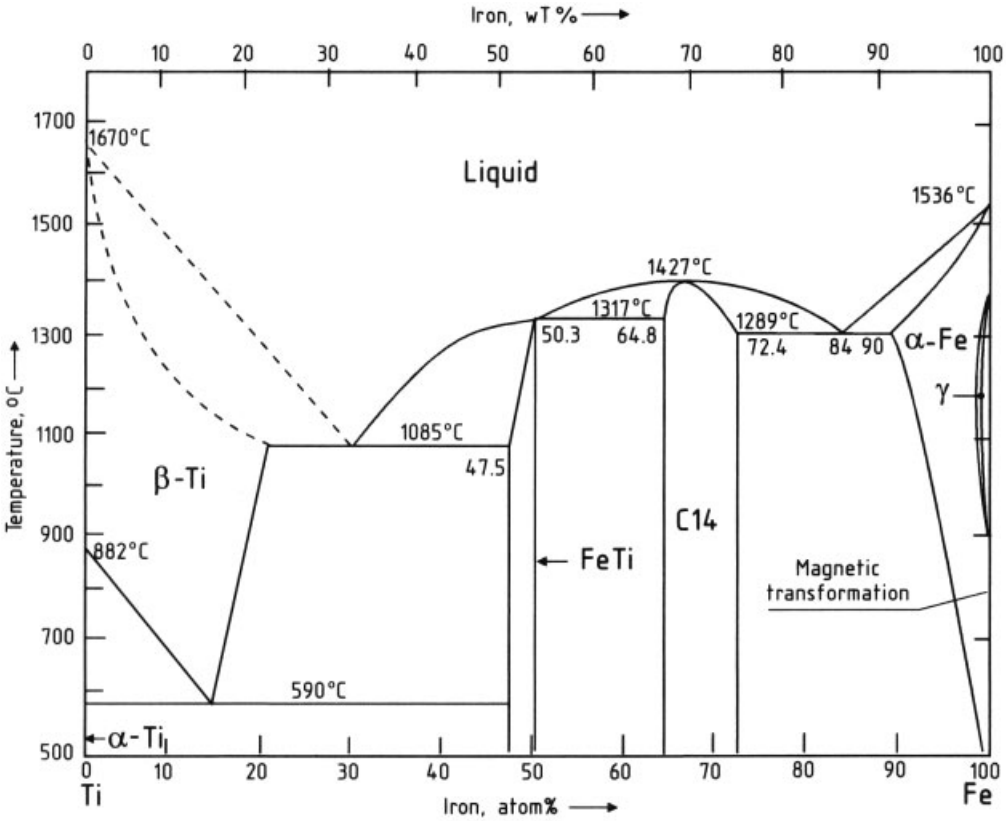

Figure 5. Ti – Fe phase diagram [147]

ement in high-strength, low-alloy (HSLA) steels [150].

The addition of titanium to cast iron causes the formation of finer graphite on supercooling, thereby improving the mechanical properties.

Powdered ferrotitanium and ferrosilicon – titanium alloys, sometimes with higher aluminum contents, are used as coatings on welding electrodes. Also, welding wires for HSLA steels are alloyed with ca. 0.5 % Ti to give a fine-grained weld bead.

13.1.2. Production

The starting materials for the production of ferrotitanium are ilmenite, leucoxene, perovskite, and slag concentrates produced from ilmenite. Because of the increasing availability of titanium scrap, this also is used to an increasing extent for the production of ferrotitanium (mostly > 70 % Ti) and other titanium master alloys.

Reduction with Carbon. Carbothermic production in electric-arc furnaces leads to high carbon contents in the ferrotitanium. However, as the main use of ferrotitanium is to combine with the carbon in steel, the presence of carbon in the ferrotitanium is undesirable, and this production method is therefore now hardly used. The carbothermic production of ferrotitanium containing 15 – 18 % Ti, 5 – 8 % C, 1 – 3 % Si, and 1 – 2 % Al, and of ferrosilicon – titanium containing ca. 30 % Ti, ca. 30 % Si, up to 0.3 % C, and 1 – 3 % Al is described in [151].

Metallothermic Production. DIN 17 566 specifies grades of ferrotitanium containing 28 – 50 wt % Ti and 4.5 – 7.5 wt % Al. These are mainly produced by the aluminothermic process. The reduction of TiO_2 by Al proceeds via TiO.

$$3\,TiO_2 + 2\,Al \longrightarrow 3\,TiO + Al_2O_3$$

$$3\,TiO + 2\,Al \longrightarrow 3\,Ti + Al_2O_3$$

If there is too much TiO_2 in the reaction mixture, TiO can be formed as a third phase besides the metal and slag [152]. In the aluminothermic production of ferrotitanium there is a high consumption of aluminum, as it reacts both with the iron oxide in the ilmenite and with the oxygen-producing substances added to increase the exothermicity.

An aluminothermic mixture, consisting, for example, of 4320 kg Australian ilmenite (58.55 % TiO_2), 480 kg rutile (96.7 % TiO_2), 190 kg calcined limestone, 107 kg potassium perchlorate, and 1693 kg Al powder, is placed in a refractory-lined combustion vessel and ignited to start the reaction. After cooling, a 2250 kg block of ferrotitanium is obtained containing 39.8 % Ti, 6.7 % Al, 3.4 % Si, and 0.02 % C, which separates well from the slag. The titanium yield is 50 %, and the specific consumption of aluminum (based on the Ti in the ferrotitanium) is 1.89 kg Al/kg Ti.

In the electro-aluminothermic process, which is carried out in an arc furnace, the raw materials contain as little iron oxide as possible, so that most of the aluminum is used to reduce titanium oxide. The consequent lack of exothermicity is compensated by electrical energy. A process developed in France is now widely used [153]. Here, an ilmenite concentrate, after addition of lime, is partially reduced with carbon in an electric-arc furnace to give a slag concentrate with a high concentration of titanium oxide, and the reduced iron is tapped off. The molten slag is then reduced in several stages by blowing in aluminum and iron, producing ferrotitanium of the desired composition.

In a process developed by the Gesellschaft für Elektrometallurgie [154], a calcium oxide/titanium oxide melt containing ca. 75 % TiO_2 is first produced from prereduced ilmenite, from rutile and lime, or from perovskite with addition of iron scrap, and is reacted with a molten aluminum – iron alloy containing 60 – 70 % Al in a reaction ladle to give ferrotitanium and slag. Yields of titanium of 60 – 67 % are obtained, and the specific aluminum consumption is 1.0 – 1.1 kg Al/kg Ti.

Production from Titanium Scrap and Sponge. As the availability of titanium scrap is increasing with the growth in the consumption of titanium, it is being increasingly used for the production of ferrotitanium (mainly high-grade material containing 65 – 75 % Ti) and ferrotitanium – silicon. Because of its low melting point (ca. 1100 °C), the alloy containing 70 % titanium melts comparatively readily when alloyed with iron in an induction furnace, in an arc furnace with consumable electrodes under vacuum or argon, or in an electroslag melting furnace. Contact of the melt with air must be prevented because of the strong affinity of titanium for oxygen and nitrogen. Ferrotitanium can also react with refractory crucibles [155]. The high-grade ferrotitanium alloys produced by melting under vacuum are designated FeTi 70 VB in DIN 17 566.

Calcium – silicon – titanium alloys are produced by melting together commercial calcium silicon and titanium sponge or scrap in an induction furnace [151, p. 579].

13.1.3. Economic Aspects

The prices of ferrotitanium are published in the London "Metal Bulletin", and the monthly average prices (from the Metal Bulletin), with the corresponding exchange rates, have been published since 1979 in the journal *Erzmetall* [156].

13.2. Titanium – Aluminum

The alloys listed in Table 11, used as master alloys for alloying aluminum with titanium and as grain-refining agents, are standardized in DIN 1725. The master alloys containing 5 – 10 wt % Ti (known in the United Kingdom as titanium – aluminum intermediate hardeners) are produced by melting aluminum and titanium scrap or sponge in crucible or drum furnaces heated by gas or oil, or in induction furnaces. They are cast into pigs, or grooved ingots or plates, and are marketed in these forms.

The alloy V-AlTi 5 B 1, which is used for grain refining aluminum, contains titanium boride (TiB_2) in the form of a very fine precipitate which acts as seed crystals [157]. The alloy is also supplied as wire for continuous grain refinement. It is produced by adding potassium titanium fluoride and potassium borofluoride to a bath of molten aluminum in an induction furnace.

Table 11. Titanium – aluminum alloys according to DIN 1725, Sheet 3

Alloy grade	Alloy number	Composition, wt %			Applications
		Alloy constituents	Other permitted constituents		
V-AlTi 5	3.0851	Ti 4.5 – 6.0 Al to 100	Cr 0.10 Cu 0.15 Fe 0.45 Mg 0.50 Mn 0.35 Ni 0.10 Pb 0.10	Si 0.50 Sn 0.10 V 0.25 Zn 0.15 others : each 0.05 total 0.15	alloying with Ti grain refinement
V-AlTi 10	3.0881	Ti 9.0 – 11.0 Al to 100	Cr 0.02 Cu 0.02 Fe 0.30 Mg 0.02 Mn 0.02 Ni 0.04	Si 0.20 V 0.30 others : each 0.05 total 0.15	alloying with Ti grain refinement
V-AlTi 5 B 1	3.0861	Ti 5.0 – 6.2 B 0.9 – 1.4 Al to 100	Cr 0.02 Cu 0.02 Fe 0.30 Mg 0.02 Mn 0.02 Ni 0.04 Si 0.20	V 0.20 Zn 0.03 Zr 0.02 others : each 0.05 total 0.10	grain refinement

Titanium – aluminum containing 50 – 60 % titanium can also be produced by the aluminothermic reaction of rutile with aluminum powder with addition of boosters such as peroxides or chlorates, or by the electro-aluminothermic method in an arc furnace. Today, these high-titanium alloys, which are mainly used in powder form as components of welding wire coatings, are also produced by melting together aluminum and titanium in an induction furnace.

13.3. Other Master Alloys

Special master alloys such as copper – titanium (26 – 35 wt % Ti) and nickel – titanium (25 – 70 wt % Ti) are produced by melting titanium sponge or titanium scrap with copper or nickel. These master alloys are used as additives (0.1 – 1.5 % Ti) to copper, nickel, and monel metal to grain refine the structure, improve the casting and welding properties, and increase the transverse strength and elasticity. A small addition of titanium (0.15 – 0.25 %) to zinc alloys containing up to 1.5 % Cu increases the yield strength and fatigue strength [158].

The high coercive field strength and low remanence of AlNiCo permanent magnets are due to the addition of 0.5 – 8 % titanium in the form of AlTi of CoAlTi master alloys (→ Magnetic Materials, Chap. 4.1.1.).

Multicomponent, NiTi-based alloys are employed in powder metallurgy for manufacturing structural materials.

14. Titanium Compounds [159, 160]

Titanium Carbide → Carbides, Chap. 2.2.

Titanium Dioxide → Pigments, Inorganic, Chap. 2.1.

Titanium is usually tetravalent in its compounds, but may also be trivalent and, in a few compounds, bivalent; the compounds of bivalent titanium are prepared only with difficulty and are unstable in aqueous solution. The existance of titanium(I) compounds has not been established, although certain observations point to their formation, e.g., the existence of maximum volatility in the Ti – S system at a composition corresponding roughly to the formula Ti_2S.

Tetravalent titanium has a strong tendency to form anions such as $[TiO_3]^{2-}$ (titanate ion), $[TiF_6]^{2-}$ (fluorotitanate ion), $[TiCl_6]^{2-}$ (chlorotitanate ion), and $[Ti(SO_4)_3]^{2-}$ (sulfatotitanate ion). The soluble titanium(IV) com-

pounds tend to undergo hydrolytic decomposition. Incomplete hydrolysis may give rise to compounds to the type $TiOX_2$. Compounds containing the radical $[TiO]^{2+}$ are called titanyl compounds.

The oxidation potential Ti^{3+}/Ti^{4+} (relative to the normal hydrogen electrode at 0 °C) is +0.04 V. For the couple Ti^{2+}/Ti^{3+} the oxidation potential $E^0 = +0.37$ V, and for the couple Ti^{2+}/Ti^{4+}, $E^0 = +0.20$ V. The normal potential of Ti in contact with a Ti(II) salt solution is +1.75 V.

14.1. Titanium(II) Compounds

Compounds of bivalent titanium can be prepared by reduction of titanium(IV) or titanium(III) compounds. Thus titanium(II) chloride, $TiCl_2$ is produced from titanium tetrachloride by means of sodium amalgam, as a black powder which is slowly decomposed by water with the evolution of hydrogen. Aqueous solutions which contain titanium(II) chloride together with titanium(III) chloride can be prepared by dissolving TiO in cold dilute hydrochloric acid. Ti^{2+} ions are rapidly oxidized to Ti^{3+} ions by water at normal temperature; the solutions are more stable at lower temperatures. In the pure state, $TiCl_2$ is best obtained by thermal decomposition of $TiCl_3$ or by heating $TiCl_4$ with Ti shavings. $TiBr_2$ and TiI_2 can be prepared by direct union of their components.

TiO, ϱ 4.94 g/cm^3, *mp* 175 °C, can be obtained by heating a mixture of TiO_2 and Ti. It has the NaCl type crystal structure, but a certain proportion of the lattice sites (distributed at random) is vacant. At the one end of the homogeneity range, these vacant sites represent missing Ti atoms, and at the other extreme O atoms.

14.2. Titanium(III) Compounds

Titanium(III) compounds are obtained by reducing soluble titanium(IV) compounds with zinc and acid or electrolytically. The solutions contain violet titanium(III) ions; these have a strong tendency to revert to titanium(IV) ions. The ability of trivalent titanium to bring about indirect oxidation by means of oxygen is so great that even water can play the part of an acceptor in the presence of titanium(III) compounds in alkaline solution, and is thereby converted to hydrogen peroxide:

$$Ti(OH)_3 + {}^1/_2O_2 + H_2O \rightarrow TiOH_4 + {}^1/_2H_2O_2$$

Titanium(III) Halides. Anhydrous is obtained as a violet powder by passing titanium tetrachloride vapor and an excess of hydrogen through a red hot tube. When it is heated in hydrogen to about 700 °C, $TiCl_3$ disproportionates into $TiCl_2$ and $TiCl_4$. Titanium(III) chloride is obtained in solution by reducing a solution of titanium(IV) salt in hydrochloric acid by means of zinc, or by dissolving metallic titanium in hydrochloric acid. A violet hexahydrate, $TiCl_3 \cdot 6\,H_2O$, crystallizes from the solution.

Titanium (III) bromide, $TiBr_3 \cdot 6\,H_2O$, forms reddish violet crystals (*mp* 115 °C) which are readily soluble in alcohol and acetone, as well as in water. Anhydrous $TiBr_3$ exists in two modifications. It decomposes reversibly at 400 °C, according to:

$$2\,TiBr_3 \rightleftarrows TiBr_2 + TiBr_4$$

Titanium(III) fluoride, TiF_3, can be obtained as a dark blue powder by treating Ti or $TiCl_3$ with HF. It is very unreactive and begins to sublime at 900 °C in a vacuum.

Titanium(III) sulfates. In the electrolytic reduction of a sulfuric acid solution of titanium(IV) sulfate, a black-violet solution is first obtained. The solution becomes lighter with continued reduction, and when practically all the titanium is converted to the trivalent state, it has a transparent pure violet color. An acid titanium(III) sulfate, with the composition $3\,Ti_2(SO_4)_3 \cdot H_2SO_4 \cdot 25\,H_2O$ can be isolated from the solution in the form of a crystalline violet powder. This can be converted to anhydrous neutral titanium(III) sulfate, $Ti_2(SO_4)_3$, by fuming with sulfuric acid.

Neutral titanium(III) sulfate is a green crystalline powder, insoluble in water but soluble in dilute acid to give violet solutions. Exchange of hydrogen in acid titanium(III) sulfate by metal ions gives rise to double or complex salts,

with the composition $3\,Ti_2(SO_4)_3 \cdot M^I_2SO_4$, or $M^ITi_3(SO_4)_5$ and variable water content, e.g., the ammonium salt $NH_4Ti_3(SO_4)_5 \cdot 9\,H_2O$, which forms blue crystals, sparingly soluble in water. Another type of double sulfate of trivalent titanium corresponds to the alums in composition and crystalline form. The only known titanium alums are the rubidium and cesium compounds, e.g., $RbTi(SO_4)_2 \cdot 12\,H_2O$ (red-violet). These salts can be recrystallized without decomposition from dilute sulfuric acid, but not from water. Yet another type is represented by the sodium double sulfate, a violet crystalline mass with the composition $NaTi(SO_4)_2 \cdot 2.5\,H_2O$.

Titanium(III) Hydroxide and Oxide. *Titanium(III) hydroxide* is formed by reaction of titanium(III) salt solutions with alkali hydroxides as a deeply colored precipitate which acts as a vigorous reducing agent and is therefore difficult to obtain pure.

Dititanium trioxide, Ti_2O_3, ϱ 4.49 g/cm^3, *mp* ca. 1900 °C, can be obtained crystalline by heating titanium dioxide at 1000 °C in a stream of hydrogen and titanium tetrachloride. It has the same crystal structure as corundum.

14.3. Titanium(IV) Compounds

Titanium(IV) Halides.

Titanium tetrachloride [*7550-45-0*], $TiCl_4$, M_r 189.73, *mp* − 23 °C, *bp* 136.5 °C, ϱ 1.76 g/cm^3 at 0 °C, can be obtained by passing chlorine over heated titanium, titanium carbide, or over a mixture of titanium dioxide and carbon. In the pure state, it is a colorless liquid which has a pungent smell, fumes strongly in moist air, and is rapidly hydrolyzed by water:

$$TiCl_4 + H_2O \rightarrow TiO_2 + 4\,HCl$$

If the hydrolysis is supressed by addition of acid or if only a limited amount water enters into reaction, oxychlorides can appear as intermediate products. Titanium tetrachloride adds chloride ions to form the complex chlorotitanate ion $[TiCl_6]^{2-}$, e.g., the ammonium salt $(NH_4)_2[TiCl_6] \cdot 2\,H_2O$ (yellow crystals). The corresponding free acid can only exist in aqueous solution; its formation therein is demonstrated by yellow coloration produced by the addition of concentrated hydrochloric acid to titanium tetrachloride.

Titanium tetrafluoride [*7783-63-3*], TiF_4, M_r 123.90, *mp* > 400 °C, ϱ 2.80 g/cm^3 is prepared by reaction of $TiCl_4$ with HF, forms a white powder that sublimes at 284 °C. It displays a much stronger tendency than does the chloride to form titanates. These correspond to the type $M^I_2[TiF_6]$ (fluorotitanates). All the alkali and alkaline earth salts of this type, as well as numerous salts of the heavy metals have been prepared.

Titanium tetrabromide [*7789-68-6*], $TiBr_4$, M_r 367.56, obtained in similar manner to $TiCl_4$ (as yellow octahedral crystals, ϱ 3.25, *mp* 40 °C, *bp* 320 °C), is similar to the chloride in its chemical behavior. It is extremely hygroscopic, soluble in alcohol and moderately soluble in ether.

Titanium tetraiodide, TiI_4, is prepared by double decomposition of $TiCl_4$ with HI. It crystallizes as red-brown octahedra (*mp* 150 °C, *bp* 365 °C).

Titanyl Sulfate. When titanium dioxide is fumed with concentrated sulfuric acid, titanyl sulfate, $TiOSO_4$, is formed, as a white powder, soluble in cold water. It is decomposed by hot water, with the deposition of gelatinous titanium dioxide:

$$TiOSO_4 + H_2O \rightarrow TiO_2 + H_2SO_4$$

In addition to titanyl sulfate, there are other titanium sulfates, both with higher and lower SO_3 contents. The neutral titanium sulfate, $Ti(SO_4)_2$, is not known with certainty in the free state. Double salts derived from it are known, in particular the sulfatotitanates of the type $M^I_2[Ti(SO_4)_2]$. Titanyl sulfate also forms double salts, e.g., $(NH_4)[TiO(SO_4)_2] \cdot H_2O$.

Titanates. Neglecting their water of crystallization, the alkali titanates correspond mostly to the formulas $M^I_2TiO_3$ and $M^I_2Ti_2O_5$. They can be obtained by evaporating solutions of α-titanic acid in concentrated alkali hydroxide solutions (β-titanic acid is insoluble). They can be prepared in the anhydrous state by fusing titanium dioxide with alkali metal carbonates. The titanates of other metals can also be prepared by high-temperature methods. Compounds of the type $M^I_4TiO_4$ (orthotitanates) are

formed in this way by a few metals, as well as those of the type $M^{I}_{4}TiO_3$ (metatitanates). Numerous polytitanates (compounds with more than one TiO_2 per one $M^{I}_{2}O$) have been obtained by fusion methods. Alkaline earth metal titanates are known corresponding to the types $M^{II}TiO_3$, $M^{II}_{2}TiO_4$, and $M^{II}_{2}TiO_5$. Calcium titanate, $CaTiO_3$, occurs in nature as perovskite. The compound Li_2TiO_3 is not a titanate, but is a double oxide with the rock salt structure.

The more commonly occurring mineral ilmenite, iron(II) titanate, $FeTiO_3$, usually contains more iron than is required by this formula. It is isomorphous with hematite, Fe_2O_3; in consequence hematite ores often contain considerable amounts of titanium dioxide. The crystal structure of ilmenite is derived from that of corundum, with which hematite is isomorphous, in such a way that the aluminum atoms are replaced alternately by titanium and iron atoms. The metatitanates of Co^{II}, Ni^{II}, Mn^{II}, Cd, and Mg likewise crystallize with the ilmenite structure, whereas $CaTiO_3$, $SrTiO_3$, and $BaTiO_3$ crystallize in the perovskite structure. The compound Fe_2TiO_5 occurs in nature as pseudobrookite. Titanates of general formula $M^{II}TiO_4$ for which the structures are known (M^{II} = Mg, Zn, Mn, Co), crystallize with the spinel structure.

Titanite or sphene, a monoclinic mineral of the composition $CaO \cdot TiO_2 \cdot SiO_2$, is best represented by the formula $Ca(Ti^{IV}, Fe^{III})(O, OH)[SiO_4]$. X-ray structure analysis has shown that titanite is a silicate.

Peroxytitanic Acid and Peroxytitanates. Titanium salts in neutral or acid solution are colored an intense orange red by hydrogen peroxide. From sufficiently concentrated solutions, peroxytitanic acid, H_4TiO_5, can be precipitated by ammonia as a brownish yellow precipitate. Like titanic acid, peroxytitanic acid forms a gel with a variable water content. The compound contains one peroxy group. It was shown to be a true peroxyacid, and not an H_2O_2 addition product. Peroxytitanyl ions are formed when any dilute, strongly acidified titanium or titanyl salt solutions are treated with hydrogen peroxide. This is used for the analytical detection of titanium.

Titanium Sulfides. Titanium disulfide, TiS_2, is formed by passing a mixture of titanium tetrachloride and hydrogen sulfide through a red-hot porcelain tube:

$$TiCl_4 + 2\,H_2S \rightarrow TiS_2 + 4\,HCl$$

A different reaction takes place between titanium tetrachloride and hydrogen sulfide in the cold, since the latter then acts as a reducing agent, according to:

$$TiCl_4 + H_2S \rightarrow TiCl_2 + 2\,HCl + S$$

Titanium disulfide forms brassy yellow scales with a metallic luster. It is stable in air at normal temperature. When heated it is converted into TiO_2, and on heating in a stream of hydrogen or of nitrogen it yields the lower sulfides Ti_2S_3 and TiS. It is not decomposed by water even on boiling and is also stable towards dilute sulfuric acid, hydrochloric acid, and ammonia. It is decomposed by nitric acid, as well as by hot concentrated sulfuric acid, with the deposition of sulfur. The disulfide is dissolved by boiling caustic potash, with the formation of potassium titanate and potassium sulfide.

14.4. Interstitial Compounds

If titanium compounds are reduced at high temperature in the presence of nitrogen, *titanium nitride* is readily formed in an impure state. Pure titanium nitride TiN is obtainable in the form of a bronze colored powder by strongly heating titanium tetrachloride in a current of ammonia:

$$3\,TiCl_4 + 16\,NH_3 \rightarrow 3\,TiN + {}^{1}/_{2}N_2 + 12\,NH_4Cl$$

It has a structure of the rock salt type.

Titanium can take considerable amounts of boron into solid solution. It also forms two borides TiB (cubic) crystallizes with the zinc blende structure, and TiB_2 (hexagonal) is isotypic with AlB_2, TiB_2 is harder than any other known metallic boride. It is prepared industrially from boron carbide, B_4C, which is manufactured on a considerable scale as an abrasive. This is mixed with B_2O_3, and treated with metallic Ti:

$$7\,Ti + 3\,B_4C + B_2O_3 \rightarrow 7\,TiB_2 + 3\,CO$$

15. Toxicology and Occupational Health

Ti metal generally shows relatively low toxicity as compared to most other heavy metals. The toxicity of its compounds varies to a large extent, being dependent on their physicochemical properties (e.g., water solubility) and chemical reactivity. Ti apparently is not an essential element for humans [161].

The most important compound of Ti industrially is titanium dioxide, TiO_2. $TiCl_4$ is a starting material for other Ti compounds.

In medicine, Ti metal is used in surgical implants and is mostly tolerated by tissue. Other Ti compounds, such as the dioxide, salicylate, and tannate, are widely used in pharmaceutical and cosmetic products. They have been reported to show practically no adverse effects. Ionic cyclopentadienyltitanium(IV) derivatives (titanocene complexes), a new type of antitumor agent, have shown antitumor activities against certain experimental animal tumors [162, 163].

Intoxications with Ti and its compounds are comparatively rare; they occur almost exclusively by occupational exposure.

Biochemistry, toxicokinetics, human and animal toxicity, levels of tolerance, evaluation of health risks to humans, and ecotoxicity of Ti and its compounds have been reviewed [161, 164].

Toxicokinetics. The average intake of TiO_2 in humans is 0.3 – 0.5 mg/d. Approximately the same amount is excreted daily: at least 0.29 mg/d via the feces and 0.01 mg/d via the urine [161, 165, 166]. The rate of gastrointestinal absorption is about 3 % of the ingested material, i.e., most of the ingested Ti is excreted unabsorbed. The mean Ti concentration in blood is 0.07 mg/L [167].

The lungs and the lymph nodes are the predominant target organs for the accumulation of TiO_2 [167, 168]. About 0.4 µg Ti is stored in the lungs per day. In experimental studies in rats, after a single exposure to TiO_2, 40 – 45 % of the deposited material is removed by pulmonary clearance after 25 d. Storage of Ti in the lung is a dose-dependent process [169].

The biological half-life of Ti is between 320 and 640 d. Excretion by other routes than via urine is unknown [161].

Acute Toxicity. Generally TiO_2 behaves as an inert substance which is practically free of acute toxicity. After short-term exposure of guinea pigs to TiO_2 aerosols, an inflammatory response could not be detected; the number of leukocytes and macrophages did not change [170]. In addition, TiO_2 did not lead to the production of hydroxyproline by fibroblasts, which indicates a lack of fibrogenicity [171].

In contrast, $TiCl_4$ represents a considerable toxic hazard to humans [172]. Contact with the skin, the mucous membranes, and the conjunctiva leads to severe irritation and necrosis due to the hydrochloric acid formed by immediate hydrolysis. Contact with the eyes can cause blindness. The vapor leads to lacrymation and to irritation of the lower respiratory tract, with the danger of lung edema. In animal experiments the development of lung edema was more rapid and the death rate after exposure to $TiCl_4$ was higher than with an equivalent concentration of HCl alone, since HCl adsorbed on the particles of hydrolyzed $TiCl_4$ is transported to deeper parts of the respiratory tract which usually are not reached by the highly water soluble HCl [173].

Chronic Toxicity. In a long-term study of up to 390 d with various species, but very low numbers of test animals (two guinea pigs, two rabbits, four cats, one dog), the animals received between 0.6 and 9 g technical grade TiO_2 per day. There were no adverse effects and no histological abnormalities in any of these animals [174]. In a long-term study with mice from weaning to natural death, the addition of 5 mg of titanium potassium oxalate per liter of drinking water led to a decrease of the survival rate after 18 months, which was slight in female mice and more pronounced in male mice [175]. The titanium-fed mice generally showed a higher body weight.

Inhalation of TiO_2 dust by rats for a period of up to 13 months (4 times per day, 5 d per week), followed by a 7-month period of fresh air, did not lead to any pathological effect on the lungs; the weight of the rats remained unchanged [176]. In another inhalation study (10 – 250 mg/m^3 for 6 h/d, 5 d/week) performed for 2 years with rats, exposed groups showed an increase in the incidence of pneumonia, tracheitis, and rhinitis with squamous metaplasia of the anterior nasal cavity.

Rats exposed to 50 and 250 mg/m^3 showed various signs of disturbance of respiratory tract function and dust deposition in the tracheobronchial lymph nodes [177] (see also "Genotoxicity"). These findings were confirmed in detail by an analogous study in which an accumulation of TiO_2 was observed in the upper respiratory tract, lungs, gastrointestinal tract, lymphatic system, liver, and spleen [178].

Similar results concerning effects on respiratory tract function were obtained in a long-term study with rats exposed to $TiCl_4$ (0 – 10 mg/m^3, 6 h/d, 5 d/week) for two years [179].

Ti compounds such as the hydride, carbide, and boride as well as potassium octatitanate and titanium phosphate fibers apparently are fibrinogenic in animal tests [161].

In humans, the most pronounced effects of a chronic occupational TiO_2 exposure appear in the respiratory system [decreased FEV 1 (Forced Expiration Volume during the first second) and FVC (Forced Vital Capacity)], resulting from deposition of TiO_2 in the lung [180].

A case of TiO_2 pneumoconiosis accompanied by lung cancer and slight fibrosis of the interstitium around bronchioles and vessels has been observed with a fifty-three-year-old man who had been employed for about thirteen years with packing of TiO_2 [181].

Genotoxicity. In an inhalation study with TiO_2 dust (10 – 250 mg/m^3 for 6 h/d, 5 d/week) performed for two years with rats, there was an increased incidence of bronchioalveolar adenomas and cystic keratinizing squamous cell carcinomas as well as excessive dust loading in the lung of the 250 mg/m^3 exposure group [177, 178]. These tumors were different from common human lung cancers with respect to tumor type, anatomic location, and tumorigenesis, and were devoid of tumor metastasis [177]. In a ^{32}P postlabeling assay for the detection of DNA adducts, TiO_2 formed one of the two DNA adducts that could be quantitated separately [179]. In contrast, up to 5 % TiO_2 in the diet during 130 weeks did not cause carcinogenic effects in rats [182].

According to the IARC [183], there is inadequate evidence for the carcinogenicity of TiO_2 in humans, and limited evidence for the carcinogenicity of TiO_2 in experimental animals. TiO_2 is not classifiable as to its carcinogenicity to humans (Group 3).

Intratracheal administration of TiO_2 combined with benzo[*a*]pyrene to hamsters causes an increase in the incidence of benign and malignant tumors of the larynx, trachea, and lungs over that in benzo[*a*]pyrene-treated controls.

A two-year inhalation study with rats exposed to $TiCl_4$ (0 – 10 mg/m^3, 6 h/d, 5 d/week) revealed an increased incidence of cystic keratinizing squamous cell carcinomas in high-dose groups [184].

The inhalation of fibrous potassium octatitanate by rats, hamsters, and guinea pigs has basically the same effects on the lungs as asbestos, although the effects are considerably less pronounced [185]. These findings were confirmed by an in-vivo test system [186].

Synthetic titanium phosphate mineral fiber shows a fibrinogenic potential in rats and hamsters after intratracheal injection, similar to that of asbestos; however, in contrast to the results with asbestos, no abnormal increase of the abdominal tumor rate after intraperitoneal injection of 2 – 10 mg/kg was observed [187].

Reproductive Toxicity. In a three-generation study with rats, the addition of titanium potassium oxalate to the drinking water (5 mg/L) led to a marked decrease in the number of animals surviving to the third generation [161]. In pregnant mice, the application of 30 or 60 mg/kg of the antitumor agent titanocene dichloride on day 10 and 12 resulted in the appearance of cleft palate in numerous fetuses and of costal malformations in some fetuses [188].

Immunotoxicity. In a group of dermatitis patients, no evidence of a skin sensitizing potential of TiO_2 could be observed [189].

Toxicological Data. Titanium dioxide, TiO_2: TDLo (rat, intramuscular) 360 mg/kg [190]; titanium tetrachloride, $TiCl_4$: LCLo (mouse, 2 h, inhalation) 10 mg/m^3 [191]; LC_{50} (rat, 4 h, inhalation) 460 mg/m^3 [192]; calcium titanate, $CaTiO_3$: LD_{50} (rat, i. p.) 5300 mg/kg; barium titanate, $BaTiO_3$: LD_{50} (rat, i. p.) 3000 mg/kg; lead titanate, $PbTiO_3$: LD_{50} (rat, i. p.) 2000 mg/kg [161]; tetrabutyl titanate: LD_{50} (mouse, i. v.) 180 mg/kg [193]; dicyclopentadienyltitanium: LDLo (rat, intramuscular) 50 mg/kg, LDLo (hamster, intramuscular) 83 mg/kg [190].

Treatment. After contact of the skin with $TiCl_4$ immediate dry removal (without use of water) is necessary. Thereafter, remaining traces of $TiCl_4$ should be removed with soap and plenty of water. After contact of the eyes with $TiCl_4$, adequate measures are rinsing with plenty of water, use of Isogutt eye drops, and contacting an ophthalmologist as soon as possible. In cases of $TiCl_4$ vapor inhalation, cortisone preparations should be applied immediately (e.g., inhalation of Auxiloson) to avoid the development of lung edema. In severe cases clinical observation for at least 24 h is necessary.

Occupational Health. The MAK value is 6 mg TiO_2/m^3, measured as total fine dust [194]. In the United States the ACGIH set a TLV value of 10 – 20 mg TiO_2/m^3 in 1980 [195].

Lung fibrosis caused by titanium and some of its compounds (carbide, hydride) is included in the list of the German Berufskrankheitenverordnung (Occupational Disease Regulations) [196].

Since TiO_2 tends to accumulate, persons heavily exposed at the workplace should undergo periodical medical examinations, including chest X-rays and ventilatory function tests.

16. References

General References

1. M. Hansen, K. Anderko: *Constitution of Binary Alloys,* McGraw Hill Book Comp., New York 1958.
2. U. Zwicker: *Titan und Titanlegierungen,* Springer-Verlag, Berlin 1974.
3. *The Economics of Titanium.* 3rd. ed., Roskill Information Service Ltd., London 1980.
4. K. Rüdinger: "Titan," in C. Rohrbach (ed.): *Werkstoffe, erforscht, geprüft, verarbeitet,* Technischewissenschaftliche Vorträge auf der Industrieausstellung, Colloquium-Vlg., Berlin 1971, pp. 129 – 172.
5. H. H. Weigand: "Titan und Titanlegierungen," *Stahl Eisen* **80** (1960) 174 – 182, 301 – 309.
6. K. Rüdinger: "Titan und Titanlegierungen – Moderne Werkstoffe – Auswahl – Prüfung – Anwendung – Übersichten über Sondergebiete der Werkstofftechnik für Studium und Praxis," *Z. Werkstofftech.* **9** (1978) 181 – 189, 214 – 218.

Specific References

7. Atomic Energy Commission (ed.): *The Reactor Handbook,* **vol. I,** AECD 3645, 1955, p. 352.
8. A. D. Schelberg, M. B. Sampson, A. C. G. Mitchell, *Phys. Rev.* **74** (1948) 1239.
9. J. S. V. Allen, M. I. Pool, J. D. Kurbatov, L. L. Quill, *Phys. Rev.* **60** (1941) 425 – 429.
10. G. T. Seaborg, I. Pearlman, *Rev. Mod. Phys.* **20** (1948) 585 – 667.
11. L. Seren, H. N. Friedlander, S. H. Turkel, *Phys. Rev.* **72** (1947) 888 – 901.
12. H. Walke, E. J. Williams, G. R. Evans, *Proc. Soc. London Ser. A* **171** (1939) 360 – 382.
13. T. H. Schofield, *Proc. Phys. Soc. London Sect. B* **67** (1954) 845 – 847.
14. K. D. Deardorff, E. T. Hayes, *J. Met.* **8** (1956) 509 –511.
15. A. D. McQuillan, M. K. McQuillan: *Titanium,* Butterworth, London 1956.
16. G. Hägg, *Z. Phys. Chem. Abt. B* **11** (1931) 433 –454.
17. S. Koncz, S. Szanto, H. Waldhauser, *Naturwissenschaften* **42** (1955) 368 – 369.
18. A. D. McQuillan, *Proc. Soc. London Ser. A* **204** (1950) 309 – 323.
19. I. Backhurst, *J. Iron Steel Inst. London* **198** (1968) 124 – 134.
20. B. W. Levinger, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **197** (1953) 195.
21. H. T. Clark, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **185** (1949) 588 – 589.
22. J. M. Blocher, I. E. Campbell, *J. Am. Chem. Soc.* **71** (1949) 4040 – 4042.
23. P. Schüler, *DEW-Tech. Ber.* **7** (1967) 5 – 12.
24. K. Bungardt, K. Rüdinger, *Z. Metallkd.* **52** (1961) 120 – 135.
25. K. Bungardt, K. Rüdinger, *Metall (Berlin)* **14** (1960) 988 – 994.
26. K. Rüdinger: "Titan," in E. Rabald, D. Behrens (eds.): *DECHEMA-Werkstoff-Tabelle, Physikalische Eigenschaften,* Deutsche Gesellschaft für Apparatewesen, Frankfurt/M. 1966.
27. R. G. Netzel, J. R. Dillinger, *Proc. Int. Conf. Low Temp. Phys. 7th* 1961, 389 – 391.
28. W. Spyra, *DEW-Tech. Ber.* **5** (1965) 20 – 24.
29. F. R. Brotzen, E. L. Harmon, Jr., A. R. Troiano, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **203** (1955) 414.
30. R. W. Powell, R. P. Tye, *J. Less Common Met.* **3** (1961) 226 – 233.
31. W. Sprya, *DWE-Tech. Ber.* **1** (1961) 156 – 161.
32. D. L. Burk, I. Estermann, S. A. Friedberg, *Z. Phys. Chem. (Munich)* **16** (1958) 183 – 193.

33. L. G. Carpenter, F. R. Reavell, *Nature (London)* **163** (1949) 527.
34. C. M. Libanati, S. F. Dyment, *Acta Metall.* **11** (1963) 1263 – 1268.
35. R. N. Jeffery, *Urbana Scientific and Techn. Aerospace Report* **9** (1971) no. 6, W 71 – 16 778.
36. Lepkowski, Holladay, TML Report No. 73, Batelle Memorial Inst., Columbus, OH, 1957.
37. B. W. Roberts: "Properties of Selected Superconducting Materials," NBS Techn. Note 983, Suppl., 1978.
38. F. B. Litton, B. W. Gonser, *Met. Prog.* **55** (1949) 346 – 347.
39. J. D. Ramsdell, D. R. Matthews, *Bur. Mines Rep. Invest.* **5701** (1960).
40. W. Knorr, L. Kopp, C. M. v. Meysenbug, K. Rüdinger, *Z. Werkstofftech.* **7** (1976) 437 – 451.
41. D. Schlain, *Bur. Mines Rep. Invest.* **4965** (1953).
42. R. Otsuka, *Sci. Pap. Inst. Phys. Chem. Res. (Jpn.)* **54** (1960) 97 – 123.
43. M. Stern, H. Wissemberg, *J. Electrochem. Soc.* **106** (1959) 755 – 759.
44. M. Stern, C. R. Bishop, *ASM Trans. Q.* **52** (1960) 239 – 256.
45. K. Rüdinger, *Werkst. Korros.* **16** (1965) 109 – 115.
46. M. Stern, H. Wissemberg, *J. Electrochem. Soc.* **106** (1959) 759 – 764.
47. R. Boyer, G. Welsch, E. W. Collings (eds.): "Titanium Alloys," *Materials Properties Handbook,* ASM, Materials Park, Ohio 1994, pp. 175 – 176.
48. K. Rüdinger, *Werkst. Korros.* **13** (1962) 401 – 405.
49. E. A. Gulbransen, K. F. Andrew, *J. Electrochem. Soc.* **96** (1949) 363 – 376.
50. D. N. Williams et al., *WADC Techn. Rep.,* Battelle Memorial Inst., Columbus, OH, 1957, pp. 54 – 661.
51. U. Zwicker, *Z. Metallkd.* **49** (1958) 179 – 184; **50** (1959) 261 – 268.
52. K. Rüdinger. "Sonderwerkstoffe für den Korrosionsschutz, " in H. Gräfen (ed.): *Die Praxis des Korrosionsschutzes,* **vol. 64,** "Kontakt u. Studium Werkstoffe," Expert-Verlag, Grafenau/Württ. 1981, pp. 111 – 144; 345 – 347.
53. K. Rüdinger, H. H. Weigand, *Metall (Berlin)* **27** (1973) 241 – 245.
54. R. N. Lyon: *Liquid Metals Handbook,* **vol. 733 (Rev.**) Atomic Energy Commission, Dept. of the Navy, Washington, D.C., Navexos 1952, pp. 144 – 183.
55. P. Eggert et al., *Metall (Berlin)* **35** (1981) 340 – 342.
56. G. P. Merrill, *Mem. Nat. Acad. Sci. Washington* **14** (1916) 28.
57. Titanium, Mineral Commodity Profiles MCP-18, U.S. Bureau of Mines, Washington D.C. (Aug. 1978).
58. O. Priess, DE 334 249, 1921.
59. D. Y. Toptygin, I. S. Morozov, *Zh. Prikl. Khim. (Leningrad)* **34** (1961) 691.
60. E. C. Perkins, H. Dolezal, D. M. Taylor, R. S. Lang, *Bur. Mines Rep. Invest.* **6317** (1963).
61. British Titanium Prod. Co., EP 745 931, 1952.
62. Illinois Inst. of Techn. Res. Inst., *Chem. Eng. News* **42** (1964) March 23, 47.
63. Pittsburgh Plate Glass Co., US 2 245 358, 1939.
64. C. K. Stoddard, E. Pietz, *Bur. Mines Rep. Invest.* **4153** (1947).
65. E. Junker, *Z. Anorg. Chem.* **228** (1936) 97 – 111.
66. R. Kieffer, F. Lihl, E. Effenberger, *Z. Metallkd.* **60** (1969) 94 – 100.
67. Z. Wyss, *Ann. Chim. (Paris)* **3** (1948) 215 – 219.
68. O. Kubaschewski, W. A. Dench, *Trans. Inst. Min. Metall.* **66** (81956/57) 1 – 6.
69. Südd. Kalkstickstoffwerke, DE 1 111 403, 1958.
70. Bayer, DE 974 210, 1954.
71. L. D. Jaffe, R. K. Pitler, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **188** (1950) 1396.
72. W. J. Kroll, *Trans. Electrochem. Soc.* **78** (1940) 35 – 47.
73. M. A. Hunter, *J. Am. Chem. Soc.* **32** (1910) 330 – 336.
74. H. Freudenberg, DE 658 995, 1936.
75. J. P. Quin, GB 717 930, 1952; GB 720 517, 1954.
76. T. Ikeshima: "Recent Development in Titanium Sponge Production in Titanium Science and Technology," in G. Lütjering, U. Zwicker, W. Bunk (eds.): *Proc. 5th Int. Conf. Titanium,* München 1984, DGM, Oberursel 1985, pp. 3 ff.
77. W. J. Kroll, US 2 205 854, 1940.
78. F. S. Wartman et al., *Bur. Mines Rep. Invest.* **4519** (1949).
79. P. J. Maddex, US 2 556 763, 1948.
80. Titangesellschaft, DE 1 045 720, 1954; DE 1 107 942, 1957.
81. R. Holst, *Angew. Chem.* **68** (1956) 154.
82. Kinzoku Zairyo Kenkysuko, EP 837 905, 1957.

83. Y. Okura: "Titanium Sponge Production Technology", *Proc. 8th World Conf. Titanium,* Birmingham 22 – 26 Oct. 1995, The Institute of Materials, London 1996.
84. Barium Steel Corp., DE 1 103 033, 1957.
85. O. J. C. Runnals, L. M. Pidgeon, *J. Met.* **4** (1952) 843.
86. Fulmer Res. Inst., DE 974 695, 1952.
87. W. E. Reid, J. M. Bisch, A. Brenner, *J. Electrochem. Soc.* **104** (1957) 21 – 29.
88. V. G. Gopenko, I. A. Ivanov, *Russ. Metall. Fuels* 1960, 21 – 23.
89. M. E. Sibert, O. H. McKenna, M. A. Steinberg, E. Wainer, *J. Electrochem. Soc.* **102** (1955) 252 – 262.
90. Horizons Titanium Corp., DE 1 060 605, 1953.
91. W. J. Kroll, *Metall (Berlin)* **9** (1955) 370 – 376.
92. W. R. Opie, O. W. Moles, *Trans. Met. Soc. AIME* **218** (1960) 646 – 649.
93. Titangesellschaft, DE 975 587, 1951; DE 1 045 667, 1955; DE 1 093 562, 1957; DE 1 103 600, 1952.
94. P. Ehrlich, H. Kühn, *Z. Anorg. Allg. Chem.* **298** (1959) 176 – 192.
95. Deutsche Norton-Ges. mbH, DE 1 072 393, 1953.
96. Horizons Titanium Corp., DE 1 063 814, 1953.
97. ICI, DE 1 115 032, 1956.
98. R. I. Jaffee: *Titanium '80, Science and Technology,* The Metallurgical Society of AIME, Warrendale, PA, 1980, pp. 53 – 74.
99. G. Cobel, J. Fisher, L. E. Snyder: *Titanium '80, Science and Technology,* The Metallurgical Society of AIME, Warrendale, PA, 1980, pp. 1969 – 1976.
100. K. Rüdinger, *Metall (Berlin)* **35** (1981) 778 – 779.
101. S. Tamamoto: "Present Features in Production of Titanium Sponge and Ingot," in F. H. Froes, I. L. Caplan: *Titanium '92, Science and Technology,* TMS Publication, Warrendale, PA, 1993, pp. 53 – 63.
102. K. Rüdinger et al.: "Investigation of Heat Exchange in the Electrode Tip of the Non-Consumable Electrode Vacuum Arc Furnace for Melting Titanium," in F. H. Froes, I. L. Caplan: *Titanium '92, Science and Technology,* TMS Publication, Warrendale, PA, 1993, pp. 2355 – 2362.
103. C. H. Entrekin, H. R. Harter: "State of the Art in Electron Beam Melting of Titanium," in F. H. Froes, I. L. Caplan: *Titanium '92, Science and Technology,* TMS Publication, Warrendale, PA, 1993, pp. 2339 –2346.
104. W. R. Chinnis, W. H. Buttrill: "Production Titanium Plasma Cold Hearth Melting," in F. H. Froes, I. L. Caplan: *Titanium '92, Science and Technology,* TMS Publication, Warrendale, PA, 1993, pp. 2363 –2370.
105. A. R. Vaia, R. R. Akers: *Titanium Science and Technology,* Plenum Press, New York 1973, pp. 331 – 341.
106. H. Stephan: *Titanium Science and Technology,* Plenum Press, New York 1973, pp. 343 – 352.
107. J. Greenspan, F. J. Rizzitano, E. Scala: *Titanium Science and Technology,* Plenum Press, New York 1973, pp. 365 – 379.
108. R. F. Geisendörfer, R. J. Saidak: *Titanium Science and Technology,* Plenum Press, New York 1973, pp. 399 – 418.
109. M. Hohmann, S. Jönsson: "New Concepts of Inert Gas Atomization Plants," *Metall Powder Report* (1990) Jan.
110. H. G. Domazer, G. Büttner, H. Eggert, DE 3 017 782 A 1, 1980.
111. K. Rüdinger, *Blech Rohre Profile* **28** (1981) 234 – 237.
112. K. Rüdinger, A. Ismer, *Schweißen Schneiden* **19** (1967) 71 – 73.
113. A. Matting, K. Ulmer, *Metall (Berlin)* **16** (1962) 2 – 6.
114. F. Hesselt, *DGLR-Jahrestagung,* Report no. 56, Bremen 1969.
115. K. Rüdinger, *Z. Werkstofftech.* **2** (1971) 169 – 174.
116. P.-J. Winkler: "Recent Advances in Superplasticity and Superplastic Forming of Titanium Alloys," in P. Lacombe, R. Tricot, G. Béranger (eds.): *Proc. 6th World Conf. Titanium,* Société Française de Métallurgie, Cannes, 6 – 9 June 1988.
117. F. Preißer, P. Minarski, F. Hoffmann: "Hochdrucknitrieren von Titanwerkstoffen," *HTM* **46** (1991) 361 – 366.
118. D. Muster (ed.): "Comparison of Surface Modifications," *Brite-Euram Final Report,* Contr. BREU-0477, CHRU BP 426, Strasbourg 1995.
119. A. Weisheit: *Lasergaslegierung von Titanwerkstoffen,* Ph. D. Thesis TU Clausthal, 1993.
120. A. Lang, H. Waldmann, H. W. Bergmann: "Cladding of Metallic Substrates with Diamonds and Cubic Boron Nitride," *ECLAT '94,* Bremen, DVS-Verlag, Düsseldorf 1994, pp. 456 – 461.
121. B. L. Mordike, R. Haude: "Titanwerkstoffe in Präzisionsbearbeitung mit Festkörperlasern," *Laser in der Materialbearbeitung,* vol. 4, VDI-Verlag, Düsseldorf 1995, pp. 173 – 184.

122. J. L. Murray: *Phase Diagrams of Binary Titanium Alloys,* ASM International, Metals Park, Ohio 1990.
123. R. I. Jaffee, *J. Met.* **7** (1955) 247 – 252.
124. G. Terlinde, G. Fischer: "Beta Titanium Alloys," *Proc. 8th World Conf. Titanium, 22 – 26 Oct. 1995,* The Institute of Materials, London 1996.
125. K. Rüdinger, *Z. Werkstofftech.* **13** (1982) 229 – 236.
126. W. H. Graft, D. W. Levinson, W. Rostoker, *Trans. Am. Soc. Met.* **49** (1957) 263 – 279.
127. S. Yoshida, S. Okamoto, T. Araki, *J. Mechan. Lab. Jpn.* **10** (1956) 67 – 73.
128. W. K. Boyd: *DMIC Memorandum,* **vol. 234,** Battelle Memorial Inst., Columbus, Ohio, 1968, pp. 35 – 44.
129. U. R. Kattner et al., *Metall. Trans. A* **23** (1992) no. 8, 2081.
130. T. Khan, S. Naka, P. Veyssiére, P. Costa: "Intermetallics for Structural Applications," *Proc. High Temp. Materials for Power Engineering 1990,* Liège 1990.
131. G. Sauthoff: "Intermetallic Phases – Materials Developments and Prospects," *Z. f. Metallkunde* **80** (1989) 337 – 344.
132. G. Sauthoff: "Internationaler Stand der Werkstoffentwicklungen auf der Basis intermetallischer Phasen," in F. J. Bremer (ed.): *Intermetallische Phasen als Strukturwerkstoffe für hohe Temperaturen,* Hagen 1990.
133. Y.-W. Kim, F. H. Fores: "Physical Metallurgy of Titanium Aluminides," in S. H. Whang, C. T. Lin, D. P. Pope, J. D. Stiegler (eds.): *High-Temperature Aluminides and Intermetallics,* TMS, Warrendale, PA, 1990, pp. 465 – 492.
134. W. Smarsly: "TiAl-Strukturwerkstoffe als Alternative für Ni-Superlegierungen im Flugtriebwerksbau," in F. J. Bremer (ed.): *Intermetallische Phasen als Strukturwerkstoffe für hohe Temperaturen,* Hagen 1990.
135. W. Smarsly, L. Singheiser: "Potential of Intermetallics to replace Superalloys for Advanced Operation Conditions in Gas Turbines," in D. Coutsouradis et al. (eds.): *Materials for Advanced Power Engineering,* part II, Kluwer Academic Publishers, 1994, pp. 1731 – 1756.
136. S. Hurta, H. Clemens, G. Frommeyer, H.-P. Nicolai, H. Sibum: "Valves of Intermetallic γ-TiAl-Based Alloys: Processing and Properties," *Proc. 8th World Conf. Titanium,* Birmingham, Oct. 1995, The Institute of Materials, London 1996.
137. H. Clemens, W. Glatz: "Mikrostruktur und Eigenschaften von γ-TiAl-Basislegierungen," *Prakt. Metallographie* **26** (1995) 257 – 269.
138. R. Wagner et al.: "Pulvermetallurgie von Titan-Aluminiden und -Silziden," in F. J. Bremer (ed.): *Intermetallische Phasen als Strukturwerkstoffe für hohe Temperaturen,* Hagen 1990.
139. *ASM Handbook,* 10th ed., **vol. 2,** ASM, Metals Park, Ohio 1990, p. 926.
140. T. W. Duerig, K. N. Melton, D. Stöckel, C. M. Wayman (eds.): *Engineering Aspects of Shape Memory Alloys,* Butterworths-Heinemann, London 1990.
141. H. Hillmann: "Entwicklung harter Supraleiter," *Metall* **27** (1973) 797 – 808.
142. R. Leucht, H. J. Dudek: "Properties of SiC – Fibre Reinforced Titanium Alloys Processed by Fibre Coating and Hot Isostatic Pressing," *Mater. Sci. Eng. A* **188** (1994) 201 – 210.
143. O. Bernauer: "Development of Hydrogen – Hydride Technology in the F.R.G.," *Int. J. Hydrogen Energy* **14** (1989) no. 10, 727 – 735.
144. K. Rüdinger, *Metall (Berlin)* **24** (1970) 1303 –1314.
145. L. P. Jahnke: *The Science, Technology and Application of Titanium,* Pergamon Press, Oxford 1970, pp. 1099 – 1115.
146. H. Sibum, D. Fisher: "Titanium-Base-Alloys, The Advanced Material in Automotive Design and Construction to Improve Environmental Performance," *Proc. 25th ISATA,* Florence, 1 – 5 June, 1992.
147. J. L. Murray, *Bull. Alloy Phase Diagrams* **2** (1981) no. 3, 393.
148. London & Scandinavian Metallurgical Co.: *Legierungselemente,* Ges. f. Elektrometallurgie, Düsseldorf 1962. Leitner-Plöckinger in E. Plöckinger, H. Straube (eds.): *Die Edelstahlerzeugung,* 2nd ed., Springer-Verlag, Berlin 1965.
149. T. Sims, W. C. Hagel: *The Superalloys,* Wiley, New York 1972.
150. "Micro Alloying" *75th Proceedings of an International Symposium on HSLA-Steels 1975,* Union Carbide Corp., New York 1977.
151. W. Dautzenberg, G. Volkert: "Ferrotitan," in G. Volkert, K. D. Frank (eds.): *Metallurgie der Ferrolegierungen,* 2nd ed., Springer-Verlag, Berlin 1972.
152. H.-G. Brandstelter, Ph. D. Thesis RWTH Aachen 1956.
153. Pêchiney Compagnie de Produits Chimiques et Électrométallurgiques, FR 1 321 503, 1962; DE-AS 1 245 134, 1972.

154. Gesellschaft für Elektrometallurgie, DE 2 242 352, 1972.
155. G. Duderstadt, Ph. D. Thesis RWTH Aachen 1960.
156. *Erzmetall* **32** (1979) 107, 251, 354, 461; **33** (1980) 63, 241, 421, 521; **34** (1981) 50, 224, 439, 566; **35** (1982) 59, 212, 412, 549.
157. G. P. Jones, J. Pearson, *Metall. Trans.* **B 7** (1976) 223 – 234.
158. E. Pelzel, *Metall (Berlin)* **23** (1969) 782 – 785.
159. *Gmelin,* System No. 41: "Titanium", Springer Verlag, Berlin 1951.
160. H. Remy: *Lehrbuch der anorgischen Chemie,* **vol. 2,** Akademische Verlagsgesellschaft, Leipzig, 1961, pp. 72 – 84.
161. World Health Organization: *Titanium, Environmental Health Criteria* **24,** Geneva 1982, pp. 1 – 68.
162. P. Koepf-Maier, T. Klapotke, *Arzneim. Forsch.* **39** (1989) no. 4, 488 – 490.
163. P. Koepf-Maier et al., *Cancer Chemother. Pharmacol.* **24** (1989) no. 1, 23 – 27.
164. R. Wennig, N. Kirsch in H. G. Seiler, H. Sigel, A. Sigel (eds.): *Handbook on Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988, pp. 705 – 714.
165. H. Valentin, K. H. Schaller, Titanium, in L. Alessio et al. (eds.): *Human Biological Monitoring of Industrial Chemical Series,* Commission of the EC, Publication EUR 8476 EN, Luxembourg 1983, pp. 15 –157.
166. H. A. Schroeder, J. J. Balassa, I. H. Tripton, *J. Chron. Dis.* **16** (1963) 55 – 69.
167. E. I. Hamilton, M. J. Minski, J. J. Cleary, *Sci. Total Environ.* **1** (1972) 341 – 347.
168. W. Röthig, J. H. Wehran, *Dtsch. Gesundheitswes.* **27** (1972) 1555 – 1559.
169. J. Ferin, M. L. Feldstein, *Environ. Res.* **16** (1978) 342 – 352.
170. R. Rylander, M. Sjöstrand, R. Bergström, *Toxicology* **12** (1979) 211 – 220.
171. A. G. Heppleston in W. H. Walton (ed.): *Inhaled Particles and Vapors,* **vol. III,** Unwin Brothers Ltd., pp. 357 – 369.
172. R. Kühn, K. Birett: *Merkblätter Gefährliche Arbeitsstoffe,* 4th ed., ecomed verlag, Landsberg/Lech 1979, Datenblatt T 050.
173. E. A. Mel'nikova, *Gig. Sanit.* **5** (1958) 27 – 31.
174. K. B. Lehmann, L. Herget, *Chem. Ztg.* **51** (1927) 793 – 794.
175. H. A. Schroeder, J. J. Balassa, W. H. Vinton, Jr., *J. Nutr.* **83** (1964) 239 – 250.
176. H. Christie, R. J. MacKay, A. M. Fisher, *Am. Ind. Hyg. Assoc. J.* **24** (1963) 42 – 46.
177. K. P. Lee, H. J. Trochimowicz, C. F. Reinhardt, *Toxicol. Appl. Pharmacol.* **79** (1985) 179 – 192.
178. DuPont Corp., EPA/OTS 88-85 008 005, 1986.
179. J. Gallagher et al., *Carcinogenesis* **15** (1994) 1291 –1299.
180. *Arbete och Halsa* **32** (1989) 44 – 54.
181. I. Yamadori, S. Ohsumi, K. Taguchi, *Acta Pathol. Jpn.* **36** (1986) 783 – 790.
182. B. K. Bernard, M. R. Osherhoff, A. Hofmann, J. H. Mennear, *J. Toxicol. Environ. Health* **29** (1990) 417 –429.
183. IARC International Agency on Cancer Research, *Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans* **47** (1989) 307 – 326.
184. Haskell Labs., EPA/OTS 88-9 680 208, 1986.
185. K. P. Lee et al., *Environ. Res.* **24** (1981) 176 – 191.
186. A. Poole, R. C. Brown, A. P. Rood, *Br. J. Exp. Pathol.* **67** (1986) 289 – 296.
187. P. Gross, R. J. Kociba, G. L. Sparschu, J. M. Norris, *Arch. Pathol. Lab. Med.* **101** (1977) 550.
188. P. Koepf-Maier, P. Erkenswick, *Toxicology* **33** (1984) 171 – 181.
189. BIBRA Working group, The British Industrial Biological Research Association: *Titanium Dioxide, Toxicity Profile,* Carshalton/London, 1990, p. 7.
190. Natl. Cancer Inst. (US), Progress Report for Contract No. PH-43-64-886, July 1968.
191. National Technical Information Service, AEC-TR-6710.
192. D. P. Kelly, K. P. Lee, B. A. Burgess, *The Toxicologist* **1** (1981) 76 – 77.
193. CSLNX, U.S. Army Armament Research and Development Command, Chemical Systems Laboratory, NX 01650.
194. Deutsche Forschungsgemeinschaft (ed.): *MAK- und BAT-Werte-Liste 1995,* Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte, Mitteilungen der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 31, VCH Verlagsgesellschaft, Weinheim 1995.
195. J. Whitehead, "Titan," in E. Merian (ed.): *Metalle in der Umwelt,* Verlag Chemie, Weinheim 1984, pp. 585 – 588.
196. Berufskrankheitenverordnung, BK Nr. 41 07: Erkrankungen durch Lungenfibrose durch Metallstäube bei der Herstellung oder Verarbeitung von Hartmetallen.

Tobacco

T. C. Tso, Institute of International Development and Education in Agriculture and Life Science, Beltsville, Maryland 20705, United States

1. History 2
2. Botany 3
2.1. Taxonomy 3
2.2. Nicotiana Species 3
2.3. Tobacco Plant Characteristics 3
3. Genetics and Propagation 3
3.1. Origin of Tobacco 3
3.2. Monosomics and Their Characteristics 5
3.3. Intraspecific and Interspecific Hybridization 5
3.4. Parasexual Hybridization 6
3.5. Techniques for Plant Propagation . 6
4. Production 6
4.1. General Classification and Environmental Requirements 7
4.2. Seed and Seedlings 7
4.3. Soil and Fertilizer 7
4.4. Pests and Pest Control 8
4.4.1. Diseases 8
4.4.2. Insects 8
4.5. Harvesting 8
4.6. Curing 9
4.6.1. Flue Curing 9
4.6.2. Air Curing 9
4.7. Fermentation and Aging 10
5. Tobacco Products 10
5.1. Cigarettes 10
5.2. Cigars 11
5.3. Pipe Tobacco 11
5.4. Snuff 11
5.5. Chewing Tobacco 11
6. Plant Physiology 12
6.1. Growth and Development 12
6.2. Primary Nutrition 12
6.3. Trace Elements and Heavy Metals . 12
6.4. Nutritional Disorders 13
6.5. Physiological Disorder 13
7. Biochemistry 13
7.1. Alkaloids 13
7.2. Proteins 15
7.3. Carbohydrates and Organic Acids . 16
7.4. Pigments and Polyphenols 17
7.5. Other Major Compounds 18
8. Chemical Characteristics 19
9. Toxicity 21
10. Economics 21
11. Utilization of Tobacco Other than Smoking 24
11.1. In Basic Research 24
11.2. Use of Nicotine 24
11.3. As a Source of Chemicals 24
11.4. Food and Feed 25
11.5. Dust, Midribs, Stems, and Stalks . . 25
11.6. Biomass 25
12. Outlook 25
12.1. Tobacco in Plant Science 25
12.2. Tobacco in Biotechnology 25
12.3. Stress Research 26
13. References 26

1. History

Tobacco is the most important economic crop that grew up with colonial America. It also helped America grow! On 11 October 1492, when Columbus landed in the New World, he was offered dried tobacco leaves at the House of the Arawaks. The chronology of early tobacco distribution can be summarized as follows. (see below)

Various names were given to the plant, including *Nicotiana*, after Jean Nicot, French Ambassador at Lisbon. Nicot was one of the first to grow tobacco in Portugal, and was largely responsible for making the plant known at the royal court in Paris; the Duc de Guise proposed the name *Nicotiana* in 1585. However, the name tobacco had become established in North America and became the common name. Tobacco grown in France and Spain was *Nicotiana tabacum*, from seed originating in Brazil and Mexico. The species first grown in Portugal and England was *N. rustica* – the seed grown in Portugal came from Florida, while that grown in England came from Virginia.

Tobacco played an important role in farm economics and international trade. In the early 1600s when tobacco was commercially culti-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_123

vated in Virginia, shipping of tobacco leaf helped the development of England's merchant marine, and opened up a valuable export. The Anglo-American tobacco colonies continued to supply the expanding tobacco market throughout the colonial period. In America, from the small beginning at Jamestown, tobacco production in Virginia and Maryland increased rapidly, for it was about the only commodity the colonists could exchange for manufactured products from Europe. Tobacco served to underwrite credit during the Revolutionary War. Independence saw the extension of tobacco growing into North Carolina, Kentucky, Tennessee, Ohio, Missouri, and later into several other States.

Date	Event
1492	(11 October) Columbus offered "dried tobacco leaves" by the Arawaks
1499	Amerigo Vespucci reported the use of chewing tobacco in an island off Venezuela
1545	Iroquois Indians near Montreal, Canada, found to be smoking tobacco
1550	Tobacco pipes later unearthed in China identified as made at this time
1556	Tobacco first grown (or became known) in France
1558	Tobacco in Brazil, Portugal
1559	Tobacco in Spain
1560	*N. rustica* in central Africa
1565	Tobacco in England
1600	Tobacco introduced to Italy, Germany, Norway, Sweden, Russia, Persia, India, Indochina, Japan, China, and west coast of Africa
1612	John Rolfe known to be the first man to grow tobacco for export, at Jamestown, Virginia
1631	Tobacco production extended to Maryland; gradual spread to other areas
1650s	Portuguese took tobacco to South Africa and other countries
	Spanish distributed tobacco to Philippines, Guatemala, and other parts of Central and South America and the West Indies
	Indonesia began tobacco cultivation
	Other European countries began tobacco cultivation
1700s	Tobacco produced in New Zealand
	N. tabacum introduced to Australia (*N. suaveolens* grown prior to this)

2. Botany

2.1. Taxonomy

Commercial tobacco *Nicotiana tabacum* is one of more than 64 known species in the genus. Classification is as follows:

Kingdom:	Plant
Subkingdom:	Embryophyta
Division:	Tracheophyta
Subdivision:	Pteropsida
Class:	Angiospermae
Subclass:	Dicotyledonae
Order:	Solanales
Family:	Solanaceae
Genus:	*Nicotiana*
Species:	*tabacum*

2.2. Nicotiana Species

Nicotiana is classified into subgenera, sections, and species (Table 1). Among those species, 45 are indigenous to North or South America, and 15 to Australia; 5 more species were subsequently identified with some revisions and further changes.

2.3. Tobacco Plant Characteristics

Tobacco is an annual plant, although it is potentially perennial in habit. Among different species, there are variations in leaf shape, size, form of leaf tip, petiolate or sessile, leaf angle to stem, and leaf asymmetry. Leaves of *N. tabacum* are generally ovate or oblong-lancolate and borne directly (sessile) on the main stem; those of *N. rustica* are usually petiolate and of regular ovate or cordate shape.

One of the outstanding characteristics of tobacco is the enormous leaf area it produces. Most varieties have a relatively large leaf and, under favorable conditions the fully mature leaf averages $0.1-0.15\,m^2$ in area. In cigar tobacco, a single plant with 18 leaves may produce $>2\,m^2$ of leaf. In American varieties the leaf does not have a bare petiole. These varieties produce a large, thick stem, the lower portion of which is woody. The flowers are about 1 cm long, and most domestic varieties are pale pink, although some varieties produce white or carmine blossoms. The flower is borne on a short stem and has a five-cleft calyx, from which emerges the slender corolla tube. The latter is much longer than the calyx, and usually expands at the top into a five-lobed stigma. The five stamens are attached to the corolla tube, and each consists of a long, slender filament and an oval anther that splits lengthwise in discharging its pollen. The

Table 1. Classification of the genus Nicotiana

Subgenus	Section	Species	Authority	Somatic chromosome number
Rustica	Paniculatae	*glauca*	Graham	24
		paniculata	Linnaeus	24
		knightiana	Goodspeed	24
		solanifolia	Walpers	24
		benavidesii	Goodspeed	24
		cordifolia	Philippe	24
		raimondii	Macbride	24
	Thyrsiflorae	*thyrsiflora*	Bitter ex Goodspeed	24
	Rusticae	*rustica*	Linneaus	48
Tabacum	Tomentosae	*tomentosa*	Ruix and Pavon	24
		tomentosiformis	Goodspeed	24
		otophora	Grisebach	24
		setchellii	Goodspeed	24
		glutinosa	Linnaeus	24
	Genulinae	*tabacum*	Linnaeus	48
Petunioides	Undulatae	*undulata*	Ruiz and Pavon	24
		arentsii	Goodspeed	48
		wigandioides	Koch and Fintelman	24
	Trigonophyllae	*trigonophylla*	Donal	24
	Alatae	*sylvestris*	Spegazzini and Comes	24
		langsdorffii	Weinmann	18
		alata	Link and Otto	18
		forgetiana	Hort ex Hemsley	18
		bonariensis	Lehmann	18
		longiflora	Cavanilles	20
		plumbaginifolia	Viviani	20
	Repandae	*repanda*	Willdenow ex Lehmann	48
		stocktonii	Brandegee	48
		nesophila	Johnston	48
	Nictiflorae	*noctiflora*	Hooker	24
		petunioides	(Griseback) Millan	24
		acaulis	Spegazzini	24
		ameghinoi	Spegazzini	?
	Acuminatae	*acuminata*	(Graham) Hooker	24
		pauciflora	Remy	24
		attenuata	Torrey ex Watson	24
		longibracteata	Philippi	?
		miersii	Remy	24
		corymbosa	Remy	24
		linearis	Philippi	24
		spegazzinii	Millan	24
	Bigelovianae	*bigelovii*	(Torrey) Watson	48
		clevelandii	Gray	48
	Nudicaules	*nudicaulis*	Watson	48
	Suaveolentes	*benthamiana*	Domin	38
		umbratica	Burbidge	46
		cavicola	Burbidge	40
		debneyi	Domin	48
		gossei	Domin	36
		amplexicaulis	Burbidge	36
		maritima	Wheeler	32
		velutina	Wheeler	32
		hesperis	Burbidge	42
		occidentalis	Wheeler	42

Table 1. (Continued)

Subgenus	Section	Species	Authority	Somatic chromosome number
		simulans	Burbidge	40
		megalosiphon	Heurck and Mueller	40
		rotundifolia	Lindley	44
		excelsior	J. M. Black	38
		suaveolens	Lehmann	32
		ingulba	J. M. Black	40
		exigua	Wheeler	32
		goodspeedii	Wheeler	40
		rosulata	(S. Moore) Domin	40
		fragans	Hooker	48
		africana	Merx and Butler	24

pistil is composed of a swollen basal element, the ovary, containing the ovules arranged on a fleshy placenta, a long slender style arising above the ovary, and a more or less two-lobed stigma. The structure of the flower favors self-fertilization.

In the leaf axils there are buds that ordinarily remain dormant during growth, but if the terminal bud is removed these axillary buds readily develop into large branches, "suckers". The plant possesses an extensive, but comparatively shallow system of fibrous roots that usually afford adequate support, though the plant may be toppled by strong winds, especially if it bears a full-size flower head. The growing leaf has a vivid green to blue-green color that varies considerably with the variety and state of nutrition. The color tends to change to yellow-green as the leaf approaches maturity.

3. Genetics and Propagation

3.1. Origin of Tobacco

The origin of *N. tabacum* has been a challenging problem for plant scientists. One hypothesis is that *N. tabacum* comes from an interspecific hybrid between *N. sylvestris* and *N. tomentosa*; another hypothesis is between *N. sylvestris* and *N. tomentosiformis*; a third between *N. sylvestris* and *N. otophora*. The basic concept is that *N. tabacum* (N = 24) is a result of hybridization between two species (N = 12) after chromosome doubling. Many experimental crosses have been conducted to examine such hypotheses on taxonomic and morphological grounds.

Tobacco fraction 1 protein is a unique genetic marker. Synthesis of this photosynthetic enzyme is regulated by both the nuclear and chloroplast genomes. Fraction 1 protein is identical to ribulose-1,5-diphosphate (RuDP) carboxylase – oxygenase, and is constructed from eight large and eight small subunits (L_8S_8). By electrofocusing fraction 1 protein isolated from *N. tabacum* and the putative progenitor species, it has been confirmed that *N. tabacum* arose from the hybridization of *N. sylvestris* females with *N. tomentosiformis* males [1, pp. 195 – 196]. Comparison with the polypeptide compositions of fraction 1 proteins from the three species indicates that *N. sylvestris* contributed the large subunit polypeptides, and therefore was the maternal parent in the original hybridization. *Nicotiana sylvestris* also contributed one of the two small subunit polypeptides of *N. tabacum*. The other polypeptide was contributed by *N. tomentosiformis*. Such evidence proved that the original hybridization was *N. sylvestris* females with *N. tomentosiformis* males, and that *N. otophora* was not involved in the origin of *N. tabacum*.

3.2. Monosomics and Their Characteristics

A plant with variations in its number of whole chromosomes is known as aneuploid. For example, $2n - 1$ is a monosomic and $2n + 1$ is a trisomic. *N. tabacum* has available a complete set of monosomic types and each type has been identified for its distinguishing characteristics.

The monosomics are therefore of special importance because they are the optimum material for locating genes on a specific chromosome. In addition, it is of great value to examine the transfer of single chromosomes between genus, species, or varieties.

3.3. Intraspecific and Interspecific Hybridization

Intraspecific hybridization is a cross between plants within the same species regardless of genetic differences among the parents or the resulting level of heterozygosity of the hybrid.

Through hybridization and other genetic manipulations, desirable characteristics can be incorporated into tobacco plants for cultivation. Intraspecific hybrids generally occur in Mendelian behavior, and interspecific hybrids essentially occur in a non-Mendelian manner. This is because hybrids between plants of the same species have complete chromosome homology, and hybrids between plants of different species have only partial or nonchromosome homology. There are three conventional approaches to interspecific hybridization for the purpose of alien transfer and incorporation of germplasm into stable diploid lines.

1) Diploid × diploid, then doubling to produce the allopolyploid and backcrossing to produce the sesquidiploid
2) Autotetraploid × autotetraploid to produce the allopolyploid directly
3) Autotetraploid × diploid to produce the sesquidiploid directly

Incorporation of alien chromosomes or segments is generally known as introgression. Interspecific hybridization and introgression in natural populations is a known source of genetic variation and adaptation. The goal of experimental interspecific hybridization is usually the introgressive transfer of a specific trait from one species into the genome of another. These techniques are widely used in tobacco genetics, both pure and applied.

3.4. Parasexual Hybridization

Parasexual hybridization was first conducted with tobacco. It is a means of developing hybrid plants by fusion of protoplasts from different species or related plants differing in their cytoplasms. It may also include inserting organelles such as nuclei, plastids, mitochondria, and chromosomes into protoplasts, or inserting microorganisms, including viruses into protoplasts, or transferring nucleic acids or genes into plant cells. Some experimental fusion has also been reported between plants and animals.

Cellular compatibility or incompatibility between species is usually a main factor in deciding the success or failure of the formation of sexual hybrids. Protoplast fusion, however, involves many other conditions; e.g., protoplast fusion and organelle uptake requires the aggregation or agglutination of cells or organelles with cells. Environmental conditions, media formulation, identification and recovery of fused cells, and availability of biochemical or physiological markers are among the many factors requiring more study.

3.5. Techniques for Plant Propagation

This discussion centers on unconventional techniques, which are generally used not for commercial tobacco production, but for plant improvement through novel approaches in varietal development, academic studies in plant science, and development of new technology, using *Nicotiana* plants as experimental material.

Plant Part Culture. Plant part culture is generally known as organ culture. Culture from vegetative organs includes root, shoot apex, stem, and leaf. Reproductive organs include embryo, ovary, and flower bud. In each system, there are requirements for clean or sterilized material, suitable media, adequate nutrient supply, and environmental conditions.

Tissue Culture. *Nicotiana* plant materials are used widely as models for tissue culture. Early tissue culture work was conducted with interspecific *Nicotiana* hybrids, *N. glauca* × *N. langsdorffii*. Tissue or cell cultures have been valuable tools for evaluating growth regulators, organic metabolism, environmental pollutants, disease resistance, chemical

constituents, genetics, and breeding for crop improvement. Different media have been developed for specific tissue. Pith tissues or young stems are commonly used to initiate tissue growth. Subcultures can be generated from the initial callus. There is wide application of tissue culture for disease resistance.

A fundamentally new possibility has been demonstrated for the polyploidization of tobacco diploids and the diploidization of haploids, with the help of tissue culture.

Cell Culture. Cell culture is a phase of tissue culture, but is discussed separately so that more attention can be devoted to individual cells.

Individual tobacco cells have been used for selection of NaCl-tolerant lines. Cell culture may provide a rapid method for producing salt-tolerant crop plants. Both mutagen-treated and untreated cultures develop salt tolerance. The occurrence of spontaneous or induced variants in somatic cell cultures could provide a source of many agriculturally desirable phenotypes.

Anther Culture. Anther-derived haploids are potentially useful in plant physiology and breeding programs. Haploid plants of *N. tabacum* have been obtained by culturing anthers at the microspore stage. Explants of pith tissue from haploid as well as diploid plants have been grown in vitro, and changes in cell number, cell size, and fresh weight determined. Anther culture is also used for research on sterile hybrids and mutants.

4. Production

Tobacco is consumed throughout the world in various forms: smoking, chewing, and snuff. Production occurs in all locations where the environment is suitable.

4.1. General Classification and Environmental Requirements

Classification of leaf tobacco is primarily based on curing methods and tobacco types. The United States system uses tobacco types as broad units, which are broken down into classes. These are further divided into grades determined by quality, size, condition, color, etc. There are many other types of tobacco which differ from the United States classification. Turkish or Oriental tobaccos are mainly "sun" cured. The Maryland tobacco produced in Switzerland is a combination of air and flue curing.

Leaf property, expressed in chemical and physical properties, is very sensitive to environmental factors. Climate and weather have significant effects on leaf type, grade, and quality. Tobacco is of tropical origin and grows most rapidly in a warm climate. However, it is grown as far north as central Sweden, and as far south as southern Australia. The most important temperature factor in determining latitudinal distribution of tobacco production is duration of the frost-free period. In cool climates, ordinary tobacco requires ca. 100 – 120 d from date of transplanting to attain full maturity, at a mean temperature of 27 °C (80°F). In the United States, the length of the growing season, as measured by average dates of the latest killing frost in Spring and the earliest killing frost in the Fall, is >160 d.

Tobacco needs a liberal, well-distributed rainfall to maintain adequate water supply for rapid growth and large size. Prolonged severe drought results in pronounced stunting of the plants and "firing" of the lower leaves. Moderate drought may not materially reduce yield. Relatively wet seasons tend to reduce the yield by producing thin, light-weight leaf. Excessive rainfall may cause injury to the crop through loss of plant nutrients in light soils and drowning of the plants in heavy soils. Relative humidity has an important effect on loss of moisture by the soil and plant. Strong winds may damage the leaf by bruising or physical destruction.

4.2. Seed and Seedlings

Tobacco seeds are extremely small – ca. 300 000 seeds per ounce (10 000 – 11 000 per gram). Each tobacco plant may produce half an ounce (15 g), ca. 150 000 seeds, sufficient for 100 ft^2 (10 m^2) of seedbed area. Under regular conditions, seeds obtained from one plant may provide enough seedlings for 2 – 5 acres (0.15 –2 ha) of field tobacco. Tobacco plants are grown in a seedbed and later transplanted to the field. The seedbed can be any size or shape convenient for

weeding, applying pesticides, pulling plants, etc. Depending on local conditions and disease problems, seedbeds may be permanent or shifted to a new location each year.

Table 2 shows the approximate seeding rate and area that can be planted from an ounce (30 g) of seed.

Table 2. Seeding rate and field acreage from 1 oz (30 g) of seed

Type	Seed bed area		Field transplants	
	square yards	m^2	acres	ha
Flue-cured	600	500	4 – 5	1.5 – 2
Fire-cured	400	350	4 – 6	1.5 – 2.5
Burley	400	350	3 – 4	1.2 – 1.5
Maryland	300	250	3 – 4	1.2 – 1.5
Dark air-cured	400	350	4 – 6	1.5 – 2.5
Cigar filler	100	85	3 – 4	1.2 – 1.5
Cigar binder	100	85	6 – 8	2.5 – 3
Cigar wrapper	100	85	3 – 4	1.2 – 1.5

4.3. Soil and Fertilizer

Tobacco can be grown in sandy soils to heavy clays, but different types of tobacco require special soil types for optimum leaf quality. Tobacco has an active root system which contributes to the rapid development of an enormous leaf area; adequate soil aeration, water, and nutrient supply are three major requirements.

Table 3 shows several well-known tobacco types and their respective soil types.

Table 3. Major tobacco types and their respective soil types

Tobacco	Soil
Flue-cured	sandy loam to fine sandy loam
Burley	silt loam
Maryland	fine sandy loam
Cigar	
Wrapper	sandy loam
Filler	silt loam
Binder	sandy loam
Dark fire-cured	clay loam, silt loam
Oriental	clay sandy topsoil, and chalky subsoil

The ideal fertilization situation is a fully adequate continuous supply of all essential elements through the growing period, with a rapid decline as the crop approaches maturity. The plants need to maintain an uninterrupted, rapid growth from the time of transplanting to the approach of maturity. Deficiency at any stage may delay plant growth and development, and oversupply of food toward the end of the growing season may prevent normal ripening.

4.4. Pests and Pest Control

Tobacco is subject to many pests. Pests may not only induce direct physical damage, but cause related chemical change. Serious attacks by disease or insects early in plant growth may destroy the whole crop. Efforts to breed disease-resistant plants have been very successful, but tobacco plants resistant to insects are still in an early stage of investigation.

4.4.1. Diseases

The causal organisms of the most common tobacco diseases are described in [1, pp. 136]. Many diseases occur in seedbeds; others occur primarily in the field. In addition to causal organisms and plant varieties, environmental conditions contribute to a large proportion of the occurrence as well as the degree of infection.

4.4.2. Insects

Insect damage similar to that of disease, occurs at various stages of tobacco growth. Among the insects which cause serious damage are budworms, hornworms, flea beetles, aphids, and wireworms. More than 100 species feed on tobacco; 35 of these cause enough damage to require treatment.

The important insects and related pests on tobacco, together with description and type of damage, and chemicals for insect control as described in detail in [1]. The level of pesticide residues on tobacco is always a concern for consumers and the tobacco industry. Regulatory actions have resulted in drastic reduction, if not complete elimination, of certain undesirable materials.

4.5. Harvesting

The metabolic activities in tobacco leaves during growth, maturation and senescense are continuous processes and shift gradually from one stage

to the next. Stage of harvesting and method of harvesting are the critical phases in the production of good quality tobacco. They differ according to tobacco type.

Flue-cured tobacco is ripe when the green leaves begin to fade and become a dull, yellow-green with a cream tip. The ripened leaves are primed in lots of three to five at each priming, beginning at the bottom. Normally, primings are made every 4 – 5 d or at weekly intervals, requiring 4 – 6 weeks to harvest an entire crop.

Fire-Cured Tobacco. When mature, the leaves of fire-cured tobacco are thick and heavy, mottled yellow and green. Maturity is commonly reached 30 – 40 d after topping. Gum accumulates on the leaves in dry weather during the interval between topping and harvest. Its presence improves leaf quality.

Burley Tobacco. The plants usually reach maturity ca. 3 – 4 weeks after topping. At maturity, plants become a lighter shade of green. The entire plant is harvested by cutting the stalk as near the ground as possible to save the lower leaves. The best cigarette grades are usually found among the lower and middle leaves.

As the plants are cut, they are at once speared onto a stick, one end of which is forced into the ground at an angle, often so that it rests on the stubble of a cut plant.

In general, the bottom leaves on burley tend to burn before it is cut. If these leaves are primed off early, they may bring a better price. Usually, only one or two primings are profitable before the stalk is cut.

Maryland tobacco is harvested much like burley. Growers may vary in exact harvesting methods; most burley growers spear immediately after cutting the stalk, but cut Maryland tobacco plants are allowed to wilt before field spearing or removal for barn spearing.

Dark air-cured tobacco is harvested like burley. After the plants are cut, they must be allowed to wilt before they are put in curing barns.

Cigar Filler Tobacco. For cigar filler tobacco, plants are harvested when fully ripe, as indicated by pale green color, mottled appearance, and increased thickness and brittleness of the leaves. The plants are cut off near the ground when the weather is clear and the plants are free from dew. The plants are speared on sticks and allowed to wilt on rocks.

Cigar Binder Tobacco. Plants of cigar binder tobacco are harvested when the middle leaves are ripe, as indicated by a lighter shade of green. At this stage, the lower leaves of the plant tend to be more or less overripe and the top leaves show varying stages of maturity. The stalks of the plants are cut near the ground with a special hatchet, and the plants are laid on the ground to wilt before they are handled.

Cigar Wrapper Tobacco. In harvesting all shade-grown tobacco, the tobacco leaves are primed as they ripen. Usually, the first priming of three or four lowermost leaves is made at the time the flower head forms. Subsequent pickings of three or four leaves are usually made at weekly intervals, progressing up the plant until all usable leaves are picked. The plants are not topped.

Oriental tobacco is always harvested by priming, with three to five leaves being taken at a time. Five or six primings are usually required; each is called a "hand" or "row". The leaves are picked at a stage just before full ripeness, when the color is yellow-green and the leaf angle to the stem has increased nearer to 90°. Leaves are threaded onto string. Generally $1.5 - 2 \times 10^6$ leaves per acre are harvested ($3.5 - 5 \times 10^6$ per hectare), and this requires much hand labor.

4.6. Curing

Curing is a vital process, and can be considered as a starvation phenomenon. The purpose of curing is to produce dried leaf of suitable physical properties and chemical composition. Respiratory losses, involving hydrolysis and oxidative deamination, may lead to 20 % loss in dry matter in stalk curing. Various regimes of ventilation, temperature, and humidity control are employed for different types of tobacco.

4.6.1. Flue Curing

Flue curing was the system used in early curing barns; it has almost disappeared. Oil, gas, or electrical heating systems may be used. A type of barn called a bulk barn has been gaining in popularity. Bulk barns are especially constructed and purchased prebuilt from the manufacturer. Air is circulated by a fan through a system of ducts and plenum chambers, rather than by natural movement.

4.6.2. Air Curing

Air curing gives satisfactory results in dry, warm weather. When tobacco is hung in the barn, water evaporates from the leaves, and the barn becomes saturated with moisture. Evaporation soon stops, unless the moisture is removed from the barn by circulating the air.

Under favorable conditions, curing periods for various air-cured tobaccos are 4 – 8 weeks. The temperature should be maintained, and adequate moisture provided, but not so much as to cause molding. The tobacco is cured if the midrib of the leaf is dry enough to snap when bent between the fingers; the stalks may still be green at this stage.

Other curing methods include open fire, dark air curing, sun curing, etc. They are used to achieve the specific characteristics of each type of tobacco. Details may be found in [1].

4.7. Fermentation and Aging

Freshly cured tobacco leaf is unfit for use because of its pungent and irritating smoke. After fermentation and aging, the leaf delivers mild and aromatic smoke. Aging is generally applied to cigarette tobacco, i.e., a mild state of fermentation, usually carried out in a hogshead in compressed conditions for one or two years, at moisture content 10 – 13 %. The loss of dry weight is ca. 1 – 2 % for flue-cured and 3 – 4 % for air-cured leaves.

Fermentation, or sweating, is usually applied to cigar tobacco. It is characterized by high initial moisture content (may reach 50 %), by heat generation, and by 10 – 20 % loss of dry weight. In addition to the improvement of leaf aroma, there is a change to a uniform, darker color, and a grainy leaf texture. During fermentation, there is evolution of CO_2, ammonia and other nitrogenous compounds, and methyl alcohol; uptake of O_2; change of pH; change in water retention; and improvement of fire-holding capacity.

During fermentation, there is a marked increase in ammonia and almost complete loss of asparagine amide nitrogen. Amide-hydrolyzing enzymes are evidently present which ultimately decompose all the asparagine synthesized in the early phase of curing. During cigar fermentation, there is a larger loss of soluble carbohydrates; a considerable decrease in organic acids, pectins, pentosans, hemicelluloses, essential oils, resins, waxes, and pigments; and a small change in cellulose and lignin. The fate of nicotine during fermentation of cigar tobacco has been widely studied.

Aging of tobacco is much milder than fermentation, although fermented tobacco leaf may continue to “age”. The first aging step for cured cigarette tobacco leaves is to subject them to a redrying process, and to bring them to a low and uniform moisture content (9 – 10 %). During aging, little self-heating (except at the beginning) takes place. Small amounts of carbon dioxide, acetic acid, formic acid, and ammonia are evolved during aging; there is an increase in moisture and a decrease in sugar, total nitrogen, water-soluble nitrogen, amino nitrogen, nicotine, total acids, and pH. It is believed that the aging of flue-cured tobacco is essentially a chemical process, the main reaction being that between sugars and amino compounds, with the formation of melanoidins and carbon dioxide.

5. Tobacco Products

Unlike other agricultural products, tobacco is mostly consumed through combustion, only a small amount being used for snuff or chewing.

The manufacture of tobacco products is a science as well as an art. Each finished product has its specific characteristics. The physical properties and chemical composition of each leaf tobacco differ with variety, location, year, environmental variables, postharvest handling, including curing, storage, and aging, and many other factors. Selection of leaf and blending formula are of the greatest importance. Flavoring

and other additives can only "modify" the natural material to a certain degree; casing with sugar may reduce harshness, fructose is used to neutralize irritancy. Other commonly used materials include liquorice, glycerine, honey, vanilla, rum, tonka beans, coumarin, certain acids, aldehydes, lactones, ketones, esters, essential oils, and hundreds of others. The formula is known only by the manufacturer.

5.1. Cigarettes

There are two distinct stages; leaf processing to arrive at the thoroughly mixed blend of cut rag, followed by the use of this rag to make the cigarette.

Where a blended cigarette is being made, from several tobacco types (e.g., flue-cured, burley, and oriental), and one or more of them may be cased, there are as many leaf processing lines as there are types and differential treatments.

The width of cut affects smoking characteristics and cigarette hardness. The finer it is the more complete the combustion, and the coarser it is, within limits, the more the physical characteristics of the cigarette can be preserved at lower tobacco densities. Very thin tobaccos with relatively poor moisture characteristics, such as burley and Maryland, have a high manufacturing loss if cut too fine. The cut for flue-cured cigarette tobacco is 33 – 37 to the inch (0.770 –0.686 mm). The United States custom is for a coarser cut, in line with the above considerations, at 28 – 32 cuts to the inch (0.907 – 0.794 mm). Cut rag must be dried, cooled, and mixed with the cut rolled stem, and stored in silos or bins prior to being fed to the making machines.

In the United States, much or all of the stem is incorporated in the cigarette as reconstituted sheet. In the United Kingdom and many other countries, it is used as cut rolled stem.

Other Cigarette Components The paper has a profound effect on the yield and composition of smoke. One critical paper characteristic is permeability; its primary physical influence is smoke dilution.

Calcium carbonate is used to influence the permeability. Magnesium carbonate may be added to improve ash color, or titanium oxide if a particularly white ash is required.

Another nontobacco part of the cigarette, which has vital influence on the smoking characteristics, is the filter. It may be made from cellulose acetate fibre or paper. The acetate and paper modify the particulate smoke phase by particle retention, and finely divided carbon modifies the gaseous phase.

5.2. Cigars

A typical cigar is 85 % filler, 10.5 % binder, and 4.5 % wrapper. These proportions refer to the weights of leaf actually performing those functions.

Manufacture is essentially forming the filler into the shape of the cigar, wrapping it in the binder to give a bunch, and then covering the whole with the wrapper.

5.3. Pipe Tobacco

"Smoking tobacco" generally refers to pipe tobacco, a mixture of flue-cured with some sun-, air-, and fire-cured leaf. In the United States, the heavy burley leaf forms either a substantial part or the whole of pipe tobacco blends. Perique and Latakia may be added in small quantities to provide their distinct flavors. United States pipe tobaccos are very heavily cased and many United Kingdom brands are top flavored.

The basic problems of mixing the various blend constituents and correctly conditioning the leaf are similar in principle to those with cigarette tobaccos, although moisture levels are much higher in smoking tobaccos than in cigarettes or cigars. In the United Kingdom, the legal maximum is 32 %; typical product mixtures are 20 – 24 %, flake 22 – 26 %, and plug and bar 28 – 30 %.

5.4. Snuff

Fire-cured and dark air-cured tobaccos are used to make snuff. Leaf destined for snuff has been packed and aged in the normal way, and generally the whole leaf is used, but some may be in strip form. Following heavy reconditioning, the

tobacco is repacked to undergo a severe fermentation. In the United States, the tobacco may be cut into strips 1 – 2 in (2.5 – 5.0 cm) wide, before fermentation. The fermented tobacco is dried completely and ground to a fine powder, which must pass a silk bolting cloth containing up to 96 threads per inch (38 per cm). In the United States, snuff may be either blended with a number of flavorings and scents, or left plain. Thee is a dry powder which is placed between the lower lip and the gum, or a moist form which is placed between the cheek and the gum. These forms of consumption differ from the European habit of inhaling through each nostril in turn.

5.5. Chewing Tobacco

In the United States, the four types of chewing tobacco are, in order of sales, loose leaf (scrap), plug, fine-cut, and twist. The basic ingredients are the dark air-cured tobaccos of Pennsylvania and Wisconsin which used to find outlets as cigar binders and fillers. The leaf is stemmed, cut, or granulated, and then either cased heavily (sweet form), cased lightly (semisweet form), or not cased at all. It has been estimated that a sweet form contains, in each 1 pound (0.45 kg) of product, 0.56 pound (0.25 kg) tobacco materials and 0.44 pound (0.20 kg) flavoring additives and moisture.

In the United Kingdom, chewing tobacco is made only as plug. It is unsweetened and usually made of fire-cured tobacco. The wrapper is a particularly fine, thin, well-textured leaf. The moisture content of manufactured plug is about 30 %.

6. Plant Physiology

6.1. Growth and Development

Morphological development of tobacco plants greatly affects the physical and chemical characteristics of tobacco leaf, and thus its quality and usability. Among the most fascinating things about higher plants are the dynamic changes that occur during the life cycle, i.e., growth and development. The ability to function as an autotroph, and to synthesize large, often complex biomolecules from inorganic ions, H_2O, and CO_2 leads not only to new cells, but to new cells that differentiate into tissues with specific functions, e.g., roots, leaves, or flowers.

Environmental and physiological factors affecting growth and development include carbon dioxide, nutrients, temperature, water, and light. Seed germination is the first step of plant growth. The three distinct stages in germination of tobacco seed are: imbibition of water; cell elongation; and increase in cell number. In a physiological sense, the start of germination depends on the coupling of respiration to growth.

In tobacco plants, yield increases with increased total root number and surface area. Root branching and elongation ratios are greater in plants with early rather than late transplanting; at maturing stage, this is reversed.

Cell division may be controlled by auxins and cytokinins; cell elongation and enlargement may be controlled by auxins and gibberelins; while maturation of cells can be influenced by ethylene.

At tissue level, auxin affects the differentiation of vascular tissues and root initiation, and may stimulate cell division in certain tissues and stimulate cell enlargement. Gibberelins may stimulate flowering and cell elongation. Cytokinins are cell division factors and ethylene stimulates aging of tissues in general and the formation of abscissian layers of aging tissues (leaves and fruit) in particular. Abscisic acid is a growth inhibitor, but more importantly, it adapts tissues to deal with stress imposed by salt and/or water.

Senescence is the phase of plant growth that extends from full maturity to actual death, and is characterized by an accumulation of metabolic products, increase in respiratory rate, and loss in dry matter. Many physiological and biochemical changes are associated with senescence of leaves, including yellowing or decrease in chlorophyll, the decline in proteins, RNA, DNA, and increased activity of hydrolytic enzymes such as amylase, invertase, and peroxidase.

6.2. Primary Nutrition

Mineral nutrition is among the most important factors in producing a high-quality leaf tobacco.

Primary elements include N, P, K, and Ca, of which nitrogen is the most important element for growth and development of the tobacco plant and leaf quality, particularly nitrogenous fractions such as protein, amino acids, or nicotine. Nitrogen moves continuously within a plant. Nitrate is the most effective form. The most critical period at which nitrogen exercises the greatest effect on final leaf area is the time prior to emergence of the leaf from the bud, which includes the period of active cell division. Nicotine content increases with increasing nitrogen rates and decreases with increasing irrigation.

Phosphorus is vital in plant metabolism. It is important for plants to have access to a readily available source of phosphate at an early growth stage. In young tobacco leaves, ca. 30 % of the phosphorus is present as RNA and 7 % as DNA.

Although potassium is essential to tobacco growth, its exact role in the metabolic system is not well defined. Potassium is the principal constituent of ash and is important in certain enzyme systems. Tobacco is known as a "luxury" user of potash.

Calcium is essential to cell wall structure and is one of the principal tobacco ash components, next only to potassium. Generally present in the plant as insoluble salts, calcium is not very mobile. It functions as a general regulator of cell metabolism, and may also prevent injury from an excess of other ions.

6.3. Trace Elements and Heavy Metals

Almost all elements are detected in tobacco, including heavy metals. Some are of very minute quantity, which may originate through contaminated soil; some are of significance to tobacco development and growth, e.g., boron, chlorine, copper, iron, magnesium, manganese, molybdenum, selenium, sodium, sulfur, and zinc. Detailed information can be found in [1–3].

6.4. Nutritional Disorders

A deficiency of any element essential for the normal metabolic processes of tobacco plants results in visible or chemical abnormalities, or both. Table 4 is a key to mineral deficiency symptoms on tobacco.

On the other hand, mineral excess can induce physiological disorder in the tobacco plant. Studies in this area are not as extensive as in deficiency. It is known that excess nitrogen produces dark green leaves, promotes vegetative growth, and delays maturity. Excess phosphorus may reduce yield and thicken the leaves. High potassium delays maturity, but increases the rate of burn. Toxicity of excessive minor elements is more frequently reported.

6.5. Physiological Disorder

Physiological disorder of tobacco plants can be induced by various factors including environmental, genetic, and biochemical abnormalities. These abnormalities disturb the delicate balance of normal metabolism, and thus induce chemical and physical changes, reflected in plant growth or leaf appearance.

Pollutants. Many factors contribute to air pollutants, but the most significant source is from energy conversion. Energy conversion products include water, carbon dioxide, carbon monoxide, sulfur dioxide, sulfuric acid, hydrogen sulfide, nitric oxide, nitrogen dioxide, hydrogen fluoroxide, ethylene, ozone, aldehydes, soot, and hydrocarbons. Other materials such as radioactive materials, and many other organic and inorganic wastes are also present as pollutants.

Tobacco "weather fleck", a physiological disorder of the leaf, is associated with the presence of high levels of ozone. Synergism between ozone and sulfur dioxide has been observed. Tobacco plants which did not show damage symptoms when exposed to low-level individual gases, displayed ozone-type injury when exposed to mixtures of ozone and sulfur dioxide at subthreshold concentrations. A separate study showed that ozone was more injurious than SO_2, but in combination the damaging effect of both pollutants on tobacco plants was increased.

Frenching is a noninfectious physiological disorder of tobacco. It usually appears as a network chlorosis of apical leaves, followed by formation of progressively narrower leaves. In extreme cases, there is no expansion of lamina in

Table 4. Key to mineral deficiency symptoms in tobacco

1. Effects localized on older or lower leaves or more or less general on whole plant	(Group 1)
1.1. Local: occurring as mottling or chlorosis with or without necrotic spotting of lower leaves; little or no drying up of lower leaves	
1.1.1. Lower leaves curved or cupped under with yellowish mottling at tips and margins; necrotic spots at tips and margins	potassium
1.1.2. Lower leaves chlorotic between the principal veins at tips, and margins of a light-green to white color; typically, no necrotic spots	magnesium
1.2. General: also yellowing and drying or "firing" of lower leaves	
1.2.1. Plant light green, lower leaves yellow, drying to light-brown color	nitrogen
1.2.2. Plants dark green, leaves narrow in proportion to length; plants immature	phosphorus
2. Effects localized on terminal growth, consisting of upper and bud leaves	(Group 2)
2.1. Dieback involving the terminal bud, preceded by peculiar distortions and necrosis at the tips or base of young leaves making up the terminal growth	
2.1.1. Young leaves making up the terminal bud first light green, followed by a typical hooking downward at tips, followed by necrosis, so that if later growth takes place, tips and margins of the upper leaves are missing	calcium
2.1.2. Young leaves constricted and light green at base, followed by more or less decomposition at leaf base; if later growth takes place, leaves show a twisted or distorted development; broken leaves show blackening of vascular tissue	boron
2.2. Terminal bud remains alive, chlorosis of upper or bud leaves, with or without necrotic spots, veins light or dark green	
2.2.1. Young leaves with necrotic spots scattered over chlorotic leaf, smallest veins tend to remain green, producing a checkered effect	manganese
2.2.2. Young leaves without necrotic spots, chlorosis does or does not involve veins so as to make them dark or light green in color	
2.2.3. Young leaves with veins of a light-green color or of same shade as intervein tissue; color light green, never white or yellow; lower leaves do not dry up	sulfur
2.2.4. Young leaves chlorotic, principal veins characteristically darker green than tissue between the veins; when veins lose their color, all leaf tissue is white or yellow	iron

younger leaves (strapping) or stalk elongation. A series of studies on tobacco concluded that frenching is caused by an organic toxin produced by *Bacillus cereus* Frankland and Frankland. It is suggested that diffusates from *B. cereus*, and perhaps other soil bacteria, may be the cause of frenching of tobacco in the field. Increased populations of the bacillus have been found in adjacent soil and rhizospheres of frenched plants. Diffusion products of these bacteria and the dead bacteria cause symptoms of frenching in aseptically grown tobacco seedlings.

Genetic Tumors. *Nicotiana* tumors may grow spontaneously on stems, roots, leaves, or flower systems on tumorous F_1 progeny or only part of the tissue. Certain *Nicotiana* hybrids produce teratoid proliferations of a genetic origin. The species involved in the production of tumorous interspecific *Nicotiana* hybrids can be separated into two groups – the plus group and the minus group. Hybrids among plus species or hybrids among minus species are free from tumors.

Wounding of leaves of the tobacco hybrid *Nicotiana glauca*×*N. langsdorffii* caused tumor formation at the site of the wound where metabolites would have collected under normal transport conditions.

7. Biochemistry

7.1. Alkaloids

General Concepts. In *Nicotiana* plants, most alkaloids are 3-pyridyl derivatives. Indole alkaloids, such as harmane and norharmane, are also reported, but in very minute quantities. Among the many alkaloid materials found in *Nicotiana*, nicotine is considered the principal alkaloid in commercial tobaccos, and a report confirmed this in 34 of 65 species [1, pp. 427]. Nornicotine appears to be the main alkaloid in 19 of 65 species of *Nicotiana*, and anabasine is the third most important. In another study on alkaloids of 60 *Nicotiana* species, the kinds and amounts, varying with each species, gave no clear-cut correlation between alkaloid amounts and subgeneric or sectional classification [1, pp. 427].

Many 3-pyridyl alkaloids have been reported in *Nicotiana* plants. In cured leaf, 24 pyridine

derivatives have been separated, including nicotine, nornicotine, anabasine, oxynicotine, myosmine, 3-acetylpyridine, 2,3′-dipyridyl, nicotinamide, anatabine, nicotinic acid, and unidentified pyridine alkaloids or derivatives. Major tobacco alkaloids are shown in Figure 1.

Many explanations have been suggested for the presence of alkaloids in plants:

1) Alkaloids serve as protection against insects and herbivores
2) Alkaloids are detoxification products
3) Alkaloids may serve as reserves
4) Alkaloids may include regulatory substances, such as nicotinic acid
5) Alkaloids are useful to the existence of the plant
6) Alkaloids are waste products of metabolism

Nature and Amount of Alkaloids in Nicotiana Plants. Tobacco types, plant parts, cultural practices, degree of ripening, and fertilizer treatment are among the factors which determine the levels of alkaloids in *Nicotiana* plants. Every step in tobacco production that affects plant metabolism influences the level of alkaloid content to a certain degree.

Tobacco leaves have the highest content of nicotine, roots have less, and stalks have the least. Within a leaf, the tip and outer area usually have higher alkaloid content than the base and inner area, respectively. Alkaloid level increases as plants mature, especially during the period after topping. Marked increase of nicotine generally accompanies increased rate of nitrogen fertilization. Under favorable conditions, *N. rustica* consistently produces more nicotine than commercial *N. tabacum*. When grown on fertile soils under irrigation over a period of years, *N. rustica* can consistently produce 150 lb (68 kg) or more of nicotine per acre (165 kg/ha), approximately double the quantity obtained from ordinary tobacco. In *N. tabacum*, a wide variation exists in alkaloid content among varieties. From a random examination of 152 cultivated varieties, a range of alkaloid content between 0.17 and 4.93 % has been found.

Biosynthesis of Alkaloids. Numerous studies indicate that nicotine is formed in the tobacco root and translocated to the top through xylem; nornicotine is the demethylation product of nicotine in the shoot of a plant; and anabasine can be formed in both root and shoot. However, there is evidence of alkaloid formation at other sites in the tobacco plant, e.g., independent nicotine synthesis in both root and aerial portions. Studies on the biosynthesis of tobacco alkaloids have been widely conducted, starting in the early 1930s, and were most active in 1950–1960. They involved areas such as the source of carbon, nitrogen, and hydrogen; the origin of the pyridine ring, the pyrrolidine ring, the piperidine ring, and the *N*-methyl group; enzyme activities, and the interconversion and degradation of alkaloids. In addition, tobacco-specific *N*-nitrosamines, many of which may be related to health problems, are being intensively examined. All these areas are of specific scientific interest as well as health concern.

7.2. Proteins

Tobacco leaf, like other higher plant leaves, contains two classes of proteins: soluble and insoluble, usually equal in quantity. The best-studied leaf protein is the soluble portion, which can be further divided into two major categories on the basis of molecular size. One having a sedimentation value of 18 *S*, representing a single protein is the fraction 1 protein (F 1 protein), whereas all the remaining smaller (4–6 *S*) soluble proteins combine to form the unfractionated proteins or fraction 2 protein (F 2 protein). Generally, there are equal amounts of F 1 and F 2 proteins in the earlier stages of development. Approaching maturation, this proportion is altered, mainly by degradation of F 1 protein. Although our knowledge of F 1 protein is rather comprehensive, there is little information on F 2 protein.

The soluble F 1 protein is found in all organisms containing chlorophyll a, and is ribulose-1,5-diphosphate (RUBP) carboxylase–oxygenase (rubisco) in higher plants. This enzyme has a dual function in that it catalyzes both the carboxylation and oxygenation of RUBP, i.e., the crucial reactions of both photosynthesis and photorespiration; the ratio of these two processes determines the plant's productivity.

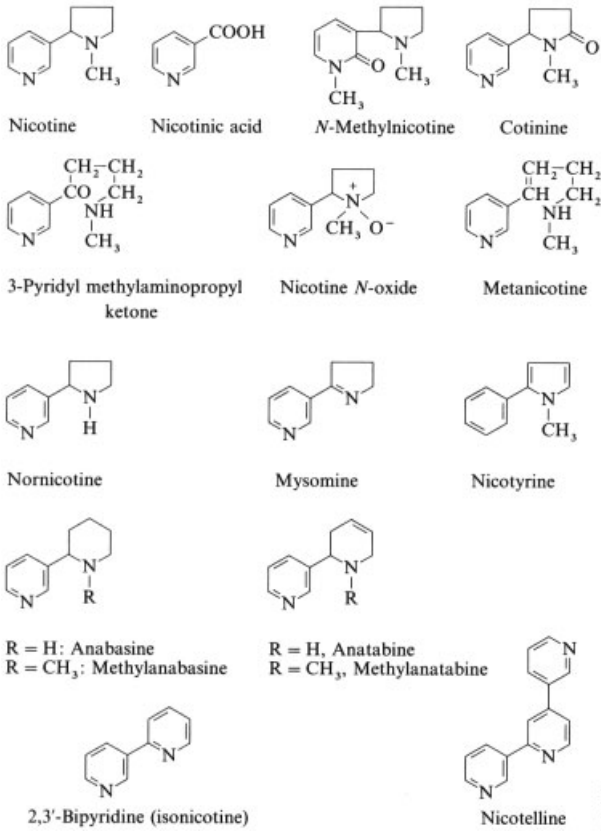

Figure 1. Major tobacco alkaloids and related compounds

The essential amino acid compositions of tobacco F 1 protein, egg protein, and milk proteins are similar.

7.3. Carbohydrates and Organic Acids

Carbohydrates and organic acids are known to affect leaf quality. They are major products of plant metabolism, with carbohydrates being present in the greatest amount of any class of compounds in tobacco leaf. Carbohydrates and organic acids are of equal significance to cultivars, cultural practices, and curing in differentiating major tobacco types. As much as 75 % of all the carbon from CO_2 reduced by the plant enters into polysaccharides. Obviously, because of the CO_2 fixation mechanism all the CO_2 fixed by the tobacco plant is in a carbohydrate for at least a brief time.

The main products of CO_2 fixation are sucrose and starch. Sucrose is the primary carbohydrate transported in tobacco, and starch accumulation in the chloroplast is a temporary storage form.

Fructose and glucose, whether free or combined in sucrose, are equally available for starch synthesis and respiration. The rate of sucrose formation may control the rate of starch formation, and the sucrose-phosphate synthetase may be involved in regulation of partitioning of carbon between sucrose and starch. Saccharide is accumulated in the leaves during maturity and decomposed during curing.

In growing plants, an important sink for carbohydrate from photosynthesis is the polysaccharides of the cell wall. Until a leaf is about two-thirds expanded, virtually all carbon fixed in the leaf remains in the leaf; only in the later

stages of leaf development is sucrose translocated to other sinks of the plant.

Primary cell walls have pectin, hemicellulose, and cellulose fractions. Primary cell wall formation utilizes five sugars – glucose, mannose, galactose, xylose, and arabinose – and five sugar derivatives – glucuronic acid, galacturonic acid, rhamnose, fructose, and methyl esters of galacturonic acid.

Nonstructural carbohydrates are sugars or starch, readily available for metabolic processes for energy production and growth; concentrations of nonstructural carbohydrates are subject to drastic changes during curing. Structural carbohydrates are polymers of monosaccharides, most frequently of glucose, galactose, mannose, arabinose, xylose, galacturonic acid, and glucuronic acid. Structural carbohydrates are outside the normal metabolic pools in the plant, and once formed have a very low turnover, and consequently change very little during curing. Midrib contains more cellulose, with higher polymerization of cellulose than in lamina. Lower grade tobacco appears to have cellulose properties more like those of midrib than higher grade leaf.

Many organic acids are known to be present in leaf tobacco. Among nonvolatile acids in leaf tobacco, the major ones are citric, malic, and oxalic; the minor ones are glycolic, succinic, malonic, fumaric, and pyruvic. Among volatile acids in tobacco, acetic and formic acids are the major ones; the minor ones are propionic, α-furanic, benzoic, α-methylbutyric, β-methylvaleric, and many others.

Salts of nonvolatile organic acids, e.g., potassium, magnesium, and calcium salts of organic acids, are always present. Oxalic acid occurs mainly as the calcium salt which, in burley varieties, has been reported to affect the availability of calcium to plant growth, with apparent calcium deficiency as a result.

From a series of studies on the metabolism of organic acids in tobacco leaves, it is evident that when the acid supplied to the plant is one that is metabolically allied with the acids normally occurring, this acid enters readily into the reactions in the cell, and alters the composition of the leaf. This is true for malic, citric, *d*-isocitric, succinic, fumaric, lactic, and pyruvic acid. The only exception is oxalic acid, which appears to be an inert end product.

7.4. Pigments and Polyphenols

Pigments and polyphenols are among the major components which influence leaf quality and usability. Field leaf color is often considered an indicator of the physiological state of growth, and cured leaf color may reflect market value. Phenolic compounds play an important role in leaf color, quality, and physiological strength of smoke. In recent research, the amount and alteration of the amount of free phenols in the smoke has become more important than polyphenols in the tobacco leaf per se.

In many aspects, pigment and polyphenol content of the plant reflect the genetic makeup, plant growth and development, response to pest infection, physical damage, biochemical changes of postharvest treatments, and thus the leaf acceptability and smoke quality. The "brown pigment" is an example which represents a complex product – mainly pigment, polyphenols, protein, and alkaloids.

The major plant pigments in green tobacco are chlorphyll a, chlorphyll b, lutein, β-carotene, neoxanthin, and violaxanthin. Generally, the chlorophyll content of green leaves is 0.5 – 4 %. The total amount of the yellow pigments, carotene and xanthophyll, is about one-fifth to one-third that of chlorophyll in mature tobacco leaves. Carotene from flue-cured tobacco leaves is composed of a mixture of β-carotene and neo-β-carotene in the proportion of 68 : 32, while xanthophyll contains 50 % lutein, 22 % neoxanthin, and 18 % violaxanthin.

The flue-cured varieties have a higher chlorophyll content than the Turkish varieties, and the Turkish tobaccos contain more chlorophyll than the burley types. Only small differences have been observed in the chlorophyll a : chlorophyll b ratio among varieties, the range being 2.14 – 2.40.

In flue-cured tobacco, chlorophyll degradation accompanies individual leaf senescence, while mature burley plants show a progressive decline in chlorophyll concentration from the top to the bottom. In addition, to chlorophyll a and b, β-carotene, lutein, violaxanthin, and neoxanthin, the leaves of certain tobaccos contain antheraxanthin in some harvests.

During leaf curing, the less mature the leaf, the longer the time required for yellowing.

Losses of yellow pigments and dry matter apparently proceed independently of chlorophyll disappearance.

Chlorophyll decreases sharply in the early states of air curing. Carotenoid content also decreases, but the change is rather gradual and incomplete. Rutin and chlorogenic acid remain substantially unchanged in the yellowing stage and decrease in the stages that follow. Loss of chlorogenic acid is almost complete.

Phenolic compounds in living plants occur almost exclusively as glycosides and esters. Polyphenols and related compounds in tobacco include tannins, coumarins, flavanols, simple phenolic derivatives, anthocyanines, and hydroxylated cyclohexanes.

Caffetannins are the only tannins found in tobacco. Chlorogenic acid, its geometric isomer, and their hydrolytic products, caffeic and quinic acids, are the predominant tannin precursors reported. Among the coumarins, scopoletin and its glycosidic derivatives are most frequently found; among the flavonoids, several flavonols are observed, including the glycosides, rutin, isoquercetin, and kaempferol-3-rhamnoglucoside. The predominant anthocyanine is cyanidin-3-rutinoside in flowers of several *Nicotiana* plants. Many other phenolic and related compounds are also reported in leaf tobacco, including catechol, *p*-coumaric acid, *p*-coumerylquinic acid, esculetin, eugenol, ferulic acid, guaiacol, melilotic acid, phenol, quinic acid, shikimic acid, syringic acid, vanillic acid, and others. Quinic acid is believed to be a normal component of leaf tobacco, and is easily lost during drying, which may be the reason for failure to detect this compound in some studies. Pelargonidine-3-rutinoside and two flavonols, quercetin and kaempferol, have also been found.

There is a greater loss of soluble polyphenols during air curing than during flue curing. Results indicate that the interaction of cultural practices with chlorophyll-deficient genes affects the metabolism of chlorophyll and polyphenols during leaf growth and maturation.

During fire curing and air curing, most of the major polyphenols except scopolin and scopoletin are destroyed. These compounds are usually present in both air- and flue-cured tobaccos. There is a progressive increase in scopoletin during flue curing. It is believed that the extra scopoletin arises through hydrolysis of the glucoside scopolin.

7.5. Other Major Compounds

Fatty compounds, lipids, waxes, phytosterols, and related compounds may contribute to leaf aroma; serve as growth regulators; serve as indicators of plant development and processing; serve as bases for chemotaxonomy; and most important, serve as precursors of polynuclear aromatic hydrocarbons (PAH) of smoke condensate.

All aliphatic carboxylic acids may be classified as fatty acids. Usually, fatty acids with ten or fewer carbons, e.g., butyric, caproic, caprylic, capric, are considered lower fatty acids.

Many long-chain fatty acids are present in tobacco seed and leaf. In the tobacco seed oils, the fatty acids are: 75 % linoleic (18 : 2), 15 % oleic (18 : 1), 7 % palmitic (C_{16}), and 3 % stearic acids (C_{18}). In cured bright leaf tobacco, 16 methyl ester compounds have been determined, including lauric (C_{12}), myristic (C_{14}), palmitic (C_{16}), stearic (C_{18}), oleic (18 : 1), linoleic (18 : 2), and linolenic (18 : 3) acids.

The presence of several lower and fatty alcohols in bright tobacco, including *n*-propanol, *n*-butanol, 1-docosanol, 1-eicosanol, ethanol, 1-heneicosanol, 1-heptadecanol, methanol, 3-methyl-1-pentanol, 1-nonadecanol, 1-octadecanol, and 1-tricosanol has been reported.

The characteristic aroma and taste of flue- or air-cured leaves are from the leaf surface lipid fraction. Both leaves contain almost the same amount of alkanes. Terpenes or their degradation products are the components which contribute to aroma and taste of tobacco.

Hydrocarbons are of special significance because they are related to leaf aroma. Analysis of the normal hydrocarbons isolated from tobacco aroma fraction shows the presence of homologues varying from C_{14} to C_{33}. The major fraction, however, is unsaturated material (neophytadiene) and branched hydrocarbons. Alkanes are the major constituent of epicuticular wax, universally distributed in plants. Tobacco leaves contain 5 – 10 mg alkanes per 100 cm^2, gradually increasing with leaf age. Leaves on the upper stalk contain more alkane than those on the lower stalk.

Tobacco terpenoids appear to be related to the flavor and aroma of the product. In the past, tobacco researchers have referred to terpenoids as compounds of hexane or petroleum-ether extractables.

Terpenoids are derived from a basic 5-carbon, isopentene unit. Some terpenes are vital to plant growth; some appear to be secondary metabolites. Terpenes are classified on the basis of number of carbon atoms: hemiterpenes (C_5), monoterpenes (C_{10}), sesquiterpenes (C_{15}), diterpenes (C_{20}), triterpenes (C_{30}), tetraterpenes (C_{40}), and polyterpenes (C_{50} and greater). A study of the diterpenes of *Nicotiana* has shown that they are synthesized within the trichomes, and excreted on to the leaf surface.

Two major classes of diterpenes have been identified in surface extracts and trichome exudates of *Nicotiana* species and *N. tabacum* cultivars. Compounds of the first class, the labdanoids, have been isolated from aged, sun-cured Turkish tobacco, and identified as diterpene lactones belonging to the group of diterpenes typified by labdanoic acid. The labdanoids have a C_6 sidechain and a perhydronaphthalene ring. In general, the labdanes are produced by Turkish, Oriental, and Greek tobacco and some species of *Nicotiana,* while they are absent from *N. tabacum* cultivars of flue-cured, burley, and Maryland types.

The second major class of diterpenes found in the surface washes and trichomes or trichome exudates are macrocyclic, with 14 carbon atoms in the ring. The first of these diterpenes were isolated from aged burley tobacco. Based on chemical and physiochemical properties, these deterpenes have been shown to be diasteroisomers of 12-isopropyl-1,5,9-trimethyl-4,8,13-cyclotetradecatriene-1,3-diol. The common name duvane was assigned to the parent saturated hydrocarbon.

8. Chemical Characteristics

Chemical composition of leaf tobacco varies with genetic makeup, environmental conditions, and every step of production and handling. Major groups and some compounds reported present in leaf tobacco are listed in the following.(see below)

With improved analytical methods and modern instrumentation, minute quantities of many more compounds are likely to be identified in tobacco. The general list summarizes the information available up to now.

In addition, major components representing the differences among principal tobacco types, such as flue-cured, burley, Maryland, and Turkish are listed in Table 5.

Acids

Acetic	Arachidic
Arachidonic	Auxin and indoleacetic acid
Azelaic	Benzoic
Butyric	$C_{10}-C_{23}^{(saturated)}$
$C_{10}-C_{14}^{(normal)}$	$C_{15}-C_{26}^{(iso,\ anteiso)}$
$C_{16}-C_{18}^{(hydroxy)}$	$C_{22}-C_{25}^{(cyclohexyl)}$
Citric	Crotonic
Decanoic	α-Fluorenecarboxylic
Formic	Fumaric
Furoic	D-Glyceric
Glycolic	Glyoxylic
Heptanoic	α-Hydroxyisocaproic
β-Hydroxyiso-caproic	α-Hydroxy-β-methylvaleric
	Hydroxypyruvic
β-Hydroxy-β-methylvaleric	Isobutyric
	2-Isopropylmalic
α-Hydroxyvaleric	α-Ketoglutaric
Isocaproic	Lauric
Isovaleric	Linolenic
Lactic	Malic
Linoleic	α-Methylbutyric
Maleic	
Malonic	
β-Methylvaleric	Octanoic
Nonanoic	Oxalacetic
Oleic	Palmitic
Oxalic	α-Phenyllactic
Phenylacetic	Propionic
Phenylpyruvic	Stearic
Pyruvic	Terephthalic
Succinic	
Valeric	

Alcohols

Aliphatic	
1-Docosanol	1-Eicosanol
Ethanol	1-Heneicosanol
1-Heptadecanol	Methanol
3-Methyl-1-pentanol	1-Nonadecanol
1-Octadecanol	1-Tricosanol
Aromatic	
Benzyl alcohol	β-Phenethyl alcohol
Diethylene glycol	Glycerol
Propylene glycol	Triethylene glycol
Cyclic	
Furfuryl alcohol	Inositol
Menthol	

Aldehydes, Ketones, Quinones

Aldehydes
- Acetaldehyde
- Acrolein
- *p*-Anisaldehyde
- Benzaldehyde
- Butyraldehyde
- Crotonaldehyde
- Formaldehyde
- Furfural
- Glycolaldehyde
- Glyoxal
- 5-Hydroxymethyl-furfural
- Isobutyraldehyde
- Isovaleraldehyde
- Mesoxaldialdehyde
- 5-Methylfurfural
- Methylglyoxal
- Methylreductone
- Propionaldehyde
- Reductone
- *m*-Tolualdehyde
- Valeraldehyde

Ketones
- Acetone
- 2-Butanone
- 4-Methyl-2-pentanone
- Methyl-α-pyrryl ketone
- 2-Pentanone

Quinones
- 9,10-Anthraquinone
- 2-Methyl-9,10-anthraquinone
- 2,3-Dimethyl-1,4-anthraquininone

Alkaloids and Other Bases
- Adenine
- Ammonia
- Anabasine
- Anatabine
- Cotinine
- Ethylamine
- Dimethylamine
- Harmane
- Guanine
- Isobutylamine
- Isoamylamine
- Methylamine
- Isonicotine (2,3′-bipyridyl)
- *N*-Methylanatabine
- *N*-Methyl-2-phenylethylamine
- *N*-Methylanabasine
- *N*-Methylpyrrolidine
- *N*-Methylnicotinamide
- Myosmine
- Nicotinamide
- 2-Methylpyrrolidine
- Nicotine *N*-oxide
- *N*-Methyl-3-pyrroline
- Nicotyrine
- Nicotelline
- Nornicotine
- Nicotine
- Pyridine
- Nicotinic acid
- 3-Pyridyl propylketone
- Norharmane
- 1,2,5,6-Tetrahydropyridine
- 2-Phenylethylamine
- Trimethylamine
- 3-Pyridyl methyl ketone
- Pyrrolidine
- Thymine

Alkenes and Alkynes
- α-Carotene
- β-Carotene
- Neo-β-carotene
- Neophytadiene
- Phytoene
- Phytofluene

Amino Acids and Related Compounds
- α-Alanine
- β-Alanine
- α-Aminoadipic acid
- α-Aminobutyric acid
- τ-Aminobutyric acid
- Arginine
- Asparagine
- Aspartic acid
- Betaine
- Choline
- Citrulline
- Cysteic acid
- Cysteine
- Cystine
- Glutamic acid
- Glutamine
- Glutathione
- Glycine
- Histidine
- Homocystine
- Homoserine
- Hydroxyproline
- Isoleucine
- Leucine
- Lysine
- Methionine
- Methionine sulfone
- 1-Methylhistidine
- Norleucine
- Ornithine
- Phenylalanine
- Pipecolic acid
- Proline
- Pyrrolidine-2-acetic acid
- Serine
- Taurine
- Threonine
- Tryptophan
- Tyramine
- Tyrosine
- Valine

Carbohydrates
- Arabinose
- Arabogalactan
- Cellulose
- 1-Deoxy-1-L-alanino-D-fructose
- 1-Deoxy-1-(*N*-τ-aminobutyric acid)-D-fructose
- 1-Deoxy-1-L-proline-D-fructose
- Deoxyribose
- Erythrose
- Fructose
- Galactan
- Galactosamine
- Galactose
- Galacturonic acid
- Glucosamine
- Glucose
- Salt of gum-like polysaccharide
- Lignin
- Maltose
- Mannose
- Pectins
- Pentoses
- Planteose
- Raffinose
- Rhamnose
- Ribose
- Rutinose
- Sorbitol
- Starchyose
- Starch
- Sucrose
- Xylan
- Xylose

Esters
- β-Amyrenyl esters
- Benzyl acetate
- Dibutylphthalate
- Di(2-ethylhexyl) phthalate
- Dipropylphthalate
- Ethyl isovalerate
- Ethyl-β-methylvalerate
- Ethyl propionate
- Glycerides
- Ethyl valerate
- Ethyl butyrate
- Ethyl acetate
- Methyl and ethyl esters of higher fatty acids
- Ethyl caproate
- Solanesyl esters
- Methyl salicylate
- Undecyl acetate
- β-Phenethyl acetate
- Steryl esters

Inorganic Elements

Aluminum	Arsenic
Barium	Beryllium
Boron	Calcium
Cesium	Chlorine
Chromium	Cobalt
Copper	Fluorine
Iodine	Iron
Lead	Lithium
Magnesium	Manganese
Mercury	Molybdenum
Nickel	Polonium
Potassium	Radium
Rubidium	Selenium
Silicon	Silver
Sodium	Strontium
Thallium	Tin
Titanium	Uranium
Vanadium	Zinc

Miscellaneous Components

$C_{10}H_{14}O$	Chlorophyll
Nucleic acids	Phosphatides
Resins	Saponins
Silicones	

***N*-Heterocyclic Compounds**

2-Acetyl-3-methylpyrazine
Acetylpyrazine
2-Acetylpyridine
2-Acetylpyrrolidine
2-Acetylpyrrole
6-Ethyl-2-methylpyrazine
2-Ethylpyridine
3-Ethylpyridine
2-Formylpyrrole
Maleimide
1-Methyl-2-acetylpyrrole
1-Methyl-3-acetylpyrrole
2-Methylpyridine
1-Methylpyrrole
Piperidine
2-Pyridone
Pyrazine
Pyridine
Pyrrolidine
Pyrrolidin-2-on
Quinoline
Succinimide
Trimethylpyrazine

Phenols and Related Compounds

4-Allylcatechol	*p*-Anisaldehyde
Caffeic acid	1-*O*-Caffeoylglucose
4-Caffeoylquinic acid	Catechol
Chlorogenic acid	*p*-Coumaric acid
p-Coumarylquinic acid	*m*-Cresol
Esculetin	Esculetin 7-glucoside
Eugenol	Ferulic acid
1-*O*-Feruloyl glucose acid	Guaiacol
Hydrocaffeic acid	Hydroxyacetophenone
	(*o*-,*m*-,*p*-) 3-Feruloylquinic acid
(*m*-,*p*-)Hydroxybenzaldehyde	3-Hydrobenzoic acid
4-Hydroxybenzoic acid	2-Hydroxyphenylacetic acid
3-Hydroxyphenylacetic acid	3-Hydroxyphenylpropionic acid
4-Hydroxyphenylacetic acid	Isoeugenol
Isoquercetin	Kaempferol 3-rhamnoglucoside
Melilotic acid	Methyl salicylate
Naringenin	Naringin
Neochlorogenic acid	Phenol
Protocatechuic acid	Protocatechuic aldehyde
Quercetin methyl esters	Quercimertrin
Quinic acid	Rutin
Salicylaldehyde	Scopoletin
Scopoletin-7-glucoside	Scopoletin rhamnoglucoside
Shikimic acid	Syringaldehyde
Syringic acid	1,2,3-Trimethoxybenzene
Vanillic acid	Vanillin

9. Toxicity

Smoking and public health is a controversial issue which has recently become more intense. Tobacco has been smoked for centuries and used for possibly millennia. The idea that smoke might cause cancer was first proposed in 1795, followed by suggestions 100 years later that chemical or physical effects of tobacco dust might cause lung disease. Not until early this century were case control studies reported, mostly on lung cancer.

Numerous studies were conducted on the association with smoking of four diseases: cancer of the lung, myocardial infarction, peripheral vascular disease, and chronic obstructive lung diseases. However, others maintained that the conclusion that smoking is one of the factors that contribute to these diseases is "more fairly described as presumptive than as proved."

In attempting to review and summarize available knowledge, the International Agency for Research on Cancer (IARC) commissioned a group of 28 invited experts in 1985 in Lyon, France, and completed the monograph *Tobacco Smoking*, as a separate volume of IARC series *Evaluation of the Carcinogenic Risk of Chemicals to Humans* [4]. This volume contains in-depth conclusions and evaluations at that time. More recently, tobacco smoke, including mainstream smoke and "environmental tobacco smoke" (ETS), has been reported to be associated with many other health problems. The US Surgeon General provides an annual report on

Sterols, Oxygenated Terpenes, and Other Isoprenoids

Sterols	
Campestrol	Cholestrol
Ergosterol	β-Sitosterol
Stigmasterol	
Monoterpenes	
Borneol	1-Linalool
Diterpenes	
3,8,13-Duvatriene-1,5-diol (α-,β-)	4,8,13-Duvatriene-1,3-diol (α-,β-)
12α-Hydroxy-13-epimanoyl oxide	α_2-Levantanolide
Levantenolide (α-,β-)	α-5,8-Oxido-3,9,13-duvatrien-1-ol
α-5,8-Oxido-3,9(17), 13-duvatrien-1-ol	β-5,8-Oxido-3,9(17),13-duvatrien-1-ol
Triterpenes	
β-Amyrin	
Tetraterpenes	
Cryptoxanthin	Flavoxanthin
Lutein	Neoxanthin
Violaxanthin	Zeaxanthin
Trisesquiterpene	
Solanesol	
Related Isoprenoids	
6,8-Dihydroxy-11-isopropyl-4,8-dimethyl-14-oxo-4,9-pentadecadienoic acid	Hexahydrofarnesylacetone
Solanones	
Vitamin K_1 (2-methyl-3-phytyl)-1,4-naphthoquinone	Solanochromene
Tocopherols	

Table 5. Representative analyses of cigarette tobaccos (leaf web after aging, moisture-free basis)

Component	Flue-cured type 13	Burley type 31	Maryland type 32	Turkish∗
Total volatile bases, as ammonia, %	0.282	0.621	0.366	0.289
Nicotine, %	1.93	2.91	1.27	1.05
Glutamine, as ammonia, %	0.033	0.035	0.041	0.020
Asparagine, as ammonia, %	0.025	0.111	0.016	0.058
α-Amino nitrogen, as ammonia, %	0.065	0.203	0.075	0.117
Protein nitrogen, as ammonia, %	0.91	1.77	1.61	1.19
Nitrate nitrogen, as NO_3, %	Trace	1.70	0.087	Trace
Total nitrogen, as ammonia, %	1.97	3.96	2.80	2.65
pH	5.45	5.80	6.60	4.90
Total volatile acids, as acetic acid, %	0.153	0.103	0.090	0.194
Formic acid, %	0.059	0.027	0.022	0.079
Malic acid, %	2.83	6.75	2.43	3.87
Citric acid, %	0.78	8.22	2.98	1.03
Oxalic acid, %	0.81	3.04	2.79	3.16
Volatile oils, %	0.148	0.141	0.140	0.248
Alcohol-soluble resins, %	9.08	9.27	8.94	11.28
Reducing sugars, as dextrose, %	22.09	0.21	0.21	12.39
Pectin, as calcium pectate, %	6.19	9.91	12.14	6.77
Crude fiber, %	7.88	9.29	21.79	6.63
Ash, %	10.81	24.53	21.98	14.78
Calcium, as CaO, %	2.22	8.01	4.79	4.22
Potassium, as K_2O, %	2.47	5.22	4.40	2.33
Magnesium, as MgO, %	0.36	1.29	1.03	0.69
Chlorine, as Cl, %	0.84	0.71	0.26	0.69
Phosphorus, as P_2O_5, %	0.51	0.57	0.53	0.47
Sulfur, as SO_4, %	1.23	1.98	3.34	1.40
Alkalinity of water-soluble ash∗∗	15.9	36.2	36.9	22.5

∗ Blend of Macedonia, Smyrna, and Samsun
∗∗ Milliliters of 1 N acid per 100 g tobacco

various aspects of smoking and health. The vast amount of findings and reports cannot be fully discussed in this limited space; current knowledge on toxicity of compounds reported in tobacco is summarized in Table 6 [5].

10. Economics

Tobacco is the most rewarding economic crop for farmers. Despite the health issue, tobacco is still an expanding industry worldwide; average demand for tobacco leaf maintains a 2 % annual increase. In highly industrialized countries, under recession, high tax, and antismoking pressure, tobacco consumption is decreasing by ca. 2 % each year. However, in the developing world, there are continuous increases, far more than the reductions of developed regions. In Asia, the greatest increase in tobacco production has occurred in China. From the mid-1980s, China showed an annual 12 % growth for several years, and recently still maintained an annual 3 % growth. China is now the leading tobacco producer, cigarette manufacturer, and consumer in the world; its exports in 1992 surpassed $\$400\times10^6$. Thailand and South Korea, have also expanded their tobacco industries significantly; Indonesia, the Philippines, Vietnam, and India have also made efforts toward further progress.

In Africa, Malawi tobacco production has promoted its national economy. Zimbabwe's tobacco is the nation's major export. Many other African countries, such as Kenya, Tanzania, Uganda, and Zambia, are eager to follow.

Croatia, Hungary, Bulgaria, Poland, and Slovakia are making progress in both production and manufacture, with varying degrees of success.

For Oriental tobacco, Turkey and Greece still dominate. Turkey has increased its production steadily in recent years, but Greece appears to face some decline.

Almost every country has tried to grow tobacco because of its high economic return, but only a few can produce good-quality, usable tobacco. The United States produces the best quality tobacco, but faces a decrease in production and export because of high production costs in comparison with Brazil and Zimbabwe. Argentina has also increased tobacco production considerably in recent years with a high production cost. At its peak in the early 1960s, the United States produced 25 % of the world total; in 1992, this was reduced to only 9 %. Even so, the United States is likely to continue to produce tobacco. No other crop can provide the same financial return to farmers. This holds true in many other countries.

11. Utilization of Tobacco Other than Smoking

11.1. In Basic Research

Tobacco has been used as a most important research tool in biological science; it is frequently referred to as the *E. coli* of the plant kingdom. A few significant examples are:

1) Photoperiodism
2) Genetics and breeding, especially in molecular genetics (interspecific hybridization, parasexual hybridization or cell fusion, disease resistance, and insect resistance)
3) Growth regulation
4) Pollutants and pollution effects
5) Viruses (tobacco mosaic virus was the foundation for virus research)
6) Plant nutrition (effects of micro- and macroelements on plant growth and development; specific symptoms for diagnosis of nutrition-related disorders)
7) Photosynthesis and photorespiration
8) Organic metabolism
9) Postharvest physiology
10) Total usage of plant products, including biomass

11.2. Use of Nicotine

Nicotine, long established as an insecticide to destroy scabies and lice on cattle, has also been used to destroy scabies, lice, and ticks on sheep. In recent years, nicotine has been used in the form of tobacco dust for pest control in poultry houses. In field plant production, "black leaf 40" – a 40 % solution of nicotine sulfate – can be used to control aphids.

Table 6. Tumorigenic agents in tobacco and tobacco smoke

Compounds	In processed tobacco, per gram	In mainstream smoke, per cigarette	IARC evaluation of evidence of carcinogenicity [a]	
			In laboratory animals	In humans
PAH [c]				
Benz[*a*]anthracene		20 – 70 ng	sufficient	
Benzo[*b*]fluoranthene		4 – 22 ng	sufficient	
Benzo[*j*]fluoranthene		6 – 21 ng	sufficient	
Benzo[*k*]fluoranthene		6 – 12 ng	sufficient	
Benzo[*a*]pyrene	0.1 – 90 ng	20 – 40 ng	sufficient	probable
Chrysene		40 – 60 ng	sufficient	
Dibenz[*a,h*]anthracene		4 ng	sufficient	
Dibenzo[*a,i*]pyrene		1.7 – 3.2 ng	sufficient	
Dibenzo[*a,l*]pyrene		present	sufficient	
Indeno[1,2,3-*c,d*]pyrene		4 – 20 ng	sufficient	
5-Methylchrysene		0.6 ng	sufficient	
Aza-arenes				
Quinoline	1 – 2 µg			
Dibenz[*a,h*]acridine		0.1 ng	sufficient	
Dibenz[*a,j*]acridine		3 – 10 ng	sufficient	
7*H*-Dibenzo[*c,g*]-carbazole		0.7 ng	sufficient	
N-Nitrosamines				
N-Nitrosodimethylamine	ND [d] – 215 ng	0.1 – 180 ng	sufficient	
N-Nitrosoethylmethylamine		3 – 13 ng	sufficient	
N-Nitrosodiethylamine		ND [d] – 25 ng	sufficient	
N-Nitrosonornicotine	0.3 – 89 µg	0.12 – 3.7 µg	sufficient	
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone	0.2 – 7 µg	0.08 – 0.77 µg	sufficient	
N′-Nitrosoanabasine	0.01 – 1.9 µg	0.14 – 4.6 µg	limited	
N-Nitrosomorpholine	ND [d] – 690 ng		sufficient	
Aromatic amines				
2-Toluidine		30 – 200 ng	sufficient	inadequate
2-Naphthylamine		1 – 22 ng	sufficient	sufficient
4-Aminobiphenyl		2 – 5 ng	sufficient	sufficient
Aldehydes [b]				
Formaldehyde	1.6 – 7.4 µg	70 – 100 µg [b]	sufficient	
Acetaldehyde	1.4 – 7.4 µg	18 – 1400 µg [b]	sufficient	
Crotonaldehyde	0.2 – 2.4 µg	10 – 20 µg		
Miscellanous organic compounds				
Benzene		12 – 48 µg	sufficient	sufficient
Acrylonitrile		3.2 – 15 µg	sufficient	limited
1,1-Dimethylhydrazine	60 – 147 µg		sufficient	
2-Nitropropane		0.73 – 1.21 µg	sufficient	
Ethylcarbamate	310 – 375 ng	20 – 38 ng	sufficient	
Vinyl chloride		1 – 16 ng	sufficient	sufficient
Inorganic compounds				
Hydrazine	14 – 51 ng	24 – 43 ng	sufficient	inadequate
Arsenic	500 – 900 ng	40 – 120 ng	inadequate	sufficient
Nickel	2000 – 6000 ng	0 – 600 ng	sufficient	limited
Chromium	1000 – 2000 ng	4 – 70 ng	sufficient	sufficient
Cadmium	1300 – 1600 ng	41 – 62 ng	sufficient	limited
Lead	8 – 10 µg	35 – 85 ng	sufficient	inadequate
Polonium (^{210}Po)	0.2 – 1.2 pCi	0.03 – 1.0 pCi	sufficient	sufficient

[a] No designation indicates that an evaluation by IARC has not been carried out.
[b] The 4th report of the Independent Scientific Committee on Smoking and Health (1988) published values for the 14 leading British cigarettes in 1986 (51.4 % of the market) of 20 – 105 µg per cigarette (mean 59 µg) for formaldehyde and 550 – 1150 µg per cigarette (mean 910 µg) for acetylaldehyde.
[c] PAH = polynuclear aromatic hydrocarbons.
[d] ND = not detected.

In animal studies, nicotine administration decreases body weight or retards the normal body weight increase. Cessation of nicotine is accompanied by marked increases in body weight. Similar to the results of human studies, animals receiving large doses of nicotine decrease their consumption of very sweet foods.

11.3. As a Source of Chemicals

Tobacco contains appreciable quantities of citric and malic acids and unique essential oils. Pectin and the enzyme, pectase, are known to occur, at least in the stems. A considerable quantity of resins, possibly suitable for coatings and plastics, is also present.

Many studies have been conducted to provide new uses for low-grade tobaccos; many usable inorganic and organic constituents, e.g., chlorophyll, furfural, sugars, nitrogen, potassium, nicotine, crude fiber, and organic acids, have been found. Development of a market for tobacco seed oil and its residue might make it profitable to produce seed along with tobacco.

11.4. Food and Feed

A class of compounds which can be removed in the pure state and with relative ease is soluble proteins. The biological efficiency of fraction 1 protein as a mammalian nutrient is very high, and its amino acid breakdown is very similar to human or cow's milk. It is far superior in this respect to soybean protein.

Fraction 2 protein has given indications that its nutrient properties as a human food are only a little below those of fraction 1 protein, and much superior to those of soybean. The insoluble proteins and valuable pigments (e.g., xanthophyll) can also be removed in a subsequent solvent extraction process, but this has yet to receive full attention.

11.5. Dust, Midribs, Stems, and Stalks

Tobacco dust can be used as a dry organic fertilizer. Tobacco stems, stalks, and fines have been used for reconstituted sheets. Many other uses are profitable for stems and stalks: animal feed; bedding in animal sheds for cattle and horses; dropping pits for poultry; cigarette paper; and pulp and paper.

11.6. Biomass

Cellulose can be used as cattle feed, or converted to glucose by acid hydrolysis or by an enzyme. The enzyme method gives the highest yields, and is becoming less expensive. Thus, prospects are good that large quantities of glucose derived from wood may become available. This product could be used as food for humans, as a carbon source for microbial formation of protein, or as a feedstock for fermentation yielding liquid fuels, chemicals, and pharmaceuticals.

12. Outlook

12.1. Tobacco in Plant Science

On the frontiers of science, tobacco maintains and expands its role as a research tool for many reasons:

1) There are abundant resources – more than 65 *Nicotiana* species and thousands of germplasm collections which are available and well documented.
2) There is abundant scientific information – tobacco-production-related studies or basic research with tobacco have generated valuable information, providing the foundation for new research.
3) Tobacco is easy to grow. It has a short growing period, has many characteristic markers, and produces a large quantity of seeds per plant.
4) Tobacco tissue and cell culture techniques are well developed for easy and mass production.
5) Tobacco covers broad areas of science – culture, physiology, biochemistry, chemistry, physics, and medical science.
6) Tobacco is a major economic crop in many parts of the world.

12.2. Tobacco in Biotechnology

Biotechnology or genetic engineering is often used to describe the technology of structuring, or purposely restructuring the genetic composition of a cell or a plant. It may be possible to induce or introduce new variation through biotechnology. One approach is protoplast tissue culture. The regeneration of plants from a single cell has been shown to alter the genetic composition of the plant; this can be used to increase resistance to diseases or insects. Another means of introducing new germplasm is by protoplast fusion.

The employment of soil microorganisms in biotechnology has been described in detail by many authors [1, pp. 654]. The tumor-inducing (Ti) plasmid of the soil microorganism *Agrobacterium tumefaciens* is the agent of crown gall disease in dicotyledonous plants.

The *Agrobacterium tumefaciens*-induced tumor cells also produce a series of unusual metabolites, called opines, which are not found in normal plant cells. Both the genes responsible for hormone-independent growth and for opine biosynthesis are encoded by t-DNA. As a result of infection, t-DNA is integrated into the nuclear genome of the host plant, and its expression is responsible for the peculiar traits of the tumor cells. Mutant Ti plasmids have been developed, allowing complete plants to regenerate from transformed cells. These findings have important implications for genetic engineering, as well as for the use of somaclonal variation techniques in plant breeding. Three major possibilities specific to tobacco include:

1) The use of tobacco cell cultures for the production of valuable drugs
2) The modification of tobacco cultivars through genetic engineering to obtain plants with improved resistance various viral, bacterial, fungal, or nematode pathogens, or with modified contents of aromatics and nitrate, or with improved morphological properties or modified senescence properties
3) The breeding of tobacco for food production

12.3. Stress Research

Environmental and biological stresses cause inhibition of growth and reduction of crop yield and quality. Biochemically, plants, like animals, respond to stress by the synthesis of different proteins that are directly related to the type of stress experienced.

Salinity and Tolerance. Cultivated crops are very sensitive to the presence of high concentrations of salt in the soil. Plants must be able to alleviate the detrimental effects of high external concentrations of salts on enzymatic and other metabolic processes necessary to sustain life. The plants must also overcome the dehydration and toxicity induced by salinity.

In general, synthesis of preexisting proteins is inhibited by high salt concentrations, while new proteins are synthesized in response to stress.

Several new proteins, synthesized in response to altered environment, have been reported as "stress" or "shock" proteins in plants.

Sodium Chloride Stress and Sulfate Stress. Sodium chloride is considered the most common source of salinity in plants. The tolerance limits between crops vary, and tobacco plants are among the least tolerant crop tested; tobacco shoot cultures are more tolerant.

Sulfates are much less active in inducing stress than chlorides, and are taken up at a slower rate; in some instances, the uptake rates are described as one-tenth those of chloride.

Sulfate-tolerant shoots also exhibit the highest percentage rooting, particularly in the presence of sulfates, and produce the most abundant roots.

Water Stress. Leaf initiation is greatly inhibited by water deficit. Rate of leaf initiation in well-watered plants is constant with time, but rapidly reduced by a small water deficit. Cell expansion shows a similar response to water shortage, but cell division is much less sensitive.

Heat Stress. Transferring living cells to an elevated temperature produces stress. At injurious temperatures, tobacco tissues respond by increased electrical resistance, followed by a rapid decrease, and irreversible injury to the tissues, with a strong inhibition of oxygen uptake. The most readily observable response to heat stress in many organisms is the induction of heat-shock proteins.

Investigations of formation of heat-shock proteins at different levels of gene expression, as well as studies to distinguish between heat-shock genes and genes for heat tolerance should lead to breeding programs that can select for heat-shock genes and their regulation.

13. References

1. T. C. Tso: *Production, Physiology and Biochemistry of Tobacco Plant,* IDEALS, Beltsville, MD, 1991, p. 785.
2. T. C. Tso: *Physiology and Biochemistry of Tobacco Plants,* Dowden, Hutchinson and Ross, Inc., Stroudsburg, PA 1972, p. 393.
3. B. C. Akehurst : *Tobacco,* 2nd ed., Longman, London 1981, p. 764.
4. World Health Organization, International Agency for Research on Cancer: "Tobacco Smoking", *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* vol. 38, Lyon 1986.
5. D. Hoffmann: *Tumorigenic Agents in Tobacco Smoke, American Health Foundation,* New York, NY, 1994, personal communication.

Toluene

Jörg Fabri, VEBA AG, Düsseldorf, Federal Republic of Germany

Ulrich Graeser, VEBA Öl AG, Gelsenkirchen, Federal Republic of Germany

Thomas A. Simo, Lurgi Öl-Gas-Chemie GmbH, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Properties 2
2.1. Physical Properties 2
2.2. Chemical Properties 2
2.3. Properties as a Motor Fuel Component 3
3. Separation, Extraction, and Processing 4
3.1. Separation 4
3.2. Processing 4
4. Integration into Refineries and Petrochemical Complexes 5
5. Economic Aspects 6
6. Quality Requirements 6
7. Storage, Transport, and Safety 6
8. Environmental Aspects and Toxicology 7
8.1. Environmental Aspects 7
8.2. Toxicology 9
9. References 10

1. Introduction

In the early industrial era, toluene was produced from aromatic fractions, mostly obtained by the coking of hard coal. The importance of toluene as a raw material increased dramatically during World War I, when it was used extensively for the production of the explosive trinitrotoluene (TNT). Since the quantity of toluene produced from coking plants was not sufficient, it was also obtained by direct distillation from aromatic crude oils from Java and Borneo, and by thermal cracking of other naphtha fractions with a narrow boiling range [1].

Sulfuric acid was initially used for refining crude toluene, leading to serious disposal problems. This process began to be replaced by hydrogen refining from ca. 1950.

During World War II, the demand for toluene rose again as a result of increased demand for high-octane aircraft fuels. Production of sufficient quantities of toluene was achieved by development of the catalytic reforming process, in which naphtha fractions from crude petroleum can be converted to reformate with high aromatic content. To isolate pure toluene from reformate, extraction and extractive distillation processes were developed. Pyrolysis gasoline, another source of toluene, became available as a side product of olefin production by steam cracking from about the mid-1960s.

Since then the development of the chemical uses of toluene has lagged behind the increase in its availability. Its relatively low status in chemicals production is mainly because of its lower economic attractiveness compared to the alternative raw material benzene, e.g., for producing phenol or caprolactam. Consequently, considerable quantities of toluene are converted to benzene by hydrodealkylation. Comparatively small quantities of toluene are used for the production of explosives, pharmaceuticals, dyes, and plastics.

A significant proportion of toluene is not isolated in pure form, but is added to motor fuels as a mixture with other aromatics (reformate or pyrolysis gasoline). Because of its high resistance to knocking, the use of toluene as a motor fuel component is frequently more economic than extraction and subsequent use in chemical processes.

History [2]. Toluene (methylbenzene) was discovered by P. S. Pelletier and P. Walter in 1837, during the preparation of coal gas from pine resin. In 1841, H. Sainte-Claire-Deville described the isolation of toluene by the dry distillation of tolu balsam, a solid resinous material with a vanilla-like odor, extracted from the trunk of *Myroxylon balsamum* Harms var. *genium*, a tall tree found in the high plateaus and mountain ranges of South America. In the extraction

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_147

process toluene is formed from benzyl cinnamate (3-phenyl-2-propenoic acid phenylmethyl ester).

C. B. MANSFIELD identified toluene in coal tar in 1849. It had been given the name toluin by J. J. BERZELIUS in 1843; later, J. S. MUSPRATT and A. W. HOFMANN changed the name to toluol. In English, toluene refers to the pure substance, whereas toluol refers to a commercial product, of variable purity.

2. Properties

2.1. Physical Properties

Toluene [*108-88-3*], methylbenzene, C_7H_8, is a colorless, flammable liquid of low viscosity with a benzene-like odor. It is a good solvent for fats, oils, tars, resins, sulfur, phosphorus, iodine, etc. It is completely miscible with most organic solvents, e.g., alcohols, ethers, ketones, phenols, esters, and chlorohydrocarbons. Toluene is only slightly soluble in water, 0.047 g/100 mL at 16 °C. At 15 °C the solubility of water in toluene is also very low: 0.4 g/100 mL [2].

Some physical data are listed in the following [2]:

Molecular mass	92.13
mp	−94.991 °C
bp (100 kPa)	110.625 °C
Critical temperature	320.8 °C
Critical pressure	4.133 MPa
Critical compressibility	0.260
Critical molecular volume	0.32 L/mol

Table 1. Data for binary azeotropes of toluene [3,4]

Component	*bp* at 0.1 MPa, °C		Toluene
	Component	Azeotrope	content of azeotrope, wt %
Water	100	84.1	79.7
Methanol	64.7	63.8	31
Acetic acid	118.5	105.0	66
Ethanol	78.3	76.6	33
Glycol	197.4	110.2	93.5
Ethylenediamine	116.9	103	70
Propanol	97.3	92.5	49
1,2-Propanediol	187.8	110.5	98.5
1,2-Propanediamine	120.9	105	68
Dioxan	101.8	101.4	20
Butanol	117.8	105.7	73
Isobutanol	108.0	100.9	55.5
sec-Butanol	99.5	95.3	45
Pyridine	115.5	110.1	78.8
tert-Amyl alcohol	101.7	100.0	44
Isoamyl alcohol	131.3	110.0	88

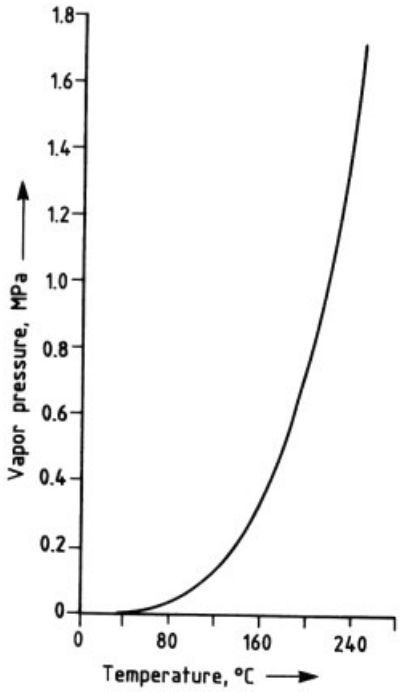

Figure 1. Vapor pressure of toluene as a function of temperature [3,4]

Table 2. Data for ternary azeotropes of toluene [3,5]

Components		*bp*, °C		Content in the azeotrope, wt %		
		Component	Azeotrope	Toluene	A	B
A	Water	100	76.3	48.7	13.1	38.2
B	Isopropanol	82.3				
A	Butanol	117.8	108.7	67.4	11.9	20.7
B	Pyridine	115.5				
A	Pyridine	115.5	110.2	87.3	8.6	4.1
B	Isoamyl alcohol	131.3				

Density at 100 kPa and 25 °C	0.8631 g/cm^3
Surface tension at 100 kPa and 25 °C	26.75 mN/m
Viscosity at 100 kPa and 20 °C	0.5864 mPa · s
Thermal conductivity of liquid at 100 kPa and 0 °C	0.1438 W m^{-1} K^{-1}
Enthalpy of evaporation at 100 kPa and *bp*	32.786 kJ/mol
Lower calorific value	
gas	40.97 kJ/g
liquid	40.52 kJ/g
Flash point	4 °C
Ignition temperature	552 °C
Explosion limits in air	
upper	1.27 vol %
lower	6.75 vol %
Refractive index n_D^{20}	1.49693

Binary and ternary azeotropes of toluene are listed in Tables 1 and 2, respectively. Figure 1 shows the vapor pressure as a function of temperature.

2.2. Chemical Properties

Toluene is similar to benzene in its chemical properties, but the methyl group provides additional reactivity. The aromatic nucleus can be hydrogenated to methylcyclohexane. Oxidizing agents preferably attack the methyl group, giving benzaldehyde and then benzoic acid. The latter can be decarboxylated to phenol (Dow and California Research) or hydrogenated to cyclohexanecarboxylic acid (cf. the Snia-Viscosa ε-caprolactam process [6], → Caprolactam, Chap. 4.3.).

Both the nucleus and the side chain can be chlorinated. Nitration leads to *o*- and *p*-nitrotoluenes, which are separated by freezing, solid–liquid separation, and distillation. Further nitration gives di- and trinitrotoluenes; TNT is used as an explosive (→ Explosives, Chap. 5.2.2.).

Some nitrotoluenes are important organic intermediates, as are the toluidines obtainable from them by reduction (starting materials for dye syntheses and vulcanization accelerators), and the dinitriles derived from dinitrotoluenes, which give toluene diisocyanates (→ Isocyanates, Organic, Chap. 4.3.). The latter are starting materials for polyurethanes.

Alkylation of toluene with propylene gives methylcumene isomers (cymenes). Saponification of the cumene hydroperoxides results in the isomeric cresols and acetone (Sumitomo Chemical [7], → Cresols and Xylenols, Chap. 1.4.3.).

Toluenesulfonic acids and acid chlorides are, like chlorotoluene, starting materials for dyes, pharmaceuticals, and saccharin.

2.3. Properties as a Motor Fuel Component

The usefulness of toluene as a motor fuel component depends on how its properties can be combined with those of the other components. Because of the high resistance to knocking and the comparatively low vapor pressure, toluene is highly valued as a motor fuel component [8,9]. In Table 3 the octane number and the vapor pressure values (Reid vapor pressure) of toluene are compared with the specification requirements for unleaded super gasoline (DIN 51 607/EN 228).

Table 3. Octane numbers and vapor pressures of toluene and motor fuel

	Toluene	Unleaded super (DIN 51 607/EN 228)
Research octane number (RON), pure	110	
Motor octane number (MON), pure	98	
Research octane number (RON), typical blend	124	95
Motor octane number (MON), typical blend	112	85
Reid vapor pressure, kPa	0.8	9 (winter) 7 (summer)

3. Separation, Extraction, and Processing

3.1. Separation

It is possible to synthesize toluene industrially by alkylation of benzene with methanol [10], and by cyclization of *n*-heptane with subsequent aromatization [11,12]. However, for economic reasons toluene is extracted from the raw materials described below. They contain toluene, its homologues, other hydrocarbons, and heteroatomic organic compounds. These raw materials originate from:

Reformates from crude petroleum distillates
Liquid products from the pyrolysis of hydrocarbons (steam cracking)

Liquid products from the gasification or coking (pyrolysis) of coal, lignite, etc.

These product streams with their characteristic compositions and treatments have been described in detail [13–16] (→ Xylenes). An important aspect of the extraction of toluene is the fact that the pyrolysis products (from steam cracking, coking, etc.) must be hydrogenated before pure toluene can be extracted. The unsaturated components are converted to saturated ones and the heteroatoms such as sulfur, nitrogen, and oxygen are removed. In the case of reformates such pretreatment is usually unnecessary.

For the separation of toluene from other components within the same boiling range several methods are available, depending on quality requirements. Fine fractionation is now suitable only for the production of toluene with lower purity, and involves significant losses in fores and tails. Azeotropic distillation uses entrainers, such as methanol [13], to separate toluene from nonaromatics; a nonaromatics – methanol fraction with a lower *bp* than the methanol – toluene azeotrope distills at the column head, while pure toluene is removed from its base. Methanol is recovered from the distillate by washing with water.

For economic reasons, extractive distillation is now used only for the separation of toluene from nonaromatics. Technical grade solvents with higher *bp* than toluene have proved to be suitable extraction agents, e.g., *N*-methylpyrrolidone (Distapex process, Lurgi Öl-Gas-Chemie [14, 17]), and morpholine (Morphylane process, Krupp-Koppers [18]). Extractive distillation essentially involves two distillation columns between which the extraction agent is circulated. The toluene-containing material is charged to the extraction column; the extraction agent is charged to the column head. The extraction agent – toluene mixture leaves the column at the bottom, and is separated into pure toluene and extraction agent in a second recovery column. The product obtained at the distillation head of the extraction column contains the nonaromatic components of the starting material and the extraction agent. This fraction is separated in the recovery column into raffinate (nonaromatics) and extraction agent. The latter is then combined with the main portion of extraction agent from the recovery column, and fed back into the extraction column.

If a toluene fraction is a complex aromatic mixture (e.g., benzene – toluene – xylene = BTX), as most raw materials are, then liquid – liquid extraction is necessary. The solvents which have proved suitable industrially are shown in Table 4.

Table 4. Aromatic extraction processes

Process	Licenser	Solvent
Sulfolan	Shell-UOP	tetrahydrothiophene dioxide
Arosolvan	Lurgi	*N*-methylpyrrolidone
Morphylex	Krupp-Koppers	*N*-formylmorpholine
Formex	Snamprogetti	*N*-formylmorpholine
IFP	IFP	dimethyl sulfoxide
Mofex	Leunawerke	methylformamide

A range of other solvents, ethylene glycol, triethylene glycol, tetraethylene glycol, propylene carbonate, etc., have been proposed, but are no longer economically important.

3.2. Processing

Most of the pure toluene produced is converted to benzene by dealkylation. The catalytic conversion takes place in a hydrogen atmosphere at 550 – 650 °C, 3 – 10 MPa. This process, used by UOP, Air Products (Houdry-Lummus), and BASF, is described in more detail elsewhere (→ Benzene, Chap. 5.3.1.).

Another route is thermal dealkylation, using technologies developed by Atlantic Richfield (HDA process), Gulf (THD process), Mitsubishi Petrochemical (MHC process), and UOP. A somewhat higher temperature is used (600 – 800 °C), with a hydrogen atmosphere. The other features of these process (yield, purity, hydrogen consumption, etc.) are similar to those of the catalytic process (→ Benzene, Chap. 5.3.1.).

Toluene is sometimes used for production of xylenes by transalkylation and disproportionation. In transalkylation, one molecule of benzene and one molecule of xylene are formed from two molecules of toluene. The ratio of *o*- to *m*- to *p*-xylene corresponds to that present at thermal equilibrium. Ethylbenzene is formed in small quantities.

Disproportionation is the typcial reaction when only toluene is used. Transalkylation is

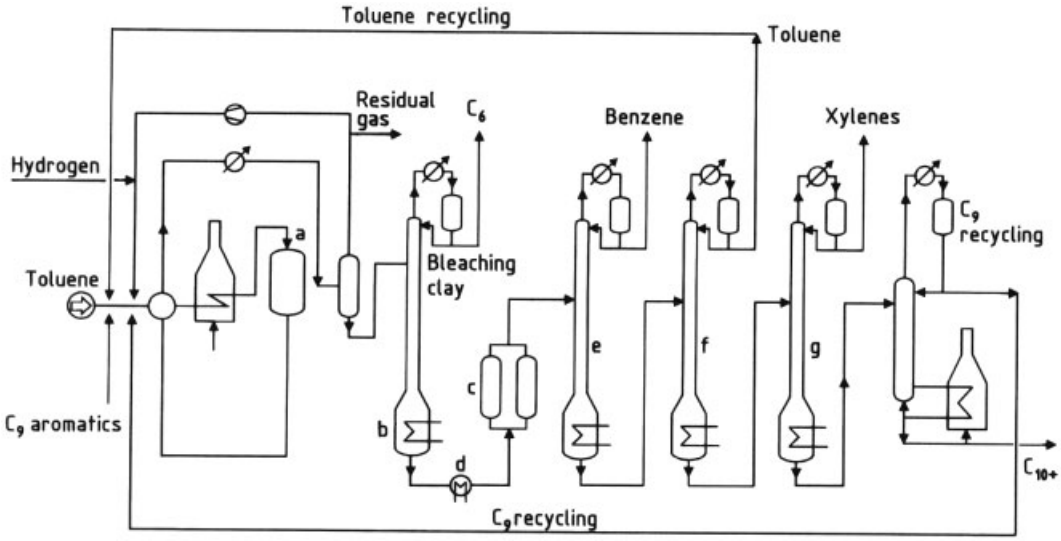

Figure 2. The Tatoray process [2, 19]
a) Reactor; b) Stripper; c) Bleaching clay columns; d) Heater; e) Benzene column; f) Toluene column; g) Xylene column; h) C_9 column

used for additional extraction of xylenes starting from toluene and C_9 aromatics (A_9). Methyl groups are transferred from higher alkylated aromatics to less alkylated ones.

The following technologies have achieved industrial importance:

1) Tatoray (UOP)
2) Xylenes-Plus (Atlantic Richfield)
3) LTD, low-temperature disproportionation (Mobil)

The Tatoray process gives ca. 38 % conversion at each pass. Consequently, after removal of the benzene from the reaction product, a toluene fraction is obtained and recycled, then the xylenes and possibly heavier aromatics are distilled from the Tatoray product (Fig. 2) [2, 19]. Publications from UOP [20] on the MSTDP (Mobil selective toluene disproportionation process) [21, 22] describe the characteristics of the Tatoray process and the integration of the disproportionation – transalkylation into an aromatics complex. The Tatoray process was commissioned in nine industrial plants between 1985 and 1993, and is under construction in a further three units. The MSTDP process was successfully commissioned in three industrial plants between 1989 and 1992.

Development work for the production of xylenes from toluene by alkylation with methanol has continued [23], although the economic environment and the availability of toluene do not justify immediate industrial applications.

4. Integration into Refineries and Petrochemical Complexes

Backward Integration into Refineries. Production of toluene from crude petroleum occurs almost exclusively in major refineries and petrochemical complexes [24]. Because of its great importance for the fuel sector, toluene is either added to the motor fuel pool or is extracted and further processed, the proportions varying according to its current economic value [25]. In addition to the capacity for isolation of toluene, a corresponding hydrodealkylation or disproportionation capacity is often retained, so that there can be a flexible response to changes in demand and price between benzene and toluene. The principle of incorporation of toluene production into the structure of a petrochemical refinery is shown in Figure 3 (see next page).

Removal of toluene from reformates or pyrolysis gasoline reduces the octane number in the motor fuel pool, and can lead to a significant gap in the boiling range of the fuel, as is also the case in benzene and xylene extraction. This is known

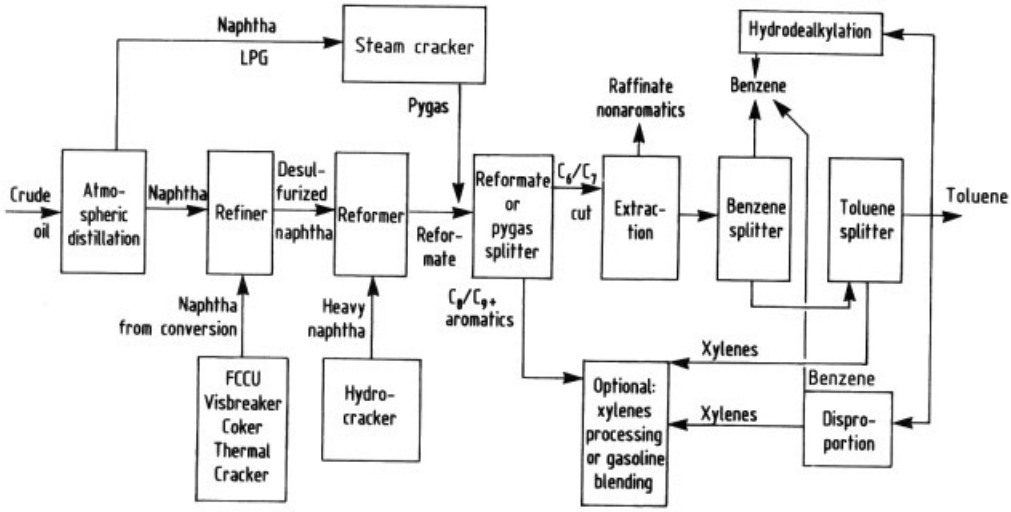

Figure 3. Incorporation of toluene production into a petrochemical refinery

as a gap fuel characteristic, and must be compensated for by adding alternative components [26]. Addition of pure toluene to the motor fuel pool reduces the benzene content by dilution, which is often welcome in terms of environmental and specification aspects. The octane number is raised. However, in view of a possible future limitation on the total aromatic content in motor fuel in the United States and possibly also western Europe [27–30], removal of toluene from reformate and pyrolysis gasoline could help meet the new specification requirements for gasoline [31–33].

Forward Integration into Chemical Processes. Toluene can be used as a solvent without further processing. In the area of petrochemistry, conversion processes (hydroalkylation or disproportionation) giving benzene and xylenes are the most important. The large number of further processing routes in the chemical industry involve a far greater extent of conversion. Toluene diisocyanate (TDI), phenol, caprolactam, nitrotoluene, and phthalates are produced in the largest quantities (Fig. 4).

5. Economic Aspects

As in the case of benzene and xylenes, toluene production and extraction is affected by the pricedifferential compared with naphtha (Fig. 5) and by the relationship, in terms of net product, between separation and further processing as a raw material for chemicals production compared with direct use as a motor fuel component, in the form of reformate or pyrolysis gasoline [25, 34]. The first aspect is essentially fixed by the economics of reforming naphtha, the second by the economic situations in the mineral oil and chemical industries.

In the 1980s, in western European refineries with appropriate extraction capacities, about one-third of the toluene produced was left in pyrolysis gasoline and reformate, in spite of the possibilities for extraction, and was added to motor fuel directly in this form. In the near future, no significant change in the production and consumption of toluene is expected (Table 5).

Table 6 lists toluene producers, production capacities, and raw material sources in western Europe. Other important production capacities for toluene are in the United States, the CIS, Japan, and Canada [36–38].

6. Quality Requirements

Quality requirements for toluene depend on the type of further processing. Apart from individual specification agreements between producers and customers, there are also generally accepted quality specifications for toluene, e.g., ASTM, EN, ISO, DIN, BS [39, 40]. In the ASTM

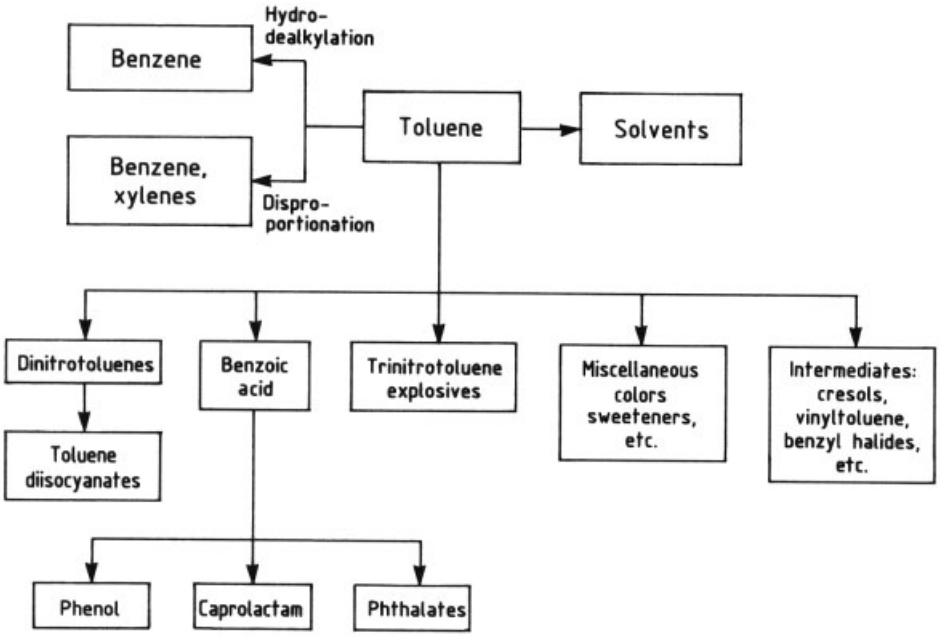

Figure 4. Further processing of toluene

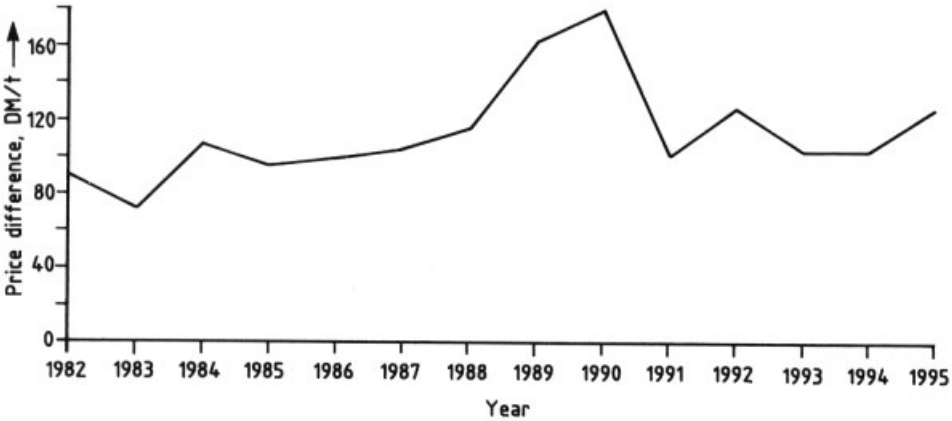

Figure 5. Price difference between toluene and naphtha in western Europe [34]

specifications there is a standard quality (Industrial Grade Toluene, ASTM D 362) and a higher purity quality (Nitration Grade Toluene, ASTM D 841) (Table 7). The German standard is DIN 51 633, comparable to ASTM D 841 (Table 7).

7. Storage, Transport, and Safety

To guarantee safe storage, transfer, and transport of flammable liquids Technical Rules for Flammable Liquids (TRbF) have been drawn up in Germany [41, 42], and are continually updated.

For toluene, the same TRbF numbers apply as for xylenes (→ Xylenes).

Because the flash point lies below 21 °C, toluene is assigned to hazard class A I. On storage and transport, toluene must be labeled as highly flammable and slightly toxic [43]. For transport in Germany, it is assigned to class 3, cipher 3b of the GGVS/GGVE regulations governing the transport of hazardous goods by road

Table 5. Production and consumption of toluene in western Europe (in 10^3 t) [25,35]

	1983	1986	1990	1992	1995	2000 [a]
Production						
From pyrolysis gasoline [b]	1460	1519	1733	1600	1667	1978
From reformate	1114	1653	1675	1378	1325	1325
From coal	23	67	66	64	62	62
Total production	2597	3239	3474	3042	3054	3365
Net imports [a]	100	120	150	150	200	200
Total supply	2697	3359	3624	3192	3254	3565
Consumption						
Dealkylation/disproportionation	993	994	1468	1028	1106	1106
Solvent	423	510	469	465	459	446
TDI	184	160	242	237	250	273
Phenol	115	100	152	116	130	159
Caprolactam	75	95	10	17	17	17
Nitrotoluenes	17		20	20	20	20
Phthalates	8		10	10	10	10
Other chemical products [c]	228	210	228	234	239	246
Total chemical consumption	2043	2069	2599	2127	2231	2277
Use as a fuel component [d]	654	1290	1025	1065	1023	1288

[a] Estimate.
[b] Estimate; only ca. 45 % of the toluene in pyrolysis gasoline is extracted.
[c] For example, benzoic acid, benzaldehyde, benzyl chloride, cresols.
[d] Difference.

Table 6. Producers, capacities, and raw material sources for toluene in western Europe in 1994 (in 10^3 t) [35]

Country	Producers	From reformate	From pyrolysis gasoline	From coal	Total
Austria	Voest	0	0	4	4
Belgium	Finaneste; Sopar	0	65	8	73
France	Atochem; Shell; Total	30	40	0	65
Germany	SOW; EC; Shell; DEA; PCK; Ruhr Öl; Wintershall	475	650	60	1185
U.K.	BP; ICI; Conoco; Shell	465	70	18	555
Italy	Enichem; Praoil; Carboquimice	290	205	0	495
Nether- lands	Exxon	255	0	0	255
Portugal	Petrogal	140	0	0	140
Spain	Cepsa	280	0	0	280
Total		1805	845	92	2727

and rail [44]. The explosion limits of toluene are 1.27 and 6.75 %.

8. Environmental Aspects and Toxicology

8.1. Environmental Aspects

Leakage of toluene into soil or bodies of water constitutes a serious pollution problem. In Germany, toluene is assigned to WGK 2 [42]. The solubility of toluene in water is very low (0.1 g/L at 20 °C). With a vapor pressure of 2.9 kPa (at 20 °C) toluene evaporates relatively rapidly in air; besides the direct toxicological effects (see Section 8.2) secondary reactions with other air pollutants are also important. The reaction with nitrogen oxides in the presence of sunlight leads to the formation of ozone and smog, and can cause severe ecotoxicological pollution [27,45]. Even though toluene undergoes these types of reaction comparatively slowly, extensive measures are increasingly being taken to reduce vapor emissions, both in solvents (where toluene is being replaced to some extent as a result of other developments, e.g., solvent-free paints), and in the fuel sector (in which the emission of toluene into the atmosphere is being significantly lowered by the installation of corresponding vapor retention systems – activated charcoal canisters in passenger cars or gas displacement devices at gasoline stations, storage depots, and refineries) [46,47]. When toluene is used a a fuel compo-

Table 7. Quality requirements for toluene

	Industrial grade	Nitration grade	Method of determination
According to ASTM	**D 362**	**D 841**	
Density at 20 °C, g/mL	0.860 – 0.874	0.8690 – 0.8730	D 891
Color (Hazen)	≤20	≤20	D 1209
Boiling range, °C	± 2	± 1	D 850, D 1078
Acidity	≤0.005 %		D 847
Color after acid treatment	≤no. 4	≤no. 2	D 848
Sulfur content	no H_2S/SO_2	no H_2S/SO_2	D 853
According to DIN 51 633			
Color (Hazen)	20		DIN 53 409
Density at 15 °C, g/mL	0.869 – 0.872		DIN 51 757
Refractive index at 20 °C	≥1.4963		DIN 51 423, part 2
bp, °C	110.6		
Reaction with sulfuric acid	≤0.2		DIN 51 762
Bromine consumption	≤0.1 g/100 mL		DIN 51 774, part 3
Total sulfur	≤0.0002 wt %		DIN-EN 41
Active sulfur	below detection limit		DIN 51 764
Hydrogen sulfide	none		DIN 51 766
Doctor test	negative		DIN 51 765

nent, it affects the composition of exhaust gases [27,48].

8.2. Toxicology

From 1 September 1993, the MAK value for toluene was reduced to 50 ppm (190 mg/m^3) [49, 50]. This corresponds to a halving of the 1992 value. With regard to the effect on pregnancy, a parallel reassignment was made to group C (previously group B). Extensive medical data on the effects of toluene were used for the toxicological evaluation. Results of animal experiments supplement the data.

General Activity Profile. Inhalation of toluene vapor affects the central nervous system, giving rise to symptoms such as headaches, dizziness, or coordination difficulties [51–55]. Loss of consciousness can occur at higher concentration [56]. On acute skin contact with liquid toluene, removal of oils and drying of the skin can occur, with development of dermatitis as a possible consequence. The direct irritant effect of toluene is relatively low [56,57].

Acute and Subacute Toxicity. In animal experiments the following toxicity values were found [58]:

Rat, oral, LD_{50} = 5000 mg/kg
Rabbit, percutaneous, LD_{50} = 12 124 mg/kg
Mouse, 8 h inhalation, LC_{50} = 5320 mg/m^3

On exposure for several hours to <1000 mL/m^3 toluene in inhaled air, no significant effects were established in rats [59]. In the concentration range ca. 2000 mL/m^3, in rats, pulse rate was increased many times and a state of exhaustion was recorded [55]; at higher concentrations, a narcotic effect predominated [60–62].

In humans, concentrations up to ca. 800 mL/m^3 affect the central nervous system, causing confusion, headache, nausea, or coordination difficulties as the most obvious effects [51–55]. Weakness and tiredness are also observed. The effect of toluene on the central nervous system can lead to hallucination and addiction (solvent sniffing). In three cases of death following frequent deliberate inhalation of large quantities of toluene-containing solvents, the lethal dose for humans was estimated as 2000 mL/m^3 within 30 min [63]. At high concentrations, disruption of kidney function is observed [64].

Chronic Toxicity. On prolonged exposure, toluene accumulates in the brain [65]. As with acute intoxication, this leads to damage to the central nervous system. However, in animal experiments this is clearly visible only at concentrations of >500 mL/m^3 [66]. Only at much higher doses were damage to the liver [67] and cardiotoxic effects [68] established in rats and mice. However, in animal experiments, after initial suspicions, there was no confirmation of a

leukemogenic effect similar to that found with benzene.

Many toxic effects which have been recorded for humans with high exposure to solvents cannot be attributed to toluene according to current knowledge, but mainly to small quantities of benzene present as an impurity [69,70]. The effects of acute intoxication on the central nervous system with the symptoms already mentioned are, however, characteristic. Damage to internal organs through chronic inhalation of comparatively low toluene concentrations (up to ca. 200 – 400 mg/m^3) is not unambiguously detectable [71]. However, the chronic inhalation of high doses of toluene (solvent abuse) leads to damage to the cerebellum, disruption of the metabolism, and muscle weakness [72,73].

Pharmacokinetics and Metabolism. In humans, the main route for taking in toluene is by inhalation of vapor. Toluene retention in the lungs on prolonged exposure is ca. 50 %, and can be raised by bodily exertion [74,75]. Half-lives between ca. 2 min and ca. 3.5 s have been reported [76,77]. According to more recent discoveries, the percutaneous resorption rate on skin contact with liquid toluene is 0.17 mg cm^{-2} h^{-1}, considerably lower than assumed previously [78].

The metabolic conversion of toluene mainly occurs in the liver. Degradation takes place over many steps and involves the formation of benzaldehyde and benzoic acid, and their conjugation with glycine to give hippuric acid, which is excreted in the urine. Phenolic metabolites (*o*- and *p*-cresol) have also been detected in the urine in small quantities (up to ca. 0.2 %) [79,80].

Mutagenicity, Teratogenicity, and Carcinogenicity. In the Ames test using *Salmonella typhimurium* strains, no mutagenic effects could be detected with toluene, with or without metabolic activation [81,82].

On testing the effect of toluene on mice, no increase in chromosomal aberrations was established [83]. In a more recent study, in humans exposed to toluene vapor in the concentration range 200 – 300 mL/m^3 for ca. 16 a, significant increase in structural changes was observed (chromatid cleavage and exchange), which could be detected up to two years after exposure had terminated [84]. In a large number of animal experiments, there was no certain evidence for teratogenic effects [85]. No indication of a carcinogenic or cocarcinogenic potential has been found for toluene [86].

9. References

1. H.-G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, pp. 102 – 103.
2. *Ullmann,* 4th ed., **23,** 301.
3. L. H. Horsley, *Anal. Chem. Ser.* **21** (1949) no. 7, 831 – 873.
4. B. Riediger: *Verarbeitung des Erdöls,* Springer Verlag, Berlin 1971, p. 89.
5. Dechema Chemistry Data Series, Vapor-Liquid Data Collection, Frankfurt 1993. C. L. Yaws, *Chem. Eng. (N.Y.),* Sept. 1975, 73 – 81. *Beilstein,* **Suppl. 3, (V) part 2,** 663 – 775. F. D. Rossini et al.: *Selected Values of Physical and Thermodynamic Properites of Hydrocarbons and Related Compounds,* Carnegie Press, Pittsburgh 1953.
6. M. B. Sherwin, *Chem. Eng. Prog. (Nov.)* **26** (1979) 31. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 3rd ed., Verlag Chemie, Weinheim – New York 1990, pp. 273 – 274. F. Asinger: *Die Petrolchemische Industrie,* Akademie Verlag, Berlin 1971, pp. 505 – 640, 1439 – 1441.
7. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 3rd ed., Verlag Chemie, Weinheim – New York 1990, p. 374, 377.
8. K. Owen, T. Coley: *Automotive Fuels Handbook,* Society of Automotive Engineers Inc. (SAE), Warrendale, PA 1990, pp. 144 – 145.
9. W. E. A. Dabelstein, A. A. Reglitzky, K. Reders, N. Lucht: *100 Jahre Kraftstoffe für den Straßenverkehr,* Shell Technischer Dienst, Hamburg 1989, pp. 7, 8, 16 – 25.
10. *Winnacker-Küchler,* 3rd ed., **vol. 3,** p. 817.
11. *Ullmann,* 3rd ed., **17,** 446.
12. *Beilstein,* **4 (5),** 652.
13. K. H. Eisenlohr et al., *Erdoel Kohle Erdgas Petrochem.* **16** (1963) no. 6, 523.
14. K.-H. Eisenlohr, *Erdoel Kohle Erdgas Petrochem.* **20** (1967) 82 – 89.
15. J. J. Wise, A. J. Silvestri, *Oil Gas J.* **74** (1976) 140. C. D. Chang, A. J. Silvestri, *J. Catal.* **47** (1977) 24a.
16. A. Chauvel, G. Lefebre: *Petrochemical Processes,* **vol. 1,** Editions Technip IFP, Paris 1989, pp. 154 – 159, 165 – 178.

17. Lurgi Öl-Gas-Chemie GmbH, company brochure, Lurgi Distapex Process – Recovery of Pure Aromatics, no. 1640e/9.91/10.
18. K. Lackner, *Erdoel Kohle Erdgas Petrochem.* **34** (1981) no. 1, 26 – 30.
19. Toray Industries, Inc.: Transalkylation of Aromatics, Tokyo, Japan, March 1982. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie,* Verlag Chemie, Weinheim – New York 1978, pp. 313 – 314. *Hydrocarbon Process* **60** (1981) no. 11, 139.
20. P. J. Kuchar, D. Y. Lin, V. Zukauskas, D. Brkie: "Isomar and Tatoray Production Flexibility," *Proc. 1988 UOP Technology Conference,* Des Plaines, Ill. 1989. E. C. Haun, M. W. Golem, S. Sapuntzakies, P. P. Piotrowski: "The Modern Aromatics Complex," *Proc. 1988 UOP Technology Conference,* Des Plaines, Ill. 1989.
21. F. Gerra, L. L. Beckenridge, W. M. Guy, R. A. Sailor, *Oil Gas J.* **90** (1992) 60 ff.
22. *Hydrocarbon Process.* **70** (1991) no. 3, 141.
23. M. W. Anderson, J. Klimowski, *Nature (London)* **339** (1989) 200 – 203.
24. W. Keim, A. Behr, G. Schmitt: *Grundlagen der Industriellen Chemie,* Verlag Salle-Sauerländer, Frankfurt/Main 1986, p. 165.
25. Chem Systems Ltd.: *Petroleum and Petrochemicals Economics,* Ann. Rep. 1 and 2, London, Aug. 1991.
26. K. Owen, T. Coley: *Automotive Fuels Handbook,* Society of Automotive Engineers Inc., Warrendale, PA 1990, pp. 564 – 565.
27. *Octane Week* 1993, July 19, 6. *Octane Week* 1993, May 24, 1, 4, 11.
28. A. M. Hochhauser et al., *SAE-Paper* **912322** (1991) 6 – 9.
29. W. J. Koehl: *SAE-Paper* **912321** (1991) 14 – 16, 28 – 29.
30. J. J. Wise, *Fuel Reformulation* 1992, May/June, 64 – 69.
31. D. B. Anthony, R. Ragsdale, *Fuel Reformulation* 1992, July/August, 13 – 18.
32. UOP: *The Clean Air Act And The Refining Industry,* Des Plaines, Ill. 1991, p. 17. UOP: *Process Solutions for Reformulated Gasoline,* Des Plaines, Ill., 1991.
33. W. P. Barry: *Gasoline Regulations in USA and Influence To Aromatics Market,* Paper presented to JAIA Annual General Meeting, Tokuyama, Japan, 14 Nov. 1990.
34. Chem Systems: *Benzene in the 1990's,* London 1991.
35. Parpinelli Technon: *West European Petrochemical Industry,* 1993 Aromatics Report and Tables, Mailand 1993, p. 202.
36. H.-G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, p. 134.
37. A. Chauvel, G. Lefebre: *Petrochemical Processes,* **vol. 1,** Editions Technip IFP, Paris 1989, p. 299.
38. F. Asinger, *Erdöl Erdgas* **106** (1990) no. 6, 263 ff.
39. H.-G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, p. 133.
40. A. Chauvel, G. Lefebre: *Petrochemical Processes,* **vol. 1,** Editions Technip IFP, Paris 1989, p. 298.
41. Degener, Krause, VbF/TRbF, vol. 5 (Vorschriften), 16. Lfg., June 1989, pp. 31 ff.
42. BArbBL, no. 9 (1990) pp. 65 ff.
43. Kühn, Birett, *Merkblätter gefährliche Arbeitsstoffe,* 58. Erg.-Lfg. 3/92-T 013-1 (1992).
44. Berufsgenossenschaft Chemie (ed.): *Unfallmerkblatt für den Straßentransport,* MED-Verlagsgesellschaft, Landsberg 1987, Ausgabe 5/87.
45. J. Fabri, A. Reglitzky, M. Voisey, *Erdöl Erdgas* **107** (1991) no. 1, 28.
46. K. Owen, T. Coley: *Automotive Fuels Handbook,* Society of Automotive Engineers Inc., Warrendale, PA 1990, pp. 158 – 160, 490.
47. DGMK-Tagungsbericht Nr. 8902, Hamburg 1989, pp. 1 – 71.
48. TÜV Rheinland: *Aromaten im Abgas von Ottomotoren,* Verlag TÜV-Rheinland, Köln 1988.
49. *Gefahrstoffe 1993,* Universum Verlagsanstalt, Wiesbaden 1993.
50. Berufsgenossenschaft der chemischen Industrie, Anlage 4 zu den Unfallverhütungsvorschriften, Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg, Sep. 1993, pp. 3 ff.
51. E. Browning: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, Amsterdam – London – New York 1965, p. 66.
52. H. W. Gerarde in G. D. Clayton, F. E. Clayton (eds.): *Industrial Hygiene and Toxikology,* 2nd ed., vol. 2, Wiley-Interscience, New York – London – Sidney 1963, pp. 1222, 1226.
53. G. G. Fodor: *Schädliche Dämpfe,* VDI-Verlag, Düsseldorf 1972, p. 97.
54. B. J. Dean, *Mutat. Res.* **47** (1978) 75.
55. V. A. Benignus, *Neurobehav. Toxicol. Teratol.* **3** (1981) 407.

56. E. O. Longley et al., *Arch. Environ. Health* **14** (1967) 481.
57. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons, Toluene,* Elsevier, Amsterdam –London – New York – Princeton 1960, p. 140.
58. MERCK Sicherheitsdatenbank MS-Safe, state as of Jan. 1, 1991.
59. M. Ikeda et al., *Toxicol. Lett.* **9** (1981) 255.
60. C. P. Carpenter et al., *Toxicol. Appl. Pharmacol.* **36** (1976) 473.
61. N. W. Lazarew, *Naunyn Schmiedebergs Arch. Exp. Path. Pharmakol.* **141** (1929) 223.
62. W. Estler, *Arch. Hyg. (Berlin)* **114** (1935) 261.
63. K. Nomiyama, H. Nomiyama, *Int. Arch. Occup. Environ. Health* **41** (1978) 55.
64. E. Reisin, A. Teicher, R. Jaffe, H. E. Eliahou, *Br. J. Ind. Med.* **32** (1975) 163.
65. A. Fabre, *Bull. Soc. Chim. Biol.* **28** (1946) 764.
66. K. Bättig, E. Grandjean, *Arch. Environ. Health* **9** (1964) 745.
67. J. V. Bruckner, R. G. Peterson, *Toxicol. Appl. Pharmacol.* **61** (1981) 27.
68. V. Morvai, A. Hudak, G. Ungvary, B. Varga, *Acta Med. Acad. Sci. Hung.* **33** (1976) 275.
69. L. Greenberg, M. R. Mayers, H. Heimann, S. Moskowitz, *JAMA J. Am. Med. Assoc.* **118** (1942) 573.
70. T. Matsushita et al., *Ind. Health* **13** (1975) 115.
71. W. Bänfer, *Zentralbl. Arbeitsmed. Arbeitsschutz* **11** (1961) 35.
72. J. W. Boor, H. I. Hurtig, *Ann. Neurol.* **2** (1977) 440.
73. C. M. Fischmann, J. R. Oster, *JAMA J. Am. Med. Assoc.* **241** (1979) 1713.
74. J. Piotrowski, *Med. Pr.* **18** (1967) 129.
75. K. H. Cohr, J. Stockholm, *Scand. J. Work Environ. Health* **5** (1979) 71.
76. K. Nomiyama, H. Nomiyama, *Int. Arch. Arbeitsmed.* **32** (1974) 85.
77. A. Sato et al., *Int. Arch. Arbeitsmed.* **33** (1974) 169.
78. A. Sato, T. Nakayima, *Br. J. Ind. Med.* **35** (1978) 43.
79. O. M. Bakke, R. R. Scheline, *Toxicol. Appl. Pharmacol.* **16** (1970) 691.
80. P. Pfäffli et al., *Scand. J. Work Environ. Health* **5** (1979) 286.
81. R. P. Bos et al., *Mutat. Res.* **88** (1981) 273.
82. Litton Bionetics Inc.: *Mutagenicity Evaluation of Toluene,* API Med. Res. Publ. 26-60020, American Petroleum Institute (API), Washington, D.C., Jan. 1978.
83. P. Gerner-Smidt, U. Friedrich, *Mutat. Res.* **58** (1978) 313.
84. M. Bauchinger et al., *Mutat. Res.* **102** (1982) 439.
85. Litton Bionetics Inc.: *Teratology Study in Rats: Toluene,* API Med. Res. Publ. 26-60019, American Petroleum Institute (API), Washington, D.C., Jan. 1978.
86. J. E. Gibson, J. F. Hardisty, *Fundam. Appl. Toxicol.* **3** (1983) 315.

Toluidines

JOSEPH S. BOWERS, JR., First Chemical Corporation, Pascagoula, MS 39568, United States

1. **Introduction** 1
2. **Physical and Chemical Properties** . . 1
3. **Production** 2
4. **Quality Specifications, Storage, and Transportation** 2
5. **Uses** 3
6. **Derivatives** 3
   - 6.1. ***N*-Alkyltoluidines** 3
   - 6.2. **Chlorotoluidines** 4
   - 6.3. **Toluidinesulfonic Acids** 5
   - 6.4. **Chlorotoluidinesulfonic Acids** 5
7. **Toluidine Toxicology** 7
8. **References** 8

1. Introduction

Early commercial production of toluidines involved the nitration of toluene followed by a dissolving metal reduction, usually with iron and hydrochloric acid, to give a mixture of toluidine isomers. These early mixed toluidines were commercially important as dye intermediates. Today the isomers are separated by fractional distillation at the nitrotoluene stage, and the individual compounds typically reduced by catalytic hydrogenation. Current commercial applications for mixed toluidines, and more importantly for the individual toluidine isomers, include their use as: intermediates for dyes and pigments, pesticides, and herbicides; vulcanizing accelerators for the rubber industry; polymer curing agents; corrosion inhibitors; pharmaceutical intermediates; and intermediates for numerous other fine and speciality chemical areas.

2. Physical and Chemical Properties

Table 1 (see next page) provides a list of selected physical properties of the three isomers of toluidine (aminotoluene), C_7H_9N, M_r 107.2.

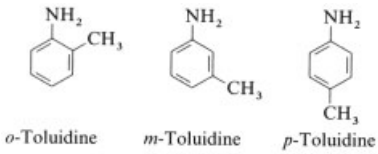

The chemical properties of the toluidines are greatly influenced by delocalization over the aromatic ring of the two unshared electrons on nitrogen, as well as, to a lesser extent, by the electron-donating inductive effect of the methyl group.

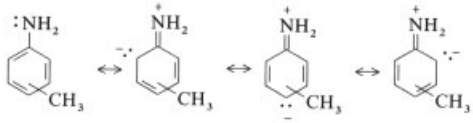

The toluidines are relatively weak bases ($pK_b \approx 9$) when compared to aliphatic amines ($pK_b \approx 2-3$) [1] (→ Amines, Aliphatic). However, toluidines do still display many properties similar to those of aliphatic amines, such as alkylation, acylation, and salt formation with strong acids. The toluidines differ from aliphatic amines in their reaction with nitrous acid, in that they form relatively stable diazonium salts which are in turn very important in the textile-dye industry.

Reactions involving the aromatic ring (e.g., substitution reactions; → Amines, Aromatic) are influenced by the strong ring-activating effect of the amine group, and to a lesser extent that of the methyl group. The amine group is a strong *ortho*–*para* director, and only if the amine group is deactivated by extensive protonation or quaternization does the methyl group exert any significant directing effect.

3. Production

The manufacture of toluidines begins with a mixed-acid (nitric/sulfuric acid) mononitration of toluene, which produces all three isomers (*ortho*, *meta*, and *para*), usually in the ratio 15:1 : 9,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_159

Table 1. Physical properties of the toluidines

	o-Toluidine	*m*-Toluidine	*p*-Toluidine
CAS registry no.	[*95-53-4*]	[*108-44-1*]	[*106-49-0*]
fp, °C	−14.7	−30.4	44.0
bp, °C (101.3 kPa)	200.2	203.3	200.5
ϱ (20 °C)	0.9984	0.989	0.9619
Vapor pressure,	0.20 (50 °C)	0.13 (41 °C)	0.13 (42 °C)
kPa	5.3 (110 °C)	3.3 (100 °C)	8.0 (140 °C)
Heat of combustion, kJ/mol (20 °C)	4038	4039	4013
Heat of vaporization, J/g (101.3 kPa)	398	534.94	524.77
n_D^{20}	1.5728	1.5670	1.5534
Solubility in water, wt % (25 °C)	1.5	1.6	sl. sol.
Flash point, °C (TOC)	85.0	86.0	87.0
Autoignition temperature, °C	482	482	482

respectively. Since the isomeric toluidines themselves cannot be effectively separated by distillation, distillative separation of the isomers is achieved at the nitrotoluene stage.

The literature describes a multitude of methods for reducing nitrotoluenes to toluidines (→ Amines, Aromatic, Chap. 3.1.; → Aniline, Chap. 3.2.2.,). These range from dissolving-metal reductions, such as the Béchamp process [2,3] of years ago, to catalytic hydrogenation. Currently, high-volume manufacture of toluidines utilizes the same type of continuous vapor-phase hydrogenation process used for aniline manufacture [4] (→ Aniline, Chap. 3.2.1.). The catalysts include various supported metals [5,6], Raney nickel [7,8], copper [4], molybdenum, tungsten, vanadium, and noble metals [9–12]. The catalyst is suspended in a column or bed, and a gaseous mixture of the nitrotoluene and excess hydrogen is passed over it at about 250 °C. The gaseous products are condensed, the aqueous and organic layers are separated, excess hydrogen is recycled, and the final crude product is dried and distilled.

Batch liquid-phase catalytic hydrogenation is a common process for the production of lower-volume specialty toluidines or amines too high in molecular mass or too sensitive for vapor-phase reactions. In these cases, liquid nitro compounds can either be reduced neat or else dissolved in appropriate aromatic solvents, aliphatic alcohols, or esters. Catalysts include Raney nickel and supported noble metals. Reduction is conducted between room temperature and 150 °C under hydrogen pressures in the range 350–3500 kPa.

There are several alternative possibilities for reducing nitrotoluenes to toluidines, utilizing such reducing agents as hydrazine [13], alkaline sulfide [14], and sodium hydrosulfite [15]. These methods rarely find commercial application unless the substrate nitro compound has a functional group that might be adversely affected by catalytic hydrogenation.

Additional manufacturing techniques involve the amination of corresponding phenol derivatives [16,17]. This method has been used commercially to produce *m*-toluidine from *m*-cresol.

4. Quality Specifications, Storage, and Transportation

Typical purity of each of the commercially available toluidine isomers is ca. 99.5 % minimum, with a maximum of 0.5 % total for all other isomers. Moisture is usually in the range 0.1 – 0.2 % maximum [18]. Additional specifications important for specific isomer applications are sometimes stated, such as a 1 °C distillation range for 5 % to 95 % of a distilled *o*-toluidine, or a freezing point minimum of 43.0 °C for *p*-toluidine. Assays and moisture content are usually established by GC analysis.

Toluidines are typically stored in steel tanks or drums. All the toluidine isomers exhibit solvent-like properties, so plastics, wax coatings, or other nonmetallic materials are rarely permitted in the context of storage or handling. Teflon TFE or FEP fluorocarbon resins are the preferred materials for gaskets and packaging. When stored in the presence of air all the toluidine isomers tend to discolor, particularly over an extended period of time.

Toluidines are typically shipped in steel drums, tank trucks, rail tank cars, or marine ves-

sels. For transportation by truck, rail, or ship the applicable hazard classification is UN 1708, RQ.

5. Uses

Since *o*-toluidine is the isomer produced in greatest volume, it has found the most commercial applications. By far the single largest use for *o*-toluidine is in the preparation of 6-ethyl-*o*-toluidine, an intermediate in the manufacture of the two very large-volume herbicides Metolachlor [19] and Acetochlor [20]. Another important use area for *o*-toluidine is rubber chemicals. Thus, *o*-toluidine is used in the manufacture of a rubber antioxidant developed by Goodyear called Wingstay as well as di-*o*-tolylguanidine, a non-staining rubber accelerator (→ Amines, Aromatic, Chap. 2.2.2.; → Guanidine and Derivatives, Chap. 2.4.2.). Acetoacet-*o*-toluidine, 3-hydroxy-2-naphthoyl-*o*-toluidine, 2-toluidine-5-sulfonic acid, and *o*-aminoazotoluene are four of the more important dye and pigment intermediates manufactured from *o*-toluidine. In addition, *o*-toluidine is used to manufacture epoxy resin hardeners such as methylene-bis-2-methylcyclohexylamine [21], fungicide intermediates such as 2-amino-4-methylbenzothiazole [22], and *o*-fluorobenzoyl chloride [23]. Certain pharmaceutical intermediates are also prepared starting with *o*-toluidine [24].

Both *m*- and *p*-toluidines still find considerable application in the dye and pigment industries, as well as in the manufacture of triarylmethane dyes. Two compounds derived from *m*- and *p*-toludine, *N,N*-dimethyl-*m*-toluidine and *N,N*-dimethyl-*p*-toluidine, have found use as accelerators for peroxide-cured epoxy and unsaturated polyester resins [25–27]. *m*-Toluidine is used in the manufacture of three important photographic color developers: color developer 2 uses *N,N*-diethyl-*m*-toluidine as an intermediate, and color developers 3 and 4 are based on *N*-hydroxyethyl-*N*-ethyl-*m*-toluidine. Just as with *o*-toluidine, the *m*- and *p*-toluidines find many applications as starting materials and intermediates in the manufacture of various fine and specialty chemicals [28–30].

6. Derivatives

6.1. *N*-Alkyltoluidines

N-Alkyltoluidines can be prepared from unalkylated toluidines with such common alkylating agents as alkyl halides, sulfates, and carbonates. However, large-scale manufacture of *N*-alkyltoluidines usually involves one of two alternative methods: acid-catalyzed alkylation with lower unbranched alcohols or ethers [31], or reductive alkylation with lower aldehydes or ketones over active-metal catalysts (e.g., nickel, platinum, or palladium) and under hydrogen pressure [32]. Toluidines can also be alkylated with reactive three-membered-ring compounds such as ethylene oxide and aziridine.

Important examples include:

N-Methyl-*o*-toluidine [*611-21-2*], M_r 121.18, *bp* 207 – 208 °C, ϱ 0.97
N-Ethyl-*o*-toluidine [*94-68-8*], M_r 135.21, *bp* 218 – 219 °C, ϱ 0.94
N,N-Dimethyl-*o*-toluidine [*609-72-3*], M_r 135.21, *bp* 221 – 222 °C, ϱ 0.94
N-Methyl-*m*-toluidine, [*696-44-6*], M_r 121.18, *bp* 206 – 207 °C, ϱ 0.95
N-Ethyl-*m*-toluidine [*102-27-2*], M_r 135.21, *bp* 221 °C, ϱ 0.96
N,N-Dimethyl-*m*-toluidine [*121-72-2*], M_r 135.21, *bp* 215 – 216 °C, ϱ 0.93
N-Ethyl-*N*-(2-hydroxyethyl)-*m*-toluidine [*91-88-3*], M_r 179.26, *bp* 114 – 115 °C (0.1 kPa), ϱ 1.02
N-Methyl-*p*-toluidine [*623-08-5*], M_r 121.18, *bp* 208 – 209 °C, ϱ 0.94
N-Ethyl-*p*-toluidine [*622-57-1*], M_r 135.21, *bp* 217 – 218 °C, ϱ 0.95
N,N-Dimethyl-*p*-toluidine [*99-97-8*], M_r 135.21, *bp* 211 – 212 °C, ϱ 0.94

6.2. Chlorotoluidines

There exist four monochlorinated isomers each for *o*- and *m*-toluidine, and two for *p*-toluidine. Some are rather difficult to prepare, and thus do not find commercial application. Chlorotoludines are usually prepared by reducing the nitro group of the corresponding chloronitrotoluene. These reductions have traditionally been

accomplished by dissolving-metal reductions, such as the Béchamp process. Today, catalytic hydrogenation is the method of choice, especially since catalysts with improved selectivity almost eliminate dehydrochlorination side reactions [33,34]. The intermediate chloronitrotoluenes are obtained by either chlorination of nitrotoluenes or nitration of chlorotoluenes. Direct chlorination of toluidines can also be accomplished by first protecting the amine function. Chlorotoluidines find commercial applications in such areas as azo dyes, agricultural chemicals, and pharmaceuticals.

3-Chloro-*o*-toluidine, 3-chloro-2-methylaniline [*87-60-5*] (**1**), M_r 141.6, *mp* 2 °C, *bp* 241 –242 °C, ϱ 1.19, is produced by the catalytic hydrogenation of 3-chloro-6-nitrotoluene [35], which is in turn obtained via the acidic chlorination of *o*-nitrotoluene followed by separation of isomers by distillation. An alternate method of preparing the nitro intermediate involves nitration of *o*-chlorotoluene, which also produces a mixture of isomers that can be separated by distillation. The compound is used in the manufacture of azo dyes, and as an intermediate in the manufacture of 2,6-dichlorotoluene.

4-Chloro-*o*-toluidine, 4-chloro-2-methylaniline [*95-69-2*] (**2**), M_r 141.6, *mp* 29 – 30 °C, *bp* 240 –241 °C, is produced by the direct chlorination of acetyl-protected *o*-toluidine. After removal of the protecting group with base, the 4-chloro-*o*-toluidine is separated by distillation from the 6-chloro-*o*-toluidine isomer, described below. This compound is employed in the manufacture of the pesticide Chlordimeform [36], and for making acid azo dyes. Use of the pesticide is declining because of concern over its carcinogenicity [37,38], and Chlordimeform has in fact been banned in several countries.

5-Chloro-*o*-toluidine, 5-chloro-2-methylaniline [*95-79-4*] (**3**), M_r 141.6, *mp* 20 – 22 °C, *bp* 237 °C, is produced by reduction of the nitro group in 4-chloro-2-nitrotoluene, which is itself prepared by nitration of *p*-chlorotoluene followed by distillatory separation of 4-chloro-2-nitrotoluene from the accompanying 4-chloro-3-nitrotoluene isomer (see below 6-chloro- *m*-toluidine).

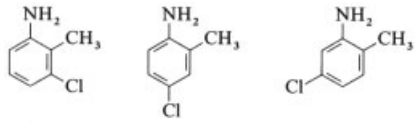

1 **2** **3**

6-Chloro-*o*-toluidine, 6-chloro-2-methylaniline [*87-63-8*] (**4**), M_r 141.6, *mp* 8 – 9 °C, *bp* 220 °C, ϱ 1.15, is co-produced along with 4-chloro-*o*-toluidine, as described above, and is used as an intermediate in the manufacture of azo and xanthene dyes [39] as well as pharmaceuticals [40,41].

6-Chloro-*m*-toluidine, 6-chloro-3-methylaniline [*95-81-8*] (**5**), M_r 141.6, *mp* 29 – 30 °C, *bp* 228 – 230 °C, is produced by the nitro reduction of 4-chloro-3-nitrotoluene, where the nitro intermediate is co-produced in the nitration of *p*-chlorotoluene described above under 5-chloro-*o*-toluidine. It is used in the manufacture of azo dyes.

3-Chloro-*p*-toluidine, 3-chloro-4-methylaniline [*95-74-9*] (**6**), M_r 141.6, *mp* 24 – 25 °C, *bp* 237 – 238 °C, is produced by the catalytic hydrogenation [8] of 2-chloro-4-nitrotoluene, obtained in the chlorination of *p*-nitrotoluene. It serves as an intermediate in the manufacture of the herbicide Chlortoluron [42,43] and the red pigment intermediate "2B-Acid."

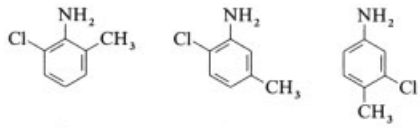

4 **5** **6**

6.3. Toluidinesulfonic Acids

There are ten distinct toluidinesulfonic acid isomers. Several of the isomers are difficult to prepare, and only five have found commercial use. Toluidinesulfonic acids find application in the manufacture of direct dyes, acid dyes, pigments, and triarylmethane dyes.

The toluidinesulfonic acids are produced either by sulfonation of the appropriate nitrotoluene, or by nitration of a toluenesulfonic acid

followed by catalytic hydrogenation or chemical reduction of the nitro group. Other methods involve direct sulfonation of the appropriate toluidine isomer.

Commercial toluidinesulfonic acids are light yellow to red. Often they are hygroscopic solids without defined melting points. Just as with other amino acids, the toluidinesulfonic acids form zwitterionic inner salts that are sparingly soluble in water and dilute mineral acids, but very soluble in aqueous caustic (→ Amino Acids, Chap. 2.).

3-Amino-4-methylbenzenesulfonic acid, 3-amino-*p*-toluenesulfonic acid [*618-03-1*] (**7**), M_r 187.2, is produced by the low-temperature sulfonation of *o*-nitrotoluene with oleum followed by catalytic hydrogenation or chemical reduction. It is used in the manufacture of azo dyes and nonlinear optical materials [44].

4-Amino-3-methylbenzenesulfonic acid, 4-amino-*m*-toluenesulfonic acid [*98-33-9*] (**8**), M_r 187.2, is produced by the sulfonation of *o*-toluidine. The conditions entail heating at 150 – 200 °C in concentrated sulfuric acid, or heating the *o*-toluidine sulfate salt either neat or in an inert solvent. The compound is used as an intermediate in the manufacture of azo and diazo acid dyes [45], pigments [46], and fluorescent brighteners [47].

4-Amino-2-methylbenzenesulfonic acid, 4-amino-*o*-toluenesulfonic acid [*133-78-8*] (**9**), M_r 187.2, is produced by sulfonation of *m*-nitrotoluene to give 4-nitro-2-methylbenzenesulfonic acid, followed by catalytic hydrogenation or chemical reduction. Much older technology involves baking *m*-toluidine sulfate or heating *m*-toluidine with sulfamic acid. The compound is used in the manufacture of azo dyes.

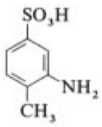

7

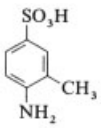

8

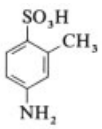

9

5-Amino-2-methylbenzenesulfonic acid, 5-amino-*o*-toluenesulfonic acid [*118-88-7*] (**10**), M_r 187.2, is produced by the sulfonation of *p*-nitrotoluene with oleum or SO_3, followed by catalytic hydrogenation or chemical reduction. It is used as an intermediate in the manufacture of azo and anthraquinone acid dyes.

2-Amino-5-methylbenzenesulfonic acid, 6-amino-*m*-toluenesulfonic acid [*88-44-8*] (**11**), M_r 187.2, is produced by direct sulfonation of *p*-toluidine with either concentrated sulfuric acid at high temperature or various concentrations of oleum [48], or by heating *p*-toluidine sulfate salt neat [49] or in an inert solvent [50]. The compound is better known as 4B Acid, and is used as an intermediate in the manufacture of several azo acid dyes, plastics, and printing-ink pigments (e.g., Pigment Red 57, Lithol Rubine), as well as fluorescent brighteners [51].

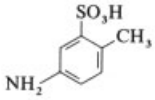

10

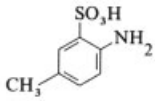

11

6.4. Chlorotoluidinesulfonic Acids

There are thirty different chlorotoluidinesulfonic acid isomers, only a few of which have commercial applications. The commercial potential of these compounds is determined by the availability of appropriate starting materials and the selectivity of the manufacturing process. Many of the possible isomers are prepared as mixtures containing other isomers, and are unusable as such for most commercial applications. Thus, only reactions that produce essentially one isomer are of significance. Commercial applications center around the manufacture of dyes and pigments for the textile, paint, printing ink, and plastics industries.

Chlorotoluidinesulfonic acids can be produced by a variety of methods depending on the choice of starting materials. Chlorination of nitrotoluenesulfonic acids or sulfonation of chloronitrotoluenes both result in chloronitrotoluenesulfonic acids, which can then be catalytically hydrogenated or chemically reduced to the desired product. Care must be taken in selecting the hydrogenation conditions, because catalytic dehalogenation can become the major reaction path with certain catalysts (e.g., very active platinum or palladium catalysts). Some nickel and

poisoned platinum and palladium catalysts minimize the extent of dehalogenation [33]. Additional manufacturing methods for chlorotoluidinesulfonic acids include the chlorination of toluidinesulfonic acids and the sulfonation of chlorotoluidines.

Chlorotoluidinesulfonic acids also form zwitterionic inner salts, with solubility properties much like those of the toluidinesulfonic acids.

2-Amino-3-chlorotoluene-5-sulfonic acid, 4-amino-3-chloro-5-methylbenzenesulfonic acid [*6387-14-0*] (**12**), M_r 221.7, can be produced by the chlorination of acyl-protected 2-aminotoluene-5-sulfonic acid [52], by baking the sulfate salt of 6-chloro-2-methylaniline, or by sulfonation of 6-chloro-3-methylaniline with oleum. It serves as an intermediate in the manufacture of such azo acid dyes as Direct Red 108, C.I. 28165.

2-Amino-6-chlorotoluene-3-sulfonic acid, 2-amino-4-chloro-3-methylbenzenesulfonic acid [*79380-22-6*] (**13**), M_r 221.7, is produced either by sulfonation of 3-chloro-2-methylaniline with oleum at 70 – 90 °C or by heating the corresponding sulfate salt at 150 – 200 °C, the "baking method" [53]. It is an intermediate in the manufacture of azo dyes.

2-Amino-6-chlorotoluene-4-sulfonic acid, 3-amino-5-chloro-4-methylbenzenesulfonic acid [*6387-27-5*] (**14**), M_r 221.7, is produced by sulfonating 6-chloro-2-nitrotoluene with 20 % oleum to give 6-chloro-2-nitrotoluene-4-sulfonic acid, followed by catalytic hydrogenation [54] or chemical reduction. It is used as the diazo component for a series of acid azo dyes intended primarily for leather dying (e.g., Acid Brown 105, C.I. 33530).

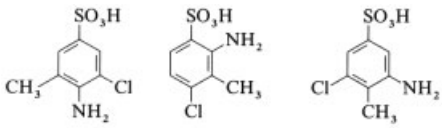

12 **13** **14**

3-Amino-6-chlorotoluene-4-sulfonic acid, 2-amino-5-chloro-4-methylbenzenesulfonic acid [*88-53-9*] (**15**), M_r 221.7, was traditionally prepared by a route that proceeds from *p*-toluenesulfonic acid via sequential chlorination, nitration, isomer separation, and reduction. More recent modifications begin with *m*-toluidine and involve acylation, chlorination, de-acylation, and sulfonation [55]. Yet another route uses *o*-chlorotoluene as starting material and proceeds via nitration, catalytic hydrogenation, and sulfonation of the amine [56]. Thiopolyols have been shown to be effective dehalogenation inhibitors during hydrogenation [57]. The compound is known commercially by several common names (e.g., Red Lake C Amine, Lake Red C Amine, C-Acid), and constitutes the diazo component for various azo dyes, metallized dyes, and pigments intended primarily for printing inks and plastics (e.g., Pigment Red 52, C.I. 15860, and Pigment Red 53, C.I. 15585).

4-Amino-6-chlorotoluene-3-sulfonic acid, 2-amino-4-chloro-5-methylbenzenesulfonic acid [*88-51-7*] (**16**), M_r 221.7, is made from *p*-nitrotoluene by first chlorinating to obtain 2-chloro-4-nitrotoluene, which is catalytically hydrogenated to the amine and finally sulfonated by heating the sulfate salt in *o*-dichlorobenzene to 100 – 120 °C [58], in sulfolane to 160 °C [59], or in 1,2-dichloroethane in an autoclave to 140 °C [60]. The compound is commonly known as Brilliant Toning Red Amine or Red 2B Acid, and is used as an intermediate in the manufacture of acid azo dyes. The azo dyes and their metallized salts are in turn used as pigments (e.g., Pigment Red 48, C.I. 15865) for paints, printing inks, and plastics [61, 62].

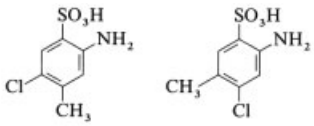

15 **16**

7. Toluidine Toxicology

The *o*-, *m*-, and *p*-toluidines as well as the chlorotoluidines belong to a group of chemicals known as potential "blue lip" compounds. This refers to the serious condition that results from overexposure to toluidines via inhalation, skin absorption, or even ingestion. All the toluidines

Table 2. Toxicity data for the isomeric toluidines [63]

	o-Toluidine	*m*-Toluidine	*p*-Toluidine
Inhalation (LC_{50})	ND∗	ND	ND
Humans (TCL_0)∗∗	25 mg/m^3	ND	ND
Ingestion, mg/kg (LC_{50})			
Rats	670	450	656
Mice	520	470	330
Rabbits	840	ND	270
Skin Contact :			
Rabbits	LD_{50} (3250 mg/kg)	mildly irritating (500 mg/24 h)	irritating (500 mg/24 h)
Eye Contact :			
Rabbits	corneal damage	mildly irritating (20 mg/24 h)	severe irritation (100 mg/24 h)

∗ ND = not determined.
∗∗ TLC_0 = lowest published toxic concentration.

are easily absorbed through the skin and mucous membranes.

Toluidines in the blood stream cause the conversion of hemoglobin to methemoglobin. The latter does not combine with oxygen, and in high concentrations renders the blood stream incapable of carrying oxygen from the lungs to the rest of the body. The earliest manifestation of poisoning is a bluish discoloration (*cyanosis*) of the lips, mucous membranes, fingernail beds, earlobes, and tongue. Minor levels of oxygen deficiency lead to a temporary sense of well-being and exhilaration. As the lack of oxygen increases, such symptoms as headache, drowsiness, and occasional nausea and vomiting appear. More profound decreases in blood oxygen lead to stupor, unconsciousness, and even death if treatment is not obtained immediately. All the toluidines are considered skin and eye irritants, and may cause an allergic skin reaction in some people. Pre-existing diseases of the central nervous system, cardiovascular system, liver, or bone marrow may be aggravated by excessive exposure to the toluidines. Toxicity characteristics are summarized in Table 2.

The December 1990 NIOSH Alert associated worker exposure to a combination of aniline and *o*-toluidine with an increased risk of bladder cancer; the effects of aniline and *o*-toluidine could not be separated. NIOSH concluded that *o*-toluidine was the more likely agent to be responsible for the observed bladder cancers [64]. No studies currently described in the literature directly link worker exposure to *m*- and *p*-toluidine to a carcinogenic effect in humans.

o-Toluidine, as the hydrochloride, has been tested for carcinogenicity in mice and rats by oral administration, and it was found to produce neoplasms at various sites in both species [65, 66]. No neoplasms were observed following subcutaneous injection in hamsters [67]. *p*-Toluidine showed cancerous potential in mice but not in rats, and *m*-toluidine did not produce definitive cancerous activity in either animal species. All three toluidines have been assigned a TLV of 2 ppm.

Chlorinated toluidines have received considerable attention as potential carcinogens. Animal tests have shown both 4-chloro-*o*-toluidine and 5-chloro-*o*-toluidine to be carcinogenic in mice, but not in rats [68, 69]. No convincing evidence was found for carcinogenicity in mice or rats with 3-chloro-*p*-toluidine [70]. One study established a strong correlation between worker exposure to 4-chloro-*o*-toluidine and increased incidence of bladder cancer [71, 72]. On the basis of this study a commission of the Deutsche Forschungsgemeinschaft categorized the compound as a human carcinogen, making it the first example of a monocyclic aromatic amine to be so designated.

All toxicity data indicate that care should be taken to minimize exposure when working with any toluidine, and extreme caution should be exercised when working with *o*-toluidine and its derivatives.

8. References

1. R. T. Morrison, R. N. Boyd: *Organic Chemistry,* 2nd ed., Allyn and Bacon, Inc., Boston 1966, pp. 720–721.
2. I. G. Farbenind., GB 274 562, 1926.
3. I. G. Farbenind., GB 279 283, 1926.

4. Lonza Ltd., DE 1 809 711, 1969 (L. Szigeth).
5. Ruhrchemie, DE 1 543 836, 1972 (W. Rottig).
6. Texaco Development Corp., DE 2 210 564, 1972 (J. Knifton, R. Suggitt).
7. Nippon Kayaku Co., JP 49 042 617, 1974 (K. Akiyoshi, S. Shimozawa, H. Danjo, S. Shizuhiko, S. Furuhashi).
8. Bayer, DE 2 743 610, 1979 (J. Zander).
9. GAF Corp., US 3 093 685, 1963 (H. Freyermuth, D. Graham, E. Hort).
10. Bayer, DE 2 135 154, 1973 (H. Thelen, W. Schwerdtel, K. Halcour, W. Swodenk).
11. Clayton Aniline Co., EP 2 308, 1979 (J. Hildreth, D. Haslam, D. Allen).
12. Du Pont de Nemours, US 4 212 824, 1980 (R. Seagroves).
13. P. M. G. Bavin, *Can. J. Chem.* **36** (1958) 238.
14. W. W. Hartman, H. L. Silloway, *Org. Synth. Coll.* **3** (1955) 83.
15. L. F. Fieser, M. Fieser, *J. Am. Chem. Soc.* **56** (1934) 1565.
16. ICI, DE 564 436, 1936.
17. Ashi Chemical Ind., JP 7 773 829, 1977 (S. Enomoto, T. Kamiyama, I. Tsutomu).
18. First Chemical Corp., Toluidine Sales Specifications, Pascagoula, Miss., 1993.
19. Ciba-Geigy, DE 2 328 340, 1973 (C. Vogel, R. Aebi).
20. Monsanto, US 4 258 196, 1981 (J. Chupp, R. Goodin).
21. BASF, FR 1 514 526, 1968.
22. Eli Lilly & Co., DE 2 250 077, 1973 (C. Paget).
23. ICI, EP 15 756, 1980 (K. Parry, P. Worthington, W. Rathmell).
24. Bayer, DE 2 262 830, 1974.
25. R. F. Storey, D. L. Smith: "Improved Room-Temperature Cure of Unsaturated Polyester Resins," *Mod. Plast.* **61** (1984) 58.
26. First Chemical Corp. and The University of Southern Mississippi, US 4 666 978, 1987 (R. Storey, M. Hogue, A. Bayer).
27. Bayer, DE 3 202 090, 1983 (E. Eimers, D. Margotte, R. Dhein, K. Kraft, W. Kloler).
28. Nippon Kayaku, JP 92 020 950, 1992.
29. Kawaguchi Chemical Ind., JP 82 139 542, 1981.
30. Ciba-Geigy, BE 800 471, 1973.
31. Biller, Michaelis & Co., US 2 991 311, 1958 (M. Thoma).
32. A. P. Bonds, H. Greenfield, "The Reductive Alkylation of Aromatic Amines With Formaldehyde," *Chem. Ind. (London)* **47** (1992) 65.
33. H. Greenfield, F. S. Dovell, "Metal Sulfide Catalysts for Hydrogenation of Halonitrobenzenes to Haloanilines," *J. Org. Chem.* **32** (1967) 3670.
34. Bayer, US 4 230 637, 1980 (J. Zander).
35. DuPont de Nemours, DE 2 615 079, 1975 (J. Kosak).
36. Ciba Corp., US 3 496 270, 1970 (C. Counselman).
37. K. Hooper, J. LaDou, J. Rosenbaum, S. Book: "Regulation of Priority Carcinogens and Reproductivity or Developmental Toxicants," *Am. J. Ind. Med.* **22** (1992) 793.
38. S. Xue et al.: "Assessment on the Carcinogenic Risk of Chlordimeform," *Huanjing Huaxue* **10** (1991) 54.
39. Hoechst, DE 2 504 925, 1976 (M. Haehnke, R. Neeb, T. Papenfuhs).
40. Beecham Group PLC, EP 61 907, 1982 (D. Smith).
41. Otsuka Pharmaceutical Co., FR 2 477 542, 1981 (K. Banno, T. Fujiok, M. Osaki, K. Nakagawa).
42. E. Gy, T. Gyogyszervegyeszeti Gyar, HU 19 950, 1981 (I. Bitter, A. Kis-Tamos, F. Jurak, L. Kocsis).
43. Valles Industria Organica, ES 2 027 074, 1992 (M. Camps-Anaya, M. Riba-Garcia).
44. Sony Corp., JP 05 210 130, 1993 (Y. To).
45. Nippon Kayaku Co., EP 190 947, 1986 (R. Ueno, S. Itoh, S. Fujimoto, Y. Komatsu, H. Tsuchiya).
46. Toyo Ink Mfg. Co., FR 2 564 099, 1985 (T. Funatsu, M. Hayashi, Y. Ohtomo).
47. Showa Chemical Ind., JP 53 102 329, 1978 (K. Suzuki).
48. Bayer, DE 3 401 572, 1985 (H. Blank, H. Emde, R. Schimp).
49. Ciba-Geigy Corp., US 4 681 710, 1987 (R. Miller, J. Billini, R. Schneider).
50. Bayer, DE 3 114 830, 1982 (H. Emde, H. Blank, P. Schnegg).
51. Ciba-Geigy, DE 2 233 429, 1973 (F. Guenter, W. Fringeli, P. Liechti).
52. BASF, DE 218 370, 1911.
53. Hoechst, DE 398 049, 1927.
54. Dinippon Ink and Chemical, Inc., JP 57 014 571, 1982.
55. Dainichiseika Color and Chemical Mfg. Co., JP 52 062 245, 1977 (K. Noguchi, T. Fukuda, S. Hitotsuyanagi).
56. Hoechst, DE 3 141 286, 1983 (T. Papenfuhs, R. Berthold).
57. Dainippon Ink and Chemicals, Inc., JP 57 021 362, 1982.

58. Combinatul Chimic, RO 76 842, 1981 (L. Vladutiu, C. Atanasiu-Meves, A. Spiliadis, D. Dobrescu, V. Aldea).
59. Bayer, DE 3 144 830, 1982 (H. Emde, U. Blank, P. Schnegg).
60. Bayer, DE 3 114 829, 1982 (H. Emde, U. Blank, P. Schnegg).
61. Toyo Ink Mfg. Co., JP 03 072 574, 1991 (Y. Otomo, H. Sugamo).
62. Hoechst, DE 3 833 226, 1990 (W. Deucker).
63. First Chemical Corp., Toxicity Data from Respective Material Safety Data Sheets, Pascagoula, Miss., 1992.
64. E. Ward et al.: "Excess Number of Bladder Cancers in Workers Exposed to Ortho-toluidine and Aniline," *JNCI J. Natl. Cancer Inst.* **83** (1991) 501.
65. *IARC Monogr.* **27** (1982) 155.
66. S. Hecht, K. El-Bayoumy, A. Rinenson, E. Fiala: "Comparative Carcinogenicity of o-Toluidine Hydrochloride and o-Nitrosotoluene in F-344 Rats," *Cancer Lett. (Shannon, Irel.)* **16** (1982) 103.
67. S. Hecht, K. El-Bayoumy, A. Rivenson, E. Fiala: "Bioassay for Carcinogenicity of 3,2′-Dimethyl-4-nitrosobiphenyl, o-Nitrosotoluene, Nitrosobenzene and the Corresponding Amines in Syrian Golden Hamsters," *Cancer Lett. (Shannon, Irel.)* **20** (1983) 349.
68. National Cancer Institute: "Bioassay of 4-Chloro-*o*-toluidine for Possible Carcinogenicity," DHEW Publ. no. NIH-79-1721, Washington, D.C., 1978, US Govt. Printing Off.
69. National Cancer Institute: "Bioassay of 5-Chloro-*o*-Toluidine for Possible Carcinogenicity," DHEW Publ. no. NIH-79-1743, Washington, D.C., 1978, US Govt. Printing Off.
70. National Cancer Institute: "Bioassay of 3-Chloro-*p*-Toluidine for Possible Carcinogenicity," DHEW Publ. no. NIH-78-1701, Washington, D.C., 1978, US Govt. Printing Off.
71. M. Stasik: "Carcinomas of the Urinary Bladder in 4-Chloro-o-toluidine Cohort," *Int. Arch. Occup. Environ. Health* **60** (1988) 21.
72. M. Stasik: "Urinary Bladder Cancer from 4-Chloro-o-toluidine," *Dtsch. Med. Wochenschr.* **116** (1991) 1444.

Toluidinesulfonic Acids → Benzenesulfonic Acids and Their Derivatives
Tomography → Visualization Techniques in Process Engineering
Toothpaste → Oral Hygiene Products

Transuranium Elements

LOTHAR KOCH, European Commission, Joint Research Centre, Institute for Transuranium Elements, Karlsruhe, Federal Republic of Germany

1.	**Isotopes of Transuranium Elements**	1
2.	**Properties of Transuranium Elements**	2
2.1.	**Physical Properties**	3
2.2.	**Chemical Properties**	3
3.	**Production**	7
4.	**Applications**	10
5.	**Transuranium Elements as Waste**	10
5.1.	**Radiotoxicity**	11
5.2.	**Partitioning and Transmutation**	12
6.	**References**	12

1. Isotopes of Transuranium Elements

Of the transuranium elements only the first ones – neptunium (Np), plutonium (Pu) (→ Plutonium and Plutonium Compounds), americium (Am), curium (Cm), and californium (Cf) – are of some technical relevance. All the other transuranium elements are short-lived, and of those with higher masses only few atoms have been synthesized.

Neptunium, plutonium, americium and curium are byproducts of nuclear energy generation. Except for plutonium, they are regarded as waste.

Neptunium, ^{239}Np, was the first transuranium nuclide to be discovered, by MCMILLAN and ABELSON in 1940. The isotope with the longest half-life, ^{237}Np with 2.14×10^6 years (Table 1), is used to study the chemical and physical properties of this element. It has accumulated on a tonnage scale in nuclear energy generation where it is formed by two reactions:

$$^{238}\mathrm{U}(n,2n)^{237}\mathrm{U}\xrightarrow[6.75\,d]{\beta^-}{}^{237}\mathrm{Np}$$

$$^{235}\mathrm{U}(n,\gamma)^{236}\mathrm{U}(n,\gamma)^{237}\mathrm{U}\xrightarrow[6.75\,d]{\beta^-}{}^{237}\mathrm{Np}$$

In light water reactor fuels, the second reaction is dominant.

^{237}Np can be regarded as the parent nuclide of the extinct decay chain $(4n+1)$ (see Fig. 5).

The americium isotope ^{241}Am was first discovered in 1944/45 by SEABORG and others. Like in nuclear power stations, it was formed by successive neutron capture by ^{239}Pu and subsequent decay of ^{241}Pu:

$$^{239}\mathrm{Pu}(n,\gamma)^{240}\mathrm{Pu}(n,\gamma)^{241}\mathrm{Pu}\xrightarrow[14.89\,a]{\beta^-}{}^{241}\mathrm{Am}\,(432\,a)$$

This nuclide, together with the longer-lived isotope ^{243}Am (7.38×10^3 a), is contained in spent nuclear fuels on a tonnage scale. The shorter-lived ^{242m}Am (141 a) is only present in low concentration in spent light water reactor fuels because of its high thermal fission cross section of 6600 barn, but accumulates in higher concentration in fast reactors.

Curium is the heaviest transuranium element produced on a tonnage scale in nuclear power stations. ^{242}Cm was the first of its isotopes to be discovered, in 1944 by SEABORG, JAMES and GHIORSO according to the reaction:

$$^{239}\mathrm{Pu}\,(\alpha,n)\,^{242}\mathrm{Cm}\,(162.8\,d)$$

They produced this nuclide by means of cyclotron-accelerated helium ions, which differs from the formation in nuclear fuels, where curium is formed by neutron capture of americium nuclides with successive β-decay. Neutron capture by the curium isotopes builds up the chain of heavier curium isotopes up to mass 248 (see Fig. 1 and Table 1).

Berkelium was first produced via a nuclear reaction by bombarding ^{241}Am with accelerated α-particles in 1949:

$$^{241}\mathrm{Am}\,(\alpha,2n)\,^{243}\mathrm{Bk}$$

There are two long-lived isotopes with mass numbers 247 and 249. The latter is part of the chain producing californium by neutron irradiation of ^{248}Cm (Table 1).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_167

Table 1. Nuclear properties of long-lived transuranium isotopes

Isotope	Half-life, a	Decay *	Main radiation, MeV	Thermal cross section, barn		Minimal critical mass in water, g	Minimal critical mass as metal sphere with steel reflector, kg
				Capture	Fission		
^{236}Np	1.55×10^5	EC, β	0.163 (γ)		2500		
^{237}Np	2.14×10^6	α	4.788 (51 %) 4.770 (19 %)	169	0.02		55
^{241}Am	432	α	5.486 (84 %) 5.443 (13 %)	83.6 + 748	3.15		34
^{242m}Am	141	α	5.207 (89 %) 5.141 (6.0 %)	1400	6600	19	3.8
^{243}Am	7.38×10^3	α	5.277 (88 %) 5.234 (11 %)	75 + 4	<0.07		36
^{242}Cm	0.45 6.1×10^6	α SF	6.113 (74 %) 6.070 (26 %)	16	<5		
^{243}Cm	28.5	α	5.785 (73 %) 5.742 (11 %)	225	600	122	
^{244}Cm	18.1 1.35×10^7	α SF	5.805 (76 %) 5.764 (24 %)	14	1.2		5.4
^{245}Cm	8.5×10^3	α	5.362 (93 %)	345	2020	41	
^{246}Cm	4.7×10^3	α	5.386 (79 %)	1.3	0.17		
^{247}Cm	1.6×10^7	α	4.869 (71 %)	60	90	2050	
^{248}Cm	3.4×10^5	α	5.078 (82 %)	4	0.34		
^{247}Bk	1.4×10^3	α	5.712 (17 %) 5.532 (45 %)	?	?		
^{249}Cf	351 6×10^{10}	α SF	5.812 (84 %)	465	1660	70	
^{250}Cf	13.1 1.7×10^4	α SF	6.031 (83 %)	2030	<350		
^{251}Cf	898	α	5.851 (27 %) 5.677 (35 %)	2850	4300	22	
^{252}Cf	2.64	α (97 %) SF (3.1 %)	6.118 (84 %)	20.4	32		

* $\alpha = \alpha$-decay, $\beta = \beta$-decay, SF = spontaneous fission, EC = electron capture.

Californium is the transuranium element with the highest atomic number so far discovered that has isotopes with half-lifes longer than years (Table 1). The first nuclide to be synthesized was ^{245}Cf, obtained by bombarding ^{242}Cm with cyclotron-accelerated helium ions according to the reaction ^{242}Cm (α,n) ^{245}Cf in 1950 by Thompson, Street, Ghiorso, and Seaborg.

The heavier transuranium elements einsteinium (99) and fermium (100) were first detected in the fall-out of an atom bomb test (Bikini Atoll, 1952). They have since been produced by multiple neutron capture of californium targets. The higher transuranium elements (up to 116) are obtained by nuclear reactions of heavy ions on Pb or Bi targets (Table 2).

2. Properties of Transuranium Elements

The elements up to the proton number 103 belong to the actinide elements where, similar to the lanthanides, an inner f shell is filled (in the case of the actinides, $5f$). From 104 to 110, the $6d$ and $7p$ shells are completed.

2.1. Physical Properties

Neptunium, americium and curium are metals with silvery appearance and malleability like uranium metal. When they are exposed to air, an oxide layer readily forms. When heated, the metals undergo one or two phase transitions (Ta-

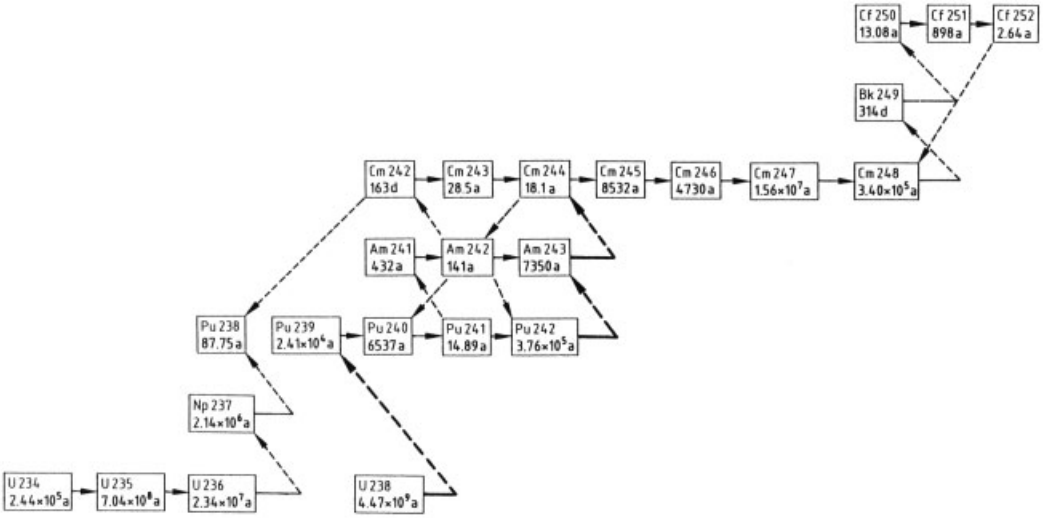

Figure 1. Buildup chain of long-lived actinides in nuclear fuels, with major and minor pathways (dashed lines symbolize short-lived radioactive decay)

ble 3) before melting. They densify under pressure, passing through different phases to an α-U type structure.

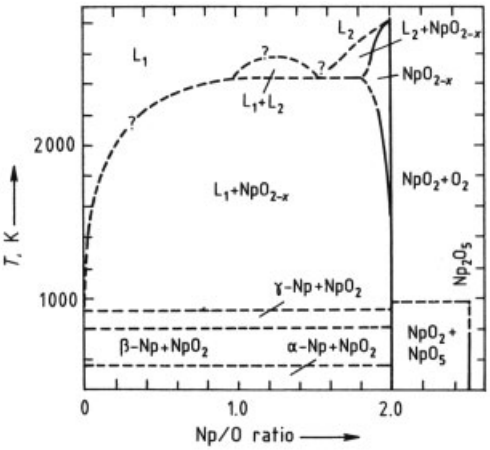

Figure 2. Phase diagram of Np – O (L = liquid phase; L_1, L_2 have different oxygen contents)

2.2. Chemical Properties

Contrary to the lanthanides, the lighter actinides show a large variety of valences and consequently form a wealth of chemical compounds.

Neptunium forms numerous compounds (hydrides, halogenides, chalgonides, pnictides) and organo-complexes, some of which are listed in Tables 4 and 5. Similar to thorium, it reacts with oxygen to give neptunium dioxide (NpO_2); oxides with lower oxidation states cannot be obtained (see Fig. 2). Only dehydration of tri- or pentavalent neptunium hydroxides yields other oxides. In ternary oxides with alkali and alkaline earth elements higher oxidation states (5+, 6+, 7+) are observed. Like plutonium, five neptunium ions with different oxidation states exist in aqueous solution, easily distinguished by their characteristic colors which range from blue or violet (Np^{3+}) over yellow-green (Np^{4+}) to green (NpO_2^+), and pink (NpO_2^{2+}) to brownish red (NpO_3^+). The oxidation potentials in acid solution are:

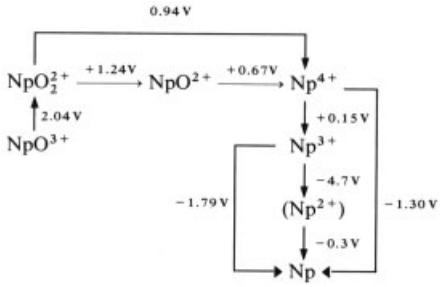

Table 2. Discovery of the transcalifornium elements [4]

Proton Number	Element ∗	Symbol	Discovered by	By the reaction	First isotope	Longest-lived isotope
99	Einsteinium	Es	Ghiorso, Thompson, Higgins	thermonuclear explosion	^{253}Es	^{254}Es
100	Fermium	Fm	Seaborg, Studier, Fields, Fried, Diamond, Mech, Pyle, Huizenga, Hirsch, Mannig, Brown, Smith, Spence (1952)	thermonuclear explosion	^{255}Fm	^{257}Fm
101	Mendelevium	Md	Ghiorso, Harvey, Choppin, Thompson, Seaborg (1955)	^{235}Es (α,n)	^{256}Md	^{258}Md
102	Nobelium (Flerovium)	No (Fl)	Ghiorso, Sikkeland, Walton, Seaborg (1958)	$^{246}_{96}$Cm($^{12}_{6}$C,4n)	^{254}No	^{259}No
			Flerov, Polikanov, Karmanjan, Pasink, Parfanovich, Tarantin, Druin, Volkov, Semchinova, Oganesian, Khalisev, Khlebnikov, Miasoedov, Gavrilov (1957/58)	$^{241}_{94}$Pu($^{16}_{8}$O,5n)	^{252}Fl	
103	Lawrencium	Lr	Ghiorso, Sikkeland, Larsh, Latimer (1959)	$^{249-252}_{98}$Cf + ^{10,11}B	^{257}Lr	^{260}Lr
			Russian group, Dubna	^{243}Am($^{18}_{8}$O,5n)	^{256}Lr	
104	Rutherfordium (Dubnium)	Rf (Db)	Flerov, Oganesian, Lobanov, Kusnezov, Druin, Ferelygin, Gavrilov, Tretjakova, Plotko (1964)	$^{242}_{94}$Pu($^{22}_{10}$Ne,4n)	^{260}Rf	^{261}Rf
			Ghiorso, Nurmia, Harris (1969)	$^{249}_{98}$Cf($^{12}_{6}$C,4n)	^{257}Db	
105	Dubnium (Joliotium)	Db (Jl)	Ghiorso, Nurmia, K. Eskola, Harris, P. Eskola (1970)	$^{249}_{98}$Cf($^{15}_{7}$N,4n)	^{260}Db	^{262}Db
			Flerov, Oganesian, Lobanov, Lazarev, Tretjakova, Kolesov, Plotko (1970)	$^{243}_{95}$Am($^{22}_{10}$Ne,4n)	^{261}Jl	
106	Seaborgium	Sg	Ghiorso, Nitschke, J. R. Alonso, C. T. Alonso, Nurmia, Seaborg, Hulet, Lougheed (1974)	^{249}Cf($^{18}_{8}$O,4n)	^{263}Sg	^{263}Sg
			Oganesian, Ilanov, Demin, Pleve, Tretjakova, Plotko, Ivanov, Danilov, Krotkin, Flerov (1974)	^{207,208}Pb + ^{54}Cr	^{259}Sg	
107	Bohrium	Bh	Oganesian, Demin, Danilov, Ivanov, Ilinov, Kolesnikov, Marfkov, Plotko, Tretjakova, Flerov (1976)	^{209}Bi(^{54}Cr,2n)	^{261}Bh	
			Münzenberg, Armbruster, Faust, Güttner, Hessberger, Hofmann, Reisdorf, Sahm, Schmidt, Schött, Thuma, Vermeulen	^{209}Bi(^{54}Cr,n)	^{262}Bh	
108	Hassium (Hahnium)	Hs (Ha)	Münzenberg et al. (1981)	^{208}Pb + ^{58}Fe $\rightarrow$ ^{265}Hs + n	^{265}Hs	
109	Meitnerium	Mt	Münzenberg et al. (1984)	^{209}Bi(^{58}Cr,n)	^{266}Mt	
110	Ununnilium	Uun	Hofmann et al. (1994)	Pb(Ni)	^{269}Uun	
111	Unununium	Unn	Hofmann et al. (1994)	^{209}Bi(^{64}Ni,n)	^{272}Unn	
112	Ununbium	Unb				
114	Ununguadium	Ung				
116	Ununhexium	Unh				

∗ Names are those proposed by IUPAC in 1995 and due for confirmation in 1997. Previous names were: 102 Nobelium (No); 104 Rutherfordium (Rf); 105 Hahnium (Ha); 106 Rutherfordium (Rf), since 1994; 107 Nielsbohrium (Ns); 108 Hassium (Hs).

Like plutonium, neptunium can be present simultaneously in different oxidation states in solution, which makes its isolation from other nuclides difficult.

Americium. A large variety of compounds of americium have been synthesized and described. Americium reacts with oxygen to give the sesqui- and dioxides (Fig. 3).

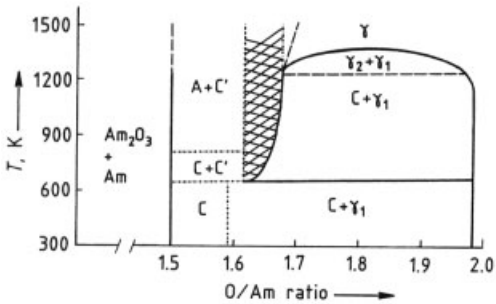

Figure 3. Am – O phase diagram. A = hexagonal Am_2O_3; C = low-temperature bcc Am_2O_3; C′ = high-temperature bcc Am_2O_3; γ, γ_1, γ_2 = fcc AmO_{2-x}

Table 3. Structural data for transuranium elements

Phase	Symmetry	Unit cell dimensions, pm	X-ray density, g/cm^3	Transformation temperature at STP, °C
Neptunium				
α	orthorhombic	a = 666.3 b = 472.3 c = 488.7	20.45	280
β	tetragonal	a = 489.7 c = 338.8	19.36	576
γ	cubic body centred (bcc)	a = 351.8	18.04	
Americium				
α	double hexagonal close packed (dhcp)	a = 346.8 c = 1124.1	13.6	ca. 658
β	face centred cubic (fcc)	a = 489.4	13.65	
Curium				
α	double hexagonal close packed (dhcp)	a = 349.6 c = 1133.1	13.5	ca. 1277
β	face centred cubic (fcc)	a = 503.9	12.9	
Berkelium				
α	double hexagonal close packed (dhcp)	a = 341.6 c = 1106.9	14.79	ca. 930
β	face centred cubic (fcc)	a = 499.7	13.24	ca. 986
Californium				
α	double hexagonal close packed (dhcp)	a = 338.4 c = 1104.0	15.10	
β	face centred cubic (fcc)	a = 574	8.74	ca. 900
Einsteinium				
α	double hexagonal close packed (dhcp)	a = ? c = ?	?	
β	face centred cubic (fcc)	a = 575	8.84	ca. 860

Like neptunium, higher oxidation states exist in ternary oxides with alkaline elements. There is an AmH_{2+x} (fcc) phase, which is isostructural with most of the lanthanide, plutonium, and neptunium dihydrides. A trigonal phase close to AmH_3 has also been observed. Halogenides, chalcogenides and pnictides of Am have been obtained; some trivalent Am compounds are listed in Table 4. Complexes with organic ligands that have been used in large-scale Am separation are listed in Table 5. In aqueous solution, americium can occur in five oxidation states: pink or yellow Am^{3+}, yellow AmO_2^+ and rum-colored AmO_2^{2+}. The colors of the other oxidation states are unknown. The oxidation potentials are given for 1 M $HClO_4$:

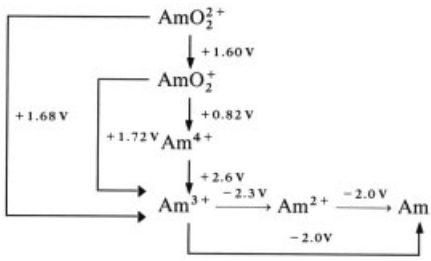

In basic media, Am can be stabilized in higher oxidation states than +3, which has been exploited for Am separation. The oxidation potentials are:

Table 4. Binary compounds of the transuranium elements

Formula	Symmetry	Lattice constants, pm			Angle (deg)
		a	*b*	*c*	
NpH_2	cubic	534.3			
NpH_3	trigonal	651		672	
NpO_2	cubic	542.5			
Np_2O_5	monoclinic	418.3	658.4	408.6	$\beta = 90.32$
NpF_3	trigonal	712.9		728.8	
NpF_4	monoclinic	1268	106.6	834	$\beta = 126.3$
NpF_5	tetragonal	653		445	
NpF_6	orthorhombic	990.9	899.7	520.2	
$NpCl_3$	hexagonal	741.3		428.2	
$NpBr_3$	hexagonal	791.6		439.0	
NpI_3	orthorhombic	430	1403	995	
NpN	cubic	489.9			
Np_2C_3	cubic	810.2			
AmSb	fcc	623.9			
AmAs	fcc	588			
AmBi	fcc	633.8			
AmB_4	tetragonal	710.5		400.6	
AmB_6	simple cubic	411.5			
$AmBr_2$	tetragonal	1159		712.1	
$AmBr_3$	orthorhombic	406.4	1266	914.4	
$AmCl_2$	orthorhombic	896.3	757.3	453.2	
$AmCl_3$	hexagonal	738.2		421.4	
AmF_3	trigonal	704.4		722.5	
AmF_4	monoclinic	1253.8	1051.6	820.4	$\beta = 126.8$
AmH_2	fcc	534.8			
AmH_3	trigonal	668		675	
$Am(OH)_3$	hexagonal	642.6		374.5	
AmI_2	monoclinic	767.7	831.1	792.5	$\beta = 98.46$
α-AmI_3	orthorhombic	428	1394	997.4	
β-AmI_3	hexagonal	742		2055	
AmN	fcc	500			
AmS	fcc	559.2			
AmO_2	fcc	537.4			
α-Am_2S_3	orthorhombic	398	739	1536	
β-Am_2S_3					
γ-Am_2S_3	bcc	843.4			
$AmTe_2$	tetragonal	436.6		896.9	
$AmTe_3$	orthorhombic	439.9	433.9	2557	
A-Cm_2O_3	hexagonal	379.2		598.5	
B-Cm_2O_3	monoclinic	1428.2	364.1	888.3	$\beta = 100.29$
C-Cm_2O_3	cubic	1100.2			
CmO_2	cubic	535.84			
CmS	fcc	557.54			
$CmS_{1.98}$	tetragonal	392.6		801	
CmN	fcc	504.1			
CmP	fcc	574.3			
CmAs	fcc	588.7			
CmSb	fcc	624.2			
CmF_3	trigonal	701.9		719.8	
$CmCl_3$	hexagonal	737.43		418.50	
$CmBr_3$	orthorhombic	404.1	1270	913.5	
CmI_3	hexagonal	744		204	
CmF_4	monoclinic	1250.0	1048.8	818.3	$\beta = 126.10$

Table 4. (Continued)

Formula	Symmetry	Lattice constants, pm			Angle (deg)
		a	*b*	*c*	
CmH_{2+x}	fcc	632.2			
$CmH_{3-\delta}$	trigonal	652.8		673.2	
Cf_2O_3	bcc	108.39			
CfF_3	trigonal	69.45		71.01	
CfF_3	orthorhombic	66.53	70.39	43.93	
$CfCl_3$	hexagonal	73.79		40.90	
$CfCl_3$	orthorhombic	38.59	117.48	85.61	

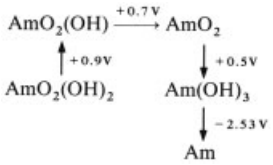

Curium. Unlike the lighter actinides, curium exists only in the trivalent and tetravalent states. Since it can be obtained in gram amounts, its compounds have been synthesized and described. Two oxides exist: the sesquioxide Cm_2O_3 and the dioxide CmO_2. Trivalent curium forms binary compounds analogous to those of americium (Table 4) and complexes with organic ligands that are employed in Cm separation (Table 5). In aqueous solution, only the pale green trivalent ion is stable, because of the high oxidation potential of the tetravalent ion:

$$Cm^{4+} \xrightarrow{+3.1\,V} Cm^{3+} \xrightarrow{-2.06\,V} Cm$$

Californium has found widespread use as a spontaneous-fission neutron source; the properties of the oxide and halides are listed in Tables 4 and 5. Organic complexes used for Cf separation are listed in Table 5.

3. Production

Spent nuclear fuels are the only significant source of the lighter transuranium elements. The most abundant element is plutonium. Its recovery and use for energy production are described in → Nuclear Technology, Chap. 5.3.3., → Nuclear Technology, Chap. 5.4.

The amount of transuranium elements contained in a spent fuel varies for each reactor type. For common pressurized water reactors with an initial ^{235}U enrichment of 3.2 % and with a typical burnup of 33 GWd/t, the transuranium content of 1 t of freshly discharged fuel is:

Np	430 g
Pu	9500 g
Am	120 g
Cm	40 g
Bk	0.6 µg
Cf	11 µg

The masses of the lighter transuranium elements which will have accumulated by the year 2000, assuming an installed nuclear electrical energy generating capacity of 514 MW_e and excluding the material in the former Soviet Union is as follows:

^{237}Np	38 t
^{241}Am	34 t
^{243}Am	7 t
^{242}Cm	0.4 t
^{244}Cm	2.2 t

The separation of neptunium from spent nuclear fuels is achieved in two ways: by a modified PUREX process (→ Nuclear Technology, Chap. 5.4.) or from the high-level waste stream of this process. By carefully adjusting the oxidation state of neptunium and extending the extraction stages of the PUREX process, neptunium can be co-extracted with uranium and plutonium. Plutonium is stripped as Pu^{3+} under conditions where neptunium remains dissolved as Np^{4+}, and neptunium can then be stripped selectively from uranium as NpO_2^+.

The americium contained in spent nuclear fuel, like curium, remains together with most of the fission products in the high-level waste of nuclear fuel reprocessing.

Table 5. Stability constants of transuranium elements with inorganic and organic ligands

Ligand	T, °C	Ionic strength, μmol/L	Medium	Log of formation constants		
				β_1	β_2	β_3
F^- (Np^{4+})	20	4.0	perchlorate	8.26	6.13	5.78 4.7 (β_4)
Cl^- (Np^{4+})	20	1.0	perchlorate	0.04		
Cl^- (Np^{3+})	25	>7	chloride	4×10^{-3}	1.04×10^{-5}	
Br^- (Np^{3+})	25	>7	bromide	4×10^{-4}	2.9×10^{-7}	
Cl^- (Np^{3+})	25	>7	chloride	2.42	4.96	
Br^- (Np^{3+})	25	>7	bromide	3.39	6.48	
OH^- (Np^{4+})	25	2.0	perchlorate	2.3 (in H_2O)		
F^- (Np^{4+})	20	4.0	perchlorate	4.82	7.57	
Cl^- (Np^{4+})	20	1.0	perchlorate	0.04	0.24	0.48
SO_4^{2-} (Np^{4+})	25	2.0	perchlorate	2.43	3.47	
OH^- ($Np^{5+}O_2^+$)	25	0.1	chloride	8.9		
Cl^- ($Np^{5+}O_2^+$)	25	2.0	chloride	0.3		
F^- ($Np^{6+}O_2^{2+}$)	25	1.0	perchlorate	5.37		
Cl^- ($Np^{6+}O_2^{2+}$)	0	3.0	perchlorate	0.21	1.01	
SO_4^{2-} ($Np^{6+}O_2^{2+}$)	21	1.0	perchlorate	1.90	2.77	
H_4EDTA (ethylenediaminetetraacetic acid) (Np^{4+})	25	1.0	perchlorate	24.55		
H_5DTPA (diethylenetriaminepentaacetic acid (Np^{4+})	25	1.0	perchlorate	30.33		
Hlact (lactic acid) ($Np^{5+}O_2^+$)	25	0.2	perchlorate	1.56	2.21	
α-hib (α-hydroxyisobutyric acid) ($Np^{5+}O_2^+$)	25	0.1	perchlorate	1.62		
H_4EDTA ($Np^{5+}O_2^+$)	20	0.05	perchlorate	9.68		
Cl^- (Am^{3+})	25	13.7	chloride	−2.21 $[AmCl]^{2+}$	−4.70 $[AmCl_2]^+$	
OH^- (Am^{3+})	25	0.005	perchlorate	10.7 $[AmOH]^{2+}$	20.9 $[Am(OH)_2]^+$	
NO_3^- (Am^{3+})	25	8.0	nitrate	−1.3 $[AmNO_3]^{2+}$		
NO_3^- (Am^{3+})	25	1.0	perchlorate	0.25		
H_5DTPA (Am^{3+})	25	0.1	chloride	23.07 $[AmDTPA]^{2-}$ 14.06 $[AmHDTPA]^-$		
H_4EDTA (Am^{3+})	25	0.1	perchlorate	18.15 $[AmEDTA]^-$		
α-hib (Am^{3+})	25	0.5	perchlorate	2.88	4.03	
Hlact (lactic acid) (Am^{3+})	25	2.0	perchlorate	2.52 $[AmLact]^{2+}$	4.77 $[Am(Lact)_2]^+$	5.98 $Am(Lact)_3$
H_2Ox (oxalic acid) (Am^{3+})	20	0.1	chloride		8.3 $[Am(Ox)_2]^-$	11.8 $[Am(Ox)_3]^{3-}$
HTTA (thenoyltrifluoroacetone) (Am^{3+})	25	0.1	perchlorate	3.4 $[AmTTA]^{2+}$	8.5 $[Am(TTA)_2]^+$	13.3 $Am(TTA)_3$
F^- (Cm^{3+})	25	0.5	perchlorate	3.34	6.18	9.08
SCN^- (Cm^{3+})	25	1.0	perchlorate	0.18	0.61	
NO_3^- (Cm^{3+})	23	1.0	perchlorate	0.34	0.11	
Cl^- (Cm^{3+})		0.5	chloride	0.20	0.09	
acetate (Cm^{3+})	20	0.5	perchlorate	2.06	3.09	
Hlact/di(2-ethylhexyl)phosphoric acid (Cm^{3+})	RT (room temperature)	0.5	perchlorate	2.74	4.47	6.11

Table 5. (Continued)

Ligand	T, °C	Ionic strength, µmol/L	Medium	Log of formation constants		
				β_1	β_2	β_3
H_4EDTA (Cm^{3+})	25	0.1	chloride	17.4		
α-hib (Cm^{3+})	RT	0.5	perchlorate	3.43	4.71	5.23
HTTA (Cm^{3+})	25	0.1	perchlorate			13.4
F^- (Cf^{3+})	25	1.0	perchlorate	3.03		
OH^- (Cf^{3+})	25	0.1	chloride	5.62		
SO_4^{2-} (Cf^{3+})	25	2.0	perchlorate	1.36	2.07	
SCN^- (Cf^{3+})	25	1.0	perchlorate	3.06		
				3.71	0.28	2.65
Hlact (Cf^{3+})	20	0.5	perchlorate			6.09
α-hib (Cf^{3+})	25	0.5	perchlorate			6.9
H_4EDTA (Cf^{3+})	25	0.1	perchlorate	19.09		
H_5DTPA (Cf^{3+})	25	0.1		22.6		
				25.19		
HTTA (Cf^{3+})	25	0.1	perchlorate	6.90		14.94

Since ^{241}Am accumulates in stockpiled plutonium by decay of ^{241}Pu, it can be easily separated, e.g., by sorption of the Pu(IV) hexanitrato complex on anion exchanger and washing Am off with 8 M HNO_3.

Curium has been produced on a kilogram scale by irradiating plutonium, with a yield of about 10 %. After several extraction and ion-exchange processes, remaining Pu is extracted by tri-*n*-butyl phosphate (TBP). Am and Cm chlorides have been extracted with tertiary amines from low acid 11 M LiCl (TRAMEX process) leaving behind the lanthanides, formed by fission of Pu.

Cm has been used as source material to produce gram amounts of californium. Analogous to the TRAMEX process, the lanthanides are removed from the actinides (Am, Cm, Bk, Cf), which are separated by HPLC, e.g., cation exchange where the lanthanides are absorbed from nitrate solution and sequentially eluted with 0.25 M α-hydroxyl isobutyric acid at pH gradually varied from 3.8 to 4.2.

Like the lighter transuranium elements, californium accumulates in spent nuclear fuels.

The longer-lived transuranium nuclides which accumulate in spent fuel as radioactive waste possess considerable capture and/or fission neutron cross sections (Table 1). This offers the possibility of transmuting the radiotoxic alpha-emitting transuranium nuclides into less radiotoxic waste. In handling transuranium elements, it must be borne in mind that a few nuclides have rather low critical masses in solution, as is the case with plutonium. In particular ^{242}Am has a critical mass of only 19 g when dissolved in a water sphere. Also, the metal of certain transuranium nuclides can have critical masses (Table 1) as low as 5.4 kg (for ^{244}Cm). Use for military purposes is limited, due to the high specific radioactivity of most of the nuclides. They would emit too many neutrons which lead to preignition and/or heating to such an extent that the chemical explosive for the implosion would degrade. Only ^{237}Np can be used for nuclear explosives.

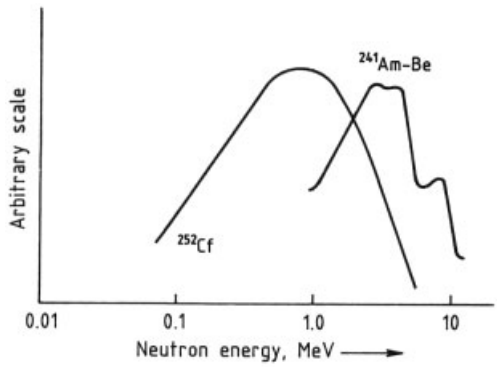

Figure 4. Energy spectrum of neutrons emitted from a ^{252}Cf and ^{241}Am-Be source

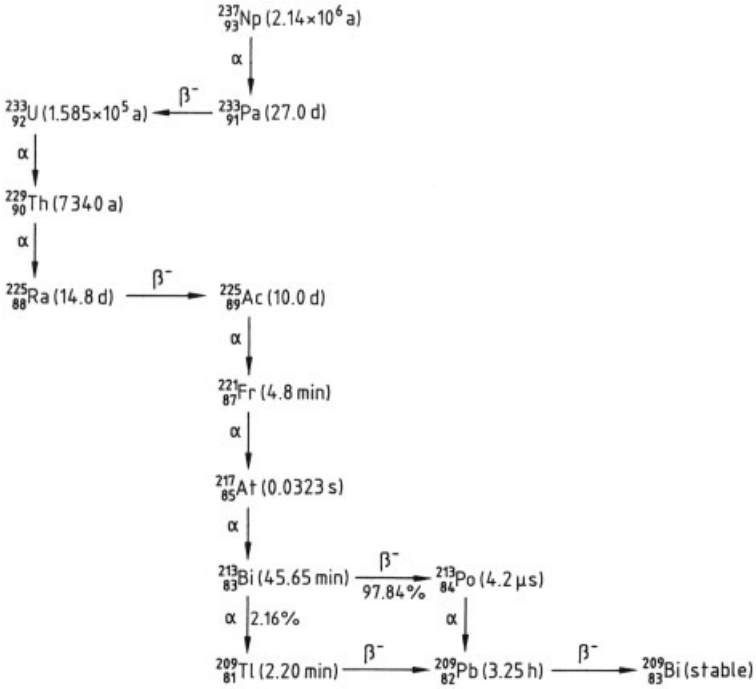

Figure 5. Extinct decay chain $4n+1$

4. Applications

A commercial application for Np, Am, and Cm – often referred to as minor actinides (MA) – so far has not been found except for negligible amounts of ^{237}Np used to produce ^{238}Pu for thermo-electric devices and for some ^{241}Am to produce pure ^{238}Pu by decay of ^{242}Cm for the same purpose (the decay heat of ^{242}Cm is 120 W/g ^{242}Cm). ^{241}Am is also in use as Am/Be neutron sources by (α,n) reaction (Fig. 4). Some smoke detectors contain ^{241}Am. From the longer-lived Cf nuclides ^{252}Cf serves as a fission neutron source (Fig. 4), and because of its spontaneous fission, fission fragments are used as an ionization source, e.g., in mass spectrometry.

5. Transuranium Elements as Waste

In general, the minor actinides are regarded as waste. All the minor actinides decay into four decay chains, of which three are natural ones with parent nuclides ^{232}Th ($4n$), ^{238}U ($4n+2$), ^{235}U ($4n+3$). (The mass numbers of the three chains obey the given algorithm with n as integer). The chain $4n+1$ (Fig. 5) had disappeared but is now being resurrected thanks to ^{237}Np. Since this nuclide is being produced on a tonne scale, it will testify far into the future that nuclear energy was once used on earth.

Table 6. Annual limits (Bq) of inhalation and ingestion as recommended by ICRP in 1990

Nuclide	Inhalation	Ingestion
^{237}Np	300	3×10^4
^{241}Am	300	3×10^4
^{242m}Am	300	4×10^4
^{243}Am	300	3×10^4
^{242}Cm	6000	9×10^5
^{243}Cm	400	5×10^4
^{244}Cm	500	6×10^4
^{245}Cm	300	3×10^4
^{246}Cm	300	3×10^4
^{247}Cm	300	4×10^4
^{248}Cm	80	9×10^3
^{249}Bk	1×10^5	1×10^7
^{250}Cf	500	6×10^4
^{251}Cf	200	3×10^4
^{252}Cf	900	1×10^5
^{253}Es	2×10^4	2×10^6
^{254}Es	3000	4×10^5
^{253}Fm	1×10^5	1×10^7
^{257}Fm	4000	7×10^5
^{258}Md	6000	9×10^5

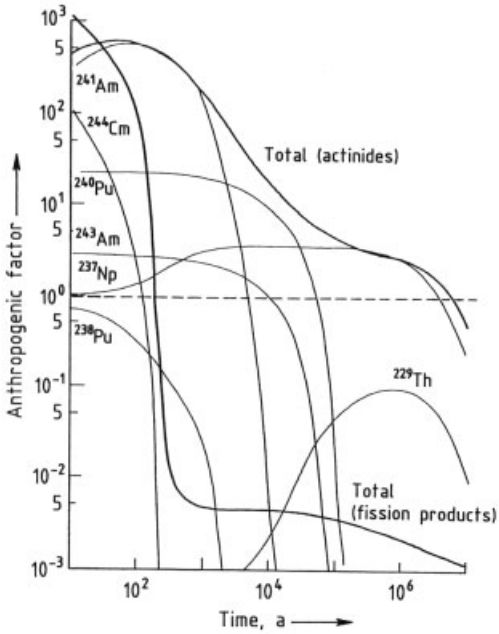

Figure 6. Time-dependence of radiotoxicity in a spent LWR fuel (burnup 33 GWd/t), normalized to the radiotoxicity of the uranium ore (dashed line) mined to produce the fuel

The present waste strategy for spent fuel is either direct disposal or to extract only plutonium and uranium, but discharge the other transuranium nuclides together with fission products as vitrified waste to a geological repository.

5.1. Radiotoxicity

The transuranium elements consist mainly of α-emitting nuclides which, when incorporated, are deposited in critical organs, where – as has been shown by animal experiments – malicious tumors may develop. Neptunium tends to stay on the surface of the bones, whereas the transplutonium elements are deposited mainly in the skeleton and liver.

The annual limits of intake for workers are given in Table 6. For the public, limits roughly a factor of 100 lower are set. Inhalation is the most critical intake path; consequently, except for μCi amounts, the transuranium elements are handled in glove-boxes. Because of the possible neutron emission by (α,n) reactions (depending on the other elements present) and spontaneous fission (especially curium and californium) additional shielding by water-filled walls is needed to reduce the occupational dose to the accepted limit. To improve the absorption of the thermalized neutrons, borates or cadmium can be added to the water.

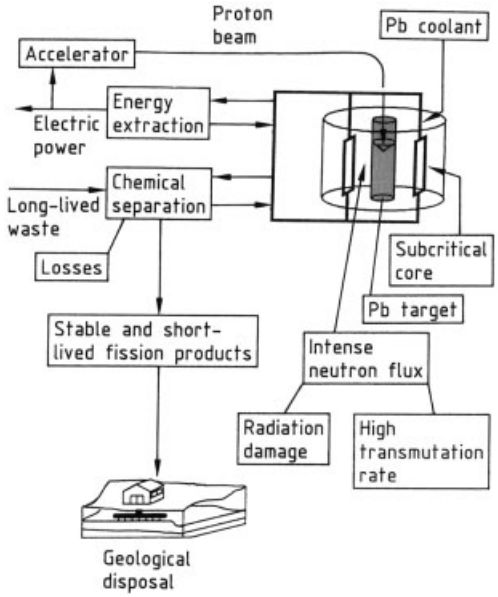

Figure 7. Concept of an accelerator-driven reactor for transmutation of long-lived waste

5.2. Partitioning and Transmutation

The time-dependence of the radiotoxicity resulting from energy generation in a typical light water reactor is shown in Figure 6. Within a few hundred years, the radiotoxicity of fission products drops below that of the uranium ore mined to produce the fresh fuel.

The accumulated transuranium elements would decay over much longer periods. Hence, their transmutation into short-lived or stable nuclides would reduce the radiotoxic hazard for future generations. This view looks only at the source of the radiotoxic hazard as it is posed by the spent fuel in a geological repository. The leaching of the fuel and the eventual migration of nuclides back to the biosphere is a more relevant description of the imminent danger. Also fission product nuclides, even if being of lower

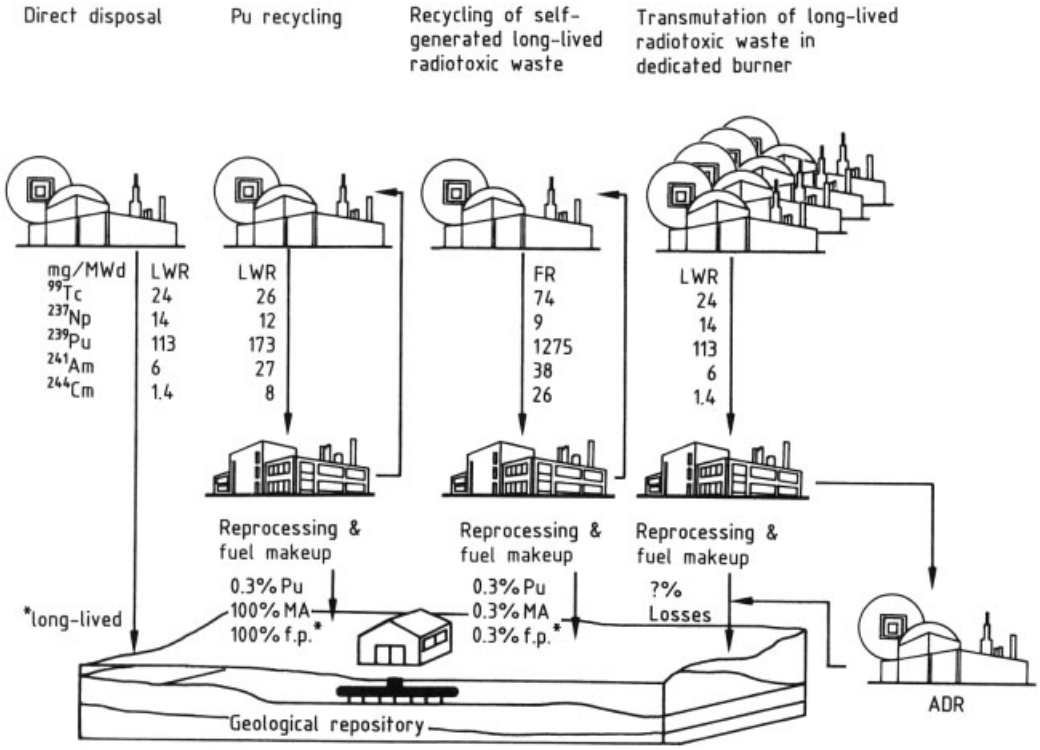

Figure 8. Spent fuel cycle options [mass (mg) of radionuclides discharged from the reactors is normalized to the generated energy (MW · d)]
MA = minor actinides;f.p. = fission produts;LWR = light water reactor;FR = fast reactor;ADR = accelerator-driven reactorIn the case of FR and ADR, the radiotoxic nuclides are transmuted in the reactor. Only the losses during recycling are discharged to the repository. For the FR these amount to ca. 0.3 % per cycle and in total less than 1 %.

quantities, could pose a problem if they are readily soluble (TcO_4^-, Cs^+, I^-) and accumulate in the biosphere.

A radiotoxic nuclide can be transmuted by nuclear reactions into a short-lived or stable one by several methods. Transmutation by photon-induced reactions or by charged particles seem to have technical difficulties. The most promising methods are transmutation by neutron capture or neutron-induced fission. Existing power stations could be used, but also newly-developed dedicated burner reactors driven by accelerators.

The main feature of the concept is a neutron-spallation source in which accelerated charged particles of about 1 GeV strike an initially non-radioactive heavy element target (e.g., Pb), the spallated nuclei produce a neutron shower, with up to 50 neutrons per incident particle. In an appropriate setup, a dense neutron flux is induced in the surrounding, previously subcritical fission reactor (Fig. 7). This is the unique safety feature of an accelerator-driven reactor compared to existing ones: as soon as the accelerator beam is turned off, the reactor is shut down.

To define a reduction of the nuclear hazard, the direct storage of spent light water reactor fuel is taken as a reference point. Figure 8 compares it to the concept of plutonium recycling in light water reactors and recycling of Np, Pu, Am, Cm, Tc in fast reactors. In the latter case, the radionuclides, which are self-generated by the nuclear reactor, are fed back. Instead of slow decay in a geological repository, they are transmuted faster by storage in the nuclear reactor. The equilibrium concentration given for each reactor fuel corresponds to a partitioning and transmutation cycle loss that gives a 100-fold overall hazard reduction.

6. References

1. J. J. Katz, G. T. Seaborg, L. R. Morss: *The Chemistry of the Actinide Elements,* 2nd ed., vol. 1, Chapman and Hall, London 1986.
2. *Gmelin,* 8th ed., System no. 71, part D 1.
3. Annual Limits of Intake of Radionuclides by Workers Based on the 1990 Recommendations, Annals of the International Committee for Radiation Protection (ICRP), (Publication 61), Pergamon Press, Oxford 1990.
4. L. F. Trueb, Die Chemischen Elemente, S. Hirzel Verlag Stuttgart (1996)

Triarylmethane and Diarylmethane Dyes

THOMAS GESSNER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

UDO MAYER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Definition and Chromophores . 2
1.1. Definition 2
1.2. Chromophores 3
2. History and Economic Importance 6
2.1. History 6
2.2. Economic Importance 7
3. Production 7
4. Uses 9
5. Diphenylmethane Dyes 9
5.1. Michler's Hydrol and Derivatives 9
5.2. Michler's Ketone, Auramines, and Derivatives 11
5.2.1. Michler's Ketone 11
5.2.2. Auramines 11
6. Triphenylmethane Dyes 12
6.1. Monoaminotriphenylmethane Dyes 12
6.1.1. Cationic Dyes 12
6.1.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones) . . 13
6.2. Diaminotriphenylmethane Dyes 13
6.2.1. Cationic Dyes 13
6.2.2. Nonionic Derivatives (Triphenylmethanol and Triphenylmethylamine Derivatives) 16
6.2.3. Anionic Dyes 17
6.3. Triaminotriphenylmethane Dyes 19
6.3.1. Cationic Dyes 19
6.3.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones) . . 24
6.3.3. Betaines and Anionic Dyes 26
6.4. Hydroxytriphenylmethane Dyes 28
6.4.1. Metal-Complex Dyes 29
6.4.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones) . . 30
7. Heteroaryldiarylmethane and Diheteroarylarylmethane Dyes . 31
7.1. Cationic Dyes 31
7.1.1. Monoaminodiphenylheteroarylmethane Dyes 31
7.1.2. Diaminodiphenylheteroarylmethane Dyes 31
7.2. Nonionic Derivatives (Triphenylmethanol Derivatives, Lactones, and Leuco Compounds) 32
8. Dyes with Several Chromophores 33
8.1. Bis(triarylmethane) Dyes 34
8.2. Oligomeric and Polymeric Triarylmethane Dyes 34
8.3. Triarylmethane Dyes with Other Chromophores 35
9. Acridine Dyes 35
10. Xanthene Dyes 35
10.1. Ring-Closed Diphenylmethane Dyes (Pyronins) 36
10.2. Ring-Closed Triphenylmethane Dyes 36
10.2.1. Monoaminoxanthene Dyes 36
10.2.1.1. Cationic Dyes 37
10.2.1.2. Fluorans 37
10.2.2. Diaminoxanthene Dyes 40
10.2.2.1. Cationic Dyes 40
10.2.2.2. Nonionic Derivatives (Lactones and Related Compounds) 42
10.2.2.3. Anionic Dyes 43
10.2.3. Hydroxyxanthene Dyes 44
11. Thioxanthene Dyes 45
12. Other Ring-Closed Triarylmethane Dyes 46
12.1. Cationic Dyes 46
12.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones) 46
13. Toxicology 47
14. References 48

1. Definition and Chromophores

1.1. Definition

Di- and triarylmethane dyes [1], [2], [3, p. 381], [4, p. 399] belong to the class of polymethine dyes [5], [6]. They can be considered branched polymethines, the latter being known as cyanines, oxonols, and merocyanines (for a definition see →Methine Dyes and Pigments, Chap. 7.4., →Methine Dyes and Pig-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_179

ments, Chap. 8., →Methine Dyes and Pigments, Chap. 10.3.). The branches are created by two aryl rings, in which the polymethine chain is incorporated, and by another R group bonded to the central methine carbon (**1**).

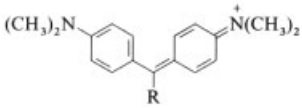

1

The R group possesses π-electrons or lone pairs of electrons that can interact with the rest of the π-electron system.

The most important electron donor is the amino group. *Triarylmethane dyes* are usually divided into mono-, di-, and triaminotriarylmethane dyes. The triphenylmethane dyes with hydroxyl groups, which form metal complexes, are not classified according to this system. Here the amino-group-containing compounds are assigned to the hydroxytriphenylmethane dyes, since the character of these dyes is determined by the complex-forming hydroxyl groups.

The donor group can also be part of a heterocyclic ring, as in the diphenylindolylmethane dyes. In some di- and triarylmethane dyes, the ring carbon atoms *ortho* to the central methine carbon are bonded via a heteroatom to form a heterocyclic six-membered ring. These include the acridine, xanthene, and thioxanthene dyes.

The suggestion has been made [7] that di- and triarylmethane dyes be called di- and triarylmethine dyes, because the central carbon atom is sp^2 hybridized and not sp^3. Despite the advantage of systematic nomenclature, the old trivial names are used here.

1.2. Chromophores

The diphenylmethyl and triphenylmethyl cations can be considered the basic chromophores of di- and triarylmethane dyes [1], [2], [7], [8]. However, the electronic state of the donor-substituted di- and triarylmethane dyes can be described better by breaking down the chromophore of these dyes into a straight- or branched-chain polymethine subchromophore, respectively, and two ethylene units (from the formal cleavage of the benzene rings) [2]. This model allows the similarities between the di- and triarylmethane dyes and the polymethines to be recognized.

The introduction of two dimethylamino groups into the diphenylmethyl cation gives Michler's hydrol blue (**1**, R = H). This compound shows a strong bathochromic shift and an increase in intensity of the absorption band at the longest wavelength compared with the diphenylmethyl cation [**1**, R = H: λ_{max} (log ε) = 607.5 nm (5.17)]. Donor substituents (R) in diarylmethane dye **1** cause a blue shift of the longest-wavelength absorption band, whereas acceptor groups generally cause a red shift [2].

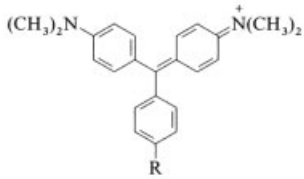

2

Malachite green (**2**, R = H) has two absorption bands in the visible spectrum at λ_{max} (log ε) = 621 (5.02) and 427.5 nm (4.30). The intense absorption band at longest wavelength, known as the *x* band, is shifted toward the red compared with Michler's hydrol blue. The second band (*y* band) is in the shorter-wavelength part of the spectrum and is significantly weaker. The *x* band is polarized normal to the plane of symmetry of the molecule, and the *y* band along the dyadic axis [1], [2], [7].

The appearance of the *y* band can be explained by using the Hückel molecular orbital (HMO) theory or the Pariser – Parr – Pople (PPP –MO) method [8]. The *x* band in malachite green arises basically from a transition from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO), and the *y* band from a transition from the next highest occupied molecular orbital (NHOMO) to the LUMO [2]. Figure 1 shows the frontier orbitals and the NHOMO of malachite green with the HMO coefficients, $c_{i,\mu}$, the squares of which correlate with the probability μ of finding the electrons around a particular atom. Positive HMO coefficients $c_{i,\mu}$ are indicated in Figure 1 by filled circles and negative ones by empty circles. The sizes of the circles represent approximately the absolute values of the coefficients.

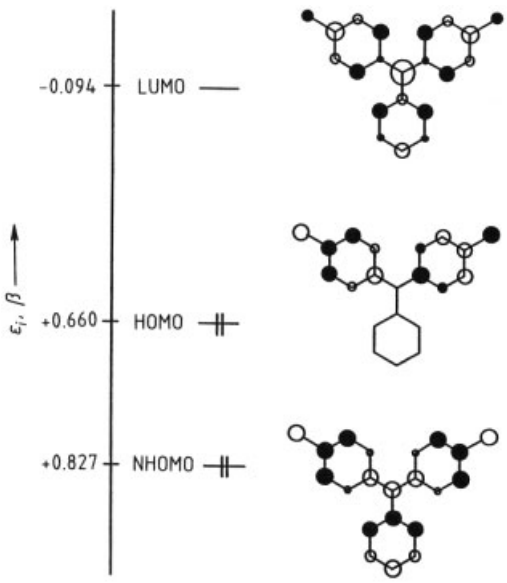

Figure 1. Frontier orbitals and NHOMO with HMO coefficients $c_{i,\mu}$ and HMO orbital energies ε_i ($\alpha_N = \alpha_C + 1.5\,\beta$, $\beta_{CN} = \beta_{CC}$, where α denotes the Coulomb integral and β the resonance integral)

Replacement of one dimethylamino group in malachite green by a methoxy group brings about energetic stabilization of the lone electron pair interacting with the conjugated π-system and thus a lowering of the energy ε_i of the molecular orbital *i* because of the higher electronegativity of oxygen compared with nitrogen. The coefficients c_i in the position of the substituent are larger in the HOMO and NHOMO than in the LUMO. Thus, the HOMO and NHOMO energies are lowered to a greater extent than the LUMO energy, and the HOMO – LUMO and NHOMO – LUMO energy differences are increased if the dimethylamino group is replaced by a methoxy group. This results in a hypsochromic shift of the *x* and *y* bands. Replacement of the dimethylamino by a diethylamino group results in a bathochromic shift and an increase in intensity of the longest-wavelength absorption band [λ_{max} (log ε) = 629.5 nm (5.08)] because of the more pronounced donor character of the diethylamino group [9].

The effect of cyclic amino groups [9] and other substituents (e.g., methyl [10]) in the amino-substituted phenyl rings on the absorption behavior of di- and triarylmethane dyes has been investigated. Bulky substituents in the *ortho*-position cause steric hindrance, which leads to rotation around the central C – C bond. Substituents in the *meta*-position cause the amino groups to rotate out of the coplanar arrangement. The simple HMO model can no longer be reliably used here [2]. Introduction of a group into the *ortho*-position of the third – the "meso"-phenyl ring in malachite green – causes rotation of the ring around the central C – C bond. This generally leads to an increase in intensity of the *x* band and a decrease in intensity of the *y* band (i.e., a color shift from green to blue [11], [12]). The same effect is observed when the meso-phenyl ring is replaced by a 1-naphthyl group [13].

Substituents in the *meta*- and *para*-positions of the meso-phenyl ring shift the position of the absorption band at longest wavelength, depending on their Hammett constants [11], [12]. If a dimethylamino group is introduced into the *para*-position of the meso-phenyl ring, crystal violet is obtained [(**2**, R = $N(CH_3)_2$], instead of malachite green. The absorption spectrum of the former has a hypsochromically shifted *x* and [λ_{max} (log ε) = 589 nm (5.06)] and no *y* band [9]. Figure 1 shows that in the HOMO, the coefficients in the meso-ring disappear. In **2**, the substituent R therefore affects only the LUMO and NHOMO. If a donor substituent R = $N(CH_3)_2$ is introduced into the phenyl ring of malachite green, the LUMO is destabilized; i.e., the energy is raised, so the HOMO – LUMO difference is increased in crystal violet compared with malachite green. The energy of the NHOMO is raised to such an extent that in crystal violet it becomes the degenerate HOMO and the *y* band thus disappears.

The *hydroxytriarylmethane dyes*, which are isoelectronic with cationic triarylmethane dyes, absorb at a shorter wavelength than the latter. The longest-wavelength absorption band of benzaurin (**3**) is at λ_{max}=585 nm compared with λ_{max}=621 nm in malachite green [2].

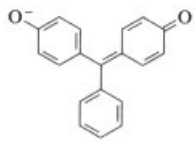

3

By bonding two phenyl rings via a *heteroatom bridge* in the 2- and 2′-positions, the yellow acridine dyes (**4a**), the red to violet xanthene dyes (**4b**), or similarly colored thioxanthene dyes (**4c**) in the case of amino-substituted di- or triarylmethane dyes are obtained. The donor effect of the occupied *p*-orbital of the heteroatom on the 2- and 2′-positions in malachite green increases the energy of the LUMO more than that of the HOMO. This results in a larger HOMO–LUMO energy difference and thus a hypsochromic shift of the absorption band at longest wavelength, which is most pronounced in the case of acridine dyes. Because of their rigid molecular skeletons these compounds fluoresce. Hydroxyxanthenes behave similarly (e.g., fluorescein) [2].

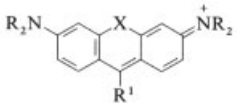

4a: X = NR² Acridines
4b: X = O Xanthenes
4c: X = S Thioxanthenes

The direct C–C bonding of the 2- and 2′-positions of malachite green to give the fluorenyl cation (**5**) destablizes the HOMO because of the antibonding interaction of the two centers, and stabilizes the LUMO, which experiences a bonding interaction [14].

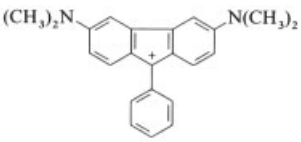

5

The electronic spectrum of **5** [λ_{max} (log ε) = 955 (4.28), 850 (4.20), 505 (4.32), 470 nm (4.30)] [8] differs drastically from that of malachite green [λ_{max} (log ε) = 621 (5.02), 427.5 nm (4.30)]. The absorption band of **5** at the longest wavelength is shifted strongly to the red and is significantly weaker in intensity compared to malachite green. The fluorenyl cation (**5**) can be considered a chromophore similar to a diradical (antiaromatic) because its frontier orbitals are almost degenerate [15].

On protonation of crystal violet, the color changes first to green and then to yellow. The green coloration is caused by protonation of one of the dimethylamino groups and can be explained by the appearance of the *y* band. The protonation of another amino group leads to the color change to yellow. Amino groups whose conjugation with the ring system is reduced for steric reasons are readily protonated so the intensity of the absorption bands is lowered in comparison with systems having unhindered conjugation because of the proportion of protonated species. In an alkaline medium, decoloration occurs because of formation of triphenylmethanol (**6**) [2].

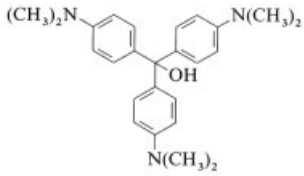

6

Color formers are very sensitive to the addition of acids and solvents. In acid, crystal violet lactone (**7**) is only partially converted to the colored carboxylic acid (**8**) with opening of the lactone ring [λ_{max} (log ε) = 603 nm (4.41) in methanol – acetic acid (1 : 1)]

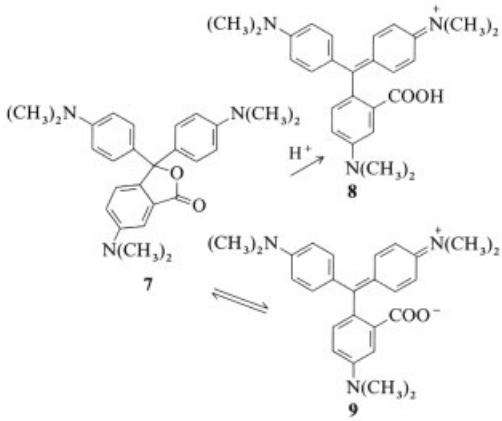

The acid is in equilibrium with the colorless lactones protonated on the nitrogen. In very polar solvents such as trifluoroethanol, hexafluoroisopropanol, or phenol, the lactone ring opens to give the zwitterion **9**. The longest-wavelength band of **9** [λ_{max} (log ε) = 593 nm (4.91) in hexafluoroisopropanol] is shifted hypsochromically compared with that of **8**. As in

the case of rhodamine dyes [16], this reaction is a solvent-dependent intramolecular Lewis acid – base equilibrium. The high intensity of the band shows that the equilibrium is shifted strongly to the zwitterionic compound. The electronic spectrum of *fluoran* (**10**), a black color former, has two bands of the same intensity in the visible region [λ_{max} (log ε) = 589 (4.24), 449 nm (4.23) in acetic acid; λ_{max} (log ε) = 554 (4.28), 448 nm (4.26) in 2,2,2-trifluoroethanol] [17], [18]. Here, the equilibrium lies completely on the side of the ring-opened compound. This is verified by comparison with the electronic spectrum of the black ester **11** [λ_{max} (log ε) = 584 (4.26), 447 nm (4.21) in acetonitrile] [17].

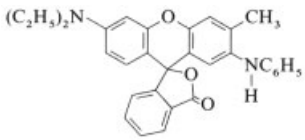

10

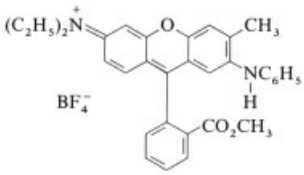

11

2. History and Economic Importance

2.1. History

The triphenylmethane dyes are among the oldest industrially product dyes. At first they were prepared by methods discovered empirically, without knowledge of their constitution. The first industrial production process for fuchsin was developed by VERGUIN in 1859. In 1861, LAUTH synthesized methyl violet by oxidizing dimethylaniline. Fuchsin was converted into blue dyes with aniline by NICHOLSON in 1862.

In 1876, MICHLER successfully reacted phosgene with dimethylaniline to give 4,4′-bis(dimethylamino)benzophenone. This compound, known as Michler's ketone, has been a starting material for many important triphenylmethane dyes [19]. In 1883, CARO and KERN produced crystal violet from phosgene and dimethylaniline. Malachite green was obtained in 1877 by O. FISCHER from benzaldehyde and dimethylaniline by a process still used industrially today [20]. SANDMEYER and SCHMID isolated a chloro-substituted product of malachite green in 1896. All these dyes are still economically important.

In 1953, ZOLLINGER, WEGMANN, and RHYNER found that Michler's hydrol [21] and mono-, di-, and triaminotriphenylmethane and xanthene dyes [22] are all surprisingly lightfast on polyacrylonitrile fibers. In 1971, KAST, BAUMANN, and MAYER developed a process for the catalytic oxidation of leuco triarylmethane dyes with atmospheric oxygen [23]. The use of crystal violet lactone for the production of self-duplicating paper has been known since 1963 [24].

Acid triphenylmethane dyes were initially produced by sulfonation of basic (cationic) dyes. Aniline blue, which is soluble in alcohol, was thus converted by NICHOLSON in 1862 into a form suitable for dyeing. Acid triphenylmethane dyes based on benzyl-substituted aniline were found to be particularly effective for dyeing wool. Introduction of a sulfonic acid group *ortho* to the methine carbon was a considerable advance in efforts to improve alkali fastness. The first dye of this type was patent blue V, which was obtained by HERMANN in 1888 by treating 3-hydroxybenzaldehyde with diethylaniline, followed by sulfonation.

The first *hydroxytriphenylmethane dye*, rosolic acid, was synthesized in 1834 by RUNGE. Dyes with practical applications, however, were produced only after carboxyl groups *ortho* to the hydroxyl group were introduced by SANDMEYER in 1889. The condensation of *ortho*-substituted benzaldehydes with 2-hydroxy-3-methylbenzoic acid gave dyes that could be converted into metal complexes on wool by aftertreatment with chromium compounds (CONZETTI, 1906). They give clear washfast blue dyes that are still used today.

In 1871, VON BAEYER prepared *fluorescein* by condensation of phthalic anhydride with resorcinol, from which H. CARO, together with VON BAEYER, produced eosin by bromination. The first *rhodamine dyes* were described by BASF in 1887 and the first *sulforhodamines* by STEINER in 1906.

Xanthene lactones have achieved considerable economic importance as chromogenic substances. The first fluorans, which are suitable for self-duplicating paper, were produced by Fuji in 1967. They were used to produce orange and green colorations. A fluoran-based uniform black dye was synthesized in 1969 by NCR and is still important today.

2.2. Economic Importance

The diphenyl- and triphenylmethane and xanthene dyes have achieved considerable economic importance because of their color strength and brilliance. In 1993, ca. 9000 t of *basic di- and triphenylmethane dyes* and ca. 4000 t of *xanthene dyes* were sold. Crystal violet, methyl violet, malachite green, auramine, and rhodamine B are suitable for many purposes and belong to the economically most important dyes.

About 2500 t of *acid diaminotriphenylmethane dyes* were sold in 1993, 1500 t of which were patent blue AE. Of the *sulfonated triaminotriphenylmethane dyes*, whose sales are estimated to be 4000 t, the alkali blue and reflex blue dyes with betaine character have achieved a market volume of ca. 3000 t. They are used mainly in the printing industry to brighten carbon black.

The demand for *chrome dyes* for wool is slightly less than 1000 t.

The demand for *color formers* has increased significantly since the late 1980s. In 1993, ca. 8000 t of color formers were used, half for producing self-duplicating paper and half for producing thermosensitive paper for fax machines. Crystal violet lactone and fluorans are particularly important here.

3. Production

Condensation Reactions. Many triaminotriarylmethane dyes can be produced by condensation of Michler's ketone (**12**) with aromatic amines. The ketone must generally be activated with phosgene or phosphorus oxychloride, whereby a reactive blue chloro compound (**13**) is formed from the intermediate geminal dichloride. With *N*-phenyl-1-naphthylamine, for example, **13** reacts to form Victoria blue B (**14**):

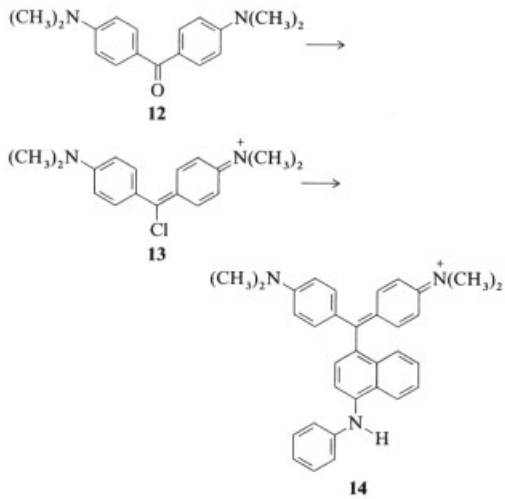

Michler's hydrol (**15**) is even more reactive. It can be obtained by reduction of Michler's ketone with zinc dust [25] or sodium borohydride [26]. Condensation of the hydrol with 3-dimethylaminobenzoic acid (**16**) gives leuco crystal violet carboxylic acid (**17**). The latter is then oxidized to crystal violet lactone.

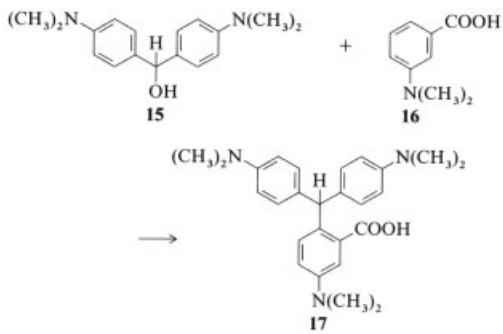

A leuco compound is also obtained when a benzaldehyde derivative is treated with 2 mol equivalents of an aniline derivative. Leuco malachite green (**20**) is thus formed from benzaldehyde (**18**) and dimethylaniline (**19**):

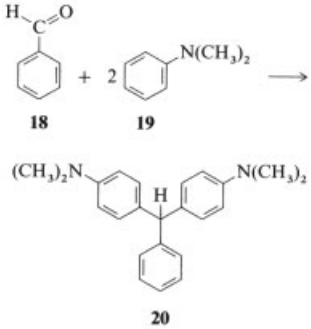

With benzaldehyde-2-sulfonic and benzaldehyde-2,4-disulfonic acids, important acid dyes are synthesized via this reaction. They are known as patent blue dyes.

Crystal violet (**23**) can be produced directly from formaldehyde and dimethylaniline. The methane base (**21**), *N,N,N′,N′*-tetramethyl-4,4′-methylenedianiline, is formed initially. This is then oxidized to Michler's hydrol (**15**), which condenses with another molecule of dimethylaniline to give leuco crystal violet (**22**). The latter is converted into the dye in a second oxidation step. Pararosaniline, methyl violet and Victoria blue can also be obtained by this reaction sequence:

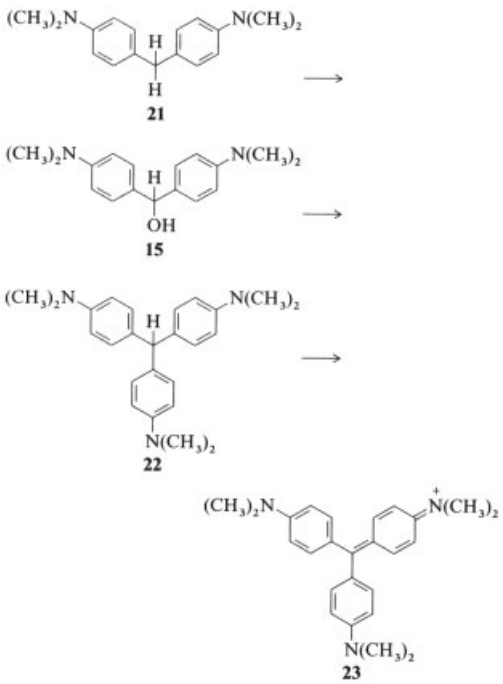

In the past, alkali dichromates and lead dioxide were used mainly for the oxidation. Because of the toxicity of heavy metals, manganese dioxide has been used instead for some processes [27–29]. The reduction product formed can be recycled by reoxidation to manganese dioxide.

Novel processes use catalytic oxidation with atmospheric oxygen or hydrogen peroxide. They are economical and environmentally acceptable. Both the oxidation of leuco malachite green [30] and the oxidative production of crystal violet [31] from formaldehyde and dimethylaniline can be carried out in glacial acetic acid in the presence of catalytic quantities of 2,3,5,6-tetrachloro-1,4-benzoquinone (chloranil) and (5,14-dihydrodibenzo[*b,i*][1, 4,8,11]tetraazacyclotetradecinato)iron(II) [(tetraaza[14]annulene)iron] using atmospheric oxygen. Vanadium and molybdenum catalysts have also been recommended for the oxidation to give triphenylmethane dyes [32]. Oxidation with atmospheric oxygen and nitrous gases in the presence of chloranil has the advantage of a homogeneous reaction medium [33–35]. A mixture of copper sulfate, sodium chlorate, and chloranil has also been used for oxidizing leuco compounds [36]. Oxidation to acid triphenylmethane dyes can be carried out by using hydrogen peroxide in the presence of (tetraaza[14]annulene)iron [37] or another transition-metal catalyst [38].

Nucleophilic substitution of chlorine, sulfonic acid, or amino groups by aromatic amines is a common process for producing triaminotriarylmethane dyes. Spirit blue (**26**) can be formed from the trichlorotriphenylmethyl cation (**24**) [39], [40] and from pararosaniline (**25**) [41], [42] by reaction with aniline:

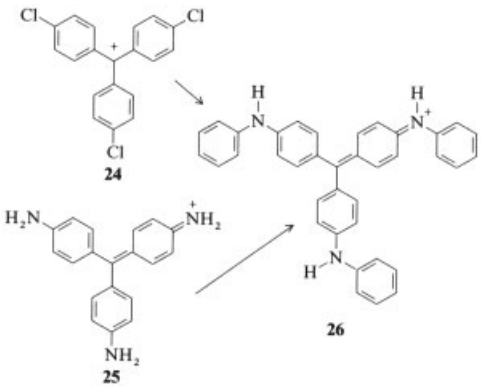

In the production of C.I. Acid Blue 90 a *p*-sulfonyl group is replaced by *p*-phenetidine [43].

Production of Xanthene Dyes. The most important starting material for xanthene dyes is phthalic anhydride. It reacts with 3-diethylaminophenol in the molten state to give rhodamine B, and with resorcinol to give fluorescein. For unsymmetrical xanthene dyes, dihydroxybenzoyl- (**27**) or an aminohydroxybenzoylbenzoic acid (**28**) is required:

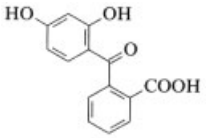

27

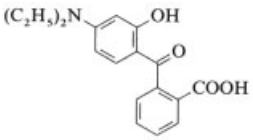

28

These compounds can condense and cyclize with resorcinol or substituted 3-aminophenols. Nucleophilic substitution of chlorine by amines leads to xanthene dyes, which contain a sulfonic acid group [44] or hydrogen [45] instead of the carboxyl group.

In the acidic condensation of (4-diethylamino-2-hydroxy)benzoylbenzoic acid (**28**) with *para*-substituted phenols or phenol ethers, attack occurs at the *ortho* position. The phthalide formed (**29**) cyclizes readily in alkali to form a fluoran (**30**). A black color former is obtained in good yield by reacting **28** with 4-methoxy-3-methyldiphenylamine. The cyclization involves nucleophilic substitution with loss of methanol [46]:

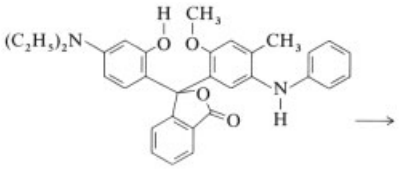

29

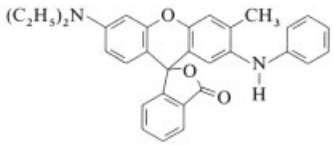

30

4. Uses

Basic triphenylmethane dyes are used for dyeing paper, coconut fibers, jute, and other lignin-containing substrates. In the office equipment industry, they are used in the production of ballpoint pastes, fountain pen inks, stamping inks, and hectographic and carbon copy paper. They are also used to dye detergents, seed dressers, plant protection agents, fertilizers, and antifreezes.

An important application for basic dyes and those with betaine character is brightening carbon black for printing inks. By incorporating the dyes into carbon black, a deep black can be produced from the otherwise brownish color. By precipitation with heteropolyacids based on phosphorus or silicon with molybdenum, tungsten, or both, brilliant Fanal pigments (trade mark of BASF) are produced for the printing industry.

Xanthene dyes are used for dyeing wood, paper, fountain pen inks, and cosmetic articles. Lightfast, brilliant red Fanal pigments can be produced from basic rhodamine dyes and heteropolyacids.

Crystal violet lactone is used in the production of self-duplicating paper. *Fluorans* are also used for this purpose and for heat-sensitive paper for fax machines.

5. Diphenylmethane Dyes

Among the diphenylmethane dyes are the salts of substituted bis(*p*-aminophenyl)methanols and diaminobenzophenoneimines. The most important representatives are Michler's hydrol and auramine.

5.1. Michler's Hydrol and Derivatives

Michler's hydrol [*119-58-4*], bis(4-dimethylaminophenyl)methanol (**15**), is produced by oxidation of bis(4-dimethylaminophenyl)methane (*N,N,N′,N′*-tetramethyl-4,4′-methylenedianiline) [*101-61-1*] with lead dioxide. A patent describes the oxidation with

hydrogen peroxide in the presence of the enzyme peroxidase [47]. Michler's hydrol can also be obtained by reduction of Michler's ketone [*90-94-8*] using zinc dust [25] or sodium borohydride [26], [48].

In the presence of acid, the hydroxyl group can be eliminated from Michler's hydrol giving *Michler's hydrol blue* [*14844-71-4*] (**31**) – a deep blue salt that is readily destructed by hydrolysis. It can be stabilized by complex anions such as tetrachlorozincate or phosphotungstomolybdate [49].

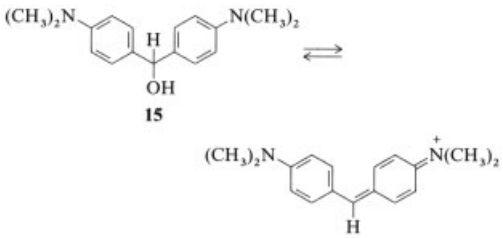

15 **31**

Michler's hydrol and related compounds are intermediates in the production of special triphenylmethane dyes.

Derivatives formed by reaction of Michler's hydrol with sulfinic acids or aromatic amines have been recommended as chromogenic compounds for the production of carbon-free copy paper. *Bis(4-dimethylaminophenyl)methyl(4-tolyl) sulfone* (**32**) has achieved industrial importance

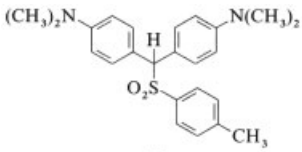

32

The compound is obtained by treating Michler's hydrol with sodium 4-toluenesulfinate in glacial acetic acid followed by precipitation with ethanol [50]. Bis(4-dimethylaminophenyl)methyl(4-tolyl) sulfone (**32**) is produced best immediately following the synthesis of Michler's hydrol. For example, Michler's ketone can be reduced with sodium borohydride and then reacted with sodium 4-toluenesulfinate [26]. Alternatively, bis(4-dimethylaminophenyl)methane (*N,N,N′,N′*-tetramethyl-4,4′-methylenedianiline) can be oxidized with atmospheric oxygen in glacial acetic acid in the presence of chloranil and salcomine [*N,N′*-bis-(salicylidene)ethylenediimino cobalt(II)] as catalysts, and the product subsequently precipitated with sodium 4-toluenesulfinate [51]. Michler's hydrol can also be produced from 4-dimethylaminobenzaldehyde and *N,N*-dimethylaniline in the presence of urea. Bis-(4-dimethylaminophenyl)methyl(4-tolyl) sulfone is then obtained by addition of 4-toluenesulfinate [52].

The reaction products of Michler's hydrol with thiophenols [53], azides [54], oximes [55], [56], esters and amides of α-hydroxycarboxylic acids [57], and compounds with active methylene groups [58] are chromogenic compounds. They are colorless but turn deep blue when treated with acidic inorganic substances (e.g., kaolin or bentonite).

Production Example. A mixture of 508 g bis(4-dimethylaminophenyl)methane, 10 g chloranil, 20 g salcomine, 1900 g methanol, and 140 g glacial acetic acid is treated with oxygen at 50 °C in a closed vessel. After 32 g oxygen has been absorbed, 193 g sodium 4-toluenesulfinate dihydrate is added. The suspension is then heated under reflux for 2 h. After cooling, it is filtered; the residue is washed with water, dried, and recrystallized from benzene. Yield: 550 g, *mp* 184 – 185 °C.

Derivatives of leuco auramine (**33**) are obtained by condensation of Michler's hydrol with aromatic amines

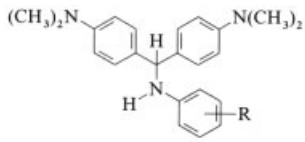

33

These compounds also give a deep blue coloration with acidic inorganic substances and are suitable for the production of colorless copier paper. The aromatic amines can contain nonionic substituents – for example, halogens [59], alkyl groups [60], hydroxyalkyl groups [61], and carboxylate ester groups [62]. They can also contain anionic groups [63], [64] and are then suitable for use in carbon paper in alco-

hol duplicating processes. The most important anionic groups are the carboxyl [65] and sulfonic acid [66] groups. Cationic groups, such as the trimethylammonium group, have also been described for this application.

Dimeric leuco auramine derivatives are obtained on reacting Michler's hydrol with hexamethylenediamine [67] or 3,3′-dimethoxybenzidine [68], for example.

Aminothiazole and aminobenzothiazole [69] are suitable heterocyclic amines. Morpholine, piperidine, or piperazine [70], and aminoalkylcarboxylic esters [71] can also be reacted successfully with Michler's hydrol. With morpholine and 2,5-dichloroaniline, leuco auramines are obtained, which have been used successfully for transfer printing on polyacrylonitrile materials [72].

5.2. Michler's Ketone, Auramines, and Derivatives

5.2.1. Michler's Ketone

Michler's ketone [*90-94-8*], 4,4′-bis (dimethylamino)benzophenone (**34**), is an important intermediate in the production of triphenylmethane dyes. It is obtained by reaction of dimethylaniline with phosgene. A second synthetic route involves the catalytic oxidation of bis(dimethylaminophenyl)methane [73].

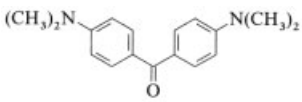

34

Production Example. A mixture of 508 g bis(4-dimethylaminophenyl)methane, 10 g chloranil, and 20 g salcomine is treated with oxygen in 1970 g methanol at 50 °C with vigorous stirring in a closed vessel. After 64 g oxygen has been absorbed, the reaction mixture is allowed to cool, and the precipitated Michler's ketone is collected by suction filtration, washed with methanol, and dried. Yield: 459 g, *mp* 172 – 174 °C.

Uses. A reactive intermediate for the production of triphenylmethane dyes is formed by treating Michler's ketone with thionyl chloride [74]. A large number of investigations have been carried out on the use of Michler's ketone as a photosensitizer [75], [76].

5.2.2. Auramines

Auramine O [*2465-27-2*], bis(4-dimethylaminophenyl)methyleneiminium chloride (**35**) (C.I. Basic Yellow 2, C.I. 41000) and its ethyl homologue, ethylauramine, are brilliant yellow dyes with high color strength. They have worldwide annual sales of ca. 1000 t.

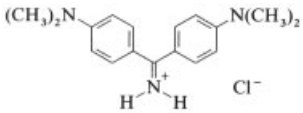

35

Auramine O can be produced from Michler's ketone and ammonium chloride in the presence of zinc chloride. However, the industrial process starts from bis(4-dimethylaminophenyl)-methane, which is heated to 200 °C with ammonium chloride, sulfur, and sodium chloride. 4,4′-Bis(dimethylaminophenyl)thiobenzophenone is formed as an intermediate; this is readily converted into the imine auramine O with ammonia [77].

Ethylauramine [*6358-36-7*], (C.I. Basic Yellow 37, C.I. 41001), the corresponding *N,N,N′,N′*-tetraethyl compound, is produced similarly.

Both dyes are obtained in particularly pure form if the crude melt is converted into the dye base in an organic solvent and the dye salt is then extracted with water after addition of acid [78].

Production Example. A mixture of 62 g bis(4-dimethylaminophenyl)methane, 120 g urea, 23 g sulfur, and 40 g ammonium sulfate is stirred for 3 h at 155 °C, for 1 h at 165 °C, and for 1 h at 175 °C. The melt is transferred to a prepared mixture of 1 L water at 60 °C, 75 g 50 % sodium hydroxide solution, and 250 g trichloroethylene. The mixture is stirred for 15 min. After phase separation, 1 L water is added to the trichloroethylene phase at 70 °C. The pH is adjusted to 5 with dilute sulfuric acid,

and the mixture is stirred for another 15 min. The trichloroethylene layer is separated, and the dye is precipitated from the filtered aqueous solution with 60 g sodium chloride. After cooling, the dye is collected by suction filtration and dried. Ca. 73.2 g auramine hydrochloride is obtained with a purity of 96 %.

Auramine G (**36**) (C.I. Basic Yellow 3, C.I. 41005) is derived from bis(3-methyl-4-methylaminophenyl)methane. It can be prepared according to the process used for auramine O. However, the dye is no longer produced.

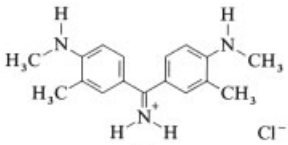

36

Uses. Auramine dyes are used for dyeing of leather, jute, tanned cotton, and paints, and as dye components in inking ribbons, ballpoint pastes, and carbon paper. The most important areas of application are in dyeing paper and in flexographic printing. For the latter, a large number of salts are produced, which have a high solubility in ethanol and ethanol – water mixtures (e.g., nitrates, nitrites [79], bromides, iodides [80], salts of alkyl-, aralkyl-, or arylsulfonic acids [81], thiocyanates [82], and phosphates [83]). For dyeing paper, solutions of the hydrochloride in organic solvents (e.g., thiodiglycol [84]) are employed. Tartaric and citric acids have been used as stabilizers here [85].

The chlorides can also be converted into more easily soluble acetates [86] or propionates [87] to obtain highly concentrated solutions. The hydrogensulfates of auramine are also readily soluble [88].

Colored salts based on auramine with tannic acid, phosphomolybdic acid, and preferably phosphotungstomolybdic acid can be used as yellow toner pigments for the development of latent electrostatic images [89]. *Hydroxyethylauramine dyes*, such as the bis(2-hydroxyethylamine) compound, have been used to dye paper [90].

The solubility of auramine dyes in linseed or mineral oil is increased by substitution at the imino nitrogen. These modified dyes are obtained by reaction of the dye base of ethylauramine with an aliphatic amine in benzene. The reaction product of ethylauramine with butylamine can be used in printing inks [91]. Auramines with aromatic substituents on the imino nitrogen can be produced from Michler's ketone and the anion of an aniline derivative, which can be obtained by reaction with ethylmagnesium bromide and subsequent loss of ethane [92].

6. Triphenylmethane Dyes

6.1. Monoaminotriphenylmethane Dyes

6.1.1. Cationic Dyes

The simplest monoaminotriphenylmethane dye (**37**) is yellow and has good lightfastness on polyacrylonitrile fibers [22]. It is obtained by treating 4-chlorobenzotrichloride with benzene in the presence of aluminum trichloride, and subsequent nucleophilic substitution of chlorine by an amino group.

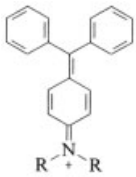

37

Monoaminotriphenylmethane dyes containing additional donor substituents exhibit a bathochromic shift giving red shades. They can be obtained by reacting alkyl- and alkoxy-substituted benzophenones with aromatic amines [93].

Aminoalkoxytriphenylmethane dyes with carboxyl-containing substituents are used in photography as antihalo dyes [94]. Triarylmethane dyes, which contain a 4-oxo-1-naphthylidene group besides an aminophenyl group, have also been described as antihalo dyes. They contain an acylated group *ortho* to the central carbon atom, which is hydrolized by alkali. A nucleophile is thus liberated, which attacks the dye intramolecularly and irreversibly decolorizes it. Suitable nucleophilic groups include oxymethylene [95], aminomethylene [96], and

in particular, the SO_2NHCH_3 group (**38**, **39**) [97], [98].

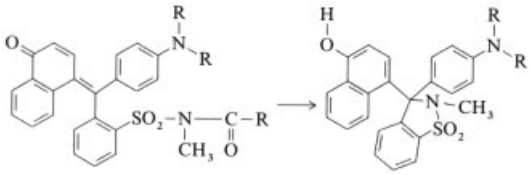

38 **39**

6.1.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones)

Substituted aminohydroxytriphenylmethanols and their ethers can be used for transfer printing on textile fiber materials containing anionic groups [99] and as color formers for pressure- and heat-sensitive recording paper [100]. Lactones of monoaminotriphenylmethane dyes are yellow to orange color formers [101].

Extensive kinetic studies were carried out on the addition of hydroxyl [102–104], cyanide [105], [106], and azido groups [107] to the central atom of variously substituted triarylmethane dyes with variation of the temperature, the ionic strength, and dielectric constant (by mixing water with aprotic solvents). These studies gave values for the activation enthalpy and information on the transition state and its agreement with a purely electrostatic theoretical model. It was also found that the rate constants for the reaction with hydroxyl ions correlate with the equilibrium constants [108].

Color formers, which give red-violet dyes on loss of a hydroxyl or alkoxy group or a nitrogen-containing base, are obtained by condensing 4,4′-dimethoxybenzophenone with *N*-methyldiphenylamine. Depending on whether a stoichiometric quantity or an excess of ketone is used, the 4,4′-dimethoxy-4″-*N*-methylanilinotriphenylmethyl cation [109], [110] or the dimeric dication containing one *N*-methyldiphenylamine moiety as a bridge is obtained [111]. These dyes can be converted with sodium hydroxide [109], [111], imidazole, or triazole [110] to colorless compounds (**40**, **41**) that can be used as color formers.

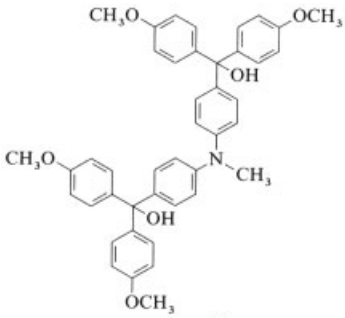

40

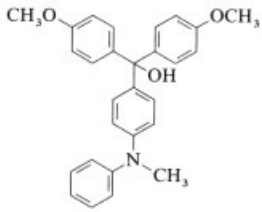

41

If an *N*-methyldiphenylamine is used, which has an acceptor (e.g., a cyano group) in one *para*-position, a dark violet color former is produced with sodium methoxide [112].

6.2. Diaminotriphenylmethane Dyes

Diaminotriphenylmethane dyes can be divided into cationic dyes; triphenylmethanol derivatives and lactones; and anionic dyes. The latter assume a zwitterionic character if they contain only one sulfonic acid group.

6.2.1. Cationic Dyes

Doebner's violet (**42**) is the simplest diaminotriphenylmethane dye.

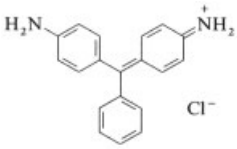

42

This dye is obtained by condensing benzotrichloride with 2 mol equivalents of chlorobenzene in the presence of aluminum trichloride and replacing the *p*-chlorine atoms with amino groups [113].

Malachite Green. A dye that is still very important is malachite green [*569-64-2*] (**43**) (C.I. Basic Green 4, C.I. 42000; oxalate [*2437-29-8*]; zinc chloride double salt [*79118-82-4*]; derivatives: C.I. Solvent Green 1, C.I. Pigment Green 4).

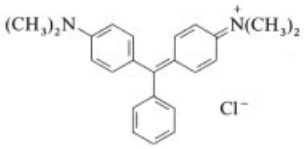

43

Malachite green is synthesized by condensation of benzaldehyde with dimethylaniline in sulfuric acid and oxidation of the leuco compound in hydrochloric acid solution with lead dioxide [82], [114] or manganese dioxide [28]. Alternatively, especially for production of liquid dyes, benzaldehyde is condensed with dimethylaniline in acetic acid to give a leuco base, which is then oxidized catalytically with atmospheric oxygen in the presence of (dihydrodibenzotetraaza[14]-annulene)iron and chloranil [30].

Production Example 1. Condensation [3, pp. 381, 389]: A mixture of 930 kg dimethylaniline, 345 kg benzaldehyde, 270 kg 78 % sulfuric acid, and 90 L water is heated to 112 °C under reflux with stirring for 12 h. After addition of a solution of 215 kg anhydrous sodium carbonate in 230 kg water, excess dimethylaniline is removed by steam distillation. The leuco base is separated.

Oxidation [28]: Leuco malachite green (90.3 g) is dissolved in a mixture of 86.5 g concentrated hydrochloric acid and 100 mL water; this solution is then diluted with 500 mL water. After cooling to 0 °C, 91.5 g of a 34 wt % aqueous suspension of manganese dioxide is added. During oxidation the temperature rises to 15 °C. After 15 min, 140 g potassium carbonate and 100 mL toluene are added. The mixture is heated to 55 – 60 °C and filtered with suction over a frit; the frit is washed with 40 mL toluene. The aqueous phase is removed, and the toluene phase is concentrated on a rotary evaporator at 2 kPa and 45 °C. A solution of 33.5 g anhydrous oxalic acid in 150 mL water is added, and the mixture is stirred for 30 min at 45 °C. The residual toluene and water are distilled off at 2 kPa and 45 °C to constant weight. Approximately 121.5 g of ca. 91 % malachite green oxalate (87 % yield) is obtained in the rotary evaporator flask as green crystals, *mp* 137 – 139 °C, Mn content: 1 mg/kg.

Production Example 2 [30]. A mixture of 18.04 kg glacial acetic acid, 7.26 kg dimethylaniline, 0.9 kg urea, and 3.3 kg benzaldehyde is heated under nitrogen to 60 °C for 1 h and then to 95 °C for 4 h. After cooling to 35 °C the nitrogen flow is switched off. 0.15 kg chloranil and 75 g (dihydrodibenzotetraaza[14]annulene)iron are then added. Air is then passed in at 40 °C with vigorous stirring until oxygen absorption is complete. 30 parts of a ca. 40 % solution of malachite green are obtained, yield ca. 95 %.

On further oxidation of malachite green (**43**) with lead dioxide, demethylation occurs; in addition, the quinoneiminium salt (**44**) and benzoic acid are formed [4, pp. 399, 402].

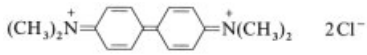

44

The chloride salt of malachite green is readily soluble in water and all polar solvents that are miscible with water. It is therefore suitable for the production of inks for flexographic printing. It cannot be isolated from aqueous solutions by salting out, but can be obtained by addition of a solvent that is immiscible with water [115]. Crystalline malachite green can be precipitated from dilute aqueous solutions as the oxalate (see production example) and as a mixed salt with carboxylic and phosphoric acids [116].

Uses. Malachite green is used to dye paper, leather, inks, and waxes; it produces deep shades on polyacrylonitrile fibers. It is employed in dye mixtures that also give deep shades on polyacrylonitrile fibers, particularly black and navy blue. Malachite green can be used as the phosphomolybdate for quantitative phosphate determination in aqueous solutions [117]. In medical diagnostics, traces of hydrogen peroxide can be detected quantitatively by using leuco malachite green in the presence of a peroxidase [118].

Trade names include Aizen Malachite Green (Hodogaya), Astrazon Green M (Bayer), and Basazol Green 14 L (BASF).

Brilliant green [*633-03-4*] (**45**) (Diamond Green BX, C.I. Basic Green 1, C.I. 42040,

derivative: C.I. Pigment Green 1) is produced from benzaldehyde and diethylaniline similarly to malachite green and is used in the same way. A Michler's ketone-free ink for inking ribbons containing brilliant green has been described [119].

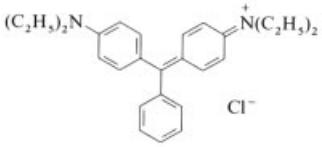

45

Trade names include Aizen Diamond Green GH (Hodogaya), Basazol Green 20 L (BASF), and Brilliant Green NSA (Zeneca).

Rhodulin blue 6 G [*3521-06-0*] (**46**) (C.I. Basic Blue 1, C.I. 42025, derivative: C.I. Pigment Blue 9) and *rhodulin blue B* [*3943-82-6*] (**47**) (C.I. Basic Blue 5, C.I. 42140, derivative: C.I. Pigment Blue 3) are produced by condensation of 2-chlorobenzaldehyde with dimethylaniline and *N*-ethyl-*o*-toluidine, respectively, followed by oxidation. They are used for dyeing polyacrylonitrile and cellulose acetate fibers.

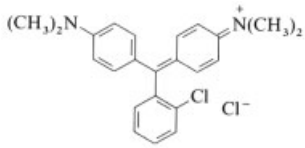

46

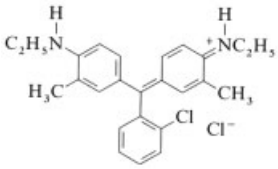

47

The nitrates of **46** and **47** are very soluble in organic solvents and are therefore suitable for the production of copier paper, stamping inks, and ballpoint pastes [120].

*Trade names.*Rhodulin blue 6 G: e.g., Aizen Brilliant Basic Cyanine 6 GH (Hodogaya), Astrazon Blue G (Bayer), and Atacryl Blue 6 G (Atlantic Industries). Rhodulin blue B: e.g., Astrazon Blue B (Bayer), Hispacril Blue 5 B (Robama), Orlamar Blue B (Amar Dye-Chem.), and Primocyanin BX (Sumitomo).

The *lightfastness* of diaminotriphenylmethane dyes on polyacrylonitrile fibers is significantly improved if the amino groups carry electron acceptor-substituted alkyl groups, such as fluoroalkyl [121], 2-cyanoethyl [122], or alkylcarboxylic acid esters [123]. Diaminotriphenylmethane dyes with methyl groups *ortho* to the central carbon atom on the phenyl rings bearing the amino groups have improved lightfastness on polyacrylonitrile fibers [124]. The lightfastness of these dyes on natural fibers is low.

Diaminotriphenylmethane dyes with phenylamino groups are obtained by treating benzotrichloride with chlorobenzene in the presence of aluminum trichloride, followed by substitution of aromatic amines for the *p*-chlorine atoms [125]. These dyes can be used for mass dyeing of polyacrylonitrile fibers. Dyes with only one phenylamino group, which are produced by condensation of 4-dialkylaminobenzophenone with diphenylamine, give blue polyacrylonitrile fibers with a green cast [126].

Diaminotriphenylmethane dyes can be used as positive charge stabilizers [127] in electrophotographic toners and as sensitizers in electrophotographic photoreceptors [128]. Diaminotriphenylmethane dyes with polyoxyalkylamino groups are used as readily washable marking inks and color markers [129–131]. Diaminotriphenylmethane dyes, which have been precipitated with phosphomolybdotungstic acid [132] (C.I. Pigment Blue 3 and 9, C.I. Pigment Green 1 and 4) or with acid metal-complex dyes [133], are used as pigments.

Leuco compounds, such as **48** [*15008-36-3*] [134], which are substituted in the aminophenyl rings *ortho* to the central carbon atom, are used as charge-transfer compounds in organic photoconductors [135], [136].

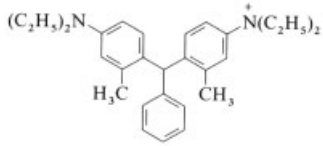

48

Leuco compounds are suitable components for electrochromic materials for displays [137] and

stabilizers for vulcanizable chloroprene rubber [138].

6.2.2. Nonionic Derivatives (Triphenylmethanol and Triphenylmethylamine Derivatives)

Lactones. Diaminotriphenylmethane dyes, which have a carboxyl group *ortho* to the central carbon atom are obtained by condensation of phthalic anhydride with aromatic amines. A lactone ring is formed on neutralization. Malachite green lactone (**49**) formed with dimethylaniline has a poor color strength when used as a color former because of the stability of the lactone. Its applications are therefore limited.

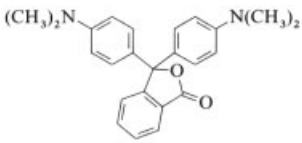

49

Higher color strengths can be achieved by incorporating additional donor groups into the molecule. Phthalic anhydride and *N,N*-diethyl-*m*-toluidine give a benzoylbenzoic acid that can be condensed with tetramethyl-*m*-phenylenediamine to give a violet-gray color former (**50**) [139].

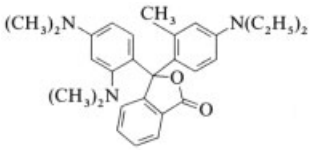

50

After treatment of 2-bromobenzaldehyde with dimethylaniline, the bromine in the leuco malachite green obtained can be replaced by diethyl butylphosphonite. After oxidation and hydrolysis, a green-blue color former (**51**) is produced in alkaline solution [140]:

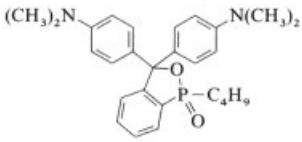

51

Benzoxazine Derivatives. Other color formers that do not contain lactone units are derived from salicylaldehyde and 4-diethylaminosalicylaldehyde. After condensation – for example, with dimethylaniline – a leuco malachite green derivative is obtained, whose *o*-hydroxyl group can react with phenyl isocyanate. The leuco dye is oxidized, and the *N*-phenylcarbamoyloxy group formed undergoes a cyclization reaction in alkaline solution. A mixture of two color formers **52** and **53** is thus obtained [141].

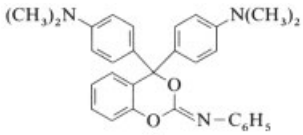

52

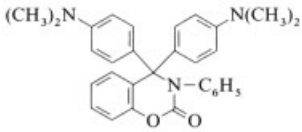

53

Another starting material, which gives a new system of color formers, is 3-dimethylaminoaniline, whose primary amino group has been reacted with ethyl chloroformate to yield ethyl-3-dimethylaminophenylcarbamate. The carbamate can be condensed with 4-diethylamino-4′-methoxybenzophenone to yield a dye that cyclizes in alkaline solution to give a 4,4-diaryl-2-oxobenzo-1,3-oxazine (**54**) [142].

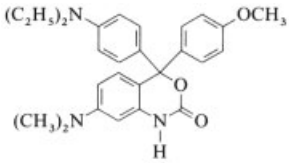

54

If 3-dimethylaminoaniline is acylated with benzoyl chloride and the 3-benzoylaminodimethylaniline is treated with the same benzophenone, a 4,4-diaryl-2-phenyl-1,3-benzoxazine (**55**) with X and Y = H is obtained [143].

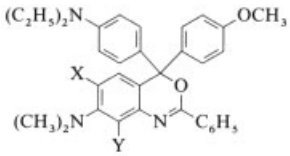

55

Color formers that only differ from **55** in that they are derived from a toluylene mixture are of particular interest. A mixture of 90 % X = methyl and Y = H and 10 % X = H and Y = methyl gives a black color former [144] that, in combination with fluorans [145] and bisindolylphthalides or with benzoxazines derived from monoaminotriphenylmethane dyes [146], shows improved intensity and lightfastness.

Quinazoline Derivatives. The *4,4-diaryltetrahydroquinazolones* are also derived from 3-dimethylaminoaniline. The synthesis gives blue color formers in the triaminotriphenylmethane series. After addition of the primary amino group of 3-dimethylaminoaniline to an alkyl or acyl isocyanate, a urea derivative is formed that is converted into a leuco dye by reaction with Michler's hydrol. After oxidation the leuco dye is cyclized to a color former in alkali. The 4,4-diaryltetrahydroquinazolones (**56**) formed have either an alkyl [147] or an acyl [148] group on the nitrogen in the 3-position.

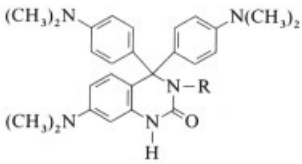

56

Blue color formers are obtained by condensation of Michler's ketone with (2-benzothiazolyl)-3-diethylaminophenylamine. Cyclization gives a *dihydroquinazoline* (**57**) [149].

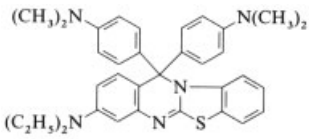

57

Diaminotriphenylmethane dyes can be converted into color formers by bonding a base to the central carbon atom of a triarylmethane dye. Examples are malachite green [150] and dyes in which the *p*-nitrogen is bridged with the *ortho* position, as in 1-ethyl-2,2,4-trimethyl-1,2,3,4-tetrahydroquinoline, 1,2,3,3-tetramethylindoline [151], or 9-ethylcarbazole [152], [153].

6.2.3. Anionic Dyes

Triarylmethane dyes must contain at least two sulfonic acid groups before they can assume anionic character. These dyes are obtained by condensation of a benzaldehyde derivative with 2 mol equivalents of a substituted aniline. A leuco compound is first formed, which must then be oxidized to give the dye by using lead dioxide, manganese dioxide, or sodium dichromate. The two sulfonic acid groups can be located either on the benzaldehyde or on the aniline derivative.

Patent Blue Range. In 1902 STEINER showed that a sulfonic acid group *ortho* to the central carbon atom imparts particularly good fastness to alkalis. Systematic development led to the patent blue dyes, which have become valuable products for dyeing wool.

One of the first of this series was *patent blue V* [*3536-49-0*] (**58**) (C.I. Acid Blue 3, C.I. Food Blue 5). It is obtained by condensation of 3-hydroxybenzaldehyde with *N,N*-diethylaniline, sulfonation of the leuco base *ortho* and *para* to the hydroxyl group, and oxidation with manganese dioxide [154].

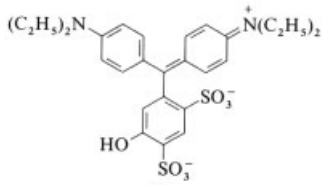

58

A dye with similarly good properties can be produced more easily. *Patent blue VE* [*129-17-9*] (**59**) (C.I. Acid Blue 1, C.I. Food Blue 3, C.I. 42 045, with R = ethyl) does not contain a

hydroxyl group and can be produced directly from 4-formyl-1,3-benzenedisulfonic acid and diethylaniline with subsequent oxidation [155].

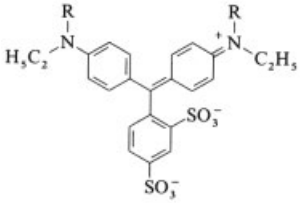

59

If diethylaniline is replaced by ethylbenzylaniline in this synthesis, *patent blue AF* (**59**) (C.I. Acid Blue 7, C.I. 42080, R = benzyl) is formed. Both dyes are used for wool.

Trade names. Acid Blue 1 is commercially available as Sulfacid Brilliant Blue 6 J (ICI), and Acid Blue 7 as Sandolan Turquoise E-AS (Sandoz).

The greatest economic importance of some patent blue dyes is in nontextile applications. Acid Blue 1 and 3 are both used in sanitary installations to dye lavatory fresheners or flushing water. *Patent blue AE* [*3844-45-9*] (**60**) (C.I. Acid Blue 9, C.I. Food Blue 2, C.I. 42090, R = SO_3H) can also be used for this purpose. Its leuco compound is obtained from 2-formylbenzenesulfonic acid and *N*-ethyl-*N*-(3-sulfobenzyl)aniline [156] and oxidized by an environmentally acceptable process in the presence of (dihydrodibenzotetraaza[14]annulene)iron as the catalyst with peroxomonosulfuric acid [157] or hydrogen peroxide [37]. In 1993, more than 1000 t of this dye were produced and used in sanitary installations.

Trade names include Basacid Blue 755 (BASF), Sulfacid Brilliant Blue 5 J (ICI), and Neolan Blue E-A (Ciba-Geigy).

Derivatives of N-Ethyl-N-(3-sulfobenzyl)aniline. Dyes that do not belong to the patent blue range and, like C.I. Acid Blue 9, are derived from *N*-ethyl-*N*-(3-sulfobenzyl) aniline include C.I. Acid Green 3 [*4680-78-8*], (**60**, R = H, C.I. Food Green 1, C.I. 42085) and C.I. Acid Green 9 [*4857-81-2*] (**60**, R = Cl, C.I. 42100).

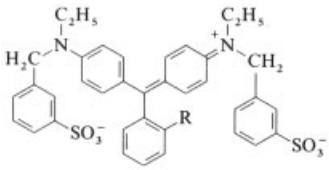

60

The starting material for C.I. Acid Green 3 is benzaldehyde and for C.I. Acid Green 9, *o*-chlorobenzaldehyde. Both dyes are used for wool.

Trade names. Hispacid Green GB (Robama) and Sandolan Brilliant Green E-6 B (Sandoz).

Production Example for Acid Green 3. N-Ethyl-*N*-(3-sulfobenzyl)aniline (1340 kg) is charged as a moist paste to a glass-lined stirred vessel containing 2000 L water, which is equipped with a reflux condenser; 300 kg benzaldehyde is then added. The mixture is heated to 98 – 100 °C for 48 h, and excess benzaldehyde is distilled off. The mixture is diluted to 8000 L with water, and the precipitated leuco acid is dissolved at 80 °C by slow addition of 500 L 30 % sodium hydroxide, resulting in a pH between 7 and 7.2. The solution of the leuco acid is then diluted to 9000 L with water and oxidized to the dye in three portions. Then, 3000 L of the solution and 4000 L of water are cooled to 0 °C in a glass-lined stirred vessel. A 35.5 wt % aqueous solution of 75 kg sodium dichromate is added with vigorous stirring. A solution of 192 kg oxalic acid in 800 L water and 96 kg 78 % sulfuric acid is then charged in one portion, and the mixture is stirred for 15 min at 0 – 10 °C. It is then heated to 50 °C, and the dye is salted out as a colored resinous material with 600 kg sodium sulfate. After stirring for 1 h at 50 °C, 2 kg of a nonionic – polymeric emulsifier [castor oil + poly(ethylene oxide)] is added. After an additional 20 min without stirring, the colored resinous material is drained on drying sheets. Yield: 1640 kg.

Derivatives of Naphthalenesulfonic Acids. Dyes containing naphthalenesulfonic acids require tetraalkylated diaminobenzhydrols (diaminodiphenylmethanols) as condensation partners. *C.I. Acid Green 16* (**61**) (C.I. 44025) is produced by condensation of 4,4′-bis(diethylamino)benzhydrol with 2,7-naphthalenedisulfonic

acid and oxidation of the leuco compound [158]. The dye is used for wool and 50 % wool blends.
Trade Names. Neolan Green E-B (Ciba-Geigy) and Sulfacid Brilliant Green 2 J (ICI).

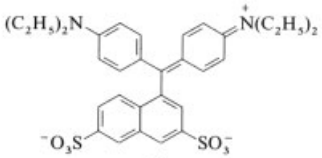

61

Marking inks can be produced from 4-formyl-1,3-benzenedisulfonic acid and polyoxyethylated aniline. The latter is obtained by treating aniline with 50 – 100 mol ethylene oxide. The marking inks can be washed readily from the substrate and do not leave any permanent coloration on the skin [159].

6.3. Triaminotriphenylmethane Dyes

6.3.1. Cationic Dyes

Pararosaniline and its ring-alkylated derivatives belong to the group of cationic triaminotriphenylmethane dyes, as do the phenylated and alkylated rosanilines and the triaminodiphenylnaphthylmethane dyes of the Victoria blue series.

Pararosaniline [*569-61-9*] (**62**) (parafuchsin, paramagenta, C.I. Basic Red 9, C.I. 42500) is the simplest member of this series.

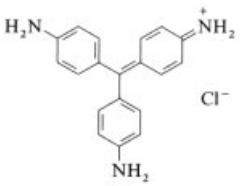

62

It is produced by condensation of 4-aminobenzaldehyde with aniline [160]; 4-aminobenzyl alcohol with aniline [161]; or 4,4′-diaminodiphenylmethane with aniline [162] in the presence of oxidizing agents such as nitrobenzene, nitrotoluene, azobenzene, or atmospheric oxygen in the presence of iron(II) chloride, arsenic acid, or ammonium or sodium metavanadate [163] at 150 – 170 °C. Pararosaniline can be purified by precipitation from the aniline solution with dodecylbenzene [164]. An aniline-free product is also obtained by converting **62** to the nitrate salt, which is dissolved in aniline and converted into the triphenylmethanol derivative with sodium hydroxide. The triphenylmethanol is isolated and converted with acids into the corresponding dye salts, which are deep red in solution [165].

Uses. Pararosaniline is an intermediate for the production of phenylated rosanilines. It gives a brilliant red dyeing on polyacrylonitrile fibers. In analytical chemistry, pararosaniline can be used to detect aldehydes in biological materials [166] and to determine the content of sulfur dioxide in air [167]. It can also be used to produce a reagent for detecting aromatic aldehydes [166].

Fuchsin [*632-99-5*] (**63**) (rosaniline, magenta, C.I. Basic Violet 14, C.I. 42501) is produced by oxidizing a mixture of aniline and toluidines with nitrobenzene, nitrotoluene, or atmospheric oxygen in the presence of zinc chloride or iron(II) chloride. The methyl group of *p*-toluidine provides the central carbon atom. Fuchsin is the oldest cationic triaminotriphenylmethane dye. The discovery of fuchsin with the methyl group *ortho* to the amino group occurred by chance, because the aniline used contained *o*-toluidine as an impurity.

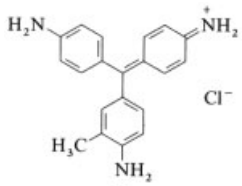

63

Uses. Fuchsin dyes silk, wool, and paper in brilliant red shades with a blue cast and is used as a microscopic stain.
Trade names include Atacryl Fuchsin (Atlantic Industries), Basic Magenta (Amar Dye-Chem.), and Magenta (W. S. Simpson & Co.).

New fuchsin [*3248-91-7*] (**64**) (C.I. Basic Violet 2, C.I. 42520) is obtained by condensing *N*-methyl-*o*-toluidine with formaldehyde [168] or methylene dichloride [169], or from *N,N*-dimethyl-*o*-toluidine in the presence of nitro compounds as oxidizing agent. Under these conditions methyl groups are eliminated from the nitrogen atom of *N,N*-dimethyl-*o*-toluidine as formaldehyde. These then form the central carbon atom of the dye.

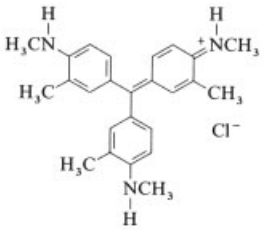

64

New fuchsin can also be produced by reaction of *N*-methyl-*o*-toluidine, *o*-toluidine hydrochloride, and *m*-xylidine with *o*-nitrotoluene and iron(II) chloride [170]. The crude dye obtained by this method can be purified by dissolving it in a mixture of chlorobenzene and dilute hydrochloric acid and cooling it slowly [171]. Stable concentrated solutions of new fuchsin are obtained in a mixture of acetic acid and mono- or polyhydric alcohols as solvent [172]. The tint of new fuchsin is bluer than that of pararosaniline.

Uses. New fuchsin can be used for dyeing polyacrylonitrile, paper, and leather. Hexaazotized new fuchsin can be used to detect proteinases in cell membranes [173].

Trade names. Basic Magenta (Dintex Dyechem PVT) and Ravi Basic Magenta II (Ravi Chem Dye Bombay).

Rosaniline dyes containing carboxyl groups can be used as color or filter layers in photographic materials [174].

Tris(4-hydrazinophenyl)methylium chloride is used for detecting aromatic aldehydes and ketones [175]. It can be produced by diazotization of pararosaniline followed by reduction in the presence of zinc chloride.

Phenylated rosaniline dyes can be produced by two different processes. Pararosaniline is the starting material for the BASF process. It is subjected to acid-catalyzed phenylation with aniline involving elimination of ammonia to give *spirit blue* [*2152-64-9*] (**65**) (Blue II T Base, Opal, C.I. Solvent Blue 23, C.I. 42760). Depending on reaction conditions, one to three phenyl groups can be introduced. The blue dye becomes greener with increasing degree of phenylation.

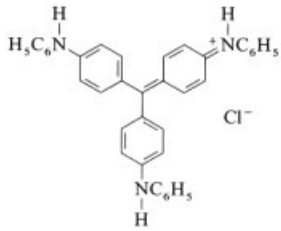

65

Production Example for Spirit Blue [176]. A mixture of 460 kg pararosaniline, 2450 kg aniline, and 8.5 kg benzoic acid is heated to 180 – 182 °C until the dye has reached the desired color that is tested in an acetic acid – alcohol mixture. This requires ca. $1\,^1/_2 - 2$ h. After the reaction mixture is cooled to 100 °C, it is added to a mixture of hydrochloric acid and water. The dye precipitates and is separated by suction filtration. To remove residual aniline, the product is heated again with dilute hydrochloric acid and the dye is filtered, washed, and dried under vacuum at 100 °C, yield: 682 kg.

The arylation of pararosaniline can also be carried out with toluidines, diamines [41], and alkoxyarylamines [42], as well as aniline. The reflex blue dyes made by Hoechst also contain spirit blue or similar compounds. The synthesis of reflex blue dyes does not start from pararosaniline but from 4,4′,4″-trichlorotriphenylmethyl tetrachloroaluminate, which is obtained from *p*-chlorobenzotrichloride and chlorobenzene in the presence of aluminum trichloride and whose chlorine atoms can be substituted by nucleophiles. Replacing the chlorines with aniline, however, leads to an undesired side reaction. After substitution of the third chlorine atom, the carbon atom *para* to the amino group of another molecule of aniline attacks the central carbon

of triphenylpararosaniline. A molecule of diphenylamine is eliminated, with the formation of diphenylpararosaniline. It is thus very difficult to obtain blue dyes with green casts in good yield according to this method. However, if *m*-toluidine is used instead of aniline, dyes having blue shades with a green cast can be obtained. These cannot be produced by any other process. The reason for the successful reaction with *m*-toluidine is that the *m*-methyl group shields the carbon *para* to the amino group and thus hinders nucleophilic attack at the central carbon atom [39].

Production Example for Tri-m-tolylpararosaniline [39]. Production of 4,4′,4″-trichlorotriphenylmethyl tetrachloro-aluminate: A mixture of 200 g 4-chlorobenzotrichloride in 190 g chlorobenzene is added to a suspension of 127.5 g anhydrous aluminum trichloride in 450 g chlorobenzene with cooling and stirring in such a way that the temperature does not rise above −5 °C. The mixture is stirred for 15 h at −5 °C and then heated to 55 – 60 °C and kept at this temperature for ca. 20 h. After cooling, excess chlorobenzene is removed by suction through an attached glass frit, leaving the condensation product in the form of a thick crystalline precipitate.

Production of tri-*m*-tolylpararosaniline: *m*-Toluidine (613 g) is added to the aluminum complex of 4,4′,4″-trichlorotriphenylmethane, and the mixture is heated to 150 °C for 5 h. The strongly bronzing blue melt is stirred in dilute hydrochloric acid. Small quantities of chlorobenzene are removed by steam distillation, and the crystalline dye is washed until neutral.

The three chlorine atoms in 4,4′,4″-trichlorotriphenylmethyl tetrachloroaluminate differ in their reactivity. Besides symmetrical spirit blue derivatives, unsymmetrical compounds can also be produced. One chlorine atom can first be replaced by a *m*- or *p*-substituted aniline derivative. Aniline or an *o*-substituted aniline can then be substituted for the others [177]. Triphenylpararosanilines can also be produced from diphenylamine derivatives and carbon tetrachloride in the presence of aluminum trichloride [178]. Addition of DMF promotes this reaction [179].

Triphenylpararosaniline sulfates have pigment character and can be used in place of alkali blue [180]. Color lakes made from triphenylpararosanilines and molybdenum- or tungsten-containing heteropolyacids are blue pigments [181], [182].

Methyl Violet. Alkylated pararosaniline dyes are obtained from secondary and tertiary aromatic amines. Methyl violet [*101357-18-0*] (C.I. Basic Violet 1, C.I. 42535, derivatives: C.I. Solvent Violet 8; C.I. Pigment Violet 3, 3 : 3, 27) can be produced by air oxidation of dimethylaniline in the presence of phenol, copper salts, and sodium chloride. A methyl group is eliminated and oxidized to formaldehyde. Further condensation with dimethylaniline and oxidation lead to the dye. Since the partially dealkylated dimethylaniline also takes part in the condensation, methyl violet is a mixture of tetra-, penta-, and hexamethylated pararosanilines.

Production Example [183]. In a cast iron, glass-lined 5-m^3 vessel, 30 kg sodium chloride and 236 kg copper sulfate are dissolved in 1000 L water. Then, 400 kg dimethylaniline and 25 kg phenol are added and the mixture is stirred for 30 min at 20 °C. After increasing the stirrer speed, 222 kg sodium chloride is added in portions of 10 kg/h. In the first 4 h, 12 m^3/h air is blown in; in the subsequent 16 h, 17 m^3/h; and in the last 4 h, 23 m^3/h. At the end of this period the melt contains only traces of dimethylaniline. Formaldehyde must no longer be detectable. After being heated to 100 °C the mixture is drained into a copper trough with stirring, and after being cooled to 50 °C, the supernatant aqueous solution is separated from the melt. The completely cooled melt is ground up and stirred for 12 h in an iron vessel with 1500 L water, 50 kg sodium chloride, and 60 kg 34.5 wt % aqueous sodium hydroxide solution to decompose the copper(I) chloride double salt of the dye. After addition of 60 kg 34.5 wt % aqueous sodium hydroxide solution, the reaction product is added to 2500 L water, and the volume is made up to 8000 L with water. After stirring for 4 h at 20 °C, 180 kg 34.5 wt % aqueous sodium hydroxide solutions added, and the mixture is heated to 95 °C in 6 h. The color base is allowed to settle, ca. 3000 L supernatant alkali is removed, and the volume is again made up to 8000 L with cold water, whereby the temperature is adjusted to 60 °C. A solution of 65 kg sodium sulfide (60 %) in 3000 L water is added with stirring to convert the copper oxide formed in the alkaline liquor into copper sulfide. After

stirring for 2 h, the color base is decanted three times by alternately adding and removing water. The color base is then stirred with 5500 L water and 80 kg 62.5 % sulfuric acid for 1 h in the cold. It is heated to 100 °C and brought into solution with sulfuric acid in a weakly acidic solution. The dye solution is separated from copper sulfide on a filter press and salted out with sodium chloride. After redissolving in water and salting out with sodium chloride, the dye is converted completely to the chloride. After isolation, it is dried in a drum dryer. Yield: 104 kg from 100 kg of reacted dimethylaniline (84.3 %).

Instead of phenol, the addition of a dispersing agent [184] or a nonionic surfactant [185] to the oxidation melt is recommended.

Uses. Methyl violet is readily soluble in organic solvents and is therefore used as the dye salt and dye base (Solvent Violet 8) in many areas of the office equipment industry (e.g., in the production of carbon paper, stamping inks, hectographic inks, and inking ribbons). It is also a component of black dyes for polyacrylonitrile materials.

Trade names include Aizen Methyl Violet RL Special (Hodogaya), Basonyl Violet 600 (BASF), and Paris Violet (Zeneca).

The phosphomolybdotungstate (Pigment Violet 3), copper hexacyanoferrate (Pigment Violet 27) and tannic acid (Pigment Violet 3 : 3) salts of methyl violet are pigments with high color strength used in the production of printing inks.

C.I. Basic Violet 13 [*8004-88-4*], C.I. 42536, is produced by benzylation of methyl violet and consists of a mixture of benzylated tetra-, and pentamethylpararosaniline and hexamethylpararosaniline.

Trade name. Benzyl Violet DSC (Dye Specialties).

Crystal violet [*548-62-9*] (**66**) (C.I. Basic Violet 3, C.I. 42555, derivatives: C.I. Solvent Violet 9, C.I. Pigment Violet 39) is produced by condensation of dimethylaniline and phosgene. 4,4′-Bis(dimethylamino)benzophenone (Michler's ketone) is formed as an intermediate, but is not isolated [186], [187].

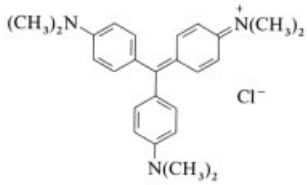

66

The production of crystal violet can also be carried out as a one-pot reaction by condensing dimethylaniline with formaldehyde to 4,4′-methylenebis(*N,N*-dimethylaniline), which is reacted with dimethylaniline and simultaneously oxidized to the dye by atmospheric oxygen in the presence of (dihydrodibenzotetraaza[14]annulene)iron and chloranil [188]. Vanadium and molybdenum compounds [32] and nitrous gases [35] can also be used as oxidation catalysts together with chloranil. Crystal violet can be produced by condensation of 4-dimethylaminobenzaldehyde [30] with 2 mol equivalents of dimethylaniline or from 4,4′-bis(dimethylamino)diphenylmethanol with 1 mol equivalent of dimethylaniline and subsequent oxidation of the leuco base.

Production Example [188]. A mixture of 1089 parts dimethylaniline, 90 parts paraformaldehyde, and 1391 parts glacial acetic acid is heated to 70 °C for 1 h. The solution is cooled to 50 °C, treated with 15 parts chloranil and 15 parts (dihydrodibenzotetraaza[14]annulene)iron, and the mixture aerated for 20 h with vigorous stirring. Ca. 800 parts glacial acetic acid are distilled off at reduced pressure and the residue is dissolved in 16 000 parts water at 95 °C. The solution is filtered hot and the dye is precipitated from the filtrate with 3000 parts sodium chloride at pH 1.5. The crystalline dye is collected by suction filtration and washed with 3000 parts dilute sodium chloride solution. Yield: 1050 parts.

Uses. Crystal violet is used in the office equipment industry for the production of indelible pencils, for making stamping and hectographic inks, and for dyeing paper. The color strength in hectographic coloring matter is particularly high if crystal violet is present as the nitrate [189]. The grain hardness of crystal violet or methyl violet is lowered by crystallizing the dye in the presence of poly(ethylene glycol) [190]. Crystal violet can also be used as a component of navy blue and black dyes for

polyacrylonitrile fibers. Ink jet inks containing crystal violet and di(ethylene glycol) butyl ether have high antibacterial activity [191]. The salt of crystal violet and a heteropolyacid has been proposed for detecting phosphorus [192].

Precipitation with phosphomolybdotungstic acid gives a printing ink pigment, C.I. Pigment Violet 39.

Trade names for methyl violet and crystal violet include Basazol Violet 57 L (BASF), Methyl Violet B (Zeneca), Dycosbasic Methyl Violet 2 B (Shanghai Dyestuffs Industries), and Crystal Violet DSC pure (Dye Specialties).

Ethyl violet [*2390-59-2*] (**67**) (C.I. Basic Violet 4, C.I. 42600, derivative: C.I. Pigment Blue 14) is obtained analogously to crystal violet from diethylaniline and phosgene or from diethylaniline and formaldehyde with catalytic oxidation [188].

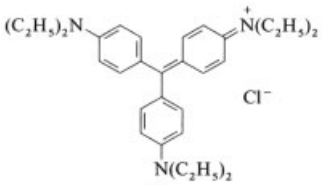

67

Uses. Ethyl violet is used to dye wood-containing paper. With phosphomolybdotungstic acid, a pigment for printing inks, Pigment Blue 14, is obtained.

Trade names are Ethyl Violet (Dye Specialties) and, for use as a pigment, Halopont Blue RNM (Mobay).

Cationic triaminotriphenylmethane dyes that contain a *phenylamino group* are obtained by treating diaminotriphenylmethane dyes, which have a chlorine atom bonded to the 4-position of the third phenyl group, with aromatic amines [193]. If cyanoethyl groups are present on the two other amino groups of these phenyl-substituted dyes, lightfast coloration of polyacrylonitrile fibers is achieved [194].

Triaminotriphenylmethane dyes with *long alkoxylated amino groups* are used as marking inks [195]. These dyes fulfill the requirements for fugitive inks for temporary marking during fiber production [130].

Victoria blue B [*2580-56-5*] (**68**) (C.I. Basic Blue 26, C.I. 44045, derivatives: C.I. Solvent Blue 4, C.I. Pigment Blue 2) is produced by condensation of Michler's ketone with *N*-phenyl-α-naphthylamine in toluene with phosphorus oxychloride [196] or by catalytic oxidation of the leuco base [188] – formed by oxidation of 4,4′-bis(dimethylaminophenyl)methane in the presence of *N*-phenyl-α-naphthylamine – with air and catalytic quantities of (dihydrodibenzotetraaza[14]annulene)iron and chloranil.

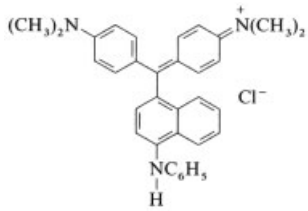

68

Uses. Victoria blue B is used to dye paper and, melted in oleic or stearic acid, as a color base for the production of printing inks.

Trade names are Aizan Victoria Blue BH (Hodogaya), Atlantic Victoria Blue B (Atlantic Industries), and Basazol C Blue 57 L (BASF), for example.

The *N*-methyl derivative of Victoria blue B is commercially available under the name *Victoria blue 4 R* [*2185-87-7*] (C.I. Basic Blue 8, C.I. 42563, derivative: C.I. Solvent Blue 2).

Trade name: Basonyl Blue 633 (BASF).

If the phenylamino group in Victoria blue B is replaced by an ethylamino group, *Victoria blue R* [*2185-86-6*] (C.I. Basic Blue 11, C.I. 44040, derivative: C.I. Pigment Blue 10) is obtained.

Trade name: Dycosbasic Victoria Blue R (Shanghai Dyestuffs Industries) and Victoria Blue RSA (Zeneca).

Victoria blue FBR [*3251-84-1*], (C.I. Basic Blue 55, C.I. 44044) differs from Victoria blue B in that it contains a cyclohexylamino group instead of a phenylamino group. It is more soluble in alcohol than Victoria blue B.

Trade name: Flexo Blue 630 (BASF) [197].

Victoria pure blue BO [*2390-60-5*] (**69**) (C.I. Basic Blue 7, C.I. 42595, C.I. Solvent Blue 5, C.I. Pigment Blue 1, 1 : 2, 62) is produced by treating 4,4′-bis(diethylamino)benzo-

phenone with phosgene to give the ketone dichloride, followed by treatment with *N*-ethyl-α-naphthylamine [198].

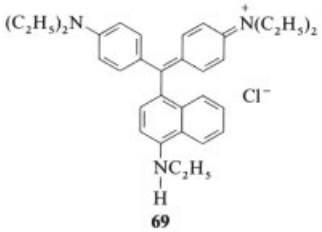

69

Victoria blue BO is used to dye paper and in stamping and flexographic printing inks.
Trade names include Basonyl Blue 636, Flexo Blue 636 (BASF), and Mitsui Victoria Pure Blue BO (Mitsui Toatsu Dyes).

Victoria pure blue BO is converted into a color lake with phosphomolybdotungstic acid (Pigment Blue 1), which is used in the printing inks industry for decorative printing. The color lake obtained from Victoria pure blue BO and copper hexacyanoferrate(II), Pigment Blue 62, is used in aqueous printing inks and for shading gravure printing inks for illustrations. A coloring material used in ballpoint inks and pastes consists of the salt of a polysulfonated phthalocyanine and Victoria pure blue BO [199]. Addition of cyanotriacetic acid or ethylenediaminetetraacetic acid gives pigments in the Victoria blue range, which are stable in nitrocellulose [200].
Trade names. Fast Blue Lake (China National Chemicals), Fanal Blue D 6340 (BASF), and Irgalit Blue FR (Ciba-Geigy).

Victoria pure blue FGA [*73309-46-3*] (C.I. Basic Blue 81, C.I. 42 598) is similar to Victoria pure blue BO, but contains a 4-ethoxyphenylamino group instead of an ethylamino group. It is insoluble in cold water and can therefore be used in printing inks with good waterfastness without addition of a laking agent.
Trade name: Flexo Blue 680 (BASF).

Crystal violet tetraphenylborate [201], Victoria blue dyes [202], and substituted triphenylpararosanilines [203] can be used as positive charge stabilizers in electrophotography. Anodized aluminum can be dyed electrolytically with triarylmethane dyes [204].

Leuco compounds of the triaminotriphenylmethane dyes, in which the amino groups are replaced by alkyl [205] or aryl groups [206], have been proposed as color formers for colorless copier paper. Leuco triaminotriphenylmethanes can be used as charge-transfer compounds [207] and photoreceptors [208] in electrophotography and as contrast dyes in photoresists [209]. Leuco bases can also be used for the quantitative detection of ozone [210] and in microencapsulated temperature indicator systems together with an oxidizing agent [211].

6.3.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones)

Colorless compounds derived from triaminotriphenylmethanols and lactones (phthalides) form deeply colored compounds on irradiation or contact with acidic substrates. A film for radiation dose measurement contains the azide of crystal violet, which is converted to the dye on irradiation. The optical density is proportional to the radiation absorbed [212]. The cyanides of crystal violet [213] and new fuchsin [214] give deep colorations under UV light and are therefore suitable for reproduction purposes.

Lactones and Related Compounds. *Crystal violet lactone* [*1552-42-7*] (**70**) is the most important triaminotriphenylphthalide. In 1993, its sales were ca. 800 t worldwide. This color former gives a reddish blue print. Crystal violet lactone can also be used in gas analysis. The blue color of thin layers, which contain, besides the lactone, an acidic phenol [e.g., bisphenol A (2,2′-bis(4-hydroxyphenyl)propane], becomes weaker on addition of polar solvents. This optochemical sensor system permits detection of the smallest quantities of polar solvents in air [215].

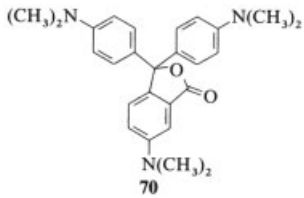

70

Crystal violet lactone can be produced by condensation of Michler's ketone with methyl *m*-diethylaminobenzoate [216]. This

reaction and those involving 5-dimethylaminophthalaldehyde acid [217] and its methyl ester [218] (**71**) are, however, not used in industry. Both compounds **71** can be obtained by formylation of the mixed anhydride of acetic acid and 3-dimethylaminobenzoic acid or of methyl-3-dimethylaminobenzoate, respectively. They can be converted into leuco crystal violet carboxylic acid (**72**).

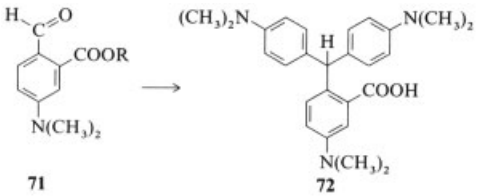

3-(4-Dimethylaminophenyl)-6-dimethylaminophthalide (**73**) has versatile synthetic possibilities. The phthalide can be produced from 3-dimethylaminobenzoic acid and 4-dimethylaminobenzaldehyde. It can be condensed with dimethylaniline in the presence of aluminum trichloride to give leuco crystal violet carboxylic acid (**72**) [219], which is then converted to crystal violet lactone. A technically simpler variant involves oxidation of the phthalide and subsequent condensation of the 4,4′-bis(dimethylamino)-benzophenone-2-carboxylic acid (**74**) with dimethylaniline to give crystal violet lactone [220]. 3-Nitrobenzenesulfonic acid has proved to be a particularly effective oxidizing agent [221].

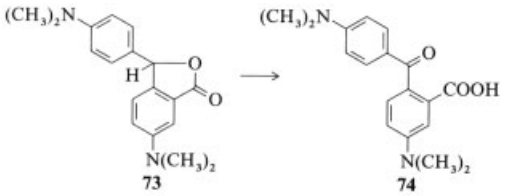

Michler's hydrol reacts readily with 3-dimethylaminobenzoic acid and is therefore used for industrial syntheses. The condensation can be carried out in sulfuric [222] or phosphoric acid [223]. Instead of Michler's hydrol, leuco auramine obtained by reduction of auramine with zinc can also be used [224].

If 4-dimethylaminobenzaldehyde, dimethylaniline, and 3-dimethylaminobenzoic acid are reacted in the presence of urea under acidic conditions, the synthesis proceeds via a urea adduct of Michler's hydrol [225]. The reaction gives the 4-dimethylaminobenzylideneurea (**75**), to which dimethylaniline adds to form Michler's hydrol – urea (**76**). The latter can be condensed with 3-dimethylaminobenzoic acid [226].

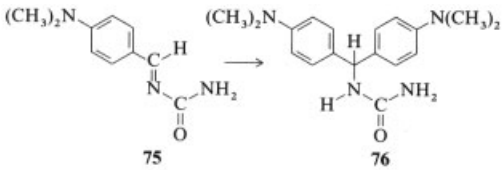

The leuco crystal violet carboxylic acid formed in the condensation is contaminated by leuco crystal violet. This can be removed by solvent extraction if the carboxylic acid is dissolved in alkali and the solution extracted with a mixture of toluene and heptane, for example [225], [227]. Oxidation of the leuco compound to crystal violet lactone can be carried out in aqueous solution with potassium peroxodisulfate [228] and potassium permanganate [229]. Hydrogen peroxide has proved particularly suitable [230]. The selectivity of oxidation with hydrogen peroxide can be increased by addition of hexacyanoferrate [231]; manganese dioxide [232]; vanadium, chromium, manganese, and iron salts [233]; and copper, cobalt, and nickel salts [237]. Oxidation can also be carried out with oxygen in the presence of salcomine [235] or water-soluble 1,2-cobalt complexes of azo dyes [236].

Besides crystal violet lactone, compounds have also been described that have a methyl group *ortho* to the central carbon atom [237] or in which the amino nitrogen carries a benzyl group [238]. *Lactams* have also been described as color formers [239] and *sulfonephthaleins* (intramolecular sulfonates) as charge-generating compounds for electrophotography [240].

If crystal violet lactone is treated with hydrazine, the *dihydrophthalazinone* (**77**) is formed. This compound has been the subject of extensive investigations.

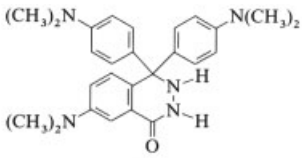

77

As model experiments with lead tetraacetate have shown, oxidation of the hydrazide leads to loss of nitrogen, followed by rearrangement to several products, which have an absorption maximum at 490 nm [241]. Because of this property the compound is a redox indicator. It is used for colorimetric detection of hydrogen peroxide in metabolic processes [242]. This compound and its derivatives are also used as color formers in electrochromic recording [243].

6.3.3. Betaines and Anionic Dyes

Betaine and anionic triaminotriarylmethane dyes have achieved considerable economic importance. They are derived from pararosaniline dyes containing an aryl or benzyl group with a sulfonic acid group on it.

Arylated pararosaniline dyes with one sulfonic acid group (**78**) are water-insoluble betaines and have pigment character. They are used under the *trade names* Alkali Blue (BASF) and Reflex Blue (Hoechst) to brighten carbon black in printing inks.

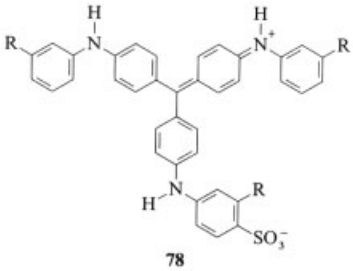

78

In 1992, more than 3000 t of this pigment were produced worldwide. *C.I. Pigment Blue 61* (**78**, R = H) (C.I. 42765) and *C.I. Pigment Blue 56* (**78**, R = methyl) (C.I. 42800) are obtained by sulfonation of arylated pararosaniline dyes, whose production is described in Section 6.3.1. *Production Example.* With vigorous stirring, 50 g tri-*m*-tolylpararosaniline is added to 25 g 89 wt % sulfuric acid at 20 – 25 °C over 1 – 2 h. The mixture is heated to 40 – 45 °C and stirred until a sample of three drops is soluble in 20 mL 4 % ammonia on boiling. The reaction mixture is first stirred on 250 g ice and then stirred for 30 min at 50 – 60 °C. The solid is collected by suction filtration and washed with water. The filter cake is suspended in 500 mL water; the suspension is heated to 80 – 85 °C, treated with 27 g 33 wt % aqueous sodium hydroxide solution, and then heated to its boiling point. After ca. 10 min a clear reddish solution is formed, which is treated rapidly with 120 g 12 % sulfuric acid. The mixture is stirred for 5 min at 95 – 100 °C and cooled rapidly with cold water to 80 – 85 °C; the solid is collected by suction filtration and washed with water. After drying at 40 °C, 55 g of the corresponding monosulfonated dye is obtained.

C.I. Pigment Blue 56 can be used to produce printing inks for paper and plastics with high color strength. The properties of the resulting pigment (e.g., purity, color strength, grain hardness, and flushing or flowing behavior) strongly depend on the workup of the sulfonation process and thus on the particle-size distribution obtained. The property profile can also be affected by use of coating agents and other specially selected additives [244], [245].

The three chlorine atoms in 4,4′,4″-trichlorotriphenylmethyl tetrachloroaluminate have differing reactivities. Direct synthesis of the corresponding monosulfonated dye is therefore possible, which makes sulfonation of spirit blue and associated disposal of the sulfuric acid used unnecessary [246]. After sulfanilic acid has been substituted for one chlorine atom the other two chlorine atoms can be replaced by aniline or *m*-toluidine.

C.I. Acid Blue 93. The more highly sulfonated triphenylpararosanilines are readily water soluble and are used for writing inks. This applies particularly to the trisulfonic acid (C.I. Acid Blue 93, C.I. 42780), which is sold by Hoechst under the *trade name* Ink Blue BJT. C.I. Acid Blue 93 can also be produced by condensing formaldehyde with 2 mol equivalents of diphenylaminosulfonic acid and oxidatively coupling the diphenylmethane derivative with another mol equivalent of diphenylaminosulfonic acid in the presence of iron(III) chloride [247].

Although triphenylpararosanilines have achieved importance mainly as betaines, the benzylated pararosanilines are used as anionic dyes for textiles.

C.I. Acid Blue 104 (**79**, R^1 = methyl, R^2 = diethylamino) (C.I. 42735, *trade name* Atlantic Alizarin Walk Blue FFR, Atlantic Industries) gives a clear blue shade on wool. It is obtained by condensation of 4-diethylaminobenzaldehyde with *N*-ethyl-*N*-(3-sulfobenzyl)-*m*-toluidine and oxidation with sodium dichromate and oxalic acid [248].

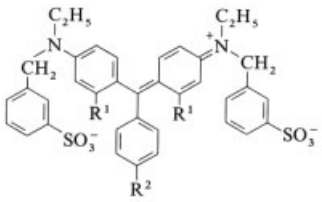

79

C.I. Acid Violet 49 [*1694-09-3*] (**79**, R^1 = H, R^2 = dimethylamino) (C.I. 42640, *trade name* Atlantic Acid Violet 4BNS, Atlantic Industries). For Acid Violet 49 a synthesis using a diphenylmethanol derivative was developed, which involves four steps without isolation of any intermediates. *N*-Ethyl-*N*-benzylaniline is first converted to the corresponding sulfonic acid, which is then condensed with formaldehyde to give the diaminodiphenylmethane derivative. After oxidation with sodium dichromate, the benzhydrol obtained is condensed with dimethylaniline to give the leuco compound, which is then oxidized to the dye [249].

Production Example. Sulfuric acid monohydrate (520 kg) is charged to a 3-m^3 glass-lined stirred vessel. *N*-Ethyl-*N*-benzylaniline (520 kg) is added slowly with stirring over ca. 20 h at 35 – 40 °C. Then, 520 kg 65 % oleum is added at 35 – 40 °C over 16 h. The reaction mixture is added to 8000 L water with cooling and stirring in a 20-m^3 stirred vessel. The pH is adjusted to 2 – 2.7 by adding of ca. 1850 kg 30 wt % aqueous sodium hydroxide solution within 2 h and keeping the temperature below 60 °C. Next, 135 kg 30 % formaldehyde solution is added and the total volume is made up to 11 000 L with water. The reaction mixture is heated to 95 °C with stirring and kept at this temperature for 4 h. After cooling to 50 – 60 °C the precipitated methane derivative is dissolved at pH 5.6 – 5.8 by adding 400 kg 30 wt % aqueous sodium hydroxide solution. The total volume is made up to 12 000 L with water, and the reaction mixture is divided into three portions. Then, 4000 L of the methane derivative solution is diluted to 13 200 L with water in a 20-m^3 stirred vessel, and 286 kg 35 wt % sodium dichromate solution, 53 kg 78 wt % sulfuric acid, and 200 L water are added at 25 – 30 °C. After 20 min, 51.6 kg *N,N*-dimethylaniline is added at 25 – 30 °C. The reaction mixture is then stirred for 8 h at 25 – 30 °C, 6 h at 45 °C, and 6 h at 75 °C. During this time the leuco compound is formed by condensation and subsequently oxidized to the dye. The total volume is made up to 15 000 L with water, and the dye solution is separated from the chromium sludge by pressing out. The dye is precipitated as a melt at 60 – 70 °C by addition of 850 kg sodium chloride. Then, 1 kg of a nonionic – polymeric emulsifier [castor oil + poly(ethylene oxide)] is added and the stirrer is switched off. The dye resin settles on the bottom of the vessel and is run out onto drying sheets via the bottom valve. Yield: 708 kg.

Oxidation can be carried out with hydrogen peroxide in the presence of a transition-metal catalyst [38] instead of sodium dichromate.

C.I. Acid Violet 17. Another dye used for wool is C.I. Acid Violet 17 [*4129-84-4*] (**79**, R^1 = H, R^2 = diethylamino) (C.I. Food Violet 1, C.I. 42650, *trade name* Hispacid Brilliant Violet S4BF, Robama). This dye can be produced by the same method used for C.I. Acid Violet 49.

Diaminotriphenylmethane dyes, which have easily replaceable substituents in the *p*-position on the third phenyl ring, react with aromatic amines to give triaminotriphenylmethane dyes. If aromatic amines with an alkoxy group *para* to the amino group are used for this reaction, dyes with better lightfastness on wool are obtained. The most commonly used reactive substituents are chlorine and the sulfonic acid group.

C.I. Acid Blue 83 [*6104-59-2*] (**80**, R = H) (C.I. 42660, *trade name* Polar Blue 6 B, CibaGeigy) is produced by condensation of *N*-ethyl-*N*-(3-sulfobenzyl)aniline with 4-

chlorobenzaldehyde, oxidation with dichromate in the presence of oxalic acid, and reaction with *p*-phenetidine in the melt [43].

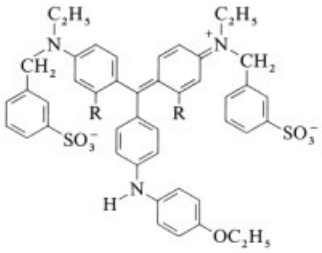

80

Production Example. A filter cake of *N*-ethyl-*N*-(3-sulfobenzyl)aniline weighing 900 kg, with a dry matter content of 670 kg, is stirred in 800 L water and dissolved by adding 50 kg 45 wt % aqueous sodium hydroxide solution to pH 5. After addition of 175 kg 4-chlorobenzaldehyde, the mixture is heated at 90 °C for 36 h. Excess aldehyde is then removed by steam distillation. The reaction mixture is rendered weakly alkaline by addition of 200 L hot water and 200 kg 45 wt % aqueous sodium hydroxide solution, and the solution is diluted to 6000 L. To each 1000 L of this solution, 600 L ice – water and a solution of 23 kg sodium dichromate in 100 L water are added at 10 °C. A solution of 48 kg oxalic acid and 23 kg 78 % sulfuric acid in 300 L water is added within 5 min. Two oxidation reaction mixtures are combined, heated to 45 °C, and the dye is salted out with sodium sulfate. The dye resin is separated and dried at 95 °C under vacuum. Yield: 950 kg (equivalent to 768 kg chlorine green, 100 %).

To replace the chlorine atom, 600 kg *p*-phenetidine is treated with 475 kg chlorine green (384 kg 100 % dye) in a glass-lined vessel at 70 – 75 °C. The mixture is heated to 100 °C and kept at this temperature for 3 h. The progress of the reaction is followed by TLC. After the reaction is complete, the melt is transferred to a mixture of 6000 L water and 500 kg 30 % hydrochloric acid at 70 °C. The pH of the solution must be between 2 and 3. After stirring for 30 min the precipitated dye resin is allowed to settle and the mother liquor is removed. The resin is washed twice with 1500 L hot water containing some hydrochloric acid. The dye is dissolved in 4000 L water at 80 °C with addition of 30 – 35 kg anhydrous sodium acetate and then salted out slowly by addition of 300 kg sodium chloride. The dye resin is separated and dried under vacuum at 95 °C. Excess *p*-phenetidine is recovered from the acidic wash waters by neutralizing the excess acid and precipitating *p*-phenetidine sulfate with sodium sulfate. *p*-Phenetidine sulfate is collected by suction filtration, converted to the base with sodium hydroxide, and distilled under vacuum. Yield: 465 kg Acid Blue 83, corresponding to 1900 kg commerical product.

C.I. Acid Blue 90 [*6104-58-1*] (**80**, R = methyl) (C.I. 42655, *trade names* Polar Blue G, Ciba-Geigy, Derma Cyanine G, Sandoz) is obtained by replacing a sulfonic acid group by *p*-phenetidine. After condensation of benzaldehyde with *N*-ethyl-*N*-benzylaniline, the product is sulfonated to the trisulfonic acid and oxidized with lead dioxide. The *p*-sulfonic acid group is then replaced by aniline [43]. The dye can be used for wool and leather, and in analysis for detecting proteins [250].

C.I. Acid Blue 13 (**81**) (C.I. 42 571, *trade name* Hispacid Fast Violet 10 B, Robama) is synthesized from Michler's hydrol. Production involves sulfonation of *N*-ethylbenzylaniline and combination of the reaction mixture with a solution obtained by oxidation of bis(4-dimethylaminophenyl)methane with manganese dioxide in sulfuric acid solution. The leuco sulfonic acid is oxidized to the dye with lead dioxide [251].

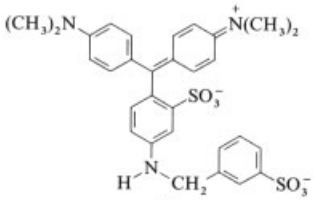

81

6.4. Hydroxytriphenylmethane Dyes

The simplest members of this series are *benzaurin* [*569-60-8*], a 4-[4-hydroxyphenyl(phenyl)methylene]-2,5-cyclohexadien-1-one, and *aurin* (C.I. 43800), a 4-[bis-(4-hydroxyphenyl)-methylene]-2,5-cyclohexadien-1-one, which contains an additional hydroxyl

group and is also known as *p*-rosolic acid. Benzaurin and aurin are relatively unimportant. Only those dyes with a carboxyl group *ortho* to the hydroxyl group, which can be complexed with metals, in particular chromium, have achieved importance in industry.

Chrome-developing dyes are used mainly in dyeing wool. They can also be used to dye polypropylene materials in the presence of water-soluble metal salts [252].

6.4.1. Metal-Complex Dyes

Hydroxytriarylmethane dyes are best produced by condensation of aromatic aldehydes with *o*-hydroxybenzoic acids and oxidation of the leuco compound formed with nitrosylsulfuric acid in concentrated sulfuric acid. To prevent the introduction of undesired sulfonic acid or nitroso groups under the oxidation conditions, the position *ortho* to the hydroxyl group in the hydroxybenzoic acid must be occupied. Therefore, *o*-cresotic acid [*83-40-9*] (2-hydroxy-3-methylbenzoic acid) is used preferably as starting material.

C.I. Mordant Blue 1 (**82**) (C.I. 43830, *trade name* Diamond Chrome Pure Blue B, Bayer) is obtained by condensing 2,6-dichlorobenzaldehyde with *o*-cresotic acid at 20 – 25 °C in 90 % sulfuric acid. After increasing the acid concentration to 95 %, oxidation is then carried out by addition of sodium nitrite.

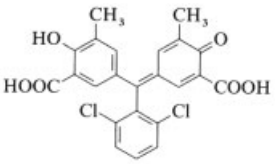

82

Production Example. A mixture of 1300 kg sulfuric acid monohydrate and 660 kg 78 wt % sulfuric acid is cooled to 20 – 25 °C in a glass-lined vessel. Then, 200 kg 84 – 90 % 2,6-dichlorobenzaldehyde is added as a moist paste containing 10 – 16 % water, and the reaction mixture is heated to 50 °C to dissolve the reactants completely. It is cooled to 30 °C, small quantities of precipitated resin are removed, and the mixture is cooled further to 15 – 20 °C. Next, 376 kg *o*-cresotic acid (2-hydroxy-3-methylbenzoic acid) is added over 6 h. The mixture is stirred for 12 h at 20 – 25 °C. Neither aldehyde nor cresotic acid should then be detectable in the solution. The sulfuric acid concentration is increased by adding 535 kg sulfuric acid monohydrate; 50 kg sodium nitrite is added at 15 – 20 °C over 6 h. The mixture is stirred for 1 h at 15 – 20 °C, heated rapidly to 60 °C, and stirred at this temperature until the colorimetrically determined color strength no longer increases. The mixture is transferred to 7000 L ice – water, and the precipitated dye is pressed out and washed with water until free of acid. The dye paste is dissolved in 1000 L water with 300 kg 45 wt % sodium hydroxide solution at 85 °C, and the dye is salted out by addition of 500 kg sodium chloride. It is collected by suction filtration and dried at 95 °C under vacuum. Yield: 615 kg, corresponding to 1500 kg commercial product.

C.I. Mordant Blue 1 dyes wool a cherry red shade, which is transformed into a clear pure blue by afterchroming with dichromate. In afterchroming, trivalent chromium is formed by reduction and chelated by the hydroxyl and carboxyl groups. Besides the bathochromic shift, lightfastness is also improved.

Other chrome mordant dyes with useful properties are obtained when benzaldehyde derivatives other than 2,6-dichlorobenzaldehyde are used for the condensation with cresotic acid. *C.I. Mordant Blue 29* (C.I. 43825, *trade name* Chrome Cyanine BLA, Mitsui) is formed with 3-formyl-2,4-dichlorobenzenesulfonic acid, *C.I. Mordant Blue 47* (C.I. 43855, *trade name* Chrome Brilliant Blue GX, Mitsui) with 2,3,6-trichlorobenzaldehyde, and *C.I. Mordant Blue 3* [*3564-18-9*] (C.I. 43820, *trade name* Omegachrome Cyanine GR, Sandoz) with 2-formylbenzenesulfonic acid. *C.I. Mordant Violet 1* (C.I. 43565, *trade name* Omegachrome Violet RC, Sandoz) is obtained from 4-diethylaminobenzaldehyde and cresotic acid.

C.I. Mordant Violet 16 (**83**) (C.I. 43865, *trade name* Omegachrome Violet R, Sandoz). The most successful route for production of **83** involves condensation of 2 mol equivalents of cresotic acid with formaldehyde. The diphenylmethane derivative formed is oxidized to the diphenylmethanol with sodium nitrite. The diphenylmethanol condenses with a third mole

equivalent of cresotic acid to give the leuco dye, which is then oxidized to the dye with sodium nitrite.

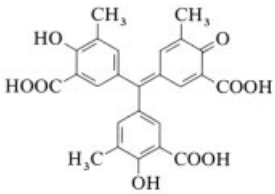

83

Production Example. To 4000 kg sulfuric acid monohydrate, 122 kg 40 wt % aqueous formaldehyde solution and 475 kg *o*-cresotic acid are added and stirred for 15 h at 80 °C. The diphenylmethane compound is precipitated by diluting with water to twice the volume. After collection by suction filtration the filter cake is washed with water until neutral and is dried at 90 °C under vacuum. Yield: 92 % methylenedi-*o*-cresotic acid.

Methylenedi-*o*-cresotic acid (177.5 kg) is added to 2250 kg 93 % sulfuric acid at 25 °C in 6 h, and oxidation to the benzhydrol is carried out with 20 kg sodium nitrite at 27 °C over 12 h. The benzhydrol is condensed to the leuco compound by adding 92 kg *o*-cresotic acid and stirring for 12 h at 25 °C. Another 30 kg sodium nitrite is added at 25 – 30 °C over 6 h. After oxidation is complete, 8000 L water and 1000 kg ice are added to the mixture. The dye is precipitated at 8 – 10 °C and pressed out, washed with water until neutral, and dried under vacuum at 95 °C. Yield: 1050 kg commercial product.

Other Uses. Hydroxytriarylmethane dyes sensitize photoconducting zinc oxide layers for electrophotography [253]. Their aluminum complexes have antihalo properties and can be used in filter layers for color photography [254].

6.4.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones)

Phenolphthalein exists in neutral and acid solutions as the colorless lactone (**84**) and in alkali as the red phenolate (**85**) with an absorption maximum at 555 nm.

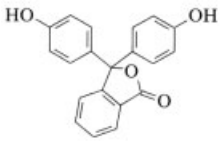

84

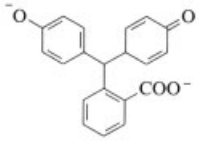

85

It is produced from phthalic anhydride and phenol in the presence of zinc chloride and toluenesulfonyl chloride. A sulfonic acid exchange resin can also be used as the condensation agent in the presence of triphenylphosphite to activate phenol [255].

Lactones derived from 1-hydroxynaphthalene-2-carboxylic acid can be used as filter dyes for instant color photography. The function of the carboxylate group is to form a hydrogen bond to the phenolic OH group and thus increase the pK_a, which determines the pH of the color change. The color is developed in daylight in an alkaline medium and protects the latent image from the nonselective entry of light.

Compound **86** forms a deeply colored dye salt in alkaline media, which becomes colorless when the pH is lowered and thus allows a picture to be reproduced in its true colors [256].

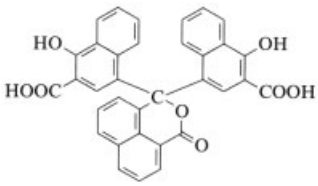

86

Symmetrical and unsymmetrical compounds are produced by condensing formylnaphthoic acid with an ester of 1-hydroxynaphthalene-2-carboxylic acid to give a lactone, which is converted to the intermediate **87** by oxidation and treatment with thionyl chloride.

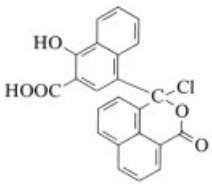

87

The filter dye is obtained by condensation with hydroxynaphthalene derivatives [257]. Intermediates can also contain acyloxy [258] or halophenoxy groups [259] instead of halogen as the leaving groups.

Tetrahalophenolphthaleins are used together with zinc oxide as photoconducting layers in electrophotography [260].

Phenol red (**88**) (phenolsulfophthalein) is obtained by condensation of *o*-sulfobenzoic acid anhydride with phenol.

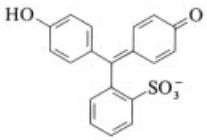

88

The corresponding tetrahydroxy compound is produced similarly from catechol [261]. Derivatives of phenolphthalein and phenol red with *halogen* and *alkyl substituents* are indicators that change color between pH 0.5 and 10.6. A thermographic process uses sulfophthalein dyes with volatile amines as salt formers. The action of heat alters the color shade [262]. Octabromosulfophthalein can be used as a sensitizer for zinc oxide photoconductors [263].

Sulfonamides (**89**), which are acylated at the amide nitrogen, can be used as filter and antihalo dyes in color photography. They have intense colors and are deacylated by alkali leading to cyclization involving the central carbon atom and, thus, irreversible decolorization to the sultam (**90**) [264].

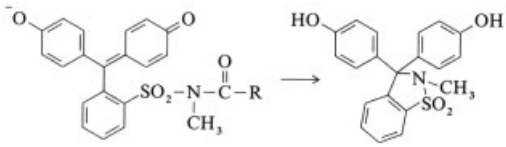

89 **90**

The sulfonamides are produced by treating leuco compounds of sulfophthalein, in which the hydroxyl groups are protected, with inorganic acid halides, then introducing the acylamino group, and finally oxidizing to the dye with cleavage of the protecting groups [97].

Hydroxytriphenylmethane dyes are also important in *biomedical analysis*. The monophosphate of phenolphthalein, in which the hydroxyl group is esterified with phosphoric acid, can be used to determine the activity of phosphatases [265]. The phosphate esters are produced from phenolphthalein and phosphorus oxychloride [265]. The monoether obtained from phenol red and galactose can be used for the detection of β-galactosidase [266].

7. Heteroaryldiarylmethane and Diheteroarylarylmethane Dyes

The aromatic rings in triarylmethane dyes can be replaced by one or two heterocycles. Examples are aniline derivatives, in which the nitrogen forms part of a heterocycle that is condensed to the phenyl ring *ortho* to the nitrogen to give a saturated five- or six-membered ring. However, indoles and carbazoles with unsaturated rings have proved particularly useful.

7.1. Cationic Dyes

7.1.1. Monoaminodiphenylheteroarylmethane Dyes

By condensation of aminobenzophenones with indole derivatives, dyes are obtained that dye polyacrylonitrile fibers blue, green, and black. Dyes containing phenylamino groups give dyeings with particularly good lightfastness [267]. The dyes can also be produced by oxidation of the corresponding leuco compounds with lead dioxide, sodium dichromate, iron(III) chloride [268], or oxygen [23], [188] in the presence of chloranil and iron or cobalt complexes. The substitution pattern of the second phenyl group has a considerable effect on the color shade. The introduction of chlorine *ortho* to the central carbon atom gives rise to a blue shift for steric reasons. A methoxy group in the *para*-position gives black color shades [3, pp. 381, 406 – 408].

7.1.2. Diaminodiphenylheteroarylmethane Dyes

Cationic dyes in this series are obtained by condensation of 4,4′-diaminobenzophenone and

its derivatives with α-substituted indoles [269]. The amino groups can also be introduced later into the dye skeleton by first condensing 4,4′-dichlorobenzophenone with an α-substituted indole and then replacing the *p*-chlorine atom with an amino group. Both aromatic amines (e.g., *o*-phenetidine [270]) and aliphatic ones (e.g., morpholine) can be used for this reaction. In the latter case the substitution reaction is carried out in the presence of aluminum trichloride [271].

7.2. Nonionic Derivatives (Triphenylmethanol Derivatives, Lactones, and Leuco Compounds)

Lactones. The reaction of phthalic anhydride with 2 mol equivalents of 1-ethyl-2-methylindole gives a *bisindolylphthalide* [*22091-92-5*] (**91**), which is an effective red color former for pressure-sensitive paper [272].

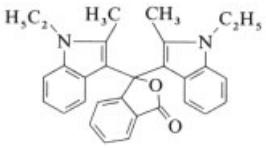

91

Phthalic anhydride can be reacted with 1-ethyl-2,2,4-trimethyl-1,2,3,4-tetrahydroquinoline to yield a green color former (**92**) [273].

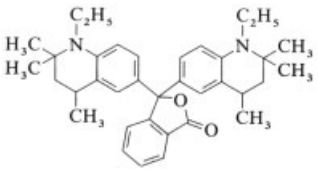

92

Color formers containing only *one heterocyclic ring* (**93**) are derived from substituted 2-(4-aminobenzoyl)benzoic acids, which are condensed with indole derivatives [274].

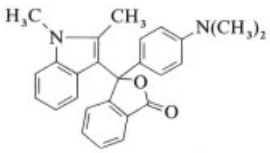

93

Color formers that contain a dimethylaminophenyl [275] or nitrophenyl ring [276], or a naphthalene group [277] instead of the phenyl ring, have been described as well.

Lactones containing heterocycles can also be obtained by using heterocyclic anhydrides instead of phthalic anhydride. Pyridinedicarboxylic acid anhydride can be reacted with aniline derivatives [278] and with equimolar quantities of an aniline and a heterocyclized aniline derivative [279] or an indole derivative [280], [281] to form an azaphthalide. Isomerically pure 4-azaphthalides containing indoles (**94**) are blue color formers, which do not exhibit self-coloration and are superior in lightfastness to crystal violet lactone [282].

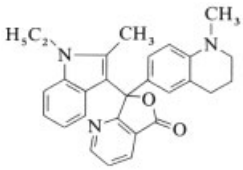

94

Lactones are also formed with pyrazine [283] and quinoxaline dicarboxylic acid anhydrides [284], which are obtained by condensation with 2 mol equivalents of an aniline derivative or contain an indole derivative besides an aniline derivative.

Ethers. Color formers based on a new principle are the chromenoazaindolizines (**95**) [285] and chromenothiazoles (**96**) [286].

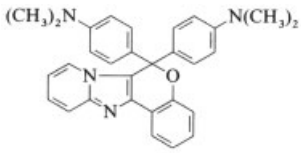

95

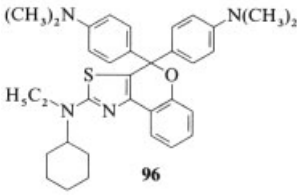

96

The chromenoazaindolizines are produced by reacting a 2-hydroxyphenacyl halide with 2-aminopyridine and condensing the *o*-hydroxyphenylazaindolizine with Michler's hydrol. After oxidation and addition of alkali, the phenolate oxygen attacks the central carbon atom to form a six-membered ring via an ether bridge. Chromenothiazoles can be obtained by the same reaction sequence. Instead of 2-aminopyridine, an *N,N*-dialkylthiourea is used and the 2-amino-4-(2-hydroxyphenyl)thiazole formed is used as starting material for the color former.

Leuco Compounds. Some leuco compounds are oxidized by air on acidic self-duplicating paper to give intense colors. Heterocycles play an important role in the synthesis of these dyes. Condensation of benzaldehyde derivatives with indole derivatives gives red color formers (**97**).

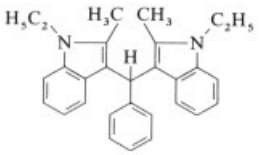

97

Heterocyclic aldehydes, such as 3-indolecarbaldehyde, furfural, 4-pyridinecarbaldehyde, and 2-pyrrolecarbaldehyde have also been used with success. The aniline derivatives give green and the indole derivatives red color formers [287].

Leuco bases of the diaminodiphenylindolylmethane dyes [288] have been proposed as color formers for heat-sensitive recordings, and those of the diaminodiphenylpyridylmethane dyes as electrochromic material [289]. 4-Aminophenyl-4′-alkoxyphenylindolylmethanes give a deep black coloration in air [290].

Of the heterocyclic aldehydes, *carbazolecarbaldehyde* is particularly important. Condensation of *N*-ethyl-3-carbazolecarbaldehyde with 2 mol equivalents of *N*-methyldiphenylamine gives a leuco compound (**98**) that turns blue on oxidation [291].

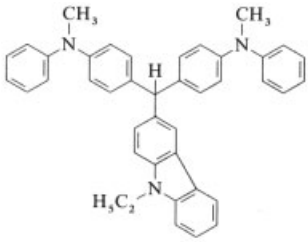

98

This compound belongs to the group of *secondary color formers*. Because of the necessary oxidation, its color develops slowly. However, for this reason, it is an ideal supplement to some primary color formers such as crystal violet lactone, whose colors develop immediately and then become weaker after a certain period. The combination of primary and secondary color formers gives optimum lightfastness. Color formers containing benzoannelated carbazoles [292] or two carbazole groups besides an aniline substituent [293] have also been described.

Leuco compounds obtained by condensing heterocyclic aldehydes with aniline derivatives are readily oxidized in the presence of peroxidase and can therefore be used to detect hydrogen peroxide [294].

8. Dyes with Several Chromophores

These dyes contain two or more triarylmethane chromophores or are polymeric. The triarylmethane chromophore can also be bonded to that of azo and oxazine dyes.

8.1. Bis(triarylmethane) Dyes

Condensation of terephthalaldehyde with aromatic amines in the presence of an acidic catalyst and subsequent oxidation gives dyes (**99**) in which the malachite green chromophore is bonded to that of Michler's hydrol blue. They can be used to dye or tint cotton, silk, leather, and paper [295].

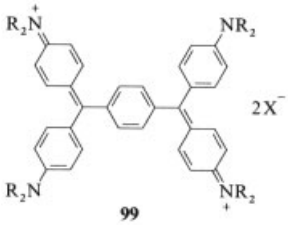

99

Bis(triaminotriphenylmethane) dyes are produced from 4,4′-diaminodiphenylmethane derivatives and bridged dianilines and subsequent oxidation with oxygen in the presence of (dihydrodibenzotetraaza[14]annulene)iron or -cobalt and chloranil as catalysts [296], [297]. Reaction of *N*-ethyl-*m*-toluidine with *p*-xylylene dichloride gives a dianiline, from which dye **100** is obtained. This dye has a very high affinity for wood-free paper and colors bleached sulfite pulp blue with a red cast.

Production Example [298]. A mixture of 508 parts 4,4′-methylenebis(*N,N*-dimethylaniline), 372 parts *N,N′*-diethyl-*N,N′*-di-3-tolyl-*p*-xylylenediamine, 1200 parts glacial acetic acid, 300 parts 1,2-propanediol, 10 parts chloranil, and 10 parts (dihydrodibenzotetraaza[14]annulene)iron is oxidized at 35 – 45 °C by passing through air with vigorous stirring until oxygen consumption can no longer be detected. Approximately 2400 parts of an ca. 45 % solution of the dye (**100**) are obtained.

Uses. Dimeric diaminotriphenylmethanes (leuco bases) can be used as photoconductors [136] and, in combination with quinones, for ink jet printing [294] and thermographic [299] recording materials.

8.2. Oligomeric and Polymeric Triarylmethane Dyes

Oligomeric and polymeric triarylmethane dyes are produced by treating formaldehyde with *N,N′*-ethylenebis(*N*-ethylaniline) in acetic acid solution, adding dimethylaniline, and oxidizing with oxygen in the presence of chloranil and (dihydrodibenzotetraaza[14]annulene)iron [300].

Polymeric triarylmethane dyes with a polyester, polycarbonate, polyurethane, or polyethylenimine chain as the backbone are also known [301]. A dye with a polyester chain can be produced from 4-dimethylaminobenzophenone, phosphorus oxychloride and the polymeric aniline **101**. The latter is made from *N,N*-di(2-hydroxyethyl)-3-methylaniline and terephthaloyl chloride.

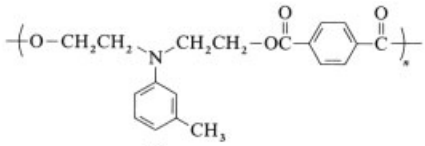

101

Triarylmethane dyes, which are bonded to bisphenol A diglycidyl ether via two secondary amino groups, are polycationic polymers that can be used as starting materials for the production of optical, electronic, photoelectronic, or electrochemical devices, films, or coatings [302]. Oligomeric and polymeric triarylmethanes (leuco bases) can be used as charge-transfer compounds in electrographic photoconductors [303].

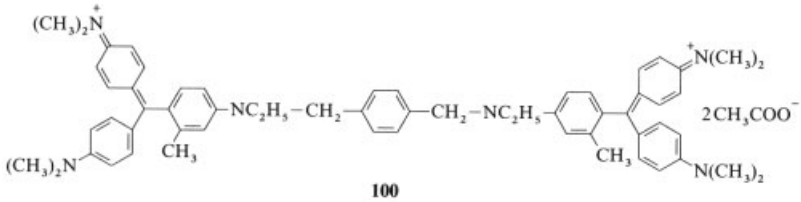

100

8.3. Triarylmethane Dyes with Other Chromophores

By coupling a triarylmethane dye with a salicylic acid derivative, an additional group that is capable of complexing metal ions is introduced. This type of dye can be produced by diazotizing 5-amino-2-chlorobenzaldehyde and coupling the product with *o*-cresotic acid. The formylazo compound thus obtained is condensed with 2 mol equivalents of *o*-cresotic acid in sulfuric acid, and the product is oxidized with nitrosylsulfuric acid [304]. Azodiaminotriphenylmethane dyes have been recommended as additives for galvanizing baths for copper plating [305].

By condensation of Michler's hydrol with oxazine dyes (which contain only one amino group and therefore have a hydrogen atom *para* to the azine nitrogen that can be replaced) followed by oxidation, biscationic dyes are obtained. These contain the diphenylmethane and oxazine systems as chromophores [306].

9. Acridine Dyes

Acridine dyes are alkyl derivatives of 3,6-diaminoacridine (**102**), 9-(4′-aminophenyl)-3-aminoacridine (**103a**), and 9-(4′-aminophenyl)-3,6-diaminoacridine (**103b**). Compound **103a** is identical to the oldest acridine dye, *chrysaniline,* discoverd by A. W. HOFMANN as a byproduct in the fuchsin synthesis.

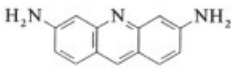

102

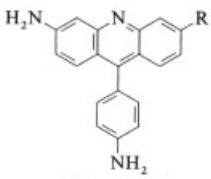

103a: R = H
103b: R = NH_2

Acridine dyes have low lightfastness and are not commercially important. They were formerly used for printing on tanned cotton because of their heat resistance. Their use as charge stabilizers [307], [308] and light-sensitive materials in electrophotography [309] has been described.

Acridine dyes derived from **102** are not synthesized from the acridines, but from 1,3-diaminobenzenes and aliphatic or aromatic aldehydes. The tetraaminodiarylmethane derivatives first formed eliminate ammonia to give leuco acridines. The latter are readily oxidized with atmospheric oxygen or iron(III) chloride to give acridine dyes. With acid, the dyes form salts that can give yellow, red, and brown shades. *Acridine yellow*, the 2,7-dimethyl derivative of **102**, is obtained in this way from 1,3-diamino-4-methylbenzene and formaldehyde. Other examples are *acridine orange 2 G* [*65-61-2*] (**104**) [C.I. Basic Orange 14, C.I. 46005, derivative: C.I. Solvent Orange 15 (free base)] and *euchrysin GGNX* [*4215-95-6*] (**105**) (C.I. Basic Yellow 9, C.I. 46040). The latter is produced from 1,3-diamino-4-methylbenzene and acetaldehyde with ring closure and subsequent oxidation by iron(III) chloride [310].

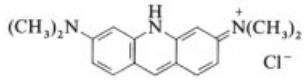

104

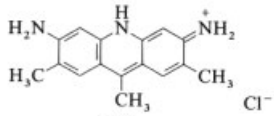

105

Acridine orange 2 G (zinc chloride double salt) can also be obtained by reduction of bis-(4-dialkylamino-2-nitrophenyl)methane with zinc in hydrochloric acid [311].

10. Xanthene Dyes

If the carbon atoms *ortho* to the central carbon in di- and triarylmethane dyes are bonded via an oxygen atom, xanthene dyes are obtained. Incorporation of this bridge has a pronounced hypsochromic effect. Although green and blue dominate in the triphenylmethane series, the characteristic color shade of xanthene dyes is red.

The xanthene dyes are classified according to the substituents in the 3- and 6-positions. Mono- and diaminoxanthenes and hydroxyxanthenes are important classes of dyes.

10.1. Ring-Closed Diphenylmethane Dyes (Pyronins)

Pyronin G and Pyronin B. The most widely used pyronin dyes are pyronin G [*92-32-0*] (**106**, R = methyl, C.I. 45005) and pyronin B [*2150-48-3*] (**106**, R = ethyl, C.I. 45010).

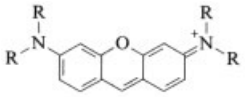

106

For both dyes, formation of the oxygen bridge occurs at the leuco stage. The leuco compound is obtained by condensation of formaldehyde with 3-dimethyl- or 3-diethylaminophenol. For pyronin G, nitrous acid has proved the best oxidizing agent and, for pyronin B, iron(III) chloride. Isopentyl nitrite has been recommended as oxidant for the production of particularly pure pyronin G [312].

Both dyes exist as the dye salt or the hydrol, depending on pH. The equilibrium position and distribution of the dye between water and hydrophobic solvents have been thoroughly investigated [313].

Uses. Pyronins G and B can be used for the quantitative determination of metals. Some metals form tetra- or hexavalent complex anions with halides, which combine with the dye cations to form ion pairs. These can be extracted from aqueous solutions with hydrophobic solvents and determined by visible spectroscopy. Complex anions that can be analyzed quantitatively by this method are $TlCl_4^-$ [314], $PdBr_4^{2-}$ [315], HgI_4^{2-} [316], PtI_6^{2-} [317], CdI_4^{2-} [318], TaF_6^- [319], and BF_4^- [320]. Sulfur dioxide [321] and carbon monoxide [322] can also be determined quantitatively through ion-pair formation. After reduction of $NaIO_3$ by CO or SO_2 in the presence of chloride ions, the complex anion ICl_4^- is formed in a quantity equimolar to that of SO_2 and CO. The ion pairs formed with pyronin G can be extracted with benzene.

Pyronin G has a higher affinity for RNA than for DNA [323], [324]. The selective coloration of RNA by pyronin G offers the possibility of determining the activity of ribonucleases and of following them spectroscopically [325].

Pyronins with heteropolyacids as the anion become positively charged through friction and can be used as pigments and charge stabilizers for toners in electrophotography [308].

Pyronins react with alkali cyanides, initially giving leuco compounds with a nitrile group in the 9-position. Oxidation gives blue cyano-substituted xanthene dyes (**107**), which can be used for dyeing cotton, silk [326], and polyacrylonitrile fibers [327].

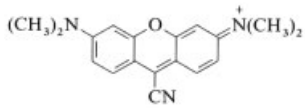

107

Both pyronins and their cyano-substituted derivatives are suitable as laser dyes [328]. How rapidly the photodegradation of xanthene nitriles occurs in various solvents and which mechanism predominates have been investigated spectroscopically [329].

Rhodamine S (**108**) (Basic Red 11, C.I. 45050) is obtained from 3-dimethylaminophenol and succinic anhydride [330].

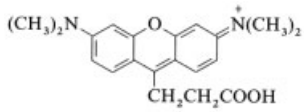

108

10.2. Ring-Closed Triphenylmethane Dyes

Unlike the pyronins, some dyes contain a third phenyl ring. This can be introduced via the aldehyde moiety, whereby leuco compounds are formed, which must then be oxidized. With phthalic anhydride, *o*-sulfobenzoic acid anhydride, or benzotrichloride, the dye is obtained directly.

10.2.1. Monoaminoxanthene Dyes

Monoaminoxanthene dyes, which have a carboxyl group *ortho* to the central carbon atom, exist as lactones and are widely used as color formers. The corresponding cationic dyes have not achieved any economic importance.

10.2.1.1. Cationic Dyes

Cationic monoaminoxanthene dyes are derived from the corresponding lactones, which are obtained by condensing 2-(4-dialkylamino-2-hydroxybenzoyl)benzoic acids with phenols (see Section 10.2.1.2). Conversion to a permanent cationically charged state involves conversion of the carboxyl into an amide [331] or an ester group. With β-naphthol a red dye (**109**) [332] is obtained and with *p*-diethylaminophenol a green dye (**110**) [333].

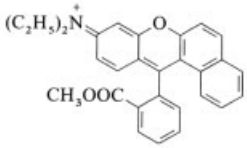

109

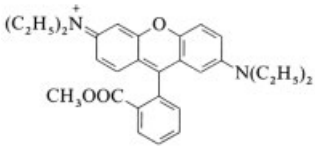

110

10.2.1.2. Fluorans

Fluorans are classed as xanthene lactones. They are spiro compounds in which a xanthene group is bonded to a benzofuranone moiety. These compounds have gained a strong position in the market because of their chromogenic behavior and the possibility that fluorans can be prepared in almost all colors, depending on the type of substitution.

Fluorans are used in the paper industry as color formers and are the color-forming components of self-duplicating and thermofax paper. Consumption is distributed approximately equally between the two applications and is estimated to be ca. 5000 t total. Self-duplicating paper has extensively replaced carbon paper and is now used for almost all forms. For production of self-duplicating paper, microdrops that contain solutions of the color former in liquid hydrocarbons are coated with a polymer layer to form microcapsules. The underside of a transmitting paper is coated with these microcapsules. When this paper is written on, the capsules break under the pressure and the color former is transferred to the upper side of the receiver paper, which is to provide the copy, and produces a color reaction with an acidic substance that coats the upper side of that paper. This acidic substance can be an inorganic material, such as bentonite or montmorillonite, or an organic such as a *p*-alkyl- or *p*-phenylphenol – formaldehyde polycondensate or an alkyl-substituted zinc salicylate.

Red and Orange Color Formers. *3-Diethylamino-7,8-benzofluoran* [*26628-47-7*] (**111**) is a lactone that becomes red on the action of acid and has practical applications.

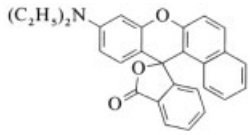

111

According to a patent published in 1985, homologues have been synthesized with an isobutyl or isopentyl group on the amino nitrogen [334].

In 1966, Fuji reported that condensation of 2-(4-diethylamino-2-hydroxybenzoyl)-benzoic acid with 2,4-dimethylphenol gives a colorless lactone, *3-diethylamino-4,6-dimethylfluoran* [*21934-68-9*] (**112**), which turns orange through the action of acid and can be used as a color former [335].

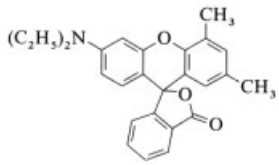

112

Other 2,4-disubstituted phenols were later used to produce fluorans [336]. Color formers giving orange and red dyes are 3-*N*-isobutylethylamino-7-chlorofluoran [334], 3-*N*-di-*n*-butylamino-6-methyl-7-bromofluoran [337], and fluorans with a phthalimido group in the 7-position [338].

Selective replacement of a chlorine atom in fluorescein chloride with cyclohexylamine or

another primary amine has also been used to synthesize fluorans [339].

The color former 3-cyclohexylamino-6-chlorofluoran [*26206-78-0*] (**113**) has useful technical properties attributed to the secondary amino group and thus to its reduced base paper coloration.

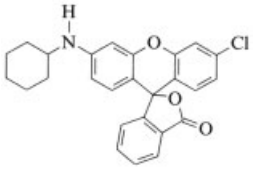

113

Green Color Formers. In 1967, Fuji discovered that fluorans with alkylamino groups in the 3- and 7-positions are green color formers. They can be obtained by condensation of 2-(4-diethylamino-2-hydroxybenzoyl)benzoic acid with a *p*-aminophenol derivative [340]. With *n*-dodecylamino-*p*-anisidine as the *p*-aminophenol derivative, *3-diethylamino-7-n-dodecylaminofluoran* (**114**) is obtained [341]. This color former is readily soluble in hydrophobic solvents. *3-Diethylamino-7-dibenzylaminofluoran* [*34372-72-0*] (**115**) has important practical applications [342].

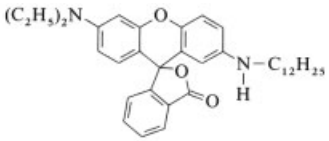

114

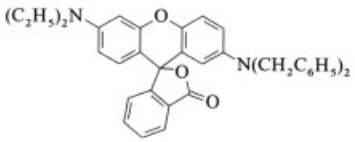

115

The color former becomes green on acidification. It is produced by reaction of 3-diethylamino-7-aminofluoran with benzyl chloride in a mixture of toluene and water in the presence of potassium carbonate. Benzyl *p*-toluenesulfonate can be used instead of benzyl chloride [343].

Black Color Formers. If 4-hydroxydiphenylamine is used for the condensation, 3-diethylamino-7-anilinofluoran, a black-green color former, is produced [344]. 3-Di-*n*-butylamino-7-*o*-chloroanilinofluoran [*85639-01-6*] (**116**) is an almost neutral black color former [345].

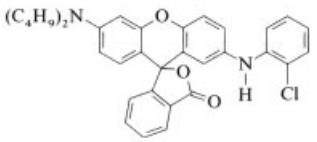

116

The aim of color former development has always been to produce black copies that can also be photocopied well. Initially, this could be achieved only by combining several color formers of different colors. Mixtures of orange, red, green, and violet give black. These mixtures are still the basis of self-duplicating paper with mineral clays as the acid component. In 1969, NCR found that fluorans with a methyl group in the 6-position are neutral black color formers [346].

3-Diethylamino-6-methyl-7-anilinofluoran. 3-Diethylamino-6-methyl-7-aminofluoran was arylated with 2-bromobenzoic acid and the product subsequently decarboxylated. The 3-diethylamino-6-methyl-7-anilinofluoran [*29512-49-0*] (**117**) obtained is black and has good application properties. Industrial production has been achieved by development of an economical synthesis of 2-methyl-4-methoxydiphenylamine [347].

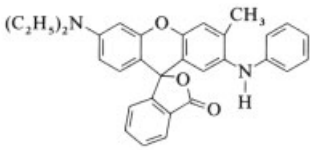

117

Synthesis of the color former itself is carried out in concentrated sulfuric acid. After condensation of 2-(4-diethylamino-2-hydroxybenzoyl)benzoic acid with 2-methyl-4-methoxydiphenylamine, a phthalide is formed with a hydroxyl and a methoxy group *ortho* to the central carbon atom. This phthalide cyclizes in alkali with loss of methanol to give the fluoran (**117**) [348]. Instead of 2-methyl-4-methoxydiphenylamine,

its *N*-formyl derivative can also be used. The formyl group is eliminated by hydrolysis of the phthalide before cyclization to the fluoran [349]. This color former is still important economically and is marketed worldwide by a number of suppliers.

Fluorans with a *pyrrolidine group* [350], [351] or an *N-methylcyclohexylamino group* [*55250-84-5*] [350] in the 3-position have similarly useful properties. Compounds with *N*-alkylarylamino groups in the 3-position [352] and aniline derivatives with an alkylated benzene ring in the 7-position [353], [354] have also been synthesized.

The black fluoran color formers are particularly useful in combination with organic developers. Whereas insufficient lightfastness was achieved with acidic or acidically modified clays, a neutral black with good fastness could be formed by using developers based on phenolic resin.

Fluorans for Thermoreactive Papers. The discovery that fluorans can also be used to produce thermoreactive paper opened up new possibilities. Incorporation of a fluoran and, for example, bisphenol A in a binder [e.g., poly(vinyl alcohol)] gives a colorless substance that forms a dye when heated [355]. The action of heat melts the thermoreactive substance, thus exposing the color former to the acidic phenol. No microcapsules are required for the production of thermoreactive paper. The starting materials are incorporated into the binder in finely ground form and applied to the carrier paper in a thin layer.

The first thermoreactive papers contained 3-diethylamino-6-methyl-7-anilinofluoran as the color former. After extensive optimization experiments, thermoreactive fax paper was produced, which satisfied most requirements. The solubility of the color former in the binder, though low, and its undesired reaction with phenol in the cold were disadvantages that made the production of pure white paper difficult. Further product development was aimed at fluorans with lower solubility in the binder. Derivatives with a *N*-ethyltetrahydrofurfurylamino group [*102232-11-1*] [356] and particularly an di-*n*-butylamino group [*89331-94-2*] (**119**) [358] in the 3-position, instead of the diethylamino group, have proved successful.

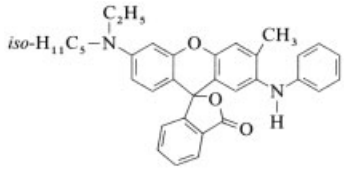

118

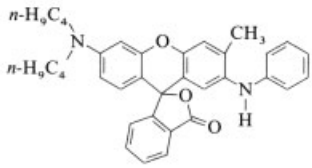

119

In the search for optimized products, fluorans have also been produced that have *N*-ethylisobutylamino [359], *N*-isobutylalkoxyalkylamino[360], *N*-propylcyclohexylamino[361], diarylamino [362], and 3,5,5-trimethylhexahydroazepinyl groups [363] in the 3-position. Phenylethyl [364], pyridyl-2-ethyl [365], and alkyl groups with more than nine carbon atoms [366] have also been introduced as substituents. New compounds have likewise been obtained by reductive alkylation of the amino group with ketones [367].

Fluorans which absorb in the IR region and can therefore be used for printing bar codes contain a polyanilino group in the 3-position [368]. Other fluorans have been obtained by introducing a methyl [369] or an ethoxypropyl group [370] in the 2-position and by replacing the 6-methyl group with phenyl, benzyl [371], halogen [372], and methoxy groups [373].

The anilino group in the 7-position has also been modified. It has been substituted by lower alkyl groups [353], [374] and by C_5 to C_8 alkyl groups [354] in the *p*-position and by a 1-naphthylamino group [375]. Alkyl groups can be introduced into the color former molecule by reaction with secondary and tertiary alcohols [376]. Substitution with ketones has led to *fluoran dimers* [377], [378]. Fluoran dimers can also be produced by using methylenebisphthalic acid anhydride as the starting material [379]. *tert*-Butyl [380], alkoxy or aryloxy [381], carboxyl, and alkoxycarbonyl groups [382] have also been described as substituents on the benzofuranone part of the molecule.

10.2.2. Diaminoxanthene Dyes

Rhodamine dyes are the most important diaminoxanthene dyes. They are xanthenes that contain a 2-carboxyphenyl ring on the central carbon atom. Diaminoxanthene dyes are unimportant in dyeing textiles because their lightfastness is usually insufficient. They are used mainly to dye paper and, after precipitation with heteropolyacids, to produce printing inks. Ca. 2000 t of rhodamine dyes are produced and processed annually.

10.2.2.1. Cationic Dyes

Symmetrical rhodamine dyes are obtained in a melt reaction, in which 2 mol equivalents of a 3-aminophenol are condensed with 1 mol equivalent of phthalic anhydride at ca. 180 °C. Unsymmetrical dyes are produced from a 2-(4-dialkylamino-2-hydroxybenzoyl)benzoic acid, which can be obtained by reacting equimolar quantities of a 3-dialkylaminophenol with phthalic anhydride in nonpolar solvents such as xylene or toluene. The benzoylbenzoic acid is condensed with another 3-dialkylaminophenol in concentrated sulfuric acid at ca. 50 °C [383].

Rhodamine B [*81-88-9*] (**120**) (C.I. Basic Violet 10, C.I. Solvent Red 49, C.I. 45170, BASF) is obtained by melting 3-diethylaminophenol with phthalic anhydride at 170 – 180 °C [384]. If the reaction is carried out in *o*-dichlorobenzene, the dye can be obtained in particularly pure form [385].

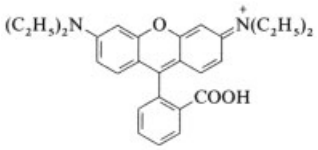

120

The dye is readily soluble in water and gives a blueish red coloration. It is characterized by its intense fluorescence.

The ethyl ester of rhodamine B (Rhodamine 3 B, C.I. Basic Violet 11, C.I. 45175) is obtained by treating rhodamine B with diethyl sulfate in the presence of base in aprotic solvents such as butyrolactone [386].

Because of their high solubility the *hydroxyethyl esters* (rhodamine 2 BU) are particularly suitable for the preparation of liquid formulations. They are obtained by reacting rhodamine B with a dialkyl sulfate in glycol and subsequent heating, whereby transesterification with the solvent glycol takes place [387]. For the production of red ballpoint pastes, esters of *long-chain alcohols*, such as nonanol, have been developed [388].

Rhodamine 2 C base is produced from 3-ethylamino-4-methylphenol and phthalic anhydride in a melt reaction. The dye is red with a yellow cast, but is unimportant because of its poor water solubility. However, the esters of rhodamine 2 C base are readily water soluble and even allow the preparation of liquid formulations [389].

Rhodamine 6 G [*3056-93-7*] (**121**) (C.I. Basic Red 1, C.I. 45160, BASF) is the ethyl ester of rhodamine 2 C base. It is obtained by esterification with diethyl sulfate.

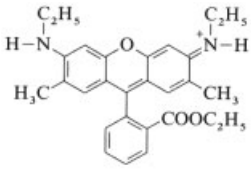

121

Production Example. Stepwise, 600 kg 3-ethylamino-4-methylphenol and then 1016 kg phthalic anhydride are added to 3310 kg *o*-dichlorobenzene in an 8100-L titanium-lined steel vessel. The temperature is raised to 177 – 180 °C within 2 – 3 h, and the mixture is stirred at this temperature until condensation is complete (6 h). After cooling to 100 °C, 2080 L water is added, and a mixture of water and *o*-dichlorobenzene is distilled off. For complete removal of the *o*-dichlorobenzene, another 1100 L water is added during distillation over 3 – 4 h. When all the *o*-dichlorobenzene has been removed, water is added until the vessel is filled to 3/4. Then, 1820 kg 30 wt % aqueous sodium hydroxide solution is added over 1 h. The mixture is heated to 85 – 90 °C, and the pH is adjusted to 12 by addition of 20 kg 30 wt % aqueous sodium hydroxide solution. After stirring for 4 h at 85 – 90 °C,

the base is allowed to settle and the aqueous phase is decanted. After addition of water and 154 kg sodium hydroxide as a 30 % solution, the mixture is heated again to 85 – 90 °C for 2 h to completely hydrolyze the phthalic acid complex. Rhodamine 6 G base is then isolated and washed with warm water at 40 – 45 °C until the alkali is completely removed.

After 2950 kg *o*-dichlorobenzene has been charged to a 7500-L titanium-lined steel vessel, 1020 kg rhodamine 6 G base is added as a moist filter cake. The temperature is raised slowly to 95 – 110 °C until water begins to distill. The temperature is increased to 140 °C over 3 – 4 h and then kept at 155 – 160 °C for 0.5 h to completely remove water. Next, 605 kg diethyl sulfate is added at 155 – 160 °C over 10 h. The mixture is stirred for 4 h at this temperature and then cooled to 100 °C. After addition of 2840 L water, *o*-dichlorobenzene is distilled off, and 1135 L water is added during the distillation. After 7 – 8 h the removal of *o*-dichlorobenzene is complete. The pH of the solution is 2 – 3 as a result of the hydrolysis of excess diethyl sulfate. The reaction mixture is transferred to a 26 000-L stirred tank and diluted to 23 000 L with water. Then, 54 kg sodium acetate is added at 75 – 80 °C, and the pH is adjusted to 4 with 180 kg sodium hydrogencarbonate. After addition of clarifying agents the solution is filtered at 75 – 80 °C through a press and the filtrate is divided into two parts. To each part, 680 kg 30 % hydrochloric acid is added, and the dye is salted out with 3 % sodium chloride at 40 °C. After stirring for several hours, rhodamine 6 G is isolated by suction filtration at 35 – 40 °C and the dye is dried at 85 –90 °C. Yield: 1020 kg rhodamine 6 G.

The dye can also be isolated as a granulate if precipitation is carried out at high temperature and the droplets formed solidify to solid spheres on cooling [390].

With heteropolyacids, water-insoluble rhodamine dyes can be precipitated; these are used as pigments in the printing industry. Phosphomolybdotungstic acid has proved particularly effective. If precipitation is carried out in the presence of special amines [391], pigments are obtained with optimum color strength, brilliance, and good dispersibility.

If rhodamine 2 C base is treated with phosphorus oxychloride and then with a primary amine, surprisingly no amide or lactam is formed but a dimeric imide in which the amine nitrogen forms a bridge between the carboxyl groups of two dye molecules. If an additional basic group is introduced via the amine, as in *N,N*-dimethylpropylenediamine, dyes are obtained that have a high affinity for paper [392].

When *N*-aryl-substituted rhodamine dyes are reacted with formaldehyde and an imidazole derivative, basic groups in the form of an imidazolylmethyl group are introduced, also giving dyes with a high affinity for paper [393].

Rhodamine dyes with polyoxyalkylene groups on the nitrogen have a fugitive character. They do not adhere to textiles or human skin and are therefore particularly suitable for marking fibers and yarns in the production of woven fabrics and tufting articles because the marks can later be removed [394].

Because of their brilliance, rhodamine dyes can also be used in *electrophotography*. They can be processed to give red toners, which – together with yellow and blue toners – allow the production of colored photocopies. Dyes with tetrafluoroborate [395] and the anions of heteropolyacids [396] and trifluoromethanesulfonic acid [397] are particularly useful.

Rhodamine dyes with two carboxyl groups are obtained by treating 3-aminophenols with 1,2,4-benzenetricarboxylic acid anhydride. They are used for fluorescent labeling of nucleotides in chain termination DNA sequencing procedures [398].

Laser Dyes. Because of their high quantum yield, rhodamine dyes can be used as laser dyes (→Laser Dyes, Chap. 2.4.) [399]. When the mobility of the *N*-alkyl groups is hindered by bonding to the benzene ring, the quantum yield is high [400]. This applies particularly to dyes based on *julolidin* (**122**) [401].

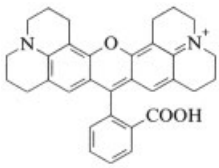

122

In the search for laser dyes with enhanced photostability, products have been synthesized with 2,2,2-trifluoroethyl substituents on the nitrogen atoms [402].

Aryl-substituted diaminoxanthene dyes give a coloration that is shifted to the blue range compared to those with alkyl substituents. Aryl-substituted dyes can be obtained by treating benzotrichloride with resorcinol [45] and replacing the chlorine atoms in 3,6-dichloro-9-phenylxanthylium chloride with aromatic amines [403]. Because of steric hindrance, red with only a slightly blue cast is obtained with 2,6-dimethylaniline. Aryl-substituted rhodamine dyes can also be obtained from phthalic anhydride and aryl-substituted 3-aminophenols [332], [404].

10.2.2.2. Nonionic Derivatives (Lactones and Related Compounds)

Lactones. Depending on the pH and the nature of the solvent, *rhodamine B* exists either as a cation, a zwitterion, or a colorless lactone. In protic solvents, the equilibrium position between the zwitterion and the lactone is influenced by hydrogen bonds and by the dielectric constant and polarizability of the solvent. In aprotic solvents, depending on the acidity only the lactone or the cation can be detected [405].

Isolating the pure, colorless lactone of rhodamine B is difficult. The equilibrium with the cationic dye can be shifted in the direction of the lactone, however, if the amine nitrogens of the xanthene dye carry phenyl substituents. Thus, tetrachlorophthalic anhydride and 3-hydroxy-*N*-methyldiphenylamine give a colorless compound (**123**) that becomes purple blue on acidic clay [406].

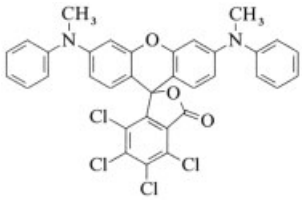

123

Lactones of diaminoxanthenes with two aryl groups on one or both nitrogens are also suitable as color formers [407]. The bis(diphenylamino)xanthenes can be obtained by phenylation of bis(monophenylamino)xanthenes with iodobenzene or by treating 3-hydroxytriphenylamine with phthalic anhydride [408]. They develop a blue color on acidic clays.

Valuable red color formers (**124**) are derived from 2,4-bisazarhodamine dyes. They lactonize to colorless compounds and can be produced by condensation of 2-(4-diethylamino-2-hydroxybenzoyl)benzoic acid with, e.g., 2-dibutylamino-6-methylpyrimid-4-one [409].

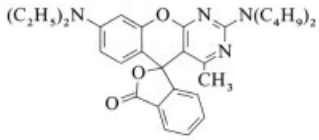

124

Lactams. Unlike most lactones, rhodamine lactams are stable compounds. The equilibrum can be shifted in the direction of the dye when aryl groups with electron-withdrawing substituents are present in the molecule. One of the first color formers, with which two-color effects were achieved together with crystal violet lactone, was the lactam (**125**) [*29199-09-5*] obtained from rhodamine B and 4-nitroaniline. It was described in 1955 by NCR [410].

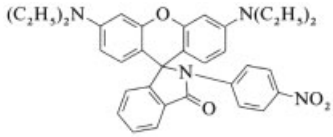

125

Thio- and dithiolactones are colorless. They are obtained from rhodamine B and phosphorus pentasulfide. The thiolactone ring can be opened only in the presence of Lewis acids [411].

Sultams have an even lower tendency to color formation than lactams. Sultams are derived from sulfonamides of primary amines and are formed by attack of the amide nitrogen on the central carbon atom with loss of a proton. The sultam ring can be forced to open only if

the amide nitrogen can be substituted in an intramolecular reaction. This is the case with sultams in which the amide nitrogen has been arylalkylated with bivalent compounds such as 1,8-bis(bromomethyl)naphthalene compound (**126**) [412] or pentafluorophenacyl bromide [413]. At higher temperature the amide nitrogen is arylalkylated or arylated to form a six- or five-membered ring, thus liberating the color chromophore (**127**).

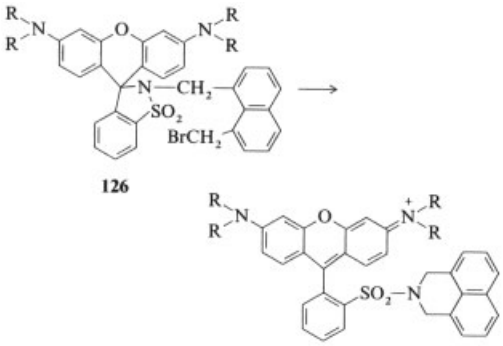

126 **127**

Sultams in which one amino group of an *o*-phenylenediamine is part of the sultam ring and the second amino group has been acylated with phenyl chloroformate (**128**) also cleave to form a dye at elevated temperature. The driving force is further acylation of the amide nitrogen, with loss of phenol and formation of a benzimidazolone (**129**) [414].

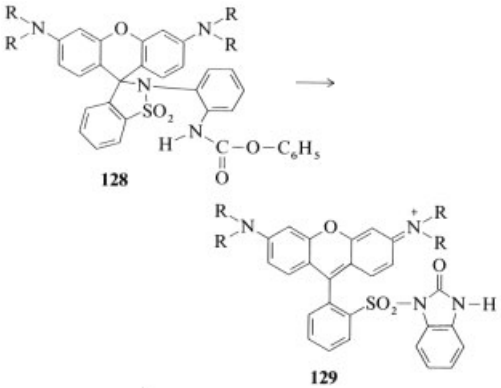

128 **129**

Whereas sultams with bivalent substituents liberate a dye in a thermal reaction, some xanthenes can undergo fragmentation in a thermally induced reaction that results in the opposite effect (i.e., a negative image). Suitable compounds for this application are xanthenes that have a sulfonamide group *ortho* to the central carbon atom and in which the nitrogen carries a methyl and a *tert*-butoxycarbonyl group. At elevated temperature, *tert*-butyl alcohol and car-bon dioxide are liberated and the *N*-methylsulfonamide cyclizes to the colorless sultam [415].

Sulfonamides of xanthene dyes that have an alkyl and an acyl group on the amide nitrogen (**130**) are readily deacylated with alkali and cyclize to colorless sultams (**131**). They are therefore valuable correction filter and antihalation dyes in photographic materials. During illumination, they contribute to a sharp image with accurate colors and can be completely bleached out during development [416].

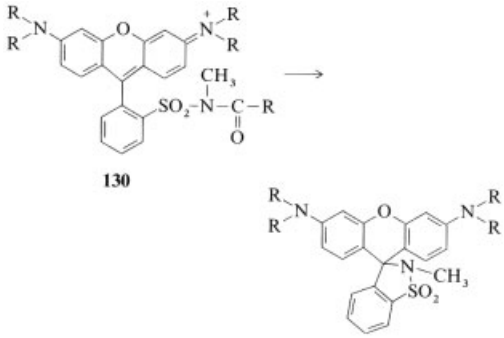

130 **131**

10.2.2.3. Anionic Dyes

Sulforhodamine B [*3520-42-1*] (**132**) (C.I. Acid Red 52, C.I. 45100, *trade name*: Duasyn Acid Rhodamine B 01, Hoechst), the first anionic xanthene dye, was synthesized by STEINER in 1906. This dye is produced by condensation of 2 mol equivalents of 3-diethylaminophenol with 1 mol equivalent of 4-formyl-1,3-benzenedisulfonic acid in sulfuric acid solution to give the leuco triphenylmethanedisulfonic acid. Heating in 90 % sulfuric acid to 135 °C causes dehydration to the leuco xanthene acid,

which is then oxidized to the dye with sodium dichromate [417].

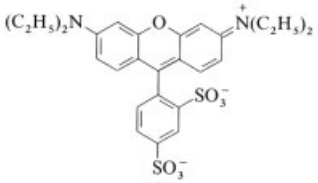

132

Sulforhodamine G (C.I. Acid Red 50) is obtained similarly from 3-ethylamino-4-methylphenol. For both dyes, iron(III) chloride can be used in the oxidation instead of sodium dichromate. The sulforhodamines are brilliant red, fluoresce, and can be used as laser dyes [418].

The reaction of 2-sulfobenzoic acid anhydride with resorcinol in the presence of phosphorus oxychloride gives the *3,6-dichloro-9-* (*2-sulfophenyl*)*xanthylium betaine* (**133**).

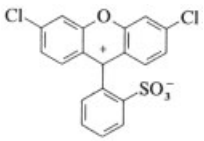

133

Both chlorine atoms can be replaced by substituted anilines. This reaction gives monosulfonic acids of the diaminoxanthene series, which, as sparingly soluble betaines, have pigment character and can be used to produce printing inks [419]. Sulfonation yields soluble dyes that give brilliant red and violet shades on wool [44].

Violamines are produced by nucleophilic substitution of chlorine. Dichlorofluoran lactone, which is used as an intermediate, is formed from fluorescein (see Section 10.2.3) and phosphorus oxychloride. After replacing both chlorine atoms with an aromatic amine and introducing a sulfonic acid group, violet dyes with high brilliance are obtained. With *o*-toluidine *Fast Acid Violet ARR* (**134**) (C.I. Acid Violet 9, C.I. 45190, *trade name* Sandolan Violet E-R, Sandoz) is obtained.

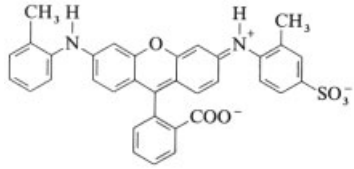

134

10.2.3. Hydroxyxanthene Dyes

Reaction of phthalic anhydride with aromatic polyhydroxy compounds gives hydroxyxanthene dyes.

Fluorescein. The most well known hydroxyanthene is fluorescein [*518-47-8*] (**135**) (C.I. Acid Yellow 73, C.I. Solvent Yellow 94, C.I. 45350, *trade name* Basacid Yellow 226, BASF). It is produced industrially by melting resorcinol and phthalic anhydride at 200 °C [420].

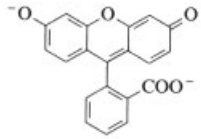

135

Production Example. Phthalic anhydride (35 kg) is added gradually to 50 kg molten resorcinol. The reaction mixture is heated to 200 – 205 °C and the melt is kept at this temperature for 30 h. After cooling, the crude melt is ground up and stirred with dilute sodium hydroxide solution. Undissolved material is removed by filtration, and fluorescein is precipitated with hydrochloric acid. Yield: 72 kg.

Uses. Fluorescein is not important in textile dyeing. Its alkali salt is used to dye liquid bath preparations. Ca. 250 t of fluorescein are produced annually. Esters of fluorescein are used for the quantitative fluorimetric determination of the activity of ester-cleaving enzymes since the fluorescence increases as cleavage of the ester bond progresses [421]. The thyroxine conjugate of 6-carboxyfluorescein (bonding to thyroxine via the carboxyl group) can be used in biochemical analysis to determine quantitatively the thyroxine content in blood. In an immunoassay the conjugate is bonded to an antibody in competition with the thyroxine to be determined. The intensity of the polarized fluorescent light emit-

ted by the conjugate is therefore inversely proportional to the quantity of free thyroxine [422].

Eosin Dyes. The reaction of fluorescein with bromine gives eosin dyes. *Eosin G* [*17372-87-1*] (**136**) (C.I. Acid Red 87, C.I. Solvent Red 43, C.I. Pigment Red 90, C.I. 45380) was produced by H. Caro and A. von Baeyer in 1874 by bromination of fluorescein in alcoholic solution [423].

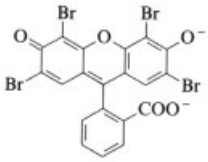

136

Eosin G is *used* to dye cosmetics [424]. The aluminum salt and the lead salt of eosin G are pigments that are used as printing inks for books and newspapers, mainly in the United States. The consumption is ca. 200 t/a.

Erythrosin B [*16423-68-0*] (C.I. Acid Red 51, C.I. Food Red 14, C.I. Pigment Red 172, C.I. Solvent Red 140, C.I. 45430) is a tetraiodo derivative of fluorescein and is obtained by treating fluorescein with iodine and iodic acid in alcohol. The dye is isolated as the disodium salt. It is used in the food and cosmetic industries. The market volume is ca. 600 t/a.

Hydroxyxanthene dyes can also be produced from di- and tetrachlorophthalic acid. Bromination gives the *phloxin* and iodination the *rose bengal dyes*.

Rose bengal GTO [*18265-55-9*] (**137**) (C.I. Acid Red 93, C.I. Solvent Red 47, C.I. 45435) is the tetraiodinated derivative of the dye produced from dichlorophthalic acid [425].

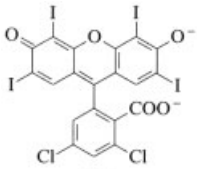

137

Rose bengal, erythrosin, and eosin are photosensitizers that are used, e.g., as catalysts for curing epoxy and poly(methyl methacrylate) resins [426]. They also have phototoxic properties and act as insecticides toward adult insects and larvae [427].

Gallein (**138**) (C.I. Mordant Violet 25, C.I. 45445) is produced by condensation of 2 mol equivalents of pyrogallol with 1 mol equivalent of phthalic anhydride [428].

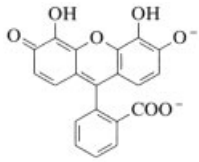

138

Gallic acid can be used instead of pyrogallol. The dye colors chrome mordanted wool pure violet.

11. Thioxanthene Dyes

Treatment of bis(4-dimethylaminophenyl)-methane with the reaction product of sulfur and oleum gives the violet thioxanthene dye thiopyronin [*2412-14-8*] (**139**) [429].

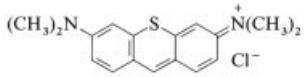

139

Thioxanthene dyes can be used in photography [430] and are of interest as sensitizers in photodynamic therapy [431].

3,6-Bis(dimethylamino) thioxanthone (**140**) can be used as starting material for the production of benzopyranothioxanthene dioxides, such as **141**, which can be used as color formers [432].

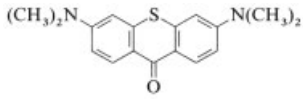

140

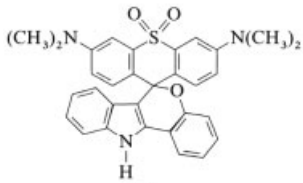

141

Similar to the xanthene series, thioxanthene lactones are obtained from phthalic anhydride and substituted 3-aminophenylthiols. The thioxanthene lactones can be used as color formers [433].

12. Other Ring-Closed Triarylmethane Dyes

12.1. Cationic Dyes

Fluorene Derivatives. If the two carbon atoms *ortho* to the central carbon in triphenylmethane dyes are bonded directly, fluorenes are obtained. Ring closure gives a significant shift of the absorption maximum toward longer wavelengths (see Section 1.2).

The dye **142** formally derived from crystal violet was produced initially from 3,6-bis(dimethylamino)fluorenone and lithium 4-dimethylaminophenylide [434].

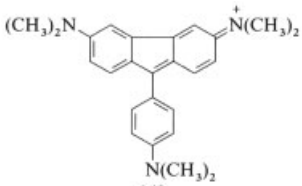

142

In an industrially more suitable synthesis, crystal violet is treated with sulfuric acid in the presence of catalytic quantities of phosphate at >150 °C. A sulfonic acid group is initially introduced *ortho* to the central carbon atom. Rearrangement and hydrolysis under oxidizing conditions lead to the fluorene dye (**142**) [435], which colors polyacrylonitrile fibers blue with a green cast [436].

Dyes with an Isopropyl Bridge. Triphenylmethane dyes that are bridged by an isopropyl group are obtained by condensation of Michler's hydrol with 3-dimethylaminopropenylbenzene and subsequent oxidation of the leuco compound. Dye **143** can be used in photography [437].

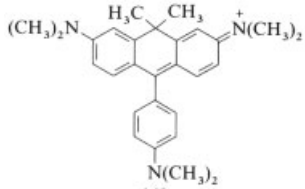

143

Dye **144** is obtained similarly by ring closure of 4,4′-bis(dimethylamino)-2-propenyldiphenylmethane.

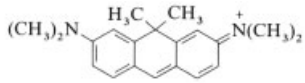

144

As with Michler's hydrol, the absorption maximum of this dye is 607 nm. Because of restricted rotation, however, it exhibits intense red fluorescence. It emits at 640 nm and is used as the active medium in dye lasers both in solution [438] and in polymer resins [439].

12.2. Nonionic Derivatives (Triphenylmethanol Derivatives and Lactones)

Fluorene Lactones. Because they absorb in the near IR, fluorenes have attracted interest as color formers. After synthesis of a derivative of crystal violet lactone with an amino group *ortho* to the central carbon atom (**145**), the amino group is diazotized and the diazonium salt is cyclized to the fluorene (**146**) in the presence of copper powder [440].

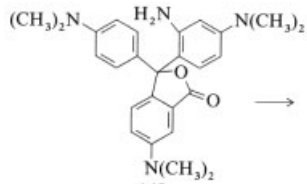

145

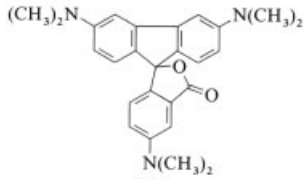

146

Fluorene lactones with a pyridine or pyrazine ring in the 9-position [440] and derivatives with alkenylamino and substituted alkylamino groups [441] have also been produced by this method. Higher-purity products with improved yields can be obtained if the cyclization is carried out in phosphoric instead of sulfuric acid [442].

An alternative process involves the cyclization of phthalides with aluminum trichloride. Under nonoxidizing conditions, a leuco fluorene dye can be produced from crystal violet [443]. If the melt reaction of triaminotriphenylmethane lactones, such as crystal violet lactone, is carried out in the presence of urea and air, oxidation to the fluorene lactone occurs in addition to cyclization. Addition of aluminum trichloride hexahydrate suppresses dealkylation and yields purer products [444].

Cyclization of 3-(4′-dimethylaminophenyl)-3-(4′-dimethylaminonaphthyl)-6-dimethylamino-phthalide does not give a fluorene. Instead, cyclization occurs with the 5-position of the 1-naphthylamine group. A spirobenzanthracenephthalide (**147**) is obtained [445].

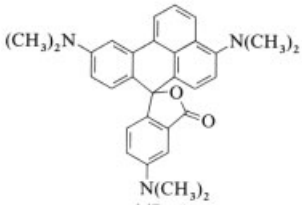

147

Dihydroanthracenes. By oxidation with atmospheric oxygen, the colorless dihydroanthracenes (**148**) are converted to dyes. They are produced by condensing Michler's hydrol with 3-dimethylaminobenzyl alcohol and cyclizing the leuco tris(dimethylaminophenyl)methane dye, which has a hydroxymethyl group *ortho* to the central carbon atom, in the presence of acid [446].

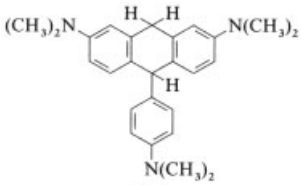

148

13. Toxicology

Toxicological data of important triarylmethane and diarylmethane dyes are summarized in the following:

Michler's ketone has been found to induce hepatocellular carcinomas and hemangiosarcomas in rats and mice [447].

Auramine. Many authors found that auramine showed no mutagenic effects in the Ames test [448–450]. A weak mutagenic activity was observed only in strains TA 98 and TA 1537 with metabolic activation. This activity is suggested to be attributed to a contaminant.

Technical-grade auramine O has been found to induce hepatomas in rats and mice on oral administration [451], [452].

In a test conducted by Kirsch and coworkers [453], auramine, however, showed no activity when fed to rats for two years at lower dose levels of 50, 100, and 200 mg/kg in the diet. The inactivity at lower doses suggests that impurities in the commercial dye preparation may be responsible for the activity observed in tests at the higher 1000-mg/kg dose [451], [452].

Other studies [454] also strongly suggest that contaminants in commercial auramine dye preparations may be responsible for mutagenic activity. A commercial preparation was found to induce DNA damage. The preparation contained as a major contaminant Michler's ketone, which is known to have carcinogenic activity. More importantly, tests on the same systems after purification of the dye sample by TLC were negative.

Case and Pearson established [455] an increased occurrence of bladder cancer among workers involved in the production of auramine O and fuchsin in the United Kingdom between 1910 and 1952. However, in a detailed study in which persons involved in auramine production in the BASF plant at Ludwigshafen participated, the trends or observations concerning auramine discussed by Case and Pearson could not be confirmed [453], [456]. According to this study there is no evidence that auramine should be regarded as carcinogenic.

Malachite green exhibits acute oral toxicity. Inhibition of algal growth, acute lethal toxicity

in fish, and inhibition of activated sludge have also been observed [457].

Pararosaniline induces tumors in rats and mice. It can definitely be classed as having mutagenic potential [457].

Methyl violet inhibits the growth of algae and is toxic to fish [457].

Crystal Violet. In animal experiments, crystal violet exhibits acute oral toxicity. The toxic effects occurring in humans, such as nosebleeds, contact sensitization, nausea, gastrointestinal irritation, and mucosal ulceration, refer to case reports and are not the results of systematic clinical tests. Crystal violet is toxic to activated sludge [457]. Long-term feeding experiments with rats and mice have led to increased occurrence of tumors in various organs [458].

Acridine Dyes. Toxicological investigations indicate that aminoacridines such as acridine orange 2 G (**104**) have mutagenic potential. Light either enhances the effect or is necessary for the activity observed [457].

Rhodamine B has been shown to cause point mutations in the Ames test and chromosome aberrations in a number of in vitro systems. Long-term feeding studies [459] in rats and mice have produced positive oncogenic responses, while short-term carcinogenicity assays [451], [459–462] have been predominantly negative. The possibility has been raised that contaminants contribute to the tumorigenicity.

14. References

1. J. Griffiths: *Colour and Constitution of Organic Molecules,* Academic Press, London 1976, p. 250.
2. J. Fabian, H. Hartmann: "Light Absorption of Organic Colorants," *Reactivity and Structure Concepts in Organic Chemistry,* vol. 12, Springer Verlag, Berlin 1980, p. 137.
3. *Ullmann,* 4th ed., **23,** 381.
4. *Kirk-Othmer,* 4th ed., **23,** 399.
5. H. A. Staab: *Einführung in die theoretische Chemie,* 3rd ed., Verlag Chemie, Weinheim 1962.
6. S. Dähne, *Z. Chem.* **5** (1965) 441 – 451.
7. H. Zollinger: *Color Chemistry,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1991, p. 71.
8. J. Griffiths, K. J. Pender, *Dyes Pigm.* **2** (1981) 37.
9. S. F. Beach et al., *J. Chem. Soc. Perkin Trans. 2* 1989, 1087.
10. G. Hallas et al., *J. Chem. Soc. Perkin 2* 1977, 450.
11. G. Hallas, D. E. Grocock, J. D. Hepworth, *J. Soc. Dyers Colour.* **86** (1970) 200.
12. A. S. Ferguson, G. Hallas, *J. Soc. Dyers Colour.* **87** (1971) 187.
13. G. Hallas, R. M. Potts, *J. Chem. Soc. Perkin Trans. 2* 1974, 59.
14. A. Barker, C. C. Barker, *J. Chem. Soc.* 1954, 1307.
15. J. Fabian, R. Zahradnik, *Angew. Chem.* **101** (1989) 693; *Angew. Chem. Int. Ed. Engl.* **28** (1989) 677.
16. C. Reichardt: *Solvents and Solvent Effects in Organic Chemistry,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1988, p. 107.
17. T. Gessner, unpublished results.
18. M. Okawara, T. Kitao, T. Hirashima, M. Matsuoka: "Physical Sciences Data," in: *Organic Colorants,* vol. 35 Kodansha, Tokyo; Elsevier, Amsterdam 1988, p. 298.
19. W. Michler, *Ber. Dtsch. Chem. Ges.* **9** (1876) 400, 716.
20. O. Fischer, *Ber. Dtsch. Chem. Ges.* **10** (1877) 1625.
21. Ciba, FR 1 097 541, 1953.
22. Ciba, FR 1 098 497, 1953.
23. BASF, DE 2 138 931, 1971 (H. Kast, H. Baumann, U. Mayer, A. Oberlinner).
24. Kanzaki Paper, US 3 244 730, 1963.
25. ICI, DE 2 243 436, 1971 (N. Hughes, G. H. Keats).
26. Bayer, DE 2 820 485, 1978 (R. Raue).
27. Hoechst, DE 3 309 726, 1983 (A. Engelmann).
28. Bayer, DE 3 842 014, 1988 (H. Berneth, R. Raue).
29. BASF, EP 491 256, 1990 (S. P. Rines, C. J. Zullig).
30. BASF, DE 2 736 679, 1977 (H. Kast, U. Mayer).
31. BASF, DE 2 152 703, 1971 (H. Kast, H. Baumann, U. Mayer, A. Oberlinner).
32. Amer. Cyanamid, US 4 321 207, 1980 (F. F. Cesark).
33. Bayer, DE 2 853 822, 1978 (K. H. Hermann).
34. Bayer, DE 2 928 466, 1979 (K. H. Hermann).
35. Bayer, DE 2 928 404, 1979 (K. H. Hermann).
36. Hodogaya, JP 73 20 603, 1970.

37. BASF, DE 4 211 783, 1992 (T. Gessner, U. Mayer).
38. Hodogaya, JP 53 74 530, 1976.
39. Hoechst, DE 1 098 652, 1956.
40. Hoechst, DE 1 644 619, 1967.
41. Dahl, DE 36 900, 1886.
42. ICI, DE 508 499, 1929.
43. Bayer, DE 287 003, 1914; DE 293 352, 1913. MLB, DE 292 998, 1914; DE 293 322, 1914.
44. Hoechst, DE 2 460 491, 1974 (M. Haehnke, F. Kohlhaas, F. Meininger, T. Papenfuhs).
45. Hoechst, DE 2 344 443, 1973 (R. Neeb, T. Papenfuhs).
46. National Cash Register, DE 2 242 005, 1971 (S. Farber, T. E. Hoover, D. B. McQuain, A. J. Wright).
47. Iwaki Seiyaku, JP 63 116 698, 1986 (T. Okada et al.).
48. Nisso Kako, JP 56 142 247, 1980.
49. IG Farben, DE 498 154, 1926.
50. Bayer, DE 2 820 486, 1978 (R. Raue).
51. BASF, DE 2 243 322, 1972 (H. Kast, H. Baumann).
52. Amer. Cyanamid, US 4 210 601, 1979 (J. H. Bright, J. P. Dundon, E. Klingsberg).
53. Allied Chem., US 3 489 800, 1966 (F. L. Tewksbury, Jr., R. B. Crawford, A. L. Cain).
54. Scott Paper, US 3 377 185, 1962.
55. Moore Business, BE 842 382, 1976.
56. Amer. Cyanamid, US 4 435 002, 1981 (F. F. Cesark, D. W. Thomas).
57. Amer. Cyanamid, US 4 343 939, 1981 (F. F. Cesark, R. J. Manfre, D. W. Thomas).
58. Ciba-Geigy, DE 2 602 404, 1975 (R. Garner, J. C. Petitpierre).
59. National Cash Register, DE 1 079 250, 1955.
60. National Cash Register, DE 1 094 386, 1956.
61. Hodogaya, DE 2 224 568, 1971 (Y. Miyazawa, M. Ozutsumi).
62. Caribonum, GB 831 852, 1958.
63. ICI, DE 2 144 323, 1970 (V. Boyd, R. A. Evans, K. A. Holt, A. H. M. Renfrew).
64. ICI, DE 2 229 031, 1971 (R. A. Evans, K. A. Holt, A. H. M. Renfrew).
65. Wiggins Teape, GB 1 417 695, 1971 (A. H. Berrie, D. Greenwood, N. Hughes, A. H. M. Renfrew).
66. ICI, DE 2 252 628, 1971 (J. F. McKellar, G. G. Warburton).
67. Hodogaya, JP 69 14 717, 1966 (H. Isao, N. Tomoyuki, O. Yoshitsugu).
68. Sumitomo, JP 71 18 400, 1967.
69. Nisso Kako, JP 72 29 660, 1969.
70. Allied Chem., DE 1 202 921, 1958.
71. Hodogaya, JP 50 32 949, 1973.
72. Bayer, DE 2 324 017, 1973 (H. Psaar).
73. BASF, DE 2 226 039, 1972 (H. Kast, H. Baumann, U. Mayer, A. Oberlinner).
74. Anilinfabriken W. Brauns, DE 1 768 927, 1968.
75. J. F. Rabek: "Sensitizing Effects of Benzophenone and Michler's Ketone in Photochemistry of Polymers," *Tech. Pap. Reg. Tech. Conf. Soc. Plast. Eng.,* Mid-Hudson Sect. (Oct.) 1973, 27 – 42.
76. S. P. Pappas: "Photosensitizers. Why and how?," *Tech. Pap. Soc. Manuf. Eng.* FC, FC 75 – 303, 1975, 5 pp.
77. BASF, DE 29 060, 1884; DE 31 936, 1884; DE 38 433, 1884; DE 53 614, 1889; DE 58 277, 1890; DE 70 908, 1893; DE 71 320, 1893. A. Kern, DE 37 737, 1886.
78. BASF, DE 2 234 836, 1972 (M. Eisert, K. Schmeidl); DE 2 261 064, 1972 (K. Schmeidl, M. Eisert).
79. BASF, DE 1 050 472, 1957; DE 2 207 609, 1972 (K. Schmeidl, M. Eisert, G. Riedel).
80. BASF, DE 2 213 848, 1972 (K. Schmeidl, M. Eisert, G. Riedel).
81. BASF, DE 1 109 291, 1959; DE 1 769 501, 1968; DE 2 502 838, 1975 (M. Eisert, K. Schmeidl).
82. BASF, DE 2 201 162, 1972 (K. Schmeidl, M. Eisert, G. Riedel).
83. Prod. Chim. Ugine Kuhlmann, DE 2 435 920, 1973; DE 2 435 922, 1973 (J. C. Hardouin, M. Vallette).
84. Prod. Chim. Ugine Kuhlmann, DE 2 342 857, 1972 (A. Arsac, R. Tosan).
85. BASF, DE 4 029 893, 1990 (W. Ball, R. Senninger).
86. Du Pont, US 3 373 199, 1962.
87. Amer. Cyanamid, US 3 209 027, 1962.
88. BASF, DE 1 198 470, 1962.
89. Sumitomo, DE 2 352 919, 1972 (S. Hotta).
90. Amer. Cyanamid, US 4 125 557, 1977 (F. F. Loffelman, F. F. Cesark).
91. BASF, DE 2 158 954, 1971 (K. Schmeidl, M. Eisert).
92. M. Okuba, S. Ueda, *Bull. Chem. Soc. Jpn.* **53** (1980), 281 – 282.
93. Bayer, DE 1 811 337, 1968 (R. Raue, W. Eifler, H. P. Kuehlthau); DE 1 811 652, 1968 (R. Raue, H. P. Kuehlthau, W. Eifler).
94. Fuji, GB 1 062 994, 1964.
95. Polaroid, US 4 277 407, 1979 (J. W. Foley).
96. Polaroid, US 4 277 406, 1979 (J. W. Foley).
97. Polaroid, EP 32 036, 1979 (J. W. Foley).
98. Polaroid, US 4 283 537, 1978 (L. Cincotta, J. W. Foley).

99. Bayer, DE 2 325 154, 1973 (H. Psaar).
100. Bayer, DE 2 750 283, 1977 (H. P. Kuehlthau, H. Psaar, R. Raue).
101. Fuji, JP 72 16 098, 1965.
102. V. V. Sinev, E. N. Smirnova, O. F. Ginzburg, *Zh. Org. Khim.* **16** (1980) no. 6, 1234 – 1238.
103. V. V. Sinev, A. V. Aleksandr, M. P. Pavlova, *Zh. Org. Khim.* **16** (1980) no. 8, 1684 – 1688.
104. A. V. Aleksandr. V. V. Sinev, O. F. Ginzburg, T. F. Stepanova, *Zh. Org. Khim.* **16** (1980) no. 8, 1688 – 1693.
105. V. V. Sinev, T. A. Nikolova, O. F. Ginzburg, *Zh. Org. Khim.* **18** (1982) no. 11, 2356 – 2360.
106. V. V. Sinev, T. A. Nikolova, *Zh. Org. Khim.* **18** (1982) no. 11, 2350 – 2355.
107. V. V. Sinev, I. T. Selivanova, *Zh. Org. Khim* **16** (1980) no. 12, 2472 – 2478.
108. V. V. Sinev, T. A. Nikolova, *Zh. Org. Khim.* **18** (1982) no. 9, 1977 – 1981.
109. Bayer, DE 3 605 552, 1986 (U. Eckstein, H. Psaar).
110. Bayer, EP 433 813, 1989 (U. Eckstein, R. Raue).
111. Bayer, EP 431 422, 1989 (U. Eckstein).
112. Bayer, EP 281 891, 1987 (U. Eckstein, H. Psaar, R. Raue).
113. Ciba, FR 1 118 967, 1954.
114. O. Fischer, *Ber. Dtsch. Chem. Ges.* **11** (1878) 950; **14** (1881) 2520.
115. BASF, DE 2 247 346, 1972 (U. Mayer, H. Kast).
116. Bayer, EP 373 432, 1988 (H. Berneth).
117. Finetech Japan, JP 20 1441, 1982.
118. Exp. Metedrology Ins., SU 1 286 995, 1984 (G. V. Vlasova, Y. U. I. Skurlatov, N. A. Kuznetsova). Akad. Wissenschaft DDR, DD 235 115, 1985 (F. Duttrich, M. Scholz).
119. Fuji Kagakushi Kogy, WO 9 204 416, 1991 (C. Takazawa, H. Akasako, M. Shini, Y. Shimazaki).
120. BASF, DE 1 925 811, 1969 (A. Tartter, G. Riedel, E. Nold).
121. Prod. Chim. Ugine Kuhlmann, DE 2 430 778, 1973 (G. Kremer, J. Pechmeze, R. Sureau).
122. Sandoz, DE 1 644 662, 1968 (U. Blass).
123. Sumitomo, JP 7 231 402, 1969.
124. Mitsubishi, FR 1 470 451, 1965 (W. Yamaya, S. Fujino, F. Kitakyushu). Sandoz, BE 711 959, 1967 (U. Blass).
125. Hoechst, BE 617 208, 1961; BE 617 209, 1961. Hoechst, DE 2 429 673, 1974 (M. Haehnke, T. Papenfuhs).
126. Hoechst, DE 2 206 476, 1972 (H. W. Bernhardt).
127. Hodogaya, JP 60 107 654, 1983. Mitsubishi Paper Mill, JP 61 236 561, 1985 (J. Yamada, K. Takaoka, H. Yamamoto). Canon, JP 01 306 858, 1988 (H. Tanigawa).
128. Minnesota Mining, EP 102 745, 1982 (D. E. Brown, K. Reynolds). Sharp, JP 63 148 270, 1986 (H. Sugimura, et al.).
129. Nippon Kayaku, JP 71 277 89, 1968.
130. Milliken, US 4 877 411, 1988 (J. B. Hines, J. R. Harris).
131. Milliken, US 5 043 013, 1989 (E. W. Kluger, P. D. Moore, J. A. Wagner).
132. IG Farben, DE 523 464, 1929.
133. IG Farben, DE 529 840, 1928; DE 572 475, 1931.
134. W. Wiedemann, *Chem. Ztg.* **106** (1982) 313. L. B. Schein in D. H. Auston, G. Ecker, W. Engl, L. B. Felsen (eds.): *Electrophotography and Development Physics,* **vol. 14,** Springer Verlag, Berlin 1988, p. 31.
135. Eastman Kodak, FR 2 099 783, 1970 (M. S. Bloom, C. J. Fox). Dainippon Ink Chem., JP 60 191 265, 1984 (H. Tanaka, O. Takenochi, T. Kawahara). Mita Ind., JP 63 210 939, 1987 (Y. Hanatani, K. Nakatani, T. Fukami).
136. James River-Graphic, WO 8 301 618, 1981 (E. W. Bennett, J. B. Suchy).
137. Tomoegawa Paper, JP 62 297 383, 1986 (M. Kamiyama).
138. Dainichi-Japan Cables, JP 51 047 037, 1974 (H. Kato, H. Fujita). Dainichi-Japan Cables, JP 51 112 858, 1975 (H. Kato, H. Fujita).
139. Sterling Drug, DE 2 361 813, 1972 (N. N. Crounse, P. J. Schmidt).
140. ICI, GB 2 177 699, 1985 (B. H. Meyrick, E. Young).
141. Hodogaya, DE 25 30 463, 1974 (M. Ozutsumi, Y. Miyazawa, M. Yamaguchi).
142. Bayer, EP 179 378, 1984 (K. Meisel).
143. Bayer, DE 3 500 361, 1985 (H. Berneth).
144. Bayer, DE 3 622 262, 1986 (H. Berneth, G. Japs).
145. Bayer, DE 3 841 668, 1988 (H. Berneth, G. Klug, G. Japs).
146. Bayer, DE 4 010 641, 1990 (G. Klug, H. Berneth, G. Japs).
147. Bayer, DE 3 420 799, 1984 (H. Berneth, A. Brack).
148. Bayer, DE 3 439 282, 1984 (H. Berneth, A. Brack, K. Meisel).
149. Ciba-Geigy, DE 3 314 195, 1982 (H. Balli, S. Gunzenhauser, I. J. Fletcher, D. Bedekovic).
150. Ciba-Geigy, DE 2 555 080, 1974 (R. Garner, J. C. Petitpierre).
151. Bayer, EP 330 040, 1988 (U. Eckstein).

152. Mitsubishi, JP 62 129 357, 1985 (F. Okumura, H. Tsukahara, H. Ikeda).
153. Bayer, EP 303 942, 1987 (U. Eckstein, H. Psaar, G. Japs).
154. MLB, DE 46 384, 1888; DE 50 286, 1889.
155. Sandoz, DE 154 528, 1902.
156. Geigy, DE 89 397, 1896.
157. BASF, DE 4 211 782, 1992. (T. Gessner, U. Mayer)
158. MLB, DE 108 129, 1889; DE 110 086, 1889; DE 111 506, 1889.
159. Deering Milliken, DE 2 125 098, 1970 (L. R. Foster, H. H. Kuhn).
160. MLB, DE 16 710, 1881.
161. Kalle, DE 93 540, 1895.
162. MLB, DE 61 146, 1889.
163. Amer. Cyanamid, US 542 544, 1948.
164. Chemetron, DE 2 355 735, 1973 (F. H. Turrill, L. E. Squire, J. L. Towle).
165. BASF, US 4 824 610, 1987 (R. J. Sappok, L. R. de Alvare, F. G. Spence).
166. Amer. Hospital Suppl., US 4 201 693, 1978 (W. D. Hurt, J. P. Koski).
167. P. K. Dasgupta, K. DeCesare, J. C. Ullrey, *Anal. Chem.* **52** (1980) 1912 – 1922.
168. MLB, DE 59 775, 1889.
169. Hoechst, DE 397 823, 1922.
170. J. T. Scanlan, *J. Am. Chem. Soc.* **57** (1935) 887.
171. Bayer, DE 2 555 747, 1975 (K. H. Hermann, H. L. Dorsch). Bayer, DE 2 723 774, 1977 (K. H. Hermann, H. L. Dorsch).
172. Orient Kagaku Kogyo, JP 73 283 90, 1969.
173. J. Schröder, R. Gossrau, *Acta Histochem. Suppl.* **25** (1982) 147 – 152.
174. Agfa-Gevaert, DE 2 128 388, 1970 (J. A. Herbots, A. J. Paesschen).
175. Khark Univ., SU 749 827, 1978 (V. D. Orlov, A. I. Roberman, V. F. Lavrushin).
176. BIOS, Final Report no. 1433, stem no. 22, p. 36.
177. Hoechst, DE 1 161 370, 1961; DE 1 161 371, 1961. Hoechst, DE 2 753 072, 1977 (M. Haehnke, E. Mundlos).
178. Minesterul Ind. Pet. Chim., DE 1 250 946, 1961.
179. Centr. Chim. Timisoara, DE 2 033 120, 1969 (W. E. Schmidt, R. Yilceanu).
180. BASF, EP 322 829, 1987 (A. S. G. Sharangpan, F. G. Spence, B. D. Teunis).
181. Sherwin-Williams, US 4 444 592, 1982 (T. E. Ludwig).
182. Hoechst, DE 2 018 169, 1970 (T. Papenfuhs, R. Rehberg, E. Spietschka); DE 2 504 934, 1975 (M. Haehnke, T. Papenfuhs, R. Schunck); DE 2 605 417, 1976 (M. Haehnke, T. Papenfuhs, C. Uhde).
183. FIAT **1313 II,** 314. Ullmann, 4th ed., **23,** 381, 396.
184. Amer. Cyanamid, US 2 816 900, 1954.
185. Nakada, JP 5117 2 21, 1974.
186. BASF, DE 26 016, 1883; DE 27 032, 1883; DE 27 789, 1883; De 29 943, 1884; DE 29 962, 1884. MLB, DE 34 463, 1884.
187. FIAT, **1313 II,** 317.
188. BASF, DE 2 334 918, 1973 (H. Kast).
189. BASF, DE 1 126 893, 1960.
190. BASF, DE 1 119 818, 1959.
191. IBM, US 4 229 219, 1979 (E. A. Metz).
192. As. Belo. Hard. Body., SU 1 140 035, 1983 (L. I. Ganago, N. N. Ishchenko, L. G. Starobinets).
193. Francolor, FR 1 307 382, 1961.
194. Mitsubishi, JP 20 976/68, 1965. Hodogaya, JP 7 1 11 1 17, 1968.
195. Milliken Res., US 4 091 034, 1976 (H. H. Kuhn).
196. BASF, DE 27 789, 1883; DE 29 962, 1884. FIAT, **1313 II,** 322.
197. BASF, GB 978 739, 1962.
198. BIOS, Final Report no. 959, stem no. 22, p. 16.
199. Du Pont, US 2 950 285, 1958.
200. BASF, DE 2 707 972, 1977 (M. Kamiyama).
201. Ricoh, JP 59 077 447, 1982.
202. Ricoh, DE 3 527 306, 1984 (S. Inoue, A. Matsui).
203. Hoechst, DE 3 641 525, 1986 (H. T. Macholdt, A. Sieber). Ricoh, JP 63 147 170, 1986 (M. Orihara, H. Tosaka, M. Tomita).
204. Henkel, EP 293 774, 1987 (W. Buchmeier, D. Brodalla).
205. Du Pont, DE 1 036 278, 1954; US 2 726 252, 1954. Allied Chem., FR 1 406 247, 1963.
206. Ciba-Geigy, DE 2 824 693, 1977 (P. Burri).
207. Xerox, US 4 304 829, 1977 (W. W. Limburg, D. M. Pai, D. S. Renfer). Mitsubishi Paper Mill, JP 59 031 960, 1982. Tomoegawa Paper, JP 59 159 166, 1983.
208. Fuji Xerox, JP 57 169 759, 1981.
209. Xidex, WO 8 001 846, 1979 (L. Roos, J. L. Jennings). Mitsui Toatsu Chem., JP 59 083 152, 1982. BASF, DE 3 824 551, 1988 (R. Aldag, et al.).
210. Duskin, JP 62 291 564, 1986 (M. Kajimaki, Y. Nakajima, H. Yamamoto, N. Matsuda).
211. Mitsui Toatsu Chem., WO 8 501 105, 1983 (M. Asano, I. Chiba, M. Iemura).
212. U.S. Atomic Energy Comm., US 3 609 093, 1968.
213. L. Chalkley, US 2 839 542, 1955.

214. L. Chalkley, US 2 829 149, 1956.
215. F. L. Dickert et al., *Anal. Chem.* **60** (1988) 1377.
216. ICI, GB 1 331 047, 1969.
217. Ciba-Geigy, DE 2 643 569, 1975 (P. Burri, J. C. Petitpierre).
218. BASF, DE 2 363458, 1973 (H. Baumann, K. Grychtol, A. Oberlinner).
219. Kanzaki Paper, DE 2 629 937, 1975 (M. Kondo et al.).
220. Ciba-Geigy, GB 2 004 294, 1978 (R. Garner).
221. Kanzaki Paper, DE 2 709 580, 1976 (M. Kondo, H. Iwasaki, K. Yasui, M. Miyake).
222. Clayton Aniline, DE 2 156 648, 1970 (G. Smith, T. Marley).
223. ICI, DE 2 141 406, 1970 (N. Hughes).
224. BASF, DE 2 215 048, 1972 (W. Seibert).
225. Amer. Cyanamid, US 4 455 435, 1979 (J. P. Dundon, E. Klingsberg).
226. Ciba-Geigy, DE 3 104 950, 1980.
227. Amer. Cyanamid, US 4 076 728, 1977 (D. R. Maulding).
228. ICI, DE 2 143 021, 1970 (N. Hughes).
229. Sterling Drug, DE 964 627, 1954.
230. Clayton Aniline, DE 2 156 662, 1970 (G. Smith, T. Marley).
231. BASF, DE 2 835 450, 1978 (W. Seibert).
232. Hodogaya, JP 50 124 931, 1974 (R. Aida, M. Nagata, I. Motono).
233. Ciba-Geigy, DE 3 227 425, 1981 (M. C. Clark, J. B. Henshall).
234. Yamada, EP 291 600, 1987 (Y. Fujino, H. Kawai, K. Tsunemitsu).
235. BASF, DE 2 226 936, 1972 (H. Kast, H. Baumann, U. Mayer, A. Oberlinner).
236. Hodogaya, DE 3 005 397, 1979 (M. Nagata, K. Wakasugi).
237. Kanzaki Paper, JP 52 107 026, 1976 (K. Yasui, M. Miyake, H. Iwasaki, M. Kondo).
238. Kanzaki Paper, JP 52 108 965, 1976.
239. Hodogaya, US 3 967 835, 1974.
240. Olivetti, EP 82 711, 1981 (F. Knirsch).
241. M. Kuzuya, F. Miyake, T. Okuda, *Tetrahedron Lett.* **21** (1980), 1043 – 1046.
242. Bayer, EP 249 103, 1986 (H. Heidenreich, G. Wolfrum, K. Wehling, H. Hugl).
243. Sterling Drug, EP 108 382, 1982 (P. J. Schmidt, W. M.-W. Hung).
244. Hoechst, DE 2 914 299, 1979 (M. Schneider, R. Deubel, R. Schunck).
245. Hoechst, FR 2 017 958, 1968.
246. BASF, EP 60 428, 1982.
247. Geigy, DE 73 092, 1892.
248. Bayer, DE 125 134, 1900.
249. Geigy, DE 59 811, 1890. Cassella, DE 62 339, 1890.
250. Otsuka Pharm, JP 56 099 566, 1979; JP 56 103 366, 1980.
251. Bayer, DE 68 291, 1891; DE 69 777, 1891.
252. Geigy, DE 1 167 789, 1961.
253. Agfa-Gevaert, BE 696 604, 1966.
254. Du Pont, US 3 406 069, 1965 (J. D. Overman).
255. Dow Chem., US 4 252 725, 1979 (H. B. Prindle, G. E. Ham).
256. Polaroid, DE 2 127 907, 1970 (E. H. Land, M. S. Simon, D. P. Waller).
257. Polaroid, US 3 772 338, 1971; EP 362 750, 1988 (D. P. Waller).
258. Polaroid, US 3 822 293, 1972 (H. Bader, M. H. Feingold, S. C. January, J. W. Sparks); US 3 822 294, 1972 (E. Chinoporos, P. S. Huyffer); US 3 806 522, 1972 (M. H. Feingold); US 3 869 473, 1972 (Y. H. Chiang, J. W. Sparks).
259. Polaroid, US 3 821 254, 1972 (P. S. Huyffer, J. J. Wasatonic).
260. Adressograph, US 3 486 888, 1966 (L. C. Browning).
261. Z. Vodak, DE 1 061 328, 1953.
262. IBM, GB 1 182 626, 1966 (U. O. Thompson, B. I. Halpering).
263. GAF, DE 2 101 250, 1970.
264. Polaroid, DE 2 841 321, 1978 (A. L. Borror, J. W. Foley, L. Cincotta, E. W. Ellis).
265. Warner-Lambert, DE 1 276 047, 1964.
266. Boehringer, EP 146 866, 1983 (M. Kuhr et al.).
267. BASF, DE 1 254 118, 1963. Bayer, GB 1 139 408, 1966 (R. Raue, H. P. Kuhlthau); BE 702 240, 1966; BE 713 385, 1967.
268. Bayer, DE 1 569 750, 1967 (H. Psaar, R. Raue).
269. Bayer, DE 121 837, 1900; DE 127 245, 1901; DE 128 660, 1901; DE 128 904, 1901.
270. Hodogaya, JP 7 220 478, 1969.
271. Hoechst, DE 2 625 944, 1976 (M. Haehnke, E. Mundlos, T. Papenfuhs).
272. Ciba-Geigy, DE 2 257 711, 1971 (R. Garner, M. Duennenberger).
273. Bayer, EP 237 915, 1986 (U. Eckstein).
274. Ciba-Geigy, DE 2 423 535, 1974 (R. Garner, J. C. Petitpierre); DE 2 514 934, 1975 (R. Garner, J. C. Petitpierre).
275. National Cash Register, US 3 540 912, 1967.
276. Ciba-Geigy, DE 2 423 534, 1973 (R. Garner, J. C. Petitpierre).
277. Mitsubishi JP 50 119 026, 1974 (F. Okumura, S. Maeda, S. Inoue, H. Honda).
278. Sumitomo, JP 74 017 489, 1970 (N. Akamatsu, K. Koga, M. Miyake).
279. Hodogaya, DE 2 435 408, 1973 (Y. Miyazawa, M. Ozutsumi, S. Ogawa); DE 2 435 640, 1973.

280. Appleton Papers, DE 3 012 201, 1979 (H. H. Baum).
281. Ciba-Geigy, EP 082 822, 1982 (D. Bedekovic, I. J. Fletcher).
282. Yamamoto, DE 3 116 815, 1980.
283. ICI, EP 192 328, 1985 (N. Hall).
284. National Cash Register, DE 2 520 148, 1974 (S. Farber).
285. Ciba-Geigy, EP 91 402, 1982 (H. Balli, S. Gunzenhauser, I. J. Fletcher, D. Bedekovic).
286. Bayer, DE 3 521 406, 1985 (H. Heidenreich, G. Wolfrum, U. Nehen).
287. Ciba-Geigy, DE 2 503 309, 1974 (R. Garner, J. B. Henshall, J. C. Petitpierre).
288. Ciba-Geigy, DE 2 602 403, 1975 (R. Garner, J. C. Petitpierre).
289. Tomoegawa Paper, JP 62 297 384, 1986 (M. Kamiyama).
290. Bayer, DE 2 917 271, 1979 (H. Psaar, H. P. Kuehlthau, R. Raue).
291. Ciba-Geigy, DE 2 747 526, 1976 (P. Burri).
292. Ciba-Geigy, DE 2 915 928, 1978 (J. C. Petitpierre, P. Burri).
293. Ciba-Geigy, DE 2 747 525, 1976 (P. Burri).
294. Eastman Kodak, EP 162 685, 1984 (B. E. Babb, D. S. Daniel).
295. Hodogaya, JP 69 015 313/69, 1966 (M. Iizuka, A. Arakawa, M. Yokobori). BASF, DE 2 138 931, 1971 (H. Kast, H. Baumann, U. Mayer).
296. BASF, DE 2 739 953, 1977 (H. Kast, K. Grychtol, F. Feichtmayr).
297. BASF, DE 3 917 601, 1989 (E. Hahn, W. Breitschaft, U. Mayer, G.-R. Schröder).
298. Mitsui Toatsu Chem., JP 60 106 858, 1983.
299. Mitsui Toatsu Chem., JP 11 14 4 81, 1987.
300. BASF, DE 2 951 353, 1979 (H. Bruder et al.).
301. Minnesota Mining, US 4 477 635, 1984 (S. Mitra).
302. Fr.-Schiller-Universität Jena, DD 231 079, 1984 (H. H. Hoerhold, N. Baumbach, J. Opfermann, M. Helbig).
303. Fuji Photo Film, DE 3 048 408, 1980 (S. Watarai, K. Ishii, T. Takayama); DE 3 146 113, 1980 (H. Tachikawa, S. Watarai); JP 58 057 134, 1981; JP 58 105 240, 1981. Canon, JP 61 151 545, 1984 (M. Matsumoto, H. Miyazaki).
304. Bayer, DE 223 879, 1909; DE 226 348, 1909.
305. Uemura Kogyo, JP 72 285 78, 1968.
306. Bayer, DE 68 381, 1892; DE 73 112, 1892; DE 80 282, 1893; DE 81 516, 1893.
307. Mitsubishi Paper Mill, JP 61 278 867, 1985 (J. Yamada, H. Yamamoto, K. Takaoka). Ricoh, JP 63 008 652, 1986 (C. Uematsu, T. Isoda, N. Nakayama). Bayer, DE 3 714 289, 1987 (R. Raue, H. Psaar).
308. Bayer, DE 3 714 288, 1987 (R. Raue, H. Psaar).
309. Canon, JP 62 279 345, 1986 (H. Toma, N. Kashimura, M. Hisamura, F. Sumino, S. Tanaka).
310. *Ullmann* 3rd ed., **3,** 71.
311. Bayer, EP 58 332, 1981 (R. Raue).
312. P. Jakobsen, A. P. Anderson, H. Lyon, S. Treppendahl, *Microsc. Acta* **87** (1983) no. 1, 41 – 47.
313. M. El Baraka, M. Deumie, P. Viallet, T. J. Lampidis, *J. Photochem. Photobiol. A* **56** (1991) nos. 2 – 3, 295 –311.
314. D. A. Mikaelyan, V. Z. Artsruni, A. G. Khachatryan, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **34** (1991) no. 7, 121 – 123.
315. N. O. Geokchyan, A. G. Khachatryan, *Arm. Khim. Zh.* **43** (1991) no. 9, 572 – 575.
316. Y. Anjaneyulu, N. Suguna, M. R. P. Reddy, K. Chandrasekhar, *Analusis* **14** (1991) no. 4, 200 – 204.
317. S. Jaya, T. P. Rao, T. V. Ramakrishna, *Analyst (London)* **109** (1986) no. 11, 1405 – 1407.
318. T. P. Rao, T. V. Ramakrishna, *Analyst (London)* **107** (1987) no. 1275, 704 – 707.
319. G. Popa, D. Cruceru, A. Tanase, *Bul. Inst. Politeh. Gheorghe Gherorghiu-Dej Bucuresti, Ser. Chim. Metal.* **41** (1979) no. 4, 49 – 53.
320. T. Naganuma, *Aichi Kyoiku Daigaku Kenkyu Hokuku Shizen Kagaku* **31** (1982) 123 – 130.
321. P. Selvapathy, T. V. Ramakrishna, N. Balasubramanian, R. Pitchai, *Analyst (London)* **112** (1987) no. 8, 1139 – 1141.
322. P. Selvapathy, R. Pitchai, T. V. Ramakrishna, *Talanta* **37** (1990) no. 5, 539 – 544.
323. H. Lyon, P. Jakobsen, P. P. Clausen, A. P. Andersen, *Histochem. J.* **15** (1983) no. 6, 605 – 606.
324. A. Pollack et al., *Cytometry* **3** (1982) no. 1, 28 – 35.
325. A. P. Wagner, M. C. Iordachel, L. P. Wagner, *J. Biochem. Biophys. Methods.* **8** (1983) no. 4, 291 –297.
326. Cassella, DE 274 358, 1913.
327. BASF, DE 1 469 639, 1965.
328. Bayer, DE 3 322 945, 1983 (K. H. Drexhage, H. Harnisch, R. Raue).
329. T. M. Grigorèva, V. L. Ivanov, M. G. Kuz'min, *Zh. Org. Khim.* **17** (1981) no. 1, 162 – 167.
330. Bayer, DE 51 983, 1888.
331. ICI, GB 1 407 331, 1973 (N. Hughes, J. L. Leng).

332. Hoechst, DE 1 942 508, 1969 (K. Loehe, R. Mohr, E. Mundlos, H. Freb).
333. ICI, GB 1 409 493, 1973 (N. Hughes, J. L. Leng).
334. Nisso Kako, DE 3 612 618, 1985.
335. Fuji, Hodogaya, DE 1 561 748, 1966.
336. Sumitomo, DE 2 016 371, 1970 (K. Koga, H. Suda, T. Akamatsu).
337. Hodogaya, EP 356 199, 1988 (M. Yamaguchi, K. Kikkawa, M. Gonda).
338. Ciba-Geigy, EP 384 895, 1989 (R. Zink).
339. Nisso Kako, DE 1 768 862, 1967. Fuji, Hodogaya, DE 1 805 697, 1967 (S. Katayama, R. Goto, T. Nita, S. Ogawa).
340. Fuji, DE 1 671 545, 1967.
341. Ciba-Geigy, DE 2 422 899, 1973 (R. Garner, J. C. Petitpierre).
342. Yamamoto, DE 2 130 845, 1970 (M. Terayama, K. Yamamoto); DE 2 166 374, 1970 (M. Terayama, K. Yamamoto).
343. Ciba-Geigy, DE 3 832 756, 1987 (R. Zink).
344. Fuji, DE 2 024 859, 1969 (S. Kimura, T. Kobayashi, S. Ishige, M. Kiritani).
345. Hodogaya, EP 48 026, 1980 (K. Motohashi, T. Obara, K. Kaneko, M. Yamaguchi).
346. National Cash Register, DE 2 001 864, 1969 (C.-H. Lin, T. E. Hoover, A. J. Wright).
347. Yamamoto, DE 2 244 873, 1971 (Y. Hatano, K. Yamamoto); DE 2 261 984, 1971.
348. National Cash Register, DE 2 242 005, 1971 (S. Farber, T. E. Hoover, D. B. McQuain, A. J. Wright).
349. Ciba-Geigy, DE 3 505 285, 1985 (R. Zink).
350. Nisso Kako, DE 2 424 935, 1973 (M. Yahagi, S. Horiuchi, T. Toyama, A. Kashiwagi).
351. Ciba-Geigy, DE 2 444 297, 1973 (R. Garner, J. C. Petitpierre).
352. Yamada, DE 2 262 127, 1971 (K. Tsunemitsu et al.).
353. Yamamoto, DE 2 202 315, 1971 (Y. Hatano, K. Yamamoto).
354. Nisso Kako, DE 2 242 693, 1971 (M. Yahagi, T. Suzuki, T. Igaki, S. Horiuchi).
355. Mitsubishi, DE 2 110 859, 1970 (K. Futaki, M. Motoki).
356. Hodogaya, EP 155 796, 1984 (H. Sensui, S. Suzuka, M. Gonda, K. Kikkawa).
357. Yamada, DE 2 834 711, 1977 (H. Kawai, K. Tsunemitsu, Y. Fujino, Y. Shimizu).
358. Hodogaya, EP 122 318, 1983 (M. Yamaguchi, M. Gonda, Y. Satoh, M. Kanasugi).
359. Nisso Kako, DE 3 618 562, 1985.
360. Mitsui Toatsu Chem., EP 420 024, 1989 (A. Otsuji et al.).
361. Mitsui Toatsu Chem., EP 411 509, 1989 (M. Nakatsuka et al.).
362. Fuji, DE 3 221 651, 1981 (K. Iwakura).
363. BASF, DE 4 002 929, 1990 (T. Gessner, U. Mayer).
364. Nippon Soda, DE 4 112 066, 1991 (N. Hirai et al.).
365. Sterling Drug, EP 110 366, 1982 (W. M.-W. Hung, E. E. Ehlinger).
366. Fuji, DE 3 529 796, 1984 (M. Satomura, K. Iwakura, A. Igarashi).
367. Toaka Chemical, EP 176 161, 1984 (T. Akamatsu).
368. Shin Nisso Kako, DE 3 725 258, 1986 (T. Obitsu et al.). Jujo Paper, EP 286 116, 1987 (T. Satake, T. Minami, T. Nagai, F. Fujimura).
369. BASF, EP 324 344, 1988 (U. Mayer, B. Landmann). Ciba-Geigy, EP 333 649, 1988 (R. Zink).
370. Jujo Paper, EP 421 278, 1989 (T. Minami, T. Fukuchi, R. Ohashi, T. Kaneko).
371. Hodogaya, EP 156 250, 1984 (M. Anzai et al.).
372. Fuji, DE 3 225 028, 1981 (K. Iwakura, S. Ishige).
373. Hodogaya, EP 107 780, 1982 (M. Anzai et al.).
374. Appleton Papers, EP 276 980, 1987 (P. Dwyer-Hallquist, W. J. Becker, R. E. Miller, K. D. Glanz). Yamamoto, EP 526 856, 1991 (S. Fujita et al.).
375. Kanzaki Paper, EP 89 752, 1982 (M. Kondo et al.).
376. ICI, EP 207 635, 1985 (N. Hughes, E. Young).
377. Showa Denko, EP 220 515, 1985 (T. Kitao et al.).
378. K. Imamura, T. Yamaguchi, T. Kitao, *Chem. Express* **2** (1987) no. 1, 81 – 84.
379. Champion Int. Corp., US 4 250 098, 1978 (J. Chang).
380. BASF, EP 138 177, 1983 (U. Mayer, A. Oberlinner).
381. Fuji, GB 2 178 440, 1985 (M. Satomura, K. Iwakura, A. Igarashi).
382. Sterling Drug, DE 3 018 546, 1979 (P. J. Schmidt, W. M.-W. Hung).
383. BASF, EP 167 998, 1984 (U. Mayer, A. Oberlinner).
384. BIOS Final Report no. 959, stem no. 22, pp. 32 –36.
385. Taoka Chemikal KK, EP 468 821, 1990 (K. Aburada, M. Akagi).
386. BASF, DE 3 531 272, 1985 (K. Schmeidl).
387. BASF, DE 2 138 178, 1971 (K. Schmeidl).
388. BASF, DE 1 644 487, 1968 (R. Schrödel, G. Riedel, F. Andree).
389. BASF, DE 2 138 179, 1971 (K. Schmeidl, F. Andree).

390. BASF, DE 3 832 739, 1988 (K. Schmeidl).
391. Mobay, EP 180 116, 1984 (V. S. Oliver, W. Tames, H. H. Prochaska, K. H. Wieser). Dainippon Ink Chem., EP 222 335, 1985 (H. Ando, Z. Takada, Y. Shigeta).
392. BASF, EP 347 697, 1988 (U. Mayer, A. Oberlinner).
393. BASF, EP 34 725, 1980 (M. Patsch, M. Ruske).
394. Milliken Res., EP 527 648, 1991 (C. N. Barry, Jr.); EP 527 649, 1991 (T. B. Gunnell, J. B. Hines, Jr., C. N. Barry, Jr.).
395. BASF, EP 392 356, 1989 (W. Breitschaft, H. Bruder, U. Mayer, G. R. Schroeder).
396. Bayer, DE 3 714 288, 1987 (R. Raue, H. Psaar).
397. Eastman Kodak, EP 248 176, 1986 (W. T. Gruenbaum, J. P. Harmon, L. C. Roberts).
398. Appl. Biosystems, EP 233 053, 1986 (S. Fung); EP 272 007, 1986 (S. M. Menchen, S. Fung).
399. Exxon Res., DE 2 638 943, 1975. Kodak, DE 2 113 127, 1970.
400. Kodak, US 3 822 270, 1975.
401. Kodak, DE 2 238 050, 1971 (K. H. Drexhage, G. A. Reynolds).
402. U.S. Dept. of Energy, US 4 945 176, 1989 (P. R. Hammond, J. F. Feeman); US 5 047 559, 1990 (P. R. Hammond, J. F. Feeman).
403. Hoechst, DE 2 504 925, 1975 (M. Haehnke, R. Neeb, T. Papenfuhs).
404. Hoechst, DE 2 002 916, 1970 (E. Mundlos, R. Mohr, K. Loehe).
405. D. A. Hinckley, P. G. Seybold, D. P. Borris, *Spectrochim. Acta Part A* **42** (1986) 747 – 754
406. Fuji, DE 2 049 527, 1969 (S. Kimura, T. Kobayashi, S. Ishige).
407. Fuji, DE 3 045 022, 1979 (K. Sato, K. Iwakura, A. Igarashi).
408. Fuji, DE 3 212 213, 1981 (S. Ishige, K. Sato, K. Iwakura).
409. BASF, DE 2 338 953, 1973 (H. Kast, G. Dunkelmann).
410. National Cash Register, DE 932 614, 1955.
411. Polaroid, EP 250 558, 1985 (E. J. Dombrowski, J. R. Freedman, P. F. King).
412. Polaroid, US 4 839 335, 1987 (E. W. Ellis).
413. Polaroid, US 4 818 742, 1987 (E. W. Ellis).
414. Polaroid, EP 224 568, 1986 (A. L. Borror, E. W. Ellis); EP 543 136, 1991 (R. S. Baker et al.).
415. Polaroid, EP 174 054, 1984 (A. B. Borror, E. W. Ellis, D. A. McGowan).
416. Polaroid, EP 31 698, 1979 (R. L. Cournoyer, J. W. Foley); DE 3 048 165, 1979 (J. W. Foley, L. Locatell, C. M. Zepp); EP 115 701, 1982 (A. L. Borror, L. Cincotta, J. W. Lee).
417. MLB, DE 205 758, 1906; DE 229 466, 1910.
418. Bayer, DE 2 421 607, 1974.
419. Hoechst, DE 2 018 168, 1970 (T. Papenfuhs, R. Rehberg, E. Spietschka).
420. A. v. Baeyer, *Ber. Dtsch. Chem. Ges.* **4** (1871) 555, 658.
421. Boehringer, DE 2 040 033, 1970 (W. Rittersdorf, H. G. Rey, W. Guethlein). Hoffmann La Roche, DE 2 359 099, 1972.
422. Abbot Lab., EP 108 403, 1982 (C. A. Flentge, C. L. Kirkemo); EP 108 399, 1982 (J. R. Fino, C. L. Kirkemo).
423. A. von Baeyer, *Ber. Dtsch. Chem. Ges.* **8** (1875) 146; *Justus Liebigs Ann. Chem.* **183** (1876) 1.
424. L'Oreal, DE 1 768 740, 1967.
425. A. H. Royer, *Justus Liebigs Ann. Chem.* **238** (1887) 350. Ciba, DE 32 564, 1885.
426. Amer. Cyanamid, US 3 351 603, 1965 (C. S. H. Chen, F. M. Gray).
427. Sterling Drug, EP 146 059, 1983 (N. N. Crounse, J. R. Heitz).
428. A. von Baeyer, *Ber. Dtsch. Chem. Ges.* **4** (1871) 457, 555, 658. O. Gürke, DE 30 648, 1884; DE 32 830, 1885.
429. Ciba Geigy, DE 65 739, 1892.
430. Ciba-Geigy, BE 874 476, 1978.
431. S. Schütze, *Radiat. Environ. Biophys.* **21** (1982) 123 –140.
432. Ciba-Geigy, DE 4 139 851, 1990 (H. Balli, S. Gunzenhauser).
433. Sankyo Chem. JP 5 589 354, 1976; DE 2 707 784, 1976 (S. Kimura, Y. Nakamura, S. Kurimoto, K. Motouchi).
434. A. Barker, C. C. Barker, *J. Chem. Soc.* 1954, no. 4, 1307.
435. Chester Davies, US 3 344 189, 1963 (C. Davies).
436. BASF, DE 2 051 318, 1970 (D. Leuchs, H. Kast).
437. Horizons Inc., US 3 980 706, 1975 (P. J. Mondron).
438. Eastman Kodak, US 3 781 711, 1972 (K. H. Drexhage, G. A. Reynolds).
439. Hughes Aircraft Comp., US 4 139 342, 1977 (S. J. Sheldrake, H. H. L. Wang).
440. Yamamoto, EP 124 377, 1983 (S. Ikegami, T. Amako).
441. Kanzaki Paper, EP 209 259, 1985 (N. Kanda, M. Kondo, T. Horiike).
442. Nippon Kayaku, JP 1 3 11 078, 1988.
443. BASF, DE 2 145 027, 1971 (H. Kast).
444. Yamamoto, EP 278 614, 1987 (S. Ikegami, S. Nakao).

445. Nisso Kako, DE 3 723 247, 1986 (Y. Ohnishi et al.).
446. Fuji, JP 3 166 267, 1989.
447. National Cancer Institute Carcinogenesis: "Bioassay of Michlers Ketone for Possible Carcinogenicity," *Tech. Rep. Ser.* no. 181, U.S. Department of Health, Education and Welfare, Public Health Service, National Institutes of Health, Washington 1979.
448. V. F. Simmon, *J. Natl. Cancer Inst. U.S.* **62** (1979) 893 – 899.
449. J. McCann, E. Choi, E. Yamasaki, B. N. Ames, *Proc. Natl. Acad. Sci. U.S.A.* **72** (1975) 5135 – 5139.
450. R. S. U. Baker, A. M. Bonin in F. J. De Serres, J. Ashby (eds.): *Evaluation of Short-Term Tests for Carcinogens,* Report of the International Collaborative Program, vol. 1, Elsevier/North Holland, New York 1981, pp. 249 – 260.
451. G. M. Bonser, D. B. Clayson, J. W. Jull: "The Introduction of Tumors of the Subcutaneous Tissues, Liver and Intestine in the Mouse by Certain Dyestuffs and Their Intermediates," *Br. J. Cancer* **10** (1956) 653 – 667.
452. M. H. C. Williams, G. M. Bonser, *Br. J. Cancer* **16** (1962) 87 – 91.
453. P. Kirsch et al.: *Arbeitsmedizin, Sozialmedizin, Präventivmedizin,* special issue no. 2, A. W. Gentner Verlag, Stuttgart 1978.
454. S. Parodi et al., *J. Toxicol. Environ. Health* **9** (1982) 941 – 952.
455. R. A. M. Case, J. T. Pearson, *Br. J. Ind. Med.* **11** (1954) 213 – 216.
456. A. M. Thiess, R. Link, G. Wellenreuther, *Proc. Int. Conf. Occup. Health Chem. Ind. 9th* 1981.
457. SRI, ETAD, *SRI Rep.* 1987.
458. N. A. Littlefield, B. Blackwell, C. C. Hewitt, D. W. Gaylor, *Fundam. Appl. Toxicol.* **5** (1985) 902 – 912.
459. SRI, ETAD, *SRI Rep.* 1987, pp. 13, 95.
460. S. Toyoshima et al.: "Effect of Environmental Substances on Production of Spontaneously Occurring Carcinoma in C3H Mice," *Jpn. J. Pharmacol.* **36** (1984) 123.
461. S. Carson, "Skin Painting Studies in Mice With 14 FD & C and D & C Colors," *J. Toxicol. Cutaneous Ocul. Toxicol.* **3** (1984) 357 – 370.
462. M. Umeda: "Experimental Study of Xanthene Dyes as Carcinogenetic Agents," *GANN* **47** (1956) 51 –78.

Trichloroacetic Acid → Chloroacetic Acids
Trichloroethanes → Chlorinated Hydrocarbons
Trichloroethylene → Chlorinated Hydrocarbons
Trimellitic Acid → Carboxylic Acids, Aromatic
Trimethylamine → Methylamines
Trioxan → Formaldehyde
Triple Superphosphate → Phosphate Fertilizers
Tritium → Hydrogen
Tryptophan → Amino Acids

Tungsten, Tungsten Alloys, and Tungsten Compounds

Erik Lassner, Graz, Austria (Chap. 1, Section 2.2, Chaps. 3 – 7, Chaps. 9 – 13)

Wolf-Dieter Schubert, Institut für chemische Technologie anorganischer Stoffe, Technische Universität Wien, Wien, Austria (Chap. 1, Section 2.2, Chaps. 3 – 7, Chaps. 9 – 13)

Eberhard Lüderitz, Fürth, Federal Republic of Germany (Chap. 8)

Hans Uwe Wolf, Universität Ulm, Ulm, Federal Republic of Germany (Chap. 14)

1. Introduction 2
2. Properties 4
2.1. Physical Properties 4
2.2. Chemical Properties 6
3. Raw Materials 6
3.1. Natural Resources 6
3.2. Tungsten Scrap 8
4. Production 8
4.1. Ore Beneficiation 8
4.2. Pretreatment of Ore Concentrates and Scrap 9
4.3. Hydrometallurgy 9
4.3.1. Digestion 9
4.3.2. Purification 10
4.3.3. Conversion of Sodium Tungstate Solution to Ammonium Tungstate Solution 11
4.3.3.1. Solvent Extraction 11
4.3.3.2. Ion Exchange Process 11
4.3.4. Crystallization of Ammonium Paratungstate (APT) 11
4.4. Calcination of APT 12
4.5. Reduction by Hydrogen to Tungsten Metal Powder 12
4.6. Production of Compact Metal . . 13
4.7. Processing of Sintered Parts 14
4.7.1. Shaping 14
4.7.2. Mechanical Bonding of Tungsten to Tungsten and Other Metals 15
4.8. Surface Treatment 15
4.9. Tungsten Coatings 16
4.10. Production of High-Purity Tungsten Metal (99.999 – 99.9999 %) 16
5. Tungsten Alloys 16
5.1. Single-Phase Mixed-Crystal Alloys 16
5.2. Multiple Phase Alloys 17
6. Uses 19
7. Analysis 21
7.1. Raw Materials 21
7.2. High-Purity Intermediate Products, Tungsten Powder, and Compact Tungsten Metal 21
7.3. Trace Elements in High-Purity Tungsten Metal 22
8. Ferrotungsten and Related Alloys 23
8.1. Composition 23
8.2. Uses 23
8.3. Production 23
8.3.1. Ferrotungsten 23
8.3.1.1. Carbothermic Production 24
8.3.1.2. Carbothermic and Silicothermic Production 25
8.3.1.3. Metallothermic Production 26
8.3.2. Production of Tungsten Melting Base and Tungsten Master Alloys . 26
9. Compounds 27
9.1. Tungsten Chemistry 27
9.2. Aqueous Solutions of Tungsten . . 27
9.3. Intermetallic Compounds 27
9.4. Compounds with Nonmetals . . . 27
9.4.1. Tungsten – Boron Compounds . . . 27
9.4.2. Tungsten – Carbon Compounds . . . 27
9.4.3. Tungsten – Silicon Compounds . . . 28
9.4.4. Tungsten – Group 15 Compounds . 28
9.4.5. Tungsten – Oxygen Compounds . . 28
9.4.6. Tungsten – Chalcogenide Compounds 30
9.4.7. Tungsten – Halogenide Compounds 31
10. Tungsten in Catalysis 31
11. Economic Aspects 32
12. Tungsten Recycling 32
13. Beneficial Effects on Human Health 33
14. Toxicology and Occupational Health 34
15. References 35

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_229

Sections 2.1 and 4.6 – 4.9 were based on the article Wolfram, Wolfram-Legierungen und Wolfram-Verbindungen in Ullmann's, 4th ed., written by Hugo M. Ortner

1. Introduction

Tungsten (wolfram, W), atomic number 74, relative atomic mass 183.85, is a transition metal in the third long period and group 6 of the periodic table. Its electron configuration is [Xe] $6s^2 4f^{14} 5d^4$ but can also be interpreted as [Xe]$6s^1 4f^{14} 5d^5$. The naturally occurring isotopes (abundances in parentheses) have relative atomic mass 180 (0.14 %), 182 (26.42 %), 183 (14.40 %), 184 (30.64 %), and 186 (28.41 %). There are also 17 known artificial isotopes and isomers from ^{137}W to ^{189}W, with half-lifes between 14 μs and 140 d In nature, tungsten occurs only in a chemically combined state (mainly as tungstate). The metal, which has a silvery white luster, has a very high density (19.3 g/cm^3 at 20 °C), and the highest melting point (3410 °C) of all the metallic elements.

History. Long before elemental tungsten was discovered (1783), the mineral wolframite was known, e.g., in the tin mines of Saxony and Bohemia. It was first described in 1574 by Lazarus Ercker, and thought to be a tin mineral that contained arsenic and iron.

The name wolfram is derived from the word wolf, i.e., the wolf that devoured tin. A foam (Latin: *lupi spuma*, wolf's foam) was formed on the molten tin during the smelting of tin ore, reducing the yield of tin.

In 1757, a new mineral with density 6 (scheelite) was described by A. F. Cronstedt. He named this "tungsten" (Swedish: *tung sten*, heavy stone) and considered it to be a calcium-containing iron mineral. In 1781, C. W. Scheele analyzed the ore and announced that it contained calcium with a hitherto unknown acid. He described digesting the mineral with K_2CO_3, precipitating the unknown acid with nitric acid, decomposing the mineral with mineral acids, and dissolving the precipitated acid in ammonia. In the same year, T. Bergmann suggested that the metal might be obtained by reduction of the acid with carbon.

In 1783, the element was produced by the brothers J. J. and F. de Elhuyar from wolframite by reduction of the oxide with carbon. They named it wolfram. In 1786, the hardening of steel was described by H. I. du Monceau. In 1821, K. C. von Leonhard proposed the name scheelite for the mineral $CaWO_4$.

The real chemistry of tungsten originated with R. Oxland (1874), who patented a method for producing sodium tungstate, tungsten oxide, and metallic tungsten. He was also the first to propose a method for the production of ferrotungsten. The possibility of using tungsten as an alloying element to increase the hardness of steel could not initially be realized because of the high price of the metal.

The developments that led to the industrial use of tungsten began in 1890 with the production of WC by H. Moissan. They may be summarized as follows:

In 1900, the Bethlehem Steel Company announced the first high-speed, high-temperature steel at the Paris Exhibition.

In 1903, the first tungsten filaments for incandescent lamps were produced in Hungary by the method patented by Just and Hanamann.

In his patent of 1909, W. D. Coolidge described the production of ductile tungsten wire by powder metallurgy, a process that has not changed in its essential features even today. This process represents the first industrial use of powder metallurgy.

In 1922, the Osram Research Association obtained a patent for the production of tungsten carbide – cobalt hard metals.

Four years later, the Krupp organization introduced the first hard metal cutting tool at the Leipzig Exhibition.

The increasing use of hard metals in the machining of steel and in the mining industry (energy production) led to a great increase in the demand for tungsten. As a result of the uncertain availability of raw materials in crisis situations and the increased use of tungsten materials in the armaments industry (hard metals, tungsten-containing heavy metals), the metal has become of strategic importance.

All later materials developments are essentially only a broadening and diversification of the early work on the production of the metal and

its alloys. Tungsten compounds are becoming increasingly important, e.g., as the active components in catalysts.

Lattice type	A 2, bcc
Lattice parameter	0.31648 nm
Atomic radius	0.139 nm
Ionic radius	
(4-valent)	0.064 nm
(6-valent)	0.068 nm
Absorption cross section for thermal neutrons	19.2 barn
1st ionization potential	8.0 eV
mp	3410 °C
bp	5900 °C
Heat of fusion	35.17 kJ/mol
Heat of sublimation	850.8 kJ/mol
Molar heat capacity at 20 °C	24.28 J mol^{-1} K^{-1}
Specific heat capacity	
at 25 °C	0.135 kJ kg^{-1} K^{-1}
at 1000 °C	0.17 kJ kg^{-1} K^{-1}
at 2000 °C	0.20 kJ kg^{-1} K^{-1}
Coefficient of thermal expansion	
at 500 °C	2.3 mm/m
at 1500 °C	7.6 mm/m
at 3000 °C	19.2 mm/m
Thermal conductivity	
at 0 °C	129.5 W m^{-1} K^{-1}
at 1000 °C	112.5 W m^{-1} K^{-1}
at 2000 °C	96.0 W m^{-1} K^{-1}
Vapor pressure	
at 1700 °C	1×10^{-10} Pa
at 2300 °C	4×10^{-1} Pa
at 3000 °C	2×10^{-3} Pa
Standard entropy	
$S_{25}^{0\ \circ C}$ at 101 kPa	32.76 J mol^{-1} K^{-1}
Standard enthalpy	
$H_{25}^{0\ \circ C} - H_{-273}^{0\ \circ C}$	5.09×10^{3} J/mol
Specific electrical resistivity	
at 20 °C	0.055 μΩ · m
at 1000 °C	0.330 μΩ · m
at 2000 °C	0.655 μΩ · m
at 3000 °C	1.40 μΩ · m
Magnetic susceptibility	
at 25 °C	0.32×10^{-6} cm^3/g
Black body temperature at 727 °C (λ = 665 nm)	693 °C
Spectral emission at 25 °C (λ = 665 nm)	0.47
Total radiation	
at 527 °C	0.173
at 1127 °C	3.82
at 2327 °C	80.60
Electron emission	
at 727 °C	1.07×10^{-15} A/cm^2
at 1627 °C	2.28×10^{-4} A/cm^2
at 2227 °C	2.98×10^{-1} A/cm^2
Electron yield	
at 727 °C	1.77×10^{-5} A/W
at 1627 °C	1.22×10^{-5} A/W
at 2227 °C	4.26×10^{-5} A/W
Density at 20 °C	19.3 g/cm^3

(Continued)

Hardness (Vickers) HV 30	
at 0 °C	450
recrystallized	300
deformed	≤650
at 400 °C	240
at 800 °C	190
Modulus of elasticity	
at 0 °C	407 kN/mm^2
at 1000 °C	365 kN/mm^2
at 2000 °C	285 kN/mm^2
Velocity of sound	
longitudinal	5320 m/s
transverse	2840 m/s
Minimum compression strength (sintered)	1150 N/mm^2
Shear modulus at 20 °C	177 kN/mm^2

Continuously increasing demand up to the 1980s led to the constant exploration of tungsten deposits and the opening of mines and processing plants. However, owing to a fall in price and demand, a large number of mines have been closed since 1985, especially in the industrialized countries.

In terms of quantity, tungsten is the second most important high-melting metal, after molybdenum; however, it is the most important refractory metal used in powder metallurgy.

General information on tungsten and its alloys and compounds can be found in [1–7]. A summary of the early history can be found in [8].

2. Properties

2.1. Physical Properties

Tungsten has the highest melting point of all metals, exceeded only by elemental carbon and the metallic monocarbides of niobium, zirconium, tantalum, and hafnium. The extremely high bonding energy (due to the half-filled 5*d* shell) leads to extreme values of other properties. For example, tungsten has the lowest vapor pressure of all metals, the lowest compressibility, an extremely high density, a high modulus of elasticity, low thermal expansion, and high thermal conductivity.

Some physical properties of tungsten follow [1], [3], [9]. (see left column)

Mechanical and Metallurgical Properties. Pure metallic tungsten has a body-centered cu-

bic structure, is brittle at room temperature, and is therefore not suitable for cold forming. At elevated temperatures (ca. 100 – 500 °C) it is transformed into the ductile state. The temperature of this sharply defined transformation is mainly dependent on the purity and the history of deformation and heat treatment of the material. The transformation temperature is increased by the presence of extremely small amounts of interstitially soluble elements, e.g., oxygen, carbon, and nitrogen, which lead to intergranular precipitation. The transformation temperature decreases with increasing deformation; highly deformed products such as thin sheet and wire are ductile at room temperature. This is mainly due to the distribution of the precipitated impurities over a larger intergranular area, and the formation of a fine-grained fibrous structure which tends to prevent intergranular failure perpendicular to the fiber direction. In the fiber direction, the tendency toward intergranular failure is more or less constant.

Addition of rhenium can cause the transformation temperature to fall below room temperature, even for only slightly deformed products (see Section 5.1). With extremely pure single crystals, brittle – ductile transformation temperatures as low as −190 °C are observed.

The grain size and grain structure have a great influence on the mechanical properties of tungsten. These can be controlled by sintering conditions, type of deformation, degree of deformation, and annealing processes during and after machining (intermediate annealing, stress-relieving annealing, soft annealing, recrystallization annealing, see Sections 4.6, 4.7).

Consequences of the high bonding energy of tungsten include its high ultimate strength and yield strength, high modulus of elasticity, and high recrystallization temperature.

The hot strength of tungsten, like its elongation, depends on its deformation and annealed condition, and can vary widely at a given temperature (Fig. 1). The hot strength of tungsten is exceeded only by that of rhenium. The upper limit for the hot tensile strength corresponds to a stress-relieved sheet, and the lower limit to a recrystallized sheet. At ca. 1500 °C, both limiting curves coincide, i.e., at this temperature the tungsten completely recrystallizes as it is heated to the tensile testing temperature (and therefore at temperatures above this) independently of the starting condition of the tungsten, so that is always has the hot strength of recrystallized tungsten. The results of hot tensile testing are to some extent influenced by the strain rate.

Figure 1. Hot tensile strength of tungsten sheet, thickness 1 mm, strain rate 15 % min^{-1} [3]

The relation between temperature and fracture strain of recrystallized tungsten sheets is shown in Figure 2.

Figure 2. Effect of temperature on fracture strain of recrystallized tungsten sheet [3]

No allotropic transformations of group 6 metals are known which could lead to changes in their properties. The solubility of the nonmetals carbon, silicon, oxygen, nitrogen, and water in these metals is very small. Therefore, impurities in the form of interstitially dissolved nonmetals have a smaller effect on the mechanical properties than is the case with other transition metals, but have a very significant effect if the solubility limits are exceeded. Carbides, silicides, nitrides, and oxides are then formed [9], and the inclusions have a marked effect on grain shape and size.

As the solubility of interstitial impurities (O, N, C) at room temperature is well below 0.1 µg/g, and the solubilities, even at the eutectic temperature, are only 0.01 – 0.1 % [10], the insoluble excess segregates at the grain boundaries on cooling. Oxide impurities, in particular, considerably reduce intergranular strength. With carbide impurities, a beneficial effect on ductility is apparent, depending on the form of the carbides. The effect of nitrides is not yet fully understood. In deformed tungsten, it leads to

marked anisotropy of the mechanical properties in the direction of deformation, so that sheet and bar tend to split. In recrystallized tungsten, the intergranular impurities enable plastic deformation to take place above the transformation temperature, as already mentioned [9].

2.2. Chemical Properties

Tungsten is stable in air up to 350 °C, but begins to oxidize above 400 °C. A thin film of blue tungsten oxides forms on the surface, and this at first prevents further attack. As the temperature increases, cracks develop in the oxide film, and these favor oxidative attack. Above 800 °C, rapid oxidation takes place with formation of WO_3, which sublimes. In oxygen, tungsten burns at 500 – 600 °C. Tungsten is stable to mineral acids in the cold, and is only slightly attacked at higher temperatures. Its stability to cold and hot hydrofluoric acid is of practical importance. Mixtures of hydrofluoric and nitric acids dissolve tungsten very readily, the tungsten remaining in solution because of the complexing effect of the fluoride ions. Hydrogen peroxide is a good solvent for finely divided tungsten; the peroxo complexes so formed keep the tungsten in solution. Alkaline solutions do not attack tungsten, which is also very stable to molten alkalis. In the presence of oxidizing agents (Na_2O_2, $NaNO_2$, or $NaNO_3$), rapid dissolution takes place.

Tungsten is stable to molten glass and silica up to 1400 °C. At higher temperatures, it reacts with many elements, e.g., B, C, Si, P, As, S, Se, Te, and the halogens. It reacts with fluorine at lower temperatures. It is stable to chlorine up to 250 °C, and to bromine and iodine up to 500 °C. It is very stable to gaseous nitrogen and ammonia up to 1400 °C. There is no reaction with hydrogen. With carbon monoxide, the hexacarbonyl is formed at low temperatures, and the carbide above 800 °C.

3. Raw Materials

3.1. Natural Resources

The average concentration of tungsten in the earth's crust is 1.5 g/t. It is thus a rare element.

Minerals. Tungsten occurs only in combined form. The most important tungsten minerals are [4]:

Anthoinite	$Al(WO_4)(OH) \cdot H_2O$
Cuprotungstite	$Cu_2(WO_4)(OH)_2$
Ferritungstite	$Ca_2Fe_2^{2+}Fe_2^{3+}(WO_4)_7 \cdot 9\,H_2O$
Raspite	$PbWO_4$
Russellite	Bi_2WO_6
Sanmartinite	$ZnWO_4$
Scheelite	$CaWO_4$
Stolzite	$PbWO_4$
Tungstenite	WS_2
Tungstite	$WO_3 \cdot H_2O$
Ferberite	$FeWO_4$ with up to 20 % $MnWO_4$
Hübnerite	$MnWO_4$ with up to 20 % $FeWO_4$
Wolframite	$(Fe, Mn)WO_4$

With the exception of tungstenite, all the minerals are tungstates, but only scheelite and the wolframite group are of industrial importance. The most important properties of scheelite and the wolframites are listed in Table 1 [4].

Scheelite is often associated with the isomorphous calcium molybdate (powellite). Pure scheelite has a blue fluorescence in ultraviolet light, becoming white with a molybdenum content of ca. 1 %, and yellow above this concentration. This property is utilized in prospecting for scheelite.

Wolframite is a general term for a mixed crystal series (with no miscibility gap) of Fe(II) and Mn(II) tungstates. The minerals between pure iron tungstate and 20 % $MnWO_4$ are known as ferberite. Manganese tungstate and mixed crystals containing up to 20 % $FeWO_4$ are known as hübnerite.

The minerals tungstite and cuprotungstite have achieved a small degree of industrial importance.

Deposits. All tungsten deposits are of magmatic – hydrothermal origin. During cooling of the magma, differential crystallization occurs, first of basic and later of slightly acidic rocks, and enrichment of tungstate and other ions takes place, accompanied by an increase in the concentration of silica in the remaining magma and the hydrothermal solution. Tungsten cannot replace the main elements such as Fe, Mg, Ca, Al, and Si in the basic and slightly acidic rocks, and is not incorporated in the first material to crystallize. Therefore, scheelite and

Table 1. Physical and chemical properties of wolframites and scheelite

	Ferberite	Wolframite	Hübnerite	Scheelite
Formula	$FeWO_4$	$(Fe, Mn)WO_4$	$MnWO_4$	$CaWO_4$
WO_3 content, %	76.3	76.5	76.6	80.6
Crystal structure	monoclinic	monoclinic	monoclinic	tetragonal
Lattice parameters				
a, nm	0.471	0.479	0.485 *a*	0.5257
b, nm	0.570	0.574	0.577 *c*	1.1373
c, nm	0.574	0.499	0.498 *a/c*	1 : 2.165 *a/c*
β	90°	90° 26′	90° 53′	
Density, g/cm^3	7.5	7.1 – 7.5	7.2 – 7.3	5.4 – 6.1
Color	black	dark gray – black	red-brown – black	brown, yellowish, white
Hardness (Mohs)	5	5 – 5.5	5	4.5 – 5
Common form	well-formed crystals or crystal masses	irregular crystal masses or radiating crystal groups	radiating groups of lamellar crystals	crystals, mainly fine grained

wolframite are often to be found in the last crystallization stage of a magma, i.e., the pegmatites and the material crystallized from hydrothermal solutions.

There are many types of deposit:

1) Pegmatite deposits are relatively rare and can contain wolframite and scheelite. The tungsten minerals are crystallized together with the pegmatitic minerals.
2) Greisen deposits contain wolframite, and are formed by the mineralization of intrusive rock (and of the rock it penetrates) by hydrothermal solution processes.
3) Contact-metasomatic deposits are the commonest type and are of pure scheelite. They are formed by the reaction between limestone and residual magma or hydrothermal solutions that penetrate into it.
4) Vein deposits can contain wolframite and scheelite. Hot magma causes cracks and faults in the surrounding rock, into which hydrothermal solution is injected. It then crystallizes.
5) Pneumatolytic deposits are rare, and are formed by the transport of volatile tungsten compounds in the gas phase.
6) Secondary enriched deposits are very rare and contain tungstite formed by weathering of wolframite.
7) Placer deposits, in which high density materials become concentrated in river sediments, are also very rare.

The gangue minerals differ not only among the various deposits, but can also vary greatly within a single mine. This is easily understood, as the various types of deposits can often occur together in one mine. The commonest types of gangue minerals are: quartzite, various silicates, limestone, magnesite, fluorspar, apatite, and various sulfide ores.

The concentration of workable ores is 0.1 – 2.5 % WO_3, contents of ca. 2 % being rare. The average content is ca. 0.5 % WO_3. The lower limit of workability depends on the ore type, gangue composition, and world market price. It fluctuates between 0.1 and 0.3 % WO_3.

Reserves. Tungsten deposits occur throughout the younger mountain ranges, e.g., the Alps, the Himalayas, and the circum-Pacific belt. World tungsten reserves (in 10^3 t) in 1985 are listed here [11]. (see below)

These figures include reserves that are known, but have so far not been found to be economically workable. This type of information is partly based on estimates, as countries such as the People's Republic of China and the former Soviet countries have not published any accurate information. By far the largest reserves are in China (35 %, but may exceed 50 %), followed by Canada (19 %), and the former Soviet Union (14 %). At the current world annual consumption rate of 30 000 t, these known reserves of ore are likely to last for ca. 120 years. Of the ore deposits, 75 % are of scheelite and 25 % wolframite, with a few in which both minerals occur together [12]. Over half the reserves of ore are in developing countries.

North America	985
United States	290
Canada	670
Mexico	20
Others	5
South America	110
Bolivia	70
Brazil	20
Others	20
Europe	665
Austria	20
France	20
Portugal	40
Former Soviet Union	490
United Kingdom	70
Others	25
Africa	20
Rwanda	5
Zimbabwe	5
Others	10
Asia	1535
Burma	15
People's Republic of China	1230
North Korea	100
South Korea	60
Malaysia	20
Thailand	30
Turkey	70
Others	10
Oceania	10
Australia	140

3.2. Tungsten Scrap

Tungsten scrap has been used as a raw material for production of pure tungsten for more than 50 years. There are several reasons for this:

The low concentration in the earth's crust results in a consistently high cost of the metal; recycling has always been economic.

The concentration of tungsten in the various types of scrap is always considerably higher (30 – 99 %) than in any ore, so there are no concentration costs.

The accompanying elements in scrap do not cause serious problems during chemical processing.

High demand and consequent high prices from 1975 to 1985 led to an increase in scrap processing, and made lower-grade scrap (grinding wheel swarf and dusts) more attractive for reuse [13]. Since then, the amount of scrap used as a percentage of the total tungsten raw materials has increased still further in spite of falling price and demand. This is largely because environmental regulations have become increasingly strict, making it impossible to dispose of scrap with even a low tungsten content, e.g., wheel swarf or similar materials. Consequently, scrap is classified as a primary tungsten raw material. This is shown in Figure 3, which also shows that approximately one-third of demand is supplied from scrap [14].

Figure 3. Tungsten flowchart

The processing of scrap is environmentally beneficial. No energy is required for mining and preparation, and no chemicals for concentration processes. This reduces both atmospheric pollution and chemical waste. Especially important for the future is the fact that scrap recycling conserves reserves of ore.

Scrap is classified as either "hard" or "soft". Soft scrap is finely divided material, e.g., powder, dust, grindings, and turnings, with tungsten content 10 – 98 %. Hard scrap is in lump form (sintered material, scrap bar and sheet, wire, etc.) with tungsten content 40 – 99.9 %.

4. Production

4.1. Ore Beneficiation

Most tungsten ores contain less than 1.5 % WO_3. Therefore, ore-dressing plants are always in close proximity to the mine to keep transport costs low. The first steps in the procedure are crushing and milling to liberate the tungsten mineral crystals from the gangue materials. The final grain size to which the ore must be reduced depends on the size of mineralization present in the tungsten ore body. This can be quite different from deposit to deposit. The finer the mineralization, the more costly the pretreatment and the lower the recovey. No more crushing and milling should be applied than is absolutely required, to minimize losses.

Scheelite Ore. After crushing, coarsely crystallized scheelite can be concentrated by gravimetric methods (hydrocyclones, tables, spirals, etc.). Finely mineralized scheelite, however, has to undergo froth flotation. A very high purity may be obtained by gravity methods (sometimes >90 % scheelite in the concentrate). The yield is normally 60 – 70 % and sometimes 80 – 90 %. The flotation process allows recoveries of 80 – 90 %. Modern plants combine gravimetric and flotation methods because on disintegration some of the scheelite crystals are broken into fine particles which are lost if only gravity methods are used (scheelite is very brittle). The fine particles can be recovered in a subsequent flotation step. Typical commercial concentrates contain >70 % WO_3. Fully downstream integrated plants that process their concentrates themselves keep the final WO_3 concentration in the flotation process low (5 – 40 %) to minimize losses, giving overall yields of >90 %.

Wolframite Ore. Gravity methods are used for wolframite ore, sometimes in combination with magnetic separation. The weakly magnetic wolframite properties enable a further concentration. Commercial concentrates should assay >65 %.

4.2. Pretreatment of Ore Concentrates and Scrap

Hydrochloric acid leaching of scheelite concentrates at room temperature is used to reduce phosphorus, arsenic, and sulfur content [4].

Calcination in air (500 – 600 °C) is used to oxidize organic flotation agents which could have a detrimental effect on subsequent processing [15], [16]. Also, in mixtures of concentrates with high sulfide and calcium carbonate contents, gypsum is formed [16], and arsenic and sulfur are removed if they are present in high concentrations in wolframites.

Grinding is necessary for scheelite concentrates produced by gravity methods, and with all wolframite concentrates.

Oxidation of soft scrap by heating in air at 700 – 800 °C allows the tungsten to be dissolved in subsequent pressure leaching processes [17].

4.3. Hydrometallurgy

This section deals with the production of high-purity ammonium paratungstate (APT), which is now the most important raw material for all other tungsten products. Only the modern process is described; a comparison with classical methods [7] is given in Figure 4. The most important differences are:

1) The use of solvent extraction instead of precipitation of $CaWO_4$ to convert the sodium tungstate solution to ammonium tungstate solution, followed by decomposition to H_2WO_4.
2) Conversion of the tungsten in aqueous solution to a solid compound by crystallization rather than precipitation. Each crystallization is a slower process than immediate precipitation, and therefore leads to a purer product.

The advantages of the modern process are:

1) The possibility of using various raw materials (wolframite, scheelite, scrap)
2) Better energy efficiency
3) Less labor required
4) Ease of automation
5) Higher yields
6) Uniformity and higher purity of the product

A flow diagram of a modern APT production plant based on various raw materials is shown in Figure 5.

4.3.1. Digestion

Wolframite concentrates can be dissolved either at atmospheric pressure with concentrated NaOH solution (40 – 50 %, 15 mol/L) at 100 – 145 °C followed by dilution [18], or by high-pressure digestion with dilute NaOH solution [19].

$$(Fe, Mn)WO_4 + 2\,NaOH \rightarrow Na_2WO_4 + (Fe, Mn)(OH)_2$$

Scheelite concentrates are always leached with Na_2CO_3 solution under pressure [4]. The process requires a large excess of reagent, but gives a milder reaction with respect to dissolution of gangue, and can be carried out in mild steel autoclaves (NaOH treatment requires Inconel cladding).

Figure 4. Comparison of classical and modern WO_3 production

Figure 5. The modern APT process

Figure 6. Tungsten solvent extraction

$$CaWO_4 + Na_2CO_3 \rightarrow Na_2WO_4 + CaCO_3$$

Mixtures of concentrates of wolframite and scheelite can be dissolved by Na_2CO_3 solutions with addition of NaOH under pressure [20]. Oxidized soft scrap is digested with NaOH solution, as in the treatment of wolframite [15].

Information on the parameters and yields for various chemical leaching processes is given in Table 2.

4.3.2. Purification

The sodium tungstate solution obtained by leaching the ore contains amounts of dissolved impurity elements depending on the composition of the raw material. If the concentration of these elements is too high, subsequent processing may be disturbed or the end product may have too high an impurity level. Purification is carried out by precipitation and filtration.

Silicate Precipitation. Silicates are common gangue minerals which dissolve on pressure leaching, at least partially. Dissolved silicate can be precipitated by the addition of an aluminum sulfate or magnesium sulfate solution, or a mixture of the two at pH 8 – 11 [21]. The chemistry of the precipitation is complex and has not been investigated in detail. With higher silicate content, a two-stage precipitation with a mixture of aluminum and magnesium sulfates is preferable, enabling the SiO_2 content to be reduced to 30 –60 mg/L [4].

This precipitation technique reduces not only the silicate content, but also the phosphate and fluoride concentration. Precipitation is first performed with magnesium sulfate at pH 10 – 11, and then with aluminum sulfate at pH 7 – 8 [22].

Molybdenum Precipitation. Molybdenum is a very common companion element in tungsten ores. It remains in the concentrate and goes into solution on chemical digestion, either completely or partially. Precipitation is carried out by adding excess sodium sulfide to the neutral or mildly alkaline solution, forming thiomolybdate:

$$MoO_4^{2-} + 4\,S^{2-} + 4\,H_2O \rightarrow MoS_4^{2-} + 8\,OH^-$$

Table 2. Typical parameters for pressure leaching of tungsten raw materials

	Scheelite concentrate	Woframite concentrate	Oxidized scrap
Particle size, µm	<44 to <90	<44	<100
Temperature, °C	190 – 225	175 – 190	150 – 200
Pressure, MPa	1.2 – 2.6	0.8 – 1.2	0.5 – 1.2
Time, h	1.5 – 4	4	2 – 4
Reagent concentration, %	10 – 18 (Na_2CO_3)	7 – 10 (NaOH)	20 (NaOH)
Molar ratio, WO_3 : reagent	1 : 2.5 – 4.5	1 : 1.05	1 : 1.4

The solution is then acidified with sulfuric acid to pH 2.5 – 3.0. This causes the molybdenum to be precipitated as the trisulfide [20]:

$$MoS_4^{2-} + 2\ H_3O^+ \rightarrow MoS_3 + H_2O + H_2S$$

This purification stage is not only effective for Mo, but also removes a number of other elements that form insoluble sulfides, e.g., As, Sb, Bi, Pb, and Co.

4.3.3. Conversion of Sodium Tungstate Solution to Ammonium Tungstate Solution

In modern plants, this reaction is now carried out exclusively by solvent extraction or ion exchange resins. The sodium ion concentration must be reduced from ca. 70 g/L to <10 mg/L, as sodium levels >10 ppm in the APT cause problems during the reduction to metal powder.

4.3.3.1. Solvent Extraction [4], [23], (Fig. 6)

The sodium tungstate solution (pH 2 – 3) is contacted with the organic phase, which extracts the isopolytungstate ions as an ion associate complex. Mixtures of tertiary or secondary aliphatic amines (e.g., trioctylamine) with isodecanol and kerosene form the organic phase. The amine is the reagent that forms the ion associate complex, the isodecanol increases the solubility of the complex (modifier), and kerosene is the solvent.

$$5\,R_3NH^+ + [HW_6O_{21}]^{5-} \rightarrow (R_3NH)_5HW_6O_{21}$$

The extract is washed with deionized water, and the isopolytungstate is then reextracted with dilute ammonia solution into an aqueous phase:

$$(R_3NH)_5HW_6O_{21} + 5\,OH^- \rightarrow [HW_6O_{21}]^{5-} + 5\,(R_3NH)OH$$

The solvent goes through a regeneration stage and is then returned to the extraction process.

4.3.3.2. Ion Exchange Process [24], [25]

This process is now used only in small plants in the People's Republic of China. A strongly alkaline ion exchange resin in the chloride form is used. On contact with the sodium tungstate solution, the tungstate is adsorbed. Desorption is carried out with ammonium chloride solution. Elements that form heteropolytungstates, e.g., Si, P, As, and Mo, can be removed.

4.3.4. Crystallization of Ammonium Paratungstate (APT) [26]

This compound has the formula $(NH_4)_{10}(H_2W_{12}O_{42}) \cdot 4\,H_2O$.

On evaporation of the isopolytungstate solution, water and ammonia are distilled, the pH of the solution decreases, and the sparingly soluble APT crystallizes [27], [28]. The ammonia – water mixture is condensed and recycled to the solvent extraction process. The crystallization conditions are used to control the physical properties that influence the subsequent processing to give tungsten powder. The process is carried out batchwise or continuously in recirculating crystallizers with partial recycling of mother liquor (to maintain the concentration of impurities in the mother liquor and hence in the crystals formed). Owing to the low solubility of APT compared with the impurities in the mother liquor, this process step represents not only a conversion from the dissolved form to a solid, but also a very important purification (Table 3).

Table 3. Maximum impurity concentrations in APT crystallization

Element	Feed solution, mg/L	Mother liquor, mg/L	APT, ppm
Al	<10	<100 ∗ <3000 ∗∗	<10
As	<50	<2000	<20
F	<250	<3000	<10
Fe	<10	<200	<10
Mo	<10	<60	<20
Na	<10	<100	<10
P	<50	<400	<20
Si	<10	<200	<20
V	<100	<1200	<20

∗ In absence of fluoride.
∗∗ For fluoride concentration of ca. 4000 mg/L.

4.4. Calcination of APT [4], [21]

On heating APT to 400 – 800 °C, ammonia and water are evolved. This decomposition becomes more complete as the temperature and duration of heating are increased. The form of the solid phase is determined by the decomposition temperature, the retention time in the reactor, and the reduction potential of the decomposition atmosphere. If heating takes place with air addition, yellow WO_3 is obtained; with exclusion of air, blue oxide. In the blue oxide, some tungsten atoms are pentavalent, owing to H_2 reduction (NH_3 decomposition). Unlike the yellow oxide, blue oxide is not a single, well-defined chemical compound. It contains WO_3 together with the oxides $W_{20}O_{58}$ and $W_{18}O_{47}$, depending on the decomposition conditions, and also ammonium and hydrogen tungsten oxide bronzes (residual NH_3 and H_2O content). Nevertheless, it is currently used more than the yellow oxide.

On an industrial scale, calcination is usually carried out in rotary furnaces, more rarely in pusher furnaces. The former yield a more uniform product. The chemical composition and micromorphology of the oxide particles are determined by the duration and temperature of heating. The micromorphology is very important in the production of submicron tungsten powders.

4.5. Reduction by Hydrogen to Tungsten Metal Powder [29–31]

The production of tungsten metal powder is today carried out almost exclusively by the hydrogen reduction of high-purity tungsten oxides (WO_3, WO_{3-x}). Reduction of the oxide by carbon is now used only for the production of tungsten carbide (direct carburization). The reduction of tungsten halides (Axel Johnson process) has not become established on a large scale.

The reduction of tungsten trioxide or blue oxide can be controlled by the temperature and the water vapor partial pressure.

At lower temperatures (600 °C) and lower moisture contents, the reduction proceeds by solid-state diffusion of the oxygen out of the oxide, and can be represented by the following equation:

$$WO_3\,(s) + 3\,H_2\,(g) \rightarrow W\,(s) + 3\,H_2O\,(g)$$

On complete reduction, an oxide-pseudomorphic metal sponge is formed.

At higher temperatures and moisture contents, the reaction proceeds stepwise via the oxides $W_{20}O_{58}$, $W_{18}O_{47}$, and WO_2 to tungsten metal. Only the first transformation to $W_{20}O_{58}$ and each nucleation of a newly formed phase on further reduction are solid-state reactions. All other transformations are linked with gas-phase transport of tungsten as the volatile oxide hydrate $WO_2(OH)_2$, and are associated with a significant change in shape of the crystal faces. The formation of the vapor-phase oxide hydrate takes place by reaction of the corresponding suboxide with water formed in the reduction. The vapor pressure of the oxide – hydrate in equilibrium with a stable compound formed under the reaction conditions is lower than that in equilibrium with the continually decreasing amount of the unstable compound. As a result of these

gradients, chemical vapor transport (CVT) takes place, so that the higher oxide decreases in quantity, and the lower oxide increases. The interphase reaction can plausibly be divided into three steps:

1) Formation of $WO_2(OH)_2$ from the oxygen-rich compound with H_2O
   $WO_2\,(s) + 2\,H_2O\,(g)$
   $\rightarrow WO_2(OH)_2\,(g) + H_2\,(g)$
2) Transport of $WO_2(OH)_2$ to the compound of lower oxygen content
3) Reduction of $WO_2(OH)_2$ by hydrogen at the surface of the compound of lower oxygen content (or the metal)
   $WO_2(OH)_2\,(g) + H_2 \rightarrow W\,(s) + 4\,H_2O\,(g)$

Formation of a new phase must be preceded by nucleation. This takes place on the surface of the reacting oxide if the water vapor pressure is lower than the equilibrium water vapor pressure for the phase transformation concerned. The nucleus then continues to grow as described above.

Reduction on an industrial scale is carried out in pusher furnaces, in which the powder passes through the furnace in boats, or in rotary furnaces. Walking beam furnaces or furnaces with internal band conveyors are less often used. Furnaces are provided with several temperature zones controllable between 600 and 1100 °C. The flow of hydrogen is usually in a countercurrent direction, more rarely cocurrent, and a drying stage is always included in the circuit, as a large excess is used. The hydrogen acts not only as a reducing agent, but also carries away the water formed. The flow rate is therefore one of the parameters for controlling the water vapor partial pressure. Other control parameters are:

1) The amount of oxide charged to the furnace per unit time
2) The depth of the oxide bed
3) The porosity of the oxide bed (particle size and particle size distribution)
4) The moisture content of the hydrogen entering the furnace

The porosity and depth of the bed, in conjunction with the H_2 flow rate, determine the rate of diffusion of water vapor from the bed.

In practice, the average grain size of the tungsten powder (0.1 – 60 µm) is controlled empirically by setting the temperature, oxide quantity, heating time, and H_2 flow rate. In general, these have the following effects:

1) Fine Powder: lower temperature, low bed depth, long dwell time, and high H_2 flow rate. These conditions correspond to a low water vapor partial pressure (and hence a more rapid nucleation rate), low crystal growth rate via the gas phase, and short transport distances of $WO_2(OH)_2$. The furnace capacity is low.
2) Coarse Powder: high temperature, greater bed depth, lower heating time, lower H_2 flow rate. These conditions correspond to a high water vapor partial pressure, a lower rate of nucleation, and more rapid transport via the gas phase over longer distances. The furnace capacity is high.

In the powder bed, there is a water vapor concentration gradient, as diffusion from the interior takes longer than from the outer parts. This leads to a grain size distribution in the tungsten powder; the powder in the interior of the bed is coarser than that in the outer zones and at the surface.

Doping with low concentrations of alkali metals (Na or Li salts) catalyzes crystal growth during the reduction process, and enables coarser tungsten powders (100 – 200 µm) to be produced [32].

4.6. Production of Compact Metal

Conversion of tungsten metal powder to compact metal is carried out exclusively by powder metallurgy, i.e., by a heat treatment (sintering) of a compacted preform.

Compaction (→Powder Metallurgy and Sintered Materials, Chap. 5.). In the first step, the powder is compacted hydraulically or mechanically by a polydirectional pressure of 200 MPa (2000 bar) to produce preforms – rods or plates that have edge stability and are of the correct dimensions. Circular blanks with a low height/diameter ratio and small items such as pins are produced in automatic presses at high outputs. The raw material for this must consist of free-flowing granules similar to the granulated hard metals produced in spray drying equipment [33].

Sintering (→Powder Metallurgy and Sintered Materials, Chap. 6.). Pure tungsten is today sintered almost exclusively by the economic indirect sintering process. Furnaces for this consist of tungsten heating elements surrounded by radiation shielding made of tungsten sheets on the inside and molybdenum on the outside, and are purged with hydrogen as a protective gas. At maximum sintering temperature (2800 °C), sintering times of several hours are required. In the production of so-called non-sag tungsten for incandescent lamp manufacture, direct sintering, i.e., by passage of an electric current, is usual. In both processes, the sintered densities achieved are ca. 95 % of theoretical.

The weight of individual sintered tungsten components is usually 1 – 100 kg. The choice of optimum sintering parameters greatly influences the structure and hence the strength and ductility of sintered tungsten.

In the direct sintering process, very high sintering temperatures, and hence short sintering times can be used. However, the practical limits on rod dimensions and production rates are a considerable disadvantage, and electricity costs per kilogram are significantly higher than with indirect sintering. Also, the fixing points, which are poorly sintered, must be knocked off the rod, causing further loss. Disadvantages of the indirect sintering process include the high cost of the plant and heating elements. Nevertheless, overall this process is considerably more economic, and it is the only process that allows the sintering of preforms.

Another important aspect is that sintering is an extremely effective purification process, removing a high proportion of impurities from the material. The hydrogen flow rate during sintering should therefore be high enough to transport the impurities away.

Melting. Approximately since the 1950s, i.e., after the development of electric arc melting plants and high-capacity vacuum equipment, it has been possible to melt highly refractory metals on an industrial scale. Some time later, trials on plasma melting and electron beam melting of tungsten showed that these processes led to a product with extremely coarse crystals and hence poor working properties, so these melting processes are of no technical value for tungsten (unlike molybdenum and its alloys). Electron beam zone melting is in fact used in the production of single crystals of tungsten [5], but these products have only a few special uses as production costs are high.

4.7. Processing of Sintered Parts

4.7.1. Shaping

Completely dense, pore-free bodies cannot be obtained by sintering, the residual porosity being 3 – 10 %. To obtain a completely dense material, a complex multi-stage forming process is required. Tungsten can be shaped by all the usual forming processes, e.g., forging, rolling, extruding, and wire drawing. The first forming step is usually carried out at 1500 – 1700 °C. Many reheating stages are necessary in the first stages of shaping, as the heat is rapidly lost at these temperatures and a stress-relieving operation below the recrystallization temperature is often advantageous. Intermediate heating operations are necessary for recrystallization. The forming temperature is progressively reduced, since the recrystallization temperature decreases as forming proceeds. To give a well-defined fiber structure (high degree of deformation), a final cold forming can be carried out, as already stated.

Chipless Forming, Including Stamping and Cutting. The temperature range for the forming of tungsten has a lower limit, set by the transformation temperature, and an upper limit, set by the recrystallization temperature. Thin, strongly deformed sheet and foil have a pronounced structure in the longitudinal direction due to elongation of the grains during rolling. The bending properties along the direction of rolling are therefore different from those across it. Tungsten sheet should therefore always be bent perpendicular to the rolling or longitudinal direction. If bending in the longitudinal direction cannot be avoided, owing to the design, a much higher temperature is required.

Tungsten can also be formed at high temperatures by pressing, flow-turning (spinning), or forging.

Die stamping of tungsten sheet is possible if the working temperatures of the sheet and the tool are kept high.

Tungsten can be punched and forged at high temperatures. Blunt edges on punching tools and shears, or low punching or cutting temperatures can cause cracking and splitting.

Machining. Tungsten can be drilled, turned, milled, planed, and ground. However, machining operations require great experience and close adherence to optimum conditions [3]. Complex shapes and holes can be produced by spark erosion; the tungsten workpiece forms the anode and the working electrode the cathode. Suitable electrode materials can be based on tungsten – copper.

4.7.2. Mechanical Bonding of Tungsten to Tungsten and Other Metals

Rivets. Tungsten components can be bonded together or to components made of other metals by means of rivets, provided the joint need not be impermeable to liquids and/or gases. Tungsten rivets and round-head and countersunk molybdenum rivets can be used.

Brazing (→Soldering and Brazing). The parts to be joined must be free from grease, oils, oxides, or other impurities, and should preferably be etched immediately before brazing. As tungsten is very sensitive to oxidation, it is preferable to carry out the operation under a protective gas, hydrogen, or vacuum. Parts with a large surface area or of complex geometric shape should be brazed in a furnace with the aid of a special jig. If brazing temperatures exceed 1150 °C, there is a possibility of recrystallization. Typical brazing metals and temperatures are: Rh (1970 °C), Pd (1550 °C), CuNi 45 (1300 °C), etc. [3].

Welding (→Welding and Cutting). Tungsten has only moderate welding properties. The weld seams have a coarse-grained structure in the hot fusion zone, owing to recrystallization, and can therefore withstand only low mechanical stresses. If welding is unavoidable, weld seams should if possible be located in regions of low internal and mechanical stress. It is preferable to locate weld seams as far as possible from the edges or ends of fabricated articles.

Tungsten inert gas (TIG) welding can be performed in protective gas chambers. However, in electron beam welding, the melting and hot fusion zones are much more limited in size, so this technique is preferable for welding tungsten. It is advantageous to preheat the parts to be joined to ca. 700 – 800 °C by a defocused electron beam immediately before welding. Welded components should be cooled as slowlyas possible. Tungsten can be electron beam welded to tungsten – rhenium and molybdenum –rhenium alloys, copper, and stainless steel. Tungsten – rhenium alloys are recommended as filler metal for welding tungsten.

As fusion welding can be used only under certain circumstances, diffusion welding of tungsten to tungsten and to other metals is important. Despite the high melting point of tungsten, temperatures of 1300 – 2000 °C and pressures of 2 – 20 N/mm^2 give satisfactory joints. The diffusion welding of tungsten is carried out in either vacuum or pure hydrogen.

Thin intermediate layers in the form of films of nickel, platinum, rhodium, ruthenium, and especially palladium, considerably accelerate the diffusion processes.

Tungsten wires and small tungsten components can be bonded to fusible alloys based on iron – nickel – cobalt and to other metals by spot welding under a protective gas blanket. The use of small pieces of foil made of platinum, niobium, tantalum, or zirconium is recommended. Laser beam welding of tungsten is likely to become very important in the future.

4.8. Surface Treatment

Tungsten components can be degreased with trichloroethylene, perchloroethylene, and similar solvents.

Tungsten can be etched with alkaline and acid media. Whereas oxidized tungsten is more easily cleaned with alkaline etchants, grinding residues and metallic contamination are more easily removed by acids; very clean tungsten surfaces can be obtained by alkaline etching followed by acid etching.

An alkaline etchant can consist of molten NaOH containing 10 % $NaNO_2$. The melt temperature is 400 – 450 °C. Thin oxide films can also be removed by dipping tungsten in alka-

line solutions of $K_3[Fe(CN)_6]$. Other common alkaline etchants are 10 % solutions of NaOH or NH_3/H_2O_2. Acid etchants usually consist of $HNO_3/HF/H_2O$ mixtures. Other commonly used mixtures are $HNO_3/HF/CH_3COOH$ and $HClO_4/H_3PO_4/H_2O$.

Electropolishing. Tungsten can be anodically electropolished in a $K_2CO_3/KClO_3/NaOH/H_2O$ mixture. Another possible electrolyte is $CH_3OH/H_2SO_4/HF$.

4.9. Tungsten Coatings

Tungsten coatings are applied to metals, glass, ceramics, semiconductors, inorganic woven products, uranium carbide nuclear fuel granules, etc., for a large number of reasons, including the improvement of wear resistance, corrosion resistance, emissivity, and high-temperature properties [5]. The most important processes are:

Plasma Spraying (→Metals, Surface Treatment, Chap. 5.1.3.). This process is often used to apply very thick tungsten coatings (e.g., in the construction of engines and rockets), as the flame spraying process (used successfully with molybdenum) cannot be used with tungsten. The cathode material in the plasma burner is also tungsten.

Vacuum Vapor Deposition, Ion Sputtering, and Physical Vapor Deposition (PVD). In most of the usual coating processes, the substrate material must be raised to a high temperature, which usually has a detrimental effect on its mechanical properties. When tungsten is deposited by vacuum vapor deposition or ion sputtering onto a relatively cool substrate, these problems can be largely avoided. These coating methods are therefore very useful for metallizing semiconductors, ceramics, glasses, and also metals. The coatings are comparatively thin, but adhere well [34].

Deposition from the Vapor Phase, Chemical Vapor Deposition (CVD). Suitable readily vaporizable tungsten compounds include chlorides, fluorides [35], carbonyl compounds, and organometallic tungsten compounds. The carrier gas is hydrogen or a noble gas. Tungsten and tungsten alloys are applied in this way to, e.g., nuclear fuels (UC, UO_2), metallic substrates, and semiconductors. Small tungsten components are today manufactured by CVD.

Electrolytic Methods. Although there is an extensive patent literature [5] on the electrolytic deposition of tungsten coatings on metallic substrates, the method is not of practical significance. The production of tungsten coatings by diffusion, metallization, or cladding (including explosion cladding), has often been proposed, but has not become established as an alternative to other methods.

4.10. Production of High-Purity Tungsten Metal (99.999 – 99.9999 %)

The starting material is ammonium paratungstate (APT) of normal purity, but with uranium and thorium contents as low as possible. Important steps to high purity include:

1) Selective extraction of uranium and thorium [36]
2) Multiple crystallization of APT (giving a general reduction in the various trace impurities)
3) Working under clean-room conditions [38]
4) Electron beam zone melting [37], [39]

5. Tungsten Alloys

The addition of one or more elements or compounds to tungsten can improve its properties in a particular respect, or can combine the useful properties of tungsten with those of the additives. Depending on the additives, single-phase or two-phase alloys are produced. Powder metallurgy offers a simple route to the desired ratio of components, simply by mixing them. Of the large number of alloys described in the literature, only those of industrial importance are described here.

5.1. Single-Phase Mixed-Crystal Alloys

Tungsten – Molybdenum Alloys. Alloys with the composition Mo-10 W and Mo-30 W

have good corrosion resistance to molten zinc. In general, additions of tungsten increase the recrystallization temperature of molybdenum.

Tungsten – Rhenium Alloys. These have a higher recrystallization temperature than that of pure tungsten, a lower brittle – ductile transformation temperature, and higher high-temperature strength. They do not become brittle on recrystallization, and have superior welding properties. The alloying range of practical interest is 5 – 26 % Re.

The powder metallurgical method of production is from mixtures of tungsten and rhenium powder, or by coreduction by hydrogen of mixtures of yellow or blue oxide with ammonium perrhenate. The latter method ensures better uniformity.

High-temperature thermocouple elements, serviceable up to 2000 °C in protective gas atmospheres or in vacuum, are made from W- 5 Re or W- 26 Re wires.

The most important application area of W – Re alloys is in medical technology, where they are used in the manufacture of high-power rotating X-ray anodes, e.g., in computerized tomographic scanning equipment. The surface layer consists of a W-(3 – 10)Re alloy, while the body of the component is a Mo – W alloy, a Mo – Ti – Zr – O – C microalloy (TZM), or graphite [4], [33]. Compared with the pure tungsten anodes formerly used, the W – Re alloy can be operated at higher electron currents, allowing for shorter irradiation times. This is due to its greater stability to extreme alternating thermal stresses. The W – Re layer prevents roughening of the heated zones, and increases tube lifetime.

Tungsten – Tantalum Alloys. As with the molybdenum alloys, the only alloys of importance are those with high tantalum content. Developments have been principally aimed at pairing the good corrosion resistance and elasticity properties of tantalum with the superior strength and high-temperature strength of tungsten.

Ta-(7 – 8)W alloys are used in the manufacture of springs for use in strongly corrosive media [40], [41].

Ta-2.5 W is a corrosion-resistant construction material used in cases where the strength of pure tantalum is inadequate [42], [43].

Tungsten – Titanium Alloys. W-10 Ti is used as a sputtering target in the manufacture of microelectronic devices, such as VLSI or ULSI (very large or ultralarge scale integration), or DRAM (dynamic random access memory) chips.

5.2. Multiple Phase Alloys

Tungsten has excellent thermal properties and high density. However, for certain applications, its electrical and thermal conductivity, sensitivity to oxidation, and relatively poor machining properties are unsatisfactory. These limitations have led to the development of two-phase alloys in which the tungsten is the principal component. These are always produced by powder metallurgy.

Heavy Metals [44]. The term "tungsten heavy metal" means two-phase tungsten alloys of the type W – Ni – Fe (Ni 1 – 7 %, Fe 0.8 – 3 %, Mo 0 –4 %, magnetic) and W – Ni – Cu (Ni 1 – 7 %, Cu 0.5 – 3 %, Fe 0 – 7 %, nonmagnetic). The density varies between 17 and 19 g/cm^3, depending on the tungsten content.

They are produced by a powder metallurgical process involving liquid-phase sintering. Dense sintering takes place only slightly above the melting temperature of the eutectic mixture (1440 –1500 °C). As a result of the solubility of tungsten in the molten binder phase and the tendency of the interphase area of the system to decrease, recrystallization of the tungsten takes place, and hence a significant grain growth. The original mean particle size of the tungsten powder (3 – 5 µm) is increased in the sintered alloy (40 –100 µm). The structure consists of spherical tungsten grains that contain a few tenths of a percent of Ni and Fe in solid solution. These grains are embedded in a tough Ni – Fe – (Cu) matrix that contains ca. 4 % W in solution (Fig. 7).

In tungsten heavy metals, high modulus of elasticity and high electrical and thermal conductivity are combined with increased elongation and good machinability. This is due to the binder. Also, heavy metals have good absorption properties for γ- and X-rays, and are therefore used for radiation shielding. The properties of various heavy metal alloys are as follows [45]:

Density at 20 °C	17.0 – 18.5 g/cm^3
Modulus of elasticity	320 – 380 kN/mm^2
Shear modulus	125 – 160 kN/mm^2
Poisson's ratio	0.28 – 0.29
Hardness	
HV 10	270 – 470
HB 30	250 – 450
Tensile strength	660 – 1350 N/mm^2
Yield strength (0.2 %)	565 – 1300 N/mm^2
Fracture strain	2 – 15 %
Compressive strength	3500 – 5500 N/mm^2
Hot strength	
at 100 °C	560 – 730 N/mm^2
at 500 °C	340 – 610 N/mm^2
at 1000 °C	90 – 260 N/mm^2
Specific electrical resistivity	0.10 – 0.18 μΩ · m
Mean linear thermal coefficient of expansion, 20 – 800 °C	(5.22 – 6.45) $\times 10^{-6}$ mm^{-1} K^{-1}
Half-value thickness	
against ^{60}Co	8.5 – 9.7 mm
against ^{137}Cs	4.5 – 5.2 mm

Tungsten – Copper and Tungsten – Silver Alloys. Depending on the application, these can contain 20 – 50 % Cu or 10 – 80 % Ag. They combine the high hardness and high-temperature strength, the low welding tendency, and the excellent resistance to burning (by sparks and electric arcs) of tungsten with the high conductivity and good machinability of copper and silver, respectively.

Figure 7. Scanning electron micrograph of the fracture surface of non-sag tungsten after recrystallization, magnification 5000×

They are produced by a powder metallurgical process involving impregnation (infiltration). The tungsten powder, or a mixture of this with small amounts of the alloying element, is compacted and sintered at low temperature. The pores are then filled by infiltrating with a liquid phase. The amount of Cu or Ag required can be found from the "green" density (density of the compact prior to sintering). Infiltration can be by immersion, dipping, or flooding.

As the components W and Cu or Ag have practically no mutual solubility (pseudoalloys), no appreciable change in the particle size of tungsten takes place during infiltration, i.e., the tungsten skeleton approximately matches the particle size of the tungsten powder, though in densified form. The most important properties are listed in Table 4.

Non-Sag Tungsten (NS-W). This is a microalloy with metallic potassium. The technologically active alloying component, potassium, is present in trace concentrations (20 – 100 μg/g). The most important application is in the filaments of conventional incandescent lamps.

Non-sag tungsten wires have a so-called staple fiber structure (long, intertwined crystallites in an axial direction). This prevents the coiled wire from sagging when incandescent, and increases the creep strength at temperatures >2000 °C. The structure originates from small pores filled with metallic potassium, formed from larger potassium-filled bubbles by deformation (wire drawing) followed by heat treatment. The pores are arranged axially, i.e., in lines following the direction of deformation; they prevent slippage of dislocations and migration of grain boundaries on recrystallization. Grain growth is possible in the axial direction, and strongly inhibited in the radial direction.

The discovery of non-sag tungsten was accidental [47]. In the early 1920s, it was noticed that wire manufactured from tungstic acid calcined in special clay crucibles made by the Battersea Company of London had a more stable shape than wire from material calcined in standard crucibles. It was later reported that K, Si, and Al were picked up from the crucibles during the process. Since then, the oxide (formerly tungstic acid, but now blue oxide) is always doped with potassium silicate and the chloride or nitrate of aluminum.

On reduction of the oxide, finely dispersed phases of the doping elements K, Al, and Si (potassium aluminum silicates) are formed, and are then to some extent incorporated into the tungsten crystallites by overgrowth [48]. The doping elements mainly remain outside the powder particles, and are removed after the reduction process by washing with hydrofluoric acid. The final sintering evaporates any remaining dop-

Table 4. The most important properties of tungsten pseudoalloys [46]

Composition	Density, g/cm^3	Electrical conductivity, % IACS∗	Specific electrical resistivity, μΩ · m	Thermal conductivity, W m^{-1} K^{-1}	Hardness, HB 10	Modulus of elasticity, kN/mm^2	Tensile strength, N/mm^2
W-20 Cu	15.4	35	0.05	134	180	225	440
W-25 Cu	14.5	36	0.042	138	160	225	390
W-30 Cu	14.0	43	0.04	147	160	230	390
W-35 Cu	13.5	47	0.037	155	150	200	340
W-40 Cu	12.7	51	0.033	176	140	180	290
W-50 Cu	11.6	58	0.026	195	120	150	250
W-20 Ag	15.6	35	0.05	138	180	205	340
W-35 Ag	14.0	47	0.037	159	110	132	290
W-6.5 Ni 3.0 Fe	17.0	10 – 12	0.15 – 0.18	76	270	350	750
W-6 Ni 4 Cu	17.0	12.5 – 13.5	0.13 – 0.14	94	260	320	700

∗ International annealed copper standard.

ing elements still adhering. The ternary silicate phase included in the powder particles dissociates at higher temperatures. The elements Si, Al, and O, which have small atomic radii, can diffuse through the tungsten crystal lattice and thus volatilize. The potassium atoms are too large, and, as potassium is insoluble in tungsten, remain behind as potassium-filled bubbles. These break up to form lines of very small bubbles on subsequent processing (Fig. 8).

Figure 8. Microstructure of a tungsten heavy metal W-7 Ni 3 Fe, magnification 250×

The amounts of doping elements added to the oxide are: K 2000 – 3000 μg/g; Al 200 – 400 μg/g; and Si 2000 – 3000 μg/g, of which the amounts remaining in the sintered microalloy are only 20 – 100 μg/g, <20 μg/g, and <10 μg/g, respectively.

In addition to their use in incandescent filaments, NS-W microalloys are used in halogen lamps, resistance wires for heating motor vehicle windows, and, owing to their excellent resistance to corrosion by molten metals, as containers and heating elements in aluminum vapor deposition plants.

Alloys with Oxide and Carbide Dispersions. Fine, homogeneously distributed dispersions of oxides of the elements Th, Y, Zr, La, and Ce increase the thermal conductivity and recrystallization temperature of tungsten and its alloys. The commercially most important alloy is thoriated tungsten (1 – 4 % ThO_2), although the radioactive thorium is replaced, where possible, by rare earth oxides. It is mainly used in welding electrodes. Thorium increases the thermal emission of electrons, and ensures a constant, stable electric arc. The dispersion-strengthened alloys W – Re – ThO_2 and W – Re – HfC have the highest high-temperature strength of all known metallic materials [36].

Porous Tungsten. This is not itself an alloy, but can become one, e.g., by impregnation. Tungsten components can be produced with 15 – 35 % porosity, and are then impregnated with Cu to give good machinability. After machining, the Cu is removed by vacuum evaporation.

Porous tungsten impregnated with alkaline earth oxides is used to make dispenser cathodes with reservoirs in high-power transmitter tubes. Ion thruster motors for space travel also use porous tungsten as ion sources for the high-temperature cesium vapor.

6. Uses

The uses of tungsten, its alloys, and compounds are extremely diverse. They can be divided into the following large groups (by consumption):

Hard metals	50 – 60 %
Steel	20 – 30 %
Tungsten metal and its alloys	6 – 20 %
Tungsten chemicals (for uses, see Chap. 9)	4 – 8 %

Hard Metals. These products represent by far the largest consumption of tungsten worldwide. With few exceptions, their main component is tungsten carbide (WC). Carbides, nitrides, or carbonitrides of Ti, Nb, Ta, and Hf can also be present as mixed crystal formers. The hard material phases are bonded together by a ductile metallic phase that surrounds them (cemented carbides), usually Co, more rarely Ni or Fe alloys.

The outstanding properties of the hard metals are their high hardness and wear resistance, especially at high temperatures, combined with good toughness properties. This unusual combination of properties is achieved by combining the hard and brittle carbide phase(s) with the ductile and deformable binder. By varying the carbide/ binder ratio, and by suitable choice of the carbide composition, the properties can be varied within wide limits. A further control parameter for certain properties is the microstructure, i.e., the grain size of the carbide phase(s), which can be controlled via the particle size of the powder used, the powder milling, and the sintering conditions.

The most important groups of applications of hard metals are:

1) Metal cutting tools (drilling, turning, milling)
2) Tools for processing wood and plastics
3) Drilling tools in mining and mineral oil and water drilling technology
4) Wear-resistant components in a wide range of machinery (a continuously increasing group with the widest diversification)
5) Elastically bonded abrasive materials

Steel. Tungsten is added to a very wide range of steels in a concentration of 1 – 18 %. It increases the hardness and wear resistance at high temperatures (carbide formation), resistance to thermal shock, and high-temperature properties. The main areas of use are in high-speed steels, high-temperature steels, and tool steels. The use of tungsten in steel is significantly greater in the People's Republic of China and the former Eastern Bloc countries, as high-speed steel is still used for many applications rather than hard metals, as in the West. In the United States, tungsten has mainly been replaced by molybdenum in high-speed steels.

Tungsten and Tungsten Alloys. This group represents the widest range of uses. It extends from everyday uses, e.g., the coil of an incandescent lamp or the contact tip of an automobile horn, to components of nuclear fusion reactors or ion drive motors in space probes. The reason for this range of uses lies in the many outstanding properties of tungsten:

High melting point
Low vapor pressure
High atomic number
Good electrical and thermal conductivity
High density and modulus of elasticity
Wide radiation band in visible light and good light yield
Good X-ray yield
Expansion coefficient comparable to those of glass and silica

The following list shows this versatility very clearly:

Lighting Technology
Incandescent lamps (NS-W)
Halogen lamps (NS-W)
Gas discharge lamps: Hg, Na vapor lamps (W)
Fluorescent tubes (W)
Xe short arc lamps (W, W – ThO_2, W – Re)

Electrical and Electronic Technology
Sputter targets in VLSI technology (W – Ti, high-purity W)
Transistors (W)
Diodes (W)
Electronic tubes (W, W – ThO_2, porous W with Ba or ThO_2)
Thyristors (W)
Switch contacts [W, W – Cu, W – Ag, W – Fe – Ni – (Cu)]
Heat sinks (W – Cu)

High-Temperature Technology (furnace construction, nuclear energy, thermal power stations)
Structural components (W)

Walls of fusion reactors (W, W – Re – ThO_2, W – Re – HfC)
Construction components in the plasma space in magnetohydrodynamic electricity production (W, W – Cu)
Thermocouple elements (W/WRe, W/Mo, W/Ir, W/graphite)
Heating elements (W)

Vacuum and Plasma Metallization, Welding, Spark Erosion
Nozzles (W – ThO_2)
Electrodes [W – ThO_2, W – Ag, W – Fe – Ni – (Mo)]
Welding electrodes (W – ThO_2, W – LaO_2, W – CeO_2, W – ZrO_2, W – Y_2O_3)

X-Ray and Radiation Technology, Medical Technology
Anodes (W)
Rotating anodes (W – Re)
Containers for radioactive materials (W – Fe – Ni)
Components for radiation shielding, e.g., collimators in computer tomography scanners [W – Fe – Ni – (Cu)]

Machinery and Motor Construction
Governor balance weights, counterbalance weights, flywheel weights and other weights, e.g., in Formula 1 racing cars [W – Fe – Ni – (Cu)]

Chemical Industry
Electrodes, nozzles, crucibles (W)
Construction materials (W – Mo)

Space Travel
Rocket nozzles (W, W – Ag)
Space nuclear reactors (W – Re)
Thermionic converters (porous W, CVD-W, W – Re)
Ionic reactive thrust motors (porous W)
Structural components (tungsten fiber-reinforced niobium matrix composites, superalloys, Al, Ti)

Armaments
Armor piercing shells, armor plating, scatter grenades, shaped charge liners, couterweights in tanks, nozzles for air-to-air rockets, gas rudders (W – Fe – Ni, W – Cu, W – ThO_2)

Aircraft
Aircraft nose counterweights (W – Fe – Ni)
Turbine blades of fiber-reinforced nickel-based superalloys (W – Re – ThO_2 and W – Re – HfC fibers)

Laser Technology
Cathodes (porous W/barium aluminate)
Components of gas lasers (W – Cu)

7. Analysis

7.1. Raw Materials

The determination of the tungsten content of ore concentrates and scrap is very important, as the price is calculated from this. Sampling procedures are also extremely important, as the high density of tungsten can easily lead to segregation [49]. Determination of tungsten itself is almost exclusively carried out by X-ray fluorescence analysis [50]. Thus, matrix effects are compensated for by dissolution and dilution, and by the addition of an internal standard. Molten lithium tetraborate is widely used as a solvent, with tantalum pentoxide as an internal standard. Matrix corrections are carried out with computer programs.

7.2. High-Purity Intermediate Products, Tungsten Powder, and Compact Tungsten Metal

In these materials, it is not sufficient to determine the tungsten. In any case, especially in the metallic state (99.97 – 99.98 % W) this would be far too inaccurate. Here, it is important to measure trace impurity contents. In all these products, the specifications applicable are very rigorous, setting maximum acceptable upper limits for concentrations of individual elements. For industrial tungsten and its high-purity precursor materials, these are all in the μg/g range, the maximum tolerable level for most elements being 10 μg/g. Exceptions include Mo (50 μg/g), Si (20 μg/g), and H (1 μg/g).

Routine determination of trace elements is mainly carried out by the following methods:

Flame atomic absorption spectrometry
Graphite tube atomic absorption spectrometry

Inductively coupled plasma optical spectrometry
Photometry
Vacuum fusion hot extraction (H, N)
Carrier gas hot extraction (O)
Ignition in oxygen combined with IR spectrometry (C)

Up-to-date literature on this topic is rare, as improved and refined analytical methods are increasingly regarded as industrial secrets. The determination of metallic trace impurities is described in [51], [52], and of nonmetallic trace impurities in [53–55].

Large samples are always used. The result therefore always represents a mean value. Also, these analyses give no information on the distribution of a trace element and its state of bonding in the compact metal [56], [57]. It is well known that impurities can exist in various forms and states of combination in metals (dissolved interstitially or as mixed crystals, as precipitates in the crystal or at grain boundaries, etc.). This knowledge is often important for subsequent treatment processes or for the final properties. Topochemical analytical methods are also used:

Scanning electron microscopy coupled with energy or wavelength dispersive X-ray spectrometry
Auger electron spectroscopy
Auger microprobes
Ion beam microprobes
UV and X-ray photoelectron spectrometry, etc.

These techniques are not used routinely, mainly because of their cost, but for special applications and in research.

Heterogeneous impurities represent a special case (though often widely occurring) of particular importance in powder metallurgy. These impurities are small foreign particles present in a certain statistical distribution alongside the matrix particles. In most cases, they are due to mechanical abrasion or corrosion of production equipment. They can become concentrated in furnaces by vaporization, and can then end up in the material. Organic materials can also be present in the powders, e.g., fibers from filter cloths, packing materials, or plastic films. The various types of heterogeneous impurities, their effects, and their analytical determination are described in [58–61].

The presence of this type of heterogeneous impurity can lead to very serious material defects on powder metallurgical processing. Unlike molten metallurgical processes, where small foreign particles can dissolve in the homogeneous melt or slag, and hence cause very little change to the composition, the sintering of a compact can still leave locally high concentrations of a foreign element. Consequently, many kinds of defects can occur, e.g., empty or filled pores, porous zones, zones with completely changed material properties, etc., leading to fractures and cracks on subsequent processing. Fine wires and thin sheets are therefore extremely sensitive to heterogeneous impurities.

In the characterization of tungsten powder, physical analysis is important as well as chemical analysis [62]. Typical powder characteristics include:

Mean grain size (Fisher sub-sieve sizer, ASTM No. 430-1970)
Morphology, heterogeneity (scanning electron microscopy, SEM)
Particle size distribution (sedimentation, laser scattering, ultracentrifuge)
Bulk, tapped, and pressed density
Specific surface area (BET method)
Degree of agglomeration

7.3. Trace Elements in High-Purity Tungsten Metal

In high-purity tungsten metal, levels of trace elements are in the ng/g range. There are therefore special analytical requirements:

Working in a clean-room laboratory
Prevention of any other contamination
Correct calibration, etc.

Comparisons between various methods and laboratories show clearly that in trace analysis of this kind the results can have considerable scatter [63], [64]. The analytical techniques are always very complex and expensive, the most important being:

Glow discharge mass spectrometry
Secondary ion mass spectrometry
Isotope dilution mass spectrometry
Trace matrix separation combined with inductively coupled plasma mass spectrometry

Very significant improvements in this type of analysis have been achieved [38].

8. Ferrotungsten and Related Alloys

8.1. Composition

Ferrotungsten. Commercial ferrotungsten is a tungsten – iron alloy containing at least 75 wt % W, and having a very fine-grained structure and a steel-gray appearance. It is supplied in 80 – 100 mm lumps in accordance with DIN/ISO standard specifications. Special sizes can be produced by further size reduction to suit customers' requirements. The standardized compositions are listed in Table 5 [65].

The binary phase diagram for Fe – W is shown in Figure 9. The phases Fe_7W_6 (μ) and FeW (δ) exist close to the edge phases. However, ferrotungsten occurs in the equilibrium state only after prolonged high-temperature treatment. Therefore, in commercial ferrotungsten, the phase Fe_2W (λ) should still be present alongside the δ-phase. The melting temperature of ferrotungsten containing ca. 80 % W is very high, the liquidus temperature being >2500 °C and the solidus temperature ca. 1640 °C. The density of ferrotungsten is 15.4 g/cm^3.

Tungsten Melting Base (TMB). The term "tungsten melting base" refers to alloys produced from secondary raw materials. These can have various compositions, depending on the type of recycled material used. Typical analyses are listed in Table 6. The alloys are supplied as granulated material with grain size <10 mm.

8.2. Uses

Ferrotungsten and Tungsten Melting Base. Ferrotungsten and tungsten melting base are mainly used as alloying materials in the steel industry as they dissolve more readily than pure tungsten in molten steel, owing to their lower melting points, and are also cheaper.

Tungsten additions increase the hardness, yield strength, and ultimate tensile strength of the steel without reducing the elongation, area reduction on breaking, and notched bar fracture toughness [70].

In steel, tungsten forms stable carbides that increase hot strength and wear resistance. Tungsten-containing alloys are therefore mainly used in the production of tool steels, especially high-speed steels [71].

High-speed steels (HSS) represent the main application area for ferrotungsten. These steels are highly alloyed with tungsten, molybdenum, vanadium, and chromium. After the correct heat treatment, they have high hardness and wear resistance, and are mainly used for machining at high cutting rates and temperatures up to 600 °C. The alloying elements and the carbon form hard, wear-resistant carbides in a matrix with good hardness retention properties. The standard high-speed steels specified in DIN 17 350 contain 6 – 18.5 % tungsten [72].

Other Tungsten-Based Master Alloys. Other master alloys, used especially in the production of nickel-based superalloys, include Ni – W, complex nickel-based alloys containing Cr, Ta, and Mo [69], and cobalt – chromium – tungsten master alloys for wear-resistant materials and protective coatings [73] (→ Cobalt and Cobalt Compounds).

The master alloys for superalloys must be produced from very pure raw materials. They are melted in vacuum induction furnaces, and are cast and cooled under vacuum. The delivered form is small lumps (<50 mm) or powder (<0.3 mm). Some typical compositions are listed in Table 6.

8.3. Production

8.3.1. Ferrotungsten

The raw materials used for ferrotungsten production are rich ores or concentrates containing the minerals wolframite, hübnerite, ferberite, and scheelite. "Synthetic scheelite", a precipitated calcium tungstate, is also used.

These raw materials contain tungsten in the form of WO_3, which can be relatively easily reduced either with carbon or silicon and/or aluminum. Reduction with magnesium and calcium is also possible, but is of no industrial importance for cost reasons. Figure 10 shows the effect of temperature on the oxygen potentials of metal oxides. Although tungsten forms carbides with carbon, the formation of tungsten carbide in the finished product can be virtually prevented by suitably controlling the CO : CO_2 ratio, so that a low-carbon ferrotungsten with a maximum carbon content of 1 % is produced. Tungsten oxides

Table 5. Specifications of the composition of ferrotungsten ∗, DIN 17 562 and ISO 5450 [65]

	% W		% C max.	% Al max.	% Sn max.	% As max.	% Sb max.		% Mo max.	
	DIN	ISO	DIN/ISO	DIN/ISO	DIN/ISO	DIN/ISO	DIN	ISO	DIN	ISO
FeW 80	75 – 80		1.0		0.10	0.10	0.08			
FeW 80		70 – 85	1.0	0.10	0.10	0.10		0.05		1.0
FeW 80 LC		70 – 85	0.1	0.10	0.10	0.10		0.05		0.5

∗ Impurities: 0.6 – 1.0 % Si, 0.6 – 1.0 % Mn, 0.20 – 0.25 % Cu, 0.05 – 0.06 % S, 0.05 – 0.06 % P.

Table 6. Composition (wt %) of tungsten melting base and tungsten master alloys for superalloys

	W	Mo	Cr	V	Co	Ni	S	P	C
TMB [67]	28 – 38	0.5	0.2	0.2	max. 0.6	0.4	0.04	0.04	1.2
TMB containing Co [67]	24 – 30	0.6	0.5	0.2	5 – 10	0.4	0.04	0.04	1.4
FeWMo [67]	7	5.5	0.5	2	0.6	0.4	0.04	0.04	1.2
FeWMoCo [67]	7	5.5	0.5	2	5 – 10	0.4	0.04	0.04	1.2
Darby melt [67]	10	0.4 – 1.5	3	1.2			0.03	0.02	0.4 – 1.3
NiW [69]	45					55	0.001	0.002	0.002
NiCrAl 12 MoW [69]	7	4	40		1	36	0.01	0.015	0.1
NiCrAl 8 WTi 4.4 Ta 3.5 Mo [69]	7	3.5	26		1	47	0.01		0.04

Figure 9. Binary phase diagram of the Fe – W system [66]

Figure 10. Effect of temperature on the oxygen potentials of oxides [74]

are reduced in preference to iron oxides, so that ferrotungsten with a high tungsten content can be produced, even if the ore has a high FeO content.

Thus, ferrotungsten can be produced by carbothermic reduction in an electric arc furnace or by metallothermic reduction with silicon and/or aluminum.

The carbothermic or silicocarbothermic method is preferred for cost reasons. Moreover, higher impurity contents (tin and arsenic) can be tolerated in the raw materials.

When high levels of arsenic and tin are present in the concentrates, these can be vaporized by roasting in rotary tube furnaces.

8.3.1.1. Carbothermic Production

Because of the high melting point of ferrotungsten, the so-called solid block melting process is normally used, as tapping off is not possible at the furnace temperatures that are required.

In this process, the ferrotungsten accumulates in the hearth of the furnace vessel, which is constructed in sections. After the desired weight has been produced, the furnace lining is removed

and the metal ingot can be removed after cooling. The solid block is cleaned, crushed, and sorted. The cleaned-off material and skin of the ingot are returned to the furnace for remelting.

The use of several melting vessels enables the process to be semicontinuous.

The process is usually operated in two stages, as an excess of tungsten ore is used in the ingot melting process (refining stage), producing a WO_3-rich slag that must be processed in a second (reduction) stage to give ferrotungsten with a low W content and a low-WO_3 slag.

The low-W ferrotungsten is reused in the refining stage. The slag, which contains $<1\,\%$ WO_3, can be dumped. The flow diagram of this two-stage process is shown in Figure 11 [75].

In the three-phase electric arc furnaces usually used today, the pitch circle diameter of the electrodes is small to ensure adequate energy concentration. Typical furnace data for a three-phase electric arc furnace for the production of FeW 80 are [76]:

Power	2000 – 3000 kV · A
Refractory lining	monolithic rammed carbon
Electrode material	
refining stage	graphite
reduction stage	Söderberg
Energy consumption per tonne FeW	7500 – 8500 kW · h
FeW ingot weight	11 – 12 t

The electric arc furnaces are provided with waste-gas purification plants both for environmental reasons and to realize the value of the WO_3 in the flue dust. This dust, enriched in arsenic and tin vaporized from the concentrates, is usually added to the reduction stage or processed in special batches. The resulting secondary dust must either be chemically processed to obtain tungsten compounds or disposed of as a special waste if it cannot be sold.

Tungsten yields of 97 – 98 % can be obtained in the solid block process.

8.3.1.2. Carbothermic and Silicothermic Production

In 1937, V. N. Gusarov developed a continuous process for the production of ferrotungsten containing ca. 75 % W. This was used in the former Soviet Union and the former Czechoslovakia [77], [78]. It is presumably used in the Czech Republic.

This metallurgically interesting process is carried out in three successive stages in a three-phase electric arc furnace lined with magnesite. Of the oxygen in the WO_3, 60 % reacts with carbon and 40 % with silicon (as ferrosilicon containing 75 % Si). In the first stage, ferrotungsten (75 % W) is produced by substoichiometric carbon reduction. This ferrotungsten is formed in a pasty consistency under a WO_3-rich slag, and can be scooped out with ladles. Details of the individual processes are as follows:

Stage 1. Reduction of tungsten ore concentrates by carbon under a WO_3-rich slag (10 – 16 % WO_3, 26 – 28 % FeO, 15 – 17 % MnO), and scooping out the ferrotungsten (e.g., 73.3 % W, 0.4 % Mn, 0.43 % Si, 0.05 % P, 0.06 % S, 25.2 % Fe).

Stage 2. Reduction of the WO_3-rich slag with FeSi 75 and addition of scrap iron (to reduce the tungsten content of the ferrotungsten produced). Tapping off the low-WO_3 slag (e.g., 0.3 % WO_3, 20.8 % MnO, 51 % SiO_2, 2.7 % FeO, 3.6 % Al_2O_3, 21.1 % CaO). The metal, with a reduced tungsten content (54 – 70 % W, 3 – 8 % Si, 1 – 2 % Mn) remains in the furnace.

Stage 3. Refining of the low-tungsten metal by adding tungsten concentrates. A WO_3-rich refining slag (18 – 25 % WO_3) is formed, and the tungsten content of the metal increases.

The advantage of the process is the continuous method of operation; thus the furnace can be operated for considerably longer periods without interruption for relining than in the solid block process. A further advantage is the lower energy consumption, i.e., ca. 4000 kW · h per tonne of ferrotungsten. The tungsten yield is 98 %, approximately the same as for the solid block process.

Disadvantages include the heavy labor of the scooping operation and the accumulation of some of the ferrotungsten in a “furnace sow” or “salamander”. This ferrotungsten can be retrieved only after furnace relining, and is recovered as product after remelting.

Figure 11. Carbothermic production of ferrotungsten [75]

8.3.1.3. Metallothermic Production

Tungsten oxide can be reduced by silicon and/or aluminum (see Fig. 10). Compared with carbothermic reduction, metallothermic production of ferrotungsten requires purer raw materials, as the reactions proceed very rapidly, and the impurities are chemically reduced as well as the raw materials. In this process, about 50 % of the tin and arsenic present, most of which would be vaporized during the longer carbothermic process, end up in the ferrotungsten.

The tungsten concentrates in finely divided form are mixed with coarsely powdered aluminum and silicon. Pure silicon or ferrosilicon cannot be used, as these would not give a selfsustaining reaction, the heat evolved being insufficient to melt the ferrotungsten and slag formed. Aluminum and silicon in the ratio 70 : 30 are therefore used.

The reaction mixture is charged into a refractory-lined furnace vessel and preheated to 400 –500 °C. The preheating can be omitted if Fe_2O_3 and Al are mixed into the reaction mixture as a booster.

The reaction is started at the top by igniting initiators, which are mixtures of, e.g., BaO_2 and aluminum powder. A purely aluminothermic mixture burns completely in a few minutes, but silicothermic – aluminothermic mixtures react more slowly. After cooling, the furnace vessel is removed, and the blocks of metal and slag are separated. This method produces ferrotungsten ingots of 700 – 2000 kg. The metal ingot is cleaned, crushed, and sorted. Pieces with adhering slag are sent back for remelting.

The metallothermic production process has lost much of its importance owing to the high costs of the aluminum and silicon reducing agents, and the necessity for using pure and therefore expensive raw materials.

The process is suitable only for special customer requirements, and for operations which also produce other alloys by the aluminothermic principle.

Advantages of the process include the simple, low-cost plant, and the minimal tying-up of materials resulting from the short processing time. The tungsten yield is ca. 96 %.

M. Riss and Y. Khodorovsky have described the electroaluminothermic production of ferrotungsten from scheelite [78]. The reaction time is extended by melting in an electric arc furnace, and separation of the metal from the slag is improved.

8.3.2. Production of Tungsten Melting Base and Tungsten Master Alloys

During World War II, melting base alloys for the production of high-speed steels were developed and marketed by the London and Scandinavian Metallurgical Company (LSM). They were produced from tungsten hammer scale, grinding wheel swarf, and other waste products by reductive smelting in electric arc steel furnaces.

A similar recycling process is today operated by the Treibacher Chemische Werke. Here, the tungsten-containing waste is pretreated (degreased, etc.), melted in an electric arc furnace (7.2 MV · A) to produce tungsten melting base (TMB), and is then granulated. At this plant, which is also used for other ferroalloys, the TMB production rate is ca. 1000 t/a in the late 1980s. The product specifications are listed in Table 6 [67], [79].

J. Becvar has described a process operated in the Czech Republic [68]. Tungsten-containing wheel swarf is first degreased and dewatered, and is then melted with other scrap in an electric arc furnace to give a so-called Darby melt, which is used as a master alloy for high-speed steel. A typical composition is given in Table 6. The

process appears favorable, as other alloying elements in the scrap, e.g., chromium and vanadium, are also recovered.

Other master alloys used for superalloys also require very pure raw materials. The alloys are melted in vacuum induction furnaces, tapped off, and cooled under inert gas (argon). Some examples of the alloys produced by the Gesellschaft für Elektrometallurgie mbH (GfE) in Nuremberg are listed in Table 6. These are supplied either as small lumps or powder [72].

9. Compounds

9.1. Tungsten Chemistry

Tungsten occurs in all oxidation states from -2 to +6 inclusive, the most important being +6. Valencies between -2 and +1 exist only in organometallic compounds and compounds with π-acceptor ligands. In the lower valencies, tungsten forms bases (sometimes unstable), but in the higher valencies it forms very stable acids.

Tungsten combines with most elements and with a large number of inorganic and organic ligands. The highest coordination number is 8, and the most important donor atoms are O, N, S, and the halogens. Tungsten has a very marked tendency to form polymeric compounds.

9.2. Aqueous Solutions of Tungsten

This chemistry is extremely complex [80], [81], as the tungstate anion exists in monomeric form only in strongly alkaline solutions. In mildly alkaline solution, the tungstate anions begin to polymerize, and this progresses with decreasing pH. Below pH 1, sparingly soluble tungsten oxide hydrate (tungstic acid) precipitates. Various degrees of polymerization (isopolytungstates, metatungstates, paratungstates, etc.) are favored as the pH changes [82]. Other parameters influencing the polymerization include concentration, temperature, and time. Time plays an important role in reactions that proceed slowly. Thus, the properties of mildly acid or almost neutral solutions of tungsten change markedly after long periods of storage (a process often referred to as aging). Foreign ions also influence various polymerization reactions. Ammonium ions favor polymerization, even at higher pH values. This is why $(NH_4)_2WO_4$ cannot be crystallized from aqueous solution (it always forms APT), but only from tungstic acid and liquid ammonia. Anions such as phosphate, silicate, arsenate, etc., copolymerize with tungstates to form heteropolytungstates which usually have a higher degree of polymerization than the isopolytungstates stable at the same pH. Several isopolytungstates group around the central heteropolyanion.

Fluoride, sulfide, and peroxo anions have a depolymerizing effect, as these replace O atoms on the W [80]. The substitution of several O atoms by organic compounds with two or more donor atoms leads to the formation of numerous chelate complexes, and hence to depolymerization [80].

9.3. Intermetallic Compounds

There are a number of compounds of tungsten with the elements Al, Be, Co, Ni, Fe, Zr, Hf, Ir, Os, Re, Ru, and Tc which can play a positive or negative role in the metallurgy of tungsten alloys. A comprehensive review of the literature on this subject can be found in [4].

9.4. Compounds with Nonmetals

9.4.1. Tungsten – Boron Compounds [4], [83]

The following compounds are known: W_2B, WB, W_2B_5, and WB_4. They are prepared by sintering the elements, from the oxides in an aluminothermic reaction, or by deposition from the gas phase. They are hard, brittle, crystalline, and electrically conducting, and are of no industrial importance.

9.4.2. Tungsten – Carbon Compounds [4]

There are three carbides (Table 7): W_2C, WC, and the cubic high-temperature phase α-WC_{1-x}.

Table 7. Physical properties of tungsten carbides

	mp, °C	Crystal system	Density, g/cm^3
W_2C	2770	hexagonal	17.2
WC	>2600 (decomp.)	hexagonal	15.6
α-WC_{1-x}	ca. 2750	cubic	

Production. The carbides are formed from tungsten or its compounds by reaction with carbon or carbon compounds. The most important industrial process is the reaction of tungsten metal powder with carbon black or graphite at 900 – 2200 °C. The reaction of WO_3 with carbon is of less importance, as it is difficult to control the carbon balance.

Industrial Importance. WC is quantitatively the most important tungsten compound because of its hardness. It is also used as a reforming catalyst.

W_2C, like WC, is present in the carbide melt used for "hard facing".

Mixed-crystal carbides such as WC – TiC (50 : 50) and WC – TiC – Ta(Nb)C (very variable composition) are widely used in hard metal production.

The η-carbides, of composition Co_6W_6C and Co_2W_4C to Co_3W_3C, and analogous compounds with Fe occur as intermediates in tungsten-containing steels and in special grades of hard metal.

Tungsten Hexacarbonyl $W(CO)_6$ [84]. This can be produced from W + CO or from WCl_6 + CO + Al. It is a white, crystalline substance that decomposes at 150 °C and is soluble in organic solvents, e.g., CCl_4. It is used as a catalyst in organic syntheses.

9.4.3. Tungsten – Silicon Compounds [4], [83–85]

The following compounds are known: WSi_2, W_5Si, W_3Si_2, W_2Si_3, and W_5Si_2. They are produced by heating the elements together. They are hard and stable to 900 °C. Tungsten silicide sputter targets are used in microelectronics to produce WSi_x films because of their exceptional characteristics and their thermal and chemical stability.

9.4.4. Tungsten – Group 15 Compounds

Tungsten – Nitrogen Compounds [84]. The following compounds are known: W_2N and WN. They are not formed by reaction of the elements, but by heating tungsten in ammonia gas. They are of no industrial importance.

Tungsten – Phosphorus Compounds [4], [85]. The following compounds are known: WP_2, WP, W_2P. These can be synthesized from the elements. They are of no industrial importance.

Tungsten – Arsenic Compounds [4]. The following compounds are known: WAs_2, W_2As_3. Synthesis is from the elements. They have no industrial importance.

9.4.5. Tungsten – Oxygen Compounds

Oxides [84], [86]. In the W – O system, there are not only the stoichiometric oxides WO_3, $WO_{2.9}$, $WO_{2.72}$, WO_2, but also nonstoichiometric structures that represent the ordered or partially ordered defect structures of the oxygen-rich oxides, in which the central W atom is octahedrally surrounded by six oxygen atoms. In WO_3, neighboring octahedra are in contact only at the corners. With increasing oxygen deficiency (reduction, conversion to lower oxides), common edges and surfaces are progressively formed.

Tungsten trioxide, WO_3, is formed on strongly heating tungstic acid or APT in air. It is yellow, but even a small oxygen deficitency leads to greenish colorations. There are several crystallographic modifications which differ structurally from each other (variable distortion of the WO_6 octahedra). Commercial WO_3 (density 7.29 g/cm^3) is monoclinic at room temperature. Heating to just below 700 °C leads to orthorhombic WO_3, and above this temperature

to the tetragonal form. It begins to vaporize appreciably below its melting point (1473 °C). It is much more volatile in the presence of water vapor. WO_3 is soluble in alkaline solutions and melts.

$WO_{2.9}$ ($W_{20}O_{58}$) is bluish-violet, and occurs as an intermediate product during the reduction of WO_3 with hydrogen. It can be prepared by the reduction of WO_3 by moist hydrogen (e.g., 500 – 550 °C, water vapor pressure 60 kPa). It is a component of commercial blue oxide.

$WO_{2.72}$ ($W_{18}O_{49}$) is reddish-violet, and above 570 °C occurs as a characteristic needle-shaped intermediate product of WO_3 reduction. It is prepared in the pure state by reducing WO_3 with moist hydrogen (e.g., 800 °C, water vapor pressure 80 kPa). It can be present at low concentrations in industrial blue oxide.

Tungsten dioxide, WO_2, is chocolate brown in color, and is also an intermediate product of the reduction of tungsten compounds. It forms monoclinic crystals with a distorted rutile structure, and is therefore structurally very different from the higher oxides. The pure material is produced by reduction of WO_3 with moist hydrogen (e.g., 900 °C, water vapor pressure 50 kPa).

β-Tungsten, W_3O, [87], [88], is a gray or black substance produced at low reduction temperatures, usually during the reduction of doped tungsten oxides. It has a cubic A_{15} structure, and was first described as an allotropic modification of the metal, later as the suboxide W_3O. It is now thought that it is a metastable metallic phase stabilized by oxygen. It is stabilized by a series of other foreign elements, e.g., K, Be, P, B, As, and Al. It occurs as a characteristic intermediate product in the reduction of tungsten oxides doped with K, Al, and Si. Pure β-tungsten (by X-ray analysis) can be prepared by reduction of a thin layer of $W_{20}O_{58}$ powder in dry hydrogen at 500 – 600 °C.

Oxide Hydrates. $WO_3 \cdot H_2O$ or $WO_2(OH)_2$ [89] occurs as a monomeric gaseous compound during the industrial reduction of tungsten oxides with hydrogen at >600 °C. It is formed by the reaction of the oxides with the water vapor formed during the reduction, and is the substance on which the chemical vapor transport of tungsten depends.

Tungstic Acids [4], [84], [90]. H_2WO_4 is formed when a hot alkali metal tungstate solution is acidified with a strong acid. It is a yellow powder, very sparingly soluble in water and acids, but very soluble in alkaline media.

White tungstic acid is obtained by precipitating in the cold and reversing the order of addition. It corresponds to the composition $WO_3 \cdot x\,H_2O$, where $x \approx 2$.

If tungstic acids are dried at <300 °C, the orange-yellow hemihydrate $2\,WO_3 \cdot H_2O$ is obtained.

Tungstic acid was formerly the most important high-purity intermediate product used in tungsten production, but has now been replaced by APT.

Tungstic acid in a highly pure form is today produced from APT by treatment with strong acids, and is used in the production of very fine tungsten powder and various tungsten-based chemicals.

Tungstates [4], [84], [91]. These are formed by most metal cations (e.g., Na_2WO_4, $NiWO_4$, Cr_2WO_6, ZrW_2O_8). The structure of the tungstates is mainly determined by the size of the metal cation. Tungstates with large bivalent cations ($\gg$0.1 nm) have the scheelite structure, and those with smaller ions (<0.1 nm) the wolframite structure. Apart from the alkali metal and magnesium tungstates, these are all sparingly soluble in water. All the most important ores are tungstates.

Sodium tungstate crystallizes with two molecules of water of crystallization. Aqueous solutions are used as heavy liquids, in fuel cells, for flame-resistant and antistatic textiles, in cigarette filters, and as corrosion inhibitors for steel. Anhydrous sodium tungstate is obtained by melting WO_3 with sodium hydroxide or carbonate.

Calcium tungstate fluoresces under ultraviolet light, a phenomenon which is utilized when prospecting for scheelite deposits. Calcium tungstate is used in the production of phosphors for lasers, fluorescent lamps, oscilloscopes, luminescent dyes, scintillation counters, X-ray screens, etc.

A wide range of tungstates are used in the ceramic industry, in catalysts, and in pigments.

Polytungstates. The tendency of tungstate ions to polymerize in aqueous solution has already been referred to. The degree of polymerization increases with decreasing pH until tungstic acid is precipitated at pH 1. Other anions, e.g., arsenate, phosphate, silicate, can be incorporated into these highly polymeric ions. According to whether the polyanions are composed only of tungstate or also include foreign ions, they are referred to as isopolytungstates or heteropolytungstates.

Isopolytungstates can have various preferred degrees of polymerization, as mentioned above [82]. The most industrially important compound in this class is APT, the most widely used high-purity raw material in the tungsten industry.

Ammonium metatungstate is produced either with the aid of ion exchangers or by electrodialysis, and, owing to its high solubility in water, is used for impregnating carriers (e.g., aluminum oxide) in the production of catalysts.

Sodium metatungstate is used as a heavy liquid when drilling for oil.

Amine Tungstates [82], [92], [93]. Organic amines form water-insoluble ion associate complexes with isopolytungstates. Complex long-chain amines ($C_8 - C_{10}$) are used as reagents in the industrial solvent extraction of tungsten. Complexes with short-chain amines (ethylenediamine, propylamine, etc.) show autoreductive and autocarburizing properties on heating. This is of interest for the production of finely powdered W or WC, which can be obtained from the precursor in a single-step process.

Heteropolytungstates. There are over 30 elements known to form the central anion in these compounds. Typical properties independent of the central heteroatom include high relative molecular mass (>3000), high degree of hydration, and high solubility in water and organic solvents. A summary of possible structures and compositions can be found in [84].

One of the most industrially important compounds in this group is $Na_3PW_{12}O_{40} \cdot x\,H_2O$ [94], [95]. This has many uses: for fixing and oxidizing in photographic processes; as an additive in electroplating; for waterproofing plastics, adhesives, and cement; in the production of organic pigments; for the surface treatment of skins; as an antistatic agent for textiles; for printing and paper dyes; for coatings; wax pigments; steel passivation; and as a catalyst in the organic chemical industry.

Tungsten Bronzes [4], [84], [91]. These are well-defined though nonstoichiometric compounds with the general formula M_xWO_3, in which x is between 0 and 1; M represents a metal cation, H^+, or NH_4^+. The name is derived from the metallic luster and intense colors (golden yellow to blue-black) of these compounds. The bronze character is determined by the amount of metal in the structure. With high values of x, interaction between the metal ions leads to the formation of metallic bonds. The stability of the bronzes is determined by the size of the metal ions. Stability decreases as the ionic radius decreases. Small metal ions, e.g., Si and Ge, do not form bronzes. Owing to the oxygen deficit, these compounds are electrical semiconductors or conductors. The sodium bronzes for which the Na/WO_3 ratio is >0.3 have a positive temperature – resistance coefficient, and vice versa.

The bronzes are produced:

1) By partial reduction of the corresponding tungstates
2) By electrolytic reduction
3) By the reaction in the fused or solid state of $WO_3 - WO_2$ mixtures with metal compounds

Tungsten bronzes are used as colored pigments, catalysts, e.g., for the oxidation of CO, and in fuel cells. Electrodes made of La_xWO_3 are used in sensors for the potentiometric determination of dissolved oxygen.

Hydrogen and alkali metal bronzes are used as electrochromic coatings for so-called dynamic glasses. Application of an electrical potential causes migration of the ions (H^+, Na^+) into or out of the coating, thereby changing the properties of the glass from transparent (oxidic) to reflecting (metallic due to bronze formation). The only practical application of this so far is in electrochromic sunglasses.

9.4.6. Tungsten – Chalcogenide Compounds

Tungsten – Sulfur Compounds [4], [84], [96], [97]. The following compounds are known: WS_2, WS_3. Tungsten disulfide is of industrial

importance. It is black, forms hexagonal or rhombic crystals, and can be synthesized from the elements. It is used as a catalyst for the desulfurization of mineral oil, and as an oil-soluble lubricant under extreme conditions (pressure, temperature, vacuum).

Tungsten – Selenium Compounds [4], [97]. The following compounds are known: WSe_2, WSe_3. The diselenide can be synthesized from the elements, and is used as an oil-soluble lubricant.

Tungsten – Tellurium Compounds [4]. WTe_2 is known, but is of no industrial importance.

9.4.7. Tungsten – Halogenide Compounds

Tungsten – Fluorine Compounds [4], [84]. The following compounds are known: WF_6, WF_5, WF_4, WOF_4, WO_2F_2, WOF_2. Only the hexafluoride is of industrial importance, being used to produce CVD tungsten. It can be produced by the reaction of tungsten hexachloride with hydrogen fluoride, or by direct combination of the elements at 350 – 400 °C. The boiling point is 17.5 °C. The colorless gas is soluble in organic solvents, and is sensitive to moisture (hydrolysis).

Tungsten – Chlorine Compounds [4], [84]. The following compounds are known: WCl_6, WCl_5, WCl_4, WCl_2, $WOCl_4$, WO_2Cl_2, $WOCl_3$, $WOCl_2$. Two of these are of industrial importance:

Tungsten hexachloride is a crystalline blue-black compound (*mp* 275 °C, *bp* 346 °C). It is obtained by direct combination of the elements at 600 °C. It is soluble in carbon disulfide and organic solvents, and is hydrolyzed by water. It has a wide range of uses in catalysis.

Tungsten oxytetrachloride forms red crystals (*mp* 211 °C, *bp* 327 °C), and is soluble in carbon disulfide and benzene. It is produced by the reaction of sulfuryl chloride with WO_3. It too has a wide range of uses in catalysis.

Tungsten – Bromine Compounds [4], [84]. The known compounds, which are of no industrial importance, are: WBr_6, WBr_5, WBr_4, WBr_3, $WOBr_4$, WO_2Br_2, $WOBr_2$, $WOBr_3$.

Tungsten – Iodine Compounds [4], [84]. The known compounds, which are of no industrial importance, are: WI, WI_3, WI_2, WO_2I_2, WOI_2, WO_2I, WOI_4.

10. Tungsten in Catalysis [69], [98–100]

Tungsten, both in metallic form and in a wide range of compounds, is used as a catalyst or catalyst component in a large number of chemical processes.

In the form of a metallic alloy with Ni, Co, or Rh, sometimes in the form of a sulfide, and sometimes on aluminum oxide carriers, it can catalyze the reaction of CO with H_2, and also hydrotreating, hydrocracking, reforming, and hydrodesulfurization reactions.

The following oxides are used as catalysts:

WO_3 is used in colloidal form for the photocatalytic reduction of organic compounds.

$W_{20}O_{58}$ catalyzes hydrogenation, dehydrogenation, hydroxylation, epoxidation, etc.

Mixed oxides of WO_3 with TiO_2 and/or V_2O_5, Al_2O_3, or SiO_2 have a wide range of uses, e.g., as DENOX catalysts for exhaust gas cleaning in thermal power stations, hydrogenation, aldol condensation, ring opening reactions, dimethyl sulfoxide synthesis, etc.

Of the halides, the compounds WCl_6 and $WOCl_4$ play a very important role in organic chemistry as catalyst components, e.g., in association with organometallic compounds such as $AlCl_2(C_2H_5)$, $Sn(C_6H_5)_4$, $Sn(CH_3)_4$, $Sn(C_4H_9)_4$, triaryl- or trialkyltin hydroxides, or the analogous lead compounds. Typical uses include: diene polymerizations, ring opening polymerizations, alkylations, etc.

WC is used for reforming *n*-heptane, as a redox catalyst for the oxidation of hydrogen in fuel cells, and for the formation of hydrogen from water.

$W(CO)_6$ catalyzes a wide range of organic reactions, e.g., the metathesis of alkenes.

Sodium tungsten bronze catalyzes the oxidation of CO.

The sodium salt of 12-tungstophosphoric acid is used as a catalyst for isomerization, polymerization, nitrile synthesis, dehydrochlorination, dehydrogenation, ketone syntheses, and ring closure syntheses.

WS_2 and WSe_2 in the form of colloidal coatings have been tested for photocatalysis.

11. Economic Aspects [101]

The price of tungsten in ore concentrate depends on supply and demand, and is therefore subject to fluctuations (Fig. 12). It is expressed either as \$/STU (short ton unit = 20 lb WO_3 = 7.19 kg W) or in \$/MTU (metric ton unit=10 kg WO_3 = 7.93 kg W). In times of warlike confrontation, the strategic importance and consequently increased demand for tungsten have led to high prices. On average, prices increased until 1977, although this was partly caused by inflation. The price maximum reached in 1977 was a consequence of increased consumption. Hand in hand with this development, the mining of tungsten shifted from the developing countries to the highly industrialized countries. This was mainly to secure supplies in times of crisis. In 1970, 49 % of tungsten came from western industrialized countries. In 1980, this figure was 58 %, and 69 % was predicted for 1990. However, this figure was never reached, as the market share of the People's Republic of China had continuously increased (Fig. 13).

In 1977, world tungsten demand was still <40 000 t. As shown in Figure 13, demand continued to increase until 1989, reaching ca. 52 000 t, but the price fell from 1977 onward. Thus, the price control mechanism that had been operating well until then was obviously no longer effective.

Several circumstances were responsible for this:

Up to 1977, China, for long the most important exporter of tungsten ore concentrates, had been price conscious in its trading, as this was only carried out by state officials. Following the liberalization of trading in China, producers attempted to obtain foreign currency directly by quoting low prices. Internal competition and pressure from dealers led to continued price reductions.

This fall in prices was exacerbated by the fact that high-purity intermediate products, e.g., APT, blue tungsten oxide, tungstic acid, and ferrotungsten, were offered at the same price per unit of tungsten as ore concentrates. At the end of 1993, the price of high-purity APT was so low that even the Chinese producers were no longer able to sell at a profit.

The drastic fall in demand that started in 1990 and which was caused by the general recession, the breakdown of the communist world, and the associated decline in the armaments industry led to stabilization of the lower prices.

Quite simply, the low tungsten price and the fact that there was no price difference between ore concentrate and intermediate product led to the enforced closure of almost all the larger mines in North America, Europe, South Korea, and Australia, and the suspension of many associated industries based on the treatment of ore concentrates.

In consequence, not only has the proportion of tungsten supplied from the People's Republic of China increased (Fig. 13), but also the makeup of the products has changed (Fig. 14). Whereas China was still an exporter exclusively of tungsten ore concentrate in the early 1980s, today it is the largest supplier of high-purity tungsten intermediate products.

Thus, none of the predictions, e.g., 4 % annual growth, opening of new mines, further shifting of mine production to the industrialized countries, self-sufficiency, etc., has been fulfilled. World demand for tungsten has decreased drastically; China is the main supplier.

12. Tungsten Recycling

As stated in Chapter 3, large amounts of tungsten scrap have been used as a raw material for several decades. This is because of:

The value of the tungsten content
The value of other metals in the alloys, e.g., Ta and Co in hard metals and Ni in heavy metals
Environmental aspects
Economic aspects

Methods for the treatment of scrap and utilization of the recovered materials are optimized where possible in the light of these factors.

Both direct and indirect recycling processes are used. In direct recycling, scrap is converted from the compact form to a powder which is then recycled only to the process that produced the material from which the scrap originated. For this, the scrap must consist of one grade only, and must not contain heterogeneous impurities.

Figure 12. Average mid-range price of wolframite 1955 – 1992
From *Metal Bulletin*

Figure 13. Tungsten supply and demand 1985 – 1992

Figure 14. Tungsten supply in China 1981 – 1992

In the indirect process, scrap is chemically converted to materials from which the impurities and other components are removed, and is then converted to APT as described in Chapter 4.

The direct method is preferable for economic and environmental reasons, but it can be used only for certain types of scrap, and the demand for the recycled product is not always certain. For this reason, the proportion of scrap tungsten treated by conversion to APT is continually increasing.

Combined processes are also used, e.g., the selective separation and recovery of one alloying component. Examples include the leach milling of hard metal scrap to liberate the cobalt binder metal, leaving behind the carbide phase(s), which can be recycled to the hard metal production process.

A proportion of tungsten scrap goes to smelting metallurgy for the production of stellites, melting base alloys, menstruum WC, and steel.

The economics of various treatment processes for the important types of scrap are compared in Table 8.

13. Beneficial Effects on Human Health [102], [103]

In contrast to certain toxic effects of metallic tungsten and its compounds and alloys, some very beneficial effects on human health have been described. An aqueous solution of sodium tungstate (ca. 2 %) was taken orally, applied to the skin, used as a gargle or an eye lotion, or injected subcutaneously. Experiments on 100 people in Japan showed remarkably successful results with a wide range of illnesses, e.g., cataract, stomach ulcers, duodenal ulcers, cancer, diabetes, herpes, migraine, gout, cardiac arrhythmia, and many others. Tungstate is also beneficial in cases of mercury or selenium poisoning, burns, and sunburn, and it can counteract the unpleasant aftereffects of excessive alcohol consumption, as is also reported for molybdate. It has furthermore been established that the natural aging process is significantly retarded. It is to be hoped that funded scientific investigations into the action of tungstate will be undertaken as soon as possible.

Table 8. Tungsten recycling ∗ [13], [14]

Treatment of scrap	Hard metal scrap	Heavy metal		W – Cu	W – AG	Pure tungsten
		Hard	Soft			
Zinc process	B					
Coldstream	B		G			L
Bloating/crushing	B					
Oxidation/reduction			B			L
Hydrometallurgy		L	L	B	B	L
Electrolytic		L	L	B	G	L
Chlorination	L	L	L	L	L	L
Melting (ferrotungsten, stellite)	G	G	G			G

∗ B best economic use; G good economic use; L least economic use.

14. Toxicology and Occupational Health

Tungsten and most of its compounds generally show relatively low toxicity compared to most other heavy metals and their compounds. However, some tungsten compounds are more toxic than they were originally considered to be.

Intoxications with tungsten and its compounds are comparatively rare; they occur almost exclusively by occupational exposure.

Physiological responses, toxicokinetics, toxic effects, levels of tolerance, and ecotoxicity of tungsten and its compounds have been reviewed [104–106].

Toxicokinetics. In animal models, nearly one-half of ingested soluble tungsten compounds and roughly one-third of deposited inhaled tungsten trioxide aerosol is readily absorbed. Most of the absorbed tungsten is rapidly excreted in the urine. The remainder is distributed among the red blood cells, spleen, kidney, and bone. About 3 months after ingestion and 6 months after inhalation, most of the complete tungsten body burden has been transferred into bone. However, this amount is very small as compared to the administered dose [105]. The biological half-life of tungsten in bone is calculated at 1100 d for the slowest component of a three-component elimination kinetics [107].

Tungsten seems to have an influence on molybdenum homoeostasis. In rats, the addition of 100 ppm tungsten to a diet containing 30 ppm molybdenum caused a decrease in the activities of sulfite oxidase and xanthine oxidase, which led to a significant increase in the susceptibility of these animals to sulfur dioxide [108].

Acute Toxicity. Acute intoxications are caused especially by tungsten hexafluoride, WF_6. Liquid or gaseous WF_6 reacts readily with water, generating highly toxic hydrogen fluoride. Consequently, the mucous membranes of the respiratory tract and the eyes are damaged severely after contact with WF_6. Acute intoxication causes a strong irritation of the upper respiratory tract, laryngitis, and bronchitis; after a period of less severe symptoms the development of cyanosis and lung edema is possible [109].

In contrast, 25 – 80 g of powdered tungsten metal did not cause adverse effects when administered orally as a substitute for barium sulfate in radiological examinations [110].

Chronic Toxicity. Cemented tungsten carbide, WC, is suspected of causing respiratory diseases in humans. However, these effects are possibly not due to WC itself, but rather to its content of cobalt metal. In workers processing WC the development of interstitial pulmonary fibrosis has been reported [111], [112]; the clinical picture includes cough, exertional dyspnea, and weight loss. Abnormal gas transfer was detected in respiratory function tests.

In the case of a 33-year-old woman who was exposed to cemented WC for several years and who died after development of pneumoconiosis, an increased content of cobalt ions in bones and lung was demonstrated after autopsy [113]. In various test series with rats there were strong indications that cobalt was the main reason for the development of the lung diseases [114].

In long-term studies with rats and mice, 5 ppm tungsten (as sodium tungstate) was added to the drinking water during the life span of the test animals. In rats, growth development increased slightly, whereas the life span decreased significantly [115]. Mice did not show these effects [116].

Genotoxicity. In a long-term study with mice, in which 5 ppm tungsten (as sodium tungstate) was added to the drinking water during their life span, no tumorigenic effect was seen [116].

Reproductive Toxicity. In rats, tungsten causes an increased embryo lethality and decreased ossification of bones in a dose range with no toxic effects to the maternal organism [117]. Tungsten accumulates in the fetus without being retained significantly by the placenta.

In another study on the embryonic and fetal development in mice [118], tungsten was injected as sodium tungstate before implantation or at early organogenesis (on day 3 or 8 of pregnancy, injection of 0.1 mL of 25 mmol/L sodium tungstate). The treatment did not show any effect on implantation, but increased the frequency of resorption.

Immunotoxicity. In a patch test, no allergic reactions to Na_2WO_4 were seen on the skin of 853 hard-metal workers [119]. However, irritant pustular reactions appeared in 2 % of the test persons.

Toxicological Data. Tungsten: LD_{50} (rat, i.p.) 5000 mg/kg [120]; sodium tungstate, Na_2WO_4 (rat, s.c.) 223 – 255 mg/kg [121]; phosphotungstic acid, $H_3P(W_3O_{10})_4$: LD_{50} (rat, oral) 3300 mg/kg [122]; tungsten carbide, WC (containing 8 % cobalt): LDLo (rat, intratracheal) 75 mg/kg [123]; tungsten carbide, WC (containing 15 % cobalt): LDLo (rat, intratracheal) 50 mg/kg [123].

In humans normal mean blood levels are considered to be 5.8 (± 3.5) µg W/L [104].

Treatment. Dimercaprol (British Anti-Lewisite) may be useful in the treatment of acute tungsten poisoning [124]. Cortisone preparations have been proposed for decreasing the development of lung fibrosis [125].

Occupational Health. The MAK value is 1 mg/m^3 for tungsten and soluble tungsten compounds, and 5 mg/m^3 for tungsten carbide and other insoluble tungsten compounds, measured as total dust [126]. In the United States, NIOSH has set exposure limits (TLV-TWA) of 1 mg/m^3 for soluble and 5 mg/m^3 for insoluble tungsten compounds [127].

Lung fibrosis caused by tungsten or tungsten carbide is included in the list of the German Berufskrankheitenverordnung (Occupation Disease Regulations) [128].

Since tungsten tends to accumulate, persons exposed at the workplace should undergo periodical medical checks including chest X-rays and ventilatory function tests for insoluble tungsten compounds, as well as examinations of the digestive tract and the central nervous system for soluble tungsten compounds [129].

15. References

1. R. Kieffer, G. Jangg, P. Ettmayer: *Sondermetalle,* Springer Verlag, Wien – New York 1971.
2. A. G. Quarrell (ed.): *Niobium, Tantalum, Molybdenum and Tungsten,* Elsevier, Amsterdam 1971.
3. Metallwerk Plansee, Wolfram, company brochure, Reutte, Austria, 1980.
4. S. W. H. Yih, C. T. Wang: *Tungsten: Sources, Metallurgy, Properties and Applications,* Plenum Press, New York – London 1979.
5. *Gmelin,* system no. 54, Suppl. vol. A 1.
6. *Gmelin,* system no. 54, part B 2; part B 3; part B 4.
7. *Ullmann,* 4th ed., **24,** 457.
8. B. Aronsson: "Some Notes about the History of Tungsten and Scheelite," booklet printed by AB Sandvik Coromant, R & D Materials and Processing Library, Stockholm 1978.
9. R. Eck, E. Lugscheider: "Molybdän, Wolfram und ihre Legierungen," Lecture on Sondermetalle at Technischen Akademie Esslingen, Esslingen 1980.
10. E. Fromm, E. Gebhardt: *Gase und Kohlenstoff in Metallen,* Springer Verlag, Berlin – Heidelberg – New York 1976.
11. Roskill Information Services: *The Economics of Tungsten,* London, Feb. 1990.

12. R. F. Stevens, Jr., *Bull. U.S. Bur. Mines* **667** (1975) 1161.
13. B. F. Kieffer, E. F. Baroch in H. Y. Sohn, O. N. Carlson, J. T. Smith (eds.): "Extractive Metallurgy of Refractory Metals," *Proc. TMS-AIME Annu. Meeting,* 1981, 273.
14. B. F. Kieffer, E. Lassner, *Tungsten Proc. Int. Tungsten Symp. 4th,* 1987, 59.
15. E. Lassner in H. Y. Sohn, O. N. Carlson, J. T. Smith (eds.): "Extractive Metallurgy of Refractory Metals," *Proc. TMS-AIME Annu. Meeting,* 1981, 269.
16. AMAX, US 4 338 287, 1983 (R. F. Hogsett, D. K. Huggins, L. B. Beckstead).
17. E. Lassner, B. F. Kieffer, *Conf. Proc. Adv. in Hard Materials Production,* London 1988, Contr. 5.
18. GTE Prod., US 4 353 878, 1982 (L. R. Quatrini, M. B. Terlizzi, B. E. Martin).
19. Engelhard Minerals and Chem., US 4 092 400, 1978 (V. Zbranek, D. A. Burnham).
20. AMAX, US 4 320 096, 1982 (P. B. Queneau, D. K. Huggins, L. W. Beckstead).
21. E. Lassner, *Österr. Chem. Ztg.* **80** (1979) 111.
22. AMAX, US 4 311 679, 1982 (P. B. Queneau, L. W. Beckstead, D. K. Huggins).
23. M. B. Macinnis, T. K. Kim, *Int. J. Refr. Hard Met.* **5** (1986) 78.
24. T. Zuo, Q. Zhao, S. Li, J. Wang, *Proc. 1st Int. Conf. Metall. Sci. W, Ti, Re and Sb* **1** (1988) 11.
25. J. Zhou, J. Xue, *Proc. 5th Int. Tungsten Symp.,* Budapest 1990, pp. 73 – 86.
26. E. Lassner in B. Lux, L. Bartha, E. Lassner, W. D. Schubert (eds.): *Chemistry of Doped Non Sag Tungsten, (in press).*
27. J. W. van Put, T. W. Zegers, A. van Sandwijk, P. J. M. van der Straten in C. S. Baird, C. Vijayan (eds.): *Proc. 2nd Int. Conf. Separation Sci. Technol.,* Hamilton, Canada, 1989, p. 387.
28. J. W. van Put, D. M. de Konig, A. van Sandwij, G. J. Witkamp in GVC.VDI/E.F.C.A. Mersman (ed.): *Proc. 11th Symp. Ind. Crystallization,* Garmisch Partenkirchen 1990, p. 647.
29. R. Haubner et al., *Int. J. Refr. Hard Met.* **2** (1983) 108, 156.
30. W. D. Schubert, *Int. J. Refr. Met. Hard Mat.* **9** (1991) 178.
31. W. D. Schubert, E. Lassner, *Int. J. Refr. Met. Hard Mat.* **10** (1991) 133, 171.
32. R. Haubner, W. D. Schubert, B. Lux, E. Lassner in H. M. Ortner (ed.): *Proc. 11th Plansee-Seminar* **2** (1985) 69.
33. F. Benesovsky (ed.): *Pulvermetallurgie und Sinterwerkstoffe,* 3rd ed., Metallwerk Plansee GmbH, Reutte, Austria, 1982.
34. W. Schintelmeister, O. Pacher, K. Kailer, W. Wallgram in H. M. Ortner (ed.): *Proc. 10th Plansee-Seminar* **2** (1981) 635.
35. Y. M. Korolev, V. A. Sorkin in H. Bildstein, H. M. Ortner (eds.): *Proc. 12th Plansee-Seminar* **1** (1989) 247.
36. J. P. Wittenauer, T. G. Niem, J. Wadsworth, *Adv. Mater. Process.* **9** (1992) 28.
37. H. S. Yu et al., *Int. J. Refr. Met. Hard Mat.* **11** (1992) 317.
38. R. Huenert, G. Winter, W. Kiliani, D. Greifendorf, *Int. J. Refr. Met. Hard Mat.* **11** (1992) 331.
39. V. G. Glebovsky in H. Bildstein, H. M. Ortner (eds.): *Proc. 12th Plansee-Seminar* **3** (1989) 379.
40. Fansteel Inc., Fansteel Corrosion Prevention Products, company brochure.
41. Fansteel Inc., Fansteel "61" Metal, company brochure, 1973.
42. Fansteel Inc., Fansteel "63" Metal, company brochure, 1973.
43. Metallwerk Plansee, Tantal- und Nioblegierungen, Planseeinformationen 16/82, Reutte, Austria, 1982.
44. D. V. Edmonds, *Int. J. Refr. Met. Hard Mat.* **10** (1991) 15.
45. Metallwerk Plansee, Densimet Schwermetall-Legierungen, company brochure, Reutte, Austria, 1982.
46. Metallwerk Plansee, Elmet Kontaktwerkstoffe, company brochure, Reutte, Austria 1982.
47. E. Pink, L. Bartha: *The Metallurgy of Doped Non-Sag Tungsten,* Elsevier, London 1989.
48. B. Zeiler, W. D. Schubert, B. Lux, *Int. J. Refr. Met. Hard Mat.* **12** (1993) 9 – 16.
49. O. Grau: "Sampling in the Non-Ferrous Industry," in G. Kraft (ed.): *Series of Bulk Material Handling,* vol. 6, Trans Tech Publ., Clausthal-Zellerfeld 1993, p. 131.
50. Gesellschaft deutscher Metallhütten- und Bergleute e. V. (eds.): *Analyse der Metalle,* Suppl. vol. 1, Springer Verlag, Berlin – Heidelberg – New York 1980, p. 213.
51. V. Scherer, D. Hirschfeld, *Erzmetall* **39** (1986) 251.
52. Gesellschaft Deutscher Metallhütten- und Bergleute e. V. (eds.): *Analyse der Metalle,* Suppl. vol. 1, Springer Verlag, Berlin – Heidelberg – New York 1980, p. 229.
53. H. M. Ortner, *Talanta* **26** (1979) 629.
54. D. Hirschfeld, H. M. Ortner, *Metall (Berlin)* **35** (1981) 434.

55. C. Vandecasteele, K. Strijckmans, C. Engelmann, H. M. Ortner, *Talanta* **28** (1981) 19.
56. H. M. Ortner in O. van der Biest (ed.): *Analysis of High Temperature Materials,* Applied Science Publ., London – New York 1983.
57. H. M. Ortner, *Erzmetall* **35** (1982) 555, 612.
58. E. Lassner, F. Benesovsky, *Mikrochim. Acta Suppl.* **5** (1974) 291.
59. H. M. Ortner, *Erzmetall* **32** (1979) 507.
60. E. Lassner, H. Petter, *Mikrochim. Acta Suppl.* **6** (1975) 133.
61. V. Scherer in H. M. Ortner (ed.): *Proc. 10th Plansee-Seminar* **2** (1981) 269.
62. W. D. Schubert, E. Lassner, *Int. J. Refr. Met. Hard Mat.* **10** (1991) 133.
63. P. Wilhartiz, H. M. Ortner, R. Krismer in H. Bildstein, H. M. Ortner (eds.): *Proc. 12th Plansee-Seminar* **3** (1989) 575.
64. K. G. Heumann, P. Herzer, H. E. Gäbler in H. Bildstein, H. M. Ortner (eds.): *Proc. 12th Plansee-Seminar* **4** (1989) 191.
65. H. J. Fischer, K.-D. Frank, *Erzmetall* **35** (1982) 403 –409.
66. T. Massalski: *ASM Binary Alloy Phase Diagrams,* 2nd ed., American Society for Metals, Materials Park, Ohio 1992.
67. Treibacher Chemische Werke, Lieferprogram Wolfram-Melting Base, Treibach 1991.
68. J. Becvar, *Freiberg. Forschungsh. B* **B 259** (1987) 61 –71.
69. Gesellschaft für Elektrometallurgie, Lieferprogramm, Nürnberg 1989.
70. *Ullmann,* 4th ed., **A 22,** 54.
71. *Winnacker-Küchler,* 4th ed., **4,** 180 – 182.
72. J. Eube in: *Stahl, Tabellenbuch für Auswahl und Anwendung, DIN,* 1st ed., Verlag Stahleisen, Düsseldorf 1992.
73. F. Binder, *Metall (Berlin)* **32** (1978) 1269 – 1273.
74. VDEH (eds.): *Grundlagen des Hochofenverfahrens,* Verlag Stahleisen, Düsseldorf 1973.
75. A. G. E. Robiette: *Electric Smelting Processes,* Charles Griffin, London 1973.
76. R. Durrer, G. Volkert: *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin – Heidelberg –New York 1972.
77. V. P. Elyutin, Y. A. Pavlov, B. E. Levin, E. M. Alekseev: *Production of Ferroalloys – Electrometallurgy,* 2nd ed., MIR Publ., Moscow 1957.
78. M. Riss, Y. Khodorovsky: *Production of Ferroalloys,* MIR Publ., Moscow 1967.
79. W. Mayer, *Erzmetall* **35** (1982) 343 – 346.
80. R. Püschel, E. Lassner: "Chelates and Chelating Agents in the Analytical Chemistry of Mo and W," in H. Flaschka, A. J. Barnard (eds.): *Chelates in Analytical Chemistry,* Marcel Dekker, New York 1967.
81. K. H. Tytko, O. Glemser, *Adv. Inorg. Chem. Radiochem.* **19** (1976) 239.
82. T. K. Kim, R. W. Mooney, V. Chiola, *Sep. Sci.* **3** (1968) 467.
83. H. J. Goldschmidt: *Interstitial Alloys,* Plenum Publ., New York 1967.
84. *Kirk-Othmer,* **23,** 426 – 438.
85. W. B. Pearson: *Handbook of Lattice Spacings and Structure of Metals,* Pergamon Press, New York 1967.
86. J. Booth, T. Eckström, E. Iguchi, R. J. D. Tilley, *J. Solid State Chem.* **41** (1982) 293.
87. A. J. Hegedüs, T. Millner, J. Neugebauer, *Z. Anorg. Allg. Chem.* **281** (1955) 64 – 82.
88. W. R. Morcom, W. L. Worrell, H. G. Sell, *Met. Trans.* (1974) 155 – 161.
89. O. Glemser, H. Ackermann, *Z. Anorg. Allg. Chem.* **325** (1963) 281 – 286.
90. G. Yidong, *Proc. 1st Int. Conf. Metall. Mat. Sci. W, Ti, Re and Sb,* **1** (1988) 213.
91. T. Eckström, R. J. D. Tilley, *Chem. Sci.* **16** (1980) 1.
92. GTE Sylvania, US 4 175 109, 1979 (T. K. Kim).
93. Tang Xinhe, Ph. D. Thesis, Technical University, Vienna 1994.
94. Sylvania, Technical Information Bull. CM-9003, no. 9, Towanda, PA, 1980.
95. Misono Makato: "Catalysis Review," *Sci. Eng.* **29** (1987) 269.
96. J. C. Wildervanch, F. Jellinek, *Z. Anorg. Chem.* **328** (1964) 309.
97. O. Glemser, H. Sauer, B. König, *Z. Anorg. Allg. Chem.* **257** (1948) 241.
98. J. Desilvestro, M. Neumann-Spallart, *J. Phys. Chem.* **89** (1985) 3684.
99. I. Bedja, S. Hotchandani, P. Kamat, *J. Phys. Chem.* **97** (1993) 11064.
100. M. Neumann-Spallart, *Hydrogen Energy Prog. VI* **12** (1986) 577.
101. International Tungsten Industry Association (eds.): *Newsletter,* 280 Earls Court Road, London SW5 9AS, June 1993.
102. K. Kase: *Tungsten Sui Soda,* Tokyo 1991.
103. E. Lassner, W. Aschenbrenner, *Proceedings of the 6th Tungsten Symposium,* Guangzhou 1993.
104. H. E. Stokinger in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., Wiley-Interscience, New York 1981, pp. 1981 –1995.

105. G. Kazantzis in L. Friberg, G. F. Nordberg, V. Vouk (eds.): *Handbook on the Toxicology of Metals,* 2nd ed., Elsevier Science Publisher, Amsterdam 1986, pp. 610 – 622.
106. R. Wennig, N. Kirsch in H. G. Seiler, H. Sigel, A. Sigel (eds.): *Handbook on Toxicity of Inorganic Compounds,* Marcel Dekker, New York 1988, pp. 731 – 738.
107. S. V. Kaye, *Health Phys.* **15** (1968) 398 – 417.
108. H. J. Cohen, R. T. Drews, J. L. Johnson, K. V. Rajagopalan, *Proc. Natl. Sci. USA* **70** (1973) 3655 – 3659.
109. R. Kühn, K. Birett: *Merkblätter gefährliche Arbeitsstoffe,* 4th ed., ecomed Verlag, Landsberg/Lech, Datenblatt W-004.
110. R. Krüger, *Muench. Med. Wochenschr.* **35** (1912) 1910.
111. E. O. Coates, J. H. L. Watson, *Ann. Intern. Med.* **75** (1971) 709 – 716.
112. N. L. Sprince, R. I. Chaberlin, C. A. Hales, A. L. Weber, H. Kazemi, *Chest* **86** (1984) 549 – 557.
113. H. Kitamura, *Yokohama Med. Bull.* **31** (1980) 103 –126
114. H. Kitamura, Y. Yoshimura, I. Tozawa, K. Koshi, *Acta Pathol. Jpn.* **30** (1980) 241 – 254.
115. H. A. Schroeder, M. Mitchener, *J. Nutr.* **105** (1975) 421 – 427.
116. H. A. Schroeder, M. Mitchener, *J. Nutr.* **105** (1975) 452 – 458.
117. V. G. Nadeenko, V. G. Lenchenko, S. B. Genkina, T. A. Arkhipenko, *Farmakol. Toksikol. (Moscow)* **41** (1978) 620 – 623.
118. M. Wide, *Environm. Res.* **33** (1984) 47 – 53.
119. I. Rystedt, T. Fischer, B. Lagerholm, *Contact Dermatitis* **9** (1983) 69 – 73.
120. American Conference of Governmental Industrial Hygienists (eds.): *Documentation of Threshold Limit Values for Substances in Workroom Air,* Cincinnati, Ohio 1980, p. 273.
121. F. W. Kinard, J. van de Erve, *Am. J. Med. Sci.* **199** (1940) 668 – 670.
122. E. H. Vernot, J. D. MacEwen, C. C. Haun, E. R. Kinkead, *Toxicol. Appl. Pharmacol.* **42** (1977) 417 –423.
123. National Technical Information Service, AEC-TR-6710. English translation of: Z. E. Izraelson: *Toxicology of the Rare Metals,* Moscow 1963.
124. M. J. Ellenhorn, D. G. Barceloux: *Medical Toxicology, Diagnosis and Treatment of Human Poisoning,* Elsevier, New York 1988, p. 77.
125. S. Moeschlin: *Klinik und Therapie der Vergiftungen,* 7th ed., Georg Thieme, Stuttgart 1986, p. 201.
126. Deutsche Forschungsgemeinschaft (ed.): *MAK- und BAT-Werte-Liste 1993,* Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte, Mitteilungen der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 29, VCH Verlagsgesellschaft, Weinheim 1993.
127. American Conference of Governmental Industrial Hygienists (eds.): *Threshold Limit Values (TLV) 1982,* Cincinnati, Ohio 1982.
128. Berufskrankheitenverordnung, BK Nr. 41 07: Erkrankungen durch Lungenfibrose durch Metallstäube bei der Herstellung oder Verarbeitung von Hartmetallen.
129. R. Roi, W. G. Town, W. G. Hunter, L. Alessio (eds.): *Occupational Health Guidelines for Chemical Risks,* CEC Publ. Eur 8513, Luxembourg 1983, pp. 128 – 129.

Turpentines

Terpenes is a separate keyword.

Manfred Gscheidmeier, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Chaps. 1 – 13, 15 – 16)

Helmut Fleig, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 14)

1. Turpentine as a Natural Product . . 1
2. History 2
3. Definition and Classification 2
4. Production 3
4.1. Gum Turpentine 3
4.2. Wood Turpentine 4
4.3. Sulfate Turpentine 4
4.4. Sulfite Turpentine 4
4.5. Destructively Distilled Wood Turpentine 5
5. Composition 5
6. Physical Properties 7
7. Chemical Properties 7
8. Analysis 8
9. Safety and Occupational Health . . . 8
10. Storage and Transportation 9
11. Ecological Aspects 9
12. Uses 9
13. Economic Aspects 10
14. Toxicology 11
15. Turpentine Substitute 12
16. Pine Oil 12
16.1. Production 12
16.2. Properties 12
16.3. Storage and Transportation 13
16.4. Uses 13
16.5. Economic Aspects 13
17. References 14

1. Turpentine as a Natural Product

Turpentine, which occurs in many natural forms, is a component of conifer balsam and plays an important role in the ecosystem of the pine forest [5]. It is part of the plant metabolism, and its individual components are important for the protective system of the conifers: for rapid closure of wounds to the bark by resin formation, for repelling insects (e.g., myrcene repels the pine beetle), and for protection against attack by bacteria (effect of, e.g., 3-carene).

The volatile terpenes (and polymers derived from them) emitted to the atmosphere are partially absorbed by the forest soil and contribute to humus formation [6]. This emission occurs to a great extent, particularly in summer. In Russian pine forests, terpene concentrations of 3.5 – 35 µg per cubic meter of air have been measured [7], [8]. In similar investigations in the state of Pennsylvania (total area 117 350 km^2, of which 62 000 km^2 is forest) an emission of 3400 t of terpene hydrocarbons was found for one day in the month of August [9]. The annual worldwide biogenic production of terpenes is estimated to be 10^9 t in 1984 [10]. Since this quantity is degraded by natural routes, the environmentally relevant terpenes can be considered problem-free with regard to air pollution. Only in combination with nitrogen oxides and with UV radiation may ozone be formed.

2. History

In ancient times turpentine obtained from conifer balsam was very important as a fuel for lamps and torches; as a solvent; as a fragrance; and as an embalming oil. Cuneiform script texts and archaeological finds indicate that turpentine was known by around 3000 BC in Mesopotamia [11] and Pakistan [12]. The preparation and use of distilled cedarwood oil ("turpentine oil" from *Pistacia terebinthus*) were described in early Egyptian culture (e.g., for embalming [13]). Turpentine was also known to the Greeks, Celts, and Romans. In alchemists' books of the 13th century, turpentine oil was described as *aqua ardens*, and in the distillery book of the Strasbourg doctor H. Brunschwygh (1450 – 1543) as *spiritus terpenthinae*.

In China, turpentine was used early on as a solvent for rosin (colophony) for impregnating

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_267

sail canvas and rigging and sealing ships' hulls. This kind of use was very important among the maritime powers in the Middle Ages.

In Europe, *gum spirit of turpentine* (also known as *gum turpentine*) was produced in the 16th century in Poland, Germany, Scandinavia, and Russia; in the 19th century in France, Spain, Portugal, and Finland; and since 1902 in Greece. In North America, turpentine has presumably been extracted since 1606 in Nova Scotia. Since 1710 it has been produced to a greater extent in North Carolina, and later in other states. For a long time a mixture of turpentine and spirit – the lamp fuel camphine – was an important commercial product.

In ca. 1870, steam distillation of turpentine was introduced, followed in 1910 by extraction from tree stumps left after deforestation. In this way, the cheaper *wood turpentine* and *pine oil* came on the market.

Sulfate turpentine, a byproduct of pulp production, was obtained in small quantities in Norway and Sweden already before 1910, in Finland since 1911, and in the United States since 1918. Only since 1930 has it been available in large quantities as a low-priced raw material [14]. From 1954 onward its production in the United States overtook that of all turpentines of other origins.

The first elemental analysis of turpentine was carried out in 1818 by J. J. Houton de Labillardiere. A little later the essential oils of gum turpentine were investigated scientifically by J. B. Dumas (1833), J. J. Berzelius, J. von Liebig, and F. Wöhler (1837). In 1862, M. Berthelot began his studies on terpene hydrocarbons (e.g., pinene hydrochloride), and 1877 W. Tilden began to characterize the hydrocarbon components of spirit of turpentine by means of their crystalline nitrosyl chlorides. In 1884, O. Wallach succeeded in structurally characterizing the components of turpentine oils, and in 1893 A. von Baeyer began exploring the field of cyclic terpenes [15]. More recently, the biosynthesis of monoterpenes has been investigated by many scientists [16–21].

3. Definition and Classification

Turpentine [*8006-64-2*] (oil of turpentine, turpentine oil, spirit of turpentine) is the collective name for the volatile components of the crude resins (balsam or oleoresin, Latin: *resina terebinthi*) from conifers, particularly the pine family. The turpentines represent a subgroup of plant and tree saps (essential oils).

Depending on the starting material and production method, the following types of turpentine are defined according to DIN 53 248 (April 1977): gum spirit of turpentine, steam-distilled wood turpentine, refined sulfate wood turpentine, and destructively distilled wood turpentine.

The use of the term turpentine without further specification is not permitted. Usually, the name of the country of origin is added (e.g., Greek gum spirit of turpentine). The addition of products that are not typial components of turpentine must be declared.

Gum spirit of turpentine (Latin: *oleum terebinthinae*, French: *essence de térébinthine*, German: *Balsamterpentinöl*) according to DIN 53 248 denotes only the pure essential plant oil obtained by distillation of the resin exudate (balsam or turpentine) from living trees of the genus *Pinus* at up to 180 °C (under normal pressure). The name must not be used for material obtained after the extraction of several components (e.g., pinene).

Steam-distilled wood turpentine (French: *essence de bois distillée à la vapeur*, German: *Wasserdampfdestilliertes Wurzelterpentinöl*) is a turpentine obtained either directly from the steam distillation of finely chopped wood from roots and stumps of trees of the *Pinus* family or after extraction with solvents, which is otherwise unchanged.

Refined sulfate wood turpentine (French: *essence de bois dite au sulfate* or *essence de papeterie de sulfaté desodorisée*, German: *Gereinigtes Sulfatterpentinöl*) is a byproduct of pulp production using the soda or sulfate pulping process (→ Paper and Pulp, Chap. 1.3.2.), that is subsequently purified.

Crude sulfate turpentine (French: *essence de papeterie de sulfate brute*, German: *Rohsulfatterpentinöl*), which is not purified, is now rarely marketed, because the sulfur compounds it contains produced a nauseating odor.

Destructively distilled wood turpentine (French: *essence de bois par destillation destructive*, German: *Gereinigtes Kienöl*) is the name given to the oil obtained by dry distillation of resin-rich pinewood followed by purification. It must not be called turpentine or oil of turpentine.

Sulfite turpentine and pine oil are not defined according to DIN 53 248, but are generally related to the other turpentines (see Section 4.4, Chap. 16). Sulfite turpentine is a byproduct of pulp production by the sulfite pulping process (→ Paper and Pulp, Chap. 1.3.2.). Its composition (main component *p*-cymene) differs significantly from that of the other turpentines [22].

The term *naval stores* relates to all rosin and oil products extracted from conifer wood, such as turpentine, rosin (→ Resins, Natural, Chap. 5.), pine oil, or tall oil (→ Tall Oil). It originates from the time when the United States was still an English colony and when these types of product were used widely in shipbuilding (e.g., in paint, to seal ships' hulls, or to treat rigging) and were therefore stored and sold in naval stores (→ Resins, Natural, Chap. 5.).

4. Production

4.1. Gum Turpentine

Gum turpentine is extracted in many countries with warm climates (more frequently in the Northern Hemisphere) from the wood resin or crude balsam of various conifers, particularly those of the pine family (see also → Resins, Natural, Chap. 5.1.) [23]. Incisions made in the trunks of fully grown trees are deep enough to allow crude balsam to flow out of the resin channels. It can then be collected in removable vessels. These mostly V-shaped cuts ("lachte") may be applied only to a small section of the circumference to protect the tree and must be extended upward or downward several times a month. A further development involves inserting poly(ethylene terephthalate) bottles or small polyethylene bags into holes that have been bored in the bark and injected with an irritant. In this way, not only is the tree protected, but oxidation and contamination of the balsam and evaporation of the terpenes are avoided [24].

The crude balsam collected is first freed of foreign material, such as sand, wood, etc. To this end, either in molten form or dissolved in turpentine, the balsam is filtered through a finely meshed sieve. To remove irritants, etc., the product is washed with hot water and distilled. Normal distillation has been replaced extensively by more gentle steam distillation [3]. Rosin remains as the turpentine-free residue. It is removed as the molten resin and transferred to metal drums.

In certain operations, solid high-boilers (pitch) can also be formed. These compounds are found in wood extracts in quantities of 2 to 8 % and consist mainly of triglycerides, fatty acids, resin acids, and waxes [25].

The annual yield of crude balsam per tree is ca. 1.5 kg for Scots pine, 2.5 – 4 kg for black pine, and ca. 0.6 kg for spruce. With Central American species, more than 5 kg can be obtained. The average proportion of turpentine is 13 – 25 %.

By means of resinification using irritants – i.e., treating cuts in the tree with mineral acid, particularly sulfuric acid; with the herbicides diquat and paraquat; or with 2-chloroethylphosphonic acid (CEPA) – the yield can be increased two- to fivefold [24], [26], [27].

4.2. Wood Turpentine

Wood turpentine is now produced in only a few places, such as the United States (see also → Resins, Natural, Chap. 5.2.). It is extracted from finely chopped stump wood either by steam distillation or with solvents such as naphtha or chlorohydrocarbons (under pressure if required). The extract is subjected to either fractional or steam distillation. Wood turpentine is obtained as the lower-boiling fraction and pine oil as the higher-boiling one. Limonene can sometimes be isolated as an intermediate fraction [28], [3]. Ca. 2.2 wt % turpentine, 1.5 wt % pine oil, and 0.4 wt % limonene can be extracted from stumps of yellow pine that are at least eight years old. From stumps of Scots pine, only ca. 0.75 wt % turpentine is obtained. The yield increase with the age of the stump because the hollows between the dead and dried-out cells fill with turpentine balsam. This enrichment can be accelerated by boring holes in the tree just above

the roots two years before felling and filling the holes with paraquat [29].

4.3. Sulfate Turpentine

Sulfate turpentine is isolated as a byproduct in pulp production (→ Paper and Pulp, Chap. 1.5.1.) and is most important in the United States, Scandinavia, and Russia.

In the kraft (sulfate) pulping process, debarked and finely chipped wood is heated to 150 – 180 °C in an aqueous digestion liquor containing NaOH, Na_2S, Na_2CO_3, and small quantities of Na_2SO_4, Na_2SO_3, and $Na_2S_2O_3$ in large pressure vessels at 7 – 13 bar for 1 – 5 h. Crude sulfate turpentine is condensed from the waste gas from the digester and is separated from the water (→ Paper and Pulp, Chap. 1.3.1.1.). The process can also be carried out continuously in horizontal or vertical tube digesters [30–34] (→ Paper and Pulp, Chap. 1.4.1.), enabling more than 74 % of the turpentine originally present to be extracted.

Crude sulfate turpentine still contains 5 – 15 wt % sulfur compounds with very nauseating odors, mainly methanethiol, dimethyl sulfide, and dimethyl disulfide [35]. These can be oxidized with sodium hypochlorite solution (bleaching liquor) at ca. 60 °C to the less volatile sulfonic acids, sulfoxides, or sulfones [36], which can be removed by washing with water or by distillation. Additional methods for removing sulfur compounds are treatment with chloride of lime [37] or sodium peroxide [38], careful distillation [39], and blowing air or inert gas through the product. These measures may be used individually or in combination. Even extensively purified sulfate turpentine contains traces of sulfur and chlorine compounds, which give a characteristic musty odor.

The yield of sulfate turpentine is 6 – 16 kg per tonne of pulp for pine, ca. 2 kg per tonne for spruce, and 2 – 3 kg per tonne for fir trees. The location and date of the wood harvest and the length of the storage period before processing affect the yield. In winter, up to 30 % more turpentine can be obtained than in summer. Wood chips can lose up to 80 % of their turpentine content in ten weeks [40].

About 55 % of the pulp factories worldwide use the sulfate digestion process.

4.4. Sulfite Turpentine

Sulfite turpentine is a byproduct of pulp production by the sulfite process (→ Paper and Pulp, Chap. 1.3.2.). Very small wood chips are heated in pressure vessels (pulp digesters) with pulping liquor. The latter consists mainly of $Ca(HSO_3)_2$ and SO_2 in acid bisulfite pulping (pH 1.5 – 2); $Mg(HSO_3)_2$ in bisulfite pulping (pH 4); and Na_2SO_3, $NaHCO_3$, $(NH_4)_2SO_3$, and NH_3 in the neutral sulfite process (pH 7 – 9). Wood chips are first impregnated at temperatures that increase to 110 °C under alternating reduced and overpressure. Then they are digested at 125 – 160 °C and 7 – 13 bar. The overall process takes 6 – 10 h [34], [41–44]. After the waste gas has been driven off, the crude oil separates on top of the pulping liquor as a yellow-brown, weakly green-fluorescing liquid. From this the almost colorless sulfite turpentine, *bp* 176 –180 °C, is obtained by neutralization with sodium hydroxide or lime and distillation (possibly using steam) in a yield of 0.3 – 1 kg per tonne of pulp [37]. Sulfite turpentine consists of 70 – 90 % *p*-cymene, which is formed from the terpene components originally present in a reaction catalyzed by hydrogensulfite. It also contains dipentene, borneol, and other mono- and sesquiterpenes [22] and, when mild digestion conditions are used, significant quantities of α-pinene.

4.5. Destructively Distilled Wood Turpentine

The production of destructively distilled wood turpentine by the dry distillation of resin-rich pinewood (predominantly stumps) is no longer important. For more information, see [45]. This product contains little α-pinene, but high quantities of limonene (dipentene), *p*-cymene, camphene, α-terpinene, and sylvestrene.

5. Composition

The main components of turpentines are mono- and bicyclic monoterpenes ($C_{10}H_{16}$; M_r 136.2). They also contain smaller quantities of terpene – oxygen compounds (e.g., terpene alcohols), sesquiterpenes ($C_{15}H_{24}$), and other compounds (see Fig. 1).

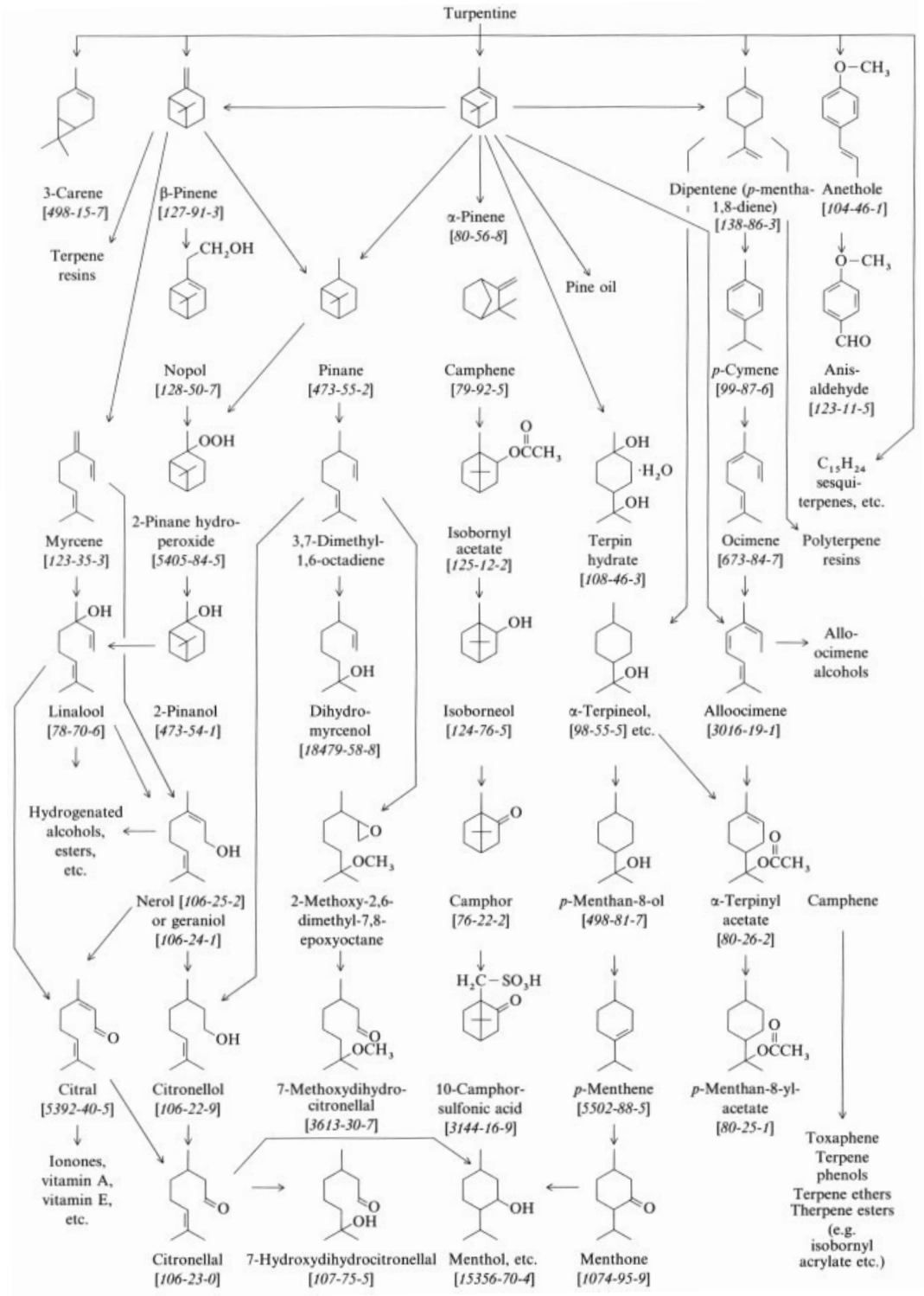

Figure 1. Turpentine constituents and downstream products

The composition of different turpentines depends on the species of pine from which they are extracted [46]. About 105 species and varieties have now been described which can be distinguished to only a limited extent on the basis of turpentine composition (their "terpene pattern"). Species of pine that are important for turpentine extraction are listed in Table 1.

Table 1. Species of pine that are important for turpentine extraction

Habitat	Systematic name	Common name
Western Europe	*P. pinaster* Sol.	French maritime pine
Southeastern Europe	*P. nigra* Arnold	black pine, Austrian pine
Greece	*P. halepensis* Mill.	Aleppo pine
Mid- and Eastern Europe and Russia	*P. sylvestris* L.	Scots pine
Siberia	*P. pumila*	
China	*P. massoniana*	Chinese red pine
	P. langbianensis	
	P. elliottii Engelm.	slash pine
India	*P. longifolia* Roxb.	
Southeast Asia	*P. armandi*	
Indonesia	*P. mercussii* Junk	Mindoro pine
United States	*P. palustris* Mill.	longleaf pine
	P. ponderosa Laws	yellow pine
Southwest USA	*P. edulis* Engelm.	
Southeast USA	*P. taeda* L.	loblolly pine
	P. elliottii Engelm.	slash pine
South Brazil	*P. elliottii* Engelm.	slash pine
	P. taeda	loblolly pine
Argentina	*P. taeda* L.	loblolly pine
	P. elliottii Engelm.	slash pine
Mexico	*P. oocarpa* Schiede	
Honduras	*P. montezuma* and others	
Northern Brazil	*P. caribaea* M.	
Venezuela	*P. caribaea* M.	

The habitat of the trees generally has little effect on the ratio of components in the turpentine oil extracted. However, for some varieties of pine (e.g., *Pinus ponderosa* Laws) the turpentine varies significantly depending on the climate, age, and habitat of the trees [47], [48].

α-Pinene is the main component of most turpentine oils. Many turpentines also contain large quantities of β-pinene, *d*- or *l*-limonene (or its racemate dipentene), 3-carene, or α-terpinene (sometimes up to 15 %). Table 2 gives a simplified overview of the average composition and physical properties of various turpentines. The turpentine oils on the market are not specified according to species of pine but according to producer country, so that rather pronounced deviations in composition are possible, depending on starting material. More details can be found in [1–3], [49], [50] (with many additional references).

Besides the components listed in Table 2, turpentines can also contain α- and γ-terpinene, α- and β-phellandrene, terpinolene, myrcene, *p*-cymene, bornyl acetate, anethole, sesquiterpenes (such as longifolene; up to 20 % in *Pinus longifolia* Roxb.), heptane, nonane, and many other compounds [51]. Some turpentines do not contain α-pinene as the main constituent but, for example, β-pinene (75 % in *Pinus clausa* Engelm.), *l*-limonene (80 % in *Pinus pinceana* Gordon), β-phellandrene (96 % in *Pinus contorta* Douglas), or heptane (95 % in *Pinus sabiniana* Douglas).

Turpentines usually contain 0.3 – 2.5 wt % of nonvolatile components in the evaporation residue, or considerably more if resinification occurs due to improper storage or handling.

Turpentines from spruce or fir contain significantly less α-pinene but higher quantities of camphene, α-terpinene, phellandrene, myrcene, cyclofenchene, and bornyl acetate, for example.

6. Physical Properties

Turpentines are colorless to pale-yellow liquids with low viscosity and a characteristic odor. In gum turpentine this odor is pleasantly aromatic. The physical data can vary within certain limits, depending on composition (see Table 2):

Boiling range (101.3 kPa)	150 – 177 °C
Solidification point	<-40 °C
ϱ (20 °C)	0.855 – 0.872 g/mL
$[\alpha]_D^{20}$	-40 to $+48°$
n_D^{20}	1.465 – 1.478
Relative evaporation rate (diethyl ether: 1)	ca. 170
Calorific value	ca. 40 000 kJ/kg

Turpentines are miscible with most organic solvents, but immiscible with water. They are also good solvents for hydrocarbon resins, waxes, paints and coatings, fats, and oils. Ca. 0.06 wt % of water is absorbed at 20 °C.

Turpentine is not listed in either DAB 10 or the European Pharmacopoeia.

Table 2. Simplified overview of the composition of turpentines of different origins

Type, origin	Proportion in wt % of					Density at	Rotation, $[\alpha]_D^{20}$	Distillation range at
	α-Pinene	β-Pinene	Camphene	3-Carene	Limonene (dipentene)	20 °C, g/mL		101.3 kPa, °C
Gum turpentine								
Greece	92 – 97	1 – 3	ca. 1	0 – 1	0 – 2	0.860 – 0.865	+34° to +48°	151 – 157
Mexico	70 – 95	2 – 15	2 – 15	1 – 2	0 – 4	0.862 – 0.868	+13° to +35°	151 – 161
China	70 – 95	4 – 15	1 – 2	0 – 5	1 – 3	0.860 – 0.865	−20° to +30°	151 – 160
Portugal	70 – 85	10 – 20	ca. 1		1 – 4	0.860 – 0.870	−24° to −38°	151 – 173
South America	45 – 85	5 – 45	1 – 3	0 – 2	2 – 4	0.860 – 0.870	−20° to	
Indonesia	65 – 85	1 – 3	ca. 1	10 – 18	1 – 3	0.865 – 0.870	+20° to +35°	152 – 162
France	65 – 75	20 – 26	ca. 1		1 – 5	0.865 – 0.871	−28° to −33°	152 – 173
Russia	40 – 75	4 – 15	1 – 5	0 – 20	0 – 5	0.855 – 0.865		152 – 172
Poland	40 – 70	2 – 15	ca. 1	0 – 25	1 – 5	0.855 – 0.865	+15° to +20°	152 – 172
USA – Canada	40 – 65	20 – 35	ca. 1	0 – 4	2 – 20	0.860 – 0.870	−10° to −30°	151 – 174
New Zealand	30 – 50	40 – 60						
India	20 – 40	5 – 20	1 – 5	45 – 70		0.850 – 0.865	−5° to −15°	153 – 174
Sulfate turpentine								
Finland	55 – 70	2 – 6	ca. 1	7 – 30	ca. 4	0.860 – 0.870	+2° to +21°	151 – 177
Sweden	50 – 70	4 – 10	ca. 1	15 – 40	1 – 3	0.860 – 0.870	+2° to +18°	150 – 175
USA	40 – 70	15 – 35	1 – 2	2 – 10	5 – 10	0.864 – 0.870	+2° to +20°	152 – 172
Russia	55 – 70	1 – 5	1 – 8	10 – 25	3 – 8	0.858 – 0.868	+2° to +20°	152 – 172
Wood turpentine								
USA	75 – 85	0 – 3	4 – 15		5 – 15	0.860 – 0.875	−26° to +10°	152 – 175

7. Chemical Properties

The chemical properties of turpentine are determined by its major components, particularly the reactive pinenes. Typical reactions include isomerization, disproportionation, and polymerization catalyzed by acids, acid salts, surfactants, or Friedel – Crafts catalysts. Hydrates can be formed in the presence of water. Addition reactions take place with halogens, hydrogen halides, and nitrosyl chloride. Turpentine reacts with phosphorus pentasulfide and hydrogen sulfide to form complex sulfur compounds and thiols. Carboxylic acids add to give esters, and alcohols to give ethers. Pyrolysis yields mixtures of acyclic and monocyclic terpenes, pyronenes, hydroterpenes, and *p*-cymene. Moist air leads to autoxidation of α-pinene to sobrerol (*trans-p*-menth-6-ene-2,8-diol) and resins. These reactions can be inhibited by the addition of 0.005 – 1.0 % of certain amines or phenols acting as antioxidants (e.g., diphenylamine, thymol, catechol, resorcinol, pyrogallol, naphthols, or 2,6-di-*tert*-butyl-*p*-cresol) [52], [53].

8. Analysis

German Standards. The usual tests that should be referred to in test reports are listed in DIN 53 248 (solvents for paints, turpentine, and destructively distilled wood turpentine, April 1977). Besides appearance, color, and odor, the following characteristics should be determined:

Density	(DIN 51 757, ISO 279)
Refractive index	(DIN 53 169, ISO 280)
Flash point	(DIN 53 213; ISO 3679)
Acid number	(DIN 53 402, ISO 709)
Distillation range	(DIN 53 171 with limitations according to DIN 53 248, ISO 3405)
Residue on evaporation and material insoluble in sulfuric acid	(DIN 53 248, ISO 4715)

Determination of the residue on evaporation should be carried out under a weak stream of nitrogen to avoid oxidation and, thus, erroneous results. Determination of the bromine number of turpentines has not been standardized internationally [54].

U.S. Standards. In the United States, specifications according to ASTM D 13-92 for appearance, color, odor, and physical data of various turpentines are used. Sampling and testing must be carried out according to ASTM D 233-92. The sulfur present in sulfate turpentine, mostly in the form of volatile compounds, can be determined according to ASTM D 1266-91 [55].

Steam-distilled wood turpentine always contains traces of benzaldehyde. Detection of the latter is characteristic of this turpentine [56].

Methods of Analysis. Gas chromatography is used most frequently for the determination of individual components (DIN 51 405, ASTM E 260-91) [57–60]. Capillary columns with surface coatings [e.g., 50-m steel capillary columns with poly(propylene glycol) or poly(phenyl ether) coating or 25- or 50-m glass capillary columns with Carbowax 20 M or dimethylsilicone coating] are particularly effective. Accurate evaluation of gas chromatograms is occasionally difficult because of the possibility of complex reactions, such as isomerization or dehydration, in the injection unit or the column [61].

Head space gas chromatography (HSGC) has proved useful for the investigation of natural products [62].

9. Safety and Occupational Health

Turpentines are volatile, flammable liquids whose vapors are 4.7 times heavier than air; they can form explosive vapor – air mixtures on being heated above the flash point.

Flash point	32 – 35 °C
Ignition temperature	220 – 240 °C
Explosion limits (101.3 kPa)	ca. 0.8 – 6 vol % (45 – 340 g per cubic meter of air)
Ignition group (according to VDE 0165)	G 3

Handling of turpentine above the flash point must be carried out in explosion-proof apparatus [63]. Furthermore, the fact that turpentine can self-ignite in contact with air on distribution over large surface areas (e.g., on cleaning cloths or insulating materials) must be taken into account. Turpentine also ignites on contact with concentrated nitric acid.

MAK value	100 ppm (560 mg/m^3) [64]
Short-term exposure limit	5 min 2 × MAK 8 times per day
Sensitization	hypersensitive reactions of skin and respiratory tract are possible
Skin absorption	poisoning through skin absorption is possible

When handling turpentine, inhalation of the vapor, swallowing, and contact with the skin or eyes should be avoided (see Chap. 14). According to GefStoffV the following container labeling for turpentine is stipulated: St. Andrew's Cross and symbol for flammable liquids. In Germany further security information is given in VBG 1 (UVV 1 of BG Chemie), Kühn-Birett Merkblatt T 02, and Hommel-Merkblatt 190.

Both "spirit of turpentine" [*8006-64-2*], EINECS no. 232-350-7 and "turpentine, its extracts and physically modified derivatives" [*9005-90-7*], EINECS no. 232-688-5 are listed in the register of existing substances (EINECS) according to ChemG (September 1981).

10. Storage and Transportation

Turpentines should be stored in locations that are as cool as possible. Light, air, and acidic substances should be excluded to minimize changes in composition. The regulations governing equipment for storage, filling, and transport of substances hazardous to water and the regulations governing flammable liquids (VbF) must be adhered to.

The following classifications are used for the storage and transportation of turpentine:

VbF	group A class II
GGVE/GGVS	class 3, item 31 c
RID/ADR/ADNR	class 3, item 31 c
IATA/DGR	2319
IMDG code	class 3.3 – 2319
UN no.	2319

According to the GGVS for road transportation of turpentine, labeling with orange hazard notices and the labeling no. 30/2319 is stipulated. The same applies to α-pinene (labeling no. 30/2368).

11. Ecological Aspects

Being truly natural products, turpentines and their individual components are completely biodegradable usually in a few days, depending on the degree of dilution, temperature, availability of air, and presence of bacteria. Nevertheless, according to current estimation of the potential

hazard to water, and in view of the great variety of turpentines they are assigned to the German water hazard class II (WGK II). Below their solubility limit, they do not represent any hazard to biological wastewater-treatment plants.

Turpentines are not classified in TA Luft, but by analogy to pinenes, they may be assigned to TA Luft hazard class III.

The ozone depletion potential (ODP) and the global warming potential (GWP) of turpentines are given as zero. Even the Clean Air Act of 1990 does not list terpenes among the 190 substances that cause air pollution [65].

12. Uses

The direct use of turpentine is now limited mainly to that as a solvent or dilution agent, e.g., for natural or modified binders (mainly alkyd resins), for oils, coatings, resins, paints (also artists' paints), and floor and shoe polishes (→ Solvents, Chap. 9.3.). It accelerates, for example, the drying of coatings and paints by peroxide formation, in particular of those based on drying oils or other film-forming agents. Higher-boiling turpentine fractions (containing large quantities of 3-carene) are suitable as pesticides for use in forestry [66].

Although the importance of unfractionated turpentine as a solvent has clearly decreased as a result of the use of cheaper petrochemical products, turpentine has attracted increasing interest as a raw material for the chemical industry. By fractionating distillation under reduced pressure, it can be separated into fractions or individual components. Separation by process-scale chromatography has also been described [67]. The main product is usually α-pinene, whose purity can reach 99 %. β-Pinene [68], 3-carene, and monocyclic terpenes [α-terpinene, limonene (dipentene), and phellandrene] are obtained in smaller quantities. Higher-boiling fractions are obtained from the residue by steam distillation. Terpene alcohols, sesquiterpenes, and diterpenes are thus obtained. The resin remaining has a composition similar to rosin and can be used for the same applications.

From the residue of sulfate turpentine distillation, estragole [*140-67-0*] (1-allyl-4-methoxybenzene) and anethole (1-propenyl-4-methoxybenzene), mostly contaminated with caryophyllene, can also be obtained.

From the different components of turpentine, particularly α- and β-pinene, a large number of substances are synthesized, some on an industrial scale. This is shown in simplified form in Figure 1 (see also [69–74] with many additional references).

The following possible uses of turpentine and its downstream products are known (→ Terpenes, Chap. 1.5.):

1) Flavors and fragdances Flavors and Fagrances, Chap. 2.2., Flavors and Fagrances, Chap. 2.3.)
2) Pharmaceuticals (e.g., camphor, menthol)
3) Disinfectants (e.g., terpineol, *p*-cymene)
4) Pesticides (e.g., toxaphene, 3-carene, terpene halides and thiocyanates)
5) Cleaning agents (e.g., dipentene, *d*- or *l*-limonene)
6) Solvents (e.g., turpentine and its individual components *p*-cymene, *p*-menthane, and pine oil)
7) Resins (e.g., β-pinene polymers)
8) Varnish resins (e.g., terpene phenols, isobornyl acrylate, and terpene – maleate resins)
9) Melt adhesives (e.g., terpene polymers, terpene phenols)
10) Polymerization accelerators for rubber (e.g., pinane or *p*-menthane hydroperoxide)
11) Antioxidants (e.g., isobornyl phenols)
12) Plasticizers (e.g., camphor, esters of pinic acid)
13) Lubricating oil additives (e.g., sulfur-containing terpenes)
14) Traction fluids (e.g., terpene ethers)
15) Textile auxiliaries (e.g., terpineol, pine oil)
16) Flotation aids (e.g., pine oil)

In case of overproduction sulfate turpentine may occasionally be used as a fuel (e.g., in Canada).

13. Economic Aspects

The worldwide production of turpentines in 1990 was ca. 315 000 t, of which ca. 190 000 t was sulfate turpentine. A breakdown of production among the most important producer coun-

Table 3. Turpentine production (1000 t/a) in the most important producer countries with percentage of sulfate turpentine; 1990

Country	1960	1970	1980	1990	Sulfate turpentine, %
United States	99	94	86	105	97
Russia	28	42	30	54	85
China	10	30	65	48	2
Portugal	15	20	22	13	5
Finland		11	12	12	100
Sweden	2	9	9	10	100
Brazil				10	
Argentina			1	7	30
Mexico	9	8	11	6	10
Indonesia				6	0
New Zealand				4	0
India	5	7	8	4	0
Spain	12	9	5	2	90

tries is shown in Table 3 [75], with the proportion of sulfate turpentine given in each case [73]. According to this table the most important producer is still the United States. Although turpentine production has stagnated or decreased in many countries, its importance has increased markedly since ca. 1980 in the People's Republic of China and South America, where significant afforestation is being carried out. In these countries, predominantly gum turpentine is produced, whereas gum and wood turpentine production has become insignificant in the United States [75] (Table 4), the former because of the high extraction costs and the latter because of the dwindling supply of tree stumps.

The price of turpentine (and downstream products) is subject to considerable variation, as can be seen for *American sulfate turpentine* between 1977 and 1993 (midyear prices in dollars per gallon, prices in DM per 100 kilogram in parentheses).

1977	0.51	(31.20)
1979	0.72	(40.30)
1980	1.16	(64.30)
1981	1.74	(120.00)
1982	0.95	(70.40)
1983	0.69	(53.80)
1984	0.64	(55.60)
1985	0.75	(67.30)
1986	0.97	(64.20)
1987	0.83	(45.50)
1988	1.09	(58.50)
1989	2.10	(120.50)
1990	1.90	(93.70)
1991	1.49	(75.50)
1992	1.10	(52.30)
1993	0.70	(35.30)

The price of *gum turpentine* can differ according to origin and quality. It is also subject to large variations that do not, however, inevitably parallel those of sulfate turpentine. The average prices for Chinese gum turpentine (cost insurance freight, cif) may be cited as an example (source: Stat. Bundesamt Wiesbaden):

1989	$ 70.30 per 100 kg
1990	$ 68.40 per 100 kg
1991	$ 76.60 per 100 kg
1992	$ 68.00 per 100 kg
1993	$ 75.90 per 100 kg

Turpentines and their downstream products are expected to be very important in the future since they represent a continually renewable source of raw materials. Their yields can be increased considerably by cultivation of pine species that are richer in turpentine and less susceptible to pests or unfavorable weather conditions; by extending the area under cultivation; and by intensification of resinification using irritants.

14. Toxicology

Oil of turpentine (in vapor form also) causes characteristic and pronounced irritation of the skin and of the mucous membranes of the respiratory organs and eyes. It also leads to sensitization [76]. The LD_{50} values of 1.4 – 5.2 g/kg for rat [77] and the human lethal dose of 60 – 100 g per person [78], which was derived from cases of poisoning, show that the acute oral toxicity is relatively low. In the rat, the *acute inhala-*

Table 4. Proportions of different types of turpentine (given in %) in the United States

Type of turpentine	1950	1960	1970	1980	1990
Gum turpentine	38.4	19.7	4.5	1.0	0.1
Wood turpentine	33.5	26.9	16.4	8.0	2.9
Sulfate turpentine	27.4	53.4	79.1	91.0	97.0
Destructively distilled wood turpentine	0.7				

tion toxicity (LC_{50}) after 1 – 6 h of exposure is 12 – 20 mg/L [77].

Apart from biological scattering, variations can be explained by the varying compositions of different types of turpentines. The LD_{50} values (oral, rat) or turpentine are in the same range as those of its main components [79], [80]:

α-Pinene	2.1 – 3.7 g/kg
β-Pinene	4.7 – 5 g/kg
Camphene, *d*- and *l*-limonene	>5 g/kg
3-Carene	4.8 g/kg

The potential *skin irritation and sensitization* is greater in the case of old resinified turpentine because of the oxidation products formed. Of the main components, 3-carene clearly causes sensitization in humans and animals. Camphene and *l*-limonene do not have a sensitizing effect in humans, and α-pinene shows this effect only after oxidation – it has little or no effect as a purified product. The results for β-pinene and *d*-limonene are contradictory. Here, too, purity plays a part in the sensitizing potency [80]. The irritating effect on the mucous membranes of humans is pronounced: liquid turpentine and its vapors irritate the eyes and the respiratory tract at concentrations of >100 ppm (560 mg/m^3).

The *oral intake* of turpentine can cause irritation of the stomach, intestines, and urogenital tract, as well as vomiting, spasms, nausea, and kidney lesions.

Turpentine oil is easily absorbed through the skin. Although this route of application can cause damage, turpentine was used in the past in human and veterinary medicine to improve blood circulation in the skin.

In industry, however, uptake of turpentine is taken in primarily via the respiratory organs. Inhalation of higher concentrations (>100 ppm) causes not only irritation of the respiratory organs, but also headache, dizziness, increased pulse rate, and in rare cases, reversible impairment of kidney functions. Hardly any information exists on the chronic damaging effects caused by inhalation of low concentrations. Part of the turpentine absorbed parenterally is excreted unchanged by respiration. The largest part is oxidatively metabolized and excreted in urine as glucuronide [81].

The odor of turpentine is perceptible at concentrations as low as ca. 200 ppm (1.12 g/m^3). Since even this concentration causes irritation of the eyes and respiratory tract, it acts as a warning signal.

In the case of the inadvertent ingestion of turpentine, vomiting must be prevented and a doctor called immediately.

15. Turpentine Substitute

Turpentine substitute is frequently used instead of turpentine, mainly for economic reasons. The name is a collective term for colorless, flammable liquids that are similar to turpentine in their volatility and solubilizing capacity for resins, coatings, waxes, etc. The term includes high-boiling naphthas (140 – 170 °C), particularly those with high aromatics contents, as well as hydrogenated hydrocarbons, decalin, and tetralin. Essential oils can be added to improve the odor. *Hydroterpine* (hydrogenated destructively distilled pine oil) and distillation and rearrangement products of turpentine, consisting mainly of monocyclic monoterpenes, are more closely related to turpentine.

Trade names include Depanol I, Dipentene, Solvenol, Terposol, Unitene, and Yarmor [82].

The orange-peel terpenes may also be considered to belong to this group. They consist mainly of *d*-limonene. Those produced industrially are being used increasingly as cleaning agents, e.g., for printed circuit boards in microelectronics (as a replacement for chlorofluorocarbons).

16. Pine Oil

16.1. Production

Originally the name pine oil was given only to the higher-boiling distillation fractions (boiling range >180 °C) obtained in the production of wood turpentine (steam-distilled pine oil, see Section 4.5) or of American pine oil (destructively distilled pine oil) [2], [83].

As a result of the decrease in production of these natural pine oils and the simultaneous increase in demand, *synthetic pine oils* have become more important. They currently account for more than 90 % of sales. Synthetic pine oils are mostly produced by hydration of α-pinene or pinene-rich turpentine fractions by use of dilute mineral acids, usually sulfuric acid, at elevated temperature [84]. After separation of the aqueous phase, they can be purified by vacuum or steam distillation. Higher-grade products are obtained by carrying out the hydration at moderate temperature to give the solid terpin hydrate. The latter is dehydrated in a weakly acidic medium, after separation of the liquid phase, and then distilled.

16.2. Properties

Pine oils are flammable, colorless to yellow, oily liquids with pleasant odors. Their main components are terpene alcohols, particularly α-terpineol.

Natural pine oils can contain the following:

50 – 70 %	α-terpineol
5 – 10 %	borneol (and isoborneol)
5 – 10 %	fenchol
5 – 15 %	other terpene alcohols (e.g., *p*-menthan-8-ol, β-terpineol, terpin)
0 – 10 %	terpinyl and bornyl acetate
0 – 3 %	monoterpenes (e.g., pinenes, dipentene, terpinolene)
5 – 10 %	other components (e.g., cineols, camphor, anethole, estragole, sesquiterpenes)

Synthetic pine oils generally do not contain borneol, camphor, esters, anethole, and sesquiterpenes. The proportions of the components can vary within wide limits:

50 – 95 %	α- and γ-terpineol
3 – 20 %	other terpene alcohols (e.g., β-terpineol, terpinen-1-ol, and fenchol)
0 – 30 %	terpenes (e.g., 3-carene, terpinene, dipentene, and terpinolene)
1 – 10 %	1,4- and 1,8-cineol

Pine oils are characterized according to their terpene alcohol content (e.g., pine oil 70). The terpene alcohol content is normally ca. 60 – 80 % but can be 20 – 95 % after appropriate fractionation.

Pine oils are insoluble in water, but can be emulsified by use of dispersion agents such as soaps. Their *physical properties* lie mostly within the following limits:

ϱ (20 °C)	0.90 – 0.94 g/mL
n_D^{20}	1.475 – 1.488
Distillation range (101.3 kPa)	180 – 225 °C
Flash point	45 – 90 °C

The *chemical properties* of pine oils are determined mainly by the α-terpineol content.

16.3. Storage and Transportation

Pine oils should be stored in closed containers with the exclusion of heat sources. The relevant regulations governing equipment for storage, filling, and transportation of substances hazardous to water and regulations governing flammable liquids (VbF) must be adhered to. The classifications listed in Table 5 are in use, depending on the flash point.

Table 5. Regulations governing storage and transportation of pine oils

Regulation	Flash point	
	Below 55 °C	Above 55 °C
VbF	group A, class II	group A, class III
GGVE/GGVS	class 3, item 31c	class 3, item 32c
RID/ADR/ADNR	class 3, item 31c	class 3, item 32c
IATA – DGR	1987, at flash point >60.5 °C	
	no limitations	
IMDG-Code	class 3.3 – 1987	class 3.4 – 1987
UN-No.	1987	1987
Transport labeling according to regulations governing hazardous substances	30/1987	1987

Pine oil is not cited in the register of existing substanes (EINECS) according to the

ChemG. However, the CAS registry no. and the EINECS no. of its main component terpineol [*8000-41-7*], EINECS no. 232-268-1, may be assigned to it.

16.4. Uses

Pine oils are used mainly as foaming agents in the flotation of copper, lead, and zinc ores [85], [86] and sylvinite [87]. They have also been employed for the separation of poly(vinyl chloride) from mixtures of plastics [88]. Pine oils are good solvents for many paints and can be used as deodorants and disinfectants, as antifoams, and as additives for insecticides and wood preservatives [89]. Furthermore, they are used for the production of lubricants, emulsions, shoe and floor polishes, metal cleaners, and textile auxiliaries [90].

In the chemical industry, pine oil is used for the isolation of α-terpineol, which crystallizes from pine oil on cooling to less than ca. $-2\,^{\circ}C$ (depending on concentration) or is obtained from pine oil distillation.

16.5. Economic Aspects

Besides a few South American and European (French and Austrian) pine oil-producing companies, the main producers are the southern states of the United States with annual sales listed in Table 6 [75]. The decrease in sales can be explained mainly by its reduced use in household disinfectants [65]. The most important pine oil-importing countries are Canada, Brazil, the United Kingdom, and Germany.

Table 6. Annual production and sales of pine oil in the southern United States

Year	Production, t	Export, t
1960	34 200	8 000
1970	48 400	13 000
1980	40 400	16 000
1990	28 500	6 600

Trade names (besides the common term pine oil): Arizole, Dertol, Flotation Terpene-S, and Unipine.

17. References

General References

1. E. Guenther: *The Essential Oils,* vol. VI, D. van Nostrand, New York 1952.
2. E. Gildemeister, F. Hoffmann: *Die ätherischen Öle,* 4th ed., vol. IV, Akademie Verlag, Berlin 1956.
3. W. Sandermann: *Naturharze, Terpentinöl, Tallöl.* Springer Verlag, Berlin 1960.
4. *Kirk-Othmer,* 2nd ed., **vol. 16,** pp. 680 ff; **vol. 20,** pp. 748 ff.

Specific References

5. G. Ourisson, *Pure Appl. Chem.* **62** (1990) no. 7, 1401.
6. R. A. Stepen, T. M. Korsunova, *Sev. Biol. Nauk* **2** (1988) 43.
7. V. A. Isidorov, I. G. Zenkevich, B. V. Joffe, *Atmos. Environ.* **19** (1985) no. 1, 1.
8. R. C. Evans et al., *Bot. Gaz. (Chicago)* **143** (1982) no. 3, 304.
9. K. B. Lamb, H. H. Westberg, T. Quarles, D. L. Flyckt, Washington State Univ. Report 1984, EPA-600/3-84-001, order no. PB84-124981.
10. *Kosmos* **4** (1986) April, 87.
11. M. Levey, *Scientia Assoc.* **91** (1956) 145.
12. P. Rovesti, *Dragoco Rep.* **24** (1977) no. 3, 59.
13. K. Volke, *Chem. Unserer Zeit* **27** (1993) no. 1, 42.
14. E. v. Romaine, *Chem. Ind. (N.Y.)* **45** (1939) 259, 402.
15. E. Gildemeister, F. Hoffmann: *Die ätherischen Öle,* 4th ed., **vol. I,** Akademie Verlag, Berlin 1956, p. 5.
16. W. Sandermann, *Holzforschung* **16** (1962) 65.
17. S. Juvonen, *Acta Bot. Fenn.* **71** (1966) 92.
18. Y. Poltavchenko, G. Rudakov, *Biol. Nauki (Moscow)* **15** (1972) 95. *Chem. Abstr.* **77** (1972) 123 879.
19. R. Croteau, *SPC Soap Perfum. Cosmet.* **53** (1980) no. 8, 428.
20. M. H. Beale, *Nat. Prod. Rep.* **8** (1991) no. 5, 441.
21. R. Croteau, Washington State Univ. Report 1991, DOE/ER/12027-6 order no. DE 92004298.
22. M. Bukala et al., *Przegl. Papier.* **13** (1957) 360.
23. W. B. Critchfield, G. L. Little: *Geographic Distribution of the Pines of the World,* United States Department of Agriculture, Mish. Publ., Washington 1971.
24. A. W. Hodges, *Naval Stores Rev.* **103/4** July/Aug. 1993.

25. G. C. Hoffmann, T. S. Brush, R. L. Farrell, *Naval Stores Rev.* 102/3, May/June 1992.
26. J. M. Conley et al., *Tappi* **60** (1977) 114.
27. O. J. Schwarz, F. J. Ryan, *Proc. Plant Growth Regul. Work Group* **7** (1980) 199.
28. G. C. Harris, *Tappi Monogr. Ser.* **6** (1948) 167.
29. R. G. Fajans: "Hercules Pinex Program," 7th Int. Naval Stores Meeting, Washington 1980.
30. W. P. Lawrence, *Pap. Trade J.* **124** (1947) 128.
31. S. Witek, *Przegl. Papier.* **14** (1958) 105.
32. H. F. J. Wenzl: *Kraft Pulping,* Lockwood Publ. Co., New York 1967.
33. G. S. Landry, C. D. Stillwell, *Naval Stores Rev.* **94** (1984) no. 5, 16.
34. G. A. Smook: *Handbook for Pulp und Paper Technologists,* 2nd ed., Angus Wilde Publ., Vancouver 1992.
35. T. T. Collins, M. G. Schmitt, *Pap. Ind. Pap. World* **26** (1944) 1137.
36. L. E. Larrabec et al., EP 194 286, 1923.
37. Chem. Fabrik vorm. Schering, SE 56 056, 1922.
38. C. Schmidt, DE 340 126, 1919.
39. J. J. Jefischew et al., *Bum. Promst.* **29** (1954) 23.
40. D. Zinkel, J. Russel: *Naval Stores – Production, Utilization,* Pulp. Chem. Assoc. Inc., New York 1989.
41. T. T. Collins, *Pap. Ind. Pap. World* **27** (1945) 537, 719.
42. N. N. Neppenin: *Chemie und Technologie der Zellstoffherstellung,* vol. 1, Akademie-Verlag, Berlin 1960.
43. H. F. J. Wenzl: *Sulphite Pulping Technology,* Lockwood Trade Journal Co., New York 1965.
44. O. V. Ingruber, M. Kocurek, A. Wong: *Pulp and Paper Manufacture,* 3rd ed., Joint Textbook Committee of the Paper and Pulp Industry, Montreal – Atlanta 1989.
45. *Ullmann,* 4th ed., **22,** 555.
46. N. T. Mirov: *The Genus Pinus,* The Ronald Press Co., New York 1967.
47. N. T. Mirov, *Physiol. For. Trees Symp.* 1957, 251 (publ. 1958).
48. J. Weck, *Forstwiss. Centralbl.* **77** (1958) 193.
49. *Miltitzer Berichte über Ätherische Öle, Riechstoffe usw.,* collection of brief reports on essential oils, flavors, and fragrances, published annually until 1986, VEB Chemisches Werk Miltitz, German Democratic Republic.
50. *Ullmann,* 4th ed., **22,** 557.
51. J. P. Vitè et al., *Dragoco Rep.* **27** (1980) no. 11/12, 247.
52. W. Lützkendorf, *Seifen Öle Fette Wachse* **77** (1951) 221.
53. S. A. Ivanov et al., *Seifen Öle Fette Wachse* **97** (1971) 171.
54. E. Hezel, *Dtsch. Farben Z.* **5** (1951) 379.
55. W. C. Smith, *Ind. Eng. Chem. Anal. Ed.* **3** (1931) 345.
56. S. R. Snider, *Ind. Eng. Chem. Anal. Ed.* **17** (1945) 108.
57. K. Miltenberger, G. Keicher, *Farbe + Lack* **69** (1963) 677.
58. R. F. Taylor, B. H. Davies, *J. Chromatogr.* **103** (1975) 327.
59. R. Hiltunen, S. Raisanen, *Planta Med.* **41** (1981) 174.
60. W. Bruhn, *Dragoco Rep.* **34** (1987) no. 6, 167.
61. W. Bruhn, *Dragoco Rep.* **21** (1974) no. 6, 115; **25** (1978) no. 9, 171.
62. J. Pohjala, R. Hiltunen, M. v. Schantz, *Flavour Fragrance J.* **4** (1989) 121.
63. Explosionsschutzrichtlinien der BG Chemie (ZH 1/10; Sept. 1990) und Merkblatt T 033 (4/92) der BG Chemie (Guidelines on explosion protection and memorandum T 033 of BG Chemie).
64. Anlage 4 "Grenzwerte" (1993) UVV (Appendix 4 "limit values" of UVV).
65. S. G. Traynor, Pulp Chemicals Association, Int. Naval Stores Conference, Wien 1992.
66. A. K. Oshkaev, G. V. Stadnitskii, *Izv. Vyssh. Uchebn. Zaved. Lesn. Zh.* 1980, no. 6, 19.
67. M. Pointet et al., *Chem. Ind. (Düsseldorf)* **32** (1980) 690.
68. Int. Flavors & Fragrances Inc., US 3 987 121, 1976.
69. J. M. Derfer, *Parfum. Flavor.* **3** (1978) 45.
70. V. Krishnasamy et al., *Chem. Petro-Chem. J.* **10** (1979) 35.
71. R. E. Close, *Manuf. Chem. Aerosol News* **12** (1980) 45.
72. J. Verghese, *Parfum. Flavor.* **6** (1981) 23.
73. F. A. Dawson, Pulp Chemicals Association, Intern. Naval Stores Conference, Orlando, Fl. 1993.
74. K. Bruns, *Parfüm. Kosmet.* **61** (1980) no. 12, 457.
75. Naval Stores Rev., Int. Yearbooks, published until 1992.
76. U.S. Dept. of Health and Human Services (ed.): Occupational Health Guideline for Turpentine, Washington D.C., Sept. 1978.
77. Hercules Inc., Hercules Terpene Hydrocarbons and Solvents – Summary of Toxicological Investigations, Bull. T. 108, Wilmington, Del., 1962.

78. S. Moeschlin: Klinik und Therapie der Vergiftungen, Thieme, Stuttgart 1980.
79. O. A. Pickett, J. M. Schantz, *Ind. Eng. Chem.* **26** (1934) 709.
80. E. L. J. Opdyke (ed.): Monographs of Fragrance Raw Materials, Pergamon Press, Oxford 1979.
81. G. D. Clayton, F. E. Clayton (eds.): *Patty,* 4th ed., **vol. 2,** 1994.
82. Th. Kunzmann, *Seifen Öle Fette Wachse* **103** (1977) 470.
83. R. W. Clements, D. N. Collins, *Naval Stores Rev.* **59** (1949) no. 40, 17.
84. *Kirk-Othmer,* 2nd ed., **vol. 19,** p. 819.
85. L. Desalbres, *Rev. Ind. Miner.* **38** (1956) 379.
86. A. W. Troizki: *Die Flotation,* Fachbuchverlag, Leipzig 1956, p. 115.
87. Kali-Chemie, DE 917 361, 1954.
88. Mitsui Mining and Smelting Co., Japan Kokai JA 7 532 274, 1973.
89. A. Davidsohn, *Seifen Öle, Fette Wachse* **83** (1957) 15, 43.
90. *Seifen Öle Fette Wachse* **94** (1968) 175.

Type Metal → Lead Alloys
Tyres → Tires
Tyrosine → Amino Acids
Ulcer Therapy → Antiulcer Drugs
Ultrafiltration → Membranes and Membrane Separation Processes
Ultramarine Pigments → Pigments, Inorganic

Uranium, Uranium Alloys, and Uranium Compounds

MARTIN PEEHS, Siemens AG, Energieerzeugung (KWU), Erlangen, Germany
THOMAS WALTER, PreussenElektra Aktiengesellschaft, Hannover, Germany
SABINE WALTER, Universität Hannover, Institut für Werkstoffkunde, Hannover, Germany
MARTIN ZEMEK, AREVA NP GmbH, Erlangen, Germany

1. Introduction 2
2. History 3
3. Physical Properties 4
3.1. Radioactivity 4
3.2. Modifications 4
3.3. Mechanical Properties 4
3.4. Thermal Properties 6
3.5. Electrical and Electrochemical Properties 8
3.6. Magnetic Properties 8
4. Chemical Properties 8
5. Occurrence, Requirement, and Production Figures 9
5.1. Occurrence 9
5.2. Resources, Requirement, and Production Figures 10
6. Production 14
6.1. Uses of Uranium and Uranium Compounds 15
6.2. From Ore to End Product – A Review of Processes 15
6.2.1. From Crude Ore to Yellow Cake . 15
6.2.2. From Yellow Cake to UF_6 17
6.2.3. From UF_6 to the Nuclear Fuel UO_2 17
6.3. Detailed Description of the Processes 17
6.3.1. Digestion and Leaching of Ores . 17
6.3.1.1. Acidic Ores 18
6.3.1.2. Alkaline Ores 18
6.3.1.3. Digestion of Phosphate Rock . . . 18
6.3.2. Treatment of the Liquor 19
6.3.2.1. Uranium Recovery by Ion Exchange 19
6.3.2.2. Uranium Recovery by Solvent Extraction 20
6.3.2.3. Eluex Process 21
6.3.3. Production of Uranium Ore Concentrate 21
6.3.3.1. From Precipitation to Yellow Cake Production 21
6.3.3.2. Processing of Phosphate Liquor and Precipitation of AUC 21
6.3.4. Final Purification of Uranium Concentrate 21
6.3.4.1. Dissolution of Yellow Cake 21
6.3.4.2. Extractive Purification 22
6.3.5. Production of UO_3 and UO_2 from Purified Uranyl Nitrate Solution . 23
6.3.5.1. Evaporation of Uranyl Nitrate Solution and Denitration by Thermal Decomposition 24
6.3.5.2. Precipitation of Uranium by the ADU and AUC Processes 26
6.3.5.3. Reduction of Precipitated Product to UO_2 Powder 27
6.3.6. Production of UF_4 27
6.3.6.1. Hydrofluorination of UO_2 28
6.3.6.2. Hydration and Hydrofluorination of UO_3 28
6.3.7. Production of UF_6 from UF_4 . . . 29
6.3.7.1. Process Description 29
6.3.7.2. Chemical Reactors 30
6.3.7.3. Removal of Excess Fluorine from UF_6 30
6.3.8. Complete Plant for Production of UF_6 from Uranyl Nitrate 30
6.3.8.1. French Process in Pierrelatte . . . 30
6.3.8.2. Allied Chemical Process 31
6.3.9. Enrichment of ^{235}U 32
6.3.9.1. Diffusion Process 32
6.3.9.2. Ultracentrifuges 33
6.3.9.3. Nozzle Process 34
6.3.9.4. Chemical Enrichment 34
6.3.9.5. Laser Separation Process 35
6.3.9.6. Plasma Processes 36
6.3.10. Production of UO_2 Pellets from UF_6 36
6.3.10.1. Conversion of UF_6 to UO_2 36
6.3.10.2. Possible Future Developments in the Conversion of UF_6 to UO_2 Powder 38
6.3.10.3. Pelletizing of UO_2 Powder 41
6.3.11. Production of Uranium Metal . . . 41
6.3.11.1. Reduction of UF_6 to UF_4 41

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_281.pub2

6.3.11.2. Reduction of UF_4 to Uranium Metal . . . 41
6.3.11.3. Production of Uranium Powder . . 43
7. Uranium Alloys . . . 43
7.1. Classification . . . 43
7.2. Production of Important Alloys 44
8. Uranium Compounds . . . 46
8.1. Halides . . . 46
8.1.1. Trivalent Halides . . . 46
8.1.2. Uranium Tetrahalides . . . 46
8.1.3. Uranium Pentafluoride . . . 46
8.1.4. Uranium Hexafluoride . . . 47
8.2. Carbides . . . 47
8.3. Nitrides . . . 48
8.4. Oxides . . . 49
8.4.1. Uranium Dioxide . . . 49
8.4.2. Uranium Trioxide . . . 49
8.4.3. Triuranium Octaoxide . . . 49
8.4.4. Peroxides . . . 50
8.5. Nitrates, Sulfates, and the Carbonato Complex . . . 51
8.5.1. Nitrates . . . 51
8.5.2. Sulfates . . . 51
8.5.3. Tricarbonatodioxouranate . . . 51
9. Safety . . . 51
9.1. Radiation Shielding . . . 51
9.2. Safety Against Uncontrolled Criticality . . . 52
9.3. Geometrically Safe Vessels . . . 52
9.4. Apparatus with Heterogeneous Neutron Absorbers . . . 53
9.5. Neutron Interactions in the UO_2 Fabrication Plant . . . 53
9.6. Transportation . . . 53
10. References . . . 54

Abbreviations:

ADU	ammonium diuranate
AUC	ammonium uranyl carbonate
AVLIS	Atomic Vapor Laser Isotope Separation
BNFL	British Nuclear Fuels Limited
DC	Direct Conversion
GECO	General Electric Dry Conversion
DEHPA	2-ethylhexyl phosphate
IDR	Integrated Dry Route
LWR	light-water reactor
MLIS	Molecular Laser Isotope Separation
MTR	materials testing reactor
RIP	resin in pulp
TBP	tributyl phosphate
TOPO	trioctylphosphine oxide
USW	uranium separative work
UTK	uranium tetrafluoride in kilns

1. Introduction

Uranium, atomic number 92, is a radioactive element that is transformed into the stable element lead via many intermediate stages involving the emission of α- and β-radiation. Uranium has the probable electronic configurations $4s^2p^6d^{10}f^{14}$, $5s^2p^6d^{10}$, $6s^2p^6d^4$, and $7s^2$, and is a member of the actinides, in which the $6d$ shell is filled until uranium is reached, and the $5f$ shell is then filled from neptunium onward to $5f^{14}$. Natural uranium consists of three isotopes whose concentrations in different resources are usually the same, even to the fourth decimal: ^{238}U, 99.2762 at %; ^{235}U, 0.7182 at %; ^{234}U, 0.0056 at %. However, uranium from one deposit in Gabon (the Oklo mine) is exceptional [8], the proportion of ^{235}U being lower (see below [9]):

Expected ^{235}U content	0.711 wt %
Measured ^{235}U content (July 7, 1972)	0.7078 wt %
Absolute discrepancy	0.0032 wt %
Relative discrepancy	0.45 %
Accuracy of measurement	±0.0001 wt %

The somewhat lower percentage of ^{235}U in the Oklo deposit led to the discovery of natural reactors and the revelation that the first nuclear fission chain reaction was not brought about by humans in 1942 but by nature long before this. Since its discovery, the Oklo phenomenon has been investigated in great detail, and the results are documented in almost 100 papers. At least ten different reactor zones have been identified in Oklo. The individual reactors were active for 100 000 to 800 000 years. These periods of activity, which occurred 1.8×10^9 years ago, resulted in the release of 100×10^9 kW · h thermal energy, equaling the output of a large-capacity power reactor over a three-year period. These natural reactors are the only ones discovered thus far in the world. During these periods of activity, natural uranium still had a ^{235}U content of ca. 3 %; geochemical processes caused

the U content to increase to 30 % and higher. The thickness of the natural deposits reached 50 cm, and the presence of water as a moderator and coolant caused the configuration to become critical. The nuclear fission that then occurred was controlled by the properties of water. High power output led to evaporation of water, with consequent reduction of the moderating effect; lower power led to condensation and increased moderation. Decay products from earlier fission products and from the plutonium formed in the active zones of the natural Oklo reactor are still present without losses. This shows that these elements have not been dispersed but have been immobilized by simple geological barriers.

2. History

Uranium was discovered in 1789 by Martin Heinrich Klaproth, who extracted it from pitchblende. He precipitated the yellow uranate and reduced this with charcoal, but obtained only a black oxide, probably UO_2. In 1790, he named the putative metal uranium after the planet Uranus, which had been discovered in 1781 by Herschel. In 1841, E. M. Peligot reported that he had obtained uranium metal by thermal reduction of anhydrous uranium tetrachloride with potassium in a platinum crucible. In 1856, he succeeded in producing the pure metal in compact form by reducing uranium tetrachloride with sodium under inert conditions. A worth-mentioning quantity of uranium metal was put on display at the Paris World Exhibition of 1867.

In the 1890s, H. Moissan developed new processes for producing metallic uranium. He obtained the purest metal (>99 % uranium) by electrolysis of sodium uranyl chloride. At the end of the 19th century, uranium found some almost unnoticed industrial applications, e.g., in glass and ceramics because of the beautiful colors of its different compounds.

The discovery of radioactivity by A. H. Becquerel on March 1, 1896, resulted from experiments with uranium. The fluorescence and phosphorescence of uranium compounds had been studied from about the mid-1800s. Following the discovery of X-rays, the question arose of whether substances showing strong fluorescence or phosphorescence emitted penetrating rays of a similar kind. Becquerel exposed potassium uranyl sulfate to sunlight and observed that it emitted rays that darkened photographic plates. An interruption of these investigations due to cloud cover led to the discovery of intense darkening of the plates by uranium salts that had not been exposed to light. He concluded from this that uranium salts produced a special kind of radiation independent of light exposure.

The history of uranium is closely linked to the discovery of polonium and radium in pitchblende by P. and M. Curie in 1898. The induced fission of the isotope ^{235}U was discovered in 1938 by O. Hahn and F. Strassmann. This discovery was followed by intense research activity, that led to the first man-made nuclear chain reaction, which was brought about in Chicago by E. Fermi on December 2, 1942. World War II and U.S. decision to have a nuclear weapon led to intense activity in uranium research and nuclear development. Uranium enrichment plant based on the electromagnetic process was started in 1943 in Oak Ridge, Tennessee, followed a year later by another plant based on the gas diffusion principle.

After World War II, the search for uranium ores for the peaceful utilization of nuclear energy was intensified in most countries, and some large deposits were discovered. In the 1960s, military applications predominated; the peaceful use of nuclear energy gradually gained in importance in the late 1970s. Enriched uranium was required for the light-water reactors favored for energy production; thus, enrichment processes and techniques to manufacture fuel elements became important research and development topics. Enrichment by gas diffusion processes was used in the United States and France, whereas Germany, The Netherlands, and the United Kingdom developed a cooperative program based on three enrichment plants using the ultracentrifuge principle. The jet nozzle separation process, developed in Germany, is used in a demonstration plant in Brazil and in the Republic of South Africa.

Literature searches carried out for the present article have revealed that research and development of nuclear energy production and related technologies were largely complete in Western societies by the early 1980s. However, research continued into the technology of enrichment (laser and chemical processes), along with

fundamental research into the band structure of uranium.

3. Physical Properties

3.1. Radioactivity

The *natural uranium isotopes* (^{238}U, ^{235}U, ^{234}U) are α-emitters. Tables 1 and 2 show the decay chains of the two most important isotopes. These chains terminate with the nonradioactive isotopes ^{206}Pb and ^{207}Pb.

Artificial uranium isotopes are formed in a nuclear reactor (e.g., by irradiation or decay of transuranics). Thus, after reprocessing of spent fuel elements, the following uranium isotopes are found in the purified uranium for a uranium burnup of 36 GW · d/t and an initial enrichment in the fuel elements of 3.5 wt % (For a definition of burnup, see → Nuclear Technology):

^{232}U: concentration 0.71 mg per tonne U; half-life 74 a; → ^{228}Th
^{233}U: concentration 7.4 mg per tonne U; half-life 1.6×10^5 a; → ^{229}Th
^{234}U: concentration 137 g per tonne U; half-life 2.5×10^5 a; → ^{230}Th
^{236}U: concentration 4490 g per tonne U; half-life 2.4×10^7 a; → ^{232}Th

All four artificial isotopes are α-emitters with a certain amount of γ-emission and are transformed into the thorium isotopes indicated. Because of the comparatively short half-life of ^{232}U, associated γ-emission requires heavily shielded equipment for radiation protection. This hot-cell technology is also needed to handle the radioactive fission products in the reprocessing of spent fuel. Because it has an odd atomic number, ^{233}U is fissile by capture of a thermal neutron when it is reused in the reactor. However, ^{236}U in particular is purely a neutron absorber, so that on use of the reprocessed uranium, the fissile fraction (^{235}U) must be increased by ca. 0.15 % to compensate for neutron capture by the 0.45 % ^{236}U.

3.2. Modifications

The phase diagram of uranium is shown in Figure 1 [10, 11]. At atmospheric pressure, uranium occurs in three modifications with different crystal structures, depending on temperature [12]:

α-Uranium, $T < 942$ K, orthorhombic
Lattice constants: $a = 0.2858$ nm, $b = 0.5876$ nm, $c = 0.4955$ nm at 298.15 K.
β-Uranium, $T = 942 - 1049$ K, tetragonal
Lattice constants: $a = 1.0759$ nm, $c = 0.5656$ nm at 942 K (lower transformation point).
γ-Uranium, $T = 1049 - 1408$ K, body-centered cubic
Lattice constants: $a = 0.3524$ nm at 1049 K (lower transformation point).

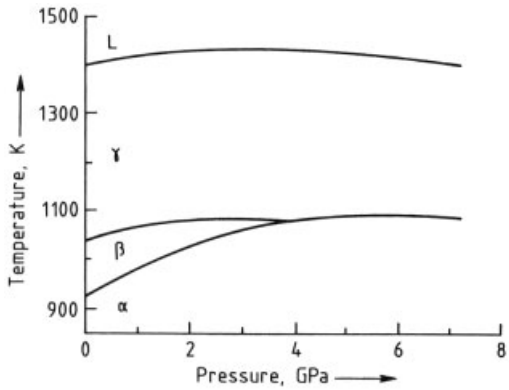

Figure 1. Phase diagram of uranium [10, 11]
Ranges of the α-, β-, and γ-modifications and of the liquid are shown

The triple point of the three modifications is at 3.6 ± 0.2 GPa and 1078 ± 3 K [10]. Figures quoted for the *mp* at atmospheric pressure cover the relatively narrow range between 1398 K (1125 °C) [13] and 1408 K (1135 °C) [14]. These small discrepancies are due to extremely low levels of impurities that cannot be avoided completely. The *bp* is said to be 4407 K. The theoretical density of α-uranium is 19.13 g/cm^3, small variations in the experimental value again being due to impurities.

3.3. Mechanical Properties

In the γ-phase (1049 K $< T <$ 1408 K), the metal is soft, ductile, and easily worked. However, since this phase is readily oxidized in an oxidizing environment (e.g., air), uranium usually

Table 1. Radioactive decay of ^{238}U

	Atomic number	Relative atomic mass	Penetration of α-particles in air at 15°C and 0.1 MPa, cm	Half-life	Amount in radioactive equilibrium with ^{238}U contained in 1000 kg uranium, g
^{238}U	92	238.05	2.63	4.515×10^9 a	9.93×10^5
α ↓					
^{234}Th	90	234		24.10 d	1.426×10^{-5}
β ↓					
^{234}Pa	91	234		1.16 min	4.76×10^{-10}
β ↓ 99.85% ; γ 0.15% → ^{234}Pa Isomeric transition	91	234		6.7 h	2.5×10^{-10}
β γ → ^{234}U	92	234	3.22	2.32×10^5 a	50.1
α ↓					
^{230}Th	90	230	3.13	8.3×10^4 a	17.6
α γ ↓					
^{226}Ra	88	226.03	3.30	1590 a	3.32×10^{-1}
α γ ↓					
^{222}Rn	86	222	4.06	3.825 d	2.147×10^{-6}
α ↓					
^{218}Po	84	218	4.67	3.05 min	1.17×10^{-9}
α 99.96% ↓ ; β 0.031% → ^{218}At					
^{214}Pb	82	214		26.8 min	1.01×10^{-8}
β γ ↓					
^{218}At	85	218	5.53	a few seconds	ca. 1×10^{-14}
^{214}Bi ← α 99.9% ; β 0.1% → ^{218}Rn	83	214	4.1	19.7 min	7.4×10^{-9}
β γ 99.96% ↓ ; α 0.04% → ^{210}Tl					
^{218}Rn α → ^{214}Po	86	218	6.1	0.019 s	1×10^{-18}
^{214}Po	84	214	6.95	1.45×10^{-4} s	9.1×10^{-16}
α ↓					
^{210}Tl	81	210		1.32 min	1.95×10^{-13}
β γ → ^{210}Pb					
^{210}Pb	82	210		22.3 a	4.32×10^{-3}
β γ ↓					
^{210}Bi	83	210		4.95 d	2.63×10^{-6}
β γ ↓					
^{210}Po	84	210	3.85	138.8 d	7.37×10^{-5}
α γ ↓ ; α 5×10^{-5}% → ^{206}Tl					
^{206}Tl	81	206		4.23 min	7.6×10^{-16}
β → ^{206}Pb	82	205.97		stable	∞

must be worked while in the α-phase at lower temperature (at 563 K).

Mechanical strength data for the α- and β-phases are given in Table 3 [15].

The hardness of the metal depends strongly on the annealing temperature (Table 4). The Rockwell hardness was measured at applied weights of 100 and 150 kg, and the Brinell hard-

Table 2. Radioactive decay of ^{235}U

	Atomic number	Relative atomic mass	Penetration of α-particles in air at 15°C and 0.1 MPa, cm	Half-life	Amount in radioactive equilibrium with ^{235}U contained in 1000 kg uranium, g
^{235}U	92	235.04	2.7	7.07×10^{8} a	7.06×10^{3}
α ↓ ^{231}Th	90	231		24.6 h	2.75×10^{-8}
β ↓ ^{231}Pa	91	231.04	3.62	3.2×10^{4} a	3.14×10^{-1}
α ↓ ^{227}Ac	89	227		21.7 a	2.09×10^{-4}
β 98.8% ↓ ^{227}Th	90	227	4.67	18.6 d	4.85×10^{-7}
α 1.2% ↓ ^{223}Fr	87	223		21 min	4.5×10^{-12}
α γ ↓ ^{223}Ra ← 99.996%	88	223	4.33	11.3 d	2.93×10^{-7}
α 0.004% ↓ ^{219}At	85	219	4.97	0.9 min	8×10^{-18}
α ↓ ^{219}Rn ← β ≈ 3%	86	219	5.75	3.92 s	1.16×10^{-12}
α ≈ 97% ↓ ^{215}Bi	83	215		8 min	6×10^{-17}
α ↓ ^{215}Po ← β	84	215	6.49	0.00183 s	5.30×10^{-16}
α 99.9995% ↓ ^{211}Pb	82	211		36.1 min	6.15×10^{-10}
β 0.0005% ↓ ^{215}At	85	215	8.0	ca. 10^{-4} s	ca. 1×10^{-22}
β γ ↓ ^{211}Bi ← α	83	211	5.46	2.16 min	3.68×10^{-11}
β 0.32% ↓ ^{211}Po	84	211	6.59	ca. 0.005 s	4.5×10^{-18}
α 99.68% ↓ ^{207}Tl	81	207		4.76 min	7.93×10^{-11}
β γ ↓ ^{207}Pb ←	82	206.98		stable	∞

ness is given in Brinell numbers (applied weight of 300 kg divided by the penetration area of a 10-mm steel sphere).

The modulus of elasticity is also very dependent on the history of the material; it ranges between 155 GPa (α-phase, rolled; β-phase, annealed) and 207 GPa (γ-phase, extruded).

3.4. Thermal Properties

The specific and molar heat capacities of the three uranium modifications are given in Table 5 [12].

Table 3. Mechanical strength of rolled uranium metal products annealed in the α- or β-region before measurement [15]

	Test temperature, °C	0.2 % Yield strength, MPa	Ultimate tensile strength, MPa	Elongation, %
Rolled at 570 °C				
α-annealed	20	298.6	766.0	6.8
β-annealed	20	169.3	439.9	8.5
Rolled at 1110 °C				
α-annealed	20	179.3	609.5	13.5
β-annealed	20	169.6	425.4	6.0
Rolled at 570 °C, α-annealed	299	120.7	242.7	49.0
Rolled at 570 °C, α-annealed	299	129.6	217.7	43.0
Rolled at 570 °C, α-annealed	499	35.2	76.7	61.0
Rolled at 1110 °C, α-annealed	499	37.6	72.0	57.0

Table 4. Hardness of annealed uranium metal

Annealed for 1 h at	Rockwell hardness		Brinell hardness number
	150-kg weight	100-kg weight	
310 °C	36.6	109	341
490 °C	24.0	101	247
600 °C	22.0	99	240
720 °C	9.0	90	185
770 °C	0	92	152

Table 5. Specific and molar heat capacities of uranium modifications

	c_p, J kg^{-1} K^{-1}	C_p, J mol^{-1} K^{-1}
α-Uranium (298.15 K)	116	27.666
β-Uranium (942 K)	180	42.928
γ-Uranium (1049 K)	161	38.284

In the α- and β-phases, uranium is anisotropic; i.e., thermal expansion depends on direction. Thus, a crystal of α-uranium expands in only two directions and contracts in the third (Fig. 2). The expansion coefficients for α-uranium in the various crystal directions are given in Figure 3 [16, 17].

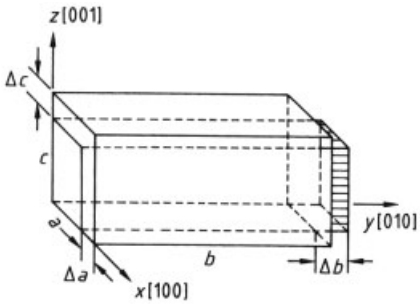

Figure 2. Dimensional change of α-uranium from 0 to 300 °C.
Original dimensions: $a = 0.2852$ nm; $b = 0.5985$ nm; $c = 0.4945$ nm; $\Delta a = +0.0015$ nm; $\Delta b = -0.0006$ nm; $\Delta c = +0.0339$ nm

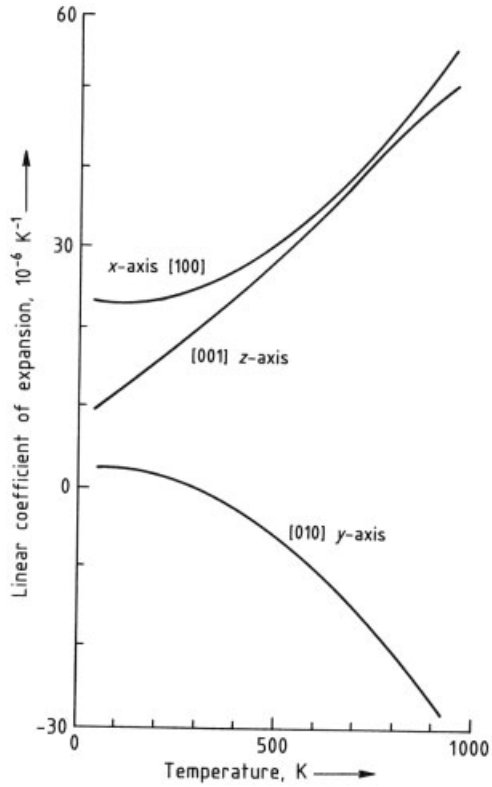

Figure 3. Linear expansion coefficient of α-uranium [16]

For β-uranium in the range 959 – 998 K, the linear coefficients of expansion are

z-axis, 4.6×10^{-6} K^{-1}
x-axis, 23.0×10^{-6} K^{-1}
y-axis, 16.9×10^{-6} K^{-1}

The linear coefficient of expansion of γ-uranium is the same for all three axes (i.e., ca. 18×10^{-6} K^{-1}). The thermal conductivity of uranium at 298 K is 27.60 W m^{-1} K^{-1}.

3.5. Electrical and Electrochemical Properties [18]

The specific electrical conductivity of uranium has been determined by several authors. The variations in the value obtained [(3.0 – 3.38)$\times10^4$ Ω^{-1} cm^{-1}] are due to variations in the purity and history of the metal. The values determined for specific electrical resistance naturally show similar variation, the mean value at room temperature being ca. 30 μΩ cm. The specific electrical resistance increases with increasing temperature to ca. 55 μΩ cm at 873 K. At the transformation points of α- to β-uranium and of β- to γ-uranium, the curve shows a downward discontinuity but then rises again. The temperature coefficient of the specific electrical resistance in the range 273 – 373 K is (2.61 – 2.76)$\times10^{-3}$.

Uranium becomes superconducting at very low temperature, but the transition temperature T_c is very pressure dependent [10]. At atmospheric pressure, T_c has its lowest value (0.2 K); it reaches its highest value (2.25 K) at 1 GPa.

The standard electrode potential ϵ_0 for the reaction

$$2\,H^+ + 2\,e^- \rightleftharpoons H_2$$

$$2\,U^{3+} \rightleftharpoons 2\,U^{4+} + 2\,e^-$$

at 25 °C with a 1 mol/L solution of metal ions and a hydrogen electrode is with water −0.414 V, with a 1 mol/L $HClO_4$ solution −0.631 V, and with a 1 mol/L HCl solution −0.640 V. The thermoelectric potential with copper is ca. 5 mV [19].

3.6. Magnetic Properties

Uranium is paramagnetic. The paramagnetic susceptibility varies with temperature and exhibits discontinuities at the phase transformation points (Fig. 4). The specific magnetic susceptibility at 298 K is 1.66×10^{-6} cm^3/g.

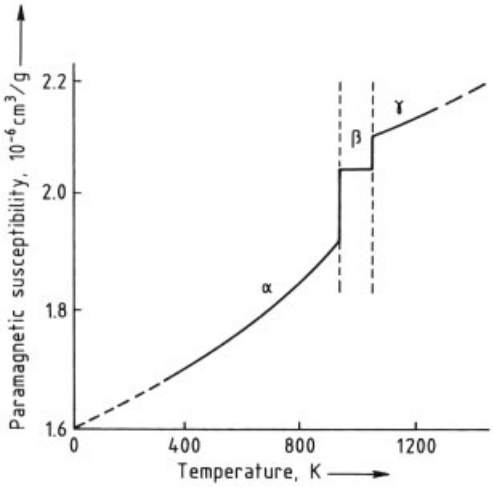

Figure 4. Paramagnetic susceptibility of uranium

4. Chemical Properties

The freshly exposed surface of massive uranium metal has a silvery luster, but it is oxidized relatively rapidly in air. Within a few hours, the metal becomes covered with a thin oxide layer of iridescent colors and eventually turns black. The same mass of metal that is oxidized by steam at 593 K is oxidized by air only above 753 K. The enthalpy and the free energies of formation of oxides from the metal at 298 K are as follows [20]:

UO_2
$\Delta H = -1084$ kJ/mol
$\Delta G = -1029$ kJ/mol
U_3O_8
$\Delta H = -3565$ kJ/mol
$\Delta G = -3360$ kJ/mol
UO_3
$\Delta H = -1218$ kJ/mol
$\Delta G = -1138$ kJ/mol

Uranium metal powder is pyrophoric and burns or smolders very readily in air, sometimes spontaneously if the powder has a large enough surface area. Mechanical machining of uranium metal by turning, lathing, and drilling must be carried out with adequate work fluids for lubrication of the cutting tools and for cooling the chippings formed. To store the chippings, they are wetted with a special oil or burned promptly in a controlled manner to form U_3O_8. Uranium does not react with *hydrogen* at room temperature but combines readily with it at 498 K to UH_3, which decomposes again at 709 K. This behavior is utilized in the manufacture of the powdered metal from its compact form, in which UH_3 is produced as an intermediate. Also, to reduce the hazard involved in transporting tritium gas, it is absorbed by depleted uranium powder in flasks.

Uranium reacts very vigorously with *halogens* and *hydrogen halides*. It burns in fluorine gas to form UF_6 ("flame reactor"). The reaction with HF gas proceeds only to the UF_4 stage; reaction with aqueous HF leads to hydrated UF_4 ($UF_4 \cdot 0.75\ H_2O$, $UF_4 \cdot 2.5\ H_2O$). Uranium burns in a stream of chlorine to form UCl_6 and UCl_4. With bromine and iodine, it reacts to only UBr_3 and UI_3. With *sulfur*, grayish black US_2 is formed, which is stable up to 1273 K.

Powdered uranium reacts very vigorously with *water*; above 308 K the compact metal also reacts with water. Uranium is attacked vigorously by *dilute acids*, with the formation of salts. *Strong solutions of alkali* also attack the metal to form uranates.

5. Occurrence, Requirement, and Production Figures

5.1. Occurrence

The mean concentration of uranium in the earth's crust is 2×10^{-4} wt %. For comparison, the abundances of cobalt, lead, and molybdenum are similar. Uranium is a typical lithophilic element, occurring mainly in the silicates of the earth's crust. Most of the uranium is distributed widely in small concentrations, both in acidic rocks such as granite, gneiss, and pegmatite, and in sedimentary rock such as schist and phosphatic ores. Basic rocks (basalt) have a lower uranium content than, for example, granite. The wide distribution of uranium and the low degrees of enrichment in ores compared with other metals of similar abundance are the result of its ability to form a wide range of chemical complexes. Uranium occurs in more than 100 minerals; those of industrial importance are listed in Table 6.

Table 6. Uranium minerals of industrial importance *

Mineral	Chemical formula	Origin
Oxides, hydrates, simple silicates (uranium, completely or primarily tetravalent)		
Pitchblende	$x\ UO_3 \cdot y\ UO_2$	Canada, Congo, Western and Central Europe
Uraninite	UO_2 (with UO_3)	Canada, Colorado Plateau, Argentina
Coffinite	$x\ USiO_4 \cdot y\ U(OH)_4$	Colorado Plateau
Uranothorite	$x\ ThSiO_4 \cdot y\ USiO_4$	Canada, Madagascar
Complex deposits of uranium oxides with rare earths (titanates, etc.)		
Brannerite	$(U, Ca, Fe, Y, Th)_3Ti_5O_{16}$	Canada
Betafite (similar to ampangabeite)	$(U, Ca)(Nb, Ta, Ti)_3O_9 \cdot x\ H_2O$	Madagascar
Davidite	$(Fe, \text{rare earths}, U, Ca, Zr, Th)_x (Ti, Fe^{3+}, V, Cr, Y)_yO_z$	
Secondary minerals with hexavalent uranium		
Gummite	$UO_3 \cdot x\ H_2O$ (also silicate, phosphate, of indefinite composition)	
Uranophane	$CaO \cdot 2\ UO_3 \cdot 3\ SiO_2 \cdot 6\ H_2O$	Colorado Plateau, Argentina
Schroeckingerite	$Na_2O \cdot 3\ CaO \cdot UO_3 \cdot 3\ CO_2 \cdot SO_3 \cdot F \cdot 10\ H_2O$	Joachimsthal, Colorado Plateau, Argentina
Carnotite	$K_2O \cdot 2\ UO_3 \cdot V_2O_5 \cdot 3\ H_2O$	Colorado Plateau, Australia, Argentina
Tjujamunite	$CaO \cdot 2\ UO_3 \cdot V_2O_5 \cdot 8\ H_2O$	Turkestan, Colorado Plateau, Argentina
Autunite (torbernite)	$CaO \cdot 2\ UO_3 \cdot P_2O_5 \cdot 12\ H_2O$	France, Portugal, Colorado Plateau, Argentina

* Detailed information on uranium minerals is given in [21].

Some materials not regarded as true uranium minerals are the poorly defined uranium-containing bituminous products known as kolm (Sweden), thucholite (Sweden, Canada, Republic of South Africa), and asphalt or asphalite (Colorado Plateau, Argentina). An industrially important mineral which contains uraninite or pitchblende as accompanying mineral is mon-

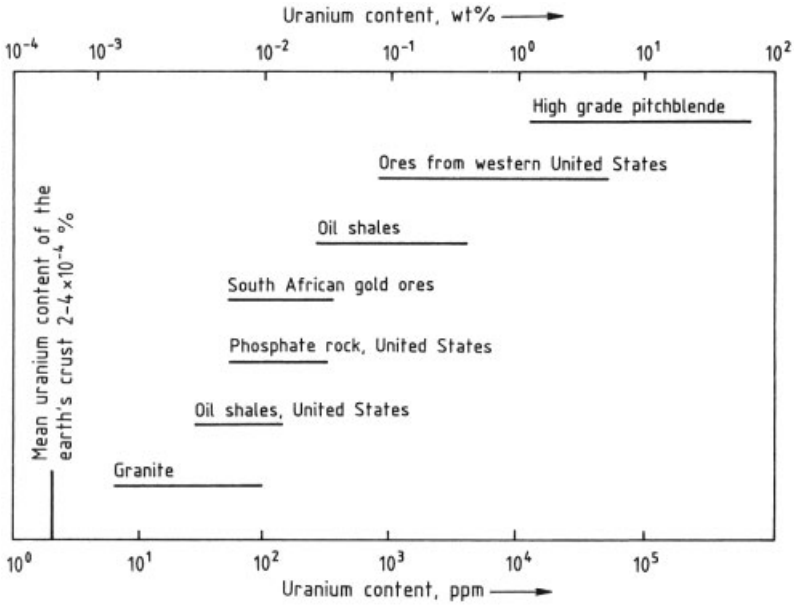

Figure 5. Uranium concentration in important uranium minerals

azite. Large deposits of monazite are found in India and Brazil.

Large reserves are also present in seawater because of its enormous volume. However, the uranium level (ca. 3 ppb = 3 μg/L) is very low. Nevertheless, attempts have repeatedly been made since the 1950s to extract uranium from seawater. An enrichment factor of 10^5 has been achieved. However, the processes will be economical only if enrichment can be increased to ca. 10^6; the concentration would then correspond to that in a minable ore [22]. Uranium concentrations in the most important uranium-bearing ores in the earth's crust are shown in Figure 5 [23]. Approximately half of the annual uranium production comes from secondary minerals. One-third comes from conglomerates in which unweathered rock has been sedimented together with oxidized uranium. The remaining uranium production is from veins or lodes and syngenetic – sedimentary ore deposits.

5.2. Resources, Requirement, and Production Figures

A 1991 report with data valid for January 1, 1991, issued jointly by the Nuclear Energy Agency of the OECD and the IAEA, gives data on deposits, production, and uranium requirements for 44 countries [24]. This report, for the first time, includes countries from the former Eastern Bloc. Since then, the data from the former Eastern Bloc have become more precisely known and have been more thoroughly evaluated. The development of uranium supply and demand has also changed considerably. Thus considerable quantities of uranium from military projects are now available for civil nuclear energy generation as a result of worldwide nuclear disarmament. Hence OECD/NEA [25] and The Uranium Institute [26] have revised and re-issued their existing reports. The market perspectives have been studied with respect to how the ever-increasing gap between uranium demand and uranium production can only be bridged by the release of military reserves [27 – 29].

The known world uranium resources, classified according to the IAEA system by the state of exploration and anticipated costs per kilogram of uranium extracted, are listed in Table 7. Known reserves amount to 3.125×10^6 t, a figure that has been confirmed by more recent surveys of the Uranium Institute [26].

The reserves of the following countries are not included in the IAEA classification:

Table 7. World uranium reserves according to IAEA classification [24]; in 1000 t uranium *

Country	Reasonably assured resources		Estimated additional resources (category I)	
	Up to $ 80/kg U	$ 80 – 130/kg U	Up to $ 80/kg U	$ 80 – 130/kg U
North America (total)	247.90	323.90	149.00	80.70
Canada	146.00	68.00	149.00	80.00
Mexico	0.00	1.70	0.00	0.70
United States	101.90	254.20	0.00	0.00
Africa (total)	547.02	138.70	380.57	75.50
Algeria	26.00	0.00		
Central African Republic	8.00	8.00		
Gabon	11.00	4.65	1.30	8.30
Namibia	84.75	16.00	30.00	23.00
Niger	166.07	6.65	295.77	10.00
Somalia	0.00	6.60	0.00	3.40
South Africa	247.60	96.80	51.80	30.80
Zaire	1.80	0.00	1.70	0.00
Zimbabwe	1.80	0.00		
Europe (total)	58.27	74.25	23.05	51.35
Austria			0.70	1.00
Denmark	0.00	27.00	0.00	16.00
Finland	0.00	1.50		
France	23.80	15.70	4.20	3.90
Germany	0.60	4.00	1.60	5.70
Greece	0.30	0.00	6.00	0.00
Hungary	1.62	1.50	9.10	9.15
Italy	4.80	0.00	0.00	1.30
Portugal	7.30	1.40	1.45	0.00
Spain	17.85	21.15	0.00	9.00
Sweden	2.00	2.00	1.00	5.30
Australia	469.00	60.00	264.00	126.00
South America (total)	172.53	2.19	96.26	2.09
Argentina	8.74	2.19	0.54	1.95
Brazil	162.00	0.00	94.00	0.00
Peru	1.79	0.00	1.72	0.14
Asia (total)	4.32	27.53	0.00	3.20
Indonesia	4.32	0.00		
Japan	0.00	6.60		
Korea	0.00	11.80	0.00	3.00
Turkey	0.00	9.13		
Vietnam			0.00	0.20
Total	1449.04	626.57	912.88	338.84

* For a definition of resource categories, see → Nuclear Technology.

Chile	0.23×10^3 t
China	55.10×10^3 t
India	66.15×10^3 t
Romania	26.00×10^3 t
Former Soviet Union	685.60×10^3 t
Bulgaria	45.00×10^3 t
Former Czechoslovakia	145.00×10^3 t
Poland	20.00×10^3 t
Total	1038.98×10^3 t

On the basis of reports on estimated and speculative reserves, it is assumed that worldwide reserves amount to $(8 - 13) \times 10^6$ t of uranium. uranium.

Uranium Production of the Western Countries. Figure 6 shows the uranium production and uranium demand of the Western world in the period 1956 to 1995. In this part of the world, the uranium industry experienced its first boom in the early 1960s. At that time, almost the entire uranium production was consumed by the weapons programmes of the western nuclear powers (United States, United Kingdom, France). The civil uranium demand for nuclear power stations began in the late 1960s and early 1970s. This was stimulated by the expansion programmes that were established as a result of the oil crisis of 1973. However, the actual expansion in Europe and the United States failed to meet the planned level. Nevertheless, between 1975 and 1980, uranium production grew to meet the increased demand from civil nuclear energy generation. At this time, production of

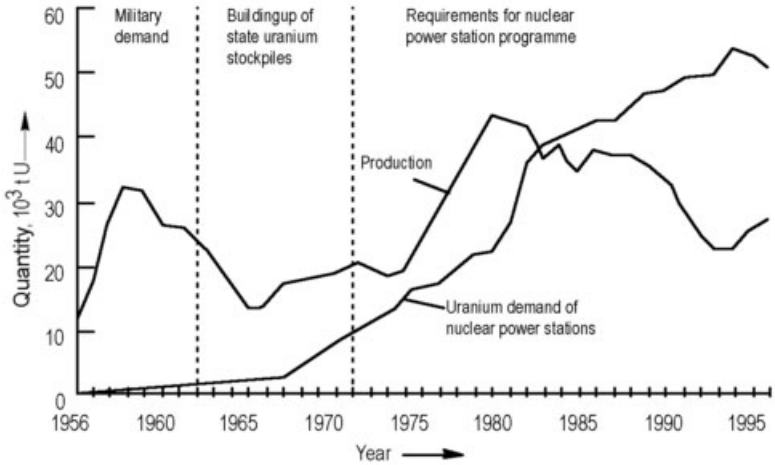

Figure 6. Uranium production and demand of the Western world (1956 – 1996)

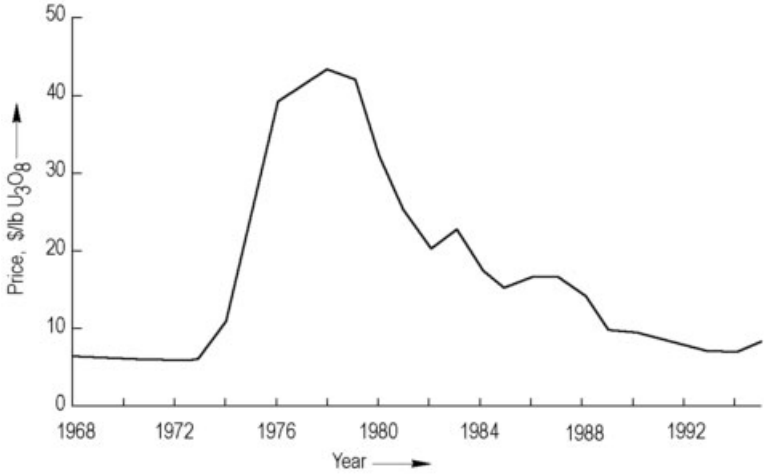

Figure 7. Uranium spot market price for European buyers (1968 – 1995)

uranium as a byproduct of phosphate-rock mining became important.

The dramatic decrease in production since 1989 is due to increasing oversupply. Thus, in 1980 the uranium production of the Western world was ca. 44 000 t U, but demand was only ca. 22 000 t U. While uranium production continued to decrease, the uranium demand of the civil nuclear energy programmes grew worldwide. The difference between supply and demand was covered by uranium from stockpiles. The price development for uranium shows clearly that sufficient quantities of previously produced uranium are reaching the market, since the uranium price has dropped (Fig. 7) despite the fact that production has failed to meet demand. Up to 1973, the uranium price remained at a constant 7 $/lb U but then increased greatly due to the oil crisis, reaching a peak of 45 $/lb U in 1976. The overcapacity existing at this time led to a price drop and eventual stabilization at ca. 23 $/lb in 1982. For the reasons discussed above, the uranium price fell slowly up to 1994 and then reached temporary stability at a low level.

Uranium Production of the Former Eastern Bloc. In the former Soviet Union, its satellite states, and China, ca. 630 000 t of uranium have been produced since the 1940s (Fig. 8).

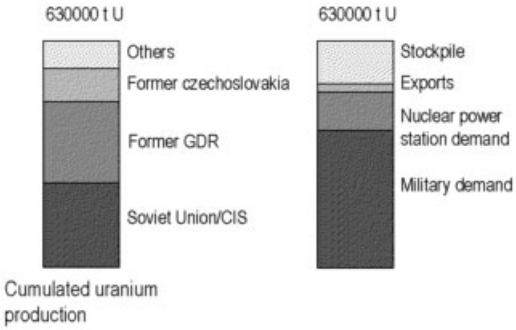

Figure 8. Cumulated uranium production in the former Eastern Bloc and use of the material (1945 – 1994)

Prior to the political and economic liberation of the Eastern Bloc in 1989/90, the majority of the uranium produced in the Soviet sphere of influence was transported to the Soviet Union as a strategic material. Only Romania retained a few thousand tonnes for its CANDU reactor progamme. Of the uranium produced in the Soviet Union and imported from its satellites up to 1994, probably 450 000 t U was made available for military purposes, and 115 000 t U was used for the nuclear reactor programmes of the Soviet Union and Eastern Europe. Thus, today Russia should possess a stockpile of ca. 75 000 t U in the form of natural and enriched uranium. Figure 9 shows the uranium production of the CIS since 1991. Two of the largest producers, Kazakhstan and Uzbekistan, have essentially no nuclear energy programme, so that their entire production should become available to the world market in the form of exports.

Figure 9. Uranium production in the CIS (1991 – 1996)

Uranium Production Today. World uranium production in 1995 was ca. 34 200 t U, 75 % of which came from the Western countries and 25 % from Eastern Europe, the CIS, and China (Fig. 10). The 13 largest mines accounted for 70 % of world uranium production. The three largest Western uranium producers – Cogemá, Cameco, and Uranerzbergbau – were responsible for ca. 44 % of world uranium production. Figure 11 shows 1994 uranium production divided according to mine owners.

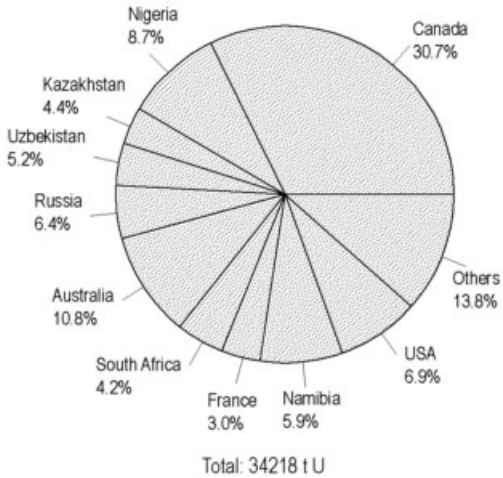

Figure 10. World uranium production in 1995 by countries

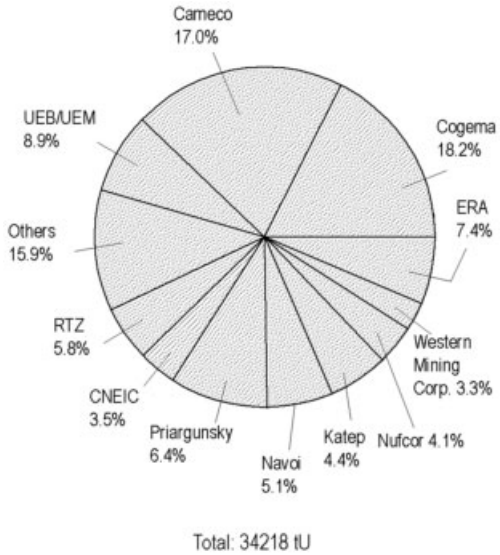

Figure 11. World uranium production in 1995, broken down by owners of mines

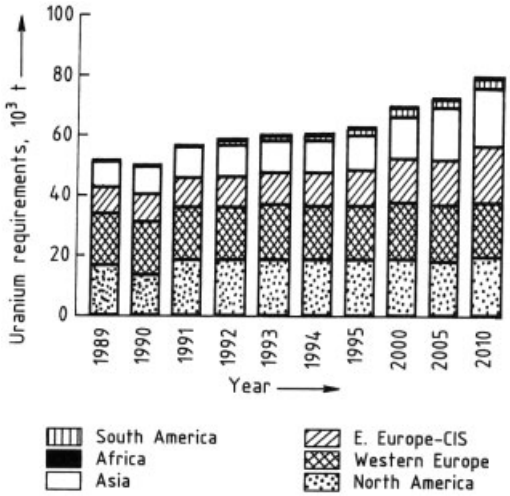

Figure 12. Projected uranium consumption for nuclear reactors [24]

Future Supply/Demand Pattern. Figure 6 shows uranium demand and uranium production up to 1996. Figure 12 shows an OECD/IAEA estimate of future uranium demand dating from 1991. World uranium demand in 1991 was 49 300 t U, and this was distributed between the continents as shown in Figure 12. On the basis of planned nuclear energy generating capacity, a uranium demand of ca. 77 700 t U is expected in the year 2010. New investigations of various reference scenarios of The Uranium Institute [26] give values between 70 000 and 80 000 t U. Schreiber [28] assumes a demand of less than 60 000 t U and predicts the uranium supply situation shown in Figure 13.

World uranium demand will increase slowly for about a decade, after which a downward trend can be expected if it is assumed that nuclear power stations are removed from the grid and not replaced when their licenses expire. This assumption is based on a projection of current energy policies. Figure 13 compares world uranium demand with maximum achievable production capacity. Production capacity is divided into potential projects, mines held in reserve, and plants in operation or for which concrete plans exist. Figure 13 shows that world uranium production could double by 2002, but also that the operating, planned, and potential projects will not be sufficient to cover the uranium demand of the nuclear power stations. Hence, in future the existing civil uranium inventories and those that become available from military programmes as a result of disarmament will have to be used to meet these demands.

6. Production

The production of uranium and its most important compounds is described in this chapter. Sec-

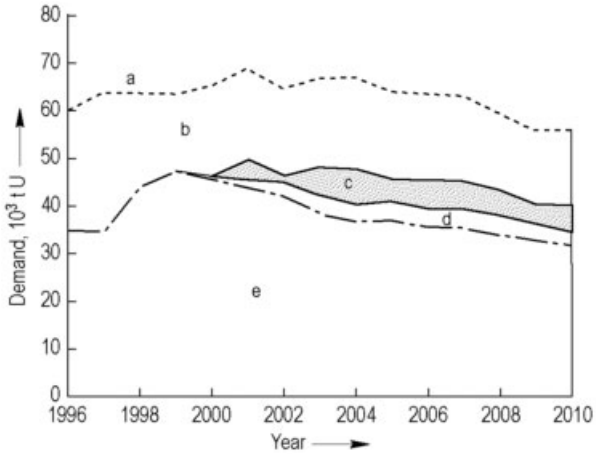

Figure 13. Contributions of the production of fresh uranium and uranium inventories to future uranium supply (1996 – 2010) a) Uranium demand; b) Stockpiles, recyclable weapon Uranium and revenues from reprocessing; c) Potential future Uranium mines, d) Mines held in reserve; e) Mines in operation or planned to be operated

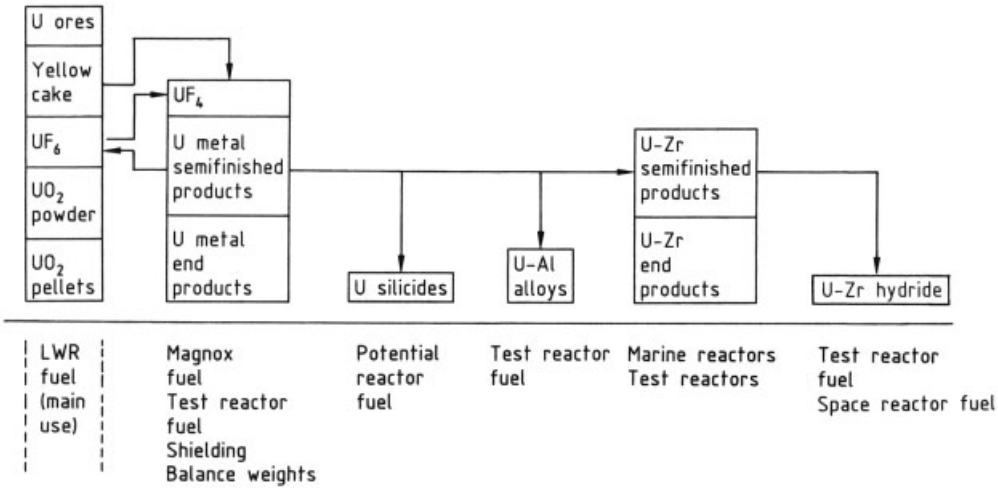

Figure 14. Uranium starting materials, intermediates, end products, and their uses

tion 6.1 is devoted to a review of starting materials, intermediates, and the most important uranium end products. Section 6.2 provides an overview of the most important processes used to convert crude ore to the end product. Section 6.3 contains detailed descriptions of all the processes mentioned in Section 6.2.

6.1. Uses of Uranium and Uranium Compounds

Uranium starting materials, typical intermediates, and industrially important end products are indicated in Figure 14. In all cases, the intermediate product "yellow cake" is produced from uranium ore. The industrially most important end product is UO_2, which is the only nuclear fuel used in power station reactors, although *uranium carbide* (UC) is sometimes used in sodium-cooled fast breeder reactors because of its higher density and superior thermal conductivity compared with UO_2.

Uranium metal is used as the fuel in materials testing reactors and Magnox reactors. Because of its extremely high density, it is also used in depleted form for trimming weights in large aircraft. A Boeing jumbo jet contains 360 kg of depleted uranium metal. The high density of uranium metal also utilized in shielding materials for protection against radiation sources (e.g., in portable equipment for weld radiography and medical X-ray equipment).

U – Zr alloys are used as nuclear fuel in marine reactors and in research and materials testing reactors. *Hydrogenated U – Zr alloys* ($UZrH_x$) are used as combined fuel – moderator systems in spacecraft reactors. *Uranium silicides* could in principle be used as fuel in light-water reactors. However, uranium silicides are not utilized because undesired reactions with water, the cooling medium, can occur as a result of defects in the fuel elements. *Uranium – aluminum alloys* are used in materials testing and research reactors.

Uranium salts are now rarely added to ceramic products to produce luminous colors (e.g., in Merano glassware or tiles) because of the radiation problems involved.

6.2. From Ore to End Product – A Review of Processes

6.2.1. From Crude Ore to Yellow Cake

The industrial routes from crude ore to yellow cake are illustrated in Figure 15. Mined crude ores are prepared for leaching by crushing and grinding. The chemical leaching process is selected according to the mineralogical nature of the ore. Acidic ores are treated with dilute H_2SO_4, and alkaline ores with an aqueous solution containing sodium carbonate and

Process step	Acidic ores	Alkaline ores	Phosphate rock
Classification	Screening		
Crushing, grinding	Mechanical crushing and grinding		
Leaching	Acid leaching with dilute H_2SO_4	Alkaline leaching with Na_2CO_3 and $(NH_4)_2CO_3$	Acid treatment
Solid–liquid separation	Filtration, multistage decantation, separation by hydrocyclones		
U extraction	Ion exchange, solvent extraction, Eluex process		Oxidation–reduction stripping process
Precipitation	By ammonia or magnesium hydroxide	By 50% sodium hydroxide	By $(NH_4)_2CO_3$ to form ammonium uranyl carbonate (AUC)
	Yellow cake ← Yellow cake	Yellow cake	AUC

Figure 15. From crude ore to yellow cake or ammonium uranyl carbonate (AUC)

Yellow cake

Process step	Wet process			Dry process
Pretreatment	Dissolution in HNO_3			Size reduction and preliminary purification
Wet purification	Solvent extraction			
Further treatment	ADU precipitation	AUC precipitation	Evaporation, denitration to UO_3	
Filtration	Yes	Yes		
Calcination	ADU → UO_2	AUC → UO_2	UO_3 → UO_2	Reduction to UO_2
Conversion to UF_4	$UO_2 + 4HF \rightarrow UF_4 + 2H_2O$			
Conversion to UF_6 Dry purification	$UF_4 + F_2 \rightarrow UF_6$			Fractional distillation of UF_6
UF_6				UF_6

Figure 16. From yellow cake to UF_6

sodium bicarbonate. Phosphatic ores are treated with acid.

Solid – liquid separation of leached ore is achieved by standard processes such as filtration, multistage decantation, and hydrocyclone separation; uranium is recovered from the solution by ion exchange or solvent extraction. Combining these two techniques, as in the Eluex process, is also advantageous. Yellow cake is then produced by precipitation from the acid liquor with ammonia or $Mg(OH)_2$ or by precipitation from the alkaline liquor with NaOH.

Uranium is obtained as an accompanying element from phosphatic ores in which the uranium

content is very low (see Section 5.2). After acid treatment, uranium is purified and recovered by the stripping process. Uranium is precipitated as ammonium uranyl carbonate (AUC).

6.2.2. From Yellow Cake to UF_6

In the wet treatment process (Fig. 16), yellow cake is dissolved in HNO_3 and purified by solvent extraction. The resulting solution of uranium in nitric acid can then be reacted chemically to form UO_2 or UO_3 by using either the ammonium diuranate (ADU) or the AUC process, or by denitrating evaporation.

UO_2 is converted into UF_6 in two stages: first UF_4 is produced by treatment with HF; UF_4 is then converted into UF_6 by treatment with fluorine gas. In principle, yellow cake can also be processed in the dry state. After size reduction and preliminary purification, yellow cake is calcined to form UO_2, which is then converted into UF_6 as described above. However, the dry UF_6 so obtained must then be purified by fractional distillation.

6.2.3. From UF_6 to the Nuclear Fuel UO_2

If UF_6 (Fig. 17) is to be converted to the nuclear fuel UO_2 for light-water reactors, an enrichment process must be carried out for reasons of reactor physics until the ^{235}U isotope content is between 2 and 6 %. Enrichment processes commonly used include the diffusion process, the nozzle process, and centrifugation.

Enriched UF_6 can be processed by either wet methods (e.g., AUC, ADU) or dry methods [e.g., Integrated Dry Route (IDR), Direct Conversion (DC), General Electric Dry Conversion (GECO)] to obtain UO_2 powder. After chemical conversion and before pelletization, the powder must be pretreated (except in the case of precipitation by the AUC process). Only after pretreatment do the various steps of pelletization (compression, sintering, and grinding) give an end product with the desired spectrum of properties.

6.3. Detailed Description of the Processes

6.3.1. Digestion and Leaching of Ores

Before ore digestion appropriate pretreatment is essential. This includes size reduction for which several stages are necessary. In the first stage, lumps of ore ca. 1 m in diameter are passed through gyratory or jaw crushers with a power consumption of 250 – 350 kW. These reduce the ore to lumps of ca. 200 mm diameter. Cone

UF_6

Process step	Process description				
^{235}U enrichment	Diffusion, centrifugation, nozzle separation, laser process				
Conversion to UO_2 powder	Wet processes		Dry processes		
	AUC	ADU	IDR	DC	GECO
Powder preparation	None	Milling, compaction, granulation			
Pressing of green pellets	Mechanical pressing, usually with rotary pelleting machine				
Sintering	Reductive sintering in H_2 atmosphere, also pilot stage oxidative-reductive process				
Surface grinding	Centerless grinding to final dimensions				

UO_2 pellets

Figure 17. From UF_6 (natural) to UO_2 pellets for light-water reactors
AUC = Ammonium uranyl carbonate; ADU = Ammonium diuranate; IDR = Integrated Dry Route; DC = Direct Conversion; GECO = General Electric Dry Conversion Process

crushers with a power consumption of 30 – 40 kW are used in the second stage, giving pieces of ore ca. 65 mm in diameter. Particle sizes of <10 mm are obtained in the third stage, in which fine cone crushers or roller crushers with power consumption of 15 – 20 kW are used [30, 31]. This degree of size reduction is adequate with most types of ore for which the much more usual sulfuric acid process is used. Finer grinding is necessary only for alkali treatment. Here screen mills and ball mills are used in succession to reduce the particle size of the ore to ca. 0.5 mm.

Consideration of the economics of the complete process of grinding and chemical treatment must include the question of whether preliminary removal of worthless material (gangue) might be beneficial. However, except for visual hand sorting of lump materials, no industrial process has thus far been used. Pilot-scale experiments in Elliot Lake, Canada, have been reported (Table 8) [32].

Table 8. Separation methods: amounts of worthless material removed and consequent uranium loss [32]

Method	Worthless material removed, wt %	U_3O_8 loss, wt %
Flotation after grinding to <0.3 mm	62 – 80	8
Gravitational method	ca. 35	8
Radiometry method	ca. 20	8
Magnetic separation	ca. 85	14

6.3.1.1. Acidic Ores

Dilute sulfuric acid is always used for the acid treatment of ores. The rate of dissolution depends on acid concentration, temperature, and surface area of the ore particles. If the ore contains feldspar and clay, highly concentrated acid should not be used since it will dissolve aluminum silicates. If the particles are ground too finely, clarification of the cloudy solution can present some problems. If tetravalent uranium is present, an oxidizing agent (sodium chlorate or manganese dioxide) is added to the liquor, with dissolved iron acting as a catalyst. The following reactions occur during the dissolution process:

$$UO_3 + 2\,H^+ \rightarrow UO_2^{2+} + H_2O$$

$$UO_2 + 2\,Fe^{3+} \rightarrow UO_2^{2+} + 2\,Fe^{2+}$$

$$UO_2^{2+} + SO_4^{2-} \rightarrow UO_2SO_4$$

At higher acid concentration, complex anions such as $[UO_2(SO_4)_3]^{4-}$ are also formed, but these do not cause problems in later processing.

6.3.1.2. Alkaline Ores

Alkaline ores require treatment with alkaline solutions. Alkaline leaching is considerably slower than acid leaching, but is more effective for ores in which the gangue materials contain calcium compounds or other acid-consuming components. For alkaline treatment the ore must be more finely ground. Since carbonate solution does not attack this type of gangue, uranium is dissolved much more selectively than when acidic solutions are used, and the subsequent concentration stage is easier. The dissolution of uranium is due to the formation of a tricarbonato complex:

$$UO_3 + Na_2CO_3 + 2\,NaHCO_3 \rightarrow Na_4\left[UO_2(CO_3)_3\right] + H_2O$$

The above ratio of carbonate to hydrogencarbonate should be maintained. If tetravalent uranium is present, oxygen is used as an oxidizing agent, and higher temperature and pressure are applied. This oxidation is catalyzed by copper sulfate and ammonia.

6.3.1.3. Digestion of Phosphate Rock

Uranium is present in phosphate rock at a mean concentration of 100 – 150 mg/kg. Florida phosphate rock, for example, contains 110 mg/kg uranium. Total uranium reserves in phosphate rock are currently estimated at ca. 7.1×10^6 t [24], which is nearly twice as much as the world reserves with beneficiation costs of <$ 130 per kilogram uranium.

Most phosphate rock minerals are members of the apatite group, having the typical composition $Ca_{10}(PO_4,CO_3)_6(OH,F,Cl)_2$, in which calcium is replaced to a small extent by U^{4+}, which is of similar size (U^{4+}, 0.097 nm; Ca^{2+}, 0.099 nm). The presence of uranium is due to the formation of phosphate deposits from seawater, which contains 3 µg/L uranium. In the presence of organic reducing agents or Fe(II) ions, U(VI)

is reduced to U(IV), which is incorporated into the apatite lattice.

When apatite is treated with dilute sulfuric acid at 80 – 135 °C to form crude phosphoric acid, uranium also goes into solution. However, if apatite is treated by the dry method or with concentrated sulfuric acid to form superphosphate, uranium is not dissolved but remains in the product. In Israel (only), uranium is first dissolved with hydrochloric acid when using the dry method. Two processes for extracting uranium from phosphoric acid solution have been put into practice, oxidation stripping and reduction stripping.

6.3.2. Treatment of the Liquor

Uranium can be removed from the liquor by ion exchange, solvent extraction, or a combination of the two processes known as the Eluex process. Undissolved solids, which would have a detrimental effect, are usually removed first from the liquor by sedimentation or decantation, or with hydrocyclones, filters, or centrifuges. The best process is vacuum filtration using plane, band, or drum filters, because the solution should be diluted as little as possible by washing of the filter cake.

Whereas slight cloudiness is tolerable for the ion-exchange process, the liquor must be clarified as much as possible before solvent extraction is carried out, because the expensive extraction agent can be lost by adsorption by the solids, which can lead to environmental pollution.

6.3.2.1. Uranium Recovery by Ion Exchange

Anionic complexes of uranium are present in both acid and alkaline uranium solutions. Strongly basic ion exchangers are generally used to extract these complexes:

$$4\,\mathrm{RX} + \left[\mathrm{UO_2(SO_4)_3}\right]^{4-} \rightleftharpoons \mathrm{R_4}\left[\mathrm{UO_2(SO_4)_3}\right] + 4\,\mathrm{X^-}$$

$$4\,\mathrm{RX} + \left[\mathrm{UO_2(CO_3)_3}\right]^{4-} \rightleftharpoons \mathrm{R_4}\left[\mathrm{UO_2(Co_3)_3}\right] + 4\,\mathrm{X^-}$$

where R denotes the matrix and X the functional basic group of the ion exchanger.

The pH is adjusted to 1.5 – 2 in sulfuric acid solutions and to 9 – 10 in alkaline solutions for the ion-exchange process, which can be operated batchwise (fixed-bed process) or continuously.

Fixed-Bed Process. In the fixed-bed process, cylindrical columns of diameter 2 – 3 m, height 4 – 5 m, containing 10 – 30 m^3 resin are used. The ion-exchange resin is loaded at a flow rate of 5 – 20 m^3/h and eluted at 1 – 5 m^3/h. The resin pebble bed usually consists of spherical particles of diameter 0.3 – 0.8 mm, but 2-mm-diameter granules may be needed to prevent blockage of the columns if solutions are incompletely clarified. For such solutions, the resin in pulp (RIP) process was developed, in which large resin granules are held in sieve baskets that are moved to and fro in a horizontal flow of unclarified liquor.

Elution can be carried out with 1 mol/L solutions of chlorides and nitrates. However, these anions interfere with subsequent loading of the resin, so elution with sulfates is often preferred. This proceeds more slowly but does not affect loading of the resin and is more cost-effective. The resin is first eluted with a 1 mol/L sulfate solution and subsequently treated with 10 % sulfuric acid or, in the case of alkaline uranium solutions, with 1 – 2 mol/L sodium carbonate – sodium hydrogencarbonate solution.

Continuous Ion-Exchange Processes. Several types of columns have been developed for continuous ion exchange. The first is exemplified by the *Higgins column*, which has been operated successfully in Texas and Wyoming since 1955 [33]. Resin is loaded as it is pushed slowly upward through the column. It is washed at the overflow at the top and then passed to elution columns. Similar columns have been used successfully by the Bureau of Mines in South Dakota. In 1977, the *Himsley column* was installed in Ontario, Canada, and the *Cloete-streat column* in South Africa. In the *Porter column*, which was installed in Namibia in 1978, uranium solution flows upward and overflows from the top into a second ion-exchange column at a slightly lower level. The laden resin is drawn off from the bottom by means of compressed air and fed via an intermediate tank into elution columns.

Although ion-exchange resins have been improved considerably, the trend now is toward solvent extraction or a combination of ion exchange and solvent extraction (Eluex process) because these processes give better selectivity and, hence, improved product purity.

6.3.2.2. Uranium Recovery by Solvent Extraction

Two types of extraction solvents are used: the first includes alkylated phosphoric acids and pyrophosphoric acids; the second, higher aliphatic amines. Both types are dissolved in inert hydrocarbons, mostly high-purity kerosene. The mechanism of extraction is based on ion exchange in the liquid phase. Tributyl phosphate (TBP), which is used in the production of high-purity uranium, cannot be used for liquid – liquid extraction of sulfuric acid liquors because neutral extraction media are not suitable for uranyl sulfate solutions.

Solvent Extraction with Phosphoric Acids. The following three phosphoric acids have been found suitable for extraction of uranium:

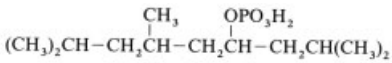

Isododecyl phosphate

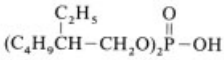

Di(2-ethylhexyl) phosphoric acid (DEHPA), bis(2-ethylhexyl) phosphate [*298-07-7*]

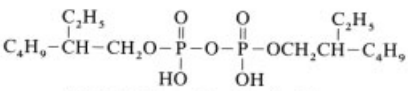

Diiosooctylpyrophosphoric acid

In the extraction process, acidic hydrogen atoms of phosphoric acids are replaced by the uranyl ion. The exchange therefore also depends on the pH of the solution. The complexes formed are soluble in the organic phase. In contrast to extraction with neutral organophosphorus compounds, the anion of the uranyl salt is not extracted at the same time:

$$UO_2^{2+}\,(aq) + 2(RO)_2PO_2H\,(org) \rightarrow UO_2\left[PO_2(OR)_2\right]_2\,(org) + 2\,H^+\,(aq)$$

Extraction can be much improved by adding tributyl phosphate to the phosphoric acid (synergistic effect). Concentrated mineral acids or sodium carbonate solutions are used to strip uranium from the organic solvent.

Solvent Extraction with Amines. Newer uranium processing plants use mainly amines as extraction solvents. These are usually commercially available mixtures of either highly branched dialkylamines or straight-chain trialkylamines (see also → Thorium and Thorium Compounds, Chap. 4.2). Solutions of these amines in hydrocarbons have a strong affinity for uranyl sulfate and form addition complexes of the type $\left(R_3NH^+\right)_2\left[UO_2(SO_4)_2^{2-}\right]$ that are soluble in organic solvents. Uranium can be stripped from these solutions by simple ion exchange with solutions of the nitrate or chloride of sodium or ammonium. Sodium carbonate solutions, which are sufficiently basic to convert the amine salt into free base, can also be used.

Technique of Solvent Extraction. Only a few column types are suitable for solvent extraction; these include the rotating disk column and the *Scheibel column*. However, the most suitable type is the mixer – settler in its various forms. A continuous mixer – settler consists of a mixing chamber and a settling chamber. The mixing chamber is provided with an agitator to mix the two phases (aqueous and organic). The organic phase is fed by gravity from the preceding settling chamber into the inlet at the top of the mixing chamber, while the aqueous phase from the next settling chamber enters the mixing chamber from below. The mixed phases pass through an outlet at an intermediate height into the settling chamber in which the two phases separate. Thus, the phases flow countercurrently, and each mixer – settler constitutes one stage of the exchange process. Mixer – settlers are used for both the extraction and the stripping stages.

Four-stage mixer – settlers of various designs are used for industrial-scale solvent extraction (→ Liquid–Liquid Extraction). For example, in the Davy Power Gas mixer – settler, the mixed phases flow out centrally in upward direction. The Israel Mining mixer – settler contains lamellar internals to promote rapid phase separation [34]; these enable throughput rates to be improved by a factor of 20. The Lurgi Multi-Tray settler is aimed at a similar result [35]. Optimization of design and operation is described in [36].

6.3.2.3. Eluex Process

In the combined Eluex process, uranium is crudely separated by means of an ion-exchange resin, followed by solvent extraction [37]. The ion-exchange resin is designed so as to tolerate as much "cloudiness" as possible, and uranium is collected almost quantitatively. Breakthrough of foreign ions is acceptable since these are removed easily in the second (solvent extraction) stage. In this combined process, the first stage has the useful effect of increasing the concentration of uranium, with consequent reduction in the mass flow. This enables the second stage to be smaller by a factor of 20 – 30, and also improves the purification effect of this step as the uranium concentration in the feed is higher. An additional advantage is that the process can be used with relatively low uranium concentrations in the leach liquor (low-grade uranium ore). Thus, for example, the ion-exchange stage increases the concentration of a starting solution containing 0.02 – 0.5 g/L U_3O_8 to 9 – 12 g/L U_3O_8 and reduces the mass flow by the same factor. Four industrial plants processing 1000 – 1200 t/d ore are operating in the United States [38]. In the Republic of South Africa, the technique was developed in 1963 under the name "BuMex Process", and full-scale operation followed.

6.3.3. Production of Uranium Ore Concentrate

6.3.3.1. From Precipitation to Yellow Cake Production

Solutions obtained by the above processes contain uranium in the form of its sulfate or the carbonato complex $Na_4[UO_2(CO_3)_3]$. Uranium is precipitated from this as a uranate by addition of base, filtered, and dried. The uranium concentrate obtained, known as "yellow cake" because of its color and form, is the basic material for the production of nuclear fuels. Commercial uranium concentrates of various origins are listed in Table 9. Precipitation is carried out in large, agitated vessels. The precipitating agent for acidic, sulfate-containing solutions is ammonia or $Mg(OH)_2$, whereas carbonato complexes are precipitated by a 50 % solution of NaOH.

6.3.3.2. Processing of Phosphate Liquor and Precipitation of AUC

Oxidation Stripping Process. In oxidation stripping, U(VI) is first converted into U(IV) by adding Fe(II) salts to the phosphoric acid; uranium is then extracted from the orthophosphoric acid by mono- and dioctylphenyl esters mixed with kerosene. Stripping is by oxidative treatment with phosphoric acid and sodium chlorate ($NaClO_3$). The U(VI) is again extracted from the stripping solution at 40 – 50 °C, this time by trioctylphosphine oxide (TOPO) and 2-ethylhexyl phosphate (DEHPA) dissolved in kerosene; it is then reductively stripped from this extract by addition of Fe(II). Multiple recycling of the stripping solution gives an enrichment factor of 40. The stripping solution is then reoxidized by $NaClO_3$ and extracted with DEHPA – TOPO – kerosene. Phosphoric acid is washed out of the extract, which is then stripped with $(NH_4)_2CO_3$ solution. An excess of ammonium carbonate is added in the cold to the complex salt solution obtained, precipitating ammonium uranyl carbonate, $(NH_4)_4UO_2(CO_3)_3$. This is filtered off, washed, and calcined to give UO_{2+x}.

Reduction Stripping Process. Reduction stripping proceeds in an almost identical manner, except that only one extraction medium (DEHPA – TOPO – kerosene) is used. Uranium from the first stage is thus not reduced but extracted as U(VI). Both stripping stages are therefore reductive. Uranium does not require further purification in these cases and is usually purer than normal yellow cake.

6.3.4. Final Purification of Uranium Concentrate

6.3.4.1. Dissolution of Yellow Cake

Yellow cake produced by mining companies is not a single chemical substance, but a mixture of uranates. Its composition depends on the mined ores, as mentioned above, and on the chemistry of the treatment process suitable for those ores. The uranium content and the nature and concentration of impurities vary. However, the manufacture of nuclear fuel requires a product of high

Table 9. Specifications for uranium concentrate (yellow cake), contents in wt %

	British Nuclear Fuel (UK)	Comurhex (France)	Eldorado (Canada)	Allied Chemical (United States)	Kerr McGee (United States)
U	40	60	50	75	60
Moisture	10	5	5.0	2.0	
Insolubles in HNO_3		0.1	0.1		0.1
Granulation, mm	6.68	6	6.25	6.25	6.25
Th	n.a. **	n.a. **	2.0		2.0
Mo	<0.6	0.45	0.15 *	0.1	0.15 *
V_2O_5	<1.00	1.80	V = 0.1	0.1 *	0.1 *
Ca		1.15	1.0	0.05	1.0
P_2O_5	<6	1.1	P = 0.35	PO_4 = 0.1	P = 0.35
Cl, Br, I	<0.5	0.25	0.25 *	0.05	0.25
F		0.15	0.15 *	0.01	0.15
SO_4		10.50	S = 3.5	S = 3.00	S = 3.5
Fe				0.15	
As	<2.00	1.0	1.0	0.05	1.00
CO_3	<2.00	2.0	2.00	0.20	2.00
B	<0.2	0.15	0.15	0.005	0.15
Na		n.a. **		0.50 *	
K		n.a. **		0.1	
SiO_2	<4.00				
Zr					2.00

* Maximum concentration without cost penalties.
** Not available.

purity and constant composition, and the fine purification processes are intended to achieve this aim.

On an industrial scale, wet processes are used mainly to treat yellow cake, which must therefore first be brought into solution. This is achieved with hot nitric acid, which dissolves the uranium as uranyl nitrate. Impurities settle out to some extent or may also be dissolved as nitrates.

6.3.4.2. Extractive Purification

In the extractive stage of purification, uranium is separated from the accompanying impurities as far as possible. Removal of elements with a high cross section for thermal neutrons (e.g., boron, rare earths, cadmium, and lithium) is especially important.

Purification is carried out by dissolving the uranium concentrate in nitric acid and extracting uranium with selective solvents. TBP diluted with hydrocarbons is used in most purification plants. A complex is formed according to

$$UO_2^{2+}\,(aq) + 2\,NO_3^-\,(aq) + 2\,TBP\,(org) \rightleftharpoons UO_2(NO_3)_2{\cdot}2\,TBP\,(org)$$

The complex is neutral, anhydrous, and undissociated, and has a higher solubility in the organic than in the aqueous phase. The uranyl nitrate solutions contain an excess of free nitric acid (sometimes nitrates of alkali metal, alkaline-earth metal, or aluminum are added), which has a salting-out effect and improves the distribution coefficient between the organic and inorganic phases. The tributyl phosphate solution is then treated with pure water, and uranyl nitrate is transferred back to the aqueous phase. The distribution coefficient is temperature dependent, and hot water (50 – 60 °C) is preferred for this stripping process.

Other heavy-metal nitrates can also form complexes with TBP. The more completely the amount of tributyl phosphate present in solution is bonded to $UO_2(NO_3)_2$, the smaller is the extent to which other heavy-metal nitrates are extracted. Therefore, a high degree of saturation of tributyl phosphate with uranyl nitrate is necessary. To reduce the density of tributyl phosphate (which is close to that of water) and to lower its relatively high viscosity, it is diluted with inert, saturated hydrocarbons. Kerosene and *n*-hexane are generally used, giving solutions that usually contain 25 – 35 % TBP.

Harmful impurities in uranium solutions include gypsum and soluble silica, which can be deposited in the solvent extraction apparatus and

reduce the efficiency of extraction. Also, some anions, especially sulfate and phosphate, cause problems by forming complexes with uranium. These complexes are not extracted and therefore lead to uranium losses. Fluorides and chlorides in the presence of nitric acid can cause severe corrosion of the vessel material.

Tributyl phosphate also contains impurities that interfere with extraction. Some of these are present in the commercial product as supplied; others are formed by chemical attack and by radiation. The main impurities are dibutyl- and monobutylphosphoric acids. Although dibutylphosphoric acid extracts uranium more effectively, it is less selective than tributyl phosphate (i.e., impurities are extracted at the same time). These make the stripping of uranium more difficult. Acids can be removed by washing them out of the tributyl phosphate (e.g., with sodium hydroxide solution).

Technique of Solvent Extraction. The concentrate is dissolved in nitric acid, and the concentrations of uranium and nitric acid in the unfiltered solution are adjusted to the desired values, usually 250 – 350 g/L uranium and 1 – 3 mol/L free nitric acid. Before being fed to the solvent extraction plant, the solution is filtered in some plants. Various types of extractors (Fig. 18) are used in solvent extraction plants. Pulsed sieve tray columns are the most popular, but mixer –settlers are also used in some cases. Pulsed columns and mixer – settlers are used both for solvent extraction and for washing and stripping. The columns contain perforated stainless steel plates ca. 50 mm apart. The holes are 3 – 5 mm in diameter, and the total area of all the holes is ca. 23 – 25 % of the area of the plate. In the first column, uranyl nitrate is extracted by the solvent (25 % TBP in kerosene). The volume ratio of the two phases is set such that maximum saturation of the organic phase with uranyl nitrate, together with almost quantitative extraction (>99.9 %), is achieved. In the second column (washing column), the organic phase is washed with a small amount of deionized water to remove impurities extracted at the same time. A volume ratio of organic phase : water = 10 : 1 is adequate. In this washing process, uranium is also partly transferred into the washing water. This solution is therefore recycled to the process. After the organic phase has been washed, it passes into the stripping column where uranyl nitrate is stripped out by hot (60 °C) deionized water. After treatment of the organic solvent with Na_2CO_3 to remove hydrolysis products of tributyl phosphate and the remaining uranium (a step not always included in the process), the organic solution is fed back to the first column.

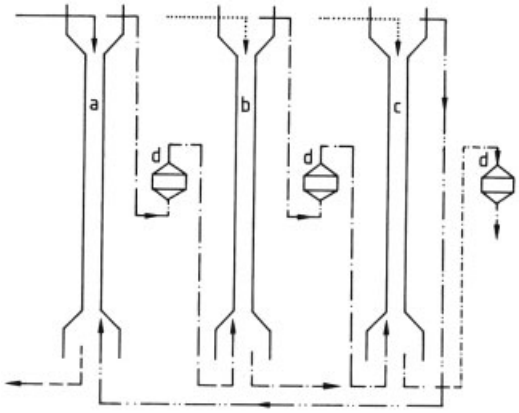

Figure 18. Schematic of a solvent extraction plant
a) Extraction column; b) Washing column; c) Stripping column; d) Diatomite filter
—— Impure $UO_2(NO_3)_2$ solution containing 275 g/L U, 3 mol/L HNO_3; - - - Depleted solution (0.1 g/L U) passing out of extraction column; - · - · Solution of uranyl nitrate in TBP; - · · - · · Depleted solution (1 – 3 g/L U) of uranyl nitrate in TBP; · · · · Water; - - · - - · Aqueous solution of $UO_2(NO_3)_2$

6.3.5. Production of UO_3 and UO_2 from Purified Uranyl Nitrate Solution

Two different methods have been developed for the production of UO_3 and UO_2 from uranyl nitrate solution. In the first of these, the uranyl nitrate solution is evaporated until uranyl nitrate hexahydrate is completely crystallized; and this is then converted into uranium trioxide by thermal decomposition. Alternatively, uranium is precipitated in the form of an insoluble compound, and this precipitate (usually an ammonium salt) is converted into uranium trioxide. Common to both processes is reduction of the uranium trioxide UO_3 obtained to uranium dioxide UO_2. The two processes are compared in Figure 19.

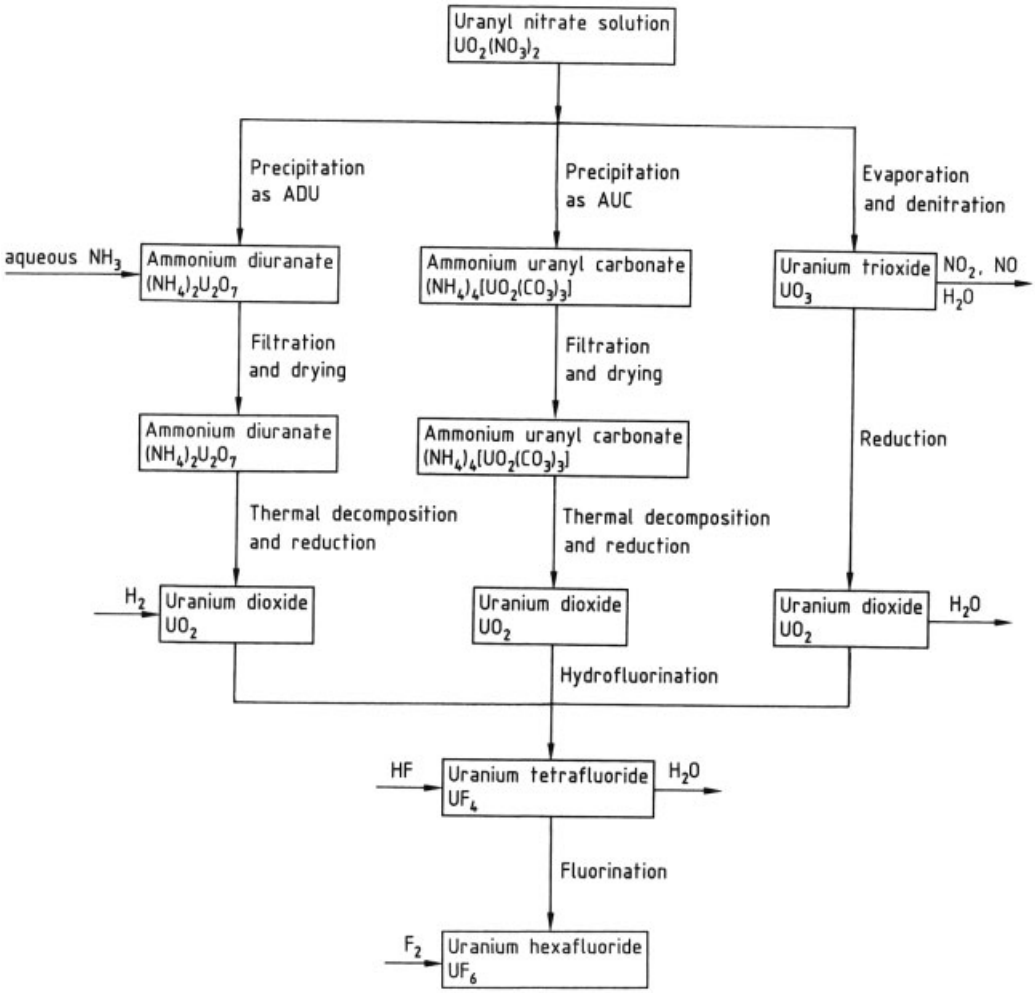

Figure 19. Methods for the production of UF_6

6.3.5.1. Evaporation of Uranyl Nitrate Solution and Denitration by Thermal Decomposition

Solutions of uranyl nitrate are evaporated until uranyl nitrate hexahydrate, $UO_2(NO_3)_2 \cdot 6\,H_2O$, begins to crystallize. Further evaporation leads to the liberation of nitrogen oxides and water:

$$UO_2(NO_3)_2 \cdot 6\,H_2O \rightarrow UO_3 + NO + NO_2 + O_2 + 6\,H_2O$$

This decomposition and the phase diagram of the hydrated uranyl nitrate have been investigated thoroughly. The phase diagram (Fig. 20) shows that the hexahydrate and the trihydrate have incongruent melting points at 60 and 113 °C, respectively. The uranyl nitrate – water system is stable up to 184 °C, the melting point of the dihydrate. Above this temperature, decomposition to uranium trioxide occurs with liberation of nitrogen oxides. This reaction becomes much more rapid above 300 °C. Since uranium trioxide loses oxygen above 430 °C [39], 400 °C is considered the optimum temperature for producing UO_3. In order not to reduce the reactivity of UO_2 in later process steps (reduction, sintering, or hydrofluorination), avoiding higher temperatures, at which the U_3O_8 phase could be formed, is desirable.

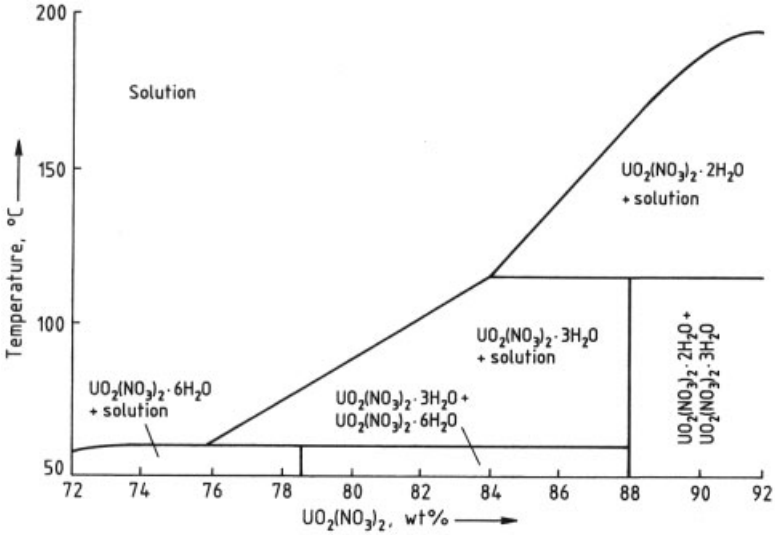

Figure 20. Phase diagram of the system $UO_2(NO_3)_2 - H_2O$

Evaporation. The design of evaporation plants depends on the initial concentration of solutions to be evaporated. Since nitric acid at low concentration (up to 0.05 mol/L) is handled, the construction material is stainless steel. Vertical tube evaporators and evaporation tanks heated by internal steam pipes are commonly used. The final concentration of uranyl nitrate for the subsequent denitration stage is established by means of temperature control. Thus, a temperature of 120 °C corresponds to a uranium concentration of ca. 1000 g/L, and 143 °C to ca. 1285 g/L, the solution fed to the denitration stage being produced at a concentration between these two. The melting point of the less concentrated solution is 60 °C, and that of the more concentrated solution 116 °C. Because these melting points are relatively high, care must be taken to prevent the melt from solidifying in any part of the system (e.g., by heating the pipework). Concentrated uranyl nitrate solution is stored in heated tanks before being pumped to the denitration stage.

Denitration. *Batchwise denitration* was carried out in stainless steel pots containing fairly powerful agitators but the batch process is rarely used today.

Continuous Denitration. The continuously operating furnace first used in 1956 in a plant in Hanford (Washington) consists of a *horizontal trough* with a central stirring device. Uranyl nitrate is fed via rotameters through the inlets to the decomposition trough, and the UO_3 formed flows out over a weir into a powder receiving vessel.

The temperature in the bed of powder largely determines the water and nitrate content of the powder. The nitrate content varies between 1.1 % at 242 °C and 0.02 % at 454 °C. At these temperatures, the water contents are 2.6 and 0.1 %, respectively. Here, too, higher bed temperature leads to powder with lower reactivity.

The mean particle diameter of the powder is ca. 150 µm. In comparison to UO_3 powder produced by other processes, all powder produced by thermal decomposition of uranyl nitrate has a relatively high bulk density (ca. 3.5 – 4.3 g/cm^3) (i.e., the individual particles are very dense).

In the course of development work on equipment, Argonne National Laboratory started to develop a *fluidized-bed denitration furnace* as early as 1953. Fluidized-bed furnaces are now used widely not only in this stage but also in the reduction stage, for example.

Denitration is carried out by spraying uranyl nitrate solution onto a fluidized bed of hot particles of UO_3. Carrier air is introduced from below through a porous metal filter and into the reaction tube where it fluidizes the bed of powder. Gaseous reaction products pass through sintered metal cartridges at the top of the reaction tube, and the nitrogen oxides are subsequently

absorbed in water. The mean particle diameter is 150 – 200 µm. A fluidized-bed reactor of a type still often used today is illustrated in Figure 21.

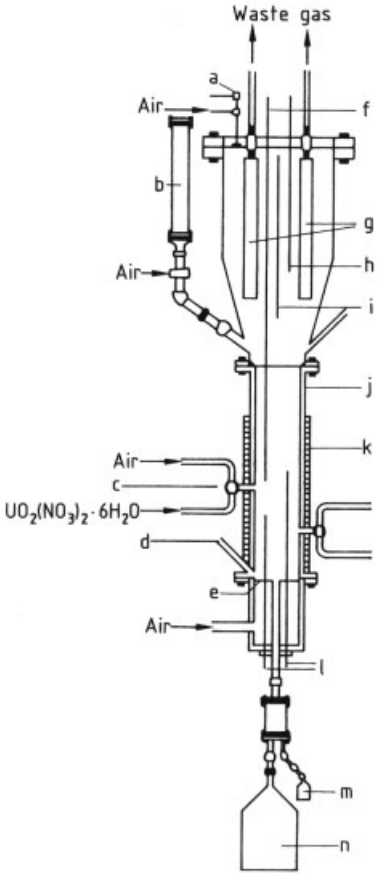

Figure 21. A 10-inch fluidized bed for denitration of $UO_2(NO_3)_2 \cdot 6\,H_2O$ (laboratory plant)
a) Safety valve; b) UO_3 for charging; c) Spray nozzle; d) Pressure compensation; e) Porous filter plate (steel); f) Bed level detector; g) Porous filter; h) Separating plate; i) Thermocouple; j) Reaction tube; k) Heating element; l) Thermocouples; m) Sampling device; n) Receiver

6.3.5.2. Precipitation of Uranium by the ADU and AUC Processes

The sparingly soluble compounds ammonium diuranate (ADU), $(NH_4)_2U_2O_7$, ammonium uranyl carbonate (AUC), $(NH_4)_4[UO_2(CO_3)_3]$ and uranium peroxide, $UO_4 \cdot 2\,H_2O$, can all be used to precipitate uranium. All three compounds form uranium trioxide on thermal decomposition in air, although precipitation of uranium peroxide is now rarely used industrially.

The ADU process was the first to be operated on an industrial scale and is still used widely today [40]. Precipitation of uranium from aqueous solution by ammonia gives a nonstoichiometric precipitate whose composition does not correspond to the chemical formula. Instead, the molar ratio NH_3 : U varies between 0.3 and 0.45. The properties of ammonium diuranate and of uranium dioxide produced from it can be influenced within limits by changing the operating conditions. Important parameters that determine particle size and particle-size distribution include temperature, precipitation time, initial concentration, and intensity of agitation. A relatively large size of agglomerated particles is obtained by using urea as the precipitating agent [41] or by passing a mixture of NH_3 and air into the solution [42]. Depending on the plant, precipitations can be carried out continuously or batchwise. The precipitate is separated from the mother liquor with the aid of continuous filters and centrifuges.

AUC Process [43, 44]. The AUC process was developed for precipitating uranium from fluorine-containing solutions (from the hydrolysis of UF_6) because the fluoride contained in precipitated AUC (produced by addition of NH_3 and CO_2) can be removed nearly completely by washing, in contrast to ammonium diuranate precipitated by NH_3. Because of the high-quality UO_2 powder obtained from the precipitated product, this process was also used for uranyl nitrate solutions. The equations of these reactions are as follows:

$$UF_6 + 5\,H_2O + 10\,NH_3 + 3\,CO_2 \rightarrow (NH_4)_4\left[UO_2(CO_3)_3\right] + 6\,NH_4F$$

$$UO_2(NO_3)_2 + 3\,H_2O + 6\,NH_3 + 3\,CO_2 \rightarrow (NH_4)_4\left[UO_2(CO_3)_3\right] + 2\,NH_4NO_3$$

Ammonia and carbon dioxide pass through a system of jets into a reaction tank filled with water, and uranyl nitrate solution is added at the same time at a controlled rate. The aqueous solution is pumped around a circulating system for the entire period during which reactants are

added. The temperature of the water and the suspension increases because of liberation of heat of solution and reaction. When it reaches 60 °C, the suspension is cooled to prevent excessive decomposition of the ammonium carbonate formed. After cooling to room temperature, the suspension is fed to a rotary disk filter, and after washing the precipitate with water, residual water is removed by suction until the moisture content is ca. 8 %. If desired, the product can be dried further by washing with alcohol. The filter is discharged by rotating it while a knife fixed to a moving arm travels from the center to the edge of the filter over a period of up to 3 – 4 h, removing the filter cake in a thin layer. The powder is fed by a pneumatic suction device through the air lock of the fluidized-bed furnace.

6.3.5.3. Reduction of Precipitated Product to UO_2 Powder

Ammonium diuranate or ammonium uranyl carbonate can be reduced to uranium dioxide either immediately or after drying. The reduction temperature is 500 – 700 °C. Hydrogen or cracked ammonia (i.e., an N_2 – H_2 mixture) can be used as reducing agent. Most reducing furnaces used today are continuously or discontinuously operated fluidized-bed furnaces.

Batchwise Process. A discontinuously operated fluidized-bed furnace is used in the AUC process (Fig. 22). The precipitated product in powder form is fed in small portions of ca. 0.5 kg through a cellular wheel sluice into the furnace, where it decomposes to uranium dioxide in the hot, reducing atmosphere of the furnace. The gases liberated are passed through porous metal filter cartridges. Water vapor is used as the carrier gas for fluidization, and hydrogen is the reducing gas; these are fed through a fritted metal plate.

Continuous plants usually consist of several fluidized beds. The product is fed by a screw feeder into the first reactor where only partial reduction takes place; the powder then passes over an overflow into the second reactor where reduction is completed. *Moving-bed furnaces*, which are electrically heated cylindrical shafts through which a powder – granular material migrates continuously downward from above, produce a heterogeneous powder because of large temperature gradients in the furnace; these are now seldom used. However, *rotary kilns* are sometimes employed and, like fluidized-bed furnaces, are in widespread use for other process stages in the thermal treatment of uranium compounds.

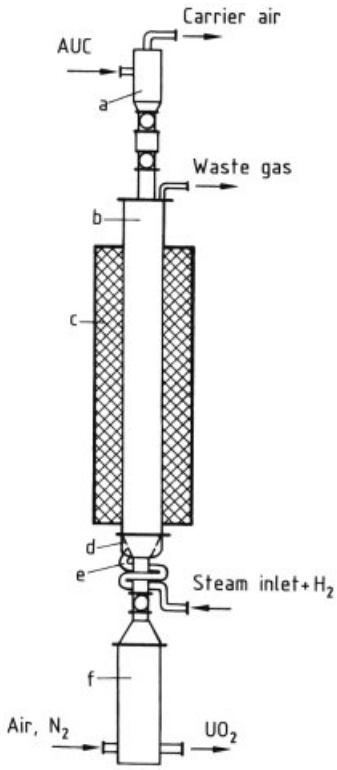

Figure 22. Fluidized-bed furnace for reduction of $(NH_4)_2U_2O_7$ (ADU) or $(NH_4)_4[UO_2(CO_3)_3]$ (AUC)
a) Feed; b) Fluidized-bed furnace; c) Electrical heating; d) Sintered metal cone; e) Exterior heating coil for steam and H_2; f) Discharge vessel

6.3.6. Production of UF_4

Uranium hexafluoride has always been produced on an industrial scale via the intermediate UF_4, which is produced in a dry process by reaction of UO_2 with anhydrous hydrogen fluoride. If UO_3 is the starting material, it is first reduced to UO_2. Uranium hexafluoride is then produced by reaction of UF_4 with elementary fluorine.

6.3.6.1. Hydrofluorination of UO_2

The reaction

$$UO_2 + 4\,HF \rightleftharpoons UF_4 + 2\,H_2O \quad \Delta H_{298} = -178.3\,kJ/mol$$

is carried out at 300 – 600 °C, depending on the reactivity of the UO_2 powder and the desired bulk density of the UF_4 product. Some difficulties are encountered due to the reversible nature of the hydrofluorination reaction and the possibility that eutectic UF_4 – UO_2 mixtures may be formed that can lead to plastic flow and, hence, sintering at higher temperature. Uranium tetrafluoride always contains a small proportion of UO_2F_2 formed from the reaction of uranium trioxide with HF:

$$UO_3 + 2\,HF \rightarrow UO_2F_2 + H_2O$$

Fluidized-bed furnaces and rotary kilns are generally used to carry out the reaction. The LC reactor, which was developed in France, is of a special type (Fig. 23), part of which resembles a moving-bed reactor (a) in which granulated uranium trioxide is reduced to uranium dioxide. The latter is then converted to UF_4 by passing anhydrous hydrogen fluoride over it in the next part of the equipment (d), which has vertical and horizontal sections. The powder flows by gravity through the moving bed and is then propelled forward in the horizontal section by a built-in screw feeder.

6.3.6.2. Hydration and Hydrofluorination of UO_3

In a process developed by British Nuclear Fuels Limited (the BNFL process) [44], uranyl nitrate solutions are concentrated continuously and denitrated to form UO_3.

Reduction of UO_3 to UO_2 was formerly carried out in semicontinuous fluidized-bed reactors, and hydrofluorination of UO_2 to UF_4 in batch-operated fluidized-bed reactors. For economic and technical reasons, a change to continuous rotary kilns was made.

The so-called UTK process (uranium tetrafluoride in kilns) has three principal stages (Fig. 24):

1) Exothermic hydration:

$$UO_3 + x\,H_2O \rightarrow UO_3 \cdot x\,H_2O$$

2) Endothermic dehydration and exothermic reduction:

$$UO_3 \cdot H_2O + H_2 \rightarrow UO_2 + 2\,H_2O$$

3) Exothermic hydrofluorination:

$$UO_2 + 4\,HF \rightarrow UF_4 + 2\,H_2O$$

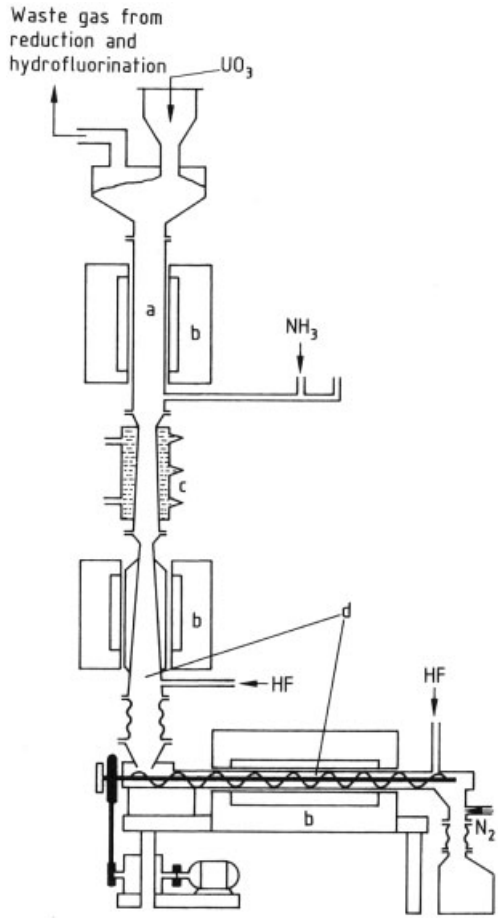

Figure 23. LC reactor for UF_4 production
a) Reduction section; b) Heating; c) Cooling; d) Hydrofluorination section

Uranium trioxide as produced by denitration in the fluidized bed consists of hard spherical particles with low surface area and inner porosity. To increase both of these parameters and thereby ensure rapid diffusion of the reacting gases, UO_3 is hydrated in jacketed vessels made of special steel. These contain screw feeders that

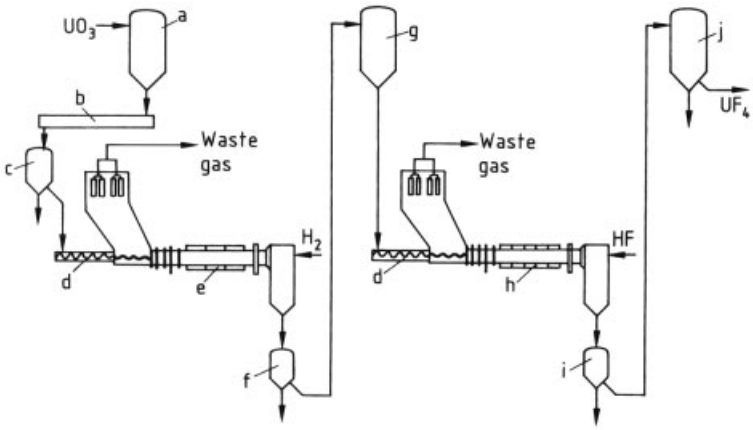

Figure 24. Flow diagram of UF_6 production plant in Springfield, United Kingdom
a) UO_3 storage vessel; b) UO_3 hydrator; c) $UO_3 \cdot H_2O$ storage vessel; d) Screw feeder; e) Rotary kiln for reduction; f) UO_2 intermediate storage vessel; g) UO_2 storage vessel; h) Rotary kiln for hydrofluorination; i) Intermediate storage vessel for UF_4; j) UF_4 receiver

operate intermittently, bringing the powder into contact with metered quantities of steam. Careful control of contact time and temperature ensures that the product is a dry, free-flowing powder.

Hydrated UO_3 powder is reduced in stainless steel rotary kilns (b) containing a system of pushers and retaining rings to give intimate contact with the countercurrent flow of hydrogen. The UO_2 formed is fed pneumatically under nitrogen into the feeder tank (c) of the hydrofluorination stage. The reaction temperature over the entire length of the kiln must be controlled carefully to ensure complete reduction of the product without any decrease in its chemical reactivity during the endothermic dehydration and exothermic reduction. Waste gases leaving the kiln contain nitrogen, steam, and excess hydrogen; they are passed first through filters that remove suspended particles of powder and then through water-cooled condensers and a flame. The gases are then discharged into the atmosphere.

UO_2 is converted into UF_4 in a rotary kiln (h) containing similar powder feeding equipment, the material of construction being Inconel. Waste gases (nitrogen and small amounts of excess HF) are filtered and scrubbed with sodium hydroxide solution before being released to the atmosphere. The NaF solutions produced are treated with hydrated lime.

6.3.7. Production of UF_6 from UF_4

6.3.7.1. Process Description

The only processes used on an industrial scale involve the use of elementary fluorine. Uranium tetrafluoride reacts with fluorine according to

$$\mathrm{UF_4\,(solid) + F_2\,(g)} \rightarrow \mathrm{UF_6\,(g)}\,;\ \Delta H_{298} = -246\,\mathrm{kJ/mol}$$

The reaction proceeds fairly rapidly above 250 °C. Reaction temperatures of 400 °C are used for fluorination in fluidized-bed reactors, while flame reactors operate at 1100 °C.

If the conversion to UF_4 is incomplete, the product will contain UO_2 and UO_2F_2, which also react with elementary fluorine. Large amounts of fluorine are consumed in these reactions, so the concentration of these compounds in UF_4 should be kept low. All metallic impurities react to give fluorides in which the metals are at their maximum oxidation states. Nonvolatile fluorides (e.g., AlF_3, FeF_3, and ThF_4) remain behind with unreacted UF_4; the volatile compound UF_6 contains volatile fluoride impurities (MoF_6, VF_6, SiF_4, CF_4, SF_6, etc.). These impurities can be removed by distilling UF_6 and condensing it as a solid. For this, several condensers are used in series. Volatile fluorides of other elements contained as impurities (C, P, Si,

B, and S) can be removed without problem during crystallization of UF_6; nonvolatile fluorides remain behind as solids at the bottom of the reactor. Most of the UF_6 can be recovered between +5 and −15 °C in relatively large condensers; the remaining UF_6 is collected in much smaller condensers at ca. −50 °C. Uranium hexafluoride is then melted out of the filled condensation vessels and collected in storage cylinders.

The heat-exchange surfaces of the condensers consist of flat ribbed plates that are fixed with baffles to the coolant pipes and steam pipes to give good contact. The tubes are U-shaped to reduce thermal stress to a minimum and to enable the condenser contents to be removed easily from the vessel.

6.3.7.2. Chemical Reactors

Fluidized beds can be used for all stages of the process, i.e., the production of UO_3 and its further reaction to UO_2, UF_4, and UF_6. All of these are gas – solid reactions for which the fluidized-bed reactor is ideally suited, providing rapid mixing of the components and having only small temperature gradients within the bed volume.

Flame reactors, developed by Union Carbide Nuclear Corp., are also used. These reactors (Fig. 25) consist of a vertical Monel tube surrounded by a cooling coil, with a device at the top to give good distribution of UF_4, which is fed by a screw feeder. Fluorine, which is in excess, is fed through nozzles to the top of the reactor. Reaction of the two components produces a flame at ca. 1100 °C inside the reactor. The upper part of the reactor, where most of the reaction occurs, must be cooled to give a wall temperature between 450 and 540 °C. This both minimizes corrosion of the structural materials and prevents the formation of low-valence uranium fluorides.

6.3.7.3. Removal of Excess Fluorine from UF_6

To achieve complete fluorination of uranium, excess fluorine must be used; therefore, excess fluorine will still be present in the reaction gas after it has passed through the condenser. This fluorine comprises ca. 20 – 40 % of the reaction gas and must be removed in an additional chemical reactor. This is similar to the fluorination reactor and operates with an excess of UF_4. Not only is UF_6 formed, but also lower-valence intermediates such as UF_5, U_2F_9, and U_4F_{17}.

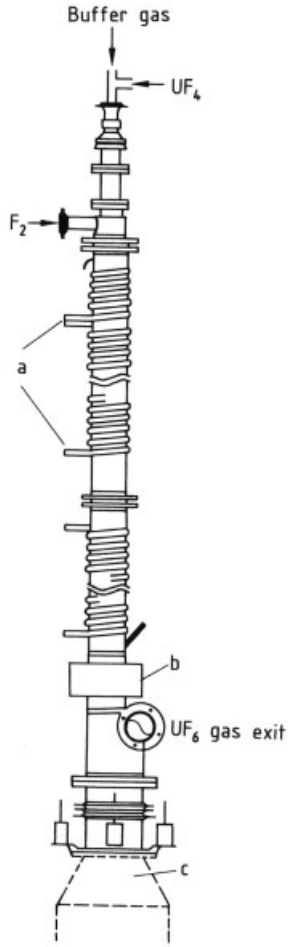

Figure 25. Flame reactor for fluorination of UF_4
a) Cooling coils; b) Expansion compensator; c) Ash receiver

6.3.8. Complete Plant for Production of UF_6 from Uranyl Nitrate

6.3.8.1. French Process in Pierrelatte [45]

Uranyl nitrate solution (see Fig. 26) is pumped into a precipitation vessel (a) together with ammonia and water, which react to form ammonium diuranate. Temperature and pH are controlled during precipitation. The precipitate is

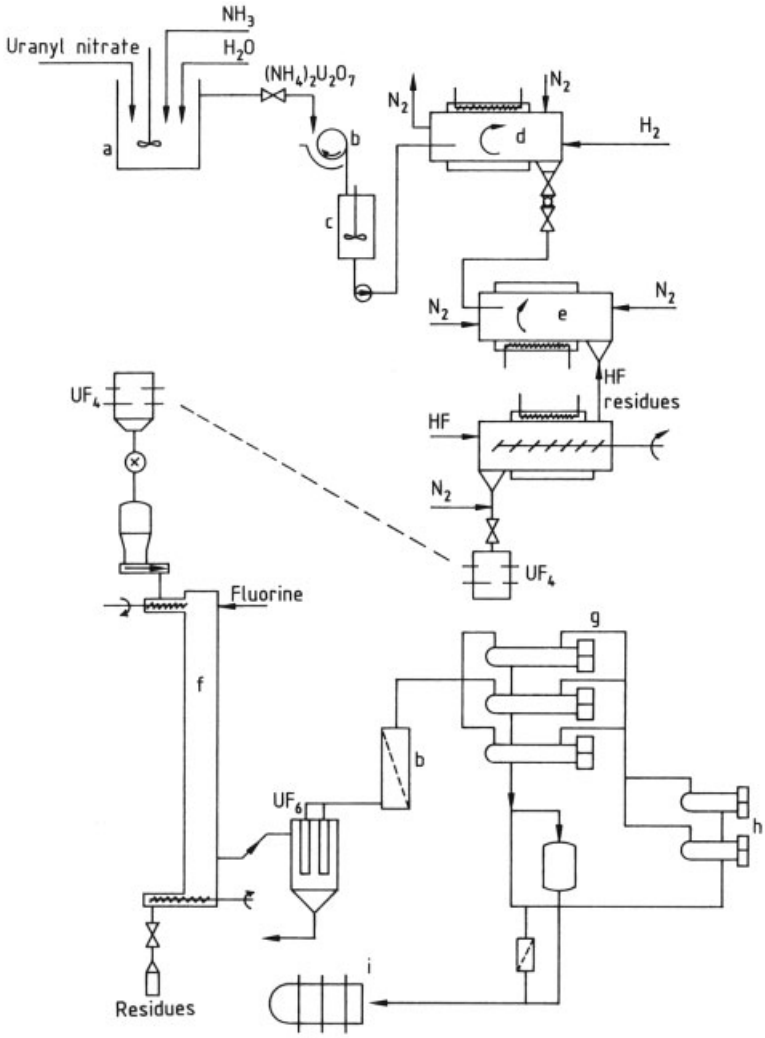

Figure 26. French process for UF_6 production
a) Precipitation; b) Filters; c) Mixer; d) Calcination and reduction; e) Hydrofluorination; f) Flame reactor; g) Primary crystallizers; h) Secondary crystallizers; i) UF_6 receiver

filtered off by using a vacuum drum filter (b) and is transferred from this to a mixer (c), which converts it into a pumpable slurry by utilizing the thixotropic properties of the material. Ammonium diuranate is calcined to UO_3 in the first rotary kiln (d) and then reduced by a countercurrent flow of hydrogen further down this kiln. Uranium dioxide passes through a seal into a second rotary kiln where hydrofluorination occurs. A third kiln containing a screw feeder completes this reaction. The HF required for fluorination is fed from the exit end of the third rotary kiln in countercurrent flow to the solid. The UF_4 formed is conditioned in drums and then burned in a flame reactor (a vertical Monel tube 4.40 m in length) to form UF_6. The solid byproducts UF_5 and U_2F_9, which are formed in very small quantities, are removed by a screw and recycled to the top of the reactor.

6.3.8.2. Allied Chemical Process

Unlike the processes described above, the Allied Chemical process does not include final purification of uranyl nitrate by liquid – liquid extraction. Instead, UF_6 is purified by distillation [46]

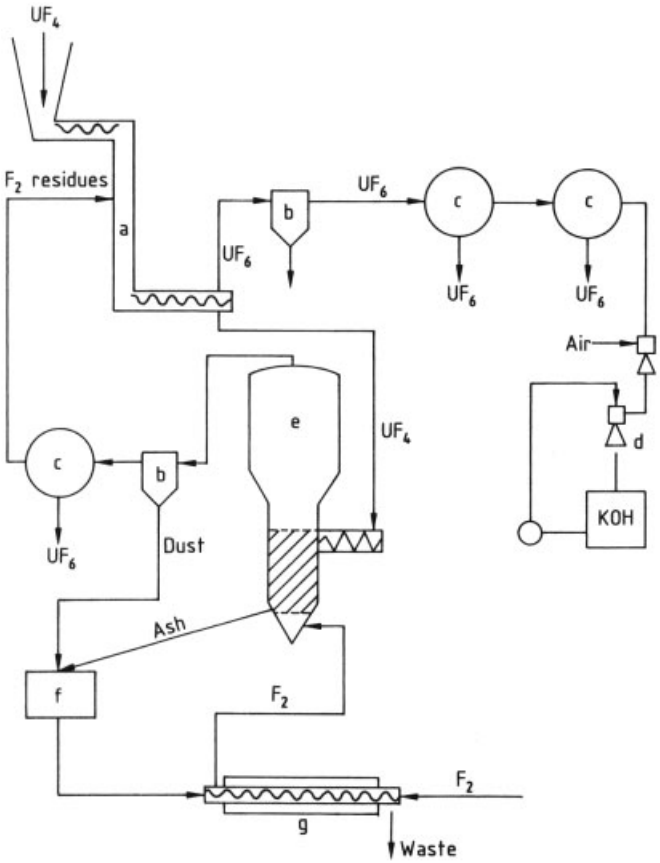

Figure 27. Flow diagram of an Allied Chemical UF_6 production plant
a) Tower reactor; b) Filter; c) Condenser; d) Scrubber; e) Fluorinator; f) Mill; g) Ash treatment

at a pressure and temperature above the triple point of UF_6 (64.05 °C at 151.1 kPa). Schematics of UF_6 production and distillation plants are shown in Figures 27 and 28, respectively.

6.3.9. Enrichment of ^{235}U

Three industrial processes are used for enrichment of uranium: the diffusion process, the ultracentrifuge process, and the less important jet nozzle process. Chemical enrichment, laser separation, and the plasma process are used on an experimental or laboratory scale. The three industrial processes employ gaseous UF_6. Chemical enrichment requires the uranium to be in solution. Two laser separation processes are in competition with each other. Metallic uranium is used in one of these, and gaseous UF_6 in the other. Metallic uranium is used in the plasma process as well.

6.3.9.1. Diffusion Process

Separation by diffusion occurs because isotopes diffuse through a semipermeable membrane at different rates owing to the $\sqrt{m}$ effect. A schematic diagram of the process is given in Figure 29 [47]. The separation factor is small (i.e., 1.002 – 1.004 per stage). Many separation stages (1000 – 1500) are arranged in series in cascade to give an integral separation effect. For improved maintenance and ease of replacement, 5, 8, 10, 16, or 20 separation cells are combined as a module. The heavier fraction (depleted in ^{235}U) that did not diffuse goes back to the previous stage. The heart of the separation cell is the

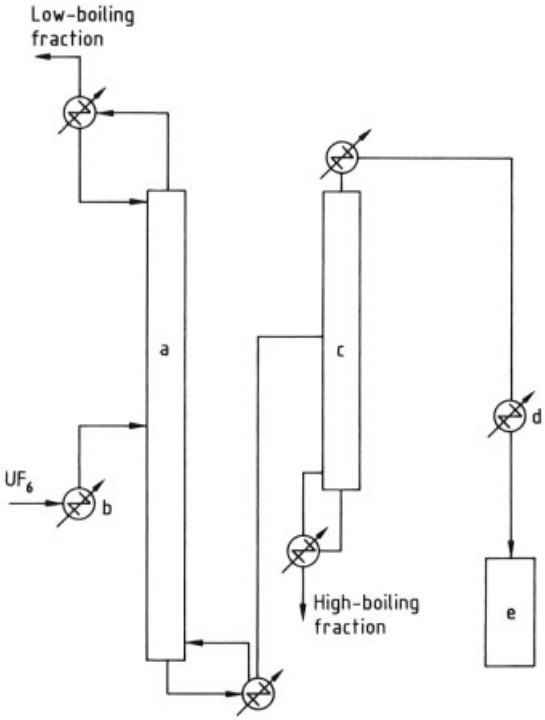

Figure 28. Flow diagram of the fine purification of UF_6 by the Allied Chemical process
a) Column I for low-boiling fraction; b) Evaporator; c) Column II for high-boiling fraction; d) UF_6 cooler; e) UF_6 receiver

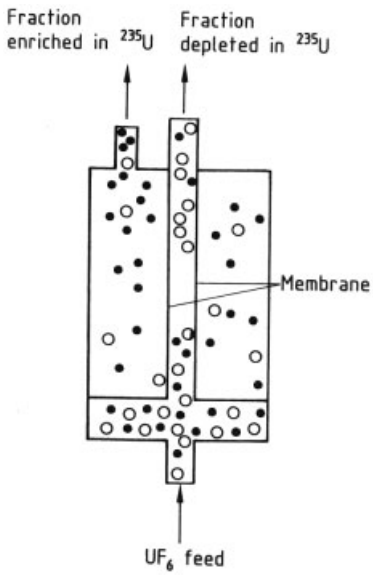

Figure 29. Schematic of a diffusion cell [47]

diffusion membrane, which usually has a tubular shape. Diffusion is accelerated by a concentration difference between the outside and the inside induced by a pressure. The pressure difference is 1 – 10 kPa, depending on the membrane material, and the operating pressure in a cell is 10 –30 kPa. Many materials have been tested as membrane materials, including Au, Ag, Ni, Al, Cu, Monel, Teflon, and Al_2O_3. Details of membranes are kept secret. They must be as thin as possible and have a pore size of 10 nm, but they must also have sufficient mechanical stability to withstand the pressure difference for years. One method of producing pores is to dissolve out a component of alloy foils (Ag – Zn, Au – Ag) selectively. Al_2O_3 membranes are produced from aluminum foil (13 – 30 μm thick) by anodically oxidizing it to completion in 5 – 10 % sulfuric acid and then heat treating at 500 – 530 °C. Water is evolved, and a porous membrane is formed with a permeability of ca. 0.1 $cm^3\ cm^{-2}\ s^{-1}$. In the EURODIF plant in Tricastin, France, diffusers (diffusion cells) and compressors have been joined to form a single unit. The main disadvantage of the diffusion process is the high energy requirement for the work of compression. For example, the EURODIF plant, which provides 10 800 t USW/a (tonnes of uranium separative work per year), requires ca. 3000 MW electric power, for which four nuclear power stations, each of 925 MW capacity, have been constructed [48].

6.3.9.2. Ultracentrifuges

In 1942, P. Harteck and W. Groth first used centrifuge technology to separate uranium isotopes. The countercurrent centrifuge is illustrated in Figure 30 [47]. The separation factor of 1.2 formerly achieved has now been increased to 1.5 in modern equipment of this type. In Germany in 1970, a trilateral centrifuge project among the United Kingdom, The Netherlands, and the Federal Republic of Germany was inaugurated (foundation of URENCO), and construction of three plants each with an enrichment capacity of 200 t USW/a was agreed upon. The German and Dutch plants are in Gronau, Lower Saxony, and Almelo, The Netherlands, respectively; the British plant is in Capenhurst,

United Kingdom. Over the long period of operation, 99 % plant availability has been achieved [49]. Ultracentrifuge plants of capacity >100 t USW/a are competitive with diffusion plants and have two important advantages:

1) The specific energy consumption using the centrifuge technique is 200 kW · h/kg USW.
2) The cascade for a low degree of enrichment (e.g., for light-water reactors) has only 15 – 20 separation stages, but diffusion equipment requires 1200 stages (EURODIF). A plant using the ultracentrifuge technique can be extended easily by the modular method using parallel connections.

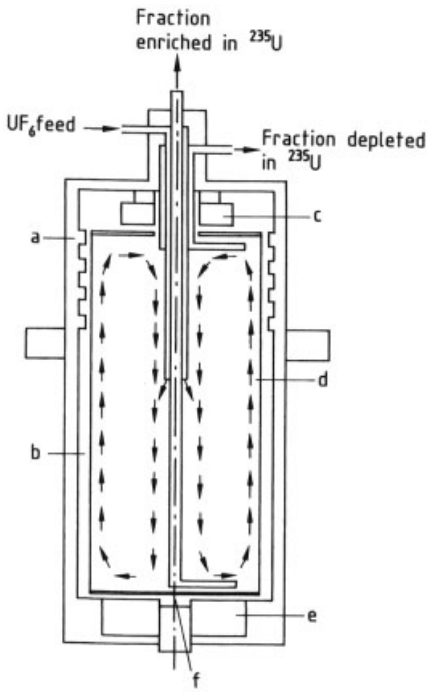

Figure 30. Cross section of a centrifuge [47]
a) Housing; b) Vacuum; c) Magnetic holding device; d) Rotor; e) Electrical drive; f) Needle bearing

6.3.9.3. Nozzle Process

This process, which was developed by E. W. Becker in Karlsruhe, is the only one to which secrecy restrictions do not apply [50]. A modern nozzle with dimensional data is described in [51]. The separation effect is improved by addition of a light gas. The nozzle principle is illustrated in Figure 31 [47].

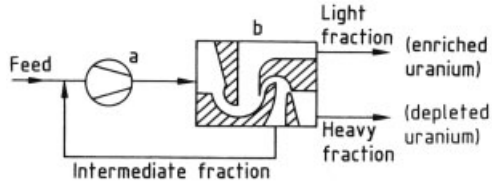

Figure 31. Schematic of the Brazilian nozzle separation system [47]
a) Compressor; b) Double deflecting system

A mixture of 4 % gaseous UF_6 and 96 % of a light carrier gas (e.g., H_2 or He) is expanded in the first jet. Due to the flow deflection in the carrier stream the heavier uranium isotopes become enriched in the outer flow region. Subsequently the outer flow region is separated by specially designed separation geometry from the inner flow region. The second jet works similarly and is used to improve the efficiency of the stage.

The production of enriched uranium for light-water reactors requires several hundred stages in a cascade; i.e., the separation effect (1.015 – 1.026) is better than that obtained by the diffusion process and worse than that obtained by ultracentrifugation. The energy requirement is comparable to that of the diffusion process since high energies of compression are required. The process is operated on an industrial scale in Brazil and the Republic of South Africa (the Helikon process).

6.3.9.4. Chemical Enrichment

The principle of chemical enrichment is based on the fact that in some chemical exchange processes (e.g., $UX + Y \rightleftharpoons UY + X$), different uranium isotopes have different tendencies to bond to substances X and Y. This mechanism is based on small differences in the preferred mean oxidation states of the particular isotopes in the given reaction environment. For industrial operation, the reaction must be rapid; the separation factor for the reaction must be acceptable; and the reverse reaction must occur readily. A pilot plant in operation in Japan [52] depends in principle on the equilibrium between U^{4+} and

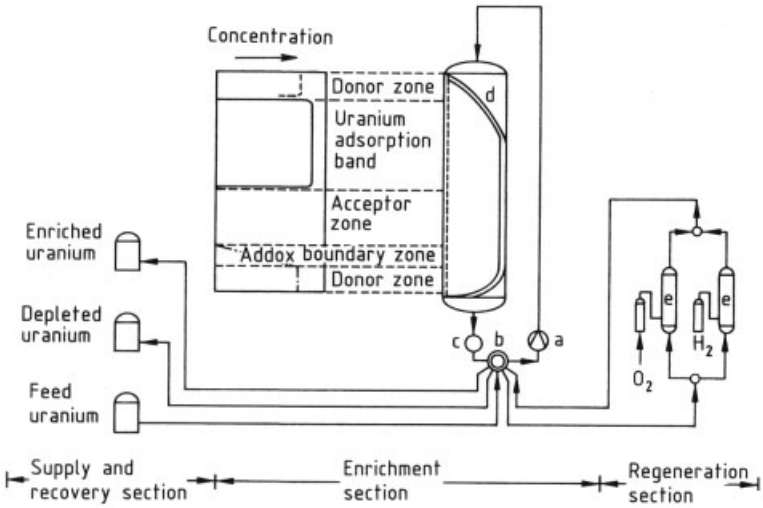

Figure 32. Schematic of the Japanese chemical enrichment plant
a) Pump; b) Valve; c) Detector; d) Enrichment column; e) Regeneration reactors

UO_2^{2+} in aqueous solution in the presence of an ion-exchange material. In this environment, the preferred valence of ^{238}U is 4 and that of ^{235}U is 6. The ion-exchange material has no preference for either isotope, but it preferably adsorbs hexavalent uranium isotopes. So the overall effect is a slight increase in ^{235}U concentration. Since the enrichment factor is very small (1.001 to 1.0014), many stages in the sequence are necessary. The integration of many stages into a continuously operating column was achieved in the pilot plant with the development of the self-generation reaction (Addox reaction) between the deactivated redox agents. A regeneration of ca. 70 % is achieved in a semi-commercial plant. Figure 32 shows a schematic diagram of the plant.

In the column the zones are migrating continuously from top to bottom. In the adsorption zone the ion exchanger is loaded with ^{235}U. If the adsorption zone arrives at the valve, the adsorber is removed together with enriched uranium. The acceptor zone contains the depleted uranium. Uranium and regenerated ion exchanger material are introduced at the donor zone. The redox agents are regenerated in the addox boundary zone. Hydrogen and oxygen are fed in the area of the valve to complete regeneration of the redox agent.

In France, the equilibrium between U^{3+} and U^{4+} is utilized [53]. Separation occurs between an aqueous hydrochloric acid phase retaining trivalent ^{238}U and the phase of an organophosphorus compound, retaining $^{235}UCl_4$. The enrichment factor is ca. 1.025. This process also requires several thousand stages.

6.3.9.5. Laser Separation Process

Laser separation of isotopes is based on the fact that small differences exist between the absorption spectra of ^{235}U and ^{238}U. By using lasers of suitable wavelength, one isotope can be excited selectively. Two methods have been investigated, in one of which selective excitation is applied to uranium metal vapor, and in the other to gaseous UF_6. The first of these is known as the Atomic Vapor Laser Isotope Separation (AVLIS) process. Here, the selectively excited uranium atom is ionized and then separated electrostatically from the stream of metal vapor [54]. In the Molecular Laser Isotope Separation (MLIS) process, selectively excited UF_6 molecules dissociate into UF_5 and F. Fluorine must be bonded rapidly to suppress the reverse reaction. The UF_5 can then be separated from the stream of molecules [54]. Both processes require several stages despite the high separation factor. The AVLIS process is regarded as the

most promising because the sequence of stages can be put together without problems. However, in the MLIS process, UF_5 must be converted back to UF_6 after each stage [54]. For this reason, research in the United States was formerly concentrated on the AVLIS process.

Two possible methods of incorporating the AVLIS process into the series of operations to convert uranium ore into nuclear fuel are illustrated in Figure 33.

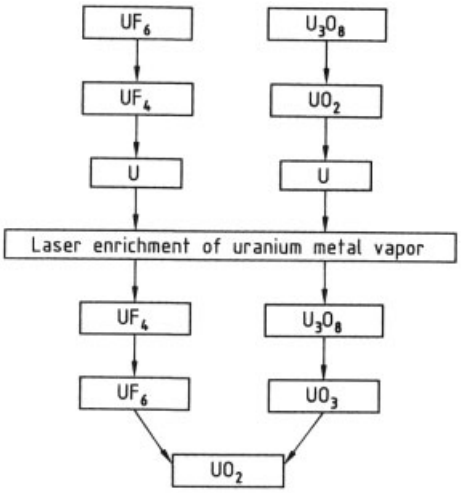

Figure 33. Example of incorporation of the AVLIS process into the chemical process chain of fuel element production

6.3.9.6. Plasma Processes

Two effects suitable for isotope separation are under investigation.

The first utilizes the fact that uranium plasmas rotating in magnetic fields achieve considerably higher rotational speeds than those produced by a centrifuge. Thus, uranium isotopes can be separated by gravity. The second effect utilizes the fact that in a homogeneous magnetic field the cyclotron frequency of the ions in the plasma depends only on magnetic field strength and ionic mass. Hence, the isotopes have different circulation paths in the cyclotron. Thus, isotopes are separated in a rotating plasma and collected separately by different collecting devices.

6.3.10. Production of UO_2 Pellets from UF_6

6.3.10.1. Conversion of UF_6 to UO_2

A schematic overview of industrially important processes for producing UO_2 pellets from UF_6 is given in Figure 34. These may be divided into wet and dry processes.

Wet Processes. Wet processes include the precipitation of ADU and AUC, as described in Section 6.3.5 for obtaining UO_2 from high-purity uranyl nitrate. The apparatus for converting UF_6 to UO_2 is illustrated in Figures 35 and 36. The difference from the processes described in Section 6.3.5 is that the starting material is UF_6. The following two reactions occur:

ADU process

$$2\,UF_6 + 2\,NH_3 + 7\,H_2O \rightarrow (NH_4)_2U_2O_7 + 12\,HF$$

AUC process

$$UF_6 + 5\,H_2O + 10\,NH_3 + 3\,CO_2 \rightarrow (NH_4)_4UO_2(CO_3)_3 + 6\,NH_4F$$

Subsequent conversion of ammonium diuranate or ammonium uranyl carbonate to UO_2 powder is carried out by methods similar to those already described. However, since conditions in the calcination stage significantly affect the properties of the powder for later finishing operations, properties are optimized by controlling those conditions.

Generally, the advantage of the ADU and AUC conversion routes is the possibility to process directly the UNH from recycling of wastes produced by fuel manufacturing.

Dry Processes. Reaction equations for dry processes are given below in simplified form:

$$UF_6 + 2\,H_2O \rightarrow UO_2F_2 + 4\,HF$$

$$UO_2F_2 + H_2 \rightarrow UO_2 + 2\,HF$$

However, these dry processes vary considerably with respect to processing details and design of the chemical reactors. In the older processes, up to four separate reaction chambers were required, whereas modern processes use one or at most two reaction chambers. A schematic diagram of the Direct Conversion (DC)

UF_6

Process step	Process				
	Wet processes		Dry processes		
	ADU	AUC	IDR	DC	GECO
UF_6 feed system	Vaporization of UF_6 from transport containers by controlled heating				
Precipitation or reaction	As ADU	As AUC	As UO_2	As UO_2 + UO_2F_2	As UO_2 + UO_2F_2
Recovery of first product	Filtration and drying		Filtration in sintered filter cartridges		
Conversion of first product into UO_2	Calcination ADU → UO_2	Calcination AUC → UO_2	↓	↓	Calcination F removal
Additions	U_3O_8, compaction auxiliaries	U_3O_8	U_3O_8, compaction auxiliaries		
Powder conditioners	Milling, compaction, granulation	↓	Milling, compaction, granulation		
Production of green pellets	Green pellets usually formed in mechanical rotary presses				
Sintering	Sintering in Mo boats under H_2 with controlled oxygen potential (H_2: H_2O ratio) at up to 1780°C to achieve good stoichiometry				
Finishing	Centerless grinding				

UO_2 pellets

Figure 34. From UF_6 to UO_2 pellets for light-water reactors

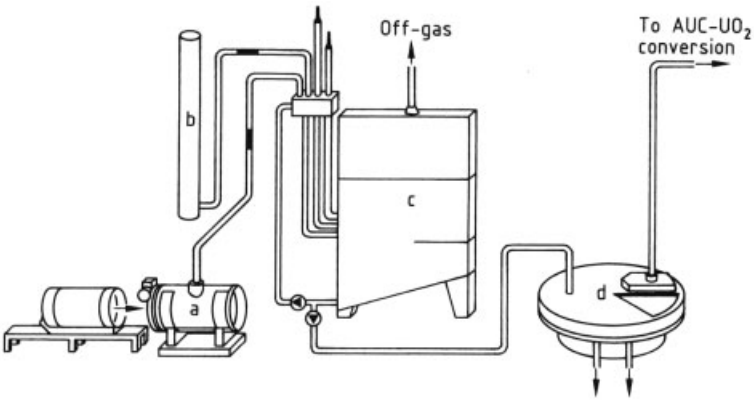

Figure 35. Conversion of UF_6 into AUC
a) UF_6 evaporation; b) NH_3, CO_2 supply; c) Reaction tank; d) Rotating filter for AUC collection

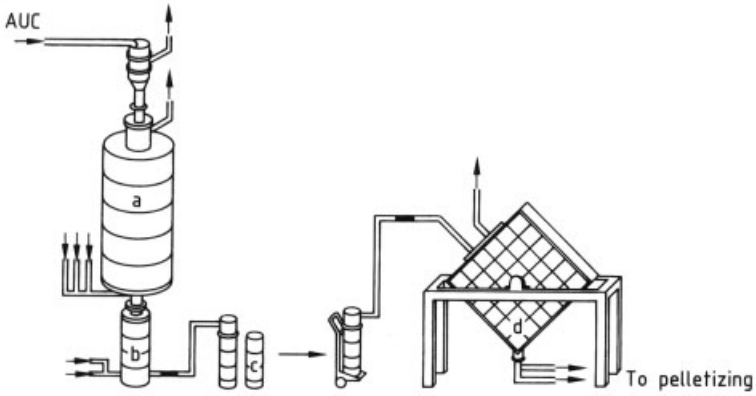

Figure 36. Conversion of AUC into UO_2 powder
a) Fluidized bed; b) Stabilization; c) Storage container; d) Homogenizer

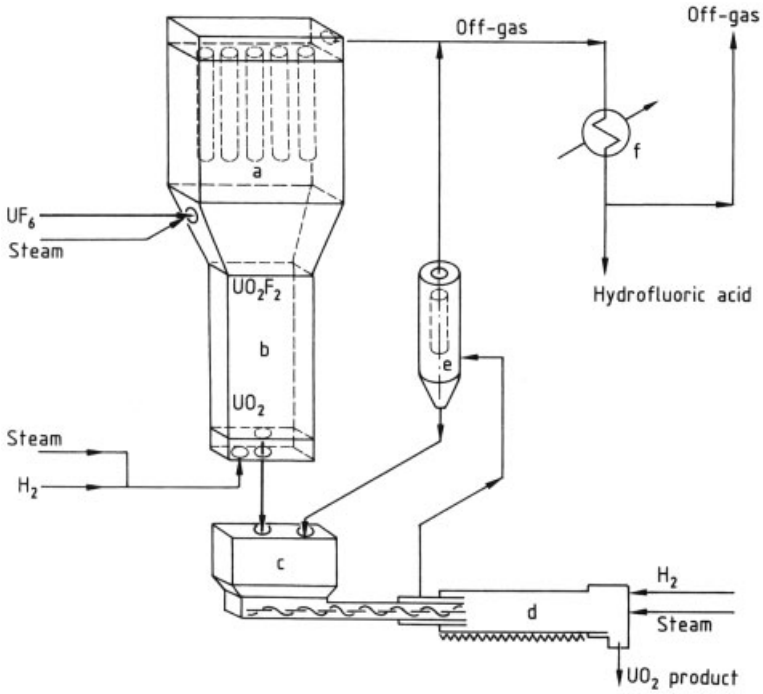

Figure 37. Schematic of Direct Conversion process
a) Reactor; b) Fluidizd bed; c) Feed bin; d) Rotary kiln; e) Filter; f) Condenser

process is given in Figure 37. Construction materials must be chosen very carefully because of the corrosive properties of the hot, gaseous hydrogen fluoride produced. However, satisfactory experience with this plant shows that reliable solutions to these problems have been achieved by materials technology.

6.3.10.2. Possible Future Developments in the Conversion of UF_6 to UO_2 Powder

Nowadays, both wet and dry processes are used worldwide to convert UF_6 to UO_2 powder for the production of nuclear fuel pellets for light-water reactors [55]. The combination of processes has

been optimized within the limits of process parameters to enable the pellets to be produced economically and in accordance with quality specifications. The standards established are applicable to all later production developments, especially the dry conversion processes.

Future Processes. In the production of UO_2 pellets, intermediate products in powder form can of course be avoided only by using *wet processes*. A possible approach is to use the sol – gel process [56, 57, 59], in which UF_6 is dissolved in aqueous uranyl nitrate solution and mixed with poly(vinyl alcohol). This solution is then sprayed into a second liquid with a very different surface tension, and ammonia is added. This causes hardening of the droplets to form particles, which are then washed in aqueous NH_3 solution to remove the undesired byproduct NH_4NO_3. A calcination process in air then decomposes the poly(vinyl alcohol) and the ammonium diuranate to UO_2. The solid particles obtained can then be compressed directly into pellets. The process is carried out in such a way that the UO_2 pellets are of similar quality and have similar properties to those produced by conventional processes.

However, analysis of this process shows that the many washing and purification steps and the considerable volume of materials used lead to a large volume of process and waste materials that must be cleaned and disposed of. Although very little dust is produced, the cost of cleaning and waste disposal is high. For this reason, the main interest in process development has now reverted to *dry processes* such as IDR, DC, and GECO, the industrial handling of dusts and aerosols having meanwhile been greatly improved.

All these processes convert UF_6 to UO_2 by the following reactions

$$UF_6 + 2\,H_2O\,(\text{steam}) \rightarrow UO_2F_2 + 4\,HF$$

$$UO_2F_2 + H_2 \rightarrow UO_2 + 2\,HF$$

These two reaction steps can be combined in one process as realized by the BNFL IDR process. Siemens prefers a two-step procedure. In a fluidized-bed reactor UO_2 is formed together with minor amounts of UO_2F_2 [58]. In a subsequent rotary kiln the $UO_2 - UO_2F_2$ mixture is converted to UO_2 which contains negligible impurities of fluorine. A slightly different approach is applied for conversion in the oxygen–hydrogen flame, which is also in use on a manufacturing scale (TVEL). Here the conversion of UF_6 to uranium oxide is performed already within a flame reactor with only small amounts of fluorides in the reaction product. A common feature of all processes is to provide a UO_2 powder with an acceptable sinterability and a low F content. Reducing the fluorine content requires a hydrogen treatment at higher temperature (500

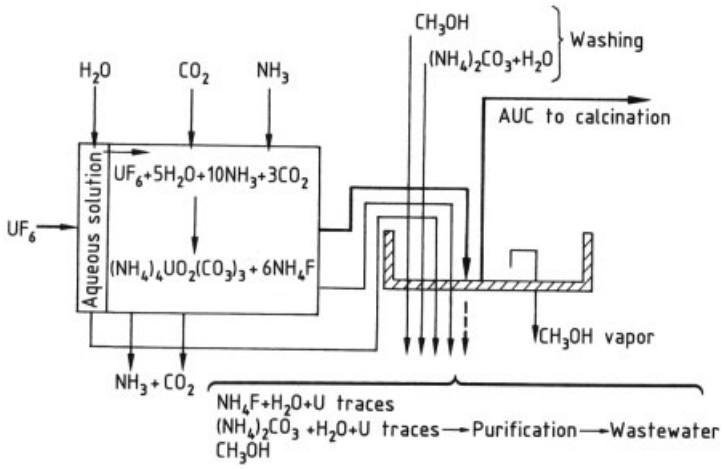

Figure 38. AUC conversion – precipitation stage

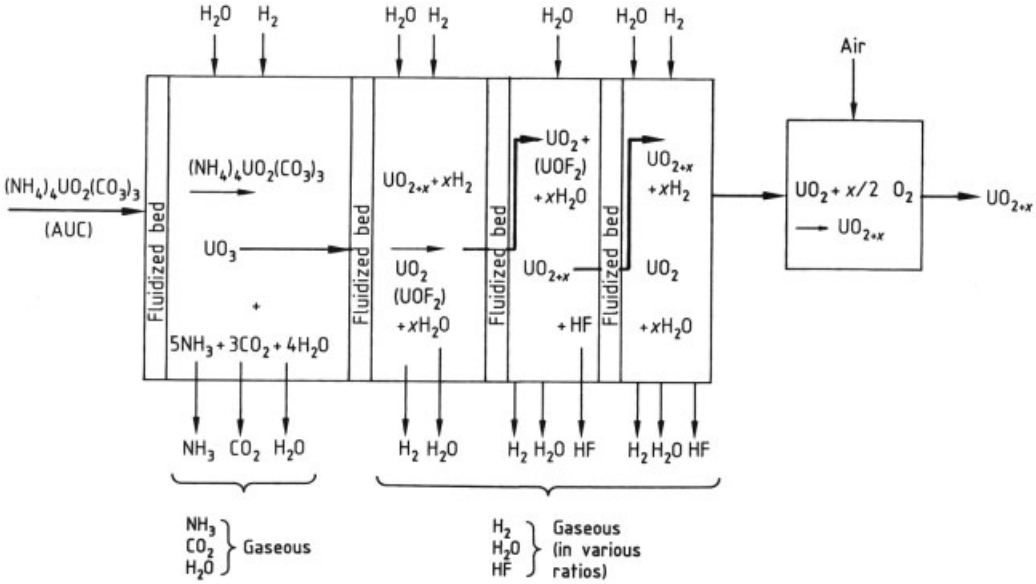

Figure 39. AUC conversion – calcination stage

Process step	Residual material	Treatment	Disposal route
Precipitation	g: NH_3 g: CO_2	Waste-gas scrubbing	Stack
Filtration	l: NH_4F l: H_2O l: U (traces) l: Other activity	Peroxide precipitation Lime precipitation	Utilization
Washing	l: $(NH_4)_2CO_3$ l: H_2O l: U (traces)	ADU precipitation	Wastewater
AUC feed (AUC → UO_{2+x})	g: NH_3 g: CO_2 g: H_2O g: H_2 g: H_2O g: HF	Waste-gas scrubbing	Stack
Calcination			

Figure 40. Treatment of processing materials and residues in the AUC process (simplified)
g = gaseous phase; l = liquid phase

– 800 °C), which results in an optimization of fluorine depletion and sinterability.

Chemicals and Residues in the AUC and DC Processes. Flows of materials in the *AUC process* are shown in Figures 38 and 39, in which the main product stream flows from left to right and the processing chemicals and residues flow from top to bottom. Figure 40 shows how the treatment (not presented in detail here) of processing chemicals and residues is integrated in such a way that it is compatible with the main process of $UF_6 \rightarrow UO_2$ conversion. Material flows are suitably interlinked to minimize consumption of raw materials and processing chemicals, and to recycle usable materials. At the same time, individual process steps must be op-

timized continuously to ensure compliance with steadily decreasing pollution limits.

Figure 41 shows the *DC process* in a diagram similar to that for the AUC process. The processing materials are steam, H_2, and N_2. In addition, HF is formed in amounts corresponding to the amounts of UF_6 treated. Gas streams passing from the reaction vessels are filtered several times to remove solids. After condensation of hydrogen fluoride, the gas is recycled as far as possible for use as a processing material. Residual gas is passed over $CaCO_3$ to remove HF completely, the traces of HF reacting to form CaF_2. The gas so obtained is flared off to remove H_2, and the waste-gas stream can be released from a stack since it contains only residual H_2 and H_2O. Fluorine introduced into the process in the form of UF_6 can be reused in the form of pure hydrofluoric acid. The small amounts of CaF_2 in $CaCO_3$ can be dumped without further treatment.

A comparison of the aqueous AUC process and the dry DC process shows that for the same product quality, the dry process has considerable advantages in the treatment of processing and waste materials. The DC process is consequently growing in importance, which shows how ongoing efforts to reduce pollution can also lead to process rationalization.

6.3.10.3. Pelletizing of UO_2 Powder

A process overview is given in Figure 34. The fuel pellets are generally fabricated by means of powder metallurgy processes: powder preparation–compaction–sintering. Some processes require preparation of the powder by compaction and granulation. When the AUC process is used, the UO_2 powder obtained can be compressed and sintered without any intermediate steps.

6.3.11. Production of Uranium Metal

The requirement for uranium metal is limited (Fig. 14). Only Magnox reactors for power generation require uranium metal as fuel (→ Nuclear Technology, Chap. 3.2.2.1). Uranium metal is also needed for the production of U – Zr, U – Zr hydride, uranium silicide, and uranium – aluminum alloys. If laser enrichment processes such as the AVLIS process attain industrial importance, the requirement for elementary uranium will increase. Uranium metal is produced from UF_6 by a two-stage reduction; the intermediate product is UF_4.

6.3.11.1. Reduction of UF_6 to UF_4

Two principal routes are used:

1) Reduction of UF_6 by H_2 in the gas phase
2) Reduction of UF_6 with chlorinated hydrocarbons (mainly unsaturated)

Reduction with Hydrogen. The throughput and degree of enrichment required are important considerations when selecting a process. Reduction of UF_6 by hydrogen in so-called hot-wall and cold-wall reactors was reported at the Geneva Conference of 1958 [60]. The reaction is exothermic:

$$UF_6 + H_2 \rightarrow UF_4 + 2\,HF \quad \Delta H = -281.7\,kJ/mol$$

The heat of this reaction is not sufficient to give quantitative conversion. The *hot-wall reactor* is therefore provided with external heating to give a wall temperature of 630 °C. The *cold-wall reactor* obtains its energy from the highly exothermic reaction of fluorine (mixed directly with UF_6) with excess hydrogen.

Reduction with Tetrachloroethylene. Reaction of UF_6 with tetrachloroethylene (C_2Cl_4) in a fluidized-bed furnace was developed by United Nuclear Corp. [61]:

$$UF_6 + Cl_2C{=}CCl_2 \rightarrow UF_4 + Cl_2FC{-}CFCl_2$$

The reaction temperature is 260 – 295 °C. Nitrogen is added to tetrachloroethylene to act as a carrier gas.

6.3.11.2. Reduction of UF_4 to Uranium Metal

Calcium, magnesium, and sodium are thermodynamically suitable agents for the reduction of UF_4 to uranium metal, but only calcium and magnesium are of practical importance – calcium for small batches because of its higher heat of reaction, and magnesium for larger quantities because it is the cheaper metal.

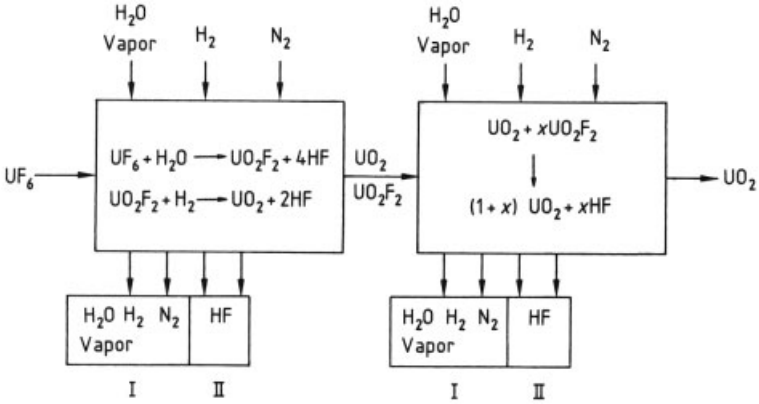

Figure 41. From UF_6 to UO_2 by the DC process
I = Gas stream is filtered and recycled to the process; II = Gas stream is recovered in the form of pure hydrofluoric acid; carrier gas is passed over $CaCO_3$ to remove residual HF, and H_2 is flared off.

Reduction with Calcium. The heat of reaction of reduction with calcium is very large:

$$UF_4 + 2\,Ca \rightarrow 2\,CaF_2 + U \quad \Delta H = -577.8\,kJ/mol$$

With a calcium excess of 15 % and adiabatic conditions during reduction, the temperature of the reaction mixture can rise to 1800 °C. This temperature is high enough to melt the slag formed (*mp* of CaF_2 1418 °C). Calcium granules (ca. 5 mm) are used, which must be very pure. High magnesium content may cause a violent reaction because the reaction temperature is higher than the boiling point of magnesium.

A mixture of UF_4 and calcium is charged into a steel reaction vessel whose inner walls are lined with high-purity CaF_2 (Fig. 42). This interior lining prevents the container material from reacting with molten uranium metal. Before reduction, the reaction apparatus is evacuated and filled with argon. The reaction mixture is then ignited electrically with an ignition pellet, and the reaction is complete after a few seconds. Molten uranium collects at the bottom of the reaction vessel and can be separated easily from the slag after the vessel is cooled. The yield of metal is ca. 98 – 99 % of theory.

Reduction with Magnesium. Uranium reduction on a larger scale (several hundred kilograms of metal) is usually carried out with the much cheaper magnesium. Also, since its relative atomic mass is only about half that of calcium, the amount of magnesium required is smaller. However, the price advantage is counterbalanced partly by the lower heat of reaction:

$$UF_4 + 2\,Mg \rightarrow U + 2\,MgF_2 \quad \Delta H = -349.6\,kJ/mol$$

which makes it necessary to heat the reaction mixture to ca. 600 °C. If the mixture is at room temperature when it is ignited, the melting point of uranium metal and of the MgF_2 slag (1263 °C) is not reached. By using an excess of 15 % magnesium, the temperature reached is only ca. 900 °C under adiabatic conditions.

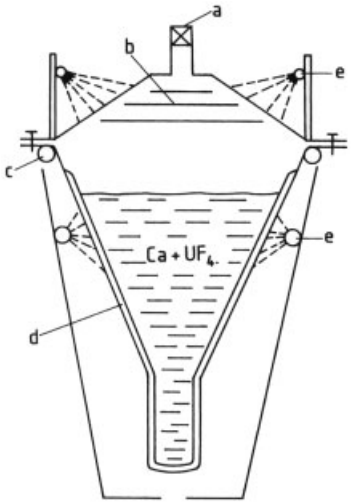

Figure 42. Reactor for reduction of UF_4 with calcium
a) Pressure control valve; b) Heat protection shields; c) Circular water pipe for flange cooling; d) CaF_2 lining; e) Circular pipe for water jets

The interior walls of the reactor are often lined with dolomite. To obtain a good yield, the bulk density of UF_4 must be relatively high (>3 g/cm^3), giving good thermal conductivity. Briquetting the reaction mixture before loading it into the reactor vessel is therefore advantageous.

6.3.11.3. Production of Uranium Powder

Uranium powder can be obtained by reduction of UO_2, electrolysis of fused salts, or decomposition of uranium hydride.

Reduction of UO_2. Reduction of UO_2 with calcium or magnesium has not become widely established in industry because separation of the product from the slag is difficult. Also, the demand for uranium powder is limited.

Reduction with Calcium. The best reducing agent for UO_2 is calcium. A carefully prepared mixture of calcium granules and UO_2 is charged to a reactor whose walls are lined with CaO. The mixture is heated to 1200 °C under argon and kept at this temperature for at least 30 min. The temperature should reach the melting point of uranium (1132 °C) but should not cause CaO to melt. The uranium particles therefore become embedded in unmelted CaO. Excess calcium, which has low solubility in CaO, acts as a grain growth accelerator for the uranium.

After reaction, the mixture, which contains uranium and CaO together with excess calcium, is broken up and slurried with water. After careful addition of acetic acid (to buffer the solution at pH 5.5 by formation of calcium acetate), dilute nitric acid is added to leach out calcium. A temperature of 30 °C should not be exceeded in this leaching process because higher temperatures cause oxidation of the surface of the metal powder.

Reduction with Magnesium. The temperature during reduction with magnesium must be kept below 1270 °C; however, it should be above 1133 °C, which is higher than the *bp* of magnesium, so that large enough metal particles will be formed. The metal powder produced by reduction with magnesium is much finer than that obtained by reduction with calcium. Magnesium chloride is often added to the reaction mixture to act as a flux.

Electrolysis of Fused Salts. The electrowinning of metal powder is carried out in fused fluoride or chloride salts. KUF_5 or UF_4 is dissolved in a molten salt mixture of 80 % $CaCl_2$ and 20 % NaCl, and electrolyzed at 900 °C in a graphite crucible, which forms the anode. When withdrawing the cathode from the molten salt, appropriate precautions must be taken to prevent burning of powdered metal that adheres to it. After cooling, soluble chlorides can be removed from the metal by washing with water.

Decomposition of Uranium Hydride. Hydrogen reacts very readily with uranium metal, forming uranium hydride (UH_3). The optimum hydrogenation temperature is 220 °C. Because of the large density difference between the hydride (10.8 g/cm^3) and the metal (19.06 g/cm^3), the brittle hydride breaks down during the reaction into a fine pyrophoric powder. This is decomposed thermally at 310 °C with application of a vacuum to draw off hydrogen. The metal powder so obtained is strongly pyrophoric and can be handled only under an inert gas atmosphere.

7. Uranium Alloys

7.1. Classification

Light-water reactors, in which the nuclear fuel is UO_2, now dominate the power generation market, so uranium metal and uranium alloys have lost much of their industrial importance, although they are still used in materials testing reactors (MTRs), high-flux reactors, submarine reactors, and as shielding materials for radiation sources and irradiation equipment. Thus, only a small number of uranium alloys are still of interest. These can be classified as follows.

Metallurgical Mixtures of Mutually Insoluble Metal Phases. In this group of alloys, the components are mutually insoluble in the molten state (if solubilities in the part-per-million range are ignored) and form neither solid solutions nor mixed phases. Metals of technical importance include calcium and magnesium, which are used for reduction of UF_4. Reduction with calcium gives uranium metal containing <20 ppm Ca. This group also includes thorium, which has

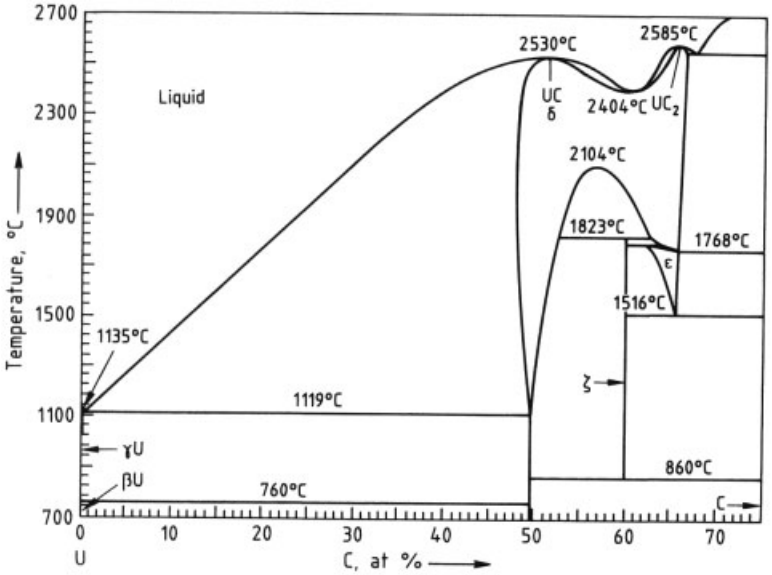

Figure 43. U – C phase diagram [12]

a maximum solubility of 0.3 % in uranium at 900 °C.

Metals That Form Definite Intermetallic Phases with Uranium. These elements form definite intermetallic phases, but not solid solutions, with uranium. The only member of the group of current interest is carbon. Zirconium is of minor importance. The U – C phase diagram (Fig. 43) shows various uranium carbides that could be of importance as breeder reactor fuels. Properties and production of uranium carbides are described in Section 8.2. Even low carbon contents in uranium (<500 ppm) cause the metal to be hard and brittle. Machining of uranium containing >0.2 % C is practically impossible.

Metals That Form Solid Solutions. These metals are more or less mutually soluble but do not form solid intermetallic phases. They include chromium, silver, lead, molybdenum, niobium, titanium, tantalum, vanadium, and tungsten. Only molybdenum is of industrial importance since additions of 2 – 8 wt % molybdenum to uranium impart improved corrosion resistance. Uranium – molybdenum alloys can be used, therefore, as shielding materials for medical radiation equipment and in aircraft trimming weights, items that should not corrode in air.

Elements That Form Both Solid Solutions and Intermetallic Phases. These include aluminum, silicon, nitrogen, iron, gold, beryllium, manganese, nickel, and hydrogen. The most important metal in this group is aluminum because Al – U alloys can be used in special applications such as materials testing and high-flux reactors. The Al – U phase diagram is given in Figure 44. Work published in the late 1980s on the phase diagram gives slightly different results for the solidus – liquidus transition in medium-concentration regions [62].

7.2. Production of Important Alloys

U – Zr alloys are produced from the elements in metallic form. The best process is arc melting with consumable electrodes [63]. A homogeneous alloy is obtained by remelting the product two or three times by the same method. Final processing is by hot pressing, cutting, or machining.

U – Zr hydride is produced by hydrogenation of the U – Zr alloy at 300 – 900 °C. Since the absorption of hydrogen causes swelling, the process must occur slowly enough to ensure

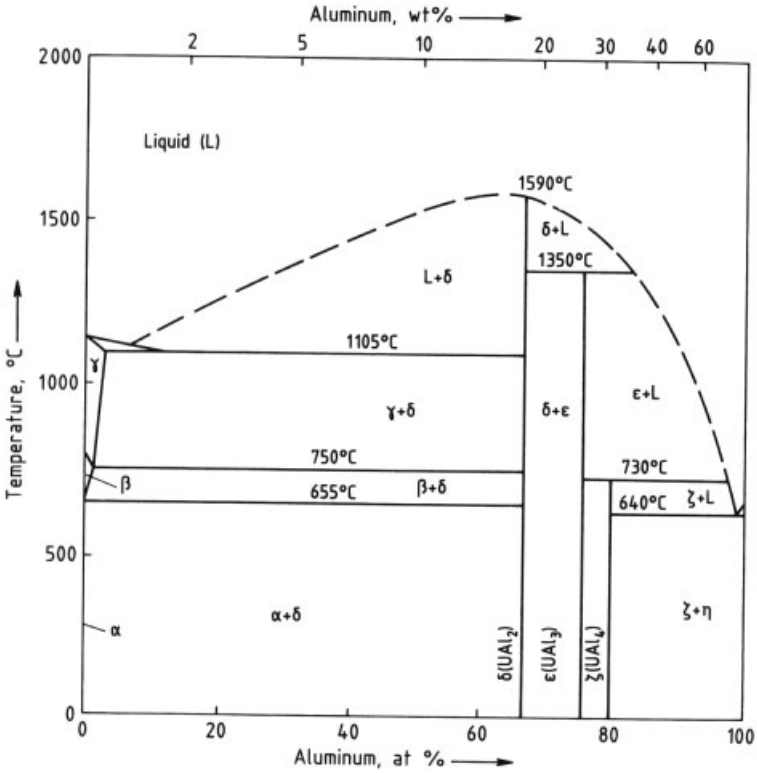

Figure 44. U – Al phase diagram

that the stresses produced relax without causing cracking [63]. To be able to control the hydrogenation process by means of the H_2 partial pressure, detailed knowledge of the U – Zr – H system is necessary, including the phase diagram and the H_2 partial pressure of the hydrides ZrH_x [64]. Also, the level of impurities such as iron must be kept as low as possible because iron melts out locally as a low-melting U – Fe alloy during conversion of the U – Zr alloy into U – Zr hydride.

Uranium Silicides. For production of U_7Si, a mixture of uranium and silicon metal in the correct proportions is melted in a lined graphite crucible at 1640 – 1700 °C for ca. 10 min. The material is then cast in a graphite mold to form rods. The cast structure consists of ϵ-U_3Si_2 in a uranium matrix. The precipitated grains of ϵ-U_3Si_2 are surrounded by a layer of δ-U_3Si. The size and shape of the grains of ϵ-U_3Si_2 depend on cooling conditions. Hollow pellets can be prepared by casting in a suitable mold, although the material shrinks onto the core and the latter is then difficult to remove. A better method seems to be to drill a solid rod with a deep hole drilling machine.

To convert the structure of the solid into a δ-U_3Si phase that is as homogeneous as possible, it is annealed at 800 – 900 °C. The transformation is diffusion controlled and depends strongly on the form of the primary ϵ-U_3Si_2 precipitate in the cast structure. Optimum homogenization condition involves annealing at 800 – 820 °C for 100 h [65].

U – Al Alloys. Aluminum alloys containing 13 – 24 % uranium are used in the production of fuel elements for materials testing reactors and high-flux reactors, or the compound UAl_3 to obtain higher uranium contents. Whereas the alloy can be melted and cast, production of UAl_3 leads to brittle button-like formations. These can be pulverized easily, and the powder is then pressed and sintered to form the fuel rod.

8. Uranium Compounds

8.1. Halides

The most important halides (see Table 10) are the fluorides; polymers such as U_4F_{17} and U_5F_{22} are not included in this table because they have little technical importance.

Table 10. Halogen compounds of uranium (excluding polymeric compounds)

		Structure
Fluorides	UF_3	hexagonal, LaF_3 type
	UF_4	triclinic (pseudomonoclinic), isomorphous with ThF_4
	U_2F_9	cubic
	α-UF_5	tetragonal, chain structure
	β-UF_5	tetragonal, three-dimensional lattice
	UF_6	rhombic, molecular lattice
Chlorides	UCl_3	hexagonal, $La(OH)_3$ type
	UCl_4	tetragonal
	UCl_6	hexagonal, molecular lattice
Bromides and iodides	UBr_3	hexagonal, $La(OH)_3$ type
	UI_3	rhombic, LaI_3 type

8.1.1. Trivalent Halides

In general, trivalent halides may be prepared by reduction of the tetravalent compounds with hydrogen, e.g., [66]:

$$2\,UF_4 + H_2 \rightarrow 2\,UF_3 + 2\,HF$$

At 950 °C all of the UF_4 does not react; above 1080 °C, disproportionation of UF_3 into UF_4 and uranium occurs. Preparation of chlorides, bromides, and iodides involves similar problems.

Uranium trifluoride, UF_3, forms red-violet crystals. It is stable to moist air at room temperature, but is converted to oxide fluorides on heating and is oxidized to U_3O_8 above 900 °C. It is insoluble in cold water but is slowly oxidized to UF_4. In boiling water, further oxidization to uranyl fluoride occurs.

Uranium trichloride, UCl_3, forms dark red, very hygroscopic needles, *mp* 842 °C. In aqueous HCl solution, rapid transformation into a tetravalent green compound occurs, with liberation of hydrogen according to:

$$2\,UCl_3 + 2\,HCl \rightarrow 2\,UCl_4 + H_2$$

Uranium Tribromide and Triiodide. Uranium tribromide, UBr_3, *mp* 730 °C, and uranium triiodide, UI_3, *mp* 506 °C, are formed similarly to UF_3, and are also soluble in water.

8.1.2. Uranium Tetrahalides

Uranium tetrafluoride is a very important intermediate in the production of UF_6 for enrichment plants (see Section 6.2.2) and for production of the metal. It is a highly polymeric green solid that melts at 1034 [67] to 1036 °C [68]. Gaseous UF_4 is monomeric. The stability of solid UF_4 is explained by its complex crystal lattice (α-ZrF_4 structure). However, oxidation occurs in a stream of oxygen at 700 – 850 °C:

$$2\,UF_4 + O_2 \rightarrow UF_6 + UO_2F_2$$

Uranium tetrafluoride is almost insoluble in water and is dissolved only by oxidizing acids and some concentrated acids.

Hydrated UF_4 can also be prepared by precipitation from U(IV) solutions with hydrofluoric acid. At room temperature, the very pale green $UF_4 \cdot 2.5\,H_2O$ [69] is formed, but at 80 – 100 °C the grass green compound $UF_4 \cdot 0.75\,H_2O$ is obtained. Both can be converted into anhydrous UF_4 by careful dehydration in a vacuum.

Other Halides. Uranium tetrachloride (*mp* 569 °C), UBr_4 (*mp* 518 °C), and UI_4 (*mp* 706 °C) can be produced by heating a mixture of uranium oxides and carbon in a stream of halogen. Uranium tetrachloride dissolves very readily in water, with evolution of heat; UBr_4 and UI_4 behave similarly.

8.1.3. Uranium Pentafluoride

The reaction of UF_4 with UF_6 produces UF_5, which is formed together with U_2F_9 and U_4F_{17} as an intermediate in the production of UF_6 (see Section 6.3.7). UF_5 is now also important as an intermediate product formed during uranium enrichment using lasers (see Section 6.3.9.5). Other pentahalides are known but are of only theoretical interest.

8.1.4. Uranium Hexafluoride

Many possible methods can be used to produce UF_6, the more important are listed below:

$$U + 3\,F_2 \rightarrow UF_6 \text{ (hazardousreaction)}$$

$$UF_4 + F_2 \rightarrow UF_6 \text{ (at } 340^\circ\text{C)}, 50\%\ F_2 \text{ excess}$$

$$UO_2 + 3\,F_2 \rightarrow UF_6 + O_2 \text{ (at } 500^\circ\text{C)}$$

$$UO_3 + 3\,F_2 \rightarrow UF_6 + \frac{3}{2}O_2 \text{ (at } 400^\circ\text{C)}$$

$$U_3O_8 + 17\,F_2 \rightarrow 3\,UF_6 + 8\,F_2O \text{ (at } 360^\circ\text{C)}$$

$$UC_2 + 7\,F_2 \rightarrow UF_6 + 2\,CF_4 \text{ (at } 350^\circ\text{C)}$$

$$UO_2 + 4\,HF \rightarrow UF_4 + 2\,H_2O \text{ (at } 550^\circ\text{C)}$$

$$UF_4 + F_2 \rightarrow UF_6$$

$$U + 2\,BrF_3 \rightarrow UF_6 + Br_2 \text{ (at } 50-125^\circ\text{C)}$$

The only industrial route is from pure UF_4. The phase diagram of UF_6 is shown in Figure 45. At normal pressure, UF_6 sublimes at 56.54 °C and melts only under a pressure of 1.517×10^5 Pa at 64.05 °C.

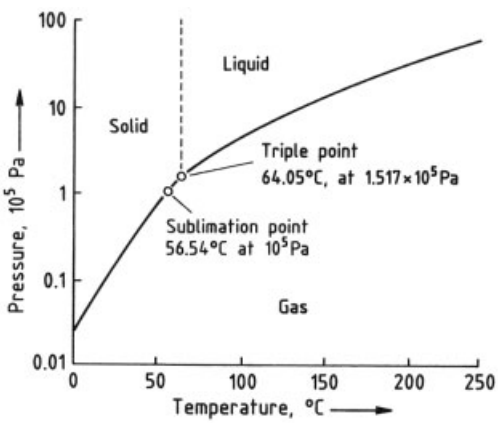

Figure 45. Phase diagram of UF_6 [70]

The critical constants for UF_6 are [70] T_c, 230.5 ± 0.2 °C; p_c, $45.50 \pm 0.50 \times 10^5$ Pa; ϱ_c, 1.375 g/m^3; V_c, 0.256 L/mol.

Solid UF_6 has similarities to white phosphorus, having a colorless to yellowish appearance and being easily cut into pieces. However, in contrast to phosphorus, it reacts rapidly with moisture, hydrolyzing even in moist air to form the bluish compound UO_2F_2, which resembles cigarette smoke:

$$UF_6 + 2\,H_2O \rightarrow UO_2F_2 + 4\,HF$$

This fine bluish smoke enables the smallest UF_6 leak in pipework to be detected immediately. At increased temperatures, UF_6 gas reacts violently or explosively with organic substances, but metal and glass are not attacked in the absence of moisture. Comprehensive data on UF_6 can be found in [70].

8.2. Carbides

The three carbides UC, U_2C_3, and UC_2 appear in the U – C phase diagram (Fig. 43) [71]. Both UC and UC_2 are stable up to their melting points; U_2C_3 is an intermediate phase that disproportionates above 1800 °C into UC and UC_2 [72]. The carbide UC_2 is unstable below 1275 –1500 °C, disproportionating to form U_2C_3 and carbon [73]. Above 1800 °C, UC_2 has the composition U/C = 1.86 ± 0.02 (atomic ratio) and a cubic structure, whereas a tetragonal form exists between 1600 and 1800 °C with composition U/C = 1.96 ± 0.04. Various temperatures are quoted for the melting point. According to [74], the *mp* of UC is ca. 2350 °C, and that of UC_2 with the composition U/C = 1.9 is 2475 °C.

The carbides react very vigorously with water and are therefore not used in water-cooled nuclear reactors because secondary damage could occur in the event of a defective fuel rod. However, they can be used as nuclear fuel if cooling is by gas or sodium, two generations of reactors that could be important in the future. Carbide fuels are therefore not yet available on the market except for testing purposes. In experimental fuel elements, UC is used mainly in the form of pellets, rods, or coated particles.

Uranium carbide is produced by three main methods: (1) reaction of uranium metal with carbon; (2) reaction of uranium metal with alkanes; and (3) reaction of uranium oxides with carbon.

Reaction of Uranium Metal with Carbon. Stoichiometric UC is obtained when a mixture of uranium powder with carbon is compressed at 800 – 1000 °C [75]. The material so produced has a density of 98 % of the theoretical value and a Vickers hardness of 700 kg/mm^2, a figure normally only obtainable with UC sintered at 2000 °C. A somewhat lower-density UC is formed by compressing uranium powder with carbon at 700 – 800 °C and then sintering the pellets at 1100 °C [76].

Very pure and dense UC is obtained by arc melting of uranium and carbon under vacuum or argon. This method gives densities of 98 – 99.5 % of theory [77].

Reaction of Uranium Metal with Alkanes. Very fine uranium powder (produced by decomposition of uranium hydride) is reacted with methane at 600 – 900 °C to give very pure and dense UC [78]. Reaction times are 0.5 – 2 h. Propane has also been used successfully.

Reaction of Uranium Oxides with Carbon. This is the only reaction of industrial importance. The following reactions can occur:

$$U_3O_8 + 11\,C \rightarrow 3\,UC + 8\,CO \quad \text{(at 1800 °C)}$$
$$U_3O_8 + 14\,C \rightarrow 3\,UC_2 + 8\,CO \quad \text{(at 2400 °C)}$$
$$UO_2 + 3\,C \rightarrow UC + 2\,CO \quad \text{(at 1800 – 1900 °C)}$$

Nuclear fuels are produced by this method in a vacuum furnace or under a protective gas since UC is pyrophoric.

Production of UC Fuel Rods. Uranium carbide rods are produced by melting UC at ca. 2500 °C and casting in rod-shaped molds [79]. Of the possible melting processes (i.e., electron beam melting, induction melting, and arc melting), only the last has been developed to an industrial scale. Induction melting is particularly unsuitable because reaction with the walls of the graphite crucible occurs, converting UC to UC_2.

The process of melting UC in an electric arc furnace is known as skull melting; a solid shell of UC forms between the water-cooled copper crucible and the melt, preventing molten UC from reacting with the metal crucible. This ensures the required purity of the nuclear fuel. To be able to cast the melt, tilting furnaces or centrifugal casting furnaces are used.

8.3. Nitrides

The nitrides UN, U_2N_3, and UN_2 are analogous to the carbides. The mononitride has a theoretical density of 14.32 g/cm^3 [80]. Under N_2 at normal pressure (100 kPa), UN decomposes at 2800 °C in a uranium-rich melt with liberation of N_2. If the N_2 pressure is reduced to 10 kPa, the decomposition temperature decreases to 2080 °C. Above 250 kPa, UN melts without decomposition [81]. At nitrogen pressures above atmospheric and 1300 °C, U_2N_3 disproportionates into UN + UN_2 with a 13 % decrease in volume [82].

The *mononitride* is produced by three main methods:

1) Reaction of uranium metal with nitrogen
2) Reaction of UH_3 with nitrogen
3) Reaction of uranium oxide with carbon and nitrogen

The equations for these reactions are as follows [83]:

$$1)\ 2U + 3/2\,N_2 \xrightarrow{850\ °C} U_2N_3 \xrightarrow[\text{Vacuum}]{1300\ °C} 2UN + 1/2\,N_2$$

$$2)\ U + 1.5\,H_2 \xrightarrow{250\ °C} UH_3 \xrightarrow[+N_2]{800\ °C}$$
$$U_2N_3 \xrightarrow[\text{Vacuum}]{1300\ °C} 2UN + 1/2N$$

$$U + NH_3 \xrightarrow{300\ °C} UH_3 \xrightarrow[+N_2]{800\ °C}$$
$$U_2N_3 \xrightarrow[\text{Vacuum}]{1300\ °C} 2UN + 1/2N$$

$$3)\ 2UO_2 + 4C + N_2 \xrightarrow[1850\ °C]{1400-} 2UN + 4CO$$

If method 3 is used, the nitride always contains some carbon and oxygen [83].

The nitrides have not yet become important nuclear fuels for power reactors, but the properties of uranium nitrides have been investigated again in detail with the aim of using them as nuclear fuels for special reactors (e.g., in space applications) [84 – 87].

8.4. Oxides

The following oxides of uranium are known:

1) Uranium monoxide, UO, discovered by X-rays, of no industrial importance
2) Uranium dioxide, UO_2, an important reactor fuel
3) $UO_{2.3}$, stable only above 500 °C, tetragonal crystal structure
4) $UO_{2.33}$ – U_3O_7, tetragonal crystal structure
5) U_4O_9, formed as an intermediate from UO_2 at high temperature
6) U_2O_5, monoclinic crystal lattice, red-brown
7) U_3O_8, stable in all temperature regions ($U_2O_5 \cdot UO_3$), black
8) UO_3, stable up to ca. 300 °C, orange
9) Uranium peroxide, UO_4, light yellow

8.4.1. Uranium Dioxide

Production of the industrially important compound uranium dioxide is described in Section 6.3.5.3. Important physical properties are

Theoretical density	10.96 g/cm^3
Melting point	2860 ± 70.5 °C
Thermal conductivity	
at 500 °C	4.3 W m^{-1} K^{-1}
at 1100 °C	2.6 W m^{-1} K^{-1}
at 1650 °C	2.15 W m^{-1} K^{-1}
at 2200 °C	4.3 W m^{-1} K^{-1}
Linear coefficient of expansion (4–2800 °C)	1.1×10^{-5} K^{-1}
Specific heat capacity (32 – 315 °C)	237.4 J kg^{-1} K^{-1}
Electrical conductivity	
at 20 °C	5.6×10^{-3} Ω^{-1} m^{-1}
at 975 °C	3.0 Ω^{-1} m^{-1}
at 1790 °C	270 Ω^{-1} m^{-1}

The physical properties, such as thermal and electrical conductivities, melting point, etc., are strongly dependent on stoichiometry and possibly on impurities. At ambient temperature UO_2 is a p-type semiconductor. The crystal lattice of UO_2 is face-centered cubic of the CaF_2 type. Stoichiometric UO_2 powder is pyrophoric and is oxidized in air to intermediate products ($UO_{2.30}$ and $UO_{2.33}$) or to U_3O_8. A lattice parameter of 0.547 nm can be given for stoichiometric UO_2.

UO_2 powder with a large specific surface area is self-igniting. Therefore, handling of UO_2 powders with BET surfaces $\geq$2 m^2/g requires special precaution measures, e.g., surface oxidation in a controlled manner.

8.4.2. Uranium Trioxide

Normal decomposition at 300 °C of intermediate products obtained during uranium production always leads to orange-colored UO_3. At >400 °C, this begins to change into black U_3O_8. Uranium trioxide occurs in six forms:

- α-UO_3 is formed by careful decomposition of UO_4 at 450 – 500 °C [88].
- β-UO_3 is formed by slow decomposition of ammonium diuranate at 450 – 500 °C [89].
- γ-UO_3 is formed by slow decomposition of uranyl nitrate hexahydrate (e.g., 4 h at 550 °C) [90].
- δ-UO_3 is red and is formed by dehydration of $UO_3 \cdot H_2O$ in air at 375 °C [89].
- ϵ-UO_3 is brick red and is formed in a few minutes by oxidation of U_3O_8 with NO_2 at 350 °C [89].
- ζ-UO_3 is a brown powder produced by rapid decomposition of uranyl nitrate hexahydrate in the presence of 0.6 % sulfamic acid [91]. The existence of this compound is not yet certain.

Uranium trioxide can be an intermediate in the production of nuclear fuel, but it is not itself used.

8.4.3. Triuranium Octaoxide

The polymorphism of U_3O_8 is extremely complex, and the data are to some extent conflicting. A summary is given below:

At normal pressure, four modifications exist denoted by α, α', α'', and β. Of these, α and β are base-centered rhombic, α' is body-centered rhombic, and α'' is hexagonal. The α-forms are stable at low temperature (room temperature) and the β-form at high temperature (>730 °C). At 50 °C, the α-form changes to the α'-form [92]. The α-form changes to the β-form at high pressure or under shear forces at 1100 – 1500 °C in the presence of atmospheric oxygen [93]. However, β-U_3O_8 has a lower density and changes back to α-U_3O_8 at 50 °C. Therefore, the reported phase transitions depend on temperature, pressure, and redox reactions, which

cause changes in density. The interrelationships between the different phases are thus very complex.

Another U_3O_8 (or $UO_{2.67}$) phase has been described, which is formed only at high pressure at ca. 627 °C and has a face-centered cubic structure. The combination of pressure and shear stress is said to be important for formation of this phase [92]. Thus, cubic U_3O_8 is obtained from α-U_3O_8 by prolonged milling in a ball mill, conversion being almost complete after one week. Considerable amounts of water are taken up during milling. Therefore, to obtain a pure product, α-U_3O_8 is first milled for 72 h; the mixture of α-U_3O_8 and cubic U_3O_8 obtained is heated to 750 °C; this is then milled again for 24 h in the absence of moisture. After milling for another 30 h in contact with atmospheric oxygen, an amorphous variety of α-UO_3 is formed.

Of four other possible modifications, the existence of two is fairly well established: γ-U_3O_8 is said to be formed at 940 °C from β-U_3O_8 with a heat of transformation of 4.63 ± 0.06 kJ/mol [94]. At 120 °C, this changes back irreversibly to α-U_3O_8, the heat of transformation being 4.40 ± 0.21 kJ/mol. The δ-U_3O_8 modification is formed above 1050 °C [95]. It crystallizes in a rhombic structure with unit cell dimensions α = 0.670 nm, b = 1.246 nm, and c = 0.851 nm. The density is said to be 7.86 g/cm^3.

Oxygen is evolved when U_3O_8 melts, so the melting point cannot be determined; U_3O_8 vaporizes at 925 – 1525 °C with formation of mainly monomeric gaseous UO_3 [96].

The specific *magnetic susceptibility* of U_3O_8 does not obey the Curie law or the Curie – Weiss law; this indicates that U_3O_8 should be written as $U_2O_5 \cdot UO_3$. The *electrical conductivity* of $U_3O_{8\pm x}$ depends strongly on temperature and the O : U ratio, so a generally valid value cannot be given. However, at room temperature stoichiometric U_3O_8 exhibits an electrical conductivity of about 100 $\Omega^{-1}m^{-1}$. The oxide U_3O_8 is an *n*-type semiconductor. Conduction is thought to occur by electrons and also by holes. Because of its mixed oxide properties, electron transport from U(V) to U(VI) is assumed.

Chemical Properties. Triuranium octaoxide dissolves readily in oxidizing acids and is oxidized to the hexavalent state and hydrolyzed to uranyl salts [e.g., $UO_2(NO_3)_2$] in the aqueous phase. Uranyl chloride, UO_2Cl_2, is formed by passing HCl gas over U_3O_8 at 700 °C. Triuranium octaoxide is insoluble in water and is oxidized rapidly to UO_4 (uranium peroxide) by hydrogen peroxide; it is reduced by hydrogen or ammonia to UO_2 at >500 °C.

8.4.4. Peroxides

Of the uranium oxides containing a peroxide group, the existence of two is well established: UO_4 (crystallizing with two or four molecules of water) and U_2O_7 with or without water of crystallization.

The hydrate $UO_4 \cdot x\,H_2O$ can be precipitated easily from aqueous uranyl nitrate solutions at room temperature by addition of H_2O_2:

$$UO_2(NO_3)_2 + H_2O_2 + x\,H_2O \rightarrow UO_4 \cdot x\,H_2O + 2\,HNO_3$$

This precipitation easily can be performed continuously, and the precipitate can be dried to give a powder with good flow properties [97].

A UO_4 powder with good sintering properties is prepared from a solution of uranyl nitrate (90 – 120 g/L) in 0.5 mol/L HNO_3 to which ammonium nitrate (80 g/L) has been added. Precipitation at pH 1.5 and 40 °C occurs at a uranium : hydrogen peroxide ratio of 1 : 3. The UO_4 powder formed is a spheroidal to spherical agglomerate with a maximum particle size of 50 µm, although most particles are ca. 20 µm. Calcination of this powder gives a product with good sintering properties.

The hydrate $UO_4 \cdot x\,H_2O$ is pale yellow; $UO_4 \cdot 4\,H_2O$ is the main product at a precipitation temperature of <50 °C, $UO_4 \cdot 2\,H_2O$ at higher temperature. Careful dehydration in a vacuum at ca. 130 °C gives hydrated U_2O_7, which decomposes with liberation of oxygen above 150 °C.

The structure of $UO_4 \cdot x\,H_2O$ has not yet been established, but U_2O_7 probably has the following structure [98]:

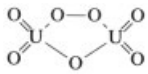

This is a brick-red to orange hygroscopic substance [99].

8.5. Nitrates, Sulfates, and the Carbonato Complex

8.5.1. Nitrates

Uranium and its oxides dissolve in nitric acid to form uranyl nitrate:

$$UO_3 + 2\,HNO_3 \rightarrow UO_2(NO_3)_2 + H_2O$$

The hexahydrate crystallizes from weakly acidic supersaturated solutions, but tri- and dihydrates can be obtained at higher concentrations of HNO_3 (see Fig. 46). Upon evaporation of uranyl nitrate solutions to >450 g/L uranium, $UO_2(NO_3)_2 \cdot 6\,H_2O$ (*mp* 60 °C) crystallizes on cooling [100]. The uranium content of a solution boiling at 120 °C is ca. 1000 g/L. The trihydrate melts at 113 °C, and the dihydrate at 184 °C. No decomposition occurs up to this temperature, but denitration begins above 300 °C. The UO_3 so produced has a relatively high bulk density (3.5 – 4.3 g/cm^3) and, after reduction to UO_2, is not very suitable for the production of compressed and sintered pellets. The nitrate of U(IV), $U(NO_3)_4$, can be obtained by treatment of uranyl nitrate with reducing agents such as hydrazine, iron(II) formate, or SO_2, or by electrolytic reduction. Uranium(IV) nitrate is a green substance, in marked contrast to yellow $UO_2(NO_3)_2$. It oxidizes readily to U(VI) in air.

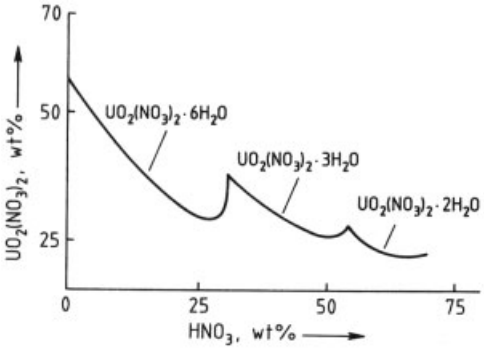

Figure 46. The $UO_2(NO_3)_2$ – HNO_3 – H_2O system

8.5.2. Sulfates

The sulfates of uranium are important because uranium ores are usually digested with sulfuric acid (see Section 6.3.1), which generally leads to the formation of uranyl sulfate, UO_2SO_4. The nature of the hydrates formed depends on the H_2SO_4 concentration (see Fig. 47). Sulfates of tetravalent uranium are more stable than nitrates. The following hydrates are known: $U(SO_4)_2 \cdot 8\,H_2O$, $U(SO_4)_2 \cdot 4\,H_2O$, $U(SO_4)_2 \cdot 9\,H_2O$, and $U(SO_4)_2 \cdot 2\,H_2O$.

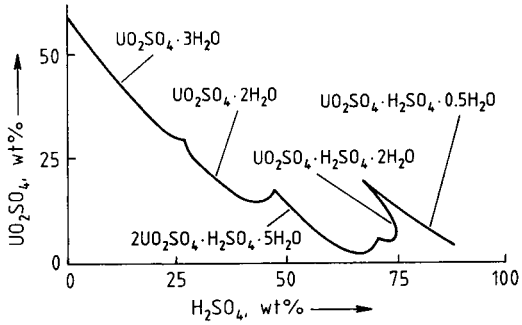

Figure 47. The UO_2SO_4 – H_2SO_4 – H_2O system

8.5.3. Tricarbonatodioxouranate

Tricarbonatodioxouranate, $[UO_2(CO_3)_3]^{4-}$, is formed by alkaline digestion and in the AUC process. The ammonium salt of this anion (AUC) decomposes above 80 °C, liberating CO_2 and forming ammonium diuranate.

9. Safety

The normal safety precautions taken in the chemical industry when handling toxic, liquid, volatile, and gaseous substances must be observed also in the case of uranium, but in addition radiation protection and criticality safety must be included. These affect plant and equipment design and may necessitate the use of apparatus with dimensions not generally employed in chemical technology (see Section 9.3), but reflecting all considerations as regards chemical, nuclear, and radiological safety.

9.1. Radiation Shielding

Handling natural uranium that has not been irradiated does not present special problems.

Table 11. Maximum permitted uranium concentrations in the workplace (with figures for other toxic elements for comparison) [101]

	Outside atmosphere, μg/m^3	In the workplace	
		Inactive area, μg/cm^3	Active area *, μg/cm^3
Natural uranium	180	150	1500
Uranium with 5 % ^{235}U	42	35	350
Uranium with 10 % ^{235}U	24	20	200
HF	2000		
Pb	200		
Be	2.0		
^{239}Pu	0.000032	0.00017	0.0017

* Workplace where radioactive materials are processed.

The danger of unacceptable emission of radiation is generally small; nevertheless, all processes are generally emboxed. Special attention must be given to cleaning UF_6 cylinders because buildup of ^{234}Th, which is in equilibrium with ^{238}U, can occur. After evaporation of UF_6, the cylinder still contains residues in the form of uranium decay products. However, providing shielding against the intense β-radiation and bremsstrahlung emitted by these residues while they are being treated is not difficult.

In addition, the uranium content of the atmosphere and contamination of the workplace are measured and controlled to protect personnel from uptake of or contamination by uranium. If legal limits are exceeded, respiratory protection must be provided. Workers must also be examined with a lung monitor at regular intervals, and the uranium content of urine must be determined. If uptake of uranium is suspected, whole-body measurements must be carried out.

Maximum permitted uranium concentrations in air at the workplace and in the form of surface contamination are given in Table 11.

9.2. Safety Against Uncontrolled Criticality

Criticality safety includes all technical and organizational measures taken to prevent an uncontrolled chain reaction. Uranium with its natural isotopic composition becomes critical only under special reactor conditions. This means that a spontaneous chain reaction under the conditions of uranium processing is impossible. However, in contrast to uranium with a natural isotopic composition, enriched uranium can have a critical configuration under considerably simpler circumstances (e.g., during manufacture or storage of fuel elements). The critical mass depends mainly on the degree of enrichment of ^{235}U, the composition of the material, and the moderating effect of the geometrical configuration.

In the Federal Republic of Germany, for example, plants for processing nuclear fuels can, therefore, be operated only if they are licensed under § 7 of the Atomic Energy Law, which specifies essential criteria for the nuclear safety of personnel and of the environment. Similar regulations apply in other countries.

In older plants, the observance of limit values was ensured mainly by measures under the control of plant personnel, but in newer plants this has been superseded by automatic control and is largely independent of human error. Exact and in licensing approved and validated computer codes enable a reduction of extra safety margins, so that larger vessels can be used with the knowledge that they are geometrically safe with respect to permitted criticality dimensions. In addition, more precise calculation of neutron interactions between individual vessels can be carried out with these methods; hence, a space-saving arrangement of components is possible.

9.3. Geometrically Safe Vessels

The nuclear safety of a critically safe (k_{eff} <0.95) thickness of material is achieved by increasing the surface of a vessel, thereby increasing the loss of neutrons to the environment and preventing the neutron population necessary to reach a critical condition. The safe geometry is determined as follows:

A hypothetical optimum neutron reflector is assumed to be present at the vessel, whether or not this assumption corresponds to actual conditions. This assumed reflector reduces calculated neutron loss considerably, and, therefore, causes a calculated increase in activity exceeding that of the actual arrangement. The smallest dimensions of the vessel in which a self-sustaining chain reaction can occur are found, based on the combined effects of concentration of fissile material, degree of moderation, and neutron loss by leakage by using this conservative method of calculation. From these conservatively smallest critical dimensions for ideally reflected uranium – water systems, the largest safe subcritical dimensions for $k_{\mathrm{eff}} = 0.95$ can be deduced, including safety factors (Table 12). Safe maximum amounts of uranium under all circumstances are listed in columns 2 and 3 of Table 12. The geometric dimensions in columns 4 – 6 apply to any amounts of uranium, provided $\varrho_{\mathrm{u}} < 3.5$ g/cm^3.

9.4. Apparatus with Heterogeneous Neutron Absorbers

The dimensions of safe vessels are often too small to allow desired throughputs to be achieved. In these cases, considerable benefit can be obtained by incorporating heterogeneous neutron absorbers (nuclear poisons), such as boron carbide or cadmium, into the equipment. Their effect is very dependent on the spatial arrangement of the no-absorption or intermediate vessel walls. For example, fluidized-bed furnaces are provided with a nuclear poison consisting of a central boron carbide rod with double cladding, and rotary UO_2 mixers with a surface covering or cladding consisting of securely attached cadmium sheets. This ensures nuclear safety even if water penetrates the fluidized bed or rotary mixer (accident safety).

9.5. Neutron Interactions in the UO_2 Fabrication Plant

A nuclear fuel production plant contains many individual processing stations. Therefore, demonstrating that these will remain subcritical as individual units is not sufficient. Equally important is ensuring that no increase in criticality is caused by interactions of the systems, plant, and equipment due to the way in which they are arranged. The entire layout of processing and storage operations must therefore be in the form of subcritical geometry.

9.6. Transportation

Plants for ore extraction, processing, and concentration are usually geographically remote from those for fuel assembly production; thus, uranium and its compounds must be transported between these plants. Since the materials being transported are often fissile and radioactive, special precautions must be taken and rules observed. For this reason, the IAEA has issued *guidelines* that are updated continuously according to the latest knowledge (Regulations for the Safe Transport of Radioactive Materials, IAEA, Vienna), in which national safety regulations are also considered. In its "Safety Series" of publications, the IAEA includes *recommendations* for methods of transporting radioactive materials, taking international and national regulations into account.

Since shipments often cross international boundaries, international harmonization of rules and licensing procedures is necessary. This is controlled under the UNO umbrella (→ Transport, Handling, and Storage, Chap. 5.1) [102]. The aim of the recommendations, guidelines, and regulations is to protect the general public, transport operatives, and private property from both the direct and the delayed effects of radiation during transportation [102].

This protection is secured by:

1) Restriction of the nature and activity of the radioactive material, which must be transported only in packaging of a specified type
2) Specification of design criteria for each type of packaging
3) Rules for handling and storage during transportation

The regulations are intended to prevent the release and uptake of radioactive materials into the human organism, to protect against radiation emitted by the materials being transported, and

Table 12. Safe masses and geometries for uranium salts and solutions as a function of degree of enrichment for a total density of <3.5 g/m^3

Degree of enrichment, % ^{235}U	Mass of ^{235}U max., kg	Total mass of U max., kg	Safe volume, L	Safe cylinder diameter, cm	Safe thickness of material, cm
100	0.350	0.35	4.4	12.70	3.81
75	0.360	0.46	5.0	13.21	4.06
50	0.390	0.76	6.0	14.48	4.83
40	0.410	1.025	6.7	15.24	5.08
30	0.440	1.467	7.7	16.00	5.59
20	0.480	2.4	9.5	17.53	6.86
15	0.520	3.467	11.0	18.80	7.87
12	0.583	4.858	12.5	19.81	8.64
10	0.600	6.0	14.0	20.83	9.14
8.0	0.6650	8.125	16.0	22.10	9.91
6.0	0.710	11.833	20.5	24.38	11.43
5.0	0.800	16.0	27.0	26.04	12.70
4.0	0.930	23.25	33.8	28.45	13.97
3.5	1.04	29.714	40.0	30.48	15.24
3.0	1.20	40.0	49.2	32.51	16.51
2.5	1.47	58.8	64.6	35.56	18.03
2.0	2.0	100.0	95.1	40.64	22.10
1.75	2.56	146.286	126	44.70	24.38

to ensure that unforeseen criticality events do not occur.

These aims are achieved in practice by:

1) Containment of the material being transported, the design and sturdiness of the packaging depending on the activity and nature of this material
2) Limitation of the maximum dose level at the outer surface of the item being transported, by taking account of the contents of the package and storage techniques used in the transport system
3) Protection of the transport container from external effects of an accident, again by taking into account the nature and activity of the material being transported
4) Prevention of a criticality event by very restrictive assumptions with respect to the items being transported and the conditions that could arise from an accident

Experience with the transport of radioactive goods has justified the practices used internationally. In industrialized Western countries, no case of injury or death has resulted due to the release of radioactive materials in a transportation accident. This is supported by data from the following countries:

United States: A detailed study of 811 reported incidents between 1971 and 1981

United Kingdom: A study of 330 000 shipments of radioisotopes

Federal Republic of Germany: A study of shipments up to 1984 [103]

The positions are underlined when assessing the *Mont-Louis* incident. In August 1984, the freighter *Mont Louis*, with a cargo of 350 t UF_6, sank in 15 m water. Within 40 days, all of the containers were recovered undamaged from the wreck without any release of radioactivity.

10. References

General References

1. *Gmelin*, Uranium, Part A: The Element, Part B: The Alloys, Part C: The Compounds, Part D: The Chemistry in Solution, Part E: The Coordination Compounds
2. Uranium Ore Processing Panel Proceedings Series, printed by IAEA (Internat. Atomic Energy Agency) Vienna, Dec. 1976.
3. S. Villani: *Uranium Enrichment, Topics in Applied Physics*, **vol. 35**, Springer Verlag, Berlin – Heidelberg – New York 1979.
4. *The Chemical Thermodynamics of Actinide Elements and Compounds*, **part I, 1976; part II, 1976; part III, 1978**, IAEA, Vienna.
5. Advances in Uranium Refining and Conversion, IAEA-TECDOC-420, Vienna 1987.
6. J. H. Cavendish in W. W. Schulz, J. D. Navratil, T. Bess (eds.): *Science and Technology of Tributyl Phosphate*, **vol. 2: "Selected Technical and Industrial Uses,"** CRC Press, Boca Raton, Fla. 1987.
7. W. D. Müller: *Geschichte der Kerntechnik in der Bundesrepublik Deutschland*, Schäffer

Verlag für Wirtschaft und Steuern, Stuttgart 1990.

Specific References

8. Int. Symposium on the Oklo-Phaenomen, Libreville, Gabon, June 23 – 27, 1975; IAEA Proc. Series – STI/PUB/405.
9. C. Keller: "Das Oklo Phänomen," *GIT Fachz. Lab.* **30** (1986) no. 5, 423 – 433; **30** (1986) no. 6, 581 – 585.
10. P. C. Sahu, M. Yousuf, K. Govinda Rajan, *Physica B + C* **183 B** (1993) 145 – 155.
11. J. F. Cannon, *J. Phys. Chem. Ref. Data* **3** (1974) 801.
12. ES Microware, TAPP, Thermochemical and Physical Properties, Hamilton, OH 1991.
13. J. D. Grogan, *J. Inst. Met.* **77** (1950) 571.
14. B. Blumenthal, *J. Nucl. Mater.* **2** (1960) 23.
15. A. N. Holden, US-Report KAPL-480 (1950).
16. R. W. Logan, *Metall. Trans. A.* **14A** (1983) 2337.
17. L. T. Lloyd, C. S. Barrett, *J. Nucl. Mater.* **18** (1966) 55.
18. M. A. Filyand, E. I. Seminova: *Handbook of the Rare Elements III*, Boston Technical Publ., Cambridge, Mass. 1970 (translated by M. E. Alferieff).
19. A. F. Hollemann, E. Wiberg; *Lehrbuch der Anorganischen Chemie,* Walter de Gruyter Verlag, Berlin 1976.
20. J. P. Coughlin, Bulletin 542 U.S. Bureau of Mines, "Contribution to the Data on Theoretical Metallurgy X, Heats and Free Energies of Formation of Inorganic Oxides," Washington D.C. 1954.
21. E. Kohl: *Uran*, Enke Verlag, Stuttgart 1954. M. Rauboult: *Géologie de l'Uranium*, Paris 1958.
22. *Gmelin*, **A1**, 75 – 77.
23. M. Benedict: *Nuclear Chemical Engineering*, McGraw-Hill, New York 1957.
24. OECD and IAEA: *Uranium, Resources, Production and Demand*, 1991.
25. OECD/NEA, *Uranium 1995 – Resources, Production and Demand,* Paris 1996.
26. The Uranium Institute Market Report, 1996.
27. A. Max, T. Mason, "Vergangenheit und Zukunft der U-Produktion," *atw* **41** (1996) no. 2, 79.
28. "Uranversorgung der Welt," *atw* **41** (1996) no. 7, 479.
29. K. Schreiber, *Stand und Perspektiven der Uranversorgung,* Lecture at Jahrestagung Kerntechnik, Mannheim, 23.5.1997.
30. H. Kirchberg: *Aufbereitung bergbaulicher Rohstoffe*, Gronau, Jena 1953.
31. E. J. Pryor: *Mineral Processing*, 3rd ed., Elsevier, Amsterdam – London – New York 1965.
32. M. E. Grimes, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **254** (1973) 312 – 318.
33. I. Higgins, *Environ. Sci. Technol.* **7** (1973) 1100 – 1140.
34. P. Pfüller, F. J. Goglke, *Aufbereit. Tech.* **14** (1973) 555 – 560.
35. G. Ritcey: "Extractive Metallurgy of Uranium," *Short Course Toronto Solvent Extraction* 1978, 23 – 31.
36. P. Niederer, *Verfahrenstechnik (Mainz)* **4** (1970) no. 12, 548 – 554.
37. J. Wyllie, *World Min.* **32** (1979) 58.
38. J. B. Rosenbaum: *Uranium Ore Processing*, IAEO Meeting, Nov. 24 – 26, 1975 Paper IAEA-AG/33–8, p. 6, Vienna 1976.
39. J. K. Dawson, E. Wait, K. Alock, D. R. Chilton, *J. Chem. Soc.* 1956, 3531 – 3540. W. Biltz, H. Müller, *Z. Anorg. Allg. Chem.* **163** (1927) 257 – 296.
40. General Electric Co., US 3 579 311, 1968. C. D. Harrington, A. E. Ruehle, *Chem. Eng. Prog.* **54** (1958) no. 3, 65 – 70. C. H. Chilton, *Chem. Eng. (N.Y.)* **65** (1958) no. 21, 138 – 141.
41. P. S. Gentils, L. H. Talley, T. J. Coloppy, *J. Inorg. Nucl. Chem.* **10** (1959) 110 – 113.
42. G. Wirths, L. Ziel, *Proc. U.N. Int. Conf. Peaceful Uses At. Energy 2nd*, **4** (1958) 16 – 21.
43. Nukem, DE 1 126 363, 1960/62; DE 1 592 477, 1966; DE 1 592 478, 1967. F. Plöger, H. Vietzke, *Chem. Ing. Tech.* **37** (1965) 692–699. K. G. Hackstein, F. Plöger, *ATW Atomwirtsch. Atomtech.* **12** (1967) 524 – 526.
44. T. J. Heal, J. E. Littlehild, H. Page, *Nucl. Eng. Int.* **25** (1980) 48 – 51.
45. A. Ducouret, B. Kalthoff, *Ber. Jahrestag. Kerntech. Kerntech. Ges. Dtsch. Atomforum* 1980, 405 – 408.
46. W. C. Ruch, D. A. Peterson, E. A. Gaskil, H. G. Tepp, *Chem. Eng. Prog. Symp. Ser.* **56** (1960) no. 28, 35 – 41.
47. F. K. Pickert, H. J. Zech: *Brennstoffkreislauf*, Deutsches Atomforum, Bonn 1981.
48. Urenco-Werkzeitschrift, Panorama, Gronau – Almelo – Capenhurst, Dec. 1980.
49. H. Mohrhauer, E. Coester, *Atomkernenerg. Kerntech.* **35** (1980) no. 3, 161 – 165.
50. W. Becker, K. Bier, *Chem. Ing. Tech.* **39** (1967) no. 1, 1 – 7.

51. P. Bley, W. Ehrfeld, U. Heiden, KFK report no. 2580, Kernforschungszentrum, Karlsruhe 1978.
52. K. Takeda, H. Onitsuka, H. Obanawa in M. Abe, T. Kataoka (eds.): *New Developments in Ion Exchange*, Kodensha Ltd., Tokyo 1991, pp. 583 – 589.
53. J. M. Lerat, C. Lorrain, *Conference on Solvent Extraction and Ion Exchange in the Nuclear Fuel Cycle*, Harwell, UK Sep. 3 – 6, 1985, pp. 53 – 61.
54. I. D. Heriot: *Uranium Enrichment by Gas Centrifuge*, Commission of the European Communities 1988.
55. N. L. Ranek et al.: Evaluation of UF_6-to-UO_2 Conversion Capability at Commercial Nuclear Fuel Facilities, ANL/EAD/TM-110, 2001.
56. H. Kühn, M. Peehs, K. Reichardt: "Thermische Brü-terstudie," part 2, *Forschungsber. BMWF Inv. Reaktor 37,* April 1969.
57. M. Peehs et al., *IAEA Tech. Committee on Utilization of Th-based Nuclear Fuel*, Vienna, Dec. 1985.
58. M. Zemek et al.: Dry Conversion Process Control and UO_2 Quality, Proceedings of the Annual Meeting on Nuclear Technology, Düsseldorf, 25 – 27 May 2004, pp. 213–216.
59. M. Peehs, W. Dörr, M. Hrovat, *IAEA Advisory Group Meeting on Advanced Fuel Technology and Performance*, EIR Würenlingen, Dec. 1984
60. S. H. Smiley, D. C. Brater, US-Report K-1379 1/18 (1958).
61. J. A. Rode, US 594 201, 1966.
62. M. E. Kassner, P. H. Adler, M. G. Adamson, D. E. Petersen, *J. Nucl. Mater.* **167** (1989) 160 – 168.
63. G. Cheliotis et al.: *Thermionic Electrical Power Generation Proc.*, **vol. 2**, Jülich, June 1972, p. 861.
64. P. Dünner, J. Jung in (KFA Jülich ed.): Thermionic Electrical Power Generation Proc., vol. 2, Jülich, June 1972, p. 1015.
65. W. Rupp, H. Aßmann, M. Peehs: "Über die Einsatzmöglichkeiten von U_3Si in Schwerwasserreaktoren," *Reaktortagung des Deutschen Atomforums*, Conference Proceedings 1972.
66. U. Berndt, B. Erdmann, *Radiochim. Acta* **19** (1973) 45 – 46.
67. R. J. Sheil, US-Report ORNL-2061-Pt. 1–2-3 Del. 71 (1959).
68. S. Langar, F. E. Blankenship, *J. Inorg. Nucl. Chem.* **14** (1960) 26 – 31.
69. J. K. Dawson, R. W. M. D'Eye, E. A. Truswell, *J. Chem. Soc.* 1954, 3922 – 3929.
70. R. DeWitt, US-Report GAT-280, p. 102 (1960).
G. D. Oliver, H. T. Milton, J. W. Grisard, *J. Am. Chem. Soc.* **75** (1953) 2827 – 2829.
71. M. W. Mallett, A. F. Gerds, H. R. Nelson, *J. Electrochem. Soc.* **99** (1952) 197 – 204.
72. W. B. Wilson, *J. Am. Ceram. Soc.* **43** (1960) 77.
73. J. M. Leitnacker, W. G. Wittemann, *J. Chem. Phys.* **36** (1962) 1445.
74. S. Imoto et al., *Proc. of the Symposium at Harwell 1963*, **vol. 1**, McMillan, Basingstoke 1963, pp. 7 – 20.
75. J. Dubuisson et al., *Proc. U.N. Int. Conf. Peaceful Uses At. Energy 2nd*, **6** (1958) 551.
76. E. Barnes, *Prog. Nucl. Energy Ser. 1* **1** (1956) 61.
77. J. K. Dawson, R. G. Sowden: *Chemical Aspects of Nuclear Reactors,* vol. 1, Butterworths, London 1963.
78. H. S. Kalish, US-Report TID-7614 (1961) 99 – 110.
79. P. Himmelstein, H. Kühn, O. Pfahls, R. Lucas: "Verbesserung der Technologie bei der Fabrikation gegossener Urancarbidstäbe," *Euratom-Report* **EUR-4273 d.**
80. P. D. Shalek, *J. Am. Ceram. Soc.* **46** (1963) 155.
81. R. M. Dell, M. Albutt, UKEA-Report AERE-R 4253 (1963).
82. R. E. Rundle, N. C. Baenziger, A. S. Wilson, R. A. McDonald: "The Structure of Carbides, Nitrides and Oxides of Uranium," *J. Am. Chem. Soc.* **70** (1948) 99 – 105.
83. M. Smith, R. I. Honeycomble, *J. Inst. Met.* **83** (1954/55) 421 – 426.
84. T. Matsui, R. W. Ohse, Commission of the European Communities, report EUR-10858, Order no. PB88–125 356, 1986.
85. R. B. Matthews, *Space Nucl. Power Syst.* **8** (1989) 201 – 205.
86. S. L. Hayes, J. K. Thomas, K. L. Peddicord, *J. Nucl. Mater.* **171** (1990) nos. 2/3, 271 – 288.
87. G. Ledergerber, Z. Kopajtic, F. Ingold, R. W. Stratton, *J. Nucl. Mater.* **188** (1992) 28 – 35.
88. V. J. Wheeler, R. M. Dell, E. Wait, *J. Inorg. Nucl. Chem.* **26** (1964) 1829 – 1845.
89. H. R. Hoestra, S. Siegel, *J. Inorg. Nucl. Chem.* **18** (1961) 154 – 165.
90. V. G. Vlasov, Y. N. Semavin, *Zh. Prikl. Khim. (Leningrad)* **40** (1967) 1210 – 1215.
91. V. J. Wheeler, R. M. Dell, E. Wait, *J. Inorg. Nucl. Chem.* **26** (1964) 1829 – 1845.

92. A. N. Tsvigunov, L. M. Kuzwetsov, *Radiokhimiya* **16** (1974) 882 – 885.
93. H. R. Hoekstra, S. Siegel, L. H. Fuchs, J. J. Katz, *J. Phys. Chem.* **59** (1955) 136 – 138.
94. G. K. Khomyakov, V. I. Spitsyn, S. A. Zhvanko, *Issled. Obl. Khim. Urana* 1961, 141 – 144.
95. A. F. Bessonov, *Kristallografiya* **15** (1970) 62 – 67.
96. H. Powers, E. Welch, J. B. Trice, NEPA-934-SCR-34 (1949 and 1961).
97. Nukem, DE 2 623 977, 1976 (P. Börner, H. Vietzke).
98. M. El-Chehabi, Dissertation University of Texas, 1957.
99. J. E. Boggs, M. El-Chehabi, *J. Am. Chem. Soc.* **79** (1957) 4258 – 4260.
100. W. L. Marshall, J. S. Gull, C. H. Secoy, *J. Am. Chem. Soc.* **73** (1951) 1867.
101. A. Pilgenröther, W. Thomas, *ATW Atomwirtsch. Atomtech.* **16** (1971) 74 – 77.
102. Bundesminister für Verkehr (eds.), bulletin "Die Beförderung radioaktiver Stoffe," Oct. 19, 1984.
103. Deutscher Bundestag, Drucksache no. 10/2160, Oct. 19, 1984.

Urea

JOZEF H. MEESSEN, DSM Stamicarbon, Geleen, The Netherlands (Chaps. 1 – 7)

HARRO PETERSEN, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 8)

1. Physical Properties 2
2. Chemical Properties 3
3. Production 4
3.1. Principles 4
3.1.1. Chemical Equilibrium 4
3.1.2. Physical Phase Equilibria 7
3.2. Challenges in Urea Production Process Design 8
3.2.1. Recycle of Nonconverted Ammonia and Carbon Dioxide 9
3.2.2. Corrosion 11
3.2.3. Side Reactions 12
3.3. Description of Processes 13
3.3.1. Conventional Processes 13
3.3.2. Stripping Processes 13
3.3.2.1. Stamicarbon CO_2-Stripping Process 13
3.3.2.2. Snamprogetti Ammonia- and Self-Stripping Processes 16
3.3.2.3. ACES Process 18
3.3.2.4. Isobaric Double-Recycle Process . 19
3.3.3. Other Processes 19
3.4. Effluents and Effluent Reduction 20
3.5. Product-Shaping Technology . . . 21
4. Forms Supplied, Storage, and Transportation 22
5. Quality Specifications and Analysis 23
6. Uses 24
7. Economic Aspects 24
8. Urea Derivatives 25
8.1. Thermal Condensation Products of Urea 25
8.2. Alkyl- and Arylureas 25
8.2.1. Transamidation of Urea with Amines 25
8.2.2. Alkylation of Urea with Tertiary Alcohols 26
8.2.3. Phosgenation of Amines 26
8.2.4. Reaction of Amines with Cyanates (Salts) 27
8.2.5. Reaction with Isocyanates 27
8.2.6. Acylation of Ammonia or Amines with Carbamoyl Chlorides 27
8.2.7. Aminolysis of Esters of Carbonic and Carbamic Acids 27
8.3. Reaction of Urea and Its Derivatives with Aldehydes 28
8.3.1. α-Hydroxyalkylureas 28
8.3.2. α-Alkoxyalkylureas 29
8.3.3. α,α'-Alkyleneureas 30
8.3.4. Cyclic Urea – Aldehyde Condensation Products 31
9. References 33

Abbreviations:

CRH, %	critical relative humidity
ΔH_S, kJ/mol	integral heat of solution
m, mol/kg	urea molality, moles of urea per kilogram of water
$P^{O}_{(s)H^2}$, Pa	water vapor pressure of a saturated urea solution
P_v, Pa	vapor pressure

Urea [*57-13-6*], $CO(NH_2)_2$, M_r 60.056, plays an important role in many biological processes, among others in decomposition of proteins. The human body produces 20 – 30 g of urea per day.

In 1828, WÖHLER discovered [1] that urea can be produced from ammonia and cyanic acid in aqueous solution. Since then, research on the preparation of urea has continuously progressed. The starting point for the present industrial production of urea is the synthesis of BASAROFF [2], in which urea is obtained by dehydration of ammonium carbamate at increased temperature and pressure:

$$NH_2COONH_4 \rightleftarrows CO(NH_2)_2 + H_2O$$

In the beginning of this century, urea was produced on an industrial scale by hydration of cyanamide, which was obtained from calcium cyanamide:

$$CaCN_2 + H_2O + CO_2 \rightarrow CaCO_3 + CNNH_2$$

$$CNNH_2 + H_2O \rightarrow CO(NH_2)_2$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_333

After development of the NH_3 process (HABER and BOSCH, 1913, →Ammonia, Chap. 2. →Ammonia, Chap. 3. →Ammonia, Chap. 4.), the production of urea from NH_3 and CO_2, which are both formed in the NH_3 synthesis, developed rapidly:

$$2\,NH_3 + CO_2 \rightleftharpoons NH_2COONH_4$$

$$NH_2COONH_4 \rightleftharpoons CO(NH_2)_2 + H_2O$$

At present, urea is prepared on an industrial scale exclusively by reactions based on this reaction mechanism.

1. Physical Properties [3], [4]

Pure urea forms white, odorless, long, thin needles, but it can also appear in the form of rhomboid prisms. The crystal lattice is tetragonal – scalenohedral; the axis ratio $a : c$=1 : 0.833. The urea crystal is anisotropic (noncubic) and thus shows birefringence. At 20 °C the refractive indices are 1.484 and 1.602. Urea has an *mp* of 132.6 °C; its heat of fusion is 13.61 kJ/mol.

Physical properties of the melt at 135 °C follow:

ϱ	1247 kg/m^3
Molecular volume	48.16 m^3/kmol
η	3.018 mPa · s
Kinematic viscosity	2.42×10^{-6} m^2/s
Molar heat capacity, C_p	135.2 J mol^{-1} K^{-1}
Specific heat capacity, c_p	2.25 kJ kg^{-1} K^{-1}
Surface tension	66.3×10^{-3} N/m

In the temperature range 133 – 150 °C, density and dynamic viscosity of a urea melt can be calculated as follows:

$$r = 1638.5 - 0.96T$$

$$\ln\eta = 6700/T - 15.311$$

The density of the solid phase at 20 °C is 1335 kg/m^3; the temperature dependence of the density is given by 0.208 kg m^{-3} K^{-1}.

At 240 – 400 K, the molar heat capacity of the solid phase is [5]

$$C_p = 38.43 + 4.98 \times 10^{-2}T + 7.05 \times 10^{-4}T^2 - 8.61 \times 10^{-7}T^3$$

The vapor pressure of the solid phase between 56 and 130 °C [6] can be calculated from

$$\ln P_v = 32.472 - 11\,755/T$$

Hygroscopicity. The water vapor pressure of a saturated solution of urea in water $P^{O}_{(s)H^2}$ in the temperature range 10 – 80 °C is given by the relation [7]

$$\ln P_{(s)H_2O} = 175.766 - 11\,552/T - 22.679 \ln T$$

By starting from the vapor pressure of pure water $P^{O}_{H^2}$, the critical relative humidity (CRH) then can be calculated as

$$\mathrm{CRH} = \left(P_{(s)H_2O}/P_{H_2O}\right) 100$$

The CRH is a threshold value, above which urea starts absorbing moisture from ambient air. It shows the following dependence on temperature:

25 °C	76.5 %
30 °C	74.3 %
40 °C	69.2 %

At 25 °C, in the range of 0 – 20 mol of urea per kilogram of water, the integral heat of solution of urea crystals in water ΔH_s as a function of molality m is given by [8]:

$$\Delta H_s = 15.351 - 0.3523m + 2.327 \cdot 10^{-2}m^2 - 1.0106 \cdot 10^{-3}m^3 + 1.8853 \cdot 10^{-5}m^4$$

Urea forms a eutectic mixture with 67.5 wt % of water with a eutectic point at −11.5 °C.

The solubility of urea in a number of solvents, as a function of temperature is summarized in Table 1 [9], [10].

2. Chemical Properties

Upon *heating*, urea decomposes primarily to ammonia and isocyanic acid. As a result, the gas phase above a urea solution contains a considerable amount of HNCO, if the isomerization reaction in the liquid phase

$$CO(NH_2)_2 \rightleftharpoons NH_4NCO \rightleftharpoons NH_3 + HNCO$$

Table 1. Solubility of urea in various solvents (solubility in wt % of urea)

	Temperature, °C					
Solvent	0	20	40	60	80	100
Water	39.5	51.8	62.3	71.7	80.2	88.1
Ammonia	34.9	48.6	67.2	78.7	84.5	90.4
Methanol	13.0	18.0	26.1	38.6		
Ethanol	2.5	5.1	8.5	13.1		

has come to equilibrium [11]. In dilute aqueous solution, the HNCO formed hydrolyzes mainly to NH_3 and CO_2. In a more concentrated solution or in a urea melt, the isocyanic acid reacts further with urea, at relatively low temperature, to form biuret ($NH_2-CO-NH-CO-NH_2$), triuret ($NH_2-CO-NH-CO-NH-CO-NH_2$), and cyanuric acid $(HNCO)_3$ [12]. At higher temperature, guanidine $[CNH(NH_2)_2]$, ammelide $[C_3N_3(OH)_2NH_2]$, ammeline $[C_3N_3OH(NH_2)_2]$, and melamine $[C_3N_3(NH_2)_3]$ are also formed [13], [14].

Melamine can also be produced from urea by a catalytic reaction in the gas phase. To this end, urea is decomposed into NH_3 and HNCO at low pressure, and subsequently transformed catalytically to melamine (→Melamine and Guanamines, Chap. 4.).

Urea reacts with NO_x, both in the gas phase at 800 – 1150 °C and in the liquid phase at lower temperature, to form N_2, CO_2, and H_2O. This reaction is used industrially for the removal of NO_x from combustion gases [15], [16].

Reactions with Formaldehyde. Under *acid conditions*, urea reacts with formaldehyde to form among others, methyleneurea, as well as dimethylene-, trimethylene-, tetramethylene-, and polymethyleneureas. These products are used as slow-release fertilizer under the generic name ureaform [17] (→Fertilizers, Chap. 4.4.2.1.). The reaction scheme for the formation of methyleneurea is given below:

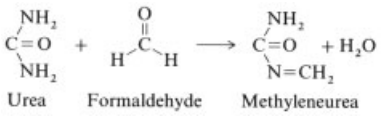

Methyleneurea reacts with additional molecules of formaldehyde to yield dimethyleneurea and other homologous products.

The reactions of urea with formaldehyde under *basic conditions* are used widely for the production of synthetic resins (→Amino Resins, Chap. 7.1.). As a first step, methylolurea instead of methyleneurea is formed:

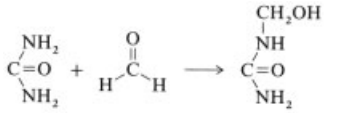

This product subsequently reacts with formaldehyde to dimethylol urea, $CO(NHCH_2OH)_2$, and further polymerization products. Since urea is also the raw material for the production of melamine, from which melamine – formaldehyde resins are produced, it is the most important building block in the production of amino resins.

When urea is applied as fertilizer to soil, it *hydrolyzes* in the presence of the enzyme urease to NH_3 and CO_2, after which NH_3 is bacteriologically converted into nitrate and, as such, absorbed by crops [17].

3. Production

3.1. Principles

3.1.1. Chemical Equilibrium

In all commercial processes, urea is produced by reacting ammonia and carbon dioxide at elevated temperature and pressure according to the Basaroff reactions:

$$2\,NH_3\,(l) + CO_2\,(l) \rightleftharpoons NH_2COONH_4$$
$$\Delta H = -\,117\,\text{kJ/mol} \qquad (1)$$

$$NH_2COONH_4 \rightleftharpoons NH_2CONH_2 + H_2O$$
$$\Delta H = +\,15.5\,\text{kJ/mol} \qquad (2)$$

A schematic of the overall process and the physical and chemical equilibria involved is shown in Figure 1. In the first reaction, carbon dioxide and ammonia are converted to ammonium carbamate; the reaction is fast and exothermic. In the second rection, which is slow and endothermic, ammonium carbamate dehydrates to produce urea and water. Since more heat is produced in the first reaction than consumed in the second, the overall reaction is exothermic.

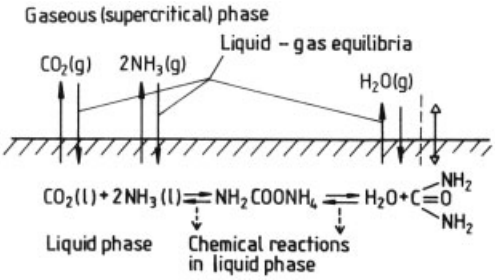

Figure 1. Physical and chemical equilibria in urea production

Processes differ mainly in the conditions (composition, temperature, and pressure) at which these reactions are carried out. Traditionally, the composition of the liquid phase in the reaction zone is expressed by two molar ratios: usually, the molar $NH_3 : CO_2$ and the molar $H_2O : CO_2$ ratios. Both reflect the composition of the so-called initial mixture [i.e., the hypothetical mixture consisting only of NH_3, CO_2, and H_2O if both Reactions (1) and (2) are shifted completely to the left].

First attempts to describe the chemical equilibrium of Reactions (1) and (2) were made by Frejacques [19]. Later descriptions of the chemical equilibria can be divided into regression analyses of measurements [20], [21] and thermodynamically consistent analyses of the equilibria [20], [22]. As far as the most important consequences of these equilibria on urea process design are concerned, the methods correspond closely to each other: The achievable conversion per pass, dictated by the chemical equilibrium as a function of temperature, goes through a maximum (Figs. 2 and 3). This effect is usually attributed to the fact that the ammonium carbamate concentration as a function of temperature goes through a maximum. This maximum in the ammonium carbamate concentration can be explained, at least qualitatively, by the respective heat effects of Reactions (1) and (2). However, this mechanism cannot explain the observed conversion maximum fully and quantitatively; other contributing mechanisms have been suggested [23].

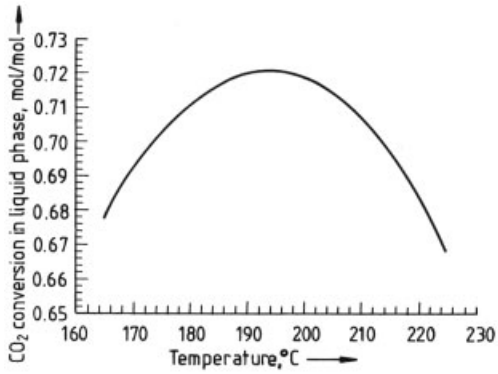

Figure 2. Carbon dioxide conversion at chemical equilibrium as a function of temperature
$NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture); $H_2O : CO_2$ ratio = 0.25 mol/mol (initial mixture)

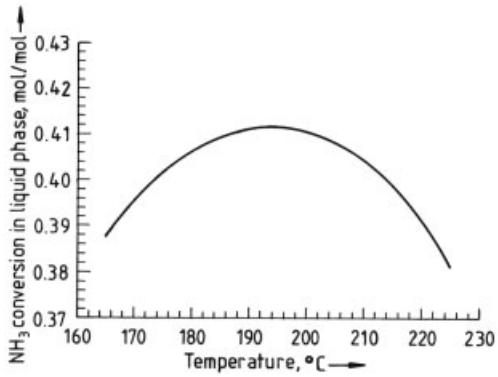

Figure 3. Ammonia conversion at chemical equilibrium as a function of temperature
$NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture); $H_2O : CO_2$ ratio = 0.25 mol/mol (initial mixture)

The influence of the composition of the initial mixture on the chemical equilibrium can be explained qualitatively by Reactions (1) and (2) and the law of mass action:

1) Increasing the $NH_3 : CO_2$ ratio (increasing the NH_3 concentration) increases CO_2 conversion, but reduces NH_3 conversion (Figs. 4 and 5).
2) Increasing the amount of water in the initial mixture (increasing the $H_2O : CO_2$ ratio) results in a decrease in both CO_2 and NH_3 conversion (Figs. 6 and 7).

In these cases, too, a full quantitative description cannot be derived simply from the law of mass action and Reactions (1) and (2). Other, not yet fully understood reaction mechanisms probably contribute to the chemical equilibria to a minor extent.

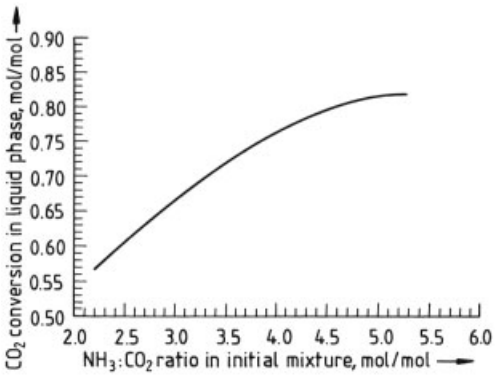

Figure 4. Carbon dioxide conversion at chemical equilibrium as a function of $NH_3 : CO_2$ ratio
$T = 190\,°C$; $H_2O : CO_2$ ratio = 0.25 mol/mol (initial mixture)

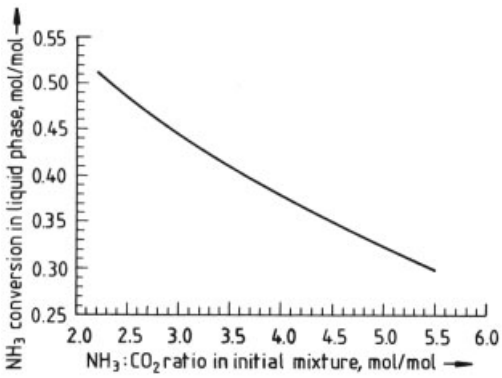

Figure 5. Ammonia conversion at chemical equilibrium as a function of $NH_3 : CO_2$ ratio
$T = 190\,°C$; $H_2O : CO_2$ ratio = 0.25 mol/mol (initial mixture)

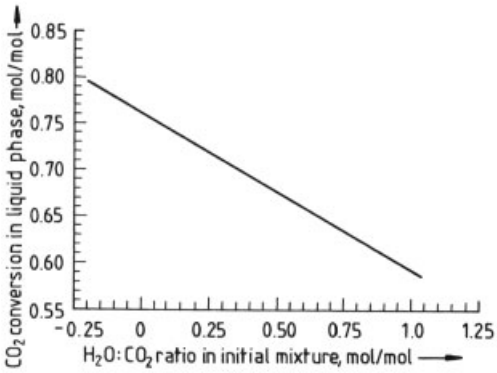

Figure 6. Carbon dioxide conversion at chemical equilibrium as a function of $H_2O : CO_2$ ratio
$T = 190\,°C$; $NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture)

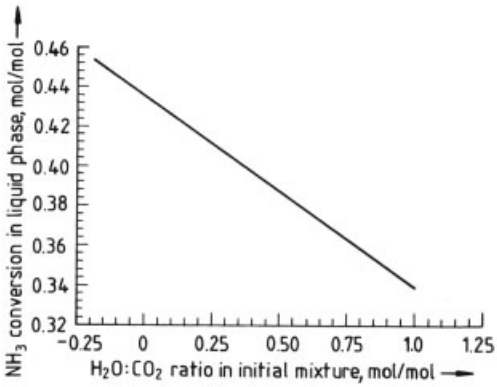

Figure 7. Ammonia conversion at chemical equilibrium as a function of $H_2O : CO_2$ ratio
$T = 190\,°C$; $NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture)

In Figures 2, 4, and 6, the conversion at chemical equilibrium is expressed as CO_2 conversion, that is, the amount of CO_2 in the initial mixture converted into urea (plus biuret), if no changes occur in overall NH_3, CO_2, and H_2O concentrations in the liquid phase. This way of representing the chemical equilibrium is consistent with the presentation usually found in the traditional urea literature. However, it is based on the arbitrary choice of CO_2 as the key component. Historically, this may be justified by the fact that in early urea processes, CO_2 conversion was more important than NH_3 conversion. For the present generation of stripping processes, however, giving a higher weight to CO_2 conversion is not justified. Comparing, e.g., Figs. 4 and 5,

shows that an arbitrary choice of one of the two feedstock components as yardstick to evaluate optimum reaction conversion can easily lead to faulty conclusions.

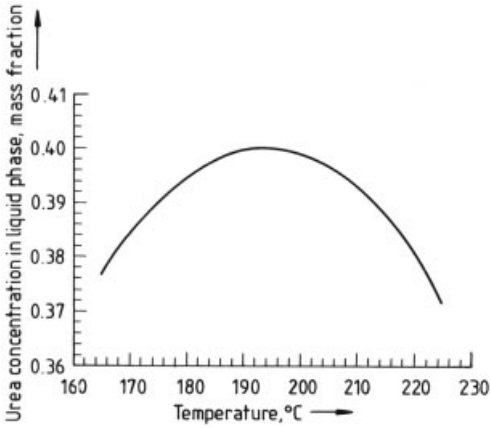

Figure 8. Urea yield in the liquid phase at chemical equilibrium as a function of temperature
$NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture); H_2O: CO_2 ratio = 0.25 mol/mol (initial mixture)

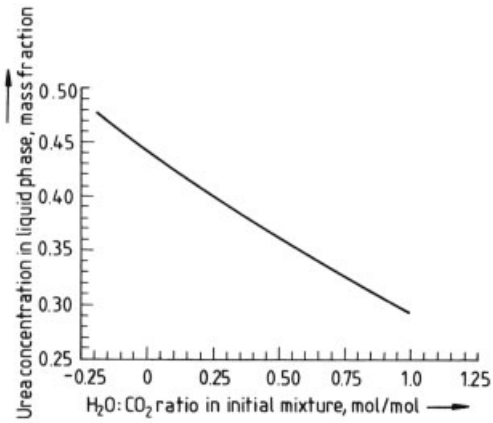

Figure 9. Urea yield in the liquid phase at chemical equilibrium as a function of $H_2O : CO_2$ ratio
$T = 190\,°C$; $NH_3 : CO_2$ ratio = 3.5 mol/mol (initial mixture)

Ultimately, project economics (investment and consumptions) will dictate the choice of process parameters in the reaction section. Without going into such time- and place-dependent economic considerations, one can argue that the urea yield (i.e., the concentration of urea in the liquid phase) is a better tool for judging optimum process parameters than CO_2 or NH_3 conversion. Figure 8 illustrates that urea yield as a function of temperature also goes through a maximum; the location of this maximum is of course composition dependent. Figure 9 again shows the detrimental effect of excess water on urea yield; thus, one of the targets in designing a recycle system must be to minimize water recycle.

Figure 10 shows that the urea yield as a function of $NH_3 : CO_2$ ratio reaches a maximum somewhat above the stoichiometric ratio (2 : 1). This is one of the reasons that all commercial processes operate at $NH_3 : CO_2$ ratios above the stoichiometric ratio. Another important reason for this can be found from the physical phase equilibria in the $NH_3 - CO_2 - H_2O -$ urea system.

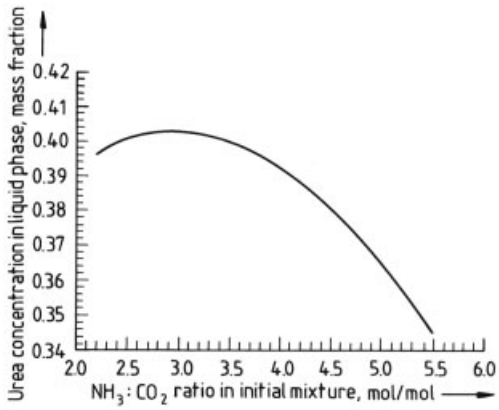

Figure 10. Urea yield in the liquid phase at chemical equilibrium as a function of $NH_3 : CO_2$ ratio
$T = 190\,°C$; $H_2O : CO_2$ ratio = 0.25 mol/mol (initial mixture)

3.1.2. Physical Phase Equilibria

In urea production, the phase behavior of the components under synthesis conditions is important. In all commercial processes, conditions are such that pressure and temperature are well above the critical conditions of the feedstocks ammonia and carbon dioxide; i.e., both components are in the supercritical state. The chemical interaction between NH_3 and CO_2 (mainly the formation of ammonium carbamate) results in a strongly azeotropic behavior of the "binary" system $NH_3 - CO_2$. An approach to the description of the phase equilibria if urea and water are added to the $NH_3 - CO_2$ system was given by Kaasenbrood and Chermin [24]. If

a less volatile solvent C (water) is added to an azeotropic system A – B ($NH_3 - CO_2$) at a pressure where both components A and B are supercritical, then the $T - X$ liquid and gas planes for the ternary system thus formed assume a special shape owing to the peculiar path described by the boiling points of the changing solutions (Fig. 11). Sections through the liquid plane for constant solvent content are analogous to the liquid line for the binary system. The liquid plane for the ternary systems appears as a ridge in the $T - X$ space. If the peak points of this ridge are linked up, the top ridge line is obtained. The points on this line do not have the same A : B ratio as the maximum for the binary azeotrope, because A and B are not soluble in solvent C to the same extent. The A : B ratio changes and the boiling point increases as the percentage of C increases.

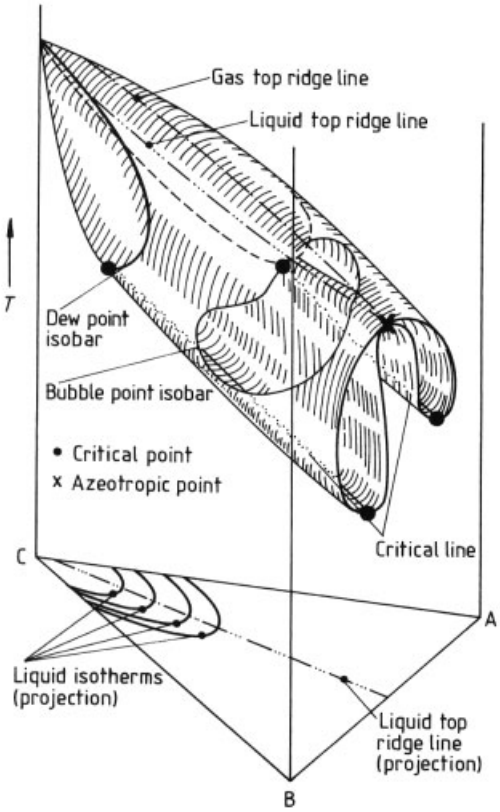

Figure 11. Liquid – gas equilibrium in a ternary system with binary azeotrope at constant pressure
The system A – B forms a binary azeotrope; C is a solvent for both A and B. The pressure is such that both A and B are supercritical, whereas the pressure is below the critical pressure of C.

Analogous to the description of Figure 11, the equilibria in the $NH_3 - CO_2 - H_2O$ – urea system under urea synthesis conditions show a maximum in temperature at a given pressure as a function of $NH_3 : CO_2$ ratio. A full description of the phase equilibria in this system is even more complex than the aforementioned hypothetical A – B – C system, since the solid – liquid (S – L) and solid – gas (S – G) equilibria interfere with the liquid – gas (L – G) equilibria.

The strongly azeotropic behavior of the $NH_3 - CO_2$ system, and the associated temperature maximum (or pressure minimum) in the ternary and quaternary systems with water and urea, are of practical importance in the realization of commercial urea processes. Carbon dioxide is less soluble than ammonia in water and urea melts. As a result, the pressure gradient at constant temperature is much steeper on the CO_2-rich side of the top ridge line. Moreover, this difference in solubility also causes the pressure minimum (or temperature maximum) to shift toward higher $NH_3 : CO_2$ ratios as the amount of solvent (water and urea) increases. Practically, this means that in order to achieve relatively low pressures at a given temperature, the $NH_3 : CO_2$ ratio in all commercial processes is chosen well above the stoichiometric ratio (2 : 1). In some processes, this ratio is chosen on the pressure minimum (on the top ridge line, i.e., at a ratio of ca. 3 : 1), whereas in other processes an even greater excess of ammonia is used.

3.2. Challenges in Urea Production Process Design

Like any process design, a urea plant design has to fulfil a number of criteria. Most important items are product quality, feedstocks and utilities consumptions, environmental aspects, safety, reliability of operation and a low initial investment. Since the urea process already has half a century of commercial scale history, it will be clear that compromises between the aforementioned, partly conflicting, criteria are well established. Also resulting from the age of urea process design is the observation that a process can only be successful if acceptable and competing solutions to all of these criteria can be combined into one process design. Apart from applying straightforward normal engineering approaches,

the challenge of finding an optimum synergy between partly conflicting criteria, focuses in urea plant design essentially on a few peculiarities:

1) The thermodynamic limit on the conversion per pass through the urea reactor, combined with the azeotropic behavior of the $NH_3 - CO_2$ system, necessitates a cunning recycle system design.
2) The intermediate product ammonium carbamate is extremely corrosive. A proper combination of process conditions, construction materials, and equipment design is therefore essential.
3) The occurrence of two side reactions – hydrolysis of urea and biuret formation – must be considered.

3.2.1. Recycle of Nonconverted Ammonia and Carbon Dioxide

The description of the chemical equilibria in Section 3.1.1 indicates that the conversion of the feedstocks NH_3 and CO_2 to urea is limited. An important differentiator between processes is the way these nonconverted materials are handled.

Once-Through Processes. In the very first processes, nonconverted NH_3 was neutralized with acids (e.g., nitric acid) to produce ammonium salts (such as ammonium nitrate) as coproducts of urea production. In this way, a relatively simple urea process scheme was realized. The main disadvantages of the once-through processes are the large quantity of ammonium salt formed as coproduct and the limited amount of overall carbon dioxide conversion that can be achieved. A peculiar aspect of this historic development is a partial "revival" of these combined urea – ammonium nitrate production facilities (UAN plants, see Section 3.3.3).

Conventional Recycle Processes. Once-through processes were soon replaced by total-recycle processes, where essentially all of the nonconverted ammonia and carbon dioxide were recycled to the urea reactor. In the first generation of total-recycle processes, several licensors developed schemes in which the recirculation of nonconverted NH_3 and CO_2 was performed in two stages. Figure 12 is a typical flow sheet of these, now called conventional, processes. The first recirculation stage was operated at medium pressure (18 – 25 bar); the second, at low pressure (2 – 5 bar). The first recirculation stage comprises at least a decomposition heater (d), in which carbamate decomposes into gaseous NH_3 and CO_2, while excess NH_3 evaporates simultaneously. The off-gas from this first decomposition step was subjected to rectification (e), from which relatively pure NH_3 (at the top) and a bottom product consisting of an aqueous ammonium carbamate solution were obtained. Both products are reycled separately to the urea reactor (c). In these processes, all nonconverted CO_2 was recycled as an aqueous solution, whereas the main portion of nonconverted NH_3 was recycled without an associated water recycle. Because of the detrimental effect of water on reaction conversion (see Figs. 6, 7, and 9), achieving a minimum CO_2 recycle (and thus maximum CO_2 conversion per reaction pass) was much more important than achieving a low NH_3 recycle. All conventional processes therefore typically operate at high $NH_3 : CO_2$ ratios (4 – 5 mol/mol) to maximize CO_2 conversion per pass. Although some of these conventional processes, partly equipped with ingenious heat-exchanging networks, have survived until now (see Section 3.3.1), their importance decreased rapidly as the so-called stripping processes were developed.

Stripping Processes. In the 1960s, the Stamicarbon CO_2-stripping process was developed, followed later by other processes (see Section 3.3.2). Characteristic of these processes is that the major part of the recycle of both nonconverted NH_3 and nonconverted CO_2 occurs via the gas phase, such that none of these recycles is associated with a large water recycle to the synthesis zone. Another characteristic difference between conventional and stripping processes in terms of the recycle scheme, can be found in the way heat is supplied to the recirculation zones. The energy balance of the conventional processes is shown in Figure 13. In this first-generation urea process, the heat supplied to the urea synthesis solution was used only once; therefore, this type of process can be referred to as an $N = 1$ process. Such a process required about 1.8 t of steam per tonne of urea.

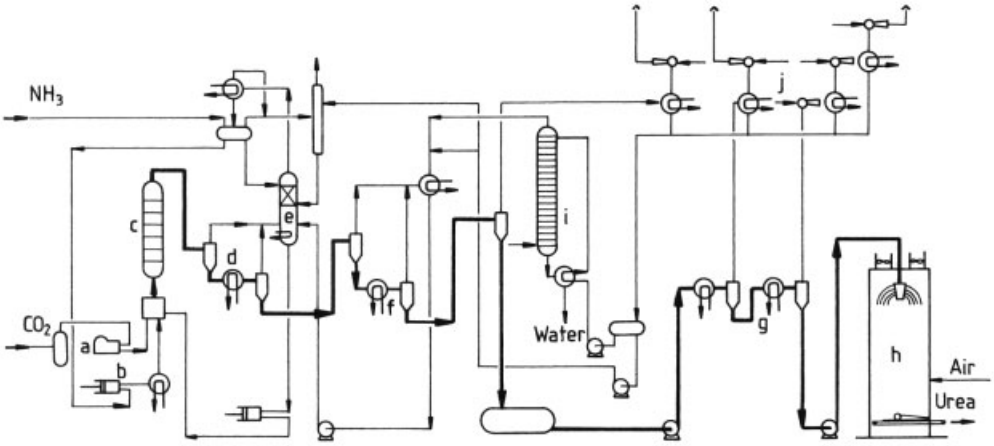

Figure 12. Typical flow sheet of a conventional urea plant
a) CO_2 compressor; b) High-pressure ammonia pump; c) Urea reactor; d) Medium-pressure decomposer; e) Ammonia – carbamate separation column; f) Low-pressure decomposer; g) Evaporator; h) Prilling; i) Desorber (wastewater stripper); j) Vacuum condensation section

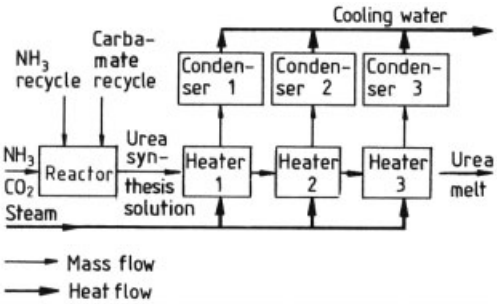

Figure 13. Conceptual diagram of the heat balance of a conventional urea process
Heat to each subsequent heater is supplied in the form of steam; the heat is used only once ($N = 1$).

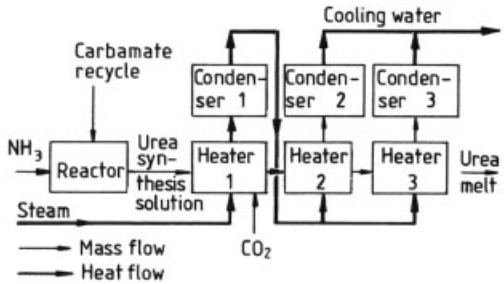

Figure 14. Conceptual diagram of the heat balance of a stripping plant
Heat supplied to the first heater (the stripper) is recovered in the first condenser (high-pressure carbamate condenser) and subsequently used again in the low-pressure heaters (decomposers and water evaporators); the heat is effectively used twice ($N = 2$).

The energy balance of a stripping plant is shown in Figure 14. As in conventional plants, heat must be supplied to the urea synthesis solution to decompose unconverted carbamate and to evaporate excess ammonia and water. However, a distinct difference in the heat balance with respect to the conventional process is that only the heat in the first heater (the high-pressure stripper) is imported. This heat is recovered in a high-pressure carbamate condenser (unconverted ammonia and carbon dioxide are condensed to form ammonium carbamate) and reused in the low-pressure heaters. The heat supplied is effectively used twice; thus, the term $N = 2$ process is used. The average energy consumption of the stripping process is 0.8 – 1.0 t of steam per tonne of urea.

In the 1980s, some processes were described that aim at a greater reduction of energy consumption by a further application of this multiple effect to $N = 3$ (Fig. 15) [25–29]. As can be seen from Figures 14 and 15, the steam requirement for process heating is reduced in these types of processes. However, whether the total energy consumption for the process is also reduced is doubtful, if the full capabilities of a $N = 2$ type of process are exploited and if the total energy supply scheme, including the energy supply to the carbon dioxide compressor drive, are taken into consideration [30]. Moreover, it seems that the emphasis in urea technology now is shifting from low energy consumption toward other

factors, such as more durable construction materials, more modern process control systems, and simple process design [31].

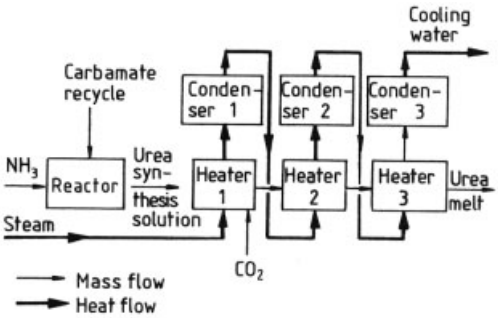

Figure 15. Further heat integration of a stripping plant in conceptual form
Heat supplied to the first heater (the stripper) is effectively used three times ($N = 3$).

3.2.2. Corrosion [32]

Urea synthesis solutions are very corrosive. Generally, ammonium carbamate is considered the aggressive component. This follows from the observation that carbamate-containing product streams are corrosive whereas pure urea solutions are not. The corrosiveness of the synthesis solution has forced urea manufacturers to set very strict demands on the quality and composition of construction materials. Awareness of the important factors in material selection, equipment manufacture and inspection, technological design and proper operations of the plant, together with periodic inspections and nondestructive testing are the key factors for safe operation for many years.

Role of Oxygen Content. Since the liquid phase in urea synthesis behaves as an electrolyte, the corrosion it causes is of an electrochemical nature. Stainless steel in a corrosive medium owes its corrosion resistance to the presence of a protective oxide layer on the metal. As long as this layer is intact, the metal corrodes at a very low rate. Passive corrosion rates of austenitic urea-grade stainless steels are generally between <0.01 and (max.) 0.10 mm/a. Upon removal of the oxide layer, activation and, consequently, corrosion set in unless the medium contains sufficient oxygen or oxidation agent to build a new layer. Active corrosion rates can reach values of 50 mm/a. Stainless steel exposed to carbamate-containing solutions involved in urea synthesis can be kept in a passivated (noncorroding) state by a given quantity of oxygen. If the oxygen content drops below this limit, corrosion starts after some time – its onset depending on process conditions and the quality of the passive layer. Hence, introduction of oxygen and maintenance of a sufficiently high oxygen content in the various process streams are prerequisites to preventing corrosion of the equipment. Although some alternatives have been mentioned [33], [34], the use of oxygen to influence the redox potential has become common practice in urea manufacture ever since it was initially suggested [35], [36].

From the point of view of corrosion prevention, the condensation of $NH_3 - CO_2 - H_2O$ gas mixtures to carbamate solutions deserves great attention. This is necessary because – notwithstanding the presence of oxygen in the gas phase – an oxygen-deficient corrosive condensate is initially formed on condensation. In this condensate the oxygen is absorbed only slowly. This accounts for the severe corrosion sometimes observed in cold spots inside gas lines. The trouble can be remedied by adequate isolation and tracing of the lines.

When condensation constitutes an essential process step – for example, in high-pressure and low-pressure carbamate condensers – special technological measures must be taken. These measures can involve ensuring that an oxygen-rich liquid phase is introduced into the condenser, while appropriate liquid – gas distribution devices ensure that no dry spots exist on condensing surfaces.

Not only condensing but also stagnant conditions are dangerous, especially where narrow crevices are present, into which hardly any oxygen can penetrate and oxygen depletion may occur.

Role of Temperature and Other Process Parameters in Corrosion. *Temperature* is the most important technological factor in the behavior of the steels employed in urea synthesis. An increase in temperature increases active corrosion, but more important, above a critical temperature it causes spontaneous activation of passive steel. The higher-alloyed austenitic stain-

less steels (e.g., containing 25 wt % chromium, 22 wt % nickel, and 2 wt % molybdenum) appear to be much less sensitive to this critical temperature than 316 L types of steel.

Sometimes, the *NH_3 : CO_2 ratio* in synthesis solutions is also claimed to have an influence on the corrosion rate of steels under urea synthesis conditions. Experiments have showed that under practical conditions this influence is not measurable because the steel retains passivity. Spontaneous activation did not occur. Only with electrochemical activation could 316 L types of steel be activated at intermediate $NH_3 : CO_2$ ratios. At low and high ratios, 316 L stainless steel could not be activated. The higher-alloyed steel type 25 Cr 22 Ni 2 Mo showed stable passivity, irrespective of the $NH_3 : CO_2$ ratio, even when activated electrochemically. Of course, these results depend on the specific temperature and oxygen content during the experiments.

Material Selection. Corrosion resistance is not the only factor determining the choice of construction materials. Other factors such as mechanical properties, workability, and weldability, as well as economic considerations such as price, availability, and delivery time, also deserve attention.

Stainless steels that have found wide use are the austenitic grades AISI 316 L and 317 L. Like all Cr-containing stainless steels, AISI 316 L and 317 L are not resistant to the action of sulfides. Hence it is imperative in plants using the 316 L and 317 L grades in combination with CO_2 derived from sulfur-containing gas, to purify this gas or the CO_2 thoroughly.

In stripping processes, the process conditions in the high-pressure stripper are most severe with respect to corrosion.

In the *Stamicarbon CO_2-stripping process*, a higher-alloyed, but still fully austenitic stainless steel (25 Cr 22 Ni 2 Mo) was chosen as construction material for the stripper tubes. This choice ensures better corrosion resistance than 316 L or 317 L types of material but still maintains the advantages of workability, weldability, reparability, and the cheaper price of stainless steel-type materials.

In the *Snamprogetti stripping processes*, titanium usually is chosen for this critical application, although mechanically bonded bimetallic 25 Cr 22 Ni 2Mo – zirconium tubes have also been suggested to improve corrosion resistance [37], [38].

In the *ACES process*, duplex alloys (ferritic – austenitic) are used as construction material for the stripper tubes.

3.2.3. Side Reactions [39]

Three side reactions are of special importance in the design of urea production processes:

Hydrolysis of urea

$$CO(NH_2)_2 + H_2O \rightarrow NH_2COONH_4 \rightarrow 2\,NH_3 + CO_2 \quad (3)$$

Biuret formation from urea:

$$2\,CO(NH_2)_2 \rightarrow NH_2CONHCONH_2 + NH_3 \quad (4)$$

Formation of isocyanic acid from urea:

$$CO(NH_2)_2 \rightarrow NH_4NCO \rightarrow NH_3 + HNCO \quad (5)$$

All three side reactions have in common the decomposition of urea; thus, the extent to which they occur must be minimized.

The *hydrolysis reaction* (3) is nothing but the reverse of urea formation. Whereas this reaction approaches equilibrium in the reactor, in all downstream sections of the plant the NH_3 and CO_2 concentrations in urea-containing solutions are such that Reaction (3) is shifted to the right. The extent to which the reaction occurs is determined by temperature (high temperatures favor hydrolysis) and reaction kinetics; in practice, this means that retention times of urea-containing solutions at high temperatures must be minimized.

The *biuret reaction* (4) also approaches equilibrium in the urea reactor [20], [22]. The high NH_3 concentration in the reactor shifts Reaction (4) to the left, such that only a small amount of biuret is formed in the reactor. In downstream sections of the plant, NH_3 is removed from the urea solutions, thereby creating a driving force for biuret formation. The extent to which biuret is formed is determined by reaction kinetics; therefore, the practical measures to minimize biuret formation are the same as described above for the hydrolysis reaction.

Reaction (5) shows that formation of isocyanic acid from urea is also favored by low NH_3 concentrations. This reaction is especially

relevant in the evaporation section of the plant. Here, low pressures are applied, resulting in a transfer of NH_3 and HNCO into the gas phase and, consequently, low concentrations of these constituents in the liquid phase. Together with the relatively high temperatures in the evaporators, this shifts Reaction (5) to the right. The extent to which this reaction occurs is again determined by kinetics. The HNCO removed via the gas in the evaporators collects in the process condensate from the vacuum condensers, where low temperatures shift Reaction (5) to the left, again forming urea. As a result of this mechanism of chemical entrainment, attempts to minimize entrainment from evaporators with physical (liquid – gas) separation devices are destined to be unsuccessful.

3.3. Description of Processes

3.3.1. Conventional Processes

As explained in Section 3.2, conventional processes have generally been replaced by stripping processes. Only two conventional processes may still have some importance in the near future.

Urea Technologies Inc. (UTI) Heat Recycle Process (see Fig. 16). Ammonia (containing passivating air), recycled carbamate, and about 60 % of the feed CO_2 are charged to the top of an open-ended coil reactor (c) operating at 210 bar. Ammonium carbamate is formed within the coil, exits the coil at the bottom, and then flows up and around it–the exothermic heat of carbamate formation in the coil driving the endothermic dehydration of carbamate to urea outside the coil. The reactor is claimed to achieve a uniform temperature profile in this way. In the reactor, a relative high $NH_3 : CO_2$ ratio (4.2 : 1) is applied. The reactor effluent is depressurized and subcooled, and the flashed gases are released before the first decomposer (f). Gases leaving the first decomposer separator (g) are mixed with about 40 % of the feed carbon dioxide and partially condensed in the heat recovery section (i – k). The two combined gas streams are then further condensed and form the carbamate recycle flow. Urea as an 86 – 88 % solution is concentrated by evaporation (i) before granulation or prilling. The process is applied in eight small-scale and two medium-scale plants.

MTC (Mitsui Toatsu Corporation) Conventional Processes of Toyo Engineering Corporation. The conventional processes developed by Toyo Engineering Corporation (TEC) were successfully commercialized until the mid-1980s. The continuous evolution of these processes is reflected in their sequential nomenclature:

TR – A	Total-Recycle A Process
TR – B	Total-Recycle B Process
TR – C	Total-Recycle C Process
TR – CI	Total-Recycle C Improved Process
TR – D	Total-Recycle D Process

Partial-recycle versions of these processes were also realized. These TEC MTC conventional processes were applied in more then 70 plants. However, the licensor of these processes has announced a stripping process (the ACES process; see Section 3.3.2.3); this probably means the end of the TEC MTC conventional process line.

3.3.2. Stripping Processes

3.3.2.1. Stamicarbon CO_2-Stripping Process (Figs. 17 and 18)

The synthesis stage of the Stamicarbon process consists of a urea reactor (c), a stripper for unconverted reactants (d), a high-pressure carbamate condenser (e), and a high-pressure reactor off-gas scrubber (f). To realize maximum urea yield per pass through the reactor at the stipulated optimum pressure of 140 bar, an $NH_3 : CO_2$ molar ratio of 3 : 1 is applied. The greater part of the unconverted carbamate is decomposed in the stripper, where ammonia and carbon dioxide are stripped off. This stripping action is effected by countercurrent contact between the urea solution and fresh carbon dioxide at synthesis pressure. Low ammonia and carbon dioxide concentrations in the stripped urea solution are obtained, such that the recycle from the low-pressure recirculation stage (h, j) is minimized. These low concentrations of both ammonia and carbon dioxide in the stripper effluent can be obtained at relatively low temperatures of the urea solution because carbon dioxide is only sparingly soluble under such conditions.

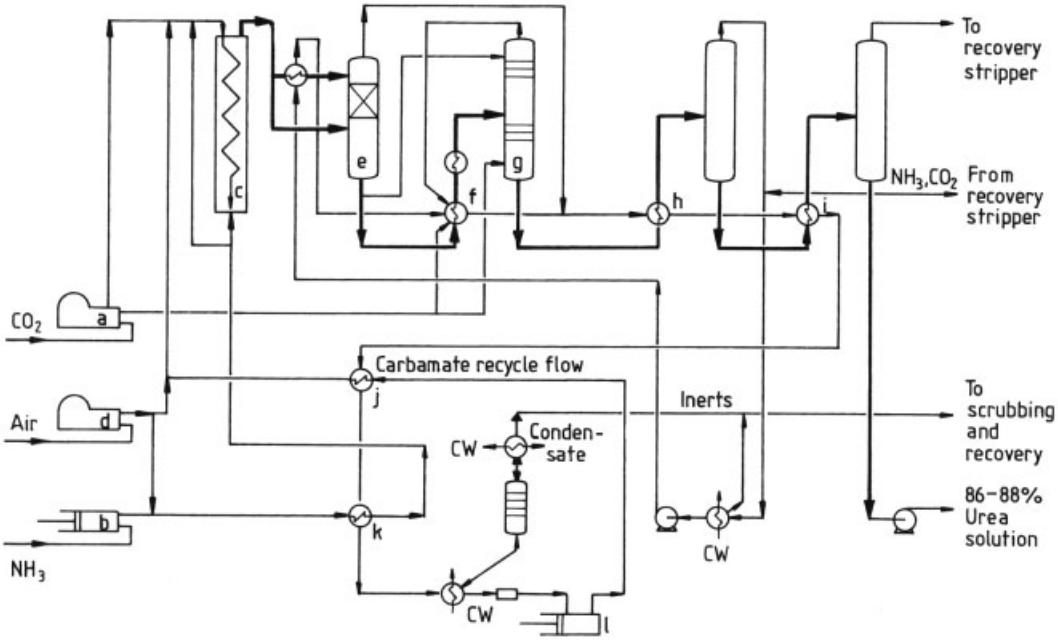

Figure 16. UTI heat recycle urea process
a) CO_2 compressor; b) High-pressure ammonia feed pump; c) Urea reactor with internal coil; d) Air compressor; e) Liquid distributor (used as first flash vessel); f) First decomposer; g) First separator; h) Second decomposer; i) Urea concentrator; j) Carbamate heater; k) Ammonia heater; l) High-pressure carbamate recycle pump
CW = Cooling water

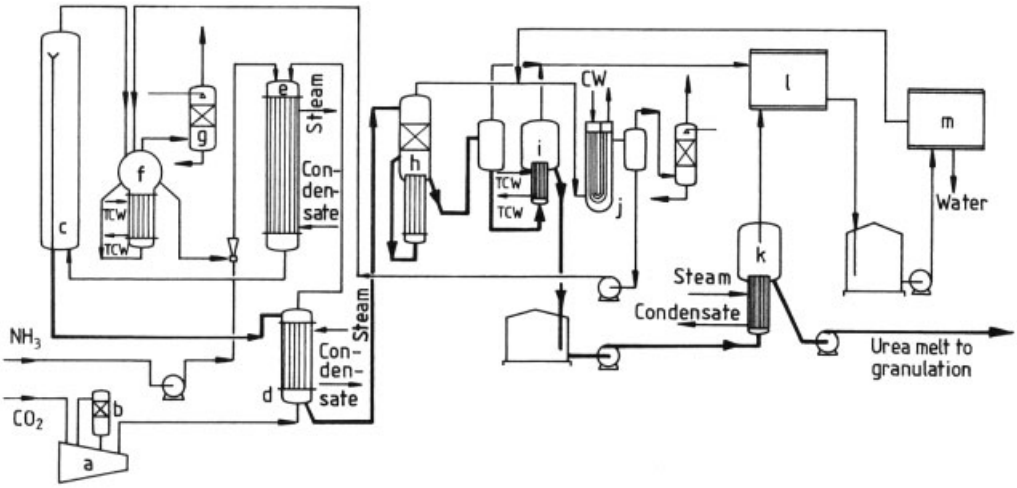

Figure 17. Stamicarbon CO_2-stripping urea process (The process suitable for combination with a granulation plant is shown here; combination with prilling is also possible.)
a) CO_2 compressor; b) Hydrogen removal reactor; c) Urea reactor; d) High-pressure stripper; e) High-pressure carbamate condenser; f) High-pressure scrubber; g) Low-pressure absorber; h) Low-pressure decomposer and rectifier; i) Pre-evaporator; j) Low-pressure carbamate condenser; k) Evaporator; l) Vacuum condensation section; m) Process condensate treatment
CW = Cooling water; TCW = Tempered cooling water

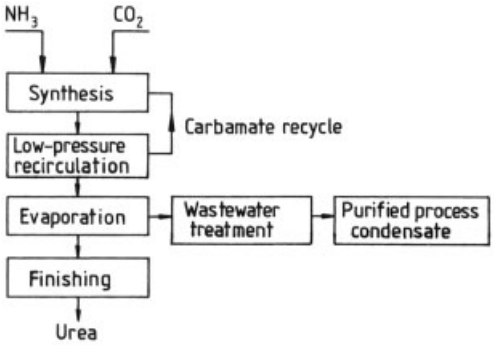

Figure 18. Functional block diagram of the Stamicarbon CO_2-stripping urea process

Condensation of ammonia and carbon dioxide gases, leaving the stripper, occurs in the high-pressure carbamate condenser (e) at synthesis pressure. As a result, the heat liberated from ammonium carbamate formation is at a high temperature. This heat is used for the production of 4.5-bar steam for use in the urea plant itself. The condensation in the high-pressure carbamate condenser is not effected completely. Remaining gases are condensed in the reactor and provide the heat required for the dehydration of carbamate, as well as for heating the mixture to its equilibrium temperature. In an improvement to this process, the condensation of off-gas from the stripper is carried out in a prereactor, where sufficient residence time for the liquid phase is provided. As a result of urea and water formation in the condensing zone, the condensation temperature is increased, thus enabling the production of steam at a higher pressure level [40].

The feed carbon dioxide, invariably originating from an associated ammonia plant, always contains hydrogen. To avoid the formation of explosive hydrogen – oxygen mixtures in the tail gas of the plant, hydrogen is catalytically removed from the carbon dioxide feed (b). Apart from the air required for this purpose, additional air is supplied to the fresh carbon dioxide input stream. This extra portion of oxygen is needed to maintain a corrosion-resistant layer on the stainless steel in the synthesis section. Before the inert gases, mainly oxygen and nitrogen, are purged from the synthesis section, they are washed with carbamate solution from the low-pressure recirculation stage in the high-pressure scrubber (f) to obtain a low ammonia concentration in the subsequently purged gas. Further washing of the off-gas is performed in a low-pressure absorber (g) to obtain a purge gas that is practically ammonia free. Only one low-pressure recirculation stage is required due to the low ammonia and carbon dioxide concentrations in the stripped urea solution. Because of the ideal ratio between ammonia and carbon dioxide in the recovered gases in this section, water dilution of the resultant ammonium carbamate is at a minimum despite the low pressure (about 4 bar). As a result of the efficiency of the stripper, the quantities of ammonium carbamate for recycle to the synthesis section are also minimized, and no separate ammonia recycle is required.

The urea solution coming from the recirculation stage contains about 75 wt % urea. This solution is concentrated in the evaporation section (k). If the process is combined with a prilling tower for final product shaping, the final moisture content of urea from the evaporation section is ca. 0.25 wt %. If the process is combined with a granulation unit, the final moisture content may vary from 1 to 5 wt %, depending on granulation requirements. Higher moisture contents can be realized in a single-stage evaporator, whereas low moisture contents are economically achieved in a two-stage evaporation section.

When urea with an extremely low biuret content is required (at a maximum of 0.3 wt %), pure urea crystals are produced in a crystallization section. These crystals are separated from the mother liquor by a combination of sieve bends and centrifuges and are melted prior to final shaping in a prilling tower or granulation unit.

The process condensate emanating from water evaporation from the evaporation or crystallization sections contains ammonia and urea. Before this process condensate is purged, urea is hydrolyzed into ammonia and carbon dioxide (l), which are stripped off with steam and returned to urea synthesis via the recirculation section. This process condensate treatment section can produce water with high purity, thus transforming this "wastewater" treatment into the production unit of a valuable process condensate, suitable for, e.g., cooling tower or boiler feedwater makeup. Since the introduction of the Stamicarbon CO_2-stripping process, some 125 units have been built according to this process all over the world.

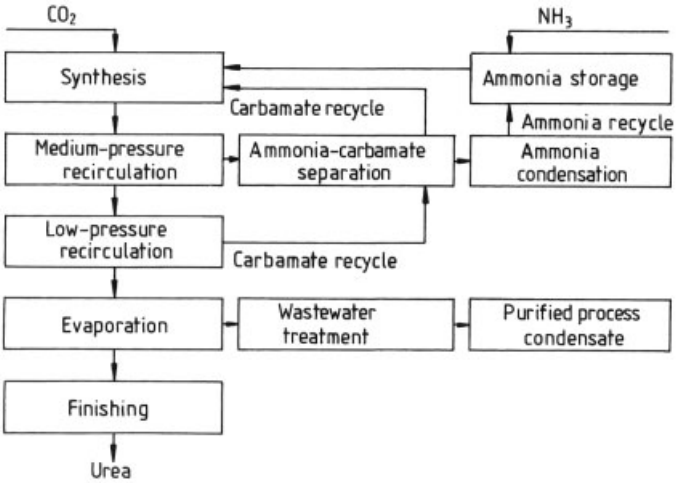

Figure 19. Functional block diagram of the Snamprogetti self-stripping process

3.3.2.2. Snamprogetti Ammonia- and Self-Stripping Processes [41–47]

In the first generation of NH_3- and self-stripping processes, ammonia was used as stripping agent. Because of the extreme solubility of ammonia in the urea-containing synthesis fluid, the stripper effluent contained rather large amounts of dissolved ammonia, causing an ammonia overload in downstream sections of the plant. Later versions of the process abandoned the idea of using ammonia as stripping agent; stripping was achieved only by supply of heat ("thermal" or "self"-stripping). Even without using ammonia as a stripping agent, the $NH_3 : CO_2$ ratio in the stripper effluent is relatively high, so the recirculation section of the plant requires an ammonia – carbamate separation section, as in conventional processes (see Fig. 19).

The process uses a vertical layout in the synthesis section. Recycle within the synthesis section, from the stripper (h) via the high-pressure carbamate condenser (f), through the carbamate separator (e) back to the reactor (b), is maintained by using an ammonia-driven liquid – liquid ejector (c) [43], [45] (see Fig. 20). In the reactor, which is operated at 150 bar, an $NH_3 : CO_2$ molar feed ratio of ca. 3.5 is applied. The stripper is of the falling film type [46]. Since stripping is achieved thermally, relatively high temperatures (200 – 210 °C) are required to obtain a reasonable stripping efficiency. Because of this high temperature, stainless steel is not suitable as construction material for the stripper from a corrosion point of view; titanium and bimetallic zirconium – stainless steel tubes have been used [37], [38].

Off-gas from the stripper is condensed in a kettle-type boiler (f) [44]. At the tube side of this condenser the off-gas is absorbed in recycled liquid carbamate from the medium-pressure recovery section. The heat of absorption is removed through the tubes, which are cooled by the production of low-pressure steam at the shell side. The steam produced is used effectively in the back end of the process.

In the medium-pressure decomposition and recirculation section, typically operated at 18 bar, the urea solution from the high-pressure stripper is subjected to the decomposition of carbamate and evaporation of ammonia (i). The off-gas from this medium-pressure decomposer is rectified. Liquid ammonia reflux is applied to the top of this rectifier (j); in this way a top product consisting of pure gaseous ammonia, and a bottom product of liquid ammonium carbamate are obtained. The pure ammonia off-gas is condensed (k) and recycled to the urea synthesis section. To prevent solidification of ammonium carbamate in the rectifier, some water is added to the bottom section of the column to dilute the ammonium carbamate below its crystallization point. The liquid ammonium carbamate – water mixture obtained in this way is also recycled to

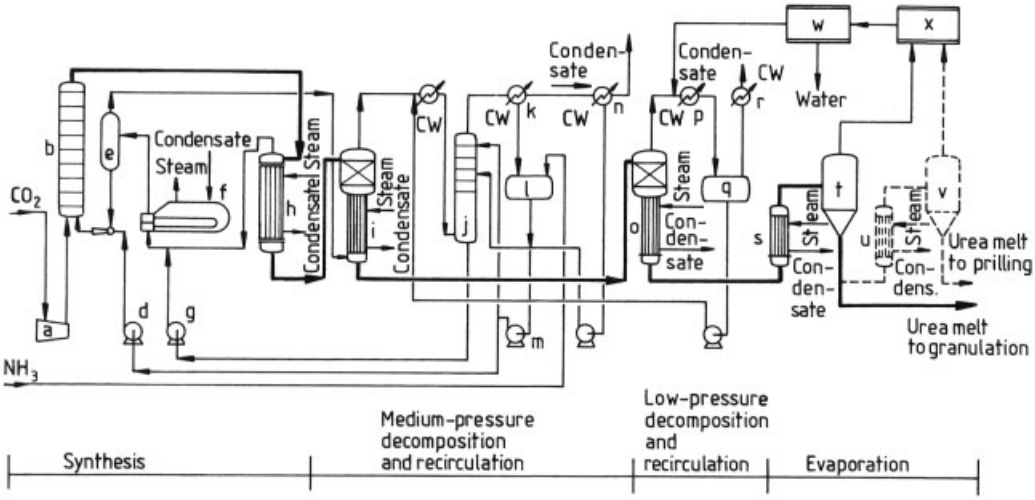

Figure 20. Schematic of the Snamprogetti self-stripping process
a) CO_2 compressor; b) Urea reactor; c) Ejector; d) High-pressure ammonia pump; e) Carbamate separator; f) High-pressure carbamate condenser; g) High-pressure carbamate pump; h) High-pressure stripper; i) Medium-pressure decomposer and rectifier; j) Ammonia – carbamate separation column; k) Ammonia condenser; l) Ammonia receiver; m) Low-pressure ammonia pump; n) Ammonia scrubber; o) Low-pressure decomposer and rectifier; p) Low-pressure carbamate condenser; q) Low-pressure carbamate receiver; r) Low-pressure off-gas scrubber; s) First evaporation heater; t) First evaporation separator; u) Second evaporation heater; v) Second evaporation separator; w) Wastewater treatment; x) Vacuum condensation section
CW = Cooling water

the synthesis section. The purge gas of the ammonia condensers is treated in a scrubber (n) prior to being purged to the atmosphere.

The urea solution from the medium-pressure decomposer is subjected to a second lowpressure decomposition step (o). Here, further decomposition of ammonium carbamate is achieved, so that a substantially carbamate-free aqueous urea solution is obtained. Off-gas from this low-pressure decomposer is condensed (p) and recycled as an aqueous ammonium carbamate solution to the synthesis section via the medium-pressure recovery section.

Concentrating the urea – water mixture obtained from the low-pressure decomposer is performed in a single or double evaporator (s – v), depending on the requirements of the finishing section. Typically, if prilling is chosen as the final shaping procedure, a two-stage evaporator is required, whereas in the case of a fluidized-bed granulator a single evaporation step is sufficient to achieve the required final moisture content of the urea melt. In some versions of the process, heat exchange is applied between the off-gas from the medium-pressure decomposer and the aqueous urea solution to the evaporation section. In this way, the consumption of low-pressure steam by the process is reduced.

The process condensate obtained from the evaporation section is subjected to a desorption – hydrolysis operation to recover the urea and ammonia contained in the process condensate.

Up to now, about 70 plants have been designed according to the Snamprogetti ammonia- and self-stripping processes.

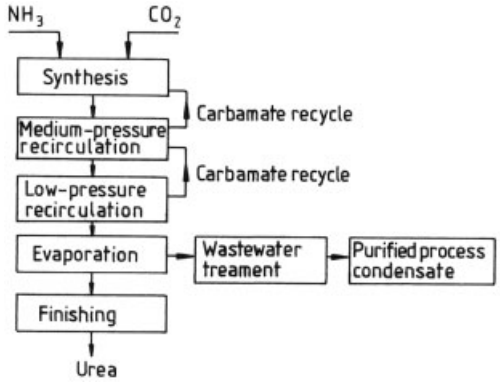

Figure 21. Functional block diagram of the ACES urea process

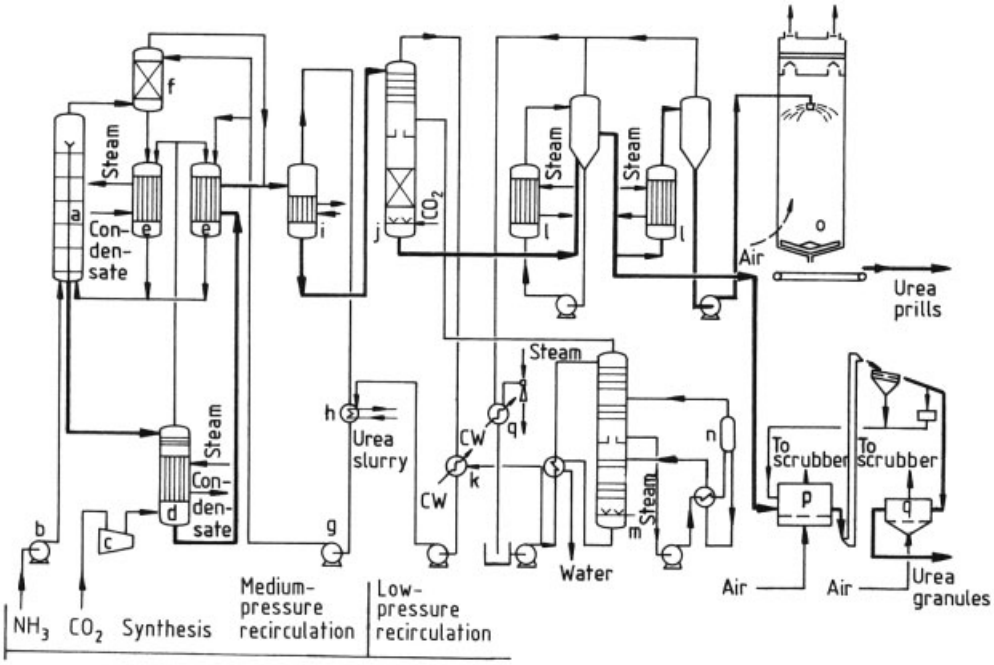

Figure 22. Schematic of the ACES process
a) Urea reactor; b) High-pressure ammonia pump; c) CO_2 compressor; d) Stripper; e) High-pressure carbamate condensers; f) High-pressure scrubber; g) High-pressure carbamate pump; h) Medium-pressure absorber; i) Medium-pressure decomposer; j) Low-pressure decomposer; k) Low-pressure absorber; l) Evaporators; m) Process condensate stripper; n) Hydrolyzer; o) Prilling tower; p) Granulation section; q) Surface condensers
CW = Cooling water

3.3.2.3. ACES Process [29], [48], [49]

The ACES (i.e., Advanced Process for Cost and Energy Saving) process has been developed by Toyo Engineering Corporation. Its synthesis section consists of the reactor (a), stripper (d), two parallel carbamate condensers (e), and a scrubber (f) – all operated at 175 bar (see Figs. 21 and 22).

The reactor is operated at 190 °C and an NH_3 : CO_2 molar feed ratio of 4 : 1. Liquid ammonia is fed directly to the reactor, whereas gaseous carbon dioxide after compression is introduced into the bottom of the stripper as a stripping aid. The synthesis mixture from the reactor, consisting of urea, unconverted ammonium carbamate, excess ammonia, and water, is fed to the top of the stripper. The stripper has two functions. Its upper part is equipped with trays where excess ammonia is partly separated from the stripper feed by direct countercurrent contact of the feed solution with the gas coming from the lower part of the stripper. This prestripping in the top is said to be required to achieve effective CO_2 stripping in the lower part. In the lower part of the stripper (a falling film heater), ammonium carbamate is decomposed and the resulting CO_2 and NH_3 as well as the excess NH_3 are evaporated by CO_2 stripping and steam heating. The overhead gaseous mixture from the top of the stripper is introduced into the carbamate condensers (e). Here, two units in parallel are installed, where the gaseous mixture is condensed and absorbed by the carbamate solution coming from the medium-pressure recovery stage. Heat liberated in the high-pressure carbamate condensers is used to generate low-pressure steam in one of the condensers and to heat the urea solution from the stripper after the pressure is reduced to about 19 bar in the shellside of the second carbamate condenser. The gas and liquid from the carbamate condensers are recycled to the reactor by gravity flow. The urea solution from the stripper, with a typical NH_3 content of 12 wt %, is purified further in the subsequent

medium- and low-pressure decomposers (i, j), operating at 19 and 3 bar, respectively. Ammonia and carbon dioxide separated from the urea solution here are recovered through stepwise absorption in the low- and medium-pressure absorbers (h, k). Condensation heat in the medium-pressure absorber is transferred directly to the aqueous urea solution feed in the final concentration section. In this final concentration section (l), the purified urea solution is concentrated further either by a two-stage evaporation up to 99.7 % for urea prill production or by a single evaporation up to 98.5 % for urea granule production. Water vapor formed in the final concentrating section is condensed in surface condensers (q) to form process condensate. Part of this condensate is used as an absorbent in the recovery sections, whereas the remainder is purified in the process condensate treatment section by hydrolysis and steam stripping, before being discharged from the urea plant.

The highly concentrated urea solution is finally processed either through the prilling tower (o) or via the urea granulator (p). Instead of concentration via evaporation, the ACES process can also be combined with a crystallization section to produce urea with low biuret content.

Until now, the ACES process has been used in seven urea plants.

3.3.2.4. Isobaric Double-Recycle Process
[26], [28], [50]

The isobaric double-recycle (IDR) stripping process, developed by Montedison, is characterized by recycle of most of the unreacted ammonia and ammonium carbamate in two decomposers in series, both operating at the synthesis pressure. A high molar $NH_3 : CO_2$ ratio (4 : 1 to 5 : 1) in the reactor is applied. As a result of this choice of ratio, the reactor effluent contains a relatively high amount of nonconverted ammonia. In the first, steam-heated, high-pressure decomposer, this large quantity of free ammonia is mainly removed from the urea solution. Most of the residual ammonia, as well as some ammonium carbamate, is removed in the second high-pressure decomposer where steam heating and CO_2 stripping are applied. The high-pressure synthesis section is followed by two lower-pressure decomposition stages of traditional design, where heat exchange between the condensing off-gas of the medium-pressure decomposition stage and the aqueous urea solution to the final concentration section improves the overall energy consumption of the process. Probably because of the complexity of this process, it has not achieved great popularity so far. The IDR process or parts of the process are used in four revamps of older conventional plants.

3.3.3. Other Processes

Urea – Ammonium Nitrate (UAN) Production. Mixtures of urea (*mp* 133 °C) and ammonium nitrate (*mp* 169 °C) with water have a eutectic point at −26.5 °C [51]. As a result solutions with high nitrogen content can be made with solidification temperatures below ambient temperature. These mixtures, called UAN solutions, are used as liquid nitrogen fertilizers.

UAN solutions can be made by mixing the appropriate amounts of solid urea and solid ammonium nitrate with water or, alternatively, in a production facility specially designed to produce UAN solutions. In this latter category the Stamicarbon CO_2-stripping technology is especially suitable. In a partial-recycle version of this process, unconverted ammonia emanating from the stripped urea solution and from the reactor off-gas is neutralized with nitric acid. The ammonium nitrate solution thus formed and the urea solution from the synthesis section are mixed to yield a product solution with the desired nitrogen content (32 – 35 wt %) directly. Such a plant designated for the production of UAN solutions is cheaper than the separate production of urea and ammonium nitrate in investment and in operating costs, because evaporation, final product shaping for both urea and ammonium nitrate, and wastewater treatment sections are not required.

Integrated Ammonia – Urea Production. Both feedstocks required for urea production, ammonia and carbon dioxide, are usually obtained from an ammonia plant. Since an ammonia plant is a net heat (steam) producer and a urea plant is a net heat (steam) consumer, it is normal practice to integrate the steam systems of both plants. Since both processes usually contain a process condensate treatment section where

volatile components are removed by steam stripping, the advantages of combining these sections have been explored [52–54]. Several attempts for further integration of mass streams of both processes have been published [55–59]. Despite the claimed reduction in both capital and raw material cost, these highly integrated process schemes have not gained acceptance mainly because of their increased complexity.

3.4. Effluents and Effluent Reduction

Gaseous Effluents. There are potentially two sources for air pollution from a urea plant: (1) gaseous ammonia emission from continuous process vents, and (2) urea dust and ammonia emissions from the finishing section (prilling or granulation).

Gaseous Emissions from Process Vents. Noncondensable gases enter the urea process as contaminants in the raw materials, as process air introduced for corrosion protection, and as air leaking into the vacuum sections of the process. At places where these noncondensable gases are vented, proper measures should be taken to minimize ammonia losses. The present state of the art allows reduction of these losses to <0.2 kg of ammonia per tonne of urea produced. This is realized by using conventional absorption techniques. Special attention is required to avoid explosive gas mixtures originating from combustibles (hydrogen, methane) present in the carbon dioxide and ammonia feedstocks, and the air introduced for anticorrosion purposes. Catalytic combustion of the hydrogen present in carbon dioxide, and dosing of nitrogen or excessive amounts of combustibles have been suggested to avoid the risks of formation of explosive gas mixtures [60].

Gaseous Emissions from the Finishing Section (Prilling/Granulation). In the prilling processes, urea dust is produced, mainly from evaporation and subsequent sublimation of urea, but also partly via a chemical mechanism: formation of ammonia and isocyanic acid in the urea melt, evaporation of these components, followed by sublimation of this ammonia – isocyanic acid mixture to form urea dust in the colder air. Urea dust formed in this way is typically very fine (0.5 – 2 μm). Removal of this fine urea dust from the prilling tower exhaust gas is a technical challenge. Because of the small particle size, dry cyclones cannot be used. Instead, wet impingement type devices have proved successful in removing a major part of the dust from the air.

Many new urea plants use granulation instead of prilling as the finishing technique. The main purpose is to improve the size and strength of the product, but less difficulty in controlling dust emissions is also an advantage. The dust produced in these granulation devices (fluidized-bed granulation or drum granulation) is typically much coarser than the fume-like dust produced in a prilling tower.

This results basically from the much shorter contact time of liquid urea with air. Because of the coarser dust and the smaller air quantities to be handled, wet scrubbers provide adequate and much simpler dust emission control for granulation units compared to the apparatus required for emission control in prilling.

Liquid Effluents. The process condensate produced from the evaporation or crystallization sections of the plant contains 3 – 8 wt % ammonia and 0.2 – 2 wt % urea. Two techniques are known to remove these pollutants:

1) Biological treatment [61–63]
2) Chemical hydrolysis and steam stripping to remove ammonia from the condensate

Biological treatment seems to have gained only slight acceptance. The method involving chemical hydrolysis and steam stripping recycles urea (in the form of ammonia) and ammonia to the synthesis section for the production of urea, whereas with biological treatment the urea and ammonia present in the feed to wastewater treatment are lost to urea production. Several chemical hydrolysis and steam-stripping systems are described below.

Stamicarbon System. In the Stamicarbon system [64] first the bulk of ammonia is removed by pre-desorption of the process condensate, followed by hydrolysis of the urea with steam at 170 – 230 °C to form ammonia and carbon dioxide via ammonium carbamate. The hydrolyzer is a vertical bubble-washer column, operated in countercurrent with respect to gas and liquid flow. Ammonia remaining in the process condensate is then removed by further steam stripping (desorption). Since both

the pre-desorber and the desorber operate at low pressure (1 – 5 bar), low-pressure steam as produced in the urea synthesis section can be used as stripping agent. The combination of pre-desorption and countercurrent operation of the hydrolyzer ensures that the chemical equilibrium of the hydrolysis reaction does not limit the minimum achievable urea content in wastewater to concentrations <1 ppm. Also, the remaining NH_3 concentration is <1 ppm.

Snamprogetti System. The Snamprogetti system [65] also includes a system of pre-desorption, hydrolysis, and final desorption with steam stripping. In this process, however, the hydrolyzer is built as a horizontal column with cross-flow operation with respect to gas and liquid flow. Hydrolysis is carried out at somewhat higher temperature (230 – 236 °C) and pressure (33 – 37 bar) than in the Stamicarbon process. Also, this process claims minimum achievable NH_3 and urea concentrations <1 ppm in the liquid effluent.

Toyo Engineering Corporation also offers a system of pre-desorption and hydrolysis, followed by final desorption, as part of its ACES and conventional urea processes. Urea and ammonia concentrations <5 ppm in the effluent are claimed.

UTI System. In the UTI system [66], desorption and hydrolysis take place in a single column, in which the hydrolysis of urea back to ammonium carbamate is carried out concurrently with the stripping of ammonia from the liquid phase. Because of this concept, both the steam required for steam stripping and the steam required to bring the process condensate to a sufficiently high temperature for hydrolysis must be at medium pressure. A small amount of carbon dioxide is fed to the bottom of the column. The system is claimed to achieve residual NH_3 concentrations <10 ppm.

Other Systems. Some systems have been proposed [52–54] in which the wastewater treatment sections of the urea and ammonia plants are combined. Like the systems described above, they also remove urea by hydrolysis to ammonia and carbon dioxide, which are subsequently removed by transferring them into a steam-containing gas phase. The principal difference between these systems and the aforementioned methods is that the ammonia and carbon dioxide produced are recycled to the ammonia plant reforming system, rather than to the urea synthesis section.

3.5. Product-Shaping Technology

Prilling. For a long time, prilling has been used widely as the final shaping technology for urea. In prilling processes, the urea melt is distributed in the form of droplets in a prilling tower. This distribution is performed either by showerheads or by using a rotating prilling bucket equipped with holes. Urea droplets solidify as they fall down the tower, being cooled countercurrently with up-flowing air. The prilling process has several drawbacks:

1) The size of the product is limited to a maximum average diameter of about 2.1 mm. Larger-size product would require uneconomically high prilling towers; moreover, larger droplets tend to be unstable.
2) Very fine dust is formed in the prilling process (see Section 3.4). Removal of this dust is a technically difficult and expensive operation.
3) The crushing strength and shock resistance of prills are limited, making prilled product less suitable for bulk transport over long distances. This problem can be largely overcome with appropriate techniques to improve the physical properties, such as seeding [67] to improve shock resistance or addition of formaldehyde to improve crushing strength, and to suppress the caking tendency.

Granulation [29], [68–70]. The drawbacks of prilling have initiated the development of several granulation techniques. These techniques deviate from the prilling technique in that the urea melt is sprayed on granules, which gradually increase in size as the process continues. The heat of solidification is removed by cooling air or, for some granulation techniques, evaporation of water. Since the contact time between liquid urea and air in these processes is much lower than in prilling, the dust formed in granulation processes is much coarser; therefore, it can be removed much more easily from the cooling air.

Drum granulation systems have been developed, for example, by C & I Girdler,

Kaltenbach – Thuring, and Montedison. *Pan granulation processes* have also been developed, for example, by Norsk Hydro and the Tennessee Valley Authority (TVA). More recently *spouted-bed* and *fluidized-bed granulation techniques* were introduced. A spouted-bed technique was developed by Toyo Engineering Corporation; however, the fluidized-bed technology from Hydro Agri seems to be most successful at this time.

All granulation processes require the addition of formaldehyde or formaldehyde-containing components. Also common to all granulation techniques is that they yield products with larger diameters compared to prilling. However, their capabilities in this respect differ from each other to quite an extent.

Although the improvements brought about by granulation are beyond doubt, prilling techniques still have a place because of the lower investment and lower variable costs associated with prilling compared to granulation.

4. Forms Supplied, Storage, and Transportation

Forms Supplied. Urea may be supplied either in solid form or as a liquid. For *liquid compound fertilizers*, urea is a favorite ingredient. It is generally used in combination with ammonium nitrate as an aqueous solution to obtain liquids containing 32 – 35 wt % nitrogen. These solutions are designated as UAN-32 to UAN-35.

The *solid forms* are generally classified as granular or prilled products, because of the differences in handling properties. Prilled product is considered less suitable for bulk transportation because prills have lower crushing strength, a lower shock resistance, and a higher caking tendency than granules. Because of this, prilled products are usually marginally cheaper than granulated product. Granulated product usually also has a larger diameter (2.0 – 2.5 mm) than prills (1.5 – 2.0 mm), making granules more suitable for bulk blending to produce compound fertilizers.

Special Grades. The majority of urea is designated as "fertilizer grade"; however, some special forms have found limited application:

Technical Grade. Technical-grade urea should be without additions; color, ash-, and metal content are sometimes also specified. For urea used to produce urea – formaldehyde resins, its content of pH-controlling trace components is important. Because of this, technical-grade urea at present is mostly traded as a performance product, rather than being bound to narrow specification limits. The fitness of the product for use is judged by application-specific tests.

Low-Biuret Grade. A maximum biuret content up to 1.2 wt % is considered acceptable for nearly all fertilizer applications of urea. Only for the relative small market segment of foliar spray to citrus crops is a lower biuret content (max. 0.3 %) desirable.

Feed Grade. Some urea is also used directly as a feed component for cattle. Urea used for this purpose should be free of additions. Feed-grade urea is supplied in the form of microprills with a mean diameter of about 0.5 mm.

Slow-Release Grades. Studies show that only 50 – 60 % of fertilizer nitrogen applied to soil is usually recovered by crop plants. Several attempts have been made to increase this percentage by slowing the release of fertilizer to the ground via coating or additions [71].

Urea Supergranules. Granulated product with a very large diameter (up to 15 mm) has found limited application for deep placement in wetland rice [72] and forest fertilization.

Storage. The shift from bagged to bulk transport and storage of prilled and granulated urea has called for warehouse designs in which large quantities of urea can be stored in bulk. These warehouses should be designed in such a way that the product suffers little degradation. Degradation may result from: (1) segregation of fines; (2) disintegration; and (3) absorption, loss, or migration of water.

Segregation of fines can be avoided through uniform product spreading during pouring. *Disintegration* can be minimized by:

1) Providing the product pouring system with a pouring height adjuster
2) Design of "product-friendly" reclaiming systems, because reclaiming the product by means of payloaders and tractor shovels invariably leads to product disintegration

Caking and subsequent product disintegration at unloading are known to result from water absorption. What is not commonly known,

however, is that excessive drying of the product during storage also leads to a higher caking tendency and that migration of water from warm product in the bulk of a pile to the cold surface leads to crust formation. Thus, attempts to decrease water absorption through refrigeration or air conditioning, dehumidification, or space heating may cause the air in the warehouse to become too dry or may result in too great a temperature difference between the product and the surrounding air. Instead, the warehouse (especially the roof) should be airtight and thoroughly insulated. The caking tendency of urea can be reduced by addition of small amounts of formaldehyde (up to 0.6 wt %) to the urea melt or by addition of surfactants to the solid product.

Transportation. Urea prills and granules are transported by bulk transport in trucks, ships, rail cars, etc. To withstand numerous and rapid loading and unloading operations, product for bulk transport should have a high initial physical stability. Great demands are made, especially on the shock resistance of the product, e.g., at seaport loading and unloading facilities. In addition, a number of "good housekeeping" rules should be adhered to:

1) Do not load or unload during rain
2) Make sure that the means of transport is clean and dry
3) Close the ship's hold when rain is imminent
4) Do not replace the air above the product or ventilate the holds
5) Cover the product (e.g., by polyethylene sheeting) during prolonged transport
6) Product should be spread rather then poured solely from one point to prevent dust coning due to segregation
7) Restrict the pouring height to avoid unnecessary disintegration

Liquid Fertilizer Transport. Liquid fertilizers are transported by tank cars, railway tanks, ships, and pipelines. Although liquid fertilizer is generally accepted as the most economic form to distribute, the solid form is still the most popular by far. Distribution of large quantities of liquid fertilizer requires a complex infrastructure and is limited at present to large farm units in developed countries. Transport and storage of UAN solutions in carbon steel lines and tanks require the addition of a corrosion inhibitor to the solution.

5. Quality Specifications and Analysis

Typical quality specifications for fertilizer-grade urea are summarized in Table 2. The capabilities of a modern urea plant are better than the typical trade data given in this table.

Table 2. Typical product specifications for fertilizer-grade urea

Specification	Prilled product	Granulated product
Nitrogen content, wt %	min. 46	min. 46
Biuret content, wt %	max. 1	max. 1
Water content, wt %	max. 0.3	max. 0.25
Crushing strength, bar	20 – 25	30 – 60
Shock resistance, wt %	min. 85	100
Product size		
1.0 – 2.4 mm, wt %	90 – 95	
1.6 – 4.0 mm, wt %		95
Bulk density (loose), kg/m^3	730	750 – 790

The total *nitrogen content* is usually determined by digesting urea with sulfuric acid toyield ammonium sulfate. The ammonia content is then determined by distillation and titration. Alternatively, the total N content may be determined by the Kjeldahl method or by using a method based on hydrolyzing urea with urease followed by titration of the ammonia formed.

The *water* content is usually determined with Karl Fischer reagent (→ Gas Production, Chap. 8.2.2.3.).

Biuret is determined by the formation of a violet-colored complex of biuret with copper(II) sulfate in an alkaline medium and subsequent measurement of the absorbance of the colored solution at 546 nm.

Crushing strength is defined as the force required per unit cross-sectional area of a granule to crush the granule or, if it is not crushed, the force at which it is deformed by 0.1 mm. A single granule is subjected to a force that is increased at a constant rate, the force at breakage (or at 0.1-mm deformation) being recorded.

The *shock resistance* of granules is defined as the weight percentage of a sample that is not crushed when subjected to a specified shock load. To determine shock resistance, a sample of prills or granules is shot against a metal plate

by means of compressed air under normalized conditions. The amount of nondamaged product that remains after the test is determined.

The *granulometry* of the product is measured by conventional sieve techniques.

6. Uses

Urea is used for soil and leaf fertilization (more than 90 % of the total use); in the manufacture of urea – formaldehyde resins; in melamine production; as a nutrient for ruminants (cattle feed); and in miscellaneous applications.

Soil and Leaf Fertilization. Worldwide, urea has become the most important nitrogenous fertilizer. Urea has the highest nitrogen content of all solid nitrogenous fertilizers; therefore, its transportation costs per tonne of nitrogen nutrient are lowest. Urea is highly soluble in water and thus very suitable for use in fertilizer solutions (e.g., "foliar feed" fertilizers). Urea is also used as a raw material for the production of compound fertilizers. Compound fertilizers may be produced by mixing in urea melts or urea solutions before shaping the compound fertilizers or by mixing solid urea prills or granules with other fertilizers (bulk blending). In the latter case, the product sizes must match to prevent segregation of the products during further handling.

Urea – Formaldehyde Resins (→Amino Resins, Chap. 7.1.). A significant proportion of urea production is used in the preparation of urea – formaldehyde resins. These synthetic resins are employed in the manufacture of adhesives, molding powders, varnishes, and foams. They are also used to impregnate paper, textiles, and leather.

Melamine Production. At present, nearly all melamine production is based on urea as a feedstock (→Melamine and Guanamines, Chap. 4.). Since ammonia is formed as a coproduct in melamine production from urea (see Chap. 1), integration of the urea and melamine production processes is beneficial.

Feed for Cattle and other Ruminants. Because of the activity of microorganisms in their cud, ruminants can metabolize certain nitrogen-containing compounds, such as urea, as protein substitutes. In the United States this capability is exploited on a large scale. In Western Europe, by contrast, not much urea is used in cattle feed.

Other Uses. On a smaller scale, urea is employed as a raw material or auxiliary in the pharmaceutical industry, the fermenting and brewing industries, and the petroleum industry. Urea can be used for the removal of NO_x from flue gases. Urea is also used as a solubilizing agent for proteins and starches, and as a deicing agent for airport runways.

7. Economic Aspects

The predicted growth in *demand* for urea until 1997 is slightly more than 3 % per year, bringing the total urea demand in 1997 to some 89×10^6 t/a. Most of the growth will occur in Asia, with China and India in the lead. A little more than 7 % of the worldwide demand for urea is from industry, in which Europe takes a leading role, ahead of North America and the industrialized countries of Asia.

The predicted growth in *capacity* up to 1997 will be around 1.5 % per year, resulting in a total installed capacity of some 100×10^6 t by 1997. Between 1994 and 1997, new plants will account for an increase in capacity of ca. 6.5×10^6 t, whereas closure of old plants will result in a reduction of the installed capacity by some 2×10^6 t. The world's capacity utilization can thus be calculated to be ca. 89 %.

Because of geopolitical changes during the early 1990s, different economic laws have become valid in the former Soviet Union and Eastern Europe, causing those countries to supply to the world market large amounts of prilled urea at prices far below the cost to many producers [e.g., sales prices, free on board (FOB) Black Sea $ 75 per tonne in 1993]. This situation is not expected to last long, although FOB prices at Black Sea ports will generally remain lower compared to other places. In the 1990s urea prices have shown large fluctuations. For instance, in 1994 – 1995 lows of $ 100 per tonne and highs of $ 250 per tonne have been reported. Predictions of future developments of urea prices are therefore highly speculative.

To cope with tough competition on the world market, producers have the tendency to build plants with very high single line capacities (2000 t/d or more), in which operating reliability is of extreme importance. Uninterrupted operating periods of more than one year are often achieved.

Furthermore, producers are increasingly shifting production facilities (both new plants and relocations) to places where natural gas is plentiful and cheap. Europe seems to have lost its competitive edge in export markets due to its expensive feedstocks.

The investment costs (in 10^6 \$) for a present state-of-the-art total-recycle urea plant are estimated to be:

1000-t/d plant (single line) 43
1500-t/d plant (single line) 52
2000-t/d plant (single line) 62

8. Urea Derivatives

Barbituric acids and derivatives, see →Hypnotics, Chap. 5.1.

8.1. Thermal Condensation Products of Urea

Thermolysis of urea gives biuret, triuret, and cyanuric acid, and in a special process, melamine is produced.

Biuret [*108-19-0*], imidodicarbonic diamide, $H_2NCONHCONH_2$, *mp* 193 °C, is produced by heating urea in inert hydrocarbons at 110–125 °C or in the melt at 127 °C [73], [74].

Triuret [*556-99-0*], diimidotricarbonic diamide, $H_2NCONHCONHCONH_2$, *mp* 231 °C, is produced by decomposing urea in a thin film at 120–125 °C [75]. It is also obtained by treating 2 mol of urea with 1 mol of phosgene in toluene at 70–80 °C [76], [77].

Cyanuric acid [*108-80-5*], 1,3,5-triazine-2,4,6 (1*H*,3*H*,5*H*)-trione (**1**), is formed on heating urea in the presence of zinc chloride, sulfuryl chloride, or chlorosulfuric acid [78].

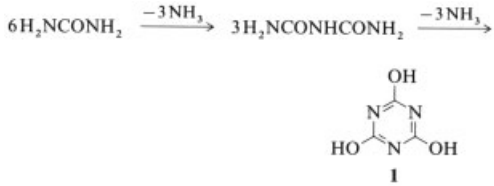

Melamine [*108-78-1*], 1,3,5-triazine-2,4,6-triamine (**2**) is produced industrially from urea [79] (→Melamine and Guanamines, Chap. 4.).

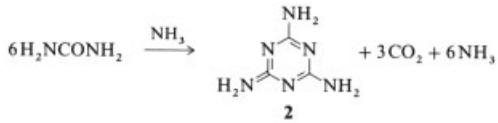

8.2. Alkyl- and Arylureas

Most simple substituted alkylureas are crystalline products. Tetramethyl- and tetraethylurea and some cyclic ureas are liquids. Alkylated and arylated ureas are used in the production of plant protection agents, in pharmaceutical and dye chemistry, as plasticizers, and as stabilizers. Alkylureas and polyalkyleneureas are used as additives in the production of aminoplastics.

Various processes can be used for the production of substituted ureas, the most important of which are listed in Table 3.

Table 3. Production processes for substituted ureas

Starting materials	Product
Urea and amines	ureas substituted at one or both nitrogen atoms and cyclic ureas
Urea and tertiary alcohols	ureas substituted at one nitrogen atom
Phosgene and amines	ureas substituted at one nitrogen atom
Isocyanates and NH_3 or amines	ureas substituted at one or both nitrogen atoms
Carbamoyl chloride and NH_3 or amines	ureas substituted at one or both nitrogen atoms
Esters of carbonic or carbamic acids	ureas substituted at one or both nitrogen atoms
Urea and aldehydes or ketones	ureas substituted at one or both nitrogen atoms and cyclic ureas

8.2.1. Transamidation of Urea with Amines

The transamidation of urea with amines is one of the most important industrial production processes for substituted ureas. Monosubstituted

ureas are produced by condensation of urea with a sufficiently basic amine in a 1 : 1 molar ratio in the melt at 130 – 150 °C [80]. Symmetrically disubstituted ureas can be produced from 2 mol of amine and 1 mol of urea at 140 – 170 °C [81]. Instead of the amines, amine salts can also be reacted with urea in the melt or by prolonged boiling in aqueous solution [82].

Monomethylurea [*598-50-5*], $CH_3NHCONH_2$, *mp* 102 °C, is produced industrially by passing monomethylamine into a urea melt [83]. Monomethylurea is used for the synthesis of theobromine.

Phenylurea [*64-10-8*], $C_6H_5HNCONH_2$, *mp* 147 °C, can be produced by heating an aqueous solution of aniline hydrochloride and urea to its boiling point [84].

Symmetric *N,N′*-dimethylurea [*96-31-1*], $CH_3HNCONHCH_3$, *mp* 105 °C, is used for the synthesis of caffeine by the Traube method and for the production of formaldehyde-free easy-care finishing agents for textiles [84].

Hexamethylenediurea [*2188-09-2*], $H_2NCONH(CH_2)_6NHCONH_2$, is obtained by heating a mixture of hexamethylenediamine with an excess of urea at 130 – 140 °C, with elimination of ammonia [85], [86].

***N,N′*-Diphenylurea** [*102-07-8*], carbanilide, $C_6H_5HNCONHC_6H_5$, *mp* 238 °C, can be produced in high yields by heating 2 mol of aniline with 1 mol of urea in glacial acetic acid [86], [87].

Polyalkyleneureas can be obtained by treating di-, tri-, and tetraalkylenamines in concentrated aqueous solution with urea at ca. 110 °C. They are used, for example, to modify melamine – formaldehyde impregnating resins and binders for derived timber products. Examples of polyalkyleneureas include the following:

1,2-Ethylenediurea, $H_2NCONH-CH_2CH_2-HNCONH_2$, *mp* 198 °C
1,3-Propylenediurea, $H_2NCONH-CH_2CH_2CH_2-HNCONH_2$, *mp* 186 °C
Diethylenetriurea, $H_2NCONH-CH_2CH_2-N(CONH_2)-CH_2CH_2-HNCONH_2$, *mp* 215 °C
Dipropylenetriurea, $H_2NCONH-(CH_2)_3-N(CONH_2)-(CH_2)_3-HNCONH_2$
2-Hydroxypropylene-1,3-diurea, $H_2NCONH-CH_2CH(OH)-CH_2-HNCONH_2$, *mp* 147 °C

Cycloalkyleneureas. *2-Imidazolidinone* [*120-93-4*], ethyleneurea (**3**), *mp* 131 °C [88], [89]; *2-oxohexahydropyrimidine* [*65405-39-2*], propyleneurea (**4**), *mp* 260 – 265 °C [90]; and *2-oxo-5-hydroxyhexahydropyrimidine*, 5-hydroxypropyleneurea (**5**), are produced industrially in the melt above 180 °C, preferably at 200 – 230 °C, by condensation of 1,2-ethylenediamine or 1,3-propylenediamine with urea and elimination of ammonia. These cycloalkyleneureas are used in the form of their *N,N′*-dihydroxymethyl compounds for easy-care finishes for cellulose-containing textiles [91].

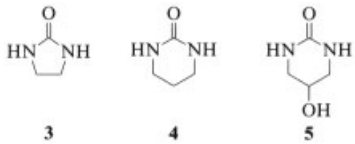

8.2.2. Alkylation of Urea with Tertiary Alcohols

Urea can be alkylated with tertiary alcohols in the presence of sulfuric acid [92], [93].

***tert*-Butylurea,** $(CH_3)_3CHNCONH_2$, *mp* 182 °C, is produced by treating urea with 2 mol of *tert*-butanol in the presence of ca. 2 mol of concentrated sulfuric acid at 20 – 25 °C with ice cooling [92–94].

8.2.3. Phosgenation of Amines

Symmetrically disubstituted ureas are produced in good yields by passing phosgene into solutions of amines in aromatic hydrocarbons [95]. In some cases, aqueous solutions or suspensions of amines can also be reacted with

phosgenes [96]. The phosgenation of mixed aliphatic – aromatic amines is carried out industrially in the presence of sodium hydroxide at 40 – 60 °C [97]. For the production of substantive dyes, aminosulfone or aminocarboxylic acid groups are bonded by means of phosgenation [98]

$$2\,RNH_2 + COCl_2 \rightarrow RNH-CO-NHR + 2\,HCl$$

Tetramethylurea [*632-22-4*], $(CH_3)_2NCON(CH_3)_2$, *bp* 156.5 °C, is produced from dimethylamine and phosgene and is used as an aprotic solvent [99].

Asymmetric Diphenylurea, $(C_6H_5)_2NCONH_2$, *mp* 189 °C, is produced from diphenylamine, phosgene, and ammonia:

$$COCl_2 + (C_6H_5)_2NH + 3\,NH_3 \rightarrow (C_6H_5)_2NCONH_2 + 2\,NH_4Cl$$

Symmetric dimethyldiphenylurea, *mp* 121 – 127 °C, is obtained by treating phosgene with monomethylaniline and sodium hydroxide.

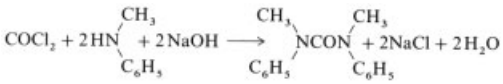

Symmetric dialkyldiarylureas are used under the name Centralite as plasticizers and stabilizers for nitrocellulose and propellants [97].

8.2.4. Reaction of Amines with Cyanates (Salts)

The salts of aliphatic or aromatic amines react with potassium cyanate at 20 – 60 °C to give substituted ureas in high yields [100], [101].

$$R-NH_2\cdot HX + KNCO + H_2O \rightarrow R-HNCONH_2 + KX + OH^-$$

4-Ethoxyphenylurea [*150-69-6*] (dulcin) is produced from potassium cyanate and *p*-phenetidine hydrochloride in aqueous solution at room temperature [100]. This production method can be applied to aminosulfonic and aminocarboxylic acids, whereby the betaine-like salts formed by these acids react with potassium cyanate to give ureas that are substituted at only one NH_2 group [101]. Sulfonamides react with potassium cyanate to give potassium salts of the corresponding sulfonylureas, from which the sulfonylureas are obtained by acidification [102]:

$$RSO_2NH_2 + KNCO \rightarrow RSO_2N^-CONH_2\,K^+ \rightarrow RSO_2HNCONH_2$$

8.2.5. Reaction with Isocyanates

Symmetrically disubstituted ureas can also be produced from isocyanates by prolonged heating in aqueous solution [103], [104]:

$$RNCO \xrightarrow{H_2O} RHNCOOH \xrightarrow{RNCO} RHNCONHR + CO_2$$

When ammonia or primary or secondary amines are reacted with isocyanates, the corresponding substituted ureas are obtained in almost quantitative yield [104–106]. This process is particularly suitable for the production of unsymmetrically substituted ureas [105], [106].

8.2.6. Acylation of Ammonia or Amines with Carbamoyl Chlorides

Ammonia and primary or secondary amines react with carbamoyl chlorides to give the corresponding urea derivatives in good yields [107].

Tetraphenylurea, $(C_6H_5)_2NCON(C_6H_5)_2$, [108] is obtained in quantitative yield by heating diphenylcarbamoyl chloride with diphenylamine.

8.2.7. Aminolysis of Esters of Carbonic and Carbamic Acids

Esters of carbonic and carbamic acids (carbonates and carbamates) react with amines at ele-

vated temperatures to give symmetrically disubstituted ureas [109–112].

8.3. Reaction of Urea and Its Derivatives with Aldehydes

8.3.1. α-Hydroxyalkylureas

The industrial production of α-hydroxyalkylureas is limited to the addition of formaldehyde or glyoxal to urea, monoalkylureas, symmetrical dialkylureas, and cyclic ureas. It involves acid- or base-catalyzed additions that are generally equilibrium reactions [113–130].

Urea can bond with up to 4 mol of formaldehyde. However, only monohydroxymethyl- and *N,N′*-dihydroxymethylurea can be isolated in pure form [131], [132]. Tri- and tetrahydroxymethylureas are formed only as nonisolable intermediates, for example, in the synthesis of trimethoxymethylurea (**6**), [133], [134]; *N,N′*-dialkoxymethyl-4-oxomethyltetrahydro-1,3,5-oxadiazine (**7**) [135], [136]; and *N,N′*-dihydroxymethyl-2-oxo-5-alkyltetrahydro-1,3,5-triazines (**8**) [137–140].

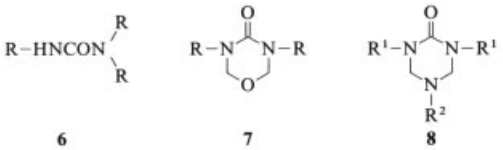

6 **7** **8**

R = Alkoxymethyl
R^1 = Hydroxymethyl
R^2 = Alkyl

***N,N′*-Dihydroxymethylurea** [*140-95-4*], $HOCH_2HNCONHCH_2OH$ [131], [132] is produced industrially by charging 2 mol of formaldehyde per mole of urea to a stirred vessel. The solution is neutralized with triethanolamine. Urea is added with cooling, and the temperature must not exceed 40 °C in this slightly exothermic reaction. After a few hours the reaction mixture is cooled to room temperature and dihydroxymethylurea crystallizes out. The product is dried in a spray-drying tower.

Hydroxymethyl derivatives of cyclic ureas can be produced by reaction of these substances with formaldehyde in an alkaline medium [141].

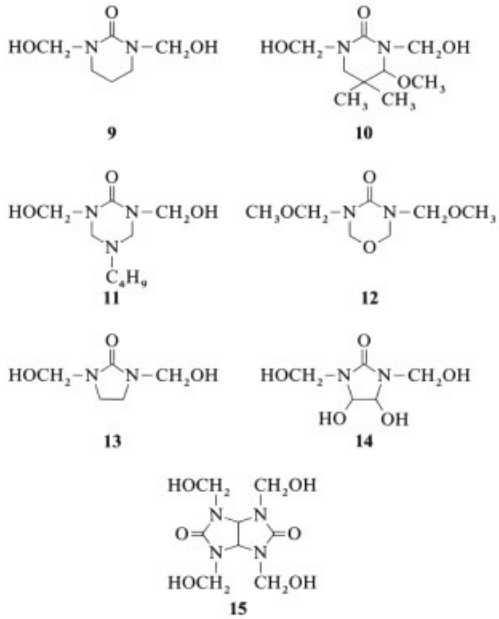

9 **10** **11** **12** **13** **14** **15**

Hydroxymethyl derivatives of urea and cyclic ureas (**9**)–(**15**) are used in easy-care finishes for textiles [91].

The addition of higher aldehydes to urea, mono- and symmetrically disubstituted ureas, and cyclic ureas generally gives unstable α-hydroxyalkyl compounds. Electron-withdrawing and electron-donating substituents next to the α-hydroxyalkyl group affect the stability of these compounds and also their ability to undergo condensations.

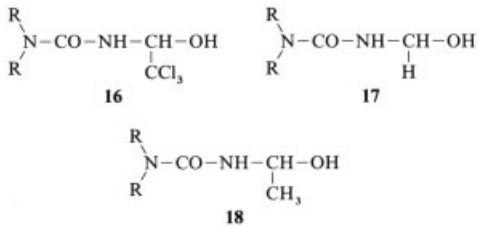

16 **17** **18**

For example, the chloral – urea derivatives (**16**) exhibit considerable differences in reactivity compared with the *N*-hydroxymethyl (**17**) and *N*-α-hydroxyethyl compounds (**18**). These differences are exemplified by a decrease in the H-acidity of the OH groups [142]. Chloral compounds (**16**) can form alkali-metal salts, whereas the corresponding salts of *N*-hydroxymethyl (**17**) and *N*-α-hydroxyethyl compounds (**18**) are

unknown. Condensation of chloral compounds with nucleophiles is possible only under extreme reaction conditions. However, N-α-hydroxyethyl compounds can be converted smoothly with alcohol into N-α-alkoxyethyl compounds in basic and sometimes even in neutral media. The reactivity of N-hydroxymethyl compounds lies between that of the chloral and N-α-hydroxyethyl compounds. The 4-hydroxycycloalkyleneureas (cyclic N-hemiacetals), which are obtained by treating suitable aldehydes with ureas, are stable. For example, 2-oxo-4,5-dihydroxyimidazolidines are formed by cyclization of urea, or its mono- or symmetrically disubstituted derivatives, with glyoxal [143–145]:

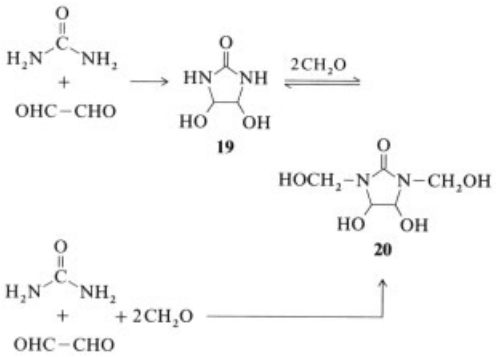

***N,N′*-Dihydroxymethyl-2-oxo-4,5-dihydroxyimidazolidine** (**20**) is produced by hydroxymethylation of 4,5-dihydroxy-2-oxoimidazolidine (**19**) in weakly acidic to weakly alkaline aqueous solution or, more elegantly, by direct reaction of urea with glyoxal and formaldehyde in the appropriate molar ratio in weakly acidic to neutral solution at 40 – 80 °C, sometimes in the presence of catalytically active buffers [130].

Compound **20** and its derivative in which the OH groups are partly acetalized with methanol are used as formaldehyde-free cross-linking agents for easy-care finishes for cellulose-containing textiles.

2-Oxo-4-hydroxyhexahydropyrimidines also belong to the group of α-hydroxyalkylureas. These compounds can be produced industrially by cyclocondensation of urea with active enolizable aldehydes (see Section 8.3.4).

8.3.2. α-Alkoxyalkylureas

Condensation of N-hydroxymethylureas with alcohols to give N-alkoxymethylureas (ureidoalkylation of alcohols) is of great industrial importance. Pure hydroxymethylureas and an excess of alcohol are reacted in the presence of catalytic amounts of acid. The nature and quantity of the acid catalyst depend on the reactivity of the N-hydroxymethyl compound, its stability to hydrolysis, and the formation of byproducts and polycondensation products. At elevated temperature the reaction can be carried out under weakly acidic conditions, whereby the equilibrium position must be adjusted by variation of the concentration and the molar ratios [91], [146], [147]. Sometimes, ureidomethylation of alcohols is better at room temperature in the presence of strong acids.

The alcohol-modified urea – formaldehyde condensation products are used as resins for heat- or acid-curing coatings. Besides the water-soluble or almost solvent-free resins, which are becoming increasingly important for environmental reasons, a wide range of aminoplastic resins for coatings exist that are readily soluble in common paint solvents. To convert urea – formaldehyde resins to resins that are soluble in organic solvents, the highly polar N-hydroxymethyl groups obtained in the initial reaction between urea and formaldehyde are acetalized with alcohols, mainly butanol and isobutanol, as well as ethanol and methanol or their mixtures.

The alcohol-modified urea – formaldehyde resins are produced industrially by passing aliphatic alcohols into aqueous solutions of urea and formaldehyde or solutions of hydroxymethylated ureas in the presence of small quantities of acid at 90 – 100 °C so that the water formed and excess alcohol distill off. The molar ratios vary between 2 and 4 mol of formaldehyde and 2 and 5 mol of alcohol per mole of urea. The process is carried out in a reactor equipped with an adequately dimensioned heat exchanger, a vacuum pump, a distillation column, and for alcohols that are sparingly soluble in water, a water separator.

Some *4-hydroxycycloalkyleneureas* are so reactive that they can be converted into the N-α-alkoxy compounds (**21**), (**22**) even in a neutral medium by heating with alcohol [148]:

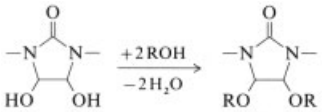

21

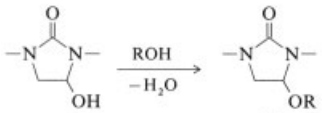

22

To shift the equilibrium in the N-α-ureidoalkylation of alcohols in the direction of the N-α-alkoxyalkyl compounds, the water formed during the reaction must be removed. In industrial production processes an aprotic entrainer (e.g., an aromatic) is used. In the ureidomethylation of alcohols that are immiscible or sparingly miscible with water, an excess of alcohol is used and water is removed by azeotropic distillation.

Higher-boiling alcohols can also be ureidomethylated by transacetalization of the N-methoxymethyl compounds.

Polymerizable compounds are obtained by ureidoalkylation of unsaturated alcohols (e.g., allyl alcohol [149], [150]). Tetraallyloxymethyltetrahydroimidazo[4,5-*d*]imidazole-2,5(1*H*,3*H*)-dione (**23**) has achieved importance as a polymerizable coating component [151], [152]:

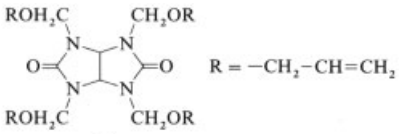

23

8.3.3. α,α'-Alkyleneureas

α,α'-Alkyleneureas are obtained by condensation of urea or its derivatives with aldehydes in a weakly acidic medium. The aldehyde group first adds to the urea to form an α-hydroxyalkylurea, which then reacts with a second molecule of urea or with another α-hydroxyalkylurea to give linear or branched α,α'-alkyleneureas:

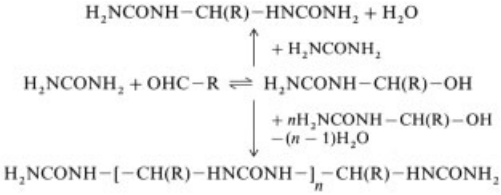

Reaction of equimolar quantities of urea and formaldehyde in acidic solution gives polymethyleneureas as a result of stepwise ureidomethylations:

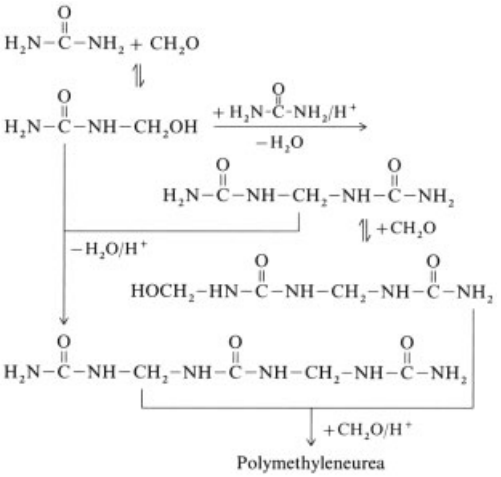

KADOWAKI obtained polymethyleneureas containing up to five urea groups joined by methylene bridges by means of a stepwise synthesis [135]. Because of their extreme insolubility, no higher polymethyleneureas have yet been isolated. Polymethyleneureas are used as slow-release nitrogen fertilizers, for example.

Isobutylidenediurea (**24**), which is sparingly soluble in water, is obtained by condensation of urea with isobutyraldehyde in a molar ratio of 2 : 1 in a weakly acidic medium:

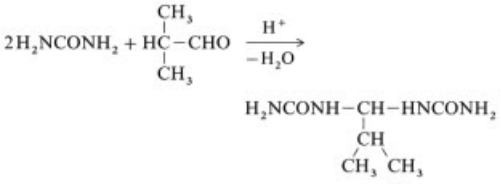

24

Isobutylidenediurea is also a slow-release nitrogen fertilizer used in various special fertil-

izer formulations. Isobutylidenediurea is produced by a continuous process. According to a patent published by Mitsubishi Chemical Industries [153], urea is charged continuously with a screw feed via a belt weigher and is reacted with a stoichiometric quantity of isobutyraldehyde in the presence of semiconcentrated sulfuric acid in a mixer. In the last section of the mixer the reaction product is neutralized by injecting dilute aqueous potassium hydroxide solution. The product is dried by using plate driers and processed by sieving, filtering, and grinding.

8.3.4. Cyclic Urea – Aldehyde Condensation Products

Almost all cyclizations of urea and its derivatives with aldehydes involve an α- or a vinylogous ureidoalkylation [113–115], [146]. If the urea bears a nucleophilic substituent on the second nitrogen atom, cyclocondensation occurs [154]. Saturated and unsaturated cyclic ureas with five, six, seven, or eight ring atoms, bicyclic and polycyclic heterocycles (with both uncondensed rings and rings anellated in the 1,2- or 1,3-position), and spiro compounds can be produced this way [154], [155].

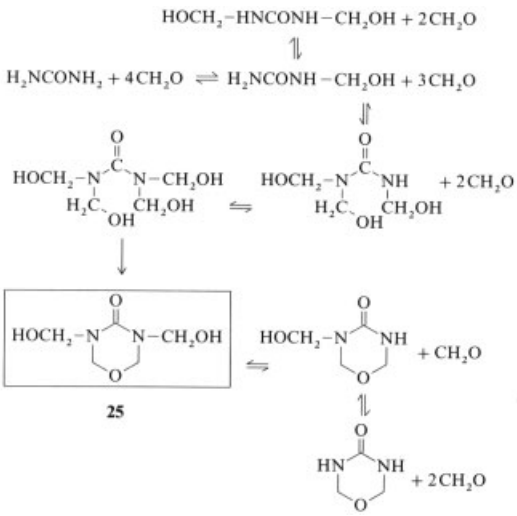

Industrially important reactions are those of urea with formaldehyde, acetaldehyde, isobutyraldehyde, and their mixtures. Treatment of urea with formaldehyde in a molar ratio of 1 : $\geq$4 gives an equilibrium mixture of hydroxymethyl derivatives. On ureidomethylation of a hydroxymethyl group bonded to the second nitrogen atom of the urea, cyclocondensation to hydroxymethylated 4-oxotetrahydro-1,3,5-oxadiazines (**25**) occurs [135], [136].

4-Oxo-3,5-dialkoxymethyltetrahydro-1,3,5-oxadiazines are used as cross-linking agents for easy-care finishing of textiles. These compounds can be obtained directly by condensation of urea with formaldehyde and alcohols [135], [136]:

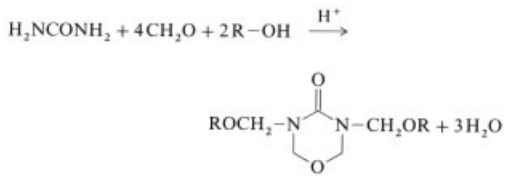

The hydroxymethyl and methoxymethyl derivatives of *2-oxo-5-alkylhexahydro-1,3,5-triazines* have achieved importance as cross-linking agents for easy-care finishing of cellulose-containing fabrics. These compounds are obtained by cyclizing ureidomethylation of urea with formaldehyde and a primary amine [137–140], [154]:

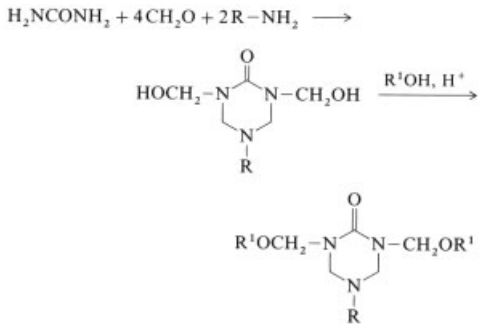

Bicyclic Ureas. 4,5-Dihydroxy- or 4,5-dialkoxyimidazolidin-2-ones can be converted into bicyclic ureas, such as *tetrahydroimidazo*[*4,5-d*]*-imidazole-2,5*(*1H,3H*)*-dione* (**26**), by means of a double α-ureidoalkylation with urea:

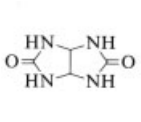

26

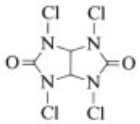

27

Tetrahydroimidazo[4,5-*d*]imidazole-2,5(1*H*, 3*H*)-dione [*496-46-8*], acetylenediurea (**26**), can be obtained directly by condensation of glyoxal with excess urea in an acidic medium [156], [157]. A solution of urea is acidified to pH <3 with sulfuric acid, and a 40 % glyoxal solution is added slowly. At the end of the reaction the molar ratio of glyoxal to urea is adjusted to at least 1 : 2.5. Reaction temperature should not exceed 70 °C. The precipitated product is neutralized by decantation and stirring with water.

In the form of its tetrahydroxymethyl derivative (**15**) the product has achieved importance as a cross-linking agent in easy-care textile finishes [158]. The saturated and unsaturated *N*-alkoxymethyl compounds (**23**) are used as cross-linking agents in the paint and coating industry.

The tetrachloro compound (**27**) is used as a chlorine-transfer agent for mild chlorination reactions and as a bleaching agent for textiles [159–161].

The synthetic possibilities for cyclizing α-ureidoalkylations with compounds containing enolizable carbonyl groups are manifold. Aldehydes and ketones are particularly important because they react not only as nucleophiles but also as carbonyl components in these cyclocondensations. Under mild conditions (i.e., in weakly acidic media), aliphatic and aromatic aldehydes generally react only as carbonyl components in reactions with urea and its derivatives, giving linear or branched condensation products. In a more strongly acidic medium and at elevated temperature, however, an α-ureidoalkylation at nucleophilic α-carbon atom in the aldehyde occurs. (see right column)

In the cyclocondensation of urea or its mono- or symmetrically disubstituted derivatives with aldehydes, which have at least one activated hydrogen in α-position to the carbonyl group, 2-oxo-4-hydroxyhexahydropyrimidines (**29**) are formed in the presence of an acid [113], [154], [162], [163]. Symmetrical di-α-hydroxyalkyl compounds (**28**) and poly-α-alkyleneureas are formed as intermediates. They all contain the groups necessary for cyclocondensation involving an α-ureidoalkylation, i.e., the NH component, 1 mol of aldehyde as the carbonyl component, and the second mole of aldehyde as the nucleophilic component with activated α-hydrogen. In cyclocondensation, a bond is formed between the nucleophilic α-carbon atom of one α-hydroxyalkyl group and the carbon α to the urea nitrogen in the second α-hydroxyalkyl group to give 2-oxo-4-hydroxyhexahydropyrimidines (**29**).

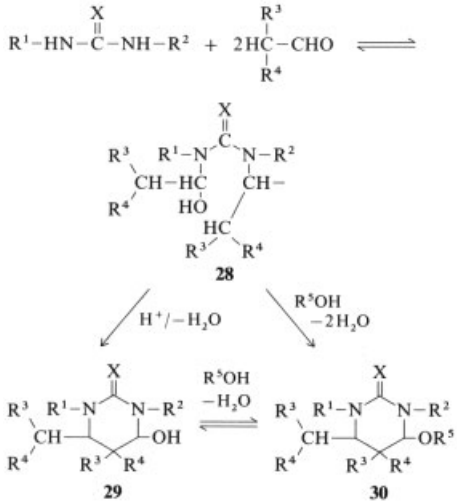

Preparation of 2-Oxo-4-hydroxy-5,5-dimethyl-6-isopropylhexahydropyrimidine (*mp* 220–222 °C). One liter of a 30 % aqueous urea solution (5 mol) is treated with 100 mL of 50 % sulfuric acid in a stirred tank reactor with reflux cooling. Then, 10 mol isobutyraldehyde is added with stirring. After being heated at 85 °C for 2 h, the product is filtered and washed with water.

If these cyclocondensations are carried out in the presence of alcohols, 2-oxo-4-alkoxyhexahydropyrimidines (**30**) are formed [163], [164].

These cyclocondensations can also be carried out between urea and two different aldehydes, at least one of which must have an enolizable carbonyl group. For example, the cyclocondensation of 1 mol of urea with 1 mol of formaldehyde and 1 mol of isobutyraldehyde in the presence of an acid gives 2-oxo-4-hydroxy-5,5-dimethylhexahydropyrimidine (**31**) [164]

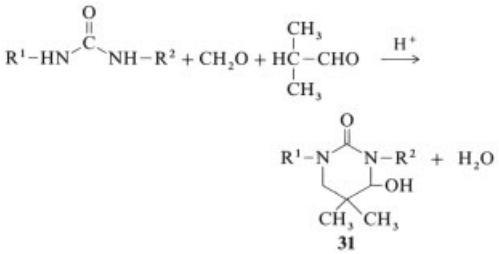

The 2-oxo-4-hydroxy-5,5-dimethylhexahydropyrimidines are cyclic *N*-hemiacetals, whose OH groups can undergo nucleophilic substitutions similar to those undergone by *N*-hydroxymethylureas. 2-Oxo-4-hydroxy- and 2-oxo-4-alkoxy-5,5-dialkylhexahydropyrimidines are used in the form of their *N,N′*-dihydroxymethyl compounds in noncrease finishes for cellulose-containing textiles [91], [165].

Preparation of *N,N′*-Dihydroxymethyl-2-oxo-4-hydroxy(methoxy)-5,5-dimethylhexahydropyrimidine. Urea is treated with formaldehyde in the molar ratio 1 : 1 at pH >9 and 50 – 60 °C in the presence of an excess of methanol. Isobutyraldehyde is added in the presence of a strong mineral acid to bring about cyclocondensation. The reaction mixture is rendered alkaline, and hydroxymethylation is carried out with formaldehyde.

2-Oxo-4-hydroxyhexahydropyrimidines (**32**), which are not or only mono-substituted in the 5-position, react with ureas to give 2-oxo-4-ureidohexahydropyrimidines (**32**).

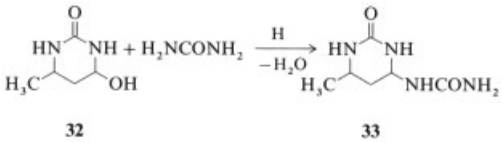

2-Oxo-4-ureido-6-methylhexahydropyrimidine [*1129-42-6*], crotonylenediurea (**33**), is obtained either by condensation of urea with crotonaldehyde in the presence of acid [166] or by the industrially more straightforward route involving condensation of urea with acetaldehyde in the molar ratio 1 : 1 in the presence of acid [167], [168]. 2,7-Dioxo-4,5-dimethyldecahydropyrimido[4,5-*d*]pyrimidine is formed as a byproduct in the second route [168–170].

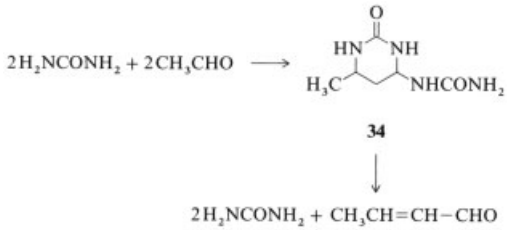

2-Oxo-4-ureido-6-methylhexahydropyrimidine is produced industrially in a continuous process employing a stirred tank cascade. A 70 % urea solution is treated with acetaldehyde at a molar ratio of 1 : 1 in the presence of a catalytic quantity of 75 % sulfuric acid. The exothermic reaction is kept at 38 – 60 °C by controlled cooling. The pH is initially kept above 3 and in the final reactors below 2 by addition of sulfuric acid. Average residence time in the cascade is 40 min. In the last stirred tank, the reaction mixture is neutralized with aqueous potassium hydroxide solution. Drying is carried out in a spray tower.

2-Oxo-4-ureido-6-methylhexahydropyrimidine is used as a slow-release nitrogen fertilizer [171]. This fertilizer is characterized by its extreme insolubility in water and is therefore not washed out of the soil by rain or irrigation. It decomposes as a result of acid hydrolysis induced by humic acids during the growth period of plants, bringing about mineralization of the nitrogen it contains.

2-Oxo-4-hydroxy-5,5-dimethylhexahydropyrimidyl-*N,N′*-bisneopentals can be produced from urea, formaldehyde, and isobutyraldehyde in acid-catalyzed cyclo- and linear condensations. These bisneopentals react further according to a Claisen – Tishchenko reaction to give soft and hard resins with good light stability for the paint and coatings industry, depending on the molar ratios of the starting materials [172], [173].

9. References

1. F. Wöhler, *Ann. Phys. Chem.* **2** (1828) no. 12, 253 – 256.
2. A. I. Basaroff, *J. Prakt. Chem.* **2** (1870) no. 1, 283.

3. J. Berliner, *Ind. Eng. Chem.* **28** (1936) no. 5, 517 – 522.
4. L. Vogel, H. Schubert, *Chem. Tech. (Leipzig)* **32** (1980) no. 3, 143 – 144.
5. A. A. Kozyro, S. V. Dalidovich, A. P. Krasulin, *J. Appl. Chem. (Leningrad)* **59** (1986) no. 7, 1353 – 1355. (*Zh. Prikl. Khim. (Leningrad)* **59** (1986) no. 7, 1456 – 1459.).
6. A. P. Krasulin, A. A. Kozyro, G. Ya. Kabo, *J. Appl. Chem. (Leningrad)* **60** (1987) no. 1, 96 – 99. (*Zh. Prikl. Khim. (Leningrad)* (1987) no. 1, 104 – 108).
7. G. Midgley, *The Chem. Eng.* (1977) Dec., 856 – 866.
8. E. P. Egan, B. B. Luff, *J. Chem. Eng. Data* **11** (1966) no. 2, 192 – 194.
9. A. Seidell: *Solubilities of Organic Compounds,* 3rd ed., Van Nostrand Company, New York 1941, **vol. 2.**
10. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Organic Compounds,* Suppl. 3rd. ed., Van Nostrand, New York 1952.
11. V. I. Kucheryavyi, G. N. Zinov'ev, L. K. Skotnikova, *J. Appl. Chem. (Leningrad)* **42** (1969) no. 2, 409 – 410 (*Zh. Prikl. Khim.* **42** (1969) no. 2, 446 – 447).
12. R. I. Spasskaya, *J. Appl. Chem. (Leningrad)* **46** (1973) no. 2, 407 – 409. (*Zh. Prikl. Khim.* **46** (1973) no. 2, 393 – 396).
13. G. Ostrogovich, R. Bacaloglu, *Rev. Roum. Chim.* **10** (1965) 1111 – 1123.
14. G. Ostrogovich, R. Bacaloglu, *Rev. Roum. Chim.* **10** (1965) 1125 – 1135.
15. W. R. Epperly, *Chem. Tech. (Heidelberg)* **21** (1991) no. 7, 429 – 431.
16. A. Lasalle, et al., *Ind. Eng. Chem. Res.* **31** (1992) no. 3, 777 – 780.
17. B. K. Banerjee, P. C. Srivastava, *Fert. Technol.* **16** (1979) nos. 3 – 4, 264 – 288.
18. *Kirk-Othmer,* 3rd. ed., **vol. 2,** pp. 440 – 469.
19. M. Frejacques, *Chim. Ind. (Paris)* **60** (1948) no. 1, 22 – 35.
20. S. Inoue, K. Kanai, E. Otsuka, *Bull. Chem. Soc. Jpn.* **45** (1972) no. 5, 1339 – 1345 (Part I); *Bull. Chem. Soc. Jpn.* **45** (1972) no. 6, 1616 – 1619 (Part II).
21. D. M. Gorlovskii, V. I. Kucheryavyi, *J. Appl. Chem. (Leningrad)* **54** (1981) no. 10, 1898 – 1901. (*Zh. Prikl. Khim.* **53** (1980) no. 11, 2548 – 2551).
22. W. Durisch, S. M. Lemkowitz, P. J. van den Berg, *Chimia* **34** (1980) no. 7, 314 – 322.
23. S. M. Lemkowitz, J. Zuidam, P. J. van den Berg, *J. Appl. Chem. Biotechnol.* **22** (1972) 727 – 737.
24. P. J. C. Kaasenbrood, H. A. G. Chermin, paper presented to The Fertilizer Society of London, 1st Dec., 1977.
25. Stamicarbon, EP 0 212 744, 1987 (J. H. Meessen, R. Sipkema).
26. Montedison, EP 0 132 194, 1985 (G. Pagani).
27. Toyo Engineering Corporation, GB 2 109 372, 1983 (S. Inoue et al.).
28. *Eur. Chem. News,* Aug. 9 (1982) 15 – 16.
29. T. Jojima, B. Kinno, H. Uchino, A. Fukui, *Chem. Eng. Prog.* **80** (1984) no. 4, 31 – 35.
30. E. Dooyeweerd, J. Meessen, *Nitrogen* **143** (1983) May – June, 32 – 38.
31. *Nitrogen* **194** (1991) Nov.– Dec., 22 – 28.
32. R. De Jonge, F. X. C. M. Barake, J. D. Logemann, *Chem. Age India* **26** (1975) no. 4, 249 – 260.
33. Montedison, EP 0 096 151, 1983 (G. Pagani, G. Faita, U. Grassini).
34. Urea Casale, EP 0 504 621, 1992 (V. Lagana).
35. Stamicarbon, US 2 727 069, 1955 (J. P. M. van Waes).
36. Stamicarbon, US 3 720 548, 1973 (F. X. C. M. Barake, C. G. M. Dijkhuis, J. D. Logemann).
37. C. Miola, H. Richter, *Werkst. Korros.* **43** (1992) 396 – 401.
38. Snamprogetti, US 4 899 813, 1990 (S. Menicatti, C. Miola, F. Granelli).
39. E. M. Elkanzi, *Res. Ind.* **36** (1991) no. 4, 254 – 259 (Eng.).
40. Unie van Kunstmestfabrieken, EP 0 155 735, 1985 (K. Jonckers).
41. Snamprogetti, GB 1 542 371, 1979 (U. Zardi, V. Lagana).
42. V. Lagana, G. Schmid, *Hydrocarbon Process.* **54** (1975) no. 7, 102 – 104 (Eng.).
43. U. Zardi, F. Ortu, *Hydrocarbon Process.* **49** (1970) no. 4, 115 – 116 (Eng.).
44. Snamprogetti, GB 1 506 129, 1978.
45. Snamprogetti, US 3 954 861, 1976 (M. Guadalupi, U. Zardi).
46. Snamprogetti, US 3 876 696, 1975 (M. Guadalupi, U. Zardi).
47. *Nitrogen* **185** (1990) May – June, 22 – 29.
48. Toyo Engineering Corp., US 4 301 299, 1981 (S. Inoue, H. Ono).
49. Toyo Engineering Corp., GB 2 109 372, 1983 (S. Inoue et al.).
50. Montedison, GB 1 581 505, 1980 (G. Pagani).
51. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Organic Compounds,* Suppl. 3rd ed., D. Van Nostrand Co., New York 1952, p. 393.
52. Snamprogetti, US 4 327 068, 1982 (V. Lagana, U. Zardi).

53. Unie van Kunstmestfabrieken, US 4 410 503, 1983 (P. J. M. van Nassau, A. M. Douwes).
54. The M. W. Kellogg Co., US 5 223 238, 1993 (T. A. Czup-pon).
55. Snamprogetti, US 4 235 816, 1980 (V. Lagana, F. Saviano).
56. Snamprogetti, GB 1 520 561, 1978 (G. Pagani).
57. Snamprogetti, GB 1 470 489, 1977 (A. Bonetti).
58. Snamprogetti, GB 1 560 174, 1980 (V. Lagana, F. Saviano).
59. V. Lagana, *Chem. Eng. (N.Y.)* **85** (1978) no. 1, 37 – 39.
60. Snamprogetti, GB 2 060 614, 1981 (V. Lagana, F. Saviano, V. Cavallanti).
61. S. K. Saxena, A. Ali, *Indian J. Environ. Prot.* **9** (1989) no. 11, 831 – 833 (Eng.).
62. R. M. Krishnan, *Fert. News* (1992) May, 53 – 57.
63. Ting-Chia Huang, Dong-Hwang Chen, *J. Chem. Technol. Biotechnol.* **55** (1992) no. 2, 191 – 199.
64. Unie van Kunstmestfabrieken, EP 0 053 410, 1982 (J. Zuidam, P. Bruls, K. Jonckers).
65. Snamprogetti, EP 0 417 829, 1991 (F. Granelli).
66. I. Mavrovic: "Pollution Control in Urea Plants," *Br. Sulfphur's 12th Int. Conf. Nitrogen 88,* Geneva, March 27 – 29, 1988.
67. Unie van Kunstmestfabrieken, EP 0 037 148, 1981 (M. H. Willems, J. W. Klok).
68. *Fert. Int.* **296** (1991) April, 32 – 37.
69. R. M. Reed, J. C. Reynolds, *Chem. Eng. Prog.* **69** (1973) no. 2, 62 – 66.
70. *Nitrogen* **95** (1975) May/June, 31 – 36.
71. R. D. Hauck, M. Koshino: *Fertilizer Technology & Use,* 2nd ed. Soil Science Society of America, Madison 1971, pp. 455 – 494.
72. N. K. Savant, E. T. Craswell, R. B. Diamond, *Fert. News* **28** (1983) no. 8, 27 – 35.
73. G. Wiedemann, *Justus Liebigs Ann. Chem.* **68** (1848) 325.
74. Kurzer, *Chem. Rev.* **56** (1956) 95 – 197.
75. *Zh. Prikl. Khim. (Leningrad)* **42** (1969) no. 13, 713.
76. IG Farbenind., DE 689 421, 1940.
77. G. Kränzlein, H. Keller, H. Schiff, *Justus Liebigs Ann. Chem.* **291** (1896) 374.
78. R. C. Haworth, F. G. Mann, *J. Chem. Soc.* 1943, 603.
79. American Cyanamid, US 2 760 961, 1956 (Mackey).
80. A. Fleischer, *Ber. Dtsch. Chem. Ges.* **9** (1876) 995.
81. A. v. Baeyer, *Justus Liebigs Ann. Chem.* **131** (1864) 252.
82. T. L. Davis, K. C. Blanchard, *J. Am. Chem. Soc.* **45** (1923) 1816. US 1 785 730, 1927 (T. L. Davis).
83. Knoll, DE 896 640, 1942.
84. T. L. Davies, K. C. Blanchard, *Org. Synth.* **3** (1923) 95.
85. Du Pont, US 2 145 242, 1937 (H. W. Arnold).
86. *Houben-Weyl,* **VIII, III,** 151.
87. A. Sonn, *Ber. Dtsch. Chem. Ges.* **47** (1914) 2440.
88. Schweizer, *J. Org. Chem.* **15** (1950) 471.
89. *Houben-Weyl,* **VIII, III,** 164.
90. McKay, Coleman, *J. Am. Chem. Soc.* **72** (1950) 3205.
91. H. Petersen: "Chemical Processing of Fibers and Fabrics, Functional Finishes," Part A, in M. Lewin, S. B. Sello (eds.): *Handbook of Fiber science and Technology,* vol. 2, Marcel Dekker, New York 1983, pp. 48 – 327.
92. *Org. Synth.* **29** (1949) 18.
93. *Houben-Weyl,* **VIII, III,** 153.
94. L. J. Smith, O. M. Emeron, *Org. Synth. Coll.* **III** (1955) 151.
95. G. M. Dyron, *Chem. Rev.* **4** (1927) 138.
96. W. Hentschel, *J. Prakt. Chem.* **27** (1883) no. 2, 499.
97. CIOS, XXVII/80, 15.
98. Bayer, DE 116 200, 1899. *Friedländer* **6,** 200. BASF, DE 46 737, 1888. *Friedländer* **2,** 450. Bayer, DE 131 513, 1901. *Friedländer* **6,** 968. Bayer, DE 216 666, 1908. *Friedländer* **9,** 372. Bayer, DE 493 811, 1926. Ges. Chem. Ind., *Friedländer* **16,** 1059.
99. A. Lüttringhaus, H. W. Dirksen, *Angew. Chem. Int. Ed. Engl.* **3** (1964) 260.
100. J. Berlinerblau, *J. Prakt. Chem.* **30** (1884) no. 3, 103.
101. Höchst, DE 205 662, 1906. *Friedländer* **9,** 395.
102. Du Pont, US 2 390 253, 1943 (C. O. Henke); *Chem. Abstr.* **40** (1946) 1876.
103. A. Wurtz, *Hebd. Seances Acad. Sci.* **27** (1848) 242.
104. A. Wurtz, *Justus Liebigs Ann. Chem.* **71** (1849) 329.
105. A. W. Hofmann, *Justus Liebigs Ann. Chem.* **74** (1850) 14.
106. A. Wurtz, *Justus Liebigs Ann. Chem.* **80** (1851) 347.
107. W. Michler, *Ber. Dtsch. Chem. Ges.* **9** (1876) 396, 711.
108. W. Michler, C. Escherich, *Ber. Dtsch. Chem. Ges.* **12** (1879) 1162.

109. *Houben-Weyl, VIII,* III, 160.
110. H. Eckenroth, *Ber. Dtsch. Chem. Ges.* **18** (1885) 516.
111. C. A. Bischoff, A. von Hedenström, *Ber. Dtsch. Chem. Ges.* **35** (1902) 3437.
112. T. Wilm, G. Wischin, *Justus Liebigs Ann. Chem.* **147** (1868) 162.
113. H. Petersen, *Angew. Chem.* **76** (1964) 909.
114. H. Petersen, *Test. Rundsch.* **16** (1961) 646.
115. H. Petersen, *Melliand Textilber.* **43** (1962) 380.
116. L. E. Smythe, *J. Phys. Chem.* **51** (1947) 369.
117. L. E. Smythe, *J. Am. Chem. Soc.* **73** (1951) 2735; **74** (1952) 2713; **75** (1953) 574.
118. L. Bettelheim, J. Cedwall, *Sven. Kem. Tidskr.* **60** (1948) 208.
119. G. A. Growe, C. C. Lynch, *J. Am. Chem. Soc.* **70** (1948) 3795; **72** (1950) 3622.
120. J. I. de Jong, J. de Jonge, *Reol. Trav. Chim. Pays-Bas* **71** (1952) 643, 661, 890; **72** (1953) 88.
121. J. Lemaitre, G. Smets, R. Hart, *Bull. Soc. Chim. Belg.* **63** (1954) 182.
122. J. Ugelstadt, J. de Jonge, *Reol. Trav. Chim. Pays-Bas* **76** (1957) 919.
123. M. Okano, Y. Ogata, *Bull. Soc. Chim. Belg.* **63** (1954)182.
124. R. Kvéton, F. Hanousek, *Chem. Listy* **48** (1954) 1205; **49** (1955) 63.
125. N. Landquist, *Acta Chem. Scand.* **9** (1955) 1127, 1459; **10** (1956) 244; **11** (1957) 776.
126. B. R. Glutz, H. Zollinger, *Helv. Chim. Acta* **52** (1969) 1976.
127. P. Eugster, H. Zollinger, *Helv. Chim. Acta* **52** (1969) 1985.
128. H. Petersen, *Textilveredlung* **2** (1967) 744.
129. H. Petersen, *Textilveredlung* **3** (1968) 160.
130. H. Petersen, *Chem. Ztg.* **95** (1971) 625.
131. A. Einhorn, A. Hamburger, *Ber. Dtsch. Chem. Ges.* **41** (1908) 24.
132. A. Einhorn, A. Hamburger, *Justus Liebigs Ann. Chem.* **361** (1908) 122.
133. H. Petersen, *Text. Res. J.* **40** (1970) 335.
134. BASF, GB 1 241 580, 1968 (H. Petersen).
135. H. Kadowaki, *Bull. Chem. Soc. Jpn.* **11** (1936) 248; *Chem. Zentralbl.* (1936 II)3535.
136. Sumitomo Chem. Co., DE 1 123 334, 1962 (T. Oshima).
137. A. M. Paquin, *Angew. Chem.* **60** (1948) 267.
138. W. J. Burke, *J. Am. Chem. Soc.* **69** (1967) 2136.
139. DuPont, US 2 304 624, 1942 (W. J. Burke).
140. DuPont, US 2 321 989, 1942 (W. J. Burke).
141. BASF, GB 1 077 344, 1965. BASF, CH 478 287, 1966. DuPont, US 2 436 355, 1946.
142. H. Petersen in: *Kunststoff-Jahrbuch,* 10th. ed., Wilhelm Pansegrau Verlag, Berlin 1968, p. 46.
143. H. Pauli, H. Sauter, *Ber. Dtsch. Chem. Ges.* **63** (1930) 2063.
144. BASF, DE 910 475, 1951 (H. Scheuermann, B. von Reibnitz, A. Wörner); *Chem. Zentralbl.* (1955) 702.
145. H. Petersen, *Textilveredlung* **3** (1968) 51.
146. H. Petersen, *Chem. Ztg.* **95** (1971) 692.
147. H. Petersen, *Text. Res. J.* **41** (1971) 239.
148. BASF, US 4 219 494, 1980 (H. Petersen).
149. E. R. Atkinson, A. H. Bump, *Ind. Eng. Chem.* **44** (1952) 333.
150. Societe Nobel Française, DE 962 119, 1952 (P. A. Talet); *Chem. Zentralbl.* (1955) 1625.
151. BASF, DE 1 049 572, 1959 (H. Willersinn, H. Scheuermann, A. Wörner).
152. BASF, DE 1 067 210, 1960 (H. Willersinn, H. Scheuermann, A. Wörner).
153. Mitsubishi Chem. Ind., US 3 322 528, 1961; DE 1 543 201, 1965 (M. Hamamoto, Y. Sakaki).
154. H. Petersen, *Synthesis* 1973, 243 – 292.
155. H. Petersen, *Angew. Chem.* **79** (1967) 1009.
156. BASF, US 2 731 472, 1956 (B. von Reibnitz).
157. H. Petersen, *Synthesis* 1973, 254.
158. BASF, DE 859 019, 1941 (J. Lintner, H. Scheuermann); *Chem. Zentralbl.* (1953) 4779.
159. Biltz, Behrens, *Ber. Dtsch. Chem. Ges.* **43** (1910) 1984.
160. US 2 638 434, 1953; US 2 649 381, 1953.
161. Chemische Werke Hüls, DE 1 020 024, 1957 (H. Steinbrink, I. Amende).
162. BASF, DE 1 230 805, 1962 (H. Petersen, H. Brandeis, R. Fikentscher).
163. G. Zigeuner, W. Rauter, *Monatsh. Chem.* **96** (1965) 1950.
164. BASF, DE 1 231 247, 1962 (H. Petersen, H. Brandeis, R. Fikentscher).
165. H. Bille, H. Petersen, *Text. Res. J.* **37** (1967) 264; *Textilveredlung* **2** (1967) 243.
166. A. M. Paquin, *Kunststoffe* **37** (1947) 165.
167. BASF, DE 1 223 843, 1962 (H. Petersen, H. Brandeis, R. Fikentscher).
168. G. Zigeuner, E. A. Gardziella, G. Bach, *Monatsh. Chem.* **92** (1961) 31.
169. G. Zigeuner, M. Wilhelmi, B. Bonath, *Monatsh. Chem.* **92** (1961) 42.
170. G. Zigeuner, W. Nischk, *Monatsh. Chem.* **92** (1961) 79.
171. BASF, DE 1 081 482, 1959; US 3 190 741, 1963 (J. Jung, H. Müller von Blumencron, C. Pfaff, H. Scheuermann).
172. BASF, EP 0 002 793, 1977; US 4 220 751, 1980 (H. Petersen, K. Fischer, H. Klug, W. Trimborn).
173. BASF, DE 2 757 220, 1977; EP 0 002 794, 1977; US 4 243 797, 1981 (H. Peterson et al.).

Urea Resins → **Amino Resins**
Vaccines → **Immunotherapy and Vaccines**
Vacuum Coating → **Thin Films**

Vanadium and Vanadium Compounds

GÜNTER BAUER, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

VOLKER GÜTHER, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

HANS HESS, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

ANDREAS OTTO, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

OSKAR ROIDL, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

HEINZ ROLLER, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

SIEGFRIED SATTELBERGER, GfE Gesellschaft für Elektrometallurgie mbH, Nürnberg, Federal Republic of Germany

1. History 1
2. Properties 2
3. Occurrence 3
4. Processing of the Raw Materials to Vanadium Compounds 6
4.1. Iron Ores and Titanomagnetites as Raw Materials 6
4.2. Processing of Other Raw Materials 7
5. Production of Vanadium, Ferrovanadium, and Other Alloys . 9
5.1. Reduction Behavior of Vanadium Oxides 9
5.2. Production of Vanadium Metal and its Alloys 11
5.3. Ferrovanadium 12
5.3.1. Production from Vanadium Oxides . . 13
5.3.2. Direct Production from Slags and Residues 13
5.4. Production of Other Vanadium Master Alloys 14
6. Uses 14
7. Vanadium Compounds 15
8. Analysis 18
9. Economic Aspects 18
10. Environmental Protection and Toxicology 19
11. References 20

1. History [1]

In 1801, MANUEL DEL RIO discovered vanadium in Mexican lead vanadate ore. In 1831, SEFTSTRÖM detected the element in converter slags from certain iron ores, and named it after the Norse goddess of beauty, Vanadis.

Vanadium metal was first produced in powder form by ROSCOE in 1867 – 1869 by reduction of vanadium dichloride with hydrogen. MARDEN and RICH obtained ductile vanadium by reducing vanadium pentoxide, V_2O_5, with calcium metal.

The first major application of vanadium was in 1903 in England, where a vanadium-alloyed steel was produced on an 18 t scale [2]. In 1905 HENRY FORD recognized the advantages of vanadium steel and promoted its use in automobile construction. Today vanadium is of major importance as an alloying component in steel and titanium alloys and as a catalyst for chemical reactions.

2. Properties

Vanadium [*7440-62-2*], V, atomic number 23, has two stable isotopes: ^{51}V (99.75 %) and ^{50}V (0.25 %). Unstable isotopes of relative atomic mass 48, 49, and 52 have half-lifes ranging from 4 min to 600 d. The electronic configuration is $1s^2\,2s^2p^6\,3s^2p^6d^3\,4s^2$. Vanadium is steel gray with a bluish tinge. It is ductile, and can be forged and rolled at ambient temperature.

Vanadium, together with niobium and tantalum, belongs to group 5 of the periodic table. It has a high melting point and good corrosion resistance at low temperature.

Physical Properties. The most important physical properties of vanadium are listed in the following [1], [3–6]:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_367

Relative atomic mass	50.9415
Crystal structure	body-centered cubic
Lattice constant a	0.30238 nm
Density	6.11 g/m^3
mp	1929 °C ± 6
Heat of fusion	21 500 J mol^{-1} ± 3000
Specific heat, J mol^{-1} K^{-1}	
at 298 K	24.35 ± 0.10 J mol^{-1} K^{-1}
for 298 – 900 K	$C_p = 24.134 + 6.196 \times 10^{-3}\,T - 7.305 \times 10^{-7}\,T^2 - 1.3892 \times 10^5\,T^{-2}$
for 900 – 2200 K	$C_p = 25.9 - 1.25 \times 10^{-4}\,T + 4.08 \times 10^{-6}\,T^2$
for liquid V	47.43 J mol^{-1} K^{-1}
Vapor pressure	
at 2190 K	3.73 Pa
at 2200 K	4.31 Pa
at 2300 K	12.53 Pa
at 2400 K	30.13 Pa
at 2500 K	87.86 Pa
at 2600 K	207.6 Pa
for liquid V, mbar	$\log p = -24.265 \times 10^3\,T^{-1} + 9.65$
Heat of vaporization	465.9 kJ mol^{-1}
Linear coefficient of expansion	
at 20 – 200 °C	7.88×10^{-6}
at 20 – 500 °C	9.6×10^{-6}
at 20 – 900 °C	10.4×10^{-6}
at 20 – 1100 °C	10.9×10^{-6}
Specific electrical resistivity at 20 °C	24.8×10^{-6} Ω cm
Temperature coefficient for 0 – 100 °C	0.0034 μΩ cm K^{-1}
Thermal conductivity	
at 100 °C	0.31 J cm^{-1} K^{-1} s^{-1}
at 500 °C	0.37 J cm^{-1} K^{-1} s^{-1}
Superconductivity, transition temperature	5.13 K
Capture cross section for thermal neutrons	4.8 barn

Mechanical Properties. The mechanical properties of vanadium are strongly dependent on purity and hence on the production method used. In particular, the elements O, H, N, and C increase the hardness and tensile strength and decrease the ductility (elongation) [7]. The most important mechanical properties are listed in Table 1 [3], [4], [7].

Table 1. Mechanical properties of vanadium metal

	Commercial purity	High purity
Tensile strength σ_B, N/mm	245 – 450	180
Extension, %	10 – 15	40
Vickers hardness HV 10, N/mm	80 – 150	60 – 70
Modulus of elasticity, N/mm	137 000 – 147 000	
Poisson's ratio	0.35	

Chemical Properties. Vanadium is stable in air below 250 °C. On prolonged storage the surface becomes bluish-gray to brownish-black, and significant oxidation takes place in air above 300 °C. Vanadium absorbs hydrogen in interstitial lattice sites at elevated temperatures (up to 500 °C). The metal becomes brittle and can easily be powdered. This hydrogen is liberated on heating to 600 – 700 °C in vacuum. At low temperature, a hydride phase exists in the V – H system [7]. Vanadium reacts with nitrogen at > 800 °C to form vanadium nitrides. It has a high affinity for carbon, forming carbides at 800 – 1000 °C. Data on the solubility of O, N, H, and C in vanadium and the reactions that occur with these elements can be found in [7], [8].

In its compounds, vanadium exhibits the oxidation state II, III, IV, or V. It is relatively stable towards dilute sulfuric, hydrochloric, and phosphoric acids, but is dissolved by nitric and hydrofluoric acids. Its corrosion resistance towards tap water is good, and towards seawater moderate to good, but pitting does not occur. Vanadium is resistant to 10 % sodium hydroxide solution, but is attacked by a hot solution of potassium hydroxide [1], [3], [4], [9], [10]. Vanadium and some vanadium alloys have good corrosion resistance towards molten low-melting metals and alloys, especially alkali metals, which are used in nuclear reactors as coolants and heat-exchange media [11].

3. Occurrence

Vanadium is present in the earth's crust at a mean concentration of 150 g/t, and is therefore one of the more common metals. It is more abundant than copper and nickel, and of similar abundance to zinc. Vanadium forms several minerals, of which the most important are listed in Table 2 [12], [13].

In the early 1900s, vanadium was obtained almost exclusively from Peruvian patronite. As these deposits became exhausted, descloizite ores were mined in southern Africa, Namibia, and Zambia.

After World War II, the continental sedimentary uranium- and vanadium-containing carnotite deposits of the Colorado Plateau and the marine sedimentary vanadium-bearing phosphate deposits of Idaho became important. There

Table 2. Principal vanadium minerals [12], [13]

Mineral and chemical formula		Vanadium content, %	V_2O_5 content, %	Occurrence
Roscoelite	$[KV_2(OH)_2/AlSi_3O_{10}]$	11.2 – 14.0	20 – 25	in uranium – vanadium ores; e.g., Colorado Plateau, United States
Montroseite	(V, Fe)OOH	45.4	81.0	
Carnotite	$K_2[(UO_2)_2/V_2O_8] \cdot 3\,H_2O$	10.3	18.3	
Tyuyamunite	$Ca[(UO_2)_2/V_2O_8] \cdot 5-8\,H_2O$	11.1	19.8	
Francevillite	$(Ba, Pb)[(UO_2)_2/V_2O_8] \cdot 5\,H_2O$	9.9	17.7	
Corvusite	$V_2^{4+} \cdot V_{12}^{5+}O_{34} \cdot n\,H_2O$	40.8	72.8	
Vanadinite	$Pb_2[Cl/(VO_4)_3]$	10.2	18.2	in Pb, Zn, Cu vanadate ores; e.g., Otavi Mountains, Namibia
Descloizite	$Pb(Zn, Cu)[OH/VO_4]$	12.7	22.7	
Mottramite	$Pb(Cu, Zn)[OH/VO_4]$	10.5	18.8	
Patronite	VS_4 or V_2O_5	16.8	ca. 30	in asphaltites; e.g., Mina Ragra, Peru
Magnetite∗	$Fe_2^{2+} \cdot Fe^{3+}O^{4-}$	$< 0.5 - 1.5$	$< 1.2 - 2.7$	in titanomagnetite ores; e.g., Bushveld, Republic of South Africa
		max. 2.8	max. 5.0	

∗ Several varieties exist in which there is partial replacement of Fe^{3+} by V, Ti, Al, and Cr, and of Fe^{2+} by Ti, Mg, Mn, and other elements. The vanadium- and titanium-containing magnetites are generally referred to as titanomagnetites.

are also uranium – vanadium reserves in Yeellirie, Western Australia, and in the Ferghana Basin the Kirghiz, Tadzhik, and Uzbek Republics [13].

Vanadium production greatly increased from the mid-1950s with the mining and processing of titanomagnetites. This type of raw material can be used directly for vanadium extraction, or it can be employed to obtain vanadium-containing pig iron from which an oxidation slag highly enriched in vanadium is produced. This development began in Finland, followed by South Africa and smaller producers in Norway and Chile. Titanomagnetites have also been mined in the former Soviet Union and China in large quantities since the early 1960s and 1970s, respectively. The vanadium-bearing titanomagnetite deposits are of magmatogenic origin, and occur in many parts of the world. The most important reserves mined today include the Bushveld deposit in South Africa, Katschkanor in the Urals of Russia, Lanshan and Chienshan in the Sichwan province of China, and a deposit in New Zealand. The titanomagnetites of the Urals and in the Sichwan province are palaeozoic complexes.

The mines at Otanmäki and Mustavaara in Finland and Rødsand in Norway were shut down during the 1980s.

Production of lead – zinc vanadates in Namibia had been discontinued by the end of the 1970s.

Since the 1980s, a new type of secondary raw materials has gained importance, i.e., residues of mineral oil processing. Most crude oils contain vanadium in amounts ranging from ca. 10 ppm (Middle East) to 1400 ppm (Central America). In petrochemical refining, vanadium is retained in boiler residues and fly ashes from incineration, with vanadium contents between a few percent and 40 %. The total amounts greatly exceed current world consumption.

Noteworthy is an aqueous emulsion of a highly viscous crude from the Orinoco basin in Venezuela (Orimulsion) whose combustion yields fly ashes containing 10 % vanadium.

With environmental legislation becoming stricter, emission and deposition of these residues will be drastically reduced so that they will have increasingly to be processed.

Table 3. Vanadium production as a byproduct

Raw material	Coupled product	Vanadium byproduct
Titanomagnetites	iron/steel	vanadium slag
Mineral oils	energy/petrochemicals	fly ashes boiler residues
Uranium – vanadium ores	uranium	petrochemical residue
Bauxite	alumina	vanadium salt
Phosphates	phosphorus	vanadium-containing Fe – P salamander
Lead vanadates	lead, zinc	vanadium slag

Other, less important vanadium sources are: vanadium-containing waste salts from bauxite

production, and spent catalysts from the chemical and petrochemical industries. The extraction of vanadium from currently available raw material sources (except direct processing of titanomagnetite) is always coupled with the production of other metals or energy. The most important sources are shown in Table 3.

Resources and Production. Known and estimated mineral reserves are listed in Table 4.

In this table the vanadium containing carbonitic shells (very low concentration) and the petroleum ashes/residues are not included, although these oils contain tremendous amounts of vanadium.

Assessment of world reserves by studying the geology of deposits suggests that it will be possible to discover further workable deposits in the region of the widely distributed titanomagnetites in the old shields. The Indian shield in the Singhbum (Bihar) and Magurbhanj (Orissa) regions, the Australian shield in Western Australia, the Canadian shield in Quebec and Ontario, and the Finno-Scandinavian shield all show great promise. However, little is so far known about the Brazilian shield. The enormous vanadium reserves in the titanomagnetites cannot yet be even approximately quantified.

The world production of vanadium by country is listed in Table 5 [14].

Comparison of Tables 4 and 5 shows that the largest producers, the Republic of South Africa and Russia, have also the largest mineral reserves. The next two producing countries China and the United States have comparatively small reserves of such origin.

Details of the Most Important Deposits. The Bushveld deposit in South Africa is an oval, bowl-shaped complex in a magmatic layer extending over an area of 65 000 km^2. The ores were deposited in the following order: chromite, platinum ores, vanadium-containing titanomagnetites, and tin ores. In all, 21 magnetite layers have been found in the main and upper zones of this deposit. The titanomagnetites have the following chemical composition (%):

Fe	55.8 – 57.5	Cr_2O_3	0.13 – 0.45
V_2O_5	1.4 – 1.6	Al_2O_3	2.5 – 3.5
TiO_2	12.2 – 13.9	SiO_2	0.9 – 1.5

The V_2O_5 content can be as high as 2.4 % in exceptional cases.

Production by Highveld Steel & Vanadium is based on ore containing 1.6 % V_2O_5 from the Mapochs mine, situated 90 km NNE of Middleburg. The reserves contain 200×10^6 t ore. Mining is by the open pit method, and the maximum ore production capacity is 8000 t/d [13].

Other South African titanomagnetite deposits include the Kennedy's Vale mine, with a production capacity of 6000 t/d (though not currently producing) and two mines near Brits, each with a production capacity of 5000 – 6000 t/d.

There are a number of lead vanadate deposits in the Otavi Mountains in Namibia. Mount Aukas, near Grootfontein, has deposits of 2×10^6 t ore containing 15 % Zn, 4 % Pb, and 0.5 % V_2O_5. Deep mining has been discontinued since 1978. The mined ore, which was processed by flotation, yielded the following products: lead sulfide concentrate, zinc sulfide concentrate, willemite concentrate, cerussite concentrate, and a vanadium concentrate containing 17 % V_2O_5, 43 % Pb, and 17 % Zn.

The titanomagnetite deposits at Otanmäki and Mustavaara in Finland had reserves of 15×10^6 t ore containing 35 – 40 % Fe, 13 % TiO_2, and 0.45 V_2O_5, and 40×10^6 t ore containing 17 % Fe, 3.1 % TiO_2, and 0.36 %V_2O_5 [14]. The production capacity of the Otanmäki deep mine was ca. 1×10^6 t/a ore. Magnetite containing 69 % Fe, 2.5 % TiO_2, and 1.07 % V_2O_5 together with ilmenite and pyrite were obtained by magnetic separation and flotation. Both mines were closed in the 1980s.

The Colorado Plateau in the United States contains the main mining area of the Uravan mineral belt, > 1100 km long, in SW Colorado and SE Utah. This belt contains irregular lenticular ore bodies, normally < 3000 t, with 0.2 – 0.3 % U_3O_8 and 0.85 – 1.4 % V_2O_5. The oxidation region contains the minerals carnotite, tyuyamunite, montroseite, and pascoite, and the unoxidized region contains the minerals coffinite, pitchblende, and corvusite. These ores are extracted in small deep mine workings which produce ca. 0.4×10^6 t/a ore.

Table 4. Estimate of the world reserves and reserve base of vanadium in thousand tonnes of vanadium content (percentage given in parentheses)

Geographic area	Reserves		Estimated reserves
	1985	1990	1985
North America	185 (4.0)	135 (3.2)	2500 (13.7)
South America	25 (0.5)		130 (0.7)
Former Soviet Union	2900 (60.4)	2631 (61.7)	4500 (24.6)
South Africa	950 (19.7)	862 (20.2)	8600 (47.0)
China	670 (14.0)	500 (14.2)	1800 (9.8)
Pacific	35 (3.1)	30 (0.7)	570 (3.1)
Total	4800 (100)	4267 (100)	18 300 (100)

Table 5. World production of vanadium-containing commodities by country, given in tonnes of vanadium content [14]

Countries	1985	1986	1987	1988	1989	1990	1991	1992	1993
China [a]	5 500	5 500	5 500	5 500	4 500	4 500	4 500	4 500	5 000
Finland	2 590								
South Africa [b]	17 000	18 600	17 300	19 900	16 500	17 000	15 500	14 300	16 000
Former Soviet Union	11 500	11 500	11 600	11 600	9 600	9 600	8 000	6 000	7 900
United States (total)	4 000	3 600	3 800	3 500			5 600	5 300	5 500
from catalysts and ashes	2 970	2 560	2 760	2 480					
Japan [c]	925	1 020	1 020	1 020			600	800	
New Zealand							2 200	1 700	
Total	45 700	44 000	42 100	45 300	31 600	31 600	35 800	33 200	34 500

[a] Estimated.
[b] South Africa; roughly 40 % is pentoxide and vanadate products and 60 % is vanadiferrous slag products.
[c] Spent catalyst and petroleum ashes.

4. Processing of the Raw Materials to Vanadium Compounds

The raw materials used today include the titanomagnetite ores and their concentrates, which are sometimes processed directly, vanadium slags derived from the ores, oil combustion residues, residues from the hydrodemetallization (HDM) process, and spent catalysts (secondary raw materials). A summary of the processing routes is given in Figure 1.

4.1. Iron Ores and Titanomagnetites as Raw Materials

Production of Vanadium Slags. The titanomagnetite ore in lump form, containing ca. 1.51 – 1.7 % V_2O_5 (Mapoch mine in South Africa), is first prereduced by coal at ca. 1000 °C in directly heated rotary kilns. A further reduction is then performed in an electric arc furnace to obtain a pig iron which contains ca. 1.4 % V_2O_5. The slag that is also formed contains ca. 30 % TiO_2, and this is deposited on slag heaps. In a heat resistant shaking ladle the molten pig iron is oxidized with oxygen lances, causing the vanadium to be transferred to the slag. The vanadium slag contains vanadium in the form of a water-insoluble trivalent iron spinel, $FeO \cdot V_2O_3$, at a concentration corresponding to ca. 25 % V_2O_5. This slag is the world's principal raw material for vanadium production [16].

The largest slag and vanadium producer in the western world is South Africa (see Table 5), most of whose production is by the Highveld Steel and Vanadium Corp. [13]. Second largest producer is the Nizhny Tagil metallurgical plant in Russia; the slag contains ca. 12 – 18 % V_2O_5. Another producer of vanadium slag is the Pannang steelworks in Panzihua, China.

Vanadium slags were produced experimentally by W. L. Goodwin and W. P. Firth in 1919, but the basic industrial process was developed by R. von Setz, and was first used by the Spigerverk plant in Oslo in the early 1930s. The process was then mainly used in conjunction with the Thomas process. The pig iron from the blast furnace was not directly converted to Thomas steel with addition of lime in the usual way. Instead, the vanadium was oxi-

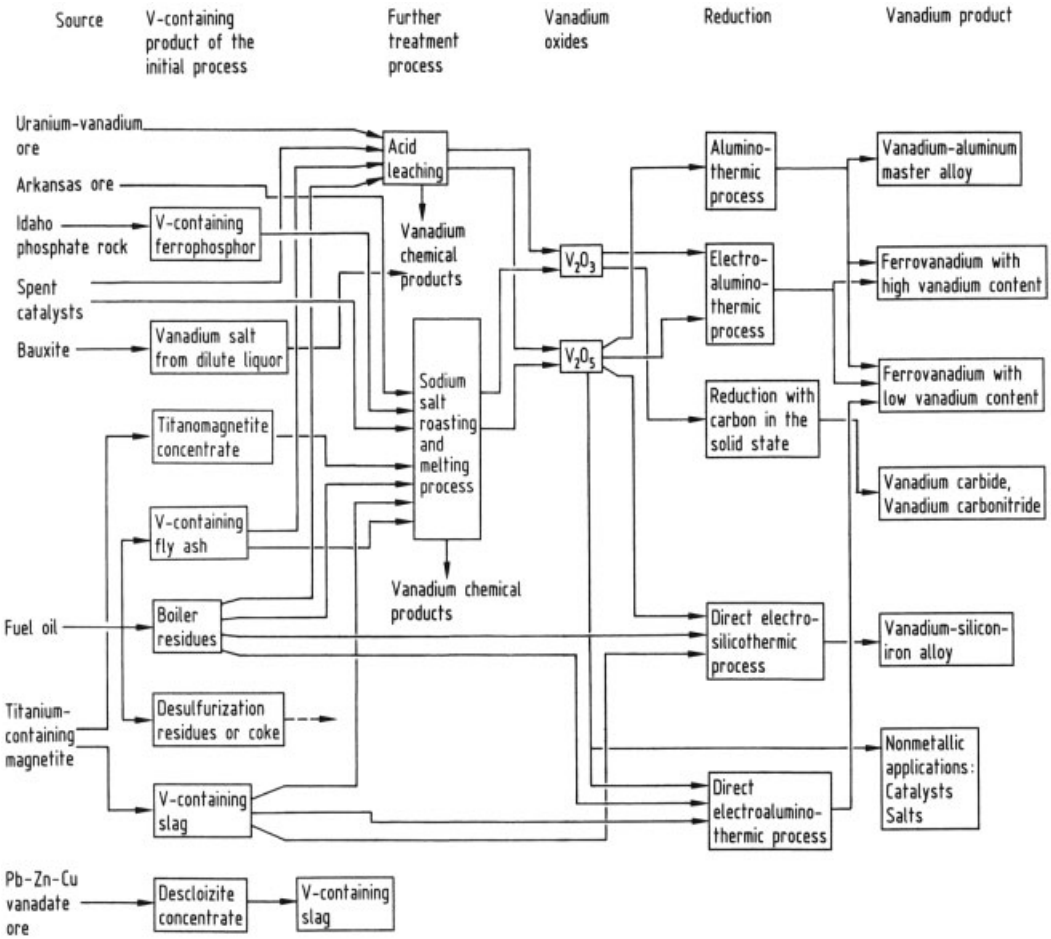

Figure 1. Treatment of vanadium raw materials [15]

dized completely or partially before lime addition. As vanadium has a high affinity for oxygen, it was oxidized before the phosphorus and transferred to the slag. Slags containing several percent of vanadium were obtained, depending on the vanadium content of the pig iron, whereas normal Thomas slags contain an average of only 0.5 % V.

Oxidation of vanadium-containing pig iron to give vanadium-containing slag was carried out by the Röchling Iron and Steel Works in Völklingen from 1937 using a converter with a very acid lining or a continuously operating oxidizing drum. For details of these abandoned processes see [17].

Processing of Vanadium Slags. The main process used today to produce vanadium oxide from vanadium slags is alkaline roasting. The same process, with minor differences, is also used for processing titanomagnetite ores and vanadium-containing residues.

A vanadium slag has the following approximate composition: 14 % V (=25 % V_2O_5), 1.5 % Ca, 2.5 % Mg, 2.0 % Al, 0.01 % P, 9 % metallic Fe, 32 % total Fe, 7 % Si, 3.5 % Mn, and 3.5 % Ti. The presence of the elements Ca, Mg, and Al, which form water-insoluble vanadates during alkaline roasting, leads to a reduction in yield. Silicon can cause filtration difficulties during leaching of the calcine. Phosphorus, unless removed in a separate treatment, reports practically quan-

titatively to the vanadium oxide, and lowers the vanadium yield in the precipitation process.

The process is shown schematically in Figure 2. The vanadium slag is first ground to $< 100\,\mu m$ (1 – 11 in Fig. 2), and the iron granules contained therein are removed. Alkali metal salts are then added, and the material is roasted with oxidation at 700 – 850 °C in multiple-hearth furnaces or rotary kilns (12 – 20) to form water-soluble sodium vanadate. During roasting, care must be taken to prevent agglomeration of the material by sintering and to ensure rapid cooling after the material leaves the furnace. The roasted product is leached with water (21 in Fig. 2), and ammonium polyvanadate or sparingly soluble ammonium metavanadate is precipitated in crystalline form from the alkaline sodium vanadate solution by adding sulfuric or hydrochloric acid and ammonium salts at elevated temperature (22 – 28). These compounds are converted to high-purity, alkali-free vanadium pentoxide by roasting. The usual commercial "flake" form of vanadium pentoxide is obtained by solidifying the melt on cooled rotating tables (29 – 31 in Fig. 2).

The aluminothermic production of ferrovanadium is shown in Figure 2 (32 – 34), as well as wastewater purification (35 – 44), including the reduction of Cr(VI) to Cr(III) (35) and the crystallization of sodium sulfate (37).

In Russia, a combination of alkaline roasting and sulfuric acid leaching is used. A disadvantage of this process is the manganese content of up to 2 % in the V_2O_5, caused by acid leaching.

Direct Production of Vanadium Oxide from Titanomagnetite Ores and Clays. If the ores contain > 1 % vanadium, they can be directly converted to vanadium pentoxide. The gangue material is first removed from the ore by flotation, and the ore is then mixed with sodium carbonate, pelletized, and roasted in shaft furnaces or rotary kilns at 1000 °C. The sodium metavanadate formed is leached out with water, and vanadium pentoxide is obtained from this by the same method as that used in the roasting process for vanadium slags. The leach residue is used to produce pig iron if its ilmenite content is sufficiently low [15].

4.2. Processing of Other Raw Materials

Hydrometallurgical methods or a combination of pyrometallurgical and hydrometallurgical processes are used to produce vanadium oxides and salts from other raw materials. In the combined processes, thermal treatment is followed by alkaline or, more rarely, acid processing.

Acid Processes. Vanadium-containing oil residues and spent catalysts are digested with sulfuric acid, often under reducing conditions with addition of SO_2. After removal of the impurities, the solution is oxidized, and polyvanadate and calcium or iron vanadate are precipitated by partial neutralization of the acid [13].

Acid digestion was formerly an important method of treating carnotite, a uranium – vanadium ore. After the uranium had been removed from the solution in sulfuric acid by solvent extraction, the vanadium was partially reduced with iron powder and solvent extracted. After re-extraction, the vanadium was in the form of an acid solution of vanadyl sulfate. This was oxidized with sodium chlorate at elevated temperature, and precipitated as red cake, a mixture of sodium and ammonium polyvanadate, by adding ammonia or ammonium salts. This precipitate was roasted to convert it into alkali-containing V_2O_5 [13], [18].

Historically, Pb – Zn – V concentrates from Namibia were subjected to acid treatment process. However, the mining of these ores became uneconomic and was discontinued in 1978.

The ore concentrate produced by flotation was ground and leached with dilute sulfuric acid (10 – 15 %). This led to the formation of insoluble lead sulfate. The solubility of the vanadium was considerably increased by simultaneously passing SO_2 through the liquor. Phosphorus was removed by adding sodium zirconate to precipitate sparingly soluble zirconium phosphate. The filtrate, which contained mainly $ZnSO_4$ and $VOSO_4$, was treated with H_2S to precipitate As and Cu. The pH of the clear solution was increased to 1 – 2 by addition of sodium carbonate, and tetravalent vanadium was oxidized to the pentavalent state by heating with sodium chlorate. More sodium carbonate was added to the iron-containing solution to increase its pH to 4, precipitating a polyvanadate containing iron and sodium. This precipitate was filtered and treated

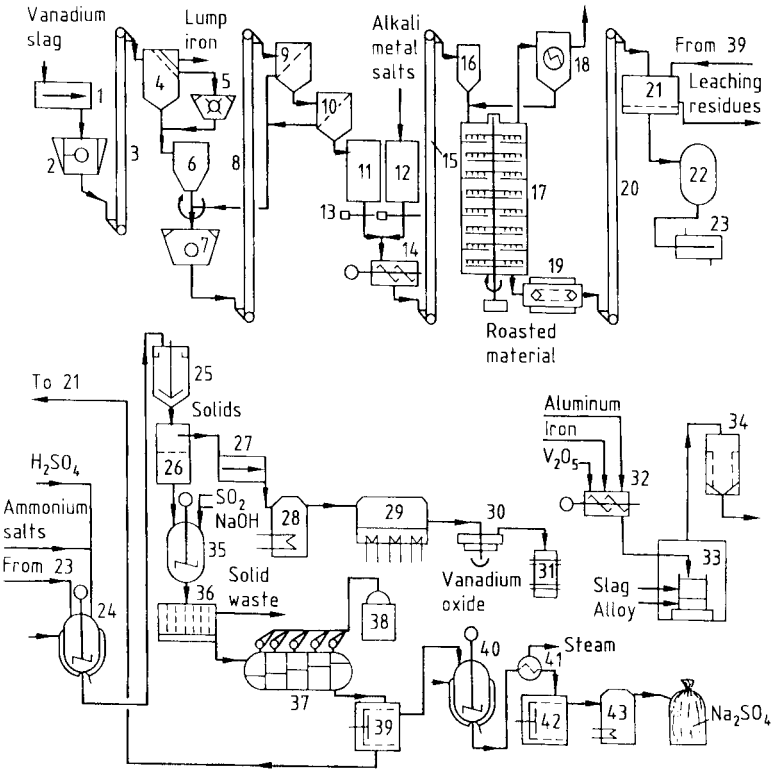

Figure 2. Production of vanadium from vanadium slags: process of the Gesellschaft für Elektrometallurgie
1) Push feeder; 2) Jaw crasher; 3) Bucket conveyor; 4) Screen; 5) Impact mill; 6) Container with discharge device; 7) Mill; 8) Bucket conveyor; 9) Screen; 10) Air classifier; 11) Container for ground material; 12) Container for alkali metal salt; 13) Metering weighing machine; 14) Mixer; 15) Bucket conveyor; 16) Container with discharge device; 17) Multideck furnace; 18) Electrostatic precipitator; 19) Chain conveyor with cooling; 20) Bucket conveyor; 21) Leaching with filtration; 22) Storage container; 23) Double-tube heat exchanger; 24) Precipitation vessel; 25) Thickener; 26) Plate filter; 27) Continuous feeder; 28) Dryer; 29) Melting furnace; 30) Cooled rotary table; 31) Packing; 32) Mixer; 33) Aluminothermic reaction; 34) Bag filter; 35) Precipitation vessel; 36) Filter press; 37) Cooling with crystallization; 38) Steam generation; 39) Pusher centrifuge; 40) Melting vessel; 41) Heat exchanger; 42) Pusher centrifuge; 43) Dryer

with a solution of an ammonium salt to replace the sodium ions by ammonium ions. It was then dried and converted to V_2O_5 by fusion.

Zinc carbonate was precipitated from the filtrate after the polyvanadate precipitation.

Combination of Pyrometallurgical and Alkaline Hydrometallurgical Methods. This combined process, which is preferred in practice, is used to treat vanadium-containing residues and spent catalysts. (It was formerly also used for V-containing clays, phosphates, and Pb – Zn vanadates.) The vanadium, which is mainly trivalent, is oxidized at high temperature and converted to alkali metal vanadate, which can be leached with water.

The best known process is alkaline roasting, in one version of which the ore is pelletized with sodium carbonate or compacted. After leaching, the solutions are acidified and ammonium salts are added, precipitating the vanadium as polyvanadate, which is then converted to vanadium pentoxide by calcination.

In the treatment of Pb – Zn vanadates, the zinc and lead were first reduced with carbon in the presence of added sodium carbonate in short drum furnaces. The zinc was distilled off and the lead tapped off as molten metal. The vanadium was then present in the alkaline slag in water-soluble form.

Production of Vanadium Pentoxide from Oil Residues and Boiler Ash. Appreciable amounts of vanadium are present in mineral oil, especially Venezuelan and Russian oils. The vanadium accumulates in the soot and ash of oil-fired boilers and in oil distillation residues and is sold in this form by power stations and refineries. These secondary raw materials, which sometimes contain $> 50\,\%$ V_2O_5, can be blended in the pyrometallurgical/hydrometallurgical process [13] or treated separately.

The extraction of vanadium from mineral oil combustion residues is gaining in importance, since the asphalt-like oil deposits of the Orinoco basin in Venezuela can now be processed into a fuel resembling mineral oil by emulsification with water with addition of magnesium nitrate. The fuel is used for energy generation in power stations. The amount of vanadium in the resulting boiler ashes is expected to reach the same order of magnitude as is currently generated in vanadium slags.

Production of V_2O_5 during the Processing of Bauxites. Some deposits of bauxite contain high levels of vanadium, and during alumina production this ends up in the salt residue, from which V_2O_5 can be obtained. This process is more expensive than vanadium production from slags and causes severe wastewater treatment problems due to the arsenic originating from the raw material [19]. Small amounts of vanadium are obtained by this method in India and Hungary [13].

5. Production of Vanadium, Ferrovanadium, and Other Alloys

5.1. Reduction Behavior of Vanadium Oxides

The most important oxide obtained on treating raw materials is vanadium pentoxide, V_2O_5 (see Chap. 6).

The oxygen partial pressure in equilibrium with the vanadium oxides in accordance with the equations

$$2\,V_2O_5 \rightleftharpoons 2\,V_2O_4 + O_2$$

$$2\,V_2O_4 \rightleftharpoons 2\,V_2O_3 + O_2$$

$$2\,V_2O_3 \rightleftharpoons 4\,VO + O_2$$

$$2\,VO \rightleftharpoons 2\,V + O_2$$

can be calculated from thermodynamic data [20]. The reactions are strongly endothermic, and have positive free enthalpies in the temperature range 298.15 – 2000 K (Fig. 3). V_2O_5 is the only oxide which dissociates to a large extent, the oxygen pressure reaching 0.08 MPa at 2000 K. In the presence of reducing gases such as CO and H_2, the higher oxides are reduced to V_2O_3:

$$V_2O_4 + CO \rightleftharpoons V_2O_3 + CO_2$$

The reduction of V_2O_3 by CO or H_2 in accordance with the equation

$$V_2O_3 + CO \rightleftharpoons 2\,VO + CO_2$$

does not go to completion [20], but reduction by carbon is almost complete:

$$V_2O_3 + C \longrightarrow 2\,VO + CO$$

The reaction of VO with carbon initially gives vanadium carbide:

$$VO + 1.88\,C \rightleftharpoons VC_{0.88} + CO$$

In the further reaction of oxide with carbide,

$$VO + 1.14\,VC_{0.88} \rightleftharpoons 2.14\,V + CO$$

it can be calculated from thermodynamic data [20] that the CO pressure will reach ca. 600 Pa at 2000 K. However, VO and C are soluble in vanadium at high temperature. This does not favor carbothermic reduction to pure vanadium metal even under high vacuum, and it is difficult to obtain a metal with low C and O contents [1].

For the metallothermic reduction of vanadium oxide silicon, aluminum, calcium, and magnesium could be considered as possible reagents, of which calcium and magnesium have an especially strong reduction potential. However, reduction with aluminum is incomplete, the vanadium produced being either oxygen-containing and low in aluminum, or aluminum-containing and low in oxygen [21], which, in addition, leads to loss of vanadium in the slag. In the production of VAl, a low-oxygen, aluminum-containing alloy is refined in a second stage using a vacuum induction furnace and is adjusted

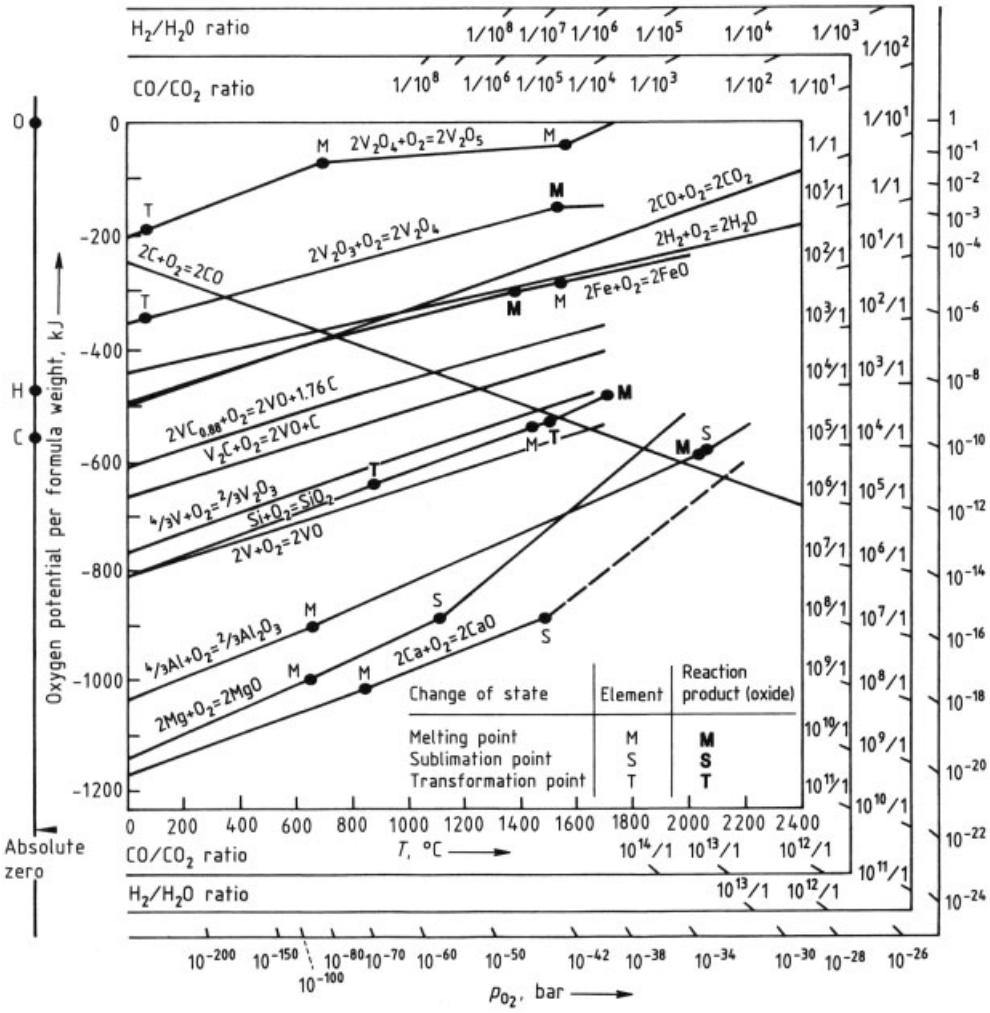

Figure 3. Oxygen potentials of oxides of vanadium and other elements as a function of temperature

to the derived final composition. The refinement of aluminum-free vanadium metal is performed in an electron-beam furnace. The even less complete reduction with silicon is only used in the production of ferrovanadium, as the reaction proceeds better in the presence of iron.

Reduction of the oxides with calcium and magnesium goes almost to completion. As these metals are volatile, it is advantageous to carry out the reaction in closed vessels.

5.2. Production of Vanadium Metal and its Alloys

Reduction of Vanadium Oxide with Carbon. Carbothermic reduction of vanadium oxides is possible (Section 4.1), but the vanadium produced is of high oxygen and carbon content; thus, this method is rarely used in practice. Also, the formation of a low-melting ternary eutectic (96.6 % V, 2.8 % C, 0.6 % O; *mp* 1500 °C) prevents the reaction going to completion [22].

Avoidance of these problems by reducing vanadium oxide with vanadium carbide (carbidothermic production) has been proposed [23], [24]. High-purity vanadium ($\geq$ 99.8 % V)

can be obtained by multiple sintering (two indirect sintering stages, one direct stage, and post-sintering).

Reduction of V_2O_5 with Calcium. A pressure vessel lined with pure magnesium oxide is used in this process. A mixture of high-purity V_2O_5, double-distilled calcium chippings, and some high-purity sulfur as a booster is caused to react by heating [25]. This technique leads to a compact metal regulus (8 – 12 kg, 99.5 – 99.8 % V) which is well separated from the slag. With a calcium excess of up to 70 %, a vanadium yield of 85 – 90 % is obtained. The Vickers hardness (HV 10) of this vanadium metal is 80 – 180, depending on the impurities present (mainly O, N, C, and Si) [1].

Reduction of V_2O_5 or V_2O_3 with Aluminum [26], [27]. The reduction

$$3\,V_2O_5 + 10\,Al \longrightarrow 6\,V + 5\,Al_2O_3$$

is self-sustaining (→ Aluminothermic Processes, Chap. 2.1.), i.e., after initiation it proceeds without additional heating. The metal and slag are melted and separate well from each other. An excess of heat is produced during reduction. For process control, only a fraction of the material is ignited primarily and further material is then charged to the melt. To further decrease the exothermicity, a mixture of V_2O_5 and V_2O_3 is used. In the reduction of V_2O_3 with aluminum, boosters such as $KClO_3$ must be added to increase the reaction heat.

In the process developed by Teledyne Wah Chang Albany [28], a vanadium – aluminum alloy containing 13 – 15 % Al is obtained by the aluminothermic reaction of high-purity V_2O_5 with aluminum powder in a closed vessel. The steel reaction vessel is lined with aluminum oxide at the bottom and in the reaction zone. After evacuation, it is purged several times with helium, and the charge is then electrically ignited with a heated resistance wire. The metal blocks obtained are refined in an electron-beam furnace by remelting several times. The high aluminum content leads to the removal of most of the oxygen by volatilization of aluminum suboxides, but nitrogen and silicon cannot be removed by this method. A purity of 99.93 % V can be achieved.

In a similar process [29], the metal produced in a closed reactor by the aluminothermic method is treated at 1700 °C and 10^{-4} Pa in a resistance-heated vacuum furnace, causing most of the aluminum to volatilize. Further purification by melting in an electron-beam furnace gives 99.93 % pure metal.

Refining. Vanadium – aluminum alloys produced aluminothermically can be refined by molten salt electrolysis (electrorefining) [26]. A metal containing ca. 4.8 % Al, 0.09 % O, and 0.08 % N can be refined by this method to give 99.6 % pure vanadium.

The most effective purification of technically pure vanadium is achieved by the van Arkel process. Vanadium and iodine are heated together in a closed, evacuated tube, forming VI_2 and VI_3. These are then sublimed and decomposed in a second reaction chamber on a tungsten or vanadium wire heated to 800 °C, yielding a 99.95 % pure vanadium. The equipment required is complex and expensive, and the industrial potential is limited.

Production of Semifinished Products (Sheet, Rods, Tubes, Wire). The cast blocks of vanadium metal and alloys have a coarse structure which must first be converted into a fine-grained structure by extruding or forging at 900 – 1200 °C and then annealing. To prevent reaction of the vanadium with atmospheric oxygen and nitrogen, the blocks are encased in an iron or stainless steel cladding by vacuum welding. Owing to the good ductility of vanadium, subsequent metal working by rolling, forging, or drawing can be carried out at 100 – 500 °C. Pure vanadium can normally be formed at room temperature into thin sheets, wires, and tubes.

Vanadium and vanadium alloys can be welded using tungsten electrodes protected by an inert gas (TIG welding). Vanadium can be machined like niobium and tantalum. Excessive heating due to machining at high speeds must be avoided [30], [31].

5.3. Ferrovanadium

According to DIN 17 563 [32], ferrovanadium is "a master alloy with a vanadium content of at least 50 wt % produced by reduction of the corresponding raw materials or their concentrates."

In the United States alloys with lower vanadium contents are also commercially available. The iron – vanadium and vanadium – aluminum master alloys and the commercial grades of vanadium alloys are listed in Table 6.

Table 6. Commercial grades of vanadium master alloys [33], [34]

Ferrovanadium	% V	% C	% Si	% Al
FeV 40	35 – 48	0.5	2.0	0.5
FeV 50	48 – 60	0.5	2.0	0.3
FeV 60	50 – 65	<0.15	<1.5	<2.0
FeV 80	78 – 82	0.15	1.5	<1.5

The Fe – V phase diagram (Fig. 4) shows almost complete mutual solubility of the elements. At compositions approximating to FeV below 1200 °C, a σ-phase exists which forms tetragonal crystals.

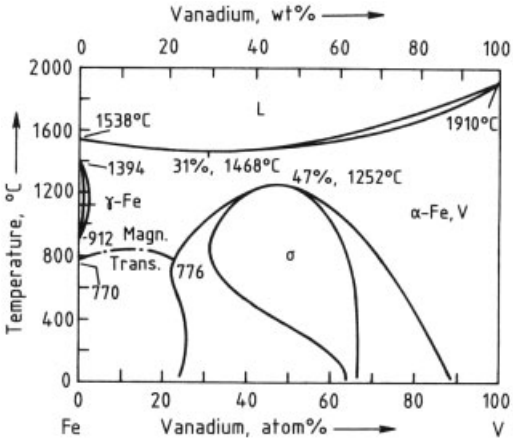

Figure 4. Phase diagram Fe – V [35]

5.3.1. Production from Vanadium Oxides

The reduction of vanadium oxides is assisted by the presence of iron, as the activity of the vanadium is reduced, and the solubility of oxygen in vanadium – iron is much lower than in pure vanadium, because of the lower melting point [21]. The vanadium oxides can therefore be reduced with carbon, silicon, and aluminum in the presence of iron, the product being ferrovanadium.

Reduction with Carbon. The high affinity of vanadium for carbon leads to carbide formation, so that carbothermic reduction can only be used when there is no requirement for vanadium with low carbon content. The Union Carbide Corporation has developed a process that gives a V_2C-containing alloy known as Carvan [36], with a C content in the range 10 – 13 % [33].

Reduction with Silicon. Under today's conditions for the processing of V_2O_5, silicothermic reduction is no longer generally profitable, as the process has multiple stages and gives a poor vanadium yield in the form of a low-vanadium ferrovanadium. The vanadium losses in this process are 10 – 25 % [37].

Reduction with Aluminum. The high-vanadium grades of ferrovanadium, FeV 60, FeV 80, and FeV 90, are now usually produced by aluminothermic treatment of V_2O_5. The aluminothermic process can readily be controlled on a large scale. It is especially suitable for the treatment of vanadium pentoxide because it gives a high yield of high-grade ferrovanadium in a single process step.

After ignition, the reaction of V_2O_5 with aluminum is self-sustaining. Some of the mixture of V_2O_5 flakes, aluminum powder or granules, fine steel shot, and an initiating mixture (e.g., BaO_2 + Al powder) is ignited, and further quantities of the mixture are then added. The combustion time for the usual batch size (to produce 0.5 – 1 t metal) is only a few minutes. After cooling for 2 – 3 d, the furnace is dismantled, and the block of metal at the bottom is cleaned and crushed or ground to the commercially desired particle size.

As the reaction proceeds very rapidly, the aluminum in the metal and the vanadium oxide in the slag may not attain equilibrium, so that the vanadium content of the slag may be too high. To prevent this, an electric arc is ignited after completion of the aluminothermic reaction to maintain the molten state of the melt and the slag until reaction is complete [38].

5.3.2. Direct Production from Slags and Residues

Enriched vanadium-containing oxidation slags (see Chap. 4) can also be processed directly to a technical-grade ferrovanadium. In the method

of Christiania Spigerverk [39], a 50 % ferrovanadium is obtained in an electric arc furnace by a two-stage reduction process. In the first stage, the FeO fraction of the slag is reduced with ferrosilicon (FeSi 75) with addition of lime, and in the second stage the slag obtained in the first stage is reduced with 90 % silicon to give a ferrovanadium alloy. The vanadium yield is 80 %, and the FeV still contains traces of Si, Cr, Mn and Ti.

Ferrovanadium can be obtained by the electrosilicothermic treatment of high-vanadium boiler ashes and enriched fly ash, either alone or in combination with vanadium-containing oxidation slags, depending on their composition. As oil residues and fly ash contain only small amounts of iron, the iron removal stage normally necessary with oxidation slags can be omitted.

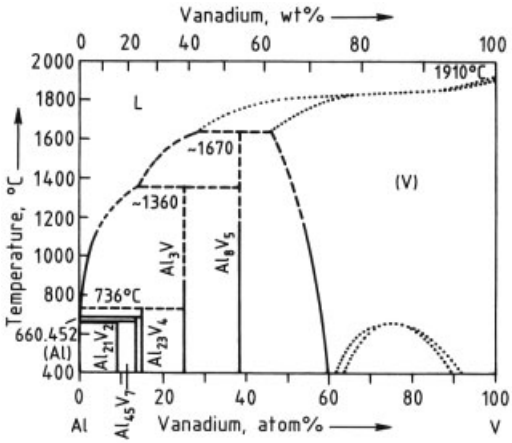

Figure 5. Phase diagram Al – V [35]

5.4. Production of Other Vanadium Master Alloys

Vanadium – Aluminum Master Alloys. In the production of vanadium-containing titanium alloys, especially TiAl 6 V 4, the vanadium is added as a vanadium – aluminum alloy. The Al – V phase diagram (Fig. 5) shows a higher solubility of aluminum on the vanadium-rich side. This results in a wide range of composition in which vanadium – aluminum alloys can be produced:

VAl 40/60	38 – 42 % V
VAl 50/50	45 – 54 % V
VAl 65/35	60 – 65 % V
VAl 80/20	78 – 82 % V
VAl 85/15	83 – 87 % V

Vanadium – aluminum alloys are produced aluminothermically. The excess heat produced in this self-sustaining reaction is controlled by slow charging of the reaction mixture. The vanadium – aluminum alloys have a lower density than vanadium (6.11 g/cm^3). The melting points of the alloys decrease with increasing aluminum content:

	Density, g/cm^3	Melting range, °C
VAl 80/20 (85 % V)	5.2	1850 – 1870
VAl 40/60 (40 % V)	3.8	1360 – 1520

Because of the small difference between the density of the metal and the slag, poor separation or inclusions of slag in the metal can occur in the aluminothermic process. As titanium alloys are used in highly stressed machinery and equipment, the master alloys used are subject to strict quality control and must be free of oxidic inclusions and impurities, especially heavy metals.

Vanadium – aluminum master alloys can also be produced in water-cooled copper molds, preventing contamination by impurities from slag and refractory oxides in the furnace lining [40]. However, the higher cooling rate can lead to inclusion of process slags.

Nitrogen-Containing Vanadium Master Alloys. As vanadium is a strong nitride former, vanadium-containing alloying agents can be nitrided and used for alloying steels with vanadium and nitrogen (see Chap. 5). Examples of nitrogen-containing alloys are ferrovanadium – nitrogen and manganese – vanadium – nitrogen [41].

Other commercial alloys are listed in the following:

VAlFe	69 % V, 19 % Al, 12 % Fe
VAlSn	43 % V, 43 % Al, 14 % Sn
VCrAl	50 % V, 40 % Cr, 10 % Al
VCo	50 % V, 50 % Co
VNi	65 % V, 35 % Ni
VMoAl	50 – 54 % V, 26 % Mo, 22 % Al

6. Uses

Vanadium is mainly used as an alloying element in the steel industry, other major areas of use are [42], [43]

Steel	ca. 85 %
Nonferrous alloys	ca. 9 %
Chemical industry	ca. 4 %
Others	ca. 2 %

The second largest use area is in nonferrous alloys, mostly vanadium-containing titanium alloys and nickel-based superalloys for the aerospace industry. Additional uses are in vanadium-containing alloys for batteries and in grain refining of aluminum alloys. Vanadium compounds, principally the oxidation catalyst V_2O_5, account for only ca. 3 % of vanadium consumption [12].

Addition to Steel. Vanadium is added to steel mainly as ferrovanadium; technical vanadium carbide and vanadium carbonitride are also used. For the simultaneous addition of vanadium and nitrogen, nitrogen-containing Fe – V alloys are available.

Even small additions of vanadium increase the tensile strength and high-temperature strength of carbon steel [1], [56], and it has a grain refining and dispersion hardening effect in tempering steels.

Addition to Titanium. Alloying with vanadium improves the properties of titanium. The most important alloy is TiAl 6 V 4, which has good strength properties at room temperature and good creep resistance. It is mainly used as a wrought alloy and casting alloy in airframe construction for load-bearing components and fixing devices, and in compressor disks and blades in jet engines. Other areas of increasing importance include power stations, shipbuilding, and reactor technology [13]. Other vanadium-containing titanium alloys with similar strength properties to TiAl6V4 include TiAl3–V2.5, TiV15Cr3Sn3Al3, TiV10Fe2Al3, TiAl6V6Sn2, as well as TiV13Cr11Al3 [44], [45].

In light, corrosion-resistant titanium aluminides (intermetallic γ-TiAl phase) with high-temperature strength, small amounts of vanadium (as well as of chromium and manganese) increase ductility at room temperature, which is an important criterion for usability and workability of this otherwise brittle material [46].

Other Uses. *Alloys for Batteries.* Intermetallic phases of the ZrV_2 type (Laves phases), in which vanadium is substituted partially by, e.g., Ni, Cr, Ti, or Mn, are used as electrode materials in metal hydride/nickel hydroxide batteries [47].

Coating materials containing vanadium and vanadium compounds are used in the electronics and glass industries as well as for wear protection.

Gettering materials based on Zr – V – Mn or Zr – V – Fe alloys are used for gas purification and for improving vacuums. The main component of these alloys, which contain up to 30 % vanadium, is zirconium [48]. Gettering materials are specially adapted to individual applications; thus, with the exception of some products of the company SAES Getters, only a few standard alloys are marketed.

Vanadium Metal and Alloys. The use of vanadium as a principal component of alloys is in its infancy. As vanadium has higher thermal conductivity and strength and lower thermal expansion and density than stainless steel [49], it has been considered for cladding nuclear fuels for sodium-cooled fast breeder reactors, in the form of either pure metal or an alloy. In the United States, in the late 1960s, a large-scale development program on vanadium alloys was undertaken by the Atomic Energy Division at Westinghouse. Wide-ranging cooperative research projects in the Federal Republic of Germany in 1964 – 1971 are reported in [31], [50], [51]. A comprehensive review of these developments can be found in [7].

Of the vanadium alloys investigated, those in which titanium, niobium, chromium, and zirconium are the alloying elements seem the most promising.

The behavior of vanadium and its alloys in contact with lithium and sodium is described in [11], [51]. The effect of irradiation by neutrons on the strength properties of vanadium and its alloys is described in [52].

A further possible application of vanadium is in superconductivity. Vanadium – gallium alloys with composition V_3Ga are well known, but

vanadium – niobium alloys containing 30 % Nb and vanadium – hafnium/zirconium alloys also have good superconducting properties.

7. Vanadium Compounds

A large variety of vanadium compounds are used industrially: halides, oxides, nitrates, carbides, silicides, hydrides, vanadates, and organometallics.

Vanadium Oxides. *Vanadium pentoxide* [*1314-62-1*], V_2O_5, M_r 181.88, *mp* 690 °C, ϱ 3.36 g/cm^3, heat of formation ΔH^{298} −1550.8 kJ/mol, is orange yellow in color and is traded as fused flakes or powder. Reaction with alkali metal hydroxides or ammonium hydroxide gives vanadates. V_2O_5 is prepared by thermal decomposition of ammonium vanadates at 500 – 600 °C in an oxidizing atmosphere (air). It is by far the most important vanadium compound. V_2O_5 is used as an oxidation catalyst in heterogeneous and homogeneous catalytic processes for the production of sulfuric acid from SO_2, phthalic anhydride from naphthalene or *o*-xylene, maleic anhydride from benzene or *n*-butane/butene, adipic acid from cyclohexanol/cyclohexanone, and acrylic acid from propane. Minor amounts are used in the production of oxalic acid from cellulose and of anthraquinone from anthracene.

V_2O_5 is used to lower the melting point of enamel frits for the coating of aluminum substrates. FeV and VAl master alloys are preferably produced from V_2O_5 fused flakes due to the low loss on ignition, low sulfur and dust contents, and high density of the molten oxide compared to powder.

Further uses of V_2O_5 are as a corrosion inhibitor in the CO_2 scrubbing solutions of the Benfield and related processes for the production of hydrogen from hydrocarbons, as cathode in primary and secondary (rechargeable) lithium batteries, as UV absorbent in glass, in YVO_4: Eu^{3+} red phosphors for high-pressure mercury lamps and TV screens, for glazes [53], for yellow (SnO_2/V_2O_5) and blue (ZrO_2/V_2O_5) pigments [13], as colloidal solution for antistatic layers on photographic material [54], and as a starting material for the production of carbides, nitrides, carbonitrides, silicides, halides, vanadates, and vanadium salts.

Vanadium trioxide [*1314-34-7*], V_2O_3, M_r 149.88, *mp* 1977 °C, ϱ 4.99 g/cm^3, heat of formation ΔH^{298} −1218.8 kJ/mol, black powder, is formed by thermal decomposition of ammonium vanadates at 600 – 900 °C in a reducing atmosphere (H_2, CO, CH_4, NH_3). The most important application is as an alternative to V_2O_5 in the production of FeV, VAl master alloys, high-purity vanadium metal, and vanadium carbides and carbonitrides [55].

Vanadium suboxides such as V_2O_4 [*12036-21-4*], M_r 82.94, *mp* 1642 °C, heat of formation ΔH^{298} −1427.37 kJ/mol, dark blue in color, or V_6O_{13} [*11099-11-9*], M_r 513.64, colored bluish black, are of minor industrial importance [56]. They are produced by heating stoichiometric mixtures of vanadium oxides or by reduction of ammonium vanadates or V_2O_5 in atmospheres of controlled reduction potential. Bronzes of V_6O_{13} or V_5O_{11} with Li_2O are used for lithium ion batteries.

Vanadates. The most important product is *ammonium metavanadate* [*7803-55-6*], NH_4VO_3, M_r 116.98, heat of formation ΔH^{298} −1551.0 kJ/mol, a white powder. Thermal decomposition starts at 70 °C. It is sparingly soluble in cold water (ca. 1 %), but readily soluble in mono- and diethanolamine. It is precipitated from neutral alkali metal vanadate solutions by the addition of ammonium chloride or ammonium sulfate. Because of its ready conversion to V_2O_5 at elevated temperatures in oxidizing atmospheres, it is used as a substitute in the production of, e.g., DENOX catalysts and zirconium vanadium oxide yellow ceramic colorants.

Ammonium polyvanadate [*11115-67-6*], $(NH_4)_2V_6O_{16}$, M_r 597.72, is an orange powder, produced by the addition of mineral acids and ammonium salts to alkali metal vanadates at pH 2 – 3 [57]. It is an intermediate product in the production of vanadium pentoxide.

Sodium ammonium vanadate [*39455-80-6*], $2\,(NH_4)_2O \cdot Na_2O \cdot 5\,V_2O_5 \cdot 15\,H_2O$, M_r1345.77, an orange powder, is precipitated from sodium vanadate solutions by the addition of ammonium salts at pH 5 – 6. It is almost exclusively used as a soluble oxidation catalyst for the desulfurization of H_2S-containing gases by the Stretford process.

Sodium metavanadate [*13718-26-8*], $NaVO_3$, M_r 121.93 and *potassium metavanadate* [*13769-43-2*], KVO_3, M_r 138.04, both white powders, are water-soluble compounds. They are added as corrosion inhibitors to CO_2 washing solutions in the production of H_2 from hydrocarbons.

Vanadium Salts. *Vanadyl sulfate* [*27774-13-6*], $VOSO_4 \cdot 5\,H_2O$, M_r 253.08, is produced by the reaction of V_2O_5 and SO_2 in aqueous media, and forms blue, water-soluble crystals. It is used in catalyst production as alternative to V_2O_5. In multivitamin/multimineral pills it is the active component for fast energy supply to the muscles by enhancing glucose metabolism in the blood. It acts as an orally applied insulin mimic.

$VOSO_4$, of blue color, and $V_2(SO_4)_3$, of green color, are the feed solutions for the vanadium redox battery. This battery contains vanadium(V) (yellow) and vanadium(II) (pink) in dilute sulfuric acid as electrodes when charged [58].

Vanadyl oxalate [*14974-48-2*], VOC_2O_4, M_r 154.96, forms blue, water-soluble crystals and is used in catalyst production as alternative to vanadates and vanadium oxides.

Vanadium Halides. Only *vanadium oxytrichloride* [*7727-18-6*], $VOCl_3$, M_r 173.30, *mp* −79.5 °C, *bp* 126.7 °C, ϱ 1.82 g/cm^3, a bright yellow liquid, *vanadium tetrachloride* [*7632-51-1*], VCl_4, M_r 192.75, *bp* 148 °C, a red brown liquid, and *vanadium trichloride* [*7718-98-1*], VCl_3, M_r 157.30, boiling range 300 – 400 °C (disproportionates), a deep purple solid, are of commercial interest. Whereas $VOCl_3$ is stable, VCl_4 slowly decomposes into solid VCl_3 and Cl_2. $VOCl_3$ is produced by the reaction of V_2O_5 and carbon with chlorine gas. For the production of VCl_4, oxygen-free vanadium compounds such as vanadium carbide or vanadium carbonitride are reacted with chlorine gas. The chlorides are used as catalysts in the production of EPDM (ethylene – propene – diene) rubber and polyethylene and are starting materials [59] for the preparation of organic vanadium compounds.

Organic Vanadium Compounds. Representatives of this class are vanadium(IV) and vanadium(III) acetylacetonates, naphthenates, octoates, and alkoxides (iso/*n*-butyl, iso/*n*-propyl), which are soluble in organic solvents. They are used in siccatives, antifouling paints, as catalysts for stereospecific organic synthesis, and in chemical vapor deposition (CVD).

Vanadium Carbides [*12070-10-9*]. The monocarbide, VC, has a theoretical carbon content of 19.08 % and melts at 2830 °C. It has a wide phase range. VC is isomorphous with VO and VN. Pure VC can be prepared by carburizing vanadium hydride powder with carbon in a vacuum [60]. Reaction at 800 °C leads to vanadium carbide in which C : V = 0.72 – 0.87.

Industrial-grade VC is produced by heating V_2O_5 or V_2O_3 with carbon at 1100 °C under a protective atmosphere of hydrogen [61], or by reacting V_2O_3 with carbon in a carbon-tube furnace. The crude carbide formed in the first stage is recarburized at 1700 – 2200 °C under vacuum. The final product contains 18.5 – 19 % total-C and up to 0.5 % free carbon. The VC is not completely free of oxygen because of its isomorphism with VO. VC is added to hard metals in amounts of 0.3 – 0.5 %. It acts as a grain growth limiter [3]. Another carbide, V_2C, with a theoretical carbon content of 10.54 %, also exists. This also has a large range of homogeneity, and decomposes at 1850 °C. It is of no industrial importance as such, but is formed in all carbon-containing vanadium alloys.

Vanadium Nitrides [*24646-85-3*]. Three nitrides exist: VN, V_2N, and V_3N. The mononitride, *mp* 2350 °C, is a grayish-brown powder, isomorphous with VC and VO [62]. Its theoretical composition is 78.45 % V, 21.55 % N, but its range of homogeneity extends from $VN_{1.0}$ to $VN_{0.71}$. It can be obtained by the reaction of vanadium tetrachloride with mixtures of nitrogen and hydrogen gas, using the technique of growth on tungsten wires at 1400 – 1600 °C. Although the preparation of VN from vanadium oxides is difficult because of the isomorphism mentioned above, a product containing 78.3 % V, 21.1 % N, and 0.5 % O can be obtained by reduction with carbon in an atmosphere of nitrogen [61].

As vanadium carbides and nitrides are mutually soluble, vanadium carbonitrides with a wide range of C : N ratios can be prepared at

1100–1400 °C. This is true for both V (C, N) and V_2 (C, N). Two broad three-phase regions extend continuously between the corresponding phases VC – VN and V_2C – V_2N.

Vanadium Silicides [*12039-87-1*]. VSi_2, *mp* 1650 °C, can be prepared by sinter metallurgy. For example, the pure substance can be prepared by exothermic reaction sintering of vanadium and silicon powder [63]. In industry, it is obtained by the reaction of vanadium oxides with SiO_2 and carbon (or with SiC) at 1200 – 1800 °C under vacuum.

Vanadium Hydrides [*13966-93-3*]. In the H – V system [64] the following phases occur: β_1, corresponding to ca. V_2H (low temperature), 33 atom % H; β_2, corresponding to V_2H or VH, 33 – 50 atom % H; δ, corresponding to V_3H_2, 40 atom % H; γ, corresponding to VH_2, 66.7 atom % H.

8. Analysis

The most important analysis is the determination of vanadium in raw materials, especially vanadium-containing slags and the commercial products vanadium oxide and ferrovanadium [65].

To determine vanadium in raw materials, the sample is fused with sodium peroxide and sodium carbonate, the solidified product is leached with water, and the vanadium-containing solution so obtained is acidified with sulfuric acid. The vanadium is completely oxidized with potassium manganate(VII) to the pentavalent state, and the solution is back-titrated with iron(II) sulfate, with diphenylamine as indicator.

In the analysis of vanadium oxide, the sample is dissolved in sulfuric acid, completely reduced with iron(II) sulfate to the tetravalent state, and back-titrated potentiometrically with potassium manganate(VII) solution, using a calomel reference electrode.

Ferrovanadium is also analyzed potentiometrically. The sample is dissolved in sulfuric/nitric acid, and is then oxidized by potassium manganate(VII). The vanadium, now in pentavalent form, is reduced with iron(II) solution, and potentiometrically back-titrated with potassium manganate(VII) solution, using a calomel reference electrode.

9. Economic Aspects

Consumption, Production Capacity, and Prices. The development of vanadium market prices is governed by consumer demand and existing production capacities. For main uses of vanadium and their shares see Chapter 6. The dependence of vanadium production on world steel production is shown in Figure 6 [66]. Total world demand is compared to available production capacities in Figure 7 [67]. This led to the price trend shown in Figure 8 [67].

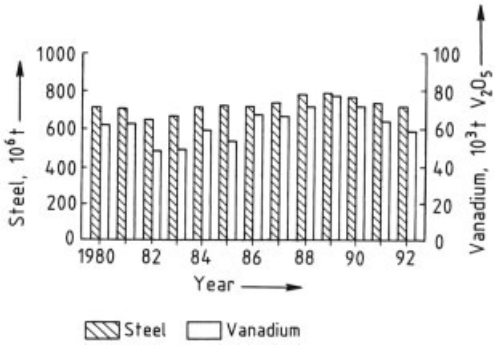

Figure 6. World production of steel and vanadium

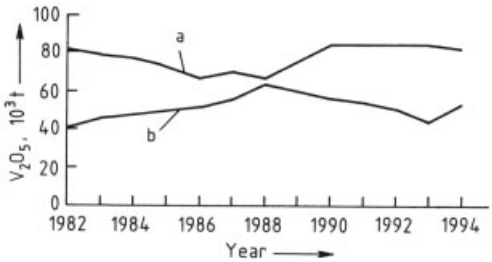

Figure 7. Worldwide vanadium capacity (a) and demand (b)

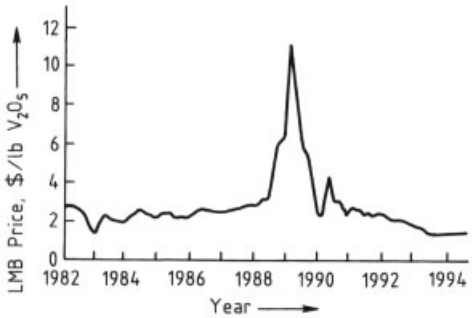

Figure 8. Vanadium price trends

For the vanadium demand of the western world, South Africa is the major exporter (Table 7). Vanadium in the form of slag and oxides was traded to converters in Europe, Japan, and the United States, for the production of ferrovanadium. Since 1994 most South African suppliers have installed their own ferrovanadium production capacity. This will eventually lead to the nonavailability of South African preproducts for the converters mentioned above. Their plants will have to use other raw materials, such as residues from mineral oil processing, or cease production.

Further suppliers of vanadium oxides are Russia, whose share for export strongly depends on its own economic development; and China, which is expected to turn from an exporter to an importer because of its high economic growth rate.

Ferrovanadium producers are listed in the following [61]:

Country	Company
Argentina	Pamet
	Stein Ferroaleaciones Sacifa
Austria	Treibacher Industrie AG
Belgium	SADACI NV
Brazil	Termoligas Metalurgicas SA
	Cia Paulista de Ferro Ligas
Canada	Masteralloy Products Inc.
	Panzihua Iron and Steel Plant
China	Emei Ferroalloy Plant
	Jinzhou Ferroalloy plant
Czech Republic	Nikom Ferroalloy Plant
Germany	GfE Gesellschaft für Elektrometallurgie
Hungary	Ötvözetgyar; Hungaraloy
India	Birla
Japan	Awamura Metal Industry Co. Ltd.
	Taiyo Mining and Industrial Co.
	NKK Corp.
Mexico	Ferroaleaciones de Mexico
	FERROMEX
	Ferralver SA
CIS	Chusovskoy Metallurgi- cal Kombinat
	Tulachermet Kombinat
Spain	Ferroaleaciones Especiales Asturianas
South Africa	Highveld Steel and Vanadium Corp. Ltd.
	Vametco
	Vantech
United States	BEAR Metallurgical Corp.
	Shieldalloy Corp.
	Strategic Metal Corp., Stratcor

10. Environmental Protection and Toxicology

Soil. In sulfidic and oxidized form, vanadium is a ubiquitous element with an average abundance of 150 mg/kg in the earth's crust. The concentrations vary widely, between a few ppm and 1000 ppm. Depending on geological conditions, even higher values can be found, e.g., 1650 ppm vanadium in oil shales of permic copper schist [68]. The vanadium content of oil shale and mineral oil, in which it is present primarily as porphyrin complexes, is due to biological processes. Higher anthropogenic burdens can be found only in a few isolated geographic locations. In the upper layers of arable soils utilization of mineral fertilizers leads to accumulation of vanadium, which is retained and is only slightly mobile [69].

Air. Vanadium concentrations in uncontaminated air can reach 1 ng/m^3, caused by whirled-up dust from weathered rock and by evaporation of water droplets driven off of surface waters by wind and waves. Increased concentrations are recorded mainly in urban areas. They originate mainly from combustion of coal and especially heavy fuel oil. Concentrations of ca. 10 – 35 ng/m^3 measured in the past (maximum values of up to 1000 ppm, recorded in London, 1962) are lower today, due to the use of efficient waste air purification systems. Industrial activities are of only minor, local importance in the pollution of the air with vanadium.

Water. Vanadium dissolved in water is present almost exclusively in the pentavalent form. In running waters, its concentration range is ca. 0.1 – 220 µg/L [70], the vanadium concentration in the water correlates to these in rocks near the river source in the river bed. The sediment of creeks also mirrors the abundance in the surrounding area. Seawater contains 0.3 – 29 µg/L vanadium [70]. Accumulation does not occur, since the vanadium input from running waters is transferred to the sediment by precipitation and biochemical processes.

Water contamination by vanadium takes place only to a minor extent. For removing vanadium from wastewater, single- or multiple-stage precipitation with iron or lime is employed.

Table 7. World exports of vanadium, 1989 – 1992 [66] (in tonnes of V_2O_5)

Country	Exports				
	1989	1990	1991	1992	%
South Africa [a]	32 950	25 989	25 561	24 000	84.7
New Zealand [b]	2 000	4 000	4 500	3 000	10.6
CIS	1 500	1 350	1 200	1 200	4.2
Others [c]	200	150	150	150	0.5
Total world	36 650	31 489	31 411	28 350	100

[a] Including Republic of South Africa and Bophuthatswana.
[b] Estimates.
[c] China is considered to be a net importer of vanadium at present. Its re-exports are, therefore, not shown in this table.

Table 8. Toxicological data [78], [79]

	Oral LD_{50} (14 d), mg/kg body weight	Inhalative LC_{50} (14 d), mg/L	Dermal LD_{50} (14 d), mg/kg body weight
V_2O_5, analytical grade	470 (male)	11.09 (male)	>2500 (male)
	467 (female)	4.3 (female)	>2500 (female)
V_2O_3, technical grade	8713 (male)	>6.65 (male)	>2500 (male)
	5639 (female)	>6.65 (female)	>2500 (female)
Ammonium metavandate	218 (male)	2.61 (male)	>2500 (male)
	141 (female)	2.43 (female)	>2500 (female)
Potassium metavanadate	318 (male)	1.85 (male)	>2500 (male)
	314 (female)	4.16 (female)	>2500 (female
$BiVO_4 - BiMoO_4$	>5000	>5.15	

Fauna and Flora. Vanadium is considered to be essential for a number of organisms [71–73]. In vertebrates the average concentration is 0.1 mg vanadium per kg dry matter. Considerably higher concentrations ($\leqq$ 0.1 g/kg) can be found in lower organisms. Especially in one order of invertebrates (ascidians), vanadium accumulates in unusually high amounts [74]. In some species of these organisms, the vanadium is present in green blood cells (vanadocites) and is complexed by pyrrole rings [75]. Data on vanadium concentrations in humans vary widely, between 100 µg [76] and 17 – 43 mg [77]. Fat tissue, bones, teeth, bone marrow, as well as serum are especially rich in vanadium.

In general, the vanadium concentration in plants is related to that in the soil. As with animals, plants vary widely in vanadium content. For example, in fly agaric (*Amanita muscaria*) concentrations 100-times as high as in other fungi are found, independent of location. The reasons for this are not yet known [74].

Toxicology. Toxicity data for some vanadium compounds are given in Table 8. The data show a correlation of toxicity with oxidation state and solubility. Toxicity decreases with decreasing oxidation state and decreasing solubility.

In investigations of neurobehavioral, neuropsychological, psychosomatic, and psychological parameters as a function of exposure to V_2O_5 in dust among workers in an industrial plant, no significant changes compared to uncontaminated persons were found, independent of the concentration of vanadium pentoxide [80].

Occupational Safety. For vanadium pentoxide, there is a total dust limit value (0.5 mg/m^3) and a limit value for fine dust (0.05 mg/m^3); in Germany, the definition of the latter is based on studies on volunteers [81].

11. References

1. R. Kieffer, G. Jangg, P. Ettmayer: *Sondermetalle,* Springer Verlag, Wien 1971.
2. A. M. Sage: "Discovery and History of Vanadium and its Contribution to Life in the Modern World," Metals Society, 1981.
3. R. Kiefer, H. Braun: *Vanadium, Niob, Tantal,* Springer Verlag, Berlin 1963.
4. W. Rostocker: *The Metallurgy of Vanadium,* Wiley, Chapman & Hall, New York 1958.

5. *Bull. Alloy Phase Diagrams* **2** (1981) no. 1, 40 – 41, 146; **2** (1981) no. 2, 172.
6. I. Barin, O. Knacke: *Thermochemical Properties of Inorganic Substances,* Springer Verlag, Berlin, Verlag Stahleisen, Düsseldorf 1973.
7. D. L. Harrod, R. E. Gold, *Int. Met. Rev.* **25** (1980) 163 – 221.
8. E. Fromm, E. Gebhardt: *Gase und Kohlenstoff in Metallen,* Springer Verlag, Berlin 1976.
9. H. E. Dunn, D. L. Edlung: "Vanadium", in C. A. Hampel (ed.): *Rare Metals Handbook,* 2nd ed., Reinhold, Chapman & Hall, London 1961, pp. 640 – 642.
10. D. Schlain, C. B. Kenehan, W. L. Ackermann, *J. Less-Common Met.* **3** (1961) 458 – 467.
11. R. L. Ammon, *Int. et. Rev.* **25** (1980) 255 – 268.
12. M. Rühle, *Metall (Berlin)* **35** (1981) 1168 – 1172, 1282 – 1285.
13. Bundesanstalt für Geowissenschaften und Rohstoffe: *Untersuchungen über Angebot und Nachfrage mineralischer Rohstoffe, XIV. Vanadium,* Berlin – Hannover 1981.
14. P. Storm: "Vanadium – Sources, Applications, and Markets," *Raw Materials Report,* **vol. 10, no. 3,** Royal Institute of Technology, Stockholm.
15. R. K. Evans: Vanitec, reprint from *Met. Mater.,* April 1978.
16. B. Rohrmann, A. G. Raper, *J. Iron Steel Inst. London* 1970, April, 336 – 341.
17. Ullmann, 3rd ed., **18,** 56.
18. L. White, *Eng. Min. J.* **177** (1976) Jan., 87 – 91.
19. H.-J. Retelsdorf, *Metall (Berlin)* **35** (1981) 1166 – 1167.
20. A. D. Mah: "Thermodynamic Properties of Vanadium and its Compounds," *Bur. Mines Rep. Invest.* **6727** (1966).
21. W. Schmidt: "Untersuchungen an Vanadiummetall und seinen aluminothermisch hergestellten Legie-rungen mit Aluminium und Eisen," Ph. D. Thesis, RWTH Aachen 1969.
22. M. F. Joly: *2nd United Nations Int. Conf. on the Peaceful Use of Atomic Energy,* Geneve, May 1958, Paper No. A/Cnf. 15/P/1274.
23. R. Kiefer, H. Bach, H. Lutz, *Metall (Berlin)* **21** (1967) 19 – 22.
24. R. Kiefer, F. Lihl, E. Effenberger, *Z. Metallk.* **60** (1969) 94 – 100
25. T. T. Campbell, F. E. Block, E. R. Anderson: "Reducing Vanadium Compounds in Bomb Reactors," *Bur. Mines Rep. Invest.* **6374** (1964).
26. P. V. S. Pillai, K. U. Nair, T. K. Mukherjee, C. K. Gupta, *Trans. Indian Inst. Met.* **26** (1973) no. 6, 24 – 30.
27. T. K. Mukherjee, C. K. Gupta, *J. Less-Common Met.* **27** (1972) 251 ff.
28. C. T. Wang, E. F. Baroch, S. A. Worcester, Y. S. Shen, *Met. Trans.* **1** (1970) no. 6, 1683 – 1689.
29. O. N. Carlson, H. R. Burkholder, G. A. Martsching, F. A. Schmidt: *Extractive Metallurgy Refractory Metals Proc. Symp. 1981,* pp. 191 – 203.
30. R. W. Buckmann, Jr., *Int. Met. Rev.* **25** (1980) 158 –162.
31. M. Rühle, *Metall (Berlin)* **33** (1979) 140 – 147.
32. DIN 17 563 Ferrovanadium, Technische Lieferbedingungen, Dec. 1965.
33. G. Roethe, W. Gocht: *Handbuch der Metallmärkte,* Springer Verlag, Berlin 1974.
34. GfE Gesellschaft für Elektrometallurgie mbH, company brochure Nürnberg, Aug., 94.
35. J. F. Smith: *Phase Diagrams of Binary Vanadium Alloys,* American Society for Metals Int., Met. Park, Ohio 1989.
36. UCC, US 3 334 992, 1962.
37. O. Smetana: "Ferrovanadin und Vanadinmetall," in *Durrer-Volkert*: *Metallurgie der Ferrolegierungen,* 2nd ed., Springer Verlag, Berlin 1972.
38. F. Goebel, OE 169 315, 1951.
39. Christiania Spigerverk, NO 115 556, 1967.
40. F. H. Perfect, *Trans. Metall. Soc. AIME* **239** (1967) 1282 – 1286.
41. Ges. f. Elektrometallurgie, DE 1 558 500, 1967.
42. R. Hähn: "Vanadin-Gewinnung und Verwendung," *Erzmetall* **40** (1987) no. 6, 298 ff.
43. O. E. Kraus, *Stahl Eisen* **109** (1989) no. 11, 547 –552.
44. H. B. Bomberger, F. H. Froes, P. H. Morton: "Titanium – A Historical Perspective, Titanium Technology: Present Status and Future Trends," *Tit. Dev. Ass.* 1985.
45. K. Rüdiger, *DVS-Ber.* **53** (1978) 39 – 47.
46. Y.-W. Kim, *J. Met.* **46** (1994) no. 7, 30.
47. A. Züttel, F. Meli, L. Schlapbach, *Z. Phys. Chem. (Munich)* **183** (1994) 355.
48. K. Ichimura, M. Matsuyama, K. Watanabe, *J. Vac. Sci. Technol. A* **5** (1985) no. 2, 3 – 4.
49. F. L. Yaggee, E. R. Gilbert, J. W. Styles, *J. Less-Common Met.* **19** (1969) 39 – 51.
50. M. Schirra, *Metall (Berlin)* **33** (1979) 455 – 465.
51. H. U. Borgstedt, *Metall (Berlin)* **33** (1979) 264 – 266.

52. R. E. Gold, D. L. Harrod, *Int. Met. Rev.* **25** (1980) 232 – 254.
53. *Nachr. Chem. Tech.* **16** (1968) no. 2, 23.
54. Eastman Kodak, FR 7 522 060, 1975.
55. R. Hähn: Dissertation, RWTH Aachen 1983.
56. Combustion Engineering, US 4 486 400, 1984.
57. U. M. Levanto, *Acta Polytech. Scand. Chem. Technol. Metall. Ser.* **82** (1969) 37.
58. M. Skyllaskazacos et al.: "Characteristics and Performance of 1 kW UNSW Vanadium Redox Battery" *J. Power Sources* **35** (1991) 399 – 404.
59. Akzo Chemicals Inc.: "Polymerization Catalysts," Bulletin 92-03, 1992.
60. G. V. Samsonov, V. Ya. Vanmanko, L. N. Okhremchuk, B. M. Rud, *Izv. Akad. Nauk SSSR Neorg. Mater* **10** (1974) no. 1, 52 – 56; *Chem. Abstr.* **80** (1974) 90 458.
61. R. Kieffer, F. Benesovski: *Hartstoffe,* Springer Verlag, Wien 1963.
62. P. Ettmayer, R. Kieffer, F. Hattinger, *Metall (Berlin)* **28** (1974) 1151 – 1156.
63. F. Binder, *Radex Rundsch.* 1975, no. 4, 539 – 540.
64. J. F. Smith, D. T. Peterson, *Bull. Alloy Phase Diagrams* **3** (1982) no. 1, 55 – 60.
65. Chemiker-Ausschuß: *Analyse der Metalle,* **vol. 1**: Schiedsanalysen, **vol. 2**: Betriebsanalysen, Springer Verlag, Berlin 1961.
66. G. Grohmann: *World Vanadium Review,* Report R 14/93, Rep. of South Africa, Department of Mineral and Energy Affairs.
67. W. Beattie: "Overview of Vanadium and Tungsten Commodity Markets," *Metals Week Presentation,* reprint, 19th Sept. 1994.
68. H. Fauth, R. Hindel, U. Siewers, J. Zinner: *Geochemischer Atlas Bundesrepublik Deutschland,* Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover 1985.
69. R. Schnabel, C. Koch, C. Bunke, G. Schmieder: "Zum Verhalten von Vanadin in der Umwelt," *Z. Gesamte Hyg. Ihre Grenzgeb.* **30** (1984).
70. Environmental Health Criteria 81: *Vanadium,* World Health Organization, Geneva 1988.
71. F. H. Nielsen, "Studies on the Essentiality of Some Elements Ascribed as Toxic – Arsenic, Boron, Lead, Tin and Vanadium," U.S. Department of Agriculture, Agricultural Research Service, Grand Forks Human Nutrition Research Center, Grand Forks, ND.
72. E. J. Underwood: "Vanadium" in: *Trace Elements in Human and Animal Nutrition,* 4th ed., Academic Press, New York 1977, pp. 388 – 397.
73. R. Nechay: "Mechanism of Action of Vanadate," *Ann. Rev. Pharmacol. Toxicol.* **24** (1984) 501 – 524.
74. Natl. Res. Council, Committee on Biolog. Effects of Atmospheric Pollutants, "Vanadium," National Academy of Science, Washington, D.C., 1977, pp. 117 104.
75. E. P. Levine, *Science (Washington, D.C.)* **133** (1961) 1352.
76. A. R. Byrne, L. Kosta: "Vanadium in Foods and in Human Body Fluids and Tissues," *Sci. Total Environ.* **10** (1978) 17 – 30.
77. H. A. Schroeder, J. J. Balassa, I. H. Tipton: "Abnormal Trace Metals in Man: Vanadium," *J. Chron. Dis.* **16** (1963) 1047 – 1071.
78. J. Leuschner, H. Haschke, G. Sturm: "New Investigations on Acute Toxicities of Vanadium Oxides," *Monatsh. Chem.* **125** (1994) 623 – 646.
79. H. Wienand, W. Ostertag: "Bismutvanadat/molybdat – ein neuartiges Farbpigment," *Farbe Lacke* **9** (1986) 118 ff.
80. Final Report Vanadium Pentoxide Investigation, Institute for Social Medicine, Innsbruck 1993/94, unpublished.
81. C. Zenz, B. A. Berg, *Arch. Environ. Health* **14** (1967) 709.

Vanillin → Flavors and Fragrances
Vapor Deposition → Thin Films
Varnishes → Paints and Coatings
Vasodilators → Antihypertensives
Vat Dyes → Anthraquinone Dyes and Intermediates
Veneer → Wood
Vermiculite → Silicates

Veterinary Drugs

RICHARD SAMS, The Ohio State University, Dept. of Veterinary Clinical Sciences, Columbus, Ohio 43210, United States

1. **Antibiotics and Antibacterial Drugs** . 2
1.1. **Sulfonamides** 2
1.2. **β-Lactam Antibiotics** 3
1.3. **Tetracyclines** 3
1.4. **Aminoglycosides/Aminocyclitols** . . . 4
1.5. **Macrolide Antibiotics** 6
1.6. **Amphenicols** 7
1.7. **Tiamulin** 8
1.8. **Peptide Antibiotics** 8
1.9. **Ionophore Antibiotics** 9
2. **Antimycotics** 10
3. **Coccidiostats** 10
4. **Parasiticides** 11
4.1. **Avermectin Derivatives** 11
4.2. **Benzimidazole Derivatives** 12
4.3. **Imidazothiazole Derivatives** 12
5. **Anti-inflammatory Drugs** 12
5.1. **Nonsteroidal Anti-inflammatory Drugs** 12
5.2. **Steroidal Anti-inflammatory Drugs** . 13
6. **Respiratory Stimulants** 15
7. **Anesthetics** 15
8. **Narcotic Agonists and Antagonists** . . 16
9. **Antihistamines** 17
10. **References** 18

Drugs used in veterinary medicine include those originally developed for use in human medicine and later investigated for the treatment of animals. Such drugs are often used to treat horses and companion animals such as dogs and cats but are generally not used to treat foodproducing animals unless drug residues in food and other food safety issues have been investigated. Human drugs used in animals include various antibiotics (e.g., penicillins, cephalosporins, aminoglycosides, and tetracyclines), drugs affecting the cardiovascular system (e.g., digoxin, lidocaine, furosemide), anticonvulsant drugs (e.g., phenobarbital and phenytoin), anesthetics (e.g., halothane, pentobarbital), and analgesics (e.g., morphine, oxycodone, phenylbutazone, and aspirin).

Other drugs have been developed specifically for use in veterinary medicine. Some of these are used to prevent or treat parasitic infections unique to animals. Other veterinary drugs are used to increase weight gain in cattle, sheep, and pigs to decrease production costs and shorten the time until the animal is slaughtered for meat. Other veterinary drugs are close structural analogues of and are used similarly to drugs used in human medicine. For example, the animal tranquilizer acepromazine is a close structural analogue of chlorpromazine and other human tranquilizers.

The following chapters describe some of the veterinary drugs approved for use in the prevention or treatment of various diseases of animals. Human drugs that are not specifically approved by a regulatory agency for use in veterinary medicine are not included.

1. Antibiotics and Antibacterial Drugs

1.1. Sulfonamides (see also → Chemotherapeutics, Chap. 2.7.3.)

Sulfonamides are derivatives of 4-aminobenzenesulfonamide. These substances are weak acids with differing degrees of water solubility.

Sulfonamides are broad-spectrum bacteriostatic drugs, active against a wide range of gram-positive and gram-negative bacteria. Their antibacterial activity is increased by various diaminopyrimidines,in combination with which they are frequently marketed. Sulfaquinoxaline has both antibacterial and coccidiostat activity.

Sulfonamides are believed to act by inhibiting the conversion of *p*-aminobenzoic acid to dihydrofolic acid in sensitive bacteria. The diaminopyrimidines such as trimethoprim and ormetoprim inhibit conversion of dihydrofolic acid to tetrahydrofolic acid. Inhibition of both steps

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_387

by administration of a combination of these two groups of drugs is common and leads to greater antibacterial efficacy than either drug alone. Some combinations are bactericidal to certain organisms.

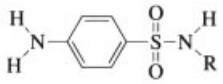

Sulfonamides

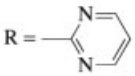

Sulfadiazine

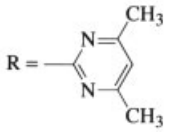

Sulfamethazine

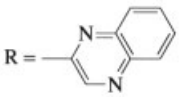

Sulfaquinoxaline

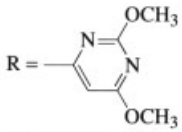

Sulfadimethoxine

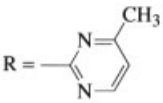

Sulfamerazine

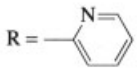

Sulfapyridine

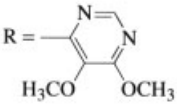

Sulfadoxine

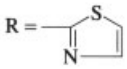

Sulfathiazole

Sulfamethazine [*57-68-1*] (sulfadimidine) is used in the treatment of bacterial enteritis and to maintain weight in the presence of atrophic rhinitis in pigs. In combination with chlortetracycline it is used to maintain weight gains in beef cattle suffering from respiratory diseases such as shipping fever. In combination with chlortetracycline and penicillin it is used to promote growth and increase feed efficiency in swine weighing up to 34 kg. Sulfamethazine is frequently found as an illegal residue in meats in the United States.

Sulfaquinoxaline [*59-40-5*] is used to treat or control outbreaks of coccidiosis caused by *Eimeria tenella, E. necatrix, E. acervulina, E. maxima*, or *E. brunetti* in chickens; by *E. meleagrimitis* or *E. adenoeides* in turkeys; and by *E. bovis* or *E. zurnii* in cattle. It is also used to treat or control fowl cholera caused by *Pasteurella multocida*, as well as fowl typhoid caused by sensitive organisms.

Sulfadimethoxine, sulfapyridine, sulfamerazine, sulfadoxine, and sulfadiazine are other sulfonamides used in veterinary medicine to treat infections caused by various bacteria including streptococci, staphylococci, escherichia, salmonella, klebsiella, proteus, or shigella organisms.

1.2. β-Lactam Antibiotics (see also → Antibiotics, Chap. 3.1.)

Penicillins are semisynthetic or synthetic derivatives of 6-aminopenicillanic acid produced by *Penicillium chrysogenum*, whereas cephalosporins are derivatives of 7-aminocephalosporanic acid, originally produced by cultures of *Cephalosporium acremonium.* Modifications of the side chains of both compounds have produced antibiotics with different antimicrobial spectra and pharmacokinetic properties.

β-Lactam antibiotics are bactericidal drugs with gram-positive activity in the older agents and broad-spectrum or gram-negative activity in some newer agents. Organisms producing β-lactamase hydrolyze the β-lactam ring of the penicillins and render them inactive. Clavulanic acid is a β-lactamase inhibitor that is marketed in combination with amoxicillin.

β-Lactam antibiotics inhibit cell wall synthesis in susceptible organisms. They inhibit transpeptidases, which prevents cell wall synthesis and eventually causes lysis of the bacterial cell.

The most commonly used penicillin in veterinary medicine is penicillin G. It is often formulated in combination with procaine or benzathine to produce a poorly soluble salt that dissolves slowly after intramuscular administration. Oral administration of penicillin results in relatively poor bioavailability due to hydrolysis in the acid environment of the stomach.

Semisynthetic penicillins such as ampicillin [*69-53-4*], amoxicillin [*26787-78-0*], hetacillin

[*3511-16-8*], and ticarcillin are used in veterinary medicine to treat infections caused by susceptible organisms. Amoxicillin is a broad-spectrum antibiotic active against many gram-positive and gram-negative organisms, except those producing β-lactamase.

Cephalosporins approved for use in veterinary medicine include cefadroxil and ceftiofur. Cefadroxil [*66592-87-8*] is a first-generation cephalosporin with good activity against gram-positive bacteria and some activity agains gram-negative organisms. Ceftiofur [*80370-57-6*] is active against gram-positive and gram-negative bacteria, including those that produce β-lactamase.

1.3. Tetracyclines (see also → Antibiotics, Chap. 3.2.)

Chlortetracycline and oxytetracycline are produced by *Streptomyces aureofaciens* and *S. rimosus*, respectively. Tetracycline is produced from chlortetracycline, and doxycycline is a semisynthetic derivative. Many of the tetracyclines form chelates with cations such as calcium and magnesium, rendering them inactive. Chlortetracycline has a chlorine atom at C-7 and oxytetracycline bears a hydroxy group at C-5, compared to tetracycline. Doxycycline has a hydroxy group and a hydrogen atom at C-5 and a methyl group at C-6, compared to tetracycline. All are weak bases that form salts with strong acids such as hydrochloric acid.

The tetracyclines are broad-spectrum antibiotics that are bacteriostatic in vitro. Good activity against various rickettsiae such as those responsible for Potomac Horse Fever has been demonstrated. Bacterial resistance often develops after exposure of bacteria to the tetracyclines by plasmid-mediated interference with transport of the drug to the ribosome, increased rate of removal from the ribosome, or decreased effect on the ribosome.

The tetracyclines inhibit bacterial protein synthesis by reversible binding to the 30 S ribosomal subunit of sensitive bacteria.

Tetracycline [*60-54-8*] is marketed in oral formulations including tablets and a soluble powder for dissolution in drinking water and milk. It is used in the treatment of bacterial enteritis caused by *Escherichia coli* and bacterial pneumonia caused by various species of *Pasteurella, Hemophilus*, and *Klebsiella*.

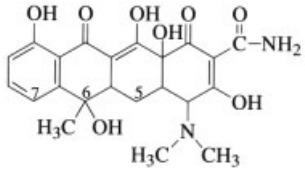

Tetracycline

Oxytetracycline [*79-57-2*] is marketed in a number of formulations including medicated feeds, parenteral products, and ointments. Terramycin Soluble Powder contains oxytetracycline hydrochloride and is used to treat bacterial enteritis in several species, shipping fever complex in cattle, liver abscesses in cattle, atrophic rhinitis in swine, leptospirosis in swine, enterotoxemia in sheep, as well as bluecomb, air-sac infection, sinusitis, fowl cholera, and hexamitiasis in poultry.

Chlortetracycline [*57-62-5*] is marketed in a number of products for veterinary use. As Aureomycin 2G Type C Medicated feed for swine, beef and dairy cattle, sheep, and horses it is used to promote growth and improve feed efficiency and as an aid in the prevention of bacterial swine enteritis. It is also employed in the treatment of bacterial enteritis and in reducing the spread of leptospirosis in swine. It is utilized in cattle to reduce losses due to shipping fever complex and for prevention of anaplasmosis, bacterial diarrhea, and bacterial foot rot. Aureomycin 70 Type A Medicated Article is used in the manufacture of various animal feeds. It is used to treat chronic air-sac infection and bluecomb and to prevent synovitis in chickens.

Doxycycline [*17086-28-1*] is used to treat infections caused by susceptible bacteria; in contrast to the other tetracyclines it is not usually administered as a feed additive.

1.4. Aminoglycosides/Aminocyclitols

(see also → Antibiotics, Chap. 3.3.)

The aminoglycoside antibiotics are produced by species of *Streptomyces* ("mycins") and *Micromonospora* ("micins"). They consist of one or more aminosugars attached to an aminocyclitol ring through a glycosidic bond. The aminocyclitol is either streptidine (streptomycin and dihydrostreptomycin) or 2-deoxystreptamine (all other available aminoglycosides).

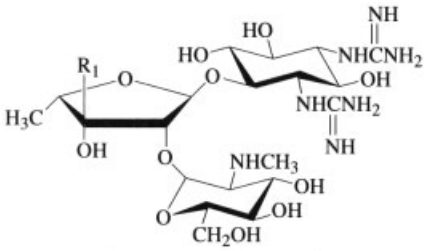

Streptomycin: $R_1 = CHO$
Dihydrostreptomycin: $R_1 = CH_2OH$

The aminoglycosides are highly polar drugs due to the large number of hydroxy and amino groups.

The aminoglycoside antibiotics are used primarily to treat infections caused by susceptible gram-negative aerobic bacteria. Bacterial resistance is plasmid mediated and results in bacterial modification of the aminoglycoside; the modified aminoglycosides are not clinically active. Strict anaerobes are resistant to aminoglycosides because they lack the mechanism necessary to transport the drug to the site of action. Due to their high polarity the aminoglycoside and aminocyclitol antibiotics are not well absorbed after oral administration. However, they are sometimes administered orally to reduce bacterial populations in the gastrointestinal tract. The aminoglycoside antibiotics most commonly used in food animal production are gentamicin, neomycin, streptomycin, and dihydrostreptomycin.

The aminoglycosides bind with the 30 S ribosomal subunit of susceptible bacteria and disrupt protein synthesis, resulting in a rapid bactericidal effect.

Aminoglycoside Antibiotics. *Streptomycin* [*57-92-1*] is produced by *Streptomyces griseus*; it is marketed as the sulfate salt in combination with bacitracin methylene disalicylate as Entromycin Powder for the treatment of bacterial enteritis and the associated diarrhea in dogs.

Dihydrostreptomycin [*128-46-1*] is a semisynthetic aminoglycoside produced by the reduction of streptomycin; it is no longer used in human medicine because of ototoxicity. It is marketed in combination with procaine penicillin G in an oil suspension as Quartermaster Suspension for intramammary use to reduce existing infection and to prevent new infection with *Staphylococcus aurea* in dry cows.

Kanamycin [*8063-07-8*] is an antibiotic complex produced by *Streptomyces kanamyceticus* and consists primarily of kanamycin A and two minor components, one of which is designated kanamycin B. Kanamycin is marketed in combination with bismuth subcarbonate and attapulgite as Amforol suspension and tablets for treatment of bacterial enteritis and the associated diarrhea in dogs.

Amikacin [*37517-28-5*] is a semisynthetic aminoglycoside derived from kanamycin A. It frequently has activity against organisms that have developed resistance to other aminoglycosides.

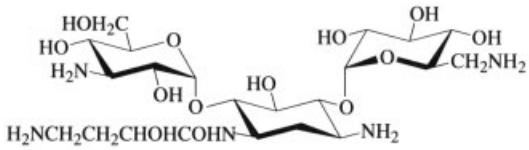

Amikacin

It is marketed as Amiglyde-V as the sulfate salt for injection and intrauterine infusion. The parenteral formulation is indicated for the treatment of skin and soft tissue infections caused by *Pseudomonas* species and *E. coli*, and urinary tract infections caused by *E. coli* and *Proteus* species. The intrauterine infusion is used for the treatment of uterine infections caused by susceptible organisms in the mare.

Gentamicin [*1403-66-3*] is an antibiotic complex produced by *Micromonospora purpurea* and *M. echinospora*; it consists of three substances designated gentamicins C_1, C_{1a}, and C_2.

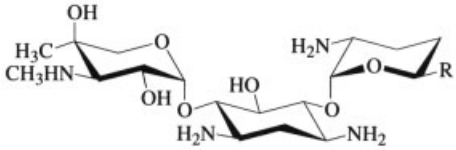

Gentamicin C_1: R = $CHCH_3NHCH_3$
Gentamicin C_{1a}: R = CH_2NH_2
Gentamicin C_2: R = $CHCH_3NH_2$

It is marketed as the sulfate salt as Garacin oral solution and in injectable form for treatment of colibacillosis in weanling swine caused by strains of *E. coli* sensitive to gentamicin, and swine dysentery associated with *Treponema hyodysenteriae*. Gentamicin is also available as Gentocin solution for parenteral administration to control bacterial infections of the uterus in horses and as an aid in improving conception in mares.

Aminocyclitol Antibiotics. *Spectinomycin* [*1695-77-8*] is an aminocyclitol antibiotic produced by *Streptomyces spectabilis*.

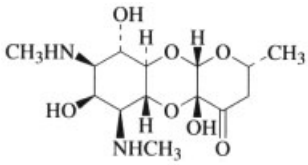

Spectinomycin

It is marketed as an oral solution as Spectam Scour-Halt for the treatment of infectious diarrhea in baby pigs caused by *E. coli*. It is also available as Spectam Injectable for treating airsaculitis and chronic respiratory disease associated with susceptible organisms in turkey poults.

Apramycin [*37321-09-8*] is an aminocyclitol antibiotic that is a component of the nebramycin complex and is produced by a strain of *Streptomyces tenebarius.* Apramycin is marketed as its sulfate salt as Apralan 75 Premix/Medicated, a type A medicated article for oral use in pigs, and Apralan Soluble Powder. Both products are used for the control of porcine colibacillosis caused by susceptible strains of *E. coli*.

1.5. Macrolide Antibiotics

Macrolide antibiotics (see also → Antibiotics, Chap. 3.5.,) all possess a macrocyclic lactone ring attached to one or more amino sugars. The macrolide ring may contain 12, 14, 16, or 17 atoms.

Macrolide antibiotics have good antibacterial activity against many gram-positive organisms such as *Mycoplasma* and *Chlamydia* but have little or no activity against most gram-negative bacteria.

The macrolide antibiotics inhibit protein synthesis by binding to the 50 S ribosomal subunits of sensitive microorganisms. Gram-positive microorganisms accumulate much more macrolide antibiotic than do gram-negative microorganisms. Since bacterial cell walls are much more permeable to the unionized form, these drugs generally exhibit greater antimicrobial activity under alkaline conditions.

Erythromycin [*114-07-8*] is a mixture of three components, designated Erythromycin A, Erythromycin B, and Erythromycin C, produced by *Streptomyces erythreus*. Erythromycin A is the major component. All are macrolide antibiotics consisting of a fourteen-membered lactone ring attached to desosamine, an amino sugar, and cladinose or mycarose. Erythromycin, isolated in 1952, was the first macrolide antibiotic widely used in medicine. It is active against a wide variety of gram-positive bacteria and mycoplasmas and is widely used in veterinary medicine. Erythromycin is degraded by stomach acids and is therefore marketed as various esters (e.g., stearate and ethylsuccinate) or enteric coated products for oral administration. Erythromycin has a relatively large distribution volume and is extensively distributed to various tissues, particularly the lungs, kidneys, and liver, where it may attain tissue concentrations higher than the corresponding plasma concentration.

Erythro-100 and Erythro-200 Injectable are erythromycin preparations in sterile solution containing 100 and 200 mg/L of erythromycin for intramuscular administration. In sheep it is indicated for the prevention of lamb dysentery and in the treatment of upper respiratory infections. In pigs it is used for treating respiratory syndrome and as an aid in the management of

mastitis, metritis, and leptospirosis in sows at farrowing time and scours in young pigs. It is indicated in cattle for the treatment of pneumonia, shipping fever, mastitis, metritis, foot rot, and as an aid in reducing weight loss during shipping.

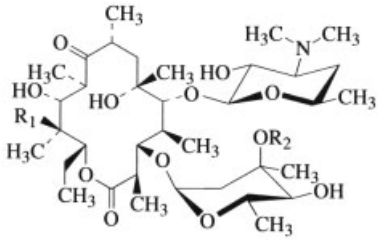

	R_1	R_2
Erythromycin A	OH	CH_3
Erythromycin B	H	CH_3
Erythromycin C	OH	H

Gallamycin-Dry Cow and Gallimycin-36 are tubes containing erythromycin for administration into the udders of nonlactating and lactating cows, respectively, for the treatment of acute and chronic mastitis caused by *Staphylococcus aureus, Streptococcus agalactiae, S. dysgalactiae,* and *S. uberis*. Uddermate is indicated for mastitis treatment in lactating cows only.

Oleandomycin [*3922-90-5*] consists of a single component produced by *Streptomyces antibioticus*. It is a macrolide antibiotic consisting of a fourteen-membered lactone ring attached to desosamine and another sugar moiety. Oleandomycin has an antibacterial spectrum similar to that of erythromycin but is less potent. It is marketed as its phosphate salt and as its triacetyl derivative, troleandomycin.

Tylosin [*1401-69-0*] is a macrolide antibiotic containing a sixteen-membered lactone ring substituted with various sugar moieties. Tylosin is a mixture of four related compounds, designated tylosin A, B, C, and D, with tylosin A the major component.

Tylosin is used as a feed additive for growth promotion and as an antibacterial drug in veterinary medicine. It is one of the most effective macrolide antibiotics against *Mycoplasma*. Tylosin was isolated from a strain of *Streptomyces fradiae* found in soil from Thailand.

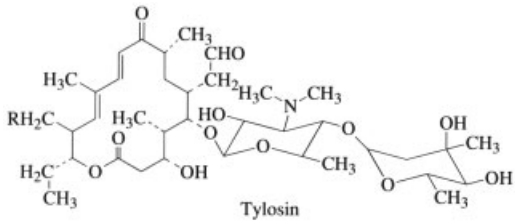

Tylosin

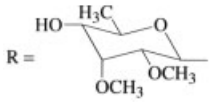

Tylan 40 is a feed additive and antibiotic containing 40 g of tylosin as tylosin phosphate per pound. Tylan 40 is used in swine to maintain weight gain and feed efficiency in the presence of atrophic rhinitis. It is also used in swine to increase the rate of weight gain and to improve feed efficiency; it is employed to prevent, treat, and control swine dysentery. It is used to reduce the incidence of liver abscesses caused by *Spherophorus necrophorus* and *Corynebacterium pyogenes* in beef cattle. It is utilized in chickens to improve feed efficiency and to control chronic respiratory disease caused by *Mycoplasma gallisepticum.*

Tylosin Injection and Tylan Injection are formulated with 50 and 200 mg of the base per milliliter. The solution is administered intramuscularly to beef cattle and nonlactating dairy cattle to treat bovine respiratory complex associated with *Pasteurella multocida* and *Corynebacterium pyogenes*, foot rot and calf diphtheria caused by *Fusobacterium necrophorum*, and metritis caused by *Corynebacterium pyogenes.* Tylosin Injection is administered intramuscularly to swine to treat swine arthritis caused by *Mycoplasma hyosynoviae,* swine pneumonia caused by *Pasteurella* spp, swine erysipelas caused by *Erysipelothrix rhusiopathiae*, and acute swine dysentery associated with *Treponema hyodysenteriae*.

Spiramycin [*8025-81-8*] is a mixture of three macrolide antibiotics containing a sixteen-membered lactone ring, substituted with amino sugars and mycarosyl. The components are designated spiramycin I, II, and III, with spiramycin I being the most important. The antimicrobial spectrum of spiramycin is similar to that of other macrolide antibiotics, but its activity

against *Treponema* is greater than that of erythromycin or oleandomycin.

1.6. Amphenicols

Chloramphenicol is produced by *Streptomyces venezuelae*; it is unique among natural compounds in having a nitrobenzene moiety and a dichloracetamide group. The active form of chloramphenicol has the *threo* configuration. Chloramphenicol is not ionizable at physiologic pH values and is relatively lipophilic. Thiamphenicol differs from chloramphenicol in that it has a methylsulfonyl group in place of the nitro group. Florfenicol differs from thiamphenicol by having a fluorine atom substitued for the hydroxy group at C-3. Thiamphenicol and florfenicol are totally synthetic.

The amphenicols are broad-spectrum antibiotics that are primarily bacteriostatic but may be bacteriocidal to certain bacteria. Gram-negative bacteria generally acquire resistance to chloramphenicol and thiamphenicol through plasmid-mediated transfer of acetyl transferase, which inactivates these drugs.

Florfenicol, lacking the hydroxy group that is acetylated in these other drugs, is not affected by the presence of acetyl transferases and therefore is effective against many chloramphenicol-resistant organisms.

Amphenicol antibiotics bind to the 50 S ribosomal subunits of susceptible bacteria, thereby inhibiting protein synthesis.

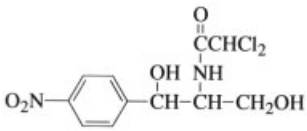

Chloramphenicol

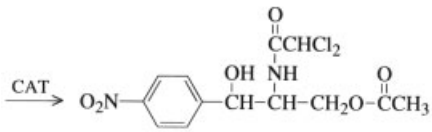

3-Acetylchloramphenicol

Chloramphenicol [*56-75-7*] is a broad-spectrum antibiotic marketed as Bemacol ointment and Chlorbiotic ointment for topical treatment of bacterial conjunctivitis due to susceptible organisms in dogs and cats. It is also available as chloramphenicol tablets and capsules for treatment of bacterial infections in dogs. Chloramphenicol palmitate is an oral suspension for treatment of bacterial infections in dogs.

Chloramphenicol use in food-producing animals is not allowed due to the possibility of food contamination.

Thiamphenicol [*15318-45-3*] is a broad-spectrum antibiotic used in veterinary medicine. The D-form is used as an antimicrobial drug, and the DL-form in the control of fowl cholera.

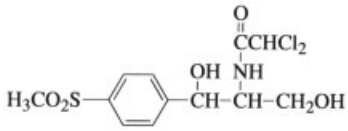

Thiamphenicol

Florfenicol [*73231-24-2*] is a broad-spectrum antibiotic used to treat bovine respiratory disease complex. It is also effective in the treatment of various fish diseases such as furunculosis in Atlantic salmon, pseudotuberculosis in yellowtail, and vibriosis in goldfish.

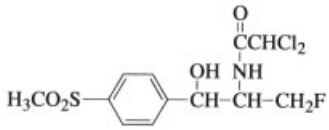

Florfenicol

1.7. Tiamulin

Tiamulin [*55297-95-5*] is a diterpene antibiotic containing an eight-membered carbocyclic ring. It is a derivative of pleuromutilin, an antibiotic substance produced by *Pleurotus mutilis* and *P. passeckerianus*.

Tiamulin is effective against gram-positive bacteria, *Mycoplasma*, and *Serpulina hyodysenteriae*. Tiamulin interferes with the metabolism and elimination of polyether ionophores (e.g., monensin, salinomycin, lasalocid, and narasin). This interaction is a potential cause of toxicity

when used in combination with these antibiotics. For this reason the combination of tiamulin and ionophore antibiotics is generally contraindicated.

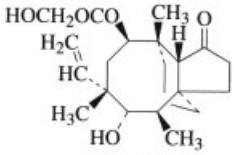

Tiamulin

Denagard soluble antibiotic is a fine, white granular powder containing 45 wt % tiamulin designed for administration in the drinking water of swine to treat infections caused by *Treponema hyodysenteriae* and *Hemophilus pleuropneumonia*. Denagard 10 is a feed additive and antibiotic containing 10 g of tiamulin as the hydrogenfumarate salt per pound. This product is used to treat swine dysentery associated with *Treponema hyodysenteriae* and as a feed additive to increase weight gain. Swine being treated with Denagard 10 should not have access to feeds containing ionophore antibiotics.

1.8. Peptide Antibiotics

The peptide antibiotics (see also → Antibiotics, Chap. 3.7.) are a diverse group of compounds having peptides arranged in linear or ring structures. Many polypeptide antibiotics have been used in veterinary medicine as feed additives and drugs. Bacitracin, enramycin, and virginiamycin are widely used for growth promotion, improvement of feed efficiency, and treatment of disease.

The various polypeptide antibiotics have different antibacterial spectra and produce their antibacterial effects by a variety of mechanisms.

Actaplanins are a complex of antibiotics produced by *Actinoplanes missouriensis*. Six components designated A, B_1, B_2, B_3, C_1, and G have been isolated and characterized. The drugs have a central peptide core, with the amino sugar ristosamine and up to four neutral sugars attached. The actaplanins are used to promote growth and increase milk production in ruminants.

Avoparcin [*37332-99-3*] is a glycopeptide antibiotic complex produced by *Streptomyces candidus*. It consists of a mixture of α- and β-avoparacin (chlorine-substituted analogue) and is used as an antibacterial agent and growth enhancer.

Bacitracin [*1405-87-4*] is a polypeptide complex produced by *Bacillus subtilis* and *licheniformis*. The commercial product is a mixture of at least nine bacitracins. It is used as local antibiotic, growth promoter, and for enteric infections.

Virginiamycin is an antibiotic complex produced by a *Streptomyces* related to *S. virginiae*. It is a mixture of two principal components designated virginiamycin M_1 [*21411-53-0*] and virginiamycin S_1 [*23152-29-6*]. Virginiamycin is used as an antibiotic and as a feed additive.

Bambermycin [*11015-37-5*] is an antibiotic complex consisting of at least four components: moenomycins A, B_1, B_2, and C with moenomycin A the major component. It is obtained from cultures of *Streptomyces bambergiensis* and others. Bambermycin is used as an antibacterial agent and as a growth promoter in poultry, swine, and calves.

1.9. Ionophore Antibiotics

The ionophore antibiotics are classified as polyethers and contain an acidic functional group at one end of the molecule. The acid group forms a hydrogen bond with a hydroxyl group at the other end of the molecule, giving an inclusion complex with alkyl groups on the outer surface and polar groups on the inner surface. Cations bind to the polar groups on the inner surface.

Ionophore antibiotics were initially developed as poultry coccidiostats but are now also used to promote growth of cattle. They enhance propionic acid production in ruminants by altering rumen fermentation. They are usually added at a rate of 5 – 30 ppm of finished ration to improve feed efficiency of feedlot cattle.

Ionophore antibiotics facilitate transport of monovalent and divalent cations across biological membranes. They selectively transport

sodium and potassium ions between extracellular and intracellular spaces and interfere with potassium ion transport across mitochondrial membranes.

Monensin [*17090-79-8*] is a major component of the antibiotic complex isolated from *Streptomyces cinnamonensis*. It has antibacterial, antifungal, and antiprotozoal activity and is a coccidiostat in chickens. Coban 45 and Coban 60 are feed additives and coccidiostats containing 45 and 60 g of monensin sodium salt per pound for use in poultry; Rumensin 60 is a feed efficiency enhancer for feedlot and pasture cattle containing 60 g of monensin sodium salt per pound. (See picture next page.)

Narasin [*55134-13-9*] is the main component of the polyether antibiotic complex produced by *Streptomyces aureofaciens*. It is used as a coccidiostat and as a growth stimulant. Monteban 45 is marketed as a feed additive and coccidiostat at a concentration of 45 g per pound. Maxiban 72 is a feed additive and coccidiostat containing both narasin and nicarbazine at a concentration of 36 g per pound for use in broiler chickens. (See picture next page.)

Maduramicin [*84878-61-5*] is a polyether antibiotic isolated from a species of *Nocardia*. It is used as a feed additive and coccidiostat. Cygro Type A Medicated Article is a feed additive and coccidiostat containing 4.54 g of maduramicin ammonium salt per pound for prevention of coccidiosis in broiler chickens caused by various species of *Eimeria*.

Lasalocid [*25999-31-9*] was originally isolated from an unknown species of *Streptomyces* from Hyde Park, Massachusetts. It is used as a feed additive and coccidiostat in chickens. Avatec is a medicated premix containing 20 % of sodium lasalocid in a carrier suitable for use in feed for broiler chickens for prevention of coccidiosis. Bovatex is a medicated premix that contains 15 % of sodium lasalocid in a carrier suitable for incorporation in feed to improve feed efficiency and increased rate of gain when used for cattle fed in confinement for slaughter. (See picture next page.)

2. Antimycotics (see also → Antimycotics)

Miconazole and clotrimazole are substituted imidazoles containing one or more aromatic rings with chlorine substituents.

Miconazole [*22916-47-8*] is marketed as its nitrate salt in a number of topical products (e.g., Conofite Cream, Lotion, and Spray and Dermazole Shampoo) for the control of fungal infections caused by *Microsporum canis, M. gypseum*, and *Trichophyton mentagrophytes* in dogs and cats.

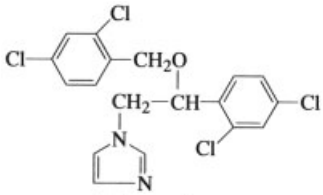

Miconazole

Clotrimazole [*23593-75-1*] is marketed as Veltrim Dermatologic Cream for treatment of fungal infections caused by *Microsporum canis* and *Trichophyton mentagrophytes* in dogs and cats.

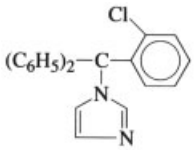

Clotrimazole

3. Coccidiostats

Coccidiostats are a diverse group of drugs, including the ionophore antibiotics and other substances useful in the control or treatment of coccidiosis.

Coccidiostats affect various stages of the life cycle of species of *Eimeria* that infect the digestive tracts of birds, causing economic losses in commercial poultry operations. Resistance to coccidiostats may develope upon repeated use. Therefore, it has become common practice in the broiler industry to use two or more coccidiostats sequentially at various intervals.

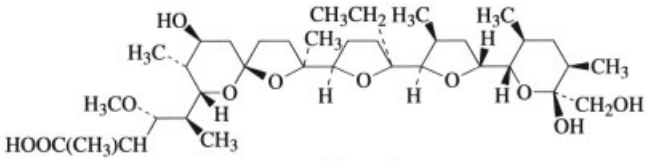

Monensin

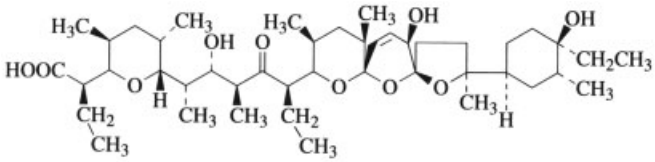

Narasin

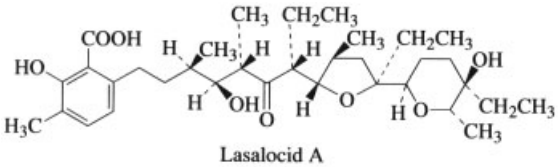

Lasalocid

Clopidol [*2971-90-6*] is mainly coccidiostatic and is active only against the sporozoite stage of *Eimeria*. It is effective only if administered on the day of exposure to the coccidial oocysts. Clopidol is marketed as a feed additive and is fed at a rate of ca. 125 ppm in the feed.

Amprolium [*121-25-5*] interferes with thiamine utilization in the parasite.

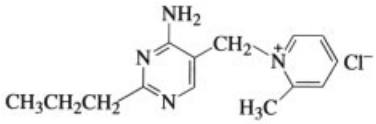

Amprolium is rarely used alone because it has relatively weak activity against *Eimeria maxima, E. mivati*, and other species. It is the only coccidiostat without a withdrawal time in the poultry industry. Amprolium is marketed as Aprol, Amprovine, and Corid in various soluble powders and oral solutions for treatment of coccidiosis in growing chickens, turkeys, and laying hens.

Decoquinate [*18507-89-6*] is a 4-hydroxyquinolone derivative that prevents development of the sporozoite stage by disruption of electron transport in the mitochondrial cytochrome system of the coccidia. Decoquinate is marketed as Deccox as a top dressing for addition to the feed for prevention of coccidiosis caused by *Eimeria bovis* and *Eimeria zurnii* in ruminating calves and cattle.

4. Parasiticides

4.1. Avermectin Derivatives

The avermectins are natural and semisynthetic derivatives of pentacyclic 16-membered lactones related to the milbemycins. The natural compounds are produced by *Streptomyces avermitilis*.

The avermectin antibiotics are very effective and highly potent anthelmintics that are effective by a number of administration routes. They are effective against a wide variety of nematodes as well as external parasites such as lice, flies, and bots. Ivermectin is also effective against the microfilarial and infective stages of canine heartworm disease caused by *Dirofilaria immitis*. Concerns have been expressed about their effects on the environment since they are excreted

in the feces and may affect insects and other organisms responsible for the normal degradation of dung.

These drugs apparently inhibit motility of parasites by increasing the release of γ-aminobutyric acid (GABA) thereby inhibiting neurotransmission. Worms are thereby paralyzed and are expelled from the animal.

Ivermectin [*70288-86-7*], 22,23-dihydroabamectin, is a semisynthetic derivative of abamectin. Ivermectin contains at least 80 % of 22,23-dihydroabamectin B_{1a} and not more than 20 % of 22,23-dihydroabamectin B_{1b}. Ivermectin is marketed as Eqvalan liquid and paste for use in horses for the treatment and control of large

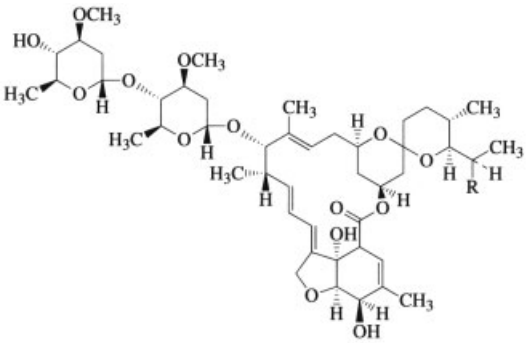

Ivermectin

Component B_{1a}: R = CH_2CH_3
Component B_{1b}: R = CH_3

strongyles, small strongyles, hairworms, pinworms, ascarids, intestinal thread-worms, large-mouth stomach worms, bots, lungworms, summer sores, and cutaneous onchocerciasis. Ivermectin is also available as Heartgard-30 tablets to prevent canine heartworm disease by eliminating the tissue stage of heartworm larvae for a month after infection, and for the treatment and control of ascarids (*Toxacara canis* and *Toxascaris leonina*) and hookworms (*Ancylostoma caninum* and *Uncinario stenocephala*). Ivomec is a 1 % ivermectin solution for treatment and control of internal and external cattle parasites and for treatment and control of gastrointestinal roundworms, lungworms, lice, and mange mites in swine. Ivomec sheep drench is a 0.08 % solution of ivermectin for treatment of gastrointestinal roundworms, lungworms, and nasal bots in sheep.

Abamectin [*71751-41-2*], avermectin B_1, is produced by *Streptomyces avermitilis* and is structurally related to ivermectin. It is a mixture containing at least 80 % of avermectin B_{1a} and not more than 20 % of avermectin B_{1b}. Abamectin is marketed as Affirm, Avomec, and Vertimec for use as an anthelmintic drug.

4.2. Benzimidazole Derivatives

Benzimidazole anthelmintics are widely used in cattle, swine, and horses for controlling various gastrointestinal roundworms, tapeworms, liver flukes, and lungworms. They have a wide range of antiparasitic activity, a high degree of efficacy, and a good safety margin. Most are effective when administered orally. Thiabendazole is also used as a fungicide on many crops.

These drugs inhibit microtubules, inhibit glycogen uptake, and inhibit fumarate reductase in susceptible organisms.

Thiabendazole [*148-79-8*] was one of the first benzimidazole anthelmintic drugs approved for use in animals. It is marketed as Equizole A as a liquid horse wormer in combination with piperazine, and as Equizole suspension for control of gastrointestinal roundworms in cattle, horses, sheep, and goats. It is commercially available as TBZ 6 suspension for treatment and control of roundworm infection in cattle, and as Tresaderm in combination with dexamethasone and neomycin sulfate in solution for treatment of various bacterial, mycotic, and dermatoses and otitis externa in dogs and cats.

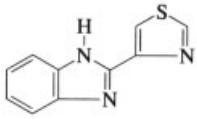

Thiabendazole

Other benzimidazole anthelmintics used in veterinary medicine include albendazole, fenbendazole, cambendazole, oxfendazole, oxibendazole, and mebendazole.

4.3. Imidazothiazole Derivatives

Imidazothiazole drugs are broad-spectrum anthelmintics used to treat and control a number of internal parasites affecting several animal species. They can be administered by various routes and are highly efficacious.

Butamisole [*54400-59-8*] is used to treat whipworm and hookworm infections in dogs. It is administered subcutaneously. It is marketed as Styquin, an aqueous solution of the hydrochloride. The product is not to be administered to heartworm-positive dogs or those treated with bunamidine because fatalities have been reported following its use in these animals.

Levamisole [*16595-80-5*] is an anthelmintic that causes paralysis of nematodes by ganglionic stimulation. It is available in several formulations including parenteral solutions, drenches, boluses, and oral gels. Tramisol injectable solution is a 13.65 % solution of levamisole hydrochloride for subcutaneous injection in cattle, and Tramisol Type A Medicated Article is a premix for addition to the feed of cattle and swine.

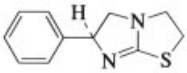

Levamisole

5. Anti-inflammatory Drugs (see also → Anti-inflammatory – Antirheumatic Drugs)

5.1. Nonsteroidal Anti-inflammatory Drugs

The nonsteroidal anti-inflammatory drugs are lipophilic acids of various structures. Ketoprofen and carprofen are members of the propionic acid class, which also includes ibuprofen, naproxen, and fenoprofen. Meclofenamic acid is a member of the fenamate class of nonsteroidal anti-inflammatory drugs.

The nonsteroidal anti-inflammatory drugs produce anti-inflammatory, antipyretic, and analgesic effects in animals.

They are believed to exert their pharmacologic effects at least in part by inhibiting cyclooxygenase activity, thereby reducing concentrations of prostaglandins, thromboxanes, and prostacyclins. They may also act centrally to reduce pain.

Flunixin [*38677-85-9*] is a nonsteroidal anti-inflammatory drug approved for use in horses. It is marketed as Banamine granules containing flunixin as its meglumine salt, equivalent to 250 or 500 mg of flunixin per 10 or 20 g packet for oral administration, for alleviation of pain and inflammation associated with musculoskeletal disorders of horses. Banamine paste contains 1500 mg of flunixin as its meglumine salt per 30 g syringe for oral administration to alleviate pain and inflammation associated with musculoskeletal disorders of horses. Banamine solution contains 50 mg of flunixin as its meglumine salt per milliliter for alleviation of pain and inflammation associated with musculoskeletal disorders and pain associated with colic in horses. Banamine solution may be administered intravenously or intramuscularly.

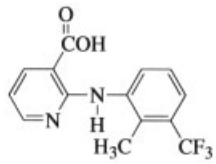

Flunixin

Meclofenamic acid [*644-62-2*] is marketed as Arquel granules containing 500 mg of meclofenamic acid per packet for oral administration to horses to alleviate pain and inflammation associated with musculoskeletal disorders.

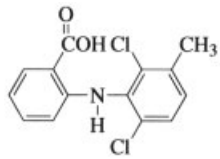

Meclofenamic Acid

Ketoprofen [*22071-15-4*] is a relatively new nonsteroidal anti-inflammatory drug. It is marketed as Ketofen solution containing 100 mg/mL

of ketoprofen for intravenous administration to horses for alleviating pain and inflammation associated with musculoskeletal disorders.

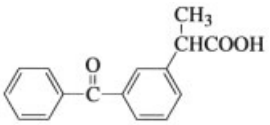

Ketoprofen

Phenylbutazone [*50-33-9*] is the oldest of the nonsteroidal anti-inflammatory drugs used in veterinary medicine. It is marketed in a number of parenteral and oral formulations. Phenylbutazone tablets containing 100 mg and 1000 mg of active ingredient are available for oral administration to dogs and horses, respectively, for treatment of inflammation associated with the musculoskeletal system. Phenylbutazone solution for intravenous injection contains 200 mg/mL of active ingredient and is indicated for the relief of inflammation associated with the musculoskeletal system of the horse.

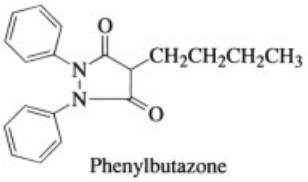

Phenylbutazone

Carprofen is the newest member of this group and is used to treat musculoskeletal injuries in horses.

5.2. Steroidal Anti-inflammatory Drugs

The steroidal anti-inflammatory drugs are fluorinated analogues of cortisol. Fluorine substitution has been shown to increase potency.

They are used locally or systemically to produce an anti-inflammatory effect. They have greater anti-inflammatory activity and less sodium-retaining activity than naturally occurring corticosteroids such as cortisol and cortisone.

The steroidal anti-inflammatory drugs inhibit release of arachidonic acid from cell membranes. They thereby interfere with the first step of the arachidonic acid cascade and prevent or reduce the biosynthesis of prostaglandins, prostacyclins, thromboxanes, and other cytokines involved in the inflammatory response.

Betamethasone [*378-44-9*] is marketed as Betasone, a mixture of betamethasone dipropionate and betamethasone sodium phosphate. The betamethasone dipropionate is in aqueous suspension and the betamethasone sodium phosphate is in solution in this product to provide immediate and sustained anti-inflammatory effects in the dog for relief of pruritis associated with dermatoses. Betamethasone is also available as its valerate ester in combination with gentamicin in a spray (Gentocin) and an ointment (Topagen) for topical use to treat superficial lesions caused by bacteria susceptible to gentamicin in the dog.

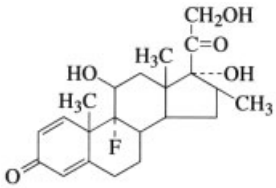

Betamethasone

Dexamethasone [*50-02-2*] is marketed as Azium powder for cattle and horses and as Azium solution for intramuscular and intravenous administration for treating primary bovine ketosis, and as an anti-inflammatory agent in cattle, horses, dogs, and cats. Dexamethasone is also marketed in combination with the diuretic trichlormethiazide as Naquasone for the treatment of physiological parturient udder edema in cattle.

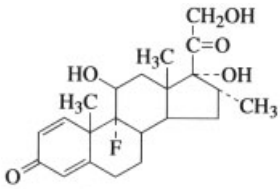

Dexamethasone

Flumethasosone [*2135-17-3*] is marketed in combination with neomycin sulfate and polymyxin B sulfate in Anaprime Ophthalmic Solution for the treatment of inflammation, edema, and secondary bacterial infections of the eye in dogs. It is also marketed as Flucort solution for intravenous, intramuscular, or subcutaneous administration to horses, dogs, and cats, and intra-articular administration to horses. Flucort tablets are available for oral therapy with flumethasone.

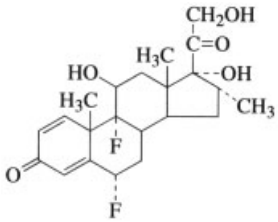

Flumethasone

Isoflupredone [*338-95-4*] is marketed as its acetate ester in combination with neomycin sulfate in Neo-Predef ointment for use in cattle, horses, dogs, and cats, and in combination with both neomycin sulfate and the local anesthetic tetracaine as Neo-Predef, as Tetracaine topical powder, and Tritop topical ointment for use in dogs, cats, and horses. Predef 2 X is a suspension of isoflupredone acetate for intramuscular or intra-synovial use in cattle, horses, and pigs.

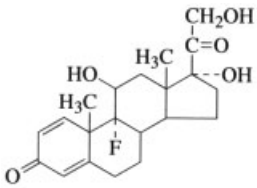

Isoflupredone

Triamcinolone [*124-94-7*] is marketed as its acetonide [*76-25-5*] in Vetalog cream, oral powder, tablets, and parenteral solution.

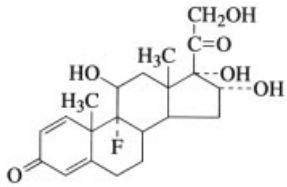

Triamcinolone

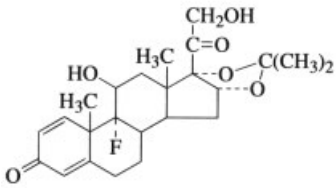

Triamcinolone acetonide

6. Respiratory Stimulants

Doxapram [*309-29-5*] is used as a respiratory stimulant during recovery from barbiturate or inhalation anesthesia in dogs, cats, and horses. It produces respiratory stimulation at doses considerably less than those producing CNS stimulation.

Doxapram stimulates chemoreceptors of the carotid and aortic regions and may stimulate the medullary respiratory center.

Doxapram is marketed as Dopram-V injectable for intravenous use in dogs, cats, and horses.

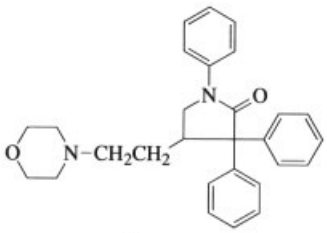

Doxapram

Amiphenazole [*490-55-1*] is a phenyl-substituted thiazolediamine that forms soluble salts with acids.

It is used as a barbiturate and morphine antagonist in animals, and is marketed as Daptizole and Fenamizol.

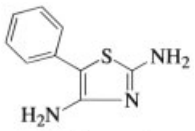

Amiphenazole

7. Anesthetics

Dissociative anesthetics are rapid-acting, nonnarcotic, nonbarbiturate drugs used to produce anesthesia in cats and dogs (tiletamine only) as well as restraint in nonhuman primates (ketamine only). These drugs produce profound analgesia, normal pharyngeal-laryngeal reflexes, as well as mild cardiac stimulation and respiratory depression.

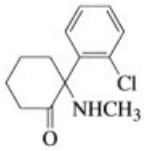

Ketamine

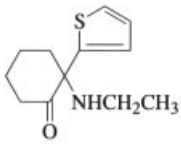

Tiletamine

Ketamine [*6740-88-1*] is marketed as Ketaset and Vetalar solutions for intramuscular administration to cats and subhuman primates.

Tiletamine is marketed as its hydrochloride salt in combination with zolazepam hydrochloride as Telazol solution for intramuscular administration to dogs and cats.

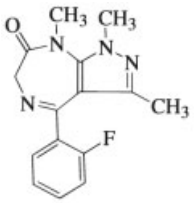

Zolazepam

α_2-Agonists. The α_2-agonists used in veterinary medicine consist of a substituted aromatic ring attached to a heterocyclic ring containing at least one nitrogen atom. Addition of a methyl group to detomidine produced medetomidine and created an asymmetric center.

The α_2-agonists are sedative and analgesic drugs used to facilitate minor surgical procedures.

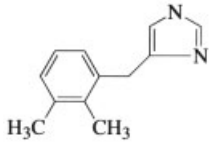

Detomidine

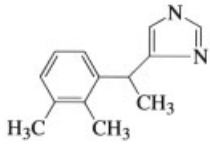

Medetomidine

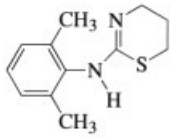

Xylazine

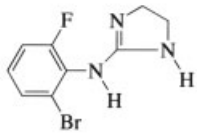

Romifidine

The α_2-adrenoreceptor agonist drugs produce sedation and analgesia via CNS depression. The action of the α_2-adrenoreceptor agonist drugs can be reversed by administration of α_2-adrenoreceptor antagonists such as yohimbine and atipamezole.

Xylazine [*7361-61-7*] is marketed as its hydrochloride salt as Rompun (100 mg/mL) and Anased (20 mg/mL) injectable solutions for intravenous administration to horses and dogs, respectively. The actions of xylazine may be reversed by the administration of yohimbine, an indolalkylamine alkaloid, that blocks those α_2-adrenoreceptors that are stimulated by xylazine.

Detomidine [*76631-46-4*] is marketed as its hydrochloride salt as Dormosedan solution for intramuscular or intravenous administration to horses.

Medetomidine [*86347-14-0*] is marketed as its hydrochloride salt as Domitor, and its dextrorotatory isomer, *dexmedetomidine*, are used to produce sedation and analgesia in many mammalian species but used mainly in the dog and domestic cat.

Romifidine is marketed as its hydrochloride salt.

Reversing Agents. *Yohimbine* [*146-48-5*] is an alkaloid found in *Corynanthe johimbe* and related trees as well as in *Raulwolfia serpentina*. It is marketed as its hydrochloride salt as Yobine to reverse the effects of the α_2-adrenoreceptor agonists.

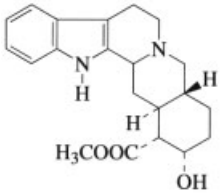

Yohimbine

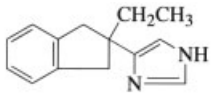

Atipamezole

Atipamezole is a synthetic compound marketed as Antisedan to reverse the effects of the α_2-adrenoreceptor agonists.

8. Narcotic Agonists and Antagonists

Narcotic analgesic drugs used in veterinary medicine are synthetic or semisynthetic derivatives of morphine (and related alkaloids) and derivatives of fentanyl.

Narcotic analgesics are used in veterinary medicine to produce tranquilization to assist in the capture of wild and zoo animals, as preanesthetic medications, and as analgesic drugs.

They are believed to exert their effects by acting as agonists at specific opiate receptors located in the brain and possibly other sites in the body. Stimulation of these receptors produces profound analgesia in animals. The actions of the narcotic analgesic drugs may be reversed by the administration of competitive antagonists such as diprenorphine and naloxone.

Numerous narcotic analgesic drugs such as morphine, codeine, oxymorphone (Numorphan), pentazocine (Talwin-V), and butorphanol (Torbugesic and Torbitrol) are used in veterinary medicine. Other narcotics with somewhat specialized application in veterinary medicine are discussed below.

Etorphine/diprenorphine is an agonist/antagonist combination used to produce tranquilization and reversal in large animals. Etorphine [*14521-96-1*] is marketed as its hydrochloride salt as M-99 for intramuscular administration. It is approximately 10 000 times more potent than morphine as an analgesic. It is usually dosed heavily to produce tranquilization and avoid hyperexcitability in wild animals, and then reversed with diprenorphine [*14357-78-9*] (M-5050). Etorphine is also marketed in combination with acepromazine, a phenothiazine tranquilizer, to produce neuroleptic analgesia.

Fentanyl [*437-38-7*] is marketed as its citrate salt in combination with droperidol, a butryphenone tranquilizer, as Innovar-Vet for intravenous or intramuscular use in the dog for various surgical procedures requiring analgesia and tranquilization.

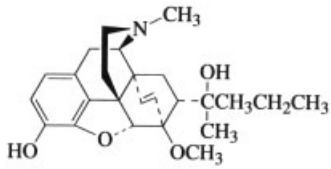

Etorphine

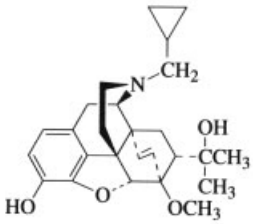

Diprenorphine

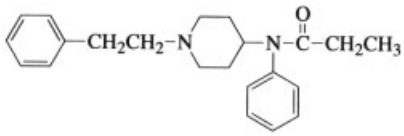

Fentanyl

Carfentanil is marketed as its citrate salt as Wildnil for immobilization of free-ranging or confined deer, elk, and moose. Carfentanil is ca. 10 000 times as potent as morphine. The effect of carfentanil is rapidly reversed by the administration of diprenorphine.

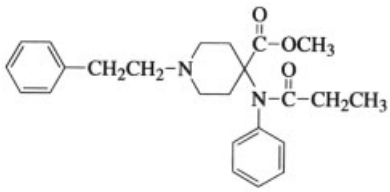

Carfentanil

9. Antihistamines (see also → Antiallergic Agents, Chap. 2.)

The antihistamines used in veterinary medicine are classified as ethylenediamine derivatives (i.e., pyrilamine and tripelennamine), ethanolamine derivatives (i.e., doxylamine), and phenothiazine derivatives (i.e., trimeprazine).

Antihistamines antagonize many of the pharmacological effects of histamine by acting as competitive inhibitors of histamine H_1 receptors.

Pyrilamine [*91-84-9*] is marketed as Histavet-P containing 20 mg/mL of pyrilamine maleate and is indicated for use in the horse. The formulation may be administered intravenously, intramuscularly, or subcutaneously.

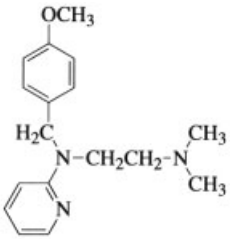

Pyrilamine

Doxylamine [*496-21-6*] is marketed as A-H Injection containing 11.36 mg/mL of doxylamine succinate for use as an antihistamine in horses, dogs, and cats. It may be administered intravenously to horses and intramuscularly or subcutaneously to dogs, cats, and horses. A-H Tablets contain 100 mg of doxylamine succinate for use as an antihistamine in horses.

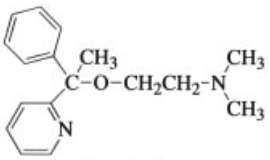

Doxylamine

Tripelennamine [*91-81-6*] is marketed as Re-Covr Injection containing 20 mg/mL of tripelennamine hydrochloride for use as an antihistamine in cattle, horses, dogs, and cats. It may be administered intravenously to cattle and intramuscularly to cattle, horses, dogs, and cats.

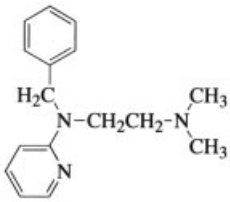

Tripelennamine

Trimeprazine [*84-96-8*] is a phenothiazine derivative marketed as Temaril-P tablets containing 5 mg of trimeprazine tartrate and 2 mg of prednisolone. The drug is used for the antipruritic and antitussive effects of trimeprazine and the anti-inflammatory effects of prednisolone in dogs.

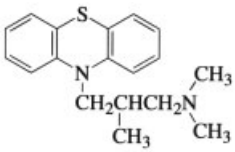

Trimeprazine

10. References

General References

1. M. Sittig: *Veterinary Drug Manufacturing Encyclopedia,* Noyes Publications, Park Ridge, NJ, 1981.
2. H. R. Adams: *Veterinary Pharmacology and Therapeutics,* 5th ed., The Iowa State University Press, Ames, IA, 1995.
3. K. Bennett (ed.): *Compendium of Veterinary Products,* 1st ed., North American Compendiums, Inc., Port Huron, MI, 1991.
4. R. B. Talbot (ed.): *Veterinary Pharmaceuticals and Biologicals,* 8th ed., Veterinary Medicine Publishing Company, Lenexa, KS, 1993/1994.

Vinegar

Heinrich Ebner, Linz, Austria

Heinrich Follmann, Heinrich Frings GmbH & Co KG, Bonn, Federal Republic of Germany

Sylvia Sellmer, Heinrich Frings GmbH & Co KG, Bonn, Federal Republic of Germany

1. Introduction 2
2. Physical and Chemical Properties 2
3. Acetic Acid Fermentation 2
4. Raw Materials 2
4.1. Alcohol 3
4.2. Water 3
4.3. Nutrients 3
5. Production 3
5.1. Microorganisms 3
5.1.1. Summary and Classification 3
5.1.2. Required Properties of Industrially Used Strains 4
5.1.3. Industrially Used Strains 4
5.1.4. Genetic Improvement 5
5.1.5. Bacteriophages 5
5.2. Biochemistry 5
5.2.1. Ethanol 5
5.2.2. Sugar 6
5.2.3. Acetate 6
5.2.4. Carbon Dioxide 6
5.2.5. Nitrogen 6
5.2.6. Growth Factors 6
5.3. Physiology 6
5.3.1. Oxygen Demand and Total Concentration 6
5.3.2. Lack of Ethanol 7
5.3.3. Specific Growth Rate 7
5.3.4. Specific Product Formation 7
5.3.5. Concentration and Temperature . . 8
5.3.6. Overoxidation 8
5.4. Industrial Processes 8
5.4.1. Submerged Vinegar Fermentation . 8
5.4.1.1. Semicontinuous and Continuous Fermentations 8
5.4.1.2. Two-Stage Processes 8
5.4.1.3. The Frings Acetator 9
5.4.1.4. Further Trends in Submerged Processes 10
5.4.1.5. Other Submerged Fermentation Processes 10
5.4.1.6. Continuous Culture and Cell Recycle 10
5.4.1.7. Semicontinuous Culture and Cell Recycle 11
5.4.2. Surface Fermentation 12
5.4.2.1. History 12
5.4.2.2. New Experiments with Surface Fermentation 12
5.4.3. Trickle Processes 12
5.4.4. Experiments with Immobilized Acetic Acid Bacteria 12
5.4.5. Environmental Protection 13
5.5. Treatment of Raw Vinegar 13
5.5.1. Storage and Quality 13
5.5.2. Filtration 13
5.5.3. Pasteurization 14
5.5.4. Bottling and Shelf Life 14
6. Constituents of Vinegar 14
7. Analysis of Vinegar 15
8. Use of Vinegar 15
9. Economic Aspects 15
10. References 15

1. Introduction

Legal Definition in the European Union. Vinegar is the product obtained from alcoholic and subsequent acetous fermentation of liquors of agricultural origin (ethanol, wine, cider, perry, or liquors obtained from cereals, including malted barley).

Legal Definition in the United States. In the United States, this definition is extended to include synthetically produced alcohol used for acetous fermentation.

Legal Definition in Japan. Brewed vinegar is a liquid condiment obtained by the acetous fermentation of mash prepared from grains or fruits where alcohol or saccharides may be added.

Use of Synthetic Acetic Acid. Although in Germany a product which has been produced by dilution of acetic acid with water may be called

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_403

vinegar, the use of diluted acetic acid for human consumption must be declared as such in all other countries.

2. Physical and Chemical Properties

Vinegar is a clear aqueous liquid, which may be colorless, the color of its raw material, or colored by caramel; the prescribed acetic acid content is 40 – 150 g/L. Further constituents derive from the raw material, added nutrients, or the water used for dilution. Depending on the acetic acid content and the raw material, density, boiling point, freezing point, surface tension, and viscosity differ more or less from the values for pure water; pH is 2.0 – 3.5.

3. Acetic Acid Fermentation

Acetic acid fermentation (acetous fermentation) is an oxidative fermentation in which dilute solutions of ethanol are oxidized by acetic acid bacteria with oxygen from the air to acetic acid and water. Oxidation proceeds according to the basic equation:

$$C_2H_5OH + O_2 \rightarrow CH_3COOH + H_2O$$

$$\Delta G^0 = -455 \text{ kJ/mol}$$

The alcohol-containing solution is called "mash". Its alcohol concentration is given as a percentage by volume. Usually, it also contains some acetic acid, expressed in grams acetic acid per 100 mL (wt/vol %). The sum of ethanol vol % and acetic acid wt/vol % is called the "total concentration" because the sum of these rather incommensurable values gives the maximal concentration of acetic acid which can be obtained by complete fermentation. For details of this curious, but commonly used calculation, see reference [1]. The quotient of the total concentration of the vinegar produced over the total concentration of the mash gives the concentration yield.

4. Raw Materials

All mashes must contain ethanol, water, and nutrients for the acetic acid bacteria.

4.1. Alcohol

By far the largest percentage of vinegar is alcohol vinegar, produced from diluted purified ethanol; other common names are white vinegar or spirit vinegar. It is customary in most countries to denature the ethanol which serves as raw material for the vinegar industry. In most European countries, denaturation is carried out with alcohol vinegar. In the United States it is mostly done with ethyl acetate, which is split into ethanol and acetic acid during fermentation.

Mashes obtained by alcoholic fermentation of natural sugar-containing liquids also serve as raw material. The vinegar is designated according to the particular raw material used. Wine vinegar is produced by acetic acid fermentation of grape wine. Cider vinegar is produced from fermented apple juice. Malt vinegar is made by alcoholic and subsequent acetous fermentation, without distillation, of an infusion of barley malt with or without addition of other cereals, the starch of which has been saccharified exclusively by malted barley diastase. Rice vinegar is made by saccharification of rice starch, followed by alcoholic and acetous fermentation.

4.2. Water

The water used for the preparation of mashes must be clear, colorless, odorless, bacteriologically clean, and without sediments or suspended particles. It must also be free of chlorine, ozone, and other compounds which damage bacteria.

4.3. Nutrients

Most natural raw materials do not require the addition of extra nutrients, but apple cider is usually low in nitrogenous materials. This can be corrected by addition of ammonium phosphate. Some grape wines also require addition of ammonium phosphate for satisfactory fermentation. In rare cases, the same nutrients must be added as for the production of alcohol vinegar.

For the production of alcohol vinegar, a mixture of the required nutrients has been developed. For vinegar with $\leq$15 % acetic acid, acetic acid bacteria require glucose and potassium, sodium, magnesium, calcium, ammonium (as

phosphate, sulfate, and chloride). The following trace minerals are needed: iron, manganese, cobalt, copper, molybdenum, vanadium, and zinc. These additions are sufficient for satisfactory fermentation.

Commercially available mixtures contain supplements, e.g., dried yeast extract, in order to restart fermentation more quickly should it be stopped, e.g., by a power failure. Nutrients should be added sparingly in order to exert a selection pressure on the bacteria.

5. Production

5.1. Microorganisms

5.1.1. Summary and Classification

The microorganisms which oxidize ethanol to acetic acid are commonly called acetic acid bacteria. This special primary microbial metabolism at low pH differentiates them from all other bacteria. Acetic acid bacteria are polymorphous. Cells are gram-negative, ellipsoidal to rod-shaped, straight or slightly curved, 0.6–0.8 µm, occurring singly, in pairs, or chains. There are nonmotile and motile forms, with polar or peritrichous flagella. They are obligately aerobic; some produce pigments, some cellulose.

Attempts to classify acetic acid bacteria up to 1980 have been summarized [1]. At that time, a major contribution to their taxonomy was made by DNA–rRNA hybridization studies [2]. An early conclusion was that the two groups *Gluconobacter* and *Acetobacter*, as classified earlier by DE LEY [3], were closely related, justifying their union in the family Acetobacteraceae, clearly distinguishable as a separate branch in superfamily IV (the α-subclass [4]). Although *Bergey's Manual* [5] gives a set of 32 features to differentiate the genera *Gluconobacter, Acetobacter,* and *Frateuria*, not all of these features are necessary for satisfactory identification.

The characteristic of not overoxidizing acetate identifies the strains from "high-acid" fermentations as members of the genus *Acetobacter*, and thus separates them from the genera *Gluconobacter* and *Frateuria*. SWINGS [6] divided *Acetobacter* into seven species. GILLIS et al. [7] described a species capable of fixing N_2 microaerobically, called *Acetobacter diazotrophicus*.

5.1.2. Required Properties of Industrially Used Strains

Considering the interest that acetic acid bacteria have inspired in microbiologists, it is astonishing how little effect the available information about these bacteria has had on vinegar manufacture. "Working" strains were isolated, but lost again, demonstrating the difficulties of isolating and growing *Acetobacter* strains from industrial vinegar fermentations on solid media.

Considerable progress has been achieved in Japan by propagating bacteria from fermenters on a special double-layer agar. The bacteria isolated on this medium have been described as *Acetobacter polyoxogenes* [8]. However, *A. polyoxogenes* was not available from the Japan Collection of Microorganisms, owing to propagation and preservation problems.

A new species, for which SIEVERS et al. [9] proposed the name *Acetobacter europaeus*, has been isolated and characterized in pure culture from vinegar fermentations at high acidity in Germany and Switzerland. All investigated strains isolated from submerged fermenters and trickle generators had very low (0–22 %) DNA–DNA similarities with traditional strains of *Acetobacter* and *Gluconobacter*. Phenotypic differentiation at species level is rather difficult for *Acetobacter*, since different strains of a single species do not necessarily utilize the same carbon source. A useful criterion for identification of *Acetobacter europaeus* from other *Acetobacter* species is the strong tolerance to the 4–8 % acetic acid contained in AE agar, and the absolute requirement of acetic acid for growth.

Independent of all these problems, the vinegar industry's first requirement is a strain of acetic acid bacteria which: tolerates high concentrations of acetic acid and total concentration; requires small amounts of nutrients; does not overoxidize the acetic acid formed; and yields high production rates. It seems that it is also necessary to make this strain phage resistant [10].

5.1.3. Industrially Used Strains

The vinegar industry has traditionally worked with acetic acid bacteria which are not derived

from pure cultures. The astonishing fact that common microorganisms used on a large scale for industrial vinegar production have not been properly described and characterized in taxonomic terms is due to the difficulty of cultivating them on semisolid media.

Industrial submerged vinegar fermentations are started by treatment with "inoculation vinegar" (microbiologically undefined residues from previous fermentations). Acetic acid fermentation can be carried out for years without interruption or decrease in efficiency or yield, if suitable conditions are chosen to allow a permanent selection of strains which tolerate high acidity and can live on a minimum of nutrients.

It has long been known that the properties of a newly isolated strain of acetic acid bacteria may change from the very first moment of cultivation in Petri dishes, and that this strain, if cultivated over a number of generations, may show other properties, especially in its adaptation to certain concentrations of acetic acid, but also in its phenotypic features.

Investigations using plasmid profile analysis to characterize and identify acetic acid bacteria derived from vinegar fermentations failed to verify this variability.

Plasmid Profile Analysis. Since DAVIES et al. [11] first used plasmid profile analysis for identification of lactic acid bacteria, this method has become a powerful tool to identify and characterize strains of all genera. Plasmids are circular, extrachromosomal DNA molecules, which can be extracted from bacteria. The plasmid profile of one strain is very specific. Plasmids can determine certain properties of bacteria (phage resistance or citrate metabolism).

The potential of this technique for vinegar fermentations has been investigated [12–14]. The majority of acetic acid bacteria contain 1 – 8 plasmids of molecular mass $(1-17)\times 10^6$ or more. SIEVERS et al. [15] used plasmid profile analysis to characterize *Acetobacter europaeus* strains and microflora in industrial vinegar plants. The microflora in submerged fermentations consists of one predominant strain, whereas in trickle generators the microflora is quite heterogeneous, containing a mixture of several *Acetobacter europaeus* strains with quite different plasmid profiles. This is probably the reason for the susceptibility of submerged fermentations to bacteriophage attack. The determination of plasmid profiles from submerged vinegar fermenters represents a new and powerful technique for characterizing the microflora, and answering questions, e.g., on the stability, origin, and identity of the prevailing microflora.

Investigations in the last years showed that the plasmid profiles of these strains can be stable for at least 5 a, and do not change their phenotypic features during this time.

Fermentations at high total concentrations are carried out continuously in the laboratories of Heinrich Frings, Germany, to increase acetic acid tolerance of the strains. Cultures are supplied to vinegar factories in all parts of the world.

In 1993 Frings established a Phage and Plasmid Laboratory for further investigations.

5.1.4. Genetic Improvement

To improve strains of *Acetobacter* genetically, recombinant DNA techniques are considered useful. Host – vector systems and an efficient transformation method for *Acetobacter* have been developed. Preparation of spheroplasts and development of a spheroplast fusion method have been described [16]. A strain growing at 37 °C in 4 % acetic acid and a strain unable to grow at 35 °C in 5 % acetic acid were spheroplasted and fused. Fusants were able to grow at 37 °C on agar plates containing 5 % acetic acid, on which both parent strains were unable to grow.

Genetic transformation systems have been described for *Acetobacter xylinum* [17], *Acetobacter aceti* [18], *Acetobacter* and *Gluconobacter* strains [13], [19–21].

5.1.5. Bacteriophages

Acetobacter -specific bacteriophages demonstrated by electron microscopy in submerged and trickle fermenters have been suspected of causing fermentation problems in industrial vinegar fermenters [14], [22–24]. In trickle generators, fermentation is sometimes slowed down, but rarely stops completely. In submerged fermentations, complete loss of productivity has been observed. Only three morphologically different *Acetobacter* phage types had been described until 1992, when SELLMER et al. [23]

demonstrated a great variety in phage types, regarding the dimensions of the phage heads and tails. All phages isolated from vinegar fermentations in Germany, Austria, and Denmark showed long contractile tails and isometric heads belonging to Bradley's group A [25] and the family Myoviridae [26], respectively. Although not all phages described have actually been shown to be infective for defined *Acetobacter* strains, owing to lack of a suitable technique, their occurrence in high numbers ($10^6 - 10^9$ per milliliter) in disturbed fermentations suggests that bacteriophages may actually be the source of fermentation problems.

5.2. Biochemistry

During acetic acid fermentation, ethanol is almost quantitatively oxidized to acetic acid. Concentration yields of 95 – 98 % are common; the remainder is mainly lost in the effluent gas. At the same time a suitable C_6 carbon source (preferably glucose) is oxidized; end products are CO_2 and H_2O.

5.2.1. Ethanol

Ethanol is oxidized by two sequential reactions catalyzed by alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH), which can be located at the outer surface of the cytoplasmic membrane, functionally linked to the respiratory chain of the organism [27], [28]. Alcohol dehydrogenase has been purified from several acetic acid bacteria and shown to contain pyrroloquinoline quinone (PQQ) as well as heme c as prosthetic groups [29], [30]. Extensive purification and characterization of the enzymes involved in ethanol oxidation of *Acetobacter aceti, A. rancens,* and *Gluconobacter suboxydans* have been reported [29–35]. ADH of acetic acid bacteria donates electrons directly to ubiquinone in the cytoplasmic membranes. Thus, the ethanol oxidase respiratory chain of acetic acid bacteria consists of only three membranous respiratory components, alcohol dehydrogenase, ubiquinone, and terminal oxidase [36].

Ameyama et al. [33] confirmed that ADH and ALDH of acetic acid bacteria have the same prosthetic group, PQQ, as glucose dehydrogenase of the bacteria. These quinoproteins – glucose, alcohol, and aldehyde dehydrogenases – are involved in sugar oxidation linked to an electron transport system in the membrane of oxidative bacteria. Fukaya et al. [37] and Tayama et al. [38] purified the ADH and the ALDH from *A. polyoxogenes* sp. nov. This strain was isolated from a vinegar factory; it had higher acid productivity and tolerance to acetic acid than strains isolated from natural sources. In contrast to the ADH and ALDH obtained from *A. aceti* and *G. suboxydans*, the enzymes of *A. polyoxogenes* were quite stable, and this high stability seems at least partially involved in the ability of *A. polyoxogenes* to produce high concentrations of acetic acid.

5.2.2. Sugar

For the breakdown of sugars, *Acetobacter* is equipped with the hexose monophosphate (HMP) pathway and the tricarboxylic acid (TCA) cycle [39], whereas glycolysis is either absent or very weak [40]. The enzymes of the Entner – Doudoroff pathway are present in *Gluconobacter* spp. and in *Acetobacter xylinum* [41]. *Gluconobacter* and *Acetobacter* are known for their direct oxidative capacity on sugars, alcohols, and steroids [42]. For further details see reference [1]. The biochemistry of the ketogenic activities of acetic acid bacteria have been reviewed [43].

5.2.3. Acetate

Strains of *Acetobacter* are able to oxidize acetate as well as lactate in both the presence and absence of ethanol by using enzymes of the TCA cycle.

5.2.4. Carbon Dioxide

It has been known since 1952 that *Acetobacter* needs CO_2 for growth [44]. CO_2 is incorporated into cell substances, where ca. 0.1 % of the cell carbon is derived from CO_2. A very small, but measurable portion of acetic acid is derived from CO_2 metabolism [45].

5.2.5. Nitrogen

If an organic carbon source is present, a number of acetic acid bacteria can use ammonium ions as the sole nitrogen source. Other strains need the presence of amino acids, others need cofactors, e.g., vitamins or purines [46]. No essential amino acids are known.

5.2.6. Growth Factors

The growth factor requirement depends on the carbon source [47]. Some strains require *p*-aminobenzoic acid, niacin, thiamin, or pantothenic acid as growth factors [48]. Ameyama and Kondo [49] found that some *Acetobacter* strains did not need vitamins in the presence of glucose. But the combined addition of glutathione and sodium glutamate had a cumulative effect on the growth of *A. aceti* [50].

The novel prosthetic group in acetic acid bacteria, PQQ, also shows an interesting growth stimulating effect on various microorganisms [50–52]. There are two PQQ effects:

1) Stimulation of both growth rate and total cell yield with only a trace amount of PQQ as an essential growth factor
2) A marked reduction of the lag phase in microbial growth, but with no increase in growth rate in the subsequent exponential phase, or in total cell yield in the stationary phase

PQQ is produced in culture media by most strains of microorganisms.

5.3. Physiology

5.3.1. Oxygen Demand and Total Concentration

During commercial use of many aerobic microorganisms, it is common to permit the inoculum to remain in the fermenter for several days without aeration, before inoculation into the main fermenter. However, this procedure is unthinkable in the case of acetic acid fermentation.

Extensive tests have been conducted [53], [54] on the damage to acetic acid bacteria by interruption of aeration as a function of total concentration, acetic acid concentration, rate of fermentation, and length of the interruption. The longer the interruption of aeration and the higher the total concentration of the mash, the more marked the effect. At total concentration 5 %, an interruption of aeration for 2 – 8 min leads to the same damage as an interruption of 15 – 60 s at total concentration 11 – 12 %. At constant total concentrations, damage increases with increasing concentration of acetic acid, and with increasing fermentation rate. The specific production of cell material is extremely low. For more details see reference [1].

Muraoka at al. [55] confirmed these results. There was a sharp drop in ethanol oxidation rate as well as in the activity of the enzymes involved (ADH and ALDH) after an interruption of the oxygen supply, with acetic acid concentration 6 %. The ADH activity in the membrane fraction was highly stable in culture broth containing $\leq$4 % acetic acid, but the purified enzyme was very unstable, even in a culture broth of 2 % total concentration. It was suggested that ADH has a membrane component contributing to its stability up to 4 % acetic acid, which becomes gradually liable to damage with further increasing acidity. The relation of the ADH system to oxygen has not been elucidated.

Under conditions of industrial acetic acid production, the microorganisms undergo considerable stress, caused by a high concentration of acetic acid and limited oxygen supply. A high utilization of oxygen, up to 70 %, can be achieved in industrial production. A sparing use of air is very important, because of the volatility of ethanol and acetic acid. However, the use of pure oxygen or highly oxygen-enriched air may damage the bacteria [53].

Hitschmann and Stockinger [56] found that, at a total concentration of 13 %, ATP content of the cell and growth show reverse behavior. The energy charge (average value of intracellular ATP, ADP, and AMP) showed a rather high value of 0.84 at the beginning of fermentation. During fermentation, while the acetic acid concentration increased from 9 to 12 %, the energy charge decreased to 0.7. After 45 s interruption of aeration, the energy charge dropped to 0.58, and after several weeks of storage, the charge of the inoculum was 0.50.

5.3.2. Lack of Ethanol

Acetic acid bacteria are also damaged if vinegar fermentation is carried out to the point where all the ethanol has been oxidized, and the addition of fresh ethanol-containing mash is delayed. This is analogous to the damage resulting from interruption of the oxygen supply, and also depends primarily on the total concentration and the duration of the ethanol shortage.

5.3.3. Specific Growth Rate

Hromatka et al. [57] were the first to calculate the specific growth rate of acetic acid bacteria. In semicontinuous fermentations, the specific growth rate did not decrease at increasing acetic acid concentrations, but at increasing total concentrations it decreased from $0.49\,h^{-1}$ at 5 % to $0.16\,h^{-1}$ at 11 %.

In a continuous culture at total concentration 12 %, the specific growth rate decreased from $0.027\,h^{-1}$ at 4.5 % ethanol to $0.006\,h^{-1}$ at 1 % ethanol. The specific growth rates obtained in continuous fermentations with constant acetic acid and ethanol concentrations are less favorable than those obtained under semicontinuous conditions.

Vera and Wang [58] carried out batch fermentations with *Acetobacter suboxydans*, and showed, at a total concentration of 7.5 %, a specific growth rate of $0.30\,h^{-1}$, in good agreement with the findings of Hromatka et al. [57].

5.3.4. Specific Product Formation

Hromatka and Ebner [59] found a specific product formation of 21 g acetic acid per gram bacterial dry substance per hour in semicontinuous tests at total concentration 10 %. This very high value was independent of acetic acid concentration from 4.5 to 7.2 %. Mori and Terui [60] found a strong dependence of specific product formation on acetic acid concentration in continuous fermentations at total concentration 7 %. The maximum specific product formation was $15\,g\,g^{-1}\,h^{-1}$. Vera and Wang [58] found that specific product formation follows a growth-associated mechanism as long as the acetic acid concentration is <3 %. At higher acetic acid concentrations, up to 7 %, an additional non-growth-associated term must be considered. The highest value obtained with cell recycle in continuous culture of *A. suboxydans* was $11.5\,g\,g^{-1}\,h^{-1}$.

5.3.5. Concentration and Temperature

When a fermentation is charged with mash, this must be done at the same total concentration under rapid mixing, because a locally high concentration gradient damages the bacteria. The next cycle starts with no lag phase. The temperature has to be kept constant because a repeated change in temperature causes a similar result [1].

5.3.6. Overoxidation

Overoxidation is the undesirable oxidation of acetic acid to CO_2 and H_2O. To prevent it, lack of ethanol must be avoided, and the total concentration should be kept high. In trickle fermenters, this is much more difficult than in submerged fermenters [1].

5.4. Industrial Processes

5.4.1. Submerged Vinegar Fermentation

5.4.1.1. Semicontinuous and Continuous Fermentations

For production of vinegar with 12 – 15 % acetic acid, the process has to be be carried out in a semicontinuous manner [61], fermentation cycles taking about equal amounts of time. The starting concentration of each cycle is 7 – 10 % acetic acid and ca. 5 % ethanol. When an alcohol concentration of 0.05 – 0.3 % has been reached in the fermenting liquid, a quantity of vinegar is discharged from the fermenter, so that refilling with new mash of 0 – 2 % acetic acid and 12 – 15 % alcohol restores the same starting concentrations for the new cycle. Discharging must be carried out quickly to avoid complete depletion of alcohol. Charging must be slow, under constant fermentation temperature and rapid mixing.

By the end of 1993 it was possible to produce vinegar of up to 17.5 % acetic acid in a modified

single-stage process. Starting with similar concentrations as above, alcohol is slowly added at a constant concentration of 2 – 3 % alcohol during a corresponding increase of total concentration. When alcohol approaches zero, part of the fermenter contents is discharged and replaced by mash with lower total concentration, to bring acetic acid and total concentration back to starting conditions, enabling the bacteria to multiply faster. Later, total concentration is increased again, by adding alcohol.

Continuous fermentation is possible up to 9 –10 % acetic acid only because the specific growth rate of the bacteria decreases with decreasing ethanol concentration; to obtain a high yield, fermentation must be carried out at low alcohol concentration.

5.4.1.2. Two-Stage Processes

The canning industry requires vinegar with a very high percentage of acetic acid. The vinegar industry is interested in producing vinegar of high acidity to save storage and transport costs. In 1978, a two-stage process for the production of vinegar with >15 % acetic acid was developed [62]. During submerged fermentation with total concentration <15 % in a first fermenter, alcohol is added slowly to increase the total concentration to ca. 18.5 %. When the acetic acid concentration has reached 15 %, ca. 30 % of the fermenting liquid is transferred to a second fermenter. The first fermenter is resupplied with new mash of lower total concentration. In the second fermenter, fermentation continues until the alcohol is almost depleted. The whole quantity of the finished vinegar is discharged. The fermenting liquid in the first fermenter is supplied with alcohol at the appropriate time, and later divided again.

Vinegar of >20 % acetic acid can be produced by this process. Automatic control for the process has been described [63].

In 1980, a similar two-stage process for the production of vinegar with >20 % acetic acid was described [64], [65], where the second fermentation stage is carried out at only 18 – 24 °C, while the first fermentation stage uses the normal temperature range of 27 – 32 °C.

Maselli and Horwath [66] reported a similar two-stage process with semicontinuous fermentation in at least two fermenters and membrane filtration of the mash from the semicontinuous stage, after combining it with an oxygen-containing gas. After the addition of ethanol and nutrients, the concentrated mash with the bacteria is subjected to batch fermentation at a substantially lower temperature until 18 – 20 % acetic acid is reached. The membranefiltered vinegar is used as the second end product.

5.4.1.3. The Frings Acetator

The Frings Acetator is the most common equipment for the production of all kinds of vinegar. At the end of 1993, >600 acetators were in operation worldwide, with total annual production 1354×10^6 L vinegar with 10 % acetic acid. Energy consumption is ca. 400 W per liter fermented ethanol, and the yield is up to 98 %.

Commercial acetators ferment up to 3600 L pure ethanol in 24 h.

The different processes demand excellent performance of the fermenter, especially of its aeration system.

The Frings Aerator [67] is self-aspirating; no compressed air is needed. The rotor is installed on the shaft of a motor mounted under the fermenter, connected to an air suction pipe, and surrounded by a stator. It sucks air and pumps liquid, creating an air – liquid emulsion which is ejected through the stator, radially outward at a given speed, chosen so that the turbulence of the stream causes a uniform distribution of the air over the whole cross section of the fermenter.

As foam formation cannot always be avoided, the Acetator is equipped with a mechanical defoamer. From 1966 to 1991, the equipment described by Ebner [68] was used. The rotor, turning in a spiral housing, separates liquid and gas, and returns the liquid to the lower part of the fermenter. Only during discharging of the finished vinegar is part of this returned liquid pumped out of the fermenter. As this liquid causes foam formation, the defoamer must operate during most of the fermentation cycle. Wittler [69] described an improvement where the separated liquid portion is not pumped back into the fermenter, but into a small collecting vessel, to be later admixed with the vinegar end product. Lysis of harmed vinegar bacteria releases foam, forming surface-active substances. By total elimination of the separated liquid, further

foam buildup is avoided. Minimal foam formation with shorter operating time of the defoamer, and consequently lower power uptake and negligibly small volume of separated liquid are the result.

The Frings Alkograph [70], an automatic instrument for measuring the amount of ethanol in the fermenting liquid, has been in use since 1965. Small amounts of liquid flow continuously through the analyzer, first through a heating vessel, and then through three boiling vessels. The boiling temperature of the incoming liquid is measured in the first boiling vessel. While alcohol is distilled off continuously from the second and third boiling vessels, the higher boiling point of the liquid from which ethanol has been removed is measured in the third boiling vessel. The difference in temperature corresponds to the ethanol content, and is recorded automatically. As flow through the vessels takes some time, there is a delay of about 15 min between the beginning of sample inflow and the time when the right value is shown on the recorder. As alcohol concentration decreases slowly during fermentation, this delay has no adverse effect on fermentation control.

To carry out the single-stage semicontinuous process, a contact in the Alkograph activates the vinegar discharge pump when a defined alcohol content is reached. As soon as a preset level has been reached, the mash pump starts adding fresh mash. This pump is controlled by the fermentation temperature, in order to refill at constant temperature. The pump is stopped when the desired level is reached, and an automatic cooling system is activated. A fermentation cycle takes 24 – 48 h.

The Alkograph has worked with great reliability for ca. 30 a in 231 acetators around the world. New biosensors, together with modern membranes, have permitted the development of a device to measure alcohol with no time delay – the Frings Alkosens. Heider et al. [71] described a device for measuring volatile constituents of the fermentation medium. A membrane serves to filter them from the fermenting liquid into a carrier medium, and an electronic sensor responds to these volatile constituents. As the measurement relies on diffusion, ambient temperature has to be maintained at the same level. Thus, the alcohol probe is situated in the fermentation broth. The alcohol value shown at the analyzer is influenced by the concentration of acetic acid, and new calibrations are needed at different total concentrations. This disadvantage has been overcome by the Frings Alkosens II, which incorporates a multilayer membrane. New sensor technologies and more sophisticated computer equipment permit fully automated process control of continuous, semicontinuous, as well as single- or two-stage high-strength fermentations, including preparation of the mash.

5.4.1.4. Further Trends in Submerged Processes

Hekmat et al. [72] applied a computerized experimental system to submerged acetic acid fermentation with industrial *Acetobacter* strains at ca. 10 % final acidity, using automated semicontinuous cycles over a period of 1 a. Based on these data, a mathematical model was developed, which could predict the time course of a cycle, even if a lag phase occurred at the beginning of the cycle, owing to a lack of substrate at the end of the previous cycle. The parameters substrate, product, and biomass are not sufficient to characterize the physiological status of the culture, since the behavior of the microorganisms also depends on their history. The RNA concentration per unit of cell mass was found to be a key internal component of the microorganisms, permitting reasonable agreement between the experimental data and the mathematical model.

Mori [73] summarized the possibilities for on-line measurement of alcohol concentration in vinegar production – the Frings Alkograph and Alcolyzer, a Japanese instrument based on the same principle, immobilized microbial sensors, immobilized enzyme sensors, and semiconductor alcohol sensors.

Nomura et al. [74–76] reported experiments on the continuous production of acetic acid by submerged fermentation with continuous removal of acetic acid by electrodialysis, continuous supply of ethanol and concentrated nutrient solution, with control of ethanol and pH. The ethanol concentration was kept constant at 1 %, acetic acid concentration ca. 2 %. Without dialysis and pH control, acetic acid production stopped after 2 d around 4 %, while with dialysis it lasted 30 d.

5.4.1.5. Other Submerged Fermentation Processes

Other processes of little practical significance and abandoned processes are described in reference [1].

5.4.1.6. Continuous Culture and Cell Recycle

Much work has been invested in increasing the productivity of acetic acid fermentation by continuous culture and cell recycle [58], [77].

At first sight, productivity in an acetator does not seem very high, e.g., in a fermenting volume of 100 m^3 4000 kg acetic acid in 24 h at a concentration up to 15 % (1.6 g L^{-1} h^{-1} on average, or ca. 2.4 g L^{-1} h^{-1} at maximum fermentation rate). The aeration rate is only 0.1 vvm (vvm denotes volume of gas per volume of liquid per minute); nutrient consumption is as described in Section 4.3.

Kinetic studies [77] have clarified the dependence of bacterial growth rate on ethanol concentration, acetic acid concentration, and concentration of dissolved oxygen; optimum values are: ethanol 23 g/L; acetic acid 36 g/L; and oxygen 2.5 – 3 mg/L. The productivity of acetic acid formation depends not only on the growth rate, but also on the cell concentration; the maximum productivity of 6 – 6.6 g L^{-1} h^{-1} occurs at ethanol 32 g/L and acetic acid 60 – 66 g/L.

Nanba et al. [78] studied the synergistic effects of acetic acid and ethanol on the growth of *Acetobacter* spp. at low concentrations, with similar results. The possible productivity is at first sight about four times higher than presently reached, and could be further increased by recycling the bacteria, but this high degree of productivity is possible only over a narrow range of concentrations. A slight change of ethanol, acetic acid, or oxygen concentration means an immediate return to lower values.

A second fermenter connected in series to the main fermenter would be necessary to complete the fermentation. To maintain the high oxygen concentration in the fermenting liquid at the higher fermentation rate, a high aeration rate would be necessary, with recovery of ethanol and acetic acid from the exhaust gas to obtain high yields; addition of oxygen would probably also be necessary. Higher nutrient consumption and increasing foam difficulties would result. The cell recycle system would need continuous microfiltration to avoid a lack of oxygen, which would harm the bacteria, and automatic control of cell concentration. The considerable expenditure on process control would merely allow production of the same amount of vinegar with $\leq$10 % acetic acid in a smaller fermenter. High productivity is not necessarily the right path to follow.

Park et al. [79], [80] tried to improve continuous acetic acid production in a laboratory fermenter equipped with a hollow fiber filter module by increasing the concentration of dissolved oxygen. The maximum acetic acid concentration reached was 5 %. The highest productivity obtained with yeast extract as nutrient was 107 g L^{-1} h^{-1}. The dissolved oxygen concentration was already lower than the critical concentration of 1 ppm, although pure oxygen was supplied at an aeration rate of 1 vvm. Furthermore, it was difficult to maintain a high concentration of viable cells in the cell recycle culture, because the cells were inactivated by the lethal conditions of acetic acid concentration, oxygen deficiency, and nutritional limitation. The effect of dissolved oxygen and acetic acid concentration on acetic acid production in continuous culture has been extensively investigated [81]. The optimum concentration of dissolved oxygen was ca. 2 ppm for acetic acid production at low acidity. But the specific acetic acid production rate was reduced to zero when the dissolved oxygen concentration was $>$2 ppm at acetic acid concentration $>$4.5 %. This is in good agreement with the fact that using pure oxygen during fermentation may completely inhibit formation of acetic acid [53].

Viability is important for high productivity, and factors affecting it have been investigated [82]. To investigate the effect of cell bleeding (withdrawal at constant flow rate) serial experiments were conducted in a bioreactor with cell recycle, and supported by mathematical analysis. Acetic acid concentrations were up to 5.6 %. Maximum productivity was 123 g L^{-1} h^{-1}. The bleed ratio and dilution rate at which concentrations of acetic acid and viable cells are relatively high can now be predicted. These results should be useful for designing a two-stage culture for acetic acid production, using the active cells in

the bleed from the first stage in a second-stage vessel beyond a lethal acetic acid concentration.

How far these results may influence the construction of industrial equipment in Japan remains to be seen. The costs of such complicated fermenters may prevent a rapid conversion to the industrial scale, although in Japan the amount of vinegar with 5 % acetic acid is relatively high. But these results are unlikely to influence the production of vinegar with more than 10 % acetic acid. A review of the Japanese development work states that the scale-up problem remains to be investigated [83].

5.4.1.7. Semicontinuous Culture and Cell Recycle

Park et al. [84] investigated semicontinuous cultures with cell recycle, and found an increase in productivity compared with semicontinuous culture without cell recycle at acidity up to 6 %. While the concentration of total cells increased in every cycle, the viable cells reached a much lower quasi-steady value, which was explained by product inhibition. Productivity was $2.9\,g\,L^{-1}\,h^{-1}$ without, and $5.0\,g\,L^{-1}\,h^{-1}$ with cell recycle at 6 % final acetic acid concentration, but $14\,g\,L^{-1}\,h^{-1}$ at 3 %.

In another study Park et al. [85] reported semicontinuous fermentations in which the ethanol concentration was held for a time between 20 and 30 g/L, by adding ethanol-rich medium. At the end of the cycle ethanol-poor medium was added. Cells were recycled by membrane filtration. The object was to obtain high concentrations of acetic acid. The highest value reached was 9 %. *Acetobacter* cells lost viability at about 6 % acetic acid, but nonviable cells were still effective for ethanol oxidation. Ebner had noted this interesting feature of *Acetobacter* cells in 1976 [86].

5.4.2. Surface Fermentation

5.4.2.1. History

Conner and Allgeier [87] have described the history of vinegar fermentation; equipment used in the famous Orleans process for the production of wine vinegar can be seen only in museums.

5.4.2.2. New Experiments with Surface Fermentation

A new approach to surface fermentation has been tried [88]. It uses a bioreactor with a shallow, horizontal flow of medium, under a bacterial film of a few hundred cm^2 surface area. Liquid depth is up to 10 mm and the acetic acid concentration of the effluent 5.76 %. The oxygen absorption rate through the microbial film is very high [89] and a kinetic study of the acetic acid production has been published [90].

A surface fermentation process in a system of series-intercommunicating fermentation vessels, with flow regulation to approximate plug flow, has been described [91]. Liquid depth under the layer of acetic acid bacteria was ca. 50 mm. The inflowing liquid had 2 % acidity and 3.5 % alcohol. The acidity of the effluent stabilized at 5 % acetic acid. The residence time was ca. 21 h.

Higashide et al. [92] reported continuous surface fermentations in a single vessel with 16 L working volume. Under optimum conditions, with a broth depth of 100 mm, 2 – 3 % acidity, and 20 – 30 g/L alcohol concentration of the inflowing liquid, vinegar with 4.5 % acetic acid was produced for 80 d. The flavor was claimed to be better than that of vinegar produced by other processes.

For the production of high-quality rice vinegar in Japan, surface fermentation has been developed to an automatic process at lower cost.

5.4.3. Trickle Processes

The old, semicontinuous generator process used beech wood shavings, birch twigs, or corn cobs as carrier material for the vinegar bacteria. Large wooden tanks with up to $100\,m^3$ shavings are still in use. The fermenting liquid is collected in the lower part of the tank, and pumped continuously under cooling over the carrier column. Air is aspirated or ventilators are installed. Generators more than 40 years old are still producing.

The inhomogeneity of the carrier column makes it impossible to distribute the air and the trickling liquid as homogeneously as desired. Variations in temperature within the column can hardly be avoided. Mashes containing

high amounts of nutrients and low in total concentration form slimy deposits in the carrier material which may clog the column.

Nevertheless, the use of a silicate carrier material with large specific inner surface area and an open pore structure has been proposed [93].

5.4.4. Experiments with Immobilized Acetic Acid Bacteria

Many Japanese scientists have investigated the fermentation of rice or fruit vinegar at the usual strength of ca. 5 % acetic acid in continuous fermentation at high productivity with immobilized acetic acid bacteria.

Mori et al. [94] reported a 460-d run in an airlift bioreactor with acetic acid bacteria immobilized on κ-carrageenan gel beads, where living cells were newly released into the reactor. These newly released bacteria had an extremely high growth rate and respiratory activity, which they retained for a few generations after leakage. The fact that bacteria, entrapped in a κ-carrageenan gel and treated in a fluidized bed reactor were gradually released from the gel beads, was already known [95].

Okuhara [96] and Nanba et al. [97] proposed polypropylene fibers as carrier material for *Acetobacter* producing rice vinegar with 3 % acetic acid. Kondo et al. [98] used a ceramic honeycomb monolith as carrier in a fixed-bed reactor, and reached a maximum of 4 % acetic acid. Saeki [99] investigated calcium alginate beads in a fluidized bed reactor, and obtained 3.5 % acetic acid. This reactor has been used in combination with an ethanol fermenter to produce rice vinegar of up to 6.5 % acetic acid from a saccharified rice medium [100], [101].

Takada and Hiramitsu [102] tested a two-column, porous ceramic bioreactor to produce vinegar from orange wine with up to 8.4 % acetic acid. Yamashita et al. [103] fixed the cells on woven cotton fabric to produce vinegar from kiwi fruit and persimmon with 4.5 % acetic acid. Because of the disadvantages of all known carriers, Sueki et al. [104] tried a new ceramic carrier, Aphrocell, with a continuous pore structure, but the oxygen supply to its structure was difficult. None of the tested carriers was ideal.

Sun et al. [105] carried out experimental and theoretical studies on continuous acetic acid production by immobilized cells in a three-phase fluidized bed bioreactor at very low acetic acid concentrations. The effects of gel-entrapped and suspended cells, gel size, volumetric mass transfer coefficient of oxygen ($k_L\,a$), and solid holdup on productivity were studied by theoretical calculations and experiments.

Immobilized acetic acid bacteria are highly oxygen dependent [106–108]; it is advantageous to supply oxygen to a closed and pressurized culture. Feeding medium at a higher dilution rate gives a higher production rate. As acetic acid inhibits the bacteria, multiple reactors in series should be the best means of increasing acidity, thereby keeping the dilution rate higher and creating the necessary prolonged residence time. An output acidity of 5.5 % was obtained in two bench-scale fluidized bed reactors in series. In scaling up to a pilot reactor, carrageenan gel beads were not used as they tended to break up. Porous chitosan beads were selected, diameter 0.3 – 3 mm, giving a higher productivity, up to 9 g L^{-1} h^{-1}, but the microbial layer was easily peeled off. More operational experience is needed for the practical application of this technology.

5.4.5. Environmental Protection

Vinegar can be produced without environmental pollution. The exhaust gas contains alcohol, acetic acid, and ethyl acetate according to their vapor pressures at fermentation temperature ca. 30 °C. To obtain high yields, the aeration rate is kept as low as possible. It is common practice to cool the exhaust gas with water and, in order to absorb the rest of the volatiles, it can be scrubbed with water, which is recycled and later used for mash preparation.

The raw vinegar contains bacteria which must be filtered off. Most modern factories use cross-flow filtration without any filter aid. The very small quantity of remaining solids (ca. 100 g per 1000 L vinegar) formed by the acetic acid bacteria, is degraded during wastewater treatment.

5.5. Treatment of Raw Vinegar

Vinegar produced by one of these processes is called "raw vinegar". It contains <0.5 % alcohol

and, depending on the raw material, up to 20 % acetic acid. It is more or less turbid as it contains acetic acid bacteria and deposits which originate from the raw material; prior to consumption, it needs further treatment.

5.5.1. Storage and Quality

In the course of acetous fermentation, the pH of the fermenting mash is reduced. Vinegars obtained from natural raw material are unstable with regard to the solubility of previously dissolved substances. This lability lasts longer, for lower values of the pH change during fermentation. A freshly produced cider vinegar may need months to stabilize. Alcohol vinegar does not show this instability. Vinegar for storage should preferably be undiluted, as discharged from the fermenter. In storage the quality improves, especially important for all types of wine vinegar. The aging of wine vinegar comprises many complicated reactions [109]. Vinegar must have a pure aroma which imparts the flavor of the raw material.

5.5.2. Filtration

The traditional methods of vinegar clarification by fining and precoat filtration are still in use [110]. Micro- and ultrafiltration are gaining ever-increasing market shares in the large European and American factories.

The basic principle of the first long-term ultrafiltration process for vinegar, which was performed in the Frings stream filter, has been described in detail [110], and further improvements have been made [111].

Vinegar is a cheap substance, so filter costs have to be low, and repeated long-term use of the filter modules is important. The unavoidable formation of a film on the membrane (reducing the filtration rate) must be delayed as long as possible. Easy, repeated and mainly automatic cleaning of the membrane is also important. A cross-flow filter, partly fulfilling these requirements, may be more or less suitable for the filtration of vinegar.

5.5.3. Pasteurization

Wine vinegars are frequently unstable, owing to enzymes and microorganisms which may later cause cloudiness in bottle. To avoid this, vinegar is pasteurized before or after bottling. Heating to 75 – 80 °C for 30 – 40 s, or to 50 – 55 °C for longer periods is common [110].

5.5.4. Bottling and Shelf Life

Vinegar for household use is filled into glass or polyethylene bottles, sealed with plastic screw caps. Air tightness is essential to guarantee preservation. Filtered spirit vinegar does not present great difficulties, but all wine vinegars do. Institutional users such as restaurants or hospitals often get their vinegar in plastic containers of ca. 25 L, industrial users in stainless steel tank cars.

6. Constituents of Vinegar

Besides acetic acid and ethanol, vinegar contains secondary constituents which play an important role in its smell, taste, and preserving qualities. These constituents have their origin in the raw material, in added nutrients, and in the water used for dilution. They are also formed by acetic acid bacteria or by interaction of the components of the vinegar.

Spirit Vinegar. Schanderl and Staudenmeyer [112] discovered 8 amino acids in a spirit vinegar produced by submerged fermentation, where only glucose and inorganic salts had been added as nutrients; Bergner and Petri [113–115] increased the number of amino acids found in spirit vinegar to 18. According to Petri [116] these are mainly products of autolysis of the acetic acid bacteria. Aurand et al. [117] and Kahn et al. [118], [119] identified 27 volatile compounds in an ether – pentane extract of neutralized spirit vinegar; more details are given in reference [110].

Wine Vinegar. Wine vinegars contain the same spectrum of amino acids as spirit vinegar, but in larger amounts. During acetous fermentation, acetylmethylcarbinol develops in varying

quantities [120], [121]. In an ether – pentane extract of wine vinegar, KAHN et al. [119] identified 42 compounds; more details are given in reference [110].

Cider Vinegar. KAHN et al. [118], [119] identified 33 compounds in the ether – pentane extract of cider vinegar, and only a few were found in the cider from which the vinegar was made. The compounds correspond approximately to those found in wine vinegar.

Malt Vinegar. JONES and GREENSHIELDS [122], [123] identified 6 esters, 6 alcohols, 2 acids, acetoin, and acetaldehyde in malt vinegar. KAHN et al. [119] found 54 compounds in malt vinegar, including halogen compounds which obviously originated from the raw material.

Rice Vinegar. MAEKAWA and KODAMA [124] identified 5 phenolic acids in rice vinegar. YANAGIDA et al. [125] found that the percentage of amino acids in rice wine decreased during acetous fermentation. KOIZUMI et al. [126] compared the free amino acids, organic acids, and aroma components in Chinese vinegars with those of Japanese rice vinegars. The Chinese vinegars were rich in alanine, glutamic acid, lysine, and leucine, in amounts ca. 10 times higher than in Japanese rice vinegar. In Chinese vinegars, lactic acid and pyroglutamic acid were the most dominant acids besides acetic acid. High levels of *n*-propyl alcohol, isobutyl alcohol, acetoin, diacetyl, *n*-butylacetate, ethylacetate, and furfural were detected among the aroma components.

Whey Vinegar. BOURGEOIS [127] analyzed the amino acids of whey vinegar; HADORN and ZÜRCHER [128] determined the specific compounds for its characterization.

7. Analysis of Vinegar

Vinegar is analyzed for process control by general routine methods, and for a comprehensive knowledge of its chemical constituents by special methods. The latter is needed to distinguish between vinegar made by fermentation and diluted synthetic acid, and between true wine or fruit vinegar and blends of such vinegars with spirit vinegar or acetic acid; older methods are described in reference [110].

Newly developed methods offer promising results. Authentication of the origin of vinegars is determined by modern isotopic methods [129]. Initial results demonstrated the ability to detect added synthetic acid, and to establish the botanical origin of the raw materials. It is now possible not only to determine whether the vinegar is made from wine, apples, malt, cane, or beets, but also to determine the country, and even the region, of the raw material. The detection threshold of adulteration with synthetic acetic acid of vinegar of known or unknown geographic origin is 5 and 10 %, respectively. The detection threshold for alcohol vinegar added to wine vinegar of known or unknown geographic origin is 10 and 20 %, respectively.

8. Use of Vinegar

NUNHEIMER and FABIAN [130] found that acetic acid, compared with citric, lactic, malic, and tartaric acids, is a stronger growth inhibitor of microorganisms at higher pH. SZAKALL [131] stated that vinegar has a specific, inhibitory effect on the growth of microorganisms, compared with diluted acetic acid, whose effect is a function of acid concentration only; the reason is as yet unknown.

Household vinegar is mainly used to dress salads. The food industry makes use of the long-familiar ability of vinegar to preserve and season food. The most important preserved vegetables prepared with vinegar are gherkins, onions, red beets, and peppers, in the form of pickles and relishes. Vinegar is an important ingredient in mustard, tomato ketchup, Worcester sauce, mayonnaise, and salad dressings. Fish and meat are also preserved with vinegar.

9. Economic Aspects

European Union. In 1992 the European Union produced 325×10^6 L of alcohol vinegar, 143×10^6 L of wine vinegar, and 45×10^6 L of other vinegars (all at 10 % acetic acid).

United States. The United States vinegar industry no longer publishes production figures, so older figures from the U.S. Department of Commerce's Census of Manufacturers must be quoted here. Production volumes in 1987 were 581×10^6 L distilled vinegar, 54×10^6 L cider vinegar, and 126×10^6 L other vinegars, all of different strength.

Japan. In 1992, Japan produced 56×10^6 L of rice vinegar, 146×10^6 L of other grain vinegars, and 161×10^6 L of other brewed vinegars, in total 394×10^6 L, at a concentration of 4 – 5 % acetic acid.

Worldwide. From statistics and personal knowledge of the situation in many countries, annual world production (excluding the former Soviet Union and China) of vinegar at 10 % acetic acid is assumed to be ca. 1900×10^6 L, or 190 000 t pure acetic acid.

10. References

1. H. Ebner, H. Follmann in H.-J. Rehm, G. Reed (eds.): *Biotechnology,* **vol. 3,** chap. 3 b, Verlag Chemie, Weinheim 1983, pp. 387 – 407.
2. M. Gillis, J. de Ley, *Int. J. Syst. Bacteriol.* **30** (1980) 7 – 27.
3. J. de Ley, *J. Gen. Microbiol.* **24** (1961) 31.
4. E. Stackebrandt, R. G. E. Murray, H. G. Trüper, *Int. J. Syst. Bacteriol.* **38** (1988) 321 – 325.
5. J. de Ley, J. Swings, F. Gossele in N. R. Krieg, J. G. Holt (eds.): *Bergey's Manual of Systematic Bacteriology,* **vol. 1: Genus Acetobacter,** Williams and Wilkins, Baltimore 1984, pp. 268 – 274.
6. J. Swings in A. Balows et al. (eds.): *The Prokaryotes,* 2nd ed., **vol. III,** Springer Verlag, New York 1992, pp. 2268 – 2286.
7. M. Gillis et al., *Int. J. Syst. Bacteriol.* **39** (1989) 361 –364.
8. E. Entani, S. Ohmori, H. Masai, K.-I. Suzuki, *J. Gen. Appl. Microbiol.* **31** (1985) 475 – 490.
9. M. Sievers, S. Sellmer, M. Teuber, *Syst. Appl. Microbiol.* **15** (1992) 386 – 392.
10. A. J. Schocher et al., *Arch. Microbiol.* **121** (1979) 193 –197.
11. F. L. Davies, H. M. Underwood, M. J. Gasson, *J. Appl. Bacteriol.* **51** (1981) 325 – 337.
12. T. Inoue, M. Fukuda, K. Yano, *J. Ferment. Technol.* **63** (1985) no. 1, 1 – 4.
13. M. Fukaya et al., *Agric. Biol. Chem.* **49** (1985) no. 5, 1349 – 1355.
14. M. Teuber, M. Sievers, A. Andresen, *Biotechnol. Lett.* **9** (1987) 265 – 268.
15. M. Sievers, A. Andresen, M. Teuber, *Food Biotechnol.* **4** (1990) 555.
16. M. Fukaya et al., *Agric. Biol. Chem.* **53** (1989) no. 9, 2435 – 2440.
17. S. Valla, D. H. Coucheron, J. Kjos-Bakken, *Arch. Microbiol.* **134** (1983) 9 – 11.
18. H. Okumura, T. Uozumi, T. Beppu, *Agric. Biol. Chem.* **49** (1985) no. 4, 1011 – 1017.
19. M. Fukaya et al., *Agric. Biol. Chem.* **49** (1985) no. 7, 2083 – 2090.
20. M. Fukaya et al., *Agric. Biol. Chem.* **49** (1985) no. 7, 2091 – 2097.
21. M. Fukaya et al., *Agric. Biol. Chem.* **49** (1985) no. 8, 2407 – 2411.
22. W. W. Stamm, M. Kittelmann, H. Follmann, H. G. Trüper, *Appl. Microbiol. Biotechnol.* **30** (1989) 41 – 46.
23. S. Sellmer, M. Sievers, M. Teuber, *Syst. Appl. Microbiol.* **15** (1992) 610 – 616.
24. C. Defives, D. Ochin, J. P. Hornez, M. Werquiin, *Microbiol. Aliments Nutr.* **8** (1990) 77 – 79.
25. D. E. Bradley, *Bacteriol. Rev.* **31** (1967) 230 – 314.
26. H.-W. Ackermann, *Microbiol. Sci.* **4** (1987) 214 – 218.
27. M. Ameyama, O. Adachi in W. A. Wood (ed.), *Methods Enzymol.* **89** (1982) 450 – 457.
28. M. Ameyama, O. Adachi in W. A. Wood (ed.), *Methods Enzymol.* **89** (1982) 491 – 497.
29. O. Adachi et al., *Agric. Biol. Chem.* **42** (1978) 2045 –2056.
30. H. Muraoka, Y. Watabe, N. Ogasawara, H. Takahashi, *J. Ferment. Technol.* **59** (1982) 247 – 255.
31. O. Adachi et al., *Agric. Biol. Chem.* **42** (1978) 2331 –2340.
32. O. Adachi et al., *Agric. Biol. Chem.* **44** (1980) 503 –515.
33. M. Ameyama et al., *FEBS Lett.* **130** (1981) no. 2, 179 –183.
34. R. Hommel, H.-P. Kleber, *Proc. 3rd Eur. Congr. Biotechnol.* **I** (1984) 133 – 137.
35. H. Muraoka, Y. Watabe, N. Ogasawara, H. Takahashi, *J. Ferment. Technol.* **59** (1981) 247 – 255.
36. K. Matsushita et al., *Biosci. Biotech. Biochem.* **56** (1992) 303 – 310.
37. M. Fukaya et al., *Appl. Environ. Microbiol.* **55** (1989) no. 2, 171 – 176.

38. K. Tayama et al., *Appl. Microbiol. Biotechnol.* **32** (1989) 181 – 185.
39. T. Asai: *Acetic Acid Bacteria,* University of Tokyo Press, Tokyo, University Park Press, Baltimore 1968.
40. T. Leisinger, Thesis, Universität Rudolfstadt, 1965.
41. K. Kersters, J. de Ley, A. van Leeuwenhoek, *J. Microbiol. Ser.* **34** (1968) 393 – 408.
42. J. de Ley, K. Kersters, *Bacteriol. Rev.* **28** (1964) 83 – 95.
43. M. Kulhanek in V. Krumphanzl, Z. Rehacek (eds.): Modern Biotechnology, vol. 2, Czechoslovak Academy of Sciences 1984, pp. 614 – 676.
44. Z. G. Razumovskaya, T. Z. Belousova, *Mikrobiologiya* **21** (1952) 403 – 407.
45. O. Hromatka, H. Gsur, *Enzymologia* **25** (1962) no. 2, 81 – 86.
46. G. D. Brown, C. Rainbow, *J. Gen. Microbiol.* **15** (1956) 61.
47. M. R. Raghavendra Rao, J. L. Stokes, *J. Bacteriol.* **66** (1953) 634.
48. F. Gossele, J. Swings, J. de Ley, *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Hyg. Abt. I: Orig.* **C 1** (1980) 348 – 350.
49. M. Ameyama, K. Kondo, *Agric. Biol. Chem.* **30** (1966) 203.
50. O. Adachi et al., *Agric. Biol. Chem.* **54** (1990) no. 10, 2751 – 2752.
51. M. Ameyama, K. Shinagawa, K. Matsushita, O. Adachi, *Agric. Biol. Chem.* **48** (1984) no. 11, 2909 –2911.
52. M. Ameyama, E. Shinagawa, K. Matsushita, O. Adachi, *Agric. Biol. Chem.* **49** (1985) no. 3, 699 – 709.
53. O. Hromatka, H. Ebner, *Enzymologia* **15** (1951) no. 2, 57 – 69.
54. O. Hromatka, W. Exner, *Enzymologia* **25** (1962) no. 1, 37 – 51.
55. H. Muraoka, Y. Watabe, N. Ogasawara, H. Takahashi, *J. Ferment. Technol.* **61** (1983) no. 1, 89 – 93.
56. A. Hitschman, H. Stockinger, *Appl. Microbiol. Biotechnol.* **22** (1985) 46 – 49.
57. O. Hromatka, G. Kastner, H. Ebner, *Enzymologia* **15** (1953) no. 6, 337 – 350.
58. F. M. Vera, D. I. C. Wang, *174th Am. Chem. Soc. Natl. Meet.,* Chicago 1977.
59. O. Hromatka, H. Ebner, *Enzymologia* **13** (1949) no. 6, 369 – 387.
60. A. Mori, G. Terui, *J. Ferment. Technol.* **50** (1972) no. 2, 70 – 78.
61. Heinrich Frings, US 4 503 078, 1982 (H. Ebner).
62. Heinrich Frings, US 4 076 844, 1975 (H. Ebner, A. Enenkel).
63. Heinrich Frings, US 4 773 315, 1987 (A. Enenkel).
64. Nakano Vinegar Co., US 4 282 257, 1980 (Y. Kunimatsu, H. Okumara, H. Masai, K. Yamada, M. Yamada).
65. Nakano Vinegar Co., US 4 364 960, 1980 (Y. Kunimatsu, H. Okumara, H. Masai, K. Yamada, M. Yamada).
66. Nabisco Brands, US 4 456 622, 1982 (J. A. Maselli, R. O. Horwath).
67. Heinrich Frings, US 3 813 086, 1967 (H. Ebner, A. Enenkel).
68. H. Ebner, US 3 262 252, 1958.
69. Heinrich Frings, US 4 997 660, 1989 (R. Wittler).
70. Heinrich Frings, US 3 290 924, 1963 (H. Ebner, A. Enenkel).
71. Vogelbusch, US 4 404 284, 1981 (M. Heider, J. Hofschneider, W. Schallenberger, F. Nodes, E. Kempe).
72. D. Hekmat, D. Vortmeyer, *J. Ferment. Bioeng.* **73** (1992) no. 1, 26 – 30.
73. A. Mori: *Instrumentation and Control Engineering in Japan,* Kogyo Gijutsu-sha, 1989, pp. 18 – 24.
74. Y. Nomura, M. Iwahara, M. Hongo, *Appl. Environ. Microbiol.* **54** (1988) no. 1, 137 – 142.
75. Y. Nomura, M. Iwahara, M. Hongo, *J. Biotechnol.* **12** (1989) 317 – 326.
76. Y. Nomura, *Hakkokogaku* **70** (1992) 205 – 216.
77. J. J. Bänsch, W. Bauer, *BioEng.* **7** (1991) no. 3, 26 – 36.
78. A. Nanba, A. Tamura, S. Nagai, *J. Ferment. Technol.* **62** (1984) no. 6, 501 – 505.
79. J. S. Park et al., *Biotechnol. Bioeng.* **33** (1989) 918 – 923.
80. Y. S. Park et al., *J. Ferment. Bioeng.* **68** (1989) no. 5, 315 – 319.
81. Y. S. Park et al., *J. Ferment. Bioeng.* **68** (1989) no. 2, 96 – 101.
82. Y. S. Park, K. Toda, *J. Gen. Appl. Microbiol.* **36** (1990) 221 – 233.
83. M. Fukaya, Y. S. Park, K. Toda, *J. Appl. Bacteriol.* **73** (1992) 447 – 454.
84. Y. S. Park et al., *Biotechnol. Lett.* **13** (1991) no. 4, 271 –276.
85. Y. S. Park et al., *Appl. Microbiol. Biotechnol.* **35** (1991) 149 – 153.
86. *Ullmann,* 4th ed., **11,** 41 – 55.
87. H. A. Conner, R. J. Allgeier, *Adv. Appl. Microbiol.* **20** (1976) 81 – 133.

88. K. Toda et al., *Appl. Microbiol. Biotechnol.* **30** (1989) 559 – 563.
89. K. Toda, M. Miyake, H. Ohtake, *J. Ferment. Bioeng.* **70** (1990) no. 2, 114 – 118.
90. Y. S. Park, H. Ohtake, K. Toda, *Appl. Microbiol. Biotechnol.* **33** (1990) 259 – 263.
91. Kewpie Jozo Kabushiki Kaisha, GB 1 305 868, 1970.
92. T. Higashide et al., *Bull. Fac. Bioresour.* **8** (1992) 43 –49.
93. Carl Kühne, EP 0 523 524 A 1, 1992 (U. Steckowski).
94. A. Mori, N. Matsumoto, C. Imai, *Biotechnol. Lett.* **11** (1989) no. 3, 183 – 188.
95. J. Osuga, A. Mori, J. Kato, *J. Ferment. Technol.* **62** (1984) no. 2, 139 – 149.
96. A. Okuhara, *J. Ferment. Technol.* **63** (1985) no. 1, 57 – 60.
97. A. Nanba, K. Kimura, S. Nagai, *J. Ferment. Technol.* **63** (1985) no. 2, 175 – 179.
98. M. Kondo, Y. Suzuki, H. Kato, *Hakkokogaku* **66** (1988) 393 – 399.
99. A. Saeki, *Nippon Shokuhin Kogyo Gakkaishi* **37** (1990) no. 3, 191 – 198.
100. A. Saeki, *Nippon Shokuhin Kogyo Gakkaishi* **37** (1990) no. 9, 722 – 725.
101. A. Saeki, *Nippon Shokuhin Kogyo Gakkaishi* **38** (1991) no. 10, 891 – 896.
102. M. Takada, T. Hiramitsu, *Nippon Shokuhin Kogyo Gakkaishi* **38** (1991) no. 10, 967 – 971.
103. S. Yamashita, S. Ohta, H. Suenaga, *Nippon Shokuhin Kogyo Gakkaishi* **38** (1991) no. 7, 608 – 613.
104. M. Sueki, N. Kobayashi, A. Suzuki, *Biotechnol. Lett.* **13** (1991) no. 3, 185 – 190.
105. Y. Sun, S. Furusaki, *J. Ferment. Bioeng.* **69** (1990) no. 2, 102 – 110.
106. A. Mori, *Process Biochem.* **20** (1985) 67 – 74.
107. A. Mori in A. Tanaka, T. Tosa, T. Kobayashi (eds.): *Industrial Application of Immobilized Biocatalysts,* Marcel Dekker, New York – Basel – Hong Kong 1992, pp. 291 – 313.
108. A. Mori, S. Tanaka, N. Matsumoto, C. Imai in S. Furusaki, I. Endo, R. Matsuno (eds.): *Biochemical Engineering for 2001,* Springer Verlag, Tokyo – Berlin – Heidelberg – New York 1992, pp. 441 – 443.
109. K. Mecca, R. Andreotti, L. Veronelli: *L'Aceto,* Edizioni AEB S.p.A.S. Polo-Brescia, Italy, 1979, pp. 1 – 433.
110. H. Ebner, H. Follmann in H.-J. Rehm, G. Reed (eds.): *Biotechnology,* **vol. 5,** chap. 9, Verlag Chemie, Weinheim 1983, pp. 425 – 446.
111. Heinrich Frings, US 4 689 153, 1985 (H. Follmann, H. Ebner).
112. H. Schanderl, T. Staudenmayer, *Z. Lebensm. Unters. Forsch.* **163** (1956) 121 – 122.
113. K. G. Bergner, H. Petri, *Angew. Chem.* **71** (1959) 31 – 32.
114. K. G. Bergner, H. Petri, *Z. Lebensm. Unters. Forsch.* **111** (1960) 319 – 333.
115. K. G. Bergner, H. Petri, *Z. Lebensm. Unters. Forsch.* **111** (1960) 494 – 504.
116. H. Petri, Thesis, Technische Hochschule, Stuttgart 1958, pp. 1 – 92.
117. L. W. Aurand, J. A. Singleton, T. A. Bell, J. L. Etchells, *J. Food Sci.* **31** (1966) no. 2, 172 – 177.
118. H. J. Kahn, G. B. Nickol, H. A. Conner, *Agric. Food Chem.* **14** (1966) 460 – 465.
119. H. J. Kahn, G. B. Nickol, H. A. Conner, *Agric. Food Chem.* **20** (1972) 214 – 218.
120. C. Galoppini, O. T. Rotini, *Ann. Fac. Agrar. Univ. Pisa* **17** (1956) 99 – 111.
121. O. T. Rotini, C. Galoppini, *Ann. Sper. Agrar.* **11** (1957) 1355 – 1372.
122. D. D. Jones, R. N. Greenshields, *J. Inst. Brew.* **75** (1969) 457 – 463.
123. D. D. Jones, R. N. Greenshields, *J. Inst. Brew.* **77** (1971) 160 – 163.
124. K. Maekawa, M. Kodama, *Agric. Biol. Chem.* **28** (1964) 436 – 442.
125. F. Yanagida et al., *Nippon Jozo Kyokai Zasshi* **69** (1974) no. 11, 759 – 764.
126. Y. Koizumi, T. Nakakoji, F. Yanagida, *Nippon Shokuhin Kogyo Gakkaishi* **32** (1985) no. 4, 288 –294.
127. J. Bourgeois, *Branntweinwirtschaft* **79** (1957) 250 –254.
128. H. Hadorn, K. Zürcher, *Mitt. Geb. Lebensmittelunters. Hyg.* **64** (1973) 480 – 503.
129. M. Dumoulin, Eurofins Laboratoires, Nantes, France, 1993.
130. T. D. Nunheimer, F. W. Fabian, *Am. J. Public Health* **30** (1940) 1040 – 1049.
131. A. Szakall, *Z. Lebensm. Unters. Forsch.* **95** (1952) 100 – 107.

Vinyl Acetate → Vinyl Esters
Vinyl Chloride → Chlorinated Hydrocarbons

Vinyl Esters

GÜNTER ROSCHER, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany

1. Introduction 1
2. Vinyl Acetate 1
2.1. Properties 1
2.2. Production 2
2.2.1. Addition of Acetic Acid to Acetylene 2
2.2.2. Ethylidene Diacetate Process 5
2.2.3. Reaction of Acetic Acid with Ethylene and Oxygen 5
2.2.4. Reaction of Methyl Acetate with CO and H_2 10
2.3. Quality Specifications, Analysis, Storage, Transport, and Toxicology 10
2.4. Use, Economic Importance 11
3. Vinyl Propionate 12
3.1. Properties 12
3.2. Production 12
3.3. Use 13
4. Vinyl Esters of Higher Carboxylic Acids 13
4.1. Reaction with Acetylene in the Gas Phase 13
4.2. Reaction with Ethylene and Oxygen 14
4.3. Transvinylation and Other Processes 14
4.4. Reaction with Acetylene in the Liquid Phase (Wacker and Shell Processes) 14
4.5. Individual Vinyl Esters 15
5. References 16

1. Introduction

The first vinyl esters, vinyl trichloroacetate and vinyl acetate, were prepared by KLATTE in 1912 by treatment of acetylene with the corresponding carboxylic acid and mercury salts as catalyst. Initially, the reaction was used for the industrial production of vinyl chloroacetate. The polymer obtained from this vinyl ester was used during World War I as a varnish for airplanes. Vinyl acetate is the most important vinyl ester. Vinyl propionate and vinyl esters of higher carboxylic acids are now also produced industrially (see Chap. 3 and 4).

2. Vinyl Acetate

2.1. Properties

Physical Properties. Vinyl acetate [*108-05-4*], $CH_3CO_2CH{=}CH_2$, M_r86.09, is a colorless, flammable liquid with a characteristic, slightly pungent odor, *bp* 72.8 °C, density at 20 °C 0.932 g/mL, *mp* −93.2 °C, viscosity 0.43 mPA · s, vapor pressure 12 kPa at 20 °C, 42.6 kPa at 50 °C, coefficient of cubic expansion 0.0014 K^{-1}, flashpoint − 8 °C, ignition temperature 385 °C. Lower/upper flammability limits in air 2.3/13.4 vol %, ignition group (VDE 0165) G 2, specific heat 1.926 kJ/kg; heat of evaporation 379.3 kJ/kg at 72.7 °C, heat of combustion 2082.0 kJ/mol, refractive index n_D^{20} 1.3956, heat of polymerization 1035.8 kJ/kg, solubility of water in vinyl acetate 0.9 wt % at 20 °C, solubility of vinyl acetate in water 2.3 wt % at 20 °C, azeotrope with water *bp* 66 °C/100 kPa, water content 7.3 wt %.

Chemical Properties. The chemical property which is exploited almost exclusively is the capacity to polymerize (see Section 2.4).

Other reactions of vinyl acetate: halogens give 1,2-dihaloethyl acetates [6], hydrogen halides give 1-haloethyl acetates [7], acetic acid gives ethylidene diacetate (see Section 2.2), hydrogen cyanide gives 2-acetoxypropionitrile [8], hydrogen peroxide gives hydroxyacetaldehyde [9], and dienes, such as butadiene or cyclopentadiene, give Diels – Alder products [10].

Transesterification with carboxylic acids produces the corresponding vinyl carboxylate and acetic acid [11] (see also Section 2.3) and with alcohols gives the corresponding acetate and acetaldehyde. Thermal cleavage gives ketene and acetaldehyde [12]. Acid-catalyzed [13] or thermal [14] hydrolysis produces acetaldehyde and acetic acid. Vinyl acetate can be epoxidized with peracetic acid (84 % yield)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_419

[15]. It undergoes addition of H-active compounds, e.g., dimethyl phosphite gives dimethyl acetoxyethylphosphonate [16].

2.2. Production

There are various possible routes for vinyl acetate production:

1) Addition of acetic acid to acetylene:
   a) in the liquid phase in the presence of homogeneous mercury salt catalysts
   b) in the gas phase in the presence of heterogeneous catalysts containing zinc salts
2) Addition of acetic anhydride to acetaldehyde giving ethylidene diacetate, and subsequent cleavage of the latter to form vinyl acetate and acetic acid
3) Reaction of ethylene with acetic acid and oxygen
   a) in the liquid phase in the presence of palladium/copper salts as homogeneous catalysts
   b) in the gas phase on heterogeneous catalysts containing palladium
4) Reaction of methyl acetate or dimethyl ether with carbon monoxide and hydrogen in the liquid phase in the presence of homogeneous catalysts, e.g., rhodium salts or noble metals of the platinum group, giving ethylidene diacetate; cleavage of the latter giving acetic acid and vinyl acetate

Acetylene, which is expensive, has mostly been replaced by the cheaper alternative, ethylene; ca. 80 % of the available capacity is used for process 3 b and ca. 20 % for process 1 b.

Processes 1 a, 2, and 3 a are no longer used, while process 4, although allegedly developed to an industrial level, has not yet been used industrially. It may become more important because the starting materials can readily be produced from coal or naphtha.

2.2.1. Addition of Acetic Acid to Acetylene

Liquid-Phase Process The addition of carboxylic acids to acetylene using mercury salts as the catalyst [17]

$$CH_3CO_2H + CH{\equiv}CH \xrightarrow[\text{Hg salts}]{60-100\,^\circ C} CH_3CO_2CH{=}CH_2 \qquad \Delta H = -117\ \text{kJ/mol}$$

is now of historical interest only. For more details see [18], [19].

Gas Phase Process

$$CH{\equiv}CH + CH_3CO_2H \xrightarrow[\text{zinc acetate on activated charcoal}]{160-210\,^\circ C} CH_3CO_2CH{=}CH_2 \qquad \Delta H = -117\ \text{kJ/mol}$$

The first process was developed in Munich by Consortium f. Elektrochemische Industrie [20]. It was further developed and used industrially by Wacker Chemie in Burghausen. Until 1965, almost all vinyl acetate was produced by the acetylene gas-phase process. Only two smaller plants used the ethylidene diacetate process [21]. Zinc salts on activated charcoal have proved to be effective catalysts. The development of suitable types of activated charcoal has improved the process [22].

Most of the industrial development work was concerned with carrying out the reaction. In the shaft furnaces used initially as reactors, controlling the heat of reaction was difficult. An occasional runaway of the reaction led to baking of the catalyst. Exothermic autodecomposition of the acetylene could not always be avoided. For better heat removal, other types of reactor employing a cooling medium were used. At Hoechst, Fischer furnaces were initially used, and at Wacker-Chemie tube furnaces.

As far as is known, all producers using the solid bed catalysts have started to use tube reactors because of the defined gas flow, the easier charging and discharging of the catalyst, and the good temperature control. Only at Kurashiki in Japan, Du Pont in the United States, and in some plants in the former Soviet Union have fluidized-bed reactors been used. Solid bed and fluidized bed processes are considered of equal value.

Process Description. The modes of operation of individual producers no longer differ significantly [23]. The process used by Hoechst until 1975 is described as an example (Fig. 1). Process data are given later in this section.

The circulated acetylene is preheated at the exit of the reactor in a countercurrent, and is then mixed with acetic acid vapor. The circulating gas can also be fed directly through the

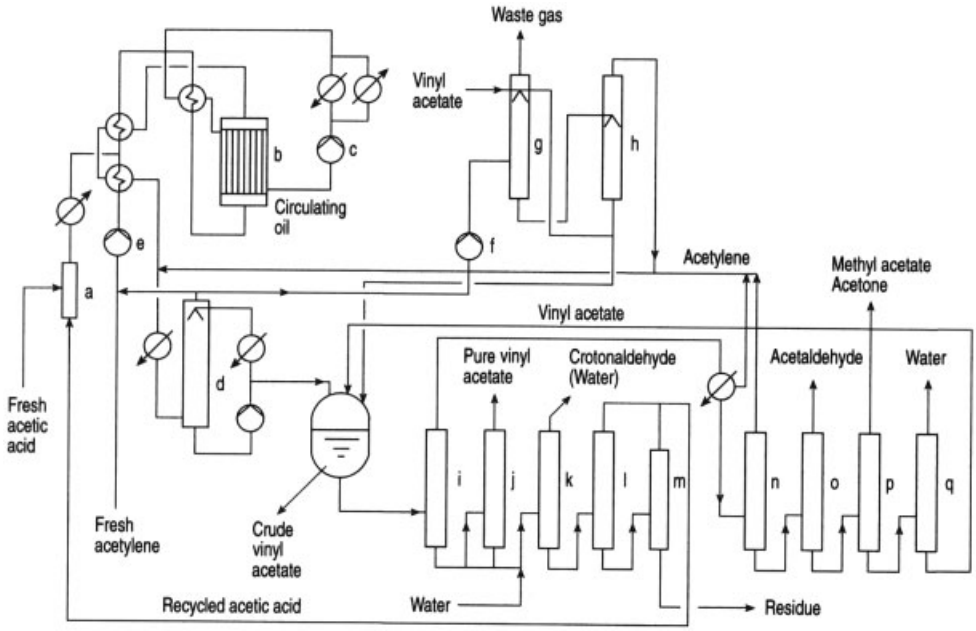

Figure 1. Acetylene gas phase process for vinyl acetate production
a) Acetic acid evaporator; b) Reactor; c) Heat transfer oil, cooling loop; d) Quenching tower; e) Recycle gas blower; f) Liquid ring pump; g) Washing column; h) Regenerating column; i) Lightends column; j) Pure vinyl acetate column; k) Crotonaldehyde column; l) Acetic acid column; m) Residue column; n) Degassing column; o) Acetaldehyde column; p) Acetone column; q) Water removal

acetic acid evaporator (a). The gas, heated to the reaction temperature, enters the tube reactor (b).

The heat of reaction is removed by heat transfer oil in the cooling loop (c).

The mixture leaving the reactor is cooled in stages. The last cooling stage takes place in the quenching tower (d). In this packed column, the gas mixture is cooled to 0 °C. The condensate itself is used for cooling. It is circulated from the bottom to the top of the quenching tower via a brine – cooled condenser. The liquid product stream is removed; excess acetylene is recycled via a circulating gas blower (e).

The crude vinyl acetate produced at the bottom of the quenching tower is distilled giving acetic acid, which is recycled, and pure vinyl acetate. Some 90 % of the circulating gas is acetylene; the remainder is CO_2, CO, and methane formed from thermal decomposition, acetaldehyde, N_2, and other inert substances. To prevent enrichment of the impurities in the circulating gas, a small portion is removed behind the quenching tower (d) and then purified.

The acetylene from the gas stream, which has been brought to ca. 100 kPa overpressure by means of a liquid ring pump (f), is extensively absorbed in a washing column (g) containing brine-cooled vinyl acetate. The inert substances are led as waste-gas to flare. The sump product from (g) is freed from dissolved acetylene in a regeneration column (h) by boiling. The acetylene is recycled. The crude vinyl acetate formed in the bottom of the quenching tower (d) contains ca. 62 – 63 wt % vinyl acetate and 30 –35 % acetic acid. It also contains dissolved acetylene, acetaldehyde, crotonaldehyde, acetone, methyl acetate, ethylidene diacetate, and acetic anhydride.

In the lightends column (i) acetaldehyde, acetone, methyl acetate, dissolved acetylene, some vinyl acetate, and water (originating from the starting materials) are first distilled overhead. The sump product is fractionated in distillation columns (j – m). Pure vinyl acetate is removed as the top product of the column (j). In (k) crotonaldehyde distills at the head as an azeotrope after addition of water. The bottom of the acetic acid

column (l) also contains, besides acetic acid, the high-boiling ethylidene diacetate and acetic anhydride (unless they have already been hydrolyzed in the crotonaldehyde column), and very small quantities of polymers. Column (m) is operated under vacuum, and to some extent batchwise; the residual acetic acid distills so extensively that the liquid sump product which remains can subsequently be incinerated. The distillate from the lightends column (i) contains mainly acetaldehyde, acetone, methyl acetate, vinyl acetate, and water. It is purified in distillation columns (o – q).

Stabilizers. To avoid polymerization during the distillative work-up of the crude vinyl acetate, polymerization inhibitors are added. The preferred stabilizer is hydroquinone. Copper resinate, phenothiazine, or methylene blue are also used.

Materials and Environmental Aspects. The plant is made from mild steel in the hot area of the reaction section and in places where no liquid acetic acid is present. These areas include the reactor, the gas – gas heat exchanger, and the circulating gas blower. For the distillation section and the equipment in the reaction section which comes into contact with liquid product, stainless steel 316 L is used. The process has almost no polluting waste streams. All gaseous or liquid byproducts (high-boiling, crotonaldehyde, and acetone – methyl acetate fractions) can be incinerated. Waste circulating gas is passed to the excess gas burner.

Process Data. The catalyst consists of zinc acetate on activated charcoal (particle size 3 – 4 mm) as the carrier material. The zinc content is 10 – 15 wt %. The catalyst is produced by dipping. As the activated charcoal can contain traces of copper, small quantities of other components are added to the catalyst to prevent the formation of cuprene, which can block the tubes.

The operating time is ca. 5000 – 7000 h, depending on the type of activated charcoal used. It also depends on the purity of the acetylene and the circulating acetic acid. When using acetylene produced from carbide, the type of carbide is important. Because of the varying contents of phosphorus hydrides, hydrogen sulfide, arsine, and ammonia in the acetylene, it sometimes has to be further purified before being used in vinyl acetate production. This involves several steps.

Acetylene produced from petrochemicals does not contain these components. The decrease in catalyst activity is caused to a small extent by migration of zinc acetate out of the catalyst, but to a greater extent by formation of byproducts or by foreign components, which either act as catalyst poisons or adhere to the catalyst surface and pores [24], [25]. Vinyl acetate itself does not appear to contribute to the deactivation of the catalyst under the reaction conditions.

The space – time yield for the vinyl acetate is normally 60 – 70 g per liter catalyst per hour.

The reaction temperature is 160 – 170 °C, depending on the type of activated charcoal used as catalyst, but increases to 205 – 210 °C as the catalyst activity decreases. At the elevated temperature in the reactor, more byproducts are formed.

The pressure can increase a little during the operating time of a catalyst charge. This is caused by a slight increase in the flow resistance of the catalyst. The maximum pressure is ca. 40 kPa overpressure. Plants are protected against overpressures greater than 40 – 50 kPa because acetylene can undergo exothermic autodecomposition at high pressures and temperatures:

$$CH{\equiv}CH \rightarrow H_2 + 2\,C \quad \Delta H = -229.4\ \text{kJ/mol}$$

The acetylene pipes leading to the plant therefore contain barriers which inhibit the acetylene autodecomposition. The plants are also provided with equipment for automatic flushing with inert gas. The molar ratio acetic acid : acetylene is normally between ca. 1 : 4 and 1 : 4.5. The load per m^3 catalyst is ca. 135 m^3 (STP) recycle gas per hour and 76 kg acetic acid per hour. The recycle gas consists of ca. 90 % acetylene. Acetylene conversion is ca. 15 % and acetic acid conversion ca. 55 %. The yields based on acetic acid can reach 99 %, or 98 % based on acetylene, if the acetaldehyde formed (2 – 3 kg per 100 kg vinyl acetate) is included in the yield.

2.2.2. Ethylidene Diacetate Process

$$CH_3CHO + (CH_3CO)_2O \rightarrow CH_3CH(OCOCH_3)_2$$

$$CH_3CH(OCOCH_3)_2 \rightarrow CH_3CO_2CH{=}CH_2 + CH_3CO_2H$$

The process was developed by Celanese Corporation of America [26] and was operated industrially in the United States from 1953 to 1970,

and then replaced by the ethylene gas-phase process. A small plant (Celmex) operated in Mexico until 1991. Acetic anhydride is converted to ethylidene diacetate with acetaldehyde in the liquid phase using catalysts such as iron(III) chloride. The ethylidene diacetate is then cleaved thermally to vinyl acetate and acetic acid, using catalysts such as toluenesulfonic acid. Some of the ethylidene diacetate is converted back to acetic anhydride and acetaldehyde.

2.2.3. Reaction of Acetic Acid with Ethylene and Oxygen

Liquid-Phase Process. The formation of vinyl acetate from ethylene and acetic acid in the presence of palladium chloride and alkali acetate in glacial acetic acid was first described by MOISEEW [27]:

$$C_2H_4 + PdCl_2 + CH_3CO_2H \rightarrow CH_3CO_2CH{=}CH_2 + 2\,HCl + Pd$$

Addition of benzoquinone to the reaction mixture was said to reoxidize the palladium to palladium chloride. The reaction corresponds to the Wacker – Hoechst process, in which acetaldehyde is obtained from ethylene and water in the presence of palladium chloride:

$$C_2H_4 + PdCl_2 + H_2O \rightarrow CH_3CHO + Pd + 2\,HCl$$

The palladium formed is reoxidized to Pd^{2+} with copper(II) chloride. The copper(I) chloride formed is reoxidized with oxygen (→ Acetaldehyde, Chap. 4.3.1.). Production of vinyl acetate by a similar route has been widely investigated [28–30]. Corresponding production plants were commissioned by ICI in England, Celanese in the United States, and Tokuyama Petrochemical in Japan [31], but later shut down. Now only the ethylene gas-phase process uses ethylene as the starting material (see below).

In the liquid-phase process, a recycle ethylene gas stream is passed through a reaction solution containing acetic acid, water, high-boiling byproducts analogous to those of the acetaldehyde process, $PdCl_2$, and $CuCl_2$. Oxygen is passed into the reaction solution at the same time to reoxidize the palladium and the CuCl. The reaction and regeneration of the catalyst take place in one step:

$$CH_2{=}CH_2 + CH_3CO_2H + PdCl_2 \rightarrow CH_3CO_2CH{=}CH_2 + 2\,HCl + Pd$$

$$Pd + 2\,CuCl_2 \rightarrow PdCl_2 + 2\,CuCl$$

$$2\,CuCl + 2\,HCl + 1/2\,O_2 \rightarrow 2\,CuCl_2 + H_2O$$

Water is formed in the reoxidation of CuCl; part of the vinyl acetate formed is hydrolyzed to acetaldehyde and acetic acid. Part of the acetaldehyde is also formed directly from ethylene, as in the acetaldehyde process.

A certain ratio of palladium ions to copper ions and of copper ions to chloride ions is necessary for reoxidation of the palladium. Chlorine, which is lost through formation of chlorinated byproducts, is replaced by hydrochloric acid. To maintain the necessary quantity of chloride ions in solution, alkali metal chloride must be added. A typical reaction solution contains ca. 30 – 50 mg/L palladium ions and 3 – 6 g/L copper ions.

Byproducts include CO_2, formic acid, oxalic acid, oxalic acid esters, chlorinated compounds, and butenes. The pressure is 3 – 4 MPa, and the reaction temperature 110 – 130 °C. The ratio of acetaldehyde to vinyl acetate can be controlled by adjusting the water concentration and the residence time [32].

Gas-Phase Reaction. The process was developed to an industrial scale only slightly later than the ethylene liquid-phase process, and has been used in industry since 1968. Currently, 80 % of world vinyl acetate capacity uses the ethylene gas-phase process. There are two variants: one developed by National Distillers Products (United States) [49], and the other independently by Bayer in cooperation with Knapsack and Hoechst (Germany) [33–35]. Most plants employ the Bayer – Hoechst variant, of which there are several versions. The original process has been further developed by various operators.

In the ethylene gas-phase process, ethylene reacts exothermically with acetic acid and oxygen on solid bed catalysts, giving vinyl acetate and water:

$$CH_2{=}CH_2 + CH_3CO_2H + 1/2\,O_2 \rightarrow CH_3CO_2CH{=}CH_2 + H_2O$$

$\Delta H = -178$ kJ/mol

All catalysts used in industry contain palladium and alkali metal salts on carrier materials, e.g., silicic acid, aluminum oxide, lithium spinel, or activated charcoal. Additional activators can include gold, rhodium, platinum, and cadmium.

The reaction mechanism is assumed to involve either pure metal catalysis [36] or a reaction sequence according to the following equations:

$$CH_2{=}CH_2 + Pd(CH_3CO_2)_2 \rightarrow Pd + CH_3CO_2CH{=}CH_2 + CH_3CO_2H$$

$$Pd + 1/2\,O_2 + 2\,CH_3CO_2H \rightarrow Pd(CH_3CO_2)_2 + H_2O$$

The finely divided palladium on the catalyst is thought to be oxidized to divalent palladium. For reoxidation of the palladium, a redox reaction analogous to the liquid-phase process is assumed. Copper, manganese, and iron are cited as redox elements [37].

In the process, which operates above 140 °C and at overpressure 0.5 – 1.2 MPa, practically no acetaldehyde is formed, even if the acetic acid used as starting material contains water. Byproducts are water, CO_2 and small quantities of ethyl acetate, ethylidene diacetate, and glycol acetates.

Process Description: Reaction Section (Fig. 2). The recycle gas stream, which consists mainly of ethylene, is saturated with acetic acid in the evaporator (a) and is then heated to the reaction temperature. The gas stream is then mixed with oxygen in a special unit.

The allowed oxygen concentration is determined by the flammability limits of the ethylene – oxygen mixture. The flammability limit depends on temperature, pressure, and composition. It is shifted by additional components, such as acetic acid, nitrogen, and argon, which are brought in with the oxygen, or by CO_2. In general, the oxygen concentration at the entry to the reactor is ≤ 8 vol %, based on the acetic-acid-free mixture. It is essential to avoid gas mixtures capable of igniting; great care is taken in mixing in oxygen and measuring the oxygen concentration. If the oxygen stream is switched off, the inlet line must be flushed with nitrogen immediately to avoid back-diffusion of the circulating gas. The mixing chamber is usually installed behind concrete walls. The heat of reaction is removed in the form of steam via (c) from the tube reactor (b) by evaporative cooling in the shell side of the reactor.

The reaction temperature is adjusted by the pressure of the boiling water. The steam formed can be used within the plant itself in the work-up section. The heat of reaction is ca. 250 kJ/mol based on vinyl acetate, because of the simultaneous formation of CO_2. Pressurized water in a circulation loop is used for cooling in some plants.

Ethylene conversion is 8 – 10 %, and that of acetic acid 15 – 35 %. Oxygen conversion can be up to 90 %. As small quantities of alkali metal salt on the catalyst migrate under the reaction conditions, traces of alkali metal salt are mixed with the gas at the entry to the reactor [38].

The gas mixture leaving the reactor is first cooled in (d) in countercurrent with the cold recycle gas, which is thus warmed. There is virtually no condensation of the acetic acid, vinyl acetate, or water. The dew point is generally not reached. The gas mixture is then led into the predehydration column (l) and then cooled to about room temperature in (e). The liquid product consists of an acetic-acid-free mixture of vinyl acetate and water. The mixture is separated in a phase separator (m) into an aqueous phase, which is removed, and an organic vinyl acetate phase, which is recycled to the head of the predehydration column.

Between 40 and 50 % of the water formed in the reaction is removed in this way without the need to supply extra energy; this quantity of water does not need to be removed in the subsequent distillation of the crude vinyl acetate. Most of the energy consumed in the distillation is used for water removal. Crude vinyl acetate, which is low in water, collects in the sump of the predehydration column. Older plants do not have this column [39]. The noncondensed vinyl acetate fraction is washed out of the circulating gas with acetic acid in column (f) [40]. The remaining gas is recycled via compressor (g), after addition of fresh ethylene.

To remove the CO_2 formed in the reaction, a partial stream of recycle gas is first washed with water in column (h) to remove the remaining acetic acid. The CO_2 is then absorbed with potash solution in column (i). The potash solution is regenerated by depressurizing to normal

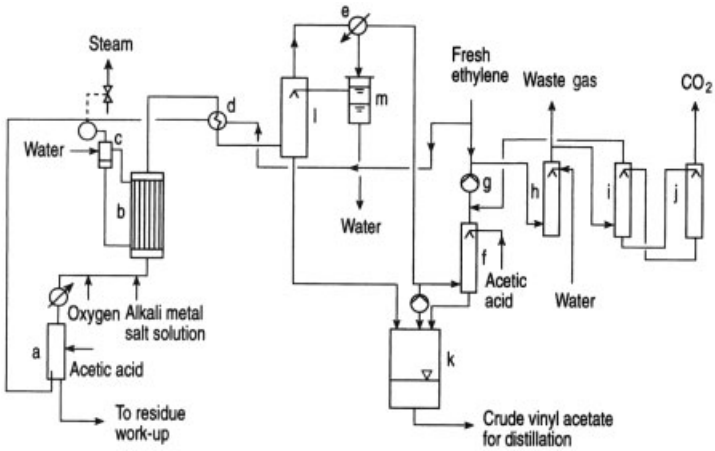

Figure 2. Ethylene gas phase process for vinyl acetate production; reaction section
a) Acetic acid evaporator; b) Reactor; c) Steam drum; d) Countercurrent heat exchanger; e) Water cooler; f) Recycle gas washing column; g) Recycle gas compressor; h) Water wash; i) Potash wash; j) Potash regeneration; k) Crude vinyl acetate collector; l) Predehydration column; m) Phase separator

pressure and boiling (j). Depending on the quantity of CO_2 formed in the reactor, the desired CO_2 content of the circulating gas can be adjusted by altering the quantity of circulating gas present in the circulating gas wash, and the degree of absorption in the potash wash. The CO_2 concentration is generally 10 – 30 vol % [41], [42]. It is also possible to perform water and CO_2 washes in the main gas stream [43].

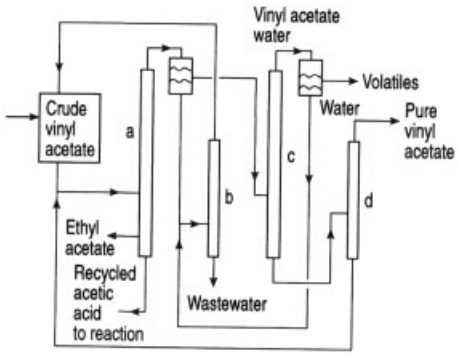

Figure 3. Ethylene gas-phase process for vinyl acetate production; work-up of crude vinyl acetate
a) Azeotrope column; b) Wastewater column; c) Drying/lightends column; d) Pure vinyl acetate column

To remove inert gases (nitrogen, argon) mainly brought in with the oxygen, a small quantity of waste-gas is removed before the CO_2 absorption column (i) and then incinerated. In some plants, part of the ethylene contained in this waste-gas is recovered by additional purification to reduce ethylene loss.

The liquid products formed, i.e., the condensate from the sump of the predehydration column (l) and the sump of the circulating gas wash, are depressurized to almost normal pressure and drained off into a collector for crude vinyl acetate (k). The circulating gas portions dissolved under the pressure of the circulating gas system are degassed and recycled to the circulating gas system after compression.

Process Description: Distillation Section. Distillative work-up of the liquid products to give acetic acid (which is recycled) and pure vinyl acetate is carried out in various ways, depending on the location of the plant and on the relative importance of energy consumption and investment costs. Besides the systems shown in Figures 3 and 4, combinations of both versions are used.

The liquid products contain 20 – 40 wt % vinyl acetate and ca. 6 – 10 vol % water. The rest consists of acetic acid and small quantities of byproducts, e.g., ethylidene diacetate and ethyl acetate. To work up the liquid products (Fig. 3) a vinyl acetate – water mixture can be distilled at the head in a first distillation (a). This mixture separates into two phases. The dissolved vinyl acetate is distilled from the water phase in wastewater column (b). The remaining water is wastewater. The organic vinyl acetate phase

is freed as the top product in a second distillation (c) from dissolved water, other volatile products, and acetaldehyde, formed by vinyl acetate hydrolysis. The sump product is dry vinyl acetate, which is distilled in a third distillation column (d) to give pure vinyl acetate as the top product.

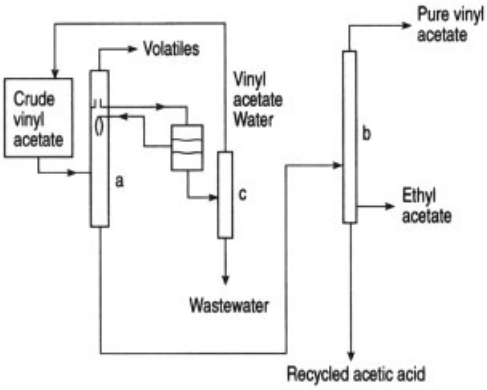

Figure 4. Ethylene gas-phase process for vinyl acetate production; variant for work-up of crude vinyl acetate
a) Dehydration column; b) Pure vinyl acetate column; c) Wastewater column

To remove polymers, a small partial stream is removed from the sump of the third distillation column and fed back to the first column. Thus all nonvolatiles and polymers produced in the distillative work-up are contained in the sump of the first column, together with the recycled acetic acid. To remove the polymers and nonvolatiles, a partial stream is removed from the acetic acid evaporator in the reaction section of the plant. From this, the acetic acid for recycling is distilled so that the residue, which is still flowable, can be incinerated.

The small quantities of ethyl acetate formed are removed through a side exit in the first column (a) as a mixture with acetic acid, water, and vinyl acetate [44].

If an additional column is used for the work-up, only the dissolved water is distilled over in the second distillation, in the third the light ends, and in the fourth the pure vinyl acetate [45].

In another version (Fig. 4) the water contained in the crude vinyl acetate is removed as an azeotrope with the vinyl acetate together with volatile products, e.g., acetaldehyde, in a first distillation column (a), which operates at increased pressure. The dry bottom product, which contains vinyl acetate, acetic acid, polymers, and nonvolatiles, is fractionated in a second column (b) into pure vinyl acetate as the top product, and acetic acid and nonvolatiles. The last two are recycled. Ethyl acetate is removed through a side exit in the second distillation column as a mixture with acetic acid and vinyl acetate [46], [47].

To avoid polymer formation during the distillative workup of the crude mixture, polymerization inhibitors, e.g., hydroquinone, benzoquinone, or *tert*-butylcatechol, must be added. Passing in oxygen-containing gases is also said to inhibit polymerization [48].

Process Data. The catalysts used always contain palladium as the metal or a salt, alkali metal salts, and additional activators, e.g., metallic gold or alkali acetoaurate, cadmium acetate, or noble metals of the platinum group. Silicic acid of various structures, aluminum oxide, spinel, or activated charcoal are mainly used as the carrier material [49–54], [55].

The space–time yield for vinyl acetate is $200\,g\,L^{-1}h^{-1}$, in older plants up to more than $1000\,g\,L^{-1}h^{-1}$, depending on the catalyst and the plant layout. The life time of the catalyst is ≤ 4 a.

The reaction pressure is 0.5–1.2 MPa overpressure. The space–time yield for vinyl acetate increases with the reaction pressure and with the oxygen concentration in the reaction gas. However, an increase in pressure shifts the flammability limits of ethylene–oxygen to lower oxygen concentrations, reducing the quantity of oxygen available, and consequently the quantity of vinyl acetate formed, so pressure limits are set; higher pressures also raise equipment costs.

The reaction temperature is generally $> 140\,°C$. It increases to $> 180\,°C$ towards the end of the catalyst life. A lower reaction temperature results in the formation of less CO_2, but then the heat produced in the reactor can no longer be used in the plant.

The gas loading of the catalyst is $2-4\,m^3$ (STP) per liter catalyst per hour. The gas mixture contains 10–20 mol % acetic acid, 10–30 % CO_2, and ca. 50 % ethylene. The maximum oxygen content is ca. 1.5 % below the flammability limit, which varies with the composition of the gas mixture and the reaction conditions. The nitrogen and argon contents are adjusted according

to the quantity of waste gas. They are generally ca. 10 %, but depend on the purity of the oxygen used.

For old plants, energy consumption is ca. 3 t of heating steam per tonne of vinyl acetate produced. As a result of process improvements, modern plants have a heating steam consumption of 1.2 t per tonne of vinyl acetate.

The yields are up to 99 % based on acetic acid, and up to 94 % based on ethylene, if the acetaldehyde, formed in small quantities by hydrolysis of vinyl acetate during the distillative work-up, is included in the yield.

The process has not yet posed any environmental problems. Volatile and nonvolatile liquid products are incinerated. The water produced in the reaction can contain traces of acetic acid formed by hydrolysis of vinyl acetate in the wastewater column. It is subjected to biological wastewater treatment. To remove inert gases brought in with the oxygen, some of the waste-gas containing nitrogen and argon is burned after partial recovery of the ethylene it contains. There are small quantities of residual ethylene in the CO_2, formed in the regeneration column of the potash absorber. Removal of ethylene, e.g., by subsequent catalytic incineration, is now necessary in Germany because of more stringent regulations. In addition, special measures are required for waste-gas incineration.

Plants are generally made from stainless steel 316 L, apart from the potash washer, which is made from normal steel or stainless steel 321.

Proposals for Process Improvement. Since the first plants for the ethylene gas-phase process were commissioned, numerous suggestions have been made for improving the process; some of these are now being introduced industrially. Proposals regarding the catalyst include: simplification of production methods [56]; activity improvement (increase in vinyl acetate space – time yield [57–60], [55], [61]; improvement in selectivity (less CO_2 formation) [62], [63]; increase in life span [64–66]; reduction in flow resistance in the reactor (form of carrier material) [67]; better utilization of the noble metal content (distribution on the carrier) [68]; regeneration of spent catalysts [69–74].

Proposals regarding the mode of operation include: increasing the vinyl acetate space – time yield by raising the oxygen concentration in the gas at the entry to the reactor by adding desensitizers [75]; facilitating separation of ethyl acetate from vinyl acetate [76–79]; increasing the selectivity of the reaction by adding other components to the reaction gas [80]; carrying out the reaction on a fluidized bed [81]; lowering the energy consumption, e.g., by making use of the heat of condensation of the gas coming out of the reactor for removal of the water formed (see Section 2.2) [39]; simplified work-up of the polymer-containing residue [82].

2.2.4. Reaction of Methyl Acetate with CO and H_2.

Several companies have attempted to develop a reaction which, in principle, has been known for a long time. It involves the conversion of methanol, acetic anhydride, dimethyl ether, or preferably methyl acetate to ethylidene diacetate with carbon monoxide and hydrogen [83–95], [96–98]:

$$2\,CH_3CO_2CH_3 + 2\,CO + H_2 \rightarrow (CH_3CO_2)_2CHCH_3 + CH_3CO_2H$$

The reaction takes place in the liquid phase at ca. 7 MPa and ca. 150 °C. Rhodium salts together with methyl iodide and amines can be used as catalyst. The ethylidene diacetate formed can be cleaved to vinyl acetate and acetic acid, using the process previously employed by Celanese (see Section 2.2). As the acetic acid formed in the process can be reesterified to methyl acetate using methanol, the process essentially involves production of vinyl acetate from methanol and synthesis gas:

$$2\,CH_3CO_2CH_3 + 2\,CO + H_2 \rightarrow (CH_3CO_2)_2CHCH_3 + CH_3CO_2H$$

$$(CH_3CO_2)_2CHCH_3 \rightarrow CH_3CO_2CH{=}CH_2 + CH_3CO_2H$$

$$2\,CH_3OH + 2\,CH_3CO_2H \rightarrow 2\,CH_3CO_2CH_3 + 2\,H_2O$$

Dimethyl ether can also be used to prepare methyl acetate:

$$CH_3OCH_3 + CO \rightarrow CH_3CO_2CH_3$$

It is said to be possible to produce dimethyl ether by partial oxidation of natural gas:

$$2\,CH_4 + O_2 \rightarrow CH_3OCH_3 + H_2O$$

Byproducts are formed in all the steps; recovering them in pure form is expensive [99–102].

The process has not yet been used in industry. It could become more important, depending on the further development of raw material prices.

2.3. Quality Specifications, Analysis,Storage, Transport, and Toxicology

Quality. The specifications of individual producers often differ only slightly with regard to the content of impurities. One specification requires: vinyl acetate ≥ 99.9 wt %, acid (as acetic acid) ≤ 0.005 wt %; carbonyl (as acetaldehyde) ≤ 0.02 wt %, water ≤ 0.04 wt %; distillation range 95 % at 72 – 73 °C within 0.5 °C; peroxide free; polymer free; capacity for polymerization (see Section 2.4)

Analysis. The following methods are used for quality testing:

Carbonyl groups: titration of the hydrochloric acid formed after conversion to the oxime using hydroxylamine hydrochloride solution
Water content: titration with Karl-Fischer solution
Acid content: neutral to litmus
Peroxide: test with potassium iodide solution
Polymer content: mixtures with petroleum ether must not produce turbidity or precipitation of solids
Polymerization test: carried out differently by individual producers; the time taken for the onset of polymerization following the addition of a defined quantity of dibenzoyl peroxide and warming to a defined temperature

Storage. As a flammable liquid, vinyl acetate is assigned to VbF, group A, class 1, and ignition group T 2 according to VDE 0165. It can be stored in steel, aluminum, or stainless steel containers under nitrogen. It is not necessary to add stabilizers at lower temperatures. If the vinyl acetate is to be warmed, stabilizers, such as hydroquinone, hydroquinone monomethyl ether, or diphenylamine are added. The quantity of stabilizer used is small, e.g., 3 – 20 ppm hydroquinone, so that it does not generally need to be removed during the later polymerization.

Table 1. World vinyl acetate capacities

Country	Process	Capacity, 10^3 t/a
United States	Ethylene Bayer – Hoechst and National Distillers	1429
Canada	Ethylene Bayer – Hoechst	54
Mexico	Ethylene Bayer – Hoechst	100
South America	Acetylene, Ethylene National Distillers	100
Australia	Ethylene Bayer – Hoechst	14
Germany	Ethylene Bayer – Hoechst	280
France	Acetylene	150
Spain	Ethylene Bayer – Hoechst, Acetylene	100
United Kingdom	Ethylene Bayer – Hoechst	100
Italy	Acetylene	60
Eastern Europe	Ethylene Bayer – Hoechst, Acetylene	288
Japan	Ethylene Bayer – Hoechst and National Distillers, Acetylene	612
Near East	Ethylene Bayer – Hoechst and National Distillers	90
India	Ethylene Bayer – Hoechst	50
Korea	Ethylene Bayer – Hoechst	150
Taiwan	Ethylene Bayer – Hoechst	100
China	Ethylene Bayer-Hoechst, Acetylene	224

Transport. Vinyl acetate is transported in practically all quantities: small containers, drums, tankers, tank cars, and ships. Stabilizers are usually added for transportation.

Transport Regulations
EEC guidelines: No. 607-023-00-0
Land transport ADR/RID and GGVS/GGVE
GGVS/GGVE class 3, cipher 3 b
ADR/RID class 3, cipher 3 b
Warning notice no. 339, substance no. 1301
Symbol F flammable liquid
Sea transport IMDG/UN class 2, U.N. no. 1086
IMOG/UN class 3.2 and no. 1301 II
EmS no. 3-07 MFAG 330
Symbol F flammable liquid/3
Air transport ICAO/IATA-DGR
ICAO/IATA-DGR class 3, UN no. 1301 PG II
Dispatch symbols FL: Double arrows FL flammable liquid/3
Packaging regulations
PAC: 305 max. net 5 L
CAO: 307 max. net 60 L
Air mail: not permitted

Safety Data

CAS [*108-05-4*]	
Flashpoint	– 8 °C
Ignition temperature (DIN)	385 °C
Hazard class (VbF)	AI
Explosion limits in air	
lower 2.6 vol %	
upper 13.4 vol %	
Explosion group (DIN)	IIA
Exothermic decomposition dTA	from 500 °C
Risk assessment depending on the European Seveso guidelines no. 2	
Emergency Response Planning Guidelines	75 ppm (ERPG-2)

Toxicological Data

Acute toxicity	
LD_{50}, oral, rat	2920 mg/kg
LC_{50}, inhalation, rat	14.2 mg/L × 2 h
LD_{50}, dermal, rabbit	2335 mg/kg

Health hazards
F easily inflammable
R 11
S 16-23-29-33
Classification according to the regulations of hazardous chemicals/TRGS 900
Carcinogenic: category K 3
Teratogenic: no (TRGS 905)
Alters genotype: no (TRGS 905)
Skin resorption: no
Sensitization: no
Effect on skin: slightly irritant
Other health hazards, e.g., possible damage to organs; narcosis; damage to liver and kidneys; 4000 ppm causes death in rats in 4-h inhalation test [103].

Protection at Work

MAK/TKR value	10 ppm 35 mg/m^3
Peak limiting category	I
Warning symptoms: Sweetish, not unpleasant odor	
Odor threshold	1 mg/m^3
Odor perception threshold	1.8 mg/m^3

Compatibility with water

Water hazard class	2
Fish toxicity LC_{50}	20 mg/L
Bacteria toxicity	100 – 1000 mg/L
Biodegradability	>70 % good
Toxicity rating number	fish 4.6; bacteria 5.2; mammals 1

2.4. Use, Economic Importance

Vinyl acetate is used mainly for the production of polymers and copolymers, e.g., for paints (mainly dispersions), adhesives, textile and paper processing, chewing gum [→ Poly(Vinyl Esters), and for the production of poly(vinyl alcohol) (→ Paints and Coatings, Chap. 2.4.5.) and polyvinylbutyral (→ Paints and Coatings, Chap. 2.4.6.).

Vinyl acetate – ethylene copolymers (→ Poly(Vinyl Esters), Chap. 6.) are processed to give resins, paints (→ Paints and Coatings, Chap. 2.4.4.), and sheeting. Floor coverings and gramophone records are made from vinyl acetate – vinyl chloride copolymers. Vinyl acetate is also used in small quantities as a comonomer in polyacrylic fiber production.

Uses differ according to the region. In Japan, ca. 70 % of vinyl acetate is used in the production of poly(vinyl alcohol), while in the United States and Europe more than half is processed to give poly(vinyl acetate).

The total capacity for vinyl acetate production was $< 10^6$ t/a in 1965. In 1994 it was ca. 3.8×10^6 t/a. The rapid increase has been promoted significantly by the development of the ethylene gas-phase process. After ethylene had become one of the cheapest raw materials in the chemical industry, it was obvious that vinyl acetate should be produced from ethylene instead of acetylene. The vinyl acetate capacities of individual countries are listed in Table 1.

A plant with a capacity of 150 000 t/y is to be commissioned by Hoechst Celanese in South East Asia in 1997.

The economic optimum of individual process variants depends on location. Investment costs, material consumption (acetic acid, ethylene, acetylene), and energy consumption need to be compared. The weighting depends on the prices of energy and starting materials, and on official regulations specific to a particular country.

3. Vinyl Propionate

3.1. Properties

Physical Properties. Vinyl propionate [*105-38-4*], $CH_3-CH_2-COO-CH{=}CH_2$, M_r 100.1, is a clear, colorless, flammable liquid with an ester-like odor; *bp* 95 °C; d_4^{20} 0.917; n_D^{20} 1.404; solidification point – 80 °C;

viscosity 0.56 mPa · s at 20 °C; heat of combustion 26 922 kJ/kg; flashpoint according (Abel-Pensky) − 2 °C; specific heat (liquid) 2.12 kJ/kg; heat of evaporation 340 kJ/kg at *bp*; heat of polymerization 882 kJ/kg; vapor pressure 1.34 kPa at 0 °C, 6.45 kPa at 20 °C; explosion limits (vapor in air at 20 °C) 2.4 – 15.5 vol %; ignition temperature 360 °C; solubility in water 6.5 mL/L at 25 °C; solubility of water in vinyl propionate 8.5 mL/L at 25 °C.

Chemical Properties. Vinyl propionate closely resembles vinyl acetate in its chemical reactions and on polymerization. The greatest differences are the more difficult saponification and its activity as an internal plasticizer in copolymers.

3.2. Production

Vinyl propionate is produced on an industrial scale at BASF from acetylene and propionic acid in the gas phase. The catalyst is zinc propionate on activated charcoal.

Process Description (Fig. 5). The recycle gas from blower (a) with an overpressure of up to 40 kPa is driven through the heat exchangers (b), connected in series, where it is heated to ca. 160 °C. It consists of ca. 95 % acetylene together with inert gases and traces of vinyl propionate, propionic acid, and other reaction products. After the second heat exchanger, propionic acid vapor from (j) is mixed into the circulating gas, and the resulting gas mixture is heated to the reaction temperature in (c). Depending on the activity of the zinc propionate catalyst, the temperature in the contact reactors (d) is 160 – 190 °C. The heat of reaction is removed via a cooling medium (l) by means of evaporative cooling, and is used for steam production in (k). The reaction gases from the reactor outlet (d) are cooled to 80 °C in the heat exchangers (b). The reaction products and unreacted propionic acid are condensed by further cooling in (e), (b), and (f). The recycle gas is washed with propionic acid in (g) to remove vinyl propionate and acetaldehyde. A small quantity is removed via the waste – gas scrubber (h).

The crude mixture in (i) contains vinyl propionate, propionic acid, and, besides dissolved acetylene, small quantities of the byproducts acetaldehyde, crotonaldehyde, diethyl ketone, and propionic anhydride. The latter is hydrolyzed by addition of water. Diethyl ketone and crotonaldehyde are removed by special methods. Water, acetaldehyde, and acetylene are removed at the head of the lightends column.

The sump product is pumped to the purification column where pure vinyl propionate is obtained at the head. The sump product here is at a temperature of 150 °C, and contains propionic acid and higher-boiling products. The propionic acid is recovered 99.5 % pure in an acid recycling column.

The yield is over 98 % based on acetylene, and over 99 % based on propionic acid.

Quality. The following quality specifications are in use:

Purity (by GLC)	99.9 %
Ester content (from saponification number)	99.8 %
Water content	≤ 0.1 %
Acid content (calculated as propionic acid)	≤ 0.03 %
Color value	≤ 10 APHA

3.3. Use

Vinyl propionate is the starting material for the production of homopolymers and copolymers with common monomers, such as (meth)acrylic compounds, vinyl chloride, and other vinyl esters [→ Poly(Vinyl Esters), Chap. 6.]. The polymers are marketed as Propiofan brands (BASF).

4. Vinyl Esters of Higher Carboxylic Acids

Apart from vinyl acetate and vinyl propionate, the vinyl esters of only a few higher carboxylic acids are important (see Section 4.5.). The properties of polymers, e.g., poly(vinyl acetate), poly(vinyl chloride), polystyrene, and esters of poly(acrylic acid), are modified using these compounds as comonomers. With this type of modification paints hydrolyze less readily and absorb less moisture. The various production methods depend on the physical and chemical properties of the carboxylic acids to be vinylated.

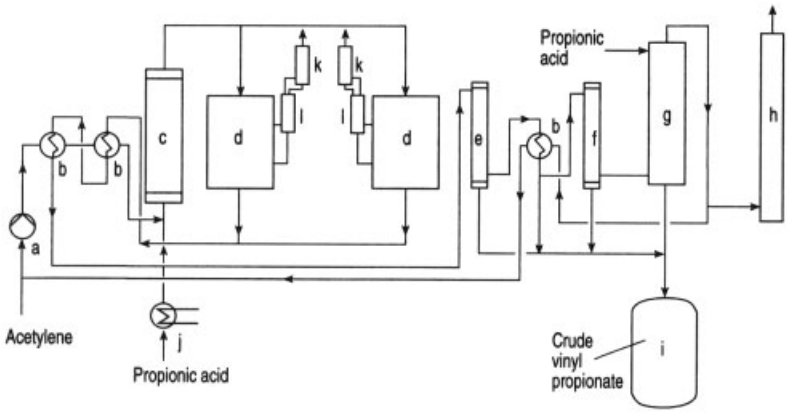

Figure 5. Production of vinyl propionate (BASF)
a) Recycle gas blower; b) Heat exchangers; c) Superheater; d) Reactors; e) Condenser; f) Brine-cooled condenser; g) Recycle gas wash; h) Wastegas wash; i) Crude product receiver; j) Propionic acid evaporator; k) Cooling medium; l) Steam generator

4.1. Reaction with Acetylene in the Gas Phase

The reaction of a thermally stable higher carboxylic acid with acetylene in the gas phase is possible only if the *bp* is sufficiently low for it to be evaporated in an acetylene stream below the reaction temperature. Catalysts containing zinc or cadmium are suitable for the vinylation itself. The production process generally does not run as smoothly as that for vinyl acetate because the zinc salts of higher carboxylic acids, particularly the branched ones, are volatile and migrate from the catalyst more readily. This problem can be circumvented to a certain extent by modification of the catalyst [104].

4.2. Reaction with Ethylene and Oxygen

This process, analogous to that used for the production of vinyl acetate, has not been used industrially, although it is possible in principle [105]. As the molecular mass of the carboxylic acid increases, the quantity required per mole of vinyl ester also increases; the higher equipment expenditure for the ethylene process is no longer compensated for by the cheapness of ethylene compared with acetylene.

4.3. Transvinylation and Other Processes

The transvinylation of the corresponding carboxylic acid with vinyl acetate can be applied to almost all carboxylic acids.

$$CH_3CO_2CH{=}CH_2 + RCO_2H \rightleftharpoons RCO_2CH{=}CH_2 + CH_3CO_2H$$

Mercury, palladium, or silver salts are suitable catalysts [106–109]. Methods involving mercury salts are appropriate only for very special and expensive vinyl esters because of the associated work-up costs. The disadvantage of the transvinylation method lies in the fact that a large excess of vinyl acetate must be used. Also, working up the mixtures of vinyl acetate, acetic acid, the carboxylic acid, and its vinyl ester by distillation is not always straightforward. Vinyl esters of higher carboxylic acids have therefore been used for transvinylation instead of vinyl acetate.

In special cases, the reaction of a carboxylic anhydride with acetaldehyde or paraldehyde, and subsequent cleavage of the ethylidene dicarboxylate have proved successful [111]. This is analogous to the production of vinyl acetate from ethylidene diacetate. The dehydration of ethylene glycol monocarboxylates has also been proposed [112].

4.4. Reaction with Acetylene in the Liquid Phase (Wacker and Shell Processes)

The vinylation of thermally stable carboxylic acids can also be achieved by passing acetylene into a melt formed by heating the corresponding zinc carboxylate or a zinc carboxylate – carboxylic acid mixture to ca. 200 °C (Wacker and Shell processes). The reaction can also be carried out with acetylene – nitrogen mixtures at pressures of up to 2.5 MPa [112], [113]. Carboxylic anhydrides may also be present [114].

If the vinyl esters of the higher carboxylic acids are thermally unstable, the reaction with acetylene can be carried out in the presence of mercury catalysts, enabling the reaction temperature to be lowered to 60 – 70 °C.

With the Wacker and Shell processes, most vinyl esters of higher saturated carboxylic acids up to ca. C_{18} can be produced in the same plant. The Shell process is particularly suitable for vinylating branched higher carboxylic acids [115–117]. In the Wacker process, excess acetylene is passed through a melt consisting of the zinc salt of the corresponding carboxylic acid and the acid itself at ca. 200 – 220 °C. In the Shell process, the reaction mixture contains a high-boiling mineral oil fraction. The vinyl ester formed is carried out with the excess acetylene.

Process Description. A recycle acetylene gas stream is preheated and then passed into the vinylation reactor. This reactor can be a short bubble column, a normal stirred vessel, or a stirred vessel with a gas entrainment stirrer because of the low initial pressure of the acetylene. Trickle bed columns can also be used. The reactor is heated or cooled by a heat transfer oil in the jacket. It contains a melt of the carboxylic acid to be vinylated mixed with its zinc salt (Wacker) or a solution of the mixture in a solvent with $bp > 300$ °C, e.g., a mineral oil fraction with a low aromatic content (Shell). The reactor must be heated or cooled depending on the mode of operation and the nature of the carboxylic acid. At equilibrium, it is also possible to use adiabatic conditions. The reaction temperature is 200 –230 °C and depends to some extent on the carboxylic acid used. The vinyl ester formed and the unreacted carboxylic acid are stripped with the acetylene stream. The conditions described for the vinyl acetate process regarding removal of inert gases and safeguarding against acetylene decomposition also apply here (see Section 2.2). The crude mixture, which condenses on cooling, contains up to 90 wt % vinyl ester. The rest consists of unreacted carboxylic acid, carboxylic anhydride, and small quantities of other byproducts. For high-boiling vinyl esters, the distillative work-up of the crude mixture is carried out in vacuo. The vinyl ester is distilled after removal of volatile products, e.g., acetaldehyde and crotonaldehyde. The sump product of the ester distillation, consisting of acid, acid anhydride, and high-boiling products is recycled to the reactor. The carboxylic anhydride formed can first be hydrolyzed thermally to the carboxylic acid. The hydrolysis can also be effected in the reactor itself by addition of small quantities of water, e.g., brought in with the acetylene as steam. Addition of water is not necessary with α-branched carboxylic acids. To avoid enrichment of high-boiling products, a small partial stream of the reactor liquid is removed [118], [119].

A large number of vinyl esters have been produced by the methods described [120]. They are not quite as pure as vinyl acetate or vinyl propionate because the starting acids are less pure than acetic or propionic acid. Plants for the vinylation of higher carboxylic acids are operated by Wacker in Burghausen and by Shell in Moerdijk and Pernis (Holland). Only a few of these vinyl esters are still produced in large quantities. For data on other vinyl esters which have been produced in the past see [121].

4.5. Individual Vinyl Esters

Vinyl laurate [*2146-71-6*] (Wacker), M_r 226.3, colorless, oily liquid with a weak odor, purity 99.9 ± 0.1 %, acid number 1 ± 0.2 mg KOH/g, density at 20 °C 0.87 g/mL, solidification point 6.5 °C, refractive index $n_D^{20} = 1.437$, viscosity 2.617 mPa · s at 20 °C, heat of evaporation 146 kJ/kg at 5.5 kPa, vapor pressure 10 Pa at 50 °C, bp (550 Pa) 123 °C/(101.3 kPa) 254 °C, flashpoint 136 °C, ignition temperature 300 °C, lower explosion limit in air 0.5 % (calculated) [114], solubility in water < 1 g/L, solubility of water in the ester < 1 g/L. The ester is not stabilized during transport or storage. Biodegradabil-

ity, good; water hazard class WGK 2. Transport: not classed as hazardous goods.

Vinyl isononanoate (Wacker), M_r 184.3, colorless liquid with a weak, ester-like odor, acid content calculated as isononanoic acid $\leq 0.3\,\%$, carbonyl compounds calculated as acetaldehyde $\leq 0.1\,\%$, water content $\leq 0.05\,\%$, *bp* 145 °C at 100 kPa, 44 °C at 50 Pa, solidification point − 70 °C, density 0.872 g/mL at 20 °C, flashpoint 68 °C, ignition temperature 390 °C.

Vinyl Esters of Versatic 911, 10, and 1519 Acids (Shell). Versatic acids are produced by the Koch synthesis from olefins, carbon monoxide, and water (→ Carboxylic Acids, Aliphatic, Chap. 10.8.). 911 refers to mixtures with 9 – 11 carbons, 1519 to mixtures with 15 – 19 carbons. The vinyl ester of versatic acid 10 with general formula

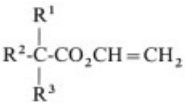

is the most important. R^1, R^2, and R^3 are straight-chain alkyl groups, one of which is always a methyl group.

The vinyl ester of versatic 10 acid, $C_{12}H_{22}O_2$, M_r 198, is a colorless liquid with a slight, ester-like odor, color value ≤ 15 APHA.

Physical Data, Properties

Density at 20 °C	0.88 g/mL
Refractive index, n_D^{25}	1.435
Kinematic viscosity:	
at 20 °C	2.2 $mm^2\,s^{-1}$
at 40 °C	1.3 $mm^2\,s^{-1}$
Vapor pressure:	
at 30 °C	130 Pa
at 110 °C	4.25 kPa
at 210 °C	101 kPa
Solubility in water	< 0.1 wt %
Solubility of water in the ester	0.05 wt %
Acid number	≤ 1
Bromine number	80 g Br/100 g
Heat of evaporation	49 kJ/mol
Heat of polymerization	96.3 kJ/mol
Specific heat	1.97 J/g
Boiling range	133 – 136 °C at 13.3 kPa
Ignition temperature	309 °C
Solidification point	− 20 °C
TA-Luft class III	
VbF AIII	

Regulations governing hazardous substances: not hazardous as understood by these regulations. *Toxicology*

LD_{50}, oral, rat	> 8800 mg/kg
LD_{50}, dermal, rat	> 3500 mg/kg
LD_{50}, inhalation, rat, 4 h	> 2.6 mg/L
Ecology	
Fish toxicity	
(rainbow trout) LC_{50}, 96 h	14 mg/L
Daphnia toxicity LC_{50}, 48 h	110 mg/L
Green algae LC_{50}, 96 h	31 mg/L
Not readily biodegradable	

5. References

General References

1. Beilstein, II_1 p.63, II_2 p.147, II_3p.268.
2. C. E. Schildknecht: *Vinyl and Related Polymers,* J. Wiley, New York 1952.
3. R. P. Arganbright, R. I. Evans: "Vinyl Acetate Still Growing Fast," *Hydrocarbon Process. Pet. Refin.* **43** (1964) no. 11, 149 – 164.
4. R. B. Stabaugh, W. C. Allen, Jr., R. H. Van Sternberg: "Vinylacetate: How, Where, Who – Future," *Hydrocarbon Process.* **51** (1972) no. 5, 153 – 161.
5. W. Schwerdtel, *Chem. Ing. Tech.* **40** (1968) 781 – 784.

Specific References

6. Rhône Poulenc, DE 531 336, 1929.
7. Chem. Fabrik Griesheim; DE 313 696, 1915.
8. I. G. Farbenind., DE 712 373, 1938 (P. Kurtz) Du Pont, US 2 397 341, 1940 (E. K. Ellingboe).
9. N. A. Miles, H. S. Mason, S. Sussmann, *J. Am. Chem. Soc.* **61** (1939) 1844 – 1847.
10. K. Alder, H. F. Rickert, *Justus Liebigs Ann. Chem.* **543** (1940) 8, 21.
11. R. L. Adelmann, *J. Org. Chem.* **14** (1949) 1057.
12. Consortium Elektrochem. Ind. DE 515 307, 1927 (H. Deutsch, W. O. Herrmann).
13. Consortium Elektrochem. Ind., DE 555 224, 1927 (H. Deutsch, W. O. Herrmann). Hoechst, DE 1 618 426, 1967 (G. Roscher, H. Schmitz).
14. Bayer, DE-AS 1 283 226, 1967 (W. Krönig, G. Scharfe).
15. Snamprogetti, DE-OS 2 521 470, 1975 (L. Re, V. Caciagli, G. E. Bianchi).
16. Hoechst, DE 2 127 821, 1971 (M. Finke, H.-J. Kleiner).
17. Chem. Fabrik Griesheim, DE 271 381, 1912.
18. *Ullmann,* 3rd ed., **18,** 77 – 78.

19. *Ullmann,* 4th ed., 23, 598.
20. Consortium Elektrochem. Ind., DE 403 784, 1921 (E. Baum, H. Deutsch, W. O. Herrmann, M. Mugdan).
21. R. P. Arganbright, R. H. Evans, *Hydrocarbon Process. Pet. Refin.* **43** (1964) no. 11, 159 – 164.
22. Monsanto, GB 651 072, 1947 (W. H. Gehrke).
23. *Pet. Refin.* **38** (1959) no. 11, 304.
24. K. Gerlach, G. Emig, H. Hofmann, *Chem. Ing. Tech.* **49** (1977) 969 – 970.
25. S. Sacharjan, *Chem. Tech. (Leipzig)* **20** (1968) 617, 620.
26. Celanese, US 2 859 241, 1956 (A. W. Schnitzer); US 2 425 389, 1945 (H. F. Oxley, E. B. Thomas). *Pet. Refin.* **40** (1961) no. 11, 308.
27. J. J. Moiseew, M. N. Vargaftik, J. K. Syrkin, *Dokl. Akad. Nauk SSSR* **133** (1960) 377 – 380.
28. ICI, GB 964 001, 1960 (D. Clark, W. D. Walsh, W. E. Jones).
29. Hoechst, DE-AS 1 191 366, 1961 (W. Riemenschneider, T. Quadflieg).
30. Union Carbide, DE 629 885, 1962 (J. E. McKeon, P. S. Starcher).
31. *Eur. Chem. News* **51** (1973) Feb. 2, 22.
32. *Ullmann,* 3rd ed., **suppl. vol.,** 137 – 139.
33. Bayer, BE 627 888, 1962 (H. Holzrichter, W. Krönig, B. Frenz).
34. Knapsack, DE-AS 1 244 766, 1965 (K. Sennewald, W. Vogt, H. Glaser).
35. Hoechst, DE-AS 1 296 138, 1967 (H. Fernholz, H.-J. Schmidt, F. Wunder).
36. S. Nakamura, T. Yasui, *J. Catal.* **23** (1971) 315 – 320.
37. Hoechst, DE-AS 1 191 366, 1961 (W. Riemenschneider, T. Quadflieg).
38. Bayer, DE 1 568 657, 1966 (K. Sennewald, G. Roscher, W. Krönig, W. Schwerdtel). DE 1 296 621, 1966 (W. Krönig, W. Schwerdtel, K. Sennewald, G. Roscher).
39. *Ullmann,* 4th ed. **23,** 602.
40. Bayer, DE 1 768 078, 1968 (W. Krönig, W. Schwerdtel).
41. Knapsack, DE 1 277 249, 1966 (K. Sennewald et al.).
42. Nat. Distillers, DE-AS 1 593 159, 1965 (I. L. Mador, I. L. McClain).
43. *Eur. Chem. News* **51** (1973) Feb. 2, 21 – 22.
44. Bayer, DE 1 768 412, 1968 (W. Krönig, W. Schwerdtel).
45. *Chem. Age Int.* 1973, (Feb.) 9, 15.
46. Hoechst, DE 1 668 063, 1967 (G. Roscher, H. Schmitz).
47. Hoechst, DE 1 282 014, 1967 (G. Roscher, K. Günther, O. Probst, J. Scholderer).
48. Bayer, DE 1 807 738, 1968, (W. Schwerdtel, W. Krönig, G. Scharfe, W. Swodenk).
49. Nat. Distillers, US 3 190 912, 1962 (R. E. Robinson).
50. Bayer, DE-AS 1 568 645, 1964 (W. Krönig, W. Schwerdtel).
51. Hoechst, DE-AS 1 296 138, 1967 (H. Fernholz, H.-J. Schmidt, F. A. Wunder).
52. Bayer, DE 1 185 604, 1962 (H. Holzrichter et al.).
53. Knapsack, DE 1 244 766, 1965 (K. Sennewald, W. Vogt, H. Glaser).
54. Hoechst, DE-OS 2 315 037, 1973 (H. Fernholz et al.).
55. UCC, US 5 185 308, 1991 (W.-J. Bartley, S. Jobson, G. G. Harkreador, M. Kitson, M. Lemanski).
56. Nat. Distillers, DE-OS 1 944 933, 1968 (D. Lum).
57. Knapsack, DE-AS 1 230 784, 165 (K. Sennewald et al.).
58. Nat. Distillers, DE-OS 1 944 933, 1968 (D. Lum).
59. Hoechst, DE-OS 2 100 778, 1971 (H. Fernholz, F. Wunder, H.-J. Schmidt).
60. Hoechst, DE-OS 2 315 037, 1973 (H. Fernholz et al.).
61. UCC, US 5 179 057, 1991 (W.-J. Bartley et al.).
62. Hoechst, DE-OS 2 057 087, 1970 (H. Fernholz, F. A. Wunder, H.-J. Schmidt).
63. Knapsack, DE 1 668 352, 1967 (K. Sennewald et al.).
64. Nat. Distillers, DE-OS 2 657 442, 1975 (R. C. Bartsch).
65. Nat. Distillers, DE-AS 1 817 213, 1967 (D. W. Lum).
66. Knapsack, DE 1 258 863, 1966 (K. Sennewald, W. Vogt, H. Glaser).
67. Hoechst, DE-OS 2 811 211, 1978 (F. A. Wunder, T. Quadflieg, G. Roscher).
68. Bayer, DE-OS 2 528 363, 1975 (W. Biedermann, H. Kösler, K. Wedenmeyer).
69. Nat. Distillers, DE-OS 1 667 206, 1966 (D. W. Lum, I. L. Mador).
70. Nat. Distillers, DE-AS 1 667 212, 1967 (F. D. Miller).
71. Nat. Distillers, US 3 634 497, 1969 (C. C. Budke).
72. Knapsack, DE 1 542 257, 1966 (K. Sennewald, W. Vogt, H. Glaser).
73. Hoechst, DE-AS 2 513 125, 1975 (H. Fernholz, H. Krekeler, H.-J. Schmidt, F. Wunder).
74. Nat. Distillers, DE-OS 2420 374, 1973 (S. Schott, D. W. Lum, I. L. Mador).

75. Celanese, DE-OS 2 361 098, 1972 (J. Severs).
76. Nat. Distillers, DE-OS 2 114 751, 1970 (J. Feldmann, F. Lerman, F. D. Miller).
77. Bayer, DE-AS 1 768 412, 1968 (W. Krönig, W. Schwerdtel).
78. Du Pont, US 3 591 463, 1968 (H. B. Copelin).
79. Celanese, DE-AS 1 618 240, 1966 (G. J. Fischer).
80. Wacker, DE-OS 3 036 421, 1980 (D. Dempf et al.).
81. Knapsack, DE 1 793 481, 1968 (K. Sennewald et al.).
82. Celanese, DE-OS 2 359 286, 1973 (G.-J. Fischer, E. N. Wheeler).
83. Halcon, DE-OS 2 610 035, 1975 (N. Rizkalla, C. N. Winnick).
84. Halcon, DE-OS 2 261 562, 1971 (R. Fernandez).
85. Union Oil, US 3 579 566, 1967 (D. M. Fenton).
86. Ajinomoto, DE-OS 2 016 061, 1969 (H. Wakamatsu, N. Yamagami, J. Furukawa).
87. Hoechst, DE-OS 2 450 965, 1974 (H. Kuckertz).
88. Halcon, DE-OS 2 610 035, 1975 (N. Rizkalla, C. N. Winnick).
89. Halcon, DE-OS 2 856 791, 1977 (C.-G. Wan).
90. Mitsubishi Gas, EP 0 025 702, 1979 (T. Isshiki, Y. Kijima, Y. Miauchi, T. Jasunaga).
91. Hoechst, DE-OS 2 941 232, 1979 (H.-K. Duebbeler, H. Erpenbach, K. Gehrmann, G. Kohl).
92. Mitsubishi Gas, EP 0 028 447, 1979 (K. Jasuhiko, J. Akira, J. Kenji).
93. Mitsubishi Gas, EP 0 028 515, 1979, T. Issihiki, J. Kijima, T. Kondoh).
94. Mitsubishi Gas, EP 0 034 062, 1980 (T. Issihiki, J. Kijima, T. Watanabe).
95. Mitsubishi Gas, EP 0 035 860, 1980 (T. Issihiki, A. Ito, J. Kijima, T. Watanabe).
96. Halcon, US 4 341 741, 1981 (W. C. Davidson).
97. Shell, EP 58 442, 1981 (E. Drent).
98. Daicel, JP 100 135, 1979 (N. Negawa, T. Yokoyama).
99. Halcon, US 267 960, 1981 (R. V. Porcelli).
100. Hoechst, DE 134 347, 1981 (H. Erpenbach, K. Gehrmann, W. Lork, P. Prinz).
101. Halcon, US 241 180, 1981 (W. C. Davidson, R. V. Porcelli).
102. Daicel, JP 177 671, 1984 (Y. Harano).
103. *Ullmann,* 4th ed., **23,** 615.
104. Hoechst, DE 1 233 387, 1964 (G. Roscher, W. Riemenschneider).
105. Bayer, DE 1 568 657, 1966 (W. Krönig, W. Schwerdtel, K. Sennewald, G. Roscher).
106. R. L. Adelmann, *J. Org. Chem.* **14** (1949) 1057, 1073.
107. Consortium Elektrochem. Ind., DE 654 282, 1938 (W. O. Herrmann, E. Baum, W. Haehnel).
108. Consortium Elektrochem. Ind., DE 1 127 888, 1960, (J. Schmidt, A. Abel).
109. Sekisui Chem. Ind. JP 19 146/64, 1961 (S. Jwai, Y. Ohzuka).
110. A. M. Sladkow, G. S. Petrow, *Zh. Obshch. Khim.* **24** (1956) 151.
111. Wacker, DE-AS 2 142 419, 1971 (H. Eck, H. Spes).
112. I. G. Farbenind., DE 588 352, 1932 (W. Reppe).
113. H. M. Teeter, *J. Am. Oil. Chem. Soc.* **40** (1963) 143 –155.
114. H. Hopff, CH 324 667, 1954.
115. Shell, BE 634 642, 1962 (W. F. Engel, G. E. Rumscheidt).
116. Shell, DE 1 237 557, 1962 (W. F. Engel, G. E. Rumscheidt).
117. Shell, DE 1 238 322, 1963.
118. Wacker, DE 1 542 574, 1965 (H. Kainzmaier, G. Hübner).
119. Shell, DE-OS 1 768 289, 1967 (A. C. H. Borsboom, P. A. Gautier, D. Medema).
120. G. Hübner, *Fette Seifen Anstrichm.* **68** (1966) 290 –292.
121. *Ullmann,* 3rd ed., **18,** 84 – 86.

Vinyl Ethers

Ernst Hofmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 1 – 8)

Hans-Joachim Klimisch, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 9)

1. Introduction 1
2. Properties 1
2.1. Physical Properties 1
2.2. Chemical Properties 3
3. Production 3
3.1. Reaction of Acetylene with Alcohols 4
3.2. Reaction Conditions 4
3.2.1. Introduction 4
3.2.2. Dependence of Reaction Conditions on the Nature of the Alcohol 4
3.3. Process Description 5
3.3.1. Production of Vinyl Ethers under Pressure 5
3.3.2. Pressure-free Production of Vinyl Ethers 6
4. Environmental Protection, Waste Disposal 6
5. Quality Specifications 6
6. Chemical Analysis 6
7. Storage and Transport 6
8. Use 6
9. Toxicology 7
10. References 8

1. Introduction

Vinyl ethers were first described by Wislicenus in 1878. He obtained ethyl vinyl ether by treating chloroacetal with sodium.

Vinyl ethers (CH_2=CH−O−R; R = alkyl, cycloalkyl, or aryl) may be regarded as derivatives of vinyl alcohol, the unstable enol form of acetaldehyde. Vinyl ethers with substituted vinyl groups are of laboratory interest only, and are not dealt with here. In the 1930s, alkyl vinyl ethers achieved industrial importance as the starting materials for polymers (→ Poly(Vinyl Ethers)), after they had been made readily accessible through the work of Reppe at I. G. Farbenindustrie, Ludwigshafen [1–4].

2. Properties

2.1. Physical Properties

The lower aliphatic vinyl ethers ($C_1 - C_4$) are colorless, combustible, mobile, volatile liquids with an ethereal odor. Methyl vinyl ether is gaseous under normal conditions, but can be liquefied under pressure (pressurized gas). Octadecyl vinyl ether, one of the higher vinyl ethers, is a low-melting solid of low volatility. Vinyl ethers are only slightly soluble in water, but readily miscible with most organic solvents (with the exception of glycol or glycerol). Some physical properties of industrially important vinyl ethers are given in Table 1.

2.2. Chemical Properties

Unlike the chemically very stable saturated ethers (→ Ethers, Aliphatic, Chap. 2.2.), vinyl ethers are extremely reactive substances because of the presence of the enol ether group. They can be readily polymerized in the presence of Friedel – Crafts catalysts, such as BF_3, $AlCl_3$, or $SnCl_4$ (→ Poly(Vinyl Ethers)). They are hydrolyzed by dilute aqueous acid at room temperature, to acetaldehyde and an alcohol [5, p. 391]. However, they are stable to alkalis, even hot aqueous caustic solutions.

Ethyl ethers are obtained by catalytic hydrogenation of vinyl ethers. In the presence of acid catalysts they react with alcohols to give acetals. Chlorine and bromine add to vinyl ethers giving 1,2-dihaloethyl ethers, but iodine causes polymerization, even in very small quantities. The reaction of halogens with vinyl ethers in

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_435

Table 1. Physical properties of representative vinyl ethers

Vinyl ether	CAS registry number	Formula	M_r	*mp*, °C	*bp*, °C	d_4^{20}	n_D^{20}	Heat of vaporization, kJ/kg	Flash point, °C	Ignition temperature, °C	Explosion limits in air, vol % vinyl ether
Methyl	[*107-25-5*]	$CH_3-O-CH=CH_2$	58.10	−122	6	0.773 [a]	1.3947 [b]	417	−60	210	1.9 – 32.0
Ethyl	[*109-92-2*]	$C_2H_5-O-CH=CH_2$	72.12	−115	36	0.754	1.3767	379	−45	180	1.35 – 12.0
Propyl	[*764-47-6*]	$C_3H_7-O-CH=CH_2$	86.13		65	0.768	1.3908		−25	185	1.8 – 10.2
Isobutyl	[*109-53-5*]	i-$C_4H_9-O-CH=CH_2$	100.18	−112	83	0.769	1.3965	332	−15	195	0.73 – 7.9
Octadecyl	[*930-02-9*]	$C_{18}H_{37}-O-CH=CH_2$	296.54	27	175 – 190 [c]	0.81 [d]		239	171	265	
Cyclohexyl	[*2182-55-0*]	$C_6H_{11}-O-CH=CH_2$	126.19	−109	150 – 152	0.891	1.4550	352	35	190	1.0 – 7.3
Butanediol-1,4-monovinyl [e]	[*17832-28-9*]	$HO-C_4H_8-O-CH=CH_2$	116.00	−33	189	0.944	1.4440		88	225	1.4 – 9.9
Butanediol-1,4-divinyl	[*3891-33-6*]	$CH_2=CH-O-C_4H_8-O-CH=CH_2$	142.20	−8	170	0.898	1.4420		63	205	0.9 – 13.9
Diethylene glycol divinyl	[*764-99-8*]	$CH_2=CH-O-C_2H_4-O-C_2H_4-O-CH=CH_2$	158.19	−21	191	0.968	1.4468		83	190	1.1 – 19.0

[a] At 0 °C.
[b] At −25 °C.
[c] At 0.6 kPa.
[d] At 40 °C.
[e] = Vinyl-4-hydroxybutyl ether.

the presence of alcohols gives 2-haloacetals [5, pp. 391 –392]. Siggia's method of determination of vinyl ethers is based on this reaction [6]. Vinyl ethers undergo addition reactions with hydrogen halides giving 1-haloethyl ethers [4, pp. 92 – 93]. The reaction of vinyl ethers with *ortho* esters gives tetraacetals of dialdehydes. With dienes and 1,2-unsaturated aldehydes, vinyl ethers undergo a Diels – Alder reaction [5, pp. 392 – 395], [7, p. 227]. Vinyl ethers react with numerous other reagents, e.g., acetals, carboxylic acids, hydrogen cyanide, phosgene, carbon tetrachloride, ammonia, and amines [5, pp. 393 – 396].

3. Production

Methods for preparing vinyl ethers, e.g., reaction of alkali alkoxides with vinyl halides or dihalohydrocarbons [4, pp. 85 – 86], and transvinylation (transfer of a vinyl group from a vinyl ether or vinyl ester to another alcohol) [5, p. 371], [7, p. 207], [8, p. 508] have not achieved any importance for the production of vinyl ethers beyond laboratory scale.

The production of monovinyl ethers of diols can take place by transfer of a vinyl group from a diol divinyl ether to a free diol:

$$\begin{array}{l} \qquad\qquad\qquad CH_2{=}CH{-}(CH_2)_4{-}O{-}CH{=}CH_2 + \\ HO{-}(CH_2)_4{-}OH \rightleftharpoons 2\,CH_2{=}CH{-}O{-}(CH_2)_4{-}OH \end{array}$$

Butanediol-1,4-monovinyl ether

The equilibrium can be shifted in favor of monovinyl ether formation by excess diol. Compounds of the platinum group metals, in particular those of palladium, are used as catalysts [9].

Vinyloxycarboxylic acid esters (not accessible by Reppe vinylation) can be produced by vinyl group transfer from vinyl ethers to hydroxycarboxylic acid esters, using special palladium compounds as catalysts:

$$HO{-}CH_2{-}\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{C}}{-}COOCH_3 + CH_2{=}CH{-}O{-}i{-}C_4H_9$$

Methyl hydroxypivalate Isobutyl vinyl ether

$$\rightleftharpoons CH_2{=}CH{-}O{-}CH_2{-}\underset{\underset{CH_3}{|}}{\overset{\overset{CH_3}{|}}{C}}{-}COOCH_3 + i{-}C_4H_9{-}OH$$

Methyl vinyloxypivalate Isobutanol

Here also, the equilibrium can be shifted in favor of vinyloxycarboxylic acid ester formation by excess vinyl ether [10].

An industrial process for vinyl ether production involves the cleavage of acetals at 270 – 400 °C in the presence of palladium asbestos, thorium oxide, or manganese sulfate on aluminum oxide as catalyst:

$$CH_3{-}CH(OR)_2 \rightarrow CH_2{=}CH{-}O{-}R + R{-}OH$$

Ethyl, butyl, and isobutyl vinyl ethers have been produced industrially by this method (Union Carbide) [5, pp. 369 – 370], [7, p. 207], [8, p. 508]. Improved processes for vinyl ether production by acetal cleavage use special zeolites as catalysts (mordenite [11], borosilicate – iron silicate zeolites [12]), but these developments have not been exploited industrially.

Vinyl ethers are also obtained by elimination of hydrogen chloride from chloroethyl ethers using alkali [13]. For example, 2-chloroethyl vinyl ether used to be produced industrially from 2,2′-dichlorodiethyl ether by heating with solid NaOH (Union Carbide) [5, p. 371]:

$$\begin{array}{l} Cl{-}CH_2{-}CH_2{-}O{-}CH_2{-}CH_2{-}Cl + NaOH \rightarrow \\ \quad CH_2{=}CH{-}O{-}CH_2{-}CH_2{-}Cl + NaCl + H_2O \end{array}$$

In a catalytic oxidation process developed by ICI for the production of methyl vinyl ether, ethylene is reacted with methanol in the presence of palladium chloride complexes or palladium metal [14–16]. It is not known whether this process has been developed on an industrial scale:

$$CH_2{=}CH_2 + CH_3{-}OH \rightarrow CH_2{=}CH{-}O{-}CH_3 + H_2O$$

3.1. Reaction of Acetylene with Alcohols

The most important industrial method for the production of vinyl ethers is the reaction of acetylene with alcohols in the presence of strong alkalis in the liquid phase (Reppe vinylation) [3], [17]:

$$CH{\equiv}CH + R{-}OH \rightarrow CH_2{=}CH{-}O{-}R \quad \Delta H = -134\ \text{kJ/mol}$$

The strongly exothermic reaction is carried out in the presence of alkali alkoxide catalysts. The reaction proceeds particularly smoothly in the presence of the corresponding potassium compound [1], [4]. The vinylation temperature is generally 150 – 180 °C. With methanol, a vigorous reaction occurs at 120 °C.

3.2. Reaction Conditions

3.2.1. Introduction

The choice of operating pressure depends on the boiling point of the alcohol to be vinylated. Lower alcohols ($C_1 - C_4$) are vinylated under pressure, as their boiling points lie below the reaction temperature. Thus, methyl vinyl ether is produced at 1.6 MPa, ethyl vinyl ether at 1.5 MPa, and isobutyl vinyl ether at 0.8 MPa. The vinylation also succeeds at normal pressure if so much KOH is dissolved in the alcohol that the boiling point of the solution rises to the vinylation temperature or above. However, this method gives poor yields because of side reactions.

When working under pressure, the acetylene must be diluted with nitrogen for safety reasons [3] (→ Acetylene, Chap. 5.1.). The acetylene concentration must not exceed 55 % for methyl and ethyl vinyl ethers, and 75 % for isobutyl vinyl ether.

Higher alcohols can be vinylated at normal pressure as their boiling points are above the vinylation temperature. In this case, expenditure on equipment is significantly lower than in the pressure process. In the normal pressure process, octadecyl vinyl ether, vinyl ethers of fatty alcohol mixtures, and decalyl vinyl ether, for example, can be produced.

The vinylation of alcohols is also possible in the gas phase. A mixture of alcohol vapor and acetylene is passed over solid-bed catalysts, e.g., soda lime [4, p. 87], KOH on activated charcoal, or magnesium or calcium oxides [18], but such processes have not achieved any practical importance.

3.2.2. Dependence of Reaction Conditions on the Nature of the Alcohol

In the case of primary and secondary aliphatic alcohols, vinylation proceeds very smoothly, even for the highest homologues. Tertiary alcohols react less readily. Polyhydric alcohols, e.g., 1,4-butanediol, can be partially or fully vinylated (to monovinyl or divinyl ethers) depending on the reaction conditions chosen. Where the distance between the hydroxyl groups is small (e.g., with ethylene glycol) the cyclic acetal is formed from the monovinyl ether before the latter can react with the acetylene to give the divinyl ether. Alicyclic alcohols (e.g., cyclohexanol, 2-decalol (2-decahydronaphthol), terpene alcohols, tertiary aminoalcohols, and heterocyclic alcohols, such as hydroxyfuran, can also be vinylated. The vinylation of aminoalcohols with primary or secondary amino groups to give the corresponding vinyl ethers can best be carried out with so-called superbases as catalysts (DMSO with alkali hydroxide/alkoxide, preferably potassium compounds) [19]. Examples are 3-vinyloxy-1-propylamine, *N*-ethyl-2-vinyloxy-1-ethylamine, and 4-vinyloxypiperidine.

Phenols and naphthols react with acetylene less readily than aliphatic alcohols. The corresponding aryl vinyl ethers can be obtained only in the presence of a large quantity of potassium phenoxide and a good solvent for acetylene, e.g., *N*-methylpyrrolidone.

3.3. Process Description

In industry, vinyl ethers are produced in a vertical reaction tower (tube reactor). The alcohol, endowed with KOH, is charged at the top, while the gaseous mixture of acetylene and nitrogen is fed in from below. The acetylene not consumed in the reaction is led back into the reactor (recycling process) and supplemented by pure acetylene (fresh gas).

3.3.1. Production of Vinyl Ethers under Pressure

The pure acetylene is precompressed in the fresh gas compressor (Fig. 1, h) to the required operating pressure and then pumped into the pressure tower (c) from below by means of the circulating gas pump (g). At the same time, the alcohol, to which 2 – 3 % KOH has been added in the feed vessel (a), is continuously pumped from above toward (c), by pressure pump (b). The vinyl ether formed at 150 – 160 °C distills together with excess unreacted alcohol from the pressure tower (c), and is condensed out of the circulating gas stream in the condenser (e). The remaining circulating gas is led back to the circulating gas pump after removal of the waste gas

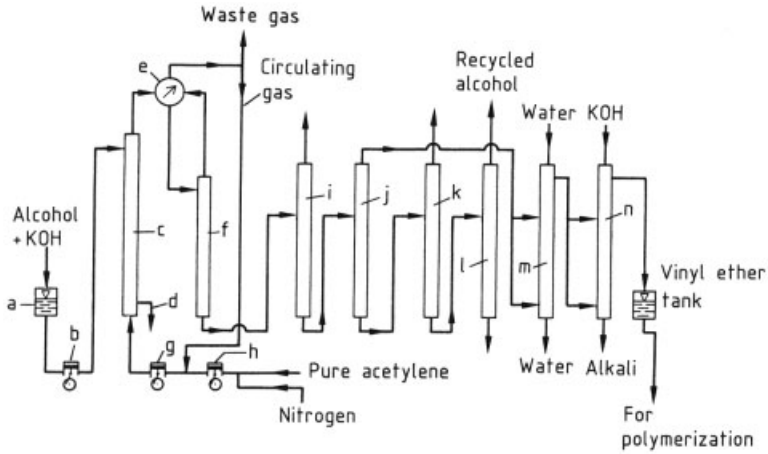

Figure 1. Production of vinyl ethers under pressure
a) Feed vessel; b) Pressure pump; c) Pressure tower; d) Sump exit; e) Condenser; f) Pressurized degasser; g) Circulating gas pump; h) Fresh gas compressor; i) Fores column; j) Purification column; k, l) Alcohol columns; m) Washing column; n) Drying column

by taking out inert components. The heat of reaction is removed by evaporation mainly of the vinyl ether – alcohol mixture and by a cooling coil in the pressure tower (c). A small proportion of the contents of the tower is removed at the lower end of the pressure tower via the drainage exit (d) to remove used catalyst and high-boiling side products from the reaction mixture.

The vinyl ether – alcohol mixture distilling from the pressure tower (c) is freed from dissolved acetylene in the pressure degasser (f) and from low-boiling fractions in the fores column (i). In the purification column (j) the vinyl ether is obtained at the distillation head by fractional distillation, and freed from alcohol residues in the washing column (m) by washing with a countercurrent of water. After drying with solid KOH in the drying tower (n) to remove residual moisture, the vinyl ether is sufficiently pure for further processing (polymerization).

Unreacted alcohol forming the sump in the purification column (j) is purified in two further distillation columns (k, l) and recycled to (a). The whole plant operates completely continuously.

The production of special vinyl ethers under pressure is carried out as a batch operation in stirred autoclaves (vinylation of cyclohexanol, butane-1,4-diol, diethylene glycol, etc.). The crude vinyl ethers produced are purified by vacuum distillation.

3.3.2. Pressure-free Production of Vinyl Ethers

Vinylation at normal pressure is carried out as a batch operation so that complete conversion of the alcohol is achieved and the purification of the reaction product is simplified.

The alcohol treated with 2 – 3 % KOH in the feed vessel (Fig. 2, a) is charged in batches to the top of the reaction tower (b). The pure acetylene is fed into the reaction tower from below by means of the circulating gas blower (d). Unreacted acetylene comes out of the top and passes back to the circulating gas blower (d) via the condenser (c). By removal of waste gas, the acetylene concentration in the gas circulation is held at ca. 90 % (the remainder is nitrogen). The reaction temperature is 170 – 180 °C. The heat of reaction is removed from the reaction tower (b) by means of a cooling coil.

After the reaction has finished (acetylene uptake falls to zero) nitrogen is blown through the crude vinyl ether to remove dissolved acetylene, and the vinyl ether is then passed into the stirred vessel (e) where it is purified by vacuum distillation using column (f) in a batch process.

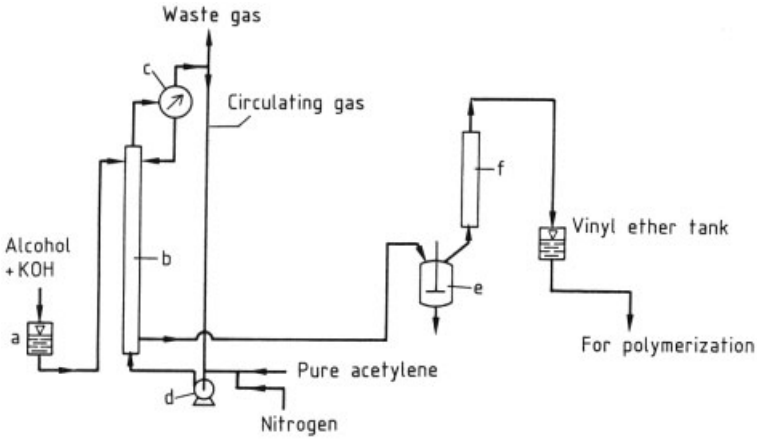

Figure 2. Pressure-free production of vinyl ethers
a) Feed vessel; b) Reaction tower; c) Condenser; d) Circulating gas blower; e) Stirred vessel; f) Column

4. Environmental Protection, Waste Disposal

To protect the environment, the residues and waste materials formed in the production of vinyl ethers must be disposed of according to the regulations. Gaseous waste products (process waste gas from the synthesis and displacement gas from the containers) are incinerated in an excess gas burner. Liquid waste and distillation residues are disposed of in an incineration plant for special wastes. Wastewater which is licensed for treatment is subjected to biological purification in a treatment plant. No environmental problems are encountered if disposal measures are carried out properly.

5. Quality Specifications

As precursors for polymer production, vinyl ethers are subject to high quality requirements. The specifications require a purity (vinyl ether content) >99.8 %, residual alcohol content <0.1 %, other side products <0.1 %, and water <0.01 %.

6. Chemical Analysis

Vinyl ethers are preferably analyzed by standard GLC methods. They may also be determined quantitatively by addition of iodine to the double bond of the vinyl group in the presence of methanol (Siggia's method) [6]. Absolutely dry vinyl ether (<0.01 water) does not attack freshly pressed sodium wire (no gas formation, no change in the lustrous metal surface within 1 h).

7. Storage and Transport

Vinyl ethers can be stored for only a limited period. They can be kept at 20 °C for several months if water and acids are excluded. They are stored in stainless or normal steel containers under dry nitrogen, which prevents the formation of combustible air – vapor mixtures. The exclusion of atmospheric oxygen inhibits the formation of peroxide in the stored product. Methyl vinyl ether (pressurized gas) is stored in pressurized containers.

High-boiling organic bases have proved to be useful stabilizers for inhibiting polymerization during transport of vinyl ethers, e.g., *N,N*-diethylaniline, or triethanolamine (soluble in the product). The quantity of stabilizer added is 0.05 – 0.1 %. The use of solid KOH in pellets as a stabilizer is limited to drums of up to 200 L.

Because vinyl ethers are readily flammable (see Table 1), the relevant safety instructions and explosion protection measures must be adhered to during production, storage, and transport. Dispatch regulations and hazard classes for transport are given in Table 2.

Table 2. Dispatch regulations for vinyl ethers: hazard classes

Vinyl ether	Rail, GGVE/ RID	Trucks, GGVS/ ADR	Ship, IMDG Code
Methyl	2,3 ct	2,3 ct	2, UN 1087
Ethyl	3,2 b	3,2 b	3,1 UN 1302
Propyl	3,3 b	3,3 b	3,1 UN 1993
Isobutyl	3,3 b	3,3 b	3,2 UN 1304
Octadecyl	none	none	none
Cyclohexyl	3,31 c	3,31 c	3,3 UN 1993
Butanediol-1,4-monovinyl ∗	3,32 c	3,32 c	none
Butanediol-1,4-divinyl	3,32 c	3,32 c	3,3 UN 1993
Diethylene glycol divinyl	3,32 c	3,32 c	

∗ = Vinyl-4-hydroxybutyl ether.

8. Use

Vinyl ethers are used principally for the production of homo- and copolymers, in the adhesives and varnish industries (→ Poly(Vinyl Ethers)). Vinyl ethers can also be used as chemical intermediates for the production of glutardialdehyde [7, p. 227], tetramethoxypropane [5, p. 392], γ-picoline [20], and 2-aminopyrimidine [21].

9. Toxicology

Investigations of toxicokinetics and metabolism have shown that, on inhalation, rapid resorption takes place, and that vinyl ethers are relatively stable and metabolized only slightly. They are mainly exhaled unchanged. Epoxidation of the vinyl group has not been described [22–25].

The acute toxicity of vinyl ethers is low. In rats the acute oral LD_{50} decreases with increasing chain length:

Methyl vinyl ether	4.900 mg/kg [22], [26]
Ethyl vinyl ether	8.160 mg/kg [27]
Butyl vinyl ether	10.300 mg/kg [28]
Isobutyl vinyl ether	13.100 mg/kg [29]

As LC_{50} indicates, the acute toxicity by inhalation in rats is also very low:

Methyl vinyl ether	150 mg/L	(4 h, rat) [26]
Ethyl vinyl ether	21.2 mg/L	(4 h, rat) [30]
Isobutyl vinyl ether	21.1 mg/L	(4 h, rat) [31]

The symptoms are characterized by the narcotic action of the vinyl ethers at high concentration. Compared with saturated alkyl ethers, the vinyl group reinforces the narcotic properties. Correspondingly, divinyl ether is a stronger narcotic than, for example, ethyl vinyl ether or diethyl ether.

On repeated inhalation (28-day test on rats) only slight systemic toxicity was observed at high concentration (25 000 ppm). With methyl vinyl ether there was an increase in the relative liver weight with no histopathological correlation and a local irritant effect on the mucosa of the nasal cavity in the form of atrophy of the olfactory epithelium. The no adverse effect concentration (NOAEC) was 1500 ppm (3.6 mg/L) for male rats and 3500 ppm (8.3 mg/L) for female rats. No other histopathological change was observed [22]. In experiments with dogs no changes in the liver, no electrocardiographic changes (dogs, monkeys), and no effect on blood pressure were found [32].

On contact with skin and with the mucous membranes of the eye, highly volatile vinyl ethers are not irritant, and those with low volatility slightly irritant [22], [29], [33], [34].

No data from skin sensitization experiments on animals are available. However, short-chain vinyl ethers are not expected to be hazardous because of their high volatility. In humans no cases of sensitization have been published, either in production of the compounds or for their earlier use as an inhaled narcotic.

In vitro investigations on genotoxicity are contradictory. On microorganisms (*Salmonella typhimurium*: Ames test) and with mammalian cells (HPRT test: [35]) vinyl ethers were not mutagenic, with or without metabolic activation. Chromosome aberrations could not be detected on somatic cells with isobutyl vinyl ether [36]. However, with ethyl vinyl ether and 2,2,2-trifluoroethyl vinyl ether possible indications of an increase in the sister chromatid exchange rate

(SCE test; [37]) have been described. In animal experiments no genotoxic activity was observed (micronucleus test; [22]).

No investigations of carcinogenicity and reproductive toxicity in animals have been reported.

From the use of ethyl vinyl ether as an inhaled narcotic in the 1950s, a number of effects on humans have been reported. In one out of 220 narcoses spasm of the larynx and flow of saliva occurred. In 12 of 220 cases there was generalized cramp and respiratory and cardiac arrest. Postnarcotic effects are: nausea/vomiting (26.7 %), reduction in circulation (4.1 %), excitation (1.6 %), and hypoxia (0.8 %). Liver function tests gave no indication of damage [38]. In another investigation indications of slight leukocytosis and slight myocardial depression were observed, but liver function tests showed no change [38].

10. References

1. I. G. Farbenind., DE 584 840, 1930; DE 639 843, 1933; DE 640 150, 1934 (W. Reppe).
2. W. Reppe: *Neue Entwicklungen auf dem Gebiet der Chemie des Acetylens und Kohlenoxids,* Springer-Verlag, Berlin 1949.
3. W. Reppe: *Chemie und Technik der Acetylen-Druckreaktionen,* 2nd ed., Verlag Chemie, Weinheim 1952.
4. W. Reppe et al., *Justus Liebigs Ann. Chem.* **601** (1956) 81 – 110.
5. N. D. Field et al.: "Vinyl Ethers," in *High Polymers,* vol. 24, Wiley, New York 1971, pp. 365 – 411.
6. S. Siggia, R. L. Edsberg, *Anal. Chem.* **20** (1948) 762.
7. S. A. Miller: "Vinyl Ethers," in *Acetylene, Its Properties, Manufacture and Uses,* vol. 2, Ernest Benn Ltd., London 1966, pp. 198 – 231.
8. D. H. Lorenz: *Encyclopedia of Polymer Science and Technology,* **vol. 14,** Interscience, New York 1971, pp. 504 – 510.
9. BASF, DE 4 134 807 A1, 1991 (M. Karcher, J. Henkelmann).
10. BASF, EP 0 538 685 A2, 1992 (M. Karcher, J. Henkelmann).
11. Degussa, DE 3 535 128 A1, 1985 (D. Arntz et al.).
12. BASF, EP 0 299 286 A1, 1988 (W. Holderich et al.).
13. C. E. Rehberg, *J. Am. Chem. Soc.* **71** (1949) 3247.
14. ICI, NL 6 411 879, 1965.
15. ICI, NL-Appl. 6 611 559, 1967.
16. *Ullmann,* 3rd ed., **18,** 96.
17. *Ullmann,* 4th ed., **7,** 45.
18. R. L. Zimmerman et al., *Ind. Eng. Chem. Prod. Res. Dev.* **2** (1963) no. 4, 293 – 296.
19. BASF, EP 0 514 710 A2, 1992 (M. Karcher, J. Henkelmann).
20. BASF, DE 944 250, 1952 (W. Reppe et al.).
21. BASF, DE 1 117 587, 1958 (H. Maisack et al.).
22. Vinylmethylether (BG no. 63), Toxikologische Bewertung, Berufsgenossenschaft der chemischen Industrie, ed. 02/92.
23. M. E. Andersen, M. L. Gargas, R. A. Jones, L. J. Jenkins, Jr.: "Determination of the Kinetic Constants for Metabolism of Inhaled Toxicants in vivo Using Gas Uptake Measurements," *Toxicol. Appl. Pharmacol.* **54** (1980) 100 – 116.
24. M. E. Andersen: "A Physiologically Based Toxicokinetic Description of the Metabolism of Inhaled Gases and Vapors: Analysis at Steady State," *Toxicol. Appl. Pharmacol.* **60** (1981) 509 – 526.
25. R. T. Williams: *Detoxication Mechanism,* 2nd ed., John Wiley & Sons, New York 1959.
26. W. B. Deichmann: *Toxicology of Drugs and Chemicals,* Academic Press, New York 1969.
27. G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd ed., **vol. 2 A,** John Wiley & Sons, New York 1981, p. 2503.
28. H. F. Smyth, Ch. P. Carpenter, C. S. Weil, H. C. Pozzani: *AMA Arch. Ind. Hyg. Occup. Med.* **10** (1954) 61.
29. H. F. Smyth, Ch. P. Carpenter, C. S. Weil, H. C. Pozzani: *Am. Ind. Hyg. J.* **23** (1962) 95 – 98.
30. BASF AG, Toxicology department: unpublished results (88/193).
31. BASF AG, Toxicology department: unpublished results (13I0192/887035).
32. J. R. Krantz et al.: *J. Pharmacol. Exp. Ther.* **90** (1947) 88.
33. BASF AG, Toxicology department: unpublished results (79/406).
34. BASF AG, Toxicology department: unpublished results (88/193).
35. BASF AG, Toxicology department: unpublished results (50M0234/924205).
36. BASF AG, Toxicology department: unpublished results (32M0234/924195).

37. A. E. White, T. Shim, E. Edmond, W. Sheldon, S. Wendell: *Anesthesiology* **50** (1979) 426 – 430.

38. Vinylethylether (BG no. 263), Toxikologische Bewertung, Berufsgenossenschaft der chemischen Industrie, unpublished data, 09/95.

Vinylidene Chloride → Chlorinated Hydrocarbons
Vinylidene Polymers → Poly(Vinylidene Chloride)
Vinyl Polymers → Fluoropolymers, Organic
Vinyl Polymers → Poly(Vinyl Chloride)
Vinyl Polymers → Polyvinyl Compounds, Others
Vinyl Polymers → Poly(Vinyl Esters)
Vinyl Polymers → Poly(Vinyl Ethers)
Vinylpyridine → Pyridine and Pyridine Derivatives
N-Vinylpyrrolidone → 2-Pyrrolidone
Violanthrone → Anthraquinone Dyes and Intermediates
Viral Infections → Chemotherapeutics
Viral Infections → Veterinary Drugs
Viscose Fiber → Cellulose

Vitamins

Manfred Eggersdorfer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1)

Geo Adam, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 1)

Michael John, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2)

Wolfgang Hähnlein, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2)

Ludvik Labler, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 3)

Kai-U. Baldenius, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Linda von dem Bussche-Hünnefeld, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Eckhard Hilgemann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Peter Hoppe, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Rainer Stürmer, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 4)

Fritz Weber, Münchenstein, Switzerland (Chap. 5)

August Rüttimann, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 5)

Gérard Moine, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 6)

Hans-Peter Hohmann, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 6)

Roland Kurth, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 7)

Joachim Paust, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 7)

Wolfgang Hähnlein, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 7)

Horst Pauling, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 8)

Bernd–Jürgen Weimann, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 8)

Bruno Kaesler, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 9)

Bernd Oster, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 10)

Ulrich Fechtel, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 10)

Klaus Kaiser, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 11)

Bernd de Potzolli, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 11)

Michael Casutt, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 12)

Thomas Koppe, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 12)

Michael Schwarz, E. Merck, Darmstadt, Federal Republic of Germany (Chap. 12)

Bernd-Jürgen Weimann, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 13)

Urs Hengartner, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 13)

Antoine de Saizieu, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 13)

Christof Wehrli, F. Hoffmann-La Roche Ltd, Basel, Switzerland (Chap. 13)

René Blum, Lonza, Basel, Switzerland (Chap. 14)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_443

1. Introduction 6
1.1. Definition 6
1.2. Substances with Vitamin-Like Character 8
1.3. History 8
1.4. Determination of Requirement . . 8
1.5. Application and Tolerance 8
1.6. Studies with Vitamins 10
1.7. Use of Vitamins in Food- and Feedstuffs 10
1.8. Antivitamins 12
1.9. Analysis of Vitamins 12
1.10. Production 13
2. Vitamin A (Retinoids) 13
2.1. Introduction 13
2.2. Historical Aspects 14
2.3. Physical Properties 15
2.4. Chemical Properties 17
2.5. Occurrence 17
2.6. Biosynthesis 18
2.7. Production 20
2.7.1. Isolation and Purification from Natural Raw Materials 20
2.7.2. Industrial Synthesis 20
2.8. Metabolism and Physiological Functions 28
2.8.1. Preformed Vitamin A 28
2.8.2. β-Carotene 29
2.9. Deficiency Symptoms and Requirements 29
2.10. Analysis and Standardization . . . 30
2.11. Trade Names and Economic Aspects 31
2.12. Tolerance 32
3. Vitamin D 32
3.1. Introduction 32
3.2. History 32
3.3. Structure and Nomenclature . . . 33
3.4. Chemical and Physical Properties 33
3.5. Biosynthesis and Occurrence . . . 35
3.5.1. Biosynthesis 35
3.5.2. Occurrence 35
3.6. Vitamin D Requirements 35
3.7. Pharmacological Effects and Uses 36
3.7.1. Calcium Homeostasis and Disorders of Calcium Metabolism 36
3.7.2. Further Activities 37
3.8. Synthesis 37
3.8.1. Synthesis of Vitamin D_3 and D_2 . . 37
3.8.2. Synthesis of Calcitriol 39
3.9. Assays for Vitamin D and Metabolites 39
3.10. Trade Names and Economic Aspects 41
4. Vitamin E (Tocopherols, Tocotrienols) 41
4.1. Introduction 41
4.2. History 42
4.3. Physical Properties 43
4.4. Chemical Properties 43
4.5. Metabolism and Importance in the Organism 43
4.6. Deficiency, Requirement, and Application 44
4.7. Analysis and Standardization . . . 45
4.8. Occurrence 46
4.9. Economic Aspects 46
4.10. Biosynthesis 47
4.11. Production 47
4.11.1. Semisynthetic Vitamin E: Isolation from Plant Oils and Methylation . . 47
4.11.2. Totally Synthetic Vitamin E: Industrial Synthesis 47
4.11.3. Trimethylhydroquinone 48
4.11.4. Phytol and Isophytol 49
4.11.5. Stereoselective Syntheses of α-Tocopherol 50
5. Vitamin K 52
5.1. Introduction; History 52
5.2. Physical Properties 53
5.3. Chemical Properties 54
5.4. Occurrence 55
5.5. Biosynthesis 56
5.6. Chemical Synthesis 57
5.7. Analysis 60
5.8. Metabolism 61
5.9. Importance for the Organism . . . 63
5.9.1. Function 63
5.9.2. Antagonists and Anticoagulants . . 65
5.9.3. Relative Activity 66
5.10. Deficiency Symptoms 68
5.10.1. Causes 68
5.10.2. Evaluation of Vitamin K Status . . . 68
5.11. Requirement 69
5.12. Application 70
5.13. Tolerance 71
5.14. Trade Names and Economic Aspects 71
6. Vitamin B_1 (Thiamin) 71
6.1. Structure and Nomenclature . . . 71
6.2. History 72
6.3. Physical Properties 72
6.4. Chemical Properties 73
6.5. Natural Occurrence and Isolation 73
6.6. Biosynthesis in Microorganisms . 75
6.6.1. Biosynthesis of the Pyrimidine Component in Prokaryotes 76

6.6.2. Biosynthesis of the Pyrimidine Component in Eukaryotes 76
6.6.3. Biosynthesis of the Thiazole Component 76
6.6.4. Biosynthesis of Thiamin Pyrophosphate 77
6.6.5. Thiamin Biosynthetic Genes from *E. coli* 79
6.6.6. Mapping and Cloning of Thiamin Biosynthetic Genes from Other Species 79
6.6.7. Regulation of Thiamin Biosynthesis in Prokaryotes 80
6.6.8. Regulation of Thiamin Biosynthesis in Yeasts 80
6.6.9. Thiamin-Overexpressing Microorganisms 80
6.7. Chemical Synthesis 81
6.7.1. Condensation of the Pyrimidine and Thiazole Rings 81
6.7.2. Construction of the Thiazole Ring on a Preformed Pyrimidine Portion . . 81
6.8. Commercial Forms 83
6.9. Derivatives, Analogues, and Antimetabolites 83
6.10. Biochemical and Physiological Functions 85
6.10.1. Metabolic Functions in the Organism 85
6.10.2. Thiamin Requirements and Deficiency 86
6.11. Analytical Methods 86
7. Riboflavin 86
7.1. Introduction 86
7.2. History 87
7.3. Physical and Chemical Properties 87
7.4. Occurrence 88
7.5. Biosynthesis 89
7.6. Production 89
7.6.1. Syntheses 89
7.6.2. Industrial Chemical Production . . . 91
7.6.3. Industrial Production by Fermentation 92
7.6.4. Syntheses of Flavin Mononucleotide (FMN) 93
7.6.5. Syntheses of Flavin Adenine Dinucleotide (FAD) 93
7.7. Importance for the Organism . . . 93
7.8. Requirements, Deficiency Symptoms, and Therapeutic Application 94
7.9. Analysis 95
7.10. Economic Aspects 96
7.11. Tolerance 96
8. Vitamin B_6 96
8.1. Introduction 96
8.2. History 96
8.3. Physical Properties 97
8.4. Chemical Reactions 98
8.5. Occurrence 99
8.6. Biosynthesis 99
8.7. Production of Vitamin B_6 Compounds 99
8.7.1. Pyridoxine 99
8.7.1.1. Oxidative Degradation of Bicyclic Heterocyclic Compounds 99
8.7.1.2. Condensation Reactions with Aliphatic Precursors 100
8.7.1.3. From Furan Compounds 101
8.7.1.4. Diels – Alder Syntheses with Oxazoles 101
8.7.1.5. Cobalt-Catalyzed [2+2+2]-Cycloaddition Reactions of Acetylenes and Acetonitrile 103
8.7.2. Pyridoxine 5′-Phosphate 103
8.7.3. Pyridoxal 103
8.7.4. Pyridoxal 5′-Phosphate 104
8.7.5. Pyridoxamine 104
8.7.6. Pyridoxamine 5′-Phosphate 104
8.8. Metabolism and Importance for the Organism 104
8.9. Deficiency Symptoms and Applications 105
8.10. Analysis 106
8.11. Economic Aspects 106
8.12. Requirements and Tolerance . . . 106
9. Vitamin B_{12} (Cobalamins) 107
9.1. Introduction 107
9.2. Properties of Vitamin B_{12} 109
9.3. Analysis 110
9.4. Biosynthesis 110
9.5. Production 111
9.5.1. Fermentation 111
9.5.2. Work-Up 113
9.5.3. Patents and Scientific Survey 113
9.6. Specifications and Legal Aspects . 114
9.7. Economic Aspects 114
10. Vitamin C (l-Ascorbic Acid) . . . 114
10.1. Introduction 114
10.2. History 115
10.3. Physical and Chemical Properties 115
10.4. Analysis 116
10.5. Occurrence and Sources of Vitamin C 118
10.6. Biosynthesis 118
10.7. Manufacture of Vitamin C 118
10.7.1. Reichstein Synthesis 119

10.7.2. Industrial Manufacture by the Reichstein Route . . . 120
10.7.3. Other Methods for Production of Ascorbic Acid . . . 122
10.8. Absorption and Metabolism . . . 125
10.9. Medical Aspects of Vitamin C . . . 125
10.10. Industrial Uses . . . 126
10.11. Economic Aspects . . . 127
11. Pantothenic Acid . . . 127
11.1. Introduction . . . 127
11.2. History . . . 127
11.3. Physical and Chemical Properties 127
11.4. Occurrence . . . 127
11.5. Biosynthesis . . . 128
11.6. Production . . . 129
11.7. Metabolism and Importance for the Organisms; Coenzyme A and its Precursors . . . 130
11.8. Deficiency Symptoms, Requirement, and Application . . . 132
11.9. Analysis . . . 134
11.10. Economic Aspects . . . 134
12. Biotin . . . 134
12.1. Introduction . . . 134
12.2. History . . . 135
12.3. Physical and Chemical Properties 135
12.4. Occurrence . . . 135
12.5. Biosynthesis . . . 136
12.6. Function as Prosthetic Group . . . 136
12.7. Isolation and Production . . . 138
12.8. Metabolism and Importance for the Organism . . . 138
12.9. Deficiency Symptoms, Requirement, and Application . . . 140
12.9.1. Symptoms and Therapy in Humans 140
12.9.2. Symptoms and Therapy in Animals 140
12.10. Biotin Analogs . . . 141
12.11. Analysis and Standardization . . . 141
12.12. Uses and Economic Aspects . . . 142
12.13. Tolerance and Environmental Protection . . . 143
13. Folic Acid . . . 143
13.1. Introduction . . . 143
13.2. Historical Notes . . . 144
13.3. Properties . . . 145
13.4. Content in Food and Bioavailability . . . 146
13.5. Biosynthesis . . . 146
13.6. Chemical Synthesis . . . 147
13.7. Metabolism and Biochemical Functions . . . 148
13.8. Nutritional Requirements and Medical Use . . . 149
13.9. Analysis . . . 150
13.10. Economic Aspects . . . 150
14. Niacin (Nicotinic Acid, Nicotinamide) . . . 150
14.1. Introduction . . . 150
14.2. Physical and Chemical Properties 151
14.3. Biochemical Functions . . . 151
14.4. Production . . . 152
14.5. Quality Specifications . . . 153
14.6. Analysis . . . 153
14.7. Uses . . . 153
14.8. Economic Aspects . . . 155
14.9. Toxicology . . . 156
15. References . . . 156

1. Introduction

1.1. Definition [1]

Vitamins are essential, organic compounds which are either not synthesized in the human and animal organism or formed only in insufficient amounts. Therefore, they must be regularly consumed with the diet either as such or as a precursor (provitamin) that can be converted to the vitamin in the body. A typical representative of the provitamins is β-carotene, which is split into two molecules of vitamin A in the organism. The metabolic functions of vitamins are mainly catalytic or regulatory. In contrast to energy-providing nutrients, vitamins are required only in exceptionally small amounts. An inadequate supply of a vitamin leads to typical pathological deficiency symptoms which can be remedied by intake of the lacking vitamin.

Vitamins are classified not chemically but by their activity. The historical distinction between fat- and water-soluble vitamins has been retained to this day [2] because the solution properties are important not only for the occurrence, but also for the behavior of vitamins in the organism (resorption, transport, excretory pathways, and storage).

However, it is not the case that fat-soluble vitamins are stored much more effectively than water-soluble vitamins. The assumption that a deficiency of fat-soluble vitamins occurs after a much longer period than in the case of water-soluble vitamins (higher reserve capacity of liposoluble substances) is incorrect. In fact, the reserve capacity depends on the special prop-

erties of the particular vitamin, irrespective of differences in solubility [3].

In the case of vitamin B_{12}, for instance, the reserve capacity amounts to three to five years, but for thiamine it is only two weeks. Although vitamin K is fat-soluble, deficiency of this vitamin occurs shortly after the interruption of food intake and suppression of enteral synthesis. Consequently, the solubility properties of the individual vitamins only provide vague information about the occurrence and stimulation of absorption by lipids in the food and bile acids.

Table 1. Active forms and functions of vitamins

Vitamin	Active form	Function [4]
Vitamin A	(11-*Z*)-retinal	involvement in vision
Vitamin D		regulation of calcium and phosphate metabolism
Vitamin E		intracellular antioxidant
Vitamin K		biosynthesis of prothrombin
Vitamin B_1	thiamin pyrophosphate (TPP), thiamin triphosphate (TTP)	decarboxylation of α-oxo acids, transfer of aldehyde groups, phosphate donor
Vitamin B_2	flavinmononucleotide (FMN), flavinadenine dinucleotide (FAD), coenzymes of flavoproteins	hydrogen or electron transport in energy-releasing metabolism
Niacin	nicotinamide adenine dinucleotide (NAD), nicotinamide adenine dinucleotide phosphate (NADP)	hydrogen or electron transport
Pantothenic acid	coenzyme A (CoA)	transfer of acetyl groups
Vitamin B_6	pyridoxal-5-phosphate	transfer of amino groups
Biotin	biocytin	transfer of carboxyl groups
Folic acid	tetrahydrofolic acid	transfer of formyl groups
Vitamin B_{12}	coenzyme B_{12}	1,2-hydrogen shift
Vitamin C		involvement in hydroxylation, ultracellular antioxidant

On the basis of the above definition, thirteen compounds or groups of compounds have been classified as vitamins for humans (Table 1):

Fat-soluble vitamins
Vitamin A (retinols)
Vitamin D (calciferols)
Vitamin E (tocopherols, tocotrienols)
Vitamin K (phylloquinone)

Water-soluble vitamins
Vitamin B_1 (thiamin)
Vitamin B_2 (riboflavin)
Vitamin B_6 (pyridoxal group)
Vitamin B_{12} (cobalamins)
Vitamin C (L-ascorbic acid)
Pantothenic acid
Biotin
Folic acid
Niacin

Other compounds exist that are also important for the organism and cannot always be produced in sufficient quantities. Hence, an alimentary additional requirement can arise under certain conditions. These are known as pseudovitamins.

1.2. Substances with Vitamin-Like Character

In the past, many authors regarded the essential fatty acids ("vitamin F"), lipoic acid, ubiquinones, choline, myoinositol, "vitamin U" (*S*-methylmethionine), and some other substances and substance mixtures as vitamins. Some characteristics of these substances are listed in Table 2.

1.3. History

The term "vitamins" was introduced by Funk [5] in 1911. Searching for the causes of beriberi, a nervous disease often encountered in East Asia, Eijkman showed in 1896 [6] that polished rice caused this disease and that it could be cured by consumption of rice bran. In 1911, Funk isolated a substance in the form of water-soluble crystals from rice bran. This substance was responsible for curing beriberi and was named "beriberi vitamin".

The designation "vitamin" was accepted for all the essential nutritional factors discovered in the following period, although it was soon recognized that not all these substances are amines. According to a proposal made by Drummond [7] in 1920, the beriberi vitamin was designated vitamin B, the fat-soluble antixerophthalmia factor was named vitamin A, and the

Table 2. Compounds with vitamin or coenzyme character

Compound	Formula	Source/compound	Biological importance
Vitamin F (essential fatty acids)	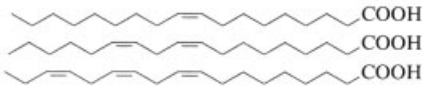	plant oils fish oils	component of membranes
Lipoic acid	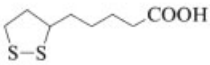	animal and plant cells	growth factor
Vitamin Q (ubiquinone)	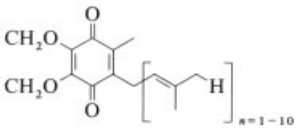	animal and plant cells	factor in the redox system of the respiratory chain
Myoinositol	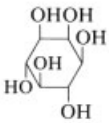	plant cells	nutrient for microorganisms, growth factor for yeasts and fungi
Vitamin U (*S*-methyl-methionine)	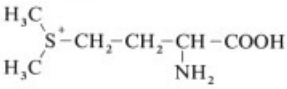	vegetables	"antipeptic ulcer factor"
Choline chloride	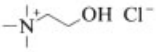	animal and plant cells	growth and metabolic factor, deficiency causes fatty liver, hypotensive
Taurine		oxidative degradation product of cysteine	conjugate formation with bile acids → fat resorption
Carnitine	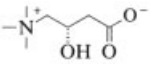	animal and plant cells	growth factor for lower life forms, involvement in fat metabolism
Orotic acid (Vitamin B_{13})	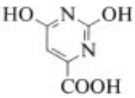	yeasts, liver, milk	intermediate in pyrimidine metabolism, growth factor
Adenosine monophosphate (Vitamin B_8)	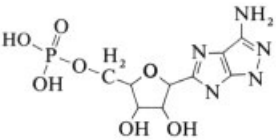	component of coenzymes (CoA, NAD^+, FAD), structural unit of ribonucleic acids	allosteric effector in the phosphofructo kinase and fructose diphosphatase reactions
Adenosine-3′,5′-monophosphate (cyclic AMP)	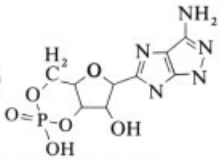	animal and plant cells	transfer of the effect of hormones and hormone-like substances

antiscurvy factor was named vitamin C. All the vitamins discovered later were included in this classification.

A survey of the discovery, isolation, and assignment of the chemical structure of the individual vitamins is presented in Table 3.

As mentioned above, the discovery of the vitamins was due to the fact that a deficiency can severely damage the organism. Therefore, the substances that should be regarded as vitamins are all known today. It is unlikely that new food constituents which fully correspond to the definition of vitamins with regard to their occurrence and action will be discovered.

Nevertheless, the "history of vitamins" has still not come to an end. A constantly increasing

Table 3. Discovery of the vitamins

Vitamin	First isolation	Discovery	Isolation	Structural elucidation
Vitamin A	fish liver oil	1909	1931	1931
Vitamin D	fish liver oil	1918	1932	1936
Vitamin E	wheat germ oil	1922	1936	1938
Vitamin K	alfalfa	1929	1939	1939
Vitamin B_1	rice	1897	1911	1936
Vitamin B_2	eggs	1920	1933	1935
Vitamin B_6	rice	1934	1938	1938
Vitamin B_{12}	liver	1926	1948	1956
Niacin	liver	1936	1935	1937
Folic acid	liver	1941	1941	1946
Panthothenic acid	liver	1931	1938	1040
Biotin	liver	1931	1935	1042
Vitamin C	lemons	1912	1928	1933

number of publications deal with the preventive effects of vitamins on vascular diseases and various types of cancer (see Section 1.6).

1.4. Determination of Requirement

According to the definition of the term vitamin, an inadequate supply of vitamins causes a certain deficiency (hypovitaminosis or avitaminosis).

In the case of avitaminosis, a minimum amount of vitamin is required to remedy these conditions. However, this amount does not correspond to the true requirement. No exact data can be given for this requirement because it varies from individual to individual and strongly depends on the physical condition of the particular person. In stress situations (e.g., illness, pregnancy, and physical exertion), the requirement is considerably higher. Even when in human experiments the selected group of persons is well balanced in this respect, substantial variations remain, which are mainly due to the difficulty of producing the same state of deficiency at the start of the investigation and in evaluating the often complex deficiency symptoms.

For this reason, it is meaningful to stipulate the "desirable supply" of a vitamin (recommended dietary intake = RDI or recommended dietary allowance = RDA) instead of the exact requirement. Experience shows that full health and productivity are observed at these levels of intake. RDA values are regularly issued by the U.S. Food and Nutrition Board, the U.K. Department of Health and Social Security, WHO, and by the German Society for Nutrition (DGE) [8]. A compilation issued in 1991 is listed in Table 4.

1.5. Application and Tolerance

Vitamins are used as single preparations or as multivitamin preparations in the treatment of hypovitaminosis and avitaminosis and in the prophylaxis of these conditions.

Recently, much attention has been paid to the megadosing of vitamins, i.e., the controlled administration of vitamins in doses which exceed by far the physiological intake.

The following applications can be differentiated:

1) Combating deficiency diseases, e.g., megadoses of vitamin K for infants and small children in developing countries
2) Prophylactic or therapeutic applications in the case of diseases which can cause vitamin deficiency
   a) Vitamin C for colds
   b) Fat-soluble vitamins in chronic disturbances of fat absorption (chronic Pauls reaction, chronic cholesterase)
   c) Megadoses of thiamin for alcoholics
3) Alleviation of vitamin metabolic diseases
   a) Megadoses of vitamin D in cases of chronic kidney insufficiency
   b) Megadoses of biotin in cases of multiple carboxylase deficiency

A survey of the use of megadoses of vitamins in prophylaxis and therapy is given in [9].

Hyperdosing can occur in the case of the fat-soluble vitamins A and D because they can be

Table 4. Recommended vitamin intake per day

Group	Vitamin A, mg RE [a]		Vitamin D,	Vitamin E,	Vitamin K, mg		Thiamine, mg		Riboflavine, mg		Niacin, mg NE [c]		Vitamin B_6, mg		Folic acid, µg		Vitamin B_{12},	Vitamin C,
	m	f	µg	mg TE [b]	m	f	m	f	m	f	m	f	m	f	[d]	[e]	µg	mg
Infants																		
0 – 4 months	0.5		10	3	5		0.3		0.3		5		0.3			40	0.5	40
4 – 12 months	0.6		10	4	10		0.4		0.5		6		0.6		80	40	0.8	50
Children																		
1 – 4 years	0.6		5	6	15		0.7		0.8		9		0.9		120	60	1.0	55
4 – 7 years	0.7		5	8	20		1.0		1.1		12		1.2		160	80	1.5	60
7 – 10 years	0.8		5	9	30		1.1		1.2		13		1.4		200	100	1.8	65
10 – 13 years	0.9	0.9	5	10	40	40	1.2	1.2	1.4	1.3	15	14	1.6	1.5	240	120	2.0	70
13 – 15 years	1.1	1.0	5	12	50	50	1.4	1.2	1.5	1.4	17	15	1.8	1.6	300	150	3.0	75
Youths and adults																		
15 – 19 years	1.1	0.9	5	12	70	60	1.6	1.3	1.8	1.7	20	16	2.1	1.8	300	150	3.0	75
19 – 25 years	1.0	0.8	5	12	70	60	1.4	1.2	1.7	1.5	18	15	1.8	1.6	300	150	3.0	75
25 – 51 years	1.0	0.8	5	12	70	60	1.6	1.3	1.8	1.7	18	15	1.8	1.6	300	150	3.0	75
51 – 65 years	1.0	0.8	5	12	80	65	1.3	1.1	1.7	1.5	18	15	1.8	1.6	300	150	3.0	75
65 years and older	1.0	0.8	5	12	80	65	1.3	1.1	1.7	1.5	18	15	1.8	1.6	300	150	3.0	75
Pregnant women		1.1	10	14		65		1.5		1.8		17		2.6	600	300	3.5	100
Nursing mothers		1.8	10	17		65		1.7		1.7		20		2.2	450	225	4.0	125

[a] 1 mg retinol equivalent = 6 mg all-*trans*-β-carotin = 12 mg other provitamin A carotinoide = 1.15 mg all-*trans*-retinyl acetate = 1.83 mg all-*trans*-retinyl palmitate.

[b] 1 mg (*RRR*)-α-tocopherol equivalent = 1.1 mg *RRR*-α-tocopheryl acetate = 2 mg (*RRR*)-β-tocopherol = 4 mg (*RRR*)-γ-tocopherol = 100 mg (*RRR*)-δ-tocopherol = 3.3 mg (*RRR*)-α-tocotrienol = 1.49 mg all-*rac*-α-tocopheryl acetate.

[c] 1 mg niacin equivalent = 60 mg tryptophan.

[d] Calculated on the basis of "total folate" (sum of folate-active compounds in normal food).

[e] Folate equivalent or free folic acid (pteroyl monoglutamate).

stored and can be eliminated only by conversion to excretable substances. To prevent toxic side effects, controlled administration is necessary.

1.6. Studies with Vitamins

The effects of overdoses of vitamins (vitamin C, E, β-carotene) in the prevention of cancer and coronary heart diseases are being studied in a series of investigations.

A summary of the studies on the effects of antioxidative vitamins is given in Table 5.

An evaluation of these studies shows that the intake of certain vitamin/carotene combinations has a prophylactic effect. Further information on these results is given in [10–12].

1.7. Use of Vitamins in Food- and Feedstuffs

For humans, the most important sources of vitamins are plant and animal foods. In the industrial countries, the vitamin requirement can be completely covered by the food consumed. Nevertheless, the danger of a suboptimal supply exists in large segments of the population of these countries. According to the nutritional report of the Deutsche Gesellschaft für Ernährung (German Society for Nutrition), the supply of the vitamins B_1, B_2, and B_6 is regarded as critical in Germany [13].

The reasons for this are the one-side preference for certain foods and a steady transition from fresh, home cooked meals to the consumption of canned foods and mass catering. The industrial production of foods, which requires preservation in most cases to give long shelf lives, results in a considerable loss of vitamins due to the effect of heat, air, and light. To compensate for these losses, fortification of basic foods (flour, bread, and cakes with B complex vitamins; margarine and milk products with vitamins A, E, and D) diet foods (e.g., baby foods containing all the vitamins) and fruit juices (vitamin C, etc.) is gaining importance. For legal aspects, see [14].

In the feed industry, vitamins are used on a large scale as additives in animal feeding to ensure an optimal supply.

Breeding diseases, fertility losses, and performance changes of all types are avoided by the selective feeding of vitamins, improving profitability in animal husbandry. This applies especially to today's intensive farming of poultry, fish, and pigs and to calf breeding with milk substitute.

In Germany, the type and quantity of feed additives that are mixed with the different kinds of feedstuffs are stipulated by the feed law [15].

1.8. Antivitamins

Antivitamins are compounds that reduce or reverse in a specific manner the effect of vitamins. A distinction is made between compounds which destroy or capture the vitamin, and substances which compete with the vitamin for the binding site on the receptor (competitive inhibition).

Examples of the first class of compounds are, e.g., thiaminase, which cleaves the vitamin thiamine, and avidin, which forms a complex with biotin, making it unavailable for the body.

Antivitamins can be used to produce defined vitamin deficiency diseases, allowing the biochemical background of vitamin activity to be studied.

Some antivitamins are therapeutically applied, e.g., folic acid antagonists in the treatment of certain forms of leukemia and vitamin K antagonists as inhibitors of clotting.

1.9. Analysis of Vitamins

Methods of determination for vitamins which are based on the specific properties of vitamins include biological, microbiological, and enzyme activation processes.

Biological methods are mostly growth tests with experimental animals fed vitamin deficient diets. They were important especially for the checking the progress of enrichment during the isolation of vitamins. Today, these processes are used only for special purposes because they are difficult and tedious to conduct and evaluate and, consequently, inaccurate.

Microbiological methods are based on the fact that various vitamins are growth factors for certain microorganisms. The photometrically measurable turbidity of the culture medium

Table 5. Summary of studies on the effects of antioxidative vitamins

Period	Vitamin	Participants	Country	Results
1970 – 1980	β-carotene	3000 men	USA	the men who died of cancer during the study had 20 % less β-carotene in the blood than control persons
1980 – 1990	β-carotene, vitamins E, C	2000 nurses recovered from cardiac infarct or angina	USA	antioxidant vitamins reduce the risk of cardiac infarct by 33 % and stroke by 71 %
1985 – 1991	β-carotene, vitamin E	29 000 smokers	Finland	no preventive effect was detected during the study
1985 – 1991	β-carotene, vitamin E, Se	30 000 smokers	China	vitamins reduce risk of cancer
1991 – 1992	β-carotene, vitamin E	1400 persons: 50 % with acute cardiac infarct, 50 % control persons	Europe	infarct patients have a lower β-carotene and vitamin E status
1988 – 1992	β-carotene, vitamins E, C	900 patients with removed adenomas	USA	the treatment does not influence the probability of formation of new adenomas
1989 – 1994	β-carotene	400 albinos with skin diseases	Tanzania	β-carotene reduces the development of skin cancer
1990	β-carotene, vitamins E, C	500 men	UK	low plasma concentrations correlate with high risk of angina
1986 – 1990	vitamin E	40 000 men	USA	vitamin E reduces the risk of severe heart diseases by 37 %
1980 – 1988	vitamin E/multi-vitamin preparation	37 000 women	USA	vitamin E reduces the risk of coronary diseases by 40 %

is a measure of the vitamin concentration. Microbiological methods are characterized by high specificity and accuracy. They are used for the determination of vitamins in biological material and in pharmaceutical preparations (e.g., vitamins of the B group).

For clinical investigations (e.g., measurement of the vitamin level in the blood and urine), enzyme activation is commonly used because of its high specificity. It is based on the addition of the specific apoenzyme of the vitamin (coenzyme) to be determined and subsequent measurement of the activity of the holoenzyme.

The chemical and physical methods of determination used correspond to the normal methods for the analysis of organic substances. Modern chromatographic processes are particularly important.

In the case of vitamin preparations, the concentrations given refer to the quantity by weight. Only for vitamins A, D, and E, it is customary to state the international units as well.

1.10. Production

Vitamins are produced on an industrial scale by chemical synthesis or partial synthesis, by fermentation, or by extraction from natural material, whereby chemical synthesis is the dominant method.

The production processes used commercially are summarized in Table 6.

Table 6. Typical processes for the commercial production of individual vitamins ∗

Vitamin	Synthesis	Fermentation	Isolation
Vitamin A	+		+
		○	
Vitamin B_1	+	+	
Vitamin B_2	+		
		○	
Vitamin B_6	+	+	
Vitamin B_{12}		+	
Vitamin C	+		
Vitamin D_3	+		+
Vitamin E	+		+
Vitamin K	+		+
Biotin	+		
		○	
Folic acid	+		
		○	
Niacin	+		
Pantothenic acid	+		
		○	

∗ + Commercially used; ○ commercially possible.

Many vitamins have become large-scale products because of the constantly increasing

demand. The economic importance of the vitamins has increased to the same extent. At present, the worldwide market value for vitamins as a bulk product is ca. 26×10^9 €/a. In terms of value, the feed industry accounts for ca. 50 %, the pharmaceutical industry for ca. 30 %, and the food industry for ca. 20 %.

2. Vitamin A (Retinoids)

2.1. Introduction

In the broadest sense, the term vitamin A refers to a group of monocyclic diterpenes with similar biological activity. Since the term vitamin A is frequently linked to vitamin A alcohol [(all-*E*)-3,7-dimethyl-9-(2,6,6-trimethyl-1-cyclohexenyl)-2,4,6,8-nonatetraen-1-ol, retinol, vitamin A_1], and a large number of other natural and synthetic vitamin A derivatives exist, the designation retinoids is often employed [31].

In addition to retinol (**1**), there are retinoids that can be distinguished by their side chain and their ring system. These include retinal (**2**), retinoic acid (**3**), retinyl phosphate (**4**), the 3-dehydro compounds (**5**), (**6**), (**7**), and retroretinoids (**8**), (**9**), (**10**).

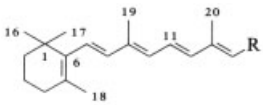

1 R = CH_2OH: retinol [*68-26-8*]
2 R = CHO: retinal [*116-31-4*]
3 R = CO_2H: retinoic acid [*302-79-4*]
4 R = $CH_2OPO_3H_2$: retinyl phosphate [*53859-19-1*]

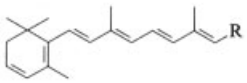

5 R = CH_2OH: 3-dehydroretinol [*79-80-1*]
6 R = CHO: 3-dehydroretinal [*472-87-7*]
7 R = CO_2H: 3-dehydroretinoic acid [*4159-20-0*]

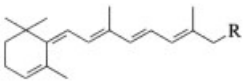

8 R = CH_2OH: retroretinol [*27597-84-8*]
9 R = CH_2OAc: retroretinyl acetate [*7631-44-9*]
10 R = CH_2: anhydroretinol [*16729-22-9*]

Vitamin A is found only in animal tissue, whereas the carotenoids known as provitamin A are found in the plant world. The latter are oxidatively degraded to vitamin A compounds in the animal organism.

The important provitamins A are β-carotene (**11**), α-carotene (**12**), γ-carotene (**13**), β-apo-8′-carotinal (**14**), β-carotene-5,6-epoxide (**15**), cryptoxanthin (**16**), and echinenone (**17**).

Provitamin A activity requires at least one cyclohexene or 5,6-epoxycyclohexane ring and one polyene chain equivalent with a chain length corresponding to that of vitamin A.

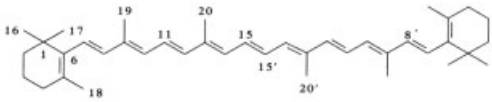

11 [*7235-40-7*]

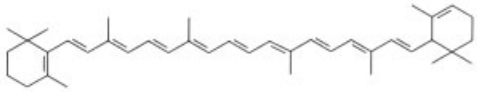

12 [*7488-99-5*]

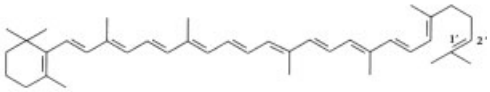

13 [*472-93-5*]

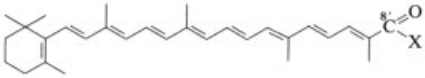

14 X = H [*1107-26-2*]
14a X = OR: β-apo-8′-carotinoic ester

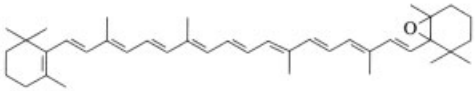

15 [*1923-89-3*]

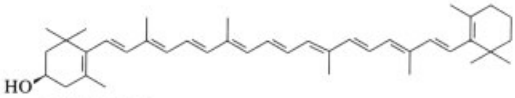

16 [*472-70-8*]

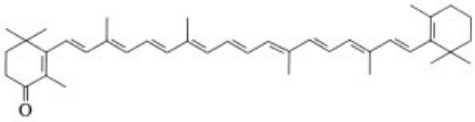

17 [*432-68-8*]

2.2. Historical Aspects

In 1876, SNELL showed that cod-liver oil could cure night blindness and xerophthalmia [32].

In 1880, LUNIN found that the growth and vital functions of experimental animals could not be maintained with carbohydrates, fats, and proteins alone, but that the addition of small quantities of milk powder was also required [33].

In 1909, STEPP extracted vital fat-soluble substances from milk [34]. In 1916, MCCOLLUM made a distinction between "fat-soluble A" and "water-soluble B" [35]. Later, the fat-soluble agent was called "vitamin A" on the proposal of DRUMMOND [36]. (The name "vitamin" had already been suggested by FUNK in 1912 when he isolated the anti-beriberi agent thiamin [37]).

Progress in the concentration of retinol (**1**) was made only after its sensitivity to oxygen and higher temperatures was recognized and the connection between the growth promoting effect and the color intensity in the Carr – Price test was discovered [38]. In 1931, KARRER isolated almost pure retinol (**1**) from the oil of a type of mackerel (*Scom-bresox saurus*) [39]. The first crystalline preparations were made in 1937 by HOLMES and CORBET [40] and in 1942 by BAXTER and ROBESON [41], who also succeeded in isolating crystalline (13*Z*)-retinol in 1946 [42]. The compound 3-dehydroretinol (**4**) was found in the liver oil of freshwater fish in 1937 [43], isolated in 1948 [44], and obtained in the crystalline form in 1962 [45]. As early as 1831, WACKENRODER [46] succeeded in isolating the orange-yellow pigment of the carrot (*Daucus carota*) which he called carotene. The wide distribution of carotene in the plant world was described by ARNAUD in 1887 [47]. It was only after improved isolation methods became available that the production of pure preparations was possible. In the course of this work, WILLSTÄTTER in 1907 was able to establish the empirical formula of carotene [48], which was thought to be a homogeneous compound.

In 1931 KUHN, KARRER et al. succeeded in isolating optically active α-carotene (**8**) by fractionated crystallization or chromatography on calcium hydroxide [49]. Of importance for the structural elucidation of the carotenoids was the discovery that the chromophoric group consists of conjugated olefinic double bonds [50]. From the amount of hydrogen absorbed in catalytic hydrogenation, it was calculated that carotene has 11 and the isomeric lycopene has 13 double bonds [51]. In 1929, the provitamin A activity of carotene was unambiguously demonstrated in animal experiments and it was simultaneously shown that the other carotenoids known at that time were completely inactive in this respect [52]. Thus, further experiments for elucidating the structures of carotene and vitamin A were closely connected. Important progress in determining the structure was made by means of oxi-dative degradation with potassium permanganate, chromic acid, and ozone to give characteristic fragments [53]. One of the first correct suggestions for the structure of retinol (**1**) was made by KARRER who had shortly before determined the structure of β-carotene [54]. The constitution of 3-dehydroretinol (**4**) was long disputed and finally established with its synthesis in 1952 [55].

In 1937, KARRER obtained β-apo-8′-carotinal by oxidative degradation of β-carotene with potassium permanganate [56]. The compound β-apo-8′-carotinoic acid ((**14a**) with X = OH) was isolated from egg yolk after feeding laying hens with **14** [57].

2.3. Physical Properties

The melting points and the UV spectroscopic data of vitamin A compounds and their most important derivatives are presented in Table 7.

Retinol (**1**), retinal (**2**), and the dehydro compounds **5** and **6** are readily soluble at room temperature in fats, oils, and common organic solvents. In contrast, retinoic acid (**3**) is appreciably soluble only in chlorinated hydrocarbons.

High-vacuum distillation can be used for the purification of retinol (**1**), its esters, and retinal (**2**) (retinol at 125 °C, retinyl acetate at 132 °C and 400 Pa). The double bonds in the side chain also allow the existence of *Z* isomers, some of which occur in nature. Thus, (all-*E*)-retinol is frequently accompanied by the (13*Z* isomer (**18**). The compound (11*Z*)-retinal (**19**) is of importance as the prosthetic group of the protein opsin [58].

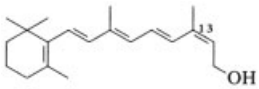

18 [*2052-63-3*]

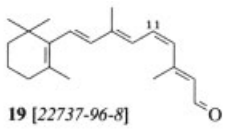

19 [*22737-96-8*]

Table 7. Characteristic data of vitamin A compounds and their most important derivatives ∗

Compound	*mp*, °C	UV absorption (in Ethanol)		Biological Activity
		λ, nm	$\varepsilon_{1\,cm}^{1\,\%}$	in % ∗∗
Retinol	62 – 64	325	1835	100
(13*Z*)-Retinol	58 – 60	328	1686	75
Retinal	61 – 62	381	1530	91
(11*Z*)-Retinal	64 – 65	254, 290, 376	614, 412, 878	48
Retinoic acid	179 – 180	350	1510	
(13*Z*)-Retinoic acid	174 – 175	354	1325	
Retinyl acetate	57 – 60	326	1550	
Retinyl palmitate	28 – 29	327	975	
3-Dehydroretinol	63 – 65	276, 286, 350	555, 715, 1455	40
3-Dehydroretinal	78 – 79	(314), 401	395, 1470	
3-Dehydroretinoic acid	183 – 184	(305), 370	443, 1395	

∗ The values refer to the (all-*E*) isomers unless otherwise noted.
∗∗ Curative growth test in rats.

In comparison with the other stereoisomers of retinoic acid, (9*Z*)-retinoic acid (**20**) exhibits the highest binding affinity for the nuclear receptor RXR α [59], resulting in an altered gene expression as a consequence of direct ligand interaction.

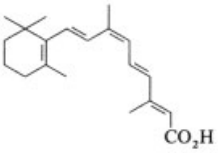

20 [*5300-03-8*]

With four double bonds in the side chain of vitamin A, 16 stereoisomeric forms (*E* and *Z* isomers) are theoretically possible. In the individual isomers, the steric effects allow a distinction to be made with regard to the stability of hindered and unhindered *cis* forms. While the trisubstituted double bonds (9 and 13) of the side chain are quite stable in the (*Z*) configuration, the disubstituted bonds (7 and 11) rearrange readily to the (*E*) form (**21**) [60].

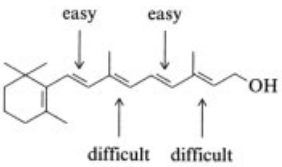

21

In the case of unhindered *cis* double bonds, the 1,4 steric interaction of two H atoms is weak, but strong interaction occurs between a H atom and a methyl group or between two methyl groups in hindered *cis* double bonds. Therefore, facile isomerization from the hindered *cis* forms to the *trans* forms is observed.

Thus, the four stable retinal isomers (all-*E*; 9*Z*; 13*Z*; 9*Z*,13*Z*) are well known. Of the remaining 12 isomers, however, only four of the labile forms (7*Z*; 11*Z*; 7*Z*,9*Z*; 11*Z*,13*Z*) are known [61]. Other labile *Z* forms have been synthesized under special conditions [62, 63].

X-ray analyses of retinol (**1**), its esters [64], and retinal (**2**) [65] assign a 6,7-*s*-*cis* conformation to the molecule in the crystal lattice.

Two different crystal modifications of retinoic acid have been determined [66]. In addition to the stable 6,7-*s*-*cis* form (**22**, triclinic modification), the labile 6,7-*s*-*trans* form (**23**, monoclinic modification) is also known.

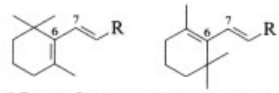

6,7-*s*-*cis*-form **22** 6,7-*s*-*trans*-form **23**

In the all-*trans* forms, the polyene chain has a saber-like curvature which results from the interaction of the methyl group in the side chain with the hydrogen on the next carbon atom but one.

The provitamins A have colors which vary from bright yellow to red, due to the system of conjugated double bonds.

With increasing numbers of double bonds and with the introduction of carbonyl groups conjugated to the double bond system, the maxima in

the UV and visible region shift to longer wavelengths.

The melting points, UV spectroscopic data, and biological activity of the most important provitamins A are listed in Table 8.

Both α- and β-carotene (**12** and **11**) are readily soluble in chlorinated hydrocarbons and in carbon disulfide, and poorly soluble in ether, lower alcohols, and in aromatic or aliphatic hydrocarbons. In comparison with the carotene hydrocarbons (**11**) and (**12**), the β-apocarotenoids are in general more readily soluble. Data on the solubility of (all-*E*)-β-carotene (**11**) in g/100 mL at 20 °C follow:

Water	0.0
Ethanol	< 0.01
Fats and oils	0.05 – 0.08
Orange oil	0.2 – 1
Cyclohexane	ca. 0.1
Ether	ca. 0.1
Acetone	ca. 0.1
Chloroform	ca. 3

The provitamins A occur in fresh plant tissue in the (all-*E*) form. During drying and storage, however, partial conversion to the *Z* forms takes place. Heating in solution or in the presence of iodine causes isomerization. This process produces all-*E* compounds as well as 9*Z* and 13*Z* forms as the main components. In comparison with the all-*E* compounds, the *Z* isomers have a lower extinction and biological activity [67].

Table 9. Typical reactions of retinoids. In principle, they also apply to the 3-dehydro forms.

Retinoids	Reagents	Products
Retinol	manganese dioxide	retinal (**2**) [68, 69]
	acetic anhydride (Ac_2O)	retinyl acetate
	palmitoyl chloride	retinyl palmitate
	maleic anhydride	Diels – Alder adducts
	singlet oxygen	5,8-peroxide (**24**) [70]
	peroxy acids	5,6-epoxide (**25**) [71]
	ozone	geronic acid
	potassium permanganate	acetic acid
Retinal	sodium borohydride	retinol [72]
	silver oxide	retinoic acid [73]
	manganese oxide	4-oxoretinal [74]
Retinoic acid	alcohols	retinoic esters
	glucuronic acid	glucuronides
	lithium aluminum hydride	retinol [75]

2.4. Chemical Properties

Retinol (**1**) and its esters are readily oxidized by atmospheric oxygen and other oxidizing agents, especially in the presence of light or at higher temperatures. The addition of antioxidants (e.g., α-tocopherol) or storage in the cold with the exclusion of light and oxygen ensures a certain stability. Some typical reactions of vitamin A are listed in Table 9.

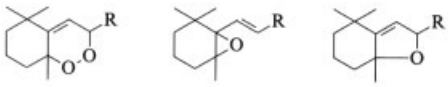

24 [*29443-90-1*] **25** [*801-72-9*] **26** [*744-98-9*]

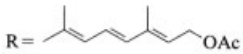

In all these reactions, the stereochemistry of the polyene side chain remains unchanged provided light and oxygen are excluded.

Vitamin A is resistant to bases, but retinol (**1**) and its esters undergo rearrangement and elimination reactions in the presence of acids. Dilute acids cause the conversion of retinol (**1**) to anhydroretinol (**10**).

In dichloromethane, retinyl acetate is converted in the presence of hydrogen bromide to retroretinyl acetate (**9**) within seconds. The treatment of vitamin A compounds with halogenating agents (e.g., *N*-bromosuccinimide) results in the preferential halogenation of the methylene group at position 4. Subsequent base-catalyzed elimination of hydrogen halide provides access to the 3-dehydroretinoids (**5**, **6**, **7**), which, however, have no significant vitamin A activity. Strong Lewis acids react with vitamin A to produce characteristically colored complexes (e.g., blue color of complex with $SbCl_3$ in the Carr – Price reaction), which form the basis of analytical identification [76]. With dienophiles (e.g., maleic anhydride), retinol (**1**) and its esters undergo typical [4 + 2] cycloaddition reactions. The 11,13-di-*trans* configuration reacts preferentially, allowing easy differentiation and separation of the all-*trans* and 9-*cis* compounds from the slower reacting 13-*cis*, 11-*cis*, 11,13-di-*cis*, and 9,13-di-*cis* stereoisomers [82].

The provitamins A are also sensitive to oxidation. The degradation of β-carotene with potassium permanganate leads to β-apo-12′- and β-apo-8′-carotinal (**14**) [77]. Chromic acid cleaves the ring double bond with formation of open-chain diketones [78]. Peroxy acids preferentially give 5,6-epoxides, which rearrange to 5,8-di-

Table 8. Characteristic data of the most important provitamins A [(all-E) isomers]

Compound	*mp*, °C	UV absorption (in Ethanol)		Biological Activity
		λ, nm	$\varepsilon_{1\,\mathrm{cm}}^{1\,\%}$	in % ∗
β-Carotene	182 – 183	456, 484	2500, 2200 (Cyclohexane)	100
(*R*)-(+)- α-Carotene	187 – 188	422, 446, 474	1790, 2720, 2500 (Hexane)	54
β-Apo-8′-carotinal	138 – 139	461, 488	2640, 2165 (Cyclohexane)	72
β-Apo-8′-carotinoic acid ethyl ester	137 – 138	449, 475	2550, 2140 (Cyclohexane)	25, 78

∗ Curative growth test in rats.

hydrofuran derivatives (**26**) under acid catalysis [79].

The reaction of provitamin A with *N*-bromosuccinimide or iodine preferentially leads to 4,4′-dihalo compounds [80], which allow the synthesis of 4,4′-dioxo-β-carotene (canthaxanthin) by continued oxidation [81].

The *E,Z* isomeric vitamin A compounds differ considerably in their biological activity [83]. The highest activity is exhibited by (all-*E*)-retinol and its esters (see Table 7). Industrially, retinyl acetate is isomerized either photochemically [84] or in the presence of catalytic amounts of iodine or palladium compounds [85]. The all-*E* form, which predominates at equilibrium, can be separated by crystallization from hexane or methanol.

2.5. Occurrence

Retinol (**1**) occurs exclusively in animal organs, especially in the liver. Generally, it is esterified with higher fatty acids (e.g., palmitic acid). Relatively high concentrations are found in milk, eggs, and blood plasma.

Vitamin A is largely formed by degradation of carotenoids ingested with food. Carotenoids are synthesized only by higher plants and microorganisms. After consumption carotenoids, also known as provitamins A, are enzymatically converted to vitamin A, especially in the cells of the intestinal mucosa [86]. The compounds α- and β-carotene (**12** and **11**) are permanent companions of chlorophyll. High concentrations are found especially in paprika, carrots, and in palm oil (400 – 800 ppm) [87], which is used to color edible fats. The β-apo-carotinals mainly occur in citrus fruits. However, their concentration is too low to permit economic commercial isolation [88].

2.6. Biosynthesis

In the animal organism, vitamin A compounds occur as metabolites of the provitamins A, especially of β-carotene (**11**).

The C_{40} carotenoids are tetraterpenes. They are biosynthesized according to the principle valid for all terpenes [89] (see Scheme 1; → Terpenes, Chap. 1.3.). The C_5 structural unit 1-isopentenyl pyrophosphate (**28**) is made from three molecules of acetyl coenzyme A via mevalonic acid (**27**). Isomerization of **28** gives the more reactive 3,3-dimethylallyl pyrophosphate (**29**). The linkage of these two components starts a chain of alkylation steps which leads to the C_{20} intermediate geranylgeranyl pyrophosphate (**30**) and is interrupted by dimerization to give (15*Z*)- or (all-*E*)-phytoene ((**32a**) and (**32b**)), the first C_{40} carotenoid. The intermediate in this reaction is prephytoene pyrophosphate (**31**). Dehydrogenation of **32** gives lycopene (**33**) via phytofluene, ζ-carotene, and neurosporin. From lycopene, α- and β-carotene (**12** and **11**) are formed by stepwise cyclization.

The subtle stereochemical problems involved in the individual steps have been elucidated by the use of mevalonic acid (**27**), stereospecifically labelled at the prochiral centers C-2, C-4, and C-5 (**27′**). Thus, it is known that the anhydrodecarboxylase step **27** → **28** proceeds as a trans elimination. In the isomerase reaction **28** → **29**, a proton is added at the *re,re* side of the double bond, followed by eliminaton of the 2-*pro-R* proton of **28** (H_{4S} of **27′**).

The prenyltransferase steps **29** →**30** also proceed with elimination of the 2-pro-R proton (Scheme 1) of **28** and with inversion at C-1 of the allylpyrophosphate components (C-5 of **27′**). An enzymatic addition – elimination mechanism leads to *E* double bonds.

In the linkage of two molecules of geranylgeranyl pyrophosphate (**30**) to form (15*Z*)-phytoene (**32a**), the 5-*pro-S* hydrogen atoms of mevalonic acid are lost. In the dehydrogenation **32** → **33**, double bonds in the *E* configuration are introduced stepwise with the elimination of one 2-*pro-S* and one 5-*pro-R* hydrogen atom of mevalonic acid (**27′**).

The cyclization **33** → **11** or **12** is started by,*re,re* addition of a proton to C-3 of the terminal double bond (C-4 of **27′**) and intramolecular alkylation of the 6,7 double bond. Cleavage of a proton from C-6 and C-4 (C-4 and C-2 of **27′**) results in the parallel formation of the α- and β-ionone rings. The methyl group from C-2 of mevalonic acid (**27′**) occupies the *pro-S* position.

From a stereochemical standpoint, the reaction sequence can be represented by the labelling scheme **27′** → **30′** → **33′** → **12′** (Scheme 2).

The conversion of the provitamins A to vitamin A compounds takes place in the intestinal mucosa. In this process, α- and β-carotene (**12** and **11**) are cleaved by oxygen in the presence of carotene 15,15′-dioxygenase into two molecules of retinal (**2**) [90]. Similar enzymes also catalyze the oxidative degradation of β-apocarotinals [91]. Retinal (**2**) is predominantly reduced to retinol (**1**) and partly oxidized to retinoic acid (**3**).

After transfer into the intestinal cells, retinol (**1**) is esterified with higher fatty acids and released into the lymph in this form [92]. After the direct vitamin requirement of the organism is covered, the retinyl esters are stored in the liver [93]. Free retinol (**1**) predominates in the blood stream; a specific protein is responsible for transport [94].

2.7. Production

2.7.1. Isolation and Purification from Natural Raw Materials

Owing to the sensitivity of retinol (**1**) to air, higher temperatures, and acids, it must be isolated under especially mild conditions (extraction with organic solvents under protective gases, e.g., CO_2, N_2). Since retinol (**1**) and its esters occur in the free form in tissues and organs, the first concentration step can be extraction with heptane. For retinal (**2**), however, special isolation methods are required because it is covalently bound to primary amino groups in the form of a Schiff's base [95].

In general, vitamin A is isolated by liquid–liquid extraction, molecular distillation [96], and HPLC [63]. Commercial isolation of retinol (**1**) from natural raw materials starts with fish oils and is conducted in six steps: (1) accumulation of crude vitamin A by the Solexol countercurrent extraction process [97], (2) short-path distillation of the retinyl esters (200 °C, 400 Pa), (3) ester saponification (KOH, 60–70 °C, 30 min, N_2), (4) extraction of retinol (**1**) with ether, (5) short-path distillation of retinol (**1**) (120 °C, 400 Pa), and (6) final purification by crystallization from ethyl formate (−35 °C).

With the development of efficient processes for the synthesis of vitamin A on a commercial scale, isolation from natural raw materials has lost importance.

2.7.2. Industrial Synthesis

Since retinol (**1**) was first synthesized by KUHN [98], numerous strategies for its synthesis have been developed. Vitamin A is formed by the linkage of various cyclic components with the corresponding open-chain compounds [99, 100]: $C_{10}+C_{10}$ [101], $C_{11}+C_9$ [102], $C_{13}+C_7$ [103], $C_{14}+C_6$, $C_{15}+C_5$, $C_{16}+C_4$, $C_{18}+C_2$, and $C_{18}+C_1+C_1$.

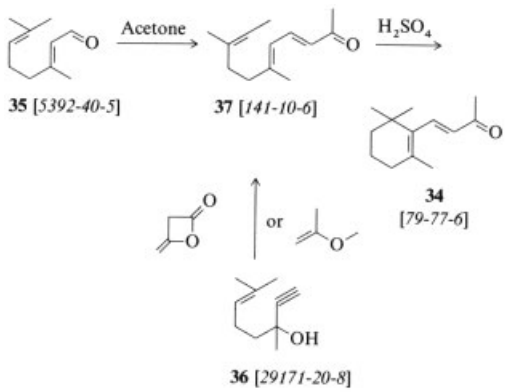

The key building block of all industrial vitamin A syntheses is β-ionone (**34**), which can be made from citral (**35**) [104] and acetone or from

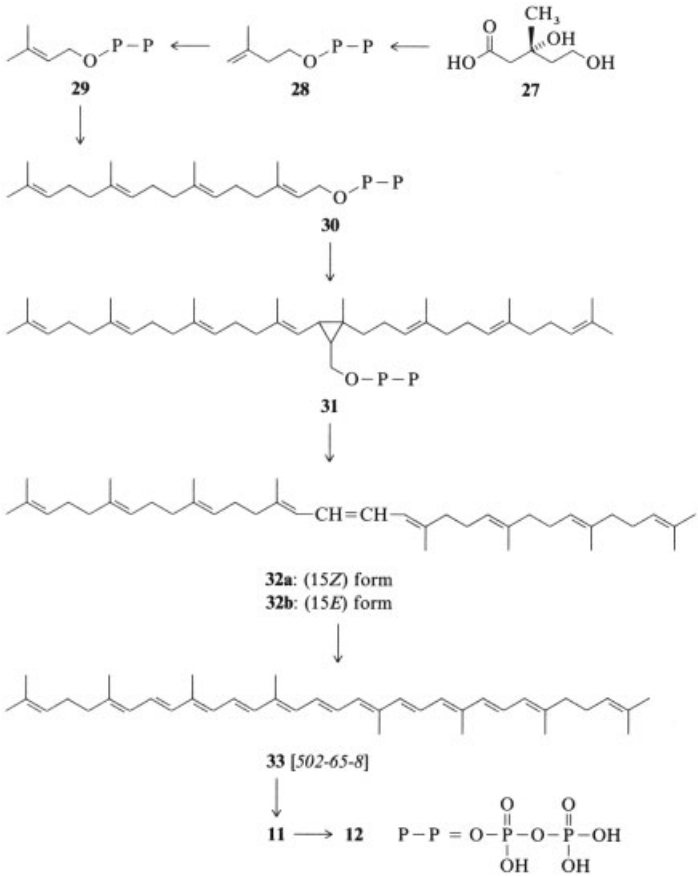

Scheme 1.

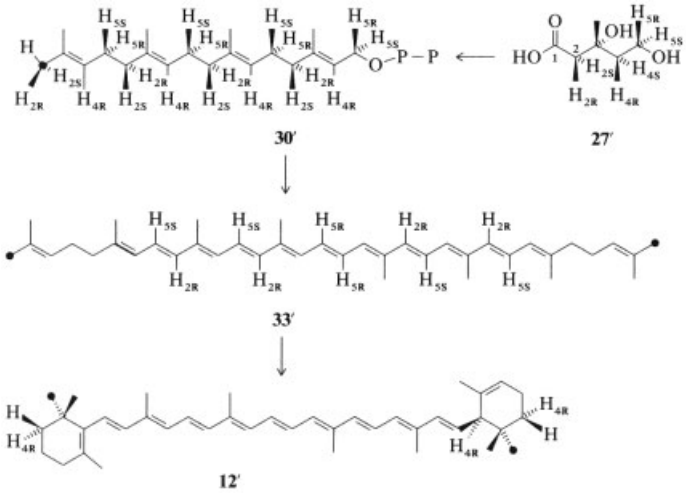

Scheme 2.

dehydrolinalool (**36**) [105] and diketene [106] or isopropenyl methyl ether [107] via pseudoionone (**37**) [108].

Synthesis of the C_{20} Skeleton of Vitamins A. $C_{15} + C_5$ *(Sumitomo)* [109]. Knoevenagel condensation of β-ionylidene acetaldehyde (**38**) [110] with (*E*)-3-methyl-2-butenoic ester using potassium amide in ammonia produces (all-*E*)-retinoic ester (**39**). Reduction with lithium aluminum hydride and subsequent acetylation yields retinyl acetate (**40**).

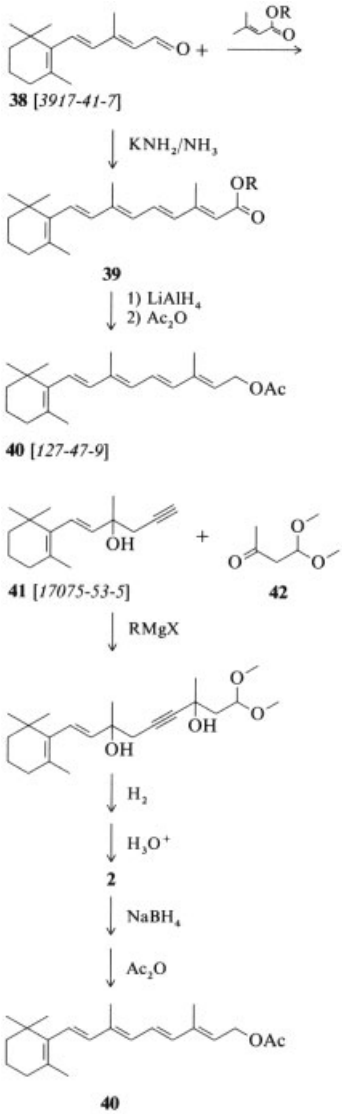

$C_{16} + C_4$ *(DPI, Glaxo)* [111]. The reaction of β-ionone (**34**) with 2-propinyl chloride in the presence of zinc yields β-propargyl-β-ionone (β-C_{16} alkinol) (**41**). Grignard reaction with 4,4-dimethoxy-2-butanone (**42**), followed by partial hydrogenation and hydrolysis of the acetal as well as elimination of water, gives retinal. Retinal is then converted to retinyl acetate (**40**) by reduction and subsequent acetylation.

$C_{18} + C_2$ *(Philips)* [112]. β-Ionone (**34**) and the homologous β-C_{18} ketone (**43**) are converted to β-ionylidene acetaldehyde (**38**) and retinal (**2**) by condensation with cyanoacetic ester and reduction of the nitriles **44** and **45** with diisobutyl aluminum hydride. The β-C_{18} ketone (**43**) is made from β-ionylidene acetaldehyde (**38**) by condensation with acetone.

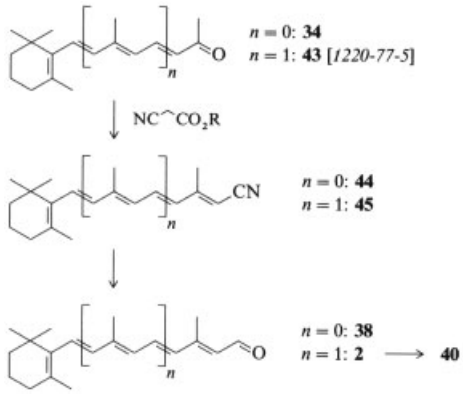

$C_{18} + C_1 + C_1$ *(AEC)* [113]. Starting with β-ionone (**34**) and β-C_{18} ketone (**43**), a stepwise conversion to β-ionylidene acetaldehyde (**38**) and retinal (**2**) is carried out. Condensation of the ketones **34** and **43** with formic ester gives the β-C_{14} and β-C_{19} enolates, which yield the hydroxyacetals **48** and **49** after acetal formation (**46** or **47**) and Grignard reaction. Subsequent acetal hydrolysis and water elimination give the final products of the reaction sequence **38** and **2**.

$C_{14} + C_6$ *(Hoffmann LaRoche)* [114]. β-Ionone (**34**) is converted to the β-C_{14} aldehyde (**50**) in a Darzens glycidic ester synthesis [115,116]. Subsequent Grignard reaction with 3-methyl-2-penten-4-in-1-ol(**51**) yields the C_{20}alkynediol (**52**).

(All-*E*)-retinyl acetate (**40**) is obtained after partial hydrogenation of the triple bond to give the (11*Z*, 13*Z*) compound **53**, acetylation of the

primary hydroxyl group, and subsequent acid-catalyzed dehydration with isomerization of the double bonds into the *E* configuration.

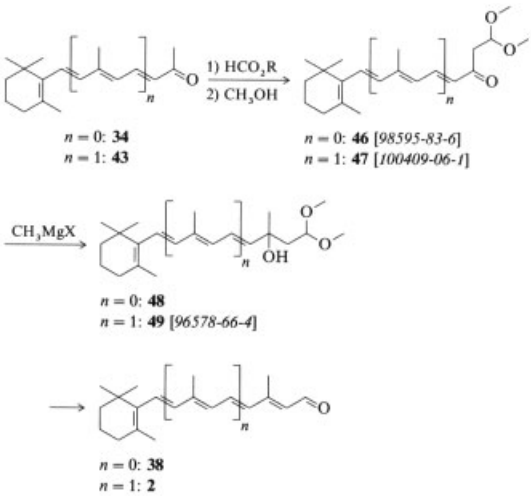

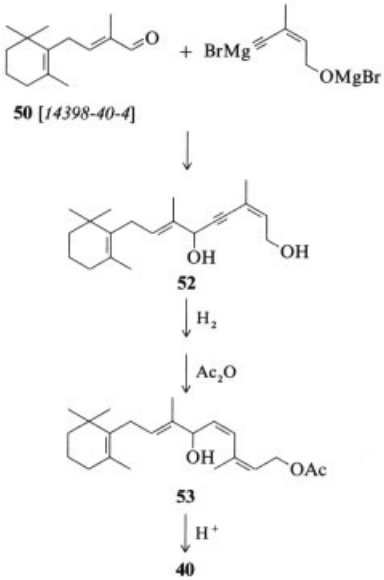

The C_6 structural unit **51** is made from methyl vinyl ketone by ethynylation and alkyl rearrangement of the tertiary alkyne carbinol **54** [117].

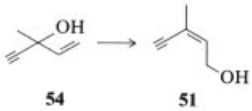

$C_{15} + C_5$ *(BASF)* [118] (Scheme 3). The synthesis of the C_{20} skeleton is achieved by using the Wittig reaction [119]. Alkylidenetriarylphosphanes are obtained from alkyltriarylphosphonium salts by treatment with base [120–122]. The formation of olefins with simultaneous release of triarylphosphine oxide is achieved by condensation of these alkylidene triarylphosphanes with carbonyl compounds [123].

The C_{15} phosphonium salt **56** can be made from β-ionone (**34**) in two different ways. β-Vinylionol (**55**) is obtained by ethynylation and subsequent partial hydrogenation [124] or by vinylation [125]. Treatment with triphenylphosphine and acid yields the C_{15} phosphonium salt **56** [126], the key building block for the synthesis of retinoids. Retinyl acetate (**40**) is prepared by Wittig reaction with the C_5 aldehyde **57**, which is accessible by various routes [127, 128], e.g., by rhodium-catalyzed hydroformylation of 1,2-diacetoxy-3-butene (**58**) [129]. Coupling of **56** with **57** is carried out in organic solvents such as alcohols or DMF, or in aqueous solution with mild bases (e.g., alkali metal carbonate or ammonia) [130]. In this way, a mixture of the double bond isomers 11*Z* and all-*E* is obtained in a yield of ca. 90 %. The labile 11*Z* isomers can be converted to the all-*E* form in various ways [131].

If the C_{15} phosphonium salt **56** is reacted with the alternative C_5 building blocks **59** [132], **60** [133], or **61** [134], retinal (**2**), retinoic acid (**3**), or (13*Z*)-retinoic acid (**62**) is obtained [if necessary after hydrolysis and isomerization of the (11*Z*) form].

$C_{15} + C_5$ *(Rhône-Poulenc)* [135] (Scheme 4). This synthesis of vitamin A is based on a two-step olefin synthesis [136] using Julia sulfone chemistry. Allyl phenyl sulfones are metalated in the α-position and alkylated with alkyl halide. The double bond is formed by elimination of benzenesulfinic acid. The coupling and elimination steps both require comparatively strong bases such as potassium *tert*-butoxide or other alkali metal alkoxides [137].

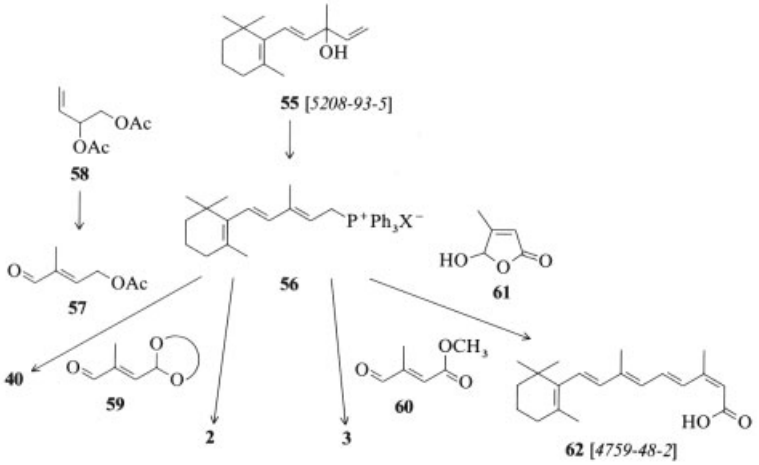

Scheme 3.

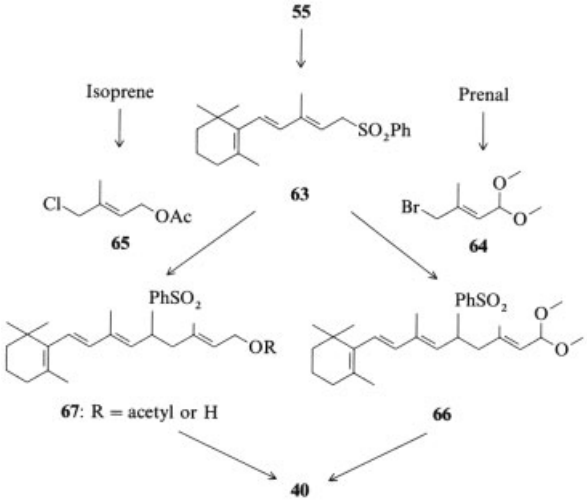

Scheme 4.

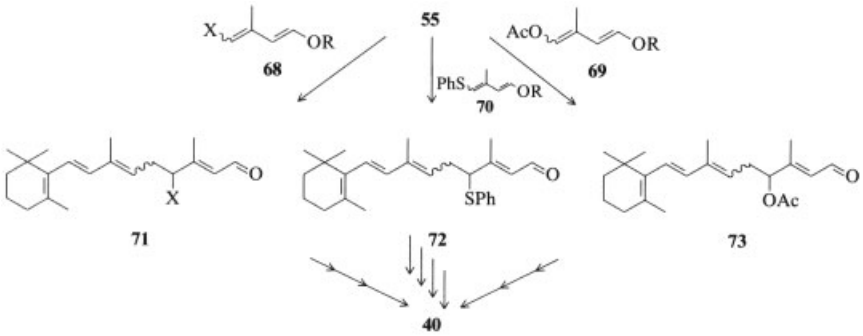

Scheme 5.

As in the BASF process, the C_{15} building block is obtained from β-vinylionol (**55**). The reaction of β-vinylionol with sodium phenylsulfinate yields the C_{15} sulfone **63** [138].

Subsequent reaction of **63** with 1,1-dialkoxy-4-bromo-3-methyl-2-butene (**64**) [139] or 1-acetoxy-4-chloro-3-methyl-2-butene (**65**) [140] yields the C_{20} sulfones **66** or **67** [141]. Elimination of benzenesulfinic acid, followed by acetylation affords retinyl acetate (**40**) (if necessary after hydrolysis of the acetal group and reduction of the aldehyde [142]). While the C_5 building block **64** is obtainable from prenal [143], **65** is derived from isoprene [144].

Formation of the polyene skeleton is also possible by treating β-vinylionol (**55**) with γ-halogeno (**68**), γ-acetoxy (**69**) or γ-thiophenyl dienol ethers (**70**), catalyzed by Lewis acids [145]. The primary coupling products **71**, **72**, and **73** can be converted to retinyl acetate (**40**) via a reaction cascade (Scheme 5). The C_5 building blocks used are obtainable from prenal. A commercial application of this synthesis has not been realized.

$C_{10} + C_{10}$ *(Kuraray, Solvay – Duphar)* [146]. On the basis of sulfone chemistry, the synthesis of retinoids is achieved by coupling two C_{10} building blocks [147]. The common starting material for both compounds is myrcene (**74**) or linalool (**75**), which is converted in a multistep sequence to β-cyclogeranyl sulfone (**76**) and 8-acetoxy-2,6-dimethyl-2,6-octadienal (**77**) [148]. The β-hydroxysulfone **78** is obtained by α-metalation of the C_{10} sulfone **76**, followed by condensation with the C_{10} aldehyde acetate **77**. The β-hydroxysulfone is then converted to retinyl acetate (**40**) via the derivatives **79**, **80** [149] (Scheme 6).

$C_{13} + C_2 + C_5$ *(L'Oreal)* [150]. Starting from β-ionone (**34**), β-ethynylionol (**81**) is produced by ethynylation. In a subsequent Grignard reaction, coupling with β-formyl crotonate (**60**) is carried out, giving the corresponding acetylene ester **82**, which after reduction to the triol is selectively etherified at the primary alcohol group.

The resulting ether diol **83** is stereospecifically reduced to retinyl ether (**84**) after partial hydrogenation of the triple bond with a mixture of titanium trichloride and lithium aluminum hydride.

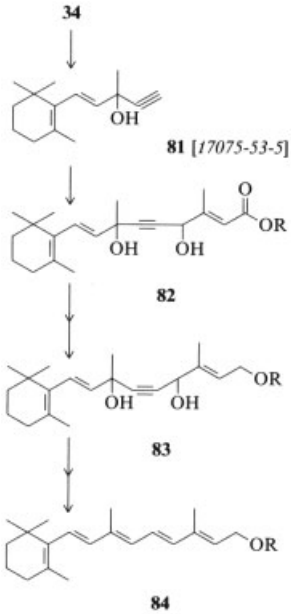

$C_{13} + C_2 + C_3 + C_2$ [151]. Another path leading to retinal (**2**) is via the vinylogization of carbonyl compounds with the formation of α,β-unsaturated aldehydes and methyl ketones. The building blocks for a synthesis of this type are 2-lithiovinyloxy(trialkyl)silanes (**85**) and 2-lithio-1-methylvinyloxy(trialkyl)silanes (**86**). In a further development of this reaction principle [152], retinal (**2**) is obtained with the C_5 building blocks **87**, **88** or with the C_7 building block **89** (Scheme 7).

More recently that concepts have been developed for the synthesis of polyenes via Heck coupling of C – C single bonds [153].

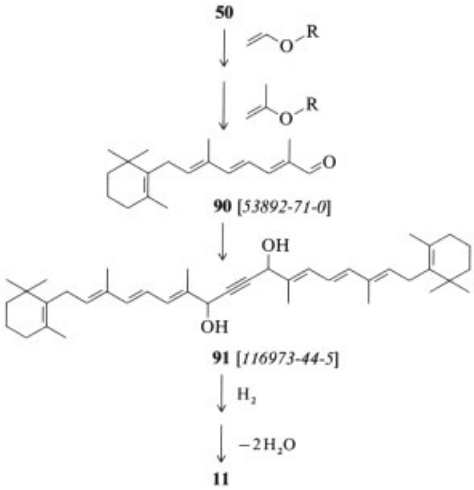

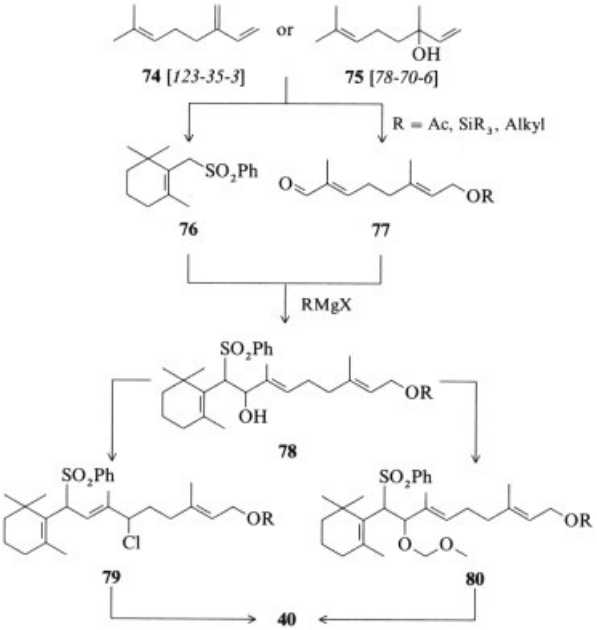

Scheme 6.

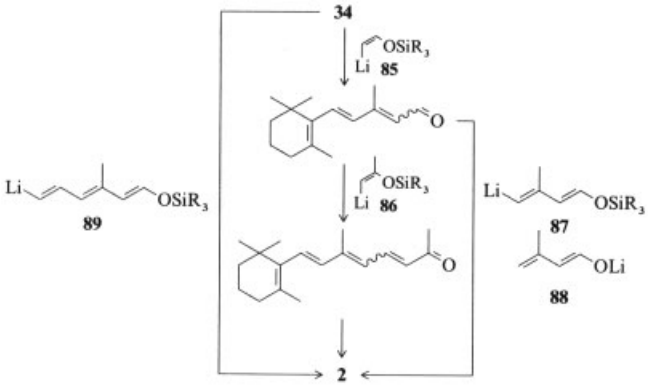

Scheme 7.

Synthesis of the C_{40} Skeleton of Provitamins A. The syntheses of the provitamins A are closely related to those of the vitamin A compounds.

$C_{19} + C_2 + C_{19}$ (Hoffmann LaRoche) [154]. β-Carotene is produced from the β-C_{14} aldehyde **50**, an intermediate in the synthesis of vitamin A. **50** is converted to the β-C_{19} aldehyde **90** in two steps: First, with ethyl vinyl and then with ethyl isopropenyl ether in a Müller – Cunradi reaction. Subsequently, two molecules of **90** are linked together by acetylene giving the C_{40} diol **91**, which after partial hydrogenation and dehydration yields β-carotene (**11**).

Other syntheses of β-carotene (**11**) are mainly based on the building principles $C_{15} + C_{10} + C_{15}$ and $C_{20} + C_{20}$. While the first synthetic sequence comprises the Wittig condensation between two molecules of C_{15} phosphonium salt **56** and 2,7-dimethyl-2,4,6-octatrien-1,8-dial (**92**) [155], β-carotene can also be made by the condensation of retinal (**2**) with retinyltriphenylphosphonium salt (**93**), obtainable from retinyl acetate (**40**) [156, 157].

The synthesis of β-carotene (**11**) is also possible by using the Julia reaction [158]. Retinyl phenyl sulfone (**94**) is made by the reaction of retinol (**1**) with sodium benzenesulfinate. α-

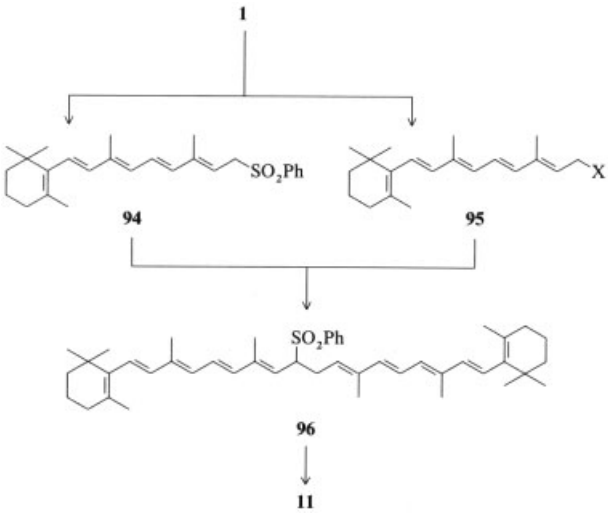

Scheme 8.

Metalation with potassium *tert*-butoxide to give the C_{40} sulfone **96** is followed by alkylation with retinylhalide (**95**). Subsequent elimination of sulfinic acid yields β-carotene (**11**) (Scheme 8).

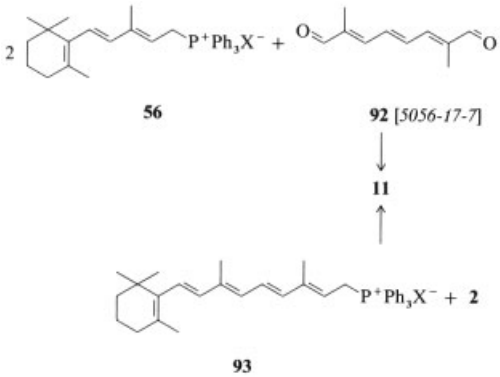

A new concept for the synthesis of β-carotene (**11**) uses the precursors of the vitamin A process of Kuraray ($C_{10} + C_{10}$) **79** or **80**. After selective oxidation to the corresponding aldehydes **97**, **98**, these compounds are subjected to Wittig condensation with retinyltriphenylphosphonium salt (**93**) [159]. Subsequent double elimination results in formation of β-carotene (**11**).

Other processes for the synthesis of β-carotene are based on the linking of two identical C_{20} components. Both the reductive dimerization of retinal in a manner analogous to the McMurray reaction [160] with lithium aluminum hydride and titanium trichloride, as well as a modified Wittig condensation with oxidative coupling of retinyltriphenylphosphonium hydrogensulfate (**93**, $X = HSO_4^-$) [161] with potassium carbonate and hydrogen peroxide in aqueous solution yield β-carotene (**11**).

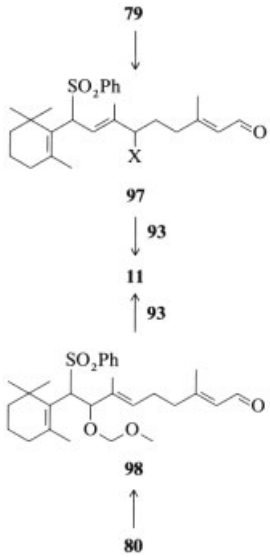

The coupling of the C_{20} sulfone **99** with retinal (**2**) and the reaction of two molecules of C_{15} sulfone **63** with the C_{10} dialdehyde **92** have been described in the literature as possible methods for the synthesis of β-carotene [162].

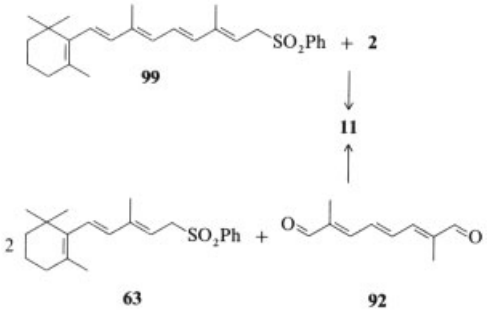

All syntheses of β-carotene initially produce mixtures of *E,Z* isomers. As a result of the good crystallization properties of the all-*E* form [163], these mixtures can be almost quantitatively isomerized to the all-*E* form under isomerizing conditions [164].

Synthesis of the C_{30} Skeleton of Apocarotenoids. The syntheses of β-apo-8′-carotenoids (**14**), (**14a**) follow the paths $C_{19}+C_6+C_2+C_3$, $C_{15}+C_{10}+C_5$, and $C_{20}+C_{10}$ [165].

$C_{19}+C_6+C_2+C_3$. Starting with the C_{14} aldehyde **50**, the C_{19} aldehyde **90** is made via enol ether condensation, followed by reaction with 2-methylpent-2-en-4-yne-1-al diethyl acetal (**100**) in a Grignard reaction. After acetal cleavage and dehydration, 15-dehydro-β-apo-12′-carotinal (**101**) is accessible. Subsequent acetal formation, enol ether condensation [166], and hydrolysis yields 15-dehydro-β-apo-10′-carotinal (**102**). Further enol ether condensation or Wittig reaction with the corresponding C_3 building blocks and subsequent partial hydrogenation allows the synthesis of **14** and **14a** [167].

$C_{15}+C_{10}+C_5$. Wittig condensation of the C_{15} phosphonium salt **56** with the C_{10} dialdehyde **92** first produces β-apo-12′-carotinal (**103**), which can be condensed to **14** or**14a** by further treatment with C_5 alkylidenephosphanes **104**, **105** [168].

$C_{20}+C_{10}$. Wittig condensation of cyclic C_{20} components with open-chain C_{10} building blocks allows the reaction of both retinal (**2**) with the corresponding C_{10} alkylidenephosphane (**106** with Y = OR or O) [169] as well as retinyl-(triphenyl)phosphonium salt (**93**) with C_{10} aldehydes (**107** with X = H or OR) [170] (Scheme 9). As with β-carotene, isolation of the

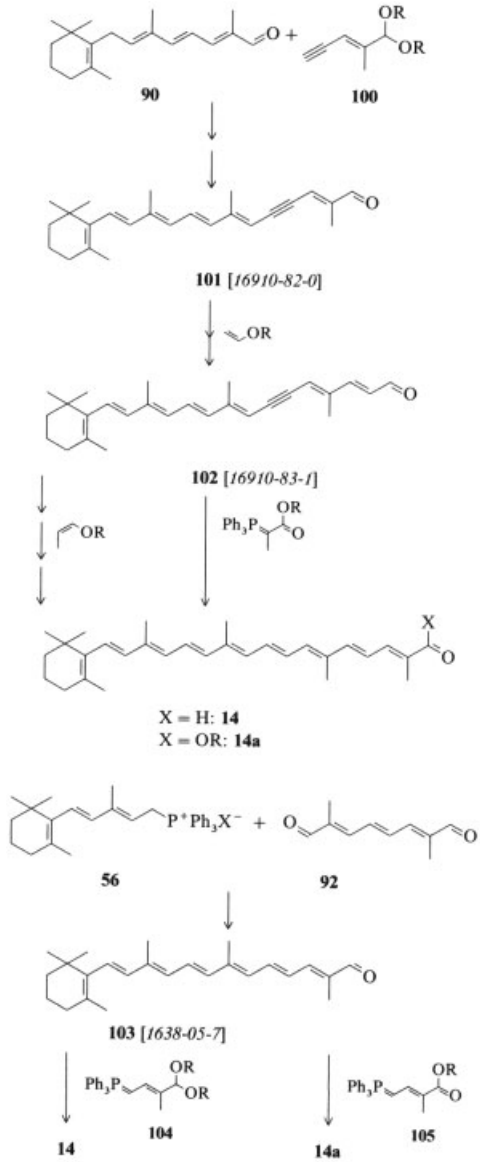

all-*E* isomer exploits its good crystallizability, which allows almost quantitative isolation under isomerization conditions [171].

2.8. Metabolism and Physiological Functions

2.8.1. Preformed Vitamin A

About 90 % of the vitamin A stored in the liver is in the form of retinyl palmitate. Free retinol

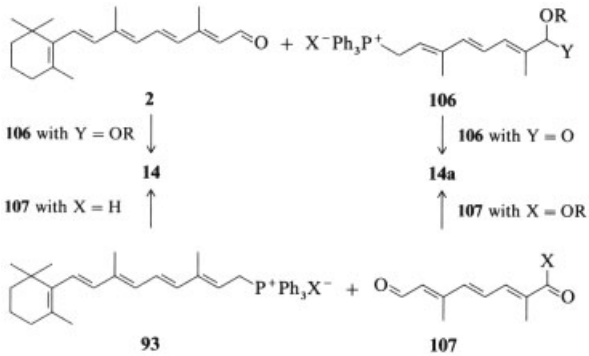

Scheme 9.

is esterified here. The liver has a high storage capacity for retinyl palmitate.

Mobilization from the liver occurs as required. Retinyl palmitate is enzymatically hydrolyzed and the resulting retinol is released into the blood plasma by means of a specific retinol binding protein (RBP) and transported to the target cell [172].

At the receptor of the target cell, retinol is passed on to a cellular RBP (CRBP). As required, retinol can now be used as such, metabolized to retinoic acid, or reesterified to retinyl palmitate, which can be stored in the cell for a short time and used to bridge a temporary deficit.

Vitamin A plays a number of roles in the organism, the most important being:

1) Growth, development, and differentiation of epithelial tissue. In these processes, retinoic acid is the active compound. Retinol only represents the precursor of retinoic acid. The mechanisms of action in this area are extremely complex and not yet fully elucidated.
2) Reproduction. Vitamin A is essential for the reproduction of higher animals. In men and male animals, it is required in the form of retinoic acid for testosterone synthesis, and as retinol for the maintenance of spermatogenesis [173]. In women and female animals, retinol is substantially involved in the healthy development of the placenta and the fetus.
3) Vision. In the retina, retinol (**1**) is dehydrogenated and isomerized to (11*Z*)-retinal (**19**) which condenses with the lysine residue of the protein opsin to give the aldimine, forming the light receptor rhodopsin. The action of light causes conversion of the chromophore from the 11*Z* to the all-*E* form (bathorhodopsin, lumirhodopsin). After a change in the conformation of the protein part (metarhodopsin), retinal (**2**) is released hydrolytically. This affects ion transport and membrane potentials, resulting in the generation of an optic nerve impulse. Vitamin A deficiency causes night blindness (nyctalopia).

In the purple membrane of *Halobacterium halobium*, (13*Z*)-retinal is bound to a protein as a Schiff's base (bacteriorhodopsin) [174]. On illumination, light energy can be converted to chemical energy ("proton pump", ATP synthesis) with the isomerization of the chromophore to the all-*E* form. Furthermore, vitamin A plays a role in the important metabolic step of glycoprotein synthesis [175] and is involved in maintenance of the immune system [176, 177].

2.8.2. β-Carotene

β-Carotene is resorbed from plant food in the small intestine to an extent of 30 – 60 %, depending on the added fat content [178]. In intestinal cells, part of the β-carotene is homolytically cleaved to retinal by the enzyme 15,15′-carotene dioxygenase. Retinal is, in turn, reversibly reduced to retinol by retinal reductase and, to a smaller extent, irreversibly oxidized to retinoic acid [179].

Apart from homolytic cleavage, heterolytic cleavage of β-carotene also occurs, producing various apo-carotinals. These substances are also converted to retinal, depending on the chain length.

That part of β-carotene that is not metabolized in the intestinal mucosa is transported to the target organs in lipoproteins, especially LDL,

and in erythrocytes and leukocytes. Here, further conversion occurs to retinol and retinoic acid via retinal or apo-carotinals. From the standpoint of the physiology of nutrition, it is assumed that the conversion of β-carotene to vitamin A is ca. 15 % [180]. The conversion decreases with increasing amount of β-carotene absorbed [181] and depends on the individual vitamin A status [182]. The activity of 15,15′-carotene dioxygenase appears to decrease with an increasing supply of vitamin A. Consequently, it is controlled by the supply of vitamin A. For this reason, hypervitaminosis A cannot result from β-carotene.

The β-carotene that is not metabolized to vitamin A can be resorbed unchanged by humans and has its own biological functions, independent of its provitamin A properties.

Like the other carotenoids, β-carotene is regarded as a natural physiological antioxidant. As a result of its ability to deactivate singlet oxygen [181] and to interrupt oxidative radical chain reactions (e.g., lipid peroxidations in cell membranes) [182], β-carotene has a protective function in cells and, therefore, has a preventive effect on a series of radical-induced degenerative diseases.

The following applications are known or are being discussed:

1) Protection of the skin from light and treatment of solar dermatosis and pigment anomalies [183, 184]
2) Prevention of various types of cancer, especially lung cancer [185]
3) Prevention of atherosclerosis [186]
4) Boosting of the immune functions [187]
5) Prevention of certain eye diseases such as cataract and macula degeneration [187]

2.9. Deficiency Symptoms and Requirements

Deficiency Symptoms. Vitamin A deficiency causes numerous symptoms such as disturbances of vision (night blindness), damage to the skin and mucous membranes, atrophies, and growth retardation.

Even if the exogenous supply of vitamin A is adequate, deficiency symptoms can arise due to a deficit of carotene 15,15′-dioxygenase and carrier protein [188], caused by insufficient protein supply (hypovitaminosis A syndrome in undernourished persons).

Requirements. In modern animal husbandry, green fodder is being increasingly replaced by ready-to-use feed. Since this industrially produced feed contains little provitamin A, it is enriched with vitamin A (especially retinyl acetate).

The vitamin A requirement for the maintenance and performance (growth, eggs, milk) of animals is largely dependent on the body weight and is in the range of 50 I.U. per kilogram of body weight per day. For optimal utilization of fodder (fattening) and optimal milking or laying capacity, daily doses of 100 – 200 I.U. per kg of body weight are recommended for working animals. Independent of its importance as provitamin A, β-carotene has a specific influence on the fertility of cattle (progesterone synthesis) and other animals [189].

German recommendations for the vitamin A supply of working and domestic animals [in I.U. retinol per kg of dry feed or per animal (∗) per day] are as follows:

Piglets	10 000 – 16 000
Porkers	5 000 – 10 000
Milk cows ∗	80 000 – 120 000
Horses ∗	30 000 – 60 000
Chicks, feedercocks	8 000 – 12 000
Layers, turkeys	8 000 – 12 000
Dogs	8 000 – 12 000
Cats	15 000 – 25 000
Fish	8 000 – 12 000
Beef cows ∗	40 000 – 70 000

For humans, the daily doses recommended by the German Society for Nutrition (DGE) are shown in Table 10.

2.10. Analysis and Standardization

Vitamin A. An unequivocal characterization of vitamin A can be achieved by its UV spectra, by color reactions, and from its biological activity. For purity testing, chromatographic methods (SFC, HPLC) have become the methods of choice and are essential for quantification and identification. Spectrophotometric determination is meaningful only at relatively high vitamin A concentrations [190] because even small

amounts of interfering impurities result in considerable errors. In this method, the UV absorption is measured and has a maximum between 325 and 328 nm depending on the solvent.

In colorimetric determination, the intensity of the blue color is measured at 620 nm after the addition of antimony(III) or antimony(V) chloride in trichloromethane (Carr – Price reaction). *Cis*, *trans* isomers give equally intense colors. Other characteristic color reactions with vitamin A are exhibited by concentrated sulfuric acid, tin chloride, arsenic trichloride, and boron trifluoride. The mechanism of these reactions has been described [191].

Although the colorimetric method is more sensitive than the spectrophotometric method, it is somewhat less accurate since the colors fade relatively quickly. Exact determination of vitamin A is possible by fluorometry, a very sensitive and specific detection method [192, 193]. However, carotenoids and tocopherol [194] interfere with the analysis. The fluorimetric test is more effective in the presence of holo-retinol binding protein (holo-RBP) [195] because retinol bound to apo-RBP (i.e., holo-RBP) fluoresces 14 times more intensively than free retinol dissolved in a hydrocarbon. Holo-RBP can easily be separated electrophoretically in polyacrylamide gel and purified. Fluorimetric determination can be carried out directly on the electrogram.

Table 10. Vitamin A requirement of humans in retinol equivalents

Age	Retinol, mg RE ∗		
	m	m/f	f
0 – 4 months		0.5	
4 – 12 months		0.6	
1 – 4 years		0.6	
4 – 7 years		0.7	
7 – 10 years		0.8	
10 – 13 years	0.9		0.9
13 – 15 years	1.1		1.0
15 – 19 years	1.1		0.9
19 – 25 years	1.0		0.8
25 – 51 years	1.0		0.8
51 – 65 years	1.0		0.8
> 65 years	1.0		0.8
Pregnant women			1.1
Nursing women			1.8

∗ 1 mg Retinol Equivalent = 6 mg
all-*trans*-β-carotene = 12 mg other provitamin A carotenoids.
mg Retinol Equivalent is the official unit of vitamin A activity.
1 mg RE corresponds to 3333 I.U.

Biological methods [196] are based on the curative and prophylactic growth test in the rat, the colpokeratosis test in the rat and mouse, the antiestrus test, the liver storage test in the rat, and the curative xerophthalmia test. These tests are laborious and are mainly used to check the accuracy of new methods for the determination of active substances and to determine the total concentration of active vitamin A compounds.

Today, the separation of retinoids of the same polarity (e.g., *cis*, *trans* isomers) is exclusively conducted with high-pressure liquid chromatography (HPLC). Detection systems range from simple UV detection to electrochemical processes [197] and laser photoionization [198, 199].

Provitamin A. For the determination of provitamins A, spectrophotometric measurement of the absorption maximum is the most important method. If necessary, the substances are first purified on basic or neutral alumina, calcium oxide, magnesium oxide, or silica gel. Separation of structural isomers (*cis*, *trans*) is achieved by using HPLC [200].

2.11. Trade Names and Economic Aspects

Retinol (**1**) is sold almost exclusively in the form of the more stable esters (mainly as retinyl acetate, also as retinyl palmitate and propionate). About 80 % of the production is used in animal feed, the remainder in the food and pharmaceutical sectors.

For practical reasons, various preparations are made and mostly stabilized with antioxidants: solutions in fats and oils (food sector, especially margarine); liquid, water-miscible dispersions containing emulsifiers (pharmaceutical sector, animal feed); and, above all, dry powders which contain the agent embedded in a matrix of gelatin and/or polysaccharides (especially animal feed, and specialities in the food and pharmaceutical sectors).

At present, world production is ca. 2700 t/a. The producers are Hoffmann LaRoche (Switzerland and United States), BASF (Germany) and Rhône-Poulenc (France) as well as smaller producers in India, China, and the CIS. The market price is 200 – 250 DM/kg (calculated on the basis of 100 % active agent) depending on the preparation.

Retinoic acid (for acne) and (13*Z*)-retinoic acid (for psoriasis) have attained importance in dermatology.

The economic importance of the carotenoids, especially of β-carotene, has increased significantly in the past years and is still increasing.

In the form of oily dispersions, β-carotene is used to color margarine and other high-fat foods. As an active agent (see Section 2.8.2.) or as provitamin A, these dispersions are being increasingly filled into soft gelatin capsules.

Water-dispersible dry powders are used to color beverages, pudding powders, and many other foods. They are also employed as provitamin A in vitamin drinks or tablets.

In the animal feed sector, other carotenoids, e.g., β-apo-8′-carotinoic acid ethyl ester or canthaxanthin, are used as a dry powder to pigment the skin and egg yolk of poultry. Carotenoids identical to the naturally occurring substances are being increasingly used to pigment salmon and trout in fish farming.

About 400 t/a of carotenoids are synthesized (calculated as pure compounds). The price is between 800 and 2000 DM/kg (calculated on the basis of 100 % purity). Producers are Hoffmann LaRoche and BASF.

2.12. Tolerance

Excessive feeding of vitamin A compounds (single doses of ca. 2×10^6 I.U. or repeated administration of 0.2×10^6 I.U. and more to adults) causes severe damage (hypervitaminosis A). Typical signs in experimental animals are skeletal fractures and bleeding. The molecular cause for these symptoms is an increase in the membrane permeability in lysosomes with the release of proteolytic enzymes (cathepsin D) [201]. This is produced by vitamin A that is not bound to a carrier protein.

High doses of provitamin A compounds do not cause hypervitaminosis A.

3. Vitamin D

3.1. Introduction

The group of vitamins D (calciferols) are compounds which play a central role in the maintenance (homeostasis) of calcium and phosphorus concentration in the body fluids of vertebrates, including humans, by controlling intestinal calcium absorption and calcium mobilization of the bones. According to present knowledge, the vitamins D are biologically inert, and before they can fulfil their physiological role they must be converted by two subsequent enzymatic hydroxylations in the liver and the kidney to the biologically active vitamin D derivative.

A selection of reviews and books is given in general references [202–210]. An excellent overview of major developments in particular vitamin D research frontiers is offered by the *Proceedings of Vitamin D Workshops* which are organized every three years [211–213].

3.2. History

Rickets, a disease of sucklings and children, is a serious disorder in bone development due to a disturbed calcium and phosphorus metabolism as a consequence of vitamin D deficiency. Rickets became a serious health problem with industrialization in geographical areas of low-incident sunlight in northern Europe and North America, as well as with the accompanying advances in urbanization of society. A relation between exposure to sunlight and its preventive action towards rickets was already recognized in the early 1800s (Sniadecki). In 1919 Melanby succeeded in producing rickets experimentally in dogs and cured the disease by adding cod liver oil, a known popular remedy for the prevention of rickets, to their food. Shortly afterwards, Huldshinski demonstrated that exposure of rachitic children to sunlight or artificial UV light also cures the disease. Food and vegetable oils exposed to UV light exhibited the same therapeutic effect as cod liver oil. The component of the vegetable oil activated by UV light was identified as a chromophoric trace impurity accompanying the plant sterols in the nonsaponifiable fraction and inseparable from the parent sterol even by repeated crystallization. Analogous impurity was also detected in the crude animal sterol cholesterol. The absorption spectra of the impurities mentioned, resembled that of ergosterol, an already known

$\Delta^{5,7}$-sterol isolated from yeast. Indeed, ergosterol when irradiated yielded a mixture of products possessing high antirachitic activity, from which two independent teams (WEBSTER, WINDAUS) succeeded, in the early 1930s, in isolating the active principle: first in impure (vitamin D_1 or calciferol) and later in pure crystalline form, named it vitamin D_2 (ergocalciferol), and elucidated its structure by means of classical chemical degradation reactions (WINDAUS, HEILBRON). Vitamin D_3 (cholecalciferol), the active principle isolated from cod liver oil (BROCKMANN) was prepared by irradiation of 7-dehydrocholesterol (cholesta-5,7-diene-3β-ol), a synthetic analogue of ergosterol (and accompanying impurity of crude cholesterol), and its structure (WINDAUS, HEILBRON) was verified by total synthesis (INHOFFEN). Further vitamins D, products of irradiation of 7-dehydro derivatives of 22,23-dihydroergosterol (D_4), β-sitosterol (D_5), stigmasterol (D_6) and campesterol (D_7), were reported, the three last named possessing only low activity. The complex photochemistry was elucidated by HAVINGA and VELLUZ who recognized previtamin D as the key product of photolysis of steroidal 5,7-dienes (provitamins D) and as the intermediate of all subsequent transformations. HAVINGA corrected further the erroneous assumption of WINDAUS, who took vitamin D for the end product of the photolytic sequence, and identified it as a product of a thermal isomerization.

The years 1966 – 1971 brought a dramatic revival to the field of vitamin D due to the pioneer investigations of the three teams of DELUCA, KODICEK, and NORMAN. By using synthetic, labelled vitamin D_3 the named authors proved that in warmblooded animals it is converted by stepwise enzymatic hydroxylations and that one of the resulting metabolites, calcitriol (1α,25-dihydroxycholecalciferol) is the true highly active principle which performs all the functions ascribed previously to vitamin D_3. Calcitriol and other metabolites were isolated by means of advanced experimental techniques, and their structure was elucidated from their chemical and spectrometric properties and verified by synthesis. As under normal physiological conditions, humans and animals produce sufficient amounts of vitamin D_3 in the skin, the latter should not, strictly speaking, be classified as a vitamin, in contrast to vitamin D_2 which is available only from the diet. (A vitamin is a dietary component essential for health which the body is not able to synthesize.) However, the use of the designation vitamin D has been preserved as a general term for compounds exhibiting qualitatively the biological activity of calciferols and for terms such as vitamin D activity, vitamin D deficiency, etc. The name vitamin D_3 or vitamin D_2 is still used as a synonym for cholecalciferol or ergocalciferol, respectively. The active metabolite calcitriol, in view of its synthesis in the organism and its function, must, *per definitionem*, be classified as a hormone, and its precursor vitamin D_3 as a prohormone.

3.3. Structure and Nomenclature

The vitamins D belong chemically to the steroids [214] in which the B-ring of the steroid nucleus is cleaved between carbon atoms 9 and 10 (9,10-secosteroids). X-Ray analysis and NMR data [215–217] indicate that vitamin D_2 and D_3 and their derivatives exist in the crystal and predominantly in solution not as the cisoid, steroid-like form (**1a**, **1b**) – though according to computer calculations, this is the lowest energy conformation [218] – but as the coplanar conformer **2** (Fig. 1). (In **2** the A-ring is rotated through ca. 180° around the $C_6 - C_7$ linkage, giving a transoid diene system extending from C_5 to C_8.) In addition, **2** is an approximately equimolar mixture of two A-ring chair conformations [219, 220].

To avoid a certain confusion in the steroid nomenclature arising with the transoid conformation **2** for which the descriptors α and β do not apply, IUPAC [221] has recommended designating the chiral centres in the A-ring consequently as *R* and *S*. Nevertheless, systematic steroid nomenclature, as well as that of Chemical Abstracts, still use the traditional steroid notation derived from the cisoid form **1**. Thus vitamin D_3 (cholecalciferol, calciol) in IUPAC notation is (3*S*,5*Z*,7*E*)-9, 10-seco-5,7,10(19)-cholestatriene-3-ol (**2a**), and in systematic nomenclature (5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-3β-ol. Analogously, vitamin D_2 (ergocalciferol, ercalciol) (**2b**), which differs from vitamin D_3 in the side chain, is (3*S*,5*Z*,7*E*,22*E*)-9,10-seco-5,7,10(19),22-ergostatetraene-3-ol or

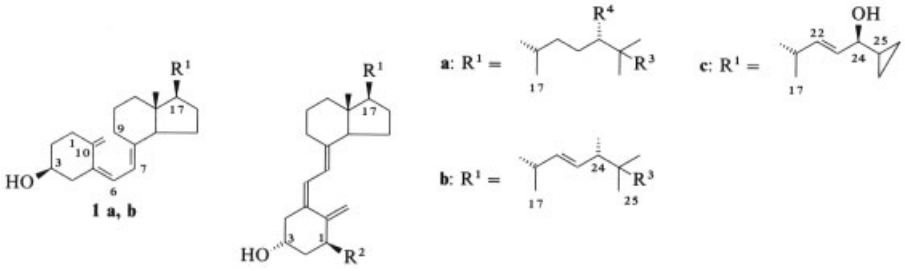

Figure 1. Structure of vitamin D_2 and D_3 and derivatives

(5*Z*,7*E*,22*E*)-9,10-secoergosta-5,7,10(19),22-tetraene-3β-ol. The abbreviated form of representation of the hydroxylated vitamin D metabolites such as 25-OH-D_3 for 25-hydroxycholecalciferol (**3a**) or 1,25-$(OH)_2$-D_3 for calcitriol (**4a**), though discouraged [221], is generally used throughout the literature. Subscript ciphers attached to trivial names (e.g., tachysterol$_3$ or lumisterol$_3$) correspond to that of the particular vitamin D.

3.4. Chemical and Physical Properties

The vitamins D, their metabolites, and analogues containing the characteristic 5,7,10(19)-triene system are highly unstable towards light, atmospheric oxygen, and acids; they are stable towards alkalies if protected from oxygen and light. Even in the crystalline state the vitamins D contain a certain amount of isomeric previtamin D with which they are in a temperature-dependent equilibrium. Preparations should therefore be kept protected from light and at low temperatures (+4°C or less) under inert gas. Heating to 180 °C leads to irreversible B-ring closure yielding pyro- and isopyrocalciferol (9α, 10α- or 9β,10β-cholesta-5,7-diene-3β-ol, respectively). Contrary to the provitamins (7-dehydrosterols), the vitamins D fail to give a precipitate with digitonin in ethanolic solution.

Vitamin D_3 (cholecalciferol, calciol) [*67-97-0*] (**2a**), $C_{27}H_{44}O$, M_r 384.6, colorless crystals, *mp* 84 – 85 °C (acetone), $[\alpha]_D$ +84° (acetone), UV (ethanol): λ_{max} 265 nm (ε 18 200) [222], NMR: [223], MS: [224], ^{13}C NMR: [223, 225].

Vitamin D_2 (ergocalciferol, ercalciol) [*50-14-6*] (**2b**), $C_{28}H_{44}O$, M_r 396.6, colorless prisms, *mp* 114.5 – 117 °C (acetone), $[\alpha]_D$ +102.5° (ethanol); UV (ethanol): λ_{max} 265 nm (ε 19 200) [222], NMR: [226], MS [227].

Previtamin D_3 [(6*Z*)-tacalciol], (3*S*,5*Z*,7*E*)-9,10-secocholesta-5 (10),6,8-triene-3-ol, (6*Z*)-9, 10-seco-5 (10),6,8-cholestatriene-3β-ol [*1173-13-3*] (**12a**), $C_{27}H_{44}O$, M_r 384.6, oily, UV (ethanol): λ_{max} 262 nm (ε 9000).

25-Hydroxycholecalciferol monohydrate (calcidiol, calcifediol), (3*S*,5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-3,25-diol, (5*Z*,7*E*)-9,10-seco-5,7,10(19)-cholestatriene-3β,25-diol [*19356-17-3*] (**3a**), $C_{27}H_{44}O_2 \cdot H_2O$, M_r 418.6, *mp* 81 – 83 °C (aq. acetone), $[\alpha]_D$ +77° (trichloromethane), UV (ethanol): λ_{max} 265 nm (ε 18 000), NMR: [228, 229], MS: [229], ^{13}C NMR: [225].

1α,25-Dihydroxycholecalciferol (calcitriol), (1*S*,3*R*,5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-1,3,25-triol, (5*Z*,7*E*)-9,10-seco-5,7,10(19)-cholestatriene-1α,3β,25-triol

[*32222-06-3*] (**4a**), $C_{27}H_{44}O_3$, M_r 416.6, white solid, *mp* 115 – 116 °C (methyl formate) [230], $[\alpha]_D$ +45° (ethanol), UV (ethanol): λ_{max} 265 nm (ε 16 600), IR: [231], NMR: [232], MS: [232], ^{13}C NMR: [225].

24*R*,25-Dihydroxycholecalciferol (secalciferol), (3*S*,24*R*,5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-3β,24,25-triol, (24*R*,5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-3β-24,25-triol [*55721-11-4*] (**5a**), $C_{27}H_{44}O_3$, M_r 416.6, white crystals, *mp* 136 – 137 °C (methyl formate), $[\alpha]_D$ +113° (ethanol), UV (ethanol): λ_{max} 265 nm, NMR: [233], MS: [234].

1α-Hydroxycholecalciferol (alfacalcidiol) (1*S*, 3*R*,5*Z*,7*E*)-9,10-secocholesta-5,7,10(19)-triene-1,3-diol, (5*Z*,7*E*)-9,10-seco-5,7,10(19)-cholestatriene-1α,3β-diol [*41294-56-8*] (**6a**), $C_{27}H_{44}O_2$, M_r 400.6, white crystals, *mp* 138 – 139.5 °C (methyl formate), $[\alpha]_D$ +29° (dioxane), UV (ethanol): λ_{max} 264 nm (ε 17 600), IR: [235], NMR: [235], MS: [235].

Calcipotriol (1*S*,3*R*,24*S*,5*Z*,7*E*,22*E*)-24-cyclopropyl-9,10-seco-5,7,10(19),22-cholatetraene-1,3,24-triol, (24*S*,5*Z*,7*E*,22*E*)-24-cyclopropyl-9,10-seco-5,7,10(19),22-cholatetraene-1α,3β,24-triol [*112965-21-6*] (**4c**), $C_{27}H_{40}O_3$, M_r 412.6, crystals, *mp* 166 – 168 °C (methyl formate), UV (ethanol): λ_{max} 264 nm (ε 17 200), NMR: [236], MS: [236].

3.5. Biosynthesis and Occurrence

3.5.1. Biosynthesis

The natural source of vitamin D_3 in warm-blooded animals and humans is the transformation of provitamin D_3 (7-dehydrocholesterol), a product of cholesterol biosynthesis, in the skin [210] when exposed to the ultraviolet component of sunlight (λ= 290 – 315 nm, UV-B radiation). The light penetrates into the epidermis and converts provitamin D_3 photochemically to previtamin D_3, and this isomerizes at body temperature to vitamin D_3. The photolysis is very efficient [206], though liable to seasonal and climatic variations. Pigmentation and aging significantly decrease the capacity of human skin to produce vitamin D_3.

The vitamin D_3 produced either endogenously or obtained dietarily enters the lymphatic system bound to the specific vitamin D binding protein (DBP) and is transferred first to the liver and hydroxylated there at C_{25} to the biologically inactive 25-hydroxycholecalciferol (calcifediol, 25-OH-D_3) (**3a**), the DBP-bound, major circulating form of vitamin D_3, and then transferred to the kidney, the site of further hydroxylation. Renal hydroxylation is strictly controlled by serum calcium level: a decrease below the physiological concentration (86 – 105 µg/mL) stimulates production of parathyroid hormone (parathormone, PTH) [237] in the parathyroid gland, which activates renal 25-OH-D_3-1α-hydroxylase. The enzyme converts 25-OH-D_3 into the 1α-hydroxylated highly active hormone calcitriol (**4a**), which carries out all the known functions ascribed to the vitamin D_3. The resulting normalization of calcium in serum suppresses PTH production, and calcitriol deactivates the 1α-hydroxylase by feed-back control. The deactivation is associated with the activation of renal 25-OH-D_3-24-hydroxylase, which starts to convert 25-OH-D_3 into 24*R*,25-dihydroxycholecalciferol [24*R*,25-$(OH)_2$-D_3] (**5a**). The physiological role of this metabolite, if any, is still the subject of controversial reports. Both metabolites are subjected to a complex catabolic process which proceeds via additional hydroxylations up to the final cleavage of the side chain [203, 206, 238]. Should calcium concentration rise above normal, the thyroid gland starts to secrete the hormone calcitonin (CT) [239], a counterregulator of PTH, which blocks mobilization of calcium from the bone and stimulates excretion of calcium in the kidney. Vitamin D_2 (**2b**), the source of which is the diet, is metabolized in mammals and humans analogously to vitamin D_3, to give 25-OH-D_2 (**3b**) and 1α, 25-$(OH)_2$-D_2 (**4b**).

3.5.2. Occurrence

Fish oils contain considerable amounts of vitamin D_3, accompanied by small amounts of vitamin D_2. Lesser concentrations of vitamin D_3 are found in egg yolk, mammalian and human breast milk (as sulfate), invertebrates, and mushrooms. Natural sources of vitamin D are summarized in [240]. Generally, vegetables and fruits

do not contain vitamin D. The glycoside of calcitriol was detected in the leaves of calcinogenic plants, e.g., *Solanum malacoxylon* [241]. The cutaneous photosynthesis of vitamin D_3 in fish is questionable because of the low transmittance of water for UV light. The mode of accumulation of large hepatic stores of vitamin D_3, the physiological significance of which is poorly understood, is most probably a result of concentration through food chains starting from zoo- and phytoplankton. Hypothetical, nonphotochemical enzymic mechanisms of vitamin D_3 synthesis in fish and some terrestrial subterranean or nocturnal herbivores have been discussed [242].

3.6. Vitamin D Requirements

The daily vitamin D requirements in human adults probably do not exceed 100 – 200 I.U. (1 mg = 40 000 I.U.); this can readily be obtained from the dietary sources (eggs, cheese, liver, oily fish) and exposure to sunlight. For children up to the age of six, the recommended allowance is 400 I.U./d; this dose may represent the upper limit of vitamin D required [203, 243]. Intake of vitamin D in excess of 1000 I.U. per day is potentially dangerous to normal children and adults and must be avoided. Vitamin D intoxication is believed to be the result of high circulating 25-OH-D_3 levels that, mimicking 1,25-$(OH)_2$-D_3 on the receptor, lead to dangerous hypercalcemia, which may result in irreversible soft tissue calcifications (kidney, lungs, heart, pancreas, aorta). Food is supplemented in many countries by vitamin D_3 or D_2 (milk, margarine); the daily portion of a certain food contains usually one third or one half of the total vitamin requirement and the vitamin amount in a definite portion of food is declared on the package. The vitaminized food products are subject to control by state institutions. In veterinary nutrition, food fortification with vitamin D is important for the normal development of young animals and for husbandry of cattle, swine, and poultry. The vitamin D requirements are given in I.U. either per animal or kilogram body weight and day or per kilogram dry food per day [244, 245]. It has, however, to be emphasized that the data related to requirement levels depend on a number of different factors and should be considered as guideline recommendations only and that generally valid data are not obtainable.

3.7. Pharmacological Effects and Uses

Calcitriol is a hormone which functions at the level of the target cell nucleus in a fashion analogous to that of steroid hormones. The actions of calcitriol are mediated by a specific 1,25-$(OH)_2$-D_3 receptor molecule which belongs to the superfamily of protein receptors for steroid hormones [246, 247]. The calcitriol – receptor complex is transported to the nucleus and stimulates the transcription of the gene coding for proteins that modify the functions of the target cell. Calcitriol exerts its main task, the control of calcium metabolism [248], on the classic target tissues, i.e., intenstine, bone, and kidney. As, however, specific 1,25-OH-D_3 receptor molecules have been identified in a series of other tissues as well, this indicates far broader calcitriol activities than formerly assumed. Some biological responses of calcitriol are generated by nongenomic pathways and two different forms of the receptors were postulated for genomic and nongenomic mechanisms [249]. The steroid-like cisoid conformation **1** is suggested to be responsible for the nongenomic function [249].

3.7.1. Calcium Homeostasis and Disorders of Calcium Metabolism

The most obvious effect of calcitriol is normal bone development by maintenance of plasma calcium and phosphorus concentrations at normal levels. This mechanism is carried out by activation of intestinal calcium and phosphorus absorption, improvement of renal calcium reabsorption, and stimulation of calcium mobilization from the bone. The most important indication in human medicine in this field is the prevention and healing of vitamin D deficiency in growing children (rickets) and in adults (osteomalacia). Patients suffering from renal failure and anephric patients on dialysis produce either insufficient amounts of calcitriol or fail to produce it at all, and as a result exhibit low intestinal calcium absorption, low plasma calcium concentration (hypocalcemia), and secondary hyperparathyroidism (overproduction of

PTH), the latter being the major factor in the development of renal osteodystrophy. Oral doses of calcitriol or its synthetic analogue 1α-hydroxycholecalciferol (**6a**) – 1α-OH-D_3, before it acts, is hydroxylated at C_{25} in the liver – normalize the hypocalcemia and the PTH concentration in the serum. The therapy requires medical surveillance and thorough control of serum calcium concentrations. Postmenopausal and age-related (senile) osteoporosis, the most frequent metabolic bone disease seen in medical practice, is a failure of calcium absorption which results in a shift from dietary sources to skeletal calcium mobilization accompanied by a continuous loss of bone substance. Clinical trials using calcitriol for treatment of postmenopausal women have been reported [250,251] but the application of calcitriol for treatment of senile osteoporosis is still the subject of controversy. Calcitriol has been reported to support increased egg production and egg shell breaking strength, and recommended for prevention and treatment of parturient paresis (milk fever), a hypocalcemic disorder of dairy cows. Substantial economic losses in farming have been caused by severe calcinoses of grazing animals after intake of leaves of calcinogenic plants containg calcitriol in glycosidic form [241]. Vitamin D_2 and 1α,25-$(OH)_2$-D_2 are approximately equally as active in humans and mammals as the corresponding vitamin D_3 derivatives but are substantially less active in avian species; the discrimination is due to the vitamin D transport protein, which has low affinity towards vitamin D_2 derivatives [252].

3.7.2. Further Activities

Nuclear 1,25-$(OH)_2$-D_3 receptors have been demonstrated in tissues not primarily related to mineral homeostasis, such as the skin, parathyroid cells, pancreas, heart muscle, placenta, and others. Further, a substantial fraction of all malignant cells contain significant amounts of the receptor. Calcitriol has been shown to influence the proliferation and differentiation of several tissues and may also have immunoregulatory properties [253]. Calcitriol-receptor complex suppresses the gene of the PTH in the parathyroid gland [254], stimulates insulin secretion in the pancreas [255,256], inhibits leukemic cell growth and is an effective antiproliferative agent with potential use in cancer therapy [257]. Similar effects have been demonstrated on human epidermal cells, where calcitriol inhibits proliferation and induces terminal differentiation of keratinocytes; this has been used for treatment of psoriasis (hyperproliferation epidermal disorder). The antiproliferative and differentiation inducing effects of calcitriol are dose dependent and require supraphysiologic doses. The obvious clinical limitation, the potentially serious hypercalcemic side effect of doses higher than a few micrograms per day, questions a long-term therapeutic use and calls for analogues having diminished calcemic activity. Analogues have been synthesized that possess high proliferation/differentiation activity combined with a significantly lower calcemic effect [236,258] and the structure – function relationship has been discussed [259]. Two synthetic derivatives, calcipotriol (**4c**) and tacalcitol [(5*Z*,7*E*,24*R*)-9,10-secocholesta-5,7,10(19)-triene-1α,3β,24-triol], have been introduced to the market for topical treatment of psoriasis.

3.8. Synthesis

3.8.1. Synthesis of Vitamin D_3 and D_2

On an industrial scale, the vitamins D_3 are manufactured by chemical synthesis only. The fish oils, the natural source of vitamin D_3, are only used for preparation of vitamin concentrates. Starting materials for vitamins D are 7-dehydrosterols ($\Delta^{5,7}$-sterols). Vitamin D_2 (**2b**) is prepared from the natural substance ergosterol (**11b**); its synthetic analogue, 7-dehydrocholesterol (**11a**) is the starting material for vitamin D_3 (**2a**).

The synthesis of 7-dehydrocholesterol (Fig. 2) starts from commercially available cholesterol (**7a**), the hydroxy group of which is protected as acetate or benzoate (**8a**). Allylic bromination to the 7-bromo derivative (**9a**) and dehydrobromination to the $\Delta^{5,7}$ ester (**10a**) followed by saponification gives the free 7-dehydrocholesterol (**11a**).

The photolysis of the 5,7-diene system of either 7-dehydrocholesterol (**11a**) or ergosterol (**11b**) (for details see → Photochemistry, Chap. 5.2.) is carried out by irradiation with ultraviolet light of a mercury lamp in an inert sol-

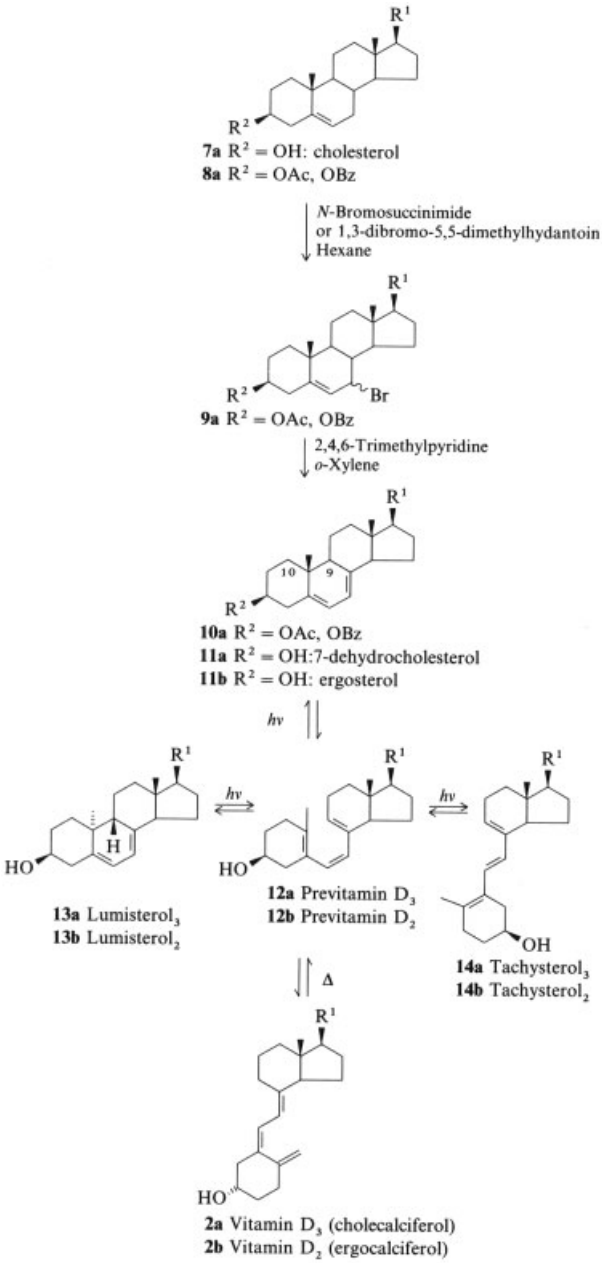

Figure 2. Synthesis of vitamin D_3 and D_2
Ac = CH_3CO, Bz = C_6H_5CO
a, **b**: R^1 see Figure 1

vent [260] yielding previtamin D ((**12a**), (**12b**)) by reversible fission of the cyclohexadiene system in the B-ring between C_9 and C_{10}. The previtamin D is in an equilibrium with its ring reclosure photoproducts – the starting diene ((**11a**), (**11b**)) and its $9\beta,10\alpha$-isomer lumisterol ((**13a**), (**13b**)) – and with the product of *E*/*Z* photoisomerization of the 6,7 double bond, tachysterol ((**14a**), (**14b**)). The ratio of products and the predominating component in the reaction

mixture depends on the wavelength of the UV light used [261]. To avoid the slow, irreversible formation of over-irradiation products, photolysis is discontinued after max. 40 % turnover of the starting diene. The subsequent step, the thermal isomerization of previtamin D into vitamin D (**12a**, **12b** → **2a**, **2b**), is a reversible sigmatropic 1,7-hydrogen shift from C_{19} to C_9, performed by heating the solution of the photoproducts at 80 °C to a vitamin – previtamin equilibrium (ratio ca. 4 : 1). The reaction solution is then concentrated and treated with methanol to recover the sparingly soluble starting $\Delta^{5,7}$-diene. Tachysterol is removed as a Diels – Alder adduct with maleic anhydride. The resulting solution is evaporated to give crude vitamin D (vitamin D resin) which either is used directly for veterinary purposes or purified to crystalline vitamin D_3 or D_2 via an ester that crystallizes well [262] (3,5-dinitrobenzoate or butanoate), subsequent saponification, and final crystallization.

The introduction of triplet sensitizers that influence the *E*/*Z* isomerization of the 6,7 double bond by reconverting the undesirable side product tachysterol to previtamin (**14** →**12**), brought decisive progress in the photochemical transformation of 7-dehydrosterols and a very efficient optimization of the previtamin D yield. In practice, the 5,7-diene is irradiated as above, then the sensitizer is added and irradiation is continued to transform the tachysterol almost quantitatively to previtamin (double irradiation technique) [263].

3.8.2. Synthesis of Calcitriol

The classical pathway starts with the synthesis of 25-hydroxycholesterol (**15**) from an appropriate steroid derivative by degradation and resynthesis of the side-chain [238, 264, 265]. The method of choice for introduction of the 1α-hydroxy group in the A-ring of 25-hydroxy-cholesterol (**15**) is the three-step procedure of Barton (**15** → **16** → **17** → **20**) [266] which in a modified form [267] was applied for the industrial production of 1α,25-dihydroxycholesterol (**20**) (Fig. 3). The chemical introduction of the 1α-hydroxyl group was replaced later by an economic and for industrial purpose adapted [268] fermentative one-step 1α-hydroxylation [269] using dehydroepiandrosterone (**18**), a commercially available compound prepared by microbiological degradation of cheap plant sterols, as starting material. The synthesis of the side chain in the 1α-androstane derivative **19** gives 1α,25-dihydroxycholesterol (**20**) [270], followed by the conversion to calcitriol analogous to the synthesis of vitamin D (Fig. 2) using the double-irradiation technique [271]. Calcitriol is isolated by preparative HPLC and crystallized. The strategy of both the above-mentioned synthetic pathways is based on multistep manipulations of intact steroid ring systems, and the final conversion of the resultingly expensive 1α,25-dihydroxycholesterol to the seco-triene system via the 5,7-diene is realized by a sequence of reactions notorious for their moderate yields. A conceptually new alternative is the synthesis starting from the commercially available bulk product ergocalciferol (vitamin D_2) (**2b**) [228]. The pathway (Fig. 4) involves protection of the triene system as sulfur dioxide adduct **21**, cleavage and resynthesis of the side chain and chemical introduction of the 1α-hydroxy group (**23** → **29**) [228]. An alternative synthesis of calcitriol [230, 272] and calcipotriol [236] is based on a similar strategy. A promising approach, the direct fermentative 1α- and 25-hydroxylation of vitamin D_3 to calcitriol [273] has been reported. The total synthesis of calcitriol or other vitamin D metabolites and analogues [264, 274] has not yet been realized on an industrial scale.

3.9. Assays for Vitamin D and Metabolites

Assays for vitamin D and metabolites have been reviewed [275]. The official method for determination of vitamin D in feedstuffs, tissues, and body fluids in the years 1922 – 1958 was the rat line test, a bioassay based on curative and preventive effects in rachitic or normal animals, kept on a rachitogenic diet containing varying amounts of vitamin D. The measure of healing of the rickets induced was evaluated after silver-nitrate staining of the split tibias. The method is sensitive (25 µg vitamin D_2 or D_3) but time consuming (7 – 14 d assay period) and nonspecific. An alternative to the line test, a subsequently adopted chemical method, is based on spectrometric estimation of the red coloration of vitamin D with antimony trichloride at 500 nm. The

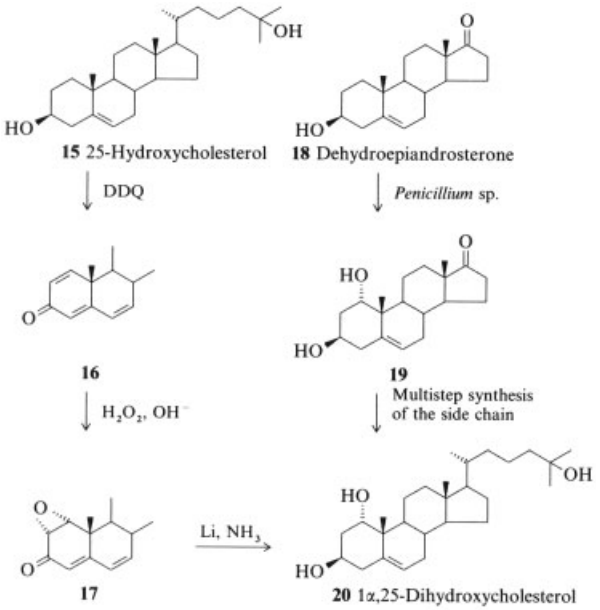

Figure 3. Synthesis of 1α,25-dihydroxycholesterol
DDQ = 2,3-Dichloro-5,6-dicyano-1,4-benzoquinone

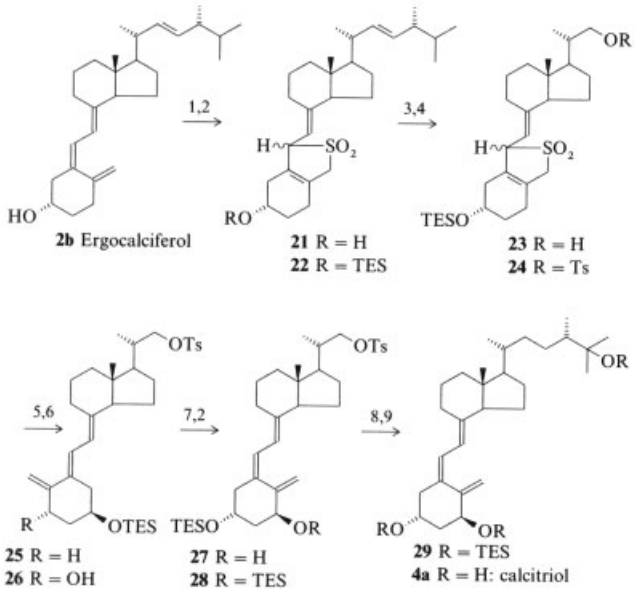

Figure 4. Synthesis of calcitriol from ergocalciferol
TES = Et_3Si, Ts = p-$CH_3C_6H_4SO_2$
1) SO_2 (liquid); 2) Et_3SiCl, imidazole, CH_2Cl_2; 3) $O_3/NaBH_4$, 4) TsCl, pyridine; 5) $NaHCO_3$, EtOH; 6) SeO_2, N-methylmorpholine N-oxide, CH_2Cl_2/MeOH; 7) $h\nu$, anthracene; 8) $BrMgCH_2CH_2C(CH_3)OTES$, Cu(I), THF; 9) Bu_4NF, THF

method was designed for estimations in pharmaceutical preparations containing larger quantities of vitamin D (> 250 µg) [276]. Official methods for vitamin D estimation are described in detail in [277].

Given the knowledge of vitamin Dmetabolism and the low concentrations of the metabolites in the plasma or serum (25-OH-D_3: 30 ng/mL, 1,25-$(OH)_2$-D_3: 30 – 40 pg/mL, 24,25-$(OH)_2$-D_3: 0.5 –2.0 ng/mL), two approaches were developed for the measurement of metabolites in plasma and tissues in vitro using ^{3}H-radiolabeled vitamin D derivatives: competitive protein-binding analysis (CPB) and radioimmunoassay (RIA). The CPB technique takes advantage of the high affinity for vitamin D metabolites of proteins occurring in serum (vitamin D-binding protein) or in intestinal target cells (cytoreceptors). For RIA, the vitamin D metabolites being themselves not antigenic, must be chemically modified to haptens which are then covalently linked to a protein molecule [mostly bovine serum albumin (BSA)]. The vitamin D derivative – BSA complex then constitutes the antigen which, injected into an animal, induces antibody formation. Both assays involve competition between a radiolabeled metabolite and the vitamin D derivative of interest for available sites on the binding protein. However, the low specificity of both the binding protein and the antibody and, furthermore, the presence of interfering substances in serum make prepurification and separation of the metabolites by extraction and chromatographic techniques indispensable. Significant advance in the isolation techniques has been the introduction of solid-phase extraction and "phase-switching" [278, 279]. The introduction of high performance liquid chromatography (HPLC) rendered separation and isolation of individual metabolites possible either for the assays mentioned above or for the direct measurement of HPLC peak areas using UV detection (for vitamin D and 25-OH-D only) or the highly sensitive fluorometric assay [280]. A combination of gas chromatography and mass spectrometry (GC – MS) [279] has also been introduced in this field. The accuracy and precision of differentassays for vitamin D and metabolites have been compared and discussed [281].

3.10. Trade Names and Economic Aspects

The pharmaceutical industry offers vitamin D_3, and to a lesser extent vitamin D_2, in pure crystalline form or as a solution in vegetable oil (1 000 000 I.U./g) for preparation of pharmaceuticals and for food supplementation. The majority of vitamin D_3 produced is the crude product, so-called vitamin D_3 resin, an intermediate for manufacturing dispersions in plant protein and dextrin matrix (about 500 000 I.U./g) used for fortification of animal feeds. Vitamin D_2 and D_3 are available in numerous drug products [282], in multivitamin specialities, and in fish oil concentrates. The main bulk producers of vitamin D are Solvay-Duphar (The Netherlands, D_3, D_2), Hoffmann-La Roche (Switzerland, D_3), BASF (Germany, D_3), Synthesia (Czech Republic, D_2). Annual worldwide vitamin D_3 production amounts to ca. 1.5×10^{15} I.U.; the production of calcitriol, the daily dose of which is 0.25 µg only, does not exceed several hundreds of grams. Trade names and the producers of synthetic metabolites and analogues [282] are summarized in Table 11. Commercially available tritiated derivatives for assays are available from Amersham (United Kingdom) and New England Nuclear Corp. (United States).

4. Vitamin E (Tocopherols, Tocotrienols)

4.1. Introduction

The eight naturally occurring substances with vitamin E activity are all derivatives of 6-chromanol. They belong to two groups of compounds. The first group (**1a**–**1d**) is derived from tocol (**1** with $R^1 = R^2 = R^3 = H$), a substance which has a saturated C_{16} isoprenoid side chain and has not yet been found in nature.

The second group (**2a**–**2d**) consists of derivatives of tocotrienol, which has a triply unsaturated C_{16} side chain (at position 3′, 7′, and 11′). The same α, β, γ, and δ structures are found as in the tocopherol group (**1a**–**1d**).

Table 11. Trade names and producers of synthetic vitamin D metabolites and analogues

Abbreviation	Generic Name	Trade Name	Producer
25-OH-D_3	calcifediol	Dedrogyl	Roussel-Uclaf (France)
		Calderol	Upjohn (United States)
		Hidroferol	Juventus (Spain)
25-OH-5E-D_3	5,6-*trans*-25-hydroxy-cholecalciferol	Delakmin	Roussel-Uclaf (France)
1α,25-$(OH)_2$-D_3	calcitriol	Rocaltrol	Hoffmann-La Roche (Switzerland, United States)
		Calcijex	Abbott (United States)
1α-OH-D_3	alfacalcidiol	Alfarol	Chugai (Japan)
		Dediol	Leo (Denmark)
		Eenalfadrie	Teva (Israel)
		Einsalpha	Thomae (Germany)
		Odinal	Hemofarm (Former Yugoslavia)
		OneAlpha	Leo (Denmark)
		Vetalpha	Crown Chem. (United Kingdom)
	calcipotriol	Daivonex	Leo (Denmark)
		Dovonex	Leo (United Kingdom)
1α,24R-$(OH)_2$-D_3	tacalcitol	Bonealfa Oint.	Teijin (Japan)
	dihydrotachysterol	A.T. 10	Bayer (Germany)
		Calcamin	Wander (Switzerland)
		DHT	Roxane (United States)
		Dihydral	Duphar (Belgium)
		Dygratyl	Ferrosan (Denmark)
		Hytakerol	Sterling (United States)
		Tachyrol	Duphar (Denmark)
		Tachystin	Ankerwerk (Germany)

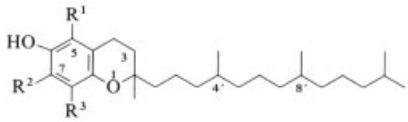

1a, α-Tocopherol: $R^1 = R^2 = R^3 = CH_3$
1b, β-Tocopherol [*148-03-8*]: $R^1 = R^3 = CH_3$, $R^2 = H$
1c, γ-Tocopherol [*54-28-4*]: $R^1 = H$, $R^2 = R^3 = CH_3$
1d, δ-Tocopherol [*119-13-1*]: $R^1 = R^2 = H$, $R^3 = CH_3$

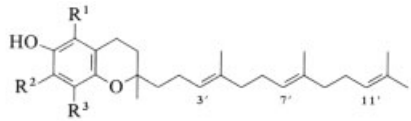

2a, α-Tocotrienol [*1721-51-3*]: $R^1 = R^2 = R^3 = CH_3$
2b, β-Tocotrienol [*490-23-3*]: $R^1 = R^3 = CH_3$, $R^2 = H$
2c, γ-Tocotrienol [*14101-61-2*]: $R^1 = H$, $R^2 = R^3 = CH_3$
2d, δ-Tocotrienol [*25612-59-3*]: $R^1 = R^2 = H$, $R^3 = CH_3$

Of the above eight compounds, α-tocopherol (**1a**), 2,5,7,8-tetramethyl-2-(4′,8′,12′-trimethyltridecyl)-6-chromanol, has the highest biological vitamin E activity and, therefore, the greatest industrial and economic importance. It is the predominant tocopherol in human and animal tissue.

The tocopherols (**1**) have three chiral centers, at positions 2, 4′, and 8′. Until now, only tocopherols having the R configuration at all three chiral centers have been found in nature.

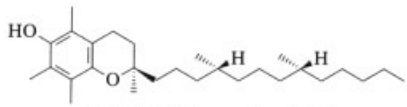

(2*R*,4′*R*,8′*R*)-α-tocopherol (**1a**)

The correct denomination for naturally occurring α-tocopherol (**1a**) is (2*R*,4′*R*,8′*R*)-α-tocopherol [*59-02-9*] or briefly (*R,R,R*)-α-tocopherol. Totally synthetic α-tocopherol, which is racemic with respect to each stereocenter, should be denominated (2*RS*,4′*RS*,8′*RS*)-α-tocopherol [*10191-41-0*], or briefly (all-*rac*)-α-tocopherol. The denominations *d*- and *dl*-α-tocopherol, which have been used frequently for the (*R,R,R*)-enantiomer and the mixture of all eight stereoisomers, respectively, should not be used anymore.

The tocotrienols (**2**) have the (2*R*,3′*E*,7′*E*) configuration.

4.2. History

While studying the connection between fertility and nutrition, EVANS and BISHOP discovered in 1922 [291] that the absence of a substance in the diet of female rats led to death of the fetus and its resorption. This substance was first designated Factor X and later vitamin E [292]. In the 1920s and 1930s, it was observed that apart from the damage done to the reproductive system in male and female rats, a diet in which vitamin E had been oxidatively destroyed by iron(III) chloride produced encephalomalacia in chickens and muscular dystrophy in guinea pigs and rabbits [293–295].

In 1936, EVANS isolated from wheat-germ oil an alcohol which was given the name α-tocopherol (Greek: *tokos*, childbirth; *phero*, to bear) [296].

In 1937, two additional substances were isolated via crystallized allophanates and named β- and γ-tocopherol; they have lower biological activity than α-tocopherol [297]. δ-Tocopherol was discovered later [298].

In 1938, FERNHOLZ elucidated the structure of α-tocopherol based on the results of pyrolytic and oxidative degradation [299]. In the same year, KARRER et al. succeeded in synthesizing α-tocopherol from trimethylhydroquinone and phytyl bromide [300].

The structures of α-, β-, γ-, and δ-tocotrienol were elucidated in the 1960s [301]. The first synthesis of the tocotrienols was achieved by ISLER et al. in 1963 [302]. This was followed in 1976 by the synthesis of α-tocotrienol in the natural 2*R* form [303].

4.3. Physical Properties

The melting point of natural α-tocopherol lies between 2.5 and 3.5 °C and that of γ-tocopherol between −3 and −2 °C [304]. The melting points of various esters, e.g., acetate [305], allophanate [297], *p*-phenylazobenzoate [306], *p*-nitrophenylurethane [297], and 2,4-dinitrobenzoate [300], can be used to determine the purity.

The optical rotation of the natural tocopherols is very small and solvent dependent. The specific optical rotations $[\alpha]^{25}_{546}$ of some stereoisomeric α-tocopherols and their derivatives are listed in Table 12 and those of the natural tocopherols are given in the following:

α-Tocopherol	+0.32 (ethanol)	−3.0 (benzene)
β-Tocopherol	+2.9 (ethanol)	+0.9 (benzene)
γ-Tocopherol	+2.2 (ethanol)	−2.4 (benzene)
δ-Tocopherol	+3.4 (ethanol)	+1.1 (benzene)

Tocopherols are light yellow viscous oils at room temperature. They are readily soluble in lipophilic solvents and ethanol, but practically insoluble in water.

The UV spectra of the tocopherols and tocotrienols in ethanol have a maximum at 292–298 nm (α-tocopherol: 292 nm). Acetylation of the phenolic hydroxyl group shifts this maximum to 276–285 nm (α-tocopherol acetate: 284 nm) [309].

Nuclear resonance spectroscopy [283,312, 313], IR spectroscopy [310,311], and mass spectroscopy [314,315] are used to differentiate differently substituted tocopherols and tocotrienols.

4.4. Chemical Properties

The tocopherols (**1**) are gradually oxidized by atmospheric oxygen, producing a reddish color. In the absence of oxygen, they are resistant to alkali and heat up to 200 °C, and can be vacuum distilled without decomposition.

Tocopherol esters are resistant to atmospheric oxygen at ambient temperature.

Other chemical properties of the tocopherols are determined by:

1) The free phenolic hydroxyl group, which can be esterified [316] or etherified
2) A free position in the aromatic ring (except in α-tocopherol and α-tocotrienol) which is available for hydroxymethylation, chloromethylation, nitrosation, diazotization, etc.
3) Ready oxidizability (as the monoether of a hydroquinone) [317]

Tocopherol exerts an antioxidative effect, terminating radical chain reactions. This is due to the ready release of a hydrogen atom with the formation of the relatively inert tocopheroxy radical [318–320]. Tocopherol can be regenerated from the tocopheroxy radical in vivo by ascorbate or coenzyme Q10–hydroquinone [319,320]. More recent studies show that α-tocopherol is a much more potent antioxidant

Table 12. Specific optical rotation of stereoisomeric α-tocopherols [284, 306, 308]

Stereoisomer	Free α-tocopherol, $[\alpha]_D^{25}$ (ethanol)	*p*-Phenylazobenzoate, $[\alpha]_{600}^{25}$ ($CHCl_3$)	Prod. of oxidation with $K_3Fe(CN)_6$, $[\alpha]_D^{25}$ (isooctane)	Acetate, $[\alpha]_D^{25}$ (ethanol) [307]
2*R*,4′*R*,8′*R*	+0.75°	+7.07°	+26°	+3.3°
2*RS*,4′*R*,8′*R*	+0.38°	0°	0°	
2*S*,4′*R*,8′*R*	+0.19°	−7.64°	−25.8°	−2.1°
2*R*,4′*RS*,8′*RS*		+6.96°	+25.8°	+1.9°
2*S*,4′*RS*,8′*RS*	−0.14°	−7.47°	−23.6°	

than γ-tocopherol and the other homologues with a lower degree of methylation [321].

4.5. Metabolism and Importance in the Organism

Vitamin E is the most important lipid-soluble biological anitoxidant [322]. It is localized in cellular and subcellular membranes which are rich in polyene fatty acids. Since it reacts with free radicals faster than polyene fatty acids, it interrupts the chain reaction of lipid oxidation. Therefore, vitamin E is essential for the morphological integrity and functionality of the membranes of all cells [323].

The antioxidative effect in vitro is not identical with the biological activity: γ-tocopherol has about half the antioxidative effect of α-tocopherol but only about one-tenth of its biological activity [324]. This discrepancy appears to be due to differences in the lipoprotein-bound transport of tocopherols in the plasma.

In the course of fat digestion, vitamin E is absorbed in the small intestine [286, 325]. Bile acids and pancreatic lipase are required for absorption. In the case of malabsorption of lipids, the risk of vitamin E deficiency exists. Transport from the intestine to the systemic circulation proceeds via the chylomicrons of the lymph. Vitamin E is transported in the plasma bound to lipoproteins (mainly low density lipoproteins, LDL) [324]. The normal plasma level is 10 µg/mL.

The functions of vitamin E are revealed by the pathological results of its deficiency (see Section 4.6) [286]. Epidemiological studies on humans have shown that a high vitamin E intake is associated with a lower risk of atherosclerosis and coronary heart disease [326]. Prophylactically, vitamin E appears to be of importance in cell ageing, in diseases caused by the involvement of free radicals (e.g., ventricle bleeding in premature infants, senile cataract, coronary heart disease, and other ischemic reperfusion damage), and in cancer of various organs [327, 328]. Vitamin E stimulates cellular and humoral immunity [329].

Vitamin E is practically nontoxic, excess vitamin being eliminated via the bile.

4.6. Deficiency, Requirement, and Application

Vitamin E deficiency can occur due to an inadequate supply of the vitamin and lipid malabsorption. A lack of selenium and a high intake of polyene fatty acids enhance this deficiency. In animals, vitamin E deficiencies mainly affect the muscles, nervous system, cardiovascular system, formation of blood, and reproduction:

1) Muscles: The most frequent and economically most important manifestation of vitamin E deficiency in farm animals is degeneration of skeletal muscles ("white muscle disease," nutritional muscular dystrophy). As a rule, newborn and young animals (young chicks, lambs, piglets, calves, foals) are affected because their vitamin E supplies are low. Adequate vitamin E supply to the dam prevents this condition, and a combined injection of vitamin E and selenium is therapeutically effective.
2) Nervous system: In young chicks vitamin E deficiency leads to encephalomalacia, a fatal disease of the cerebellum which is caused by edema and necrosis.
3) Reproduction: In the female rat, vitamin E deficiency results in fetal death and resorption of the fetus. The classical biotest for the determination of the activity of vitamin E preparations is based on this manifestation. In the male, testicular function is disturbed

by irreversible degeneration of the germ cells (rat, rabbit, dog, monkey, pig).

4) Cardiovascular system: Piglets develop mulberry heart disease, an acute and usually fatal cardiomyopathy.

Clinical vitamin E deficiency in humans is rare [324]. This disease is usually associated with prolonged deficiency and lipid malabsorption. In rare cases, an inborn deficiency of plasma lipoproteins (abetalipoproteinemia) is the cause [324]. Neurological disorders such as cerebellar ataxia, a disturbed sense of position, loss of reflexes, and loss of the sensation of vibration are the main symptoms. They are caused by degenerative changes in the large motoneuron axons and in the myelin sheath [330].

In animal feeding, the following vitamin E additions are recommended (per animal and day): dairy cows 200 mg, young cattle 40–70 mg, beef cows 60–100 mg, and horses 600 mg.

The following additions are recommended per kilogram of complete feed: laying hens 20 mg, growing/finishing broilers 40 mg, turkeys 45 mg, piglets 45 mg, growing/finishing pigs 30 mg.

The recommended supply of vitamin E in the human diet in Germany is listed in Table 13.

Table 13. Recommended levels of vitamin E in the diet in milligrams tocopherol equivalent ∗ per day as stipulated by the Deutsche Gesellschaft für Ernährung (German Society for Nutrition) 1985

	RDI
Infants up to 2 months	3
Infants 3 to 11 months	4
Children, 1 to 3 years	5
Children, 4 to 6 years	7
Children, 7 to 9 years	8
Children, 10 to 12 years	10
Adolescents (15 years up)	12
Adults, 19 to 65 years	12
Pregnant women	14
Nursing mothers	17

∗ 1 mg (*R,R,R*)-α-tocopherol equivalent = 1.1 mg (*R,R,R*)-α-tocopheryl acetate = 1.49 mg (all-*rac*)-α-tocopheryl acetate = 2.0 mg (*R,R,R*)-β-tocopherol = 4 mg (*R,R,R*)-γ-tocopherol = 1.49 I.U.

4.7. Analysis and Standardization

Determination. Apart from photometric methods, chromatographic methods are primarily employed for analysis of vitamin E today.

An older method used to determine vitamin E in biological material is the Emmerie–Engel reaction, which is based on the colorimetric measurement of colored Fe(II) complexes with bipyridine [331] or 4,7-diphenyl-1,10-phenanthroline [332]. Spectrofluorometry is a very sensitive analytical method for free and esterified tocopherols. For instance, this method is used for vitamin E determinations in the spinal cord and in erythrocytes [333]. Photometric determination of tocopherol with the 2,2-diphenyl-1-picrylhydrazyl radical is also possible [334].

Of the chromatographic methods, gas chromatography (GC) and high performance liquid chromatography (HPLC) allow the most accurate determinations [335]. HPLC permits the simultaneous determination of vitamin E homologues [336], even directly in plasma and tissue samples [337].

The measurement of the optical rotation of $K_3Fe(CN)_6$ oxidation products can be used to differentiate the individual stereoisomers of α-tocopherol (see Table 12). However, only the configuration at C-2 can be determined with reasonable reliability. The diastereomers can be differentiated by gas chromatography [338] and ^{13}C NMR spectroscopy [339]. All eight stereoisomers of (all-*rac*)-α-tocopherol can be separated by using a combination of chiral HPLC and GC [340].

Standardization of Biological Activity. The relative biological activity of vitamin E is expressed in International Units (I.U.), 1 I.U. corresponding to the activity of 1 mg of racemic (2*RS*,4′*RS*,8′*RS*)-α-tocopherol acetate [*7695-91-2*]. The relative biological activities were mainly derived from fetus resorption tests on rats. They have been established by the National Formulary as follows [341]:

1 mg (2*RS*,4′*RS*,8′*RS*)-α-tocopherol acetate = 1.00 I.U.
1 mg (2*RS*,4′*RS*,8′*RS*)-α-tocopherol = 1.10 I.U.
1 mg (2*R*,4′*R*,8′*R*)-α-tocopherol acetate = 1.36 I.U.
1 mg (2*R*,4′*R*,8′*R*)-α-tocopherol = 1.49 I.U.

According to [342], 2*R*,4′*R*,8′*R*-α-tocopherol has a higher biological activity than that established by the National Formulary. As far as farm animals and humans are concerned,

the biological activity of the different forms of vitamin E is a controversial subject.

The relative biological activity of the eight α-tocopherol stereoisomers, determined by fetus resorption tests in rats [343] is as follows:

2*R*,4′*R*,8′*R*	100
2*R*,4′*R*,8′*S*	90
2*R*,4′*S*,8′*S*	73
2*R*,4′*S*,8′*R*	57
2*S*,4′*S*,8′*S*	60
2*S*,4′*R*,8′*R*	31
2*S*,4′*S*,8′*R*	21
2*S*,4′*R*,8′*S*	37

The *R* configuration at C-2 of the chromane ring seems to be important for high biological activity.

Biotests. Data on the relative biological activities of the tocopherol stereoisomers are obtained in animal experiments. The biological activity is determined by the ability of the test substances to prevent or remedy symptoms of vitamin E deficiency in comparison with a standard substance. The following test systems (bioassays) are preferentially used:

1) The fetus resorption test [344] in pregnant rats with vitamin E deficiency is the classical method for the determination of the biological activity of tocopherols.
2) The erythrocyte hemolysis test in the rat [345] is an in vitro process which measures the protective function of vitamin E against peroxide-induced hemolysis of erythrocytes. This test is also applied to humans [346].
3) In the muscular dystrophy test in the rat, rabbit, and chicken, muscular damage due to vitamin E deficiency is measured either directly [347] or via the determination of enzyme activities, e.g., plasma pyruvate kinase activity [348] as a measure of the course of muscular dystrophy.

4.8. Occurrence

All eight tocopherol (**1a**–**1d**) and tocotrienol homologues (**2a**–**2d**) occur in nature. Appreciable amounts are contained, above all, in plant oils. The highest concentration by far is found in wheat germ oil (ca. 0.2 %) [349]. Considerable amounts are also contained in soybean, sunflower, rape, corn, and peanut oil.

A survey of the total concentration of tocopherols in foods (in milligrams per 100 g) is provided in the following:

Wheat germ oil	215.4
Soybean oil	14.6
Sunflower oil	55.8
Peanut oil	17.2
Olive oil	12.0
Cocoa butter	1.0
Coconut fat	0.8
Pollard	2.4
Corn	2.0
Wheat	1.4
Barley	0.2
Oats	1.5
Soybeans	1.2
Beans (white)	2.1
Asparagus	1.5
Tomatoes	0.9
Carrots	0.6
Beans (green)	0.3
Garlic	0.1
Cauliflower	0.1
Walnut	20.0
Hazelnut	26.0
Peanut	9.0
Chestnut	1.2
Coconut	1.0

The concentrations of vitamin E in animal foods are much lower than those in plant products. The concentration is subject to major fluctuations caused by the seasonal concentration variations in green forage plants.

The concentration of various tocopherol and tocotrienol isomers in foods and the effect of processing and storage on the tocopherol content are described in [350].

4.9. Economic Aspects

Animal feeds represent the most important market for vitamin E. Of the world production, ca. 75 % is added to industrially produced feed mix for animals. The main product is a 50 % adsorbate of tocopherol acetate on silica gel. Formulations dispersible in cold water also exist.

The pharmaceutical, food, and cosmetic sectors account for ca. 25 % of world consumption. Special formulations are used here, e.g., spray formulated dry powders or adsorbates. The compound mainly used in the encapsulation of vitamin E is 98 % (all-*rac*)-α-tocopherol acetate. (All-*rac*)-α-tocopherol is also employed to prevent oxidation processes in oils and fats.

While totally synthetic (all-*rac*)-α-tocopherol is represented in all areas of application, the use of semisynthetic (*R,R,R*)-α-tocopherol from natural sources is almost exclusively restricted to the pharmaceutical, food, and cosmetic industries.

The most important producers of (all-*rac*)-α-tocopherol acetate are Hoffmann-La Roche (Switzerland) which has plants in Switzerland and the United States, and BASF (Germany), with plants in Germany and the United States. Other producers are Rhône-Poulenc (France) and Eisai (Japan). Vitamin E from natural sources is produced by ADM and Henkel in the United States. Smaller amounts are also produced in Japan.

In 1994, ca. 20 000 t of synthetic vitamin E and ca. 2000 t of vitamin E from natural sources (calculated as 100 % product) were consumed by the world market. Regions of high demand are the United States and Western Europe. Owing to consumer habits, demand in the human sector in the United States is more pronounced than in other regions of the world.

4.10. Biosynthesis

Tocopherols are synthesized by higher plants and algae. The precursor of the aromatic nucleus of the tocopherols is shikimic acid (**3**) which is converted to homogentisic acid (**4**) [351]. This synthesis can proceed in plants, algae, and bacteria, but not in the animal organism.

Starting with homogentisic acid (**4**), there are two pathways for the biogenesis of vitamin E (Scheme 10).

In the "tocopherol pathway," [352] homogentisic acid (**4**) reacts with phytyl pyrophosphate with elimination of CO_2 to give the intermediate **6** [353]. The "tocotrienol pathway" starts with the condensation of homogentisic acid (**4**) with geranylgeranyl pyrophosphate to give the intermediate **5**. Enzymatic cyclization of **5** or **6** yields δ-tocotrienol or δ-tocopherol, respectively. Some of the enzymes involved in these synthetic pathways have been isolated [354].

The *S*-methyl group of *S*-adenosylmethionine methylates δ-tocotrienol to give β-, γ-, and α-tocotrienol, and δ-tocopherol to give the corresponding tocopherols [355]. In some organisms, the second methyl group in the aromatic ring is introduced before cyclization. The enzyme responsible for subsequent cyclization to γ-tocopherol has been isolated from algae [356], and the cyclization mechanism elucidated [357].

The hydrogen source for the conversion of tocotrienols to the corresponding tocopherols is NADPH [358].

4.11. Production

4.11.1. Semisynthetic Vitamin E: Isolation from Plant Oils and Methylation

The most important natural sources of vitamin E are plant oils and fats. The deodorization of edible oils [359] yields as a byproduct deodorizer sludges which contain tocopherols in economically interesting concentrations (3–15 %) [360]. The deodorizer sludge from alkaline chemical refining of soybean oil, for instance, contains ca. 10 % of tocopherols, 10–20 % of free sterols and their fatty acid esters, 40–50 % of free fatty acids, triglycerides, and hydrocarbons (e.g., squalene).

Various processes are available for the purification of tocopherols. The fatty acids can be extracted [361] and separated in the form of sparingly soluble calcium salts [362]. Alternatively, they can be eliminated by distillation, preferably after esterification [363]. The sterols

can be separated by treatment with hydrogen chloride [364] or $CaCl_2$ [365], by crystallization [361, 366], or by extraction [367]. Alternatively, the free sterols can be esterified with the excess fatty acids, and the tocopherols can be separated from the high-boiling sterol fatty acid esters by distillation [368].

After removal of the free fatty acids, the tocopherols can be concentrated by adsorption on basic ion exchangers.

The amount of α-tocopherol and, consequently, the specific vitamin E activity in the mixture of tocopherols isolated from soybean oil are low (high proportion of δ- and γ-tocopherol).

To improve the natural mixed tocopherols, methyl groups are substsituted for the aromatic hydrogens. As a rule, halogen-, nitrogen-, or oxygen-substituted methyl groups are introduced and the substituents subsequently eliminated by catalytic hydrogenation or reduction

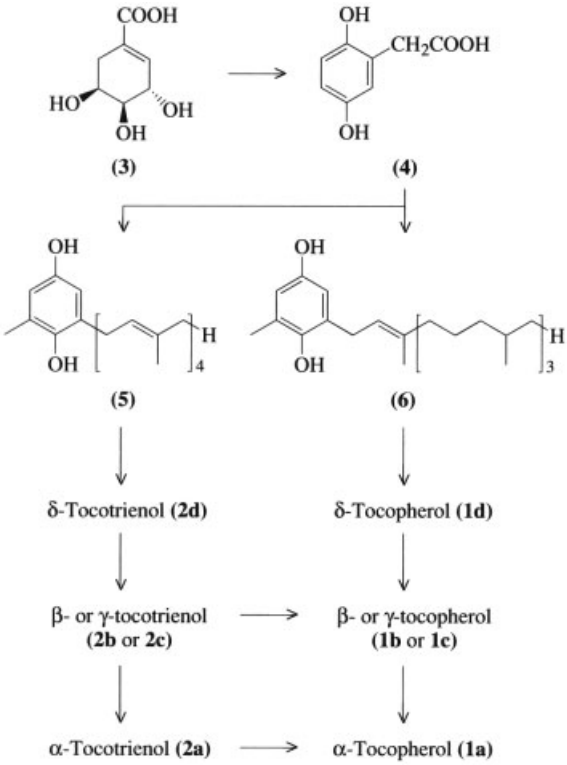

Scheme 10. Biosynthesis of tocotrienols and tocopherols

with zinc/hydrochloric acid. Hydroxymethylation with formaldehyde under acidic [370], alkaline [371], or neutral [372] conditions is the method of choice. Other processes are chloromethylation with formaldehyde and HCl [373] or $POCl_3$ [374]; formylation with zinc cyanide/ HCl, chloroform/alkali, or hexamethylenetetramine [375]; and Mannich aminomethylation [376].

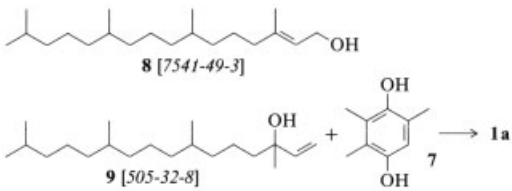

4.11.2. Totally Synthetic Vitamin E: Industrial Synthesis

The industrial total synthesis of α-tocopherol (**1a**) is based on the condensation of 2,3,6-trimethylhydroquinone [*700-13-0*] (**7**) with phytol (**8**), phytyl halides, or preferentially isophytol (**9**).

Lewis and Brönsted acids [306], especially zinc chloride with a mineral acid [377], serve as catalysts for condensation. Boron trifluoride and $AlCl_3$ [378], Fe/HCl [379], trifluoracetic acid [380], and boric acid/carboxylic acid [381] are also good catalysts. A wide range of solvents can be used, e.g., ethyl acetate, acetic acid, (chlorinated) hydrocarbons or supercritical CO_2 [382].

When synthetic phytol (**8**) or isophytol (**9**) is used, the product consists of equal amounts of all eight α-tocopherol stereoisomers and is thus denominated (all-*rac*)-α-tocopherol.

The majority of synthetic α-tocopherol is treated with acetic anhydride to give the acetate, which is stable on storage. In addition, the succinate (monoester of succinic acid) [383] and the nicotinate are produced commercially. Purification of the crude product is possible at the free tocopherol or acetate stage by short-path distillation under high vacuum or extraction with supercritical CO_2 [384].

4.11.3. Trimethylhydroquinone

The starting material in the most important industrial syntheses of trimethylhydroquinone (**7**) is 2,3,6-trimethylphenol [*2416-94-6*] (**11**) [385], which is oxidized by oxygen or peroxides to give trimethylquinone (**13**) [*935-92-2*]. The catalysts mainly used are salcomin [386], copper chloride with diverse cocatalysts [387], or heteropoly acids [388]. Trimethylhydroquinone (**7**) is finally obtained by the catalytic hydrogenation of trimethylquinone (**13**) [389].

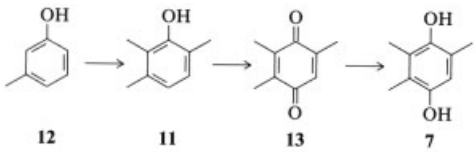

The intermediate 2,3,6-trimethylphenol (**11**) is produced from *m*-cresol (**12**) by catalytic alkylation with methanol [390].

Another important large-scale process for the synthesis of trimethylphenol (**11**) is the condensation of diethyl ketone with crotonaldehyde to yield trimethylcyclohexenone [*20030-30-2*] (**10**) [391] or diethyl ketone with methyl vinyl ketone to give trimethylcyclohexenone [*20030-29-9*] (**10′**) [392], followed by catalytic dehydrogenation of trimethylcyclohexenone to 2,3,6-trimethylphenol (**11**).

Still another synthesis of trimethylhydroquinone starts with mesitol (**14**) [393], which is

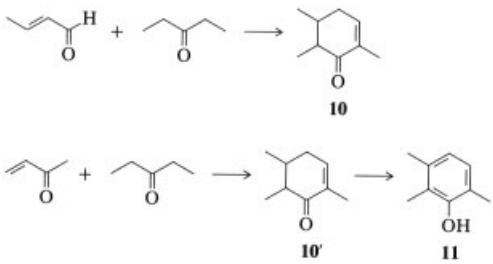

oxidized with oxygen [394] or hypochlorite or chlorine/water [395] to give trimethylquinol (**15**). Compound **15** is then rearranged by treatment with alkali [396] to yield trimethylhydroquinone (**7**). This pathway was long used by Rhône-Poulenc, but has now been replaced by oxidation of 2,3,6-trimethylphenol (**11**) with oxygen because oxidation with chlorine causes wastewater problems [397].

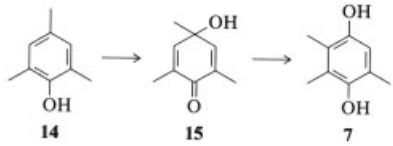

Trimethylhydroquinone (**7**) can also be made from the inexpensive compound isophoron (**16**). Oxidation to trimethylquinone (**13**) can be performed directly [398] or via the intermediate 2,3,5-trimethylphenol (**11′**), which gives **7** on hydrogenation [399]. Oxidation of **16** to oxoisophoron (**17**), followed by isomerization produces trimethylhydroquinone (**7**) directly [400]. These processes are, however, not used commercially.

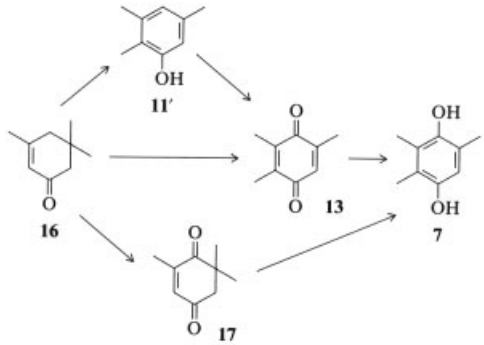

4.11.4. Phytol and Isophytol

The key sequence in the formation of the vitamin E side chain is a series of C_2 and C_3 chain extensions, starting with a ketone.

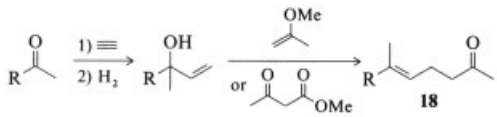

The C_2 extension of the ketone proceeds via vinylation with vinyl Grignard or by ethynylation with subsequent partial hydrogenation. The allyl alcohol thus formed

is extended to give the γ,δ-unsaturated ketone (**18**) by means of acid catalyzed ketal formation/elimination/rearrangement with the C_3 building block 2-methoxypropene (Saucy–Marbet reaction [401]) or by transesterification/rearrangement with methyl acetoacetate (Carroll reaction [402]). The ketone (**18**) can be subjected to the same sequence again. The Saucy–Marbet reaction can also be applied to tertiary propargyl alcohols, forming the β-ketoallene (**18′**). In this case, partial hydrogenation after ethynylation is not required [403].

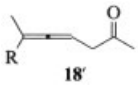

In this procedure, acetone is first converted to methyl heptenone [*110-93-0*] (**18**, R = Me), and then via linalool [*78-70-6*] (**19**) to geranylacetone [*689-67-8*] (**18**, R = prenyl), nerolidol [*1119-38-6*] (**20**) and farnesylacetone [*1117-52-8*] (**18**, R = geranyl), or their hydrogenation products. Hexahydrofarnesylacetone [*502-69-2*] is converted to isophytol (**9**), which can be isomerized to phytol (**8**) by acid-catalyzed rearrangement.

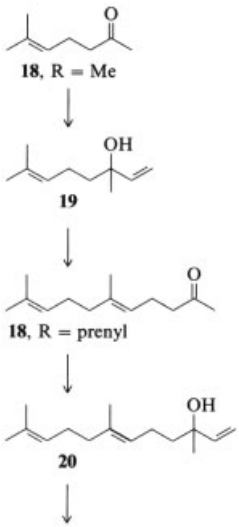

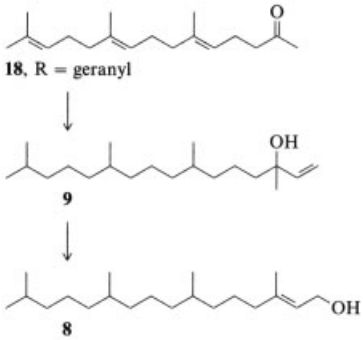

One of the four isophytol producers (Kuraray) makes methylheptenone (**18**, R = Me) from prenyl chloride and acetone [404]. In the late 1980s, Rhône-Poulenc changed its process, basing it on the inexpensive raw material myrcene (**21**) [*123-35-3*] [397], which yields geranylacetone (**18**, R = prenyl) in a rhodium-catalyzed reaction with methyl acetoacetate [405]. This compound is then further processed in the conventional way.

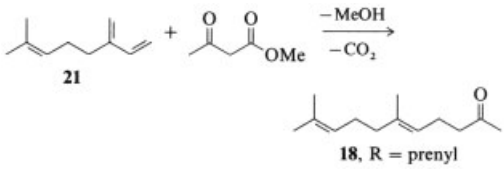

4.11.5. Stereoselective Syntheses of α-Tocopherol

In 1938, the first synthesis of (2*RS*,4′*R*,8′*R*)-α-tocopherol (**1a**) was conducted by KARRER et al. by the condensation of trimethylhydroquinone (**7**) and phytyl bromide (made from natural phytol) with zinc chloride [300].

This reaction was frequently modified [306]. Most of the α-tocopherol syntheses proceed via acid-catalyzed condensation of hydroquinone derivatives with phytol or isophytol derivatives. Tocopherols produced in these syntheses are always racemic at the C-2 center of the chromane ring.

The first synthesis of (2*R*,4′*R*,8′*R*)-α-tocopherol (**1a**) was conducted by MAYER et al [308] in conjunction with the elucidation of the absolute configuration. In this process, trimethylhydroquinone (**7**) was

converted to (2*R*,4′*R*,8′*R*)-tocopherol via chromanyl(ethinylcarboxylic acid) (**22**) – racemate separation was achieved via the quinine salt – and via the chromane carbaldehyde (**23**). The optically active side chain originated from natural phytol (**8**), which was converted to the corresponding Wittig reagent (Scheme 11).

Several modern total syntheses of (2*R*,4′*R*,8′*R*)-α-tocopherol were described in the late 1970s and early 1980s [406, 407]. These processes depended on the separation of racemic mixtures for the stereoselective formation of the chromane skeleton.

According to more recent results, the quaternary center at C-2 can be made by asymmetric synthesis, which uses either chiral auxiliaries in stoichiometric amounts [408, 409] (Scheme 12) or enantioselective catalytic processes [409, 410] (Scheme 13).

Enzymatic methods using tocopherol cyclases have also been successfully employed to generate the C-2 asymmetric center [356, 357, 411]. The γ-tocopherol initially obtained can be converted to the α-derivative by methylation (see Section 4.11.1).

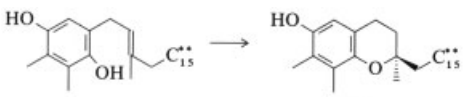

1c

In the mid-1980s [412], an asymmetric catalytic hydrogenation process was developed for the synthesis of the enantiomerically pure side chain (Scheme 14).

This method has greatly improved the synthetic availability of the side chain with regard to both the number of reaction steps and the total yield. The enantiomeric excess achieved is 98.4 %. Nevertheless, compared with the availability of natural vitamin E, the effort involved in the commercial synthesis of (*R,R,R*)-tocopherol makes it uneconomical.

5. Vitamin K

5.1. Introduction; History

In 1929, while studying the metabolism of cholesterol in chickens fed a diet extracted with nonpolar solvents, Dam observed bleeding in various organs and tissues and a prolonged blood clotting time [413]. The symptoms could be prevented or cured by the feeding of plant or animal feedstuffs. Thus, Dam had discovered a new, essential, fat-soluble factor that controls the blood prothrombin level, which is important for blood clotting. In 1935, he suggested calling this factor vitamin K. At that time, the letter K was not only the next free letter for the designation of an existing or postulated vitamin, but K is also the first letter of the word for coagulation in German and the Scandinavian languages.

At the same time, Almquist found that an antihemorrhagic factor is also formed by bacteria in decaying feedstuffs (e.g., fish meal) [414]. In 1939/40, isolation, structural elucidation, and synthesis in various laboratories (Dam, Almquist, Karrer, Doisy, Fieser; see [413, 414]) showed surprisingly that two different forms of vitamin K occur in nature. They differ in the length and degree of saturation of the isoprenoid side chain linked to 2-methyl-1,4-naphthoquinone at C-3. Vitamin K_1 from plants has a mono-unsaturated C_{20} side chain, and bacterial vitamin K_2 has a polyunsaturated side chain of different length.

It was shown in 1974/75 that vitamin K functions as a cofactor for the γ-carboxylation of specific glutamic acid residues of the precursor protein of prothrombin, converting the protein to biologically active prothrombin [415].

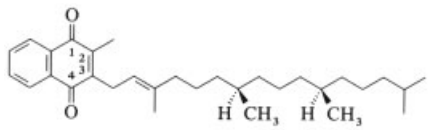

1

The term vitamin K is now used as the generic name for 2-methyl-1,4-naphthoquinone and all derivatives that qualitatively exhibit the biological activity of phylloquinone (vitamin K_1) [416]. The following names have been recommended by IUPAC for the individual vitamins K [417]:

Phylloquinone, abbreviated K, previously vitamin K_1, for 2-methyl-3-phytyl-1,4-naphthoquinone. The U.S. Pharmacopoeia uses the name phytonadione [418]. The synonym phytomenadione is occasionally found in the

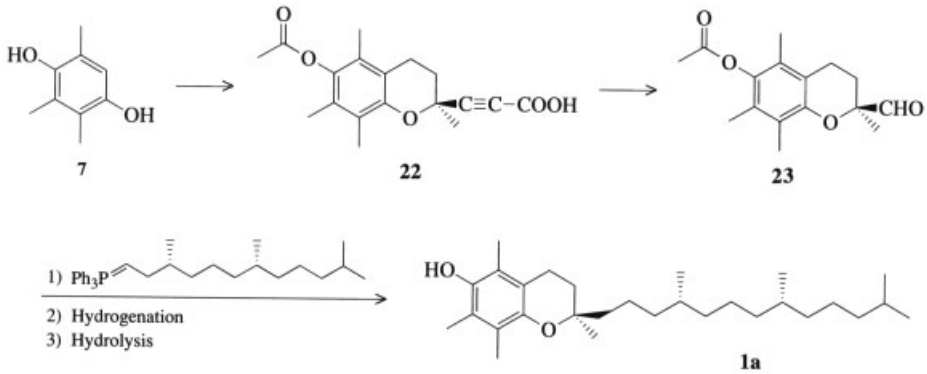

Scheme 11. Synthesis of (*R,R,R*)-α-tocopherol (**1a**)

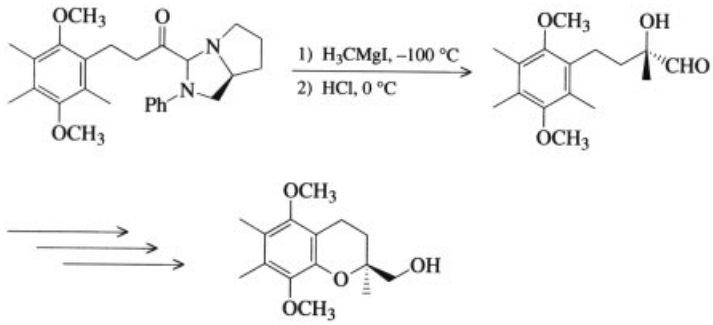

Scheme 12. Asymmetric chromanol synthesis with chiral auxiliary

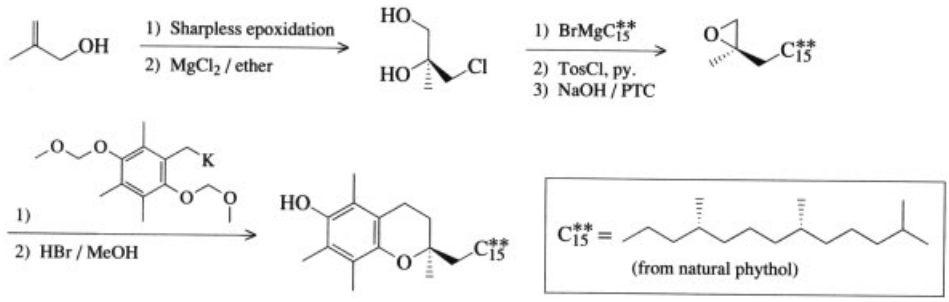

Scheme 13. Asymmetric synthesis of (*R,R,R*)-α-tocopherol
PTC = phase transfer catalyst

literature. Natural phylloquinone [*84-80-0*] (**1**) has the configuration 2′*E*,7′*R*,11′*R* [419].

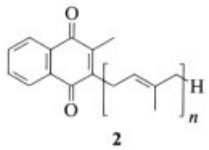

2

Menaquinone-n (**2**), abbreviated MK-*n*, previously vitamin K_2, for 2-methyl-3-polyprenyl-1,4-naphthoquinone, *n* being the number of prenyl residues.

For instance, the compound originally isolated from rotting fish meal and named vitamin K_2 was later identified as menaquinone-7 (MK-7) [*2124-57-4*] (2-methyl-3-farnesylgeranylgeranyl-1,4-naphthoquinone) [419]. For MK-7, the older literature contains the designation vitamin $K_{2(35)}$ or vitamin K_2(35), which is no longer used. Menaquinones found in nature have side chains of 4 – 13 prenyl residues and usually exhibit the all-*trans* configuration.

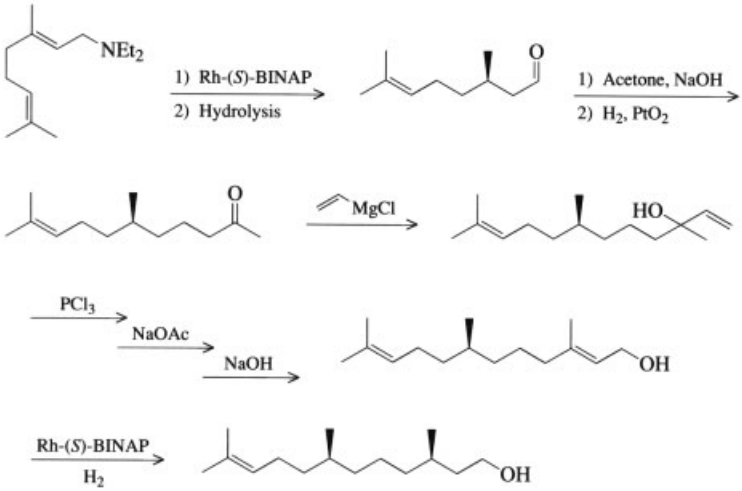

Scheme 14. Synthesis of *R,R* side chain by catalytic hydrogenation

However, menaquinones with the *cis* configuration and partially saturated side chains also exist [420].

Menadione [*58-27-5*] (**3**), previously vitamin K_3, for 2-methyl-1,4-naphthoquinone. It does not occur in nature but can be converted to menaquinone-4 in the organism [421].

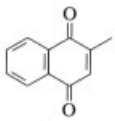

3

5.2. Physical Properties

All the vitamins K are lipophilic compounds which are readily soluble in ether, petroleum ether, benzene, hexane, and acetone, poorly soluble in methanol and ethanol, and insoluble in water. Due to the naphthoquinone chromophore, they have characteristic UV maxima (Table 14).

Table 14. UV maxima and extinctions ($E^{1\,\%}_{1\,cm}$) of phylloquinone and menaquinone-7 ($n = 7$) in petroleum ether [419]

Phylloquinone	Menaquinone-7
243 (395)	243 (278)
249 (420)	248 (295)
261 (390)	261 (266)
270 (392)	270 (267)
325 (68)	325 – 328 (48)

The infrared spectrum of vitamin K_1 and of all the K_2 vitamins shows intensive bands at 1660, 1618, 1300, and at 700 cm^{-1}, characteristic of the naphthoquinone ring [422].

In the 1960s, the ^1H NMR spectra were used for the determination of the configuration of the trisubstituted double bond in the side chain of vitamin K_1 [CH_3-C(3′): *trans*: 1.78 ppm, *cis*: 1.68 ppm] [423].

Vitamin K_1, phylloquinone 1 [422] is a viscous yellow oil at room temperature. It can be crystallized out at low temperatures in acetone or ethanol, *mp* −20 °C, *bp* 140 – 145 °C at 133 Pa, $[\alpha]^{20}_{D}$ −0.3° (dioxane), $C_{31}H_{46}O_2$, M_r 450.71, and ϱ 0.967 g/cm^3. It decomposes at normal pressure at 100 – 120 °C, but can be subjected to high vacuum distillation without appreciable decomposition. The redox potentials are 0.005 V (25 °C, 80 % ethanol, 0.02 molar in NaOAc and 0.02 molar in AcOH) and 0.363 V (20 °C, 95 % ethanol, 0.2 N in HCl and 0.02 molar in LiCl).

Vitamins K_2, menaquinone-n 2. A vitamin K_2 (menaquinone-7; **2**, $n = 7$) with *mp* 54 °C was first isolated in the pure form from rotting fish meal by Doisy in 1939 [424]. Most of the K_2 vitamins are light yellow, crystalline compounds. The melting points and the extinctions ($E^{1\,\%}_{1\,cm}$) at the UV maximum (248 nm) of menaquinone-*n* (**2**, $n = 1 – 10$) are presented in Table 15.

Table 15. Melting points and UV absorptions at 248 nm $E^{1\%}_{1cm}$ of menaquinones-n (n = 1 – 10) in petroleum ether [422]

n	$E^{1\%}_{1cm}$ (248 nm)	*mp*, °C
1	784	
2	617	53
3	496	35
4	440	35
5	363	39
6	320	50
7	295	54
8	268	56
9	246	56 – 57
10	224	62

The extinction $E^{1\%}_{1cm}$ of the K_2 vitamins decreases continuously with increasing chain length. In the structural elucidation, a comparison of these extinction values provided important information about the number of isoprene units in the side chain.

5.3. Chemical Properties [419]

General Aspects. The vitamins K are relatively heat resistant (≤100 °C) and stable in air. However, they are very sensitive to alkali and light. Therefore, they should always be protected from light.

Color Reactions. With sodium ethoxide, the vitamins K give a violet-blue color which, with time, changes to red and subsequently to brown (Dam – Karrer test) [425]. An intense blue color is also obtained with sodium diethyldithiocarbamate [426]. A stable orange color is produced by the reaction of the vitamins K with 5-imino-3-thioxo-1,2,4-dithiazolidine (Schilling – Dam test) [427].

Oxidation/Reduction. The structural elucidation of the vitamins K is based primarily on the oxidative degradation of the quinones and their hydroquinone diacetates. Chromic acid degradation of vitamin K_1 (**1**) yields a mixture of phthalic acid (**4**) and 2-methyl-1,4-naphthoquinone-3-acetic acid (**5**). Reductive acetylation of **1** gives dihydrovitamin K_1 diacetate [*604-87-5*] (**6**) which yields a mixture of diacetoxyaldehyde **7** and 6,10,14-trimethyl-2-pentadecanone (**8**) after ozonization (Scheme 15). On ozonolysis, dihydrovitamin K_2 diacetates (**9**) give the same diacetoxyaldehyde **7**, variable amounts of levulinic aldehyde (**10**) (depending on the chain length), and one equivalent of acetone (**11**) (Scheme 15).

Treatment of vitamin K_1 (**1**) with H_2O_2 in EtOH in the presence of Na_2CO_3 produces vitamin K_1 2,3-epoxide [*25486-55-9*] (**12**), which can be reduced to dihydrovitamin K_1 [*572-96-3*] (**13**) with sulfurous acid. Reduction of vitamin K_1 with sulfurous acid or with H_2/Lindlar catalyst also yields dihydrovitamin K_1 (**13**), which is, in turn, convertible to vitamin K_1 by reaction with oxidizing agents.

Table 16. Phylloquinone content of some vegetables ∗

Vegetable	Phylloquinone content, µg/100 g fresh weight	
	According to [428]	According to [429]
Kale	724	817
Spinach	415	385
Purslane		381
Brussel sprouts		175
Outer leaves	400	475
Sprouts	177	
Broccoli	147	205
Cos lettuce	120	123
White cabbage		55
Outer leaves	137	
Inner leaves	83	
Cress	88	
Beans	46	53
Peas	39	33
Cauliflower	27	25
Red cabbage	19	57
Cucumber	15	22
Leek	10	18
Tomato	6	6
Carrot	5	11

∗ Determined by high performance liquid chromatography (HPLC).

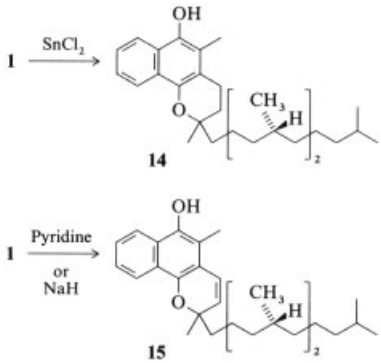

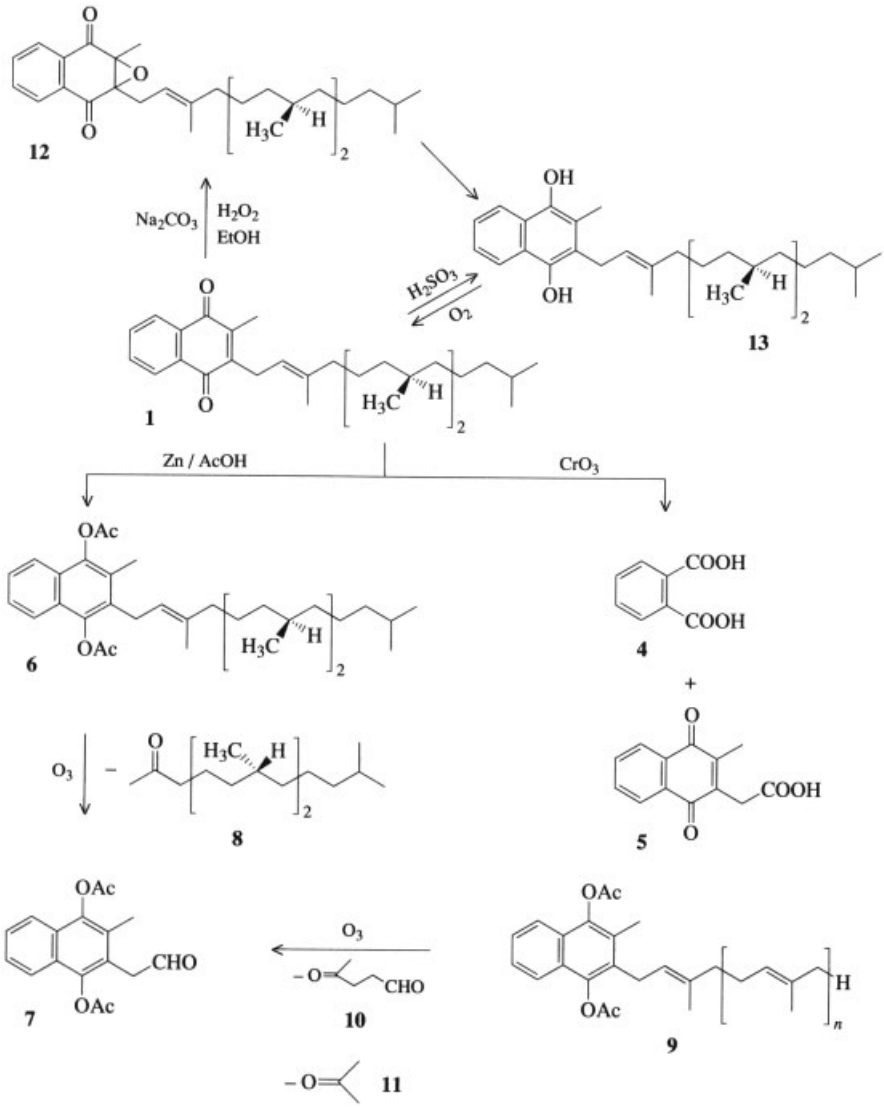

Scheme 15. Oxidation/reduction of vitamin K_1

Formation of Chromanol and Chromenol. Chromanols and chromenols regained importance after it was shown that the corresponding 6-phosphate derivatives of these compounds appear as intermediates in oxidative phosphorylation in cells.

Chemical conversion of vitamin K_1 (**1**) to chromanol (**14**) is achieved with tin(II) chloride. Cyclization to chromenol (**15**) is possible with pyridine or with sodium hydride.

5.4. Occurrence

Phylloquinone is widely distributed in higher plants and in green and blue algae. It is localized mainly in the chloroplasts [420]. Thus, green leafy vegetables contain relatively high amounts of phylloquinone (Table 16). The phylloquinone content is not considerably altered by cooking [429]. Comparatively little phylloquinone is present in fruit.

The vitamin K_1 content of some fruits (in μg/100 g fresh weight) is as follows [430]:

Avocados	20
Apples	5
Peaches	5
Bananas	1
Oranges	1
Pears	1
Pumpkins	1

Plant oils contain greatly varying amounts of vitamin K_1 (in µg/100 g) [431]:

Soybeans	193
Rapeseed	141
Olive	55.5
Sesame	15.5
Walnut	15
Safflower	9.1
Sunflower	9
Almond	6.7
Corn	2.9
Peanut	0.6

The phylloquinone concentration in plants can vary depending on genetic factors or on the site and degree of ripeness [432].

There is little information available on the vitamin K content of milk and meat products (Table 17). Apart from phylloquinone, the liver also contains long-chain menaquinones in concentrations several times higher than vitamin K_1. Relatively large amounts of menaquinones, especially of MK-9, are contained in cheese [434].

Table 17. Vitamin K content of milk and meat products [a] [430, 433]

Foods	Vitamin K content, µg/100 g
Cow's milk	0.5 – 1 [b]
Cheese	10 – 50 [c]
Butter	10 [c]
Skeletal meats	<1 – 5 [b]
Liver	20 – 100 [c]

[a] Determined by HPLC.
[b] Phylloquinone.
[c] Phylloquinone and menaquinones.

Menaquinones occur in certain, predominantly gram-positive bacteria. Other bacteria contain ubiquinones (**16**) with chain lengths of 6 – 10 prenyl residues instead of menaquinones. Some bacteria exhibit both quinones, usually with the same chain length. Among the menaquinones, MK-7 to MK-9 predominate. Those found in minor amounts are MK-4 to MK-6, and MK-10 to MK-13 [420]. In humans, some anaerobic intestinal bacteria produce mainly MK-6 to MK-13; MK-8 to MK-10 with two saturated isoprene units in the side chain (**17**) was also detected [435]. The menaquinones are localized in the cytoplasmic membranes of the bacteria. Their concentrations are in the range 0.6 – 1.7 mg/kg of bacterial dry weight [436].

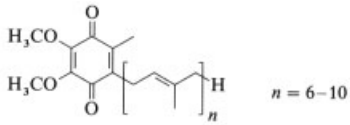

16

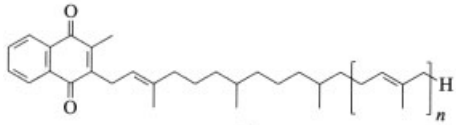

17

$n = 5–7$

5.5. Biosynthesis

Both plants [437] and bacteria [438] use the same biosynthetic pathway for the formation of the naphthoquinone ring of vitamin K (Scheme 16). Shikimic acid (**18**) is made from the precursors erythrose 4-phosphate and phosphoenolpyruvic acid. After phosphorylation and reaction with phosphoenolpyruvic acid, shikimic acid is converted via chorismic acid (**19**) to isochorismic acid (**20**). In a complex reaction of the 2-succinylbenzoic acid synthase system [439], 2-oxoglutaric acid is decarboxylated in the presence of thiamine diphosphate (TPP), forming a succinylsemialdehyde-TPP carbanion (**21**). This substance then reacts with **20** yielding 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylic acid (**22**) after elimination of pyruvic acid. Aromatization of **22** produces 2-succinylbenzoic acid (**23**), which is converted to 1,4-dihydroxy-2-naphthoic acid (**25**) after activation via a coenzyme A ester **24** and a second aromatization step.

In the further biosynthesis of the vitamins K (Scheme 17), e.g., in the formation of bacterial menaquinones, **25** is condensed with polyprenyl diphosphate (**26**) to give 2-polyprenyl-1,4-naphthoquinol (**27**) on decarboxylation [436]. The methyl group of (*S*)-adenosyl-L-methionine is transferred to **27**, producing 2-methyl-3-polyprenyl-1,4-naphthoquinol (**28**), which yields menaquinone-*n* (**2**) on oxidation. The for-

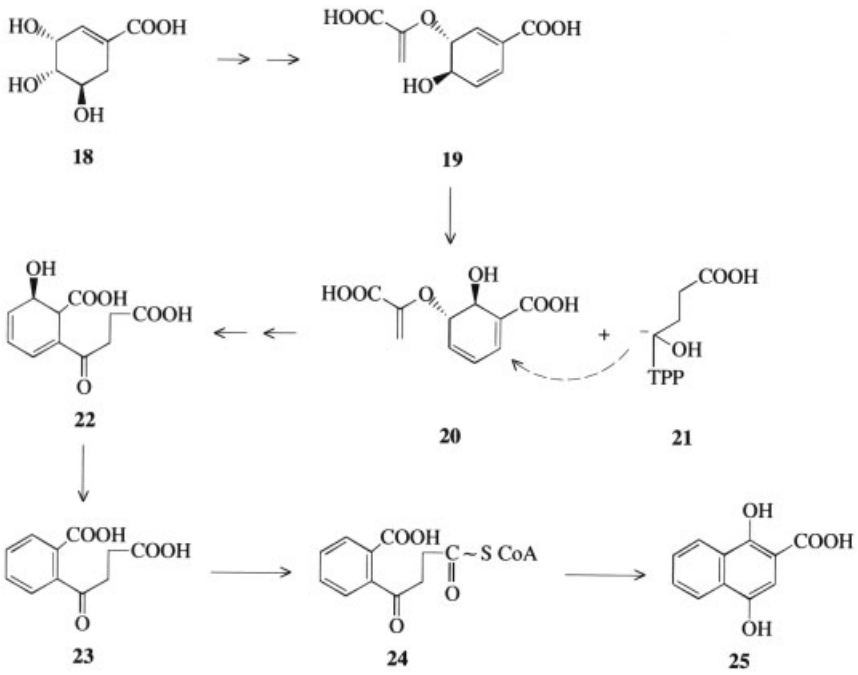

Scheme 16. Biosynthesis of the naphthoquinone ring of vitamin K

mation of phylloquinone (**1**) proceeds analogously. In the same reaction sequence, phytyl diphosphate (**29**) is incorporated instead of the polyprenyl side chain. In the chloroplasts of plants, phytyl diphosphate is made from geranylgeranyl diphosphate by hydrogenation. However, phytylation and methylation of the corresponding vitamin K_1 precursors occur in separate membrane systems of the chloroplasts [440]. The synthesis of polyprenyl disphosphate starts with condensation of 1-isopentenyl diphosphate, which is formed from mevalonic acid, with the isomeric 3,3-dimethylallyl diphosphate to give geranyl diphosphate. Geranyl diphosphate is then extended by further coupling reactions with 1-isopentenyl diphosphate until the appropriate number of isoprene units in the side chain are obtained [420].

5.6. Chemical Synthesis

In 1939, the first syntheses of vitamin K_1 were published almost simultaneously by four groups [441] in connection with the structural elucidation. The starting materials were menadione (**3**) or menadiol (**30**) as the aromatic component and natural phytol (**31**) or one of its derivatives.

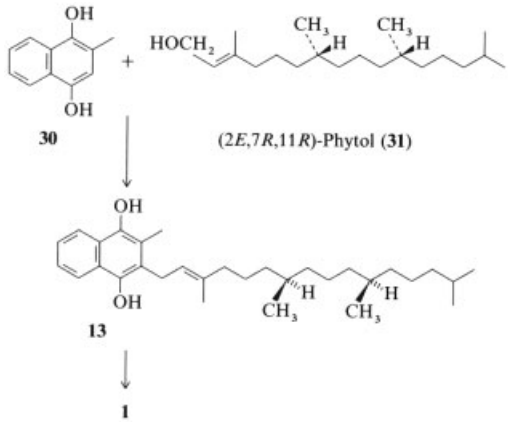

In the Fieser synthesis [443], which was the preferred method for a long time, an excess of menadiol [*481-85-6*] (**30**) was condensed with natural phytol [*150-86-7*] (**31**) in dioxane at 75 °C in the presence of oxalic acid as catalyst. The excess menadiol was subsequently extracted with dilute alkali and the alkylation product **13** was precipitated with petroleum ether.

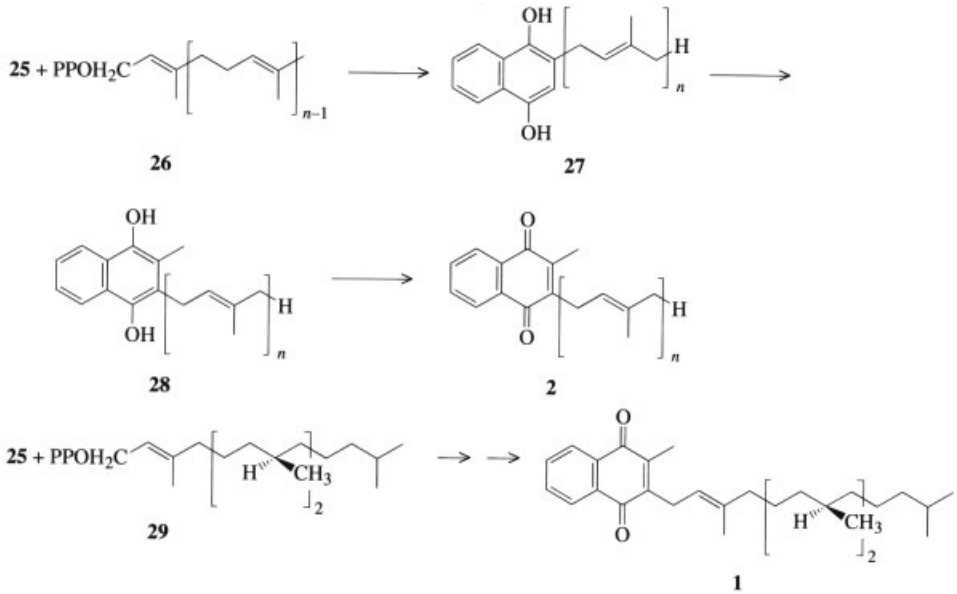

Scheme 17. Biosynthesis of vitamins K

Oxidation of this waxy precipitate with silver oxide gave vitamin K_1 (**1**) in a total yield of ca. 30 % based on phytol (**31**).

A disadvantage of this synthesis is the Friedel – Crafts alkylation, which does not proceed regioselectively with regard to positions 2 and 3 of menadiol. In addition to the desired alkylation product **13**, almost equal amounts of the isomer **32** are produced. Furthermore, the acidic alkylation conditions cause partial isomerization of the trisubstituted double bond in the side chain. Under these conditions, phyllochromanol (**14**), a cyclization product of **13**, and phytadienes of type **33** are also formed.

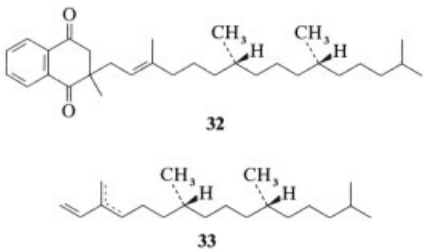

With regard to commercial synthesis, a breakthrough was achieved in the 1950s when Isler et al. [445] at Hoffmann-LaRoche and Hirschmann et al. [446] at Merck found that monoacylated menadiols, e.g., the monoacetate **35a** or the monobenzoate **35b**, could be advantageously used in the alkylation step (Scheme 18). Thus, the undesired alkylation in position 2 of menadiol (**30**) is prevented. Moreover, it was found that boron trifluoride etherate is a much better catalyst than anhydrous oxalic acid and that natural phytol (**31**) can be replaced by isophytol [*505-32-8*] (**36**), which is easy to synthesize.

There are various processes available for the production of menadione (**3**). The most common is the oxidation of 2-methylnaphthalene with, e.g., CrO_3 or H_2O_2 in acetic acid or $Na_2Cr_2O_7$ in sulfuric acid [447].

Starting with menadione (**3**), the monobenzoate **35b** can be produced in three steps: catalytic reduction of menadione (**3**), e.g., in ethanol or methanol, followed by acylation with two equivalents of benzoyl chloride to give the dibenzoate **34b**, and partial saponification, e.g., with KOH or NaOH, to produce the desired monobenzoate **35b**. Menadiol monoacetate (**35a**) can be produced in an analogous manner and used for the synthesis of vitamin K_1 (**38**).

In the Isler – Lindlar method [445], excess menadiol monobenzoate (**35b**) is condensed with isophytol (**36**) in dibutyl ether at 85 °C in

the presence of boron trifluoride etherate as catalyst. Unreacted monobenzoate **35b** is precipitated with petroleum ether, filtered off, and recovered. After evaporation of the filtrate, the alkylation product **37b** is obtained as an *E*/*Z* mixture (*E*/*Z* 70 : 30). The *E* form can be enriched by recrystallization (e.g., from methanol/ethanol; USP requirement: $\leq$20 % of *Z* form in the final product vitamin K_1). **37b** can be obtained in the practically pure *E* form by repeated recrystallization.

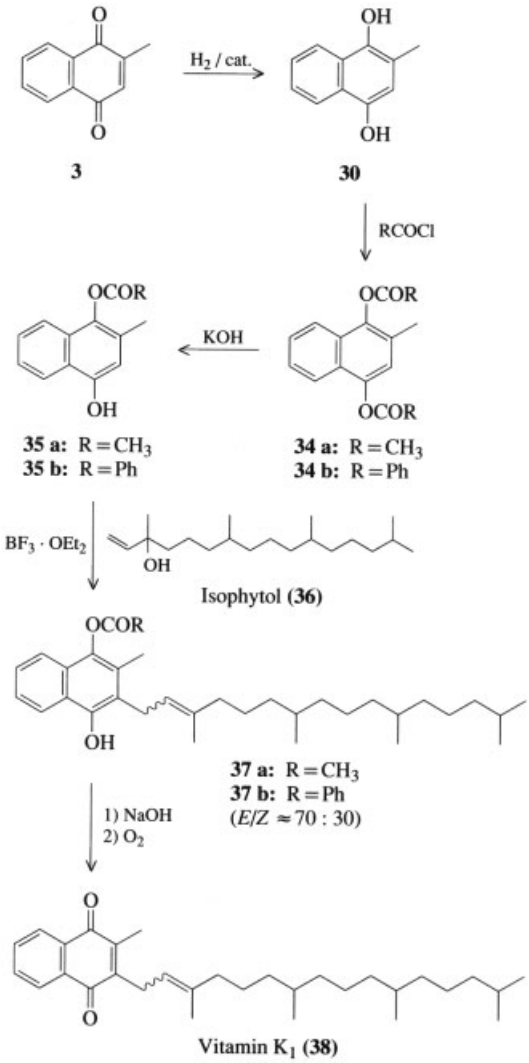

Scheme 18. Synthesis of vitamin K_1

The *E*-enriched alkylation product **37b** (*E*/*Z* 9 : 1) is saponified with potassium hydroxide and oxidized to vitamin K_1 (**38**) with oxygen. In this method, *E*/*Z*-vitamin K_1 (**38**) is produced as a yellow oil in a yield of ca. 60 – 70 % based on isophytol (**36**).

For the synthesis of isophytol, see 4.11.4.

The fact that only *E*-vitamin K_1 exhibits K-activity in biological tests led to intensive work on this synthesis in the 1970s and 1980s, with the aim of synthesizing pure *E*-vitamin K_1.

Many multistage syntheses with organometallic reagents, e.g., Grignard reagents [448]; zinc reagents [449]; aryl [450] and vinyl cuprates [451]; trialkyltin compounds [452]; π-allylnickel [453]; phythaloylcobalt [454]; and chromium carbene complexes [455], have been published and patented. These and other synthetic possibilities are discussed in a review article [451].

In all these variants, a disadvantage is the use of stoichiometric amounts of organometallic reagents which must be recovered for economical and ecological reasons.

A potentially economic, short synthesis of vitamin K_1 and vitamins K_2 has been patented by F. Hoffmann-LaRoche [451,456] (Scheme 19).

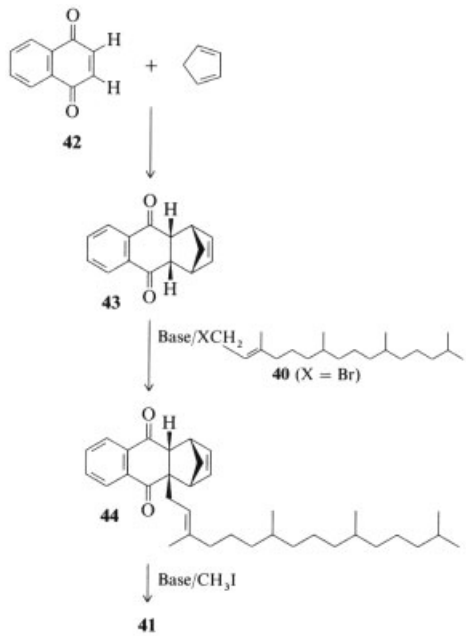

Menadione (**3**) is first subjected to a Diels–Alder reaction with cyclopentadiene, producing the crystalline *endo*-Diels–Alder adduct [*97804-50-7*] **39** in 93 % yield. Subsequently, the proton α to the carbonyl group can be abstracted with a strong base (e.g., potassium *tert*-butoxide) and the resulting enolate alkylated with phytyl bromide (**40**) ($E/Z \geq 98:2$). The alkylation product **41** is obtained in a yield of $\geq$90 %. As a result of the basic alkylation conditions, the *trans* geometry of the double bond in phytyl bromide [*76524-59-9*] (**40**) ($\geq$98 % E) is completely retained during alkylation.

The alkylation product **41** [*97804-51-8*] is thermally unstable and undergoes a retro-Diels–Alder reaction. On heating in toluene (110 °C), it decomposes rapidly into vitamin K_1 (**38**) ($\geq$98 % E) and cyclopentadiene. The yield of vitamin K_1 is practically quantitative. The cyclopentadiene can be recovered by distillation and reused.

The vitamins K_2 (menaquinones) (**2**) and ubiquinones (**16**) can be synthesized according to the same scheme.

A patent application for a similar synthetic route to vitamins K and analogous compounds has been filed by Eisai Chemical Co. [457]. Instead of menadione (**3**), naphthoquinone (**42**) is used as starting material.

Analogously, the isoprenoid side chain [e.g., phytyl bromide (**40**)] and then the methyl groupare introduced via the cyclopentadiene adduct **43**. The intermediate **41** [451,456] described above is obtained via compound **44** and gives the desired quinone [e.g., vitamin K_1 (**38**)] after elimination of cyclopentadiene.

5.7. Analysis

Physicochemical Methods. As a result of its high selectivity and sensitivity, high performance liquid chromatography (HPLC) is the method of choice for the determination of phylloquinone and menaquinones in the blood, tissues, milk, and in foods. Various procedures for extraction and preliminary purification, internal standardization, HPLC conditions (with normal or reversed phase) and possibilities for UV, electrochemical, and fluorescence detection (both after electrochemical or chemical reduction and after photochemical decomposition) of the vitamins K have been described [458]. The detection limits for phylloquinone are in the range 25–500 pg, depending on the detection method employed [459]. Similar values that vary according to the length of the side chain apply to the menaquinones [460]. HPLC methods are also available for the determination of menadione and water-soluble derivatives (see Section 5.12) in feedstuffs, premixes, and in vitamin concentrates [458,461].

Alternative methods are thin layer chromatography (TLC) [462], high performance thin layer chromatography (HPTLC), and gas chromatography (GC) [458]. The spectrophotometric, fluorometric, and colorimetric methods previously used without chromatographic purification of the samples to be analyzed have the disadvantage of low sensitivity and specificity. They are unable to distinguish between phylloquinone and menaquinones [436,462]. The detection limits are ca. 25 µg/mL for phylloquinone and menadione.

Biological methods measure the biological activity of a specific vitamin K compound or detect vitamin K activity in a sample under analysis. It is not possible to differentiate between definite forms of vitamin K in the sample. The biological methods are based on the measurement of the prolonged blood clotting time (prothrombin time) in vitamin K deficient animals (chicks, rats) and its normalization after the administration of the sample exhibiting antihemorrhagic activity [463]. The bioavailability and biopotency of formulated vitamin K compounds are tested mainly in the prophylactic test, i.e., after intermixing with vitamin K deficient feed and feeding for a period of 24 d [464]. On the other hand, the vitamin K activity of pure substances is determined after a single dose in the curative test [465]. The biological determination of vitamin K can also be conducted in animals that have prolonged blood clotting times caused by the feeding of anticoagulants (see Section 5.9.2). Strictly speaking, however, this test does not measure the reversal of vitamin K deficiency, but the competition between the vitamin K compound to be tested and the antagonist used.

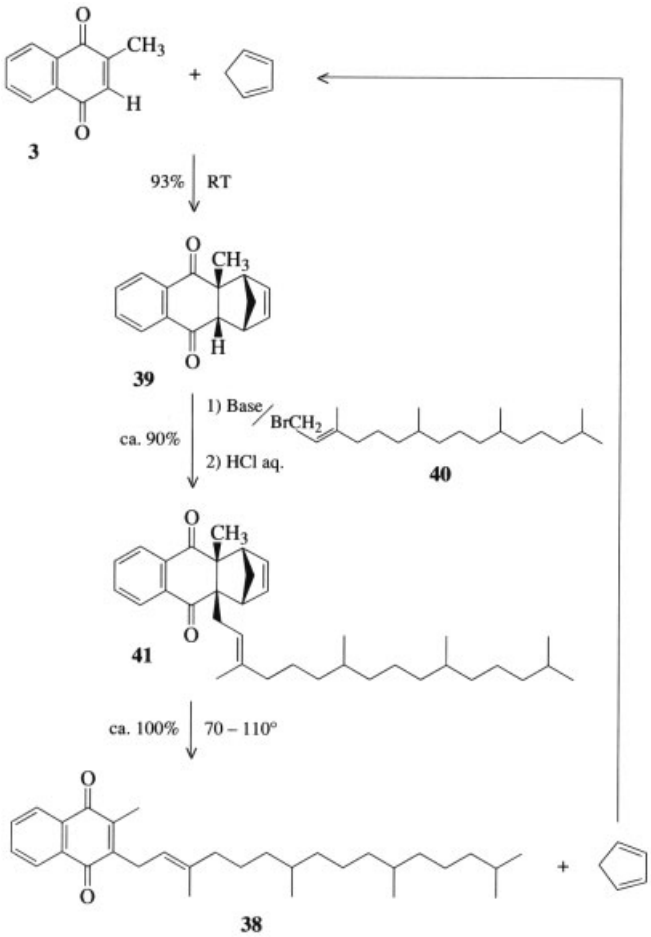

Scheme 19. Synthesis of vitamins K (Hoffmann-LaRoche)

5.8. Metabolism

Alkylation of Menadione. Menadione (vitamin K_3) itself has no antihemorrhagic activity. However, it can be prenylated to active menaquinone-4 in the liver and in other animal tissues [421].

Intestinal bacteria can form menadione by eliminating the phytyl side chain of phylloquinone. After absorption, menadione is converted to MK-4 in the animal organism [433]. It appears possible that dealkylation of phylloquinone and, consequently, conversion of vitamin K_1 to MK-4 can occur in animal tissues as well [466, 467].

Absorption. Phylloquinone and menaquinones present in food are absorbed in the proximal small intestine. Sufficient amounts of bile and pancreatic juice enzymes are required in the intestinal lumen so that the vitamins K can be emulsified in the mixed micelles consisting of bile salts, fatty acids, and monoglycerides, before they are absorbed by the intestinal mucosa cells [468]. A saturable, energy-dependent mechanism that cannot be influenced by the menaquinones seems to be responsible for the absorption of phylloquinone [433]. In the case of the menaquinones produced by intestinal bacteria, absorption occurs in the large intestine by passive diffusion [434]. In the intestinal cells, the absorbed vitamins K are incorporated into chylomicrons and transported to the liver via the lymph. The degree of absorption for vitamin K in the small intestine is 40 – 70 % [469].

The absorption of menadione occurs in the distal small and large intestine by a passive,

energy-independent process. It enters the blood stream directly from the intestine [470].

Distribution. A part of the vitamin K transported to the liver is incorporated into VLDL (very low density lipoproteins), released into the blood stream, and transferred via LDL (low density lipoproteins) to other organs [436] (e.g., lungs, kidneys, adrenals, bone marrow, lymph nodes [421], testicles, and skin [471]). The part remaining in the liver, which is the main site of activity of vitamins K, is mainly taken up by the mitochondria and the microsomes [433].

The distribution of the various forms of vitamin K in the organism varies greatly. Phylloquinone is only moderately retained in the liver. It has a half life of ca. 17 h in rat liver [433]. In human liver, the average concentration of phylloquinone is 5.5 ng/g (range: 1.1 – 21.3 ng/g for 32 individuals tested: $N = 32$) [472]. In contrast to phylloquinone, the more lipophilic menaquinones have a longer retention time or slower turnover in the liver [434], which can lead to a certain accumulation (Table 18). The menaquinones account on an average for 92 mol % of the total vitamin K content of the adult human liver (individual variations of 75 –97 %) [472]. With the exception of horse liver, in which only phylloquinone could be detected, considerable amounts of menaquinones are also present in the liver of other animal species. Bovine liver contains mainly MK-10 to MK-12, while in dog and pig liver MK-8 to MK-10 predominate [436]. Chicken liver contains much more MK-4 than phylloquinone [466]. In rats, relatively high levels of MK-4 are found in the pancreas and salivary gland [467].

Compared to liver, the ratio of phylloquinone to menaquinones in human blood plasma is substantially shifted (Table 18). The menaquinones exceed the amount of phylloquinone only by a factor of 1.2 – 2.3, and very long chain menaquinones (MK-9 to MK-13) are no longer found. This is another indication of a slower hepatic turnover of the highly lipophilic menaquinones and their greater affinity for intracellular membranes. The average phylloquinone concentration in the plasma is 0.485 ± 0.314 ng/mL ($N = 332$) [475].

Breast milk can contain low concentrations of MK-5 to MK-7, but the main form of vitamin K present is phylloquinone [459,476]. Its concentration fluctuates between 1.1 and 6.5 ng/mL (average 2.1 ng/mL). Cow's milk has a much higher phylloquinone content: 10 – 20 ng/mL [436].

In the organism, menadione is distributed over many tissues, but very rapidly eliminated. Only a small part of menadione is converted to menaquinone-4 [421].

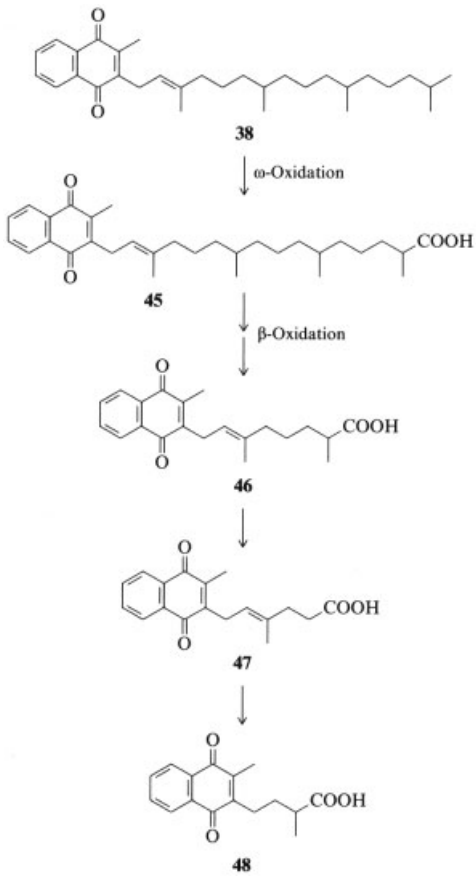

Degradation and Excretion. In studies on the catabolism of phylloquinone (**38**) in humans and laboratory animals, the following metabolites were identified [477]: 2-methyl-3-(7′-carboxy-3′,7′-dimethyl-2′-heptenyl)-1,4-naphtho- quinone (**46**), 2-methyl-3-(5′-carboxy-3′-methyl-2′-pentenyl)-1,4-naphthoquinone (**47**), and 2-methyl-3-(3′-carboxy-3′-methyl-propyl)-1,4-naph-thoquinone (**48**). They are produced from **38** by ω-oxidation to the corresponding carboxylic acid **45**, followed by

Table 18. Concentration of vitamin K in human liver and blood plasma (ng/g liver or ng/mL plasma)

Vitamin K form	Liver, $N = 7$ [473]	Plasma $N = 22$ [473]	Plasma $N = 11$ [474]
K_1	12.6 ± 1.9	0.59 ± 0.06	0.51 ± 0.37
MK-4	0.4 ± 0.2	0.05 ± 0.01	
MK-6		0.05 ± 0.01	
MK-7	21.8 ± 7.8	0.57 ± 0.08	0.29 ± 0.18
MK-8	6.3 ± 1.2	0.08 ± 0.05	0.54 ± 0.32
MK-9	1.5 ± 0.5		
MK-10	62.6 ± 8.0		
MK-11	89.0 ± 13.4		
MK-12	13.3 ± 2.4		
MK-13	5.0 ± 1.4		

repeated β-oxidation of the side chain. After reduction to 1,4-naphthohydroquinone and conjunction with glucuronic acid, the metabolites are excreted in the urine as mono- or diglucuronides. A part of the less polar

metabolites is also excreted in the conjugated form via the bile and feces [478]. Menaquinones (like ubiquinones (**16**) and tocopherols) are degraded in the same manner.

Menadione is also excreted in the urine in the conjugated form as the phosphate, sulfate, or glucuronide of menadiol (**30**).

5.9. Importance for the Organism

5.9.1. Function

Vitamin K plays a vital role as a cofactor in the γ-carboxylation of specific glutamic acid residues near the amino terminal end of seven protein precursors of the blood clotting system (see also → Blood, Chap. 2.2.2.) [479]. This post-translational modification of inactive preproteins produces active coagulation proteins (Table 19) with 10 – 13 γ-carboxyglutamic acid residues per molecule. In the presence of calcium ions, these residues allow the proteins to bind to phospholipid membrane surfaces [436]. The classical vitamin K dependent factors II, VII, IX, X, and protein C can thereby perform their specific functions (Table 19) within the complex blood clotting system. The overall function of this system is maintenance of normal hemostasis, i.e., the absence of hemorrhagic or thrombic events. In principle the reactions of the blood clotting system proceed in such a manner that in conjunction with additional components of an activation complex, an activated clotting factor converts an inactive clotting factor to an active proteolytic enzyme. This enzyme, in turn, acts as a proteinase in the next step of the blood clotting cascade. The last activation step involves the conversion of prothrombin (factor II) to thrombin, which finally causes the formation of fibrin from its precursor fibrinogen, producing a stable fibrin clot. The coagulation proteins are synthesized in the liver and then released into the blood. In vitamin K deficiency, the blood clotting proteins are produced, but they remain inactive because of insufficient or lacking γ-carboxylation of the glutamic acid residues [469].

The form of vitamin K that is "active" as a cofactor for the γ-glutamyl carboxylase located in the microsomes (E1 in Scheme 20) is 1,4-dihydrovitamin K (**50**). Simultaneous with the carboxylation of a protein-bound glutamyl residue (**51**) to a γ-carboxyglutamyl residue (**52**), 1,4-dihydrovitamin K is oxidized to the 2,3-epoxide of vitamin K (**53**) [481]. It is assumed that carboxylation and epoxidation are catalyzed by the same enzyme [434]. The epoxide **53** is then reduced to vitamin K (**49**) by a dithiol-dependent vitamin K epoxide reductase (E2). Another dithiol-dependent vitamin K reductase (E3) reforms vitamin K hydroquinone (**50**). It is possible that the reduction (E3) of **49** to **50** is also carried out by vitamin K epoxide reductase (E2). However, vitamin K can also be reduced to the quinol **50** by a NAD(P)H-dependent vitamin K reductase (E4). The activity of the dithiol-dependent reductases (E2 and E3) is inhibited by anticoagulants (A) (see Section 5.9.2).

Table 19. Vitamin K dependent proteins of the blood clotting system [480]

Protein	Function	Molecular mass	Concentration in human blood, ng/mL
Factor VII (proconvertin)	activation of factor X to X a	48 000	0.6
Factor IX (PTC ∗)	activation of factor X to X a	57 000	5
Factor X (Stuart – Prower Factor)	as factor X a: conversion of factor II to thrombin	55 000	10
Factor II (prothrombin)	precursor of thrombin	72 000	100
Protein C	inhibition of coagulation	62 000	4 – 5
Protein S	cofactor of protein C	80 000	25
Protein Z	still unknown	55 000	<1 ∗∗

∗ Plasma thromboplastin component.
∗∗ According to [436].

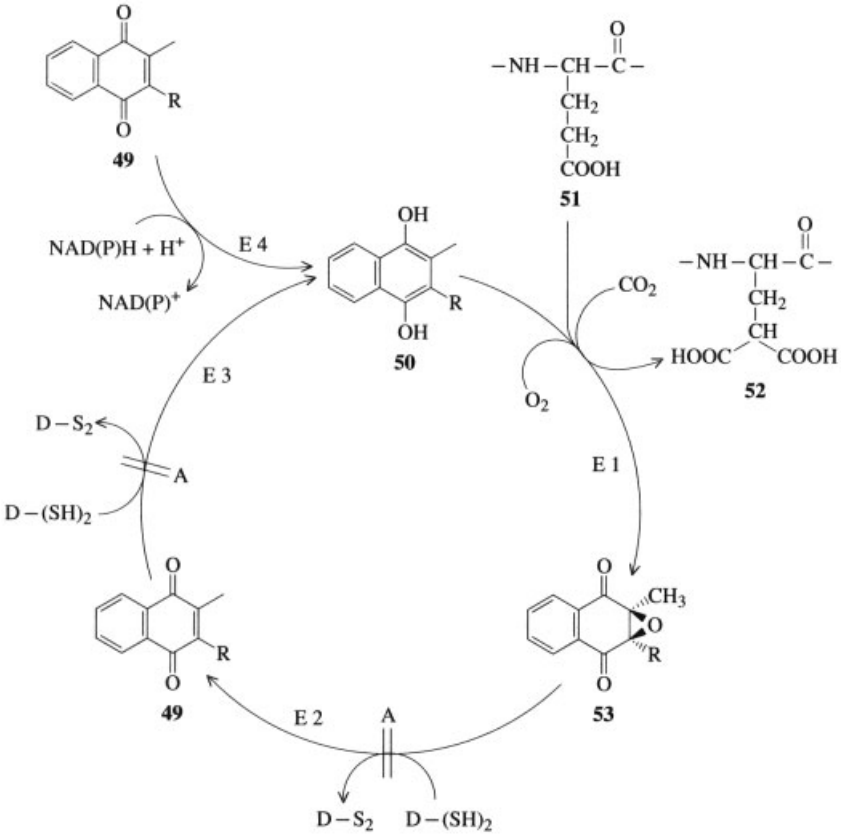

Scheme 20. Reactions of the vitamin K cycle
R = phytyl (phylloquinone) or geranylgeranyl (MK-4); 1E1 = γ-glutamyl carboxylase; E2 = vitamin K epoxide reductase; E3 = dithiol-dependent vitamin K reductase; E4 = NAD(P)H-dependent vitamin K reductase; D-$(SH)_2$ = reduced dithiol; D-S_2 = oxidized dithiol (functional dithiols: dithiothreitol and thioredoxin); NAD(P)H = reduced nicotinamide adenine dinucleotide (phosphate); $NAD(P)^+$ = oxidized nicotinamide adenine dinucleotide (phosphate); A = inhibition by anticoagulants

This vitamin K (or vitamin K epoxide) cycle acts very effectively. It is assumed that one molecule of vitamin K is recycled several thousand times before it is converted to inactive metabolites [482]. Approximately 10 % of the phylloquinone in the liver is present as phylloquinone-2,3-epoxide [483]. Menaquinones, especially MK-3 and MK-4, are also active in the vitamin K cycle. However, their activity decreases with increasing length of the side chain (see Section 5.9.3). Menadione cannot replace the vitamins K.

After the elucidation of the cofactor role of vitamin K in the formation of γ-carboxyglutamic acid (Gla) in the vitamin K dependent blood clotting factors in the liver, it was observed that the vitamin K cycle is active in other tissues as well (e.g., bone, cartilage, kidney, spleen, and lung). In fact, various organs form Gla-containing proteins which are not involved in blood clotting [434, 484]. Little is known about the function of these proteins. Until now, the bone Gla-protein (osteocalcin) and the matrix Gla-protein have been relatively well studied. Osteocalcin is produced by the osteoblasts of bone tissue, approximately 30 % of this protein being released into the blood [482]. Osteocalcin, which is subject to regulation by 1,25-dihydroxyvitamin D_3, is quantitatively the most common noncollagenic protein component of the bone matrix and accounts for ca. 2 % of the total proteins of the bone. In addition, it is quantitatively the most common vitamin K dependent protein in the body [436]. It is assumed that the Gla-containing part of osteocalcin plays an important role in mineralization or demineralization of the bone. The matrix Gla-protein, which is formed in a series of organs (especially heart,lung, kidney, and cartilage tissue), is enriched above all in the bone and cartilage. It probably has a mediator function in cell growth and differentiation [485].

These examples show that vitamin K not only plays a specific role in blood clotting, but this vitamin apparently has a wider biological significance (see Section 5.10.1) which is connected with the strong calcium-binding properties of the negatively charged Gla residues in vitamin K dependent proteins.

In plants, phylloquinone acts as an electron carrier in the chloroplasts. In the plasma membrane of bacteria, menaquinones are involved as redox substances in electron transport and oxidative phosphorylation [486].

5.9.2. Antagonists and Anticoagulants

Around 1930, a hemorrhagic disease was observed in cattle in Canada and the United States and was put down to spoilt sweet clover hay fodder. In 1941, the substance responsible for this disease was identified by isolation and structural elucidation as 3,3′-methylene-bis-(4-hydroxycoumarin), usually called dicumarol (**54**) [433]. This led to the synthesis of other derivatives of 4-hydroxycoumarin which can be used in the treatment of thromboembolic diseases and for the prevention of thrombus formation in vessels. Active oral anticoagulants are, e.g., phenprocumarol [3-(1′-phenylpropyl)-4-hydroxycoumarin] (**55**), warfarin [3-(α-acetonylbenzyl)-4-hydroxycoumarin] (**56**), acenocumarol [3-(α-acetonyl-4-nitrobenzyl)-4-hydroxycoumarin] (**57**), and ethyl-biscoumacetate [3,3′-carboxymethylene-bis-(4-hydroxycoumarin) ethyl ester] (**58**) [487].

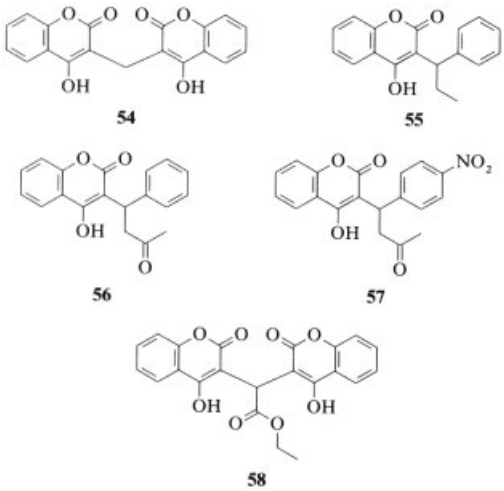

The action of all oral anticoagulants is based on the inhibition (A in Scheme 20) of vitamin K epoxide reductase (E2) and of dithiol-dependent vitamin K reductase (E3) in the vitamin K cycle. This leads to an increase in the concentration of vitamin K 2,3-epoxide (**53**) in the liver, a decrease in the activity of the blood clotting proteins due to lacking γ-carboxyglutamic acid formation and, consequently, to an increase in the blood clotting time [434]. Since oral anticoagulants inhibit the dithiol-dependent vitamin K reductase (E3) but not the NAD(P)H-dependent reductase (E4), inhibition of the vitamin K cycle can be avoided by large doses of phylloquinone. Under these conditions, however, the 1,4-dihydrophylloquinone (**50**) formed can only act to a limited extent because the recycling of vitamin K_1 epoxide is inhibited or blocked. The result is a relatively high vitamin K epoxide level in the blood. Oral anticoagulants interfere not

only with the vitamin K cycle in the liver, but also in other organs. A decrease of at least 50 % in the blood level of osteocalcin and a reduction of the γ-carboxyglutamic acid content of osteocalcin by 50 % have been observed under therapy conditions [482].

More effective, extremely lipophilic derivatives of 4-hydroxycoumarin have now been synthesized for use as rodenticides (→ Rodenticides, Chap. 5.3.). Examples are difenacoum (**59a**) and brodifacoum (**59b**). These substances have a very long biological half life of 100 – 200 d. They are potentially harmful to humans [434].

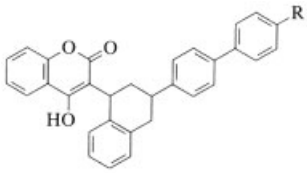

59a: R = H
59b: R = Br

A competitive inhibitor of vitamin K is 2-chloro-3-phytyl-1,4-naphthoquinone (**60**), also known as chloro-K. It inhibits the γ-glutamyl carboxylase reaction (E1 in Scheme 20) by preventing epoxidation of vitamin K. 2,3,5,6-Tetrachloro-4-pyridinol (**61**) which is structurally related neither to vitamin K nor to 4-hydroxycoumarin, inhibits the carboxylase apparently as a direct antagonist of the vitamin [433, 484].

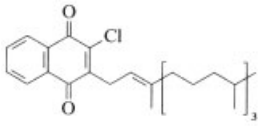

60

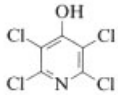

61

5.9.3. Relative Activity

Studies of the biological activity of vitamin K clearly show that two structural elements are required: a 2-methyl-3-substituted 1,4-naphthoquinone ring which can be reduced to hydroquinone, and a side chain at the C-3 atom which guarantees binding to or interaction with microsomal membranes.

Totally synthetic phylloquinone (**38**) consists of an equimolar mixture of four stereoisomers ((**62a**), (**62b**),(**62c**), (**62d**)). In curative prothrombin-time chick assays (see Section 5.7), no substantial differences were found in their antihemorrhagic activity [488] (Table 20).

Table 20. Antihemorrhagic activity of the stereoisomers of phylloquinone and of (E, all-rac)-phylloquinone [488]

Compound	Configuration	Activity, %	Confidencelimits
62a	2′*E*,7′*R*,11′*R* (standard)	100	88 – 116
62b	2′*E*,7′*R*,11′*S*	93	75 – 115
62c	2′*E*,7′*S*,11′*S*	119	110 – 128
62d	2′*E*,7′*S*,11′*R*	99	88 – 111
38	2′*E*,7′*RS*,11′*RS* (*E*, all-*rac*)	111	100 – 122

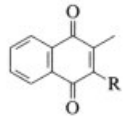

62

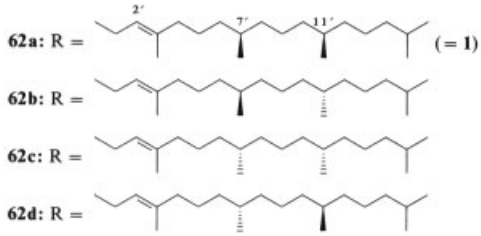

Although **62c** and (*E*, all-*rac*)-phylloquinone (**38**) have a slightly higher biological activity than natural phylloquinone (**62a = 1**), all four stereoisomers and **38** are to be considered identical with regard to biological activity within the statistical confidence limits [488]. Totally synthetic commercial phylloquinone always contains a low proportion (up to 10 %) of (*Z*, all-*rac*)-phylloquinone (**65**). However, the antihemorrhagic activity of **65** is only 0.3 – 1 % (Table 21) and thus, negligible [489].

Phylloquinone-2,3-epoxide (**12**), the main component of the vitamin K cycle (see Section 5.9.1), is 1.66 times as active as phylloquinone (Table 21). Changes in the side chain, e.g., 6′-hydroxylation (**63**) or hydrogenation of the double bond of the side chain (**64**), result in a great decrease in the biological activity. If, in addition, the naphthoquinone ring

Table 21. Antihemorrhagic activity of structural analogs of phylloquinone (K_1) [465,489]

Compound	Structure	Activity, %	Confidence limits
38	(2′*E*,7′*RS*,11′*RS*)-K_1 (standard)	100	87.8 – 116.7
12	(2′*E*,7′*RS*,11′*RS*)-K_1 2,3-epoxide	166	137.3 – 200.2
63	(2′*E*,7′*RS*,11′*RS*)-6′-hydroxy-K_1	20.5	16.1 – 26.0
64	(3′*RS*,7′*RS*,11′*RS*)-2′,3′-dihydro-K_1	6.7	5.1 – 8.8
65	(2′*Z*,7′*RS*,11′*RS*)-K_1	<1	

is also partially hydrogenated to 2′,3′,5,6,7,8-hexahydrophylloquinone (**66**), the biological activity is totally lost [465]. Substitution of the phytyl side chain of phylloquinone by an unbranched alkyl side chain (e.g., 2-methyl-3-octadecyl-1,4-naphthoquinone) also causes a major loss of biological activity [420,421].

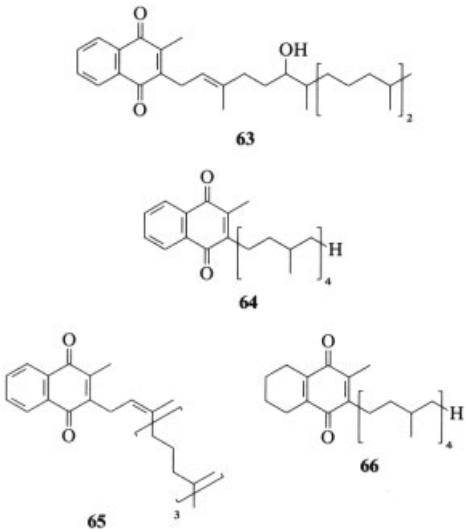

The length of the isoprenoid side chain is also of importance. In the prothrombin-time assay in vitamin K deficient chicks (Table 22), comparisons of the antihemorrhagic activity of isoprenologs of phylloquinone and menaquinones (on a molar basis) showed that a C_{20} side chain (four isoprene units) is optimal for the phylloquinone series. In the case of the menaquinones, MK-5 is 20 % more active than phylloquinone, while MK-4 and MK-6 have the same activity as vitamin K_1. The biological activity of the menaquinones decreases with increasing length of the side chain [487]. Kinetic studies on the cofactor activity of the menaquinones, measured as the reaction rate of partially purified γ-glutamic acid carboxylase at saturating vitamin concentrations, confirm that the activity depends on the length of the side chain (Table 23). In this study MK-3 had optimal activity, MK-2 and MK-4 were almost as active as phylloquinone, while the activity steadily decreases from MK-5 to MK-10 [490].

Table 22. Antihemorrhagic activity of isoprenologs of phylloquinone and menaquinones [487]

Phylloquinone series		Menaquinone series	
Isoprene units in the side chain	Activity, %	MK-*n*	Activity, %
2	10	2	15
3	30	3	40
4	100 ∗	4	100
5	80	5	>120
6	50	6	100
		7	70
		8	68
		9	60
		10	25

∗ Standard: (2′*E*,7′*RS*,11′*RS*)-phylloquinone.

Table 23. Activity of menaquinones as a cofactor of vitamin K dependent γ-glutamyl carboxylase from liver. Measurement of the reaction rate at saturating vitamin concentrations (V_{sat}) [490]

Vitamin	V_{sat}, µmol bound CO_2/h	V_{sat}, %
Phylloquinone	25.4	100.0
MK-1	4.0	15.9
MK-2	24.9	98.1
MK-3	30.5	120.0
MK-4	24.7	96.7
MK-5	22.2	87.1
MK-6	19.6	77.0
MK-7	18.0	70.7
MK-8	15.8	62.0
MK-9	14.0	55.8
MK-10	13.0	51.2

5.10. Deficiency Symptoms

5.10.1. Causes

Vitamin K deficiency leads to the secretion of incompletely carboxylated, biologically less active or inactive forms of the blood clotting factors II, VII, IX, and X (see Section 5.9.1) and, consequently, to subcutaneous or organ bleeding. Vitamin K deficiency can result from inadequate dietary supply or from diseases which interfere with the absorption of the vitamin in the intestine. Examples are pancreatic insufficiency and biliary obstruction (lack of pancreatic enzymes and bile salts for the solubilization of vitamin K, see Section 5.8), intestinal diseases (malabsorption of vitamins), or parenteral nutrition (if not supplemented with vitamin K) [468]. In liver diseases, the protein synthesis of the blood clotting factors can be reduced so that in spite of an adequate supply of vitamin K, there can be a quantitative lack of circulating blood clotting factors [480]. As in true vitamin K deficiency, therapeutic treatment of patients with anticoagulants produces a qualitative lack of the blood clotting factors II, VII, IX, and X (see Section 5.9.2). The effect of anticoagulants can be reversed by high doses of vitamin K. Whether a decrease in the synthesis of menaquinones in intestinal bacteria due to antibiotics or sulfonamides can cause vitamin K deficiency is still unknown (see Section 5.11).

In healthy adults, the risk of vitamin K deficiency is very low under normal nutritional conditions. Newborn infants, however, have a relatively high risk of developing hemorrhagic disease of the newborn (HDN). If newborn infants are not prophylactically treated with vitamin K, 2.5 – 5 per 1000 infants suffer from HDN [491]. This disease is manifested in bleeding (e.g., in the gastrointestinal tract), which can appear on the first day (early HDN) or in the first week after birth (classical HDN). In late HDN (2nd to 10th week), dangerous and often fatal intracranial bleeding occurs [434]. The cause of HDN is a lack of blood clotting factors in the blood of the newborn as a result of an inadequate supply of vitamin K. This is due to the low permeability of the placenta for vitamin K, limiting the transport of the vitamin from the blood of the mother to the fetus. The lack of an intestinal flora, i.e., absence of menaquinone production in the newborn during the first days of life, can also play a part [471]. Another serious factor is that breast milk is a relatively poor source of vitamin K. Consequently, newborns can cover only about 20 % of their daily vitamin K requirement [469]. The requirement of newborn infants is relatively high because vitamin K is not only needed in the liver for the formation of active blood clotting factors, but rapidly growing bones also require a considerable amount of vitamin K for the synthesis of active osteocalcin [471]. In many countries, newborns are prophylactically given 1 mg of phylloquinone at birth to prevent HDN [492].

It is possible that postmenopausal women also bear a risk of becoming deficient in vitamin K. In patients suffering from osteoporosis, a decrease in the plasma concentration of phylloquinone of up to 30 % of the normal value [493] and insufficient γ-glutamyl carboxylation of the circulating osteocalcin at normal prothrombin values have been observed [482]. Administration of vitamin K to these patients increased not only the plasma level of osteocalcin and the content of γ-carboxyglutamic acid, but also reduced by 50 % the previously increased excretion of calcium in the urine [471]. However, the physiological significance of these results and the association between dietary or subclinical vitamin K deficiency and the pathogenesis of osteoporosis must be further investigated [494].

With regard to animal nutrition, in particular pigs deficient in vitamin K suffer from subcutaneous hemorrhage [495]. Shortly after weaning, young piglets can develop severe internal and external bleeding, a condition called porcine hemorrhagic syndrome. The causes are an insufficient vitamin K content of the feed and sulfonamide/antibiotic medication [496].

5.10.2. Evaluation of Vitamin K Status

In principle, two methods are available for the evaluation of vitamin K status in humans and animals: the quantitative determination of vitamin K in the blood or organs by HPLC (see Section 5.7) or the mostly indirect detection of prothrombin or other vitamin K dependent proteins in the plasma.

The general blood clotting tests for the determination of the prothrombin time [480] have a low sensitivity. They are either more sensi-

tive to the concentration of factor VII than to prothrombin or they do not respond to a mild vitamin K deficiency [433, 482]. For this reason, it is better to determine under-γ-carboxylated prothrombin, i.e., prothrombin containing little or no γ-carboxyglutamic acid residues (often called PIVKA, prothrombin induced by vitamin K absence or antagonists). The prothrombin time is measured in one sample after activation of prothrombin with thromboplastin (only completely carboxylated prothrombin is activated) and in another sample, after activation of prothrombin with a certain snake venom (which activates both normal and under-γ-carboxylated prothrombin, i.e., total prothrombin). The ratio of the two test results is normally ca. 1, but <1 under conditions of vitamin K deficiency [497, 498]. PIVKA can also be directly determined cross-immuno-electrophoretically or by using monoclonal or monospecific antibodies [498]. There are, however, two limitations on the detection of PIVKA: abnormal prothrombin is a very heterogeneous molecule and, therefore, it is unknown to what extent various forms of PIVKA react in the different methods. In addition, abnormal prothrombin is also found in the plasma of patients suffering from hepatic cancer or other liver diseases [499]. Parallel determinations of the vitamin K concentration in the blood and of intact and/or abnormal prothrombin may be necessary to evaluate the vitamin K status more accurately.

In vitamin K deficiency, under-γ-carboxylated osteocalcin has been detected in the plasma by means of a radio-immunological method [498]. However, further studies are necessary to clarify whether this parameter is suitable for the evaluation of vitamin K status or can even be used as an early indicator of vitamin K deficiency. This also applies to the excretion of γ-carboxyglutamic acid (Gla), which results from the degradation of vitamin K dependent proteins and could serve as an indicator for the assessment of vitamin K deficiency. In combination with the determination of the plasma vitamin K_1 level, the measurement of Gla in the urine appears to be useful [475].

5.11. Requirement

The daily vitamin K requirement of a healthy human is generally estimated at 1 – 2 µg/kg of body weight [500]. In 1989, the U.S. National Research Council first published recommended dietary allowances for vitamin K of 1 µg per kg of body weight per day (Table 24) [469]. For babies, the recommended daily dose is 5 µg of vitamin K for the first six months and 10 µg for the second six months on condition that newborn infants are prophylactically given at least 1 mg of vitamin K at birth to prevent bleeding (see Section 5.10.1). For pregnant or nursing women, an increase in the vitamin K requirement is not considered necessary [469].

Table 24. Daily vitamin K intake recommended (1989) by the U.S. National Research Council for man [469]

Category	Age, a	Weight, kg	Vitamin K, µg
Small children	0 – 0.5	6	5
	0.5 – 1	9	10
Children	1 – 3	13	15
	4 – 6	20	20
	7 – 10	28	30
Young males and men	11 – 14	45	45
	15 – 18	66	65
	19 – 24	72	70
	25 – 50	79	80
	>51	77	80
Young females and women	11 – 14	46	45
	15 – 18	55	55
	19 – 24	58	60
	25 – 50	63	65
	>51	65	65

The amount of vitamin K in the diet consumed daily by a healthy American adult is estimated to be 300 – 500 µg. This estimate is probably too high [469]. In general, however, it can be assumed that the quantity of vitamin K in the daily diet is fully sufficient to cover the requirement under normal nutritional conditions. However, it is doubtful whether the earlier assumption that the menaquinones formed by intestinal bacteria contribute partially or completely to meeting the human vitamin K requirement is correct. Studies on volunteers showed that a restriction of the phylloquinone supply to 2 – 5 µg/d for a period of two weeks caused a significant decrease in the blood clotting factor VII and an increase in the concentration of insufficiently γ-carboxylated (abnormal) prothrombin in the plasma (see Section 5.10.2), although

the relatively insensitive measurements of the prothrombin time remained in the normal range [501]. In a second study on students, a reduction by 50 % of the daily supply of vitamin K_1 (from 82 to 32 – 40 µg/d) for a period of three weeks resulted in a significant decrease in the serum level of phylloquinone, formation of abnormal prothrombin, and a significant reduction in the excretion of γ-carboxyglutamic acid in the urine [497]. These results show that dietary phylloquinone is very important for humans and that long-chain bacterial menaquinones are not able to prevent the appearance of vitamin K deficiency symptoms. The reasons for this include the highly lipophilic properties, lower bioavailability [499], slower turnover, and lower biological activity of long-chain menaquinones (see Section 5.9.3).

The vitamin K requirement of some animal species is given in Table 25. The high requirement of poultry is remarkable and was attributed to insufficient absorption of the vitamin from their short digestive tract. However, more recent results have shown that the activity of vitamin K epoxide reductase (see also Section 5.9.1) in chicken liver is ten times lower than in the rat. This results in high concentrations of vitamin K 2,3-epoxide in the liver and serum of the chicken [466]. In pigs, an increase in the calcium content of the feed results in a higher vitamin K requirement [502], indicating that there is a relationship between calcium homeostasis and vitamin K supply.

Table 25. Vitamin K requirement of various animal species [433]

Species	Requirement, µg/kg body weight per day	Amount in feed, µg/kg
Dog	1.25	60
Pig	5	50 ∗
Rat (male)	11 – 16	100 – 150
Chicken	80 – 120	530
Turkey	180 – 270	1200

∗ Other authors recommend a concentration of 500 –2000 µg vitamin K_3/kg feed [495, 496].

5.12. Application

Phylloquinone is the preferred form of vitamin K for clinical application. The main indications are:

1) Prophylactic treatment of newborn infants to prevent hemorrhagic disease of the newborn
2) Treatment of vitamin K deficiency caused by disturbances of absorption in various diseases (e.g., pancreatic insufficiency, biliary obstruction, lipid malabsorption), insufficient alimentary supply, or parenteral nutrition
3) Countertreatment of anticoagulant therapy
4) Prophylactic administration before operations in which bleeding can become a problem (e.g., gallbladder operations)

In Japan, menaquinone-4 (menatetrenone) is used for the prevention of hemorrhagic disease of the newborn [490].

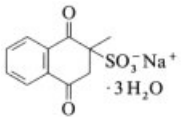

67

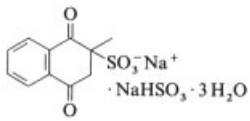

68

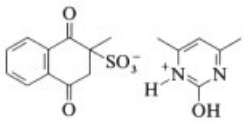

69

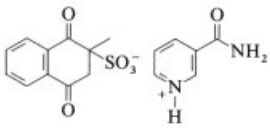

70

In animal nutrition (pigs and especially poultry), water-soluble forms of menadione are mainly used to cover the vitamin K requirement because phylloquinone is too expensive for this application. Menadione itself is less suitable for this purpose because it is instable in feed and its absorption in the intestine depends on the presence of lipids in the feed [433]. The most common water-soluble forms of menadione (**3**) are menadione sodium bisulfite (MSB) (**67**); menadione sodium bisulfite complex (MSBC) (**68**); menadione dimethylpyrimidinol bisulfite, known as menadione pyrimidinol bisulfite (MPB) (**69**) [495]; and menadione nicotinamide bisulfite (MNB) (**70**) [503]. MSBC (**68**) is still more stable than MSB (**67**) and, therefore, is widely used in poul-

try feeding. In chickens, all these compounds exhibit approximately the same biological activity on a molar basis as phylloquinone [504]. Menadione is readily converted to menaquinone-4 in the body of the chicken.

5.13. Tolerance

Toxic effects of phylloquinone and menaquinones are so far unknown. Very rare cases of cardiopulmonary reactions observed on intravenous administration of phylloquinone were due to the use of auxiliary agents for the emulsification of the vitamin [470]. Intramuscular injections of water-miscible formulations of phylloquinone are unproblematic [495].

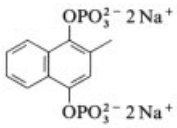

71

Today, parenteral administration of menadione or of water-soluble forms such as menadiol sodium diphosphate (**71**) is no longer employed. Menadione can react with free sulfhydryl groups in various tissues, causing hemolytic anemias, hyperbilirubinemia, and kernicterus [433]. This side effect of menadione is not connected with its conversion to menaquinone-4 in the organism, but is due to its chemical properties as a quinone. For this reason, phylloquinone (or menaquinone-4 in Japan) is used exclusively today in the medical treatment of hemorrhagic diseases and in the prophylactic treatment of newborn infants.

In animal nutrition, menadione and water-soluble forms are well tolerated when administered orally within the dosage range. While the LD_{50} for a single parenteral application of menadione is 75 – 200 mg/kg of body weight in the chicken, mouse, or rat, the LD_{50} for a single oral dose is 600 – 800 mg/kg of body weight. This means that side effects would only be expected if the amount of menadione ingested daily with the feed exceeds the daily requirement by a factor of 1000 [504].

5.14. Trade Names and Economic Aspects

Phylloquinone is available for parenteral application in the form of colloidal suspensions, emulsions, and aqueous suspensions and for oral application in the form of tablets and mixed micelle solutions. In Japan, preparations containing menaquinone-4 are also available. The products sold are, e.g., Konakion, Konakion MM, Mephyton, Aquamephyton, and Kay-two (menatetrenone).

World production of phylloquinone is estimated at 3000 – 3500 kg/a. The main producers are Roche, Merck, Eisai, and Nisshin Chemical.

Water-soluble forms of menadione are mainly produced in feed-grade quality as feed additive for poultry feed and other animal feed (e.g., for pigs). The worldwide demand for menadione derivatives is ca. 500 t/a. The main producers are Vanetta and Heterochemical.

6. Vitamin B_1 (Thiamin)

6.1. Structure and Nomenclature

Vitamin B_1 (**1a**) has the IUPAC – IUB name thiamin [*59-43-8*], although the term thiamine is used in many official and commercial documents.

The basic skeleton of thiamin contains a thiazole ring and a pyrimidine ring. The exact chemical nomenclature is 3-[(4-amino-2-methylpyrimidin-5-yl)methyl]-5-(2-hydroxyethyl)-4-methylthiazolium chloride, $C_{12}H_{17}N_4OSCl$.

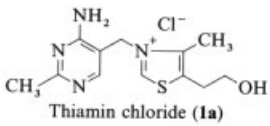

Thiamin chloride (**1a**)

Thiamin is used chiefly in the form of its chloride hydrochloride (**1b**) [*67-03-8*] and nitrate (**1c**) [*532-43-4*].

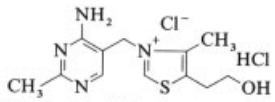

Thiamin chloride hydrochloride (**1b**)

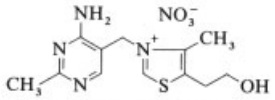

Thiamin nitrate (**1c**)

6.2. History

The history of the discovery of thiamin is closely associated with beriberi, a disease prevalent in many parts of the world where the diet consisted mainly of milled or polished rice. The first "cure" was obtained by TAKAKI, then surgeon general of the Japanese Navy, who suggested that the patients should reduce their carbon/nitrogen intake ratio by increasing their protein intake with a "Western-style" diet [506]. In the 1890s the Dutch physician EIJKMAN noticed by chance that hens kept on a diet of dehusked rice developed a disease very similar to human beriberi [507], and that adding rice bran to the feed could prevent it. He thus postulated that beriberi was caused not by a pathogen or by noxious agents in the food, but by the lack of a vitally important food constituent. However, this conclusion was firmly established only in 1901 by GRIJN who showed that this "polyneuritis preventive factor" was destroyed in meat or bran by heat [508].

GRIJN extracted the factor from rice bran, showing that it was water soluble. The pure factor was isolated from rice bran by FUNK, who obtained a crystalline substance that prevented beriberi [509]. Having found that the substance had the character of an amine, he coined the term "vitamin" (vitally important amine). Pure crystals of thiamin hydrochloride were isolated from rice bran extracts in 1926 by Dutch chemists JANSEN and DONATH [510] who, however, omitted the sulfur atom in the formula they published, delaying elucidation of the chemical structure of vitamin B_1 [511]. This was achieved in 1936, independently by WILLIAMS [512,513] and GREWE [514], who discovered that the molecule could be split quantitatively by sulfite into a 6-aminopyrimidinesulfonic acid (**2b**) and a thiazole ring (**3a**) (see section 6.4).

Confirmation of the structure of thiamin was obtained by WILLIAMS who accomplished the first total synthesis by building up each ring independently and then linking the two together [515]. The synthetic product was shown to be identical to the natural compound by its UV spectrum and biological tests on microorganisms and animals. In 1937 TODD and BERGEL synthesized thiamin by a different approach [516,517], soon followed by ANDERSAG and WESTPHAL [518].

Industrial production was started in 1937 by Hoffmann-La Roche in Switzerland and Merck in the United States. Demand and production rose rapidly.

6.3. Physical Properties

Thiamin chloride hydrochloride (**1b**) $C_{12}H_{18}N_4OSCl_2$ is a white crystalline material, *mp* 248 – 250 °C (decomp.). It crystallizes as colorless monoclinic needles. It has a characteristic odor and a slightly bitter taste. It is soluble in water (1 g/mL) and glycerin (1 g/18 mL), sparingly soluble in alcohol or acetone, and almost insoluble in other organic solvents. On exposure to air of average humidity, dry thiamin is converted to the hydrate by water absorption (up to 4 wt %). The water can be removed by heating at 100 °C, or under vacuum over conventional drying agents. If protected from light and humidity, thiamin is relatively stable against air oxidation, even on heating.

Thiamin nitrate (**1c**) $C_{12}H_{17}N_5O_4S$ is a white crystalline material with *mp* 190 – 200 °C (with decomposition). It shows weak solubility in water (about 2.7 g per 100 mL).

The UV spectrum of thiamin chloride hydrochloride (**1b**) in aqueous solution shows maxima at 234 nm ($\varepsilon_1 = 307$) and 266 nm ($\varepsilon_1 = 277$) [519, 520]. Its IR spectrum in KBr is reported in [521].

The ^{1}H NMR spectrum of thiamin chloride hydrochloride (**1b**) in D_2O is reported in [522], and the ^{13}C and ^{15}N NMR spectra in [523–525].

Since the mass spectrum of thiamin chloride hydrochloride (**1b**) shows no molecular ion under standard EI conditions, literature refers to data obtained using the fast atom bombardment technique [526–528].

X-ray crystallographic data [529] and molecular orbital calculations of the electron density [530] have also been published.

6.4. Chemical Properties

Hydrolysis. Under mildly alkaline aqueous conditions (at pH 7.0 or above), thiamin is converted to the thiol form (**4**) (Fig. 5). The rate-limiting step appears to be ring opening [520, 531,532].

A 5 % aqueous solution of the hydrochloride (**1b**) has pH ca. 3.5, whereas the corresponding solution of the nitrate (**1c**) has pH 6.5 – 7.1.

Under normal conditions, an aqueous solution of thiamin is stable below pH 5.5, even to oxidation, but decomposes to (4-amino-2-methyl-5-pyrimidinyl)methanol (**2a**) [*73-67-6*] and 4-methyl-5-(2-hydroxyethyl)thiazole (**3a**) [*137-00-8*] when heated in a sealed tube at 140 °C (Fig. 5) [533].

By treatment with sulfite in weakly acidic solutions, thiamin chloride (**1a**) is split into the methanesulfonate derivative (**2b**) and the thiazole compound (**3a**) (Fig. 5) [511,512]. Under strongly acidic conditions, thiamin chloride (**1a**) is converted to oxythiamin (**5**), which has no vitamin activity (Fig. 5).

Various aspects of the chemistry of thiamin are still under investigation, especially the reversible opening of the thiazole ring in aqueous alkaline solution [528,534–537].

Intramolecular Aminolysis. Under strongly basic, but anhydrous conditions (2 mol of sodium ethoxide in ethanol), the 4′-amino group of thiamin adds to the thiazole ring to give tricyclic dihydrothiochrome (**6**), which then eliminates a thiolate ion to give the sodium salt of the yellow form of thiamin (**7**) [531, 538–540]. Addition of acid converts **7** to the thermodynamically more stable thiol form (**4**) (Fig. 6).

Oxidation. In the presence of an oxidant, the thiolate (**7**) is irreversibly converted to thiochrome (**8**) (Fig. 6). Thiochrome (**8**) is a yellow crystalline compound which has been isolated from yeast but has no physiological importance. In solution it exhibits a strong blue fluorescence, a property used for the quantitative determination of thiamin (see Section 6.11). Reduction of thiochrome (**8**) converts it back to thiamin, in up to 60 % yield.

Ylid Formation. Upon abstraction of the hydrogen atom from the 2-position of the thiazole ring, the ylid of thiamin (**9**) is formed (Fig. 6), which plays a central role in both the coenzymatic reaction of vitamin B_1 as well as in nonenzymatic reactions such as the acyloin condensation. The ability of thiamin to form an ylid can be rationalized in terms of molecular orbital theory and resonance structures [534, 541]. There is, however, widespread disagreement about whether the 4′-amino group acts as an intramolecular acid or base in either the chemistry or the enzymology of ylid formation from thiamin. Isolation of the ylid in the solid state after deprotonation of thiamin chloride with one equivalent of sodium ethoxide has been the subject of much controversy [531,534], the tricyclic structure (**6**) being favored over a tetracyclic one, based on NMR studies. Kinetic and thermodynamic studies have been made to interpret the lability of the 2-proton in the thiazole ring of thiamin [542–544].

The pK_a of thiamin in aqueous solution at 25 °C is estimated at 12.6 for the CH in the thiazole ring and 4.8 for the NH_2 of the pyrimidine moiety [531,541].

Reduction. Thiamin undergoes reduction of the thiazole ring by a variety of reducing agents (lithium aluminum hydride, sodium borohydride, etc.) to afford, via the intermediate dihydrothiamin (**10**), tetrahydrothiamin (**11**) [545, 546] (Fig. 7).

6.5. Natural Occurrence and Isolation

Thiamin is widespread in nature, but only in relatively small quantities [547–549].

Average thiamin concentrations in foodstuffs (in μg/100 g) are as follows:

Wheat germ	2050
Brewer's yeast, dried	1820
Soyabeans	1300
Dried beans and peas	680

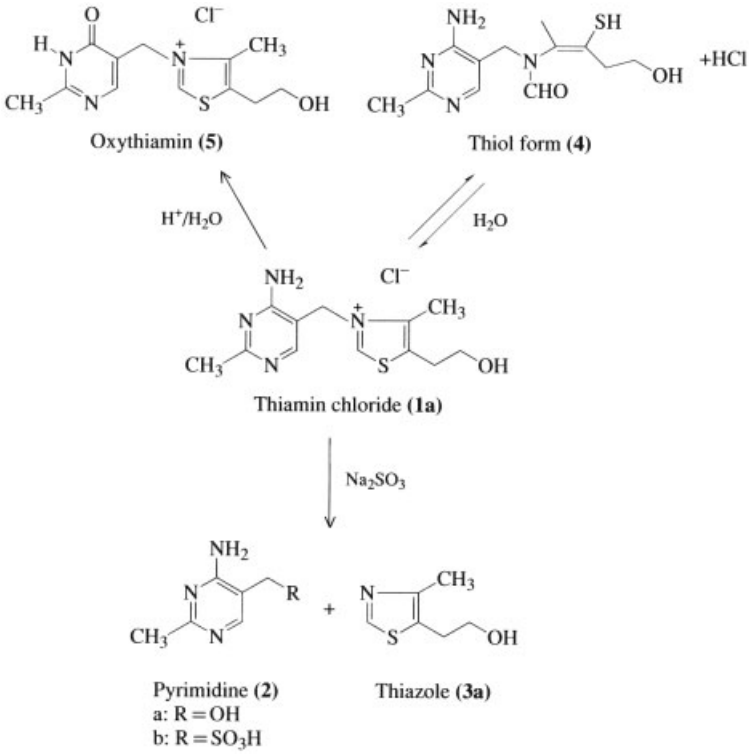

Figure 5. Reactions of thiamin chloride (**1a**)

Nuts	300 – 560
Rice, brown	300
Rice, white	50
Potatoes	170
Beef heart	600
Beef liver	300
Pork	600 – 950
Milk whey, dried	500
Whole milk	30 – 70
Eggs	70
Fish	50 – 90
Vegetables (flower and fruit)	70
Vegetables (leaf and stem)	70
Vegetables (root and tuber)	60

Thiamin can be found in animal tissue in phosphorylated forms, mainly as the pyrophosphate (**12b**) [*154-87-0*] (see section 6.9), where it is bound to an enzyme as a protein complex. The heart, liver, kidney, and the brain are organs with a high thiamin concentration of ca. 100 mg/100 g. Normal human blood contains ca. 90 ng of thiamin per liter, although there are strong variations among individuals, with 49 ng/L being considered as indicative of a possible deficiency.

In plant products, free thiamin is the most abundant form. It is found in the pericarp and the seeds of grains, cereal grains, yeast, dried vegetables, rice, and potatoes. Oils, fats, and highly processed foods such as refined sugars are essentially devoid of thiamin.

Although thiamin is widespread in foodstuffs, its concentration in individual foods varies widely and is relatively low since considerable amounts are destroyed in cooking, either by heat, the presence of metals, chlorine in water, or by reactive organic substances. Therefore, to restore the nutritional quality of processed foods to the original or acceptable levels, they have to be supplemented with synthetic thiamin or its derivatives. Thus in developed countries most white rice and white flour are fortified with vitamin B_1. Since extraction of thiamin from natural sources would not be economically profitable, it has to be manufactured by chemical synthesis.

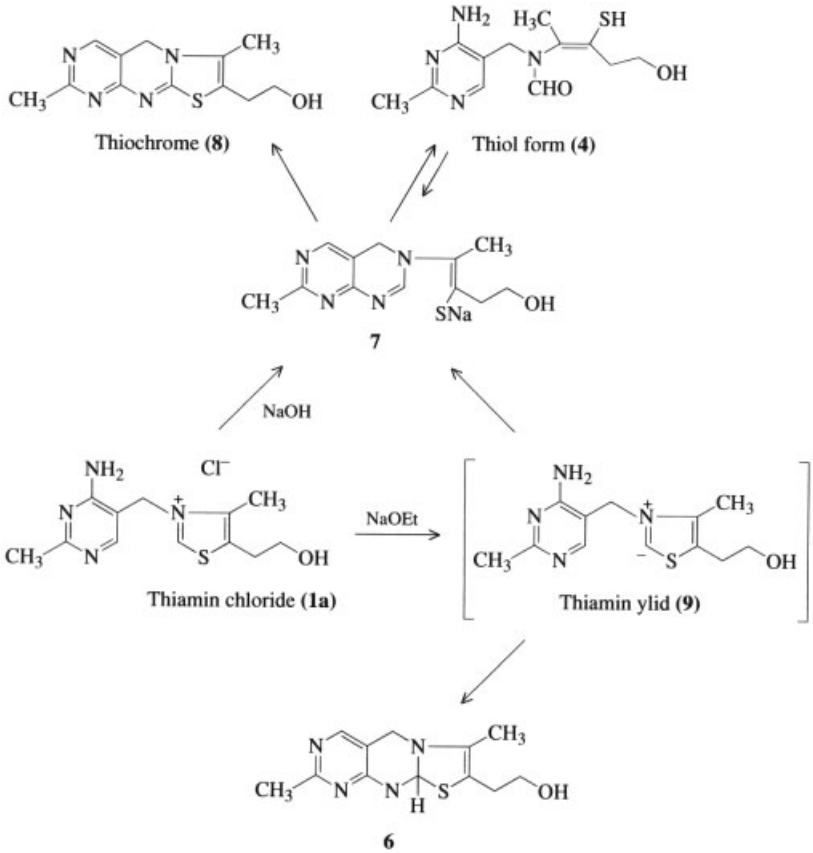

Figure 6. Reactions of thiamin chloride (**1a**)

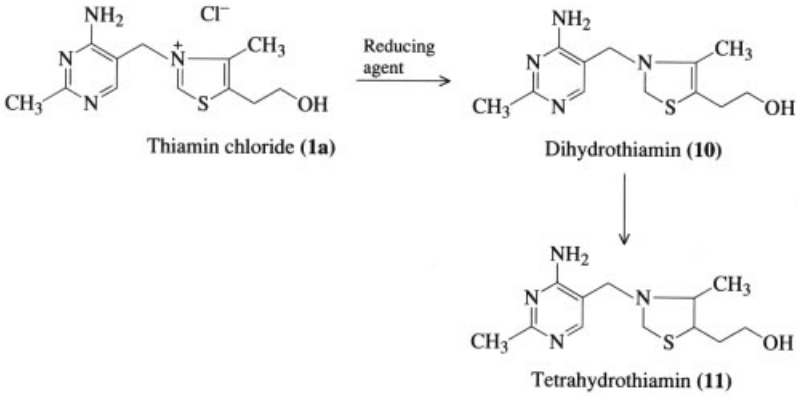

Figure 7. Reduction of thiamin

6.6. Biosynthesis in Microorganisms

Thiamin pyrophosphate (**12b**) is synthesized in prokaryotic and eukaryotic microorganisms from the precursor compounds hydroxymethylpyrimidine (**2a**) and hydroxyethylthiazole (**3a**). The biosynthetic pathways leading to these precursor compounds are different in facultative anaerobic bacteria, e.g., *Escherichia coli* or *Salmonella typhimurium*, and in lower eu-

karyotes such as *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*. Reviews covering thiamin biosynthesis up to the mid-1980s can be found in [550, 551]. The latter is restricted to *E. coli* and *S. typhimurium*.

6.6.1. Biosynthesis of the Pyrimidine Component in Prokaryotes

One of the key findings revealing the origin of the pyrimidine component (**2a**) of thiamin in prokaryotes was the observation that certain mutants of *S. typhimurium* defective in one of the early steps of purine biosynthesis were thiamin auxotroph as well [552–555].

Specifically, mutants defective for conversion of phosphoribosyl pyrophosphate (PRPP) to phosphoribosyl aminoimidazole (**13**) (AIR) (e.g., *purF* mutants) required thiamin supplementation in addition to purine nucleotides for growth (Fig. 8). Purine mutants impaired in conversion of AIR (**13**) to inosine monophosphate (IMP), however, were independent of exogenous thiamin supply. These and other observations [550, 551] suggested that purine nucleotide and thiamin biosynthesis use the same route up to production of AIR (**13**). Conversion of AIR (**13**) to the pyrimidine component (**2a**) is a complex intramolecular rearrangement, which in *E. coli* depends on the *thiC* gene product (see below).

Additional pathways to the pyrimidine component (**2a**) in *S. typhimurium*, independent of *purF*, have been suggested [556–558].

6.6.2. Biosynthesis of the Pyrimidine Component in Eukaryotes

An initial indication that bacteria and yeast employ different biosynthetic pathways to the pyrimidine component (**2a**) of thiamin was the finding that in yeast cultivated with ^{14}C-labelled formate the ^{14}C atom was recovered at the C4 position of the pyrimidine component (**2a**) [559].

In contrast, the pyrimidine component from ^{14}C-formate-fed prokaryotes contained the ^{14}C label in the C2 position. Feeding of *S. cerevisiae* with ^{14}C- or ^{15}N-labelled histidine resulted in nonrandomly labelled thiamin [560, 561] and indicated that N3, C4, and the amino nitrogen N9 of the pyrimidine moiety (**2a**) were derived as a unit from the N – C – N part of the imidazole ring of histidine. The amino nitrogen of histidine was not incorporated. The source of C2, C5, C6, the methoxy C7, and the methyl C8 atom of the pyrimidine component (**2a**) was ribose [562–564]. A minor pathway to the pyrimidine component (**2a**) in eukaryotes has been suggested in which the single carbon atom of formate is incorporated into the C2 position. This pathway, however, is not identical to the hydroxymethylpyrimidine biosynthesis pathway in prokaryotes [562].

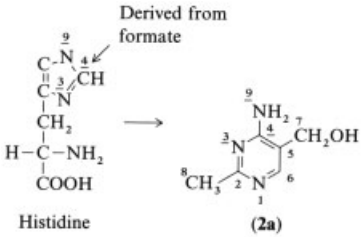

6.6.3. Biosynthesis of the Thiazole Component

Cysteine is the common sulfur source for biosynthesis of the thiazole component (**3c**) of thiamin [565]. In yeasts the N and C2 atoms are obtained from glycine (Fig. 9) [566, 567]. In the yeast *Candida* the source of the C4 to C8 atoms is ribulose-5-P (**14**) derived from the oxidative branch of the pentose phosphate pathway [568]. In *S. cerevisiae* the pentose contributing C5 to C8 of the thiazole component (**3c**) is derived from the oxidative and nonoxidative branches of the pentose phosphate pathway to a similar extent [569, 570].

In facultative anaerobic bacteria such as *E. coli* and *S. typhimurium* the C2 and N atoms of the thiazole component (**3c**) are derived from the α-carbon and nitrogen atom of tyrosine (Fig. 10) [571–573]. The C_5 sugar D-1-deoxyxylulose (**15**), derived from the glycolysis intermediates pyruvate and glyceraldehyde-3-phosphate, serves as donor for the C4 to C8 carbon atoms of the thiazole moiety (**3c**) [574].

In aerobic bacteria such as *B. subtilis* the C2 and N atoms of the thiazole component (**3c**) are derived from glycine, as in yeasts [575].

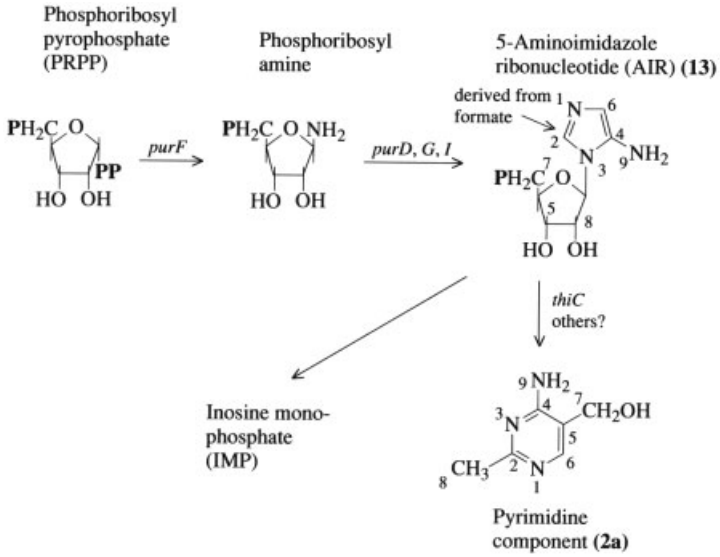

Figure 8. Biosynthesis of the pyrimidine component in prokaryotes

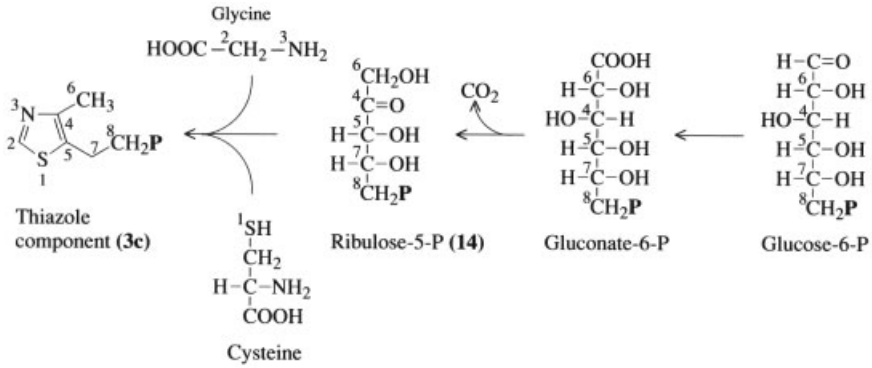

Figure 9. Biosynthesis of the thiazole component in yeasts

6.6.4. Biosynthesis of Thiamin Pyrophosphate

Biosynthesis of thiamin pyrophosphate (**12b**) from the pyrimidine (**2a**) and thiazole (**3a**) precursors (Fig. 11) depends on the activity of five enzymes, four of which are kinases. Hydroxymethylpyrimidine kinase and phosphomethylpyrimidine kinase convert pyrimidine component (**2a**) to the mono- (**2c**) and diphosphorylated (**2d**) derivatives, respectively. The enzymes have been purified from yeast [576,577]. Hydroxyethylthiazole kinase catalyzes phosphorylation of the thiazole component (**3a**) to the phosphorylated derivative (**3c**). An *S. cerevisiae* mutant defective for this enzyme grew well on minimal medium without thiamin supplementation [578]. It was concluded that in yeast, hydroxyethylthiazole kinase is not required for de novo synthesis of thiamin but seems to be involved in a salvage pathway of thiamin biosynthesis using preformed hydroxyethylthiazole (**3a**). The coupling enzyme thiamin phosphate synthase has been purified from *S. cerevisiae* and partially purified from *E. coli* [579,580]. Hydroxyethylthiazole kinase and thiamin phosphate synthase of *S. cerevisiae* mapped in the same gene, and the two enzymatic activities could not be separated during purification to apparent homogeneity. This finding indicated a bifunctional enzyme in the thiamin pyrophosphate biosynthetic pathway of yeast [581].

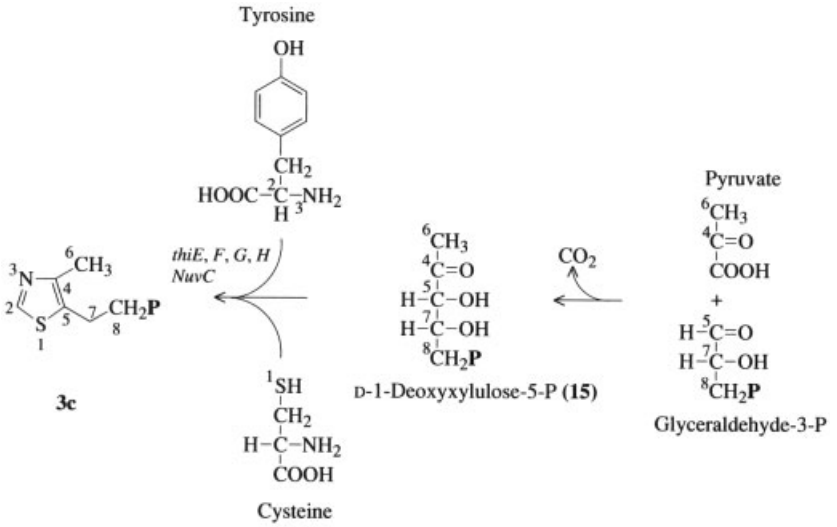

Figure 10. Biosynthesis of the thiazole component in facultative anaerobic bacteria

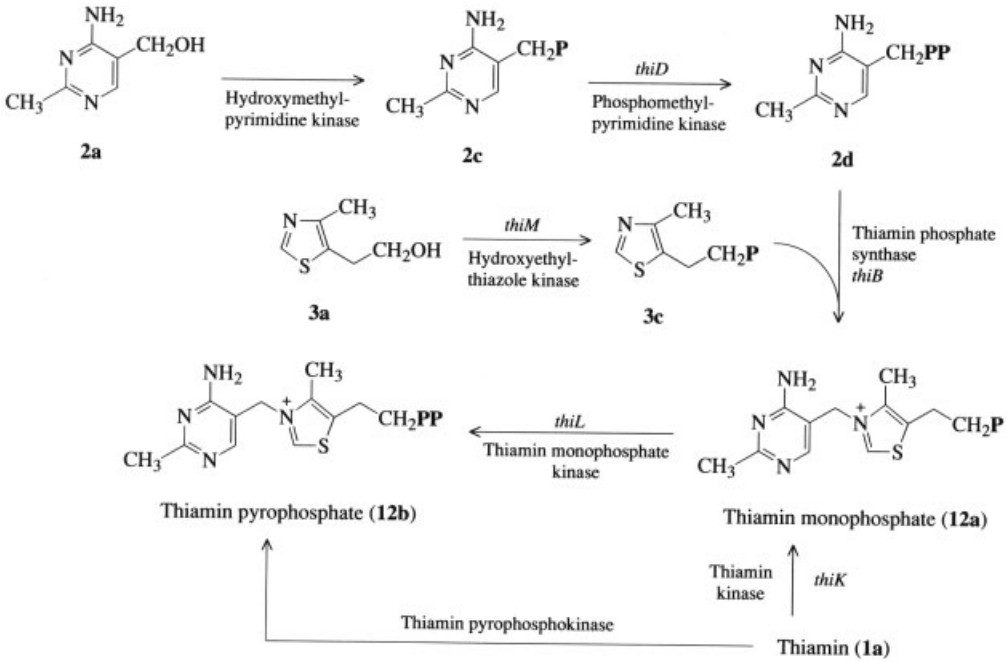

Figure 11. Biosynthesis of thiamin pyrophosphate

In *E. coli*, thiamin monophosphate (**12a**) can be derived from exogenous thiamin (**1a**) by means of the enzyme thiamin kinase. Conversion of thiamin monophosphate to the active coenzyme thiamin pyrophosphate (**12b**) is catalyzed by thiamin monophosphate kinase. In yeasts and many other organisms, including humans (for references, see [582]), thiamin pyrophosphate (**12b**) can be obtained from exogenous thiamin (**1a**) by a one-step phosphorylation reaction catalyzed by thiamin pyrophosphokinase.

6.6.5. Thiamin Biosynthetic Genes from *E. coli*

A number of mutations leading to thiamin auxotrophy in *E. coli* (*thi* mutants) mapped at 90′ of the *E. coli* chromosome [583,584]. A 6.5 kilobase genomic DNA fragment was isolated from *E. coli* that complemented all of the mutations clustered in the 90′ *thi* locus [585]. Five tightly linked open-reading frames, designated as *thiC, thiE, thiF, thiG* and *thiH* were identified. To evaluate the function of the genes encoded by these open-reading frames mutants were constructed, each of them carrying another defect in one of the five open-reading frames. The growth requirements of these mutants indicated that the *thiC* gene product catalyzes an essential step in biosynthesis of the pyrimidine component (**2a**) of thiamin from 5-aminoimidazole ribonucleotide (**13**) (AIR) (see Section 6.6.1). The *thiE, thiF, thiG* and *thiH* mutants responded to thiazole (**3a**), but not to D-1-deoxyxylulose (**15**), indicating that the *thiE, thiF, thiG*, and *thiH* genes are involved in synthesis of the thiazole component (**3a**) from the precursor compounds tyrosine, cysteine, and D-1-deoxyxylulose (**15**) (see Fig. 10).

The *thiG* gene exhibits significant homology to the *chlN* gene of *E. coli* [586], encoding an enzyme of the molybdopterin biosynthetic pathway [587]. Molybdopterin complexed with a Mo atom is the cofactor of Mo-containing enzymes, e.g., respiratory chain nitrate reductase, formate dehydrogenase, or biotin-D-sulfoxide reductase [588, and references therein]. Molybdopterin, like the thiazole component (**3a**) of thiamin, contains a sulfur atom, suggesting a role of the *chlN* and the *thiG* gene products in sulfur incorporation into molybdopterin and into the thiazole component (**3a**), respectively.

The kinases encoded by *thiD, thiM, thiL* and *thiK*, which are involved in thiamin pyrophosphate biosynthesis from the pyrimidine (**2a**) and thiazole component (**3a**) or from thiamin (**1a**) (see Fig. 11), have been mapped on the *E. coli* chromosome [589,590].

The *nuvC* gene located at 45′ of the *E. coli* chromosome was found to be involved in thiamin pyrophosphate biosynthesis, as well [591]. The *nuvC* gene product is one of the two enzymes, designated factors A and C, responsible for modification of the uridine moiety of certain tRNAs with sulfur. *E. coli nuvC* mutants lacking factor C were found to be thiamin auxotroph in addition to being unable to produce 4-thiouridine. The mutations could be complemented by administration of hydroxyethylthiazole (**3a**). One could speculate that factor C, which transfers the sulfur atom from cysteine to tRNA uridine to form 4-thiouridine, is also responsible for delivering a sulfur atom from cysteine to the thiazole component (**3a**).

6.6.6. Mapping and Cloning of Thiamin Biosynthetic Genes from Other Species

Thi genes complementing *thiA32* (hydroxyethylthiazole (**3a**) requirement), *thiC34* (hydroxymethylpyrimidine (**2a**) requirement) and to some extent *thiB33* mutations (coupling defect) of *E. coli* mutants have been mapped on the second symbiotic megaplasmid of *Rhizobium meliloti* [592].

Three classes of *B. subtilis thi* mutants can be distinguished according to the dependence on hydroxymethylpyrimidine (**2a**) (*thiA* mutants), hydroxyethylthiazole (**3a**) (*thiB* mutants) or thiamin (*thiC* mutants) for growth [593]. (Note that the nomenclature of *thi* mutants in *B. subtilis* and *E. coli* is not consistent.) Whereas in *E. coli* the genes involved in biosynthesis of the two thiamin precursor components and the coupling enzyme are tightly clustered, the *thiA*, *-B*, and *-C* mutations of *B. subtilis* mapped at 70°, 102°, and 331° of the bacterial chromosome, respectively [594]. During the course of the *Bacillus subtilis* genome sequencing project a 97 kb region from 325° to 333° including the *thiC* locus has been cloned and sequenced [595].

Schizosaccharomyces pombe thi2 mutants required hydroxyethylthiazole (**3a**) for growth. The *thi3* and *thi4* mutants could be complemented with hydroxymethylpyrimidine (**2a**) and thiamin (**1a**), respectively [596]. The *thi3* gene, which is identical to the *nmt1* gene [597], has been isolated and sequenced. The *thi2* gene has been cloned [598], but the sequence has not been published so far. The thiamin biosynthetic *nmt2* gene of *Schizosaccharomyces pombe* has been cloned and sequenced [599]. The *thi4* gene [600], which should be involved in phosphorylation of one of the two thiamin precursor components or in the coupling step, has some

weak homology to the *E. coli thiE* gene, which is required for biosynthesis of the thiazole component (**3a**). There was no further sequence homology observed between the *thi* genes of *E. coli* and *Schizosaccharomyces pombe* in accordance with the different biochemical routes leading to thiamin pyrophosphate (**12b**) in yeast and bacteria. Recently the DNA sequence of the *thi80* gene encoding thiamin pyrophosphate kinase from yeast has been reported [582].

6.6.7. Regulation of Thiamin Biosynthesis in Prokaryotes

Thiamin biosynthesis is tightly controlled by feedback and repression mechanisms. In *S. typhimurium*, thiamin, which has no effect on PRPP aminotransferase, interfered with biosynthesis of pyrimidine (**2a**) and thiazole component (**3a**) at 40 ng/mg dry cell weight, and with the coupling reaction at 100 ng/mg dry cell weight [601,602]. In addition, hydroxyethylthiazole (**3a**) inhibited its own synthesis at 1 ng/mg dry cell weight, whereas hydroxymethylpyrimidine (**2a**) did not interfere with its own synthesis.

The hydroxymethyl- and phosphomethylpyrimidine kinases, hydroxyethylthiazole kinase, and the coupling enzyme thiamin phosphate synthase (see Section 6.6.4) are repressed in *E. coli* by intracellular thiamin [584,603,604]. Mutant *E. coli* strains resistant towards the thiamin antimetabolite pyrithiamin exhibited similar activities of the coupling enzyme and hydroxyethylthiazole kinase, irrespective of the intracellular thiamin concentration. The mutation in one of these strains, designated *PT-R1*, mapped at 90′ of the *E. coli* genome and is probably very close to the *E. coli thi* locus.

6.6.8. Regulation of Thiamin Biosynthesis in Yeasts

In *S. cerevisiae* two positive regulatory genes of thiamin biosynthesis have been described and cloned [605–607]. These genes were designated *thi2* (*pho6*) and *thi3* similar to the structural *thi* genes of *S. pombe*. Mutations in *S. cerevisiae thi2* and *thi3* resulted in thiamin auxotrophy and affected synthesis of thiamin monophosphate (**12a**) from the pyrimidine (**2a**) and thiazole component (**3a**). In addition, *thi3* mutants had a markedly impaired thiamin transport system.

As in *S. typhimurium* and *S. cerevisiae*, biosynthesis of thiamin is negatively regulated in *S. pombe* by thiamin [596]. Expression of the thiamin biosynthetic genes *thi2* and *thi3* involved in biosynthesis of the thiazole (**3a**) and pyrimidine (**2a**) components in *S. pombe*, respectively (see Section 6.6.6), is strongly repressed by thiamin [598]. In addition, the *thi2* gene is repressed by hydroxyethylthiazole (**3a**), but not by hydroxymethylpyrimidine (**2a**), similar to the situation in *S. typhimurium*. Several mutations designated *tnr*1, 2, and 3 have been mapped in the genome of *S. pombe*, resulting in de-repressed expression of genes involved in thiamin biosynthesis [608]. Intracellular levels of thiamin (**1a**), but not of thiamin monophosphate (**12a**) or thiamin pyrophosphate (**12b**), in *tnr*3 were 10- to 20-fold higher (ca. 10 ng/mg dried cells) than in *tnr*1, 2 or in the wildtype. In addition to mutations in the *tnr* genes that might encode negative regulators of thiamin biosynthesis, a positive regulatory gene designated *thi1* has also been identified in *S. pombe*. *Thi1* mutants are thiamin auxotroph. They can be complemented with hydroxyethylthiazole (**3a**), but not with hydroxymethylpyrimidine (**2a**), indicating that these mutants can not produce the thiazole component (**3a**). Accordingly, no *thi2* mRNA encoding a structural gene of thiazole biosynthesis could be detected in *S. pombe thi1* mutant strains.

6.6.9. Thiamin-Overexpressing Microorganisms

E. coli PT-R1 mutants which are deregulated for thiamin biosynthesis (Section 6.6.7) accumulated ca. 100 ng thiamin/mg dry cellular weight [584], which corresponded to a three- to fourfold increase of thiamin levels compared to wildtype *E. coli* strains. Similar levels of thiamin pyrophosphate (**12b**) were reached in *S. typhimurium* 1.5 h after derepression of thiamin biosynthesis caused by adenosine preincubation [602]. An *S. typhimurium thi* mutant defective in thiazole biosynthesis accumulated hydroxymethylpyrimidine (**2a**) in the medium.

Certain *Neurospora crassa* mutants containing a defect in biosynthesis of the thiazole component (**3a**) secreted up to 0.2 mg/L of substances closely related to the pyrimidine component (**2a**) into the culture broth. The chemical nature of one of them was identified as being the diamine (**2e**) [609, 610]. A broad search among many yeast species has been reported [611]. The *Saccharomyces cerevisiae* strain 1271 secreted up to 200 mg/L thiamin into the culture broth.

6.7. Chemical Synthesis

Several syntheses of thiamin (**1**) have been published, and basically two general methods have evolved over the years [612–616].

6.7.1. Condensation of the Pyrimidine and Thiazole Rings

The first approach, developed after the original method reported by WILLIAMS et al. [514], consists of separate syntheses of 4-amino-5-bromomethyl-2-methylpyrimidine hydrobromide (**2f**) [*2908-71-6*] and of the thiazole moiety (**3a**) or its acetate (**3b**) [*656-53-1*]. Condensation of the two intermediate heterocycles gives thiamin bromide hydrobromide [*4234-86-0*] (**1d**) which is converted to thiamin chloride hydrochloride (**1b**) by treatment with silver chloride in methanol or with an ion-exchange resin.

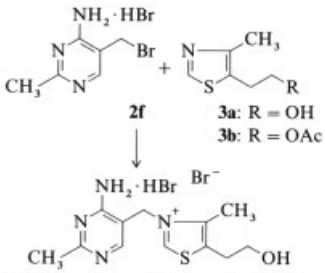

Thiamin bromide hydrobromide (**1d**)

The first industrial synthesis of vitamin B_1, developed by Merck-Rahway (United States), followed this synthetic route.

6.7.2. Construction of the Thiazole Ring on a Preformed Pyrimidine Portion

The second general method, based on the procedure of TODD and BERGELL [516, 517] and ANDERSAG and WESTPHAL [518], constructs the thiazole ring on a preformed pyrimidine intermediate. Thiamin manufacturers generally follow this approach. Some use malonitrile [*109-77-3*] as the starting material for construction of the 4-amino-5-aminomethyl-2-methylpyrimidine ring (**2e**) [*95-02-3*] (also called Grewe diamine). Others start with acrylonitrile but all proceed via the key Grewe diamine intermediate (**2e**) through to thiothiamin [*299-35-4*] (**16**) (see page 82).

For the preparation of malonitrile, β-aminopropionitrile (**17a**), obtained by addition of ammonia to acrylonitrile, is subjected to oxidative dehydrogenation in the gas phase at high temperature and in the presence of molecular oxygen and a metallic catalyst [617].

Pyrimidine Moiety. A convenient synthesis of the pyrimidine moiety starts from acetamidine hydrochloride [*124-42-5*] or the corresponding imino ether and a suitably substituted C_4 unit carrying the required functional groups, such as methoxymethylenemalononitrile [*672-25-3*] (**18a**), aminomethylenemalononitrile [*672-81-1*] (**18b**), 2-dimethoxymethyl-3-methoxypropionitrile (**19**), or α-amino-β-formylaminopropionitrile (**20**) (aniline, *o*-chloroaniline or *N*-methylaniline are the most frequently used α-amino groups).

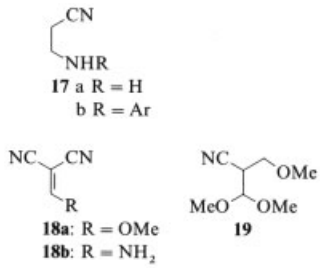

Subsequent conversion of the resulting intermediates – 5-cyanopyrimidine [*698-29-3*] (**21a**), *N*-[(4-amino-2-methyl-5-pyrimidinyl)-methyl]-acetamide [*23676-63-3*] (**2g**), *N*-[(4-amino-2-methyl-5-pyrimidinyl)-methyl]-form-

amide (**2h**) – provides the desired Grewe diamine (**2e**) [613, 618–623].

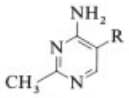

21a: R = CN
21b: R = CH(OMe)$_2$
21c: R = CHO

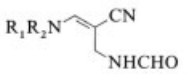

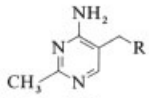

20a: R$_1$ = H R$_2$ = H
20b: R$_1$ = phenyl R$_2$ = H
20c: R$_1$ = phenyl R$_2$ = Me
20d: R$_1$ = *o*-chlorophenyl R$_2$ = H

2e: R = NH$_2$
2g: R = NHC(O)CH$_3$
2h: R = NHCHO
2i: R = NHCHS

The Takeda and Ube in Japan have jointly developed a new process to prepare the key intermediate Grewe diamine (**2e**) [624, 625]. Acrylonitrile and carbon monoxide are used as basic raw materials. Nitrite-catalyzed oxidation of acrylonitrile in the presence of an alcohol affords 3,3-alkoxypropanenitrile (**22**). This same intermediate can also be obtained, probably more expensively, by formylation of acetonitrile, followed by acid-catalyzed acetalization [626, 627]. Alternatively, **22** can be generated by electrooxidation of β-aminopropionitrile (**17b**) [628].

22: R = Me or Bu

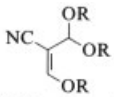

23: R = Me or Bu

Carbonylation of dialkoxypropanenitrile (**22**) followed by acid-catalyzed acetalization and subsequent thermal elimination of alcohol yields 2-(dialkoxymethyl)-3-alkoxy-2-propenenitrile (**23**). On reaction with acetamidine, **23** gives in high yield the acetal (**21b**), which after hydrolysis to the 2-methyl-4-amino-5-pyrimidinecarboxaldehyde [*73-68-7*] (**21c**) and reductive amination provides the desired Grewe diamine (**2e**) [629–631].

In a different approach [632], 5-cyano-1,6-dihydropyrimidine (**25**), prepared by Thorpe–Ziegler cyclization of N'-cyano-N-(2-cyanoethyl)-acetamidine (**24**), is converted directly to the desired Grewe diamine (**2e**) by dehydrogenation in the presence of Raney cobalt and ammonia.

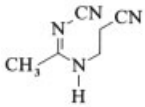

24

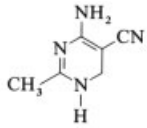

25

Thiamin Synthesis. Having obtained 4-amino-5-aminomethyl-2-methylpyrimidine (**2e**), three chemical steps are still required to obtain thiamin (**1a**): extension of the aminomethyl side chain at the 5-position, cyclization to the thiazole ring, and conversion to thiamin. Commonly used synthons for extending the side chain are 3-chloro-5-hydroxypentan-2-one [*13045-13-1*] (**26a**), the 3-mercaptoketone [*15678-01-0*] (**26b**) or their corresponding acetates [*13051-49-5*] (**26c**) and [*55289-66-2*] (**26d**) [633, 634]. Besides the multistep preparation of chloroketone (**26c**), a catalytic procedure has been developed in which a mixture of α-chloroacetylacetone (**27**) and ethylene oxide is reacted at 80 °C in an autoclave in the presence of nickel(II) acetylacetonate as catalyst [635].

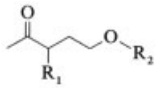

26a: R$_1$ = Cl, R$_2$ = H
26b: R$_1$ = SH, R$_2$ = H
26c: R$_1$ = Cl, R$_2$ = C(O)CH$_3$
26d: R$_1$ = SH, R$_2$ = C(O)CH$_3$

27

Condensation of 4-amino-5-aminomethyl-2-methylpyrimidine (**2e**) with carbon disulfide and chloropentanone (**26a**) provides in very good yield the dithiocarbamate (**28**), which is then converted by acidic treatment to the 4-thiazoline-2-thione derivative (**16**) [636–639]. Oxidative desulfurization with hydrogen peroxide followed by anion-exchange resin completes the synthesis of thiamin hydrochloride (**1a**).

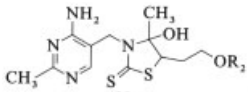

28

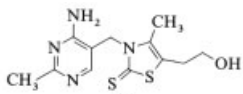

16

Similarly, thiamin hydrochloride (**1b**) can be prepared by condensation of *N*-[(4-amino-2-methyl-5-pyrimidinyl)methyl]methanethioamide (**2i**) [*31375-20-9*], or *N*-formyl Grewe diamine (**2h**) with ester (**26c**) or (**26d**), respectively [633,634,640]. Reaction of Grewe diamine (**2e**) with α-mercaptoketone (**26b**) and formaldehyde provides dihydrothiamin (**10**) which can then be converted to thiamin (**1a**) [641,642].

A more convergent synthesis of thiamin (**1b**) has been developed by Hoffmann-La Roche [643,644]. Condensation of 3,4-dihydro-7-methylpyrimido[4,5-*d*]pyrimidine (**29**) [*31375-19-6*] with the mercaptopentanone (**26d**) in formic acid followed by treatment with hydrochloric acid gave thiamin in high yield. **29** is obtained by heating Grewe diamine (**2e**) in dimethylformamide dimethyl acetal [645–647] or alternatively, in an excess of triethyl orthoformate with a catalytic amount of *p*-toluenesulfonic acid.

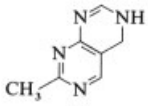

29

6.8. Commercial Forms

The commercially available forms of thiamin are the chloride hydrochloride (known as thiamin hydrochloride) (**1b**) and the mononitrate (**1c**). World production of thiamin in 1993 was estimated at ca. 4200 t; the principal producers are Hoffmann-La Roche, Takeda Chemical Industries, and Chinese State Companies. The price for thiamin in 1993 on the U.S. market was ca. $ 30 – 35/kg.

The discovery in 1951 by Fujiwara et al. [648] that treatment of thiamin with an extract of garlic or other *Allium* species converts it to a product that is physiologically very active opened the door to derivatives of thiamin that act as "prodrugs." The mixed thiamin allyl disulfide, TAD, (**30a**) (allithiamin [*554-44-9*]) was the first representative of a variety of thiamin alkyl sulfides. The most important commercial forms, almost exclusively marketed in Japan (they are not yet approved in the United States) are TAD, thiamin propyl disulfide, TPD, (**30b**) (prosulthiamin [*59-58-5*]), thiamin tetrahydrofurfuryl disulfide (**30c**) (fursulthiamin [*804-30-8*]), *O*-benzoylthiamin disulfide (**31**) (bisbenthiamin [*2667-89-2*]), and dibenzoylthiamin [*299-88-7*] (**32**). These forms are more stable, less water soluble and more lipid soluble, being thus more rapidly absorbed and better retained in the body than the more common thiamin hydrochloride. In vivo they are converted back to thiamin [649].

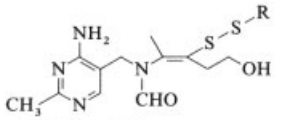

30a: R = allyl
30b: R = *n*-propyl
30c: R = tetrahydrofurfuryl

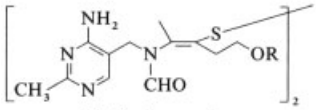

31: R = benzoyl

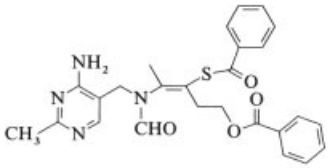

Dibenzoylthiamin (**32**)

6.9. Derivatives, Analogues, and Antimetabolites

Thiamin forms esters at the hydroxyethyl side chain with various acids, the most important being the mono- (**12a**), di- (**12b**), and triphosphates (**12c**). The pyrophosphate (**12b**) acts as a coenzyme (cocarboxylase) in the decarboxylation of α-keto acids in association with carbohydrate metabolism. The palmitate (**33**) is as effective as thiamin and is used for supplementation of various feed and food products due to its stability and good lipid solubility [650].

In Japan, derivatives of thiamin with sugars, especially with saccharide, have also been prepared as food additives. They are readily hydro-

lyzed to physiologically active thiamin and, unlike thiamin, have no unpleasant smell.

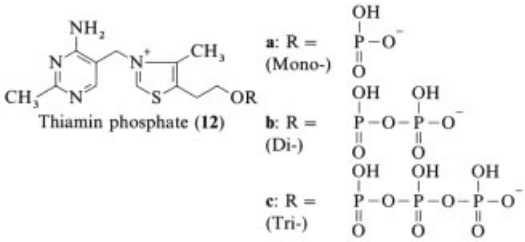

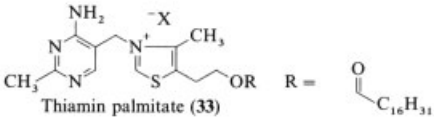

Other derivatives are obtained from the thiol form of thiamin (see Section 6.8).

Several analogues have been synthesized by structural modification of the pyrimidine or thiazole ring, but none exceeds the biological activity of thiamin. The 2-ethyl analogue (**34**) shows activity comparable to that of thiamin, whereas the 2-butyl (**35**), 2-methoxy (**36**), and 2-methylthio (**37**) analogues are potent antagonists [651–653].

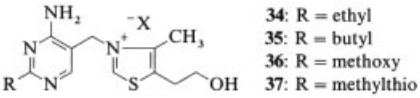

Replacement of the 4-amino group by hydrogen or an alkylamino group generally results in loss of activity [652,654]. Oxythiamin (**5**) (see section 6.4), the 4-hydroxy (or 4-oxo) analogue is a known antagonist.

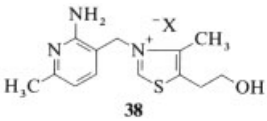

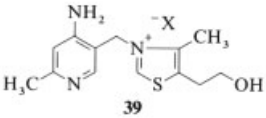

From analogues **38** and **39** it was concluded that the nitrogen atom in the 1-position is more important for coenzyme activity than that at the 3-position [655,656].

Compounds **40** and **41** [657], the 2-methyl substituted analogue **42** [658], oxothiamin (**43**) and thiothiamin (**16**) are completely inactive [659–661], whereas the homothiaminglycol (**44**) is an antagonist [662], demonstrating that hydrogen at the 2-position and the 3-hydroxyethyl group at the 5-position in the thiazole ring are essential for biochemical activity.

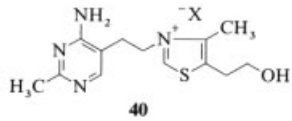

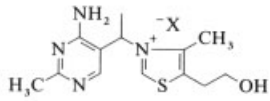

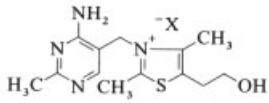

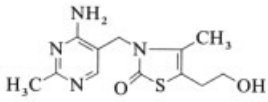

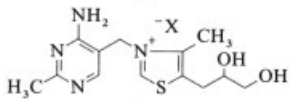

Replacement of the thiazole ring by other heterocycles, as in imidazolthiamin (**45**) [652,663], pyrithiamin (**46**) [659,664,665] or Amprolium (**47**) [666] provides potent thiamin antagonists. The latter is a good inhibitor of thiamin diphosphate dependent enzymes. In small doses, it affects thiamin transport across bacterial cell walls and is therefore an effective coccidiostat. It is used commercially for the treatment of poultry [667,668].

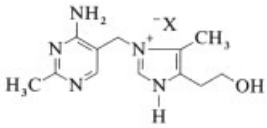

Imidazolthiamin (**45**)

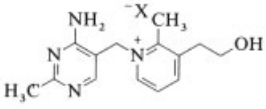

Pyrithiamin (**46**)

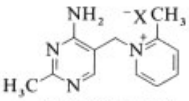

Amprolium (**47**)

6.10. Biochemical and Physiological Functions

6.10.1. Metabolic Functions in the Organism

Thiamin serves a number of essential metabolic functions and its deficiency is associated with imbalances in carbohydrate status, with consequent deleterious effects on nerve functions. In living systems, the only known biologically active form of thiamin is the diphosphate ester (**12b**) – otherwise known as thiamin pyrophosphate, TPP, or cocarboxylase – which is formed by the reaction between thiamin and ATP in liver cells [669]. As a cofactor of enzymes in intermediate metabolism, TPP participates in the decarboxylation of α-keto acids (pyruvate and α-ketoglutarate dehydrogenase complex) and in the reversible α-ketol transfer reactions catalyzed by transketolase in the pentose phosphate cycle [670, 671].

Oxidative Decarboxylation. Pyruvate is an important intermediate product of glycolysis, a largely anaerobic process. Under these conditions, pyruvate is decarboxylated to acetyl-coenzyme A (acetyl-CoA), a universal product of the metabolism of carbohydrates, fats, and amino acids. The decarboxylation reactions are initiated by addition of pyruvic acid to the 2-position of the thiazole ylid (**9**) of TPP to form the adduct (**48a**). The adduct after the splitting-off of carbon dioxide forms the so-called active aldehyde (**49**) from which the subsequent biochemical reactions are derived, leading ultimately to acetyl-coenzyme A [672] (see bottom of page).

To support this mechanism, the protonated form of the adduct between thiamin and acetaldehyde – 2-(1-hydroxyethyl)thiamin (**50**) – has been synthesized [673] as well as the thiamin – pyruvate adduct and its pyrophosphate (**48b**) [674, 675].

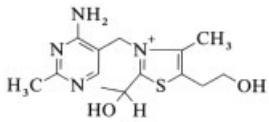

50

In the citric acid cycle, TPP similarly takes part in the oxidative decarboxylation of α-ketoglutaric acid to succinyl coenzyme A. Analogously, the oxidative decarboxylation of the threebranched-chain α-ketoacids derived from the deamination of leucine, isoleucine, and valine is catalyzed by TPP among a multienzyme complex.

Transketolase Reactions. In the transketolase reactions (direct glucose oxidation), TPP adds to appropriate α-ketosugars to break the carbon – carbon bond between C2 and C3 to

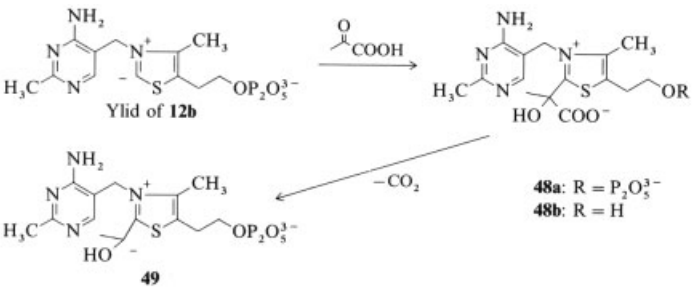

form an active intermediate which is then transferred to a suitable acceptor. This metabolic pathway is important for the production of pentoses for RNA and DNA synthesis.

Other Functions. Thiamin may be an active substance in the nervous system. It has been postulated that thiamin, probably as thiamin triphosphate (**12c**), could play an essential role in the stimulation of peripheral nerves [676–678].

This could be a distinct function of thiamin, independent from its role as a coenzyme in the metabolism of carbohydrates. This remains, however, to be established.

6.10.2. Thiamin Requirements and Deficiency

The maintenance requirements for thiamin depend on carbohydrate intake. An average figure of 20 – 30 µg per kilogram body weight is given for humans. This corresponds to a recommended daily allowance ranging from 0.3 to 1.5 mg/d, depending on age and sex; this recommendation differs from country to country. Factors such as pregnancy, lactation, fever, alcoholism, liver disease, stress, or dietary practices can substantially raise the minimum requirement.

Beriberi is the most prominent pathological condition resulting from a thiamin deficiency. Symptoms of the disease can vary, depending on individual, diet, duration and severity of the deficiency. The disorders affect the central nervous system, but are also accompanied by cardiovascular damage.

In most cases, parenteral administration of thiamin in large doses is sufficient to restore the normal supply status of thiamin. For better absorption from the digestive organs, derivatives of the thiol form of thiamin are used in Japan (see Section 6.8).

6.11. Analytical Methods

For the determination of thiamin in foods, feeds, pharmaceutical preparations and other substances, physicochemical [679, 680] and microbiological [680] methods are employed.

Since thiamin occurs principally as phosphate esters and a significant part is therefore protein bound, extraction procedures have been designed to free thiamin before analytical methods are employed. For some more complex biological materials, further treatment or purification may be necessary to remove compounds which might interfere with the analyses.

The fluorometric thiochrome method is at present the most widely used method for measuring free thiamin. It is based on the oxidation of thiamin (**1a**) to thiochrome (**8**) by an alkaline solution of potassium ferricyanide (see Section 6.4). Mercury(II) chloride or cyanogen bromide can also be used as oxidant. The thiochrome is extracted into isobutanol and the fluorescence of the extract at an emission wavelength of 436 nm compared with that of a standard thiochrome solution.

Thiamin can also be determined spectrophotometrically by measuring its UV absorption at 266 nm, but only in cases where no other materials absorbing at this wavelength are present in significant amounts. A less sensitive and less specific method is a colorimetric test based on the formation of a diazotized aromatic amine from thiamin. Methods based on high performance liquid chromatography offer the advantages of specificity, simplicity, ease of use, and reduced time of analysis [681].

Although earlier bioassays were performed on animals, modern methods make use of microorganisms, a large variety of which have been shown to require thiamin for their growth and reproduction (see Section 6.6). The main disadvantage of using microorganisms is their relative lack of specificity, as some respond either to the thiazole moiety or to the pyrimidine precursor or to intact thiamin. Assays based on *Lactobacillus viridescens* and *Lactobacillus fermenti* are presently the most widely used, partly because both are thiamin specific.

Microbiological assays are simple, inexpensive and quite sensitive (detection limit of 5 – 50 ng thiamin), but their main drawback is the longer period of time to obtain the results.

7. Riboflavin

7.1. Introduction

According to IUPAC rules [691], riboflavin [*83-88-5*] (**1**) is the valid designation for 7,8-

dimethyl-10-(D-1′-ribityl)isoalloxazine, also known as vitamin B_2 or lactoflavin.

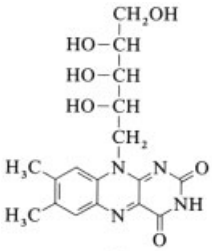

1

Until about 1970, this compound was generally called 6,7-dimethyl-9-(D-1′-ribityl)-isoalloxazine in the literature because of a different numbering of isoalloxazine.

Riboflavin occurs in nature as the free vitamin, as the 5′-phosphate (**2**) or flavin mononucleotide [*130-40-5*] (FMN), and as the 5′-adenosine diphosphate (**3**) or flavin adenine dinucleotide [*146-14-5*] (FAD).

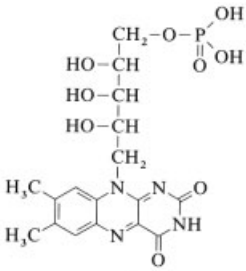

2

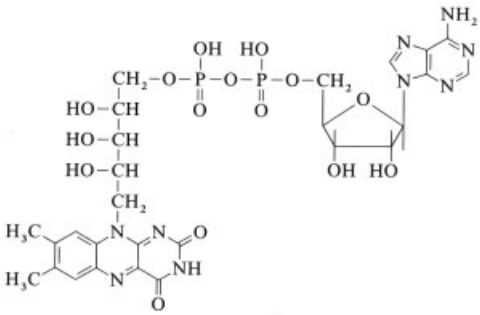

3

7.2. History

A water-soluble growth factor present in milk was mentioned for the first time in 1913 [692]. In 1917, this heat-stable growth factor was isolated from rice bran and separated as a homogeneous fraction from another, heat-sensitive compound having vitamin character. The second compound was able to cure a certain type of polyneuritis [693]. Thus, vitamin B, which was considered to be homogeneous until then, was separated into the vitamins B_1 and B_2.

In 1932, WARBURG and CHRISTIAN described a yellow enzyme which they obtained from aqueous extracts of yeast [694]. In the following year, they separated this enzyme into a protein and a dye component and showed that neither component alone is enzymatically active [695]. The isolation of the protein-free dye in the crystalline state from chicken egg white was achieved in 1933 [696].

It was shown later that the flavins isolated from various sources, which were called ovoflavin, lactoflavin, uroflavin, hepatoflavin, vitamin B_2, and vitamin G, depending on the source, were all identical. The name riboflavin gained acceptance after 1937 when it was recognized that this vitamin is a derivative of ribose.

The structure of riboflavin was established in 1933 – 1935 by the groups of KUHN and KARRER. The key reactions were:

1) Formation of lumiflavin (**4**) on exposure to light under alkaline conditions. The identification of one of the cleavage products as a tetrahydroxybutyl group [697].
2) Alkaline degradation of lumiflavin to 4,5-dimethyl-2-(methylamino)aniline [698], which allowed the identification of lumiflavin as 7,8,10-trimethylisoalloxazine.
3) Total synthesis of riboflavin, whereby the synthesis of different stereoisomeric flavin-10 derivatives led to the establishment of the configuration of the tetrahydroxybutyl side chain of riboflavin [699–701].

7.3. Physical and Chemical Properties

Riboflavin (**1**), M_r 376.36, crystallizes as yellow, bitter-tasting needles. The melting point with decomposition is between 271 and 293 °C,

depending on the heating rate. The solubility in 100 mL water is 5 mg at 7 °C and 13 mg at 27 °C. It can be improved by the addition of, e.g., boric acid, urea, nicotinamide [702], or xanthine compounds [703]. At 27 °C, 4.5 mg of riboflavin dissolves in 100 ml of ethanol. Riboflavin is poorly soluble in pentanol, cyclohexanol, benzyl alcohol, and phenol. It is insoluble in acetone, ether, chloroform, and in aliphatic or aromatic hydrocarbons. It is freely soluble in dilute aqueous alkali solutions, but decomposes rapidly. Neutral aqueous solutions of riboflavin have an intense yellow color (λ_{max} at 445, 372, 265, and 220 nm) and exhibit a green fluorescence (maxima at 530 and 565 nm).

The specific rotation $[\alpha]_D^{25}$ is +59° ($c = 0.5$ in 37 % HCl). Light decomposes riboflavin in neutral or weakly acidic solutions to give lumichrome [*1086-80-2*].

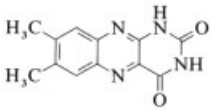

Lumichrome

Irradiation of alkaline solutions with visible light produces lumiflavin [*1088-56-8*] (**4**).

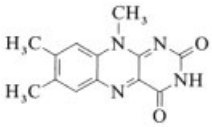

4

In acid solution, riboflavin is resistant to oxidizing agents (e.g., H_2O_2, HNO_3, or halogens) up to ca. 100 °C. This property is used in the purification of crude products during production. Sodium dithionite ($Na_2S_2O_4$) reduces riboflavin with the uptake of two hydrogen atoms to give the colorless leucoriboflavin, which can readily be reoxidized to riboflavin by shaking with air. This reaction can be used in the analytical determination of riboflavin (see Section 7.9.).

Riboflavin binds heavy metals (e.g., Fe, Mo, Cu, Ag, Cd, Ni, Zn, Co), forming deeply colored complexes (metal chelates) [704], for example:

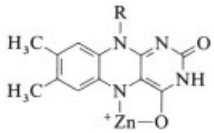

Flavin mononucleotide (**2**) crystallizes as the monosodium salt with two molecules of water of crystallization. Its solubility (pH 4.5) in water, 5 g/100 mL, is much higher than that of riboflavin. It is poorly soluble or insoluble in ethanol, ether, chloroform, and benzene. Owing to its high solubility in water, it is used instead of riboflavin in the production of solutions of the vitamin.

Like riboflavin, FMN is instable in alkaline solution. In an acidic medium, this phosphate ester is hydrolyzed to riboflavin. FMN is most stable at pH 6.

7.4. Occurrence

Riboflavin is one of the vitamins that occur ubiquitously. It is contained in all plant and animal cells. Certain bacteria, yeasts, fungi as well as the organs liver, heart, and kidneys are especially rich in riboflavin. Appreciable amounts are also found in egg white, meat, and milk.

The concentration of riboflavin in foods (in µg/g) is as follows [683]:

White bread	20 – 100
Rye bread	73 – 250
Cows milk	20 – 300
Cheese	330 – 565
Hen's egg	275 – 350
Pork	90 – 350
Beef	40 – 350
Pork liver	2900 – 4400
Spinach	57 – 340
Potatoes	7.5 – 200

The concentration of riboflavin in human tissues (in µg/g) [682] is:

Kidney	20.0
Liver	16.0
Heart	8.3
Adrenal gland	8.2
Stomach	5.2
Ovary	4.3
Ileum	4.2
Spleen	3.6

Brain	2.5
Mammary gland	2.4
Muscle	2.0 – 2.3
Colon	2.1
Testes	2.0
Lung	1.9
Skin	1.2
Spermaduct	1.0

In the culture media of various species of bacteria and fungi (e.g., *Ashbya gossypii*, *Eremothecium ashbyii*, *Bacillus subtilis*, *Candida flareri*), riboflavin is present in concentrations far above 10 g/L (see Section 7.6.3).

7.5. Biosynthesis [684]

Riboflavin is synthesized by many plants and microorganisms. In the animal kingdom, many herbivores which do not form this vitamin in their tissues can cover their riboflavin requirement with the help of bacteria in their digestive tract.

According to BACHER [684, 705–710], the biosynthesis of riboflavin starting with guanosine triphosphate (GTP) involves the following steps (Scheme 21): opening of the imidazole ring of GTP (**5**), catalyzed by the enzyme GTP cyclohydrolase II, produces the pyrimidine derivative (**6**). Conversion of **6** to 5-amino-6-ribitylamino-2,4(1*H*,3*H*)-pyrimidinedione (**10**) proceeds via **8** in fungi and via **7** in bacteria, respectively. The condensation of **10** with the C_4 building block L-3,4-dihydroxy-2-butanone-4-phosphate (**12**) leads to 6,7-dimethyl-8-ribityllumazine (**13**). In *Bacillus subtilis*, this step is catalyzed by the β-subunit of riboflavin synthetase. The α-subunit of riboflavin synthetase catalyzes the dismutation of the lumazine (**13**) to give riboflavin (**1**) and the pyrimidinedione (**10**).

7.6. Production

7.6.1. Syntheses

The structural elucidation of riboflavin (**1**) was completed with the first syntheses conducted by KARRER [711–713] and KUHN [714–716]. KARRER et al. reacted 2-(ethoxycarbonylamino)-4,5-dimethylaniline (**14**) with D-ribose, hydrogenated the resulting Schiff base to give the corresponding *N*-ribitylamine (**15**), and obtained *N*-(2-amino-4,5-dimethylphenyl)-D-1′-ribitylamine (**16**) after alkaline hydrolysis. Reaction with alloxan [*2244-11-3*] (**17**) in hydrochloric acid medium gave riboflavin (**1**) in 15 % yield, based on D-ribose. KUHN et al. directly hydrogenated the Schiff base formed from nitroxylidine (**18**) and D-ribose to compound **16** (Scheme 22).

An improved yield of **16** was attained by introducing the amine group at the stage of *N*-(3,4-dimethylphenyl)-D-1′-ribitylamine (**19**), which is obtainable from 3,4-dimethylaniline and D-ribose. This was achieved by coupling **19** with benzenediazonium chloride to give *N*-(2-phenylazo-4,5-dimethylphenyl)-D-1′-ribitylamine (**20**), followed by reductive cleavage of the azo double bond with sodium dithionite [717]. Cyclocondensation of **16** with alloxan (**17**) produced riboflavin (**1**) in a yield of 38 %, based on D-ribose.

Substantial progress in the synthesis of riboflavin was made by TISHLER [718] who found that in a weakly acidic medium, the azo compound (**20**) can be directly reacted with barbituric acid [*67-52-7*] (**21**) to give riboflavin (**1**), with elimination of aniline. Thus, the reduction of **20** to **16** was no longer required, and instead of the relatively expensive alloxan (**17**), the more readily available barbituric acid (**21**) was used in the synthesis. The total yield of riboflavin (**1**) was increased to 46 %, based on D-ribose.

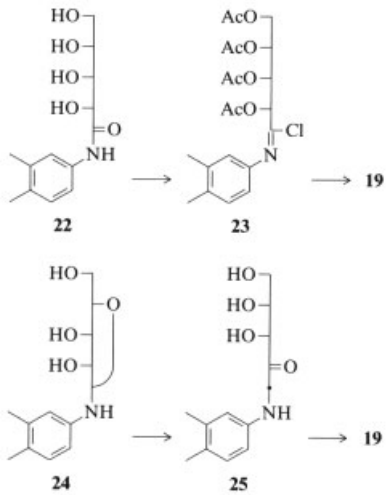

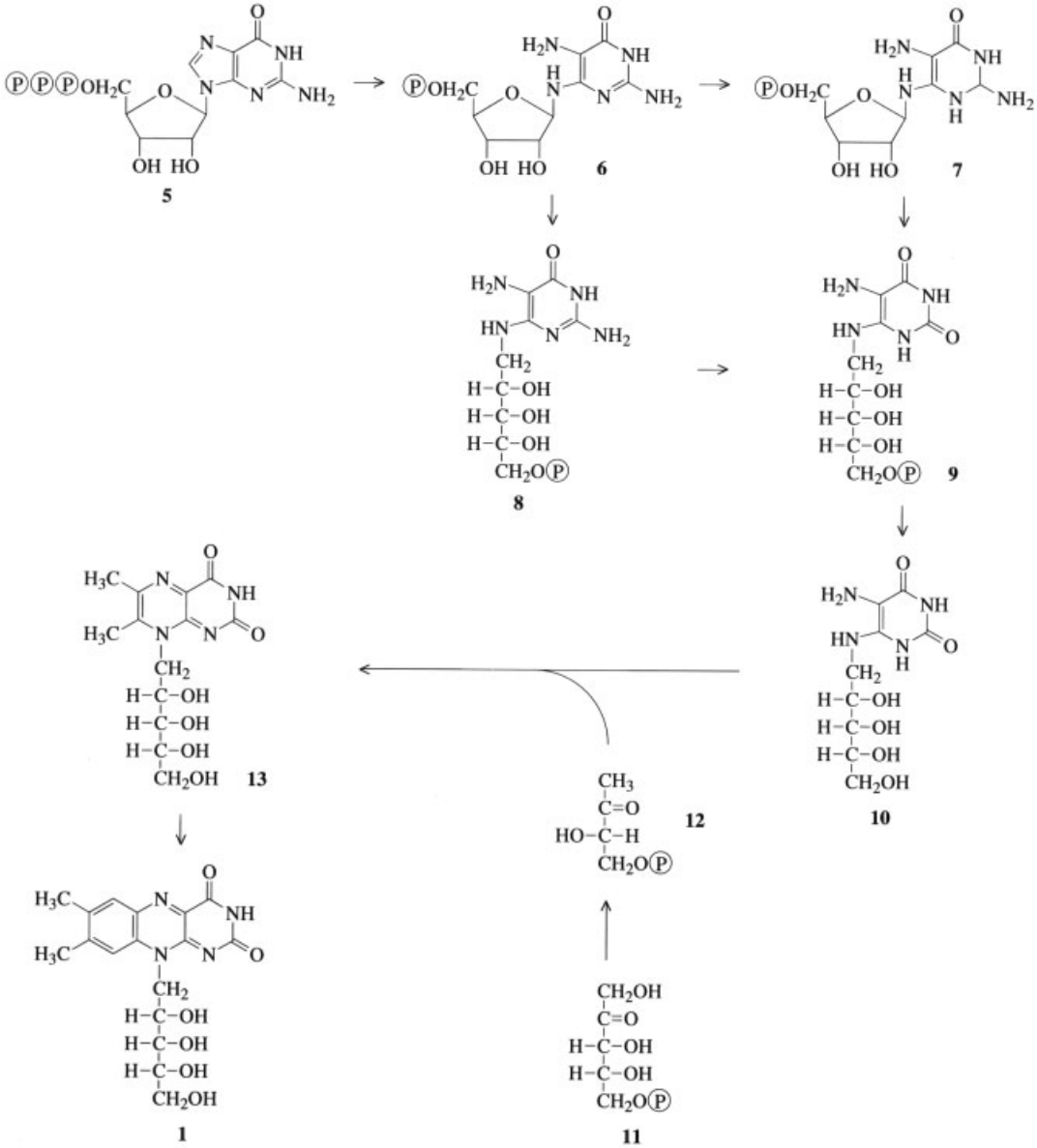

Scheme 21. Biosynthesis of riboflavin

Subsequently, major efforts were made to replace D-ribose, which is not widely available, by other sugars. For example, ribamide (**22**), made from D-ribonolactone and 3,4-xylidine, was directly [719,720] or, after acetylation via the imide chloride (**23**), catalytically hydrogenated to give **19** [721]. D-Arabinose, which is readily available, was reacted with 3,4-xylidine to yield arabinoside (**24**). Amadori rearrangement with benzoic acid gave the keto compound (**25**), which could then be catalytically hydrogenated to **19** in alkaline solution [722,723]. As a result of moderate total yields, none of these syntheses has achieved importance in the production of flavins.

In the 1970s, a novel synthesis of flavin was described by YONEDA et al. They condensed **19** with 6-chlorouracil (**26**) to give 6-(*N*-3,4-di-

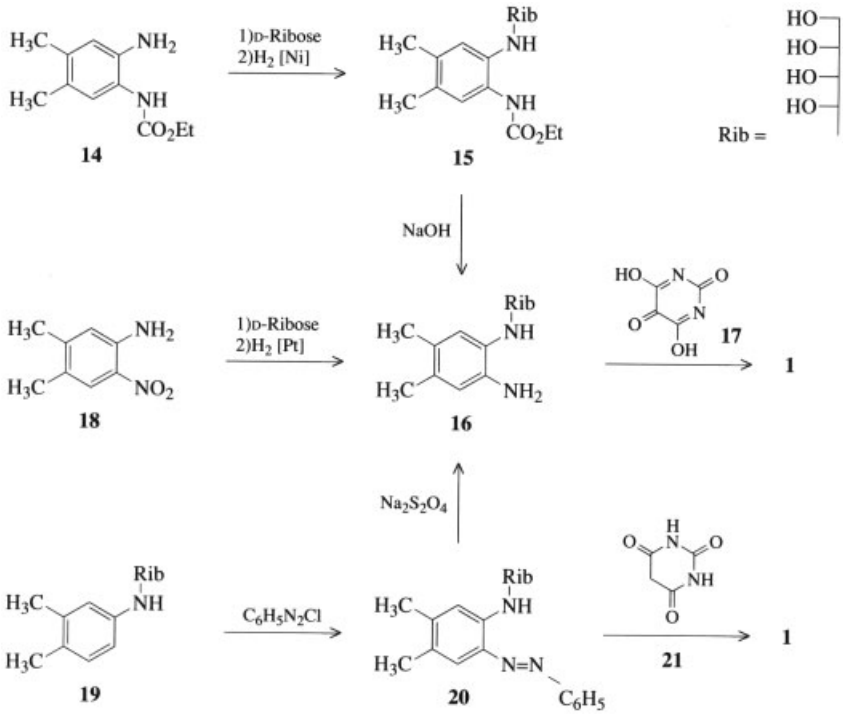

Scheme 22. Synthesis of riboflavin

methylphenyl)-*N*-D-1′-ribitylamino)uracil (**27**). Nitrosation with nitric acid and reduction of the resulting *N*-oxide (**28**) with sodium dithionite yielded riboflavin (**1**) in a total yield of ca. 70 % [724].

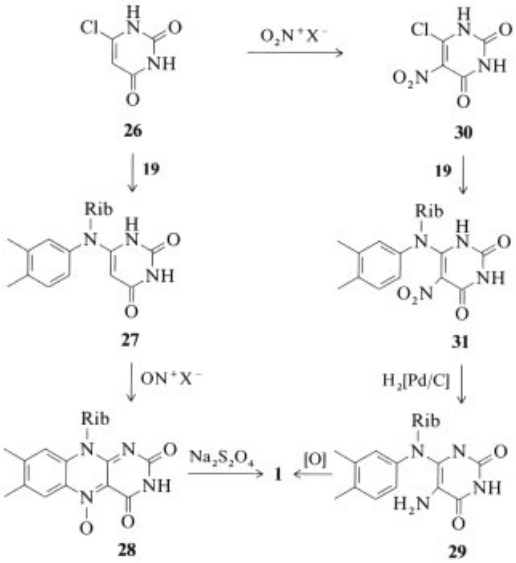

Riboflavin (**1**) was also obtained in good yield by autoxidation of the amine (**29**), which was made from **19** and 6-chloro-5-nitrouracil (**30**), followed by catalytic hydrogenation of the nitro compound (**31**) [725]. Nevertheless, these syntheses have not yet attained economic importance, mainly because of the multistep production of the uracils (**26**) and (**30**) from barbituric acid (**21**).

7.6.2. Industrial Chemical Production

Commercial syntheses of riboflavin require economic pathways for the production of D-ribose (**32**). Three processes, all starting with D-glucose [*492-62-6*] (**33**) should be mentioned here.

Oxidation of D-glucose with oxygen in aqueous potassium hydroxide solution [726] yields potassium arabonate (**34**) in ca. 90 % yield. **34** is converted to crystalline calcium arabonate (**35**). Heating **35** in aqueous solution to ca. 140 °C leads to an equilibrium mixture of epimeric salts, containing ca. 30 % calcium D-ribonate (**36**).

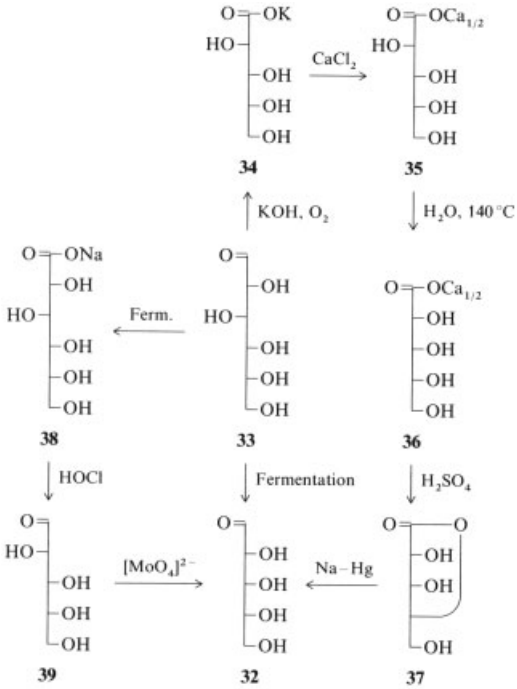

After separation of high-molecular colored side products by adsorption on activated carbon, **35** is isolated by crystallization [727], the mother liquor is acidified with sulfuric acid, and the poorly soluble calcium sulfate is separated. The D-ribonolactone [*5336-08-3*] (**37**) in the mother liquor is purified by crystallization from *n*-butanol and then reduced with sodium amalgam in aqueous solution in the presence of boric acid [728]. The sodium hydroxide solution formed is neutralized with sulfuric acid, sodium sulfate being precipitated with methanol and separated. A D-ribose solution is obtained which contains low amounts of starting material and ribitol and is directly further processed. The total yield of D-ribose (**32**) is 40 %, based on **33**.

In a more recent chemical process, D-ribose (**32**) is made from D-glucose (**33**) in three steps with a total yield of ca. 56 % [729]. **33** is fermentatively oxidized to sodium D-gluconate [*527-07-1*] (**38**), which is, in turn, oxidatively degraded to D-arabinose [*28697-53-2*] (**39**) with sodium hypochlorite. After separation of sodium chloride, **39** is crystallized from methanol by electrodialysis. Epimerization of **39** with molybdic acid in aqueous solution gives an equilibrium mixture containing ca. 25 % D-ribose (**32**), from which the starting material can be separated by crystallization and recycled [730]. The resulting D-ribose mother liquor can be used without further purification for the synthesis of riboflavin (**1**). In an advantageous process variant, molybdic acid is fixed to a basic ion exchanger and the process is conducted continuously [731].

D-Ribose (**32**) is obtained directly from D-glucose (**33**) by fermentation with microorganisms of the genus *Bacillus* which have no transketolase activity and do not form spores [732]. Especially productive are genetically engineered strains which overexpress the gluconate operon [733]. After separation of the biomass, decoloration, desalting, and evaporation of the filtrate, **32** is obtained by crystallization from water – ethanol mixture in a yield of ca. 45 %.

In the commercial synthesis of riboflavin, the three step reaction sequence, **32** → **19** → **20** → **1**, conceived by KARRER and TISHLER, is used. The total yield has been increased to >60 % by optimizing the original reaction conditions.

The readily available compound 3,4-xylidine is reacted with D-ribose (**32**) in methanol. The resulting riboside is catalytically hydrogenated to give *N*-(3,4-dimethylphenyl)-D-1′-ribamine (**19**) without isolation, and the product is purified by crystallization.

Coupling of **19** with phenyl diazonium salt in a buffered aqueous solution yields crystalline *N*-(2-phenylazo-4,5-dimethylphenyl)-D-1′-ribamine (**20**). Dry **20** is converted to riboflavin (**1**) by cyclocondensation with barbituric acid in a weakly acidic medium. Mixtures of dioxane and acetic acid or di-α-substituted carboxylic acids have proved suitable as solvents [734]. Crude **1** can be purified by reprecipitation from hydrochloric acid or dilute sodium hydroxide solution, possibly with the addition of hydrogen peroxide [735, 736].

7.6.3. Industrial Production by Fermentation

Riboflavin (**1**) can be produced microbiologically. At present, the microorganisms *Bacillus subtilis*, the ascomycetes *Eremothecium ashbyii*, *Ashbya gossypii*, and the yeasts *Candida flareri*

and *Saccharomyces cerevisiae* are used [737–740]. The nutrient media employed are molasses or plant oils as carbon source, inorganic salts, amino acids, animal or plant peptones and proteins, as well as vitamin additives. In a sterile aerobic submerged process, yields much higher than 10 g of riboflavin per liter of culture broth are obtained in a few days with good aeration and stirring at temperatures below 30 °C.

After separation of the biomass, evaporation and drying of the concentrate, an enriched product with a vitamin B_2 content of up to 80 % is obtained. The majority of the riboflavin produced by fermentation is used in animal feeds. The companies ADM (United States) and BASF (Germany) produce riboflavin microbiologically.

7.6.4. Syntheses of Flavin Mononucleotide (FMN)

Flavin mononucleotide, FMN, (**2**) was first produced by KUHN by reacting riboflavin (**1**) with phosphorus oxychloride [741]. Of the numerous phosphorylation reagents tested [742], partially hydrolyzed phosphorus oxychloride has been used most often [743]. A complex mixture of various riboflavin phosphates is always produced in this process. These can be separated by RP-HPLC and identified with the aid of authentic preparations [744].

In a commercial process employed since 1954 by Hoffmann-La Roche [745, 746], after phosphorylation, repeated crystallization is used to separate **2** as the diethanolamine salt from isomeric phosphates and unreacted **1**. Resalting with sodium acetate and crystallization from water – ethanol yields FMN (**2**) in yields of >80 %.

In a more recent process, unreacted riboflavin (**1**) is separated on polar adsorption resins made by the copolymerization of acrylic acid derivatives with divinylbenzene [747]. Subsequent fine purification can be achieved by evaporative crystallization or by chromatography on RP silica gel with water – methanol. Products purified by crystallization contain ca. 75 % FMN (**2**) and ca. 10 % 4′-FMN. Chromatographically purified preparations exhibit a ca. 90 % content of active ingredient but still contain ca. 5 % of 4′-FMN.

7.6.5. Syntheses of Flavin Adenine Dinucleotide (FAD)

The synthesis of flavine adenine dinucleotide FAD (**3**) comprises the condensation of FMN (**2**) with adenosine 5′-phosphate (AMP). The main problem is preventing formation of symmetrical phosphate esters; this is achieved by using activated AMP. The compounds morpholid [748], imidazolid [749], and products of reaction of AMP with α-pyridone [750] or with diphenylphosphoryl chloride [751] are used as activated derivatives of AMP. FAD yields of up to 70 % are attained, depending on the method employed.

FAD can also be isolated from the mycelium of cultures of *Eremothecium ashbyii* [752].

FAD can also be produced from riboflavin and AMP by fermentation with genetically engineered *E. coli* strains that overexpress the enzymes flavokinase and FAD synthetase [753, 754] (Scheme 23).

Table 26. Flavoproteins and their functions

Enzyme	Coenzyme	Function
Acyl-CoA dehydrogenase	FAD	first dehydrogenation step in the β-oxidation of fatty acids
Electron transferring flavoprotein (ETF)	FAD	transfer of hydrogen from acyl-CoA dehydrogenase to ubiquinone
Xanthine oxidase	FAD	oxidation of hypoxanthine to xanthine and xanthine to uric acid in purine degradation
Succinate dehydrogenase	FAD	dehydrogenation of succinate to fumarate in the citric acid cycle
Dihydroliponamide dehydrogenase	FAD	transfer of hydrogen from dihydroliponamide to NAD in the 2-oxo acid oxidase system
NADH cytochrome c reductase	FMN	transfer of hydrogen from NADH to ubiquinone in the respiratory chain
Monoamino oxidase	FAD	oxidation of monoamines and diamines
Aldehyde oxidase	FAD	oxidation of aldehydes to carboxylic acids
Glutathione reductase	FAD	reduction of oxidized to reduced glutathione

7.7. Importance for the Organism

In the human and animal organism, riboflavin acts as the coenzyme of a series of oxidation and reduction enzymes classified as

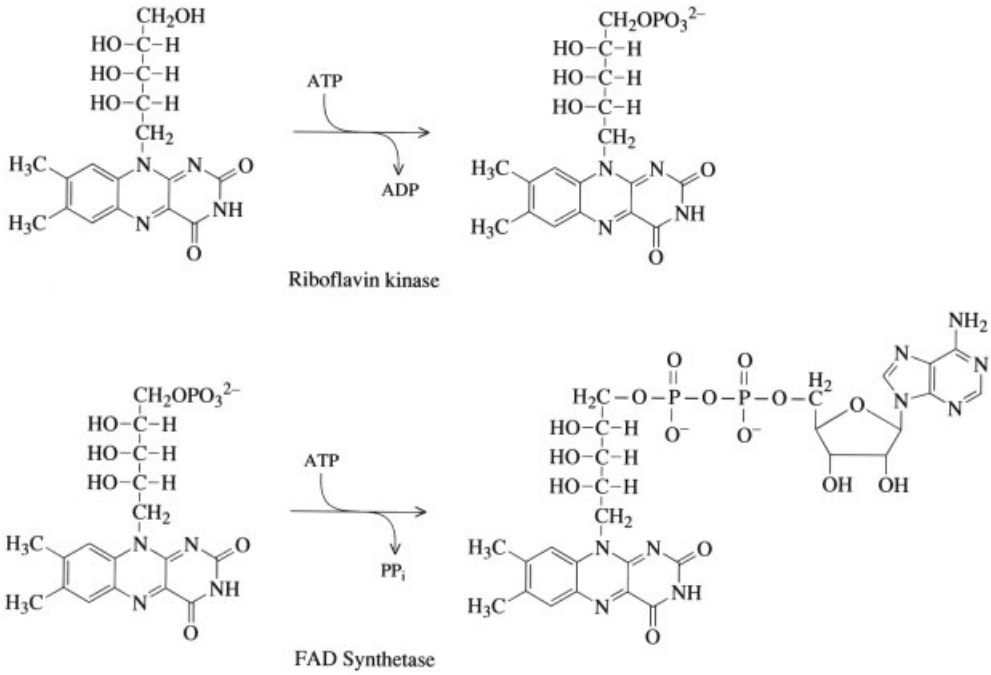

Scheme 23. Fermentative production of FAD

flavoproteins. More than 100 flavoproteins are known and more than 40 are found in the human organism. Examples are succinate dehydrogenase, lactate oxidase, glutathione reductase, and NADH/NADH$_2$ cytochrome c reductase (see Table 26).

Riboflavin is specifically bound to these enzymes as flavin mononucleotide (FMN) or flavin adenine dinucleotide (FAD) via the 8-methyl group of isoalloxazine.

In energy producing metabolism, flavoproteins act as intermediate hydrogen carriers in the oxidation of substrates and are reoxidized by enzymes with a higher oxidation potential (e.g., cytochrome c). In this process, the hydrogen or electron acceptor and donor are the coenzymes FAD or FMN [685, 755]. Since flavoproteins are involved in intermediary metabolism in diverse ways, riboflavin deficiency in the diet has an effect on many processes and causes deficiency symptoms.

Riboflavin is resorbed exclusively in the free form, mainly in the small intestine. After resorption, riboflavin is rephosphorylated to FMN by riboflavin kinase in the mucosa cells. FMN is then converted to FAD, primarily in the liver.

The elimination of riboflavin proceeds predominantly via the kidneys (urine).

7.8. Requirements, Deficiency Symptoms, and Therapeutic Application [686]

To prevent clinical deficiency symptoms in adults, 0.8 – 0.9 mg of riboflavin per day are generally sufficient. However, this amount cannot be considered adequate in the sense of an optimal supply.

The German Society for Nutrition (1991) recommends the following amounts: women 1.5 mg, men 1.7 mg, pregnant women 1.8 mg, nursing mothers 2.3 mg.

In humans, the most characteristic symptoms of a diet deficient in riboflavin (ariboflavinosis) are rhagades and fissures at the corners of the mouth and lips (cheilosis), atrophic glossitis, magenta colored tongue, seborrheic dermatitis in the facial area, and dystrophy of the finger nails. Of the diverse eye problems, conjunctivitis, corneal vascularization with the feeling of a foreign body and photophobia are noteworthy

[687]. Neurological symptoms such as paresthesia in the legs, a burning sensation in the feet, ataxia, and tremor have been observed. On controlled administration of only 0.5 mg of riboflavin per day, deficiency symptoms occur after 4 – 10 months [756].

Although clinical B_2 deficiency symptoms are now practically unknown in the industrial countries, the supply of this vitamin in all groups of society cannot be regarded as satisfactory. The 1991 national study on consumption and the nutritional report of 1992 show that the riboflavin intake of a large percentage of young girls and women in Germany is much less than the amounts recommended by the German Society for Nutrition.

Proven applications of vitamin B_2 are in the prophylaxis and treatment of clinical riboflavin deficiencies. These can have various causes:

1) Inadeqate and false nutrition
2) Increased requirement (e.g., during pregnancy and lactation)
3) Malabsorption
4) The result of phototherapy in newborn hyperbilirubinemia

In prophylaxis, vitamin B_2 is generally administered in daily doses of 1 – 5 mg, usually in the form of multivitamin or vitamin B complex tablets. In the treatment of an established ariboflavinosis, daily doses of up to 25 mg are administered.

In the food industry, riboflavin is used on a large scale together with other vitamins in the fortification of various foods, such as flour, baby foods, cereals, and multivitamin juices.

Riboflavin is of great importance in animal feeds, especially as an additive to poultry and pig feed. In these animals, a lack of riboflavin causes growth retardation, poor utilization of feed, and diarrhea. For laying hens, which have a greatly increased riboflavin requirement because of the high production of eggs, the addition of 5 mg per kilogram of mixed feed is recommended. The typical manifestation of riboflavin deficiency in the chicken is toes bent inwards.

7.9. Analysis [688, 689]

Microbiological, enzymatic, and physicochemical methods are used for the analytical identification of flavins. The growth tests conducted earlier with experimental animals on deficiency diets no longer play a role because they are time consuming and inaccurate.

Microbiological methods are especially suitable for the determination of low concentrations in biological media and in pharmaceutical preparations. They have the advantage of high specificity and relatively high accuracy. They are based on the effect of riboflavin on the growth of certain microorganisms, e.g., the unicellular organism *Tetrahymena pyriformis* or the lactic acid bacterium *Lactobacillus casei*. In the widely used assay with *Lactobacillus casei* [690], either the turbidity produced or the lactic acid liberated is used as the measure of the riboflavin concentration. If FMN or FAD is present, prior release of riboflavin is necessary.

The enzymatic method is based on the increase in the activity of glutathione reductase on addition of the cofactor FAD. This method provides an indication of the riboflavin status via the determination of the degree of saturation in the blood. It is a suitable method for the diagnosis of riboflavin deficiencies [757].

Physicochemical methods include the photometric measurement of the yellow color of riboflavin and the measurement of the characteristic yellow-green fluorescence. In both cases, riboflavin is determined either directly or as lumiflavin (**4**), which is formed on irradiation. Polarography is also employed for determination.

In the direct photometric method, the absorption is measured at 445 nm, the maximum exhibited by the yellow color of riboflavin. Since the specific extinction ($\varepsilon^{1\,\%}_{1\,\mathrm{cm}} = 323$) is relatively low, this method is not very sensitive. In addition, it requires the absence of other yellow substances. Therefore, this method is restricted to the measurement of high-percentage products without interfering components [688].

The fluorometric method is more sensitive. It involves the measurement of the fluorescence with a maximum at 530 nm which appears on irradiation with UV light. Its intensity is constant in the pH range 3 – 5 and depends only on the concentration of riboflavin. In some cases, interference by accompanying substances with a similar fluorescence can be eliminated by quenching the fluorescence of riboflavin with $Na_2S_2O_4$. This method is suited to the determination of ri-

boflavin in pharmaceutical preparations and in biological media [688].

In the lumiflavin method, riboflavin is converted to lumiflavin (**4**) by irradiation. This compound can be extracted by shaking with trichloromethane and measured either photometrically at 450 nm or fluorometrically at 513 nm.

The specificity and the sensitivity, especially of the fluorometric lumiflavin method, are higher than those of the direct methods. For this reason, this method is used more frequently [688].

Riboflavin can also be determined by using a straightforward polarographic method [758].

High-pressure liquid chromatography (HPLC) is the most important method for the analysis of riboflavin, FMN, and FAD. HPLC is the fastest and most sensitive method, especially for the determination of these components in complex mixtures, such as foods, and in the analysis of fermentation samples [759–761].

7.10. Economic Aspects

The present worldwide requirement of riboflavin is ca. 2400 t/a. Of this, ca. 70 % is used as feed additive, 20 % for pharmaceutical purposes, and 10 % in foods.

The world market price has remained fairly constant over the past ten years, at ca. 45 €/kg for feed quality and ca. 54 €/kg for pharmaceutical quality.

Producers of riboflavin are Hoffmann-La Roche (Switzerland), BASF (Germany), ADM (United States), and Takeda (Japan). There are also local producers in the Commonwealth of Independent States and in China.

7.11. Tolerance

Riboflavin causes no pharmacological side effects in animals or human beings and can be regarded as completely nontoxic. The administration of oral doses of up to 10 g/kg to rats and 2 g/kg to dogs produced no acute toxicity. Intraperitoneal administration to rats gave an LD_{50} value of 560 mg/kg.

8. Vitamin B_6

8.1. Introduction

Vitamin B_6 is a water-soluble vitamin. The term vitamin B_6 refers to a group of six compounds that are chemically closely related and, under physiological conditions, biochemically (enzymatically) interconvertible. These compounds (**1**) have in common the structural element 2-methyl-3-hydroxypyridine and bear two other substituents in positions 4 and 5 (Table 27).

Table 27. The vitamins B_6

R^1	R^2	Name
CH_2OH	OH	pyridoxine (**2**)
CH_2OH	OPO_3H_2	pyridoxine 5′-phosphate (**3**)
CHO	OH	pyridoxal (**4**)
CHO	OPO_3H_2	pyridoxal 5′-phosphate (**5**)
CH_2NH_2	OH	pyridoxamine (**6**)
CH_2NH_2	OPO_3H_2	pyridoxamine 5′-phosphate (**7**)

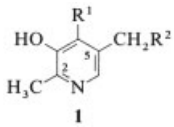

1

In 1973, the IUPAC – IUB Commission on Biological Nomenclature [762] proposed a change in the numbering of the pyridine ring. Accordingly, the common structural unit (**1**) should be called 5-hydroxy-6-methylpyridine. Until now, however, this proposal has been largely ignored in the literature (especially with regard to the 5′-phosphates which should now be named 3′-phosphates).

The commercial form of vitamin B_6 is pyridoxine hydrochloride [*58-56-0*] (**8**), which is more stable than the other compounds in this series.

8.2. History

In 1926, a syndrome similar to pellagra (pellagra is characterized by skin changes and disturbances of the nervous system, digestion, growth, fertility, blood count, etc.) was induced in rats on

a diet deficient in "vitamin B complex" [763]. This syndrome was called rat acrodynia. From 1934 to 1936, GYÖRGY et al. [764] showed that vitamins B_1 and B_2, which were both known at that time, could not cure this rat dermatitis. They showed that another vitamin, designated antidermatitis vitamin, adermin, or vitamin B_6, must be responsible for the cure. Based on GYÖRGY's exact animal model tests, the isolation (from yeast and rice bran) and structural elucidation of pyridoxine hydrochloride (**8**) were achieved simultaneously by five groups in 1938 [765].

Pyridoxine hydrochloride (**8**) was synthesized in 1939 by two teams working independently [766]. Between 1942 and 1945, SNELL et al. showed by using microbiological experiments that vitamin B_6 is not a homogeneous substance, but a group of six interconvertible compounds **2**–**7** [767].

Although **8** was isolated as early as 1932 [768], its vitamin function was not recognized at that time.

8.3. Physical Properties

A compilation of the physical properties of **2**–**8** is given in [769–772]. The properties of the commercial product pyridoxine hydrochloride (**8**), including its spectra, are described in detail in [770, 771]. All vitamin B_6 compounds are sensitive to light and, as pyridine bases, form salts with acids. They decompose when heated for longer periods in neutral or basic aqueous solutions.

Pyridoxine [*65-23-6*] (**2**), (CA: 3,4-pyridinedimethanol, 5-hydroxy-6-methyl), 2-methyl-3-hydroxy-4,5-bis(hydroxymethyl)pyridine, $C_8H_{11}NO_3$, M_r 169.18. Colorless needles (from acetone), *mp* 159 – 160 °C, soluble in water, methanol, ethanol, and acetone; thermolabile as solid and in solution.

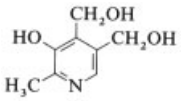

2

Pyridoxine hydrochloride [*58-56-0*] (**8**), (CA: 3,4-pyridine-dimethanol, 5-hydroxy-6-methyl, hydrochloride), $C_8H_{12}ClNO_3$, M_r 205.64.

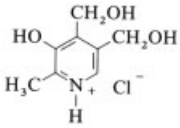

8

Colorless platelets or rods from ethanol – acetone, *mp* 205 – 212 °C (decomp.). Soluble in water (22 g/100 mL at room temperature), ethanol (1.1 g/100 mL), and propylene glycol. Poorly soluble in methanol and acetone, and insoluble in ether and chloroform. Aqueous solutions are stable at room temperature (120 °C for at least 30 min). The pH of a 10 % aqueous solution is 3.2.

Pyridoxine 5′-phosphate [*447-05-2*] (**3**), [CA: 3,4-pyridine-dimethanol, 5-hydroxy-6-methyl, α-3-(dihydrogen phosphate)], $C_8H_{12}NO_6P$, M_r 249.16.

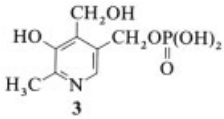

3

White needles from concentrated aqueous solutions, *mp* 212 – 213 °C (decomp.), soluble in water, and insoluble in ethanol.

Pyridoxal [*66-72-8*] (**4**), (CA: 4-pyridinecarboxaldehyde, 5-hydroxy-3-(hydroxymethyl)-6-methyl), 3-hydroxy-5-hydroxymethyl-2-methyl-4-pyridinecarboxaldehyde, $C_8H_9NO_3$, M_r 167.16.

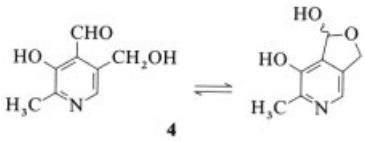

4

The aldehyde form is in equilibrium with the cyclic hemiacetal form. Readily decomposing compound, independent of pH.

Pyridoxal hydrochloride [*65-22-5*] (**9**), (CA: 4-pyridinecarboxaldehyde, 5-hydroxy-3-(hydroxymethyl)-6-methyl, hydrochloride), $C_8H_{10}ClNO_3$, M_r 203.63.

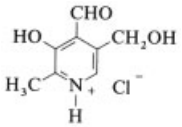

9

Almost colorless rhombic crystals, *mp* 177–181 °C (decomp.), soluble in water (50 g/100 mL) and 96 % ethanol (1.7 g/100 mL), insoluble in acetone, ether, and chloroform. In solution, **9** is less stable than **8**.

Pyridoxal 5′-phosphate [*54-47-7*] (**5**), (CA: 4-pyridinecarboxaldehyde, 5-hydroxy-6-methyl-3-[(phosphonooxy)-methyl]), $C_8H_{10}NO_6P$, M_r 247.14.

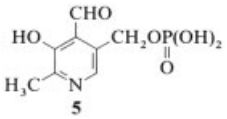

5

The monohydrate of **5** [*41468-25-1*], $C_8H_{12}NO_7P$, M_r 265.15, forms yellowish needles, *mp* 139–142 °C (decomp.), solubility in water 0.5 %, less soluble in ethanol, and insoluble in acetone and ether.

Pyridoxamine [*85-87-0*] (**6**) (CA: 3-pyridinemethanol, 4-(aminomethyl)-5-hydroxy-6-methyl), 3-hydroxy-4-aminomethyl-5-hydroxymethyl-2-methylpyridine, $C_8H_{12}N_2O_2$, M_r 168.20. Crystals, *mp* 193 °C, soluble in water and ethanol.

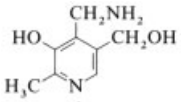

6

Pyridoxamine dihydrochloride [*524-36-7*] (**10**), $C_8H_{14}Cl_2N_2O_2$, M_r 241.12. Colorless hygroscopic platelets, *mp* 228 °C (decomposition), soluble in water (50 g/100 ml) and in 96 % ethanol (0.65 g/100 mL), insoluble in ether and chloroform.

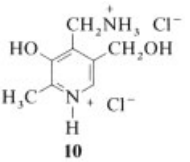

10

Pyridoxamine 5′-phosphate [*529-96-4*] (**7**) [CA: 3-pyridinemethanol, 4-(aminomethyl)-5-hydroxy-6-methyl, α-(dihydrogenphosphate)], $C_8H_{13}N_2O_5P$, M_r 248.18.

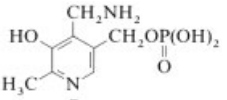

7

The dihydrate of **7** [*84878-64-8*], $C_8H_{17}N_2O_7P$, M_r 284.2, forms rhombic platelets; *mp* of the hydrochloride 240 °C (decomp.).

8.4. Chemical Reactions [770, 772, 773]

Many of the reactions and syntheses in vitamin B_6 chemistry date from the years following the discovery and isolation of pyridoxine. These reactions often aided the structural elucidation and proof of structure of compounds **2** to **8**. Since the commercial form of vitamin B_6 is pyridoxine hydrochloride (**8**), some characteristic reactions of pyridoxine (**2**) are described here (see Scheme 24).

Pyridoxine (**2**) contains three reactive hydroxyl groups. The OH group in position 3 of the pyridine ring reacts with iron(III) ions to give the red or orange color typical of aromatic hydroxyl groups. This OH group can be esterified and etherified. Together with the 4-hydroxymethyl group, six-membered cyclic ketals or acetals (**10**) can be formed with ketones and aldehydes. The two hydroxymethyl groups in positions 4 and 5 undergo reactions typical of groups of this type, e.g., ester formation. Under acid catalysis, reactions with ketones and aldehydes produce, apart from **10**, seven-membered cyclic ketals or acetals (**11**). Of the two CH_2OH groups, the more reactive is the one in position 4. For example, oxidation

of **2** with MnO_2 gives pyridoxal (**4**), which can be further oxidized by O_2 under catalysis with cyanide to give pyridoxine 4-carboxylic acid (**12**) [*82-82-6*]. Although acidic aqueous solutions of **2** are quite stable even at 100 °C, it reacts with inorganic acid halides or hydrohalic acids under drastic conditions. With HCl, for instance, depending on the reaction conditions, the compounds **13** [*5196-20-3*], **14** [*102694-13-3*], and **15** [*13983-22-7*] are formed. Acetylation can give the 4′,5′-diacetate **16** [*26280-83-1*] or the 3,4′,5′-triacetate **17** [*10030-93-0*] depending on the reaction conditions. Heating **2** with aqueous ammonia produces pyridoxamine (**6**).

8.5. Occurrence

Vitamin B_6 occurs in plants mainly as pyridoxine. Grains, potatoes, green vegetables, and legumes are especially rich in vitamin B_6. A considerable part of the vitamin occurs in plants as 5′-*O*-(*β*-glucopyranosyl)pyridoxine. Animal experiments show that this glucosylated form is apparently well absorbed but appears to be not entirely bioavailable. The main forms observed in the animal kingdom are pyridoxal and pyridoxamine. Accordingly, these forms are found in animal foods, such as meat, fish, and eggs. The concentration of vitamin B_6 in some foods is given in Table 28 [774–777].

Table 28. Vitamin B_6 content of some foods [774–777]

Food	Vitamin B_6, mg/100 g	Pyridoxine *β*-glucoside %
Potatoes	0.14 – 0.39	42
Spinach	0.208 – 0.22	50
Carrots	0.17 – 0.70	51
Cabbage	0.14 – 0.27	46
Bananas	0.26 – 0.313	3
Legumes	0.381 – 0.57	42
Meat (chicken beef, pork)	0.30 – 0.70	0
Fish	0.316 – 0.45	0
Eggs	0.19 – 0.25	

8.6. Biosynthesis

Pyridoxine is synthesized in bacteria, fungi, and plants. Experiments with radioactively labelled or deuterated compounds, such as glucose and glycerol, contributed to the elucidation of the biosynthesis of vitamin B_6. Hill et al. found later that D-1-deoxyxylulose is incorporated into the pyridoxine ring and supplies the carbon atoms C 2, 2′, 3, 4, and 4′ [778,779]. According to Dempsey, 4-hydroxythreonine is a central intermediate in the biosynthesis of pyridoxine [780]. Lam and Winkler reported that 4-hydroxythreonine is made from erythrose 4-phosphate in four enzymatic steps and subsequently condensed with D-1-deoxyxylulose to give pyridoxine [781]. The enzymes that catalyze these reactions are as yet not completely known. In *E. coli*, the corresponding genes are spread over the entire chromosome and organized in complex operons. At present, biotechnological production by genetically manipulated microorganisms or the industrial isolation from natural sources is not yet of interest.

8.7. Production of Vitamin B_6 Compounds

8.7.1. Pyridoxine [770,772,782]

Pyridoxine hydrochloride (**8**) is the form of vitamin B_6 that is produced on a large scale. It is mainly used in pharmaceutical preparations and as an additive in food and feed. The other B_6 compounds, pyridoxal (**4**), pyridoxamine (**6**), and the accompanying 5′-phosphates (**3**), (**5**), and (**7**) can be made from **8**.

The five synthetic pathways known for **2** and **8** are outlined here in the order of their discovery. At first, **8** was produced entirely by the methods described in Section 8.7.1.2. Today, the methods described in Section 8.7.1.4. are exclusively employed. Intensive work on these synthetic pathways began in 1939 and 1957 and led to many patents because of the commercial interest in B_6. Since the expiry of all the basic patents on these processes, the number of new patents in this field has clearly decreased in the past years.

8.7.1.1. Oxidative Degradation of Bicyclic Heterocyclic Compounds [773]

In 1939, the teams of Kuhn [766] and Ichiba [783] published the first syntheses of **2** starting with suitably substituted isoquinoline or

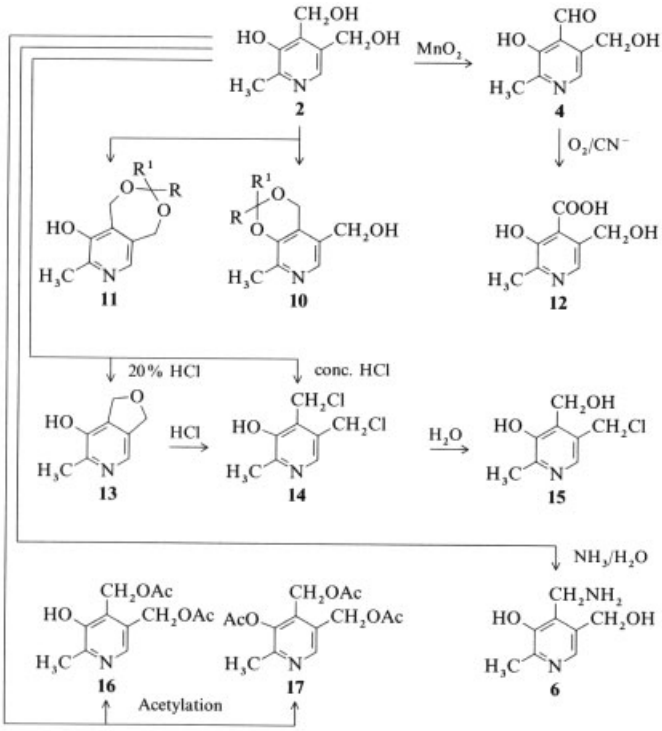

Scheme 24. Reactions of pyridoxine (**2**)

quinoline derivatives [784], e.g., 4-methoxy-3-methylisoquinoline [*103988-78-9*] (**18**).

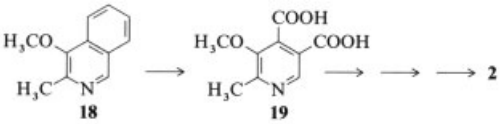

The synthesis starts with the oxidative degradation of the carbon ring to give the dicarboxylic acid (**19**) [*7442-22-0*], which is subsequently converted in several stages to **2**. The commercialization of syntheses of this type was prevented by the poor availability of the starting materials, the multiple stages involved, and the still unsolved problem of finding an inexpensive reduction of such carboxyl functions to alcohols.

8.7.1.2. Condensation Reactions with Aliphatic Precursors [770, 773, 782]

In 1939, Harris and Folkers [785] also published a synthesis of **2** which was subsequently developed into the first commercial process. The reaction of ethoxyacetylacetone [*20754-01-2*] (**20**) with cyanoacetamide [*107-91-5*] (**21**) gives the pyridone (**22**) [*17718-67-1*].

Following the same synthetic principle, the reaction of 2-amino-4-oxo-5-ethoxypent-2-ene and malodinitrile provides improved commercial

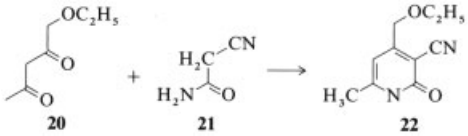

access to **22**. This method was described by Schnider as early as 1941 [786].

The conversion of **22** to **2** occurs in many stages and involves relatively heavy losses. In the following 20 – 25 years, a number of synthetic pathways for **2**, which use many different aliphatic building blocks, were described. A survey is given in [773].

8.7.1.3. From Furan Compounds

Synthesis of 3-hydroxypyridine from 2-aminomethyl-2,5-dimethoxydihydrofuran [787] was published in 1950 and used in the mid-1950s to synthesize **2** [788]. In the key step, *N*-[1-[3,4-bis[acetyloxymethyl]-2-furanyl]ethyl]acetamide [*84878-62-6*] (**23**) is converted to its 2,5-dimethoxy derivative (**24**) [*84878-63-7*] by anodic oxidation in methanol. Compound **24** is then converted to pyridoxine (**2**) by alkaline saponification and acidification.

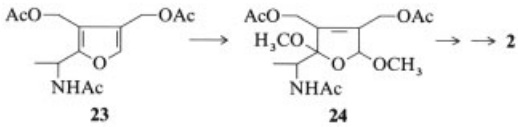

Even after the availability of the starting compounds was improved, the cost of the starting materials, the multiple stages, and the modest total yield of the synthesis hindered its commercialization [789, 790].

8.7.1.4. Diels – Alder Syntheses with Oxazoles [770, 772]

This synthetic principle described by Kondratyeva in 1957 [791] is now used exclusively for the industrial production of pyridoxine (**2**). She found a new synthesis for pyridines via the Diels – Alder reaction of oxazoles (**25**) with maleic acid (**26**) (or derivatives). Under the reaction conditions, the Diels – Alder adduct (**27**) is aromatized to the cinchomeronic acid derivative (**28**) by the elimination of water.

An important discovery was that if good leaving groups Z are incorporated into position 5 of 4-methyloxazole (**29**), after reaction with the dienophile (**30**), the bridging oxygen atom of the bicyclic Diels – Alder adduct remains in the subsequent product and forms an OH group in position 3 of pyridine (**31**) [792].

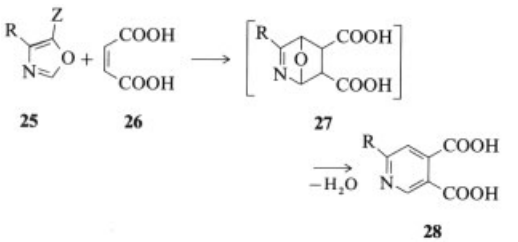

For good results, these oxazoles should exhibit low steric hindrance and high thermal stability.

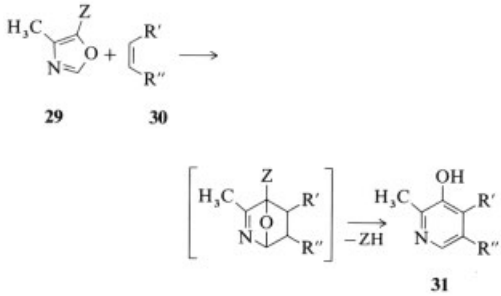

Diene Components. From the extensive patent literature, it appears that the compounds that are industrially applied are mainly 5-ethoxy-4-methyloxazole (**29**, Z = OEt) and 5-cyano-4-methyloxazole (**29**, Z = CN).

The starting material for the production of 5-ethoxy-4-methyloxazole [*5006-20-2*] (**34**) is D,L-alanine ethyl ester (**32**) [*17344-99-9*], its *N*-formyl derivative (**33**) [*4289-99-0*] being dehydrated to **34** by P_2O_5 [793].

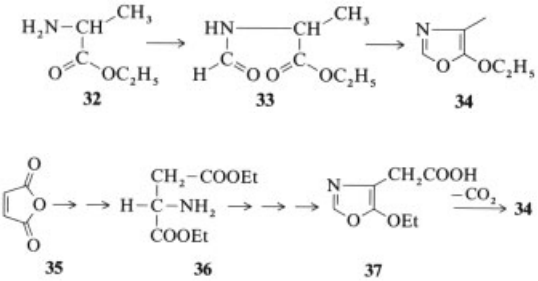

Attempts were made to find alternatives to the formerly expensive D, L-alanine and the oxazole (**34**), which is patented. One synthesis started with inexpensive maleic anhydride (**35**) and produced D, L-aspartic acid [*617-45-8*] by NH_3 addition to the double bond. The diethyl ester of aspartic acid (**36**) [*43101-48-0*] was then used to produce 5-ethoxy-4-oxalacetic acid [*10500-75-1*] (**37**) via a known reaction sequence (*N*-formylation and dehydration) and saponification [794]. Under Diels – Alder reaction conditions, compound **37** is decarboxylated to **34**.

Another precursor of **34** is 5-ethoxy-4-methyl-oxazole-2-carboxylic acid [*23429-05-2*] (**40**), which is also decarboxylated to **34** in the Diels – Alder reaction. Here, propionic acid (**38**) is halogenated and converted to D,L-alanine with ammonia. Esterification to give **32** and reaction with ethyl oxalate gives the intermediate **39** [*23460-73-3*] which on cyclocondensation and saponification produces **40** [795].

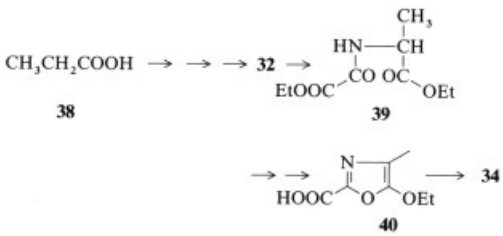

The starting material for 4-methyl-5-cyanooxazole [*1003-52-7*] (**41**) is diketene, which is converted to ethyl acetoacetate with alcohol. Chlorination produces chloroethyl acetoacetate (**42**), which is reacted with formamide (**43**) to give 4-methyl-5-oxazolecarboxylic acid ethyl ester [*20485-39-6*] (**44**), which is dehydrated to the nitrile (**41**) via the 4-methyl-5-oxazole carboxamide (**45**) [*4866-00-6*]. This dehydration can be conducted with P_2O_5/quinoline [796], acetic anhydride [797], in a catalytic gas-phase reaction [798], or under very mild condition with cyanuric chloride [799].

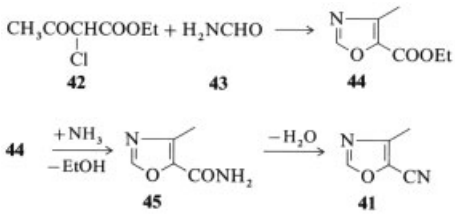

Dienophiles. Derivatives of maleic acid or fumaric acid (e.g., their esters or nitriles) react as dienophiles with the oxazoles used as dienes in the Diels – Alder reaction (e.g., **34** or **41**), giving very good results. However, the above-mentioned expensive reduction steps of the resulting pyridoxine dicarboxylic acid (derivatives) have prevented industrial application.

With regard to the oxidation state, 2-butene-1,4-diol would be the ideal dienophile if the oxazole diene **34** or **41** is used. However, the first-named compound reacts very poorly. Its openchain derivatives, e.g., the ether or ester, react better but are still fairly unreactive. (Nevertheless, (*Z*)-2-butene-1,4-diol-diacetate [*25260-60-0*] appears to have been used in a production process by a Japanese company). In contrast, the cyclic seven-membered acetals and ketals of (*Z*)-2-butene-1,4-diol [*6117-80-2*] (**45**) [800] – e.g., the isopropyldioxepin (**47**) [*5417-35-6*] obtainable from isobutyraldehyde (**46**) [*78-84-2*] – have proved to be sufficiently reactive and stable in an inert-gas atmosphere. The compound *n*-propyldioxepin (**48**), derived from *n*-butyraldehyde, is also a suitable reagent.

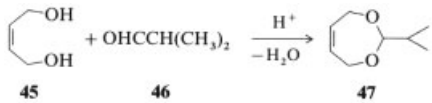

Diels – Alder Reaction. The reaction of the oxazole (**34**) or (**41**) with the dioxepin (**47**) or (**48**) requires high temperatures (ca. 115 – 180 °C) and longer periods of time (ca. 10 – 20 h or more). The primary Diels – Alder adducts formed have three chiral carbon centers, resulting in mixtures of eight stereoisomers in the form of four diastereomeric pairs. Excess or unreacted starting material can be separated and recycled.

The use of 5-cyano-4-methyloxazole (**41**) and isopropyldioxepin (**47**) results in a Diels – Alder adduct that is not stable under the reaction conditions and decomposes with the elimination of hydrogen cyanide and aromatization to give **49** [*1622-67-9*] in > 90 % yield [800].

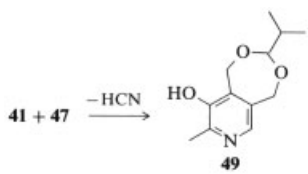

5-Ethoxy-4-methyloxazole (**34**) reacts with the dienophile (**47**) under somewhat milder conditions than **41**. In this case, the Diels – Alder adduct is much more stable and can be isolated (yield > 95 %). Under mildly acidic con-

ditions, aromatization occurs with the elimination of ethanol to give the bicyclic acetal (**49**) [772]. Heating **49** with hydrochloric acid results in almost quantitative conversion to isobutyraldehyde (**46**), which is separated and reused, and pyridoxine hydrochloride (**8**).

The above reactions of the oxazoles (**34**) and (**41**) with the dienophile (**47**) are "inverse" Diels – Alder reactions in which electron-deficient dienes react with an electron-rich dienophile. The variant corresponding to the "normal" Diels – Alder reaction has also been published [801]. Here, 4-methyloxazole [*693-93-6*] (**50**) reacts with 2,5-dihydro-3-(methylsulfonyl)furan [*41409-84-1*] (**51**).

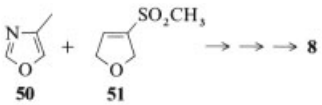

In this combination, the dienophile and not the oxazole carries a good leaving group in the form of a sulfonyl group. This group makes the Diels – Alder reaction possible and is also required for the aromatization of the Diels – Alder adduct. However, this synthesis has problem steps which make industrial application doubtful.

Another Diels – Alder reaction for the synthesis of pyridoxine uses 1,2,4-triazine derivatives as dienes instead of oxazoles [802]. However, the elaborate synthesis of the triazines makes this synthetic principle appear totally uninteresting for an industrial process. Since all the basic patents for the various oxazole routes to **2** have expired and the price of D,L-alanine has fallen considerably in the past years, it appears that the combination **34**+**47** is used by Takeda and Daiichi, **41**+**47** by Roche, and **40**+**48** by at least one of the Chinese producers for the industrial production of pyridoxine hydrochloride (**8**).

8.7.1.5. Cobalt-Catalyzed [2+2+2]-Cycloaddition Reactions of Acetylenes and Acetonitrile

Independently, two groups have developed a new pathway to pyridoxine using cobalt-catalyzed [2+2+2]-cycloaddition reactions of protected α,ω-diyne ethers and acetonitrile [803].

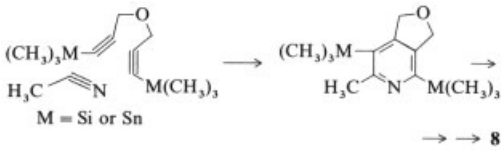

However, the yields of **8** obtained were less than 10 %. For instance, the problem of the conversion of the $(CH_3)_3Si$ group at position 3 of pyridoxine to an aromatic OH group could not be satisfactorily solved.

8.7.2. Pyridoxine 5′-Phosphate

The OH groups in positions 3 and 4 of pyridoxine (**2**) can be protected with acetone/hydrochloric acid, forming the ketal (**52**). Phosphorylation of **52** at 60 °C with H_3PO_4 – P_2O_5, followed by gentle acid hydrolysis, yields **3** [804]. Pyridoxine 5′-phosphate (**3**) can also be obtained by the reduction of pyridoxal 5′-phosphate (**5**) with $NaBH_4$ [805].

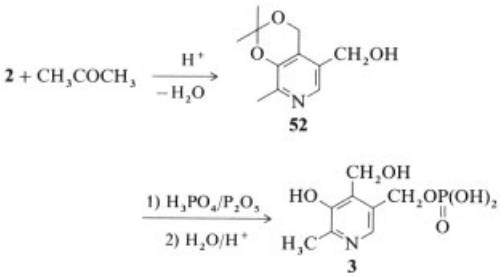

Phosphorylation of **52** is also achieved with metaphosphoric acid $(HPO_3)_3$. The resulting mixture of oligophosphates can be hydrolyzed to **3** [806]. Compound (**3**) can also be produced by treating pyridoxamine 5′-phosphate (**7**) with HNO_2 [807].

8.7.3. Pyridoxal [808]

Oxidation of pyridoxine (**2**) with permanganate in the presence of hydroxylamine produces pyridoxal oxime (**53**), which can be converted to

pyridoxal (**4**) by treatment with nitrite and hydrochloric acid [809].

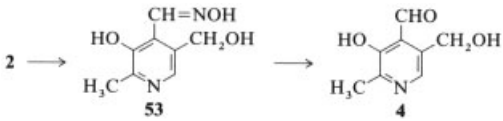

Direct oxidation of **2** to **4** can be conducted with $MnO_2 - H_2SO_4$ [806,810]. Isolation of **4** from solution can be advantageously achieved via its bisulfite addition product [811]. A total synthesis of **4** from aliphatic building blocks has been described [809].

8.7.4. Pyridoxal 5′-Phosphate

Pyridoxal 5′-phosphate (**5**) ("codecarboxylase") [772] can be made from pyridoxine 5′-phosphate (**3**) [803, a (Baddiley; Targa)] or from pyridoxamine 5′-phosphate (**7**) [803, a (Wilson)] [804,805] by oxidation with MnO_2 or with pyruvic acid in the presence of Cu, Fe, Ni, or Co salts [812]. Direct phosphorylation of pyridoxal (**4**) is described in [813]. Derivatives of pyridoxal in which the aldehyde group is protected by readily removable groups can be phosphorylated and **5** released by mild hydrolysis. The following protective groups have been used: *N*-dimethylglycylhydrazone [806,814], Schiff bases of aromatic amines [815] and alkenediamines [816], oxazolidines [817], and thiazolidines [818]. Starting with pyridoxine 5′-phosphate (**3**) or pyridoxamine 5′-phosphate (**7**), pyridoxal 5′-phosphate (**5**) was made by enzymatic oxidation with immobilized pyridoxamine 5′-phosphate oxidase [819]. Starting with pyridoxine 5′-phosphate (**3**), the synthesis of **5** using immobilized bacterial cells has been described [820].

8.7.5. Pyridoxamine [772,808]

Pyridoxamine (**6**) can be produced from pyridoxal oxime (**53**) by catalytic hydrogenation with PtO_2 in acetic acid [821], Pd/C [822], or by reduction with zinc dust [823]. Pyridoxamine (**6**) is obtained from pyridoxine (**2**) by reaction with NH_3 in CH_3OH [824]. It is also produced by the reductive amination of pyridoxal (**4**) [825]. Pyridoxamine (**6**) is also synthesized in a Diels – Alder reaction of 5-ethoxy-4-methyloxazole (**34**) with γ-hydroxycrotonitrile, followed by reduction of the nitrile group in position 4 of the pyridine ring to an aminomethyl group [826].

8.7.6. Pyridoxamine 5′-Phosphate

Pyridoxamine 5′-phosphate (**7**) is obtained from pyridoxal 5′-phosphate (**5**) by reaction with glutamic acid in an autoclave [827]. Starting with pyridoxamine (**6**), **7** is made by treatment with $POCl_3$ in aqueous solution [808]. However, better results are obtained when **6** is phosphorylated by heating with metaphosphoric acid, followed by degradation of the resulting 5′-triphosphate by mild acid hydrolysis [828] or by the action of H_3PO_4/P_2O_5 at 60 °C on **6** [807,829].

8.8. Metabolism and Importance for the Organism

The three nonphosphorylated B_6 vitamins are absorbed in the region of the upper small intestine (proximal jejunum) by a nonsaturable process (passive diffusion) [830]. They are phosphorylated in the liver by a pyridoxal kinase and ATP, bound to albumin, and released into the blood [831]. As a result of the binding to albumin, pyridoxal 5′-phosphate is protected from hydrolysis in circulation and is able to reach the tissues. The two phospho forms, pyridoxine 5′-phosphate and pyridoxamine 5′-phosphate, are converted to pyridoxal 5′-phosphate (PLP) by a FMN-dependent oxidase [832]. After dephosphorylation by a membrane-bound alkaline phosphatase, pyridoxal can enter the cell through the cell wall. Pyridoxal is irreversibly converted to 4-pyridoxic acid by an NAD-dependent dehydrogenase or an FAD-dependent aldehyde oxidase. This acid is excreted in the urine as the main metabolite of B_6 metabolism [833]. The enzymatic interconversion of the vitamins B_6 are shown in Scheme 25.

Pyridoxal 5′-phosphate acts as a coenzyme for many enzymes which, above all, regulate amino acid metabolism. The coenzyme is bound to an ε-amino group of a lysine residue of

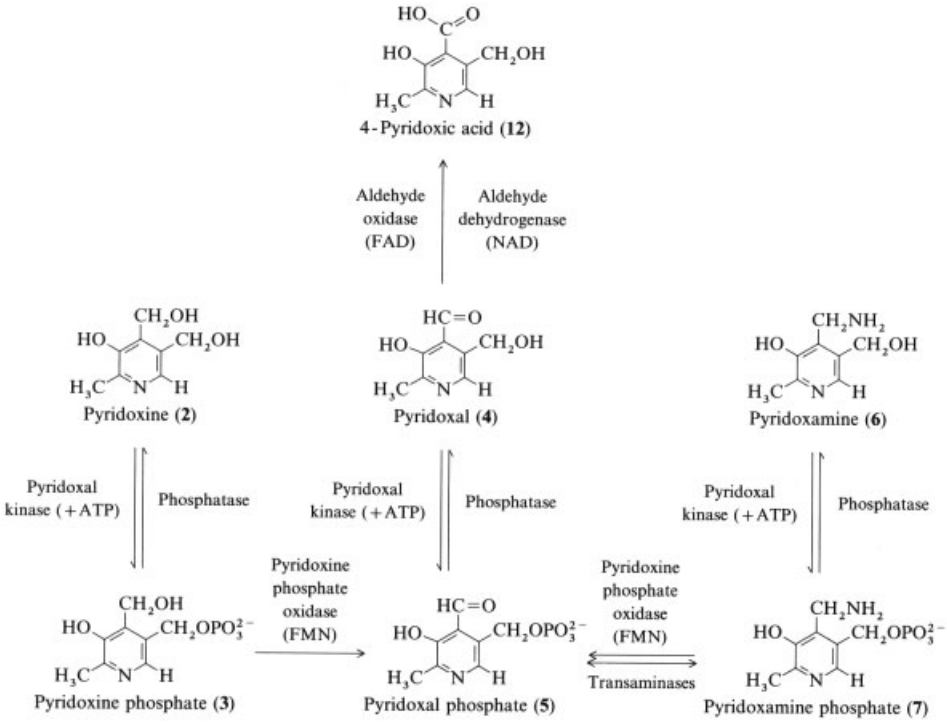

Scheme 25. Metabolic interconversion of the vitamins B_6

the apoenzyme through its formyl group. Thus, the active enzyme complex is formed via the resulting Schiff base. In this way, reactions such as transamination, nonoxidative deamination, decarboxylation, and hydrodesulfurization are catalyzed. The majority of pyridoxal 5′-phosphate (75 – 80 %) is found in glycogen phosphorylases, which are involved in gluconeogenesis in plants and animals. However, the reaction mechanism is still unknown. Furthermore, pyridoxal 5′-phosphate acts in microbial racemases which catalyze the conversion of L-amino acids to D-amino acids. In addition, it appears to have a regulatory effect on steroid hormone receptors. Under physiological conditions, pyridoxal 5′-phosphate binds reversibly to steroid receptors and can modulate or inhibit the interaction between the steroid receptor complex and the specific DNA sequence [834, 835]. The physiological importance of the interaction between PLP and steroid receptors is not yet understood.

8.9. Deficiency Symptoms and Applications

A food-related vitamin B_6 deficiency manifesting itself in clinical symptoms is largely unknown in humans. Since vitamin B_6 plays an essential role in numerous biochemical processes, a lack of this vitamin is easily detectable from the resulting metabolic disturbances. Deficiencies are produced by an unbalanced diet, interference with medicines, excessive loss due to functional kidney disorders, intestinal diseases (diarrhoea, disturbances of absorption), hereditary disorders (cystathioninuria, homocysteinuria), or neurological diseases.

Patients with liver cirrhosis (e.g., alcoholics) often have a low vitamin B_6 level in the blood [836, 837]. In 15 – 20 % of women, estrogen-containing contraceptives cause an impairment of tryptophan metabolism which can be normalized by 2 – 25 mg of vitamin B_6 per day [838]. In the treatment of tuberculosis with isoniazid, side effects appear (neuritis, convulsions) which can be prevented by 25 – 50 mg of vitamin B_6 per day [839, 840]. Treatment of the rare Wilson's

disease with D-penicillamine causes impaired tryptophan metabolism which responds well to 100 mg of vitamin B_6 per day [841]. The very low plasma values found in patients with kidney diseases (chronic kidney failure, hemodialysis) or after kidney transplantation can be normalized by administration of 5 – 50 mg of vitamin B_6 per day. Some genetic diseases result in B_6 malfunctions. Included in this category are cases of infantile convulsions caused by the poor binding of vitamin B_6 to the apoenzyme glutamate decarboxylase; primary cystathioninuria, which is marked by a greatly reduced affinity of the enzyme cystathionase for vitamin B_6; or homocysteinuria, which is caused by low amounts of cystathionine β-synthase. In general, intake of 5 – 50 mg of vitamin B_6 per day results in improvements. High doses or megadoses of vitamin B_6 are often applied in the treatment of rheumatic diseases, carpal tunnel syndrome, autism, schizophrenia, learning difficulties, and mental retardation [842, 843].

8.10. Analysis

Numerous methods have been developed for the determination of vitamin B_6. Preparations enriched with pyridoxine can be determined by using UV spectra or color reactions (e.g., the reaction between pyridoxine and diethyl-*p*-phenyldiamine). The analysis of vitamin B_6 in foods, animal organs, and body fluids is complicated. Reasons for this are the different chemical forms (pyridoxine, pyridoxal, pyridoxamine, and their phosphate esters), sensitivity to heat and light, and the covalent binding of these forms to proteins. Microbiological methods are often used for the determination of vitamin B_6. The yeast strain *Saccharomyces uvarum* is generally used for this purpose [844, 845]. Frequently, all the vitamin B_6 forms can be determined by using combinations with other microorganisms, e.g., *Lactobacillus casei*, *Streptococcus faecium*. The detection limit is 0.5 – 1.0 ng. However, these methods are very time consuming. The microorganisms can mutate or suffer growth inhibition under the influence of the extracts. Pyridoxal 5′-phosphate is enzymatically determined in the blood, serum, or in erythrocytes by the decarboxylation of amino acids [846, 847]. Pyridoxal 5′-phosphate can be indirectly determined in erythrocytes by measuring the enzyme activities of erythrocyte glutamate oxaloacetate transaminase (EGOT) and erythrocyte glutamate pyruvate transaminase (EGPT) [847–849]. Gas chromatography and HPLC have also been successfully used [850, 851].

8.11. Economic Aspects

World production of pyridoxine hydrochloride (**8**) is estimated at 2550 t/a for 1993. The main producers are Roche (Germany), Takeda (Japan), and Daiichi (Japan). In the past years, the production capacity in several plants has been drastically increased, especially in China, and the estimated world capacity of $>$ 4100 t/a for 1993 greatly exceeds demand. From 1992, the market has been increasingly supplied by Chinese companies (especially by New Asiatic, Shanghai). The list price for pyridoxine hydrochloride given in [852] is \$ 50/kg. However, the actual retail prices on the world market fluctuate and are considerably lower than the list price (as of October 1994).

The historically highest retail price for **8** was 7600 \$/kg in the year of its introduction and the lowest price was \$ 30/kg in 1967.

Trade names of vitamin B_6 and its producers are, e.g., Benadon (Roche), Hexobion (Merck), Beelith (Beach Pharmaceuticals, USA), Vitamin B_6 Richard (Lab. M. Richard, France), Vitamin B_6 Streuli (Streuli, Switzerland), B_6-Vicotrat (Heyl, Germany), Pyragamma (Wörwag, Germany), and Vitamin B_6 Jenapharm (Jenapharm, Germany). All these products contain pyridoxine hydrochloride as the active agent in doses of 20 to 300 mg, and in one case 600 mg.

8.12. Requirements and Tolerance

The requirement for vitamin B_6 is influenced by various factors, especially by the daily intake of protein. The requirement increases with increasing protein absorption. In the United States, the average daily intake of vitamin B_6 in 1985 was 1.87 mg for men, 1.16 mg for women, and 1.22 mg for children of 1 – 5 years [853]. Various methods are used to estimate the requirement, e.g., direct determination of the different

forms in the blood and/or urine (excretion of 4-pyridoxic acid), various load assays (e.g., determination of the tryptophan metabolites xanthurenic and kynurenic acid in the urine after oral tryptophan loading with 2 – 5 g), or the measurement of vitamin B_6 dependent enzymes (indirect functional determination). Since 1989, the recommended daily allowance (RDA) for vitamin B_6 in the United States has been set at 2.0 mg for men, 1.6 mg for women, 2.1 mg for pregnant and nursing women, 1.0 mg for children of age 1 – 3, 1.1 mg for 4 – 6 year old children, and 1.4 mg for children at the age of 7 – 10 [777, 853].

Pyridoxine and related compounds having vitamin B_6 activity show only low acute toxicity [854, 855]. Teratogenic and carcinogenic effects of high pyridoxine doses are unknown [856, 857]. After ingestion of excessively high doses (2 – 6 g of pyridoxine per day) over longer periods of time, peripheral neurological disorders have been observed in humans [858]. A daily intake of up to 250 mg of pyridoxine over longer periods is harmless [859, 860]. The limit at which toxic effects can appear is ca. 500 mg/d or more [861].

9. Vitamin B_{12} (Cobalamins)

9.1. Introduction

Vitamin B_{12} [*68-19-9*] belongs to the group of cobalamins, which also includes adenosylcobalamin [*13870-90-1*], methylcobalamin [*13422-55-4*] and hydroxocobalamin [*13422-51-0*]. The cobalamins and other compounds similar to B_{12} are corrinoids (Fig. 12). The corrinoids have a tetrapyrrole ring system which, compared to the porphyrins, lacks the methine bridge between rings A and D; central atom is cobalt.

The cobalt atom is coordinated by two ligands perpendicular to the plane of the tetrapyrrole ring. One of these ligands is a nucleotide linked to a propionyl group of the tetrapyrrole ring by an aminopropanol group.

All corrinoids that contain 5,6-dimethylbenzimidazole [*582-60-5*] as the nucleotide ligand are known as cobalamins. Depending on the anionic ligand, a distinction is made between vitamin B_{12} (cyanide as ligand), coenzyme B_{12} (deoxyadenosyl group as ligand), etc. The cobalamins are presented in Figure 13.

Deoxyadenosylcobalamin and methylcobalamin are physiologically active as coenzymes. Another compound found in the bodies of mammals is hydroxocobalamin, a precursor of the coenzymes. Cyanocobalamin is a form made from the naturally occurring cobalamins. It is reconverted to the coenzyme forms in the body.

Other B_{12} derivatives are found in bacteria. They have in common that the ribazole group (5,6-dimethylbenzimidaole) is replaced by imidazoles or purines. Important representatives are pseudovitamin B_{12} (Co α-[α-(5,6-dimethylbenzimidazoyl)]cobamide) of methanogenic bacteria. In addition, other corrinoids are found in acetogenic bacteria, clostridia, and sulfate-reducing and sulfur-metabolizing bacteria [862].

In 1925, Whipple and Robscheit-Robbins postulated for the first time that vitamin B_{12} is an animal protein factor. In 1926, Minot and Murphy showed that raw liver cured pernicious anemia. Subsequent work showed that this disease is caused by vitamin B_{12} deficiency. In 1948, vitamin B_{12} was isolated from liver samples by Rickes and Smith. The structure was elucidated by Hodgkin et al. in 1955. In 1972, the total chemical synthesis, with more than 70 synthetic steps, was achieved by Woodward and Eschenmoser.

Neither animals nor plants can synthesize vitamin B_{12}. According to present knowledge, the biosynthesis of corrins is restricted to microorganisms.

In humans, vitamin B_{12} is absorbed in two ways: by diffusion at higher vitamin B_{12} concentrations, and via the intrinsic factor, a mucoproteid that is formed in the mucous membrane of the stomach and specifically binds cobalamins. [The name "intrinsic factor" comes from the fact that two factors can be responsible for vitamin B_{12} deficiency: lack of B_{12} (extrinsic factor) and lack of binding protein (intrinsic factor).] The mucoprotein complex reaches the ileum, where resorption is mediated by a transport system. Even without intrinsic factor, vitamin B_{12} can be absorbed by diffusion throughout the small intestine. However, the doses needed to cover the daily requirement in this way are 10 to 100 times larger than the amounts available in foods. In the liver, vitamin B_{12} is converted to coenzyme B_{12}

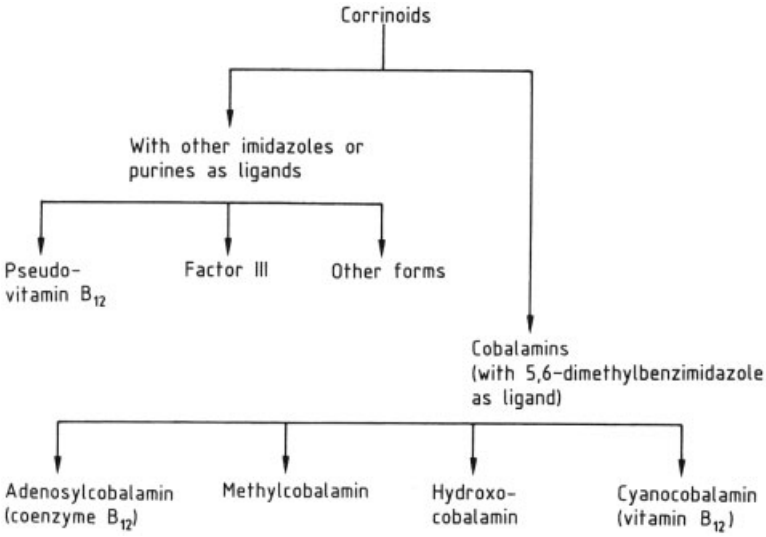

Figure 12. Classification of corrinoids

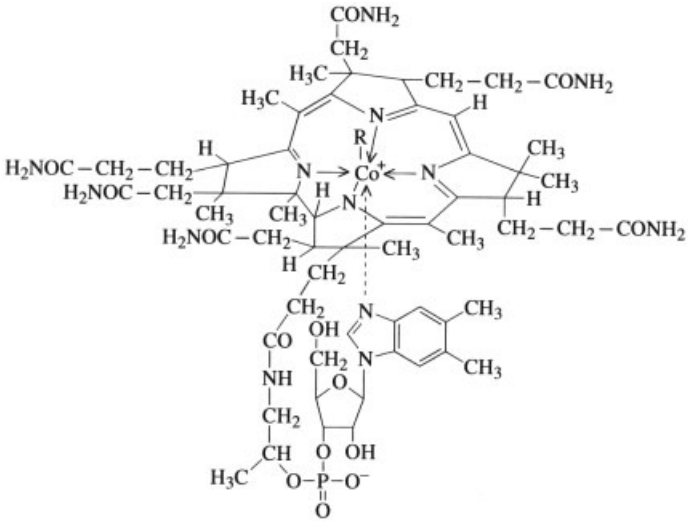

Figure 13. Structure of cobalamins
R = CN^- : cyanocobalamin = vitamin B_{12}R = 5′-desoxyadenosyl group: adenosylcobalamin = coenzyme B_{12}R = methyl group: methylcobalamin = mecobalaminR = hydroxyl group: hydroxocobalamin = aquocobalamin = vitamin $B_{12\,a}$

in a reaction dependent on adenosyl triphosphate (ATP) and flavin adenine dinucleotide (FAD), and stored. Vitamin B_{12} is excreted with the bile.

With a balanced diet, humans ingest 10–30 µg of vitamin B_{12} or other cobalamins per day, of which ca. 5 µg is resorbed. This amount has also been stipulated by the German Society for nutrition and by the U.S. Food and Nutrition Board as the daily requirement for humans.

Cobalamins are predominantly ingested with animal products. In experiments with pigs, feeding exclusively with plant proteins led to stoppage of growth due to B_{12} deficiency.

In the large intestine of humans, the intestinal flora produce ca. 50 µg of cobalamins per day, which are, however, not resorbed.

In humans vitamin B_{12} deficiency causes pernicious anemia, which is characterized by a large reduction in the erythrocyte count due to a disturbance in the maturation of red blood cells. The anemia is accompanied by neurological symptoms. In most cases, pernicious anemia is due

to a lack of intrinsic factor, resulting in reduced B_{12} absorption.

In the treatment of pernicious anemia, vitamin B_{12}, hydroxocobalamin, or adenosylcobalamin is administered in a dosage of ca. 100 µg. In the case of cyanide poisoning, high doses of hydroxocobalamin are given to bind the cyanide.

Vitamin B_{12} is used as a component of diet foods, children's foods, vitamin mixtures, cereals, beverages, vitamin preparations, and as a feed additive [863].

In pig breeding, vitamin B_{12} is used in a concentration of 10 – 30 mg/t of feed. The German Feed Law permits the addition of 10 – 20 µg of vitamin B_{12} per kilogram of feed for young pigs, hens, turkeys, and ducks. Ruminants cover their requirements via the formation of B_{12} by rumen bateria.

9.2. Properties of Vitamin B_{12}

Cyanocobalamin forms dark red, odorless, tasteless crystals (needles and prisms). The crystals are hygroscopic and can absorb up to 12 % of moisture. All the corrinoids are soluble in water (ca. 1 %), poorly soluble in lower alcohols and phenols, and insoluble in acetone, chloroform, ether, and other organic solvents. Aqueous solutions are optically active. The following optical rotations are measured in dilute aqueous solutions of cyanocobalamin: $[\alpha]^{20}_{656}$: $-59 \pm 9°$, $[\alpha]^{20}_{644}$: $-110 \pm 10°$.

Aqueous solutions of cyanocobalamin are stable. Decomposition occurs only at temperatures above 120 °C at pH values below 3 or above 8. Solutions of adenosyl-, methyl-, and hydroxocobalamin are, however, sensitive to light. The decomposition temperature and solubility of some B_{12} derivatives (solids) are listed in Table 29.

Table 29. Decomposition temperature and solubility of some important derivatives of vitamin B_{12} [864]

Name	Decomposition temperatures, °C	Solubility in water, g/L
Vitamin B_{12}	210	14.1
Adenosylcobalamin	150	15.0
Methylcobalamin	210	13.4
Hydroxocobalamin	200	107.0

According to IUPAC – IUB nomenclature, the Co complexes are designated as Co α-agylkonyl-(Co β-ligandyl) cobamides. The meaning of α and β is the same as in the steroid series. Thus, cyanocobalamin is [(Co α-(5,6-dimethylbenzimidazoyl)]-Co β-cyano-cobamide, abbreviated:

Apart from cyanide, other ions can complex cobalt, in the following order of stability: $CN^- \gg SO_3^{2-} > OH^- \gg$ halides.

In alkaline solution, the cobalamins can be reduced by thiols, giving reduced vitamin B_{12r}, a Co(II) cobalamin. Further reduction gives rise to Co(I) cobalamin (vitamin B_{12s}).

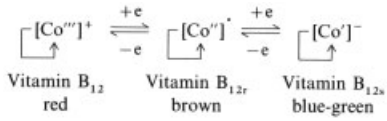

Vitamin B_{12s} can be converted by methyl iodide to methylcobalamin with the formation of an organometallic bond:

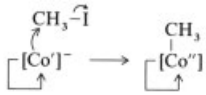

Reaction with 5′-tosyl-2′3′-isopropylidene adenosine gives adenosylcobalamin.

Most of the known biochemical reactions involving adenosyl- or methylcobalamin have been found to occur in microorganisms. These include ribonucleotide reductase (DNA synthesis), propanediol dehydratase (isopropanol degradation), glutamate mutase (glutamate degradation), ethanolamine ammonium lyase (ethanolamine degradation), lysine mutase (lysine degradation), ornithine mutase (ornithine degradation), and glycerol dehydratase (glycerol degradation). In human metabolism, methyl-cobalamin is required by the enzyme methionine synthase (E.C. 2.1.1.13), and adenosylcobalamin is required by methylmalonyl CoA mutase (E.C. 5.4.9.9.2). It is probable that leucine mutase (E.C. 5.4.3.7) is also dependent on adenosylcobalamin.

9.3. Analysis

A large number of methods for the detection of vitamin B_{12} have been described [865]:

1) Microbiological growth tests (detection limit of 10^{-11} mol/L $\approx$ 10 ng/L) with *Lactobacillus leichmanii* [American Type Culture Collection (ATCC) 4797 and 7830], which reacts very specifically to cobalamin. Growth tests with *Escherichia coli* (ATCC 9637) are also possible. These are either turbidimetric assays in test tubes or growth tests on agar plates. The plate tests are mainly used in screening for a quick assessment of the productivity of mutants.
2) Chromatographic processes (detection limit of 10^{-6} mol/L $\approx$ 1 mg/L) with HPLC methods (XAD-2 columns or reversed phase RP-18 columns and photometric detection).
3) Radioimmunological methods (RIA); commercial test sets are available for serial assays.
4) UV – VIS spectroscopic methods (detection limit of 10^{-5} mol/L $\approx$ 10 mg/L). This method is only suitable for pure substances.

Aqueous solutions of cyanocobalamin have absorption maxima at (Fig. 14): $\lambda = 278$ nm ($\pm$ 1 nm), $\varepsilon^{1\,\%}_{1\,\mathrm{cm}} = 93$; $\lambda = 360$ nm ($\pm$ 1 nm), $\varepsilon^{1\,\%}_{1\,\mathrm{cm}} = 168$; $\lambda = 550$ nm ($\pm$ 1 nm), $\varepsilon^{1\,\%}_{1\,\mathrm{cm}} = 52$.

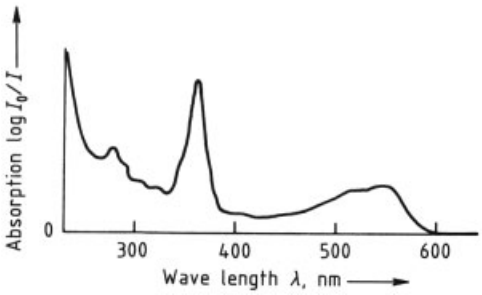

Figure 14. UV-VIS spectrum of cyanocobalamin

The absorption maxima of the other corrinoids are similar to those of cyanocobalamin.

Owing to the very low concentrations of vitamin B_{12} in foods and feeds, its detection by spectroscopic methods is possible only after purification or isolation. Microbiological methods allow very specific B_{12} detection without pretreatment of the samples. These tests are, however, less accurate and require a microbiological laboratory.

Before analysis, each of the corrinoids in corrinoid mixtures must be isolated and determined separately. If the total amount of corrinoids is sufficient, the corrinoids are converted to the cyano form and determined as vitamin B_{12}. If coenzyme B_{12} or methylcobalamin is isolated, it is necessary to work under red light because these B_{12} variants are sensitive to light.

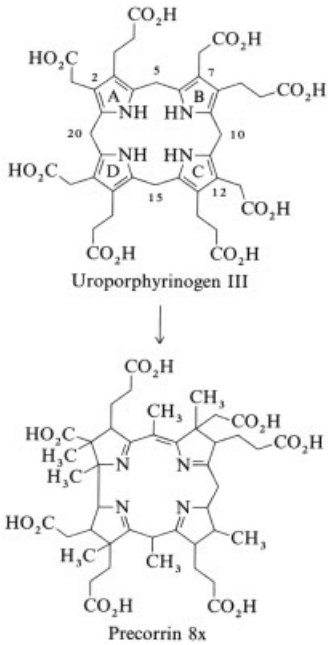

9.4. Biosynthesis

The biosynthesis of vitamin B_{12} is one of the most complex metabolic pathways known. Many of the steps have been studied [866–868]. The metabolism has, however, not yet been fully elucidated, and various groups are working on the biochemistry and genetics of B_{12} biosynthesis. The synthesis of B_{12} in the bacteria *Escherichia coli*, *Propionibacterium shermanii, Pseudomonas denitrificans*, and *Salmonella typhimurium* has been thoroughly studied.

As in the case of hemes and chlorophylls, the precursor of the corrinoids is uroporphyrinogen III [*1976-85-8*] (uro'gen III). Uro'gen III is converted to precorrin 8$\times$ by a series of methylation

steps and elimination of the methine bridge between rings A and D.

Cobalt insertion gives rise to cobyrinic acid, insertion of a molecule of threonine [*72-19-5*] produces cobinamide [*14709-02-5*], and linkage of the "nucleotide loop" with 5,6-dimethylbenzimidazole from riboflavin (vitamin B_2) [*83-88-5*] and adenylation yields coenzyme B_{12}.

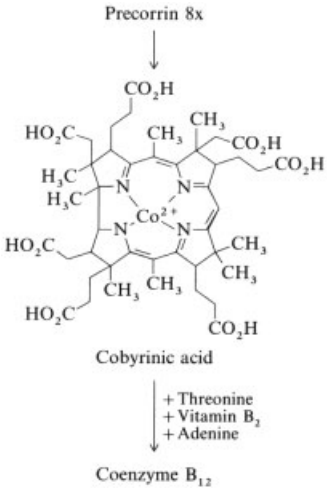

The biosynthesis of vitamin B_{12} is controlled by ca. 30 genes. In the bacterium *Salmonella typhimurium*, most of these genes are organized in a functional unit, the cob operon. The genes of *Pseudomonas denitrificans* are similarly organized.

9.5. Production

Vitamin B_{12} is produced exclusively by fermentation. With more than 70 steps, chemical synthesis is not economical.

In the past, vitamin B_{12} was isolated from fermentation residues obtained from the production of antibiotics, e.g., from streptomycin production with *Streptomyces griseus*, from neomycin fermentation with *Streptomyces fradiae*, or from chlorotetracycline production with *Streptomyces aureofaciens* [869]. In comparison with fermentation processes, however, this process is no longer economical.

Clarification sludge from anaerobic wastewater treatment is rich in vitamin B_{12} and pseudovitamin B_{12}, but extraction is not profitable.

Although more than 100 processes for the fermentative production of vitamin B_{12} are described in the literature and in patents, only five to ten are used in practice worldwide.

The most important processes are fermentation with *Pseudomonas denitrificans* and with bacteria of the genus *Propionibacterium*. Fermentations with other species have also been described [870].

Bacteria which produce B_{12} or B_{12} derivatives are listed in the following:

Propionibacteria	*Propionibacterium shermanii*
	Propionibacterium freudenreichii
Pseudomonada	*Pseudomonas denitrificans*
	Pseudomonas methanolignica
	Pseudomonas aeroginosa
Methanogenic bacteria	*Methanosarcina barkeri*
	Methanobacterium thermo-autotrophicum
Acetogenic bacteria	*Acetobacterium woodii*
	Acetobacterium spec.
Other species	*Arthrobacter hyalinus*
	Micromonaspora purpurea
	Streptomyces olivaceus
	Bacillus megaterium
	Nocardia methanophilus
	Nocardia gardeneri
	Klebsiella spec.
	Eubacterium spec.
	Rhodopseudomonas spec.
	Butyribacterium metholo-trophicum

All fermentations produce a mixture of methyl-, hydroxo-, and adenosylcobalamin. The methanogenic bacteria produce pseudovitamin B_{12} or Factor III. Isolation of the individual corrinoids is tedious. For this reason, work-up is usually carried out after the corrinoids have been converted to the cyano form by addition of cyanide.

9.5.1. Fermentation

Vitamin B_{12} is produced in fermenters which are filled with a production medium in which the bacteria grow and produce. The production medium is suited to the needs of the strains used. The fermenter is inoculated from a preculture which, in turn, comes from a working cell bank [871].

The bacterial strains used to produce vitamin B_{12} are improved in a strain-development pro-

gram. In this way, the productivity (amount of B_{12} per liter of fermentation medium per unit time) of the strains can be greatly improved. SPELLA et al. [872] have described in a program which was used to improve the productivity of *Pseudomonas denitrificans* in eight steps from 4 mg/L of B_{12} to 150 mg/L of B_{12}. In this program, the strains were treated with an agent that causes mutations (e.g., UV radiation or nitrosomethylurea [*684-93-5*]). Among the mutants, strains with improved characteristics were selected (mutation – selection process). Improved characteristics can include not only higher productivity, but also improved growth rate, improved nutrient utilization, higher tolerance to end products (e.g., propionic acid) and starting materials (e.g., dimethylbenzimidazole), or improved genetic stability. The search for improved mutants is very tedious (1 improved strain per $10^5 - 10^6$ strains).

The number of hits can be increased if it is possible to specifically search for characteristics which are connected with improved production. For instance, it is known that porphyrin synthesis reduces B_{12} production because both compounds have a common precursor in their biosynthesis. Mutants with reduced porphyrin synthesis can be recognized, e.g., by their reduced catalase activity. In the screening process, strains with reduced catalase activity are searched for first. These strains are then tested for B_{12} productivity.

It is also possible to search for improved strains by using antimetabolites (e.g., ethionine) which permit the selection of strains with altered enzyme activities.

Propionibacterium Process. The propionibacterium process is the oldest described fermentation process. *Propionibacterium shermanii* (ATCC 6707) and *Propionibacterium freudenreichii* (ATCC 13 673) are the preferred species. The carbon and energy sources are carbohydrates that are fermented to propionic acid without oxygen. The acid produced (ca. 10 % of the fermentation volume) must be neutralized (pH 7) during fermentation. An example of a production medium follows:

Glucose	120 g
Corn steep liquor	80 mL
Glycine	5 g
NH_4NO_3	3 g
$MgSO_4 \cdot 5\,H_2O$	0.5 g
Na_2MoO_4	10 µg
Calcium pantothenate	5 mg
$CoCl_2 \cdot 6\,H_2O$	10 mg
Dimethylbenzimidazole	20 g
Tap water	1 L

Propionibacteria are anaerobic organisms which grow and produce B_{12} very poorly in the presence of oxygen. However, the synthesis of 5,6-dimethylbenzimidazole, a component of B_{12}, is dependent on oxygen. Furthermore, the biosynthesis of B_{12} is inhibited by the presence of B_{12} (feedback inhibition). These problems are avoided by using a two-stage process. In an anaerobic fermentation that takes ca. 3 d, the bacteria grow and produce the vitamin B_{12} precursor cobamide (B_{12} without dimethylbenzimidazole). Subsequently, the culture is moderately aerated for 1 – 3 d to promote the biosynthesis of dimethylbenzimidazole and to complete the formation of B_{12}. As a rule, dimethylbenzimidazole is added in the second stage of fermentation, increasing the yields and reducing the fermentation time. In this way, B_{12} yields of > 150 mg/L can be achieved.

Pseudomonas Process. MILLER and ROSENBLUM [873] have isolated mutants of *Pseudomonas denitrificans* characterized by good productivities. Taxonomically, *P. denitrificans* no longer exists, having been assigned to the species *Pseudonomas* fluorescence and *Alcaligenes xylosoxidans* [874].

Pseudomonas denitrificans converts glucose to carbon dioxide, consuming oxygen in the process. In contrast to the Propionibacterium process, growth occurs with parallel synthesis of B_{12}. The fermentation is aerated and conducted at ca. pH 7 and 30 °C. After a fermentation period of 1 – 2 d, a yield of > 150 mg/L is achieved. The culture is well aerated during the logarithmic growth phase. However, the oxygen concentrations in the medium must not be high because otherwise a decrease in growth and production occurs. Dimethylbenzimidazole as B_{12} precursor is added to the fermentation medium. The growth of *Pseudomonas denitrificans* is improved by glutamate, and the synthesis of B_{12} is stimulated by betaine, choline, dimethylbenzimidazole, and cobalt. Sugar beet molasses is often used in industrial fermentations because of

its high betaine content. An example of a B_{12} production medium is present in the following:

Sugar beet molasses	105 g
Sucrose	15 g
Betaine	3 g
Ammonium sulfate	2.5 g
Magnesium sulfate	0.2 g
Zinc sulfate	0.08 g
5,6-Dimethylbenzimidazole	0.025 g
Tap water	1 L

9.5.2. Work-Up

Vitamin B_{12} and its derivatives are present inside the cell. Vitamin B_{12} is either isolated and purified or the dried cell mass is used as a concentrate in animal feeds (Fig. 15).

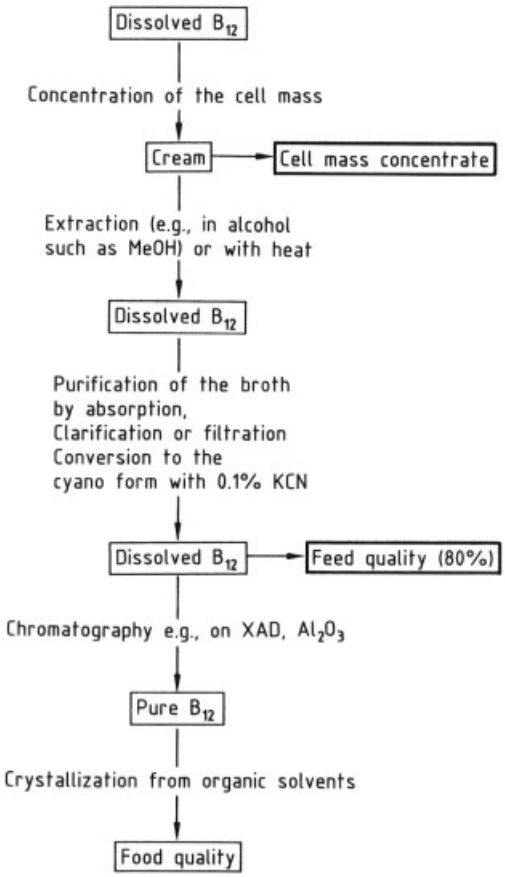

Figure 15. Work-up of vitamin B_{12}

Hydroxocobalamin is obtained if the natural cobalamins are converted to the chloro, sulfate, or nitrate form during extraction and then subjected to alkaline ion exchange [e.g., on Amberlite IRA 400 (OH^-)] [875]. Adenosyl- and methylcobalamin are isolated directly from the fermentation broth.

To produce a concentrate, the cells are separated from the medium with separators or centrifuges, and the cell mass is dried. Alternatively, the entire contents of the fermenter can be evaporated or concentrated and spray dried. The concentrates produced in this manner contain ca. 1 g of corrinoids per kilogram of dry matter.

For use in the feed industry, the centrifuged cell mass is heated as an aqueous suspension to lyse the cells. Depending on the production organism, other processes can also be used for lysis. The dissolved corrinoids are converted to cyanocobalamin by the addition of KCN, usually in the presence of sodium nitrite and heat. The vitamin B_{12} solution is then clarified by filtra-tion, treatment with zinc hydroxide, or similar processes. Precipitation of this aqueous solution produces a product of ca. 80 % purity which is a suitable animal feed additive. The precipitation can be performed by concentration and addition of auxiliaries such as tannic acid or cresol.

For further purification to give B_{12} of food and pharmaceutical quality, the clarified vitamin B_{12} solution can be extracted with organic solvents (cresol, carbon tetrachloride), and then with water after the addition of butanol, and again with organic solvents. Adsorption processes, e.g., with ion exchangers, aluminum oxide, or activated carbon, can be used alternatively or in addition. Pure vitamin B_{12} is obtained by crystallization after the addition of organic solvents. A detailed description is given in [871].

9.5.3. Patents and Scientific Survey

More than 1000 patents and articles on the fermentative production of vitamin B_{12} have been published.

A general protection by patent for the Propionibacterium and Pseudomonas processes no longer exists. However, a number of special process and preparation variants have been patented. Furthermore, other production organisms and substrates have been decribed. A survey is presented in Table 30.

Table 30. Survey of the patent and literature on the fermentative production of vitamin B_{12}

Company/year	Organism	Productivity	Remarks
Asahi Chemical (J) 1988			fermentation with cell recycling
Daicel Chemicals (J) 1987 – 91	*Acetobacterium, Methanosarcina*	4.1 mg/L in 124 h, 23 mg B_{12}/g DM∗	fermentation with methanol mutants
Dainippon (J) 1978 – 79	*Pseudonocardia, Nocardia*	1.85 mg/L	fermentation with methanol
Gedeon Richter (HUN) bis dato	*Propionibacterium shermanii*	26 mg/L in 10 d	different fermentation and preparation patents
Kobe Steel (J) 1992	*Methanobacterium thermoautotrophilum*		fermentation with methane
Kureha Chemicals (J) 1983	*Protami-nobacterium+ Rhodospirillum*		cell fusion
Kyowa Hakko (J) 1975	*Klebsiella*		methanol medium
Mitsubishi Gas Chemicals (J) 1987	*Eubacterium/Butyribacterium*	37 – 39 mg/L in 40 h	methanol medium, partly mixed culture
Nippon Oil (J) 1976 – 1995	*Arthrobacter, Propionibacterium shermanii*	43 mg/L in 7 d	fermentation with glucose, isopropanol, patents on work-up
Nippon Point (J) 1988			isolation of B_{12}
Nippon Zeon (J) 1980	*Candida tropicalis*	0.2 mg/L in 80 h	ethanol medium
Rhône-Poulenc, Rhône-Poulenc Rohrer (F) 1979 – 1985	*Pseudomonas denitrificans*		studies on metabolic elucidation, patents on cloning of metabolic pathways of pseudomonas
Shimizu Construction (J) 1990 – 1993	*Methanogenic*	1.4 mg/L in 4 d	
Shiseido (J) 1979 – 1987			reinfarction of B_{12}

∗ DM = dry matter.

9.6. Specifications and Legal Aspects

A USP cyanocobalamin reference standard is available (USP 207) with the following UV – VIS spectroscopic properties: maxima at 278 ± 1 nm, 361 ± 1 nm, and 550 ± 2 nm. The quotient of the absorptions at 361 and 278 nm must be 1.70 – 1.90.

The toxicity of vitamin B_{12} is extremely low. In animal experiments, vitamin B_{12} in doses of up to 1 g/kg of body weight showed no toxic effects. In humans, allergic reactions have been observed in a few cases. A connection between methylcobalamin and tumor development has also been reported. However, these effects are still disputed [876]. There is no limitation to the dosage of B_{12}; excess B_{12} is excreted.

9.7. Economic Aspects

With reference to the pure substance, the total sales of vitamin B_{12} amount to more than 10 t/a and the market volume is ca. € 77×10^6. The feed sector accounts for ca. 55 % of the sales, and the food/pharmaceutical sector for ca. 45 %. Apart from cyanocobalamin, hydroxo- and adenosylcobalamin are traded, especially on the pharmaceutical market.

North America and Western Europe represent the largest market potentials. The Asian countries including Japan are gaining importance. The remaining market volume is almost equally distributed among Brazil, the other countries in Latin America, and Eastern Europe. The African and Western Asian markets are of less importance.

More than 80 % of the total annual production of vitamin B_{12} comes from France. The company Rhône-Poulenc occupies a dominant position in both feed and pharmaceutical sectors. Roussel Uclaf, another French manufacturer, is active exclusively in the pharmaceutical sector.

Other producers are: Gedeon Richter, Hungary, with an estimated share of the market of ca. 10 %; Nippon Petrochemicals, Japan; Merind, India; and smaller producers in Eastern Europe.

10. Vitamin C (l-Ascorbic Acid)

10.1. Introduction

Ascorbic acid [*50-81-7*] (**1**) (or L-ascorbic acid, L-*xylo*-ascorbic acid, L-*threo*-hex-2-enoic acid γ-lactone) is the official IUPAC designation for vitamin C [893]. Ascorbic acid is the naturally occurring antiscorbutic vitamin. Vitamin C defi-

ciency causes scurvy, one of the oldest diseases known to humankind.

In 1928 vitamin C was isolated for the first time by SZENT-GYÖRGYI [894] and five years later its chemical structure was determined by HAWORTH and HIRST [895]. At the same time industrial production of ascorbic acid was developed, and has since reached a volume of ca. 60 000 t/a. The most significant characteristic of L-ascorbic acid is its oxidation to dehydro-L-ascorbic acid [*490-83-5*] (**2**), with which it forms a reversible redox system. This reducing property, together with its nutritional qualities and low toxicity, is the main reason for the numerous applications of vitamin C in the food and pharmaceutical industries.

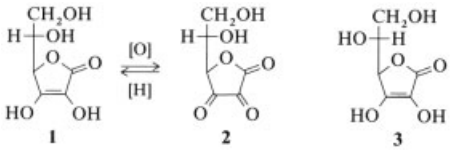

Of the other three diastereomers of ascorbic acid only D-*arabo*-ascorbic acid [*89-65-6*] (**3**) (isoascorbic acid, erythorbic acid, D-*erythro*-hex-2-enoic acid γ-lactone) shows antiscorbutic activity (ca. 5 % of that of vitamin C). In some countries it is used as an antioxidant in the food industry, as a cheaper substitute for ascorbic acid.

10.2. History [882]

Although a disease remarkably similar to scurvy was known by the ancient Egyptians, it was not until 1753 that JAMES LIND described the prevention of scurvy by dietary means. He revealed scurvy to be a dietary deficiency arising from a lack of fresh fruits or vegetables. Nevertheless, scurvy incidents still occurred in times when fruits and vegetables were unavailable.

A breakthrough in scurvy research was the discovery by HOLST and FRÖLICH [896] in 1907 that guinea pigs are susceptible to scurvy. Thus, they had an animal to test various diets for their antiscorbutic activity, and they conclusively confirmed that scurvy was produced by a defective diet.

In 1928 SZENT-GYÖRGYI accomplished the isolation of pure vitamin C from the adrenal cortex of cattle, orange juice, cabbage water, and later from paprika. He demonstrated its relationship to antiscorbutic factor and named it hexuronic acid [894]. In 1933 the structure was elucidated by HAWORTH and HIRST [895] and confirmed by a synthetic route to D-ascorbic acid and L-ascorbic acid (REICHSTEIN, HAWORTH [897]). In the same year HAWORTH and SZENT-GYÖRGYI suggested the name be changed to L-ascorbic acid to reflect its antiscorbutic properties [898]. Both scientists were awarded the Nobel Prize in 1937 for their work on vitamin C.

10.3. Physical and Chemical Properties [883, 884]

A summary of physical properties of L-ascorbic acid is given in the following [899]:

Appearance	white, odorless, acidic taste
Formula	$C_6H_8O_6$
M_r	176.13
Crystalline form	monoclinic, usually plates, sometimes needles
mp	190 – 192 °C (decomp.)
Density	1.65 g/cm^3
Optical rotation	$[\alpha]_D^{25}$ in H_2O: +20.5° to +21° ($c = 1$)
	$[\alpha]_D^{20}$ in EtOH: +52° ($c = 0.5$) [900]
	$[\alpha]_D^{20}$ in MeOH: +49.5° ($c = 0.5$) [900]
pH	3 (5 g/L); 2 (50 g/L)
pK_1	4.17 (3-OH)
pK_2	11.57 (2-OH)
Redox potential	first stage: $E_0^1 = +0.127$ V (pH 5)
Solubility, g/mL	0.33 (H_2O)
	0.02 (absolute EtOH)
	0.01 (glycerol USP)
	0.05 (propylene glycol)
	insoluble in ether, benzene, oils, fats
UV	$\lambda_{max} = 245$ nm (acid solution)
	$\lambda_{max} = 265$ nm (neutral solution)
Spectral data	UV [901, 902]; IR [901, 903]; ^{1}H NMR [901, 904]; ^{13}C NMR [905, 906]; MS [907]

L-Ascorbic acid exists in at least five tautomeric forms [885], of which tautomer **1** represents crystalline L-ascorbic acid [908] and also predominates in aqueous solution at pH 2 [909]. The 2-OH and 3-OH groups are acidic. With bases ascorbic acid forms monobasic salts; the dibasic salts are unstable. In sodium ascorbate ($C_6H_7O_6Na$) and calcium ascorbate [$(C_6H_7O_6)_2Ca \cdot 2\,H_2O$] the metal atom is associated with the O atom in the 3-position [908]. The negative charge of the ascorbate monoanion is delocalized in a π-bonding system, which

explains the stability of its lactone ring toward alkali:

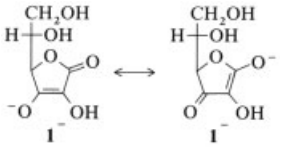

When boiled in methanolic potassium hydroxide L-ascorbic acid isomerizes to L-isoascorbic acid [910].

L-Ascorbic acid (**1**) is a strong reducing agent due to its enediol structure. Two-step oxidation of **1** proceeds via intermediate radical anion semidehydroascorbic acid (**4**) to dehydroascorbic acid (**2**) [911]. Semidehydroascorbic acid is a strong acid (pK = 0.45) and a radical scavenger; it reacts with itself, yielding **1** and **1a** [912], or with other free radicals.

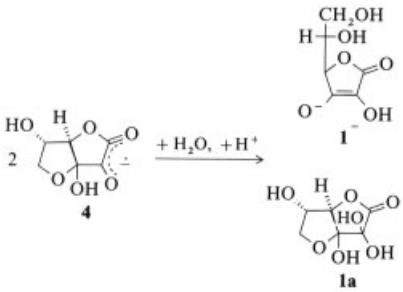

The autoxidation of ascorbic acid to dehydroascorbic acid (**2**) is catalyzed by transition metals (e.g., Cu, Fe). In the presence of oxygen they form a ternary intermediate complex (e.g., **5**) which dissociates to dehydroascorbic acid (**2**), hydrogen peroxide, and the metal ion [913].

The autoxidative process accelerates above pH 7 and can be retarded by complexing the metal ions with chelating agents such as EDTA, metaphosphoric acid, citric acid, or oxalic acid. Treatment of ascorbic acid with oxygen in methanol containing activated carbon gives a syrupy methanol adduct, which on refluxing in methyl ethyl ketone forms dehydroascorbic acid (**2**) in 64 % yield as a crystalline dimer [914]. **2** is stable for several days at 4 °C in aqueous solution at pH 2.5 – 5.5 and can be quantitatively reduced back to ascorbic acid by, e.g., hydrogen sulfide, dithiothreitol, or cysteine [915]. Dehydroascorbic acid (**2**) still exhibits vitamin C activity, which is lost on irreversible opening of the lactone ring to give 2,3-diketogulonic acid [*3445-22-5*] (**6**).

Ascorbic acid readily undergoes degradation in aqueous solution under aerobic and anaerobic conditions (Fig. 16). The decomposition depends on pH, temperature, light, concentration, and heavy metal catalysts. The loss of ascorbic acid is often ten times higher in the presence of oxygen than the loss under anaerobic conditions. Final oxidation products are L-threonic acid (**7**) and oxalic acid. Anaerobic degradation proceeds to furfural (**8**) and CO_2 [883, 916].

In the absence of moisture and in darkness solid ascorbic acid is stable for a long time. Protection against oxidative degradation in aqueous applications is achieved by coating solid ascorbic acid with fats and polymers or by chemical substitution of the enediol hydroxyl groups (L-ascorbate 2-sulfate [917], L-ascorbate 2-phosphate [918], 2-*O*-methyl ascorbic acid [919]).

10.4. Analysis [886, 887]

Methods for determination of ascorbic acid are numerous. The first test was the biological assay: prevention or cure of scurvy in guinea pigs [920]. Despite being time consuming it is still used to measure the antiscorbutic activity of products. Chemical methods are mainly based on the reducing character or derivatization of ascorbic acid. They provide less expensive and faster assays, but do not distinguish between the different stereoisomers, and often other reducing substances interfere. Therefore, the selection of an analytical procedure for ascorbic acid in biological fluids, foods, and pharmaceutical preparations should be made carefully. Increased specificity can be achieved by using commercially available ascorbate oxidase. Ascorbic acid is selectively oxidized by this plant enzyme. HPLC allows rapid separation and detection of ascorbic acid and isoascorbic acid. High analytical specificity can be achieved by using a HPLC method with a post-column oxidation/reduction system in conjunction with UV, electrochemical (EC), or fluorometric measurement of ascorbic acid. HPLC and electrochemical methods are fully

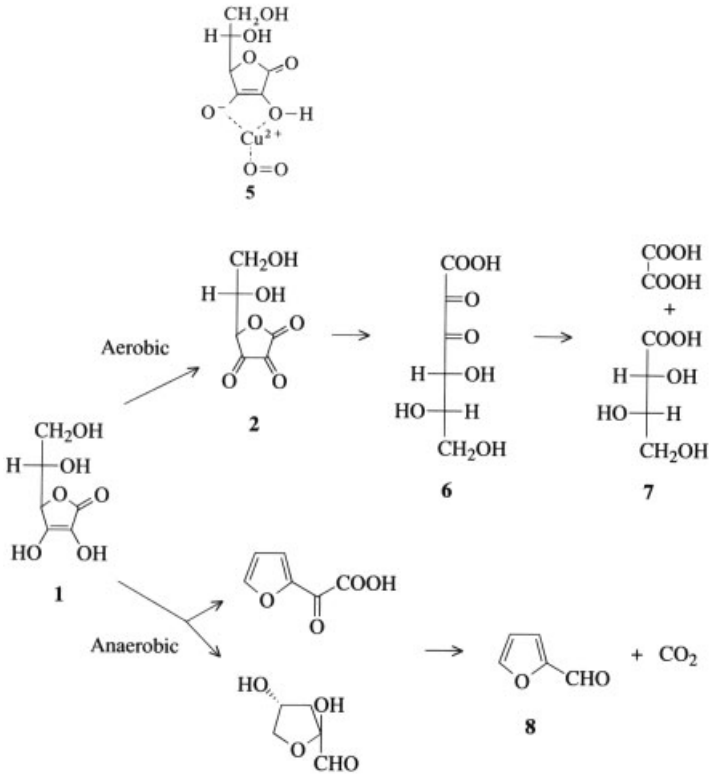

Figure 16. Degradation of ascorbic acid

Table 31. Selected analytical procedures for ascorbic acid

Method, reagents	Notes	Ref.
Iodine	standard method in pharmacopeias, wide analytical range, not specific	[921]
Chloramine T	KI/starch, mixtures with sulfite, cysteine, glutathione, pharmaceutical preparations	[922]
2,6-Dichlorphenol-indophenol	standard redox reagent, automated procedures, poor specificity, dye instability	[923]
MTT	thiazolyl blue, ascorbic acid oxidase, marketed in kit form	[924]
2,4,6-Tris-(2-pyridyl)-*s*-triazine	colored iron complex, clinical method, determination of AA/DHA	[925]
2,4-Dinitro-phenylhydrazine	classical colorimetric method, interfering substances, clinical samples	[926]
Diazotized 4-methoxy-2-nitroaniline	pharmaceutical stability studies, automated serum and urinary analysis, blank correction	[927]
o-Phenylenediamine	official method, automatable, food analysis	[928]
Ascorbic acid oxidase	enzyme isolation, increase of specificity, food and clinical analysis	[929]
Electrochemical methods	specific oxidation at an electrode, automated procedure, 0.4 ng detection limit	[930]
GLC	trimethylsilyl derivatives of ascorbic acid, sugars and related substances	[931]
TLC	qualitative analysis of AA and degradation products	[932]
HPLC	detection by UV/refractive index, IAA/AA/diketo-gulonic/gluconic acid	[933]
HPLC, derivatization with DNPH	electrochemical detection, biological fluid analysis, IAA/AA, 2 ng detection limit	[934]
HPLC, reversed phase, ion pair	fluorescence detection, biological fluid analysis, AA/DHA/IAA, 10 ng detection limit	[935]

automatable and therefore suitable for routine assays or process control in production.

A selection of analytical methods is presented in Table 31.

10.5. Occurrence and Sources of Vitamin C

Ascorbic acid is considered to be an ubiquitous constituent of green plants. Ascorbic acid has not been encountered in yeast and prokaryotes, other than cyanobacteria, and it is not clear whether it occurs in colorless forms of plants [888]. During winter potatoes and cabbage were probably the most important sources of vitamin C for the Western population. Today, fresh fruits and vegetables are available all year. Storage and cooking partly destroy ascorbic acid [936].

The content of vitamin C varies in the different plant tissues. High concentrations are encountered in tropical fruits and in the leaves of gladiolus. The latter source was used industrially before the first chemical synthesis had been worked out. Some plants contain ascorbic acid in bound form: ascorbigen *(Brassica* species) [937] or elaeocarpusin [938].

In the animal kingdom ascorbic acid is also widespread, but in lower concentrations. The highest contents are found in the endocrinic organs (cow: hypophysis 126, adrenal gland 100–160 mg/100 g). Some species of animals, e.g., primates (including humans), guinea pig, flying mammals, some birds, and certain fishes and insects, cannot produce ascorbic acid.

The vitamin C content of some foods (in mg/100 g) follows [939]:

Fish	0–3	Asparagus	15–30
Liver, kidney	10–40	Beans	10–30
Meat (beef, pork)	0–2	Broccoli	70–160
Milk (human)	3–6	Brussels sprouts	90–150
Milk (cow)	1–2	Cabbage	30–60
Acerola	1300	Carrots	5–7
Apples	10–30	Cauliflower	60–80
Bananas	5–10	Oat, rye, wheat	0
Black currant	150–230	Parsley	170
Cherries	10	Peas	10–30
Guava	300	Peppers	125–200
Hawthorn berries	160–800	Potatoes	10–30
Lemons	50–80	Rice	0
Oranges	40–60	Spinach	50–90
Rose hips	1000	Tomatoes	10–30
Strawberries	40–90	Turnips	15–40

10.6. Biosynthesis [888]

Tracer experiments with radiolabeled sugars have revealed different biosynthetic pathways for ascorbic acid in animals and plants.

Animals convert D-glucose via glucuronic acid to vitamin C (Fig. 17). The biosynthesis includes an inversion of the carbon chain sequence, whereby C-1 and C-6 of D-glucose become C-6 and C-1, respectively, of ascorbic acid [940]. In scurvy-prone animals such as the guinea pig, the final step, which is catalyzed by L-gulonolactone oxidase, is missing. Administration of this enzyme enables guinea pigs to live on an ascorbic acid deficient diet [941]. Humans also lack L-gulonolactone oxidase. The ability to produce ascorbic acid was probably lost in the progress of evolution [942]. L-Gulonolactone oxidase resides in the liver of mammals and in the kidneys of other vertebrates (amphibians, reptiles, etc.) [943].

Biosynthesis in plants follows a noninversion pathway. D-Glucose [*50-99-7*] (**9**) is oxidized at C-2 to D-glucosone [*1854-25-7*] (**10**), epimerized at C-5 and finally oxidized at C-1. A speculative scheme of ascorbic acid synthesis in plants based on results of tracer studies is presented in Figure 18 [888]. Since plants also convert L-galactono-1,4-lactone to ascorbic acid very efficiently, an inversion-type pathway was postulated, too [944].

Metabolic degradation in plants leads to oxalic and tartaric acid [945], and in animals to CO_2, oxalic acid, C_5 aldonic acids, and undetermined products [946].

Biosynthesis is inhibited by deficiencies of vitamins such as vitamin A, E, or biotin, and is stimulated by xenobiotics and drugs (3,4-benzpyrene, barbiturates, etc.) [947].

10.7. Manufacture of Vitamin C

The first synthesis of L-ascorbic acid [895, 948] was accomplished before complete elucidation of its chemical structure. The starting material, L-xylosone (**11**), was treated with potassium cyanide, forming L-xylonitrile (**12**). After lactonization and enolization L-ascorbic acid (**1**)

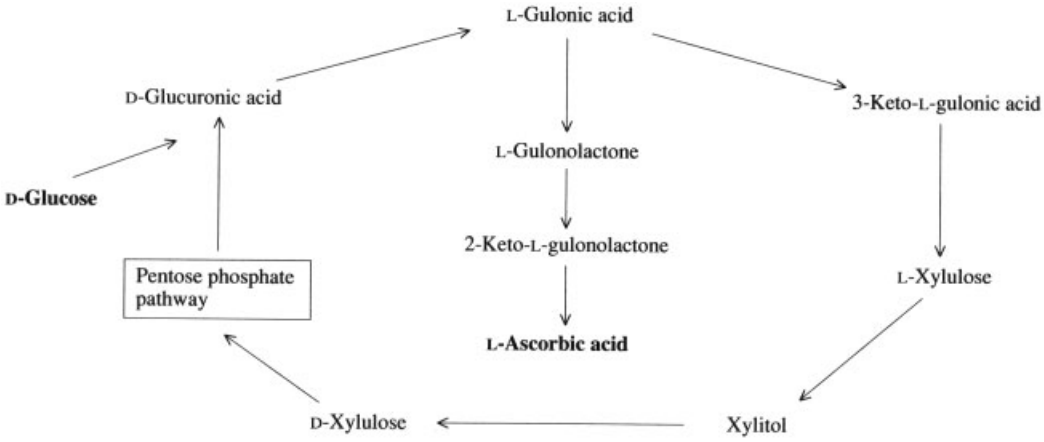

Figure 17. Biosynthesis of ascorbic acid in animals

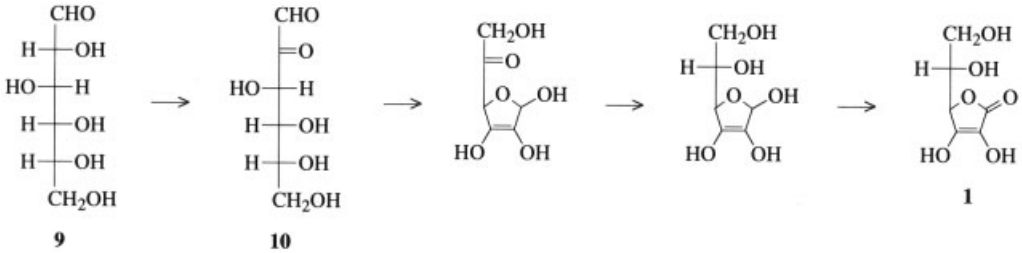

Figure 18. Biosynthesis of ascorbic acid in plants

was obtained in moderate yields (ca. 40 % from L-xylosone). However, this synthesis never attained economical importance, because of the lack of an easy and inexpensive access to L-xylosone.

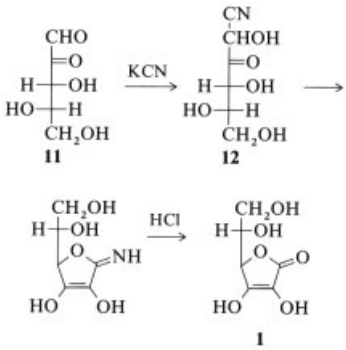

About a year after publication of the first synthesis REICHSTEIN and GRÜSSNER described their route to L-ascorbic acid from the readily available C_6 sugar D-glucose [949]. This method remains the most important synthetic pathway for commercial manufacture. Today, ascorbic acid is produced by slight modifications of the scheme first devised in 1934.

10.7.1. Reichstein Synthesis [684]

The basic principle of the Reichstein synthesis is a sequence of easy, high-yield steps without cleavage or formation of a carbon – carbon bond. The essence of this pathway is reduction of C-1 of D-glucose and oxidation at positions 5 and 6. The chirality at C-2 and C-3 is preserved to give the L-*threo*-configuration at C-4 and C-5 of L-ascorbic acid.

The classical Reichstein – Grüssner synthesis starts with reduction of D-glucose (**9**) to D-sorbitol (**13**) by hydrogenation over a nickel catalyst. The microbiological oxidation of D-sorbitol [*50-70-4*] (**13**) to L-sorbose [*87-79-6*] (**14**) is carried out with *Acetobacter xylinum*. On treatment of L-sorbose (**14**) with acetone at low temperature in the presence of sulfuric acid, 2,3 : 4,6-di-*O*-isopropylidene-α-L-sorbofuranose [*17682-70-1*] (**15**) is formed.

The di-*O*-isopropylidenyl protection of the hydroxy-groups at C-2, C-3 and C-4, C-6 allows high-yield oxidation to di-*O*-isopropylidene-2-ketogulonic acid [*18467-77-1*] (**16**), without over-oxidation or other side reactions. The oxidation is carried out with potassium permanganate in alkaline solution. Treatment of **16** with hot water affords 2-keto-L-gulonic acid [*526-98-7*] (**17**), which is converted to L-ascorbic acid (**1**) by heating in water at 100 °C (20 % yield) or by esterification and treatment with sodium methoxide in methanol followed by acidification with hydrogen chloride, yielding ca. 70 % of **1**. The overall yield of ascorbic acid from D-glucose is 15 – 18 %.

10.7.2. Industrial Manufacture by the Reichstein Route

The development of REICHSTEIN's classical procedure into an industrial process is marked by great efforts to improve each reaction step. As a result of many technical and chemical modifications each step gives over 90 % yield. The overall yield of ascorbic acid from D-glucose is now ca. 60 %.

d-Sorbitol. The catalytic hydrogenation of D-glucose (**9**) to D-sorbitol (**13**) is accomplished at high pressures and elevated temperatures in the presence of a Raney nickel catalyst (see bottom of page).

The hydrogenation is carried out in batch or continuous operations and affords almost quantitative yields with minimal formation of D-mannitol and L-iditol. After removal of the catalyst, the sorbitol solution is employed in fermentation without any further purification. Therefore, manufacturers of ascorbic acid make high demands on the quality of D-glucose (**9**); crystalline dextrose or purified starch hydrolysates are used.

The cathodic and microbiological reduction of D-glucose (**9**) to D-sorbitol (**13**) have also been reported [950].

l-Sorbose is obtained in the pyranose form by microbiological oxidation of sterile aqueous solutions of D-sorbitol (**13**) in batch or continuous operations in the presence of air. The most frequently used strains belong to the *Gluconobacter oxydans* family, which converts D-sorbitol (**13**) to L-sorbose (**14**) with more than 90 % efficiency. Large-scale fermentations have to be carried out at pH 4 – 6 and at 30 – 35 °C under sterile conditions to avoid loss of product during oxidation and work-up by filtration and crystallization.

2,3 : 4,6-Di-O-isopropylidene-α-l-sorbofuranose. The protection of the 2,3- and 4,6-

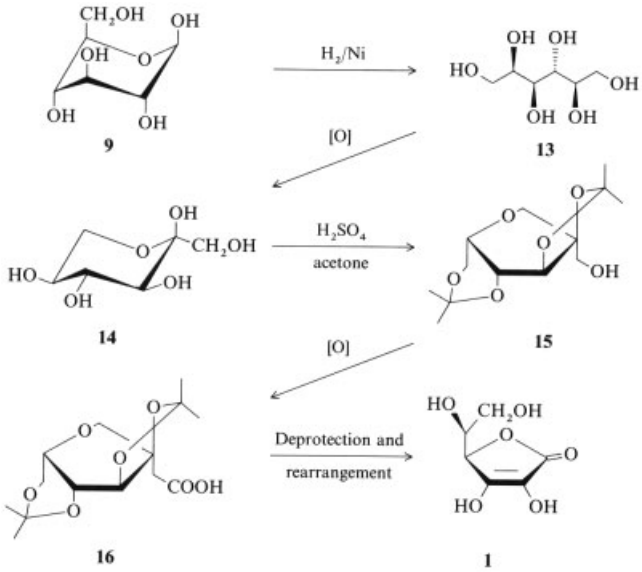

hydroxyl groups by forming cyclic ketals is achieved in acetone with excess sulfuric acid as catalyst and dehydrating agent, generally at low temperatures (e.g., 4 °C). In the final reaction mixture 2,3 : 4,6-di-*O*-isopropylidene-α-L-sorbofuranose (**15**) is obtained as main product. Byproducts are 2,3-*O*-isopropylidene-α-L-sorbofuranose and 1,2-*O*-isopropylidene-α-L-sorbopyranose. After neutralization, excess acetone is recovered by distillation. 2,3 : 4,6-di-*O*-isopropylidene-α-L-sorbofuranose (**15**) is extracted from the aqueous solution with aromatic solvents (e.g., toluene). The remaining monoisopropylidenesorboses are also recovered from the aqueous solution and returned to the process.

2,3 : 4,6-Di-O-isopropylidene-2-keto-l-gulonic acid. Originally, the oxidation of 2,3 : 4,6-di-*O*-isopropylidene-α-L-sorbofuranose (**15**) to 2,3 : 4,6-di-*O*-isopropylidene-2-keto-L-gulonic acid (**16**) was performed at elevated temperatures in dilute sodium hydroxide with $KMnO_4$, yielding ca. 90 % of product. Much less expensive oxidation methods are applied in modern continuous processes: sodium hypochlorite, electrochemical oxidation, or catalytic air oxidation.

Oxidation with hypochlorite in the presence of catalytic amounts of nickel chloride or sulfate at 60 °C yields >93 % of **16**. The active oxidant is presumably nickel peroxide [951].

Electrochemical oxidation is performed with nickel or nickel oxide electrodes in alkaline solution [952, 953]. Hydrogen, evolved at the cathodes, can be used for hydrogenation of D-glucose. It can be advantageous to stop electrochemical conversion to **16** at ca. 90 % to avoid a drop in selectivity at the end of electrolysis. Oxidation can be completed with sodium hypochlorite solution. At the end of the electrooxidation, selectivity can be improved by electrolytic cells with high electrode surfaces, such as Swiss-roll cells [954] or cells with three-dimensional electrodes [955].

Alternatively, the oxidation has been performed by using air or oxygen and a metal catalyst in alkaline solution. Good yields are achieved with palladium or platinum on carbon [956, 957]. The reaction product 2,3 : 4,6-di-*O*-isopropylidene-2-keto-L-gulonic acid (**16**) is isolated by acidification and precipitation as the monohydrate.

l-Ascorbic acid. The conversion of 2,3 : 4,6-di-*O*-isopropylidene-2-keto-L-gulonic acid (**16**) to L-ascorbic acid (**1**) is achieved by two different procedures:

1) Deprotection to give 2-keto-L-gulonic acid (**17**), followed by esterification with methanol and base-catalyzed cyclization
2) Acid-catalyzed cyclization to ascorbic acid directly from the protected or released 2-keto-L-gulonic acid.

The starting material for base-catalyzed reactions is methyl 2-keto-L-gulonate [*3031-98-9*] (**18**), prepared by treatment of the acid **16** with acidic methanol. Finally, reaction of the methyl ester **18** with sodium hydrogencarbonate or sodium acetate affords sodium ascorbate in high yield [958, 959].

The first acid-catalyzed route to ascorbic acid was published only few years after discovery of the Reichstein process [960, 961]. Today, the industrial process is performed with an inert solvent (e.g., trichloromethane or toluene) in the presence of hydrochloric acid. The advantage of this method is that ascorbic acid precipitates from the mixture as it is formed, minimizing decomposition during reaction. The crude product is obtained by filtration in high yield and high purity. After dissolution of crude ascorbic acid in water, impurities are removed by refining with activated carbon, decolorizing resins, or ionexchange resins, followed by crystallization.

Environmental Aspects. The economical efficiency of the Reichstein synthesis is strongly influenced by environmental protection costs.

Wastewater is fed into a sewage plant. The discharged COD load is similar to the amount of ascorbic acid produced. Most constituents (sugar derivatives) are easily biodegradable. Effluents containing ketalized sugar derivatives have to be pretreated to hydrolyze nondegradable ketals. Loss of nickel catalyst in wastewater is nearly quantitatively eliminated by filtration and adsorption on ion-exchange resins. The inorganic load consists of sodium chloride and sodium sulfate.

Contaminations in exhaust gases are mainly removed by absorption in water scrubbers.

Gases containing water-insoluble compounds such as chlorinated solvents are purified by adsorption on activated carbon and/or by low-temperature condensation.

Distillation residues are disposed of by incineration. Solid residues are mainly sewage sludge, spent carbon, and filter aids.

10.7.3. Other Methods for Production of Ascorbic Acid [879, 885, 889]

All syntheses of L-ascorbic acid (**1**) are partial syntheses in which the chirality at C-4 and C-5 is derived from naturally occurring sugars. Starting material is most frequently cheap D-glucose (**9**), and to some extent galactose and galacturonic acid obtained from natural sources. Syntheses involving the coupling of a C_1 fragment and a C_5 fragment (e.g., L-arabinose [962]) or a C_2 fragment and a C_4 fragment are less important, but useful for the preparation of radiolabeled derivatives and analogues of ascorbic acid (review in [879]). D-Glucose (**9**) is converted to L-ascorbic acid (**1**) either by inversion of the carbon chain sequence (C-2/C-3 of glucose become C-5/C-4 of ascorbic acid) or by inversion of the configuration of C-5. A schematic overview of selected partial syntheses of L-ascorbic acid is given in Figure 19.

Many efforts have been made to reduce production costs by optimizing and shortening the various reaction routes. These attempts include the combination of chemical and microbiological reaction sequences, direct chemical oxidations (avoiding the use of protecting groups), enzymatic catalysis, the use of genetically modified microorganisms or plant tissues, and cofermentations. Main emphasis in the last decade was on the development of the fermentative transformation of D-glucose (**9**) to 2-keto-L-gulonic acid (**17**) via D-sorbitol (**13**), L-sorbose (**14**) or 2,5-diketogluconic acid [*2595-33-7*] (**28**), and the direct conversion of D-glucose (**9**) to L-ascorbic acid (**1**) by mutagenized microalgae. However, the Reichstein synthesis still predominates in commercial production, although it is reported that a single fermentation step to 2-keto-L-gulonic acid (**17**) has already reached production scale [963].

Synthesis via l-Sorbose (Fig. 19 B). *Modification of Protecting Groups.* The replacement of acetone as a protecting agent for the oxidation of L-sorbose (**14**) has been intensively studied in the past [885]. However, the use of other reagents such as formaldehyde, benzaldehyde, cyclohexanone, etc. is not economic and little attention is paid to further improvements of these methods. Syntheses using partly protected L-sorbitol [964] or L-sorbose (e.g., methylsorboside [965]) are also not viable alternatives.

Chemical Oxidation. Direct oxidation of L-sorbose (**14**) to 2-ketogulonic acid (**17**) has been tried by many authors. Best results are obtained by catalytic oxidation with noble-metal catalysts, promoted by heavy metals such as bismuth or lead. The reaction however is not sufficiently selective, the catalyst is rapidly deactivated, and published yields (up to 87 %) are hardly reproducible [966].

Microbiological Oxidation. Fermentative production of 2-keto-L-gulonic acid (**17**) from L-sorbose (**14**) with microorganisms of the genera *Pseudomonas, Bacillus, Gluconobacter*, etc. is the subject of numerous patents. The biosynthesis of 2-keto-L-gulonic acid (**17**) in bacteria proceeds via L-sorbosone [*49865-02-3*] (**19**). L-Idonic acid, reported earlier to be an intermediate, is probably a degradation product of 2-keto-L-gulonic acid (**17**) [967].

Oxidation can be accomplished with single microorganisms, mixed-culture fermentations, or by enzymatic catalysis. Continuous improvements to these methods over the last decades by using bioengineering techniques has led to remarkable results. A Chinese patent claims a process which comprises the mixed-culture fermentation of L-sorbose (**14**) to 2-keto-L-gulonic acid (**17**) in yields of 70 – 80 % [968]. Even better results are obtained by fermentation of L-sorbose (**14**) with *Pseudogluconobacter saccharoketogenes/Bacillus megaterium* and supplementing the culture medium with rare earth elements (yield 86 %); isolation of the crystalline free acid from the fermentation broth gave an overall yield of ca. 70 % [969]. Further improvement is achieved when D-sorbitol (**13**) is used as starting material instead of L-sorbose (**14**). Mixed-culture fermentation of D-sorbitol (**13**) with strains of *Gluconobacter suboxydans* and *Gluconobacter oxydans* results in yields of 75 – 90 % [970]. A purification method for 2-keto-L-

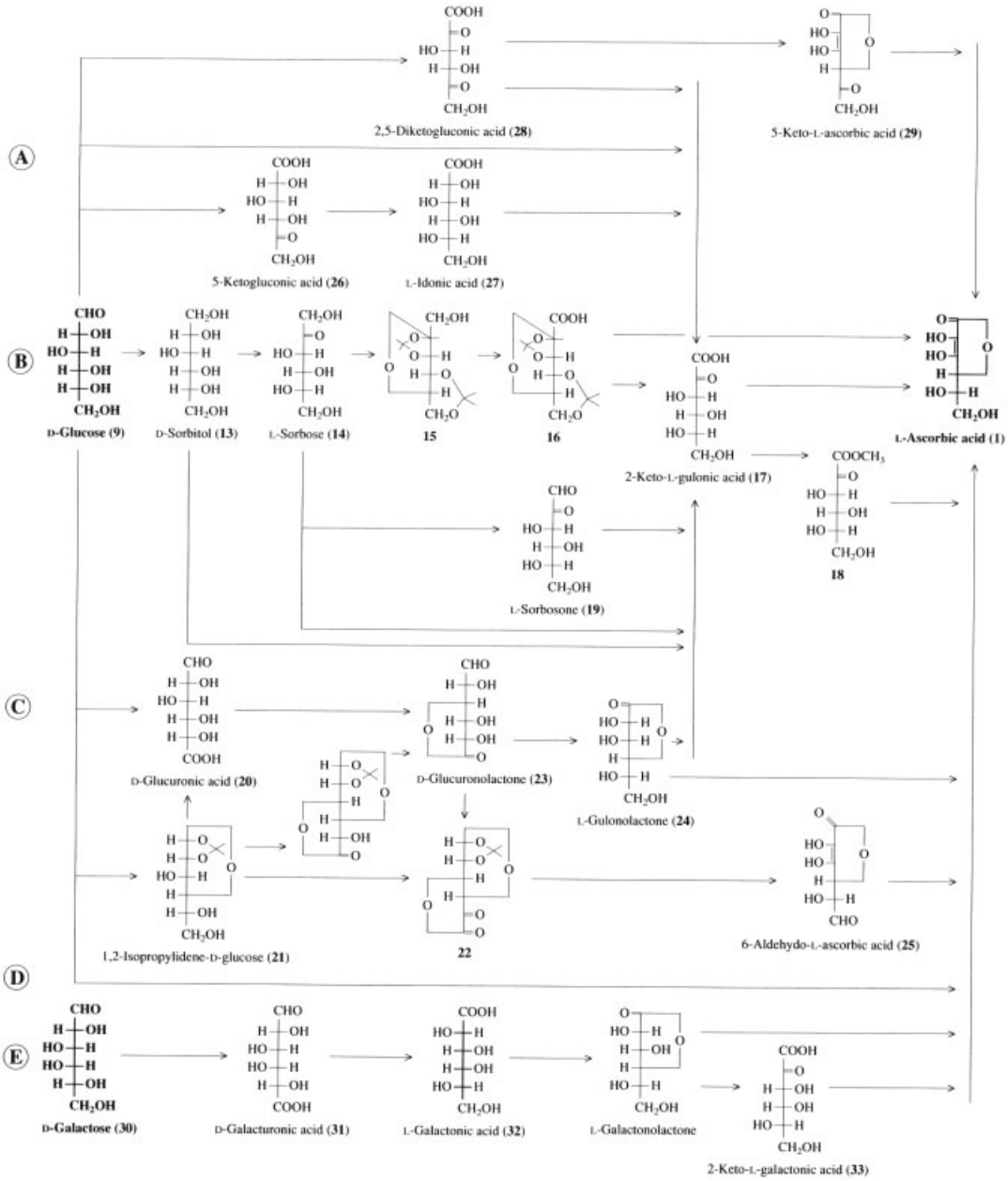

Figure 19. Synthetic pathways to L-ascorbic acid

gulonic acid (**17**) is given in [971]. Once scaled up, the microbiological oxidation of D-sorbitol (**13**) or L-sorbose (**14**) to 2-keto-L-gulonic acid (**17**) will enable a cheaper production route to L-ascorbic acid.

In a wider sense all these reaction sequences via L-sorbose represent modifications of the Reichstein process, and the final step – conversion of 2-keto-L-gulonic acid (**17**) to ascorbic acid (**1**) – still has to be performed chemically.

Synthesis via d-Glucuronic Acids (Fig. 19 C). Other routes from D-glucose (**9**) to L-ascorbic acid (**1**) start with oxidation at C-6,

leading to D-glucuronic acid derivatives. The C-1 aldehyde group is reduced in subsequent steps. Key intermediates of these syntheses are the ketolactone (**22**) and L-gulonolactone [*1128-23-0*] (**24**).

In the synthesis of BAKKE and THEANDER (**9** → **21** → **22** → **25** → **1**) L-ascorbic acid is obtained in 28 % overall yield from D-glucose [972]. A modification of this method has been published by KITAHARA, who synthesized ketolactone (**22**) from D-glucuronolactone [*32449-92-6*] (**23**) [973]. Due to low yields and high reagent costs these syntheses have not attained industrial significance.

A shorter production route to L-ascorbic acid proceeds via L-gulonolactone (**24**), which is obtained by hydrogenation of D-glucuronolactone (**23**) [974]. L-Gulonolactone (**24**) can be oxidized chemically [975] and microbiologically [976] to 2-ketogulonic acid (**17**) in good yield, or directly converted to L-ascorbic acid (**1**) with L-gulonolactone dehydrogenase [977]. To compete with the Reichstein process these syntheses require cheap access to D-glucuronolactone (**23**). However, **23** is obtained from protected glucose [978] or starch [979] only in low yields.

Synthesis via Gluconic Acids (Fig. 19 A). In this pathway D-glucose is converted to L-ascorbic acid without carbon-chain inversion, and the correct configuration at C-5 is achieved by a oxidation/reduction sequence. Key intermediates of these pathways are 5-ketogluconic acid [*5287-64-9*] (**26**) and 2,5-diketogluconic acid (**28**).

Microbiological oxidation of D-glucose to 5-keto-D-gluconic acid (**26**) is carried out with strains of *Acetobacter suboxydans* in almost quantitative yields. **26** is reduced by catalytic hydrogenation to a mixture of D-gluconic and L-idonic acid [*1114-17-6*] (**27**) (77 %). Fermentation of **27** with *Pseudomonas* strains yields 2-keto-L-gulonic acid (**17**) (90 %) [885]. The lack of a selective reduction method for **26** seems to be the reason that little attention has been paid to further development of this synthesis.

2,5-Diketogluconic acid (**28**) is efficiently produced from D-glucose (**9**) by several species of *Acetobacter*, *Gluconobacter*, or *Erwinia* [980]. Subsequent chemical ($NaBH_4$) or microbiological reduction converts **28** to 2-ketogulonic acid (**17**) [885]. An alternative chemical route proceeds via the 5,5-dimethyl ketal of **28** and 5-keto-L-ascorbic acid (**29**) to **1** [981]. **29** and its 5,5-dimethyl ketal exhibit antiscorbutic activity. The microbiological process → **17** has been optimized by introducing a two-stage fermentation method with mutant strains of Erwinia sp. and *Corynebacterium sp.* (73 % overall yield from glucose) [982]. The mutagenic treatment is necessary to repress the gene for 2-keto-D-gluconate formation, because the isolation of **17** from a mixture containing 2-keto-D-gluconate is difficult [983]. Addition of nitrate to the culture medium has a stimulative effect on the production of **17** [984]. Further simplification of the process by combining the relevant traits of *Erwinia sp.* and *Corynebacterium sp.* into a single organism was investigated by two different research groups, but the results obtained still remain far from industrial application [985].

Synthesis via d-Galacturonic Acid (Fig. 19 E). This approach is similar to synthesis via D-glucuronic acid [*6556-12-3*] (**20**): inversion of carbon-chain sequence, similar reaction conditions, and difficult access to low-cost starting material. The differing chirality at C-4 of D-glucuronic acid (**20**) is lost in the final reaction step (C-3 of L-ascorbic acid).

D-Galacturonic acid [*685-73-4*] (**31**) is obtained by enzymatic hydrolysis of pectic acid [885] or by chemical transformation of D-galactose [*59-23-4*] (**30**) which is available from lactose-containing whey [986]. Catalytic reduction of the aldehyde group affords L-galactonic acid [*28278-08-2*] (**32**), which is transformed to L-ascorbic acid (**1**) by fermentation with *Candida sp.* [987] or by chemical oxidation to **33** and subsequent cyclization to **1** [885].

Synthesis by Microalgae (Fig. 19 D). Direct synthesis of L-ascorbic acid (**1**) from D-glucose (**9**) can be achieved with mutagenized microalgae of the genus *Chlorella* [988]. Up to now yield and concentration are too low for industrial application, but the L-ascorbic acid (**1**) enriched biomass obtained may be useful as an aquaculture fish feed or feed additive.

10.8. Absorption and Metabolism

Ascorbic acid is readily absorbed intestinally and to a lesser extent in the mouth and stomach. The absorption rate of low supplemental doses is up to 80 %; for example, 75 % is absorbed at an oral intake of 1 g. There is an inverse relationship between supplement and absorption rate because of limited intestinal absorption capacity. The absorption rate is lowered to ca. 20 % at an oral intake of 5 g [989]. Excess ascorbic acid is excreted unaltered in the urine.

Vitamin C is the most abundant water-soluble antioxidant in the body. The polarity that encourages its presence in aqueous fluids also hinders its passage across the hydrophobic cell membrane. This, and the high concentrations within the cell, makes it almost certain that the form of transport of ascorbic acid is active and facilitated, i.e., requiring membrane carriers such as sodium ions [990].

Although widely distributed in the body, ascorbic acid is concentrated in adrenal and pituitary glands, in the brain, eye lens, and in leukocytes [991]. A dosage of ca. 60 mg of ascorbate per day is necessary to create and maintain a body pool of 1500 mg, which protects against scurvy for at least 30 d. However, the body pool can be expanded to 2300 – 2800 mg by 200 mg/d doses [992].

The main metabolites of ascorbate in the human body are oxalate, dehydroascorbic acid, 2,3-diketogulonic acid and ascorbic acid 2-sulfate. Large intakes of ascorbic acid do not induce higher levels of oxalate because humans are not able to indefinitely increase the metabolism of ascorbate.

10.9. Medical Aspects of Vitamin C

The most significant chemical property of L-ascorbic acid is its oxidation to dehydro-L-ascorbic acid. This reversible oxidation is the basis for many of its known physiological activities.

A number of enzymes are stimulated by ascorbic acid. Compared with other reducing agents vitamin C shows the greatest stimulation of dioxygenases that contain prosthetic Fe^{2+} and monooxygenases with prosthetic Cu^{+}. Dioxygenases are responsible for several hydroxylation reactions. In collagen synthesis, proline and lysine are hydroxylated on the growing polypeptide chain. Many of the clinical signs of scurvy are explained by defects in collagen synthesis. Keeping iron in dioxygenase in its required reduced form seems to be the role of ascorbic acid [992].

The requirement for ascorbic acid as a cofactor in hydroxylation reactions is also suggested in the biosynthesis of norepinephrine from dopamine, of carnitine, of homogentisic acid, of 5-hydroxytryptophan, and of 3,4-dihydroxyphenylalanine.

Ascorbic acid is an essential cofactor for the activation of peptidylglycine α-amidating monooxygenase. It provides reducing equivalents to the copper-containing enzyme which is responsible for biosynthesis of neuropeptides and endocrine peptides. Thus, ascorbic acid plays an important role in the proper functioning of the nervous and endocrine systems [993].

Much literature deals with the relationship between vitamin C and the common cold. A high vitamin C level in the body does not decrease the number of incidences of the common cold. However, it has consistently decreased the duration of cold episodes and the severity of symptoms. Ascorbic acid seems to decrease inflammatory effects by reduction of oxidizing compounds released from activated leukocytes [994].

Vitamin C appears to play many important roles in protecting the body against cancer. The ability of ascorbic acid to inhibit formation of carcinogenic nitrosamines in the stomach is the best documented cancer-protecting effect of vitamin C [995]. It takes part in the protection of DNA and lipid membranes from oxidative damage, an important factor in cancer incidence. Vitamin C in connection with other antioxidants may also have beneficial effects on immune response, thereby reducing cancer risk by enhancing tumor surveillance by the immune system [996].

Ascorbic acid serves as an intracellular and extracellular quencher of free radicals and oxidants. Together with other antioxidants such as carotenes, glutathiones or vitamin E, ascorbic acid acts as a scavenger of oxygen species. In contrast to redox reactions, in which vitamin C donates two electrons, in the inactivation of highly reactive free radicals, ascorbic acid do-

nates only one electron. The reduction of tocopheroxy radical to tocopherol, retaining the level of vitamin E in tissues, follows this pathway. Vitamin C reduces vitamin E radical not only in homogeneous solution but also in liposomal membrane systems [997, 998].

Vitamin C influences the catabolism of cholesterol. Studies in humans show that for individuals with high total cholesterol concentrations a supplement of ascorbic acid may have a beneficial effect. These results suggest an inverse relationship between vitamin C level in the body and cardiovascular diseases [999].

Ascorbic acid possesses an essential physiological function in iron absorption. It enhances the absorption in a logarithmic fashion. A dosage of 25 – 50 mg vitamin C per meal secures an optimum promoting effect on absorption [1000]. Despite some warnings, the ingestion of large quantities of ascorbic acid does not result in calcium oxalate stones, increased uric excretion, impaired vitamin B_{12} status, iron overload, systemic conditioning (dependence after prolonged ingestion of high doses), or increased mutagenic activity in healthy individuals [1001]. High doses of ascorbic acid over long periods of time do not lead to any side effects. In several clinical studies daily intakes up to 10 g were applied for several years without any adverse effects [1002].

Vitamin C is an essential nutrient for the maintenance of human health. The main application is the prevention and therapy of scurvy. L-Ascorbic acid supports wound healing and it enhances the body's immune response in preventing and fighting infection. It has been suggested that L-ascorbic acid is helpful in preventing cancer and cardiovascular diseases. A daily supplement to enhance iron-absorption from food is also recommended in treating anemia [1003].

Some high risk groups, such as smokers, elderly people, diabetics, and alcoholics, have a lower vitamin C status than the rest of the population. Pregnant and lactating women, women taking oral contraceptives, sportsmen, sick and stressed people require a higher daily intake of vitamin C. In these cases, supplementary vitamin C is beneficial.

The current recommended ascorbic acid intakes of 60 mg/d (United States, European Community) [1004, 1005] or 75 mg/d (Germany) [1006] are based only on the prevention of scurvy and on the urinary excretion threshold. The optimum amount of ascorbic acid for human health is under discussion, and many experts recommend a much higher daily intake.

10.10. Industrial Uses [890–892]

The chemical properties of L-ascorbic acid provide a wide range of industrial applications. About one-third of total production is used for vitamin preparations in the pharmaceutical industry. The rest is mainly applied as an additive to food and feed to enhance product quality and stability.

L-Ascorbic acid is an active component in a variety of pharmaceutical products, including multivitamin preparations, tablets, syrups, elixirs, and effervescent tablets. Solid dosage forms are quite stable. In liquid formulations ascorbic acid is subject to oxidative decomposition and interaction with other ingredients.

The largest field of application is in the food industry, particularly in the processing of meat, fruits, and flour. The antioxidant and physiological vitamin activity of ascorbic acid is used to preserve foods, avoiding undesirable odors, flavors, and colors due to oxidative degradation, to enhance the nutrient value, or to improve processing:

1) Inhibition of nitrosamine formation in cured meat products
2) Protection from enzymatic browning in processed fruits and vegetables
3) Nutritional restoration and fortification of foods and beverages
4) Improvement of flour and dough quality
5) Retardation of oxidative rancidity in fats and oils
6) Increased clarity of wine and beer
7) Synergistic activity in applications with other antioxidants (e.g., tocopherols)

In the United States, ascorbic acid may not be added to fresh meats, because it prevents discoloration and may thus give a false appearance of freshness. Ascorbic acid, its salts, and ascorbyl palmitate are registered food additives in the European Community and classified as generally recognized as safe (GRAS) substances by FDA.

Ascorbic acid or the more stable 2-sulfate are used in fish farming to prevent scurvy and enhance immune response.

The use of ascorbic acid in plant applications (improvement of seed germination, ozone protection, harvesting aid, disease resistance, growth regulation) is limited by its economics.

Further use in the polymer, photographic, and cosmetic industries, etc. is less important.

10.11. Economic Aspects

Total demand for ascorbic acid in 1995 is estimated at 60 000 t/a. Since 1980 world production has increased by 50 %. After a long consolidation period at ca. 24 DM/kg in Europe and $ 15 in overseas trade, the price of ascorbic acid began to fall in 1995 because of the appearance of new suppliers, especially from China. The development of world market price is as follows:

1954	36 €/kg	1980	10 €/kg
1957	20 €/kg	1982	12 €/kg
1964	8 €/kg	1993	12.50 €/kg
1974	9 €/kg	1995	10 €/kg

Principal manufacturers include: Hoffmann-La Roche; Dalry (United Kingdom), Belvidere (United States); Takeda Ind. Chem., Osaka (Japan), Wilmington (United States); Merck, Darmstadt (Germany); BASF, Grenaa (Denmark); Pliva, Zagreb (Croatia); and several Chinese companies, including Northeast General Pharmaceutical Factory, Sunve Shanghai No. 2, Xing Xang Pharmaceutical Factory, Jingshu Pharma, and Jiang Xi.

The bulk of production is sold as ascorbic acid. However, specialities such as salts, esters (ascorbyl palmitate, sulfate), coated ascorbic acid, and granulated acid for direct compression are becoming increasingly important.

11. Pantothenic Acid

11.1. Introduction

The structure of pantothenic acid, [(*R*)-(+)-*N*-(2,4-dihydroxy-3,3-dimethyl-1-oxobutyl)-β-alanine] (**1**), is characterized by a carboxyl group, a primary and a secondary hydroxyl group, an ω-peptide bond, and a chiral center. Only the *R* enantiomer occurs in nature; the *S* form has no vitamin activity.

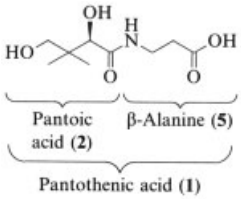

11.2. History

Pantothenic acid was discovered by WILLIAMS in 1931 as a growth factor for microorganisms [1012]. In 1938, it was isolated in a purity of ca. 40 % from sheep liver [1013]. It was then identified with the aid of two "active factors," required for normal growth in rats [1015] and to prevent dermatitis in chickens [1014]. The name pantothenic acid results from its ubiquitous distribution (Greek: from everywhere) [1016].

11.3. Physical and Chemical Properties

Pantothenic acid (**1**) is a hygroscopic oil which is instable in acidic or alkaline solution. The commercial form is the calcium salt. The sodium salt and the corresponding alcohol, (*R*)-panthenol, are used in special pharmaceutical formulations.

Calcium and sodium (*R*)-pantothenate are white powders (*mp* ca. 200°C, decomp., and 160 – 165°C, respectively). The sodium salt is hygroscopic.

(*R*)-Panthenol is a colorless, viscous oil. Other physical data are presented in Table 32.

11.4. Occurrence

Pantothenic acid (**1**) is widely distributed in plant and animal tissues in varying concentrations. Liver, kidneys, egg yolk, muscle, brain, yeast, rice and wheat bran, green leafy vegetables, and legumes are especially rich sources of this vitamin (Table 33).

Table 32. Physical data of pantothenic acid and some of its derivatives

Compound	Empirical formula	M_r	$[\alpha]_D^{20}$ ($c = 4$, water)	Solubility in		
				water	ethanol	ether
(*R*)-Pantothenic acid	$C_9H_{17}NO_5$	219.2	+ 37.5	readily soluble	soluble	slightly soluble
Calcium (*R*)-pantothenate	$(C_9H_{16}NO_5)_2Ca$	476.5	+ 27.0	readily soluble	poorly soluble	insoluble
Sodium (*R*)-pantothenate	$C_9H_{16}NO_5Na$	241.2	+ 27.5	readily soluble	slightly soluble	insoluble
(*R*)-Panthenol	$C_9H_{19}NO_4$	205.3	+ 30.5	readily soluble	readily soluble	poorly soluble

Table 33. Occurrence of pantothenic acid in foods according to the German Bundeslebensmittelschlüssel (BLS), 1990 and Souci, Fachmann, Kraut (SFK) 1989 [1017]

Food	Content, mg/100 g		
	BLS	[1018]	SFK
Fruit			
Apple	0.1		0.1
Strawberries	0.3		0.3
Oranges	0.25		0.24
Vegetables			
Broccoli	1.3		1.29
Cauliflower	0.9		1.01
Corn	0.49		0.65
Potatoes	0.35		0.4
Tomatoes	0.32		0.31
Milk products			
Skimmed milk powder		3.4	
Cows' milk	0.35		0.35
Hens' egg	1.6		1.6
Meat			
Calf liver	8.0		7.9
Pork liver	6.8		6.8
Beef heart	2.6		2.78
Pork	0.62		0.7
Beef	0.5		0.6
Cereals			
Wheat bran		2.1	
Wheat (whole grain)	1.2		1.18
Wheat flour	0.21		0.21
Soymeal		1.4	
Rolled oats	1.1		1.09
Oats (whole grain)		0.8	
Polished rice	0.25		0.63

Pantothenic acid mainly occurs in nature in the bound form as a component of coenzyme A. It is not formed in the human or animal organism. Only the conversion of pantothenic acid to the metabolically active molecule coenzyme A is biosynthetically possible. In contrast, most microorganisms are capable of synthesizing pantothenic acid from pantoic acid and β-alanine.

(*R*)-Panthenol does not occur in nature. It is rapidly oxidized to pantothenic acid by warm-blooded animals [1019]. In feeding experiments on chickens, (*R*)-panthenol exhibits ca. 60 % of the vitamin activity of an equimolar amount of calcium (*R*)-pantothenate [1020].

11.5. Biosynthesis

The biosynthesis of pantothenic acid proceeds via enzyme-catalyzed condensation of pantoic acid with β-alanine in the presence of adenosine triphosphate. The intermediate enzyme complex of pantothenic acid synthetase, pantoic acid, and AMP has been thoroughly studied [1021].

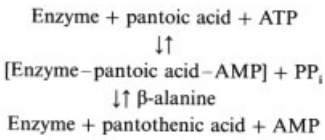

The biosynthesis of pantoic acid (**2**) branches off from that of valine at the 2-oxo-3-methylbutyric acid (**3**) stage [1022]. The introduction of a hydroxymethyl group by means of 5,10-methylenetetrahydrofolic acid [1023] results in 4-hydroxy-3,3-dimethyl-2-oxobutyric acid (**4**), which is enzymatically reduced to **2** with formation of an asymmetric center.

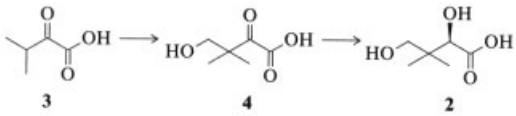

At least four pathways are known for the biosynthesis of β-alanine (**5**):

1) Hydrogenation and hydrolysis of uracil (**6**) via β-ureidopropionic acid (in *Chlostridium uracilicum*) [1024]
2) Decarboxylation of L-aspartic acid (**7**) [1025]

3) Oxidative degradation of spermine or spermidine (**8**) (in *Pseudomonas aeruginosa*) [1026]
4) Transamination of malonylaldehydic acid (**9**) with amino acids, especially with glutamic acid in mammals and microorganisms [1027]

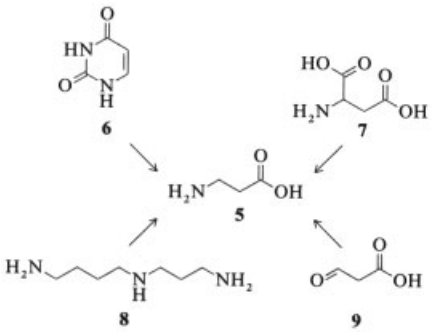

11.6. Production

The starting materials for the large-scale production of calcium (*R*)-pantothenate, the commercial form of pantothenic acid, are (*R*)-pantolactone and calcium β-alaninate.

Pantolactone. Racemic pantolactone (**10**) can be produced in a yield of ca. 90 % from isobutyraldehyde, formaldehyde, and hydrogen cyanide without the isolation of intermediates [1028]. The intermediates are 3-hydroxy-2,2-dimethylpropanal (**11**) and 2,4-dihydroxy-3,3-dimethylbutyronitrile (**12**). Saponification of **12** with strong acids directly produces the lactone (**10**), which can be isolated by extraction and/or distillation.

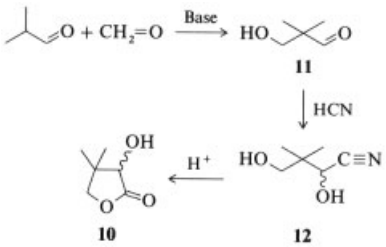

Several processes have been developed for the preparation of optically active (*R*)-pantolactone (**13**): (1) resolution of racemates with recycling, (2) asymmetric hydrogenation, and (3) microbiological synthesis.

Diastereomeric salts or amides of pantoic acid can be separated by fractional crystallization. The separation can be achieved with alkaloids [1029] and with synthetic chiral amines [1030]. Amines such as dehydroabietylamine (from rosin) [1031] and (+)-3-aminoethylpinane (from (−)-α-pinene) [1032], which are easily obtainable from inexpensive starting materials, are especially suited to industrial application.

Frequently, the enantiomers are separated by fractional crystallization after alternately seeding with the *R* or *S* form. Such processes are known for pantolactone [1033], acetylpantolactone [1034], and salts of pantoic acid [1035].

Racemization and recycling of undesired (*S*)-pantolactone are achieved either by heating with aqueous NaOH for several hours at 150°C to give sodium pantoate, or in the presence of strongly basic amines or alkoxides [1036].

Alternative processes for (*R*)-pantolactone begin with the corresponding 2-oxolactone (**16**), which is obtainable by oxidation of **10** [1037] or of the 2-aminolactone **14** [1038], or by reaction of dimethylpyruvic acid (**15**) with formaldehyde [1039]. 2-Oxopantolactone can be enantioselectively hydrogenated to (*R*)-pantolactone (**13**) microbiologically [1040] or catalytically with optically active rhodium complexes.

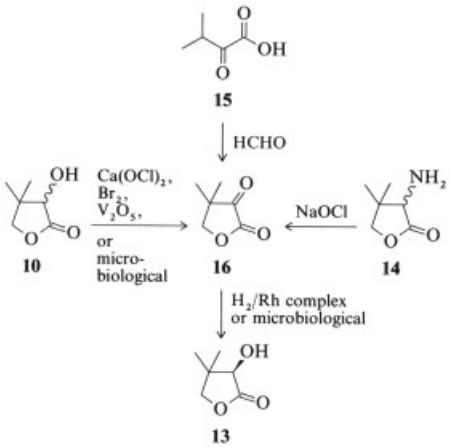

The biphosphines **17** and **18**, obtainable from 4-hydroxyproline, are the preferred chiral ligands [1041].

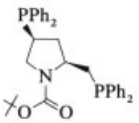

17

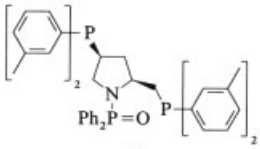

18

A large number of microbiological processes for the production of (*R*)-pantolactone have been described [1042]. For instance, stereoselective enzymatic oxidation of racemic pantolactone (**10**), followed by asymmetric reduction of the resulting 2-oxolactone, affords quantitative conversion to (*R*)-pantolactone [1043]. However, these processes have not yet gained economic importance.

β-Alanine. Two processes are mainly used for the industrial synthesis of β-alanine (**5**).

The addition of ammonia to acrylonitrile (**19**) at 20 – 30 bar and ca. 125°C yields β-aminopropionitrile (**20**), which is hydrolyzed with sodium hydroxide solution to give sodium β-alaninate [1044]. After acidification with sulfuric acid, evaporation, and extraction of the residue with methanol, crystalline β-alanine is obtained [1045].

Starting with acrylic acid (**21**), the addition of ammonia directly produces β-alanine with the avoidance of salt formation [1046].

Until now, microbiological processes [1047] have not been used on a large scale.

Aqueous solutions of β-alanine are obtained by chromatography on weakly acidic cation

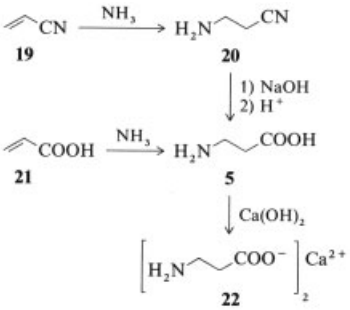

exchangers and, more recently, by electrodialysis of aqueous alkaline solutions [1048].

β-Alanine is converted to the corresponding calcium salt (**22**) before further reaction with (*R*)-pantolactone. After neutralization of **5** with $Ca(OH)_2$ in water, filtration, and drying, a product is obtained which is free of water-insoluble components. Consequently, it fulfils the strict quality requirements of Pharm. Eur. II and USP XXII.

Pantothenic Acid. Calcium (*R*)-pantothenate is obtained by the reaction of calcium β-alaninate with (*R*)-pantolactone in boiling methanol or ethanol. Depending on the crystallization and drying conditions, three solventless polymorphous forms, two solvates, and an amorphous product can be isolated [1049]. Hygroscopic prisms with four moles of methanol and one mole of water are formed on crystallization from 90 % aqueous methanol at 0°C. Drying of these prisms at 80°C and 30 mbar produces X-ray amorphous calcium (*R*)-pantothenate while maintaining the crystal form.

Calcium (*R*)-pantothenate can be produced by seeded crystallization in 90 % aqueous methanol by exploiting the higher solubility of the racemate compared with the solvates of the enantiomers [1050]. Racemization of the *S* form is achieved with sodium methoxide in anhydrous methanol [1051].

Sodium (*R*)-pantothenate is isolated by reacting sodium β-alaninate with (*R*)-pantolactone in methanol or ethanol (1 h, 65°C) [1052]. On cooling, it crystallizes in the form of square platelets (*mp* 160 – 165°C) and is highly hygroscopic.

Panthenol. (*R*)-Pantolactone reacts with 3-amino-1-propanol at 70°C (2 h) to give (*R*)-panthenol [1053]. On distillation at 170°C and 10^{-3} bar, a colorless product is obtained which meets the requirements of Pharm. Eur. II and USP XXII.

11.7. Metabolism and Importance for the Organisms; Coenzyme A and its Precursors

Pantothenic acid exhibits its biological activity as a component of coenzyme A (**23**) [1054] and of some peptide coenzymes [1055] which, in the form of energy-rich thiol esters, participate in the metabolism of carboxylic acids, especially of acetic acid.

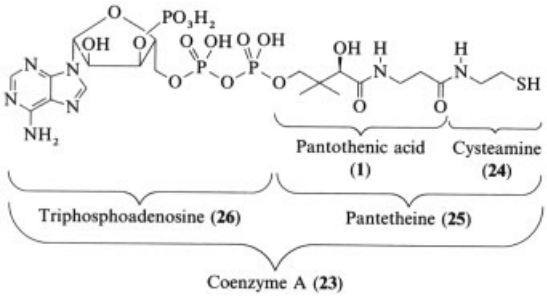

Pantothenic acid (active form: acetyl coenzyme A and derivatives) occupies a central position in intermediary metabolism (Fig. 20). It is involved in numerous biochemical reactions and enzymatic processes and, via coenzyme A, is of fundamental importance in the entire metabolism:

1) Reactive esters are formed from unreactive fatty acids
2) Degradation of carbohydrates, fats, and amino acids occurs mainly at the acetyl coenzyme A stage
3) Acetyl coenzyme A provides active acetate groups for the biosynthesis of fatty acids, phosphatides, and sterols (cholesterol, bile acids, steroid hormones)
4) Acetyl groups bound to coenzyme A enter the citric acid cycle (energy producing oxidative degradation of substances to CO_2, H_2O, and synthesis of ATP)
5) Acetyl coenzyme A transfers the acetyl group to certain acceptors (e.g., formation of the transfer agent acetylcholine at the nerve ends)
6) Amino sugars are acetylated in the synthesis of mucopolysaccharides (cartilage), hyaluronic acid (connective tissue), and other (mesenchymal) skeletal substances
7) Acetyl coenzyme A supplies the acetyl groups for acetylation reactions that are involved in the detoxification of foreign substances (e.g., chemotherapeutic agents, sulfonamides)
8) Coenzyme A participates in the synthesis of the porphyrin skeleton of hemoglobin by providing succinic acid in an activated form

Further physiological functions in the organism are attributed to pantothenic acid. It is important for:

1) Formation and normal function of body tissues
2) Maintaining the resistance of mucous membranes to infections
3) The optimal course of metabolic processes, especially in the skin and epithelial tissue
4) Promotion of growth and pigmentation of the hair

The absorption of pantothenic acid in the intestine proceeds rapidly, absorption being mainly by diffusion, less by active transport [1056]. It is distributed among practically all the cells in the entire body (tissue concentrations between 2 and 45 µg/g).

Especially high concentrations of coenzyme A are found in the liver, kidneys, adrenals, and erythrocytes. The pantothenic acid present in blood serum is bound to plasma proteins. In the human organism, pantothenic acid is hardly metabolized. Absorption and elimination of this vitamin are almost in balance (ca. 70 % of the pantothenic acid absorbed is excreted unchanged in the urine and ca. 30 % in the feces) [1057]. Thus, the vitamin status can be determined from either the plasma level or the rate of excretion in the urine.

The structural elucidation of coenzyme A in 1954 [1054] was confirmed by its synthesis in 1959 [1058]. The key product was the derivative (**27**) of adenosine diphosphate which was linked to pantetheine 4′-phosphate (**28**). The *R* configuration was unambiguously assigned in 1970 by chemical correlation of (−)-pantolactone with (*S*)-(+)-ethyl pinacolyl ether [1059].

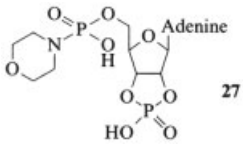

In body cells, the biosynthesis of coenzyme A (**23**) from pantothenic acid (**1**) proceeds via five intermediate steps [1060].

By using high performance liquid chromatography (HPLC), coenzyme A and its precursors can now be separated and quantitatively determined on a picomole scale [1061].

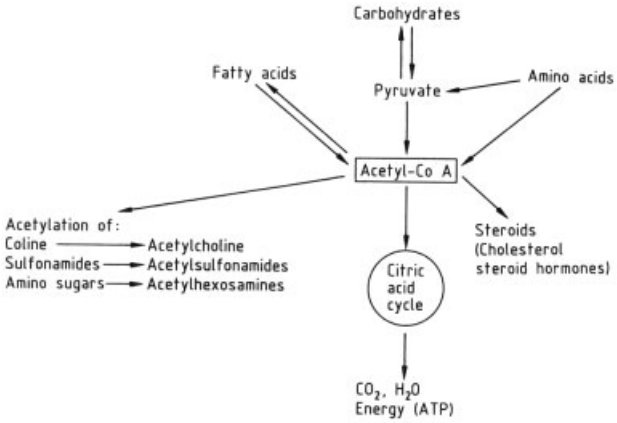

Figure 20. Intermediary metabolism

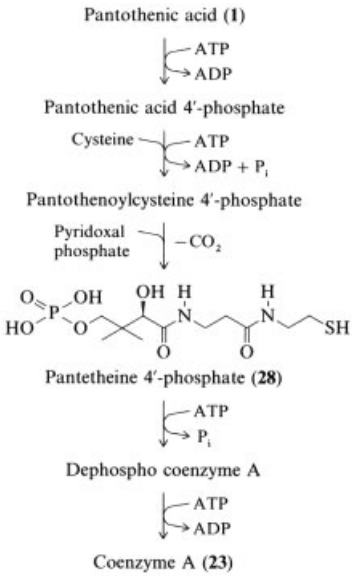

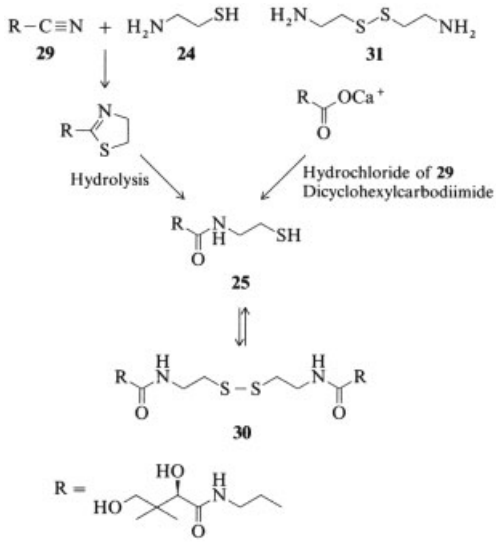

Pantetheine (**25**) was discovered as a growth promoting substance in lactic acid bacteria and was thus named Lactobacillus bulgariens factor (LBF) [1062]. Chemical synthesis was first achieved by the reaction of methyl (*R*)-pantothenate with cysteamine (**24**) [1063]. Later, higher-yield processes were developed which start with pantothenic acid nitrile (**29**) or calcium pantothenate [1064]. Analogously, the corresponding disulfide, pantethine (**30**), is obtainable from cystamine (**31**) and has interesting pharmaceutical properties [1064].

11.8. Deficiency Symptoms, Requirement, and Application

Spontaneous pronounced deficiency symptoms in humans are unlikely because of the ubiquitous occurrence of pantothenic acid. Clear deficiency symptoms appear in humans only after application of the strong pantothenic acid antagonist ω-methylpantothenic acid. A deficiency of pantothenic acid can be manifested as follows:

1) Impaired adrenal function (insufficiency symptoms), deficiency symptoms in the nervous system (peripheral perception disor-

ders, coordination problems, degeneration of the central/peripheral nerve tracts) [1065]
2) Damage to the skin (dermatitis) and to the hair and nails
3) Gastrointestinal disorders (nausea, vomiting, motility, disturbances) [1066]
4) Impaired reproduction
5) Retardation of growth or loss of weight
6) Degeneration of the liver cells (fat infiltration)
7) Inhibition of immune resistance/antibody formation (increased susceptibility to infection, e.g., infections of the upper respiratory tract)
8) Burning feet syndrome in undernourished persons [1067]

A balanced mixed diet of 10 000 kJ contains ca. 10 mg of pantothenic acid. In Germany, the average daily intake of pantothenic acid is ca. 5 mg/d [1068]. The German Society for Nutrition (1991) set the daily requirement for adults at 6 mg. The Food and Drug Administration in the United States stipulates a daily dose of 4 – 7 mg of pantothenic acid (1989). Symptoms due to large doses of pantothenic acid or panthenol are not known. Extremely high doses are tolerated without side effects, so that pantothenic acid is designated as atoxic.

Apart from the application in acute (*R*)-pantothenic acid deficiency and in preventive treatment against such a vitamin deficiency (e.g., multivitamin preparations), favorable effects in the local treatment of wounds have been reported in the literature.

Calcium (*R*)-pantothenate is preferably used in solid drug forms. The hygroscopicity of the compound must be taken into account. It is also suitable for direct tabletting. In liquid and semisolid forms – especially in combination with other vitamins – (*R*)-panthenol (provitamin B_5) is preferred to calcium (*R*)-pantothenate for reasons of stability. In the topical application of (*R*)-panthenol, a suitable form of this substance (e.g., 5 g/100 g of ointment or solution) is applied to or sprayed on the affected areas once or several times a day.

Calcium (*R*)-pantothenate is added to various foods to prevent pantothenic acid deficiency due to false nutrition or malnutrition or for certain nutritional requirements (baby food; athletes' products; low-calorie, reduced-calorie, and vitamin-rich foods).

(*R*)-Pantothenic acid improves the pigmentation and growth of hair and prevents skin damage. As a result of its higher stability and easier applicability, (*R*)-panthenol is usually preferred to calcium (*R*)-pantothenate, e.g., in hair-care products, sunscreens, shaving lotions, and baby care products.

Pantothenic acid deficiency occurs more often in working animals. The symptoms vary in individual animal species. Apart from a poor general condition, deficiency is characterized by low utilization of fodder, retardation or even arrest of growth, diverse impairments of the skin, mucous membranes, nervous system, blood, digestive tract, reproductory system, and of the adrenals [1069]. In addition, severe pathological changes can occur, including convulsions and paralytic symptoms, especially in the pig [1070], and a higher sensitivity to stress. In the hen, the egg yield is reduced and the brooding results are unsatisfactory [1071]. In severe cases, dermatitis, decolorization and loss of feathers, incrustations of the beak and around the eyes, as well as degeneration of the myelinated nerves occur [1072]. Pantothenic acid is also a vitamin for insects [1073] and acts as a growth substance in plants [1074].

Native pantothenic acid in feedstuffs usually does not cover requirements. Therefore, the composition of the total feed plays an important role. With modern feeding processes and performance demands, pigs and poultry are dependent on supplementation of the feed with pantothenic acid. In the case of ruminants with fully formed rumena, according to present knowledge, the addition of pantothenic acid is not required if the total feed is suited to ruminants. The microbial synthesis of pantothenic acid in these animals is sufficient. However, in calves fed on the basis of milk surrogates, the supply of pantothenic acid from the feedstuff is not adequate. Therefore, appropriate pantothenic acid additives are necessary. Since mixed feed contains on average 6 ppm of pantothenic acid, addition of ca. 50 % of the total requirement is recommended as a safety margin [1076].

Recommended doses in milligrams per kilogram of solo feed are [1075]:

Laying hens	8
Fattening turkeys	15
Porkers	8
Calves	20
Breeding and show horses	60

11.9. Analysis

For calcium (*R*)-pantothenate, the following methods are stipulated in the pharmacopeias (Pharm. Eur. II, US XXII): complexometric determination of calcium, determination of nitrogen by the Kjeldahl method, and measurement of optical rotation.

Furthermore, a 5 % aqueous solution must be clearly water soluble and have a pH value in the range of 6.8 – 8.

As a result of their low sensitivity, chemical detection methods can be applied only to simple mixtures (e.g., multivitamin preparations) [1077]. These analytical techniques include chromatographic preliminary purification (ion exchanger) and hydrolytic cleavage of the active agent, the pantolactone and β-alanine liberated being detected spectrophotometrically [1078], titrimetrically [1079], or colorimetrically (color reaction with ninhydrin or 1,2-naphthoquinone-4-sulfonic acid) [1080].

Pantolactone forms a purple-red hydroxamate with hydroxylamine in alkaline solution in the presence of Fe(III) [1081] and a yellow-green complex with naphthalenediol in sulfuric acid [1079].

The enantiomeric purity of (*R*)-pantolactone can be determined from its specific optical rotation and, more accurately, by gas chromatography on columns with a chiral stationary phase [1082].

As a component of pharmaceutical preparations, (*R*)-panthenol is detected by thin layer chromatographic separation [1083] or directly by ion-pair chromatograpy [1084]. Determination of optical purity can be achieved by gas chromatography after derivatization [1085].

For the determination of pantothenic acid in complex substrates (e.g., in foods and feeds and in tissues and body fluids), microbiological test systems give reliable results. The test organism used is usually *Lactobacillus plantarum* (ATCC 8014) [1086]. Pantothenic acid present in the form of coenzyme A must first be enzymatically released. Combinations of alkaline phosphatase with liver enzymes that cleave pantetheine to pantothenic acid and cysteamine are generally used [1087, 1077].

(*R*)-Panthenol can be determined microbiologically by its growth-inhibiting effect on *Leuconostoc mesenteroides* (ATCC 8042 P-60) [1088]. More recently, immunoassays have been used for the determination of pantothenic acid in blood and tissues. Both the RIA method (radioimmunoassay) [1089] and the ELISA method (enzyme linked immunosorbent assay) [1090] yield results comparable to microbiological tests.

The curative chick growth test [1091] is a biological determination method for free and bound pantothenic acid and for (*R*)-panthenol.

11.10. Economic Aspects

Pantothenic acid is supplied almost exclusively in the form of the calcium salt. The world market volume in 1993 was ca. 6000 t/a. The most important producers are: Hoffmann-La Roche (Scotland), Daiichi (Japan), and BASF (Germany) for calcium (*R*)-pantothenate; Alps (Japan), Terapia (Rumania) and Polfa (Poland) product is also supplied by other smaller producers for the feed sector [1076]. World market prices in 1993 were: calcium (*R*)-pantothenate ca. € 18/kg, calcium (*R,S*)-pantothenate ca. € 7 – 8/kg.

About 80 % of the calcium (*R*)-pantothenate is used in animal feeds and the rest in the food and pharmaceutical sectors.

World annual production of (*R*)-panthenol is ca. 1000 t with a price of ca. DM 43 – 45/kg in 1993. The most important producers of (*R*)-panthenol are Hoffmann-La Roche, Daiichi, and BASF. This product has found diverse applications in the pharmaceutical and cosmetic sectors, especially in the treatment of wounds and for skin- and hair-care products [1092, 1093].

12. Biotin

12.1. Introduction

Biotin [*58-85-5*], 2,3,3a,4,6,6a-hexahydro-2-oxo-1*H*-thieno[3,4-*d*]imidazole-4-pentanoic

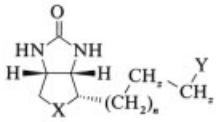

(1) Biotin	X = S	$n = 2$	$z = 2$	Y = COOH
(2) Biocytin	X = S	$n = 2$	$z = 2$	Y = $CONH(CH_2)_4CH(NH_2)COOH$
(3) Dethiobiotin	X = 2H	$n = 2$	$z = 2$	Y = COOH
(4) Selenobiotin	X = Se	$n = 2$	$z = 2$	Y = COOH
(5) Biotin sulfoxide	X = SO_2	$n = 2$	$z = 2$	Y = COOH
(6) Oxybiotin	X = O	$n = 2$	$z = 2$	Y = COOH
(7) Biotinol	X = S	$n = 2$	$z = 2$	Y = OH
(8) Norbiotin	X = S	$n = 1$	$z = 2$	Y = COOH
(9) Homobiotin	X = S	$n = 3$	$z = 2$	Y = COOH
(10) α-Dehydrobiotin	X = S	$n = 2$	$z = 1$	Y = COOH
(11) Biotin sulfone	X = SO	$n = 2$	$z = 2$	Y = COOH

acid, has three asymmetric centers and can occur in eight stereoisomeric forms. However, only the isomer with the configuration (3a*S*,4*S*,6a*R*), D-(+)-biotin (**1**), has full biological activity as vitamin H. Some important derivatives of biotin are presented in Formulas **1** – **11** above.

12.2. History

Biotin was first isolated in 1935. In addition to mesoinositol (Bios I) [*87-89-8*] and pantothenic acid (Bios IIA) [*79-83-4*], a factor Bios IIB was separated from a mixture of substances [1100] known as growth factor Bios for microorganisms [1101]. Bios IIB was named biotin. A year later crystalline biotin methyl ester [*608-16-2*] was isolated from dried egg yolk [1102]. It was verified [1103] shortly afterwards that biotin was identical to both vitamin H [1104], a protective factor against egg white injury caused by one-sided consumption of large amounts of raw eggs [1105], and coenzyme R [1106], a factor necessary for the growth and respiration of the bacteria of the root tubercles of leguminous plants.

In 1942, the structure of biotin was established by degradation reactions [1107, 1108]. This was followed by the first, albeit stereochemically nonspecific, total synthesis of biotin [1109]. The first stereospecific synthesis of pure enantiomers and diastereoisomers of biotin was achieved in 1949 [1110–1112].

12.3. Physical and Chemical Properties

D-(+)-Biotin (**1**) $C_{10}H_{16}N_2O_3S$, M_r 244.31, is readily soluble in dilute alkalis, hot water, and 95 % ethanol, slightly soluble in cold water and dilute acids, and almost insoluble in organic solvents. It crystallizes from water as long thin needles, *mp* 230 – 232 °C (decomp.), $[\alpha]_D^{22}$ +92.0° ($c = 1$, 0.4 % NaOH) [1113]. Crystalline biotin has an orthorhombic structure [1114]. The thiophane ring has the same conformation in solution and in the solid state. The sulfur lies above the plane of the ring and the valeric acid side chain occupies a quasi-equatorial position [1115]. Crystalline biotin is stable in air, thermally stable, and resistant to sunlight. It is, however, instable when exposed to UV light, oxidizing agents, and strong acids and strong bases.

With avidin [*1405-69-2*], biotin forms a molecular compound, in the ratio of 4 : 1, having one of the most stable noncovalent bonds known [1116, 1117]. Avidin, a glycoprotein contained in raw egg white, has a molar mass of ca. 70 000. This molecular compound can be cleaved neither enzymatically (peptidases) nor by acids, but only by the denaturation of avidin by radiation or by heating at 100°C for a longer period of time. Biotin forms similarly strong complexes with the related proteins streptavidin [*9013-20-1*] and stravidin [*97276-57-8*] which are formed in the fungi *Streptomyces* and *Saccharomyces* [1118].

12.4. Occurrence

Biotin occurs in varying amounts, either in the free state or bound to proteins, in most foodstuffs. The richest natural sources of biotin are yeasts [1119], lower fungi [1120], and microorganisms [1121]. Microorganisms excrete such large amounts of biotin into the surrounding

medium that most animals can cover their biotin requirement through microbiological synthesis within the intestinal tract. In humans, the fecal excretion of biotin exceeds external intake, indicating endogenous biosynthesis by intestinal flora. In plants, biosynthesis of biotin is localized in the leaves; therefore, cultivated green cells of *Lavendula vera* have an especially high biotin content [1122]. Since biotin can also be resorbed from the soil through the roots, it is distributed throughout the entire plant. Some types of phytoplankton are capable of biosynthesizing biotin as well, e.g., *Skeletonema costatum, Stephanopyxis turris* or *Gonyaulax polyedra*.

Fresh cow's milk, with a biotin content of 2.0 – 3.5 µg/mL, is one of the most important exogenous natural sources for the biotin requirement of humans [1123]. Depending on the stage of lactation, breast milk also contains varying amounts of biotin (0.06 – 0.29 µg/mL in the colostrum stage and 0.81 – 1.30 µg/mL during the nursing period) [1124]. In contrast to meat, beef liver (up to 100 µg/100 g) and kidney have an especially high biotin content. Among human foodstuffs, other fairly good sources of biotin are eggs, some vegetables (e.g., soya with 60 µg/100 g or lentils), cereals (e.g., rolled oats with 20 µg/100 g), and nuts [1123]. In smaller amounts (0.1 – 5.0 µg/100 g), biotin is a ubiquitous component of all other foodstuffs. Remarkably, the biotin content of cheese increases fourfold during maturation due to bacterial synthesis.

In some foodstuffs, such as meat or plant shoots, biotin is present in the bound form as biocytin (ε-*N*-biotinyllysine) [*576-19-2*]. Since the body can resorb only free biotin, this amide is quantitatively hydrolyzed to biotin and L-lysine [*56-87-1*] in the gastrointestinal tract by the enzyme biotinidase [1125, 1126]. However, humans utilize only ca. 50 % of ingested biotin; the rest is excreted in the feces. Accordingly, sedimentation sludge with 50 – 70 µg of biotin/100 g of solids represents a further, albeit not directly exploitable, source of biotin.

12.5. Biosynthesis

The biosynthesis of biotin in individual microorganisms is now well studied and described in detail in the literature [1127, 1128]. Although there is no uniform precursor, biosynthesis usually starts with pimelic acid [*111-16-0*] and acetyl coenzyme A, which are converted by pimelyl CoA synthetase to pimelyl coenzyme A, which undergoes decarboxylation with L-alanine [*56-41-7*] to give 7-oxo-8-aminopelargonic acid [*4707-58-8*]. This reaction is catalyzed by the pyridoxal phosphate dependent enzyme 7-oxo-8-aminopelargonic acid synthetase. Subsequently, 7,8-diaminopelargonic acid aminotransferase transfers the amino group from *S*-adenosylmethionine [*29908-03-0*] to give 7,8-diaminopelargonic acid [*21738-21-6*]. This compound then undergoes ring closure with dethiobiotin synthetase and bicarbonate, ATP, and Mg^{2+} to give dethiobiotin [*533-48-2*]. Finally, sulfur is incorporated into biotin with the aid of thiobiotin synthetase (Scheme 26) [1132].

In this unusual reaction, the sulfur presumably comes from L-cysteine [*56-89-3*] [1129] and is incorporated at the prochiral C-4 of dethiobiotin with retention of configuration, eliminating the pro-4*S* hydrogen [1130]. Accordingly, cystine, pimelic acid, and dethiobiotin promote biosynthesis in biotin culture media. However, biotin itself causes negative feedback inhibition of biosynthesis, probably by repressing formation of most of the enzymes required for synthesis [1131].

Other precursors include glutaric acid [*110-94-1*] and pelargonic acid [*112-05-0*], the latter being the precursor for biosynthesis in a strain of *Pseudomonas* [1133].

For conversion to the biologically active biotin enzyme, biotin is subsequently activated by adenosine triphosphate with the formation of biotin adenosinylate [*4130-20-5*]. This substance reacts with the ε-amino group of an L-lysine residue of apoenzymes to give the active cofactor biocytin.

12.6. Function as Prosthetic Group

Enzymes which contain biotin as cosubstrate or coenzyme are classified as carboxylases (carboxylation via hydrogencarbonate), transcarboxylases (transfer of carboxyl function), and decarboxylases (Na transport with decarboxylation), depending on their metabolic function. However, only carboxylases play a role in higher organisms [1134, 1135]. All three classes have in

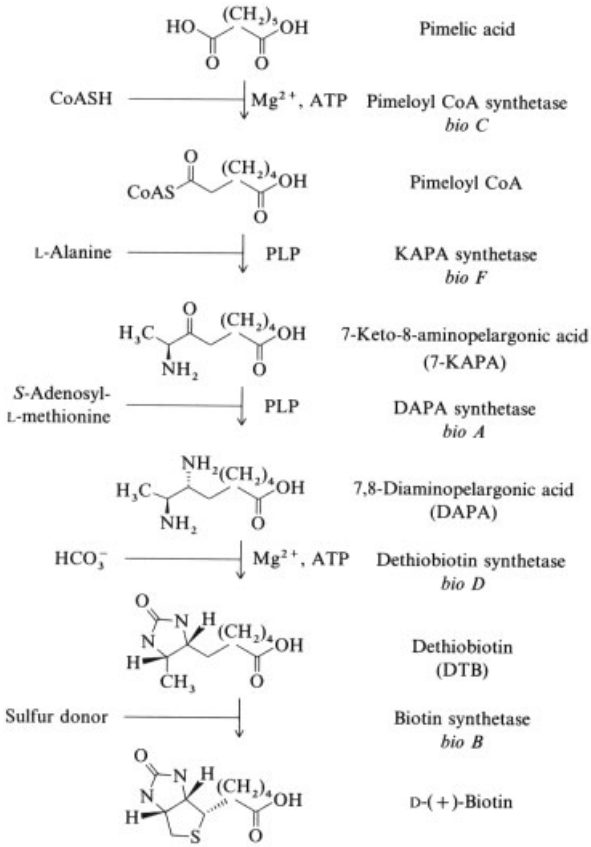

Scheme 26. Biosynthesis of biotin (enzymes and *E. coli* genes)

common a two-step mechanism. The carboxyl group is first transferred to the biotin enzyme, and then reacts further in a substrate-specific manner. In this process, the carboxylases must be activated with ATP, while the trans- and decarboxylases are independent of ATP [1136]:

$$\text{E-Biotin} + \text{ATP} + \text{HCO}_3^- \rightleftharpoons \text{E-Biotin-CO}_2^- + \text{ADP} + \text{P}$$

$$\text{E-Biotin-CO}_2^- + \text{R-H} \rightleftharpoons \text{E-Biotin} + \text{R-CO}_2^-$$

$$\sum: \text{R-H} + \text{ATP} + \text{HCO}_3^- \rightleftharpoons \text{RCO}_2^- + \text{ADP} + \text{P}$$

In addition, metal cations such as Mg^{2+}, K^+ or Mn^{2+} are required for enzyme activity. However, they play a more structural than participative role. As expected, avidin inhibits all the biotin-dependent steps of the carboxylases.

Some of the important biotin-dependent carboxylases are listed in Table 34. A series of review articles are available for detailed information on biotin enzymes [1134, 1135].

Table 34. Biotin dependent enzymes (carboxylases)

Enzyme	Catalyzed reaction	Role in metabolism
Acetyl CoA carboxylase	acetyl CoA → malonyl CoA	fatty acid biosynthesis
Pyruvate carboxylase	pyruvate → oxalacetate	gluconeogenesis, lipogenesis
Propionyl CoA carboxylase	propionyl CoA → methylmalonyl CoA	propionate metabolism
3-Methylcrotonyl CoA carbo- xylase	3-methylcrotonyl CoA → 3-methylglutaconyl CoA	leucine catabolism
Geranyl CoA carboxylase	Geranyl CoA → carboxygeranyl CoA	catabolism of isoprenoid compounds
Urea carboxylase	urea → *N*-carboxyurea	metabolization of urea in yeasts to a utilizable form

12.7. Isolation and Production

Isolation from Natural Sources. Biotin was first isolated from the natural sources egg yolk [1102], ox liver [1137], and milk [1138]. The extracts are subjected to an elaborate purification procedure, and crude biotin is finally precipitated with phosphotungstic acid [*1343-93-7*]. After conversion to the methyl ester, purification is carried out by high-vacuum sublimation or molecular distillation. However, the amounts isolated are out of all proportion to the effort required. Genetically modified strains of *E. coli* and *Bacillus sphaericus* have proved to be a richer source of natural biotin [1128]. Concentrations of up to 230 mg/L have already been obtained in fermenter charges [1139].

However, all the biotin used in the animal feed or pharmaceutical industry is still produced synthetically and corresponds to the natural product in all respects.

Industrial Synthesis [1099]. The first nonstereospecific total synthesis of biotin started with L-cysteine and involved the formation of the thiophane ring, followed by cyclization of the imidazolone segment. It produced all eight isomers of biotin [1109]. Following this proof of structure, GOLDBERG and STERNBACH were able in their synthesis to reduce the number of isomers to one separable pair of diastereomers (3a*S*,4*S*,6a*R*-(+)- and 3a*R*,4*S*,6a*S*-(−)-biotin) [1110–1112]. Their process provided the foundation for the modified synthesis of GERECKE [1140], which is the basis for almost all industrial syntheses of biotin. This improved process has the advantage of the economically more favorable separation and recycling of the biologically inactive L-(−) isomer to an earlier stage in the synthesis (Scheme 27).

More recent efforts to optimize this synthesis include shifting the isomeric ratio to favor (+)-biotin by asymmetric induction with optically active amines [1141] or alcohols [1142], recycling and isomerization of (−)-biotin [1143], or a simpler introduction of the exocyclic valeric acid group by means of the Wittig reaction [1144] or by using ortho esters [1145].

The recent literature contains numerous efforts to find new methods for the synthesis of racemic and diastereoisomerically pure biotin (Table 35).

In the latter case, efforts are made to transfer asymmetric centers from natural chiral substrates, such as sugars or amino acids, to the target molecule. Until now, however, none of these syntheses can compete with the modified Gerecke process in industrial application.

In the past years, rapid progress has been made in the biosynthetic preparation of biotin via genetically modified bacterial strains (see above). To seriously compete with the stereospecific total chemical synthesis, biotin concentrations of at least 1 g/L of fermentation solution must be attained. Therefore, the current biosyntheses which give concentrations of not more than 230 mg/L are still not feasible.

12.8. Metabolism and Importance for the Organism

Biotin ingested with food is present in the body either in the free state or protein-bound in the form of biocytin. The latter is cleaved into biotin and ε-lysine by the enzyme biotinidase, which is found in most mammals [1168]. The uptake of free biotin in the intestinal tract is not yet well understood. However, it appears that an active, carrier-mediated transport mechanism plays a role at low biotin concentrations, and a passive diffusion mechanism at high biotin concentrations [1169, 1170]. The intracellular accumulation of biotin in various bacterial strains also proceeds in an energy-dependent protein carrier process [1171].

In the plasma, 80 % of the biotin present is protein bound [1172]. Transport probably occurs via biotinidase and is not linked to any energy source [1173]. Biotin cannot be degraded in mammalian metabolism. The only reaction that has been observed is oxidation to biotin sulfoxide [*3376-83-8*] and tetrabisnorbiotin in the mitochondria after activation of the carboxyl group [1174]. Both metabolites as well as biotin are excreted in the urine and feces. The total excretion of biotin usually exceeds the intake [1175].

Various bacterial strains and molds are capable of metabolizing biotin completely. Thus, strains of *Pseudomonas* use biotin as the sole source of C, N, and S. Degradation proceeds via stepwise β-oxidative decomposition of the side chain and the thiophane ring. Finally, the imidazolidine ring is metabolized to urea [1128].

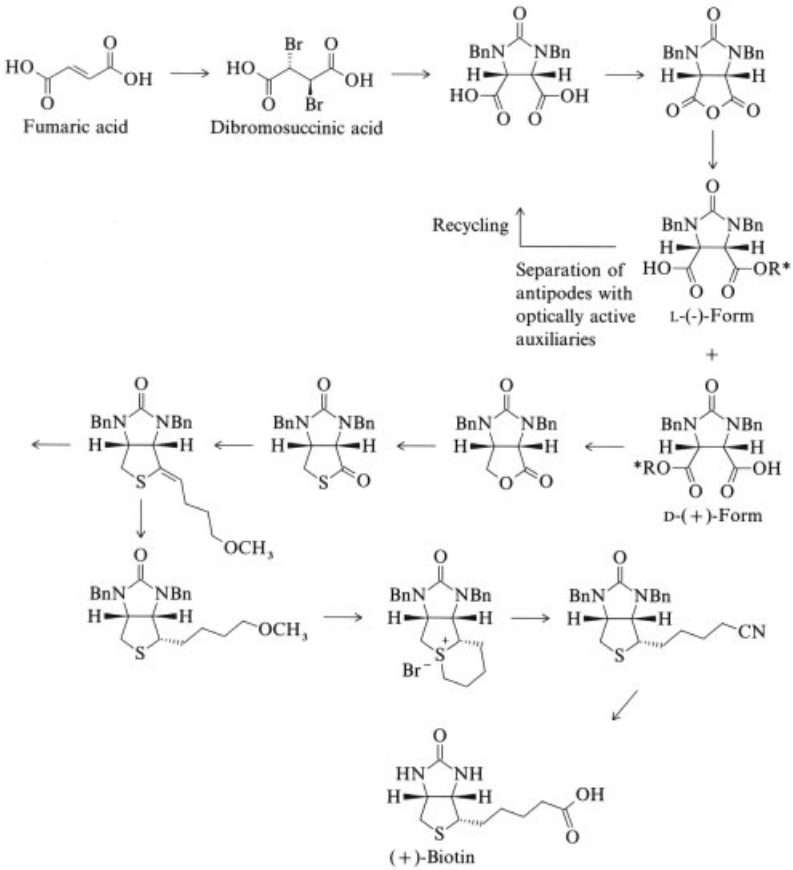

Scheme 27. Industrial preparation of (+)-biotin from fumaric acid
R∗ = oxyalkyl, aminoalkyl; Bn = benzyl

Table 35. Alternative syntheses for (±)- or (+)-biotin

Starting material	Product	Isomer separation	Steps	Refs.
L-Cysteine	(+)-biotin	+	13	[1146]
L-Cysteine	(+)-biotin	–	12/14	[1147–1149]
L-Cystine	(+)-biotin	–	11	[1150]
D-Glucose	(+)-biotin	–	23/18	[1151,1152]
D-Glucos-amine [*3416-24-8*]	(+)-biotin	–	10	[1153]
D-Mannose	(+)-biotin	–	16	[1154]
D-Arabinose	(+)-biotin	–	11/–	[1155,1156]
3-Amino-2-carbmethoxythiophene [*22288-78-4*]	(±)-biotin	–	9	[1157]
Chromene	(±)-biotin	–	12	[1158]
Cycloheptene	(±)-biotin	–	14/10	[1159,1160]
6-Methoxy-2*H*-pyran [*4454-05-1*]	(±)-biotin	–	10	[1161]
Fumaric acid	(+)-biotin	+	11	[1162]
Pimelic acid	(±)-biotin	–	11	[1163]
ω-Ketohexanoic acid [*928-81-4*]	(±)-biotin	–	9	[1164]
ω-Ketohexanoic acid	(+)-biotin	+	12	[1165]
ω-Ketoheptanoic acid [*35923-65-0*]	(+)-biotin	+	9	[1166]
1,4-Butenediol	(±)-biotin	–	14	[1167]

As a coenzyme (see above) for a series of important metabolic functions (e.g., CO_2 fixation and transfer) biotin is of essential importance for both micro- and macroorganisms. Therefore, several metabolic functions are simultaneously disturbed in biotin deficiency.

12.9. Deficiency Symptoms, Requirement, and Application

Biotin deficiency can be produced in humans and animals by diets low in this vitamin, consumption of large amounts of raw egg white, or by the administration of sulfonamides, which act as antivitamins by competitive enzyme inhibition [1175]. Manifestations of biotin deficiency in mammals include alopecia, pruritus, dermatitis, central nervous disorders, formation of crusts and rhagades, and hypercholesterinemia [1176].

12.9.1. Symptoms and Therapy in Humans

Biotin deficiency can be produced in humans by a one-sided diet of raw eggs [1177], alcohol abuse [1178], side effects of some drugs (e.g., sulfonamides or antiepileptics) [1179], inborn genetic defects (absence of biotin enzymes) [1180], or by incomplete parenteral nutrition [1181]. In babies, biotin deficiency can also occur after an excessively long nursing period (>4 months) due to the sinking biotin levels in breast milk [1182].

In 1942, biotin avitaminosis [1183] was first described by Sydensticker. He fed four adult test persons with a one-sided diet rich in egg white. The first typical symptoms, such as depresion, muscular pain, and irritability, appeared after 3 – 4 weeks. Longer term biotin deficiency leads to desquamative, maculosquamous, and seborrheic dermatitis; sleepiness; local paraesthesiae, anorexia; alopecia; ataxia; and glossitis. The normal blood values with regard to hematocrit, cholesterol, and bile pigments are altered; the excretion of biotin in the urine decreases from 30 – 50 to 3 – 7 µg/d.

Is is also possible that sudden infant death syndrome (SIDS) is also due to vitamin deficiency [1184], as suggested by the low hepatic biotin level found in the dead babies.

All the symptoms, which are biochemically caused by a disturbance in glucogenesis and potassium deficiency in the muscles, can be reversed in a very short time by oral or parenteral administration of ca. 150 µg/d of biotin. Biokinetic studies on the turnover of biotin and validated deficiency experiments are not yet available. Hence, all the known biotin requirement values are estimates. The recommendation made by the DGE (German Society for Nutrition) is 30 – 100 µg of biotin per day for adults.

12.9.2. Symptoms and Therapy in Animals

Symptoms of biotin avitaminosis in higher organisms were first observed in rats given a diet rich in egg white [1104]. After 2 – 4 weeks, the rats suffered from egg white injury, which exhibits the typical biotin deficiency symptoms. It starts with progressive dermatitis, especially of the genitals, spreading to the neck and muzzle. In the further course of the syndrome, the dermatitis becomes pronounced and resembles Leiner's disease. This is followed by alopecia, pruritus, edema, and a characteristic spastic gait (kangaroo posture). A low fat content in the diet increases the symptoms, whereas administration of ascorbic acid (vitamin C) slows down the syndrome [1185].

Before the outer signs become significant, there is a marked decrease in the activity of the biotin-dependent enzymes pyruvate and acetyl CoA carboxylase. The enzymes are transformed into vitamin-free apoenzymes, which, however, can be reactivated by addition of biotin in an ATP-dependent process [1186]. Apart from potassium deficiency in the striated neuromuscles, a further pathological symptom is the increase in the C_{16} fatty acid content and the cholesterol level in the tissue [1187].

As with other higher organisms, all the symptoms can be completely reversed within 2 – 4 weeks by the external supply of 3 – 8 µg of biotin per 100 g of animal food.

Most of these avitaminosis symptoms, such as dermatitis, loss of hair and feathers, depigmentation, encrustations of the eye lids, growth inhibition, and weight loss, are also found in other animals, including fish, hens, pigs, dogs, and monkeys [1175, 1188]. In the case of chicks, marginal deficiency symptoms also include per-

osis as well as fatty liver and kidney syndrome. The latter results in a very high mortality of the young animals due to disturbed hepatic glucogenesis via pyruvate carboxylase. The symptoms are similar to those of SIDS (see above) [1189].

12.10. Biotin Analogs

Of the 8 possible stereoisomers of biotin, only the D-(+) form (**1**) has biological activity. Its naturally occurring form, having the same activity, is biocytin, biotin ε-lysine (**2**). Biotin can be liberated from this compound by acid hydrolysis or enzymatically with biotinidase [1190].

Some analogs also show biological effects, but only in the configuration corresponding to the D-(+) form. Dethiobiotin (**3**), which can readily be made by reduction of biotin, can fully substitute for biotin as a growth factor in some microbial substrates, such as *E. coli* or *Bacillus sphaericus* [1191]. Therefore, it is certain that dethiobiotin is a precursor of biotin [1130]. After incorporation into acetyl CoA carboxylase, selenobiotin [*57956-29-3*] (**4**) exhibits coenzymatic properties comparable to those of biotin [1192]. Selenobiotin can be made synthetically or isolated microbially from *Phycomyces blakesleanus*. Like biotin sulfoxide (**5**), it can totally replace biotin in lower organisms. In the case of higher organisms, oxybiotin [*14474-91-0*] (**6**) and biotinol [*53906-36-8*] (**7**) are the only analogs capable of replacing biotin in the diet without the appearance of deficiency symptoms. Vitamin deficiency symptoms in rats are also modified by these analogs, but to a limited extent.

Biotin derivatives that are modified in the side chain, such as norbiotin [*669-72-7*] (**8**), homobiotin [*1784-22-1*] (**9**) or α-dehydrobiotin [*10118-85-1*] (**10**) are generally active antagonists for microorganisms such as *Lactobacillus casei* or *Saccharomyces cerevisiae*. α-Dehydrobiotin is one of the strongest antagonists known until now. When incubated with apocarboxylases, it irreversibly prevents the binding of biotin to the enzyme [1193]. Biotin sulfone [*40720-05-6*] (**11**) has a minor resorption-inhibiting effect. The most effective biotin antagonists are the glycoprotein avidin found in raw egg white and the proteins stravidin and streptavidin produced by microorganisms [1194].

12.11. Analysis and Standardization [1098]

For the determination of biotin concentrations in serum, bioassays using biotin-deficient animals were mainly employed in the past. The disadvantages of this method are the low accuracy and high time consumption [1195].

Highly specific biotin determinations are possible with microbiological and radioisotope dilution assays. Microbiological determination is based on the measurement of the increase in the amount of the corresponding substrates (see Table 36) after administration of a sample of unknown biotin concentration compared with several standard solutions. The biotin contained in the sample must be converted to a bioavailable form by treatment with acid or papain [*9001-73-4*], and unsaturated fatty acids must be removed because they interfere with the assay. After inoculation and incubation, the determination of the biotin concentration (detection limit 0.1 ng) is carried out by turbidimetry or titrimetry. The disadvantage of this process is that diverse biotin analogs are also included, depending on the microbe [1196].

More recent methods, including the radioisotope assay, make use of biotin – avidin technology. Identical samples of radioactively labelled avidin are treated in parallel with the unknown sample and standard biotin solutions. Subsequently, the samples are fixed via immobilized biotin. The concentration – amounts as low as 0.1 ng can be determined [1199] – is directly obtained from the count rate [1200]. Apart from labelling by means of doping, fluorescence [1201] and chemiluminescence measurements [1202] have also been used for detection.

In general, chemicophysical methods are employed for the purification and definition of purity. The purification is conveniently conducted by ion exchange [1203, 1204], paper, thin-layer, or liquid chromatography [1098]. Detection is carried out refractometrically or by reaction with color reagents. The sensitivity of the latter methods, which detect either visually or by UV/fluorescence measurements, varies between 0.3 and 50 ng [1205]. Finally, purity testing can

Table 36. Microorganisms for biotin bioassays and their activity towards biotin analogs ∗

Microorganism	Biotin	Biotin methyl ester	Biotin sulfoxide	Biotin sulfone	Oxybiotin	Norbiotin	Dethiobiotin	Biocytin	7-Oxo-8-aminonopelargonic acid	7,8-Diaminopelargonic acid
Saccharomyces cerevisiae ∗∗	+	+	+	inhibitor	+/−	inhibitor	+	+	+	+
Lactobacillus casei	+	−	−	inhibitor	+/−	inhibitor	−	+		
Lactobacillus lantarum ∗∗∗	+	−	+	inhibitor	+/−	inhibitor	−	−	−	−
Neurospora crassa	+			inhibitor		inhibitor				
Ochromonas danica	+			inhibitor		inhibitor	+	−		
Allescheria boydii	+		+/−	inhibitor	+/−	inhibitor	+/−	+		
Streptococcus faecalis	+			inhibitor	−	inhibitor				

∗ + shows activity similar to pure biotin; +/− shows activity, but less than pure biotin; − shows no activity. ∗∗ Detection of "total biotin" [1197]. ∗∗∗ Detection of "true biotin" [1198].

be completed by using NMR, LC/GC – MS, and polarography.

There is no international standard for biotin.

12.12. Uses and Economic Aspects

The annual production of 25 t of biotin is largely used in the animal feed, pharmaceutical, and food industries. To a minor extent, it is also used together with avidin for analytical purposes in immunoassays, as an additive in nutrient media for microbial substrates, and in the cosmetic industry.

In the large-scale husbandry of domestic animals, the use of biotin as a feed additive is based on the prevention of the readily occurring biotin deficiency symptoms caused by the lack of this coenzyme, which is essential in a series of important metabolic functions. In pharmacy, biotin is used – in the form of multivitamin tablets, dragees, capsules, or syrup – exclusively to prevent vitamin deficiency symptoms, which can occur, e.g., after operations, during pregnancy and lactation, and in growth phases.

In immunology, diagnostics, and biochemical research, the use of biotin – avidin technology [1096, 1206] has undergone rapid development in the last few years. The method is based on two principles. First, biotin can easily be functionalized via its carboxyl group and, secondly, the bicyclic unit with which binding to avidin occurs is spatially far removed due to a spacer of four methylene groups. Avidin can also be activated with suitable substrates. Thus, two otherwise chemically inert substrates can be joined together by means of the biotin – avidin complex. In tumor therapy, for instance, antitumor antibodies are biotinylated and the tumor cells are treated with these immunoglobulins. Subsequently, the damaged cells are incubated with an avidin – cytostatic agent complex. The resulting formation of the avidin – biotin complex results in the drug being brought directly to the cancer cell without interfering with healthy cells [1207, 1208].

The application of biotin – avidin technology in numerous immunoassays [1209], in the separation and purification of immunoglobulins from sera [1210] (see Fig. 21), and in the detection of DNA or RNA fragments in hybridization experiments [1211] is based on the same principles.

These uses and pharmaceutical and food technological purposes, however, require only 5 – 10 % of the annual production of 25 t of biotin (100 %). The majority is employed in the animal feed industry.

The main producers of biotin are Hoffmann-La Roche (Switzerland), Tanabe (Japan), Sumitomo (Japan), Merck (Germany), Il Sung (Korea), Lonza (United States), and, for feed-quality only, BASF (Germany).

The world market price of 1 g of 100 % pure D-(+)-biotin is ca. 5 €(1995).

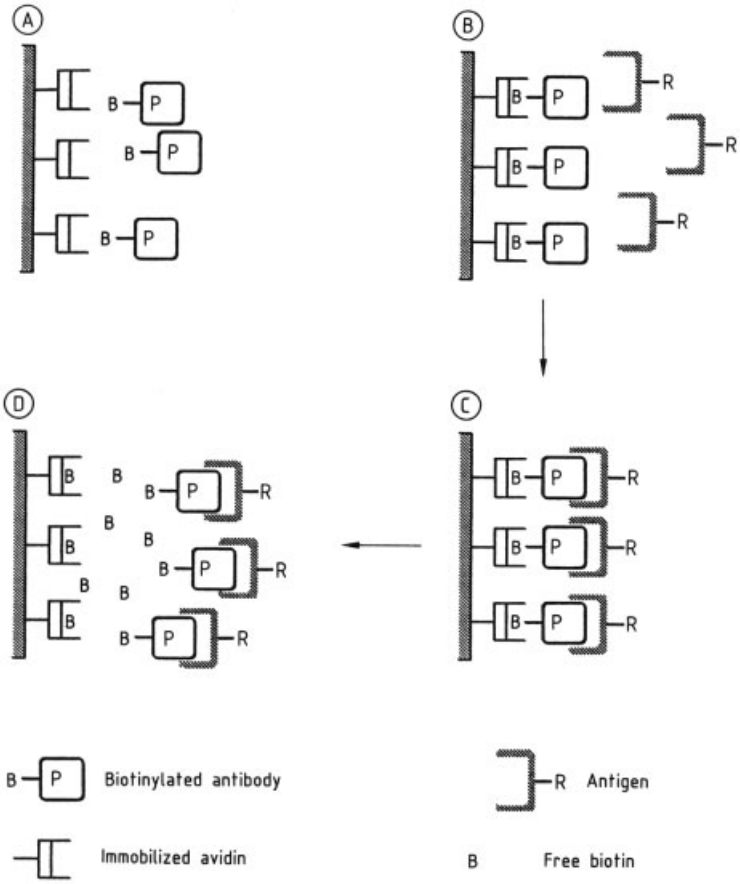

Figure 21. Isolation of antibody/antigen complex by means of biotin – avidin technology
A) Binding of biotinylated protein (antibody) to immobilized avidin; B) Addition of specific antigen; C) Binding of immobilized antibody – antigen; D) Elution of biotinylated antibody – antigen complex via displacement of avidin with free biotin

12.13. Tolerance and Environmental Protection

Even after the administration of large amounts of biotin, no acute toxicity in humans or animals has been observed [1212]. Thus, doses of up to 10 mg of biotin per 100 g of body weight showed no side effects in rats. A six-month therapeutic treatment of humans with 10 mg/d resulted in no harmful effects on body functions. Injections of 1 mg of biotin/kg of body weight in pregnant rats, however, caused resorption of the fetus and placenta and disturbances of estrogen formation [1213].

Biotin is considered environmentally harmless and is subject to no special regulations with regard to storage and transport. No explosion limits exist for biotin. Water, carbon dioxide, foam, or powder can be used as extinguishing agents in the case of fire. As a result of its acidic carboxyl function, classification as a weakly water endangering substance is being considered. Normal personal and hygienic measures (protective glasses, washing of hands) should be taken when working with biotin. Ecological problems are not to be expected if biotin is properly handled and used [1214].

13. Folic Acid

13.1. Introduction

Folates are a group of biologically active compounds which are structurally derived from

pteroylglutamic acid or folic acid (**1**). The chemical structure can be divided into the three substructures 6-methylpterin, *p*-aminobenzoic acid and L-glutamic acid. The Chemical Abstracts name is *N*-[4-[[(2-amino-1,4-dihydro-4-oxo-6-pteridinyl)methyl]amino]benzoyl]-L-glutamic acid.

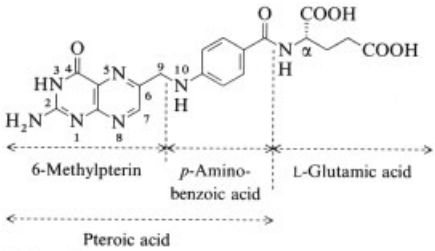

Folic acid (**1**)
Pteroylglutamic acid (PteGlu)

Table 37. H_4folate cofactors

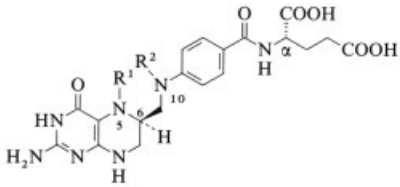

R^1	R^2	Nomenclature *	Configuration of natural stereoisomer
2 H	H	tetrahydrofolic acid (H_4folate)	(6*S*,α*S*)
3 CH_3	H	5-methyltetrahydrofolic acid (5-CH_3-H_4folate)	(6*S*,α*S*)
4 $-CH_2-$		5,10-methylenetetrahydrofolic acid (5,10-CH_2-H_4folate)	(6*R*,α*S*)
5 HC=O	H	5-formyltetrahydrofolic acid (5-HCO-H_4folate)	(6*S*,α*S*)
6 H	HC=O	10-formyltetrahydrofolic acid (10-HCO-H_4folate)	(6*R*,α*S*)
7 HC=NH	H	5-formiminotetrahydrofolic acid (5-NHCH-H_4folate)	(6*S*,α*S*)
8 $-\overset{+}{C}H-$		5,10-methenyltetrahydrofolic acid (5,10-CH^+-H_4folate)	(6*R*,α*S*)

* [1215].

Enzymatic reduction of folic acid leads to 7,8-dihydrofolic acid (H_2folate), a key substance in biosynthesis. Further reduction, catalyzed by the enzyme dihydrofolic acid reductase, provides (6*S*)-5,6,7,8-tetrahydrofolic acid (H_4folate) (**2**). This is the parent compound of the folates **3**–**8** possessing at N-5 and/or N-10 an additional C_1 substituent (Table 37). These substances and the corresponding polyglutamates are essential coenzymes for the transfer of C_1 units of different oxidation levels. In cofactor **8** with a methenyl group ($-CH^+-$) bridging positions 5 and 10, a positive charge is associated with the substructure $[N(5)-CH-N(10)]^+$. All natural folates have the same absolute configuration at position 6 as (6*S*)-tetrahydrofolic acid (**2**).

In contrast, chemical reduction of folic acid gives (6*RS*)-5,6,7,8-tetrahydrofolic acid, a diastereomeric mixture of (6*S*,α*S*)- and (6*R*,α*S*)-H_4folate. Nomenclature and symbols for folic acid and related compounds are subject of recommendations published by the IUPAC–IUB Joint Commission on Biochemical Nomenclature (1986) [1215].

In living cells folates are predominantly stored as polyglutamates in which the second and subsequent molecules of L-glutamate are each linked by amide bonds to the γ-carboxyl of the preceding glutamate unit.

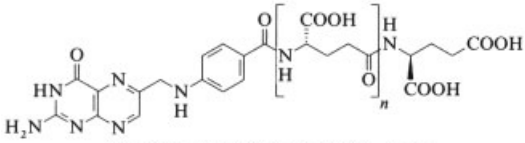

Pteroylpolyglutamic acid, $PteGlu_n$ $n \geq 1$

13.2. Historical Notes

In the 1930s, a dietary factor in yeast extracts and crude liver was found to cure megaloblastic anemia in pregnant women [1216]. Experimentally, this disorder could also be induced in monkeys and cured by antianemic factor (vitamin M). Two bioassay procedures, the microbiological growth assay and the antianemic factor assay in chickens (vitamin B_c), were developed and used to follow the isolation and purification of folates [1217–1219]. A close relationship between the yeast growth factor and the antianemic factor was soon established, but it became evident that the yeast growth factor and vitamin B_c were different entities. A crystalline preparation of a high molecular mass derivative of folate was

isolated [1220] and later shown to be a peptide [1221]. The unequivocal elucidation of the structure of pteroylglutamic acid (**1**) was achieved in 1946 by ANGIER et al. [1222]. The discovery of folates and developments during the following decade are reviewed in [1223].

13.3. Properties

This chapter describes some properties of selected folates [1224–1227]. UV data are given as λ_{max} (ε) in nanometers.

Folic acid (**1**), pteroylglutamic acid, PteGlu [*59-30-3*], $C_{19}H_{19}N_7O_6$, M_r 441.41; yellow thin platelets, no *mp*, chars above 250°C; UV (pH 13): 256 nm (30 000), 282 nm (26 000), 365 nm (9800); $[\alpha]_D^{27} = +19.9°$ (0.1 N NaOH). Contains ca. 8 % water which is lost in vacuum at 90°C. Soluble in aqueous alkali metal hydroxides and carbonates and in aqueous mineral acids, less soluble in acetic acid and pyridine, slightly soluble in hot water, very slightly soluble in cold water, insoluble in acetone and chloroform. Stable in crystalline form in the dark at room temperature; unstable in alkaline and acidic solution.

Dihydrofolic acid, 7,8-dihydropteroylglutamic acid, H_2folate [*4033-27-6*], $C_{19}H_{21}N_7O_6$, M_r 443.42; UV (pH 7.2): 282 nm (28 600). Preparation from folic acid by catalytic hydrogenation or sodium dithionite reduction in presence of ascorbic acid [1228]. Highly air sensitive compound.

Tetrahydrofolic acid (**2**), 5,6,7,8-tetrahydropteroylglutamic acid, H_4folate [*135-16-0*], $C_{19}H_{23}N_7O_6$, M_r 445.44; UV (pH 7.85): 296 nm (28 000). Synthesized from folic acid by catalytic hydrogenation in acidic solution or by reduction with sodium cyanoborohydride [1229]. Very sensitive to oxygen. (6*RS*)-H_4folate has $[\alpha]_D^{27} = +14.9°$ (0.1 N NaOH). Natural (6*S*)-H_4folate [*71963-69-4*] has $[\alpha]_D^{27} = -16.9°$ (0.1 N NaOH).

5-Methyltetrahydrofolic acid (**3**), 5-CH_3-H_4-folate [*134-35-0*], $C_{20}H_{25}N_7O_6$, M_r 459.46; UV (pH 7): 290 nm (32 000). Synthesis from H_4-folate by condensation with formaldehyde and reduction of the resulting 5,10-CH_2-H_4-folate with sodium borohydride [1230]. **3** is more stable towards oxygen than H_4folate.

5,10-Methylenetetrahydrofolic acid (**4**), 5,10-CH_2-H_4folate [*3432-99-3*], $C_{20}H_{23}N_7O_6$, M_r 457.45; UV (pH 7.2): 294 nm (32 000). Preparation from H_4folate and formaldehyde. Very sensitive to hydrolysis. Synthesis of the natural (6*R*)-stereoisomer by enzymatic reduction of H_2folate and treatment with formaldehyde [1231]. Coenzyme in thymidylate biosynthesis.

10-Formyltetrahydrofolic acid (**6**), 10-HCO-H_4-folate [*2800-34-2*], $C_{20}H_{23}N_7O_7$, M_r 473.45; UV (pH 7.5): 260 nm (17 000). The unstable compound **6** is prepared by hydrogenation of 10-HCO-folate or by hydrolysis of 5,10-CH^+-H_4folate [1232]. Coenzyme in purine synthesis.

5-Formyltetrahydrofolic acid (**5**), (6*RS*)-5-HCO-H_4folate, folinic acid, leucovorin [*58-05-9*], $C_{20}H_{23}N_7O_7$, M_r 473.45; UV (pH 13): 282 nm (32 600); $[\alpha]_D^{25} = +16.0°$ (H_2O, Ca salt). Most stable of the H_4 folate cofactors. Preparation by hydrolysis of 5,10-CH^+-H_4folate in neutral or alkaline solution or by direct formylation of H_4folate [1233]. (6*S*)-5-HCO-H_4folate, citrovorum factor, *l*-leucovorin [*68538-85-2*]; $[\alpha]_D = -13.3°$ (H_2O). Enzymatic synthesis of the natural (6*S*)-stereoisomer from H_2folate [1234, 1235]. Preferential crystallization of the Ca salt of natural (6*S*)-5-HCO-H_4folate from (6*RS*)-5-HCO-H_4folate in water [1236]. 5-HCO-H_4folate is used in medicine as an antidote to folic acid antagonists such as methotrexate.

5-Formiminotetrahydrofolic acid (**7**), 5-NHCH-H_4folate [*2311-81-1*], $C_{20}H_{24}N_8O_6$, M_r 472.46; UV (pH 7): 285 nm (35 400). Fast hydrolysis in aqueous solution.

5,10-Methenyltetrahydrofolic acid (**8**), 5,10-CH^+-H_4folate, anhydroleucovorin [*65981-89-7*], $C_{20}H_{22}N_7O_6^+$ (cation); UV (pH 1): 352 nm (23 900). Preparation by dehydration of 5-HCO-H_4folate in acidic medium [1235].

13.4. Content in Food and Bioavailability

Folic acid is synthesized de novo by microorganisms and plants; mammals require it in their diet. Folates are present in all food products of animal and plant origin, especially in green-leaved vegetables. Liver has a notably high content. The folate content (in µg/100 g) of some foods is given in the following [1237–1241]:

Lettuce	106 – 200
Spinach	78 – 194
Asparagus	50 – 195
Cabbage	30 – 79
Broccoli	71 – 189
Tomatoes	6 – 73.4
Beans	68 – 100
Apples	3 – 6
Oranges	24 – 37
Bananas	22 – 36
Avocados	30 – 36
Beef	5 – 18
Beef liver	141 – 330
Kidney	63
Chicken	5 – 28
Eggs	65 – 140
Milk	5 – 20.5
Cheese	20 – 66
Fish	4.5 – 15

Most sources list the total folate content [1237–1240], but occasionally a distinction is made between mono- and polyglutamates. To take into account the lower bioavailability of polyglutamates, a folate equivalent can be calculated [1241]. The wide range of folate contents most likely reflects different methods used for determination.

Food processing, storage, and cooking reduce the content of folates considerably [1242–1244]. Different forms of folate vary in stability. In particular, oxidation results in inactive cleavage products. Folates can be stabilized for longer periods in the presence of reducing agents such as ascorbate.

In humans, folic acid is almost completely absorbed under fasting conditions. The absorption of folylpolyglutamates is 50 – 80 % of that of monoglutamates as judged by urinary excretion [1245, 1246]. This percentage may be lowered in the presence of many foods.

13.5. Biosynthesis

Folic acid is synthesized in microorganisms and plants from the precursors guanosine 5′-triphosphate (GTP), *p*-aminobenzoic acid, and L-glutamic acid. The enzymatic steps involved in folate biosynthesis in enteric bacteria have been extensively reviewed by Brown and Williamson [1247, 1248]. The pathway is shown in Figure 22 A. GTP is converted by GTP cyclohydrolase I (EC 3.5.4.16) to 7,8-dihydroneopterin triphosphate. This transformation includes the hydrolysis of the annelated imidazole ring, elimination of the C-8 atom as formate, Amadori rearrangement, and ring closure to the pyrazine ring involving the ribose C-2′ atom. The pyrophosphate group is subsequently hydrolyzed chemically [1249] or by a nonspecific pyrophosphatase so far only described in *E. coli* [1250]. The residual phosphate is removed by an as yet uncharacterized phosphomonoesterase. Hydroxymethyl-7,8-dihydropterin pyrophosphate is then obtained by a retro-aldol reaction with removal of a hydroxyacetaldehyde moiety and subsequent addition of a pyrophosphate group. Condensation of *p*-aminobenzoic acid with hydroxymethyl-7,8-dihydropterin py-rophosphate is carried out by dihydropteroate synthase (E.C. 2.5.1.15) leading to 7,8-dihydropteroate. L-Glutamate is then attached to this intermediate by dihydrofolate synthetase. In some microorganisms, this enzyme also has folylpolyglutamate synthetase activity [1251].

Organization of genes encoding for the folate biosynthetic enzymes varies greatly among bacteria. In *Bacillus subtilis* and *Streptococcus pneumoniae* folate operons were shown to carry six and four biosynthetic genes, respectively [1252, 1253], while in *E. coli* biosynthetic genes are spread along the chromosome. In *B. subtilis*, genes for folate biosynthesis were found at three different locations on the chromosome (Figure 22B). They have been cloned and sequenced [1252, 1254, 1255]. First, the *mtr* operon maps at position 204° and encodes GTP cyclohydrolase I and tryptophan RNA-binding attenuation protein (TRAP) [1254]. Second, the *folC* gene encoding dihydrofolate synthetase maps at position 258° and is not linked to the other genes of folate biosynthesis [1255]. Third, a folic acid biosynthetic operon mapping at po-

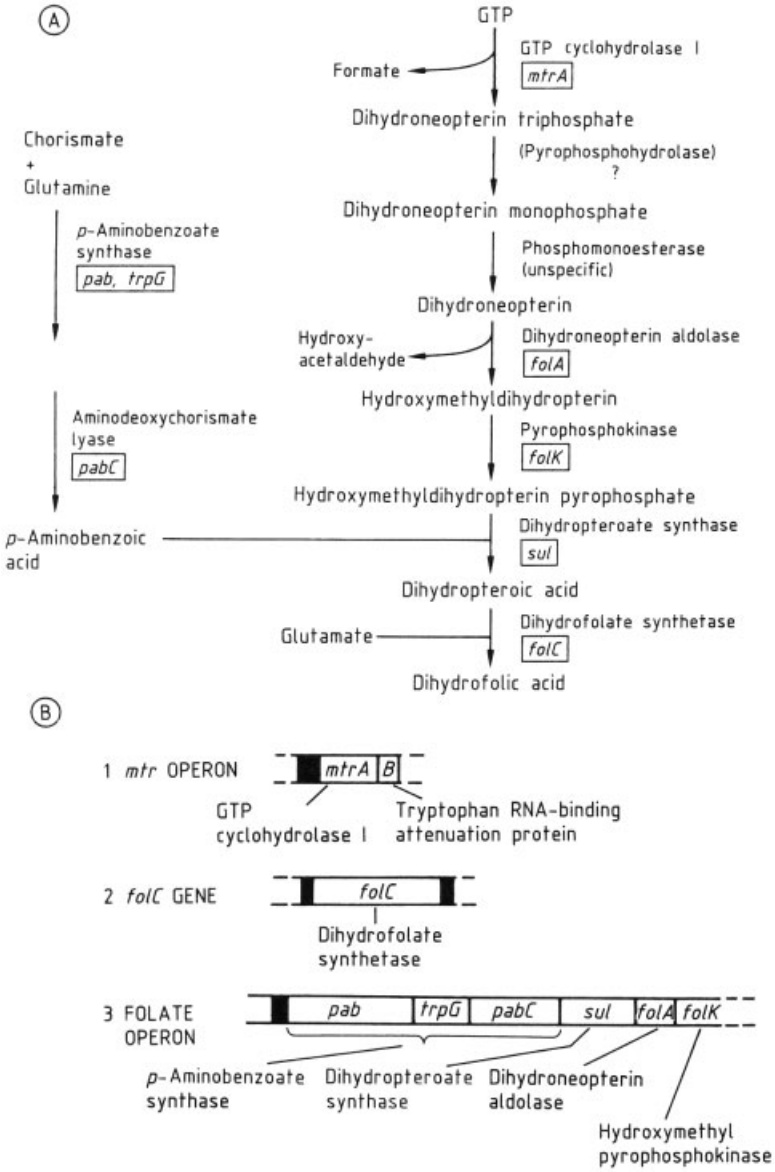

Figure 22. Biosynthesis of folic acid
A) Biochemical pathway; B) Genes encoding enzymes of folate biosynthesis in *Bacillus subtilis*

sition 10° encodes for the three subunits of *p*-aminobenzoate synthase, dihydroneopterin aldolase, hydroxymethyl pyrophosphokinase, and dihydropteroate synthase [1252].

Industrial production of folic acid by genetically engineered microorganisms or extraction from natural sources is not yet economically viable.

13.6. Chemical Synthesis

Folic acid (**1**) was first synthesized by condensation of 2,5,6-triamino-4(3*H*)-pyrimidinone (**9**), *p*-aminobenzoyl-L-glutamic acid (**10**), and 2,3-dibromopropanal (**11**) [1222, 1256].

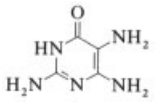
9

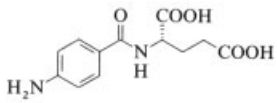
10

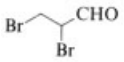
11

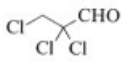
12

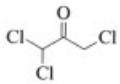
13

Later variants of this one-pot process replaced the C_3 component **11** by various other compounds, such as 1,1,3-tribromo-2-propanone [1257], 2,2,3-tribromopropanal [1257], 2,2,3-trichloropropanal (**12**) and 1,1,3-trichloro-2-propanone (**13**) [1258]. Addition of sodium bisulfite to the reaction medium substantially increased the yield of folic acid [1258].

Syntheses of folic acid by stepwise condensation of **9**, **10**, and a C_3 component have also been reported. For example, 2-amino-6-(hydroxymethyl)pteridinone (**14**), prepared by condensation of **9** with 1,3-dihydroxy-2-propanone, was reacted with **10** to give folic acid (**1**) [1259]. In a different two-step synthesis, 2-hydroxymalondialdehyde was condensed with two equivalents of **10**. Subsequent reaction of the resulting diimine with **9** provided folic acid [1260].

Condensation of triaminopyrimidinone (**9**) with unsymmetrical C_3 compounds invariably leads to regioisomeric byproducts substituted in position 7.

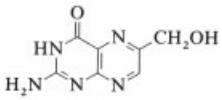
14

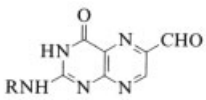
15a: R = H
15b: R = $COCH_3$

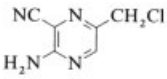
16

6-Formylpterin (**15b**) is another viable intermediate for the synthesis of folic acid. It can be prepared from triaminopyrimidinone (**9**) and 2-bromo-3,3-diethoxypropanal with H_2O_2 as oxidizing agent, followed by acetylation and hydrolysis [1261]. A synthesis of **15a** from the pyrazine derivative **16** has been developed as a route giving a single regioisomer [1262]. Finally, condensation of **15b** with the diester of **10**, reduction of the Schiff base with sodium borohydride and hydrolysis leads to folic acid [1263].

The various schemes to synthesize folic acid are extensively discussed in [1225].

13.7. Metabolism and Biochemical Functions

Prior to absorption by the jejunum, dietary folylpolyglutamates are hydrolyzed to monoglutamates by folylpolyglutamate hydrolases (conjugases). Deconjugation occurs at the surface of the absorptive cells and within lysosomes in enterocytes [1264–1267]. Intestinal absorption and transport are optimal in a slightly acidic environment at pH 6 [1268]. A single protein seems to be responsible for the transport of the monoglutamates, as shown by competitive inhibition [1269]. Some conversion of folates, mostly to 5-CH_3-H_4folate, can occur during the passage across the intestinal mucosa and before release into the portal circulation. At pharmacological doses ($\geq$1 mg), significant amounts of folate appear unchanged in portal blood. They are converted to methylated forms in the liver [1270, 1271]. Apart from H_4folate and 10-HCO-H_4folate, mainly 5-CH_3-H_4folate is observed in blood, bound to plasma proteins such as albumin, α-macroglobulin, and transferrin (low-affinity binders). High-affinity folate-binding proteins are also known, but their functions are not yet well understood [1272, 1273]. The enyzme folylpoly-γ-glutamate synthetase converts folic acid to folylpolyglutamates and is located in the cytosol as well as in mitochondria [1274, 1275]. The liver exhibits the highest activity of this enzyme; appreciable amounts are also observed in most other mammalian tissues [1276]. Folylpolyglutamates seem to be the storage forms, but they can also increase the specificity of some folate-dependent enzymes. The enzyme dihydrofolate reductase transforms the different forms of folate to the active coenzymes, the tetrahydrofolylglutamates, whereby NADPH functions as hydrogen donor. Tetrahydrofolates are involved in a variety of biochemical processes, such as methylation reactions, amino acid metabolism and biosynthesis of purines and pyrimidines [1239, 1277]. They

acquire C_1 units at N-5 and/or N-10 and function as coenzymes in the transfer and utilization of C_1 fragments of different oxidation states, such as methanol, formaldehyde, and formate.

In humans, plasma levels of 7 – 17 µg/L are found. The concentration of folate in erythrocytes is ca. 200 – 500 µg/L. The total folate pool in adult humans is estimated to be ca. 5 – 10 mg, half of which is stored in the liver. The daily excretion of folate in urine is ca. 40 µg and in feces 200 µg [1278, 1279].

13.8. Nutritional Requirements and Medical Use

Recommended Daily Intake (RDI). Approaches that have been used to estimate folate requirements in humans are the determination of folate amounts needed to normalize the folate status after depletion, the evaluation of the average daily folate intake, the measurements of urinary metabolites, and kinetic studies. Based on several studies, the official recommendation in the United States was lowered in 1989 and is set at 3 µg/kg body weight per day, resulting in ca. 200 µg folate for adult males and 180 µg for adult females [1280]. Since pregnancy and lactation are often associated with poor folate status, 400 µg/d and 260 – 280 µg/d, respectively, is recommended. The recommended intake for infants from birth until 1 year of age is set at 3.6 µg/kg body weight per day (25 – 35 µg). This amount is also recommended for children between 1 and 10 years of age (50 – 150 µg). As pointed out by Bailey [1240], the present U.S. value [1280] may not be sufficient to maintain normal folate levels and will most likely have to be reevaluated.

Medical Use. Folate deficiency leads to impaired amino acid metabolism, protein synthesis, and cell division. The effects are most noticeable in growing tissue. A major clinical problem associated with folate deficiency is macrocytic anemia in which the bone marrow produces giant, immature red blood cells. An increased incidence of megaloblastic anemia has been reported during pregnancy and following child birth. Folate antagonists (e.g., methotrexate and certain antiepileptics) are used in therapy for various diseases, but their administration can lead to a functional folate deficiency. In cases of infections (e.g., protozoal infections in AIDS patients) or certain types of cancer (e.g., choriocarcinoma, cancer of the head and neck, childhood acute lymphocytic leukemia) high doses of 5-HCO-H_4folate (5 – 30 mg) may be given as "rescue dose" [1281–1283]. Precancerous conditions such as bronchial squamous metaplasia improved in some smokers after taking 10 mg/d [1284].

Homocystinemia is a heterogeneous group of inherited diseases of folate, vitamin B_{12}, and homocysteine metabolism and is associated with severe arteriosclerosis and thromboembolism as well as mental retardation. Of the inherited folate disorders, deficiency of methylenetetrahydrofolate reductase [1285] reduces or eliminates the generation of 5-CH_3-H_4folate, which is needed as cosubstrate to convert homocysteine to methionine by methionine synthase. In men with homocysteinemia, daily administration for 6 weeks of 1 mg folic acid, 10 mg vitamin B_6, and 0.4 mg vitamin B_{12} normalized serum homocysteine levels [1286]. In psoriasis patients receiving methotrexate, increased plasma homocysteine levels could be lowered with folic acid [1287, 1288]. In healthy men and women, a significant inverse correlation between plasma levels of folate and vitamin B_{12} and total plasma homocysteine was observed [1289, 1290].

Neural tube defects (NTDs), including spina bifida and anencephaly, are polygenic conditions. However, epidemiological studies have shown an increased incidence of NTDs associated with environmental factors. Several studies have provided evidence that multivitamin or folic acid (0.8 – 4 mg/day) supplementation prevents the majority of these NTDs [1291–1294]. Health authorities in the United States and United Kingdom advise women of childbearing age to take daily folate supplements of 0.4 mg [1295, 1296].

Toxicity. Daily oral supplementation with 5 – 10 mg of folic acid is not toxic for normal non-pregnant humans [1297]. Problems that might occur are the masking of vitamin B_{12} deficiency resulting in irreversible damage to spinal cord (subacute combined degeneration of the cord), and recurrence of epilepsy in epileptics treated with drugs with antifolate activity. A hypervitaminosis has not yet been reported. The

acute toxicity (LD_{50}) is 500 and 600 mg/kg body weight for rats and mice, respectively [1298].

Animal Nutrition. In animal nutrition, folic acid supplementation is required to obtain optimal performances of farm animals [1299]. Accordingly, doses of 0.2 – 0.5 mg/kg feed (dry matter) and 0.3 mg/kg for chickens and pigs are required, respectively. Under commercial production conditions, however, increased folic acid supplementation is recommended [1300]. Thus, the recommended amounts are 0.5 – 1.0 mg/kg feed for chickens and 0.5 – 2.0 mg/kg for swine.

13.9. Analysis

Analytical procedures, such as physicochemical, chromatographic, radioisotopic, and microbiological methods, are available to determine folate concentrations. However, analysis is still very difficult, because of the many naturally occurring derivatives and their rapid conversion by numerous enzymes, their lability, and the low amounts in biological samples. Methods for determining the C_1 unit, the reductive state of the cofactor, or the chain length of folylpolyglutamates are also important. For general discussions and detailed procedures see [1301–1303]. Pure folic acid, folate in enriched preparations or various monoglutamates can be measured by spectrophotometry, polarography, or HPLC [1304–1307]. Several microorganisms such as *Lactobacillus casei*, *Streptococcus faecalis*, or *Pediococcus cerevisiae* do not synthesize folate and can, therefore, be used to measure folate from natural sources [1308]. Since folylpolyglutamates with more than 3 glutamate residues are less active or inactive, hydrolysis of polyglutamate chains by γ-glutamyl hydrolase is required to obtain maximal growth of the microorganisms. Differential analysis of folates can be obtained by using *P. cerevisiae*, as this microorganism does not respond to nonreduced folate or 5-CH_3-H_4folate. To determine folateactive substances in erythrocytes and plasma, a number of radioisotope kits based on competitive protein binding are commercially available. Various folates, however, bind with different affinities to the corresponding folate binding proteins, resulting in determination of preferential folates. In addition, the results obtained with different test kits fluctuate considerably. Interpretation of the results with regard to the folate status may, therefore, pose problems.

13.10. Economic Aspects

The present world market for folic acid is ca. 400 t/a, commercially available as pure crystalline material (water content 8 %) or as premix. The major part (ca. 80 %) is used for feed enrichment in animal nutrition. The pharmaceutical industry offers folic acid for therapeutic and prophylactic use, either as a monopreparation or in combination with iron and/or other vitamins. Folic acid is also increasingly used for food enrichment (flour, cornflakes). The major producers of folic acid are Hoffmann-La Roche; the Japanese companies Takeda, Sumica Fine Chemicals (previously Yodogawa Pharmaceuticals), and Kongo; the Chinese companies Changzhou Pharma and Changshu Huagang Pharma; as well as some smaller companies in China, India, and Russia. The sales price in 1995 varied between € 36 and € 77/kg.

14. Niacin (Nicotinic Acid, Nicotinamide)

14.1. Introduction

Niacin is one of the vitamins of the B complex. In accordance with the rules on nomenclature the Institute of Nutrition [1309] suggested that niacin be used as the generic name for both nicotinic acid and nicotinamide.

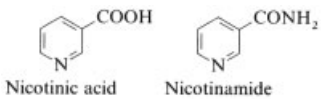

Other commonly used names include pyridine 3-carboxylic acid and vitamin PP for nicotinic acid, and pyridine 3-carboxylic acid amide, pyridine 3-carboxamide, vitamin PP, and niacinamide for nicotinamide.

Nicotinic acid was first synthesized in 1867 by oxidative degradation of nicotine. It had been chemically identified long before its importance

as an essential nutrient (pellagra-preventing vitamin) was recognized [1310].

In the early 1900s it was discovered that pellagra, a disease with symptoms affecting the skin, gastrointestinal tract, and central nervous system, was caused by nicotinic acid deficiency. Experiments in 1917 with yeast, known for its high natural nicotinic acid content of up to 500 mg/kg, showed a curative effect on pellagra [1310, 1311].

In 1935 nicotinamide was shown to be a moiety of the coenzymes NAD (nicotinamide adenine dinucleotide) and NADP (nicotinamide adenine dinucleotide phosphate). These coenzymes are indispensable for many biochemical reactions in all living cells [1310].

Niacin is a naturally occurring substance. Nicotinic acid is the form present in food of plant origin, whereas nicotinamide occurs in animal products. Not all of the natural niacin is fully bioavailable for humans and animals. Nicotinic acid in cereals and in oil seeds is present in a bound complex. Average contents of niacin are in the range of 10 – 200 mg/kg of product.

The symptoms of niacin deficiency are usually described as a disease of the "three Ds": dermatitis, diarrhea, and dementia. However, the three Ds represent the last stage of the disease. The early stages include lassitude, impaired digestion, and inflammation of the mucous membranes of the mouth, stomach, or intestine.

In developed countries, pellagra symptoms are found in connection with chronic alcoholism and one-sided nutrition. Animals exhibit skin problems, diarrhea, weight loss and decreased feed intake.

Niacin is a component of multivitamin preparations, breakfast cereals, and soft drinks. Flour is enriched with niacin. For domestic animals niacin is an important part of compound feed. More than 60 % of the niacin produced is required by poultry, swine, ruminants, fish, and pets.

14.2. Physical and Chemical Properties

Physical properties of nicotinic acid and nicotinamide are summarized in Table 38. The acid and the amide differ substantially in melting point and solubility. Most producers guarantee stability of 2 – 3 years. However, nicotinamide is a hygroscopic substance and tends to form lumps. It is therefore recommended that storage, in a closed package and under cool, dry conditions, should not exceed six months. Aqueous solutions are stable, and sterilization is possible without special precautions.

Table 38. Physical properties of niacin

	Nicotinic acid	Nicotinamide
IUPAC name	pyridine 3-carboxylic acid	pyridine 3-carboxamide
CAS no.	[*59-67-6*]	[*98-92-0*]
Formula	$C_6H_5NO_2$	$C_6H_6N_2O$
Molecular mass	123.11	122.13
Appearance	white crystals	white crystals
Melting point	234 – 237 °C (>238 °C decomp.)	128 – 131 °C
Solubility in water	1.3 g/100 mL (15 °C)	38.3 g/100 mL (20 °C)
	9.8 g/100 mL (100 °C)	

14.3. Biochemical Functions

Niacin is absorbed almost completely by simple diffusion across the intestinal mucosa. The absorbed niacin is taken up by the tissues, and is thought to be incorporated directly into coenzyme forms.

Niacin serves as the precursor of two essential coenzymes: nicotinamide adenine dinucleotide (NAD) and nicotinamide adenine dinucleotide phosphate (NADP) [1310, 1312, 1313]. Both coenzymes catalyze the metabolic transfer of hydrogen – one of the basic functions in the metabolism of proteins, fats, and carbohydrates. This function is required for both the synthesis and degradation of amino acids, fatty acids, and carbohydrates.

Another important task of the niacin coenzymes is their repeated intervention in the citric acid cycle. The citric acid or Krebs cycle comprises many steps in which activated acetate is repeatedly oxidized.

Compounds originating from the degradation of fatty acids, carbohydrates and amino acids are first broken down to pyruvate oxalacetate and then enter the cycle with the active acetate. The various oxidative steps of the cycle yield a considerable amount of energy, which is stored in adenosine triphosphate (ATP). This energy can be released by conversion of ATP to adenosine

diphosphate (ADP), which is then ready to pick up more energy. Niacin thus plays an essential part in the metabolic production and utilization of energy.

14.4. Production

Industrial production of nicotinic acid and nicotinamide are based on the raw materials 5-ethyl-2-methylpyridine and 3-methylpyridine (3-picoline).

3-Methylpyridine is obtained by catalytic gas-phase reaction of acetaldehyde, formaldehyde, and ammonia (Figure 23). However, the principal product is pyridine, while 3-methylpyridine is obtained as a byproduct in 30 – 50 % yield. 5-Ethyl-2-methylpyridine is produced by reaction of paraldehyde with ammonia (→ Pyridine and Pyridine Derivatives, Chap. 2.2.6.).

Nicotinic acid is obtained directly by nitric acid oxidation of 5-ethyl-2-methylpyridine (Fig. 23) or by total hydrolysis of 3-cyanopyridine. Nicotinamide is available either from the amidation of nicotinic acid or from partial hydrolysis of 3-cyanopyridine. The latter is obtained by ammoxidation of 3-methylpyridine.

Oxidation of 5-Ethyl-2-methylpyridine. The world's largest plant for the production of niacin is located in Visp, Switzerland, and is operated by Lonza.

The principles of the Lonza process are summarized in the following and in the flow diagram of Figure 24 [1314]:

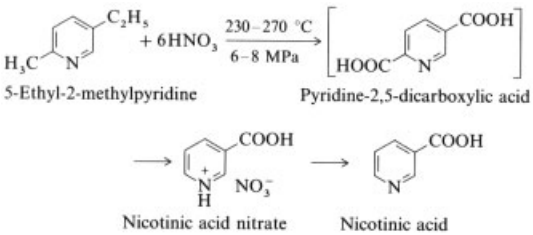

In an aqueous solution 5-ethyl-2-methylpyridine is oxidized to nicotinic acid in a single-tube reactor by using an excess of nitric acid. The continuous process takes place at 230 – 270 °C and 6 – 8 MPa. Pyridine-2,5-dicarboxylic acid is formed as an intermediate, but is not stable under the reaction conditions, undergoing decarboxylation to nicotinic acid nitrate. Neutralization of the nitrate with 5-ethyl-2-methylpyridine yields nicotinic acid. Water and carbon dioxide formed in the reaction are separated from the product. The nitrogen oxide formed is oxidized with air to nitrogen dioxide and then absorbed in water to yield nitric acid, which is recycled into the process.

Ammoxidation of 3-Methylpyridine and Hydrolysis of 3-Cyanopyridine (→ Pyridine and Pyridine Derivatives, Chap. 3.7.).

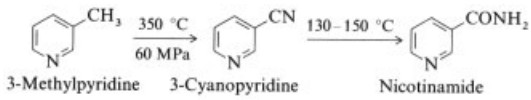

In a multitubular reactor 3-methylpyridine, air, ammonia, and hydrogen react at ca. 350 °C and moderate pressure to give 3-cyanopyridine [1315]. Heterogenous catalysts containing oxides of antimony, vanadium and titanium [1316] or antimony, vanadium and uranium [1317] are highly effective and commonly used. For instance, with an antimony – vanadium – titanium catalyst, a reactor temperature of 360 °C, and a molar feed ratio of 3-methylpyridine : ammonia: air : water of 1 : 6 : 30 : 6 yields 96 % of 3-cyanopyridine [1318]. 3-Cyanopyridine is converted to nicotinamide by alkaline hydrolysis. This reaction has the advantage that saponification to the amide is fast compared to total hydrolysis to nicotinic acid. The hydrolysis to the amide is normally carried out with catalytic amounts of bases, mainly sodium hydroxide, at 130 – 150 °C [1319]. Other catalysts are also mentioned. A copper – chromium oxide catalyst gives good yields of nicotinamide [1320].

Conversion of 2-Methylglutaronitrile. 2-Methylglutaronitrile, a byproduct of adiponitrile production, is converted to 2-methyl-1,5-diaminopentane. Cyclic hydrogenation gives 3-methylpiperidine. Dehydrogenation yields 3-methylpyridine [1321], which is then ammoxidated and partly hydrolyzed to nicotinamide.

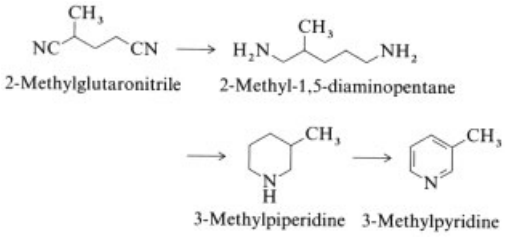

14.5. Quality Specifications

Nicotinic acid and nicotinamide are available in various forms. Pharmaceutical grade meets the requirements of current pharmacopoeias, and food grade, which is usually equal to pharmaceutical grade in quality, meets the standards of USP and FCC (Food Chemical Codex). These products have an assay of >99.8 % of the active substance. The technical and feed grades of the leading producers have a purity of min. 99.5 % (nicotinic acid) and min. 99 % (nicotinamide). Lonza offers a DC quality (DC = direct compression), which is preferably used in the tabletting of vitamin or multivitamin tablets without auxiliary substances, thus improving the process in terms of energy consumption and production time.

There are only minor differences in product quality caused by variations in production processes or in the raw material used.

14.6. Analysis

Assays of the pure substances are most readily determined by titration. Nicotinic acid is determined by titration with sodium hydroxide [1322] or UV spectroscopy [1323]. Nicotinamide is determined by titration with perchloric acid in acetic acid [1324] or UV spectroscopy [1325]. For the pharmaceutical grade, the official standards of purity of the pharmacopoeias, such as USP [1323, 1325] or European Pharmacopoeia [1324, 1326] apply.

To ensure a vitamin supply oriented to meeting requirements in human and animal nutrition, foods and feeds are enriched with niacin or mixtures of vitamins. Microbiological [1327], spectrophotometric [1328], and chromatographic [1329] procedures are described for the quantitative determination of nicotinic acid or nicotinamide in foods and feeds. A most suitable method is the AOAC spectrophotometric method [1328]. Microbiological methods and the above-mentioned photometric method have the disadvantage that only the total amount of nicotinic acid and nicotinamide can be detected. With HPLC it is possible to determine the single compounds [1330, 1331].

14.7. Uses

Pellagra in humans was formerly widespread in many parts of the world where corn was the main source of nutrition. Subclinical symptoms of deficiencies still appear under conditions of malnutrition or one-sided nutrition. Recommended dietary allowances for humans are listed in Table 39 [1332].

Table 39. Recommended dietary allowances for niacin

	Age, a	Weight, kg	Niacin, mg/d
Infants	0 – 0.5	6	5
	0.5 – 1	9	6
Children	1 – 3	13	9
	4 – 6	20	12
	7 – 10	28	13
Men	11 – 14	45	17
	15 – 18	66	20
	19 – 24	72	19
	25 – 50	79	19
	>51	77	15
Women	11 – 14	46	15
	15 – 18	55	15
	19 – 24	58	15
	25 – 50	63	15
	>51	65	13
Pregnant women			17
Nursing mothers			20

Enrichment of cereal products with B vitamins, especially with nicotinic acid or nicotinamide, became feasible in the early 1940s when the vitamin became commercially available, and is now standard practice [1333]. Niacin is present in breakfast cereals, multivitamin drinks, and multivitamin tablets [1334].

Niacin is also essential to most animals [1335]. The requirements can be covered by various sources. Biosynthesis from tryptophan is not of importance since the conversion rate is very low and most natural feedstuffs are relatively poor in this amino acid. Nicotinic acid

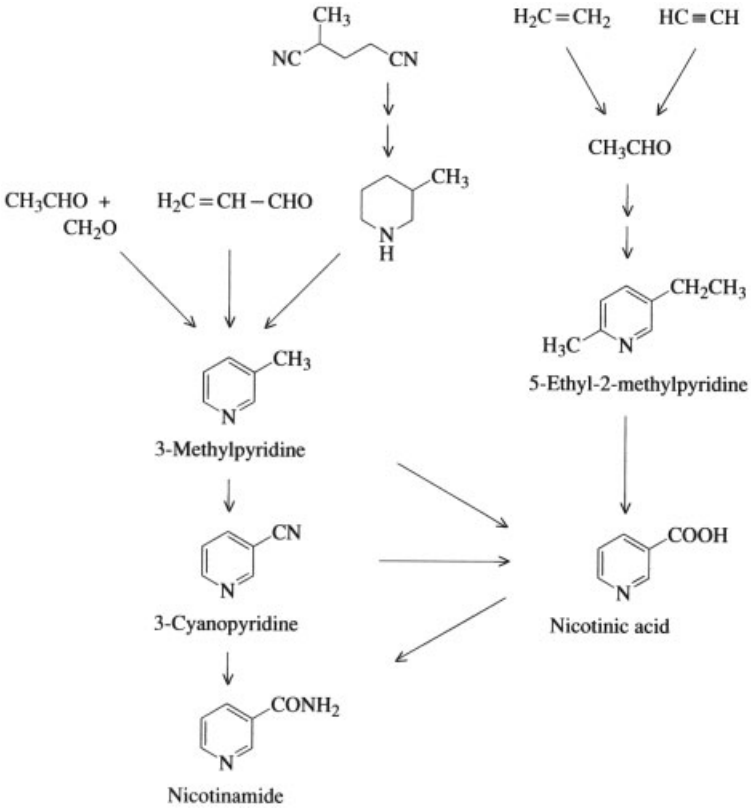

Figure 23. Industrial production of nicotinic acid and nicotinamide

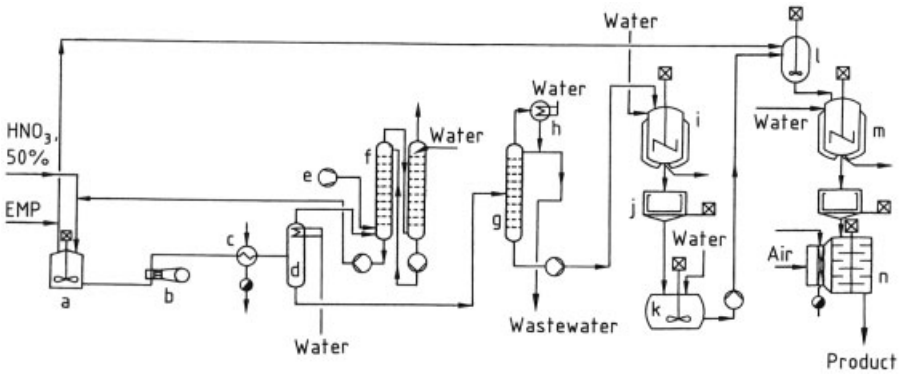

Figure 24. Flow diagram of the oxidation of 5-ethyl-2-methylpyridine to nicotinic acid (Lonza process)
a) Reactor feed preparation; b) Piston pump; c) Steam-heated reactor; d) Separator; e) Air compressors; f) Absorption columns for nitrogen oxides; g) Concentration column; h) Condensor; i) Crystallization; j) Centrifugation; k) Dissolver; l) Neutralization; m) Crystallization; n) Tray dryer

is synthesized by the microflora in the large intestineof monogastrics and in the rumen of ruminants. In monogastrics, biosynthesis is insignificant because it occurs in the large intestine, located after the main niacin-absorption sites of the duodenum and small intestines. Most of the nicotinic acid synthesized by the microflora is therefore excreted in the feces. Synthesis of nicotinic acid in the rumen is suboptimal when the animals are under stress for high production or when the rumen microflora has been disturbed. The nicotinic acid production by the ru-

men microflora may be affected by antiobiotics and other feed additives.

Niacin deficiency reduces the activity of the coenzymes NAD and NADP, resulting in metabolic disorders and hence in reduced weight gain, decreased feed intake, reduced feed efficiency, health problems (e.g., ketosis), and impaired reproduction.

To achieve peak performance and to maintain good health all domestic animals and pets need a dietary source of niacin to meet their requirements. Niacin levels should therefore be higher than the NRC (National Research Council) recommendations [1336], because it is generally accepted that the NRC recommendations are based on ideal dietary and management conditions, since no safety margin is included or limited bioavailability considered. Recommendations for niacin addition to animal feed (in mg/kg feed) are summarized in the following [1335]:

Poultry	
Chicks	50
Broilers	60
Hens (laying)	40
Hens (breeding)	50
Turkeys (starting)	90
Turkeys (growing and fattening)	80
Turkeys (breeding)	80
Ducks and geese (starting)	50
Pigs	
Piglets (early-weaned)	40
Piglets (starting)	35
Pigs (growing)	35
Pigs (fattening)	30
Sows (breeding)	35
Ruminants	
Calves (0 – 3 months)	60
Cattle (fattening)	1000 mg per animal per day
Dairy cows	3000 – 6000 mg per animal per day
Sheep and goats	1000 mg per animal per day
Horses	
Foals and yearlings	40
Working horses	90
Race horses	90
Pets	
Dogs	25
Cats	70
Fish	
Cyprinids	100
Salmonids	200
Shrimps	200
Others	
Rabbits	50
Mink	50
Foxes	50

Nicotinic acid and nicotinamide used in doses in excess of nutritional requirements generate a number of readily recognizable pharmacological responses, some of which are of therapeutic significance. In view of the relationship of nicotinic acid to nicotinamide it is remarkable that few if any of the considerable pharmacological activities of nicotinic acid are shared by nicotinamide.

Some of the observed effects of pharmacological doses of nicotinic acid are as follows:

1) Effects on lipid metabolism: Plasma cholesterol and triglycerides concentrations fell significantly with 1 – 3 g/d doses of nicotinic acid [1310, 1337, 1338]. Different products with nicotinic acid or esters of nicotinic acid are commercially available.
2) Vasodilation: The flush reaction caused by the increased blood flow rate appears very quickly [1310, 1313]. This effect is also used in medicine [1310, 1313] (e.g., in treatment of smokers' legs).
3) Activation of fibrinolysis [1310].

Pharmacological doses of nicotinamide have been shown to improve islet-cell regeneration and appear to have a protective effect on residual insulin secretion in Type 1 diabetic patients. Thus, it is conceivable that nicotinamide may have a place in preventing diabetes in predisposed subjects [1339].

Niacin is used in electroplating baths. Zinc electroplating baths contain niacin in a quaternized form obtained by the reaction of niacin with benzyl chloride, resulting in uniform and brilliant surfaces.

14.8. Economic Aspects

In 1995, ca. 22 000 t of nicotinic acid and nicotinamide were produced worldwide. 16 000 t or ca. 75 % of total niacin production was used in animal nutrition and 6000 t as food additive. The most important markets are the United States with a market share of 35 – 40 %, followed by Europe with 30 %, and Japan and South East Asia with 20 %.

Table 40 shows the development of the niacin market over the last 25 years. Producers of nicotinic acid and nicotinamide and their capacities are presented in Table 41.

Table 40. Niacin market 1970 – 1995 in t

	1970	1980	1995
Nicotinic acid	3000	9 000	14 000
Nicotinamide	1000	3 000	8 000
Total	4000	12 000	22 000

Table 41. Producers and their capacities of nicotinic acid and nicotinamide in 1995

Producer	Location	Nicotinic acid, t/a	Nicotinamide, t/a
Lonza	Visp, Switzerland	13 000	1500
	China ∗∗		3000
Vitachem ∗	Antwerpen, Belgium	1 000	3500
	Indianapolis, USA		3500
Nepera	Harriman, USA		3000
Yuki Gosei	Japan	200	1000
Others		1 000	500

∗ Joint Venture of Degussa and Reilly Tar.
∗∗ Joint Venture of Lonza/GPF Guangzhou, start-up planned for 1998.

Demand for niacin is rising by ca. 2 – 3 %/a in the industrialized countries, while it is soaring by as much as 10 %/a in the developing and newly industrialized countries.

14.9. Toxicology

Nicotinic acid and nicotinamide are essential for human and animal health. The daily requirement to avoid deficiencies in humans is in the range of 15 – 30 mg. For the treatment of hyperlipoproteinemia and hypercholesterinemia daily doses up to 6000 mg of nicotinic acid are used [1337].

Nicotinic acid and nicotinamide are acutely nontoxic (oral LD_{50} greater than 5000 and 3000 mg/kg in rats, respectively), not irritant to the skin but moderately irritant to the eye of rabbits [1340].

Nicotinic acid, up to 1000 mg kg^{-1} d^{-1}, did not show either cumulative toxicity in a 28 d oral study or teratogenicity in rats [1340]. No mutagenicity was found in a battery of mutagenicity tests [1340], and no carcinogenicity was observed in mice treated for a lifetime with nicotinic acid in drinking water [1341].

Rare cases of skin flushing may occur in humans, but this effect is reversible after termination of exposure to nicotinic acid.

Nicotinic acid is practically nontoxic to fishes, daphnia, algae, and bacteria and is readily biodegradable [1340].

Based on the available information nicotinic acid and nicotinamide do not represent a hazard to humans or the environment.

15. References

1. K. Folkers, *Int. J. Vitam. Nutr. Res.* **39** (1969) 334.
2. J.-C. Laubscher, SLZ Schweiz. Lab.-Z. **52** (1995) 42.
3. H. van Elen Berg, *Ber. Bundesforschungsanst. Ernähr.* **BFE** (1993) Bioavailability 193, part 2, 267.
4. R. B. Rucker, T. Stites, *Nutrition* (Syracuse, N. Y.) 1994, 507.
5. C. Funk, *J. Physiol. (London)* **43** (1911/12) 395.
6. C. Eijkmann, *Virchows Arch. Pathol. Anat. Physiol.* **148** (1897) 523.
7. C. Drummond, *Biochem. J.* **14** (1920). 660.
8. Deutsche Gesellschaft für Ernährung: *Empfehlungen für die Nährstoffzufuhr,* 5th. ed. Umschau Verlag, Frankfurt 1991.
9. Lemoine, A., C. le Dehevat, in Elevated Dosage of Vitamins (D. Walter, G. Brumbacher, H. Stälrelin, Editors) S. 129 – 147. Hans Huber Verlag, Bern 1989.
10. C. H. Hennekens, J. E. Burins in (I. Pacher, E. Cadenas (editors): *Biol. Oxid. Antioxid.,* Hippokrates Verlag, Stuttgart 1994, p. 249.
11. P. Laupheimer: "Smoking and Free Radicals," *Dtsch. Apoth. Ztg.* **134** (1994) no. 10, 28, 31-2, 35-6.
12. R. A. Riemersma, *Arteroscler. Rev.* **25** (1993) 287 –292.
13. H. Heseker et al.: " Vitaminversorgung Erwachsene in der Bundesrepublik Deutschland", in: W. Kübler, H. J. Anders, W. Heeschen, M. Kohlmeier (eds.): *VERA-Schriftreihe,* Wissenschaftlicher Fachverlag Dr. Fleck, Niederkleen, 1992.
14. Verordnungen für vitaminisierte Lebensmittel vom 01. 09. 42 in der geltenden Fassung.
15. Futtermittelgesetz; Neufassung vom 05. 08. 95, Bundesgesetzblatt 1995, part 1, p. 990.

General References

16. *Ullmann,* 4th ed., **23,** 628.
17. W. H. Sebrell, R. S. Harris: *The Vitamins,* 2nd ed., vols. I – IV, Academic Press, New York 1967.

18. R. A. Morton: "Fat Soluble Vitamins," *Int. Encycl. Food Nutr.* **9** (1970).
19. O. Isler, *Experientia* **26** (1970) 225; **33** (1977) 555.
20. O. Isler, *Carotinoids,* Birkhäuser, Basel 1971.
21. E. C. Grob in R. Ammon, W. Dirscherl (eds.): *Fermente, Hormone, Vitamine,* **vol. III/1,** Thieme Verlag, Stuttgart 1974, p. 162.
22. R. S. Harris, E. Diczfalury, P. L. Munson, J. Glover, *Vitam. Horm. (N.Y.)* **32** (1974) 131.
23. L. M. DeLuca: "The Fat-Soluble Vitamins," in H. F. DeLuca (ed.): *Handbook of Lipid Research,* Plenum Press, New York 1978.
24. D. S. Goodman et al.: "Vitamin A and Retinoids," *Recent Adv. Fed. Proc. Fed. Am. Soc. Exp. Biol.* **38** (1979) no. 11, 2501 – 2543.
25. O. Isler, G. Brubacher: *Vitamine I, Fettlösliche Vitamine,* Thieme Verlag, Stuttgart 1982.
26. W. Friedrich: *Vitamins,* de Gruyter, Berlin 1988.
27. L. J. Machlin: *Handbook of Vitamins,* Marcel Dekker, New York 1984.
28. J. Ganguly, *Biochemistry of Vitamin A,* CRC-Press, Boca Raton, FL, 1989.
29. J. A. Olson, *Food Sci. Technol.* **40** (1991) 1 – 57.
30. K. H. Bässler, E. Grühn: *Vitamin-Lexikon für Ärzte, Apotheker und Ernährungswissenschaftler,* G. Fischer Verlag, Stuttgart 1992.

Specific References

31. Commission on the Nomenclature of Biological Chemistry (IUPAC), *J. Am. Chem. Soc.* **82** (1960) 5581, 5583.
32. E. E. Snell, *Lancet* **17** (1876) no. 3, 8.
33. W. N. Lunin, *Hoppe. Seyler's Z. Physiol. Chem.* **5** (1881) 31.
34. W. Stepp, *Biochem. Z.* **22** (1909) 406. W. Stepp, *Z. Biol.* **57** (1911) 135.
35. E. V. McCollum, C. Kennedy, *J. Biol. Chem.* **24** (1916) 491.
36. J. C. Drummond, *Biochem. J.* **14** (1920) 660.
37. C. Funk, *J. State Med.* **20** (1912) 314.
38. F. H. Carr, E. A. Price, *Biochem. J.* **20** (1926) 497.
39. P. Karrer, R. Morf, K. Schöpp, *Helv. Chim. Acta* **14** (1931) 1040, 1431.
40. H. N. Holmes, R. E. Corbet, *J. Am. Chem. Soc.* **59** (1937) 2042.
41. J. G. Baxter, C. D. Robeson, *J. Am. Chem. Soc.* **64** (1942) 2407, 2411. H. Chatein, M. Debodard, *C. R. Seances Acad. Sci. Ser.* **232** (1951) 355.
42. C. D. Robeson, J. G. Baxter, *J. Am. Chem. Soc.* **69** (1947) 136.
43. G. Wald, *Nature (London)* **139** (1937) 1017. E. Lederer, V. Rosanova, A. E. Gillam, J. M. Heilbronn, *Nature (London)* **140** (1937) 233.
44. E. M. Shantz, *Science (Washington D.C.)* **108** (1948) 417. K. R. Farrar, J. C. Hamlett, H. B. Hambest, E. R. H. Jones, *J. Chem. Soc.* 1952, 2657.
45. U. Schwieter et al., *Helv. Chim. Acta* **45** (1962) 517.
46. H. Wackenroder Geiger's, *Mag. Pharm.* **33** (1831) 144.
47. A. Arnaud, *C. R. Hebd. Seances Acad. Sci.* **104** (1887) 1293.
48. R. Willstätter, W. Mieg, *Justus Liebigs Ann. Chem.* **355** (1907) 1.
49. P. Kuhn, E. Lederer, *Ber. Dtsch. Chem. Ges.* **64** (1931) 1349. P. Karrer, R. Morf, *Helv. Chim. Acta* **14** (1931) 833. P. Karrer, A. Helfenstein, W. Wehrli, B. Pieper, R. Morf, *Helv. Chim. Acta* **52** (1969) 1739. R. Buchecker, *Helv. Chim. Acta* **56** (1973) 2548.
50. R. Kuhn, A. Winterstein, *Helv. Chim. Acta* **11** (1928) 87, 116, 123, 144.
51. L. Zechmeister, L. v. Cholnoky, V. Vrabely, *Ber. Dtsch. Chem. Ges.* **61** (1928) 566. P. Karrer, R. Widmer, *Helv. Chim. Acta* **11** (1928) 751.
52. T. Moore, *Biochem. J.* **23** (1929) 803. H. Euler, V. Demole, P. Karrer, O. Walker, *Helv. Chim. Acta* **13** (1930) 1078.
53. P. Karrer, W. E. Bachmann, *Helv. Chim. Acta* **12** (1929) 28. P. Karrer, A. Helfenstein, *Helv. Chim. Acta* **12** (1929) 1142. P. Karrer, R. Morf, *Helv. Chim. Acta* **14** (1931) 1033; **13** (1930) 1084. R. Kuhn, H. Roth, *Ber. Dtsch. Chem. Ges.* **66** (1933) 1274. P. Karrer, A. Rüegger, A. Geiger, *Helv. Chim. Acta* **21** (1938) 1171. R. Kuhn, H. Brockmann, *Ber. Dtsch. Chem. Ges.* **67** (1934) 1408.
54. P. Karrer, R. Morf. K. Schöpp, *Helv. Chim. Acta* **14** (1931) 1036, 1431. P. Karrer, A. Helfenstein, H. Wehrli, A. Wettstein, *Helv. Chim. Acta* **13** (1930) 1094.
55. K. R. Farrar, J. C. Hamlet, H. B. Henbest, E. R. H. Jones, *J. Chem. Soc.* 1952, 2657.
56. P. Karrer, O. Solmessen, *Helv. Chim. Acta* **20** (1937) 682. P. Karrer, O. Solmessen, W. Gugelmann, *Helv. Chim. Acta* **20** (1937)

1020. G. Brubacher, U. Gloor, O. Wiss, *Chimica* **14** (1960) 19.
57. O. Wiss, H. Thommen, *Wiss. Veröff. Dtsch. Ges. Ernähr.* **9** (1963) 179.
58. G. Wald, R. Hubbardin, P. D. Boyer, H. Lardy, K. Myrbäck (eds.): *The Enzymes,* 2nd ed., Academic Press, New York 1960, vol. 3, p. 369.
59. M. Rosenberger, A. Lovey, J. F. Grippo, *Nature (London)* **355** (1992) 359 – 361. A. Wyss, L. J. Sturzenbecker, J. F. Grippo, P. LeMotte, *Nucleic Acids Res.* **21** (1993) 1231 – 1237.
60. F. Kienzle, *Pure Appl. Chem.* **47** (1976) 183.
61. O. Isler, *Experientia* **33** (1977) 555.
62. H. Matsumoto, R. S. H. Liu, C. J. Simmons, K. Seff, *J. Am. Chem. Soc.* **102** (1980) 4259. M. Denny, M. Chun, R. S. H. Link, *Photochem. Photobiol.* **33** (1981) 267.
63. O. Isler, G. Brubacher: *Vitamine I, Fettlösliche Vitamine,* Thieme Verlag, Stuttgart 1982.
64. W. E. Oberhänsli, H. P. Wagner, O. Isler, *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **30** (1974) 161.
65. T. Hamanaka, T. Mitsui, T. Ashida, M. Kakudo, *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **28** (1972) 241.
66. C. H. Stam, *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **28** (1972) 2936.
67. O. A. Roels in W. H. Sebrell, Jr., R. S. Harris (eds.): *The Vitamins,* **vol. I,** Academic Press, New York 1967, p. 182. L. Zechmeister: *Cis-trans-Isomeric Carotenoid, Vitamin A and Aryl Polyenes,* Springer Verlag, Wien, 1962. J. P. Sweeney, A. C. Marsh, *J. Nutr.* **103** (1973) 20. J. C. Bauernfeind, C. R. Adams, W. L. Marusich, in C. Bauernfeind (ed.): *Carotenoids as Colorants and Vitamin A Precursors,* Academic Press, New York 1981, p. 577.
68. P. Karrer, E. Jucker, *Helv. Chim. Acta* **29** (1946) 229.
69. S. Ball, T. W. Goodwin, R. A. Morton, *Biochem. J.* **42** (1948) 516.
70. B. Frank, M. Dust, A. Stange, P. P. Hoppe, *Naturwissenschaften* **69** (1982) 401.
71. F. B. Jungalwala, H. R. Cama, *Biochem. J.* **95** (1965) 19.
72. Eastman Kodak, US 2 839 585, 1958 (O. D. Hawks).
73. Hoffmann-La Roche, DE-OS 2 061 507, 1969 (R. Marbet).
74. H. B. Henbest, E. R. H. Jones, T. C. Owen, *J. Chem. Soc.* 1957, 4909.
75. G. Wald, P. K. Brown, R. Hubbard, W. Oroshnik, *Proc. Natl. Acad. Sci. USA* **41** (1955) 438. P. K. Brown, G. Wald, *J. Biol. Chem.* **222** (1956) 865.
76. F. H. Carr, E. H. Price, *Biochem. J.* **20** (1926) 497.
77. P. Karrer, U. Solmssen, *Helv. Chim. Acta* **20** (1937) 682, 1020.
78. R. Kuhn, H. Brockmann, *Ber. Dtsch. Chem. Ges.* **65** (1932) 894; *Justus Liebigs Ann. Chem.* **516** (1935) 95.
79. P. Karrer, E. Jucker, *Helv. Chim. Acta* **28** (1945) 427, 471.
80. F. J. Petracek, L. Zechmeister, *J. Am. Chem. Soc.* **78** (1956) 1427. R. Entschel, P. Karrer, *Helv. Chim. Acta.* **41** (1958) 402.
81. H. Thommen in O. Isler (ed.): *Carotenoids,* Birkhäuser Verlag, Basel 1971.
82. C. D. Robeson et al., *J. Am. Chem. Soc.* **77** (1955) 4111.
83. S. R. Ames, W. J. Swanson, P. L. Harris, *J. Am. Chem. Soc.* **77** (1955) 4134, 4136. S. R. Ames, *Ann. Rev. Biochem.* **27** (1958) 375. D. W. Stainer, T. K. Murray, *Can. J. Biochem. Physiol.* **38** (1960) 1467. O. R. Braekan, G. Lambertsen, L. R. Njaa, F. Utne, *Int. Z. Vitaminforsch.* **30** (1960) 363. J. Brüggemann, J. Tiews, *Int. Z. Vitaminforsch.* **30** (1960) 279.
84. BASF, US 3 838 029, 1974 (M. Fischer, W. W. Wiersdorf, A. Nürrenbach). Hoffmann-La Roche, US 4 026 778, 1977 (1974) (M. Lalende, H. Stoller). W. H. Waddel, K. Chihara, *J. Am. Chem. Soc.* **103** (1981) 7389. A. H. Levin, F. I. Carrol, *J. Am. Chem. Soc.* **103** (1981) 6527.
85. Hoffmann-La Roche, US 4 051 174, 1977 (H. Stoller, H. P. Wagner).
86. K. H. Bässler, E. Grühn: *Vitamin-Lexikon,* G. Fischer Verlag, Stuttgart 1992, pp. 211 – 233.
87. W. Müller-Mulot, *Fette, Seifen, Anstrichm.* **78** (1976) 18.
88. A. Winterstein, *Angew. Chem.* **72** (1960) 902. H. Thommen, O. Wiss, *Z. Ernährungswiss. Suppl.* **3** (1963) 18.
89. T. W. Goodwin: *The Biochemistry of the Carotenoids,* 2nd ed., **vol. I: Plants,** Chapman and Hall, London 1980.
90. J. C. Bauernfeind, C. R. Adams, W. L. Marusich in J. C. Bauernfeind (ed.): *Carotenoids as Colorants and Vitamin A Precursors,* Academic Press, New York 1981, p. 564. H. S. Huang, D. S. Goodman, *J. Biol. Chem.* **241** (1965) 2839. D. S.

Goodman, H. S. Huang, M. Kanai, T. Shiratori, *J. Biol. Chem.* **242** (1967) 3543. N. H. Fidge, F. R. Smith, D. S. Goodman, *Biochem. J.* **114** (1969) 689. G. Brubacher, U. Gloor, O. Wiss, *Chimia* **14** (1960) 19.
91. J. A. Olson, *Am. J. Clin. Nutr.* **22** (1969) 953.
92. N. H. Fidge, T. Shiratori, J. Ganguly, D. S. Goodman, *J. Lipid Res.* **9** (1968) 103. D. F. Moffa, F. J. Lotspeich, R. F. Krause, *J. Biol. Chem.* **245** (1970) 439.
93. D. S. Goodman et al., *J. Clin. Invest.* **45** (1966) 1615.
94. H. B. Sewell, G. E. Mitchel, C. O. Little, B. W. Hayes, *Int. Z. Vitaminforsch.* **37** (1967) 301. J. A. Boling et al., *J. Nutr.* **99** (1969) 502. M. Kanai, A. Raz, D. S. Goodman, *J. Clin. Invest.* **47** (1968) 2025.
95. G. W. T. Grönendijk, P. A. A. Jansen, S. L. Bonting, F. J. M. Daemen, *Methods Enzymol.* **67** (1980) 203.
96. K. C. D. Hickman, *Ind. Eng. Chem.* **29** (1937) 968. K. C. D. Hickmann, *Chem. Rev.* **34** (1944) 51.
97. H. J. Passino, *Ind. Eng. Chem.* **41** (1949) 280.
98. R. Kuhn, C. J. O. R. Morris, *Ber. Dtsch. Chem. Ges.* **70** (1937) 853. J. F. Arens, D. A. van Dorp, *Recl. Trav. Chim. Pays-Bas* **67** (1948) 973.
99. H. Mayer, O. Isler: *Carotenoids,* Birkhäuser, Basel 1971, p. 373. O. Isler, *Pure Appl. Chem.* **51** (1979) 447.
100. J. C. Bauernfeind: "Aspects of Human and National Nutrition," *World Rev. Nutr. Diet.* **41** (1983) 110. W. Bollag, *Lancet* **1, 8329** (1983) 860.
101. BASF, DE 951 212, 1956 (H. Pommer, G. Wittig). H. Pommer, *Angew. Chem.* **72** (1960) 811.
102. G. W. H. Cheeseman et al., *J. Chem. Soc.* 1949, 2031. J. Attenburrow et al., *J. Chem. Soc.* 1952, 1094.
103. BASF, DE 950 551, 1956 (H. Pommer); DE 1 050 763, 1959 (H. Pommer, W. Sarnecki); DE 1 068 704, 1959 (H. Pommer, W. Sarnecki); DE 1 109 671, 1958 (H. Stilz). H. A. M. Jacobs, H. M. Berg, L. Brandsma, J. F. Arens, *Recl. Trav. Chim. Pays-Bas* **84** (1965) 1113. Organon, NL 6 404 175, 1965 (H. A. M. Jacobs).
104. Glidden, US 3 076 839 (R. L. Webb) 1963; US 3 062 875, 1962 (J. P. Bain); US 3 278 623, 1964 (J. M. Deifer). H. Pauling, D. A. Andrews, N. C. Hindley, *Helv. Chim. Acta* **59** (1976) 1233.
105. O. Isler, R. Rüegg, A. Studer, R. Jürgens, *Hoppe-Seyler's Z. Physiol. Chem.* **295** (1953) 305.
106. W. Kimel et al., *J. Org. Chem.* **23** (1958) 153.
107. G. Saucy, R. Marbet, *Helv. Chim. Acta* **50** (1967) 1158, 2091, 2095.
108. E. E. Royals, *Ind. Eng. Chem.* **38** (1946) 546. W. Hoffmann, *Seifen, Öle, Fette, Wachse* **101** (1975) 89. W. Hoffmann, *Chem. Ztg.* **97** (1973) 23.
109. Sumitomo Chem., US 2 951 853, 1960 (M. Matsui). M. Matsui et al., *J. Vitaminol.* **4** (1958) 178.
110. H. O. Huisman, A. Smit, S. Vromen, L. G. M. Fischer, *Recl. Trav. Chim. Pays-Bas* **71** (1952) 899. A. E. C., FR 1 234 824, 1960 (R. Courbron).
111. Eastman Kodak, US 2 676 990, 1954 (W. J. Humphlett, D. M. Burness); US 2 676 992, 1954 (W. J. Humphlett); US 2 676 994, 1954 (D. M. Burness, C. D. Robeson); US 2 683 746, 1954 (C. H. Beuten, C. D. Robeson); US 2 688 747, 1954 (E. Marx).
112. Philips Gloeilampen-Fabrieken, DE 1 041 950, 1958; NL 6 409 913, 1966 (H. O. Huisman, A. Smit). H. O. Huisman, A. Smit, H. P. von Leeuwen, J. H. van Rij, *Recl. Trav. Chim. Pays-Bas* **75** (1956) 977.
113. A. E. C., FR 1 243 824, 1960 (G. J. M. Nicolaux, E. A. Gray, J. Matet).
114. J. Cymerman, I. M. Heilbron, A. W. Johnson, E. R. H. Jones, *J. Chem. Soc.* 1944, 141.
115. O. Isler, W. Huber, A. Ronco, M. Kofler, *Helv. Chim. Acta* **30** (1947) 1911. O. Isler et al., *Helv. Chim. Acta* **32** (1949) 489. C. v. Planta et al., *Helv. Chim. Acta* **45** (1962) 517, 548.
116. O. Isler et al., *Helv. Chim. Acta* **30** (1947) 1922. I. M. Heilbronn, A. W. Johnson, E. R. H. Jones, A. Spinks, *J. Chem. Soc.* 1942, 727. S. Ishikawa, T. Matsuura, *Chem. Zentralbl.* **108 II** (1937) 3452. N. Milas, *Science (Washington D.C.)* **103** (1946) 581.
117. O. Isler, R. Rüegg, P. Schudel, *Chimia* **15** (1961) 212.
118. H. Pommer, *Angew. Chem.* **89** (1977) 437. W. Reif, H. Grassner, *Chem. Ing. Tech.* **45** (1973) 646. H. Pommer, A. Nürrenbach, *Pure Appl. Chem.* **43** (1975) 527.
119. G. Wittig, G. Geissler, *Justus Liebigs Ann. Chem.* **580** (1953) 44.
120. BASF, DE 943 648, 1956 (G. Wittig, H. Pommer); DE 1 003 730, 1957 (G. Wittig, H. Pommer).
121. BASF, DE 950 552, 1956 (G. Wittig, H. Pommer); DE 1 026 745, 1958 (G. Wittig, H. Pommer).

122. BASF, DE 1 046 046, 1958 (W. Sarnecki, H. Pommer).
123. U. Schöllkopf, *Angew. Chem.* **71** (1959) 260.
124. W. Oroshnik, G. Karmas, A. D. Mebane, *J. Am. Chem. Soc.* **74** (1952) 295. W. Oroshnik, A. D. Mebane, *J. Am. Chem. Soc.* **71** (1949) 2062. R. G. Gould, Jr., A. F. Thompson, Jr., *J. Am. Chem. Soc.* **57** (1935) 340.
125. L. A. Yanovskaya, *Izv. Akad. Nauk SSSR Ser. Khim.* 1960, 1435. Y. Ishikawa, *Bull. Chem. Soc. Jpn.* **37** (1964) 207.
126. BASF, DE 1 060 368, 1959 (H. Haeussler, M. Eucken); US 2 950 321, 1960 (W. Sarnecki, H. Pommer); DE 1 068 710, 1959 (W. Sarnecki, H. Pommer); NL 6 405 660, 1964 (W. Sarnecki, H. Pommer); Hoffmann-La Roche, US 3 932 485, 1976 (J. D. Surmatis).
127. H. Mayer, O. Isler in O. Isler (ed.): *Carotenoids,* Birkhäuser, Basel 1971, p. 394. BASF DE-OS 2 004 675, 1970 (W. Himmele); DE-OS 1 941 632, 1969 (W. Himmele, A. Aquila). G. Eletti-Bianchi, F. Centini, L. Re, *J. Org. Chem.* **41** (1976) 1648. J. H. Babler, M. J. Coghlan, M. Fang, P. Fries, *J. Org. Chem.* **44** (1979) 1716. P. A. Wehrli, B. Schaer, *Synthesis* **1977,** 649. Hoffmann-La Roche, DE 2 621 224, 1976 (P. Fitton, H. Moffet). J. H. Babler, US 4 175 204, 1979. P. A. Wehrli, US 4 219 660, 1980. BASF EP 268 931, 1987 (F. Merger, R. Fischer, H. Hasler). Eastman Kodak, WO 90/11989, 1990 (J. R. Mannier, P. J. Muehlbauer); US 5 120 877, 1992 (A. T. Fanning, H. F. Goss); US 5 138 077, 1992 (J. R. Mannier, P. J. Muehlbauer).
128. T. Onoda, *Petrotech Tokyo* **10** (1987) 810. Y. Tanabe, *Hydrocarbon Process.* **60** (1981) 187. J. Haji, J. Yokotake, T. Yanraguchi, M. Sato, N. Murai, EP 289 725, 1988. A. V. Devekki, M. J. Yakushkin, *Kinet. Katal.* **28** (1987) 1099. P. R. Stapp, US 4 044 041 (1977).
129. BASF, DE 950 552, 1956 (G. Wittig, H. Pommer); DE 1 026 745, 1958 (H. Pommer, G. Wittig, W. Sarnecki).
130. BASF, US 3 838 029, 1974 (M. Fischer, A. Nürrenbach). Hoffmann-La Roche, US 4 026 778, 1977 (M. Lalonde, H. Stoller). W. H. Waddel, K. Chihara, *J. Am. Chem. Soc.* **103** (1981) 7389. A. H. Lewin, F. I. Carrol, *J. Am. Chem. Soc.* **103** (1981) 6527.
131. Hoffmann-La Roche, DE-OS 2 733 231, 1978 (K. Schleich, H. Stoller).
132. J. Paust, W. Reif, H. Schumacher, *Justus Liebigs Ann. Chem.* 1976, 2194.
133. BASF, US 2 831 884, 1958 (H. Pommer, W. Arend). K. Sisido, K. Kondo, H. Nozaki, M. Tuda, Y. Udo, *J. Am. Chem. Soc.* **82** (1960) 2286.
134. W. J. Conradi, C. F. Garbers, P. S. Steyn, *J. Chem. Soc. C* 1964, 594; G. Pattenden, B. C. L. Weedon, *J. Chem. Soc. C* 1968, 1984.
135. E. J. Corey, M. Chaykowsky, *J. Org. Chem.* **28** (1960) 254. G. W. Fenton, C. K. Ingold, *J. Chem. Soc.* 1970, 705.
136. M. Julia, D. Arnould, *Bull. Soc. Chim. Fr.* 1973, 743, 746.
137. P. Chabardes, J. P. Decor, J. Varagnat, *Tetrahedron* **33** (1977) 2799. A. Fischli, H. Mayer, W. Simon, H. J. Stoller, *Helv. Chim. Acta* **59** (1976) 397. G. C. Olson, H. Ch. Cheung, K. D. Morgan, Ch. Neukom, C. Saucy, *J. Org. Chem.* **41** (1976) 3287.
138. Rhône-Poulenc, DE-OS 2 202 689, 1972 (M. Julia).
139. Rhône-Poulenc, DE-OS 2 708 282, 1977 (J. P. Decor).
140. SCM-Corp., US 4 048 220, 1977 (C. G. Cardeus). G. Allmang, F. Grass, J. M. Grosselin, C. Mercier, *J. Mol. Catal.* **66** (1991) L 27 – L 31. E. Facke, C. Mercier, N. Pagnier, B. Despeyroux, P. Pauster, *J. Mol. Catal.* **79** (1993) 117 – 131. C. G. Cardenas, US 4 001 307, 1977; US 4 048 220, 1977.
141. A. Fischli, H. Mayer, *Helv. Chim. Acta* **58** (1975) 1492. Rhône-Poulenc, DE-OS 2 305 267, 1973 (P. Chabardes, M. Julia); DE-OS 2 355 898, 1974 (P. Chabardes, M. Julia, A. Menet); DE-OS 2 708 210, 1977 (J. P. Decor).
142. Rhône-Poulenc, EP 319 407, 1988. J. M. Grosselin, C. Mercier, *Organometallics* **10** (1991) 2126.
143. Rhône-Poulenc, DE 2 708 282, 1976 (J. P. Decor); DE 2 708 304, 1977 (J. P. Decor); EP 430 307, 1989 (J. P. Decor).
144. J. H. Babler, US 4 175 204, 1979. J. H. Babler, W. Buttner, *Tetrahedron Lett.* 1976, no. 4, 239 – 242. Dunlop BP 978 892, 1974; US 3 927 076, 1975 (J. Babler). J. E. Bäckvall, *J. Am. Chem. Soc.* **107** (1985) 3676 –3686.
145. Rhône-Poulenc, EP 544 588, 1993 (P. Duhamel, L. Duhamel).
146. Duphar International, EP 456 287, 1991 (G. J. Lagerweg, C. Bakker); EP 489 454, 1991 (C. Bakker). Kuraray, EP 187 259, 1985 (J. Otera, T. Mandai); EP 348 813, 1989 (S. Suzuki, T. Takigawa); EP 349 991, 1989 (T. Mori, S. Suzuki).

147. J. Otera, H. Misawa, T. Onishi, S. Suzuki, Y. Fujita, *J. Org. Chem.* **51** (1986) 3834. F. Chemla, M. Julia, D. Ugnen, *Bull. Soc. Chim. Fr.* 1993, 130, 200. J. Otera et al., *J. Am. Chem. Soc.* **106** (1984) 3670. J. Otera et al., *Chem. Lett.* 1985, 1883.
148. Kuraray, EP 282 913, 1988 (T. Onishi, S. Suzuki, T. Mori).
149. Kuraray, EP 412 551, 1990 (T. Mori, S. Suzuki, T. Onishi); EP 282 914, 1988 (S. Suzuki, T. Mori, T. Onishi); EP 234 496, 1987 (J. Otera, S. Suzuki, T. Onishi).
150. L'Oreal, EP 292 350, 1988 (G. Solladie, S. Forestier). G. Solladie, A. Girardin, P. Metra, *Tetrahedron Lett.* **29** (1988) 209. G. Solladie, A. Girardin, *Tetrahedron Lett.* **29** (1988) 213; *J. Org. Chem.* **54** (1989) 2620.
151. L. Duhamel, P. Duhamel, J. P. Lecouve, *J. Chem. Res. Synop.* 1986, 34 – 35.
152. L. Duhamel, P. Duhamel, J. P. Lecouve, *Tetrahedron* **43** (1987) 4339; **43** (1987) 4349. L. Duhamel, J. Guillemont, J. M. Poirier, P. Chabardes, *Tetrahedron Lett.* **32** (1991) 4499. L. Duhamel, P. Duhamel, Y. LeGallic, *Tetrahedron Lett.* **34** (1993) 319. Rhône-Poulenc, EP 237 438, 1987 (L. Duhamel, P. Duhamel); EP 242 247, 1987 (L. Duhamel, P. Duhamel).
153. E. Negishi, Z. Owozarczyk, *Tetrahedron Lett.* **32** (1991) 6683. N. Okukado, D. E. Van Horn, W. E. Klima, E. Negishi, *Tetrahedron Lett.* 1978, 1027 – 1030. W. J. Scott, G. T. Crisp, J. K. Stille, *J. Am. Chem. Soc.* **106** (1984) 4630. J. F. Normant, J. F. Alexakis, N. Jabri, *Tetrahedron Lett.* **22** (1981) 959, 3851; **23** (1982) 1589. N. Miyaura, K. Yamada, A. Suzuki, *Tetrahedron Lett.* **30** (1979) 3437. N. Miyaura, K. Yamada, H. Suginome, A. Suzuki, *J. Am. Chem. Soc.* **107** (1985) 972. E. Negishi et al., *J. Am. Chem. Soc.* **109** (1987) 2393. E. Negishi et al., *J. Am. Chem. Soc.* **100** (1978) 2254.
154. Hoffmann-La Roche, DE 855 399, 1952; DE 858 059, 1952; DE 953 073, 1956 (O. Isler, M. Montavon); DE 953 074, 1956 (O. Isler, M. Montavon).
155. H. Pommer, *Angew. Chem.* **72** (1960) 911; **89** (1977) 440.
156. BASF, DE 945 247, 1956 (H. Pommer, G. Wittig); DE 1 068 705, 1959 (H. Pommer, W. Sarnecki).
157. BASF, DE 1 068 709, 1959 (H. Pommer, W. Sarnecki); DE 1 158 505, 1963 (W. Sarnecki, A. Nürrenbach); DE 1 155 126, 1963 (H. Pommer, W. Sarnecki). U. Schwieter et al., *Helv. Chim. Acta* **49** (1966) 369. J. D. Surmatis, J. Gibas, R. Thommen, *J. Org. Chem.* **34** (1969) 3039. BASF, NL 6 909 081, 1969 (W. Sarnecki, A. Nürrenbach).
158. Rhône-Poulenc, DE 2 224 606, 1972 (P. Chabarides, M. Julia).
159. Kuraray, EP 461 653, 1991 (T. Mori, H. Fujii, O. Onishi).
160. The Regents of the University of California, DE-OS 2 515 011, 1975 (J. E. McMurray).
161. H. J. Bestmann, L. Kisielowski, W. Distler, *Angew. Chem.* **88** (1976) 297. A. Nürrenbach et al., *Justus Liebigs Ann. Chem.* **1977,** 1146. BASF, DE 2 505 869, 1976 (B. Schenk, J. Paust).
162. B. C. L. Weedon in O. Isler (ed.): *Carotenoids,* Birkhäuser, Basel 1971, p. 270.
163. Hoffmann-La Roche, US 3 367 985, 1968 (J. D. Sumatis).
164. Hoffmann-La Roche, DE 2 440 747, 1975 (J. D. Sumatis).
165. K. Bernhard, H. Mayer, *Pure & Appl. Chem.* **63** (1991) 35.
166. O. Isler et al., *Helv. Chim. Acta* **39** (1956) 249. O. Isler, R. Rüegg, P. Schudel, *Chimia* **15** (1961) 208. I. G. Farbenind., US 2 165 962, 1939 (M. Müller-Conradi, K. Pieroh).
167. R. Rüegg et al., *Helv. Chim. Acta* **42** (1959) 874. R. Rüegg et al., *Helv. Chim. Acta* **42** (1959) 854. Hoffmann-La Roche DE 1 112 730, 1962 (H. Lindlar, M. Montavon); DE 1 081 454, 1960;DE 1 090 658, 1961 (H. Lindlar, M. Montavon); DE 1 092 466, 1961 (O. Isler, H. Lindlar, M. Montavon); DE 1 088 951, 1961 (W. Guex, R. Rüegg, O. Isler).
168. H. Pommer, A. Nürrenbach, *Pure & Appl. Chem.* **43** (1975) 546. H. Pommer, *Angew. Chem.* **89** (1977) 440. BASF, DE 1 216 862, 1966 (H. Pommer, A. Nürrenbach); DE 1 211 616, 1966 (H. Pommer, A. Nürrenbach); DE 1 008 729, 1957 (H. Pommer, A. Nürrenbach).
169. O. Isler et al., *Helv. Chim. Acta* **42** (1959) 864. Hoffmann-La Roche, DE 1 096 349, 1961 (O. Isler, R. Rüegg); DE 1 197 085, 1963 (V. Schwieter, R. Rüegg).
170. Hoffmann-La Roche, DE 1 247 299, 1968 (V. Schwieter, R. Rüegg).
171. B. C. L. Weedon in O. Isler (ed.): *Carotenoids,* Birkhäuser, Basel 1971, p. 270.
172. J. H. Prystowsky, *J. Biol. Chem.* **256** (1981), 4498 –4503. R. Blomhoff, *Physiol. Rev.* **71** (1991) 951.

173. D. R. Applung, F. Chytil, *Endocrinology (Baltimore)* **108** (1981) 2120–2133.
174. T. G. Ebrey, T. Yoshisawa, *Photochem. Photobiol.* **33** (1981) 417. K. Nakanishi, V. Balogh-Nair, M. Arnaboldi, *J. Am. Chem. Soc.* **102** (1980) 7945. W. Stockenius, R. H. Lozier, R. A. Bogomolni, *Biochim. Biophys. Acta* **505** (1979) 215.
175. A. B. Roberts, M. B. Sporn: "Cellular Biology and Biochemistry of Retinoids," in: *The Retinoids,* Academic Press, Orlando 1984.
176. S. Moriguchi, *Immunology* **56** (1985) 169–177.
177. R. K. Chandia, *Br. Med. J.* **297** (1988) 123–124.
178. N. G. Rao, *Am. J. Clin. Nutr.* **11** (1968) 1306–1309.
179. J. L. Napoli, K. R. Race, *J. Biol. Chem.* **263** (1988) 420. J. A. Olson, *Am. Inst. Nutr.* **105** (1988) no. 3, 243–247.
180. Food and Nutrition Board: "Vitamin-A," in: *The Recommended Dietary Allowances,* National Academy of Sciences, Washington, DC, 1980.
181. C. S. Foote, R. W. Denny, *J. Am. Chem. Soc.* **90** (1968) 1193–1197.
182. G. W. Burton, K. U. Ingold, *Science (Washington, DC)* **224** (1984) 189.
183. A. Vahlquist, *J. Invest. Dermatol.* **79** (1982) 73–74.
184. M. M. Mathews-Roth, *Biochemie* **68** (1968) 111–114.
185. H. K. Biesalski, *Dtsch. Apoth. Ztg.* **132** (1992) 24–27.
186. K. Seelert, *Dtsch. Apoth. Ztg.* **131** (1991) 13–17.
187. A. Bendich, *Clin. Nutr.* **7** (1988) 82–83.
188. D. McLaren in W. H. Sebrell, R. S. Harris (eds.): *The Vitamins,* **vol. I,** Academic Press, New York 1967, p. 267. D. S. Goodman in H. P. DeLuca, J. W. Suttis (eds.): *Fat Soluble Vitamins,* The University of Wisconsin Press, Madison, Milwaukee–London 1969, p. 203. G. Brubacher et al., *J. Nutr.* **6** (1982) 211.
189. K. H. Lotthammer, L. Ahlswede, H. Meyer, *DTW Dtsch. Tierärztl. Wochenschr.* **83** (1976) 353. D. Dressler, *Die Mühle-Mischfuttertechnik* **120** (1983) 192.
190. C. S. P. Stastry, *Indian Drugs* **23** (1986) 242.
191. R. F. Bayfield, E. R. Cole, *Methods Enzymol.* **67 F** (1980) 189; S. Grys, *Methods Enzymol.* **67 F** (1980) 195; L. K. Oliver, *Methods Enzymol.* **67 F** (1980) 199.
192. J. N. Thompson, *Eur. J. Cancer Clin. Oncol.* **19** (1983) 1645.
193. J. Glover, *Methods Enzymol.* **67 F** (1980) 282.
194. B. E. Giuliany, *J. Assoc. Off. Anal. Chem.* **64** (1981) 225.
195. N. Usada, M. Kanneko, M. Kanai, T. Nagata, *Histochemistry* **78** (1983) 487.
196. C. J. Bliss, O. A. Roels in P. György, W. N. Pearson (eds.): *The Vitamins VI,* Academic Press, New York 1967. K. H. Bässler, K. Lang, *Vitamine,* Steinkopff, Darmstadt 1981, p. 9. *Vitamin Compendium,* 2nd ed., Hoffmann-La Roche, Basel 1980.
197. P. D. Bryan, J. L. Honigberg, *J. Liq. Chromatogr.* **14** (1991) 2287. W. A. MacGehan, E. Schoenberger, *J. Chromatogr. Biomed. Appl.* **61** (1987) 65.
198. T. Ogawa, T. Matsumoto, *Anal. Sci.* **4** (1988) 473. S. Yamada, *Anal. Chem.* **61** (1989) 612.
199. A. F. El Waliby, F. El Anwar, *Anal. Chim. Acta* **248** (1991) 583. S. Oetles, Y. Hisil, *Nahrung* **35** (1991) 391. A. Apel, E. Ohst, *GIT Fachz. Lab.* **35** (1991) 218. R. G. Pozo, E. S. Saitna, *J. Food Sci.* **55** (1990) 77. A. P. Sancez, C. P. Gomez, *J. Chromatogr.* **623** (1992) 69. B. L. Lee, S. C. Chua, *J. Chromatogr. Biomed. Appl.* **119** (1992) 41. A. Baillet, L. Corbeau, *J. Chromatogr.* **634** (1993) 251. D. C. Woolard, H. Indyk, *J. Micronutr. Anal.* **2** (1986) 125.
200. D. B. Rodriguez-Amaya, *Food Chem.* **35** (1990) 187–195; *Chromatographie* **28** (1989) 249–252. F. W. Quackenbush, R. L. Smallidge, *J. Assoc. Off. Anal. Chem.* **69** (1986) 767. B. R. Premachandra, *Inst. J. Vitam. Nutr. Res.* **55** (1985) 139.
201. L. Thomas, R. T. McCluskey, J. L. Potter, G. Weissmann, *J. Exp. Med.* **111** (1960) 705. J. A. Lucy, J. T. M. Dingle, *Biochem. J.* **79** (1961) 500. G. Weissmann, L. Thomas, *J. Clin. Invest.* **42** (1963) 661.

General References

202. R. Kumar (ed.): *Vitamin D: Basic and Clinical Aspects,* Nijhoff Publ., Boston 1984.
203. B. E. Miller, A. W. Norman in L. J. Machlin (ed.): *Handbook of Vitamins, Nutritional, Biochemical and Clinical Aspects,* Marcel Dekker, New York 1984, p. 45.
204. N. H. Bell, *J. Clin. Invest.* **76** (1985) 1.
205. T. Suda, C. Miyaura, E. Abe, T. Kuroki, *Bone Miner. Res.* **4** (1986) 1.
206. H. F. DeLuca in M. E. Shils, V. R. Young (eds.): *Modern Nutrition in Health and Disease,* 7th ed., Lea & Febiger, Philadelphia 1988, p. 313.

207. H. Reichel, H. P. Koeffler, A. W. Norman, *N. Engl. J. Med.* **320** (1989) 980.
208. A. W. Norman, *J. Cell. Biochem.* **49** (1992) 1.
209. G. R. Mundy, T. J. Martin (eds.): "Physiology and Pharmacology of Bone," *Handb. Exp. Pharmacol.* **107** (1993).
210. M. F. Holick in M. E. Shils, J. A. Olson, M. Shike (eds.): *Modern Nutrition in Health and Disease,* 8th ed., **vol. 1,** Lea & Febiger, Philadelphia 1994, p. 308.
211. A. W. Norman, K. Schaefer, H. G. Grigoleit, D. v. Herrath (eds.): "Vitamin D, Chemical, Biochemical and Clinical Update," *Proc. Workshop Vitamin D* **6th** (1985).
212. A. W. Norman, K. Schaefer, H. G. Grigoleit, D. v. Herrath (eds.): "Vitamin D, Molecular, Cellular and Clinical Endocrinology," *Proc. Workshop Vitamin D* **7th** (1988).
213. A. W. Norman, R. Bouillon, M. Thomasset (eds.): "Vitamin D, Gene Regulation, Structure-Function Analysis and Clinical Application," *Proc. Workshop Vitamin D* **8th** (1991).

Specific References

214. L. F. Fieser, M. Fieser: *Steroids,* Reinhold Publ. Corp., New York 1959. F. J. Zeelen (ed.): *Medicinal Chemistry of Steroids, Pharmacochemistry Library,* **vol. 15,** Elsevier, Amsterdam 1990, p. 251. G. Habermehl, P. E. Hammann: *Naturstoffchemie,* Springer Verlag, Berlin 1990, p. 53. R. H. Thomson: *The Chemistry of Natural Products,* Blackie Academic & Professional, Glasgow 1993.
215. D. Crowfoot-Hodgkin, B. M. Rimmer, J. D. Dunitz, K. N. Trueblood, *J. Chem. Soc.* 1963, 4945.
216. Trinh-Toan, H. F. DeLuca, L. F. Dahl, *J. Org. Chem.* **41** (1976) 3476.
217. V. Delaroff, P. Rathle, M. Legrand, *Bull. Soc. Chim. Fr.* 1963, 1739.
218. S. R. Wilson, R. Umwalla, W. Cui, in [3.11], p. 78.
219. W. H. Okamura, R. M. Wing in A. W. Norman (ed.): *Vitamin D, Molecular Biology and Clinical Nutrition,* Marcel Dekker, New York-Basel 1980, p. 59.
220. O. Hofer, H. Kählig, W. Reischl in [213] p. 190.
221. IUPAC, *Pure Appl. Chem.* **54** (1982) 1511; **61** (1989) 1783.
222. J. L. Napoli, N. J. Koszewski, R. L. Horst in F. Chytil, D. B. McCormick (eds.): *Methods in Enzymology,* **vol. 123,** Academic Press, London 1986, p. 127.
223. E. Berman, Z. Luz, Y. Mazur, M. Sheves, *J. Org. Chem.* **42** (1977) 3325.
224. W. H. Okamura, M. L. Hammond, *Tetrahedron Lett.* 1976, 4807.
225. T. H. Williams et al. *Helv. Chim. Acta* **63** (1980) 1609.
226. N. J. Koszewski et al. *Anal. Biochem.* **162** (1987) 446.
227. R. Lauber, U. P. Schlunegger, Z. V. I. Zaretski, *Org. Mass Spectrom.* **18** (1983) 315.
228. D. R. Andrews, D. H. R. Barton, R. M. Hesse, M. M. Pechet, *J. Org. Chem.* **51** (1986) 4819.
229. J. W. Blunt, H. F. DeLuca, *Biochemistry* **8** (1969) 671.
230. Hoffmann-La Roche, US 5 182 393, 1993 (G. Yannikouros, P. S. Manchand).
231. L. Vanmaele, P. J. DeClercq, M. Vandewalle, *Tetrahedron* **41** (1985) 141.
232. T. Sato et al. *Chem. Pharm. Bull.* **26** (1978) 2933.
233. H. Takayama, M. Ohmori, S. Yamada, *Tetrahedron Lett.* **21** (1980) 5027.
234. N. Ikekawa et al. *Chem. Pharm. Bull.* **23** (1975) 695.
235. A. Fürst, L. Labler, W. Meier, K.-H. Pfoertner, *Helv. Chim. Acta* **56** (1973) 1708.
236. M. J. Calverley, *Tetrahedron* **43** (1987) 4609.
237. D. V. Cohn, R. Kumarasamy, W. K. Ramp in G. D. Aurbach (ed.): *Vitamins and Hormones,* **vol. 43,** Academic Press, London 1986, p. 283.
238. R. Pardo, M. Santelli, *Bull. Soc. Chim. Fr.* 1985, 98.
239. M. Zaidi et al. in G. D. Aurbach (ed.): *Vitamins and Hormones,* **vol. 46,** Academic Press, London 1991, p. 87.
240. H. Bekenmeier, G. Pfennigsdorf in R. Amon, W. Dirscherl (eds.): *Fermente Hormone Vitamine,* **vol III/1,** "Vitamine," Thieme Verlag, Stuttgart 1974, p. 222.
241. R. L. Boland, *Nutr. Rev.* **44** (1986) 1.
242. T. C. Norman, A. W. Norman, *BioMed. Chem. Lett.* **3** (1993) 1785.
243. R. W. Chesney, *J. Pediatr. (St. Louis)* **116** (1990) 159.
244. G. Behm et al., *Vitamins in Animal Nutrition,* 3rd ed., Arbeitsgemeinschaft für Wirkstoffe in der Tierernährung, Bonn 1991.
245. G. Krampitz, Vitamin D in Animal Nutrition, Hoffmann-LaRoche, Basel 1980.
246. H. F. DeLuca, J. Burmester, H. Darwish, J. Krisinger in C. Hansch (ed.): *Comprehensive Medicinal Chemistry,* **vol. 3,** Pergamon Press, Oxford 1990, p. 1129.

247. J. W. Pike, *Steroids* **49** (1987) 3.
248. Drug Evaluations Subscription, *Metabolic Drugs,* sect. 19, American Medical Assoc., Chicago 1990, chap. 4. A. M. Perault-Staub, J. F. Staub, G. Milhaud in J. N. M. Heersche, J. A. Kanis (eds.) *Bone Miner. Res.* **7** (1990).
249. A. W. Norman et al. *J. Biol. Chem.* **268** (1993) 13811.
250. J. C. Gallagher, D. Goldgar, J. O'Neill in [212] p. 836.
251. B. E. Nordin, A. G. Need, H. A. Morris in [213] p. 787.
252. S. Hurwitz in G. D. Aurbach (ed.): *Vitamins and Hormones,* **vol. 45,** Academic Press, London 1989, p. 173.
253. E. P. Amento, *Steroids* **49** (1987) 55. D. D. Bikle, *Endocr. Rev.* **13** (1992) 765.
254. J. Silver, J. Russell, L. M. Sherwood, *Proc. Natl. Acad. Sci. USA* **82** (1985) 4270.
255. B. S. Chertow et al. *Endocrinology* (*Baltimore*) **113** (1983) 1511.
256. S. Kadawaki, C. Cade, T. Fujita, A. W. Norman in [211] p. 863.
257. L. Binderup, *BioMed. Chem. Lett.* **3** (1993) 1891. J. A. Eisman in J. N. M. Heersche, J. A. Kanis (eds.), *Bone Miner. Res.* **8** (1994) 45.
258. L. Binderup, E. Bramm, *Biochem. Pharmacol.* **37** (1988) 889. K. Kragballe et al., *Lancet* **337** (1991) 193. M. F. Holick et al. in: "Vitamin D, Molecular, Cellular and Clinical Endocrinology," *Proc. Workshop Vitamin D* **7th** (1988) p. 362. M. J. Calverley, C. A. S. Bretting, *BioMed. Chem. Lett.* **3** (1993) p.1841. M. J. Calverley, L. Binderup, *BioMed. Chem. Lett.* **3** (1993) 1845. J. Abe et al. in [212] p. 310. C. Camisa: *Psoriasis,* Blackwell Scientific Publ., Oxford 1994, p. 321.
259. A. W. Norman et al. in [213] p. 146. M. M. Midland, J. Plumet, W. H. Okamura, *BioMed. Chem. Lett* **3** (1993) 1799. W. H. Okamura, J. A. Palenzuela, J. Plumet, M. M. Mudland, *J. Cell. Biochem.* **49** (1992) 10.
260. E. Toromanoff, *Actual. Chim.* 1989, 71.
261. W. G. Dauben et al. *J. Am. Chem. Soc.* **113** (1991) 8367.
262. UCLAF, US 2 693 475, 1954; *Chem. Abstr.* **49** (1954) 14818 c; US 2 707 710, 1955; *Chem. Abstr.* **50** (1955) 6517 a (L. Velluz, G. Amiard). Philips, US 3 157 678, 1964 (M. P. Rappoldt). M. Nepras et al. CS 249 443, 1988; *Chem. Abstr.* **110** (1988) 75889 v.
263. A. E. C. Snoeren, M. R. Daha, J. Lugtenburg, E. Havinga, *Recl. Trav. Chim. Pays-Bas* **89** (1970) 261. S. C. Eyley, D. H. Williams, *J. Chem. Soc. Chem. Commun.* **1975,** 858. Solarchem Research. US 4 686 023, 1987 (R. D. S. Stevens). K. H. Pfoertner, *J. Chem. Soc. Perkin Trans. 2* **1991,** 523. K. H. Pfoertner, M. Voelker, *J. Chem. Soc. Perkin Trans. 2* **1991,** 527.
264. M. R. Uskokovic, J. J. Partridge, T. A. Narwid, E. G. Baggiolini in A. W. Norman (ed.): *Vitamin D, Molecular Biology and Clinical Nutrition,* Marcel Dekker, New York-Basel 1980, p. 1. G. Quinkert (ed.), *Synform* **4** (1986) 131 – 250; **5** (1987) 1 – 17. T. Kametani, H. Furuyama, *Med. Res. Rev.* **7** (1987) 147.
265. A. Mouriño et al. in [213] p. 199. Y. Tachibana, M. Tsuji in Atta-ur-Rahman (ed.): *Studies in Natural Products Chemistry,* **vol. 11,** Elsevier, Amsterdam 1992, p. 379. A. B. Turner, *Nat. Prod. Rep.* **5** (1988) 311; **6** (1989) 539; **8** (1991) 17 (always Section 2.2).
266. D. H. R. Barton, R. H. Hesse, M. M. Pechet, E. Rizzardo, *J. Am. Chem. Soc.* **96** (1973) 2748. D. H. R. Barton, R. H. Hesse, M. M. Pechet, E. Rizzardo, *J. Chem. Soc. Chem. Commun.* 1974, 203.
267. A. Fürst, L. Labler, W. Meier, *Helv. Chim. Acta* **64** (1981) 1870.
268. Hoffmann-La Roche, US 4 379 842, 1983 (A. Fujiwara, C. Miyamoto, T. Okada).
269. R. M. Dodson, A. H. Goldkamp, R. D. Muir, *J. Am. Chem. Soc.* **82** (1960) 4026.
270. W. G. Dauben, T. Brookhart, *J. Am. Chem. Soc.* **103** (1981) 237. A. D. Batcho, D. E. Berger, M. R. Uskokovic, *J. Am. Chem. Soc.* **103** (1981) 1293. A. D. Batcho et al. *Helv. Chim. Acta* **64** (1981) 1682. Hoffmann-La Roche, US 4 310 467, 1982 (A. D. Batcho, D. E. Berger, M. R. Uskokovic); US 4 360 470, 1982 (A. D. Batcho, M. R. Uskokovic); FR 2 475 557, 1981 (A. Fürst, L. Labler, W. Meier). A. Fürst, L. Labler, W. Meier, *Helv. Chim. Acta* **65** (1982) 1499. Kuraray Co., JP 04 13, 695, 1992 (Y. Ando, S. Sakane, S. Nagakawa, M. Shiono);*Chem. Abstr.* **117** (1992) 26921k.
271. Hoffmann-La Roche DE-OS 2 607 322, 1976 (M. R. Uskokovic, T. A. Narvid, J. A. Jacobelli, E. Baggiolini).
272. Hoffmann-La Roche, US 5 225 569, 1993 (G. P. Yiannikouros, P. S. Manchand).
273. Taisho Pharmaceutical Co., EP-A 298 757, 1989 (S. Omura, A. Mikami, J. Sasaki, K. Mizoue); *Chem. Abstr.* **111** (1989) 230660 e; JP-Kokai 02 231 089, 1989 (S. Omura, J.

Sasaki); *Chem. Abstr.* **114** (1991) 120272 j; US 4 892 821, 1990 (S. Omura, J. Sasaki, A. Mikami, K. Mizoue); EP-A 461 921, 1991 (K. Takeda, K. Kimura, K. Okamura, R. Okamoto, J. Sasaki, T. Adachi, S. Omura); *Chem. Abstr.* **116** (1992) 126979 h.
274. B. Lythgoe, *Chem. Soc. Rev.* **9** (1980) 449. A. Mouriño, et al. in [212] p. 34.
275. D. A. Seamark, D. J. H. Trafford, H. L. J. Makin, *J. Steroid Biochem.* **14** (1981) 111. D. D. Bikle: *Assay of Calcium-Regulating Hormones,* Springer Verlag, New York 1983. C. E. Porteous, R. D. Coldwell, D. J. H. Trafford, H. L. J. Makin, *J. Steroid Biochem.* **28** (1987) 785. K. Shimada, N. Kobayashi, *Trends Anal. Chem. (Pers. Ed.)* **10** (1991) 103. G. Jones, D. J. H. Trafford, H. L. J. Makin, B. W. Hollis in J. Cazes (ed.): *Modern Chromatographic Analysis of Vitamins, Chromatographic Science Series,* 2nd ed., **vol. 60,** Marcel Dekker, New York 1992, p. 73.
276. P. P. Nair, *Adv. Lipid Res.* **4** (1966) 227.
277. US Pharmacopeia, US Pharmacopeial Convention, Inc., Rockville 1990, p. 1551. K. Hartke, E. Mutschler (eds.): "DAB9-Kommentar," *Deutsches Arzneibuch,* 9th ed., vol. 4, Govi-Verlag, Frankfurt 1990, pp. 3975, 4038.
278. B. W. Hollis, *Clin. Chem. (Winston-Salem, N.C.)* **32** (1986) 2060. B. W. Hollis, T. Kilbo in [212] p. 710.
279. R. D. Coldwell, C. E. Porteous, D. J. H. Trafford, H. L. J. Makin, *Steroids* **49** (1987) 155.
280. M. Shimizu, S. Kamachi, Y. Nishii, S. Yamada, *Anal. Biochem.* **194** (1991) 77.
281. H. L. J. Makin, R. D. Coldwell, D. J. H. Trafford in [213] p. 637.
282. Swiss Pharmaceutical Society (ed.): *Index Nominum,* Medpharm, Stuttgart 1992. USAN and the USP Dictionary of Drug Names, US Pharmacopeial Convention Inc., Rockville, USA, 1992.

General References

283. O. Isler, G. Brubacher: *Vitamine I,* Georg Thieme Verlag, Stuttgart 1982, p. 126.
284. P. Schudel, H. Mayer, O. Isler, in W. H. Sebrell, R. S. Harris (eds.): *The Vitamins,* vol. V, Academic Press, New York 1972.
285. L. J. Machlin (ed.): *Vitamin E – A Comprehensive Treatise,* Marcel Dekker, New York 1980.
286. W. Friedrich: *Handbuch der Vitamine,* Urban & Schwarzenberg, München 1987, p. 151.
287. M. Mino et al. (eds.): *Vitamin E,* Jpn. Sci. Soc. Press, Tokyo; Karger, Basel 1993.
288. L. Packer, J. Fuchs (eds.): *Vitamin E in Health and Disease,* Marcel Dekker, New York 1993.
289. Vitamin E Research & Information Service: *Vitamin E Fact Book,* La Grange, Illinois 1994.
290. *Kirk-Othmer,* **vol. 21,** p. 574.

Specific References

291. H. M. Evans, K. S. Bishop, *Science (Washington D.C.)* **56** (1922) 650.
292. H. M. Evans, *Proc. Natl. Acad. Sci. USA* **11** (1925) 373.
293. A. M. Pappenheimer, M. Goetsch, *J. Exp. Med.* **53** (1931) 11.
294. M. Goetsch, A. M. Pappenheimer, *J. Exp. Med.* **54** (1931) 145.
295. J. Wadell, H. Steenbock, *J. Biol. Chem.* **80** (1928) 431.
296. H. M. Evans, O. H. Emerson, G. A. Emerson, *J. Biol. Chem.* **113** (1936) 319.
297. O. H. Emerson, G. A. Emerson, A. Mohammad, H. M. Evans, *J. Biol. Chem.* **122** (1937) 99.
298. Distillation Products, US 2 486 541, 1946 (J. G. Baxter et al.).
299. E. Fernholz, *J. Am. Chem. Soc.* **60** (1938) 700.
300. P. Karrer et al., *Helv. Chim. Acta* **21** (1938) 520.
301. J. Green, P. Mamalis, S. Marcinkiewicz, D. McHale, *Chem. Ind. (London)* **73** (1960). D. McHale, J. Green et al., *J. Chem. Soc.* 1963, 784. J. F. Pennock et al., *Biochem. Biophys. Res. Commun.* **17** (1964) 542.
302. P. Schudel, H. Mayer, J. Metzger, R. Rüegg, O. Isler, *Helv. Chim. Acta* **46** (1963) 2517.
303. J. W. Scott, F. T. Bizzarro, D. R. Parrish, G. Saucy, *Helv. Chim. Acta* **59** (1976) 290.
304. C. D. Robeson, *J. Am. Chem. Soc.* **65** (1943) 1660.
305. C. D. Robeson, *J. Am. Chem. Soc.* **64** (1942) 1487.
306. P. Schudel, H. Mayer, J. Metzger, R. Rüegg, O. Isler, *Helv. Chim. Acta* **46** (1963) 333.
307. J. G. Baxter, C. D. Robeson, J. D. Taylor, R. W. Lehman, *J. Am. Chem. Soc.* **65** (1943) 918.
308. H. Mayer, P. Schudel, R. Rüegg, O. Isler, *Helv. Chim. Acta* **46** (1963) 650.
309. P. W. R. Eggitt, F. W. Norris, *J. Sci. Food. Agric.* **6** (1955) 689.

310. J. Green et al., *J. Chem. Soc.* 1959, 3362.
311. H. Rosenkrantz, *J. Biol. Chem.* **173** (1948) 439.
312. H. Finegold, H. T. Slover, *J. Org. Chem.* **32** (1967) 2557.
313. K. Tsukida, M. Ito, F. Ikeda, *J. Vitaminol.* **18** (1972) 24.
314. H. Mayer et al., *Helv. Chim. Acta* **50** (1967) 1168.
315. S. E. Scheppele et al., *Lipids* **7** (1972) 297.
316. R. P. Evstigneva, E. I. Zakharova, *Izv. Vyssh. Uchebn. Zaved. Khim. Khim. Tekhnol.* **34** (1991) 3.
317. J. Kohar, C. Suarna, P. T. Southwell-Keely, *Lipids* **28** (1993) 1015. S. Matsumoto, Y. Iitaka, S.-I. Nakano, M. Matsuo, *J. Chem. Soc., Perkin Trans. 1* 1993, 2727. C. Martius, H. Eilingsfeld, *Justus Liebigs Ann. Chem.* **607** (1957) 159. P. Schudel, H. Mayer, R. Rüegg, O. Isler, *Helv. Chim. Acta* **46** (1963) 637.
318. G. W. Burton, K. U. Ingold, *Acc. Chem. Res.* **19** (1986) 194. M. Foti, K. U. Ingold, J. Lusztyk, *J. Am. Chem. Soc.* **116** (1994) 9440. S.-I. Nagaoka et al., *J. Phys. Chem.* **96** (1992) 2754, 8184. R. H. Bisby, A. W. Parker, *J. Am. Chem. Soc.* **117** (1995) 5664.
319. G. W. Burton, *Proc. Nutr. Soc.* **53** (1994) 251.
320. V. W. Bowry, K. U. Ingold, R. Stocker, *Biochem. J.* **288** (1992) 341. K. U. Ingold, V. W. Bowry, R. Stocker, C. Walling, *Proc. Natl. Acad. Sci. USA* **90** (1993) 45. V. W. Bowry, D. Mohr, J. Cleary, R. Stocker, *J. Biol. Chem.* **270** (1995) 5756.
321. W. A. Pryor et al., *J. Org. Chem.* **58** (1993) 3521. K.-H. Ha, O. Igarashi, *J. Nutr. Sci. Vitaminol.* **36** (1990) 411. G. W. Burton et al., *J. Am. Chem. Soc.* **107** (1985) 7053. G. W. Burton, K. U. Ingold, *J. Am. Chem. Soc.* **103** (1981) 6472. E. Niki et al., *Bull. Chem. Soc. Jpn.* **59** (1986) 497. J. Lehmann, H. T. Slover, *Lipids* **11** (1976) 853.
322. B. Halliwell, J. M. C. Gutteridge: *Free Radicals in Biology and Medicine,* Clarendon Press, Oxford 1987.
323. G. W. Burton, K. U. Ingold: "Vitamin E: Biochemistry and Health Implications," *Ann. N.Y. Acad. Sci.* **570** (1988) 7.
324. H. J. Kayden, M. G. Traber, *J. Lipid Res.* **34** (1993) 343.
325. H. E. Gallo-Torres in L. J. Machlin in [285] p. 193.
326. F. Gey, P. Puska: "Vitamin E: Biochemistry and Health Implications," *Ann. N.Y. Acad. Sci.* **570** (1988) 268.
327. M. K. Horwitt, *Nutr. Rev.* **38** (1980) 105.
328. J. R. Walton, L. Packer in [4.3, p. 495].
329. R. S. Panush, J. C. Delafuente, *World Rev. Nutr. Diet.* **45** (1985) 97.
330. H. J. Kayden, *Neurology* **43** (1993) 2167.
331. A. Emmerie, C. Engel, *Recl. Trav. Chim. Pays-Bas* **57** (1938) 1351.
332. C. C. Tsen, *Anal. Chem.* **33** (1961) 849.
333. G. T. Vatassery, G. A. Martenson, *Clin. Chem. (Winston Salem, N.C.)* **18** (1972) 1475. S. L. Taylor, M. P. Lambden, A. L. Tappel, *Lipids* **11** (1976) 530.
334. W. Boguth, R. Repges, *Int. Z. Vitaminforsch.* **39** (1969) 289. F. Linow, J. Pohl, *Nahrung* **14** (1970) 269.
335. J. D. Desai in [285] p. 68.A. P. De Leenheer, W. E. Lambert et al. (eds.): *Modern Chromatographic Analysis of the Vitamins,* 2nd ed., Marcel Dekker, New York 1992.
336. T.-S. Shin, J. S. Godber, *J. Am. Oil Chem. Soc.* **70** (1993) 1289.
337. W. Schüep, R. Rettenmaier, *Methods Enzymol.* **234** (1994) 294.
338. N. Cohen et al., *Helv. Chim. Acta* **64** (1981) 1158.
339. W. Bremser, F. Vogel, *Org. Magn. Reson.* **14** (1980) 155.
340. M. Vecchi et al., *Helv. Chim. Acta* **73** (1990) 782. G. Riss et al., *Methods Enzymol.* **234** (1994) 302.
341. *National Formulary* **14** (1975) 758.
342. S. R. Ames, *J. Nutr.* **109** (1979) 2198.
343. H. Weiser, M. Vecchi, *Int. J. Vitam. Nutr. Res.* **52** (1982) 351.
344. S. R. Ames, M. J. Ludwig, D. R. Nelan, C. D. Robeson, *Biochemistry* **2** (1963) 188. H. Weiser, M. Vecchi, M. Schlachter, *Int. J. Vitam. Nutr. Res.* **56** (1986) 45.
345. J. Bruggemann, K. H. Niesar, C. Zentz, *Int. Z. Vitaminforsch.* **33** (1963) 180.
346. M. K. Horwitt, *Am. J. Clin. Nutr.* **8** (1960) 451.
347. M. L. Scott, I. D. Desai, *J. Nutr.* **83** (1964) 39.
348. L. J. Machlin, E. Gabriel, M. Brian, *J. Nutr.* **112** (1982) 1437.
349. E.-L. Syväoja et al., *J. Am. Oil Chem. Soc.* **63** (1986) 328.
350. J. Bauernfeind in [285] p. 99.
351. O. Isler, *Experientia* **33** (1977) 555.
352. W. Janiszowska, J. F. Pennock, *Vitam. Horm. (N.Y.)* **34** (1976) 77.
353. G. Thomas, O. R. Threlfall, *Biochem. J.* **142** (1974) 437.

354. H. D'Harlingue, B. Camara, *J. Biol. Chem.* **260** (1985) 15200. O. Dogbo, B. Camara, *Biochim. Biophys. Acta* **920** (1987) 140.
355. P. S. Marshall, S. R. Morris, D. R. Threlfall, *Phytochemistry* **24** (1985) 1705.
356. A. Stocker, A. Rüttimann, W.-D. Woggon, *Helv. Chim. Acta* **76** (1993) 1729.
357. A. Stocker et al., *Helv. Chim. Acta* **77** (1994) 1721.
358. A. R. Welburn, *Phytochemistry* **9** (1970) 743.
359. A. M. Gavin, *J. Am. Oil Chem. Soc.* **58** (1981) 175.
360. T. Rubel: *Vitamin-E-Manufacture,* Noyes Development Corp., Park Ridge, N.J., 1969.
361. Eastman Kodak, US 3 153 054, US 3 153 055, 1962.
362. Hoffmann-La Roche, EP 316 729, 1987 (C. Fizet). Eastman-Kodak, US 3 108 120, 1963.
363. Eastman-Kodak, GB 774 855, 1955. Refining, US 2 704 764, 1955. Nisshin Oil Mills, JP 57 144 281, 1981. Nisshin Flour Milling, US 4 454 329, 1980 (Y. Kai, Y. Takagi). S. K. Kim, J. S. Rhee, *Korean. J. Food. Sci. Technol.* **14** (1982) 174.
364. General Mills, US 3 418 335, 1966.
365. Henkel, US 4 148 810, 1976. A. Struve, R. Schuh, *Fette Seifen Anstrichmittel* **87** (1985) 103.
366. Nisshin Oil Mills, GB 2 145 418, 1983. Eastman-Kodak, US 2 843 610, 1955; US 3 335 154, 1963. Nisshin Oil Mills, JP 055 959, 1974.
367. Yoshihara Seiyu, JP 057 729, 1981. Henkel, EP 171 009, 1984 (S. M. Willing, R. R. Swanson). Distillation Products, US 2 349 275, 1941 (K. C. D. Hickman); US 2 454 692, 1946 (N. D. Embree, N. H. Kuhrt). Showa Denko, JP 60 048 981, 1983.
368. Hoffmann-La Roche, EP 610 742, 1993 (C. Fizet).
369. Eisai, US 3 122 565, 1960 (S. Kijima et al.); US 3 402 182, 1966 (S. Kijima, T. Nakamura). Nisshin Flour Milling, JP 60 109 586, 1983. Bidec Bioind Dev Center, US 5 190 618, 1988 (T. Kawada et al.).
370. General Mills, US 3 819 657, 1972 (N. S. Baldwin et al.); DE 2 351 272, 1972. Eastman-Kodak, US 4 239 691, 1979. Ajinomoto, JP 56 068 677 = JP-AS 86/48838, 1979 (D. R. Neler, C. H. Foster). Henkel, EP 178 400, 1984 (S. M. Willing).
371. Eastman Kodak, US 2 640 058, 1949 (L. Weisler).
372. BASF, EP 338 429, 1989 (P. Lechtken, U. Hörcher, B. Jessel).
373. Distillation Products, US 2 486 539, 1946 (L. Weisler); US 2 486 542, (L. Weisler, A.-J. Chechak).
374. Vitamins Lim., US 2 992 235, 1957 (J. Green, S. Z. Marcinkiewicz).
375. Eastman Kodak, US 2 592 531, US 2 592 628, US 2 592 629, US 2 592 630, 1949 (J. G. Baxter). Eastman Kodak, US 2 843 604, 1955 (O. D. Hawks).
376. Henkel, EP 159 018, 1984 (N. S. Baldwin et al.).
377. Hoffmann-La Roche, US 2 411 967, US 2 411 968, US 2 411 969, 1946 (P. Karrer et al.). Eastman Kodak, US 3 444 213, 1967 (D. R. Nelan); US 3 631 068, 1971. Diamond Shamrock, US 3 708 505, 1971 (S. B. Greenbaum, W. Hacke, H. Horn). Nisshin Flour Milling, DE 2 743 920, 1976 (Y. Yoshino, K. Kondo, S. Kawagoe). Sumitomo, JP 60 054 380, 1983, 1985; *Chem. Abstr.* **103** (1985), 123 731. Eisai, JP 60 072 881, 1983, 1985; *Chem. Abstr.* **103** (1985) 160 390. A. Skoda, B. Proksa, M. Majar, CS 254 402, 1983; *Chem. Abstr.* **111** (1989) 233 286. BASF, EP 100 471, 1982 (J. L. Finnan); DE 3 203 487, 1983 (B. Schulz, E. Renauer).
378. Hoffmann-La Roche, US 2 723 278, 1954 (J. D. Surmatis et al.); DE-OS 1 909 164, 1968 (W. Metlesics, P. A. Wehrli). P. A. Wehrli, R. I. Fryer, W. Metlesics, *J. Org. Chem.* **36** (1971) 2910.
379. Hoffmann-La Roche, US 3 789 086, 1974.
380. Hoffmann-La Roche, EP 012 824, 1978 (P. Fitton, R. Propper).
381. BASF, DE 4 208 477, 1992 (P. Grafen et al.).
382. BASF, DE 4 243 464, 1992 (R. Lowack et al.).
383. Eisai, US 3 459 774, 1964. Eastman Kodak, US 3 538 119, 1970.
384. Rhône-Poulenc, FR 2 602 772, 1986 (A. Beind, F. Sagi).
385. G. S. Chernikov, N. G. Baranova, U. M. Azizov, I. B. Afanas'er, *Khim. Farm. Zh.* **22** (1988) 1244.
386. BASF, DE 1 739 183, 1968 (L. Schuster, H. Pommer); DE 1 668 874, 1968 (A. Arnold, H. Pasedach, H. Pommer). Rhône-Poulenc, DE 2 450 908, 1974 (M. Jouffret).
387. Hoffmann-La Roche, DE 2 221 624, 1972 (W. Brenner); EP 35 635, 1980 (D. Bartoldus, B. Lohri). Kuraray, JP 55 072 136, 1978. Sun Refining, EP 070 665, 1981(C. X. Hsu, J. F. Lyons); EP 107 427, 1982 (C. Y. Hsu, J. F. Lyons). BASF, EP 93 880, 1982 (F. Thoemel,

W. Hoffmann); DE 4 029 198, 1990 (B. Bockstiegel et al.). Mitsubishi Gas, EP 127 888, 1983, EP 167 153, 1986 (T. Ishiki et al.). Eisai, EP 294 584, 1987 (N. Hirose et al.). Jpn. Agency of Ind. Science and Techn., EP 369 823, 1988 (K. Takehira et al.). M. Shimizu et al., *Bull. Chem. Soc. Jpn.* **65** (1992) 1522. CK Fine Chemicals, JP 087 993, 1989. British Petroleum, EP 175 574, 1984 (C. G. Griggs).

388. S. Fujibayashi, K. Nakayama, Y. Nishiyama, Y. Ishii, *Chem. Lett.* **1994,** 1345. O. A. Kholdeeva, A. V. Golovin, R. I. Maksimovskaya, I. V. Kozhevnikov, *J. Mol. Catal.* **75** (1992) 235. H. Lissel, H. Jansen in de Wal, R. Neumann, *Tetrahedron Lett.* **33** (1992) 1795. M. Shimizu, H. Orita, T. Hayakawa, K. Takehira, *Tetrahedron Lett.* **30** (1989) 471. Jpn. Agency Ind. Sciences Techn., EP 338 666, 1988 (M. Shimizu et al.).
389. BASF, DE 1 940 386, 1969 (L. Schuster, R. Oster). Kuraray; DE 2 250 066, 1971 (T. Kawaguchi et al.); DE 2 309 051, 1972 (K. Murakami, T. Yasui). Union Rheinische Braunkohlen Kraftstoff, DE 225 543, 1972 (H. Hoever, E. Biller). BASF, EP 152 842, 1984 (F. J. Broecker et al.). Mitsubishi Gas, EP 198 476, 1985 (T. Yui, A. Ito, T. Takata); EP 264 823, 1986 (T. Yui, A. Ito).
390. Hoffmann-La Roche, DE 2 415 930, 1974. Mitsubishi Gas, EP 419 045, 1989. V. Durgakumari, G. Sreekanth, S. Narayanan, *Res. Chem. Intermed.* **14** (1990) 223. Asahi, JP 06 116 189, 1992 (K. Sakai, H. Oohashi).
391. BASF, DE 1 793 037, 1968 (L. Arnold et al.). Hoffmann-La Roche, US 3 857 892, 1972 (P. A. Wehrli).
392. BASF, DE 1 668 874, 1968 (L. Arnold et al.). P. A. Wehrli, R. I. Fryer, F. Pigott, G. Silverman, *J. Org. Chem.* **37** (1972) 2340.
393. V. A. Bushmelev, T. A. Kondrat'eva, M. A. Lipkin, A. V. Kondrat'ev, *Khim. Farm. Zh.* **25** (1991) 65.
394. Teijin, DE 2 314 600, 1973. Y. Ichikawa et al., *Ind. Eng. Chem. Prod. Res. Dev.* **18** (1979) 373.
395. Rhône-Poulenc, US 4 612 401, 1986 (M. Constantini et al.). C. K. Fine Chemicals, JP 62 192 337, 1986 (D. Yoshida). Mitsubishi Gas, EP 84 158, 1986 (T. Tomita); JP 59 227 835, 1984; US 4 477 682, 1981 (T. Tomita et al.).
396. Teijin, DE 2 345 062, 1973 (Y. Ichikawa et al.). Kuraray, JP 62 263 136, 1986 (T. Kuwimoti, H. Tamai).
397. C. Mercier, P. Cabardes, *Pure Appl. Chem.* **66** (1994) 1509. C. Mercier, P. Chabardes, *Chem. Ind. (Dekker)* **62** (1995) 213.
398. Rhône-Poulenc, DE-OS 2 418 342, 1973 (D. Michelet).
399. Rhône-Poulenc, DE-OS 2 518 028, 1974 (M. Jouffret).
400. Rhône-Poulenc, US 4 247 720, 1975 (M. M. Baudouin, R. M. Perron). Eastman Kodak, DE-OS 2 149 159, 1970 (J. G. Thweatt, F. M. Rash). Y. A. Joe et al., *Bull. Korean Chem. Soc.* **12** (1991) 253.
401. G. Saucy, R. Marbet, *Helv. Chim. Acta* **50** (1967) 2091.
402. BASF, DE 1 068 696, 1955; DE 1 073 476, 1960 (H. Pasedach, M. Seefelder); US 4 310 705, 1979 (A. Nissen et al.). W. Hoffmann, H. Pasedach, H. Pommer, *Justus Liebigs Ann. Chem.* **729** (1969) 52. W. Hoffmann, *Chem. Ztg.* **97** (1973) 23. Hoffmann-La Roche, US 2 628 250, 1951; 2 638 484, 1952; 2 660 608, 1952; 2 795 617, 1957 (W. Kimel). M. F. Carroll, *J. Chem. Soc.* 1940, 704, 1266;1941, 507.
403. G. Saucy, R. Marbet, *Helv. Chim. Acta* **50** (1967) 162; *Angew. Chem.* **72** (1960) 869.
404. K. Itoi in R. Noyori (ed.): *Organic Synthesis in Japan, Past, Present, and Future,* Tokyo Kagaku Dozin, Tokyo 1993, p. 165.
405. Rhône-Poulenc, EP 44 771, 1980 (D. Morel). Rhône-Poulenc, EP 441 708, 1991 (P. Chabardes, C. Mercier). D. Morcel, G. Mignani, Y. Colenille, *Tetrahedron Lett.* **26** (1985) 6337. C. Mercier, G. Mignani, M. Aufrand, G. Allmang, *Tetrahedron Lett.* **32** (1991) 1433.
406. C. Fuganti, P. Grasselli, *J. Chem. Soc., Chem. Commun.* **1982,** 205; **1979,** 995. N. Cohen et al., *J. Org. Chem.* **46** (1981) 2445. N. Cohen et al., *J. Am. Chem. Soc.* **101** (1979) 6710. K. Chan et al., *J. Org. Chem.* **43** (1978) 3435. K. Chan, G. Saucy, *J. Org. Chem.* **42** (1977) 3828. K. Chan et al., *J. Org. Chem.* **41** (1976) 3497. N. Cohen et al., *J. Org. Chem.* **41** (1976) 3505, 3512. H. Mayer et al., *Helv. Chim. Acta* **61** (1978) 837. H. Leuenberger et al., *Helv. Chim. Acta* **62** (1979) 455. M. Schmid, R. Barner, *Helv. Chim. Acta* **62** (1979) 464. R. Zell, *Helv. Chim. Acta* **62** (1979) 474. R. Barner, M. Schmid, *Helv. Chim. Acta* **62** (1979) 2384. J. W. Scott et al., *Helv. Chim. Acta* **59** (1976) 290. L. Crombie et al., *J. Chem. Soc., Chem. Commun.* **1978,** 59. D. L. Coffen et al., *Heterocycles* **39** (1994) 527.

407. BASF, EP 1 217 73, 1984 (H. Ernst et al.); EP 122 426, 1983 (H. P. Gehrken et al.). BASF, EP 76 448, 1983 (H. Ernst et al.).
408. N. Cohen et al., *Helv. Chim. Acta* **64** (1981) 1158.
409. J. Hübscher, R. Barner, *Helv. Chim. Acta* **73** (1990) 1068.
410. K. Sato et al., *J. Org. Chem.* **52** (1987) 5497. S. Takano et al., *Synlett* **1990,** 451. Hoffmann-La Roche, EP 392 389, 1989 (A. Knierzinger, M. Scalone). Hoffmann-La Roche, EP 183 042, 1984 (H. Imfeld).
411. Hoffmann-La Roche, WO 90/01554, 1988 (A. Stokker, W. D. Woggon). Hoffmann-La Roche, EP 531 639, 1991 (F. Grüninger et al.).
412. Hoffmann-La Roche, EP 104 376, 1982 (R. Schmid et al.); EP 257 411, 1987 (Heiser, Stoller). K. Takabe et al., *Tetrahedron Lett.* **26** (1985) 5153. R. Noyori et al., *J. Am. Chem. Soc.* **109** (1987) 1596.
413. H. Dam, *Vitam. Horm. (N.Y.)* **24** (1966) 295 – 306.
414. H. J. Almquist, *Am. J. Clin. Nutr.* **28** (1975) 656 –659.
415. C. T. Esmon, J. A. Sadowski, J. W. Suttie, *J. Biol. Chem.* **250** (1975) 4744 – 4748.
416. IUNS Committee 1/I, Nomenclature, *Nutr. Abstr. Revs.* **48** (1978) 831 – 835.
417. IUPAC-IUB Commission on Biochemical Nomenclature (CBN), *Biochem. J.* **147** (1975) 15 – 21; *Eur. J. Biochem.* **53** (1975) 15 – 18.
418. *U.S. Pharmacopoeia, USP 23, The National Formulary, NF 18,* U.S. Pharmacopeial Convention, Rockville, MD, 1995, p. 1224.
419. H. Mayer, O. Isler in W. H. Sebrell, Jr., R. S. Harris (eds.): *The Vitamins,* 2nd ed., **vol. III,** Academic Press, New York 1971, p. 418.
420. F. Weber, B. Pfister in R. Ammon, W. Dirscherl (eds.): *Fermente, Hormone, Vitamine,* 3rd ed., **vol. III/3,** Thieme Verlag, Stuttgart 1982, p. 183.
421. J. W. Suttie in A. T. Diplock (ed.): *Fat-Soluble Vitamins,* Technomic, Lancaster, PA, 1985, p. 225.
422. P. Sommer, M. Kofler, *Vitam. Horm. (N.Y.)* **24** (1966) 349 – 399.
423. L. M. Jackman et al., *Helv. Chim. Acta* **48** (1965) 1332 – 1349.
424. R. W. MacKee et al., *J. Biol. Chem.* **131** (1939) 327 –344.
425. H. Dam et al., *Helv. Chim. Acta* **22** (1939) 310 – 313.
426. F. Irreverre, M. X. Sullivan, *Science (Washington D.C. 1883)* **94** (1941) 497 – 498.
427. K. Schilling, H. Dam, *Acta Chem. Scand. (1947 –1973)* **12** (1958) 347 – 348.
428. M. J. Shearer, V. Allan, Y. Haroon, P. Barkhan in J. W. Suttie (ed.): *Vitamin K Metabolism and Vitamin K-Dependent Proteins,* University Park Press, Baltimore 1980, p. 317.
429. J. P. Langenberg et al., *Acta Aliment.* **15** (1986) 187 –198.
430. J. W. Suttie, *J. Am. Diet. Assoc.* **92** (1992) 585 – 590.
431. G. Ferland, J. A. Sadowski, *J. Agric. Food Chem.* **40** (1992) 1869 – 1873.
432. G. Ferland, J. A. Sadowski, *J. Agric. Food Chem.* **40** (1992) 1874 – 1877.
433. J. W. Suttie in L. J. Machlin (ed.): *Handbook of Vitamins,* 2nd ed., Marcel Dekker, New York 1991, p. 145.
434. M. J. Shearer, *Blood Revs.* **6** (1992) 92 – 104.
435. F. Fernandez, M. D. Collins, M. J. Hill, *Biochem. Soc. Trans.* **13** (1985) 223 – 224.
436. W. Friedrich: *Vitamins,* De Gruyter, Berlin 1988, p. 287.
437. J. W. Seeger, Jr., R. Bentley, *Phytochemistry* **30** (1991) 3585 – 3589.
438. A. Weische, M. Johanni, E. Leistner, *Arch. Biochem. Biophys.* **256** (1987) 212 – 222.
439. M. Simantiras, E. Leistner, *Z. Naturforsch. C: Biosci.* **46** (1991) 364 – 370.
440. S. Kaiping, J. Soll, G. Schultz, *Phytochemistry* **23** (1984) 89 – 91.
441. S. B. Binskley et al., *J. Am. Chem. Soc.* **61** (1939) 2558 – 2559.
442. H. J. Almquist, A. A. Klose, *J. Am. Chem. Soc.* **61** (1939) 2557 – 2558.
443. L. F. Fieser, *J. Am. Chem. Soc.* **61** (1939) 2559 –2561.
444. Hoffmann-LaRoche CH 545 286, 1940.
445. O. Isler, K. Doebel, *Helv. Chim. Acta* **37** (1954) 225 – 233. F. Hoffmann-LaRoche & Co. AG, CH 320 582, 1953 (H. Lindlar).
446. R. Hirschmann, R. Müller, N. L. Nendler, *J. Am. Chem. Soc.* **76** (1954) 4592 – 4594. Merck & Co., US 2 768 650, 1954.
447. R. T. Arnold, R. Larson, *J. Org. Chem.* **5** (1940) 250 – 252. BASF, DE 09 2 952 709, 1981 (H. Puetter, W. Bewert).
448. C. D. Snyder, H. Rappoport, *J. Am. Chem. Soc.* **96** (1974) 8046 – 8054. S. Araki et al., *Bull. Chem. Soc. Jpn.* **57** (1984) 3523 –3525.
449. Y. Tachibana, *Chem. Lett.* 1977, 901 – 902.
450. B. L. Chenard et al., *J. Org. Chem.* **45** (1980) 378 –384.

451. A. Rüttimann, *Chimia* **40** (1986) 290 – 306.
452. Y. Naruta, *J. Org. Chem.* **45** (1980) 4097 – 4104.
453. S. Inoue et al., *Bull. Chem. Soc. Jpn.* **47** (1974) 3098 –3101.
454. L. S. Liebeskind, S. L. Baysdon, V. Goedken, *Organometallics* **5** (1986) 1086 – 1092.
455. K. H. Dötz et al., *Chem. Ber.* **115** (1982) 1278 – 1285.
456. F. Hoffmann-LaRoche AG, EP 0 132 741, 1984 (A. Rüttimann, G. Büchi). Hoffmann-LaRoche Inc., US 4 603 223, 1984 (A. Rüttimann, G. Büchi).
457. Eisai Chemical Co., EP-A 0 613 877, 1993 (H. Kimio et al.).
458. W. E. Lambert, A. P. De Leenheer in A. P. De Leenheer, W. E. Lambert, H. J. Nelis (eds.): *Modern Chromatographic Analysis of Vitamins,* 2nd ed., Marcel Dekker, New York 1992, p. 197.
459. M. Shino, *Analyst (London)* **113** (1988) 393 – 397.
460. Y. Usui et al., *J. Chromatogr.* **489** (1989) 291 – 301.
461. U. Manz, R. Maurer, *Int. J. Vitam. Nutr. Res.* **52** (1982) 248 – 252.
462. M. F. L. Lefevere, A. E. Claeys, A. P. De Leenheer in A. P. De Leenheer, W. E. Lambert, M. G. M. De Ruyter (eds.): *Modern Chromatographic Analysis of the Vitamins,* Marcel Dekker, New York 1985, p. 201.
463. J. W. Suttie in L. J. Machlin (ed.): *Handbook of Vitamins,* Marcel Dekker, New York 1984, p. 147.
464. H. Weiser, F. Tagwerker in E. Brummerstedt (ed.): *Papers Dedicated to Professor Johannes Moustgaard,* Royal Danish Agricult. Soc., Copenhagen 1981, p. 138.
465. H. Weiser, A. W. Kormann, *Int. J. Vitam. Nutr. Res.* **53** (1983) 143 – 155.
466. B. H. Will, Y. Usui, J. W. Suttie, *J. Nutr.* **122** (1992) 2354 – 2360.
467. H. H. W. Thijssen, M. J. Drittij-Reijnders, *Br. J. Nutr.* **72** (1994) 415 – 425.
468. F. Weber in H. Bickel, Y. Schutz (eds.): "Digestion and Absorption of Nutrients," *Int. J. Vitam. Nutr. Res. Suppl.* **25** (1983) 55 – 65.
469. National Research Council (U.S.): *Recommended Dietary Allowances,* 10th ed., National Academy Press, Washington, D.C., 1989, p. 107.
470. W. Bayer, K. Schmidt: *Vitamine in Prävention und Therapie,* Hippokrates Verlag, Stuttgart 1991, p. 89.
471. C. Vermeer, *Biochem. J.* **266** (1990) 625 – 636.
472. M. J. Shearer, P. T. McCarthy, O. E. Crampton, M. B. Mattock in J. W. Suttie (ed.): *Current Advances in Vitamin K Research,* Elsevier, New York 1988, p. 437.
473. Y. Usui et al., *Am. J. Clin. Nutr.* **51** (1990) 846 –852.
474. S. J. Hodges et al., *Clin. Sci.* **78** (1990) 63 – 66.
475. J. A. Sadowski et al. in J. W. Suttie (ed.): *Current Advances in Vitamin K Research,* Elsevier, New York 1988, p. 453.
476. R. v. Kries et al. in J. W. Suttie (ed.): *Current Advances in Vitamin K Research,* Elsevier, New York 1988, p. 515.
477. A. McBurney, M. J. Shearer, P. Barkhan, *Biochem. Med.* **24** (1980) 250 – 267.
478. M. J. Shearer, A. McBurney, P. Barkhan, *Vitam. Horm. (N.Y.)* **32** (1974) 513 – 542.
479. P. A. Friedman, *N. Engl. J. Med.* **310** (1984) 1458 –1460.
480. D. E. G. Austen in D. L. Williams, V. Marks (eds.): *Biochemistry in Clinical Practice,* 2nd ed., Butterworth-Heinemann, Oxford 1994, p. 495.
481. P. C. Preusch, J. W. Suttie, *J. Org. Chem.* **48** (1983) 2291 – 2293.
482. C. Vermeer, K. Hamulyák, *Thromb. Haemostasis* **66** (1991) 153 – 159.
483. L. Uotila, J. W. Suttie, *Med. Biol.* **60** (1982) 16 – 24.
484. C. Vermeer, *Mol. Cell. Biochem.* **61** (1984) 17 – 35.
485. P. A. Price, J. S. Rice, M. K. Williamson, *Protein Sci.* **3** (1994) 822 – 830.
486. K. Ramotar et al., *J. Infect. Dis.* **150** (1984) 213 –218.
487. F. Weber, O. Wiss in W. H. Sebrell, Jr., R. S. Harris (eds.): *The Vitamins,* 2nd ed., **vol. III,** Academic Press, New York 1971, p. 457.
488. R. Schmid et al., *Helv. Chim. Acta* **73** (1990) 1276 –1299.
489. A. Hanck, H. Weiser in A. Hanck (ed.): "Vitamins in Medicine," *Int. J. Vitam. Nutr. Res. Suppl.* **24** (1983) 155 – 170.
490. H. C. Buitenhuis, B. A. M. Soute, C. Vermeer, *Biochim. Biophys. Acta* **1034** (1990) 170 – 175.
491. T. H. Tulchinsky et al., *Am. J. Public Health* **83** (1993) 1166 – 1168.
492. R. von Kries, Y. Hanawa, *Thromb. Haemostasis* **69** (1993) 293 – 295.
493. J. P. Hart et al., *J. Clin. Endocrinol. Metab.* **60** (1985) 1268 – 1269.

494. N. C. Binkley, J. W. Suttie, *J. Nutr.* **125** (1995) 1812 – 1821.
495. D. E. Ullrey in E. R. Miller, D. E. Ullrey, A. J. Lewis (eds.): *Swine Nutrition,* Butterworth-Heinemann, Boston 1991, p. 215.
496. P. P. Hoppe, *Pig Int.* **18** (1988) 16, 18.
497. J. W. Suttie et al., *Am. J. Clin. Nutr.* **47** (1988) 475 – 480.
498. R. von Kries, F. R. Greer, J. W. Suttie, *J. Pediatr. Gastroenterol. Nutr.* **16** (1993) 231 – 238.
499. J. J. Lipsky, *Mayo Clin. Proc.* **69** (1994) 462 – 466.
500. R. E. Olson, *Annu. Rev. Nutr.* **4** (1984) 281 – 337.
501. J. W. Suttie, C. G. Kindberg, J. L. Greger, N. U. Bang in J. W. Suttie (ed.): *Current Advances in Vitamin K Research,* Elsevier, New York 1988, p. 465.
502. D. D. Hall, G. L. Cromwell, T. S. Stahly, *J. Anim. Sci.* **69** (1991) 646 – 655.
503. G. W. Oduho, T. K. Chung, D. H. Baker, *J. Nutr.* **123** (1993) 737 – 743.
504. National Research Council (U.S.): *Vitamin Tolerance of Animals,* National Academy Press, Washington, D.C., 1987, p. 31.

General References

505. W. H. Sebrell, Jr., R. S. Harris (eds.): *The Vitamins,* 2nd ed., **vol. V,** Academic Press, New York 1972, p. 98. H. Z. Sable, C. J. Gubler (eds.), *Ann. N.Y. Acad. Sci.* **378** (1982) 1. C. J. Gubler in L. J. Machlin (ed.): *Handbook of Vitamins,* 2nd ed., Marcel Dekker, New York 1991.

Specific References

506. N. Shimozono, E. Katsura (eds.), *Rev. Jpn. Lit. Beriberi Thiamine* 1961.
507. C. Eijkman, *Geneeskd. Tijdschr. Ned. Indie* **36** (1896) 214.
508. G. Grijn, *Geneeskd. Tijdschr. Ned. Indie* **41** (1901) 3.
509. C. Funk, *J. Physiol. (London)* **43** (1911) 395.
510. B. C. P. Jansen, W. F. Donath, *Proc. K. Ned. Akad. Wet.* **29** (1926) 1390.
511. P. Karlson, *Trends Biochem. Sci.* 1984, 536.
512. R. R. Williams, R. E. Waterman, J. C. Keresztesky, E. R. Buchman, *J. Am. Chem. Soc.* **57** (1935) 536.
513. R. R. Williams, J. K. Cline, *J. Am. Chem. Soc.* **58** (1936) 1063.
514. R. Grewe, *Hoppe Seyler's Z. Physiol. Chem.* **242** (1936) 89.
515. R. R. Williams, J. K. Cline, *J. Am. Chem. Soc.* **58** (1936) 1504.
516. A. R. Todd, F. Bergel, *J. Chem. Soc.* 1937, 364.
517. A. R. Todd, F. Bergel, *J. Chem. Soc.* 1938, 26.
518. H. Andersag, K. Westphal, *Ber. Dtsch. Chem. Ges. A* **70** (1937) 2035.
519. G. D. Maier, D. E. Metzler, *J. Am. Chem. Soc.* **79** (1957) 4386.
520. J. M. El Hage Chahine, J.-E. Dubois, *J. Am. Chem. Soc.* **105** (1983) 2335.
521. H. Lenormant, C. de Lozé, *Bull. Soc. Chim. Fr.* 1954, 375.
522. H. Z. Sable, J. E. Biaglow, *Proc. Natl. Acad. Sci. USA* **54** (1965) 808.
523. R. E. Echols, G. C. Levy, *J. Org. Chem.* **39** (1973) 1321.
524. A. H. Cain, G. R. Sullivan, J. D. Roberts, *J. Am. Chem. Soc.* **99** (1977) 6423.
525. A. Lycka, J. Cizmarik, *Pharmazie* **45** (1990) 371.
526. M. Barber, R. S. Bordoli, R. D. Sedgwick, A. N. Tyler, *J. Chem. Soc. Chem. Commun.* 1981, 325.
527. K. L. Rinehart, Jr., *Science (Washington, D.C. 1883)* **218** (1982) 254.
528. I. Hormann, *Helv. Chim. Acta* **67** (1984) 1478.
529. K. Kraut, H. J. Reed, *Acta Crystallogr.* **15** (1962) 747.
530. B. Pullman, C. Spanjaad, *Biochim. Biophys. Acta* **46** (1961) 576.
531. J. A. Zoltewicz, G. Uray, *Bioorg. Chem.* **22** (1994) 1.
532. J. Fournier, *Bull. Soc. Chim. Fr.* 1988, 854.
533. J. J. Windheuser, T. Higuchi, *J. Pharm. Sci.* **51** (1962) 354.
534. H. Sugimoto et al., *J. Org. Chem.* **55** (1990) 467.
535. H. Sugimoto, K. Hirai, *Heterocycles* **27** (1988) 877.
536. R. Breslow, *J. Am. Chem. Soc.* **80** (1958) 3719.
537. J. Herrmann, W. Knoche, R. Neugebauer, *J. Chem. Soc. Perkin Trans. 2* 1995, 463.
538. M. B. Doughty, D. S. Lawrence, *J. Chem. Soc. Chem. Commun.* 1985, 454.
539. M. B. Doughty, G. E. Risinger, *J. Chem. Soc. Chem. Commun.* 1985, 1551.
540. H. Sugimoto, K. Hirai, *Tetrahedron Lett.* **26** (1985) no. 7, 883.
541. R. Kluger, *Chem. Rev.* **87** (1987) 863.
542. J.-M. El Hage Chahine, J.-E. Dubois, *J. Chem. Soc. Perkin Trans. 2* 1988, 1409.

543. J.-M. El Hage Chahine, J.-E. Dubois, *J. Chem. Soc. Perkin Trans. 2* 1989, 25.
544. R. F. W. Hopmann, G. P. Brugnoni, B. Fol, *J. Am. Chem. Soc.* **104** (1982) 1341.
545. P. Karrer, H. Krishna, *Helv. Chim. Acta* **33** (1950) 555.
546. G. E. Bonovicino, D. J. Hennessy, *J. Am. Chem. Soc.* **79** (1957) 6325.
547. B. T. Burton: *The Heinz Handbook of Nutrition,* McGraw Hill, New York 1965.
548. M. H. Dong, E. L. MacGown, B. W. Schwenneker, H. E. Sauberlich, *J. Am. Diet. Assoc.* **76** (1980) 156.
549. B. K. Watt, A. L. Merril: *Composition of Foods,* Handbook no. 8, U.S. Department of Agriculture, Washington, D.C. 1963.
550. D. W. Young, *Nat. Prod. Rep.* **3** (1986) 395.
551. G. M. Brown, J. M. Williamson in F. C. Neidhardt (ed.): *Escherichia coli and Salmonella typhimurium,* American Society for Microbiology, Washington, D.C. 1987, p. 521.
552. T. Yura in: *Genetic Studies with Bacteria,* Carnegie Inst. Washington Publ., Washington, D.C. 1956, pp. 63 – 75.
553. P. C. Newell, R. G. Tucker, *Nature (London)* **215** (1967) 1384.
554. P. C. Newell, R. G. Tucker, *Biochem. J.* **106** (1968) 279.
555. P. C. Newell, R. G. Tucker, *Biochem. J.* **106** (1968) 271.
556. D. M. Downs, J. R. Roth, *J. Bacteriol.* **173** (1991) 6597.
557. D. M. Downs, *J. Bacteriol.* **174** (1992) 1515.
558. D. M. Downs, L. Petersen, *J. Bacteriol.* **176** (1994) 4858.
559. S. David, B. Estramareix, H. Hirshfeld, *Biochim. Biophys. Acta* **148** (1967) 11.
560. K. Tazuya, K. Yamada, H. Kumaoka, *Biochim. Biophys. Acta* **990** (1989) 73.
561. K. Tazuya, M. Morisaki, K. Yamada, H. Kumaoka, *Biochem. Int.* **16** (1988) 955.
562. G. Grue-Sorensen, R. L. White, I. D. Spenser, *J. Am. Chem. Soc.* **108** (1986) 146.
563. K. Tazuya et al., *Biochem. Int.* **12** (1986) 661.
564. K. Tazuya, K. Yamada, H. Kumaoka, *Biochem. Int.* **30** (1993) 893.
565. K. Tazuya, K. Yamada, K. Nakamura, H. Kumaoka, *Biochim. Biophys. Acta* **924** (1987) 210.
566. P. E. Linnett, J. Walker, *J. Chem. Soc. C*1967, no. 9, 796.
567. P. E. Linnett, J. Walker, *Biochem. J.* **109** (1968) 161.
568. K. Yamada et al., *Biochem. Int.* **10** (1985) 689.
569. R. L. White, I. D. Spenser, *J. Am. Chem. Soc.* **101** (1979) 5102.
570. R. L. White, I. D. Spenser, *J. Am. Chem. Soc.* **104** (1982) 4934.
571. R. H. White, F. B. Rudolph, *Biochim. Biophys. Acta* **542** (1978) 340.
572. E. Bellion, D. H. Kirkley, J. R. Faust, *Biochim. Biophys. Acta* **437** (1976) 229.
573. B. Estramareix, M. Therisod, *Biochim. Biophys. Acta* **273** (1972) 275.
574. R. H. White, *Biochemistry* **17** (1978) 3833.
575. K. Tazuya et al., *Biochem. Int.* **14** (1987) 153.
576. L. M. Lewin, G. M. Brown, *Fed. Proc.* **20** (1961) 447.
577. G. W. Camiener, G. M. Brown, *J. Biol. Chem.* **235** (1960) 2411.
578. A. Iwashima, K. Nosaka, H. Nishimura, Y. Kimura, *J. Gen. Microbiol.* **132** (1986) 1542.
579. I. G. Leder, *Methods Enzymol.* **18A** (1970) 207.
580. Y. Kayama, T. Kayasaki, *Arch. Biochem. Biophys.* **158** (1973) 242.
581. Y. Kawasaki, *J. Bacteriol.* **175** (1993) 5153.
582. K. Nosaka, Y. Kaneko, H. Nishimura, A. Iwashima, *J. Biol. Chem.* **268** (1993) 17440.
583. T. Kawasaki, T. Nakata, Y. Nose, *J. Bacteriol.* **95** (1968) 1483.
584. T. Kawasaki, Y. Nose, *J. Biochem. (Tokyo)* **65** (1969) 417.
585. P. B. Vanderhorn, A. D. Backstrom, V. Stewart, T. P. Begley, *J. Bacteriol.* **175** (1993) 982.
586. T. Nohno, Y. Kasai, T. Saito, *J. Bacteriol.* **170** (1988) 4097.
587. J. L. Johnson, M. M. Wuebbens, K. V. Rajagopalan, *J. Biol. Chem.* **264** (1989) 13440.
588. M. E. Johnson, K. V. Rajagopalan, *J. Bacteriol.* **169** (1987) 117.
589. N. Imamura, H. Nakayama, *Experientia* **37** (1981) 1265.
590. N. Imamura, H. Nakayama, *J. Bacteriol.* **151** (1982) 708.
591. J. Ryals, R. Y. Hsu, M. N. Lipsett, H. Bremer, *J. Bacteriol.* **151** (1982) 899.
592. T. M. Finan, B. Kunkel, V. G. F. De, E. R. Signer, *J. Bacteriol.* **167** (1986) 66.
593. W. Walter, A. Bacher, *J. Gen. Microbiol.* **103** (1977) 359.
594. C. Anagnostopoulos, P. J. Piggot, J. A. Hoch in A. L. Sonenshein, J. A. Hoch, R. Losick (eds.): *Bacillus subtilis,* American Society for Microbiology, Washington, D.C. 1993, p. 425.

595. P. Glaser et al., *Mol. Microbiol.* **10** (1993) 371.
596. A. M. Schweingruber, J. Dlugonski, E. Edenharter, M. E. Schweingruber, *Curr. Genet.* **19** (1991) 249.
597. K. Maundrell, *J. Biol. Chem.* **265** (1990) 10857.
598. A. Zurlinden, M. E. Schweingruber, *Gene* **117** (1992) 141.
599. A. G. O. Manetti, M. Rosetto, K. G. Maundrell, *Yeast* **10** (1994) 1075.
600. A. Zurlinden, M. E. Schweingruber, *J. Bacteriol.* **176** (1994) 6631.
601. P. C. Newell, R. G. Tucker, *Biochem. J.* **100** (1966) 512.
602. P. C. Newell, R. G. Tucker, *Biochem. J.* **100** (1966) 517.
603. T. Kawasaki, A. Iwashima, Y. Nose, *J. Biol. Chem.* **65** (1969) 407.
604. T. Kawasaki et al., *J. Biochem. (Tokyo)* **79** (1976) 1035.
605. H. Nishimura et al., *J. Bacteriol.* **174** (1992) 4701.
606. H. Nishimura et al., *FEBS Lett.* **297** (1992) 155.
607. Y. Kawasaki et al., *J. Bacteriol.* **172** (1990) 6145.
608. A. M. Schweingruber et al., *Genetics* **130** (1992) 445.
609. A. F. Diorio, L. M. Lewin, *J. Biol. Chem.* **243** (1968) 3999.
610. A. F. Diorio, L. M. Lewin, *J. Biol. Chem.* **243** (1968) 4006.
611. Y. Haj-Ahmad, C. A. Bilinsky, I. Russell, G. G. Stewart, *Can. J. Microbiol.* **38** (1992) 1156.
612. J. Fragner (ed.): *Vitamine: Chemie und Biochemie,* **vol. II,** Fischer Verlag, Jena 1965, p. 1445.
613. *Kirk-Othmer,* 2nd ed., **20,** 173.
614. *Kirk-Othmer,* 3rd ed., **24,** 124.
615. D. J. Brown: *The Pyrimidines,* Suppl. II, J. Wiley, New York 1985.
616. O. Isler, G. Brubacher, S. Ghisla, B. Kräuler: *Vitamine II,* Thieme Verlag, Stuttgart 1988, p. 17.
617. Asahi Kasei Kogyo Kabushiki Kaisha, EP-A 0 053 519 A1, 2, 1981 (T. Katsumata, T. Dozono, M. Honda).
618. A. Takamizawa, K. Tokuyama, K. Tori, *Bull. Soc. Chim. Jpn.* **32** (1959) no. 2, 188.
619. A. Takamizawa, K. Hirai, *Chem. Pharm. Bull.* **12** (1964) no. 4, 393.
620. BASF, US 4 226 799 A1, 1978 (W. Bewert, W. Littmann).
621. BASF, DE-OS 3 511 273 A1, 1985 (H. Ernst, J. Paust).
622. BASF, DE-OS 3 520 982 A1, 1985 (W. Bewert, H. Kiefer).
623. BASF, DE-OS 3 431 270 A1, 1984 (H. Ernst, W. Littmann).
624. K. Nishihira, M. Nakai, Y. Chiba, *Kagaku Kogaku* **55** (1991) 433.
625. Ube Industries, EP-A 55 108 A1, 1982 (K. Matsui, S. Uchimumi, A. Iwayama, T. Umeza).
626. Hüls Troisdorf, DE-OS 3 641 605 A1, 1988 (F. A. Von Itter, W. Vogt).
627. Dynamit Nobel, DE-OS 3 211 679 A1, 1983 (H. Peeters, W. Vogt, C. Theis).
628. S. Torii, T. Inokuchi, M. Kubota, *J. Org. Chem.* **50** (1985) 4157.
629. Ube Industries, EP-A 0 279 556 A1, 1988 (K. Nishihira, S. Fujikawa, M. Yamashita).
630. Ube Industries, DE-OS 3 303 789 A1, 1983 (K. Fujii, K. Nishihira, H. Sawada, S. Tanaka, M. Nakai, H. Yoshida, T. Inoue, K. Oomori).
631. Ube Industries, DE-OS 3 303 815 A1, 1983 (K. Fujii, K. Nishihira, H. Sawada, S. Tanaka, M. Nakai, H. Yoshida, T. Inoue).
632. A. Edenhofer, H. Spiegelberg, H. Oberhänsli, *Helv. Chim. Acta* **58** (1975) 1230.
633. T. Matsukawa, M. Ohta, T. Matsuno, *Yakugaku Zasshi* **63** (1943) 216.
634. D. Szlompek-Nesteruk, S. Paszkowski, PL 125 556, 1979.
635. BASF, EP-A 0 234 586 A2, 1987 (H. Eckhardt, K. Halbritter, F.-K. Pape).
636. T. Matsukawa, T. Iwatsu, *Yakugaku Zasshi* **69** (1949) 550.
637. T. Matsukawa, T. Iwatsu, *Yakugaku Zasshi* **70** (1950) 28.
638. T. Matsukawa, T. Iwatsu, *Yakugaku Zasshi* **71** (1951) 720.
639. T. Matsukawa, T. Iwatsu, *Yakugaku Zasshi* **71** (1951) 1215.
640. Takeda, JP 7 474 588, 1972 (T. Maruyama, M. Tsurushima, Y. Ando).
641. H. Hirano, Y. Yonemoto, *Yakugaku Zasshi* **76** (1956) 230.
642. H. Hirano, T. Iwatsu, S. Yurugi, *Yakagaku Zasshi* **76** (1956) 1332.
643. Hoffmann-La Roche, EP 0 342 482 B1, 1989 (U. Hengartner, G. Moine).
644. P. Contant, L. Forzy, U. Hengartner, G. Moine, *Helv. Chim. Acta* **73** (1990) 1300.
645. S. K. Chatterji, N. J. Anand, *J. Sci. Ind. Res. Sect. B* **18** (1959) 278.
646. Shionogi & Co., DE-OS 2 040 554, 1970 (K. Ikawa, F. Takami).

647. R. F. Evans, A. V. Robertson, *Aust. J. Chem.* **26** (1973) 1599.
648. M. Fujiwara, *J. Nutr. Sci. Vitaminol.* **22** (1976) Suppl., 57.
649. H. Baker, O. Frank, *J. Nutr. Sci. Vitaminol.* **22** (1976) Suppl., 63; R. Bitsch, I. Bitsch, *Dtsch. Apoth. Ztg.* **129** (1989) no. 2, 65.
650. T. Matzukawa, H. Hirano, S. Yurugi, *Methods Enzymol.* **18A** (1970) 146.
651. H. C. Koppel, B. H. Springer, R. K. Robins, C. C. Chang, *J. Org. Chem.* **27** (1962) 3614.
652. H. Neef, K. D. Kohnert, A. Schellenberger, *J. Prakt. Chem.* **315** (1973) 701.
653. S. D. Verma, A. N. Day, *J. Indian Chem. Soc.* **40** (1963) 283.
654. A. Fumagelli, G. Checchi, P. Pasotti, *Farmaco Ed. Sci.* **26** (1971) 736.
655. A. Schellenberger, K. Wendler, P. Creuzburg, G. Hübner, *Hoppe Seyler's Z. Physiol. Chem.* **348** (1967) 501.
656. R. Golbik et al., *Bioorg. Chem.* **19** (1991) 10.
657. J. Biggs, P. Sykes, *J. Chem. Soc.* 1959, 1849.
658. A. Takamizawa, K. Hirai, T. Ishiba, *Chem. Pharm. Bull.* **19** (1971) 2222, 2229.
659. P. Sykes, A. R. Todd, *J. Chem. Soc.* 1951, 534.
660. P. Sykes, P. Neslitt, *J. Chem. Soc.* 1954, 4585.
661. N. Ninomiya, T. Suzuoki, M. Kurihara, *Bitamin* **6** (1953) 429.
662. A. Schellenberger et al., *Hoppe Seyler's Z. Physiol. Chem.* **346** (1966) 123.
663. K. Masuda, *Yakugaku Zasshi* **77** (1957) 153.
664. R. F. Raffauf, *Helv. Chim. Acta* **33** (1950) 102.
665. A. N. Wilson, S. A. Harris, *J. Am. Chem. Soc.* **73** (1951) 2388.
666. BASF, EP-A 0 358 081 A2, 1989 (J. Paust, H. Ernst).
667. E. F. Rogers, *Ann. N.Y. Acad. Sci.* **98** (1962) 412.
668. W. H. Ott, A. M. Dickinson, A. van Iderstine, *Poult. Sci.* **44** (1965) 920.
669. L. O. Krampitz in: *Thiamin Diphosphate and its Catalytic Functions,* Marcel Dekker, New York 1970.
670. M. Vilkas: *Vitamines Mécanismes d'Action Chimique,* Ed. Hermann, Paris 1994, p. 25.
671. A. Schellenberger, *Chem. Ber.* **123** (1990) 1489.
672. E. J. Crane III, J. A. Vaccaro, M. W. Washabaugh, *J. Am. Chem. Soc.* **115** (1993) 8912.
673. G. E. Risinger, W. E. Gore, K. Pulver, *Synthesis* 1974, 659.
674. R. Kluger, J. Chin, T. Smyth, *J. Am. Chem. Soc.* **103** (1981) 884.
675. R. Kluger, T. Smyth, *J. Am. Chem. Soc.* **103** (1981) 1214.
676. A. Plaitakis et al., *Ann. N.Y. Acad. Sci.* **378** (1982) 367.
677. R. G. Mair, C. D. Anderson, P. J. Langlais, W. J. MacEntee, *Brain Res.* **360** (1985) 273.
678. G. E. Gibson, L. Barclay, J. Blass, *Ann. N.Y. Acad. Sci.* **378** (1982) 382.
679. K. A. M. Al-Rashood, F. J. Al-Shammary, N. A. Aziz Mian, *Anal. Profiles Drug Subst.* **18** (1989) 414.
680. C. J. Blake: *The Determination of Vitamins B1, B2, B6 and B12. A Literature Review,* part 1; Microbial Assays; part 2: Chemical Methods, Leatherhead, U.K. 1977.
681. K. Takashi, *Chromatogr. Sci. Ser.* **60** (1992) 319.

General References

682. R. I. Williams, R. E. Eakin, E. Beerstecher, W. Shive: *The Biochemistry of B-Vitamins,* Reinhold Publ., New York 1950, p. 26.
683. J. Fragner: *Vitamine, Chemie u. Biochemie,* G. Fischer, Jena 1965, p. 1406.
684. F. Müller (ed.): *Chemistry and Biochemistry of Flavoenzymes,* vol. I, CRC Press, Boca Raton 1991.
685. P. D. Boyer (ed.): *Reviews in the Enzymes,* 3rd ed., vol. XI – XIII, Academic Press, New York 1975.
686. L. J. Machlin (ed.): *Handbook of Vitamins, Nutritional, Biochemical and Clinical Aspects,* Marcel Dekker, New York 1984.
687. S. G. Flavell Matto: *Vitamins in Medicine,* **vol. 1,** W. Heinemann Verlag, London 1980, p. 426.
688. R. Strohecker, H. M. Henning: *Vitaminbestimmungen,* Verlag Chemie, Weinheim 1963, pp. 101 ff.
689. H. Baker, O. Frank in R. S. Rivlin (ed.): *Riboflavin,* Plenum Press, New York 1975, pp. 49 ff.
690. W. N. Pearson: "Riboflavin," in P. Gyorgy, W. N. Pearson (eds.): *The Vitamins,* vol. 7, Academic Press, New York 1967, pp. 99 ff.

Specific References

691. The Commission of the Nomenclature of Biological Chemistry (IUPAC), *J. Am. Chem. Soc.* **82** (1960) 5581.
692. T. B. Osborne, L. B. Mendel, *J. Biol. Chem.* **15** (1913) 311.
693. A. D. Emmet, L. H. McKin, *J. Biol. Chem.* **32** (1917) 409.

694. O. Warburg, W. Christian, *Biochem. Z.* **254** (1932) 438.
695. O. Warburg, W. Christian, *Biochem. Z.* **266** (1933) 377.
696. R. Kuhn, P. György, T. Wagner-Jauregg, *Ber. Dtsch. Chem. Ges. A* **66** (1933) 576.
697. R. Kuhn, H. Rudy, T. Wagner-Jauregg, *Ber. Dtsch. Chem. Ges. A* **66** (1933) 1950.
698. R. Kuhn, H. Rudy, *Ber. Dtsch. Chem. Ges. A* **67** (1934) 1125, 1298.
699. P. Karrer, K. Schöpp, F. Benz, *Helv. Chim. Acta* **18** (1935) 426.
700. P. Karrer, K. Schöpp, K. Pfaehler, *Ber. Dtsch. Chem. Ges. A* **68** (1935) 216.
701. R. Kuhn, *Ber. Dtsch. Chem. Ges. A* **68** (1935) 1765.
702. D. V. Frost, *J. Am. Chem. Soc.* **69** (1947) 1064.
703. D. E. Guttmann, M. Y. Athalye, *J. Am. Pharm. Assoc. Sci. Ed.* **49** (1960) 687.
704. A. Ehrenberg, *Vitam. Horm. (N.Y.)* **28** (1970) 489 – 504.
705. G. Richter et al., *J. Bacteriol.* **174** (1992) no. 12, 4050 – 4056.
706. G. Richter et al., *J. Bacteriol.* **175** (1993) no. 13, 4045 – 4051.
707. R. Volk, A. Bacher, *J. Biol. Chem.* **266** (1991) no. 31, 20610 – 20618.
708. K. Schott, R. Ladenstein, A. König, A. Bacher, *J. Biol. Chem.* **265** (1990) no. 21, 12686 – 12689.
709. K. Schott, J. Kellermann, F. Lottspeich, A. Bacher, *J. Biol. Chem.* **265** (1990) no. 8, 4204 – 4209.
710. A. Bacher in F. Müller (ed.): *Chemistry and Biochemistry of Flavoenzymes,* **vol. I,** CRC Press, Boca Raton 1991, pp. 215 – 259.
711. P. Karrer, K. Schöpp, F. Benz, *Helv. Chim. Acta* **18** (1935) 426.
712. H. v. Euler et al., *Helv. Chim. Acta* **18** (1935) 522.
713. P. Karrer et al., *Helv. Chim. Acta* **18** (1935) 1435.
714. R. Kuhn et al., *Naturwissenschaften* **23** (1935) 260.
715. R. Kuhn, *Chem. Ztg.* **59** (1935) 604.
716. R. Kuhn, K. Reinemund, F. Weygand, R. Ströbele, *Chem. Ber.* **68** (1935) 1765.
717. P. Karrer, H. Meerwein, *Helv. Chim. Acta* **18** (1935) 1130; **19** (1936) 264.
718. M. Tishler et al., *J. Am. Chem. Soc.* **69** (1947) 1487.
719. Hoffmann-La Roche, US 2 422 997, 1947 (H. M. Wüst).
720. BASF, DE 2 558 515, 2 558 516, 1975 (H. Scholz).
721. Merck & Co., US 2 420 210, 1947 (M. Tishler).
722. F. Weygand, *Chem. Ber.* **53** (1940) 1259.
723. Miles Lab., US 2 406 774, 1946 (J. Kamlet).
724. F. Yoneda, Y. Sakuma, M. Ichiba, K. Shinomura, *Chem. Pharm. Bull.* **20** (1972) 1832.
725. F. Yoneda, Y. Sakuma, K. Shinomura, *J. Chem. Soc. Perkin Trans. 1,* 1978, 348.
726. BASF, US 4 125 559, 1978 (H. Scholz, G. Gotsmann).
727. Hoffman-La Roche, US 2 438 881, 1948 (L. H. Sternbach).
728. Tokyo Tanabe Pharm. Co, JP-Kokai 7 3 28 410, 1971 (Y. Takahashi, T. Sekine, Y. Kikushi).
729. V. Bilik, *Chem. Zvesti* **26** (1972) 372.
730. Merck, EP 20959, 1980 (R. Wolf, F. Reiff, R. Wittmann, J. Butzke).
731. BASF, EP 288 847, 1988 (W. Dobler, J. Paust).
732. Takeda Chem. Ind., DE 2 454 931, 1975 (K. Sasajima et al.).
733. Takeda Chem. Ind., EP 501 765, 1992 (K. Miyagawa, J. Miyazaki, N. Kanzaki).
734. BASF, EP 115 295, 1984 (H. Ernst, W. Schmidt, J. Paust).
735. Chas. Pfizer & Co., US 2 324 800, 1941 (P. Pasternack, E. V. Brown).
736. BASF, DE 3 421 714, 1984 (J. Grimmer, H. C. Horn).
737. Coors Biotech Products Co., WO 92/01060, 1992 (R. Bailey et al.).
738. Hoffmann-La Roche, EP-A 0 405 370, 1989 (J. Perkins et al.).
739. Ajinomoto Co., EP-A 0 531 708, 1991 (N. Usui et al.).
740. M. J. Buitrago, G. A. Gonzales, J. E. Saiz, J. Revuelta, *Yeast* **9** (1993) 1099 – 1102.
741. R. Kuhn, H. Rudy, *Ber. Dtsch. Chem. Ges. B* **68** (1935) 383.
742. V. M. Berezovskii, R. V. Artemkina, E. D. Komuteva, *J. Gen. Chem. USSR (Engl. Transl.)* **34** (1963) 2813.
743. G. Scola-Nagelschneider, P. Hemmerich, *Eur. J. Biochem.* **66** (1976) 567.
744. P. Nielsen, P. Rauschenbach, H. Bacher, *Methods Enzymol.* **122** (1986) 209.
745. *Chem. Eng. N.Y.* **61** (1954) no. 8, 146; **61** (1954) no. 11, 120.
746. Hoffmann-La Roche, US 2 610 177, 1952 (L. A. Flexser, W. G. Farkas).
747. BASF, EP 336 214, 1988 (W. Dobler, M. Eggersdorfer, J. Paust).

748. I. G. Moffat, H. G. Khorona, *J. Am. Chem. Soc.* **83** (1961) 649.
749. F. Cramer, H. Neunkreffer, *Chem. Ber.* **95** (1962) 1664.
750. Y. Akahabe, JP 7 3 86 898, 1972.
751. Yamasa Shoyu Co., DE-OS 2 038 262, 1970 (K. Tamura et al.).
752. Takeda Chem. Ind., JP 9037 (65), 1961 (T. Suzuki et al.).
753. Kyowa Hakko Kogyo, EP-A 542 240, 1993 (M. Arimoto et al.).
754. A. Bacher in F. Müller (ed.): *Chemistry and Biochemistry of Flavoenzymes,* **vol. I,** Chap. 15, CRC Press, Boca Raton 1990, pp. 349 – 370.
755. R. D. Williams, H. L. Mason, P. L. Cusik, R. M. Wilder, *J. Nutr.* **35** (1943) 361.
756. M. K. Horwitt, *J. Nutr.* **39** (1949) 357.
757. I. A. Tillotson, E. M. Baker, *Am. J. Clin. Nutr.* **25** (1972) 425.
758. A. Maquinay, N. Broukou, *J. Pharm. Belg.* **39** (1957) 350.
759. R. P. Hansinger, J. F. Honeck, Ch. Walsh, *Methods Enzymol.* **122** (1986) (Vitam. Coenzyms, PT. G) 199 – 209.
760. C. Vidal-Valerde, A. Reche, *J. Liq. Chromatogr.* **13** (1990) no. 10, 2089 – 2101.
761. P. Nielsen, P. Rauschenbach, A. Bacher, *Anal. Biochem.* **130** (1983) 359 – 368.
762. IUPAC – IUB Commission on Biochemical Nomenclature: Nomenclature for Vitamins B_6 and Related Compounds, Recommendations, *Eur. J. Biochem.* **40** (1973) 325.
763. J. Goldberger, C. A. Wheelen, R. D. Lillie, L. M. Rogers, *Public Health Rep.* **41** (1926) 297.
764. P. György, *Nature (London)* **133** (1934) 498; *Biochem. J.* **29** (1935) 741. T. W. Birch, P. György, *Biochem. J.* **30** (1936) 304.
765. S. Lepkovsky, *Science (Washington D. C. 1883)* **87** (1938) 169; *J. Biol. Chem.* **124** (1938) 125. J. C. Keresztesy, J. R. Stevens, *Proc. Soc. Exp. Biol. Med.* **38** (1938) 64; *J. Am. Chem. Soc.* **60** (1938) 1267. P. György, *J. Am. Chem. Soc.* **60** (1938) 983. R. Kuhn, G. Wendt, *Ber. Dtsch. Chem. Ges.* **71** (1938) 780, 1118. A. Ichiba, K. Mití, *Sci. Pap. Inst. Phys. Chem. Res. Jpn.* **34** (1938) 623; *Chem. Abstr.* **32** (1938) 7534.
766. S. A. Harris, K. Folkers, *Science (Washington D.C. 1883)* **89** (1939) 347. S. A. Harris, E. T. Stiller, K. Folkers, *J. Am. Chem. Soc.* **61** (1939) 1245. R. Kuhn, K. Westphal, G. Wendt, O. Westphal, *Naturwissenschaften* **27** (1939) 469.
767. E. E. Snell, B. M. Guirard, R. J. Williams, *J. Biol. Chem.* **143** (1942) 519. E. E. Snell, *J. Am. Chem. Soc.* **66** (1944) 2082; *J. Biol. Chem.* **154** (1944) 313; **157** (1945) 491. E. E. Snell, A. N. Rannefeld, *J. Biol. Chem.* **157** (1945) 475.
768. S. Ohdake, *Bull. Agric. Chem. Soc. Jpn.* **8** (1932) 11.
769. *Kirk-Othmer,* 2nd ed., **16,** 806.
770. O. Isler, G. Brubacher, S. Ghisla, B. Kräutler in: *Vitamine II,* Thieme Verlag, Stuttgart 1988.
771. H. Y. Aboul-Enein, M. A. Loutfy, *Anal. Profiles Drug Subst.* **13** (1984) 447.
772. *Ullmann,* 4th ed., **23,** 671.
773. J. M. Osbond, *Vitam. Horm. N.Y.* **22** (1964) 387.
774. J. Marks: *A Guide to the Vitamins,* MTP Press, Lancaster 1975, p. 93.
775. H. E. Sauberlich, M. J. Kretsch, H. L. Johnson, R. A. Nelson in D. C. Beitz, R. G. Hansen (eds.): *Animal Products in Human Nutrition,* Academic Press, New York 1982, p. 339.
776. W. Friedrich: *Handbuch der Vitamine,* Urban & Schwarzenberg, München 1987, p. 349.
777. J. E. Leklem in L. J. Machlin (ed.): *Handbook of Vitamins,* 2nd ed., Marcel Dekker, New York 1991, p. 341.
778. R. E. Hill et al. *J. Biol. Chem.* **262** (1987) 7463.
779. R. E. Hill, B. G. Sayer, I. D. Spencer, *J. Am. Chem. Soc.* **111** (1989) 1916.
780. W. B. Dempsey in F. C. Neidhardt et al. (eds.): *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology,* American Society for Microbiology, Washington 1987, p. 539.
781. H.-M. Lam, M. E. Winkler, *J. Bacteriol.* **172** (1990) 6518.
782. H. König, W. Böll, *Chem. Ztg.* **100** (1976) 105.
783. A. Ichiba, K. Miti, *Sci. Pap. Inst. Phys. Chem. Res. Jpn.* **36** (1939) 173; *Chem. Abstr.* **33** (1939) 8201.
784. IG Farbenind., DE 699 555, 1939 (W. Salzer). IG Farbenind., DE 710 396, 1939 (W. Salzer).
785. S. A. Harris, K. Folkers, *J. Am. Chem. Soc.* **61** (1939) 1245, 3307. S. A. Harris, K. Folkers, *Science (Washington D. C. 1883)* **89** (1939) 347.
786. Hoffmann-La Roche, US 2 384 136, 1941 (O. Schnider).

787. K. Aso, *Tohoku J. Agric. Res.* **1** (1950) 125.
788. Sadolin und Holmblad, DE 1 134 379, 1954 (N. Clauson-Kaas, N. Elmig) N. Clauson-Kaas, N. Elmig, Z. Tyle, *Acta Chem. Scand.* **9** (1955) 1, 9, 14, 17, 23.
789. BASF, DE 1 935 009, 1969 (F. Graf, H. König).
790. T. Shono, Y. Matsumara, K. Tsubata, J. Takata, *Chem. Lett.* 1981, 1121.
791. G. Y. Kondratyeva, *Khim. Nauka Promst.* **2** (1957) 666.
792. G. Y. Kondratyeva, *Dokl. Akad. Nauk SSSR* **141** (1961) 628, 861. Hoffmann-La Roche, US 3 250 778, 1962 (W. Kimel, W. Leimgruber). E. E. Harris, R. A. Firestone, *J. Org. Chem.* **27** (1962) 2705.
793. P. Karrer, G. Granacher, *Helv. Chim. Acta* **7** (1924) 763.
794. Takeda, JP-035 887, 1966 (T. Miki, T. Matsuo). T. Miki, T. Matsuo, *Yakugaku Zasshi* **87** (1967) 323.
795. Ajinomoto, FR 1 533 817, 1968 (R. Yoshida, I. Maeda, S. Asai, M. Takehara). I. Maeda et al. *Bull. Chem. Soc. Jpn.* **42** (1969) 1435. Shanghai Med. Indust. Inst., CN 86 101 512, 1986 (Zhou Houyuan).
796. Hoffmann-La Roche, US 3 222 374, 1964 (G. O. Chase).
797. Hoffmann-La Roche, DE 2 847 625, 1978 (H. Hoffmann-Paquotte).
798. Hoffmann-La Roche, EP 492 233, 1990 (P. Nösberger).
799. Hoffmann-La Roche, CH 578/93, 1993 [=EP 612 736] (W. Bonrath, R. Karge, H. Pauling).
800. Hoffmann-La Roche, US 3 250 778, 1962 (W. Kimel, W. Leimgruber).
801. BASF, DE 2 143 989, 1971 (W. Böll, H. König). BASF, DE 2 717 478, 1977 (W. Böll, H. König).
802. Daiichi, JP 57 134 467, 1981 (S. Shimada, M. Oki).
803. R. E. Geiger, M. Lalonde, H. Stoller, K. Schleich, *Helv. Chim. Acta* **67** (1984) 1274. C. A. Parnell, K. P. C. Vollhardt, *Tetrahedron* **41** (1985) 5791.
804. A. N. Wilson, S. A. Harris, *J. Am. Chem. Soc.* **73** (1951) 4693. J. Baddiley, A. P. Mathias, *J. Chem. Soc.* **1952,** 2583. Targa S. A., ES 516 847 A 1, 1982 (F. Montoro, J. Calatayad).
805. C. J. Argoudelis, *J. Agric. Food Chem.* **34** (1986) 995.
806. M. Viscontini, C. Ebnöther, P. Karrer, *Helv. Chim. Acta* **34** (1951) 1834.
807. E. A. Peterson, H. A. Sober, *J. Am. Chem. Soc.* **76** (1954) 169.
808. M. Iwata, *Bull. Chem. Soc. Jpn.* **54** (1981) 2835.
809. S. A. Harris, D. Heyl, K. Folkers, *J. Biol. Chem.* **154** (1944) 315. S. A. Harris, D. Heyl, K. Folkers, *J. Am. Chem. Soc.* **66** (1944) 2088.
810. D. Heyl, *J. Am. Chem. Soc.* **70** (1948) 3434.
811. Merck, GB 852 398, 1960.
812. Hoffmann-La Roche, US 2 755 284, 1954 (R. F. Long).
813. Nisshin Flour Milling Co., JP 75 131 965, 1974 (J. Hamashima, K. Minami).
814. M. Viscontini, P. Karrer, *Helv. Chim. Acta* **35** (1952) 1924.
815. M. Iwanami, T. Numata, M. Murakami, *Bull. Chem. Soc. Jpn.* **41** (1968) 161. S. Wang, *Huadong Huagong Xueyuan Xuebao,* **3** (1983) 369; *Chem. Abstr.* **100** (1984) 209 986.
816. K. Nakagawa, M. Matsui, *Nippon Nogei Kagaku Kaishi* **42** (1968) 300.
817. Merck, US 3 124 587, 1960 (G. Schorre).
818. Daiichi, JP 68 15 437, 1969 (T. Yoshikawa, T. Kabayashi, F. Ishikawa, T. Naito).
819. H. Tsuge, J. Sakuma, K. Ozeki, K. Ohashi, *Vitamin* **55** (1981) 437; *Chem. Abstr.* **95** (1981) 202 054.
820. H. Yamada, S. Shimizu, Y. Tani, T. Hino, *Enzyme Eng.* **5** (1980) 405; *Chem. Abstr.* **94** (1981) 28 873.
821. P. Karrer, M. Viscontini, D. Forster, *Helv. Chim. Acta* **31** (1948) 1016.
822. M. V. Balyakina, N. L. Yakovleva, V. I. Gunar, *Khim. Farm. Zh.* **14** (1980) 79; *Chem. Abstr.* **93** (1980) 204 416.
823. Lepetit S.p.A, DE 1 012 604, 1955 (G. Carrara, E. Testa).
824. Takeda, JP 52 128 378, 1976 (Y. Fujiwara, T. Niinobe).
825. Unitika Ltd., JP 52 005 772, 1975 (J. Tabuse, Y. Yabushita).
826. Takeda, BE 648 226, 1964 (A. Euchi, K. Kunimitsu, M. Tetsuo, T. Noriaki, Y. Mikio).
827. W. S. McNutt, E. E. Snell, *J. Biol. Chem.* **182** (1950) 557.
828. M. Viscontini, C. Ebnöther, P. Karrer, *Helv. Chim. Acta* **34** (1951) 2199.
829. P. F. Mühlradt, Y. Morino, E. E. Snell, *J. Med. Chem.* **10** (1967) 341.
830. L. M. Henderson in R. D. Reynolds, J. E. Leklem (eds.): *Vitamin* B_6. Its Role in Health and Disease, A. R. Riss, New York 1985, p. 22.

831. L. Lumeng, T.-K. Li in G. P. Tryfiates (ed.): *Vitamin B_6* Metabolism and Role in Growth, Food and Nutrition Press, Westpoint, CT, 1980, p. 27.
832. H. Wada, E. E. Snell, *J. Biol. Chem.* **236** (1961) 2089.
833. A. H. Merrill et al. *J. Nutr.* **114** (1984) 1664.
834. V. E. Allgood, J. A. Cidlowski, *J. Nutr. Biochem.* **2** (1991) 523.
835. G. Litwack, N. M. Robertson, A. B. Maksymowych, M. Celiker in M. M. Jacobs (ed.): *Vitamins and Minerals in the Prevention and Treatment of Cancer,* CRC Press, Boca Raton 1991, p. 19.
836. J. P. Bonjour, *Int. J. Vitam. Nutr. Res.* **50** (1980) 215.
837. J. M. Henderson et al. *Hepatology (Baltimore)* **6** (1986) 464.
838. R. R. Brown et al. *Am. J. Clin. Nutr.* **28** (1975) 10.
839. H. B. Carlsson, E. M. Anthony, W. F. Russel, G. Middlebrook, *N. Engl. J. Med.* **225** (1956) 118.
840. S. Wason, P. G. Lacouture, F. H. Lovejoy, *JAMA J. Am. Med. Assoc.* **246** (1981) 1102.
841. L. E. Hollister, F. F. Moore, F. Forrest, J. L. Bennett, *Am. J. Clin. Nutr.* **19** (1966) 307.
842. K. H. Bässler, *Int. J. Vitam. Nutr. Res.* **58** (1988) 105.
843. A. H. Merrill, J. M. Henderson, *Annu. Rev. Nutr.* **7** (1987) 137.
844. M. L. Orr: "Pantothenic Acid, Vitamin B_6, and Vitamin B_{12} in Foods," *Home Econ. Res. Rep.* **36** (1969) 53.
845. J. F. Gregory, *J. Food Comp. Anal.* **1** (1988) 105.
846. S. P. Colowick, N. O. Kaplan: *Methods in Enzymology,* Academic Press, New York-London 1970, pp. 505, 509.
847. J. P. Vuilleumier, J. P. Keller, R. Rettenmeier, F. Hunziker, *Int. J. Vitam. Nutr. Res.* **53** (1983) 359.
848. N. Raica, H. E. Sauberlich, *Am. J. Clin. Nutr.* **15** (1965) 67.
849. H. E. Sauberlich, P. D. Dowdy, J. H. Skala: *Laboratory Tests for the Assessment of Nutritional Status,* CRC Press, Cleveland, OH, 1974, p. 37.
850. G. Brubacher, W. Müller-Mulot, D. A. T. Southgate: *Methods for the Determination of Vitamins in Food,* Applied Science Publ., London 1984, p. 129.
851. D. Hess, J. P. Vuilleumier, *Int. J. Vitam. Nutr. Res.* **59** (1989) 338.
852. *Chem. Mark. Rep.* (1994) Sept. 12.
853. *Recommended Dietary Allowances,* 10th ed., National Academic Press, Washington 1989.
854. M. Brin in: *Human Vitamin B6 Requirements,* National Academy of Sciences, Washington 1978, p. 1.
855. J. C. Bauernfeind, O. N. Miller in: *Human Vitamin B6 Requirements,* National Academy of Sciences, Washington 1978, p. 78.
856. K. S. Khera, *Experientia* **31** (1975) 469.
857. M. R. Marathe, G. P. Thomas, *Indian J. Physiol. Pharmacol.* **30** (1986) 264.
858. H. Schaumberg et al., *N. Engl. J. Med.* **309** (1983) 445.
859. M. Cohen, A. Bendich, *Toxicol. Lett.* **34** (1986) 129.
860. A. Bendich, M. Cohen, *Ann. N. Y. Acad. Sci.* **585** (1990) 321.
861. K. H. Bässler in P. Walter, H. Stähelin, G. Brubacher (eds.): *Elevated Dosages of Vitamins,* Hans Huber Publ., Toronto; Lewinston, New York Bern, Stuttgart 1989, p. 120.
862. B. Kräutler, D. Kohler, E. Stupperich, *Eur. J. Biochem.* **176** (1988) 461 – 469.
863. G. Barrère et al., *Pour la Science* **49** (1981) 56 – 64.
864. B. Kräutler: "Wasserlösliche Vitamine" in O. Isler (ed.): *Vitamine,* vol. 2, Thieme Verlag, Stuttgart 1986, pp. 340 – 388.
865. R. Strohecker, H. M. Hening: *Vitamin Assay – Tested Methods,* Verlag Chemie, Weinheim 1966.
866. A. I. Scott, *Tetrahedron* **50** (1994) no. 47, 13 315 –13 333.
867. A. R. Battersby, *Science* **264** (1994) 1551 – 1557.
868. J. R. Roth et al., *J. Bacteriol.* **175** (1993) no. 11, 3303 – 3316.
869. C. Spalla et al. in E. J. Vandame (ed.): *Biotechnology of Vitamins, Pigments and Growth Factors,* **vol. 1,** Elsevier, Amsterdam 1989, pp. 257 – 284.
870. D. Perlman, *Adv. Appl. Microbiol.* **1** (1959) 87 – 122.
871. J. Florent, L. Ninent in H. J. Peppler, D. Perlmann (eds.) *Microbial Technology,* Academic Press, New York 1979, pp. 497 – 519.
872. Rhône-Poulenc Biochimie, WO 91/11 518, 1991 (F. Blanche et al.).
873. Merck und Co., Inc., US 2 939 822, 1960 (J. M. Miller, C. Rosenblum).
874. M. Doudoroff et al., *Int. J. Syst. Bacteriol.* **24** (1974) no. 2, 294 – 300.

875. Pierrel, GB 1 007 972, 1965 (I. Rolovich, G. Ragri).
876. W. Friedrich: *Handbuch der Vitamine,* Urban + Schwarzenberg, München 1987, p. 581.

General References
877. P. A. Seib, P. M. Tolbert (eds.): *Ascorbic Acid: Chemistry, Metabolism, and Uses,* American Chemical Society, Washington, D.C., 1982.
878. U. Moser, A. Bendich: "Vitamin C," in L. J. Machlin (ed.): *Handbook of Vitamins,* 2nd ed., Marcel Dekker, New York 1991, pp. 195 – 232.
879. O. Isler, G. Brubacher, J. Kiss: "Vitamin C," in O. Isler, G. Brubacher, S. Gishla, B. Kräutler (eds.): *Vitamine II,* Thieme Verlag, Stuttgart – New York 1988, pp. 396 – 444.
880. M. B. Davies, J. Austin, D. A. Partridge: *Vitamin C, Its Chemistry and Biochemistry,* The Royal Society of Chemistry, Cambridge 1991.
881. U. Wintermeyer, H. Lahann, R. Vogel: *Vitamin C: Entdeckung, Identifizierung und Synthese – Heutige Bedeutung in Medizin und Lebensmitteltechnologie,* Deutscher Apotheker-Verlag, Stuttgart 1981.
882. K. J. Carpenter: *The History of Scurvy and Vitamin C,* Cambridge University Press, Cambridge 1986.
883. M.-L. Liao, P. A. Seib: "Chemistry of L-Ascorbic Acid Related to Foods," *Food Chem.* **30** (1988) 289 – 312.
884. P. A. Seib: "Oxidation, Monosubstitution and Industrial Synthesis of Ascorbic Acid," *Int. J. Vitam. Nutr. Res. Suppl.* **27** (1985) 259 – 306.
885. T. C. Crawford, S. A. Crawford: "Synthesis of L-Ascorbic Acid," *Adv. Carbohydr. Chem. Biochem.* **37** (1980) 79 – 155.
886. L. A. Pachla, D. L. Reynolds, P. T. Kissinger: "Analytical Methods for Determining Ascorbic Acid in Biological Samples, Food Products, and Pharmaceuticals," *J. Assoc. Off. Anal. Chem.* **68** (1985) 1 – 12.
887. M. T. Parviainen, K. Nyyssönen: "Ascorbic Acid," *Chromatogr. Sci.* **60** (1992) 235 – 260.
888. F. A. Loewus, M. W. Loewus: "Biosynthesis and Metabolism of Ascorbic Acid in Plants," *CRC Crit. Rev. Plant Sci.* **5** (1987) 101 – 119.
889. J. Boudrant: "Microbial Processes for Ascorbic Acid Biosynthesis," *Enzyme Microb. Technol.* **12** (1990) 322 – 329.
890. J. C. Bauernfeind: Ascorbic Acid Technology in Agricultural, Pharmaceutical, Food, and Industrial Applications," in [877] pp. 395 – 497.
891. R. L. Newsome: "Use of Vitamins as Additives in Processed Foods," *Food Technol.* **9** (1987) 163 – 168.
892. A. Deifel: "Die Chemie der L-Ascorbinsäure in Lebensmitteln," *Chem. Unserer Zeit* **27** (1993) 198 – 207.

Specific References
893. Commission on Biochemical Nomenclature IUPAC-IUB, *Biochim. Biophys. Acta* **107** (1965) 4.
894. A. Szent-Györgyi, *Biochem. J.* **22** (1928) 1387.
895. W. N. Haworth, E. L. Hirst, *J. Chem. Soc. Ind.* **52** (1933) 645.
896. A. Holst, T. Frölich, *J. Hyg.* **7** (1907) 634.
897. T. Reichstein, A. Grüssner, R. Oppenhauer, *Helv. Chim. Acta* **16** (1933) 561. W. N. Haworth, *Chem. Ind. (London)* **52** (1933) 482.
898. A. Szent-Györgyi, W. N. Haworth, *Nature (London)* **131** (1933) 24.
899. S. Budavari, M. J. O'Neil, A. Smith, P. E. Heckelman (eds.): *The Merck Index,* 11th ed., Merck & Co., Rahway, N.J., 1989, p. 130.
900. I. K. Thege, E. Pungor, E. Schulek, *Pharm. Zentralhalle* **99** (1960) 231.
901. W. Jabs, W. Gaube, C. Fehl, *J. Prakt. Chem.* **329** (1987) 933.
902. B. H. J. Bielski, D. A. Comstock, R. A. Bowen, *J. Am. Chem. Soc.* **93** (1971) 5624.
903. M. Falk, M. J. Wójcik, *Spectrochim. Acta Part A* **35 A** (1979) 1117.
904. D. T. Sawyer, J. R. Brannan, *Anal. Chem.* **38** (1966) 192.
905. T. Radford, J. G. Sweeny, G. A. Iacobucci, *J. Org. Chem.* **44** (1979) 658.
906. S. Berger, *Tetrahedron* **33** (1977) 1587.
907. A. Benninghoven, W. K. Sichtermann, *Anal. Chem.* **50** (1978) 1180.
908. J. Hvoslef: "Crystallography of the Ascorbates," in [877] pp. 37 – 57.
909. J. V. Paukstelis, D. D. Mueller, P. A. Seib, D. W. Lillard: "NMR Spectroscopy of Ascorbic Acid and Its Derivatives," in [877] pp. 125 – 151.
910. G. S. Brenner, D. F. Hinkley, L. M. Perkins, S. Weber, *J. Org. Chem.* **29** (1964) 2389.
911. B. H. J. Bielski: "Chemistry of Ascorbic Acid Radicals," in [877] pp. 81 – 100.
912. B. H. J. Bielski, A. O. Allen, H. A. Schwarz, *J. Am. Chem. Soc.* **103** (1981) 3516.
913. A. E. Martell: "Chelates of Ascorbic Acid," in [877] pp. 153 – 178.

914. M. Ohmori, H. Higashioka, M. Takagi, *Agric. Biol. Chem.* **47** (1983) 607.
915. B. M. Tolbert, J. B. Ward: "Dehydroascorbic Acid," in [877] pp. 101 – 123.
916. B. R. Hajratwala, *S.T.P. Pharma* **1** (1985) no. 4, 281. K. Niemelä, *J. Chromatogr.* **399** (1987) 235. J. C. Deutsch, C. R. Santhosh-Kumar, K. L. Hassell, J. F. Kolhouse, *Anal. Chem.* **66** (1994) 345.
917. P. A. Seib et al., *J. Chem. Soc. Perkin Trans. I* 1974, 1220.
918. C. H. Lee et al., *Carbohydr. Res.* **67** (1978) 127.
919. P. W. Lu et al., *J. Agric. Food Chem.* **32** (1984) 21.
920. M. Oliver: "Ascorbic Acid: Estimation," in W. H. Sebrell, R. S. Haris (eds.): *The Vitamins,* vol. I., Academic Press, New York 1967, pp. 338 – 359.
921. British Pharmacopoeia, HM Stationery Office, London 1988, p. 47.
922. K. K. Verma, A. K. Gulati, *Anal. Chem.* **52** (1980) 2336.
923. J. Tillmanns, P. Hirsch, J. Jackish, *Z. Lebensm. Unters. Forsch.* **63** (1932) 241. H. Sauberlich, W. C. Goad, J. H. Skala, P. P. Waring, *Clin. Chem. (Winston – Salem, N.C.)* **22** (1976) 105.
924. H. O. Beutler, G. Beinstingl, *Dtsch. Lebensm. Rundsch.* **76** (1980) 69. G. J. Kelly, E. Latzko, *J. Agric. Food Chem.* **28** (1980) 1320.
925. J. V. Lloyd, P. S. Davis, H. Lander, *J. Clin. Pathol.* **22** (1969) 453. V. Zannoni, M. Lynch, G. Goldstein, P. Sato, *Biochem. Med.* **11** (1974) 41.
926. J. H. Roe, C. A. Kuether, *J. Biol. Chem.* **147** (1943) 399.
927. M. Schmall, C. W. Pfifer, E. G. Wollish, *Anal. Chem.* **25** (1953) 1486. S. S. Wilson, R. A. Guillan, *Clin. Chem. (Winston – Salem, N.C.)* **15** (1969) 282.
928. M. J. Deutsch, C. E. Weeks, *J. Assoc. Off. Agric. Chem.* **48** (1965) 1248. M. J. Deutsch, *J. Assoc. Off. Anal. Chem.* **67** (1984) 380.
929. M. H. Lee, C. R. Dawson, *Methods Enzymol.* **62** (1979) 30. T. Tono, S. Fujita, *Agric. Biol. Chem.* **45** (1981) 2947. T. Z. Liu, N. Chin, M. D. Kiser, W. N. Bigler, *Clin. Chem. (Winston – Salem, N.C.)* **28** (1982) 2225.
930. S. P. Perone, W. J. Kretlow, *Anal. Chem.* **38** (1966) 1760. A. N. Strohl, D. J. Curran, *Anal. Chem.* **51** (1979) 1045. U. Korell, R. B. Lennox, *Anal. Chem.* **64** (1992) 147.
931. C. C. Sweeley, R. Bentley, M. Makita, W. W. Wells, *J. Am. Chem. Soc.* **85** (1963) 2497.
932. J. C. Saari, E. M. Baker, H. E. Sauberlich, *Anal. Biochem.* **18** (1967) 173. M. Otsuka, T. Kurata, N. Arakawa, *Agric. Biol. Chem.* **50** (1986) 531. R. Klaus, W. Fischer, H. E. Hauck, *LG-GC* **13** (1995) 816.
933. L. W. Doner, K. B. Hicks, *Anal. Biochem.* **115** (1981) 225.
934. C. S. Tsao, S. L. Salimi, *J. Chromatogr.* **245** (1982) 355.
935. A. Lopez-Anaya, M. Mayersohn, *Clin. Chem. (Winston – Salem, N.C.)* **33** (1987) 1874.
936. J. W. Erdmann, B. P. Klein: "Harvesting, Processing, and Cooking Influences on Vitamin C in Foods," in [877] pp. 499 – 532.
937. G. Kiss, H. Neukom, *Helv. Chim. Acta* **49** (1966) 989.
938. T. Tanaka, G. Nonaka, I. Nishioka, *J. Chem. Soc. Perkin Trans. 1* 1986, 369.
939. M. Olliver: "Ascorbic Acid: Occurrence in Food," in W. H. Sebrell, R. S. Harris (eds.): The Vitamins, vol. I, Academic Press, New York 1967, pp. 359 –367. A. A. Paul, D. A. T. Southgate, *McCance and Widdowson's The Composition of Foods,* 4th ed., Medical Research Council's Special Report no. 297, London, HMS, 1978, 40 – 272. S. Nobile, J. M. Woodhill: *Vitamin C: The Mysterious Redox System, A Trigger of Life,* MTP Press, Boston 1981, pp 38 – 55.
940. C. G. King, *World Rev. Nutr. Diet.* **18** (1973) 47.
941. P. H. Sato, L. A. Roth, D. M. Walton, *Biochem. Med. Metabol. Biol.* **35** (1986) 59.
942. P. Sato, S. Udenfriend, *Vitam. Horm. (N.Y.)* **36** (1978) 33.
943. I. B. Chatterjee, *Science (Washington, D.C. 1883)* **82** (1973) 1271.
944. F. A. Isherwood, Y. T. Chen, L. W. Mapson, *Biochem. J.* **56** (1954) 1. K. Ôba et al., *Biochem.* **117** (1995) 120 and refs. cited therein.
945. F. A. Loewus, J. P. F. G. Helsper: "Metabolism of L-Ascorbic Acid in Plants," in [877] pp. 249 –261.
946. S. T. Omaye, J. A. Tillotson, H. E. Sauberlich: "Metabolism of L-Ascorbic Acid in the Monkey," in [877] pp. 317 – 334.
947. A. H. Conney, G. A. Bray, C. Evans, J. J. Burns, *Ann. N.Y. Acad. Sci.* **92** (1961) 115.
948. T. Reichstein, A. Grüssner, R. Oppenhauer, *Helv. Chim. Acta* **16** (1933) 1019; *Nature (London)* **132** (1933) 280.
949. T. Reichstein, A. Grüssner, *Helv. Chim. Acta* **17** (1934) 311.

950. A. T. Kuhn, *J. Appl. Electrochem.* **11** (1981) 261; Agency of Ind. Sci. and Technology, JP 60 083 588 (1985).
951. K. Nakagawa, R. Konaka, T. Nakata, *J. Org. Chem.* **27** (1962) 1597.
952. E. Merck, DE 3 019 321, 1981 (R. Wittmann, W. Wintermeyer, J. Butzke).
953. P. Seiler, P. M. Robertson, *Chimia* **36** (1982) 305.
954. Hoffmann-La Roche, DE 2 503 819, 1980 (P. M. Robertson).
955. J. Nanzer, S. Langlois, F. Coeuret, *J. Appl. Electrochem.* **22** (1993) 471.
956. Hoffmann-La Roche, DE 2 123 621, 1971 (G. M. Jaffe, E. J. Pleven).
957. L. O. Shnaidman, I. N. Kushehinskaya, *Tr. Vses. Nauchno-Issled. Vitam. Inst.* **8** (1961) 13.
958. Hoffmann-La Roche, US 2 179 978, 1939 (F. Elger).
959. Hoffmann-La Roche, GB 469 157, 1937 (T. Reichstein).
960. Hoffmann-La Roche, GB 428 815, 1935 (T. Reichstein).
961. Hoffmann-La Roche, GB 466 548, 1937 (T. Reichstein).
962. A. A. Othman, U.S. Al-Timari *Tetrahedron* **36** (1980) 753.
963. Hoffmann-La Roche: "Vitamins, Part VIII: Vitamin C," in Y. H. Hui (ed.): *Encyclopedia of Food Science and Technology,* vol. 4., J. Wiley, New York 1992, p. 2732.
964. P. A. Garcia, R. Velasco, F. Barba, *Synth. Commun.* **21** (1991) 1153.
965. Merck & Co., DE-OS 1 813 757, 1968; US 3 721 663, 1973 (D. F. Hinkley, A. M. Hoinowski).
966. Á. Lengyel-Mészáros, B. Losonczi, J. Petró, I. Rusznák, *Acta Chim. Acad. Sci. Hung.* **97** (1978) no. 2, 213. Mitsui Toatsu Chemicals, JP 57 163 340, 1983 (T. Kiyoura).C. Brönnimann, T. Mallat, A. Baikev, *J. Chem. Soc., Chem. Commun.* **1995,** 1377 and refs. cited therein.
967. S. Makover et al., *Biotechnol. Bioeng.* **17** (1975) 1485.
968. Beijing Pharmaceutical Plant Guanghuala, EP 0 278 447, 1988 (W. Ning et al.).
969. Takeda, EP 0 295 861, 1988 (I. Nogami, T. Yamaguchi, M. Oka, H. Shirafuji).
970. Hoffmann-La Roche, EP 0 518 136, 1992 (T. Hoshino et al.)
971. Takeda, GB 2 217 714, 1989 (M. Tsuda, K. Iwai).
972. J. Bakke, O. Theander, *J. Chem. Soc. D* 1971, 175.
973. T. Kitahara, T. Ogawa, T. Naganuma, M. Matsui, *Agric. Biol. Chem.* **38** (1974) 2189.
974. ICI, US 4 326 072, 1982 (W. M. Kruse, M. H. Meshreki). C. Brönnimann, T. Mallat, A. Baiker, *J. Chem. Soc. Chem. Commun.* 1995, 1377 and refs. cited therein.
975. P. P. Regna, B. P. Calwell, *J. Am. Chem. Soc.* **66** (1944) 243.
976. Pfizer US 4 155 812, 1979 (D. Kita).
977. Hoffmann-La Roche, EP 0 476 442, 1992 (T. Hoshino, P. K. Matzinger, S. Ojima, T. Sugisawa); *Biosci. Biotech. Biochem.* **59** (1995) 190.
978. C. L. Mehltretter, *Methods Carbohydr. Chem.* **2** (1963) 29.
979. K. Heyns, G. Graefe, *Chem. Ber.* **86** (1953) 646.
980. T. Sonoyama, S. Yagi, B. Kageyama, *Agric. Biol. Chem.* **52** (1988) 667. BASF, DE 4 238 905, 1994 (J. Meyer, B. Hauer).
981. Pfizer, DE 2 947 741, 1979 (G. Andrews). Pfizer, EP 0 024 114, 1981 (G. Colton). Hoffmann-La Roche, EP 0 133 493, 1985 (P. Fahrni, T. Siegfried).
982. T. Sonoyama et al., *Appl. Environ. Microbiol.* **43** (1982) 1064 – 1069.
983. T. Sonoyama, B. Kageyama, S. Yagi, K. Mitsushima, *Agric. Biol. Chem.* **51** (1987) 3039.
984. S. Yagi, B. Kageyama, T. Sonoyama, *Agric. Biol. Chem.* **54** (1990) 1597.
985. R. A. Lazarus et al.: "A Biocatalytic Approach to Vitamin C Production," in D. A. Abramowicz (ed.): *Biocatalysis,* Van Nostrand Reinhold, New York 1990, pp. 135 – 155. Genentech, US 5 032 514, 1991 (R. A. Lazarus et al.). J. F. Grindley, M. A. Payton, H. Van de Pol, G. Hardy, *Appl. Environ. Microbiol.* **54** (1988) 1770. Biogen, WO 87/00863, 1987 (J. F. Grindley et al.).
986. Bernard Wolnak and Assoc., US 4 259 443, 1981 (J. P. Danehy).
987. Kraft, WO 85/01745, 1985 (J. F. Roland, T. Cayle, R. C. Dinwoodie, D. W. Mehnert).
988. Bio-Technical Resources, US 5 001 059; WO 93/19192, 1993; WO 95/21933, 1993 (R. J. Huss et al.).
989. D. Hornig, J. P. Vuilleumier, D. Hartmann, *Int. J. Vitam. Nutr. Res.* **50** (1980) 309.
990. C. J. Schorah, *Proc. Nutr. Soc.* **51** (1992) 189.
991. J. Meyle, W. Heller, W. Schulte, *Medwelt* **38** (1987) 902.
992. M. Levine, *N. Engl. J. Med.* **314** (1986) 892.

993. C. C. Glembotski, *Ann. N.Y. Acad. Sci.* **498** (1987) 54.
994. H. Hemilä, *Br. J. Nutr.* **67** (1992) 3.
995. S. R. Tannenbaum, J. S. Wishnok, C. D. Leaf, *Am. J. Clin. Nutr.* **53** (1991) 247 S.
996. T. Byers, G. Perry, *Annu. Rev. Nutr.* **12** (1992) 139.
997. E. Niki, *Ann. N.Y. Acad. Sci.* **498** (1987) 186.
998. H. Sies, W. Stahl, A. R. Sundquist, *Ann. N.Y. Acad. Sci* **669** (1992) 7.
999. J. A. Simon, *J. Am. Coll. Nutr.* **11** (1992) 107.
1000. L. Hallberg, M. Brune, L. Rossander-Hulthén, *Ann. N.Y. Acad. Sci.* **498** (1987) 324.
1001. J. M. Rivers, *Ann. N.Y. Acad. Sci.* **498** (1987) 445.
1002. A. Bendich, *Ann. N.Y. Acad. Sci.* **669** (1992) 300.
1003. H. J. Roth: "Vitamin C," *Dtsch. Apoth. Ztg.* **131** (1991) 414.
1004. Food and Nutrition Board: *Recommended Dietary Allowances,* 10th ed., National Academic Press, Washington 1989.
1005. Richtlinie des Rates der EG 0/496; Amtsblatt der Europäischen Gemeinschaft L 276, 1990.
1006. Deutsche Gesellschaft für Ernährung, Empfehlungen für die Nährstoffzufuhr, Frankfurt/Main 1991.

General References

1007. *Ullmann,* 4th ed., **23,** 692.
1008. O. Isler, G. Brubacher, S. Ghisla, B. Kräutler (eds.): *Vitamine II,* Wasserlösliche Vitamine, Thieme Verlag, Stuttgart 1988.
1009. W. Friedrich: *Handbuch der Vitamine,* Urban + Schwarzenberg, München 1987.
1010. L. J. Machlin (ed.): *Handbook of Vitamins,* Marcel Dekker, New York 1984.
1011. K. H. Bässler, E. Grühn, D. Loew: *Vitamin Lexikon,* G. Fischer Verlag, Stuttgart 1992.

Specific References

1012. R. J. Williams, E. M. Bradway, *J. Am. Chem. Soc.* **53** (1931) 783. E. E. Snell, F. M. Strong, W. H. Peterson, *Biochem. J.* **31** (1937) 1789.
1013. R. J. Williams et al., *J. Am. Chem. Soc.* **60** (1938) 2719.
1014. T. H. Jukes, *J. Am. Chem. Soc.* **61** (1939) 975.
1015. Y. Subbarow, C. H. Hitchings, *J. Am. Chem. Soc.* **61** (1939) 1615.
1016. R. J. Williams et al., *J. Am. Chem. Soc.* **55** (1933) 2912.
1017. S. W. Souci, W. Fachmann, H. Kraut: *Die Zusammensetzung der Lebensmittel,* Nährwert-Tabelle, Wissenschaftliche Verlagsgesell. Stuttgart 1989.
1018. D. A. Roth-Maier, M. Kirchgessner, *Z. Tierphysiol. Tierernähr., Futtermittelkd.* **41** (1979) 275.
1019. H. Pfaltz, *Z. Vitaminforsch.* **13** (1943) 236. S. H. Rubin, J. M. Cooperman, M. E. Moore, J. Scheiner, *J. Nutr.* **35** (1948) 499.
1020. J. L. Milligan, G. M. Briggs, *Poult. Sci.* **28** (1949) 202.
1021. G. M. Brown, *Adv. Enzymol.* **53** (1982) 371. D. J. Aberhart, *J. Am. Chem. Soc.* **101** (1979) 1354. D. R. Wilken, R. F. Dyar, *Arch. Biochem. Biophys.* **189** (1978) 251. J. H. Teller, S. G. Powers, E. F. Snell, *J. Biol. Chem.* **251** (1976) 3780, 3786. K. Miyatake, Y. Nakano, S. Kitaoka, *Methods Enzymol.* **62D** (1979) 215. A. Neufarth-Kreiling, W. Ludwig, G. Pfleiderer, *Biochem. Z.* **336** (1962) 241. R. W. Kaplan, A. Neufahrt-Kreiling, G. Pfleiderer, *Biochem. Z.* **342** (1965) 161.
1022. W. K. Maas, H. J. Vogel, *J. Bacteriol.* **65** (1953) 388.
1023. R. H. Wightmann, *J. Chem. Soc. Chem. Commun.* 1979, 818. D. Wasmuth, D. Arigoni, D. Seebach, *Helv. Chim. Acta* **65** (1982) 344.
1024. L. L. Campbell, *J. Biol. Chem.* **227** (1957) 693. A. J. Virtanen, T. Laine, *Biochem. J.* **33** (1939) 412.
1025. D. Billen, H. C. Lichstein, *J. Bacteriol.* **58** (1949) 215.
1026. S. Razin, U. Bacharach, J. Gery, *Nature (London)* **181** (1958) 700.
1027. Y. Nishizuka, M. Takeshita, S. Kuno, O. Hayaishi, *Biochim. Biophys. Acta* **33** (1959) 591. C. F. Baxter, E. Roberts, *J. Biol. Chem.* **233** (1958) 1135.
1028. E. Glaser, *Monatsh. Chem.* **25** (1904) 46. Jenapharm, DL 2 228 641, 1976 (J. Schmidt, H. Grunert, C. Weigelt). BASF, DE-OS 2 758 883, 1979 (H. Distler, W. Goetze).
1029. T. Reichstein, A. Grüssner, *Helv. Chim. Acta* **23** (1940) 650. E. T. Stiller et al., *J. Am. Chem. Soc.* **62** (1940) 1785. Merck, US 2 390 281, 1945 (M. Tishler, R. Beutel); US 2 474 719, 1949 (R. Beutel, M. Tishler). Nopco Chem. Co., US 3 009 922, 1961 (H. C. Klein, R. Kapp). Hoffmann-La Roche, US 2 423 062, 1947 (F. Bergel, A. Cohen, A. L. Morrison, A. R. Moss). Nopco Chem. Co., US 2 460 239, US 2 460 240, 1949 (F. D. Pickel, J. I. Fass, S. Chodroff).

1030. Nopco Chem. Co., US 3 185 710, 1965 (M. Dunkel, J. Loter, H. C. Klein). Jenapharm, DD 16 482, 1959 (W. Ozegowski, H. Haering). Am. Cyanamid Co., US 3 541 139, 1970 (R. Winterbottom, W. M. Ziegler). The Upjohn Co., US 2 805 219, 1957 (F. Kagan, J. W. Greiner, D. Weisblat). Jenapharm, DD 16 982, 1959 (R. Ring); DL 37 505, 1965 (W. Braune, W. Bamberg, D. Schmidt, R. Ring, K. Knabe). C. Fizet, *Helv. Chim. Acta* **69** (1986) 404.
1031. Hoffmann-La Roche, DE-AS 1 568 755, 1972 (S. F. Schaerer).
1032. J. Paust, S. Pfohl, W. Reif, W. Schmidt, *Justus Liebigs Ann. Chem.* 1978, 1024.
1033. Daiichi Seiyaku K.K., JP-AS 22 220, 1972. Kyowa Gas Kakaku Kogyo K.K., JP-OS 79 263, 1979 (K. Hazama).
1034. Daiichi Seiyaku Co., DE-AS 1 948 368, 1973 (S. Nabeta, M. Inagaki).
1035. Fuji Chem. Ind., FR 1 522 111, 1967. Jenapharm, DE-AS 2 350 610, 1975 (J. Schmidt, C. Weigelt, W. Bamberg, W. Scheider).
1036. E. T. Stiller et al., *J. Am. Chem. Soc.* **62** (1940) 1785. Nissan Chem. Ind. K.K., JP-OS 19 780, 1985 (K. Arai, Y. Obara, H. Matsumoto, S. Tsuchiya).
1037. Takeda Chem. Ind., JP-OS 95 085 (T. Kaneko, H. Sumiya, K. Kashiwa); JP-OS 95 086 (N. Kuroda, K. Kashiwa); JP-OS 95 087, 1992 (N. Kuroda, K. Kashiwa). Fuji Pharm. Ind. K.K., JP-OS 208 488, 1989 (H. Sato, S. Takeda, M. Yuya). Seitetsu Chem. Ind. K.K., JP-OS 106 479, 1984; JP 27 882, 1984 (K. Kawamura, S. Akutsu, M. Takahashi, J. Oonishi). BASF, DE 3 242 560, 1982 (J. Paust, H. Leininger). Sagami Chem. Res. Centre, JP-OS 126 880, 1983 (M. Matsumoto, S. Itou); JP-OS 88 257, 1979 (J. Oshima). Hoffmann-La Roche, EP 86 322, 1982 (P. Nösberger); EP 6156, 1978 (M. Schmid).
1038. O. Nagase, Y. Hosogawa, M. Shimizu, *Chem. Pharm. Bull.* **17** (1969) 398.
1039. C. Fizet, *Helv. Chim. Acta* **65** (1982) 2024. Hoffmann-La Roche, EP 50 721, 1980 (C. Fizet). Seitetsu Chem. Ind. K.K., JP-OS 227 875, 1983 (M. Kawamura et al.). S. Shimizu et al., *Agric. Biol. Chem.* **51** (1987) 289. BASF, DE 3 229 026, 1982 (K. Halbritter, M. Eggersdorfer).
1040. Syntex Corp., US 3 850 750, 1974 (R. Lanziolotta). Takeda Yakuhin Kogyo K.K., JP-AS 6369, 1971. T. Wieland, *Chem. Ber.* **81** (1948) 323.
1041. Fuji Yakuhin Kogyo, EP 336 123, 1988 (K. Achiwa, H. Takeda). H. Takahashi, T. Morimoto, K. Achiwa, *Chem. Lett.* 1987, no. 5, 855. T. Morimoto et al., *Chem. Lett.* 1986, no. 12, 2061. H. Takahashi, et al., *Tetrahedron Lett.* **27** (1986) 4477. Hoffmann-La Roche, EP 158 875, 1984 (E. A. Broger, Y. Crameri); EP 218 970, 1985 (E. A. Broger, Y. Crameri). K. Achiwa, T. Kogure, J. Ojima, *Chem. Lett.* 1978, 297.
1042. Seitetsu Chem. Ind. K.K., JP-OS 25 690, 1984; JP-OS 130 192, 1984; JP-OS 98 695, 1984 (H. Yamada, A. Shimizu, H. Hato). M. Kataoka, S. Shimizu, H. Yamada, *Agric. Biol. Chem.* **54** (1990) no. 1, 177. H. Hata, S. Shimizu, H. Yamada, *Agric. Biol. Chem.* **51** (1987) 3011; **48** (1984) 2285. Fuji Pharm. Ind. K.K., JP-OS 100 692, 1990 (H. Yamada et al.). Takeda Chem. Ind., JP-OS 271 789, 1992 (Y. Hikichi et al.); JP-OS 4 234 995, 1992 (T. Moriya et al.).
1043. S. Shimizu, S. Hattori, H. Hata, H. Yamada, *Enzyme Microbiol. Technol.* **9** (1987) no. 7, 411.
1044. J. D. Hulse, *J. Biol. Chem.* **253** (1978) 1654. Jenapharm, DE 2 223 236, 1971 (J. Schmidt et al.).
1045. F. Poppelsdorf, R. C. Lemon, *J. Org. Chem.* **26** (1961) 262.
1046. Tokyo Fine Chem. Co., DE-AS 2 232 090, 1972 (H. Uesugi, T. Takeda). BASF, DE-OS 1 939 710, 1971 (H. Leitner, N. Kutepow).
1047. Mitsui Toatsu Chem. Inc., JP-OS 100 295, 1987 (S. Tawaki et al.). Nitto Chem. Ind. K.K., JP-OS 201 992, 1982 (Y. Takayanagi, J. Watanabe). Mitsubishi Rayon K.K., EP 187 680, 1985 (K. Enomoto et al.).
1048. Musashino Chem. Res. Inst., EP 201 925, 1985 (K. Kishida, T. Tanabe). Tokuyama Soda Co., JP-OS 89 488, 1982 (T. Sada, T. Deo). Dow Chemical Co., US 4 589 963, 1984 (R. A. Cipriano, B. R. Ezzell).
1049. M. Inagaki, H. Tukamoto, O. Akazawa, *Chem. Pharm. Bull.* **24** (1976) 3097. Lederle Lab., GB 571 915, 1945. Nopko Chem., US 3 150 175, 1964 (R. Griffith). Hoffmann-La Roche, CH 244 468, 1947.
1050. Daiichi Seyaku Co., DE-AS 1 593 817, 1976 (N. Okuda, J. Koniyoshi, H. Tsukamoto).
1051. Daiichi Seyaku Co., DE-OS 1 593 832, 1967 (N. Okuda et al.).
1052. Hoffmann-La Roche, DE-OS 2 919 515, 1979 (D. Bartoldus, E. Broger).

1053. G. S. Kozlowa, T. J. Erokhina, V. J. Gunar, *Pharm. Chem. J. (Engl. Transl.)* **11** (1977) 505.
1054. F. Lipmann, *Science (Washington D.C. 1883)* **120** (1954) 855.
1055. P. W. Majerus, P. R. Vagelos, *Adv. Lipid Res.* **5** (1967) 1. F. Lynen, *Biochem. J.* **102** (1967) 381. H. Kleinkauf, W. Gevers, R. Rokoski, Jr., F. Lipmann, *Biochem. Biophys. Res. Commun.* **41** (1970) 1218. C. C. Gilhuus-Moe et al., *FEBS Lett.* **7** (1970) 287.
1056. W. Friedrich: *Handbuch der Vitamine,* Urban & Schwarzenberg, München 1987. D. K. Fenstermacher, R. C. Rose, *Am. J. Physiol.* **250** (1986) 155 – 160.
1057. E. Höhne: *Vitamine,* Otto Hoffmanns Verlag, München 1985, pp. 63 – 69.
1058. J. G. Moffatt, M. G. Khorana, *J. Am. Chem. Soc.* **81** (1959) 1265. M. Shimiza, *Method. Chim.* **11** (1977) 73.
1059. R. K. Hill, T. M. Chan, *Biochem. Biophys. Res. Commun.* **38** (1970) 181.
1060. G. M. Brown, *Compr. Biochem.* 1971, no. 21, 73.
1061. O. Halverson, S. Skrede, *Anal. Biochem.* **107** (1980) 103.
1062. W. L. Williams, E. Hoff-Joergensen, E. E. Snell, *J. Biol. Chem.* **177** (1949) 933.
1063. E. E. Snell et al., *J. Am. Chem. Soc.* **72** (1950) 5349.
1064. M. Shimizu et al., *Chem. Pharm. Bull.* **13** (1965) 180. Sogo Pharm. Co., JP 3342, 1980.
1065. W. B. Bean, R. E. Hodges, *Proc. Soc. Exp. Biol. Med.* **86** (1954) 693. R. E. Hodges, W. B. Bean, M. A. Ohlson, R. E. Bleiler, *J. Clin. Invest.* **38** (1959) 1421.
1066. R. M. Herman, F. B. Stifel, H. L. Greene, *Sci. Pract. Clin. Med.* **1** (1976) 386. L. D. Wright, *Present Knowl. Nutr. 4th Ed.* 1976, 226. A. Fidanza, *Acta Vitaminol. Enzymol.* **31** (1977) 85.
1067. R. E. Hodges, M. A. Ohlson, W. B. Bean, *J. Clin. Invest.* **37** (1958) 1642.
1068. K. Pietrzik: "Untersuchungen zur Ermittlung des Versorgungszustandes und des Bedarfs an Pantothensäure," Habilitationsschrift, Landwirtschaftliche Fakultät der Rhein.-Friedrich-Wilhelm-Universität, Bonn 1977
1069. R. W. Winters, R. B. Schultz, W. A. Krehl, *Endocrinology (Baltimore)* **50** (1952) 377, 388.
1070. M. Kirchgessner, D. A. Roth-Maier, *NMV Not. Med. Vet.* 1977, 135; *VMR Vet. Med. Rev.* 1977, 135.
1071. J. C. Bauernfeind, L. C. Norris, *Science (Washington D.C. 1883)* **89** (1939) 416.
1072. P. H. Philipps, R. W. Engel, *J. Nutr.* **18** (1939) 227.
1073. G. Fränkel, M. Blewett, *Nature (London)* **157** (1946) 697.
1074. J. Bonner, G. Axtmann, *Proc. Natl. Acad. Sci.* **23** (1937) 453.
1075. BASF, Tierernährung: Zum Thema Futterzusatzstoffe, technical report, Ludwigshafen 1991, pp. 46 – 49.
1076. C. S. Sodano in: *Vitamins, Synthesis, Production and Use,* Noyes Data Corp., Park Ridge, N.J. 1978, p. 237.
1077. O. D. Bird, R. Q. Thompson: *The Vitamins,* **vol. 7,** Academic Press, New York 1967, p. 209.
1078. E. G. Wollish, M. Schmall, *Anal. Chem.* **22** (1950) 1033.
1079. C. R. Szalkowski, J. H. Davidson, Jr., *Anal. Chem.* **25** (1953) 1192.
1080. M. Schmall, E. G. Wollish, *Anal. Chem.* **29** (1957) 1509.
1081. E. Illner, *Pharmazie* **35** (1980) 186.
1082. D. R. Wilken, R. E. Dyar, *Anal. Biochem.* **112** (1981) 9. D. Schmalzing et al., *J. High Resolut. Chromatogr.* **15** (1992) no. 11, 723. H. Brunner, S. Forster, *Monatsh. Chem.* **123** (1992) nos. 6 – 7, 659. W. Keim, A. Koehnes, W. Meltzow, H. Roemer, *J. High Resolut. Chromatogr.* **14** (1991) no. 8, 507. Advanced Separation Technologies, Inc., US 4 948 395, 1990 (D. W. Armstrong). A. Takasu, K. Ohya, *J. Chromatogr.* **389** (1987) 251. W. Koenig, U. Sturm, *J. Chromatogr.* **328** (1985) 357.
1083. S. Nag, S. K. Das, *J. AOAC Int.* **75** (1992) no. 5, 898. H. Thielemann, *Pharmazie* **36** (1981) 783.
1084. M. H. Bui-Nguyen, *J. Chromatogr.* **303** (1984) 291.
1085. W. Koenig, U. Sturm, *J. Chromatogr.* **328** (1985) 357. A. Takasu, K. Ohya, *J. Chromatogr.* **389** (1987) 251.
1086. O. Solberg, I. K. Hegna, *Methods Enzymol.* **62D** (1974) 201.
1087. G. D. Novelli, N. O. Kaplan, F. Lipmann, *J. Biol. Chem.* **177** (1949) 97. L. Johnston, L. Vaughan, H. M. Fox, *Am. J. Clin. Nutr.* **34** (1981) 2205.
1088. O. D. Bird, L. McCready, *Anal. Chem.* **30** (1958) 2045. H. E. Sauberlich in R. S. Goodhart, M. E. Shils (eds.): *Modern Nutrition in Health and Disease,* 5th ed., Lea and Fabiger, Philadelphia 1973, p. 203.

1089. R. K. Airas, *Anal. Biochem.* **134** (1983) 122. V. Srinivasan, N. Christensen, B. W. Wyse, R. G. Hansen, *Am. J. Clin. Nutr.* **34** (1981) 1736. D. K. Reibel, B. W. Wyse, D. A. Berkich, J. R. Neely, *J. Nutr.* **112** (1982) 1144.
1090. W. O. Sang et al., *Nutr. Res. (N.Y.)* **10** (1990) no. 4, 439. H. C. Morris, P. M. Finglas, R. M. Faulks, M. R. Morgan, *J. Micronutr. Anal.* **4** (1988) no. 1, 33. P. M. Finglas et al., *J. Micronutr. Anal.* **4** (1988) no. 1, 47.
1091. C. J. Bliss, O. D. Bird, R. Q. Thompson in P. György, W. N. Pearson (eds.): *The Vitamins,* **vol. VII,** Academic Press, New York 1967, p. 237.
1092. E. Negwer: *Organisch-chemische Arzneimittel und ihre Synonyme,* 5th ed., Akademie Verlag, Berlin 1978.
1093. S. H. Rubin, C. Magid, J. Scheiner, *Drug. Cosmet. Ind.* **86** (1960) 107. Hoffmann-La Roche, US 2 983 650, 1961 (S. H. Rubin).

General References
1094. J. P. Bonjour: "Biotin," Food Sci. Technol. (Handb. Vitam.), 2nd ed., vol. 2, 1991, 393.
1095. K. Dakshinamurti, H. N. Bhagavan (eds.), *Ann. N.Y. Acad. Sci.* **447** (1985) p. 1.
1096. M. Wilchek, E. A. Bayer (eds.): *Methods in Enzymology: Avidin-Biotin Technology,* vol. 184, Academic Press Inc., San Diego 1990.
1097. S. P. Mistry in B. M. Barker, D. A. Bender (eds.): *Vitamins in Medicine,* **vol. 1,** W. Heinemann Medical Books Ltd., London 1980, p. 381.
1098. M. Gaudry, P. Ploux: "Biotin," *Chromatogr. Sci.* **60** (1992) 441 – 467.
1099. *Kirk-Othmer,* **24,** 41.

Specific References
1100. E. Wildiers, *Cell. Rec. Cytol. Histol.* **18** (1901) 313.
1101. F. Kögl, *Ber. Dtsch. Chem. Ges. A* **68** (1935) 16 – 28.
1102. F. Kögl, B. Tönnis, *Hoppe Seyler's Z. Physiol. Chem.* **242** (1936) 43 – 73.
1103. P. György, D. B. Melville, D. Burk, *Science (Washington D.C. 1883)* **91** (1940) 243 – 245.
1104. P. György, *Z. Ärztl. Fortbild.* **28** (1931) 377 – 380.
1105. F. Kögl, L. Pons, *Hoppe Seyler's Z. Physiol. Chem.* **269** (1941) 61 – 80.
1106. F. E. Allison, S. R. Hoover, *Science (Washington D.C. 1883)* **78** (1933) 217 – 218.
1107. D. B. Melville, A. W. Mayer, K. Hoffmann, V. Du Vigneaud, *J. Biol. Chem.* **146** (1942) 487 – 492.
1108. V. Du Vigneaud, *Science (Washington D.C. 1883)* **96** (1942) 455 – 461.
1109. S. A. Harris, D. E. Wolf, R. Mozingo, K. Folkers, *Science (Washington D.C. 1883)* **97** (1943) 447 –448.
1110. Hoffmann-La Roche, US 2 489 232, 1949 (M. W. Goldberg, L. H. Sternbach).
1111. Hoffmann-La Roche, US 2 489 235, 1949 (M. W. Goldberg, L. H. Sternbach).
1112. Hoffmann-La Roche, US 2 489 238, 1949 (M. W. Goldberg, L. H. Sternbach).
1113. V. Du Vigneaud, K. Hofmann, D. B. Melville, J. R. Rachele, *J. Biol. Chem.* **140** (1941) 763 – 766.
1114. W. Traub, *Nature (London)* **178** (1956) 649 – 650.
1115. G. T. DeTitta, J. W. Edmonds, W. Stallings, J. Donohue, *J. Am. Chem. Soc.* **98** (1976) 1920 – 1926.
1116. E. Pinn et al., *Eur. J. Biochem.* **123** (1982) 545 – 546.
1117. G. T. DeTitta, J. W. Edmonds, W. Stallings, J. Donohue, *J. Am. Chem. Soc.* **98** (1976) 4983 – 4990.
1118. K. H. Baggaley, B. Blessington, C. P. Falshaw, W. D. Ollis, *J. Chem. Soc. Chem. Commun.* 1969, 101 – 102.
1119. L. H. Leonian, V. G. Lilly, *Science (Washington D.C. 1883)* **95** (1942) 658.
1120. B. M. Pearson, D. A. MacKenzie, M. H. J. Keenan, *Lett. Appl. Microbiol.* **2** (1986) 25 – 28.
1121. B. C. J. G. Knight, *Vitam. Horm. (N.Y.)* **3** (1945) 105 – 228.
1122. K. Watanabe, S. I. Yano, Y. Yamada, *Phytochemistry* **21** (1982) 513 – 516.
1123. W. Friedrich: *Handbuch der Vitamine,* Urban & Schwarzenberg, München 1987, pp. 486 – 519.
1124. D. M. Mock, *Contemp. Issues Clin. Nutr.* **15** (1992) 211 – 222.
1125. G. S. Heard, J. R. S. McVoy, B. Wolf, *Clin. Chem. (Winston* -Salem N.C.) 30 (1984) 125 – 127.
1126. D. V. Craft, N. H. Gross, N. Chandramouli, H. G. Wodd, *Biochemistry* **24** (1985) 2471 – 2476.
1127. Y. Izumi, K. Ogata, *Adv. Appl. Microbiol.* **22** (1975) 145 – 176.
1128. S. W. Brown, K. Kamogawa, *Biotech. Gen. Eng. Rev.* **9** (M. P. Thomas, Ed.), Intercept, Andover, 1991, 295 – 326.

1129. D. A. Trainer, R. J. Parry, A. Guttermann, *J. Am. Chem. Soc.* **102** (1980) 1467 – 1468.
1130. A. Marquet et al., *J. Am. Chem. Soc.* **115** (1993) 2139 – 2145.
1131. M. A. Eisenberg, *Ann. N.Y. Acad. Sci.* **447** (1985) 335 – 349.
1132. R. Gloeckler et al., *Gene* **87** (1990) 63 – 70.
1133. M. Ohsugi et al., *Agric. Biol. Chem.* **47** (1983) 2725 – 2730.
1134. J. R. Knowles, *Annu. Rev. Biochem.* **58** (1989) 195 –221.
1135. K. Dakshinamurti, J. Chauhan, *Vitam. Horm. (N.Y.)* **45** (1989) 337 – 384.
1136. D. Samols et al., *J. Biol. Chem.* **263** (1988) 6461 –6464.
1137. P. György, R. Kuhn, E. Lederer, *J. Biol. Chem.* **131** (1939) 745 – 759.
1138. D. B. Melville, K. Hofmann, E. Hague, V. Du Vigneaud, *J. Biol. Chem.* **142** (1942) 615 – 618.
1139. M. Masuda et al., *Process Biochem.* **30** (1995) no. 6, 553 – 562.
1140. Hoffmann-La Roche, CH 556 867, 1969 (M. Gerecke, J. P. Zimmermann).
1141. Hoffmann-La Roche, CH 2458, 1984 (H. Pauling, C. Wehrli).
1142. Sumitomo Chemical Co., JP 95 818, 1980 (M. Hazama, T. Aratani, G. Suzukamo, T. Takahashi).
1143. E. Merck, EP 0 226 043, 1985 (M. Casutt, E. Poetsch).
1144. Sumitomo Chemical Co., JP 7253, 1982 (N. Ohashi, T. Ikeda, K. Shimago, T. Takahashi, K. Ishizumi).
1145. E. Merck, EP 0 502 392, 1991 (M. Schwarz, M. Casutt, J. Eckstein).
1146. H. L. Lee, E. G. Baggiolini, M. R. Uskokovic, *Tetrahedron* **43** (1987) 4887 – 4903.
1147. E. Merck, EP 0 242 686, 1986 (E. Poetsch, M. Casutt).
1148. F. D. Deroose, P. J. De Clercq, *Tetrahedron Lett.* **34** (1993) 4365 – 4368.
1149. E. Merck, WO 91/02734, 1989 (W. N. Speckamp, M. Casutt, E. Poetsch).
1150. E. J. Corey, M. M. Mehrotra, *Tetrahedron Lett.* **29** (1988) 57 – 60.
1151. T. Ravindranathan, S. V. Hiremath, D. Rajagopala Reddy, A. V. Rama Rao, *Carbohydr. Res.* **134** (1984) 332 – 336.
1152. T. Ogawa, T. Kawano, M. Matsui, *Carbohydr. Res.* **57** (1977) C31 – C35.
1153. H. Ohrui, N. Sueda, S. Emoto, *Agric. Biol. Chem.* **42** (1978) 865 – 868.
1154. H. Ohrui, S. Emoto, *Tetrahedron Lett.* **32** (1975) 2765 – 2766.
1155. R. R. Schmidt, M. Maier, *Synthesis* 1982, 747 – 748.
1156. F. G. M. Vogel, J. Paust, A. Nürrenbach, *Liebigs Ann. Chem.* 1980, 1972 – 1977.
1157. F. G. M. Vogel et al., *Tetrahedron Lett.* **22** (1981) 3493 – 3496.
1158. A. Fliri, K. Hohenlohe-Oehringen, *Chem. Ber.* **113** (1980) 607 – 613.
1159. P. N. Confalone, G. Pizzolato, D. Lollar-Confalone, M. R. Uskokovic, *J. Am. Chem. Soc.* **102** (1980) 1954 – 1960.
1160. H. J. Manteiro, *An. Acad. Bras. Cienc.* **52** (1980) 493 – 501.
1161. R. A. Whitney, *Can. J. Chem.* **61** (1983) 1158 –1160.
1162. S. Bory et al., *J. Am. Chem. Soc.* **100** (1978) 1558 –1563.
1163. P. N. Confalone, G. Pizzolato, M. R. Uskokovic, *J. Org. Chem.* **42** (1977) 135 – 139.
1164. M. Marx et al., *J. Am. Chem. Soc.* **99** (1977) 6754 –6755.
1165. J. Vasilevskis et al., *J. Am. Chem. Soc.* **100** (1978) 7423 – 7424.
1166. R. A. Volkmann, J. T. Davis, C. N. Meltz, *J. Am. Chem. Soc.* **105** (1983) 5946 – 5948.
1167. A. V. Rama Rao, T. Ravindranathan, S. V. Hiremath, D. Rajagopala Reddy, *Indian J. Chem. Sect. B* **22B** (1983) 419 – 420.
1168. J. M. Schreiner, *Ann. N.Y. Acad. Sci.* **447** (1985) 420 – 422.
1169. R. C. Rose, *Annu. Rev. Physiol.* **42** (1980) 157 –171.
1170. F. A. Hommes, *World Rev. Nutr. Diet.* **48** (1986) 34 – 84.
1171. A. Piffeteau, M. Gaudry, *Biochim. Biophys. Acta* **816** (1985) 77 – 82.
1172. J. P. Bonjour, *Ann. N.Y. Acad. Sci.* **447** (1985) 97 –104.
1173. J. Chauhan, K. Dakshinamurti, *Biochem. J.* **256** (1988) 265 – 270.
1174. J. M. Westendorf, D. D. McCormick, *Int. J. Vitam. Nutr. Res.* **50** (1980) 150 – 155.
1175. J. P. Bonjour, *Int. J. Vitam. Nutr. Res.* **47** (1977) 107 – 118.
1176. S. P. Mistry in B. M. Barker, D. A. Bender (eds.): *Vitamins in Medicine,* **vol. 1,** William Heinemann Medical Books Ltd., London 1980, pp. 381 – 397.
1177. K. S. Roth, *Am. J. Clin. Nutr.* **34** (1981) 1967 –1974.
1178. L. Sweetman et al., *Pediatrics* **68** (1981) 553.
1179. K. H. Krause, W. Kochen, P. Berlit, J. P. Bonjour, *Int. J. Vitam. Nutr. Res.* **54** (1984) 217 – 222.

1180. W. L. Nyhan, *Int. J. Biochem.* **20** (1983) 363 – 370.
1181. D. M. Mock et al., *Ann. N.Y. Acad. Sci.* **447** (1985) 314 – 334.
1182. L. Salmenpera, J. Perheentupa, J. Pispa, M. A. Simes, *Int. J. Vitam. Nutr. Res.* **55** (1985) 281 –285.
1183. V. P. Sydensticker et al., *JAMA J. Am. Med. Assoc.* **118** (1942) 1199 – 1200.
1184. G. S. Heard, R. L. Hood, A. R. Johnson, *Med. J. Aust.* 1983, no. 2, 305 – 306.
1185. K. Dakshinamurti, S. P. Mistry, *Arch. Biochem. Biophys.* **99** (1962) 254 – 257.
1186. R. Bitsch, A. Dersi, D. Höltzel, *Nutr. Metab.* **21** (1977) Suppl. 1, 27.
1187. L. Völker, *Übers. Tierernähr.* **5** (1977) 185 – 220.
1188. E. T. Kornegay, *Ann. N.Y. Acad. Sci.* **447** (1985) 112 – 121.
1189. C. C. Whitehead, C. J. Randall, *Br. J. Nutr.* **48** (1982) 177 – 184.
1190. D. B. McCormick, L. D. Wright, *Compr. Biochem.* **21** (1971) 81 – 110.
1191. Y. Izumi et al., *Agric. Biol. Chem.* **45** (1981) 1983 –1989.
1192. C. Lindblow-Kull, F. J. Kull, A. Shrift, *Biochem. Biophys. Res. Commun.* **93** (1980) 572 – 576.
1193. A. Piffeteau et al., *Biochemistry* **19** (1980) 3069 –3073.
1194. T. D. Tran, *Kraftfutter* **9** (1981) 453 – 455.
1195. J. P. Bonjour, *Food Sci. Technol.* **40** (1991) 393 –436.
1196. R. B. Ferguson, H. C. Lichstein, *J. Bacteriol.* **75** (1958) 366.
1197. E. E. Snell, R. E. Eakin, R. J. Williams, *J. Am. Chem. Soc.* **62** (1940) 175 – 178.
1198. H. R. Skeggs in F. Kavanagh (ed.): *Analytical Microbiology,* Academic Press, New York 1963, p. 421.
1199. Konica Corp., JO 1 203 975, 1989 (A. Onishi et al.).
1200. D. M. Mock, D. B. Dubois, *Anal. Biochem.* **153** (1986) 272 – 278.
1201. D. L. Mock et al., *Anal. Biochem.* **151** (1985) 178 –181.
1202. E. J. Williams, A. K. Campbell, *Anal. Biochem.* **155** (1986) 249 – 255.
1203. K. Ogata et al., *Agric. Biol. Chem.* **29** (1965) 895 –901.
1204. Mitsubishi Petrochem. Co., JO 5 078 369, 1993; JO 5 078 370, 1993 (M. Terasawa et al.).
1205. T. Yoshida et al, *J. Chromatogr.* **456** (1988) 421 –426.
1206. E. P. Diamandis, T. K. Christopoulos, *Clin. Chem. (Winston Salem N.C.)* **37** (1991) 625 – 636.
1207. M. Wilchek, E. A. Bayer, *Immunol. Today* **5** (1984) 39 – 43.
1208. G. Paganelli, M. Malcovati, F. Fazio, *Nucl. Med. Commun.* **12** (1991) 211 – 234.
1209. E. A. Bayer, M. Wilchek, *Methods Enzymol.* **184** (1990) 467 – 469.
1210. E. A. Bayer, M. Wilchek, *J. Chromatogr.* **510** (1990) 3 – 11.
1211. P. R. Langer, A. A. Waldrop, D. C. Ward, *Proc. Natl. Acad. Sci. USA* **78** (1981) 6633 – 6637.
1212. W. F. Körner, J. Völlm in H. P. Kuemerle, E. R. Garrett, K. H. Spitzy (eds.): *Klinische Pharmakologie und Pharmakotherapie,* Urban & Schwarzenberg, München 1976, p. 381.
1213. D. R. Miller, K. C. Hayes in J. N. Hathock (ed.): *Nutritional Toxicity,* **vol. 1,** Academic Press, New York 1982, p. 81.
1214. E. Merck, Sicherheitsdatenblatt, Darmstadt 1992. References **for Chapter 13**
1215. IUPAC-IUB Joint Commission on Biochemcial Nomenclature (JCBN): "Nomenclature and Symbols for Folic Acid and Related Compounds," Recommendations 1986, *Eur. J. Biochem.* **168** (1987) 251 – 253; *Pure Appl. Chem.* **59** (1987) 834 – 836.
1216. L. Wills, P. W. Clutterbuck, P. D. F. Evans, *Biochem. J.* **31** (1937) 2136 – 2147.
1217. A. G. Hogan, E. M. Parrott, *J. Biol. Chem.* **132** (1940) 507 – 517.
1218. E. E. Snell, W. H. Peterson, *J. Bacteriol.* **39** (1940) 273 – 285.
1219. H. K. Mitchell, E. E. Snell, R. J. Williams, *J. Am. Chem. Soc.* **63** (1941) 2284.
1220. S. B. Binkley et al., *Science (Washington D.C. 1883)* **100** (1944) 36 – 37.
1221. J. J. Pfiffner, D. G. Calkins, E. S. Bloom, L. B. O'Dell, *J. Am. Chem. Soc.* **68** (1946) 1392.
1222. R. B. Angier et al., *Science (Washington D.C. 1883)* **103** (1946) 667 – 669.
1223. A. D. Welch, *Perspect. Biol. Med.* **27** (1983) 64 – 75.
1224. R. L. Blakley, S. J. Benkovic (eds.): *Folates and Pterins,* **vols. 1 and 2,** J. Wiley & Sons, New York 1984/1985.
1225. O. Isler, G. Brubacher in O. Isler, G. Brubacher, S. Gishla, B. Kräutler (eds.): *Vitamine II,* Thieme Verlag, Stuttgart – New York 1988, pp. 264 – 308.
1226. T. Maruyama, T. Shiota, C. L. Krumdieck, *Anal. Biochem.* **84** (1978) 277 – 295.

1227. L. B. Bailey (ed.): *Folate in Health and Disease,* M. Dekker Inc., New York – Basel – Hong Kong 1995.
1228. P. A. Charlton et al., *J. Chem. Soc., Perkin Trans. 1* 1985, 1349 – 1353.
1229. C. M. Tatum, M. G. Fernald, J. P. Schimel, *Anal. Biochem.* **103** (1980) 255 – 257.
1230. J. A. Blair, K. J. Saunders, *Anal. Biochem.* **34** (1970) 376 – 381.
1231. C. Zarow, A. M. Pellino, P. V. Danenberg, *Prep. Biochem.* **12** (1983) 381 – 393.
1232. J. E. Baggott, C. L. Krumdieck, *Biochemistry* **18** (1979) 1036 – 1041.
1233. E. Khalifa, A. N. Ganguly, J. H. Bieri, M. Viscontini, *Helv. Chim. Acta* **63** (1980) 2554 – 2558.
1234. P. Stover, V. Schirch, *Anal. Biochem.* **202** (1992) 82 – 88.
1235. T. Eguchi et al., *Biosci. Biotech. Biochem.* **56** (1992) 701 – 703.
1236. D. B. Cosulich, J. M. Smith, H. P. Broquist, *J. Am. Chem. Soc.* **74** (1952) 4215 – 4216.
1237. W. Friedrich: *Handbuch der Vitamine,* Urban & Schwarzenberg, München – Wien – Baltimore 1987, pp. 398 – 485.
1238. B. Bötticher, R. Kluthe in K. Pietrzik (ed.): *Folsäure-Mangel,* W. Zuckschwerdt Verlag, München 1987, pp. 15 – 24.
1239. T. Brody in L. J. Machlin (ed.): *Handbook of Vitamins,* M. Dekker Inc., New York – Basel – Hong Kong 1991, pp. 453 – 489.
1240. L. B. Bailey in L. B. Bailey (ed.): *Folate in Health and Disease,* M. Dekker Inc., New York – Basel – Hong Kong 1995, pp. 123 – 151.
1241. K. H. Bässler, E. Grühn, D. Loew, K. Pietrzik: *Vitamin-Lexikon,* G. Fischer Verlag, Stuttgart –Jena – New York 1992, pp. 127 – 154.
1242. J. Leichter, A. F. Landymore, C. L. Krumdieck, *Am. J. Clin. Nutr.* **32** (1979) 92 – 95.
1243. S. P. Rothenberg, M. P. Iqbal, M. Da Costa, *Anal. Biochem.* **103** (1980) 152 – 156.
1244. J. F. Gregory, *Adv. Food Nutr. Res.* **33** (1989) 1 – 101.
1245. H. E. Sauberlich et al., *Am. J. Clin. Nutr.* **46** (1987) 1016 – 1028.
1246. J. F. Gregory in L. B. Bailey (ed.): *Folate in Health and Disease,* M. Dekker Inc., New York – Basel – Hong Kong 1995, pp. 195 – 235.
1247. G. M. Brown, H. Williamson, *Adv. Enzymol. Relat. Areas Mol. Biol.* **53** (1982) 345 – 381.
1248. G. M. Brown, H. Williamson in F. C. Neidhardt (ed.): *Escherichia coli and Salmonella typhymurium,* **vol. 1,** American Society for Microbiology, Washington, D.C. 1987, pp. 521 – 538.
1249. A. de Saizieu, P. Vankan, A. P. G. M. van Loon, *Biochem. J.* **306** (1995) 371 – 377.
1250. Y. Suzuki, G. M. Brown, *J. Biol. Chem.* **249** (1974) 2405 – 2410.
1251. J. Toy, A. L. Bognar, *J. Biol. Chem.* **265** (1990) 2492 – 2499.
1252. J. Slock et al., *J. Bacteriol.* **172** (1990) 7211 – 7226.
1253. S. A. Lacks, B. Greenberg, P. Lopez, *J. Bacteriol.* **177** (1995) 66 – 74.
1254. P. Babitzke, P. Gollnick, C. Yanofsky, *J. Bacteriol.* **174** (1992) 2059 – 2064.
1255. P. S. Margolis, A. Dirks, R. Losick, *J. Bacteriol.* **175** (1993) 528 – 540.
1256. C. W. Waller et al., *J. Am. Chem. Soc.* **70** (1948) 19 – 22.
1257. F. Weygand, V. Schmied-Kowarzik, *Chem. Ber.* **82** (1949) 333 – 336.
1258. S. Uyeo, S. Mizukami, T. Kubota, S. Takagi, *J. Am. Chem. Soc.* **72** (1950) 5339 – 5340.
1259. P. Karrer, R. Schwyzer, *Helv. Chim. Acta* **31** (1948) 777 – 782.
1260. F. Hoffmann-La Roche, EP-A 608 693 A2, 1994 (C. Wehrli).
1261. M. Sletzinger et al., *J. Am. Chem. Soc.* **77** (1955) 6365 – 6367.
1262. E. C. Taylor, R. N. Henrie, R. C. Portnoy, *J. Org. Chem.* **43** (1978) 736 – 737.
1263. J. H. Bieri, M. Viscontini, *Helv. Chim. Acta* **56** (1973) 2905 – 2911. E. Khalifa, P. K. Sengupta, J. H. Bieri, M. Viscontini, *Helv. Chim. Acta* **59** (1976) 242 – 247.
1264. C. M. Baugh, C. L. Krumdieck, H. J. Baker, C. E. Butterworth, *J. Nutr.* **105** (1975) 80 – 89.
1265. T. T. Y. Wang, A. M. Reisenauer, C. H. Halsted, *J. Nutr.* **115** (1985) 814 – 819.
1266. A. M. Reisenauer, C. L. Krumdieck, C. H. Halsted, *Science (Washington D.C. 1883)* **198** (1977) 196 – 197.
1267. C. J. Chandler, T. T. Y. Wang, C. H. Halsted, *J. Biol. Chem.* **261** (1986) 928 – 933.
1268. J. Zimmerman, Z. Gilula, J. Selhub, I. H. Rosenberg, *Int. J. Vitam. Nutr. Res.* **59** (1989) 151 – 156.
1269. J. Zimmerman, J. Selhub, I. H. Rosenberg, *Am. J. Clin. Nutr.* **46** (1987) 518 – 522.
1270. J. Kiil, M. Jaegerstad, L. Elsborg, *Int. J. Vitam. Nutr. Res.* **49** (1979) 296 – 306.

1271. V. Herbert, K. C. Das in M. E. Shils, J. A. Olson, M. Shike (eds.): *Modern Nutrition in Health and Disease,* 8th ed., Lea & Fabiger, Philadelphia 1993, pp. 402 – 425.
1272. C. Wagner, *Annu. Rev. Nutr.* **2** (1982) 229 – 248.
1273. G. B. Henderson, *Annu. Rev. Nutr.* **10** (1990) 319 –335.
1274. J. J. McGuire, J. R. Bertino, *Mol. Cell. Biochem.* **38** (1981) 19 – 48.
1275. J. J. McGuire, J. K. Coward in R. L. Blakley, S. J. Benkovic (eds.): *Folates and Pterins,* **vol. 1,** J. Wiley & Sons, New York 1984, pp. 135 – 190.
1276. R. G. Moran, P. D. Colman, *Anal. Biochem.* **140** (1984) 326 – 342.
1277. C. Wagner in L. B. Bailey (ed.): *Folate in Health and Disease,* M. Dekker Inc., New York – Basel –Hong Kong 1995, pp. 23 – 42.
1278. V. Herbert, *Am. J. Clin. Nutr.* **21** (1968) 743 – 752.
1279. V. Herbert et al., *Trans. Assoc. Am. Physicians* **97** (1984) 161 – 171.
1280. *Recommended Dietary Allowances,* 10th ed., National Academy Press, Washington 1989.
1281. C. J. Allegra et al., *N. Engl. J. Med.* **317** (1987) 978 –985.
1282. J. Jolivet et al., *N. Engl. J. Med.* **309** (1983) 1094 –1104.
1283. W. E. Evans et al., *N. Engl. J. Med.* **314** (1986) 471 –477.
1284. D. C. Heimburger et al., *J. Am. Med. Assoc.* **259** (1988) 1525 – 1530.
1285. S. H. Mudd et al., *Biochem. Biophys. Res. Commun.* **46** (1972) 905 – 912.
1286. J. B. Ubbink, W. J. H. Vermaak, A. van der Merve, P. J. Becker, *Am. J. Clin. Nutr.* **57** (1993) 47 – 53.
1287. H. Refsum, P. M. Ueland, S. Kvinnsland, *Cancer Res.* **46** (1986) 5385 – 5391.
1288. E. H. Broxson et al., *Cancer Res.* **49** (1989) 5879 –5883.
1289. D. W. Jacobsen, V. J. Gatautis, R. Green, *Clin. Chem. (Winston Salem N.C.)* **40** (1994) 873 – 881.
1290. A. Andersson et al., *Eur. J. Clin. Invest.* **22** (1992) 79 – 87.
1291. J. Mulinare, J. F. Cordero, J. D. Erickson, R. J. Berry, *J. Am. Med. Assoc.* **260** (1988) 3141 – 3145.
1292. A. F. Czeisal, J. Dudas, *N. Engl. J. Med.* **327** (1992) 1832 – 1835.
1293. MRC Vitamin Study Research Group: "Prevention of Neural Tube Defects: Results of the Medical Research Council Vitamin Study," *Lancet* 1991, no. 338, 131 – 137.
1294. R. W. Smithells et al., *Lancet* 1983 vol. I, 1027 –1031.
1295. Centres of Disease Control and Prevention: "Recommendations for the Use of Folic Acid to Reduce the Number of Cases of Spina Bifida and Other Neural Tube Defects," *J. Am. Med. Soc.* **269** (1993) 1233 – 1238.
1296. Expert Advisory Group to U.K. Department of Health: *Folic Acid and the Prevention of Neural Tube Defects,* Department of Health, London 1992, pp. 1 – 33.
1297. C. E. Butterworth, T. Tamura, *Am. J. Clin. Nutr.* **50** (1989) 353 – 358.
1298. A. Hanck in: *Spektrum Vitamine,* vol. 42, Aesopus Verlag, Zug 1986, pp. 81 – 85.
1299. Nutrient Requirements of Poultry, 9th revised ed., NRC 1994, Nutrient Requirements of Swine, 9th revised ed. 1988, NRC, National Academy Press, Washington, D.C.
1300. Arbeitsgemeinschaft für Wirkstoffe in der Tierernährung e. V. (AWT): *Vitamine in der Tierernährung,* Bonn 1991.
1301. C. L. Krumdieck, T. Tamura, I. Eto, *Vitam. Horm. (N.Y.)* **40** (1983) 45 – 104.
1302. G. Brubacher, W. Müller-Mulot, D. A. T. Southgate: *Methods for Determination of Vitamins in Food,* Elsevier Appl. Sci. Publ., London – New York 1985, pp. 158 – 159.
1303. D. B. McCormick, L. D. Wright (eds.), *Methods Enzymol.* **66E** (1980) 429 – 727.
1304. L. V. Fenys, K. D. Thakker, V. D. Reif, L. T. Grady, *J. Pharm. Sci.* **71** (1982) 1242 – 1246.
1305. H. E. Sauberlich, *Annu. Rev. Nutr.* **4** (1984) 377 –407.
1306. S. D. Wilson, D. W. Horne, *Anal. Biochem.* **142** (1984) 529 – 535.
1307. S. A. Kashani, B. A. Cooper, *Anal. Biochem.* **146** (1985) 40 – 47.
1308. J. Kas, J. Cerna, *Methods Enzymol.* **66E** (1980) 443 – 451.
1309. American Institute of Nutrition, *J. Nutr.* **105** (1975) 134.
1310. M. Weiner, J. van Eys: "Nicotinic Acid (Nutrient –Cofactor – Drug)," *Clinical Pharmacology,* vol. 1, Marcel Dekker, New York 1983.
1311. L. M. Henderson, *Annu. Rev. Nutr.* **3** (1983) 289 – 307.
1312. D. Giesecke, *Uebers. Tierernähr.* **11** (1983) 133 – 154.
1313. K. F. Gey, L. A. Carlson (eds.): *Metabolic Effects of Nicotinic Acid and its Derivatives,* Hans Huber Verlag, Bern 1971.

1314. Lonza, CH 495 990, 1968 (A. Stocker, O. Marti, T. Pfammatter, G. Schreiner); CH 504 437, 1969 (A. Stocker, O. Marti, T. Pfammatter, G. Schreiner); CH 511 848, 1969 (A. Stocker, O. Marti, T. Pfammatter, G. Schreiner); CH 527 193, 1969 (A. Stocker, O. Marti, T. Pfammatter, G. Schreiner); CH-A 548 394, 1969 (A. Stocker, O. Marti, T. Pfammatter, G. Schreiner).
1315. Yuki Gosei, JP 45 013 572, 1970 (M. Takeishi, K. Fuji, T. Matsui); Koei Chemical, JP 46 041 545, 1971 (M. Imazu, Y. Higata); Degussa, DE 1 948 715, 1969 (T. Lüssling, H. Schaefer, W. Weigert); Degussa, DE 2039 497, 1970 (T. Lüssling, H. Schaefer, W. Weigert); Lonza, CH 582 151, 1973 (F. Moulin, K.-J. Boosen).
1316. Degussa, DE-OS 3 107 755, 1982 (H. Beschke, H. Friedrich, J. Heilos).
1317. Takeda, EP 0 037 123, 1981 (O. Isatsugu, F. Itsuo, M. Ichiro).
1318. Degussa, DE-OS 3 128 956, 1983 (H. Beschke, F. L. Dahm, H. Friedrich).
1319. Degussa, DE-OS 2 517 054, 1976 (H. Beschke, H. Friedrich, K.-P. Müller, G. Schreyer).
1320. Yuki Gosei, JP 74 127 976, 1973 (Y. Watabiki et al.)
1321. Lonza, WO 94/22 824, 1994 (J. Heveling, E. Armbruster, W. Siegrist).
1322. *DAB 10* (1991), Suppl. 1993.
1323. *United States Pharmacopoeia,* **vol. 23,** USP Convention, Inc., Princeton, N.J., 1990, p. 1080.
1324. *European Pharmacopoeia,* 2nd ed., Maisonneuve, Sainte Ruffine, France 1980, p. 47
1325. *United States Pharmacopoeia,* **vol. 23,** USP Convention, Inc., Princeton, N.J., 1990, p. 1082.
1326. *European Pharmacopoeia,* 2nd ed., Maisonneuve, Sainte Ruffine, France 1986, p. 459.
1327. K. Helrich (ed.), *Association of Analytical Chemists* : Official Methods of Analysis, 15th ed., Arlington, Virginia 1990, Microbiology 960.46 and 985.43.
1328. K. Helrich (ed.), *Association of Analytical Chemists* : Official Methods of Analysis, 15th ed., Arlington, Virginia 1990, no. 961.14.
1329. R. T. Nuttall, B. Bush, *Analyst (London)* **96** (1971) 875.
1330. G. W. Chase et al., *J. Micronutr. Anal.* **7** (1990) 15 – 25.
1331. J. Fontaine, J. Hörr, *Agribiol. Res.* **46** (1993) 10 – 19.
1332. National Academy of Sciences: *Recommended Dietary Allowances (RDA),* 10th ed., National Academy Press, Washington 1989, p. 285.
1333. C. L. Brooke, *J. Agric. Food. Chem.* **16** (1968) 163 – 167.
1334. R. H. Anderson et al., *Food Technol. (Chicago)* **30** (1976) 110.
1335. Lonza, company brochure, Niacin in Animal Nutrition, Basel 1991, 20 p.
1336. *Nutrient Requirements of Swine,* 9th rev. ed., National Academy Press, Washington, D.C., 1988. *Nutrient Requirements of Poultry,* 9th rev. ed., National Academy Press, Washington, D.C., 1994.
1337. Bundesanzeiger, no. 76, April 21, 1990.
1338. Lonza, company brochure, Niacin and Cholesterol, Basel 1990, 14 p.
1339. S. Sandler, *J. Clin. Endocrinol. Metab.* **77** (1993) no. 6, 1574 – 1576.
1340. Lonza, unpublished data.
1341. *Anticancer Res.* **2** (1982) 71 – 74.

Vodka → Spirits
Vulcanization → Rubber, 4.Chemicals
Vulcanization → Rubber, 5.Technology
Washing Agents → Laundry Detergents

Visualization Techniques in Process Engineering

Georg Petritsch, Institut für Verfahrenstechnik, Universität Hannover, Hannover, Federal Republic of Germany

Nicolas Reinecke, Mc Kinsey & Comp., Berlin, Federal Republic of Germany

Dieter Mewes, Institut für Verfahrenstechnik, Universität Hannover, Hannover, Federal Republic of Germany

1. Introduction 1
2. Optical Visualization Techniques . . . 1
3. Tomographic Visualization Techniques 2
3.1. Tomographic Measurement Systems . 3
3.2. Capacitance Tomography 8
3.3. Conductive Tomography 10
3.4. Optical Tomography 11
3.5. X-Ray Tomography 12
3.6. Example of the Application of Tomography in a Trickle-Bed Reactor . . . 13
4. Future Perspectives 15
5. References 16

1. Introduction

For the examination and control of a process, the process parameters must be measured without interfering with the process. Nevertheless, it is necessary to obtain as much information on the physical properties of the process as possible. Often the spatial distribution of a fluid, a temperature, or a concentration has an important impact on product quality and process efficiency. Local measurement techniques, can only measure the physical properties at certain locations, and often interference of the process by the measurement sensor can not be avoided. To obtain information on a complete cross section of an apparatus, numerous measurements must be taken by placing the measurement probe at different locations. Therefore, only time-averaged information is available, and this can lead to misinterpretations, especially for nonstationary flows. To overcome these intrinsic measurement problems, in process engineering there is a need for measurement techniques which provide high time and spatial resolution over a complete cross section of an apparatus without interfering with the process.

2. Optical Visualization Techniques

Numerous optical techniques have been used to obtain two-dimensional images of different kinds of flow. Their major advantages are nonintrusive measurement and high spatial and time resolution. However, their application is limited because the optical accessibility of the measurement object has to be guaranteed. Besides visible light, as in video or high-speed camera observations, nonvisible light is used for various optical measurement techniques. Details on optical measurement techniques for process engineering applications are described in [111].

In Raman spectroscopy the measurement plane is illuminated by monochromatic light. If a molecule is hit by a light photon it is lifted into a higher energy state. After a very short time, the molecule emits the photon and returns to its original state. If the emitted photon has a different energy to the absorbed photon, its wavelength shifts. Depending on the shift, different types of scattering (e.g., Stokes and anti-Stokes Raman scattering, Rayleigh scattering) can be observed. The analysis of the different types of scattering leads to the measurement of temperature or concentration fields.

The effect of thermal radiation is used in measurement techniques like thermography and pyrometry. Depending on its temperature, a body has a certain radiant emittance. This is measured by radiation detectors, which are suited to different radiation wavelengths. For calculating the temperature of a body, various influences on the measurement result have to be taken into account. These are the material and surface properties of the examined material, the Kirchhoff radiation laws, and the dependence of the radiation on the emission angle. Such measurement

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a27_b21

systems are used for the two-dimensional imaging of temperature fields.

Three-dimensional information can be collected by using holographic measurement techniques (see also → Holography). The object to be measured is illuminated by monochromatic light, which is scattered and reflected by the measurement object onto a photographic plate. At the same time, part of the monochromatic light is led directly from the light source to the photographic plate, so that it does not interact with the measurement object. On the photographic plate the two light beams, called the object and reference beams, form a stable interference pattern, which is stored on the plate. When the plate is illuminated by the reconstruction beam, which has the same physical properties as the reference beam, the measured object is reconstructed three dimensionally and can be viewed from different angles.

For the measurement of temperature or concentration profiles, the reference and the measurement beam are collected at different times. First the reference beam is led through the measurement plane and stored for the initial state of flow. After a change in temperature or concentration, and therefore a change in the refractive index of the flow, has occurred, the measurement beam is recorded. This causes an interference pattern on the photographic plate. Analysis of the patterns leads to the temperature or concentration field in the flow.

By taking holograms of an object from two directions, its position can be assessed by cutting the two holograms into slices and comparing them to each other. This kind of technique provides high spatial and temporal information on the process and has been successfully employed for measuring particle sizes and velocities in sprays and bubbly flows [123]. To measure the particle speed, two holograms are taken at short intervals. By comparing the two holograms, the paths of the particles and their velocity vectors can be measured.

For measuring two-dimensional velocity fields in flows of different types, particle image velocimetry (PIV) has become a valuable tool. From the velocity field, further information such as vorticity, pathlines, and Reynolds stresses can be derived. For imaging, the measurement volume is illuminated with a laser sheet for a short time. The illuminated sheet is photographed by a CCD camera and stored on a PC. This procedure is repeated after a certain time to obtain information on the change of flow. By processing the stored images (in the simplest way, by cross-correlation of two images) the velocity field in the illuminated part of the flow can be calculated. For the application of PIV in two-phase flow, special algorithms for the detection of gaseous or liquid particles have been developed. In addition, various algorithms for the calculation of the flow field are available. For an overview of PIV techniques, see [124, 125].

3. Tomographic Visualization Techniques

Tomographic measurement techniques have become increasingly popular in chemical and process engineering. With such techniques, fields of density, concentration, temperature, velocity, or local void fraction can be measured without interfering with the flow in the measurement plane. The measured fields are often nonstationary and may even have a transient character. Therefore, the tomographic measurement techniques must have high spatial and temporal resolution.

The term tomography originates from the Greek *tomos* (slice) and *graph* (image). Thus, goal of tomographic measurement is to generate a two-dimensional image of a cross section of the measurement object. The two-dimensional image contains information on the distribution of a physical property. For medical applications this is the texture of tissue or the water fraction therein. Tomographic measurements have two major advantages over other techniques. First, the measurements are conducted without interfering with the measurement object, and second, that the target quantity is measured simultaneously at many places in the measurement plane.

The application of tomographic measurement techniques in process and chemical engineering is, in comparison to the medical sector, relatively new. Developments in the design of sensors and in reconstruction techniques have been the subject of international conferences [2–10]. These symposia dealt with the measurement of density fields and the measurement of the void fraction field in single- and multiphase flows. Currently these techniques are mainly used for scientific and research applications, but there are

already potential applications in industry. For example, they are used for the measurement of the phase distribution in the exploitation of natural oil and gas.

In Figure 1 the tomographic measurement is depicted schematically. On the left, a chemical reactor filled with a solid catalyst is depicted. At the top of the reactor, gas and liquid are introduced and flow cocurrently through the catalyst, on which a chemical reaction takes place. The homogeneous distribution of the liquid phase is an important factor affecting the quality and the efficiency of the chemical reaction. The tomographic measurement allows a slice to be taken out of the reactor, as is depicted on the right of Figure 1. The solid catalyst is depicted in gray, the liquid phase in blue, and the gas phase in white. The location of the three different phases can be readily detected. Therefore, a detailed analysis of the phase distribution is possible, which helps to improve such chemical reactors with regard to the quality of the product and the consumption of the starting materials.

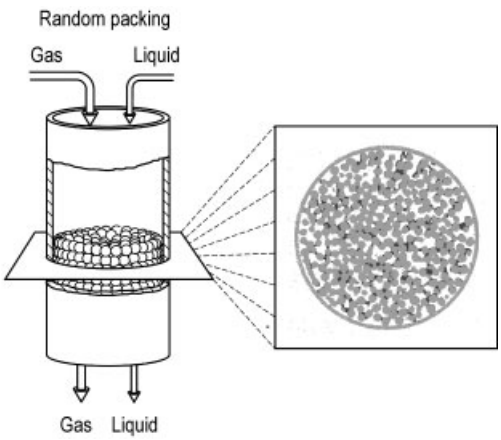

Figure 1. Schematic of tomographic measurements in a trickle-bed reactor

The choice of the tomographic techniques for a given measurement task depends on the physical properties of the fluids to be measured and the required time and spatial resolution. Medical tomography systems often cannot be used, in spite of their high spatial resolution, because their measurement time is too long for nonstationary multiphase flows. Therefore, new measurement techniques with high temporal and sufficient spatial resolution must be developed. Goals of current research and development programs are improving the spatial resolution of the electrical tomographic techniques and the temporal resolution of the X-ray techniques.

3.1. Tomographic Measurement Systems

The tomographic measurement is performed in two consecutive steps, depicted schematically in Figure 2. Figure 2a shows the steps which have to be performed for the tomographic imaging, and Figure 2b shows the tool used. In the first step integral measurements of the property field in the measurement plane are performed by the tomographic sensor and the data-acquisition system. The measured values contain information on the property field inside the sensor. After the measurements have been made, the local measurement property is calculated from the measurement values (reconstruction). The obtained measurement result represents the property field in the measurement plane.

Integral Measurement. Figure 3 shows the integral measurement of the property field for a linear tomographic system. The goal of the integral measurement is to obtain the measurement values $\boldsymbol{\Phi}_{\mathrm{M}}$ which are dependent on the property field $g(x,y)$ in the measurement plane. The measurement plane is illuminated by a beam, which is attenuated by the material inside. The beam can be a X-ray or a γ-ray beam as well as a beam of light, depending on the material in the measurement plane. If many measurements are taken in one direction, a projection is obtained. By rotating the measurement device or by mounting many beam sources and detectors, linearly independent projections in different directions are obtained. This yields the measurement vector $\boldsymbol{\Phi}$, which contains the information on the distribution of the physical property inside the measurement plane.

To generate the measurement vector, different techniques can be used. In Tables 1, 2, 3 an overview of the tomographic techniques is given. The techniques are classified according to the physical interaction between the sensor and the material inside. In principle, different physical interactions can be used for the measurement as depicted in Figure 4. In the case of transmission (Figure 4a) the physical property

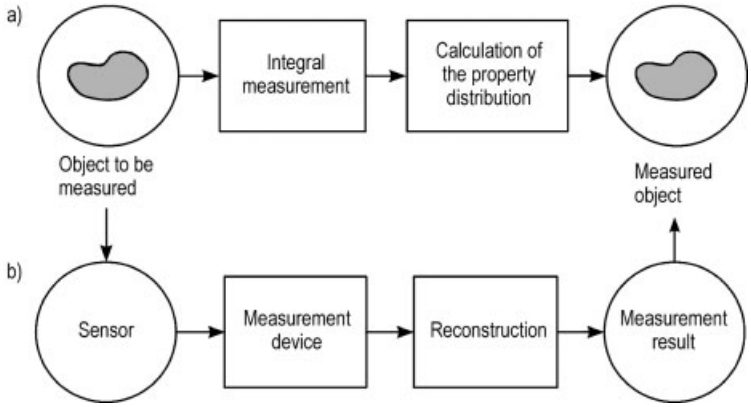

Figure 2. Schematic description of the tomographic measurement

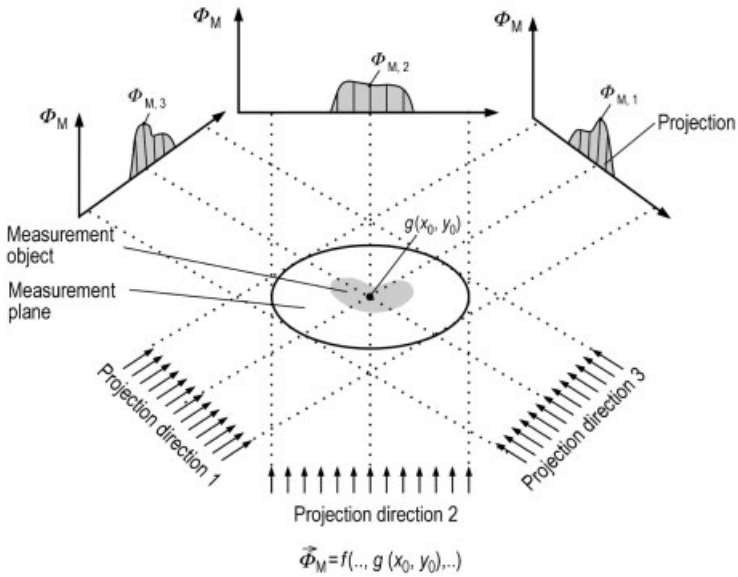

Figure 3. Integral measurement for a linear tomography system

of the electromagnetic wave or particle radiation emitted by the source is influenced by the material in the measurement plane. The change in the physical property is measured by the receiver mounted on the opposite side of the sensor. The physical property that changes can be the intensity, the polarization, or the run time of the beam. In emission techniques (Figure 4b) the intensity and the spatial orientation of the radiation emitted from the measurement plane is measured. The material in the measurement plane is itself the source of the radiation. The radiation can be natural or the result of external excitation. In reflection techniques the position and the change in the physical property of a wave or a particle that has been reflected from an interface are measured. Similar to the reflection techniques, for the refraction and diffraction techniques, the change of the direction or the physical property of a wave or particle radiation at a discrete or continuous interface in the measurement plane is measured. Particle radiation or high-energy electromagnetic waves cause scattering effects when they strike molecules or atoms. These scattering effects can also be used for tomographic measurements. In this case the statistical distribution of the property vector as well as its orientation are measured.

Table 1. Electrical tomographic techniques in process engineering [11]

	Electrical capacitance tomography (ECT)	Electrical resistance tomography (ERT)	Electromagnetic inductive tomography (EMT)	Microwave tomography
Parameters measured	capacitance and loss angle at high frequencies (1 MHz)	resistance and phase angle at low frequency (1 – 500 kHz)	independent measurements of magnetic permeability and eddy current loss (50 Hz – 5 MHz)	real and imaginary components of complex permittivity (2 – 3 GHz)
Parameters derived	solids: volumetric concentration, moisture profiles gases: concentration profiles	solids and gas volumetric concentration profiles	separate volumetric concentration profiles of ferrous and nonferrous materials	structural integrity, physical parameters, e.g., water content, temperature
Applications	pneumatic conveyors, fluidized beds, combustion imaging, foams, oil/gas flows	hydraulic conveyors, solid/liquid mixers and separators, liquid/gas flows, liquid/liquid flows	metallurgical processes, location of ferrous and nonferrous contaminants in products, particle population tomography (ferrous labels)	conveyed products, wood, paper, pipes, buried objects
Speed	100 frames/second	200 frames/second	100 frames/second	100 frames/second
Resolution	5 – 10 % of vessel diameter	3 – 6 % of vessel diameter	5 – 10 % of vessel diameter	10 – 20 % of vessel diameter
Slice thickness	> 25 cm or diameter/ 10	diameter/ 10	diameter/ 5	diameter/ 10
Vessel diameter	with present electronics max. 1 m, no theoretical limit	no theoretical limit, typically 1 mm to 3 m	200 mm	no theoretical limit
Cost	low	low	low (expensive if magnetic screen is used, without screening similar to ECT)	medium
Notes	electrodes are not in contact with the fluid, no electrically conducting material is allowed in the measurement plane, reconstruction of the images is complicated due to the nonlinearity of the electrical field	electrodes are in contact with the fluid, special: wiremesh sensor	at research stage, no electrodes	most suitable for thin slabs of material
References	[12–15]	[16–18]	[19]	[20–27]

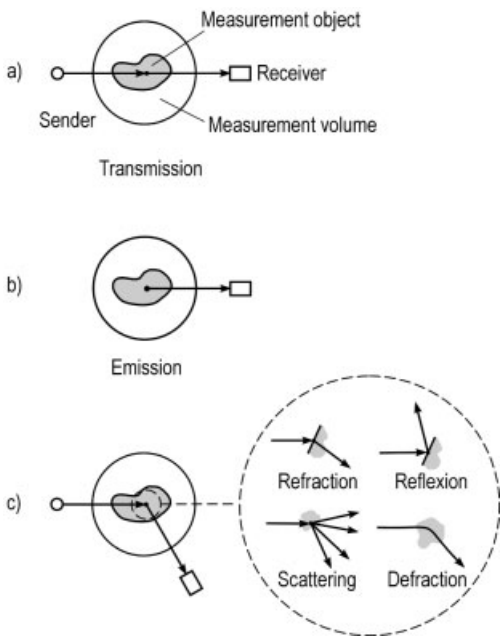

Figure 4. Physical interactions used for tomographic measurement

Reconstruction. The measurement vector $\boldsymbol{\Phi}$ implicitly contains the information of the measured property field $g(x, y)$. The information is extracted in the second step of the tomographic measurement process, known as reconstruction. During reconstruction, the property field $g(x, y)$ is calculated by a mathematical inversion of the measurement process. If the measurement vector is as large as the number of unknown elements of the property field, a complete direct calculation of the property field is possible. In process engineering applications, the number of measurements is very low, to achieve a short measurement time. Therefore, the reconstruction of the property field has to be conducted with mathematical approximations.

In Figure 5a the measurement process is depicted schematically. The goal of reconstruction is the mathematical inversion of the measurement process, which is depicted in Figure 5b. Usually there are more unknown properties in the property field, than measurement values. Therefore, to improve the quality of the reconstruction, iterative techniques have to be used.

Table 2. Radiation tomography in process engineering [11]

	γ-ray	X-ray	Optical attenuation/emission	Optical diffraction	Positron emission	Positron emission/particle tracking
Parameters measured	photon attenuation	photon attenuation, Compton scattering	received light intensity	refractive indices of fluides	distribution of pairs of back-to-back γ-rays	pairs of back-to-back γ-rays from single tracer particle
Parameters derived	mass concentration profiles	mass concentration profiles	profile of absorbent concentration or temperature	density, temperature, concentration profiles	concentration field or radioactive tracer	trajectory of tracer particle
Applications	gas – solid – liquid processes	gas – solid – liquid processes	discontinuous phase in liquid or flames, mixing	mixing, heating of viscous fluids, flames, density fields near impeller flames	steady state segregation	solids mixing
Speed	depending on the spatial resolution needed: 90 to 30 frames/second	depending on the resolution needed: few seconds to 2000 frames/second	high	ca. 15 s	1 h for 3D volume image	100 locations/s
Resolution	1 mm for rotating scanner	$<$ 1 mm	1 – 2 mm	2 mm	8 mm in 3D	depends on trace speed: 5 mm at 1 m/s
Slice thickness	$<$ 5 mm	$<$ 5 mm	1 – 2 mm	2 mm	8 mm in 3D	depends on trace speed: 5 mm at 1 m/s
Vessel diameter	200 mm	$<$ 1 m	no theoretical limit	$<$ 100 mm	$<$ 600 mm	$<$ 600 mm
Cost	high	high	low	high	high	high
Notes	radiation safety precautions required	radiation safety precautions required	optical access to the measurement plane is necessary	optical access to the measurement plane is necessary	skilled operation using radiolabeled material	skilled operation using radiolabeled material
References	[28]	[29–39]	[40–47]	[48–51, 111–115]	[52, 53]	[54]

Table 3. NMR, acoustic and triboelectric tomographic techniques in process engineering [11]

	Nuclear Magnetic Resonance	Acoustic Tomography		Triboelectric Tomography
	Imaging	High frequencies (liquids)	Low frequencies (gases)	
Parameters measured	nuclear spin density of various chemical species	transmission, reflection, and scattering at 1 MHz	velocity of sound	electrostatic charge on particles
Parameters derived	quantity of individual species	size of gas-filled zones	temperature, vector velocity	flow regime and flow velocity
Applications	concentration profiling in various processes and products	measuring gas profiles of liquid or liquid/solid mixture	temperature and velocity profiles in furnaces	flow detection, velocity, and regime detection in gas/solid flows
Speed	25 frames/second	limited by speed of sound	limited by speed of sound	high, depends on particle count
Resolution	10 – 40 µm	3 % of vessel diameter	8 $\times$ 8 pixels	3 mm near vessel wall, insensitive to particles more than 5 m from wall
Slice thickness	$<$ 1 mm	15 mm	not specified	3 mm
Vessel diameter	maximum practicable diameter is the dimension of the human body	no theoretical limit	power station boilers have been used to test the system	no theoretical limit
Cost	high	medium	high	low
Notes	small amount of ferromagnetic or significant amount of diamagnetic material will prevent the system from working	large number of small gas bubbles block the ultrasound beam	ca. 150 dB sonic pulses	in development, only sensitive close to the wall of the measurement plane
References	[55, 56]	[57–62]	[63]	[64, 65]

These require a mathematical description of the measurement process, which is depicted in Figure 5c. With this mathematical description, the theoretical measurement vector is calculated and compared to the actual measurement vector. The difference between the two leads to a change in the property field from which a new theoretical measurement value can be calculated. This iterative process of calculating and comparing the measurement vectors and correcting the property field is continued until the difference between the two measurement vectors is below a certain error margin.

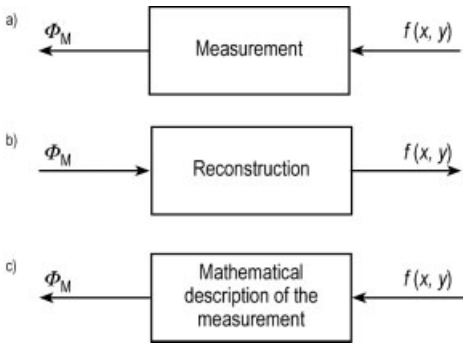

Figure 5. Schematic of the reconstruction process

In Table 4 the known reconstruction techniques are listed. In principle, they can be divided into series expansion, inversion, and adaptive techniques. In series expansion techniques, the field function $g(x, y)$ is discretized. By using a mathematical description of the measurement process, a set of linear equations with the known measurement values and the unknown property field is then developed. In inversion techniques, the measurement process is mathematically inverted by using simplifications and preprocessing the measurement values. The discretization of the property field $g(x, y)$ is often not necessary. For the adaptive techniques, a detailed mathematical description of the measurement process is often not necessary. The algorithm is trained by a priori knowledge to reconstruct the property field.

Spatial and Temporal Resolution. The choice of the tomographic measurement system which is most suitable for a given measurement task is not only dependent on the technical limitations arising from the surroundings of the measurement (pressure, temperature, and material properties of the fluids and the apparatus), but also on the desired spatial and temporal resolution. The latter two parameters are coupled to each other. High temporal resolution results in a low spatial resolution, and vice versa. The spatial and temporal resolutions of the tomographic techniques are qualitatively compared to each other in Figure 6.

The classical tomography systems, such as X-ray tomography, positron emission tomography, and nuclear magnetic resonance tomography, have high spatial resolution. In contrast, their time resolution of 1 s to 30 min is rather low. For material science applications, particle-radiation techniques are quite commonly used because of their high energies, but their time resolution is even worse than those of the medical techniques. Current high-speed X-ray tomography systems are very expensive and have a relatively low spatial resolution.

For the application of tomographic measurement techniques in multiphase flows, especially those with rapidly changing flow patterns, high time resolution is essential. This can be realized by using optical techniques, which are limited to transparent measurement locations and to fluids that have the same refractive index. This can only be achieved for a few applications. In comparison to the above-mentioned techniques, the electrical tomographic techniques have a high temporal resolution and can be applied in most technical surroundings. The high temporal resolution can be achieved because the sensors have no moving parts and electrical components have short switching times. The relatively low spatial resolution of these techniques is generally sufficient for the applications.

3.2. Capacitance Tomography

The measurement principle of capacitance tomography is schematically depicted in Figure 7. Electrodes are mounted peripherally along the circumference of the measurement plane. The integral measurement values are taken by measuring the capacitance between two or more electrodes. The measured capacitance is dependent on the permittivity field inside the sensor. If the permittivities of the fluids inside the sensor are different, they can be distinguished from

Table 4. Reconstruction techniques

Reconstruction techniques [66–68]						
Series expansion [69, 70]			Inversion [71]			Adaptive
Direct	Iterative		Direct	Sequence		
Matrix methods [72, 73]	ART [78] (grid, sample)	Constraining	Radon [82]	Filtered backprojection [86]	Filter [85]:	Neural network [89]
Linear backprojection [74]	SIRT [79] (grid, sample)	Relaxation	Convolution methods [83, 84]	Iterative Fourier transformation [89]	Band limit	Pattern association [14]
Annular harmonic decomposition [75]	SART [80]	Rearrangement Normalization	2D/3D Fourier transformation [85]		sin/cos	Generic algorithm [90]
Polynomial decomposition [76]	ILST [81]	Selective smoothing			Hamming	
NPDM [77]	Richardson [81]	Thresholding			Kernels: Shepp and Logan [87] Rama. and Laksh. [88]	

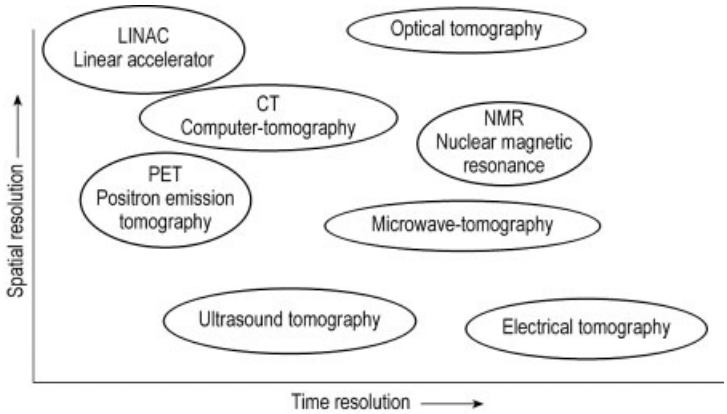

Figure 6. Qualitative comparison of the spatial and temporal resolution of tomographic measurement systems

each other by the tomographic measurement (e.g., water and air have a relative permittivity ratio of 1 : 80). Capacitance can be measured by various electrical measurement circuits, e.g., the sender electrode is charged and the current at the receiver electrode is measured. Table 5 lists applications of capacitance tomography in process engineering.

The electrical field inside the measurement plane is described by the Poisson equation

$$\nabla^2 \varphi(x,y,z) + \frac{1}{\varepsilon(x,y,z)} \operatorname{grad} \varphi(x,y,z) \operatorname{grad} \varepsilon(x,y,z) = \frac{\varrho(x,y,z)}{\varepsilon(x,y,z)} \quad (1)$$

The right side of Equation 1 contains the source terms for a free charge carrier in the measurement plane. For most applications the fluids are without electrical charge and therefore the right side of the equation can be set to zero. The left side of the equation consists of the Laplace and Newton equations. The Laplace equation is valid for homogeneous permittivity fields and is extended to nonhomogeneous permittivity fields by the Newton equation. The capacitance is calculated by Equation 2

$$C = \frac{\int\int_A \varepsilon(x,y,z) \operatorname{grad} \varphi(x,y,z) \, \mathrm{d}\hat{A}}{\int_s \operatorname{grad} \varphi(x,y,z) \, \mathrm{d}\hat{s}} \quad (2)$$

which is the shift of the charge on the surface of an electrode divided by the potential difference along the line between a pair of electrodes. It is evident from Equation 1 that the relation between potential and permittivity is linear with

Table 5. Applications of capacitance tomography [15]

Phases	Application	Fluids	Dimensions	Refs.
Gas – liquid	horizontal plug – flow	kerosene – nitrogen	$d_i = 154$ mm	[91] [12,92]
		oil – air water – air	$d_i = 54$ mm	[93]
Gas – solid	pneumatic transport	air – sand	$d_i = 84$ mm	[94] [95]
	fluidized bed	air – sand air – PS	$d_i = 150$ mm	[96] [97]
			$d_i = 800$ mm	[98] [99]
Gas – gas	combustion	air – methane	$d_i = 76$ mm	[100]
Gas – liquid – solid	phase separator	air – foam oil – water – slurry		[101] [102]
Gas – liquid – liquid	three – phase – flow	oil – water – gas	$d_i = 82$ mm	[103]
Gas – liquid	trickle-bed reactor	water – air	$d_i = 120$ mm	[104]

respect to the potential, but nonlinear with respect to the permittivity. In addition the capacitance between a pair of electrodes depends on the permittivity field in the entire measurement plane. This makes the reconstruction of the permittivity field more complicated in comparison to other sonographic techniques, as depicted in Figure 8. In Figure 8a the relation between the property field and the measured physical value is depicted for a linear tomographic system, e.g., X-ray or optical tomography. The measured attenuation of the beam is only dependent on a small part of the measurement plane, because the radiation beam follows a straight path from the sender to the receiver. The relation between the attenuation of the beam and the density field inside the sensor is only dependent on the density field. Because of this, the reconstruction of such measurement data is comparatively easy.

In Figure 8b the measurement process for capacitance tomography is depicted. The measured capacitance is dependent on the permittivity field of the entire measurement plane. In addition, the electrical field is distorted by the permittivity field, which results in a change in the field lines or the potential lines. According to Equation 2, this results in a change in the capacitance. Therefore, the reconstruction of the capacitance tomography systems has to be performed with more sophisticated algorithms. The two algorithms most widely used for the reconstruction of capacitance measurement are linear backprojection [92, 105–107]and the ART algorithm. Other methods include model-based reconstruction [103, 108, 109], neural networks [110], and pattern association [14, 93].

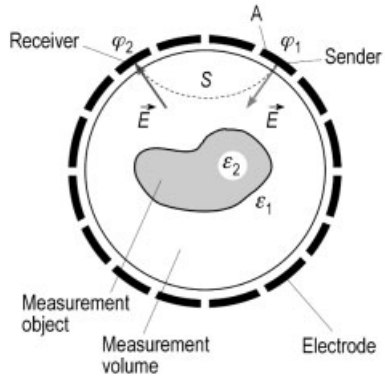

Figure 7. Measurement principle of capacitance tomography

Before the reconstruction is performed, it has to be taken into account that the measurement values are taken segmentally and that the flow is therefore moving on while the measurements are taken. This is implemented, by a time and spatial correction [15]. The measurement values of each combination of electrodes is approximated by a fitting function. Then the measurement value of each combination is available at any desired time. If the velocity of the flow is known, the measurement values can be calculated at the time when the slice of the flow which is to be reconstructed was in front of the electrodes.

3.3. Conductive Tomography

In conductive tomography the electrodes are generally mounted peripherally around the measurement volume. They are mounted flush with

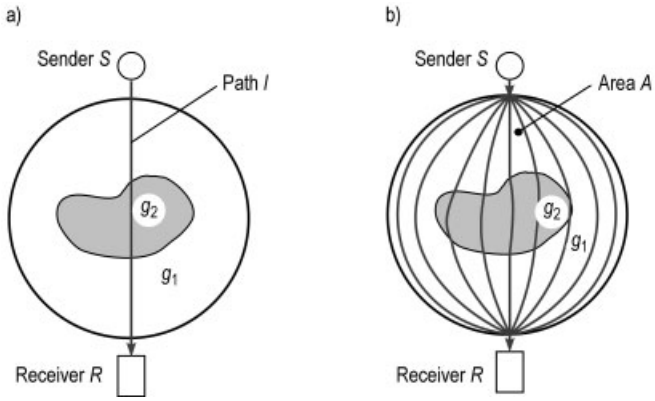

Figure 8. Comparison of linear and nonlinear tomographic measurement techniques

the wall and in direct electrical contact with the flowing fluid phases [16, 17]. Depending on the techniques, either a constant current is induced and the resulting potential at the other electrodes is measured, or a constant potential is applied to two electrodes and the resulting current measured. Therefore, the electrical properties measured have a nonlinear dependence on the phase distribution. In addition, this techniques requires full contact of the electrodes with the conducting phase.

A different conductive tomographic technique is depicted schematically in Figure 9 [18]. The sensor consists of three planes of parallel wires which are strung across the measurement plane. The three planes are shifted relative to each other such, that the measurement plane is segregated into small equilateral triangles. The conductance measured between two adjacent wires is dependent on the distribution of the fluids between the wires. By measuring the conductance between all wires, three projections of integral measurement values are obtained, which are used for the reconstruction of the conductivity distribution inside the measurement plane.

Figure 10 shows the measurement result for a single slug in horizontal gas – liquid slug flow. By stacking the measured and reconstructed images obtained by the tomographic measurement, an Eulerian representation of the phase distribution of the slug is obtained. This Eulerian representation can be used to compare the measured and reconstructed slugs to visually observed ones. In Figure 10a the side view of a slug is given. The direction of flow is from right to left. The mixing zone at the front of the slug is clearly visible. In order to compare this visual observation to the result of the tomographic measurement, two Eulerian plots are shown in Figure 10b and c. Two surfaces with different liquid contents are shown. The dark colored surface encloses regions with a liquid holdup over 0.5. The light colored surface represents a liquid holdup of 0.1. The slug is also represented in a side view, with the direction of flow from right to left. The surfaces are transparent, so that the mixing zone at the slug front and the gas bubbles entrained there can be seen. From visual comparison it is evident that the characteristic slug properties, in particular the shape of the mixing zone, are captured by the tomographic technique.

3.4. Optical Tomography

An overview of optical tomography is given in [111]. Jet mixing processes have been investigated by measuring the concentration of a dye by optical tomography [112, 113]. Temperature fields have been obtained from the phase shift of four laser beams passing through the flow field [114]. The phase shift depends on the refractive index, which in turn depends on the temperature of the fluid in the mixing tank.

Two lasers of different wavelength have been used to measure the micro- and macrotransport processes in the mixing of viscous liquids [115]. A mixture of an inert dye (Duasyn Acid Ponceau: red) and a dye that is decolorizes by a chemical reaction with the vessel contents (iodine: blue) are fed into the vessel. The rate of the

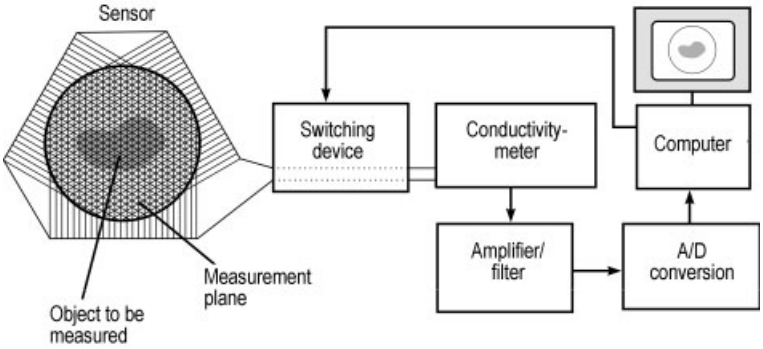

Figure 9. Measurement chain for conductive tomography with a wire-mesh sensor [18]

decolorization reaction is limited by micromixing. The local intensity of segregation is therefore obtained by comparing the local concentration of the reacting dye with that of the inert dye. Macromixing leads to equal dilution of both dyes, where as micromixing only reduces the concentration of the reacting dye. The three-dimensional transient concentration fields of both dyes are measured by tomographic absorption spectroscopy. The mixing vessel is illuminated from three directions with two superimposed laser beams of different wavelength. The absorption of the laser beams depends on the absorption coefficients of the dyes and on their concentrations along the beams. The light absorption is measured for all projections. From the measurement data the dye concentration is reconstructed in 1200 elements per plane of the vessel, and the vessel is divided into 30 planes. Thus, the local dye concentrations and the local intensity of segregation can be detected at 36 000 locations in the vessel.

In Figure 11 the initial measurement result is depicted in a quasi three-dimensional representation. The impeller was a Rushton turbine. The inert and reacting dyes can be clearly distinguished from each other. Five seconds after injection of the dye into the vessel, effective mixing has not taken place. The dyes stay in the main vortex caused by the impeller, so that the concentration of red dye in the middle plane is still high. In addition, due to the slow macrotransport, micromixing is also not effective. The blue iodine droplet is only stretched in the circumferential direction, resulting in a slow diffusive transport. With the tomographic measurement system, the mixing process can be observed fully three dimensionally, and insight into the efficiency of micro- and macromixing can be obtained at the same time.

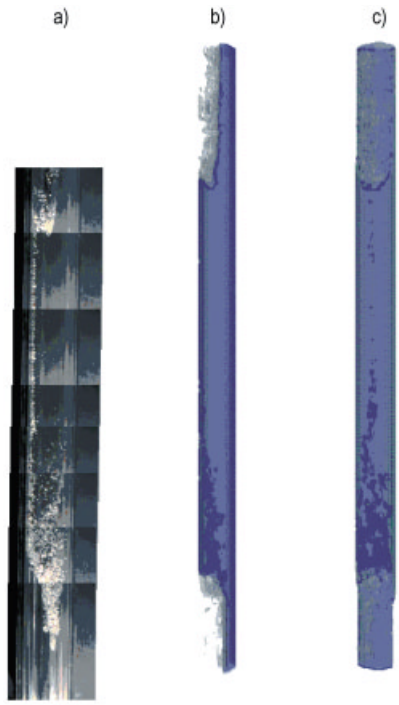

Figure 10. Tomographic measurement result for horizontal gas – liquid slug flow

3.5. X-Ray Tomography

X-ray tomography is quite widely distributed for medical diagnosis and material science applications but is rarely used for the measurement

and control of processes. As the available equipment has rather long measurement times, it only can be used for the measurement of stationary or quasistationary processes. With new developments in positioning systems and computer technology, measurement times in the order of a few seconds are possible. With a high technical and financial effort measurement times of a few milliseconds can be attained [33]. With such systems, X-ray tomography becomes a valuable tool for the measurement of nonstationary multiphase flows.

X-ray tomography has already been successfully used for process engineering applications, for example, investigations of porosity and dispersion coefficients [116, 117] nonstationary multiphase flows [118], and gas – liquid flows in pipes [119, 120]. The temporal resolution of these measurements is about one minute with a spatial resolution of a few centimeters.

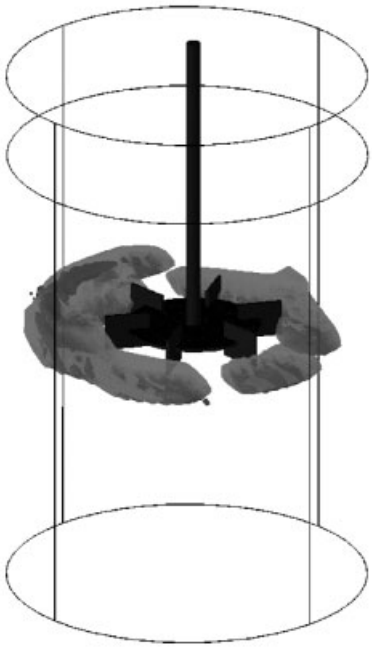

Figure 11. Measurement of micro- and macromixing by optical tomography [115]

Higher spatial resolution is required for measurements in regular and irregular packings. Here the distribution of the liquid and the void fraction are investigated as a function of the different packings, the flow rates of the fluids, and the physical properties of the fluids. The distribution of liquid in a trickle-bed reactor has been investigated by using a medical tomograph [121]. Because the energy of this tomograph is rather low, a random packing of glass spheres of diameter 0.4 mm was used in a reactor of 45 mm diameter. The parameter is the liquid flow rate. Because the spatial resolution of the tomograph is on the order of the size of the spheres, the local liquid holdup can be measured, but the paths of the liquid on the spheres cannot be resolved. In similar experimental investigations [122], vertical slices through a packing of 60 mm diameter and 300 mm height were obtained. In more recent investigations, an X-ray tomograph with an acceleration voltage of 120 kV, was used and gave a spatial resolution of about 1.1 mm^2 with a measurement time of 2.5 min [38]. The experimental investigations are carried out in a column of 0.6 m diameter and 2 m height. The packing consists of plastic spheres. The measured liquid holdups show a good agreement with those obtained by an empirical correlation.

3.6. Example of the Application of Tomography in a Trickle-Bed Reactor

A trickle-bed reactor, depicted schematically in Figure 1, consists of a random packing of catalyst spheres. At the top of the reactor, gas and liquid are introduced and flow cocurrently to the bottom. Such reactors are used in the chemical industry for hydrofinishing and hydrocracking. In the packing, four different flow regimes can become established (Figure 12), depending on the flow rates of the fluids. At low flow rates a continuous liquid film flows over the spheres of the packing (Figure 12a). With increasing flow rates the liquid film becomes unstable, eventually forms liquid bridges between the spheres, and the pulse-flow regime (Figure 12b) becomes established. The liquid phase is transported in large plugs through the packing. At high gas flow rates the liquid becomes dispersed and flows in drops through the column (Figure 12c), while at high liquid flow rates the gas phase is dispersed in bubbles. The pulse flow regime is supposed to increase the mass transfer, because of the transient change of gas and liquid phases and the turbulence, and is therefore of major industrial interest.

In the film flow regime, X-ray tomography is applicable, as the time resolution need not necessarily be high. The distribution and the thickness of the liquid films can be measured with high spatial resolution. For the pulse flow regime, high time resolution is essential. The spatial resolution need not to be too high, as the most important parameters to be measured are the location and size of the liquid slugs in the measurement plane, and these are large in comparison to the spheres.

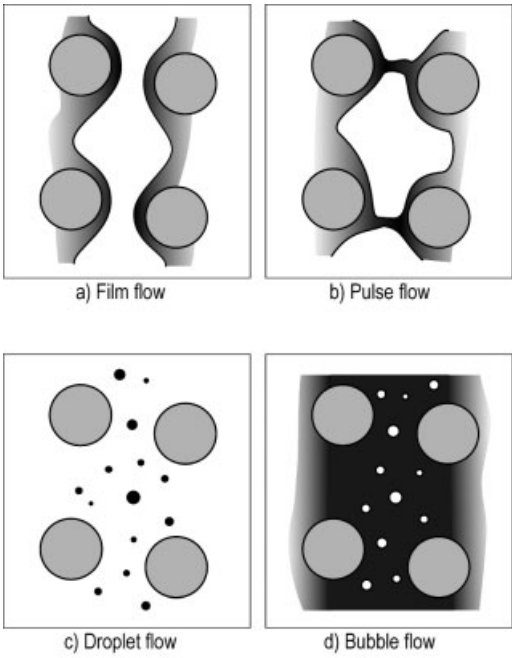

Figure 12. Flow regimes in trickle-bed reactors

Capacitance Tomography. The liquid distribution in the packing for the pulse flow regime has been investigated with the tomographic measurement chain depicted in Figure 13 [15]. It consists of a sensor with 16 electrodes on which electrical microswitches are mounted. These microswitches connect the electrodes with the capacitance meter. With this arrangement only one capacitance meter need be used, and measurement wires are relatively short. The capacitance meter supplies a voltage which is directly proportional to the measured capacitance. This voltage is digitized and stored in a PC for the post-processing of the measurement values. The same PC controls an EPROM, which in turn controls the microswitches attached to the electrodes. The reconstruction can be performed with the backprojection algorithm on the PC or on a workstation, where more sophisticated reconstruction techniques, like ART, are available.

On the left of Figure 14, the sequentially measured pictures are depicted. Regions with a high liquid holdup are shown in dark colors. The ceramic spheres are not visible because they have been removed from the measurement values by calibration of the measurement system. The slices have a time interval of 3 ms and show the change of the phase distribution in the sensor. They can be added to one other to give a quasi-three-dimensional representation of the measurement results. A vertical slice through this extracts the temporal change of the phase distribution in the measurement plane. The liquid plugs which are formed in the reactor are clearly visible in this representation. They have a velocity of about 1.5 m/s and a frequency of about 5 – 10 Hz. The quasi-three-dimensional representation is depicted in Figure 15. Five slices are made through the measurements. The three horizontal slices show the phase distribution as measured by the tomographic system. The larger, vertical slices show the temporal change of the phase distribution.

X-ray Tomography. Figure 16 shows an X-ray tomographic measurement chain. The sensor consists of a X-ray tube which emits radiation in a fan beam. The attenuation of the X-rays in the measurement plane is measured by 15 detectors. The detectors supply a voltage which is directly proportional to the measured attenuation. The voltage is digitized and stored in a computer, which controls the measurement process. The measurement object is mounted on a positioning table, which is turned and shifted to obtain the number of projections needed for the reconstruction. The complete reconstruction is performed on a workstation. The spatial resolution is 0.4 mm^2 at a measurement time of 15 min.

Figure 17 shows the pictures obtained in a trickle bed-reactor with a diameter of 190 mm. The liquid flow rate was $30\,m^3m^{-2}h^{-1}$. Details of the measurement setup and the results are given in [39]. The liquid phase is depicted in black, and the ceramic spheres of diameter 10 mm in gray. From these slices and the slice

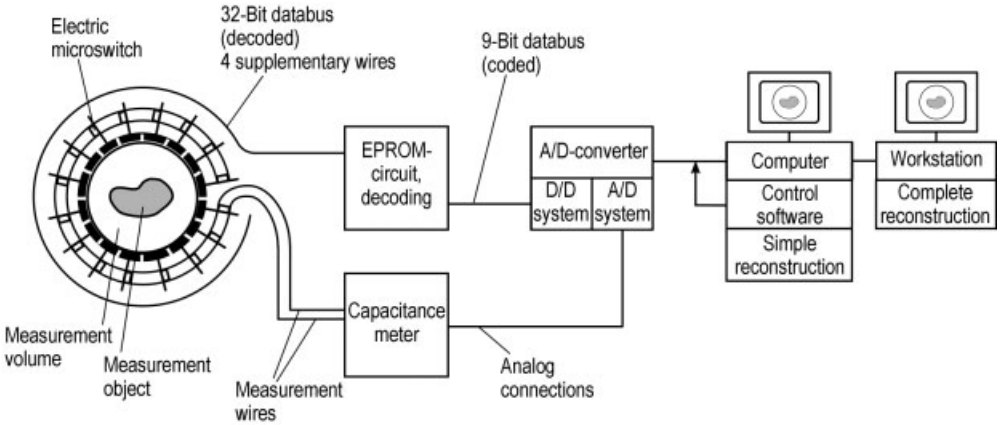

Figure 13. Measurement chain for capacitance tomography [15]

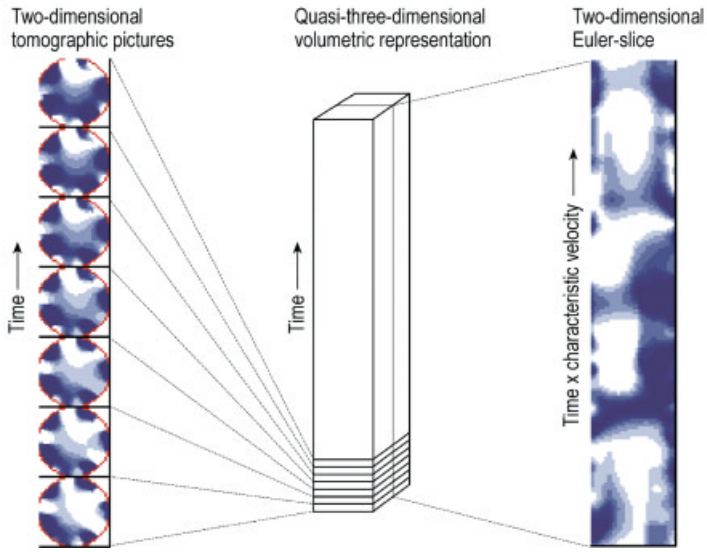

Figure 14. Measurement results obtained by capacitance tomography

on the right of Figure 17, it can be seen that the liquid phase is distributed very inhomogeneously in the packing. This is mainly due to the random arrangement of the spheres, which block the flow of the water at some locations, which results in the accumulation of water. To give an impression of the high spatial resolution, Figure 18 shows a three-dimensional representation of the measurement result for ceramic spheres of diameter 10 mm. The figure shows a surface through points in the measurement volume of equal density, which corresponds to the surface of the spheres.

4. Future Perspectives

Capacitance tomography has the advantage of a time resolution which is in the range of 100 slices of the local void fraction per second. The spatial resolution is about 10 % of the sensor diameter and is mainly dependent on reconstruction. The liquid plugs which occur in the pulse-flow regime in trickle-bed reactors, can be measured well. The location and the frequency of the plugs can be identified, and this leads to a better understanding of the formation and the establishment of the pulse-flow regime. Further research in the

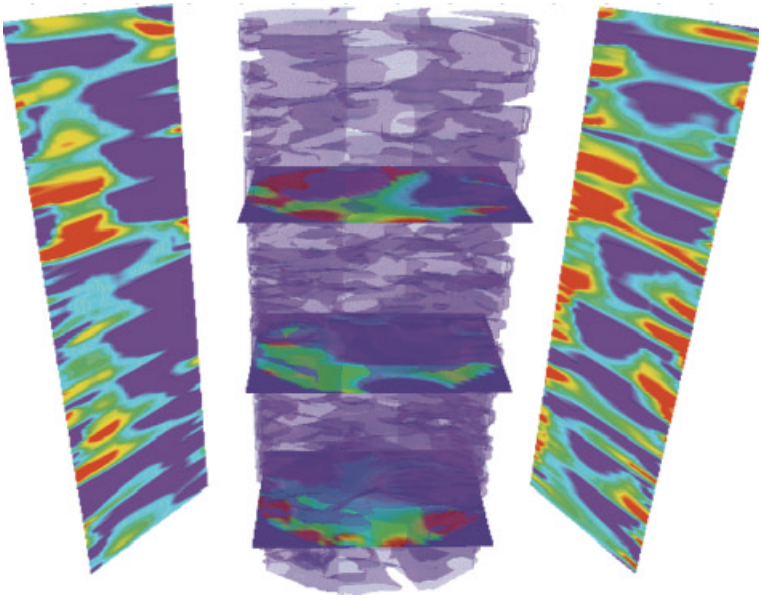

Figure 15. Quasi-three-dimensional representation of measurement results

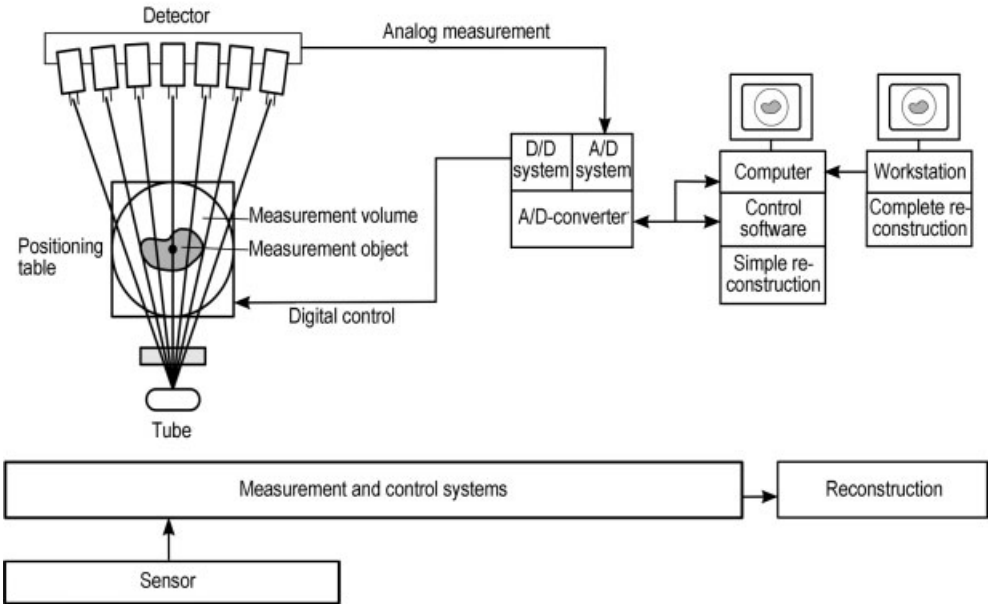

Figure 16. Measurement chain for X-ray tomography

field of capacitance tomography will deal mainly with improving the reconstruction method, decreasing the errors, and therefore increasing the spatial resolution.

With X-ray tomography, a high spatial resolution, on the order of 0.4 mm^2, can be achieved. Because the measurement times are often in the range of minutes, this technique is only applicable to stationary flows. The application of X-ray tomography in the film-flow regime of a trickle-bed reactor shows the uneven distribution of the liquid phase inside the reactor. To decrease the measurement time of these systems, up to now, a considerable technical and monetary effort was

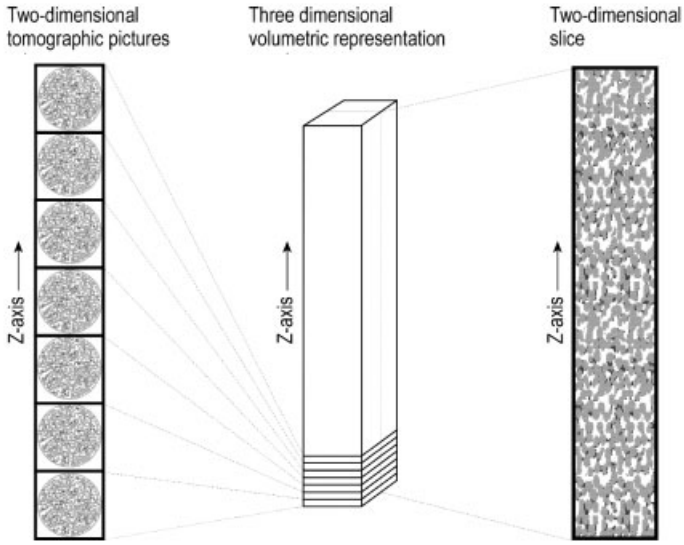

Figure 17. Measurement results obtained by X-ray tomography

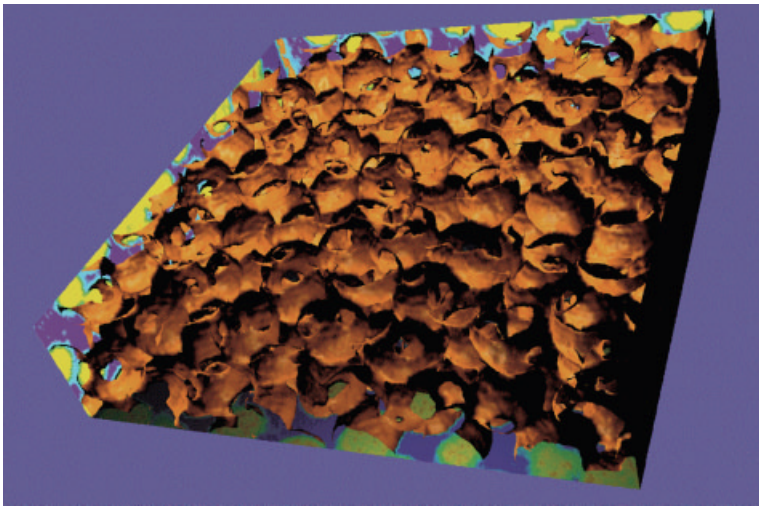

Figure 18. Three-dimensional representation of the measurement result for ceramic spheres

necessary. Therefore, these systems are rarely used in research and in process engineering. Further developments will involve shortening the measurement time of such systems without making them more expensive.

5. References

General References

1. For further information on the authors and on tomography, see http://www.ifv.uni-hannover.de/
2. M. S. Beck, R. A. Williams, E. Compogrande, M. A. Morris, R. C. Waterfall (eds.): *Tomographic Techniques for Process Design and Operation,* Computational Mechanics Publications, Southampten and Boston 1992.
3. M. S. Beck, R. A. Williams, E. Compogrande, M. A. Morris, R. C. Waterfall (eds.): *Process Tomography – a Strategy for Industrial Exploitation,* UMIST, Manchester, 1993.
4. M. S. Beck, R. A. Williams, E. A. Hammer, E. Compogrande, M. A. Morris, R. C. Waterfall (eds.): *Tomography – a Strategy for Industrial Exploitation,* UMIST, Manchester, 1994.

5. M. S. Beck, R. A. Williams, B. S. Hoyle, M. A. Morris, R. C. Waterfall (eds.): *Tomography – 95 Implentation for Industrial Process,* UMIST, Manchester, 1995.
6. R. A. Williams, M. S. Beck (eds.): *Tomography – Principles, Techniques and Applications,* Butterworth – Heinemann, Oxford, 1995.
7. M. S. Beck, E. A. Hammer, B. S. Hoyle, M. A. Morris, R. C. Waterfall, R. A. Williams (eds.): Process Tomography, *Meas. Sci. Technol, Special Issue,* **March** (1996).
8. D. M. Scott, R. A. Williams, *Frontiers in Industrial Process Tomography,* Engineering Foundation, New York 1997.
9. B. S. Hoyle, D. M. Scott, *Frontiers in Industrial Process Tomography,* Engineering Foundation, New York 1997.
10. *Internationales Symposium über Computertomographie für die industrielle Anwendung, Berichtsband 44,* Bundesanstalt für Materialprüfung, Berlin 1994.
11. M. S. Beck, T. Dyakowski, R. A. Williams, "Process Tomography – The state of the Art," *Proc. frontiers in Process Tomography II,* 1997, 357 – 362.
12. W. Q. Yang, A. L. Scott, M. S. Beck, C. G. Xie: "Development of Capacitance Tomographic Imaging Systems for Oil Pipeline Measurements", *Rev. Sci. Instrum.,* **66** (1993) 8, 4326 – 4332.
13. Ø. Isaksen, J. E. Nordtvedt, "Capacitance Tomography; Reconstruction Based on Optimization Theory," *Proc. ECAPT '92, Manchester, Process Tomography – A Strategy for Industrial Exploitation,* 1992, 178 – 189.
14. F. Klug, "Ein Meßverfahren zur impedanzgestützten Rekonstruktion der Gemischzusammensetzung in Mehrphasenströmungen," *Fortschr.-Ber. VDI* **3** 327, (1993).
15. N. Reinecke, *Kapazitive Tomographie für transiente Mehrphasenströmungen,* Dissertation, Universität Hannover, 1996.
16. F. J. Dickin, M. Wang, "Impendance Sensors – Conducting Systems" in R. A. Williams, M. S. Beck (eds.): *Process Tomography – Principles, Techniques and Applications,* Butterworth – Heinemann, Oxford, 1995, pp. 63 – 84.
17. M. Z. Abdullah, T. Dyakowski, F. J. Dickin, R. A. Williams, "Observation of Hydrocyclone Seperator Dynamics using Electrical Impedance Tomography", *Proc. ECAPT '93, Karlsruhe, Germany, Process Tomography – A Strategy for Industrial Exploitation,* 1993, 93 – 96.
18. N. Reinecke, G. Petritsch, M. Boddem, D. Mewes: "Tomographic Imaging of the Phase Distribution in Two-Phase Slug Flow", *Int. J. Multiphase Flow* **24** (1998) 4, 617 – 634.
19. A. J. Peyton, "Mutual Inductance Tomography" in R. A. Williams, M. S. Beck (eds.): *Process Tomography – Principles, Techniques and Applications,* Butterworth – Heinemann, Oxford, 1995, pp. 85 – 100.
20. R. Aitmehdi, A. P. Anderson, M. Ferrando: "The Determination of Dielectric Loss Tangent by Microware Phase Tomography", *Inrose Problems* **4** (1988) 5, 333 – 345.
21. J. H. Jacobi, L. E. Larsen, C. T. Hast: "Water-Immersed Microwave Antennas and Their Application to Microwave Interrogation of Biological Targets", *IEEE Trans. Microwave Theory Tech.* **27** (1979) 1, 70 – 78.
22. M. F. Adams, A. P. Anderson: "Synthetic Aperture Tomographic Imaging for Microwave Diagnostics", *Proc. Inst. Elec. Eng.* **129** (1982) 83 – 88.
23. T. C. Guo, W. W. Guo: "Physics of Image Formation by Microwave Scattering", *Med. Imag. SPIE Proc.,* **767** (1987) 1 – 10.
24. C. Pichot, L. Jofre, G. Peronnot, J. C. Boloney: "Active Microwave Imaging of Inhomogeneous Godies", *IEEE Trans. Ant. Propojat.,* **AP-33** (1985) 416 – 425.
25. J. C. Boloney, M. S. Hawley: "Noninvasive Control of Hyperthermia" in M. Gautherie (ed.): *Methods of Hyperthermia Control,* Springer, Berlin, 1990, 35 – 111.
26. F. Caspers, J. Conoway, "Measurement of Power Density in a Lossy Material by Means of Electromagnetically Acoustic Signals for Non-Invasive Determination of Spatial Thermal Absorption in Connection With Pulsed Hyperthermia", *12th European Microwave Conference Helsinki,* 13. – 17. Sept., 1982, Microwave Exhibitions and Publishers Ldt, pp. 565 – 568.
27. R. G. Olsen, J. C. Lin, "Microwave Puls Induced Resonance in Spherical Head Models"; *IEEE Trans. Microwave Theory Tech.* **29** (1981) 10, 1114 – 1117.
28. G. A. Johanson, T. Froystein, B. T. Hjertaker, O. Isaksen, O. Olsen, B. McKibben, A. Valle, E. A. Hammer: "Current Status of the Bergen Dual Mode Tomographic Project" in M. S. Beck, R. A. Williams, B. S. Hoyle, M. A.

Morris and R. C. Waterfall (eds.): *Process Tomography – 95 Implementation for Industrial Process,* UMIST, Manchester 1995.
29. R. Grimm, *Industrielle CT-Anlagen,* 1991, München, DGZfP, 64 – 73.
30. R. Link, W. Nuding, H. Wiacker: "Industrielle Computer Tomographie", *Atomwirtschaft* **10** (1992) 495 – 497.
31. N. Takenaka, T. Fujii, A. Ono, Y. Motomura, A. Toruno: "Visualization and Measurement of Multi-phase and Liquid Metal Flow by Real-Time Neutron Radiography", *Proc. German-Japanise Symp. on Multi-Phase Flow,* Aug. 23. – 25. 1994, Karlsruhe, Germany, pp. 147 – 161.
32. D. C. Copley, J. W. Eberhard, G. A. Mohr: "Computed Tomography Part I: Introduction and Industrial Applications", *JOM* **46** (1994) 1, 14 – 26.
33. K. Hori, T. Fujimoto, H. Nishikawa: Advanced High Speed X-Ray CT Scanner for Measurements and Visualization of Multiphase Flow, OECD/CSNI Specialist Meeting, Santa Barbara 1997.
34. H. Riesemeier, J. Goebbels, B. Illerhaus, Y. Onel, P. Reimers: "3-D Microcomputertomographie für die Werkstoffentwicklung und Bauteilprüfung", *DGZfP* **37** (1993) 280 – 287.
35. J. A. Stokes, K. R. Alvar, R. L. Corey, D. G. Corstello, J. John, S. Kocimski, N. A. Lurie, D. D. Thayer, A. P. Trippe, J. C. Young: "Some New Applications of Collimated Photon Scattering for Non-Destructive Examination", *Nucl. Instr. and Methods* **193** (1982) 261 – 267.
36. G. Harding, J. Kosanetzky: "X-ray Scatter Image in Non-Destructive Testing", *10P Short Meetings Series,* **19** (1988) 1 – 15.
37. P. Zhu, G. Peix, D. Babot, J. Muller: "X-Ray Imaging and Characterization of Materials: Compton Backscatter Techniques", in M. S. Beck et al. (eds.): *Proc. ECAPT '95, Bergen, Norway, April 6. – 8. 1995, Process Tomography – 1995 Implementation for Industrial Processes,* pp. 331 – 340.
38. D. Toye, P. Marchot, M. Crine, E. L'Homme: "The Use of Large Scale Computer Assisted Tomography for the Study of Hydrodynamics in Trickling Filters", *Chem. Eng. Sci. Vol. 49* **24b** (1994), 5271 – 5280.
39. D. Schmitz, N. Reinecke, G. S. Petritsch, D. Mewes: "High Resolution X-Ray Tomography for Statinary Multiphase Flows", *OECD/CSNI Specialist Meeting on Advanced Instrumentation and Measurement Techniques,* March 17. – 20., 1997, Santa Barbara, CA.
40. W. Ostendorf, D. Mewes, F. Mayinger: "A Tomographical Method Using Holographic Interferometry for the Registration of Three-Dimensional Unsteady Temperature Profiles in Laminar and Turbulent Flow", in C. L. Tiens (eds.): *Proc. 8th Int. Heat Transfer Conference,* vol. 2, Hemisphere Publ. Corp., New York, (1986), pp. 519 – 524.
41. Y. Zhang, C. A. Ruff: "Three-Dimensional Temperature Measurements in Enclosures by Using Multiview Interferometric Tomography", *Meas. Sci. Technol.* **5** (1994) 495 – 502.
42. F. Mayinger, D. Lübbe: "Ein tomographisches Meßverfahren und seine Anwendung auf Mischvorgänge und Stoffaustausch", *Wärme und Stoffübergang* **18** (1984) 49 – 59.
43. D. H. Burns: "Optical Tomography for Three-Dimensional Spectroscopy", *Appl. Spectroscopy* (1994) 12A – 19A.
44. H. Phillipp, E. Winkelhofer, G. K. Fraidl: "Verbrennungsdiagnostik mit Computertomographie", *MTZ* **56** (1995) 3, 148 – 157.
45. M. Kawaji, W. Ahmad, J. M. DeJesus, B. Sutharsham, C. Lorencez, M. Oiha: "Flow Visualization of Two Phases", *Nuc. Eng. Des.* **141** (1993) 343 – 355.
46. H. Yan, W. W. Yuen, T. G. Theofamous: "The Use of Flourescence in the Measurement of Local Liquid Content in Multiphase Flows", *Nuc. Eng. Des.* **141** (1993) 315 – 328.
47. H. Neumann, S. Lorenz, W. Leirer: "Infrarot-Thermographie durch Fenster bei niedrigen Temperaturen", *Wärme- und Stoffübertragung* **29** (1994) 219 – 225.
48. M. P. Mengüç, P. Dutta: "Scattering Tomography and its Application to Soothing Diffusion Flames", *Trans. ASME J. Heat Transfer,* 1994, 144 – 151.
49. T. C. Wetberg, J. J. Stamnes: "Recent Results in Quantitive Optical Diffraction Tomography", in M. S. Beck et al. (eds.): *Proc. ECAPT '95, Bergen, Norway, April 6. – 8., 1994, Processes Tomography – 1995 Implementation for Industrial Processes,* 118 – 126.
50. T. T. Charalampopoulos, R. Chang: *Agglomerate Parameters and Fractal Dimension of Soot using Light Scattering: Effects on Surface Growth, Comb and Flame,* 1991, Vol. 89 – 99 .

51. M. H. Maleki, A. J. Devaney: "Phase Retrieval and Intensity-Reconstruction Algorithms for Optical Diffraction Tomography", *J. Opt. Soc. Am.* **A10** (1993) 1086 – 1092.
52. D. S. Parker, M. R. Hawkesworth, C. J. Broadbent, P. Fowles, T. D. Fryer, P. A. McNeal: "Industrial Positron-Based Imaging: Principles and Applications", *Nucl. Instr. and Meth.* **A349** (1994) 583 – 592.
53. T. Bücherl, C. Lierse, W. Nuding, C. Sauerwein, H. Wiacker: "Bildgebende Radiometer an radioaktiven Abfallgebinden", *DGZfP* **47** (1995) 1, 329 – 336.
54. G. Daries, N. M. Spyron, I. G. Hutchinson, J. Huddleston, *Nucl. Instr. and Meth.* **A242** (1986) 615.
55. E. J. Fordham, L. D. Hall, T. S. Ramakrishnan, M. R. Sharpe, C. Hill, "Saturation Gradients in Drainage of Porous Media: NMR Imaging Measurements", *AIChE J.* **39** (1993) 1431 – 1443.
56. L. F. Gladden: "Industrial Applications of NMR Imaging", *Eng. J.* **56** (1995) 149 – 158.
57. M. A. Seiraffi: "Ultraschall-Computertomographie zur Messung von Zweiphasenströmungen", *Fortschr.-Ber. VDI* **8** (1993), 327.
58. J. Wolf, Ein Meßverfahren zur Untersuchung von Blasensäulenreaktoren. Die Ultraschalltomographie, *Fortschr.-Ber. VDI,* **8** (166), VDI-Verlag 1988.
59. D. Hiller, H. Ernest: "Tomography Using Transmission and Reflection Mod Application to Medical Diagnostics", Ac. Imag. Vol. 12, *Proc. 12th Int. Symp.,* Plenum Press, New York 1982, pp. 557 – 563.
60. S. C. Morris, A. D. Hill: "Ultrasonic Imaging and Velocimetry in Two-Phase Flow", Trans ASME, *J. Energ. Res Tech.,* **6** (1993), 115, 108 – 116.
61. S. S. Norton, M. Linzer: "Ultrasonic Reflectivity Tomography: Reconstruction with Circular Transducer Array", *Ultrasonic Imaging* **1** (1979) 154 – 184.
62. M. Moshfeghi: "Ultrasound Reflection-Mode Tomography using Fan-Shaped-Beam Insonification", *IEEE Trans. on Ultrasonics, Ferroelectrics and Frequency Control,* **UFFC-33.3** (1986) 5, 299 – 314.
63. A. Schwarz: "Acoustic Measurement of Temperature and Velocity Fields in Furnaces" in M. S. Beck et al. (eds.): *Tomographic Techniques for Process Design and Operation,* Computational Mechanics Publications, Southhampton1993, pp. 381 – 389.
64. A. R. Bidin, R. G. Green, M. E. Shackleton, A. L. Stott, R. W. Taylor: "Process Tomography using Electrodynamic Sensors", in M. S. Beck et al. (eds.): *Proc. ECAPT '93, Karlsruhe, Germany, Process Tomography – A Strategy for Industrial Exploitation – 1993.*
65. J. Cross: *Electrostatics: Principles, Problems and Applications,* A. Hilger, Bristol 1987.
66. F. Natterer: *The Mathematics of Computerized Tomography,* B. G. Teubner Verlag, Stuttgart 1986.
67. G. T. Herman: *Image Reconstruction from Projections – The Fundamentals of Computerized Tomography,* Academic Press, New York 1980.
68. A. C. Kak, M. Slaney: *Principles of Computerized Tomography Imaging,* IEEE Press, New York 1988.
69. G. T. Herman, "Iterative Reconstruction Algorithms," *Comput. Biol. Med.* **6** (1976) 273 – 294.
70. Y. Censor, "Finite Series-Expansion Reconstruction Methods," *Proc. IEEE* **71** (1983) 409 – 419.
71. A. K. Louis: *Inverse und schlechtgestellte Probleme,* B. G. Teubner Verlag, Stuttgart 1989.
72. J. C. Hunter, M. W. Collins, "Three -Dimensional Refractive Index Field Reconstruction from Holographic Interferograms," *Int. J. Optoelectronics* **4** (1989) 2, 95 – 132.
73. A. Ralston, P. Rabinowitz: *A First Course in Numerical Analysis,* Mc Graw-Hill, New York 1978.
74. A. Gordon, G. T. Herman, "Three-Dimensional Reconstruction from Projections: A Review of Algorithms," *Int. Rev. Cytol.* **38** (1974) 111 – 151.
75. M. D. Altschulter, G. T. Herman: "Fully Three–Dimensional Image Reconstruction Using Series Expansion Methods" in A. B. Brill (ed): *A Review of Information Processing in Medical Imaging,* Oak Ridge 1977, pp. 124 – 142.
76. R. B. Marr, "The Reconstruction of a Function on a Circular Domain from a Sampling of its Line Integrals," *J. Math. Anal. Appl.* **45** (1974) 357 – 374.
77. M. H. Buonocore, W. R. Brody, A. Macovski, "A Natural Pixel Decomposition for Two-Dimensional Image Reconstruction," *IDEE Trans. Biomed. Eng. BME-* **28** (1981) 2, 69 – 78.

78. R. Gordon, R. Bender G. T. Herman, "Algebraic Reconstruction Techniques (ART) for Three-Dimensional Electron Microscopy and X-Ray Photography," *J. Theor. Biol.* **29** (1970) 471 – 481.
79. P. Gilbert, "AIterative Methods for the Three-Dimensional reconstruction of an Object from Projections," *J. Theor. Biol.* **36** (1972) 105 – 117.
80. A. H. Anderson, A. C. Kak, "Simultaneous Algebraic Reconstruction Technique (SART): A Superior Implementation of the ART Algorithm," *Ultrasonic Imaging* **6** (1984) 81 – 94.
81. M. D. Young: *Iterative Solutions of Large Linear Systems,* Academic Press, New York 1971.
82. J. Radon, "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten," *Ber. Verb. Sächs. Akad. Wiss. Leipzig, Math. Phys.* **69** (1917) 262 – 277.
83. H. P. Lutz, *Fourier Image reconstruction Incorporating Three Simple Interpolation Techniques,* Tech. Rep. 104, Dept. of Comp. Sci, Sony Buffalo, 1974.
84. R. N. Bracewell, A. C. Riddle, "Inversion of Fanbeam Scans in Radio Astronomy," *Astrophys. J.* **150** (1967) 427 – 434.
85. R. M. Mersereau, "Direct Fourier Transformation Techniques in 3-D Image Reconstruction," *Comput. Biol. Med.* **6** (1976) 247 – 258.
86. S. W. Rowland: "Computer Implementation of Image Reconstruction Formulas", in G. T. Herman (ed.): *Image Reconstruction from Projections: Implementation and Application,* Springer Verlag, Berlin 1979, pp. 9 – 80.
87. L. A. Shepp, B. F. Logan, "The Fourier Reconstruction of a Head Section", *IEEE Trans. Nucl. Sci. NS-* **21** (1974) 21 – 43.
88. G. N. Ramachandran, A. V. Lakshminarayanam, "Three Dimensional Reconstruction from Radiographs and Electron Micrographs: Application of Convolutions Instead of Fourier Transforms", *Proc. Nat. Acad. Sci. U.S.* **68** (1971) 2236 – 2240.
89. A. Y. Nooralahiyan, B. S. Hoyle, N. J. Bailey, "Pattern Association and Feature Extraction in Electrical Capacitance Tomography", in M. S. Beck et al. (eds.): *Process Tomography – A Stategy for Industrial Exploitation,* UMIST, Manchester 1994, pp. 266 – 275.
90. H. S. Ko, D. P. Lyons, K. D. Kihm: A Comparative Study of Algebraic Reconstruction Technique (ART) and Generic Algorithms (GA) for Beam Deflection Tomography, *Proc. ASME Fluids Engineering Division Summer Meeting,* Vancouver 1997.
91. S. M. Huang, A. Plaskowski, C. G. Xie, M. S. Beck: "Tomographic Imaging of Two-Phase Flow using Capacitance Sensors", *J. Phys. E. Sci. Instrum.* **22** (1989) 173 – 177.
92. C. G. Xie, S. M. Huang, D. G. Hayes, I. A. Gregory, M. S. Beck: "Transputer/DSP-based Capacitance Tomography for Real-Time Imaging and Velocity Profile Measurement of Oilfield Flow Pipelines," in M. Valero et al. (eds.): *Parallel Computing and Transputer Appl.,* IOS Press, Barcelona 1992.
93. F. Klug, F. Mayinger: "Impedance Based Flow Reconstruction, A Novel Measuring Technique for Multiphase Flows," *Proc. NURETH-5,* Salt Lake City 1992, pp. 1263 – 1270.
94. S. L. McKee, R. A. Williams, T. A. Bell: "Industrial Application of Electrical Tomography to Conveying Processes", *Trans. Inst. Chem. Eng.* **71** (1993) 347 – 349.
95. K. Brodovic, L. Marynika, T. Dyakowski: "Application of Capacitance Tomography for Pneumatic Conveying Process," in M. S. Beck, R. A. Williams (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1992, 309 – 318.
96. T. Dyakowski, S. J. Wang, C. G. Xie, R. A. Williams, M. S. Beck: "Real-Time Capacitance Imaging of Bubble Formation at the Distributor of a Fluidized Bed," in M. S. Beck et al. (eds.): *Tomographic Techniques for Process Design and Operation,* Computational Mechanics Publications, Southampton 1993, pp. 333 – 346.
97. F. T. Kühn, J. C. Schouten, R. F. Mudde, C. M. van den Bleek, B. Scarlett: "Analysis of Chaos in Fluidization Using Electrical Capacitance Tomography", *Meas. Sci. Technol.* **7** (1996) 3, 361 – 368.
98. J. S. Halow, G. E. Fasching, P. Nicoletti, J. L. Spenik: "Observations of a Fluidized Bed using Capacitance Imaging", *Chem. Eng. Sci.* **4** (1993) 48, 643 – 659.
99. G. E. Fasching, N. S. Smith: "A Capacitance System for Three-Dimensional Imaging of Fluidized Beds", *Rev. Sci. Instrum.* **62** (1991) 9, 2243 – 2251.
100. R. He, C. M. Beck, R. C. Waterfall, M. S. Beck: "Development of Capacitance Measurements Towards Tomographic Imaging of Flames," in K. T. V. Grattan and A. T.

Agousty (eds.): *Sensors VI: Technology, Systems and Applications,* Iop Pub./Inst. of Physics, Bristol 1993 pp. 365 – 368.

101. P. W. Southern, R, J. Deloughry: "Imaging of Oil/Water/Sand Interface Levels in an Oil Separation Vessel", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1993, pp. 181 – 184.
102. W. Q. Yang, H. X. Wang, M. R. Brant, C. G. Xie, Y. W. Chen, R. C. Waterfall, M. S. Beck: "Multi-Interface Level Measurement System using Segmented Capacitance Sensor", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1993, pp. 185 – 188.
103. G. A. Johansen, T. Frønstein, B. T. Hjertaker, Ø. Olsen: "A Dual Sensor Flow Imaging Tomographic System", *Meas. Sci. Technol.* **7** (1996) 3, 297 – 307.
104. N. Reinecke, D. Mewes, "Tomographic of Trickle-Bed Reactors", *Chem. Eng. Sci.* **51** (1996) 10, 2131 – 2138.
105. C. G. Xie, S. M. Huang, B. S. Hoyle, R. Thorn, C. Lenn, M. S. Beck: "Process Tomography – System Model for Development of Image Reconstruction Algorithms for Capacitance Imaging", in Y. Wu et al. (eds.): *ICEMI '92,* Electronic Industrial Publishing House, 1992, pp. 105 – 112.
106. M. S. Bair, J. P. Oakley: "A Quadratic Sensor Model for the Use in Electrical Capacitance Tomography", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1993, pp. 136 – 139.
107. C. G. Xie, S. M. Huang, B. S. Hoyle, R. Thorn, C. Lenn, M. S. Beck: "Electrical Capacitance Tomography for Flow Imaging – System Model for Development of Reconstruction Algorithms and Design of Primary Sensors", *IEE Proc. G* **139** (1992) 1, 89 – 98.
108. Ø. Isaksen, J. E. Nordtvedt: "An Implicit Model based Reconstruction Algorithm for Use With a Capacitance Tomography System", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1994, pp. 215 – 226.
109. Q. Chen, B. S. Hoyle, H. J. Strangeways: "Electric Field Interaction and an Enhanced Reconstruction Algorithm in Capacitance Tomography", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1992, pp. 170 – 177.
110. A. Y. Nooralahiyan, B. S. Hoyle, N. J. Bailey: "Application of a Neural Network in Image Reconstruction for Capacitance Tomography", in M. S. Beck et al. (eds.): *Process Tomography – A Strategy for Industrial Exploitation,* UMIST, Manchester 1993, pp. 144 – 147.
111. F. Mayinger (ed.): *Optical Measurement-Techniques and Applications,* Springer-Verlag, Berlin 1994.
112. W. Harde: *Das Vermischen mit Hilfe von Flüssigkeitsstrahlen*, Dissertation, Universität Hannover, 1989
113. R. Renz: *Das Diskontinuierliche Vermischen von Flüssigkeitsstrahlen in zylindrischen Behältern,* Dissertation, Universität Hannover, 1992
114. A. Fellhölter: *Kontinuierliches Mischen von Gasen in Kanälen,* Dissertation, Universität Hannover, 1994
115. M. Buchmann, D. Mewes: "Experimental Investigation of Micro- and Macromixing with the Tomographical Absorption Spectroscopy", *IMECHE Conference: Optical Methods and Data Processing in Heat and Fluid Flow,* April 16 – 17, 1998, City University, London.
116. E. J. Peters, W. D. Hardham: "Visualization of Fluid Displacements in Porous Media using Computed Tomography Imaging", *J. Petrol. Sci. and Eng.* **4** (1990) 155 – 168.
117. S. Y. Wang, Y. B. Huang, V. Pereira, C. C. Gryte: "Applications of Computed Tomography to Oil Recovery from Porous Media", *Applied Optics* **24** (1985) 23, 4021 – 4027.
118. T. Ikeda, K. Kotani, Y. Meada, H. Kohno: Preliminary Study on Application of X-ray CT Scanner to Measurement of Void Fractions in Steady State Two-Phase Flows, *J. Nuclear Science and Technology* **20** (1983) 1, 1 – 12.
119. J. R. Fincke, M. A. Vince, C. L. Jeffrey: "Measurement of Time-averaged Density Distribution in Horizontal Multiphase Flow using Reconstructive Tomography", *Plasma Thermophysics and Heat Transfer Conference,* St. Louis, June 7. – 11., 1982.
120. A. C. De Vuono, P. A. Schlosser, F. A. Kulacki, P. Munshi: "Design of an Isotopic Scanner for Two Phase Flow Measurements", *IEEE Trans. Nucl. Sci.,* **NS-27,** (1980) 814 – 820.
121. A. Kantzas: "Computation of Holdups in Fluidized and Trickle Beds by

Computer-Assisted Tomography", *AIChE J.* **40** (1994) 7, 1254 – 1261.

122. P. G. Lutran, K. M. Ng, E. P. Delikat: "Liquid Distribution in Trickle Beds: An Experimental Study Using Computer Assisted Tomography", *Ind. Eng. Chem. Res.* **30** (1991) 1270 – 1280.
123. O. Feldmann, P. Gebhard, F. Mayinger: "Evaluation of Pulsed Laser Holograms of Flashing Sprays by Digital Image and Holographic Particle Image Velocimetry," *Proc. OECD/CSNI Specialist Meeting on Advanced Instrumentation and Measurement Techniques,* Santa Barbara, March 17 – 20, 1997.
124. Y. A. Hassan: "Multiparticle Imaging Velocimetry Measurements in Two-Phase Flow," *Proc. OECD/CSNI Specialist Meeting on Advanced Instrumentation and Measurement Techniques,* Santa Barbara, March 17 – 20, 1997.
125. L. Gui, W. Merzkirch, R. Lindken: "Phase-separated PIV Measurements of the Flow Around Systems of Bubbles Rising in Water," *Proc. ASME Fluid Engineering Division Summer Meeting,* Vancouver, June 22 – 26, 1997.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Water

Standard Article

Hermann Weingärtner[1], Ernst Ulrich Franck[2], Gabriele Wiegand[3], Nicolaus Dahmen[4], Georg Schwedt[5], Fritz H. Frimmel[6], Birgit C. Gordalla[7], Klaus Johannsen[8], R. Scott Summers[9], Wolfgang Höll[10], Martin Jekel[11], Rolf Gimbel[12], Robert Rautenbach[13], William H. Glaze[14]

[1]Ruhr-Universität Bochum, Physikalische Chemie II, Bochum, Federal Republic of Germany

[2]Institut für Heiße Chemie, Forschungszentrum Karlsruhe, Karlsruhe, Federal Republic of Germany

[3]Institut für Heiße Chemie, Forschungszentrum Karlsruhe, Karlsruhe, Federal Republic of Germany

[4]Institut für Heiße Chemie, Forschungszentrum Karlsruhe, Karlsruhe, Federal Republic of Germany

[5]Institut für Anorganische und Analytische Chemie, Technische Universität Clausthal, Clausthal-Zellerfeld, Federal Republic of Germany

[6]Engler-Bunte-Institut der Universität Karlsruhe, Bereich Wasserchemie, Karlsruhe, Federal Republic of Germany

[7]Engler-Bunte-Institut der Universität Karlsruhe, Bereich Wasserchemie, Karlsruhe, Federal Republic of Germany

[8]Technische Universität Hamburg-Harburg, Arbeitsbereich Wasserwirtschaft und Wasserversorgung, Hamburg, Federal Republic of Germany

[9]University of Cincinnati, Department of Civil Engineering, Cincinnati, Ohio, United States

[10]Forschungszentrum Karlsruhe, Institut für Technische Chemie, Bereich Wasser- und Gastechnologie, Karlsruhe, Federal Republic of Germany

[11]Technische Universität Berlin, Fachgebiet Wasserreinhaltung, Berlin, Federal Republic of Germany

[12]Gerhard-Mercator-Universität Gesamthochschule Duisburg und Rheinisch-Westfälisches Institut für Wasserchemie und Wassertechnologie, Mülheim, Federal Republic of Germany

[13]Rheinisch-Westfälische Technische Hochschule, Institut für Verfahrenstechnik, Aachen, Federal Republic of Germany

[14]Department of Environmental Sciences and Engineering, School of Public Health, University of North Carolina, Chapel Hill, NC, United States

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a28_001
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Water as a Solvent
1.1. Properties of Pure Water
1.1.1. Molecular Properties
1.1.2. Hydrogen Bonding and Water Structure
1.1.3. Bulk Properties of Liquid Water
1.2. Aqueous Solutions
1.2.1. General Solvent Properties
1.2.2. Solutions of Simple Nonpolar Gases
1.2.3. Solutes with Hydrophilic Groups
1.2.4. Electrolyte Solutions
2. Water at High Pressure and Temperature
2.1. Properties of Ice
2.1.1. Thermodynamic Properties
2.1.2. Transport Properties

2.1.3. Electrolytic Properties
2.1.4. Mechanical Properties
2.2. Properties of Fluid Water
2.2.1. Thermodynamic Properties
2.2.2. Transport Properties
2.2.3. Electrolytic Properties
2.2.4. Other Physical Properties
2.3. Properties of Water Vapor
2.4. Water in the Supercritical State
3. Water Analysis
3.1. Sampling and Sample Preservation
3.2. Physicochemical and Sum Parameters
3.3. Inorganic Analysis
3.3.1. Determination of Cations
3.3.2. Anion Analysis
3.3.3. Determination of Dissolved Gases
3.3.4. Quick Test Processes
3.4. Organic Analysis
3.4.1. Spectrometric Methods
3.4.2. Gas and Liquid Chromatography
3.5. Biochemical Methods
4. Hydrological Cycle and Water Use
4.1. World Water Balance
4.2. Hydrological Cycle
4.3. Demand for Water
4.4. Source of Water Used
4.5. Water Treatment
5. Adsorption Processes in Water Treatment
5.1. Introduction
5.2. Properties of Activated Carbon
5.3. Adsorption Theory
5.4. Adsorption Equilibrium
5.5. Adsorption Kinetics
5.6. Biological Processes
5.7. Design of GAC Systems
5.8. Design of PAC Systems
5.9. Reactivation of GAC
6. Ion Exchange
6.1. Fundamentals
6.1.1. Characterization of the Chemical State of Water
6.1.2. Representation of the Chemical State of Water
6.1.3. Ion Exchangers
6.2. Cation-Exchange Processes
6.2.1. Strongly Acidic Resins in the Na^+ Form
6.2.2. Strongly Acidic Resins in the H^+ Form
6.2.3. Weakly Acidic Resins in the Hydrogen Form
6.3. Anion-Exchange Processes
6.3.1. Strongly Basic Resins in the Cl^- Form
6.3.2. Strongly Basic Resins in the OH^- Form
6.3.3. Strongly Basic Resins in the HCO_3^- Form
6.3.4. Weakly Basic Resins
6.4. Combined Processes
6.4.1. Demineralization
6.4.2. Partial Demineralization
7. Flocculation
7.1. Introduction
7.2. Definitions
7.3. Chemicals Used in Coagulation and Flocculation

7.3.1. Inorganic Chemicals
7.3.2. Organic Chemicals
7.4. Kinetics of Aggregation
7.5. Process Technology for Coagulation and Flocculation
7.5.1. Dosing
7.5.2. Rapid Mixing
7.5.3. Reactors for Floc Formation
7.5.4. Integrated Flocculation Systems
7.5.5. Operational Aspects
8. Filtration
8.1. Introduction
8.2. Slow Sand Filters
8.3. Polishing Filters
8.4. Rapid Filters
8.4.1. Design
8.4.2. Areas of Application and Modes of Operation
8.4.3. Filter Performance
9. Membrane Separation Processes in Water Treatment
9.1. Principles
9.1.1. Flow, Selectivity, Driving Forces
9.1.2. Mass Transport Resistance in Front of the Membrane
9.2. Membranes
9.2.1. Organic Membranes
9.2.2. Inorganic Membranes
9.3. Modules
9.3.1. Modules with Tubular Membranes
9.3.2. Modules with a Flat-Sheet Membrane
9.4. Fouling and Scaling
9.4.1. Membrane Blockage due to Crystallization (Scaling)
9.4.2. Membrane Blockage due to Contaminants (Fouling)
9.5. Module Arrangements (Plant Design)
9.6. Use of Membrane Processes in Water and Wastewater Technology
9.6.1. Desalination of Seawater and Brackish Water
9.6.2. Recovery of ∈-Caprolactam
9.6.3. Leachates from Dump Sites
9.6.4. Wastewater from the Dye Industry
10. Oxidation Processes in Water Treatment
10.1. Physical and Chemical Properties of Chemical Oxidants
10.2. Production of Chemical Oxidants
10.3. Uses of Chemical Oxidants
10.3.1. Drinking Water Treatment
10.3.1.1. Disinfection
10.3.1.2. Chemical Oxidation
10.3.2. Wastewater Treatment
10.3.3. Oxidation in Pulp and Paper Treatment
10.3.4. Advanced Oxidation Processes
10.4. Toxicology and Environmental Health
11. Acknowledgement

[Top of Page]

1. Water as a Solvent

Hermann Weingärtner

1.1. Properties of Pure Water

1.1.1. *Molecular Properties*

In comparison to other solvents, water and aqueous solutions in many respects show unique properties that are highly sensitive to temperature and pressure [1-3]. These properties result from the structure and the charge distribution of the water molecule. In a simple four-point-charge model, this charge distribution is depicted as a quadrupole with two protons and two lone electron pairs at the corners of a tetrahedron with oxygen in its center (Fig. 1). In the isolated molecule the O–H

internuclear distance is 0.096 nm, with an H–O–H bond angle of 104.3°, which is close to the tetrahedral angle [2], [3]. This nonspherical charge distribution gives rise to electric moments. The dipole moment of the isolated water molecule is 1.85 D (6.14×10^{-30} C · m), the dipole vector acting along the bisector of the H–O–H angle with the negative end directed toward the oxygen. Thus, water is a highly polar solvent.

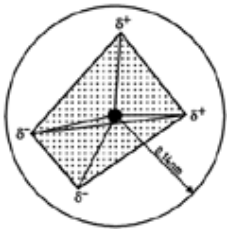

Figure 1. Four-point-charge model of water after [4]. The partial charges δ^+ refer to the protons, the charges δ^- to the lone electron pairs. The experimental dipole moment is reproduced with $\delta^+ = \delta^- = 0.236\ e$, where e is the elementary charge.

When considering intermolecular interactions, the electrostatic forces have to be supplemented by terms accounting for the core repulsion of the molecules and the attractive dispersion (van der Waals) forces. For water, these terms are often approximated by the corresponding figures for the isoelectronic neon, which has a core radius of 0.14 nm. In computer simulations, models of this type are able to reproduce the major features of the properties of liquid water [4].

1.1.2. *Hydrogen Bonding and Water Structure*

The molecular geometry and charge distribution of water favor the formation of a hydrogen-bonded network with four O–H–O bonds per molecule. Hydrogen bonding can be proved by infrared and Raman spectroscopy. Three normal vibrations of the isolated H_2O molecule occur in the vapor phase: the symmetric stretching mode at a wave number $\bar{\nu}$ = 3675 cm^{-1}, the asymmetric stretching mode at 3756 cm^{-1}, and the bending mode at 1595 cm^{-1} (Fig. 2). Hydrogen bonding leads to shifts in the wave numbers of these modes, a change in their band contours, and the appearance of new bands associated with intermolecular vibrations [3]. These spectral features are often used to monitor changes in hydrogen bonding as a function of temperature and pressure or induced by added solutes. In addition, the chemical shift in the ^1H NMR spectrum of water provides a sensitive tool for measurement of hydrogen bonding.

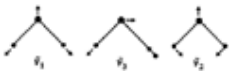

Figure 2. Normal vibrations of the water molecule: $\bar{\nu}_1$ is the symmetric stretching mode, $\bar{\nu}_3$ the antisymmetric stretching mode, and $\bar{\nu}_2$ the bending mode

The structure of *hexagonal ice* Ih (Fig. 3), which is the stable polymorph of ice at atmospheric pressure, is an ideal representation of this hydrogen-bonded network [3]: Each oxygen is surrounded tetrahedrally by four others at a distance of 0.276 nm, the structure being isomorphous with the tridymite structure of silica. The protons lie approximately along the O–O axes. Because of hydrogen bonding, the intramolecular OH distance (0.101 nm) is greater than in the isolated molecule. If compared with closely packed structures with up to 12 nearest neighbors, the ice Ih structure is very open. The large molar volume of ice Ih is, for example, responsible for the volume contraction of ice upon melting.

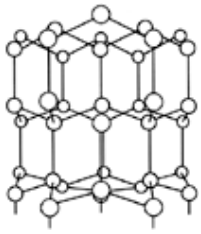

Figure 3. Structure of hexagonal ice Ih. The spheres give the positions of the oxygen atoms.

In the *liquid state* the long-range order is lost, but short-range positional and orientational correlations are retained, which can be studied by X-ray and neutron scattering [3]. At room temperature a structure is perceptible in the local environment, which extends to about 0.8 nm. On average, one molecule has 4.4 nearest neighbors at a distance of 0.284 nm. These figures and the location of second-nearest neighbors correspond closely to the requirements imposed by the ice Ih structure. However, intermolecular O–O and O–H distances and H–O–H angles show a broad distribution around the average value, and due to thermal reorientation, the lifetime of a given configuration is only of the order of some picoseconds. With increasing temperature, the three-dimensional network gradually breaks down. Thus, with rising temperature, water loses its unique properties, although Raman spectroscopy shows that even at the critical point at 374 °C, some degree of hydrogen bonding is still retained [5].

Any model of liquid water must accommodate these features, but none of the existing models seems to be immune to criticism. For a long time, so-called mixture models have been used to explain the properties of water and aqueous solutions. These models assume the existence of two (or more) physically distinguishable water species, corresponding to molecules with “broken” and “unbroken” hydrogen bonds [2], [3], according to equilibria of the general form H_2O (icelike) $\rightleftharpoons$ H_2O (gaslike), which depend on temperature and pressure and are changed by solutes. From the present point of view, the need for such models is overcome largely by the possibility of performing more rigorous model-free analyses based on statistical mechanical methods and computer simulations, so that many of these models are only of historical interest.

1.1.3. *Bulk Properties of Liquid Water*

A knowledge of the bulk properties of pure water and their anomalous behavior is a prerequisite for understanding its solvent properties, because many methods and rules for evaluating and predicting data in nonaqueous systems are of little value for aqueous solutions. Most of these anomalies occur near room temperature or up to 100 °C. The significance of these anomalies for processes in nature and in sustaining life has been outlined [3]. Doubtless, these anomalies result from the strong hydrogen bonding that occurs, however, in many other substances as well. Hence, the existence of strong hydrogen bonds cannot fully explain the uniqueness of some properties of water. Rather, it is the existence of the three-dimensional network that distinguishes water from other hydrogen-bonded liquids. Despite great progress in liquid-state theory, the treatment of such hydrogen-bonded systems remains difficult.

Isotopic Composition. Natural ("ordinary") water is a mixture of isotopic species made up by the hydrogen isotopes ^{1}H and ^{2}H (deuterium) and the oxygen isotopes ^{16}O, ^{17}O, and ^{18}O. Their masses and natural abundances are compiled in Table 1. Proton exchange equilibria of the form $^{1}H_2O + {}^{2}H_2O \rightleftharpoons 2\,{}^{1}H{}^{2}HO$ yield further mixed species. The ^{1}H and ^{2}H are usually assumed to be distributed randomly on the oxygens. Then, the four major species of ordinary water occur in the proportions [6]

Table 1. Isotopic masses and natural abundances of hydrogen and oxygen isotopes [11]

	Atomic mass	Natural abundance	
		Mol %	wt %
H (natural)	1.0080		
^{1}H	1.00783	99.985	99.970
^{2}H	2.01410	0.015	0.030
O (natural)	15.9994		
^{16}O	15.9995	99.759	99.732
^{17}O	16.9991	0.037	0.039
^{18}O	17.9992	0.204	0.229

$$^{1}H_2{}^{16}O : {}^{1}H_2{}^{18}O : {}^{1}H_2{}^{17}O : {}^{1}H_2{}^{2}H_2{}^{16}O = 997\,290 : 2000 : 400 : 320$$

The molar mass of ordinary water is 18.012 g/mol. The variation of these proportions in water from different natural sources is usually smaller than variations as a result of water purification, and except in work of the highest accuracy, the properties of ordinary water can be set equal to the properties of $^{1}H_2{}^{16}O$. Practically pure $^{2}H_2O$ (heavy water) and highly enriched ^{17}O- and ^{18}O-labeled water are available commercially, so that some data also exist for water of other isotopic compositions [7]. In the following, the term "water" is used for ordinary water with natural isotopic composition.

p – V – T Data and Invariant Points. Water is the only inorganic substance that, under ambient conditions found in nature, occurs in the solid, liquid, and gaseous states. Its state of aggregation is characterized by the pressure p, the temperature T, and volume V (or density ρ). Only the $p - V - T$ properties of the liquid phase are discussed here. In most applications, the properties of the liquid along the saturation curve can be set equal to the properties at atmospheric pressure.

The *liquid range* is limited by the triple point at which solid ice Ih, liquid water, and steam are in equilibrium and by the critical point. The triple point of ordinary water is bound by the definition of the temperature scale to 273.16 K = 0.01 °C. The normal melting and boiling points at 1.01325 bar are 0 and 100 °C, respectively. The triple-point pressure is $p_t = 6.113\times10^2$ Pa. Recommended values for the *critical temperature, critical pressure*, and *critical density* are [8] t_c = 373.98 °C, p_c = 22.05 MPa, and ρ_c = 322 kg/m^3. Table 2 lists volumetric properties of the liquid from 0 to 100 °C [8], [9]. Here, ρ is the mass density; $\alpha_p = (1/V)(\partial V/\partial T)_p$, the thermal expansion coefficient; and $\chi_T = (1/V)(\partial V/\partial p)_T$, the isothermal compressibility. The thermal pressure coefficient $\gamma_V = (\partial p/\partial T)_V$ can be calculated by $\gamma_V = \alpha_p/\chi_T$.

Table 2. Density ρ, thermal expansion coefficient α_p, and isothermal compressibility χ_T of ordinary water at atmospheric pressure [11]

t, °C	ρ, kg/m^3	$\alpha_p\times10^6$, K^{-1}	χ_T, bar^{-1}
0	999.84	–68.14	50.88
10	999.70	87.90	47.82
20	998.20	206.6	45.90
25	997.05	257.1	45.25
30	995.65	303.1	44.77
	992.23	385.4	44.24

40			
50	988.03	457.8	44.18
60	983.19	523.4	44.51
70	977.76	584.0	45.17
80	971.79	641.3	46.15
90	965.31	696.3	47.43
100	958.36	750.0	49.02

At atmospheric pressure, *ice Ih* contracts by ca. 8.2 % on melting — from a density of 916.8 kg/m^3 to 999.87 kg/m^3 in the liquid. The slope of the melting pressure curve depends on the volume change on melting. The anomalous volume contraction of water leads to a depression of its melting point with rising pressure. This depression continues up to 207 MPa and –22 °C, where the ice polymorph changes to ice II. Close to 0 °C, a pressure of 133 bar is required to lower the melting point by 1 °C.

The volume contraction on melting continues in the liquid phase up to 3.98 °C, where the density reaches a maximum value and the thermal expansion coefficient changes sign from negative to positive. Over the entire temperature range considered, the thermal expansion coefficient of water is one order of magnitude smaller than that of other liquids. The isothermal compressibility is unusually large and possesses a minimum near 46.5 °C. Both extremes represent anomalous properties of pure water that tend to disappear at high pressure.

Actually, liquid water may be supercooled below the equilibrium melting point [10]. Experiments can be performed fairly easily down to –20 °C, which is roughly the limit of heterogeneous nucleation. With special techniques, temperatures as low as –40 °C have been achieved with liquid water. At 1 bar, homogeneous nucleation sets a limit at –43 °C. Many of the anomalous properties of liquid water become even more unusual in the supercooled, metastable regime [10].

Specific Heat. At the normal melting point, the isobaric specific heat c_p increases from 2.072 J K^{-1} g^{-1} in the solid to 4.228 J K^{-1} g^{-1} in the liquid state, where it remains almost unchanged up to 100 °C [8], [9] (a very shallow minimum is found near 35 °C) (Table 3). The specific heat is particularly amenable to molecular interpretation and has often been used to estimate the extent of hydrogen bonding in water [2], [3].

Table 3. Isobaric specific heat c_p, vapor pressure p, latent heat of vaporization $\Delta_l^g h$, and surface tension σ of water at saturation pressure

t, °C	c_p, J g^{-1} K^{-1}	p, bar	$\Delta_l^g h$, kJ/kg	$\sigma \times 10^3$, N/m
0	4.217	0.00611	2500.5	75.64
10	4.191	0.01228	2467.9	74.23
20	4.182	0.02338	2453.4	72.75
30	4.178	0.04245	2429.6	71.20
40	4.178	0.07382	2405.9	69.60
50	4.180	0.12346	2309.9	67.94
60	4.184	0.19936	2357.7	66.24
70	4.189	0.31181	2333.1	64.47
80	4.196	0.47379	2308.2	62.67
90	4.205	0.70123	2282.7	60.82
100	4.216	1.01325	2256.6	58.91

Liquid – Vapor Equilibria. The vapor pressure at saturation is given in Table 3 [9]. The slope of the vapor pressure curve yields the latent heat of vaporization $\Delta_l^g h$ [8], [9], which is ca. 2500 kJ/kg at the normal melting point. Up to 100 °C, a slight decrease in $\Delta_l^g h$ is observed (Table 3). Compared with other liquids, the latent heat of vaporization is anomalously high, and its values have often been used to estimate the energy of hydrogen bonds in water. Based on comparison with data for similar molecules, the heat of vaporization of non-hydrogen-bonded H_2O should be ca. 550 kJ/kg, about one-fourth of the actual values between 0 and 100 °C. The difference is ascribed to the energy of the hydrogen bonds.

The vapor pressure is well investigated for other isotopic compositions — above all heavy water. Vapor pressure data for other isotopic forms usually are represented by the partition coefficient $\alpha = p_x / p_{H_2O}$, where H_2O represents light water and x the heavier isotopic form. Figure 4 shows the partition coefficient of heavy water as a function of temperature [7]. Over a

large range of temperature, heavy water is less volatile than light water, so its melting and boiling points (3.82 and 101.42 °C) are higher than those of ^1H_2O. The inverse is true above 200 °C, and the critical point of ^2H_2O is about 3 °C lower.

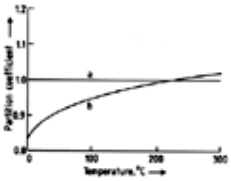

Figure 4. Partition coefficients α of light (a) and heavy water (b) as a function of temperature

Surface Tension. Below 100 °C, the surface tension σ [8], [9] (Table 3) refers to the interface between water saturated with air and air saturated with water. It is distinctly higher than that of other liquids, including most hydrogen-bonded ones. On the basis of simple theories, the quantity $\sigma/\varrho^{2/3}$ should be a function of the temperature distance $T - T_c$ from the critical point. This is observed approximately, but for water the temperature function of $\sigma/\varrho^{2/3}$ does not exhibit the predicted universal slope found with most other liquids.

Transport Properties. The *dynamic viscosity* η of liquid water (Table 4) is known with great accuracy [6], [8], [9], and the data are often used to calibrate viscometers. The dynamic viscosity is higher than that of comparable substances without hydrogen bonds. The decrease of dynamic viscosity with *increasing temperature* cannot be described with sufficient accuracy by the conventional Arrhenius-type expression with two adjustable constants η_0 and E_η

$$\eta = \eta_0 \exp\left(E_\eta / RT\right)$$

unless the energy of the viscous flow E_η is assumed to be temperature dependent. In other words, the plot of the logarithm of viscosity versus the inverse temperature does not yield a straight line, as found for many other liquids. More complex expressions are therefore in use [10]. The temperature dependence of viscosity is of some interest in interpreting processes in supercooled water [10]. The *pressure dependence* of the dynamic viscosity of water is anomalous as well: Whereas the viscosities of liquids usually increase with increasing pressure, the viscosity of water initially decreases up to a minimum at ca. 60 MPa. This decrease has clear structural implications: the hydrogen-bonded network becomes deformed with rising pressure, and this tends to facilitate molecular motion. An unusually large increase of viscosity (ca. 23 % at 25 °C) is found when ordinary water is replaced by heavy water. At higher temperature, this isotope effect diminishes. This unusually large effect is attributed to a higher structuredness of heavy water.

Table 4. Viscosity η and thermal conductivity λ of liquid water at atmospheric pressure

t, °C	$\eta \times 10^6$, Pa · s	λ, mW K^{-1} m^{-1}
0	1792	565
10	1307	584
20	1002	602
25	890.3	
30	797.5	617
40	653.2	631
50	547.1	642
60	466.6	652
70	403.9	660
80	353.8	669
90	312.8	675
100	278.3	679

In gases, the *thermal conductivity* λ is proportional to viscosity, but this correlation is lost in the liquid phase, where for most liquids a linear decrease is found with increasing temperature. In contrast, water shows an anomalous increase in λ up to ca. 130 °C [6], [8], [9] (Table 4). Despite its considerable technical importance, understanding of the thermal conductance of liquids on a molecular basis is poorly developed.

The *self-diffusion coefficient D* characterizes the mean-square displacement of water molecules due to thermal agitation. In ordinary water at 25 °C, $D = 2.30 \times 10^9$ m^2/s [11], corresponding to a displacement of the order of 10^{-4} m during 1 s. The temperature and pressure dependence of *D* and the isotope effects are correlated strongly with those of the reciprocal viscosity.

Autoprotolysis. Self-dissociation of water is characterized by the ion product $K_w = a_{H^+} a_{OH^-}$ or its negative decadic

logarithm. Here a_{H^+} and a_{H^-} are the activities of the H^+ and OH^- ions. At 25 °C, $K_w = 1.008\times10^{-14}$. Accordingly, the pH value of pure water is $pH = -\log a_{H^+} = 7$. The ion product increases with increasing temperature and increasing pressure [12] (see Section Electrolytic Properties). Autoprotolysis is of paramount importance for the Brønsted acid – base concept. According to this scheme, water is classified as an amphiprotic solvent that, depending on the solute, exhibits acidic or basic properties.

Electric Properties. Water exhibits a nonvanishing *electrical conductance*. Values of the specific conductance κ as low as $3\times 10^{-8}\ \Omega^{-1}\ cm^{-1}$ have been achieved, but in practice, water with $\kappa \approx 1\times10^{-6}\ \Omega^{-1}\ cm^{-1}$ is used even in accurate experiments. The discrepancy results from impurities not removable by simple distillation or ion exchange and from dissolved carbon dioxide.

The *dielectric constant* (relative permittivity) ϵ of water (ϵ = 87.81 at 0 °C) is very high and decreases substantially with increasing temperature to 55.44 at 100 °C [13]. The value at 25 °C is 78.46. These high values are attributed to a cooperative alignment of molecules in the hydrogen-bonded network. The dielectric constant is a key quantity for interpreting solute – solvent interactions, above all in electrolyte solutions. Its change with temperature and pressure gives rise to drastic changes in the solvent properties of water.

Other Properties.
Refractive Index. At 25 °C, the refractive index of water is 1.3340. It decreases slightly with increasing temperature.

Standard Enthalpy and Entropy of Formation. The standard enthalpy of formation of water at 25 °C and 1 bar is –286.2 kJ/mol, and the corresponding entropy is 69.98 J $mol^{-1}\ K^{-1}$ [14].

Cryoscopic and Ebullioscopic Constants. The cryoscopic constant K_f = 1.853 g $K^{-1}\ mol^{-1}$ characterizes the freezing-point depression caused by 1 mol dissolved species in 1000 g H_2O. The ebullioscopic constant K_b = 0.515 g $K^{-1}\ mol^{-1}$ characterizes the boiling-point elevation caused by 1 mol dissolved species in 1000 g H_2O.

1.2. Aqueous Solutions

1.2.1. *General Solvent Properties*

At ambient conditions, water is a good solvent for many ionic and polar inorganic and organic substances. To some degree, it can even accommodate nonpolar particles and larger nonpolar molecular groups in the interstitial voids of its hydrogen-bonded network. At high temperature the solvent behavior of water can change drastically. Many comparatively inert substances that mix only poorly with water at room temperature become highly soluble or even completely miscible. In contrast, due to the decreasing dielectric constant, water loses its favorable characteristics as a solvent for ionic compounds.

A useful classification distinguishes among three types of hydration, i.e., ion hydration; hydration of organics with polar functional groups such as OH, C=O, and NH; and “hydrophobic” hydration of inert particles and substances with nonpolar molecular groups. Of these, ion hydration has traditionally received the most interest and is comparatively well characterized. The hydration of substances containing polar functional groups is determined by the extent of solute – solvent hydrogen bonding. In such cases, a highly complex and specific behavior is found. Hydrophobic hydration is a nonspecific “nonbonded” interaction that depends on whether water can build up a stable structure around the inert particle, thereby retaining or strengthening its hydrogen bonds. Hydrophobic effects are therefore determined by the nature of water – water interactions rather than by water – solute interactions, and are driven by a favorable entropy change rather than by energetic contributions. Such processes are very sensitive to temperature, and this strong temperature dependence is one of the essential characteristics of hydrophobic hydration.

1.2.2. *Solutions of Simple Nonpolar Gases*

The solubility of simple low-polarity gases (noble gases, aliphatic hydrocarbons such as methane or ethane, or inert inorganic compounds such as SF_6) in water at 25 °C is lower than in most organic liquids [15] (Fig. 5). Moreover, in contrast to the normal behavior found with other solvents, the solubility decreases as a function of temperature. Only at high temperature (e.g., >100 °C) does the solubility pass through a minimum and then have a positive temperature coefficient like that of normal solvents [16]. Extensive tabulations of gas solubilities in water are available [17].

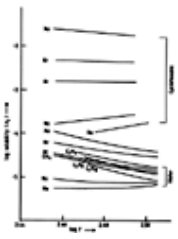

Figure 5. Solubility of simple gases in water and cyclohexane; x_2 is the mole fraction of the gas (reproduced from [21] by permission)

The thermodynamics of dissolution of inert gases in water and the structural properties of such systems have received wide attention [18], [19], because they are often used to model the hydration of nonpolar groups in more complex molecules. The same interactions play an important role in determining the properties of colloidal and macromolecular solutions, and in the manner in which proteins acquire their native structure [19], [20]. The various properties may be investigated by considering the transfer of 1 mol inert solute from a nonaqueous to an aqueous environment. The latter solutions are nonideal in the sense that the changes in volume and enthalpy on dissolution are negative. However, this is counteracted by a very large negative change of the excess entropy over the entropy of ideal mixing. This dominance of the entropy over the enthalpy term leads to a strongly positive free energy of mixing and, hence, to the low solubility of inert gases in water.

FRANK and EVANS [21] have explained this process by an ordering of water molecules and an increase in the degree of hydrogen bonding around the inert solute. This assumption is supported mainly by the existence of solid hydrates of gases and of other small molecules with definite cage structures (clathrates) [22]. In liquid water, the environment of inert gases and inert molecular groups such as the methyl group are now believed to exhibit some residues of the clathrate structure.

At higher solute concentrations, a second important effect occurs: so-called *hydrophobic interactions* or *hydrophobic bonding* [18-20]. This term denotes the formation of close-encounter configurations between two inert solutes or molecular groups that are stabilized by a favorable common water structure around the pair (Fig. 6). With larger solutes the configuration shown in Figure 6 may, for example, correspond to the encounter of two methyl groups or larger side chains accompanied by a decrease in the number of nearest water molecules. Hydrophobic bonding may include not only configurations with direct solute contact but also configurations in which one layer of water molecules is located between the hydrophobic pair. Since this association process annihilates solute – water contacts, the association is analogous to that expected if an inert solute is removed from an aqueous environment. Thus, the thermodynamic parameters characterizing hydrophobic bonding have signs opposite to those discussed for hydrophobic hydration.

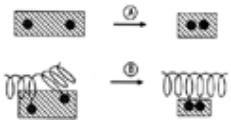

Figure 6. Schematic representation of a hydrophobic association process

A) Two methane molecules form a dimer; B) Stabilization of the conformation of a biopolymer by hydrophobic bonding between two methyl groups or side chains

The spheres represent methyl groups or side chains. The shaded area indicates that the processes occur only in an aqueous environment.

1.2.3. *Solutes with Hydrophilic Groups*

The hydration of polar groups is highly specific, and with molecules carrying both polar groups and inert residues, distinguishing between the various effects is usually difficult. Some insight is often obtained by considering the variation of solution properties within homologous series. Typical examples of such systems include aqueous solutions of alcohols, amines, ketones, ethers, fatty acids, nitriles, or polar gases.

A useful classification that may establish some order in this confusing field is based on the strength and number of proton-accepting groups of the nonaqueous interaction partner [23]. For example, nitriles and ketones are weak proton acceptors, and mixing with water is endothermic. In combination with the entropic term, such solutions are slightly stable thermodynamically and liquid – liquid immiscibilities with upper consolute points are usually found, except for the lowest members of the homologous series. If the solute is capable of accepting stronger hydrogen bonds, the mixing process may become exothermic over some temperature ranges, and a tendency exists for a lower consolute point or a closed miscibility gap. Such behavior is often found with alcohols, ethers, or amines. Finally, in some mixtures, hydrogen bonding is very strong. These are thermodynamically very stable and never separate into two liquid phases. Examples are mixtures with polyhydric alcohols such as glycerol and amides.

Even though knowledge of the hydration of comparatively simple systems is imperfect, the results are of major importance for interpreting the role of water in more complex systems of biological interest. There, hydrophobic hydration and hydrophobic interactions have gained wide attention because these nonbonded interactions play a major role in determining the stability of conformations of biomolecules (e.g., proteins) in solution [19], [20].

1.2.4. *Electrolyte Solutions*

Traditionally, electrolyte solutions have received far more attention than nonelectrolyte systems [24]. The dissolution of a salt in a solvent may be visualized by two steps consisting of (1) dissociation of the salt into gaseous ions and (2) solvation of these ions in the solvent. Consequently, to achieve dissolution, the free energy of solvation must overcome the lattice energy. The polarity and high dielectric constant of water near room temperature favor this process. This, at least qualitatively, explains why the solubility of many salts is higher in water than in nonaqueous solvents. The resulting heats of solution are low, compared to the lattice energies, and in total, the solution process can be exothermic or endothermic.

The dissolution of ions changes the water structure, because the electric field associated with the ionic charge favors definite orientation of the water molecules near the ions. However, other effects also must be accounted for, such as ion – water hydrogen bonds and conditions imposed by the need to form hydrogen bonds with water molecules beyond the first sphere. For anions, most experiments, including neutron-scattering studies [25], give evidence that a hydrogen-bonded configuration in which one of the water protons is directed toward the anion (Fig. 7, right) is preferred to that favored by electrostatic arguments (Fig. 7, left).

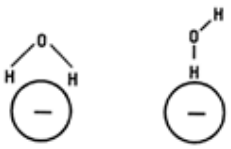

Figure 7. Possible configurations of an anion – water pair

With larger univalent ions such as K^+ or I^-, even the water configuration in the first hydration sphere may be quite isotropic. This leads to the *structure-breaking effect*: On dissolution of the salt, water behaves as if its temperature were increased, corresponding to a decrease in the solvent structure. This picture is derived from thermodynamic data such as hydration entropies [21], transport coefficients such as viscosity [24], and spectroscopic data [26], [27]. An illustrative example of the quite unusual effects obtained by the addition of structure-breaking salts is the concentration dependence of viscosity at room temperature. For salts such as potassium iodide, the viscosity decreases with increasing salt concentration, in contrast

to the usual expectations. Salts that follow this pattern are termed structure breakers as opposed to structure formers. This structure-breaking effect exhibited by alkali-metal ions (except Li^+) and halide anions (except F^-) constitutes an anomalous property of aqueous solutions. Multivalent cations and most multivalent anions are *structure formers.*

Because of the high dielectric constant of water near room temperature, most salts of univalent ions are dissociated completely even at high concentration. At high dilution of the salt, the Debye – Hückel theory of electrolyte thermodynamics and its extensions to transport processes provide a correct description of such systems [24]. In the strict sense, these are, however, limiting laws for vanishing salt concentration, as exemplified by the famous Debye – Hückel limiting law for the mean ionic activity coefficient, the logarithm of which exhibits a square-root dependence on salt concentration [24].

Progress in statistical mechanics has also led to an understanding of concentrated electrolyte solutions. Although such theories are notoriously complex, PITZER and coworkers have derived rather simple semiempirical equations for electrolyte thermodynamics that fit, within experimental accuracy, a variety of data on pure and mixed electrolytes in aqueous solutions over wide ranges of salt concentration and temperature [28].

At elevated temperature, where the dielectric constant of water is low, extensive ion pairing may occur [16]. The same is true for electrolytes of higher valence such as sulfates of alkali, alkaline-earth, and transition-metal ions, because the Coulombic interactions between the ions are stronger than in the case of unielectrolytes [24]. In addition, very specific association can be observed, for example, with zinc, cadmium, or mercury halides, where stable complexes are formed [24]. To a certain degree, such ion pairing can be incorporated in PITZER's approach. A more common way, however, is a description in terms of equilibria between distinct ionic species, which are then treated by a simple law of mass action and the corresponding thermodynamic relations [24].

[Top of Page]

2. Water at High Pressure and Temperature

Ernst Ulrich Franck, Gabriele Wiegand, Nicolaus Dahmen

Water plays a part in many technical processes and is used in a wide range of temperatures and pressures. The critical point above which liquid phase and vapor phase no longer coexist is determined by the critical data t_c = 373.98 °C, p_c = 22.05 MPa, and ρ_c = 322 kg/m^3 (see Section Bulk Properties of Liquid Water).

2.1. Properties of Ice

2.1.1. *Thermodynamic Properties*

Twelve modifications of ice are known [29-31]. The phase diagram of ice in the *p* – *T* projection in the temperature ranges –100 to +600 °C and pressures up to >10^4 MPa is shown in Figure 8 [32], [33]. The densities at 500 °C have also been entered for comparison. Natural ice exists as ice Ih and crystallizes hexagonally. In general, only this modification is of interest for technical processes. A metastable cubic variant of this modification also exists. In Figure 8, ice IV (a metastable modification, which occurs in the phase region of ice V) and ice VIII (with a triple point at ca. 5 °C and 2100 MPa, which is formed from ice VI or ice VII at higher pressure and lower temperature) are not shown. Ice IX and ice XII have also been excluded from Figure 8 because they lie outside the temperature and pressure region depicted in this figure. All known triple points are listed in Table 5 [31], [34].

Table 5 shows that modification II, for example, is not formed directly from liquid water on cooling or compression but can originate only from modifications Ih, III, V, or VI. An anomaly of water is that the melting pressure curve for ice Ih exhibits a negative slope up to the triple point Ih – III – liquid at 207 MPa and –22 °C. The values of the melting pressure curve ice Ih – liquid water are summarized below in tabular form down to this triple point [35]. e

t, °C	*p*, MPa
0.0098	6.1156×10^{-4}
0.0	0.10132
–5.0	59.8
–10.0	110.8
–15.0	155.9
–20.0	193.2
–22.0	207.4

The following equation describes the melting pressure curve of ice Ih in equilibrium with water [36]:

$$\frac{p}{\mathrm{MPa}} = -3.952 \times 10^2 \left[\left(\frac{T/\mathrm{K}}{273.16} \right)^9 - 1 \right]$$

The sublimation pressure curve of ice Ih – water vapor is known down to –110 °C:

t, °C	*p*×10^3, hPa
0	6108

–10	2597
–20	1032
–30	380.1
–50	39.37
–70	2.59
–90	0.093
–110	0.0013

Table 5. Triple points of ice and water modifications

Coexistent phases	Pressure, MPa	Temperature, °C
Ih – liquid – gaseous	6.1×10^{-4}	0.01 *
Ih – liquid – III	207	–22 *
Ih – II – III	200	–30.15 **
Ih – II – VI	1050	–120.15 **
Ih – VI – XII	1550	–168.15 **
II – III – V	350	–21.15 **
II – V – VI	970	–100.15 **
III – liquid – V	346	–17.0 *
V – liquid – VI	626	0.16 *
VI – liquid – VII	2200	81.6 *
VI – VII – VIII	2100	≈5 *

* from [34].
** from [31].

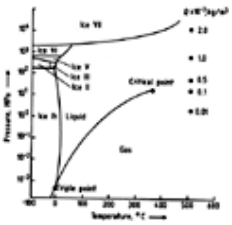

Figure 8. Pressure – temperature phase diagram of water in the solid and liquid states

Densities at 500 °C have been included for comparison.

The heat of fusion of ice Ih – water is 333.69 kJ/kg at 0 °C, and the heat of sublimation of ice Ih – water vapor is 2838 kJ/kg at 0 °C [37], [38].

The molar heat capacity C_p of ice was investigated for ice Ih. Down to –140 °C, the following equation is valid [39]:

$$\frac{C_p(\text{ice Ih})}{\text{J mol}^{-1}\text{K}^{-1}} = 2.369 + 0.130\frac{T}{\text{K}}$$

The parameters of this equation were determined by fitting the data of GIAUQUE and STOUT [35].

T, K	C_p, J mol^{-1} K^{-1}
10	0.276
20	2.050
50	7.931
100	15.879
150	22.025
200	28.211
250	34.829
270	37.481

2.1.2. *Transport Properties*

For the thermal conductivity λ of ice, the following equation is valid in the temperature range 0 to –170 °C [37], [38]:

$$\frac{\lambda}{\mathrm{W\,m^{-1}K^{-1}}} = \frac{488.19}{T/\mathrm{K}} + 0.4685$$

The self-diffusion coefficient *D* is [34]

$$D = 2 \times 10^{-15}\mathrm{m^2/s\,at}\ -7\,^\circ\mathrm{C}$$

2.1.3. *Electrolytic Properties*

Relative Permittivity. The relative permittivity $\in$ was determined on polycrystalline ice at a frequency of ν = 0 Hz:

t, °C	–0.1	–20.9	–44.7	–65.8
$\in$	91.5	97.4	104	133

The relative permittivity increases with decreasing temperature. The phase change from one ice modification to the next-denser one is always accompanied by a jump in $\in$ to considerably higher values along the phase equilibriumcurves with liquid water [32]. Within one modification region, the relative permittivity increases only slightly with pressure. The more densely the water molecules are packed, the higher is the value of $\in$.

The relative permittivity $\in$ depends on the frequency ν of an applied electrical field. This frequency dependence provides information on the rotation of dipole molecules in the lattice. The *dielectric relaxation time* is a measure of how fast the polarization caused by an electrical field decays after the field is turned off. Reorientation times of the water molecules can also be determined by NMR studies [34].

The electrical conductivity κ of ice was determined, e.g., at –10 °C [34], [40]:

$$\kappa = 1.1 \pm 0.5 \times 10^{-10}\Omega^{-1}\mathrm{cm}^{-1}$$

2.1.4. *Mechanical Properties*

The mechanical properties of ice provide information on its stability and load capacity. The mechanical properties of ice Ih are listed below [37], [38]:

Mohs scratch hardness	
at freezing point	≤2
at –78.5 °C	≈6
Tensile strength (–10 °C)	1 – 2 MPa
Compressive strength (–10 °C)	3 – 6 MPa
Young's modulus (–16 °C)	9.42 GPa
Shear strength (–16 °C)	3.55 GPa
Poisson ratio (–16 °C)	0.33 (average)

2.2. Properties of Fluid Water

2.2.1. *Thermodynamic Properties*

For water in the fluid state, equations of state are developed on the basis of experimental *p* – *V* – *T* data and fixed points (e.g., critical data). Exact and sometimes very comprehensive equations of state exist, especially for technical applications [41]. They must sometimes by subdivided for different phase regions. Equations of state that are restricted to parameters interpretable by molecular physics must take into account the high dipole moment of the water molecules and hydrogen bonds between them, especially when the equations apply to moderate temperatures and very high densities [42], [43].

The three-dimensional phase diagram of water is depicted in Figure 9.

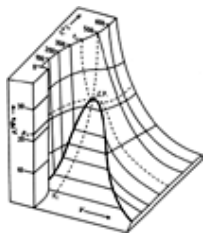

Figure 9. Three-dimensional *p* – *V* – *T* phase diagram of water; the coexistence liquid – gaseous region lies in the darkly marked area [44]

C.P. = Critical point

Tabular compilations list *p* – *V* – *T* data up to a maximum pressure of 3000 MPa and maximum temperature of 2000 °C [44-

46]. They are based on International Association for the Properties of Steam (IAPS) recommendations and on the International Formulation Committee (IFC) Formulation for Industrial Use. Selected data for temperature and pressure under saturation conditions are presented in Table 6. Corresponding volumes of the gas and liquid phase along the vapor-pressure curve up to the critical point are included [45]. A comparison between the data listed in the SI steam tables [45] and those compiled in the NBS – NRC steam tables [46] shows that for <300 °C, the values deviate from each other by only ca. 0.1 %. The deviations increase with increasing temperature and can reach 1 %. The enthalpy of vaporization $\Delta_l^g h$ is calculated from the saturation data; it is zero at the critical point. Table 7 shows specific volumes for selected pressures and temperatures in technically relevant regions and also in the supercritical state.

Table 6. Dependence of pressure *p*, specific volume *v* of water (l) and water vapor (g), and enthalpy of vaporization $\Delta_l^g h$ on the temperature in the saturation state at temperatures <100 °C [45]

t, °C	*p*, MPa	$v_1 \times 10^3$, m³/kg	v_g, m³/kg	$\Delta_l^g h$, kJ/kg
110	0.14327	1.0519	1.210	2230.0
120	0.19854	1.0606	0.8915	2202.3
130	0.27013	1.0700	0.6681	2173.6
140	0.3614	1.0801	0.5085	2144.6
150	0.4760	1.0908	0.3924	2113.3
160	0.6181	1.1022	0.3068	2081.2
170	0.7920	1.1145	0.2426	2048.0
180	1.0027	1.1275	0.1938	2013.2
190	1.2551	1.1415	0.1563	1976.8
200	1.5549	1.1565	0.1272	1938.5
210	1.9077	1.1726	0.1042	1898.4
220	2.3198	1.1900	0.08604	1856.2
230	2.7976	1.2087	0.07145	1811.8
240	3.3478	1.2291	0.05965	1764.6
250	3.9776	1.2513	0.05004	1714.5
260	4.6943	1.2756	0.04213	1661.5
270	5.5058	1.3025	0.03559	1604.7
280	6.4202	1.3324	0.03013	1543.6
290	7.4461	1.3659	0.02554	1477.6
300	8.5927	1.4041	0.02165	1406.0
310	9.8700	1.4480	0.01833	1327.6
320	11.289	1.4995	0.01548	1241.1
330	12.863	1.5615	0.01299	1143.7
340	14.605	1.6387	0.01078	1030.7
350	16.535	1.7411	0.008799	895.8
360	18.675	1.8959	0.006940	721.2
370	21.054	2.2136	0.004973	452.6
371	21.306	2.2778	0.004723	407.4
372	21.562	2.3636	0.004439	351.3
373	21.820	2.4963	0.004084	273.5
374	22.081	2.8407	0.003458	108.7
374.15	22.120	3.1700	0.003170	0.0

Table 7. Pressure and temperature dependence of specific volume *v* (10^{-3} m³/kg) of water and superheated steam up to 800 °C and 100 MPa [45]*

p, MPa	*t*, °C											
	0	100	150	200	250	300	350	400	500	600	700	800
0.1	1.0002	1696	1936	2172	2406	2639	2871	3102	3565	4028	4490	4952
1.0	0.9997	1.0432	1.0904	205.9	232.7	258.0	282.4	306.5	354.0	401.0	447.7	494.3
5.0	0.9977	1.0412	1.0877	1.1530	1.2494	45.30	51.94	57.79	68.49	78.62	88.45	98.09
10	0.9953	1.0386	1.0843	1.148	1.2406	1.3979	22.42	26.41	32.76	38.32	43.55	48.58

15	0.9928	1.0361	1.0811	1.1433	1.2324	1.3779	11.46	15.66	20.80	24.88	28.59	32.09
20	0.9904	1.0337	1.0779	1.1387	1.2247	1.3606	1.6662	9.947	14.77	18.16	21.11	23.85
25	0.9881	1.0313	1.0748	1.1343	1.2175	1.3453	1.6000	6.014	11.13	14.13	16.63	18.91
30	0.9857	1.0289	1.0718	1.1301	1.2107	1.3316	1.5540	2.831	8.681	11.44	13.65	15.62
40	0.9811	1.0244	1.0660	1.1220	1.1981	1.3077	1.4896	1.909	5.616	8.088	9.930	11.52
50	0.9767	1.0200	1.0605	1.1144	1.1866	1.2874	1.4438	1.729	3.882	6.111	7.720	9.076
60	0.9723	1.0157	1.0552	1.1073	1.1761	1.2698	1.4083	1.632	2.952	4.835	6.269	7.460
70	0.9682	1.0116	1.0501	1.1005	1.1663	1.2541	1.3793	1.567	2.467	3.972	5.257	6.321
80	0.9641	1.0076	1.0452	1.0941	1.1573	1.2401	1.3547	1.518	2.188	3.379	4.519	5.481
90	0.9602	1.0037	1.0405	1.0880	1.1488	1.2273	1.3335	1.479	2.013	2.967	3.964	4.841
100	0.9565	0.9999	1.0359	1.0821	1.1408	1.2155	1.3148	1.446	1.893	2.668	3.536	4.341

* The horizontal lines denote the transition between liquid and gaseous state.

Figure 10 shows the specific heat capacity at constant pressure c_p as a function of pressure and temperature. The specific heat capacity at the critical point approaches infinity. In the critical region, the isobars show a pronounced maximum as a function of temperature. For increasing pressures, the maxima become smaller but disappear only above 300 MPa.

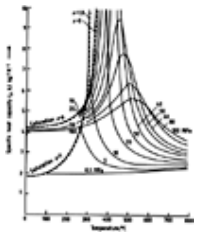

Figure 10. Specific heat capacity of water up to 800 °C and 100 MPa; saturated vapor: x = 1, saturated liquid: x = 0 [45]

Values for the isobaric expansion coefficient for temperatures in the range 200 – 1000 °C and pressures between 1 and 100 MPa are listed in Table 8. They pass through a maximum between 300 and 600 °C. The isothermal compressibility χ_T is defined as

$$\chi_T = 1/V(\delta V/\delta p)_T$$

Table 9 shows the dependence of χ_T on temperature and pressure in the pressure range 1 – 100 MPa. The maximum error is $\pm 0.016\times10^{-5}$ MPa^{-1}. The isothermal compressibility and the isobaric expansion coefficient are quantities derived from the $p - V - T$ data.

Table 8. Isobaric expansion coefficient α_p in 10^{-3} K^{-1} of water and water vapor as a function of pressure and temperature [44] *

p, MPa	*t*, °C					
	200	300	400	600	800	1000
1.0	2.714	1.926	1.563	1.168	0.941	0.789
10	1.317	3.167	2.711	1.398	1.023	0.825
20	1.255	2.63	7.030	1.720	1.121	0.866
30	1.201	2.295	36.75	2.105	1.223	0.906
40	1.153	2.060	7.972	2.538	1.325	0.944
50	1.111	1.884	4.890	2.976	1.423	0.982
60	1.073	1.745	3.711	3.343	1.512	1.016
70	1.038	1.631	3.068	3.563	1.589	1.047
80	1.007	1.537	2.655	3.602	1.650	1.073
90	0.979	1.456	2.363	3.491	1.694	1.095

100	0.952	1.386	2.144	3.288	1.721	1.112

* The horizontal line denotes the transition between liquid and gaseous state.

Table 9. Pressure and temperature dependence of the isothermal compressibility χ_T of water (in 10^{-5} MPa^{-1}) [47]

p, MPa	*t*, °C							
	0.0	**10**	**20**	**30**	**40**	**60**	**80**	**100**
10	49.461	46.564	44.725	43.633	43.100	43.287	44.778	47.426
20	48.093	45.358	43.604	42.546	42.015	42.135	43.492	45.942
30	46.772	44.196	42.527	41.507	40.982	41.048	42.285	44.555
40	45.498	43.077	41.492	40.514	39.998	40.020	41.151	43.257
60	43.076	40.955	39.542	38.652	38.167	38.129	39.080	40.902
80	40.815	38.979	37.737	36.943	36.500	36.434	37.243	38.830
100	38.699	37.135	36.063	35.371	34.980	34.911	35.612	37.002

In the area of technical thermodynamics, the enthalpy – entropy diagram (*h* – *s*) presented in Figure 11 is of interest. It is especially important in the design and operation of power plants.

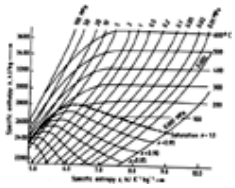

Figure 11. Enthalpy – entropy diagram of water [44]. *x* = 1: saturated vapor (*x* = 0 would correspond to the saturated liquid)

The velocity of sound *w* in water also depends on pressure and temperature. In the liquid state and at atmospheric pressure (0.1 MPa), it increases with increasing temperature. In the gaseous region and at constant higher pressure (isobaric), the sound velocity passes through a maximum at ca. 100 °C. At constant temperature, the velocity of sound increases with increasing pressure. Values for the sound velocity up to 600 °C and 100 MPa are listed in Table 10 [46].

Table 10. Sound velocity *w* in m/s in water and water vapor up to 600 °C and 100 MPa [46]

p, MPa	*t*, °C						
	0.0	**20.0**	**100**	**200**	**300**	**400**	**600**
0.1	1401.0	1483.2	472.8	533.7	585.7	632.2	713.9
1.0	1402.6	1484.8	1543.8	517.9	578.4	628.2	712.5
10	1418.1	1500.6	1564.3	1361.3	919.7	581.0	698.9
20	1434.9	1517.6	1586.2	1396.4	1004.4	505.8	684.8
30	1451.3	1534.2	1607.1	1429.1	1073.0	419.0	672.7
40	1467.9	1550.6	1627.3	1459.7	1131.3	627.3	664.1
60	1501.9	1583.0	1665.7	1515.8	1228.1	850.7	666.1
80	1538.1	1615.5	1701.9	1566.4	1307.6	990.8	702.9
100	1577.1	1648.6	1736.2	1612.7	1375.8	1096.7	767.3

* The horizontal lines denote the transition between liquid and gaseous state.

2.2.2. *Transport Properties*

Thermal Conductivity. The thermal conductivity λ of water as a function of pressure and temperature in the range 200 – 800 °C and 10 – 100 MPa is shown in Table 11. Experimental data for higher temperatures and pressures exist: for 250 °C up to 300 MPa [48] and for 250 to 510 °C up to 95 MPa [49]. The values in Figure 12 have been taken from international

tabular compilations [45]. The tolerances are around 1.5 % for 0.1 MPa and 0 °C, but they reach 8 % for 100 MPa and 600 °C. A characteristic feature is that the thermal conductivity of water shows a very pronounced peak at the critical point. This indicates that the change in structure of water is especially great at the critical point. For the supercritical region, the isobars pass through a minimum up to about 50 MPa. This minimum is more pronounced, the lower the pressure. Above ca. 50 MPa, the isobars decrease monotonously without passing through a minimum.

Table 11. Thermal conductivity of water and water vapor (10^{-3} Wm^{-1} K^{-1}) as a function of pressure and temperature in the range 200 – 800 °C and 1 – 100 MPa [45] *

p, MPa	*t*, °C					
	200	300	400	500	600	800
1.0	35.9	44.3	56.0	68.0	81.0	108.6
10	675	557	66.9	75.6	89.4	118.1
20	684	576	104.9	91.6	98.6	122.7
30	692	593	337	114.1	112.3	130.2
40	700	608	399	152.9	129.2	139.3
60	715	634	476	277	176	161
80	729	653	521	346	235	185
100	742	672	561	412	288	215

* The horizontal line denotes the transition between liquid and gaseous state.

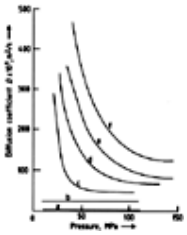

Figure 12. Self-diffusion coefficient of water up to 700 °C and 150 MPa [51]

a) 25 °C; b) 200 °C; c) 400 °C; d) 500 °C; e) 600 °C; f) 700 °C

Dynamic Viscosity. The dynamic viscosity η of water has been studied experimentally in the high-pressure region up to 300 MPa and 500 °C [50]. Figure 13 shows the temperature dependence of dynamic viscosity for various pressures in the sub- and supercritical regions. These data have been taken from SI steam tables [45]. Like other liquids, liquid water has the highest viscosity at low temperature and high pressure (i.e., the region of high density). The viscosity of water decreases rapidly with increasing temperature, even at high pressure. In water vapor, the viscosity increases at low pressure and density with increasing temperature. Between these two limits with, on the one hand, a high and, on the other hand, a low density, a large region exists in which the viscosity of water barely changes (see Section Bulk Properties of Liquid Water).

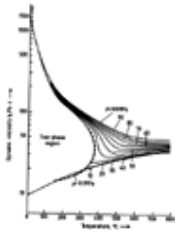

Figure 13. Dynamic viscosity of water up to 800 °C and 100 MPa [45]

Table 12 shows the viscosity, thermal conductivity, and isothermal compressibility of water [45] along the saturation curve of liquid and vapor in the temperature range 150 – 374 °C.

Table 12. Dynamic viscosity η, thermal conductivity λ, and isothermal compressibility χ_T of water (′) and steam (″) in the saturation state [45]

t, °C	$\eta'\times10^6$, Pa · s	$\eta''\times10^6$, Pa · s	$\lambda'\times10^3$, W K^{-1} m^{-1}	$\lambda''\times10^3$, W K^{-1} m^{-1}	$\chi_T'\times10^3$, MPa^{-1}	$\chi_T''\times10^3$, MPa^{-1}
0.0	1793	9.216	561.0	17.07	0.5101	1 636 700

25	890.5	9.866	607.1	18.55	0.4523	316 020
50	547.1	10.62	643.5	20.36	0.4421	92 124
100	281.9	12.27	679.1	25.09	0.4909	10 029
150	182.5	13.99	682.1	31.59	0.6215	2 205
200	134.4	15.71	663.4	40.10	0.8856	722.4
250	106.2	17.49	621.4	51.23	1.470	322.1
300	85.96	19.65	547.7	69.49	3.211	195.0
350	65.88	23.81	447.6	134.6	16.67	215.2
374	43.2		∞		∞	

At the critical point, the thermal conductivity and specific heat capacity extend to infinity.

With increasing temperature, the compressibility increases in liquids but decreases in the gas phase. At the critical point, it approaches infinity.

Self-Diffusion. Very few investigations have been conducted on the diffusion coefficient of water in technically relevant regions of high pressure and temperature [51]. The self-diffusion coefficient as a function of pressure up to 150 MPa and temperature from 25 to 700 °C is shown in Figure 12. In this region, a pressure increase always results in a decrease in the self-diffusion coefficient.

2.2.3. *Electrolytic Properties*

The ion product (see Section Bulk Properties of Liquid Water) has been the subject of diverse studies and discussions. The dependence of the ion product on temperature and pressure is presented in Table 13 [52]. It shows that the logarithmic ion product (e.g., at 400 °C and 100 MPa) assumes a value of approximately 10.8 compared with a value of about 14.0 under normal conditions. The autoionization of water is favored by increasing temperature and increasing density. Thus, an increase in pressure at constant temperature always results in an increase in ion product. Shock-tube experiments have shown that water is ionized completely above 1000 °C and 20 GPa.

Table 13. Ion product –log K_w/(mol kg^{-1})^2 of water up to 1000°C and 1000 MPa [52]

p, MPa	*t*, °C								
	0.0	25.0	100	200	300	400	600	800	1000
p_{Sat}	14.938	13.995	12.265	11.289	11.406				
25	14.83	13.90	12.18	11.16	11.14	19.43	23.27	24.23	24.93
50	14.72	13.82	12.10	11.05	10.86	11.88	18.30	19.92	20.80
75	14.62	13.73	12.03	10.95	10.66	11.17	15.25	17.35	18.39
100	14.53	13.66	11.96	10.86	10.50	10.77	13.40	15.58	16.72
250	14.08	13.28	11.61	10.45	9.91	9.74	10.18	11.11	12.02
500	13.60	12.83	11.22	10.00	9.34	8.99	8.85	9.11	9.42
1000	12.91	12.21	10.68	9.45	8.71	8.25	7.78	7.68	7.85

For a liquid, water has a very high relative permittivity of ca. 80 under normal conditions because of its unusual structure as a result of the formation of hydrogen bridges (see Section Bulk Properties of Liquid Water).

The relative permittivity ∈ of water and steam up to 500 °C and 500 MPa [53] is given in Table 14. The relative permittivity decreases with increasing temperature (isobaric). If the pressure is increased isothermally (i.e., if density is increased), ∈ increases. In the critical region, the relative permittivity is ca. 6. Pressure and temperature influence not only the value of relative permittivity, but also the solution behavior of water. High relative permittivities favor solubility and ionization of electrolytes.

Table 14. Temperature dependence of the relative permittivity of ∈ water and steam for selected pressures [53]

p, MPa	*t*, °C						
	0.0	25.0	100	200	300	400	500

0.1	87.81	78.46	1.00	1.00	1.00	1.00	1.00
1.0	87.86	78.49	55.44	1.02	1.02	1.01	1.01
10	88.28	78.85	55.76	35.11	20.39	1.17	1.11
20	88.75	79.24	56.11	35.52	21.24	1.64	1.32
40	89.64	80.00	56.77	36.28	22.56	10.46	2.34
60	90.49	80.72	57.39	36.96	23.58	13.28	4.90
80	91.29	81.42	57.98	37.59	24.43	14.88	7.50
100	92.04	82.08	58.55	38.17	25.17	16.05	9.29
200	95.20	84.94	61.00	40.66	28.07	19.69	13.86
300	97.69	87.34	63.10	42.65	30.15	21.94	16.25
400	99.72	89.39	64.94	44.33	31.81	23.62	17.89
500	101.42	91.16	66.57	45.82	33.21	24.96	19.14

** The horizontal lines denote the transition between liquid and gaseous state.

2.2.4. *Other Physical Properties*

The interfacial tension σ of water as a function of temperature at the saturation vapor pressure in the temperature range 150 – 374 °C is listed in Table 15. At 20 °C and 0.1 MPa, σ of water reaches a value of 72.75 mN/m, which is very high for liquids. With increasing temperature, the interfacial tension decreases almost linearly until it reaches zero at the critical point [45]. In principle, the isothermal pressure dependence of the interfacial tension of pure water cannot be measured. To generate pressure, the interfacial tension of water must always be measured in relation to a second partner. If gases are used for this purpose (e.g., nitrogen, methane, carbon dioxide, neon, or argon), a decrease in interfacial tension with increasing pressure is observed because the gases dissolve in water and also are adsorbed at the interface. Only at >150 MPa does the interfacial tension increase slightly [54].

Table 15. Surface tension σ of water at 0.1 MPa [45]

t, °C	σ, N/m
150	48.74
200	37.69
250	26.06
300	14.30
350	3.65
374	0

The refractive index *n* of water increases with increasing pressure from 1.334 at 0.1 MPa to 1.347 at 100 MPa.

2.3. Properties of Water Vapor

Most of the physical properties of water vapor have been described in Section Properties of Fluid Water. Its standard enthalpy of formation from the elements at 25 °C and 0.1013 MPa, $\Delta H°_{298.15\ K}$ is 241.99 kJ mol^{-1} K^{-1}; its standard entropy of formation $S°_{298.15\ K}$ is 188.85 J mol^{-1} K^{-1}.

For estimation of *p* – *V* – *T* data at moderate pressure, a virial equation is often used:

$$\frac{pv}{RT} = 1 + \frac{B}{v} + \frac{C}{v^2} + \ldots$$

where *v* is the specific volume in m^3/kg.

The serial development is often stopped after the second virial coefficient *B*. A compilation of values for the second virial coefficient as a function of temperature is given in Table 16 [55].

Table 16. Second virial coefficient of water vapor [55]

t, °C	*B*, 10^6 m^{-3} mol^{-1}	*t*, °C	*B*, 10^6 m^{-3} mol^{-1}	*t*, °C	*B*, 10^6 m^{-3} mol^{-1}
20	–1251.5	200	–206.0	575	–37.5
50	–812.2	275	–132.4	675	–26.2

100	–459.8	375	–82.0	775	–18.1
150	–295.4	475	–54.5	975	–7.6

2.4. Water in the Supercritical State [34], [56], [57]

Above the critical temperature of 374 °C, water possesses many unusual properties at high density. In contrast to the liquid state, for example, mixing with nonpolar gases and many organic compounds with large, nonpolar molecules is generally possible. In addition, supercritical water can dissolve electrolytes with the formation of ions if its density and relative permittivity are sufficiently high. Thermophysical properties such as viscosity and relative permittivity of supercritical water can be changed continuously in wide regions by pressure variation. Thus, water under supercritical conditions can be used as an extraction agent as well as a versatile reaction medium (e.g., for hydrogenation and oxidation). Combinations of pyrolysis, hydrolysis, and oxidation have already been proposed and developed for the emission-free annihilation of pollutants.

[Top of Page]

3. Water Analysis

Georg Schwedt

3.1. Sampling and Sample Preservation (see also ⟶ Sampling; ⟶ Sample Preparation for Trace Analysis).

Sampling [58-60] is frequently the crucial step in the flawless analysis and evaluation of water. Depending on the type of problem, a distinction must be made between single or random samples and mixed, composite, or average samples (⟶ Sampling – Probability Sampling). In view of representative sampling, the parameters to be considered are very complex. From a homogeneous body of water, a single sample is sufficient to achieve representative sampling. In contrast, in the case of surface waters and stagnant piping systems, several individual samples must be taken. In flowing water, both nonconstant currents and nonhomogeneous water bodies are present. In general, composite samples provide information on certain periods of time, but top loads cannot be discerned. For this reason, the type of sample collection to be used is often stipulated by law to maintain limiting and guideline values. The following DIN processes (no. 38 402) exist for the sampling of unpolluted and polluted waters. Corresponding U.S. processes are found in [60].

Sampling of wastewater	A 11
Stagnant water	A 12
Groundwater aquifers	A 13
Raw water and drinking water	A 14
Flowing water	A 15
Seawater	A 16
Falling, wet precipitation in the liquid state of aggregation	A 17
Water from mineral and medicinal springs	A 18
Water from swimming and bathing pools	A 19
Tidal waters	A 20
Cooling water for industrial use	A 22

Parameter-specific sampling is required, especially in chemical wastewater analysis. Here, too, a legal basis has been established [61], [62]. Further information is contained in ISO standards 5667–1 and -2 (corresponding to EN 25 667–1 and -2) “Water quality sampling: Guidance on the design of sampling programme” and “Guidance on sampling techniques.”

The type and requirements of *sample preservation* [58-60] depend on the constituents to be determined, their concentrations, their possible mutual interactions, and especially the activity of microorganisms contained in the sample. Keeping the sample in darkness and cooling it represent the most general and the most effective processes. For transport to the laboratory, a temperature of 2 – 5 °C is usually sufficient, but for longer storage times, deep freezing at –20 °C is required. Chemical preservation processes depend on the parameters or substances to be determined. A compilation of the preservation processes and techniques for 75 chemical parameters, as well as microbiological examinations, is given in EN ISO 5667–3 (1994, German version 1995).

3.2. Physicochemical and Sum Parameters

In the broadest sense, the term *in situ analysis* refers to analytical methods that can be applied directly to the sample to be investigated for the measurement of physicochemical and sum parameters, or individual components. In this chapter, in situ analyses refer only to processes that must be conducted directly at the sampling site because changes can occur on the way to the laboratory.

In situ analytical processes for water are based mostly on electrometric measuring principles. They include the measurement of temperature, pH, oxygen and redox potential, and conductivity. Moreover, determination of light absorption in the visible and UV regions belongs to the general physicochemical analysis of water samples. A survey is given in Table 17.

Table 17. Physical and physicochemical parameters in water analysis

Parameter	Method	Remarks
Temperature, K (°C)	electronic measurement with temperature probe (DIN 38 404-C 4)	for temperature-dependent quantities: solubility of gases, pH value, electrical conductivity, density
Density, g/cm^3	pycnometer, precision aerometer, bending vibration method (DIN DEV-C 9)	in stronger mineralized waters
Electrical conductivity κ, Ω^{-1} cm^{-1}	conductivity cell with cell constant *K*	measure electrolyte concentration
Conductance *G*, Ω^{-1}	(EN 27888-C 8, ISO 7888)	check *K*, determine temperature coefficient $\alpha_{\theta,25}$
pH value *a* (H_3O^+)	electrometric (potentiometric) measurement with glass electrode and saturated calomel or silver – silver chloride reference electrode (DIN 38 404-C 5)	measure at sampling site because of rapid changes due to physical, chemical, and biological processes: asymmetry balance, correction of electrode slope, consideration of temperature compensation, measurement in low-electrolyte waters with long adjustment times
Redox potential, mV	electrometric (potentiometric) measurement with Pt electrode against reference electrode (DIN 38 404-C 6)	interference by electrode coating and in low-oxygen waters by small changes in concentration of dissolved oxygen
Absorption in visible (436 nm) or UV range (254 nm)	measurement of spectral decadic absorbance *A* (λ), determination of the spectral absorption coefficient (SAC) α (λ) or the specific spectral absorption coefficient (SAC/DOC *) corresponding to α (λ) ρ_{DOC} (436 nm: EN ISO 7887-C 1; 254 nm: DIN 38 404-C 3)	especially a measure of the concentration of aromatics (254 nm); determination of color (436 nm)
Turbidity	by nephelometry: measurement of attenuation of transmitted (860 nm) monochromatic IR light or of scattered radiation (860 nm, scattering angle 90°) (EN 27 027-C 2, ISO 7027)	detection of sand, loam, and clay particles, of bacteria and other microorganisms

* DOC = dissolved organic carbon; see Table 18.

As a result of the large number of substances present in water of various types, *cummulative determinations* have been developed. These provide preliminary information on the degree of pollution or the presence of certain groups of substances. In the narrower sense, *sum parameters* more or less completely identify organic substances after their oxidation to carbon dioxide (Table 18). The *total organic carbon* (TOC) value gives the total content of organic substances, the *permanganate index* detects only easily oxidizable substances, and the amount of biochemically degradable substances is given by the biological oxygen demand (BOD). Determination of parameter ratios provides useful information for water monitoring self-purification processes [58], [63]. For instance, the BOD_5 : COD ratio represents a characteristic value that is specific for wastewater and characterizes its biodegradability. Cumulative determinations such as electrolytic conductivity, residue on evaporation, cation exchange, and pH (Tables 17 and 18) can also be used for *plausibility testing* (e.g., via ion balances) in conjunction with the results of individual determinations. Table 19 shows the differentiation of the sum parameters for haloorganic substances, which also contribute to the characterization of a group of substances. The determination procedure is based on the combustion of a sample at ca. 1000 °C in a stream of oxygen. The combustion products enter an electrolysis cell, and chloride is titrated coulometrically (the sum of all halides is given as chloride value) [64].

Table 18. Chemical and biochemical sum parameters

Parameter	Definition: determination method
Total dry residue ρ_{GT}, mg/L	volume based, mass of dissolved and undissolved (nonvolatile) water constituents (based on unfiltered sample) left after a prescribed drying process (DIN 38 409-H 1)
Filtrate dry residue ρ_{FT}, mg/L	residue on evaporation; volume-based mass of dissolved (nonvolatile) water constituents left after a prescribed filtration and drying process, according to EC drinking water guideline at 180 °C (DIN 38 409-H 1: 105 ± 2 °C)
Substances removable by	according to DIN 38 409-H 2

filtration P_A, mg/L	
Residue on ignition P_{GA}, mg/L	ignition (550 °C) of the filtrate dry residue (DIN 38 409-H 2)
Cumulative determination by cation exchange, mmol/L	exchange of cations by oxonium ions
Total hardness, degrees of hardness [e.g., d GH [b] or as d (Ca^{2+} + Mg^{2+})], mmol/L	EDTA-Na_2 [a] and Eriochrome Black T (pH 10) (DIN 38 406-E 3–3) or by AAS [c]: DIN 38 406-E 3–1 or ISO 7980
Acid and base capacity ($K_{A,4.3}$, $K_{A,8.2}$ and $K_{B,4.3}$, $K_{B,8.2}$), mmol/L	neutralization titration with HCl or NaOH against methyl orange or phenolphthalein or with electrometric end point indication (pH 4.3 or 8.2)
Calcuable therefrom: *m* and *p* alkalinity, mmol/L [d]	*acid capacity*: volume-based quantity of hydronium ions that a sample can absorb until it reaches a certain pH; *base capacity*: volume-based quantity of hydroxyl ions that a sample can absorb until it reaches a certain pH. An aqueous CO_2 solution with a HCO_3^- ion concentration (mmol/L) of 1 % of the concentration of dissolved CO_2, a temperature of 25 °C, and an ion strength of 10 mmol/L has a pH of 4.3. An aqueous HCO_3^- solution with >1 mol/L HCO_3^- ions, a temperature of 25 °C and an ionic strength of 10 mmol/L has a pH of 8.2. Details in DIN 38 409-H 7. Acid and base capacity are determined for the calculation of $c(CO_2)$, $c(HCO_3^-)$, and $c(CO_3^{2-})$. Influenced by the presence of other weak acids
TIC, mg/L (calculated as C)	total inorganic carbon to be derived from the carbon sum $TIC = c(CO_2 + c(HCO_3^-) + c(CO_3^{2-})$, which is determinable for Q_C > 0.5 mmol/L, according to TIC = *m* alkalinity – *p* alkalinity (DEV D 8); or determination in the course of TOC determination [d]
TOC, mg/L (calculated as C)	total organic carbon determination by oxidation of all C compounds to CO_2 (UV irradiation in the presence of an oxidant such as peroxodisulfate at ca. 60 °C (measurement of CO_2 by IR spectrometry, coulometry, or conductometry, or after reduction to methane by GC/TCD [e] or FID [f]) (DIN 38 409-H 3)
DOC, mg/L (calculated as C)	dissolved organic carbon (mostly after membrane filtration – 0.45 µm pore size)
POC, mg/L (calculated as C)	particulate (undissolved) organic carbon
VOC, mg/L (calculated as C)	volatile organic carbon
Permanganate index, mg/L	mass per unit volume of oxygen that is equivalent to the mass per unit volume of consumed permanganate ions (permanganometric, detects easily oxidizable substances) (DIN 38 409-H 5, ISO 8467)
COD, mg/L	determination of oxidizability with potassium dichromate, volumetrically or photometrically (at 340, 445, or 585 nm) (DIN 38 409-H, ISO 6060)
BOD, mg/L	determination of the volume-based mass of oxygen that is consumed for oxidation of the biochemically oxidizable constituents contained in a 1-L water sample in *n* days (generally *n* = 5, BOD_5) with the metabolic activity of a corresponding microbiocenosis at 20 °C (DIN 4045; DIN 38 409-H 51, ISO 5815); oxygen determinations volumetrically, manometrically, or electrometrically with oxygen electrode

[a] Ethylenediaminetetraacetic acid disodium salt.
[b] GH = Gesamthärte (total hardness).
[c] Atomic absorption spectrometry.
[d] For definition of *m* and *p* alkalinity see Section Characterization of the Chemical State of Water.
[e] TCD = thermal conductivity detector.
[f] FID = flame ionization detector.

Table 19. Differentiation of sum parameters for haloorganic compounds (parameters in µg/L, calculated as Cl)

Parameter	Detectable substances

TOX	total halogens in organic compounds (total organic halogens)
DOX	dissolved organic halogens
AOX	adsorbable (mostly on activated carbon) organic halogens
POX	purgeable organic halogens (DEV H 25)
VOX	volatile organic halogens
EOX	extractable organic halogens
TOCl	total of organic chlorine
DOCl	dissolved organic chlorine

Most of the processes mentioned also exist in a variant for on-line analysis. Equipment for on-line analysis allows the course of processes in the waste treatment plant to be controlled with robust (i.e., often also no- or low-maintenance) measuring devices.

3.3. Inorganic Analysis

The analytical procedures for inorganic cations and anions range from simple analytical colorimetric testing methods in the field, spectrophotometry, electrochemical methods (especially polarography and inverse voltammetry), and ion chromatography (IC) to atomic absorption spectrometry (AAS), optical atomic emission spectrometry with inductively coupled plasma (ICP – OES), as well as inductively coupled plasma coupled with high-resolution mass spectrometry (ICP – MS) [65], [66].

3.3.1. *Determination of Cations*

Because of the efficiency of instrumental methods, cation analysis is carried out mostly in water without any special sample preparation. However, trace enrichment processes [67], [68] or digestion (usually UV digestion for the destruction of organic substances) must be applied in the case of less selective methods, such as spectrophotometry, or because of interference by organic substances. [Trace enrichment is conducted with simultaneous matrix separation on selective ion exchangers and, after chelation with complexing agents such as dithiocarbamates, on chemically modified silica gel or polymer adsorbents or on activated carbon, which can also be used as flow injection analysis (FIA) technique.] A survey of the generally accepted analytical methods and of the cations that can be analyzed is presented in Table 20. The analytical methods standardized for water analysis are collected in the “German Standard Methods for the Examination of Water, Wastewater, and Sludge” (Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung, DEV).

Table 20. Methods of cation analysis

Method	Determinable cations
Spectrophotometry (with selective reagents, partly as FIA-technique), colorimetrically as quick test	Al, Ag, As. Cd, Co, Cr(III, VI), Cu, Fe, Hg, K, Mn, Ni, Pb, Zn
Fluorescence spectrometry (with selective reagents)	Al, Be, U
Atomic absorption spectrometry (AAS), flame or graphite tube technique	
Digestion/extraction with dithiocarbamate (DIN 38 406-E 21)	Ag, Bi, Cd, Co, Cu, Ni, Pb, Tl, Zn
Hydride AAS-technique	As, Bi, Ge, Pb, Sb, Se, Sn, Te
Cold vapor technique (DEV-E 12)	Hg
Atomic emission spectrometry with inductively coupled plasma (ICP – OES) (DIN 38 406-E 22)	Ag, Al, As, B, Ba, Be, Bi, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, P, Pb, S, Sb, Se, Si, Sn, Sr, Ti, V, W, Zn, Zr
ICP mass spectrometry (ICP – MS)	Ag, Al, As, B, Ba, Bi, Ca, Cd, Co, Cr, Cs, Fe, Hg, K, Li, Mg, Mn, Mo, Na, P, Pb, Rb, S, Sb, Se, Sn, Sr, Th, Ti, Tl, U, V, W, Zn
Polarography and voltammetry (DIN 38 406-E 16)	Cd, Co, Cu, Ni, Pb, Tl, Zn
Potentiometric stripping analysis	Ag, As, Bi, Cd, Co, Cu, Fe, Hg, In, Mn, Ni, Pb, Sb, Se, Sn, Tl, U, Zn

Apart from classical spectrophotometry (and fluorescence spectrometry for a few elements), high-performance methods of atomic spectrometry are available today. These include atomic absorption spectrometry with flame AAS (⟶ Atomic Spectroscopy – Flame Atomic Absorption), graphite furnace technique (flameless AAS; ⟶ Atomic Spectroscopy – Electrothermal Atomic Absorption), the hydride technique (for elements that form volatile hydrides), the cold vapor technique (for Hg, ⟶ Atomic Spectroscopy – Hydride and Cold Vapor Techniques, and as a multielement determination method,

optical emission spectrometry with inductively coupled plasma as the excitation source (⟶ Atomic Spectroscopy – Inductively Coupled Plasmas). The latter can be coupled with high-resolution mass spectrometry (ICP – MS technique) to achieve still lower detection limits and larger linearity ranges and to measure the abundance ratios of isotopes [69]. Despite the advances in and efficiencies of these methods, matrix effects and interference still must be taken into account. They must often be eliminated by special sample preparation, such as selective extraction or concentration. In addition, electrochemical methods such as polarography and voltammetry [70], especially the stripping technique — after concentration on the electrode — (⟶ Analytical Voltammetry and Polarography – Stripping Voltammetry.) and potentiometric stripping analysis (PSA, ⟶ Analytical Voltammetry and Polarography – Stripping Chronopotentiometry (SCP).), are important in water analysis. In PSA, metal ions capable of amalgam formation are first deposited on a working (usually glassy carbon) electrode by potentiostatic electrolysis. The metals are then oxidized and again dissolved (stripped), the time required being proportional to their concentration in solution. The potential – time curves are evaluated.

In cation analysis, besides determination of total concentrations, element species analysis (*speciation*), i.e., differential analysis of one element according to different chemical bonding and physical form of state, is becoming more and more important [71]. Examples are the determination of chromium as Cr(III) and Cr(VI) [72], arsenic as As(III) and As(V) [73], and inorganic and organic tin compounds [74]. For speciation, coupling processes with a chromatographic separation system (gas and liquid chromatography) and atomic spectrometric detection have gained importance [75].

3.3.2. *Anion Analysis*

Spectrophotometry is the classical analytical method in anion analyis, as well. For some elements present in the anionic state in water (e.g., B, V, Mo), methods of cation analysis are also used (Table 20). In addition, for some special anions such as fluoride, nitrate, and cyanide or sulfide, *ion-selective electrodes* [76] have proved useful in direct potentiometric determination (with use of the standard addition process and integrated in FIA systems for calibration). Most of the processes can be automated; i.e., they can also be employed as FIA systems (Table 21). Furthermore, special problems such as determination of the total nitrogen or total phosphorus concentration in wastewater samples can also be solved by using the FIA technique [77], [78]. In these processes, digestion of all nitrogen compounds (oxidation under UV radiation with peroxidisulfate) to nitrate or of organic phosphorus compounds to orthophosphate (by acid and microwave radiation) is integrated into the flow system.

Table 21. Methods of anion analysis

Method	Determinable anions
Spectrophotometry (after chemical conversions with selective reagents)	borate, chromate, cyanide, fluoride, iodide, nitrate, nitrite, phosphate, silicate, sulfate, sulfide, sulfite, thiocyanate (others such as borate/B, molybdate/Mo, vanadate/V; see also cation analysis in Table 20)
Potentiometry with ion-selective electrodes	fluoride, cyanide, nitrate
Ion chromatography	
(DIN 38 405-D 19) in slightly polluted water	fluoride, chloride, nitrite, phosphate, bromide, nitrate, sulfate
(DIN 38 405-D 20) in wastewater	chloride, nitrite, phosphate, bromide, nitrate, sulfate
(DIN 38 405-D 22)	chromate, iodide, sulfite, thiocyanate, thiosulfate

Since the early 1980s, *ion chromatography* has been of utmost importance in anion analysis [79]. Being a type of high-performance liquid chromatography, it uses separation columns with anion exchangers on a polymer or silica gel basis, as well as ion-pair separation systems in conjunction with different detection methods such as conductivity, UV, and amperometric detection [59]. If solutions with carbonate – hydrogencarbonate as eluting ions are employed, the suppressor technique for neutralization of the mobile phase (by means of an ion-exchange membrane in the H^+ form) is used in conjunction with conductivity detection. Ions that absorb UV, such as nitrate, nitrite, bromide, iodide, or thiocyanate, can be detected directly (by using UV transparent eluents); ions that do not absorb UV can be detected by indirect UV detection (UV-adsorbing eluent such as phthalate buffer or complexons [80]); and electrochemically active (oxidizable) anions, such as iodide and nitrite, can also be detected amperometrically [81]. This wide range of separation and detection methods is a characteristic feature of ion chromatography. In anion analysis, the latest method is capillary electrophoresis in the form of *zone electrophoresis* [82].

3.3.3. *Determination of Dissolved Gases*

The most frequently determined dissolved gases in water are oxygen, ozone, carbon dioxide, and chlorine.

Oxygen. The classical *Winkler process* for the measurement of oxygen is based on the conversion of manganese(II) to manganese(IV) hydroxide under alkaline conditions (“oxygen fixation”), with subsequent iodimetric determination (EN 25 813, ISO 5813). Measurements with an *oxygen electrode* (as oxygen sensor) according to Clark, are based on the following principle. In an electrolyte-filled reaction chamber, a gold cathode and a silver electrode, which is involved in the redox processes and acts as the anode, are polarized with a direct voltage of 790 ± 10 mV. The reaction chamber is sealed with a gas-permeable membrane. Oxygen from a water sample reaches the cathode by diffusion [e.g., via a teflon (polytetrafluoroethylene) film] and is reduced. Silver dissolves correspondingly at the anode. The current flowing in the cell is

proportional to the partial pressure of oxygen [58].

Carbon Dioxide (see Section Characterization of the Chemical State of Water). The carbon dioxide content ("free carbonic acid") can be calculated from the total inorganic carbon (TIC = *m* alkalinity — *p* alkalinity), with the *m* and *p* alkalinities (see Section Physicochemical and Sum Parameters) determined by titration taking the pH and conductivity into account (in the pH range 6.0 – 8.4) [58]. Carbon dioxide can also be measured with the aid of a *gas-sensitive electrode*. In the case of a CO_2 electrode, as with the oxygen sensor, gas molecules diffuse through a semipermeable membrane into an electrolyte solution (hydrogencarbonate). On entry of CO_2, the pH of this hydrogencarbonate solution is shifted, corresponding to the equilibrium between carbon dioxide and hydrogencarbonate. The pH shift is measured. The standard processes DEV-D8 [calculation of dissolved carbon dioxde (free carbonic acid), carbonate, and hydrogencarbonate ions] and DIN 38 404-C 10 (calcium carbonate saturation of a water sample) are available for investigations of the carbonate balance including the calcium carbonate system that are important in practice.

Chlorine. For the determination of *free chlorine* and *total chlorine* (e.g., in swimming pool water), both chemical and physicochemical processes (using an electrode) are available. The photometric or colorimetric (also possible volumetrical) process (DIN 38 408-G 4) is based on the reaction of elemental chlorine (as well as hypochlorous acid and hypochlorite ions, where the sum = free chlorine) with *N,N*-diethyl-*p*-phenylenediamine to give a red oxidation product. Chloramines can be detected only after conversion with iodide (bound chlorine). The content of chlorine can also be determined with a *gas-sensitive electrode* (measurement of redox potential).

Ozone. Like free chlorine, ozone is determined photometrically.

3.3.4. *Quick Test Processes*

For most of the parameters described in Sections Anion Analysis and Determination of Dissolved Gases, chemical quick tests — also called alternative tests compared to laboratory processes — are available on the market. They include test papers (for "semiquantitative" analysis of various metal ions), test strips, colorimetric tests, and gas testing tubes for substances that can be purged out of water [83].

As an analysis involving carrier-bound reagents, *test strips* are square reaction zones made of paper applied to plastic strips. With the help of a pocket instrument, test strips can also be evaluated by reflectometry.

The application of *gas testing tubes* has been extended from workplace monitoring to include water analysis. In water analysis, volatile substances are purged out of water with the help of air and a gas-washing bottle. A certain volume of air is sucked through the testing tube (located between gas-washing bottle and pump) by using a bellows pump. The length of the colored zone is a measure of a definite concentraton of the gas to be determined.

Colorimetric quick tests are available in the form of cuvette tests with color comparison, as a rotary-disk comparator, or a color-scale sliding comparator. If two cuvettes are used — one cuvette for the untreated sample — the self-color of a water sample can be compensated for. Visual colorimetry according to the Lambert – Beer law based on the principle of remission measurement (in a direct-light process) with greater layer depths (compared to the rotary-disk comparator) and higher sensitivity finds application in color chart tests. The color of a sample of water after the addition of reagents is compared with the colors on a plastic color chart valid for one concentration range in each case (lower and upper limits). The reagent sets used in these processes can also be applied in photometric determinations for on-the-spot analysis with the help of battery-driven pocket photometers.

3.4. Organic Analysis

The analysis of organic substances in water is characterized on the one hand by cumulative detection on the basis of sum parameters (see Section Physicochemical and Sum Parameters) and by biochemical and biological test methods (see Section Biochemical Methods) for total effect-oriented analysis. On the other hand, analysis of individual substances is carried out by using efficient chromatographic separation methods [84]. The groups of substances most often analyzed are pesticides, polycyclic aromatic hydrocarbons, haloorganic hydrocarbons, phenols, hydrocarbons from mineral oils and mineral oil products, and surfactants.

3.4.1. *Spectrometric Methods*

Unlike the analysis of inorganic water constituents, UV - visible (UV – VIS) spectrometry plays a much smaller role for organic substances. Substances determined by UV – VIS spectrometry include, e.g., uric acid and pyridine as single substances; phenols (phenol index DIN 38 409 H 16), organic complexing agents such as ethylenediaminetetraacetic acid (EDTA), nitrilotriacetic acid (NTA), and phosphonic acids (as bismuth complexation index I_{BIK}; humic acids; lignosulfonic acids [85]; and surfactants (DIN 38 409 H 20 and 23) as groups of substances.

For turbid water, *derivative spectrometry* is the method of choice not only to eliminate the interference of colloids and particles, but also to obtain characteristic spectra ("shoulders" of the original spectrum give minima in the first derivation), e.g., in the determination of phenol [59].

Fluorescence spectrometry (⟶ Ultraviolet and Visible Spectroscopy – Fluorimetry is based on the property of substances absorbing energy (excitation) in the form of light (UV – VIS range) and emitting it as radiation with lower energy, i.e., longer wavelength (fluorescence). Fluorimetry is most important as a detection method in conjunction with separation methods. It is also used to determine the concentration of chlorophyll A (DIN 38 412 L 33) in algae in the examination of the trophic state of waters; in the qualitative analysis of polycyclic aromatic hydrocarbons (PAHs); in surfactant analysis; and in the characterization of humic substances. Humic substances are characterized spectroscopically [86] by spectral absorption coefficients at 245 and 436 nm and by fluorescene with an excitation wavelength of 330 (or 341) nm and fluorescence radiation of 410 (or 460) nm. Complexable paramagnetic metal ions quench this fluorescence.

Infrared spectroscopy is based on the absorption of electromagnetic radiation for the initiation of mechanical vibrations of atoms or functional groups in a molecule. This method is used primarily in the characterization and quantification of hydrocarbons. Sample preparation involves extraction with CCl_4 or 1,1,2-trichloro-trifluoroethane (DIN 38 409 H 18). Nonhydrocarbons (such as surfactants) that are also extracted in this process are separated on a polar adsorbent. The characteristic absorption bands used are 3.38 µm ($\tilde{\nu}$ = 2958 cm^{-1}) for the CH_3 group, 3.42 µm ($\tilde{\nu}$ = 2924 cm^{-1}) for the CH_2 group, and 3.30 µm ($\tilde{\nu}$ = 3030 cm^{-1}) for the aromatic CH group (⟶ Infrared and Raman Spectroscopy – The Concept of Group Frequencies, ⟶ Infrared and Raman Spectroscopy – Methyl and Methylene Groups, ⟶ Infrared and Raman Spectroscopy – Alkene Groups, ⟶ Infrared and Raman Spectroscopy – Aromatic Rings). Humic substances and surfactants [87] can also be characterized by IR spectroscopy.

3.4.2. *Gas and Liquid Chromatography*

As physicochemical separation methods, chromatographic methods [84] based on the distribution of substances between a stationary and a mobile phase in conjunction with various detectors play an important role in qualitative and quantitative analysis, especially of organic substances (see also ion chromatography, Section Anion Analysis). *Gas chromatography* with a gas as the mobile phase is especially suitable for volatile substances (i.e., substances that evaporate without decomposition). The stationary phase consists of a solid (adsorbent) or fluid. The highest separation efficiencies are attained in capillary columns. Liquid chromatography with a liquid mobile phase is conducted as *thin layer (planar) chromatography* (TLC) or as *high-performance liquid chromatography* (HPLC) in columns (with small particles of 3 – 10 µm diameter as separating materials (for details, see [84]).

Gas Chromatography. In general, sample preparations are required for gas chromatography. These preparations involve direct transfer of the substances to be analyzed from the water sample to the separation column or first to an organic (easily evaporable) solvent. *Headspace analysis* (⟶ Gas Chromatography – Static and Dynamic Head Space Analysis) is used for volatile organic water constituents such as halogenated hydrocarbons [88], [89]. The sample of water is heated to a given temperature in a headspace sample vial provided with a silicone membrane and rubber seal. A controllable equilibrium is established between the liquid and the gaseous phase. With a gastight syringe, a defined volume can be withdrawn from the headspace via the silicone membrane and rubber seal of the sample vial and used for GC analysis without the interfering matrix water. In the *purge-and-trap process*, a highly pure stream of helium gas is first passed through the water sample. The highly volatile organic compounds enter the helium gas phase (purge process) and are then retained and concentrated in an adsorption tube (with an adsorbent such as tenax, i.e., a linear polymer of *para*-2,6-diphenylphenylene oxide; trap process). The substances are desorbed by rapid heating (thermal desorption) and then transferred directly to the sample inlet of a gas chromatograph (⟶ Gas Chromatography – Thermal Desorption Units). Other sample preparation processes for GC analysis are liquid – liquid (e.g., with *n*-pentane) and solid – liquid extraction (concentration and separation from water on adsorbents made of chemically modified silica gels with functional groups such as *n*-alkyl, usually C_{18}, phenyl, aminopropyl etc.). Microextraction is also possible [90], [91]. The efficiency of a GC analysis system is determined not only by the separation column but also by the type of detector used. For relevant water constituents, a compilation of GC processes with the detector used in each case is presented in Table 22 (examples: pesticides [92], polychlorinated biphenyls [93], coupling with a mass selective detector [94], and mass spectrometry [88]).

Table 22. Applications of GC analysis

Group of substances	GC process or detector (DIN)
Halogenated hydrocarbons with low volatility	electron capture detector (ECD) (DIN 38 407-F 2)
Highly volatile halogenated hydrocarbons	
After extraction with, e.g., *n*-pentane or *n*-hexane	ECD (DIN 38 407-F 4)
Headspace analysis	ECD (DIN 38 407-F 5)
Phenoxyalkane carboxylic acids with liquid – liquid extraction and derivatization	mass spectrometric detector (MSD) (DIN 38 407-F 14)
Phenols (selected) after derivatization	ECD (DIN 38 407-F 15)
Aliphatic hydrocarbons	flame ionization detector (FID)
Organic solvents	FID
Selected organic nitrogen and phosphorus compounds (pesticides) after concentration by solid-phase extraction	phosphorus – nitrogen detector (PND) (DIN 38 407-F 6) combination of GC and mass spectrometer
Aromatic hydrocarbons	photoionization detector (PID)
Hydrocarbons	combination of GC and FTIR (Fourier transform infrared spectrometry)
Benzene, toluenes, xylenes (BTX)	FID (DIN 38 407-F 9)
Vinyl chloride	ECD (DIN 38 413-P 2)
Polychlorinated biphenyls (PCBs)	ECD

Thin layer chromatography is widely applied, especially in the analysis of polycyclic aromatic hydrocarbons. For this purpose, one-dimensional or two-dimensional separation is conducted after extraction with chloroform in accordance with

DIN 38 409-H 13. This analysis is limited to six representative PAHs. Quantitative evaluations can be achieved by means of a thin layer scanner via UV absorption or fluorescence measurement.

Automated multiple development (AMD) is used as a special development technique in the separation of selected organic pesticides (DIN 38 407-F 11). This is a modern technique for stepwise development, which — unlike the classical process — starts with the solvent having the highest polarity and ends with the solvent having the lowest elution strength. Development is carried out in a vertical chamber using high-performance (HP) TLC plates. It is started with a run of 8 mm and is continued up to 25 times with a 3-mm run each time. After each step, the solvent is removed from the chamber by suction, and each step is preceded by layer conditioning. The AMD technique with the important effect of spot focusing is one of the most efficient developments in TLC.

High-Performance Liquid Chromatography. In comparison with GC and TLC, HPLC has the advantage that aqueous phases can be used on reversed-phase materials (silica gels with *n*-alkyl groups such as C_{18}, RP-18, ⟶ Liquid Chromatography). Thus, extraction of the substances to be analyzed from water samples and transfer to an organic phase are usually not required. Concentration can also be conducted in a precolumn directly before the separation column. Another advantage (especially over TLC) is the large selection of different types of detectors (UV – VIS detectors with fixed or variable wavelengths or diode array systems for the detection of spectra; and fluorescence, amperometric, and conductivity detectors) and the possibility of coupling with mass spectrometry [95]. The level of instrument technology makes a high degree of automation possible — starting with sample application, precolumn concentrations, gradient elution (e.g., increasing methanol concentration in a water — methanol eluent mixture), and recording of spectra, and ending with evaluation. In water analysis, this gradient elution is applied mainly to separation of pesticides (DIN 38 407-F 12) [96] and polycyclic aromatic hydrocarbons (ISO 7981–1, DEV process F 8) [97].

3.5. Biochemical Methods

Biochemical and biological testing methods [98] are used to test the effects of chemical and physical parameters, i.e., especially of substances dissolved (present) in water, on enzymes (biochemical) or on certain test (water) organisms. The term " effects" refers not only to the acute and chronic toxicity of individual substances, but also to detrimental biological effects in general caused by all the noxious agents together (i.e., total noxiousness). Synergistic and antagonistic effects are also discerned. Although the qualitative and quantitative identification of single substances is possible in chemical and physicochemical analyses (Sections Sampling and Sample Preservation , Physicochemical and Sum Parameters, Inorganic Analysis, Organic Analysis), no information can be provided on the joint action of several potentially toxic substances. If a damaging effect is determined in the proceses described below, the substances responsible must be identified by chemical analysis. Thus, biochemical and biological test methods are of significance as a precursor to instrumental analysis. They represent an important preliminary test and supplement the analytical determination of substance concentrations.

The enzymatic and biochemical processes employed include enzyme inhibition tests and enzyme-coupled immunochemical tests, i.e., enzyme-linked immunosorbent assays (ELISA; ⟶). ELISA tests use the ability of organisms to form antibodies against foreign higher-molecular mass substances and to bind the foreign substances with these antibodies. By coupling environmentally relevant low-molecular mass substances to larger proteins, antibodies are produced that respond specifically to these low molecular substances (haptens) [99]. The antibodies obtained are immobilized on the inner wall of plastic tubes (or in the cavities of microtiter plates) for an antigen – antibody reaction. The water sample and an enzyme conjugate, which consists of the analyte coupled with an enzyme, are then placed in a tube of this type. The free analyte molecules in the sample and the analyte molecules bound to the enzyme compete for the free binding sites on the immobilized antibodies. After a rinsing step, a solution containing the enzyme substrate is added. A dye is produced as a result of the enzymatic reaction. The higher the intensity of the color, which can be measured photometrically, the lower is the concentration of analyte molecules in the sample. The ELISA tests are available for a series of pollutants, especially for herbicides such as alachlor, cyanazine, atrazine, 2,4-dichlorophenoxyacetic acid, and metolachlor; for insecticides such as aldicarb, carbofuran, and chlorodienes, such as chlorodane, aldrin, dieldrin, heptachlor, endrin, and endosulfan; for benomyl, metalaxyl, and the fungicide captan; and also for PAH and polychlorinated biphenyls (PCBs).

Enzyme Inhibition Tests. Of the enzyme inhibition tests, the acetylcholinesterase inhibition test for phosphate esters, phosphate thioesters, and methyl carbamate is of special interest ⟶ Enzyme and Immunoassays – Enzyme Assays: Determination of Inhibitors) [100]. In this test, enzyme activity is inhibited by substances that act as neurotoxins. The inhibitory effect can be determined photometrically on the basis of the substrate reaction (e.g., with thiocholine). A *biosensor* is produced by combining an enzyme immobilized on a membrane with a pH (flat-bed) electrode [101]. The substrate reaction can then be determined from potential measurements. For the analysis of a drinking water sample, the activity of the enzyme membrane of the electrode is determined by dipping the membrane into a substrate-containing (acetylcholine) solution. As a result of enzymatic hydrolysis, acetic acid is released in the enzyme membrane. The resulting shift in pH is measured by a pH electrode and is a measure of the enzyme activity. After incubation in an inhibitor-containing sample of drinking water, the active center of the enzyme is blocked irreversibly, e.g., by phosphate ester molecules. Consequently, the activity of the enzyme is reduced. Subsequent determination of activity with the same substrate solution as before shows a lower activity. In this manner, detection limits of 0.1 µg/L can be achieved under optimum incubation conditions without previous analyte enrichment.

Biological Tests. Of the biological tests, the luminescent bacteria test DIN 38 412-L 34 and L 341) [102] is widely applied because it is easy to handle. A NaCl solution is added to the deep frozen test charges to give a temperature of 15 °C, and the bioluminescence (as a part of the bacterial metabolism) is measured. After addition of a wastewater sample and an incubation time of 30 min, the luminescence is measured again. The decrease in luminescence is a measure of bacterial toxicity. For wastewater analysis, dilution series are made to determine the dilution for an inhibition of 20 % by means of interpolation.

The reviews "Water Analysis" in the journal *Analytical Chemistry* survey the developments in methods and analyses over a period of two years [103].

[Top of Page]

4. Hydrological Cycle and Water Use

Fritz H. Frimmel and Birgit C. Gordalla

4.1. World Water Balance

The appearance of the earth is characterized by the natural occurrence and physical properties of water. As the basis of all life, water is an essential component of nature. Although the total global water volume of 1.38×10^9 km^3 accounts for only about 0.23 ppm of the mass of the earth, it covers 71 % of the earth's surface. The distribution of water among different regions of occurrence is shown in Table 23. Fresh water, including ice deposits, accounts for only 2.6 % (36×10^6 km^3) of the total volume. With reasonable effort, only about 10 % of the fresh water available is technically accessible to man at present (Table 24). Thus, the water supply of the earth's population will be based essentially on this volume of about 3.7×10^6 km^3.

Table 23. Assignment of water according to regions of occurrence according to [104]

Region	Volume, 10^6 km^3	Percentage
Oceans	1348	97.39
Polar icecaps, icebergs, and glaciers	27.82	2.01
Groundwater and soil moisture	8.06	0.58
Lakes and rivers	0.225	0.02
Atmosphere	0.013	0.001
Total	1384	100

Table 24. Breakdown of proportions of total fresh water usable for water supply according to [104]

Region	Percentage	Volume, 10^6 km^3
Underground (up to a depth of 800 m)	9.86	3.55
Lakes	0.35	0.13
Rivers	0.003	0.001
Atmosphere	0.04	0.014
Total	10.25	3.7

4.2. Hydrological Cycle

Many of the physicochemical properties of water, which lead to its normal appearance in nature, are due to the strong dipole character of the H_2O molecule (dipole moment: 6.13×10^{-30} Cm, 1.85 D, see also Chap. Water as a Solvent).

The fact that water occurs in nature in all three states of aggregation is of special importance. The flow of energy that reaches the earth's surface with incident solar radiation results in an average surface temperature of 15 °C [105], with considerable variation as a function of time and location. This energy input causes the evaporation of water from oceans, other surface waters, and vegetation. In the atmosphere, water vapor condenses and can be transported as visible clouds by wind. Water returns to the surface of the earth as rain, snow, and ice. It then either superficially flows away, seeps away, or contributes directly to surface water. Thus, it is again available for evaporation, partly via vegetation (Fig. 14). On an annual average, the earth's 149×10^6 km^2 of dry land receives 745 mm of precipitation (111 000 km^3), of which 477 mm (71 000 km^3) evaporates from the land and 40 000 km^3 flows into the sea [106].

Figure 14. Schematical representation of the hydrological cycle

From a global point of view, this cycle is closed (i.e., no water is lost). In individual regions, however, considerable deviations from the theoretical mean values for the mass fluxes occur due to long-range transport and climatic differences. Even in a relatively small country such as Germany, which belongs to the temperate climate zone, substantial differences are observed. Thus, the average temperature in July is between +17 °C and +18 °C in the North German lowlands, but about +20 °C in the Upper Rhine rift. The amounts of annual precipitation are between 500 and 700 mm in the North German lowlands, between 700 and 1500 mm in the low mountain ranges, and up to more than 2000 mm in the Alps [107]. The long-term average annual temperature in Germany is +9 °C. Since 1990, increases of up to 2 °C have been observed in the annual temperature and deviations of up to 5 °C in the monthly average values compared with the long-term values, have

been recorded. Despite these temperature variations with time, the precipitation balance remained almost unchanged. The long-term average value for precipitation over the area covered by the western German states (248×10^3 km^2) is 837 mm/a, while the value for the eastern states (former German Democratic Republic, 109×10^3 km^2) is 612 mm/a. Taking the small distance (a few hundred kilometers) between these two regions — that are in the same climatic zone — into consideration, this variation is remarkable. For the entire Federal Republic of Germany, this gives an average precipitation level of 768 mm/a, which corresponds to a precipitation of 274×10^9 m^3/a [105]. The water supply for Germany ensues from the difference between the long-term average values for precipitation and evaporation plus the long-term average value for inflow from other countries (Table 25) [107].

Table 25. Annual supply of water in Germany (based on 30 years of balance data, reference period 1931 – 1960)

	Precipitation levels and amounts of water	Relative values, %
Average supply of water	163.5× 1 0^9 m^3/a	100
Precipitation minus evaporation	95×10^9 m^3/a	58
Average inflow from other countries	68.5×10^9 m^3/a	42
Average precipitation	768 mm/a (274×10^9 m^3/a)	100
Average evaporation	501 mm/a (179×10^9 m^3/a)	65

4.3. Demand for Water

About 70 wt % of the human body consists of water. The WHO works from the fact that an adult weighing 60 kg must consume 2 L water per day in the form of drinking water, beverages, or food to avoid damage to his health. The quantities that apply to small children (body weight 10 kg) and infants (body weight 5 kg) are 1 and 0.75 L, respectively. A much higher water consumption per person is observed, depending on the standard of living and the economic structure. The FAO speaks of a shortage of water in a certain country or region when the annual water supply from domestic sources is <1000 m^3 per resident. Thus, a world population of 6×10^9 people will require at least 6×10^{12} m^3 water per year. Although this amount is only about 0.16 % of the total water available for drinking purposes, clearly discernible bottlenecks occur in some countries. They result from an inadequate regional supply of water. Examples of countries having a shortage of water (FAO, as of 1990) are given below, together with the amounts of water per citizen per year:

Rwanda	885 m^3
Botswana	797 m^3
Algeria	754 m^3
Burundi	663 m^3
Syria	615 m^3
Kenya	613 m^3
Tunisia	465 m^3
Israel	372 m^3
Jordan	222 m^3
Yemen	222 m^3
Mauretania	203 m^3
United Arab Emirates	188 m^3
Libya	154 m^3
Saudi Arabia	148 m^3
Egypt	35 m^3

Germany has an annual water supply of about 2000 m^3 per person. According to data released by the Statistical Federal Office (Statistisches Bundesamt), the total water requirement in 1991 was 47.9×10^9 m^3 and was distributed as shown in the following:

Cooling water, thermal power stations	28.8×10^9 m^3

Industry	11×10^9 m^3
Public water supply	6.5×10^9 m^3
Agriculture, irrigation	1.6×10^9 m^3

A breakdown of the utilization of water provided by the public water supply is given below:

Households and small businesses	68 %
Industry	13 %
Loss in the mains and consumption by waterworks	13 %
Public facilities	6 %

In households and small businesses, a daily water consumption of about 145 L per person was estimated over the past ten years. With 140 L per person (as of 1994), a downward trend has been observed in the past two years [108]. A breakdown of water consumption in households is given in the following:

Baths, showers, hygiene	40 %
Toilets	31 %
Laundry	13 %
Dish washing	7 %
Cleaning	5 %
Garden	2 %
Cooking, drinking	2 %

The length of the mains employed by the public water supply to distribute water covers about 386 000 km (estimation for 1994) [108]. The materials used for pipes (as of 1992) are listed below [109]:

Cast iron	49 %
Plastic	26.5 %
Cement	13 %
Steel	8.7 %
Other	2.8 %

In Europe, the statistical data for water consumption per person are in the range of 116 to 260 L/d (see below):

Austria	215 L
Belgium	116 L
Denmark	175 L
Finland	150 L
France	161 L
Germany	145 L
Hungary	153 L
Italy	214 L
Luxembourg	183 L
The Netherlands	173 L
Norway	167 L
Spain	131 L
Sweden	195 L
Switzerland	260 L
United Kingdom	161 L

4.4. Source of Water Used

Sources used by the public water supply in Germany as of 1994 include [108]

Groundwater	63.6 %
Springs	7.8 %

Riverbank filtration and replenishment of groundwater	15.3 %
Surface water	13.3 %

Groundwater is the most important natural source of drinking water in Germany [110].

Water resources are used regionally in different ways, depending on geographic location and on geological, pedological, and hydrological conditions (Fig. 15).

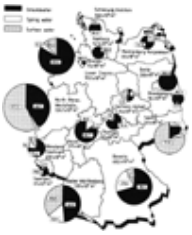

Figure 15. Regional differences in the use of water resources in Germany (as of 1994) [108]. By courtesy of the Bundesverband der deutschen Gas- und Wasserwirtschaft e. V. (Federal Association of German Gas and Water Suppliers)

As a result of their earlier history, different waters differ in their chemical composition and biological nature. Water pollution protection was the first issue of European Community environmental policy [111] and has been most extensively dealt with. The quality requirements demanded for the production of drinking water are stipulated in a European guideline [112] and, in Germany, in technical regulations [113], [114]. The requirements to be fulfilled by drinking water are also formulated in an EC guideline [115] which has to be transferred into national law of the member states. In the German decree on drinking water [116], [117], these requirements are translated into national law. The guiding principles for technical realization of the production and distribution of drinking water contained in DIN 2000 [118] have also been included partly in the decree on drinking water.

In the United States, the EPA establishes maximum contaminant level goals (MCLG) and maximum contaminant levels (MLC) as well as best available technologies for regulated water contaminants that are stipulated in the Safe Drinking Water Act (SDWA) with current amendments. A review of promulgated, proposed, and anticipated EPA drinking water regulations, which also concern disinfectants/ disinfection byproducts, analytical methods, information collection, rankings of regulatory priorities, and public notification requirements, is given in [119].

If the requirements stipulated for drinking water are not met by raw water, it must be treated before it is used as drinking water.

4.5. Water Treatment

A wide range of processes is available for the treatment of water (e.g., [120], [121]). The selection of individual treatment stages and the entire concept of treatment depend primarily on the water quality of the source used, its productivity, and consumer requirements. This means that optimum water treatment depends in each case on a particular situation, which is usually unique. For this reason, many different concepts are encountered in practice. In principle, the individual units of many treatment systems are largely similar. The water treatment concept of the Wiesbaden – Schierstein waterworks is presented in Figure 16. The large number of technical stages used permits the conversion of Rhine water to drinking water, which not only is qualitatively perfect but corresponds quantitatively to the demands of the population.

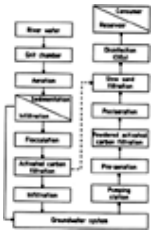

Figure 16. Principle of water treatment in the Wiesbaden – Schierstein waterworks

The versatile use of the Rhine and — despite relevant decrees and laws — the poor raw water quality found, as well as the effects of disasters and disturbances on water quality, are a real challenge to a water treatment strategy. The individual treatment stages shown in Figure 16 comply with the state of the art. In the case of better raw water conditions, some of these stages can be omitted. In the total concept of the Wiesbaden – Schierstein waterworks, the key function of the subsoil as a natural or, at least, close-to-natural treatment stage is discernible (see also [122], [123]). The bank filtrate of the river water supplements the dominant role played by groundwater in the public water supply of Central Europe. Through utilization of subsoil, nature takes on the function of technical measures in wide areas by means of its cleansing power. The reliable use of this natural self-cleansing power requires a knowledge of its capacity and responsible, future-oriented management. This includes a knowledge of the hydrological and geological situation, the allotment of protective zones with restricted use of the land, and extensive prevention of groundwater pollution [124]. The entire water utilization cycle with water above and below ground must also be regarded and assessed as a closely linked system. The German water management law (Wasserhaushaltsgesetz) [125] provides the legal framework for this purpose.

Typical concentrations of water constituents as present in some types of raw and finished water (drinking water) are given in Table 26. The limiting values found in the German decree on drinking water are given for comparison. Larger German water-supplying companies have been selected as examples. Note that the drinking water supply of the city of Munich is based on

the direct utilization of *groundwater* without technical treatment: i.e., the drinking water quality is the same as the raw water quality. The Lake Constance water supply is cited as an example of the utilization of *standing surface waters* (lakes, reservoirs), and the water supply of Düsseldorf reflects the use of the Rhine *riverbed filtrate* (riverbank filtrate) as a typical example of utilization. Accordingly, characteristic compositions are obtained for the raw waters. Groundwater has a relatively high water hardness (calcium and magnesium ions), especially when it comes from limy subsoil, has a slightly elevated nitrate content compared with the geogenic background, and otherwise shows the absence of larger influences of civilization, which is guaranteed by the protective function of the ground. These influences would be discernible from increased values for boron and phosphate. The water from Lake Constance (raw water) is correspondingly softer and exhibits the characteristic chemical features of surface water, e.g., a noticeable concentration of organic water constituents (DOC). In addition, slight anthropogenic influences are discernible, which can be seen from the fact that phosphate and the widely used EDTA are detected. The raw water quality of the Rhine bed filtrate is characteristic of the load of a variedly and greatly used flowing body of water, a load that is reduced by the natural cleansing power of the ground.

Table 26. Typical concentrations of constituents and properties of raw and drinking water compared to selected limiting values in the decree on drinking water

Parameter	Calculated Limit as		Groundwater Munich [a]	Surface water Lake Constance [b]		Riverbank filtrate (Rhine near Düsseldorf) [c]	
			Raw/drinking water	Raw water	Drinking water	Raw water	Drinking water [d]
Arsenic, mg/L	As	0.01	<0.001	0.0014	0.0014	0.002	0.002
Cadmium, mg/L	Cd	0.005	<0.001	<0.00002	<0.00002	<0.0001	<0.0001
Chromium, mg/L	Cr	0.05	<0.003	0.00020	0.00018	<0.0005	<0.0005
Cyanide, mg/L	CN	0.05	<0.0005	<0.0005	<0.0005	<0.01	<0.01
Fluoride, mg/L	F	1.5	0.13	0.083	0.08	0.09	0.09
Lead, mg/L	Pb	0.04	<0.003	<0.0002	<0.0002	<0.001	<0.001
Mercury, mg/L	Hg	0.001	<0.0005	<0.000002	<0.0000013	<0.0003	<0.0003
Nickel, mg/L	Ni	0.05	<0.003	0.00065	0.00067	<0.0008	<0.0008
Nitrate, mg/L	NO_3	50	8.5	4.55	4.65	16.3	16.6
Nitrite, mg/L	NO_2	0.1	<0.005	<0.005	<0.005	<0.01	<0.01
Polycyclic aromatic hydrocarbons (sum), mg/L	C	0.0002	<0.00003	<0.00005	<0.00005	<0.00005	<0.00005
Organochlorine compounds (sum), mg/L		0.01	<0.0005	<0.0001	<0.001	<0.005	<0.005
Tetrachloromethane, mg/L	CCl_4	0.003	<0.0001	<0.00002	<0.00002	<0.0001	<0.0001
Herbicides and pesticides, mg/L		single substance 0.0001	single <0.00003	single <0.00005	single <0.00005	single <0.0001	single <0.0001
		total 0.0005				total <0.00016	total <0.0001
Polychlorinated, polybrominated biphenyls and terphenyls, mg/L			n. deter.	single <0.000005	single <0.000005	single <0.00001	single <0.00001
Antimony, mg/L	Sb	0.01	<0.001	0.00013	0.00013	<0.0018	<0.0018
Selenium, mg/L	Se	0.01	<0.001	0.00014	0.00013	<0.0005	<0.0005
Color, m^{-1}		0.5	<0.1	n. detect.	n. detect.	0.08	<0.01
Turbidity, TE/F		1.5	0.05	0.20	0.07	0.2	0.1
Threshold odor number		2 at 12 °C		n. detect.	n. detect.	3 at 20 °C	1 at 20 °C
		3 at 25 °C	1 at 25 °C				
Temperature, °C		25	10.4	5.0	5.2	13.0	13.0
pH value at *T*		6.5<pH<9.5	7.62	8.05 (5.0 °C)	7.93 (5.2 °C)	7.24 (13.0 °C)	7.26 (13.0 °C)
pH after calcite saturation at *T*	pH_C			7.91	7.89	n. deter.	n. deter.
Electric conductivity (at 25 °C), µS/cm		2000	537	322	320	702	698
Oxidizability, mg/L	O_2	5	0.5	3.7	2.7	3.11	0.60
Aluminum, mg/L	Al	0.2	<0.02	0.0073	0.0035	0.003	0.009

Ammonium, mg/L	NH_4	0.5	<0.05	<0.01	<0.01	<0.01	<0.01
Barium, mg/L	Ba	1	<0.005	0.018	0.018	0.09	0.08
Boron, mg/L	B	1	<0.05	0.025	0.025	0.09	0.09
Calcium, mg/L	Ca	400	82.4	48.9	48.6	78	76
Chloride, mg/L	Cl	250	5.5	5.3	5.7	66	65
Iron, mg/L	Fe	0.2	<0.01	0.0055	0.0019	<0.06	<0.06
Kjeldahl nitrogen, mg/L	N	1	n. deter.	<1.0	<0.3	n. deter.	n. deter.
Magnesium, mg/L	Mg	50	19.6	8.0	8.0	10.8	10.7
Manganese, mg/L	Mn	0.05	<0.001	0.0008	0.00014	0.01	<0.01
Phenols, mg/L	C_6H_5OH	0.0005	n. deter.	n. deter.	n. deter.	<0.02	<0.02
Phosphorus, mg/L	PO_4	6.7	<0.01	0.074	0.077	0.32	1.22
Potassium, mg/L	K	12	1.0	1.3	1.3	4.8	4.8
Silver, mg/L	Ag	0.01	<0.001	<0.00005	<0.00005	<0.0002	<0.0002
Sodium, mg/L	Na	150	3.6	4.5	4.5	41	41
Sulfate, mg/L	SO_4	240	25.3	33.4	33.3	65	66
Hydrocarbons, mineral oils, mg/L		0.01	n. deter.	n. deter.	n. deter.	<0.01	<0.01
Surface active substances							
Anionic substances, mg/L	MBAS [e]	0.2	n. deter.	<0.1	<0.1	<0.01	<0.01
Nonionic substances, mg/L	BiAS [f]		n. deter.	n. deter.	n. deter.	n. deter.	n. deter.
Copper, mg/L	Cu	3	<0.001	0.00075	0.0007	0.025	0.001
Zinc, mg/L	Zn	5	<0.01	0.0006	0.0012	0.008	<0.004
Number of colonies at 20 °C		100	0 – 2	21	0	0 – 1	0 – 1
at 36 °C		100	0 – 2	9	0	3	0 – 1
Escherichia coli/100 mL at 36 °C		0	n. detect.	10	n. detect.	0 – 1	0
Coliforms/100 mL at 36 °C		0	n. detect.	10	n. detect.	26	0
Residue on evaporation, mg/L			328.5	194		n. deter.	n. deter.
Residue on ignition, mg/L			n. deter.	158		n. deter.	n. deter.
Base capacity up to pH 8.2, mmol/L			0.26 (10.4 °C)	0.04 (9 °C)	0.06 (9 °C)	0.51 (16 °C)	0.51 (16 °C)
Acid capacity down to pH 4.3, mmol/L				2.45 (20 °C)	2.44 (20 °C)	3.26 (18 °C)	3.21(18 °C)
Carbonate hardness, °dH			14.2 °KH	6.86	6.83	9.0	8.8
Sum alkaline earth (total hardness), mmol/L			2.869	1.58	1.58	2.39	2.34
°dH			16.1	8.85	8.85	13.4	13.1
Silicon dissolved, mg/L	Si		1.9	1.34		4.2	4.7
Strontium, mg/L	Sr		n. deter.	0.44		n. deter.	n. deter.
Oxygen, mg/L	O_2		9.2	10.0	26.6	2.8	4.7
DOC, mg/L	C		<0.2	1.2	1.1	1.17	<0.55
Spectral absorption coefficient (254 nm), m^{-1}			n. deter.	3.06	1.32	3.48	0.76
AOX, mg/L	Cl		n. deter.	0.005	0.026	<0.010	<0.010
Nitrilotriacetic acid (NTA), mg/L			n. deter.	<0.0005	<0.0005	<0.0005	<0.0005
Ethylenediaminetetraacetic acid (EDTA), mg/L			n. deter.	0.0022	0.0015	0.0036	0.0005
Total γ activity, Bq/L	$^{137}Cs + ^{134}Cs$		<5	<0.1	<0.1	n. deter.	n. deter.

[a] By courtesy of Stadtwerke München.
[b] By courtesy of Bodensee Wasserversorgung.
[c] By courtesy of Stadtwerke Düsseldorf.
[d] n. deter. = not determined; n. detect. = not detectable.
[e] MBAS = Methylene blue-active substance.

[f] BiAS = bismuth-active substance.

For the two treated waters, comparison of raw water and drinking water data shows clearly that many parameters are not significantly changed after treatment. The effects of oxidative and adsorptive water treatment steps are observable in the improvement of both the organoleptic values, such as color, turbidity, and odor, and the microbiological parameters (decrease in coliform organisms and number of colonies). These quality criteria also represent the primary target values of drinking water treatment. Furthermore, a decrease is observed in the values of those parameters that are partially (spectral absorption coefficient, oxidizability) or exclusively (DOC) a measure of the concentration of organic water constituents [126]. The elimination of organic water constituents is important to reduce the danger of future regrowth of bacteria in the mains during water distribution and also to restrict the formation of trihalomethanes during future chlorination. Halogen-containing disinfection byproducts can also increase the AOX value produced by industrial wastewater. The increase of phosphorus in the finished water from the riverbank filtrate is due to the addition of a phosphate – silicate mixture as a corrosion inhibitor.

Despite the quality improvement of surface water (e.g., [127] discernible since the 1980s, the need for efficient water treatment continues to exist. This results from the obligation of the public water supply system to provide sufficient amounts of qualitatively perfect drinking water at all times. In addition, precautions against accidents and disasters that can impair the quality of surface waters and are not excludable require appropriate protective measures.

Water for *technical use* is also of particular significance. Great demands are made on boiler feed water, process water, water for pharmaceutical and food technology uses, and water for swimming pools. These demands can be met only by purpose-oriented water treatment processes (see e.g., [128], [129] and Chaps. Ion Exchange and Membrane Separation Processes in Water Treatment). As a rule, a combination of various process steps is required.

[Top of Page]

5. Adsorption Processes in Water Treatment

Klaus Johannsen and R. Scott Summers

5.1. Introduction

Adsorption processes are used for the removal of dissolved organic substances for production of drinking water from both surface and groundwaters. The most important adsorbent for water treatment is activated carbon, and it was initially applied to improve the taste and odor of water. Activated carbon is also used to remove micropollutants, such as pesticides and halogenated hydrocarbons, and to limit the formation of disinfection byproducts (DBPs), such as trihalomethanes, by removal of natural organic matter (NOM), which can form disinfection byproducts upon chlorination. Activated carbon is applied in two forms: powdered activated carbon (PAC) and granular activated carbon (GAC). Powdered activated carbon is added as a suspension to the water to be purified and removed by sedimentation or filtration. In GAC treatment, fixed beds are used that contain the activated carbon. Monographs give an overview of the fundamentals and application of activated carbon in drinking water treatment [130-134].

5.2. Properties of Activated Carbon (see also ⟶ Carbon, ⟶ Carbon – Mechanical Properties, ⟶ Carbon – Adsorption Properties)

The most common raw materials used in the production of activated carbons are bituminous coal, peat, wood, lignite, coke, and coconut shells. Typical production processes include charring by pyrolytic carbonization of the raw material at <700 °C, followed by thermal activation with oxidizing gases such as steam, CO_2, and air at 800 – 1000 °C. Development of the pore structure in this manner yields internal surface areas of 500 – 1100 m^2/g. The ability of activated carbon to adsorb large quantities of material is related directly to its porous nature. Pore volume and surface area distributions are determined by mercury porosimetry and nitrogen and benzene absorption. The intitial portions of the nitrogen adsorption isotherms are often utilized to calculate the total pore surface area of an adsorbent. Standardized aqueous test methods, such as the molasses test, methylene blue number, phenol test, and iodine number, have been developed to evaluate the equilibrium adsorption capacity of PAC or GAC. In particular, the iodine number gives a good indication of the microporosity and is often correlated with the BET surface area [130], [131].

To minimize operational problems when using GAC in packed beds, the particle size and particle-size distribution must vary only within certain limits. Typical average particle diameters of *granular activated carbon* are in the 0.9 – 1.6-mm range. Smaller particles yield faster overall adsorption kinetics; however, too many particles smaller than 0.5 mm may cause high pressure losses in the fixed bed. Typical average particle diameters of *powdered activated carbon* are between 0.01 and 0.04 mm. Smaller particle sizes impede handling and separation of PAC. Values for the material density of GAC are in the 1800 – 2000 kg/m^3 range. The dry particle density of GAC (i.e., dry mass per particle volume) is between 500 and 850 kg/m^3, while its wet particle density is between 1300 and 1500 kg/m^3. The apparent (bed, bulk, or filter) density (i.e., dry mass per vessel volume) of GAC is in the 300 – 500-kg/m^3 range; that for PAC is slightly higher (400 – 700 kg/m^3) [130], [131], [133], [134].

5.3. Adsorption Theory (⟶ Adsorption)

Adsorption is the accumulation of a substance, the adsorbate, onto a solid, the adsorbent; it is driven by physical or chemical forces.

Physical adsorption is based on nonspecific electrostatic forces, such as dipole – dipole interactions, hydrogen bonding, and van der Waals forces. In general, nonpolar compounds dissolved in a polar solvent (water) are more strongly adsorbed by a nonpolar adsorbent (activated carbon). Compared with chemical adsorption (chemisorption), physical adsorption is less specific, operates over longer distances, has weaker forces and energies of bonding, and is reversible to a greater degree.

Chemisorption forces approach that of a covalent or electrostatic chemical bond, and because of the specificity of the bond

between adsorbate and surface, the adsorbate cannot accumulate at more than one molecular layer; i.e., a monolayer of adsorbate is formed. In water treatment, physical adsorption dominates.

In general, adsorbability on activated carbon is inversely related to the water solubility of the adsorbate and the pH. The adsorption of a substance increases with increasing number of double bonds (π-bonds) in the molecule, increasing molecular size, and increasing ionic strength of the solution. However, for high molar mass compounds such as humic substances, size exclusion can be observed [131]. A good relationship between the surface area or volume of micropores of activated carbons and their adsorption capacity for several compounds has been found [131], [134].

5.4. Adsorption Equilibrium (⟶ Adsorption – Thermodynamics, Equilibria, and Heat of Adsorption)

Adsorption equilibrium studies — termed isotherms, because they are conducted at constant temperature — establish the relationship between the concentration of adsorbate in the liquid phase *c* and its concentration in the solid phase *q*. For determination of equilibria, the *bottle point technique* [131], [134] is normally used. Here, defined quantities *m* of PAC or ground GAC are added to a series of bottles containing a volume of solution *L* with an initial adsorbate concentration c_0. The samples are mixed until equilibrium is reached: i.e., no measurable change in the liquid-phase concentration occurs. Activated carbon is separated from the liquid phase, and *c* is determined. Based on a mass balance for fresh carbon, the solid-phase concentration *q* is calculated according to

$$q = L(c_0 - c)/m \tag{5.1}$$

The equilibration time depends on particle size. For carbon with particle sizes <0.1 mm, an equilibration time of 2 – 5 d is adequate.

The two most commonly used models to describe the equilibrium relationship between *c* and *q* are the Langmuir [135] and the Freundlich equations [136]. Both are two-parameter equations. The Langmuir equation has a maximum solid-phase concentration value. In the low-concentration range, a linear relationship exists between the concentration of adsorbate in the solid and the liquid phase. The Freundlich equation is a empirical power function of

$$q = K_F c^n \tag{5.2}$$

where K_F is the Freundlich constant and *n* is the Freundlich exponent. These two parameters can be determined from the measured data by nonlinear regression of *c* and *q* or by linear regression of log *c* and log *q*. The slope of log *c* versus log *q* is equal to the exponent *n*, and the value of *q* at *c* = 1 equals K_F. The Freundlich equation is most often used to describe experimental data, but it is valid only in the concentration range of these data. Experimental adsorption data for a variety of compounds and carbons are summarized in [131], [137], [138].

When more than one adsorbable component is present in a solution, the compounds compete with each other for available adsorption sites on the carbon surface. If the complete reversibility of adsorption is assumed, the ideal absorbed solution (IAS) theory accounts for these interactions by using data only from the single-solute isotherms [139-143]. The ideal absorbed solution theory has the flexibility to incorporate many of the single-solute isotherm model equations; thus, for multicomponent systems it can be used with the model that best describes the single-solute data. The use of the IAS theory coupled with the Freundlich equation was applied successfully to multicomponent systems with solutes of similar characteristics [131]. However, in other situations, the IAS theory was not as successful and modifications were made to account for unequal competition for active sites and irreversible adsorption [144]; changes in the Freundlich parameters at low concentration [141], [145]; and differences in accessible surface area [146].

Most (80 – 95 %) of the natural organic matter in water is adsorbable. However, the composition of natural organic matter in water is complex; only a small percentage of the components can be identified specifically, and quantification of the few identifiable compounds is problematic because of their low concentration. Typically, NOM is quantified by the amount of dissolved organic carbon or the UV absorbance at 254 nm. This complexity of natural organic matter presents a problem in describing adsorption equilibria. The *adsorption analysis approach*, developed to describe the adsorption of an unknown mixture, divides NOM into fictive components of different adsorbabilities and uses the IAS theory to quantitate their competition [131]. The adsorption analysis approach has been applied successfully to several waters to demonstrate the influence of initial NOM concentration, pH, treatment processes, carbon type, and presence of tracer compounds. One drawback is that it does not account for the effects of size exclusion, which is likely important for the molecular size range of NOM [131].

In most polluted raw water, the NOM concentration is much higher than that of specific micropollutants to be removed. The adsorption capacity of carbon for the specific micropollutants can be lowered significantly in the presence of adsorbing NOM. Figure 17 illustrates the influence of background NOM on the adsorption isotherms of the pesticide metazachlor [*N*-(2,6-dimethylphenyl)-*N*-(1-pyrazolylmethyl)-chloroacetamide] [147]. At both pesticide initial concentrations c_0, the presence of 10 mg/L DOC causes a reduction of ca. 60 % in the adsorption capacity. Furthermore, Figure 17 shows the impact of the initial concentration of pesticide on the isotherm. At c_0 = 2 µg/L, the adsorption capacity is much lower than at c_0 = 200 µ/L for both DOC values. Thus, even the presence of 0.3 mg/L DOC causes a capacity reduction at the low initial concentration of the micropollutant.

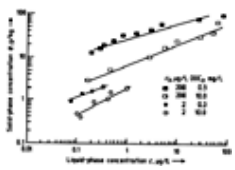

Figure 17. Influence of initial concentration and background DOC concentration on metazachlor adsorption isotherms [147]

These effects can be explained by the occurrence of competitive adsorption between metazachlor and background NOM. Models have been developed that calculate the competitive adsorption between target compound and background NOM using the IAS theory [147-150]. These competitive effects have been found for most pesticides, but not for some chlorinated hydrocarbons [131].

5.5. Adsorption Kinetics (⟶ Adsorption – Kinetics)

Although adsorption at an active site of the carbon surface is a fast reaction, the resistance to mass transfer of the adsorbate from the bulk liquid to the adsorption sites is high enough that the overall kinetics of the adsorption process must be considered in the design of both PAC and GAC systems. The driving force for mass transfer is the concentration difference of the adsorbate in the bulk solution and in the carbon particle. This exists as long as the solid-phase concentration on carbon is not in equilibrium with the liquid-phase concentration.

For the fixed-bed application of GAC, the breakthrough curve for the ideal plug-flow reactor (PFR) with no mass-transfer resistance is shown in Figure 18 (curve a). This ideal breakthrough curve may be used to define the time at which the amount of adsorbate applied to the GAC bed equals the amount that can be adsorbed at equilibrium. This time is sometimes referred to as the stoichiometric breakthrough time t_{stoich} and can be calculated from the following mass balance equation and isotherm data:

$$c_0 t_{stoich} \dot{V} = q_0 m \tag{5.3}$$

where c_0 is the influent concentration, q_0 is the solid-phase concentration in equilibrium with c_0, $\dot{V}$ is the volumetric flow rate, and m is the adsorbent mass. Mass-transfer resistances, axial dispersion, and unfavorable equilibrium conditions can cause the breakthrough curve to deviate from ideal PFR behavior as shown for a single solute in Figure 18 (curve b). The total amount of solute adsorbed should be the same for the ideal PFR and the case with mass-transfer resistance and is the difference between the quantity applied and that in the effluent. This can be calculated by the following mass balance:

$$\dot{V} = \int_0^\infty (c_0 - c_e(t))\mathrm{d}t = q_0 m \tag{5.4}$$

where c_e is the effluent concentration. The value of the integral corresponds to the area bounded by the $c = c_0$ line and the breakthrough curve. It represents the total capacity of the GAC in the reactor. Since the total amount adsorbed satisfies Equation (5.4) regardless of the shape of the breakthrough curve, the shaded areas F_1 and F_2 in Figure 18 must be equal. If they are not equal, the calculated capacity from the isotherm q_0 does not agree with the column capacity.

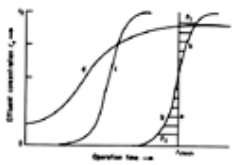

Figure 18. Comparison of breakthrough curves

a) Ideal (negligible mass-transfer resistance); b) Single solute; c) Specific solute in a mixture; d) NOM

In practice, solutes to be removed from water are present in mixtures and, as discussed in Section Adsorption Equilibrium , the presence of other adsorbates — both more weakly and more strongly adsorbing — will decrease adsorption of the target compound. This is also illustrated in Figure 18 (curve c). The concentration and adsorbability of the different components of the mixture determine the shape and time of the breakthrough curve of the target compound. If the mixture contains high concentrations of solutes that are adsorbed more strongly, c_e of the target compound may become greater than c_0 (i.e., “overshoot” may occur). If the mixture contains mostly specific compounds, breakthrough of the target compound can be calculated by using competitive adsorption models. If significant concentrations of NOM are present, competitive adsorption models are not viable, when a component of NOM is adsorbed irreversibilly. Compared to the target compound, this NOM component progresses deeper into the GAC bed and is adsorbed first. Because of its irreversible adsorption, this NOM component “blocks” the GAC surface (fouling), thus reducing the adsorption capacity for the target compound. Currently, only site-specific calibrated models can simulate this adsorption behavior [131], [134].

The breakthrough behavior of NOM is important in controlling the formation of DBPs and is illustrated in Figure 18 (curve d). Typically, an immediate breakthrough of 5 – 20 % is displayed, and if this level is maintained for any length of time, this fraction is considered nonadsorbable as shown in Figure 18. If c_e begins to rise immediately, then the immediate breakthrough may also include a fraction due to external mass-transfer resistance. The NOM breakthrough curve normally begins to plateau at a value 10 – 30 % lower than c_0, which can be attributed to very slow adsorption or biological degradation [131], [134].

A wide range of mathematical models have been developed to account for equilibria and kinetic components such as axial dispersion, external (film) mass transfer, and intra particle pore and surface diffusion [131]. The most commonly used model for micropollutants is the plug-flow homogeneous surface diffusion model, which includes external mass transfer and intra particle surface diffusion, and can utilize the IAS theory and the Freundlich equilibria model [131]. A simplified model considers the internal mass transfer as a linear driving force [151], [152]. For NOM adsorption the pore diffusion component seems to be important.

5.6. Biological Processes

Most drinking water sources contain a biodegradable NOM component in a concentration of 5 – 20 %, which is increased an additional 5 – 20 % by ozonation. Since this fraction is not removed in most treatment plants, biological degradation occurs to

some degree in all GAC contactors. In addition to removing some of the biodegradable NOM fraction and associated DBP precursors, biological degradation activity on GAC can extend its adsorption capacity by reducing the overall organic load to the GAC. The negative aspects of biological activity on GAC include growth of zooplankton in the GAC bed, increased microorganism concentrations in the effluent, release of colonized GAC fines that are difficult to disinfect, and under anaerobic conditions, the formation of taste- and odor-causing compounds. Depending on temperature and biodegradability, component bioacclimation times are in the range of 5 – 30 d. Those GAC columns that are run for extended periods of time and reach a steady-state removal level, as shown in Figure 18 for NOM, are often termed biologically activated carbon (BAC) systems. Biological activated carbon systems rely on biodegradation for removal of pollutants, but the GAC adsorption capacity attenuates influent pulses, which allows the biomass to degrade desorbing substances. Compared to other support media such as sand, GAC also has the advantage of providing larger accessible surface area for attachment of biomass. When preozonation of water is used, the interaction with GAC must be considered, because ozonation increases biodegradability but decreases the adsorbability of NOM on GAC by formation of polar substances such as aldehydes, ketones, and ketoacids. An ozone dose in the range of 1.0 – 0.5 g O_3 per gram DOC is sufficient for disinfection and NOM oxidation without significantly decreasing adsorbability [131], [134].

5.7. Design of GAC Systems

The important parameters and typical values used to characterize the design and operation of GAC adsorbers are listed below [131], [134]:

Bed volume V_F	10 – 50 m^3
Cross-sectional area A_F	5 – 30 m^2
Length l	1.8 – 4 m
Filter velocity v_F	5 – 15 m/h
Empty bed contact time *EBCT*	5 – 30 min
Operation time t_F	100 – 600 d
Throughput *BV*	4000 – 30 000 m^3/m^3
Carbon use rate *CUR*	5 – 50 g/m^3

Of particular importance are the filter velocity v_F (also termed the superficial linear velocity or surface loading rate), which is defined as

$$v_F = \frac{\dot{V}}{A_F} \tag{5.5}$$

and the empty-bed contact time *EBCT*, which is defined as

$$EBCT = \frac{V_F}{\dot{V}} = \frac{l}{v_F} \tag{5.6}$$

For a given flow rate, v_F determines the filter area, while *EBCT* determines the length and volume of the carbon bed, both of which influence the capital costs of a GAC system. The *EBCT* also has an impact on operation and maintenance costs, because shorter *EBCT*s lead to faster breakthrough of pollutants and subsequently more frequent replacement of the GAC bed is required. The throughput parameter in bed volume *BV* is a normalized expression of the volume of water treated V_L or the filter operation time t_F and is defined as

$$BV = \frac{V_L}{V_F} = \frac{t_F}{EBCT} \tag{5.7}$$

The carbon use rate *CUR* is the mass of GAC required to treat a volume of water to achieve a given treatment objective and is defined by

$$CUR = \frac{m}{V_L} = \frac{\varrho_F}{BV} \tag{5.8}$$

where ϱ_F is the density of the filter layer. For a given flow rate, *CUR* dictates the GAC demand from which the cost of GAC replacement or reactivation can be calculated. Background NOM decreases the GAC adsorption capacity for specific organic compounds and thus increases *CUR*. Values of *CUR* predicted from distilled water isotherm data were compared with the observed *CUR* of pilot and full-scale column runs for which breakthrough data were available [153]. Figure 19 shows the relationship between the natural water correction factor (observed *CUR*: distilled water *CUR*) and the distilled water *CUR*. For compounds that are strongly absorbed by the carbon bed (i.e., low *CUR*), the natural water correction factors are higher (10 – 100) than for substances that are adsorbed only weakly. This is due to the fact that in treating strongly adsorbed compounds, the GAC is exposed to NOM for a longer time, thus, more fouling occurs.

Figure 19. Relationship between natural water correction factor and distilled water carbon usage rate [153]

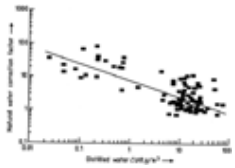

In groundwater treatment plants, the GAC is normally placed at the end of the treatment train but before final disinfection. In surface water treatment plants, GAC can be used as a filter adsorber or placed as a stand-alone unit after filtration, termed the postfilter adsorber. In both cases, predisinfection of water with chlorine-based disinfectants should be avoided because the disinfectants can react with the GAC and thus reduce its adsorption capacity for organic compounds and form undesired chlorinated compounds [134]. Old water filters that have been retrofitted as filter adsorbers tend to have low *EBCT*s, 3 – 9 min, which limits their application in most cases to taste and odor improvement and chemical spill control. Postfilter adsorbers generally have longer *EBCT*s, 10 – 30 min. They can be used effectively to control a wider range of compounds, including micropollutants, NOM, and DBP precursors.

Prediction and Simulation of GAC Breakthrough Behavior. For site-specific design purposes, a variety of approaches have been applied to predict or simulate GAC breakthrough behavior. Parallel pilot columns that utilize the same GAC and influent water as the full-scale system are accurate and reliable predictors of breakthrough behavior [154], [155]. They also have the advantage of not requiring the use of numerical models in interpretation of the data, but they have the same operation time as the full-scale system. Another approach is to collect isotherm and adsorption kinetic data and use the data as input to fixed-bed numerical models for predicting breakthrough [131], [151], [152], [156], [157]. Some extensions of these models have been made to take into account the impact of NOM on the breakthrough behavior of micropollutants [131], [147].

Minicolumns utilize small particle and column sizes to decrease the operating time of a column to a small percentage of that required in the full-scale column. *Small-scale columns* are minicolumns in which similitude to full-scale GAC systems has been maintained. The relationship among particle size, column length or *EBCT*, and operation time of small-scale and full-scale columns is determined by dimensional analysis. Successful application of small-scale columns yields breakthrough curves that are equivalent to those of a full-scale adsorber. The method was pioneered by FRICK [131] and was developed further and applied extensively by CRITTENDEN and coworkers [158-160], leading to the rapid small-scale column test (RSSCT). The major problem is that the impact of background NOM on the adsorption capacity and diffusion kinetics of micropollutants is not always reflected accurately in small-scale columns. Rapid small-scale column tests have been used successfully to predict the breakthrough of NOM and DBP precursors [160], [161].

5.8. Design of PAC Systems

Powdered activated carbon can be applied effectively to control periodic taste and odor problems and spills of synthetic organic compounds in raw source water. The PAC dosage (i.e., the mass of PAC per volume of water treated) typically ranges from 5 to 50 mg/m^3 for the improvement of taste and odor. Long-term removal of specific compounds, such as pesticides and volatile organic carbons, NOM, and DBP precursors is possible but usually only at uneconomically high activated carbon dosages. If long-term high dosages of PAC are required, GAC systems are usually more economical. The advantages of PAC compared to GAC are (1) low capital costs, because PAC is added to existing contactors; (2) rapid utilization of the adsorption capacity, because PAC is 25 times smaller than GAC; and (3) flexibility in operation, because the dosage, and to some degree the point of application of PAC, can be controlled easily. Its disadvantages are that PAC cannot be reactivated and that it can be carried over into the distribution system [131], [134].

Powdered activated carbon can be applied dry or as a slurry at several points in a surface water treatment plant, as shown in Figure 20 (points A – E). The most common points of addition are at the coagulant rapid mixing tank (B) and the flocculation tank (C). The PAC is removed from the process stream during sedimentation and filtration, and effective contact times of 30 – 60 min can be achieved. Addition of PAC at the raw water intake (A) results in longer contact times, especially if presedimentation is employed. Each PAC addition point has advantages and disadvantages with respect to contact time, kinetics, and competing organic compounds. The PAC added at point A is not incorporated initially within the floc particles, as it is when added at points B and C; i.e., addition at point A allows for faster adsorption kinetics. When PAC is added at point A, the disadvantage is increased competition for active sites from organic compounds that are lated removed by coagulation. The PAC added at point D will have the shortest contact time, but incorporation of floc and competing organic compounds is decreased. When PAC is added at point E, incorporation of competing organic compounds and floc is at a minimum and the PAC is retained within the filter for longer times. This all results in an increased adsorption capacity, thus improving effectiveness. However, filter run times are shortened, and PAC breakthrough into the distribution system must be avoided because the presence of PAC in the distribution system can lead to bacterial regrowth problems. The retention of PAC particles within filters can be aided with the addition of polyelectrolytes [131], [134].

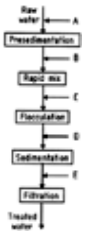

Figure 20. Flow sheet for PAC application in a conventional surface water treatment plant

5.9. Reactivation of GAC

Activated carbon has a finite adsorption capacity, and after exhaustion of this capacity, the used GAC must be replaced with fresh or reactivated GAC. Reactivation conditions are similar to those employed in manufacturing activated carbon by thermal activation, which means that part of the carbon surface may be included in the reaction. Typically, thermal reactivation of GAC occurs in four stages: (1) thermal desorption of the absorbed compounds; (2) thermal decomposition of the adsorbed compounds followed by desorption of the decomposition products; (3) carbonization of the nondesorbed products formed during thermal decomposition or chemisorbed during the adsorption step; and (4) surface reactions between the carbonaceous residuals and water vapor or oxidizing gases to form gaseous products. Each of these stages is associated with a particular temperature range and consists of several simultaneous steps. If properly designed and operated, reactivation can yield 100 % recovery of the adsorption capacity. However a 5 – 15 % loss of GAC occurs during a reactivation cycle, which includes backwashing, transportation, and reactivation; this necessitates the use of fresh makup GAC [131], [134].

[Top of Page]

6. Ion Exchange

Wolfgang Höll

(see also ⟶ Ion Exchangers)

In water technology, ion-exchange resins are used to eliminate cations or anions and replace them with other cations or anions. In this way the composition of the components of the water is changed, and as a consequence, a change in water properties occurs. Among the components of natural water carbonic acid and its dissociation products play a particular role because these species form the dominant buffer system of water. This buffer system is affected by almost every ion-exchange process. Beyond the mere exchange of ionic species the properties of water are therefore influenced in a way that is characteristic for each type of exchange. The fundamentals of ion exchange and aspects of its application have been presented in several books and innumerable publications [162-167]. This chapter illustrates the influence of ion exchange on the properties of the water treated, i.e., its chemical state, which may be characterized by pH value and alkalinity.

6.1. Fundamentals

6.1.1. *Characterization of the Chemical State of Water*

Inorganic Components of Natural Water. The main inorganic components of natural water include the following:

Cations of strong electrolytes	Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}
Anions of strong electrolytes	SO_4^{2-}, NO_3^-, Cl^-
Species of carbonic acid	CO_2, H_2CO_3, HCO_3^-, CO_3^{2-}
Water and its dissociation products	H_2O, H^+, OH^-
Silica	SiO_2

In addition, natural water usually contains nonionic impurities (dispersed substances that cause turbidity, colorants, colloids, or organic matter), which are not discussed in this chapter.

Dissociation of Carbonic Acid and Water and the Solubility of Calcium Carbonate. The above list shows that carbonic acid is present in water in the form of four different species. If nondissociated H_2CO_3 is ignored because of its low concentration, the sum of the concentrations of carbonic acid species — usually designated the total inorganic carbon (TIC) — is

$$\mathrm{TIC} = c(\mathrm{CO_2}) + c(\mathrm{HCO_3^-}) + c(\mathrm{CO_3^{2-}}) \tag{6.1}$$

The concentrations of CO_2, HCO_3^-, and CO_3^{2-} are linked by the dissociation of carbonic acid [dissociation constants K_1 (T, I) and K_2 (T, I), where I denotes ionic strength and T is temperature].

The ionic strength I is defined by

$$I = 0.5 \sum z(i)\, 2c(i)$$

where z (i) is the valence of species i. Further dependencies are normally negligible. By using the ionic strength, activity coefficients of monovalent species can be calculated according to

$$\log f(i) = -\frac{0.5\sqrt{I}}{1 + 1.5\sqrt{I}} + CI$$

For I < 0.1 mol/L the second term is negligible. Activity coefficients of bivalent species result from f (z = 2) = f (z = 1)^2.

The dissociation constants K_1 and K_2 (for Θ = 25 °C, I = 0 mol/L) can be calculated according to [168], [169]

$$pK_1^0 = \frac{17052}{T} + 215.21 \cdot \log T - 0.12675\,T - 545.560$$
$$pK_2^0 = \frac{2902.39}{T} + 0.02379\,T - 6.4980$$

The relative amount of negative charges of carbonic acid species is expressed by the so-called equivalent factor Φ:

$$\phi = \frac{c(HCO_3^-) + 2c(CO_3^{2-})}{TIC} \tag{6.2}$$

which lies in the range $0 \leq \Phi \leq 2$ [170].

The dissociation of water molecules is characterized by the ionic product of water:

$$c(H^+) \cdot c(OH^-) = K_W(T, I) \tag{6.3}$$

If the liquid phase is in equilibrium with solid $CaCO_3$, the concentrations of the respective ions are linked by the solubility product:

$$c(Ca^{2+}) \cdot c(CO_3^{2-}) = K_L(T, I) \tag{6.4}$$

The solubility product K_L^0 of $CaCO_3$ (for Θ = 25 °C, I = 0 mol/L) can be calculated from [171]

$$\log K_L^0 = -8.468 - 507.59\left(\frac{1}{T_0} - \frac{1}{T}\right) - 18.65\left(\ln\frac{T_0}{T} + \frac{T}{T_0} - 1\right)$$

with T_0 being 298.15 K.

m and p Alkalinities. One of the fundamental properties of each electrolyte solution is its electroneutrality. Given the species on the above list, this condition can be expressed as

$$\begin{aligned} &2c(Ca^{2+}) + 2c(Mg^{2+}) + \\ &c(Na^+) + c(K^+) + c(H^+) \\ &= 2c(SO_4^{2-}) + c(NO_3^-) + c(Cl^-) \\ &+ c(HCO_3^-) + 2c(CO_3^{2-}) + c(OH^-) \end{aligned} \tag{6.5}$$

Rearrangement of this equation with respect to the origin of the species from either strong or weak electrolyes yields

$$\begin{aligned} &2c(Ca^{2+}) + 2c(Mg^{2+}) + c(Na^+) + c(K^+) \\ &-2c(SO_4^{2-}) - c(NO_3^-) - c(Cl^-) \\ &= c(HCO_3^-) + 2c(CO_3^{2-}) + \\ &c(OH^-) - c(H^+) \end{aligned} \tag{6.6}$$

The algebraic sum of equivalents on the right-hand side of this equation can be determined approximately by titration with methyl orange (color change at pH 3.1 – 4.4) as indicator. Both sides of Equation (6.6), therefore, define the so-called methyl orange alkalinity (*m* alkalinity, total alkalinity, TAC) [170]. By means of the abbreviation

$$\Delta = c(OH^-) - c(H^+) \tag{6.7}$$

the right-hand side of Equation (6.6>) can be written as

$$m = TIC \cdot \phi + \Delta \tag{6.8}$$

Carbonic acid-bearing solutions are usually characterized also by a second quantity which is derived from titration with phenolphthalein as the indicator (color change at pH 8.2–9.8), the so-called phenolphthalein alkalinity (*p* alkalinity):

$$\begin{aligned} &p = -c(CO_2) + c(CO_3^{2-}) + \\ &c(OH^-) - c(H^+) \end{aligned} \tag{6.9}$$

This relationship results from the definition of *m* alkalinity and from the condition

$$TIC = m - p \tag{6.10}$$

For natural water with pH values between 6 and 8.5 the *m* alkalinity approximately equals the concentration of HCO_3^- ions and the value of the *p* alkalinity that of carbon dioxide. The advantage of using *m* and *p* alkalinities is that they allow characterization of the water without tedious analyses of all inorganic components [170].

Interdependence Between Carbonate Balance and Solubility of Calcium Carbonate. In natural water the solubility of calcium carbonate is almost never adjusted correctly. The water is either oversaturated or undersaturated with $CaCO_3$. The state of saturation can be characterized by the saturation index *S I*:

$$SI = \log \frac{c(Ca^{2+}) \cdot c(CO_3^{2-})}{K_L} \tag{6.11}$$

For positive values of SI, $CaCO_3$ can be precipitated from water; for negative values, it can be dissolved.

The content of the components unaffected by either the precipitation or the dissolution of $CaCO_3$ is characterized by the parameter $m - 2\,c\,(Ca^{2+})$, which remains constant in both cases. Subtracting $2\,c\,(Ca^{2+})$ from both sides of Equation (6.8) and expressing $c\,(Ca^{2+})$ on the right-hand side by Equation (6.11) finally yield the relationship

$$\begin{aligned} &TIC^2 \cdot \phi + TIC \cdot \{\Delta - [m - 2c(Ca^{2+})]\} \\ &-2 \cdot K_L \left[1 + \frac{c(H^+)}{K_1} + \frac{c(H^+)^2}{K_1 K_2}\right] \cdot 10^{SI} \end{aligned} \tag{6.12}$$

which gives the interdependence between the carbonate balance and the solubility of calcium carbonate for an arbitrary value of the parameter $m - 2\,c\,(Ca^{2+})$ [170], [172].

6.1.2. *Representation of the Chemical State of Water*

One of the classical graphic representations of the carbonate balance and thus of the chemical state of water was introduced by TILLMANNS, who plotted the concentration of "free" carbonic acid [$c\,(CO_2)$] as a function of $c(HCO_3^-) + 2\,c(CO_3^{2-})$ [173]. With respect to this familiar plot, a generalized modification giving the (negative) p alkalinity as a function of the m alkalinity was chosen to illustrate the changes of the state of water in ion-exchange processes. The generalization is obtained by combining Equations (6.8) and (6.10) to yield the following relationship:

$$p = m \cdot \frac{\phi - 1}{\phi} + \frac{\Delta}{\phi} \tag{6.13}$$

Since both Φ and Δ are functions of pH, Equation (6.13) represents the equation of a straight line at constant pH [170]. The set of lines is plotted in Figure 21. The state of equilibrium with solid $CaCO_3$ corresponding to Equation (6.12) is given by the set of curves for which $m - 2\,c\,(Ca^{2+})$ has a given constant value. Since TIC cannot be negative, all possible states of solutions of salts of carbonic acid must lie above the line with $p = m$. The plot allows a simple representation of most of the changes of the chemical state of water.

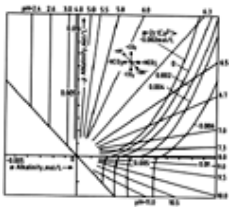

Figure 21. Interdependence of m alkalinity, p alkalinity, and pH (temperature 25 °C, ionic strength I = 0 mol/L)

Any mathematical calculation of carbonate balance requires that four quantities be known (i.e., temperature, ionic strength, and m and p alkalinities). By this means, an arbitrary fifth quantity can be calculated. Graphical plots allow representation of the interdependence of only three quantities; two more must be assumed to be constant. To fulfill this condition, all calculations were made for 25 °C and a constant ionic strength of 0 mol/L. The latter assumption is a crude simplification in view of the changes in composition of water. However, ionic strength affects only the calculated pH values (through the activity coefficients), whereas the m and p alkalinities remain unaffected.

Further graphical representations of the chemical state have been developed by HALLOPEAU and by CALDWELL and LAWRENCE [174], [175]. However, these plots are not discussed here.

6.1.3. *Ion Exchangers*

(⟶ Ion Exchangers, ⟶ Ion Exchangers – Properties) [165]

Resin Structure and Type. An ion-exchange resin consists of a polymer matrix containing functional groups. The matrix consists of polystyrene or of acrylic polymers, both cross-linked by means of divinylbenzene (DVB). The DVB concentration normally varies between 4 and 12 %. However, for particular purposes, resins with either less than 4 % or up to 25 % DVB are available.

The ability to exchange ionic species is due to the functional groups present in the resin. Either these must be fixed within the matrix structure by a suitable functionalization step during synthesis or they are already part of the organic monomers. According to the types of functional groups and their degree of either dissociation or protonation, strongly and weakly acidic or basic ion-exchange resins can be distinguished. Functional groups of common exchange resins are summarized below:

Strongly acidic $-SO_3^-H^+$
Weakly acidic $-COOH$
Strongly basic

	$-N^+(CH_3)_3OH^-$ (Type I)
	$-N^+(CH_3)_2C_2H_4OH^-$(Type II)
Weakly basic	$-NR'_2 \cdot H_2O$
	$-NRH' \cdot H_2O$
	$-NH_2H_2O$

Dissociation of Functional Groups. The dissociation of functional groups of cation and anion exchangers can be written as

$$\underset{\text{(cation-exchange resin)}}{\overline{R-H} \rightleftharpoons \overline{R^-} + \overline{H^+}} \tag{6.14a}$$

$$\underset{\text{(anion-exchange resin)}}{\overline{R-H^+} \rightleftharpoons \overline{R^-} + \overline{H^+}} \tag{6.14b}$$

Overbars refer to the resin phase. By applying the law of mass action, the following constants are obtained:

$$K_c = \frac{\overline{R^-} \cdot \overline{H^+}}{\overline{R-H}} \tag{6.15a}$$

$$K_a = \frac{\overline{R} \cdot \overline{H^+}}{\overline{R-H^+}} \tag{6.15b}$$

where the indices c and a denote cationic and anionic resins.

The p*K* values can be derived from these constants:

$$pK = -\log K \tag{6.16}$$

The p*K* values of cation- and anion-exchange resins are summarized in the following:

Strongly acidic	<1
Weakly acidic	4 – 6
Strongly basic	>13
Weakly basic	6 – 9

The range of operation of different resin types can be derived by means of p*K* values. If pH $\simeq$ p$\bar{H}$, the percentage of dissociation of charge-bearing functional groups exceeds 50 % in cation exchangers for pH > p*K* and in anion exchangers for pH < p*K*. As a crude rule, the p*K* therefore represents the limit of applicability of an exchange resin.

6.2. Cation-Exchange Processes

6.2.1. *Strongly Acidic Resins in the Na^+ Form*

Application of strongly acidic resins in the sodium ion form is the classical ion-exchange process for water softening. Softening makes possible the elimination of alkaline-earth ions from water. These species may cause troublesome $CaCO_3$, $Mg(OH)_2$, and $CaSO_4$ scales on heat-transfer surfaces. Furthermore, increased concentrations of alkaline-earth ions require increased consumption of detergents in washings. Softening processes have been introduced in industrial and drinking water treatment, both in municipal water treatment systems and in home softeners [176-178]. An equilibrium diagram for water softening is shown in Figure 22.

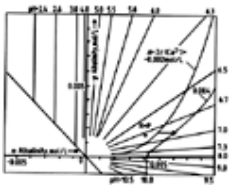

Figure 22. Equilibrium diagram for water softening. The following raw water composition is assumed: *c* (Ca^{2+}) = 3 mmol/L, *c* (Mg^{2+}) = 0.5 mmol/L, *c* (Na^+) = 1 mmol/L, *c* (Cl^-) = 2 mmol/L, TIC = 5.5 mmol/L

R = Raw water; P = Product water

In terms of a formal chemical reaction, the process reads as follows:

$$\overline{R-SO_3^-Na^+} + (Ca^{2+}, Mg^{2+}) \rightleftharpoons \overline{R-SO_3^-(Ca^{2+}, Mg^{2+})} + Na^+ \tag{6.17}$$

Symbols in parentheses indicate the stoichiometry of the exchange. Because of the greater affinity of the resin for bivalent species, calcium and magnesium ions are removed very efficiently. Sodium chloride solutions (brine) are used to regenerate exhausted resin. In practice, the resin is converted only partly to the sodium form; the relative amount of loading is about

50 % of the exchange capacity [176].

The pure neutral exchange of alkaline-earth species for Na^+ ions does not change the concentrations of the species involved in the dissociation of carbonic acid. However, the saturation index is strongly affected. The decreased calcium concentration results in a decrease in the numerical value of $m - 2\,c\,(Ca^{2+})$. Since the $CaCO_3 - H_2CO_3$ equilibrium is now described by the respective curve, the product water is far away from saturation and becomes lime agressive.

6.2.2. *Strongly Acidic Resins in the H^+ Form*

Strongly acidic resins in the H^+ form are normally applied in demineralization processes prior to the elimination of anions by means of anion exchangers. Because of the exchange of practically all cations for H^+ ions, salts are converted to the corresponding mineral acids:

$$\begin{aligned}&\overline{R-SO_3^-H^+}+\\&(Ca^{2+},Mg^{2+},Na^+,SO_4^{2-},Cl^-,HCO_3^-)\\&\rightleftharpoons\overline{R-SO_3^-(Ca^{2+},Mg^{2+},Na^+)}+\\&H_2SO_4,HCl,H_2CO_3\end{aligned}\qquad(6.18)$$

Regeneration is carried out by using HCl or H_2SO_4. However, with respect to the precipitation of $CaSO_4$, the concentration of sulfuric acid used for regeneration must be relatively low. As a consequence, the regenerant volume is greater. The amount of acid that must be applied depends on the tolerable leakage, mainly of sodium, during the service cycle.

As long as no carbon dioxide is degassed from the water, TIC remains constant. Therefore, the development of the state of water follows a straight line with a gradient equal to –1. The final state depends on the concentration of strong acids (Figure 23).

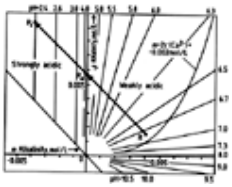

Figure 23. Change of chemical state of water during cation exchange with strongly or weakly acidic resins in the hydrogen form. The raw water composition in Figure 22 was assumed.

R = Raw water; P_w = Product water after weakly acidic resin; P_s = Product water after strongly acidic resin

6.2.3. *Weakly Acidic Resins in the Hydrogen Form*

If weakly acidic resins in the hydrogen form are allowed to come in contact with natural water, a so-called reaction-coupled exchange occurs: protons from carboxylic groups of the resin react with anions of weak acids, normally HCO_3^- and CO_3^{2-}. Thus, nondissociated carbonic acid is formed. To maintain electroneutrality, the resin adsorbs cations. Corresponding to their stronger affinity, alkaline-earth ions are mainly eliminated; Na^+ and K^+ are exchanged only if the raw alkalinity exceeds the sum of the equivalents of calcium and magnesium ions. With respect to the simultaneous elimination of alkaline-earth ions and of alkalinity, the process — called *softening with dealkalization* — represents a partial demineralization [176].

The process can be written as

$$\begin{aligned}&\overline{R-COOH}+\\&(Ca^{2+},Mg^{2+},Na^+,SO_4^{2-},Cl^-,HCO_3^-)\\&\rightleftharpoons\overline{R-COO^-(Ca^{2+},Mg^{2+})}\\&+(Na^+,H_2SO_4,HCl,H_2CO_3)\end{aligned}\qquad(6.19)$$

Softening with dealkalization can be applied as the first step during demineralization to eliminate a substantial part of the total salinity. Furthermore, it can be combined with a subsequent softening using a strongly acidic resin in the Na^+ form. In this way, cooling water or feed water for low-pressure boilers can be produced. Finally, softening with dealkalization also has been proposed to soften drinking water.

Strong acids must be used for complete conversion of the resin to the hydrogen form during regeneration. Since the functional groups show a very strong preference for protons, only slightly more than the theoretical amount of acid has to be applied. Furthermore, the acid need not be concentrated or very pure. Weakly acidic resins can also be regenerated by using weak acids (i.e., citric or carbonic acid). In the latter case, however, only partial conversion to the hydrogen form is achieved [179-182].

During the combined softening – dealkalization with weakly acidic resins, the chemical state of water changes along a line with a gradient of –1 until $m = 0$ and pH $\simeq$ 4, provided no CO_2 degasses. In practice, however, acrylic resins split part of the neutral salts. Therefore, the final m value is slightly negative at a pH of ca. 3.5. Figure 23 demonstrates the change of state of water for the application of strongly acidic and weakly acidic resins in the hydrogen form.

6.3. Anion-Exchange Processes

6.3.1. *Strongly Basic Resins in the Cl^- Form*

Strongly basic resins in the Cl^- form allow the elimination of troublesome anionic impurities such as NO_3^- and SO_4^{2-}. They have also been proposed and applied to the elimination of chromate, arsenate, and selenate [183-185]. In contact with

natural water, SO_4^{2-} and NO_3^- are exchanged easily. However, an uptake of nonpreferred species such as HCO_3^- by the resin also results:

$$\overline{R-N(CH_3)_3^+Cl^-} + (SO_4^{2-}, NO_3^-, HCO_3^-) \rightleftharpoons \overline{R-N(CH_3)_3^+(SO_4^{2-}, NO_3^-, HCO_3^-)} + Cl^- \quad (6.20)$$

Corresponding to the sequence of selectivity, subsequent zones with different resin loading are formed in filters. The boundaries between these regions pass the filter as traveling waves. Each nonpreferred species is replaced by stronger preferred ones and dumped out of the filter. As a consequence, the effluent undergoes variations mainly in HCO_3^- concentration. Initially, part of the bicarbonate is eliminated from water, accompanied by a decrease of pH. However, during a later part of the filter operation, bound bicarbonate species are replaced in the resin, which leads to higher concentration and pH values in the effluent than in the feed. These variations depend strongly on the individual properties of the resin and the composition of raw water. Therefore, this cannot be generalized. Consequently, the exchange of anions for Cl^- ions cannot be demonstrated graphically.

Strongly basic resins in the chloride form are frequently used to eliminate nitrate ions from drinking water. Unfortunately, standard strongly basic resins exhibit a marked preference for sulfate species. To decrease the predominant uptake of nonhazardous sulfate ions, so-called *nitrate-selective resins* have been developed. These are strongly basic resins of type I with ethyl or butyl instead of methyl groups [186], [187].

Apart from acting as ion exchangers the resins can also be used as adsorbents for organic matter such as humic or fulvic acids. For normal anion-exchange processes, regeneration is carried out by means of 4 – 8 % brine solution. To force the uptake of nonpreferred Cl^- ions, substantial amounts of brine in excess of the stoichiometric quantity are required [176]. These scavenger resins are regenerated by using alkaline brine solution.

6.3.2. *Strongly Basic Resins in the OH^- Form*

Strongly basic resins in the OH^- form replace all anions with OH^- ions. Normally, OH^- loaded resins are applied almost exclusively during demineralization in combination with a strongly acidic cation exchanger, either in subsequent filters or in mixed beds. If the feed water for the basic resin is the effluent of a strongly acidic resin, the anions of strong acids are exchanged for OH^- and water molecules are generated:

$$\overline{R-N(CH_3)_3^+OH^-} + (H^+, SO_4^{2-}, NO_3^-, Cl^-, H_2CO_3) \rightleftharpoons \overline{R-N(CH_3)_3^+(SO_4^{2-}, NO_3^-, Cl^-, HCO_3^-)} + H_2O \quad (6.21)$$

Because of the high degree of protonation of nitrogen atoms in quaternary ammonium groups even at relatively high pH, these resins are also capable of removing nondissociated weak acids such as carbonic acid or silica according to

$$\overline{R-N(CH_3)_3^+OH^-} + SiO_2 \cdot H_2O \rightleftharpoons \overline{R-N(CH_3)_3^+HSiO_3^-} + H_2O \quad (6.22)$$

Regeneration is carried out by means of 4 – 8 % aqueous NaOH solutions. Because of the weak affinity of the resins for OH^-, the regenerant solution must be of high purity.

In a graphical representation the uptake of anions of strong acids and of H_2CO_3 can be subdivided into two parts: (1) During the elimination of SO_4^{2-}, NO_3^-, and Cl^-, no change in TIC occurs. Therefore, the state changes along a line with TIC = constant until $m = 0$. (2) During the sorption of carbonic acid, the state changes along the ordinate toward $p = 0$. The real development runs as a superposition of both parts on an arbitrary line between the states of the feed and of the product water (Fig. 24, curve a). The elimination of silica cannot be represented in such a diagram.

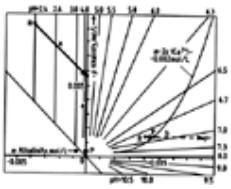

Figure 24. Change of chemical state of water during exchange of anions for OH^- (curve a) and for HCO_3^- ions (curve b). The raw water composition in Figure 22 was assumed.

R = Raw water; P = Product water; P′ = (Hypothetical) state of product water without precipitation of $CaCO_3$

6.3.3. *Strongly Basic Resins in the HCO_3^- Form*

Strongly basic resins of the HCO_3^- form exchange all anions for HCO_3^- ions:

$$\overline{R-N(CH_3)_3^+HCO_3^-} + (H^+, SO_4^{2-}, NO_3^-, Cl^-, H_2CO_3) \rightleftharpoons \overline{R-N(CH_3)_3^+(SO_4^{2-}, NO_3^-, Cl^-)} + HCO_3^- \quad (6.23)$$

According to Equation (6.23), anion exchangers in the bicarbonate form may be used to eliminate nitrate from drinking water [188], [189]. However, this kind of exchange can be applied only if the concentration of calcium ions is low. Otherwise, the

solubility product of calcium carbonate is exceeded, yielding uncontrollable precipitation of solid $CaCO_3$. Apart from their application as single exchangers, anion exchangers in the bicarbonate form are used in the Desal and Carix processes [190-192].

Regeneration can be carried out either in a single step by using $NaHCO_3$ or H_2CO_3 (+ MgO) or in two subsequent steps by means of NaOH or aqueous NH_3 and CO_2 [189], [193].

During the exchange of anions for HCO_3^- species, an increase of *m* at constant *p* is encountered. Thus, the state moves on a horizontal line. If the theoretical end point lay beyond the saturation curve, $CaCO_3$ would be precipitated, which limits the further increase of *m*. The state of the product water is then represented by the intersection of a line *p* = constant with the saturation curve (see Fig. 24, curve b).

6.3.4. *Weakly Basic Resins*

In neutral or alkaline media, weakly basic resins in the free-base form are only weakly protonated. Because of the corresponding lack of positive charges they cannot act as anion exchangers. As a consequnce, they can adsorb an-ions only from acidic media with pH <6. These resins are therefore able to eliminate strong acids from the effluent of a strongly acidic cation exchanger in the H^+ form. The uptake of carbonic acid by the resin is, however, negligible, with the exception of acrylic weakly basic resins. The exchange, which can also be classified as an adsorption of acids, can be represented as

$$\begin{aligned}&\overline{R-NR_2'\cdot H_2O}+(H^+,SO_4^{2-},NO_3^-,Cl^-,H_2CO_3)\\&\rightleftharpoons\overline{R-NR_2'H+(SO_4^{2-},NO_3^-,Cl^-)}\\&+H_2CO_3+H_2O\end{aligned}\qquad(6.24)$$

Regeneration requires alkaline solutions such as NaOH or aqueous NH_3. As with weakly acidic resins, the regenerant solution need not be concentrated or extremely pure. Because of the pH-dependent protonation of the amino groups, regeneration requires only slightly more than the theoretical amount. Like strongly basic exchangers, weakly basic resins can be used to remove humic substances.

The development of the chemical state of water treated with weakly basic resins closely resembles that of water treated with strongly basic resins. The negative *m* value of the feed water is increased until *m* = 0. Thus, the development follows a straight line with TIC = constant. The uptake of small amounts of carbonic acid by the resin leads to a decrease in TIC according to individual resin properties.

Figure 25 demonstrates the advantage of degassing CO_2 only after using the weakly basic resin. Because of the leakage of sodium in cation exchange, the *m* alkalinity of the effluent of a weakly basic resin becomes slightly positive. As follows from Figure 25, for this case the pH of the product water is high for low CO_2 concentrations and lower at high concentrations. Because of the pH-dependent protonation of the functional group the exchange capacity of the resin is necessarily greater at a high TIC.

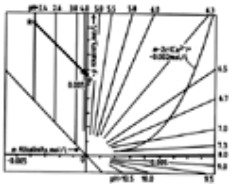

Figure 25. Change of chemical state of water during treatment of an acidic solution with a weakly basic resin (no uptake of CO_2). The raw water composition in Figure 22 was assumed.

R = Raw water; P = Product water

6.4. Combined Processes

6.4.1. *Demineralization*

The simplest form of demineralization of groundwater (e.g., for boiler feed water) consists of the subsequent application of a strongly acidic resin in the H^+ form and a strongly or weakly basic resin in the OH^- form or free-base form, respectively. Depending on the composition and quality of the fresh water to be treated and the desired product water quality, a weakly acidic resin can also be applied in the first step followed by a strongly or weakly basic anion-exchange resins [176]. This process design also saves regeneration costs. The development of the state of the water can be compiled easily from the previous sections.

Figure 26 demonstrates changes of the state of water. Furthermore, it clearly shows the advantages of using weak electrolyte resins for the removal of alkalinity and strong acids, and of efficient degassing prior to the use of a strongly basic resin.

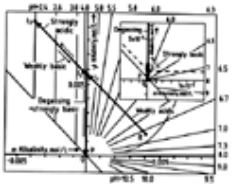

Figure 26. Change of chemical state of water during demineralization by combined application of weakly and strongly acidic resins, a weakly basic resin, a degassing step, and a strongly basic resin. The raw water composition in Figure 22 was assumed

R = Raw water; P = Product water; I_1 – I_4 = Intermediate states; – – – Equilibrium with atmospheric CO_2

For the demineralization of brackish water, other concepts have been proposed, for example, the Desal process [189], [190] (⟶ Ion Exchangers). The process is unique because of its exclusive use of weak electrolyte resins. Its key is the application of an acrylic weakly basic resin, which in contrast to styrene-based resins can adsorb HCO_3^- ions. In the first step, an anion exchanger converts all the salts to salts of carbonic acids, which can be split by a weakly acidic resin applied in the second step. Carbon dioxide evolved in this step is used for conversion of a second weakly basic resin from the free base to the HCO_3^- form. Despite its advantages, the Desal process has never been used on a pilot or technical scale.

6.4.2. *Partial Demineralization*

Partial demineralization is achieved either by the application of a sequence of (completely regenerated) exchange resins not allowing full demineralization or by the use of incompletely regenerated resins. The simplest partial demineralization using completely regenerated exchangers involves softening with dealkalization by using a weakly acidic resin in the hydrogen form (see Section Weakly Acidic Resins in the Hydrogen Form). Another example is the combination of a weakly acidic resin in the hydrogen form and a strongly acidic resin in the Na^+ form. The change of the state of the water can be derived easily from the corresponding plots.

Among the processes using incompletely regenerated resins, the Sirotherm and Carix processes (⟶ Ion Exchangers) are the most important.

The Sirotherm process, also developed for the treatment of brackish water, uses a resin with a mixed weakly acidic and weakly basic functionality [194], [195]:

$$\overline{R-COOH} + \overline{R-NR_2'} + NaCl \rightleftharpoons \overline{R-COO^-Na^+} + \overline{R-NH^+R_2'Cl^-} + H_2O \qquad (6.24)$$

The service cycle runs at ambient temperature; regeneration, however, is carried out by means of hot water, thus making use of the temperature-dependent dissociation of water. Since the treatment of NaCl solutions does not influence carbonate balance, no change in chemical state occurs as defined in Section Characterization of the Chemical State of Water.

The Carix process consists of the combined application of a weakly acidic resin in the free-acid form and of a strongly basic resin in the HCO_3^- form. Both resins are used as mixed beds. In the service cycle, the resin mixture replaces neutral salts with carbonic acid. The resins are regenerated as a mixture by using carbonic acid. The process is applied for partial demineralization of drinking water. The simplified formal representation of the process is [192]:

$$\begin{matrix} \overline{R-COOH} \\ \\ \overline{R^+-HCO_3^-} \end{matrix} + (Ca^{2+} + SO_4^{2-}) \rightleftharpoons \begin{matrix} \overline{R-COO^-(Ca^{2+})} \\ \\ R^+-(SO_4^{2-}) \end{matrix} + (H_2CO_3) \qquad (6.25)$$

In the regeneration step the exchange capacity generated on the cation exchanger is greater than that on the anion exchanger. As a consequence, the elimination of neutral salts in the service cycle is accompanied by a softening or dealkalization by the cation exchanger.

By using ion-exchange resins completely converted to the hydrogen and HCO_3^- forms, process development can be divided into three steps. First, the cation exchanger reduces the *m* alkalinity to about zero, accompanied by a corresponding increase in concentration of dissolved carbon dioxide; second, the anion exchanger replaces anions of strong acids with HCO_3^- species to yield an increase of *m* alkalinity. Thus, the cation exchanger can again replace cations by protons (Fig. 27). Because of the limited efficiency of carbonic acid, only partial regeneration of both resins is achieved. As a consequence, the exchange on both resins remains incomplete, yielding considerable leakage of both anions and cations. The possible development of the chemical state of the water is plotted in Figure 27.

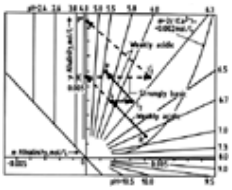

Figure 27. Development of the chemical state of water in the Carix process. The raw water composition in Figure 22 was assumed.

R = Raw water; P = Product water; P′ = Hypothetical product water; – – – = Hypothetical development with completely regenerated resins; ——— = Development under realistic conditions

[Top of Page]

7. Flocculation

Martin Jekel

7.1. Introduction

Most surface waters contain finely dispersed substances that cannot be separated completely by simple sedimentation or rapid filtration. These dispersed particles are responsible for the turbidity of water and may consist of minerals such as silica and clays, algae cells, bacteria, viruses, organic detritus, or insoluble pollutants [196]. In groundwater, particles can be formed by a preceding precipitation (e.g., lime precipitation for hardness removal) or by oxidation of iron(II) and manganese ions. The finely dispersed particles can be aggregated to form removable flocs by flocculation techniques using suitable

chemicals known as coagulants and flocculants.

The aggregates formed can be removed to a wide degree by subsequent separation processes such as sedimentation, flotation, or filtration. In the production of high-quality water such as drinking water, sedimentation and flotation usually are followed by rapid filtration, while direct filtration of flocs (see Chap. Filtration) is limited to raw water with low suspended solids and coagulant demand.

The removal of *particulate substances* is usually the primary objective of flocculation, but important cases exist in which *dissolved* inorganic and organic substances must be removed effectively by coagulants and flocculants via precipitation, coprecipitation, or adsorption onto the flocs. Examples are phosphates, some heavy metals, or dissolved organic substances of natural or anthropogenic origin. In view of the new standards for lower concentrations of organic substances in drinking water, the removal of DOC has become a major task in water treatment, especially in raw water colored by natural humic substances or having a high DOC [197].

7.2. Definitions

Scientific and technical terms that are used frequently in flocculation are listed and explained below [198-201]:

Colloids are particles with diameters in the range of 0.001 – 1 µm. They do not settle, and they cause turbidity. Their high specific surface areas induce important interfacial forces and phenomena between the colloid particles.

Suspended particles have diameters >1 µm and are detected under a light microscope. They may be removed by settling if their density is greater than that of water.

Aggregation denotes the general process of forming larger flocs from primary particles.

Stabilization. Colloidal and suspended particles can be stabilized against aggregation. Stabilization is caused by van der Waals attractive forces, by the electrostatic surface charge of the particles, by an adsorbed layer of polymers, or by forces of hydration. Most important in aqueous systems are the negative surface charges formed by the dissociation of protons from hydroxyl functions belonging to silica groups or from acidic carboxylic functions [196], [198], [199].

Destabilization. Stable particles are rendered unstable through neutralization of their negative surface charges by adsorption of positively charged coagulants.

Coagulation refers to (1) the reduction or elimination of electrostatic repulsion forces between particles via addition of certain coagulants, and (2) in technical terms, the first phase of floc formation after chemical mixing and destabilization, but before dosing of flocculants (see below) [199].

Flocculation. (1) Flocculation is the mechanism of floc aggregation by high molar mass polymers that adsorb on the particles and form bridges between them. (2) In water treatment, flocculation is the phase of formation of large flocs by addition of a secondary chemical (after addition of the coagulant) called a flocculant or flocculant aid [199]. (3) The term can be used for the entire water treatment process (coagulation plus flocculation), as in this chapter. In English-speaking countries, the terminology “coagulation and flocculation” is also used frequently to describe this process.

Coagulants are chemicals used for destabilization in the first phase of floc formation. Most common are the salts of Al^{3+} and Fe^{3+} ions; sometimes, organic cationic polymers are used.

Flocculants designate chemicals used to improve the formation of larger flocs. Flocculants are high molar mass water-soluble polymers. Flocculant aid or coagulant aid is used as a synonym.

Precipitation is the formation of insoluble substances from dissolved matter and the chemicals added. The important coagulants Al^{3+} and Fe^{3+} salts can precipitate as hydroxides in the neutral pH region and are thus removed from water.

Coprecipitation is the inclusion of dissolved or particulate substances into a precipitate (e.g., of aluminum or iron(III) hydroxide). Particles are then aggregated by the important *sweep coagulation mechanism*, with the amorphous, gelatinous hydroxide flocs acting as particle collectors.

7.3. Chemicals Used in Coagulation and Flocculation

7.3.1. *Inorganic Chemicals (* ⟶ Flocculants – Inorganic Flocculants*)*

The most important coagulants in water treatment are the salts of iron(III) and aluminum ions, available in various commercial forms and purities. Iron(III) chloride or aluminum sulfate (alum) is used frequently because of low cost. Mixed coagulants are also available, especially for applications in wastewater treatment. These coagulants dissolve readily in water, and the metal ions form hexaquo complexes, $Me(H_2O)_6^{3+}$, which are acidic species and give up protons [202].

The aquo complexes are stable below pH 3 – 4 (Al^{3+}) and 1 – 2 (Fe^{3+}). After stepwise formation of hydroxo – aquo complexes with increasing pH, insoluble hydroxides are formed at higher pH (Al^{3+}, pH > 5.5 – 6.0; Fe^{3+}, pH > 4.5 – 5) as voluminous, amorphous flocs. The residual concentrations of dissolved metal ions are well below the drinking water standards of 0.2 mg/L for Fe and Al. The $Al(OH)_3$ dissolves as aluminate $[Al(OH)_4^-$ at pH > 7.5 – 8.0, which must be avoided in water treatment [203]; $Fe(OH)_3$ redissolves at pH > 10 – 11 to form $Fe(OH)_4^-$.

The cationic partially hydrolyzed species present in a slightly acidic medium can condense to form polymeric products with a high positive charge. Various oligomeric and polymeric products are known [e.g., $Al_{13}(OH)_{32}^{7+}$ and $Fe_3(OH)_4^{5+}$]. For more

details on hydrolysis and condensation reactions, see [202]. Cationic aluminum polymers are available as commercial products or can be produced by partial hydrolysis and polymerization. These coagulants are usually called polyaluminum salts, but their composition and degree of hydrolysis vary widely.

Iron(III) and aluminum salts exhibit two distinctly different mechanisms in particle aggregation [204]:

1. Cationic polymeric species adsorb on negatively charged particles and neutralize them at an optimum dosage (charge and neutralization). Overdosing leads to restabilization because of charge reversal.
2. Insoluble hydroxides precipitate and enmesh the particles by the sweep coagulation mechanism. Optimum coagulation is achieved under conditions of maximum precipitation of the hydroxides. The nature and concentration of particles have no effect.

The sweep coagulation process is the dominant mechanism in most cases, whereas charge neutralization may be important in the coagulation of algae, humic colloids, or more concentrated suspensions and at a pH of <6.0 – 6.5.

The optimum dosages of Al^{3+} or Fe^{3+} may vary with water source and, because of frequent changes in raw water quality, even in the same plant within a few hours. The typical dosages given below are a preliminary guide for frequent cases:

1. Treatment plants for river water with settling: 0.05 – 0.2 mol/m^3 Me^{3+}
2. Treatment plants for lake and impounded water with direct or in-line filtration: 0.01 – 0.1 mol/m^3 Me^{3+}
3. Treatment plants for color and DOC removal: 0.01 – 0.05 mmol Me^{3+} per milligram DOC; optimum pH values for Al^{3+}: 5.5 – 6.5; for Fe^{3+}: 4.5 – 6 [205], [206]

Laboratory tests are usually performed to find the most suitable coagulation conditions before adjusting dosages in the plant.

Other inorganic chemicals are much less common, such as “activated silica,” a polymeric flocculant, or $Mg(OH)_2$, formed at pH >10.5 [200].

7.3.2. *Organic Chemicals (* ⟶ Flocculants – Natural Organic Flocculants, ⟶ Flocculants – Chemical Structure*)*

Organic substances added in flocculation are linear, water-soluble polymeres (synthetic, natural, or modified natural polymers) with mean molar masses in the range of 10^4 – 10^7 g/mol. The following monomers are included [200]:

Acrylamide (nonionic), acrylic acid (anionic), *N,N*-(dimethylaminopropyl)methacrylate (cationic), and ethyleneamine (cationic).

The resulting polymers can be nonionic, cationic, or anionic; the charge density can vary from low to very high, depending on copolymer composition. The molecular diameter of the polymers in solution is 10 – 500 nm.

Most commercial polymers are copolymers, containing at least two monomer types. The monomer content of these products is limited to 0.05 – 0.1 % if they are used in drinking water treatment [207].

Highly charged cationic polymers can be applied as coagulants, instead of Al^{3+} and Fe^{3+} salts because of their effective charge neutralization mechanism at optimum dosage [208]. The dosages required can be rather high (1 – 10 g/m^3).

High molar mass polymers with low to medium charge densities are flocculants, acting by the adsorption-bridging mechanism [198]. Typical dosages are in the range of 0.05 – 1.0 mg/L, and the polymer should be added at least 30 s after addition of the metal salts. Flocs formed with the aid of these polymers are larger, are more resistant to erosive and destructive forces, and settle faster.

In addition to synthetic polymers, natural or modified natural polymers such as starch, modified starch, or alginates can be used. Since natural polymers tend to be less effective, are unstable, and may support bacterial growth, they must be applied with more care.

7.4. Kinetics of Aggregation

The particle transport processes necessary for particle collisions are usually slower than mixing, chemical reactions (e.g., hydrolysis), adsorption, and destabilization. The kinetic rate of aggregation is thus determined by the particle collision rate, which itself depends on three basic transport processes [198], [209].

Particle diffusion is based on Brownian movement and dominates in concentrated colloidal systems if the particle diameter is <0.5 – 1 µm. This *perikinetic coagulation* is probably not important in water treatment or is limited to the first short phase until microflocs appear with diameters >1 µm.

Hydrodynamic shear gradients imposed by stirring or turbulent flow are responsible for different particle velocities. This important collision mechanism (*orthokinetic coagulation*) can be described for the initial period by

$$\ln N/N_0 = -4/\pi \cdot \alpha \cdot \phi \cdot \overline{G} \cdot t \qquad (7.1)$$

where N and N_0 are the total particle concentrations at time t or $t = 0$; α is the collision efficiency factor; $\bar{G}$ is the mean shear gradient (s^{-1}); Φ is the volume concentration of solids; and t is time.

The collision efficiency factor depends on the extent of destabilization of the particles (chemical factor), while $\bar{G}$ represents

shear conditions. The mean value $\bar{G}$ is used for turbulent conditions and can be calculated [198] according to the following: For back-mix reactors

$$\overline{G} = \left(\frac{P}{\eta \cdot V}\right)^{1/2} \tag{7.2}$$

where *P* is the power input of stirrer; *V* the reactor volume; and η the dynamic viscosity.

For pipes and channels

$$\overline{G} = \left(\frac{\Delta p \cdot v}{l \cdot \eta}\right)^{1/2} \tag{7.3}$$

where Δp is the pressure loss; *v* the flow rate; *l* the length of the pipe; and η the dynamic viscosity.

The orthokinetic equation is not longer valid if larger flocs are formed and destroyed again. In this case, the formula must be extended by a floc destruction term [210], describing the fact that larger flocs are formed at low $\bar{G}$ values, while high $\bar{G}$ values induce smaller flocs. In designing flocculation reactors, the use of two or more reactors can reduce overall reactor volume. The first reactor is operated at high $\bar{G}$ values for high collision rates but small flocs, whereas the second reactor is needed to produce large flocs at low $\bar{G}$ values (see Section Process Technology for Coagulation and Flocculation).

The final flocs are separated by either settling, flotation, or filtration, whereby the different floc velocities induce further collisions and floc growth. This third transport process is significant in settling and flotation tanks.

7.5. Process Technology for Coagulation and Flocculation

The major process steps in coagulation and flocculation are listed in Table 27. Steps 1 and 2 are physically different processes but are quite rapid and usually occur in the same reactor, the rapid mix system. Steps 3 and 4 are performed in one, two, or more reactors, depending on plant design.

Table 27. Process steps for coagulation and flocculation systems

Step	Task
Dosing of coagulants and rapid mixing	homogeneous distribution of chemicals
Destabilization	adsorption of coagulants on colloids and suspended particles, precipitation of dissolved substances
Aggregation to microflocs (coagulation)	rapid formation of small flocs from destabilized particles in high shear gradients, without flocculants (polymers)
Aggregation to macroflocs (flocculation)	formation of removable large flocs, with or without a flocculant (polymer)

7.5.1. *Dosing*

Coagulants and flocculants are generally prepared as a stock solution or delivered as a concentrated solution. Some restrictions exist on the concentration of stock solutions and their predilution before dosing. *Iron(III) and aluminum salts* may start to precipitate or hydrolyze at concentrations <1 % (Fe, Al) and pH values >1 – 2. Polymeric aluminum salts should be diluted only to the ratios given by the producer. All metal salts are acidic or alkaline solutions and require safety precautions. Iron(III) salts are especially corrosive.

Organic polymer solutions exhibit high viscosities at concentrations >0.2 – 0.5 %, depending on their molar mass. Dissolution sometimes requires several hours, and the stock solution may not be highly turbulent because of loss of efficiency and should be stored for only a limited time.

National and upcoming international standards exist regarding the purity of all applied chemicals to avoid secondary pollution in the finished drinking water. Cationic polymers, especially, are not approved for use in several countries because their risks are unknown.

7.5.2. *Rapid Mixing*

The small flow of a dissolved coagulant and flocculant must be distributed rapidly in the large water flow by suitable reactors. Efficient, rapid mixing is very important for destabilization of colloids. The present state of knowledge and technology is summarized in [211]. Rapid mixing reactors can be designed as plug-flow or back-mix systems. The latter apparently have some deficiencies in terms of coagulation, because the coagulant species may not be transported evenly to all particles in a short time. Mixing studies indicate that the metal salts should be mixed completely within 0.1 – 1 s, which may be difficult to achieve in large plants [211].

Possible rapid mixers are listed in Table 28. If polymeric flocculants are dosed after the primary coagulants (usually metal salts), the retention time between dosing points should be at least 30 s or more to avoid a higher polymer demand. Polymers should not be mixed with excessive intensity to limit microfloc destruction and polymer breakage.

Table 28. Rapid mixer and destabilization units

Basic type	Examples, characteristics
Pipes and channels (plug flow)	without internals, with orifices or in-line mixers
	with pipe restrictions and expansions
	with static mixers
	with high-speed injection
Stirred vessels (back mix)	detention time 10 – 120 s
	shear gradient >1000 s^{-1}
	special design to control back mixing and retention time distribution
	mixing energy can be varied
Turbulent zones	before, at, or after pumps
	hydraulic jumps
	venturi tubes
	weirs with overfall

7.5.3. *Reactors for Floc Formation*

A variety of reactors for aggregation exist in water treatment, depending on historical developments, national regulations and experience, and the technical solutions of plant designers. The basic principles are as follows [199]:

1. Stirred tanks (back-mix reactors)
2. Pipes (plug-flow reactors)
3. Static systems (channels with baffles, plug flow, and some back mixing)
4. Sludge contact systems (sludge recycling or retention in flocculators)
5. One or more of these reactors in series

Stirred tanks are used widely and allow adjustment of the mean shear gradient independently of plant flow. The tanks are designed in rectangular form or as round vessels with baffles. Retention times at nominal flow are generally between 10 and 30 min (total); longer times are needed if only one reactor is used. The required retention time is lower (2 – 10 min) in direct filtration plants. The mean shear gradients adjusted by the stirrer speed are in the range of 10 – 100 s^{-1}. In two or more sequential reactors, the $\bar{G}$ value declines typically from ca. 100 s^{-1} to 10 – 30 s^{-1}. If polymeric flocculants are added, the final shear gradient can be higher, ca. 25 – 50 s^{-1}, because of more rigid flocs. Inlets and outlets must be arranged properly to avoid short-circuits. The stirrers used are mostly of the blade or grid type, and stirrer speed should be variable to optimize floc formation under various conditions of raw water and flow.

Flocculation in pipes [212] offers the possibility of using existing piping and saving space, but the efficiency of flocculation depends on flow rate. At a low flow rate, the kinetics of aggregation are slower and solids may settle, while at high flow the flocs tend to be destroyed. Polymeric flocculants are required, and the pipe design should avoid narrow bends and other high turbulence. Pipe lengths required are 20 – 50 m, corresponding to ca. 0.5 – 2-min detention time.

Static Channel Flocculators. Similar reactor characteristics are found in static channel flocculators equipped with baffles to induce back mixing and establish shear gradients by the head loss. These systems are in wide use in developing countries because of their simple design and construction [213], [214]. The channels are usually arranged one by one in flat, rectangular tanks, requiring more surface area compared to other flocculators. Easy cleaning of the channel bottom is essential. The width of the channels may be constant or increasing in the direction of flow to lower the $\bar{G}$ value. Total detention times are generally about 20 – 40 min. Design parameters and examples are given in [214].

Sludge contact systems use the positive kinetic effect of high volume concentrations of suspended solids (see Φ in Eq. 7.1) by retaining the flocs in reactors by various methods. Besides more rapid floc formation, denser and larger flocs are formed, which settle faster. Sludge contact is established only in combination with sedimentation. Basically, two techniques are applied:

1. Recycling of settled sludge to the flocculators (stirred tanks)
2. Sludge blanket systems in upflow tanks with vertical or inclined walls

A *sludge blanket* is established when the flocs have a settling rate equal to the upflow rate. Raw water enters the bottom after rapid mixing via a pipe system, and destabilized particles are captured by the large flocs of the blanket. Excess sludge is removed via sludge hoppers. Sludge blanket reactors need an extended period of startup and are somewhat sensitive to rapid flow increase. They have advantages for low-turbidity and algae-laden raw water. Their upflow rate is in the range of 2 – 4 m/h. A commercial system uses a pulsating raw water flow [215]. Equipped with parallel plates, a more recent version can operate at an 8 – 12-m/h upflow rate. New flocculators use a stirred upflow tank to form pellet-like flocs of high density [216].

7.5.4. *Integrated Flocculation Systems*

Because of the great variety of reactors and floc separators employed only a few examples are shown here. *Classic process schemes* including settling are depicted in Figures 28 and 29. The total detention time is 2 – 4 h, and the systems are most suitable for treatment of river water with a variable content of mineral solids. The settler itself can be designed as a rectangular (Fig. 28) or a circular (Fig. 29) tank with a circular flocculator in the inner section [200]. Settled sludge may be recycled to the flocculators by a pump.

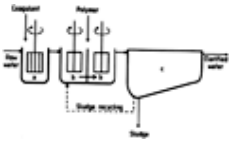

Figure 28. Rectangular settlers

a) Rapid mixing; b) Flocculator; c) Sedimentation

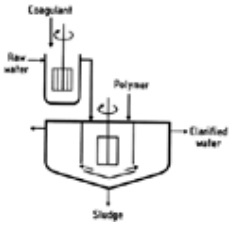

Figure 29. Circular system

The introduction of *parallel plates* or tubes into sedimentation units can reduce retention times. Such a system is shown in Figure 30, including coagulant mixing (a), sludge recycle, and polymer mixing (b); two stirred flocculators (c); parallel plate settlers (d); and a sludge thickener (e) below the plates. The total detention time is only 30 min at full capacity [200], [217].

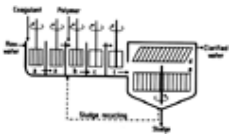

Figure 30. Process scheme of treatment plant with multiple stirring, parallel plate settler, and thickener

a) Coagulant mixing; b) Polymer mixing; c) Stirred flocculator; d) Parallel plate settler; e) Sludge thickener

Direct filtration of flocculated water can be employed for low-turbidity raw water from lakes and dams. The flocs can be quite small (microflocs) to be filtered off in dual- or triple-media filters, at rates of 10 – 15 m/h. An example is shown in Figure 31 [218].

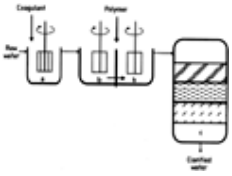

Figure 31. Direct filtration plant with three-layer filter

a) Rapid mixing; b) Flocculator; c) Three-layer filter

Flotation is a feasible floc removal process for low-density flocs, formed in colored, humic acid water or in algae-laden raw water. The version used mostly is dissolved air flotation (DAF, ⟶ Wastewater), recycling 5 – 10 % of the clarified water, which is saturated with air under a pressure of 5 – 10 bar. The flocculated water and recycled pressurized water are mixed; gas bubbles are formed, attach to the flocs, and cause them to float [219]. This process typically requires less space than classical flocculation processes, no polymers, somewhat shorter flocculation times, and no additives for flotation.

Since the 1970s, the introduction of preozonation before coagulant addition has proved successful to decrease turbidity and increase the removal of algae, to lower coagulant demand, or to extend filter run times in direct filtration [220]. Ozone is applied typically in dosages of 0.5 – 2 mg/L, depending on raw water conditions. Other oxidants such as chlorine or chlorine dioxide are less efficient or are avoided because of the formation of harmful byproducts.

7.5.5. *Operational Aspects*

The selection of suitable coagulants and flocculants and the continuous task of adjusting their dosages to optimum values are based mainly on experience, rule of thumb, and extensive laboratory and pilot-plant testing. The great number of parameters that influence flocculation and the lack of quantitative models require adapted testing procedures. The *jar-test* is the most widely used laboratory test, suitable for conventional plants and chemical selection [198], [199], [221]. Other tests include filtration or batch flotation. Flow-through pilot plants are usually designed according to full-scale systems and allow more experiments, especially regarding hydrodynamic parameters. However, problems occur in scaling-up pilot systems [222] because of different flow patterns. Large-scale plants are therefore often equipped with variable-speed stirrers or changeable reactor geometry to adjust the reactors to flow variation. On-line control and dosage adjustment have been implemented sucessfully in some plants in the 1980s, based on novel techniques and sensor systems [223], but their application require careful evaluation.

[Top of Page]

8. Filtration

Rolf Gimbel

8.1. Introduction

In this chapter, only the elimination of turbidity-causing solids (colloidal or finely suspended particles) by filtration is described. Consequently, this is clearly distinguished from filtration through layers of activated carbon (see Chap. Adsorption Processes in Water Treatment) or ion exchangers (see Chap. Ion Exchange), which is used primarily for the removal of fully dissolved water constituents.

Surface Filtration. The removal of turbidity from water can be carried out by using surface filtration (either *cake filtration* or *cross-flow filtration*; see Fig. 32 A and B). In surface filtration, the dimensions of the particulates are greater than the pore size of the filter medium (or of the filter cake possibly already present on it). Therefore, particles are retained because of sieve effects, while water passes through the filter medium under a pressure gradient. Surface filtration is used in many different areas of industry for solid – liquid separation (⟶ Filtration – Washing of Filter Cakes, ⟶ Filtration – Optimal Cycle Time). The liquids treated range from extremely dilute solutions to highly concentrated suspensions or sludge (dewatering). Therefore, a large number of different processes and equipment are available, which operate according to the basic principle of surface filtration. These also include membrane filtration processes (see Chap. Membrane Separation Processes in Water Treatment).

Figure 32. Schematic representation of various types of turbidity filtration

A) Cake filtration; B) Cross-flow filtration; C) Deep-bed filtration

a) Suspension flux; b) Filter medium or filter layer; c) Filtrate flux; d) Filter cake

Deep-Bed Filtration. In the treatment of wastewater and drinking water and the production of cooling and process water, by far the predominant amount of water is subjected to deep-bed or depth filtration to eliminate turbidity (see Fig. 32 C) (⟶ Filtration – Deliquoring of Filter Cakes). In contrast to surface filtration, particles are retained in the interior (i.e., trapped in the depths) of a porous filter layer, which usually consists of a bed of granular material. To maintain the depth effect of filters of this type, the particulates to be separated must be considerably smaller than the pore size of the filter layer. Otherwise, the top part of this layer changes to cake filtration after a very short filtration time, resulting in an excessive pressure drop in the uppermost filter layer.

Deep-bed filtration is used especially for the purification of liquids having very low solid concentrations or when the filtrate is required in larger quantities and with a high purity (i.e., practically free of turbidity). For this reason, deep-bed filtration is applied mainly in the aforementioned areas of water technology. A volume concentration of solids of <0.05 % can serve as a rough upper limit for the economically meaningful use of depth filters.

To achieve separation, particles must be transported to the surface of the filter material by various transport mechanisms in the interior of the depth filter. At the surface, contact between filter material and particles occurs. In a second step, sufficiently stable adhesion of particles to the filter material must take place. The entire separation process is very complex and can vary greatly depending on application. This variability is due to the complicated flow conditions in the filter layer and to the large range of properties of normally used filter materials. Moreover, the turbidity-causing solids to be separated are diverse and can range from solid mineral particles to particulate biomasses, particle agglomerates, and hydroxide flocs (see Section Inorganic Chemicals) of very low shear strength. Finally, dissolved inorganic and organic water constituents (including polymeric filter aids) greatly influence particle separation. In addition, biological processes can occur.

The design and mode of operation of depth filters are as diverse as the factors influencing deep-bed filtration. Filters can be divided into different groups according to the typical filtration rate in each case (water volume per unit time and total filter cross-sectional area). These groups are discussed in detail below. Today, *rapid filtration* plays a dominant role in water filtration. With rapid filtration as an example, the separation mechanisms of depth filters, which can also be influenced by appropriate pretreatment processes (flocculation, see Chap. Flocculation), and the possibilities of describing total filter behavior (as a basis for optimal filter design) are discussed.

8.2. Slow Sand Filters

Slow sand filters are operated at relatively low filtration rates of about 0.1 m/h. This type of filtration played a dominant role in the treatment of surface waters for the production of drinking water in the beginning of the 20th century and is of great importance until now. Here, purification processes are very similar to those in natural ground filtration. Apart from mechanical sieve effects and physicochemical deep-bed filtration mechanisms (transport and adhesion processes), microbiological processes are especially important [224-226].

Slow sand filters are usually employed in the form of open basins, which can have a filtration area up to 10 000 m^2. The filters are provided with a suitable drainage system or infiltrate the subsoil directly to artificially recharge groundwater (see Fig. 33). The active filter layer consists of a bed of sand with a depth of 0.5 – 1.5 m and a grain size that can vary from 0.2 to 1 mm. The head of water can reach ca. 1.5 m. A new filter develops its full efficiency only after a so-called “Schmutzdecke” (dirt blanket) has been formed in the uppermost filter layer. This is formed from the retained water constituents by a very complicated combination of diverse physical separation mechanisms and biochemical processes. The “Schmutzdecke” can reach a height of a few centimeters. The purification processes involved in slow sand filtration proceed mainly in this layer.

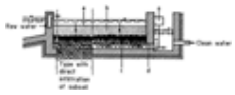

Figure 33. Schematic representation of slow sand filtration

a) Biological filter layer (Schmutzdecke); b) Fine sand, height 0.8 – 1.5 m; c) Supporting layers, height 0.2 – 0.4 m; d)

Drainage system; e) Outflow controller

Excellent water purification can be achieved with slow sand filters. Not only the finest inorganic particles but also particulate organic substances, especially protozoa, bacteria, and viruses, are retained. In addition, dissolved degradable organic substances can be mineralized and nitrogen compounds oxidized by aerobic biological processes. Moreover, smaller amounts of dissolved interfering substances (e.g., heavy metals) can be retained in the “Schmutzdecke”.

A ripening period of several days is required to form the Schmutzdecke. The length of a filter run (the time between two cleaning steps) is usually some months. At the end of a filter run, the pressure drop in the topmost filter layer is increased to such an extent that the required water throughput can no longer be maintained. In general, the necessary filter regeneration is then carried out by peeling off the uppermost 3 – 5 cm of the filter layer.

The disadvantages of the slow sand filter are a large area requirement and laborious filter regeneration. For this reason, new slow sand filtration plants are rarely built in industrialized countries. In many existing plants, attempts are being made to lengthen the filter run by using appropriate pretreatment processes for turbidity removal (flocculation, sedimentation, rapid filtration).

8.3. Polishing Filters

Polishing filters, which include *precoat, disk*, and *cartridge filters*, are characterized by the relatively small radii of the filter pores (micrometer range) and the shallow depth of the active filter layer (millimeter to centimeter range). In filters of this type, apart from physicochemical transport and adhesion processes, sieve effects on the surface of the filter layer can also be significant, depending on the ratio of the size of suspended particles to pore size. Thus, these filters cannot be regarded as pure depth filters in many cases. In contrast to slow sand filtration, however, biological processes are negligible in polishing filtration. With pressure differences up to several bar, polishing filters can be operated at average filtration rates of about 1 m/h. Depending on the specific application, larger deviations from this average can occur [227].

Precoat Filters. In precoat filters, turbidity removal occurs in a layer of a fine granular or fibrous filter aid having a characteristic size of a few micrometers. This layer has previously been filtered onto a support. The aids employed are mainly aluminum silicate (e.g., perlite) and diatomaceous earths. Other filter aids are also used in some cases. For example, powdered activated carbon or a finely dispersed ion-exchange resin gives not only highly efficient separation of colloidal sustances, but also retention of dissolved organic material or ions [228].

In precoat filtration, depending on the conditions, sieve effects on the surface of the precoat layer can dominate after a relatively short filter run, resulting in clogging of the surface. To prevent this and guarantee sufficient permeability of the resulting filter cake, filter aids are often added continuously to the suspension to be treated. In addition, the filter aids of the precoat layer can be conditioned with suitable polymers in such a way that they exhibit especially favorable properties for turbidity removal. The same polymers used in flocculation and rapid filtration are applied.

Precoat filters are usually available as closed apparatus in which the supports have a space-saving flat or tubular arrangement as far as possible. After a filter run that lasts from hours to days depending on the application, the precoat layer is generally discharged by means of a pressure surge. After renewed precoating, which requires a few minutes, the filter is ready for use again.

Disk and Cartridge Filters. In contrast to precoat filtration, the active filter layer in leaf and cartridge filters need not be built up by a filter aid but is available in a preprepared form. Apart from this characteristic difference, the modes of action and operation of different types of polishing filters are similar. The preprepared filter layers can be made of fabric, fleece, felt, or wound string especially in the case of cartridge filters. Finely porous sintered materials such as fibrous or powdery metals, ceramic materials, and plastics are also used. In many cases, disk and cartridge filters require not only mechanical but also chemical cleaning. Alternatively, filter materials are used only once [229], [230].

Filter Performance. In general, excellent filtrate quality can be achieved with polishing filters. They are used especially in the treatment of boiler feed water and ultrapure water, such as that required in the electronic, chemical, and pharmaceutical industry. In addition, these filters are often employed in smaller drinking water treatment plants, in emergency water supply facilities, and in the treatment of water for swimming pools. Polishing filters are used frequently in the food industry for clarification of beverages. In special cases, filters of this type are also used in wastewater treatment and in the purification of various solutions (e.g., in surface technology).

8.4. Rapid Filters

The term rapid filter is derived from the relatively high filtration rate through the filter, which is on average ca. 10 m/h. Filters of this type consist of a relatively coarse-grained filter material (millimeter range) having a filter bed height of 1 – 2 m. The deposition of particles within the filter layer — which generally acts as a real depth filter — is based primarily on physicochemical transport and adhesion mechanisms, which can also be influenced by biological processes.

In industrialized countries, rapid filtration has the greatest practical importance of all filtration processes used in the treatment of drinking water and industrial water or in advanced wastewater treatment. This importance is attributed not only to the low space requirement as a result of high filter rate, but also to the simple design and mode of operation of rapid filters and to the largely automated filter regeneration in which filter material is backwashed within the filter tank.

8.4.1. *Design*

The active layer of rapid filters consists of a bed of granular materials having an average diameter between 0.5 and 4 mm and a bed height of 0.5 – 2.5 m. If a fine-grained filter material is used, low bed heights are sufficient for good removal

efficiency. However, the disadvantage is that the capacity for the particulate load on the filter is exhausted quickly, which results in a short filter running time. Quartz sand, anthracite, filter coke, activated carbon, or pumice serve as filter materials [228].

The *single-medium* design of a classic rapid filter is represented schematically in Figure 34. It consists of a distributing device for raw water (a), an active filter bed (b), which is to be regarded as homogeneous; and a filter bottom (d), which supports the filter material and allows the evenly distributed passage of filtered water into the clean-water chamber (e). The filter bottom made of steel or reinforced concrete either is equipped with filter nozzles provided with fine slits or is made of special porous components. To guarantee perfect functioning of the elements of the filter bottom, especially if fine filter materials are used, one or several supporting layers (c) are included between the active filter bed and the filter bottom. Each supporting layer has a grain size ca. two to four times coarser than the material above it.

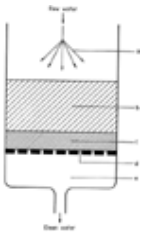

Figure 34. Basic design of a rapid filter

a) Raw water distribution; b) Active filter layer (filter bed); c) Supporting layer; d) Filter bottom; e) Clean water chamber

As a rule, the direction of flow in rapid filters is from the top downward. The filter material should have a particle-size distribution as uniform as possible. Otherwise, fine material becomes concentrated in the top filter layer during filter backwashing. This can result in a change to surface filtration, which leads to an excessive pressure drop. To prevent this effect and distribute particle deposition over the greatest possible bed depth, *dual-media* filters are used frequently (see Fig. 35). In this type, the upper filter layer consists of a relatively coarse material of low density (filter coke in Fig. 35), while the lower layer has a fine grain size and high density. If the grain sizes and densities of the filter materials in the two layers are well suited to each other, this structure remains even after filter backwashing. Because of its large pore size, the upper layer has a high load capacity for particulates, whereas particle retention is not necessarily complete. In comparison, the behavior of the lower layer is the opposite. As a result, especially in the case of high-turbidity loads, longer filter running times can be attained with dual-media filters than with single-medium filters. *Three-media filters* have also been used successfully. Typical examples of the design of multimedia filters are given in Table 29.

Table 29. Exemplary design of multimedia filters [228]

Material	Bed depth, m	Bed weight, kg/m^3	Grain size, mm
Dual-media filters			
Filter coke, pumice, etc.	0.5 – 1.2	500 – 700	1.7 – 2.5
Sand	0.6 – 1.5	1000	0.8 – 1.2
Three-media filters			
Activated carbon	0.3 – 0.6	250 – 350	3.0 – 5.0
Filter coke, etc.	0.6 – 1.2	500 – 750	1.5 – 2.5
Sand	0.5 – 0.8	1000	0.6 – 0.8

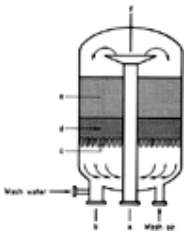

Figure 35. Exemplary design of a closed dual-media filter

a) Inlet; b) Outlet; c) Filter bottom; d) Sand; e) Filter coke; f) Water distribution

Figure 35 shows the design of *closed rapid filters* (*pressure filters*). Filters of this type are usually made in the form of cylindrical steel tanks. Filter surfaces up to 50 m^2 are achieved in vertical tanks. Horizontal tanks provide much larger filter surfaces but usually have relatively low bed heights. The usual pressure drops vary between 0.2 and 2 bar at filtration rates of about 5 – 30 m/h.

Open rapid filters (gravity filters) are often used in large plants. The pressure drop occurring in the filter bed must be compensated by a head of water above the bed. The height varies between 0.3 and 3 m of water column (corresponding to a pressure drop of 0.03 – 0.3 bar). Usual filter flow rates are in the range of 3 – 15 m/h. Open filters are usually made in the form of rectangular reinforced concrete tanks. Filter surfaces up to 150 m^2 have been achieved. Apart from the classic type of rapid filter shown here, several special designs exist with upward or horizontal flow through the filter layer. Continuously operated rapid filters are also being used increasingly. The loaded filter material is withdrawn continuously from the bottom and returned at the top after a washing stage [231], [232].

8.4.2. *Areas of Application and Modes of Operation*

In water and wastewater treatment, the different applications of rapid filters can be divided into the following main groups [228].

1. *Separation of Single Particles.*
   Applications of this type can be encountered, e.g., in the removal of turbidity from surface water that is to be used as drinking water and industrial water. The particles are inorganic substances (e.g., clay minerals) and organic particles such as plant cells, algae, or bacteria. Further applications are the separation of $CaCO_3$ particles after a water-softening step and the treatment of boiler feed water and swimming pool water. In all these cases, biological processes are usually of secondary importance.
2. *Separation of Particle Agglomerates and Hydroxide Flocs.*
   If turbidity-causing solids are difficult to filter out as single particles, agglomeration is usually carried out before rapid filtration, or the particles are embedded in iron or aluminum hydroxide flocs by means of a flocculation process. Dissolved substances can also be bound in the flocs (see Section Introduction). Applications of this type are encountered frequently in the production of drinking water and industrial water from surface waters and in advanced wastewater treatment. In the latter, the hydroxide flocs to be separated contain not only turbidity but also colloidal and dissolved phosphorus compounds. Biological processes in the filter can greatly influence turbidity removal, especially in advanced wastewater treatment.
3. *Removal of Iron, Manganese, and Ammonium.*
   In the treatment of groundwater to produce drinking water and industrial water, the rapid filter usually has the task of eliminating divalent iron and manganese ions. These ions must be converted to a separable form by chemical or biochemical oxidation before or during actual filtration. In this widely encountered application of rapid filters, chemical – catalytic as well as biological processes are important. Ammonium can also be eliminated by biochemical oxidation.

In many cases, single-medium filters are sufficient for relatively low levels of turbidity in raw water. For higher-turbidity loads, multimedia filters can be of advantage. The importance of multimedia filters has increased since the 1970s because the *direct filtration process* is often used for economic reasons. In direct filtration, water (without preliminary separation of flocs by sedimentation or flotation) is passed directly through rapid filters after a flocculation stage.

Especially in direct filtration, pretreatment of water is very essential. Apart from the type and amount of flocculant used, the conditions of addition (residence time up to the filter, energy input in the flocculation stage, etc.) are also important. Moreover, the flocculation or filter aids added, which are mostly synthetic, water-soluble polymers, can clearly improve the properties of flocs with regard to filtration. Oxidative pretreatment of raw water (e.g., with ozone) can have a similar effect.

In normal depth filtration, the space between the grains of filter material is filled completely with particulates and water. In *dry filtration,* however, air is passed through the filter bed together with water. This special process can be used in water treatment when the aerobic degradation or oxidation processes proceeding in the filter bed are so intense that a continued supply of oxygen via air is required. This applies, for example, to the treatment of water containing high concentrations of ammonium or iron. Dry filtration is usually operated as a preliminary stage in normal rapid filtration because turbidity retention in the dry filter tends to be unstable [228].

Operation. In the operation of rapid filters, it is frequently preferred to keep the filtration rate constant during a run. This requires suitable control because the resistance to flow in the filter bed increases with increased filter running time due to turbid deposits. A constant filtration rate can be achieved by throttling the filter outflow at the beginning of the run. Throttling is then reduced, corresponding to the increased pressure drop due to loading of the filter bed. In open filters, especially; the same result can be achieved by an appropriately increasing head of water. In some cases, operation with variable filtration rates is preferred, the rate decreasing to about half of the starting value during a filter run. Depending on the raw water properties, the average filtrate quality can be somewhat better than that obtained with a constant filtration rate in the same filter unit. This results from the fact that the lower the flow rates, the lower is the probability of turbidity deposits being detached and washed out due to the shear forces in the filter.

The running time of rapid filters is normally between 10 and 150 h. In this time, both filtrate quality (expressed as volume concentration of turbidity c) and pressure drop over the filter bed (Δp_V) change, as shown qualitatively in Figure 36. In the continuous monitoring of filtration plants, a simple turbidimetry is generally used in practice to determine the turbidity level. As shown in Figure 36 three characteristic phases of a filter run can be distinguished. The level of turbidity in the filter outflow decreases up to time t_1. This is the period of ripening, which can be very pronounced, especially when chemical – catalytic or biological processes are important for filtering action. The level of turbidity in the filtrate is approximately constant between t_1 and t_2. The breakthrough phase of the filter starts at t_2. The filter run can continue until the maximal permissible concentration of turbidity-causing particles is reached (e.g., at t_3); then the filter must be backwashed. Furthermore, Figure 36 shows that the pressure drop over the filter bed increases with filter running time and reaches its maximum permissible value at t_4. One of the important optimization tasks in rapid filtration consists of making t_3 and t_4, the two times that limit the filter run, coincide as much as possible. For safety reasons (i.e., to make sure that the turbidity in the filtrate never becomes too high), the t_3 value chosen in practice is often ca. 10 – 20 % greater than t_4, and attainment of the maximum permissible pressure drop is used as the primary criterion for terminating a filter run.

Figure 36. Change in pressure drop and in turbidity with time in the filtrate of a rapid filter at a constant filtration rate

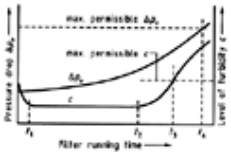

Regeneration. At the end of its running time, a classic rapid filter must be regenerated. This is usually carried out in a backwashing process in which the upward flow through the filter layer is so strong that the turbid deposits on the filter grains are detached and discharged. In *single-medium filters*, backwashing is often carried out with water only. The flow rate must be so high that the filter bed is converted from a fixed bed to a fluidized bed. To achieve more intensive purification of the filter grains, air can be used in addition as backwash medium. In the first phase, the filter layer with a slight water head is whirled up only with air, causing a reduction of possibly hardened filter grains. Subsequently, backwashing is often conducted with air and water, followed by water alone.

Multimedia layers are treated analogously. In most cases, however, simultaneous air – water backwashing is omitted because the danger of losing filter material with the wash water is too great. In multimedia filtration, the last phase involves a reclassifying process, which causes the different filter media to separate so that the original layer setup of the filter is reproduced. In this process relatively high wash water rates are applied; expansion of the filter bed can reach ca. 40 % in the process.

A filter backwash usually requires 10 – 20 min. Depending on grain size and density, the wash air and wash water rates vary from ca. 30 to 100 m/h with a filter bed expansion of 10 – 50 %. In a well-designed filter wash program, the water consumption based on the filter area should be not greater than 2 – 5 m^3/m^2 per wash. If the removal of turbidity in the filter depends greatly on biological or chemical – catalytic processes, excessive filter backwashing can be a disadvantage because a relatively long period of filter ripening is required.

8.4.3. *Filter Performance*

To design and operate a rapid filter optimally, both the volume concentration *c* of a certain type of particle and the filter load σ, which are functions of filter layer depth *z* and filter running time *t*, should be predeterminable under given feed conditions. The load σ determines the pressure distribution *p* in the filter layer and, finally, the total pressure drop. To describe *c*, σ, and *p* as a function of *z* and *t*, a mass balance, a kinetic equation for the mass transfer from the flowing phase to the stationary filter material, and a relation to describe the pressure distribution in the filter are required for a differential filter element [228], [233], [234].

The simplified *mass balance* is as follows:

$$v_f \frac{\delta c}{\delta z} + \frac{\delta \sigma}{\delta t} = 0$$

where v_f is the filtration rate; *c* the volume concentration of particles; σ the filter load (volume of particles removed, based on filter layer volume); *z* a space coordinate (filter layer depth); and *t* the filter running time.

For the *kinetic equation*, the relation proposed by IWASAKI [235] is generally used. It corresponds formally to first-order reaction kinetics:

$$\frac{\delta c}{\delta z} = -\lambda \cdot c$$

The filtration coefficient λ is a measure of filter efficiency. It corresponds to the probability of a certain particle being retained in the filter layer. The filtration coefficient depends on the properties of the filter material (e.g., grain size) and of the particles (e.g., particle size, density, and surface charge) and on filter operating conditions (e.g., filtration rate). However, λ also changes with increasing filter running time because of the increase in filter load with time. For this reason, λ is usually formulated as follows [236]:

$$\lambda = \lambda_0 \cdot f_\lambda$$

λ_0 represents the filtration coefficient in the initial phase of filtration (i.e., as long as the influence of the filter load is still negligible); f_λ is a correction function, which essentially describes the effect of the filter load on particle deposition. In the idealized model case, f_λ decreases linearly with filter load, i.e., $f_\lambda = 1 - \sigma / \sigma_s$ (σ_s: saturation or maximum load). Then, the following analytical solution for the distribution of the particle concentration in the filter layer can be obtained from the two partial differential equations for the mass balance and for the kinetic equation

$$\frac{c}{c_0} = \frac{\exp\{\lambda_0 \cdot v_f \cdot c_0 \cdot t/\sigma_s\}}{\exp\{\lambda_0 \cdot z\} + \exp\{\lambda_0 \cdot v_f \cdot c_0 \cdot t/\sigma_S\} - 1}$$

where c_0 is the volume concentration of particles in the filter inflow. In case of a more complicated dependence $f_\lambda = f(\sigma)$, numerical solutions for describing the distribution of the concentration and load must be applied. The same holds true in describing the pressure distribution in the filter layer, which is usually based on the Carman – Kozeny equation [228], [233], [234] (⟶ Filtration – Filtration Models).

For the initial phase of filtration ($t \longrightarrow 0$), the above equation results in an exponential decrease in the volume concentration

of particulates with bed depth, according to the following:

$$\frac{c}{c_0} = e^{-\lambda_0 \cdot z}$$

where λ_0 is determined by various forces that act within the filter layer on a particle to be deposited. These forces must not only bring a particle into contact with the surface of the filter material (transport mechanisms), but also make possible sufficiently stable particle adhesion (adhesion mechanisms) on contact. Particle separation is thus influenced by many parameters, including the filtration rate and the size, shape, and surface charge of the filter material. The properties of water such as density and viscosity, the pH value, and the other dissolved water constituents are also important. In addition, the properties of the turbidity-causing solids, such as density, particle size and shape, surface charge, and shear strength, sometimes play a dominant role. The important classic separation mechanisms in deep-bed filtration of liquids are shown in Figure 37. In the case of *interception* (Fig. 37 A), a particle is assumed to move through the filter layer with its center along a stream line. It can come in contact with the surface of a grain as a result of the constriction of stream lines due to neighboring filter grains. In the case of *sedimentation* (Fig. 37 B), the particle can leave its original stream line and come in contact with the filter material under the effect of gravity. If the turbidity-causing particle has a size <1 µm, a particle – grain contact can also occur due to *particle diffusion* (Fig. 37 C).

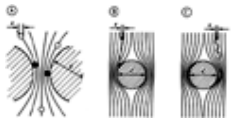

Figure 37. Schematic representation of classic particle deposition mechanisms in deep-bed filtration of liquids

A) Interception; B) Sedimentation; C) Particle diffusion*d* = diameter of filter grain; *x* = diameter of turbidity-causing particle

Despite the extremely diverse influencing factors, the success of particle deposition in rapid filters usually increases with larger particle size, higher particle density, lower filtration rate, and smaller filter grain size. However, a decreasing filter grain size results in an increasing pressure drop. In addition, excessively high electrostatic respulsion forces as a result of surface charges of the same sign on turbidity-causing particles and filter materials must not occur. These repulsion forces can prevent particle deposition. In this respect, appropriate water pretreatment can be of considerable significance.

Under normal conditions of water treatment, a range of ca. 1 – 10 m^{-1} can be given as a rough guide for λ_0. The maximum load σ_s attainable is usually < 2 % [228], [236].

[Top of Page]

9. Membrane Separation Processes in Water Treatment

Robert Rautenbach

(see also, ⟶ Membranes and Membrane Separation Processes)

Since the 1980s, membrane processes have been used increasingly in separation technology. The main emphasis was put on the treatment of aqueous solutions. In many cases, such as the desalination of brackish water, membrane processes have set the standard against which all other alternatives must be compared with regard to operating and investment costs.

In general, membrane processes are characterized by the following properties:

1. They can operate at ambient temperature.
2. They separate on a purely physical basis so that the components are not altered and can, in principle, be reused.
3. They have a modular setup so that they can be adapted to any separation capacity relatively easily and later expansion is possible (in contrast to biological processes).

With regard to the production and treatment of drinking water and wastewater treatment, the microfiltration – ultrafiltration and nanofiltration – reverse osmosis are of special significance (⟶ Membranes and Membrane Separation Processes – Ultrafiltration, Microfiltration, and Reverse Osmosis). This chapter is restricted to these processes.

In these processes, the component that passes preferentially through the membrane is the solvent (i.e., water in the case of aqueous solutions). In all the aforementioned processes, the driving force for this transport results from a transmembrane pressure difference. However, the magnitude of pressure difference and, especially, the type of membrane used differ in these processes. Whereas pore membranes are used in microfiltration and ultrafiltration (UF), nonporous solution-diffusion membranes are employed in nanofiltration (NF) and reverse osmosis (RO).

9.1. Principles

9.1.1. *Flow, Selectivity, Driving Forces*

Microfiltration and Ultrafiltration. In pore membranes, *selectivity* results basically from the pore-size distribution of the membrane. Characterization of the membrane is carried out by permeation experiments with aqueous solutions of macromolecular substances of varying molar masses (Fig. 38).

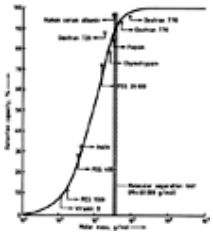

Figure 38. Separation characteristics of ultrafiltration membranes

PEG = Poly(ethylene glycol)

Like the flow through a particle bed, the *flow of solvent* through pore membranes (i.e., the performance of the membrane) is a viscous flow proportional to the transmembrane pressure difference.

$$\dot{m}_w'' = A\varrho_w \Delta p \tag{9.1}$$

where $\dot{m}''$ is the mass flow rate per unit area (flux), ϱ_w is the density of water, and Δp is the pressure difference across the membrane. The constant A is membrane and substance specific (influence of solvent viscosity) and is determined experimentally. In most applications of micro- and ultrafiltration, a *surface layer* forms on the membrane that consists of retained components. This formation can rarely be prevented and is often even desirable. This layer is more important than the real membrane with regard to the separation characteristic and performance (permeate flux). Therefore, in the case of micro- and ultrafiltration, experiments under operating conditions are mandatory — processes cannot be designed based on performance data of the membrane alone.

Since the thickness and structure of the surface layer are influenced by conditions of flow of the feed solution along the membrane, the hydrodynamics in the membrane module is of great importance. In many cases (e.g., the separation of protein from whey), the permeate flux can in practice be influenced only by hydrodynamics and is independent of the transmembrane pressure difference (Fig. 39, see also Section Mass Transport Resistance in Front of the Membrane).

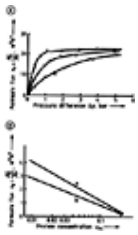

Figure 39. Influences on the performance of ultrafiltration

A) Influence of pressure difference (operating conditions: w_p = 0.8 %, ϑ = 25 °C, n = 500 min^{-1}; membranes: △ = UF-00 Kalle; ○ = UF-0 Kalle; □ = CA-19000 Kalle); B) Influence of flow conditions (operating conditions: ϑ = 25 °C, Δp = 4 bar

a) n = 550 min^{-1}; b) n = 300 min^{-1})

n = impeller speed

The measured values shown in Figure 39 are steady-state values, i.e., values obtained after a certain operating time, corresponding to Figure 40, curve a.

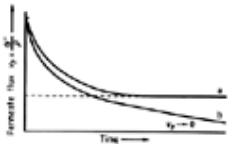

Figure 40. Behavior with time of surface layer-controlled flux

a) Reversible formation of surface layer; b) Irreversible formation of surface layer

When a reversible surface layer is formed, an equilibrium exists between deposition of the retained components on the upper side of the surface layer as a result of permeate flux through the membrane and removal of surface layer components due to shear stresses caused by tangential flow. The use of ultrafiltration instead of sedimentation for the separation of biomass after a biological treatment stage represents an example of this behavior. In this case, a constant permeate flux of ca. 100 L m^{-2} h^{-1} has been observed over several weeks with tubular modules at tube flow rates of ca. 4 m/s [237]. Formation of a reversible surface layer is often encountered in ultrafiltration (and microfiltration), but the behavior with time shown in curve b (formation of an irreversible surface layer) is encountered just as often. When an irreversible surface layer is formed, the operating time after which the unit should be cleaned is then determined by the economics of the process.

Reverse Osmosis. In nonporous membranes for reverse osmosis and nanofiltration, transport generally is determined by the mass transport resistance of the membrane and not by a surface layer. However, modeling mass transport in the membrane is more difficult than in ultrafiltration, especially in the case of organic – aqueous solvent mixtures and, in nanofiltration, in the case of a multicomponent salt solution with monovalent and multivalent anions.

If the membrane is regarded as a continuum in which the permeating components are first sorbed, then transported by diffusion, and finally desorbed, equations can be derived for any permeating component "i ", e.g., i = W (water), i = S (salt or dissolved organic substance) by the integration of the generalized Fick law over the membrane thickness.

$$\dot{n}_i'' = -c_i \frac{D_{io}}{RT} \frac{d\mu_i}{dz} \tag{9.2}$$

where $\dot{n}$ is the molar flux, D_{io} the thermodynamic diffusion coefficient, μ the chemical potential, and z the thickness of the membrane. Depending on the degree of simplifying assumptions, these equations vary with regard to complexity and

descriptive accuracy. Table 30 shows three such equation systems for reverse osmosis, as well as cases in which they describe flux and selectivity with sufficient accuracy.

Table 30. Models for describing mass transport in reverse osmosis membranes

Model	Transport relation	Application
1	$\dot{m}_p'' \approx \dot{m}_w'' = A\varrho_w(\Delta p - \Delta\pi_w)$	production of ultrapure water
	$R \equiv \frac{w_{S_F} - w_{S_P}}{w_{S_F}}$	from drinking water
2	$\dot{m}_w'' = A_{\varrho w}(\Delta p - \Delta\pi_w)$	production of drinking water
	$\dot{m}_S'' = B\left(w_{S_F} - w_{S_P}\right)$	from brackish water
3	$\dot{m}_w'' = A^* \frac{1}{1+kw_{S_F}}\left(\Delta p - \Delta\pi_w\right)$	concentration of aqueous – organic solutions
	$\dot{m}_w'' = B^* \frac{kw_{S_F}}{1+kw_{S_F}}\left(\Delta p - \Delta\pi_S\right)$	wastewater treatment

The following assumptions are common to all three relations:

1. Equilibrium of the chemical potentials μ_i at the phase boundaries on both sides of the membrane: i.e., $\mu_i^F = \mu_i^{FM}$, $\mu_i^P = \mu_i^{PM}$ (where the index F denotes feed, M denotes membrane, P permeate, FM feedside of the membrane, and PM permeate side of the membrane).
2. No coupling of the permeate flows of different components: i.e., only the gradient of the chemical potential of the component μ_i is regarded as the driving force for transport of component *i* of a mixture through the membrane

The question which equation can or must be used to describe local mass transport (i.e., the mass transport at a certain point of the membrane) for a certain application can only be solved experimentally. For instance, in the reverse osmosis of dilute sodium chloride solutions, a constant rejection *R* can be assumed — i.e., a constant ratio of the (local) salt concentration in the permeate to the local feedside concentration. In contrast, this ratio (or the rejection) is highly concentration dependent in the case of an aqueous butanol solution.

The problem can be explained by the example shown in Figure 41. In the case of a phenol – water mixture, significant differences between experiments and calculations are observed even at very low phenol concentrations when the calculation is based on model 2. The three-parameter model 3 must be used here.

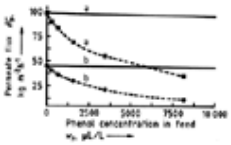

Figure 41. Decrease in flux in reverse osmosis caused by traces of organic compounds (phenol) [238] (operating conditions: phenol – water mixture, Δp = 30 bar; Θ = 25 °C)

a) Measurements performed with membrane SU 700; b) Measurements performed with membrane FT 30

——— = calculated according to model 2; – – – = measured

Nanofiltration. In principle, nanofiltration membranes are nonporous solution – diffusion membranes like reverse osmosis membranes [239]. However, they have a low rejection for low molar mass substances. Nanofiltration is characterized by two extremely interesting properties:

1. The ability to fractionate organic components of varying molar mass and/or structure in aqueous solutions (e.g., mixtures of mono- and polyhydric alcohols)
2. The possibility of utilizing the Donnan effect (⟶ Membranes and Membrane Separation Processes; ⟶ Ion Exchangers – Principles) whereby through the addition of one ionic species (e.g., SO_4^{2-} ions), the rejection capacity for another ionic species (e.g., NO_3^-) is drastically decreased.

This selectivity of nanofiltration for organic components in the molar mass range 100 g/mol < *M* < 300 g/mol is shown in Figure 42. For example, the rejection capacity of the membrane “Desal 5” is only ca. 30 % for ethylene glycol, but more than 80 % for glucose. These results were obtained from experiments with binary solutions of the same osmolarity (i.e., the same driving force) [238].

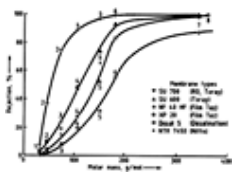

Figure 42. Separation characteristics of nanofiltration in organic aqueous systems (operating conditions: Δp = 10 bar; Θ = 25 °C; osmotic concentration x_F = 200 mOsmol per kilogram H_2O)

1 = Methanol; 2 = Ethanol; 3 = *n*-Butanol; 4 = Ethylene glycol; 5 = Triethylene glycol; 6 = Glucose; 7 = Sucrose; 8 = Lactose

The Donnan effect is illustrated in Figure 43. The addition of SO_4^{2-} ions, in the form of Na_2SO_4, reduces the rejection capacity for Cl^- ions significantly, even to negative values, while the rejection capacity for Na^+ ions increases slightly.

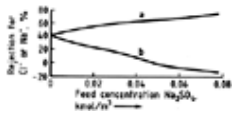

Figure 43. Influence of the addition of SO_4^{2-} in the form of Na_2SO_4 on Cl^- rejection (operating conditions: membrane XP 45; *c* NaCl = 0.05 mol/L) [239]

a) Na^+ rejection; b) Cl^- rejection

9.1.2. *Mass Transport Resistance in Front of the Membrane*

In reverse osmosis and nanofiltration, where mass transport is membrane controlled in normal operation, the separation performance is usually overestimated if only the resistance of the membrane is considered. A concentration boundary layer on the side of the raw mixture (in front of the membrane) can, in unfavorable cases, clearly reduce the performance of the membrane and the quality of separation. The steady-state local conditions are shown qualitatively in Figure 44. For the purpose of this discussion, a binary solution is used.

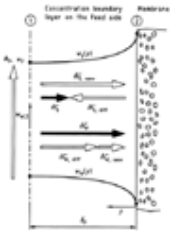

Figure 44. Concentration polarization on the feed side of a reverse osmosis or nanofiltration membrane

$\dot{m}''_{S,\,diff}$ = mass flux of salt per unit area by diffusion; $\dot{m}''_{S,\,conv}$ = mass flux of salt per unit area by convection; $\dot{m}''_{W,\,diff}$ = mass flux of water per unit area by diffusion; $\dot{m}''_{W,\,conv}$ = mass flux of water per unit area by convection; δ_F = thickness of the concentration layer

The concentration profiles indicated schematically in Figure 44 are formed in front of the membrane due to the selectivity of the membrane. The concentration *w* of the preferentially permeating component W (water) decreases toward the membrane, and the concentration of the preferentially rejected component S, corresponding to $w_S = 1 - w_W$, increases. This phenomenon is called *concentration polarization* (⟶ Membranes and Membrane Separation Processes – Concentration Polarization in Filtration Processes). On condition that concentration gradients in the direction of flow are negligibly small compared with the orthogonal gradients, the concentration profile is described by

$$\frac{w_{S2} - w_{SP}}{w_{S1} - w_{SP}} = e^{\frac{\dot{m}''_P}{\rho_P \cdot k}} \tag{9.3}$$

with $k \equiv D/\delta$ as mass-transfer coefficient for negligible wall flux ($m''_P \longrightarrow 0$) and with

$$w_{SP} = \frac{\dot{m}''_S}{\dot{m}''_S + \dot{m}''_W} \approx \frac{\dot{m}''_S}{\dot{m}''_W} \tag{9.4}$$

the subscripts S1 and S2 denote the salt at point 1 and point 2, w_{SP} is the salt concentration of the locally produced permeate. The mass-transfer coefficient can be calculated with sufficient accuracy by means of equations valid for heat transfer, using the Sherwood number instead of the Nusselt number and the Schmidt number instead of the Prandtl number. The correlations valid for tubes and rectangular channels are listed below.

Characteristic numbers

Reynolds number = $Re = d_h\, v/\nu$

Schmidt number = $Sc = \nu/D_{ij}$

Sherwood number = $Sh = k\, d_h/D_{ij}$

where

d_h = hydraulic diameter

d_h = d (tube with diameter d)

d_h = $2\,h$ (channel with height h and width b, $b \gg h$)

v = flow rate

ν = kinematic viscosity

Laminar flow fully developed hydrodynamic but developing concentration boundary layer (Sieder and Tate [240])

$$Sh = 1.62(Re\,Sc\,d_h/L)^{1/3}\; 30 < Re\,Sc\,d_h/L < 10^4$$

Turbulent flow (Linton and Sherwood [241])

$$Sh = 0.04\,Re^{3/4}Sc^{1/3}\; 10^4 < Re$$

The degree of concentration polarization is definitely determined by the hydrodynamics in front of the membrane. Good

hydrodynamics is of great importance for the development of modules (technical membrane configuration).

Estimation of the effect of concentration polarization in a technical membrane module is shown in Figure 45. The ratio of real flux to maximal possible flux (pure membrane resistance) is plotted against the ratio of the mass-transfer coefficient k to the pure water flux $A \cdot \Delta p$ (see Fig. 45). Clearly, the more efficient the membrane, the more important are proper hydrodynamics. (In order to control concentration polarization, k must increase with increasing values of $A \cdot \Delta p$).

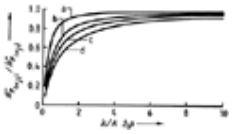

Figure 45. Influence of mass-transfer coefficient on the local performance of a membrane module [operating conditions: system water – NaCl – FT 30 SW; $A = 4\times10^{-7}$ m s^{-1} bar^{-1}; $B = 9.33\times10^{-5}$ kg m^{-2} s^{-1}; b = 8 bar per wt % NaCl; ρ_F = 1000 kg/m^3; w_S^* = 0.035 %; p_F = 60 bar; p_p = 1 bar; R = 1; $\dot{m}''_p = A\,\rho_i\,(\Delta p - b\,w_S)$; $\frac{w_{S,2}}{W_{S,1}} = \exp\left(\frac{\dot{m}''_P}{\rho_i k}\right)$

$m''_p\,(w_{S2})/m''_p\,(w_{S1})$ = permeate flux with concentration polarization/permeate flux without concentration polarization

a) $bw_{S1}/\Delta p$ = 0.1; $bw_{S1}/\Delta p$ = 0.3; c) $bw_{S1}/\Delta p$ = 0.5; d) $bw_{S1}/\delta p$ = 0.8

A special case of Equation (09.3) describes ultrafiltration in the surface layer-controlled region. In this region the stationary final value of the transmembrane flow (see Fig. 40, curve a) becomes fully independent of the transmembrane pressure difference (Fig. 39 A).

If the concentration in the boundary layer reaches saturation concentration, the system compensates for any further increase in driving force by an increase in surface layer. For a binary system consisting of water and dissolved macromolecules (e.g., proteins), which are quantitatively retained by the UF membrane, Equation (9.3) becomes

$$\frac{w_{P,max}}{w_{P,1}} = e^{\frac{\dot{m}''_P}{\rho_P k}} = \text{const.} \qquad (9.5)$$

In this case, the permeate flux is determined only by the saturation concentration $w_{P,max}$, the concentration of the feed, and the hydrodynamics.

9.2. Membranes

(⟶ Membranes and Membrane Separation Processes – Membrane Preparation and Membrane Module Constructions)

For the projecting and process development of any membrane separation process, a large range of selective and stable membrane materials is available. These materials have been optimized for many different applications. Because of the great variety of materials, the nature and efficiency of membranes must be described separately for each process.

Nevertheless, to better understand the processes occurring in a membrane, a short general introduction to the materials and structures of commercially available membranes is given. This discussion is limited to synthetic solid membranes, which are made of both organic and inorganic materials.

9.2.1. *Organic Membranes*

In organic membranes [242-244], the choice of membrane-forming polymer for a concrete separation problem is not arbitrary but is oriented to the following properties of the polymer:

1. Molar mass
2. Chemical structure and spatial arrangement of the macromolecules
3. Interactions between different macromolecules

The structural properties determine both the macroscopic properties of the membrane, such as thermal, chemical, and mechanical stability, and the microscopic “inner” properties, such as the permeability of the polymer to a certain component.

The demand for the highest possible permeate flow requires that the actual selective layer of a membrane be as thin as possible because the flow of a component is inversely proportional to membrane thickness.

As a consequence, membranes usually have an *asymmetric structure*. They consist of a very thin skin (active layer) and a porous support layer. As the actual selective barrier for mass transport, the active layer essentially determines the separation efficiency of the membrane.

In the case of asymmetric polymer membranes, integral-asymmetric and composite-asymmetric structures can be distinguished. *Integral-asymmetric membranes* are made by phase inversion (i.e., precipitation of the polymer from a homogeneous solution). The active layer and the substructure consist of the same material. The most successful materials used for integral-asymmetric reverse osmosis membranes are, e.g., cellulose acetate and polyamide.

In the case of *composite-asymmetric membranes*, a homogeneous polymer layer, which is as thin as possible, is applied to a microporous or porous structure. These membranes, called composite membranes, allow individual optimization of the selective layer and the porous support layer with regard to separation efficiency and mechanical, thermal, and chemical stability. In reverse osmosis, the support layer of many membranes is made of polysulfone, and the selective layer is made of, e.g., polyethyleneimine.

9.2.2. *Inorganic Membranes*

Apart from organic membrane materials, inorganic membranes have become increasingly important since the 1980s [245]. Their advantages are:

1. High temperature stability
2. Chemical stability
3. No aging; long service lives
4. Backwashing possible
5. Controllable separation limit and separation efficiency

Their disadvantages are

1. Sensitivity to tension and bending
2. High investment costs
3. Use of temperature-sensitive sealing materials, which limits full utilization of the temperature stability of the membranes

Inorganic membranes are not nearly as important as organic membranes. Materials used are carbon and zirconium oxide – aluminum oxide.

9.3. Modules

(⟶ Membranes and Membrane Separation Processes – Membrane Preparation and Membrane Module Constructions)

As mentioned above, membrane systems are modular (i.e., made up of individual modules). In module development, the following requirements, which are partly contradictory, are essential:

1. Good hydrodynamics (e.g., no dead zones)
2. Mechanical, chemical, and thermal stability of the module
3. High packing density of the module
4. Inexpensive production
5. Good cleaning possibilities
6. Inexpensive replacement of a membrane
7. Low friction losses

Module design must be oriented to these requirements. Since the main emphasis is on one or another requirement, depending on application, many completely differently designed module types are available on the market. Disregarding design details, the modules can be divided into two classes and six types:

Tubular Membranes	Flat-sheet Membranes
Tubular module	Plate and frame module
Capillary module	Spiral-wound module
Hollow-fiber module	Disk module

9.3.1. *Modules with Tubular Membranes*

Tubular Module. In this type of module, the membrane lies in the form of a tube on the inner side of pressure-stable pipes having a diameter of 6 – 24 mm. The mixture to be separated is pumped through the inside of the tube; the permeate is collected outside. If the supporting pipe is made of a material that is impermeable to the filtrate, a thin porous tubular fleece (e.g., of porous polyethylene) is placed between the supporting pipe and the membrane. This fleece does not impede the transport of the filtrate to the bores arranged at short intervals in the supporting pipe and provides the necessary support for the membrane, especially in the region of these bores. Membranes are partly replaceable and partly attached firmly to the supporting material. To increase the relatively low packing density (<80 m^2 of membrane area per cubic meter of space enclosed), many producers arrange several modules in one jacket pipe (Fig. 46).

Figure 46. Tubular modules in jacket pipe

Capillary Module. Unlike tubular modules and modules with flat-sheet membranes, in which the membrane film is mechanically supported by a porous supporting structure, capillary and hollow-fiber membranes are pressure-stable.

The capillary module consists of larger (d_i = 0.5 – 6 mm) membrane capillaries, which have an asymmetric structure with the selective layer on the inside. The capillary module can be compared with a shell-and-tube heat exchanger. The membrane capillaries are arranged in parallel bundles potted in epoxy resin, thus forming a head plate at both ends. Capillary modules have a higher packing density than tubular modules. However, mass transfer conditions are inferior because of laminar flow (see Section Mass Transport Resistance in Front of the Membrane).

Hollow-Fiber Module. Hollow-fiber modules have much smaller diameters (outer diameters 85 – 200 µm) than capillary modules. Hollow-fiber modules are employed in reverse osmosis.

In reverse osmosis, the mixture to be separated is always on the outside of the fiber (i.e., the selective layer is also on the outside), while the permeate flows in the lumen. In contrast to capillaries, hollow fibers are exposed to external pressure. These are favorable conditions, and accordingly hollow-fiber modules can resist pressure differences up to 100 bar. The individual fibers of a hollow-fiber module are assembled to a bundle and installed in a pressure vessel.

9.3.2. *Modules with a Flat-Sheet Membrane*

Plate and Frame Modules. Plate and frame modules are used in reverse osmosis, microfiltration, and ultrafiltration. All designs have the following essential elements: the flat membrane; the support plate for the membrane; and plates for guiding the flow on the feed side. The elements are assembled in form of stacks.

An advantage of plate modules is that the permeate from each membrane pair is collected separately. Thus, membrane defects can be recognized and repaired by disconnecting the defect membrane pair without shutting down the entire module. In all plate modules, a disadvantage is the high number of individual seals and the relatively high friction losses caused by 180° changes in flow direction.

Spiral-Wound Modules. The spiral-wound module is the most widely used module type. It was first developed for reverse osmosis and is also used in nanofiltration and ultrafiltration. The design of a spiral-wound module is shown in Figure 47. In this module, one or several membrane pockets, each with a netlike spacer made of plastic, are wound spirally around the permeate-collecting tube. The membrane pocket consist of two membranes with a porous plastic fleece between them (permeate spacer). The membrane pocket is sealed on three sides and is connected to the perforated permeate-collecting tube on the fourth open side. Feed solution enters at the front surface and flows in axial direction through the module, while the permeate flows spirally within the porous support layer toward the collecting tube.

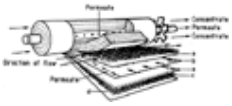

Figure 47. Schematic of a spiral-wound module

a) Membrane; b) Permeate-release layer; c) Spacer; d) Cover

Spiral-wound modules have gained acceptance to a great extent because they combine a simple and inexpensive design with relative insensitivity to fouling and a relatively high packing density (up to 1000 m^2/m^3).

Disk Module. The concept of the disk module is comparable with that of the spiral-wound module. Here, too, two individual membranes are combined with a fabric fleece between them. However, the entire periphery of the disk is sealed. The permeate is collected centrally through an opening provided with an O-ring. A clamping bolt is also located in the central hole. It compresses the membrane disks and spacers in such a way that the permeate flow is sealed off from the raw mixture. The advantages of the disk module are, above all, the low pressure drops on the permeate side and a low susceptibility to plugging on the feed side (this is due to the more “open” spacers of the disk module, compared to the feedspacer of the spiral-wound module). The disadvantage is a relatively low packing density (<400 m^2/m^3) (Fig. 48).

Figure 48. Disk module

Disk modules are used today up to pressure differences of Δp = 200 bar. This is of interest when, in the solutions to be treated, the osmotic pressure determines the limit of concentration. Spiral-wound modules are operated today at maximum pressure differences of 120 bar.

9.4. Fouling and Scaling

(⟶ Membranes and Membrane Separation Processes – Concentration Polarization and Membrane Fouling)

The efficiency of reverse osmosis and nanofiltration is often limited not by the osmotic pressure or the viscosity of the feed solution, but by two other phenomena:

1. *Scaling*: Membrane blockage due to increase in concentration of the dissolved constituents in the feed until the solubility limit is exceeded and the constituents precipitate on the membrane surface
2. *Fouling*: Membrane blockage because of suspended or colloidally dissolved substances carried in or because of growth of microorganisms

9.4.1. *Membrane Blockage due to Crystallization (Scaling)*

If the solubility limits of certain substances are exceeded in an aqueous system during the concentration of constituents, the substance precipitates and can form a layer on the membrane surface. In many cases, these layers can be removed from the

membrane by washing with acid (e.g., citric acid). Especially in the case of spiral-wound and hollow-fiber modules, however, flushing the crystal sludge out of the module is difficult or impossible. For this reason, scaling must be avoided by, e.g.,

1. Chemical conversion, elimination, or stabilization of the substances
2. Restriction of concentration increase in such a way that the saturation concentration of none of the dissolved components is reached even at the critical spot of highest concentration (i.e., the membrane surface at unit exit)

To estimate this concentration limit or to estimate the required raw water pretreatment, the following properties in water analysis must be taken into account:

1. Temperature, pH, conductivity, total dissolved solids (TDS)
2. Acid – base capacity (*m* alkalinity, *p* alkalinity)
3. Concentration of Na^+, Ca^{2+}, Ba^{2+}, Sr^{2+}, Mn^{2+}, Fe^{2+}
4. Concentration of Cl^-, SO_4^{2-}, NO_3^-, F^-, HCO_3^-, CO_3^{2-} (from *m* and *p* alkalinities)
5. Concentration of SiO_2, free Cl_2, O_2, and CO_2 (from *m* and *p* alkalinities)

For the most important dissolved solids, calculations of the solubility limit or maximum possible concentration in the module are available in the literature [246-252].

Important water pretreatment measures (e.g., against $CaCO_3$-scaling) are acid dosing $(HCO_3^- \rightarrow CO_2\uparrow)$; precipitation; and softening (substitution of Na^+ for Ca^{2+}).

Measures against $CaSO_4$ scaling are limiting or even lowering of the water recovery rate so that $CaSO_4$ precipitation is avoided in all events; softening (substitution of Na^+ for Ca^{2+}); precipitation ($Ca^{2+} \longrightarrow CaCO_3 \downarrow$); and stabilization by polyphosphate or organic antiscalants.

9.4.2. *Membrane Blockage due to Contaminants (Fouling)*

In practice, the formation of a surface layer due to suspended or colloidally dissolved substances carried in (and often not eliminated despite very careful pretreatment) or to the growth of microorganisms is the weak point of reverse osmosis. Although the thickness of the surface layer formed during operation is barely measurable, it represents a considerable additional resistance for the permeating components. Therefore, for safe operation of reverse osmosis plants, the fouling behavior must be studied in practice-related experiments with the real system “prefiltered raw water – original membrane.” Pretreatment measures for the prevention of fouling in reverse osmosis are [253-257] (⟶ Flocculants)

1. Filtration through a sand or multimedia deep bed, usually in combination with flocculation
2. Polishing filtration using cartridge filters
3. Microfiltration or ultrafiltration
4. Flotation

9.5. Module Arrangements (Plant Design)

Especially in reverse osmosis, nanofiltration, and ultrafiltration, proper hydrodynamics on the feed side are of central importance. In a continuously operating plant, the mass flow rate (decreasing) and the concentration (increasing) on the feed side change because essentially pure water is transferred from the feed to the permeate at every point of the membrane. In this connection, two module arrangements that take this problem into account are especially important (Figs. 49 and 50): (1) the tapered design and (2) the feed and bleed design, a plant structure made of several module blocks with circulation pumps.

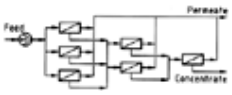

Figure 49. Flow diagram of tapered design

Figure 50. Flow diagram of feed and bleed design (module blocks with circulation pumps)

The *tapered design* is normally used in very large plants for desalination of sea- and brackish water. Plants with *feed and bleed design* are used especially in wastewater treatment where smaller mass flow rates with a greater contamination potential must be treated.

9.6. Use of Membrane Processes in Water and Wastewater Technology

9.6.1. *Desalination of Seawater and Brackish Water*

Fresh water can be separated from sea- or brackish water by (1) evaporation, if the phase transition of water during the separation process is accepted, and (2) membrane processes, where no phase transition occurs [258]. At present and in the near future, seawater will be desalinated mainly by evaporation processes. However, reverse osmosis is an attractive alternative to thermal processes.

The membranes used in the desalination of seawater and brackish water are either composite membranes (e.g., Dow-

Filmtec's FT 30, the membrane Desal 3 of Desalination Industries) or asymmetric phase inversion membranes made of polyamide (e.g., by DuPont). The modules used are spiral-wound, hollow-fiber, and disk modules.

In all RO desalination plants in operation, the great effort involved in the pretreatment of water is conspicuous. The extent of treatment depends, first, on the composition of the raw water and, second, on the module concept employed. The required pretreatment is directly proportional to the sensitivity to fouling and scaling of the module type employed. In disk modules with a relatively open feed channel, adjustment of the pH, chlorination, and separation of the suspended solids by sand filtration are sufficient. In spiral-wound or hollow-fiber modules, additional fine filtration by cartridge filters is required. In general, treatment includes the following steps:

1. Chlorination to prevent growth of algae
2. Filtration (e.g., by a combination “sand filter and subsequent precoat filter”)
3. Adjustment of pH
4. Addition of hexametaphosphate to prevent $CaSO_4$ precipitation in the modules
5. Fine filtration with a 10-µm cartridge filter (5-µm cartridge filter in the case of hollow-fiber modules)

Based on a large-scale operating plant, main design features of a reverse osmosis plant for the desalination of seawater are discussed below.

Seawater Desalination Plant in Las Palmas, Gran Canaria. This seawater desalination plant was put into operation in 1989 (Fig. 51). It has a capacity of about 36 000 m^3/d and consists of six parallel trains of identical design. Each train consists of 128 pressure vessels, which contain six spiral-wound modules each. These are arranged in such a way that eight pressure vessels form one unit. This unit is divided into five pressure pipes in the first block and three pressure pipes in the second block (tapered design).

Figure 51. Seawater desalination plant in Las Palmas, Gran Canaria

The entire plant achieves a water yield of 45 %. The product has a concentration of dissolved substances of 450 – 500 mg/L (seawater: 38 300 mg/L). The specific energy consumption of the plant is 6.16 kW · h/m^3. A turbine at the brine discharge side of the reverse osmosis plant recovers 1.65 kW · h/m^3. The total costs of the plant including pretreatment amounted to ca. \$ 47×10^6 in 1989.

Apart from large-scale plants of this type, reverse osmosis is used especially in desalination plants with lower capacities. Reverse osmosis plants with disk modules have, for instance, fully replaced the evaporators previously used on passenger ships (cruise ships) and warships for the following reasons: (1) savings in floor space, and, even more important, in height, and (2) independence on heat sources such as ship's diesel engines.

9.6.2. *Recovery of ∈-Caprolactam*

∈-Caprolactam, $C_6H_{11}ON$, M = 113.69 g/mol, is the monomer used for the production of polyamide 6. In the production of polyamide 6, a wastewater stream containing 5 wt % of ∈-caprolactam must be treated. Besides the treatment of the water, the objective is to recover the ∈-caprolactam as a concentrated solution. This cannot be achieved with reverse osmosis because osmotic pressure limits the water recovery rate. Assuming 45 bar as the maximum acceptable osmotic pressure of the concentrate ($\Delta\pi < \Delta p$), the concentration of the concentrate is limited to 20 wt % and correspondingly, the water recovery to 75 %. In the example considered here, 254 m^3 of ∈-caprolactam solution accumulates per day so that after the membrane unit a concentrate flow rate of 63.5 m^3/d is obtained which would have to be further concentrated (e.g., by evaporation). In addition, elaborate pretreatment would be required because wastewater has a high fouling potential. Thus, evaporation is carried out before reverse osmosis. Figure 52 shows the flow diagram of the process. The recovered ∈-caprolactam can be reused for the production of polyamide 6. The condensate, which still contains ca. 0.1 % of ∈-caprolactam is concentrated to 5 % in a reverse osmosis stage and recycled to the evaporation stage. The reverse osmosis permeate can be used in the producton.

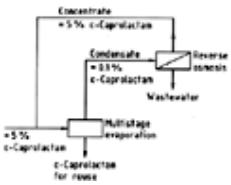

Figure 52. Flow diagram of a plant for separation of ∈-caprolactam from evaporator condensate

The membrane plant in use is equipped with 15 pressure vessels, which contain four 8-inch spiral-wound modules each. The water recovery is 98.4 %, and the rejection of ∈-caprolactam is >99 %.

9.6.3. *Leachates from Dump Sites*

According to the latest version of the German water management law (Wasserhaushaltsgesetz) [259] and successive decrees, leachates from dump sites must be treated before discharge.

Since garbage, industrial solid waste, and special waste have different constituents, the composition of leachates from dump sites depends not only on the age, mode of operation, and construction of the disposal site, but especially on the disposed substances. The average composition of leachates from garbage disposal sites (operating in the stable methane phase) is given in Figure 53.

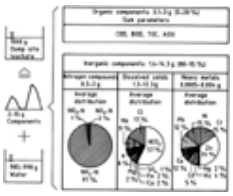

Figure 53. Average composition of dumpsite leachate

In last ten years, the following process variants have proved especially favorable for the treatment of leachates from dump sites:

1. Combination of biological treatment and adsorption
2. Combination of biological treatment and nanofiltration with recirculation of the concentrate from nanofiltration into the bioreactor, partial removal of organics from the NF concentrate by adsorption or ozonation, if necessary.
3. Combination of membrane processes, e.g., RO – high pressure RO – drying of the concentrate – nitrogen removal

Biological Processes. In treating leachates from disposal sites, membrane bioreactors are increasingly replacing traditional techniques (i.e., open-tank bioreactors combined with sedimentation). Instead of sedimentation, ultrafiltration in combination with an aerobic – anoxic biological treatment stage (Fig. 54) has the following advantages [260]: (1) 100 % retention of biomass, especially of the nitrifying bacteria that are problematic in sedimentation; (2) operation at significantly increased biomass concentrations (solids up to 20 – 30 mg/L), resulting in significantly increased volume-specific capacities (space – time yield).

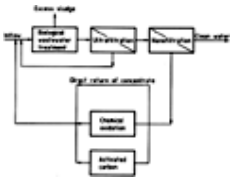

Figure 54. Flow diagram of “BioMembrat Plus” process

With 1-inch tubular modules and feed flow rates of ca. 4 m/s, a constant permeate flux is obtained over a period of weeks (ca. 100 L m^{-2} h^{-1}). An ultrafiltration unit with a membrane area of 44 m^2 combined with anoxic – aerobic closed-vessel bioreactors (plant capacity: 50 – 65 m^3/h) is shown in Figure 55.

Figure 55. Ultrafiltration plant for separation of biomass from biologically clarified leachates

In the biological stage, easily degradable organics (BOD) and ammonium – ammonia are degraded preferentially [261]. The discharge from the bioreactor still contains

1. Recalcitrants
2. Nonbiodegradable organic compounds
3. Nonbiodegradable inorganic compounds (e.g., chlorides or heavy-metal salts)

Substances that are nonbiodegradable or poorly biodegradable are adsorbed conventionally on activated carbon. However, a combination of a membrane bioreactor with nanofiltration is an interesting alternative when the concentrate is recycled to the bioreactor, possibly after partial removal of organics from the NF concentrate by adsorption or ozonation. The mostly higher molar mass recalcitrants are rejected by NF membranes, whereas degradation products and low molar mass salts permeate the membrane with the water. If the concentrate from nanofiltration is recycled to the bioreactor, the hydraulic residence time and the residence time of the recalcitrants are decoupled, in contrast with classic biological treatment. The bioreactor then operates at much higher concentration of recalcitrants than in the raw water. As a consequence the rate of degradation increases according to the Monod law. This rate increases with increasing concentration until inhibition by toxic substances occurs.

In some cases, the amount of water discharged with excess sludge is sufficient to guarantee stable operation of the bioreactor at high concentration (prevention of inadmissible concentration of nondegradable substances in the biological treatment stage). If not, intolerable concentrations in the loop can be reduced via adsorption or ozonation. Compared to a

state-of-the-art combination “bioreactors – adsorption” or “bioreactor – ozonation”, adsorption and ozonation in combination with nanofiltration are much more efficient, because these combinations can operate at favorably high concentration levels instead of the low discharge concentration level.

Reverse Osmosis – Evaporation or Reverse Osmosis – Nanofiltration. A successful alternative to the biological treatment of dumpsite leachate is treatment by reverse osmosis in combination with further concentration steps.

In most cases the RO unit is designed as a two-stage cascade with tubular or disk modules in the first and spiral wound or disk modules in the second stage. Using a two-stage cascade with the permeate of the first stage cascade serving as feed for the second stage the legal limiting values for discharge can be safely guaranteed [262].

Reverse Osmosis – >Evaporation. Usually the RO concentrate is concentrated further by evaporation and, finally, dried when the disposal of a dry residue is stipulated. Furthermore, the process must contain a step for the separation of NH_3/NH_4^+ either a stripping unit or a bioreactor. This process combination can be considered as state of the art. Disadvantageous are the high specific energy consumption and the high specific investment costs of evaporation – dryer stages.

From the composition of leachate (Fig. 53) follows that the osmotic pressure is a limiting factor for concentration by RO. Modules are now available for operation pressures of 120 and 200 bar. By implementation of these modules into the process the evaporation stage can be eliminated. Instead of concentration factors of 2 – 3, concentration factors of up to 10 can be achieved by the RO stage.

Since only small amounts of water are removed in the drying step, which, however, gives rise to 35 – 38 % of the overall treatment costs, it must be investigated whether this process combination can be improved. A simple further increase of the water recovery rate by RO cannot be expected for two reasons:

1. Danger of calcium sulfate scaling
2. Unacceptably high membrane compaction at transmembrane pressure differences above 200 bar

As shown in Figure 53, chlorides form a major part of the inorganic components of leachates. Due to the high rejection of RO membranes for chlorides, they remain almost completely in the concentrate, resulting in a high osmotic pressure and causing disposal problem.

Reverse Osmosis – Nanofiltration. The implementation of a nanofiltration stage into the treatment process can considerably extend the limits of RO set by scaling and/or osmotic pressure. Figure 56 shows the flow diagram of the treatment process. Essentially, the concentrate of the 120 bar RO stage is treated by a combination of NF and crystallization. At moderate transmembrane pressure differences (20 – 50 bar) the nanofiltration unit produces a permeate containing mainly chlorides. This permeate can be concentrated further without danger of scaling by a 200-bar high-pressure reverse osmosis stage. Essential for the process is that the NF – crystallizer cycle is able to operate in the supersaturation range with respect to calcium sulfate and at high concentrations of organics. For this reason the success of the process depends to a large extent on a proper module design — the modules must be insensitive to fouling and to the presence of crystals. Such modules have been developed by Rochem, Germany. Stacks of rectangular membrane cushions and matching spacer plates are arranged in series in the pressure vessel in such a way that stagnant areas are avoided and internal friction losses are minimized. Feed flow is strictly parallel to the vessel axis, the velocity in the feedsite module channels is ca. 1.5 m/s.

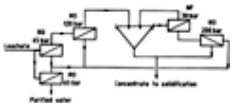

Figure 56. Flow diagram of the combination reverse osmosis – nanofiltration – high pressue reverse osmosis for leachate treatment (Ihlenberg dumpsite, Germany)

The process has been installed and tested on a technical scale at the Ihlenberg dumpsite. A nanofiltration stage consisting of four blocks with nine modules each and a total membrane area of 180 m^2 has been added to the existing reverse osmosis – high-pressure reverse osmosis combination and comissioned in September 1994 [263]. The unit is designed for the treatment of about 4 m^3/h RO concentrate with a volumetric concentration factor of $CF_V^{RO} = 10$. Depending on the concentration factor of the 60 and 120 bar RO stage, the NF stage achieves a further concentration of $CF_V^{NF} = 10 – 20$. A sludge consisting of organics, precipitated inorganics, and water is discontinuously withdrawn from the bottom of the crystallizer – sedimentation tank.

Together with the concentrate of the subsequent 200 bar high-pressure reverse osmosis stage this concentrate is solidified with fly ash and disposed of at the dumpsite.

Cleaning of the NF modules is performed by flushing with feed at zero transmembrane pressure difference for 30 s every hour and an alkaline cleaning every 250 – 300 h. The NF rejection rate for sulfates is 92 – 95 %, for the dissolved solids 20 – 35 %.

At the Ihlenberg dumpsite the specific energy consumption of the NF and the subsequent 200 bar RO stage is 32 kW · h per cubic meter permeate. With figures of 8.5 kW · h per cubic meter of totally produced permeate and an overall water recovery rate of 97 %, the overall specific power consumption of the process is extremely low compared to other processes. With respect to residues the process can be regarded as an almost zero discharge process.

9.6.4. *Wastewater from the Dye Industry*

The production of textile dyes involves a precipitation by salting-out as the final stage. After separation of the dye, the

wastewater of the process contains the salt and, unavoidably, small amounts of dye. Treatment of this wastewater by reverse osmosis is not possible because of the high osmotic pressure and because the resulting concentrate would contain both salts and organics. However, the ability of nanofiltration to fractionate low molar mass salts and higher molar mass dyes can be used here.

A plant installed in a Swiss chemical company is shown in Figure 57. It has a total membrane area of 960 m^2 and is equipped with tubular modules because of the high fouling potential of feed. It operates at a volumetric concentration factor CF_V = 10 and has an average capacity of ca. 16 m^3/h. Since the wastewater of several dye-production plants is treated simultaneously, the NF concentrate cannot be recycled, but must be disposed of. To this end, the concentrate is subjected to wet oxidation. The permeate, containing the major part of the salts, can be discharged into the Rhine river. Here, nanofiltration has two advantages:

1. The mass flow rate of the stream subjected to wet oxidation is significantly lower.
2. The corrosion potential of this stream is strongly reduced, because the major part of the salts is removed.

Figure 57. Nanofiltration plant for treatment of wastewater from the production of textile dyes

Obviously, nanofiltration offers a possibility of recovering the dyes, at least in principle. This would be of economic interest where large amounts of only one dye are produced.

[Top of Page]

10. Oxidation Processes in Water Treatment

William H. Glaze

Oxidation processes are used in water treatment for disinfection and removal of obnoxious or potentially toxic contaminants in the water, and for some industrial purposes such as pulp bleaching and purification of wastewater before discharge into the environment. The most familiar chemical oxidant is oxygen from air; other examples of oxidants used in water treatment are chlorine, chloroamines, ozone, chlorine dioxide, hydrogen peroxide, and potassium permanganate. Sometimes these oxidants are used in combination with each other (e.g., ozone with hydrogen peroxide) or with photons from irradiation lamps. Oxidation in water treatment is often carried out with the assistance of microbiological organisms, in which case the oxidant is oxygen (⟶ Wastewater – Biological Principles, ⟶ Wastewater – Basic Biotechnological Considerations, ⟶ Wastewater). Only chemical oxidation is considered in this chapter.

10.1. Physical and Chemical Properties of Chemical Oxidants

The power of a chemical oxidant to cause a chemical oxidation reaction is measured by two properties: (1) the thermodynamic driving force and (2) the rate constant of oxidation. The thermodynamic driving force of the oxidation process for one oxidant compared to another is determined by its oxidation potential. Table 31 is a list of the standard potentials E^0 for common chemical oxidants in aqueous solution at 298 K [265], [266].

Table 31. Standard electrode potentials for some common oxidizing agents in aqueous solution at 298 K [265], [266]

Oxidant	Half reaction	E^0, V
Oxygen [7782-44-7]	$O_2 + 4\,H^+ + 4\,e^- \rightleftharpoons 2\,H_2O$	1.23
Ozone [10028-15-6]	$O_3(g) + 2\,H^+ + 2\,e^- \rightleftharpoons O_2(g) + H_2O$	2.08
Chlorine [7782-50-5]	$Cl_2 + 2\,e^- \rightleftharpoons 2\,Cl^-$	1.36
Hypochlorous acid [7790-92-3]	$HOCl + H^+ + 2\,e^- \rightleftharpoons Cl^- + H_2O$	1.50
Hypochlorite	$ClO^- + H_2O + 2\,e^- \rightleftharpoons Cl^- + 2\,OH^-$	0.90
Monochloroamine [10599-90-3]	$ClNH_2 + H_2O + 2\,e^- \rightleftharpoons Cl^- + NH_3 + OH^-$	0.75
Chlorine dioxide [10049-04-4]	$ClO_2 + 2\,H_2O + 5\,e^- \rightleftharpoons Cl^- + 4\,OH^-$	1.71
Permanganate	$MnO_4^- + 4H^+ + 3e^- \rightleftharpoons MnO_2 + 2H_2O$	1.68
Hydrogen peroxide [7722-84-1]	$H_2O_2 + 2\,H^+ + 2\,e^- \rightleftharpoons 2\,H_2O$	1.78

The oxidation potential is a measure of a chemical's oxidizing (or reducing) power; the more positive the value of E^0, the more likely is the chemical reaction to take place as written. The parameter E^0 is related to the free energy ΔG^0 by the

formula $\Delta G^0 = -nFE^0$, where n is the number of electrons transferred and F is the Faraday constant (9.648×10^4 J/V). For example, the large positive value of E^0 (2.08 V) for Equation (10.1) below indicates that ozone is a powerful oxidizing agent.

The values in Table 31 are written as half reactions; thus, another half reaction is needed to complete the chemical equation. If the compound being oxidized by ozone is formate (Equation 10.2), the overall process would be as follows:

$$O_3(g) + 2H^+ + 2e^- \rightleftharpoons O_2(g) + H_2O \qquad E^0 = 2.08\ \text{V} \quad (10.1)$$

$$HCOO^- \rightleftharpoons CO_2(g) + H^+ + 2e^- \qquad -E^0 = 0.31\ \text{V} \quad (10.2)$$

$$O_3(g) + HCOO^- + H^+ \rightleftharpoons CO_2(g) + O_2(g) + H_2O \quad E^0_{net} = 2.39\ \text{V} \quad (10.3)$$

The overall chemical free energy change for the process represented by Equation (10.3) (i.e., the driving force) is calculated from

$$\Delta G^0 = nFE^0_{net} = -(2)(9.648\times10^4\ \text{J/V})(2.39\ \text{V}) = -461\ \text{kJ}$$

The large negative value of ΔG^0 indicates that the reaction should proceed spontaneously.

All common chemical oxidants shown in Table 31 are relatively powerful and should oxidize most organic compounds to carbon dioxide and water. The fact that this does not always occur (e.g., organic compounds on our planet are stable to oxidation by oxygen in most cases) indicates that thermodynamics is not the only important factor in determining whether a chemical reaction will occur spontaneously.

The other factor that is important in rating chemical oxidants is how fast they cause a chemical oxidation to proceed. This is the domain of chemical kinetics. In fact, in very few cases do the oxidants listed in Table 31 oxidize environmental contaminants as completely as would be expected from thermodynamic considerations. Chemical reactions that should occur, as indicated by the thermodynamic driving force, may do so very slowly if the rate constant for the process is low. For example, the reaction of formate with ozone is actually a very slow reaction, whereas that of another common pollutant, phenol, is very fast [267]. Another example is the rate of oxidation of two important natural products that contribute to bad taste and odor in water supplies: geosmin and methylisoborneol. The rates of chemical oxidation of geosmin and methylisoborneol by chemical oxidants vary substantially, making some impractical for use in water treatment. Of the common oxidants, ozone is more effective than others for the control of obnoxious taste and odor.

Chlorine. The solubility of chlorine in water is high compared to that of oxygen because in water, Cl_2 hydrolyzes to form hypochlorous acid:

$$Cl_2 + H_2O \rightleftharpoons HOCl + H^+ + Cl^- \quad (10.4)$$

Hypochlorous acid is a weak acid with a pK_a of 7.50 at 25 °C, dissociating to a hydrogen ion and the hypochlorite ion:

$$HOCl \rightleftharpoons H^+ + OCl^- \quad (10.5)$$

This means that at pH 7.50, equal amounts of HOCl and hypochlorite ion (OCl^-) are present. At higher pH, OCl^- predominates; at lower pH, HOCl. Hypochlorite ion is a weaker oxidant than HOCl (see Table 31) and also a weaker disinfectant. Both are powerful oxidizing agents as indicated by the E^0 values in Table 31, so aqueous chlorine is capable of oxidizing many organic and inorganic compounds [268], [269]. For example, chlorine has been used for years as a bleaching agent in pulp and paper treatment, for the oxidation of sulfides and cyanides, and for the oxidation of metals such as iron and manganese. Aqueous chlorine also reacts with organic compounds to substitute chlorine for hydrogen, which yields new chlorinated compounds (see Section Toxicology and Environmental Health). Chlorine is a toxic gas, and worker exposure must be avoided (⟶ Chlorine – Toxicology).

Chloroamines. Chloroamines are formed by reaction of chlorine with ammonia [268], [269]. The reaction occurs in three steps as shown below, leading to mono-, di-, and trichloroamines. These have slightly different oxidation potential and disinfection power, but in general, chloroamines are not as potent in both respects as chlorine. However, chloroamines are more stable in water than chlorine so they can maintain an active disinfection capacity for longer periods of time.

$$HOCl + NH_3 \rightarrow NH_2Cl + H_2O \quad (10.6)$$

$$HOCl + NH_2Cl \rightarrow NHCl_2 + H_2O \quad (10.7)$$

$$HOCl + NHCl_2 \rightarrow NCl_3 + H_2O \quad (10.8)$$

Ozone is always produced when oxygen is decomposed in air, e.g., in an electric discharge or when shortwave (<190 nm) UV radiation is absorbed by oxygen. This decomposition occurs naturally in the atmosphere, but it may also be used to generate ozone for water treatment by the reaction shown below using either air, oxygen-enriched air, or pure oxygen.

$$O_2 \xrightarrow{\text{Energy}} O^{\bullet} + O^{\bullet} \qquad (10.9)$$

$$O^{\bullet} + O_2 \rightarrow O_3 \qquad (10.10)$$

Ozone is more soluble in water than oxygen but less soluble than chlorine. In water it decomposes slowly, with a half-life that depends on the other substances present in water. Ozone is a toxic gas, and monitors are necessary where ozone is made and used, to minimize worker exposure.

Ozone is a powerful oxidant but reacts with different chemicals at various rates that range over several orders of magnitude. For example, the half-lifes of ozone's reactions with phenol and formate under similar conditions differ by a factor of about 100 000 [267]. Like chlorine, ozone has been used for the treatment of drinking water for disinfection, color and taste control and other purposes for over nine decades (ozone since 1906, installation at the Bon Voyage plant in Nice, France) [270]. Ozone has been more popular in Europe, especially France, than in North America, but this has been changing lately because of some of the environmental impacts of chlorine (see Section Toxicology and Environmental Health). Ozone alone is not always effective in oxidizing organic substances in water such as pollutants or taste and odor compounds, but its effectiveness may be enhanced by combining it with other oxidants (such as hydrogen peroxide) or with UV radiation [270].

Chlorine dioxide (⟶ Chlorine Oxides and Chlorine Oxygen Acids – Chlorine Dioxide) is a clear, colorless gas that dissolves readily in water without dissociation. It is a powerful disinfectant and oxidant (Table 31) and has been used extensively for water treatment. Chlorine dioxide is becoming an increasingly important alternative to chlorine as a bleaching agent in pulp and paper production. However, chlorine dioxide reacts rapidly with many substances in natural water, such as natural organic matter, in a reduction – oxidation reaction that produces chlorite ion (ClO_2^-). Chlorite has some documented health effects such as irreversible binding to hemoglobin, so the total amount of chlorine dioxide used in drinking water treatment is regulated in some countries (e.g., United States). Also, residuals of chlorine dioxide in drinking water have an objectionable taste and odor to some people.

Potassium permanganate has been used in water treatment for many years, primarily for oxidation of manganese(II) and iron(II) to the corresponding insoluble oxides or hydroxides, for color removal, and for taste and odor control. Potassium permanganate is usually not recommended for drinking water disinfection. Also, care must be taken not to overdose and produce a pink-colored water, which is objectionable to consumers.

Hydrogen peroxide (⟶ Hydrogen Peroxide) is a milder oxidant than chlorine or ozone. It is used in many industrial and medical applications as a mild disinfectant or oxidant, but worldwide its principal application is in wood pulping.

10.2. Production of Chemical Oxidants

Chlorine is produced in very large quantities worldwide, only a smal fraction of which is used for water treatment. Chlorine gas is produced mainly by electrolysis of brine (sodium chloride solution) according to the chlor – alkali, the mercury cell, or the diaphragm process (⟶ Chlorine, ⟶ Chlorine – Mercury Cell Process, ⟶ Chlorine – Diaphragm Process, ⟶ Chlorine – Membrane Process). Pure chlorine can be transported as liquefied gas and added to water under pressure, where it dissolves readily and hydrolyzes to hypochlorous acid and hypochlorite ion (see Eqs. 10.4 and 10.5). Commercial units are available to meter a specified amount of chlorine into water, usually into a slipstream which is then added to the total water stream to be treated so as to give the desired dose. Alternatively, aqueous chlorine (hypochlorite) may be prepared at the site of chlorine production and shipped to the location where it can be metered into the water to be treated. The latter is preferred in some highly populated areas where transport of gaseous chlorine is forbidden.

Chloroamines. The formation of chloroamines is carried out on-site by combining a chlorine solution with aqueous ammonia. To minimize byproducts, the chloroamines may be preformed by using an excess of ammonia so that no chlorine remains. A chlorine – ammonia weight ratio of 3: 1 corresponds roughly to the stoichiometric amount to form monochloroamine plus some dichloroamine, and a pH of 8.4 is optimal. Excess chlorine leads to the formation of trichloroamine, which has an objectionable taste.

Ozone is always produced on-site because it is an unstable gas that cannot be stored or shipped. Ozone is usually produced by an electric discharge in air, oxygen, or oxygen-enriched air (⟶ Ozone – Production). The most commonly used generators are of the silent discharge type in which discharge occurs across a dielectric barrier [270]. Small ozone generators may utilize discharge between ceramic electrodes or irradiation lamps.

Chlorine dioxide is also formed on-site, usually by reaction of sodium chlorite and chlorine (hypochlorous acid):

$$HOCl + 2\,ClO_2^- + H^+ \rightarrow 2\,ClO_2 + Cl^- + H_2O \qquad (10.11)$$

Commercial generators are available to regulate the ratio of chlorine and chlorite carefully or to recycle the stream so as to minimize the content of chlorine and byproducts in the product and to consume all of the chlorite ion. However, chlorine dioxide used in water treatment always contains some chlorite and usually chlorate, as well.

Potassium permanganate is produced by electrolysis of potassium manganate(VI) (K_2MnO_4), which is obtained by melting manganese dioxide and potassium hydroxide in the presence of air (⟶ Manganese Compounds – Fusion Processes , ⟶ Manganese Compounds – Anodic Oxidation of Manganate(VI)).

Hydrogen peroxide is usually produced from oxygen by using a catalytic process involving anthraquinone (⟶ Hydrogen Peroxide – Anthraquinone Process (AO Process)). It is marketed as a solution in water, usually containing 35 – 50 % H_2O_2. The compound is unstable, decomposing to oxygen and water by a reaction that is catalyzed by several substances, even

common soil. Hydrogen peroxide must therefore be shipped and stored in vessels made of passivated stainless steel, for example, in which it is more stable.

Mixed Oxidants. Oxidants are sometimes used in combination, the most common example being a mixture of ozone and hydrogen peroxide. These are always mixed at the site of application because the chemical reaction that occurs produces free-radical intermediates that have a very short lifetime (see Section Advanced Oxidation Processes).

10.3. Uses of Chemical Oxidants

Oxidants are used in water treatment for a variety of purposes, including disinfection and oxidation, in the treatment of drinking water and wastewater, oxidation in pulp and paper treatment, and disinfection and scale control in the treatment of cooling tower water.

10.3.1. *Drinking Water Treatment*

Chemical oxidants are used for treatment of drinking water, mostly for disinfection and removal of obnoxious chemical compounds or potential toxicants. The most common drinking water oxidant worldwide is chlorine (hypochlorous acid). In parts of Europe, ozone has been used for as long as chlorine, and it is growing in use worldwide.

10.3.1.1. *Disinfection*

Treatment with chemical oxidants has become the traditional method for disinfection of water for distribution in municipal water systems. Ozone and chlorine were first used for this purpose at about the beginning of this century and have been used continuously since that time [268], [269]. Chlorine is the more preferred of the two because it is more stable and maintains its disinfection capability for longer periods of time. Ozone decomposes quickly but is a more powerful disinfectant.

Drinking water sources can be classified generally as groundwater or surface water. Groundwater often does not require treatment except for disinfection, usually by addition of a chemical oxidant such as chlorine. In a drinking water plant for treating surface water, disinfection is accompanied by other unit processes such as coagulation (see Chap. Flocculation), sedimentation, and filtration (see Chap. Filtration). Also, the chemical oxidant that is applied as a disinfectant may be used to achieve other treatment objectives such as chemical oxidation (see Section Chemical Oxidation). The oxidant – disinfectant may be added at the beginning of treatment (peroxidation) or at other points in the treatment train. The amount of oxidant added and its contact time with the water are chosen such that the kill of microorganisms is highly efficient. At the end of the treatment train, a smaller amount of disinfectant (e.g., chlorine) is usually added so that water being sent to the customer will have a residual amount of disinfection protection. In Western Europe, the amount of residual disinfectant used is lower (≈0.1 mg/L) than is traditional in the United States (≈ 1 mg/L).

Disinfection efficacy is usually measured in terms of the dose (mg/L) of oxidant required to kill a certain fraction of a microorganism. Since the kill is also determined by the time the microorganism is exposed to the dose, the product of dose (C) times contact time (T) is an important process control parameter. In the United States, the Environmental Protection Agency has published values of CT for each disinfectant required to kill a certain fraction of various microorganisms. For tests used in other countries, see ⟶ Disinfectants – Efficacy Testing Methods, Requirements for Efficacy.

10.3.1.2. *Chemical Oxidation*

Oxidants are also used in drinking water treatment to remove substances that give unwelcome color to the water, compounds that cause poor taste or odor, and micropollutants that may have deleterious health effects. In addition, oxidants are used to remove iron and manganese, which are not toxic but may cause discoloration of household appliances.

Color in natural water is caused by natural organic matter (NOM) and sometimes by metallic elements complexed with NOM. Color is measured by comparison with a standard solution of a colored chemical compound such as chloroplatinic acid. The color of the water is expressed as “color units” (i.e., the amount of chloroplatinic acid necessary to give the same color). In drinking water plants using water sources that are highly colored, the water is usually treated with a chemical oxidant so that product water does not have a discernible color. Oxidants destroy color by chemically oxidizing the color center in the macromolecular NOM. Of the common chemical oxidants, ozone appears to be most effective for removing color. However, since the chemical structures in NOM that cause color vary from source to source, another oxidant may be more or less successful than ozone in removing the color, depending on the water source.

Taste and odor in water supplies are usually caused by chemical compounds produced by algae and other organisms but may also result from pollution due to industrial discharge. Again, the source of the bad taste or odor may vary from source to source, so a particular chemical oxidant may vary in its ability to remove the problem. Two compounds produced by algae, geosmin and methylisoborneol, are particularly obnoxious in that they can be sensed by humans at the ng/L level. Ozone is virtually the only common chemical oxidant that is effective in removing geosmin and methylisoborneol.

Iron and manganese are usually present in natural water. Higher oxidation states of both metals form rather insoluble hydroxides or oxides [e.g., $Fe(OH)_3$ and MnO_2], but lower oxidation states are more soluble (Fe^{2+} and Mn^{2+}). Therefore, these soluble forms are likely found under anoxic (reducing) conditions, such as in groundwater and below the hypolimnion in surface water. These substances are preferably removed in drinking water treatment before the water is sent to customers, where oxidation may cause precipitation of unwanted, colored oxides in plumbing. Oxidation of iron is a rapid reaction that can be accomplished easily in water treatment, even with molecular oxygen. All of the common oxidants can also easily oxidize Fe(II) to Fe(III), which is then precipitated and removed by filtration. Stronger oxidants are required for oxidation of Mn(II) to Mn(IV). Chlorine, potassium permanganate, and ozone are most commonly used, usually at the beginning of treatment prior to coagulation and sedimentation.

Micropollutants are synthetic organic or inorganic chemicals present in natural surface water sources because of discharges from industries, municipal waste treatment plants, and other sources. For example, surface waters in Europe and the United States are contaminated with agricultural chemicals, especially in the spring when herbicides are used in

abundance. Among the common oxidants, ozone is the most effective for removal of these and other organic micropollutants, but addition of hydrogen peroxide just before ozonation increases removal. The chemical mechanism responsible for the synergy between ozone and hydrogen peroxide is discussed in Section Advanced Oxidation Processes.

Groundwater may be contaminated with solvents, petroleum distillates, pesticides or herbicides, and other compounds that migrate into aquifers from spills, unwise use, and other human activities. A variety of treatment methods have been developed to clean contaminated groundwater, including activated carbon adsorption and air stripping. Oxidants may also be used to treat contaminated groundwater, but again the efficacy of treatment will vary depending on the molecular form of the contaminant. With regard to the most common groundwater contaminants, oxidation has been most successful in treating tri- and tetrachloroethylene, solvents that have leaked into reservoirs from storage tank, and accidental spills. The most successful oxidation methods for this purpose are ozone with hydrogen peroxide and ozone or hydrogen peroxide with UV radiation.

10.3.2. *Wastewater Treatment*

The use of chemical oxidants is not as common in wastewater treatment as in drinking water treatment because wastewater usually contains more impurities and, therefore, would consume more oxidants. Since oxidants are rather costly commodities, they are generally used for polishing (i.e., treating wastewater just before discharge) after most impurities have been removed [268], [269].

Chlorine has been used traditionally for treatment of domestic wastewater after biological treatment. The rationale was to minimize the concentration of microorganisms and ammonia being discharged into receiving streams. The chemistry of ammonia removal is essentially the same as that used to form chloroamines (see Section Physical and Chemical Properties of Chemical Oxidants), but in this case an excess of chlorine is added to oxidize the ammonia completely to nitrogen. This is referred to as *breakpoint chlorination*:

$$3\,HOCl + 2\,NH_3 \rightarrow N_2 + 3\,Cl^- + 3\,H^+ + 3\,H_2O \qquad (10.12)$$

Breakpoint chlorination of wastewater is still practiced, but concern over the chlorine byproducts produced is increasing so alternative oxidants such as ozone are gaining in favor.

10.3.3. *Oxidation in Pulp and Paper Treatment*

Chemical oxidants are used for bleaching pulp and paper (⟶ Paper and Pulp – Conventional Bleaching, ⟶ Paper and Pulp), and for decolorizing wastewater prior to discharge. Chlorine was the chemical oxidant of choice, but during the last two decades evidence of the formation of chlorinated byproducts including chlorinated dioxins and furans, has forced a reexamination of this practice. Of particular concern is the formation of chlorinated dioxins such as the highly carcinogenic 2,3,7,8-tetrachloro isomer. Modification of pulp and paper treatment is proceeding on a worldwide basis. Chlorine dioxide, oxygen, hydrogen peroxide, and ozone are being used instead of chlorine.

10.3.4. *Advanced Oxidation Processes*

The term advanced oxidation processes (AOPs) is used for water treatment processes that involve the formation of highly reactive, short-lived chemical intermediates. These intermediates (e.g., the hydroxyl radical) are powerful oxidants — more powerful than the conventional oxidants discussed above — and are used to oxidize water contaminants that resist other forms of treatment. Several AOPs have been developed and commercialized including the following:

1. Ozone in combination with hydrogen peroxide (sometimes called Peroxone process) with UV radiation, or with both hydrogen peroxide and UV
2. Hydrogen peroxide in combination with iron(II) salts (Fenton process), with UV radiation, or with UV radiation and other modifiers such as iodine ion, iron(III) salts, etc.
3. Oxygen in combination with high-energy, high-frequency sound waves (sonication), electron-beam irradiation, or gamma radiation

Oxidation with AOPs is used worldwide in the treatment of drinking water for oxidation of herbicides such as atrazine and simazine, and for oxidation of taste and odor compounds that are formed in rivers, lakes, and reservoirs. AOPs are also used in groundwater treatment to oxidize contaminants such as halogenated solvents (trichloroethylene and tetrachloroethylene). In addition, AOPs are being investigated for treatment of industrial wastewater, but they appear to be substantially more expensive than conventional biological treatment. Thus, AOPs may be most useful for treating low-volume streams containing compounds that are not easily biodegraded. In this case, oxidizing the target compounds only partially (i.e., not completely to carbon dioxide) may be most economical. Partially oxidized compounds are likely to be more easily biodegradable than their precursors, and the stream may then be given to a biological treatment unit process for final polishing before discharge into the environment.

Advanced oxidation processes that involve UV radiation utilize lamps such as mercury arc lamps that produce radiation of various wavelengths, depending on the power density of the current in the lamp. Low-pressure mercury arc lamps produce UV radiation primarily at 254-nm wavelength, whereas high-pressure lamps (operated at higher current densities and therefore at higher internal pressure and temperature) produce UV radiation with variable intensity at several different wavelengths from the visible region down to less than 200 nm. However, lamps are usually housed in quartz envelopes that cut off all radiation below 200 nm or, in the case of highly purified quartz, below 185 nm.

Low-pressure mercury arc lamps are suitable for disinfection of water because the 254-nm wavelength is absorbed by the DNA in bacterial and viral cells. This wavelength is also suitable for decomposition of ozone in the ozone – UV AOP, since ozone absorbs strongly at 254 nm. The following chemical equations summarize this complex process, which undoubtedly involves several steps not shown:

$$O_3 \xrightarrow{\text{UV radiation}} O^{\bullet} + O_2 \qquad (10.13)$$

$$O^{\bullet} + H_2O \rightarrow H_2O_2 \qquad (10.14)$$

$$H_2O_2 + 2\,O_3 \rightarrow 2\,OH^{\bullet} + 3\,O_2 \qquad (10.15)$$

$$OH^{\bullet} + \text{Organic contaminant} + O_2 \rightarrow \text{Partially oxidized byproducts} \qquad (10.16)$$

$$\text{Partially oxidized byproducts} + OH^{\bullet} + O_2 \rightarrow CO_2 + H_2O \qquad (10.17)$$

When ozone is used in combination with hydrogen peroxide, the last three equations summarize the chemical reaction.

High-pressure mercury arc sources are more suitable for other AOPs such as the H_2O_2 – UV process where UV radiation decomposes peroxide into OH radicals, which then oxidize water contaminants. The decomposition of hydrogen peroxide is described by the equations shown below:

$$OH^{\bullet} + \text{Organic contaminants} + O_2 \rightarrow \text{Partially oxidized byproducts} \qquad (10.16)$$

$$\text{Partially oxidized byproducts} + OH^{\bullet} + O_2 \rightarrow CO_2 + H_2O \qquad (10.17)$$

$$H_2O_2 \xrightarrow{\text{UV radiation}} 2\,OH^{\bullet} \qquad (10.18)$$

Similarly, when hydrogen peroxide is used in combination with iron(II) salts, OH radicals are produced by an oxidation – reduction reaction:

$$H_2O_2 + Fe^{2+} \rightarrow OH^{\bullet} + OH^{-} + Fe^{3+} \qquad (10.19)$$

The same radicals can be produced by direct decomposition of water by high-energy radiation or sonication:

$$H_2O \xrightarrow{\text{high-energy radiation}} OH^{\bullet} + H^{\bullet} \qquad (10.20)$$

As noted above, OH radicals in combination with oxygen oxidize organic contaminants in water eventually to form oxidized byproducts and, if enough oxidant is used, to form carbon dioxide and water. Whether the AOP is used to oxidize the contaminant completely is determined by economics.

10.4. Toxicology and Environmental Health

Chlorine and other chemical oxidants have had a very positive influence on human health because they are the principal disinfectants used in drinking water treatment, disinfection of sanitary facilities, and other applications. The incidence of waterborne diseases such as cholera and typhoid fever decreased dramatically in the early decades of the 20th century as water disinfection was practiced more extensively. In parts of Europe, particularly France, ozone served as the principal water disinfectant, whereas chlorine was preferred worldwide.

Chemical oxidants are generally toxic compounds themselves, so workers and the general public should be protected from large doses. However, low levels of oxidants, particularly chorine and chloroamines, are consumed regularly in drinking water with no apparent ill effects.

Disinfection Byproducts. In 1974, workers in the Netherlands and the United States discovered that chlorination of water produces a set of previously undetected byproducts known as the trihalomethanes (THMs) [271]. Trihalomethanes in this context usually mean four compounds: $CHCl_3$, $CHCl_2Br$, $CHClBr_2$, and $CHBr_3$, but other analogous containing mixtures of iodine, bromine, and chlorine (but not fluorine) as substituents are produced in low quantities. As a class, these compounds are called disinfection byproducts (DBPs). Subsequent studies showed that these and many other chlorinated and brominated byproducts are formed from the reaction of aqueous chlorine and the organic matter present in all natural water.

$$HOCl + NOM \rightarrow CHCl_3 + \text{Other chlorinated DBPs} \qquad (10.21)$$

$$HOCl + Br^{-} \rightarrow HOBr + Cl^{-} \qquad (10.22)$$

$$HOBr + HOCl + NOM \rightarrow \text{Mixed Cl — Br THMs} + \text{Other DBPs} \qquad (10.23)$$

Other chlorinated DBPs include halogenated acetic acids such as trichloroacetic acid and mixed chloro – bromo analogues, chlorinated acetonitriles such as Cl_2HCCN, halogenated aldehydes and ketones, and many others. Studies have shown that the yield of these compounds is generally in the microgram – milligram per liter range and depends on the concentration of NOM in the water, the pH, temperature, and chlorine dose.

Concern over the formation of DBPs was immediate because some of the compounds had previously been shown to be carcinogenic in test animals or to have other toxic effects. As a result, THMs are now the subject of water quality regulations in many countries, and maximum concentration levels are recommended by the World Health Organization. Depending on the country, the limit of THMs allowed in finished drinking water is in the range of 100 – 200 µg/L.

As a result of these regulations, which may be strengthened in the future, the use of chlorine in water treatment is being

reexamined. Many water treatment facilities are now using lesser amounts of chlorine; others have shifted to an alternative disinfectant such as ozone, chloroamines, or chlorine dioxide. However, all chemical oxidants react with naturally occurring compounds in water and form chemical byproducts [272]. For example, ozone reacts with naturally occurring bromide ion to form bromate ion, a substance that causes cancer in animal bioassays. The oxidation of bromide proceeds through the formation of hypobromite ion, which also may form byproducts on conversion to its corresponding acid form:

$$Br^- + O_3 \rightarrow BrO^- + O_2 \quad (10.24)$$

$$BrO^- + 2\,O_3 \rightarrow BrO_3^- + 2\,O_2 \quad (10.25)$$

$$BrO^- + H^+ \rightarrow HOBr \quad (10.26)$$

$$HOBr + NOM \rightarrow \text{Brominated Organic byproducts} \quad (10.27)$$

Chlorine dioxide and chloroamines form byproducts on reaction with NOM, although these products have not been as well characterized.

Other Oxidation Byproducts. Chemical oxidants used in wastewater treatment also form chemical byproducts and, in some cases, may affect the natural species in the receiving streams into which these waters are discharged. For example, residual ammonia in domestic wastewater is chlorinated to form chloroamines (see Section Physical and Chemical Properties of Chemical Oxidants). Chloroamines are toxic to fish and other aquatic organisms. In addition, chlorine may react with unidentified substances to form chlorinated dioxins, furans, furanones, and thiophenes. Some of these may have direct toxic effects, others may bioaccumulate in organisms and have long-term effects.

Oxidation byproducts may be produced even when oxidants are used in groundwater treatment. For example, oxidation of tri- and tetrachloroethylene, common groundwater contaminants, produces small amounts of chlorinated acetic acids, which may be more potent carcinogens than their precursors. Also, oxidation of simazine and atrazine, two common contaminants in surface water, produces various partially oxidized derivatives of the original compound.

In summary, all chemical oxidants form byproducts when they are used in water treatment. In some cases, such as drinking water chlorination, these byproducts have been well characterized and their toxicological properties are well known. World health authorities agree that the levels of these byproducts should be minimized, and regulations to do so have been promulgated in many countries. For other oxidants, very little information on the formation of chemical oxidation byproducts is available and further research is needed.

Finally, the risks from chemical byproducts formed by the use of oxidants as water disinfectants are rather low. Toxicologists estimate that the risk of obtaining cancer from the consumption of water containing chloroform at a concentration of 100 µg/L for a lifetime is only about 1 out of 100 000 – 10 000. In contrast, the absence of water disinfection can result in large outbreaks of microbiologically mediated diseases such as cholera, typhoid fever, and gastroenteritis. The risks of disinfection byproducts should always be compared with the risks of the microbiological species that disinfection removes [273].

[Top of Page]

11. Acknowledgement

The entire topic was coordinated by Fritz H. Frimmel

[Top of Page]

References

1. N. E. Dorsey: *Properties of Ordinary Water Substance*, Reinhold Publ. Co., New York 1968.
2. D. Eisenberg, W. Kauzmann: *The Structure and Properties of Water*, Oxford University Press, Oxford 1969.
3. F. Franks (ed.): *Water. A Comprehensive Treatise*, vol. 1, Plenum Press, New York 1972.
4. A. Rahman, F. H. Stillinger, *J. Chem. Phys.* **55** (1981) 3334.
5. E. U. Franck, K. Roth, *Disscuss. Faraday Soc.* **43** (1967) 108.
6. G. S. Kell in [1], chap. 7, p. 363.
7. V. A. Kirillin (ed.): *Heavy Water, Thermophysical Properties*, U.S. Department of Commerce, National Technical Information Service, Springfield, Va., 1971.
8. L. Haar, J. S. Gallagher, G. S. Kell: “NBS/NRC Steam Tables: Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units,” *Nat. Bur. Stand. (U.S.) Nat. Stand. Ref. Data Ser.* 1984
9. U. Grigull, J. Straub, P. Schiebener: *Steam Tables in SI-Units*, 3rd ed., Springer Verlag, Berlin 1990.
10. F. Franks (ed.): *Water. A Comprehensive Treatise*, vol. 7: “Water and Aqueous Solutions at Subzero Temperatures,” Plenum Press, New York 1982.
11. H. Weingärtner, *Z. Phys. Chem. (Wiesbaden)* **132** (1982) 129.
12. W. L. Marshall, E. U. Franck, *J. Phys. Chem. Ref. Data* **10** (1980) 295.
13. M. Uematsu, E. U. Franck, *J. Phys. Chem. Ref. Data* **9** (1980) 1291. Links
14. JANAF Thermochemical Tables, *Nat. Stand. Ref. Data Ser. (U.S. Natl. Bur. Stand.)* **37** (1971)
15. W. Gerrard: *Gas Solubilities*, Pergamon Press, Oxford 1980.
16. R. Fernandez-Prini, M. L. Japas, F. Corti: *High Temperature Aqueous Solutions*, CRC Press, Orlando 1992.

17. International Union of Pure and Applied Chemistry (ed.): *Solubility Data Series*, Pergamon Press, Oxford.
18. F. Franks (ed.): *Water. A Comprehensive Treatise*, vol. 4: “Aqueous Solutions of Amphiphiles and Macromolecules,” Plenum Press, New York 1975.
19. C. Tanford: *The Hydrophobic Effect*, Wiley, New York 1973.
20. W. Kauzmann, *Adv. Protein Chem.* **14** (1959) 1. Links
21. H. S. Frank, M. W. Evans, *J. Chem. Phys.* **13** (1945) 507. Links
22. F. Franks (ed.): *Water. A Comprehensive Treatise*, vol. 2: “Water in Crystalline Hydrates and Aqueous Solutions of Simple Nonelectrolytes,” Plenum Press, New York 1973.
23. J. S. Rowlinson, F. L. Swinton: *Liquids and Liquid Mixtures*, Butterworths, London 1984.
24. R. A. Robinson, R. H. Stokes: *Electrolyte Solutons*, Butterworths, London 1969.
25. J. E. Enderby, G. W. Neilson in F. Franks (ed.): *Water. A Comprehensive Treatise*, vol. 6: “Recent Advances,” Plenum Press, New York 1979, chap. 1, p. 1.
26. G. Engel, H. G. Hertz, *Ber. Bunsen Ges. Phys. Chem.* **72** (1968) 808.
27. W. A. Luck, *Ber. Bunsen Ges. Phys. Chem.* **69** (1965) 69.
28. K. S. Pitzer in R. M. Pytkowicz (ed.): *Activity Coefficients in Electrolyte Solutions*, CRC Press, Boca Raton 1979.
29. C. W. F. T. Pistorius, E. Rapoport, J. B. Clark, *J. Chem. Phys.* **48** (1968) 5509 – 5514. Links
30. T. B. Bizhigitov, N. N. Sirota, *JETP Lett. (Engl. Trans.)* **44** (1986) 417 – 419.
31. N. N. Sirota, T. B. Bizhigitov, *Russ. J. Phys. Chem. (Engl. Transl.)* **62** (1988) 134 – 136.
32. Landolt-Börnstein, **VI.4**, pp. 16 – 18.
33. K. Tödheide, *Ber. Bunsen. Ges. Phys. Chem.* **70** (1966) 1022.
34. F. Franks: *Water. A Comprehensive Treatise*, vol. 1, Plenum Press, New York 1972.
35. N. E. Dorsey: *Properties of Ordinary Water-Substance*, Hafner Publ.Co., New York 1940.
36. O. V. Nagornov, V. E. Chizov, *PMTF Zh. Prikl. Mekh. Tekh. Fiz.* 1990, no. 3, 41 – 48.
37. Landolt-Börnstein, **V.1b**, pp. 482 – 507.
38. P. V. Hobbs: *Ice Physics*, Clarendon Press, Oxford 1974.
39. J. V. Leyendekkers, R. J. Hunter, *J. Chem. Phys.* **82** (1985) 1440 – 1446. Links
40. B. Bullemer *et al.*, *Solid State Commun.* **6** (1968) 663. Links
41. A. Saul, W. Wagner, *J. Phys. Chem. Ref. Data* **18** (1989) 1537 – 1564. Links
42. D. Eisenberg, W. Kauzmann: *The Structure and Properties of Water*, Clarendon Press, Oxford 1969.
43. M. Christoforakos, E. U. Franck, *Ber. Bunsen Ges. Phys. Chem.* **90** (1986) 780 – 789.
44. U. Grigull, J. Straub, P. Schiebener: *Steam Tables in SI-Units*, 3rd ed., Springer Verlag, Berlin 1990.
45. U. Grigull: *Properties of Water and Steam in SI-Units*, 4th ed., Springer Verlag, Berlin 1989.
46. L. Haar, J. S. Gallagher, G. S. Kell: “NBS/NRC Steam Tables: Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units,” *Nat. Stand. Ref. Data (U.S. Natl. Bur. Stand.)* (1984).
47. R. A. Fine, F. J. Millero, *J. Chem. Phys.* **59** (1973) 5529 – 5536. Links
48. F. J. Dietz, J. J. de Groot, E. U. Franck, *Ber. Bunsen Ges. Phys. Chem.* **85** (1981) 1005 – 1009.
49. R. Tufeu, B. Le Neindre, *Int. J. Thermophys.* **8** (1987) 283 – 292. Links
50. K. H. Dudziak, E. U. Franck, *Ber. Bunsen Ges. Phys. Chem.* **70** (1966) 1120 – 1128.
51. W. J. Lamb, G. A. Hoffmann, J. Jonas, *J. Chem. Phys.* **74** (1981) 6875 – 6880. Links
52. W. L. Marhsall, E. U. Franck, *J. Phys. Chem. Ref. Data* **10** (1981) 295 – 304.
53. M. Uematsu, E. U. Franck, *J. Phys. Chem. Ref. Data* **9** (1980) 1291. Links
54. G. Wiegand, E. U. Franck, *Ber. Bunsen Ges. Phys. Chem.* **98** (1994) 809 – 817.
55. J. H. Dymond, E. B. Smith: *The Virial Coefficients of Pure Gases and Mixtures. A Critical Compilation*, Oxford University Press, New York 1980.
56. E. U. Franck, *J. Chem. Thermodynamics* **19** (1987) 225.
57. R. W. Shaw *et al.*, *Chem. Eng. News* (1991), December 23, 26 – 39.
58. L. A. Hütter: *Wasser und Wasseruntersuchung*, Salle + Sauerländer, Frankfurt 1994.
59. H. Hein, W. Kunze: *Umweltanalytik mit Spektrometrie und Chromatographie. Von der Laborgestaltung bis zur Dateninterpretation*, VCH Verlagsgesellschaft, Weinheim, Germany 1994.
60. American Water Works Association (eds.): Standard Methods for the Examination of Water and Wastewater, Published jointly by American Public Health Association (APHA), American Water Works Association (AWWA), Water Pollution Control Federation (WPCF), Washington 1994.
61. H. P. Hesse, F. Malz: “Die parameterspezifische Probenahme in der chemischen Abwasseranalytik,” *Abwassertechnik* **2** (1991) 21 – 26.
62. H. Gudernatsch: “Die Probenahme als wesentlicher Bestandteil der Wasser- und Abwasseranalytik,” *Vom Wasser* **60** (1983) 95 – 105.
63. C. Schöneborn: “Die Praxis der BSB-Bestimmung nach DIN 38 409-Teil 51,” *Korresp. Abwasser* **36** (1989) 424, 427 – 431.
64. M. Hahn *et al.*: “Eine neue Meßanordnung zur schnellen coulometrischen Bestimmung von extrahierbaren organischen Halogenverbindungen (EOX),” *Z. Wasser Abwasser Forsch.* **24** (1991) 191 –194.
65. G. Schwedt: *Taschenatlas der Analytik*, Thieme Verlag, Stuttgart 1996.

66. G. Schwedt: *Analytische Chemie, Grundlagen, Methoden, Praxis*, Thieme Verlag, Stuttgart 1995.
67. H. Berndt, U. Harms, M. Sonneborn: "Multielement-Spurenanreicherung aus Wässern an Aktivkohle zur Probenvorbereitung für die Atomspektroskopie (Flammen-AAS, ICP/OES)," *Fresenius Z. Anal. Chem.* **322** (1985) 329 – 333.
68. M. Sperling, X. Yin, B. Welz: "Flow Injection On-line Separation and Preconcentration for Electrothermal Atomic Absorption Spectrometry," Part 1: Determination of Ultratrace Amounts of Cadmium, Copper, Lead and Nickel in Water Samples," *J. Anal. At. Spectrom.* **6** (1991) 293 – 300.
69. R. Herzog, F. Dietz: "ICP-Massenspektrometrie — Erfahrungen mit einer neuen Analysenmethode in der Wasseruntersuchung," *Vom Wasser* **73** (1989) 67 – 109.
70. M. Weidenauer, K. H. Lieser: "Bestimmung von Spurenelementen im Flußwasser mit Hilfe der Voltammetrie," *Fresenius Z. Anal. Chem.* **320** (1985) 550 – 555.
71. F. H. Frimmel, T. Gremm: "The Importance of Element Speciation in Water Analysis — a Plea for Further Investigations," *Fresenius J. Anal. Chem.* **350** (1994) 7 – 13. Links
72. K. Stein, G. Schwedt: "Speciation of Chromium in the Waste Water From a Tannery," *Fresenius J. Anal. Chem.* **350** (1994) 38 – 43. Links
73. M. Sperling, Yin Yuefeng, B. Welz: "Differential Determination of Arsenic(III) and Total Arsenic Using Flow Injection On-line Separation and Preconcentration for Graphite Furnace Atomic Absorption Spectrometry," *Spectrochim. Acta Part B* **46 B** (1991) 1798 – 1801. H. H. Dieter: "Vorkommen und Toxikologie von Arsen und seine gesundheitliche Bedeutung im Trinkwasser," in K. Aurand, U. Hässelbarth, H. Lange-Asschenfeldt, W. Steuer (eds.): Die Trinkwasserverordnung. Einführung und Erläuterungen für Wasserversorgungsunternehmen und Überwachungsbehörden, 3rd revised ed., Erich Schmidt Verlag, Berlin 1991, pp. 154 – 171. F. W. Pontius, K. G. Brown, C.-J. Chen: "Health Implications of Arsenic in Drinking Water," *J. Am. Water Works Assoc.* **86** (1994) no. 9, 52.
74. R. D. Wilken, J. Kuballa, E. Jantzen: "Organotins: Their Analysis and Assessment in the Elbe River System, Northern Germany," *Fresenius J. Anal. Chem.* **350** (1994) 77 – 84. Links
75. J. A. C. Broekaert, S. Gücer, F. Adams (eds.): "Metal Speciation in the Environment," *NATO ASI Ser. G* **23** (1990)
76. P. C. do Nascimento, G. Schwedt: "Comparative Studies of the Determination of Cyanide at Low Concentration Levels in Waste Waters," *Anal. Chim. Acta* **283** (1993) 755 – 761. Links
77. S. Hinkamp, G. Schwedt: "Online-Verfahren zur Bestimmung von löslichem Gesamt-Stickstoff in Abwässern durch Kopplung eines UV-Aufschlusses mit der Fließinjektionsanalyse," *Z. Wasser Abwasser Forsch.* **24** (1991) 60 – 65.
78. S. Hinkamp, G. Schwedt: "Determination of Total Phosphorus in Waters With Amperometric Detection by Coupling of Flow-injection Analysis With Continuous Microwave Oven Digestion," *Anal. Chim. Acta* **236** (1990) 345 – 350. Links
79. B. Rössner, J. Behnert, A. Kiplinger: "Die Ionenchromatographie in der routinemäßigen Wasser- und Umweltanalytik," *Fresenius Z. Anal. Chem.* **327** (1987) 698 – 700.
80. G. Schwedt, B. Kondratjonok: "Simultane Ionen-Chromatographie von Kationen und Anionen mit Komplexonen als Elutionsmittel und selektiver UV-Detektion," *Fresenius Z. Anal. Chem.* **332** (1989) 855 – 861.
81. G. Schwedt, B. Rössner: "Ionen-chromatographische Spurenanalyse amperometrisch detektierbarer Anionen in Wässern," *Fresenius Z. Anal. Chem.* **327** (1987) 499 – 502.
82. J. Harms, L. Dunemann, G. Schwedt: "Comparison Between Capillary Electrophoresis and IonChromatography for Analysis of Inorganic Anions," *Mikrochim. Acta* **108** (1992) 251 – 260. Links
83. A. Meyer, A. Kelling: "Schnelltestverfahren im Vorfeld der Umweltanalytik," *CLB Chem. Labor. Betr.* **45** (1994) 414 – 417.
84. G. Schwedt: *Chromatographische Trennmethoden*, Thieme Verlag, Stuttgart 1993.
85. R. Götz: "Zur spektralphotometrischen Bestimmung von Ligninsulfonsäuren und Huminsäuren in Wässern," *Fresenius Z. Anal. Chem.* **296** (1979) 406 –407.
86. G. Abbt-Braun, F. H. Frimmel: "Spektroskopische Strukturaufklärung aquatischer Huminstoffe," *Vom Wasser* **77** (1991) 292.
87. H. Hellmann: "Aufgabe und Einsatz der IR-Spektroskopie im Rahmen der Tensidanalytik in Abwasser und Gewässern," *Vom Wasser* **66** (1986) 111 – 125.
88. P. Belouschek *et al.*: "Bestimmung von chlorierten KW mit kombinierter Headspace- und GC/MS-Technik," *Vom Wasser* **79** (1992) 1 – 8.
89. H.-J. Neu, W. Ziemer, W. Merz: "Ein neuartiges Derivatisierungsverfahren zur Bestimmung von Spuren polarer organischer Substanzen in Wasser mittels statischer Dampfraum-Gaschromatographie," *Fresenius J. Anal. Chem.* **340** (1991) 65 – 70. Links
90. C. L. Arthur *et al.*: "Environmental Analysis of Organic Compounds in Water Using Solid Phase Microextraction," *J. High Resolut. Chromatogr.* **15** (1992) 741 – 744.
91. C. L. Arthur *et al.*: "Solid Phase Microextraction for the Direct Analysis of Water: Theory and Practice," *LC-GC* **10** (1992) 656, 658, 660 – 661.
92. H.-J. Brauch: "Gas Chromatography for Determination of Pesticides in Aquatic Systems — Analytical Basis," *Acta Hydrochim. Hydrobiol.* **21** (1993) 84 – 88. Links
93. J. Brodesser, H. F. Schöler: "Die Analyse von polychlorierten Biphenylen in Wasseer im ng/l-Bereich," *Vom Wasser* **72** (1989) 145 – 150.
94. L. Braunstein, K. Hochmüller, K. Spengler: "Analyse gaschromatographierbarer Wasserinhaltsstoffe durch Kopplung von GC und massenselektivem Detektor," *Vom Wasser* **73** (1989) 167 – 189.
95. H. F. Schröder: "Hochdruck-Flüssigkeit-Chromatographie gekoppelt mit Tandemmassenspektrometrie — eine schnelle und zukunftsweisende Analysenmethode in der Wasser- und Abwasseranalytik," *Vom Wasser* **73** (1989) 111 – 136.
96. R. Reupert, E. Plöger: "Bestimmung stickstoffhaltiger Pflanzenbehandlungsmittel im Trink-, Grund- und Oberflächenwasser, Analytik und Ergebnisse," *Vom Wasser* **72** (1989) 211 – 233.

97. T. Gremm, F. H. Frimmel: “Systematische Untersuchung der PAK-Bestimmung mittels HPLC,” *Vom Wasser* **75** (1990) 171 – 182.
98. U. Obst, A. Holzapfel-Pschorn: *Enzymatische Tests für die Wasseranalytik*, Oldenbourg Verlag, München 1990.
99. B. Hock: “Immunchemische Verfahren in der Umweltanalytik und ihre Perspektiven,” *Z. Umweltchem. Ökotox.* **5** (1993) 309 – 319.
100. P. Herzsprung, L. Weil, K. E. Quentin, I. Zombola: “Bestimmung von Phosphorpestiziden und insektiziden Carbamaten mittels Cholinesterasehemmung,” *Vom Wasser* **74** (1990) 339 – 350.
101. K. Stein, G. Schwedt: “Einsatz eines Acetylcholinesterase-Biosensors aus pH-Elektrode und Enzymmembran zum Pestizid-Screening,” *Vom Wasser* **79** (1992) 211 – 224.
102. R. Kanne: “Aussagekraft ökotoxikologischer Testverfahren zur Beurteilung von Abwässern,” *GWF Gas — Wasserfach: Wasser/Abwasser* **129** (1988) 394 – 397.
103. P. MacCarthy, R. W. Klusman, S. W. Cowling, J. A. Rice: “Water Analysis,” *Anal. Chem.* **65** (1993) 244R – 292R. Links
104. A. Baumgartner, E. Reichel: *Die Weltwasserbilanz*, R. Oldenbourg Verlag, München – Wien 1975.
105. Umweltbundesamt (eds.): *Daten zur Umwelt 1992/93*, Erich Schmidt Verlag, Berlin 1994.
106. H. Bretschneider, K. Lecher, M. Schmidt (eds.): *Taschenbuch der Wasserwirtschaft*, Paul Parey, Hamburg – Berlin 1982.
107. U. de Haar (ed.): *Hydrologischer Atlas der Bundesrepublik Deutschland im Auftrag der Deutschen Forschungsgemeinschaft* unter der Gesamtleitung von R. Keller, Boldt-Verlag, Boppard 1979.
108. Bundesverband der deutschen Gas- und Wasserwirtschaft e. V. (BGW) (eds.): *106. Wasserstatistik Bundesrepublik Deutschland, Berichtsjahr 1994*, Wirtschafts- und Verlagsgesellschaft Gas und Wasser, Bonn 1995.
109. Bundesverband der deutschen Gas- und Wasserwirtschaft e. V. (BGW) (eds.): *104. Wasserstatistik Bundesrepublik Deutschland, Berichtsjahr 1992*, Wirtschafts- und Verlagsgesellschaft Gas und Wasser, Bonn 1993.
110. R. Schleyer, H. Kerndorff: *Die Grundwasserqualität westdeutscher Trinkwasserressourcen*, VCH Verlagsgesellschaft, Weinheim, Germany 1992.
111. St. P. Johnson, G. Corcelle: “Water,” *The Environmental Policy of the European Communities*, Graham & Trotman, London 1989, pp. 25 – 108.
112. Council Directive of June 16, 1975 Concerning the Quality of Surface Water Intended for the Abstraction of Drinking Water in the Member States (75/440/EEC), *O. J.* no. L 294, July 15, 1975.
113. Deutscher Verein des Gas- und Wasserfaches e. V. (DVGW): “Eignung von Oberflächenwasser als Rohstoff für die Trinkwasserversorgung,” *DVGW-Arbeitsblatt* **W 151**, ZfGW-Verlag Frankfurt/Main, July 1975.
114. Deutscher Verein des Gas- und Wasserfaches e. V. (DVGW): “Eignung von Wasser aus Fließgewässern als Rohstoff für die Trinkwasserversorgung,” *DVGW-Merkblatt* **W 251**, DVGW, Eschborn, June 1994.
115. Council Directive of July 15, 1980 Concerning the Quality of Water Intended for Human Consumption (80/778/EEC). *O. J.* no. L 229, Aug. 30, 1980. Proposal for a Council Directive Concerning the Quality of Water Intended for Human Consumption (95/C 131/03). Presented by the Commission April 28, 1995. *O. J.* no. C 131, May 30, 1995.
116. Der Bundesminister für Jugend, Familie, Frauen und Gesundheit: Verordnung über Trinkwasser und über Wasser für Lebensmittelbetriebe (Trinkwasserverordnung — TrinkwV) vom 5.12.1990. BGBl. I, 2612 – 2629 (1990).
117. K. Aurand, U. Hässelbarth, H. Lange-Asschenfeldt, W. Steuer (eds.): *Die Trinkwasserverordnung; Einführung und Erläuterungen für Wasserversorgungsunternehmen und Überwachungsbehörden*, 3rd ed., Erich Schmidt Verlag, Berlin 1991.
118. Fachnormenausschuß Wasserwesen (FNW) im Deutschen Normenausschuß (DNA), DIN 2000: Zentrale Trinkwasserversorgung. Leitsätze für Anforderungen an Trinkwasser, Planung, Bau und Betrieb der Anlagen, November 1973, Beuth-Verlag, Berlin.
119. F. W. Pontius, *J. Am. Water Works Assoc. (J. AWWA)* **88** (1996) no. 3, 36 – 46.
120. S. D. Faust, O. M. Aly: *Chemistry of Water Treatment*, Butterworths, Boston 1983.
121. T. H. Y. Tebbutt: *Principles of Water Quality Control*, Pergamon Press, Oxford 1992.
122. Bundesminister des Inneren (eds.): *Künstliche Grundwasseranreicherung, Stand der Technik und des Wissens in der Bundesrepublik Deutschland*. Bearbeitet vom BMI-Fachausschuß “Wasserversorgung und Uferfiltrat”. Erich Schmidt Verlag, Berlin 1985.
123. H. Sontheimer: “Trinkwasser aus dem Rhein?” Bericht über ein Verbundforschungsvorhaben zur Sicherheit der Trinkwassergewinnung aus Rheinuferfiltrat bei Stoßbelastungen, Academia Verlag, Sankt Augustin 1991.
124. Council Directive of December 17, 1979 Concerning the Protection of Groundwater Against Pollution by Certain Dangerous Substances (80/86/EEC). *O. J.* no. L 20, Jan. 26, 1980.
125. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit: Gesetz zur Ordnung des Wasserhaushalts (Wasserhaushaltsgesetz — WHG) vom 23.9.1986. BGBl. I, 1529 – 1544, ber. S. 1654 (1986), 6. Novelle (18. März 1994).
126. J. Mallevialle, I. H. Suffet, U. S. Chan (eds.): *Influence and Removal of Organics in Drinking Water*, Lewis Publ., Boca Raton 1992.
127. Rehabilitation of the River Rhine. Proceedings of the International Conference on Rehabilitation of the River Rhine March 15 – 19, 1993, Arnhem, The Netherlands, *Water Sci. Technol.* **29** (1994) no. 3, 1 – 394.
128. H. Bendlin: *Reinstwasser von A bis Z*, VCH Verlagsgesellschaft, Weinheim, Germany 1995.
129. H.-D. Held: *Kühlwasser, Verfahren und Systeme der Aufbereitung, Behandlung und Kühlung von Süßwasser, Brackwasser, Meerwasser in der Industrie*, neubearbeitet von H. G. Schnell, 4th ed., Vulkan-Verlag, Essen 1995.
130. H. von Kienle, E. Bäder: *Aktivkohle und ihre industrielle Anwendung*, Enke Verlag, Stuttgart 1980.
131. H. Sontheimer, J. C. Crittenden, R. S. Summers: *Activated Carbon for Water Treatment*, 2nd ed., DVGW-Forschungsstelle, Karlsruhe 1988.

132. R. M. Clark, B. W. Lykins: *Granular Activated Carbon: Design, Operation and Cost*, Lewis Publ., Michigan 1989.
133. R. Kümmel, E. Worch: *Adsorption aus wäßrigen Lösungen*, Deutscher Verlag für Grundstoffindustrie, Leipzig 1990.
134. V. L. Snoeyink in: *Water Quality and Treatment: A Handbook of Community Water Supplies*, McGraw Hill, New York 1990.
135. I. Langmuir, *J. Am. Chem. Soc.* **40** (1918) 1361 –1403. Links
136. H. Freundlich, *Z. Phys. Chem.* **A 57** (1906) 385 – 470.
137. R. A. Dobbs, J. M. Cohen, *U.S. Environ. Prot. Agency*, Report EPA-600/8–80–023 (1980).
138. T. F. Speth, R. J. Miltner, *J. Am. Water Works Assoc.* **82** (1990) 72 – 75.
139. A. L. Myers, J. M. Prausnitz, *Am. Inst. Chem. Eng. J.* **11** (1965) no. 1, 121 – 127.
140. C. J. Radke, J. M. Prausnitz, *Am. Inst. Chem. Eng. J.* **18** (1972) no. 4, 761 – 768.
141. J. C. Crittenden *et al.*, *Environ. Sci. Technol.* **19** (1985) no. 11, 1037 – 1043.
142. J. C. Crittenden, P. Luft, D. W. Hand, *Water Res.* **19** (1985) no. 12, 1537 – 1548. Links
143. D. W. Hand *et al.*, *J. Am. Water Works Assoc.* **81** (1989) no. 1, 67 – 77.
144. D. R. Yonge, T. M. Keinath, *J. Water. Pollut. Control. Fed.* **58** (1986) no. 1, 77 – 81.
145. P. C. Singer, C.-Y. Yen in I. H. Suffet, M. J. McGuire (eds.): *Activated Carbon Adsorption of Organics from the Aqueous Phase*, vol. I, Ann Arbor Science, Ann Arbor, Michigan 1980.
146. J. Fettig, H. Sontheimer, *J. Environ. Eng. (N.Y.)* **113** (1987) no. 4, 780 – 794.
147. B. Haist-Gulde: "Zur Adsorption von Spurenverunreinigungen aus Oberflächenwässern," PhD Thesis, University Karlsruhe (1991).
148. H. Sontheimer *et al.*, *Vom Wasser* **75** (1990) 183 – 200.
149. B. Haist-Gulde *et al.*, *GWF Gas Wasserfach: Wasser/Abwasser* **132** (1991) no. 1, 8 – 15.
150. I. N. Najm, V. L. Snoeyink, Y. Richard, *J. Am. Water Works Assoc.* **83** (1991) no. 8, 57 – 63.
151. E. Worch, *Chem. Tech. (Leipzig)* **43** (1991) no. 3, 111 – 114.
152. E. Worch, *Chem. Tech. (Leipzig)* **43** (1991) no. 6, 221 – 224.
153. R. Ford *et al.*, *Proc. Annu. Conf. Am. Water Works Assoc*. June 18 – 22 (1989).
154. T. F. Speth, R. J. Miltner, *J. Am. Water Works Assoc.* **81** (1989) no. 4, 141 – 148.
155. R. S. Summers *et al.*: *Standardized Protocol for the Evaluation of GAC*, Amer. Water Works Assoc. Research Found., Denver, CO, 1992.
156. J. C. Crittenden, W. J. Weber, *J. Environ. Eng. Div. (Am. Soc. Civ. Eng.)* **104 (EE3)** (1978) 433 – 443.
157. J. C. Crittenden, W. J. Weber, *J. Environ. Eng. Div. (Am. Soc. Civ. Eng.)* **104 (EE6)** (1978) 1175 – 1195.
158. J. C. Crittenden, J. K. Berrigan, D. W. Hand, *J. Water. Pollut. Control. Fed.* **58** (1986) no. 4, 312 –319.
159. J. C. Crittenden *et al.*, *J. Environ. Eng. (N.Y.)* **113** (1987) no. 2, 243 – 259.
160. J. C. Crittenden *et al.*, *J. Am. Water Works Assoc.* **83** (1991) no. 1, 77 – 87.
161. L. Cummings, R. S. Summers, *J. Am. Water Works Assoc.* **86** (1994) 6.
162. R. Kunin: *Ion Exchange Resins*, J. Wiley & Sons, New York 1958.
163. S. B. Applebaum: *Demineralization by Ion Exchange*, Academic Press, New York – London 1968.
164. F. Helfferich: *Ion Exchangers*, Mc Graw-Hill, New York 1962.
165. K. Dorfner: *Ion Exchangers*, De Gruyter, Berlin 1991.
166. M. J. Slater: *The Principles of Ion Exchange Technology*, Butterworth-Heinemann Ltd., Oxford 1991.
167. B. A. Bolto, L. Pawlowski: *Wastewater Treatment by Ion Exchange*, E. & F. Spon, London – New York 1987.
168. T. Shedlowsky, D. A. MacInnes: "The First Dissociation Constant of HCO_3^- of Carbonic Acid, 0 to 38 °C from Conductivity Measurements," *J. Am. Chem. Soc.* **57** (1935) 1705.
169. H. S. Harned, S. R. Sholes: "The Ionization Constant of HCO_3^- from 0 to 50 °C," *J. Am. Chem. Soc.* **63** (1941) 2030.
170. H. Sontheimer, P. Spindler, U. Rohmann: *Wasserchemie für Ingenieure*, ZfGW-Verlag, Frankfurt/Main 1980.
171. S. H. Eberle: "A Correctness Test of the Computation of Chemical Speciation for the Main Constituents of Natural Waters," *Water Res.* **23** (1989) 1373. Links
172. W. Höll: *Der Einfluß von Ionenaustauschvorgängen auf das Gleichgewichtssystem der Kohlensäure im Wasser*, KfK-Report 4367, Kernforschungszentrum, Karlsruhe 1987.
173. J. Tillmans, O. Heublein: "Über die kohlensauren Kalk angreifende Kohlensäure natürlicher Wässer," *Gesund. Ing.* **35** (1912) 669.
174. Degrémont: *Water Treatment Handbook*, 6th ed., Lavoisier Publ., Paris 1991, p. 262.
175. D. H. Caldwell, W. B. Lawrence: "Water Softening and Conditioning Problems," *Ind. Eng. Chem.* **45** (1935) 535.
176. K. Dorfner: *Ion Exchangers*, De Gruyter, Berlin 1991, chap. 2.2.
177. R. Kunin: Amber-Hi-Lites **95** (1966); **96** (1966);**97** (1967).
178. *LEWATIT-LEWASORB-Handbook*, Bayer AG, Leverkusen 1972.
179. K. R. Gray, H. L. Crosby, US 2 656 245, 1953.
180. R. Kunin: "Regeneration of Carboxylic Cation Exchange Resins With Carbon Dioxide," *Ind. Eng. Chem. Prod. Res. Dev.* **2** (1963) 1.
181. C. Berger-Wittmar: "Regeneration schwach saurer Ionenaustauscher mit Kohlendioxid," *Vom Wasser* **47** (1976) 297.
182. in Ref. [176] , chap. 2.4.
183. D. A. Clifford, W. J. Weber: "Nitrate Removal From Water Supplies," EPA Report **600/8 77 015** (1977) R-31,

Cincinnatti, Ohio.

184. A. K. Sengupta, D. A. Clifford: “Chromate Ion Exchange Mechanism for Cooling Water,” *Ind. Eng. Chem. Fundam.* **25** (1988) 249.
185. D. A. Clifford: “Processes for Removal of Inorganic Contaminants from Water,” *Water Eng. Manage*. (1982) R-31.
186. G. A. Guter, US 4 479 877, 1980.
187. F. X. McGarvey, B. Bachs, S. M. Ziarkowski: “Removal of Nitrates from Natural Water Supplies,” paper presented at *American Chemical Society Meeting*, Dallas, Texas, 1989.
188. J. P. v. d. Hoek, A. Klapwijk: “Nitrate Removal From Ground Water,” *Water Res.* **21** (1987) 989.
189. CHRIST AG, Technical Information TI-A-203.
190. R. Kunin, B. Vassiliou: “New Dionization Techniques Based Upon Weak Electrolyte Ion Exchange Resins,” *Ind. Eng. Chem. Process Des. Dev.* **3** (1964) 404; Further Studies on the Weak Electrolyte Ion Exchange Resin Desalination Process (DESAL Process), *Desalination* **4** (1968) 38.
191. Kernforschungszentrum Karlsruhe, EP 81 109 498.6, 1984 (W. H. Höll, B. Kiehling).
192. W. H. Höll, B. Kiehling: “Regeneration eines Ionenaustauschermischbetts mit CO_2 zur Teilentsalzung von Trinkwasser,” *Vom Wasser* **59** (1982) 207.
193. W. H. Höll, J. Horst: “Simultane Elimination von Schwermetallen und Komplexbildnern aus Wässern mit Anionenaustauschern in Hydrogencarbonatform,” *Vom Wasser* **74** (1990) 51.
194. D. E. Weiss et al.: “The Sirotherm Demineralization Process — An Ion Exchange Process with Thermal Regeneration,” part I, *J. Inst. Eng. Aust.* **37** (1966) 193.
195. D. E. Weiss: “Pilot Plant Studies of Partial Demineralization of Brackish Waters by the ‘Sirotherm' Process,” *Ion Exch. Membr.* **1** (1972) 109.
196. L. Sigg, W. Stumm: *Aquatische Chemie*, 3rd ed., B. G. Teubner Verlag, Stuttgart 1994.
197. Committee Report, *J. Am. Water Works Assoc.* **71** (1979) 588 – 603.
198. K. J. Ives: *The Scientific Basis of Flocculation*, Sijthoff & Noordhoff, Alphen aan den Rijn 1978.
199. J. Bratby: *Coagulation and Flocculation*, Uplands Press, Croydon 1980.
200. M. Jekel in DVGW (ed.): “Wasseraufbereitung für Ingenieure,” 2nd ed., *DVGW-Schriftenr. Wasser* **206** (1987).
201. M. Jekel, R. Ließfeld (eds.): “Die Flockung in der Wasseraufbereitung,” *DVGW-Schriftenr. Wasser* **42** (1985)
202. C. F. Baes, R. E. Mesmer: *The Hydrolysis of Cations*, J. Wiley & Sons, New York 1976.
203. M. Jekel, *J. Water SRT-Aqua* **40** (1991) 18 – 24.
204. A. Amirtharajah, K. M. Mills, *J. Am. Water Works Assoc.* **74** (1982) 210 – 216.
205. J. K. Edzwald, *AICHE Symp. Ser. Water.* **75** (1979) 54 – 62.
206. B. A. Dempsey, R. M. Ganko, C. R. O'Melia, *J. Am. Water Works Assoc.* **76** (1984) 141 – 150.
207. DIN:Standard 19 622, Polyacrylamides for Water Treatment, Beuth Verlag, Berlin – Köln 1977.
208. J. Gregory, *J. Colloid Interface Sci.* **55** (1976) 35 – 44. Links
209. C. R. O'Melia in W. J. Weber, Jr. (ed.): *Physico-chemical Processes for Water Quality Control*, Wiley-Interscience, New York 1972.
210. D. S. Parker, W. J. Kaufman, D. Jenkins, *J. Sanit. Eng. Div. Am. Soc. Civ. Eng.* **98** (1972) 79 – 99.
211. R. Klute in [212] , 53 – 65.
212. A. Grohmann, H. H. Hahn, R. Klute: *Chemical Water and Wastewater Treatment*, G. Fischer Verlag, Stuttgart – New York 1985, pp. 113 – 131.
213. G. Smethurst: *Basic Water Treatment for Application World-wide*, Thomas Telford Ltd., London 1979.
214. C. R. Schulz, D. A. Okun, *J. Am. Water Works Assoc.* **75** (1983) 212 – 219.
215. Degrémont: *Water Treatment Handbook*, 5th ed., J. Wiley & Sons, New York 1979.
216. N. Tambo, Y. Watanabe, *Water Res.* **13** (1979) 429 –448. Links
217. P. Wölfel in [201], 173 – 190.
218. H. Bernhardt, H. Schell, *Z. Wasser Abwasser Forsch.* **12** (1979) 123 – 133.
219. A. J. Rees, D. J. Rodman, T. F. Zabel, *Aqua* **8** (1980) 170 – 182.
220. M. Jekel, *Ozone Sci. Eng.* **16** (1994) 55 – 66.
221. H. E. Hudson, E. G. Wagner, *Proc. Water Qual. Technol. Conf.* **2A-3** (1979)
222. H. H. Hahn, R. Klute, *Z. Wasser Abwasser Forsch.* **12** (1979) 111 – 119.
223. M. Jekel, R. Gimbel (eds.): “Optimal Dosing of Coagulants and Flocculants,” Proceedings of the IWSA-IAWG Joint Specialist Group on Coagulation, Flocculation, Filtration, Sedimentation and Flotation in Water and Wastewater Treatment Workshop, Jan. 12 – 13, 1994, Mülheim an der Ruhr, Germany. Berichte aus dem Rheinisch-Westfälischen Institut für Wasserchemie und Wassertechnologie GmbH, Institut an der Gerhard-Mercator-Universität Duisburg — Gesamthochschule Duisburg, vol. 10, 1994.
224. K. H. Schmidt, “Biologische Verfahren,” *DVGW-Schriftenr. Wasser* **206** (1983) 11.1 – 11.42.
225. Degrémont: *Water Treatment Handbook*, 5th ed., Halsted Press (Wiley and Sons), New York 1979.
226. N. J. D. Graham: *Slow Sand Filtration: Recent Developments in Water Treatment Technology*, Ellis Horwood, Chichester 1988.
227. S. Vigneswaran, R. Ben Aim: *Water, Wastewater, and Sludge Filtration*, CRC Press, Boca Raton, FL, 1989.
228. R. Gimbel: “Abscheidung von Trübstoffen aus Flüssigkeiten in Tiefenfiltern,” *Veröffentlichungen des Bereichs und des Lehrstuhls für Wasserchemie am Engler-Bunte-Institut der Universität Karlsruhe*, Heft 25, ZfGW-Verlag, Frankfurt 1984.

229. P. A. Schweitzer (ed.): *Handbook of Separation Techniques for Chemical Engineers*, McGraw-Hill, New York 1979.
230. L. Svarovsky: *Solid-Liquid Separation*, 2nd ed., Butterworths, London 1981.
231. Degrémont: *Water Treatment Handbook*, vols. 1 + 2, Lavoisier Publ., Paris 1991.
232. K. W. Roennefahrt: “Filtration,” *DVGW-Schriftenr.* **206** (1983) 6–1 – 6–34.
233. R. Gimbel: “Theoretical Approach to Deep Bed Filtration,” in: *Water, Wastewater, and Sludge Filtration*, CRC Press, Boca Raton, FL, 1989.
234. C. Tien: *Granular Filtration of Aerosols and Hydrosols*, Butterworth Publish., Boston 1989.
235. T. Iwasaki: “Some Notes on Sand Filtration,” *J. Am. Water Works Assoc.* **29** (1937) 1591 ff.
236. J. P. Herzig, D. M. Leclerc, P. Le Goff: “Flow of Suspension through Porous Media – Application to Deep Filtration,” *Ind. Eng. Chem.* **62** (1970) no. 5, 8 ff.
237. R. Rautenbach, R. Mellis, W. Dahm, J. St. Kollbach; “Membranverfahren zur Aufarbeitung organisch/anorganisch hochbelasteter Abwässer — Kombinationen mit anderen Verfahrensstufen wie Biologie und Eindampfung,” Berichte der 27. Essener Tagung für Wasser und Abfallwirtschaft 1994.
238. A. Gröschl: “Umkehrosmose organisch/wäßriger Systeme — Stofftransport in Membranen und Verfahrensentwicklung,” PhD Thesis, RWTH Aachen 1991.
239. G. Schneider: “Trennverhalten von Nanofiltrationsmembranen,” PhD Thesis, RWTH Aachen 1993.
240. E. N. Sieder, G. E. Tate: “Heat Transfer and Pressure Drop of Liquids in Tubes,” *Ind. Eng. Chem.* **28** (1936) 1429.
241. W. H. Linton, T. K. Sherwood, *Chem. Eng. Prog.* **46** (1950) 258. Links
242. M. Mulder: *Basic Principles of Membrane Technology*, Kluwer Academic Publ., Dordrecht, 1991.
243. G. Menges: *Werkstoffkunde Kunststoffe*, 3rd ed., Hanser Verlag, München, Wien 1990.
244. E. Staude: *Membranen und Membranprozesse*, VCH, Weinheim, Germany 1992.
245. R. R. Bhave: *Inorganic Membranes — Synthesis, Characteristics and Applications*, Van Nostrand Reinhold, New York 1991.
246. Fachgruppe Wasserchemie in der Gesellschaft Deutscher Chemiker: Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung. Insbesondere: Verfahren der Gruppe C: Physikalische und physikalisch-chemische Kenngrößen (DIN 38 404); Verfahren der Gruppe D: Anionen (DIN 38 405); Verfahren der Gruppe E: Kationen (DIN 38 406); Verfahren der Gruppe G: Gasförmige Bestandteile (DIN 38 408); Verfahren der Gruppe H: Summarische Wirkungs- und Stoffkenngrößen (DIN 38 409).
247. W. Liebig: “Grundlagen der biologischen Abwasserreinigung am Beispiel einer Anlage für hochbelastete Abwässer eines Chemiewerkes,” *ChED Chem. Exp. Didakt.* **1** (1975) no. 1, 239 – 246.
248. P. Schuler, R. Degner: *Kleines Handbuch über die photometrische CSB-Bestimmung und Analysen von Wasserinhaltsstoffen*, Wissenschaftl. Techn. Werkstätten GmbH.
249. Permasep Products, Engineering Manual 1982.
250. A. Grohmann: “pH-Wert und Calcitsättigung des Wassers,” in K. Aurand, U. Hässelbarth, H. Lange-Asschenfeldt, W. Steuer (eds.): *Die Trinkwasserverordnung*, 3rd ed., Erich Schmidt Verlag, Berlin 1991, pp. 338 – 348.
251. H. E. Hömig: *Seawater and Seawater Distillation*, Vulkan-Verlag, Esssen 1978.
252. DIN 50 930, Korrosion der Metalle, Beuth Verlag, Berlin, Feb. 1993.
253. Babcock: *Handbuch Wasser*, 6th ed., Vulkan-Verlag, Essen 1982.
254. E. Heymann: “Filtration”, in DVGW Deutscher Verein des Gas- und Wasserfaches e. V. (eds.): “DVGW-Fortbildungskurse Wasserversorgungstechnik für Ingenieure und Naturwissenschaftler, Kurs 6: Wasseraufbereitungstechnik für Ingenieure,” 3rd ed., *DVGW-Schriftenr. Wasser* **206** (1987) 5–1 – 5–21.
255. *Ullmann*, 4th ed., **11**, 581 – 586.
256. E. Müller: *Mechanische Trennverfahren*, vol. I, Grundzüge der Verfahrenstechnk, Otto Salle Verlag/Verlag Sauerländer, Frankfurt/Main 1980.
257. K. Marquardt: “Flocculation, Precipitation, Sedimentation and Floatation for Use as Pretreatment Stages for Brackish Water and Sea Water in Desalination Plants,” in: H. G. Heitmann (ed.): *Saline Water Processing*, VCH Verlagsgesellschaft, Weinheim 1990, pp. 121 – 133.
258. R. Rautenbach, R. Albrecht: *Membrantrennverfahren — Ultrafiltration und Umkehrosmose*, Otto Salle Verlag/Sauerländer Verlag, Frankfurt/Main 1981.
259. Gesetz zur Ordnung des Wasserhaushalts (Wasserhaushaltsgesetz — WHG) vom 23. September 1986. BGBl. I, p. 1529 ff., Berichtigung p. 1654.sowieAllgemeine Rahmen-Verwaltungsvorschrift über Mindestanforderungen an das Einleiten von Abwasser in Gewässer (Rahmen-Abwasser-VwV — R-Abw.Vwv) vom 8. September 1990. Gemeinsames Ministerialblatt, Nr. 25, vom 22.9.1989, pp. 518 –520. Geändert am 29. Oktober 1992, Gemeinsames Ministerialblatt, Nr. 42, November, 1992, pp. 1065 –1066.
260. R. Mellis: “Zur Optimierung einer biologischen Abwasserreinigungsanlage mittels Nanofiltration,” Dissertation, RWTH Aachen 1994.
261. R. Rautenbach: Umkehrosmose auf der Deponie Schönberg — Betriebsergebnisse 1992/93, Interner Bericht, RWTH Aachen 1994.
262. T. Peters: “Reinigung von Deponie-Sickerwasser mit druckgestufter Umkehrosmose in DT-Modultechnik und fraktionierter Reststoff-Ausschleu-sung,” Preprints zum Aachener Membran Kolloquium, Aachen 1993, GVC-VDI — Gesellschaft Verfahrenstechnik und Chemieingenieurwesen, Düsseldorf 1993.
263. R. Rautenbach, T. Linn: “Aufarbeitung von Konzentraten aus der Deponiesickerwasser-Behandlung mittels Nanofiltration,” *Entsorgungspraxis* **9** (1995) 44 – 48.

General References

264. J. M. Montgomery: *Water Treatment Principles and Design*, J. Wiley & Sons, New York 1985. F. M. M. Morel, J. G.

Hering: *Principles and Applications of Aquatic Chemistry*, Wiley-Interscience, New York 1993. C. H. Tate, K. F. Arnold: "Health and Aesthetic Aspects of Water Quality," *Water Quality and Treatment*, 4th ed., American Water Works Association (AWWA) 1990, 63 – 156.

Specific References

265. W. M. Latimer: *The Oxidation States of the Elements and their Potentials in Aqueous Solution*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1952.
266. A. J. Bard, R. Parsons, J. Jordan (eds.): *Standard Potentials in Aqueous Solution*, Marcel Dekker, New York 1985.
267. J. Hoigné, H. Bader: "Rate Constants of Reactions of Ozone with Organic and Inorganic Compounds in Water," *Water Res.* **17** (1983) 185.
268. J. C. Morris in S. D. Faust, J. V. Hunter (eds.): *Principles and Application of Water Chemistry*, J. Wiley and Sons, New York 1967.
269. G. C. White: *The Handbook of Chlorination*, 2nd ed., Van Nostrand Reinhold, New York 1986.
270. W. Glaze : Ozone in Drinking Water Treatment, *Environ. Sci. Technol.* **21** (1987) 224 – 230.
271. J. J. Rook: Formation of Haloforms During Chlorination of NaturalWaters, *Water Treat. Exam.* **23** (1974) 234 – 243.
272. W. H. Glaze: "Reaction Products of Ozone: A Review," *Environ. Health Perspect.* **69** (1986) 151 – 157. Links
273. American Water Works Association Research Foundation: *Drinking Water and Health in the Year 2000*, American Water Works Association (AWWA), Denver, CO, 1993.

[Top of Page]

Symbols and Abbreviations

AAS: atomic absorption spectrometry

AMD: automated multiple development

AOP: advanced oxidation process

AOX: adsorbable organic halogens

BAC: biologically activated carbon

BiAS: bismuth-active substance

***BV*:** throughput

***CUR*:** carbon use ratio

DBP: disinfection byproduct

DOC: dissolved organic carbon

DOCl: dissolved organic chlorine

DOX: dissolved organic halogens

EBCT: empty-bed contact time

ECD: electron capture detector

EDTA: ethylenediaminetetraacetic acid

ELISA: enzyme-linked immunosorbent assay

EOX: extractable organic halogens

FIA: flow injection analysis

FID: flame ionization detector

FM: feed side of the membrane

***G*:** mean shear gradient

GAC: granular activated carbon

***I*:** ionic strength

IAS: ideal adsorbed solution

IC: ion chromatography

ICP: inductively coupled plasma

K_b**:** ebullioscopic constant

K_f**:** cryoscopic constant

K_F**:** Freundlich constant

***l*:** length of the pipe

***L*:** volume of solution (adsorption)

$\dot{m}$**:** mass flow

***m* alkalinity:** methyl orange alkalinity

MBAS: methylene blue-active substance

MSD: mass spectrometric detector

***n*:** Freundlich exponent (Chap. Adsorption Processes in Water Treatment); impeller speed (Chap. Ion Exchange)

$\dot{n}$**:** molar flow per unit area

***N*:** particle concentration

NF: nanofiltration

NOM: natural organic matter

NTA: nitrilotriacetic acid

OES: optical atomic emission spectrometry

***P*:** power input of stirrer

***p* alkalinity:** phenolphthalein alkalinity

PAC: powdered activated carbon

PAH: polycyclic aromatic hydrocarbon

PCB: polychlorinated biphenyl

PFR: plug-flow reactor

PM: permeate side of the membrane

PND: phosphorus – nitrogen detector

POC: particulate (undissolved) organic carbon

POX: purgeable organic halogens

PSA: potentiometric stripping analysis

***q*:** concentration in the solid phase (adsorption)

***R*:** rejection

***Re*:** Reynolds number

RO: reverse osmosis

***Sc*:** Schmidt number

***Sh*:** Sherwood number

***SI*:** saturation index

TAC: titre alcalemetrique complet (total alkalinity)

TCD: thermoionic detector

TDS: total dissolved solids

THM: trihalomethane

TIC: total inorganic carbon

TOC: total organic carbon

TOCl: total organic chlorine

TOX: total organic halogens

UF: ultrafiltration

$\dot{V}$**:** volumetric flow rate

VOC: volatile organic carbon

VOX: volatile organic halogens

***w*:** concentration

***z*:** filter layer depth, membrane thickness

Greek symbols
α**:** collision efficiency factor

λ**:** filtration coefficient, filter load

σ**:** equivalent factor, volume

Φ**:** concentration of solids

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Waxes

Uwe Wolfmeier, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Chap. 1)

Hans Schmidt, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Chap. 2)

Franz-Leo Heinrichs, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Chap. 3)

Georg Michalczyk, DEA Mineralöl AG, Hamburg, Federal Republic of Germany (Chap. 4, Chap. 7 in part)

Wolfgang Payer, Hoechst Aktiengesellschaft, Werk Ruhrchemie, Oberhausen, Federal Republic of Germany (Chap. 5)

Wolfram Dietsche, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Sections 6.1.1, 6.1.2, and 6.1.4)

Klaus Boehlke, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Sections 6.1.1, 6.1.2, and 6.1.4)

Gerd Hohner, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Sections 6.1.3, 6.1.5, 6.2, and 6.3)

Josef Wildgruber, Hoechst Aktiengesellschaft, Werk Gersthofen, Augsburg, Federal Republic of Germany (Chap. 7)

1. Introduction 3
1.1. History of Waxes and Their Applications 3
1.2. Definition 4
1.3. Classification 4
1.4. Properties and Uses 6
1.5. World Market Volume for Waxes 7
1.6. International Wax Organizations 8
2. Recent Natural Waxes 8
2.1. Introduction 8
2.2. Vegetable Waxes 9
2.2.1. Carnaúba Wax 9
2.2.2. Candelilla Wax 12
2.2.3. Ouricury Wax 14
2.2.4. Sugarcane Wax 14
2.2.5. Retamo Wax 15
2.2.6. Jojoba Oil 16
2.2.7. Other Vegetable Waxes 17
2.3. Animal Waxes 17
2.3.1. Beeswax 17
2.3.2. Other Insect Waxes 20
2.3.3. Wool Wax 21
3. Montan Wax 21
3.1. Formation and Occurrence 21
3.2. Extraction 21
3.3. Properties and Composition 22
3.4. Refining and Derivatization 23
3.4.1. Deresinification 23
3.4.2. Bleaching 23
3.4.3. Derivatization 24
3.5. Uses and Economic Aspects 25
4. Petroleum Waxes 26
4.1. Introduction 26
4.2. Macrocrystalline Waxes (Paraffin Waxes) 26
4.2.1. Chemical Composition and General Properties 26
4.2.2. Division into Product Classes 27
4.2.3. Occurrence of Raw Materials and Processing 30
4.2.3.1. Dewaxing Lubricating Oil Distillates 30
4.2.3.2. Deoiling Slack Waxes 31
4.2.3.3. Refining Deoiled Slack Waxes . . . 33
4.2.4. Storage, Transportation, Commercial Forms, and Producers 34
4.2.5. Quality Specifications and Analysis 35
4.2.6. Uses 37
4.3. Microcrystalline Waxes (Microwaxes) 38
4.3.1. Chemical Composition and General Properties 38
4.3.2. Division into Product Classes 38
4.3.3. Occurrence of Raw Materials and Processing 39
4.3.4. Commercial Products and Producers 41

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_103

4.3.5. Quality Specifications and Analysis 41
4.3.6. Uses 42
4.4. Legal Aspects 43
4.5. Ecology 43
4.6. Economic Aspects 44
4.7. Candles 44
5. Fischer – Tropsch Paraffins 45
6. Polyolefin Waxes 46
6.1. Production and Properties 46
6.1.1. Polyethylene Waxes by High-Pressure Polymerization . . . 46
6.1.1.1. Production 47
6.1.1.2. Properties 49
6.1.2. Copolymeric Polyethylene Waxes by High-Pressure Polymerization . 49
6.1.3. Polyolefin Waxes by Ziegler – Natta Polymerization 50
6.1.3.1. Production 51
6.1.3.2. Properties 51
6.1.4. Degradation Polyolefin Waxes . . . 52
6.1.4.1. Production 52
6.1.4.2. Properties 54
6.1.5. Polar Polyolefin Waxes 54
6.2. Uses 55
6.2.1. Nonpolar Polyolefin Waxes 55
6.2.2. Polar Polyolefin Waxes 56
6.3. Economic Importance 56
7. Toxicology 57
8. References 59

1. Introduction

Waxes are among the oldest worked materials used by humans. Their value as versatile construction materials ("man's first plastic") was discovered very early. Today, waxes are used mostly as additives and active substances. The use of waxes is expected to increase in the future because of their generally favorable toxicological and ecological properties.

The historic prototype of all waxes is beeswax. Since it could be obtained with relatively little effort, it was popular in antiquity, and even now the term wax is occasionally used in everyday speech as a synonym for beeswax. However, the scientific and commercial definition of wax covers a much wider area.

1.1. History of Waxes and Their Applications

Utilization of waxes was probably begun in prehistoric times, but because of their transitory nature, no definite archaeological evidence exists. Thus, the utilization of wax and related substances in mummification and as protective coverings in ancient Egypt (from ca. 3000 B.C.) represents the earliest scientific proof of the use of waxes. Ancient written documents contain many indications that waxes found many different applications. The most well known is the story of Daedalus and Icarus, who used wax as an adhesive to make wings by attaching feathers to each other. In antiquity, wax was used as a raw material for modeling, in the production of casting molds, as a pigment carrier, and for surface protection.

In colonial times, hitherto unknown waxes, such as carnaúba, candelilla, and Chinese insect wax, were introduced in Europe. From the occurrence of wax, COLUMBUS inferred the riches of the Caribbean islands: "Where there is wax, there are also thousands of other things" [1].

For a long time, not much was known about the chemical nature of wax. Only in the 18th century did the discovery occur that beeswax is an animal secretion and not a plant product collected by bees. Wax research was established as a scientific discipline in 1823 [2]. It became part of the new research area of soaps, oils, fats, and waxes. The real breakthrough of wax as an important raw material, in terms of quantity as well, occurred at the beginning of the Industrial Revolution. Ozocerite (fossil wax) was mined and refined to give ceresin (1875), montan wax was obtained from Eocene lignite (1897), and paraffin waxes were obtained from crude petroleum. In 1935 the first fully synthetic waxes were produced by the Fischer – Tropsch process. Polyethylene wax has been synthesized by the high pressure process since 1939, and became available by the low-pressure Ziegler process after 1953. On a laboratory scale polyolefin waxes can also be synthesized by using modern metallocene catalysts.

Degradation processes for the production of waxes, which start from high molar mass plastics (mainly polypropylene), have also achieved a certain degree of importance. Finally, low molar

mass substances, which would otherwise have to be disposed of in plastics production, can be processed or refined to give industrially utilizable waxes.

A large and still increasing number of natural and synthetic waxes and related substances exists, as well as applications for these materials. No decrease in overall demand is apparent, so waxes are likely to remain important during the coming decades. The history of waxes has been described in great detail in [3].

1.2. Definition

No generally accepted definition exists for the term wax. All attempts to formulate a precise, comprehensive, and scientifically verifiable definition of wax must take the large number of waxlike products and the chemical complexity of individual types into account. Nevertheless, use of the term wax for different chemical species with common properties is still reasonable.

Typically waxes do not consist of a single chemical compound, but are often very complex mixtures. Being oligomers or polymers in many cases, the components differ in their molar mass, molar mass distribution, or in the degree of side-chain branching. Functional groups (e.g., carboxyl, alcohol, ester, keto, and amide groups) can be detected in waxes, sometimes several different groups.

The academic definition still quoted in chemistry text books – that waxes are esters of long-chain carboxylic acids with long-chain alcohols – is no longer useful. It applies fairly well only to some classical waxes, such as beeswax; others (e.g., petroleum waxes) do not fall in this category. Today, *physical* and *technical* definitions are preferred. Several attempts were made to differentiate between waxes and other classes of substance, particularly fats, resins, and high molar mass polymers, by using several criteria. These primarily *physical* definitions are to some extent arbitrary and are not generalled accepted. Waxes can also be classified according to their *applications*.

Probably the most conclusive definition has been drawn up in Europe by the Deutsche Gesellschaft für Fettwissenschaft (DGF, German Association for Fat Science). It was used in modified form in the customs tariff of the European Union [4], [5]. According to this definition, waxes must have:

1) A drop point (mp) >40 °C
2) Their melt viscosity must not exceed 10 000 mPa · s at 10 °C above the drop point
3) They should be polishable under slight pressure and have a strongly temperature-dependent consistency and solubility
4) At 20 °C they must be kneadable or hard to brittle, coarse to finely crystalline, transparent to opaque, but not glassy, or highly viscous or liquid
5) Above 40 °C they should melt without decomposition
6) Above the *mp* the viscosity should exhibit a strongly negative temperature dependence and the liquid should not tend to stringiness
7) Waxes should normally melt between ca. 50 and 90 °C (in exceptional cases up to 200 °C)
8) Waxes generally burn with a sooting flame after ignition
9) Waxes can form pastes or gels and are poor conductors of heat and electricity (i.e., they are thermal and electrical insulators).

In the case of higher molar mass polyolefin waxes, differentiating between a wax and a thermoplastic polymer is sometimes difficult. This differentiation is particularly important for applications where a food approval is essential (FDA, EC, MITI in Japan, etc.). In case of doubt, relevant national regulations must be consulted.

1.3. Classification

Waxes can be classified according to various criteria such as origin; chemical, physical, and engineering properties; or applications (see Section 1.2). The primary differentiation is usually made according to origin or occurrence and synthesis. A proposal for a classification is shown in Figure 1. Here, waxes are divided into two main groups: natural and synthetic. Natural waxes exhibit their wax character without chemical treatment, whereas synthetic waxes generally acquire their waxy nature in the course of synthesis. There are no clear borderlines.

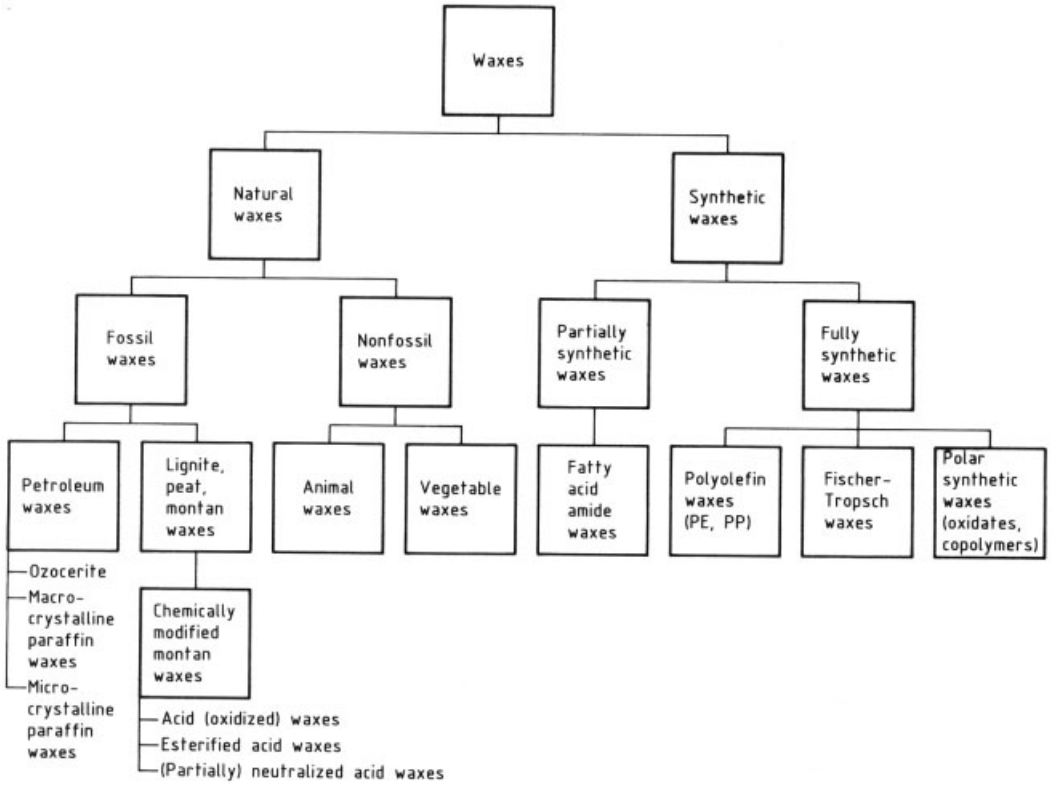

Figure 1. Classification of waxes

Natural waxes are formed through biochemical processes and are products of animal or plant metabolism. Natural waxes formed in earlier geological periods are known as *fossil waxes*. All petroleum, lignite (montan wax), and peat waxes belong to this largest category. Fossil waxes occur predominantly as minor components of oil, coal, and peat. Some deposits with a high wax content were formed by sedimentation. *Ozocerite* separated from oil, and *pyropissite*, one of the few organic minerals, is a mixture of montan wax and resins. Whereas petroleum waxes are predominantly nonpolar hydrocarbons, lignite and peat waxes consist of oxygen-containing compounds.

Biological wax synthesis is still taking place in nature. Many plants, particularly carnaúba palm, and animals – most importantly insects, such as bees – produce waxes. These are known as *nonfossil* or *recent natural waxes.* Natural waxes are seldom used industrially in their original form. They are generally converted into *modified natural waxes* by refining, e.g., by distillation, or extraction. Chemical processes, such as hydrogenation, bleaching, and oxidation, can also be used. Physical and chemical treatment can be combined; the objective of all these processes is to obtain a wax that is as pure as possible.

Partially Synthetic Waxes. If natural waxes or waxlike materials are modified by chemical reactions such as esterification, amidation, or neutralization of acidic waxes (with aqueous alkali or alkaline-earth metal hydroxides), partially synthetic waxes are obtained. These can be tailored to particular applications. In *esterified waxes* the relationship to natural waxes is still easily recognizable. For example, natural montan waxes, which consist mainly of esters of long-chain acids with long-chain alcohols, can be converted into acid waxes by cleavage of the esters and oxidation of the alcohols. Subsequent reaction with dihydric alcohols gives es-

ters that have chemical structures comparable to those of the starting materials. *Amide waxes* do not occur naturally to any significant extent. This group consists of reaction products of fatty acids with ammonia, amines, and diamines. An industrially important amide wax is distearoyl ethylenediamine.

Alcohol waxes (*lanette waxes*) also belong to the partially synthetic waxes. They are mixtures of long-chain alcohols that are optimized for their main area of use, in ointments and creams for pharmaceuticals and cosmetics, by the addition of fatty acid esters and emulsifiers. Alcohol waxes can also be used for industrial emulsions with good long-term stability.

Wool wax (*lanolin*) is obtained in crude form from sheep's wool. It is refined by chemical and adsorptive bleaching, and sometimes by subsequent hydrogenation or separation of the acidic fractions after hydrolysis. Wool wax is a mixture of esters, some being derived from branched or cyclic waxy or fatty acids and alcohols.

Fully synthetic waxes were developed only in the 20th century. Low molar mass compounds are used as starting materials. The products can be waxes in the narrower sense of the definition (see Section 1.2), or substances with only partial wax character. The two main groups of fully synthetic waxes are the Fischer – Tropsch and the polyolefin waxes. These waxes can be classified according to the starting material used for production (C_1 or C_2 chemistry).

C_1 Building Blocks. The Fischer – Tropsch (FT) synthesis is described elsewhere in detail (→ Coal Liquefaction, Chap. 2.2.). The oldest FT variant is the ARGE process (Arbeitsgemeinschaft Ruhrchemie – Lurgi). Synthesis gas is obtained by reaction of steam with carbon (→ Gas Production, Chap. 4.) and is then converted catalytically into a broad spectrum of saturated and unsaturated hydrocarbons. Among these products are the Fischer – Tropsch waxes. They are characterized by low branching of the hydrocarbon chain and a narrow molar mass distribution. They are relatively hard products. The ARGE process has been used commercially by SASOL in the Republic of South Africa.

Besides carbon, natural gas can also be used as a raw material for C_1 syntheses. The Shell Middle Distillate Synthesis (SMDS) process utilizes synthesis gas produced by steam reforming of natural gas [6]. Fischer – Tropsch wax fractions are produced and marketed at the new SMDS plant in Bintulu (Sarawak/Malaysia). This process is a variant of the classical FT process and employs specially developed catalysts.

Polyolefin waxes are, in most cases, products of C_2 or C_3 chemistry. Low molar mass α-olefins, usually ethylene, are polymerized. The other route to polyethylene waxes – thermomechanical degradation of polyethylene plastic – is now of minor importance. The two most important production processes are the high-pressure and the Ziegler processes. More recently, polyolefin waxes have also been synthesized by using modern metallocene catalysts.

Polyolefin waxes can be homopolymers or copolymers of C_2, C_3, C_4 (ethylene, propene, and butene), or higher α-olefins. Polypropylene (PP) waxes are produced by the Ziegler process or by thermomechanical degradation of polypropylene plastic. Polypropylene waxes are generally partially crystalline. In addition, predominantly amorphous and atactic low molar mass polypropylene (APO or APAO) with a waxlike character is now produced and marketed by several companies. Besides this deliberate production, atactic PP (APP) is still produced as a byproduct of PP plastics.

All FT and polyolefin waxes are nonpolar. Polar waxes are obtained by subsequent reactions, such as oxidation by air or grafting (e.g., with maleic anhydride), optionally followed by neutralization to metallic soaps, esterification, or amide formation. Polar polyolefin waxes are also accessible via direct copolymerization or addition reactions (e.g., of olefins with unsaturated carboxylic acids or esters).

A considerable proportion of industrially processed waxes is used not in pure form but as wax compounds. For example, the viscosity and *mp* of paraffin waxes can be increased to the desired degree by mixing with polyolefin waxes.

1.4. Properties and Uses

The number of possible and actual uses of waxes is extraordinarily high. The physical and chemical properties of various types of wax can be combined fairly freely through mixtures and preparations. Therefore, waxes are represented

in almost all possible areas of use, although to differing extents. Any selection of applications thus has to be somewhat arbitrary.

For the user, the application-oriented function of the wax is important, not so much its chemical structure. The latter is defined primarily by the characteristic hydrocarbon chain, the basic building block of almost all waxes. It imparts hydrophobing properties to waxes. The carbon chain length of fatlike waxes is $n = 16 - 18$ (e.g., amide wax); for esterified and paraffin waxes, $n = 20 - 60$; and for polyethylene waxes, less than hundred to several thousand. Besides the chain length itself (and thus the molar mass), molar mass distribution and degree of branching also affect the properties of waxes. The melting and softening points, melt viscosity, and degree of hardness increase with increasing chain length. High degrees of branching lower the *mp* and the hardness. Polar groups, such as carboxyl, ester, and amide groups, strengthen intermolecular forces and generally raise the *mp* and hardness.

Particular features of many waxes are their good absorption capacities and their ability to bind solvents. Hot wax solutions form stable, homogeneous pastes on cooling. These can be applied to surfaces and polished after evaporation of the solvent. The resulting wax films are shiny or glossy, hard, and resistant to mechanical stress.

When a suitable emulsifier is added, *polar waxes* can form fine stable dispersions in water. Mixed aqueous – organic emulsions can also be produced. These mostly stable dispersions and emulsions are used for surface protection. After application, the wax-containing phase dries and a protective, sometimes even glossy, coating is formed.

In addition to surface protection, hydrophobing, and gloss-giving, other functions of waxes include surface protection, hydrophobing, gloss-giving, matting, mold releasing, lubricating, binding, imparting compatibility or flexibility, regulating viscosity, adjusting consistency, emulsifying, dispersing, and adjusting drop points. Waxes are also used as combustible and illuminating materials.

For special applications only a few – frequently only one – of these functions are important. Some occasionally opposite effects are noteworthy: wax gives polish and gloss to various surfaces but has a matting effect as an additive in paints and coatings. Waxes are used as binders in master batches but as release agents in molding plastics. Table 1 gives some idea of the versatility of waxes but does not claim to be complete. For the specific suitability of waxes to individual applications, monographs, company brochures, and prospectuses should be consulted.

Table 1. Branches of industry in which waxes are used

Branch	Examples of applications
Adhesives, hot melts	viscosity regulation, lubricants, surface hardening
Building	modification of bitumen, antigraffiti treatment
Candles	fuel, drop point regulation
Ceramics and metal	binders for sintering
Cosmetics	binders and consistency regulators for ointments, pastes, creams, lipsticks
Electrical and electronics industries	release agents, insulating materials, etching bases
Explosives	stabilization
Foods	citrus fruit and cheese coating, chewing gum base, confectionery
Matches, pyrotechnics	impregnation, fuel
Medicine and pharmaceuticals	molding and release agents in dental laboratories, retardants, surface hardening of pills
Office equipment	dispersing agents and binders for carbon paper and selfduplicating paper; antioffset for toners for photocopiers
Paints and coatings	matting, surface protection
Paper and cardboard	surface hardening
Plastics	lubricants (PVC ∗), release agents (PA ∗∗), pigment carriers (masterbatch)
Polishes	surface protection of leather, floors, cars
Printing inks	improvement of rub resistance, slip
Recycling	compatibilizing
Rubber industry	release agents enhancing rigidity, surface hardening

∗ PVC = poly(vinyl choride).
∗∗ PA = polyamide.

1.5. World Market Volume for Waxes

No reliable global capacity or production data are available for waxes. The main reasons for this are the many types of wax and the difficulties in differentiating among them; the comparatively low transparency of the market; and the dependence of wax production on external factors, such as crude petroleum consumption and annual harvest of natural waxes. Some statistical

data exist for certain markets (e.g., the United States [7]). Thus, the following figures can be regarded only as rough indications, with a correspondingly large degree of fluctuation, of the magnitude of annual worldwide wax consumption data:

Paraffin waxes, including microcrystalline waxes	3 000 000 t/a
Polyolefin waxes including oxidates and copolymers	200 000 t/a
Fischer – Tropsch waxes	100 000 t/a
Amide wax	50 000 t/a
Montan wax including crude montan wax	25 000 t/a
Carnaúba wax	15 000 t/a
Beeswax	10 000 t/a
Candelilla and other plant waxes	3 000 t/a

Forecasts for future market development anticipate positive growth factors. Specialties, such as special waxes or compounds for critical applications (e.g., in the electrical, electronics, or ceramics industries), will probably profit more from this development than waxes, that are tend to be continuously replaced in their applications. The decreasing demand for waxes in the packaging sector, for example, is becoming noticeable. The proportion of waxes used in some other classical applications (e.g., carbon paper) is continually declining. However, waxes generally have favorable ecological and toxicological properties, in addition to the variety of applications described, so their market chances may outweigh the risks in the long term.

1.6. International Wax Organizations

In Europe the interests of all wax producers have been represented by the European Wax Federation since 1979 (EWF, Avenue van Nieuwenhuyse 4, Brussels, Belgium). The EWF is a member of the Conseil Européen des Federations de l'Industrie Chimique (CEFIC) and is open to all European wax producers and some outside Europe. It represents the interests of its members, with regard to both technical questions such as specifications and legal matters referring in particular to the European Union. In other regions, no organizations exist that cover the entire wax spectrum.

2. Recent Natural Waxes

2.1. Introduction

Many plant and animal organisms produce waxes with extraordinarily complex compositions. The main components are usually those that correspond to the classical definition of waxes – i.e., esters of long-chain aliphatic alcohols and acids, which belong to homologous series in the ca. $C_{16} - C_{36}$ (predominantly $C_{22} - C_{36}$) range, those with an even number of carbon atoms predominating. Recent natural waxes also contain smaller quantities of esters derived from acids and alcohols with an uneven number of carbon atoms as well as bifunctional components such as diols, hydroxycarboxylic, and dicarboxylic acids. Natural waxes contain very small proportions of branched carboxylic acids, mainly those with ω-methyl branching. Aromatic acids, such as cinnamic, 4-hydroxycinnamic, and ferulic (4-hydroxy-3-methoxycinnamic) acids, have also been found in some waxes. Because of the presence of bifunctional components, natural waxes can contain not only the simple, classical wax esters, but also oligoesters with two or more ester functions and other functional groups in the molecule.

Free acids and alcohols are present in natural waxes in widely varying quantities. They are generally identical to those present as ester components.

Homologous *n*-alkanes in the ca. $C_{15} - C_{37}$ range, mainly those with an uneven number of carbon atoms, are also components of natural waxes in proportions of 1 – 50 %. Branched-chain and olefinic hydrocarbons are present in insignificant quantities. Glycerides (fats), phytosterols, terpenes, resins, long-chain carbonyl compounds, and flower pigments are also present in very small proportions. Because of the complex mixtures involved and the relatively simple analysis techniques of earlier years, the inconsistency of literature data on the composition of natural waxes is understandable.

Modern analytical instruments are required for complete analysis of a natural wax. High-temperature capillary GC, possibly coupled to a mass spectrometer, is a suitable the best analytical method.

Beeswax and carnaúba and candelilla waxes have a total annual market volume of ca. 28 000 t worldwide. These are economically the most important natural waxes [8].

Among the large number of other known waxes, ouricury, esparto, bamboo, and Japan wax (actually a special vegetable tallow) are of regional or local importance. Their availability is estimated to be 3000 – 4000 t/a [8], [9].

In all areas of application for waxes, recent natural waxes are now mostly used together with modified fossil and synthetic waxes. Applications include preparations for cleaning, polishing, and preserving floors, furniture, and car bodywork surfaces; the candle industry; craft work; production of pharmaceuticals, cosmetics, and confectionery (observing appropriate legal approval); metal and ceramic sintering technology; coating techniques; the paint and coating industries; and the production of carbon paper compounds. The latter, which was important in terms of quantity, has been replaced almost completely by new duplicating processes. Despite the market recession this has caused [10], an annual growth of ca. 4 % is expected in the U.S. market for natural and synthetic waxes [11].

Special, ion-free or low-ion natural wax raffinates have also been used in electronics.

In his standard book *The Chemistry and Technology of Waxes*, A. J. Warth included a comprehensive chapter on all known recent natural waxes. The second edition of the book was published in 1956 and is currently out of print [12].

2.2. Vegetable Waxes

In contact with the environment, plants form a compressed epidermis on their surfaces by depositing secondary layers of cellulose, except in the case of inflorescence [13]. This mechanical reinforcement is supplemented by deposition of cutin in the outer epidermal layers and an outer coating of cutin (cuticula) to regulate water content. As a result of climatic conditions, many plants in tropical regions also store waxes in cutinized membrane layers in the cuticula as an additional protection against evaporation of water. In many cases, waxes exude outward and form wax coatings, which can be up to several millimeters thick. Xerophytes, plants found in exceptionally arid regions, are however not the greatest wax producers [14]. Specific soil types, temperature, dry periods, and heavy rainfall, as well as temporal changes in these, are responsible for enhanced wax formation. Vegetable waxes can be classified according to their origin as deciduous and coniferous tree waxes, shrub and herb waxes, and grass and flower waxes, and within these groups into leaf, needle, stem, and root waxes; bark and skin waxes; and seed and fruit waxes.

2.2.1. Carnaúba Wax

The leaf wax of the carnaúba palm is by far the most important recent vegetable wax both economically and with regard to applications [12], [13], [15], [16]. The carnaúba palm is a fan palm found mainly in Brazil and given the botanical name *Copernicia cerifera* by v. Martius in 1819 in honor of the astronomer Copernicus. It was apostrophized as the life tree of Brazil by v. Humboldt.

History. The existence of carnaúba wax was first reported in 1648 [17], and some data concerning its composition were published in 1811 in *Transactions of the Royal Society, London*.

In 1836 the Brazilian Macedo arranged for chemical investigations to be carried out on carnaúba wax at the Sorbonne in Paris. He began small-scale extraction and purification of carnaúba wax in 1856. In 1862, 1280 t of carnaúba wax was exported from Brazil.

Occurrence and Isolation. The carnaúba palm occurs mainly in north eastern Brazil in the provinces of Bahia, Rio Grande do Norte, Paraiba, Pernambuco, and particularly Ceara and Piaui, where there is a total stand of ca. 90×10^6 palms in a surface area of 10^6 km^2 [18], [19]. Smaller areas exist in southern Brazil, northern Argentina, Paraguay, and Bolivia. Different climatic conditions in other tropical and subtropical countries, in which carnaúba palms were planted as an experiment, allow satisfactory growth but markedly reduced formation of wax deposits on the leaves.

The palms grow to a height of 6 – 12 m, occasionally to 20 m, and are said to achieve a lifetime of up to 200 years. When the palm has reached an age of ca. eight years, one leaf gives a wax yield of 5 – 7 g, which is worth harvesting. The annual yield of wax per palm is ca. 150 g.

To isolate the wax, ca. 20 – 30 leaves at different stages of development are removed from each palm twice a year (September and December) [12].

The unopened heart leaves, of which only a few are removed per palm, carry particularly light-colored wax (prime yellow). Wax on the outer palm leaves is yellow, gray-green, or gray-brown, depending on climatic conditions during the vegetation period and on leaf age.

After drying the leaves in the sun or on steam-heated racks, the wax is chipped off manually or nowadays mostly by machine. The primary purification process involves melting over water and gives a crude wax that can be purified further by centrifugation, filtration with additives, or solvent extraction. Additional bleaching with hydrogen peroxide is also carried out.

The central ribs are often removed from the leaves before drying. The dried leaves are then chopped and the chips separated by air classification to give a mixture of ca. 60 % wax and 40 % chopped plant, which is stirred with water and oxalic acid to form a paste. The latter is heated to its boiling point and pressed through a filter while hot. Centrifugation of the filtrate gives virtually anhydrous wax. The remaining wax is extracted from the dried filter cake with aliphatic hydrocarbons (e.g., heptane).

The crude product is normally supplied in lumps that are obtained by breaking up the wax blocks into which the melt was cast. Carnaúba wax is also sold now in other forms.

Types. Internationally recognized classifications and specifications for genuine, unfalsified types of carnaúba wax were worked out by the American Wax Importers and Refiners Association according to established methods of determination [20].

The classification used in Brazil, shown in Table 2, has been valid worldwide since the beginning of the 1980s.

The highest-value type, *T 1*, is a prime yellow wax purified by filtration. To obtain *T 3*, suitable crude waxes (colored) are filtered and then bleached with hydrogen peroxide. The *T 4* wax is exported predominantly in the form of centrifuged or filtered variants.

Chemically refined or derivatized carnaúba waxes are no longer available commercially because of lack of demand [8]. Chemical refining processes, such as heating with alkali in the presence of paraffins, gave carnaúba wax residues in addition to the corresponding raffinates [21–27]. Where these processes are used, they account for <1 % of the total volume of carnaúba wax.

Properties. Carnaúba wax is one of the hardest and highest-melting natural waxes. Because of its fine crystalline structure, its fracture edges are smooth and have a macroscopically homogeneous structure. At room temperature the wax has a weakly aromatic odor and a characteristic haylike scent (similar to coumarin) in the molten state [28]. Carnaúba wax is compatible with almost all natural and synthetic waxes and a number of natural and synthetic resins. It is readily soluble in most nonpolar solvents on warming [12] and is miscible with them in all proportions above its *mp*. On cooling, the wax precipitates from solution to form a solid paste. Carnaúba wax is only partially soluble in polar solvents, even on warming, and is generally sparingly soluble at room temperature (0.15 – 0.6 %) [12].

Melting, solidification, and drop points as well as hardness of other waxes increase on addition of small quantities of carnaúba wax. With paraffins, the addition of carnaúba wax suppresses their tendency toward coarse crystallinity and reduces possible tack in the case of soft waxes.

Composition. Many analyses of carnaúba wax has been carried out, with conflicting results [12], [14], [15]. According to a thorough investigation, carnaúba wax prime yellow (T 1) has the composition given below [29]:

Aliphatic esters	40.0 wt %
Diesters of 4-hydroxycinnamic acid	21.0 wt %
Esters of ω-hydroxycarboxylic acids	13.0 wt %
Free alcohols	12.0 wt %
Diesters of 4-methoxycinnamic acid	7.0 wt %
Free aliphatic acids	4.0 wt %
Free aromatic acids	1.0 wt %
Hydrocarbons (paraffins)	1.0 wt %
Free ω-hydroxycarboxylic acids	0.5 wt %
Triterpene diols	0.5 wt %
Unsaponifiable components	56.4 wt %
Saponifiable components	39.2 wt %
Aromatics and/or resins	4.4 wt %

The aliphatic esters contain monocarboxylic acids of average chain length C_{26} and mono-

Table 2. Carnaúba wax types and specifications

Characteristic	T 1 (prime yellow)	T 3 (fatty gray/light)	T 4 (fatty gray/crude)	T 4 (fatty gray/centrifuged)	T 4 (fatty gray/filtered)
Minimum *mp*, °C	83.0	82.5	82.5	82.5	82.5
Minimum flash point, °C	310	299	299	299	299
Moisture content, %	0	0	2.0 (max.)	1.0 (max.)	0.1 (max.)
Minimum acid number, mg KOH/g	2	4	4	4	4
Maximum acid number, mg KOH/g	6	10	10	10	10
Minimum saponification number, mg KOH/g	78	78	78	78	78
Maximum saponification number, mg KOH/g	88	88	88	88	88
Ester number, mg KOH/g	75 – 85	75 – 85	75 – 85	75 – 85	75 – 85
Maximum ash content, %	0.04	0.15	0.20	0.20	0.10
Solidification point, °C	80	79	79	79	79
Drop point, °C	84	83	83	83	83
Minimum particle size, mesh	95	95	95	95	95

hydric alcohols of average chain length C_{32}. The ω-hydroxyesters contained are mixtures of ca. 90 % esters of ω-hydroxyacids (C_{26}) and monohydric alcohols (C_{32}) and 10 % esters of monocarboxylic acids (C_{28}) and α,ω-diols (C_{30}).

Table 3. Chain-length distributions for straight-chain monocarboxylic acids and monohydric alcohols, and paraffins in carnaúba wax

Chain length (C-number)	Acids ∗, % (approx.) ∗∗	Alcohols ∗, % (approx.) ∗∗	Paraffins ∗, % (approx.) ∗∗
16	2.0		
18	3.0		
20	8.5		0.4
21			0.5
22	7.5	0.5	0.7
23			1.3
24	23.2	0.7	1.6
25			2.6
26	10.5	0.5	3.5
27			8.7
28	16.5	2.0	5.4
29			13.1
30	5.5	10.6	13.9
31			29.3
32	2.0	60.8	9.6
33			7.3
34	0.5	16.4	1.5
35			0.3
36		0.6	0.4

∗ Values based on 100 % total acid, total alcohols, and total paraffins in each case (deviations from 100 % = sum of all other acids and alcohols).
∗∗ Approximate.

Esters containing 4-hydroxy- and 4-methoxycinnamic acids are present mainly as oligomers and polymers. The monomer units of these are diesters of the above-named cinnamic acids with mono- and polyhydric alcohols and ω-hydroxycarboxylic acids.

The free alcohols are similar in composition to those in aliphatic esters. The wax also contains a small proportion of secondary alcohols.

The even-numbered *n*-monocarboxylic acids, *n*-monohydric alcohols, and *n*-alkanes present in all carnaúba wax types have typical, distinctive chain-length distributions (see Table 3).

The important difference between Type 1 carnaúba wax and the other types is that the latter contain no triterpenediols and the esters of 4-hydroxy- and 4-methoxycinnamic acids have higher degrees of polymerization compared with Type 1. This may be attributed partly to differences in climate but also to the fact that Type 1 is harvested from genetically younger leaves, whereas the others come from older leaves that have been exposed to more sunshine.

Uses.

T 4 Variants. The wide use of carnaúba wax in the production of *preserving and cleansing agents* for floors (floor polishes, ionogenic and nonionogenic day-bright emulsions, wiping waxes and cleaners, sealing emulsions, spray cleaners, etc.); and polishes for furniture, cars, and shoes (solvent – water or straight solvent products) is based on its ability to form pastes and glossy and repolishable films and its ability to be dispersed. In this classical area a large proportion of T 4 variants of carnaúba wax are used.

Type 1 carnaúba wax and some of Type 3 (after absorptive purification to remove peroxide components arising from hydrogen peroxide bleaching) are approved for use in the *pharmaceutical industry* (e.g., as polishes for pills) and the *cosmetics industry* (as lipsticks and lip salves).

In the *food sector*, T 1 and T 3 waxes are used as release agents for bakery and confectionery products and as additives in chewing gum production.

In *polymer processing*, carnaúba wax is used in release agent preparations and, to a small extent, as a lubricant. The *varnish industry* uses carnaúba wax as an additive in the interior coating of food containers [8].

In the *leather industry*, carnaúba wax is used for cutting waxes, cleaning waxes, and edging inks. It is also approved for coating citrus fruit.

Economic Aspects. About 10 000 – 16 000 t/a carnaúba wax is available worldwide, of which ca. one-fifth is Type 1. Brazil, the sole exporter, supplies mainly Europe (ca. 2000 t/a), the United States (3000 – 4000 t/a), and Japan [8]. About 3000 – 3500 t/a is for indigenous supply.

Carnaúba wax consumption in the United States decreased markedly from 1955 to 1976 and then remained constant [9]. Brazil is increasingly exporting purified or bleached products in flaked form to achieve higher prices. Harvests, exports, and price structuring are subject to government control.

Waxes with Similar Compositions. *Copernicia australis*, a palm that produces large quantities of leaf wax, is found in Paraguay and around its border with Argentina (Gran Chaco). It is similar to *Copernicia cerifera* in many respects. A few small factories for isolating this palm wax are said to have been erected since 1945 [30].

The Brazilian *cauacu bush*, found mainly in Pará province and by the upper Amazon, produces a wax that is said to be similar to carnaúba wax in chemical composition. However, the wax is isolated only sporadically and to a limited extent [31].

2.2.2. Candelilla Wax

Occurrence and Isolation. The climatic conditions of the semideserts in southern California, Arizona, south western Texas, northern Mexico, and parts of Central and South America favor enhanced wax formation in *Euphorbia* (*E. cerifera, E. antisyphilitica*) and *Pedilanthus* (*P. pavonis, P. aphyllus*) species. The wax deposited on the stalks and leaf stems of these plants, which grow as bushes or shrubs, is known as candelilla wax [12], [14], [32], [33].

Numerous and dense stands of these plants are found mainly in northern Mexico in the states of Chihuahua, Coahuila, Durango, Nuevo León, San Luis Potosi, Tamaulipas, and Zacatecas over a total area of more than 100 000 km^2. These are the only areas in which candelilla wax is obtained [34].

In manual harvesting the plants are pulled out, leaving most of the root system in the soil. Together with scattered seeds, roots remaining partially in the soil guarantee the growth of new stands, which takes several years.

In collecting and processing locations, plant material is boiled with ca. 0.2 % sulfuric acid in open vessels, which are heated using already dewaxed and dried stalks. Metal grids keep the plants under the surface of the liquid, and addition of sulfuric acid inhibits emulsification of the wax. The latter collects in molten form on the surface of the liquid. After the wax is skimmed off into open vats or barrels, it solidifies to form the crude wax known as cerote. Yields based on plant material are 3 – 4 %.

Wax from outer areas of the solidified wax blocks is removed with impurities that have concentrated there and is subjected to the boiling process again. The remainder of the wax is purified in a central refinery located in Saltillo by melting with sulfuric acid and filtering in filter presses. The crude wax broken down into lumps is the Mexican commercial product (Mexican Standard Grade Candelilla Wax).

Candelilla wax raffinates are obtained from crude wax by adsorptive purification or hydrogen peroxide bleaching. They are produced and sold in the United States and Europe. The properties of these products, apart from color, are identical to those of crude candelilla wax.

Following the transfer of Mexican candelilla wax production from state monopoly to private

enterprise, candelilla wax raffinates produced in Mexico will also be put on the market [8].

By deresinification and subsequent oxidative chromic acid bleaching, products that are almost white can be obtained. Because of the high price of genuine wax, these processes have no practical importance.

Properties and Composition. Candelilla wax is a hard, brittle wax that is very similar to carnaúba wax with regard to solubility in polar and nonpolar organic solvents. A large proportion of the resin present in the wax can be extracted at room temperature with hot 70 % ethanol, acetone, or low-boiling, chlorinated organic solvents.

The *physical and chemical properties*, the color (brown – brownish yellow – pale yellow), and the degree of purity vary depending on climate, time of harvest, region, and age of harvested plants. The recognized specification for genuine crude candelilla wax is given in the following [34].

mp	68.5 – 72.5 °C
Refractive index	1.4550 – 1.4611
d_4^{15}	0.950 – 0.990
Acid number	12 – 22 mg KOH/g
Saponification number	43 – 65 mg KOH/g
Hydrocarbon content	30.6 – 45.6 %
Total acid content	20.6 – 29.0 %
Color	amber
Flash point	235.4 – 248.4 °C
Components insoluble in xylene/toluene (1 : 1)	0.0 – 0.1 %

The data agree with those previously laid down by the American Wax Importers and Refiners Association for prime crude candelilla wax [20].

The average *composition* of candelilla wax gives below is derived from the large number of values found in the literature.

Hydrocarbons (ca. 98 % paraffins + 2 % alkenes)	42.0 wt %
Wax + resin + sitosteroyl esters	39.0 wt %
Lactones	6.0 wt %
Free wax and resin acids	8.0 wt %
Free wax and resin alcohols (terpene alcohols)	5.0 wt %
Saponifiable components	23.0 – 29.0 wt %
Unsaponifiable components	71.0 – 77.0 wt %

Candelilla wax differs fundamentally from carnaúba wax in its high hydrocarbon content of 45 % (max.) and resin content of 20 %. Whereas carnaúba wax contains >80 % wax according to the classical chemical definition, candelilla wax consists of <35 % of these components.

As in all natural waxes, in candelilla wax the paraffins and wax acids or alcohols with even-numbered carbon chains have typical and distinctive chain-length distributions (Table 4).

Table 4. Chain-length distribution of even-numbered carboxylic acids and alcohols, and paraffins in candelilla wax

Chain length (C-number)	Acids ∗, % (approx.) ∗∗	Alcohols ∗, % (approx.) ∗∗	Paraffins ∗, % (approx.) ∗∗
22	0.3		
23			
24	0.1		
25			
26	0.3	2.4	0.1
27			0.2
28	3.2	17.4	0.5
29			5.6
30	36.4	58.0	1.3
31			80.0
32	48.6	19.0	1.8
33			9.0
34	8.0		0.2

∗ Values based on resin-free total acid, total alcohol, and total paraffin (deviations from 100 % = sum of all other acids, alcohols, and hydrocarbons).
∗∗ Approximate.

Uses. Candelilla wax is used in classical *cleaning preparations and polishes* (e.g., as a component of shoe and other leather polishes) and in furniture, car, and floor polishes. Because of its particular composition, this wax can be used in the applications only in combination with other suitable natural and synthetic waxes. Its use in floor polish is very limited because the high resin content leads to increased dirt pickup.

Candelilla wax is also used in the production of *candles, coatings* for paper and cardboard, *hotmelt adhesives*, and in *polymer processing*.

It is also approved for use in the *cosmetics*, *pharmaceutical*, and *food industries*. The main use of candelilla wax, principally in the United States and Europe, is by the cosmetics industry in the production of lipstick. Here, properties such as outstanding framework formation with castor oil, gloss, and uniform shrinkage on cooling (very good demolding after casting lipsticks) are very important [8]. Candelilla wax is also

used as a polish for pills and in the production of chewing gum and confectionery.

Economic Aspects. For nomadic peasants in the areas of Mexico mentioned, proceeds from the candelilla harvest were an important part of their income for a long time. Because of continuing improvements in infrastructure, the uneconomical isolation of candelilla wax has decreased since 1950 and been replaced increasingly by agricultural production. Exports to the main consumer, the United States, fell from 2500 t in 1950 [14] to 1800 t in 1974 [33], 1100 t in 1980, and ca. 310 t in 1992 [35]. About 800 – 1000 t of candelilla wax is available on the world market, of which ca. 200 t is consumed in Europe [8].

In the United States, demand for candelilla wax is increasing because of the shift of use from classical areas to those of foods, cosmetics, and pharmaceuticals [35]. Privatization of candelilla wax isolation in Mexico and efforts to produce raffinates there will probably lead to greater availability of these waxes.

2.2.3. Ouricury Wax

Occurrence and Isolation. Ouricury wax is deposited on the leaf stems of feather palms (*Syagros coronata*), which grow mainly in the Brazilian state of Bahia. There it is also known as urucury, licuri, aricuri, nicuri, and coqueiro.

The isolation process for ouricury wax is very similar to that used for carnaúba wax, but it is made more difficult by strong adhesion of the wax to the plant, and mechanical scratching or scraping off is necessary. The annual yield per palm is ca. 1 kg of crude wax.

Preliminary purification involves melting crude wax in hot water and filtering the melt through filter cloths or suitable sieves [12], [14], [36]. Additional purification in filter presses is used to obtain *USA Pure Refined wax.* The crude wax is sold in lumps (Brazilian Crude).

Properties and Composition. Ouricury wax resembles carnaúba wax very closely with regard to its hardness, gloss, solubility in polar and nonpolar solvents, and hardening capacity when used in combination with soft waxes. Its color is somewhat darker than that of dark carnaúba wax types. The *mp* is 82.5 °C. The most important properties for genuine commercial products have been established [20]. Thorough investigation of ouricury wax gave the following composition [37]:

Simple aliphatic esters	23.5 wt %
Monoesters of hydroxycarboxylic acids	22.4 wt %
Diesters of hydroxycarboxylic acids	17.2 wt %
Acidic hydroxypolyesters	5.4 wt %
Free acids	8.7 wt %
Free alcohols	3.0 wt %
Resins	14.8 wt %
Water	1.4 wt %
Ash	0.4 wt %

Alkaline saponification of ouricury wax yields varying quantities of total acids (69 – 74 %) and unsaponifiable components (26 – 31 %), as usual in the case of natural waxes. Despite these facts, ouricury wax, like other vegetable waxes, can be characterized definitely by the chainlength distributions of its *n*-paraffin and acids and alcohol components.

Uses and Importance. Ouricury wax was previously used in floor and shoe polishes and in carbon paper coloring agents, instead of carnaúba wax, mainly because of its low price. As a result of increasing price, the importance of ouricury wax has decreased sharply. According to Brazilian statistics, 39 t was exported in 1980. In 1991 the consumption in the United States was ca. 4.5 t/a [38].

2.2.4. Sugarcane Wax

Occurrence and Isolation. Sugarcane wax is formed as a powdery, pale to dark yellow deposit on sugarcane stalks (particularly in the thickened nodes) in quantities of 0.1 – 0.25 % based on cane weight, depending on country of origin. The main sugarcane-growing countries are India (ca. 96×10^6 t/a), Brazil (ca. 80×10^6 t/a), and Cuba (ca. 40×10^6 t/a).

Sugarcane wax is generally extracted from the dried filter cake (ca. 1 % based on sugarcane) [12] formed during sugarcane processing by using alcohols, carbon disulfide, carbon tetrachloride, or aromatic and aliphatic hydrocarbons [39].

The crude wax content of the filter cake (in weight percent) varies between the following limits, depending on sugarcane origin and method of processing:

India	8.0 – 18.0 wt %
Louisiana	4.4 – 17.9 wt %
Puerto Rico	12.0 – 14.0 wt %
Hawaii	9.6 – 11.0 wt %
Philippines	10.5 – 11.0 wt %
Cuba	12.4 – 22.0 wt %
Java	5.0 – 15.0 wt %
Republic of South Africa	6.9 – 14.6 wt %
Argentina	11.9 – 15.5 wt %
Brazil	8.9 – 17.8 wt %

Crude wax can be freed from most glycerides (fats and oils) by treatment with acetone, methyl ethyl ketone, methanol, ethanol, 1-propanol and isopropanol, 1-butanol, or ethyl acetate. In a subsequent recrystallization step, preferably in isopropanol, resinous and colored components can be removed. Further treatment with bleaching earths or oxidative bleaching of the wax melt with air gives waxes ready for use in a variety of applications. Bleaching processes for prepurified wax using potassium chlorate – sulfuric acid or chromic acid – sulfuric acid would be effective but are not used for cost reasons.

Properties and Composition. *Crude sugarcane wax* is a black-brown to green-brown semisolid with an unpleasant, rancid odor.

Semi- and completely purified products are brown- to yellow-colored hard waxes, whose characteristics vary considerably. Melting points of 68 – 81 °C, acid numbers of 7 – 22 mg KOH/g saponification numbers of 32 – 65 mg KOH/g and proportions of unsaponifiable material of 52 – 62 % have been published in the literature. The following values are typical for a semiraffinate [12]:

Wax esters (aliphatic and sterol esters)	78 – 82 wt %
Free fatty and wax acids	ca. 14 wt %
Free alcohols (wax alcohols and sterols)	6 – 7 wt %
Hydrocarbons	3 – 5 wt %

On comparing crude wax with a raffinate sample [40] in terms of chain-length distributions of the even-numbered acids and wax alcohols and of the paraffins, characteristic differences can be seen, particularly with the acids.

Importance and Uses. From an estimated annual worldwide harvest of ca. 510×10^6 t of sugarcane, ca. 5.1×10^6 t/a of dry filter cake material is theoretically obtained. From this, ca. 510 000 t of crude sugarcane wax could be extracted, given an average wax content of ca. 10 % in the filter cake. If refining losses are assumed to be 50 %, ca. 255 000 t of sugarcane wax raffinate per year remains theoretically available worldwide.

In the past, many attempts have been made to produce crude sugarcane wax on an industrial scale. As early as 1841, isolation of the wax from filter cakes was begun, and in 1909 a corresponding extraction process was patented.

Around 1916, crude sugarcane wax was extracted as a commercial product in South Africa and Java. In 1958 the wax was produced in two Cuban and one Australian plant. However, in 1960, production in Australia was halted and only one plant was still operating in Cuba [39]. In India, an extraction plant is thought to have been operating continuously since 1950 [39]. This plant is attached to a sugar refinery with a daily throughput of 1200 t of sugarcane.

Difficulties resulting from the necessary centralization of production locations, the necessarily rapid and complete drying of large quantities of filter cake material (which has a tendency to ferment), and last, but not least, the increasing costs of extraction have made sugarcane wax raffinate incapable of competing with carnaúba wax on the world market in terms of price. Consequently, sugarcane wax raffinates have only regional importance.

In producer countries, crude sugarcane wax can be used in floor and shoe polishes, carbon paper production, impregnating agents, and production of wax compounds.

2.2.5. Retamo Wax

Occurrence and Isolation. In arid areas of Argentina, *Bulnesia retama* grows wild as trees and shrubs. Branches and twigs are coated with a wax deposit that can be collected mechanically after drying the plant material harvested in summer. The crude wax is purified by melting with dilute sulfuric acid and subsequent filtration through cloths. It is sold exclusively on the home market as retamo wax [12].

Properties and Composition. Retamo wax is a light to medium brown, odorless, hard wax. It is very similar to other vegetable waxes, such as carnaúba wax, with regard to its solubility in polar and nonpolar organic solvents.

Data found in the literature, such as *mp* (76 – 78 °C), acid number (ca. 49 mg KOH/g), saponification number (ca. 87 mg KOH/g), proportion of hydrocarbons (15 – 25 %), and content of unsaponifiable components (38 – 42 %), characterize the wax less precisely than the typical chain-length distributions of even-numbered wax acids and alcohols and of paraffins [40].

Uses and Importance. With a decreasing annual production of several hundred tonnes (1975, ca. 500 t/a), retamo wax is of limited importance on the Argentinian market. It is used for floor, car, and shoe polishes and in the production of colored polishing inks.

2.2.6. Jojoba Oil

According to the DGF definition, jojoba oil is not a wax. However, its chemical composition allows it to be classified as a "liquid wax" in the sense of the classical definition of waxes.

Occurrence and Isolation. The jojoba bush (*Simmondsia chinensis* = *Simmondsia californica* = *Buxus chinensis*) grows wild over an area of ca. 250 000 km^2, most of which lies in the Sonora desert. The population density varies from a few to 500 plants per hectare. In certain areas, millions of jojoba bushes can be found.

The bushes carry fruit capsules that can contain one to three nuts of differing sizes. The yield from the harvest of wild jojoba bushes can range from a few nuts to 14 kg of nuts per plant. The average yield is ca. 1.8 kg per bush.

The nuts contain 50 – 60 wt % oil, which can be extracted almost completely from the deshelled, granulated fruit. Around 35 – 40 % of a particularly high-value oil is obtained by cold pressing and a further 15 – 20 % of a lower-value oil by extraction of the pressed cake. The oil is purified by simple filtration [41].

Properties and Composition. Jojoba oil is a virtually colorless to golden yellow, odorless, unsaturated oil, for which the most important characteristics are as follows:

mp	6.8 – 7.0 °C
Acid number	2 mg KOH/g
Saponification number	92 mg KOH/g
Iodine number	82 g I_2/100 g

A highly crystalline, relatively hard wax is obtained from jojoba oil by hydrogenation. Saponification of the oil gives 51 % unsaponifiable material and 52 % total acid.

Jojoba oil has the following composition with regard to substance groups:

Wax esters	ca. 97.0 wt %
Free alcohols	ca. 1.1 wt %
Free acids	ca. 1.0 wt %
Sterols	ca. 0.9 wt %

Unlike other vegetable waxes, the acids and alcohols in jojoba oil are monounsaturated, straight-chain representatives of the chain-length range $C_{16} - C_{24}$. Among the acids the C_{18} (ca. 12.5 % of the total acids), C_{20} (ca. 67 % of the total), and C_{22} components (ca. 14.5 % of the total) predominate, and among the alcohols, the C_{20} (ca. 48 % of the unsaponifiable components) and C_{22} components (ca. 41 % of the unsaponifiable components). With this unusual composition, jojoba oil is very similar to sperm oil contained in the skull a bone of the sperm whale.

Uses and Importance. An extraordinarily high market demand exists for a sperm oil substitute, which can be satisfied to a great extent by jojoba oil. This demand is a result of the worldwide prohibition of whaling, even though the latter is not obeyed by all countries.

The first experimental jojoba plantations were begun in the United States (Arizona, Texas, and California) to effect infrastructural improvements. Many other jojoba oil plantations were started in Mexico, Argentina, Australia, Brazil, Africa, Egypt, Israel, and countries whose climatic conditions allowed desert farming.

Because management of these plantations is difficult and the time required to achieve economical harvests is relatively long (eight to ten years), jojoba oil is available in insufficient quantities at a very high price. Its use is therefore limited to cosmetics and pharmaceuticals. As soon as reliable quantities can be supplied at

acceptable prices, jojoba oil may attract extraordinarily great interest as a raw material, particularly in the production of high-performance lubricants and in modified form in plasticizers and stabilizers. *Hydrogenated jojoba oil*, which corresponds to fatty esters such as stearyl stearate in its properties, will remain noncompetitive with these esters in terms of price and will therefore not achieve market importance.

2.2.7. Other Vegetable Waxes

Many other vegetable waxes are known, and the composition of some of them has been investigated [12]. However, they are either of very limited local importance or of purely academic interest. These waxes include:

Caranday wax
Raffia wax
Columbia wax
Esparto wax
Alfalfa wax
Bamboo wax
Hemp wax
Douglas fir wax
Cork wax
Sisal wax
Flax wax
Cotton wax
Dammar wax
Cereal wax
Tea wax
Coffee wax
Rice wax
Ocatilla wax
Oleander wax
∗ Palm waxes

Japan tallow wax, which is available commercially, fulfills some application-oriented criteria as a wax but, because of its chemical nature, is a vegetable tallow and thus belongs to the group of other tallows such as myrica, bayberry, ucuhuba, Borneo, Malabar, and illipe.

2.3. Animal Waxes

Animal waxes are obtained mainly from bees (Apoidae), particularly the honeybee (Apidae), stingless bee (Meliponinae), and bumblebee (Bombinae). Some scale insects (Coccidae) also produce waxes.

Waxes obtained from land mammals (sheep, goats, llamas, and dromedaries) and marine animals, as well as those of microbiological origin, are also known.

2.3.1. Beeswax

Beeswax is the oldest natural wax used by humans and still the most important animal wax.

Occurrence. Beeswax is an end product of the metabolism of a honeybee class (*Apis mellifica*, *A. carnica*), which belongs to the *Apis* genus. This species of honeybee is found worldwide and domesticated to a high degree. Wax is secreted from the wax glands in the lowest abdominal segments of the worker bee as white, slightly transparent flakes (virgin wax). From there it is taken to jaw, plasticized by chewing, and then used in this form as a building material for incubating combs and honeycombs.

If plenty of food is available and honey production is increasing, the wax glands are stimulated to enhance production for building new honeycombs. If the beekeeper makes preformed honeycombs available to the bees, honey production is increased [13], [36], [42], [43].

A bee population of ca. 30 000 worker bees produces ca. 270 g of wax from 3 kg of sugar or 3.6 kg of honey in one-fifth of a generation – ca. 50 d. A bee population consumes ca. 50 kg of pollen and 62 kg of honey per year.

Beeswax is produced in almost all countries of the world. In Europe, wax produced in a particular country is consumed there and supplemented by imports. Suppliers are African countries, such as Tunisia, Morocco, Kenya, Tanzania, Zambia, and the Central African Republic. North America, Australia, New Zealand, Russia, and the People's Republic of China also export beeswax.

Isolation. To obtain wax, the combs are freed from honey as far as possible by centrifuging. Then the wax is isolated either thermally by using sunlight or with boiling water. Particularly heavy impurities become enriched in the aqueous phase. Insoluble impurities remaining in the wax melt can be removed by filtration through

a fine mesh sieve of filter cloths, or by pressure filtration or centrifuging.

Colored, wax-soluble impurities can often be removed by treating the wax melt with activated carbon, aluminum or magnesium silicates, or other bleaching earths. As in pressure filtration and centrifugation, wax-containing residues are formed from which lower-quality waxes can be isolated by solvent extraction. The oldest method of lightening the color, which is said to have been practiced in 1000 B.C., involves bleaching beeswax, that has been processed to thin strips or shavings (large surface), in air and sunshine.

In the aforementioned purification processes, beeswax remains chemically unchanged and its characteristic, honeylike odor is retained. The success of the process depends mainly on the source of the beeswax.

Posttreatment of beeswax is now limited to washing with water or adsorptive purification with bleaching or diatomaceous earths. Bleaching with hydrogen peroxide is used out because peroxy compounds are formed [8].

If the process is carried out skillfully, chemical bleaching processes, such as refining with chromic acid – sulfuric acid, give very light, almost chemically unchanged products, which have lost the aromatic odor of the starting material. However, these processes are not employed. The same holds true for those involving mild bleaching with permanganate and sulfuric acid or the use of chlorine-containing bleaching agents such as chlorine bleach, sodium chlorite, and chloramine. The latter give waxes that do not have stable colors and often contain chlorine.

Beeswax is sometimes derivatized by ethoxylation or esterification. The products obtained do not have any market importance.

Properties. Crude, mechanically purified beeswax contains, besides wax, flower and pollen pigments and propolis resin, which acts as a cement in honeycomb construction. Therefore, depending on the source, the wax can have a yellow, orange, or dark brown color.

Beeswax is moderately hard. It becomes plastic and kneadable on warming in the hands, without sticking. It has a noncrystalline, finely grained fracture pattern. On cutting, matt surfaces are produced. Beeswax has a honeylike odor that intensifies on melting. It is moderately soluble in polar and nonpolar organic solvents in the cold, and completely soluble when heated to its boiling point.

Chemically bleached beeswaxes are cream to ivory in color and practically odorless and tasteless. They have a somewhat harder and more brittle consistency than chemically untreated products.

The following limiting values have been established for the properties of mechanically purified and bleached beeswax [20]:

mp	62 – 65 °C
Acid number	17 – 24 mg KOH/g
Ester number	72 – 79 mg KOH/g
Ester number/acid number	3.3 – 4.2
Saponification turbidity point	65 °C

Extensive investigations show that properties vary even among waxes from the same source [44].

Composition. The following values are given in the literature for the approximate composition of beeswax with regard to wax classes:

Wax esters	70 – 80 %
Free acids	10 – 15 %
Paraffins	10 – 20 %

Alkenes, isoparaffins, cholesteryl esters, and pollen pigments are present in minor amounts [12], [45].

Since the typical data for beeswax vary between relatively wide limits, misrepresentation of beeswax is difficult to detect by using these data [46]. A large number of publications exist on the analysis of beeswax [47]. Beeswax can be separated into the various substance classes by chromatographic methods [48–53]. Beeswax of different origin has been shown to have similar composition.

Comparative GC studies, in which ca. 65 % of the waxes investigated are analyzed (residues mostly involatile di- and triesters), also show great similarity among beeswax of widely varying origin [54] (see Table 5).

Parallel GC investigations on a genuine beeswax sample and the same sample after hydrolysis and silylation show that $C_{22} - C_{36}$ wax acids are present predominantly as the free acid, while all $C_{12} - C_{20}$ fatty acids are esterified [54].

Table 5. Composition of beeswax of different origins ∗[54]

Components	Eastern Switzerland	Brazil	China	Canada	Czech Republic
Hydrocarbons, %	14	15	16	14	13
Fatty (wax) acids, %	12	14	9	(8) – 12	13
Fatty alcohols, %	1	1	1	1	
Palmitates, %	22	20	24		
				35 ∗∗	31 ∗∗
Oleates, %	4	3	5		
Hydroxymonoesters, %	12	12	11	(4) – 8	13
Total, %	65	65	66	(62) – 70	70

∗ Values in parantheses were found only in a few cases.
∗∗ Palmitates and oleates.

Gas chromatography of hydrolyzed and silylated beeswax allows a total analysis in which 95 % of its components can be identified [54] (Table 6).

Table 6. Percentages of individual substance classed in a hydrolyzed wax sample from the honeybee *Apis mellifica* (Switzerland) [54]

Chain length (C-number)	Hydrocarbons	Fatty acids	Hydroxy fatty acids	Fatty alcohols
16		17.98	11.60	
17				
18		2.44	0.71	
19				
20		0.15		
21				
22		0.33	1.66	0.12
23	0.25			
24	0.10	4.94	0.59	5.59
25	0.58			
26	0.11	1.75	1.38	3.37
27	4.42			
28	0.08	1.78	0.10	4.37
29	2.64			
30	0.07	1.58		9.58
31	3.27			
32		1.05		7.98
33	2.24			
34		0.94		0.88
35	0.12			
36		0.08		
Total	14.23	33.02	16.04	31.89

Studies using GC show that a spectrum of wax samples from *Apis mellifica* has the same characteristic composition independent of the breed of bee. Small differences exist only in the percentage of individual components and substance classes, so that in general the following ranges of composition are correct for various classes of substance [54]:

Hydrocarbons	13 – 16 %
Fatty acids	30 – 35 %
Palmitic acid	15 – 20 %
Hydroxy fatty acids	13 – 17 %
Fatty alcohols	30 – 35 %
Other substances	4 – 8 %

The remaining substances are present mainly in concentrations of <0.05 %.

By using these facts, beeswax misrepresentation can be detected by a corresponding total analysis of the wax even in the presence of small quantities of blending materials [54].

Uses. Beeswax, mainly the yellow variety, is one of the raw materials used in the production of household *candles*, expensive decorative candles, and candles for religious purposes. For the latter, the Roman Catholic Church prescribes a minimum beeswax content of 10 %. Beeswax is also used for artistic and handicraft purposes (wax flowers, wax sculptures, painting, and batik). Most bleached and purified beeswax is used in *cosmetics and pharmaceuticals* (to regulate the consistency of lipsticks, creams, ointments, and suppositories) and in the *food sector* (as a release agent and a polish in confectionery production). Its use in these areas is favored by corresponding licensing regulations.

A significant proportion of beeswax, ca. 40 % of the total market volume, is sold back to beekeepers for the production of *synthetic honeycombs*, in which blends with paraffins and ceresin, which are not allowed for pharmaceuticals, are often used.

Economic Aspects. From the worldwide honey production figures for 1976, production of 11 000 – 19 000 t of beeswax was esti-

mated. Whether the quantities used in the candle and beekeeping industries were included in the worldwide consumption figure of ca. 10 000 t for 1981, is not known. At that time, total imports for the Federal Republic of Germany amounted to 1000 t.

At present (1994), ca. 15 000 t of beeswax is available on the world market. Ca. 6000 t is used in industry, and ca. 4000 t is sold back to beekeepers [8].

The annual availability of beeswax depends strongly on climatic variations in exporter countries and bee diseases that break out occasionally. In 1993, for example, a marked drop in exports from the United States occurred because of catastrophic weather conditions in the Mississippi area [8].

2.3.2. Other Insect Waxes

The following waxes are not market products and only in some cases have limited regional importance in the countries of origin.

Ghedda Wax. Ghedda wax is produced by some types of wild bee bound in Asia (*Apis indica*, *A. florea, A. dorsata*). It was formerly exported as such but more frequently now is blended with beeswax from the East Indies.

Ghedda wax has a yellow-brown color and varying characteristic data (e.g., acid number 4 – 11 mg KOH/g, saponification number 86 – 130 mg KOH/g, iodine number 5 – 11 g I_2/100 g, *mp* 60 – 66 °C). It has a consistency similar to beeswax with a fatty feel, which can be attributed to its glyceride content. Ghedda wax consists of ca. 78 – 80 % wax esters, 4 % glycerides (fats), 5 – 6 % free acids, 8 – 9 % hydrocarbons, and 1 – 2 % partly resinous impurities [12].

Among the acids are 7- and 16-hydroxypalmitic acids and 7,16-dihydroxymargaric acid, which can be used for identification of ghedda wax or other beeswaxes blended with ghedda wax in modern umpire assays.

Shellac Wax. The resinous exudate of the scale insect *Laccifer lacca* (formerly *Tachardia lacca*) of the Coccidae family is an important commercial product, known as shellac (→ Resins, Natural, Chap. 4.19.). Shellac contains 4 – 5 % wax, which can be obtained by treatment with dilute soda solution or 90 – 95 % ethanol (spirit extraction) as an insoluble component.

Shellac wax is a hard, yellow to brown product. Very light, bleached types are also known. The widely varying data found in the literature (e.g., acid number 12 – 24 mg KOH/g, saponification number 63 – 126 mg KOH/g, iodine number 6 – 9 g I_2/100 g, solidifying point 58 – 80 °C) may be attributed to the different sources of shellac and different methods of extraction and post-treatment of the wax component. Thus, the composition with regard to substance classes also varies as follows [12]:

Wax esters	70 – 82 %
Free acids	10 – 24 %
Free alcohols	< 1 %
Hydrocarbons	1 – 6 %
Shellac	1 – 4 %

Investigations [55] of bleached shellac waxes show that the total acid content of the wax, similar to that of beeswax, consists essentially of a mixture of even-numbered fatty acids ($C_{12} - C_{18}$, 21 – 26 % of the total acid) and waxy acids ($C_{28} - C_{34}$, mainly C_{32} and C_{34}). A significant component is trihydroxypalmitic acid (aleuritic acid), even though it occurs only in quantities of 0.5 – 1.0 %. Among the alcohols ($C_{28} - C_{32}$) the C_{28} component predominates (62 – 65 % of the total alcohol). The main hydrocarbon components are paraffins with 27, 29, and 31 carbon atoms.

Chinese Insect Wax. The East Asian wax scale insect (*Ericerus pela*, subgenus Lecaniinae, family Coccidae) secretes a wax on the twigs and leaves of trees and shrubs on which it is a parasite. The wax can be extracted by melting with hot water. It is hard, brittle, crystalline, and odorless, with a light color, and consists of 95 – 97 % wax esters, 0.5 – 1.0 % free acids, and 2 – 3 % hydrocarbons [12]. The high proportion of wax esters is significant. The East Indian wax scale insect (*Ceroplastes ceriferus*) is said to produce a wax similar to beeswax in composition [12].

2.3.3. Wool Wax

Sheep secrete a substance from their sebaceous glands to protect the epidermis and the wool from adverse weather conditions and from their own acidic perspiration. This substance is incorrectly referred to in the literature as wool fat. Because of its chemical composition it should definitely be classified as a wax.

Wool wax is obtained in emulsified form from raw wool by machine washing with soap liquor. Extraction of wool with suitable organic solvents is not employed because it would lead to complete degreasing of the wool and thus make it nonelastic.

Many processes exist for obtaining crude wool wax from the emulsion, some of them patented [36]. Wax is first precipitated along with the acids of the soap solution by addition of sulfuric acid. The fatty acids remain in the crude product regardless of whether the precipitates are pressed out, centrifuged, or extracted with solvents. They can be removed almost completely by dissolving the crude product (e.g., in hexane) and subsequent extraction with a mixture of alcohol and dilute aqueous sodium hydroxide. Further purification steps (e.g., bleaching with hydrogen peroxide in the presence of phosphoric acid [56]) give neutral wool wax, which is usually treated subsequently with adsorbents such as bleaching earth or activated carbon and sold in a highly pure form as Adeps lanae.

Crude wool wax is a greasy, glutinuous, brown-yellow to brown-black substance with a penetrating goatlike odor (*mp* 34 – 38 °C). *Neutral wool wax* is yellow to light brown in color (*mp* 38 – 42 °C), with a milder odor, whereas *Adeps lanae* is a pale yellow, almost odorless substance (*mp* 40 – 42 °C).

The *composition* of wool wax varies depending on the origin of the wool, breed of sheep, climatic conditions, farming method, and type of pasture.

Wool wax usually consists of a mixture of ca. 48 % wax esters, 33 % sterol esters, 6 % free sterols, 3.5 % free acids, 6 % lactones, and 1 – 2 % hydrocarbons [12]. The low *mp* and appearance of wool wax may be attributed to the high proportion of branched acids.

Uses. Wool wax is used mainly as the crude product, as neutral wool wax, or as hydrogenated neutral wool wax and in the form of wool wax alcohols (obtainable by hydrolysis) in leather polishes, cosmetics (e.g., creams, baby-care products, and toilet soap), pharmaceuticals (e.g., plasters, ointments, and suppositories), and lubricants. The purified, commercial product is known as wool wax or lanolin.

3. Montan Wax

3.1. Formation and Occurrence

Montan wax is a vegetable fossil wax. It forms part of the extractable, bituminous components of lignite and peat. The composition of bitumen depends on the carbonized plant material and on the geology and conditions of the deposit. The main components of bitumen are waxes, resins, asphaltenes, and dark residues. If the proportion of wax is >60 %, the bitumen is generally known as crude montan wax. Prerequisites for formation of this wax were the same or similar to those now necessary for formation of recent natural waxes (e.g., carnaúba wax [57–60]). Lignite, which originates from plant material of the Eocene period, has a high and economically utilizable proportion of wax components. Deposits of lignite, which are used for wax extraction, are found mainly in eastern Germany (Röblingen), Ukraine (Alexandrija), Russia (Baschkiren), the United States (California), and China. Other deposits are of only local importance or are not even exploited. The largest crude montan wax producers are the Montanwerke Romonta in Röblingen (Germany) and the American Lignite Co. in the United States.

3.2. Extraction

The most economically important bitumen-rich lignite deposits are found in the area around the city of Halle in Germany. An industrial process for extracting montan wax from this type of coal was based on investigations bei Riebeck. The original patent [61], [62] for this process, published in 1880, describes how montan wax can be obtained by solvent extraction. In 1900 the production of montan wax was begun in Völpke, and later in other locations. Romonta in Amsdorf

[62] is currently the most important production site. There, lignite with a bitumen content between 10 and 20 % is used for industrial extraction of montan wax. The process used involves the following steps: size reduction, drying, extraction, evaporation of the extract, and formulation [57], [60]. The yield and composition of the extract are determined by coal quality (clay and mineral content); physical parameters such as water content, particle size, and particle-size distribution; and properties of the solvent.

Granulating and Drying. Run-of-mine, pit-wet lignite contains ca. 50 % water. Before extraction it is ground in roll mills and dried to a residual moisture level of ca. 15 % in steam heated tubular dryers. The particle-size distribution is ca. 7 % >8 mm, 75 % 1 – 8 mm, and 18 % fines.

Lignite Extraction and Wax Production. Various machines such as band, chamber, carousel, or bucket chain conveyor extractors are used for extraction. Many solvents and solvent mixtures have been tested and used to extract wax from lignite. They include aliphatic, aromatic, and unsaturated hydrocarbons and hydrocarbon mixtures, mono- and polyhydric alcohols, ketones, esters, ethers, and mixtures of hydrocarbons and alcohols. The yield of extract is highest when polar solvents (e.g., ethanol) or solvent mixtures (e.g., ethanol – toluene) are used. However, the extract then contains a high proportion of polar substances, such as resin acids, sterols, triterpenes, or dark residues, and only a small proportion of wax. With hydrocarbons the overall yield is lower, but the proportion of wax in the extract is significantly higher. In industry, only hydrocarbons or hydrocarbon mixtures are therefore used (e.g., toluene at Romonta in Röblingen). Mixtures of aromatic and aliphatic hydrocarbons and naphthenic benzine are used by other producers.

If toluene is used, extraction is carried out at 85 °C. After extraction, the solvent contains ca. 8 – 10 % dissolved crude montan wax. The mixture is worked up in a three-stage distillation. The bitumen content is increased to 50 % in a continuous evaporator and then to 90 % in a circulation evaporator. The remaining solvent is removed by blowing in steam directly or in a falling-film vaporizer. The wax melt is pelletized or poured into molds. The extracted coal granulates are freed from solvent by steam stripping and used for electricity production in a power station. The toluene – water mixture is worked up, and toluene is recycled to the extraction process.

3.3. Properties and Composition

Crude montan wax is a black-brown, hard, brittle product with a conchoidal fracture pattern. It cannot be scratched by fingernails at room temperature. Crude montan wax has a penetrating terpene-like odor, which becomes very intense on melting. The wax is soluble in many organic solvents, particularly aromatic or chlorinated hydrocarbons, even on moderate heating. Crude montan wax consists of a mixture of wax acids, wax esters, resins, asphaltenes, and dark residues. The ester and acid components can be saponified and emulsified by treatment with aqueous alkali [57], [59], [60]. The waxes form the crystalline component of montan wax. They effect the relatively sharp transition from solid to liquid state and vice versa. The resins and asphaltenes have a more amorphous nature and exhibit glassy behavior. The qualitative and quantitative composition of crude montan wax is determined by the carbonized plants, extent of carbonization, the solvent, and extraction conditions.

A typical crude montan wax contains ca.:

Wax acids	35 %
Wax alcohols	20 %
Hydroxycarboxylic acids	10 %
Dicarboxylic acids	3 %
Wax ketones	1 %
Hydrocarbons	2 %
Resin acids	15 %
Sterols	10 %
Dark residues	4 %

Characteristic properties are:

Acid number	20 – 40 mg KOH/g
Saponification number	70 – 120 mg KOH/g
Iodine number	30 – 45 g I_2/100 g
Drop point	ca. 85 °C
Density (100 °C)	ca. 0.86 g/cm^3
Thermal expansion (20 – 100 °C)	18 %

Wax Components. Crude montan wax contains 70 – 75 % waxlike components, of which the following have been isolated:

Straight-chain carboxylic acids	$C_{16}-C_{34}$
Straight-chain alcohols	$C_{18}-C_{34}$
Straight-chain ω-hydroxycarboxylic acids	$C_{16}-C_{32}$
Straight-chain 2-hydroxycarboxylic acids	$C_{16}-C_{32}$
Straight-chain ω-dicarboxylic acids	$C_{22}-C_{32}$
Esters of $C_{16}-C_{34}$ carboxylic acids and $C_{18}-C_{34}$ alcohols	
Straight-chain aliphatic hydrocarbons	$C_{25}-C_{33}$

These compounds form homologous series, in which chain lengths in the range of $C_{28}-C_{30}$ predominate. The chain-length distribution for straight-chain carboxylic acids and alcohols found by GC analysis is shown in Figure 2.

Resins and Dark Residues. Besides the wax components, crude montan wax also contains ca. 25 – 30 % resins and dark residues. The quantity and composition of these components are determined mainly by the properties of the lignite, the extraction solvent, and the extraction temperature. The resin can be removed from crude montan wax by exhaustive extraction with methanol. In industry, this method is not used to determine resin content because it is expensive and time-consuming. Instead, the proportion of soluble compounds (i.e., resins) is determined by a standardized procedure such as solubility in acetone or in a toluene – ethanol or benzene – ethanol mixture. The resin is characterized by its chemical and physical properties, of which the following are typical:

Drop point	ca. 75 °C
Acid number	ca. 33 mg KOH/g
Saponification number	ca. 75 mg KOH/g
OH number	ca. 70 mg KOH/g
Iodine number	ca. 90 g I_2/100 g
Sulfur content	ca. 2 – 2.5 %

Determination of all individual components in resin extracts has not yet proved possible. However, substance classes and typical compounds in these classes have been identified: e.g., triterpenes such as betulinol, sterols, diterpenoid acids and their esters, and diterpenoid alcohols and their esters.

3.4. Refining and Derivatization

Because of its dark color and its resin content, which sometimes lead to problems, direct use of crude montan wax is limited. For most applications the wax must be refined. Refining involves three steps: extractive deresinification, oxidative bleaching, and subsequent derivatization [57–60], [63].

3.4.1. Deresinification

The resin components of the crude wax are soluble in many solvents even at low temperature (10 – 20 °C). Treatment of pulverized crude wax with methylene chloride, ethyl acetate, or methanol reduces the resin content to 5 – 10 %. Other solvents, such as ketones, ethers, hydrocarbons, mixtures of alcohols and aromatics, liquid SO_2 or supercritical CO_2, can be used but have not achieved industrial importance. After removal of the solvent, the extracted resin contains ca. 10 – 20 % wax. Deresinified wax can be used directly or as a mixture with other wax components for various applications. Such products are sold by Romonta, Amsdorf, and the American Lignite Co. Most deresinified crude wax is subjected to oxidative bleaching.

3.4.2. Bleaching

A purification process for the crude wax was sought very early on. Of the many processes investigated, only three have actually been used in production [57], [60], [64–67].

Distillation was the first method for obtaining light, hard waxes from crude montan wax. The crude wax was heated to 420 °C in stills and thus partially cracked. The distillate therefore consisted of 50 – 80 % hydrocarbons and 20 – 50 % wax acids. The hydrocarbons were removed in a further workup step, leaving the montan wax acids.

Oxidation with Nitric and Sulfuric Acids. In this process, paraffin had to be added as a diluting agent. The resin was precipitated as the nitroresin and removed. The dark residues

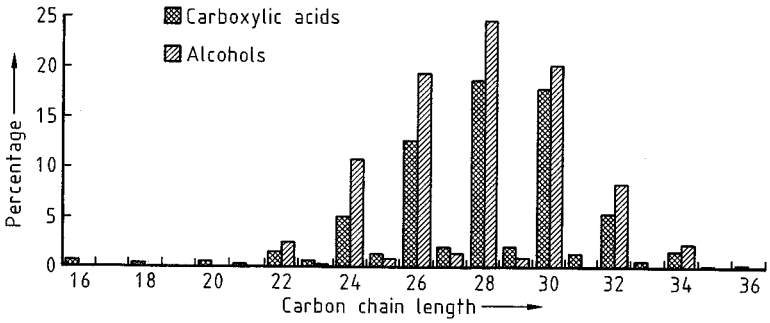

Figure 2. Chain-length distribution of acids and alcohols in montan wax

were coked and the wax acids extracted from this coke. However, the wax still had a certain amount of dark residues and nitrate-containing products.

Oxidation with Chromic Acid or Chromates in Sulfuric Acid. In processes described thus far, the wax is damaged considerably and the chainlength distributions of its constituents are altered. Oxidative bleaching with aqueous chromic – sulfuric acid mixtures, developed by I.G. Farben, is much more suitable. The starting material is deresinified crude montan wax. The resin components and dark residues are oxidized to carbon dioxide or low molar mass, water-soluble compounds in a multistage oxidation process. The wax esters are simultaneously hydrolyzed; the wax alcohols formed are oxidized to wax acids; and the hydroxycarboxylic acids are oxidized to dicarboxylic acids. Some esters remain unaltered.

$$CH_3-(CH_2)_n-COO-CH_2-(CH_2)_n-CH_3 \xrightarrow{H_2O} CH_3-(CH_2)_n-COOH + CH_3-(CH_2)_n-CH_2OH$$

$$CH_3-(CH_2)_n-CH_2OH \xrightarrow{CrO_3} CH_3-(CH_2)_n-COOH$$

$$HOCH_2-(CH_2)_n-COO-CH_2-(CH_2)_n-CH_3 \xrightarrow{H_2O} HOCH_2-(CH_2)_n-COOH + CH_3-(CH_2)_n-CH_2OH$$

$$HOCH_2-(CH_2)_n-COOH \xrightarrow{CrO_3} HOOC-(CH_2)_n-COOH$$

where n is an even number between 24 and 32.

A mixture of light wax acids, which are very similar in chain-length distribution to the crude wax, is obtained in a yield of 80 – 85 % based on crude wax. Minor components of the mixture are unchanged esters and long-chain dicarboxylic acids. The spent oxidizing agent is removed and worked up, and the bleached wax is freed from residual chromium salts in a subsequent step.

Depending on the quantity of chromic acid, the acid number of the bleached wax is between 100 and 135 mg KOH/g. Oxidation can be performed batchwise or continuously. In the *Gersthofen process* used by Hoechst, the oxidizing agent is regenerated electrochemically. BASF, Völpke, and other producers use the spent oxidizing agent in an associated plant for the production of chromium-containing tanning agents.

Development is still being carried out on refining processes [68] such as alkaline dehydrogenating oxidation with anhydrous alkali hydroxide. Apparently, these processes have not been used in industry up to now. An overview of the current processes used for extraction and refining of montan wax is given in Figure 3.

3.4.3. Derivatization

Reconstruction of the original wax structure is attempted by treating the wax acids obtained with dihydric alcohols or $Ca(OH)_2$. By varying the proportions of free acid and soap and the degree of esterification, the reaction can be carried out in such a way that the products formed are optimized for a particular application [57], [64], [65], [69].

$$CH_3-(CH_2)_n-COOH + HOCH_2-CH_2OH \rightarrow CH_3-(CH_2)_n-COO-CH_2-CH_2-OR$$

$$R = H \text{ or } CH_3-(CH_2)_n-CO-$$

$$CH_3-(CH_2)_n-COOH + Ca(OH)_2 \rightarrow (CH_3-(CH_2)_n-COO)_2^- Ca^{2+}$$

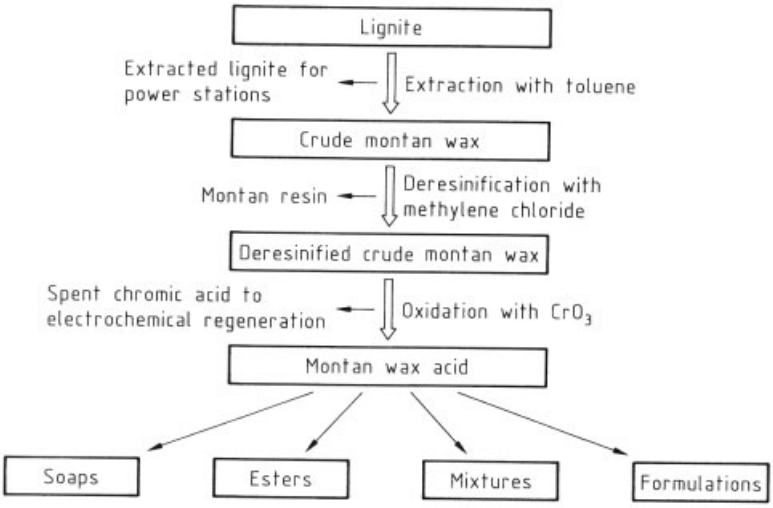

Figure 3. Flow sheet for extraction and refining of montan waxes

The most important products in this series are the esters of ethylene glycol (e.g., Hoechst Wax E or Hoechst Wax KSL) and mixtures of esters of 1,3-butanediol with calcium montanate (e.g., Hoechst Wax OP). Broad variation in properties can be achieved by using tri- and polyhydric alcohols, aromatic alcohols, aminoalcohols, and ethoxylated or fluorinated alcohols. Reactive waxes can be produced from partial esters of polyols by treatment with dicarboxylic acids, isocyanates, epoxides, or unsaturated carboxylic acids.

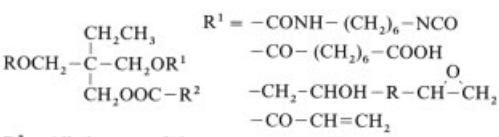

R^2 = Alkyl group of the montan wax acid
R = H, R^2CO, R^1

These waxes exhibit interesting properties and have been produced and used for special applications. If the reaction is not stopped at the reactive wax stage and a suitable stoichiometry is chosen, oligomeric waxes are produced.

3.5. Uses and Economic Aspects

Montan wax acid and its derivatives are versatile materials [64] , [65], [69–71]. They are used in the form of flakes, powders, pastes with solvents, or aqueous emulsions. The main areas of use are in polishes for floors, cars, and leather; in plastics processing as lubricants, release agents, and nucleating agents; in the paper and building industries; and in wood and metal processing. Montan wax derivatives are also used in small quantities in pharmaceutical formulations; cosmetics; and the production of adhesives, resins, and office equipment. In addition to their technical properties, the low toxicity of montan wax derivatives and the fact that BGA or FDA approval has already been granted for some derivatives are important criteria favoring their use.

Production figure of montan wax for 1994 are given below, together with a breakdown of montan wax consumption in different application areas:

Crude montan wax	
Romonta	ca. 19 000 t/a
Alpco	ca. 2 500 t/a
Ukraine	ca. 1 500 t/a
China	ca. 1 000 t/a
Bleached montan waxes	
Hoechst Gersthofen	ca. 11 000 t/a
Völpker Montanwachs	ca. 3 500 t/a
China	ca. 500 t/a
Poland	<500 t/a
Czechia	<500 t/a
Application areas	
Plastics, lubricant	ca. 7 000 t/a
Polishes, coatings	ca. 5 500 t/a
Technical applications	ca. 2 500 t/a

Table 7. Classification of waxes from crude petroleum

	Origin			
	light, medium, heavy lubricating oil distillates		residues from vacuum distillation	crude oil
Group	(macrocrystalline) paraffin waxes		microcrystalline waxes (microwaxes)	settling waxes
Subgroup	paraffin waxes	intermediate waxes	residue waxes	pipe waxes tank bottom waxes
Crude products	crude waxes (slack waxes)		petrolatum	raw waxes
Deoiled and refined products	scale waxes deoiled slack waxes filtered (decolorized) waxes fully refined waxes		plastic microwaxes hard microwaxes	
Side products from deoiling	soft waxes		soft petrolatum (microwax slacks)	

4. Petroleum Waxes

4.1. Introduction

The quantity of waxes obtained from crude petroleum has increased continuously for two reasons: (1) the demand for lubricating oils with low pour points and (2) the large proportion of paraffinic crudes in total crude oil production that have to be dewaxed for the production of lubricating oils. This development has increased the need to find more industrial applications for petroleum waxes. When the processing of lubricating oil began in the second half of the 19th century, waxes were still inconvenient side products. Since then, worldwide consumption of petroleum waxes has increased to 3×10^6 t/a. This consumption corresponds to more than 95 % of the waxes of all types produced worldwide. Depending on their natural occurrence and their crystallinity, petroleum waxes are divided into:

1) Macrocrystalline waxes (paraffin waxes)
2) Microcrystalline waxes (microwaxes)

Table 7 gives a detailed classification of paraffin and microcrystalline waxes based on origin and method of refining [80].

Paraffin waxes are obtained from light and middle lubricating oil cuts of vacuum distillation. Paraffin waxes also include waxes from heavy lubricating oil distillates, which are intermediates between macrocrystalline and microcrystalline waxes with regard to structure and composition (*intermediate waxes*).

Microcrystalline waxes originate from vacuum residues and from the sediments of paraffinic crude oil (settling waxes). Waxes that are liquid at room temperature are mostly contained in diesel oil or gas oil fractions (→ Heating Oils, → Aviation Turbine Fuels) and can be isolated from them. These are not dealt with here.

4.2. Macrocrystalline Waxes (Paraffin Waxes)

4.2.1. Chemical Composition and General Properties

Chemical Composition. *Paraffin waxes* consist predominantly of mixtures of straight-chain alkanes in a typical distribution of the homologous series whose molar masses depend on the boiling range of the lubricating oil distillate from which they are obtained. Long-chain, weakly branched isoalkanes are present in a much lower proportion, along with a very small fraction of monocyclic alkanes.

The *intermediate waxes* have a similar composition, but the molar masses of the *n*- and isoalkanes are higher. Intermediate waxes contain a higher proportion of cycloalkanes and isoalkanes; the latter more strongly branched than those in paraffin waxes.

According to the European Wax Federation (EWF), paraffin waxes have a C-number distribution of *n*-alkanes from 18 to 45 and a total content of iso- and cycloalkanes of 0 – 40 %. Typical data for the intermediate waxes are an *n*-alkane C-number of 22 to 60 and a total content of iso- and cycloparaffins of 30 – 60 %.

General Physical Properties. Paraffin waxes are insoluble in water and sparingly soluble in low molar mass aliphatic alcohols and ethers. They are more soluble in ketones,

chlorohydrocarbons, petroleum spirit, solvent naphtha, benzene, toluene, xylene, and higher aromatics, especially at elevated temperature. The solubility decreases markedly with increasing molar mass (higher melting point) of the waxes.

Chemical Properties. Paraffin waxes are extremely unreactive under normal conditions. Oxidation reactions occur only at elevated temperatures (e.g., on storage and processing above 100 °C), particularly in the presence of oxygen and catalytically active metals. These reactions can be recognized from the burnt odor produced and the yellow to brown coloration of the waxes. Nevertheless, under certain thermally and catalytically controlled conditions, these waxes can undergo chemical reactions such as chlorination (→ Chlorinated Hydrocarbons, Chap. 7.3.), oxidation, dehydrogenation, and cracking (→ Oil Refining, Chap. 3.6.3.2., → Oil Refining, Chap. 3.6.4.), of which chlorination and cracking are important in industry.

4.2.2. Division into Product Classes

Depending on the degree of refining, paraffin waxes are divided into the following product classes:

1) Crude waxes, also known as slack waxes
2) Slack wax raffinates (scale waxes)
3) Deoiled slack waxes
4) Soft waxes
5) Semirefined waxes
6) Filtered (decolorized) waxes
7) Fully refined waxes

Table 8 gives an overview of the change in physical characteristics with degree of refining.

Crude waxes (slack waxes) [*64742-61-6*] consist of a mixture of alkanes, which can be solid, semisolid, or liquid at room temperature; alkyl-substituted cyclopentanes and cyclohexanes (naphthenes); and alkyl-substituted aromatics. They also contain, as impurities, the typical contents of the lubricating oil cuts from which they originate, such as asphaltenes, resins, olefins, and sulfur and nitrogen compounds. The oil content of slack waxes is usually between 5 and 12 %, but can be as high as 25 %.

n-Paraffins and isoparaffins predominate in slack waxes in terms of quantity; both are present in continuously distributed homologous series. The chain-length spectrum (C-number distribution) is determined from the width of the boiling range, the distribution maximum (C-number maximum) from the boiling level of the lubricating oil fraction. In isoparaffins, compounds with terminal methyl branches predominate, followed by other methyl-substituted alkanes. The concentration of these compounds decreases as the branching point moves toward the middle of the chain. Other components are compounds with terminal ethyl branches and multiply branched structures, which can be detected only in slack waxes from heavy vacuum distillates [81], [82].

Table 9 shows the composition of some slack waxes as a function of origin.

Physical Properties. Depending on the origin and production process, slack waxes are light- to dark-brown, soft, unctuous to semisolid materials without a clear crystal structure because of their high oil and soft wax content. The congealing points lie between 35 and 65 °C; the drop points, between 35 and 68 °C; the needle penetration at 25 °C (a reciprocal measure of the hardness) is between 40 and 80 (at greater oil content, even higher); the viscosities at 100 °C are between 3 and 10 mm^2/s; the densities at 70 °C, between 775 and 815 kg/m^3; and the flash points between 190 and 250 °C.

Scale Waxes. Normally being physically decolorized slack waxes, scale waxes [*90669-78-6*] essentially have the same composition as the raw materials from which they are obtained. Only the dark material contents of the slack waxes (asphaltenes and resins) are removed by the decolorization process, and the content of alkylaromatics, and of sulfur and nitrogen compounds, is reduced. In physical properties, scale waxes are similar to slack waxes. Decolorization renders the products white to pale yellow.

Deoiled Slack Waxes (Crude Hard Waxes, Raw Waxes). In deoiled slack waxes [*8002-74-2*] the predominant proportion of the hydrocarbons that are liquid (oils) under normal conditions, and a certain proportion of the

Table 8. Variation in physical data with degree of refining of paraffin wax (starting material : slack wax from a medium machine oil)

Characteristics	Slack wax	Deoiled wax	Filtered (decolorized) wax	Fully refined wax
Congealing point, °C	59	62	62	62
Needle penetration, 0.1 mm				
at 25 °C	57	18	18	17
at 30 °C	80	26	26	25
at 35 °C	105	36	36	36
Oil content, %	9.9	0.5	0.5	0.4
Viscosity (at 100 °C), mm^2/s	6.8	6.1	6.0	5.8
Color	brown	brown	whitish	white
Fluorescence	very strong		weak	none

Table 9. Composition of slack waxes from various lubricating oil distillates of the same origin (80 % Arabian, 10 % German, and 10 % Norwegian crude oil)

Characteristic	Slack waxes from			
	Heavy spindle oil	Light machine oil	Medium machine oil	Heavy machine oil
Congealing point, °C	47	50	58	62
Oil content, %	9.5	9.0	9.8	8.0
n-Paraffin content, %	73	58	52	39
C-number range	16 – 36	19 – 43	21 – 48	22 – 58
C-number max.	24	29	34	41
C_{max} content, %	17	14	10	8

semisolid ones (soft waxes), have been removed. Liquid and semisolid waxes consist mainly of low molar mass *n*- and isoparaffins and of isoparaffins with centrally located and strongly branched side chains, as well as naphthenes and alkylaromatics. Therefore, in deoiled slack waxes the *n*- and weakly branched isoparaffins are enriched. By removal of the low molar mass portion a narrower C-number range results along with a more pronounced C-number maximum.

Table 10 shows these relationships with two waxes from heavy spindle oil and medium machine oil used as examples.

Physical Properties. In structure and consistency, deoiled slack waxes are coarse to medium crystalline, brittle to weakly plastic and depending on the degree of deoiling, hard to very hard. These waxes are light to dark brown and darken on heating. The congealing points are between 48 and 72 °C; the drop points between 48 and 75 °C; and the needle penetrations at 25 °C between 10 and 30. At the same degree of deoiling, the course of the penetration – temperature curve (increase in penetration in the penetrogram [83] essentially depends on the origin of the waxes (Fig. 4).

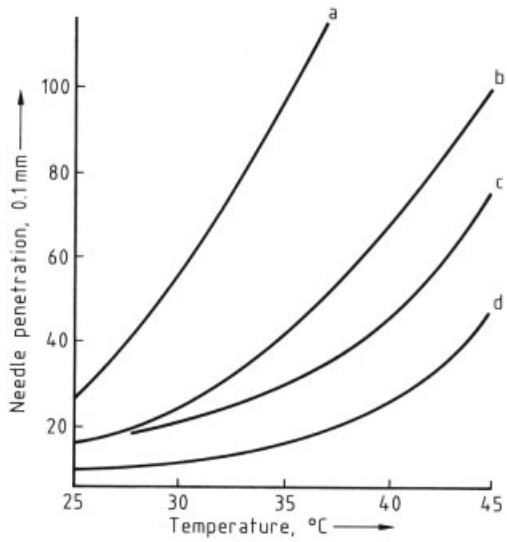

Figure 4. Variation of needle penetration with temperature for deoiled slack waxes of different origins
a) From heavy spindle oil (North Sea); b) From light machine oil (North Sea); c) From heavy machine oil (North Sea); d) Wax from Indonesia

Crude soft waxes [*64742-67-2*] are sometimes known as foot oils (see Section 4.2.3.2) and are formed in the deoiling of slack waxes. They consist predominantly of low molar mass

Table 10. Composition of deoiled slack waxes from two lubricating oil distillates (origin – see Table)

Characteristic	Slack wax ∗	Deoiled	Slack wax ∗∗	Deoiled
Congealing point, °C	47	50	58	67
Oil content, %	9.5	0.3	9.8	0.4
n-Paraffin content, %	73	84	52	64
C-number range	16 – 36	19 – 36	21 – 48	23 – 48
C-number max.	24	24	34	34
C_{max} content, %	17	19	10	11

∗ From heavy spindle oil.
∗∗ From medium machine oil.

n-paraffins and strongly branched isoparaffins; higher molar mass, weakly branched isoparaffins; naphthenes; and alkylaromatics. The *n*-paraffin content of soft waxes is generally <25 wt %.

Soft waxes are very soft, unctuous to inhomogeneous materials at room temperature because of their high oil content, which can be 30 wt % or greater. The oil is included in voids of the crystal lattice of the solid hydrocarbons. Soft waxes have a light- to dark-brown color and congealing points <40 °C. Their viscosities at 100 °C are between 3 and 12 mm^2/s; their flash points, between 180 and 220 °C.

Semi-refined, filtered (decolorized), and fully refined waxes are similar to deoiled slack waxes in the composition of their main components. Through refining they are freed from more highly condensed alkylaromatics and naphthenes, olefins, and sulfur and nitrogen compounds. Depending on the origin, their content of *n*-paraffins varies between 92 wt % (East-Asian waxes) and 45 wt % (waxes from heavy machine oil distillates of German origin). The composition of the refined waxes is determined not only by the origin of the crude (Table 11), but also by the boiling range of the starting distillate. As the boiling ranges of the lubricating oils rise, the molar masses of the hydrocarbons contained in waxes obtained from them increase. The concentration ratio of iso- and cycloparaffins to *n*-paraffins also increases. Thus, for example, a refined wax from Tuismasy (Ural) crudes contains 81.6 % *n*-paraffins with an average molar mass of 360 g/mol (average C-number 25.7) and 9.3 % isoparaffins (average molar mass 384 g/mol; average C-number 27.4). The naphthenes (9.1 %) consist of 81.5 % monocyclic and 15.5 % bicyclic alkanes (average C-number 28.9) [84].

Table 11. Composition of fully refined waxes (congealing point 62 °C in each case) from different origins

Characteristic	Wax from machine oil distillate		
	Saudi Arabia	Russia	East Asia
Congealing point, °C	62	62	62
Oil content, %	0.3	0.6	0.2
n-Paraffin content, %	52	78	92
C-number range	23 – 48	23 – 41	25 – 38
C-number max.	34	32	30
C_{max} content, %	10	14	21

Physical Properties. Waxes in this group differ from deoiled slack waxes and from each other in their residual oil content and, thus, in hardness, color, and color stability.

Semirefined waxes have an oil content of 1.5 – 3 %, and the needle penetration at 25 °C can be 20 – 60. They are somewhat plastic and kneadable, colorless to white, have good color stability under light, and are virtually odorless and tasteless.

Filtered (decolorized) waxes have an oil content of 1.5 % (max.) and a needle penetration at 25 °C of 10 – 26. They have a whitish color and are relatively color stable and generally odorless.

Fully refined waxes (oil content and penetration similar to decolorized waxes) are crystalline, no longer kneadable, pure white, transparent to slightly opaque, color stable, light-resistant, odorless, and tasteless. They fulfill purity criteria for the production of packaging materials for foods and for the formulation of cosmetics and pharmaceuticals. Detailed information on the chemical composition of fully refined waxes and methods for their analytical determination can be found in the CONCAWE report [85].

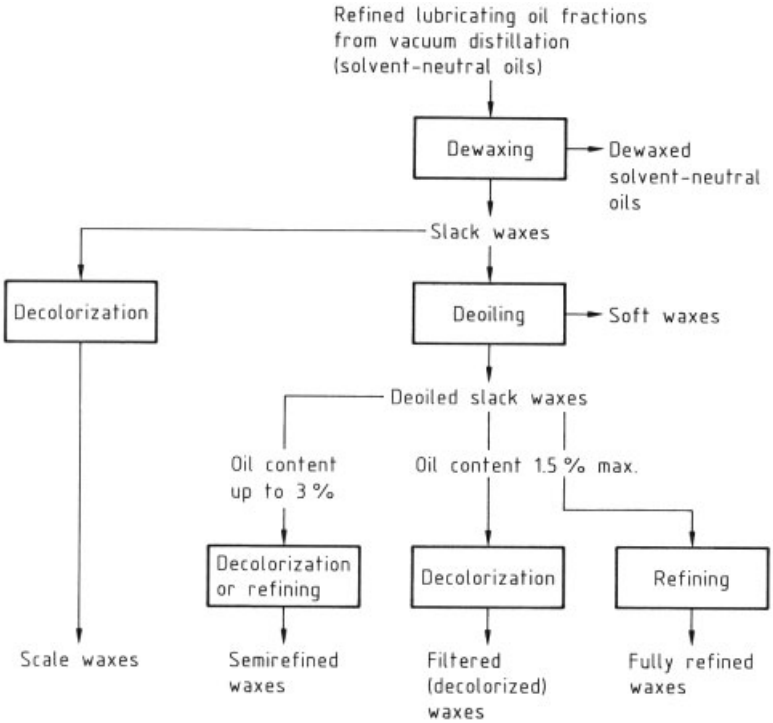

Figure 5. Process schematic for the production of paraffin waxes

According to a more recent definition, fully refined wax products are those that fulfill various national purity requirements (see Section 4.2.5). The oil content, which was once limited to 0.5 % (max.), is no longer a criterion since the oils contained in the wax also consist of *n*- and isoparaffins of the required purity. Depending on the literature, the oil content of fully refined waxes can be 1.5 % (max.) [86] or 2.0 % (max.) [87].

4.2.3. Occurrence of Raw Materials and Processing

Paraffin waxes are contained in crude petroleum and are obtained during oil refining. Depending on the source of the petroleum, the content of solid waxes can vary between 2 % (Rumanian origin) and 30 % (Indonesian origin); generally, however, it is between 3 and 15 % in crude petroleum from the main production areas.

Because of their higher molar masses and thus higher boiling ranges, waxes become enriched in the residues from atmospheric distillation (long residues). In processing the residues to lubricating oils, waxes are obtained in all fractions of the vacuum distillation. Because of their poor solubility, they give rise to the high pour points of the basic oils and must therefore be removed. The yields of wax vary depending on the origin and quality requirements of the oils. For a wax content of crude petroleum of ca. 5 %, 8 – 18 % solid waxes in the individual lubricating oil fractions can be expected.

Slack waxes are produced in the dewaxing of lubricating oil distillates (→ Lubricants and Related Products). The type of process used and process parameters used are according to the desired quality of the lubricating oils.

The process steps necessary for the production of refined waxes are shown in Figure 5.

4.2.3.1. Dewaxing Lubricating Oil Distillates

The principle of dewaxing lubricating oil distillates is based on the different crystallization temperatures of straight-chain and weakly branched paraffins and the oil phase. The wax-containing

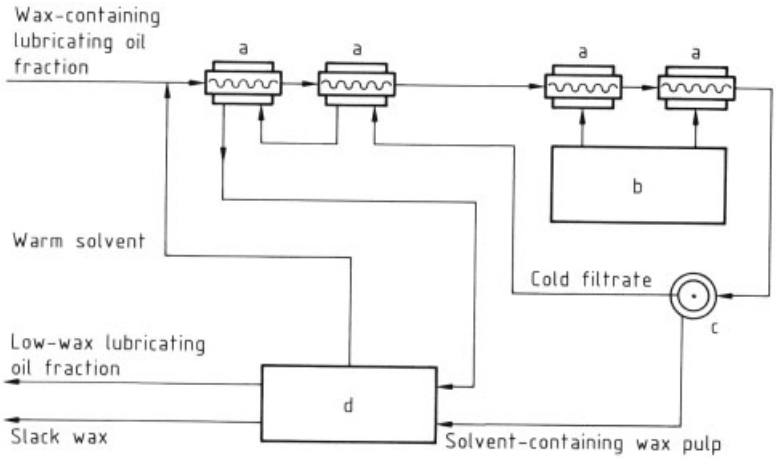

Figure 6. Dewaxing of lubricating oil distillates
a) Scraping chiller; b) Cooler; c) Vacuum rotating filter; d) Solvent recovery

solvent-neutral oils (i.e., the raffinates from solvent refining) are mixed with suitable solvents such as propane, naphtha, chlorohydrocarbons, ketones, or (most frequently) a toluene – methyl ethyl ketone mixture. The oil – solvent mixture is subsequently warmed to obtain a homogeneous solution and then cooled continuously in scraping chillers to obtain the wax crystals in a loose suspension. In the subsequent filtration, crystallized wax is separated in vacuum rotary filters, and washed with fresh solvent; the low-wax solvent-neutral oil is then freed from the solvent by distillation. The solvent-containing wax is either separated from the solvent by distillation if it is to be obtained as such or fed directly into the deoiling process for production of hard waxes.

Process conditions are controlled in such a wax that high throughput rates (i.e., short filtration times) and as low an oil content of the slack waxes as possible are achieved [88].

The selectivity of the dewaxing process is controlled by the cooling temperature in the scraping chiller and by the oil – solvent ratio. For common lubricating oil fractions the cooling temperature is between −50 and −20 °C at oil yields of 65 – 85 %. The most important process parameter is the targeted pour point of the dewaxed oil. Figure 6 shows a typical dewaxing process.

In dewaxing, all hydrocarbons with crystallization temperatures above the chosen cooling temperature are removed. *n*-Paraffins are thus almost completely separated, along with a suitable fraction of the long-chain, weakly branched isoparaffins. Even under optimum filtration and washing conditions, impurities consisting of highly branched isoparaffins and cycloparaffins are present in slack waxes, thus giving rise to their oil content [89].

A second process for producing lubricating oils with the required low-temperature properties is *catalytic dewaxing*. *n*-Paraffins are preferentially cracked on zeolite catalysts to give low molar mass hydrocarbons. Wax thus cannot be obtained from this process. Catalytic dewaxing is said to be very selective in the case of heavy distillates, whereas in light distillates the branched paraffins also react (loss of yield). Since no solvents are used in this process it does not require solvent recovery and is therefore environmentally friendly [90].

4.2.3.2. Deoiling Slack Waxes

Refining slack waxes begins with deoiling to obtain harder, higher-value products (hard waxes). The two main processes are solvent and sweat deoiling. The oils produced as byproducts, which contain a higher or lower proportion of soft waxes, are known as foot oils or soft waxes, depending on the consistency and the process used.

Solvent deoiling is the most commonly used process, involving one of three possible techniques: pulping, crystallization, or spray deoiling. Starting materials for all deoiling processes are either molten slack waxes (if deoiling is carried out in a location other than that of slack wax production) or slack wax – solvent mixtures, formed in dewaxing lubricating oil distillates (see Table 12).

Slack waxes from spindle oils to heavy machine oils, and to some extent petroleum as well (see Section 4.3), can be deoiled by the three solvent deoiling methods. The yields of deoiled slack waxes depend on their origin, the process used, and the degree of deoiling. Yields are between 80 % for spindle oil distillates and ca. 60 % for heavy machine oil distillates, based on the slack wax used.

Pulping Process. The inhomogeneous mixture of crystallized wax, oil, and solvent is diluted or repulped by addition of the same solvent or a solvent mixture (normally by using part of the wash filtrates produced later). The oils, soft waxes, and part of the low molar mass hard waxes thus dissolve. The pulp is cooled with stirring in scraping chillers, and the temperature is adjusted to that necessary for the desired degree of deoiling. Hard waxes, which are partly undissolved and have partly recrystallized, are removed on vacuum rotary filters and washed with solvent; the solvent is subsequently distilled off.

The quality of the hard waxes produced is determined by the deoiling temperature and the quantity of solvent used. With a 1,2-dichloroethane – methylene chloride mixture as solvent (Edeleanu Process), for example, typical deoiling temperatures are −10 °C for slack waxes from heavy spindle oil, −5 °C for those from a light machine oil, and +15 °C for those from medium to heavy machine oil.

Crystallization Process. The molten slack wax or solvent-containing wax crystallizate from the dewaxing process is dissolved completely, or almost completely, in solvents by warming, and the mixture is cooled to a given temperature in one or several steps, depending on the desired degree of deoiling. The temperature is chosen such that only the hard wax crystallizes. The crystallizate is filtered and washed, and the solvent is removed by distillation.

In both the pulping and the crystallization processes, technical problems arise in attempting to produce readily filterable hard wax crystals that allow high filtration capacity with good washability. These problems can be solved in various ways: e.g., through the nature and quantity of the solvent; the construction of the crystallization apparatus; successive solvent dosing during cooling; addition of cold solvent to the warm, intensively stirred wax – solvent mixture [91]; and use of filter aids [88]. Other improvements have been brought about by the development of economical cooling and filtration sysems [92].

Spray Deoiling. Molten slack wax is sprayed into a countercurrent of cold air, and precipitated wax particles are washed with solvents in mixers whereby the oil is largely diluted. After settling the washed wax particles are centrifuged and washed, and the adhering solvent is distilled off. The washing temperature is between 5 and 15 °C [93].

Sweat Deoiling. Molten slack wax is charged to chambers equipped with sieve bottoms and heating coils. The wax is solidified by cooling and then warmed very slowly. The oil – and, at higher temperatures, the low-melting soft waxes, as well – sweat from the wax block and run through the perforated bottom (run-off slack wax). At the end of the process the remaining hard wax is melted to remove it from the sweating chamber.

No new plants are being built for this classical deoiling process because of the low selectivity (poor yields of hard wax), time-consuming warmup, need to use a batch process, and inapplicability of the method to strongly oil-binding slack waxes from medium and heavy machine oil distillates. In existing plants an attempt is being made to improve yields of hard waxes by partially recycling the run-off slack [94].

Overview of Processes in Current Use. The deoiling processes most widely used in industry are listed in Table 12.

Of the deoiling plants operating in Germany, half employ solvent deoiling (using mainly the Edeleanu dichloroethane – methylene chloride

Table 12. Processes for deoiling slack waxes [95]

Company	Process	Solvent	Starting material
Exxon	crystallization process (Dilchill deoiling)	methyl ethyl ketone – methyl isobutyl ketone (or – toluene)	slack wax – solvent mixture *
Edeleanu	pulping process (Di-Me deoiling)	1,2-dichloroethane – methylene chloride	slack wax – solvent mixture *
Edeleanu	spray deoiling	1,2-dichloroethane	slack wax
Texaco	pulp process (one- and two-step)	benzene (or benzene – toluene)	slack wax – solvent mixture *
Texaco	crystallization process (wax fractionation)	methyl ethyl ketone	slack wax and slack wax – solvent mixture *
Union oil	crystallization process	water-saturated methyl isobutyl ketone	slack wax
	sweat deoiling		slack wax

* From the dewaxing step.

process) and the others use spray deoiling (Edeleanu) and sweat deoiling.

In the United States, more than 90 % of the existing plants use solvent deoiling. The most common solvents are methyl ethyl ketone and methyl ethyl ketone – aromatics mixtures. Propane and methyl isobutyl ketone are also used.

4.2.3.3. Refining Deoiled Slack Waxes

Deoiled slack waxes whose maximum oil content is 0.5 %, 0.5 – 1.5 %, or 2 – 3 %, depending on the use anticipated, still contain impurities and are mostly dark in color. To improve their quality they are purified further by adsorbents (decolorizing) or chemically. The choice of refining process depends on economic factors and the quality of raffinates required. Process combinations have also been introduced into industry. The preferred process in new plants is hydrotreating.

Refining with adsorbents (decolorizing) (→ Lubricants and Related Products) can be performed batchwise, semicontinuously, or continuously. The decolorizing temperature can be between 70 and 120 °C, depending on starting material.

In *batch decolorizing*, molten wax is stirred in heated vessels with decolorizing clays (chemically activated clays, bentonite, bauxite) in one or more steps until the desired lightening of color is achieved. The used clay is removed in heated filter presses and thermally regenerated.

Percolation Process. In the percolation process, molten wax flows downward through a tower containing 10 – 50 t of decolorizing clay, depending on plant size. After the clay has been exhausted, the wax flow is stopped; wax adhering to the clay is washed off with naphtha; the remaining solvent is driven off with steam; and the clay is burnt off in a separate tower, activated, and reused [96].

The *semicontinuous decolorizing process* operates with two or three percolator towers, which are in different phases of the adsorption cycle. While the adsorption process occurs in one tower, in the others the used decolorizing clay is washed, stripped, removed for regeneration, and replaced by new activated clay.

In the *continuous process* the regenerated clay flows downward through the adsorption tower in countercurrent to the molten wax. The loaded clay passes in the form of a sludge to a washing tower where adsorbed wax is extracted with naphtha and then to a second tower where naphtha is driven off wich steam. The clay is then burnt off in a rotary kiln, activated, and fed back to the top of the adsorption tower after cooling [97].

The decolorized waxes are then stripped with steam in packed columns at 110 – 170 °C, either under vacuum (0.6 – 1.6 kPa), under reduced pressure (up to 80 kPa), or under pressure (up to 0.8 MPa) to remove odorous substances. The weight ratio of steam to wax varies between 0.5 : 1 and 1 : 1 at detention times up to 15 min. To improve the aging resistance of the waxes, apparatus with aluminum or stainless steel lining is recommended for stripping [98].

Depending on the starting material, the wax qualities listed in Table 13 are obtained by the decolorizing process.

The quantity of decolorizing clay required depends on the origin and desired quality of

the waxes and is normally between 2 and 4 %. Refining with adsorbents removes dark substances, condensed alkyl aromatics, compounds containing hetero atoms, and metallic impurities.

Nevertheless, producing fully refined waxes fulfilling all of the purity requirements (see Section 4.2.5) by adsorption is possible only using extremely complex processes (multistage decolorizing) and large quantities of decolorizing clay, particularly in the case of the high molar mass wax fractions. To produce these types of high-purity waxes, chemical refining is the process of choice.

Table 13. Wax qualities obtained by decolorization

Starting material	Oil content, %	Decolorized product
Slack wax	>3	scale wax
Deoiled slack wax	1.5 – 3	semirefined wax
Deoiled slack wax	1.5 (max.)	filtered wax [*64742-43-4*]

Chemical Refining. Concentrated sulfuric acid, fuming sulfuric acid (oleum), or hydrogen is used as chemical refining agent.

Sulfuric acid refining can only be carried out batchwise because of the slow separation of the wax and acid sludge phases. Molten wax is mixed with sulfuric acid at 80 – 140 °C for some time (between 0.5 and 2 h, depending on the origin of the wax). After separation of the heavier acid sludge, the wax is washed with alkali, treated with decolorizing clay, and stripped with steam to remove the last residues of impurities and odorous substances.

All compounds that react chemically with the aggressive acid, such as unsaturated aliphatic and aromatic hydrocarbons, metal compounds, and those containing hetero atoms, are removed by sulfuric acid refining. Labile quaternary carbon atoms are also attacked, with bond cleavage and formation of aliphatic sulfonic acids.

The process can be optimized by variation of the number of reaction steps, degree of acidity, reaction temperature, ratio of wax to acid used, reaction time, and intensity of mixing. The higher the average molar mass of the wax fraction, the slower is the separation between the wax and the acid sludge phase.

The process has several disadvantages such as the necessity for batch operations, poor yields of raffinates, formation of polluting byproducts, and waste gas and corrosion problems. For these reasons, almost all new plants are based on hydrotreating.

Hydrotreating (Fig. 7). Deoiled slack wax is heated to the required temperature together with fresh and recycyled hydrogen in the preheater (b) and passed over a sulfur-resistant fixed-bed catalyst (c). After cooling, the reaction mixture and hydrogen are separated in the high-pressure gas separator (d), and hydrogen is recycled into the process. After depressurizing in the low-pressure gas separator (e), the wax passes into the stripper (f), where all the light crack products, odorous substances, and reaction gases are stripped off completely with compressed steam under vacuum. The wax is subsequently dried with nitrogen (g). A combination of metals from groups 6 and 10 on an inert carrier is generally used as the catalyst (e.g., nickel – tungsten or nickel – molybdenum on neutral aluminum oxide).

Reaction conditions can vary within wide limits, depending on starting material, degree of refining required, and composition of the catalyst – e.g., between 200 and 350 °C, 20 and 200 bar, and throughputs of 0.2 – 0.81 L of wax per liter of catalyst and hour. Yields are almost 100 %, based on starting wax [99], [100].

Under the severe reaction conditions of high-pressure hydrogenation, all aromatic compounds are hydrogenated to naphthenes, all dark substances are decomposed, and all sulfur and nitrogen atoms are removed as hydrogen sulfide and ammonia. Metallic impurities become bonded to the catalyst.

The process can be used to refine all classes of wax (paraffin waxes, intermediate waxes, and microwaxes).

4.2.4. Storage, Transportation, Commercial Forms, and Producers

Paraffin waxes are transported in liquid form (molten) in road tank trucks, rail tankers, or liquid containers. Solid waxes are marketed in slabs in small and large cartons and sacks, and as pastilles, flakes, and powder in sacks or big bags. They are also supplied in fiberboard and steel drums and in molded cartons.

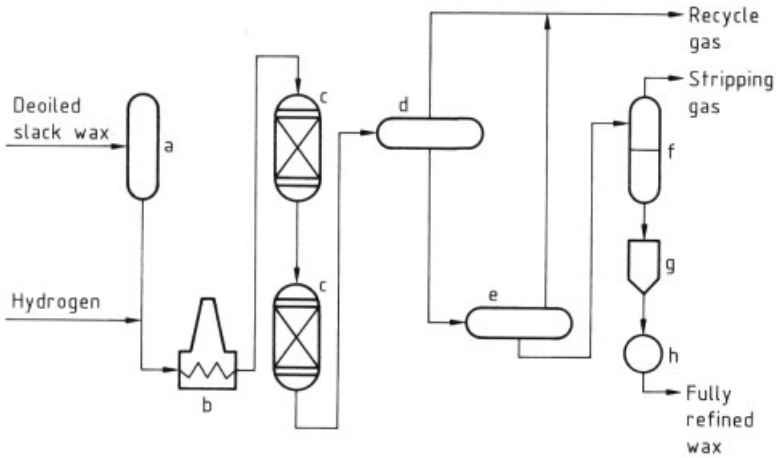

Figure 7. Two-step hydrotreating of waxes
a) Deaerator; b) Preheater; c) Reactors; d) High-pressure gas separator; e) Low-pressure gas separator; f) Stripper; g) Vacuum dryer; h) Filter

Even refined waxes can become yellow or decompose to form odorous substances on prolonged heating, particularly in the presence of air or catalytically active metals. On storage, transportation, and processing, the waxes are therefore preferably heated in vessels, fitted with warm water equipment, and the containers and piping are well insulated. Storage tanks are often blanketed with inert gas. In some cases, antioxidants are added to fully refined waxes in quantities up to 0.01 %. Deoiled and refined waxes are differentiated according to melting range (gradation). Some examples of commercial waxes are given below.

Europe	United States
(Melting gradation in °C)	(AMP∗ gradation in °F)
48/50	120 – 122
50/52	122 – 124
52/54	126 – 130
54/56	130 – 132
56/58	132 – 134
58/60	134 – 136
60/62	136 – 140
62/64	143 – 145

∗ AMP = American Paraffin Wax

Important producers of paraffin waxes include: Shell (United Kingdom, United States, Germany), Texaco (United States), BP (France, Germany), Mobil (United States, United Kingdom), Lützendorf (Germany), DEA (Germany), Wintershall (Germany), CFP (France), Total (France), Empetrol (Spain), Petrogal (Portugal). In Germany, for example, the companies either process the waxes to refined waxes (DEA, Wintershall) themselves or sell them to wax refineries (H. O. Schümann). Waxes are *sold* mostly by DEA, H. O. Schümann, and Wintershall.

4.2.5. Quality Specifications and Analysis

The quality specifications of paraffin waxes are determined mainly by their uses. They are differentiated in terms of physical characteristics, chemical composition, purity requirements, and application properties. The following standard specifications have been published for waxes that come in contact with foods:

1) Food and Drug Administration: Code of Federal Regulations (United States) [101]
2) FAO Food and Nutrition Paper (EC) [102]
3) Specifications of the European Wax Federation (EC) [86]
4) Kunststoffe im Lebensmittelverkehr (Plastics which come into contact with foods, Recommendations of the German BGA) [103]
5) Hydrocarbons in Food Regulations (United Kingdom) [104]

For waxes in cosmetic and pharmaceutical preparations, standards are found in the following publications: DAB (Germany), Codes

Français (CF) (France), B.P. (United Kingdom), ISCI Monograph (Japan), U.S. Formulary (United States), and CTFA International Cosmetics Ingredient Dictionary (United States).

The most common test standards for the determination of physical data are summarized in Table 14. The most important test standards for determining chemical composition can be found in Table 15. Some examples of important test methods for determination of the purity of waxes are listed below:

Fluorescence	49th communication of the BGA
Alkali- or acid-reacting substances	DAB, B.P., U.S.P.
Behavior toward sulfuric acid	DAB, BGA, CF, U.S.P.
Presence of polycyclic aromatics	DAB, BGA, FDA

Table 14. Standardized test methods for determination of physical data

Determination	Method	
mp (cooling curve), °C	DIN ISO 3841	ASTM D 87
Congealing point, °C	DIN ISO 2207	ASTM D 938
Needle penetration, 0.1 mm	DIN 51 579	ASTM D 1321
Oil content, %	DIN ISO 2908	ASTM D 721
Viscosity, mm^2/s	DIN 51 562/T1	ASTM D 445
Density, kg/m^3	DIN 51 757	ASTM D 1298
Refractive index	DIN 51 423/T2	ASTM D 1747
Color		
ASTM color index	DIN ISO 2049	ASTM D 1500
Lovibond color	IP * 17	
Saybolt color	DIN 51 411	ASTM D 156
Odor		ASTM D 1833

* IP = Institute of Petroleum, United Kingdom.

Table 15. Standardized test methods for determining chemical composition

Determination	Method	
Ash content, %	DIN EN 7	ASTM D 482
Sulfur content, ppm	DIN 51 400/T7	ASTM D 2622
Heavy metal content (As, Cr, Cd, Pb), ppm	XRF [a] analysis	
C-number distribution	DGF [b] M-V9	EWF [c] method
C-number max.	DGF [b] M-V9	IP [d] draft
n-Paraffin content, %	DGF [b] M-V9	ASTM draft

[a] XRF = X-ray fluorescence spectromety.
[b] DGF Einheitsmethoden der Deutschen Gesellschaft für Fettwissenschaften [Standardized methods of the Deutsche Gesellschaft für Fettwissenschaften (German Fat association)].
[c] EWF = European Wax Federation.
[d] IP = Institute of Petroleum, United Kingdom.

A looseleaf collection of national and international purity regulations for waxes has been issued by the European Wax Federation and is supplemented continuously [105].

Some of the methods cited are basic test procedures for gaining accreditation according to the series of standards DIN EN 45 000 and the regulations of the Deutsche Akkreditierungsstelle Mineralöl GmbH (DASMIN). With accreditation, DASMIN recognizes the competence of testing laboratories to test the quality of waxes.

The determination of *application properties* is described mostly in in-house methods and includes, for example, oil binding properties, contraction, polishability, abrasion resistance, solvent retention, blocking point, steam permeability, flexibility, extendability, and color stability. Melting and congealing characteristics and latent heats are determined by using differential thermoanalysis [standardized DGF method M-III 14 (1989); ASTM-D 4419-84].

Special analytical methods have been developed to determine the *chemical composition* of paraffin waxes.

The content of *n*- and isoparaffins is determined most accurately by using computer-supported high-temperature gas chromatography [106], [107]. Here, the DGF method M-V9 (1988) is the universal standard. This method allows not only sufficient separation of the *n*- and isoparaffin signals (depending on the column length and injector temperature, waxes up to a C-number of ca. 80 can be determined quantitatively), but also determination of the C-number distribution and position of the C-number maximum.

The determination of *n*- and isoparaffins using urea adduct formation and molecular sieve adsorption has been shown to be a nonquantitative separation method. Only enrichment of the *n*-paraffins results. This, however, clearly diminishes with increasing molar mass and complexity of the waxes investigated [108].

For individual analysis of the different, paraffinic substance groups, coupled GC – MS and GC – NMR methods have been described [85], [109].

The determination of individual polycyclic aromatic hydrocarbons in waxes is currently possible into the μg/kg (ppb) region [85], [110], [111].

Table 16. Uses of paraffin waxes worldwide, 1992 (in 1000 t)

Use	United States	Asia	Western Europe	of which Germany
Candles and decorative modeling	100	450	230	55
Paper, cardboard, wood	450	220	120	25
Hot melt adhesives	60	20	20	10
Rubber and cables	40	20	30	10
Industrial products	50	30	50	
Pharmaceuticals and cosmetics	20	30	10	
Other uses	130	130	40	50 ∗

∗ Including industrial products, pharmaceuticals and cosmetics

4.2.6. Uses

Paraffin waxes are used in very different fields of industry. They not only are used as such, but are also important components of wax compounds (i.e., mixtures of paraffin waxes with other waxes, plastics, resins, etc.). The consumption figures listed in Table 16 cover the most important areas of use.

Slack waxes that are not processed further to refined waxes or subjected to dehydrogenation or cracking processes – for example, to give α-olefins (→ Hydrocarbons, Chap. 2.1.2.1.) – are used mostly in the wood and mineral oil processing industries.

Wood processing requires slack waxes, either molten or in the form of emulsions, mainly for the production of particle board and fiberboard. In the United States the production of artificial logs for barbecues and fires is also important.

In the *mineral oil processing industry*, slack waxes are used to produce petroleum jellies, wire-drawing agents, lubricants, spray emulsions, and anticorrosives.

In addition to wax industry products with a wide area of applications, slack waxes are also used to produce chlorinated paraffins (→ Chlorinated Hydrocarbons, Chap. 7.3.), papers and cardboards for technical uses, cable coverings, solid anticorrosives, fire lighters, and wax torches.

Soft waxes have only a limited market in the production of industrial petroleum jellies and emulsions (wood and building industry), torches, and fire lighters and for conditioning fertilizers. The deoiling of soft waxes gives highly plastic and extendable products (isowaxes).

Refined Waxes [Scale Waxes, Filtered (Decolorized), Semi- and Fully Refined Waxes]. In Germany the producers of candles and wax goods (pictures, sculptures, ornaments) are the main users of refined waxes, which can be processed by all common production methods for candles (molding, drawing, extrusion molding, and powder compressing). In terms of quantity, paraffin wax is the most important raw material for the candle industry.

The second most important consumer of refined waxes is the wax industry, which requires paraffin waxes as an important raw material for compounding wax products for the most widely varying applications.

In the paper, cardboard, and packaging industries, refined waxes are used in the form of emulsions for paper sizing as antifoams in the production of paper, cardboard, and corrugated cardboard; and as impregnating and coating waxes for paper and packaging materials [112].

The viscosity of heat-seal waxes and hot-melt adhesives is adjusted with paraffin waxes. In rubber articles, tires, and solid cable coverings, waxes act as external lubricants/plasticizers, mold-release agents, and rapidly migrating protection agents of rubber against sunlight. They are therefore used predominantly for statically stressed rubber articles (profiles, gaskets, linings, and cable jackets).

Other important applications are in cleansers and polishes (→ Cleansing Agents, Chap. 1.4.; → Shoe Polishes, Chap. 3.1.); cosmetics (creams, ointments, lipsticks); matches; plastics

(external lubricants); office papers; and waterproof textiles (dry impregnation).

In the United States, by far the greatest proportion of refined waxes is used for paper and cardboard finishing, mainly for deep-freeze packaging and wrapping for meat, candy, and flowers; paper cups; yogurt and soft cheese pots; and corrugated cardboard [113].

In the former Soviet Union, refined waxes still play an important role in the production of synthetic fatty acids via catalytic oxidation processes (→ Fatty Acids, Chap. 4.1.). Fatty acids are mainly used in the soap industry.

4.3. Microcrystalline Waxes (Microwaxes)

4.3.1. Chemical Composition and General Properties

Like paraffin waxes, microcrystalline waxes consist of a mixture of saturated hydrocarbons that are predominantly solid at room temperature, such as *n*- and isoalkanes, naphthenes, and alkyl- and naphthene-substituted aromatics. Unlike paraffin waxes, isoparaffins and naphthenic compounds predominate here. The microcrystalline structure can be explained by the presence of strongly branched isoparaffins and naphthenes, which inhibit crystallization.

General Physical Properties. Microcrystalline waxes are insoluble in water and most organic solvents at room temperature. They are moderately to readily soluble in solvents such as chlorohydrocarbons, benzene, toluene, xylene, solvent naphtha, and turpentine oil, especially at elevated temperature. Solubility decreases markedly as molar mass increases. Solvents and oils are retained very strongly by microcrystalline waxes and therefore evaporate very slowly. The quality of some consumer products, such as petroleum jellies or floor and shoe polish, is determined by this retention capacity of microwaxes.

Chemical Properties. Microwaxes are more reactive than paraffin waxes because of the higher concentration of complex branched hydrocarbons with tertiary and quaternary carbon atoms. These C–C bonds are not very thermally stable (i.e. the waxes darken and resinify) on prolonged heating. In addition, they form black tar-like substances on contact with aggressive chemicals such as concentrated sulfuric acid or antimony pentachloride.

The reaction of microwaxes with oxygen at elevated temperature and in the presence of catalytically acting heavy-metal soaps is used for the production of oxidized microwaxes.

4.3.2. Division into Product Classes

Depending on the degree of refining, microwaxes are divided into the following classes:

1) Bright stock slack waxes (petrolatum)
2) Plastic microwaxes
3) Hard microwaxes

Table 17 and Figure 8 show some typical characteristics of microwaxes.

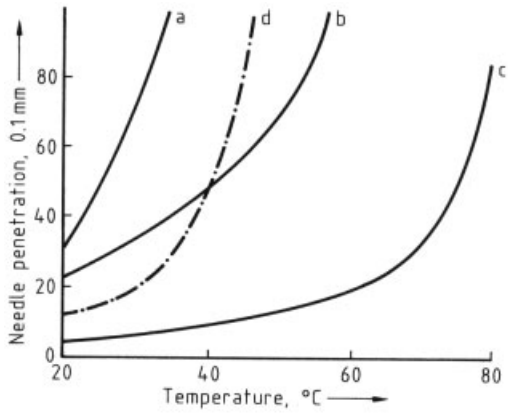

Figure 8. Variation of needle penetration with temperature for various microwaxes compared with a paraffin wax
a) Bright stock slack wax; b) Plastic microwax; c) Hard microwax; d) Paraffin wax 60/62

Bright stock slack waxes (petrolatum) constitute the most complex mixture in the group of microcrystalline waxes. The molar masses of the hydrocarbons can be between 320 and 1200 g/mol. The *n*-alkane content is seldom >10 %. Asphaltenes, hydrocarbon resins, and sulfur, nitrogen, and oxygen compounds are present as impurities. The oil content is normally between 3 and 15 %, but it can be as high as 30 %. Bright stock slack waxes are white

Table 17. Typical physical characteristics of microwaxes

Characteristic	Bright stock slack wax	Plastic microwax		Hard microwax
Congealing point, °C	67	72	75	87
Drop point, °C	76	78	80	91
Needle penetration, 0.1 mm				
at 25 °C	65	30	28	8
at 40 °C	140	85	66	18
Oil content, %	3.8	1.8	1.6	0.4
Viscosity (at 100 °C), mm^2/s	18.7	15.7	14.9	16.2
Refractive index (at 100 °C)	1.4505	1.4458	1.4412	1.4410
Color	dark brown	yellow	white	white

or yellow to dark brown, greasy to softly plastic, unctuous, smooth products with a typical odor, depending on the refining process. They have a very good oil-staining capacity, and some grades have a pronounced stickiness. The congealing points are between 60 and 75 °C; the drop points between 62 and 80 °C; and the hardness (needle penetration at 25 °C), between 45 and 160, depending on the oil content. The viscosities (10 – 30 mm^2/s at 100 °C), densities (810 – 860 kg/m^3 at 80 °C), and flash points (290 – 320 °C) are clearly higher than those of paraffin waxes.

Plastic Microwaxes. In plastic microwaxes the higher molar mass *n*- and isoparaffins and the weakly branched isoparaffins and naphthenes are enriched compared with the petrolatum from which they are obtained by deoiling. Molar masses are between 450 and 1200 g/mol. The *n*-paraffin content can be as high as 40 %. Depending on the degree of refining the impurities present in petrolatum are removed completely or to a great extent. The oil content is usually between 1.5 and 5 %.

The plastic microwaxes, although relatively hard, are definitely plastic, readily moldable, and kneadable. The plasticity is due to the presence of very long carbon chains in the *n*-paraffins and terminally branched isoparaffins. The waxes feel dry to slightly adhesive, and are sold as white to brown products. Congealing points are 65 – 80 °C, and the hardness (needle penetration at 25 °C) is 10 to 60. Highly refined microwaxes are odorless and tasteless.

Hard microwaxes also essentially consist of high molar mass alkanes (predominantly *n*-alkanes and isoalkanes with short side chains statistically distributed over the main carbon chain) and naphthenes. Average molar masses are between ca. 500 and 1200 g/mol. Thus, for example, a refined, hard microwax of North American origin contains 89 % nonpolar hydrocarbons (*n*-, iso-, and cycloparaffins) and has an average C-number of 48 [85]. The *n*-paraffin content can be as high as 70 %.

The extremely microcrystalline to almost amorphous structure of hard microwaxes is associated with the formation of supercooled melts during the setting process.

The waxes are tough and hard, barely ductile and weakly adhesive. Depending on the degree of refining, hard white, yellow and brown microwaxes are on the market, the white to pale yellow products being of highest purity. They solidify between 80 and 94 °C. The needle penetration at 25 °C is <10. The hardness of hard waxes decreases only slightly with temperature until close to their softening point (Fig. 8).

4.3.3. Occurrence of Raw Materials and Processing

Like paraffin waxes, microwaxes are components of crude petroleum. Because of their high molar masses, and thus high boiling range and poor solubility, they either are enriched in the vacuum residues (short residues) from lubricating oil distillation (residual waxes) or separate during the transportation and storage of crude oils (settling waxes). Settling waxes form pasty to solid deposits on the walls of pump tubes and pipelines and on the bottom of storage tanks (pipe and tank bottom waxes). Vacuum residues and pipe and tank bottom waxes are the raw materials for recovery of microcrystalline waxes.

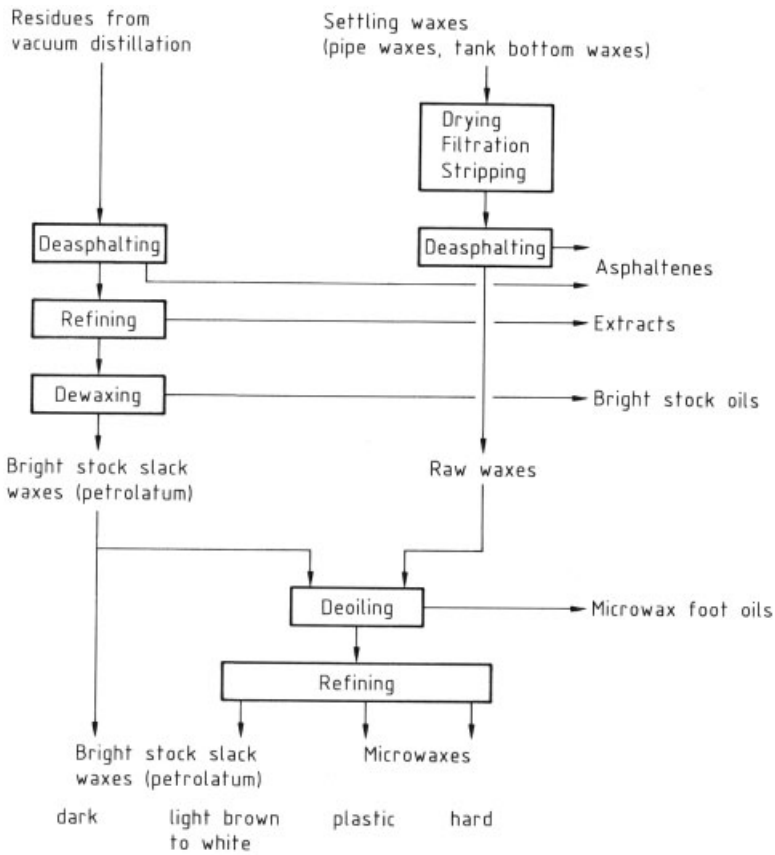

Figure 9. Process schematic for the production of microwaxes

Figure 9 shows schematically the process steps involved in the production of microcrystalline waxes.

In the United States, the most important producer country for microwaxes, the vacuum residues are subjected to deasphalting, refining, and dewaxing to produce high-value bright stock oils. Bright stock slack waxes (petrolatum) are produced in the process as economically important byproducts. However, only in the rarest cases they are processed further to microwaxes in the lube oil refineries themselves. Petrolatum is either marketed directly or used, for example, for the production of petroleum jellies. For further processing it is sold to companies specializing in the production of microwaxes. The same companies also process settling waxes purchased from oil companies.

Deasphalting is the first step in the processing of tank bottom waxes, pipe waxes, and residues, which have been freed from water, mechanical impurities, and volatile components. When propane is used as the deasphalting agent, considerable proportions of resins, polycyclic aromatics, and nitrogen and sulfur compounds are removed. To improve the selectivity with as low a propane throughput as possible, rotating disk contactors are now used instead of packed or sieve bottom columns (→ Lubricants and Related Products). In addition, the extraction process is carried out in two steps [114].

Deasphalting of residues is followed by solvent extraction and dewaxing. The products formed here are the bright stock oils and bright stock slack waxes (petrolatum).

Deoiling. The crude and bright stock slack waxes thus obtained are deoiled by fractional

crystallization (see page 32) if they are to be processed further to microwaxes. Preferred solvents are methyl ethyl ketone and methyl isobutyl ketone, often as a mixture with benzene or toluene. Propane and naphtha, frequently used earlier, have become less important because of their low selectivity.

The crude waxes (deasphalted settling waxes) and bright stock slack waxes are mixed either molten or as a pulp (if the deoiling is to follow immediately after dewaxing the brightstock oils), without isolation of the slack wax with fresh, warm solvents and warmed until only the less soluble (high molar mass) waxes remain undissolved. These waxes are then filtered, washed, and freed from solvent (hard microwax fraction). To obtain plastic microwaxes the filtrate is cooled further. The properties of reprecipitated and refiltered waxes are determined by the filter temperature. The low molar mass, very plastic, and soft microwax foot oils [*64742-67-2*] remains in the filtrate. If the hard microwaxes are not extracted, a one-step crystallization process is sufficient for obtaining plastic microwaxes.

Deoiled microwaxes are still dark yellow to dark brown and not completely odorless, so they must subsequently be refined to satisfy a broad application range.

Refining. Refining is carried out either with adsorbents (decolorizing) or with hydrogen (hydrotreating). Refining with concentrated sulfuric acid, which is possible in principle, has not become important economically because of the high loss of material.

For refining microwaxes with *adsorbents*, the same process principles apply as for decolorizing paraffin wax (see page 33). Since microwaxes are more difficult to decolorize than paraffin waxes, countercurrent percolation is used predominantly. Here, the wax melt continually comes in contact with fresh decolorizing clay [115]. Bauxite is also frequently used as a decolorizing agent.

Hydrotreating is carried out as described for paraffin waxes, in one or two steps in reactors packed with sulfur-resistant, heavy-metal-doped mixed catalysts (see page 34). Reaction conditions are ca. 350 °C and 150 bar, depending on the properties of the catalyst, at a space velocity of 0.3 L of wax per liter of catalyst and hour. Hydrogenated products are treated further by adsorptive decolorizing or stripping to completely remove odorous materials and volatile components.

In the United States, a whole range of microwaxes is produced according to the processed described. Products differ in their congealing point, drop point, hardness, viscosity, and color. They are sold in the liquid (molten) form and as solids in the form of slabs, pellets, and powder.

4.3.4. Commercial Products and Producers

Table 18 shows the most important producers of microwaxes in the United States and the corresponding commercial products.

Table 18. Producers of microwaxes in the United States and commercial products

Producer	Production site	Commercial products
Bareco Div. of Petrolite Corp.	Kilgore, Tex. Barnsdale, Okla.	Petrolite Crown, Be Square, Mekon, Stawax, Cardis, Victory, Ultraflex
Mobil Oil Co.	Beaumont, Tex. Paulsboro, N.J.	Wax Rex, Mobilwax
Quaker State Oil Ref. Corp.	Emlenton, Pa. Farmers Valley, Pa. Newell, W.V.	Microwax, Superflex
Shell Oil Co.	Deer Park, Tex. Wood River, Ill.	Shellwax
Sun Oil Co.	Tulsa, Okla.	Sunoco Wax, Sunolite
Witco Chem. Corp.	Bradford, Pa.	Multiwax

In Europe, H. O. Schümann (Germany), Mobil (United Kingdom), CFR/Total, Mobil, and BP (France), Shell and Witco (Holland), Repsol (Spain), and Petrogal (Portugal), for example, produce microwaxes by deoiling bright stock slack waxes.

4.3.5. Quality Specifications and Analysis

The quality required depends on the industrial use. In principle, the specifications given for paraffin waxes apply here.

Basic product specifications can be found in the following sources:

1) Food and Drug Administration; Code of Federal Regulations (United States) [101]
2) European Wax Federation (EC) [86]

3) Kunststoffe im Lebensmittelverkehr; Empfehlungen des Bundesgesundheitsamtes Deutschland (Plastics for Contact with Foods; Recommendations of the Federal Health Authority, Germany) [103]

The EWF specifies microwaxes for use in packaging materials for foods, as shown in Table 19.

Table 19. Microwaxes for food packaging

Determination	Method	Specification
Congealing point, °C	DIN ISO 2207 ASTM D 938	65 (min.)
Needle penetration, 0.1 mm	DIN 51 579 ASTM D 1321	<50
Extractables, %	ASTM D 721	2.5 (max.)
Viscosity at 100 °C, mm^2/s	DIN 51 562/T1 ASTM D 445	>13
Density at 100 °C, kg/m^3	DIN 51 757 ASTM D 1298	790 – 840
Refractive index at 100 °C	DIN 51 423/T2 ASTM D 1747	1.430 – 1.450
Color	DIN ISO 2049 ASTM D 1500	2.0 (max.)
Odor	ASTM D 1833	weak
Ash content, %	DIN EN 7 ASTM D 482	<0.05
Sulfur content, ppm	DIN 51 400/T7 ASTM D 2622	<300 [a] <4000 [b]
Heavy-metal content (As, Cd, Cr, Pb), ppm	XRF analysis [c]	<1
C-number distribution of *n*-paraffins	DGF M-V9 (88) IP 372-85	23 – 85
Non-*n*-paraffin content, %	DGF M-V9 (88)	25 – 75
UV-Absorption	§ 172.886 FDA	
at 280 – 289 mm		0.15 (max.)
at 290 – 299 mm		0.12 (max.)
at 300 – 359 mm		0.08 (max.)
at 360 – 400 mm		0.02 (max.)

[a] Hydrotreated.
[b] Refined with sulfuric acid/decolorizing clay.
[c] XRF = X-ray fluorescence spectrometry.

In the use of microwaxes to produce food packaging materials, and cheese and chewing gum waxes, purity and physiological nontoxicity are particularly important. For microwaxes the absence of specific polycyclic aromatics is required. These substances can occur as impurities in insufficiently refined waxes, and some of them have been found to be carcinogenic. Polycyclic aromatics can be detected by UV extinction after a special enrichment process (detection limit in microwaxes 10^{-7} g per gram of wax substance). Microwaxes have been shown to be toxicologically and physiologically harmless in extensive animal experiments, when UV adsorption values specified according to test regulations of the FDA [101] and the BGA [103] are not exceeded.

The physical characteristics of microwaxes are determined according to the same standardized methods used for paraffin waxes (see Table 14). In addition, some special analytical methods have been introduced, for example:

Determination of gloss and gloss stability	ASTM D 2895
Determination of blocking point	ASTM D 3234
Determination of sealing stability	TAPPI [Tech. Assoc. Pulp & Paper Ind. (United States)] T 642
Determination of water vapor permeability	ASTM D 988 DIN 53 413
Determination of specific penetration resistance	DIN 53 482
Determination of dielectric loss factor	DIN 53 483

Oil-binding capacity, adhesiveness, and light stability are additional quality factors.

4.3.6. Uses

Bright Stock Slack Waxes. The greatest proportion of bright stock slack waxes is processed further to microwaxes. Important consumers are the wax-producing industry, which requires crude and refined bright stock slack waxes for compounding adhesive waxes, cheese waxes, chewing gum base, cosmetic preparations, sealing and cable compounds, impregnating insulating materials and pesticide sprays; and the mineral oil processing industry, which produces anticorrosives, cavity and underfloor protection agents for motor vehicles, and petroleum jellies. Other uses include the production of artificial fire logs and oxidized petrolatum, which is suitable for the formulation of rust preventives.

Oxidized petrolatum and microwaxes are mainly produced industrially in the United States. Air at normal or at slight overpressure is blown through the wax, which is heated to 150 – 180 °C until the desired degree of oxidation is reached. Darkening of the wax is avoided by adding fresh wax to the reaction zone. Oxidized microwaxes are specified and traded according to their melting range and degree of oxidation (expressed in terms of acid

and saponification numbers). The acid and ester groups, which are distributed statistically over the hydrocarbon chains, can be reacted with inorganic and organic bases to form soaps that can be dispersed in water (emulsified waxes). These compounds also impart to the oxidized waxes an excellent adhesion to metals.

Plastic Microwaxes. The main consumer of plastic microwaxes is the wax industry. Plastic microwaxes are added to wax compounds that require either pronounced plastic or oil- and solvent-staining properties for their industrial applications.

Typical applications include the following:

1) Paper and cardboard coating for packaging meat and sausage products, cereals, bread, candy, ice cream, cookies, yogurt, and frozen food
2) External lubricant (sliding) for extrusion of plastics
3) Formulation of cheese waxes and chewing gum bases
4) Decorative for candles, production of wax pictures and sculptures
5) Polishes and grinding agents
6) Lost wax casting and dental compounds
7) Foam regulators for detergents
8) Protectives against crack formation in tires and rubber articles (sun-proof waxes) [116]
9) Additives for explosives and propellant materials.

Hard microwaxes have numerous applications based on their high melting points and great hardness; favorable hardness – temperature behavior; and good solvent retention. They are used as release agents for pressing plastics and fiberboard, in the production of ceramic articles and in molding polyurethane foam, in solvent-containing cleansing agents, in grinding and polishing pastes, and flat varnishes and paints.

The Committee on Toxicity (COT) in the United Kingdom has recommended that only hard microwaxes be approved for use in the production of cheese waxes, poultry defeathering compounds, and chewing gum [117].

On an industrial scale, hard microwaxes are *oxidized* to an acid number of 30 and are used in this form to produce emulsions (polishes and release agents) and as additives for paints and coatings, printing inks, and office transfer papers.

4.4. Legal Aspects

Apart from environmental effects, petroleum waxes can be absorbed by the human organism via foods (paraffins either are contained in them naturally, have been added to them, or have migrated there from the packaging materials), from chewing gum (only on swallowing or the simultaneous consumption of fat-solubilizing chocolates), and through the use of lipstick. To protect human health, national and international regulations and recommendations have been enacted to limit the oral consumption of waxes.

The use of waxes in packaging materials is subject of *EEC legislation* [118]. In the appendix lists (lists of additives), waxes are permitted, provided that specific migration limits of 0.3 mg per kilogram food for waxes treated with sulfuric acid or decolorizing clay and 3 mg per kilogram for high-pressure hydrotreated waxes are not exceeded.

Extensive literature on the migration of paraffins into foods has been published [119–122].

The permitted paraffins must correspond to published specifications (see Section 4.2.5).

In *Germany* the addition of petroleum waxes to food and animal feed is not allowed. The use of waxes as cheese coatings and in chewing gum is controlled by supplementary regulations.

In the *United Kingdom* the Ministry of Agriculture, Fisheries, and Food (MAFF) proposes a prohibition of petroleum waxes as food additives in England, Scotland, and Northern Ireland [117]. Only specified microwaxes may still be added to foods for a limited period.

In the *United States*, not indication has been given of any intention to ban paraffins in food.

4.5. Ecology

Paraffin and microcrystalline waxes are insoluble in water and can be separated as solids, together with sludge, in oil separators and sewage treatment plants. In the German catalogue of substances that are hazardous to water, they are therefore listed in water hazard class 0 (Wassergefährdungsklasse, WGK 0).

Since paraffins have a water solubility of $< 1\,\mu g/L$, they have no toxic effects on aquatic organisms. Used paraffin-containing packaging materials can be combusted without any problem in waste incineration plants. No polluting substances are formed, and the high calorific value of the paraffins improves the energy balance of the plants.

The *biodegradability* of petroleum waxes satisfies OECD guideline 301 B (1981) [123]. With the exception of microwaxes, biodegradabilities of more than 60 % after 28 d were measured. Bacterial degradation mechanisms are described in [124]. According to this, paraffin-containing packaging materials can also be composted [125–127]. No substances that are toxic or hazardous to groundwater are formed.

The biodegradation of waxes under anaerobic conditions (covered landfills) proceeds much more slowly.

A large number of investigations have been carried out on the *recycling* of used wax-containing packaging materials. The water-insoluble, hydrophobic petroleum wax tends to form particles in the paper mills and give rise to blockages in the sizing equipment [128].

In Europe, ca. 20×10^6 t of waste paper was recycled as a raw material for paper production in 1991. This figure includes 0.35×10^6 t of wax-containing paper and cardboard with average wax content of 12 %. Assuming an even distribution, this gives a wax content of 0.21 % wax in wastepaper [129]. The practice has shown that in modern wastepaper processing plants, wax-containing packaging materials do not interfere with the dilution process,but nearly increase the quantity of waste products. These waste products can be used to produce water-resistant cardbord.

4.6. Economic Aspects

In 1992, 9.5×10^5 t of paraffin and microcrystalline waxes were *produced* in the United States, 8×10^5 t in Asia, 6.5×10^5 t in Western Europe, 2.6×10^5 t in Germany alone, 2.0×10^5 t in Eastern Europe. and 4.0×10^5 t in the rest of the world.

In industrialized countries, a stagnation (probably a decrease) in petroleum wax production is anticipated for the coming years, independent of the growth in demand. This is because of the further development of car engine and tribology technology (reduced specific lubricant consumption and longer periods between oil changes), an increase in the proportion of partially or fully synthetic and biodegradable lubricants, the recycling of used oils, and the closing down of excess capacity in distillation plants despite an increase in the number of motor vehicles and the distances traveled [130].

However, in Asia, particularly China and Malaysia, an increase in wax production to 1.5×10^6 t in 2000 is predicted, because these waxes can be inexpensively produced from indigenous wax-rich crude petroleum by direct dewaxing. According to careful estimates, worldwide petroleum wax production will be ca. 3.8×10^6 t in 2000 (40 % in Asia, 25 % in the United States, 17 % in Western Europe).

The worldwide *consumption* of petroleum waxes was ca. 2.8×10^6 t in 1992, of which (in million tonnes) 0.5 was in Western Europe, 0.2 in Eastern Europe, 0.85 in the United States, and 0.9 in Asia.

4.7. Candles

According to German commercial quality class RAL 040 A2 (Issue 02/1993), candles are sources of heat and light and consist of a wick and a solid combustible material surrounding the wick. Therefore tea lights, lights for graveyards, and oil lamps are also included in this category.

The first candles have been traced to shortly after the birth of Christ. With Christianity, and later with the rise of principalities and the upper middle class, the use of candles developed rapidly. Whereas initially, tow wicks, beeswax, tallow, and spermaceti were used for candle production, from the 19th century onward the use of cotton wicks, petroleum wax, and stearin permitted the problem-free handling of candles.

Today, refined (petroleum) wax is the main raw material used in candle production. Candles are also made from stearin, beeswax, composites (mixtures of various waxes), and ceresin (mixtures of paraffin and hard waxes). Candles are dyed with wax-soluble organic aniline dyes (full dyeing), organic pigments, and (less often) natural pigments (dip dyeing). Fashionable candles with a colored dipping-paint coating are also available. Wicks consist of twisted cotton

threads and are sold as round or flat wicks of different strengths. The wicks are twisted and prepared so as to guarantee uniform burning, good wick curling position, and good suction capacity for the molten combustible material and to prevent afterglow of the wick once the candle has been extinguished. Candles are sometimes perfumed.

A range of mechanical and hand-worked processes exist for industrial production of candles. In the *drawing process* a candle strand is produced by passing the wick several times through the hot wax bath. The strand then passes onto the cutting machine, where it is cut to the desired length of the candle. With continuously operating candle-drawing machines, production of more than 10 000 candles per hour is possible. In the *molding process*, molten wax is poured into molds, in the middle of which the wicks are held taut. Molding machines are available in different sizes. Modern rotating molding machines have an hourly capacity up to 10 000 candles. The *powder pressing process* uses powdered wax as the starting material, which is either molded together with the wick in an extruder to form a candle strand or pressed into molds in a stamp press. At the same time, the wick is introduced with a tubular needle. Another process exists, known as the *dunking process*. Here, the wick is dipped repeatedly into molten wax.

Blank candles are made pointed by using milling cutters and are coated with dyed dip-coating material if necessary.

The quality control of candles involves, for example, determining their weight and dimensions, burning period, and burning properties (no drips or soot). The production of candles in Germany (East and West) was 80 000 t in 1988, and 17 000 t were imported. Thus, with an export of 22 000 t, a total of 75 000 t was consumed within Germany.

5. Fischer – Tropsch Paraffins

Production. The sole producer of Fischer – Tropsch paraffins is South African Coal, Oil and Gas (SASOL) in the Republic of South Africa (→ Coal Liquefaction, Chap. 2.2.). The formerly used ARGE (Ruhrchemie/Lurgi consortium) process, which employed iron catalysts arranged as a fixed bed in multitubular flow reactors, was changed to the SASOL slurry bed proces (SSBP), which afforded considerable cost savings. Production capacity has been doubled [131].

In the slurry or bubble column reactor, finely divided iron catalyst (average particle size 50 µm) is suspended in a Fischer – Tropsch wax with low viscosity at the reaction temperature, and synthesis gas (a mixture of carbon monoxide and hydrogen) is bubbled through. At reaction temperatures of 220 – 240 °C and pressures around 2 MPa, *n*-alkanes and *n*-alkenes are formed preferentially, with up to 40 % crude wax in the synthesis product.

Crude wax is separated from the other products of synthesis by fractional condensation. The low-boiling constituents (gasoline and diesel fuel) are removed by distillation under atmospheric pressure, followed by vacuum distillation to separate the *soft waxes*. Besides *n*-alkanes, distillation bottoms also contain alkenes, hydrocarbons with hydroxyl and carbonyl groups, and colored components. For purification and stabilization, the bottom products are subjected to a hydrofining step employing a nickel catalyst to yield a white *hard wax* that is practically free of alkenes, aromatics, functional hydrocarbons, and sulfur compounds.

Soft waxes are also subjected to hydrogenation.

Properties. Fischer – Tropsch waxes consist essentially of *n*-paraffins with chain lengths between 20 and 50 carbon atoms. Products with an average molar mass of 400 g/mol are marketed as soft, and those with an average molar mass of 700 g/mol as hard waxes. The paraffins have a fine crystalline structure and, because of the narrow molar mass distribution, a small melting range and very low melt viscosities. Congealing point, density, and hardness increase with increasing mean molar mass. The low molar mass compared with polyolefin waxes is the reason for a certain displaceability of the crystal layers relative to each other and, associated with this, the polishability.

Synthetic paraffins are fully compatible with refined waxes, polyolefin waxes, and most vegetable waxes. They are soluble at elevated temperature in the usual wax solvents (e.g., naphtha, turpentine, and toluene) to give clear solutions. Addition of 10 – 20 % synthetic paraffin

to other waxes increases their congealing point and hardness without significantly influencing melt viscosity. In wax pastes, the tendency of Fischer – Tropsch waxes to form microcrystals increases solvent retention. Some typical data for Fischer – Tropsch waxes are listed in Table 20.

Table 20. Typical properties of Fischer – Tropsch soft and hard paraffins

Characteristic	Sasol Wax M	Sasol Wax H2 Vestowax SP 1002
Drop point, °C		106 – 112
Needle penetration, 0.1 mm		1 – 3
Color	white	white
Molar mass, g/mol	400	700
Density (at 23 °C), g/cm^3	0.94	0.94
Viscosity (120 °C), mPa · s	<20	<20

Uses. Fischer – Tropsch waxes are used in plastics processing as lubricants for poly(vinyl chloride) and polystyrene, as well as mold-release agents; as melting point improvers, hardeners, and viscosity reducers in hot melts and candles; and, because of their good polishability, for the production of cleaning agents and polishes.

Micronized waxes improve the abrasion resistance of paints and printing inks.

Oxidized waxes containing fatty acids and fatty acid esters are produced from synthetic paraffin by oxidation and partial saponification. Major areas of application are as mold-release agents in plastics processing, in polishes and cleaning agents, and as auxiliaries in the textile and paper industries.

Trade Names. Fischer – Tropsch waxes are marketed, e.g., as Sasol Wax (Sasol Marketing Co., Johannesburg, South Africa) and Vestowax SH/SP (Hüls AG, Marl, Germany).

6. Polyolefin Waxes

6.1. Production and Properties

6.1.1. Polyethylene Waxes by High-Pressure Polymerization

High-pressure polyethylene (PE) waxes are produced, like high-pressure polyethylene plastic, at high pressure and elevated temperature in the presence of radical formers. As waxes, their molar masses are considerably lower than those of plastics. The molar mass range is adjusted during polymerization by the addition of regulators.

High-pressure polyethylene waxes are partially crystalline and therefore consist mainly of branched molecular chains in which shorter side chains, such as ethyl and butyl, predominate. They generally have low densities (low-density polyethylene waxes, LDPE waxes). Less branched polyethylene waxes with higher crystallinity and density (high-density polyethylene waxes, HDPE waxes) can be produced by increasing the pressure.

High-pressure polyethylene waxes with molar masses between 3000 and 20 000 g/mol [weight-average molar mass $\bar{M}_w$, determined by gel permeation chromatography (GPC)] dominate the market.

The development of high-pressure PE waxes proceeded in parallel to that of LDPE plastics (→ Polyolefins, Chap. 1.1.). In 1939, waxlike ethylene polymers were formed by using a variant of the ICI high-pressure polymerization process [133]. Industrial production of high-pressure ethylene waxes was started in the 1940s [134], [135]. Parallel to this, the technology of thermal depolymerization – the thermal degradation of polyethylene plastics to polyethylene waxes – was developed (see Section 6.1.4).

With the rapid growth of petrochemistry and the expansion of polyethylene plant capacities, the economic importance of PE waxes also increased rapidly. Waxes with very versatile property profiles could be synthesized by varying the density and molar mass of the homopolymers and functionalizing polyethylene by copolymerization with various monomers (e.g., vinyl acetate or acrylic acid) or by melt oxidation.

Because PE waxes are rather inexpensive, have improved applicability and consistency in quality, and are in constant supply they have displaced expensive natural waxes (e.g., carnaúba and montan waxes) in many areas.

The high-pressure process is not applicable for production of polypropylene waxes, because only soft or oily products can be obtained. These waxes are therefore produced by the Ziegler process (see Section 6.1.3) or by depolymerization of high-density polypropylene (see Section 6.1.4).

6.1.1.1. Production

The technology involved in the production of high-pressure PE waxes (see Fig. 10) is analogous to that used for high-pressure polyethylene (→ Polyolefins, Chap. 1.5.1.). The only differences are in the product finishing process. Unlike viscous plastic melts, which must be granulated underwater in granulators, for example, the mobile wax melt can be converted into powder by spraying or into granules by using dicers.

The reaction vessels are stirred autoclaves or tubular reactors (→ Polyolefins, Chap. 1.5.1.1., → Polyolefins, Chap. 1.5.1.2.), the design of which can vary considerably.

Reaction Mechanism. The high-pressure polymerization of ethylene is a highly exothermic radical chain reaction (→ Plastics, General Survey, Chap. 3.3.2.), which is initiated by reaction of ethylene with radicals formed by decomposition of an initiator [136]. Chain propagation occurs through addition of further ethylene units. Chain growth is terminated, for example, by reaction with a regulator molecule such as propene:

$$R-CH_2\dot{C}H_2 + CH_2=CH-CH_3 \rightarrow R-CH{=}CH_2 + CH_3-\dot{C}H-CH_3$$

Polymerization is terminated by recombination or disproportionation of two macroradicals:

$$2\,R-CH_2-\dot{C}H_2 \rightarrow R-CH_2-CH_2-CH_2-CH_2-R$$
$$2\,R-CH_2-\dot{C}H_2 \rightarrow R-CH_2-CH=CH-CH_2-R$$

Long-chain branching occurs by intermolecular chain-transfer reactions:

$$R^1-CH_2\dot{C}H_2 + R^2-CH_2CH_2R^3 \longrightarrow R^1-CH_2CH_3 + R^2-\dot{C}H-CH_2R^3$$
$$\xrightarrow{nCH_2=CH_2} R^2-\underset{\underset{(CH_2CH_2)_{n-1}-CH_2\dot{C}H_2}{|}}{CH}-CH_2R^3$$

A particular feature of polyethylene wax synthesis by the high-pressure process is the increased formation of short ethyl and butyl side chains. These are formed by intramolecular radical transfer:

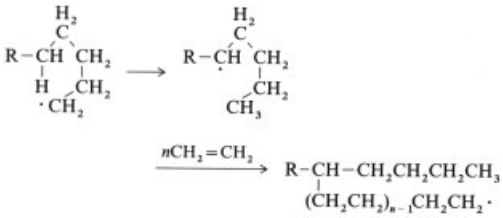

At the high pressures necessary for polymerization, ethylene is in a supercritical state. Polymerization therefore takes place in a one-phase system. After leaving the reactor, the reaction mixture is decompressed in high- or low-pressure separators, and unreacted ethylene evaporates. The wax remains as a melt and ethylene is recycled.

Reaction Conditions. The structure, molar mass, and thus properties of polyethylene wax are determined mainly by reaction pressure, reaction temperature, type and quantity of initiator and molar mass regulator, and reactor type and geometry [137–141].

In the homo- and copolymerization of ethylene the *reaction pressure* is usually 150 – 320 MPa. Somewhat lower pressures are used in autoclaves than in tubular reactors. Other process variants involve much lower pressures of 70 MPa (max.) with isopropanol as the molar mass regulator [139], [140].

Higher pressure favors chain propagation and thus leads to very short residence times in the reactor. It also inhibits chain-transfer reactions so that with increasing pressure the degree of branching decreases and the density, crystallinity, hardness, and *mp* increase. In tubular reactors, rising pressure increases the polymer yield, whereas in autoclaves this effect is smaller.

For homopolymerizaion the *reaction temperature* is 200 – 350 °C and for copolymerization 200 – 300 °C. High reaction temperature favors chain-transfer reactions. For this reason the polyethylene waxes formed have many short-chain and only a few long-chain branches [142] and thus low densities. Higher-density waxes with lower degrees of branching are formed at lower reaction temperature.

Organic peroxides and molecular oxygen (the latter exclusively in tubular reactors) are used as *initiators*. They decompose into radicals un-

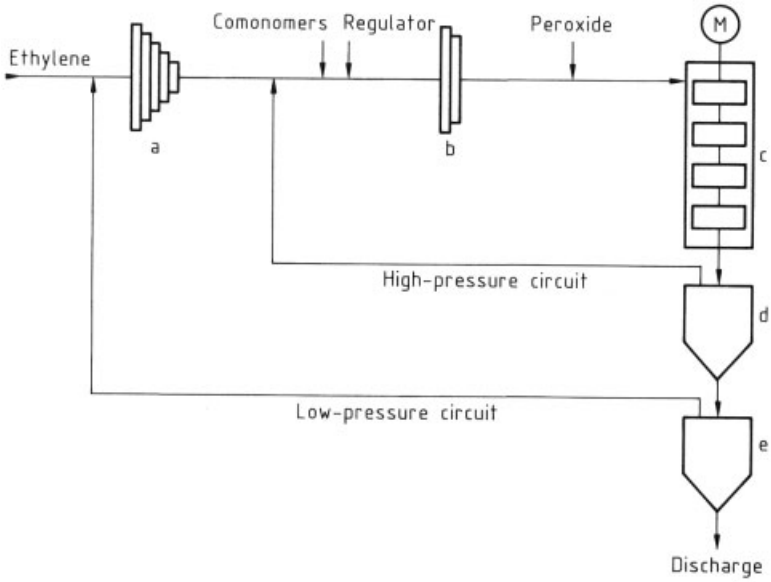

Figure 10. Schematic of production of polyethylene waxes by the high-pressure process in stirred autoclaves
a) Precompressor; b) Postcompressor; c) Autoclave reactor; d) High-pressure separator; e) Low-pressure separator

der polymerization conditions. To ensure that the peroxides are metered reproducibly, they are dissolved in organic solvents.

Hydrogen and almost all organic compounds can act as *molar mass regulators* (chain terminators). Hydrogen, lower alkanes (e.g., propane), lower alkenes (e.g., propene or butene), alkyl aromatics, lower aldehydes (e.g., propionaldehyde), and lower alcohols (e.g., isopropanol) or mixtures of these substances are mainly employed. The concentration and reactivity of these regulators determine the average molar mass, the degree of branching, and thus the density of polyethylene wax. The activity of the regulator increases with increasing temperature and decreases with increasing pressure [136].

Reactor type and geometry also significantly influence polymer structure and properties. Because of the backmixing occurring in *autoclaves*, spherical molecules with many long-chain branches are predominantly formed. In *tubular reactors* with plug flow, long, straight molecules with little long-chain branching predominate [143]. High-density, highly crystalline polyethylene waxes can therefore be produced only in tubular reactors.

Just as the geometries of the autoclave and the tubular reactor differ, so do operating procedures. In the case of autoclaves the initiator solution and the molar mass regulator are charged directly to the preheated pressurized vessel by using high-pressure pumps. In tubular reactors the initiator (e.g., air) and the molar mass regulator are added to the reaction mixture before the compression stage or at the entrance to the reactor. With autoclaves, comonomers can be charged to different parts of the reaction vessel (see Fig. 10).

Product Finishing. Shaping of the molten polyethylene can be performed directly after removal of gaseous products in high- and low-pressure separators and subsequent fine degassing and filtration. Since no metal-containing catalysts are used in high-pressure polymerization of ethylene, no catalyst removal is necessary. Low molar mass fragments and residues of the initiator and molar mass regulator evaporate mainly in the separators. Only an extremely small proportion is incorporated into the polymer.

Granules can be produced by using strip dicers. The wax melt flows in thin strips on a continuous, water-cooled metal band and, after solidification, is chopped with a rotating knife. *Pastilles* can be obtained by metered dripping of

the melt from a distributor head onto this type of cooling band.

A particular characteristic of wax melts is their low melt viscosity. Thus, spraying the melt to form *powders* of spherical particles is also possible. By using special nozzle arrangements (binary nozzles) with nitrogen as the atomizing gas, *micronized waxes* (very fine powders) with particle sizes in the micrometer range can be produced.

Powders and micronized waxes can also be obtained by grinding granules in jet mills. With soft or very viscoelastic waxes, cooling with dry ice or liquid nitrogen is necessary. Coarse particles must be removed by subsequent classification.

Like all organic dusts (e.g., coal or flour), wax powders, and particularly micronized waxes, are highly susceptible to dust explosions. Appropriate safety precautions must be taken during processing, in particular grounding all installations to avoid spark formation through electrostatic charge buildup.

6.1.1.2. Properties

The transition from polyethylene plastic to polyethylene wax is flexible. If the molar mass is lowered, the thermoplastic materials is gradually changed to a wax. The DGF definition of wax (see Section 1.2) gives an approximate boundary: For waxes, an upper limit to the melt viscosity of ca. 20 000 mm^2/s at 120 °C is defined, which corresponds to an average molar mass (weight-average molar mass $\bar{M}_w$) of ca. 37 000 g/mol.

The properties of polyethylene waxes are determined strongly by $\bar{M}_w$ as a measure of the average chain length and by the degree of branching as a measure of the shape of the molecules. Melt viscosity increases with increasing molar mass. Crystallinity, hardness, *mp*, and solidification point increase as the degree of branching decreases. These data are important for application-oriented properties.

The molar mass can be determined by GPC. This method gives the weight-average ($\bar{M}_w$) and the number-average ($\bar{M}_n$) molar mass. The ratio $\bar{M}_w/\bar{M}_n$ is a measure of molar mass distribution and is known as the polydispersity index (→ Polymerization Processes, Chap. 2.1.1.). For high-pressure polyethylene waxes currently produced, this value lies between 2.0 and 2.5. The waxes have a much narrower molar mass distribution than high-pressure polyethylene, for which the values are 3 – 8 (tubular reactor) and 12 – 16 (autoclave) [136].

Table 21 gives a series of structure and property data for three high-pressure polyethylene waxes with increasing density from two different producers (some of this represents unpublished data supplied by the authors).

The following conclusions can be drawn from these data:

1) The degree of branching is generally low with ca. ten branches (max.) per molecule. The side chains are mostly ethyl and butyl groups with a very small proportion of long-chain branches [144–146].
2) The density, crystallinity, hardness, solidification point, and drop point increase as degree of branching decreases [147]. All density-dependent properties vary correspondingly.
3) High-pressure PE waxes contain a small proportion of double bonds, whose distribution in the macromolecule depends on the production process, as the large differences between the levels of central double bonds show. The formation of double bonds is attributed to depolymerization of chain radicals at high reaction temperature [136], [148], e.g.:

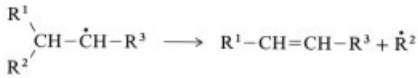

High-density polyethylene waxes are colorless, white to transparent and form clear melts. Like other waxes, they dissolve in nonpolar solvents (e.g., aliphatic, aromatic, and chlorinated hydrocarbons) on heating and generally crystallize as very fine particles on cooling. Depending on the type and concentration of the wax, they then form mobile dispersions or paste-like gels, which frequently exhibit thixotropic properties.

6.1.2. Copolymeric Polyethylene Waxes by High-Pressure Polymerization

(→ Polyolefins, Chap. 1.5.1.3.)

In the high-pressure process, many other monomers can copolymerize with ethylene giv-

Table 21. Structure and properties of high-pressure polyethylene waxes A – C and D – F from two different producers

Characteristic	A	B	C	D	E	F
Density (23 °C), g/cm^3 (DGF-M-III 2a)	0.92	0.92	0.945	0.92	0.92	0.94
CH_3 per 1000 C atoms	26	22	17	32	23	20
Branches per molecule	3.4	1.8	1.0	3.7	1.9	1.8
Crystallinity (from IR data), %	54	59	67	52	62	64
Central double bonds per 1000 C atoms	0.1	0.4	0.4	0	0	0
Terminal double bonds per 1000 C atoms	0.1	0.1	0	1.3	1.2	1.3
Side-chain double bonds per 1000 C atoms	0.6	0.4	0.2	0.9	0.5	0.4
Ball indentation hardness (23 °C), bar (DGF-M-III 9a)	200	370	550	314	413	482
Solidification point, °C (DGF-M-III 4a)	96	99	106	93	101	104
Drop point, °C (DGF-M-III 3)	106	112	116	106	111	113
Melt viscosity (120 °C), mm^2/s (DGF-M-III 8)	1310	1160	1200	780	925	890
Weight-average molar mass $\bar{M}_w$, g/mol	6140	5930	6000	5200	5500	5500
Number average molar mass $\bar{M}_n$, g/mol	2920	2390	2440	2490	2380	2660
$\bar{M}_w/\bar{M}_n$ (polydispersity index)	2.1	2.5	2.5	2.1	2.3	2.1

ing rise to considerable changes in product properties [139], [141]. For copolymerization, autoclaves with their intensive mixing, constant temperature, and consequently stable reaction process are particularly suitable. In industry, vinyl acetate and acrylic acid are used mainly as comonomers, giving waxes with higher polarity and lower crystallinity. Here also, the degree of branching decreases with increasing pressure and decreasing temperature. Density, *mp*, and hardness increase accordingly.

Table 22 gives characteristic data for two important copolymeric high-pressure polyethylene waxes with vinyl acetate and acrylic acid as comonomers.

Table 22. Physical and chemical data for copolymeric high-pressure polyethylene waxes

Characteristic	Ethylene – vinyl acetate copolymer wax	Ethylene – acrylic acid copolymer wax
Weight-average molar mass $\bar{M}_w$, g/mol	ca. 6800	ca. 6100
Number-average molar mass $\bar{M}_n$, g/mol	3000	3000
Melt viscosity (120 °C), mm^2/s (DGF-M-III 8)	2000	1300
Solidification point, °C (DGF-M-III 4a)	85	93
Drop point, °C (DGF-M-III 3)	96	102
Ball indentation hardness (120 °C), bar (DGF-M-III 9a)	120	410
Acid number, mg KOH/g		45
Vinyl acetate content, %	10	

Ethylene – vinyl acetate copolymer is a relatively polar but nevertheless hydrophobic wax that can be dispersed in organic solvents particularly well. It is used as an additive in metallic automotive paints, as a dispersing agent in pigment concentrates, and as a component of hot melts.

Ethylene – acrylic acid copolymer is suitable for aqueous systems because it can be emulsified readily and is chemically and thermally intensive. Its emulsions are used mainly in modern floor polishes and mold-release agents.

6.1.3. Polyolefin Waxes by Ziegler – Natta Polymerization

Like high molar mass polyethylene plastics, PE waxes can be produced by the Ziegler low-pressure process using organometallic catalysts (→ Polyolefins, Chap. 1.5.2.), in addition to the radical process at high pressure and temperature. Like high-density polyethylene, Ziegler waxes have a mainly linear molecular structure. They can contain short side chains but not long ones like high-pressure polyethylene waxes.

For direct synthesis of polymer waxes from propylene or higher α-olefins, only the Ziegler – Natta process is suitable (→ Polyolefins, Chap. 2.5.).

Ziegler PE waxes currently on the market have a molar mass between 800 and 8000 g/mol ($\bar{M}_n$). They thus bridge the gap between Fischer – Tropsch and paraffin waxes and HDPE thermoplastics. Besides controlling chain length, the polymerization process allows adjustment of the degree of crystallinity so that both hard – brittle, high-melting products and

softer, flexible products with low melting points can be obtained.

The soluble, soft wax fractions inevitably produced in the suspension process for HDPE are more similar to paraffin waxes than to PE waxes because of their low molar masses. For normal hard wax applications, they can be used only after processing by distillation or extraction [149].

6.1.3.1. Production

For the production of Ziegler – Natta polyolefin waxes, Ziegler catalysts that give polymers with low average chain lengths and narrow chain-length distributions, and are sufficiently active under the special conditions of wax synthesis, are particularly suitable. The low degree of polymerization typical of waxes is achieved by carrying out polymerization in the presence of hydrogen as a molar mass regulator at comparatively high temperature (usually between 100 and 200 °C, sometimes even higher) [150–154]. Aliphatic hydrocarbons are used as reaction medium. Because of the high reaction temperature the waxes are formed in solution (solution polymerization). If the product viscosity is sufficiently low the molten polymer formed can itself function as the solvent (bulk polymerization) [155], [156]. Unlike modern HDPE processes in which the catalyst is not removed, in the case of wax synthesis the catalyst must usually be decomposed and filtered off [157], [158]. This applies at least to the heterogeneous titanium catalysts still used today (see below) and is necessary because waxes have higher purity requirements than PE plastics as a consequence of the applications for which they are used. After the solvent has been distilled off, the wax melt is shaped (e.g., by spraying or pastille formation).

Catalyst Systems. Classical Ziegler catalysts, consisting of titanium tetrachloride and alkylaluminum compounds, can be used for wax synthesis, but their activity is low [155], [156], [159–161]. The degree of branching and thus also the density, drop point, and hardness can be adjusted by copolymerization, by the special type of catalyst preparation [160], and by varying the polymerization temperature [155], [161].

The state of the art for production of *PE homo- and copolymer waxes* involves, as in plastics production, the use of supported catalysts, which contain titanium atoms as the active species and magnesium compounds as the carrier material (→ Polyolefins, Chap. 1.3.2.2.). Catalysts derived from titanium tetrachloride and magnesium chloride, oxide, hydroxide [151–154], or alkoxide [150], for example, are suitable.

Catalysts based on Ti – Mg compounds can also be used in the production of *polypropylene waxes.* Through an appropriate choice of catalyst, the degree of crystallinity (isotacticity) (→ Polyolefins, Chap. 2.2.) can be varied within wide limits and thus adapted to a particular application. Flexible products with average crystallinity can be obtained [162]. Highly crystalline polypropylene waxes can now be produced economically by using stereoregulating silanes as additional catalyst components [163], [164]. Formerly, these waxes could be produced only by thermal degradation of high molar mass polypropylene (Section 6.1.4).

Indications are that the homogeneous metallocene – alumoxane catalyst systems [165], [166], which became known at the end of the 1970s and nowadays are used for plastics synthesis in industry, will also become important in wax synthesis. Since the 1980s, many patent applications have appeared in this field, relating to both PE [167–172] and PP [173–175] waxes. Metallocene catalysts have a number of advantages compared with titanium catalysts. They are more active and can thus be employed in lower quantities, so catalyst removal is not necessary. Polymerization can occur at lower temperature, and the process is therefore more economical. Above all, the specific polymerization properties and great structural variability of metallocene systems are likely to give polyolefin waxes with optimized or completely new property profiles.

6.1.3.2. Properties

The most important molecular structure parameters, which determine the macroscopic property profile of PE waxes, are the average molar mass (average chain length) and the degree of branching. Melt viscosity rises exponentially with in-

creasing molar mass (Fig. 11). Unbranched, *homopolymeric PE waxes* have a high degree of crystallinity as a result of their undistorted linear molecular structure. They are brittle and hard (needle penetration numbers <1) and have high densities and melting points (up to ca. 127 °C). These products can be obtained only by Ziegler polymerization. Radical polymerization always gives branched waxes.

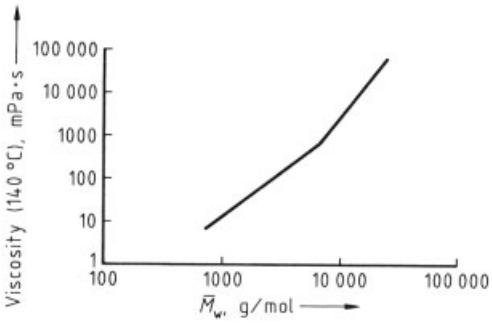

Figure 11. Relationship between average molar mass $\bar{M}_w$ and melt viscosity for PE waxes

By using *1-alkenes* (generally propene) *as copolymers*, defects can be introduced into the polymer chain in the form of short-chain branches, which lead to a decrease in crystallinity. The density, hardness, drop point, and heat of fusion are also decreased, while flexibility and plasticity increase.

In Table 23, some important physical properties of typical commercially available Ziegler polyolefin waxes are compared.

Ziegler waxes are colorless and odorless and form clear, transparent melts with high thermostability. Like most other waxes, they are soluble on warming in nonpolar aliphatic and aromatic solvents without leaving a residue, and on cooling often crystallize out as pastes. Commercial products have a melt viscosity range from ca. 10 to 30 000 mPa · s (at 140 °C). Their densities are between 0.92 and 0.97 g/cm^3.

6.1.4. Degradation Polyolefin Waxes

By heating in the absence of air, the molecular chain of high molar mass polyolefins can be cleaved into smaller molecules with wax-like character. The starting materials are usually low- and high-density polyethylene, isotactic polypropylene, and polybutene. The production capacities of degradation polyolefin waxes have diminished since ca. 1980 in favor of the built products (see Sections 6.1.1, 6.1.2, 6.1.3). While the proportion of degradation PE waxes has diminished, that of PP waxes has increased. Despite the relatively unfavorable energy balance of the degradation process, it is still used because of its technological simplicity and, for PE waxes, occupies third place with regard to the production quantity after the high-pressure and low-pressure PE waxes. The degradation process was developed approximately in parallel to high-pressure polymerization [176–178].

6.1.4.1. Production

Polyethylene is a thermally stable polymeric hydrocarbon. In the absence of oxyen it can withstand temperatures up to ca. 290 °C [179]. At higher temperature, thermal degradation begins with a decrease in the molar mass; from 400 °C, degradation occurs with high rate and yield [179]. The cleavage products consist almost exclusively of a wax fraction with a molar mass distribution that is barely wider than that of high-pressure PE waxes. Low-boiling and gaseous crack products are formed only in small quantities. Individual process steps are shown in Figure 12.

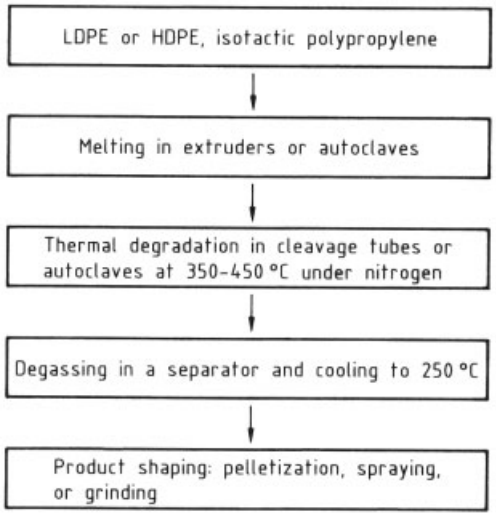

Figure 12. Process schematic for the production of degradation polyolefin waxes

Table 23. Structure and properties of typical commercial Ziegler polyolefin waxes

	A Ethylene homopolymer wax	B Ethylene copolymer wax	C Ethylene homopolymer wax	D Ethylene homopolymer wax	E Propylene homopolymer wax
Number-average molar mass $\bar{M}_n$, g/mol (GPC)	1600	2200	850	5000	3000
Weight-average molar mass $\bar{M}_w$, g/mol (GPC)	4800	6300	920	17 000	18 000
Molar mass distribution ($\bar{M}_w/\bar{M}_n$)	3.0	2.9	1.1	3.4	6.0
Density (23 °C), g/cm^3 (DGF-M-III 2a)	0.97	0.93	0.96	0.97	0.88
Heat of fusion, J/g (DSC)	240	120	230	200	60
Viscosity (140 °C), mPa · s	400	600	10	25 000	1700 *
Drop point, °C (DGF-M-III 3)	125	118	107	135 **	158 **
Needle penetration, 0.1 mm (DGF-M-III 9b)	<1	3	2	<1	<1

* Temperature of measurement 170 °C.
** Softening point ring/ball (DGF-M-III 13).

The starting materials for thermal degradation to polyolefin waxes are LDPE and HDPE [180] and HD polypropylene [181–184]. Degradation of copolymeric polyolefins [185] and polybutenes [186] has also been described but has not achieved great economic importance.

In the industrial process [180], polyolefin pellets are melted under nitrogen in a single-screw extruder and extruded into a steel tube heated to ca. 400 °C. The residence time is 15 – 30 min. The product is then cooled by evaporative cooling in ca. 5 min to <250 °C. It is freed from low molar mass components by degassing in a separator. The product can then be finished as described for high-pressure PE wax (Section 6.1.4.1).

The degree of branching of degradation waxes formed, and thus all structure-dependent properties such as crystallinity, density, hardness, and *mp*, are controlled by the choice of starting material (LDPE or HDPE, stabilizer free). The average molar mass depends on residence time in the hot-tube reactor and the average molar mass of the starting material. The degradation temperature affects mainly the reaction time and thus the throughput rate.

Batchwise, thermal degradation in autoclaves has also been described [182], [185]. It is performed in certain cases in the presence of hydrogen and hydrogenation catalysts to saturate double bonds formed on chain cleavage [176].

Like other polyethylene waxes, degradation waxes can be modified by air oxidation and thus rendered emulsifiable (see Section 6.1.5). Preliminary hydrogenation of double bonds can be necessary to avoid cross-linking reactions during oxidation.

Because of their double bonds, degradation waxes are particularly suited to grafting with unsaturated carboxylic acids (e.g., with maleic anhydride [187], [188]).

Little information is available in the literature on the mechanism of thermal degradation of polyolefins to form waxes [179]. Presumably, radicals form from the few oxygen-containing groups in the polyolefin and initiate a radical chain reaction, in which the chain is cleaved through radical transfer and a double bond is formed:

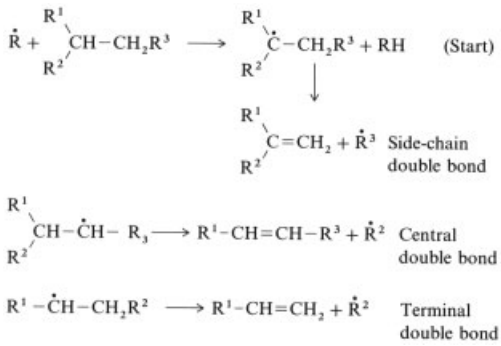

Since tertiary hydrogen atoms are particularly susceptible to radical transfer, the degradation reaction begins at the carbon atoms of the branches, to form fragments containing side-

chain double bonds. Only later are linear chains cleaved, to form central and terminal double bonds.

6.1.4.2. Properties

In Table 24, properties (in some cases unpublished results supplied by the authors) of three degradation and three high-pressure polyethylene waxes, with about the same average molar mass to decreasing branching and increasing density, are compared. The two types of wax can be seen not to differ significantly. For both, the same relationships are valid between molar mass and melt viscosity and between degree of branching and density, crystallinity, hardness, and solidification and drop points. The width of the average molar mass distribution (polydispersity) is also comparable. A significant difference is that degradation polyethylene waxes contain considerably more double bonds than comparable polyethylene waxes formed by high-pressure polymerization. Determination of the type and level of double bonds from the IR spectrum can therefore be used for analytical identification of degradation polyethylene waxes. The technical properties are affected only slightly by double bonds.

Compared with degradation polyethylene waxes, *degradation polypropylene waxes* usually have extremely high melting points and hardness because they are produced from highly crystalline, isotactic high-density polypropylene. Some characteristics of a relatively low molar mass degradation polypropylene wax are listed below:

mp (DSC)	130 – 160 °C
Ball indentation hardness (23 °C) (DGF-M-III 9a)	500 – 1900 bar
Melt viscosity (200 °C) (DGF-M-III 8)	800 – 1900 m^2/s
Weight-average molar mass (GPC, standard: PP)	17 000 – 37 000 g/mol

Degradation polyolefin waxes are colorless, white to transparent, and form clear melts. On warming, they dissolve in nonpolar solvents such as aliphatic, aromatic, and chlorinated hydrocarbons, and on cooling, they crystallize as dispersions or gels (pastes).

6.1.5. Polar Polyolefin Waxes

Polar polyolefin waxes generally contain carboxyl and ester groups, which give them special technical properties, in particular emulsifiability in aqueous media. They can be produced by a variety of methods:

1) Oxidation of nonpolar PE waxes (melt oxidation) or PE plastics (oxidative degradation) with air
2) Radical high-pressure polymerization of ethylene with oxygen-containing comonomers (acrylic acid, acrylates, vinyl acetate, etc.) (see Section 6.1.1.1)
3) Radical grafting of polar unsaturated monomers onto nonpolar PE and PP waxes

To a great extent, PE waxes are converted into wax oxidates by treatment with atmospheric oxygen in the melt at ca. 150 – 160 °C [189–191]. Oxygen functions (carboxyl, ester, carbonyl, hydroxyl groups, etc.) are introduced into the wax molecule by complex mechanisms involving peroxy intermediates and chain-shortening reactions [192]. The usual commercial products with acid numbers between 25 and 30 mg of KOH per gram contain ca. 3 – 4 wt % oxygen, mainly in the form of carboxyl and ester groups.

Oxidation of the molten wax is limited to waxes whose molar mass is not too high. In the case of longer-chain starting materials, intensive dispersion of air in the melt, which is necessary for the reaction to run smoothly, is inhibited by the melt's high viscosity. In this case, waxes can be oxidized in *inert dispersion agents*. Since a (statistical) chain cleavage is always associated with oxidation (i.e., a lowering of molar mass), PE plastics can also be converted into a polar wax by this route (*oxidative degradation*). For example, molten polyethylene in an aqueous dispersion can be converted into emulsifiable waxes by passing in air at 130 – 160 °C with intensive stirring. These waxes are harder and have higher drop points than the usual melt oxidates with the same acid number because of their relatively high molar masses [193]. The reaction can be accelerated by addition of peroxides. High-density polyethylene gives hard brittle products. Tough, flexible waxes, which can be used as antislip components of floor polishes, can be obtained from LDPE containing vinyl acetate [194].

Table 24. Characteristics of three degradation and three high-pressure polyethylene waxes

Characteristic	Degradation PE wax			High-pressure PE wax		
	G	H	I	A	B	C
Density (23 °C), g/cm^3 (DGF-M-III 2a)	0.92	0.93	0.95	0.92	0.93	0.945
CH_3 per 1000 C atoms	32	16	8	26	22	17
Branches per molecule	3.5	0.5	0	3.4	1.8	1.0
Crystallinity (from IR data), %	43	63	74	54	59	67
Central double bonds per 1000 C atoms	1.2	1.5	2.0	0.1	0.4	0.4
Terminal double bonds per 1000 C atoms	3.3	2.6	2.6	0.1	0.1	0
Side-chain double bonds per 1000 C atoms	1.7	1.1	0.5	0.6	0.4	0.2
Ball indentation hardness (23 °C), bar (DGF-M-III 9a)	450	500	600	200	370	550
Solidification point, °C (DGF-M-III 4a)	96	100	110	96	99	106
Drop point, °C (DGF-M-III 3)	105	112	121	106	112	116
Melt viscosity (120 °C), mm^2/s (DGF-M-III 8)	930	1230	1300	1310	1160	1200
Weight-average molar mass $\bar{M}_w$, g/mol	5400	6040	6100	6140	5930	6000
Number average molar mass $\bar{M}_n$, g/mol	2400	2160	2440	2920	2390	2440
$\bar{M}_w/\bar{M}_n$ (polydispersity index)	2.3	2.8	2.5	2.1	2.5	2.5

Table 25. Properties of typical commercial oxidized polyolefin waxes

Characteristic	A	B	C	D	E
Raw material	high-pressure PE wax	Ziegler PE wax	HDPE	HDPE	ethylene – vinylacetate copolymer
Acid number, mg KOH/g (DGF-M-IV 2)	22	18	16	60	21
Saponification number, mg KOH/g (DGF-M-IV 2)	50	38		103	80
Viscosity (120 °C), mPa · s	400	200	8500 (150 °C)	300	3000
Drop point, °C (DGF-M-III 3)	100	116	140	111	100
Needle penetration, 0.1 mm (DGF-M-III 9b)	5	5	<0.5	1	2
Density, g/cm^3 (DGF-M-III 2a)	0.96	0.98	0.98	1.02	0.96

Oxidative degradation of polyethylene can also be carried out as a *gas – solid reaction*. In this case, no dispersing agent is required. Air, oxygen-enriched air, or pure oxygen in the presence of radical initiators is reacted with the substrate in powder form a few degrees below the softening point [195–198]. Because a low temperature (<130 °C) must be maintained, the oxidation period is comparatively long. For example, it requires at least 20 h to achieve acid numbers of 25 – 30 mg of KOH per gram. Use of air – ozone mixtures as the oxidizing agent has also been described [199], [200]. Another process variant involves oxidation of polyethylene powder with air in *aqueous suspension* under pressure [201]. The properties of some representative commercial wax and plastic oxidates are given in Table 25.

Oxidized PE waxes can be modified further by esterification [202], amidation [203–205], saponification, and other derivatization reactions [206] and thus adapted to special technical requirements.

Polyolefin waxes can also be converted into emulsifiable products by *grafting unsaturated polar compounds* (maleic anhydride, etc.) [161], [207–210]. For *polypropylene wax* [211–213], this is the only method of introducing oxygen functions that is used in practice, since air oxidation would lead to low molar mass, colored, soft degradation products.

6.2. Uses

6.2.1. Nonpolar Polyolefin Waxes

Polyolefin waxes are used in widely differing applications, mostly as auxiliaries in products and production processes. The areas of use of high-pressure, Ziegler, and degradation waxes overlap extensively. More details can be found

in leaflets and brochures of various production companies. Important areas of use include the following:

1) *In Additive and Pigment Master batches.* PE and PP waxes are carrier materials and binders in concentrates used as plastics additive and for coloration of plastics.
2) *As Additives to Improve Rub Resistance and as Slip Agents* [214]. PE waxes are incorporated into printing inks after micronizing by milling or spraying, or in the form of dispersions.
3) *In Paints and Coatings*, PE waxes are used for matting effects and to increase resistance to scratching [215], [216].
4) *In Plastics Processing*, PE waxes are used as lubricants and release agents in molding.
5) *In Polishes*. PE waxes are components of heat-resistant pastes for floor, shoe, and car polishes.
6) *In Hotmelt Coatings.* PE and PP waxes are heat-sealable components of coatings and adhesives for paper and metal sheeting because they improve insulation, gloss properties) and heat resistance.
7) *For Corrosion Protection*. PE waxes are used as hydrophobing components for temporary corrosion protection of motor vehicles, machines, and instruments.
8) *Photocopying.* PE and highly crystalline PP waxes are used in toner preparations for photocopying machines.
9) *In the Rubber Industry*, PE waxes are used as release agents in processing.

Also, PE and PP waxes are used as components of insulating compounds, as carriers and binders for mechanically and chemically reacting self-duplicating paper and wax crayons, to coat granular fertilizer, and to increase the heat resistance of candles.

6.2.2. Polar Polyolefin Waxes

Polar polyolefin waxes are used mainly in the form of aqueous emulsions (dispersions). Depending on the nature of the emulsifier system used, nonionic, nonionic – ionic, anionic, and cationic formulations can be produced. Emulsions are produced by mixing molten waxes and emulsifiers with hot water. On rapid cooling with intensive stirring, wax droplets solidify and form stable, finely divided dispersions. High-melting waxes must be emulsified under pressure. After the emulsion is applied, water evaporates and leaves a homogeneous wax film, often with a high gloss, on the treated surface. This effect is used in floor polishes where polar waxes are applied in combination with acrylate- and styrene-based polymer dispersions. Waxes obtained by oxidative degradation of plastics are primarily used in this area because they form harder-wearing films due to their greater hardness than lower molar mass melt oxidates. Polyethylene wax emulsions are also used as leather and shoe polishes and in many other areas of industry. In textiles, they are used for wash and wear finishes and to improve sewing properties. They are used as hydrophobing agents in paint dispersions, as release agents in the building industry, and in surface finishing of paper products. Citrus fruit is provided with a glossy surface and protected from drying out through the application of wax films [217]. In the plastics processing industry, polar PE waxes are used as lubricants and release agents.

6.3. Economic Importance

As indispensable additives with a secure raw material supply and broad application potential, polyolefin waxes play an important role, mainly in the industrial, but also in the consumer market sectors.

The quantity of polyolefin waxes annually consumed worldwide is ca. 200 000 t (1995), including oxidized and grafted products. High-pressure waxes (PE and polar copolymers) account for ca. 70 %. The remainder consists essentially of Ziegler products (PE, PP). Degradation waxes (mainly PP) are less important in terms of quantity.

Ca. 12 000 t of polar waxes are produced worldwide by oxidative degradation of PE plastics.

The main consumers of nonpolar polyolefin waxes are the producers of pigment concentrates and of plastic and rubber articles, as well as the paint and printing ink industry. Polar waxes are used mainly in the form of emulsions in the textile, polish, and paper industries.

Producers and Trade Names.

High-Pressure Waxes. Allied-Signal, United States (A-C Polyethylene). Eastman Kodak, United States (Epolene Wax). BASF, Germany (Polyethylenwachs BASF, Luwax). Leuna, Germany (LE-Wachs).

Ziegler Waxes. Hoechst, Germany (Hoechst-Wachs). Mitsui Petrochemical Ind., Japan (Hiwax). Hüls, Germany (Vestowax). Petrolite, United States (Polywax).

Degradation Waxes. Sanyo Chemical Industries, Japan (Viscol, Youmex). Mitsui Petrochemical Ind., Japan (NP-Types). Lion Chemical, Korea (L-C Wax). Eastman Kodak, United States (Epolene). Ceralit, Brazil (Cerit Poly).

7. Toxicology

On the basis of experience accumulated over thousands of years, waxes are generally considered harmless, nontoxic, environmentally friendly products. The results of scientific toxicological and ecotoxicological tests on natural waxes, montan waxes, polyethylene waxes, and other wax products generally support this.

Natural Waxes. From the group of animal and vegetable waxes, *carnaúba wax* was subjected to toxicological tests. In a two-generation feeding experiment, 1 % wax (max.) based on the feed, was administered first to pregnant rats and then to their first-generation offspring over a period of 90 d. The results gave no indication of dose-dependent effects in the rats treated compared to the untreated control group [218].

Montan Waxes and Their Derivatives. The toxicological properties of acid, ester, and partially saponified ester waxes (partially neutralized with calcium hydroxide) derived from montan wax are determined mainly by the indigestibility of the products, as in the case of natural waxes. Besides determining acute oral toxicity, extensive feeding experiments on rats and dogs over periods of 90 d, 140 d, and 2 a have been carried out with montan wax derivatives [219]. The LD_{50} values are >15 000 and 20 000 mg per kilogram of body weight, respectively. According to the usual classification, these products are considered practically nontoxic and harmless [220].

Long-term feeding experiments led to no macroscopically or microscopically recognizable toxicological damage to the animals, even at very high dosages of 5 % in the feed (Table 26).

Dermatological examination of montan wax derivatives in the percutaneous test and the mucous membrane test on rabbits (Hoechst waxes S, E, and OP) did not lead to any irritation of the skin or mucous membranes.

In the Ames mutagenicity test with and without external metabolic activation, a montanic acid – ethylene glycol ester wax (Hoechst wax E) did not show any mutagenicity (tested dosages 4 – 5000 µg per plate) [221].

The *ecotoxicological behavior* of montan waxes is attributed to their water insolubility. Mixtures with water are attacked very slowly by microorganisms and cause no perceptible biological oxygen demand (BOD) in water bodies and sewage treatment plants. The waxes are separated from wastewater as ballast materials together with sludge. In the determination of biodegradability according to OECD test method 302 B (static test/Zahn – Wellens test), the finely dispersed montan wax derivatives in wastewater show good elimination of chemical oxygen demand (→ Ecology and Exotoxicology). They are separated with sludge and are nontoxic to fish and bacteria.

In the case of wax emulsions, when the usual test methods for determination of COD and BOD are used, a very slow oxygen consumption can be detected, which is higher than the oxygen demand of the emulsifiers [219]. In wastewater, these emulsions break up as a result of great dilution, altered pH, and the presence of calcium and magnesium ions. The findings described for pure wax – water mixtures also apply to the agglomerates formed.

Polyolefin Waxes. Polyethylene waxes, oxidized polyethylene waxes, and polypropylene waxes are similar to those based on montan wax and natural waxes with regard to toxicology. The LD_{50} values for rats are >5000 and 15 000 mg per kilogram of body weight, respectively. In a long-term toxicity test (a 9-month feeding experiment on rats), polyethylene waxes caused no toxic damage to the animals even at the highest dosage of 8.5 % in the feed. In a 90-d feeding experiment on rats, oxidized polyethylene wax up to acid numbers 70 in dosages of 0.2, 1, 1.6, and

Table 26. Acute oral, subchronic, and chronic toxicity of montan wax derivatives

Wax	Chemical composition	Acute oral toxicity, LD_{50}, mg/kg	Subchronic toxicity, 90-d no-effect level ∗, mg/kg	Subchronic toxicity, 140-d no-effect level ∗, mg/kg	Chronic toxicity, 2-a no-effect level ∗, mg/kg
Acid wax (Hoechst wax S)	montanic acid $C_{16}-C_{36}$	>15.000 (rat)			
Ester wax (Hoechst wax E)	esters of montanic acid with ethanediol	>20 000 (mouse)		50 000 (dog)	50 000 (rat)
Ester wax (Hoechst wax KPS)	esters of montanic acid with ethanediol and 1,3-butanediol	>15 000 (rat)	50 000 (dog)		50 000 (rat)
Ester wax (Hostalub WE 4)	esters of montanic acid with glycerol		50 000 (rat)		
Partially saponified ester wax (Hoechst wax OP)	esters of montanic acid with 1,3-butanediol and calcium salts of montanic acid	>20 000 (mouse)		50 000 (dog)	50 000 (rat)

∗ Proportion by weight in the feed.

5 % in the feed produced no adverse effects in the animals. Histopathological tests showed no morphologically recognizable damage or changes caused by the waxes. No accumulation was observed in tissues, liver, or lymph glands. Fat and iron metabolism and spermatogenesis and oogenesis were unaffected (Table 27) [219]. Practically no negative effects were found in dermatological tests on rabbits. Polyethylene waxes and oxidized polyethylene wax show good COD elimination under sewage treatment plant conditions and no toxicity to fish or bacteria.

Paraffins and Microwaxes. Petroleum waxes have been used for decades as food additives, as wax coatings for cheese, in chewing gum, for upgrading food packaging, and in cosmetics and pharmaceutical products, without damaging human health. The toxicological properties of paraffins and microwaxes extracted from crude petroleum depend essentially on their composition and purity, particularly their residual aromatics content.

In a study of the toxicity of petroleum waxes, rats were fed with several paraffin products for 2 years. At a dosage of 10 % wax in the feed, no toxicological or carcinogenic effects were observed. Even on repeated application of petroleum waxes to the skin of mice and rabbits, no carcinogenic effects were detected [222].

Toxicological tests have been carried out with paraffins and microwaxes are mainly 90-d feeding experiments on rats. In these, varying toxicological effects were observed. For low molar mass paraffins, accumulation in the tissues and increased organ weights of the liver, kidney, spleen and lymph glands were observed. The most important histopathological findings were granulomata in the liver and histiocytosis in the lymph glands. These effects were not found with high molar mass waxes.

No toxic effects were found at a dose of 0.6 mg per kilogram of body weight and day for low molar mass paraffins (C-number maximum 25) and 1.8 g per kilogram of body weight per day for high molar mass paraffin (microwaxes, C-number maximum 42).

The latest results are published in BIBRA reports [223], [224], CONCAWE reports [225], and reports of the Toxicology Forum [226], [227].

Amide Waxes. The most well known amide wax, bis(stearoyl)- bis(palmitoyl)ethylenediamide, has an LD_{50} of >20 000 mg/kg (mouse). In determination of the chronic toxicity in a 2-a feeding experiment on rats, a non-effect level of 5 % in the feed was established [218].

Approval for Food Contact. On the basis of their low toxicity, a range of natural waxes, montan wax derivatives, amide waxes, polyethylene waxes, paraffins, and microcrystalline waxes have been approved for food contact by the health authorities of many countries [e.g., the German BgVV (formerly BGA), the U.S. FDA, and in Australia and Japan]. The products are also contained in the draft version of the EC additives list (Synoptic Document no. 7 of May, 15, 1994). Some of the waxes are approved in

Table 27. Acute oral, subchronic, and chronic toxicity of polyolefin waxes

Wax	Acute oral toxicity, LD_{50}, mg/kg	Subchronic toxicity, 90-d (rat) no-effect level ∗, mg/kg	Subchronic toxicity, 9-month (rat) no-effect level ∗, mg/kg
Polyethylene wax (Hoechst wax PE 520)	>15 000		85 000
Polyethylene wax oxidates			
(Hoechst wax PED 121)		16 000	
(Hoechst wax PED 522/PED 136)	>15 000	50 000	
Polypropylene wax (Hoechst wax PP 230)	>5000		

∗ Proportion by weight in the feed.

the European Directives 95/2/EC and 95/3/EC for food contact.

Registration According to EINECS/TSCA/AICS/DSL. Waxes are registered under the CAS numbers given in Table 28 in the European Inventory of Existing Chemical Substances (EINECS), the TSCA inventory in the United States, the Australian AICS inventory, and the Canadian DSL inventory.

Table 28. Registration according to EINECS, TSCA, AICS, DSL

Type of wax	CAS registry no.
Natural waxes	
Beeswax	[*8012-89-3*]
Carnaúba wax	[*8015-86-9*]
Candelilla wax	[*8006-44-8*]
Waxes based on montan wax	
Montan wax fatty acids	[*68476-03-9*]
Ethylene esters of montan wax fatty acids	[*73138-45-1*]
1-Methyl-1,3-propanediyl esters of montan wax fatty acids	[*73138-44-0*]
Calcium salts of montan wax fatty acids	[*68308-22-5*]
Montan wax glycerides	[*68476-38-0*]
Montan wax (crude montan wax)	[*8002-53-7*]
Polyolefin waxes	
Polyethylene wax	[*9002-88-4*] ∗
Ethylene	[*74-85-1*]
Polypropylene wax	[*9003-07-0*] ∗
Propylene	[*115-07-1*]
Polyethylene wax oxidates	[*68441-17-8*] ∗
Petroleum waxes	
Paraffin and hydrocarbon waxes	[*8002-74-2*]
Microcrystalline waxes	[*63231-60-7*]
Amide wax	
N,N′-Ethylenedistearamide	[*110-30-5*]

∗ In EINECS, polymers are registered only under the CAS registry no. of the monomer.

Waxes are not in the *MAK value list* for 1995.

Labeling According to GefStoffV, ChemG, and EC Guidelines. If they are pure and are not in the form of preparations with components for which labeling is compulsory (e.g., emulsifiers), waxes do not have to be labeled according to GefStoffV (October 26, 1993), the ChemG, Technische Regeln für Gefahrstoffe (technical regulations governing hazardous substances, TRGS) 200, and the EC Guideline 67/548/EC and adaptations and amendments.

Status According to OSHA. Paraffins are regulated by OSHA. Labeling is prescribed on packaging of paraffins (right-to-know label), in which warning of paraffin vapors is given. Paraffin vapors are classified as eye and respiratory irritants; they have the following exposure limits:

OSHA-PEL	2 mg/m^3
ACGIH-TLV	2 mg/m^3

Other waxes are not regulated by OSHA.

8. References

1. Chr. Columbus, Bordbuch, November, 29, 1492. Reprint: edition Erdman, Thienemann, Stuttgart 1983.
2. M. E. Chevreul: "Recherches chimiques sur les corps gras d'origine animale," Frankreich 1823.
3. Hoechst AG: "Vom Wachs," Hoechster Beiträge zur Kenntnis der Wachse, Frankfurt 1959 ff.
4. DGF-Einheitsmethoden, Abteilung M: Wachse und Wachsprodukte, Stuttgart 1975.
5. Erläuterungen des Rates für die Zusammenarbeit auf dem Gebiet des Zollwesens 3404/1 (1992) und 3404/2 (1988): International Customs Tariff, Harmonized System HS 3404: "Artificial Waxes (Including Water-Soluble Waxes); Prepared Waxes, not Emulsified or Containing Solvents".

6. J. Haggin: "Asian Natural Gas Pipeline Proposed for Easing Energy Feedstocks Strains," *Chem. Eng. News*(1994) no. 6, 17 – 18.
7. "Wax Data," Newsletter for Marketers and Consumers of Petroleum, Natural and Synthetic Waxes, Rauch Associates Inc., Bridgewater, New Jersey (published 13 times annually).
8. Kahl & Co., Otto-Hahn-Str. 2, 22946 Trittau, personal communication, 1994.
9. "Animal and Vegetable Waxes – A Review," *Wax Data 1993,* Rauch Associates Inc., P.O. Box 6802, Bridgewater, N.J. 08807.
10. *Chem. Mark. Rep.* **244,** (1993) no. 12, 7, 11.
11. "Study Forecasts Growth for Natural and Synthetic Waxes," *Wax Data 1994,* Rauch Associates Inc., P.O. Box 6802, Bridgewater, N.J. 08807.
12. A. H. Warth: *The Chemistry and Technology of Waxes,* Reinhold Publ. Corp., New York 1956.
13. R. Büll et al.: "Vom Wachs," vol. 1; *Hoechster Beiträge zur Kenntnis der Wachse,* vol. 1 (contributions 1 – 8); vol. 2 (contributions 1 – 10), Hoechst AG, Werk Gersthofen, Augsburg 1958 – 1970.
14. E. B. Kurtz, Jr., *Fette Seifen Anstrichmittel* **61** (1959) no. 6, 505.
15. E. J. Fischer, W. Presting: *Laboratoriumshandbuch für die Untersuchung technischer Wachs-, Harz- und Ölgemenge,* VEB Wilhelm Knapp Verlag, Halle/Saale 1958.
16. L. Ivanovszki: *Wachs-Enzyklopädie,* Verlag für Chem. Industrie H. Ziolkowsky, Augsburg, vol. 1, 1954; vol. 2, 1960.
17. W. Piso, *Historia Naturalis Brasiliae* (1648).
18. J. B. De Moraes: *Ensaios sobre à Carnanbeira,* Nat. Ölinstitut, Rio de Janeiro 1942.
19. J. B. Nogueira, R. D. Machado: *Glossario da Palmeiras,* Nat. Ölinstitut, Rio de Janeiro 1950.
20. Am. Wax Importers and Refiners Assoc.: *Blue Book,* New York 1973.
21. Chem. Fabrik Dessau, DE 581 130, 1931.
22. Dtsch. Hydrierwerke, DE 658 749, 1932.
23. Degussa, DE 938 501, 1942.
24. E. Eisenring, CH 271 936, 1950.
25. W. A. Sprague, US 1 658 062, 1926.
26. I.G. Farbenind., US 1 989 626, 1932.
27. S. C. Johnson & Son, US 2 527 481, US 2 531 785, 1948.
28. G. Kronbach, *Seifen Öle Fette Wachse* **99** (1973) 97.
29. L. E. Vandenburg, E. A. Wilder, *J. Am. Oil Chem. Soc.* **44** (1967) 659.
30. S. Markley, *Econ. Bot.* **9** (1955) 39.
31. *Seifen Öle Fette Wachse* **86** (1960) 422.
32. E. A. Wilder: "Chemical Specialties Manufacturers Assn.," *61st. Mid-Year Meeting,* Washington 1975.
33. G. Illmann, *Fette Seifen Anstrichmittel* **81** (1979) no. 8, 322.
34. The Candelilla Wax Industry FONCAM; Fideicomisos Relacionados con la Cera de Candelilla; Banco Nacional de Crédito Rural, S.A. Mexico, company brochure.
35. *Chem. Mark. Rep.* **243** (1993) no. 2, 21.
36. C. Lüdecke, L. Ivanovszky: *Taschenbuch für die Wachsindustrie,* 4th ed., Wissenschaftl. Verlagsgesellschaft, Stuttgart 1958, pp. 33 – 36.
37. L. J. N. Cole, J. B. Brown, *J. Am. Oil Chem. Soc.* **37** (1960) 359.
38. *Chem. Mark. Rep.* **241** (1992) no. 6, 10, 11.
39. P. J. Manohar Rao, *Coop. Sugar/India* **13** (1981) no. 2, 1.
40. H. Schmidt: "Analytik von Naturwachsen am Beispiel des Zuckerrohrwachses," *DGF-Vortragstagung,* Mainz 1967.
41. A. Schwab: *Jojoba, ein hochwertiges Pflanzenöl aus der Wüste,* Ledermann-Verlag, Bad Wörishofen 1981.
42. E. Zander: *Handbuch der Bienenkunde in Einzeldarstellung,* Verlag E. Ulmer, Stuttgart 1946.
43. K. V. Frisch: *Aus dem Leben der Bienen,* Springer Verlag, Heidelberg 1956.
44. G. Titschak, *Fette Seifen Anstrichmittel* **71** (1969) no. 5, 369 – 379.
45. R. J. Holloway, *J. Am. Oil Chem. Soc.* **46** (1969) 189 – 190.
46. A. P. Tulloch, *J. Am. Oil Chem. Soc.* **50** (1973) 269.
47. A. P. Tulloch, *Bee World* **61** (1980) 47.
48. A. P. Tulloch, *Lipids* **5** (1970) 247.
49. A. P. Tulloch, *Chem. Phys. Lipids* **6** (1971) 235.
50. A. P. Tulloch, L. L. Hoffmann, *J. Am. Oil Chem. Soc.* **49** (1971) 696.
51. K. Stransky, M. Streibl, *Collect. Czech. Chem. Commun.* **36** (1971) no. 6, 2267.
52. K. Stransky, M. Streibl, V. Kubelka, *Collect. Czech. Chem. Commun.* **36** (1971) no. 5, 2281.
53. K. Stransky, K. Ubik, M. Streibl, *Collect. Czech. Chem. Commun.* **37** (1972) no. 12, 4099.
54. H. Brüschweiler, H. Felber, F. Schwager, *Fett Wiss. Technol.* **91** (1989) no. 2, 73 – 79.

55. H. Schmidt: "Analytische Untersuchungen an Schellackwachs," *DGF-Vortragstagung,* Freiburg 1981.
56. C. A. Anderson, *J. Am. Oil Chem. Soc.* **40** (1963) 333 – 336.
57. V. Vcelák: *Chemie und Technologie des Montanwachses,* Verlag der tschechoslowakischen Akademie der Wissenschaften, Prague 1959.
58. A. H. Warth: *The Chemistry and Technology of Waxes,* Reinhold Publ. Corp., New York 1956.
59. E. J. Fischer, W. Presting: *Laboratoriumsbuch für die Untersuchung technischer Wachs-, Harz- und Ölgemenge,* VEB Wilhelm Knapp Verlag, Halle (Saale) 1958.
60. A. Lissner, A. Thau: *Die Chemie der Braunkohle,* vols. 1 and 2, VEB Wilhelm Knapp Verlag, Halle (Saale) 1952.
61. E. Riebeck, Dissertation, Freiberg 1880.
62. Von Boyen, DE 101 373, 1897.
63. VEB Braunkohlekombinat Röblingen: *Festschrift, "75 Jahre Herstellung von Montanwachs, "* Röblingen 1972.
64. R. Büll: *Informationsschrift "Vom Wachs,"* 2 vols, Hoechst AG, Frankfurt-Höchst 1959.
65. BIOS Final Report 1759 and 1831, Technical Information and Document Unit, London 1947.FIAT Final Report 737, Field Information Agency, London 1946.
66. VEB Braunkohlekombinat Gustav Sobotka, DD 135 089, 1979 (H. J. Grauen et al.).
67. Hoechst AG, DE 954 903, 1953 (R. Schirmer); DE 957 508, 1955 (R. Schirmer); DE 1 007 918, 1955 (R. Schirmer); DE 2 855 263, 1979 (A. Riedel); DE 2 915 764, 1979 (H. Stetter); DE 2 915 802, 1979 (H. Stetter); DE 2 915 755, 1979 (H. Stetter).
68. BASF, DE 3 726 514, 1987 (L. Schuster).
69. O. Malitschek: *Hoechst Wachse, Ein Leitfaden für Wachsanwender,* Hoechst AG, SBU Wachse und Additive, Werk Gersthofen 1990.
70. R. Gächter, H. Müller: *Kunststoffadditive,* Hanser Verlag, München 1993.
71. P. Piesold: "Montanwachs als Kunststoffadditiv," in J. Edenbaum (ed.): *Plastics, Additives and Modifiers Handbook,* Van Nostrand Reinhold, New York 1992, pp. 757 – 772.

General References

72. DGF-Einheitsmethoden: *Untersuchung von Fetten, Fettprodukten und verwandten Stoffen,* Abteilung M: "Wachse und Wachsprodukte," Wissenschaftl. Verlagsges., Stuttgart 1975.
73. H. Bennet: *Industrial Waxes,* vol. 2, *Compounded Waxes,* Chem. Publ. Comp., New York 1963.
74. A. H. Warth: *The Chemistry and Technology of Waxes,* Reinhold Publ. Co., New York 1956.
75. J. Teubel, W. Schneider, R. Schmiedel: *Erdölparaffine,* VEB Dtsch. Verlag f. Grundstoffind., Leipzig 1965.
76. F. Asinger: *Die petrochemische Industrie,* Akademie-Verlag, Berlin 1971, part I, pp. 33 – 70.
77. C. Zerbe: *Mineralöle und verwandte Produkte,* 2nd ed., Springer Verlag, Berlin 1969, part 1, pp. 469 –488; part 2, pp. 581 – 593.
78. R. Sayers: *Wax – An Introduction,* Gentry Books Ltd., London 1983.
79. M. Freund et al.: *Paraffin Products,* Elsevier Sc. Publ. Co., Amsterdam – Oxford – New York 1982.

Specific References

80. European Wax Federation (EWF): *Glossary of Wax Terminology,* Brussels 1981.
81. C. Kajdas, *Seifen Öle Fette Wachse* **96** (1970) 251 –260.
82. C. Kajdas, *Seifen Öle Fette Wachse* **112** (1986) 471 –474.
83. G. Matscholl, *Coating* **14** (1981) 117 – 118.
84. I. A. Michailov, *Chem. Tech. (Leipzig)* **20** (1968) 351 – 353.
85. "Assessment and Comparison of the Composition of Food-Grade White Oils and Waxes Manufactured from Petroleum by Catalytic Hydrogenation Versus Conventional Treatment," *CONCAWE Rep.* **84/60** (1984) 18 – 26, 33 – 46.
86. European Wax Federation (EWF): *Specifications of Petroleum Waxes,* Brussels 1991.
87. F. V. Synder: Presentation at National Petroleum Refiners Assoc., Houston, Texas 1979.
88. Röhm GmbH, Spektrum 39, company brochure, Darmstadt 1992.
89. CONCAWE 1991: Report no. 91/ 33, 58 – 62 CONCAWE, Brussels.
90. R. J. Taylor, *Oil Gas J.* **91** (1993) no. 4, 48 – 50.
91. D. A. Gudelis et al., *Hydrocarbon Process.* **52** (1973) no. 9, 141.
92. M. Soudek, *Hydrocarbon Process.* **53** (1974) no. 12, 64, 66.
93. *Hydrocarbon Process.* **59** (1980) no. 9, 201.
94. L. M. L. Tobing, *Hydrocarbon Process.* **56** (1977) no. 9, 143 – 145.

95. *Hydrocarbon Process.* **59** (1980) no. 9, 180, 181, 192, 200, 201, 206.
96. *Hydrocarbon Process.* **59** (1980) no. 9, 218.
97. Shell Oil, US 4 701 254, 1987 (L. W. Maas, R. A. Geimann, C. T. Adams).
98. F. Wolf et al., *Erdöl Kohle Erdgas Petrochem.* **24** (1971) 79 – 82, 153 – 155.
99. W. Himmel et al., *Erdöl Kohle Erdgas Petrochem.* **39** (1986) 408 – 414.
100. P. Polanek et al., *Erdöl Erdgas Kohle* **109** (1993) 324 – 326.
101. U.S. Depart. of Health, Education and Welfare; Food and Drug Admin. (FDA), Part 21, Paragraph 172.886.
102. FAO Food and Nutrition Paper 52, Rome 1992.
103. BGA-Regulation XXV.
104. Mineral Hydrocarbons in Food Regulations, SI 1966, No. 1073 (1966).
105. European Wax Federation: Summaries of National Regulations Affecting Wax Usage, Brussels 1991.
106. A. Case, *Adhes. Age* **33** (1990) 28 – 31.
107. A. Aduan et al., *J. Pet. Res.* **6** (1987) 63 – 76.
108. G. Michalczyk, *Fette Seifen Anstrichmittel* **87** (1985) 481 – 486.
109. M. P. Goutam et al., *Seifen Öle Fette Wachse* **118** (1992) 1277 – 1281.
110. D. T. Kaschani, DGMK-Projekt 242, Hamburg 1992.
111. G. Grimmer et al., *Chromatographia* **9** (1976) 30 –40.
112. *Chem. Mark. Rep.* **26** (1980) no. 5, 7.
113. *Chem. Purch.* **5** (1980) 74 – 85.
114. R. Hournac, *Hydrocarbon Process.* **60** (1981) no. 1, 207 – 211.
115. A. E. Dustan, *Soap Cosmet. Chem. Spec.* **56** (1980) no. 12, 42B – 42E.
116. K. Marche, Cleveland Rubber Group Meeting, Los Angeles 1984.
117. Ministery of Agriculture, Fisheries and Food News Release 278/93, London 1993.
118. EEC-Directive 90/128 (EC 1990) and appendix lists.
119. D. Monzie, Document of Chemistry and Technology of Petroleum Waxes (CTP) No. 776 (1976), Special report of the Institute of Petroleum, London.
120. R. Franz et al., *Z. Lebensm. Technol. Verfahrenstech.* **43** (1992) 291 – 296.
121. D. van Battum, *TNO-Rep.* **B 77/2694 A** (1979) 26 –30.
122. J. Rustige, Presentation at Toxicology Forum;Special Meeting on Mineral Hydrocarbons, Green College, Oxford 1992.
123. A. O. Hanstveit, *TNO-Rep.* **R 90/1986** (1990) 18 –40.
124. DGMK Forschungsbericht 461-01, Hamburg 1991.
125. J. Davie: "Polymers, Laminations and Coatings Conference," *TAPPI Proc.* 1992, 179 – 183.
126. A. O. Hanstveit, *TNO-Rep.* **R 90/243 a** (1991) 18 –46.
127. R. Wenzel-Hartung, *Erdöl Erdgas Kohle* **108** (1992) 424 – 426.
128. J. G. E. McEwen, *Progr. Pap. Recycl.* 1992, 11 – 20
129. J. T. E. Lemmen, *TNO-Rep.* **PK 92.1051** (1992) 33, 38 – 46.
130. A. Slaughter, *Pet. Rev.* **98** (1993) 258 – 260.
131. *Hydrocarbon Process.* **72** (1993) no. 8, 36.

General References

132. A. Davidsohn, B. M. Milwidsky: *Polishes,* Leonard Hill, London 1968. A. Renfrew, P. Morgan; *Polyethylene,* 2nd ed., Iliffe & Sons, London; Interscience Publ., New York 1960.

Specific References

133. I.G. Farbenind., DE 745 425, 1939 (H. Hopff, S. Goebel).
134. I.G. Farbenind., DE 864 150, 1942 (R. Kern).
135. Allied Chemical, US 2 504 400, 1947 (M. Erchak Jr.).
136. G. Luft, *Chem. Ing. Tech.* **51** (1979) 960 – 969.
137. BASF, DE 1 720 273, 1967 (O. Büchner et al.); DE 1 927 527, 1969 (H. Pfannmüller, F. Urban, W. Hensel, O. Büchner); DE 1 951 879, 1969 (K. Stark, H. Gropper, F. Urban, H. Pfannmüller); DE 2 034 534, 1970 (F. Urban, O. Büchner, W. Hensel, K. Stark).
138. Eastman Kodak, US 2 772 259, 1953 (H. J. Hagemeyer Jr.); GB 861 277, 1959 (D. C. Hull, R. J. Schrader).
139. Allied Chemical, US 2 683 141, 1952 (M. Erchak Jr.).
140. Allied Chemical, BE 780 023, 1971 (J. E. Dench, H. Knutson, M. K. Seven).
141. Allied Chemical, DE 2 233 360, 1971 (J. E. Dench, H. Knutson, M. K. Seven).
142. M. J. Roedel, *J. Am. Chem. Soc.* **75** (1953) 6110 –6112. W. M. D. Bryant, R. C. Voter, *J. Am. Chem. Soc.* **75** (1953) 6113 – 6118. F. W. Billmeyer, *J. Am. Chem. Soc.* **75** (1953) 6118 – 6122.
143. R. Kuhn, H. Krömer G. Rossmanith, *Angew. Makromol. Chem.* **40/41** (1974) 361 – 389.
144. G. Baum, J. Boelter, *Fette Seifen Anstrichmittel* **70** (1968) 15 – 19.

145. N. Nakajima, C. F. Stark, R. D. Hoffmann, *Trans. Soc. Rheol.* **15** (1971) 647 – 662.
146. E. Brauer, H. Wiegleb, M. Helmstedt, W. Eifert, *Plaste Kautsch.* **24** (1977) 463 – 468.
147. H. Hendus, G. Schnell, *Kunststoffe* **51** (1961) 69 – 74.
148. G. Ehrlich, G. A. Mortimer, *Adv. Polym. Sci.* **7** (1970) 386 – 448.
149. Chusei Oil Co., US 5 223 122, 1992 (Y. Katayama).
150. Hoechst, DE-OS 1 929 863, 1969 (B. Diedrich, K. Rust, R. Klein).
151. Hoechst, DE-OS 1 942 734, 1969 (B. Diedrich, R. Klein).
152. Mitsui Petrochemical, DE 2 241 057, 1972 (T. Tomoshige).
153. Mitsui Petrochemical, EP-A 0 021 700, 1980 (T. Ueda).
154. Leuna-Werke, DE-OS 4 235 404, 1992 (P. Adler, J. Krüger, M. Gebauer, R.-P. Pfitzer).
155. Scholven-Chemie, DE-AS 1 301 524, 1965 (K. Schmitt et al.).
156. VEBA-Chemie, DE-OS 1 520 927, 1963 (K. Schmitt, F. Gude).
157. VEBA-Chemie, DE-OS 1 645 409, 1965 (K. Schmitt, G. Baum, F. Gude, W. Nagengast).
158. Hoechst, DE-OS 2 257 917, 1972 (M. Engelmann, R. Höltermann).
159. VEBA-Chemie, DE-OS 1 645 423, 1965 (K. Schmitt, F. Gude).
160. VEBA-Chemie, DE-OS 1 520 925, 1963 (K. Schmitt, F. Gude).
161. Scholven-Chemie, DE-OS 1 520 914, 1962 (K. Schmitt, G. Keller).
162. Hoechst, DE-OS 2 329 641, 1973 (M. Engelmann).
163. BASF, EP-A 480 190, 1991 (R. Schlund et al.).
164. Hoechst, EP-A 584 586, 1993 (G. Hohner).
165. F. Kübler, *New Sci.* (1993) Aug., 28.
166. R. Mühlhaupt, *Nachr. Chem. Tech. Lab.* **41** (1993) 1341.
167. Mitsui Petrochemical, JA 61-236 804, 1985 (T. Tsutsui, M. Kioka, N. Kashiwa).
168. Mitsui Petrochemical, JA 62-129 303, 1985 (T. Tsutsui, A. Toyota, N. Kashiwa).
169. Hoechst, DE-OS 3 743 322, 1987 (H. Lüker).
170. Exxon Chemical Patents, US 4 914 253, 1988 (M. Chang).
171. Hoechst, DE-OS 4 217 378, 1992 (H.-F. Herrmann et al.).
172. Hoechst, EP 602 509, 1993 (R. Hess, H. Voigt, L. Böhm, H.-F. Herrmann).
173. Hoechst, DE-OS 3 743 320, 1987 (A. Winter, M. Antberg, J. Rohrmann).
174. Hoechst, DE-OS 3 904468, 1989 (A. Winter et al.).
175. Mitsui Toatsu Chemicals, EP-A 563 834, 1993 (Y. Sonobe, M. Mizutani, T. Shiomura, N. Hirayama).
176. BASF, DE 922 618, 1952 (H. Hopff, H. Eilbracht).
177. Eastman Kodak, FR 1 289 766, 1960 (H. W. Coover Jr., M. A. McCall, J. E. Guillet).
178. W. G. Oakes, R. B. Richards, *J. Chem. Soc.* 1949, 2929 – 2935.
179. A. Schwarz, G. Cramer in *Polyethylene,* 2nd ed., Iliffe & Sons, London; Interscience Publ., New York 1960,, pp. 232 – 239.
180. BASF, DE-AS 1 940 686, 1969 (H.-G. Trieschmann et al.).
181. Allied Chemical, US 3 345 352, 1964 (J. J. Baron, J. P. Rahus).
182. Eastman Kodak, US 2 835 659, 1957 (J. E. Guillet).
183. Eastman Kodak, US 3 519 609, 1967 (R. L. McConnell, D. A. Weemes).
184. J. E. Guillet, H. W. Coover, *SPE J.* **16** (1960) 311 –314.
185. Eastman Kodak, US 3 562 788, 1967 (D. A. Weemes, R. L. McConnell).
186. H. W. Coover, R. L. Combs, J. E. Guillet, D. A. Weemes, *J. Appl. Polym. Sci.* **11** (1967) 1271 – 1281.
187. BASF, DE 1 364 747, 1962.
188. Eastman Kodak, US T 869 011, 1966 (M. O. Bronson).
189. A. Thalhofer in R. Vieweg, A. Schley, A. Schwarz (eds.): *Kunststoff-Handbuch,* **vol. VI,** Polyolefine, Hanser Verlag, München 1969, p. 161.
190. VEB Leuna-Werke, DD 283 730, 1988 (I. Klimek et al.).
191. Mitsui Petrochemical, DE-OS 2 126 725, 1971 (A. Mori et al.).
192. N. M. Emanuel, E. T. Denisov, Z. K. Maizus: *Liquid-Phase Oxidation of Hydrocarbons,* Plenum Press, New York 1967.
193. Hoechst, DE-OS 2 035 706, 1970 (H. Korbanka, K.-H. Stetter).
194. Hoechst, EP 28 384, 1980 (H. Korbanka et al.).
195. Allied Chemical Corp., DE-OS 1 495 938, 1963 (R. W. Bush, C. L. Kehr).
196. Allied Chemical Corp., DE-OS 1 520 008, 1964 (C. L. Kehr).
197. Allied Chemical Corp., DE-OS 1 570 652, 1965 (G. L. Braude, S. T. Phoa).
198. BASF, DE-OS 3 112 163, 1981 (A. Hettche et al.).
199. BASF, DE-OS 3 720 952, 1987 (L. Schuster et al.).

200. BASF, DE-OS 3 720 953, 1987 (L. Schuster et al.).
201. BASF, DE-OS 3 238 652, 1982 (R. W. Reffert et al.).
202. Hoechst, DE-OS 2 201 862, 1972 (H. Korbanka, K.-H. Stetter).
203. Th. Goldschmidt, EP 0 257 506, 1987 (E. Ruf).
204. Th. Goldschmidt, EP 0 258 724, 1987 (E. Ruf).
205. Th. Goldschmidt, EP 0 259 652, 1987 (E. Ruf).
206. Hüls, EP-A 0 451 352, 1990 (A. Kühnle, V. Schrenk, M. Vey, E. Wolf).
207. Mitsui Petrochemical, DE-OS 285 3862, 1978 (A. Tachi, T. Tomoshige, H. Furuta, N. Matsuzawa).
208. Petrolite Corp., US 3 590 076, 1967 (W. J. Heintzelmann, M. I. Naiman).
209. Eastman Kodak Co., US 4 028 436, 1975 (R. T. Bogan, C. M. Shelton).
210. Mitsui Petrochemical, EP 0 024 034, 1980 (T. Tomoshige, H. Furuta, A. Tachi, N. Kawamoto).
211. Eastman Kodak Co., US 3 480 580, 1965 (F. B. Joyner, R. L. McConnell).
212. Eastman Kodak Co., US 3 481 910, 1965 (M. O. Brunson).
213. Eastman Kodak Co., US 5 290 954, 1992 (T. D. Roberts, T. W. Smith, C. A. Galvin).
214. K. Rieger, *Fette Seifen Anstrichmittel* **73** (1971) 231.
215. K. Rieger, G. Grimm, *Adhäsion* **11** (1977) 311.
216. R. Köhler, *Fette Seifen Anstrichmittel* **87** (1985) 214.
217. J. Wildgruber, *Fette Seifen Anstrichmittel* **84** (1982) 75.
218. Results of investigation of Food and Drug Research Ltd., private communication of J. Maehlmann & Cie., Fortaleza, Brazil 1982.
219. J. Lange, J. Wildgruber, *Fette Seifen Anstrichmittel* **78** (1976) 62 – 69.
220. W. S. Spector: *Handbook of Toxicology,* W. B. Saunders, Philadelphia – London 1956.
221. Hoechst, Pharma Research Toxicology, unpublished results, Frankfurt 1986.
222. P. Shubik et al.: *Toxicology and Applied Pharmacology,* vol. 4, Supplement, Academic Press, New York – London 1962.
223. BIBRA 1992: "A 90 Day Feeding Study in the Rat, With Six Different Mineral Oils and Three Different Mineral Waxes," BIBRA Project no. 3, 1010, BIBRA Toxicology, Carshalton, Surrey, U.K.
224. BIBRA 1993: "A 90 Day Feeding Study in the Rat, With Two Mineral Waxes," (an Intermediate and a Micro-paraffin Blend), BIBRA Project no. 3, 1205, BIBRA Toxicology, Carshalton, Surrey, U.K.
225. CONCAWE 1993: "Report no. 93/56, White Oil and Waxes, Summary of 90 Day Studies," CONCAWE, Brussels.
226. Toxicology Forum 1992: Special Meeting on Mineral Hydrocarbons Sept. 21 – 23, 1992, Green College, Oxford, U.K. The Toxicology Forum, Washington, D.C.
227. Toxicology Forum 1993: Mineral Hydrocarbons, 1993 Winter Meeting, Feb. 15 – 17, 1993, Toxicology Forum, Washington, D.C.

Weed Control

Arnold P. Appleby, Oregon State University, Corvallis, Oregon 97331-3002, United States (Chaps. 1 – 6, 8, Sections 7.1 – 7.7, 7.11)

Franz Müller, (formerly Novartis Crop Protection, Basel), Allschwil, Switzerland (Sections 7.8 and 7.9)

Serge Carpy, Novartis Crop Protection, Basel, Switzerland (Section 7.9)

1. Introduction 4
1.1. Definitions of Weeds 4
1.2. Losses from Weeds 5
2. Biology of Weeds 5
2.1. Plant Types 5
2.1.1. Land Plants 5
2.1.2. Aquatic Plants 5
2.1.3. Parasitic Seed Plants 6
2.2. Life Cycles 6
2.2.1. Annuals 6
2.2.2. Biennials 6
2.2.3. Perennials 6
2.3. Reproduction 7
2.4. Interference Between Crops and Weeds 8
2.4.1. Competition 8
2.4.2. Allelopathy 11
2.5. Biological Variability and Adaptability in Relation to Control Systems 11
2.5.1. Species Shifts 11
2.5.2. Resistance Development within a Species 11
3. General Concepts Underlying Weed Control 12
4. Physical Methods of Weed Control 13
4.1. Manual Methods 13
4.2. Tillage Methods 13
4.3. Mowing and Cutting 14
4.4. Flooding 14
4.5. Mulching 14
4.6. Use of Heat 15
5. Biological Control of Weeds 16
5.1. Herbivore Management 16
5.2. Classical Biological Control 17
5.3. Inundative Biological Control . . 18
6. Controlling Weeds by Habitat Management 18
7. Chemical Control of Weeds 19
7.1. History and Introduction 19
7.2. Adjuvants 20
7.2.1. Surfactants 20
7.2.2. Oils 21
7.2.3. Drift Retardants 21
7.2.4. Other Adjuvants 22
7.3. Herbicide Formulations 22
7.3.1. Spray Formulations 23
7.3.1.1. Soluble Concentrates 23
7.3.1.2. Emulsifiable Concentrates 23
7.3.1.3. Wettable Powders 24
7.3.1.4. Flowables 24
7.3.1.5. Dispersible Granules or Dry Flowables 24
7.3.1.6. Invert Emulsions 25
7.3.1.7. Encapsulated Formulations 25
7.3.2. Formulations Applied Dry 25
7.3.3. Active Ingredient vs. Acid Equivalent 25
7.4. Factors Influencing Foliage-Applied Herbicides 26
7.5. Factors Influencing Soil-Applied Herbicides 30
7.5.1. Microbiological Effects 30
7.5.2. Adsorption on Soil Colloids 31
7.5.3. Chemical Decomposition 32
7.5.4. Leaching 32
7.5.5. Volatilization 32
7.5.6. Photodecomposition 32
7.5.7. Uptake and Translocation from the Soil 33
7.6. Selective Action of Herbicides . . 33
7.7. Examples of Important Herbicides and Herbicide Groups 35
7.7.1. Foliage-Active Contact Herbicides 35
7.7.2. Foliage-Active Translocated Herbicides 36
7.7.3. Soil-Active Herbicides 37
7.8. Properties of Herbicides 38
7.8.1. Diphenyl Ethers 38
7.8.2. Sulfonylureas 40
7.8.3. Chloroacetanilides 46
7.8.4. 1,3,5-Triazines 48
7.8.5. Ureas 51
7.8.6. 2,6-Dinitroanilines 55
7.8.7. Imidazolinones 58
7.8.8. Alkanoic Acids 59
7.8.9. Triazoles 62
7.8.10. Uracils 63

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_165

7.8.11. Pyridines 64
7.8.12. Hydroxybenzonitriles 65
7.8.13. (Thio)Carbamates 66
7.8.14. Organophosphates 69
7.8.15. Cyclohexanedione Oximes 71
7.8.16. Benzonitriles 72
7.8.17. Triazinones 73
7.8.18. Miscellaneous Herbicides 74
7.9. Toxicology 78
7.9.1. Diphenyl Ethers 78
7.9.2. Sulfonylureas 79
7.9.3. Chloroacetanilides 85
7.9.4. 1,3,5-Triazines 87
7.9.5. Ureas 89
7.9.6. 2,6-Dinitroanilines 92
7.9.7. Imidazolinones 94
7.9.8. Alkanoic Acids 95
7.9.9. Triazoles 97
7.9.10. Uracils 98
7.9.11. Pyridines 98
7.9.12. Hydroxybenzoniriles 99
7.9.13. [Thio] Carbamates 100
7.9.14. Organophosphorous Compounds . . 103
7.9.15. Cyclohexanedione Oximes 104
7.9.16. Benzonitriles 105
7.9.17. Triazinones 105
7.9.18. Miscellaneous Herbicides 106
7.10. Economic Aspects 109
7.11. Regulatory Aspects of Pesticide Residues 109
8. General Comments 111
9. References 111

1. Introduction

1.1. Definitions of Weeds

Weeds are unwanted plants. They have been a problem since humans first tried to grow their own crops, only to find that other plants wanted to grow there as well. Several definitions of weeds are used. A common one is, "A weed is a plant out of place." EMERSON noted that, "A weed is a plant whose virtues have not yet been discovered." A useful definition of a weed is, "A plant whose negative values outweigh its positive values." The last definition introduces the concept that not every plant often considered a weed is necessarily all bad. It can have positive values as well, such as erosion control, the ability to fix nitrogen, feed and shelter for wildlife, etc. And all the definitions imply that being a weed is not a natural condition; Mother Nature does not decide whether a plant is good or bad. This decision is made by humans. Sometimes the decision is nearly unanimous, as with Canada thistle or field bindweed, two widespread, deep-rooted noxious weeds, while people may disagree about the relative positive and negative aspects of other plants. The point is that the designation of a plant as a weed is a human decision, not a natural one.

Genus and species of weeds mentioned by common name are listed in the following:

Arrowhead	*Sagittaria* sp.
Barnyardgrass	*Echinochloa crus-galli*
Bracken fern	*Pteridium aquilinum*
Broomrape	*Orobanche* sp.
Bull thistle	*Cirsium vulgare*
Bulrush	*Scirpus* sp.
Canada thistle	*Cirsium arvense*
Cattail	*Typha* sp.
Chara	*Chara* sp.
Cheat	*Bromus secalinus*
Chickweed	*Stellaria media*
Cocklebur	*Xanthium* sp.
Coontail	*Ceratophyllum demersum*
Crabgrass	*Digitaria sanguinalis*
Curly dock	*Rumex crispus*
Dandelion	*Taraxacum officinale*
Dodder	*Cuscuta* sp.
Duckweed	*Lemna* sp.
Elodea	*Elodea* sp.
False dandelion	*Hypochaeris radicata*
Fiddleneck	*Amsinckia* sp.
Field bindweed	*Convolvulus arvensis*
Field horsetail	*Equisetum arvense*
Foxtail	*Setaria* sp.
Henbit	*Lamium amplexicaule*
Johnsongrass	*Sorghum halepense*
Kochia	*Kochia scoparia*
Lambsquarter	*Chenopodium album*
Leafy spurge	*Euphorbia esula*
Mullein	*Verbascum blatteria*
Mustard	*Brassica* sp.
Naiad	*Najas* sp.
Nightshade	*Solanum* sp.
Nitella	*Nitella* sp.
Northern jointvetch	*Aeschynomene virginica*
Pigweed	*Amaranthus* sp.
Plantain	*Plantago* sp.
Pondweed	*Potamogeton* sp.
Prickly pear	*Optunia* sp.

(Continued)

Prostrate knotweed	*Polygonum aviculare*
Quackgrass	*Elytrigia repens* or *Elymus repens*
Rush	*Juncus* sp.
Russian knapweed	*Centaurea repens*
Russian thistle	*Salsola iberica*
Sacred lotus	*Nelumbo nucifera*
Sagebrush	*Artemesia tridentata*
St. Johnswort	*Hypericum perforatum*
Strangler vine	*Morrenia odorata*
Tansy ragwort	*Senecio jacobaea*
Waterhyacinth	*Eichhornia crassipes*
Waterlettuce	*Pritia stratiotes*
Waterlily	*Nymphaea* sp.
Watermilfoil	*Myriophyllum* sp.
Wild carrot	*Daucus carota*
Wild garlic	*Allium vineale*
Witchweed	*Striga* sp.
Yellow nutsedge	*Cyperus esculentus*

1.2. Losses from Weeds

The costs of unwanted plants to human activity are enormous (→ Crop Protection). They obviously interfere with the yield and quality of the world's crops, but their detrimental effects go far beyond that. They can also clog essential waterways and canals, poison livestock, cause dermatitis in humans, induce hayfever and other respiratory problems, create fire hazards, damage pavement in parking lots and along roadways, be unsightly, interfere with utility lines, block highway signs and reduce visibility at intersections, and occasionally produce attractive but poisonous fruit, leading to human health problems or even death. The total cost of weeds in monetary terms is unknown, and most estimates are highly inexact, but there is widespread agreement that the annual cost of weeds to society worldwide greatly exceeds 10^{10} dollars.

2. Biology of Weeds

2.1. Plant Types

2.1.1. Land Plants

Most weeds on land are grasses, sedges, or broadleaves.

Grasses. Grass seedlings have only one leaf as they emerge from the seed. Their leaves are generally narrow and upright, with parallel veins. Most grasses have fibrous root systems. The growing point on seedling grasses is sheathed, and located below the soil surface during the early part of their life cycle. Some grass species are annuals; others are perennials.

Sedges are similar to grasses except that they have triangular stems and three rows of leaves. They often are listed with the grasses on pesticide labels. Most sedges are found in wet places, but principal pest species are found in fertile, well-drained soils. Yellow nutsedge is a perennial weed species that produces rhizomes and tubers.

Broadleaves. Broadleaf seedlings have two leaves as they emerge from the seed. Their leaves are generally broad with netlike veins. Broadleaves usually have a taproot and a relatively coarse root system. All actively growing broadleaf plants have exposed growing points at the end of each stem and, in some cases, in each leaf axil. Perennial broadleaf plants also may have growing points on roots and stems above and below the surface of the soil. Broadleaves include species with annual, biennial, and perennial life cycles.

Others. Weeds not in the above groups include spore-bearing plants, as field horsetail and bracken fern; conifers, such as *Juniperus* sp.; and certain monocots, such as wild garlic, *Iris* sp., cattails, and rushes.

2.1.2. Aquatic Plants

Many aquatic plants are similar to land plants, and have stems, leaves, flowers, and roots. Most act as perennial plants – dying back and becoming dormant in the fall, and beginning new growth in the spring. They are classified as:

1) Emergent (emersed) – most of the plant extends above the water surface. Examples are cattails, bulrushes, and arrowheads.
2) Floating – all or part of the plant floats on the surface. Examples are waterlilies, duckweeds, waterlettuce, and waterhyacinth.

3) Submergent (submersed) – all of the plant grows beneath the water surface. Examples are watermilfoil, elodea, naiads, pondweeds, and coontails.

Emergent and floating plants, like some land plants, have a thick cuticle on their leaves and stems, which hinders herbicide absorption. Submergent plants have a very thin cuticle on their leaves and stems, and are very susceptible to herbicide injury.

Algae are aquatic plants without true stems, leaves, or vascular systems. For control purposes, they may be classified as:

1) Plankton algae – microscopic plants floating in the water. They sometimes multiply very rapidly, and cause "blooms" in which the surface water appears soupy green, brown, or reddish brown, depending on the algal type.
2) Filamentous algae – long, thin strands of plant growth that form floating mats or long strings extending from rocks, bottom sediment, or other underwater surfaces.
3) Macroscopic freshwater algae – larger algae that look like vascular aquatic plants. The two should not be confused, however, because their control is different. Many macroscopic algae attach themselves to the bottom and grow to be as much as 60 cm tall; nevertheless, they have no true roots, stems, or leaves. Examples are chara and nitella.

2.1.3. Parasitic Seed Plants

Dodders, broomrape, witchweed, and certain mosses are important weeds found on some agricultural, ornamental, and forest plants, living on and obtaining their food from the host plants. They can severely stunt and even kill the host plants by consuming the host plants' water, food, and minerals. These plants reproduce by seeds. Some also can spread from plant to plant in close stands by vining and twining.

2.2. Life Cycles

2.2.1. Annuals

Plants with a one-year life cycle are called annuals. They grow from seed, mature, and produce seed for the next generation all within one year or less. There are two types: summer annuals and winter annuals.

Summer annuals are plants that grow from seeds that sprout in the spring. They grow, mature, produce seed, and die before winter. Examples include crabgrass, foxtails, pigweeds, lambsquarters, Russian thistle, kochia, barnyardgrass, nightshades, and many others.

Winter annuals grow from seeds that sprout in the fall. They grow, mature, produce seed, and die before summer. Examples are cheat, henbit, fiddleneck, and chickweed.

2.2.2. Biennials

Plants with a two-year life cycle are biennials. They grow from seed and develop a heavy root and compact cluster of leaves (called a rosette) during the first year. In the second year they mature, produce seed, and die. Examples are mullein, bull thistle, tansy ragwort, and wild carrot.

2.2.3. Perennials

Plants that live more than two years, and may live indefinitely, are known as perennials. Perennial plants may mature and reproduce in the first year, and then repeat the vegetative, seed production, and maturity stages during several subsequent years. In other perennials the seed maturity and production stages may be delayed for several years. Some perennial plants die back each winter; others, such as trees, may lose their leaves, but do not die back to the ground. Most perennials grow from seed; many also produce tubers, bulbs, rhizomes (below-ground root-like stems), or stolons (above-ground stems that produce roots).

Simple perennials send down a tap root and form crown buds. They normally reproduce by seeds. However, root pieces left by cultivation can also produce new plants. They do not spread vegetatively without help from humans, or machinery. Examples include dandelion, plantains, curly dock, and false dandelion.

Bulbous perennials may reproduce by seed, bulblets, or bulbs. Wild garlic, for example, produces seed and bulblets above ground, and bulbs below ground.

Creeping perennials produce seeds, but also rhizomes or stolons. Many spread by lateral roots as well. Examples of creeping perennials include johnsongrass, field bindweed, Canada thistle, Russian knapweed, yellow nutsedge, and leafy spurge.

2.3. Reproduction

Seed distribution is a major method of disseminating weed species over space and time. Anderson [11] describes a seed as "a tiny encapsulated embryonic plant." From an evolutionary standpoint, plants do not necessarily gain much advantage by producing large amounts of biomass, so in general, those plants that produce the most biomass represent developed, cultivated crops. However, there is a great evolutionary advantage associated with producing many seeds, and weeds have developed this to a high art. There are many reports in the literature of astounding levels of seed production, each one seemingly exceeding the previous. Production of over 100 000 seeds per plant is relatively common, and at least one paper reports over 11×10^6 seeds from a single specimen of tumbling pigweed [12].

If one is concerned about next year, and most of us are, it is obviously advantageous to reduce or eliminate weed-seed production. This can be accomplished in a number of ways, including tillage, applying chemicals, or mowing. A word of caution is in order about mowing, however. Sometimes large weeds that have entered the bloom stage at the time of mowing may be able to form viable seeds even after they have been cut off. Mowing should be done prior to the bloom stage if at all possible.

Many weeds form seeds in response to day-length signals. Again, plants have evolved under the pressure to reproduce their kind, so it is not uncommon to see late-germinating annual weeds forming seed heads when they are only 5 – 8 cm tall.

Weed seeds can be spread in a number of ways. A common route is contamination in crop seed, which might be transported hundreds or thousands of miles. Seeds might also become attached to vehicles, articles of clothing, etc. Besides relying on man as a disseminator, weeds can be spread by natural means as well. Many weeds have developed ingenious adaptations to help them spread. For wind dispersal, several plant families have evolved attachments that permit them to float in the wind. Composites have a tiny parachute attached to each seed that helps them cover long distances. Russian thistle, the original "tumbleweed," is well-adapted for spreading seed by wind. It grows in a generally round form, and develops an abscission layer near the soil surface so that it can readily detach when it is mature. The seeds are held with varying degrees of firmness; thus, as the Russian thistle is bounding across the land, driven by the wind, some of the seeds are dropped more quickly than others, assuring that seed will spread over a considerable area.

Many weeds are spread by movement in water. Curly dock, for example, has corky "pontoons" around each seed. Numerous species are readily moved through irrigation water or in rivers. Any grower will confirm that weed problems are always more severe following a flood.

Many weed seeds have a sufficiently hard coat so that they can pass through the digestive tract of a bird, which of course means they can be transported for many kilometers. Some weeds have adaptations such that they are transported by animals to whose fur they have become attached. A good example is cocklebur.

The way weed seeds germinate may also be very important. Seed germination is a large topic in itself, and the biochemistry and physiology of the process will not be discussed here, but certain concepts do need to be explored. For example, seeds might not germinate because they lack the proper conditions to do so. They may lack oxygen, light, or moisture. Seeds that are buried too deeply might be perfectly capable of germinating immediately if exposed to the proper conditions, but fail to do so in the soil because they lack some particular requirement. These seeds are said to be "quiescent," a condition that might also be called "enforced dormancy." When a vegetation manager tills the soil, he may in the process inadvertently bring up a new crop of weed seeds that have only been quiescent.

Other weed seeds fail to germinate even under perfect conditions because they are "dormant."

This can be caused by a number of internal conditions. Weed-seed dormancy serves as a method of dispersing seeds in time rather than space. It is easy to see how this might be an evolutionary advantage. If 100 % of a plant's seeds germinated in a particular spring, they could be vulnerable to a sudden frost, a flood, or the farmer's plow. Unfortunately for us, Mother Nature has assured that some of the seeds will germinate the following spring, some the second year, some the third year, etc. This makes the grower's job much more difficult. The person who finds a way to destroy all the dormant weed seeds in the soil cheaply and efficiently might make a more major contribution to civilization than anyone who has ever lived, and would certainly become rich and famous.

The longevity or persistence of weed seeds in the soil can significantly influence management practices. Is plowing down weed seeds a good idea or a bad idea? It all depends on the seed persistence. Burying seed that tend to degrade rapidly in the soil is an excellent way of controlling that species. However, the seeds of many species can survive for years or even decades if held in the dormant or quiescent stage. Plowing those types of seeds down merely prolongs the problem. In general, grass seeds tend to be less persistent than broadleaf seeds. Grass seeds seldom persist more than ten years, and most of them are probably degraded within the first three years. Seed from many broadleaves, on the other hand, are able to remain viable for decades. Seventy percent of mullein seeds buried in the soil were found still to be viable after 80 years [13]. Of course, the record seems to be the seed of the sacred lotus, found in a Manchurian lakebed, believed to be about 1000 years old and still viable [14].

Many weeds, mostly perennials, can spread vegetatively. Stolons are above-ground, horizontal stems, rhizomes are underground stems, and roots can spread laterally to form new plants as well. Most broadleaf creeping perennials spread by lateral roots, whereas most creeping perennial grasses spread by rhizomes.

The productivity of the underground organs of a plant can be amazing. By the end of the third season of growth, a single plant of field bindweed [15] or Canada thistle [16] is capable of penetrating to a depth of 6 – 7 m. A single Canada thistle seedling has produced a patch 18 m in diameter in three years [17]. A single tuber of yellow nutsedge planted in soil produced 1900 plants together with 6900 tubers in a patch 2 m across in one year [18]. All of this represents a tremendous productive capacity, making elimination of creeping perennial plants very challenging. It also helps explain why controlling a perennial plant while still a seedling and before it produces underground propagules is highly desirable. Understanding the methods by which weeds reproduce is the first step in designing intelligent control programs.

2.4. Interference Between Crops and Weeds

Interference is the detrimental effect of one plant on another. This may include: competition, which is the contest for limited resources within a habitat, and/or allelopathy, which is the harmful effect of one plant on another through the production of chemical compounds. Whether parasitism (dodder in alfalfa) or mechanical damage (lodging ryegrass in wheat) should be classed as interference is open for discussion.

2.4.1. Competition

Physiologically, crops and weeds are equivalent, in that both require the same resources from the environment for growth and development. Primary resources are water, nutrients, and light. When unwanted plants use up some of the resources, they are no longer available to the crop. Nearly always, this restricts crop growth and reduces yield.

Competition occurs when a resource becomes limiting. Sometimes a resource is so abundant that competition never enters the picture. If one species (crop) is able to use a resource more effectively, it may "out-compete" another species (the weed), but some fraction of the resource has nevertheless been diverted to the weed, and is thus not available to the crop, so the crop will have been reduced. For example, maize may reduce weed growth more than the weed reduces the maize, but the yield from weedy maize is still less than that from clean maize. If no use is made of the weeds for feed

or some other purpose then those resources have been wasted.

Sometimes a crop species is selected precisely because of its competitive nature. Sorghum sudangrass, for example, is rapid growing, grows very tall and thick, and can outcompete many weeds. Fall-planted canola requires weed control early in its cycle to ensure an adequate stand, but once established it outcompetes most weeds. Barley is a better competitor than wheat, and growers sometimes choose to raise barley if confronted with a weed problem that cannot be controlled by other methods. At the other extreme, onions grow slowly, barely cast a shadow, and prove very poor competitors indeed.

Two major factors are important in competition. One is the level of weed infestation. It is no surprise that as weed density increases, crop yields decline, and the whole purpose of a weed control system is to prevent that from happening. Establishing an economic threshold level is desirable, but very difficult. Leaving a few weeds in a field may have a sufficiently small effect on the crop yield in that particular season to make it non-economical to apply a control measure, but the long-term effect of allowing weeds to go to seed is difficult to predict. If a 40-year supply of seeds of a particular species is already present anyway, the grower might be well advised to forego a spraying operation during the year in question.

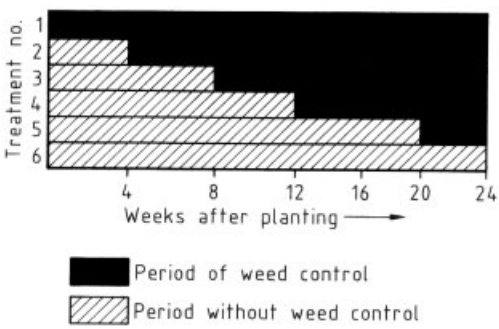

Figure 1. Treatments in an experiment to determine how long weeds can be left undisturbed before they are removed without reducing crop yield

The other major factor involves the time of competition. There are two important questions related to that subject. First, how long can one leave the weeds in place before they must be removed in order to prevent crop damage? Figure 1 illustrates the treatments involved in studying this question. Certain plots are kept weed-free all season. In other plots, weeds are allowed to grow for varying lengths of time before they are removed. A typical result would be that shown in Figure 2, from which it is apparent that if the weeds were removed within about four weeks they would not reduce the crop yield at all in this case. Other situations might lead to very different results.

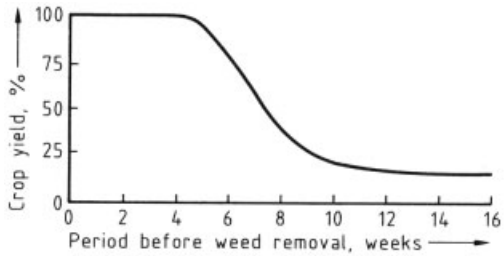

Figure 2. Typical results from the experiment shown in Figure 1

How might this information influence management decisions? It could help the manager decide between a preemergence and a postemergence herbicide. If competition begins very early in the growth of the crop, perhaps too much crop loss would have occurred by the time a postemergence herbicide was applied and had become effective. On the other hand, if the weeds do not compete for 4 – 10 weeks there might be advantages in considering use of a postemergence herbicide, waiting to see which weeds are present and to what extent. A long-standing "theorem" of weed control is that "weeds cause most damage to the crop early in the life cycle of the crop." It is true that weeds that emerge early and are left uncontrolled will be the ones that ultimately cause the most damage – but it is not always true, indeed often is not true, that competition begins early. A better generalization is "competition occurs when a resource becomes limited." That may be weeks or months after the crop emerges.

The second question involving timing is: How long must one keep a field clean of weeds before the crop has developed to the point that it can take over? This is studied in experiments like that illustrated in Figure 3, in which the plots

are kept clean for varying lengths of time, after which the weeds are allowed to grow. This may result in a graph similar to that depicted in Figure 4, which shows that the crop in question needs to be kept clean for about eight weeks; thereafter, the crop itself can prevent further weed seedlings from developing.

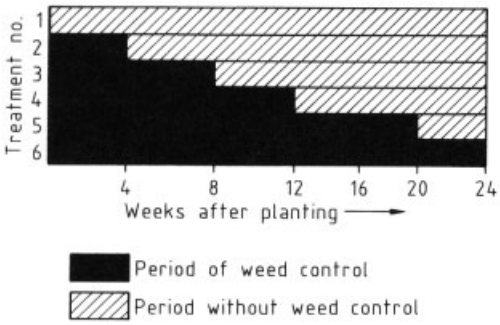

Figure 3. Treatments in an experiment to determine how long weed control must be maintained before the crop can prevent weed invasion

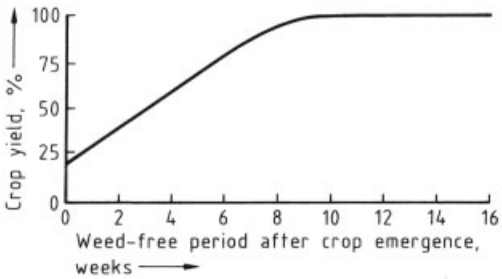

Figure 4. Typical results from the experiment in Figure 3

Why might this information be helpful in making management decisions? It might influence whether one would select a soil herbicide of long or short persistence in the soil. For a crop that develops and takes over rapidly, a short-persistence herbicide, perhaps the cheapest one, might be quite satisfactory. The information could also influence how late cultivation must be continued, and whether or not a late postemergence treatment should be applied.

When these two questions are answered, the results can be plotted on the same graph, as shown in Figure 5. This graph may at first appear confusing, but remember that the two lines were obtained from separate experiments. The two plots together reveal the "critical period of competition": the period between the time when the first line starts downward and the second line levels off, in this case approximately 4 – 8 weeks after crop emergence.

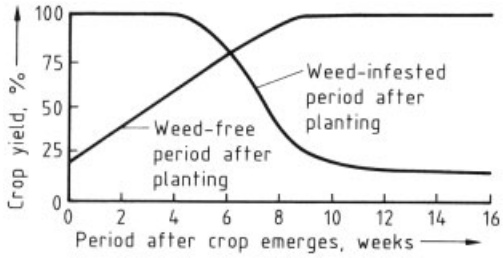

Figure 5. Results shown in Figures 2 and 4 superimposed on one graph, showing the "critical period of competition"

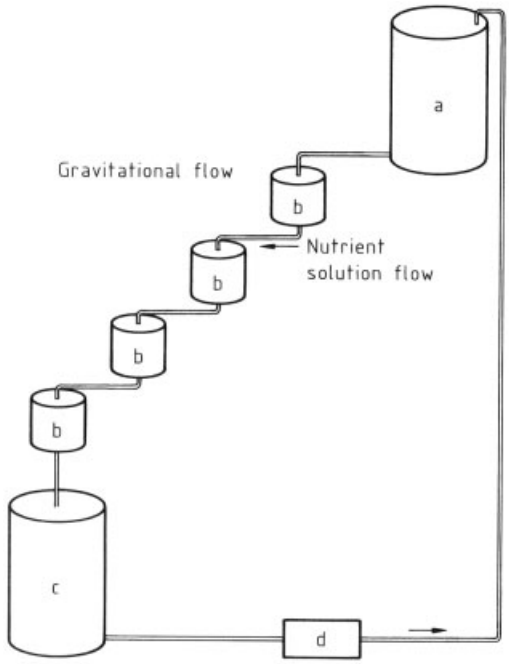

Figure 6. One line of a stairstep apparatus used to separate mechanisms of allelopathy from competition. Test lines contain pots of maize alternating with pots of *Setaria faberii*. Control lines contain only pots of maize [19].
a) Supply reservoir; b) Culture pots; c) Collection reservoir with float switch; d) Pump

2.4.2. Allelopathy

Allelopathy is the term used to describe chemical warfare between plants. This phenomenon may in fact be very common, but it is largely unrecognized. A well-known example of allelopathy involves the black walnut, which is known to inhibit plant growth around its base, but there are many other examples as well. Thus, ground-up quackgrass rhizomes distributed over the surface of the soil will essentially prevent emergence of alfalfa seedlings. Allelopathic chemicals may be released from a plant's roots, leaves, stems, berries – even as a gas from the leaves. A very large number of plants have been shown to be allelopathic.

Distinguishing in the field between interference due to competition and that from allelopathy is very difficult. Sometimes this is attempted in a greenhouse by growing the weed and the crop in separate pots but on a stairstep arrangement so that leachate from the weed is transferred downward into the crop (Fig. 6). Crop growth in a series of that type is compared to the growth of crop in a series in which the weed is not included. Competition is no longer a factor, because all the plants have plenty of minerals, nutrients, and sunlight, but the crop might still be inhibited by leachate from the weed roots.

2.5. Biological Variability and Adaptability in Relation to Control Systems

2.5.1. Species Shifts

There are literally hundreds of species in a particular climatic zone that might become important as weed species. Those species that are more adapted than others to a particular set of cultural practices will tend to become dominant. Therefore, when such practices are changed it may lead to drastic changes in the species present. This might involve tillage, flooding, drainage, mowing, grazing, or use of a herbicide. If whatever practices are applied to the field prove detrimental to the dominant species, then other species more adapted to the new practices will replace the original ones. For example, flooding for 2 – 4 months might eliminate quackgrass, but it offers excellent conditions for the invasion of aquatic-type weeds. Similarly, the use of a herbicide can cause a rapid buildup of species that are tolerant to it.

Species shifts should be expected regardless of the weed-control method employed. They can occur rapidly, even within a single season, and they must be accepted as a fact of life. Avoiding detrimental shifts requires astute management. This might include herbicide combinations, herbicide rotations, and the use of a combination of nonchemical methods to help make life miserable for surviving weeds.

2.5.2. Resistance Development within a Species

Most species contain many different individual genotypes. It should not be surprising that these may differ in their susceptibility to a particular herbicide. If a herbicide is applied consistently over a number of years, individuals within the species that have genetic resistance to the herbicide will survive and flourish after the demise of its sensitive siblings.

This process takes more time than the species shifts discussed above. Certain herbicide families seem to develop resistant populations rapidly, whereas other families develop resistance slowly or not at all. It seems to be a question of gene frequency; i.e., the number of individuals in a given field with the altered gene for resistance.

Most scientists now believe that genes for resistance to various herbicides are already present in the population even before a herbicide is introduced. Control initially may be excellent, 99.9 % or greater, but a few surviving individuals will remain. Most of the cases of herbicide resistance now being studied are examples of incomplete dominance, meaning that when homozygous resistant and suceptible individuals cross, all the offspring are resistant enough to survive another herbicide application. The resistant survivors can then proliferate at a relatively rapid pace, sometimes resulting in a field-wide infestation of the resistant weed within five or six years.

What can one do about this? Weeds are not like insects. They are anchored, and they do not fly away. Methods of preventing the development of resistance are primarily extensions of

common sense, and they have not changed over the years.

1) Use other methods of weed control in conjunction with herbicides. This may include cultivating row crops, flaming, delaying seeding to allow mechanical control of the first flush of weeds, drainage, etc.
2) Rotate crops regularly, when possible. Switching from a winter crop to a spring-planted one provides an opportunity to prepare the land at a different time of the year and perhaps kill the resistant individuals. Or by shifting to another crop it may be possible to use herbicides to which the resistant weeds are susceptible.
3) If crop rotation is not a good option, rotate herbicides annually. To be effective, the alternate herbicide must be of a different type than the one originally used, one whose biochemical mechanism of lethal action is different.
4) Resist the temptation to use higher application rates when the effectiveness of weed control starts to decline. Uncontrolled individuals in the population are genetically resistant, and increasing the herbicide dosage will probably not kill them anyway. Heavier application will simply eliminate the less resistant types, and perhaps injure the crop enough to reduce its competitiveness, permitting the most resistant types to flourish. Higher than normal application rates may actually increase the rate of resistance development.
5) Finish the job if at all possible. Hand-weeding on a 2000-ha wheat ranch is not possible. But in small fields with row crops this may be an extremely inexpensive remedy in the long term. If a new herbicide provides 99 % control the first year, the remaining 1 % might be genetically resistant, and this is the proper time to eliminate them, while they are still few in number.
6) Herbicide combinations are generally used to increase the number of weed species controlled. Such a combination will slow the development of resistance only in weeds that both herbicides affect. A species controlled by only one of the herbicides will develop resistance just as rapidly as if that herbicide were used alone.

With all the methods outlined above, it is important that preventative procedures be instituted before resistance becomes evident. By the time resistance becomes obvious there will be large numbers of resistant seeds in the soil, so even eliminating all the resistant-type plants at this late date will fail to solve the problem. Especially if the seeds are long-lasting, the infestation may persist for many years.

3. General Concepts Underlying Weed Control

An important principle to remember is that weed control is part of a more general problem – vegetation management. This involves fostering beneficial vegetation and suppressing undesirable forms. It is a broader concept than simply seeking out plants that are considered generally to be undesirable and destroying them. Growing a vigorous and healthy crop can also contribute greatly to the suppression of undesirable plants.

Specific components within a plant population respond to changes in other components. It would be short-sighted to believe that it is possible to control one plant species without affecting other components. The effects in question might well be desirable (i.e., more productive maize or sugarbeets), but they might also lead to the invasion of even more serious weed pests.

A single plant species is seldom able to exploit all the resources of a habitat, and one rarely finds a natural habitat consisting of only one species, because a mixture of species can better take advantage of the resources available. Yet much of agriculture, is devoted to growing monocultures. This puts great pressure on the system for invasion of the crop by other species that are taller, deeper-rooted, shallower-rooted, etc. In other words, a monoculture leaves unfilled "niches." *Niche* can be defined as a microhabitat potentially suitable for supporting a particular organism – a species' "place" within a community; i.e., its place with respect to space, time, or function within that community. Weed control largely consists of reducing the number of niches, or disturbing unfilled niches in such a way as to make them unfavorable for invasion by weeds. Reducing the number of unfilled

niches might be accomplished through changes in planting arrangement, adjustment of the planting time, selection of more competitive varieties, alteration of irrigation systems, etc. Niches might be disturbed in any of several ways: physically, biologically, or chemically.

Intelligent weed control thus involves the management of a complete system, not just the simple act of killing an individual species.

4. Physical Methods of Weed Control

Physical/mechanical weed control includes the physical or mechanical act of hand-pulling, hoeing, cutting, digging, tilling, burning, flooding, or mulching weeds as a way of removing them or preventing their growth. Humans have been fairly inventive in devising methods for combatting weeds, but many of the physical methods used today differ little from those used 100 years ago.

4.1. Manual Methods

Hand Weeding. In some cases, grasping a weed and pulling it out of the soil may be the only practical method for removing weeds without simultaneously damaging the crop. Even so, small seedlings of crops may be uprooted and damaged when nearby weeds are pulled out. Hand-weeding is labor-intensive and slow. Because hand-weeding is usually delayed until weeds are easily seen and also large enough to grasp, competition with crop plants and yield losses may already have occurred before the weeds are removed. In rainy weather, hand-weeding becomes difficult, impractical, and possibly damaging to the crop.

Hand-Held Tools. Use of hand-held tools makes weeding a faster and easier process compared to hand-pulling of weeds, but it retains all the general advantages and disadvantages of hand-weeding. Also, the act of hoeing or cutting may simply transplant a weed, and the use of hand-held tools may damage certain root and tuber crops. In small areas, such as home gardens, hand-weeding (with or without tools) may nevertheless be the only practical way to get rid of weeds.

4.2. Tillage Methods

Animal-Pulled Tools. The use of animals to pull a cultivator makes weeding faster and easier than hand methods. Animal-traction cultivation between crop rows may not control weeds in the crop row, but it can greatly reduce the time required for in-row hand-weeding. Crops to be managed in this way must be planted in rows or hills, and the rows and hills must be spaced far enough apart to allow the animal, tool, and operator to pass without damaging the crop. In some crops, the right implement operated at the correct speed can throw enough soil into the crop row to smother and control small weeds without harming the crop. Even so, injury to the crop may result through cutting roots located below the soil surface, or damaging edible rhizomes or roots.

Cultivation may be difficult to accomplish and also ineffective in wet weather, and it is impractical for steep lands. Though the method requires no fossil fuel to power a tractor, it does presuppose feed for the draft animals. Additionally, a farmer must invest specifically in an animal (or two) and an appropriate implement, or contract the work to someone else.

Tractor-Powered Tillage. Various sizes and types of tractors have in many places come to replace draft animals for cultivation in an effort both to control weeds and to provide improved crop-growing conditions. Tractors greatly increase the amount of land area one person can cultivate, but they also have disadvantages: they are expensive to purchase, require petroleum fuel (also expensive), require maintenance and spare parts, need trained operators, are unsuitable for operating on steep slopes, and sometimes lead to soil compaction.

Delayed tillage is a technique whereby a seedbed is prepared, weeds are allowed to germinate, and then shallow tillage is performed before most of the weeds have time to emerge or grow very large. This practice controls many weeds before the crop is planted, but is effective only when there is enough moisture to encourage weed-seed germination, and where enough time and labor are available for conducting an extra operation.

Blind Tillage. With blind tillage, fields are subjected to shallow tillage after the seed has been planted but before crop plants emerge. Success in this case depends on (1) weeds germinating before or with the crop, and (2) a crop – such as maize or sorghum – that is tolerant of physical disturbance.

Various types of implements can help control weeds growing in the crop row. Some, such as "finger weeders," are already commercially available. Others, which may involve color discriminators, TV cameras, and other "high-tech" instruments, are still under development.

There are limitations associated with all types of tillage. While dry season tillage can bring rhizomes, tubers, roots, and other plant parts to the soil surface to be killed by drying, the process requires energy, and plant parts are sometimes dragged into uninfested areas of the field, thus spreading the problem.

4.3. Mowing and Cutting

Weed plants can be mowed or cut off near ground level using hand-held, animal-pulled, or power-driven tools. While this practice does effectively kill some annual weeds, it is ineffective against weeds that are able to regenerate after being mowed.

Removing part of a weed by mowing often helps control weed-caused problems (competition, interference with the harvest, and the production of weed seed). Also, the dead plants can help protect the soil against erosion caused by wind and rain, and from compaction from vehicular or foot traffic.

While often suitable for pastures, ditch banks, roadsides, and the space between rows of woody perennial crops, mowing is ineffective for controlling creeping or low-growing weeds that escape most of the mowing action.

Principle of Carbohydrate Starvation. Removing the top part of perennial plant (shoots and leaves) by tillage or mowing interrupts the vital flow of carbohydrates to below-ground storage organs (roots, tubers, rhizomes). When the aerial part of a plant is removed, the plant uses root reserves to produce new top growth, and it continues to use root reserves for a short period even after new growth is visible. Repeated hand-cutting or mowing to remove weed top growth 10 to 15 days after new growth appears is the most efficient method for depleting root reserves in most species. To be effective, the process must be repeated several times without ever allowing the above-ground portion to grow for more than approximately 15 days after emergence, or more than about 3 weeks between tillage operations.

4.4. Flooding

Flooding, a centuries-old weed control method for rice production, is also used in nurseries, for fields receiving transplanted crops, and for direct-broadcast seeded fields. For crops such as rice that can withstand aquatic conditions, flooding can be an effective method of controlling weeds. For crops that cannot withstand aquatic conditions, flooding can nevertheless be used to control weeds during a fallow season or a rotation period. For example, a perennial grass might be killed by flooding so that after the water has been removed an upland crop can be planted.

Flooding requires relatively level fields with soils that prevent the water from quickly percolating into the ground. Accompanying requirements include a perimeter barrier (levee or dike) to contain the water, a method for draining the water off the field, a water source that provides enough volume for the depth and duration needed to prevent weed emergence or to kill existing weeds, and ability on the grower's part to practice flooding effectively.

4.5. Mulching

Mulching for weed control involves the use of any of several materials to prevent light from reaching the soil surface. Germinating weeds then die for lack of necessary light. Mulching also may conserve moisture by reducing runoff and evapotranspiration, and by lowering the temperature of the soil. Mulching may add organic matter to the soil as well. It is most frequently practiced with high-value crops grown on small areas.

Plant parts (leaves, stems, etc.) are commonly used as the mulching agent. Plants may

be mowed and left where they fall, or gathered, stored, and used for mulching later. Rice straw, rice hulls, sawdust, bark chips, newspaper, and large but immature weeds that have not formed seed are all potential mulching materials. Care must nevertheless be taken to assure that plant mulch does not contain weed or crop seeds (weed prevention).

Thin sheets of polyethylene are used as a mulch in some farming operations. Black plastic can absorb heat, raising the air temperature around crop plants and accelerating their growth, whereas clear plastic traps the sun's energy to heat the soil and air under the plastic.

These conditions can be helpful in certain situations, but damaging in others. The temperature under clear plastic can become high enough to kill many organisms, including seeds and vegetative parts of some weeds. Plastic mulch involves the cost of purchase as well as that of the labor (and possibly specialized equipment) to place it in the field. If the plastic does not self-disintegrate, labor will also be required later to remove it. Some perennial weeds, such as yellow nutsedge, can penetrate plastic mulch.

Mulching can be combined with inter-row cultivation in the case of row crops. Mulch is placed in the crop row, either around an established crop or with holes cut to allow crop seedlings to emerge. Cultivation is then used to control weeds between the rows. Obviously, mulch placed between crop rows would not control weeds within the row.

4.6. Use of Heat

Heat application is often an economical and practical means of controlling weeds. It greatly reduces the amount of hand labor required, particularly in combination with other weed-control practices. Heat is also used widely to dispose of accumulated vegetation that is unsightly or poses a health hazard.

Burning cannot be equaled for controlling weeds in some situations. For example, existing vegetative cover with low economic value can be eliminated in this way from rangelands, permitting the establishment of more desirable grasses. Controlled burning has been used for many years as a cultural practice in forest management to eliminate excessive fuel accumulation, thus reducing the risk of forest fires. Site-preparation to establish new plantings and to release desirable species is often accomplished by controlled burning. A regimen of three to four annual burns is one of the least expensive ways to remove hardwood undercover in a pine forest. Pine trees are capable of withstanding heat that kills undesirable hardwood species. Suitable precautions are of course required to prevent the fire from escaping.

In seeking to take advantage of flame, man has tried to improve upon nature. Special equipment designed to burn petroleum products as fuel is able to create temperatures of about 1 100 °C that kill only young plants. Flame may be used either as a nonselective method of killing all plants, or to kill weeds selectively without injuring a crop. A hooded burner is often used for selective flaming, but it can also be adapted for nonselective use.

Flaming does not necessarily kill by direct combustion. Burning of plant tissue occurs only after repeated flamings, and it should be kept at a minimum to avoid crop injury. The death of a plant by flame appears to be associated with injury to cells of the leaves and stem. With the cells destroyed, the seedling has no means of recovery or regeneration. Selective killing of weeds and crop plants with flame is accomplished in several ways. Different types of plants are able to withstand different amounts of heat, depending on their stem structure, age, size, and shape. Weed kill is accomplished by moving an intense flame along the base of a weed-infested crop. The intensity and duration of exposure are important factors in obtaining effective results. Desired exposures are obtained by varying the tractor speed. The size and intensity of the flame are controlled by an orifice tip and varying the fuel pressure. The critical temperature above which cambium cells are subject to injury appears to be about 55 °C. An initial flaming on cotton is normally conducted when stems are about 0.5 cm in diameter at ground level. At this stage, the cambium layer in the cotton is protected by an insulating layer. Succulent weed seedlings have no such protection, and the heat needs to travel only a short distance to reach vital cells. The larger leaf-surface area of the weeds absorbs more heat than the round surface of the cotton stem. Selective flaming is performed in a wide variety of crop, including cotton, corn, soybeans, and grain

sorghum, as well as such other crops as blueberries, raspberries, and grapes. Experimental use of flame has shown promise in such vegetable crops as onions, carrots, asparagus, and potatoes.

Alfalfa is flamed to control dodder and other weeds; flame is also used to control weeds in certain grasses and to obtain cleaner grass-seed crops; potato vines are flamed when the tubers are mature to facilitate mechanical harvesting; and drainage and irrigation ditches are kept free from weeds by nonselective flaming.

The proposed objectives of using heat for weed control should be examined carefully. Very often, too much is expected from burning or flaming. People not fully acquainted with the limitations of these methods are then disappointed by the results. For example, open-field burning is more effective in disposing of accumulated vegetation than in the killing of weeds. It is often used after weeds have already matured and viable seeds are present. In such cases, fire seldom reduces the viability of scattered seeds enough to produce any practical benefit.

Several flaming applications may be necessary to ensure satisfactory weed control. The initial cost of a flamer is often much greater than that of a sprayer. In addition, the heat generated by the burner may prove to be more than the operator can handle. If the weeds in question are as resistant to flame as the crop, then selective flaming is not feasible. A flamer is also generally more difficult to adjust than a sprayer. In most crops, weed control cannot be achieved with flame alone. It is most effective when used as a supplement to other weed-control practices, such as pre-emergence and postemergence herbicides (see Chap. 7).

Despite its limitations, flaming also has a number of important advantages in weed control. With the increased use of herbicides and increasing awareness and concern regarding residues in soil and crops, the fact that flaming is free of any residue hazard has special merit. With a well-adjusted flamer, results are usually not greatly influenced by environmental conditions, and there are no drift hazards to adjacent crops. Flaming can often reduce other plant pests such as insects and diseases, facilitate mechanical harvesting by eliminating trash, serve as a control measure when the field is too wet for mechanical methods, and permit almost immediate inspection of weed control.

The use of hot water has been tested as an alternative approach to weed control. It appears to kill some weeds more effectively than others and is not expected to be effective on perennial weeds.

5. Biological Control of Weeds

Biological control of weeds consists of the deliberate use of natural enemies to reduce the population of a target weed (→ Biological Control). In other words, it is the control of weeds by biological organisms (excluding man). Biocontrol methods include herbivore management as well as classical and inundative approaches.

5.1. Herbivore Management

This approach involves controlling weeds by managing animals, which can include geese, sheep, goats, grass carp, manatees, marisa snails, and others. The animals are in turn manipulated directly by man.

There are reports of success in cleaning up brushy, weedy pastures with a combination of cattle, sheep, and goats. These animals tend to eat different plants, so with careful management many undesirable weeds can be kept suppressed. Sheep have been used to control tansy ragwort and leafy spurge. Use of grazing animals requires considerable organisation and supervision. Grazing must be intense enough to force the animals to eat some of the less desirable weeds; on the other hand, they must not be allowed to overgraze a pasture to the detriment of the grass. Some innovative work is being conducted in "bonding" animals of different species so that they will graze together and perhaps provide some protection against predators. Also, methods are being developed for cheaper fencing methods based on a remote "shock" treatment, where a receiver and shocking device around the animal's neck senses a single wire on the soil surface. The animals soon learn to avoid getting too close to such a "fence."

Before the introduction of selective herbicides, geese were commonly used to control weeds in peppermint, strawberries, and other

crops. Herbivorous fish have been used for successful control of aquatic weeds. Grass carp, which are sterilized before release, have been used with varying degrees of success. "Sea cows" or manatees have been used to help keep canals clean. These are large aquatic mammals that are able to eat large quantities of aquatic vegetation. Unfortunately, vandalism and injury from boats have reduced their number and limited their success.

Weed control by herbivore management is not without costs. The animals must be bought, watered and fed, protected from predators, and controlled so that they eat weeds in all areas of the field (not just the shady spots). Other practical problems arise as well, such as the fact that the animal population remains constant and does not respond to increases or decreases in weed populations.

5.2. Classical Biological Control

Biological control involves the importation of exotic natural enemies of a particular weed species. These have mostly been insects, but can also include microorganisms. Such biocontrol agents are most successful against imported weeds, because the corresponding biocontrol agents are imported with added care to ensure that their natural enemies are left behind.

Preliminary determinations must be made on the feasibility of importing an insect for a particular weed. This involves establishing the severity of the weed problem and the weed's relative taxonomic relationship with desirable crops.

Foreign exploration is carried out in that area of the world where the weed itself originated. Associated enemies of the weed are studied, and the specificities of various insects are determined. A geographic distribution of the proposed biocontrol agent that coincides exactly with the geographic distribution of the weed, provides relatively good assurance that the insect in question is restricted to that particular weed. The taxonomy of the weed is studied as well, and if it is closely related to a desirable crop, extreme care must be taken to make sure that the insect will not attack the crop. For example, an insect known to control mustard weeds (*Brassica* sp.) would be viewed with great skepticism because of the existence of numerous *Brassica* crops, such as cabbage, cauliflower, and broccoli.

The most critical test is the "starvation" test. In this case, the biocontrol agent is caged with a large variety of desirable plants. To pass the test, it must starve to death before it will eat any plant other than the target weed. On rare occasions, limited feeding on another species is allowed if it does not present an economic threat. For example, the cinnabar moth, imported to control tansy ragwort, also eats dusty miller, an ornamental that was judged not sufficiently important to halt the project.

Once a biocontrol agent has passed its specificity tests it is imported, after which it undergoes a period of quarantine. This is a time to inspect for accompanying predators and pathogens, and to make quite sure that no other insect has inadvertently been included in the shipment. The species is then introduced into the area where control is intended, and nurtured as much as possible. If the insect survives but spreads only slowly, governmental agencies may physically assist in the spread of the insect.

Classical biological weed control generally follows a cyclical pattern; i.e., as the predator multiplies, the weed population declines. This causes some of the predators to starve, and the insect population then declines, with a subsequent increase in the population of the weed. The cycle continues indefinitely. Success in such a program depends upon whether or not the predator can maintain the weed infestation mostly below its economic threshold level. Classical biological control generally does not work in cropland, because the economic thresholds are too low and the process is too slow. It is most successful in rangeland and waste areas.

Examples of successful programs include prickly pear cactus control in Australia by the moth *Cactoblastis cactorium*. In the United States, St. Johnswort has been controlled with the Chrysolina beetle, and tansy ragwort has been controlled with the cinnabar moth and the flea beetle. A rust has been used for control of rush skeletonweed in Australia. Several other agents have been released as well, and early evaluation suggests eventual success.

Some of the desirable aspects of classical biological control include the fact that there are no chemical residues and no drift hazard, it can work well in areas inaccessible to other meth-

ods, it is self-perpetuating, it is highly energy efficient, and it is almost always highly selective. Possible disadvantages include disagreements over whether a particular species is a weed in all situations. Moreover, biological control is usually not effective in low threshold situations, it controls only one weed species (whereas a complex of weeds is usually present), it may be too slow for cropland, any errors in introduction may prove much less reversible than mechanical or chemical mistakes, and natural enemies and weather may limit its effectiveness. Another consideration, neither advantageous nor disadvantageous, is that there is no profit motive for private enterprise to become involved in classical biological control, so the government must pay for the anticipated benefits to private businessmen (growers). There is also always the possibility of the development of resistance. This is likely to occur very much more slowly than with chemicals, because as the target weed evolves, the predator might also evolve, which is of course impossible with a herbicide.

5.3. Inundative Biological Control

Inundative control agents are often called "bioherbicides." This method involves the periodic release of a very large population of a particular biocontrol agent so that it simply overwhelms the target weed. Suitable agents are microbes, either bacteria or fungi. Such a biocontrol agent is not self-perpetuating at a level sufficient to maintain adequate long-term control.

Marketing and usage is the same as in the case of chemical herbicides; the agent is mixed with water and sprayed on the field. At least three such products have been introduced commercially. Collego is used in rice for control of northern jointvetch. Devine is used for control of strangler vine in Florida citrus. Luboa-2 is used for control of dodder on soybeans in China. A number of other similar products are nearing commercialization.

Some of the desirable aspects of inundative control agents are that they are probably safer to humans than many chemical herbicides, they are not likely to harm non-target plants to a significant extent, and they may control certain weeds for which no chemical herbicides are available. Disadvantages include the control of only one weed species, so that it will probably still be necessary to mix the control agent with chemical herbicides, thus adding to the cost. Bioherbicides are probably affected more by the environment than are chemicals, though the latter can also be affected. Their shelf life is an additional concern for the marketing organization. In contrast to classical control methods, the person who benefits from inundative control pays the cost. The process of preparing a bioherbicide might be patentable, but the agent itself is not.

Clearly, biological control is an important approach to weed control in many situations. It can be a safe, highly cost efficient method, and it can provide an alternative to chemical herbicides. Because of the nature of the process, it is unlikely ever completely to replace other methods of weed control. Considerable research is being conducted around the world to supplement existing biocontrol agents.

6. Controlling Weeds by Habitat Management

Managing the habitat to favor desired vegetation and hamper undesirable vegetation is at the very core of weed control. Yet, it is the place where the "art" of weed management comes in, rather than the "science." Certainly everything that happens is based upon scientific principles, but habitat management sometimes is difficult to describe in a "1 – 2 – 3" manner. More than anything, it is the "art of good farming," and some people are naturally better at it than others.

There are several specific ways in which one might possibly encourage desirable vegetation and hinder weeds. One of these involves crop competition. Selecting a more competitive species (e.g., planting barley instead of wheat), growing a taller variety, growing earlier-season crops so that they emerge before most of the weeds, etc., are all measures designed to take advantage of competition from the crop on the weeds. Because the objective is to grow high-yielding and high-quality crops anyway, this approach amounts essentially to "free" weed control, and it can be quite significant. Sudangrass is sometimes planted primarily as an interim crop, selected because it emerges quickly and grows tall, thus protecting the soil from erosion

and competing effectively with later-emerging weeds. Irrigated alfalfa grown for hay can noticeably reduce the stand and vigor of patches of Canada thistle. The alfalfa is also deep-rooted, and it can compete directly, allowing the alfalfa to respond more quickly to periodic mowing than the thistle.

Planting density and arrangement can be manipulated to increase the competitiveness of a crop. Many row crops are traditionally grown in 1-m rows, primarily because that is what was required to permit a horse to walk between the rows for tillage purposes. Clearly, that is no longer needed, and a considerable trend toward narrower row spacing has been evident. Narrow-planted maize is more competitive to weeds, and it can lead to higher yields than maize planted in wide rows. Snap beans have been successfully grown in a solid seeding, but this method is hazardous because it apparently reduces the air movement among the beans to such an extent that white mold becomes a more serious problem.

Crop rotation has been shown to be highly desirable in many ways, including insect control, disease control, unknown soil factors, and certainly weed control. If alternative crops are available, ones that will yield an economic return to the grower, then this is a cheap and simple weed-control method. Planting in the spring rather than in the fall allows spring tillage, thus destroying winter annual weeds that may have built up with previous winter annual crops. Switching to a row crop may allow for different tillage systems than with a solid-seeded crop. Rotating to a different crop may also permit use of a different herbicide program, thus helping prevent species shifts and the development of resistant weed populations.

Other methods have proved successful for good managers as well. Placing the fertilizer in such a way as to benefit the crop while benefiting the weeds as little as possible has helped. Using delayed tillage or blind tillage – with careful attention to timing – also is useful. Adjusting mowing height can be important in turf areas. On golf courses, controlling traffic may reduce a problem with prostrate knotweed, allowing the grass itself to control the weed. Drip irrigation tends to water a crop more efficiently than it does a weed. Adroit grazing management, including rotation of pastures, proper placement of water and salt, etc., may be beneficial to pasture grasses at the expense of grass weeds.

Although habitat management is difficult to define and direct in specific terms, it may in fact be the most significant method of all for controlling weeds.

7. Chemical Control of Weeds

7.1. History and Introduction

The science of herbicides is relatively new, but the use of herbicides is not. Indeed, chemicals were used even in ancient times to control unwanted vegetation. In the Fourth Century B.C., THEOPHRASTUS [20] reported that trees could be killed by pouring oil (presumably olive oil) over their roots. PLINY [21] referred to the Greek DEMOCRITUS (Fifth Century B.C.), who proposed that forests could be cleared by sprinkling the tree roots with hemlock juice in which lupin flowers had been soaked. Application of salt to the fields of enemies was often used as means of punisment; thus, in 146 B.C., following the sack of Carthage, the Romans spread salt over the surrounding fields to prevent the growing of crops.

Chemicals were first used for selective weed control in the late 1890s. Several copper salts, iron sulfate, sulfuric acid, and other chemicals were used rather extensively in Europe for control of certain broadleaf weeds in small grains. Usage of these materials never became highly popular in the United States, however. Through the 1920s and 1930s, considerable amounts of nonselective-type herbicides were used, particularly to control perennial weeds. These included sodium chlorate, the borates, and inorganic arsenicals. During the 1920s and 1930s plant physiologists were also working diligently with several growth-regulator compounds. One of these, 2,4-D (see Section 7.8.8) proved to be very potent as a growth regulator, and it was patented for that purpose in the early 1940s. But 2,4-D had one major disadvantage as a growth regulator: If dosages were too high, it killed the plants to which it was applied. At first, plant physiologists using these materials did not recognize this property as being useful, but during World War II the United States government began efforts toward finding chemical tools for destroying an

enemy's crops. It was during this period of secrecy that the weed-killing properties of 2,4-D were first recognized. The British reached the same conclusion regarding MCPA during the same period. Immediately after the war these phenoxy herbicides became highly publicized, and were widely used for the control of broadleaf weeds in grass crops. The impact on American and European agriculture was so dramatic that chemical companies immediately began searching for other chemical tools, and the modern age of herbicides had begun. A trickle of new herbicides quickly became a stream; until today 150 – 200 compounds worldwide are used commercially as herbicides. The dollar value of herbicides used in the United States now exceeds that of insecticides and fungicides combined.

Herbicide use has been controversial, particularly in the developed countries, since about the 1970s or 1980s. The use of any biologically active chemical carries risks, and there are considerable differences of opinion on how extensive and acceptable those risks are with respect to herbicides. With any method, the risks involved must be compared to the benefits obtained. Not surprisingly, many members of the public are less impressed with the potential benefits of herbicides than are vegetation managers.

Clearly, there are many advantages to using chemicals for weed control compared to other methods. It takes much less time to spray a field than to till it or hoe it. It may also be the least expensive method, especially when compared to hand-weeding. Many herbicides can damage the root system as well as the above-ground parts of weeds. It may actually be the only practical technique for controlling weeds that emerge with a solid-seeded crop, such as a small-grain crop, in which tillage is seldom practical and fields are too large for hand-weeding.

Nevertheless, benefits from herbicides must be balanced against the risks involved. Herbicides can drift from a target area and harm non-target plants. Some can persist in the soil and harm subsequent crops, be diverted to non-target areas, or enter the groundwater. They can be used carelessly, resulting in the denuding of a site, which may in turn lead to erosion, loss of wildlife cover, etc. Some herbicides are moderately toxic, and if used carelessly they can lead to inadvertent poisoning. The governments in most developed countries examine the potential risks and benefits of specific herbicides, regulate their registration, and monitor their use after introduction. This matter is discussed more thoroughly in Section 7.11 dealing with regulatory aspects.

7.2. Adjuvants

An adjuvant can be defined as "a substance added to a prescription to aid in the functioning of the main ingredient." Possible reasons for incorporating additives into a spray tank for pesticides include improved wetting, evaporation reduction, increased penetration, improved translocation, extended weatherability, retarded release, pH adjustment, enhanced deposition, compatibility, drift retardation, or odor inhibition. A few adjuvant types that are of particular importance for herbicides are discussed below.

7.2.1. Surfactants

Surfactant molecules are made up of a lipophilic portion and a hydrophilic portion. The lipophilic portion usually consists of long-chain hydrocarbons or benzenoid ring structures, which are very low in water solubility but high in oil solubility. The hydrophilic end has a high affinity for water.

The primary classes of surfactants are anionic, cationic, and nonionic based on the chemical structure of the hydrophilic portion of the molecule. An anionic surfactant ionizes in water to form a negatively charged hydrophilic portion; a cationic surfactant ionizes to form a positively charged hydrophilic part. These two types can be very useful for specific purposes, but they may also react with contaminants in a spray solution. For example, ionic-type surfactants might form insoluble precipitates in hard water. Most of the commonly used agricultural surfactants are nonionic. These are easy to use, nontoxic, and unaffected by hard water. An example of each type of surfactant is given in the following:

Anionic:	sodium stearate $C_{17}H_{35}COO^{-}Na^{+}$
Cationic:	cetyltrimethylammonium bromide $C_{16}H_{33}(CH_3)_2N^{+}Br^{-}$
Nonionic:	poly(ethylene oxide) lauryl alcohol $C_{12}H_{25}O(CH_2CH_2O)_{23}H$

A newer form of nonionic surfactant is the silicone type, an example of which is shown here:

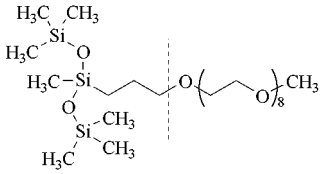

Specific surfactants can be used in many ways, including as wetting agents, dispersing agents, emulsifiers, foaming agents, detergents, etc. Some of these uses are primarily related to the formulation of a commercial product, as discussed in Section 7.3.

A wetting agent is added to herbicides to cause spray droplets to spread over the surface of the leaves, thus increasing coverage and eventual uptake of the herbicide. All the details of the process are not yet well understood, although it is known that maximum reduction of spray-droplet surface tension can usually be achieved with a wetting agent concentration of less than 0.1 %. This means that increasing the surfactant content will not cause the spray droplet to spread out any farther. But it is also known that both uptake and effectiveness of many herbicides continue to increase as the wetting agent concentration is increased above 0.1 %. This must mean that there are other ways in which a surfactant improves herbicide uptake besides simply reducing the surface tension of the spray solution.

Intensive research with surfactants has usually shown that a particular surfactant is ideal for a specific herbicide on a particular species. This of course greatly complicates the problems involved in making recommendations for herbicides that might be used on a wide range of weed species. As a result, recommendations usually refer to a small number of common surfactants shown to be reasonably effective over a wide range of herbicides and species. Household detergents are designed for an entirely different purpose, and are usually considerably less effective as adjuvants for herbicide sprays.

Increasingly popular are formulations containing phytobland oils and a significant concentration of surfactant (generally 4 – 10 %). These products are more effective in improving herbicide retention and uptake than surfactants alone. Some herbicide labels actually specify their use, usually at a rate of about 2 L/ha.

7.2.2. Oils

Oils have been used in herbicide sprays for many years. Such oils can be divided into three types:

1) Phytotoxic, nonselective: These are heavy oils with a high degree of unsaturation (i.e., many double and triple bonds and aromatic groups). The high degree of unsaturation makes the molecules readily subject to sulfonation, hence giving them a low unsulfonatable residue (USR) rating. Oils in this category are generally added to knockdown-type herbicides for road spraying or canal and ditchbank spraying, or to brush killers to facilitate uptake through the bark. Diesel fuel is an example.
2) Phytotoxic, selective: These oils are not actually adjuvants because they may be sprayed directly, without mixing with water or combination with herbicides. They are intermediate in terms of viscosity and percent USR, and are used mostly for broadleaf weed control in certain crops such as those of the carrot family. The carrot family apparently is better able to withstand the effects of oil on their membrane systems than most common weeds, so the crop plant is not harmed even though the weeds are controlled.
3) Phytobland oil – water emulsions: These are light oils that are nearly saturated chemically, and thus have a high %-USR. The oils themselves are not phytotoxic, but are often added to herbicides that require help in penetrating plant leaves. They are sometimes called corn oils, superior spray oils, supreme spray oils, and other common names.

7.2.3. Drift Retardants

When a spray solution is atomized by being forced through a spray orifice under pressure, a large number of extremely fine spray droplets accompany the normal-size droplets. These "fines" can easily be diverted from the target area with even a slight breeze. If the herbicide is sufficiently phytotoxic, if the breeze is sufficiently strong, or if the numbers of fines produced by the spraying operation is sufficiently large, serious damage to non-target species can result. For this

reason, several products have been developed to reduce the number of fines, thus reducing drift. These are mostly long-chain polymers, such as anionic polyacrylamide. Most such products can simply be added to the spray tank, and do not require any special application equipment. Care must be taken to ensure that the resulting change in character of the spray solution does not result in increased damage to the sprayed crop.

7.2.4. Other Adjuvants

Foaming agents are most often used as a way of generating a glob of foam at either end of the spray boom to serve as a marker.

An emetic causes immediate vomiting to reduce the hazard of accidental ingestion.

A stenching agent is sometimes added to the more toxic products to help prevent a pesticide from being mistaken for a soft drink, etc.

7.3. Herbicide Formulations

Once a commercial pesticide has been selected a decision must be made about which formulation will be used. There may be several available formulations for any given product. In this case, proper formulation selection can mean the difference between failure and success. Proper handling and application varies from one type of formulation to another. Therefore, it is essential that the operator have a basic knowledge of pesticide formulations.

In general, formulating means preparing a product for practical use. The formulation can influence the accuracy of application, the effectiveness of pest control in a particular situation, the selectivity of certain pesticides, convenience to the operator, and, very importantly, the cost of the treatment.

Pesticides are applied mainly as sprays or as dry materials such as granules or pellets. Some fumigants such as bromomethane are applied in gaseous form. Spray formulations include water solutions, oil-soluble liquids, soluble powders, emulsifiable concentrates, flowable formulations, wettable powders, dispersible granules or dry flowables (two designations for the same thing), invert emulsions, and encapsulated formulations. Dry materials are formulated as granules, dusts, or water-soluble pellets.

Many factors must be considered in deciding how to formulate a specific active ingredient. These can include:

1) Chemical and physical properties of the pesticide: The solubility of the pesticide in various solvents is of major importance, as will be discussed later. The volatility of the material may also dictate the preferred type of formulation.
2) Intended use: If a herbicide is intended for use in a foliar or bark application for control of waxy or brushy species it will likely be part of some type of oily formulation, whereas if it is intended as a selective treatment it is more likely to be formulated with less oil and less wetting agents. If designed to control small weeds in a growing crop through soil activity it may be formulated as granules to allow it to filter through the foliage canopy and reach the soil.
3) Geographical area of use: The customs of growers differ in various parts of the world. Many farmers own granular spreaders, and have beome accustomed to using granular products, whereas in more arid areas granules tend to be less consistent in performance, and are less readily accepted by growers.
4) Materials available for formulation: Some manufactures have switched formulations simply due to a lack of particular solvents, emulsifiers, or containers.
5) Economics: An economic analysis could include several of the factors cited above. The cost of formulating agents and containers as well as relative shipping costs and other factors must all be considered. In general, formulations with a low percentage of active ingredient (such as granules) are most expensive to the grower. Large increases in the costs of solvents and emulsifiers in recent years has changed the economic picture relative to emulsifiable concentrates. The relative cost of formulation may not be of major importance with certain high-value crops, but in other crops the cost of the product will be crucial for acceptance on the part of the grower.
6) Convenience: Some formulations are easier to measure, result in fewer dust problems,

come in more readily disposable containers, etc., and these factors may also influence acceptability in the marketplace.

7.3.1. Spray Formulations

7.3.1.1. Soluble Concentrates

With products of this type the active ingredient can be dissolved readily in either water or oil to form a true solution. In the case of a water-soluble pesticide the manufacturer may dissolve the material in a small amount of water and sell the product as a concentrate for further dilution by the user. In general, such a pesticide should be soluble at least to the extent of 10 % in water in order to be suitable for distribution as a liquid concentrate.

The manufacturer may alternatively choose to sell a water-soluble pesticide as a dry powder for mixing with water by the user. Usually the content of active ingredient ranges from 50 – 80 %, the remainder being adjuvants and diluent.

Adjuvants may be added to water-soluble products to improve their application and effectiveness. Wetting agents are commonly incorporated, and sequestering agents are sometimes added to prevent precipitation in hard water.

Major advantages of water-soluble pesticides are relatively low cost of formulation and the fact that no agitation is required in the spray tank once the materials have been thoroughly dissolved. Problems that may arise from the use of soluble pesticide formulations are as follows:

1) They may react with hard water, causing clogging of the spray equipment or rendering the pesticide nontoxic.
2) The active ingredient may not enter the plant foliage readily, due perhaps to the excessive surface tension of water, which may leave pesticide crystals on the foliage. This actually may be a desirable characteristic when selectivity of the compound is based upon differential wetting, but it would be undesirable if the goal is nonselective treatment, or if the weeds being treated are waxy.
3) The chemical may be readily leached through the soil by water. This can be an advantage when applying a herbicide to control deep-rooted perennial weeds; however, with selective pre-emergence herbicides it can be an undesirable property because the herbicide is subject to leaching by excess rain or irrigation, resulting in only a brief residual control, ineffective performance, or injury to the crop.

7.3.1.2. Emulsifiable Concentrates

Some pesticides cannot be directly dissolved in water. However, they may be soluble in a nonpolar organic solvent such as xylene, which can be mixed with water to form an emulsion.

An emulsifiable concentrate consists of a nonpolar solvent in which the pesticide is dissolved, together with an emulsifying agent. When an emulsifiable concentrate (EC) is added to water in a spray tank, a stable emulsion is formed. Tiny solvent droplets are dispersed throughout the water carrier, with each solvent droplet containing some of the dissolved pesticide. Thus, the pesticide is not dissolved in the water, it is simply suspended in the water while remaining dissolved in the nonpolar solvent. The emulsifying agent keeps the solvent droplets dispersed. Hundreds of emulsifiable concentrates are available on the market.

Formulating a pesticide as an emulsifiable concentrate requires that the active ingredient display adequate solubility in nonpolar solvents. A solubility of at least 10 % is required for preparing a cost-effective EC. Most ECs contain at least 25 % active ingredient, and a few contain as much as 0.9 kg/L.

An EC penetrates the waxy foliage of leaves better than other formulations. This means that it may be more effective against hard-to-kill weeds, but also less selective than other formulations. ECs can be applied in hard water without adverse reactions, and they are less likely to be washed off the foliage by rain or sprinkler irrigation. They are less abrasive to pumps and sprayer parts than are wettable powders, and they remain suspended in the tank for longer periods of time than wettable powders. However, the user must remember that the emulsion in the spray tank is still a suspension, and that some agitation is required to preserve its emulsion character.

Emulsifiable concentrates and emulsions are more easily absorbed by the skin than wettable

powders and dusts, and are thus more hazardous if spilled or splashed on the operator.

7.3.1.3. Wettable Powders

Some new pesticides are insoluble in water and insufficiently soluble in organic solvents to permit EC formulation. In that case, the material can be ground very finely and formulated as a wettable powder (WP). A WP is defined as a powder that will readily form a suspension in water.

Wettable powders are formulated by impregnating the active pesticide on an inert material such as clay. A wetting agent and a dispersing agent are also added to the formulation. The wetting agent helps wet the pesticide itself upon addition to water, preventing the material from simply floating on the surface. A dispersing agent aids in spreading the individual particles throughout the water. A typical 50 % WP may contain 42 % clay, 2 % wetting agent, 2 % dispersing agent, and 4 % other impurities (inert ingredients) along with 50 % active pesticide (active ingredient).

Since wettable powders are suspensions and not true solutions, considerable agitation is necessary to prevent settling out. An analogy would be silt carried along by a river. As long as the water remains turbulent, the materials stay suspended. Wettable powders generally have less foliar activity than other types of spray formulations. In fact, some herbicides have been explicitly formulated as WPs even though they could also be formulated as ECs because WPs generally cause less injury to the crop (and sometimes cost less as well). When applied to the soil they generally require activation by rain or sprinkler irrigation. They tend to be abrasive on pumps and nozzle tips of spray equipment. Since they generally contain a rather high percentage of active ingredient, and no solvents or metal containers are needed, the cost per unit of active ingredient tends to be lower than with many other formulations.

Some wettable powders are sold in water-soluble bags, which are then simply introduced into the spray tank. This can be useful when (1) the required application rate is very low, making weighing difficult, or (2) some toxicological question has been raised, since the bag allows the user to avoid all contact with the herbicide.

Wettable powders should be mixed with a small amount of water to form a slurry before mixing with the water in a spray tank. This helps disperse it in the water, and assures that a maximum amount of active pesticide will remain in suspension.

7.3.1.4. Flowables

Flowable formulations are concentrates of liquid or solid pesticides, suspended in a liquid. The most common flowables on the market contain very finely ground pesticides in the form of powders suspended in water. In some respects, such a formulation might be regarded as a prepackaged slurry ready for addition to the spray tank. Since this formulation is a suspension and not a solution, thorough agitation of the container is required before the material is measured out.

7.3.1.5. Dispersible Granules or Dry Flowables

This relatively recent type of formulation is becoming increasingly popular. The two names "dispersible granules" and "dry flowable" refer to the same thing. Such a formulation consists of small beads that disperse into a suspension when added to water. The spray mixture must be agitated, just as in the case of wettable powders. The dispersible granules create no dust problem and they can be measured by volume as well as by weight. A graduated cylinder is included with some products for accurate measurement. Some difficulties have been encountered in mixing certain dry flowables with liquid fertilizer. One should try suspending the formulation in water first before mixing it with a liquid fertilizer. This is conveniently accomplished by working with appropriate concentrations in a 1-L jar prior to mixing up a whole tankful.

Some herbicides applied at very low dosages per unit area are formulated as effervescent tablets that can be counted out and tossed into the spray tank. These quickly become dispersed, and essentially function like a wettable powder or dispersible granules.

7.3.1.6. Invert Emulsions

An invert emulsion is one in which water is the discontinuous phase and oil is the continuous phase. This results in a thick mayonnaise-like product. The primary reason for formulating an invert emulsion is to reduce drift from airplane applications or handgun applications along roadsides and railroads. Special emulsifiers are required to create an invert emulsion, and special equipment is needed for its application.

7.3.1.7. Encapsulated Formulations

New methods have been developed to trap small amounts of pesticide in tiny capsules, which are then suspended in a liquid. Such a concentrate can be mixed with water and applied through a regular sprayer. One purpose of encapsulating pesticides is to facilitate controlled release of the material over a period of time in order to prevent excessive loss by evaporation or from extensive leaching. Also, a formulation of this type may have less tendency to adhere to plant residues on the soil surface.

Poly(vinyl chlorides) and other plastics are being studied as slow release agents for pesticides. The pesticide is mixed with the plastic while the latter is still in liquid form. As the plastic hardens, the pesticide is trapped in tiny compartments, to be released slowly as the plastic breaks down. Mixtures containing such slow release materials as starch are under development.

7.3.2. Formulations Applied Dry

Granules. Granular formulations contain 2 – 20 % concentrations of pesticide in particulate solid carriers suitable for direct field application. They are prepared by impregnating the pesticide on inert particles such as clays, or on such botanical materials as ground corn cob or walnut shells. The particles are screened so that the majority usually fall between 15 and 40 mesh (6 to 15 wires per linear cm).

Granules may be applied with less expensive equipment than sprays, and their application is more convenient. They are able to penetrate a plant canopy, and reach the soil easily. They may also release pesticides gradually over a period of time, which can be advantageous in certain cases. Furthermore, there is less drift hazard with granules, and it is not necessary that the pesticide be soluble in either water or nonpolar solvents for formulation as granules. The cost of granules per unit of active ingredient tends to be higher than with other formulations. There is also a greater chance of poor distribution on the soil than with sprays, since granules may roll or blow from the tops of ridges into furrows, for example. Some pesticides that require greater amounts of water for activation in the soil may be less effective when applied in granular form.

Pellets. Pellet formulations contain larger particles than granules, and are basically used for spot treatment. Many pellets are formulated in water-soluble materials such as borates rather than the insoluble inert materials used with granules. They have most of the same advantages and disadvantages as granular formulations.

Dusts. Dusts usually contain only two ingredients, an inert diluent, usually a clay, and the toxicant. The latter rarely accounts for more than 10 % of the weight of the mixture, and is more often about 1 %. For this reason dusts are more costly per pound of toxicant (because of shipping costs). Herbicides are seldom, if ever, sold as dusts because dusts drift readily.

7.3.3. Active Ingredient vs. Acid Equivalent

Two technical terms need to be discussed at this point. *Active ingredients* refers to those chemicals in a product that are responsible for the pesticidal effects. In a dry formulation, the active-ingredient content is expressed as a percentage by weight. A wettable powder may contain 80 % active ingredient, and a granular product may contain 4 % active ingredient, for example. The concentration of a pesticide whose active form is an acid is usually expressed as an *acid equivalent*. This refers to the theoretical yield of parent acid from the true active ingredient present. Thus, a particular herbicide such as 2,4-D may be active only in the acid form, but for various reasons the acid may have been changed into a salt or an ester. The ester or salt is regarded

as the active ingredient, but dosage recommendations are based on the number of molecules of acid per hectare, whether in the acid, salt, or ester form (Fig. 7). In other words, it woud be expressed as "1 kg of 2,4-D acid per hectare" rather than "1 kg of 2,4-D ester per hectare." A molecule of 2,4-D ester might of course be considerably heavier than one molecule of 2,4-D acid, so a dosage of 1 kg of active ingredient per hectare would release considerably fewer molecules than a dosage of 1 kg of acid equivalent. To avoid confusion, everything is converted to a weight in terms of acid rather than considering a total true weight of active ingredient.

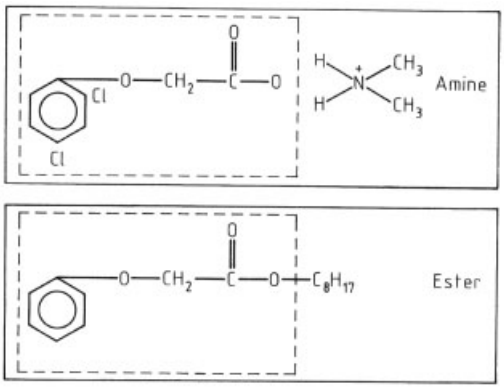

Figure 7. Everything within the solid line is active ingredient. Everything within the dotted line is acid equivalent. The acid equivalent may be identical, but the active ingredient of the ester is higher (heavier total molecule).

7.4. Factors Influencing Foliage-Applied Herbicides

For a foliage-applied herbicide to be effective, the following chain of events must be successfully accomplished: the herbicide must (1) reach the plant, (2) be retained on the leaf, (3) penetrate the leaf, (4) move to its site of action, and (5) remain toxic long enough to exert its action.

Reaching the plant seems an obvious requirement for effective herbicide use. However, this is often overlooked in practical field situations. There are at least three ways in which proper amounts of herbicides fail to reach the leaves.

One of these involves spray drift, which is the lateral movement of spray particles, including the carrier. In general, large nozzles and low pressure will reduce spray drift, because larger spray droplets are formed. Also, drift retardants may be added to a spray solution.

Volatilization of a material prior to reaching the plant is another possibility, particularly with aerial applications. This refers to a change in the herbicide from the solid or liquid state into the gaseous form. Several herbicides are sufficiently volatile that significant percentages could be lost in this way. At this point it is important to stress the difference between spray drift and volatility. Volatility refers to a change to a gas, whereas spray drift implies the movement of spray droplets. Some people attempt to disclaim any responsibility for damage caused by spray drift on the grounds that they have used a nonvolatile formulation, but this is not a compelling argument, because even nonvolatile materials can drift.

A third way in which a herbicide might be prevented from reaching the target leaf would be shading from taller plants. This is generally referred to as the "canopy" effect. Complaints are commonly lodged that a particular contact herbicide has failed to work properly because there are many surviving plants. Often this is a result of the fact that the herbicide contacted only the top overlying layer of leaf surfaces, and failed to reach the smaller plants underneath.

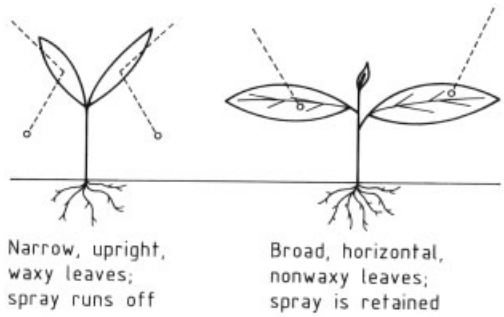

Figure 8. Leaf arrangement and type can influence retention

Once a herbicide reaches the leaf, it must be retained on the surface long enough to be absorbed. Several factors can be involved in retention. The morphology of the plant has a considerable effect. Whether the leaves are upright or horizontal might determine whether or not spray droplets would remain on the leaf or run off (Fig. 8). Also, retention may depend on whether the leaf is waxy or nonwaxy. Extreme exam-

ples are wild garlic and mustard. Wild garlic has waxy, upright leaves, while mustard has broad, horizontal, nonwaxy leaves.

The characteristics of the spray solution are important as well. These can be altered by the inclusion of additives or adjuvants, such as adhesives or "stickers." Wetting agents often act as sticking agents when applied in small volumes of water. For example, the use of wetting agents is always recommended with paraquat on grasses. Otherwise, retention on the leaves is poor.

The volatility of the herbicide also may be important. Some materials evaporate too rapidly to provide adequate retention time.

Herbicides can enter the leaves through either the lower or upper leaf surfaces. Usually the lower surfaces are more permeable than the upper sides. The herbicide can penetrate the leaf either through the stomates or directly through the cuticle. The relative importance of each of these routes is open to debate, but it almost certainly depends upon the particular species involved and the environmental conditions (light, humidity, and whether the stomates are open or closed). Other relevant factors include the nature of the herbicide, its formulation, etc. Most authorities agree that although stomatal uptake can be important, cuticular uptake is of most vital concern. The discussion that follows will focus primarily on cuticle uptake, but it must be remembered that uptake can also occur through the stomates.

Assuming that a herbicide is retained on the leaf, at least four things might happen to it.

1) It can remain on the outside, either as a crystal or a liquid. This can occur with many salts when the aqueous carrier evaporates rapidly.
2) It can enter the cuticle and remain dissolved in the nonpolar portion. This can happen with various weed oils.
3) It can enter and move in the aqueous phase along the cell walls to the vascular system. An example showing this behavior is amitrole.
4) It could enter and move directly into living cells, and through them to the vascular system. The best example of this mode of transport would be 2,4-D.

Nature of the Leaf Cuticle. It is important to understand something about the composition and structure of a leaf cuticle. The leaf cuticle is a complex layer that covers the outside of a leaf. It constitutes a barrier to water loss from the leaves in dry weather, but it also presents a barrier to the penetration of applied herbicides. The constituents of a typical plant cuticle are as follows:

Cutin waxes – nonpolar, lipophilic. Short-chain esters and alcohols of relatively low molecular mass and lacking reactive end groups; some may be cyclic in nature.

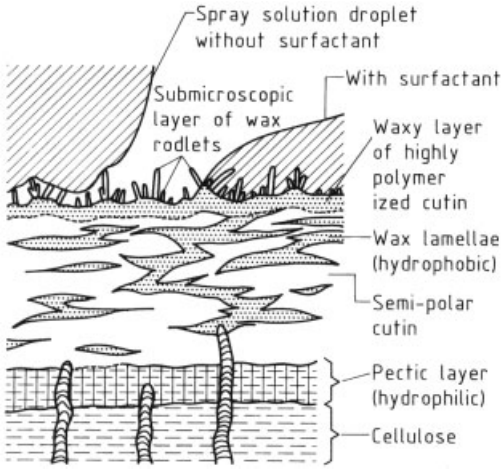

Figure 9. Hypothetical structure of a cuticular layer

Cutin – partly hydrophilic and partly lipophilic. Polymerized acids and alcohols of high molecular mass, and also characterized by reactive end groups.

Pectins – amorphous and highly hydrophilic. Made up of polyuronides.

Cellulose – hydrophilic, fibrillar in nature. Permeable to water and polar in nature.

Aqueous and Lipoidal Routes of Uptake. Polar water-soluble herbicides follow a different route through the cuticle than nonpolar herbicides. The cutin waxes are quite lipophilic in nature, and are difficult for water-soluble herbicides to penetrate. Penetration thus depends upon cracks, insect punctures, etc. A water-soluble herbicide then traverses the hydrophilic part of the cutin, avoiding the suspended wax platelets shown in Figure 9. When it reaches the pectin and cellulose, its progress is much easier, because these materials are hydrophilic and generally quite aqueous in nature. This pathways has

been termed the aqueous route. Nonpolar herbicides follow a lipoidal route. Because these materials are lipophilic in nature, they can easily penetrate the waxy barrier and move through the lipophilic portions of the waxes and cutin. If a herbicide is too lipophilic in nature, however, it may have a difficult time penetrating the hydrophilic pectin and cellulose, and will partition instead into the waxes, thus inhibiting uptake. Esters may hydrolyze to the more water-soluble acids and thereby move into the hydrophilic areas.

Environmental conditions can drastically affect the uptake of herbicides, particularly those that are water soluble. When relative humidity is high, spray droplets evaporate less rapidly, allowing more time for the dissolved herbicides to penetrate. Also, the hydrophilic portions of the cutin and pectin strands can absorb moisture and become swollen. This forces the wax platelets farther apart and provides a much easier route for the water-soluble herbicides to penetrate. Relative humidity has less effect on nonpolar, lipophilic-type herbicides.

In summary, important points regarding cuticular absorption are as follows:

1) The existence of both aqeuous and lipoidal access routes
2) Cuticles are thicker when developed under dry conditions, thinner under wet conditions, as illustrated by the extremes of cactus and aquatic plants; this can influence uptake of foliar-applied herbicides
3) Uptake can be affected by chemical properties of herbicides, e.g., polar vs. nonpolar, water solubility, etc.
4) Formulation is important; emulsifiable concentrates are more readily absorbed than wettable powders, etc.
5) Adjuvants, such as wetting agents and penetrants, can increase uptake, particularly of polar materials
6) High humidity conditions can increase uptake, especially of polar materials by
   a) Reducing evaporation of carrier from the leaf surface
   b) Moisture absorption by hydrophilic substances in the cuticular layer (hydrophilic parts of cutin and strands of pectin), causing it to swell and making the aqueous route more favorable, a swelling which also forces the wax platelets farther apart

Translocation in plants involves two different systems – the phloem and the xylem (Fig. 10). Some foliage-applied herbicides are considered contact-type materials; i.e., they kill only that tissue with which they come in contact. However, many important herbicides can move from the point of application on the foliage to other parts of the plant. Systemic herbicides of this type include such compounds as 2,4-D, amitrole, glyphosate, and many others.

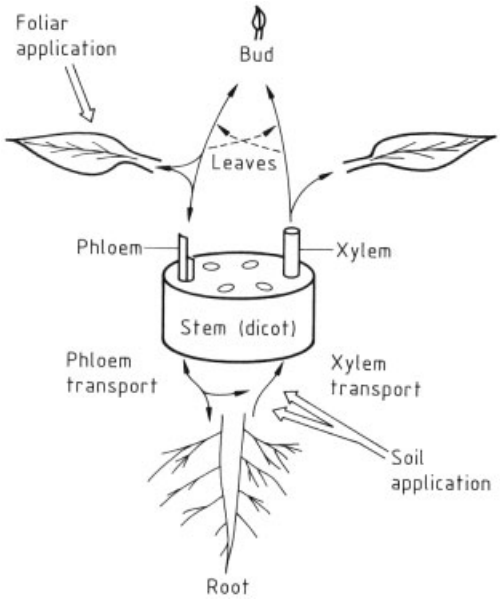

Figure 10. Herbicide transport in the phloem and xylem systems

An important concept with respect to translocation centers around the symplast and the apoplast. The symplast refers to the total interconnected protoplasm of all plant cells, functioning as a unit. This includes phloem, living cells, and the plasmodesmata (strands of protoplasm connecting living cells). The apoplast consists of nonliving tissues together with water surrounding the symplast. This includes the xylem and the secondary walls. Herbicides can move short distances by simple diffusion, but for true systemic action they must move in either the xylem or the phloem.

The movement of herbicides in the symplast has become much better understood in recent years. Movement of solutes in the phloem seems

to occur by a "mass flow" mechanism. Sugars are actively loaded into phloem cells in the leaves through the expenditure of energy from ATP. This increases the osmotic pressure, causing water to pass into those cells, leading to what is called the "source." At the "sinks" sugars are unloaded for metabolism or storage purposes, causing the osmotic potential to fall. Therefore, water simply moves from the source to the sink hydrostatically, like water through a hose. Herbicides seem to diffuse into the phloem in the leaves, permitting them to be carried along with the water to the sink.

The mechanism of herbicide movement has been studied extensively. Some herbicides, such as atrazine, seem not to move in the phloem. Yet, it is known that atrazine is fully capable of entering living cells, so why does it not move in the phloem? Recent information indicates that atrazine can move through cell membranes almost without resistance. Likewise, it can diffuse back out just as easily. Because the xylem stream is more voluminous and swift than the phloem stream, atrazine (and other such herbicides) is simply swept upward in the xylem at the expense of moving downward in the slower phloem stream.

A high percentage of the herbicide load moving in the phloem seems to be acidic. The pH outside the phloem is lower (i.e., 5 – 5.5) than inside the phloem (7.5 – 8). More of the acidic components are in the acid form at pH 5, causing them to be more lipophilic and thus able to diffuse through the cell membrane into the phloem. Once inside the phloem they tend to dissociate into the anionic form at the higher pH that prevails there, thus becoming less lipophilic (Fig. 11). In other words they become "trapped" inside the phloem, and much less able to escape than atrazine and similar substances, which means they are carried along with the water and sugar in the phloem. This mechanism does not explain the phloem movement of certain herbicides, however, such as glyphosate, that do not change form between pH 5 and pH 8. Glyphosate is able to penetrate the membrane into the phloem only with great difficulty, and therefore it also diffuses out only with difficulty, so it simply moves along with the phloem stream and does not escape into the xylem. This represents an illustration of the "intermediate permeability" hypothesis.

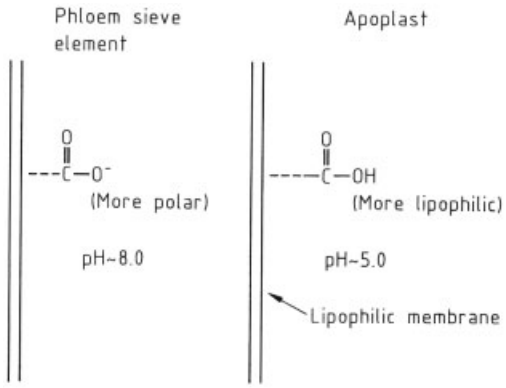

Figure 11. "Trapping" of weak acids in the phloem
The acid form can diffuse through the plasmalemma. Inside the sieve element, the high pH causes more of the acid to ionize. The more polar ion is much less lipophilic than the acid, and cannot diffuse back through the lipophilic plasmalemma as readily, thus trapping it inside where water movement carries it to the sink

There can be varying degrees of partitioning between the symplast and the apoplast. Some herbicides seem to move entirely in one or the other, but others are able to move in both types of tissue. Several compounds – including dicamba and picloram – can move from the foliage in the symplast, along with sugars, or else upward in the xylem.

In annual weeds, herbicide translocation over long distances is generally not of major importance compared to other factors influencing effectiveness. With many species of biennials it is critical that treatment occur at the seedling stage or in the rosette stage, but before the plant sends up a flower stalk, whereas in perennials, the source-to-sink concept can be of major importance in the timing of spray application. In a plant newly emerged from vegetative parts, most of the sugar produced in the leaves is necessarily used to form new vegetative growth, and more is supplied from root reserves; i.e., sugar movement is mostly upward, and spraying should be delayed (Fig. 12). When the plant reaches the bud or early bloom stage, vegetative growth slows down or ceases entirely, and most of the sugar produced in the leaves begins to move toward underground storage tissue. Application of herbicides at this point is often most effective for long-term control, because source-to-sink movement results in more complete translocation throughout the plant.

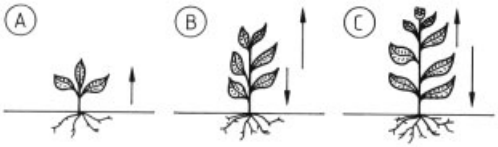

Figure 12. Carbohydrate movement in perennial weeds A) Newly emerged; B) Vegetative growth; C) Bud or bloom-stage

The other season during which downward sugar movement occurs rapidly is in the fall, when perennial plants are storing root reserves for the winter.

Most translocation of herbicides from the leaves occurs in the phloem. Since such translocation depends upon living tissue, rapid burning of the leaves and stems can be detrimental to good translocation, and would be a case of "burning your bridges in front of you." therefore, application to perennial weeds of the most toxic formulation available of a translocated material should be avoided, because it would kill the tops of the plants too quickly for good distribution into the roots.

7.5. Factors Influencing Soil-Applied Herbicides

Many herbicides now available on the market are intended for application to the soil for uptake by emerging shoots or the roots of weeds. Unlike leaves, roots are highly adapted as absorbing organs, and they are very effective at taking up herbicides. Nevertheless, soil-applied herbicides are not always available for uptake into the plant. The issue is one of availability as the major factor influencing the effectiveness of soil-applied herbicides.

7.5.1. Microbiological Effects

Before examining the effects of microbes on herbicides it is appropriate to mention briefly the effect of herbicides on microbes in the soil. This is an issue that has not been completely resolved. Most studies show that normal rates of herbicide application sometimes cause a temporary shift in the microbe population, but seldom an overall reduction. Any such reductions are usually very temporary. Changing the kinds of microbes available in the soil could have a considerable effect, and it is worthy of more thorough investigation, but so far very little detrimental influence of this type has been noted.

Microbes are able to use herbicides for food. They also have the ability to break down a herbicide rather incidentally, perhaps while working on another food source. At any rate, the breakdown of herbicides in soils by microbes is a very important phenomenon. Without microbes, most herbicides would remain in the soil indefinitely, and perhaps cause serious problems by persisting from one year to the next.

Some herbicides are decomposed very rapidly (i.e., within three to four weeks). This includes such compounds as 2,4-D and EPTC. Other compounds are more resistant to microbial breakdown, and can persist even more than one year, including such herbicides as diuron and atrazine.

In general, any condition favorable to the growth of microbes will hasten herbicide decomposition. Temperature, moisture, aeration, organic-matter content, pH, minerals present, concentration of the herbicide, species of microbes present, the particular herbicide involved, and preconditioning (soil enrichment) can all be important factors. A warm, moist, fertile soil is generally ideal for herbicide breakdown, but these conditions are not always met. In many arid regions herbicides are applied to the surface of furrow-irrigated fields. Unless rainfall occurs, the herbicide remains on the dry surface of the soil. Microbes obviously are not active under such circumstances, and the herbicide not only fails to provide good weed control but may persist until the next season and cause crop damage. Even if the herbicide is incorporated into the soil, the upward movement of furrow-applied water may bring it to the surface (Fig. 13).

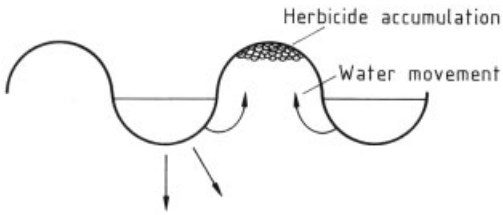

Figure 13. Upward movement of water often forces herbicides to the soil surface in furrow-irrigated fields

Adaptive Breakdown. Particular strains of microbes that can use a herbicide for food may gain an advantage over other microbes in the soil, and therefore increase rapidly in numbers. As the numbers of this particular type of microbe increase, the herbicide is broken down more and more rapidly. This increase in microbe population is sometimes termed "soil enrichment," although obviously has nothing to do with soil fertility. When a herbicide is first applied to a soil, initial breakdown usually is slow, but as herbicide-consuming microbes proliferate, breakdown occurs much more rapidly. The initial period is called the "lag phase," as illustrated in Figure 14. If the same herbicide is reapplied to the same soil after the first application has disappeared, breakdown will occur much more rapidly, and the lag phase may be eliminated completely. This is of course because the microbe population has already been built up, and it is ready to attack the herbicide at once. Some few thiocarbamate herbicides such as EPTC and vernolate have become much less effective in certain "problem" soils because of the soil enrichment phenomenon.

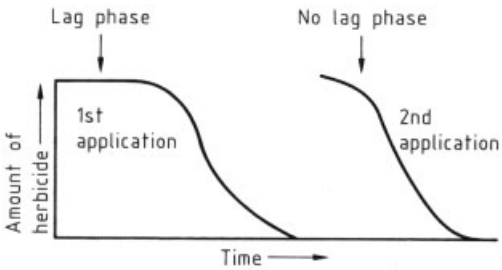

Figure 14. Adaptive breakdown of soil-applied herbicides by microbes

Incidental Breakdown. Many herbicides are subject to degradation by microbes despite the fact that the microbes appear not to be favored in any way. These herbicides simply seem to "get in the way," and they are degraded incidentally as the microbes ingest other materials. The microbe population does not increase, and there is no lag phase, and thus no soil enrichment. Herbicides that undergo this type of microbial breakdown tend to be more persistent in the soil than those broken down by the adaptive mechanism. These include the ureas, triazines, uracils, sulfonylureas, and many miscellaneous herbicides.

7.5.2. Adsorption on Soil Colloids

Adsorption (not *absorption*) is a very important process with respect to soil-applied herbicides. Adsorption refers to the chemical and/or physical attraction of a substance to a surface. When a herbicide is applied to a soil, a considerable fraction of it can become attached so firmly to the soil colloids that the molecules are unavailable for uptake by plants. This is an equilibrium process, and as herbicides are removed from the surrounding soil solution, more herbicide will be desorbed and dispersed in the solution (Fig. 15).

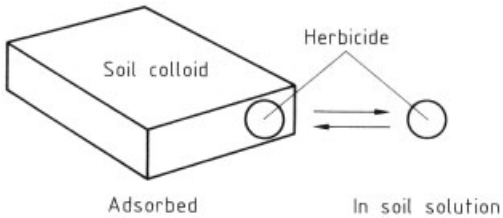

Figure 15. Equilibrium process of herbicide adsorption on soil

The extent of herbicide adsorption depends heavily on the nature of the particular herbicide in question. Some herbicides are quickly and firmly adsorbed (paraquat), while others are adsorbed to a much lower extent (sodium chlorate).

There are four primary soil factors that can affect the extent of adsorption:

1) Texture (sand vs. clay): Texture influences the extent of adsorption by affecting the surface area available. Clay particles are much smaller than sand particles, so the available surface area per unit of volume is much greater in a clay soil than in a sandy soil. Therefore, herbicides are adsorbed to a greater extent in clays than in sands, and higher application rates are necessary in clay soils to ensure adequate activity.
2) Type of clay: Clays vary greatly with respect to exchange capacity. Montmorillonite clays have greater exchange capacity, and can adsorb larger amounts of herbicide, than illite or kaolinite.

3) Organic-matter content: This is perhaps the most important soil factor affecting herbicide adsorption, probably more important than texture in determining the amount of adsorption. Some soils are very high in organic matter, such as muck soils or areas around sawmills, and these have such a tremendous adsorptive capacity that soil-applied herbicides are seldom effective. At the other extreme, some desert soils have an organic matter content $< 1\,\%$, and great care must be exercised to prevent crop damage, because applied herbicides are extremely effective.
4) Soil moisture: It is now known that water can compete with herbicides for adsorptive sites on soil colloids (Fig. 16). In the case of volatile materials applied to the soil surface, loss is much greater from a moist soil than from a dry soil. Volatile herbicide can be adsorbed more quickly and more completely on dry soil, resulting in little loss into the atmosphere, whereas on wet soil the water inhibits adsorption, and herbicide loss is extensive. On the other hand, some herbicides tend to be adsorbed too much for optimum activity, and if they are not volatile it may be desirable that adsorption be minimized. Application of many nonvolatile herbicides is more effective if delayed until after the soil has been moistened by rainfall. The principles are the same in both cases; the goal in one case is simply adequate adsorption with slow release, but less adsorption in the other case.

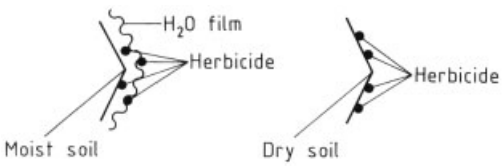

Figure 16. Water can compete with herbicides for adsorption sites

7.5.3. Chemical Decomposition

Several chemical reactions common in the soil are capable of breaking a herbicide down. These can occur even in sterile soils where microbes are not a factor. Relevant chemical reactions include oxidation, reduction, hydrolysis, and hydration. The first inactivation step with chloro-*s*-triazines has been shown to be a nonmicrobial hydrolysis, with loss of the chlorine atom. Sulfonylureas are degraded both by hydrolysis and by microbial action, and it proceeds much more slowly in high-pH soils (> ca. 7.8).

7.5.4. Leaching

Leaching refers to the downward movement through the soil of a substance in solution. The extent of herbicide leaching through a soil depends on its solubility, the volume of water passing through the soil, the distribution of the water, and the adsorptive relationships between the herbicide and the soil. In general, some movement of the chemical is always required in order to ensure that it will pass into the soil where it can be absorbed by weed roots. However, it is usually not desirable that the herbicide move too deeply, because it would then be ineffective on the weeds, or injurious to deep-rooted crops.

7.5.5. Volatilization

Some herbicides are subject to evaporation and loss in the form of a gas or vapor, including such compounds as EPTC, metham, cycloate, trifluralin, dichlobenil, and others. Such materials must usually be incorporated mechanically into the soil soon after application. Some herbicides that are less volatile can still evaporate over a period of time if the environmental conditions are hot and dry. Examples include 2,4-D high-volatile esters and even atrazine.

7.5.6. Photodecomposition

Herbicide molecules lying on the soil surface may absorb light and then be incapable of releasing the resulting energy. In this case, the energy causes chemical bonds in the molecule to rupture, and the herbicide is destroyed. Most herbicides are not highly susceptible to photodecomposition, but it can occur over a period of time. Urea herbicides such as diuron are known to be broken down by light when they remain on the soil surface for several weeks. Trifluralin, paraquat, and napropamide are other herbicides

known to photodecompose. Napropamide especially must be incorporated quickly for satisfactory activity. Movement into the soil induced by rainfall or irrigation helps prevent this photodecomposition.

7.5.7. Uptake and Translocation from the Soil

Soil-applied herbicides are naturally expected to be taken up by the roots of weeds. However, several important herbicide groups (including the thiocarbamates, dinitrobenzene amines, acetanilides, and DCPA) are effective on grasses as a result of uptake into the emerging shoot (see also Section 7.6). Some, such as the thiocarbamates, show almost no toxicity when only the roots are exposed. Others, such as the dinitrobenzenamines, are able to cause severe damage to roots, but because weeds can survive drastic reductions in root growth, the primary lethal action occurs through uptake into the emerging shoot. Shoot-uptake herbicides seem not to be translocated significantly.

Herbicides that are effective through the roots – ureas, triazines, uracils – move to the foliage in the xylem. Herbicide movement in the xylem seems to be completely passive; i.e., the herbicides simply go where the water goes, when the water goes, and to the extent that the water goes. It is analogous to a wood chip in a river; there is no particular control, and movement is dictated completely by the streaming water. This means that conditions of high transpiration (i.e., high soil moisture and low relative humidity in the air) favor rapid herbicide uptake from the soil. Herbicides also can move in the apoplast subsequent to foliar treatment. This results in concentration at the tips and edges of the leaves, and of course does not result in movement to the roots.

The soil is a complex medium. Many of the factors discussed above can interact one with another (Fig. 17). Any one factor can be sufficiently important to lead to the success or failure of a herbicide application. Of all environmental factors, soil moisture is likely to be the key factor in a successful treatment. Poor moisture conditions probably cause more failures with soil-applied herbicides than any other single problem. Growers with access to sprinkler irrigation have a considerable advantage when it comes to the use of soil-applied herbicides because they are able to assure that the compound will be leached from the soil surface into the weed uptake zone for optimum control. A herbicide lying on the soil surface, unless it is a volatile compound, simply cannot do a satisfactory job of controlling germinating weeds. Even if a herbicide has been mechanically incorporated into the soil, in the absence of sufficient moisture it may not be available for uptake. Weed seedlings emerging from deeper, moist soil can emerge through the dry soil layer unharmed. Proper placement in the soil and adequate moisture are critical concerns.

7.6. Selective Action of Herbicides

Selectivity is the result of a chemical reaching and disrupting a vital function in one plant but not in another. The selectivity of pesticides is not particularly unique. For example, mouthwash is employed in an attempt to kill certain organisms (bacteria) without killing another organism (the user). But with herbicides, the differences between target species and desirable species are often quite limited. Nevertheless, there are numerous reasons why weeds can sometimes be controlled selectively in a desired crop.

Contacting the Plant. Selectivity is often achieved mechanically through the use of directed sprays, or shields that restrict the spray to the weeds while protecting the crop. A common mechanism of selectivity involves a differential rooting depth between the weed and crops; selectivity in this case depends on the use of herbicides of sufficiently low water-solubility and/or strong soil-adsorption characteristics such that they do not penetrate deeply, and are thus able to control shallow-rooted weeds without harming deeper-rooted crops, particularly perennials. The use of simazine in orchards is one example.

Selectivity might also be achieved by timing the spray application. The application of paraquat to control emerged weeds immediately prior to emergence of the crop plant can be considered an example of such selectivity.

Penetration into the Plant. If the crop plant has waxy leaves and an upright growth form, her-

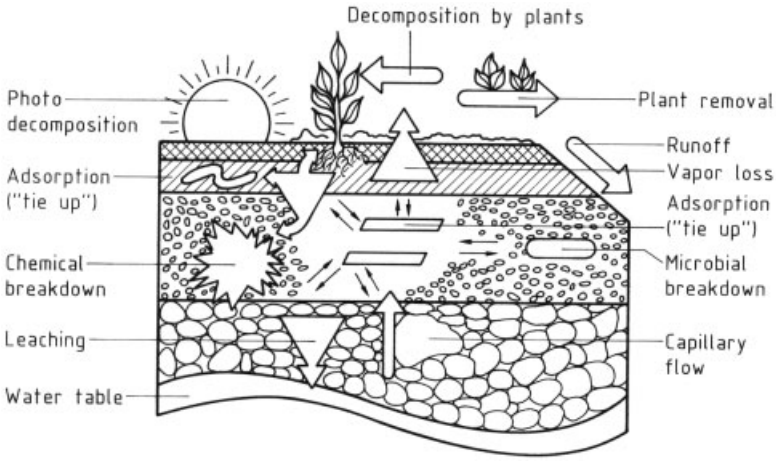

Figure 17. Interaction of many soil factors that determine the fate of herbicides; from [22]

bicides that are prone to run off might be used to control weeds lacking those characteristics. For example, dinoseb can be used to control mustards and other easily-wet weeds selectively in peas because the spray does not penetrate into the pea leaves. Onions can withstand some herbicides such as bromoxynil because of their upright, waxy leaves.

Differences in Grass-Seedling Morphology. The morphology of various grass seedlings differs as they emerge. With some species (Type I), the mesocotyl elongates, pushing the base of the coleoptile upward to a point near the soil surface. Most grasses tend to germinate in this way. On the other hand, in certain grasses (Type III) the mesocotyl does not elongate, and the base of the coleoptile remains at the seed. The coleoptile extends from the seed to the soil surface. Only later does an internode elongate, bringing the inflorescence primordia and leaf primordia near the soil surface, but this time *within* the coleoptile. Some herbicides are highly active when the region around the base of the coleoptile is exposed, because this region is rich in metabolic activity. It is here that the inflorescence primordia, leaf primordia, and future nodes are located. In Type III grasses this region approaches the soil surface very late, and it is somewhat protected by being within the coleoptile sheath. Type II grasses (oats, maize) are intermediate between Type I and Type III. Type I grasses include a great many species, whereas Type III consists largely of wheat, barley, rye, goatgrass, and a very few others. Consequently, selectivity might be achieved as a result of a difference in seedling morphology. For example, wheat seeded directly in trifluralin-treated soil will almost certainly be killed. However, trifluralin is registered (see Section 7.11) for the selective control of Type I grasses in wheat. This selectivity can be achieved only by planting the wheat seed more deeply than the trifluralin-treated soil. The base of the coleoptile of the Type I grasses (*Setaria, Echinochloa, Digitaria*) passes into the trifluralin-treated soil, whereas the base of the coleoptile in wheat remains at the seed and escapes injury.

Rapid Metabolism. Much more rapid degradation of a herbicide by the crop plant than by a weed is a common mechanism for biochemical selectivity. There are numerous examples of this type of selectivity, including atrazine in maize, terbacil in mint, diclofop in wheat, etc. In some cases the herbicide may not be rapidly destroyed, but rather conjugated with some cell component. Biochemical selectivity is generally more favorable, because it is less affected by environmental conditions than are mechanisms such as differential rooting depth.

Lack of Activation. A special type of biochemical selectivity is illustrated by the phen-

oxybutyric acids 2,4-DB and MCPB. Those two compounds are inactive as applied, and must be broken down within the plant through β-oxidation to 2,4-D and MCPA, respectively. Fortunately, most legume seedlings are unable to accomplish that conversion, and are spared herbicidal damage. Most broadleaf weeds can carry out β-oxidation readily, in effect committing chemical suicide.

Differences in Enzyme Form. In investigating the mechanism of resistance in populations that had become highly resistant to the *s*-triazine herbicides, it was discovered that the attachment protein within the chloroplast, where these materials exert their effect, had undergone a change. The herbicides could no longer attach to their site of action, and were therefore rendered ineffective. The fatty acid synthesis inhibitors (fluazifop, diclofop, sethoxydim, quizalofop) are highly selective in broadleaf crops because the corresponding acetyl CoA carboxylase enzyme displays differences such that the herbicides cannot attach to it, and is thus rendered ineffective. Unfortunately, normally sensitive species may include in their populations an appreciable number of individual plants with an altered form of enzyme, with the result that after only a few applications of an ACC-ase inhibitor, a herbicide-resistant population may be selected out.

7.7. Examples of Important Herbicides and Herbicide Groups

Thousands of formulations of 150 – 200 herbicides are used commercially around the world. Coverage here is limited to a brief introduction to selected groups and to properties of selected herbicides given in Section 7.8. Fortunately, excellent herbicide manuals containing more detailed information are available in many countries, including the United States [10], the United Kingdom [6], Germany [4], and France [3].

7.7.1. Foliage-Active Contact Herbicides

This group includes herbicides whose action occurs primarily through the foliage, and whose movement within the plant is very restricted. Damage from these herbicides is limited primarily to those tissues with which they make contact, so good spray coverage is especially important.

The *bipyridiliums* include paraquat, diquat, and a few others that are less widely used. Paraquat and diquat can injure or kill both grasses and broadleaves. They are rapidly absorbed into leaf tissue, and thus are not readily washed off by rainfall. They kill plants by accepting an electron in the photosynthetic process, leading to a proliferation of strong oxidizing agents that rapidly destroy plant membranes and other cell components. Consequently, bypyridiliums exert their toxic effects more rapidly in bright sunlight than in subdued light. Because they exist in the form of divalent cations, they are rapidly and tightly bound to soil particles, thus providing no soil activity in mineral soils. Bipyridiliums are used for controlling weeds prior to the planting of a crop, as directed sprays in ornamentals and orchards, in aquatics, as desiccants, as dormant applications to perennial crops, and in other applications. They are among the herbicides most toxic to mammals.

The *benzonitriles* include bromoxynil and ioxynil, which are primarily foliage-active nontranslocated herbicides. Dichlobenil is also a benzonitrile, but its mode of action is different. Bromoxynil and ioxynil are used primarily to control small broadleaves in grains and other grass crops, but may also be used in crops such as legumes. These materials inhibit respiration and perhaps photosynthesis as well.

Propanil is widely used for controlling weeds, primarily grasses, in rice. It is rapidly degraded in plants and soil, and inhibits photosynthesis.

Bentazon selectively controls certain broadleaf and sedge weeds in most grass crops and many large-seeded legume crops. Its activity in soil is relatively low, and its persistence is generally limited to six weeks or less. It inhibits photosynthesis.

Dinoseb is used in some countries (but not in the United States, where it is unavailable) both in the form of the parent phenol and as salts. The parent phenol is relatively nonselective, and is used as a general desiccant, in directed sprays for perennial crops, and for broad-spectrum knockdown on canal banks, etc. The salts are much more selective, and may be used

to control broadleaf weeds in many crops, including small grains, legumes, and flax. It is general biocide, and can inhibit growth of pathogens and such soil fauna as earthworms. It also is relatively toxic to mammals. Dinoseb inhibits respiration in both plants and animals.

Various phytotoxic oils can also be used directly for weed control. Unlike the nonphytotoxic oils used as adjuvants, phytotoxic oils are more highly unsaturated, and they tend to have a high content of aromatics. They are also sometimes used in conjunction with herbicides to improve uptake into brushy species.

7.7.2. Foliage-Active Translocated Herbicides

The *growth regulators*, or auxin-type herbicides, have been widely used around the world for several decades (→ Chlorophenoxyalkanoic Acids). This group includes 2,4-D, MCPA, 2,4,5-T, picloram, dicamba, clopyralid, triclopyr, and others. Their selectivities differ, but the mechanism of herbicidal action, which is not completely understood, seems to be uniform. Auxin-type herbicides primarily succeed in controlling broadleaf weeds selectively in grass crops. Their discovery in the early 1940s was largely responsible for a tremendous increase in interest in herbicides, and it is widely regarded as a turning-point in weed-control history. These materials are translocated in the phloem in a source-to-sink manner. For this reason, timing and moisture status of the plant are critical in controlling deep-rooted perennial broadleaves. Soil persistence ranges from as little as 2–4 weeks in the case of 2,4-D to perhaps several months for picloram. Their acute mammalian toxicity varies widely, from 300–1000 mg/kg in rats for the phenoxy group, less toxicity with dicamba, to very low toxicities with clopyralid and picloram. The exact mechanism of physiological action remains elusive, but evidence indicates an association with a rapid effect on the proton pump in the cell membrane, and a slower effect on RNA synthesis, perhaps leading to abnormal enzyme synthesis.

The *fatty acid synthesis inhibitors* include diclofop, fluazifop-P, sethoxydim, quizalofop, tralkoxydim, and clethodim. These materials control a wide spectrum of grasses, but display very low toxicity with respect to broadleaf species. Unfortunately, this group tends to develop resistant populations quite rapidly. Most of the members are translocated both in the xylem and the phloem. Compounds in this category block fatty acid synthesis by inhibiting acetyl CoA carboxylase. Broadleaves as well as fine fescues (*Vulpia* sp.) are tolerant, because they have a different form of this enzyme than most grasses.

Glyphosate is widely used throughout the world. It is a nonselective, general weed killer, controlling annual and perennial broadleaves as well as grasses. It is used in both cropland and noncropland situations. Selectivity can be achieved by differential timing or placement of herbicide application in relation to crop growth. Because glyphosate has little or no soil activity, registered crops can be planted immediately after application. It is absorbed through the foliage, and moves throughout the plant in the symplast in a source-to-sink manner. It kills plants by inhibiting an enzyme in the synthetic pathway for aromatic amino acids (→ Plant Growth Regulators). Because this pathway is not present in mammals, glyphosphate's toxicity to mammals is very low.

The *sulfonylureas* (SU) include chlorsulfuron, metsulfuron, sulfometuron, rimsulfuron, chlorimuron, primisulfuron, triasulfuron, and several others. The *imidazolinones* include imazamethabenz, imazapyr, imazaquin, imazethapyr, and others. Characteristics of individual herbicides within these groups vary widely, as do selectivity and uses. Soil persistence ranges from very brief to greater than a year. The sulfonylurea group in particular illustrates the variation in selectivity that can be achieved by changes in molecular structure. Some of the SUs are extremely toxic to sugar beets, for example, whereas other related compounds can be used selectively with sugar beets. Both groups inhibit plant growth by blocking an enzyme in the biochemical pathway for synthesis of the branched-chain amino acids valine, leucine, and isoleucine. This pathway is unique to plants, so these compounds display very low toxicity in mammals.

Other foliage-active, translocated herbicides include amitrole and the organic arsenicals.

7.7.3. Soil-Active Herbicides

Photosynthesis inhibitors include the phenylureas (diuron, chlorbromuron, isoproturon, etc.), the *s*-triazines (atrazine, simazine, cyanazine), uracils (bromacil, terbacil), and pyrazon. These materials are commonly called Hill-reaction inhibitors because they block electron transport in photosystem II. The compounds are absorbed by roots and transported upward in the xylem, after which they move into greenleaf tissue where they exert their toxic effects. Commonly, selectivity is a consequence of differences in rooting depth between the crop and the weed, but there are notable examples of differential metabolic rates, as with the *s*-triazines in corn and terbacil in mint. None are highly volatile nor readily photodecomposed. Soil persistence varies, but frequently can reach one or more years, depending on climatic conditions and initial application rate. Because the compounds operate primarily on a system unique to plants, toxicity to mammals is considered very low. Plant death does not result from starvation alone. Chlorophyll 680 continues to absorb solar energy even when electron transport is blocked, but it is unable to dissipate the energy. This results in highly oxidized cell products, leading to membrane destruction. Therefore, phytotoxicity is a result of both electron-transport blockage and the absorption of light, explaining why the products work more rapidly in bright sunlight than in dim light. The first *s*-triazines introduced, simazine and atrazine, have been used extensively in maize, but use of these two herbicides, as well as others in the group, has expanded to selective applications in perennial crops, orchards, and noncrop situations.

The *dinitroanilines* or *dinitrobenzeneamines* include trifluralin, benefin, oryzalin, pendimethalin, and ethalfluralin, among others. The strength of these materials is in controlling annual grasses, but they can also control certain broadleaf weeds. Most are so volatile that they must be incorporated into the soil to prevent vapor loss. Although these substances are effective at inhibiting root growth, the primary mechanism of plant death requires uptake into the emerging grass shoots. None of the materials are effective when applied postemergence to existing weeds, but they are commonly applied before a crop is planted for incorporation into the soil, or applied following crop planting and incorporated above the crop seed. Uses for members of this group can include weed control in soybeans, cotton, small-seeded legumes, beans, cole crops, umbelliferous crops, potatoes, sugarcane, and many others. The herbicides are orange – yellow in color, and they kill plants by preventing the polymerization of tubulin to form microtubules. Their toxicity to mammals is quite low.

The *diphenyl ether* herbicides include acifluorfen, nitrofen, and oxyfluorfen. They mostly control annual broadleaf weeds, but can provide control over some grass species as well. One or more members of the group can be used for weed control in a wide variety of crops, including soybeans, cole crops, umbelliferous crops, and mint. Toxicity to mammals is considered moderate to low. These materials kill plants by inhibiting protoporphyrinogen oxidase, which probably causes unprocessed protoporphyrinogen to diffuse out of its normal localized site in the chloroplast to become oxidized by molecular oxygen in an unregulated manner. This results in a large accumulation of protoporphyrin IX, leading to excessive absorption of light and subsequent oxidation of membranes. Light is required for the phytotoxic action of this group.

The *thiocarbamate* or *carbamothioate* group includes EPTC, butylate, cycloate, triallate, and vernolate. This group is used primarily for controlling annual and perennial grasses in broadleaf crops and in maize. EPTC was the first representative of this group, and it is used in a wide variety of crops, including small- and large-seeded legumes, maize, potatoes, flax, sunflower, and many others. Cycloate is used primarily for weed control in sugar beets and spinach, whereas triallate is widely used for control of wild oats in small grains and in peas and lentils. These materials are volatile, and must be incorporated into the soil surface fairly soon after application. They are low in toxicity to mammals, and they kill plants by entry into the shoots of emerging grasses. The exact mechanism of herbicidal action is not known, although considerable evidence points toward an involvement with lipid metabolism.

The *substituted amides* include alachlor, metolachlor, butachlor, propachlor, napropamide, and pronamide. Characteristics of the several members of this group vary

widely. The substituted acetanilides (alachlor, butachlor, metolachlor) are most effective on annual grasses as a result of surface application following planting of the crop. Alachlor and metolachlor are used extensively in maize and soybean, whereas butachlor is primarily used in rice. Their activity requires uptake into the shoots of emerging grasses. Napropamide is one of the few herbicides in which photodecomposition is so significant that it must be incorporated into the soil within a day or two during sunny weather. Toxicity to mammals is moderate for a few of the herbicides, but very low with most of the members of this group.

7.8. Properties of Herbicides

7.8.1. Diphenyl Ethers

Acifluorfen [*50594-66-6*], 5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-nitrobenzoic acid, $C_{14}H_7ClF_3NO_5$, M_r 361.7, *mp* 142 – 160 °C, is a light brown solid which is sparingly soluble in water and soluble in acetone, ethanol, dichloromethane, xylene, and kerosene [23, p. 12].

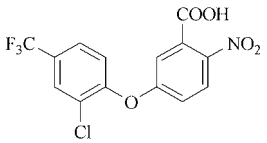

Acifluorfen is produced by reaction of 2-chloro-4-trifluoromethylphenol with 2-nitro-5-chlorobenzoic acid [25].

Acifluorfen is a selective contact herbicide which is absorbed by foliage and roots, with negligible translocation. It is used post-emergence for the control of annual broad-leaved weeds, with some effects on grasses.

Trade Names. Blazer, Doble, Galaxy, Storm (BASF).

Aclonifen [*74070-46-5*], 2-chloro-6-nitro-3-phenoxyaniline, $C_{12}H_9ClN_2O_3$, M_r 264.7, *mp* 81 – 82 °C, consists of yellow crystals which are practically insoluble in water but soluble in methanol, hexane, and toluene [23, p. 14].

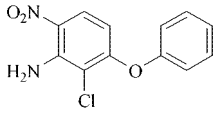

Aclonifen is manufactured by reaction of 1,2-dichloro-3-amino-4-nitrobenzene with phenolate [26].

Aclonifen is a selective, systemic herbicide used pre-emergence to control grass and broad-leaved weeds in winter wheat, potatoes, sunflowers, and other crops.

Trade Name. Challenge (Rhône-Poulenc).

AKH-7088 [*104459-82-7*], methyl (*E,Z*)-1-[5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-nitrophenyl]-2-methoxyethylidene amino-oxyacetate, $C_{19}H_{16}ClF_3N_2O_7$, M_r 476.8, *mp* 57.7 – 58.1, forms colorless crystals which are sparingly soluble in water [23, p. 22].

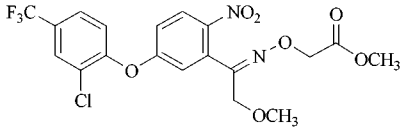

AKH-7088 is produced by reaction of 2-chloro-*p*-trifluoromethylphenolate with 3-methoxyethylidineamino-oxyacatate 4-nitrochlorobenzene [27].

AKH-7088 is used post-emergence to control broad-leaved weeds, and velvet leaf in soya beans.

Trade Name. To be announced (Asahi).

Bifenox [*42576-02-3*], methyl 5-(2,4-dichlorophenoxy)-2-nitrobenzoate, $C_{14}H_9Cl_2NO_5$, M_r 342.1, *mp* 84 – 86 °C, consists of yellow crystals which are practically insoluble in water, readily soluble in acetone, chlorobenzene, xylene, and ethanol, and slightly soluble in aliphatic hydrocarbons [23, p. 116].

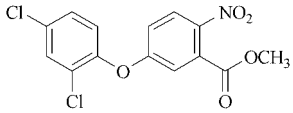

Bifenox is produced by reaction of potassium 2,4-dichlorophenolate with methyl 5-chloro-2-nitrobenzoate [28].

Bifenox acts by cellular membrane disruption and by inhibition of photosynthesis. It is a selective herbicide used to control broad-leaved weeds and some grasses in cereals, maize, and soya beans.
Trade Names. Modown, Tolkan (Rhône-Poulenc).

Chlomethoxyfen [*32861-85-1*], 5-(2,4-dichlorophenoxy)-2-nitroanisole, $C_{13}H_9Cl_2NO_4$, M_r 314.1, *mp* 113 – 114 °C, consists of yellow crystals which are practically insoluble in water, and are readily soluble in acetone, dimethyl sulfoxide, and benzene [23, p. 200].

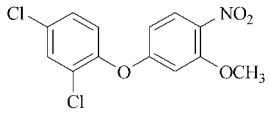

Chlomethoxyfen is obtained by reaction of 2,4-dichlorophenolate with 1-chloro-3-methoxy-4-nitrobenzene [29].

Chlomethoxyfen is a selective contact herbicide used pre-emergence for the control of annual weeds in transplanted rice.
Trade Names. Diphenex (Ishihara Sangyo); Ekkusugoni (Nihon Nohyaku, Ishihara Sangyo).

Fluoroglycofen-ethyl [*77501-90-7*], *O*-[5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-nitrobenzoyl]glycolic acid, $C_{18}H_{13}ClF_3NO_7$, M_r 447.8, *mp* 65 °C, forms a dark amber solid which is practically insoluble in water and is readily soluble in most organic solvents [23, p. 580].

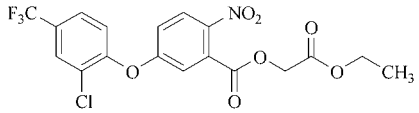

Fluoroglycofen-ethyl is produced by reaction of 5-[2-chloro-4-(trifluoromethyl)phenyl]-2-nitrobenzoyl chloride with ethyl hydroxyacetate in the presence of trimethylamine [30].

Fluoroglycofen-ethyl is a selective leaf and root herbicide used to control broad-leaved weeds and grasses in wheat, barley, oats, peanuts, rice, and soya beans.
Trade Names. Compete (Rohm & Haas); Satis (Novartis).

Fomesafen [*72178-02-0*], 5-(2-dichloro-α,α,α-trifluoro-*p*-tolyloxy)-*N*-methylsulfonyl-2-nitrobenzamide, $C_{15}H_{10}ClF_3N_2O_6S$, M_r 438.8, *mp* 220 – 221 °C, forms colorless crystals which are moderately soluble in water [23, p. 616].

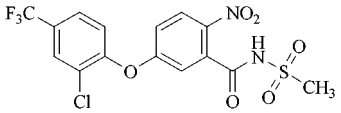

Fomesafen is manufactured by reaction of 2-chloro-4-trifluoromethylphenolate with 4-chloro-*N*-methylsulfonyl-2-nitrobenzamide [31].

Fomesafen is a selective herbicide which is absorbed by both leaves and roots. It is used in early post-emergence for the control of broad-leaved weeds in soya beans.
Trade Names. Flex, Reflex (Zeneca).

HC-252 [*131086-42-5*], ethyl *O*-[2-chloro-5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy) benzoyl]-L-lactate, $C_{19}H_{15}Cl_2F_3O_5$, M_r 451.0, is a brown oily viscous liquid which is readily soluble in acetone, methanol, and xylene [23, p. 663].

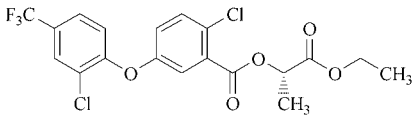

HC-252 is produced by reaction of 2-chloro-5-[2-chloro-4-(trifluoromethyl)phenoxy]benzoyl chloride with ethyl acetate [32].

HC-252 acts by accumulation of tetrapyrrole and protoporphyrin. It is a contact herbicide for the control of broad-leaved weeds in winter wheat, barley, peas, soya beans, and peanuts.
*Trade Name.*To be announced (Budapesti Vegyimüvek).

Lactofen [*77501-63-4*], ethyl *O*-[5-(2-chloro-α,α,α-trifluoro-*p*-tolyloxy)-2-nitrobenzoyl]-DL-lactate, $C_{19}H_{15}ClF_3NO_7$, M_r 461.8, is a dark brown to tan solid which is sparingly soluble in water [23, p. 747].

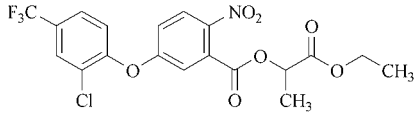

Lactofen is produced by reaction of 2-chloro-4-trifluoromethylphenolate with ethyl 2-nitro-5-chlorobenzoyllactate [33].

Lactofen is applied to control broad-leaved weeds in cereals, potatoes, soya beans, rice, and peanuts.

Trade Name. Cobra (Valent, PPG, Sumitimo).

Oxyfluorfen [*42874-03-3*], 2-chloro-α,α,α-trifluoro-*p*-tolyl 3-ethoxy-4-nitrophenyl ether, $C_{15}H_{11}ClF_3NO_4$, M_r 361.7, *mp* 85 – 90 °C, is an orange crystalline solid which is practically insoluble in water and readily soluble in most organic solvents [23, p. 919].

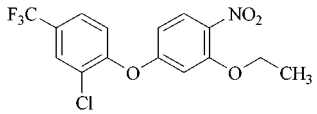

Oxyfluorfen is manufactured by reaction of 3,4-dichlorobenzotrifluoride with resorcinol; the resulting product is then treated with ethanol in the presence of potassium hydroxide [34].

Oxyfluorfen is a selective contact herbicide used pre- and post-emergence for the control of annual broad-leaved weeds in a variety of tropical and subtropical crops.

Trade Names. Goal (Rohm & Haas); Galigan (Markhteshim-Agan), etc.

7.8.2. Sulfonylureas

Amidosulfuron [*120923-37-7*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-mesyl(methyl)sulfamoylurea, $C_9H_{15}N_5O_7S_2$, M_r 369.4, *mp* 160 – 163 °C, is a colorless crystalline powder which is sparingly soluble in water and moderately soluble in isopropyl alcohol, methanol, and acetone [23, p. 37].

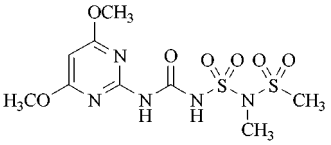

Amidosulfuron is produced by reaction of *N*-methylmesylsulfonamide with chlorosulfonyl isocyanate, followed by treatment with 2-amino-4,6-dimethoxypyrimidine [35].

Amidosulfuron is a selective, systemic herbicide used post-emergence for the control of a wide range of broad-leaved weeds, especially cleavers, in winter wheat, durum wheat, barley, rye, and oats.

Trade Name. (AgrEvo).

Azimsulfuron [*120162-55-2*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-[1-methyl-4-(2-methyl-2*H*-tetrazol-5-yl)pyrazol-5-ylsulfonyl]urea, $C_{13}H_{16}N_{10}O_5S$, M_r 424.4, *mp* 170 °C, is a colorless solid which is moderately soluble in water, acetone, acetonitrile, ethyl acetate, and methanol [23, p. 63].

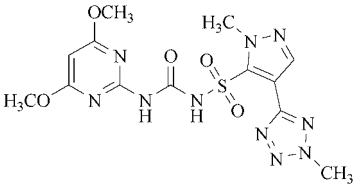

Azimsulfuron is produced by reaction of 2-(1-methyl-1*H*-diazol)-4-(1-methyl-1*H*-tetrazol)-sulfonamide with 2-(4,6-dimethoxypyrimidine)isocyanate [37].

Azimsulfuron is a post-emergence herbicide with mainly foliar uptake. It is used to control annual and perennial broad-leaved and sedge weeds in southern European rice.

Trade Name. Gulliver (Du Pont).

Bensulfuron-methyl [*83055-99-6*], methyl α-(4,6-dimethoxypyrimidin-2-ylcarbamoylsulfamyl)-*o*-toluate, $C_{16}H_{18}N_4O_7S$, M_r 410.4, *mp* 185 – 188 °C, is a colorless to pale yellow solid which is sparingly soluble in water and is fairly soluble in acetone, acetonitrile, chloromethane, and ethyl acetate [23, p. 104].

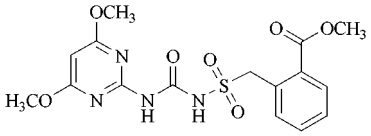

Bensulfuron-methyl is produced by the reaction between 2-carboxymethyl benzylsulfamaoyl chloride and 2-amino-4,6-dimethoxypyrimidine [37].

Bensulfuron-methyl is a selective, systemic herbicide used pre- and post-emergence for the control of annual and perennial weeds and sedges in continuously flooded rice.

Trade Names. Fujigrass (Du Pont, Zeneca), Londax, Zark (Du Pont); Wolf-Ace (Kumiai), etc.

Chlorimuron-ethyl [*90982-32-4*], ethyl 2-(4-chloro-6-methoxypyrimidin-2-ylcarbamoylsulfamoyl)benzoate, $C_{15}H_{15}ClN_4O_6S$, M_r 414.8, *mp* 181 °C, consists of colorless crystals which are moderately soluble in water and slightly soluble in organic solvents [23, p. 217].

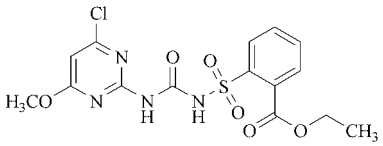

Chlorimuron-ethyl is produced by reaction of ethyl 2-sulfamoylchloride benzoic acid with 2-amino-4-chloro-6-methoxypyrimidin [38].

Chlorimuron-ethyl is used post-emergence for the control of broad-leaved weeds in soya beans and peanuts.

Trade Names. Classic (Du Pont); Darban (IPESA), etc.

Chlorsulfuron [*64902-72-3*], 1-(2-chlorophenylsulfonyl)-3-methyl-1,3,5-triazin-2-yl)urea, $C_{12}H_{12}N_5O_4S$, M_r 357.8, *mp* 170 – 173 °C, forms a colorless crystalline solid which is sparingly soluble in water and readily soluble in dichloromethane, acetone, methanol, toluene, and hexane [23, p. 239].

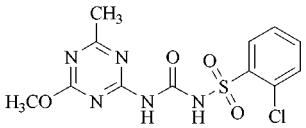

Chlorsulfuron is produced by reaction of 2-sulfonylisocyanate chlorobenzene with 2-amino-4,6-dimethoxy-1,3,5-triazine [39].

Chlorsulfuron is a selective, systemic herbicide used pre- and early post-emergence for the control of most broad-leaved weeds and some grasses in wheat, barley, oats, rye, and noncropland.

Trade Names. Glean, Telar (Du Pont); Lasher (Sanonda), etc.

Cinosulfuron [*94593-91-6*], 1-(4,6-dimethoxy-1,3,5-triazin-2-yl)-3-[2-(2-methoxyethoxy)phenylsulfonyl]urea, $C_{15}H_{19}N_5O_7S$, M_r 413.4, *mp* 127.0 – 135.2 °C, is a colorless crystalline powder which is fairly soluble in water and is readily soluble in acetone, ethanol, toluene, and *n*-octanol [23, p. 148].

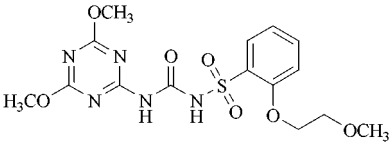

Cinosulfuron is produced by reaction of 2-methoxyethoxy sulfamoyl chlorobenzene with 2-amino-4,6-dimethoxy-1,3,5-triazine [40].

Cinosulfuron is used post-emergence to control many weeds in transplanted, directly seeded, wet-sown, water-sown, and dry-sown flooded rice crops. It is also used in tropical plantation crops.

Trade Name. Setoff (Novartis).

Ethametsulfuron-methyl [*97780-06-8*], methyl 2-[(4-ethoxy-6-methylamino-1,3,5-triazin-2-yl)carbamoylsulfamoyl]benzoate, $C_{15}H_{18}N_6O_6S$, M_r 410.4, *mp* 194 °C, consists of colorless to light tan crystals which are sparingly soluble in water and are moderately soluble in acetone, acetonitrile, ethanol, methanol, and dichloromethane [23, p. 475].

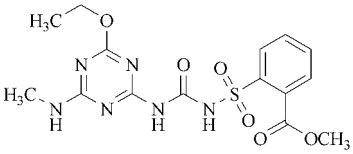

Ethametsulfuron-methyl is produced by reaction of methyl 2-sulfamoylbenzoate with 4-

ethoxy-6-methylamino-1,3,5-triazin-2-yl isocyanate [41].

Ethametsulfuron-methyl is used post-emergence to control wild mustard, hempnettle, and other broad-leaved weeds in oilseed rape.
Trade Name. Muster (Du Pont).

Ethoxysulfuron [*126801-58-9*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-(2-ethoxyphenoxysulfonyl)urea, $C_{15}H_{18}N_4O_7S$, M_r 398.4, *mp* 144 – 147 °C, is a colorless to beige powder which is moderately soluble in water [23, p. 488].

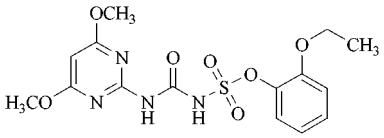

Ethoxysulfuron is produced by reaction of 2-ethoxyphenoxysulfonyl isocyanate with 2-amino-4,6-dimethoxypyrimidine [42].

Ethoxysulfuron is under development for the control of broad-leaved and sedge weeds in cereals, rice, and sugar cane.
*Trade Name.*To be announced (AgrEvo).

Flazasulfuron [*104040-78-0*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-(3-trifluoromethyl-2-pyridylsulfonyl)urea, $C_{13}H_{12}F_3N_5O_5S$, M_r 407.3, *mp* 166 – 170 °C, consists of colorless crystals which are readily soluble in water, methanol, acetone, and toluene [23, p. 550].

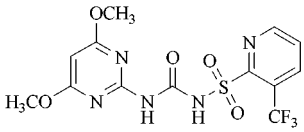

Flazasulfuron is produced by reaction of 2-sulfamoylchloride-3-trifluoromethylpyridine with 2-amino-4,6-dimethoxypyrimidine [43].

Flazasulfuron is used post-emergence for the control of grass and broad-leaved weeds and sedges in warm-season turf.
Trade Names. Katana, Shibagen (Ishihra Sangyo).

Flupyrsulfuron-methyl-sodium [*144740-54-5*], methyl 2-(4,6-dimethoxypyrimidin-2-ylcarbamoylsulfamoyl)-6-trifluoromethylnicotinate monosodium salt, $C_{15}H_{13}F_3N_5NaO_7S$, M_r 487.3, *mp* 165 – 170 °C, is a colorless powder which is sparingly soluble in water, and moderately soluble in dichloromethane, acetone, ethyl acetate, and acetonitrile [23, p. 586].

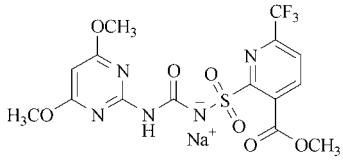

Flupyrsulfuron-methyl-sodium is produced by reaction of methyl 2-aminosulfonyl-6-trifluoromethyl-3-pyridine carboxylate with 4,6-dimethoxypyrimidin-2-yl isocyanate [44].

Flupyrsulfuron-methyl-sodium is under development for selective post-emergence control of black grass and other weeds in cereals.
*Trade Name.*To be announced (Du Pont).

Halosulfuron-methyl [*100784-20-1*], methyl 3-chloro-5-(4,6-dimethoxypyrimidin-2-ylcarbamoylsulfamoyl)-1-methylpyrazole-4-carboxylate, $C_{13}H_{15}ClN_6O_7S$, M_r 434.8, *mp* 175.5 – 177.2 °C, is a colorless powder which is moderately soluble in water and methanol [23, p. 657].

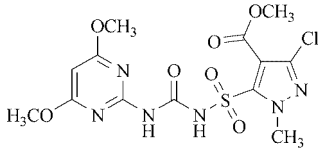

Halosulfuron-methyl is produced by reaction of *N*-1-methyl-3-chloro-4-carboxymethyl-5-sulfonylisocyanate pyrazole with 2-amino-4,6-dimethoxypyrimidine [45].

Halosulfuron-methyl is used to control annual broad-leaved weeds and nutsedge sp. in maize, sugar cane, and turf.
Trade Name. (Nissan, Monsanto).

Imazosulfuron [*122548-33-8*], 1-(2-chloroimidazo[1,2-*a*]pyridin-3-ylsulfonyl)-3-(4,6-dimethoxypyrimidin-2-yl)urea, $C_{14}H_{13}ClN_6O_5S$, M_r 412.8, *mp* 183 – 184 °C, is a crystalline powder which is sparingly soluble in water and is fairly soluble in acetonitrile,

acetone, xylene, dichloromethane, and ethyl acetate [23, p. 703].

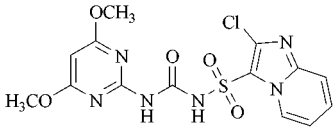

Imazosulfuron is produced by reaction of 2-chloroimidazo[1,2-*a*]pyridine sulfonylisocyanate with 2-amino-4,6-dimethoxypyrimidine [46].

Imazosulfuron is used to control most annual and perennial broad-leaved weeds and sedges in paddy rice and turf.

Trade Names. Sibatito, (Takeda).

Metsulfuron-methyl [*74223-64-6*], methyl 2-(4-methoxy-6-methyl-1,3,5-triazin-2-ylcarbamoylsulfamoyl)benzoate, $C_{14}H_{15}N_5O_6S$, M_r 381. 4, *mp* 158 °C, consists of colorless crystals which are fairly soluble in water, xylene, ethanol, methanol, and acetone [23, p. 842].

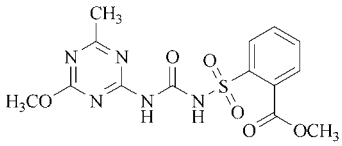

Metsulfuron-methyl is produced by reaction of 2-sulfamoylchloride benzoic acid with 2-amino-4-methoxy-6-methyl-1,3,5-triazine [47].

Metsulfuron-methyl is a selective systemic herbicide used to control a wide range of annual and perennial broad-leaved weeds in wheat, barley, rice, and oats, by pre- or post-emergence application.

Trade Names. Ally (Du Pont); Malban (IPESA); Quit (Sanonda), etc.

Nicosulfuron [*111991-09-4*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-(3-dimethylcarbamoyl-2-pyridinylsulfonyl)urea, $C_{15}H_{18}N_6O_6S$, M_r 410.4, *mp* 141 – 144 °C, consists of colorless crystal which are moderately soluble in water and are readily soluble in acetone, ethanol, chloroform, dimethylformamide, acetonitrile, and dichloromethane [23, p. 877].

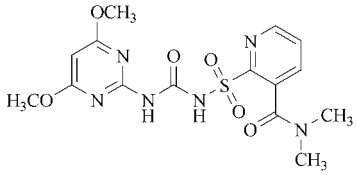

Nicosulfuron is produced by reaction of 2-sulfamoylchloride-3-*N*-dimethylamino-carboxypyridine with 2-amino-4,6-dimethylpyrimidine [48].

Nicosulfuron is a selective, systemic herbicide used post-emergence to control annual grass weeds in maize.

Trade Names. Accent (Du Pont); Dasul (Ishihara Sangyo, Dr. Maag); Nostoc (IPESA); Milagro (Ishihara Sangyo, Zeneca), etc.

Oxasulfuron [*144651-06-9*], oxetan-3-yl 2-[(4,6-dimethylpyrimidin-2-yl)-carbamoylsulfamoyl]benzoate, $C_{17}H_{18}N_4O_6S$, M_r 406.4, *mp* 158 °C (decomp.), is a colorless powder which is moderately soluble in water and fairly soluble in methanol, acetone, toluene, and dichloromethane [23, p. 911].

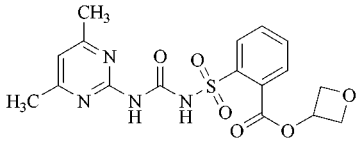

Oxasulfuron is produced by reaction of 2-[[(3-oxetanyl)oxy]carbonyl]benzenesulfonamide with 4,6-dimethylpyrimidine-2-yl carbamate [49].

Oxasulfuron is under development for the control of broad-leaved and grass weeds in soya beans.

Trade Names. Expert, Dynam (Novartis).

Primisulfuron-methyl [*86209-51-0*], methyl 2-[4,6-bis(difluoromethoxy)pyrimidin-2-ylcarbamoylsulfamoyl]benzoate, $C_{15}H_{12}F_4N_4O_7S$, M_r 468.3, *mp* 194.8 – 197.4 °C, is a fine colorless powder which is sparingly soluble in water and readily soluble in acetone, ethanol, and toluene [23, p. 997].

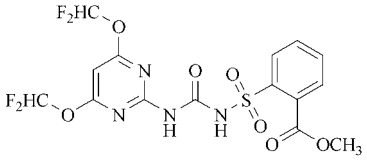

Primisulfuron-methyl is produced by reaction of 2-sulfamoylchloride benzoic acid with 2-amino-4,6-difluoromethoxypyrimidine [50].

Primisulfuron-methyl is a selective, systemic herbicide used post-emergence for the control of problem weeds and many broad-leaved weeds in maize.

Trade Names. Beacon, Tell (Novartis).

Prosulfuron [*94125-34-5*], 1-(4-methoxy-6-methyl-1,3,5-triazin-2-yl)-3-[2-(3,3,3-trifluoropropyl)phenylsulfonyl]urea, $C_{15}H_{16}F_3N_5O_4S$, M_r 419.4, *mp* 155 °C (decomp.), consists of colorless crystals which are sparingly soluble in water and readily soluble in ethanol, acetone, toluene, ethyl acetate, and dichloromethane [23, p. 1041].

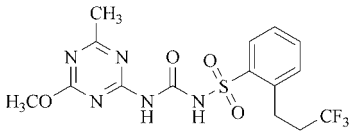

Prosulfuron is produced by reaction of 2-(3,3,3-trifluoropropyl)phenylsulfamoylchloride with 2-amino-4-methyl-6-methoxypyrimidine [51].

Prosulfuron is used post-emergence on maize, sorghum, cereals, pasture, and turf for broad-spectrum control of annual broad-leaved weeds.

Trade Names. Exceed, Peak (Novartis).

Pyrazosulfuron-ethyl [*93697-74-6*], ethyl 5-(4,6-dimethoxypyrimidin-2-ylcarbamoylsulfamoyl)-1-methylpyrazole-4-carboxylate, $C_{14}H_{18}N_6O_7S$, M_r 414.4, *mp* 181 – 182 °C, consists of colorless crystals which are sparingly soluble in water and fairly soluble in benzene, chloroform, and acetone [23, p. 1052].

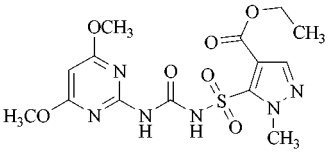

Pyrazosulfuron-ethyl is produced by reaction of *N*-1-methyl-4-carboxy-5-sulfonylisocyanate pyrazole with 2-amino-4,6-dimethoxypyrimidine [52].

Pyrazosulfuron-ethyl is a systemic herbicide used pre- and post-emergence to control annual and perennial broad-leaved weeds and sedges in wet-sown and transplanted rice crops.

Trade Names. Sparkstar (Novartis, Nisson, Zeneca); Sirius, (Nissan); Act (Nihon Bayer, Nissan); Billy (Sanonda), etc.

Rimsulfuron [*122931-48-0*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-(3-ethylsulfonyl-2-pyridylsolfonyl)urea, $C_{14}H_{17}N_5O_7S_2$, M_r 431.4, *mp* 176 – 178 °C, consists of colorless crystals which are fairly soluble in buffered solution (pH 7) [23, p. 1095].

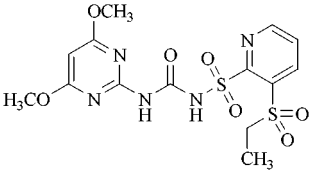

Rimsulfuron is produced by reaction of 2-sulfamoylchloride-3-sulfonylethylpyridine with 2-amino-4,6-dimethoxypyrimidine [53].

Rimsulfuron is used post-emergence to control annual and perennial grasses and several broad-leaved weeds in maize.

Trade Name. (Du Pont).

Sulfometuron-methyl [*74222-97-2*], methyl 2-(4,6-dimethylpyrimidin-2-ylcarbamoylsulfamoyl)benzoate, $C_{15}H_{16}N_4O_5S$, M_r 363.4, *mp* 203 – 205 °C, is a colorless solid which is sparingly soluble in water and fairly soluble in acetone, acetonitrile, dichloromethane, and dimethyl sulfoxide [23, p. 1128].

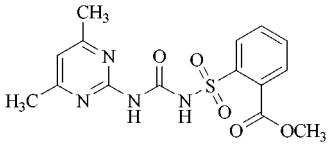

Sulfometuron-methyl is produced by reaction of 2-sulfonylisocyanate methylbenzoate with 2-amino-4,6-dimethylpyrimidine [54].

Sulfometuron-methyl is used to control annual and perennial grasses and broad-leaved weeds in noncropland, as well as in forestry to control woody tree species in pine forests.
Trade Name. Oust (Du Pont).

Sulfosulfuron [*141776-32-1*], 1-(4,6-dimethoxypyrimidin-2-yl)-3-(2-ethylsulfonylimidazo[1,2-*a*]pyridin-3-yl)sulfonylurea, $C_{16}H_{18}N_6O_7S_2$, M_r 470.8, *mp* 201.1 – 201.7 °C, is a colorless solid which is sparingly soluble in water [23, p. 1130].

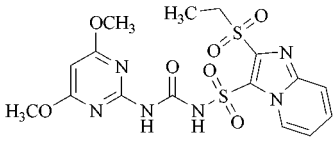

Sulfosulfuron is produced by reaction of 2-ethylsulfonylimidazo[1,2-*a*]pyridine 3-sulfonamide with 4,6-dimethoxypyrimidin-2-yl isocyanate [55].

Sulfosulfuron is under development for the control of grasses and broad-leaved weeds in cereals.

To be announced (Monsanto, Takeda).

Thifensulfuron-methyl [*79277-27-3*], methyl 3-(4-methoxy-6-methyl-1,3,5-triazin-2-ylcarbamoylsulfamoyl)thiophen-2-carboxylate, $C_{12}H_{13}N_5O_6S_2$, M_r 387.4, *mp* 176 °C, consists of colorless crystals which are sparingly soluble in water and moderately soluble in methanol, ethyl acetate, acetone, and dichloromethane [23, p. 1188].

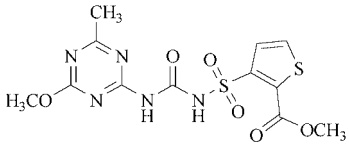

Thifensulfuron-methyl is produced by reaction of 2-(methoxycarbonyl)-3-thiophenesulfonyl isocyanate with 2-amino-4,6-dimethoxypyrimidine [56].

Thifensulfuron-methyl is used in winter and spring wheat, winter and spring barley, rye, triticale, oats, soya beans, etc.
Trade Names. Pinnacle, Harmony (Du Pont).

Triasulfuron [*82097-50-5*], 1-[2-(2-chloroethoxy)phenylsulfonyl]-3-(4-methoxy-6-methyl-1,3,5-triazin-2-yl)urea, $C_{14}H_{16}ClN_5O_5S$, M_r 401.8, *mp* 178.1 °C, is a fine colorless powder which is moderately soluble in water, ethanol, and *n*-hexane and is readily soluble in acetone, dichloromethane, and ethyl acetate [23, p. 1222].

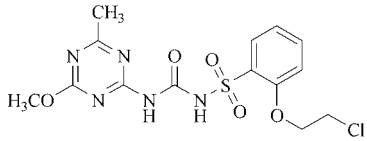

Triasulfuron is produced by reaction of 2-chloroethoxyphenylsulfamoyl chloride with 2-amino-4-methoxy-6-methyl-1,3,5-triazine [57].

Triasulfuron is a selective herbicide used pre- and post-emergence for the control of broad-leaved weeds in wheat, barley, and triticale.
Trade Names. Logran, Satis (Novartis).

Tribenuron-methyl [*101200-48-0*], methyl 2-[4-methoxy-6-methyl-1,3,5-triazin-2-yl-*N*-methylcarbamoylsulfamoyl]benzoate, $C_{15}H_{17}N_5O_6S$, M_r 395.4, *mp* 141 °C, is a light brown solid which is moderately soluble in water (pH-dependent, increasing with pH) and is sparingly soluble in acetone, acetonitrile, carbon tetrachloride, ethyl acetate, hexane, and methanol [23, p. 1230].

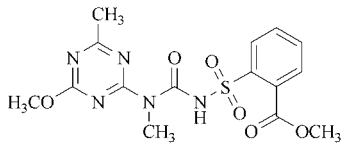

Tribenuron-methyl is produced by reaction of methyl-2-sulfamoylchloride benzoic acid with 2-methylamino-4-methoxy-6-methyl-1,3,5-triazine [58].

Tribenuron-methyl is used post-emergence for the control of broad-leaved weeds in spring and winter cereals.
Trade Names. Express, Granstar, Pointer (Du Pont).

Triflusulfuron-methyl [*126535-15-7*], methyl 2-[4-dimethylamino-6-(2,2,2-trifluromethoxy)-1,3,5-triazin-2-ylcarbamoylsulfamoyl]-*m*-toluate, $C_{17}H_{19}F_3N_6O_6S_2$, M_r 492.4, *mp* 160 – 163 °C, is a colorless crystalline solid which is sparingly soluble in water (pH 7), dichloromethane, methanol, toluene, and acetonitrile [23, p. 1250].

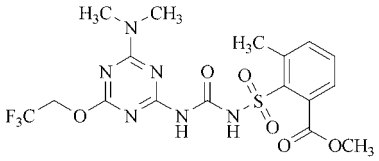

Triflusulfuron-methyl is produced by reaction of methyl 2-sulfamoylchloride-3-methylbenzoate with 2-amino-4-dimethylamino-6-(2,2,2-trifluoroethoxy)-1,3,5-triazine [59].

Triflusulfuron-methyl is used post-emergence for the control of many annual and perennial broad-leaved weeds in sugar beet.
Trade Names. Debut, Safari, Upbeet (Du Pont).

7.8.3. Chloroacetanilides

Acetochlor [*34256-82-1*], 2-chloro-*N*-ethoxymethyl-6′-ethylaceto-*o*-toluidide, $C_{14}H_{20}ClNO_2$, M_r 269.8, *mp* < 0 °C, *bp* 162 °C/7 mmHg, is a wine red to yellow or amber oil which is sparingly soluble in water and readily soluble in diethyl ether, acetone, benzene, chloroform, ethanol, ethyl acetate, and toluene [23, p. 10].

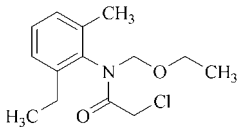

Acetochlor is produced by reaction of chloroacetyl chloride with the azomethine of 2-ethyl-6-methylaniline and formaldehyde, followed by treatment with ethanol [60].

Acetochlor is a selective herbicide used pre-emergence or pre-planting to control annual grasses, certain annual broad-leaved weeds, and yellow nutsedge in maize, peanuts, soya beans, cotton, potatoes, and sugar cane.
Trade Names. Harness (Monsanto); Surpass, Trophy (Zeneca); Acenit (Nitrokemia), etc.

[*15972-60-8*], 2-chloro-2′,6′-diethyl-*N*-methoxymethylacetanilide, $C_{14}H_{20}ClNO_2$, M_r 269.8, *mp* 40.5 – 41.5 °C, is a yellow to wine red solid which is sparingly soluble in water and *n*-hexane and is readily soluble in diethyl ether, acetone, benzene, chloroform, ethanol, and ethyl acetate [23, p. 23].

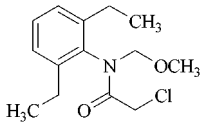

Alachlor is synthesized by reaction of chloroacetyl chloride with the azomethine of 2,6-diethylaniline and formaldehyde, followed by treatment with methanol [60].

Alachlor is a selective, systemic herbicide used pre-emergence to control annual grasses and many broad-leaved weeds in cotton, brassicas, maize, oilseed rape, peanuts, etc.
Trade Names. Lasso (Monsanto); Alanex (Markhteshim-Agan); Lacorn (Efthymiadis); Satochlor (Sagrochem), etc.

Butachlor [*23184-66-9*], *N*-butoxymethyl-2-chloro-2′,6′-diethylacetanilide, $C_{17}H_{26}NO_2$, M_r 311.9, *mp* – 2.8 to 1.7 °C, *bp* 156 °C/0.5 mmHg, is a light yellow to purple liquid which is sparingly soluble in water and is readily soluble in most organic solvents [23, p. 159].

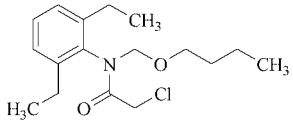

Butachlor is produced by reaction of chloroacetyl chloride with the azomethine of 2,6-diethylaniline and formaldehyde, followed by treatment with *n*-butanol [60].

Butachlor is a selective systemic herbicide used pre-emergence for the control of an-

nual grasses and certain broad-leaved weeds in seeded and transplanted rice.
Trade Names. Machete (Monsanto); Butataf (Ralis); Butanex (Makhteshim-Agan); Formachlor (Sanonda), etc.

Dimethachlor [*50563-36-5*], 2-chloro-*N*-(2-methoxyethyl)aceto-2′,6′-xylidide, $C_{13}H_{18}ClNO_2$, M_r 255.7, *mp* 45.8 – 46.7 °C, forms colorless crystals which are fairly soluble in water and are readily soluble in most organic solvents [23, p. 406].

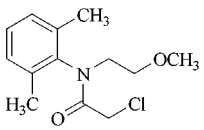

Dimethachlor is produced by reaction of 2,6-dimethylaniline with α-chloroethylmethyl ether, followed by reaction with chloroacetyl chloride [61].

Dimethachlor is a selective soil herbicide used pre-emergence for the control of annual grasses and broad-leaved weeds in oilseed rape.
Trade Name. Teridox (Novartis).

Metazachlor [*67129-08-2*], 2-chloro-*N*-(pyrazol-1-ylmethyl)acet-2′,6′-xylidide, $C_{14}H_{16}ClN_3O$, M_r 277.8, *mp* 85 °C, consists of yellowish crystals which are moderately soluble in water and are readily soluble in acetone, chloroform, ethyl acetate, and ethanol [23, p. 801].

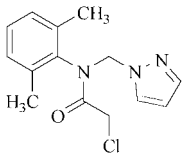

Metazachlor is produced by reaction of 2,6-dimethylaniline with formaldehyde, followed by treatment with chloroacetyl chloride, and then with potassium pyrazole [62].

Metazachlor is a selective herbicide used pre- and early post-emergence for the control of winter and annual grasses and broad-leaved weeds in a variety of crops.
Trade Name. Butisan S (BASF).

Metolachlor [*51218-45-2*], 2-chloro-*N*-(2-methoxy-1-methylethyl)aceto-*o*-toluidide, $C_{15}H_{22}ClNO_2$, M_r 283.8, *mp* −62.1 °C, *b*p 100 °C/0.001 mmHg, is a colorless to light tan liquid which is moderately soluble in water and is miscible with most organic solvents [23, p. 833].

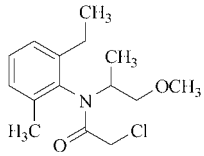

Metolachlor is produced by reaction of 2-methyl-6-ethylanilin with 2-chloroisopropyl methyl ether, followed by treatment with chloroacetyl chloride [63].

Metolachlor is a selective herbicide used pre-emergence and is incorporated pre-planting to control annual grasses and some broad-leaved weeds in many crops.
Trade Names. Dual, Dual S, Dual II (Novartis).

Pretilachlor [*51218-49-6*], 2-chloro-2′,6′-diethyl-*N*-(2-propoxyethyl)acetanilide, $C_{17}H_{26}ClNO_2$, M_r 311.9, *bp* 135 °C/0.001 mmHg, is a colorless liquid which is moderately soluble in water and readily soluble in benzene, hexane, methanol, and dichloromethane [23, p. 995].

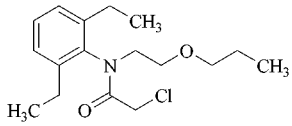

Pretilachlor is produced by reaction of 2,6-diethylamine with α-chloroethyl propyl ether, followed by reaction with chloroacetyl chloride [64].

Pretilachlor is a selective herbicide effective against the main annual grasses, broad-leaved weeds, and sledges in transplanted and seeded rice.
Trade Names. Rifit, Sofit, Solnet (Novartis); Gorbo (Takeda); Sparkstar (Nissan), etc.

Propachlor [*1918-16-7*], 2-chloro-*N*-isopropylacetanilide, $C_{11}H_{14}ClNO$, M_r 211.7, *mp* 77 °C, is a light tan solid which is moderately soluble in water and readily soluble in most common organic solvents [23, p. 1013].

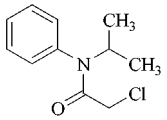

Propachlor is produced by reaction of aniline with 2-chloroisopropane, followed by treatment with chloroacetyl chloride [65].

Propachlor is a selective herbicide used pre-emergence and pre-planting-incorporated to control annual grasses and some broad-leaved weeds in a range of crops.

Trade Names. Ramrod (Monsanto); Albrass (Zeneca); Satecid (Sagrochem); Prolex (Makhteshim-Agon), etc.

Propisochlor [*86763-47-5*], 2-chloro-6′-ethyl-*N*-isopropoxymethylacetotoluidide, $C_{15}H_{22}ClNO_2$, M_r 283.8 *mp* 21.6 °C, *bp* 243 °C (decomp.), is a light brown to purple oil which is moderately soluble in water and soluble in most organic solvents [23, p. 1034].

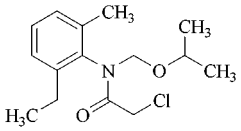

Propisochlor is produced by reaction of 2-methyl-6-ethylaniline with chloromethyl isopropyl ether, followed by reaction with chloroacetyl chloride [66].

Propisochlor is a selective herbicide used pre-emergence and pre-planting-incorporated for the control of annual grasses and some broad-leaved weeds in maize, sunflowers, soya beans, potatoes, and peas.

Trade Name. Proponit (Nitrokémia).

Thenylchlor [*96491-05-3*], 2-chloro-*N*-(3-methoxy-2-thenyl)-2′,6′-dimethylacetanilide, $C_{16}H_{18}ClNO_2S$, M_r 323.8, *mp* 72 – 74 °C, is a colorless solid which is slightly soluble in water [23, p. 1182].

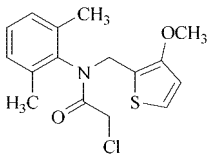

Thenylchlor is produced by reaction of 2,6-dimethylaniline with 3-methoxy-2-thenylmethyl chloride, followed by treatment with chloroacetyl chloride [67].

Thenylchlor is used pre-emergence for the control of annual grasses and broad-leaved weeds in paddy rice.

Trade Names. Kusamets, Onebest (Tokuyama).

7.8.4. 1,3,5-Triazines

Ametryn [*834-12-8*], N^2-ethyl-N^4-isopropyl-6-methylthio-1,3,5-triazine-2,4-diamine, $C_9H_{17}N_5S$, M_r 227.3, *mp* 84 – 86 °C, is a colorless powder which is moderately soluble in water and readily soluble in acetone, methanol, toluene, and *n*-octanol [23, p. 35].

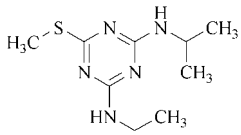

Ametryn is produced by reaction of atrazine with methylmercaptan in the presence of sodium hydroxide [68].

Ametryn is a selective, systemic herbicide used pre- and post-emergence for the control of most annual grasses and broad-leaved weeds in pineapples, sugar cane, bananas, citrus fruit, maize, coffee, tea, etc, and in noncropland.

Trade Names. Evik, Gesapax (Novartis); Mebatryne (Rhône-Poulenc); Amesip (Sipcam), etc.

Atrazine [*1912-24-9*], 6-chloro-N^2-ethyl-N^4-isopropyl-1,3,5-triazine-2,4-diamine, $C_8H_{14}ClN_5$, M_r 215.7, *mp* 178.5 °C, forms a colorless powder which is moderately soluble in water, is fairly soluble in toluene and *n*-octanol, and is readily soluble in ethyl acetate, acetone, dichloromethane, and ethanol [23, p. 55].

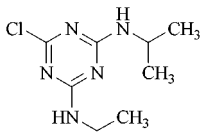

Atrazine is produced by reaction of cyanuric chloride with one equivalent of ethylamine, followed by treatment with one equivalent of isopropylamine [69].

Atrazine is a selective, systemic herbicide used pre- and post-emergence to control annual broad-leaved weeds and annual grasses in maize, sorghum, sugar cane, etc, and in industrial weed control.
Trade Names. Atrex, Bicep, Gesaprim, Primagram, Primextra (Novartis); Mebazine (Rhône-Poulenc); Vectal (AgrEvo), etc.

Cyanazine [*21725-46-2*], 2-(4-chloro-6-ethylamino-1,3,5-triazin-2-ylamino)-2-methylpropiononitrile, $C_9H_{13}ClN_6$, M_r 240.7, *mp* 167.5 – 169 °C, is a colorless crystalline solid which is moderately soluble in water, fairly soluble in benzene and carbon tetrachloride, and readily soluble in chloroform, acetone, and ethanol [23, p. 280].

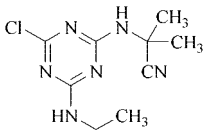

Cyanazine is produced by successive reaction of cyanuric chloride with ethylamine and 2-methyl-2-cyanopropylamine [70].

Cyanazine is a selective herbicide used pre-emergence for general weed control.
Trade Names. Bladex (Cyanamid, Du Pont); Cy-Pro (Griffin).

Desmetryn [*1014-69-3*], N^2-isopropyl-N^4-methyl-6-methylthio-1,3,5-triazine-2,4-diamine, $C_8H_{15}N_5S$, M_r 213.3, *mp* 84 – 86 °C, is a colorless powder which is moderately soluble in water and readily soluble in most organic solvents [23, p. 351].

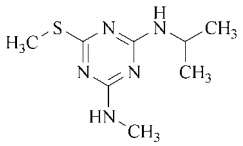

Desmetryn is produced by reaction of cyanuric chloride with, in turn, isopropylamine, methylamine, and methanethiol [69].

Desmetryn is a selective systemic herbicide used post-emergence for the control of fat-hen and other annual broad-leaved weeds and grasses in brassica, herbs, onions, leeks, and conifer seedbeds.

Trade Name. Semeron (Novartis).

Dimethametryn [*22936-75-0*], N^2-(1,2-dimethylpropyl)-N^4-ethyl-6-methylthio-1,3,5-triazine-2,4-diamine, $C_{11}H_{21}N_5S$, M_r 255.4, *mp* 65 °C, forms colorless crystals which are moderately soluble in water and readily soluble in acetone, dichloromethane, methanol, *n*-octanol, and toluene [23, p. 407].

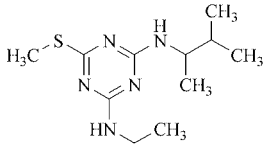

Dimethametryn is produced by reaction of cyanuric chloride with ethylamine, followed by treatment with isobutylamine [71].

Dimethametryn is a selective herbicide used for the control of annual broad-leaved weeds in transplanted- and seeded rice.
Trade Names. Avirosan (Novartis); Sparkstar (Nissan).

Prometon [*1610-18-0*], N^2,N^4-diisopropyl-6-methoxy-1,3,5-triazine-2,4-diamine, $C_{10}H_{19}N_5O$, M_r 225.3, *mp* 91 – 92 °C, is a colorless powder which is moderately soluble in water and is readily soluble in benzene, methanol, acetone, dichloromethane, and toluene [23, p. 1010].

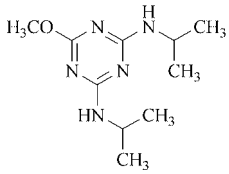

Prometon is produced by reaction of propazine with methanol in the presence of sodium hydroxide [68].

Prometon is a nonselective, systemic herbicide which is used to control most annual and many perennial broad-leaved weeds, grasses, and brush weeds on non-crop areas.

Trade Name. Pramitol (Novartis).

Prometryn [*7287-19-6*], N^2,N^4-di-isopropyl-6-methylthio-1,3,4-triazine-2,4-diamine, $C_{10}H_{19}N_5S$, M_r 214.4, *mp* 18 – 120 °C, is a colorless powder which is moderately soluble in water, fairly soluble in hexane, and readily soluble in acetone, ethanol, toluene, and *n*-octanol [23, p. 1011].

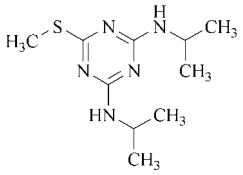

Prometryn is produced by reaction of propazine with methylmercaptan in the presence of sodium hydroxide, followed by treatment with dimethyl sulfate [68].

Prometryn is a selective, systemic herbicide used pre-emergence in cotton, sunflowers, peanuts, cereals, peas, and beans and used post-emergence in cotton, potatoes, carrots, celery, and leeks.

Trade Names. Caparol, Gesagard (Novartis); Cotton-Pro (Griffin); Efmetryn (Efthymiadis), etc.

Propazine [*139-40-2*], 6-chloro-N^2,N^4-diisopropyl-1,3,5-triazine-2,4-diamine, $C_9H_{16}ClN_5$, M_r 129.7, *mp* 212 – 214 °C, is a colorless powder which is slightly soluble in water and fairly soluble in benzene, toluene, diethyl ether, and carbon tetrachloride [23, p. 1024].

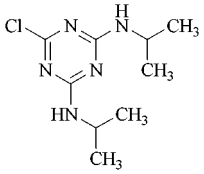

Propazine is produced by the reaction between cyanuric chloride and two equivalents of isopropylamine [72].

Propazine is a selective, systemic herbicide used pre-planting and pre-emergence for the control of grasses and broad-leaved weeds in sorghum and umbelliferous crops.

Trade Names. Gesamil, Milogard (Novartis); Milo-Pro (Griffin); etc.

Simazine [*122-34-9*], 6-chloro-N^2,N^4-diethyl-1,3,5-triazine-2,4-diamine, $C_7H_{12}ClN_5$, M_r 201.7, *mp*225 – 227 °C, is a colorless powder which is slightly soluble in water, *n*-hexane and is moderately soluble in ethanol, acetone, toluene, and *n*-octanol [23, p. 1106].

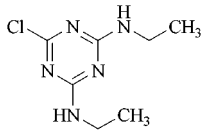

Simazine is produced by reaction of cyanuric chloride with two equivalents of ethylamine [73].

Simazine is a selective, systemic herbicide used for the control of most germinating annual grasses and broad-leaved weeds in many crops.

Trade Names. Caliber, Gesatop, Princep, Princep Caliber (Novartis); Amizina (Sipcam); Sanasim (Sanachem), etc.

Simetryn [*1014-70-6*], N^2,N^4-diethyl-6-methylthio-1,3,5-triazine-2,4-diamine, $C_8H_{15}N_5S$, M_r 213.3, *mp* 82 – 83 °C, consists of colorless crystals which are moderately soluble in water, readily soluble in methanol, acetone, and toluene, and fairly soluble in *n*-hexane [23, p. 1108].

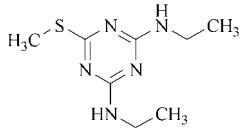

Simetryn is produced by reaction between simazine and methanethiol [68].

Simetryn is a selective herbicide used in combination with thiobencarb to control broad-leaved weeds in rice.

Trade Name. Gy-bon (Novartis, Nihon Nohyaku).

Terbumeton [*33693-04-8*], N^2-*tert*-butyl-N^4-ethyl-6-methoxy-1,3,5-triazine-2,4-diamine, $C_{10}H_{19}N_5O$, M_r 225.3, *mp* 123 – 124 °C, forms

colorless crystals which are moderately soluble in water and are readily soluble in acetone, toluene, methanol, dichloromethane, and *n*-hexane [23, p. 1167].

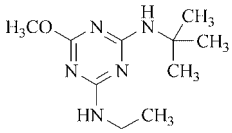

Terbumeton is produced by reaction of 2-*tert*-butylamino-4-chloro-6-ethylamino-*s*-triazine with methanol and sodium hydroxide [68].

Terbumeton is a selective herbicide used to control annual and perennial grasses and broad-leaved weeds in citrus orchards. In combination with terbuthylazine, it is used for the herbicidal control in vineyards, apple- and citrus orchards, and in forestry.
Trade Name. Caragard (Novartis).

Terbuthylazine [*5915-41-3*], N^2-*tert*-butyl-6-chloro-N^4-ethyl-1,3,5-triazine-2,4-diamine, $C_9H_{16}ClN_5$, M_r 229.7, *mp* 177 – 179 °C, is a colorless powder which is slightly soluble in water and fairly soluble in acetone, ethanol, and *n*-octanol. [23, p. 1168].

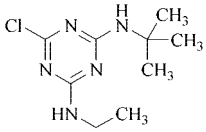

Terbuthylazine is produced by reaction of cyanuric chloride with *tert*-butylamine, followed by treatment with ethylamine in the presence of sodium hydroxide [74].

Terbuthylazine is a broad-spectrum herbicide used pre- and post-emergence for the control of weeds in many crops, in forestry, and in tree nurseries.
Trade Names. Gardoprim, Folar, Topogard (Novartis); Tyllanex (Makhteshim-Agan), etc.

Terbutryn [*886-50-0*], N^2-*tert*-butyl-N^4-ethyl-6-methylthio-1,3,5-triazine-2,4-diamine, $C_{10}H_{19}N_5S$, M_r 241.4, *mp* 104 – 105 °C, is a colorless powder which is slightly soluble in water and is readily soluble in many organic solvents. [23, p. 1170].

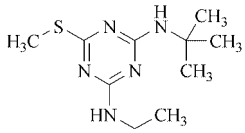

Terbutryn is produced by reaction of 2-chloro-4-ethylamino-4,6-butylamino-1,3,5-triazine with methyl mercaptan in the presence of sodium hydroxide [68].

Terbutryn is a selective herbicide used pre-emergence as well as post-emergence to control black grass and annual meadow grass in winter cereals, sugar cane, sunflowers, and other crops.
Trade Names. Igran (Novartis); Terbutrex (Makhteshim-Agan), etc.

Trietazine [*1912-26-1*], 6-chloro-N^2,N^2,N^4-triethyl-1,3,5-triazine-2,4-diamine, $C_9H_{16}ClN_5$, M_r 229.7, *mp* 102 – 103 °C, consists of colorless crystals which are moderately soluble in water and are readily soluble in acetone, benzene, chloroform, dioxane, and ethanol [23, p. 1243].

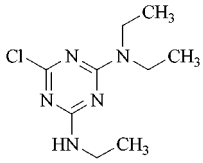

Trietazine is produced by reaction of trichloro-1,3,5-triazine with ethylamine, followed by treatment with diethylamine [75].

Trietazine is a selective herbicide, used for weed control in combination with linuron in potatoes, in combination with simazine in peas, and in combination with terbutryn in potatoes, peas, and field beans.
Trade Names. Remtal (AgrEvo); Aventox (DowElanco).

7.8.5. Ureas

Chlorbromuron [*13360-45-7*], (4-bromo-3-chlorophenyl)-1-methoxy-1-methylurea, $C_9H_{10}BrClN_2O_2$, M_r 293.5, *mp* 95 – 97 °C, is a colorless powder which is moderately soluble in water and is readily soluble in acetone, dichloromethane, hexane, benzene, and isopropyl alcohol [23, p. 204].

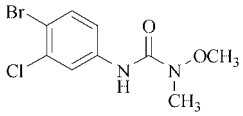

Chlorbromuron is made by bromination of *N*-(3-chlorophenyl)-*N′*-methoxy-*N′*-methyl urea [76].

Chlorbromuron is a selective herbicide used pre-emergence on carrots, peas, potatoes, sugar cane, and sunflowers; it is used post-emergence on carrots and transplanted celery.
Trade Name. Maloran (Novartis).

Chlorotoluron [*15545-48-9*], 3-(3-chloro-*p*-tolyl)-1,1-dimethylurea, $C_{10}H_{13}ClN_2O$, M_r 212.7, *mp* 148.1 °C, is a colorless powder which is moderately soluble in water and hexane and is fairly soluble in acetone, dichloromethane, ethanol, *n*-octanol, and ethyl acetate [23, p. 229].

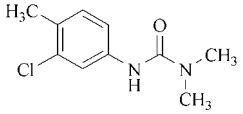

Chlorotoluron is produced by reaction of 3-chloro-*p*-tolyl isocyanate with dimethylamine [77].

Chlorotoluron is a selective herbicide used as a residual soil-acting herbicide and as a contact foliar spray against many broad-leaved weeds and grass weeds of winter cereals.
Trade Names. Dicuran (Novartis); Cekutoluron (Cequisa); Chlortophyt (Agriohar), etc.

Daimuron [*42609-52-9*], 1-(1-methyl-1-phenylethyl)-3-*p*-tolylurea, $C_{17}H_{20}N_2O$, M_r 268.4, *mp* 203 °C, forms colorless needles which are sparingly soluble in water and hexane and are fairly soluble in acetone and methanol [23, p. 330].

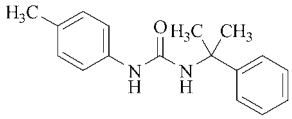

Daimuron is produced by reaction of isopropylaniline with *p*-toluene isocyanate [78].

Daimuron is a selective herbicide used pre- and post-emergence to control cyperaceous weeds and annual grass weeds in paddy rice.
Trade Name. Showrone (SDS Biotech).

Dimefuron [*34205-21-5*], [4-(5-*tert*-butyl-2,3-dihydro-2-oxo-1,3,4-oxadiazol-3-yl)-3-chlorophenyl]-1,1-dimethylurea, $C_{15}H_{19}ClN_4O_3$, M_r 338.8, *mp* 193 °C, consists of colorless crystals which are slightly soluble in water, benzene, toluene, and xylene, are readily soluble in chloroform, and are moderately soluble in acetone, acetonitrile, and ethanol [23, p. 403].

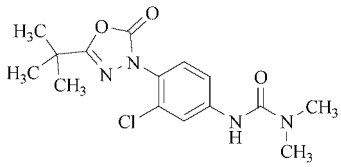

Dimefuron is produced by the reduction of the nitro group in 4-(5-*tert*-butyl-2,5-dihydro-2-oxo-1-oxadiazol-3-yl)-3-chloronitrobenzene, followed by treatment with phosgene and then with dimethylamine [79].

Dimefuron is a selective herbicide used pre- and post-emergence against different annual broad-leaved weeds in many crops.
Trade Names. Pradone (Rhône-Poulenc); Scorpion (AgrEvo).

Diuron [*330-54-1*], 3-(3,4-dichlorophenyl)-1,1-dimethylurea, $C_9H_{10}Cl_2N_2O$, M_r 233.1, *mp* 158 – 159 °C, forms colorless crystals which are slightly soluble in water, are sparingly soluble in hydrocarbons, and are fairly soluble in acetone, and benzene [23, p. 443].

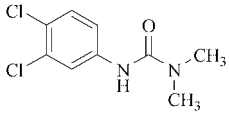

Diuron is produced by reaction of 3,4-dichlorophenyl isocyanate with dimethylamine [80].

Diuron is a systemic herbicide which is used to control weeds and mosses on noncrop areas and is used for selective control of germinating grass and broad-leaved weeds in many crops.

Trade Names. Karmex (Du Pont); Seduron (Rhône-Poulenc); Unidron (Unicrop); Toterbane (Diachem), etc.

Fenuron [*101-42-8*], 1,1-dimethyl-3-phenylurea, $C_9H_{12}N_2O$, M_r 164.2, *mp* 133 – 134 °C, consists of colorless crystals which are fairly soluble in water, diethyl ether, acetone, and benzene and are readily soluble in ethanol, acetone, and chloroform [23, p. 537].

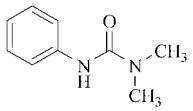

Fenuron is produced by reaction of phenyl isocyanate with dimethylamine [80].

Fenuron is a nonselective, systemic herbicide used for the control of woody plants and deep-rooted perennial weeds on noncropland.

Trade Name. Croptex Chrome (Hortichem), etc.

Fluometuron [*2164-17-2*], 1,1-dimethyl-3-(α,α,α-trifluoro-*m*-tolyl)urea, $C_{10}H_{11}F_3N_2O$, M_r 232.3, *mp* 163 – 164 °C, consists of colorless crystals which are moderately soluble in water and hexane, are fairly soluble in dichloromethane and *n*-octanol, and are readily soluble in methanol and acetone [23, p. 578].

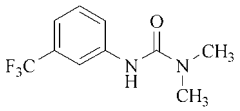

Fluometuron is produced by reaction of 3-trifluoromethylphenyl isocyanate with dimethylamine [81].

Fluometuron is a selective, systemic herbicide used to control annual broad-leaved weeds and grasses in cotton and soya beans.

Trade Names. Cotogard, Cotoran (Novartis); Meturon (Griffin), etc.

Isoproturon [*34123-59-6*], 3-(4-isopropylphenyl)-1,1-dimethylurea, $C_{12}H_{18}N_2O$, M_r 206.3, *mp* 158 °C, consists of colorless crystals which are moderately soluble in water and hexane, are fairly soluble in benzene and xylene, and are readily soluble in methanol, dichloromethane, and acetone [23, p. 732].

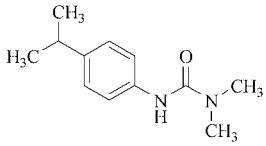

Isoproturon is produced by reaction of 4-isopropylphenyl isocyanate with dimethylamine [82].

Isoproturon is a selective, systemic herbicide used post-emergence for the control of annual grasses and many annual broad-leaved weeds in spring and winter wheat, spring and winter barley, winter rye, and titricale.

Trade Names. Panron (Sanonda), Alon, Arelon (AgrEvo); Strong (Rhône-Poulenc); Pasport (Searle India), etc.

Isouron [*55861-78-4*], 3-(5-*tert*-butylisoxazol-3-yl)-1,1-dimethylurea, $C_{10}H_{17}N_3O_2$, M_r 211.3, *mp* 119 – 120 °C, consists of colorless crystals which are moderately soluble in water and readily soluble in ethanol, acetone, and xylene [23, p. 734].

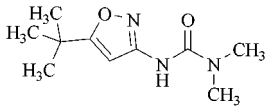

Isouron is produced by reaction of 3-*tert*-butyl-isoxazol-2-yl isocyanate with dimethylamine [83].

Isouron is a selective, systemic herbicide used pre- and post-emergence for the control of broad-leaved weeds and grasses in sugar cane and pineapples.

Trade Name. Isoxyl (Shionogi).

Linuron [*330-55-2*], 3-(3,4-dichlorophenyl)-1-methoxy-1-methylurea, $C_9H_{10}Cl_2N_2O_2$, M_r 249.1, *mp* 93 – 95 °C, forms colorless crystals which are moderately soluble in water and aromatic hydrocarbons, are sparingly soluble in aliphatic hydrocarbons, and are readily soluble in acetone, ethanol, chloroform, and dimethylformamide [23, p. 751].

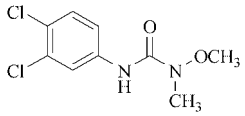

Linuron is produced by reaction of 3,4-dichlorophenyl isocyanate with methoxymethylamine [84].

Linuron is a selective, systemic herbicide used pre- and post-emergence to control annual grasses and broad-leaved weeds, and some seedlings of perennial weeds in a wide range of crops.

Trade Names. Lorox (Du Pont); Afalon (AgrEvo); Linex (Griffin); Siolcid (Siapa); etc.

Methabenzthiazuron [*1891-97-9*], 1-(1,3-benzothiazol-2-yl)-1,3-dimethylurea, $C_{10}H_{11}N_3OS$, M_r 221.3, *mp* 119 – 121 °C, consists of colorless crystals which are moderately soluble in water, fairly soluble in methanol, isopropyl alcohol, and toluene, and readily soluble in acetone, dimethylformamide, and dichloromethane [23, p. 805].

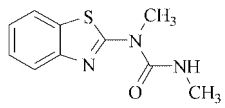

Methabenzthiazuron is produced by reaction of 2-methylaminobenzothiazole with methylisocyanate [85].

Methabenzthiazuron is a selective herbicide used to control a broad spectrum of broad-leaved weeds and grasses in cereals, legumes, maize, garlic, and onions.

Trade Name. Tribunil (Bayer).

Metobenzuron [*111578-32-6*], (±)-1-methoxy-3-[4-(2-methoxy-2,4,4-trimethylchroman-7-yloxy)phenyl]-1-methylurea, $C_{22}H_{28}N_2O_5$, M_r 40.5, *mp* 101.0 – 102.5 °C, is a colorless powder which is sparingly soluble in water [23, p. 830].

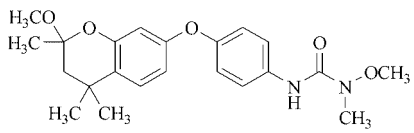

Metobenzuron is produced by reaction of 2-(2-methoxy-2,4,4-trimethylchroman-7-yloxy)phenyl isocyanate with dimethylamine [86].

Metobenzuron is a selective herbicide used post-emergence to control a wide range of broad-leaved weeds in maize.

Trade Name. To be announced (Mitsui).

Metobromuron [*3060-89-7*], 3-(4-bromophenyl)-1-methoxy-1-methylurea, $C_9H_{11}BrN_2O_2$, M_r 259.1, *mp* 95.5 – 96 °C, consists of colorless crystals which are moderately soluble in water, fairly soluble in hexane, and readily soluble in acetone, dichloromethane, methanol, toluene, octanol, and chloroform [23, p. 831].

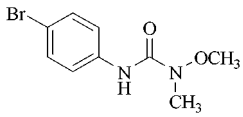

Metobromuron is produced by bromination of 3-phenyl-1-methoxy-1-methylurea [87].

Metobromuron is a selective, systemic herbicide used pre-emergence for the control of annual broad-leaved weeds and grasses in beans, potatoes, tomatoes, tobacco, soya beans, maize, artichokes, and sugar beet.

Trade Names. Patoran (Novartis); Paltonex (Makhteshim-Agan), etc.

Metoxuron [*19937-59-8*], 3-(3-chloro-4-methoxyphenyl)-1,1-dimethylurea, $C_{10}H_{13}ClN_2O_2$, M_r 228.7, *mp* 126 – 127 °C, forms colorless crystals which are moderately soluble in water, diethyl ether, benzene, and toluene, are practically insoluble in petroleum ether, and are readily soluble in acetone, cyclohexanone, acetonitrile, and hot ethanol [23, p. 838].

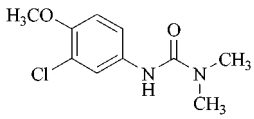

Metoxuron is produced by reaction of 3-chloro-4-methoxyphenyl isocyanate with dimethylamine [76].

Metoxuron is a selective herbicide used pre- and post-emergence for the control of some grasses and annual broad-leaved weeds in winter wheat, winter barley, winter rye, and some varieties of spring wheat.
Trade Names. Dasonex (Cyanamid); Deftor (AgrEvo); Sulerex (Siagra), etc.

Monolinuron [*1746-81-2*], 3-(4-chlorophenyl)-1-methoxy-1-methylurea, $C_9H_{11}ClN_2O_2$, M_r 214.6, *mp* 80 – 83 °C, consists of colorless crystals which are moderately soluble in water, and are readily soluble in common organic solvents [23, p. 851].

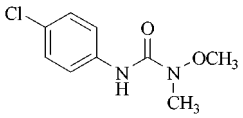

Monolinuron is produced by reaction of 4-chlorophenyl isocyanate with *N*-methylhydroxylamine, followed by treatment with dimethyl sulfate [88].

Monolinuron is a selective, systemic herbicide used pre- and post-emergence to control broad-leaved weeds and some annual grasses in vegetables, ornamental shrubs, and trees.
Trade Names. Arresin, Aresin (AgrEvo).

Neburon [*555-37-3*], 1-butyl-3-(3,4-dichlorophenyl)-1-methylurea, $C_{12}H_{16}Cl_2N_2O$, M_r 275.2, *mp* 102 – 103 °C, forms colorless crystals which are slightly soluble in water and are sparingly soluble in hydrocarbons [23, p. 872].

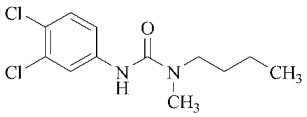

Neburon is produced by reaction of 3,4-dichlorophenyl isocyanate with *N*-methylbutylamine [80].

Neburon is a selective herbicide used pre-emergence for the control of broad-leaved weeds and grasses in beans, peas, cereals, beet, ornamentals, and in forestry.
Trade Names. Granurex (Rhône-Poulenc); Neburex (Makhteshim-Agan).

Siduron [*1882-49-6*], 1-(2-methylcyclohexyl)-3-phenylurea, $C_{14}H_{20}N_2O$, M_r 232.3, *mp* 133 – 138 °C, consists of colorless crystals which are moderately soluble in water and readily soluble in dimethylformamide, ethanol, and dichloromethane [23, p. 1103].

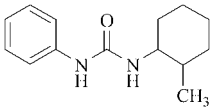

Siduron is produced by reaction of phenyl isocyanate with 2-methylcyclohexylamine [89].

Siduron is a selective herbicide used pre-emergence to control Digitaria sp. and annual grass weeds in newly seeded grass or established turf.
Trade Name. Tupersan (Gowan).

Tebuthiuron [*34014-18-1*], 1-(5-*tert*-butyl-1,3,4-thiadiazol-2-yl)-1,3-dimethylurea, $C_9H_{16}N_4OS$, M_r 228.3, *mp* 161.5 – 164 °C (decomp.), forms colorless crystals which are fairly soluble in water, benzene, and hexane, and are readily soluble in acetonitrile, acetone, methanol, and chloroform [23, p. 1153].

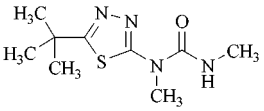

Tebuthiuron is produced by reaction of 5-*tert*-butyl-2-*N*-methyl isocyanate-1,3,4-thiadiazole with methylamine [90].

Tebuthiuron is a systemic broad-spectrum soil herbicide used for the control of herbaceous- and woody plants, as well as broad-leaved weeds. It is also used for the control of total vegetation in non-crop areas, undesirable plants in pasture and rangeland, and broad-leaved weeds and grass in sugar cane.
Trade Names. Spike (DowElanco); Bushwacker (Rhône-Poulenc); Tebusan (Sanachem).

7.8.6. 2,6-Dinitroanilines

Benfluralin [*1861-40-1*], *N*-butyl-*N*-ethyltextitα,α,α-trifluoro-2,6-dinitro-*p*-toluidine, $C_{13}H_{16}F_3N_3O_4$, M_r 335.3, *mp* 121 – 122 °C,

forms yellow-orange crystals which are practically insoluble in water, fairly soluble in hexane and methanol, and readily soluble in acetone, ethyl acetate, dichloromethane, chloroform, toluene, and acetonitrile [23, p. 94].

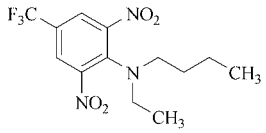

Benfluralin is produced by nitration of 4-chloro-trifluorotoluene, followed by treatment with *N*-ethylbutylamine [91].

Benfluralin is a selective soil herbicide that is incorporated pre-emergence in the soil to control annual grasses and some annual broad-leaved weeds in peanuts, lettuce, cucumbers, chicory, endive, field beans, French beans, etc.
Trade Names. Balan, Team, XL-2G (DowElanco); Benefex (Makhteshim-Agan).

Butralin [*33629-47-9*], *N-sec*-butyl-4-*tert*-butyl-2,6-dinitroaniline, $C_{14}H_{21}N_3O_4$, M_r 295.3, *mp* 61 °C, forms yellow-orange crystals which are sparingly soluble in water and readily soluble in methanol and hexane [23, p. 165].

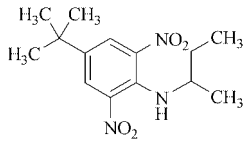

Butralin is produced by nitration of *tert*-butyl-4-chlorobenzene and subsequent reaction with *sec*-butylamine [92].

Butralin is a selective herbicide used pre-emergence for the control of annual broad-leaved weeds and grasses in cotton, soya beans, rice, barley, beans, vines, ornamentals, orchards of fruit and nut trees, etc.
Trade Names. Amexine (CFPI), Tamex (CFPI, Rhône-Poulenc).

Dinitramine [*29091-05-2*], *N,N*-diethyl-2,6-dinitro-4-trifluoromethyl-*m*-phenylenediamine, $C_{11}H_{13}F_3N_4O_4$, M_r 322.2, *mp* 98 – 99 °C, consists of yellow crystals which are sparingly soluble in water, moderately soluble in *n*-hexane, and readily soluble in acetone, chloroform, benzene, xylene, and ethanol [23, p. 423].

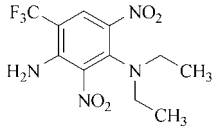

Dinitramine is produced by reaction of 2-amino-3,5-dinitro-4-chlorotrifluorotoluene with diethylamine [93].

Dinitramine is a selective soil herbicide used pre-planting and soil-incorporated for the control of many grass- and broad-leaved weeds in cotton, vegetables, sunflowers, transplanted tomatoes, aubergines, etc.
Trade Name. Cobex (Wacker).

Ethalfluralin [*55283-68-6*], N-ethyl-α,α,α-trifluoro-*N*-(2-methylallyl)-2,6-dinitro-*p*-toluidine, $C_{13}H_{14}F_3N_3O_4$, M_r 333.3, *mp* 55 – 56 °C, forms yellow-orange crystals which are sparingly soluble in water and are fairly soluble in acetone, acetonitrile, benzene, chloroform, dichloromethane, xylene, and methanol [23, p. 473].

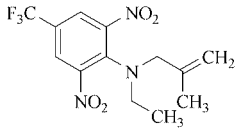

Ethalfluralin is produced by reaction of 2,6-dinitro-4-trifluoromethylchlorobenzene with *N*-ethylmethylallylamine [94].

Ethalfluralin is a selective soil herbicide used pre-planting and soil-incorporated for the control of most germinating annual grasses and broad-leaved weeds in cotton, soya beans, dry beans, lentils, peanuts, etc.
Trade Names. Sonalan, Edge (DowElanco).

Fluchloralin [*33245-39-5*], *N*-(2-chloroethyl)-2,6-dinitro-*N*-propyl-4-trifluoromethylaniline, $C_{12}H_{13}ClF_3N_3O_4$, M_r 355.7, *mp* 42 – 43 °C, is a yellow-orange solid which is sparingly soluble in water and readily soluble in acetone, benzene, chloroform, diethyl ether, ethyl acetate, cyclohexane, and ethanol [23, p. 560].

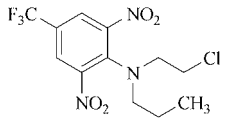

Fluchloralin is produced by reaction of 2,6-dinitro-4-trifluoromethylbenzene with *N*-chloroethylpropylamine [95].

Fluchloralin is used pre- and post-emergence and soil-incorporated for the control of annual grasses and some broad-leaved weeds in several crops.

Trade Name. Baselin (BASF).

Oryzalin [*19044-88-3*], 3,5-dinitro-N^4,N^4-dipropylsulfanilamide, $C_{12}H_{18}N_4O_6S$, M_r 346.4, *mp* 141 – 142 °C, forms yellow-orange crystals which are moderately soluble in water, fairly soluble in benzene and xylene, and readily soluble in acetone, acetonitrile, methanol, and dichloromethane, [23, p. 900].

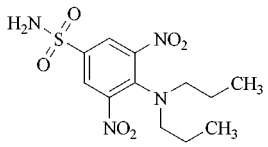

Oryzalin is produced by reaction of 4-dipropylamino-3,5-dinitrobenzene sulfonylchloride with ammonium hydroxide [96].

Oryzalin is a selective herbicide used pre-emergence for the control of many annual grasses and broad-leaved weeds in a variety of crops and in noncrop areas.

Trade Name. Surflan (DowElanco).

Pendimethalin [*40487-42-1*], *N*-(1-ethylpropyl)-2,6-dinitro-3,4-xylidine, $C_{13}H_{19}N_3O_4$, M_r 281.3, *mp* 54 – 58 °C, consists of yellow-orange crystals which are slightly soluble in water and readily soluble in many organic solvents [23, p. 937].

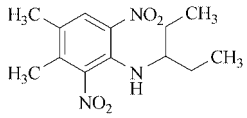

Pendimethalin is produced by reaction of 2,6-dichloro-4,5-dimethyl chlorobenzene with 1-ethylpropylamine [97].

Pendimethalin is a selective herbicide used pre-planting incorporated, and pre- and post-emergence for the control of most annual grasses and many broad-leaved weeds in a number of crops.

Trade Names. Herbadox, Prowl, Stomp (Cyanamid).

Prodiamine [*29091-21-2*], 5-dipropylamino-α,α,α-trifluoro-4,6-dinitro-*o*-toluidine, $C_{13}H_{17}F_3N_4O_4$, M_r 350.3, *mp* 122.5 – 124 °C, forms yellow crystals which are practically insoluble in water, are fairly soluble in acetonitrile, benzene, ethanol, hexane, and xylene, and are readily soluble in acetone, and chloroform [23, p. 1004].

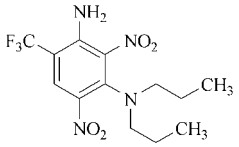

Prodiamine is produced by reaction of 2-amino-3,5-dinitro-4-chloro-α,α,α-trifluorotoluene with dipropylamine [98].

Prodiamine is a selective herbicide used pre-planting or pre-emergence for the control of annual grasses and broad-leaved weeds in cotton, nonbearing fruit trees, vines, nuts, ornamentals, turf, etc.

Trade Names. Barricade, Kusablock (Novartis).

Trifluralin [*1582-09-8*], α,α,α-trifluoro-2,6-dinitro-*N*,*N*-dipropyl-*p*-toluidine, $C_{13}H_{16}F_3N_3O_4$, M_r 335.3, *mp* 48.5 – 49 °C, forms yellow-orange crystals which are sparingly soluble in water, fairly soluble in methanol and hexane, and readily soluble in acetone, chloroform, acetonitrile, toluene, and ethyl acetate [23, p. 1248].

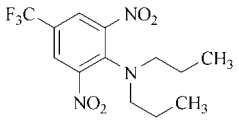

Trifluralin is produced by nitration of 4-chloro-trifluorotoluene, followed by treatment with dipropylamine [99].

Trifluralin is a selective soil herbicide used pre-emergence for the control of annual grasses and broad-leaved weeds in many crops.

Trade Names. Treflan (DowElanco); Commerce (FMC); Zeltoxone (Zeneca); Tri-4 (Cyanamid); Trifsan (Sanachem); Trilin (Griffin), Triplen (Sipcam), etc.

7.8.7. Imidazolinones

Imazamethabenz-methyl [*81405-85-8*], a mixture of (±)-6-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)-*m*-toluate and (±)-2-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)-*p*-toluate, $C_{16}H_{20}N_2O_3$, M_r 288.3, *mp* 113 – 153 °C, is a fine to lumpy powder which is moderately soluble in water, *n*-hexane, and *n*-heptane, and is readily soluble in acetone, dimethyl sulfoxide, isopropyl alcohol, methanol, and toluene [23, p. 694].

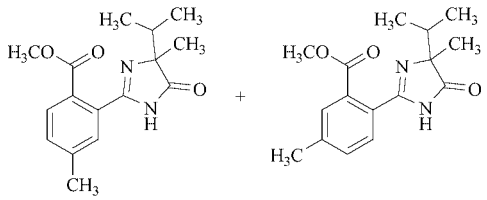

Imazamethabenz-methyl is produced by reaction of 3,4-(4,5)-dicarboxymethyltoluene with 2-aminoamide-3-methylbutane in the presence of two equivalents of potassium *tert*-butoxide [100].

Imazamethabenz-methyl is a selective, systemic herbicide which is effective post-emergence against some broad-leaved and some dicotyledonous weeds in wheat, barley, rye, and sunflowers.

Trade Names. Assert, Dagger (Cyanamid).

Imazamox [*114311-32-9*], (*R*,*S*)-2-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)-5-methoxymethylnicotinic acid, $C_{15}H_{19}N_3O_4$, M_r 305.3, *mp* 166.0 – 166.7 °C, forms colorless crystals which are fairly soluble in acetone [23, p. 696].

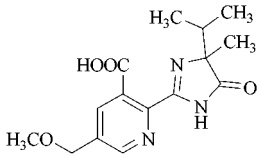

Imazamox is produced by reaction of 2,3-dicarboxymethyl-5-methoxymethylpyridine with 2-aminoamide-3-methylbutane in the presence of two equivalents of potassium *tert*-butoxide [100].

Imazamox is a selective herbicide under development which is to be used post-emergence for weed control in soya beans and other vegetables grown in rotation with sugar beet and other crops.

Trade Names. Odyssey, , Sweeper (Cyanamid).

Imazapic (AC 263 222) [*104098-48-8*], (*R*,*S*)-2-(4,5-dihydro-4-isopropyl-4-methyl-5-oxoimidazol-2-yl)-5-methylnicotinic acid, $C_{14}H_{17}N_3O_3$, M_r 275.3, *mp* 204 – 206 °C, is a tan powder which is moderately soluble in water and acetone [23, p. 5].

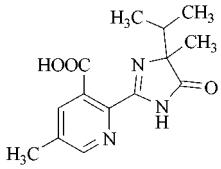

Imazapic is produced by reaction of 2,3-dicarboxymethyl-5-methylpyridine with 2-aminoamide-3-methylbutane in the presence of two equivalents of *tert*-butoxide [101].

Imazapic is a herbicide used early post-emergence for the control of a wide range of weeds in peanuts.

Trade Name. Cadre (Cyanamid).

Imazapyr [*81334-34-1*], 2-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)nicotinic acid, $C_{13}H_{15}N_3O_3$, M_r 261.3, *mp* 169 – 173 °C, is a colorless to tan powder which is fairly soluble in water, acetone, methanol, dichloromethane, and dimethylformamide, slightly soluble in benzene, and moderately soluble in toluene [23, p. 697].

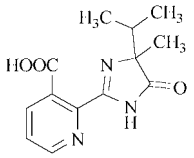

Imazapyr is produced by reaction of 2,3-dicarboxymethylpyridine with 2-aminoamide-3-methylbutane in the presence of two equivalents of *tert*-butoxide [102].

Imazapyr is a nonselective, systemic herbicide used pre- and post-emergence for the control of annual and perennial grasses, sledges, broad-leaved weeds, many bushes, and deciduous tree species on noncrop areas.

Imazaquin [*81335-37-7*], (*R*,*S*)-2-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)quinoline-2-carboxylic acid, $C_{17}H_{17}N_3O_3$, M_r 311.3, *mp* 219 – 224 °C (decomp.), is a tan solid which is moderately soluble in water and toluene, and is fairly soluble in dimethylformamide, dimethyl sulfoxide, and dichloromethane [23, p. 699].

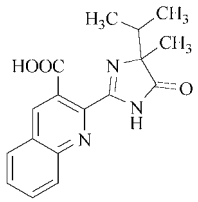

Imazaquin is produced by reaction of 2,3-dicarboxymethylquiniline with 2-aminoamide-3-methylbutane in the presence of two equivalents of *tert*-butoxide [102].

Imazaquin is a selective, systemic herbicide used pre-planting, pre-emergence, and early post-emergence to control broad-leaved weeds in soya beans.

Trade Name. Scepter (Cyanamid).

Imazethapyr [*81335-77-5*], (*R*,*S*)-5-ethyl-2-(4-isopropyl-4-methyl-5-oxo-2-imidazolin-2-yl)nicotinic acid, $C_{15}H_{19}N_3O_3$, M_r 289.2, *mp* 169 – 173 °C, forms colorless crystals which are fairly soluble in water, toluene and heptane, and are readily soluble in acetone, methanol, dichloromethane, and dimethyl sulfoxide [23, p. 701].

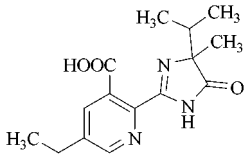

Imazethapyr is produced by reaction of 2,3-dicarboxymethyl-5-ethylpyridine with 2-aminoamide-3-methylbutane in the presence of two equivalents of *tert*-butoxide [102].

Imazethapyr is a systemic herbicide used pre-plant-incorporated, pre- and post-emergence for the control of many annual and perennial grasses and broad-leaved weeds in soya beans and other leguminous crops.

Trade Names. Hammer, Overtop, Pivot, Pursuit (Cyanamid).

7.8.8. Alkanoic Acids

Clomeprop [*84496-56-0*], (*R*,*S*)-2-(2,4-dichloro-*m*-tolyloxy)propionanilide, $C_{16}H_{15}Cl_2NO_2$, M_r 324.2, *mp* 146 – 147 °C, consists of colorless crystals which are sparingly soluble in water and are fairly soluble in xylene, acetone, cyclohexane, and dimethylformamide [23, p. 258].

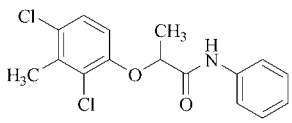

Clomeprop is produced by reaction of the potassium salt of 2,4-dichloro-3-methylphenol with 2-chloropropionanilide [103].

Clomeprop is an auxin-type herbicide used pre- to early post-emergence to control broad-leaved weeds and cyperaceous weeds in paddy rice.

Trade Name. Yukahope (Mitsubishi).

2,4-D [*94-75-7*], (2,4-di-chlorophenoxy)acetic acid, $C_8H_6Cl_2O_3$, M_r 221.0, *mp* 140.5 °C, forms colorless crystals which are moderately soluble in water, heptane, toluene, and xylene, and are readily soluble in ethanol and diethyl ether [23, p. 323].

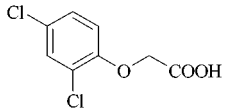

2,4-D is produced by interaction of 2,4-dichlorophenolate with monochloroacetic acid [104].

2,4-D is a selective, systemic herbicide, it acts as a growth inhibitor, and is used post-emergence for the control of annual and perennial weeds in many crops.

Trade Names. Ded-Weed (Uniroyal); Desormone (Rhône-Poulec); U46D (BASF); Weedtox (Aimco), etc.

Dalapon [*75-99-0*], 2,2-dichloropropionic acid, $C_3H_4Cl_2O_2$, M_r 143.0, *mp* 185 – 190 °C, is a colorless liquid. Its sodium salt is readily soluble in water, fairly soluble in ethanol and methanol, and moderately soluble in acetone [23, p. 331].

$CH_3CCl_2CO_2H$

Dalapon is produced by chlorination of propionic acid [105].

Dalapon is a selective, systemic herbicide, used as its sodium salt to control annual and perennial grasses on noncropland and in orchards.

Trade Names. Dalacide (Diachem); Diserbo Canali (Siapa), etc.

2,4-DB [*94-82-6*], 4-(2,4-dichlorophenoxy)butyric acid, $C_{10}H_{10}Cl_2O_3$, M_r 249.1, *mp* 117 – 119 °C, consists of colorless crystals which are moderately soluble in water, are readily soluble in acetone, ethanol, and diethyl ether, and are slightly soluble in benzene and toluene [23, p. 337].

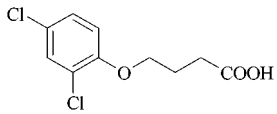

2,4-DB is produced by reaction of 2,4-dichlorophenol with trimethylene chlorobromide, followed by reaction with sodium cyanide and hydrolysis of the resulting cyano compound to the acid [106].

2,4-DB is a selective, systemic, hormone-type herbicide used post-emergence for the control of many annual and perennial broad-leaved weeds in alfalfa, clovers, under-sown cereals, grassland, etc.

Trade Names. Embutox (Rhône-Poulenc); Butormone (Unicrop); Butoxone (Cedar), etc.

Dichlorprop [*7547-66-2*], (*R*,*S*)-2-(2,4-dichlorophenoxy)propionic acid, $C_9H_8Cl_2O_3$, M_r 235.1, *mp* 116 – 117.5 °C, forms colorless crystals which are moderately soluble in water and are readily soluble in acetone, isopropyl alcohol, benzene, toluene, and xylene [23, p. 368].

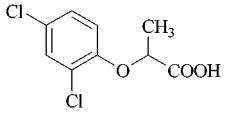

Dichlorprop is produced by reaction of 2-chloropropionic acid with 2,4-dichlorophenol [107].

Dichlorprop is a selective, systemic, hormone-type herbicide used post-emergence to control annual and perennial broad-leaved weeds in cereals and grassland, as well as for brush control on noncropland.

Trade Names. Weedone (Rhône-Poulenc); U46 DPC (BASF); Polymone (Unicrop), etc.

Dichlorprop-P [*15165-67-0*], (*R*)-2-(2,4-dichlorophenoxy)propionic acid, $C_9H_8Cl_2O_3$, M_r 235.1, *mp* 121 – 123 °C, forms colorless crystals which are moderately soluble in water, are readily soluble in acetone, ethanol, and ethyl acetate, and are fairly soluble in toluene [23, p. 370].

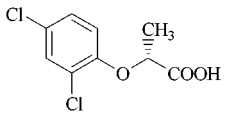

Dichlorprop-P is produced analogously to dichlorprop, but from the (+)-stereoisomer of propionic acid, obtained by fermentation [108].

Dichloprop-P is a selective, systemic, hormone-type herbicide used post-emergence for the control of various weed species.

Trade Name. (BASF).

Flupropanate [*756-09-2*], 2,2,3,3-tetrafluoropropionic acid [*22898-01-7*]. Its sodium salt, $C_3HF_4NaO_2$, M_r 168.0, *mp* 165 – 167 °C, consists of colorless crystals which are fairly soluble in water and sparingly soluble in nonpolar organic solvents [23, p. 585].

$F_2CHCF_2CO_2H$

Flupropanate is produced by oxidation of 2,2,3,3-tetrafluoropropanol with chromium trioxide [109].

Flupropanate is a systemic herbicide with low contact activity, which is used for the control of annual and perennial grasses in pastures and in uncultivated land.

Trade Name. Frenock (Daikin, Sankyo, Hodogaya, Zeneca, Sumitomo).

Fluroxypyr [*69377-81-7*], 4-amino-3,5-dichloro-6-fluoro-2-pyridyloxyacetic acid, $C_7H_5Cl_2FN_2O_3$, M_r 255.0, *mp* 232 – 233 °C, forms a colorless crystalline solid which is moderately soluble in water, dichloromethane, toluene, and xylene, and is fairly soluble in acetone, methanol, and diethyl ether [23, p. 597].

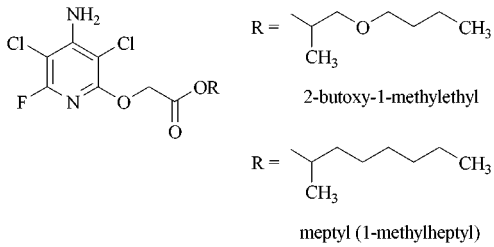

Fluroxypyr is produced by reaction of 4-amino-3,5-dichloro-2-(methylsulfonyl)pyridine with methylglycolate [110].

Fluroxypyr is applied as fluroxypyr-methyl [*81406-37-3*] and produces characteristic auxin-type responses. It is used post-emergence to control a number of broad-leaved weeds in all small grains, in orchards, and in plantations.

Trade Names. Starane (DowElanco); Herber (Barclay).

MCPA [*94-74-6*], (4-chloro-2-methylphenoxy)acetic acid, $C_9H_9ClO_3$, M_r 200.6, *mp* 119 – 120.5 °C, consists of colorless crystals which are moderately soluble in water, are fairly soluble in toluene, xylene, and heptane, and are readily soluble in ethanol and diethyl ether [23, p. 767].

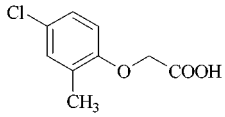

MCPA is produced by chlorination of *o*-cresol and coupling with monochloroacetic acid [111].

MCPA is a selective, systemic, hormone-like herbicide used post-emergence for the control of annual and perennial broad-leaved weeds in cereals, vegetables, grassland turf, roadside verges, embankments, forestry, etc.

Trade Names. Agrotone (Marks); Meadowman (Barclay); Empel (Unicrop); Sanphen-M (Sanachem), etc.

MCPA-thioethyl [*25319-90-8*], *S*-ethyl 4-chloro-*o*-tolyloxythioacetate, $C_{11}H_{13}ClO_2S$, M_r 244.7, *mp* 41 – 42 °C, forms colorless needles which are moderately soluble in water and readily soluble in acetone, xylene, and hexane [23, p. 770].

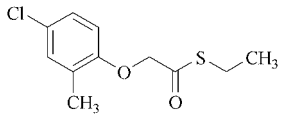

MCPA-thioethyl is produced by reaction of ethylmercaptan with 4-chloro-2-methylphenoxymethylcarboxychloride [112].

MCPA-thioethyl is a selective, systemic, hormonelike herbicide used post-emergence to control annual and perennial broad-leaved and cyperaceous weeds in orchards, paddy rice, and cereals.

Trade Names. Fenobit, Herbit, Zero One (Hokko).

MCPB [*94-81-5*], 4-(4-chloro-*o*-tolyloxy)butyric acid, $C_{11}H_{13}ClO_3$, M_r 228.7, *mp* 100 °C, consists of colorless crystals which are moderately soluble in water, fairly soluble in toluene, and readily soluble in acetone, dichloromethane, ethanol, and hexane [23, p. 771].

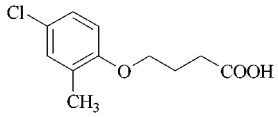

MCPB is produced by chlorination of *o*-cresol and coupling with γ-chlorobutyric acid [113].

MCPB is used as a sodium salt [*6062-26-6*] and is a selective, systemic, hormone-type herbicide used post-emergence for the control of annual and perennial broad-leaved weeds in cereals, clovers, peas, peanuts, grasslands, and forestry.

Trade Name. Tropotox (Unicrop).

Mecoprop [*7085-19-0*], (*RS*)-2-(4-chlorotextito-tolyloxy)propionic acid, $C_{10}H_{11}ClO_3$, M_r 214.6, *mp* 94 – 95 °C, forms colorless crystals which are moderately soluble in water and readily soluble in acetone, diethyl ether, ethanol, ethyl acetate, and chloroform [23, p. 776].

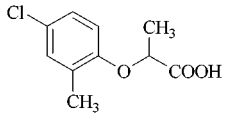

Mecoprop is produced by chlorination of 4-chloro-*o*-cresol and coupling with 2-chloropropionic acid [113].

Mecoprop is a selective, systemic, hormone-like herbicide used post-emergence for the control of broad-leaved weeds in wheat, barley, oats, seed crops, grassland, etc.

Trade Names. Actril (Rhône-Poulenc); Propal (Unicrop); Propionyl (Agriphar), etc.

Mecoprop-P [*16484-77-8*], (*R*)-2-(4-chloro-*o*-tolyloxy)propionic acid, $C_{10}H_{11}ClO_3$, M_r 214.6, *mp* 94.6 – 96.2 °C, forms colorless crystals which are moderately soluble in water, are fairly soluble in hexane, and are readily soluble in acetone, diethyl ether, ethanol, dichloromethane, and toluene [23, p. 778].

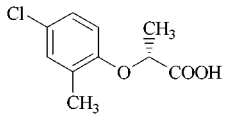

Mecoprop-P is produced analogously to Mecoprop-P from the corresponding propionic acid stereoisomer [114].

Mecoprop-P is a selective, systemic, hormonelike herbicide used post-emergence for the control of broad-leaved weeds in wheat, barley, oats, seed crops, grassland, etc.

Trade Names. Duplosan KV (BASF); Optica (Marks).

Triclopyr [*55335-06-3*], 3,5,6-trichloro-2-pyridylacetic acid, $C_7H_4Cl_3NO_3$, M_r 256.5, mp. 150.5 °C, is a fluffy, colorless solid which is fairly soluble in water, hexane, toluene, and dichloromethane and is readily soluble in acetone, acetonitrile, methanol, and ethyl acetate [23, p. 1237].

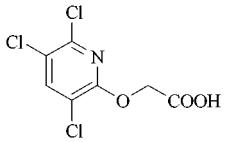

Triclopyr is produced by reaction of 2,3,5,6-tetrachloropyridine with formaldehyde in dimethyl sulfoxide containing potassium cyanide, followed by hydrolysis of the cyano compound to the acid [115].

Triclopyr is a selective, systemic herbicide used for the control of woody plants, many broad-leaved weeds in grassland, uncultivated land, industrial areas, coniferous foresst, plantation crops, and rice fields.

Trade Name. Trident (Agriphar).

7.8.9. Triazoles

Amitrole [*61-82-5*], 1*H*-1,2,4-triazol-3-ylamine, $C_2H_4N_4$, M_r 84.1, *mp* 157 – 159 °C, consists of colorless crystals which are readily soluble in water, are moderately soluble in methanol, acetonitrile, and chloroform, are sparingly soluble in ethyl acetate, and are practically insoluble in ethyl acetate, diethyl ether, acetone, and nonpolar solvents [23, p. 40].

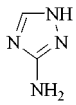

Amitrole is produced by condensation of formic acid with aminoguanidine [116].

Amitrole is a nonselective herbicide which inhibits the biosynthesis of chlorophyll and regrowth from buds. It is used for the total control of annual and perennial grasses and broad-leaved weeds around established fruit trees and bushes in orchards, in vineyards, and in many other applications.
Trade Names. Weedazal (Rhône-Poulenc, CFPI); Superzol (Productos OSA); Oriflam (CFPI), etc.

Azafenidin [*68049-83-2*], 2-(2,4-dichloro-5-prop-2-ynyloxyphenyl)-5,6,7,8-tetrahydro-1,2,4-triazolo[4,3-*a*]pyridin-3(2*H*)-one, $C_{15}H_{13}Cl_2N_3O_2$, M_r 338.2, *mp* 168 – 168.5 °C, forms a rust-colored solid which is sparingly soluble in water [23, p. 60].

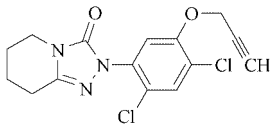

Azafenidin is produced by reaction of *N*-(2-piperidyl)-*N′*-(2,4-dichloro-5-oxy-methylacetylene)hydrazone with methylchloroformate [117].

Azafenidin is an inhibitor of the biosynthesis of porphyrin. It is under development and will be used pre-emergence in citrus, grapes, sugar cane, and other perennial crops.
*Trade Name.*To be announced (Du Pont).

Chloransulam-methyl [*147150-35-4*], (its acid: [*159518-97-5*]), methyl 3-chloro-2-(5-ethoxy-7-fluoro[1,2,4]triazolo[1,5-*c*]pyrimidin-2-ylsulfonamido)benzoate, $C_{15}H_{13}ClFN_5O_5S$, M_r 429.8 [23, p. 265].

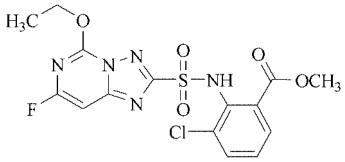

Chloransulam-methyl is produced by condensation of 5-oxyethyl-7-fluoro-1,2,4-triazolo[1,5-*c*]pyrimidine-2-sulfonyl chloride with 2-carboxymethyl-6-chloroaniline [118].

Chloransulam-methyl is a acetolactate-synthase inhibitor, and is used pre- and post-emergence for the control of broad-leaved weeds in soya beans.
Trade Name. First Rate (DowElanco).

7.8.10. Uracils

Bromacil [*314-40-9*], 5-bromo-3-*sec*-butyl-6-methyluracil, $C_9H_{13}BrN_2O_2$, M_r 261.1, *mp* 157.5 – 160 °C, consists of colorless crystals which are moderately soluble in water and are fairly soluble in ethanol, acetone, and xylene [23, p. 140].

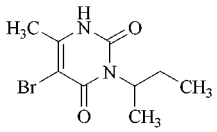

Bromacil is produced by reaction of isopropyl isocyanate with methyl 3-aminocrotonate, followed by cyclization in sodium hydroxide at elevated temperature, and bromination in acetic acid [119].

Bromacil is used for the total control of weeds and brushes on noncropland, and also for the selective control of annual and perennial weeds and grasses in citrus and pineapple plantations.
Trade Names. Hyvar X, Hyvar X-L (Du Pont); Rokar XL (Siapa); Uragan (Makhteshim-Agan), etc.

Lenacil [*2164-08-1*], 3-cyclohexyl-1,5,6,7-tetrahydrocyclopentapyrimidine-2,4-(3*H*)-dione, $C_{13}H_{18}N_2O_8$, M_r 234.3, *mp* 315.6 – 316.8 °C, forms colorless to tan crystals which are slightly soluble in water and readily soluble in most organic solvents [23, p. 748].

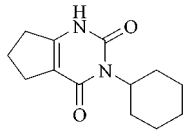

Lenacil is produced by condensation of cyclohexylurea and ethyl 2-oxocyclopentane-1-carboxylate with phosphoric acid [120].

Lenacil is a selective, systemic herbicide used for the control of annual grasses and broad-leaved weeds in sugar cane, beet, fodder beet, sweet potatoes, etc., and in ornamental plants and shrubs.
Trade Names. Venzar (Du Pont); Adol (Gedeon Richter).

Terbacil [*5902-51-2*], 3-*tert*-butyl-5-chloro-6-methyluracil, $C_9H_{13}ClN_2O_2$, M_r 216.7, *mp* 175 – 177 °C, consists of colorless crystals which are moderately soluble in water and readily soluble in dimethylformamide, cyclohexanone, methyl isobutyl ketone, and butyl acetate [23, p. 1163].

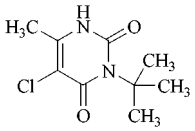

Terbacil is produced by reaction of *tert*-butyl isocyanate with methyl 3-aminocrotonate, followed by cyclization and chlorination [120].

Terbacil is a selective herbicide used for the control of many annual and some perennial weeds in apples, citrus, peaches, and other crops.
Trade Names. Sinbar (Du Pont); Geonter (Chemol).

7.8.11. Pyridines

Clopyralid [*1702-17-6*], 3,6-dichloropyridine-2-carboxylic acid, $C_6H_3Cl_2NO_2$, M_r 192.0, *mp* 151 – 152 °C, forms colorless crystals which are fairly soluble in water, acetonitrile, methanol, and *n*-hexane [23, p. 260].

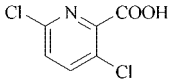

Clopyralid is produced by photochlorination of picoline, followed by further photochlorination and oxidation [121].

Clopyralid is a selective, systemic herbicide with auxin-type activity, used post-emergence for the control of many annual and perennial broad-leaved weeds in sugar beet, fodder beet, oilseed rape, maize, etc.
Trade Names. Cyranal (AgrEvo); Clio (Agriphar).

Diflufenican [*83164-33-4*], 2′,4′-difluoro-2-(α,α,α-trifluoro-*m*-tolyloxy)nicotinanilide, $C_{19}H_{11}F_7N_2O_2$, M_r 394.3, *mp* 159 – 161 °C, forms colorless crystals which are practically insoluble in water, but are soluble in most organic solvents [23, p. 397].

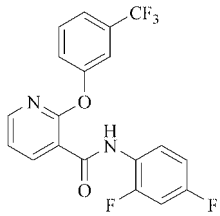

Diflufenican is produced by reaction of 2-(3-trifluoromethylphenoxy)nicotinic acid with 2,4-difluoroaniline and sulfonyl chloride [122].

Diflufenican is a carotenoid- and photosynthesis electron-flow inhibitor. It is a selective contact and residual herbicide used pre- and early post-emergence to control grass and broad-leaved weeds in autumn-sown wheat and barley.
Trade Names. Fenican, Javelin (Rhône-Poulenc).

Diflufenzopyr (BAS 654 00H) [*109293-97-2*], 2-{1-[4-(3,5-difluorophenyl)semicarbazono]ethyl}nicotinic acid, $C_{15}H_{12}F_2N_4O_3$, M_r 334.3, *mp* 155 °C, is a tan to colorless solid which is sparingly soluble in water [23, p. 81].

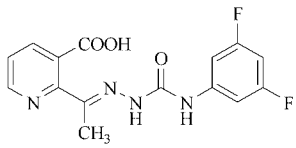

Diflufenzopyr is produced by reaction of 2-acetylnicotinic acid with 3,5-difluorobenzene-*N*-2-methylsemicarbazone in the presence of pyridyl tosylate [123].

Diflufenzopyr is a polar auxin-transport inhibitor. This systemic herbicide is used post-emergence to control annual broad-leaved and perennial weeds in maize. An initial commercialization of a mixture with dicamba is planned.
Trade Name. To be announced (BASF).

Dithiopyr [*97886-45-8*], *S,S′*-dimethyl-2-difluoromethyl-4-isobutyl-6-trifluoromethylpyrimidine-3,5-dicarbothioate, $C_{15}H_{16}F_5NO_2S_2$, M_r 401.4, *mp* 65 °C, consists of colorless crystals which are sparingly soluble in water [23, p. 442].

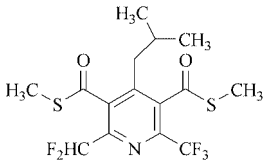

Dithiopyr is produced by dehydration of diethyl 2,6-dihydro-2-difluoromethyl-6-trifluoromethyl-4-(2-methylisopropyl) piperidine dicarboxylate, followed by hydrolysis of the pyridine dicarboxylate, and by treatment with sulfonyl chloride and methanol [124].

Dithiopyr is a cell-division inhibitor used pre- and early post-emergence to control annual grasses and broad-leaved weeds in turf, and annual grasses in rice.

Trade Names. Dictran, Dimension (Rohm & Haas).

Picloram [*1918-02-1*], 4-amino-3,5,6-trichloropyridine-2-carboxylic acid, $C_6H_3Cl_3N_2O_2$, M_r 241.5, *mp* 215 °C (decomp.), is a colorless powder which is moderately soluble in water, dichloromethane, benzene, and carbon disulfide, and is fairly soluble in acetone, ethanol, and diethyl ether [23, p. 977].

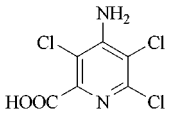

Picloram is produced from α-picoline by chlorination, followed by hydrolysis of the α-trichloropicoline derivative, and amidation [125].

Picloram is a selective, systemic herbicide used to control most annual and perennial broad-leaved weeds on grassland and noncrop areas.

Trade Name. Tordon (DowElanco).

Thiazopyr [*117718-60-2*], methyl 2-difluoromethyl-5-(4,5-dihydro-1,3-thiazol-2-yl)-4-isobutyl-6-trifluoromethylnicotinate, $C_{16}H_{17}F_5N_2O_2S$, M_r 396.4, *mp* 77.3 – 79.1, is a light tan crystalline solid which is moderately soluble in water, and readily soluble in acetone and hexane [23, p. 1185].

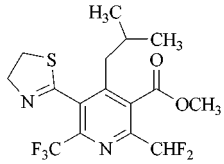

Thiazopyr is produced by amination of 3-chlorocarbonyl-5-carboxymethyl-4-(2-methylpropyl)-2-trifluoromethyl-6-difluoromethylpyridine with ethanolamine, followed by treatment with phosphorus oxytrichloride, and cyclization [126].

Thiazopyr inhibits cell division, and is used pre-emergence for the control of annual grasses and some broad-leaved weeds in tree fruit, vines, citrus, sugar cane, and pineapples.

Trade Name. Visor (Rohm & Haas).

7.8.12. Hydroxybenzonitriles

Bromofenoxim [*13181-17-4*], 3,5-dibromo-4-hydroxybenzaldehyde 2,4-dinitrophenyloxime, $C_{13}H_7Br_2N_3O_6$, M_r 461.0, *mp* 196 – 197 °C, forms cream-colored crystals which are sparingly soluble in water, are moderately soluble in hexane, isopropyl alcohol, and *n*-octanol, and are fairly soluble in acetone [23, p. 145].

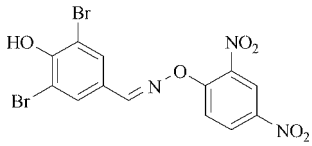

Bromofenoxim is produced by reaction of 3,5-dibromoanisaldehyde oxime with 2,4-dinitrochlorobenzene in the presence of sodium ethoxide [127].

Bromofenoxim, a precursor of bromoxynil, is a selective contact herbicide used post-emergence to control many annual broad-leaved weeds in cereals, maize, grass-seed crops, on grassland, and newly seeded turf.

Trade Name. Faneron (Novartis).

Bromoxynil [*1689-84-5*], 3,5-dibromo-4-hydroxybenzonitrile, $C_7H_3Br_2NO_2$, M_r 276.9, *mp* 194 – 195 °C, forms a colorless solid which is sparingly soluble in water, is fairly soluble in methanol, ethanol, and benzene, and is readily soluble in dimethylformamide, tetrahydrofuran, and acetone [23, p. 149].

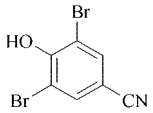

Bromoxynil is produced by reaction of sodium hypobromide with 4-hydroxybenzonitrile [128].

Bromoxynil is a selective contact herbicide used post-emergence to control annual broad-leaved weeds in cereals, maize, sorghum, onions, garlic, mint, turf, noncropland, etc.

Trade Names. Brominal, Buctril (Rhône-Poulenc); Bromox, Bromoxan (Sanachem); Kaleis (Novartis).

7.8.13. (Thio)Carbamates

Asulam [*3337-71-1*], (sodium salt: [*2302-17-2*]), sulfanilcarbamate, $C_8H_{10}N_2O_4S$, M_r 230.2, *mp* 142 – 144 °C, consists of colorless crystals which are fairly soluble in water, hydrocarbons, and chlorinated hydrocarbons, and are readily soluble in acetone, methanol, methyl ethyl ketone, and dimethylformamide [23, p. 53].

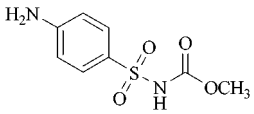

Asulam is produced by reaction of 4-aminobenzenesulfonamide, after protection of the amino group, with ethyl chloroformate in the presence of a base [129].

Asulam, a cell-division inhibitor, is a selective, systemic herbicide used for the control of annual and perennial grasses and broad-leaved weeds in a variety of crops, grassland, noncropland, and forestry.

Trade Names. Asulox (Rhône-Poulenc); Asilan (Shionogi); Sanulam (Sanchem), etc.

Butylate [*2008-41-5*], *S*-ethyl di-isobutylthiocarbamate, $C_{11}H_{23}NOS$, M_r 217.4, *bp* 137.5 – 138 °C/21 mmHg, is a colorless liquid which is moderately soluble in water and is miscible with common organic solvents [23, p. 170].

$[(CH_3)_2CHCH_2]_2NC(O)SCH_2CH_3$

Butylate is produced by reaction of diisobutylcarbamoyl chloride with ethylmercaptan [130].

Butylate, which inhibits lipid metabolism, is a selective, systemic herbicide used pre-planting and soil-incorporated for the control of annual grass weeds in maize, pineapples, etc.

Trade Names. Sutan Plus (Zeneca); Anelda (Sagrochem), etc.

Carbetamide [*16118-49-3*], (*R*)-1-(ethylcarbamoyl)ethyl carbanilate, $C_{12}H_{16}N_2O_3$, M_r 236.3, *mp* 119 °C, consists of colorless crystals which are fairly soluble in water, moderately soluble in cyclohexane, and readily soluble in acetone, ethanol, methanol, and dimethylformamide [23, p. 184].

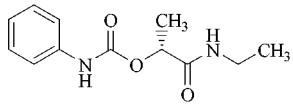

Carbetamide is produced by condensation of phenyl isocyanate with *N*-ethyl lactamide [131].

Carbetamide, a cell-division inhibitor, is a selective herbicide used to control annual grasses and some broad-leaved weeds in clover, some vegetables, oilseed rape, vines, and fruit orchards.

Trade Name. Leguram (Rhône-Poulenc).

Chlorpropham [*101-21-3*], isopropyl 3-chlorocarbanilate, $C_{10}H_{12}ClNO_2$, M_r 123.7, *mp* 41.4 °C, is a colorless solid which is moderately soluble in water and readily soluble in most organic solvents [23, p. 233].

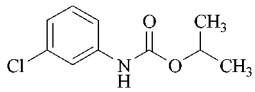

Chlorpropham is produced by reaction of *m*-chloroaniline with isopropyl chloroformate [132].

Chlorpropham, a cell-division inhibitor, is a selective herbicide used pre-emergence for the control of annual grasses and some broad-leaved weeds in many crops.
Trade Names. Decco Aerosol 273 (Elf Atochem); Prevenol (AgrEvo), etc.

Cycloate [*1134-23-2*], *S*-ethyl cyclohexyl(ethyl)thiocarbamate, $C_{11}H_{21}NOS$, M_r 215.4, *bp* 145 – 146 °C/10 mmHg, is a colorless liquid which is moderately soluble in water and miscible with most organic solvents [23, p. 285].

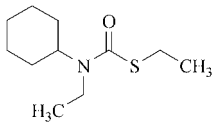

Cycloate is produced by reaction of *N*-cyclohexyl-*N*-ethylcarbamoyl chloride with ethylmercaptan [133].

Cycloate is a lipid-metabolism inhibitor which is used pre-planting and soil-incorporated for the control of annual grass weeds and some perennial grasses and broad-leaved weeds in sugar beet, fodder beet, spinach, etc.
Trade Names. Ro-Neet (Zeneca); Sabet (Sanachem), etc.

Desmedipham [*13684-56-5*], ethyl 3-phenyl- carbamoyloxyphenylcarbamate, $C_{16}H_{16}N_2O_4$, M_r 300.3, *mp* 120 °C, forms colorless crystals which are moderately soluble in water and readily soluble in polar organic solvents [23, p. 349].

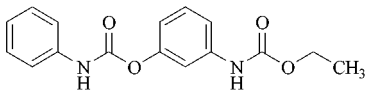

Desmedipham is produced by reaction of *m*-aminophenol with ethyl chloroformate, followed by treatment with phenyl isocyanate [134].

Desmedipham, a photosynthesis electron-transport inhibitor, is a selective herbicide used post-emergence to control broad-leaved weeds in beet crops, etc.
Trade Names. Betanal AM, Kemifam (AgrEvo), etc.

Dimepiperate [*61432-55-1*], *S*-(1-methyl-1-phenylethyl)piperidine-1-carbothioate, $C_{15}H_{21}NOS$, M_r 263.4, *mp* 38.8 – 39.3 °C, is a waxlike solid which is moderately soluble in water and fairly soluble in acetone, chloroform, cyclohexanone, ethanol, and hexane [23, p. 404].

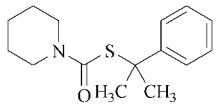

Dimepiperate is produced by reaction of benzene isopropylmercaptan with piperidine carbamoyl chloride [133].

Dimepiperate, a lipid-metabolism inhibitor, is a systemic herbicide used for the control of *Echinochloa crusgalli* in flooded rice.
Trade Name. Yakamate (Mitsubishi).

EPTC [*759-94-4*], *S*-ethyl dipropylthiocarbamate, $C_9H_{19}NOS$, M_r 189.3, *bp* 127 °C/20 mmHg, is a colorless liquid which is moderately soluble in water and is miscible with common organic solvents [23, p. 466].

$[CH_3(CH_2)_2]_2NC(O)SCH_2CH_3$

EPTC is produced by reaction of di-*n*-propylamine with phosgene, followed by treatment with ethylmercaptan [135].

EPTC, a lipid-metabolism inhibitor, is a selective herbicide used pre-planting with soil-incorporation for the control of annual and perennial grasses and some broad-leaved weeds in many crops.
Trade Names. Eptam, Eradicane (Zeneca); Alirox (Sagrochem), etc.

Esprocarb [*85785-20-2*], *S*-benzyl 1,2-dimethylpropyl(ethyl)thiocarbamate, $C_{15}H_{23}NOS$, M_r 265.4, *bp* 135 °C/35 mmHg, is a liquid which is moderately soluble in water and is fairly soluble in ethanol, chloroform, xylene, acetone, and acetonitrile [23, p. 472].

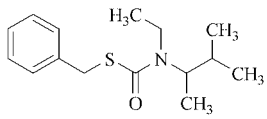

Esprocarb is produced by reaction of *N*-ethyl-*N*-(methyl)isobutylcarbamoyl chloride with mercaptotoluene [133].

Esprocarb is a lipid-metabolism inhibitor, which is used pre- and post-emergence for the control of annual weeds and *Echinocloa* sp. in paddy rice.
Trade Names. Fujigrass (Du Pont, Zeneca); Sparkstar (Nissan, Zeneca), etc.

Molinate [*2212-67-1*], *S*-ethyl azepane-1-carbothioate, $C_9H_{17}NOS$, M_r 187.3, *bp* 136.5 °C/10 mmHg, is a colorless liquid which is moderately soluble in water and is miscible with most organic solvents [23, p. 847].

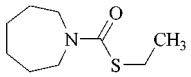

Molinate is produced by reaction of chlorothiolformate with hexamethyleneimine [136].

Molinate, a lipid-metabolism inhibitor, is a selective, systemic herbicide used to control germinating broad-leaved weeds and grasses, and is applied either before planting, to water-seeded-, to shallow-soil-seeded-, or to post-flood rice.
Trade Names. Ordram (Zeneca); Sakkimol (Sagrochem), etc.

Orbencarb [*34622-58-7*], *S*-2-chlorobenzyl diethylthiocarbamate, $C_{12}H_{16}ClNOS$, M_r 257.8, *mp* 9.0 °C, *bp* 158 °C/1 mmHg, is a colorless liquid which is moderately soluble in water and readily soluble in most organic solvents [23, p. 898].

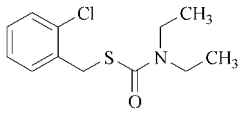

Orbencarb is produced by reaction of 2-chloro-mercaptotoluene with *N*-diethylcarbamoyl chloride [133].

Orbencarb is a selective herbicide used pre-emergence to control broad-leaved weeds in barley, wheat, rye, maize, soya beans, cotton, and turf.
Trade Name. Lanray (Kumiai).

Pebulate [*1114-71-2*], *S*-propyl butyl(ethyl)thiocarbamate, $C_{10}H_{21}NOS$, M_r 203.3, *bp* 142 °C/21 mmHg, is a colorless to yellow liquid which is moderately soluble in water and is miscible with most organic solvents [23, p. 930].

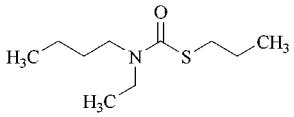

Pebulate is produced by reaction of ethylbutylamine with phosgene, followed by treatment with propylmercaptan [137].

Pebulate, an inhibitor of the biosynthesis of very long chain fatty acids, is a selective, systemic herbicide used pre-emergence and soil-incorporated to control annual grasses and some broad-leaved weeds in sugar beet, tomatoes, etc.
Trade Name. Tillam (Zeneca).

Phenmedipham [*13684-63-4*], methyl 3-(3-methylcarbaniloyloxy)carbamate, $C_{16}H_{16}N_2O_4$, M_r 300.3, *mp* 143 – 144 °C, forms colorless crystals which are moderately soluble in water and fairly soluble in polar organic solvents [23, p. 948].

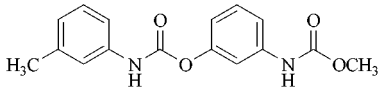

Phenmedipham is produced by reaction of 3-aminotoluene with 3-methylcarbamic benzoyl chloride [134].

Phenmedipham, a photosynthesis electron-transport inhibitor, is a selective, systemic herbicide used post-emergence for the control of most annual broad-leaved weeds in beet crops, strawberries, peas, spinach, etc.
Trade Names. Betanal (AgrEvo); Alegro (Rhône-Poulenc); Beetup (United Phosphorus), etc.

Prosulfocarb [*52888-80-9*], *S*-benzyl dipropylthiocarbamate, $C_{14}H_{21}NOS$, M_r 251.4, *bp* 129 °C/33 Pa, is a colorless liquid which is moderately soluble in water and is miscible with acetone, chloroform, chlorobenzene, ethanol, xylene, and ethyl acetate [23, p. 1039].

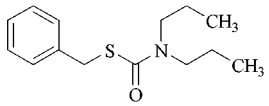

Prosulfocarb is produced by reaction of *N*-dipropylcarbamoyl chloride with mercaptotoluene [133].

Prosulfocarb is a lipid-metabolism inhibitor, and is a selective herbicide used pre- and early post-emergence for the control of a wide range of grass and broad-leaved weeds in winter wheat, winter barley, and rye.
Trade Names. Boxer, Defi (Zeneca).

Pyributicarb [*88678-67-5*], *O*-3-*tert*-butylphenyl 6-methoxy-2-pyridyl(methyl)thiocarbamate, $C_{18}H_{22}N_2O_2S$, M_r 330.4, *mp* 85.7 – 86.2 °C, forms colorless crystals which are sparingly soluble in water, fairly soluble in methanol and ethanol, and readily soluble in acetone, xylene, chloroform, and ethyl acetate [23, p. 1060].

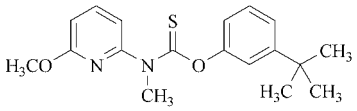

Pyributicarb is produced by reaction of 3-*tert*-butylphenol with 6-methoxy-2-*N*-methyl-*N*-thionyl chloride [138].

Pyributicarb is a systemic herbicide used pre- and early post-emergence for the control of annual and perennial grasses and weeds in turf.
Trade Names. Eigen, Isshintasuke (Dainippon).

Thiobencarb [*28249-77-6*], *S*-4-chlorobenzyl diethylthiocarbamate, $C_{12}H_{16}ClNOS$, M_r 257.8, *bp* 126 – 129 °C/0.008 mmHg, is a pale yellow to brownish yellow liquid which is moderately soluble in water and is readily soluble in acetone, ethanol, xylene, methanol, benzene, and acetonitrile [23, p. 1192].

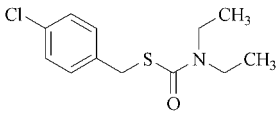

Thiobencarb is produced by reaction of *p*-chloromercaptotoluene with *N*-diethylcarbamoyl chloride [133].

Thiobencarb is a protein-synthesis inhibitor, and is a selective herbicide used pre- and early post-emergence for the control of a number of grasses and weeds in rice.
Trade Names. Saturn (Kumiai); Bigturn (Sanonda); Bolero (Valent), etc.

Triallate [*2303-17-5*], *S*-2,3,3-trichloroallyl diisopropyl(thiocarbamate), $C_{10}H_{16}Cl_3NOS$, M_r 304.7, *mp* 29 – 30 °C, is a yellow to brown solid which is moderately soluble in water and readily soluble in common organic solvents [23, p. 1220].

$[(CH_3)_2CH]_2NCOSCH_2CCl{=}CCl_2$

Triallate is produced by reaction of isopropylamine with phosgene, followed by coupling with 1,1-dichloro-2-chloro-3-mercapto-1-propylene [139].

Triallate inhibits fatty acid and lipid biosynthesis, and is a selective herbicide used pre- or post-planting and soil-incorporated for the control of wild oats and some annual grasses in wheat, barley, rye, field beans, etc.
Trade Names. Abrade BW (Monsanto, Du Pont); Frau-Go (Monsanto), etc.

Vernolate [*1929-77-7*], *S*-propyl dipropylthiocarbamate, $C_{10}H_{12}NOS$, M_r 203.3, *bp* 150 °C/30 mmHg, is a colorless liquid which is moderately soluble in water and miscible with common organic solvents [23, p. 1264].

$(CH_3CH_2CH_2)_2NCOSCH_2CH_2CH_3$

Vernolate is produced by reaction of *N*-dipropylcarbamoyl chloride with propylmercaptan [133].

Vernolate inhibits lipid metabolism, and is a selective herbicide used soil-incorporated, either pre-planting or post-emergence, for the control of germinating broad-leaved weeds and grasses in peanuts, maize, soya beans, tobacco, sweet potatoes, etc.
Trade Name. Vernam (Zeneca).

7.8.14. Organophosphates

Anilofos [*64249-01-0*], *S*-(4-chloro-*N*-isopropylcarbaniloxymethyl)-*O*,*O*-dimethyl phosphorodithioate, $C_{13}H_{19}ClNO_3PS_2$, M_r 367.8, *mp* 50.5 – 52.5 °C, is a crystalline solid which is moderately soluble in water and is readily soluble in acetone, chloroform, toluene, benzene, ethanol, dichloromethane, and ethyl acetate [23, p. 47].

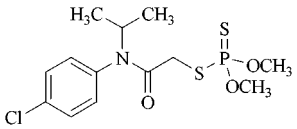

Anilofos is produced by reaction of *O*,*O*-dimethylphosphorodithioate with *p*-chlorobenzene-*N*-isopropylcarbamoyl chloride [140].

Anilofos is a selective herbicide used to control annual grasses and sedges in transplanted rice.

Trade Names. Arizon, Rico (AgrEvo); Anilogard (Gharda).

Bensulide [*741-58-2*], *S*-2-benzenesulfonamidoethyl-*O*,*O*-diisopropyl phosphorodithioate, $C_{14}H_{24}NO_4PS_2$, M_r 397.5, 34.4 °C, is a colorless solid which is moderately soluble in water and is readily soluble in acetone, ethanol, and xylene [23, p. 106].

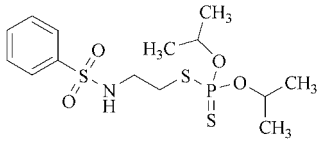

Bensulide is produced by reaction of *O*,*O*-diisopropylphosphorodithioate with *N*-chloroethylbenzene sulfonamide [141].

Bensulide is a selective herbicide used pre-emergence on cotton, on some vegetables, and also on established turf.

Trade Names. Betasan, Prefar (Gowan); Bensumac, Pre-San (PBI/Gordon); Lawnpur (Tomen).

Bilanafos [*35597-43-4*], 4-[hydroxy(methyl)- phosphinoyl]-L-homoalanyl-L-alanine, $C_{11}H_{22}N_2O_6P$, M_r 323.3, is readily soluble in water and practically insoluble in organic solvents [23, p. 120].

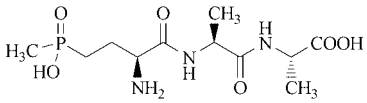

Bilanafos is produced by reaction of 2-nitro-5-phenol with phosphorus trichloride, followed by treatment with sulfur, and then with ethanol and 2-butylamine [142].

Bilanafos is a glutamine-synthase inhibitor, and is used post-emergence for the control of annual weeds in vines, apples, brassicas, rubber, etc. It is also used to control annual and perennial weeds in uncultivated land.

Trade Names. Meiji, Herbiace (Meiji Seika).

Butamifos [*36335-67-8*], *O*-ethyl-*O*′-6-nitro-*m*-tolyl *sec*-butylphosphoramidothioate, $C_{13}H_{21}N_2O_4PS$, M_r 332.4, is a yellow-brown liquid which is moderately soluble in water and is readily soluble in acetone, methanol, and xylene [23, p. 161].

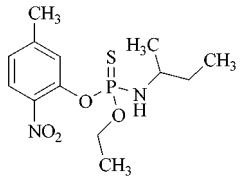

Butamifos is produced by reaction of 2-nitro-5-methylphenol with *O*-ethyl-*N*-*sec*–butylphosphoramidothionyl chloride [143].

Butamifos is a selective, nonsystemic herbicide used pre-emergence to control annual weeds in beans, rice, vegetables, etc.

Trade Name. Cremart (Sumitomo).

Glufosinate-ammonium [*77182-82-2*], ammonium 4-[hydroxy(methyl)phosphinoyl]-DL-homoalaninate, $C_5H_{15}N_2O_4P$, M_r 198.2, *mp* 125 °C, forms a colorless crystalline solid which is readily soluble in water and moderately soluble in acetone, ethanol, ethyl acetate, toluene, and hexane [23, p. 643].

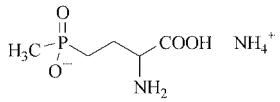

Glufosinate-ammonium is produced by reaction of *O*,*O*-diethylmethylphosphonate with 2-amino-4-bromobutanoic acid [144].

Glufosinate-ammonium, a glutamine-synthase inhibitor, is a nonselective contact herbicide used for the control of a wide range of annual and perennial broad-leaved weeds and grasses in fruit orchards, vineyards, rubber and oil palm plantations, ornamental trees and brushes, noncropland, and pre-emergence in vegetables.

Trade Names. Basta, Liberty (AgrEvo).

Glyphosate [*1071-83-6*], *N*-(phophonomethyl)glycine, $C_3H_8NO_5P$, M_r 169.1, *mp* 189.5 °C, forms colorless crystals which are fairly soluble in water and are insoluble in common organic solvents [23, p. 646].

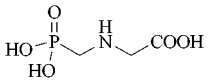

Glyphosate is produced by heating a mixture of phosphorous acid and α-amino acetic acid, and then adding formaldehyde [145].

Glyphosate inhibits enolpyruvylshikimi-3-phosphate synthase and is a nonselective herbicide used pre-harvest, post-planting/pre-emergence to control annual and perennial grasses and broad-leaved weeds in stubble, cereals, peas, beans, oilseed rape, flax, mustard, etc.

Trade Names. Rodeo, Roundup, Sting (Monsanto); Touchdown (Zeneca); Barbaran (Barclay); Sanos (Sanonda), etc.

Piperophos [*24151-93-7*], *S*-2-methylpiperidinocarbonylmethyl *O*,*O*-dipropyl phosphorodithioate, $C_{14}H_{28}NO_3PS_2$, M_r 353.5, *bp* > 250 °C, is a pale yellow slightly viscous colorless liquid which is moderately soluble in water and is miscible with benzene, hexane, acetone, dichloromethane, and ethanol [23, p. 983].

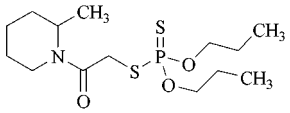

Piperophos is produced by reaction of *O*,*O*-diisopropylphosphorodithioate with 1-methylpiperidinylcarboxy chloride [146].

Piperophos is a selective, systemic herbicide used to control annual grasses and sedges in directly seeded or transplanted rice.

Trade Name. Rilof (Novartis).

7.8.15. Cyclohexanedione Oximes

Alloxydim-sodium [*66003-55-2*], methyl (*E*)-(*R*,*S*)-3-[1-(allyloxyimino)butyl]-4-hydroxy-6,6-dimethyl-2-oxocyclohex-3-enecarboxylate, $C_{17}H_{24}NNaO_5$, M_r 345.4, *mp* 185.5 °C, forms colorless crystals which are readily soluble in water, dimethylformamide, and methanol and are fairly soluble in ethanol, methyl ethyl ketone, acetone, cyclohexanone, ethyl acetate, and xylene [23, p. 30].

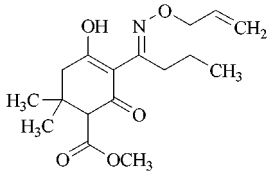

Alloxydim-sodium is produced by reaction of mesityloxide with dimethyl malonate; the resultant product is then treated with *n*-butyryl chloride, followed by reaction with hydroxyallylamine; finally the ester is hydrolyzed to the acid [147].

Alloxydim-sodium inhibits mitosis and fatty acid synthesis; it is a selective, systemic herbicide used post-emergence for the control of grass weeds and volunteer cereals in sugar beet, vegetables, and broad-leaved crops.

Trade Names. Clout (Rhône-Poulenc); Kusagard (Nippon Sosa); Frevin (AgrEvo), etc.

Clethodim [*99129-21-2*], (±)-2-[(*E*)-1-[(*E*)-3-chloroallyloxyimino]propyl]-5-[2-(ethylthio)propyl]-3-hydroxycyclohex-2-enone, $C_{17}H_{26}ClNO_3S$, M_r 359.9, is an amber liquid which insoluble in most organic solvents [23, p. 250].

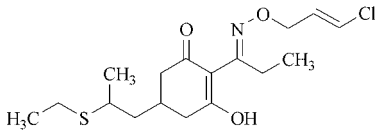

Clethodim is produced by reaction of ethylmercaptan with crotonal, followed, successively, by treatment with acetone, dimethyl malonate, hydrolysis reagents, propionic acid chloride, and α-chlorohydroxyallylamine [148].

Clethodim, a fatty-acid-synthesis inhibitor, is a systemic herbicide used post-emergence for the control of annual and perennial grasses in a wide range of broad-leaved crops, vegetables, trees, and vines.

Trade Name. Select (Tomen, Valent).

Cycloxydim [*101205-02-1*], (±)-2-[1-(ethoxyimino)butyl]-3-hydroxy-5-thian-3-ylcyclohex-2-enone, $C_{17}H_{27}NO_3S$, M_r 325.5, *mp* 41 °C, consists of colorless crystals which are moderately soluble in water, are readily soluble in acetone, ethanol, dichloromethane, and ethyl acetate, and are fairly soluble in *n*-hexane [23, p. 290].

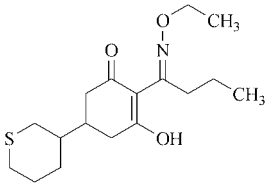

Cycloxydim is produced by reaction of 2-butyryl-5-[2-(1,3-dioxan-2-yl)ethyl]-1,3-cyclohexanedione with hydroxyethylamine [149].

Cycloxydim, which inhibits fatty acid synthesis, is a selective herbicide used post-emergence to control annual and perennial grasses in broad-leaved crops.

Trade Names. Focus, Lasre, Stratos (BASF).

Sethoxydim [*74051-80-2*], (±)-(*E*,*Z*)-2-(1-ethoxyiminobutyl)-5-[2-(ethylthio)propyl]-3-hydroxycyclohex-2-enone, $C_{17}H_{29}NO_3S$, M_r 327.5, *bp* $> 90\,°C/3 \times 10^{-5}$ mmHg, is an oily liquid which is fairly soluble in water and readily soluble in most organic compounds [23, p. 1101].

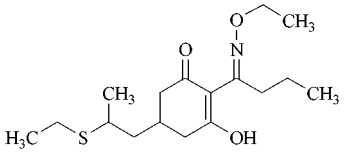

Sethoxydim is produced by reaction of ethylmercaptan with crotonal, followed by reaction with acetone, dimethylmalonate, hydrolysis reagents, butyryl chloride, and hydroxyethylamine [148].

Sethoxydim, which inhibits fatty acid synthesis, is a selective herbicide used to control annual and perennial grasses in broad-leaved crops.

Trade Names. Naba (Nippon Soda); Fervinal (AgrEvo); Poaset (BASF).

Tepraloxydim (BAS 620H) [*149979-41-9*], (*E*,*Z*)-(*R*,*S*)-2-{1-[(2*E*)-3-chloroallyloxyimino]propyl}-3-hydroxy-5-perhydropyran-4-ylcyclohex-2-en-1-one, $C_{17}H_{24}ClNO_4$, M_r 341.8 [23, p. 80].

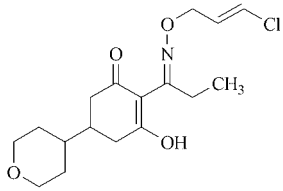

Tepraloxydim is prepared by condensation of 5-perhydropyranyl-3-hydroxy-2-carbonylpropylcyclohex-2-ene-1-one with chloroallyloxyamine [150].

Tepraloxydim is under development for use in soya beans and cotton (BASF).

Tralkoxydim [*87820-88-0*], 2-[1-(ethoxyimino)propyl]-3-hydroxy-5-mesitylcyclohex-2-enone, $C_{20}H_{27}NO_3$, M_r 329.4, *mp* 106 °C, consists of colorless crystals which are fairly soluble in water, hexane, methanol, and acetone and are readily soluble in ethyl acetate, toluene, and dichloromethane [23, p. 1211].

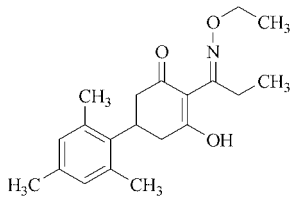

Tralkoxydim is produced by condensation of 2,4,6-trimethylbenzaldehyde with acetone, followed by condensation with diethyl malonate, acylation with propionic acid anhydride, and treatment with thioxyamine [151].

Tralkoxydim, an inhibitor of cell division and fatty-acid synthesisr, is a selective herbicide used for the control of *Avena* sp. and other grass weeds in wheat and barley.

Trade Names. Achieve, Grasp (Zeneca).

7.8.16. Benzonitriles

Dichlobenil [*1194-65-6*], 2,6-dichlorobenzonitrile, $C_7H_3Cl_2N$, M_r 172.0, *mp* 145 – 146 °C,

forms colorless crystals which are moderately soluble in water and are fairly soluble in acetone, dioxane, benzene, dichloromethane, xylene, and ethanol [23, p. 359].

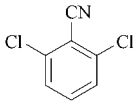

Dichlobenil is produced by conversion of 2,6-dichlorotoluene into 2,6-dichlorobenzal chloride, which is converted into 2,6-dichlorobenzaldehyde, 2,6-dichlorobenzaldoxime, and 2,6-dichlorobenzonitrile [152].

Dichlobenil, a cellulose-biosynthesis inhibitor, is a systemic herbicide used for selective weed control of many annual perennial weeds in woody ornamentals, fruit orchards, vineyards, brush fruit, forest plantations, etc.
Trade Name. Sibenil (Siagra).

Ioxynil [*1689-83-4*], 4-hydroxy-3,5-diiodobenzonitrile, $C_7H_3I_2NO$, M_r 370.9, *mp* 212 – 213 °C, is a colorless solid which is moderately soluble in water, fairly soluble in acetone, ethanol, methanol, and chloroform, and readily soluble in cyclohexanone, tetrahydrofuran, and dimethylformamide [23, p. 718].

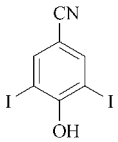

Ioxynil is produced by iodination of 4-hydroxybenzonitrile with iodine monochloride [128].

Ioxynil is a selective herbicide used post-emergence to control a wide range of annual broad-leaved weeds in cereals, leeks, onions, garlic, on lawns, and on newly sown turf.
Trade Names. Actril, Totril (Rhône-Polenc); Sanoxynil (Sanachem); Trionyl (Agriphar), etc.

7.8.17. Triazinones

Hexazinone [*51235-04-2*], 3-cyclohexyl-6-dimethylamino-1-methyl-1,3,5-triazine-2,4(1*H*,3*H*)-dione, $C_{12}H_{20}N_4O_2$, M_r 252.3, *mp* 115 – 117 °C, forms colorless crystals which are fairly soluble in water and hexane, and are readily soluble in chloroform, methanol, benzene, dimethylformamide, acetone, and toluene [23, p. 677].

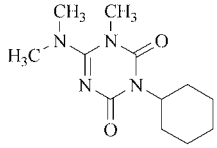

Haxazinone is produced by reaction of cyanoamine with methyl chloroformate, followed by methylation, and treatment with dimethylamine and cyclohexyl isocyanate [153].

Hexazinone, which inhibits photosynthesis, is a nonselective herbicide used post-emergence to control many annual and biennial weeds in alfalfa, pineapples, sugar cane, in plantations of certain coniferous species, and on noncrop areas.
Trade Name. Velpar (Du Pont).

Metamitron [*41394-05-2*], 4-amino-4,5-dihydro-3-methyl-6-phenyl-1,2,4-triazin-5-one, $C_{10}H_{10}N_4O$, M_r 202.2, *mp* 166.9°C, is a yellowish crystalline solid which is fairly soluble in water, dichloromethane, methanol, ethanol, chloroform, and toluene [23, p. 799].

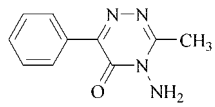

Matemitron is produced by reaction of *N′*-acetylbenzylhydrazone with hydroxylamine in pyridine containing potassium hydroxide [154].

Metamitron, a photosynthesis electron-transport inhibitor, is a selective, systemic herbicide used pre-drilling-incorporated and pre- and post-emergence for the control of grasses and broad-leaved weeds in sugar, fodder beets, etc.
Trade Names. Goltix (Bayer); Tornado (Feinchemie Schwebda).

Metribuzin [*21087-64-9*], 4-amino-6-*tert*-butyl-4,5-dihydro-3-methylthio-1,2,4-triazin-5-one, $C_8H_{14}N_4OS$, M_r 214.3 *mp* 126.3 °C, consists of colorless crystals which are moderately

soluble in water and hexane and are readily soluble in dimethylformamide, methanol, and many other organic solvents [23, p. 840].

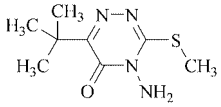

Metribuzin is produced by reaction of 4-amino-5-oxo-3-thioxo-6-*tert*-butyl-2,3,4,5-tetrahydro-1,2,4-triazine with methyliodide in the presence of methoxide [155].

Metribuzin, a photosynthesis electron-transport inhibitor, is a selective, systemic herbicide used pre- and post-emergence to control many grasses and broad-leaved weeds in soya beans, potatoes, tomatoes, sugar cane, asparagus, maize, and cereals.

Trade Names. Sencor (Bayer); Lexone (Du Pont), etc.

7.8.18. Miscellaneous Herbicides

Bispyribac-sodium [*125401-92-5*], sodium 2,6-bis(4,6-dimethoxypyrimidin-2-yloxy)benzoate, $C_{19}H_{17}N_4NaO_8$, M_r 452.4, *mp* 223 – 224 °C, is a colorless powder which is readily soluble in water and methanol and is sparingly soluble in acetone [23, p. 129].

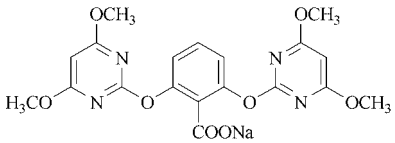

Bispyribac-sodium is produced by reaction of 2,6-dihydroxybenzoic acid with 2-chloro-4,6-dimethoxypyrimidine [156].

Bispyribac-sodium, which inhibits acetolactate synthase, is a selective, systemic post-emergence herbicide used to control a range of weeds in directly seeded rice.

Trade Names. Nominee, Grass-short (Kumiai); Short-keep (Riken Green).

Butroxydim [*138164-12-2*], 5-(3-butyryl-2,4,6-trimethylphenyl)-2-[1-(ethoxyimino)propyl-3-hydroxycyclohex-2-enone, $C_{24}H_{33}NO_4$, M_r 399.5, *mp* 80.8 °C, is an almost colorless powdery solid which is sparingly soluble in water [23, p. 167].

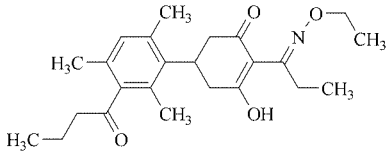

Butroxydim is produced by reaction of 5-(2,4,6-trimethylphenyl)-2-(1-ethoxyimino)-3-hydroxy-cyclohex-2-enone with butanoic acid chloride in the presence of aluminum trichloride [157].

Butroxydim is a post-emergence grass herbicide used on broad-leaved crops.

Trade Name. Falcon (Zeneca).

Cafenstrole [*125306-83-4*], *N*,*N*-diethyl-3-mesitylsulfonyl-1*H*-1,2,4-triazole-1-carboxamide, $C_{16}H_{22}N_4O_3S$, M_r 350.5, *mp* 114 – 116 °C, forms colorless crystals which are sparingly soluble in water [23, p. 173].

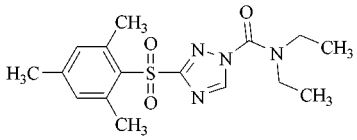

Cafenstrole is produced by reaction of 2-sulfonyl-(2′,4′,6′-trimethylphenyl)-1,3,4-triazole with *N*-diethylcarbamoyl chloride [158].

Cafenstrole is used pre- and early post-emergence for the control of annual weeds in paddy rice.

Cyclosulfamuron [*136849-15-5*], 1-[2-(cyclopropylcarbonyl)phenylsulfamoyl]-3-(4,6-dimethoxypyrimidin-2-yl)urea, $C_{17}H_{19}N_5O_6S$, M_r 421.4, *mp* 160.9 – 162.9 °C, is an almost colorless solid which is practically insoluble in water [23, p. 288].

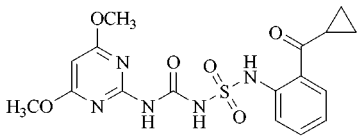

Cyclosulfamuron is produced by reaction of 2-amino-4,6-dimethoxypyrimidine with isocy-

anate sulfonyl chloride, followed by treatment with *o*-aminophenyl cyclopropyl ketone [159].

Cyclosulfamuron, which inhibits acetohydroxyacid synthase, is used to control a variety of dicotyledonous and sedge weeds in rice, wheat, and barley.
Trade Names. Jin-Qui, Savior (Cyanamid).

Dicamba [*10007-66-5*], 3,6-dichloro-*o*-anisic acid, $C_8H_6Cl_2O_3$, M_r 221.0, *mp* 114 – 116 °C, consists of colorless crystals which are moderately soluble in water and readily soluble in ethanol, cyclohexanone, acetone, dichloromethane, and toluene [23, p. 356].

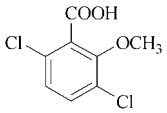

Dicamba is produced by hydrolysis of 1,2,4-trichlorobenzene with sodium hydroxide, followed by carboxylation, esterification with methanol, methylation of the hydroxyl group, and hydrolysis of the ester to the acid [160].

Dicamba is a selective systemic herbicide, with auxin-like growth-regulator activity, used to control annual and perennial broad-leaved weeds and brush species in cereals, maize, sugar cane, etc., and on turf, pastures, rangeland, and noncropland.
Trade Names. Banvel, Banvel SGF, Marksman (Novartis, BASF), etc.

Diclofop-methyl [*51338-27-3*], (*R*,*S*)-2-[4-(2,4-dichlorophenoxy)phenoxy]propionic acid, $C_{16}H_{14}Cl_2O_4$, M_r 341.2, *mp* 39 – 41 °C, forms colorless crystals which are sparingly soluble in water and are readily soluble in acetone, ethyl acetate, toluene, etc. [23, p. 347].

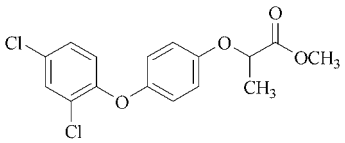

Diclofop-methyl is produced by reaction of 2,4-dichlorophenol with methyl 4-bromophenoxy-2-propionate [161].

Diclofop-methyl, a fatty-acid-synthesis inhibitor, is a selective, systemic herbicide used post-emergence to control wild oats, wild millet, and other annual grass weeds in many crops.
Trade Names. Hoegrass, Hoelon, Illoxan (AgrEvo).

Dimethenamid [*87674-68-8*], (*R*,*S*)-2-chloro-*N*-(2,4-dimethyl-3-thienyl)-*N*-(2-methoxy-1-methylethyl)acetamide, $C_{12}H_{18}ClNO_2S$, M_r 275.8, *bp* 127 °C/26.7 Pa, is a yellow-brown viscous liquid which is fairly soluble in water and readily soluble in heptane, ethanol, and diethyl ether [23, p. 409].

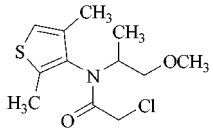

Dimethenamid is produced by reaction of 2,4-dimethyl-3-aminothiene with 2-chloro-3-methoxypropane, followed by treatment with chloroacetyl chloride [162].

Dimethenamid is used pre-emergence for the control of annual grasses and many broad-leaved weeds in maize, soya beans, and other crops.
Trade Name. Frontier (BASF).

Flupoxam [*119126-15-7*], 1-[4-chloro-3-(2,2,3,3,3-pentafluoropropoxymethyl)phenyl-5-phenyl-1*H*-1,2,4-triazole-3-carboxamide, $C_{19}H_{14}ClF_5N_4O_2$, M_r 460.8, *mp* 144 – 148 °C, forms light beige crystals which are sparingly soluble in water and are readily soluble in methanol, acetone, and ethyl acetate [23, p. 583].

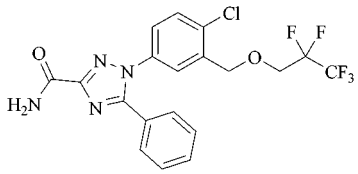

Flupoxam is produced by reduction of the nitro group of 2-chloro-5-nitrobenzene, followed by diazotization, addition of benzylamino ethylcarboxylate, and cyclization of the product [163].

Flupoxam is a cell-elongation inhibitor, used pre- and post-emergence to control annual broad-leaved weeds in winter wheat and winter barley.
Trade Name. Ovation (Synexus).

Fluthiacet-methyl [*117337-16-6*], methyl [2-chloro-4-fluoro-5-(5,6,8-tetrahydro-3-oxo-1*H*,3*H*-[1,3,4,]thiadiazolo[3,4-*a*]pyridazin-1-ylideneamino)phenylthio] acetate, $C_{15}H_{15}ClFN_3O_3S_2$, M_r 403.9, *mp* 105.0 – 106.5 °C, is a colorless powder which is sparingly soluble in water, moderately soluble in methanol and dichloromethane, and dissolves well in acetone, acetonitrile, and ethyl acetate [23, p. 606].

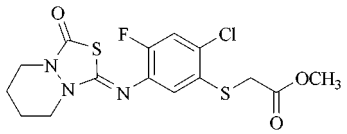

Fluthiacet-methyl is produced by reaction of 1-[(2-fluoro-4-chloro-5-thiomethylacetate)phenylthiocarbamoyl]-2-methoxycarbonylhexahydropyridazine with methanolic potassium hydroxide, followed by treatment with trichloromethyl chloroformate [164].

Fluthiacet-methyl, which inhibits protoporphyrinogen oxidase, is a selective contact herbicide used post-emergence for the control of broad-leaved weeds in maize and soya beans.
Trade Name. Action (Novartis).

Fluthiamide (BAY FOE 5043) [*142459-58-3*], 4′-fluoro-*N*-isopropyl-2-(5-trifluoromethyl-1,3,4-thiadiazol-2-yloxy)acetamide, $C_{14}H_{13}F_4N_3O_2S$, M_r 363.4, *mp* 75 – 77 °C, is a colorless to tan solid which is moderately soluble in water [23, p. 82].

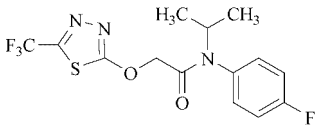

Fluthiamide is produced by reaction of *p*-fluoro-*N*-propylalanine with 5-trifluoromethyl-1,3,4-thiadiazol-2-oxyacetyl chloride [165].

Fluthiamide, which inhibits cell division and growth, is used pre- and early post-emergence. It is under development for the control of a broad spectrum of grass in a wide range of crops.
Trade Name. Axiom (Bayer).

Indanofan [*133220-30-1*], (*R*,*S*)-2-[2-(3-chlorophenyl)-2,3-epoxypropyl]-2-ethylindan-1,3-dione, $C_{20}H_{17}ClO_3$, M_r 340.8, *mp* 60.0 – 62.5 °C, is sparingly soluble in water [23, p. 715].

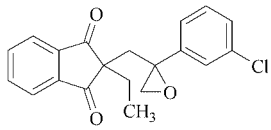

Indanofan is produced by reaction of 2-[2-(3-chlorophenyl)propen-3-yl]-2-ethyllindan-1,3-dione with peracetic acid [166].

Indanofan is under development by Mitsubishi, for pre-and post-emergence use, for the control of weeds in transplanted rice.

Isoxaflutole [*141112-29-0*], 5-cyclopropyl-1,2-oxazol-4-yl α,α,α-trifluoro-2-mesyl-*p*-tolyl ketone, $C_{15}H_{12}F_3NO_4S$, M_r 359.3, *mp* 140 °C, is a colorless to pale yellow solid which is sparingly soluble in water [23, p. 737].

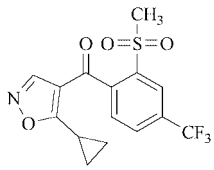

Isoxaflutole is produced by reaction of 1-(4-trifluoromethyl-2-methylsulfonylphenyl)-3-cyclopropyl-2-ethoxymethylenepropane-1,3-dione with hydroxylamine [167].

Isoxaflutole, which inhibits *p*-hydoxyphenyl pyruvate dioxygenase, is a systemic herbicide used pre-emergence or pre-planting for the control of a broad spectrum of grass and broad-leaved weeds in maize.
Trade Names. Balance, (Rhône-Poulenc).

Mefenacet [*73250-68-7*], 2-(1,3-benzothiazol-2-yloxy)-*N*-methylacetanilide, $C_{16}H_{14}N_2O_2S$, M_r 298.4, *mp* 134.8 °C, consists of colorless crystals which are sparingly soluble in water, moderately soluble in toluene and isopropyl alcohol, and readily soluble in dichloromethane [23, p. 779].

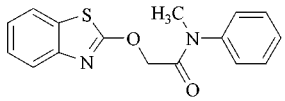

Mefenacet is prepared by reaction of 2-chlorobenzthiazol with *N*-methylbenzene hydroxyacetamide [165].

Mefenacet, a selective herbicide which inhibits cell growth and cell division, is used pre- and post-emergence mainly in transplanted rice.
Trade Names. Act Nihon (Bayer, Nissan, Yashima); Hinochloa, Rancho (Bayer); Zark (Du Pont, Kumiai, Nihon Bayer).

Metosulam [*139528-85-1*], 2′,6′-dichloro-5,7-dimethoxy-3′-methyl[1,2,4]triazolo[1,5-a]pyrimidine-2-sulfonanilide, $C_{14}H_{13}Cl_2N_5O_4S$, M_r 418.3, *mp* 210 – 211.5 °C, is a cream- to tan-colored powder which is moderately soluble in water, hexane and toluene and is fairly soluble in acetone, acetonitrile, and dichloromethane [23, p. 836].

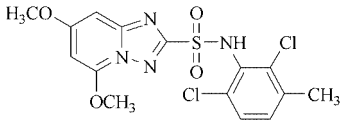

Metosulam is produced by reaction of 4,6-dimethoxy-(2-sulfonyl)-2,3-triazolyl pyrimidine chloride with 2,6-dichloro-5-methylaniline [168].

Metosulam, which inhibits acetolactate synthase, is used pre- and post-emergence for the control of broad-leaved weeds in wheat, barley, rye, and maize.
Trade Names. Pronto, Sansac, Sinal (DowElanco).

Propaquizafop [*111479-05-1*], 2-isopropyl-ideneamino-oxyethyl (*R*)-2[4-(6-chloroquinoxalin-2-yloxy)phenoxy]propionate, $C_{22}H_{22}ClN_3O_5$, M_r 443.9, *mp* 66.3 °C, forms colorless crystals which are sparingly soluble in water, readily soluble in acetone and toluene, and fairly soluble in *n*-hexane and ethanol [23, p. 1021].

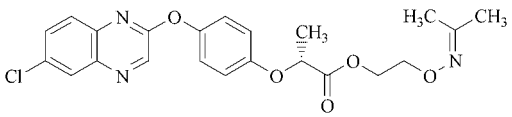

Propaquizafop is produced by reaction of 2-(6-chloroaminoxalinyl)phenoxypropionic acid ethoxycarbonyl chloride with isopropylidene hydroxylamine [169].

Propaquizafop, which inhibits fatty acid synthesis, is a systemic, post-emergence herbicide used for the control of a wide range of annual and perennial grasses in soya beans, sugar cane, cotton, vegetables, etc.
Trade Name. Agil (Novartis).

Pyridate [*55512-33-9*], 6-chloro-3-phenylpyridazin-4-yl *S*-octyl thiocarbonate, $C_{19}H_{23}ClN_2O_2S$, M_r 378.9, *mp* 27 °C, forms colorless crystals which are sparingly soluble in water and are readily soluble in many organic solvents [23, p. 1064].

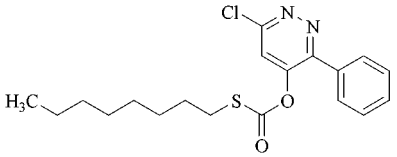

Pyridate is produced by reaction of acetophenone with diethyloxalate, followed by reaction with hydrazine, phosphorus pentachloride, and octylmercaptocarboxy chloride [170].

Pyridate, which inhibits photosystem II electron transport in photosynthesis, is a selective herbicide used post-emergence to control annual broad-leaved weeds in maize, sweetcorn, cereals, rice, and other crops.
Trade Names. Lentagran (Novartis); Pirate (Barclay).

Pyrithiobac-sodium [*123343-16-8*], sodium 2-chloro-6-(4,6-dimethoxypyrimidin-2-ylthio) benzoate, $C_{13}H_{10}ClN_2NaO_4S$, M_r 348.7, *mp* 233.8 – 234.2 °C, is a colorless solid which is readily soluble in water, acetone, and methanol and is fairly soluble in dichloromethane [23, p. 1073].

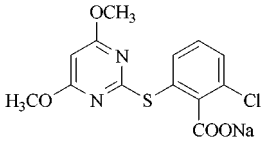

Pyrithiobac-sodium is produced by diazotization of 2-amino-6-chlorobenzoic acid ethyl ester, followed by treatment with 4,6-dimeth-

oxy-2-mercaptopyrimidine, and hydrolysis of the ester to the acid [171].

Pyrithiobac-sodium, which inhibits acetolactate synthase, is a selective herbicide used pre- and post-emergence for the control of a wide range of broad-leaved weeds in cotton.
Trade Name. Staple (Du Pont, Kumiai).

7.9. Toxicology

7.9.1. Diphenyl Ethers

Acifluorfen Na salt: [*62476-59-9*], [23, p. 12]

LD_{50} (rat, oral)	1540	mg/kg	
LD_{50} (mice, oral)	1370	mg/kg	
LD_{50} (rabbit, oral)	1590	mg/kg	
LD_{50} (rabbit, dermal)	>2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	>6.91	mg/L air	
Rabbit: moderately irritating to skin			
Rabbit: severely irritating to eyes			
Nonmutagenic [173]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	1.25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[173]
Teratogenicity.			
Rat (gavage)	90	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[173]
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[173]
Rabbit (gavage)	>36	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[173]
Chronic.			
Rat (2-year feed)	25	$mg\,kg^{-1}\,d^{-1}$	[173]
Dog (2-year feed)	1.25	$mg\,kg^{-1}\,d^{-1}$	[173]
Mouse (18-month feed)	1	$mg\,kg^{-1}\,d^{-1}$	
ADI/RfD: $0.013\,mg\,kg^{-1}\,d^{-1}$ [23] [173]			

Aclonifen [*74070-46-5*], [23, p. 14]

LD_{50} (rat, oral)	>5000	mg/kg
LD_{50} (mouse, oral)	>5000	mg/kg
LD_{50} (rat, dermal)	>5000	mg/kg
LC_{50} (rat, 4-h inhalation)	>5.06	mg/L air
Rabbit: slightly irritating to skin		
Rabbit: nonirritating to eyes		
Nonmutagenic (Ames test)		
NOEL/NOAEL		
Reproduction.		
Rat (2-generation feed)	130	$mg\,kg^{-1}\,d^{-1}$
Subchronic.		
Rat (90-day feed)	28	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Dog (6-month feed)	12.5	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$		

AKH-7088 [*104459-82-7*], [23, p. 22]

LD_{50} (rat, oral)	5000	mg/kg
LD_{50} (rat, dermal)	2000	mg/kg
Rabbit: moderately irritating to skin		
Rabbit: nonirritating to eyes		
Nonmutagenic (Ames test)		

Bifenox [*42576-02-3*], [23, p. 69]

LD_{50} (rat, oral)	>5000	mg/kg	
LD_{50} (mouse, oral)	>4556	mg/kg	
LD_{50} (rabbit, dermal)	>2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	>0.91	mg/L air	
Rabbit: nonirritating to skin and eyes			
Nonmutagenic (in vitro)			
NOEL/NOAEL			
Teratogenicity.			
Rabbit (gavage)	10 – 20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 35]
	200	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 35]
Chronic.			
Rat (2-year feed)	80	$mg\,kg^{-1}\,d^{-1}$	
Dog (2-year feed)	ca. 145	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	50	$mg\,kg^{-1}\,d^{-1}$	[174, p. 35]

Chlomethoxyfen [*32861-85-1*], [23, p. 200]

LD_{50} (rat, oral)	>10000	mg/kg
LD_{50} (mouse, oral)	>10000	mg/kg
LD_{50} (rat, dermal)	>5000	mg/kg
LC_{50} (rat, 4-h inhalation)	>1.77	mg/L air

Fluoroglycofen-ethyl [*77501-90-7*], [23, p. 580]

LD_{50} (rat, oral)	1500	mg/kg	
LD_{50} (rabbit, dermal)	>5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	>7.5	mg/L air	
Rabbit: slightly irritating to skin and eyes			
Guinea pig: not a skin sensitizer [174, p. 138]			
Nonmutagenic [174, p. 138]			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	1.0	$mg\,kg^{-1}\,d^{-1}$	[174, p. 138]
Dog (1-year feed)	9	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	1.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 138]
ADI: $0.009\,mg\,kg^{-1}\,d^{-1}$			

Fomesafen [*72178-02-0*], [23, p. 616]

LD_{50} (rat, oral)	1250 – 2000	mg/kg	
LD_{50} (rabbit, dermal)	1600	mg/kg	
LC_{50} (rat, 4-h inhalation)	4.97	mg/L air	
Rabbit: mildly irritating to skin			
Rabbit: moderately irritating to eyes			
Guinea pig: a potential skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Teratogenicity.			
Rabbit (gavage)	2.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	0.25	$mg\,kg^{-1}\,d^{-1}$	
Dog (6-month feed)	1.0	$mg\,kg^{-1}\,d^{-1}$	[174, p. 143]
Mouse (2-year feed)	1.0	$mg\,kg^{-1}\,d^{-1}$	
ADI: 0.0025 $mg\,kg^{-1}\,d^{-1}$			

HC-252 [*131086-42-5*], [23, p. 663]

LD_{50} (rat, oral)	843 – 963	mg/kg
LD_{50} (mouse, oral)	1113 – 1269	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 14-d inhalation)	9.34	mg/L air
Rabbit: nonirritating to skin		
Rabbit: moderately irritating to eyes		
Nonmutagenic (in vitro assays)		

Lactofen [*77501-63-4*], [23, p. 747]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
Rabbit: severely irritating to skin			
Nonmutagenic [175]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	2.5	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[175]
Teratogenicity.			
Rat (gavage)	50	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[175]
Rabbit (gavage)	4	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[175]
	> 20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[175]
Subchronic.			
Chimpanzees (93-day capsule)	75	$mg\,kg^{-1}\,d^{-1}$	[175]
Chronic.			
Rat (2-year feed)	ca. 25	$mg\,kg^{-1}\,d^{-1}$	[175]
Dog (2-year feed)	ca. 5.7	$mg\,kg^{-1}\,d^{-1}$	[175]
Mouse (18-month feed)	< 1.4	$mg\,kg^{-1}\,d^{-1}$	[175]
ADI/RfD: 0.002 $mg\,kg^{-1}\,d^{-1}$ [175]			

Oxyfluorfen [*42874-03-3*], [23, p. 919]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	2700 – 5900	mg/kg	[174, p. 224]
LD_{50} (dog, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 5000	mg/kg	[174, p. 224]
LD_{50} (rabbit, dermal)	> 10 000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.4	mg/L air	
Rabbit: mildly irritating to skin			
Rabbit: moderately irritating to eyes			
Guinea pig: not a skin sensitizer [174, p. 224]			
Mutagenicity: no chromosomal aberations *in vivo* but some genotoxicity potential (*in vitro* assays) [174, p. 224]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	5	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 224]
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 224]
Teratogenicity.			
Rat (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 224]
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 224]
Chronic.			
Rat (2-year feed)	ca. 2	$mg\,kg^{-1}\,d^{-1}$	[23] [174, p. 224]
Dog (2-year feed)	3	$mg\,kg^{-1}\,d^{-1}$	[23] [174, p. 224]
Mouse (20-month feed)	ca. 0.3	$mg\,kg^{-1}\,d^{-1}$	[174, p. 224]
ADI/RfD: 0.003 $mg\,kg^{-1}\,d^{-1}$			

7.9.2. Sulfonylureas

Amidosulfuron [*120923-37-7*], [23, p. 37]

LD_{50} (rat, oral)	≈ or > 5000	mg/kg
LD_{50} (mouse, oral)	≈ or > 5000	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.8	mg/L air
Nonmutagenic		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	19.5	$mg\,kg^{-1}\,d^{-1}$
ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$		

Azimsulfuron [*120162-55-2*], [23, p. 63]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.9	mg/L air	
Rabbit: nonirritating to skin and eyes			
Guinea pig: not a skin sensitizer			
Nonmutagenic (Ames test)			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	ca. 50	$mg\,kg^{-1}\,d^{-1}$	
ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ (proposed)			

Bensulfuron-methyl [*83055-99-6*], [23, p. 104]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 10 985	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 7.5	mg/L air	
Rabbit: nonirritating to skin and eyes			
Guinea pig: not a skin sensitizer [174, p. 28]			
Nonmutagenic (set of standard assays) [174, p. 28]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	ca. 20	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 28]
Teratogenicity.			
Rat (gavage)	2000	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 28]
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 28]
Rabbit (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 28]
Chronic.			
Rat (2-year feed)	ca. 37.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	ca. 20	$mg\,kg^{-1}\,d^{-1}$	[174, p. 28]
Mouse (2-year feed)	ca. 455	$mg\,kg^{-1}\,d^{-1}$	[174, p. 28]
ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$			

Chlorimuron-ethyl [*90982-32-4*], [23, p. 217]

LD_{50} (rat, oral)	4102 – 4236	mg/kg	[174, p. 309]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.6	mg/L air	[174, p. 309]
Rabbit: nonirritating to skin and eyes [174, p. 309]			
Guinea pig: not a skin sensitizer [174, p. 309]			
Nonmutagenic [174, p. 309]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	12.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[176]
	1.25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[176]
Teratogenicity.			
Rat (gavage)	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[176]
	150	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[176]
Rabbit (gavage)	60	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[176]
	300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[176]
Chronic.			
Rat (2-year feed)	12.5	$mg\,kg^{-1}\,d^{-1}$	[176]
Dog (1-year feed)	6.25	$mg\,kg^{-1}\,d^{-1}$	[176]
Mouse (18-month feed)	3.75	$mg\,kg^{-1}\,d^{-1}$	[176]
ADI: 0.02 $mg\,kg^{-1}\,d^{-1}$ [23]; 0.062 $mg\,kg^{-1}\,d^{-1}$ [176]			

Chlorsulfuron [*64902-72-3*], [23, p. 239]

LD_{50} (rat, oral)	5545 – 6293	mg/kg	
LD_{50} (guinea pig, oral)	1363	mg/kg	[174, p. 58]
LD_{50} (dog, oral)	> 2500	mg/kg	[174, p. 58]
LD_{50} (rat, i.p.)	1450	mg/kg	
LD_{50} (rabbit, dermal)	2500	mg/kg	[23]
	3400	mg/kg	[174, p. 58]
LC_{50} (rat, 4-h inhalation)	> 5.9	mg/L air	
Rabbit: nonirritating to skin			
Rabbit: mildly irritating to eyes			
Guinea pig: not a skin sensitizer			
Nonmutagenic [174, p. 58]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 58]
Teratogenicity.			
Rat (gavage)	165	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 58]
Rabbit	ca. 75 – 200	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 58]
Chronic.			
Rat (2-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	ca. 55	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	ca. 71	$mg\,kg^{-1}\,d^{-1}$	
ADI: 0.05 $mg\,kg^{-1}\,d^{-1}$			

Cinosulfuron [*94593-91-6*], [23, p. 248]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 5000	mg/kg	[177]
LC_{50} (rat, 4-h inhalation)	> 5.6	mg/L air	
Rabbit: nonirritating to skin and eyes			
Guinea pig: not a skin sensitizer			
Nonmutagenic (*in vitro*, various assays) [177]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	36.9	$mg\,kg^{-1}\,d^{-1}$ (parental)	[177]

Teratogenicity.

Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[177]
	> 2400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[177]
Rabbit (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[177]
	> 120	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[177]

Chronic.

Rat (2-year feed)	ca. 20	$mg\,kg^{-1}\,d^{-1}$	[177]
Dog (1-year feed)	ca. 65 – 92	$mg\,kg^{-1}\,d^{-1}$	[177]
Mouse (2-year feed)	ca. 60 – 404	$mg\,kg^{-1}\,d^{-1}$	[177]

Ethametsulfuron-methyl [*97780-06-8*], [23, p. 475]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.7	mg/L air

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic [174, p. 123]
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	ca. 300	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 123]

Teratogenicity.

Rat (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 123]
Rabbit (gavage)	250	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 123]

Chronic.

Rat (1-year feed)	ca. 25	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 80	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 710	$mg\,kg^{-1}\,d^{-1}$

Ethoxysulfuron [*126801-58-9*], [23, p. 488]

LD_{50} (rat, oral)	> 3270	mg/kg
LD_{50} (rat, dermal)	> 4000	mg/kg

Rabbit: nonirritating to skin and eyes
Nonmutagenic (Ames test)

Flazasulfuron [*104040-78-0*], [23, p. 550]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	5.99	mg/L air

Rabbit: nonirritating to skin
Rabbit: moderately irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic

Chronic.

Rat (2-year feed)	ca. 1.3	$mg\,kg^{-1}\,d^{-1}$

Flupyrsulfuron-methyl-sodium [*144740-54-5*], [23, p. 586]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.0	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL

Chronic.

Rat (2-year feed)	ca. 17.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	13.6	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.035\,mg\,kg^{-1}\,d^{-1}$

Halosulfuron-methyl [*100784-20-1*], [23, p. 657]

LD_{50} (rat, oral)	8865	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 6.0	mg/L air

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL

Chronic.

Rat (2-year feed)	ca. 50	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	1.0	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 430	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.1\,mg\,kg^{-1}\,d^{-1}$

Imazosulfuron [*122548-33-8*], [23, p. 703]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.4	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic [178]
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	77.8	$mg\,kg^{-1}\,d^{-1}$ (parental)	[178]

Teratogenicity.

Rat (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[178]
	1500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[178]
Rabbit (gavage)	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[178]
	125	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[178]

Chronic.

Rat (2-year feed)	106	$mg\,kg^{-1}\,d^{-1}$	[23] [178]

Dog (1-year feed)	75	$mg\,kg^{-1}\,d^{-1}$	[23] [178]
Mouse (18-month feed)	72.7	$mg\,kg^{-1}\,d^{-1}$	[178]

ADI: $0.7\,mg\,kg^{-1}\,d^{-1}$

Metsulfuron-methyl [*74223-64-6*], [23, p. 842]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.3	mg/L air	[174, p. 203]

Rabbit: moderately irritating to eyes
Guinea pig: mildly irritating to skin
Guinea pig: not a skin sensitizer
Nonmutagenic (test battery)
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$	[174, p. 203]

Teratogenicity.

Rat (gavage)	40	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 203]
Rabbit (gavage)	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 203]

Chronic.

Rat (2-year feed)	ca. 25	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	ca. 13	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	ca. 710	$mg\,kg^{-1}\,d^{-1}$	[174, p. 203]

ADI: $0.0125\,mg\,kg^{-1}\,d^{-1}$

Nicosulfuron [*111991-09-4*], [23, p. 877]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	5.9	mg/L air	[174, p. 216]

Rabbit: nonirritating to skin
Rabbit: moderately irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (test battery) [179]
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	287	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	[179]

Teratogenicity.

Rat (feed)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[179]
Rabbit (feed)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[179]
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[179]

Chronic.

Rat (2-year feed)	> 786	$mg\,kg^{-1}\,d^{-1}$	[179]
Dog (1-year feed)	125	$mg\,kg^{-1}\,d^{-1}$	[179]
Mouse (18-month feed)	> 993	$mg\,kg^{-1}\,d^{-1}$	[179]

ADI/RfD: $1.25\,mg\,kg^{-1}\,d^{-1}$ [179]

Oxasulfuron [*144651-06-9*], [23, p. 911]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.08	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL

Chronic.

Rat (2-year feed)	8.3	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	1.3	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	1.5	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.0026\,mg\,kg^{-1}\,d^{-1}$

Primisulfuron-methyl [*86209-51-0*], [23, p. 997]

LD_{50} (rat, oral)	> 5050	mg/kg	
LD_{50} (rat, dermal)	> 2010	mg/kg	
LD_{50} (rabbit, dermal)	> 2010	mg/kg	[174, p. 238]
LD_{50} (mouse, dermal)	> 2100	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 4.8	mg/L air	

Rabbit: practically nonirritating to skin [174, p. 238]
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (in vitro/in vivo) [174, p. 238]
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	ca. 65	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 238]

Teratogenicity.

Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 238]
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 238]
	600	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 238]

Chronic.

Rat (2-year feed)	15	$mg\,kg^{-1}\,d^{-1}$	[174, p. 238]
Dog (1-year feed)	25	$mg\,kg^{-1}\,d^{-1}$	[174, p. 238]
Mouse (19-month feed)	40	$mg\,kg^{-1}\,d^{-1}$	[180]

ADI: $0.13\,mg\,kg^{-1}\,d^{-1}$

Prosulfuron [*94125-34-5*], [23, p. 1041]

LD_{50} (rat, oral)	986	mg/kg
LD_{50} (mouse, oral)	1247	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.4	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Neurotoxicity.

Rat (acute gavage)	10	mg/kg	[181]

Rat (90-day feed)	ca. 13	$mg\,kg^{-1}\,d^{-1}$ (systemic)	[181]
	> 330	$mg\,kg^{-1}\,d^{-1}$ (neurotoxicity)	[181]
Reproduction.			
Rat (2-generation feed)	13	$mg\,kg^{-1}\,d^{-1}$ (systemic and reproduction)	[181]
Teratogenicity.			
Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[181]
Rabbit (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[181]
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[181]
Chronic.			
Rat (2-year feed)	7.9	$mg\,kg^{-1}\,d^{-1}$	[181]
Dog (1-year feed)	1.84	$mg\,kg^{-1}\,d^{-1}$	[181]
Mouse (18-month feed)	1.7	$mg\,kg^{-1}\,d^{-1}$	[174, p. 50]

ADI: 0.019 $mg\,kg^{-1}\,d^{-1}$ [23], RfD: 0.02 $mg\,kg^{-1}\,d^{-1}$ [181]

Pyrazosulfuron-ethyl [*93697-74-6*], [23, p. 1052]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.9	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL

Chronic.		
Mouse (18-month feed)	4.3	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.043 $mg\,kg^{-1}\,d^{-1}$

Rimsulfuron [*122931-48-0*], [182]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.4	mg/L air

Rabbit: nonirritating to skin
Rabbit: moderately irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	ca. 200	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1052]
Teratogenicity.			
Rat (gavage)	2000	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 6000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	170	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	ca. 15	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1052]
Dog (1-year feed)	ca. 1.25	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1052]
Mouse (18-month feed)	ca. 360	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1052]

Sulfometuron-ethyl [*74222-97-2*], [23, p. 1128]

LD_{50} (rat, oral)	> 5000	mg/kg	[23]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[23]
LC_{50} (rat, 4-h inhalation)	> 11	mg/L air	[174, p. 274]

Rabbit: slightly irritating to skin [23]
Rabbit: mildly irritating to eyes [174, p. 274]
Guinea pig: not a skin sensitizer [23]
Nonmutagenic (*in vitro*) [174, p. 274]
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	ca. 33	$mg\,kg^{-1}\,d^{-1}$	[23]
Teratogenicity.			
Rat (feed?)	65	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 274]
Rabbit (gavage)	> 300	$mg\,kg^{-1}\,d^{-1}$	[174, p. 274]
Chronic.			
Rat (2-year feed)	2.5	$mg\,kg^{-1}\,d^{-1}$	[23]
Dog (1-year feed)	9	$mg\,kg^{-1}\,d^{-1}$	[23]
	ca. 5.7	$mg\,kg^{-1}\,d^{-1}$	[174, p. 274]
Mouse (18-month feed)	1.6	$mg\,kg^{-1}\,d^{-1}$	[23]
	ca. 14.3	$mg\,kg^{-1}\,d^{-1}$	[174, p. 274]

ADI: 0.02 $mg\,kg^{-1}\,d^{-1}$ [23]

Sulfosulfuron [*141776-32-1*], [23, p. 1130]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg

LC_{50} (rat, 4-h inhalation): ″practically nontoxic″
Rabbit: nonirritating to skin
Rabbit: moderately irritating to eyes
Guinea pig: not a skin sensitizer

Thifensulfuron-methyl [*79277-27-3*], [183]

LD_{50} (rat, oral)	> 5000	mg/kg	[23, p. 1188]
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[23, p. 1188]
LC_{50} (rat, 4-h inhalation)	> 7.9	mg/L air	[23]
	> 5.7	mg/L air	[174, p. 282]

Rabbit: nonirritating to skin [23, p. 1188]
Rabbit: moderately irritating to eyes [23] [174, p. 282]
Guinea pig: not a skin sensitizer [23, p. 1188]
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	125	$mg\,kg^{-1}\,d^{-1}$	
Teratogenicity.			
Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	350	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 650	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	0.96	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	19.7	$mg\,kg^{-1}\,d^{-1}$	
Mouse (90-day feed)	979	$mg\,kg^{-1}\,d^{-1}$	

ADI: $0.026\,mg\,kg^{-1}\,d^{-1}$ [23, p. 1188]

Triasulfuron [*82097-50-5*], [23, p. 1222]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.18	mg/L air

Rabbit: mildly irritating to skin
Rabbit: non to slightly irritating to eyes [174, p. 287]
Guinea pig: not a skin sensitizer
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 65	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 287]
	ca. 300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 287]
Teratogenicity.			
Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 287]
	300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 287]
Rabbit (gavage)	120	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 287]
	240	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 287]
Chronic.			
Rat (2-year feed)	32.1	$mg\,kg^{-1}\,d^{-1}$	[23]
	0.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 287]
Dog (1-year feed)	33	$mg\,kg^{-1}\,d^{-1}$	[23] [174, p. 287]
Mouse (2-year feed)	1.2	$mg\,kg^{-1}\,d^{-1}$	[23] [174, p. 287]

ADI: $0.012\,mg\,kg^{-1}\,d^{-1}$

Tribenuron-methyl [*101200-48-0*], [23, p.]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.0	mg/L air

Rabbit: nonirritating to skin
Rabbit: mildly irritating to eyes
Guinea pig: a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 1.6	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 290]
Teratogenicity.			
Rat (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 290]
Rabbit (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 290]
Chronic.			
Rat (2-year feed)	ca. 1.25	$mg\,kg^{-1}\,d^{-1}$	[174, p. 290]
Dog (1-year feed)	8.2	$mg\,kg^{-1}\,d^{-1}$	[174, p. 290]
Mouse (18-month feed)	30	$mg\,kg^{-1}\,d^{-1}$	[174, p. 290]

ADI: $0.011\,mg\,kg^{-1}\,d^{-1}$

Triflusulfuron-methyl [*126535-15-7*], [23, p. 1250]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.1	mg/L air

Rabbit: nonirritating to skin and eyes
Rabbit: moderately irritating to eyes [174, p. 300]
Guinea pig: not a skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 5.8	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 300]
Teratogenicity.			
Rat (gavage)	120	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 300]
Rabbit (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 300]
	> 800	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 300]

Chronic.

Rat (2-year feed)	4.1	$mg\,kg^{-1}\,d^{-1}$	[174, p. 300]
Dog (1-year feed)	26.9	$mg\,kg^{-1}\,d^{-1}$	[174, p. 300]
Mouse (18-month feed)	20.9	$mg\,kg^{-1}\,d^{-1}$	[174, p. 300]

ADI: $0.05\,mg\,kg^{-1}\,d^{-1}$

7.9.3. Chloroacetanilides

Acetochlor [*34256-82-1*], [174, p. 3]

LD_{50} (rat, oral)	2148	mg/kg	[23, p. 10]
LD_{50} (rabbit, dermal)	4166	mg/kg	[23, p. 10]
LC_{50} (rat, 4-h inhalation)	>3.0	mg/L air	[23, p. 10]

Rabbit: practically nonirritating to skin and eyes [23, p. 10]
Guinea pig: a skin sensitizer [23, p. 10]
Nonmutagenic (*in vivo*)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$ (parental)

Teratogenicity.

Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	50	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	10	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	12	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.01\,mg\,kg^{-1}\,d^{-1}$ [23, p. 10]

Alachlor [*15972-60-8*], [184]

LD_{50} (rat, oral)	930 – 1350	mg/kg	[23, p. 23]
LD_{50} (mouse, oral)	1100	mg/kg	[23, p. 23]
LD_{50} (rabbit, dermal)	13 300	mg/kg	[23, p. 23]
LC_{50} (rat, 4-h inhalation)	1.04	mg/L air	[23, p. 23]

Rabbit: practically nonirritating to skin and eyes [23]
Rabbit: slightly to moderately irritating to skin and eyes
Guinea pig: a skin sensitizer [23, p. 23]
Mutagenicity: overall nonmutagenic *in vivo*
NOEL/NOAEL
Reproduction.

Rat (3-generation feed)	10	$mg\,kg^{-1}\,d^{-1}$ (parental)

Teratogenicity.

Rat (gavage)	150	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	150	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$	[23]
Dog (1-year feed)	ca. 1	$mg\,kg^{-1}\,d^{-1}$	[23]
Mouse (18-month feed)	26	$mg\,kg^{-1}\,d^{-1}$	[174, p. 3]

ADI: $0.0025\,mg\,kg^{-1}\,d^{-1}$
RfD: 0.01 (EPA 1995)

Butachlor [*23184-66-9*], [23, p. 159]

LD_{50} (rat, oral)	2000	mg/kg	
LD_{50} (mouse, oral)	4747	mg/kg	
LD_{50} (rabbit, oral)	>5010	mg/kg	
LD_{50} (rabbit, dermal)	13 100	mg/kg	[23] [174, p. 41]
LC_{50} (rat, 4-h inhalation)	>3.34	mg/L air	

Rabbit: moderately irritating to skin [174, p. 41]
Rabbit: non to slightly irritating to eyes [174, p. 41]
Guinea pig: a skin sensitizer
NOEL/NOAEL
Chronic.

Rat (2-year feed)	ca. 100	$mg\,kg^{-1}\,d^{-1}$	[174, p. 41]
Dog (1-year feed)	5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 41]

Dimethachlor [*50563-36-5*], [23, p. 406]

LD_{50} (rat, oral)	1600 to >2000	mg/kg
LD_{50} (rat, dermal)	>3170	mg/kg
LC_{50} (rat, 4-h inhalation)	>4.45	mg/L air

Rabbit: slightly irritating to skin and eyes
NOEL/NOAEL
Subchronic.

Dog (90-day feed)	10.4	$mg\,kg^{-1}\,d^{-1}$

Chronic.

Rat (2-year feed)	12	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	31.8	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.12\,mg\,kg^{-1}\,d^{-1}$

Metazachlor [*67129-08-2*], [23, p. 801]

LD_{50} (rat, oral)	2150	mg/kg
LD_{50} (rat, dermal)	>6810	mg/kg
LC_{50} (rat, 4-h inhalation)	>34.5	mg/L air

Rabbit: nonirritating to eyes
NOEL/NOAEL
Chronic.

Rat (2-year feed)	3.6	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	19	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.036\,mg\,kg^{-1}\,d^{-1}$

Metolachlor [*51218-45-2*], [185]

LD_{50} (rat, oral)	1200 – 2780	mg/kg	
LD_{50} (rat, dermal)	> 3170	mg/kg	[23, p. 833]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[174, p. 197]
LC_{50} (rat, 4-h inhalation)	> 1.75	mg/L air	

Rabbit: mildly irritating to skin and eyes [174, p. 197]
Guinea pig: a skin sensitisation [174, p. 197]
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	15	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 197]

Teratogenicity.

Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 197]
	300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 197]
Rabbit (gavage)	120	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 197]
	360	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 197]

Chronic.

Rat (2-year feed)	15.7	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	9.2	$mg\,kg^{-1}\,d^{-1}$	[174, p. 197]
Mouse (18-month feed)	100	$mg\,kg^{-1}\,d^{-1}$	[174, p. 197]

ADI: 0.1 $mg\,kg^{-1}\,d^{-1}$ [23, p. 833]; RfD: 0.1 $mg\,kg^{-1}\,d^{-1}$ [186]

Pretilachlor [*51218-49-6*], [187]

LD_{50} (rat, oral)	2200 – 3600	mg/kg
LD_{50} (mouse, oral)	1800 – 2300	mg/kg
LD_{50} (rat, i.p.)	1120 – 1300	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.8	mg/L air

Rabbit: moderately irritating to skin
Rabbit: slightly irritating to eyes
Nonmutagenic (*in vitro, in vivo*)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 200	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)

Chronic.

Rat (2-year feed)	1.86	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 7.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	24	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.018 $mg\,kg^{-1}\,d^{-1}$ [23, p. 995]

Propachlor [*1918-16-7*], [174, p. 250]

LD_{50} (rat, oral)	550 – 1700	mg/kg	[23]
	1800	mg/kg	
LD_{50} (mouse, oral)	290 – 306	mg/kg	[188]
LD_{50} (rabbit, dermal)	> 20 000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 1.2	mg/L air	

Rabbit: slightly irritating to skin
Rabbit: moderately [23, p. 1013] to severely irritating to eyes
Guinea pig: a skin sensitizer
Nonmutagenic (multiple assays)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	3	$mg\,kg^{-1}\,d^{-1}$ (parental)	[188]
	30	$mg\,kg^{-1}\,d^{-1}$ (developmental)	

Teratogenicity.

Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$	[188]
Rabbit (gavage)	5	$mg\,kg^{-1}\,d^{-1}$	[188]

Chronic.

Rat (2-year feed)	ca. 2.6	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	9	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	1.6	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.02 $mg\,kg^{-1}\,d^{-1}$ [188]

Propisochlor [*86763-47-5*], [23, p. 1034]

LD_{50} (rat, oral)	2088 – 3433	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5	mg/L air

NOEL/NOAEL
Subchronic.

Rat (90-day feed)	25	$mg\,kg^{-1}\,d^{-1}$

ADI: 2.5 $mg\,kg^{-1}\,d^{-1}$

Thenylchlor [*96491-05-3*], [23, p. 1182]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.67	mg/L air

NOEL/NOAEL
Sub/Chronic.

Rat	6.8	$mg\,kg^{-1}\,d^{-1}$

7.9.4. 1,3,5-Triazines

Ametryn [*834-12-8*], [189]

LD_{50} (rat, oral)	1420 – 1810	mg/kg	
LD_{50} (mouse, oral)	1370 – 1910	mg/kg	
LD_{50} (rat, dermal)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 2020	mg/kg	[174, p. 12]
LC_{50} (rat, 4-h inhalation)	> 5.17	mg/L air	

Rabbit: non [23, p. 35] to slightly [174, p. 12] irritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	ca. 1.3	$mg\,kg^{-1}\,d^{-1}$	
Teratogenicity.			
Rat (gavage)	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 250	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 60	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	2.0	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	7.2	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	1.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 12]
Mouse (102-week feed)	285	$mg\,kg^{-1}\,d^{-1}$	

ADI: 0.015 $mg\,kg^{-1}\,d^{-1}$ [23, p. 35]

Atrazine [*1912-24-9*], [174, p. 20]

LD_{50} (rat, oral)	1869 – 3090	mg/kg	
LD_{50} (mouse, oral)	> 1332 – 3992	mg/kg	[23, p. 55]
LD_{50} (rat, dermal)	> 3100	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.8	mg/L air	[23, p. 55]

Rabbit: slightly [174, p. 20] to mildly irritating to skin [23, p. 55]
Rabbit: nonirritating to eyes
Guinea pig: a skin sensitizer
Human: not a contact sensitizer [23, p. 55]
Nonmutagenic
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	ca. 3.3	$mg\,kg^{-1}\,d^{-1}$ (parental)	
Teratogenicity.			
Rat (gavage)	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
Rabbit (gavage)	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	5	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	0.7 – 3.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	3.75	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	1.4	$mg\,kg^{-1}\,d^{-1}$	[23, p. 55]
Mouse (21-month feed)	1.2	$mg\,kg^{-1}\,d^{-1}$	

ADI: 0.005 $mg\,kg^{-1}\,d^{-1}$

Cyanazine [*21725-46-2*], [174, p. 72]

LD_{50} (rat, oral)	182 – 234	mg/kg	
LD_{50} (mouse, oral)	380	mg/kg	[23, p. 280]
LD_{50} (rabbit, oral)	141	mg/kg	[23, p. 280]
LD_{50} (rat, dermal)	> 1200	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 2.46	mg/L air	[23, p. 280]

Rabbit: nonirritating to skin and eyes [23, p. 280]
Rabbit: mildly irritating to skin
Guinea pig: a skin sensitizer
Mostly nonmutagenic
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	25	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	12	$mg\,kg^{-1}\,d^{-1}$	[23]
Dog (2-year feed)	25	$mg\,kg^{-1}\,d^{-1}$	

Desmetryn [*1014-69-3*], [23, p. 351]

LD_{50} (rat, oral)	1390	mg/kg
LD_{50} (mouse, oral)	1750	mg/kg
LD_{50} (rat, dermal)	2000	mg/kg
LC_{50} (rat, 1-h inhalation)	> 1.56	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: a potential skin sensitizer
NOEL/NOAEL

Subchronic.		
Rat (90-day feed)	1.5	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	6.6	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0075 $mg\,kg^{-1}\,d^{-1}$

Dimethametryn [*22936-75-0*], [23, p. 407]

LD_{50} (rat, oral)	3000	mg/kg
LD_{50} (rat, dermal)	> 2150	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.4	mg/L air

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
NOEL/NOAEL

Subchronic.		
Rat (90-day feed)	ca. 27	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 29	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	ca. 1.25	$mg\,kg^{-1}\,d^{-1}$
Mouse (23-month feed)	ca. 4	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$

Prometon [*1610-18-0*], [174, p. 243]

LD_{50} (rat, oral)	1518 – 4345	mg/kg	
LD_{50} (mouse, oral)	2160	mg/kg	[23, p. 1010]
LD_{50} (rabbit, dermal)	> 2020	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 3.26	mg/L air	

Rabbit: slight [174, p. 243] to mildly irritating to skin [23, p. 1010]
Rabbit: mildly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (multiple assays)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	1	$mg\,kg^{-1}\,d^{-1}$ (parental)

Teratogenicity.

Rat (gavage)	< 36	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	120	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	3.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	24.5	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	1	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	5	$mg\,kg^{-1}\,d^{-1}$
Mouse (21-month feed)	70	$mg\,kg^{-1}\,d^{-1}$

Prometryn [*7287-19-6*], [174, p. 245]

LD_{50} (rat, oral)	4550	mg/kg	
LD_{50} (rat, dermal)	> 3100	mg/kg	[23, p. 1011]
LD_{50} (rabbit, dermal)	> 2020	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.17	mg/L air	

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 0.7	$mg\,kg^{-1}\,d^{-1}$

Teratogenicity.

Rat (gavage)	50	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	12	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	72	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	ca. 37.5	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	4	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 1.4	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [23, p. 1011]

Propazine [*139-40-2*], [23, p. 1024]

LD_{50} (rat, oral)	> 7000	mg/kg
LD_{50} (rat, dermal)	> 3100	mg/kg
LD_{50} (rabbit, dermal)	> 10 200	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.04	mg/L air

Rabbit: mildly irritating to skin and eyes
NOEL/NOAEL
Subchronic.

Rat (90/130-day feed)	13	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	7	$mg\,kg^{-1}\,d^{-1}$

Simazine [*122-34-9*], [174, p. 270]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	[23, p. 1106]
LD_{50} (rabbit, oral)	> 5000	mg/kg	[23, p. 1106]
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 1106]
LD_{50} (rabbit, dermal)	> 3100	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.5	mg/L air	

Rabbit: nonirritating to skin and eyes [23, p. 1106]
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (multiple assays)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 0.7	$mg\,kg^{-1}\,d^{-1}$ (parental)
	ca. 30	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	10	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	5	$mg\,kg^{-1}\,d^{-1}$

Chronic.

Rat (2-year feed)	0.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	0.7	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	5.7	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.005 $mg\,kg^{-1}\,d^{-1}$ [23, p. 1106]

Simetryn [*1014-70-6*], [190]

LD_{50} (rat, oral)	750 – 1195	mg/kg	[23, p. 1108]
LD_{50} (mouse, oral)	1600 – 1710	mg/kg	
LD_{50} (rat, i.p.)	196 – 236	mg/kg	
LD_{50} (rat, dermal)	> 3100	mg/kg	[23]
LC_{50} (rat, 4-h inhalation)	> 4.88	mg/L air	

Rabbit: slightly irritating to skin
Rabbit: nonirritating to eyes [23]
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 10	$mg\,kg^{-1}\,d^{-1}$ (parental)
	> 50	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	3	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 75	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	60	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	1.15	$mg\,kg^{-1}\,d^{-1}$	
Dog (2-year feed)	10.5	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1108]
Mouse (2-year feed)	56	$mg\,kg^{-1}\,d^{-1}$	

ADI: 0.025 $mg\,kg^{-1}\,d^{-1}$ [23, p. 1108]

Terbumeton [*33693-04-8*], [23, p. 1167]

LD_{50} (rat, oral)	651	mg/kg
LD_{50} (mouse, oral)	2343	mg/kg
LD_{50} (rat, dermal)	> 3170	mg/kg
LC_{50} (rat, 4-h inhalation)	> 10	mg/L air

Rat: nonirritating to skin
Rabbit: slightly irritating to eyes
NOEL/NOAEL

Subchronic.		
Dog (90-day feed)	25	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	7.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	25	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.075 $mg\,kg^{-1}\,d^{-1}$

Terbuthylazine [*5915-41-3*], [23, p. 1168]

LD_{50} (rat, oral)	1590 to > 2000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.3	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (2-generation feed)	0.4	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (lifetime feed)	0.22	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	0.4	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	15.4	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0022 $mg\,kg^{-1}\,d^{-1}$

Terbutryn [*886-50-0*], [23, p. 1170]

LD_{50} (rat, oral)	2500	mg/kg
LD_{50} (mouse, oral)	500	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LD_{50} (rabbit, dermal)	> 20 000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.2	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL

Chronic.		
Rat (2-year feed)	ca. 15	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 2.7	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	ca. 400	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.027 $mg\,kg^{-1}\,d^{-1}$

Trietazine [*1912-26-1*], [23, p. 1243]

LD_{50} (rat, oral)	494 – 841	mg/kg
LD_{50} (rat, dermal)	> 600	mg/kg

Rat: nonirritating to skin

Subchronic.		
Rat (90-day feed)	≈ 1	$mg\,kg^{-1}\,d^{-1}$

7.9.5. Ureas

Chlorbromuron [*13360-45-7*], [23, p. 204]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LD_{50} (rabbit, dermal)	> 10 000	mg/kg
LC_{50} (rat, 6-h inhalation)	> 1.05	mg/L air

Rabbit: nonirritating to skin and eyes
NOEL/NOAEL

Subchronic.		
Rat (90-day feed)	21	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	> 10.5	$mg\,kg^{-1}\,d^{-1}$

Chlorotoluron [*15545-48-9*], [23, p. 229]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.3	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL

Chronic.		
Rat (2-year feed)	5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	11.3	$mg\,kg^{-1}\,d^{-1}$

Daimuron [*42609-52-9*], [23, p. 330]

No information in WSSA-Herbicide Handbook, 7th Edition, 1994.

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.25	mg/L air
NOEL/NOAEL		
Reproduction.		
Rat (2-generation)	> 500	$mg\,kg^{-1}\,d^{-1}$
Subchronic.		
Rat (90-day feed)	3.1	$mg\,kg^{-1}\,d^{-1}$
Mouse (90-day feed)	6.6	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Dog (1-year feed)	≈ 0.8	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.3\,mg\,kg^{-1}\,d^{-1}$		

Dimefuron [*34205-21-5*], [23, p. 403]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	$> 10\,000$	mg/kg
LD_{50} (dog, oral)	> 2000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
Rabbit: nonirritating to skin and eyes		
NOEL/NOAEL		
Subchronic.		
Dog (90-day feed)	ca. 0.5	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	ca. 7.5	$mg\,kg^{-1}\,d^{-1}$

Diuron [*330-54-1*], [174, p. 113]

LD_{50} (rat, oral)	3400	mg/kg	[23, p. 443]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[23, p. 443]
LC_{50} (rat, 4-h inhalation)	> 5	mg/L air	[23, p. 443]

Rabbit: nonirritating to skin [23, p. 443]
Rabbit: mildly irritating to eyes [23, p. 443]
Guinea pig: not a skin sensitizer [23, p. 443]
Nonmutagenic

NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	ca. 15	$mg\,kg^{-1}\,d^{-1}$	
Teratogenicity.			
Rat (gavage)	16	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	80	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	50	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	ca. 12.5	$mg\,kg^{-1}\,d^{-1}$	[23, p. 443]
Dog (2-year feed)	ca. 4	$mg\,kg^{-1}\,d^{-1}$	[23, p. 443]

ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$ [23, p. 443]

Fenuron [*101-42-8*], [23, p. 537]

LD_{50} (rat, oral)	6400	mg/kg
Rabbit: nonirritating to skin		
NOEL/NOAEL		
Subchronic.		
Rat (90-day feed)	≈ 35	$mg\,kg^{-1}\,d^{-1}$

Fluometuron [*2164-17-2*], [174, p. 135]

LD_{50} (rat, oral)	6416	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 578]
LD_{50} (rabbit, dermal)	$> 10\,000$	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 2	mg/L air	

Rabbit: mildly irritating to skin and eyes [23, p. 578]
Guinea pig: not a skin sensitizer
Nonmutagenic (multiple assays)

NOEL/NOAEL		
Reproduction.		
Rat (2-generation feed)	ca. 0.7	$mg\,kg^{-1}\,d^{-1}$ (parental)
Teratogenicity.		
Rat (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Chronic.		
Rat (2-year feed)	19	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	10	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	1.2	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.013\,mg\,kg^{-1}\,d^{-1}$ [23, p. 578]

Isoproturon [*34123-59-6*], [23, p. 732]

LD_{50} (rat, oral)	1826 – 2417	mg/kg
LD_{50} (mouse, oral)	3350	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.95	mg/L air
Rabbit: nonirritating to skin and eyes		
NOEL/NOAEL		
Subchronic.		
Rat (90-day feed)	ca. 27	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 1.4	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	ca. 4	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.0062\,mg\,kg^{-1}\,d^{-1}$		

Isouron [*55861-78-4*], [191]

LD_{50} (rat, oral)	630 – 760	mg/kg
LD_{50} (mouse, oral)	520 – 530	mg/kg
LD_{50} (rat, i.p.)	270 – 315	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg

LC_{50} (rat, 8-h inhalation)	> 0.415	mg/L air

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes [23, p. 732], nonirritating
Guinea pig: not a skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	20	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)

Teratogenicity.

Rat (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	60	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	10	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	7.26	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	3.42	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0342 $mg\,kg^{-1}\,d^{-1}$ [23, p. 734]

Linuron [*330-55-2*], [174, p. 177]

LD_{50} (rat, oral)	1196 – 1254	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg [23, p. 751]
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 6.15	mg/L air

Rabbit: mildly irritating to skin [23, p. 751]
Rabbit: mildly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (3-generation feed)	ca. 1.7	$mg\,kg^{-1}\,d^{-1}$

Teratogenicity.

Rat (feed)	8	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	0.8	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	ca. 21	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.008 $mg\,kg^{-1}\,d^{-1}$ [23, p. 751]

Methabenzthiazuron [*18691-97-9*], [23, p. 805]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 2500	mg/kg
LD_{50} (rabbit, oral)	> 1000	mg/kg
LD_{50} (dog, oral)	> 1000	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	5.12	mg/L air

Rabbit: nonirritating to skin
NOEL/NOAEL
Chronic.

Rat (2-year feed)	ca. 7.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 21	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.05 $mg\,kg^{-1}\,d^{-1}$

Metobenzuron [*111578-32-6*], [23, p. 830]

LD_{50} (rat, oral)	> 10 000	mg/kg
LD_{50} (mouse, oral)	> 10 000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Nonmutagenic (Ames test)

Metobromuron [*3060-89-7*], [23, p. 831]

LD_{50} (rat, oral)	2603	mg/kg
LD_{50} (rat, dermal)	> 3000	mg/kg
LD_{50} (rabbit, dermal)	> 10 200	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.1	mg/L air

Rabbit: nonirritating to skin and eyes
NOEL/NOAEL
Chronic.

Rat (2-year feed)	17	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	3	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.008 $mg\,kg^{-1}\,d^{-1}$

Metoxuron [*19937-59-8*], [23, p. 838]

LD_{50} (rat, oral)	3200	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 2-week inhalation)	> 5	mg/L air

NOEL/NOAEL
Subchronic.

Rat (90-day feed)	ca. 83	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 63	$mg\,kg^{-1}\,d^{-1}$

Monolinuron [*1746-81-2*], [23, p. 851]

LD_{50} (rat, oral)	1430 – 2490	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.39	mg/L air

NOEL/NOAEL
Chronic.

Rat (2-year feed)	0.5	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.005 $mg\,kg^{-1}\,d^{-1}$

Neburon [*555-37-3*], [23, p. 872]

LD_{50} (rat, oral)	> 11 000	mg/kg	
Guinea pig: mildly irritating to skin			

Siduron [*1982-49-6*], [23, p. 1103]

LD_{50} (rat, oral)	> 7500	mg/kg	[174, p. 268]
LD_{50} (rabbit, dermal)	> 5500	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.8	mg/L air	
Rabbit: severely irritating to skin [174, p. 268]			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	ca. 25	$mg\,kg^{-1}\,d^{-1}$	
Dog (2-year feed)	ca. 63	$mg\,kg^{-1}\,d^{-1}$	

Tebuthiuron [*34014-18-1*], [174, p. 276]

LD_{50} (rat, oral)	644	mg/kg	
LD_{50} (mouse, oral)	579	mg/kg	[23, p. 1153]
LD_{50} (rabbit, oral)	286	mg/kg	[23, p. 1153]
LD_{50} (dog, oral)	> 500	mg/kg	[23, p. 1153]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
Rabbit: non [23] to slightly irritating to skin and eyes			
Guinea pig: not a skin sensitizer			
Nonmutagenic (multiple assays)			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	20	$mg\,kg^{-1}\,d^{-1}$ (parental)	
Teratogenicity.			
Rat (gavage)	80	$mg\,kg^{-1}\,d^{-1}$	
Rabbit (gavage)	25	$mg\,kg^{-1}\,d^{-1}$	
Subchronic.			
Mouse (90-day feed)	83	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	50	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	25	$mg\,kg^{-1}\,d^{-1}$	
ADI: $0.07\,mg\,kg^{-1}\,d^{-1}$ [23, p. 1153]			

7.9.6. 2,6-Dinitroanilines

Benfluralin [*1861-40-1*], [23, p. 94]

LD_{50} (rat, oral)	> 10 000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rabbit, oral)	> 2000	mg/kg
LD_{50} (dog, oral)	> 2000	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.16	mg/L air
Rabbit: moderately irritating to skin and eyes		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	≈ 50	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	6.5	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.065\,mg\,kg^{-1}\,d^{-1}$		

Butralin [*33629-47-9*], [23, p. 165]

LD_{50} (rat, oral)	1049 – 1170	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 9.35	mg/L air
Rabbit: slightly irritating to skin		
Rabbit: moderately irritating to eyes		
Guinea pig: a skin sensitizer		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	ca. 25	$mg\,kg^{-1}\,d^{-1}$
ADI (provisional): $0.5\,mg\,kg^{-1}\,d^{-1}$		

Dinitramine [*29091-05-2*], [23, p. 423]

LD_{50} (rat, oral)	3000	mg/kg
LD_{50} (rabbit, dermal)	> 6800	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.16	mg/L air
Rabbit: mildly irritating to skin and eyes		
NOEL/NOAEL		
Subchronic.		
Rat (90-day feed)	> 130	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	> 50	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	ca. 15	$mg\,kg^{-1}\,d^{-1}$

Ethalfluralin [*55283-68-6*], [174, p. 121]

LD_{50} (rat, oral)	> 10 000	mg/kg
LD_{50} (mouse, oral)	> 10 000	mg/kg
LD_{50} (dog, oral)	> 200	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 1-h inhalation)	> 0.94	mg/L air
Rabbit: slightly [174, p. 121], severely irritating to skin [23, p. 473]		
Rabbit: slightly [174, p. 121], moderately irritating to eyes [23, p. 473]		
Guinea pig: a skin sensitizer		
NOEL/NOAEL		
Reproduction.		
Rat (2-generation feed)	37.5	$mg\,kg^{-1}\,d^{-1}$ (reproduction)
Teratogenicity.		
Rat (gavage)	250	$mg\,kg^{-1}\,d^{-1}$ (maternal)

Rabbit (gavage)	75	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	
Chronic.			
Rat (2-year feed)	3.9 – 4.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	4	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	10.3	$mg\,kg^{-1}\,d^{-1}$	
ADI: $0.042\,mg\,kg^{-1}\,d^{-1}$ [23]			

Fluchloralin [*33245-39-5*], [23, p. 560]

LD_{50} (rat, oral)	1550	mg/kg	
LD_{50} (mouse, oral)	730	mg/kg	
LD_{50} (rabbit, oral)	8000	mg/kg	
LD_{50} (dog, oral)	6400	mg/kg	
LD_{50} (rabbit, dermal)	> 10 000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 8.4	mg/L air	
Rabbit: mildly irritating to skin and eyes			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	ca. 16	$mg\,kg^{-1}\,d^{-1}$	
Dog (90-day feed)	> 20	$mg\,kg^{-1}\,d^{-1}$	

Oryzalin [*19044-88-3*], [174, p. 220]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (dog, oral)	> 1000	mg/kg	[23, p. 900]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 3.1	mg/L air	
Rabbit: slightly irritating to skin			
Rabbit: nonirritating [23, p. 900] to slightly irritating [174, p. 220] to eyes			
Guinea pig: not a skin sensitizer			
Likely nonmutagenic (multiple assays)			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	112	$mg\,kg^{-1}\,d^{-1}$ (reproduction)	
Teratogenicity.			
Rat (gavage)	112	$mg\,kg^{-1}\,d^{-1}$	
Rabbit (gavage)	125	$mg\,kg^{-1}\,d^{-1}$	
Subchronic.			
Dog (90-day feed)	18.75	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	14	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	100	$mg\,kg^{-1}\,d^{-1}$	
ADI: $0.12\,mg\,kg^{-1}\,d^{-1}$ [23, p. 220]			

Pendimethalin [*40487-42-1*], [192]

LD_{50} (rat, oral)	> 10 000	mg/kg	
LD_{50} (mouse, oral)	> 12 000	mg/kg	
LD_{50} (rabbit, oral)	> 5000	mg/kg	[23, p. 937]
LD_{50} (dog, oral)	> 5000	mg/kg	[23, p. 937]
LD_{50} (rat, dermal)	> 2500	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[23, p. 937]
LC_{50} (rat, 4-h inhalation)	320	mg/L air	[174, p. 230]
Rabbit: nonirritating to skin and eyes [23, p. 937]			
Rabbit: mildly irritating to eyes [174, p. 230]			
Guinea pig: not a skin sensitizer [174, p. 230]			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (3-generation feed)	41.5	$mg\,kg^{-1}\,d^{-1}$ (parental)	
Teratogenicity.			
Rat (gavage)	500	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
Rabbit (gavage)	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[192]
	60	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 230]
Chronic.			
Rat (2-year feed)	4.3	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year capsule)	12.5	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	75	$mg\,kg^{-1}\,d^{-1}$	[174, p. 230]

Prodiamine [*29091-21-2*], [174, p. 241]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 1004]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 0.256	µg/L air	
Rabbit: nonirritating to skin			
Rabbit: mildly irritating to eyes			
Guinea pig: not a skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	ca. 13	$mg\,kg^{-1}\,d^{-1}$ (parental)	
Teratogenicity.			
Rat (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	100	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	7.2	$mg\,kg^{-1}\,d^{-1}$	[23]
	10	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	20	$mg\,kg^{-1}\,d^{-1}$	[23]
	ca. 5.7	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	60 – 70	$mg\,kg^{-1}\,d^{-1}$	

Trifluralin [*1582-09-8*], [174, p. 296]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (dog, oral)	> 2000	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 1-h inhalation)	> 4.8	mg/L air
slightly		
Rabbit: nonirritating to skin		
Rabbit: slightly irritating to eyes		
Guinea pig: a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL		
Reproduction.		
Rat (2-generation feed)	> 130	$mg\,kg^{-1}\,d^{-1}$ (reproduction)
Teratogenicity.		
Rat (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Rat (2-year feed)	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	2.4	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	73	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.024\,mg\,kg^{-1}\,d^{-1}$ [23, p. 1248]

7.9.7. Imidazolinones

Imazamethabenz-methyl [*81405-85-8*], [174, p. 159]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.8	mg/L air
Rabbit: nonirritating to skin		
Rabbit: mildly irritating to eyes		
Guinea pig: not a skin sensitizer		
NOEL/NOAEL		
Reproduction.		
Rat (2-generation feed)	93.2	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	< 250	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	250	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	750	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	250	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	12.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	6.25	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	19.5	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.0625\,mg\,kg^{-1}\,d^{-1}$ [23, p. 694]

Imazamox [*114311-32-9*], [23, p. 696]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 4000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 6.3	mg/L air
Chronic.		
Dog (1-year feed)	1165	$mg\,kg^{-1}\,d^{-1}$

Imazapic (AC 263,222) [*104098-48-4*], [23, p. 5]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	4.83	mg/L air
Rabbit: slightly irritating to skin		
Rabbit: moderately irritating to eyes		
Guinea pig: not a skin sensitizer		
Nonmutagenic		
NOEL/NOAEL		
Teratogenicity.		
Rat (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	350	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Subchronic.		
Rat (90-day feed)	1625	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.50\,mg\,kg^{-1}\,d^{-1}$

Imazapyr [*81334-34-1*], [193]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 2000	mg/kg	[23, p. 697]
LD_{50} (rabbit, oral)	4800	mg/kg	[23, p. 697]
LD_{50} (rat, i.p.)	2500 – 3200	mg/kg	[23, p. 697]
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 697]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 1.3	mg/L air	
Rabbit: nonirritating [174, p. 161] to mildly irritating to skin [193]			
Rabbit: mildly [23] [193] to severely irritating to eyes [174, p. 161]			
Guinea pig: not a skin sensitizer [174, p. 161]			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	500	$mg\,kg^{-1}\,d^{-1}$	[174, p. 161]

Teratogenicity.			
Rat (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	> 400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	500	$mg\,kg^{-1}\,d^{-1}$	[174, p. 161]
Dog (1-year feed)	250	$mg\,kg^{-1}\,d^{-1}$	[174, p. 161]
Mouse (18-month feed)	1500	$mg\,kg^{-1}\,d^{-1}$	[174, p. 161]

Imazaquin [*81335-37-7*], [174, p. 163]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	2363	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.7	mg/L air

Rabbit: mildly irritating to skin
Rabbit: nonirritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (3-generation feed)	650 – 1000	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	500	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	250	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	500	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	250	$mg\,kg^{-1}\,d^{-1}$
	500	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	25	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	150	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.25 $mg\,kg^{-1}\,d^{-1}$ [23, p. 699]

Imazethapyr [*81335-77-5*], [174, p. 166]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	[23, p. 701]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	3.27	mg/L air	

Rabbit: slightly to mildly irritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (3-generation feed)	500	$mg\,kg^{-1}\,d^{-1}$
Teratogenicity.		
Rat (gavage)	> 1125	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	375	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	500	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	25	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	750	$mg\,kg^{-1}\,d^{-1}$

7.9.8. Alkanoic Acids

Clomeprop [*84496-56-0*], [23, p. 258]

LD_{50} (rat, oral)	3250 to > 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LD_{50} (mouse, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.5	mg/L air
Chronic.		
Rat (2-year feed)	0.62	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0062 $mg\,kg^{-1}\,d^{-1}$

2,4-D acid [*94-75-7*], [194]

LD_{50} (rat, oral)	443 – 699	mg/kg	
LD_{50} (mouse, oral)	138	mg/kg	[23, p. 323]
LD_{50} (rat, dermal)	> 1600	mg/kg	[23, p. 323]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[174, p. 77]
LC_{50} (rat, 4-h inhalation)	1.79	mg/L air	[174, p. 77]

Rabbit: irritating to skin [23, p. 323]
Rabbit: severely irritating to eyes [174, p. 77]
Guinea pig: not a skin sensitizer [174, p. 77]
Nonmutagenic
NOEL/NOAEL

Neurotoxicity.		
Rat (acute and 1-year feed)	75	$mg\,kg^{-1}\,d^{-1}$
Reproduction.		
Rat (2-generation feed)	5	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)
Teratogenicity.		
Rat (gavage)	25	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 90	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	1	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	1	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	5	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ (sum of 2,4-D and of its salts and esters expressed as 2,4-D)

Dalapon [*75-99-0*], [23, p. 331]

LD_{50} (rat, oral)	7570 – 9330	mg/kg	
LD_{50} (mouse, oral)	> 4600	mg/kg	
LD_{50} (rabbit, oral)	3860	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 8-h inhalation)	> 20	mg/L air	
Rabbit: moderately irritating to skin and eyes			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	15	$mg\,kg^{-1}\,d^{-1}$	

2,4-DB [*94-82-6*], [174, p. 83]

LD_{50} (rat, oral)	1960	mg/kg	
Nonmutagenic (Ames test)			
NOEL/NOAEL			
Subchronic.			
Dog (90-day feed)	8	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	3	$mg\,kg^{-1}\,d^{-1}$	[23, p. 337]
ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$ [23, p. 337]			

Dichlorprop [*7547-66-2*], [23, p. 368]

LD_{50} (rat, oral)	825 – 1470	mg/kg	
LD_{50} (mouse, oral)	400	mg/kg	[174, p. 98]
LD_{50} (rat, dermal)	> 4000	mg/kg	
LD_{50} (mouse, dermal)	1400	mg/kg	[174, p. 98]
LC_{50} (rat, 4-h inhalation)	> 0.65	mg/L air	
Rabbit: irritating to skin and eyes			
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	5	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	3.6	$mg\,kg^{-1}\,d^{-1}$	

Dichlorprop-P [*15165-67-0*], [23, p. 370]

LD_{50} (rat, oral)	825 – 470	mg/kg
LD_{50} (rat, dermal)	> 4000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 7.4	mg/L air
Chronic.		
Rat (2-year feed)	3.6	$mg\,kg^{-1}\,d^{-1}$

Flupropanate [*756-09-2*], [23, p. 585]

LD_{50} (rat, oral)	11 900	mg/kg
LD_{50} (mouse, oral)	9600	mg/kg
LD_{50} (rat, dermal)	> 5500	mg/kg
LD_{50} (rabbit, dermal)	> 4000	mg/kg
LC_{50} (rat, 14-day inhalation)	> 1.74	mg/L air
NOEL/NOAEL		
Subchronic.		
Rat (90-day feed)	ca. 0.3	$mg\,kg^{-1}\,d^{-1}$
Chronic.		
Mouse (1-year feed)	ca. 0.9	$mg\,kg^{-1}\,d^{-1}$

Fluroxypyr [*69377-81-7*], [23, p. 597]

LD_{50} (rat, oral)	2405	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.296	mg/L air
Rabbit: nonirritating to skin		
Rabbit: mildly irritating to eyes		
Nonmutagenic		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	80	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	320	$mg\,kg^{-1}\,d^{-1}$
ADI: 0.8 $mg\,kg^{-1}\,d^{-1}$		

MCPA [*94-74-6*], [174, p. 181]

LD_{50} (rat, oral)	900 – 1160	mg/kg	
LD_{50} (mouse, oral)	800	mg/kg	
LD_{50} (rat, dermal)	> 4000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 6.36	mg/L air	[23, p. 767]
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	ca. 1.33	$mg\,kg^{-1}\,d^{-1}$	[23, p. 767]
Mouse (2-year feed)	ca. 18	$mg\,kg^{-1}\,d^{-1}$	[23, p. 767]

MCPA-thioethyl [*25319-90-8*], [23, p. 770]

LD_{50} (rat, oral)	790 – 877	mg/kg
LD_{50} (mouse, oral)	749 – 811	mg/kg
LD_{50} (mouse, dermal)	> 1500	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.044	mg/L air
Rabbit: nonirritating to skin and eyes		
Nonmutagenic		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 3	$mg\,kg^{-1}\,d^{-1}$

MCPB [*94-81-5*], [23, p. 771]

LD_{50} (rat, oral)	4700	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 1.14	mg/L air	
Rabbit: nonirritating to skin			
Rabbit: irritating to eyes			
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	ca. 6	$mg\,kg^{-1}\,d^{-1}$	

Mecoprop [*7085-19-0*], [23, p. 776]

LD_{50} (rat, oral)	930 – 1166	mg/kg	
LD_{50} (mouse, oral)	650	mg/kg	[174, p. 186]
LD_{50} (rat, dermal)	> 4000	mg/kg	
LD_{50} (rabbit, dermal)	900	mg/kg	[174, p. 186]
LC_{50} (rat, 4-h inhalation)	> 12.5	mg/L air	
Rabbit: irritating to skin			
Rabbit: highly irritating to eyes			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	4.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (90-day feed)	4	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	1.1	$mg\,kg^{-1}\,d^{-1}$	

Mecoprop-P [*16484-77-8*], [23, p. 778]

LD_{50} (rat, oral)	1050	mg/kg
LD_{50} (rat, dermal)	> 4000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.6	mg/L air
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	1.1	$mg\,kg^{-1}\,d^{-1}$

Triclopyr [*55335-06-3*], [174, p. 292]

LD_{50} (rat, oral)	577 – 713	mg/kg	
LD_{50} (rabbit, oral)	550	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 256	ppm in air	[23, p. 1237]
Rabbit: nonirritating to skin			
Rabbit: mildly irritating to eyes			
Guinea pig: a potential skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	30	$mg\,kg^{-1}\,d^{-1}$	
Subchronic.			
Dog (6-month feed)	2.5	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	3.0	$mg\,kg^{-1}\,d^{-1}$	
Mouse (22-month feed)	5.3	$mg\,kg^{-1}\,d^{-1}$	
ADI: $0.005\,mg\,kg^{-1}\,d^{-1}$ [23, p. 1237]			

7.9.9. Triazoles

Amitrole [*61-82-5*], [195]

LD_{50} (rat, oral)	> 2500	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	[174, p. 14]
LD_{50} (rat, dermal)	> 2500	mg/kg	[23, p. 40]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[174, p. 14]
LC_{50} (rat, 4-h inhalation)	> 0.439	mg/L air	
Rabbit: minimally irritating to skin [23, p. 40]			
Rabbit: mildly irritating to eyes [23, p. 40]			
Guinea pig: a potential skin sensitizer [23, p. 40]			
Mutagenicity: genotoxic potential equivocal, but nonmutagenic *in vivo*			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	0.12	$mg\,kg^{-1}\,d^{-1}$ (parental)	
	0.9	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Teratogenicity.			
Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	
Rabbit (gavage)	5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	0.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	0.29	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	1.5	$mg\,kg^{-1}\,d^{-1}$	
Hamster (18-month feed)	1	$mg\,kg^{-1}\,d^{-1}$	
Rat (2-year inhalation)	0.5	$mg\,kg^{-1}\,d^{-1}$	
ADI: $0.002\,mg\,kg^{-1}\,d^{-1}$			

Azafenidin [*68049-83-2*], [23, p. 60]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.3	mg/L air
Rabbit: nonirritating to skin and eyes		
Guinea pig: not a skin sensitizer		
NOEL/NOAEL		
Subchronic.		
Rat (90-day feed)	ca. 3.3	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 0.3	$mg\,kg^{-1}\,d^{-1}$
Mouse (90-day feed)	ca. 7	$mg\,kg^{-1}\,d^{-1}$

Cloransulam-methyl [*147150-35-4*], [23, p. 265]

No toxicological data in the public domain

7.9.10. Uracils

Bromacil acid [*314-40-9*], [197]

LD_{50} (rat, oral)	5175	mg/kg	
LD_{50} (mouse, oral)	3040	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 4.8	mg/L air	[174, p. 37]

Rabbit: moderately irritating to skin
Rabbit: mildly irritating to eyes
Guinea pig: not a skin sensitizer [174, p. 37]
Nonmutagenic [174, p. 37]
NOEL/NOAEL

Reproduction.			
Rat (3-generation feed)	ca. 12.5	$mg\,kg^{-1}\,d^{-1}$	
Teratogenicity.			
Rat (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[174, p. 37]
	75	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 37]
Rabbit (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	[174, p. 37]
Chronic.			
Rat (2-year feed)	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 37]
Dog (1-year feed)	ca. 17	$mg\,kg^{-1}\,d^{-1}$	[174, p. 37]
Mouse (18-month feed)	12.5	$mg\,kg^{-1}\,d^{-1}$	

ADI: $0.13\,mg\,kg^{-1}\,d^{-1}$ [23, p. 140]

Lenacil [*2164-08-1*], [23, p. 748]

LD_{50} (rat, oral)	> 11 000	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.2	mg/L air

Rabbit: slightly irritating to skin

Terbacil [*5902-51-2*], [197]

LD_{50} (rat, oral)	934 – 1255	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 4.4	mg/L air	[174, p. 278]

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer [174, p. 278]
Nonmutagenic (multiple assays) [174, p. 278]
NOEL/NOAEL

Reproduction.			
Rat (3-generation feed)	30	$mg\,kg^{-1}\,d^{-1}$	[174, p. 278]
Teratogenicity.			
Rat (gavage)	62.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
Rabbit (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	
Chronic.			
Rat (2-year feed)	ca. 12.5	$mg\,kg^{-1}\,d^{-1}$	
Dog (2-year feed)	ca. 1.25	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	7	$mg\,kg^{-1}\,d^{-1}$	[174, p. 278]

ADI: $0.0125\,mg\,kg^{-1}\,d^{-1}$

7.9.11. Pyridines

Clopyralid [*1702-17-6*], [174, p. 66]

LD_{50} (rat, oral)	4300	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	1.3	mg/L air	

Rabbit: slightly irritating to skin
Rabbit: severely irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (multiple assays)
NOEL/NOAEL

Reproduction.			
Rat (2-generation feed)	500	$mg\,kg^{-1}\,d^{-1}$	
Teratogenicity.			
Rat (gavage)	> 250	$mg\,kg^{-1}\,d^{-1}$	
Rabbit (gavage)	110	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Rat (2-year feed)	15	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	100	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	500	$mg\,kg^{-1}\,d^{-1}$	[23]

ADI: $0.15\,mg\,kg^{-1}\,d^{-1}$ [23, p. 260]

Dilufenican [*83164-33-4*], [23, p. 397]

LD_{50} (rat, oral)	> 2000	mg/kg
LD_{50} (mouse, oral)	> 1000	mg/kg
LD_{50} (rabbit, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.34	mg/L air

Rabbit: nonirritating to skin and eyes
Nonmutagenic (Ames test)
NOEL/NOAEL

Subchronic.		
Rat (90-day feed)	ca. 33	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$

Diflufenzopyr (BAS 654 00 H) [*109293-97-2*], [23, p. 81]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.14	mg/L air

Dithiopyr [*97886-45-8*], [198]

LD_{50} (rat, oral)	> 5000	mg/kg	[174, p. 111]
LD_{50} (mouse, oral)	> 5000	mg/kg	[23, p. 442]
LD_{50} (rat, dermal)	> 5000	mg/kg	[23, p. 442]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[174, p. 111]
LC_{50} (rat, 4-h inhalation)	> 6	mg/L air	[174, p. 111]

Rabbit: nonirritating to skin [174, p. 111]
Rabbit: slightly irritating to eyes [174, p. 111]
Guinea pig: not a skin sensitizer [174, p. 111]
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	ca. 1.5	$mg\,kg^{-1}\,d^{-1}$ (parental)
	ca. 150	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	500	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	0.36	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 0.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	ca. 0.31	$mg\,kg^{-1}\,d^{-1}$

Picloram [*1918-02-1*], [174, p. 235]

LD_{50} (rat, oral)	4012 to > 5000	mg/kg
LD_{50} (mouse, oral)	2000 – 4000	mg/kg
LD_{50} (rabbit, oral)	ca. 2000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.035	mg/L air

Rabbit: nonirritating to skin
Rabbit: moderately irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	200	$mg\,kg^{-1}\,d^{-1}$ (parental)
	1000	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	> 400	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	> 400	$mg\,kg^{-1}\,d^{-1}$

Chronic.

Rat (2-year feed)	20	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	> 175	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	1000	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.2 $mg\,kg^{-1}\,d^{-1}$ [23, p. 977]

Thiazopyr [*117718-60-2*], [174, p. 280]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 1.2	mg/L air	

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Chronic.

Rat (2-year feed)	0.36	$mg\,kg^{-1}\,d^{-1}$ (parental)	[23, p. 1185]
Dog (1-year feed)	0.5	$mg\,kg^{-1}\,d^{-1}$	[23, p. 1185]

7.9.12. Hydroxybenzoniriles

Bromofenoxim [*13181-17-4*], [23, p. 145]

LD_{50} (rat, oral)	1217	mg/kg
LD_{50} (mouse, oral)	940	mg/kg
LD_{50} (dog, oral)	> 1000	mg/kg
LD_{50} (rat, dermal)	> 3000	mg/kg
LC_{50} (rat, 6-h inhalation)	> 0.242	mg/L air

Rabbit: mildly irritating to skin and eyes
NOEL/NOAEL
Subchronic.

Dog (90-day feed)	3	$mg\,kg^{-1}\,d^{-1}$

Chronic.

Rat (2-year feed)	3.8	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.0013 $mg\,kg^{-1}\,d^{-1}$

Bromoxynil [*1689-84-5*], [174, p. 39]

LD_{50} (rat, oral)	440	mg/kg	
LD_{50} (mouse, oral)	160	mg/kg	
LD_{50} (rabbit, oral)	335	mg/kg	
LD_{50} (rdog, oral)	ca. 100	mg/kg	[23, p. 149]
LD_{50} (rat, dermal)	> 3660	mg/kg	
LD_{50} (rabbit, dermal)	3660	mg/kg	[23, p. 149]
LC_{50} (rat, 4-h inhalation)	0.41	mg/L air	[23, p. 149]

Rabbit: mildly irritating to skin and eyes
Guinea pig: not a skin sensitizer [23, p. 149]
NOEL/NOAEL

Chronic.

Rat (2-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$	[23, p. 149]

7.9.13. [Thio] Carbamates

Asulam [*3337-71-1*], [174, p. 18]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 4000	mg/kg	[23, p. 53]
LD_{50} (rabbit, oral)	> 4000	mg/kg	[23, p. 53]
LD_{50} (dog, oral)	> 4000	mg/kg	[23, p. 53]
LD_{50} (rat, dermal)	> 2000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 6-h inhalation)	> 1.8	mg/L air	

Rabbit: slightly irritating to skin
Rabbit: irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	1250	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Teratogenicity.

Rat (gavage)	> 2000	$mg\,kg^{-1}\,d^{-1}$
Rabbit (gavage)	750	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	1500	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	50	$mg\,kg^{-1}\,d^{-1}$
Dog (6-month feed)	60	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	713	$mg\,kg^{-1}\,d^{-1}$

Butylate [*2008-41-5*], [174, p. 43]

LD_{50} (rat, oral)	4660 – 5430	mg/kg	
LD_{50} (guinea pig, oral)	1660	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[23, p. 170]
LC_{50} (rat, 4-h inhalation)	4.64	mg/L air	

Rabbit: mildly irritating to skin
Rabbit: nonirritating to eyes
Guinea pig: a potential skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL

Subchronic.

Rat (90-day feed)	32	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	40	$mg\,kg^{-1}\,d^{-1}$

Chronic.

Rat (2-year feed)	50	$mg\,kg^{-1}\,d^{-1}$

Carbetamide [*16118-49-3*], [23, p. 184]

LD_{50} (rat, oral)	> 2000	mg/kg
LD_{50} (mouse, oral)	1720	mg/kg
LD_{50} (dog, oral)	900	mg/kg
LD_{50} (rabbit, dermal)	> 500	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.13	mg/L air

Rabbit: nonirritating to skin
NOEL/NOAEL

Subchronic.

Rat (90-day feed)	ca. 210	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	ca. 350	$mg\,kg^{-1}\,d^{-1}$

Chlorpropham [*101-21-3*], [23, p. 233]

LD_{50} (rat, oral)	5000 – 7500	mg/kg
LD_{50} (rabbit, oral)	5000	mg/kg

NOEL/NOAEL

Chronic.

Rat (2-year feed)	ca. 100	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	ca. 50	$mg\,kg^{-1}\,d^{-1}$

Cycloate [*1134-23-2*], [174, p. 75]

LD_{50} (rat, oral)	2275 – 3200	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	0.0047	mg/L air

Rabbit: mildly irritating to skin and eyes
Guinea pig: a potential skin sensitizer
NOEL/NOAEL

Subchronic.

Rat (90-day feed)	55	$mg\,kg^{-1}\,d^{-1}$
Dog (90-day feed)	240	$mg\,kg^{-1}\,d^{-1}$

Desmedipham [*13684-56-5*], [174, p. 87]

LD_{50} (rat, oral)	> 10 250	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	[23, p. 349]
LD_{50} (rabbit, dermal)	> 4000	mg/kg	[23, p. 349]
LC_{50} (rat, 4-h inhalation)	> 7.4	mg/L air	[23, p. 349]

Guinea pig: not a skin sensitizer [23, p. 349]
NOEL/NOAEL

Chronic.

Rat (2-year feed)	∼ 3.2	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	22 – 25	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.00125 $mg\,kg^{-1}\,d^{-1}$ [23, p. 349]

Dimepiperate [*61432-55-1*], [23, p. 404]

LD_{50} (rat, oral)	946 – 959	mg/kg	
LD_{50} (mouse, oral)	4519 – 4677	mg/kg	
LD_{50} (rat, dermal)	> 5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 1.66	mg/L air	
Rabbit: nonirritating to skin and eyes			
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	0.1	$mg\,kg^{-1}\,d^{-1}$	
ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$			

EPTC [*759-94-4*], [174, p. 119]

LD_{50} (rat, oral)	1652	mg/kg	
LD_{50} (mouse, oral)	3160	mg/kg	
LD_{50} (rabbit, oral)	2640	mg/kg	
LD_{50} (rat, dermal)	3200	mg/kg	
LD_{50} (rabbit, dermal)	5000	mg/kg	
LC_{50} (rat, 4-h inhalation)	3.8 – 4.3	mg/L air	[23, p. 466]
Rabbit: mild-moderately irritating to skin and eyes			
Guinea pig: a weak skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Teratogenicity.			
Rat (gavage)	> 300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
Subchronic.			
Rat (90-day feed)	16	$mg\,kg^{-1}\,d^{-1}$	
Dog (90-day feed)	20	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Mouse (2-year feed)	20	$mg\,kg^{-1}\,d^{-1}$	

Esprocarb [*85787-20-2*], [23, p. 472]

LD_{50} (rat, oral)	3700	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 4.0	mg/L air	
Rabbit: mildly irritating to skin and eyes			
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	1.1	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	1.0	$mg\,kg^{-1}\,d^{-1}$	

Molinate [*2212-67-1*], [174, p. 205]

LD_{50} (rat, oral)	720	mg/kg	
LD_{50} (mouse, oral)	795	mg/kg	
LD_{50} (rat, dermal)	4000 – 4800	mg/kg	[23, p. 847]
LD_{50} (rabbit, dermal)	> 4640	mg/kg	[23, p. 847]
LC_{50} (rat, 4-h inhalation)	1.36	mg/L air	[23, p. 847]
Rabbit: mildly irritating to skin			
Rabbit: severely irritating to eyes			
Guinea pig: a skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	< 3.2	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	
Subchronic.			
Rat (90-day feed)	8	$mg\,kg^{-1}\,d^{-1}$	
Dog (90-day feed)	20	$mg\,kg^{-1}\,d^{-1}$	

Orbencarb [*34622-58-7*], [23, p. 898]

LD_{50} (rat, oral)	800 – 820	mg/kg	
LD_{50} (mouse, oral)	935 – 1010	mg/kg	
LD_{50} (rat, dermal)	> 10 000	mg/kg	
LC_{50} (rat, 4-h inhalation)	2.94 – 4.32	mg/L air	
Rabbit: nonirritating to skin and eyes			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	1.7	$mg\,kg^{-1}\,d^{-1}$	

Pebulate [*1114-71-2*], [174, p. 229]

LD_{50} (rat, oral)	1675	mg/kg	
LD_{50} (mouse, oral)	1529 – 1784	mg/kg	
LD_{50} (rabbit, dermal)	4640	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 3.5	mg/L air	[23, p. 930]
Rabbit: mildly irritating to skin and eyes			
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			
Subchronic.			
Rat (90-day feed)	16	$mg\,kg^{-1}\,d^{-1}$	
Dog (90-day feed)	20	$mg\,kg^{-1}\,d^{-1}$	
Chronic.			
Dog (1-year feed)	5	$mg\,kg^{-1}\,d^{-1}$	[23, p. 930]

Phenmedipham [*13684-63-4*], [23, p. 948]

LD_{50} (rat, oral)	> 8000	mg/kg	[174, p. 233]
LD_{50} (mouse, oral)	> 8000	mg/kg	[174, p. 233]
LD_{50} (guinea pig, oral)	> 4000	mg/kg	[174, p. 233]
LD_{50} (dog, oral)	> 4000	mg/kg	[174, p. 233]
LD_{50} (rat, i.p.)	> 5000	mg/kg	
LD_{50} (rat, dermal)	2500	mg/kg	
LD_{50} (rabbit, dermal)	1000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 7.0	mg/L air	
Guinea pig: not a skin sensitizer			
NOEL/NOAEL			

Subchronic.
Rat (90-day feed) ca. 13 mg kg^{-1} d^{-1}
Dog (90-day feed) ca. 5 mg kg^{-1} d^{-1}
Chronic.
Rat (2-year feed) ca. 5 mg kg^{-1} d^{-1}
Dog (2-year feed) ca. 2.8 mg kg^{-1} d^{-1}
ADI: 0.03 mg kg^{-1} d^{-1}

Prosulfocarb [*52888-80-9*], [23, p. 1039]

LD_{50} (rat, oral) 1820 – 1958 mg/kg
LD_{50} (rabbit, dermal) > 2000 mg/kg
LC_{50} (rat, 4-h inhalation) > 4.7 mg/L air
Rabbit: slightly irritating to skin
Rabbit: mildly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (Ames test)
NOEL/NOAEL
Subchronic.
Rat (90-day feed) 1 – 10 mg kg^{-1} d^{-1}
Dog (90-day feed) 1 – 10 mg kg^{-1} d^{-1}
Chronic.
Rat (2-year feed) 0.5 mg kg^{-1} d^{-1}
Mouse (18-month feed) > 65 mg kg^{-1} d^{-1}

Pyributicarb [*88678-67-5*], [23, p. 1060]

LD_{50} (rat, oral) > 5000 mg/kg
LD_{50} (mouse, oral) > 5000 mg/kg
LD_{50} (rat, dermal) > 5000 mg/kg
LC_{50} (rat, 4-h inhalation) > 6.52 mg/L air
Rabbit: slightly irritating to skin
Rabbit: nonirritating to eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL
Chronic.
Rat (2-year feed) 0.75 mg kg^{-1} d^{-1}

Thiobencarb [*28249-77-6*], [174, p. 284]

LD_{50} (rat, oral) 1033 – 1130 mg/kg
LD_{50} (mouse, oral) 2745 mg/kg
LD_{50} (rat, dermal) > 2000 mg/kg
LD_{50} (rabbit, dermal) > 2000 mg/kg [23, p. 1192]
LC_{50} (rat, 1-h inhalation) > 3 mg/L air
Rabbit: mildly irritating to skin
Rabbit: moderately irritating to eyes
Nonmutagenic [23, p. 1192]
NOEL/NOAEL
Reproduction.
Rat (2-generation feed) < 2 mg kg^{-1} d^{-1}
Teratogenicity.
Rat (gavage) 25 mg kg^{-1} d^{-1}
Rabbit (gavage) 100 mg kg^{-1} d^{-1}
Chronic.
Rat (2-year feed) 0.9 – 1.0 mg kg^{-1} d^{-1}
Dog (1-year feed) 1.0 mg kg^{-1} d^{-1}
Mouse (18-month feed) 25 mg kg^{-1} d^{-1}
ADI: 0.009 mg kg^{-1} d^{-1} [23, p. 1192]

Triallate [*2303-17-5*], [174, p. 285]

LD_{50} (rat, oral) 1100 mg/kg
LD_{50} (rabbit, dermal) 8200 mg/kg
LC_{50} (rat, 12-h inhalation) > 5.3 mg/L air
Rabbit: slightly [23] to moderately irritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic (*in vivo*, but positive responses *in vitro*)
NOEL/NOAEL
Reproduction.
Rat (2-generation feed) 7.5 mg kg^{-1} d^{-1} (developmental)
Teratogenicity.
Rat (gavage) 30 mg kg^{-1} d^{-1} (developmental)
Rabbit (gavage) 15 mg kg^{-1} d^{-1} (developmental)
Chronic.
Rat (2-year feed) 2.5 mg kg^{-1} d^{-1}
Dog (1-year feed) 2.5 mg kg^{-1} d^{-1}
Mouse (2-year feed) 4 mg kg^{-1} d^{-1}
Hamster (22-month feed) 18 mg kg^{-1} d^{-1}
ADI: 0.025 mg kg^{-1} d^{-1} [23, p. 1200]

Vernolate [*1929-77-7*], [174, p. 202]

LD_{50} (rat, oral) 1200 – 1900 mg/kg
LD_{50} (rabbit, dermal) 4640 mg/kg
LC_{50} (rat, 4-h inhalation) > 5 mg/L air [23, p. 1264]
Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Mutagenicity: Essentially non-genotoxic
NOEL/NOAEL
Reproduction.
Rat (2-generation feed) 1 mg kg^{-1} d^{-1} (maternal and developmental) [199]
25 mg kg^{-1} d^{-1} [199]
Teratogenicity.
Rabbit (gavage) > 200 mg kg^{-1} d^{-1} (maternal and developmental) [199]
Subchronic.
Dog (90-day feed) 38 mg kg^{-1} d^{-1}
Chronic.
Rat (51-week feed) 5 mg kg^{-1} d^{-1} [23, p. 1264]
Mouse (2-year feed) 100 mg kg^{-1} d^{-1} [23, p. 1264]

7.9.14. Organophosphorous Compounds

Anilofos [*64249-01-0*], [23, p. 47]

LD_{50} (rat, oral)	472 – 830	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	26	mg/L air	

Rabbit: slightly irritating to skin and eyes

Bensulide [*741-58-2*], [23, p. 106]

LD_{50} (rat, oral)	270 – 360	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	
LD_{50} (rabbit, dermal)	3950	mg/kg	[174, p. 31]
LC_{50} (rat, 4-h inhalation)	> 1.75	mg/L air	

Rabbit: mildly irritating to skin and eyes
Guinea pig: not a skin sensitizer
NOEL/NOAEL
Subchronic.

Rat (90-day feed)	25	$mg\,kg^{-1}\,d^{-1}$	[174, p. 31]
Dog (90-day feed)	2.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 31]
Mouse (90-day feed)	30	$mg\,kg^{-1}\,d^{-1}$	

Bilanafos [*35597-43-4*], [23, p. 120]

LD_{50} (rat, oral)	268 – 404	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg

Rabbit: nonirritating to skin and eyes
Nonmutagenic (Ames test)

Butamifos [*36335-67-8*], [23, p. 161]

LD_{50} (rat, oral)	845 – 1070	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.2	mg/L air

Rabbit: nonirritating to skin and eyes

Glufosinate-ammonium [*77182-82-2*], [200]

LD_{50} (rat, oral)	1510 – 2000	mg/kg
LD_{50} (mouse, oral)	416 – 464	mg/kg
LD_{50} (dog, oral)	> 200 – 400	mg/kg
LD_{50} (rat, i.p.)	83 – 204	mg/kg
LD_{50} (rat, dermal)	ca. 4000	mg/kg
LD_{50} (rabbit, dermal)	ca. 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 0.621	mg/L air (aerosol)
	1.26 – 2.6	mg/L air (dust)

Rabbit: nonirritating to skin and eyes [23, p. 643], slightly irritating to skin [174, p. 147]
Nonmutagenic
Not neurotoxic (hens)
Reproduction.

Rat (2-generation feed)	6	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)

Teratogenicity.

Rat (gavage)	2.2	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	2	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	6.3	$mg\,kg^{-1}\,d^{-1}$ (developmental)

NOEL/NOAEL
Chronic.

Rat (2-year feed)	2	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	4.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	11	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$

Glyphosate [*1071-83-6*], [201]

LD_{50} (rat, oral)	5600	mg/kg	[174, p. 149]
LD_{50} (mouse, oral)	11 300	mg/kg	[23, p. 646]
LD_{50} (rat, dermal)	> 5010	mg/kg	[23, p. 646]
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[174, p. 149]
LD_{50} (mouse, dermal)	7030	mg/kg	[23, p. 646]

Rabbit: nonirritating to skin [174, p. 149]
Rabbit: slightly irritating to eyes [174, p. 149]
Guinea pig: not a skin sensitizer
Nonmutagenic
Non-neurotoxic (multiple assays) [174, p. 149]
NOEL/NOAEL
Reproduction.

Rat (2-generation feed)	30	$mg\,kg^{-1}\,d^{-1}$

Teratogenicity.

Rat (gavage)	1000	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	175	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	350	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (26-month feed)	> 31	$mg\,kg^{-1}\,d^{-1}$	
Rat (2-year feed)	400	$mg\,kg^{-1}\,d^{-1}$	[174, p. 149]
Dog (1-year capsules)	500	$mg\,kg^{-1}\,d^{-1}$	
Mouse (2-year feed)	814	$mg\,kg^{-1}\,d^{-1}$	

ADI: $0.03\,mg\,kg^{-1}\,d^{-1}$

Piperophos [*24151-93-7*], [202]

LD_{50} (rat, oral)	315 – 335	mg/kg	
LD_{50} (mouse, oral)	330 – 410	mg/kg	
LD_{50} (rat, i.p.)	94 – 140	mg/kg	
LD_{50} (mouse, i.p.)	250 – 290	mg/kg	
LD_{50} (rat, dermal)	> 2150	mg/kg	[23]
	> 4000	mg/kg	
LC_{50} (rat, 1-h inhalation)	> 1.96	mg/L air	

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Guinea pig: a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (2-generation feed)	6.7	$mg\,kg^{-1}\,d^{-1}$ (parental)
	> 65	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Teratogenicity.		
Rat (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 20	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	1	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 16	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	0.37	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	0.036	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	27.6	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$

7.9.15. Cyclohexanedione Oximes

Alloxydim-sodium [*66003-55-2*], [23, p. 30]

LD_{50} (rat, oral)	2260 – 2322	mg/kg
LD_{50} (mouse, oral)	3000 – 4600	mg/kg
LD_{50} (rat, dermal)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 4.3	mg/L air

NOEL/NOAEL

Chronic.		
Rat (2-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	ca. 14	$mg\,kg^{-1}\,d^{-1}$

Clethodim [*99129-21-2*], [203]

LD_{50} (rat, oral)	1360 – 1630	mg/kg	[23, p. 250]
LD_{50} (mouse, oral)	2430 – 2570	mg/kg	
LD_{50} (rabbit, oral)	> 2000, but < 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 5000	mg/kg	[23, p. 250]
LC_{50} (rat, 4-h inhalation)	> 3.9	mg/L air	

Rabbit: slightly irritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (2-generation feed)	39	$mg\,kg^{-1}\,d^{-1}$ (parental)
	> 190	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Teratogenicity.		
Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	25	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	> 300	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year feed)	16	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year capsule)	1	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	30	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.01 $mg\,kg^{-1}\,d^{-1}$

Cycloxydim [*101205-02-1*], [204]

LD_{50} (rat, oral)	3830 – 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, i.p.)	1940 – 2370	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.28	mg/L air

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.		
Rat (multigeneration, drinking water)	10	$mg\,kg^{-1}\,d^{-1}$ (parental)
	38	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Teratogenicity.		
Rat (gavage)	200	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)
Rabbit (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)
	200	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Chronic.		
Rat (2-year water)	7	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	20	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year water)	> 32	$mg\,kg^{-1}\,d^{-1}$

ADI: 0.07 $mg\,kg^{-1}\,d^{-1}$

Sethoxydim [*74051-80-2*], [205]

LD_{50} (rat, oral)	2676 – 3500	mg/kg	
LD_{50} (mouse, oral)	5600 – 6300	mg/kg	[23, p. 1101]
LD_{50} (rat, dermal)	> 5000	mg/kg	[23, p. 1101]
LD_{50} (mouse, dermal)	> 5000	mg/kg	[23, p. 1101]
LC_{50} (rat, 4-h inhalation)	> 6.28	mg/L air	[23, p. 1101]

Rabbit: nonirritating to skin and eyes [23, p. 1101]
Guinea pig: not a skin sensitizer [23, p. 1101]
Nonmutagenic
NOEL/NOAEL

Teratogenicity.

Rat (gavage)	> 250	mg kg^{-1} d^{-1} (developmental)	
Rabbit (gavage)	160	mg kg^{-1} d^{-1} (maternal and developmental)	
Chronic.			
Rat (2-year feed)	17.2	mg kg^{-1} d^{-1}	[23, p. 1101]
Dog (1-year feed)	8.9	mg kg^{-1} d^{-1}	
Mouse (2-year feed)	18	mg kg^{-1} d^{-1}	

ADI: 0.09 mg kg^{-1} d^{-1}

Tepraloxydim (proposed name: BAS 620H) [*149979-41-9*], [23, p. 80]

No toxicological data available

Tralkoxydim [*87820-88-0*], [23, p. 1211]

LD_{50} (rat, oral)	934 – 1324	mg/kg
LD_{50} (mouse, oral)	1100 – 1231	mg/kg
LD_{50} (rabbit, oral)	> 519	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.5	mg/L air

Rabbit: mildly irritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Subchronic.

Rat (90-day feed)	ca. 0.8	mg kg^{-1} d^{-1}
Dog (90-day feed)	ca. 0.125	mg kg^{-1} d^{-1}

7.9.16. Benzonitriles

Dichlobenil [*1194-65-6*], [174, p. 94]

LD_{50} (rat, oral)	4460	mg/kg	
LD_{50} (mouse, oral)	1014 – 1621	mg/kg	
LD_{50} (rat, i.p.)	442	mg/kg	[23, p. 359]
LD_{50} (rat, dermal)	> 1000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 0.25	mg/L air	

Rabbit: nonirritating to skin and eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	ca. 3	mg kg^{-1} d^{-1} (parental)

Teratogenicity.

Rat (gavage)	180	mg kg^{-1} d^{-1}
Rabbit (gavage)	45	mg kg^{-1} d^{-1} (maternal and developmental)
Mouse (gavage)	60	mg kg^{-1} d^{-1}

Chronic.

Rat (2-year feed)	ca. 2.5	mg kg^{-1} d^{-1}
Dog (1-year feed)	ca. 1.25	mg kg^{-1} d^{-1}
Hamster (20-month feed)	ca. 10	mg kg^{-1} d^{-1}

ADI: 0.025 mg kg^{-1} d^{-1} [23, p. 359]

Ioxynil [*1689-83-4*], [174, p. 168]

LD_{50} (rat, oral)	110	mg/kg	
LD_{50} (mouse, oral)	230	mg/kg	[23, p. 718]
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 718]
LD_{50} (rabbit, dermal)	1050	mg/kg	
LC_{50} (rat, 6-h inhalation)	0.4	mg/L air (4-h LC_{50})	

Rabbit: mildly irritating to skin
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL

Reproduction.

Rat (2-generation feed)	> 10	mg kg^{-1} d^{-1}

Teratogenicity.

Rat (gavage)	5	mg kg^{-1} d^{-1}
Rabbit (gavage)	15	mg kg^{-1} d^{-1}

Chronic.

Rat (2-year feed)	0.5	mg kg^{-1} d^{-1}
Dog (30-week feed)	1	mg kg^{-1} d^{-1}
Mouse (18-month feed)	< 1.5	mg kg^{-1} d^{-1}

7.9.17. Triazinones

Hexazinone [*51235-04-2*], [206]

LD_{50} (rat, oral)	1690	mg/kg	[23, p. 677]
LD_{50} (guinea pig, oral)	860	mg/kg	[23, p. 677]
LD_{50} (dog, oral)	> 3400	mg/kg	
LC_{50} (rat, 1-h inhalation)	> 7.48	mg/L air	[23, p. 677]

Guinea pig: nonirritating to skin [23, p. 677]
Rabbit: corrosive
Guinea pig: not a skin sensitizer [174, p. 156]
Nonmutagenic (*in vivo*) [174, p. 156]
NOEL/NOAEL

Reproduction.

Rat (3-generation feed)	ca. 10	mg kg^{-1} d^{-1}	[174, p. 156]

Teratogenicity.

Rat (gavage)	100	mg kg^{-1} d^{-1} (maternal and developmental)
Rabbit (gavage)	50	mg kg^{-1} d^{-1} (maternal and developmental)

Chronic.

Rat (2-year feed)	ca. 10	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 156]
Mouse (2-year feed)	ca. 30	$mg\,kg^{-1}\,d^{-1}$	

Metamitron [*41394-05-2*], [23, p. 799]

LD_{50} (rat, oral)	ca. 1200	mg/kg
LD_{50} (mouse, oral)	ca. 650	mg/kg
LD_{50} (dog, oral)	>1000	mg/kg
LD_{50} (rat, dermal)	>4000	mg/kg
LC_{50} (rat, 4-h inhalation)	>0.33	mg/L air

Rabbit: nonirritating to skin and eyes
NOEL/NOAEL
Chronic.

Rat (2-year feed)	ca. 12.5	$mg\,kg^{-1}\,d^{-1}$
Dog (2-year feed)	ca. 2.5	$mg\,kg^{-1}\,d^{-1}$
Mouse (87-week feed)	ca. 8	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.025\,mg\,kg^{-1}\,d^{-1}$

Metribuzin [*21087-64-9*], [174, p. 200]

LD_{50} (rat, oral)	1090 – 1206	mg/kg	
LD_{50} (mouse, oral)	ca. 700	mg/kg	[23, p. 840]
LD_{50} (guinea pig, oral)	245 – 274	mg/kg	[23, p. 840]
LD_{50} (rabbit, dermal)	$>20\,000$	mg/kg	
LC_{50} (rat, 4-h inhalation)	>0.65	mg/L air	[23, p. 840]

Rabbit: nonrritating to skin and eyes [23, p. 840]
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Reproduction.

Rat (3-generation feed)	ca. 2	$mg\,kg^{-1}\,d^{-1}$ (parental)	
	20	$mg\,kg^{-1}\,d^{-1}$ (developmental)	

Teratogenicity.

Rat (gavage)	ca. <25	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	
	>135	$mg\,kg^{-1}\,d^{-1}$	

Chronic.

Rat (2-year feed)	ca. 5	$mg\,kg^{-1}\,d^{-1}$	
Dog (2-year feed)	ca. 2.8	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	ca. 114	$mg\,kg^{-1}\,d^{-1}$	[23, p. 480]

ADI: $0.013\,mg\,kg^{-1}\,d^{-1}$ [23, p. 840]

7.9.18. Miscellaneous Herbicides

Bispyribac-sodium [*125401-92-5*], [23, p. 129]

LD_{50} (rat, oral)	2635 – 4111	mg/kg
LD_{50} (rat, dermal)	>2000	mg/kg
LC_{50} (rat, 4-h inhalation)	>4.48	mg/L air

Rabbit: nonirritating to skin
Rabbit: slightly irritating to eyes
Nonmutagenic (Ames test)
NOEL/NOAEL
Chronic.

Rat (2-year feed)	1.1	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	1.7	$mg\,kg^{-1}\,d^{-1}$

ADI: $0.011\,mg\,kg^{-1}\,d^{-1}$

Butroxydim [*138164-12-2*], [23, p. 167]

LD_{50} (rat, oral)	1635 – 3476	mg/kg
LD_{50} (rat, dermal)	>2000	mg/kg
LC_{50} (rat, 4-h inhalation)	>2.99	mg/L air

Rabbit: moderately irritating to skin
Rabbit: mildly irritating to eyes
Guinea pig: not a skin sensitizer
Nonmutagenic
NOEL/NOAEL
Teratogenicity.

Rat (gavage)	5	$mg\,kg^{-1}\,d^{-1}$ (developmental)
Rabbit (gavage)	15	$mg\,kg^{-1}\,d^{-1}$ (developmental)

Chronic.

Rat (2-year feed)	2.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	5	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	10	$mg\,kg^{-1}\,d^{-1}$

Cafenstrole [*125306-83-4*], [23, p. 173]

LD_{50} (rat, oral)	>5000	mg/kg
LD_{50} (mouse, oral)	>5000	mg/kg
LD_{50} (rat, dermal)	>2000	mg/kg
LC_{50} (rat, 4-h inhalation)	>1.97	mg/L air

Nonmutagenic (Ames test)
ADI: $0.003\,mg\,kg^{-1}\,d^{-1}$

Cyclosulfamuron [*136849-15-5*], [23, p. 288]

LD_{50} (rat, oral)	> 5000	mg/kg	
LD_{50} (mouse, oral)	> 5000	mg/kg	
LD_{50} (rabbit, dermal)	> 4000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 5.2	mg/L air	
Rabbit: nonirritating to skin			
Rabbit: mildly irritating to eyes			
Nonmutagenic (Ames test)			
NOEL/NOAEL			
Chronic.			
Rat (2-year feed)	50	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	3	$mg\,kg^{-1}\,d^{-1}$	

Dicamba [*1918-00-9*] acid, [174, p. 91]

LD_{50} (rat, oral)	1707	mg/kg	[23, p. 356]
LD_{50} (mouse, oral)	1190	mg/kg	[23, p. 356]
LD_{50} (rabbit, oral)	2000	mg/kg	[23, p. 356]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[23, p. 356]
LC_{50} (rat, 4-h inhalation)	> 9.6	mg/L air	[23, p. 356]
Rabbit: moderately irritating to skin [23, p. 356]			
Rabbit: severely irritating to eyes (*corrosive*) [23, p. 356]			
Guinea pig: a potential skin sensitizer			
Nonmutagenic [23]			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	40	$mg\,kg^{-1}\,d^{-1}$ (parental)	
	400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Teratogenicity.			
Rat (gavage)	160	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	> 400	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	30	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	300	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	110	$mg\,kg^{-1}\,d^{-1}$	[23]
	125	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	52	$mg\,kg^{-1}\,d^{-1}$	[23]
	60	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	115	$mg\,kg^{-1}\,d^{-1}$	

Dimethenamid [*87674-68-8*], [207]

LD_{50} (rat, oral)	1300 – 2140	mg/kg	
LD_{50} (rat, dermal)	> 2000	mg/kg	[23, p. 409]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[174, p. 264]
LC_{50} (rat, 4-h inhalation)	> 4.99	mg/L air	[174, p. 264]
Rabbit: slightly irritating to skin			
Rabbit: mildly irritating to eyes [23, p. 409]			
Guinea pig: not a skin sensitizer [174, p. 264]			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (2-generation feed)	36	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	
	10	$mg\,kg^{-1}\,d^{-1}$ (parental)	[174, p. 264]
	25	$mg\,kg^{-1}\,d^{-1}$ (developmental)	[174, p. 264]
Teratogenicity.			
Rat (gavage)	50	$mg\,kg^{-1}\,d^{-1}$ (maternal)	[23]
	215	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	37.5	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	75	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	6	$mg\,kg^{-1}\,d^{-1}$	[23]
	5	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	9.6	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	4.5	$mg\,kg^{-1}\,d^{-1}$	[174, p. 264]
ADI/RfD: 0.05 $mg\,kg^{-1}\,d^{-1}$			

Flupoxam [*119126-15-7*], [23, p. 583]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
Rabbit: nonirritating to skin		
Rabbit: mildly irritating to eyes		
Nonmutagenic (Ames test, micronucleus assay)		

Fluthiacet-methyl [*117337-16-6*], [23, p. 606]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (rabbit, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 5.05	mg/L air
Rabbit: nonrritating to skin		
Rabbit: slightly irritating to eyes		
Nonmutagenic		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	2.1	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	0.1	$mg\,kg^{-1}\,d^{-1}$
ADI: 0.001 $mg\,kg^{-1}\,d^{-1}$ (*preliminary*)		

Fluthiamide (BAY FOE 5043) [*142459-58-3*], [23, p. 82]

LD_{50} (rat, oral)	589	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 3.74	mg/L air
Rabbit: nonrritating to skin and eyes		
Nonmutagenic (Ames test)		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	ca. 1.25	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	ca. 1.1	$mg\,kg^{-1}\,d^{-1}$
Mouse (20-month feed)	ca. 7	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.011\,mg\,kg^{-1}\,d^{-1}$		

Indanofan [*133220-30-1*], [23, p. 715]

LD_{50} (rat, oral)	> 1000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
Rabbit: nonirritating to skin		
Rabbit: mildly irritating to eyes		
Nonmutagenic (Ames test)		

Isoxaflutole [*141112-29-0*], [208]

LD_{50} (rat, oral)	> 5000	mg/kg	[23, p. 737]
LD_{50} (rabbit, dermal)	> 2000	mg/kg	[23, p. 737]
LC_{50} (rat, 4-h inhalation)	> 5.23	mg/L air	[23, p. 737]
Rabbit: nonirritating to skin [23, p. 737]			
Rabbit: minimally irritating to eyes [23, p. 737]			
Guinea pig: not a skin sensitizer [23, p. 737]			
Nonmutagenic			
NOEL/NOAEL			
Neurotoxicity.			
Rat (acute gavage)	125	mg/kg	
Rat (90-day feed)	< 25	$mg\,kg^{-1}\,d^{-1}$	
Reproduction.			
Rat (2-generation feed)	1.76	$mg\,kg^{-1}\,d^{-1}$ (parental)	
	> 437	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Teratogenicity.			
Rat (gavage)	100	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	10	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Rabbit (gavage)	20	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	< 5	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	2.0	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year feed)	44.8	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	3.2	$mg\,kg^{-1}\,d^{-1}$	
RfD: $0.002\,mg\,kg^{-1}\,d^{-1}$			

Metosulam [*139528-85-1*], [23, p. 836]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	> 5000	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 1.9	mg/L air
Guinea pig: not a skin sensitizer		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	5	$mg\,kg^{-1}\,d^{-1}$
Mouse (18-month feed)	1000	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.02\,mg\,kg^{-1}\,d^{-1}$		

Propaquizafop [*111479-05-1*], [23, p. 1021]

LD_{50} (rat, oral)	> 5000	mg/kg
LD_{50} (mouse, oral)	3009	mg/kg
LD_{50} (rat, dermal)	> 2000	mg/kg
LC_{50} (rat, 4-h inhalation)	> 2.5	mg/L air
Rabbit: nonirritating to skin		
Rabbit: moderately irritating to eyes		
Guinea pig: a potential skin sensitizer		
NOEL/NOAEL		
Chronic.		
Rat (2-year feed)	1.5	$mg\,kg^{-1}\,d^{-1}$
Dog (1-year feed)	20	$mg\,kg^{-1}\,d^{-1}$
Mouse (2-year feed)	1.5	$mg\,kg^{-1}\,d^{-1}$
ADI: $0.015\,mg\,kg^{-1}\,d^{-1}$		

Pyridate [*55512-33-9*], [24]

LD_{50} (rat, oral)	3544 – 5993	mg/kg	
LD_{50} (mouse, oral)	ca. 10 000	mg/kg	
LD_{50} (rabbit, dermal)	> 2000	mg/kg	
LC_{50} (rat, 4-h inhalation)	> 4.37	mg/L air	[23, p. 1064]
Rabbit: slight to moderately irritating to skin			
Rabbit: nonirritating to eyes			
Guinea pig: a potential skin sensitizer			
Nonmutagenic			
NOEL/NOAEL			
Reproduction.			
Rat (3-generation feed)	10.8	$mg\,kg^{-1}\,d^{-1}$ (parental and developmental)	
Teratogenicity.			
Rat (gavage)	165	$mg\,kg^{-1}\,d^{-1}$ (maternal and developmental)	
Rabbit (gavage)	300	$mg\,kg^{-1}\,d^{-1}$ (maternal)	
	600	$mg\,kg^{-1}\,d^{-1}$ (developmental)	
Chronic.			
Rat (2-year feed)	10.8	$mg\,kg^{-1}\,d^{-1}$	
Dog (1-year capsule)	20	$mg\,kg^{-1}\,d^{-1}$	
Mouse (18-month feed)	< 47.7	$mg\,kg^{-1}\,d^{-1}$, non-carcinogenic	
RfD: $0.11\,mg\,kg^{-1}\,d^{-1}$			

7.10. Economic Aspects

In 1997, herbicides accounted for 49 % of the world agrochemical market, with an end user value of US $ 14.76×10^9 [172]. When the herbicide market is considered according to its chemical classes, the amino acid derivatives made up 18 % of the total herbicide market, followed by sulfonylureas (10.8 %), triazines (8.5 %), amides (7.5 %), imidazolinones (6.7 %), aryloxyphenoxypropionates (5.4 %), bipyridyls (5.0 %), toluidines (4.8 %), ureas (4.3 %), carbamates (4.2%), hormones and diazines (3.2 and 3.4 %, respectively), diphenyl ethers and pyridines (each 2.8 %), cyclohexanediones (2.0 %), benzonitriles (1.4 %), and others (8.7 %).

7.11. Regulatory Aspects of Pesticide Residues

Because herbicides represent a powerful tool that can sometimes have detrimental as well as beneficial effects, their introduction and use must be regulated. A wide range of regulatory systems is used around the world, varying from one country to the next. Many countries, particularly the more highly developed ones, have stringent requirements related to the introduction and use of pesticides, but a complete discussion of these is beyond the scope of this article. Therefore, the following discussion will be limited to procedures applicable in the United States, representative of countries with extensive and rigorous systems.

The registration of new pesticides is administered by the Environmental Protection Agency. This agency establishes policies and requirements, and reviews environmental, toxicological, and analytical data submitted to support registration of a product. The Food and Drug Administration monitors residues of commercial pesticides that may be present in the food supply.

Many extensive studies are required to support an application for registration of a new pesticide. Those include: analytical methods; storage stability; metabolism in plants, animals, and soil; general behavior in soil, including leaching, adsorption, evaporation, photodecomposition, and degradation; identification and quantification of residues in food and feed products; interaction with activated sludge products; and an extensive array of toxicology tests. Completing all the mandatory studies requires several years and many millions of dollars.

If the product in question is to be used on food or feed, a legal tolerance must be established. The legal tolerance is the maximum amount of pesticide legally permitted on a commodity as it travels in commerce. An example of general procedure of this type is outlined in Figure 18.

From the toxicological data a reference dose is determined. While the reference dose is being determined, studies are being carried out in the field under the conditions of proposed usage. Residue studies must be conducted on each crop for which the pesticide is to be registered, and in each area of the country where usage will be allowed. Pesticide formulations are applied according to the proposed label using good agricultural practices. The agricultural products are then harvested at the appropriate time, and the crop is analyzed for evidence of the new pesticide. The proposed legal tolerance is generally set somewhat higher than the actual level of detected residues to allow for differences in weather or application technique as a way of making sure that a crop produced according to the label and good agricultural practices will not inadvertently contain residues above the legal tolerance. If no residues can be detected under the application conditions specified on the proposed label, a low legal tolerance is assigned to assure the user that a new instrument detecting tiny amounts of residue will not make such a crop illegal sometime in the future.

Dietary exposure estimates are then developed based upon the level of the proposed legal tolerance using the tolerance assessment system (TAS). This is a computerized system that includes many subgroups established according to estimates of the dietary intake of that particular crop. The subgroups might include females of childbearing age, children under one year of age, specific ethnic groups, geographical population distributions, etc. The resulting dietary exposure estimate is then added to an estimate of exposure of the pesticide in question when used on other registered crops. The total exposure from all crops taken together must not exceed the reference dose. If all these requirements are satis-

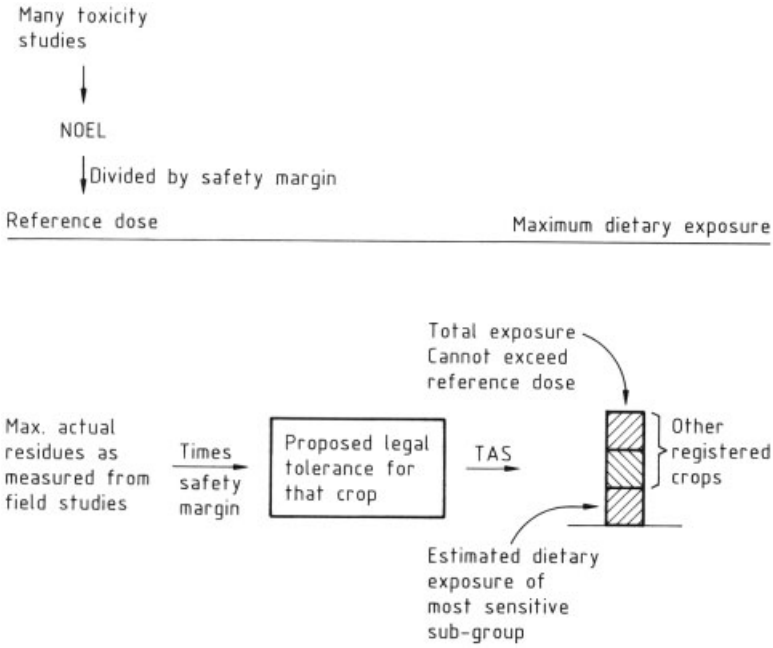

Figure 18. Schematic diagram of procedures used to establish a legal tolerance

factorily met, EPA may issue a legal tolerance for the product to be used on a given crop.

This procedure incorporates a number of safety factors. The reference dose is many-fold smaller than the most sensitive no-observed-effect level (NOEL). If residue levels in treated crops are undetectable or very small, the dietary exposure may be tens or thousands of times lower than the reference dose. In estimating dietary exposure, an assumption is made that every hectare of that particular crop in the country is to be treated with the herbicide under consideration at the maximum recommended dosage, and that every treatment will result in a residue equal to the legal tolerance. These assumptions are essentially never accurate. Moreover, residue levels are measured in raw agricultural products, but residues often are reduced by processing, washing, or cooking. All of these factors combine to provide considerable assurance that pesticide residues in the food supply will not be harmful.

However, there remain certain unavoidable weaknesses in the system. The NOEL is based upon tests on animals, not humans. It is impossible to determine a NOEL level for every possible combination of different pesticides. Some individuals might be highly sensitive to low doses of any substance, including pesticide residues.

The Departments of Agriculture in the various states become involved in monitoring and enforcing regulations on pesticides in their respective states. Pesticides are divided into those for general and restricted use. Restricted pesticides may be sold only to individuals who have received appropriate training, and who have passed an examination in pesticide usage. Consultants, government employees, and commercial applicators must receive more extensive training, and successfully complete correspondingly more rigorous examinations.

Regulations may also extend to such peripheral activities as transport, application technology, storage, and container disposal.

As a result of the large investment in both time and money associated with registration of a new pesticide, pesticides are seldom introduced unless the manufacturer can expect to acquire a significant fraction of the market associated with such extensive crops as maize and soybeans. Registrations for minor, low-volume crops are more difficult to obtain, most products certified for minor crops are ones that have already been registered in conjunction with major crops.

Another disincentive with respect to a manufacturer's registering compounds to be used on low-volume crops, especially high-value ones, is the threat of liability. If sales for a minor crop prove relatively small, even one lawsuit might cost the manufacturer more than the total profits earned from many years of sales.

The question as to whether the existing extensive requirements and regulatory measures are adequate to protect the general public depends largely on one's point of view. Some people insist on zero risk unless a compelling direct benefit can be established. Others consider the risks associated with pesticides to be small and acceptable. A number of countries in the world still lack rigorous regulatory agencies, and the corresponding risks to the environment and to the local population are considerably higher than those in much of the developed world.

8. General Comments

Controlling weeds has been a major problem since agriculture began. Besides agriculture, there are many other situations in which vegetation management is important. New techniques for controlling weeds have developed over the centuries, while few of the old techniques have been discarded. There is wide agreement among weed scientists that the use of a combination of methods is far more efficient and sustainable than dependance on only one. Relying on any one method – tillage, flooding, chemicals, even hand-weeding – can lead to a rapid shift to species that are tolerant of that method, or to the development of tolerant subtypes within a normally sensitive species.

The development of highly effective and selective herbicides, primarily since the 1940s, has resulted in an enormous improvement in the ability of vegetation managers to control unwanted plants. Weeds can be controlled faster, more cheaply, and arguably more safely than with machetes, mowers, or fire, but the use of herbicides often is not looked upon with favor.

As with any method, herbicides of course present some risks. They constitute a powerful tool, and it is impossible to say with complete assurance that they might not influence the environment in some detrimental way perhaps 20 years in the future. One cannot ever prove a negative proposition. There might certainly be some few people who are extremely sensitive to small amounts of herbicides in the environment or food. And, of course, exposure from accidents is a potential cause of physical damage, and nontarget plants may be damaged from soil residues or from drift.

Many scientists believe that the risks associated with pesticides are small and often exaggerated. Rotary mowers that can throw rocks and other projectiles a considerable distance may pose a significant hazard, but people understand this, and become accustomed to their use. On the other hand, herbicides are invisible and largely misunderstood, and very few people are aware of the safeguards already in place in the registration process. In fact, the toxicity of natural components in the food supply almost certainly exceeds by a wide margin the potential toxicity from herbicide residues.

The hazards due to herbicide use must be compared to the hazards connected with the alternative methods, including doing nothing. Doing nothing is not risk-free when one takes into account traffic obstructions, poisonous plants, and dermal and respiratory poisons in weeds, to say nothing of potential economic damage.

The most efficient weed control results from an intelligent combination of all the appropriate methods. Herbicides can play an important role in the process if used wisely.

9. References

General References

1. R. J. Aldrich: *Weed-Crop Ecology,* Breton Publ., North Scituate, Minn. 1984.
2. F. M. Ashton, T. J. Monaco: *Weed Science: Principles and Practices,* John Wiley & Sons, New York 1991.
3. Association De Coordination Technique Agricole: *Index Phytosanitaire,* Paris 1994.
4. Biologische Bundesanstalt für Land- und Forstwirtschaft Braunschweig: *Pflanzenschutzmittel-Verzeichnis* Teil I, Saphir Verlag, Ribbensbüttel 1994.
5. P. Bogar, G. Sandmann: *Target Sites of Herbicide Action.* CRC Press, Boca Raton, Fla. 1989.
6. British Crop Protection Council: *The UK Pesticide Guide,* BPCC Hazell Books Ltd., 1994.

7. M. Devine, S. O. Duke, C. Fedtke: *Physiology of Herbicide Action,* Prentice-Hall, Englewood Cliffs, N.J. 1993.
8. Industrieverband Agrar e.V. (ed.): *Wirkstoffe in Pflanzenschutz- und Schädlingsbekämpfungsmitteln.* BLV Verlagsgesellschaft, München 1990.
9. M. A. Ross, C. A. Lembi: *Applied Weed Science,* Burgess Publ. Co., Minneapolis 1985.
10. Weed Science Society of America: *Herbicide Handbook.* Champaign, Ill. 1994.

Specific References

11. W. P. Anderson: *Weed Science: Principles,* West Publ. Co., St. Paul 1983.
12. A. L. Stone: "The Weed Seed Content of Seeds," *Assoc. Off. Seed. Anal. No. Amer. Proc.* 1915, 23–27.
13. H. T. Darlington, G. T. Steinbauer: "The Eighty-Year Period for Dr. Beal's Seed Viability Experiment," *Amer. J. Bot.* **48** (1961) 321–325.
14. W. F. Libby: "Radiocarbon Dates II," *Science* **114** (1951) 291–296.
15. B. F. Kiltz: "Perennial Weeds Which Spread Vegetatively," *J. Am. Soc. Agron.* **22** (1930) 216–234.
16. I. Paczosky: "Investigations on Roots," *Bul. Appl. Bot.* **No. 8** (1914) 816–820.
17. T. K. Pavylchenko, L. E. Kirk, W. Kossan: "Eradication of Perennial Weeds by the Shallow Cultivation Method," *Univ. Saskatchewan Col. Agric., Agric. Ext. Bul.* **100** (1940) 1–8.
18. M. E. Tumbleson, T. Kommedahl: "Reproductive Potential of Cyperus esculentus by Tubers," *Weeds* **9** (1961) 646–653.
19. D. T. Bell, D. E. Keoppe, *Agron. J.* **64** (1972) 321–325.
20. Theophrastus: *Enquiry into Plants,* Translated by A. F. Hart, Loeb Classical Library, Wm. Heinemann Ltd., London 1968.
21. Pliny: *Natural History (Books 17–19),* Translated by H. Rackham, Loeb Classical Library, Wm. Heinemann Ltd., London 1971.
22. A. H. Martin et al.: *Herbicide Carryover,* Nebraska Guide 674-180, Univ. of Nebraska, Lincoln 1977.
23. C. D. S. Tomlin, *The Pesticide Manual,* 11th ed., British Crop Protection Council, Farnham, UK, 1997.
24. *US Federal Register* **63** (1998) no. 53, 837.
25. Rohm & Haas, DE 2 311 638, 1973 (H. O. Bayer, C. Swithenbank, R. Y. Yih).
26. Rhône-Poulenc, DE 2 831 262, 1980 (W. Buck, R. Sehring, G. Linden, S. Lust).
27. M. Sumida, H. Kohno, K. Shonda, H. Fukami, T. Tanaka, K. Wakabayashi, P. Böger, *J. Pesticide Sci.* **21** (1996) 317.
28. Mobil Oil, DE 2 019 821, 1970 (R. Y. Theissen).
29. T. Yoshimoto, T. Fujita, T. Harayama, *J. Pesticide Sci.* **15** (1990) 341.
30. Rohm & Haas, EP 40 898, 1981 (C. R. Downing, R. J. Theissen, R. E. Holm).
31. I. L. Adler, B. M. Jones, J. P. Wargo, *J. Agric. Food Chem.* **25** (1977) 1339.
32. Budapesti Vegyimüvek, DE 3 943 015, 1990 (J. Bakos et al.).
33. PPG, FR 2 463 119, 1981 (W. S. Grove).
34. R. Y. Yih, C. Swithenbank, *J. Agric. Food Chem.* **23** (1975) 592.
35. AgrEvo, DE 3 324 802, 1985 (L. Willims, H. Miedenberger, K. Bauer, H. Burstell, H. Bieringer).
36. Du Pont, EP 204 513, 1986 (G. Levitt).
37. Du Pont, EP 51 466, 1982 (R. F. Sauers).
38. Du Pont, EP 5 986, 1979 (G. D. Grantham, G. Levitt).
39. Du Pont, US 4 169 719, 1979 (G. Levitt).
40. Novartis, EP 44 809, 1982 (W. Meyer, W. Föry).
41. Du Pont, EP 136 061, 1985 (G. Aloysious).
42. AgrEvo, EP 342 563, 1989 (L. Willims, K. Bauer, H. Bieringer, H. Burstell).
43. Ishihara, EP 184 385, 1986 (F. Kimura et al.).
44. Du Pont, US 5 393 734, 1991 (T. A. Andrea, P. H. Liang).
45. S. Yamamoto, T. Sato, Y. Iwasawa, F. Suzuki, T. I. Kai, K. Suzuki, T. Nawamaki, *J. Pesticide Sci.* **15** (1990) 536.
46. N. Shimuzu, J. Sakamoto, H. Kamizono, K. Ohta, S. Tashiro, *J. Pesticide Sci.* **21** (1996) 287.
47. Du Pont, EP 30 139, 1981 (G. Levitt, W. C. Peterson).
48. Du Pont, EP 237 292, 1987 (M. A. Hanagan).
49. Novartis, EP 496 701, 1992 (W. Meyer).
50. Novartis, EP 72 347, 1983 (W. Meyer, W. Föry).
51. Novartis, EP 120 814, 1984 (W. Meyer, K. Örtle).
52. S. Kobayashi, *Jpn. Pesti. Info* **55** (1989) 17.
53. Du Pont, EP 13 480, 1978 (G. Levitt).
54. Du Pont, EP 13 481, 1980 (G. Levitt).
55. Takeda, EP 477 808, 1992 (Y. Ishida, K. Otha, H. Yoshikawa).
56. Du Pont, US 4 448 029, 1984 (G. Levitt).

57. Novartis, EP 44 808, 1982 (W. Meyer, W. Föry).
58. Ishihara, EP 496 608, 1992 (N. Sakashita et al.).
59. Du Pont, WO 89 09 214, 1989 (M. P. Moon).
60. Monsanto, BE 622 131, 1962 (J. F. Olin).
61. Novartis, DE 2 305 495, 1973 (C. Vogel).
62. BASF, DE 2 648 008, 1978 (K. Eicken, W. Rohr, B. Zeeh, B. Würzer).
63. P. C. Hamm, A. J. Speziale, *J. Agric. Food Chem.* **4** (1956) 518.
64. Novartis, DE 2 328 340, 1973 (C. Vogel, R. Aebischer).
65. Monsanto, US 2 863 752, 1958 (P. C. Hamm, A. J. Speziale).
66. Nitrokémia, HU 53 483, 1990 (J. Pelyva et al.).
67. Tokuyama, US 4 802 907, 1989 (T. Takematsu, S. Kato, M. Ishizaki, T. Suyama).
68. Novartis, CH 337 019, 1959 (H. Gysin, E. Knüsli).
69. Novartis, CH 329 277, 1958 (H. Gysin, E. Knüsli).
70. Deutsche Gold- und Silber-Scheideanstalt, GB 1 132 306, 1968.
71. Novartis, S. Africa 68 02 977, 1968 (E. Nikles).
72. H. Gysin, E. Knüsli, *Advance Pest. Control Res.* **3** (1960) 289.
73. W. M. Pearlman, C. K. Banks, *J. Am. Chem. Soc.* **70** (1948) 3726.
74. Novartis, GB 814 947, 1959.
75. Novartis, CH 329 277, 1958 (H. Gysin, E. Knüsli).
76. Novartis, FR 1 325 926, 1963.
77. Novartis, DE 1 905 598, 1969 (H. Martin, S. Janiak, W. Toepfl et al.).
78. Showa Denko, JP 73 35 454, (1973) (T. Takematsu et al.).
79. Rhône-Poulenc, DE 2 101 938, 1970 (R. Boesch).
80. H. C. Bucha, C. W. Todd, *Science* **114** (1951) 493.
81. Novartis, GB 914 779, 1963.
82. Novartis, GB 1 407 587, 1975.
83. Shionogi, JP 78 86 033, 1978 (H. Yikinaga, S. Sumimoto, I. Ishizuka, J. Sugita).
84. K. Härtel, *Mededel. Fac. Landbouwwet. Rijksuniv. Gent* **27** (1962) 1275.
85. Bayer, GB 1 085 430, 1967 (H. Hack, L. Eue, W. Schäfer).
86. Mitsui, WO 87 00 840, 1987 (T. Takematsu, D. Fukuoka, K. Takahashi, I. Hashimoto).
87. W. Berg, *Pflanzenkr. Pflanzenpath. Pflanzensch.* **75** (1968) 233.
88. K. Maier-Bode, K. Härtel, *Residue Rev.* **77** (1981) 1.
89. Du Pont, US 3 309 192, 1967 (R. W. Luckenbough).
90. DowElanco, DE 1 912 543, 1969 (T. Cebelo, R. Walde).
91. DowElanco, US 3 257 190, 1966 (Q. F. Soper).
92. Anchem, DE 2 058 201, 1971.
93. Wacker, DE 2 013 509, 1970 (D. L. Hunter, C. W. Lefevre, W. G. Woods, J. D. Stone).
94. DowElanco, US 3 257 190, 1966 (Q. F. Soper).
95. BASF, S. Africa 68 08 409, 1969 (K. Kiehs, K. H. König, A. Fischler).
96. DowElanco, US 3 367 949, 1968 (Q. F. Soper).
97. Cyanamid, DE 2 241 408, 1973 (A. W. Lutz, R. E. Diehl).
98. Novartis, US 3 966 816, 1976 (W. Woods, D. L. Hunter).
99. DowElanco, US 3 257 190, 1966 (Q. F. Soper).
100. Novartis, EP 233 150, 1987 (H. Szczepansk, D. Dürr).
101. J. Finn, D. Bosley, S. Rodaway, L. Quakenbush in: D. R. Baker, J. G. Feynes, G. S. Basarab (eds.): *Synthesis and Chemistry of Agrochemicals IV,* American Chemical Society, Washington, DC, 1995, p. 46.
102. D. W. Ladner, *Pestic. Sci.* **29** (1990) 317.
103. K. Ikeda, A. Goh, *Jpn. Pesticide Info* **55** (1989) 15.
104. J. Pokorny, *J. Am. Chem. Soc.* **66** (1941) 768.
105. DowElanco, US 2 642 354, 1953 (K. C. Barrons).
106. M. E. Synerholm, P. W. Zimmermann, *Contr. Boyce Thompson Inst.* **14** (1947) 369.
107. M. S. Smith, R. L. Wain, F. Wightman, *Nature* **169** (1952) 883.
108. W. O. G. Nuykken, B. H. Menck, E. Haden, *Mededel. Fac. Landbouwwet. Rijksuniv. Gent* **52** (1987) 1139.
109. H. L. Albert, J. F. Charles, *J. Am. Chem. Soc.* **76** (1954) 479.
110. DowElanco, US 4 110 104, 1978 (S. D. McGregor).
111. R. E. Shade, W. G. Templeman, W. A. Saxton, *Nature* **155** (1945) 498.
112. T. Tsuda, Y. Tamura, H. Ueda, I. Ichimoto, *J. Pesticide Sci.* **17** (1992) 261.
113. Zeneca, GB 804 505, 1958 (G. W. Kitchingman, A. C. Tucker).
114. H. J. Nestler, H. Bieringer, *Z. Naturforsch.* **85B** (1980) 366.
115. DowElanco, US 3 862 952, 1975 (L. D. Markley).
116. A. Kroller, *Residue Rev.* **12** (1966) 163.
117. Du Pont, DE 2 801 429, 1978 (A. D. Wolf).

118. DowElanco, US 5 163 995, 1995 (J. C. van Heertum, B. C. Gerwick, W. A. Kleschick, T. C. Johnson).
119. B. C. Bucha, W. E. Cuoery, J. H. Harrod, H. M. Loux, L. M. Ellis, *Science* **137** (1962) 537.
120. Du Pont, US 3 235 357, 1962.
121. Dow Elanco, US 3 317 549, 1963 (H. Johnston).
122. Rhône-Poulenc, EP 53 011, 1982 (M. C. Cramp, J. Gilmour, E. W. Parnell).
123. BASF, JP 62 45 570, 1987.
124. Monsanto, EP 133 612, 1985 (L. F. Lee).
125. DowElanco, US 3 285 925, 1966 (H. A. Johnston, M. S. Tomita).
126. Monsanto, EP 278 944, 1988 (L. F. Lee, Y. L. L. Sing).
127. Novartis, NL 6 600 834, 1966.
128. K. Carpenter, B. J. Heywood, R. L. Wain, *Nature* **200** (1963) 28.
129. H. J. Cottrell, B. J. Heywood, *Nature* **207** (1965) 655.
130. W. G. Templeman, W. A. Saxton, *Nature* **156** (1945) 630.
131. Rhône-Poulenc, FR 1 249 434, 1959 (J. Metivier).
132. Columbia, US 2 695 225, 1954 (E. D. Witman).
133. H. Tilles, *J. Am. Chem. Soc.* **81** (1959) 714.
134. AgrEvo, NL 6 604 363, 1966.
135. Zeneca, US 2 913 327, 1959 (H. Tilles, J. Antognini).
136. Zeneca, US 3 198 786, 1965 (H. Tilles, R. Curtis).
137. Zeneca, US 3 185 720, 1957 (H. Tilles, J. Antognini).
138. K. Tsuzuki, *Jpn. Pestic. Info* **57** (1990) 30.
139. H. C. Bucha, C. W. Todd, *Science* **114** (1951) 493.
140. AgrEvo, PI 78 04 448, 1979 (N. R. Patel).
141. Zeneca, US 3 205 253, 1965 (L. W. Fancher, C. L. Dewald).
142. S. Mase, *Jpn. Pestic. Info* **45** (1984) 27.
143. Sumitomo, DE 2 147 873, 1972 (T. Satomi et al.).
144. AgrEvo, JP 73 91 019, 1973 (Y. Inoue, T. Nida).
145. K. Moedritzer, R. R. Irani, *J. Org. Chem.* **31** (1966) 1603.
146. Novartis, DE 1 812 497, 1969 (E. Töpfl).
147. A. Formigomi, Y. Horono, *Mededel. Fac. Landbouwwet. Rijksuniv. Gent* **42** (1977) 1597.
148. Nippon Soda, DE 2 822 304, 1978 (I. Iwataki, H. Shibuya, H. Ishikawa, T. Kawana).
149. BASF, DE 3 121 355, 1982 (R. Becker, D. John, W. Rohr, W. Himmele, H. Siegel, B. Würster).
150. BASF, DE 4 222 261, 1993 (J. Kast et al.).
151. Zeneca, EP 80 301, 1983 (R. W. Warner et al.).
152. B. C. Sharman, *Nature* **186** (1960) 89.
153. Du Pont, US 3 897 429, 1975 (F. R. Haglid).
154. Bayer, DE 2 107 757, 1972 (K. Dickore, W. Draber, L. Eue).
155. A. Dornow, H. Manzel, P. Marx, *Ber.* **97** (1964) 2173.
156. Kumiai, EP 321 846, 1989 (N. Wada, S. Kusano, Y. Toyokawa).
157. Zeneca, WO 92 21 649, 1992 (J. D. Jones).
158. Chugai, EP 332 133, 1989 (M. Takeuchi, M. Yasuda, M. Kanzaki).
159. Cyanamid, US 5 009 699, 1991 (T. E. Brady, M. E. Condon, P-E. Marc).
160. Novartis, US 3 013 054, 1958 (S. B. Richter).
161. AgrEvo, DE 2 223 894, 1973 (W. Becker, L. H. Peter, H. Nahm, F. Schwerdtle).
162. BASF, DE 3 303 388, 1983 (K. Seckinger, F. Kuhnen, K. Milzner).
163. Kureha, EP 282 303, 1988 (T. Shida et al.).
164. Y. Sato, T. Hoshi, T. Iida, C. Ogino, B. Nicolaus, K. Wakabayashi, P. Böger, *Z. Naturforsch.* **49** (1994) 49.
165. H. Föster, R. R. Schmidt, H. J. Santel, R. Andree, *Pflanzenschutz Nachr. Bayer* **50** (1997) 105.
166. Mitsubishi, EP 398 258, 1990 (T. Jikihara et al.).
167. Rhône-Poulenc, EP 527 036, 1993 (P. A. Cain, S. M. Cramp, G. M. Little, B. M. Lascombe).
168. M. Snel, P. Watson, N. R. Gray, W. A. Kleschick, C. M. Carson, *Mededel. Fac. Landbouwwet. Rijksuniv. Gent* **58** (1993) 845.
169. Novartis, EP 52 798, 1982 (G. Frater, M. Suchy, J. Wenger, P. Winternitz).
170. Novartis, DE 2 331 398, 1975 (R. Schoenbeck, E. Kloimstein, H. Mayer, A. Diskus, E. Auer).
171. Y. Nezu, N. Wada, Y. Saitoh, S. Takahashi, T. Miyazawa, *J. Pesticide Sci.* **21** (1996) 293.
172. Wood McKenzie & Co., *Agrochemical Service,* Edinburgh, October 1998.
173. *US Federal Register* **61** (1996) 16 781.
174. W. H. Ahrens (ed.): *Herbicide Handbook,* 7th ed., Weed Science Society of America, Champaign, IL., 1994.
175. *US Ferderal Register* **61** (1996) 65 395.
176. *US EPA Chemical Fact Sheet,* No. 83, 1986.
177. Novartis Nippon, *J. Pesticide Sci.* **17** (1992) 165.
178. Takeda, *J. Pesticide Sci.* **20** (1995) 381.
179. *US Federal Register* **61** (1996) 4 399.
180. *US EPA Pesticide Fact Sheet,* No. 214, 1990.
181. *US Federal Register* **61** (1996) 16 782.

182. Du Pont Nippon, *J. Pesticide Sci.* **22** (1997) 356.
183. Du Pont Nippon, *J. Pesticide Sci.* **18** (1993) 219.
184. *US Federal Register* **53** (1988) 39 785.
185. *US EPA Chemical Fact Sheet,* No. 106, 1987.
186. *US Federal Register* **61** (1996) 60 617.
187. Novartis Nippon, *J. Pesticide Sci.* **11** (1986) 132.
188. *WHO Environmental Health Criteria* **147,** 1993.
189. Novartis Nippon, *J. Pesticide Sci.* **18** (1993) 213.
190. Nippon Nohyaku/Novartis, *J. Pesticide Sci.* **17** (1986) 315.
191. Shionogi, *J. Pesticide Sci.* **11** (1986) 131.
192. Cyanamid Nippon, *J. Pesticide Sci.* **11** (1986) 663.
193. Cyanamid Nippon, *J. Pesticide Sci.* **22** (1997) 360.
194. Pesticide Residues in Food-1996. *WHO/PCS* **97.1** (1997) 45.
195. Pesticide Residues in Food-1997. *WHO/PCS* **98.6** (1998) 21.
196. Bromacil: *US EPA Health Advisory,* 1998.
197. Terbacil: *US EPA Pesticide Fact Sheet,* No. 206, 1989.
198. Rohm & Haas Nippon, *J. Pesticide Sci.* **18** (1993) 91.
199. Vernolate: US EPA Office of Pesticides. *TOX Onliners,* 1993.
200. Pesticide Residues in Food-1991. *WHO/PCS* **92.52** (1992) 209.
201. Pesticide Residues in Food-1986. *FAO Plant Protect. Prod. Paper* **78/2** (1987) 63.
202. Novartis Nippon, *J. Pesticide Sci.* **16** (1991) 713.
203. Pesticide Residues in Food-1994. *FAO Plant Protect. Prod. Paper* **127** (1995) 45.
204. Pesticide Residues in Food-1992. *FAO Plant Protect. Prod. Paper* **126** (1993) 115.
205. *US Federal Register* **56** (1991) 11 677.
206. Hexazinone: *US EPA Pesticide Fact Sheet,* No. 183, 1988.
207. *US Federal Register* **58** (1993) 14 316.
208. *US Federal Register* **63** (1998) 50 773.

Weed Killers → Weed Control

Welding and Cutting

Heinz Günther Brandt, Messer Griesheim GmbH, Frankfurt/Main, Federal Republic of Germany

Hans Weiler, Messer Griesheim GmbH, Frankfurt/Main, Federal Republic of Germany

1. Definitions 2
2. Fundamental Principles 2
2.1. Weldability 2
2.2. Details of Welding and Cutting Processes 3
2.3. Influence of Physical and Chemical Properties 4
2.3.1. Material Properties 4
2.3.2. Design of Welded Components 5
3. Welding Processes 5
3.1. Fusion Welding 5
3.1.1. Manual Gas Welding (Autogenic Welding) 5
3.1.2. Arc Welding 7
3.1.3. Electroslag Welding 14
3.1.4. Electron Beam Welding 14
3.1.5. Laser Beam Welding 15
3.1.6. Aluminothermic Welding 16
3.2. Pressure Welding 16
3.2.1. Resistance Welding 16
3.2.2. Arc Pressure Welding 17
3.2.3. Induction Welding 18
3.3. Special Welding Processes 18
4. Fusion Cutting Processes 18
4.1. Flame Cutting (Autogenic Cutting) 18
4.2. Plasma Cutting 19
4.3. Laser Cutting 20
5. Welding of Various Materials 20
5.1. Plain Carbon Steels 20
5.2. Alloy Steels 21
5.3. Cast Steel 22
5.4. Cast Iron 22
5.5. Nickel and its Alloys 22
5.6. Copper and its Alloys 23
5.7. Aluminum and its Alloys 23
5.8. Titanium and its Alloys 24
5.9. Thermoplastic Materials 24
6. Testing and Control Methods 24
7. Health Protection and Safety 25
7.1. Health Protection 25
7.2. Safety 26
8. Economic Importance 26
9. References 26

1. Definitions

Welding is defined (DIN 1910, Part 1: 07.81 and ISO 857: 1990) as the bonding of several parts made of the same or different materials to form a single unit, where, in the weld region, either the temperature is raised to the melting point, the parts are pressed together, or both. Weld filler materials are also used as necessary. The welded joint can be separated only by destroying it. This definition applies to the welding of both metals and thermoplastic materials.

In most welding processes, the materials are in the molten state where they become joined in the weld region; in a small number of widely used processes they are in the plastic state. When the two workpieces are made of the same material, the properties of the cooled weld are similar to those of these materials. When the materials are different, the welded joint generally behaves like a mixture of the two.

Welding is also considered to include weld overlay cladding, roll cladding, and explosive cladding. Here, material that can be the same as or different from that of the workpiece is applied to an area of the surface of the workpiece by welding. Cladding with the same material is usually reserved for repair purposes, while cladding with a different material is used to protect the parent metal from wear or corrosion. The methods of welding metals can be divided into two groups:

In *fusion welding* of metals, the junction area is melted to form a pool of molten metal which consists of parent metal, either alone or with added filler metal (minus metal lost in the process). The zone surrounding the molten pool is to a greater or lesser extent affected by changes to its structure and properties, depending on the degree and duration of heating and the rate of cooling. This depends on the heat flux, which in turn is very dependent on the welding method.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_203

In *compression welding* of metals, the workpieces are pressed together at the welding region, and, in most processes, are also heated. The heat is applied by various means, e.g., electrically (resistance or induction heating), or by flames.

In *friction welding*, the heat is generated by friction between workpieces pressed together and sliding relative to each other.

In *explosive cladding* (see → Corrosion), the very high kinetic energy of detonation of the explosive material is transformed into a plastic impulse (and only to a very limited extent into heat) as the sheet of cladding material is impelled almost perpendicularly onto the parent metal. The relatively small increase in temperature, combined with the very high explosive pressure, leads to plastic deformation (which advances with the detonation front) of the thin layers brought into contact, and forms a weld with wavelike indentations.

Welding of thermoplastics (→ Plastics, Processing, Chap. 3.2.1.) is similar to pressure welding of metals. Welding temperatures are low because of the nature of the material, which is normally softened until it exhibits plastic flow. Various theories have been proposed regarding the processes that occur in the weld zone [12].

Cutting of metals is the opposite of welding. It consists in producing a gap by a fusion process. The various methods of generating the necessary heat characterize the three cutting processes:

In the *burning* process, the cutting point moves, material is heated to its ignition temperature, part of it is burnt to form metal oxide by part of the oxygen stream, and cutting follows immediately, the molten material being blown away from the gap by the rest of the oxygen.

In *plasma cutting*, electrical energy causes partial dissociation and ionization of a gas stream to form a plasma stream which then transfers most of its energy to the workpiece.

In *laser cutting*, the laser beam – monochromatic light focused to give a high energy density – is absorbed by the workpiece. Laser cutting in particular causes the material not only to melt, but also partially to vaporize. A supplementary gas stream drives the molten material out of the gap formed. Only laser cutting is suitable for cutting plastics, most of the material removed being vaporized.

2. Fundamental Principles

2.1. Weldability

A component fabricated by welding is weldable if, after fabrication, it fulfills all the functions required of it at its point of use for the duration of its expected lifetime. It is not weldable if it does not fulfill these functions as a result of any defects arising from a welding process. Weldability (see DIN 8528, Part 1: 06.73) does not depend solely on the materials used. Each welded component is made by welding together several smaller components, possibly made of different materials.

The following considerations influence the design (e.g., wall thickness, arrangement, and form of the welded joints) [13]; these are not always independent:

1) The functions of and stresses on the welded component at its operating temperature during its lifetime
2) The physical and metallurgical properties of the material used for the component (determined by its chemical composition and method of manufacture)
3) The welding method or methods and associated processes (e.g., preheating, controlled cooling during welding, and subsequent heat treatment) used to produce the welded component

One of the most important preconditions for a good welded joint is that impurities on the surface of the workpiece such as thick scale, rust, oil, and paint should be removed as completely as possible. It is certainly possible to weld over a protective coating, but this can affect the integrity of the weld.

2.2. Details of Welding and Cutting Processes

In fusion welding, pressure welding, and fusion cutting, the workpiece is heated at the point of the operation. The rate of heat flow to the workpiece being welded or cut determines the temperature changes and temperature distribution within the workpiece. This heat flux brings

the metal to its melting point where the welding is to take place. In the subsequent process of solidification, crystallites grow from the cooler regions outward (usually starting from the parent metal). This solidification structure is always surrounded by a zone of parent metal that is affected by heat. The extent of the fusion zone and heat-affected zone depend on the method used. In pressure welding, unlike fusion welding, the workpiece is not only heated, but the components are pressed together in accordance with a given pressure – force – time regime. Thus, there is not only a change in the metallic structure (recrystallization) due to heating, but a superimposed structural change due to deformation, and a characteristic "worked" structure is formed.

The metal used for filling the gap or cladding the parent metal must be compatible with both the parent metal and the welding process, must be appropriate to the form of the semifinished product used (e.g., rod, wire, or strip), and must be applied in a suitable manner, e.g., in a small number of comparatively thick welding beads or in many thin beads, in order to achieve the objective of the welding operation. The mixing effect of the welding operation may be required to affect a large or a small volume [14, 15]. Desirable metallurgical effects, such as annealing due to multiple welding passes, as well as the effects of possible reactions (usually undesirable) between the components of the workpiece in the fusion region of the weld and the surrounding gaseous atmosphere (flame gases, gases from the welding rod coating or welding powder, or the air) or the molten slag lying above it, must be considered and steered in a favorable direction. The temperature profiles and temperature changes during welding and subsequent solidification are also important. It is not usually possible to ensure satisfactory welding by using filler metal that is identical to the parent metal.

As the composition or structure (or both) of the material [16] in the weld bead is always different from that of the original parent metal, the properties of the weld bead differ from those of the parent metal. Also, in the heat-affected zones of the unmelted parent metal on both sides of the weld, changes to the structure and hence to the properties take place as a result of heating. The extent of these changes is determined by the original structure, the heating rate, the time for which temperatures are maintained, and the cooling rate. The parent metal is least affected by the welding operation if this is rapid and the volume of fused material is small. High-power welding processes therefore conserve the properties of the material better than conventional processes.

Finally, heating and melting lead to expansion, and solidification and cooling to contraction. These volume changes lead to interior stresses in the welded component if deformation is prevented, but if the interior stresses exceed the elastic limit of the material, permanent deformations are produced.

All these effects determine the performance of the welded component under operating stresses, so that, when selecting the welding process, not only the type of material and wall thickness must be considered, but also the stresses that the finished welded component is likely to undergo.

In fusion cutting, the effects on the material are similar to those caused by welding, as here, too, localized areas of the material are heated strongly and rapidly, and cooled equally rapidly. In cutting processes by both burning and plasma methods, the gap formed is swept clear by a high-velocity gas stream. The type of gas used influences the reaction in the molten metal, and sometimes also that in the surface region near the gap. Laser cutting creates a gap, even without auxiliary gases, but a higher rate of cutting metals is achieved if oxygen is used, as this oxidizes the metal and sweeps the gap clear. Nitrogen is usually used when laser cutting flammable thermoplastic materials, to extinguish the flames.

A basic rule applies to all fusion cutting processes; higher energy density processes lead to smaller heat-affected zones in the parent metal.

2.3. Influence of Physical and Chemical Properties

2.3.1. Material Properties

The chemical composition of a metallic material, including any impurities present, determines its properties and the changes to these properties that will take place on welding and cutting, and also the performance of the finished welded or cut zone when the component is put into service. The welding or cutting

method should be selected according to the melting method, the time spent at high temperature, the solidification time, etc.

The material most commonly welded is steel, and, as the properties of this material can be modified by a wide range of influences that lead to slight changes in its chemical composition, it is used here to illustrate the most important of these influences. Other metallic materials show analogous behavior. Recommendations for the welding of alloys are given in Chapter 5.

1) The percentage of carbon in the steel determines its hardening properties.
2) The principal alloying elements in the steel (chromium, nickel, molybdenum) determine the critical cooling rate and hence the extent of through tempering.
3) The number of seed nuclei (e.g., aluminum, titanium) and the extent of deformation and recrystallization determine the microstructure. This is of decisive importance for the strength and toughness properties.
4) The presence of a high concentration of seed nuclei can have a better grain refining effect on the material than the melting process.
5) The chemical composition also influences remelting and solidification, the viscosity of the melt, and the wetting properties, and hence the formation of the solidified welding pool.
6) Hot cracking is mainly caused by impurities (mainly sulfur).
7) Reaction with the environment (e.g., corrosion, scale formation, uptake of foreign gases) is also a function of the chemical composition of the material.

Other important properties include segregation behavior, the nature and formation of nonmetallic inclusions, and the anisotropy of the solid (which is increased by hot working).

The physical properties of a material also influence its weldability. These include its coefficient of thermal expansion, thermal conductivity, specific heat capacity, melting point and melting range, hardness, and toughness. These properties depend on the chemical composition and sometimes the preceding heat treatment of the workpiece. Components that have been manufactured in preparation for welding have already been subjected to heat treatments such as annealing, tempering, solution annealing, or solution annealing and hardening, and may also have been cold-hardened by forging, rolling, or drawing. These initial conditions are more or less disturbed by the welding process, especially in the region of the fusion zones and the neighboring heat-affected zones.

Most plain steels and austenitic steels can be welded without significant changes to their properties, and do not require subsequent heat treatment. Steels with good hardening properties and tempering steels are heated after the welding process to remove hard spots. Hardenable steels can also be welded if they are first preheated and then welded, slowly cooled, and annealed. Steels which can be solution hardened may regain their hardness after welding if they are welded in the solution-annealed and hardened state and subsequently aged, or if they are welded in the solution-annealed state and subsequently hardened.

If shrinkage is prevented during cooling after welding, a heat treatment is required to relieve interior stresses. The severity of these stresses is also a function of the design of the welded component and the welding sequence used in the fabrication process.

To sum up, the chemical composition of a workpiece is of critical importance for its weldability.

2.3.2. Design of Welded Components

At the planning stage, welded products, especially if they are likely to be highly stressed, should be designed so the welds are placed in areas of low stress as far as possible. Welds should be located where there is a minimum change of cross section at the junction between the workpieces to be welded. The welds should nevertheless be easily accessible. The materials of the workpieces (parent metals) should be easily welded together, and the filler metal should be similarly compatible with the parent metals. As costs of welded components are always important, compromises are often necessary.

These requirements must be matched to the welding process used, including preparation of the surfaces to be welded and of the neighboring material. The sequence of the various welds or cutting operations must also be established.

If different materials are welded together, e.g., austenite – ferrite joints, other measures are sometimes necessary, to guard against increased sensitivity to corrosion or the effect of differences in heat transfer coefficients, etc. These effects and the diffusion of individual components of a material into other materials at high service temperatures can be minimized by including intermediate welded layers. When welding pipelines in which the components have different carbon contents, intermediate welded layers (e.g., nickel alloy) are included to prevent diffusion.

Consideration should also be given at the planning stage to the necessity for later treatment of a welded component, e.g., by stress relief heat treatment (support to prevent collapse is sometimes necessary) or by pickling (draining off of pickling solution must be considered). Also, the welds or the complete welded item may need to be tested after finishing. Tests may include: approval of the combination of materials by organizations such as, e.g., the TÜV in Germany, testing of pressure vessels, determination of tensile strengths and notched-bar impact values of the welded joint, or X-ray or ultrasound testing of the weld. To ensure problem-free fabrication of welded components, the process must be suitable and the personnel concerned must be trained.

In addition to the preparation of the joint to be welded, it is also necessary to remove harmful materials on the surface (dirt, rust, and oil). Auxiliary materials must be kept dry and protected from oxidation; coated electrodes and powder for powder welding must be stored under dry conditions. These auxiliary materials may sometimes need to be dried before welding is carried out.

3. Welding Processes

Welding processes can be divided into three main groups:

1) Fusion welding (Fig. 1)
2) Pressure welding (Fig. 2 A)
3) Other processes (Fig. 2 B)

The methods most widely used in the chemical industry are gas welding, manual arc welding, various gas-shielded welding methods, and submerged arc welding, all members of the first group.

3.1. Fusion Welding

3.1.1. Manual Gas Welding (Autogenic Welding)

In manual gas welding [17, 18], the heat required is produced by burning a mixture of fuel gas and oxygen. The only fuel gas used is acetylene. The flame impinges directly on the welding area, and the filler metal is fed by hand.

There are two types of manual gas welding [1, p. 233]:

1) The filler wire is fed in the direction of welding, preceding the welding torch (leftward welding, with the torch held in the right hand)
2) The filler wire is fed in the direction of welding, following the torch (rightward welding)

Leftward welding is used for sheet metal thicknesses ≤ 3 mm, and rightward welding, which is used when penetration to the root of the joint is very important, for thicknesses >3 mm. Rightward welding produces a "key-lock" or "hook-and-eye" formation, enabling penetration to be controlled visually during welding.

In manual gas welding, the rate of heat flow to the workpiece is low compared with other welding methods, and the process is comparatively slow. For this reason, and because of the high thermal conductivity of metal, the structure of the neighboring parent metal is considerably affected. In spite of its disadvantages, ca. 10 % of all welding is by the autogenic method.

The autogenic torch is simple, lightweight, and cheap. It is much used for repairs and for simple jobs in the workshop. The gas supply consists simply of oxygen and acetylene bottles with tubing to convey these gases to the welding area. The main uses of autogenic torches are for welding pipelines in chemical plant construction, general industrial plants, shipbuilding, and heating plants, where its value depends more on the manual skill of the welder than on high melting rates.

Gas supplies for fixed welding machines for series production are usually from fixed oxygen and acetylene ring mains in which the oxygen

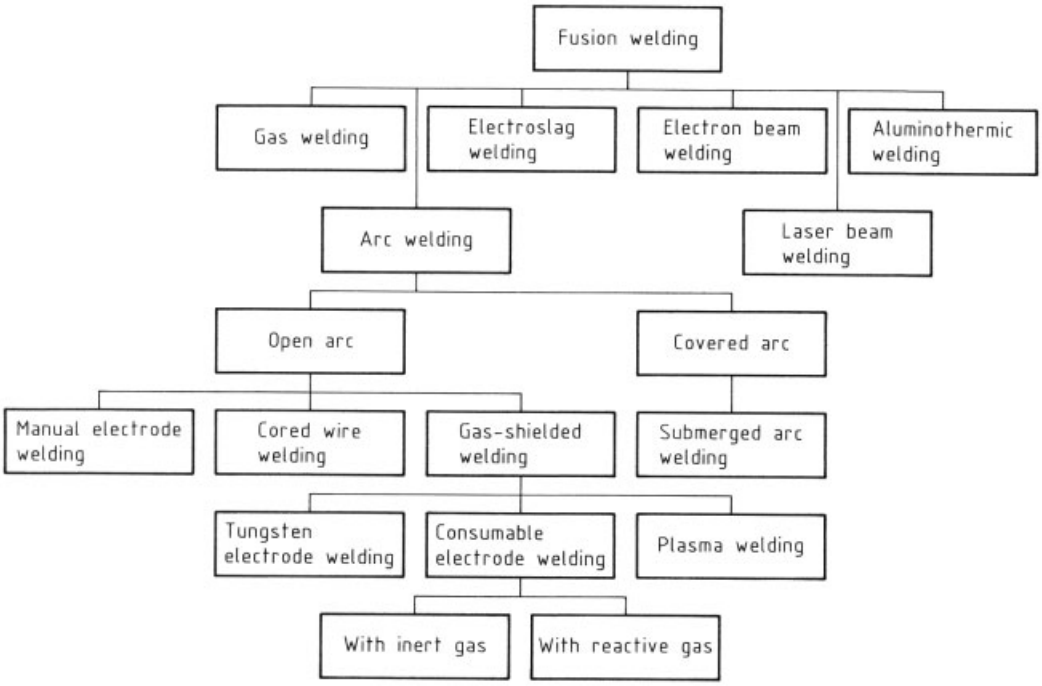

Figure 1. Fusion welding processes

is fed from banks of bottles or cryogenic gasifiers, and the acetylene from banks or bundles of bottles, or from larger transportable gas supply equipment.

Acetylene is used exclusively as the fuel gas in gas welding, as of all the hydrocarbons it has the highest heat of formation per unit mass, and gives the highest combustion temperature (Fig. 3), combustion rate, and hence heat flux and heat transfer from the flame to the workpiece. This is especially useful for the rapid heating of small surfaces. To prevent scaling and decarburization of the workpiece, the acetylene flame is adjusted to a neutral state, i.e., with no excess of either oxygen or fuel gas (volume ratio of oxygen to acetylene ca. 1.05). The rest of the oxygen required by the stoichiometry is provided by the surrounding air.

3.1.2. Arc Welding

In arc welding [19, 20], the heat source is the electric arc between the welding electrode and the workpiece. The arc may be exposed, but is then nearly always surrounded by a gas shield (see below), or it may be covered by a layer of powder (submerged arc welding). In both cases, access of air (oxygen, nitrogen, and water vapor) is restricted.

Protective gases (→ Protective Gases, Chap. 5.) [1, p. 263] are either inert (almost exclusively argon), reactive (carbon dioxide), or mixed gases with a high proportion of inert gas, e.g., argon, and 1 – 9 vol % oxygen or ≤18 vol % carbon dioxide. These not only have a protective effect, but in some cases a metallurgical effect also.

Welding powders used in the submerged arc process (DIN 8557, Part 1: 04.81 and EN 756: 1992) stabilize the arc and reduce the rate of heat loss and also have a protective effect. Metallurgical effects include deoxidation, prevention or replacement of burn-off losses, and sometimes addition of alloying elements. The powders consist mainly of oxides of silicon, manganese, calcium, barium, magnesium, titanium, sodium, potassium, and iron, and also some fluorides, e.g., of calcium and sodium/aluminum. The many types of powder can be classified according to their

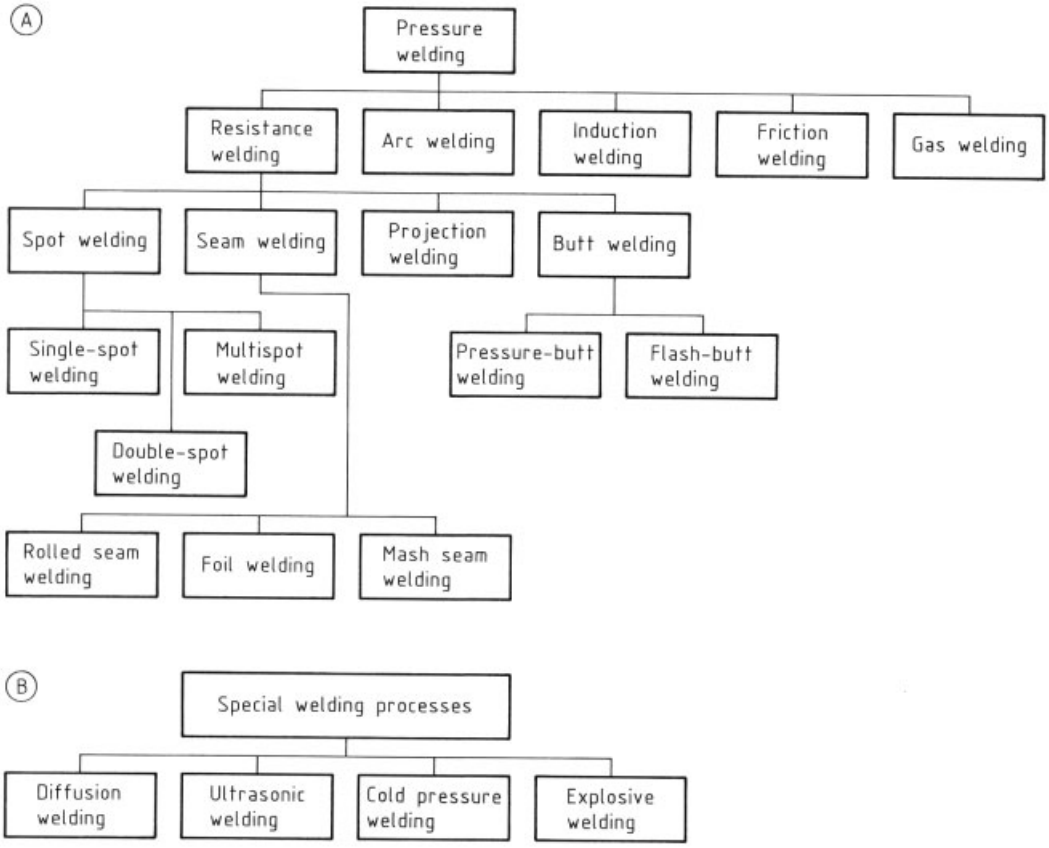

Figure 2. Pressure welding (A) and special welding (B) processes

principal components (manganese silicate type, calcium silicate type, basic aluminate type, etc.).

The electrical energy is fed to the arc from controllable and variable sources of electric current, e.g., transformers, rectifiers, or converters. Arc welding processes operate at a higher heat flow rate then gas welding, so that higher welding and melting rates can be achieved and the heat-affected zones are less extensive.

Manual Arc Welding. In manual arc welding with coated electrodes, the arc is formed between the melting end of the electrode and the workpiece, causing the coating to melt and partially vaporize. The weld pool and the molten droplets during transfer to the workpiece are protected from the surrounding air by the gases produced from the coating.

The coating also performs a number of other functions:

1) It facilitates metallurgical reactions that clean the weld pool
2) It stabilizes the arc
3) It compensates for the loss of alloying elements from the cored wire and the parent metal
4) It can introduce seed nuclei into the weld pool and hence grain refine the weld metal
5) It can control the basicity or acidity of the melt
6) It can alloy with the material of the welded joint and hence modify its chemical composition

The electrode coating [21] can form a tough skin on the weld pool. This is useful when welding in difficult situations, e.g., overhead, as it precents low-viscosity constituents of the weld pool from running off. When welding in normal situations, the tough skin improves the appearance of the surface of the weld bead.

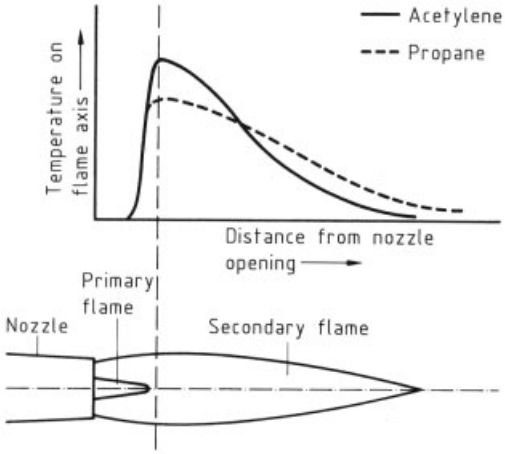

Figure 3. Temperature profile along the flame axis when burning acetylene and propane

A suitable electrode coating composition can also make the weld bead less likely to form edge notches, and the presence of reducing agents in the coating can reduce the harmful effects of partial scaling and of protective coatings on the sheet metal surface. The coating should also form a layer of slag that is easily removed after solidification, or even breaks away spontaneously.

It is often impossible for the various components of the electrode coating to perform all the many different functions desired of it.

Cored wire is usually used to achieve a composition that matches that of the alloy being welded as closely as possible. The special requirements of manual electrodes are met with the aid of the coating used. The coatings on manual welding electrodes have compositions similar to those of submerged arc welding powders, and usually contain silicon, manganese, calcium, titanium, magnesium, aluminum, potassium, and sodium, as well as gas formers such as cellulose for certain applications.

Electrodes classified according to each individual range of requirements (DIN 1913, Part 1: 04.84 and EN 499: 1992). Two extremely important types of steel electrode are:

1) The rutile (titanium dioxide) type with a high proportion of titanium dioxide in the coating, with which simple manual welding tasks can be quickly carried out
2) The basic lime (calcium carbonate) type with a high proportion of calcium carbonate in the coating, which can produce welds with the best quality characteristics, especially toughness

High-performance electrodes have coatings containing a high proportion of metal in powder form. These give a higher melting rate than normal electrodes for the same electricity consumption.

Welding with Cored Wire. Cored wire is a tubular metal electrode filled with material similar to that used for the coating on normal electrodes. Its particular advantage is that it can be fed from a long coil, and can be welded continuously. This gives a faster average welding speed than with manual electrode welding. During welding, the cored wire slides through a burner close to the welding area. Shielding gas can also be injected. The use of cored wires containing gas-producing substances can enable welding to be carried out without a gas shield under difficult conditions, e.g., in wind.

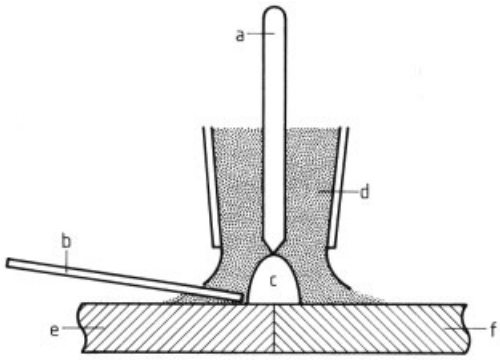

Figure 4. Principle of TIG welding
a) Tungsten electrode; b) Filler metal; c) Electric arc; d) Shielding gas; e, f) Workpieces

Tungsten Inert Gas (TIG) Welding. In these processes (Fig. 4), the arc is formed between the tip of an infusible tungsten electrode and the workpiece. The function of the electrode is only to conduct electricity and produce the arc. The filler metal is fed to the weld pool from one side, either by hand in rod form or by an automatic system as wire. Argon is usually used as the protective gas, occasionally helium or argon – helium mixture. The tungsten electrode is

alloyed with up to 4 % thorium dioxide to reduce the work of electron emission at the tip of the electrode and to increase the maximum current.

Welds of the highest quality can be produced by TIG welding. It is best to use direct current with the workpiece as the positive pole when welding high-alloy (chromium – nickel) steels, nickel-based alloys, titanium, zirconium, tantalum, and their alloys. Alternating current is used when welding metals that form high-melting oxides, e.g., aluminum and its alloys, as continuous reversal of the current direction continuously removes the oxide film from the weld pool.

Pulsed current arc welding is a variation of the TIG process. This is very useful for welding thin sheet metal, as it minimizes the extent of the heat-affected zone and hence the distortion of the sheet. The weld pool is so small that dripping of molten metal does not occur, even when welding in difficult situations. For high-quality TIG welding, transistorized pulsed current welding robots are available that can be very precisely controlled by a variety of computer programs. A pulsed current TIG weld bead is illustrated in Figure 5.

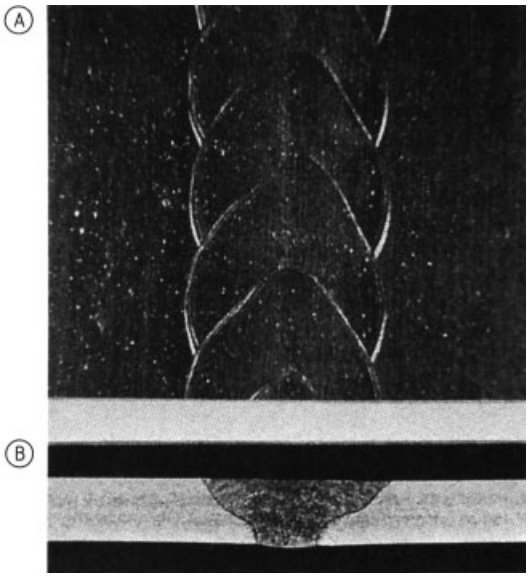

Figure 5. Weld bead produced by the TIG process
Pulsed technique; 2-mm thick Cr – Ni sheet; square butt joint without filler metal
A) Top view; B) Polished section

Plasma Welding. With the plasma arc, the gas rate is usually much higher than with the TIG arc [22–24]. Also, the temperature of the arc is increased by a factor of 2 – 4 and the energy density by a factor of ca. 10. This is achieved by greatly constricting a TIG arc together with the protective gas stream surrounding it with the aid of a water-cooled copper nozzle, forcing the entire gas stream to pass through the arc (Fig. 6).

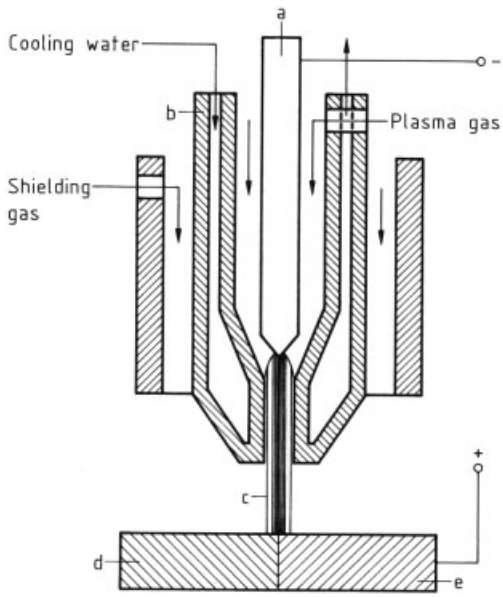

Figure 6. Principle of plasma welding
a) Electrode (negative); b) Water-cooled copper nozzle, electrically isolated from cathode; c) Plasma jet; d, e) Workpieces

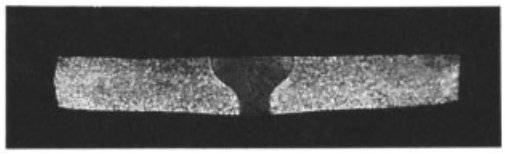

Figure 7. Polished section of weld bead produced by the plasma jet process
5-mm thick Cr – Ni sheet; square butt joint without filler material

In plasma welding, the plasma jet is always surrounded by a protective gas layer. As the plasma jet usually penetrates into the sheet, the result is a key-lock effect (Fig. 7) similar to that produced by autogenic rightward welding.

There is an upper limit to the weldable sheet thickness, but the welding speed is significantly higher than in other arc welding processes. The amount of heat applied per unit length of weld is small.

A very wide range of power outputs can be used in plasma welding, and the process can therefore be used for sheet thicknesses in the approximate range 0.01 – 10 mm. So-called microplasma welding is a low-power process using an "untransferred" arc, i.e., between the tungsten electrode and the nozzle wall. In the medium to higher power range, the plasma welding arc is "transferred," i.e., it is between the tungsten electrode and the workpiece. The usual plasma gases are argon and hydrogen.

Plasma welding is also used for resurfacing metal, which can be plated over large areas to a thickness of ca. 2 – 8 mm. It is mainly used in the construction of chemical equipment. There are two variations: plasma welding of powder coatings and plasma welding of wire.

In plasma welding of powder coatings, the powder is fed from a holding vessel to the plasma stream, where it melts. In the hot wire process, two wires are fed in parallel to the welding zone. Electric current from an auxiliary source is passed through the wires, heating them (because of their ohmic resistance) to a temperature just below their melting point close to the welding zone. The plasma burner oscillates and can therefore apply metal to a wide fusion zone.

Metal Inert Gas (MIG) Welding. In MIG welding [19], a high-power process, the arc is produced between a wire electrode, fed from a coil, and the workpiece. This method of supplying the electric current gives a high electrode current density (ca. 100 A/mm^2) and hence a high melting rate.

The wire electrode is fed to the electric arc by a motor whose speed can be varied according to the wire diameter and electric current (Fig. 8). The current to the welding wire is fed through a tubular contact in the "welding pistol" close to the arc. Only direct current is used in the MIG process, the electrode usually being the positive pole. The metal is then transferred as small droplets and penetration is deep. If the electrode is negative, larger drops of metal are transferred and there is less penetration.

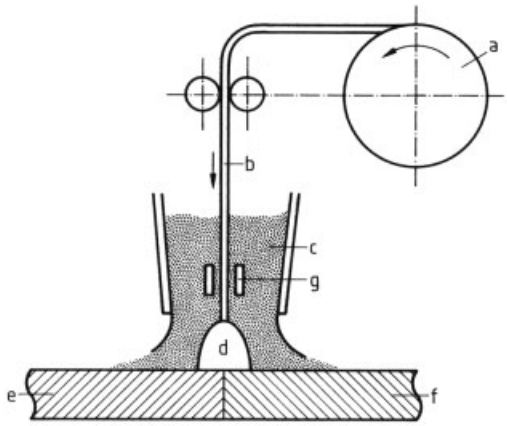

Figure 8. Principle of MIG/MAG welding
a) Coil of wire; b) Wire electrode; c) Gas shield; d) Electric arc; e, f) Workpieces; g) Electrical contact device

There are two variations of the process [25]:

1) The metal inert gas (MIG) process, in which an inactive, i.e., noble gas shield is used
2) The metal active gas (MAG) process in which the protective gas is active, e.g., carbon dioxide or mixtures or argon with carbon dioxide and/or oxygen

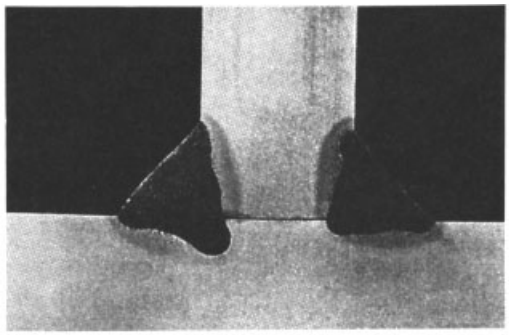

Figure 9. Polished section of a double fillet produced by the MAG process
5-mm thick St 37-2 steel plate; shielding gas 90 % argon, 5 % CO_2, 5 % O_2

The MIG process, with argon (or sometimes helium) as the protective gas, can be used to weld workpieces of thickness >3 mm made of nonferrous metals and alloy steels. With pulsed direct current of variable magnitude and pulse frequency, the transfer of droplets can be controlled. In this pulsed arc or pulsed MIG process, each pulse of current causes a droplet of metal that has been melted in the background current

Figure 10. Portable TIG welding machine (inverter, courtesy of Oerlikon, Eisenberg)
Direct or alternating current; maximum 160 A at 35 % operating duty factor

Figure 11. MIG/MAG pulsed welding machine (courtesy of Oerlikon, Eisenberg)
Maximum 400 A at 60 % operating duty factor

phase to be thrown off. The droplet size and frequency can be influenced by the type of pulse used. The pulsed MIG process is very suitable for welding in difficult situations.

The MAG process is mainly used for welding plain carbon and low-alloy steels. As the protective gas has an oxidizing effect, the filler metal must contain a high proportion of deoxidizing components (e.g., manganese, silicon, aluminum). The metallurgical reaction that takes place gives increased nucleation and hence grain-refined weld metal. A double fillet weld made by the MAG process is illustrated in Figure 9.

Sources of Electric Current. Power transistors are increasingly used for controlling the electric current. They can be used as digital switches or analog regulators. As their power losses are higher, analog regulators are used mainly for lower-power equipment. As switches (high frequency), they can be in the primary or secondary circuit, i.e., the transistors can be between the mains supply and the transformer or the transformer can be between the mains and the transistors. With the first arrangement (Fig. 10), the equipment is lightweight, but very sensitive to voltage peaks. Equipment of the second type (Fig. 11) is heavier.

Submerged Arc Welding. In submerged arc welding, the arc produced between the automatically fed filler wire and the workpiece (Fig. 12) is covered by powder, and is located in a cavity formed within powder that has become molten [26]. The powder flows in the direction of welding, preceding the welding zone of the metal sheet. Only a small amount of powder is consumed by the welding process, and the unused powder is sucked away and reused. The powder blanket essentially performs the same function as the coating on manual electrodes. The welding speed is high, and in equipment with several filler wires the melting capacity is also high. The process can be used only in situations where the sheet metal is not steeply sloped. A weld pro-

duced by this process is shown in Figure 13, and a coating of weld metal in Figure 14.

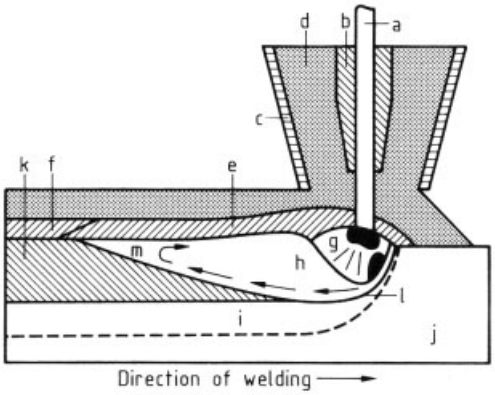

Figure 12. Principle of the submerged arc welding process
a) Electrode; b) Electrical contact device; c) Powder feed funnel; d) Welding powder; e) Molten slag; f) Solid slag; g) Cavity with electric arc; h) Weld pool (arrows indicate flow); i) Heat-affected zone; j) Parent metal; k) Solidified weld metal; l) Melting front; m) Solidification front

Figure 13. Polished section of a weld produced by the submerged arc process
Butt weld with two passes; four-wire weld per pass; metal X60; wall thickness 36 mm; welding speed 1300 mm/min

Narrow Gap Welding. Heavy section workpieces can be joined economically and with minimum effect on the parent metal by the so-called narrow gap welding process. The best of the various methods for filling a vertical narrow gap with weld metal is the MIG/MAG process. In this gas-shielded process, the filler wire and protective gas are cooled in a sword-shaped sheathed tube and fed to the welding point in the narrow gap. There are two systems:

1) Two wires, which have been subjected to plastic deformation to give a predetermined internal stress, are guided into the two corners of the welding gap, giving a defect-free root position
2) In the second system, this function is performed by a wire preformed to a waved shape so that it alternately fills each corner of the region to be welded inside the narrow gap

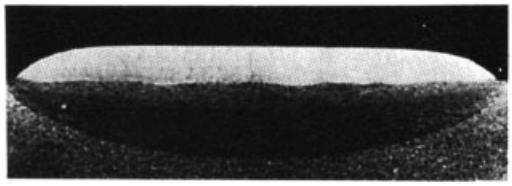

Figure 14. Polished section of a metal coating applied by the submerged arc process used for cladding a nuclear reactor pressure vessel
Parent metal 22NiMoCr37; cladding metal X2CrNiNb2413

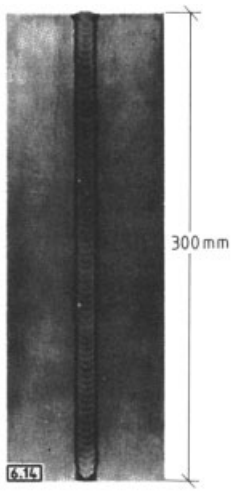

Figure 15. Narrow gap weld (vertical gap)
Metal ASME SA 387, group 22; metal thickness 300 mm; gap width 9 mm; welding speed 220 mm/min; 76 weld passes

Narrow gap welding can also be carried out by the submerged technique. While gas-shielded narrow gap welding can be used for gap widths

of ca. 10 mm (Fig. 15), the gap must be wider for submerged welding.

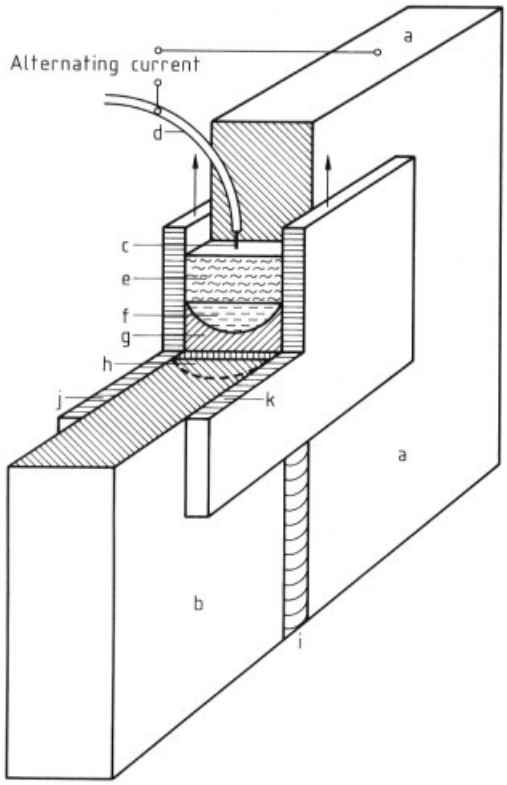

Figure 16. Principle of electroslag welding
a, b) Workpieces; c) Wire electrode; d) Wire feed; e) Molten slag; f) Weld pool; g) Solidified weld metal; h) Heat-affected zone; i) Weld bead; j, k) Water-cooled copper plates (moved upward as welding proceeds)

3.1.3. Electroslag Welding

Electroslag welding is a machine-operated resistance-heated fusion welding process for joining vertical metal plate >20 mm thick (Fig. 16). The filler metal is fed from above in wire form and melts in the molten slag layer. The electrical resistance of the slag layer causes heat to be generated by the current, heating and fusing both the wire and the side walls of the metal sheets being joined. The current passes through the filler wire to the slag layer. Water-cooled copper plates, which are moved upward as welding proceeds, retain the weld pool and give a constant weld thickness. The weld section is filled in a single pass. As the weld pool is comparatively large and the cooling rate is slow, a coarse metal structure can be formed; iron powder or additional cold wires are added to counteract this effect.

3.1.4. Electron Beam Welding

In electron beam welding (Fig. 17) [27, 28], the energy source is an electron beam with a high energy density (up to 10^8 W/cm^2). The electrons are emitted from an incandescent tungsten or tantalum cathode in a vacuum of ca. 1.3×10^{-2} – 1.3×10^{-3} Pa (1.3×10^{-4} – 1.3×10^{-5} mbar), accelerated by a high voltage (up to 150 kV), and focused by electromagnetic lenses to a narrow beam. The electrons penetrate the workpiece at high velocity, and a high proportion of their kinetic energy is converted to heat. This region of the workpiece melts, and the metal at the center evaporates, so that a metal vapor capillary is formed, surrounded by an envelope of molten metal. As the energy is released almost uniformly across the whole thickness of the sheet (deep welding effect), an extremely narrow, needle-shaped melting zone is formed (Fig. 18). The heat-affected zone in the parent metal is very small.

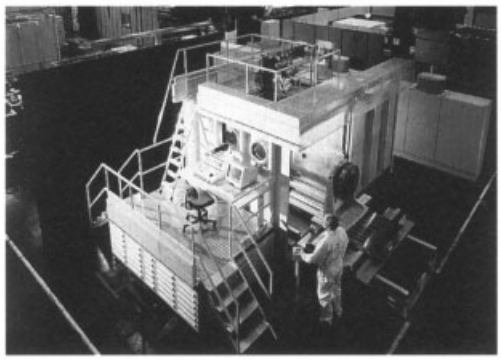

Figure 17. Electron beam welding machine
Vacuum chamber for individual welding of large components with complex weld geometries, e.g., turbine wheels (courtesy of Messer-Griesheim)

The wide range of weldable metal thicknesses, from foil with a thickness of micrometers to plates over 100 mm thick, is not matched by any other process. The deep welding effect of the

electron beam enables welds to be produced at high speeds where access is very difficult, e.g., in deep recesses.

Electron beam welding is used in the construction of equipment and machinery, and in aircraft technology. As the process is operated in a vacuum, it can be used to weld gas-sensitive, high-melting, and chemically active metals such as tungsten, molybdenum, tantalum, nickel, zirconium, titanium, and special high-temperature steels. The process is also used for welding combinations of different materials and for dimensionally accurate low-distortion welding of machined precision components.

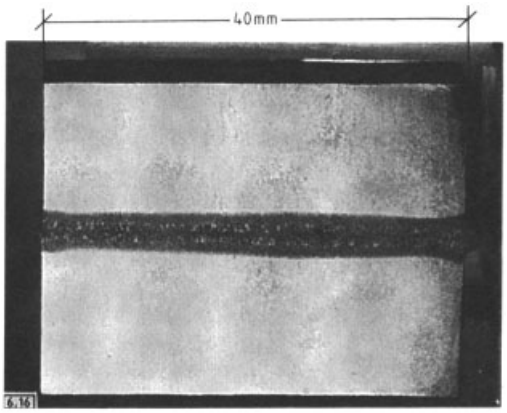

Figure 18. Polished section of a weld produced by the electron beam process
Metal Ck 45; wall thickness 40 mm; width of weld bead ca. 2 mm; width of heat-affected zone ca. 1.5 mm

3.1.5. Laser Beam Welding

In laser beam welding [27, 29], a laser produces coherent monochromatic light beams with low divergence. These beams can be focused with lenses or concave mirrors to give a power density in the focal spot of up to 10^8 W/cm^2.

Solid state lasers with power outputs of a few watts can be used for welding, but CO_2 gas lasers in the kilowatt range are more usual. The various types of laser differ with respect to the wavelength of emitted light, efficiency, and mode of operation (pulsed or continuous).

In the CO_2 laser, emission is excited by an electrical glow discharge in the laser gas, a mixture of carbon dioxide, helium, and nitrogen, and is contained in a discharge tube. This is an optical resonator, containing plane parallel mirrors at each end. One of these is semireflecting, so the laser beam passes through it and out of the laser (Fig. 19). The efficiency of the CO_2 laser is comparatively high, 15 – 20 %. The energy of the beam absorbed by the material is converted to heat. The absorption depends on the wavelength of the laser beam and on the properties of the metal and its surface.

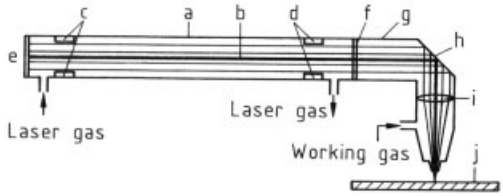

Figure 19. A CO_2 gas laser for metal working (cutting, welding, hardening, remelting)
a) Cooled discharge tube; b) Laser beam; c) Anodes; d) Cathodes; e) End mirror; f) Semireflecting mirror; g) Cutting head; h) Deflection mirror; i) Lens; j) Workpiece

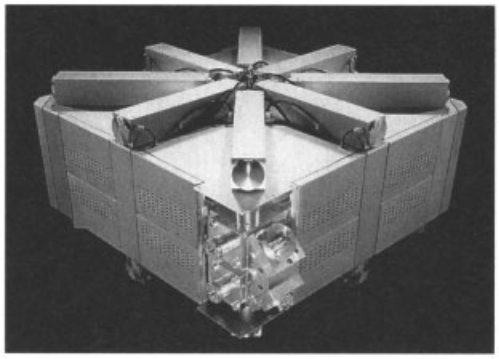

Figure 20. High-power CO_2 laser with maximum output 6 kW (courtesy of Trumpf)
Resonator formed into two squares on two levels; excitation by high frequency

An important advantage of laser welding is that the heat-affected zone is extremely small, so the workpiece is hardly distorted at all. As the heated zone is so thin, weld preparation must be precise and the beam must be precisely controlled during welding.

Laser welding (and cutting) has been developed into a high-power process, and is a likely candidate for industrial-scale production. Most equipment operates at power outputs of 3 –

6 kW (Fig. 20), but considerably higher powers are also available. Achievable welding speeds, which depend on the metal thickness, are given in Figure 21. A further great advantage of the process is that it does not require a vacuum.

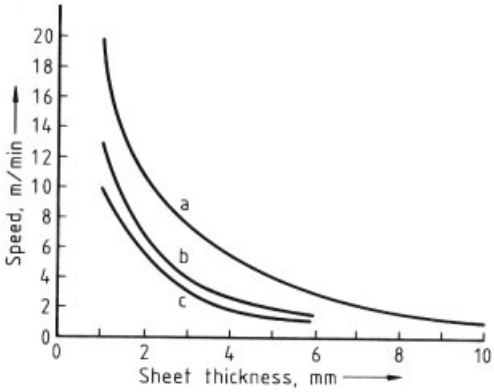

Figure 21. Laser welding and cutting speeds in structural steels
a) Welding speed at 6 kW; b) Welding speed at 3 kW; c) Cutting speed at 3 kW with N_2 cutting gas in special steel

3.1.6. Aluminothermic Welding

In this process, the heat required is produced by the reaction (which requires external ignition) between iron oxide and powdered aluminum. The ends of the steel components to be welded are clamped together inside a prepared closed form, and the aluminothermic powder mixture is added and ignited. The aluminum oxide formed in the reaction floats on the surface and protects the molten steel from the atmosphere. This process is mainly used for welding railway lines.

3.2. Pressure Welding

Modern pressure welding processes derive from the hammering methods used in the blacksmith's forge. The forge fire has been almost completely replaced by such techniques as electrical resistance heating in resistance welding, the electric arc, e.g., in flash-butt welding, electrical induction in induction welding, friction in friction welding, and sometimes the gas flame in gas pressure welding.

3.2.1. Resistance Welding

In resistance pressure welding [30], heat is produced by passing an electric current through the workpieces, which are in contact at the area to be welded, and the welded joint is produced by pressing them together. The electric current and compression force used, both of which may be varied during the process, depend on the nature of the weld.

Resistance Spot Welding. In resistance spot welding, flat components in contact with each other, usually sheet metal, are welded together at single points or rows of points. The welding current (nowadays nearly always alternating current) is fed through rod-shaped copper electrodes (Fig. 22). The pressure/force applied by one or both electrodes can vary with time. As the interfacial electrical resistance between the copper and workpiece materials (copper excluded) is always less than that between the workpiece materials, the spot weld is formed only between the workpiece components. In this process, only enough heat is applied to form the spot welds, so that heating effects on the parent metal are very small.

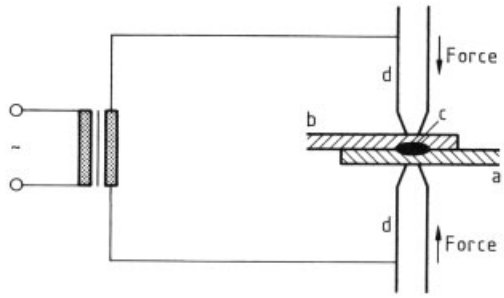

Figure 22. Principle of resistance spot welding
a, b) Workpieces; c) Spot weld; d) Copper electrodes

The force exerted by the machine must be adjusted in accordance with the thickness of the workpieces. The maximum thickness for steel is currently ca. 2×12 mm, and for light metals ca. 2×5 mm. To produce a satisfactory welded joint, the magnitude and duration of the welding current and the force exerted by the electrodes must be matched to each other for a given electrode cross-sectional area.

Spot welding can be either direct or indirect, according to the arrangement of the electrodes at the workpiece:

1) In direct spot welding, electrodes are applied to both sides of the sheet metal
2) In indirect spot welding, an electrode is applied from one side only, and the circuit is completed by a copper base

The process can be described as single-spot, double-spot, or multispot welding according to the number of spot welds produced in a single operation. In multispot welding, the strength of the joint depends on the arrangement of and distance between the spot welds. Many types of equipment and machinery are available for resistance spot welding, from small hand-operated welding tongs for a single-spot weld to the large machines used in automobile production lines.

Transistorized electric power sources are also widely used in resistance welding technology. The use of direct current for welding avoids harmful interactions with the electric network, and saves energy.

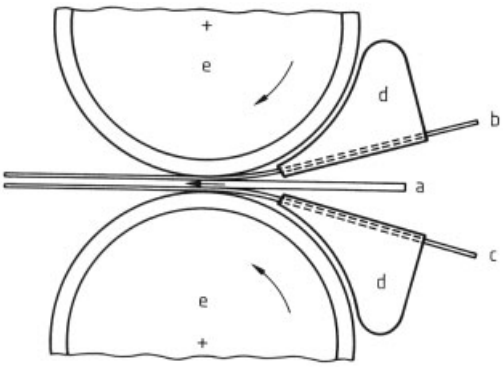

Figure 23. Principle of foil resistance roller seam welding a) Workpiece; b, c) Foil strips; d) Foil feed guides; e) Current-carrying rollers

Resistance Roller Seam Welding. In resistance roller welding, the workpieces pass between two continuously or intermittently rotating copper rollers, these being the electrodes. This system enables uniform welded seams of any desired length to be produced using a constant alternating current. Good welds are also produced by an intermittent current if the welding points overlap. One of the rollers may be replaced by a copper rail. The main application is in the production of hollow products, e.g., heating equipment such as calandrias or radiators.

In foil resistance roller seam welding, the edges of the sheet are butt welded together. On both sides of the sheet, above and below the joint, strips of corrosion-resistant metal foil are applied and welded as they are compressed between the pair of rollers (Fig. 23). Zinc-coated steel sheet can be welded with corrosion-resistant foil without reducing its resistance to chemical attack.

Resistance Projection Welding. This differs from resistance spot welding in that one of the two parts being welded is shaped to produce projections, and the electric current then becomes concentrated at these points. Several spot welds can then be formed simultaneously using simple copper electrodes with flat surfaces. This process is especially suitable for welding relatively small mass-produced components.

Resistance Butt Welding (Pressure Butt Welding). In pressure butt welding, plane parallel machined surfaces of workpiece components are pressed together, and an electric current is passed through this area with the aid of copper jaws clamped on close to the butt joint. Maximum heating occurs at the interface as the resistance is a maximum at this point. The compressive force produces a bulge during the welding operation. This process is especially suitable for very simple profiles with cross-sectional areas up to ca. 500 mm^2.

In flash-butt welding, the surfaces are not machined, and initially come into contact at a few points only. However, the whole of the interfacial area is eventually affected during the course of the welding operation. The process is used for butt welding sheet, profiles, and pipes with cross-sectional area up to ca. 70 000 mm^2.

3.2.2. Arc Pressure Welding

In this process, an electric arc between the two faces of the workpieces to be butt welded produces the heat required to melt the faces, and the workpieces are then pressed together. This squeezes out the slag which forms in the gap, so

that a good metallic joint is formed. The compressive force produces a welding burr which must be machined off in the case of larger components.

In arc flash-butt welding, considerably larger sections identical to each other are joined together by large machines using a basically similar process, but with a longer operating cycle.

In stud welding, studs are individually welded onto one side of large surfaces using manual equipment. This process is used especially in the construction of heat exchangers to increase the area of the heat transfer surface.

3.2.3. Induction Welding

Induction welding (Fig. 24) is a type of pressure welding in which the heat is provided by electrical induction. It is mainly used in the production of tubes by longitudinal welding.

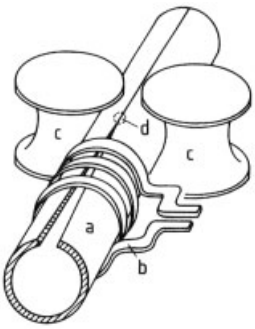

Figure 24. Principle of induction welding of a tube
a) Workpiece being welded to form a tube; b) Induction heating coil; c) Feed and compression rollers; d) Pressure welding point

Friction Welding. In friction welding, the heat for the welding process is generated by rubbing the workpieces together at high speed. For example, the end faces of two shafts can be rubbed together until softening (at least) occurs, after which the rubbing operation is stopped and the two shafts are pressed together. The pressure/force program must be very precisely matched to the friction program.

Gas Pressure Welding. In gas pressure welding, the components are heated with gas flames to the fusion temperature and then pressed together. The process is mainly used for welding small tubes. The equipment required is usually simple.

3.3. Special Welding Processes

Some welding processes cannot be classified as either fusion or pressure welding (see Fig. 2 B), e.g. diffusion welding.

Diffusion welding is an important special welding process in which the joining surfaces of the two workpieces, which fit together exactly, are very carefully smoothed and then heated to the diffusion temperature in a high vacuum while pressing them together with only a small force. The diffusion temperature (in °C) is ca. 2/3 of the fusion temperature. The process is very expensive owing to the careful surface preparation required and the high-vacuum technology. It has so far been little used, and only for very high value materials, but produces an ideally welded joint.

4. Fusion Cutting Processes

4.1. Flame Cutting (Autogenic Cutting)

In flame cutting, the surface of the workpiece is heated by the fuel gas – oxygen flame of the cutting torch at least to the temperature at which scale forms. Part of the cutting oxygen then oxidizes part of the metal, and the very strongly exothermic combustion reaction provides the heat to melt the adjacent metal. The molten material, which consists of oxides and unoxidized metal, is blown out of the gap by the remainder of the cutting oxygen.

The process of flame cutting is therefore governed by two preconditions:

1) The scaling temperature of the material to be cut must be lower than its fusion temperature
2) The oxides formed must not form a continuous skin, so that the oxygen can penetrate to the surface of the melt

The most important part of the cutting torch is the nozzle (Fig. 25) which shapes the jet of cutting oxygen, the actual cutting tool in this process. The exit openings for the flames are usually

rotationally symmetrical about the cutting oxygen jet, so that with the jet directed vertically downward it can cut in all horizontal directions.

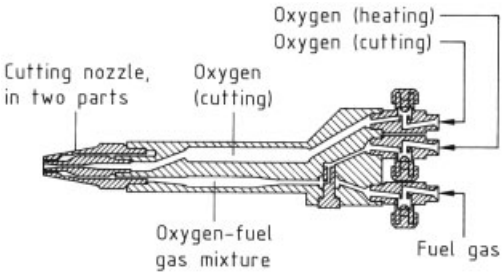

Figure 25. Mechanical cutting torch

In very thick walled workpieces, the flames also heat the side walls of the gap produced in the direction of cutting immediately behind the cutting oxygen jet. The viscosity of the molten material produced then remains low and it can be blown out of the gap almost completely. Slower burning gases such as propane are better than acetylene for achieving this result.

All the plain carbon and low-alloy steels with a wide range of thicknesses (3 mm to >2 m) can readily be flame cut by this method, but not chromium steels (see below). The maximum cutting speed depends not only on the properties of the cutting oxygen jet, but also very much on the thickness of the workpiece (Fig. 26). The required quality of the cut can limit the cutting speed. Faster cutting speeds can be used if the quality of cut need not be high.

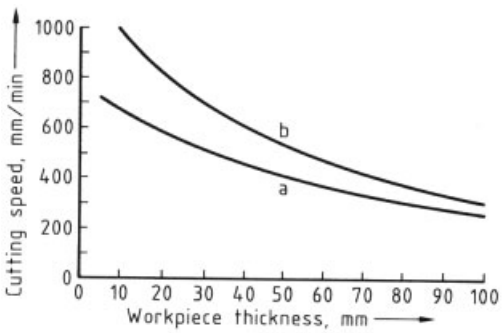

Figure 26. Speed of flame cutting plain carbon steel with a high-power nozzle as a function of workpiece thickness
a) Results achievable with optimum quality cut; b) Results if there is no quality requirement

Chromium steels cannot be flame cut conventionally, as the chromium in the steel increases the solidification temperature of the oxide melt, and an oxide skin forms. This chromium oxide skin disappears if iron powder is injected at an angle from the sides through the flame and into the oxygen jet. The iron then burns on the molten material and provides more heat. This trick enables chromium steels to be flame cut.

The range of flame cutting equipment extends from simple hand torches and portable flame cutting machines to large stationary machines. There are many ways of controlling these machines, including photoelectric scanning of a pattern (which may be a line drawing) and control by mainframe computer.

4.2. Plasma Cutting

Plasma cutting continues to grow in importance. All the electrical energy required is supplied externally, so that, unlike the flame cutting process, this process generates no additional heat. The design of the plasma cutting burner (Fig. 27) is almost identical to that of the plasma welding burner (see Fig. 6), except that there is no gas shield. Also, the power outputs of plasma cutting equipment (up to ca. 50 kW) are higher than those of welding equipment, enabling 100-mm thick steel to be cut.

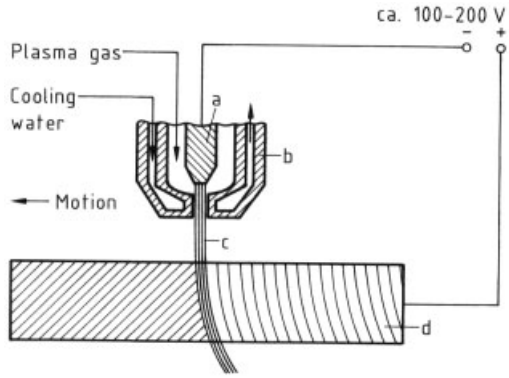

Figure 27. Principle of plasma cutting
a) Tungsten cathode; b) Copper nozzle; c) Plasma jet; d) Workpiece

Plasma cutting can be used with all electrically conducting materials. Cutting rates for thin sections are higher than with flame cut-

ting. A mixture of argon and hydrogen has been found to be a suitable plasma gas for cutting austenitic chromium – nickel steels and aluminum. For high-carbon structural steels, for which high cutting speeds are an advantage, the plasma gases used are air or other nitrogen-containing mixtures. With air as the plasma gas, a water-cooled copper electrode fitted with a small lanthanum plate insert is used instead of the usual pointed tungsten electrode.

An alternative process, the water plasma process, can also be used for cutting structural steels. Water is radially injected into the cutting nozzle immediately behind the narrowest cross section of the jet. This prevents burr formation when cutting plain carbon steel (Figs. 28).

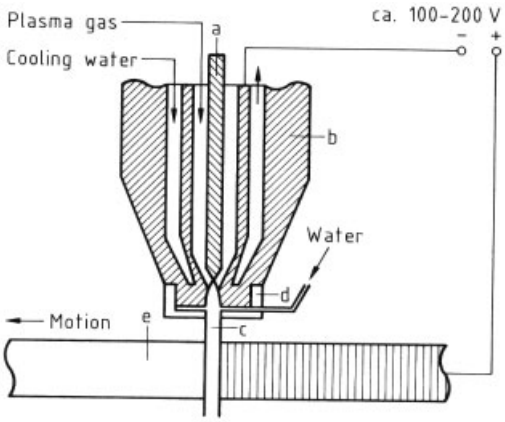

Figure 28. Principle of water plasma cutting
a) Tungsten cathode; b) Copper nozzle; c) Plasma jet; d) Ceramic nozzle; e) Workpiece

In all the plasma cutting processes described above, the noise level is high, and harmful substances are emitted. In stationary plasma cutting equipment, these substances are usually removed by air extraction. Noise levels can be much reduced by surrounding the plasma jet with a shielding envelope of water or by carrying out the process under water.

4.3. Laser Cutting

The laser beam can be focused by relatively simple optical means to give an extremely small spot (with an effective beam diameter of ca. 10^{-1} mm). Part of the material from the gap is evaporated, and the fused material formed is blown out by a coaxial gas stream. If the latter consists of oxygen, the cutting effect is considerably improved. The gap produced is narrow and smooth, and the heat-affected zone is small. Power outputs of metal-cutting lasers are in the range 500 – 3000 W. High-pressure nitrogen has been found suitable for cutting stainless steels (see Fig. 21).

Solid-state lasers with power outputs of up to 1 kW are available. Solid state lasers combined with light guides have proved to be very useful for metal cutting by robots operating in poorly accessible locations.

5. Welding of Various Materials

It was shown in Section 2.3 that the welding behavior of a metallic construction material is mainly determined by its chemical composition, and that each alloy type has its own special characteristics.

5.1. Plain Carbon Steels

The compositions, mechanical properties, and commercially available qualities of plain carbon steels (→ Steel, Chap. 6.1.1.4., → Steel) are listed in EN 10 025 : 01.91 (general structural steels). Their welding properties are determined by several factors: carbon content (small additions of silicon and manganese have a small effect), degree of deoxidation (killed or rimming steel, → Steel, Chap. 3.4.1.1., → Steel, → Steel), and the level of impurities (sulfur, phosphorus, arsenic, copper, etc.). Other important factors include the smelting process [Siemens – Martin or open-hearth process, oxygen blast process (LD process), electric furnace process], the casting process (Shell casting, continuous strand casting), and the delivered state (normalized, cold-hardened).

Plain carbon steels with carbon content up to ca. 0.25 % and an impurity level corresponding to the state of the art can be welded without problems, especially if they are fully killed, so that they do not completely lose their grain fineness on welding.

Steels with a high carbon content can also be welded if hardening is prevented by welding

when hot and following this with an additional heat treatment.

5.2. Alloy Steels

Most so-called low-alloy steels, i.e., those containing a total of up to 10 % alloying elements (→ Steel, Chap. 6.2.1.2.), can be welded like plain carbon steels. However, there is a limit to the acceptable carbon content, owing to its hardening effect, and the other alloying elements also affect the welding properties from the point of view of through tempering. This means that, with increasing levels of alloying elements, the size of the section to be welded is also important.

Whereas in plain carbon steels the type of filler metal depends mainly on the properties required (strength, toughness, gap-filling property, high yield), in alloy steels, the chemical composition must also be considered. Further details on steels and the appropriate filler metals are available in steel manufacturers' recommendations.

The important alloying elements in low-alloy steels include chromium, nickel, molybdenum, vanadium, tungsten, copper, and nitrogen. The family of low-alloy steels includes tempering steels, low-temperature steels, high-temperature structural steels, and high-strength structural steels. Further details can be found in EN 10 083, Parts 1 and 2 (tempering steels).

The high-alloy steels, in which the total content of alloying elements exceeds 10 %, require special care in welding. These steels include:

1) Low-temperature steels with 9 % nickel
2) Corrosion-resistant ferritic chromium steels containing at least 12 % chromium
3) Corrosion-resistant austenitic chromium–nickel steels containing at least 17 % chromium and 8 % nickel [31]
4) The high-temperature and very high temperature steels based on 16 % chromium and 13 % nickel

It is possible to use 9 % nickel steel at liquid nitrogen temperature (−196 °C) for components in the cryogenics industries. It cannot be welded with filler metal of the same composition for strength reasons, but must be welded either with a nickel-based alloy, or, for increased strength, with a filler metal containing chromium, nickel, and tungsten as alloy components.

The corrosion-resistant ferritic chromium steels are generally difficult to weld. The main reasons for this are the very considerable grain growth that occurs during welding and the consequent loss of toughness and increased incidence of intergranular corrosion problems (→ Corrosion, Chap. 3.2.3.1.). In spite of these difficulties, these steels are weldable if certain precautions are taken.

The corrosion-resistant austenitic chromium–nickel steels have become extremely important in the construction of chemical plant. Because they do not undergo transformations, they have good welding properties in all welding processes if their individual characteristics are borne in mind. Intercrystalline corrosion is minimized by suitable choice of the alloy composition. For example, steels with a low carbon content can be used, or those in which the carbon is chemically combined with niobium or titanium. The required degree of stabilization of this type depends on the time during which the metal is within the main precipitation temperature range (550–800 °C).

The high-temperature and very high temperature 16 : 13 chromium–nickel steels have very similar welding properties to those of the austenitic chromium–nickel steels. If it is intended to heat treat these metals before use by solution annealing and hardening, they are welded in the solution-annealed state and hardened after welding.

The corrosion-resistant austenitic–ferritic steels are seldom used, but are valuable because of their resistance to wear and corrosion. However, their precipitation behavior leads to welding problems. The ferritic steels are superior from the point of view of grain growth. However, such welding operations should be limited to the smallest possible areas.

If metals of different structure types are welded together, e.g., austenites to ferrites (black-to-white joint), the Schaeffler diagram provides useful information on the resulting metal structure [1, p. 128], [32]. In this diagram, the ferrite-forming and austenite-forming elements are summarily represented (including a weighting factor) in such a way that the phase field of the probable structure types can be estimated. Each steel alloy is represented by a particular point in the diagram. If the points for the

two different metals to be welded are joined together, the expected mixing ratio of the mixed structure lies on this line.

5.3. Cast Steel

Like rolled steel, normal cast steel can be welded using filler metal of the same type and composition. With large sections and cast steel components of complex design, it is necessary to look out for the possible presence of interior stresses. These are much more likely to be detected by experienced than inexperienced welders.

5.4. Cast Iron

All three types of cast iron, i.e., malleable iron, gray iron (with lamellar graphite), and SG iron (with spheroidal graphite), can be welded, with the exception of blackheart malleable iron (not decarburized by heat treatment). Whiteheart malleable iron (decarburized by heat treatment) can be welded without problems at its decarburized edge zone to produce structural welds. The structure of the edge zone corresponds to that of low-carbon steel.

Gray iron can be welded for repair purposes by two methods. In welding with preheating, the casting is heated to ca. 650 °C, welded at this temperature with filler metal of the same composition, and then cooled very slowly to prevent cracking. The low-viscosity weld metal is prevented from running off by fitting shaped carbon plates around the welding area. The solid weld metal is of identical quality to that of the parent metal. Welding of gray iron with filler metal of the same type without preheating is impossible. Instead, pure nickel and nickel – iron alloys are used because their thermal expansion coefficients are similar and they have very good ductility. In the welding process, the parent metal near the welding zone must be only slightly heated. Therefore, in manual arc welding, thin electrodes are welded on as short seams in various places. These seams are spread out in longitudinal and transverse directions by hammering during cooling to minimize the interior stresses. The joint produced by welding gray iron without preheating is not of such high quality as that obtained by welding with preheating.

Cast iron containing spheroidal graphite (SG iron), unlike gray iron, has generally good ductility, the breaking elongation being controllable by varying the composition. The weld metal and the parent metal should have approximately similar elongation properties. By using nickel – iron alloys as filler metal, the weld produced is always ductile. By using weld metal identical to the parent metal, the elongation properties of the weld metal can be exactly matched to those of the parent metal. When welding without preheating, the hardening effect produced must be counteracted by subsequent heat treatment [1]. Welding of cast iron and SG iron is not only used for repair purposes, but sometimes also as a fabrication method.

5.5. Nickel and its Alloys

Nickel and nickel alloys are characterized not only by their good physical properties (strength and toughness), but also by their excellent corrosion resistance. Additions of chromium, molybdenum, and various other alloying elements further improve the hot strength. These properties have led to the widespread use of nickel and nickel alloys in the construction of chemical equipment. Pure nickel [33] and the following groups of nickel alloys are particularly important: nickel – copper (monel), nickel – chromium – iron (Inconel), nickel – chromium – molybdenum (Hastelloy), and the hardenable chromium-containing nickel-based alloys based on Ni_3TiAl, the so-called Nimonic alloys, for very high hot strength.

Nickel can be welded by all the welding processes. For the most important alloys, especially the high-temperature alloys, identical filler metals are available both for manual arc welding and for welding processes with shielding gases. For welding tube joints in reactor construction, in gas turbines, and in chemical equipment construction, most processes are fully automated, in particular electron beam welding.

Because of its physical properties, especially its poor thermal conductivity, the weld preparation, the welding electrodes, and the protective gas selected must enable long periods of degassing time to be available during the solidification process. Small-section electrodes with small current consumptions are usually used to

minimize the amount of heat applied, and the protective gas should have a low thermal conductivity (e.g., argon).

For good weld quality, it is important that the workpieces and filler metals should be free from oil, dirt, oxides, and sulfides; nickel has an extremely strong affinity for sulfur, so a low-melting nickel sulfide (637 °C) can form in the weld seam and lead to cracks in the joints, even if the quantity is extremely small. The solubility of sulfur in nickel is <0.005 %. Nickel can take up trace amounts of sulfur from substances in the form of gases or vapors or from greases and lubricants, even after extremely small exposure times. Carbon can lead to graphite precipitation at grain boundaries, so the level of this element must be limited if niobium or titanium has not been included in the alloy to give carbide formation (Inconel alloy).

Unlike steel, nickel and its alloys are not very sensitive to hydrogen. As nickel is expensive, it is often used only as a plating or cladding material in chemical equipment. The welding of plated components requires great care to prevent the parent metal from mixing with the plating layer.

5.6. Copper and its Alloys

One of the most important properties of copper, its very good electrical and thermal conductivity, has a decisive influence on its welding properties. There must be a very plentiful heat supply for the welding process. If this is allowed for, copper can be welded by all the welding processes. Autogenic welding is still often used with copper, mainly for welding tubes; a flux is usually used (see [1, p. 167], EN 29454-1: 1994, and ISO 9454-1: 1990). Thorough preheating is important for producing a penetrating weld.

A condition for good weldability of copper and its alloys is low oxygen content, as copper and oxygen react to form a copper/copper(I) oxide eutectic which precipitates at the grain boundaries on solidification of the melt and causes embrittlement.

Copper – zinc alloys (brass) are usually soldered, not welded. This is also true of special brasses with high strength and corrosion resistance.

Copper – tin alloys (tin bronzes) have extremely good corrosion resistance and are much used in freshwater treatment plants containing highly chlorinated solutions. In spite of their susceptibility to hot cracking, these alloys can be welded with manual electrodes and also by the TIG process. They are also used in the production of armor plating by the explosive method.

Copper – aluminum alloys can also be welded by all the electrical welding processes, but the aluminum oxide layer adheres very strongly and causes welding difficulties (see Section 5.7).

5.7. Aluminum and its Alloys

Aluminum and its alloys are increasingly used in the construction of chemical equipment, especially for moving parts for which low density is an advantage. These materials have good corrosion resistance, and their low-temperature toughness properties enable them to be used at liquid nitrogen temperature.

The main problem in welding is dealing with the film of aluminum oxide scale which adheres extremely strongly and has a very high melting point. As the thermal conductivity of aluminum metal is also very high, there must be an ample heat supply so the weld pool does not solidify too rapidly or contain nonmetallic particles or gas-filled pores.

The simplest process, welding with manual electrodes, is widely used and is completely satisfactory for simple joints. Qualitatively good welds can be produced by the TIG and MIG processes. The preferred protective gas is helium. The alternating current removes the layer of aluminum oxide. If the weld is carefully prepared and the auxiliary materials have been correctly chosen (surface cleanliness), mechanically sound low-porosity welded joints are obtained. High-power processes such as submerged arc welding are also suitable.

Owing to the low melting point of aluminum, the low viscosity of the molten metal, and the small difference in density between the metal and its oxide, manual welding of aluminum and its alloys makes extremely high technical demands on the welder.

5.8. Titanium and its Alloys

The physical properties and corrosion resistance of titanium and its alloys have aroused interest in its use in the chemical industry as well as aircraft construction. As is the case with steel, the main welding problem is its reactivity with oxygen, nitrogen, hydrogen, and carbon, so electron beam welding in vacuum is very useful for titanium, although titanium and its alloys can also be welded by the TIG and MIG processes if suitable precautions are taken.

As even the small amount of gas uptake that occurs at slightly elevated temperatures leads to embrittlement, the weld zone must be very rapidly cooled. The heat supplied can be rapidly removed by welding on a copper base, and by using cooling sprays in the region of the heataffected zone. Essential preconditions include an absolutely clean welding surface, oxide-free filler metals, and pure protective gases. The gas nozzles used are specially designed to retain the protective gas for as long as possible at the weld as it cools and in the region of the heat-affected zone. If these precautions are taken, physically sound joints can be produced.

5.9. Thermoplastic Materials

Of all the various types of plastic materials, only some thermoplastics are weldable [34]. No thermosetting materials can be welded. The weldable thermoplastics include polyethylene, polypropylene, poly(vinyl chloride), poly(methyl methacrylate), and polyisobutylene. Polytetrafluoroethylene and polytrifluorochloroethylene are not weldable.

In all welding processes for plastics (see also → Plastics, Processing, Chap. 3.2.1.), the material is softened by supplying heat to the surfaces to be welded. This heat is usually provided by a current of hot gas (air, nitrogen) or by applying heatable metal plates (heating elements or heat reflectors). Other techniques include the application of wires through which an electric current is passed, heat pulses, friction, high-frequency heating, and ultrasound. The workpieces are then pressed together at the weld surfaces, or softened filler material is pressed into the surfaces to be welded.

The welding of thermoplastic materials is used mainly in the construction of pipework (Fig. 29), equipment, machinery, and packing materials. The conditions for a good welded joint include similarity of the types of materials to be welded, a design suitable for the service requirements, choice of the most suitable welding method, and cleanliness of the parts to be joined.

6. Testing and Control Methods

In principle, the testing methods used in welding technology are the same as those used in general materials testing. However, there are additional procedures designed specifically for testing welded joints and the weldability of materials.

The following standards are for destructive testing:

DIN 50 120 Parts 1 and 2: Tensile testing of welded joints; also ISO 4136: 1989
DIN 50 121 Parts 1 – 3: Technical testing of transverse properties of welded joints; also ISO 5173: 1981; ISO 5177: 1981; ISO/DiS 5173: 1992
DIN 50 122: Notched-bar impact testing of welded joints; also ISO/DiS 9016: 1992
DIN 50 124: Testing of shear strength of spot welded joints; also ISO 10 447: 1991
DIN 50 127: Notch-tensile strength
DIN 50 129: Cracking tendency of filler materials

The following standards are for nondestructive testing:

DIN 54 109: Picture quality in radiographic testing
DIN 54 111: Testing with X-rays and γ-rays; also ISO 1027 : 1983; ISO 1106-1/3

The most important nondestructive testing methods are radiography with X-rays and γ-rays (→ Nondestructive Testing, Chap. 2.2.) and ultrasonic testing (→ Nondestructive Testing, Chap. 2.3.). Ultrasonic testing is comparatively difficult and still in the development stage. Interpretation of the ultrasonic pictures is difficult, for example, differences in grain size may not be shown or may be interpreted as defects. Current developments in computer-controlled ultrasonic testing methods should lead to improvements in picture interpretation.

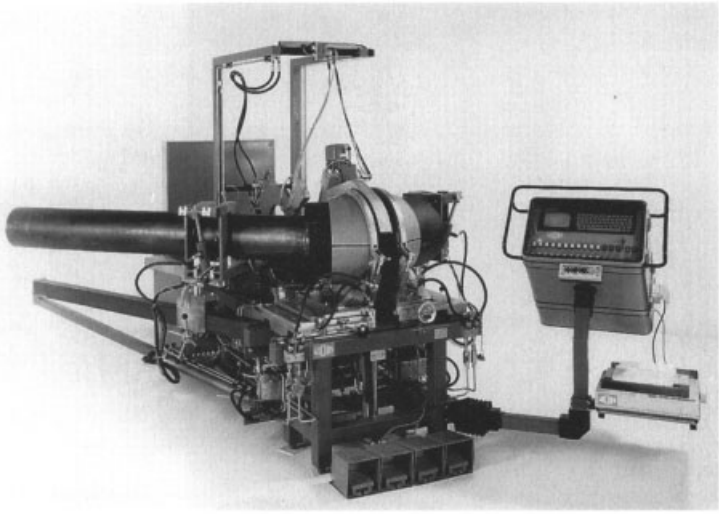

Figure 29. Welding of polyethylene pipe

There are considerable developments in control systems that continuously monitor the welding process by detecting deviations of welding parameters from standard tolerance ranges and draw conclusions concerning the quality of the weld. There are already many successful applications of these methods in spot welding, but they are in an early stage of development for arc welding. For the methods to be successful, it is necessary to know the tolerance limits within which the fluctuations of the process parameters do not cause any harmful effect on the weld. Manufacturers of welding equipment are working hard on this problem, and progress is expected in the near future.

7. Health Protection and Safety

7.1. Health Protection

Incorrect operation of a welding process can lead to hazards to personnel, especially from electric current, explosive gases, atmospheric pollutants (smoke, vapors, and gases), radiation, and noise [37].

To minimize the hazard from electric current, machines should conform to the relevant safety guidelines, e.g., in Germany, those of the VDE. The special hazards associated with restricted working areas are reduced, for example, by limiting the no-load voltage to 42 V. If higher no-load voltages are necessary for technical reasons, e.g., for plasma cutting, the potential hazard is minimized by the use of suitable protective circuits.

Strong light can be emitted by electric arcs, plasma jets, direct and reflected laser beams, and bright flames. These are mainly hazardous to the eyes and the exposed skin. The eyes should be protected by filters (see DIN EN 167, ISO 4849, DIN EN 169, and ISO 4850: welding protection filters, DIN 58 211, Part 6: safety spectacles, DIN 58 125: optical protection filters against laser light), and the skin is protected by suitable covering (textiles and leather).

The concentrations of atmospheric pollutants such as airborne particulates (fume), vapors, and gases must not exceed the legal limits, i.e., the MAK values [36]. The airborne particulates contain lead, chromium, cadmium, nickel, mercury, copper, and zinc. They are formed as metal vapor during the processes of welding and soldering, and usually become oxidized and suspended in the air as fine dust. Fume formation can be excessive, especially when welding with coated electrodes. The harmful gases formed can include nitrogen oxides (NO, NO_2) from flames, ozone and sometimes phosgene from the intense light emitted by electric arcs, and carbon monoxide when welding with a carbon dioxide gas shield. Protection from harmful vapors and gases must be provided, especially when working in confined spaces, e.g., air extraction plant and equip-

ment for supplying clean air [37]. The use of pure oxygen is prohibited because of the danger of fire and explosion.

Noise can be produced by the high-velocity gas jets and flames used in welding and cutting technology, e.g., those used in cutting torches and plasma equipment. Effective countermeasures include simple sound-absorbing shields, which absorb the high-intensity, high-frequency components.

7.2. Safety

When welding and cutting, it is extremely important to take precautions against fire, which can be produced directly by flames, or can begin slowly due to unnoticed molten material impelled over long distances, often only becoming serious after some hours.

Fuel gases constitute a further source of danger (e.g., acetylene, propane, etc.). Acetylene has a wide range of explosive concentrations and requires special safety precautions (→ Acetylene, Chap. 5.). Propane is considerably denser than air, and, after escaping from leaks in equipment, can collect in hollows or spread out for long distances over a level floor, and can then be accidentally ignited far from the leak source.

When working with oxygen, it is essential that the surfaces of the equipment with which it comes into contact (especially manometers, valves, tubes, and hoses) are free from oil and grease in order to prevent spontaneous ignition.

Accident prevention guidelines must be observed [37].

8. Economic Importance

Welding is a basic operation in the fabrication of equipment and machinery from individual components. Although the autogenic technique is still important, most welding operations are now carried out by gas-shielded processes. The growth of computer methods for the control and regulation of processes and the establishment of data banks and expert systems is typical of the progress being made in solving welding problems. Modern computer technology is used not only for controlling the process of welding, but also for establishing weld geometry and for controlling the burners.

Manual electrodes constitute an important part of welding technology, as they can readily be modified by changing the coating or the cored wire to suit the composition of a new material.

Beam welding methods are now of great economic importance. The increased speed of this welding process gives not only improved economics, but also a reduction in the extent of the heat-affected zone, and therefore continuing improvement of the conservation of the quality of the material in the region of the welding or cutting zone.

The spot welding technique occupies a special position. Its effect on added value is very significant, e.g., in automobile construction. Economic comparison with the other welding processes is not easy. The development of spot welding technology is also dominated by electronics.

Developments in the technology of cutting go hand-in-hand with those in welding. Although autogenic cutting is still the most widely used process, beam methods show the greatest growth rate. Three-dimensional cutting with high-power lasers is already an important feature of modern fabrication processes.

9. References

General References

1. J. Ruge: *Handbuch der Schweißtechnik, Werkstoffe – Verfahren – Fertigung,* Springer Verlag, Berlin – Heidelberg – New York 1985.
2. U. Dilthey: *Schweißtechnische Fertigungsverfahren,* part 1: Schweiß- und Schneidtechnologien; part 2: Verhalten der Werkstoffe beim Schweißen, VDI-Verlag, Düsseldorf 1994.
3. A. W. Kleiber: *Fachwörterbuch Schweißen, Schneiden und verwandte Verfahren,* part 1, Englisch/Deutsch; *Dictionary Welding, Cutting and Allied Processes,* part 1, English/German; Deutscher Verlag für Schweißtechnik (DVS), Düsseldorf 1985.
4. A. W. Kleiber: *Fachwörterbuch Schweißen, Schneiden und verwandte Verfahren,* part 2, Deutsch/Englisch; *Dictionary Welding, Cutting and Allied Processes,* part 2, German/English; DVS, Düsseldorf 1986.

5. A. Dollinger: *Technik Wörterbuch "Schweißtechnik," Englisch, Deutsch, Französisch, Russisch, Polnisch, Slowakisch,* Verlag Technik, Berlin 1974.
6. DIN-DVS-Taschenbuch 290, Europäische Normung Schweißtechnik, 1st ed., Beuth-Verlag, Berlin – Köln, DVS, Düsseldorf 1995.
7. N. N. Rykalin: *Die Wärmegrundlagen des Schweiß-vorganges,* Verlag Technik, Berlin 1957.
8. J. F. Lancaster: *The Physics of Welding,* Pergamon Press, Oxford 1984.
9. H. Munske: *Handbuch des Schutzgasschweißens, Elektrotechnische Grundlagen – Schweißanlagen – Einstellpraxis,* 2nd ed., DVS, Düsseldorf 1975.
10. A. Neumann: *Schweißtechnisches Handbuch für Konstrukteure,* **Fachbuchreihe Schweißtechnik, vol. 80/1 – 3,** parts 1 – 3, DVS, Düsseldorf 1986 – 1990.
11. G. Aichele: *Kalkulation und Wirtschaftlichkeit in der Schweißtechnik,* **Fachbuchreihe Schweißtechnik, vol. 83,** DVS, Düsseldorf 1985.

Specific References

12. R. Killing: *Handbuch der Schweißverfahren,* part 1, Lichtbogenverfahren, Fachbuchreihe Schweißtechnik, vol. 76/1, DVS, Düsseldorf 1991.
13. A. Neumann, D. Kluge: *Fertigungsplanung in der Schweißtechnik,* **Fachbuchreihe Schweißtechnik, vol. 106,** DVS, Düsseldorf 1992.
14. F. Rapatz: *Die Edelstähle,* 5th ed., Springer Verlag, Berlin – Göttingen – Heidelberg 1962.
15. V. Horn: *Schweißtechnischer Gefügeatlas,* Düsseldorf 1974.
16. G. R. Lohrmann, H. Lueb: *Kleine Werkstoffkunde für das Schweißen von Stahl und Eisen,* **Fachbuchreihe Schweißtechnik,** DVS, Düsseldorf 1984.
17. DIN-Taschenbuch 65, Schweißtechnik 2., Normen über Autogenverfahren, Thermisches Schneiden, Thermisches Spritzen und Arbeitsschutz, 6th ed., Beuth-Verlag, Berlin – Köln 1993.
18. C. Hase, W. Reitze: *Lehrbuch des Gasschweißens und verwandter Autogenverfahren,* Girardet-Verlag, Essen 1980.
19. W. Conn: *Die technische Physik der Lichtbogenschweißung,* Springer Verlag, Berlin – Göttingen – Heidelberg 1959.
20. C. Hase, W. Reitze: *Lehrbuch des Lichtbogenschweißens,* Girardet-Verlag, Essen 1976.
21. DIN-Taschenbuch 8, Schweißtechnik 1., Normen über Schweißzusätze, Fertigung, Güte und Prüfung, 11th ed., Beuth-Verlag, Berlin – Köln 1990.
22. "Physik und Technologie des Plasmastrahls in der Schweiß-, Schneid- und Spritztechnik", *DVS Ber.* **37** (1975).
23. "Qualitätssteigerung durch Strahlverfahren," *DVS Ber.* **58** (1979).
24. "Strahltechnik," *DVS Ber.* **63** (1980).
25. G. Aichele: *Schutzgasschweißen, Verfahren – Anwendung – Wirtschaftlichkeit,* 4th ed., Messer-Griesheim, Frankfurt/Main 1981.
26. P. Müller, L. Wolff: *Handbuch des Unterpulverschweißens,* **parts 1 – 5, Fachbuchreihe Schweißtechnik, vol. 63 S – V,** DVS, Düsseldorf 1979 – 1983.
27. D. Böhme, F. Hermann: *Handbuch der Schweißverfahren,* part 2, Autogenverfahren – Thermisches Schneiden – Elektronenstrahl- und Laserstrahlschweißen – Reib-, Ultraschall- und Diffusionsschweißen, Fachbuchreihe Schweißtechnik, vol. 76/2, DVS, Düsseldorf 1992.
28. H. Schultz: *Elektronenstrahlschweißen,* **Fachbuchreihe Schweißtechnik, vol. 93,** DVS, Düsseldorf 1989.
29. *Laserstrahltechnologien in der Schweißtechnik,* **Fachbuchreihe Schweißtechnik, vol. 86,** DVS, Düsseldorf 1989.
30. L. Pfeifer: *Fachkunde des Widerstandsschweißens,* Girardet-Verlag, Essen 1969.
31. F. W. Strassburg: *Schweißen nichtrostender Stähle,* **Fachbuchreihe Schweißtechnik, vol. 67,** DVS, Düsseldorf 1982.
32. A. L. Schaeffler: "Constitution Diagram for Stainless Steel Weld Metal," *Met. Prog.* **37** (1949) Nov., 680, 680 B.
33. S. Zschötke: *Kleine Werkstoffkunde für Nichteisenmetalle,* **Fachbuchreihe Schweißtechnik, vol. 50,** DVS, Düsseldorf 1967.
34. W. Künnecke: "Das Heizelementstumpfschweißen,"

Schweisstechnik (Zürich) **70** (1980) no. 5, 101 – 111.
35. G. Kraume, A. Zober: *Arbeitssicherheit und Gesundheitsschutz in der Schweißtechnik,* **Fachbuchreihe Schweißtechnik, vol. 105,** DVS, Düsseldorf 1989.
36. Maximale Arbeitsplatzkonzentrationen (MAK) 1994 der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe der Deutschen Forschungsgemeinschaft, Mitteilung XVII, Boldt-Verlag, Boppard 1994.
37. VDI-Richtlinie 2084, Lüftung von Schweißräumen und Schweißplätzen, VDI Verlag, Düsseldorf 1965.

Wheat → Cereals and Cereal Products
Whey → Cheese, Processed Cheese, and Whey

Whiskers

AZAR PARVIZI-MAJIDI, Department of Mechanical Engineering and the Materials Science Program, University of Delaware, Newark, Delaware 19716, United States

1. **Introduction and Historical Development** 1
2. **Applications and Commercial Products** 2
3. **Asbestos Whiskers for Thermal Resistance/Insulation Applications** 3
4. **Whiskers for Reinforcement Applications** 5
4.1. **Silicon Carbide Whiskers** 5
4.2. **Silicon Nitride and Sialon Whiskers** . 9
4.3. **Alumina and Mullite Whiskers** 10
4.4. **Aluminum Borate Whiskers** 11
4.5. **Other Types of Reinforcing Whiskers** 11
5. **Whiskers for Electronics Applications** 11
5.1. **Carbon Whiskers** 11
5.2. **Other Types of Whiskers for Electronics Applications** 12
6. **Toxicology and Occupational Health** . 13
7. **References** 13

1. Introduction and Historical Development

Whiskers are filamentary single crystals with an aspect ratio (i.e., length-to-diameter ratio) of usually over 10 and a diameter in the range of 0.01 – 10 µm [1]. Such crystal shapes have a very large surface area to volume ratio. The high degree of chemical purity and structural perfection of whiskers, which is due to their single-crystal morphology and small diameter, make them the strongest form of materials. The strength of some whiskers approaches the theoretical strength of the material predicted by OROWAN:

$$\sigma_{\text{theo}} = (E\gamma)^{1/2}/a$$

where E is the Young's modulus, γ is the surface energy per unit area, and a is the equilibrium spacing of the atomic planes. The absence of grain boundaries and line defects in whiskers also ensures better creep resistance and, therefore, higher temperature capability compared to polycrystalline fibers or bulk materials.

Being single crystals, most whiskers have a noncircular cross-section. The shape of the cross-section depends on the crystal structure and the growth orientation of the whisker. Thus, it can be triangular, hexagonal, rhombohedral, or another shape. For a noncircular cross-section, whisker diameter is defined as the square root of the cross-sectional surface area. The shape of the whisker affects the total whisker – matrix interfacial area for whiskers used as reinforcements, conducting fillers, etc. in ceramics, metals, or polymeric materials. The interface in turn has a strong influence on the physical and mechanical properties of the composites.

The existence of crystal whiskers has been known for centuries. In his *Treatise on Ores and Assaying*, published in 1574, ERCKER discusses the natural and artificial occurrence of silver whiskers, known as hair silver. However, the importance of whiskers as a form of materials with unique properties was recognized in 1952 when HERRING and GALT at Bell Laboratories discovered that whiskers have mechanical properties close to those of dislocation-free crystals [2]. A rush of activities then followed in the growth, modeling, and property characterization of whiskers. In 1958, The Air Force Office of Scientific Research, Air Research and Development Command, and the General Electric Research Laboratory sponsored an international conference on crystal growth, of which two major sessions were dedicated to whiskers [3]. In a comprehensive review presented at that conference, NABARRO and JACKSON have documented the history of natural and synthetic whiskers and the state of science, up to that point, of whisker growth from vapor, melt and solution, by chemical decomposition, by electrolysis, and under mechanical stress [4]. The majority of whiskers grown then were whiskers of metals, including tin, zinc, cadmium, copper, and iron [5–7].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_229

Early investigations of whiskers primarily focused on dislocation structures and their role in the mechanical properties, although some studies of the ferromagnetism in iron whiskers were also conducted [8,9]. Also at that same conference, Brenner gave an extensive review of the mechanical properties of whiskers, including whiskers of tin grown from solid; whiskers of zinc, cadmium, silicon, and mercury obtained from vapor; whiskers of iron, copper, nickel, cobalt, and silver grown from the reaction of metal halides; and whiskers of sodium chloride and potassium chloride prepared by precipitation from solution [10]. An example of the stress – strain curves for iron and copper whiskers are shown in Figure 1. Figure 1 illustrates a large deviation from linearity in the elastic behavior of iron whiskers, which show an elastic strain close to 5 %. Similar deviations from Hooke's law have also been observed in whiskers of zinc and some other metals, but not in copper [11].

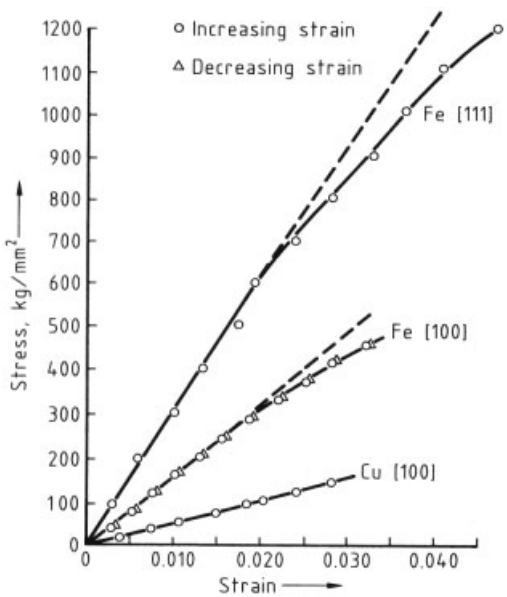

Figure 1. Stress – strain curves of iron and copper whiskers pulled in tension [10]

With the advent of metal- and ceramic-matrix composites, a new wave of activity began in the 1980s towards the development of ceramic whiskers as reinforcements for metals, ceramics, and some polymers. A fair amount of research has been also going on in the area of whiskers of conducting, semiconducting, and superconducting materials.

This article provides an overview of the present day whiskers with actual or potential use in thermal, mechanical, and electronics applications. Effort is made to cover not only whiskers that are commercially available but also those that are still in early stages of investigation and development.

2. Applications and Commercial Products

Asbestos whiskers and fibers have traditionally been used for thermal insulation in buildings as well as for heat-shield applications. The discovery of the health hazards of asbestos, however, has led to a severe limitation and regulation of asbestos use in the construction industry. At present, the most active area of whisker development is for reinforcement applications in metal- and ceramic-matrix composites. The bulk of the work in whisker-reinforced metal-matrix composites has been with aluminum matrices [12,13]. In ceramic-matrix composites, in addition to alumina [14,15], which constitutes the major matrix material investigated, other matrices such as boron nitride [16] and zirconia [17] have also been considered. Although whiskers are usually quite short, they possess large aspect ratios, usually over 20, due to their small diameters and are thus efficient reinforcements. Whisker diameter is also a critical parameter in terms of strength; increasing the diameter raises the probability of the presence of defects, hence lowering the strength.

Whiskers are incorporated in metals to increase the stiffness, creep resistance and wear resistance of the matrix, while in ceramics their role is to increase toughness, creep resistance, and wear resistance. Toughness enhancement in whisker-reinforced ceramic-matrix composites occurs through a variety of mechanisms, which include crack deflection [18,19], crack bridging [20], and whisker pullout [21,22]. Three- to fourfold increases in the toughness of alumina have been reported on incorporation of SiC whiskers [14,23]. Figure 2 illustrates a SiC whisker pulled out in SiC(w)/Al_2O_3 composite. Whiskers are used both as the sole reinforce-

ment and in combination with fibers to provide added toughness. Whisker-reinforced ceramics are used in cutting tools and considered as high-temperature materials for engine components.

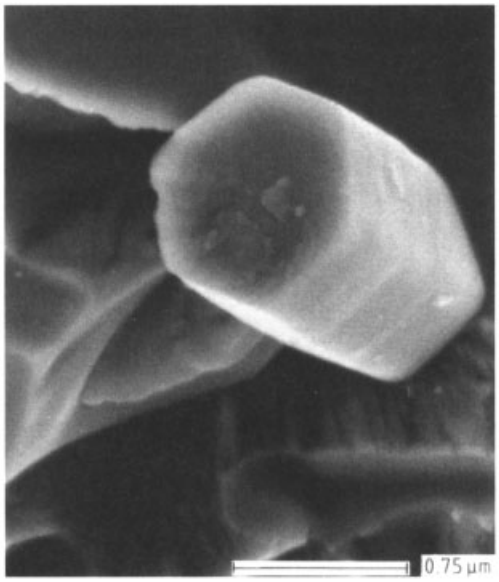

Figure 2. Whisker pullout in the SiC(w)/Al_2O_3 ceramic composite systems

The potential use of whiskers as conducting, semiconducting, or superconducting fillers for some polymers has also been recognized and is under investigation.

The range of whiskers commercially available is still quite limited. Table 1 summarizes some commercially available reinforcing ceramic whiskers. In most cases, whisker prices are very high, at least several hundred dollars per pound, because of the small production volume and/or the processing route. The high cost of whiskers is still a major barrier to their widespread use in composites. Among the commercially available whisker reinforcements for composites aluminum borate is one of the cheapest.

3. Asbestos Whiskers for Thermal Resistance/Insulation Applications

Asbestos (→ Asbestos) is a general name for a variety of fibrous minerals known as amphiboles, which are hydrated silicates containing chiefly Mg, Na, Fe, and Al. The majority of commercial asbestos whiskers or fibers come from the mineral chrysotile [$Mg_6Si_4O_{11}(OH)_6 \cdot H_2O$] but other minerals, including actinolite [$CaMg_nFe_{3-n}(SiO_3)_4$], tremolite [$Ca_2Mg_5H_2(SiO_3)_8$], crocidolite [$NaFe(SiO_3)_2$], and anthophyllite [(MgFe)-SiO_3] are also important sources [24].

The size and morphology of asbestos filaments are as broad as their compositions. Filament size can range from <1 µm to >2 mm in diameter, and from a few micrometers to nearly 100 mm in length. From their structures, asbestos filaments may or may not be considered as genuine whiskers. Some fibrous silicates, such as actinolite and tremolite, have chain structures based on the amphibole (SiO_3) chain [25]. Chrysotile fibers, however, are hollow tubes, as shown in Figure 3 [26, 27]. A structural model has been proposed for chrysotile asbestos which consists of alternating silicon – oxygen layers in tetrahedral coordination and magnesium – hydroxyl layers in octahedral coordination [28]. Each silicon – oxygen layer is strongly bonded to a magnesium-hydroxyl layer, forming a double layer. The double layers, however, are weakly bonded together. Because the alternate tetrahedral and octahedral layers have different lattice parameters, the layered structure curls, forming a hollow scroll that looks like a whisker or fiber.

Long asbestos fibers are woven into fabrics and used in brake linings, fire-proof articles, curtains, heat shields, etc. Because of the high chemical resistance of asbestos, asbestos fabrics have also been used as acid filter cloths. Short fibers have been traditionally mixed with a binder and made into insulating boards, paper, shingles, piping blankets, etc. However, since the discovery of asbestosis, a form of lung cancer caused by the inhalation of asbestos dust, the use of asbestos in construction has been severely curtailed and regulated.

Asbestos fibers have tensile strengths ranging from 0.5 GPa to over 4 GPa and a melting point above 1480 °C. Above 980 °C, the hydrous silicate loses its water, thus producing a cooling effect.

Because of the health hazards of micron-size asbestos fibers, a large effort was launched in the 1970s towards the development of larger-diameter refractory fibers. This led to the de-

Table 1. Selected whiskers and their properties

Material	Designation	Phase	Manuf.	Diameter, µm	Density, g/cm^3	Coefficient of thermal expansion, K^{-1}	Young's modulus, GPa	Tensile strength, GPa	Length, mm
SiC	Silar Sc-9	α	ACMC	0.6	3.2	4.7×10^{-6}	689	6.89	10 – 80
SiC	SCW	95 % β – 5 % α	Tateho Chemical	0.05 – 0.2	3.21	$(4.4-4.7)\times10^{-6}$	483	20.68	10 – 40
SiC	Tokama	β	Tokai Carbon	0.1 – 0.5	3.19		400 – 700	3 – 14	50 – 200
SiC	VLS	β	Los Alamos	5 – 6			578	8.40	ca. 5 000
SiC	Versite		Versar	1 – 5			ca. 560	ca. 11	up to 3 000
Si_3N_4	SNW	>97 % α	Tateho Chemical	0.2 – 0.5					50 – 300
Si_3N_4	Versite		Versar				ca. 245	ca. 5.6	
Si_3N_4	UBE-SNWB	β	UBE Industries	0.1 – 2					1 – 20
C			NiKKiso	<1.0					up to 2 000
AlBO *	ALBOREX-G		Shikoku Kasei Industries	0.5 – 1.0	2.93	1.9×10^{-6}	400	8	10 – 30
SiO_2	XPVI	Amorphous	J. M. Huber Corp.	up to 0.02	2.25				10 – 20 **

* Aluminum borate: $Al_{18}B_4O_{33}$.

** Aspect ratio.

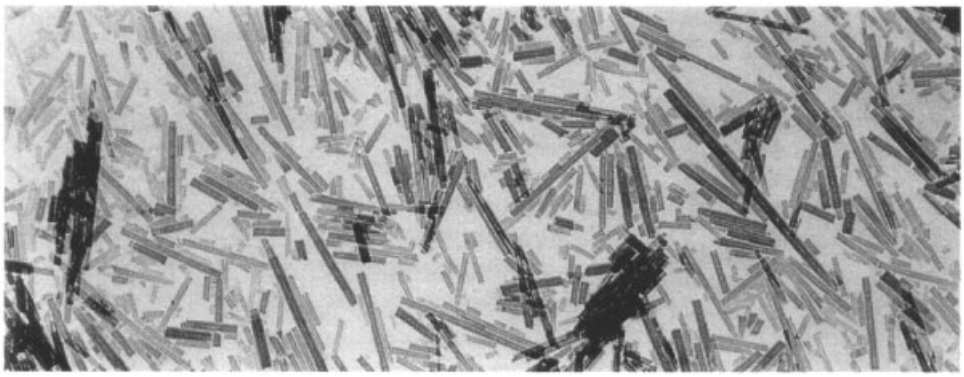

Figure 3. Synthetic chrysotile whiskers [26]

velopment of alumina fibers and staples. Since the primary application of alumina and mullite whiskers and fibers is as reinforcements, they are discussed in Chapter 4.

4. Whiskers for Reinforcement Applications

4.1. Silicon Carbide Whiskers

General Characteristics and Uses. SiC whiskers are presently the most widely used whisker reinforcements for ceramic-matrix composites [15, 29] as well as metal-matrix, particularly aluminum [12, 30, 31], composites.

Silicon carbide has two polymorphs: α-SiC with a wurtzite hexagonal or a rhombohedral structure and β-SiC with a zinc blende cubic structure. A wide variety of SiC polytypes have been reported [32]. These polytypes are usually described using Ramsdell notation [33]. The Ramsdell notation (e.g., 2 H, 1200 R, 3 C, etc.) consists of a number, representing the number of silicon and carbon double layers in the unit cell, followed by a letter identifying the Bravais lattice type (C for cubic, H for hexagonal, R for rhombohedral).

The majority of commercially available SiC whiskers are of β type grown in the $\langle 111 \rangle$ direction, although some have α or a mixture of α and β structures. The β-SiC whiskers usually have mechanical properties superior to those of α-SiC whiskers, probably due to their smaller defect density and better consistency. The physical appearance (e.g., shape, size, and color) and structural characteristics of these whiskers are highly dependent on the manufacturing process. β-SiC whiskers from various manufacturers, even when produced by basically the same processing route, show colors ranging from white to light-brown to dark-green, and aspect ratios in the range of 1 – 100 µm [1]. Such variations, in turn, result in a broad range of the mechanical properties. Table 1 includes SiC whiskers from a variety of manufacturers, along with some of their physical characteristics and properties.

Production. There are several mechanisms for SiC crystal growth. One method is the vapor-solid (VS) mechanism [34]. β-SiC whiskers have been grown using chemical vapor deposition of β-silicon carbide onto a carbon substrate by hydrogen reduction of methyltrichlorosilane gas at 1500 °C [35]. At present, however, the most commonly applied methods of production are the carbothermal reduction of silica [36] and the vapor – liquid – solid (VLS) process [37, 38].

The majority of commercial SiC whiskers are obtained by the carbothermal reduction of silica. The overall carbothermal reaction is written as [39]:

$$SiO_2(s) + 3\,C(s) \rightarrow SiC(s) + 2\,CO(g) \quad (1)$$

Below 1400 °C most of the SiC is produced by reaction (1). At higher temperatures, however, the solid-state reaction is replaced by the gas – solid reactions:

$$SiO_2(s) + CO(g) \rightarrow SiO(g) + CO_2(g) \quad (2)$$

$$SiO(g) + 2\,C(s) \rightarrow SiC(s) + CO(g) \quad (3)$$

$$2\,SiO(g) \rightarrow Si(s) + SiO_2(g) \quad (4)$$

Other possible reactions for SiC whisker formation are [36,40]:

$$SiO(g) + 3\,CO(g) \rightarrow SiC(s) + 2\,CO_2(g) \quad (5)$$

and [41]:

$$3\,SiO(g) + CO(g) \rightarrow SiC(s) + 2\,SiO_2(s) \quad (6)$$

The most economic and widely used production method based on carbothermal reduction of silica is the pyrolysis of silica and carbon obtained from rice hulls [41–45]. The Silar SC-9 whiskers (Table 1) are manufactured in this way. Rice hulls contain ca. 15 – 20 % silica and are a cheap and abundant source of silicon carbide whiskers. Sub-micron whiskers are produced by heating the rice hulls initially at 700 – 900 °C for 1 – 3 h in the absence of air or in an inert atmosphere, and subsequently above 1500 °C in a flowing inert or reducing atmosphere. The maximum yield of SiC is ca. 60 %, of which only 10 – 15 % is in the form of whiskers, while the remainder is SiC particulates [46]. Higher yields have been reported for the production of SiC whiskers by direct pyrolysis of rice hulls in a single-stage vacuum-pyrolysis process [44]. Four competitive processes have been identified during direct pyrolysis of rice hulls: (1) formation of SiC whiskers, (2) formation of SiC polycrystals, (3) crystallization of amorphous silica, and (4) formation of graphite from amorphous carbon [44]. More recently, Krishnarao et al. have reported on the formation of SiC whiskers from compacts of raw rice hulls without coking and catalyst [40]. Using a pyrolysis temperature of 1600 °C, they obtained α- and β-SiC whiskers, believed to have been formed by reaction of SiO(g) with CO(g).

Cerovic et al. (1995) investigated the effect of boric acid as a catalyst on the carbothermal reduction of silica [39]. They synthesized β-SiC whiskers by carbothermal reduction of silica, which was obtained from silica sol and saccharose in the presence of boric acid, in a carbon crucible. The presence of boric acid enhanced the crystallization of SiC into long, wide, regularly shaped whiskers. This was attributed to the formation of B_2O_3 which is thoroughly miscible in SiO_2 and provides for more intimate contact between SiO_2 and carbon.

In the VLS process particles of a catalyst metal are placed on a suitable substrate. The particles melt at the high temperatures of the reaction chamber. A mixture of gaseous C and Si compounds is then passed over the catalyst, whereby Si and C are extracted from the gases by the liquid catalyst. Crystalline SiC grows under the catalyst droplets, lifting the droplets off the surface of the substrate, once the liquid catalyst becomes supersaturated in C and Si. Figure 4 illustrates a schematic of the VLS process for SiC whisker growth. The VLS process has been adopted by Los Alamos National Laboratory and by Carborundum Company for the manufacture of β-SiC whiskers. A summary of advances made in the VLS process over the twenty five years of its development is given in [47].

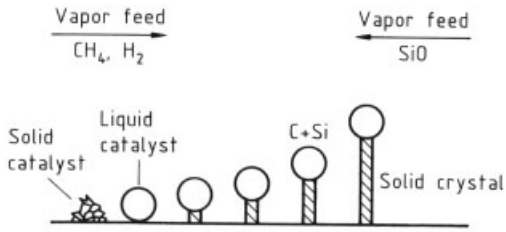

Figure 4. Schematic of the VLS process for SiC whisker growth [72]

Bootsma et al. studied SiC whisker growth by the VLS method, using iron as the catalyst, and found that the growth rate was controlled by the rate of supply of the reaction feed gases [48]. They also observed that the SiC phases obtained depended on the substrate used. Milewski and co-workers [49,50] further investigated and optimized the process. They used stainless steel catalyst particles, a graphite substrate, and a mixture of SiO (generated by the reduction of SiO_2 with carbon), CH_4, H_2, CO, and N_2 gases. The CO gas was added in order to control the rate of production of SiO(g) from Reaction (1). SiC whiskers grew by the following reaction

$$CH_4(g) + SiO(g) \rightarrow SiC(s) + H_2O(g) + H_2(g) \quad (7)$$

It has been proposed that Reaction (3) may also take place, leading to SiC whisker formation [36].

It is observed that the iron droplets found at the tips of VLS-grown SiC whiskers are smaller than the original particles placed on the substrate, indicating that iron transport may have

taken place. URRETAVIZCAYA and PORTO LOPEZ investigated the mechanism of iron transport, using 20 – 30 µm Fe catalyst particles on a graphite substrate [51]. They also distributed small cubes, made of a mixture of silicon and carbon in isopropyl alcohol and serving as SiO generators, along the substrate surface. For comparison, in some tests separate cubes of pure SiO_2 and pure C were also distributed on the substrate alongside the SiO generator cubes. It was observed that extensive SiC whisker growth took place only when the Fe particles were placed at the upstream end of the substrate rather than being spread uniformly over it. Whiskers, carrying a droplet of catalyst notably smaller (2 – 3 µm) than the initial Fe particles at their growth end, grew in the high-temperature zone (1445 – 1475 °C) remote from the initial location of the particles. These observations clearly indicated that iron had been transported within the reaction chamber. The authors concluded that among the various possible transport mechanisms, vapor-phase transport was the only one likely to occur under the processing conditions. They further observed that the most extensive whisker growth happened on the surface of SiO generators where a high concentration of SiO existed. Whisker growth also took place, albeit to a smaller extent, on the surfaces of pure C cubes and the substrate but not on the pure SiO_2 surface, where carbon was not present.

Silicon carbide whiskers with much larger dimensions (diameter of 20 – 30 µm and lengths up to 2 mm) than those reported above have been grown from the vapor phase in the temperature range 2200 – 2600 °C [52]. These whiskers are reported to have a hexagonal shape and grow in the direction of heat flow.

Because of the difficulty in homogeneous mixing of whiskers with ceramic powders and the potential health hazards of handling small whiskers, several attempts have been made to form SiC whiskers in situ in a ceramic matrix. WANG and FISCHMAN prepared β-SiC whiskers in situ in a Si_3N_4 matrix by the reaction of silicon nitride with graphite in an argon atmosphere [53]. β-SiC formation began at 1400 – 1450 °C. The whiskers were needle-like with lengths and diameters depending on temperature. The product phases were dependent upon the Si_3N_4 : C ratio of the raw materials and the processing temperature. This "chemical mixing process" route was also followed by YAMADA et al. for in situ preparation of SiC(w) – Si_3N_4 composites [54]. JAGOTA and PARVIZI-MAJIDI used sol – gel processing to produce in situ composites of SiC whiskers in a mullite matrix [55]. This was achieved by mixing mullite precursor gel with carbon-doped silica gel and crystallization above 1400 °C.

Structure and Properties. As described above, SiC whiskers obtained from rice hulls are a mixture of α- and β-phases and are characterized by a hexagonal cross-section. The structural and chemical characteristics of Silar whiskers (Table 1) have been studied in detail [42, 56]. They contain a high density of planar faults lying on close-packed planes perpendicular to the whisker axis. The crystal structure of the whiskers ranged from faulted and twinned 3C polytype to thin lamellae of α and β polytypes normal to the whisker axis. Core regions of whiskers are often filled with small cavities ranging in size from 1 to 20 nm, accompanied by partial dislocations. Figure 5 shows the defect structure of a Silar SiC whisker. PICKARD et al. studied structural defects including twins, stacking faults, grain boundaries, and microvoids in β-SiC whiskers obtained from various manufacturers and found them to be highly dependent on the manufacturing process [1].

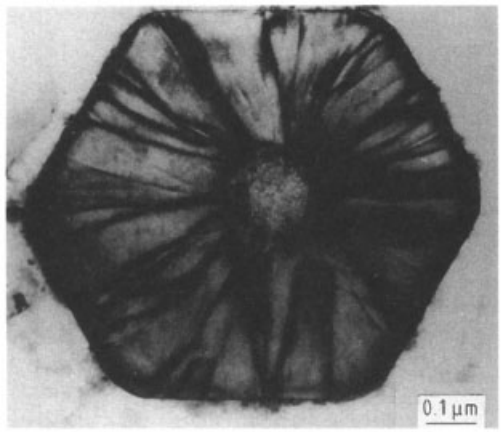

Figure 5. The defect structure of Silar SC-9 SiC whisker (a) perpendicular, and (b) parallel to the whisker axis [56]

Silicon carbide whiskers show excellent oxidation resistance in air up to at least 1000 °C. The thermal stability of Silar SC-9 whiskers un-

der oxidizing conditions has been studied [57]. Minimal oxidation occurred for whiskers heat treated for 5 h below 1000 °C. Heat treatment at 1000 °C for up to 96 h resulted in partial oxidation and formation of amorphous silica. Whiskers exhibited extensive oxidation after exposure at 1300 °C for 24 h, accompanied by the emergence of a nonstoichiometric crystalline silica phase. The same silica was observed at 1500 °C where whiskers were oxidized severely after only 1 h of exposure.

Whiskers grown by the VLS method consist almost completely of the cubic β-phase with their growth axis along the $\langle 111 \rangle$ direction. They have a rounded triangular cross-section with an effective diameter of 0.5 – 1.0 µm and lengths ranging from 1 mm to several centimeters. The VLS-grown SiC whiskers have a lower concentration of defects and therefore a higher strength and thermal conductivity than whiskers from rice hulls. These whiskers have been found to be remarkably uniform in size, a characteristic that has been attributed to the existence of a threshold catalyst droplet size for whisker growth [51]. The threshold size, however, depends on the substrate and the processing parameters used.

Petrovic et al. (1985) measured the tensile strength and Young's modulus of VLS-grown SiC whiskers by direct experiments and found quite a large scatter in both properties [58]. The strength decreased with increasing whisker length. Whiskers exhibited a considerable variation in their diameter, which contributed to the scatter in strength. The strength averaged at 8.40 GPa but reached as high a value as 23.74 GPa, which is almost half of the calculated theoretical strength of 46.3 GPa for SiC. The authors predicted a Young's modulus of 511 GPa for cubic SiC along the $\langle 111 \rangle$ direction. The experimentally determined modulus was about 15 % higher at an average value of 581 GPa.

Whisker surface chemistry and morphology affect the fabrication and properties of the resulting composites. Karasak and co-workers examined the chemistry and morphology of commercial SiC whiskers from various sources, using photoelectron spectroscopy (XPS) and scanning transmission electron microscopy (STEM) [59–61]. They found that all whiskers contained some carbon in their bulk composition. These investigators identified four categories of whisker surface chemistry:

1) A high-oxygen (>20 atom %) surface layer resembling SiO_2, prevalent in the earlier versions of whiskers when a HF acid wash was not employed to remove the silica surface layer
2) A low-oxygen (10 – 20 %) layer resembling Si – O – C glass (black glass)
3) A high-oxygen layer resembling Si – O – C glass
4) A predominantly hydrocarbon layer in those cases where a carbon layer had been deposited on the surface of the whiskers

All commercial SiC whiskers contain particulate matter, defined as particles of aspect ratio <5. Particle sizes typically range from submicron to >50 µm. Lee and Sheargold measured the particulate levels of SiC whiskers using an optical microscopy method coupled with wet sieve analysis [62]. They showed that the total particulate level ranged from 19 to 46 % among whiskers produced by different manufacturers. The authors further examined the effect of the particulate content on the flexural strength and fracture toughness of SiC(w) – borosilicate glass composites and concluded that particulate levels found in commercial whiskers did not significantly degrade the mechanical properties of the composite.

4.2. Silicon Nitride and Sialon Whiskers

General Characteristics and Uses. Silicon nitride whiskers, although still far less established than silicon carbide whiskers, have received increasing attention as reinforcements for composites because of silicon nitride's good high-temperature strength and thermal-shock resistance. Also, whisker – matrix compatibility is a major factor in the selection of the reinforcement for ceramic- and metal-matrix composites where the high temperatures encountered during processing or applications can lead to undesirable interfacial reactions. Therefore, for some matrices, Si_3N_4 whiskers may be a better choice than SiC whiskers. For example, Matsubara et al. have reported that Si_3N_4(w) – Al composites offer better tensile strength and machinability than SiC(w) – Al composites [63].

Silicon nitride (Si_3N_4) whiskers have two crystallographic modifications: α-Si_3N_4 with a

trigonal structure and β-Si_3N_4 with a hexagonal structure. Whiskers are available in a wool form with average diameter of 10 µm and in small-diameter needle form. Table 1 lists some commercially available Si_3N_4 whiskers and their characteristics.

Production and Structure. A variety of methods have been reported for the formation of α- and β-Si_3N_4 whiskers. In the majority of these methods, the VLS mechanism was found to be responsible for whisker growth. GRIBKOV et al. formed α-Si_3N_4 whiskers from a mixture of SiO_2 and Si in an $N_2 + H_2$ atmosphere and found droplets of Fe – Si or Al – Si alloy at whisker tips [64]. HAYASHI et al. nitrided the $SiO_2 - C - Na_3AlF_6$ system and obtained a whisker product containing ca. 85 % β'-sialon whiskers and 15 % α-Si_3N_4 whiskers [65]. Again, the presence of droplets at whisker ends pointed to a VLS growth mechanism.

Following a similar process to that of HAYASHI et al., WANG and WADA also obtained β'-sialon and α-Si_3N_4 whiskers, in both wool-like and short-whisker forms [66]. However, they attributed the growth mechanism to a VS process. They observed that the phases formed depended on the starting materials and the gas flow rate. The major whisker phase was β'-sialon when Na_3AlF_6 was used as the catalyst in the nitridation of SiO_2 and C. These whiskers, which had a rectangular cross-section, an effective diameter ranging from 0.5 to 1.2 µm, and a $\langle 210 \rangle$ growth axis, were believed to have formed by the following reaction:

$$(6-z)SiO(g) + (4-z/2)N_2(g) + (6-2z)CO(g) + zAl(g) \rightarrow Si_{(6-z)}Al_zO_zN_{(8-z)} + (6-2z)CO_2(g) \quad (8)$$

When NaF was substituted for Na_3AlF_6, the major whisker phase became α-Si_3N_4 with a ribbon-like morphology, a width of about 1 µm, a thickness of less than 0.1 µm, and a growth axis along $\langle 211 \rangle$. The following reaction was proposed as the most likely for silicon nitride whisker formation:

$$3\,SiO(g) + 3\,CO(g) + 2\,N_2(g) \rightarrow Si_3N_4(g) + 3\,CO_2(g) \quad (9)$$

WANG and WADA also observed oxynitride (Si_2N_2O) whiskers along with sialon and silicon carbide whiskers, except when SiO_2 was replaced by Si in the starting materials.

The synthesis of β'-sialon whiskers has also been reported in [65,67]. In the latter work, wool-like β'-sialon whiskers, about 1 cm long with diameters of 0.2 – 2 µm, were prepared via a VLS mechanism by carbothermal reaction of silica fume and α-alumina under nitrogen gas. A small amount of $FeCl_3$ was used as catalyst.

GOPALAKRISHNAN and LAKSHMINARASIMHAM prepared α-Si_3N_4 whiskers by nitridation of silicon whiskers with ammonia gas [68,69]. They obtained the silicon whiskers from a gas-phase reaction between $SiCl_4$ and zinc vapor.

HOMENY et al. [70] characterized commercial β-silicon nitride whiskers from Ube Industries (Table 1). These whiskers are found to be close to stoichiometic silicon nitride of an entirely β-Si_3N_4 phase, with oxygen and yittrium as the major impurities. They grow along the [001] axis and exhibit remarkable perfection.

4.3. Alumina and Mullite Whiskers

General Characteristics and Uses. Among the oxide whiskers, mullite and alumina whiskers are the most widely studied for the reinforcement of ceramics, particularly zirconia [17]. The low thermal conductivity of alumina and mullite also makes them excellent candidates for thermal insulation applications. Indeed, the development of Saffil staple alumina fiber with a diameter of about 3 µm by ICI in the late 1960s was in response to the need for larger diameter refractory fibers to replace submicron asbestos fibers as thermal insulating materials.

Mullite is the only crystalline phase in the aluminosilicate system that is stable over a wide range of composition and temperature; the composition can range from $3\,Al_2O_3 \cdot 2\,SiO_2$ to $2\,Al_2O_3 \cdot SiO_2$. Its low coefficient of thermal expansion, low creep rate, and good thermal and chemical stability (it is stable under both oxidizing and reducing atmospheres) make it a very attractive material as reinforcement for ceramics. The common structure for mullite is orthorhombic, with the lattice parameters a, b, and c changing almost linearly with composition. However, tetragonal mullite has also been observed under certain conditions. The coefficient of thermal expansion of mullite also varies almost linearly with its composition [71]. The coefficient

of thermal expansion of mullite with a 2 : 1 composition (i.e., $2\,Al_2O_3 \cdot SiO_2$) is lower and less anisotropic than that of the 3 : 2 composition.

Production and Structure. Sapphire (α-Al_2O_3) whiskers have been manufactured in a wool form with a diameter of 1 – 3 µm as well as in ribbon and needle forms. The strength of sapphire whiskers decreases with increasing whisker diameter [72]. The strength improves considerably when the surface flaws are removed by chemically polishing the whiskers in, for example, orthophosphoric acid [73].

Okada and Otsuka were among the first to prepare mullite whiskers [74]. They used tetraethoxysilane and aluminum nitrate nonahydrate as starting materials to make mullite precursor xerogels. The xerogels were then calcined, mixed with AlF_3, and pyrolyzed at 700 – 1600 °C to give mullite whiskers. The role of AlF_3 was to generate SiF_4 which served as a catalyst for the reaction and lowered the conversion temperature. The whiskers were reported to have grown along the $\langle 001 \rangle$ axis perpendicular to the $\{110\}$ plane, similar to the orientation of natural mullite. Whiskers fired at 1100 °C had a high aspect ratio and a thin cross-section of ca. 0.1 µm; those fired at 1600 °C grew to several tens of micrometers in length. The whiskers were found to be rich in alumina (ca. 80.2 wt % Al_2O_3), which was attributed to loss of silica.

Moyer and Rudolf prepared alumina-rich mullite whiskers by the pyrolysis of fluorotopaz [75]. Fluorotopaz was obtained by heating a mixture of κ-alumina (from the calcination of aluminum hydroxide) and silica above 600 °C in a tetrafluorosilane atmosphere:

$$2\,Al_2O_3 + SiO_2 + SiF_4 \rightarrow Al_2F_2SiO_4 \qquad (10)$$

Mullite whiskers, formed during the pyrolysis of fluorotopaz in the 1050 – 1400 °C temperature range, had about 76 wt % alumina (i.e., 1.88 $Al_2O_3 \cdot SiO_2$). Alumina enrichment was believed to have occurred during the pyrolysis stage as the fluorotopaz had a similar composition to the starting materials. Moyer and Rudolf further reported similarities between mullite compositions prepared from the pyrolysis of fluorotopaz and from melt. Slow cooling of a melt containing 70 – 77.5 wt % Al_2O_3 yields whiskers with 77.2 wt % alumina. Rapid quenching of the melt containing 71.8 wt % alumina, on the other hand, produced mullite with 80 wt % alumina. This led Moyer and Rudolf to suspect a liquid-phase mechanism for mullite formation from fluorotopaz.

Merk and Thomas [76] characterized the structure and composition of mullite whiskers made by pyrolysis of fluorotopaz [77] and the ancient Indian "wootz" obtained from a mixture of clay and rice hulls [78]. They reported a strong dependence of the structure of mullite whiskers on the processing conditions. Whiskers derived from fluorotopaz had a tetragonal symmetry and a thickness ranging from 0.2 to 0.4 µm; the metastable tetragonal phase transformed to orthorhombic upon heating the whiskers in vacuum at 1400 °C. Those obtained from wootz were orthorhombic with a thickness of only 10 – 30 nm. They observed that annealing in vacuum of both types of whiskers resulted in a continuous loss of silica and finally the complete conversion of mullite whiskers to α-alumina powder.

4.4. Aluminum Borate Whiskers

Aluminum borate ($Al_{18}B_4O_{33}$) whiskers were developed in 1987 by Shikoku Chemical Co. in Japan and are commercially available under the tradename ALBOREX-G. The mechanical properties of these whiskers are comparable to those of SiC and Si_3N_4 whiskers while their coefficient of thermal expansion is much lower (Table 1). The major advantage of aluminum borate whiskers over other commercially available reinforcing whiskers is their cost, which is significantly less, by a factor of at about 20, than other whiskers. Aluminum borate whiskers are very suitable for reinforcement of a variety of aluminum alloys [13, 79]. Aluminum-matrix composites reinforced with aluminum borate whiskers have shown a very low coefficient of thermal expansion and good abrasion resistance.

4.5. Other Types of Reinforcing Whiskers

Besides the whiskers described above, a number of other whiskerous materials have been introduced for use as reinforcement in ceramics,

metals or polymers. A cobweb whisker of amorphous SiO_2 with a silicon core has been commercialized by J. M. Huber Corporation under the designation XPV1 [72]. Cobweb whiskers are defined as single crystal fibrous materials with diameters below 0.1 µm. The XPV1 whisker has a diameter of about 10 nm. These whisker-like fibers are expected to have extremely high strengths due to their small size and have demonstrated great efficiency in reinforcing polymer matrices, where they can be used on their own or together with large-diameter fibers.

5. Whiskers for Electronics Applications

5.1. Carbon Whiskers

General Characteristics and Uses. Carbon whiskers have the highest specific strength and specific Young's modulus among all existing reinforcements. In addition, their very high electrical conductivity makes them excellent candidates as conducting fibers for polymeric materials. Polymers with carbon fillers have been considered for a variety of applications such as electrodes for carbon – polymer batteries, pressure sensors, electromagnetic shielding, etc. [80, 81]. The percolation theory has been widely used to predict the transformation of the conducting filler – polymer composite from an insulator to a conductor [82, 83].

Production and Morphology. One method of carbon whisker growth is by catalytic chemical vapor deposition in which C_2H_4 gas is decomposed in the presence of Cu – Ni catalyst at 400 – 600 °C [80, 81]. The degree of graphitization of carbon increases with temperature. Hence, whiskers grown at 600 °C are expected to have lower resistivity than those grown at 400 °C. Whiskers formed at 400 °C following the above procedure are reported to have an average diameter of 0.25 µm and a highly coiled morphology, while those prepared at 600 °C are less coiled, have a larger average diameter of 0.5 µm and a greater aspect ratio. A somewhat similar process was also followed by Motojima et al. who reacted a gas mixture of acetylene, sulfur compound, hydrogen, and argon at a temperature of 750 °C in the presence of nickel powder as catalyst [84]. The product of their reaction, however, was a highly microcoiled carbon fiber, rather than whisker, with a variety of coiling patterns and a low bulk resistivity. One of these patterns is shown in Figure 6.

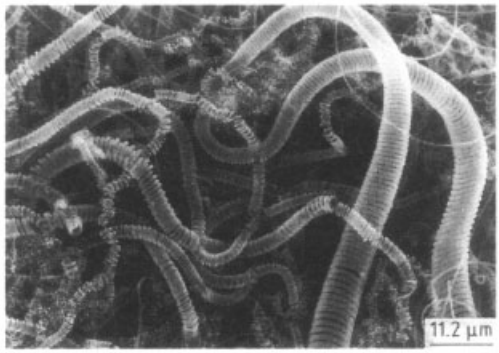

Figure 6. Coiled carbon fibers [84]

Since the discovery of fullerenes [85, 86], a host of novel carbon structures and morphologies have been obtained. Among these are buckytubes, which are nanotubes with closed ends [87, 88] and fullerene onions [89]. Wang et al. have been able to prepare open-ended carbon nanotubes in a one-step process using hydrogen arc discharge [90].

5.2. Other Types of Whiskers for Electronics Applications

Besides carbon whiskers, which have potential uses as both reinforcements and conducting fillers, a number of other whiskers of conducting, semiconducting, and superconducting materials have also been prepared. However, none of these whiskers is yet commercially available.

Aluminum nitride whiskers have been prepared by the carbothermal reaction of alumina and carbon powder at 1800 °C [91]. Whiskers thus grown are reportedly long and straight single crystals free of planar defects. They have a diameter of about 2 to 30 µm and a length in the centimeter range. Because of its high thermal conductivity, low thermal expansion coefficient, high electrical resistivity, and good dielectric properties, aluminum nitride is considered for semiconductor substrate applications.

Thick whiskers of tin oxide (SnO_2), which is used as a semiconducting sensor to detect H_2, CO, town gas, alcohol, etc., are prepared by heating a mixture of SnO_2 and tin powders at 1200 °C in air [92]. Depending on the processing conditions, these whiskers can be ribbon-shaped with a thickness of ca. 150 µm and growth direction of [010] or dendritic with rugged surfaces [93]. The dendritic whiskers exhibit a much greater hydrogen sensitivity than ribbon-shaped whiskers with smooth surfaces. The greater hydrogen sensitivity of the former morphology is attributed to its stepped surfaces.

Cadmium oxide (CdO) whiskers are grown on a CdS substrate via a VLS mechanism [94]. In this process, the surfaces of the CdS crystal are coated with a thin layer of metal catalyst (Si, Pd, Ag, Au, Al, Ni, or Pt). The coated crystals are then annealed in air at 670 – 730 °C, whereby whiskers appear on the CdS surface. In terms of the growth efficiency, the best results are reported with Au, Ag, and Pd coatings. Whiskers thus grown have a face-centered cubic lattice and their morphology largely depends on the metal catalyst used. The largest whiskers, ca. $2\times0.1\times0.01$ mm, are obtained with Au; these whiskers are parallelepiped with $\{100\}$ faces. Figure 7 shows CdO whiskers grown on CdS substrate with Ag coating.

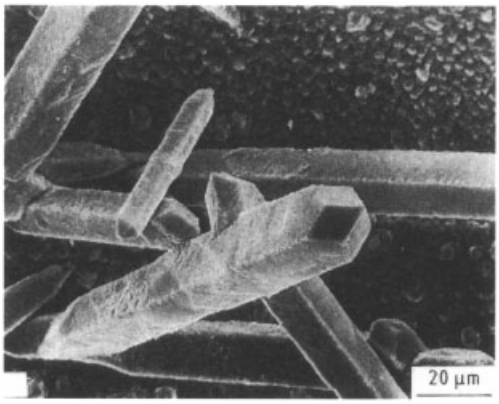

Figure 7. CdO whiskers grown on CdS crystal substrates using Ag as the catalyst [94]

In high-temperature superconductivity grain boundaries are a hindrance to achieving high critical currents. Approaches are hence taken to eliminate or minimize grain boundaries. One such approach is the use of whiskers. Upon the discovery of glass formation in the Bi – Sr – Ca – Cu – O (BSCCO) superconducting system, efforts have been made to form whiskers of these materials via a glass-ceramic processing route. Matsubara and co-workers have successfully prepared whiskers of $Bi_2Sr_2CaCu_2O_x$, known as the 2212 phase, by heating glassy melt-quenched plates of the BSCCO system doped with Al [95], Li [96], and Ga in flowing oxygen. The phases, morphology, and superconductivity of the whiskers thus obtained depend greatly on the processing conditions and the type of dopant used. Whiskers obtained with Al are reported to be plate-like single crystals about 1 – 5 µm thick, 10 – 100 µm wide, and up to 20 mm long. These are primarily the 2212 phase with about 1 % $Bi_2Sr_2Ca_2Cu_3O_x$ (2223) phase. Ga doping produces much shorter (ca. 1 mm) whiskers. Whiskers exhibit enhanced superconductivity with Li as the dopant. Kasuga et al. also prepared 2212 whiskers by heating a compacted mixture of melt-quenched $Bi_2Sr_yCa_1Cu_2O_x$ powder and alumina powder [97]. The magnitude of the pressure applied for compacting the powders was found to be an important factor in whisker growth. Whiskers obtained from specimens compacted at 20 MPa were less than 1 mm long while those grown from specimens compacted at 180 MPa were 1 – 5 mm in length.

6. Toxicology and Occupational Health

It is now well established that asbestos fibers which have a diameter less than 1 µm and aspect ratio greater than five cause irritation of the eyes, skin, and pulmonary system and may even lead to pulmonary fibrosis and cancer. Studies on the potential health hazards of other types of fibrous particles with submicron diameters are very limited. However, it is likely that whiskers may also pose health hazards similar to those of asbestos [98]. Therefore, it is highly recommended that precautions be taken during processing and handling of whiskers to avoid breathing them or releasing them into the work area and clothing. Similar care must also be exercised during cutting, machining, or other manufacturing opera-

tions that could release them into the work environment.

7. References

1. S. M. Pickard, B. Derby, E. A. Feest, *J. Mater. Sci.* **26** (1991) 6207.
2. C. Herring, J. K. Galt, *Phys. Rev.* **85** (1952) 1060.
3. R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Perfection of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958.
4. F. R. N. Nabarro, P. J. Jackson: "Growth of Crystal Whiskers," in R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Properties of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958, pp. 13 – 101.
5. R. M. Fisher, L. S. Darken, K. G. Carroll, *Acta Metall.* **2** (1954) 368.
6. S. M. Arnold, *Tech. Proc. Am. Electroplat. Soc.* **43** (1956) 26.
7. W. C. Ellis, D. F. Gibbons, R. G. Treuting: "Growth of Metal Whiskers from the Solid," in R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Properties of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958, pp. 102 – 120.
8. D. S. Rodbell: "Ferromagnetic Resonance in Iron Whiskers," in R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Properties of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958, pp. 247 – 252.
9. R. W. De Blois, C. P. Bean: "Nucleation of Ferromagnetic Domains in Iron Whiskers," in R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Properties of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958, pp. 247 – 252.
10. S. S. Brenner: "Properties of Whiskers," in R. H. Doremus, B. W. Roberts, D. Turnbull (eds.): *Growth and Properties of Crystals,* Proceedings of the International Conference on Crystal Growth, Cooperstown, NY, John Wiley and Sons, New York 1958, pp. 157 –190.
11. N. Cabrera, Cambridge Conference on Strength of Whiskers and Thin Films, 1958.
12. D. L. McDanels, *Metall. Trans. A* **16 A** (1985) 1105.
13. K. Suganuma, T. Fujita, K. Suzuki, K. Niihara, **9** (1990) 633.
14. G. C. Wei, P. F. Becher, *Am. Ceram. Soc. Bull.* **64** (1985) 298.
15. P. D. Shalek, J. J. Petrovic, J. F. Hurely, F. D. Gac, *Am. Ceram. Soc. Bull.* **65** (1986) 351.
16. H. Tan, T. J. Ahrens, *J. Mater. Res.* **3** (1988) 1010 – 1020.
17. K. Okada, N. Otsuka, *J. Mater. Sci. Lett.* **8** (1989) 1052.
18. K. T. Faber, A. G. Evans, *Acta Metall.* **31** (1983) 565 –576.
19. K. T. Faber, A. G. Evans, *Acta Metall.* **31** (1983) 577 –584.
20. M. G. Jenkins, A. S. Kobayashi, K. W. White, R. C. Bradt, *J. Am. Ceram. Soc.* **70** (1987) 393 – 395.
21. D. K. Shetty, *Ceramic Matrix Composites,* **Current Awareness Bulletin, vol. 118, no. 12,** Metals and Ceramics Information Center, Battele Columbus Laboratories, Columbus, OH, 1982.
22. R. W. Rice, *Ceram. Eng. Sci. Proc.* **6** (1985) 589 – 607.
23. T. N. Tiegs, P. F. Becher, *Am. Ceram. Soc. Bull.* **66** (1987) 339.
24. G. S. Brady, H. R. Clauser, *Materials Handbook,* McGraw-Hill Book Company, London, 1977.
25. W. L. Bragg, *Z Kristallogr Kristallgeom Kristallphy Kristallchem* **74** (1930) 237.
26. W. Noll, H. Kircher, *Naturwissenschaften* **37** (1950) 540.
27. T. F. Bates, L. B. Sand, J. F. Mink, *Science (Washington, D.C.)* **111** (1950) 512.
28. T. F. Bates, F. A. Hilebrand, A. Swineford, *Am. Mineral.* **35** (1950) 463.
29. S. T. Buljan, J. G. Baldoni, M. L. Huckabee, *Am. Ceram. Soc. Bull.* **66** (1987) 347.
30. D. Webster: "Effect of Lithium on the Mechanical Properties and Microstructure of SiC Whisker Reinforced Aluminum Alloys," *Metall. Trans. A.* **13 A** (1982) 1511 – 1519.
31. R. J. Arsenault, R. M. Fisher, *Scr. Metall.* **17** (1983) 67 – 71.
32. W. F. Knippenburg: "Growth Phenomena in Silicon Carbide," *Philips Res. Rep.* **18** (1963) 161 – 274.
33. L. S. Ramsdell, *Am. Mineral.* **32** (1947) 64.
34. P. T. Schaffer, *Acta Crystallogr. Sect. B. Struct. Crystallogr. Cryst. Chem.* **B 25** (1969) 477.

35. C. E. Ryan et al., *J. Cryst. Growth* **1** (1967) 255.
36. M. Saito, S. Nagashima, J. Kato, *J. Mater. Sci. Lett.* **11** (1992) 373.
37. K. M. Merz, in J. R. O'Connor, J. Smittens (eds.): *Silicon Carbide, A High Temperature Semi-Conductor,* Pergamon, London 1960, pp. 73 – 84.
38. R. S. Wagner, W. C. Ellis, *Trans. Metall. Soc. AIME* **233** (1965) 1053.
39. Lj. S. Cerovic, S. K. Milonjic, N. M. Bibic, *J. Mater. Sci. Lett.* **14** (1995) 1052 – 1054.
40. R. V. Krishnarao, M. M. Godkhindi, M. Chakraborty, P. G. Mukunda, *J. Mater. Sci.* **29** (1994) 2741 –2744.
41. J. Lee, I. B. Cutler: "Formation of Silicon Carbide from Rice Hulls," *Am. Ceram. Soc. Bull.* **54** (1975) 195 – 198.
42. N. K. Sharma, W. S. Williams, A. Zangvil, *J. Am. Ceram. Soc.* **67** (1984) 715.
43. M. Patel, A. Kareera, *J. Mater. Sci. Lett.* **8** (1989) 955.
44. R. V. Krishnarao, M. M. Godkhindi, M. Chakraborty, *J. Mater. Sci.* **27** (1992) 1227.
45. R. V. Krishnarao, M. M. Godkhindi, *Ceram. Int.* **18** (1992) 185 – 191.
46. K. J. Bowman, S. K. El-Rahaiby: "Silicon Carbide Whiskers," Newsletter **2** no. 1, 1.
47. W. E. Hollar, Jr., J. J. Kim, *Ceram. Eng. Sci. Proc.* **12** (1991) nos. 7 – 8, 979.
48. G. A. Bootsma, W. F. Knippenberg, G. Verspui, *J. Crystal Growth* **11** (1971) 297.
49. J. J. Shyne, J. V. Milewski, R. G. Shaver, A. L. Dunningham (1968), AFML-TR-67-402.
50. J. V. Milewski, F. D. Gac, J. T. Petrovic, S. R. Skaggs: "Growth of Beta-Silicon Carbide Whiskers by the VLS Process," *J. Mater. Sci.* **20** (1985) 1160 – 1166.
51. G. Urretavizcaya, J. M. Porto Lopez, *J. Mater. Res.* **9** (1994) 2981 – 2986.
52. S. K. Lilov, *J. Mater. Sci. Lett.* **14** (1995) 1546.
53. H. Wang, G. S. Fischman: "In Situ Synthesis of Silicon Carbide Whiskers from Silicon Nitride Powders," *J. Am. Ceram. Soc.* **74** (1991) 1519 – 1522.
54. S. Yamada et al., *J. Mater. Res.* **3** (1988) 538.
55. S. Jagota, A. Parvizi-Majidi: "Crystallization of In-Situ Composites from Multi-Component Alumina-Silica Gels," in: Proceedings of the Annual Meeting of the Materials Research Society, April 17 – 20, San Francisco, CA, 1990.
56. S. R. Nutt: "Defects in Silicon Carbide Whiskers," *J. Am. Ceram. Soc.* **67** (1984) 428 – 431.
57. G. P. LaTorre, R. A. Stokell, R. H. Krabill, D. E. Clark: "Use of FT-IRRS for Characterizating Thermal Stability of SiC Whiskers and Composites," *Ceram. Eng. Sci. Proc.* **7** (1987) 702 – 711.
58. J. J. Petrovic, J. V. Milewski, D. L. Rohr, F. D. Gac: "Tensile Mechanical Properties of SiC Whiskers," *J. Mater. Sci.* **20** (1985) 1167 – 1177.
59. K. R. Karasek, *J. Mater. Sci.* **24** (1989) 1617.
60. K. R. Karasek et al., *J. Am. Ceram. Soc.* **72** (1989) 1907.
61. K. R. Karasek et al., *Ceram. Bull.* **70** (1991) 224.
62. K. W. Lee, S. W. Sheargold: "Particulate Matters in Silicon Carbide Whiskers," *Ceram. Eng. Sci. Proc.* **8** (1987) 702 – 711.
63. M. Matsubara et al., *J. Mater. Sci. Lett.* **6** (1987) 1313.
64. V. N. Gribkov, V. A. Silaev, B. V. Schetanov, E. L. Umantsev, A. S. Isaikin, *Sov. Phys. Crystallogr. (Engl. Transl.)* **16** (1972) 852.
65. T. Hayashi, S. Kawabe, H. Saito, *Yogyo Kyokaishi* **94** (1986) 19.
66. M. J. Wang, H. Wada, *J. Mater. Sci.* **25** (1990) 1690.
67. Z. Chen, *J. Mater. Sci.* **28** (1993) 6021.
68. P. S. Gopalakrishnan, P. S. Lakshminarasimham, *J. Mater. Sci. Lett.* **12** (1993) 1422.
69. P. S. Gopalakrishnan, P. S. Lakshminarasimham, *J. Mater. Sci. Lett.* **14** (1995) 31.
70. J. Homeny, L. Neergaard, K. R. Karasek, J. Donner, S. A. Bradley, *J. Am. Ceram. Soc.* **73** (1990) 102.
71. H. Schneider, E. Eberhard, *J. Am. Ceram. Soc.* **73** (1990) 2073.
72. J. V. Milewski, H. S. Katz: "Whiskers," in: J. V. Milewski, H. S. Katz (eds.): *Handbook of Reinforcements for Plastics,* Van Nostrand Reinhold, New York 1987, pp. 205 – 229.
73. P. D. Bayer, R. E. Cooper, *J. Mater. Sci.* **2** (1967) 347.
74. K. Okada, N. Otsuka, R. Brook, A. J. Moulson, *J. Am. Ceram. Soc.* **72** (1989) 2369.
75. J. R. Moyer, P. R. Rudolf, *J. Am. Ceram. Soc.* **77** (1994) 1087.
76. N. Merk, G. Thomas, *J. Mater. Res.* **6** (1991) 825.
77. I. G. Talmy, D. A. Haught, USA 4 911 902, 1990.
78. T. L. Lowe, N. Merk, G. Thomas, in: "Alloy Phase Stability and Design," *Mater. Res. Soc. Symp. Proc.* **186** (1990).

79. Z. Y. Ma, J. Pan, X. G. Ning, J. H. Li, J. Bi, *J. Mater. Sci. Lett.* **13** (1994) 1731.
80. V. Chellappa, Z. Chiou, B. Z. Jang, *J. Mater. Sci.* **30** (1995) 4263.
81. V. Chellappa, B. Z. Jang, *J. Mater. Sci.* **30** (1995) 4879.
82. S. Radhakrishnan, *Polym. Commun.* **26** (1985) 153.
83. D. S. Blaszkiewicz, D. S. McLachlan, R. E. Newnham, *Polym. Eng. Sci.* **32** (1992) 421.
84. S. Motojima, Y. Itoh, S. Asakura, H. Iwanaga, *J. Mater. Sci.* **30** (1995) 5049.
85. H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, R. E. Smally, *Nature (London)* **318** (1985) 162.
86. W. Kratschmer, L. D. Lamb, K. Fostiropoules, D. R. Huffman, *Nature (London)* **347** (1990) 354.
87. S. Iijima, *Nature (London)* **354** (1991) 56.
88. T. W. Ebbesen, P. M. Ajayan, *Nature (London)* **358** (1992) 220.
89. T. Cabioc'h, J. P. Riviere, J. Delafond, *J. Mater. Sci.* **30** (1995) 4787.
90. X. K. Wang et al., *J. Mater. Res.* **10** (1995) 1977.
91. P. G. Caceres, H. K. Schmid, *J. Am. Ceram. Soc.* **77** (1994) 977 – 983.
92. T. Matsushita, I. Yamai, *Nippon Kagaku Kaishi* **10** (1974) 1893.
93. T. Yamazaki et al., *J. Mater. Sci. Lett.* **14** (1995) 1479.
94. N. Koparanova, Z. Zlatev, D. Genchev, G. Popovich, *J. Mater. Sci.* **29** (1994) 103.
95. I. Matsubara et al., *Jpn. J. Appl. Phys.* **28** (1989) L1121.
96. I. Matsubara et al., *J. Cryst. Growth* **110** (1991) 973.
97. T. Kasuga et al., *J. Mater. Res.* **9** (1994) 1098.
98. R. Bogoroch, S. R. Luck: "Workplace Handling Requirements and Procedures for ACMC Silicon Carbide Whiskers Based on a Sub-chronic Inhalation Study in Rats," in R. A. Bradley, D. E. Clark, D. C. Larsen, J. O. Stiegler (eds.): *Whisker and Fiber Toughened Ceramics,* ASM International, Metals Park, OH, 1988, pp. 81 – 89.

Whisky → Spirits

Whitewares

Karl-Heinz Schüller, Fachhochschule Nürnberg, Fachbereich Werkstofftechnik, Nürnberg, Federal Republic of Germany

1. Introduction 1
2. Definitions and Classification 2
3. Porous Whitewares: Earthenware . 4
3.1. Types of Earthenware 4
3.2. Structure and Properties 4
3.3. Raw Materials 6
3.4. Fabrication of Tableware and Fancy Goods 7
3.5. Fabrication of Wall and Floor Tiles 8
3.6. Glazing of Earthenware 8
3.6.1. Composition and Properties 8
3.6.2. Glaze Stresses 9
4. Vitrified Whitewares: Porcelain and China 9
4.1. Types of China and Porcelain 9
4.2. Physical Properties 9
4.3. Chemical Properties 11
4.4. Raw Materials 11
4.5. Special Types of China and Porcelain 13
4.6. Porcelain Structure: Theoretical Background 14
4.7. Fabrication of China and Porcelain 14
4.8. Standards 16
5. Technical Ceramics 16
5.1. Steatite Whitewares 16
5.2. Cordierite Whitewares 16
6. References 17

1. Introduction

The term whiteware was originally applied to white tableware and artware. It is now used for all ceramic ware that is ivory colored or has a light grey appearance in the fired state. Fine ceramic whitewares are conveniently divided into two classes:

1) Triaxial bodies, consisting primarily of the three components: clays, feldspathic materials, and quartz
2) Nontriaxial bodies made entirely or predominantly of other materials

The whiteware group can be divided into five categories: earthenware; stoneware; china; porcelain; and technical ceramics.

This article covers mainly earthenware, china, porcelains, and technical ceramics (the latter in so far as they are based on clays and other silicates). Stoneware is mostly a colored type of ceramic ware and it is treated elsewhere (→ Construction Ceramics). Technical ceramics comprise silicates and materials based on oxides, carbides, nitrides, and similar compounds (→ Ceramics, Electronic).

2. Definitions and Classification

The ASTM standard C 242-92 gives the following definition of whitewares:

Ceramic Whiteware. A fired ware consisting of a glazed or unglazed ceramic body which is commonly white and of fine texture; designating such product classifications as tile, china, porcelain, semivitreous, and earthenware.

Some definitions taken from ASTM C 242-92 and essential to this article follow:

Body. The structural portion of a ceramic article, or the material or mixture from which it is made.
Bone china.
A translucent china made from a ceramic whiteware body composition containing a minimum of 25 % bone ash.
Ceramic article. An article having a glazed or unglazed body of crystalline or partly crystalline structure, or of glass, which is produced from essentially inorganic, nonmetallic substances and either is formed from a molten mass which solidifies on cooling, or is formed and simultaneously or subsequently matured by the action of heat.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_243

Ceramic process. The production of articles or coatings from essentially inorganic, nonmetallic materials, the article or coating being made permanent and suitable for utilitarian and decorative purposes by the action of heat at temperatures sufficient to cause sintering, solid-state reactions, bonding, or conversion partially or wholly to the glassy state.
Ceramics. A general term applied to the art or technique of producing articles by a ceramic process, or to the articles so produced.
China. A glazed or unglazed vitreous ceramic whiteware made by the china process and used for nontechnical purposes, designating such products as dinnerware, sanitary ware, and artware when they are vitreous.
China sanitary ware.
Glazed, vitrified whiteware fixtures having a sanitary service function.
Decoration.
Inglaze decoration. A ceramic decoration applied on the surface of an unfired glaze and matured with the glaze.
Overglaze decoration. A ceramic or metallic decoration applied and fired on the previously glazed surface of ceramic ware.
Polychrome decoration. A multicolor decoration.
Underglaze decoration. A ceramic decoration applied directly on the surface of ceramic ware and subsequently covered with a transparent glaze.
Dinnerware. Ceramic whiteware made in a given pattern and in a full line of articles comprising a dinner service.
Earthenware. A glazed or unglazed nonvitreous ceramic whiteware.
Electrical porcelain. Vitrified whiteware having an electrical insulating function.
Frit. A product made by quenching and breaking up a glass of a specific composition, used customarily as a component of a glaze, body, or porcelain enamel.
Glaze. A ceramic coating matured to the glassy state on a formed ceramic article, or the material or mixture from which the coating is made.
Impervious. That degree of vitrification evidenced visually by complete resistance to dye penetration.
Jasper ware.
A vitreous, opaque, colored, unglazed ceramic ware having white or contrasting relief decorations and containing a substantial amount of barite.
Majolica.
Formerly an earthenware with an opaque luster glaze and overglaze colored decorations, but currently designating any decorated earthenware having an opaque glaze.
Mullite porcelain.
A vitreous ceramic whiteware for technical applications in which mullite ($3\,Al_2O_3 \cdot 2\,SiO_2$) is the essential crystalline phase.
Mullite whiteware.
Any ceramic whiteware in which mullite ($3\,Al_2O_3 \cdot 2\,SiO_2$) is the essential crystalline phase.
Oven ware.
Ceramic whiteware for culinary oven use.
Porcelain. A glazed or unglazed vitreous ceramic whiteware made by the porcelain process, and used for technical purposes, designating such products as electrical, chemical, mechanical, structural, and thermal wares when they are vitreous.
Porcelain process. The method of producing glazed ware by which a ceramic body and glaze are matured together in the same firing operation.
Pottery.
All fired ceramic wares that contain clay when formed, except technical, structural, and refractory products.
Semiporcelain.
A trade term designating semivitreous dinnerware.
Semivitreous (semivitrified). That degree of vitrification evidenced by a moderate or intermediate water absorption.
Stoneware.
A vitreous or semivitreous ceramic ware of fine texture, made primarily from nonrefractory fireclay.
Tableware.
All utensils and decorative articles used on the table for meal service.
Vitreous (vitrified). That degree of vitrification evidenced by low water absorption.

Certain terms have developed different meanings in various countries. In the United Kingdom and the United States, ceramics comprises all

inorganic nonmetallic materials. In Central Europe the term *Keramik* is restricted to materials which meet the following conditions [21–23].

1) They are inorganic and nonmetallic
2) They are partially or completely crystalline, 30 % crystallinity being the lower limit
3) They are insoluble in water
4) They are heated to high temperatures during production or use; in modern ceramic materials, heating above 900 °C is necessary to produce the desired structure and properties

In this article, ceramics has the meaning of the German term *Keramik*, i.e., it excludes glass and glass ceramics.

Ceramics are classified according to grain size (body texture), vitrification, body color, translucency, and porosity. Bodies with grain size >0.1 mm appear inhomogeneous to the naked eye; the German term for products with this coarse body texture is *Grobkeramik* (= coarse ceramic). This generic term includes various groups of products, e.g., bricks, tiles, and refractories. In the field of tableware, the nearest equivalent in English is common pottery, a term recognized internationally for tariff purposes. Bodies with grain size <0.1 mm appear uniform in texture, even when they are really inhomogeneous on a microscopic scale. The German term for products with this fine body texture is *Feinkeramik* (= fine ceramics). There is no equivalent in English. The term fine ceramics has been introduced in Japan to summarize materials based on oxides, carbides, nitrides, and other materials, used for advanced applications. This differs fundamentally from the German concept of *Feinkeramik*. The clay-based materials in the group *Feinkeramik* are further classified according to color and porosity:

Common pottery: colored, porous
Earthenware:
white or at least bright, porous
Stoneware: colored, vitrified
China or porcelain: white, translucent, vitrified

The distinction between porous and dense structures is made on the basis of water absorption; dense materials have water absorption < 2 wt %.

In the United Kingdom, the terms china and porcelain are synonyms when applied to the classification of tableware. Bone china is defined simply as a special composition with a minimum tricalcium phosphate content within the china and porcelain classification. The key property requirements for tableware in this classification are a very low water absorption (corresponding to a high degree of vitrification), a white body, and translucency in thin sections. Thus, the same properties are recognized in the United Kingdom as in Germany in the definition of *Porzellan*.

The definitions of ASTM C 242-92 are similar in scope to the German definitions, but distinctions between materials are not made on a quantitative basis, and difficulties may arise in finding equivalent English and German terms.

The term porcelain is also used for technical ceramics in the United Kingdom and the United States. For example, in the United Kingdom, it is used for electrical porcelain which is vitrified, but not translucent.

Porcelain, as defined in ASTM C 242-92, comprises certain special types of technical ceramics, e.g., cordierite, forsterite, steatite, titania (titanium dioxide), and zircon porcelain. In some countries, porcelain is generally classified as a luxury article for tariff purposes, regardless of the real type of ware; this can be a serious obstacle to international trade in some types of technical ceramics, and whiteware seems to be the preferable term. These materials can be defined as special types of ceramic whitewares, avoiding their classification as types of porcelain. A European standard for classification of materials used for tableware is in preparation. No generally accepted agreement has been reached so far, partly because of the uncertain classification of "borderline" materials, and partly because of traditional assignments of certain materials to special groups. Porcelains may have a residual porosity up to 0.5 % according to the draft European standard, although any porosity impairs translucency.

Bruno Sanders [24] introduced the following German terms for the characterization of materials:

Gefüge:
all characteristics which can be used for the description of a material, e.g., structure, including microstructure
Struktur:

characteristics which are independent of direction
Textur:
characteristics which depend on direction, e.g., orientation

No systematic use of analogous terms can be observed in British or American English; texture may be used to describe properties that are independent of direction while structure is preferred to summarize properties which depend on direction. To avoid confusion, German terms are used here when they are more clearly defined than the English ones.

3. Porous Whitewares: Earthenware

Porous ceramics are often classified according to their body color. The colored types are sometimes called pottery, whereas the white-bodied types discussed here are classified as earthenware.

Earthenware was developed by Josiah Wedgwood in the 18th century. He introduced china clay to obtain a cream-colored ceramic body.

Nowadays earthenware is used for wall and floor tiles, and for tableware and fancy goods. Its main advantages are a broad range of decorative colors and low production cost. The lower mechanical strength compared with china can be a disadvantage for certain applications. For other uses earthenware has been replaced by stoneware or soft porcelain, avoiding problems associated with the porous structure of earthenware.

3.1. Types of Earthenware

Earthenware can be classified according to body composition. The main types are clay earthenware, lime earthenware, and feldspathic earthenware.

Clay earthenware is the oldest type; it uses clays which are not too high in iron oxide, but fairly rich in fluxing impurities. Flint or sand is generally added.

Lime earthenware contains 5 – 35 % lime, introduced either by the use of natural marl (*Mergel*) or by the addition of carbonates such as limestone, chalk, or marble to the clay body. Dolomite is also sometimes used. Free silica is added as a further ingredient.

Feldspathic earthenware contains 5 – 20 % feldspathic material, introduced as beneficiated soda or potash feldspar, or nepheline syenite, or, in an impure form, as arcose sands; the remainder consists of ca. 50 % ball clays and china clays, and up to 45 % free silica. Feldspathic earthenware is the type preferred in England.

Other types of earthenware depend on the use of special raw materials: talc earthenware with the addition of up to 60 % talc, and mixed earthenware which contains a mixture of lime and feldspar as a flux.

3.2. Structure and Properties

Firing of earthenware is executed at fairly low temperatures; the mixture of raw materials consolidates to a ceramic body, but does not vitrify completely, so that the fired body has some porosity. For most applications the pores have to be sealed by glazing the body. Reliable sealing demands a glaze without any crazing, even over long periods; this requires a glaze well matched in its thermal expansion to the body.

The mineral phases in the final structure of earthenware (excluding residues of the minerals introduced in the crude body) develop partly by thermal transformation of the original minerals and partly by thermal reactions between them (Fig. 1).

The mineral phase composition of the final body, including the pores, controls the properties of earthenware.

Glazes with a melting temperature $<1100\,^{\circ}C$ have a rather high thermal expansion coefficient, exceeding that of mullite or amorphous phases in the ceramic body. To match the thermal expansion coefficients of body and glaze, special measures must be adopted, either by actuating reactions in the ceramic body during firing (which result in phases of higher thermal expansion) or by addition of materials with high thermal expansion coefficients which are preserved during firing. Suitable reactions during firing include the formation of cristobalite or anorthite. Mineral phases introduced via the raw materials and capable of increasing the thermal expansion of

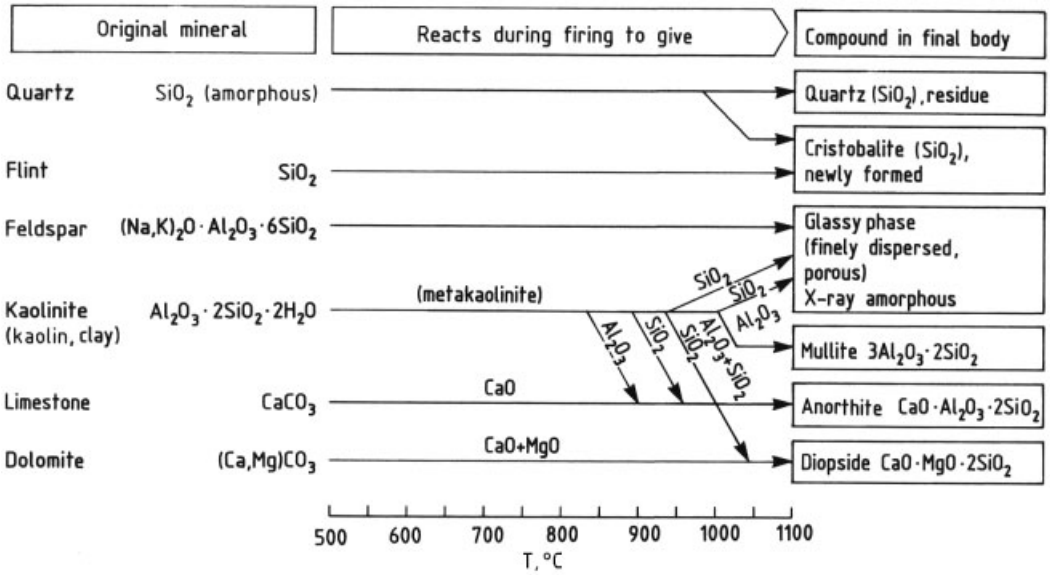

Figure 1. Reactions occurring in bodies of earthenware during firing

the body include quartz and especially cristobalite.

Higher amounts of fluxes in the crude body and increasing firing temperatures result in an enhanced formation of melting phases during firing. The process has to be balanced between sufficient consolidation of the body and controlled dissolution of minerals which have been added to yield high thermal expansion.

Even when the glaze adheres initially under slight compressive stresses, the system can develop crazing over longer periods. This is a consequence of moisture absorption and of reactions of water vapor with the amorphous phases in the ceramic body, causing a slight volume increase known as moisture expansion. If this moisture expansion of the body puts the glaze under tension, it crazes. Complete sealing of the body can prevent it from absorbing moisture. The reactivity of the amorphous phase with water vapor depends on its chemical composition: alkaline earth oxides, iron and manganese oxide reduce the sensitivity to moisture expansion. Lime earthenware and the iron-bearing types of pottery are less susceptible to moisture expansion.

3.3. Raw Materials

Ball Clays. Clays are the essential raw materials for earthenware because they render the body plastic and thus suitable for ceramic fabrication processes; highly plastic clays are preferred. Other important requirements are that they develop a white or bright fired color, i.e., they have low iron oxide content, and do not contain pyrite, which causes dark spots in the fired body. Earthenware clays should not vitrify when fired at ca. 1200 °C. Suitable clays have restricted amounts of minerals which act as fluxes. Kaolinitic clays of the sedimentary type containing some quartz are preferred for earthenware production. Some organic matter may increase their suitability for casting processes.

China clays can replace ball clays up to around half of the total plastic component, contributing to the bright fired color. To reduce costs, crude kaolins can be used instead of washed kaolins.

Quartz and Flint. Quartz is used in earthenware bodies to control the drying sensitivity and to match the thermal expansion of the fired material to that of the glaze. Flint tends to transform rapidly to cristobalite, even when fired at rather low temperatures; transformation starts at

ca. 1050 °C. They are the preferred raw materials in earthenware production in England. Siliceous earth from Neuburg on the Danube is used occasionally as a raw material in earthenware for similar reasons.

Feldspathic Raw Materials. Feldspar is used as a flux in earthenware. All types of sodium feldspar, potassium feldspar, and mixed feldspars are suitable.

For reasons of cost, arcose sands (occasionally incorrectly named pegmatites) can be used as a source of feldspars. They have to fire brightly, without spots.

Sources of Alkaline Earth Oxides. Alkaline earth carbonates (e.g., limestone, chalk, dolomite) are the preferred sources of alkaline earth oxides; they have to be rather fine grained to stimulate complete reaction with other components during firing, anorthite ($CaAl_2Si_2O_8$) and diopside [$CaMg(SiO_3)_2$] being the main reaction products; free calcia and magnesia cause the subsequent formation of cracks because of their reaction with water.

In the United States, natural wollastonite ($CaSiO_3$) is sometimes used as a raw material in earthenware bodies. Attempts to use synthetic alkaline earth silicates in earthenware production have failed because of their high cost.

In countries with suitable sources of talc [$Mg_3Si_4O_{10}(OH)_2$] this mineral can be used as a source of magnesia in earthenware bodies, which leads to the special type of talc earthenware.

Availability of Raw Materials. In temperate climates raw materials for earthenware production are generally available.

3.4. Fabrication of Tableware and Fancy Goods

The preferred type of earthenware for tableware and fancy goods is feldspathic earthenware.

Body Preparation. The feldspar and filler materials (quartz, sand, calcined flint) are ground in wet ball mills or purchased in a suitable grain size. The clays and kaolins are suspended in a blunger. The slurry is mixed thoroughly with the ground nonplastic materials, and subsequently dewatered in a filter press. The filter cakes are then homogenized in a deairing pug mill to obtain the plastic body for forming. Slips needed for casting processes are prepared from the filter cakes by adding an appropriate amount of water and deflocculant.

The more modern process of isostatic pressing needs granulates which are preferably produced by spray-drying batch slurries. Spray-dried granulates are also used to reduce transport costs; they are redispersed in the tableware factory to form casting slips or plastic bodies.

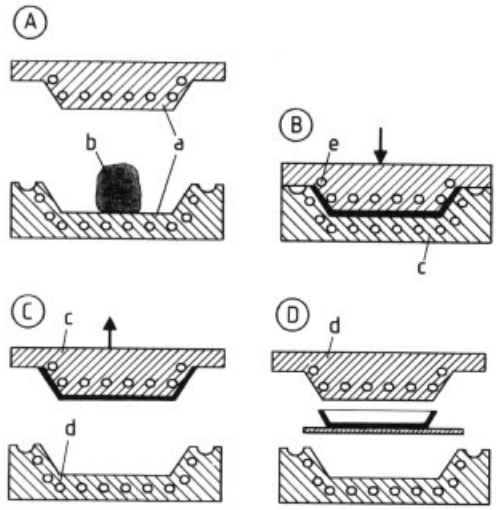

Figure 2. Principles of ram-presssing of earthenware articles
Sequence of steps during ram-pressing: A) Insertion of a body lump into the mold; B) Pressing process for execution of forming, water removal by vacuum in the mold; C) Opening of the mold, vacuum at the upper part, pressure at the lower part of the matrix; D) Removal of the pressed part, pressure at the upper part of the mold
a) Mold made from plaster of Paris; b) Lump of body; c) Mold with applied vacuum; d) Mold with applied pressure; e) Pipe system in the mold

Forming Processes. Casting is performed in the usual manner, with plaster molds.

Plastic forming of earthenware comprises extrusion and jiggering and jollying with roller machines. A further forming process especially for earthenware, is ram-pressing (Fig. 2) which uses a porous mold. Water is sucked off by vacuum; removal of the pressed body from the matrix is assisted by compressed air. Isostatic pressing has

also become an important forming process for earthenware.

Drying and Firing. The drying process is equivalent to that for all ceramic articles.

Firing of earthenware has its own characteristics: biscuit firing is executed at higher temperatures than glost firing. In biscuit firing, the ware is heated to Seger Cone 8 (ca. 1240 °C) where the body develops its final microstructure. Biscuit firing of plates can be executed in stacks because of the lack of molten phase at the surface. Subsequent glazing is easy because of the porous structure of the biscuitware. The subsequent glost firing is executed at Seger Cone 1a – 5 (ca. 1100 – 1180 °C). In order to avoid sticking of the articles to the saggars, they are placed on pins. The contact areas between articles and pins have to be ground smooth after firing.

3.5. Fabrication of Wall and Floor Tiles

Body Preparation. The preferred production method for wall and floor tiles is dry pressing with spray-dried granulate.

Forming Processes. The granulate is compressed in steel dies on special hydraulic presses. The pressing operation has to be adapted to the properties of the granulate to avoid crack formation. This is mainly a matter of experience.

Drying and Firing. Owing to the low moisture content, drying can be rapid. Biscuit firing is preferably carried out in an oxidizing atmosphere, at Seger Cone 6a – 8 (ca. 1210 – 1250 °C), glost firing at 100 – 200 °C lower. The tiles are laid on supports and fired in one layer in fast-firing kilns. Single firing is being introduced to save energy.

3.6. Glazing of Earthenware

3.6.1. Composition and Properties

Glaze is applied, e.g., by dipping (tableware), or by spraying (tiles). Multiple glazing can be applied to give patterned designs.

Glaze properties are controlled by chemical composition, by the types of starting materials used in the glaze, and by details of the application, e.g., thickness of the glaze layer. Important features are:

1) High resistance to crazing, even after long periods
2) Translucency and high gloss, especially for underglaze decoration
3) Hardness and wear resistance
   Resistance to the acids and caustic solutions encountered in household use
4) No health risk from release of toxic ingredients from the fired glaze

The low temperature of glost firing of earthenware necessitates the use of strong fluxes in the glazes which also have to meet other requirements; lead oxide and boron oxide are usual components of earthenware glazes. Alkali oxides are strong fluxes; they involve less health risk, and give a good appearance. However, they increase the thermal expansion of the glaze considerably; it becomes difficult to match it to that of the body, and they can be used only in restricted amounts.

The chemical characteristics of earthenware glazes, as expressed by the Seger formula (→ Ceramics, General Survey, Chap. 5.1.) can range within the following limits:

0.0 – 0.4 K_2O	
0.0 – 0.4 Na_2O	2.0 – 3.0 SiO_2
0.0 – 0.4 CaO	0.2 – 0.3 Al_2O_3
0.3 – 1.0 PbO	0.2 – 0.5 B_2O_3
0.0 – 0.1 MgO, ZnO	

Typically, earthenware glazes use frits as starting materials, offering certain special features:

1) The glost firing temperature is lower than with raw glazes
2) Water-soluble compounds can be fixed in an insoluble frit, so that wet grinding becomes possible
3) Toxic components can be safely fixed in insoluble frits

Colored glazes are obtained by addition of suitable metal oxides, which are dissolved in the melt during firing. The influence on glaze properties has to be considered in the Seger formula. Another possibility is the addition of pigments (→ Ceramic Colorants) which are insoluble in the fused glaze.

Opaque glazes are obtained by addition of special opacifiers such as zircon ($ZrSiO_4$) or by the selection of compositions in which certain components are insoluble, or at most sparingly soluble, in the fused glaze. In other types of glazes, crystallization occurs, which also causes opacity. In crystal glazes the crystals grow to a visible size.

3.6.2. Glaze Stresses

The thermal expansion coefficients of ceramic body and glaze affect the properties critically. Crazing can be avoided when the glaze has a thermal expansion coefficient somewhat lower than that of the body. A suitable difference is a matter of experience. The stresses caused by differences in thermal expansion depend not only on the two thermal expansion coefficients, but also on the relative thickness of the two layers.

Thermal expansion coefficients can be measured with a dilatometer, but it is better to measure the stresses in the combined system, because thermal expansion may be modified by reactions between the two layers during firing. An instrument to provide this information has been designed by STEGER. It measures the bending caused by stresses resulting from the mismatch in thermal expansion.

4. Vitrified Whitewares: Porcelain and China

Porcelain and china are types of whitewares which have a white, translucent, dense body. The type used for tableware and artware is termed china or porcelain. The term porcelain is also used to cover those types used in technical applications. Chinas and porcelains are usually coated with a transparent glaze.

China was developed by improving the color of stoneware to white, by an increasing use of white-firing raw materials. A ceramic body which can be considered real china was first produced in China during the T'ang Dynasty (9th century). The fabrication of extremely thin china reached perfection in the Sung and Ming Dynasties. Huge quantities of china were exported to Europe during late Middle Ages and early modern era.

European attempts to develop an equivalent whiteware succeeded in the early 1700s when JOHANN FRIEDRICH BÖTTGER invented hard porcelain (the Chinese produced a soft-paste type).

4.1. Types of China and Porcelain

Two basic types of china can be distinguished: the classical type is based on kaolin, feldspar, and quartz. It is used for tableware as well as in technical applications. The other type, used only for tableware and artware, contains considerable amounts of bone ash as an essential raw material, flux (feldspar) and clay being the other main components. Special porcelains contain other components (e.g., alumina).

According to the firing temperature two types are known. Hard porcelains are fired at ca. 1400 °C (ca. Seger Cone 14), soft porcelains at ca. 1250 °C (ca. Seger Cone 8). The higher the firing temperature, the harder the glaze. The firing range of whitewares is controlled by the amount of flux in the batch.

4.2. Physical Properties

Translucency is an essential property of china. The less light scattered during transmission through the body, the more translucent the china. This depends on the amount and composition of crystals, and also on the number of micropores in the fired structure. This in turn depends on the batch composition, on the purity of the raw materials, and on the firing cycle, which controls dissolution of crystals and removal of pores. Traces of heavy metal oxides impair translucency considerably [25]. To give a translucent fired body, the raw materials should have as little iron oxide and titanium dioxide as possible; these oxides react to form the dark ilmenite ($FeTiO_3$). Ilmenite formation seems to depend on the chemical bonding of iron oxide in the raw materials; iron oxide in the clay minerals seems to be less critical than that present as separate iron minerals, because it becomes incorporated in the mullite structure and does not form ilmenite [26].

Translucency also depends on the dissolution of quartz during firing, which is influenced by

the geological origin of the quartz. Vein quartz is dissolved more rapidly than quartz sand, but its high price prevents its widespread use.

The melt in the porcelain body picks up quartz during firing; it becomes depleted in alumina, and thus takes up mullite (alumina silicate) to reach equilibrium; this increases translucency. Oversaturation of the melt with silica causes the crystallization of cristobalite, increasing the number of scattering centers in the body, and thus reducing translucency [27].

Alumina, introduced as corundum, is scarcely dissolved in the melt during firing, and thus decreases translucency.

Bone china is fairly high in translucency, despite its rather high crystallinity. The refractive index of the crystalline phases (β-tricalcium phosphate and anorthite) approaches that of the glass phase, and thus scattering of light is low.

Whiteness of chinas is affected mainly by the purity of the raw materials. Besides the iron oxide content, the titanium dioxide content has proved to be critical because titanium dioxide intensifies the coloring effect of iron oxide, although colorless itself [27,28]. The chemical bonding of iron oxide in the raw materials also affects the whiteness of porcelains.

Another important feature for the color of the fired body is the firing atmosphere. In a reducing atmosphere, the iron oxide present in traces in all raw materials is reduced to the divalent state, the color of which is negligible. When the atmosphere at maximum firing temperature is oxidizing, iron remains in the trivalent state which causes an ivory to yellow color of the fired body.

Thermal Properties. Porcelains and chinas are used mainly at room temperature. In principle, their use at elevated temperatures ends when the vitreous phase begins to soften, at ca. 800 °C. An increase in mechanical strength is observed as porcelain is heated to this temperature range because of healing of microcracks.

Thermal shock resistance determines the use of china as ovenware or cookware and in chemical engineering, as well as the suitability of porcelain insulators for outdoor use. Thermal shock resistance is controlled mainly by thermal expansion behavior. The thermal expansion coefficient of porcelains ranges around $(4-6)\times 10^{-6}\,\mathrm{K}^{-1}$, which is rather low compared with other ceramic materials or even metals or plastics. Thus thermal shock resistance is quite good. It is however impaired by the sudden volume change of quartz during its $\alpha-\beta$ transformation at 573 °C, or the cristobalite phase change above 200 °C. Porcelains intended for applications where thermal shock can occur, e.g., in chemical engineering, should have low quartz content.

The increase in volume of quartz and cristobalite during heating makes it easier to find glaze compositions which ensure a lower thermal expansion coefficient of the glaze compared with the body, thus providing surface compressive stresses. This can be used to produce glazed porcelain articles with increased mechanical strength.

Mechanical Strength (Table 1) is important, especially with porcelains used for electrical insulators. The various porcelain types used in this field are standardized in IEC 672. Type C-110 is a traditional triaxial quartz porcelain. Type C-111 is a triaxial porcelain produced by pressing, containing up to 3 % porosity, whereas otherwise porcelains have to be completely vitrified to zero open porosity. Type C-112 is a porcelain which contains cristobalite instead of quartz in its fired structure. Types C-120 and C-130 are alumina porcelains with different amounts of corundum in their structures. The higher the alumina content, the better the mechanical strength, but the fabrication process is more difficult. Type C- 120 is preferred for fabrication of electrical insulators and in other engineering applications.

The figures in Table 1 are valid only for special test bars (see part 2 of IEC 672). Corrections have to be made if the mechanical strength of real components is to be estimated. Strength decreases as the diameter of the component increases.

Mechanical strength is improved considerably by glazing when the thermal expansions of glaze and body are well matched (see above). A glaze somewhat lower in thermal expansion than the body causes compressive stresses at the surface, which improve mechanical strength. When an external tensile stress is applied, the surface compressive stresses in the glaze must to be overcome before a tensile stress of a magnitude

Table 1. Minimum bending strength of porcelains according to IEC 672

Type	C-110	C-111	C-112	C-120	C-130
Bending strength, unglazed, N/mm^2	50	40	80	90	140
Bending strength, glazed, N/mm^2	60		100	110	160

sufficient to cause a catastrophic breakage can occur.

No requirements for tensile or compressive strength are standardized in IEC 672 because of the high scatter of measurements of these properties. It is estimated that the tensile strength of ceramic components is about half the bending strength, and the compressive strength is about ten times the bending strength. This suggests the design of ceramic components which are preferably stressed by compression.

The mechanical strength of various porcelain types is related to their batch compositions. In triaxial porcelains, a minimum strength is observed for the traditional types (ca. 25 % quartz); types of lower and higher quartz content have considerably higher strength [29]. The relationship between strength and quartz content can be explained by structural stresses resulting from the mismatch in thermal expansion of quartz and the vitreous phase in the porcelain body [30–32]. The structural stresses in the vitreous phase are tensile stresses perpendicular to the surface of the quartz grains, but compressive shear stresses along the boundaries. Increasing the amount of quartz results in a decrease of the tensile stress and an increase of the compressive shear stress around each quartz grain. Taking into account the amount of quartz, it is possible to deduce that a minimum strength will occur at a medium quartz content [30–32]. This agrees with practical experience [29].

The influence of alumina on strength can be ascribed to another mechanism. Corundum has a high Young's modulus; therefore, external stresses cause only little strain. The surrounding vitreous phase is relieved of deformation when external stresses are applied. As corundum crystals have a high internal strength, breakage occurs only at rather high external stresses.

The mechanical strength of bone china ranges between types C-110 and C-120 (IEC 672). This can be explained by the lack of crystals which would cause structural stresses and cracks, as well as absorbing external stresses by their high Young's modulus. No values are standardized as bone china is used only for tableware and artware.

Electrical Properties. Porcelain has been widely used for electrical insulators since the beginning of electric power transmission. Similar to other dense ceramic materials, porcelains have an excellent electric strength; the 1-min withstand voltage at 50 Hz under standardized test conditions is >30 kV. Another important feature of porcelains is the high leakage current resistance. Electric spark discharges do not cause damage which would result in a conductive surface layer.

The electrical dissipation factor at 10^6 Hz is ca. 10^{-2}; porcelains are not normally used as insulating materials in high-frequency electrical engineering. More suitable in this field are special types such as the alkaline earth porcelains. (No other uses of these porcelains are known because of their narrow range of firing temperature, which restricts them to articles of small size.)

4.3. Chemical Properties

Porcelains are generally resistant to attack by acids (except hydrofluoric) and organic liquids. Concentrated alkaline solutions attack porcelains; attack increases exponentially with concentration of the corrosive solution and with temperature [33]. A combination of chemical attack and mechanical abrasion may be more severe, and should be considered when using porcelain devices in chemical engineering.

4.4. Raw Materials (→ Ceramics, General Survey, Chap. 1.3., → Ceramics, General Survey, Chap. 2.)

Kaolins (→ Clays). Kaolins (china clays) are residual clays, usually found in primary deposits.

The kaolins used for the production of china and porcelain should have high plasticity, high dry strength, good deflocculation behavior for use in casting slips, and high whiteness in the fired state. Whiteness is controlled by low iron oxide and titanium dioxide, and by chemical bonding (Section 4.2). Dry strength and plasticity are influenced by type and amount of mixed layer minerals, which are present in small amounts in many kaolins, especially in those suitable for use in porcelain fabrication. Deflocculation behavior is closely related to ion-exchange capacity. The broad variety of deflocculants available makes it possible to find a suitable one for most kaolins.

The kaolins used in china and porcelain production are normally processed to remove sand and micas, which can be present in natural kaolins in large quantities. Small residual amounts are less critical than in filler applications so that a classification with a cut at 20–40 μm and thus with less preparation steps gives a satisfactory quality.

Normally, mixtures of various kaolins are used in china and porcelain batches to ensure reasonably constant composition and properties, and to adjust these to production requirements. To improve plasticity, small amounts of ball clays may be added; the amount is limited by the need to keep iron oxide and titanium dioxide low. In technical porcelains, this is less critical; the amount of white-firing ball clays can be increased up to ca. 15 % to improve workability of the crude body; kaolinitic secondary clays are suitable.

Feldspars. The favored type of feldspar in china and porcelain production is potassium feldspar ($KAlSi_3O_8$), which forms a melt of higher viscosity at a given temperature compared with sodium feldspar, and thus causes less deformation during firing. Moreover it allows a broader range of firing temperature [34, 35]. Calcium feldspars are scarcely used in ceramics because of their extremely narrow range of firing temperature.

The high fluxing effect of sodium feldspar is taken advantage of in glazes to assist their flow at temperatures where the body has to remain rigid.

The feldspars used in ceramics are preferably those generated in the pegmatitic phase of magmatic differentiation, because of their high purity and easy processing. Arcose sands, derived from granites by mechanical and chemical weathering, can be used when available in sufficient purity; their fineness reduces the cost of crushing. Moreover they contain feldspars with higher potassium/sodium ratios because sodium feldspars tend to be kaolinized during weathering; these feldspars are even better in preserving stability against deformation during firing. Their quartz content has to be considered in calculating the batch composition.

Nepheline syenite may be used in porcelain bodies instead of feldspar because of its higher alkali content and thus higher fluxing effect; because of its lower availability its use is confined to special types.

Quartz. The preferred silica raw material in china and porcelain production is quartz sand because of its granular structure which reduces crushing costs. Sands of sufficient purity are necessary.

In countries with suitable flint reserves, these are used as a silica source. On preheating, flint becomes brittle which makes crushing easier. The quartz microcrystals are partially transformed to cristobalite which may offer certain advantages because of its influence on thermal expansion.

Vein quartz could be attractive as a raw material because it produces a more translucent china [27]. The high price and the need for expensive crushing processes restricts its use to special products.

Bone ash is an essential ingredient in bone china. The source of supply is mainly ox bone, which has to be freed from organic constituents. The bone is then calcined at ca. 1000 °C. The main crystalline phase in bone ash is hydroxyapatite [$Ca_5(PO_4)_3OH$]. For its use in bone china fabrication, it seems essential to retain a small amount of organic matter in the ash, but too much can be troublesome.

Other Ingredients. Technical porcelains may contain free alumina up to 40 wt % to in-

crease mechanical strength. Alumina is supplied by a specialized chemical industry which offers various types adapted to the needs of the ceramic application.

Zircon ($ZrSiO_4$) has a similar effect on properties to that of alumina [36,37]. Its use is restricted to special cases, especially in countries with large zircon resources.

4.5. Special Types of China and Porcelain

Hard porcelain is the preferred type produced and used in Central Europe. It is designed to be fired at Seger Cone 13 – 16 (1380 – 1460 °C). Its classical type, with ca. 50 % kaolin, 25 % feldspar, and 25 % quartz combines good mechanical strength, low deformation during firing, and good translucency and whiteness.

For applications with thermal shock load, e.g., in cookware or chemical engineering, special types with low quartz content have been developed. Types with increased quartz content up to 40 % exhibit improved mechanical strength; they have tended to be replaced by alumina porcelains, with even higher strength.

Cristobalite porcelain contains cristobalite instead of quartz in the fired structure. This can be achieved by the use of special raw materials, such as chinastone (Japanese, *toseki*) which is available chiefly in East Asia. Chinastone is a rock containing high amounts of micaceous minerals. This contributes to a high plasticity and causes the rapid transformation of the original quartz to cristobalite. Cristobalite porcelains have a mechanical strength exceeding that of quartz porcelains. The special properties can be explained as a consequence of the structural stresses, which are higher than in quartz porcelains, because of the greater mismatch of the thermal expansion of cristobalite crystals and the vitreous phase in the fired body [31]. Cristobalite porcelains are quite common in East Asia because of the availability of chinastone; otherwise alumina porcelains are preferred.

Alumina porcelains are characterized by high mechanical strength. This established their use in the field of technical ceramics, especially for high-voltage insulators. In composition they resemble the hard porcelains, but quartz is replaced by free alumina. The higher the alumina content, the higher the mechanical strength. Superior types contain up to 40 % free alumina. Consequently, fabrication is difficult and wear of fabrication equipment is high.

Soft porcelains contain a higher amount of flux, so they can be vitrified at ca. 1250 °C. Increased flux means a reduction of the clay components, reducing plasticity. To ensure suitable fabrication properties, the kaolin has to be replaced partly by plastic, white-firing ball clays. Another consequence is the need for an oxidizing firing atmosphere to prevent precipitation of carbon in the body during firing (soft porcelains often have an ivory-colored body, but the lower firing temperature allows a larger range of decorative colors).

A special type of soft porcelain is vitreous china. It contains distinct amounts of ball clays, which improves workability. Sanitary porcelains are a special type of vitreous china. They are designed for the fabrication of large pieces, which necessiates the partial use of ball clays and a single-firing technique. Sanitary articles are mostly formed by casting, which has to be considered when selecting the raw materials.

Bone china is the usual type of china for dinnerware and artware in the United Kingdom, but it is also produced in the United States, Scandinavia, and more recently in Germany. Its special features are high translucency, high whiteness, and good mechanical strength. The batch of bone china contains about 40 – 50 parts bone ash, 20 – 30 parts feldspars, and 20 – 30 parts clays. Because of the rather low clay content, workability of the crude body is more difficult than with other types of porcelains.

FritPorcelain. In frit porcelains, alkali-rich frits are used as flux, in combination with chalk and marls. Developed originally in 18th century Europe, they are used nowadays for buttons and costume pearls. Frit porcelains are self-glazing.

Dental Porcelain. The composition of dental porcelains is designed to develop high translucency. This is realized by batches with a feldspar content up to 70 % or even higher, together with some quartz and lime.

4.6. Porcelain Structure: Theoretical Background

The Kaolin – Feldspar – Quartz Phase Diagram. Kaolin, feldspar, and quartz are essential for the properties of triaxial porcelain bodies. The plastic kaolin promotes workability of the crude body. Feldspar acts as a flux, and allows full vitrification of the body at the selected firing temperature. Quartz ensures a reasonable drying behavior; it also controls the development of the structure (Gefüge) and thus the properties of the porcelain. During firing, it is dissolved and thus compensates the drop in viscosity during heating, by inreasing the silica content of the melt.

Porcelain bodies can be obtained only within a certain region of the kaolin – feldspar – quartz phase diagram (Fig. 3) [38]; the boundaries are 35 – 65 % kaolin, 20 – 40 % feldspar, and 0 – 45 % quartz.

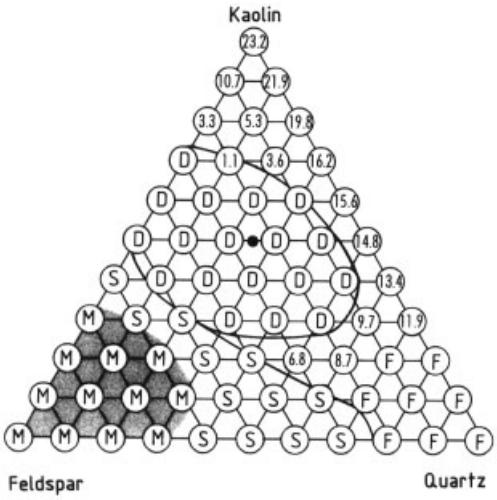

Figure 3. Properties of bodies in the kaolin – feldspar – quartz sytem when fired to Seger Cone 14
Numbers indicate percentage porosity
D = dense structure, porcelain type; S = softened, slightly deformed; M = molten, F = friable, brittle

Reactions During Firing. The reactions in the kaolin – feldspar – quartz system are partly the reactions of the single components during heating, especially at low temperatures, before a molten phase has been formed; at higher temperatures, the components react with each other to form new phases [39, 40]. The reactions are described in detail elsewhere [41].

The reactions occurring in bone china bodies can be deduced from the calcium phosphate – anorthite – silica system (Fig. 4) [42]. Tricalcium phosphate and anorthite are found as crystalline phases in the fired body. Mullite and cristobalite do not occur, although stable according to the phase diagram; they react with impurities, e.g., alkaline oxides, forming a highly viscous melt, which solidifies as a glass during cooling [42].

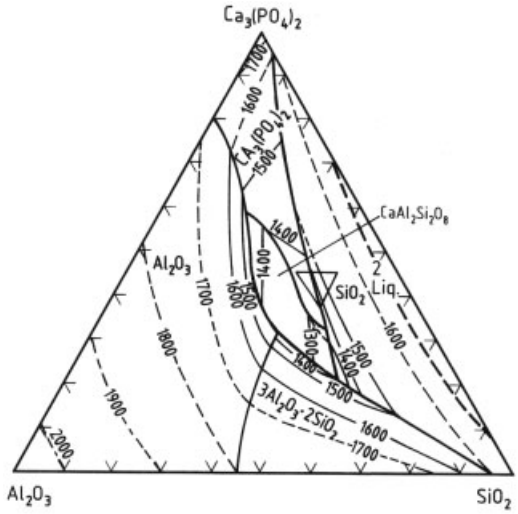

Figure 4. The $Ca_3(PO_4)_2 - Al_2O_3 - SiO_2$ system
The trapezoidal area to the right of center indicates the limits of the composition of bone chinas
Isotherms (— — —) are estimated (in °C)

Firing Atmosphere. It has been mentioned in Section 4.2 that hard porcelains are fired in a reducing atmosphere to develop a white color, whereas soft porcelains are fired under oxidizing conditions to avoid precipitation of carbon with its grey discoloration of the body. However, the reducing atmosphere for hard porcelain is also important to avoid secondary porosity which would result from partial reduction of trivalent iron at high temperature on release of gaseous oxygen.

4.7. Fabrication of China and Porcelain

(→ Ceramics, General Survey, Chap. 4.)

Body Preparation. The preferred preparation of china and porcelain bodies is wet ball milling. Costs can be reduced by using preground raw materials, so that body preparation

is reduced to homogenization in the ball mill. Sieving and magnetic removal of iron contamination from the equipment are standard practice. Further treatment depends on the intended forming process. Casting bodies are immediately deflocculated. Bodies for plastic forming are dewatered in a filter-press and homogenized in a deairing pug mill. Granulates for pressing are preferably obtained by spray-drying.

Dry preparation of china and porcelain bodies has not proved successful, because of the impossibility of removing iron contamination other than by passing the slurry over a magnetic separator.

Scrap from certain forming processes is redispersed, sieved, and added to the crude body. This not only saves raw material costs, but also improves plasticity and workability.

Forming Processes. For the fabrication of china and porcelain all the various forming methods in ceramics can be applied. Casting is used for hollow ware and complicated shapes. Pressure casting is tending to replace normal casting. Plastic forming of china is mainly executed with roller machines. Extrusion is a major process for technical porcelains. Isostatic pressing is increasingly applied for the production of tableware. Joining of component parts is executed in the "leather-hard" state.

Firing. Porcelain is usually fired twice, first at ca. 900 °C (biscuit firing) and then after applying the glaze at ca. 1250 °C (soft porcelain) or ca. 1400 °C (hard porcelain, glost firing). Glazing of the crude articles and single firing is commonly used for technical porcelain articles, especially when they are large. Attempts are being made to use single firing for tableware; dry-pressing and application of the glaze by spraying are promising fabrication steps to reach this goal.

The firing schedule of bone china and vitreous china is different: biscuit firing is executed at ca. 1250 °C and glost firing at ca. 1100 °C.

Tunnel kilns are used as well as shuttle kilns for firing chinas and porcelains; the trend is toward fast-firing kilns with a minimum of kiln furniture.

Hard porcelains are fired in a reducing atmosphere, soft porcelains under oxidizing conditions for reasons already discussed (Section 4.6).

Glazing. Glazes for hard and soft porcelains are normally transparent crude glazes. Frit glazes are used when glost firing is executed at low temperatures (bone china and vitreous china).

The Seger formula of the glazes has to be adapted to the body to adjust thermal expansion; the intended firing temperature is particularly important.

Glazes are applied to the ceramic body by dipping, pouring, or spraying. Automatic spraying equipment has also been introduced for tableware.

Processing After Firing. Further processes may follow firing, e.g., grinding and decoration.

Grinding and polishing is used to smooth areas free of glaze, e.g., saggar contact areas from which the glaze has been removed to avoid sticking. High-quality polishing avoids scratching of plates when they are stacked.

Articles for technical applications may be metallized or fitted with metal mountings, e.g., the shackles which are cemented onto certain types of electrical porcelain insulators.

Decoration (→ Ceramic Colorants). Decoration on china can be achieved by relief, color, and decorative patterns. Reliefs are applied during the forming process. Colors may be added to the body or to the glaze. Decorative patterns are applied to the unfired, biscuit-fired, or glazed body by painting, printing, or transfer techniques. Coating of the decoration with a glaze protects it from mechanical or chemical attack. Decoration applied to the glazed articles is fired under conditions where the glaze softens again; the decoration can sink into the glaze so that it is protected against attack.

The variety of color is greater the lower the firing temperature, but decoration fired at high temperature is more resistant to attack. The variety of colors capable of being fired at higher temperatures has been extended considerably by the introduction of the fast-firing technique and by the development of special pigments.

An essential requirement for decoration is resistance to attack by acids present in foods and beverages. They must not release toxic components such as lead or cadmium. A test with acetic acid is specified in consumer-protection standards.

Decorations can also consist of metallic patterns or ribbons of gold or platinum along the rim of plates or cups.

4.8. Standards

Requirements for porcelains used as electrical insulators are standardized in IEC 672. The standard comprises test methods and mechanical, thermal, electrical, and chemical minimum requirements for various types of materials. Porcelains form group C-100 of this standard.

European standards for articles in contact with food are presently in preparation. For tableware, standards on nomenclature and test methods for lead and cadmium release, water absorption, thermal shock resistance, translucency, crazing resistance, mechanical strength, and dishwashing resistance are being developed.

5. Technical Ceramics

From the wide range of technical ceramics classified as whitewares, two types are briefly considered here: steatite whitewares and cordierite whitewares.

Lithium aluminum silicates are also of technological interest because of their extremely low or even negative thermal expansion coefficients ensuring excellent thermal shock resistance. However, they have an extremely narrow range of firing temperatures when produced by ceramic fabrication procedures; they are preferably produced by crystallization of glasses, and so are better considered as glass ceramics (→ Glass Ceramics).

5.1. Steatite Whitewares (see also → Insulation, Electric, Chap. 3.1.2.2.)

Steatite whitewares are included in IEC 672 in group C-200 "materials based on magnesium silicates". Two types are important; type C-210, steatite, and type C-221, low-loss steatite. They differ in the kind of flux and thus in their properties.

Properties. Steatites have a dense structure with no water absorption, and a high mechanical strength approaching that of alumina porcelains.

Among the electrical properties, the low dielectric dissipation factor is remarkable, especially for low-loss steatite (type C-221) which is the preferred type in use.

Because of their thermal stability up to 1000 °C and their good electrical properties at elevated temperatures, steatites can also be used for insulating devices in electrical heating technology.

Among other properties, the high wear resistance is remarkable.

Fabrication. The main raw material is the steatite, the fine-grained variety of talc [$Mg_3Si_4O_{10}(OH)_2$]. Both types C-220 and C-221 are composed of ca. 85 % steatite, 5 – 10 % plastic clay, and 5 – 10 % flux (feldspar in steatite whitewares type C-220, and barium carbonate in low-loss steatites type C-221).

All common ceramic forming methods are used; the preferred method is dry pressing, because the excellent gliding properties of talc crystals aid compaction.

Steatite whitewares are fired at ca. 1300 °C, mainly in oxidizing atmospheres. For some applications, steatite components are coated with metal layers, preferably as a base for soldering them to other components.

Uses. Steatite components are widely used as electrical insulators, including high-temperature and high-frequency applications. Their use for high-voltage insulators of high mechanical strength has tended to disappear with the introduction of the alumina porcelains.

Steatite has a high wear resistance, hence the application for grinding balls and mill linings.

5.2. Cordierite Whitewares (see also → Insulation, Electric)

In IEC 672 two types of cordierite whitewares are standardized: C-410 dense cordierite and C-520 porous cordierite.

Properties. The main property of cordierite is its low thermal expansion, ca. $2\times10^{-6}\,K^{-1}$. This ensures very good thermal shock resistance.

Cordierite ceramics have a narrow range of firing temperatures. To mimimize fabrication problems, cordierite whitewares are preferably produced with a porous structure. This offers certain additional advantages: avoidance of fluxes in the body improves electrical resistance at all temperatures, the porous structure reduces the Young's modulus, and thermal shock resistance is further increased.

For certain applications, a vitrified type of cordierite whiteware is preferred. It can be obtained by the addition of flux to the body.

Fabrication. Natural supplies of cordierite are insufficient to meet the demand for cordierite whitewares. These are therefore produced from clay and steatite mineral in appropriate proportions. Fluxes are added when a dense structure is required. Feldspar has proved useful because it causes self-glazing of the body.

Production of cordierite whitewares follows the usual paths of ceramic fabrication technology. All forming methods can be applied, but extrusion has become the preferred method. Firing is at ca. 1300 °C.

No glaze is available with a thermal expansion coefficient lower than that of the cordierite body; self-glazing is a valuable characteristic.

Uses. Cordierite whitewares have been used for decades as insulators in electrical heating engineering because of the favorable combination of thermal and electrical properties. Cordierite components are used in huge quantities as catalyst supports for emission control; their excellent thermal shock resistance is the decisive property for this application. Dense cordierite whitewares are used for the production of oven-proof table- and kitchenware.

6. References

General References

1. H. Dettmer (ed.): *Ceramic Monographs. Handbook of Ceramics,* **Supplement to Interceram, since 1979,** Verlag Schmid, Freiburg i. Brg.
2. A. Dinsdale: *Pottery Science,* Ellis Horwood, Chichester 1986.
3. T. Haase: *Keramik,* 3rd ed., VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1970.
4. D. Hülsenberg, H.-G. Krüger, T. Röthig, G. Ferrière: *Maschinelle Formgebung von Keramik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1980.
5. G. Jackson: *Introduction to Whitewares,* Elsevier, Amsterdam 1969.
6. W. D. Kingery, H. K. Bowen, D. R. Uhlmann: *Introduction to Ceramics,* 2nd ed., Wiley & Sons, New York 1976.
7. E. Krause, I. Berger, J. Nehlert, J. Wiegmann: "Verfahren, Rohstoffe, Erzeugnisse," *Technologie der Keramik,* 2nd ed., **vol. I,** VEB Verlag für Bauwesen, Berlin 1985.
8. E. Krause, I. Berger, T. Plaul, W. Schulle: "Mechanische Prozesse," *Technologie der Keramik,* **vol. II,** VEB Verlag für Bauwesen, Berlin 1982.
9. E. Krause, I. Berger, O. Kröckel, P. Maier: "Thermische Prozesse," *Technologie der Keramik,* **vol. III,** VEB Verlag für Bauwesen, Berlin 1985.
10. E. Krause, G. Doll, G. Kalich, W. Schulle: "Gesamtprozeß," *Technologie der Keramik,* **vol. IV,** VEB Verlag für Bauwesen, Berlin 1988.
11. W. Lehnhäuser: *Glasuren und ihre Farben,* 3rd ed., Wilhelm Knapp Verlag, Düsseldorf 1973.
12. W. E. Matthes: *Keramische Glasuren,* Rudolf Müller Verlag, Köln 1985.
13. F. H. Norton: *Fine Ceramics, Technology and Application,* McGraw-Hill, New York 1960.
14. A. Palatzky: *Technische Keramik,* VEB Verlag Technik, Berlin 1954.
15. P. Rado: *An Introduction to the Technology of Pottery,* Pergamon Press, Oxford 1969.
16. H. Salmang, H. Scholze: *Keramik,* part 1 and 2, 6th ed., Springer Verlag, Berlin 1982 and 1983.
17. F. Singer, S. S. Singer: *Industrial Ceramics,* Chapman & Hall, London 1963.
18. S. Stefanov, S. Batschwarow: *Ceramic Glazes,* Bauverlag, Wiesbaden 1988.
19. M. S. Tite, M. Bimson, *Archaeometry* **33** (1991) 3 – 27.
20. W. E. Worrall: *Clays and Ceramic Raw Materials,* 2nd ed., Elsevier, London 1966.

Specific References

21. A. Dietzel, *Ber. Dtsch. Keram. Ges.* **41** (1964) 41.
22. H. W. Hennicke, *Ber. Dtsch. Keram. Ges.* **44** (1967) 209 – 211.

23. K. H. Schüller, H. W. Hennicke, *cfi/Ber. DKG* **62** (1985) 259 – 263.
24. B. Sander: *Gefügekunde der Gesteine,* Verlag Julius Springer, Wien 1930, p. 143.
25. F. Krauss, *Ber. Dtsch. Keram. Ges.* **17** (1936) 125 – 131.
26. H. Kromer, K.-H. Schüller, *Ber. Dtsch. Keram. Ges.* **50** (1973) 15 – 19, 39 – 40.
27. A. Dietzel, N. N. Padurow, *Ber. Dtsch. Keram. Ges.* **31** (1954) 7 – 18.
28. R. Rieke, W. Faust, *Ber. Dtsch. Keram. Ges.* **10** (1929) 567 – 576.
29. T. Wiedmann, *Sprechsaal Keram. Glas Email* **92** (1959) 2 – 5, 29 – 30, 52 – 55.
30. K.-H. Schüller, *Ber. Dtsch. Keram. Ges.* **39** (1962) 286 – 293.
31. K.-H. Schüller, *Ber. Dtsch. Keram. Ges.* **44** (1967) 212 – 223, 284 – 293, 387 – 391.
32. K.-H. Schüller, K. Stärk, *Ber. Dtsch. Keram. Ges.* **44** (1967) 458 – 462.
33. H. J. Ullrich, *Hermsdorfer Tech. Mitt.* 1979, 53, 1705 –1708.
34. H. Rasch, *Ber. Dtsch. Keram. Ges.* **54** (1977) no. 4, GW 1 – 4;no. 5, GW 5 – 8; no. 7, GW 9 – 12.
35. K.-H. Schüller, H. Jäger, *Ber. Dtsch. Keram. Ges.* **56** (1979) 29 – 35.
36. V. Frith, R. O. Heckroodt, K.-H. Schüller, *cfi/Ber. DKG* **64** (1987) 211 – 215.
37. V. Frith, R. O. Heckroodt, K.-H. Schüller, *cfi/Ber. DKG* **64** (1987) 379 – 383.
38. F. Zapp, *Ber. Dtsch. Keram. Ges.* **33** (1956) 358 –362.
39. K.-H. Schüller, *Ber. Dtsch. Keram. Ges.* **38** (1961) 150 – 157, 208 – 211, 241 – 246; **40** (1963) 320 – 326; **41** (1964) 527 – 531; **42** (1965) 299 – 310.
40. K.-H. Schüller, *Trans. Br. Ceram. Soc.* **63** (1964) 103 – 117.
41. K.-H. Schüller: *Porcelain, Ceramic Monographs,* part 2.1, Verlag Schmid, Freiburg i. Brg. 1979.
42. P. D. S. St. Pierre, *J. Am. Ceram. Soc.* **37** (1954) 243 – 258; **38** (1955) 217 – 222; **39** (1956) 147 – 150.

Wind Energy

Armin Keuper, Wilhelmshaven, Federal Republic of Germany

1. Introduction 1
2. Power Content of Wind 2
2.1. Physical Power 2
2.2. Efficiency 2
3. The Wind in Wind Energy 2
3.1. Wind Maps 2
3.2. Roughness 3
3.3. Frequency Distribution of Wind . . 3
3.4. Wind Direction 4
4. Technology 4
4.1. Design of a WEC System 4
4.2. Designs of Modern Turbines 5
4.2.1. Horizontal-Axis Systems 5
4.2.2. Vertical-Axis Systems 7
4.3. Power and Speed Control in WEC Systems 7
4.3.1. Stall Control 7
4.3.2. Pitch Control 7
4.4. Electrical Systems 8
4.4.1. Generators 8
4.4.2. Operating Modes 8
5. Economic Aspects 8
5.1. Factors in the Investment Decision . 8
5.2. Operating Expenses 9
5.3. Factors Involved in Economic Analysis 9
5.4. Present Cost of Wind Energy 10
6. References 10

1. Introduction

Wind energy has become a recognized and proven form of regenerative energy generation, offering an environmentally friendly way to replace conventional, polluting energy sources in part (coal, petroleum, natural gas, nuclear energy). Depending on site or available subsidies, wind energy can now be put to economic use in many areas (especially the coast and coastal regions, as well as highlands). In contrast to conventional energy sources, wind energy inherently minimizes external cost arising from pollution damage to public health and disposal of toxic waste. The economic appraisal also includes cost for dismantling of the wind power plant after service life.

As wind energy utilization becomes more widespread, potential operators and other interested parties are confronted with general questions about the principles of wind energy technology. Detailed explanations of the overall topic of wind energy can be found in relevant literature. A comprehensive description of wind energy utilization is given in [7–9]. The discussion here provides an introduction plus brief explanations of some key concepts.

2. Power Content of Wind

2.1. Physical Power

Utilization of wind energy requires a knowledge of power contained in the wind:

$$P_{\mathrm{wind}} = \frac{1}{2} r\, \mathrm{v}^3 A \ (\mathrm{watts}) \tag{1}$$

where ρ is the density of the air, v is wind speed, and A is the cross-section of the area through which the wind flows. Whereas two of the quantities appearing in this formula, the air density and the wind speed are subject to changes in metereological conditions; the cross-sectional area can be arbitrarily chosen. The air density ρ is approximately 1.22 kg/m^3 at sea level, decreasing by ca. 10 % for every 1000 m of altitude above sea level. This naturally has an effect on the utilization of wind energy in the highlands and high mountains.

The effect of air density variation is slight in comparison to that of wind speed, since Equation (1) contains the cube of the latter. When the instantaneous wind speed increases by a factor of two, the power contained in the wind increases by a factor of eight. An increase in wind speed from 5 to 6 m/s results in a gain of ca. 70 % in the power content of wind.

For wind turbines, the reference cross-sectional area A is determined by the area swept

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_259

by the rotor. Doubling rotor blade length doubles the radius of the rotor circle and thus increases available wind power by a factor of four. Accordingly, the length of the rotor blades is a way of controlling the power that can be extracted from wind. As blade length increases, however, static and dynamic loads increase, and this must be considered in blade design, as well.

2.2. Efficiency

As with any energy source, one key criterion is how much of the primary energy can be converted to electrical energy. The rotor blade cross-sections of modern wind energy conversion (WEC) systems are shaped similarly to aircraft airfoils. In the usual operating range, these WEC airfoils generate aerodynamic lift forces that have a component in the rotational direction of the rotor. In theory, up to 60 % of the total power can be extracted from wind [1]. This value is referred to as the power coefficient c_p. The overall efficiency $c_\mathrm{p}\,\eta$ (η being the mechanical and electrical efficiency) is markedly lower because all components of a system contribute to power losses. Currently available $c_\mathrm{p}\,\eta$ values are ca. 40 – 45 %.

Wind energy conversion systems utilizing aerodynamic drag force are far less efficient. A panel set with its maximum area perpendicular to the wind direction will be subject to the aerodynamic drag force originating from the wind flow hitting the panel. The maximum power coefficient c_p for such a device is ca. 20 %. Because of the efficiency η of the machinery attached, even this 20 % is not actually achievable [7].

3. The Wind in Wind Energy

Earth's atmosphere is continuously exposed to the radiation of the Sun. About 2.5 % of solar energy is transformed into air currents (wind). The total power of all wind is around 4.3×10^{12} kW, although only a small fraction of this amount can be utilized [7]. Wind energy is a renewable energy source.

Wind is produced by local or large-area compensating movements of the air between regions receiving different amounts of solar warming or between regions of high and low atmospheric pressure. These compensating flows take place mainly near the ground. Air movements over land are much more attenuated than those over water because of the greater friction in the former. For this reason, wind over the sea has significantly higher speeds and weakens more slowly than wind over the land. In wind distribution maps, the coast and coastal regions normally have higher annual average wind speeds than inland regions.

3.1. Wind Maps

Because the energy content of wind is strongly dependent on its speed, an "isovent" map (a map with lines of constant wind speed) is often used to assess the potential of economical wind energy conversion.

Wind speed is determined in a variety of ways, for example, as 10-min average values or as annual average wind speed: 10-min and 1-h averages are calculated from measurements of wind speed at 1-s intervals. The 10-min averaging interval has become a standard and is used in most situations.

The yearly average wind speed is determined from the 10-min or 1-h averages over an entire year. A multiyear average is found by averaging over more than one year. The deviation from multiyear average is less than 0.2 m/s in ca. 50 % of years [10].

Although an isovent map gives a first indication of whether a region is suitable for wind energy conversion, it cannot be used in site selection because local conditions can differ greatly from plotted wind speeds. The yearly average wind speed depends on many factors; the most important are examined in the following sections.

3.2. Roughness

Wind describes the motion of air over the Earth's surface. Moving air loses speed through friction mainly in the boundary layer just above the Earth's surface. The degree of attenuation of wind speed at the surface strongly depends on the surface roughness. This roughness can be characterized by the roughness length z_0,

which is a normalization factor for the vertical wind speed profile in the boundary layer. The roughness length is much smaller than the actual dimensions of the roughness elements. The boundary-layer profile of wind speed v (Fig. 1) may be calculated with the formula:

$$\frac{\mathrm{v}_x}{\mathrm{v}_{\mathrm{hub}}} = \frac{\ln\,(h_x/z_0)}{\ln\,(h_{\mathrm{hub}}/z_0)} \tag{2}$$

where v_x is wind speed at height h_x and v_{hub} is the speed at the hub height h_{hub}. The smaller the roughness length z_0, the less is the wind slowed down in the boundary layer with decreasing height (Fig. 1). At very great heights above ground (ca. 1 – 3 km) wind speed is independent of the roughness of the terrain.

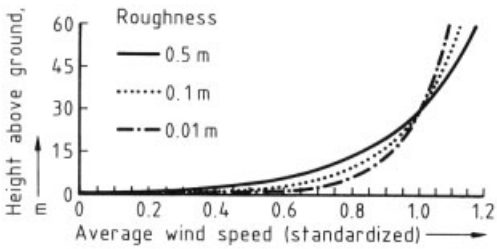

Figure 1. Boundary-layer profiles for various roughness lengths z_0.
Average wind speed is normalized to the value at hub level of 30 m (v_x/v_{hub}). Offset of boundary layer: 0 cm (according to [7], modified; source DEWI)

For shorter roughness lengths z_0 (i.e., over ice, water, or short-mowed grass), wind experiences less friction against Earth's surface. As the surface roughness increases (e.g., over unmowed meadow, stones, scrub, trees, buildings, woods, or built-up areas), wind speed declines near the surface. Typical roughness lengths (in meters) are given below:

Water or ice	0.0001
Mowed grass	0.01
Long grass or rocky surface	0.05
Suburban construction	0.6
Woods, city	1 – 5

In general, the wind grows stronger with increasing height above Earth's surface. On a large scale this effect creates opportunities for wind energy utilization in the highlands and high mountains. On a local scale, it makes tall towers for installation of wind energy conversion systems economically attractive. In addition, the friction of the wind against the surface depends on roughness: as roughness increases, the wind becomes weaker; making wind energy utilization economically doubtful in densely populated urban areas.

3.3. Frequency Distribution of Wind

One type of basic information needed to assess a potential site is annual average wind speed, but this alone tells nothing about fluctuations in wind speed and the frequency of light or hard winds. These variations, however, govern the amount of wind power available for conversion.

Theoretically, a yearly average wind speed of 6 m/s could mean that the wind blows at 6 m/s all of the time or that it has a speed of 0 m/s half of the time and 12 m/s the rest of the time, giving an average of 6 m/s. In the second case, the wind is four times more powerful because the formula for wind power contains the cube of the speed.

Information about the distribution of wind speed is obtained by measurement. Measured wind speeds are normally classified in classes 1 m/s wide and plotted as a frequency distribution (Fig. 2).

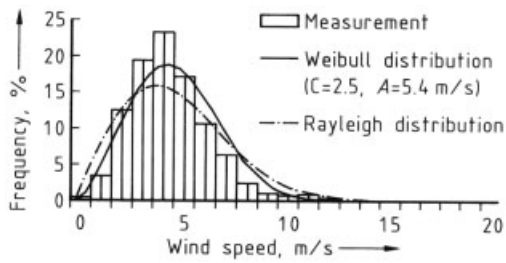

Figure 2. Typical frequency distribution of wind speed (according to [7], modified; source DEWI)

The experimentally determined distribution of wind speed is necessary because it is significant for determination of the expected energy yield. The distribution must be put in a suitable form before it can be used in energy calculations. Speed measurements can be modelled mathematically by the Weibull distribution:

$$f\,(\mathrm{v}) = \frac{C}{A}\left(\frac{\mathrm{v}}{A}\right)^{(C-1)} \mathrm{e}^{-\left(\frac{\mathrm{v}}{A}\right)^{c}} \tag{3}$$

where v is the wind speed, C is a form parameter ($1 < C < 3$), and A is a scale factor. The Rayleigh

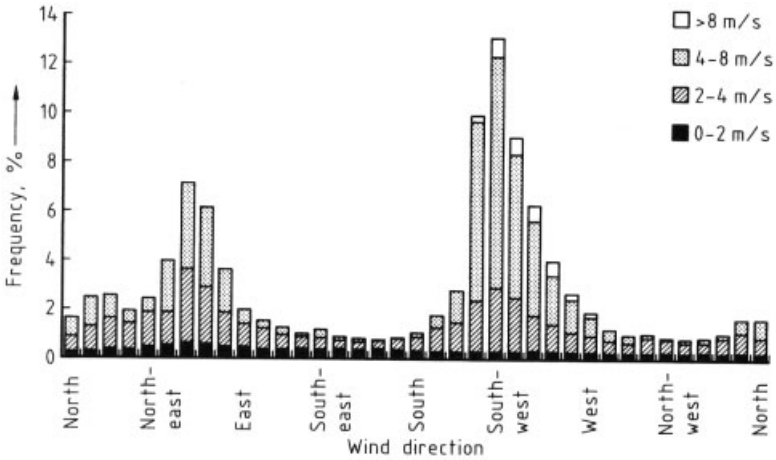

Figure 3. Typical distribution of wind direction (according to [7], modified; source DEWI)

distribution, a special case of the Weibull distribution, may also be used when only the annual average wind speed is known for a site.

The wind speed distribution is also influenced by other phenomena such as diurnal variation of wind speed, calm periods, gusts, and turbulence. More information on these items is given in [10].

3.4. Wind Direction [11, 12]

Another property of wind to be considered is its variation in direction. Many institutions have conducted measurements over periods of years to establish a qualitative description of the changes in wind direction. The measurements show "preferred" wind directions, from which the wind blows more often and more strongly than from others. In Central Europe, the prevailing wind direction is southwest to west (Fig. 3); local deviations from this pattern are possible.

Selection of a good site for a wind power installation must take prevailing wind directions into account. Local conditions may cause the prevailing wind in a certain area to differ from the large-scale distribution of wind direction. Whether this difference is due to local obstructions, which could be avoided through choice of a better site, or whether it is a general topographic feature of the region must be ascertained.

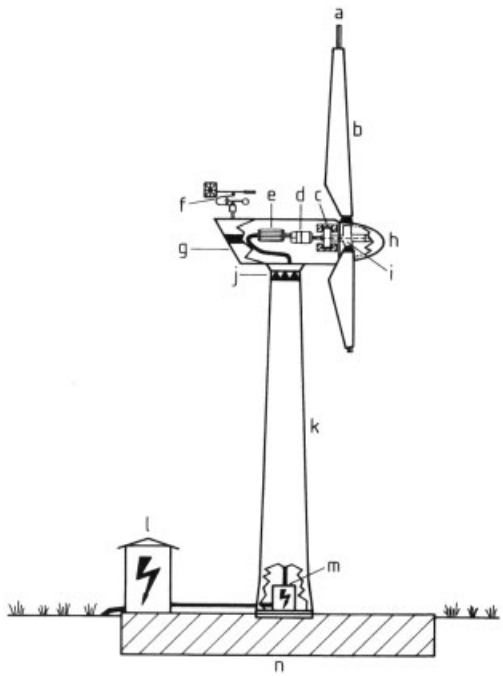

Figure 4. Main components of a wind energy conversion system (source DEWI).
a) Blade-tip brake; b) Rotor blade; c) Disk brake; d) Gear box; e) Generator; f) Wind measurement; g) Nacelle; h) Hub; i) Blade pitch; j) Yaw control; k) Tower; l) Grid connection (transformer); m) Electronic turbine control; n) Foundation

4. Technology

4.1. Design of a WEC System

Figure 4 shows a horizontal-axis wind turbine. The schematic drawing includes only the essential components, some of which are not to scale for reasons of clarity. Not every WEC system has all the components shown; actual systems may differ from the one illustrated both in the components and in their exact position in the plant.

Some vital components of a wind energy conversion plant are explained in what follows: the basic design concepts of modern wind turbines, control systems, and generators. For other components such as yaw system, tower, safety systems, and the gear box, see [7].

4.2. Designs of Modern Turbines

Wind energy systems are designed either with a horizontal or vertical axis.

4.2.1. Horizontal-Axis Systems

Most horizontal-axis systems are wind turbines with one, two, or three blades. Several reasons exist for the use of rotors with few blades:

1) High rotor speeds, desirable for electricity generation, cut down gear box and generator costs.
2) Fewer blades lead to lower material costs.
3) Blade attachments for single-bladed and two-bladed rotors can be made in hinged form.

Three-Bladed Rotor. The three-bladed rotor (Fig. 5 A) is the most common type used in modern WEC systems. It represents a good compromise between the urge to minimize the number of blades and at the same time retain a simple dynamic behavior of the rotor. The mass distributions are so favorable that forced yaw motions of the rotor and nacelle are smooth and uniform. The rotor blades are usually attached to a rigid hub, which is inexpensive to produce, and requires little maintenance. All forces arising in the blade are transmitted to the hub, leading to severe loading on the attachment points and other mechanical parts of the machine. Three-bladed rotors run at lower speeds than two-bladed and single-bladed rotors of the same dimension. If they were to turn at the same speed as other rotor types, the blades would have to be narrower, which creates structural problems and is possible to only a limited extent.

Two-Bladed Rotor. The two-bladed rotor (Fig. 5 B) can operate at higher speeds, so gear box and generator costs are reduced. Because of the reduced size of the gear box and the reduced number of blades the nacelle and rotor mass is less than that of the three-bladed type. The two-bladed rotor, however, poses more complex aerodynamical problems than the three-bladed type. The change of the rotor's moment of inertia during one revolution may cause significant dynamic yawing and tilting loads that must be considered in the design of the system. The use of teetered hubs helps to control these problems.

Single-Bladed Rotor. The single-bladed rotor (Fig. 5 C) runs at even higher speeds than two- or three-bladed rotors can attain. The cost of gear box and generator is thus reduced further. The single-bladed rotor therefore has a lighter nacelle and rotor than systems with more blades. In addition to the dynamic forces that may occur in the two-bladed rotor, it has the drawback of unbalanced, rotating aerodynamic loads on the nacelle. The negative effects of these loads can be mitigated through proper choice of nacelle suspension and hub design. The higher speed of the rotor leads to a higher blade-tip speed for a given rotor diameter. High blade-tip speeds are associated with comparatively high noise emission from the turbine which must be taken into account in site selection.

4.2.2. Vertical-Axis Systems

At least for the present, vertical-axis wind turbines are of secondary importance. Their design follows one of two main trends. The *Savonius rotor* (not discussed further here) is a resistance-type rotor; it is used where the system must start at a low wind speed or where little power is required. Like horizontal-axis machines, the Darrieus rotor operates on airfoil lift forces.

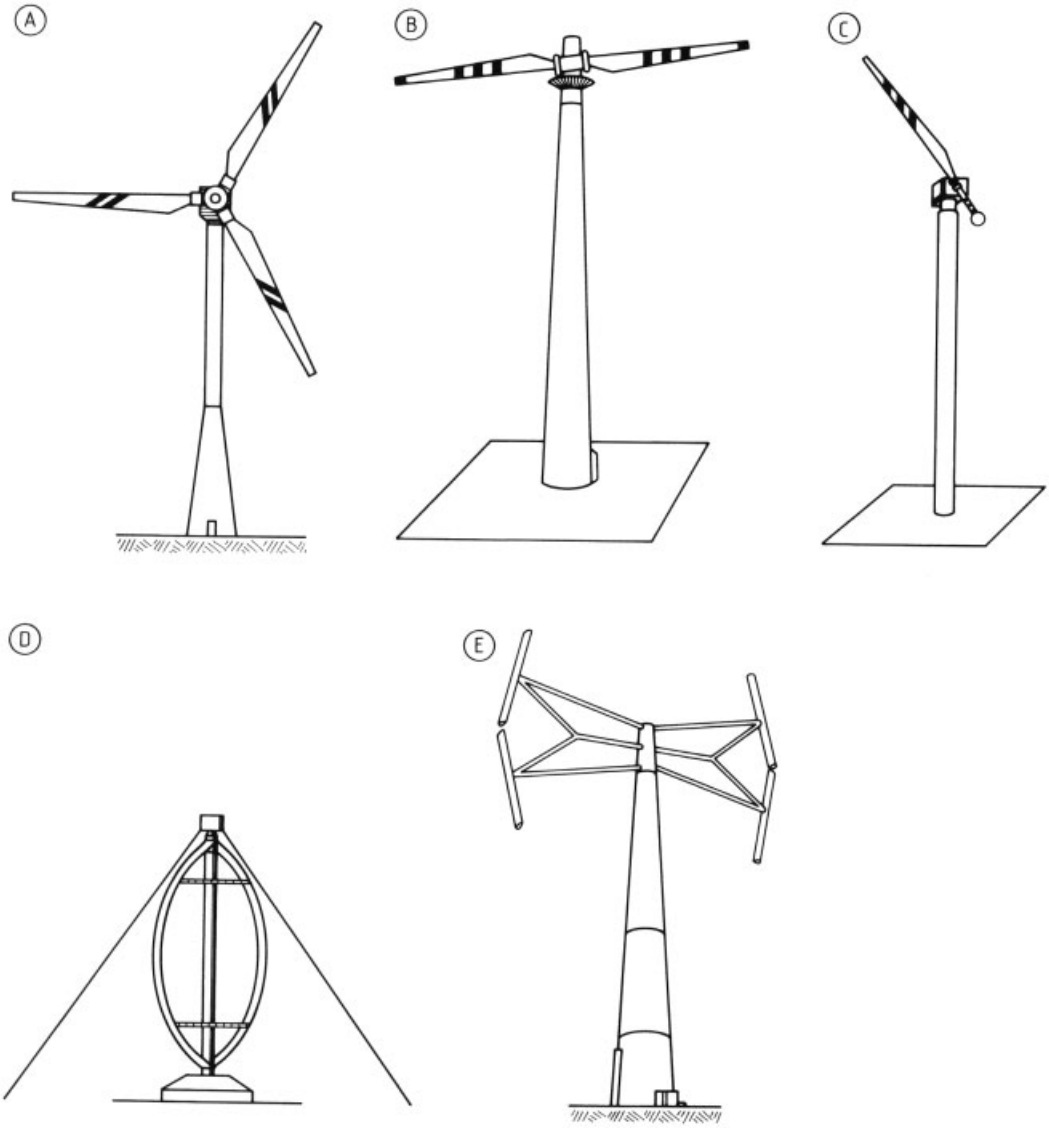

Figure 5. Rotor design (source DEWI). A) Three-bladed rotor; B) Two-bladed rotor; C) Single-bladed rotor; D) Darrieus rotor; E) H-Darrieus rotor

Darrieus Rotor. The Darrieus rotor (Fig. 5 D) has the advantage of always "facing" the wind and thus requiring no yaw control. The generator and the control system are placed at ground level where they can be operated and maintained easily. The catenary-shaped rotor blades are easy to fabricate. Because the rotor is closer to the ground than in horizontal-axis systems, however, the energy yield is lower. These machines are more difficult to control and are generally not self-starting but require starting assistance.

H-Darrieus Rotor. Compared to the "conventional" Darrieus rotor, the H-configuration of the blades in the H-Darrieus rotor (Fig. 5 E) offers the advantage of still easier blade fabrication. The blades are located at a height compa-

rable to that of a horizontal-axis turbine, so the available energy yield is also comparable. Furthermore, power control (see Section 4.3) can be implemented in a similar way as with horizontal-axis rotors.

4.3. Power and Speed Control in WEC Systems

Wind energy conversion systems are designed to reach their rated power at wind speeds that are exceeded relatively frequently. To avoid excessive loading on mechanical or electrical components of a system when wind speeds exceed rated value (i.e., the wind speed at which the system delivers its rated power), precautions must be taken to limit the speed and/or the power output. Two principal types of control are used, both based on aerodynamical principles.

4.3.1. Stall Control

Passive flow separation along the rotor blade increases blade drag and reduces the driving lift forces. When operating below rated wind speed (partial-load operation), the flow pattern around the aerodynamically shaped cross-section of the rotor blade, which is mounted at a fixed pitch position, is optimal. If rotor speed is constant or nearly constant and wind speed increases, the local angle of attack (angle between the airfoil chord and the upstream flow) grows larger and larger until the flow separates from the airfoil surface; the process is similar to what is known as stall in aircraft applications. Flow separation reduces the lift and increases the drag. If the rotational speed must stay constant, the blade can be positioned such that the power output remains approximately constant above the rated wind speed. Often, power characteristics of stall systems show a drop in power above rated wind speed followed by a subsequent rise as wind speed increases.

The advantages of stall control lie in its simplicity and low cost. The disadvantages are high aerodynamic and mechanical loads on the components, the need for a dependable brake, and the comparably poor starting performance of the system after a period of standstill. This poor starting performance refers to the suboptimal fixed blade settings with respect to the flow conditions when the rotor is not yet operating at its nominal speed.

Stall control is found in all sizes of plants.

4.3.2. Pitch Control

Active blade pitch control alters the blade pitch angle continuously to control the lift force and hence the power extracted by the rotor. Pitch control permits easy starting of the rotor without auxiliary energy. Above rated wind speed and up to the "cutout speed" power extraction is controlled by adjusting the blade pitch to keep the output at the rated value.

Blade pitch control requires actuating equipment in the hub and thus higher costs than stall-controlled systems. For a proper design the forces acting on the rotor are generally lower than in stall machines. Pitch-controlled turbines are found in all sizes and are recommended wherever the power output of the WEC system must be deliberately controlled or should follow a prescribed value that is subject to changes.

4.4. Electrical Systems

4.4.1. Generators

Wind energy plants use essentially two types of generators, synchronous and asynchronous, together with their variants.

Synchronous generators offer high efficiency, do not require reactive power, and can be hooked up to the grid either directly or through an inverter (see Section 4.4.2). If connected directly, all power variations in the WEC system are transmitted to the grid without damping, so costly auxiliary devices are needed for synchronization with the grid.

Asynchronous generators are robust and require little maintenance. They permit easy synchronization with the grid but load the grid with reactive power demand. In comparison with synchronous types, they can be connected to the grid more softly. However they have a somewhat lower efficiency.

4.4.2. Operating Modes

Modern wind energy conversion systems are designed to generate electrical energy. They do so in three basic modes.

In the *island mode*, where the WEC system is the only or one of few electricity generating plants in a small grid synchronous or d. c. generators are commonly used. Grid frequency is governed by rotor speed, so turbine speed control is necessary to keep the grid frequency as stable as possible. The grid cannot always accept all the power produced, due to lack of power demand from connected consumers. Therefore, the system must be capable of partial-load operation (operation below the rated output). Conversely, if wind is lacking, the power demand may not be covered by the WEC so energy storage devices (batteries, hot-water reservoirs, etc.) or other energy sources (diesel generator, photovoltaic systems) must be included in the island network.

In *direct grid-coupled mode*, the rotor must turn at a virtually constant speed (with an asynchronous generator) or at constant speed (with a synchronous generator) and must have power output control. In a "strong" grid (where wind energy contributes a small fraction to power supply), all the power offered by the WEC is accepted; in a "weak" grid (large fraction of wind energy), all the power may not be accepted, again requiring controlled partial-load operation.

In the *indirect grid-coupled mode*, an inverter allows variable speed operation of the wind turbine; fluctuations in power output are greatly reduced, and grid coupling becomes very soft. The inverter first rectifies current from a synchronous generator that is variable in frequency and voltage; and then conditions it to the proper grid frequency and voltage. Important drawbacks of this arrangement are harmonics injected into the grid (whole-number multiples of the 50-Hz grid frequency), which must be suppressed with additional filters; losses of efficiency in the inverter; and additional costs due to the inverter itself. The potential advantage of this mode is its higher energy yield compared to directly coupled systems and the reduction of loads on the machine due to wind gusts.

5. Economic Aspects

This section examines the economic aspects of wind energy utilization from the viewpoint of a wind energy investor in Germany. However, the considerations apply in general. The economic analysis is based essentially on [13].

5.1. Factors in the Investment Decision

Capital Expenditure. Aside from the available energy yields, capital expenditure is the most important factor in the economic operation of a WEC system. Plant expenses can be expected to be dominant; additional expenses associated with the installation of the system along with plant expenses represent the amount carried as assets in the corporate balance sheet and written off over the economic life.

Even though a wind power system can be depreciated over 10 years, most manufacturers design their systems for an expected service life of 20 years. The key positions of capital expenditure are listed below:

1) Cost of plant ex works
2) Foundation costs (ca. 8 – 12 % of plant costs: pile foundation, flat footings)
3) Grid connection costs (ca. 8 – 12 % of plant costs: to low-, medium-, or high-voltage grid)
4) Connection costs by electric power utility
5) Site development costs (ca. 3 – 4 % of plant costs: access roads, data communications)
6) Transportation costs
7) Installation and commissioning
8) Planning costs (ca. 1 – 3 % of plant costs: building permit application, consulting, engineering inspection, notarial services, grading, expert opinion on topics such as noise, site, wind analysis, soil, and environmental safety)
9) Additional financing costs (bank consulting, interest for prefinancing of value-added tax, exchange premiums)
10) Land costs

5.2. Operating Expenses

Operating expenses are also an important factor in economic analysis. For wind power systems,

the items listed in the following are regarded as operating expenses (ca. 1 – 2.5 % of plant costs annually):

1) Costs of maintenance contracts
2) Insurance costs
   Liability insurance
   Machinery insurance
   Breakdown or operation interruption insurance
   Transport insurance
   Construction performance insurance
   Installation insurance
   Guarantee insurance
3) Energy costs (costs of energy purchased for operation of WEC system)
4) Tenancy costs

5.3. Factors Involved in Economic Analysis

Based on the discussion in the preceding sections, the starting parameters in economic analysis are as follows:

1) Reimbursement (in Germany 0.0866 €/kW · h, i.e., ca $ 0.1159/kW · h)
2) Grid-coupled operation
3) Company funds (ca. 10 – 15 % of WEC plant costs)
4) Additional costs (27 % of WEC plant costs)
5) Interest rate for costing (ca. 7 %)
6) Financing, interest, and amortization payments under market conditions
7) Power category of WEC system (500 kW)
8) WEC plant costs (1994 list price ca. 450 000 €)

Power characteristics taken from the manufacturers' product descriptions literature have been used to determine energy revenues for a reference site with various annual average wind speeds at a height of 10 m. Given a logarithmic wind profile and a roughness length of 0.05 m, the wind speed at hub height has been calculated by assuming a Rayleigh distribution. The long list of parameters shows that no general statement about individual WEC projects can be made within the scope of this brief analysis. Individual analyses are therefore indispensable.

Figure 6 shows energy generation costs (on the actual value principle) plotted against yearly average wind speed for 0 and 17 % subsidies, respectively.

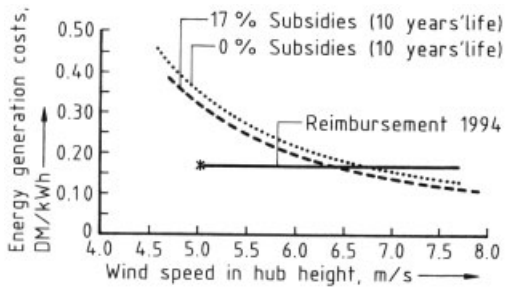

Figure 6. Energy generation cost of wind energy converters (wind speed in hub height); rated power ca. 500 kW, rotor diameter ca. 40 m (according to [13])

Table 1. Present costs of wind energy, recalculated after [14]

Country	Costs per kilowatt-hour, $
Denmark [a]	0.05 – 0.07
Netherlands	0.07
Germany [b]	0.08 – 0.25
United Kingdom	0.09
Italy	0.10
Spain	0.11
Sweden	0.07
United States [c]	0.04 – 0.08

[a] Present information from [15].
[b] Cost history in 1994; range allows for differences in wind and financing conditions.
[c] Data from [16].

To determine energy generation costs per kilowatt-hour, projected interest rates, projected depreciation, and yearly operating costs have been considered. The most important parameters in the calculation are

Costs of the WEC system
Capital service
Operating costs
Energy yield

The last of these is the sole and decisive factor for the WEC system operator's receipts. The value depends on site selection, hub height, rotor diameter, efficiency, and availability. Information on planning, capital expenditure, and energy yield of different WEC systems is given in [9].

5.4. Present Cost of Wind Energy

Table 1 surveys the present costs of wind energy as presented at the 1993 European Community Wind Energy Conference in Travemünde (Germany) [7]; the information has been modified slightly, updated, and completed.

6. References

General References

1. A. Betz: *Wind-Energie und ihre Ausnutzung durch Windmühlen,* Unveränd. Nachdruck d. Originalausg. von 1926: Verlag Ökobuch, Freiburg 1982.
2. E. W. Golding: *The Generation of Electricity by Wind Power,* Spon, London 1976.
3. D. Le Gouriérès: *Energía Eólica,* Masson S. A., Barcelona 1982.
4. E. Hau: *Windkraftanlagen,* Springer Verlag, Berlin 1988.
5. V. D. Hunt: *Windpower,* Van Nostrand Reinhold, New York 1981.
6. E. H. Lysen: *Introduction to Wind Energy,* SWD Steering Committee Wind Energy Developing Countries, Amersfoort.
7. J. P. Molly: *Windenergie. Theorie – Anwendung – Messung,* Müller, Karlsruhe 1990.
8. A. Keuper: *Der Wind in der Windenergie,* Infoblatt des Deutschen Windenergie-Instituts, Wilhelmshaven 1991.
9. A. Keuper: *Kurze Einführung in die Windenergietechnologie,* Infoblatt des Deutschen Windenergie-Instituts, Wilhelmshaven 1993.

Specific References

10. H. A. Panofsky, J. A. Dutton: *Atmospheric Turbulence,* J. Wiley & Sons, New York 1984.
11. H. Dilger, K. Nester, S. Vogt: *Meteorologie. Statistische Auswertung,* KFK Karlsruhe, 1973.
12. I. B. Troen, E. L. Petersen: *European Wind Atlas,* European Commission + Risoe National Laboratory (eds.), Roskilde 1989.
13. K. Rehfeldt, B. Schwenk: " Entwicklung der Energieerzeugungskosten von Windenergieanlagen", *DEWI-Magazin* **8** (1996) 36 – 44.
14. E. J. v. Zuylen, A. J. M. v. Wijk, C. Mitchell: "Comparison of the Financing Arrangements and Tariff Structures for Wind Energy in European Community Countries," in: *European Community Wind Energy Conference,* Travemünde 1993, pp. 108 – 111.
15. P. Nielsen: "Development of Wind Energy in Denmark," in: *Windpower '93,* San Francisco 1993, pp. 401 – 410.
16. K. E. Conover, R. Z. Poore, P. J. F. Hurley, B. Bailey, M. Markus: "Advanced Wind Turbine Optimization for New York State," in: *Windpower '93,* San Francisco 1993, pp. 125 – 133.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Wine

Standard Article

Ralph E. Kunkee[1] and Heinz R. Eschnauer[2]

[1]Dept. of Viticulture and Enology, University of California, Davis, CA 95616, United States

[2]Institut für Önologie, Ober-Ingelheim, Federal Republic of Germany

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.a28_269
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Wine Classifications, Types, and Labels
2.1. Major Categories of Wines
2.2. Table Wines
2.2.1. Still Table Wines
2.2.1.1. Dry, Still Table Wines
2.2.1.2. Nondry, Still Table Wines
2.2.2. Sparkling (Table) Wines
2.3. Dessert/Appetizer Wines
2.3.1. Sweet Dessert/Appetizer Wines
2.3.2. Maderized Wines
2.3.2.1. Oxidized by Film Yeast
2.3.2.2. Oxidized with Age
2.3.2.3. Oxidized by Heat Treatment
2.3.3. Flavored Dessert Appetizer Wines
2.3.3.1. Vermouths
2.3.3.2. Specialty and Proprietary Dessert/Appetizer Wines
2.4. Dealcoholized Wines
2.4.1. Nonalcohol Wines
2.4.2. Wines with Partial Removal of Ethanol
2.5. Wine Labeling
2.5.1. Importance of Region
2.5.2. Controlled Appellations
2.5.3. Imitations
3. Chemical and Physical Composition of Wine
3.1. Extract
3.2. Physical Characteristics
3.2.1. Density
3.2.2. Temperature
3.2.3. Viscosity
3.2.4. Surface Tension and Tears
3.2.5. Refractive Index
3.3. Chemical Composition
3.3.1. Organic Constituents
3.3.1.1. Carbohydrates
3.3.1.2. Aldehydes

3.3.1.3. Organic Acids and Carbonic Acid
3.3.1.4. Alcohols
3.3.1.5. Esters and Other Aroma Substances
3.3.1.6. Polyphenols
3.3.1.7. Nitrogenous Compounds
3.3.1.8. Vitamins
3.3.2. Inorganic Constituents
3.3.2.1. Enological Trace Concentrations
3.3.2.2. Metal Depletion in the Fermentation and Preparation of Wine
3.3.2.3. Enological Sample Bank
3.3.2.4. Essential Trace Elements in Wine
3.3.2.5. Micronutrients for *Vitis vinifera*
3.3.2.6. Enological Trace Element Taxonometry
3.3.2.7. Enological Isotopic Taxonometry
3.3.2.8. Enological Element Analysis
4. Grapes and Grape Varieties
4.1. *Vitis vinifera* Varieties
4.2. Rootstocks
5. Viticulture
5.1. Dormancy and Its Utilization
5.2. Agreement between *Terroir*, Climate and Grape Variety
5.3. Grapevine Diseases
5.4. Other Viticultural Practices
6. Alcoholic Fermentation
6.1. Pre-fermentation Operations
6.2. Yeast Growth and Fermentation
6.3. Post-fermentation Operations
7. Secondary Fermentations
7.1. Malolactic Fermentation
7.2. Sparkling Wine
7.3. Flor Sherry
7.4. Microbial Spoilage
8. Typicité and Style
9. Sensory Evaluation
10. Economic Aspects
11. Wine and Health

[Top of Page]

1. Introduction

Ralph E. Kunkee

Wine has been defined as “a beverage resulting from the fermentation by yeasts of the juice of the grape, with appropriate (and legal) processing and additions” [1]. This stark definition for wine becomes even starker when given in chemical terms: wine is an ethanolic solution of tartaric and malic and other organic acids, glycerol, phenolics, higher alcohols, sugars, and some 500 additional volatile and nonvolatile chemicals in trace amounts (see Chap. Chemical and Physical Composition of Wine). *Enology* (or oenology) is the study or science of wine and winemaking, while *viticulture* is the study or science of (wine) grape growing.

The rigid definitions for wine evade the distinctive and extraordinary position of wine in western civilization — in its history and culture. The literature on wine is enormous. It is estimated that more than 20 000 books and pamphlets, in various languages, dealing with wine or wine grapes have been published [1]. The library at the University of California, for instance, has nearly 5000 books with word “wine” in the title and receives nearly 500 periodicals having “wine” in the title. The vast interest reflects the giant intensity of devotion and adoration which the subject inspires — and also a disdain in some quarters.

Wine is not only an item of produce or a commodity, or perhaps a food or beverage (or even a health nostrum), but also has its mysterious, romantic, artistic and religious facets, which must not be ignored. These ephemeral facets become less ephemeral with the obvious recognition of the physiological effects of the alcohol (ethanol) in the wine — the ensuing euphoria of the partaker. Other alcoholic beverages have not brought forth the cadre of devotees as wine has. For example, the frequencies of the use in prose and poetry of words for various commodities and beverages shows wine to be an order of magnitude most prominent, with objects of art showing some approach to this.

Between the starkly chemical and physical characterizations of wine and the charismatic and enchanting features, the

subject of “quality” is also important: how good is the wine? or how bad? and how are these criteria established, if it all? Thus the topic of wine must to be presented on several layers, some of which are needed as in the usual descriptions of commodities, such as, for instance, cheese, or bread; others which might be applied to works of art; and then some layers, which are obviously unique to wine, embracing its charming and enigmatic attributes. Nevertheless, the emphasis of this keyword is on the physical and chemical components of wine.

[Top of Page]

2. Wine Classifications, Types, and Labels

Ralph E. Kunkee

A good and reasonable beginning in becoming acquainted with wine is an academic one — to become familiar with the various types or categories of wines, and to learn the associated vocabulary. In this approach, it is assumed that wines of each of the categories described are of “good quality”, leaving the difficulties associated with the words “quality” and “good quality” until later (see Chap. Sensory Evaluation). However, “bad quality” wine, whatever that is, does not mean dangerous wine. PASTEUR's counsel that “wine is the most healthy and hygienic of all beverages” [20] has stood the test of time and it still stands. That is, no human pathogen has ever been isolated from wine.

2.1. Major Categories of Wines

Several ways to categorize and subdivide the various classes of wines exist: In the following the system of AMERINE and SINGLETON is used [1], where the divisions are based, in order, on content of alcohol, dissolved gas, sugar, and then on color, and eventually on grape variety, geographical origin and age.

Also derived from fermentation of grape juice are other beverage products. Distilled spirits or brandies are described elsewhere (⟶ Spirits – Introduction, ⟶ Spirits – Wine Spirit. , ⟶ Spirits – Brandy, ⟶ Spirits – Cognac); dealcoholized wines are discussed in Section Dealcoholized Wines.

Table Wines Versus Dessert/Appetizer Wines. Employing the system just suggested [1] the first main division is based on the wine's content of ethanol [64-17-5]. Ethanol is the common end product of most yeast fermentations of sugars in all sorts of beverage and food fermentations. Those wines with lower amounts of ethanol, i.e., with ≤ 14 vol % are classed, in the English language, as “table wines”; because they are generally associated with consumption with meals. Those with a higher ethanol content (usually 16 to 20 vol %) are designated as “dessert/appetizer wines”. In several non-English languages, wines of lesser quality are commonly called something similar to “table wine”, such as: *vin de table*, *Tafelwein*, *vino da tavola*, etc. In the following the English term “table wine” is used, not pejoratively, but properly, even for the best wines. The wines with the higher ethanol content are usually consumed before or after meals, and thus are appropriately named either appetizer or dessert wines. Often these kinds of wines have special names, such as *vin doux natural*; the term “fortified” is sometimes been employed for these wines (see Section Dessert/Appetizer Wines).

The 14 vol % ethanol boundary is not arbitrary; it represents the limit of ethanol production possible by most strains of wine yeast under the usual conditions of grape juice vinification; although under other conditions, such as in saké making [21], these same yeast might produce as much as 20 to 21 vol % ethanol [22]. The 14 vol % limit is also operative in the storage and handling of wine. The higher concentrations of ethanol prevent oxidative spoilage by vinegar bacteria and further fermentation of sweet wines. Table wines ought to be consumed rather soon after they have been opened, as compared to the longer storage allowed for dessert/appetizer wines.

The 14 vol % ethanol boundary between the table and dessert/appetizer wines also has a regulatory and legal function: the collection of taxes. Under most governmental regimes, lower taxes, or none at all, are received from production and/or sales of wines with lower ethanol content. For example, in the United States, the federal taxes on table wines are $ 0.28 per liter and for dessert/appetizer wines $ 0.41 per liter. Furthermore most governmental regulations require that wines of any classification must contain at least 5 vol % ethanol. Some specialty wines may contain between 14 and 16 vol % ethanol [1]. Regulations on “dealcoholized” wines generally require that they contain < 0.5 vol % ethanol.

Of these two major wine categories, table wines represent by far the greater volume of wine production globally.

2.2. Table Wines

Divisions of table wines can be further based on color of the wine, grape variety, content of sugar, etc. Here the convenient system employed above [1] is used, basing the next major subdivission of the table wine category on the content of carbon dioxide [124-38-9]; that is, wines containing no evident amount (*still wines*) or those containing appreciable amounts of CO_2 under pressure (*sparkling wines*).

2.2.1. *Still Table Wines*

Governmental regulations usually allow still wines to contain a measurable quantity of carbon dioxide, but not enough to be obviously noticeable by sight or taste, which in effect means less than can be retained in the wine under standard pressure conditions (1 bar), i.e., less than 3.92 g CO_2 per liter. This category includes the largest volume and the most kinds of wine.

2.2.1.1. *Dry, Still Table Wines*

Dry is an enological term for wines having either no perceptible sweetness and/or wines containing no fermentable sugar (fermentable by wine yeast). Often these characterizations mean the same thing; but in this respect, yeast are more sensitive than humans. Wines which seem to been completely fermented to dryness by taste might in fact contain enough residual sugar to make the wine microbiologically unstable.

The dry table wines can then be easily further subdivided on basis of color: the words white, pink, and red are usually employed. In English, the French word *rosé* is often substituted for pink — where as in French, the word for black (*noir*) is used for red. Although these color descriptors are not exactly accurate (each of these categories has a wide range of hues, shades, and tints), usage has overcome the difficulties. That is, a green or yellow wine, of whatever intensity, is classified as white.

Dry, Still, Varietal Table Wines. This final categorization of table wines is based on grape variety, and indeed may supersede the classification based on color — since in almost all cases the grape variety dictates the wine color category. For example, wines made from Sauvignon blanc grapes are all classified as white, and those from Merlot as red. Although hundreds of grape varieties have been identified for production of wine, some thirty are commonly used. The geographical region often dictates the variety of grape to be used, and globally the regional names usually supersedes categorization by variety.

2.2.1.2. *Nondry, Still Table Wines*

The definitions of dryness are not precise (see Section Dry, Still Table Wines). To be microbiologically stable, the wine should be free of sugars which can be fermented by wine yeast. In fact that means that the residual concentrations of glucose [50-99-7] and fructose [57-48-7] should be zero or nearly so. In effect, wines fermented to dryness and considered “stable”, generally finish up devoid of glucose, but often with one to several grams per liter of residual fructose. These wines may also contain as much as 2 g of pentoses per liter, but pentoses are not fermented by wine yeast. For the consumer, the more operative definition of dryness in wine is the lack of perceptible sweetness. Again, that is not precise and will depend on the subject taster, the kind of wine, the setting, etc. Most experts assign 5 g/L glucose in wine as the sensory recognition threshold for sweetness.

In addition to wines with barely perceptible sweetness (as opposed to those not at all perceptibly sweet), wines with noticeable or obvious sweetness also exist.

Slightly sweet wines are often called *mellow wines*. These are often specialty wines, sometimes with proprietary names, but can be further classified by color and/or by variety. An important class of sweet wines are those produced from “late harvest” grapes, where the initial sugar concentration is so high that the fermentation is not completed and several percent of sugar remain in the wine. Usually these late harvest grapes develop the very high concentrations of sugar by a benign infection of “noble rot”. This rot, actually the fungus *Botrytis cinerea*, is dreadfully devastating except under the special conditions of late harvest. The infection allows for a gentle loss of water in the berries and thus an increase in the content in soluble solids, i.e., sugar. Furthermore this infection brings about excess formation of glycerol [56-81-5]. The subsequent inhibition of the alcoholic fermentation is due to the very high concentrations of sugar (which in itself is surprisingly inhibitory to yeast activity) and to the ethanol produced. The final products in these cases may contain only moderate amounts of ethanol, but because of the large amounts of sugar in the presence of ethanol, are microbiologically stable. The fungus attacks the acids in the berries, so the final acidity of the wine is normal. The fungus also seems to provide special fruit and honey-like flavors to the product. Wines in this classification are the Trockenbeerenauslesen of Germany and the Sauternes from Bordeaux.

2.2.2. ***Sparkling (Table) Wines***

Sparkling wines contain appreciable concentrations of carbon dioxide and are thus stored and packaged under pressure. By definition the term “sparkling wines” applies only to those wines wherein the excess carbon dioxide has come from yeast fermentation. However, wines to which carbon dioxide has been added artificially (carbonated wines), are sometimes included in this category. As far as is known only table wines are marketed with this excess gas, and thus to classify sparkling wines as table wines is probably superfluous.

The naming of sparkling wines, as a class, is highly controlled. Depending on country or region of origin and the procedure used in the production only certain terms may be employed, for example, *Champagne*, *cava*, or *fermented in this bottle*.

Dry and Sweet Sparkling Wines. Sparkling wines are rarely marketed completely dry. Some sweetness is required to balance the general high acidity in these wines, coming inherently from the base wine and also from the resident carbon dioxide (carbonic acid). In addition, the extra sweetness, in form of residual sugar, does not make these wines microbiologically unstable, because of the high acidity, high ethanol, and high carbon dioxide content. Apparently no standard agreement exists as to the labeling names used to indicate the sweetness, probably because the perception of it is so dependent on the other acidic attributes. However, AMERINE and SINGLETON [1] have provided a guide: 15, 25 to 45, 50 and 100 g/L sugar for *brut*, *sec*, *demi-sec*, and *doux*, respectively; even drier wines, with 6 g/L sugar, might be called *extra brut* [16].

The sparkling wines are subdivided on the basis of color as are the still table wines: white, pink, and red, although the categorization is not always as obvious as with the still wines. By far the greatest proportion of the sparkling wines fall in the “white” category; however, a large fraction of these are made starting with red grapes, so it is quite common to find them having a pinkish or bronze hue. Actual pink wines are also made. In addition, a small amount of red sparkling wines is produced. Red sparkling wines often have a manufacturing difficulty in that a rapid loss of carbon dioxide (gushing) occurs after opening of the bottle, probably due to particulate material arising from precipitated pigments.

As with the still wines, grape variety may be used to categorize these wines further. Again, the regional names may dictate the varieties and supplant the variety categorization.

2.3. Dessert/Appetizer Wines

A very conspicuous difference between table and dessert/appetizer wines, is that the latter are microbiologically stable in the presence of air (i.e., oxygen) and fermentable sugar (glucose and fructose). This is due to their higher content of ethanol, which has come about by addition of wine spirits, of “fortification” — the wine spirits being a distillate of wine, and often highly rectified, i.e., “silent spirits” with no congeners. The spirits must be distilled from wine, but depending on the

product, the wine distilled need not be of high quality, and indeed the distillation may represent an operation for salvaging ethanol from substandard wines. The fortification (*mutage*) can be made at the beginning of the fermentation; in the middle, “arresting” the fermentation; or at the end, depending on the kind of dessert/appetizer wine desired. This stabilization property of the fortified wines, as they are sometimes called, allows for the storage of them in open decanters at room temperature, for easy and spontaneous serving, whereas an opened bottle of table wine is susceptible to microbiological attack. Even so, there can be some problems with spoilages of the high ethanol wines, also.

Further categorizing of the high-ethanol wines is different from that of the table wines. Continuing with the preceding scheme [1], the dessert/appetizer wines will be divided into three groups: sweet wines, maderized wines, and flavored or specialty wines. The sweet dessert/appetizer wine category here is only for convenience, since sweet wines can also be found in the maderized wine and flavored wine categories.

2.3.1. *Sweet Dessert/Appetizer Wines*

In fact most of the dessert/appetizer wines are conspicuously sweet. As mentioned before, they are produced by addition of wine spirits, sometimes to grape juice which just begun alcoholic fermentation, or more usually after the ethanol concentration has reached several percent. However, in many cases the wines may be fermented to dryness and are then sweetened in conjunction with being fortified by spirits addition. The sweetening agents can be either conserved grape juice or fortified grape juice.

The nonflavored (see below) *white sweet wines* in this category are sometimes produced by the addition of wine spirits directly to grape juice that has only started to be fermented, or not fermented at all. The *mistelles* California Angelica, French Pineau des Charantes, and Ratafia de Champagne fall into this grouping. However, other white sweet dessert/appetizer may not be fortified until the fermentation has reached several percent of ethanol, examples are the White Ports of Portugal, the French Vins Doux Naturals and other Muscatels. *Pink sweek wines*, generally made by fortification well after the beginning of fermentation, might also include some wines with brownish hues. The most famous of the *red sweet wines* are the Portuguese Ruby, Vintage or Tawny Ports, but wines emulating these are also made elsewhere. These Port wines, especially the Vintage Ports, have the special attribute of improvement with age after being bottled; bottle bouquet otherwise generally found only in specially produced table wines.

2.3.2. *Maderized Wines*

The term maderized denotes those wines which have been intentionally oxidized by various methods to produce a characteristic flavor. For essentially all other kinds of wine, oxidation is an abhorrence — during production or in the flavor. Thus for these special wines, a term more appealing than “oxidation” is needed; the generic term “maderized” is derived from one famous group of this type from the Madeira Islands. This category is further subdivided on the basis of the procedure used in bringing about the maderization: film yeast, aging, or heat. The sweetness of the maderized wines is variable, they are thus considered as a separate category. All maderized wines should be classified as white wines; however, because of their production treatments they are generally dark in color, ranging from amber and tawny to very dark brown.

2.3.2.1. *Oxidized by Film Yeast*

The classic type of these wines are the flor Sherries from the Jerez region of southern Spain. Their complicated production involves the growth of film yeast (usually strains of *Torulaspora delbruekii*) on the top of the wine in partly filled barrels. The ethanol content must be high enough to prevent spoilage by vinegar bacteria, but low enough to allow these flor yeast to oxidize the ethanol and other alcohols. The *vins jeunes* of the Jura region of French are similarly made. Flor Sherries are naturally dry, because the flor yeast will not carry out oxidation in presence of glucose, but often they are subsequently adjusted to make them slightly sweet and with increased ethanol content. The flavors of these wines are complex but show little variation from year to year, due to a complicated fractional blending procedure, called in Jerez the solera system (solera coming from the Spanish word *sole*, or the ground, on which the lowest barrels of the system are placed).

2.3.2.2. *Oxidized with Age*

These wines are typically produced with enough ethanol to prevent any further microbial activity, while the wine is stored, again, in partly filled barrels. The classic type of these wines are the Olorosos of Jerez. These wines are of variable sweetness, sometimes coming from unfermented sugars of the juice, or from juice added after fortification. These wines also enjoy the benefits of the solera system. The special aging flavors in the these wines are produced before they are bottled. Only Port wines, among the dessert/appetizer wines, showing benefit from aging bottling.

2.3.2.3. *Oxidized by Heat Treatment*

These baked wines, the classic Madeira wines, are typically fermented as dry wines and then heated in open vats or tanks at ca. 60 °C for three to four months, and then sweetened and fortified. This style of treatment originates from wines that were shipped in the open on the open sea, where they were oxidized by the agitation and the heat from the sun. Some wines are still heated in the sun in glass containers, the process also being called a solera system — here the word is derived from Portuguese for sun. The system of sun heating of glass containers is also used for production of specialty wine in south western France, in Banyuls. The former California imitations of Madeira wines, the misnamed “baked sherries”, have now lost their popularity.

2.3.3. *Flavored Dessert Appetizer Wines*

This category is divided into two classes: the familiar Vermouths and the remaining specialty products.

2.3.3.1. *Vermouths*

These wines apparently had their origins as vermifuges, infusions of strange and flavorful botanicals in neutral wine spirits. Nowadays, the pretext for healing is gone, but wonderful, and sometimes secret mixtures of all sorts of aromatic herbs and spices are used, making the wines, even the sweet ones, popular as aperitives. The mixtures of herbs is so complex that

none of them is expected to be readily recognized. The infusions are generally made by the maceration of the herbs in fortified grape juice. The vermouths are often divided into two classes, the “French style” and the “Italian style”, having to do with their presumed geographical origin, not affinity. In comparison to the Italian style, the French vermouths tend to be of lighter color, not as heavily spiced, and are less sweet, having only a few percent sugar. Italian vermouths are much darker, sometimes almost reddish-brown, much sweeter (containing up 15 % sugar) and often with a noticeable bitter aftertaste.

2.3.3.2. *Specialty and Proprietary Dessert/Appetizer Wines*

An important class of dessert/appetizer wines are those to which particular herbs or flavors are added to give the product a distinctive, and often a proprietary, character. These wines usually have a definite flavor, which is, however, difficult to identify. These wines are generally not dry and often are quite sweet. A generic example would be sangrias of Spain, while brand names include French Dubonnet and Italian Campari. Some of the major firms producing these kinds of wines have started research programs to identify chemically the essences being used in these delightful concoctions, and also to search for biological activity associated with them.

2.4. Dealcoholized Wines

2.4.1. *Nonalcohol Wines*

The clientele for wines with lowered content of ethanol is continually increasing along with the modern emphasis on safer driving, diets with lowered calories, and increased health awareness. For whatever reasons, one factor has undoubtedly led to the increased appeal for dealcoholized wines: for many consumers, the new wines of this sort taste more and more like untampered wine. However, the removal of ethanol from the wine brings with it problems of microbiological stability. The stabilizing procedure of choice for protection of dealcoholized wine is sterile filtration and sterile bottling; but combinations of heat and addition of sulfur dioxide and sorbic acid, with any accompanying deleterious effects on flavor, are also used.

Three main dealcoholizing techniques are employed, each starting with table wines made in the usual way and of any of the categories given in Section Table Wines. Indeed, some producers commence with varietal wines of very high quality. The earliest dealcoholized procedure on a production scale employed the vacuum pan technology, which was developed for concentration of grape juice. For this, the liquid is put under a vacuum and the temperature raised to 30 – 40 °C, allowing the more volatile materials to distill off. For the grape juice concentrate the distillate is mostly water; but for wine, water and other low boiling fermentation products, and indeed most of the ethanol, is removed. More recently *reverse osmosis* (⟶ Membranes and Membrane Separation Processes – Principle of Reverse Osmosis) or *ultrafiltration techniques* (⟶ Membranes and Membrane Separation Processes – Principle of Ultrafiltration) have been employed, which probably are least violent to the product. The molecules with the lowest molecular mass, i.e., water and ethanol, migrate through the membrane leaving behind the other materials, including the volatile flavor components, in a small volume fraction of the original. The latter is then reconstituted with pure water approximately back to the original volume. A third method uses either *pervaporation* (⟶ Membranes and Membrane Separation Processes – Principle of Pervaporation) or *spinning cone technologies*, which are in effect very gentle adaptations of vacuum pan distillation. With these procedures, the wine is spread out into very thin surfaces, either by use of special membranes, or by a type of centrifugation, which under vacuum allows removal of some of the water and essentially all of the most volatile material, ethanol. Water is then added to bring the volume back to the original.

2.4.2. *Wines with Partial Removal of Ethanol*

Sometimes the final concentration of ethanol in table wines is higher than 14 vol %. Generally this situation arises from harvesting of overripe grapes, sometimes intentionally. Some winemakers are sure that with at least some grape varieties, greater ripeness brings with it more varietal flavor. In any case, excess ethanol can cause problems with imbalance in flavor and can also lead to excessive taxation. The methods described in Section Nonalcohol Wines, especially the ultrafiltration technology, have been well adapted commercially, to solve this problem. In these cases, only part of the ethanol and water are removed, usually to allow the final concentration of ethanol to be somewhat less than 14 vol %.

Ultrafiltration can also have commercial importance in the salvaging of wine with excessive volatile acidity, that is high concentration of *acetic acid* [64-19-7], arising from microbial spoilage (see Section Organic Acids and Carbonic Acid). Acetic acid is inhibitory to wine yeast (*Saccharomyces*) metabolism, and thus excess concentrations of it are not only objectional, in itself, but can be the cause of incomplete (“stuck”) fermentations. Acetic acid is a small molecule and thus can be removed by ultrafiltration. Treatment of the ultrafiltrate with ion exchange resin removes the acetic acid, and the deacidified material is returned to the main body of wine. The volatile acidity is removed, and in cases of incomplete fermentations, reinoculation with wine yeast should bring about dryness.

2.5. Wine Labeling

Understanding of wine labels can be the easiest or the most esoteric segment in learning about wines. Often, the word “wine” itself does not appear. Labeling of old world wines generally gives only minimal explicit information: the region where the grapes were grown, the year of the vintage, the final ethanol concentration (given on a percent by volume basis) and some indication of the quality (ordinary wine, first growth, etc.). To the uninformed, or the novice wine enthusiast, these labels can seem very mysterious and foreboding; but in fact, because of tradition and government regulations, the information on the old world label can imply much more: including the variety, or blend of varieties, of grape; the yield; the ripeness of the grape at harvest and the methods of production. In the new world (in the Americas, Australia/New Zealand and the Republic of South Africa) the labels usually explicitly tell more — but here again, the full disclosures may be lacking. For example, in California, the major grape variety used in the production may be divulged, but not necessarily how much of this variety was used, if it were not 100 %. Again, in California, the word “wine” may not appear if the grape variety and the concentration of ethanol is given; but if “wine” is used, then the concentration of ethanol need not appear. For all wines, the label will indicate the volume of the contents and the name and address of the producer; but again, these latter data may be deficient in explicit information.

2.5.1. *Importance of Region*

It is beyond the scope of this section to discuss the details of the wine labels of the world, or even of any definite regions. Competent guidebooks are available. However, a general rule can be given. For wines of higher quality, the grape growing location is of prime importance. The higher the quality of the grapes, and the wine produced from them, at least in some people's judgement, the narrower the delimitation of the vineyard region is likely to be. For example, a large region might be “France”, given wine of an ordinary type; while a very restricted location might be a vineyard in La Domaine de la Romanée-Conti in Burgundy, producing some of the most expensive wines of the world. In some regions much more than minimal information is given. For German wines, for instances, an indication of the harvest time and the ripeness of the grapes is implicitly given. Furthermore, many of the new world winemakers add more information than is required, often on their so-called back labels. In some regions further restrictive labels are required, supposedly for health or safety reasons; for example, the American requirement for “contains sulfites” labels.

2.5.2. *Controlled Appellations*

The sparseness of explicit information on the labels of high-quality wines gives them almost brand name status. Government regulations, local and international, have been instituted to protect the use of the prestigious, and often historical or folklorical names. Some sorts of controls of appellations of origin were put in place centuries ago, but the modern application begins with France's Appellation d'Origin Contrôlée in 1935. These regulations not only control the vineyard location and area, themselves being delimited not only on the basis of geographic but also political boundaries. They also include viticultural and enological considerations, as mentioned above. Although these restrictions have been initiated primarily for protection of the appellation and the property owners, and for the continued integrity of the product therefrom, they also provide a protection, a guarantee of product quality, for the clientele. These modern regulation systems had their origin in Europe, and they have been extended in some form to virtually all parts of the wine world. There seems to be no quarrel with the intention of these regulations, wherever enacted. However, neither the product nor its production or conception remain static. Thus, regulations such as these need occasional review and adjustment — which is very difficult to do, times after the fact. Many examples can be given where the appellation controls, unchanged for decades, are modern anachronims.

2.5.3. *Imitations*

Government regulations and international treaties can be strict to prevent expropriation of appellations of origin and other carefully guarded labeling characterizations. The most strict example is with the words: *Champagne* and *méthode champenoise*. International agreements limit the use of these labeling words to wines not only actually produced by the traditional method (fermented in this bottle), but also to wines produced by this method only in the Champagne districts of France. Thus, in most other regions of the world, the wine producers must name their product otherwise, e.g., “cava”, “spumante”, or “fermented in this bottle”. Sometimes the word champagne can be used elsewhere, but needs to be modified, such as “American Champagne”.

Several decades ago it was rather usual to find wines produced in the new world to be named after famous wine regions of the old — where the winemakers in developing wine districts, might be hoping to emulate the archetype wines, if not to fool their clientele. The terms Burgundy, Rhine, Sauterne, Rhône, etc., for these wines were more than common. This practice has practically disappeared, hopefully because the modern new world wine producers are endeavoring to set up their own standards and identities.

Some of the most misused appellation names and labels are imitations of dessert/appetizer wines with geographical identifications, such as Port, Madeira, and Sherry. Even in Spain in regions outside of Jerez, wines similar to Sherry wines are produced, for domestic consumption and for export. These cannot be called Sherry, so names reminiscent are used, such as “Solera”. For the wines of the new world such niceties are often ignored. Where enforcement of applicable treaties are not rigorous, the actual words, such as port and sherry are used.

[Top of Page]

3. Chemical and Physical Composition of Wine

Ralph E. Kunkee and Heinz R. Eschnauer

The main components of wine are water and ethanol. Important physical properties of such mixtures are described in Section Physical Characteristics. The components of next abundance are sugars, acids, other alcohols, and phenolics (see Section Organic Constituents) all being present in the magnitude of grams per liter, and as such are rather easy to be determined. These components are important in several aspects of the sensory qualities of the wine; in its sweetness or acidity, in its astringency or bitterness, in its mouthfeel (the viscosity or body), certainly in its color, and in its aging potential. Another group of wine constituents are volatile esters and acetals (see Section Esters and Other Aroma Substances). These occur in much lower concentrations and are much more difficult to assess; however, their evaluation is of preeminent importance, because they contribute to the odor and the taste, i.e., the vinous quality. Some other trace volatile compounds are explicitly identified with the aromas of individual grape varieties. Besides these, a very complex grouping of potential flavor components seem to be the glucosidic phenolics, which have operationally been named “flavor precursors” (see Section Polyphenols). In addition to all of these are the mineral components of wine, at various concentrations, and contributing various characteristics (see Section Inorganic Constituents).

3.1. Extract

The assessment of wine extract, a physical measurement, provides the enologist with indication of the chemical composition of the wine, and provides also information about how it was processed. Simply defined, the quantity of extract is a measure of the soluble solids in a wine after removal of ethanol and other volatile components [3]. Extract is composed mostly of glycerol and organic acids. Residual fermentable sugars, as dissolved solids, may or may not be included in the extract depending on the purposes of the measurements. Extract measurement may be helpful in assessment of sensory mouthfeel, and it can also be used in the detection of fraud. For example, illegal addition of sugar

to a limited volume of grape juice before fermentation would result in an excessive production of ethanol, which can be "corrected" by extension of the volume of the wine by addition of water. Flavor components are also diluted thereby and a sensory examination should detect this fraud; but if not, measurement of the improperly low extract value would.

3.2. Physical Characteristics

Wine shows, of course, all of the physical characteristics (densities, viscosities, surface tensions, and boiling points) associated with water – ethanol mixtures. The determination of the density of wine is about the only physical measurement routinely made, but this measurement is of paramount importance for the winemaker. Except for temperature, other physical measurements are almost never used or even made (see Section Temperature, Viscosity, Surface Tension and Tears); however, grape ripeness can also be determined refractometrically (see Section Refractive Index).

3.2.1. *Density*

Density measurements are performed not only on the wine and wine extract, but are also taken before, during, and after fermentation, to determine the fermentation behavior (see Chap. Alcoholic Fermentation). Densities are nominally given in grams per milliliter but several other scales are actually used by winemakers [3], depending on the country. Since the volume of a liquid is temperature dependent, the temperature of the wine needs to be determined as well when density measurements are made. The *Oechsle scale values* are derived from the multiplication by 1000 of the remainder of 1.000 subtracted from the specific gravity of wine; *Brix scale values* (and the closely related Balling values) give the soluble solids in wine in grams per 100 g (actually calibrated with sucrose as the soluble solid); and the *Baumé scale values* are calculated as the difference between 145 and the quotient of 145 divided by the specific gravity of wine [3]. Each of the scales is calibrated at 20 °C. Densities of wines, employing any of these scales, are measured with hydrometers. A hydrometer, in this case, is a weighted glass tube, constructed to float upright in the liquid to be measured, and in which has been placed the appropriate scale. The scale is calibrated so that the density, in whatever units, can be read directly, followed by a temperature correction.

The density of ethanol, under standard conditions, is 0.79 g/mL. Thus, the densities of table wines are usually lower than that of water, depending on the concentration of ethanol in the wine. The soluble solids in the extract (see Section Extract) will tend to increase the density. Thus a very dry table wine with 14 vol % ethanol and 2 % extract might have a density as low as 0.994 (–6 Oechsle; –1.5 Brix, –8.7 Baumé); while a Port wine with 18 vol % ethanol and 7 or 8 % of residual sugar might have a density of 1.024 (24 Oechsle; 6.1 Brix; 3.4 Baumé) [3], [23]. Density measurements are important in assessments of ripeness of grapes and of the course of grape juice fermentation.

3.2.2. *Temperature*

In addition to the density only the temperature of the wine is measured habitually by winemakers. The temperature is needed first of all to correct the density values. The temperature reading is also important for controlling both the alcoholic and the malolactic fermentations, but also to assure correct storage of the wine in the cellar and in bottles.

3.2.3. *Viscosity*

Viscosity of wine is very seldom determined, although it is often talked about by wine enthusiasts, especially in the context of its being correlated with sensory appraisals of the mouthfeel or the body of a wine. Wine has a higher viscosity than water (typically 1.5 mPa·s as compared to 1.0), due both to its content of ethanol (1.2 mPa·s) and of the extract materials (the viscosity of a 20 % sucrose solution is 1.96 mPa·s). The viscosity of water – ethanol solutions first increases with increasing concentration of ethanol, reaching a maximum of 2.89 mPa·s at 50 vol % ethanol, and then decreases to the value of pure ethanol. Viscosity can be sensed as a form of resistance when a solution is rinsed around the mouth, and thus seems to approximate what wine teasters call body [16]. Glycerol is a major component of wine extract, and thus it could be expected that enological practices which bring about higher concentrations of glycerol (application of selected yeast strains or fermentation temperatures) would increase both the measured and perceived values of viscosity. However, precise examinations of wines with high concentrations of added gylcerol, or of other viscous components, reveal no sensory detection of the increased material, even up to concentrations of 10 g/L glycerol [24], which is rarely found in wines, except possibly those made from *Botrytis* infected grapes (see Section Nondry, Still Table Wines).

3.2.4. *Surface Tension and Tears*

Again, surface tension forces are not readily measured in wine. However, a combination of surface forces of the wine and the forces of the interfacial tension between the wine and a wine glass result in the phenomenon of "tears", where the wine wets the glass, climbs up the glass surface, and eventually rolls back down to the liquid. Occurrence of tears seems to be related to the ethanol content of the wine. However, as pointed out in [16], evaporation of the wine and change in refractive index are also important, but not the viscosity.

3.2.5. *Refractive Index*

The refractive index is measured by winemakers to assess the ripeness of grapes at harvest time. Determination of the refractive index is perhaps even easier than the measurements of density and gives the content of sugar. The measured value is very sensitive to error in the presence of extraneous materials, especially ethanol. Thus, refractometric measurements are not used for following the process of fermentation, in contrast to density readings.

3.3. Chemical Composition

Chemical analysis of wine is an old branch of analytical chemistry. The quality of wine has undoubtedly improved with the application of analytical chemistry [25], although the understanding of the fermentation process itself, which preceded the chemistry [22], [25], surely played an equally important role. However, an understanding of the fermentation process would not have been possible without the continued improvement in the identification of the various wine constituents both qualitatively and quantitatively.

3.3.1. *Organic Constituents*

The major organic chemical components of wine are listed in Table 1. Some of the more interesting, well-studied, or unexpected components are discussed in the following sections. The listing in Table 1 has been gathered from several sources [3], [26-28], none of them very recent. The individual entries represent values from all classifications of wine and should be considered as estimates of orders of magnitudes rather than as precise measurements. For details on wine analyses, see [29-31].

Table 1. Organic constituents in wines

Carbohydrates	
Glucose	see text
Fructose	see text
Sucrose	see text
Pentoses	
Arabinose	0.8 – 2.0 mg/L
Rhamnose	0.5 – 1.0 mg/L
Ribose	0.04 mg/L
Xylose	0.2 – 0.4 mg/L
Galactose	0.13 mg/L
Mannose	0.04 mg/L
Fucose	0.005 mg/L
Pectin	trace
Inositol	0.3 – 0.5 mg/L
Alcohols	
Ethanol	see text
Methanol	0.1 – 0.2 mg/L
Fusel	see text
n-propanol	7 – 56 mg/L
Isobutanol	49 – 213 mg/L
Active primary amyl alcohol	111 – 193 mg/L
Isoamyl alcohol	627–773 mg/L
2-Phenethyl alcohol	5 – 74 mg/L
2,3-Butylene glycol	0.1 – 1.5 mg/L
Acetoin	0 – 0.03 mg/L
Glycerol	3 – 14 mg/L
Sorbitol	trace
Diacetyl	see text
Aldehydes	0.01 – 0.5 mg/L
Organic acids	
Tartaric	3 – 11 mg/L
Malic	1 – 6 mg/L
Citric	0 – 6 mg/L
Succinic	0 – 0.5 mg/L
Lactic	0.5 – 1.5 mg/L
Acetic	0.2 – 5 mg/L
Formic	0.3 – 0.5 mg/L
Glyoxylic	0.0012 mg/L
Mesoxalic	0.001 – 0.003 mg/L
Glyceric	trace
Saccharic	trace

Quinic	trace
Shikimic	trace
Glycolic	0.012 mg/L
α-Hydroxyisocaproic	0.002 mg/L
2-Methyl-2,3-dihydroxybutryic	0.09 mg/L
Oxalic	0.07 mg/L
Malonic	0.01 mg/L
Fumaric	see text
Citramalic	0.02 mg/L
α-Hydroxyglutaric	0.23 mg/L
Aconic	0.02 mg/L
Pyroracemic	0.01 mg/L
α-Ketoglutaric	0.03 mg/L
Galacturonic	0.19 mg/L
Esters	
Isoamyl acetate	2.0 mg/L
Ethyl hexanoate	0.6 mg/L
Hexyl acetate	0.1 mg/L
Ethyl octanoate	1.3 mg/L
2-Phenethyl acetate	0.1 mg/L
Ethyl decanoate	0.6 mg/L
Polyphenols	200 – 1800 GAE/L*
Nitrogenous compounds	
Proteins	0.01 – 0.03 mg/L
Amino acids	0.1 – 2 mg/L
Amides	0.01 – 0.08 mg/L
Ammonia	0 – 0.7 mg/L
Vitamins	see text
Carbonic acid	see text

* GAE=gallic acid equivalent.

3.3.1.1. *Carbohydrates*

Except for pectins, essentially all of the carbohydrates found in wine are sugars.

Sugars. The term *residual sugar* is most commonly applied to the sugars remaining at the end of the alcoholic fermentation. For dry table wines, the residual sugar is expected to be very low; but for all other wines the concentration of sugars may be quite high. Glucose and fructose are the natural sugars found in unfermented grape juice, and high concentrations of them will, of course, be found in wines which have not completed fermentation. Most wine yeast strains are glucophilic and ferment glucose faster than fructose; so that in table wines, near the end of the alcoholic fermentation, the glucose may be essentially gone, but the fructose might still be present in concentrations of 10 – 55 g/L [3]. Another complicating factor is that in many table wines, conserved grape juice (sweet reserve) or grape juice concentrate (see Section Sweet Dessert/Appetizer Wines), both having about equal concentrations of glucose and fructose, may have been added to provide sweetness. Still more complicating, in many locations sucrose is added at the beginning of the fermentation to provide for a higher concentration of ethanol (chaptalization). Some residue of sucrose may be present in the wine, although most of it would probably by utilized by the yeast or hydrolyzed to glucose and fructose by the high acidity of the grape juice. Sucrose is sometimes added in the form of caramel to provide color, and is used as the substrate for the secondary fermentations in sparkling wines. In addition, pentoses are present in the grape juice and remain in the wine (see Table 1), not being utilized by the wine yeast, at least under fermentative conditions. Thus “residual sugar” can have many different meanings.

The measurements of residual sugars and their interpretations depend on the analytic methodology used, which depends on the kind of sugar. With the exception of sucrose, all of the sugars mentioned above are *reducing sugars*, containing a terminal aldehyde, an α-hydroxyketone, or a hemiacetal group. These groups are easily oxidized, which is the basis of the common laboratory determination: reduction (although not stoichiometrically) of copper(II) sulfate [29]. A handy, portable method, based on this principle, are the copper(II) sulfate tablets (Clinitest), which are added to a tiny volume of wine,

giving a bright blue in the absence of reducing sugar.

In a sensory context, “dryness” means absence of perceptible sweetness, which for most individuals means less than 5 g/L of reducing sugar, easily evaluated with the Clinitest tablets.

Although “dryness” in wine is a subjective term, “dry wines with no fermentable sugar” (terminology very important to the cellar worker) normally contain some 2 g/L of pentoses, and when measured by these chemical means are designated as having 2 g/L of reducing sugars.

Values of reducing sugars higher than this indicate the presence of fermentable sugars. If the content of glucose is essentially zero, the small level of fructose contained (see above) probably does not affect microbial stability. Furthermore, some spoilage yeast (especially *Brettanomyces*) and some diacetyl-producing malolactic bacteria are thought to be inactivated only in stored wine which is completely free from glucose. The specific determination of glucose and of the other fermentation sugars is a little more involved. Potentiometric methods are available for selective measurement of glucose and/or fructose; and glucose, sucrose, and fructose can be determined by enzymatic methods. Clinical kits for home diagnoses of diabetes can determine glucose in blood confidentally at 100 mg/dL, and probably at one half at double that concentration. These procedures should be easily adaptable to enological needs.

The *pentoses* are not utilized by the wine yeast during the fermentation, but possibly may become carbon sources for the lactic acid bacteria which carry out the malolactic fermentation (see Section Malolactic Fermentation).

Pectins. Grape juice contains high concentrations of pectins, which consist mostly of polymeric chains of galacturonic acid and its methyl esters. Most of the pectins are precipitated during fermentation by the increasing concentrations of ethanol. Finished wines are often devoid of pectin, having been treated before or after the fermentation with pectinase enzymes.

3.3.1.2. *Aldehydes*

The last step in fermentative production of ethanol is the reduction of *acetaldehyde* [75-07-0] (which itself arises from decarboxylation of pyruvic acid). Acetaldehyde is highly reactive, binding with anthocyanin pigments and with added sulfur dioxide. It has a strong flavor, which is described as pungently nutty or as a raw pumpkin odor. Acetaldehyde should not be evident in properly produced table wines, but can be found in those which have not been protected from oxygen, either before the onset of fermentation or at the end of an overly long and protracted fermentation. Wines bottled without protection may also develop aldehydic flavors later from oxidation of ethanol. Wine enthusiasts are especially sensitive to wines showing any hints of oxidation; and when so, it is usually acetaldehyde that is noticed. However, in the various kinds of oxidized wines, i.e., sherries, madeiras, etc., acetaldehyde is the operative and the most desirable sensory constituent. In “submerged sherry” — which is produced by a suspension of yeast under a slight air overpressure to increase the oxygen concentration [3] rather than by a yeast film (see Section Oxidized by Film Yeast) — the concentration of acetaldehyde may reach as much as 1 g/L.

In addition to acetaldehyde, other low molecular mass aldehydes are present in wine. In maderized wines they probably contribute somewhat to the flavor and odor.

3.3.1.3. *Organic Acids and Carbonic Acid*

The two major organic acids associated with wine are L(+)-*tartaric acid* [87-69-4] and L(–)-*malic acid* [6915-15-7]. They are formed at about equal amounts in the grape berry, but as the berry becomes more and more ripe respiration of the malic acid occurs, which is more pronounced in warm climate. Thus at harvest, these two acids may be at nearly equal concentrations; or in very cold regions, the malic acid content may be substantially greater than that of the tartaric acid; and in very warm regions, the malic acid level may be reduced essentially to zero. Furthermore, in the wine the malic acid is biologically unstable. Here, a secondary fermentation by malolactic bacteria (see Section Malolactic Fermentation) often decarboxylates L(–)-malic acid to L(+)-lactic acid [79-33-4]. Thus the presence of lactic acid in wine primarily results from this fermentation, but can also arise from yeast and lactic acid bacteria metabolism of hexoses. Tartaric acid is essentially biologically stable, except toward attack by *Botrytis* in late harvested grapes (see Section Nondry, Still Table Wines). However, the concentration of tartaric acid is far from constant. Since pKa_1 of tartaric acid is 3.02 at the pH of grape juice and wine (generally between pH 3 and 4), most of the tartaric acid is in the form of its monopotassium salt, potassium bitartrate [868-14-4]. Potassium bitartrate is sparingly soluble in ethanol solutions. Thus most of the salt precipitates during fermentation, and much more of it is lost during cold storage. The latter operation is performed intentionally to stabilize the wine and prevent cloudiness whenever the bottled wine might be chilled.

Citric acid contributes less to the acidity of wine because it is present in low amounts.

The above acids are often equated with various flavors, such as fruity, lemony, milky sour, but in fact the pure acids all have the same characteristic acidic, but bland, flavor, although the intensities vary with the amount of acid (carboxyl) groups per molecule. Enologists refer to these acids collectively as *fixed acids*, in distinction to the *volatile acids*, such as acetic acid, given below. *Total acidity*, then, refers to the sum of the fixed and volatile acids, calculated as the nonionized acids. However, the term *titratable acidity* takes into account that these acids are partly dissociated at wine pH, that is part of the titratable protons of the total acid have been neutralized. For example, the p*Ka*'s of malic acid in wine are 3.45 and 5.33, respectively, and thus at a pH 3.4, a typical pH of wine, 52.5 % of the malic acid is not ionized, 46.9 % is present as hydrogenmalate, and 0.55 % as malate, and only the former two values are part of the titratable acidity.

Acetic acid [64-19-7] is a volatile acid, easily distilled off when wine is boiled. Titration of the distillate is the common procedure to determine the volatile acidity of wine, of which acetic acid is the major portion. Small concentrations of acetic acid are always present in wine, originating from the yeast fermentation, but a high concentration, distinctly above 0.5 g/L, is an indication of serious microbial spoilage. This most commonly arises from oxidation of ethanol by acetic acid bacteria, especially by the vinegar bacterium *Acetobacter aceti*, which therefore must be controlled by SO_2 and anaerobiosis

treatment. Anaerobiosis treatment requires periodic filling (topping) of developing head space (ullage) in wine barrels and addition of inert gas (N_2) to the head space of wine tanks. The indicated spoilage concentration of acetic acid is not easily sensorially detected, but is usually accompanied by another bacterial metabolism product, ethyl acetate [141-78-6], which is more difficult to measure, but which has a lower sensory threshold and is, in fact, the sensory attribute of vinegar. Small amounts of acetic acid are usually formed during the malolactic fermentation (see Section Malolactic Fermentation). Large amounts of acetic acid (5 g/L), which inhibit the wine yeast fermentation, are sometimes found in freshly crushed red grapes that are improperly stored before yeast or SO_2 treatment. These large amounts originate surprisingly from infection by lactic acid bacteria, not acetic acid bacteria.

Carbonic acid is hydrated carbon dioxide, which is of course one of major end products of alcoholic fermentation. Sparkling wines are produced under pressure and contain carbon dioxide under the pressure of several bars (see Section Sparkling (Table) Wines). Still wines are generally allowed to contain 3.92 g/L CO_2, the quality soluble under standard pressure, and young wines usually contain this amount. When the wine ages, the CO_2 dissipates. However, it is sometimes added back to certain wines, especially slightly sweet and fruity wines, at the legal concentrations, to add a fresh quality to the flavor.

3.3.1.4. *Alcohols*

Ethanol. Important physical characteristics of ethanol are discussed in Section Physical Characteristics. High concentrations of ethanol in wine may contribute a slight sweet taste and also give a slight burning sensation, described as hot.

Methanol [67-56-1] is not a direct fermentation product but may arise from demethylation of pectins (see Section Carbohydrates) by treatment of grape juice with pectinase. The concentrations of methanol may be considerably higher after distillation. Most countries have set legal limits of several grams per liter to the maximum concentration of methanol in brandies because of its oral lethal dose in humans of 320 mg/kg [29].

Fusel Alcohols. The higher alcohols, with higher boiling points and higher molecular masses than ethanol, are often called collectively fusel alcohols or fusel oil, because of their rank odor [3]. Generally, 2-phenethyl alcohol is not included with the fusel oil, on account of its low concentration and also because of it pleasant rose-like odor. Although the fusel alcohols represent a major fermentation constituent, their concentrations in wine are generally too low for them to play an important sensory role. However, their concentration increases during distillation which means they can be enriched in brandies. These alcohols are formed by the yeast during fermentation by decarboxylation and reduction of precursor α-oxocarboxylic acids. Isobutanol, for example, is formed from α-ketoisovaleric acid via isobutyraldehyde [3]. The α-oxocarboxylic acids themselves are part of the anabolic pathway of amino acid formation; in this case, α-ketoisovaleric acid is transaminated to valine. The higher alcohols can also be formed from their corresponding amino acids, using the same enzymatic pathways. Isobutanol, for example, is formed from valine (Ehrlich reaction) [3]. The final concentrations of these alcohols depend on several factors, including yeast strain, temperature, redox potential, and amine availability.

3.3.1.5. *Esters and Other Aroma Substances*

Some of the most well-studied *esters* contained in wine are listed in Table 1. These esters are present in wine in meager amounts, but because of their low sensory thresholds and their volatility, certainly contribute to the aroma of wines. However, they represent only a small portion of the total aroma perception. RAPP has indicated that the aromas can be separated into 600 – 800 single components, but also gives a table naming more than 1000 aroma substances [33]. It is beyond the scope of this keyword to go into detail, but aroma compounds in wine belong to the following groups of substances: hydrocarbons, alcohols, aldehydes, ketones, acids, esters, lactones, bases, organic sulfur compound, acetals, ethers, haloorganic compounds, nitriles, phenols, furans, epoxides, and anhydrides. The total content of these compounds amounts to about 1 g/L, or only 1 % of the ethanol concentration, and the fusel alcohols account for half of the aroma components. These components arise almost exclusively during the fermentation by the wine yeast, although the yeast strain itself seems to have little influence on the final sensory profile [3].

Grape Varietal Characterization. Considerable efforts have been made to correlate mixtures of various volatile components with specific and universally recognizable flavors. Just as important in this context is the discovery of aroma components of various varieties of grapes. Here, some progress has been made [3], [33]. The monoterpene alcohols (linalool, geraniol, nerol, citronellol and α-terpinene-4-ol) seem to define the muscat varieties such as Riesling and Gewürztraminer, and several 2-methoypyrazines (3-isobutyl-, 3-isopropyl- and 3-ethyl-) are responsible for the green pepper character of Cabernet Sauvignon and Sauvignon blanc varieties; Chardonnay contains vitispirane, 1,2,6-trimethyl-1,2-dihydronaphthalene, and damascenone.

Corkiness. One of the factors in the odor coming from defective corks, bringing out the “corked” sensory note in spoiled wines has been identified as 2,4,6-trichloroanisole. However, origin of this material, from the trees and their treatment, from the stored bark, from the bleaching and redyeing of the stamped corks or some other sources has not been established.

Bottle Bouquet. Storage of bottled wines over decades can bring about a highly prized maturity resulting in the development of a delectable aroma called bottle bouquet, variously described as the aroma from freshly cooked artichokes to that of freshly washed and air- and sun-dried bed linen. The conditions for attaining this special aging are not perfectly specified, and the compounds being developed are also not certain. However, again, some progress has been made: at least dimethyl sulfide and diethyl succinate have been implicated.

3.3.1.6. *Polyphenols*

Wine polyphenols consist of benzene rings usually with two vicinal hydroxyl groups, and include phenolic acids, cinnamates, flavonoids and tannins [3]. The kinds and amounts of polyphenols contained in a wine initially depend on

grape variety, especially on grape color. The following approximate distribution of polyphenols in wines is given [34]: hydroxycinnamates are in the range of 150 – 200 GAE/L (gallic acid equivalents per liter), while the anthocyanins and the polymeric tannins were found only in red wines at 200 – 400 and 600 – 900 GAE/L, respectively. The level of catechins and other monmeric flavonoids is low (25 GAE/L) in white wines but 150 – 200 GAE/L in reds.

The polyphenols and their reaction products form the basis of essentially all of the color, astringency, and much of the bitterness of wines. As a class, polyphenols are surprisingly reactive, if slowly so. They are the primary substrates for wine oxidation. The oxidation reactions are complicated, involving peroxy radical and hydrogen peroxide intermediates, resulting in among other things phenol dimers. Furthermore, the anthocyanins initially responsible for the red wine color are readily incorporated into phenolic polymers. Full understanding of these reactions, which can occur during storage after the wines have been bottled, will allow some prediction of changes in color and in astrigency of wines, and perhaps in the essence of bottle bouquet (see Section Esters and Other Aroma Substances).

Copigmentation. Although it is well-known that anthocyanins are responsible for the red color of wine, the reactions and equilibria involved are not completely understood. More pigment is held in solution than would normally be found based on solubility alone. The anthocyanins seem to form “stacks” in conjunction with other phenolics, and are considerably brighter in red color in solution than in the free form [35].

Flavor Precursors. It was mentioned above that perhaps more than 1000 volatile compounds are involved as flavor components of wine. New research suggests that there may be many thousands more. The polyphenols are generally glycosylated by α-bridges to two molecules of glucose, which renders them nonvolatile. New technology involves the hydrolysis of the glucoside bonds and the identification of the remaining volatile component [36]. It is supposed that these hydrolyses occur naturally and selectively during storage of wine, perhaps influenced by the pH. Rendered volatile, the components may act as odor and flavor substances. This could be specific for a grape variety, and also might involve enzymes of certain wine yeast strains and occur during fermentation.

3.3.1.7. *Nitrogenous Compounds*

The amounts and kinds of nitrogenous compounds in grape juice have the ultimate influence on the fermentations, especially by the wine yeast, but also by the wine bacteria. However, during the fermentations, most of the nitrogenous components are removed from the juice, and also from the wine when the microorganisms have settled and are separated [3]. The uptake of these components by the microorganisms has been studied intensively, especially the progression of uptake of the various amino acids, and the subsequent role this succession has on the rate and completeness of the yeast fermentation [3]. In the wine, problems with turbidity from proteins can arise after the fermentation, because the resulting ethanol brings about precipitation of the formerly soluble proteins. Although the precipitated fraction is small, it is sufficient to produce cosmetic problems unless stability operations are made. Increases in the level of small molecular mass nitrogenous materials, especially amino acids, does occur in wine with the practice of *sur lies*, that is, the storage of the wine on the yeast lees after alcoholic fermentation. Originally this practice was made just for this purpose, to increase the nutrients in the wine, and thus to stimulate the secondary bacterial fermentations, in cold regions of northern France, where the malolactic fermentation was difficult to induce because of the naturally low pH of these wines [3]. It is now practiced in other wine regions, especially in the new world, apparently to emulate the original product.

3.3.1.8. *Vitamins*

Similar to the nitrogenous components, most of the vitamins contained the grape juice are taken up by the yeast during fermentation. The results for vitamins found in wine given here are not recent, and ranges are from a wide assortment of wines, and are generally the results of microbiological assays. Nevertheless they give an indication of order of magnitude for some of the most important vitamins [26]:

Biotin	1 – 3.6 µg/L
Choline	17 – 41 mg/L
Inositol	220 – 730 mg/L
Nicotinic acid	990 – 2190 µg/L
Pantothenic acid	470 – 1870 µg/L
Pyridoxine	220 – 820 µg/L
Riboflavin	8 – 245 µg/L
Thiamin	2 – 58 µg/L
Cobalamine	0.0 – 0.16 µg/L.

3.3.2. ***Inorganic Constituents***

Approximately 1000 organic and inorganic constituents have so far been detected and determined in wine, of which ca. 50 are inorganic.

On careful evaporation of wine over a water bath, a syrup-like residue is formed. This extract consists of the nonvolatile inorganic and organic components of wine. Careful ashing (mineralization) of the extract gives the ash, which is composed of the nonvolatile inorganic constituents of wine, including the carbonates and oxides formed during ashing.

According to *definitionem vinum*, the inorganic constituents in wine are classified as mineral substances, with concentrations of up to 1000 mg/L, trace elements in the ppm range (ca. 1 mg/L or less), and ultra-trace elements (ppb range or ca. 1 µg/L).

Typical total concentrations per liter of wine are: mineral substances ≤ 5 g, trace elements < 50 mg (1 % of the mineral

substances), and ultra-trace elements ca. 50 µg (0.001 % of the mineral substances).

The ca. 50 inorganic constituents analytically determined in wine include eight mineral substances (four cations and four anions), ca. 25 trace elements, and ca. 20 ultra-trace elements, as well as rare earths and radioactive elements. About 500 articles have been published on this subject up to 1965 [37] and 20 – 30 new publications per annum have appeared since then.

A survey of the typical concentrations of the inorganic constituents in commercial wines is presented in Table 2, listed in order of decreasing concentration in the ppm and ppb ranges [38].

Table 2. Inorganic constituents in wine (typical normal concentrations in mg/L)

Mineral substances (1000 – 10)		Trace elements								Ultra-trace elements (≤ 0.001)			
		10 – 1		1 – 0.1		0.1 – 0.01		0.01 – 0.001					
Cations		B	2.0 – 5.0	Al	0.25 – 1.0	Ba	0.01 – 0.3	As	0.003 – 0.02	Sb	0.006	La	0.000175
K	370 – 1220	Fe	0.5 – 3.0	F	0.05 – 0.5	Pb	0.03 – 0.1	Co	0.001 – 0.02	Be	0.00008	Ce	0.00052
Ca	70 – 140	Cu	0.1 – 1.5	I	0.1 – 0.6	Br	0.01 – 0.7	Mo	0.001 – 0.05	Cd	0.0005	Nd	0.00026
Mg	60 – 140	Mn	1.0 – 5.0	Rb	0.2 – 4.2	Cr	0.03 – 0.08	Ag	0.005 – 0.02	Cs	0.003	Sm	0.00005
Na	7 – 15	Si	1.0 – 25.0	Sr	0.1 – 3.5	Li	0.01 – 0.5			Au	0.0001	Eu	0.00001
		Zn	0.1 – 3.5	Ti	0.04 – 0.3	Ni	0.02 – 0.06			Hf	0.0007	Gd	0.00006
Anions				V	0.06 – 0.3	Sn	0.01 – 0.3			Nb	0.001	Tb	0.00001
C	100 – 120									Pt	0.000001	Dy	0.00012
P	130 – 230									Hg	0.00003	Ho	0.00003
S	50 – 100									Se	0.0006	Tm	0.00003
Cl	20 – 80									Ta	0.0005	Yb	0.00026
										Tl	0.0001	Lu	0.00005
										Th	0.0001		
										U	0.001		
										Bi	0.00015		
										W	0.003		

3.3.2.1. *Enological Trace Concentrations*

The total concentration of a trace or ultra-trace element in commercial wines (normal concentration) is made up of the primary concentration plus the secondary concentration. Generally, the total concentration is analyzed. Correct determination of the primary concentration is difficult and laborious since all steps from grape harvest to finished wine must be conducted completely free of contamination. The secondary concentration is important for evaluating impurities (“contamination”).

The primary concentration of a trace or ultra-trace element in wine is the natural concentration that enters the grapes and wine from the vineyard soil via the roots of the grapevine. This is generally the characteristic and larger part of the total concentration. The primary concentration often varies over a wide range because vineyards belong to various geological formations.

The secondary concentration of a trace or ultra-trace element refers to an impurity (“contamination”) of geogenic (natural) or anthropogenic (man-made) origin (Fig. 1). The known cases of contamination are listed in Table 3. An excessive secondary concentration is an indication not only of the cause, but also the extent of contamination [38].

Table 3. Secondary concentrations of trace elements in wine (contamination)

Origin	Trace element	Excessive secondary concentration, ppm	Normal concentration, ppm

Unavoidable geogenic contamination			
Ocean, S. France	Br	1.0	0.3
	Br	1.5	0.2
Salt lake, Algeria	Br	2.6	
Volcano, Aeolian Islands	B	13.6	4.5
Volcano, Etna	Al	0.2	0.05
	Fe	18.3	5
	Mn	11.7	
	Zn	3.9	
Exhalation, Kaiserstuhl (grapes)	Hg	0.0016	0.0008
Avoidable anthropogenic contamination			
Aluminum plant			
Switzerland	F	0.5	0.1
S. Germany	F	0.7	0.1
Lead smelter, Rhein	Pb	0.5	0.2
Pigment plant, Neckar	Pb	0.3	0.1
	Cd	0.03	0.005
Italien	F	2.3	1.3
Cement works			
S. Germany	Tl	0.0009	
Franken	Tl	0.0008	0.00015
Coal-fired power station, Baden-Württemberg	Hg	0.0001	0.00005
Mining dump, S. Germany	Tl	0.0008	0.00015
Roads			
S. Germany	Pb	0.5	0.1
Italy	Pb	0.35	0.2
Greec	Pb	5.6	0.6
Brickworks (vine leaves)	F	~ 200	0.02
Enological contamination (viniculture and wine production)			
Pest control			
Albanien (arsenate)	As	0.92	< 0.1
USA (arsenate)	As	0.13	
Australia (lead arsenate)	Pb	0.31	0.03
	As	0.08	0.01
Spain (Hg fungicide)	Hg	0.67	
Wine storage			
Concrete tanks	F	15.1	0.22
Cr – Ni steel tanks	Cr	0.06	0.02
	Ni	0.05	0.04
Al tanks	Al	1.90	0.63
Wine technology			
Wine pipes	Sb	0.09	0.006

Wine pump	Cr	0,2	
Wine treatment			
Removal of H_2S odor	Ag	0.63	0.006
Kieselguhr – asbestos filtration	Cr	0.063	0.022
Bentonite fining	Al	0.94	0.63
Wine bottle	Cr	5 – 12	
Glass	Pb	0.96	0.30
	Pb	0.56	0.22
Lead foil	Pb	0.73	
	Pb	0.41	0.12
	As	0.016	0.005
	Sb	0.037	0.004
“Ecological Wines”	Al	1.75	0.63
	Si	13.5	9.75

Figure 1. Secondary concentration of trace elements in wine

Natural contamination is of geogenic nature and practically unavoidable. It depends on the location of the vineyard. If grapevines grow near the sea or salt lakes, or close to active or extinct volcanos (exhalations), the wines obtained from these locations usually contain insignificant, but readily detectable, increased secondary concentrations of aluminum, arsenic, boron, bromine, chlorine, iron, gold, iodine, manganese, sodium, niobium, mercury, or zinc. These substances enter the finished wine via the grapes rather than through the roots.

Anthropogenic contamination, the true contamination of wine, can be caused in diverse ways, including industrial sites, enological contamination, or the addition of adulterating agents. It can almost always be prevented or at least largely reduced.

In wines from vineyards located near factories, steelworks, cement works, brickworks, coal-fired power stations, mining industries, highways, or cities, increased secondary concentrations of arsenic, lead, cadmium, boron, mercury, thallium (and fluorine on vine leaves) can be detected. Old (pre-1975) and younger wines from grapes grown near busy roads do not differ in their ultra-trace content of platinum (ca. 0.5 – 1.0 ppt = 0.0005 – 0.001 ppb). This cannot be said with certainty of the soil or the leaves, stems, and grapes of *Vitis vinifera* [39].

Enological contamination can be caused during winegrowing (fertilizers and pesticides), grape harvesting, and wine production (making of must and wine, storage, treatment, and transport of wine).

Contamination of wine cannot be excluded if crop-protection products that contain lead and cadmium impurities (137 – 1037 ppm Pb and 22.5 – 351 ppm Cd) are used. However, this contamination does not exceed the maximum values of 0.3 mg/L Pb and 0.01 mg/L Cd permitted by German wine regulations [40]. Until now, excess secondary concentrations due to crop-protection agents having Cu, Fe, Mn, Zn, or F, Cl, Br, or P formulations have not been detected. The average aluminum content of ecologically produced wines of 1.75 mg/L, which are treated with alaun-containing crop-protection agents, is more than 1 mg/L higher than the natural aluminum content of wine [41].

It is difficult to assess the possible transfer of trace elements to the vineyard soil by fertilizers (natural and artificial mineral fertilizer, manure, garbage compost, garbage sewage sludge compost, etc.). It results in hardly any or no secondary concentration in wine. In the case of *Vitis vinifera*, the heavy metal uptake from vineyard soil contaminated with artificial fertilizer and garbage sewage sludge fertilizer (outdoor experiments) by vine and grape components is very low. The uptake in must and wine is negligible. The vine roots act as a natural filter against heavy metals, especially lead and cadmium [42].

In grape harvesting, must production, wine preparation and wine storage, a range of materials are used, including nonferrous metals and alloys. As a result of corrosion and wear, excess secondary concentrations of antimony, arsenic, lead, cadmium, iron, manganese, silver, bismuth, zinc, tin, etc. can appear in wine. The rigorous use of high-grade steel and plastic results in trace-element free wines (e.g., 0.05 mg/L Pb and 0.5 µg/L Cd). On long-term storage, tanks made of stainless Cr – Ni steel can release slight traces of Cr and Ni, cement and concrete containers can release traces of Si and F, and glass-lined vessels and glass bottles can release traces of Si into wine.

Today, trace contamination of wine-treatment agents is the exception. Studies with model solutions and with wine show that contamination by these substances, which are frequently “natural” products such as bentonite, kieselguhr, lime, or activated carbon, is usually negligible.

A high secondary concentration in wine results from the deliberate addition of forbidden adulterating agents, usually in larger quantities, e.g., compounds of aluminum, antimony, barium, lead, bromine, boron, fluorine, manganese, strontium, or bismuth. These additions are intended to improve the wine via preservation, clarification, or correction of color, odor, and taste.

The enological topography of inorganic constituents can provide information not only on the distribution of primary concentrations over parts of the vine and grapes of *Vitis vinifera*, but also on a rare contamination of the grape skin. Trace impurities barely adhere to the smooth grape skins, which are covered by a thin wax coating, and can easily be washed off by rain. Thus, a distinction can be made between “external” and “internal” trace element content.

In element species analysis, the concentrations of elements can be defined according to chemical and physical states in a matrix. Statements can be made about the mobilizability, transport processes, plant availability, resorption, toxicity, etc. In enological element species analysis, wine can be regarded as the matrix.

Due to the large number and concentration of organic constituents in wine, element species analysis is more difficult and demanding than the trace analysis of the total concentration. Noninvasive and nondestructive analytical methods must be used to avoid disturbing the complex equilibrium in the wine sample.

In the case of calcium in wine, the magnitude of the element species stability ($E\text{-}St_{Ca}$) is used as a measure of the binding strength of the wine matrix. With this factor, a characterizing value for calcium stability in wine is obtained. Consequently, the storage stability of a wine with regard to calcium (crystalline deposits of tartrate salts) can be assessed after the determination of the calcium species stability [43].

Other element speciation results are available, e.g., for the trace elements aluminum [41], lead [44], copper, nickel, iron, and for the ultra-trace element bismuth. Bismuth is present in a water-soluble state in the flesh and juice of grapes, presumably ca. 90 % in the form of a complex with sugar acid [45].

3.3.2.2. *Metal Depletion in the Fermentation and Preparation of Wine*

The term depletion (dilution) refers to the decrease in concentration of mineral substances, trace elements, and ultra-trace elements during fermentation and preparation of wine. This depletion can be followed quantitatively and represented graphically and it can be element-specifically characterized by a depletion factor or the degree of depletion.

The fermentation of sugar to alcohol involves a natural depletion process for numerous metals, with a degree of depletion of up to 100 %. Little is known about how the metals are precipitated into the yeast sediment during fermentation. This occurs partly via sparingly soluble sulfides or hydroxides and partly by adsorption on suspended and turbid material. During fermentation and the subsequent steps leading to finished wine, metals are eliminated with various degrees of depletion. Cobalt, nickel, manganese, and zinc are eliminated at a low degree of depletion of up to 25 %; antimony, arsenic, iron, molybdenum, titanium, and tin at a moderate degree of ca. 50 %; while aluminum, lead, cadmium, copper, mercury, silver, and bismuth are eliminated to almost 100 %. Therefore, fermentation is an excellent natural self-purification process for many (toxic) heavy metals.

This purification process is supplemented by enological depletion via wine fining, wine casse, and wine sedimentation (crystals and “depot”).

In blue fining with potassium ferrocyanide (Möslinger fining), iron, cobalt, copper, and zinc are eliminated almost completely, and lead, cadmium, manganese, and vanadium are eliminated in considerable amounts as sparingly soluble metal ferrocyanides in the fining sediment.

Heavy metals can also be depleted in the filtration sediment by formation of metal turbidities, which appear at a threshold value of 10 mg/L Al, 5000 mg/L Zn, 3 mg/L Fe, 1 mg/L Sn, 500 mg/L Pb, 100 mg/L Ni, or 0.3 – 0.5 mg/L Cu, and in crystal deposits and the “depot” (insoluble settlings of old red wines).

The yeast, fining, and filtration sediments are separated as completely as possible from the pure wine. The clear filtered wine is completely or largely free of (toxic) heavy metals. It is the purest wine in the entire wine producing process. Any subsequent metal contamination is to be strictly avoided because a renewed elimination is possible, if at all, only with considerable effort and usually with a reduction in wine quality.

This contributed to the fact that strict maximum values for numerous metals in wine could be stipulated very early (Table 4) long before other environmental standards for beverages or foods.

Table 4. Maximum contents of trace elements (mg/L wine)

Element	Wine regulations * 1. 9. 1993	Drinking water regulations * 5. 12. 1990	Total content
Al	8	0.2	1
As	0.1	0.01	0.007

Pb	0.3	0.04	0.05
B	6.2	1	3
Br	0.5		0.3
F	0.5	1.5	0.1
Cd	0.01	0.005	0.0005
Cu	2	(3)	0.25
Zn	5	(5)	1.5
Sn	1		0.1
Sb		0.01	0.006
Ba		1	0.1
Ca		400	70 – 140
Cr		0.05	0.05
Fe		0.2	3
K		12	370 – 1120
Mg		50	60 – 140
Mn		0.05	0.7
Na		150	7 – 15
Ni		0.05	0.04
Hg		0.001	0.00005
Se		0.02	0.005
Ag		0.01	0.01
P_2O_5		5	130 – 230
SO_4		240	60 – 120
Cl		250	20 – 80
SiO_2		40	20

* Regulations in Germany.

Important characteristic data relevant to wine can be summarized in an element vinogram (for all mineral substances, trace elements, and ultra-trace elements). It contains, above all, element-specific and element-characteristic details about the concentrations in wine, including topography and speciation, about the secondary concentrations (contamination), about metal turbidity and depletion, etc.

3.3.2.3. *Enological Sample Bank*

Wines from earlier centuries can be used to compare trace elements in old and new wines. Trace analyses by modern methods allow incorrect results from older work to be eliminated. Old wines were previously almost never tested for trace elements or ultra-trace elements. In this evaluation, some uncertainty remains because the history of the sample bottle over long periods of time is not known to the last detail (refilling, recorking, storage conditions, etc.). Possible interactions between wine and glass must be taken into account.

Seven wines from the 1700s, 13 wines from the 1800s, and 37 wines from 1900 – 1950 have been tested for cadmium, lead, cobalt, copper, nickel, mercury, and zinc (Table 5).

Table 5. Concentration of trace elements in wine (average values in µg/L)

Period	Number	Cd	Pb	Co	Cu	Ni	Hg	Zn
1700s	7	4.56	2468	6.69	275	40.8	0.10	1340
1800s	13	3.66	4050	13.29	282	54.3	0.13	2040
1900 – 1950	37	2.53	602	6.25	746	87.7	0.17	2587
Modern wines		0.5	30 – 50	1 – 10	250	30 – 50	0.02 – 0.07	1500
Maximum values*		10	300		2000			5000

* According to German wine regulations.

The lead concentrations in wine have clearly decreased in the course of the last three centuries (factor ca. 100) to values of 0.05 mg/L today. The same applies to cadmium concentrations, which have decreased from 4 to 0.5 µg/L (factor ca. 10). These trends best demonstrate the development in wine producing techniques towards wines of increasing purity. A slight increase in nickel and chromium concentrations is observed, especially since the introduction of stainless Cr – Ni steels in the mid-1950s.

The increasing concentrations of the ultra-trace element mercury in wine was a trend for centuries. It was only in the 1970s

that it decreased rapidly with the widespread decrease in the use of mercury in industry and agriculture.

The measurement of ^{14}C and ^{3}H activities, which can be of natural or synthetic origin, also provides information on the wine year, vintage blend, falsification of the year, and addition of synthetic alcohol to wine.

3.3.2.4. *Essential Trace Elements in Wine*

In earlier times the health value of wine was held in high regard. Wine offers an abundance of mineral substances, trace elements, and ultra-trace elements in an especially favorable composition [46]. Two to three glasses of 1953 pure Ober-Ingelheimer late Burgundy contains largely the same type and amount of mineral substances and trace elements as two preparations of the pharmaceutical industry recommended especially for old and convalescent people[47].

The combination of a very high potassium concentration in wine with very low concentrations of sodium and chlorine is therapeutically interesting, especially in a low-salt diet.

Stress, a term introduced into medical terminology in 1936, can be prevented by the anti-stress minerals calcium and magnesium, which are both present in considerable amounts in wine.

Wine can contribute to meeting the human requirement of mineral substances and trace elements (Table 6).

Table 6. Essential trace element contents

	Recommended daily intake, mg	Wine, mg/L	Mineral water, mg/L	Milk, mg/L	Pharmaceutical preparation, mg/tablet
K	2000 – 5500	370 – 1120	3 – 663		
Na	1100 – 3300	7 – 15	2 – 2000		
Ca	800 – 1200	70 – 140	5 – 1044		37.6
Mg	350 – 400	60 – 140	2 – 536		5
Cr	0.05 – 0.2	0.03 – 0.06			
Co		0.001 – 0.02			
Fe	10 – 18	1	0.09 – 4.5	0.4	6.85
F		0.05 – 0.5	0.08 – 17.4	0.15	0.025
J	0.15	0.1 – 0.6	0.006 – 34.8	0.06	0.002
Cu	2 – 3	0.5	0.005 – 0.1	0.12	0.115
Mn	2.5 – 5	1.5 – 5	0.05 – 1.1	0.03	0.049
Mo	0.15 – 0.5	0.001 – 0.01		0.06	0.05
Ni		0.03 – 0.05		0.25	
Se	0.05 – 0.2	0.0006 – 0.001			
Zn	15	0.5 – 3.5		4	0.05
Sn		0.01 – 0.7			
Si		1.5 – 6	4.6 – 10.7		
V		0.06 – 0.26			

3.3.2.5. *Micronutrients for Vitis vinifera*

For the growth of grapevines, essential micro- or trace nutrients having potent growth factors are iron, zinc, boron, manganese, copper, and perhaps molybdenum. According to the minimum law of LIEBIG, any deficiency jeopardizes the effect of all the other nutritional and growth measures, even the yield and the quality of the grapes, produced by the macro- or primary nutrients [48].

Micronutrients are normally present in adequate amounts in the vineyard soil. If, however, depletion of the vineyard soil occurs due to the harvesting of grapes and this requirement can no longer be met, the vine suffers from acute nutritional disturbances. These can be remedied by suitable fertilizers containing trace nutrients.

Reports are frequently made about the favorable effects of (leaf) fertilization with salts of chromium, molybdenum, tungsten, etc. on vine growth, grape yield, wine quality, and even on wine preparation and ageing [49], without practical significance so far.

3.3.2.6. *Enological Trace Element Taxonometry*

In enological trace element taxonometry, the origin of a wine can be narrowed down and established from the primary concentrations (natural concentrations) of inorganic constituents by using monovariate selection methods (one element) or multivariate selection methods (several elements). If other selection criteria are to be combined, such as type of grapevine, year, and quality levels, the detection of multiple elements can be supplemented by additional data, such as organic constituents, organoleptic properties, or physical parameters.

Enological trace element taxonometry is a two-step process. In the first step, as many elements as possible are determined in multielement methods and, perhaps, other data of the wine are consulted. In the second step, all the data are used for group formation and selection in mathematical processes (discriminant, cluster, pattern analyses, etc.).

The idea of using the inorganic composition of must and wine to differentiate location and type of vine is not new. As early as 1939, the manganese concentration was used to differentiate French red wines obtained from European vines and from hybrid vines. However, a correlation of eight trace elements in soil and must determined by spectral analysis was not sufficient to identify the origin in 1961. In 1969, the determination of the location of Greek red wines grown near the sea was possible via sodium and chlorine.

According to more recent studies with multivariate selection processes of several elements, which are determined simultaneously with modern physical multielement methods, the geographical origin of wines can be classified according to the vine-growing land, vine-growing region, location of the vineyard, and soil formation. This classification is less suitable for the type of grapevine, year, quality level, or climate.

In the case of German wines, a distinction can be made between Alzey and Oppenheim in Rheinhessen, between Mosel and Rheinpfalz, and even between northern and southern Pfalz. In the latter case, it is possible to distinguish between Riesling and Müller-Thurgau wines. It is also possible to distinguish between Pinot noir wines grown in France and in the United States, and in the latter case, between the regions California and Northwest Pacific. In another study, a significant geographical classification was made of Bordeaux and Bourgogne wines as well as of French wines according to five different types of grapevine. Hungarian Riesling wines are differentiated according to three wine-growing areas. Italian red wines from Tuscany and four locations of Venetian white wines can be differentiated. The same applies to Spanish red wines from four Catalan wine-growing regions [50]. Some trace elements, such as manganese, iron, aluminum, alkali metals and alkaline earth metals, are especially suitable for identifying the geographical origin of wines.

Until now, a generally applicable method for the reliable determination of the origin of wine by enological trace element taxonometry has not been developed.

3.3.2.7. *Enological Isotopic Taxonometry*

If the abundance ratios of a series of standard wines are recorded in a data bank as an isotopic vinogram, the corresponding data of unknown wine samples can then be compared.

Enological isotopic taxonometry is an improved and supplementary process for enological trace element taxonometry. It allows the detection of wine adulteration by the addition of sugar (cane, beet, or liquid sugar), alcohol (fermentation and synthetic alcohol), water, and wine of lower quality, and incorrect labeling of the origin.

The radioactive isotopes in wine include natural geogenic ^{40}K, ^{87}Rb, U, and Th isotopes and daughter products of the U and Th decay series, anthropogenic isotropes ^{90}Sr, ^{134}Cs, ^{137}Cs, ^{226}Ra and other fission products from nuclear weapons (detectable, if at all, after Tschernobyl in 1987 and 1988 in grapes, must, and wine in concentrations far below EC limits), and the natural (atmosphere) and/or artificial (from nuclear tests) isotopes ^{3}H and ^{14}C.

Radioactive isotopes such as ^{3}H and ^{14}C are only of limited use in the determination of age and origin (after nuclear tests). The same applies to ^{137}Cs (after a reactor accident) for the determination of location.

The natural variations in the stable isotopes, especially of the light elements hydrogen, carbon, and oxygen, but also the heavy elements strontium and lead have proved to be very useful in solving many problems, such as the addition of sugar, alcohol, and water, the origin and blending, and the purity and contamination of wine. The following abundance ratios are used: D/H, $^{13}C/^{12}C$ [51], $^{18}O/^{16}O$ [52], $^{87}Sr/^{86}Sr$ [53], and $^{208, 207, 206}Pb/^{204}Pb$ [54]. The isotopes are determined by means of high-performance NMR spectroscopy and mass spectrometry.

3.3.2.8. *Enological Element Analysis*

The methods used up to the early 1970s were mainly colorimetry, spectrophotometry, gravimetry, and titrimetry. Since then, modern single-, oligo-, and multi-element methods have been predominantly applied in enological element analysis. With these methods, even the smallest concentrations of trace elements can be accurately determined.

In modern enological element analysis, a few instrumental processes of high detection efficiency and sensitivity dominate. Atomic absorption spectrometry (AAS) with its numerous variants (flame, graphite furnace, hydride and cold-vapor processes) is of great importance as a single-element method for a large number of trace and ultra-trace elements. Simultaneous multielement determination methods are being developed.

Frequently used oligo-element processes are based on electro-analysis. Differential pulse anodic stripping voltametry (DPASV) allows the determination of As, Bi, Cd, Co, Cu, Ni, Pb, Sb, Se, Tl, Zn (and Hg on gold wool).

Pb, Cd, Cu, Zn etc. can be directly determined in wine by means of potentiometric stripping analysis (PSA) [55], [56].

Multielement methods for simultaneous determination are wavelength and energy disperse X-ray fluorescence analysis and inductively coupled plasma atomic emission spectroscopy (ICP-AES), optical emission spectroscopy (ICP-OES), and mass spectromery (ICP-MS) with the highest sensitivity and possibility of isotopic separation.

Some special methods also find application for special problems in enological element, isotope, and speciation analysis [57].

[Top of Page]

4. Grapes and Grape Varieties

Ralph E. Kunkee

The grapevines from which wines are made are all members of the genus *Vitis* and belong to the botanical family Vitaceae. This family includes some ornamental vines, but no other food plants besides grapes [1]. The species *Vitis vinifera* makes up essentially all wine grapes of the world, the main exceptions being the species *V. labrusca* (perhaps a natural hybrid of *V. labrusca* and *V. vinifera* [58]) and *V. rotundifolia* (or *Muscadinia rotundifolia* [58]), both native to North America. These latter two species have intense, if not highly regarded, non-*vinifera* odors and flavors; their use for winemaking is generally confined to their geographical origins. Koshu is the native non-*vinifera* variety of Japan, and there are other Asian non-*vinifera* varieties. Some winemaking is also based on hybrid vines from genetic crossings of *vinifera* and non-*vinifera* parents, whereby some of the disease resistance of the non-*vinifera* parent, is inherited but these are also not important in the global wine picture. The discussion here is devoted to *V. vinifera*, often referred to as the wine grape or the European grape, since the vast quantity of the world's wine is produced in Europe from this grape. However, the vine's origin seems to be, rather, more Middle Eastern, perhaps near modern-day Iraq [1]. (Some authorities divide the *Vitis* genus into two subgenera, *Euvitis* and *Muscadiniae* [59], the latter including only some North American species and the *Euvitis* being subdivided into North American and Eurasian species, in which *vinifera* is placed [59]. However, the single genus *Vitis* is used here).

4.1. *Vitis vinifera* Varieties

The “wine grape” (*Vitis vinifera*) does not tolerate extreme climatic conditions, and its growth is limited to the temperate zones, 30 – 50° N and 30 – 40° S, (10 – 20 °C isotherms) [58], [60].

The varieties are simply divided into “white” and “red” grapes. White varieties range in color from light green to dark yellow, and the red ones from pink to deep purple or black, the degree or depth of color being highly dependent upon cultural conditions.

There are some 15 000 named sub-species (cultivars or varities) of *V. vinifera*, but probably only 1000 have been exploited for wine production, and of these, perhaps only a few dozen are actually in use commercially. The varieties are often identified with famous old-world wine regions, the appellations thereof indicating the varieties. Some of the most important *V. vinifera* varieties are listed here, indicating those of greatest global planting (*) [1], [3], [16], [59-61]:

Airén*
Aligoté
Aramon
Barbera*
Black Malvoisie
Boal
Bobal *
Brunello
Bual
Burger
Cabernet franc
Cabernet Sauvignon*
Canaiolo Nero
Carignane*
Charbono
Chardonnay
Chasselas doré
Chenin blanc
Cinsaut
Colombard
Criolla
Durif
Ehrenfelser
Elbing
Fendant
Flora
Folle blanche
Franken Riesling
Fumé blanc
Gamay
Garnacho

Gewürztraminer
Golden Chasselas
Grenache*
Grenache blanc*
Grignolino
Gutedel
Johannisberg Riesling
Kerner
Macadeo*
Malbec*
Malvasia
Malvasia bianca
Mataro*
Mazuelo
Melon
Merlot*
Meunier
Mission*
Montalcino
Montepulciano
Mourisco
Mourvèdre
Müller-Thurgau
Muscat*
Muscat of Alexandria
Muscat Canelli
Muscat Frontignana
Muscat Hamburg
Natari
Nebbiolo
Orange Muscat
Palomino
Petit Sirah
Pic poule
Pineau blanc de la Loir
Pinot noir
Pinot grigeo
Pinot gris
Pinot blanc
Pinotage
Piquepoul
Portuguese blue
Red Malaga
Regosco
Rhine Riesling
Riesling*
Riesling Sylvaner
Rkatsiteli*
Saint Émilion
Sangiovese
Sauvignon vert
Sauvignon blanc
Scheurebe
Sémillon*
Shiraz
Souzão
Spätburgunder
Steen
Stein

Sultana
Sylvaner
Syrah
Tempranillo
Thompson Seedless
Tinta Roriz
Tinta Madeira
Tinta Cão
Touriga
Trebbiano*
Ungi blanc
Valdiguié
Veltiner
Verdelho
Verdicchio*
Viognier
Viura
Walschriesling*
Xarel-lo*
Zinfandel

Longer lists can be found in [1], [3] and more-or-less complete compendia are given in the amelographies of GALET [61] and MORTON [62].

Worldwide use of *V. vinifera* is not limited to wine production; the grapes are also used as fresh table fruit and dried for raisins. The relative proportions depend upon the market value of the three commodities, religious prohibitions against wine and brandy, and seasonal and transport considerations.

V. vinifera in Winemaking. The most prominent attribute of *V. vinifera* — besides the properties of its juice which allow the production of a beverage of high ethanol content and high acidity, giving wine its microbiological stability — is that the product does not have overstated flavors, but rather a subtlety that allows the flavors to remain unobtrusive.

The tannin and anthocyanin pigments (see Chap. Chemical and Physical Composition of Wine), which provide the red color and astringent and bitter flavors to wines are highly localized in the skin of *V. vinifera*. For red wine production these elements must be extracted, which is generally done by fermenting these grapes “on the skins”, the heat and ethanol releasing them from the skin cells. For white wine production from white grapes, the skins are generally removed before fermentation, to decrease the bitter and astringent components. White wines can also be made from red grapes by complete and careful removal of the skins before the onset of fermentation. Some two-thirds of Champagne production is from Pinot noir, a red grape. In all cases, extraction of the very harsh tannic materials from the grape seeds is to be avoided, preferably by early removal or by preventing crushing of the seeds.

V. vinifera Identification. Until recently, precise varietal identification has been extremely difficult because of large phenotypic variations caused by environmental factors and cultural practices. However, DNA profiling methods have begun to clarify some of the historical varietal confusion [63]. Varieties can be further subdivided into clones. Clonal selections can be made from examination of vines in sections of vineyards containing a single variety, noting the variations thereof, and then making vegetative, rather than cross-breeding, propagations of vines with superior characteristics. The DNA profiling methods used to date do not generally detect clonal differences, but there is hope that more sensitive genetic techniques will eventually allow the identification of clones [64].

4.2. Rootstocks

The major disadvantage of *V. vinifera* varieties is their susceptibility to various diseases, (see Section Grapevine Diseases), mainly originating from the new world, and including especially the root-sucking aphid Phylloxera, commonly called the root louse, but more accurately *Dactylasphaera vitifoliae* [16], [58], [65]. This insect, literally sucking the plant to death, caused devastation of European vineyards, especially in France, starting in 1863, with reoccurrences from time to time, in France in the 1910 s, and currently in California. The life cycle of the aphid and its relationship to the grapevine's roots are complex. Currently the only definite antidote is to graft the desired variety of *V. vinifera* (as the top part, or scion) onto a rootstock (bottom part), which is completely free from any *vinifera* parentage. The most successfull rootstocks are those having evolutionary resistance to the pest, that is, *Vitis* species originating in North America, such as varieties of *V. riparia* and *V. rupestris*, and hybrids thereof. The modern phylloxera problem, especially in California, is further complicated by the appearance of a new form of the insect, the so-called biotype B [65].

Nematode roundworms are another form of vineyard devastation. They also attack the roots of the vine. Not only do they bring about root damage, they can also act as virus transmittters, or vectors. The solution to nematode damage is not as developed as that for phylloxera. The roots of some non-*vinifera* species (*V. solin* and *V. champini*) show resistance and some chemicals are helpful in control [16], [58].

[Top of Page]

5. Viticulture

Ralph E. Kunkee

Viticulture, the science of grape growing, is very complex, both because of the great global variations in its culture, and because of the complexity of the grapevine itself. Its importance cannot be overstated; it is truly said that the production of great wines starts in the vineyard.

Commonalities in these considerations are the dormancy stages of the annual growth cycles of *V. vinifera* and how these are utilized by vineyardists and the importance of the agreement of grape variety with the *terroir*, the total natural environment of the region, including soil and climate.

5.1. Dormancy and Its Utilization

The dormancy period of *V. vinifera* occurs with the beginning of winter's short days and cold temperatures [58]. (In fact, in warm or equatorial regions, such as India and Venezuela, there are no natural dormant periods, and dormancy must be induced by heavy pruning of the vegetative growth.) During the dormant period, the viticulturalist has the opportunity to prune parts of the dormant wood. This is done first to control the number of buds to be left on each vine, which dictates the number of new shoots and potential clusters, and berries, and thus the size of the next harvest. Equally important, the dormant period allows the shaping, the type of trellising, or canopy management of the new growth coming after “bud break”, that is, after the reversal of dormancy with the advent of spring.

Crop Size. The extent of dormant wood removal and the number of buds to be retained is an important decision of the vineyard manager, as this influences the next crop size. The crop size, the weight of berries produced per vine, is generally considered to have a profound effect on the quality of the wine produced. The arguments are first, that a vine can only support so much crop, and if “overcropped” the berries will not ripen sufficiently; and second, small yields (“undercropping”) tend to produce larger berries, with lower surface area to volume ratios – the color and flavors being concentrated in the berry skins.

Interestingly, in Australia, where pruning labor costs are relatively high, there is a new proclivity for minimal or no pruning [16]. The expectations seem to be that after a few years, the grapevines, left to themselves, adjust and bring into balance crops which are large enough and with berries small enough to warrant this “benign neglect”.

Canopy Management. The last few years have seen new interest in shaping and trellising systems, the positioning of buds and the pruning of the new growth after dormancy — all coming under the term canopy management [69]. The main idea is to manipulate the areas and positions of the leaves of the grapevine in relation to the sunlight (and wind and rain), to engage the best aspects of region's soil and climate, bringing about the full potential of the vine in producing its healthiest and heaviest crop, while at the same time giving a crop fully ripened and with full distinct varietal flavor [66].

5.2. Agreement between *Terroir*, Climate and Grape Variety

Terroir is the French word for soil, but for the viticulturalist, the word has come to mean more than just the composition and structure of the earth of the vineyard, or of its drainage and layout, etc., and generally includes the total natural environment of the region, and often the climate [16]. In vineyard site selection, regardless of the canopy management to be employed, all aspects of *terroir* ought to be considered.

There are two schools of thought on the importance of *terroir*, in its original, narrow definijtion, as compared to the importance of climate in the production of high-quality wine. Arguments on the one side have been that grapes from adjacent vineyards with two types of soil often produce two types of wine in the same climatic year. Nevertheless, only recently have actual studies and comparisons of compositions and structures of soils from designated vineyard sites been made. [67]. Arguments on the other side have been that grapes from a single vineyard often produce varying kinds of wine from year to year, obviously due to variations in climate, not in the soil.

Viticulturalists in old world wine regions have not given great importance to the role of climate in vineyard designations, since the establishment of appellations and their proper varieties have been historical and evolutionary actualities. Trial and error and production of high- and low-quality wines over many centuries dictated success. In the new world wine regions, this long experience is not available; it has been necessary to provide some sorts of climate classifications. One of the most successful has been the California heat summation system, where the vineyard regions have been divided into five climatic zones, depending on the product of number of days with temperatures greater than 10 °C multiplied by the differences in actual temperatures and 10 °C for those days [26].

Modern viticulturists see the importance of the total concept of *terroir*, in order to provide the winemaker with grapes which have their full potential, free from disease and expressing all of their varietal distinctiveness.

5.3. Grapevine Diseases

V. vinifera grapevines are susceptible to many pests besides phylloxera and nematodes (see Section Rootstocks). MULLINS et al. [58] list — and give symptoms and control measures for — six fungal diseases of the leaves and fruit, four fungal diseases of the vascular system, three bacterial diseases, three viral diseases, and one miscellaneous disease. *Botrytis cinerea* devastation is a global vineyard problem, often named “bunch rot” or “gray rot”. This fungus does not require accidental lesion of the berry skin for infection, as do many other fungi. *B. cinerea* is only the fungus which aids in production of the highly acclaimed “late harvest” wines (see Chap. Wine Classifications, Types, and Labels), and is in that case named “noble rot”. Some microbial infections depend on weather conditions and some on cultural practices. For example, downy mildew fungus (*Plasmopara viticola*) is generally found in regions of summer rainfall; the bacterial infection of Pierce's disease (*Xylella fastidiosa*) is found near the riparian habitats of its insect vector, a leaf hopper. The fungi producing powdery mildew (*Uncinula necator*) and black rot (*Guignardia bidwellii*) both originated in the new world; and arrived in Europe (last century) to start, and then to finish off, respectively, the devastation from the co-immigrant Phylloxera.

The other listed diseases [58] can be just as destructive as the ones especially mentioned here.

5.4. Other Viticultural Practices

For further information on viticultural practices, not discussed here, such as the use and roles of irrigation, fertilization, integrated pest management, grapevine biology, physiology and anatomy, etc., see [3], [16], [58], [68].

[Top of Page]

6. Alcoholic Fermentation

Ralph E. Kunkee

“Must” is the enologic term for grape juice which is ready for fermentation. For white wine production, must usually refers to juice, without skins and seeds. It can also refer to crushed grapes, that is, the juice, skins and seeds, which is normal for red wine production. Alcoholic fermentation is fermentation of the grape sugars to ethanol and carbon dioxide by wine yeast. Malolactic fermentation by wine bacteria is discussed under secondary fermentations (Section Malolactic Fermentation).

6.1. Pre-fermentation Operations

Harvest. For best quality of wine, the grapes should be ripe enough that the sugars are adequate to ensure the high concentration of ethanol desired, but not so overripe that the acidity is too low (i.e., that too much of the malic acid in the berry has been respired). Much of the world's wine grape harvest is done mechanically, which from a quality standpoint has several advantages: more precise scheduling of harvest because of the smaller labor force needed and in warm regions, the capability of harvesting in the coolness of the night. The main disadvantages, and they are important ones, are that the juice is prematurely exposed to air and the oxidation problems which this brings and that the rough treatment of the skins is more likely to allow extraction of harsh and bitter materials (see Section Polyphenols) [3].

Crushing and Stemming. Often crushing and (de)stemming are performed by the same of equipment. The stems are removed since, like the seeds, they can add harsh flavor. Also it is important that all of the grapes be broken open so that the yeast will have contact with the juice for fermentation [3]. There is increasing interest in mild crushing operations that avoid introduction of harsh tannic materials. Furthermore, the idea of “whole berry” fermentations has some adherents, that the presence of unbroken berries, at least until near the end of the fermentation, provides some special flavors. The ultimate in this concept is so-called “carbonic maceration” [3]. In this operation, most of the berries are not crushed, but completely immersed in carbon dioxide gas. At somewhat elevated temperatures, the berry enzymes begin fermentation of the sugars to ethanol. Eventually, the grapes are broken open and the fermentation is finished in the normal way. This gives a special flavor and provides wines ready for bottling much sooner [3]. This process is associated with the wines of Beaujolais, marketed as early as November as Premeur wines.

Sulfur Dioxide Addition. Sulfur dioxide is universally used in wine production, both to inhibit oxidation reactions and as an antiseptic [3]. The latter activity is important to inhibit indigenous yeast and bacteria, especially when a yeast starter culture is not used or if the grapes are infected with rot. Sulfur dioxide is a slow scavenger of oxygen but shows rapid anti-oxidation activity in the inhibition of polyphenol oxidases of the must, which can cause severe browning. Since much of the oxidized phenols become further polymerized, these brown pigments precipitate. Winemakers nowadays are careful to use low concentrations of sulfur dioxide for this gentle method of removing browning. Sulfur dioxide is also added during the storage of wine, for its antiseptic and anti-browning activities, and generally just before the wine is bottled. Sulfur dioxide has some objectionable qualities in regard to flavor, and perhaps to health, and thus needs to be used judiciously. There have lately been tendencies to decrease the amounts used.

Draining, Pressing and Settling. For white wine production, the juice is drained to separate it from the broken skins and seeds. The remainder of the juice is separated with a press; the milder the pressure, the lower the extraction of astringent and bitter components. For the highest quality of product, much of the remaining particulate material is removed by allowing the must to settle or by centrifugation. For red wine production, there is no removal of the skins (and no pressing or settling operations) until well after the beginning of fermentation.

To provide some stylistic imprints to their winemaking, some winemakers allow the crushed grapes, usually red ones, to be stored for several days at cool temperatures, without addition of yeast and usually without addition of sulfur dioxide. This procedure is not without difficulties. If the acidity is too low (pH > 3.5), the process can lead to infections by lactobacilli, bringing about ruinous concentrations of acetic acid.

Wine yeasts can be defined as those yeasts which carry out complete fermentation of grape juice (i.e., those having high tolerance to acid and to ethanol) and give an end product with an obvious vinous character with no off-flavors or odors [3]; they are limited to strains of *Saccharomyces cerevisiae* [69] or *S. bayanus* [70]. It was previously thought that wine yeasts were found indigenously in high concentrations in the vineyard, and had something special to do with the style of wine coming from adjacent wineries, in so-called “natural” fermentations, where yeast starter cultures are not used. It is now rather accepted that the indigenous aspect in natural fermentations is the winery itself and the winery equipment [3], [71].

Wild yeasts can be defined as yeasts which are not so tolerant to ethanol and carry out fermentation of grape juice to a few percent of ethanol. These yeast are easily isolated from the surfaces of fresh grapes, and are members of several genera: *Kloeckera*, *Hansenula*, *Hanseniaspora*, *Pichia*, *Candida*, *Metchnikowia*, *Debaryomyces*, *Kluyveromyces*, as examples. In laboratory experiments, these yeasts often produce fermentation products with captivating and intriguing odors which, however, are always very short-lived and disappear when the fermentation is allowed to go to completion by addition of wine yeast.

Grape Juice as a Wine-Yeast Medium. Wine yeasts grow on minimal media; they are capable of making all the cellular components necessary for growth, with the exception of the vitamin biotin. However, wine fermentations require not only

growth of the yeast, but production of cells having sufficient vigor to complete the vinification and to produce the high concentration of the inhibitory product, ethanol. Grape juices generally are abundantly supplied with nutritional components for the yeast biomass and for the fermentation. Furthermore, the high acidities of grape juice, the anaerobic vinification conditions and the resulting high concentrations of ethanol prevent the growth and activity of unwelcome microorganisms.

Yeast Inoculation. In the new-world wine regions, and more and more in the old world, addition of commercially produced wine yeast, in an “active dry” form, as starter culture is standard. The yeasts are manufactured under highly aerobic conditions, which provide them with high concentrations of sterols and unsaturated long-chain fatty acids, called “survival factors”. These give the yeast high tolerance to ethanol. They are also growth factors, but this role is not as demonstrable as their effect on increased yeast viability at the end of fermentation [3]. The active dry yeast usually must be rehydrated with warm water under special conditions. Enough yeast is usually added to give a starting concentration of ca. 10^6 cells/mL. Once fermentation has commenced, the fermenting must can itself be used a starter culture, and transferred to the vats and tanks to start other fermentations. This is a time-honored system, but it has it perils in that too many such anaerobic serial transfers may lead to dilution of the survival factors and loss of ethanol tolerance.

More traditional methods involve the use of indigenous wine yeasts, most likely resident in the winery and winery equipment. Often the winemaker makes a *pied de cuve*, where fresh juice is obtained earlier and the yeast therein allowed to grow to obtain concentrations similar to commercial starters. During the first day or so, it is common practice to stir (to rouse) the *pied de cuve*, allowing aeration and formation of survival factors. Even though the high concentrations of glucose in the juice are inhibitory to yeast growth and fermentation (the Crabtree effect [3]), the formation of survival factors is not.

A third method finding some favor in the new-world regions is the so-called “natural” fermentation, where the winemaker, as in the above procedure, relies on indigenous yeast but usually without the benefit either of the historic situation or the preparation of the pre-starter. The advantages are said to be wines with more complex flavor and greater mouthfeel, and wines thought to have a special style from the source winery. The problems, well understood and guarded against, are the same as with multiple serial transfers.

6.2. Yeast Growth and Fermentation

The first phase of alcoholic vinification is yeast growth, essentially under nonrespiratory conditions, producing ethanol and carbon dioxide. After a several-hour lag period, the yeast begins exponential growth, which continues up to 10^8 cells/mL. The yeast rapidly takes up ammonia and most of the available amino acids. Lysine, histidine, and glycine are not well used. Proline is not used at all under these conditions. The order of uptake of the amino acids depends on strain and the overall nitrogenous complement of the juice [3]. Branched-chain amino acids are sometimes used intact, but in cases of nitrogen deficiency, they are stripped of their amino groups and converted to the corresponding higher alcohols (see Section Alcohols).

When the yeast has reached its resting phase, about one-third of the sugar has been fermented, giving ca. 5 – 6 % ethanol. This concentration has little obvious effect on yeast activity, and the yeast continues to produce ethanol and carbon dioxide and remains suspended. But some time thereafter, the inhibitory effects of ethanol become noticeable, fermentation diminishes, and nonfunctional yeast begins to settle out. Ethanol expresses its toxicity by altering the fluidity of the cell membrane, causing increased permeability and leakage of solutes into and out of the cell [3]. The survival factors seem to reverse the fluidity effect, preventing increased permeability. Sluggish or stuck fermentations can arise from shortage of survival factors or from inadequate biomass vigor due to initial nutritional deficiency [3]. Fermentation continues, if slowly, until the fermentable sugar is gone, or until the ethanol reaches a concentration that is finally toxic to the cells' activity. By this time, most of the yeast has settled to the bottom of the vessel.

Most or all of the wine yeast strains are glucophilic, in that they take up, and ferment, glucose faster than fructose [3]. Thus, near the end of the fermentation, when tests for glucose might show none present, there may still be as much as 40 g/L fructose, as fermentable sugar (Section Carbohydrates).

The apparent conversion of sugar to ethanol is variable when the sugar is given in terms of total soluble solids, as with the Brix and Baumé, etc., scales (see above), rather than in terms of actual sugar, itself. Conversion depends on the initial nitrogenous complement of the juice, since this influences the proportion of sugar going to form biomass or to be used for ethanol. One formula gives the concentration of ethanol (in volume percent) as equal to the initial Brix reading multiplied by a factor varying from 0.55 to 0.64 [72].

Temperature. If the initial temperature of the juice is very high, or if there is no mechanism (refrigeration or water curtains) to cool the fermentation, the heat produced (an increase of ca. 1.3 °C for 1 % sugar fermented [1]) may raise the temperature to a point lethal to the yeast (ca. 37 °C in wine), or too high for producing good quality wine. Kinetic models have been devised to predict ensuing temperatures, based on initial temperature and sugar concentrations [73]. These models take into account the loss of ethanol by evaporation and by the sweeping effect of large volumes of carbon dioxide, which is also temperature dependent [3]. Generally, lower fermentation temperatures are desired for white wines (10 – 18 °C), and higher ones for red wines (20 – 30 °C). Generally, the lower the temperature, the greater the formation of esters and acetals, the “fruity” flavor components [3], often more desirable in simple wines without varietal distinction. With red wine fermentations, the temperatures during the early parts of the fermentation are often allowed to rise for short periods of time to 30 °C, the idea being to extract special color pigment components.

6.3. Post-fermentation Operations

The pressing of red wine vinifications is delayed until sometime toward the end of the fermentation. If it has not been done during the fermentation, if may be done now. If the pressing is delayed too long the wine may be too harsh and rough for early consumption, but these tannin components allow the wine to age well [74]. Sometimes the pressing is delayed for several weeks (“delayed maceration”). These wines are said to have softer tannins.

Protection from Acetobacter. At the end of fermentation, the wine is normally transferred (racked) from its sediment (lees) to a full container (cooperage), or to one with inert gas in the head space (ullage). The anaerobic conditions protect the wine from vinegar spoilage (see Section Organic Acids and Carbonic Acid.) by the ubiquitous obligate aerobic *Acetobacter.* Sulfur dioxide treatment is usually also carried out at this time to help protect the wine from this and other spoilages.

Storage. Wine is stored for variable lengths of time, from several months to several years, depending upon the kind of wine. Treatments also vary, including the encouragement, or not, of secondary fermentations. Sometimes, storage is in the presence of spent yeast lees, often for several months, with stirring of the lees from time to time. Sometimes the wines are stored in inert containers, but often in wooden, usually oak, casks and barrels. This cooperage allows some removal of harsh materials from the wine, but more importantly the extraction of flavor components from the wood, which are complementary to the wine. There is much information and some mysticism regarding the kind of wood, its origin and treatment, the kind of production of the barrel, including the amount of firing (toasting), and the length of times the wines should be stored therein [3].

Stabilization: Racking, Fining and Filtration. Wines often clarify naturally by the settling action during storage, especially when they are also racked from time to time. This process can be accelerated by use of filtration. As with any treatment, filtration has the potential to do some damage to the wine, but there is no damage if filtration is done correctly. Sometimes insoluble materials with adsorptive properties, such as bentonite, egg white, or isinglass, are added. As they precipitate, they remove the unwanted material and settle with it to the bottom, after which the wine can be racked and filtered.

The above treatments are used to render the wine stable with respect to proteins, metal ions (Section Metal Depletion in the Fermentation and Preparation of Wine), tartrates, and microbes. During storage of the wine, any of these problems are usually easily solved with the above-mentioned techniques. However, the problem of continued stability after the wines are bottled and are in the consumers' hands is another matter: the protein becomes a haze, the metal ions catalyze oxidation reactions and the wine becomes brown, the potassium bitartrate crystallizes out, and the microbes cause turbidity and probably off-flavors and odors. Thus winemakers require a battery of stability tests, to assure that none of their wines will come in harm's way.

[Top of Page]

7. Secondary Fermentations

Ralph E. Kunkee

7.1. Malolactic Fermentation

Malolactic fermentation is caused by special, wine-tolerant lactic acid bacteria which can decarboxylate L-malic acid to L-lactic acid and carbon dioxide [3], [75]. This usually occurs after the end of alcoholic fermentation. The required decarboxylating enzyme is found uniquely in certain strains of *Leuconostoc*, *Pediococcus*, *Lactobacillus*, and *Lactococcus*, the first three having been found in wine. Strains of *Leuconostoc oenos* are by far the most important for commercial wine production.

The main consequences of malolactic fermentation/conversion are:

1. Deacidification of the wine due to loss of one of the carboxyl groups, which is liberated as carbon dioxide gas
2. Utilization and depletion of micronutrients by the growth of these bacteria, rendering the wine biologically stable and generally unable to support further growth of these microorganisms
3. Subtle, but real changes in flavor, in addition to those coming from the change in acidity

The first consequence is of prime importance where grape juices and wines are initially too acidic (e.g., in northern France). The second consequence is of concern at time of bottling of wine, which might be susceptible to malolactic bacteria. Either it should have been stabilized by having had this fermentation or it should now be rendered stable, e.g., by sterile filtration. The third consequence seems to be strain dependent and is desirable, unless the strain produces spoilage flavors or odors. The most commonly identified flavor end product is diacetyl, which, depending on the concentration, has odors associated with sweet and/or spoiled butter. If the amounts made are above the recognition threshold, the wine is properly considered spoiled; but if below, and still detectable, its presence may increase the complexity of the wine and thus be considered a positive attribute [3], [75]. Diacetyl is more noticeable in white wines, where its presence is also more controversial. One of the main substrates for diacetyl formation, at least for leuconostocs, is citric acid but complete management of the formation of this end product is not presently possible.

In the past, winemakers have controlled malolactic fermentation by using the bacteria resident in used cooperage as inocula. Historically, noticeable changes in the wine occurred in the springtime. In modern times, at least in the new world regions, noticeable bacterial activity comes much sooner, usually being complete within a month or two after primary fermentation. The process has been accelerated by suitable adjustments of pH, sulfur dioxide, and temperature. Furthermore, originating in Burgundy, but widely mimicked, wines are retained on their lees (*sur lies*) to provide the bacteria with additional micronutrients extracted from the yeast. However, the ultimate quickening factor has been the California technique of inoculation with starter culture. This has led to the commercial availability of several strains of lyophilized bacteria for this purpose.

Malic acid is not an energy source for these bacteria, and other sources have to be supplied. Nevertheless, the presence of malic acid does stimulate the initial growth rate [3], [75].

7.2. Sparkling Wine

Sparkling wine, of which Champagne is the prime example, is produced by a secondary fermentation of a sweetened base wine to which yeast has been added. The mixture is placed in a heavy walled bottle that withstands the several atmospheres of pressure produced during fermentation. Special tricks, including freezing the necks of bottles after they have been shaken and stored upside down for several months, are used to remove the yeast formed during secondary fermentation [25]. A bulk method (Charmat) is also used, where the yeast is removed by filtration. A new procedure, involving embedding of yeast in alginate gels for the secondary fermentation, shows great promise [22], [76].

7.3. Flor Sherry

The flor sherry process, wherein special yeast form a film on top of “shermat” wine (sherry material), and metabolize it, and similar processes, are described in Section Dessert/Appetizer Wines. The metabolism is actually an oxidation rather than a fermentation. This process can also be adapted to a bulk “fermentation”, in which the yeast are submerged (“submerged flor” process) in the shermat. The wine is put into a pressure tank and subjected to ca. 1 bar with occasional stirring, which serves to increase the oxygen tension in the wine enough to allow the yeast to respire. While this process certainly gives many of the same end products as those produced in Jerez, the later small-scale natural production in wooden cooperage is considered superior.

7.4. Microbial Spoilage

Not all secondary fermentations are desirable. Attention must be paid to the undesirable ones, the spoilages. However, modern commercial wine cellars are relatively free from microbial spoilages, because of regard to established cellar sanitation methodology, along with judicious use of sulfur dioxide and with employment of rough filtration from time to time.

Wine Yeast and Wine Bacteria. Spoilage can originate from otherwise desirable wine yeasts and wine bacteria if they are in the wrong place or at the wrong time. For example, turbidity coming from refermentation of a semidry bottled wine, in the bottle, by a strain of wine yeast, certainly must be classified as spoilage. The same could be true of an esteemed strain of malolactic bacteria being in the wrong place at the wrong time.

Spoilage Yeast. Many wild yeasts (Section Pre-fermentation Operations) will survive in wine and provide some cosmetic problems in bulk cellar storage of wine, problems easily corrected as mentioned above. However, a few genera are damaging, causing turbidity or off-flavors and odors. The most important of these are *Zygosaccharomyces*, *Saccharomycodes*, and *Brettanomyces*. *Zygosaccharomyces* are resistant to sorbic acid, *Saccharomycodes* are resistant to sulfur dioxide, and *Brettanomyces* seem to thrive on wines with minuscule amounts of residual sugar or in presence of new wood [3]. Thus, the definitive control of each of these is filtration. *Brettanomyces* has received most attention, probably because of it ubiquity, the difficulty of completely eliminating it, and the intense controversies over desirability of the putrefactive flavors and odors of its end products [3], [77]. *Dekkera* are the ascosporogenous forms of *Brettanomyces*; their importance in wine has not yet been established.

Spoilage Bacteria. Already mentioned is the potential problem with malolactic bacteria in the improper place. Another type of spoilage by these bacteria is that coming from strains which produce foul-smelling or foul-tasting materials. The identification of these strains is empirical.

The problems with lactobacillus spoilage during improper cold soaking has been mentioned (Section Pre-fermentation Operations). Another spoilage problem with lactobacilli comes from infections of Port wines and other high alcohol sweet wines. Interesting, some strains of *L. fructivorans* (*trichodes*) and *L. hilgardii* can thrive in and spoil wines with 20 % ethanol. This is an old problem that has resurfaced. In all cases, past and present, it is easily controlled with sulfur dioxide [3] .

The problems with and control of aerobic vinegar bacteria *Acetobacter* have are discussed in Chapter Chemical and Physical Composition of Wine. Other potentially dangerous aerobes are wine-tolerant bacilli. This kind of spoilage is problematic; it seems to be generally cosmetic and very localized [3].

[Top of Page]

8. Typicité and Style

Ralph E. Kunkee

One sometimes hears the saying, “wine makes itself”. It is true that when perfect attention is paid to the viticultural practices alluded to above (Chaps. Grapes and Grape Varieties, Viticulture), the grapes reach the winery with the highest possible distinctiveness of variety and when further perfect attention has been paid to avoidance of all kinds of defects (Chap. Secondary Fermentations), the chances of a superb wine “making itself” may be realized. However, more than that, these perfect conditions in fact set the stage for the ultimate in wine crafting — the development of individual styles, or expressions of the winery's (or winemaker's) *typicité*. These exciting modes include exploitation of unusual yeast and bacterial strains, and perhaps the development and isolation of new ones; innovative maceration and fermentation regimes; and imaginative *sur lies* and other storage treatments. The strength of the appellation, or estate bottling, systems has been long thought to reflect not only the *terroir* of the select vineyards, but also to have something to do with the special and historical style, or *typicité*, of the winery methods. Whether this has been true in the past or not, it is reasonable to expect that imaginative use of the above techniques, and ones not yet thought of, will provide the future consumer with wonderful wines not as yet envisioned.

[Top of Page]

9. Sensory Evaluation

Ralph E. Kunkee

"Traditionally, sensory evaluation of wine is perceived as being that done by a wine expert, who evaluates the 'quality' of wine" [78]. Indeed, that is one type of sensory evaluation, which asks which wine is preferred, or more sophisticatedly, but possibly just as imprecisely, which wine has the higher quality. NOBLE [78] points out that these kinds of hedonic or quality evaluations have their purposes, and if done under well controlled conditions by aware and like-minded judges can give some approximation to the meaning of "quality". This is of course the basis for competitive judging of wines for medals — local and international — and the ratings of wines by "scores" in wine periodicals and by wine consultants.

More definitive tests are those asking if the wines are different, what is the difference, and how big is the difference. If done in a sequential fashion it can provide detailed information about wines [78]. These analyses must be done under controlled conditions and with the use of trained subjects. These kinds of tests can also be used to obtain unspecified differences, thresholds, and time – intensity perceptions.

However, the most important new technology for precise sensory evaluation is descriptive analysis. For this a panel initially attempts to identify the most important sensory attributes and descriptive terms of a wine or group of wines, and trains themselves in doing so. Then under blind tasting conditions, wines are rated in terms of these attributes by this panel. After statistic analyses of the panelists' judgements have been made, profiles of the wine odors or flavors are developed which permit an evaluation of the differences among the wines [78].

[Top of Page]

10. Economic Aspects

Ralph E. Kunkee

In 1994, world wine production was 25.6 $\times 10^9$ L, with ca. 74 % (19×10^9 L), coming from Europe, of which only 4 % (10^9 L) was from Germany [79]. However, the consumption of wine in 1994 was: world 22.2×10^9 L, Europe 16.5×10^9 L, Germany 1.8×10^9 L. Thus for the world, and for Europe as a whole, consumption was substantially less than production. And this is not a new situation: European consumption in the early 1970 s was as much as 18.8×10^9 L, and has been decreasing ever since [22]. The only countries to show a substantial increase in consumption over the last several years are Germany, but that is from a very low base when some wine-contamination scandals cut consumption drastically, and the United Kingdom, but even that increase has leveled off in the 1990s [79]. Part of the drop in consumption is due to the displacement of wine by other beverages, especially soft drinks, but also coffee, fruit juices, and milk products [79]. Some of the problems in attempts to analyze or to offer solutions to the downward consumption have to do with the international complexity of the situation. For example, SPAHNI [80] discussed the multidimensional product, the multiregulated industry, the multinationals and the multilateral trade. Whatever the explanation, the consequences are serious. Vineyards are not quickly transformed from grape production to other crops, and thus the decreased consumption has led slowly to substantial inventories. Some reasonable suggestions involve changing strategies in marketing, in public relations, and reaching out to new markets [23]. However, the most consequential effect on consumption has come from an unexpected source, an American television news magazine, *60 Minutes*, on the subject of the "French Paradox", and the publication of the book of the same title [81]. The announcements of the benefits of ethanol in general, or wine in particular, and especially red wine, on human health (Chap. Wine and Health) had a dramatic effect on wine sales and consumption in the United States, at least in the short term.

[Top of Page]

11. Wine and Health (For toxicology of ethanol, see → Ethanol – Toxicology)

Ralph E. Kunkee

Moderate use of wine has long been touted as a health-care giver, long before St. Paul's dictum to Timothy [82]: "Drink no longer water, but use a little wine for thy stomach's sake" and up to its current use as a mild diuretic, tranquilizer, appetite stimulant, soporific and source of non-glucose calories [1]. Wine as a provider of trace elements has also long been realized (Section Essential Trace Elements in Wine). However, it has only been in the last several years the there has been substantial evidence for what many practitioners have long thought was true: that regular, but moderate, consumption of ethanol tends to reduce the incidence of coronary heart disease, apparently by raising the level of high-density lipoproteins in the blood and lowering the low-density lipoproteins (LDL) [83]. Studies have shown some of the mechanisms for these actions, at least from the moderate consumption of red wine [84]. Wine flavonoids from red wines are antioxidants and inhibit human LDL oxidation; resveratrol, another pigment from red wine, also inhibits human LDL oxidation.

[Top of Page]

References

General References

1. M. A. Amerine, V. L. Singleton: *Wine An Introduction*, 2nd ed., University of California Press, Berkeley 1977.
2. M. W. Baldy: *The University Wine Course*, The Wine Appreciation Guild, San Francisco 1993.
3. R. B. Boulton, V. S. Singleton, L. F. Bisson, R. E. Kunkee: *Principles and Practices of Winemaking*, Chapman & Hall, New York 1996.
4. O. Clarke: *Oz Clark's Encyclopedia of Wine*, Simon and Schuster, New York 1993.
5. G. Debuigne: *Larousse les Vins*, Librariri Larousse, Paris 1988.
6. G. Debuigne: *Larousse Dictionary of Wines of the World*, Larousse and Co., New York 1976.
7. Edita: *The Great Book of Wine*, World Publishing, New York 1970.

8. G. Ford: *Ford's Illustrated Guide to Wines, Brews, & Spirits*, Wm. C. Brown, Dubuque, Iowa, 1983.
9. P. G. Garoglio: *Enciclopedia Vitivinicola Mondiale*, Edizione Scientifiche UIV, Milan 1973 – 1976.
10. P. G. Garoglio: *Nuova Enologia*, Edizioni AEB, Brescia 1981.
11. H. J. Grossman: *Grossman's Guide to Wines, Spritis, and Beers*, H. Lembeck, rev., 7th ed., Charles Scribner's Son, New York 1983.
12. H. Johnson: *The World Atlas of Wine: A Complete Guide to the Wines and Spirits of the World*, enl. & rev., Simon and Schuster, New York 1985.
13. H. Johnson: *Hugh Johnson's Modern Encyclopedia of Wine*, rev. & updated, Simon and Schuster, New York 1987.
14. A. Lichine (ed.): *Alexis Lichine's New Encyclopedia of Wines & Spirits*, 5th ed., Alfred A. Knopt, New York 1987.
15. D. Muscatine, M. A. Amerine, B. Thompson (eds.): *The University of California/Sotheby Book of California Wine*, University of California Press/Sotheby Publications, Berkeley 1984.
16. J. Robinson (ed.): *The Oxford Companion to Wine*, Oxford University Press, Oxford 1994.
17. F. Schoonmaker: *Frank Schoonmaker's Encyclopedia of Wine*, 6th ed., Hastings House, New York 1978.
18. Sommelier Executive Council: *Vintage Wine Book*, 2nd ed., Food Products Press, Binghamton, NY, 1992.
19. W. Younger: *Gods, Men, and Wine*, The Wine and Food Society and the World Publishing Co., Cleveland, Ohio, 1966.

Specific References

20. L. Pasteur: *Études sur le Vin*, Victor Masson et Fils, Paris 1866.
21. K. Kodama: “Saké Yeasts,” in A. H. Rose, J. S. Harrison (eds): *The Yeasts*, 2nd ed., vol. **5**, Academic Press, London 1993, pp. 129 – 168.
22. R. E. Kunkee, L. F. Bisson: “Wine-making Yeasts,” in A. H. Rose, J. S. Harrison (eds): *The Yeasts*, 2nd ed., vol. **5**, Academic Press, London 1993, pp. 69 –127.
23. R. W. Goswell, R. E. Kunkee: “Fortified Wines,” in A. H. Rose(ed.): *Alcoholic Beverages, Economic Microbiology*, vol. **1**, Academic Press, London 1977, pp. 477 – 535.
24. A. C. Noble, G. F. Bursick: “The Contribution of Glycerol to Perceived Viscosity and Sweetness in White Wine,” *Am. J. Enol. Vitic.* **35** (1984) 110 – 112.
25. H. F. Linskens, J. F. Jackson: “Wine Analysis,” in H. F. Linskens, J. F. Jackson (eds.): *Wine Analysis, Modern Meth* Springer Verlag, Berlin 1988.
26. M. A. Amerine et al.: *The Technology of Wine Making*, 4th ed., Avi Publishing Co., Westport, Connecticut, 1980.
27. A. D. Webb, J. L. Ingraham: “Fusel Oil,” *Adv. Appl. Microbiol.* **5** (1963) 317 – 353. Links
28. R. M. Soles, C. S. Ough, R. E. Kunkee: “Ester Concentration Differences in Wine Fermented by Various Species and Strains of Yeasts,” *Am. J. Enol. Vitic.* **33** (1982) 94 – 98.
29. M. A. Amerine, C. S. Ough: *Methods for Analysis of Musts and Wines*, J. Wiley & Sons, New York 1980.
30. H. F. Linskens, J. F. Jackson (eds.): *Wine Analysis, Modern Methods of Plant Analysis*, **New Series, vol. 6**, Springer Verlag, Berlin 1988.
31. B. W. Zoecklein, K. C. Fugelsang, B. H. Gump, F. S. Nury: *Production Wine Analysis*, 2nd ed., Chapman & Hall, New York 1996.
32. R. E. Kunkee: “Malo-lactic Fermentation,” *Adv. Appl. Microbiol.* **5** (1967) 235 – 279.
33. A. Rapp: “Wine Aroma Substances from Gas Chromatographic Analysis,” in H. F. Linskens, J. F. Jackson (eds.): *Wine Analysis, Modern Methods of Plant Analysis*, **New Series, vol. 6**, Springer Verlag, Berlin 1988.
34. V. L. Singleton: “Tannins and the Qualities of Wines,” in R. W. Hemingway, P. E. Laks (eds.): *Plant Poloyphenols*, Plenum Press, New York 1992.
35. R. B. Boulton, personal communication.
36. M. A. Sefton, I. L. Francis, P. J. Williams: “The Volatile Composition of Chardonnay Juices: A Study by Flavor Precursor Analysis,” *Am. J. Enol. Vitic* **44** (1993) 359 – 370.
37. H. Eschnauer: *Spurenelemente in Wein und anderen Getränken*, Verlag Chemie, Weinheim, Germany 1974.
38. H. Eschnauer, *Naturwissenschaften* **73** (1986) 281 –290. Links
39. F. Alt, J. Messerschmidt, G. Tölg, H. R. Eschnauer, *Fresenius J. Anal. Chem.*, 1996.
40. W. Gätel, *Nachrichtenbl. Dsch. Pflanzenschutzdienstes (Braunschweig)* **36** (1984) 154 – 157.
41. H. R. Eschnauer, G. R. Scollary, *Wein-Wiss.* **50** (1995) 24 – 30.
42. H. D. Mohr: *Schwermetalle in Boden, Rebe und Wein*, Reihe A: Angewandte Wissenschaft, Heft 308, Landwirtschaftsverlag, Münster-Hiltrup 1985.
43. G. Schwedt, A. Schweizer, G. Hendrich, *Z. Lebensm. Unters. Forsch.* **187** (1988) 229 – 234.
44. R. Lobinski, J. Szpunar-Lobinska, F. C. Adams, P.-L. Teissedre, J.-C. Cabanis, *J. AOAC Int.* **76** (1993) 1262 – 1267. Links
45. K.-H. Bauer, R. Neeb, H. Eschnauer, *Fresenius Z. Anal. Chem.* **335** (1989) 808 – 812.
46. H. Kliewe: *Wein und Gesundheit*, Verlag D. Meininger, Neustadt/Weinstr. 1965.
47. H. Kliewe, H. Eschnauer, *Wein-Wiss.* **23** (1968) 30 –44.
48. W. Gärtel, *Weinberg Keller* **212** (1974) 435 – 508.
49. C. F. Pereira, *Z. Lebensm. Unters. Forsch.* **186** (1988) 295 – 300.
50. H. Eschnauer, L. Jakob, H. Meierer, R. Neeb, *Mikrochim. Acta* 1989, no. 3, 291 – 298. Links
51. M. P. Day, B. Zhang, G. J. Martin, *J. Sci. Food Agric.* **67** (1995) 113 – 123.
52. H. Förstel, *Naturwissenschaften* **72** (1985) 449 – 455.

53. P. Horn, P. Schaaf, B. Holbach, St. Hölzl, H. Eschnauer, *Z. Lebensm. Unters. Forsch.* **196** (1993) 407 –409.
54. B. L. Gulson, T. H. Lee, K. J. Mizon, M. J. Korsch, H. R. Eschnauer, *Am. J. Enol. Vitic.* **43** (1982) 180 –190.
55. H. R. Eschnauer, P. Ostapczuk, *Wein-Wiss.* **47** (1992) 206 – 209.
56. G. N. Chen, G. R. Scollary, V. A. Vicente-Beckett, *Am. J. Enol. Vitic.* **45** (1994) 305 – 311.
57. H. Eschnauer, R. Neeb: "Wine Analysis" in H. F. Linskens, J. F. Jackson (eds.): *Micro-Element Analysis in Wine and Grapes in Modern Methods of Plant Analysis*, **New Series, vol. 6**, Springer Verlag, Berlin 1988, pp. 67 – 91.
58. M. G. Mullins, A. Bouquet, L. E. Williams: *Biology of the Grapevine*, Cambridge University Press, Cambridge 1992.
59. H. J. De Blij: *Wine A Geographic Appreciation*, Rowman & Allanheld, Totowa, N. J., 1983.
60. M. A. Amerine, M. A. Joslyn: *Table Wines, The Technology of Their Production*, 2nd ed., University of California Press, Berkeley 1970.
61. P. Galet: *Précis d'Ampélographie Pratique*, 4th ed., Imprimerie Paul Déhan, Montpellier 1976.
62. L. T. Morton: *A Practical Ampelography*, Cornell University Press, Ithaca, N. Y., 1979.
63. M. R. Thomas, P. Cain, N. S. Scott: "DNA Typing of Grapevines: A Universal Methodology and Database for Describing Cultivars and Evaluating Genetic Relatedness," *Plant Mol. Biol.* **25** (1994) 939 – 949. Links
64. C. P. Meredith, personal communication.
65. J. Wolpert et al.: "Use of Phylloxera-Resistant Rootstocks in California: Past, Present and Future," *Grape Grower* **1994,** Jan., 16 – 19;Feb., 4 – 7; Mar., 10 –17.
66. R. Smart, M. Robinson: *Sunlight into Wine, A Handbook for Winegrape Canopy Management*, Winetitles, Underdale SA, Australia 1991.
67. D. L. Elliott-Fisk, A. C. Noble: "The Diversity of Soils and Environments in Napa Valley, California and Their Influence on Cabernet Sauvignon Wine Flavors" in H. J. de Blij (ed.): *Viticulture in Geographic Perspective,* Proc. 1991 Miami Symposium, Am. Assoc. Geol. 1992, pp. 45 – 71.
68. A. J. Winkler, J. A. Cook, W. M. Kliewer, L. A. Lider: *General Viticulture*, 2nd ed., University of California Press, Berkeley 1974.
69. N. J. W. Kreger-van Rij: *The Yeasts, A Taxonomic Study*, Elsevier Science Publishers, Amsterdam 1984.
70. A. Vaughn-Martini, C. P. Kurtzman: "Deoxyribonucleic Acid Relatedness among Species of the Genus Saccharomyces sensu stricto," *Int. J. Syst. Bacteriol.* **35** (1985) 508 – 511.
71. A. Martini: "Origin and Domestication of the Wine Yeast Saccharomyces cerevisiae," *J. Wine Res.* **4** (1993) 165 – 176.
72. R. S. Jones, C. S. Ough: "Variations in the Percent Ethanol V/V per °Brix Conversions of Wines from Different Climatic Regions," *Am. J. Enol. Vitic.* **36** (1985) 268 – 270.
73. R. B. Boulton: "Prediction of Fermentation Behavior by a Kinetic Model," *Am. J. Enol. Vitic.* **31** (1980) 40 – 45.
74. V. L. Singleton: "Oxygen with Phenols and Related Reactions in Musts, Wines and Model Systems: Observations and Practical Implications," *Am. J. Enol. Vitic.* **38** (1987) 66 – 77.
75. R. E. Kunkee: "Some Roles of Malic Acid in the Malolactic Fermentation in Wine Making," *FEMS Microbiol. Rev.* **88** (1991) 55 – 72. Links
76. C. Diviès, P. Lenzi, J. Beaujeu, F. Herault, EP-A 88–9249, 1988.
77. S. Hock: "Coping with Brettanomyces," *Pract. Wine Vine.* Jan/Feb (1990) 26 – 31.
78. A. C. Noble: "Analysis of Wine Sensory Properties" in H. F. Linskens, J. F. Jackson (eds.): *Wine Analysis,* Modern Methods of Plant Analysis, New Series, vol. 6, Springer Verlag, Berlin 1988, pp. 9 – 28.
79. R. Tinlot, M. Rousseau: "Situation et Statistiques du Secteur Vitivinicole Mondial en 1994," *Bull. O. I. V.* **68 (777 – 778)** (1995) 969 – 1056.
80. P. Spahni: *The International Wine Trade*, Woodhead Publishing Ltd., Cambridge 1995.
81. L. Perdue: *The French Paradox and Beyond*, Renaissance Publishing, Sonoma, CA, 1992.
82. Holy Bible, King James' Version: I Tim. 5 : 23.
83. A. Klatsky: "Alcohol, Coronary Disease, and Hypertension," *Ann. Rev. Med.* **47** (1996) 149 – 160.
84. A. Waterhouse: "Wine and Heart Disease," *Chem. Ind.* **9** (1995) 338 – 341. Links

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Wood

Horst H. Nimz, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Holzchemie und chemische Technologie des Holzes, Hamburg, Federal Republic of Germany (Sections 1.1 and 1.3)

Uwe Schmitt, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Holzbiologie und Holzschutz, Hamburg, Federal Republic of Germany (Section 1.2)

Eckart Schwab, Bundesforschungsanstalt für Forst- und Holzwirtschaft, Institut für Holzphysik und mechanische Technologie des Holzes, Hamburg, Federal Republic of Germany (Section 1.4)

Otto Wittmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2)

Franz Wolf, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chap. 2)

1. Morphology and Properties 3
1.1. Introduction 3
1.2. Biology of Wood 3
1.2.1. Anatomy 5
1.2.1.1. Cell Types 6
1.2.1.2. Softwoods 6
1.2.1.3. Hardwoods 7
1.2.2. Secondary Changes 8
1.2.3. Cell Wall Formation and Architecture 9
1.3. Chemistry of Wood 10
1.3.1. Chemical Components of Wood . . 10
1.3.2. Distribution of Components in the Cell Wall 13
1.3.3. Macromolecular Structure of Cell Wall Components 14
1.4. Physical Properties 15
2. Wood-Based Materials 16
2.1. Introduction 16
2.2. Laminate Bonding 17
2.2.1. Laminated Structural Timber (Glulam) 17
2.2.1.1. Development 17
2.2.1.2. Construction and Production 18
2.2.1.3. Properties 19
2.2.2. Veneers, Plywood, and Derived Products 20
2.2.2.1. Development 20
2.2.2.2. Production of Veneers 20
2.2.2.3. Principal Types of Plywood 20
2.2.2.4. Adhesives and Additives 22
2.2.2.5. Production of Veneer Plywood . . . 22
2.2.2.6. Properties, Testing, and Uses 23
2.2.2.7. Engineered Wood 24
2.3. Particle Bonding 25
2.3.1. Particle Board 25
2.3.1.1. Development 25
2.3.1.2. Types of Particle Board 26
2.3.1.3. Raw Materials 26
2.3.1.4. Adhesives and Additives 27
2.3.1.5. Production 28
2.3.1.6. Properties, Testing, and Uses 30
2.3.2. Fiberboard 33
2.3.2.1. Development 33
2.3.2.2. Types of Fiberboard 33
2.3.2.3. Raw Materials 34
2.3.2.4. Binders and Additives 35
2.3.2.5. Production Processes 35
2.3.2.6. Properties, Testing, and Uses 39
2.3.3. Wood-Based Materials with Mineral Binders 41
2.3.3.1. Cement-Bonded Boards 41
2.3.3.2. Gypsum-Bonded Board 41
2.3.3.3. Magnesia-Bonded Boards 42
2.4. Surface Treatment 42
2.4.1. Veneers 43
2.4.2. Liquid Paints and Coatings 43
2.4.3. Coating with Sheets, Short-Cycle Films, and Decorative Laminates . . 43
2.4.3.1. Coating with Foils 43
2.4.3.2. Decorative Films (Short-Cycle Films) 44
2.4.3.3. Decorative Laminates 45
2.4.4. Properties, Testing, and Uses 47
2.5. Environmental and Toxicological Aspects 47
2.5.1. Wood Dust 47
2.5.2. Emissions 48
2.5.2.1. Emissions during Production of Wood-Based Materials 48
2.5.2.2. Emissions from Finished Materials 49
2.5.3. Disposal 50
2.6. Polymer Wood 51
2.7. Economic Aspects 52
3. References 52

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_305

Abbreviations:

CPL continuous pressed laminates
DP degree of polymerization
DS defibrator-seconds
F formaldehyde
G guaiacylpropane
GDP guanosine diphosphate
H hydroxyphenylpropane
HDF high-density fiberboard
HPL high-pressure laminate
LSL laminated strand lumber
LVL laminated veneer lumber
M melamine
MDF medium-density fiberboard
MDI methylene diphenylene isocyanate
MF melamine – formaldehyde
ML middle lamella
MUF melamine – urea – formaldehyde
MUPF melamine – urea – phenol – formaldehyde
OSB oriented structural board
P primary wall; phenol
PF phenol – formaldehyde
PMDI polymeric methylene diphenylene isocyanate
PRF phenol – resorcinol – formaldehyde
PSL parallel strand lumber
S syringyl
S1S smooth one side
S2S smooth two sides
U urea
UF urea – formaldehyde

1. Morphology and Properties

1.1. Introduction

Not only is wood the most important renewable natural resource in terms of quantity, but because of its composition and versatile properties, its use as a raw material for pulp, structural timber, sawn wood, panels, furniture and other purposes is also economically important. In industrialized countries, wood plays only a minor role as an energy carrier [19]. The total phytomass of the earth is estimated to be 1.24×10^{12} t, of which 80 % is attributed to wood [20]. The potentially utilizable annual wood growth is 1.1×10^{10} t [21].

Data from the Food and Agriculture Organization (FAO) of the UN on forested areas and roundwood removals worldwide are given in Table 1.

The forested area of Germany, for example, covers 10.8×10^6 ha. This is 30.4 % of the total area of Germany, of which 55.1 % is agricultural. Of the forested area, 32.8 % is occupied by spruce, 27.5 % by pine, 14 % by beech, and 8.5 % by oak [23]. According to these statistics, Germany is among the more highly forested countries of the world, but with regard to the forested area per capita, it is one of the less forested, with 0.14 ha per inhabitant.

Finland, in comparison, with 33.8×10^6 ha total land area and 20×10^6 ha forests, has about the same size but nearly twice as much forests as Germany. It has 4 ha forested area per inhabitant, consisting mainly of spruce and pine (40 % each) and birch (16 %). Similarly Sweden, with a total land area of 45×10^6 ha, of which 58 % is forested, has 3.1 ha forests per inhabitant. In the southern part of Sweden spruce is predominant (51 %), followed by pine (29 %). This ratio is reversed in Northern Sweden with 48 % pine and 39 % spruce.

A quite different distribution of wood species is found in southern and tropical countries, where deciduous woods (hardwoods) often outweigh coniferous woods (softwoods).

Germany is a typical wood-importing country. In 1992, 80.3×10^6 m^3 of wood was imported and 27.7×10^6 m^3 of indigenous wood was cut. These figures take into account the different units used to quantify wood products, e.g., tonnes of pulp are expressed as raw wood equivalents in cubic meters (Table 2).

Finland, in contrast, as a typical wood exporting country, in 1992 had 55×10^6 m^3 roundwood removal from indigenous forests (cut wood) and additional 6×10^6 m^3 from Russia. Out of these 20×10^6 m^3 went into the domestic timber industry and 28×10^6 m^3 (plus 6×10^6 m^3 industrial wood residues) into the pulping industry. Finland exports some 1.5×10^6 t pulp and 6×10^6 t paper and cardboard, mainly to Western Europe, covering ca. 32 % of its national exports by wood products.

Important importing and exporting countries of wood, semifinished wood products, pulp, and paper are listed in Table 3.

Table 1. Forests and roundwood removals worldwide according to [22]

Area	Forested area (1990)			Roundwood removals (1991), 10^6 m^3				
	10^6 ha	% of land area	ha/inhabitant	Total	Coniferous wood	Deciduous wood	Wood for processing	Wood for combustion
North/Central America	530 744	25	1.24	737	494	243	593	144
South America	898 184	51	3.03	345	61	285	103	242
Africa	535 848	19	0.84	527	16	511	59	468
Europe	140 196	30	0.28	335	224	112	284	51
Former Soviet Union	754 958	34	2.61	355	291	65	274	81
Asia	463 221	18	0.15	1086	189	897	252	834
Pacific Rim	87 700	10	3.29	42	21	22	34	9
Total	3 410 851	26	0.64	3429	1295	2134	1599	1830

Table 2. Total wood balance (wood and wood products) in Germany in 10^6 m^3, according to [24]

Supply	1991	1992	Demands	1991	1992
Wood removal	32.1	27.7	Increase in stored wood	0	0
Waste paper (inland)	22.0	22.9	Export	50.9	51.0
Waste wood	0.3	0.4	Consumption	87.2	88.3
Import	77.3	80.3			
Decrease in stored wood	6.3	8.0			
Total supply	138.0	139.3	Total demands	138.1	139.3

Table 3. Economics of forest products in the most important countries in 1993, according to [25]

	Production, 10^6 \$	% of GDP ∗	Imports, 10^6 \$	Exports, 10^6 \$	% of trade	Consumption, 10^6 \$
World	391 327	2	106 742	99 618	3	398 450
United States	93 189	2	16 873	13 401	3	96 661
China	34 659	2	4 648	1 121	1	38 186
Canada	30 655	6	2 082	19 295	13	13 442
Japan	21 828	1	16 767	1 648	0	36 912
Former Soviet Union	18 443		124	2 061		16 506
Brazil	17 437	5	308	1 995	5	15 751
India	16 543	4	262	17	0	16 788
Germany	14 481	1	9 502	5 751	2	18 232
Indonesia	13 519	6	556	5 158	14	8 917
Sweden	9 792	5	845	7 483	15	3 154
Finland	9 230	10	475	7 411	32	2 294
France	8 415	1	4 962	3 997	2	9 381
Nigeria	5 281	4	77	20	0	5 337
Malaysia	5 126	7	411	4 190	9	1 347
Italy	4 491	1	5 545	2 008	1	8 028
Korea, Rep.	4 097	2	3 126	570	1	6 652
United Kingdom	4 055	0	8 192	1 932	1	10 315
Spain	3 658	1	2 479	1 032	2	5 106

∗ GDP = Gross domestic product.

The versatile technological properties of wood result from its anatomy and the chemical properties of its three main components, cellulose, lignin, and hemicelluloses, which, in turn, are affected in different ways by wood extractives occurring as minor components.

1.2. Biology of Wood

All woody plants reproduce by seeds and are part of the Spermatophyta. This systematic division is split into the Gymnospermae, to which the softwoods belong, and the Angiospermae, to which the hardwoods belong. The Monocotyledoneae are part of the Angiospermae; some of them (e.g., bamboo and rattan) develop woody

tissue but are for various reasons maintained as a separate entity.

Wood, except for the Monocotyledoneae, is defined as the inner tissue of stems (see Fig. 1), branches, and roots of perennial plants. It acts as a transport system for water and mineral solutes between the root tips and the parts above ground, serves as storage tissue for reserve material, and provides the plant with mechanical stability. Wood is surrounded by the *cambium* (Fig. 2), which consists of a few layers of meristematic cells, and the bark, which functions both as a transport system for assimilates and as protection against environmental influences. The cambial cells undergo frequent division during the vegetation period, producing wood cell initials in the adaxial direction and bark cells in the abaxial direction. With time, the wood cell initials differentiate and form the taxon-specific arrangement of functionally and morphologically different cell types.

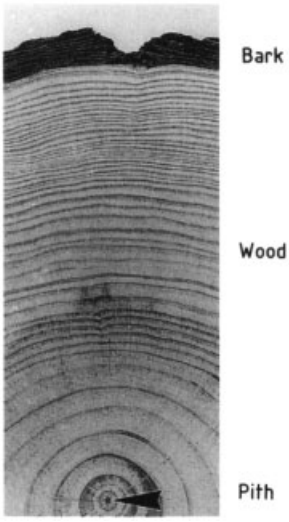

Figure 1. Macroscopic appearance of a *Pinus sylvestris* wood segment with wood and bark

1.2.1. Anatomy [26]

The histological term for wood as the tissue inside the cambium is *xylem*. Its formation is restricted to the annual vegetation period and therefore represents a periodic process. In the northern temperate zone, typical growth rings are laid down with subdivision into *earlywood* and *latewood* that determine the beginning and end of the growing season. A growth ring is strictly correlated with one year, whereas in the tropics and subtropics the growth increments may reflect alternating wet and dry seasons, not necessarily an annual cycle.

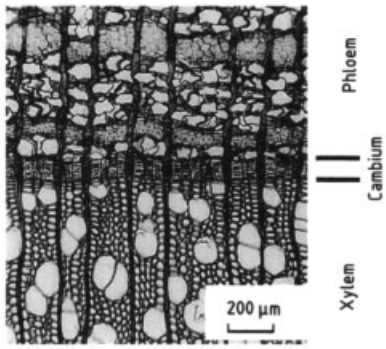

Figure 2. *Populus robusta*: Cross-section through the cambial area, adaxial xylem, and abaxial phloem
Light micrograph (LM)

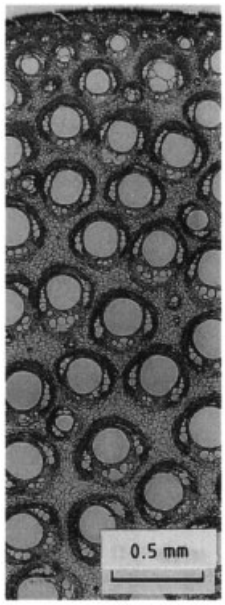

Figure 3. *Calamus caesius*: Cross-section through a rattan stem (LM)

Monocotyledoneae do not possess a cambium and therefore lack the secondary thickening characteristic of gymnosperms and most

angiosperms. Their woody tissue consists of numerous axially oriented vascular bundles, which are embedded in a parenchymatous ground tissue (Fig. 3). Within the vascular bundles, different cell types are associated with the main functions of water and assimilate transport, mechanical support, and to a certain extent, storage of reserve material. In the axial direction, the stem is subdivided into long internodes and short nodes.

The following sections describe wood formed by gymnosperms and angiosperms.

1.2.1.1. Cell Types

Wood in its unique structure is composed of vast numbers of cells. According to their morphology and function, four groups of cell types exist: parenchyma cells, tracheids, fibers, and vessels. *Parenchyma cells* in living wood contain cytoplasm (Fig. 4) and are involved in various processes such as metabolic pathways, short-distance transport, and storage; they form horizontal rays as well as axially oriented strands. Axial parenchyma is much more frequent in hardwoods, where it tends to form distinctive patterns, particularly in tropical timber species. *Tracheids* are the predominant axial cell element in softwoods that provide both water conduction and mechanical support. Tracheids also occur in some hardwoods. They are located adjacent to the vessels (vasicentric tracheids), or they resemble vessels (vascular tracheids) with the sole function of water transport. *Fibers* occur exclusively in hardwoods and represent specialized elements for mechanical support. Slight differences in morphology have given rise to a differentiation between "fiber-tracheids" and "libriform fibers." However, since continuous intergrading exists between them, the general term fiber is commonly used. *Vessels* are specialized elements for long-distance transport of water and mineral solutes in hardwoods; they are composed of individual vessel members that are associated to form a continuous channel with a length of 50 – 60 cm in some genera (*Populus, Fagus*) or greater (*Quercus*). During maturation of the vessels the end walls of the individual members abutting on other members above and below are removed partly or entirely by enzymatic action; this process results in a rather different appearance of the so-called vessel perforation plates.

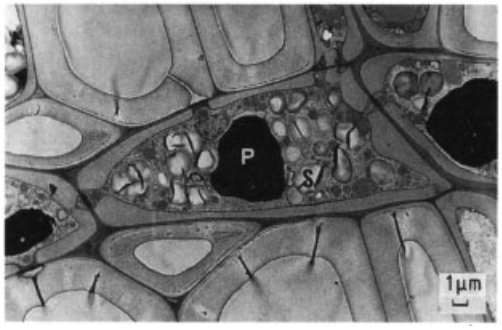

Figure 4. *Betula pendula*: Ray parenchyma cell with numerous starch grains (S); P = Phenolic body
Electron micrograph (EM)

Tracheids and vessels reach functionality only after the death of their cytoplasm. Fibers have commonly been considered dead cells too, but in some species, a few cells retain their cytoplasm over a long period after completion of wall thickening, in the extreme, up to heartwood formation. These fibers are termed living fibers and may store starch.

1.2.1.2. Softwoods

Softwoods (coniferous woods) exhibit a simpler structure than hardwoods (deciduous woods). They are composed mainly of tracheids as the principal cell type in the axial direction (Fig. 5 A). Tracheids of softwoods are extremely elongated elements with a length of 2.0 – 5.0 mm. Cell size, wall thickness, and lumen diameter vary greatly, with thin walls and wide lumina in earlywood tracheids but extreme wall thickening and narrow lumina in the latewood tracheids. The latter contribute mainly to mechanical stability, whereas the earlywood tracheids are conductive rather than supporting elements. Tracheids may also be horizontally oriented, forming lower and upper margins of the rays in some genera of Pinaceae. In some softwoods, axial parenchyma is absent (e.g., *Picea, Pinus*) or sparse (e.g., *Larix)*; in others it is of regular occurrence (e.g., *Podocarpus, Juniperus*). Axial and radial resin ducts are characteristic features of some genera of Pinaceae (e.g., *Picea, Larix, Pinus, Pseudotsuga*) and are lined

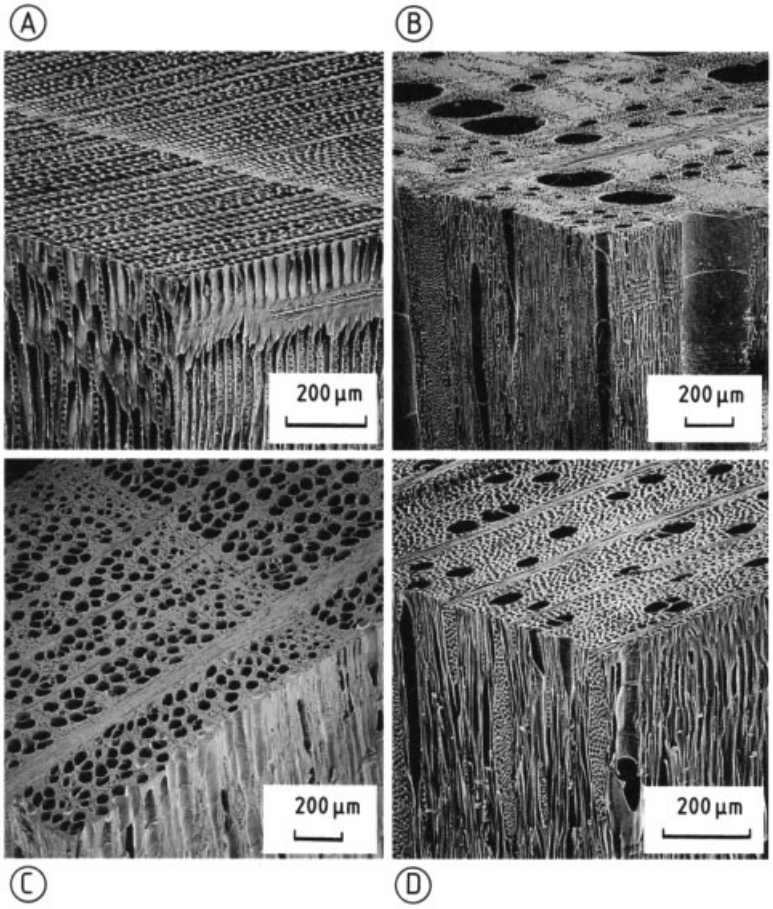

Figure 5. A) *Cedrus libani*: Arrangement of tracheids and parenchymal cells in softwoods (EM); B) *Quercus cerris*: Ring-porous hardwood showing large-diameter earlywood pores and narrow latewood pores (EM); C) *Fagus sylvatica*: Semi-ring-porous hardwood (EM); D) *Acer platanoides*: Diffuse-porous hardwood (EM)

by parenchymatous epithelial cells responsible for the production and posterior secretion of resin into the ducts.

1.2.1.3. Hardwoods

The evolutionary advanced hardwoods possess a functionally and morphologically more diversified structure. The most distinctive feature is the occurrence of vessels, alternatively called "pores" when observed on transverse surfaces. Vessels vary in size and distribution, especially within a discernible growth increment. This diversity is called porosity and represents a very important feature for the identification of hardwoods attributed to three groups: *ring-porous*, *semi-ring-porou,*, and *diffuse-porous*. *Ring-porous woods* (Fig. 5 B) show distinct differences in the size of vessels within one growth increment; earlywood vessels are larger in diameter than latewood vessels, with an abrupt transition between both. Hence, in some cases the earlywood portion becomes clearly visible as a pore-ring even with the unaided eye. *Semi-ring-porous woods* (Fig. 5 C) are characterized by a gradual change from large-diameter earlywood vessels to small latewood vessels. *Diffuse-porous woods* (Fig. 5 D) display vessels of nearly the same diameter throughout one growth increment. The variation within a species, however, may range sometimes from ring porous to semi-ring porous or from semi-ring porous to diffuse porous. Vessels may be arranged in tangential bands, may exhibit a diag-

onal or radial pattern, or may reveal a branched appearance to form a dendritic pattern. Vessels may also occur solitarily, as groups in radial multiples of four or more, or as irregular clusters.

Axial parenchyma cells are much more abundant in hardwoods than in softwoods (Table 4). The amount and various distributional patterns display a marked diversity. The position of axial parenchyma cells in relation to vessels is used for a general subdivision into two types of parenchyma. The term *apotracheal parenchyma* reflects the isolation of vessels against axial parenchyma by other cells, whereas *paratracheal parenchyma* describes the association of axial parenchyma strands with vessels. Rays in hardwoods may be large or small in width and height, or both types may occur within one species.

Table 4. Percentage of cell-type volume in some wood species

Species	Vessels	Fibers	Ray parenchyma	Axial parenchyma
Douglas fir (*Pseudotsuga menziesii*)		93	7	
European spruce (*Picea abies*)		95	5	
Pine (*Pinus sylvestris*)		93	6	
Aspen (*Populus tremula*)	26	61	13	
Birch (*Betula alba*)	25	65	8	2
Oak (*Quercus robur*)	8 – 40	44 – 58	16 – 29	sparse – 5
Ash (*Fraxinus excelsior*)	12	62	15	11
Beech (*Fagus sylvatica*)	39	40	16	5
Framiré (*Terminalia ivorensis*)	20	53	17	10
Kosipo (*Entandrophragma candollei*)	12	45	25	18
Mahogany (African) (*Khaya ivorensis*)	18	57	21	4
Dark red meranti (*Shorea macrophylla*)	29	41	12	17
Teak (*Tectona grandis*)	12	66	16	6

1.2.2. Secondary Changes

With increasing age and secondary thickening of a woody plant the more centrally located xylem of many species undergoes natural changes. Consequently, the xylem must be subdivided into two parts. In a living tree the xylem portion close to the cambium serves as a pathway mainly for water and mineral solutes, but it also has a storage function; its parenchyma cells are physiologically active and involved in various metabolic processes. The more centrally located wood portion in mature trees, however, does not contain living parenchyma. In general, the terms *sapwood* for the outer, and *heartwood* for the central portions are used. Transition from sapwood to heartwood usually occurs within a narrow zone of a few cell rows, resulting in a slightly undulating sapwood – heartwood boundary. In most cases, the heartwood has a significantly darker color than the surrounding sapwood (Fig. 6), with certain variation of the width of the sapwood between and within species. Besides the obligatory formation of such a distinctly colored heartwood, some tree species develop less colored or nearly colorless heartwood; all of those species are therefore grouped into regular heartwoods.

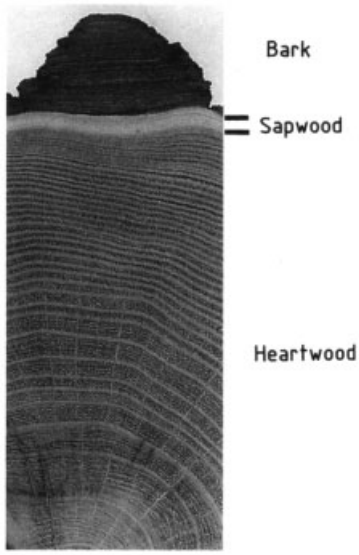

Figure 6. *Robinia pseudoacacia*: Division of the wood into light sapwood and dark heartwood

Other tree species such as *Betula* or *Fraxinus* are not able to form heartwood under unaffected growth conditions. However, all woody plants develop *discolored wood* as a result of environmental influences (e.g., injury, attack by microorganisms, severe frost). The boundary between unaffected and discolored wood then often moves across a larger number of growth increments with an irregular periphery. Discolorations may also be induced in regular heartwood as a secondary process [27]. The misleading terms "false heartwood," "pathological heartwood," and "frost heartwood" are often used for these environmentally initiated changes. The main criteria for the differentiation between heartwood and discolored wood are as follows: heartwood is characterized by increased natural durability and lower moisture content in comparison to sapwood; in discolored wood, however, the durability does not increase, and the moisture content may rise even above the sapwood level.

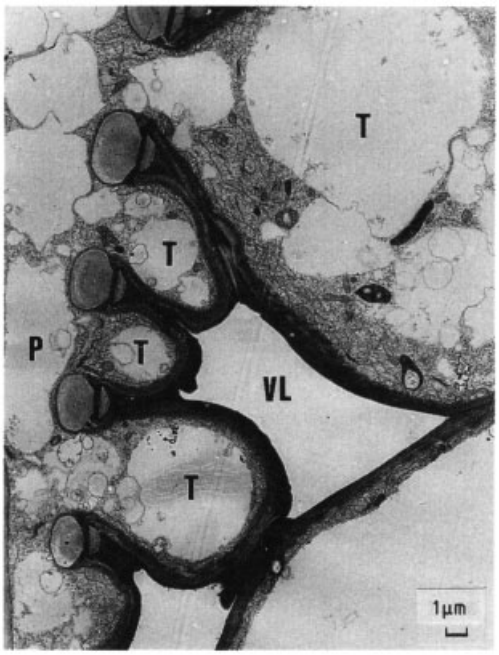

Figure 7. *Quercus robur*: Tylosis development – tylosis initials protrude from a parenchyma cell into a vessel (EM)
T = Tylosis; P = Parenchyma cell; VL = Vessel lumen

The change from sapwood to heartwood during natural heartwood formation in hardwoods is frequently accompanied by the occlusion of vessels. The lumina of some species are more or less filled with often dark-staining material secreted from adjacent parenchyma cells. Other species develop *tyloses*, representing balloon-like protrusions from parenchyma cells through the pits into adjacent vessels (Fig. 7).

1.2.3. Cell Wall Formation and Architecture

Walls typically envelop the cytoplasm of all plant cells to impart protection and firmness. Stabilization of a multicellular plant body is sustained by the continuum of adhering walls of all cells. The cell walls consist primarily of an amorphous polysaccharide matrix of pectins and hemicelluloses as well as crystalline cellulose – the basic structural polysaccharide, which is aggregated in the form of microfibrils. Non-polysaccharide components, of which the most important is lignin, are finally incorporated. This combination of different chemical constituents determines the physical properties of cell walls. Details on the chemical composition of cell walls are given in Section 1.3.

Cell wall formation starts with the division of a cambial initial. A daughter cell becomes separated first at the equatorial plane by a thin, tangential wall, which consists mainly of pectic material. This initial layer, common to both adjacent cells, is termed the *middle lamella* (ML) and functions as a cementing substance between cells. At this early developmental stage the young wood cell begins to deposit new wall material against the ML. In contrast to the ML, this new layer contains cellulosic material embedded in a pectin- and hemicullose-containing matrix. This layer represents the first-formed cell wall layer and is called the *primary wall* (P). Here, the cellulose microfibrils show a dispersed texture. Both middle lamella and primary wall are extensible to allow cell expansion during differentiation. Cellulose fibrils and matrix material are continuously incorporated into the growing wall. As the developing wood cell approaches final size, deposition of the stiff secondary wall occurs. Secondary walls in most cases show a subdivision into three layers (S 1, S 2, S 3), with the narrow S 1 as the oldest layer next to the primary wall, the S 2 layer mostly representing the thickest portion, and the sometimes barely

discernible or thin S 3 layer forming the interface with the cytoplasm in living cells or the cell lumen in dead cells. The main constituents of the developing secondary wall are hemicelluloses and, in a distinctly higher proportion than in the primary wall, cellulose. The microfibrils of all secondary wall layers are laid down in parallel arrays: in the S 1, the mean angle with the longitudinal cell axis is 50 – 75°; microfibrils of the S 2 appear more closely packed, with an orientation of 10 – 30° to the cell axis; in the S 3 layer, microfibrils exhibit a mean angle of 60 – 80° with the cell axis (Fig. 8). Sculpturing on the lumen surface of the S 3 layer is sometimes observed (e.g., the "warty layer" in some softwoods and hardwoods). Incrustation with lignin completes wall development.

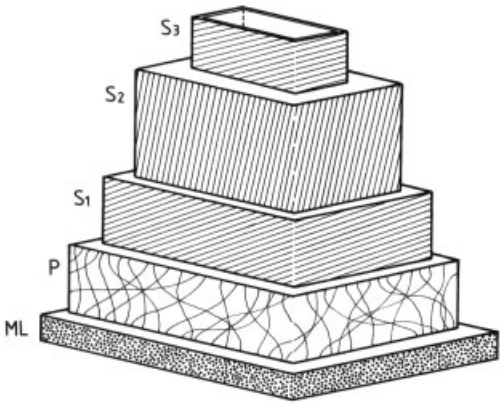

Figure 8. Schematic representation of wall architecture

The formation of microfibrils during wall thickening takes place at the interface between plasma membrane and cell wall. Fine structural investigations on differentiating plant cells performed in the 1970s and 1980s revealed specific structures within the bilayered plasma membrane: terminal globules in the outer leaflet [28] and rosettes in the inner leaflet [29]. Both terminal globules and rosettes are probably part of the cellulose-synthesizing apparatus. The position of the terminal globules at ending imprints of microfibrils in the membrane implies that they may represent sites of cellulose synthases. The membrane rosettes, in most cases a complex of six subunits, are also believed to be structures with cellulose synthase activity; their frequent occurrence in certain phases and sites of wall synthesis may reflect involvement in microfibril synthesis. Future research must clarify the details of this putative structure – function relationship. Besides the obviously membrane-bound synthesis of cellulose microfibrils, their orientation also appears dependent on membrane-associated structures, the membrane-underlying microtubules. These cortical microtubules may act as guidelines for the synthase complexes within the plasma membrane during microfibril elongation.

At the onset of secondary wall thickening, the deposition of wall material is interdicted along some primary wall areas. These develop into *pits* that represent specific pathways for the conduction of water and mineral solutes. Pits therefore appear as well-defined areas connecting neighboring cells through a thin membrane composed of a middle lamella and a primary wall layer on each side. With rare exceptions, a conduit between two cells consists of a pair of pits. *Bordered pits* show the secondary wall overarching the membrane on both sides, thus reducing the pit aperture. This is true both for pits between the tracheids in softwoods (Fig. 9 A) and for those connecting hardwood tracheids, fibers, and vessels with each other. *Half-bordered pits* generally occur between those cell types and a parenchyma cell with the unbordered part at the parenchyma side (Fig. 9 B). *Simple pits* are without border and connect two adjacent parenchyma cells (Fig. 9 C). Direct cytoplasm contact between parenchyma cells is maintained by *plasmodesmata*, small strands of cytoplasm that extend from cell to cell through perforations in the pit membrane.

1.3. Chemistry of Wood

1.3.1. Chemical Components of Wood

The main components of the wood cell wall (wood substance) are cellulose (→ Cellulose), lignin (→ Lignin), and hemicelluloses. The proportion of cellulose in the most common European woods is 41 – 43 %, whereas the lignin content in coniferous woods (27 – 28 %) is higher than that in deciduous woods (22 – 24 %). Correspondingly, hemicelluloses occur in higher concentrations in deciduous woods (30 – 35 %)

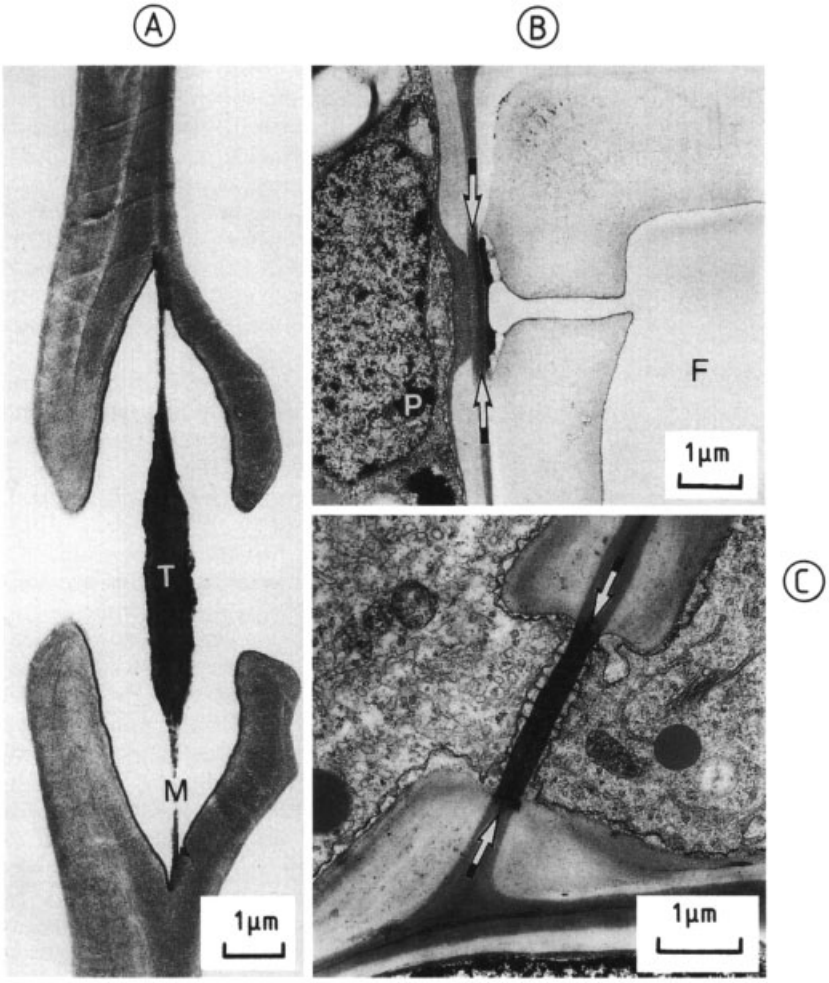

Figure 9. Electron micrographs of different types of pits
A) *Picea abies*: Bordered pit between tracheids; the pit membrane is divided into the central torus (T) and the peripheral margo (M); B) *Betula pendula*: Half-bordered pit between a parenchyma cell (P) and a fiber (F); arrows indicate the pit membrane; C) *Quercus robur*: Simple pit between parenchyma cells; arrows indicate the pit membrane.

than in coniferous woods (25 – 30 %). Compression wood contains more lignin (ca. 40 %), and tension wood (e.g., poplar) more cellulose (> 50 %). Differences are also found in tropical or subtropical types of wood. Within one wood species, age, location, and growth conditions can also play a role.

Most European woods also contain 1 – 3 % extractives and 0.1 – 0.5 % mineral components, which can be determined as ash. Higher contents of extractives have been found in certain types of wood (up to 10 % resins in pines, and tannins in quebracho and eucalyptus); in others, larger quantities of inorganic salts are found. The chemical compositions of some important types of wood are given in Table 5. Due to the lack of analytical methods that allow a clean separation of the wood components, the percentages given in Table 5 suffer from different isolation conditions used by different authors.

The *extractives* consist of a large number of low molecular mass compounds, which can be extracted from wood with organic solvents (terpenes, fats, waxes, and phenols) or hot water (tannins and inorganic salts). Depending on the quantity and the class of compound, which are specific to the tree type and genus to some extent, the extractives affect the chemical, bio-

Table 5. Chemical comparison of various wood species (% of dry wood weight) [30]

Species	Common name	Extractives *	Lignin	Cellulose	Glucomannan	Glucuronoxylan	Other polysaccharides
Softwoods							
Abies balsamea	balsam fir	2.7	29.1	38.8	17.4	8.4	2.7
Pseudotsuga menziesii	Douglas fir	5.3	29.3	38.8	17.5	5.4	3.4
Tsuga canadensis	eastern hemlock	3.4	30.5	37.7	18.5	6.5	2.9
Juniperus communis	common juniper	3.2	32.1	33.0	16.4	10.7	3.2
Pinus radiata	Monterey pine	1.8	27.2	37.4	20.4	8.5	4.3
Pinus sylvestris	Scots pine	3.5	27.7	40.0	16.0	8.9	3.6
Picea abies	Norway spruce	1.7	27.4	41.7	16.3	8.6	3.4
Picea glauca	white spruce	2.1	27.5	39.5	17.2	10.4	3.0
Larix sibirica	Siberian larch	1.8	26.8	41.4	14.1	6.8	8.7
Hardwoods							
Acer rubrum	red maple	3.2	25.4	42.0	3.1	22.1	3.7
Acer saccharum	sugar maple	2.5	25.2	40.7	3.7	23.6	3.5
Fagus sylvatica	common beech	1.2	24.8	39.4	1.3	27.8	4.2
Betula verrucosa	silver birch	3.2	22.0	41.0	2.3	27.5	2.6
Betula papyriferea	paper birch	2.6	21.4	39.4	1.4	29.7	3.4
Alnus incana	gray alder	4.6	24.8	38.3	2.8	25.8	2.3
Eucalyptus camaldulensis	river red gum	2.8	31.3	45.0	3.1	14.1	2.0
Eucalyptus globulus	blue gum	1.3	21.9	51.3	1.4	19.9	3.9
Gmelina arborea	yemane	4.6	26.1	47.3	3.2	15.4	2.5
Acacia mollissima	black wattle	1.8	20.8	42.9	2.6	28.2	2.8
Ochroma lagopus	balsa	2.0	21.5	47.7	3.0	21.7	2.9

* Extraction with CH_2Cl_2, followed by C_2H_5OH.

logical, physical, and optical properties of the wood to varying degrees. As the swelling and biodegradability may be significantly reduced by lipophilic resins they improve the dimensional stability and durability of wood, while other physical and chemical properties, e.g., during mechanical and chemical pulping, sawing, or painting, are negatively affected.

Coniferous woods contain mainly *terpenes*, which can be obtained from pines by resinification of the living tree (→ Resins, Natural, Chap. 5.2.1.) or as turpentine (monoterpenes) (→ Turpentines) and tall oil (diterpenes, resin acids) from kraft (sulfate) pulping (→ Tall Oil). *Long-chain fatty acids* are also present in tall oil and arise from fats and waxes in the wood. In other types of wood (oak, chestnut, eucalyptus, and quebracho), *tannins* and *flavonoids* predominate. *Phenolic compounds* (lignans and stilbenes) are found mainly in coniferous wood (spruce and pine), whereas *sterols* (mainly β-sitosterol) are found in smaller quantities in almost all types of wood. The *mineral components* are mainly carbonates or glucuronates of calcium (40 – 70 %), potassium (10 – 30 %), magnesium (5 – 10 %), iron (up to 10 %), and sodium. Other metals, such as manganese and aluminum, are also present as cations in smaller quantities, which are often dependent on the location. Oxalate, phosphate, and silicate anions can also be found.

Lignin is a cross-linked macromolecule made up of phenylpropane units (C_9), which are linked together by at least ten different C−C and C−O bonds. The irregular structure of lignin arises from its biosynthesis, in which the last step is a nonenzymatic, random recombination of phenoxy radicals of coniferyl, sinapinyl, and *p*-coumaryl alcohols. The dimeric and oligomeric intermediates, obtained in vitro in the dehydrogenative polymerization of coniferyl alcohol, formed the basis of the constitutional scheme of spruce lignin proposed by Freudenberg (→ Lignin, Chap. 3.) [31]. Freudenberg's hypothesis of lignin biosynthesis was later confirmed for beech lignin by degradation with thioacetic acid and ^{13}C-NMR spectroscopy [32]. Variations in the chemical reactivity of lignin arise from the proportions of the three structural units. Whereas coniferous lignin consists mainly of guaiacylpropane (4-hydroxy-3-methoxyphenylpropane) units (G), deciduous wood lignins also contain up to 50 % syringyl (3,5-dimethoxy-4-hydroxyphenyl) groups (S), and compressive wood lignins up to 40 % *p*-hydroxyphenylpropane (H) units. Up to 30 % of the latter also occur in grass lignins.

On an industrial scale, lignin and hemicelluloses are removed from wood by pulping (→ Paper and Pulp, Chap. 1.3.) under partial degradation and isomerization. Lignin can be removed from wood more selectively under laboratory conditions by using chlorine, chlorine dioxide, or peracetic acid. From the residue known as holocellulose, hemicelluloses can be extracted either with 17.5 % sodium hydroxide – with α-cellulose as extraction residue – or with dimethyl sulfoxide.

Hemicelluloses are polysaccharides and differ from cellulose in that they consist of several sugar moieties, are mostly branched, and have lower molecular masses with a degree of polymerization (DP) of 50 – 200. The two main types of hemicelluloses are xylans and glucomannans. Structural and quantitative differences exist between hemicelluloses in deciduous and coniferous woods. *Deciduous woods* contain 10 – 35 % *O*-acetyl-4-*O*-methylglucuronoxylan and 3 – 5 % glucomannan, whereas *coniferous woods* contain 10 – 15 % arabino-4-*O*-methylglucuronoxylan and 15 – 20 % *O*-acetylgalactoglucomannan [33]. The upper part of Figure 10 shows a section of the structure of a coniferous wood galactoglucomannan. The molecule chain ("backbone") consists of β-(1,4)-glycosidically bonded glucose and mannose units in the ratio 1:3, with 0.1 – 1 α-(1,6)-bonded galactose moiety and 0.24 acetyl group per glucose unit as side groups. In the lower part of Figure 10, a deciduous wood xylan is depicted with β-(1,4)-glycosidically bonded xylose units in the chain, 0.1 α-(1,2)-4-*O*-methylglucuronic acid moiety, and 0.7 *O*-acetyl group per xylose unit as side groups.

Deciduous wood glucomannans contain only ca. 1.5 mannose units and no side groups per glucose unit; coniferous wood xylans have 0.2 α-(1,2)-4-*O*-methylglucuronic acid and 0.13α-(1,3)-L-arabinofuranose groups per xylose unit. Larch wood contains 10 – 25 % arabinogalactan, which consists of β-(1,3)-bonded galactose moieties in the main chain, with β-(1,6)-bonded short side chains of D-galactose and L-arabinose units. This compound is usually classified as a pectin. Pectins (→ Polysaccharides, Chap. 3.1.) occur in small quantities in the primary walls of all wood cells.

The biogenesis of D-glucuronic acid, D-galacturonic acid, D-mannose, D-galactose, D-xylose, and L-arabinose starts from guanosine diphosphate(GDP) – glucose and involves epimerization, dehydrogenation, and decarboxylation.

Cellulose. In terms of quantity, cellulose is the most important wood component. It consists of unbranched long chains of β-(1,4)-glycosidically bonded D-glucose units, which exist as pyranose rings in the $^4C_1$ conformation with equatorial OH and CH_2OH groups. The DP of native cellulose is between 10 000 and 14 000. In industrial pulping however, cellulose is degraded to DP values of 1000 – 3000. Annually ca. 130×10^6 t of pulp is used for paper production worldwide, and only ca. 5×10^6 t for the production of viscose and cellulose esters (acetate, nitrate, and mixed esters) and ethers (methyl-, carboxymethyl-, and hydroxyethylcellulose). The byproducts lignin and hemicelluloses present in the waste liquor are incinerated to recover the pulping chemicals. They pollute the wastewater from pulping factories to a great degree.

1.3.2. Distribution of Components in the Cell Wall

The distribution of cellulose, lignin, and hemicelluloses varies among the individual layers of the wood cell walls. Lignin is formed at a distance of ca. 10 cell layers from the cambium only after formation of the cells themselves. It is deposited in the spaces between the cellulose fibrils in the secondary wall and between the cells in the middle lamella, so that its concentration in the middle lamella and in the corner sections is highest (70 – 90 %), while only 20 % is found in the secondary wall. Because of the larger volume of the secondary wall, the main quantity of lignin (63 – 74 %) is found in this layer, which also contains ca. 90 % of the cellulose and 70 – 80 % of the hemicelluloses, with the concentration decreasing toward the lumen. The lignin distribution in the wood cell wall can be determined from the absorption of UV light (240 nm) by very thin (0.1 µm) microtome sections under the microscope [34] (Fig. 11).

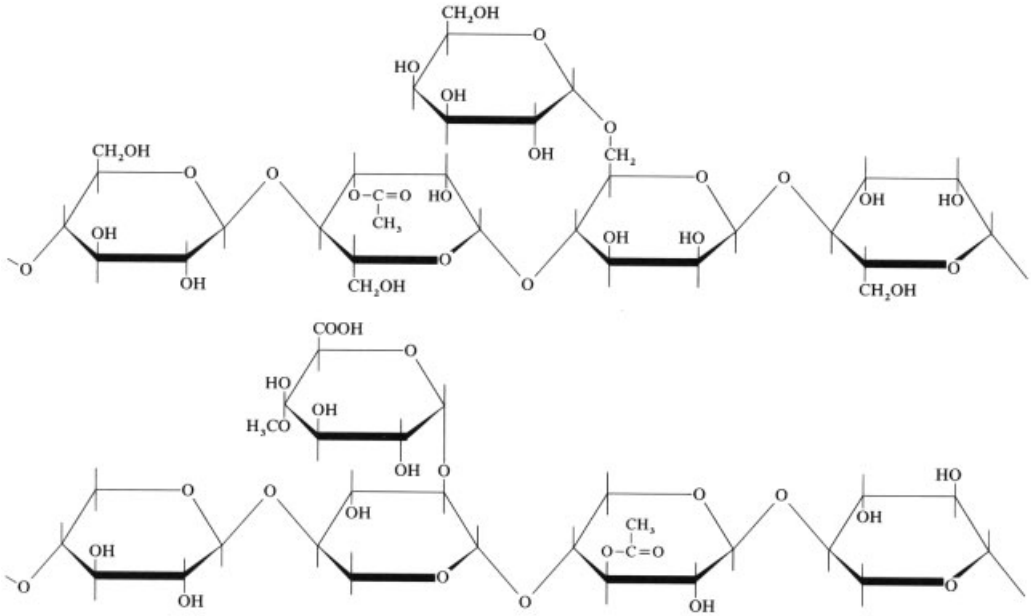

Figure 10. Sections of coniferous wood galactoglucomannan molecules (above) and deciduous wood 4-*O*-methylglucuronoxylan molecules (below)

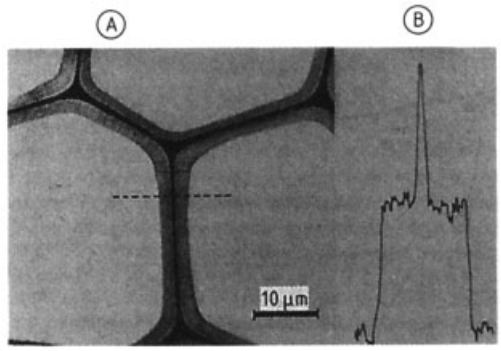

Figure 11. A) UV micrograph of microtome cross-section of an earlywood tracheid of black spruce (*Picea mariana*); B) Densitometric lignin distribution curve along the dotted line [34].

The proportions of lignin and tissue in the secondary wall, and the narrow and corner sections of the middle lamella in early and late spruce tracheids, are shown in Table 6. More lignin is found in the narrow and corner sections of the middle lamella in earlywood (28 %) than in latewood (19 %).

1.3.3. Macromolecular Structure of Cell Wall Components

Of the three cell wall components, only cellulose has a partially (ca. 70 %) ordered crystalline structure, whereas the hemicelluloses and lignin are amorphous. Because of regular hydrogen bonds, the cellulose chains are ordered to form elemental fibrils with a diameter of ca. 3.5 nm and alternating crystalline (micelles) and amorphous regions. These, in turn, are associated to form microfibrils, which are ca. 10 times thicker. The microfibrils form layers (Fig. 8) between which the lignin and the hemicelluloses are deposited. Chemical bonds are formed between the α-C atoms of lignin and the hydroxyl groups of the hemicelluloses (benzyl ethers), which prevent the leaching of lignin from the wood by neutral organic solvents. Irregular hydrogen bonds also exist between the hydroxyl groups of lignin, the hemicelluloses, and cellulose, through which a close, but irregular network of the three components is formed.

The cellulose microfibrils are first formed biogenetically as the framework of the cell wall, whose growth is finished before lignification begins. Lignification occurs in the secondary wall by filling the spaces between the cellulose mi-

Table 6. Distribution of lignin in black spruce tracheids (*Picea mariana*) [35]

Wood	Morphological differentiation	Relative absorbance	Tissue volume fraction, %	Lignin	
				% of tissue	% of total
Earlywood	secondary wall	1.0	87	22	72
	narrow sections of middle lamella	2.2 *	9	50	16
	corner sections of middle lamella	3.8	4	85	12
Latewood	secondary wall	1.0	94	22	81
	narrow sections of middle lamella	2.7 *	4	60	10
	corner sections of middle lamella	4.5	2	100	9

* Measured for the compound middle lamella. Measurements on oblique sections suggest absorbance values of 3.8 and 2.0 for the true middle lamella and primary wall sections, respectively.

crofibrils, whereby monomeric phenoxy radicals add to already formed polymers (endwise polymerization).

1.4. Physical Properties

Wood is hygroscopic; i.e., it exchanges water vapor with its environment. In a uniform climate, the moisture in wood is in equilibrium with this climate. The *moisture content of wood* is defined as the mass of moisture contained in the wood relative to the mass of the absolutely dry wood substance and is given as a percentage. Figure 12 A shows the relation of the equilibrium moisture content to the relative humidity of the air at 20 °C. This sorption isotherm is approximately valid for most wood species. Only species that contain hydrophobic components (e.g., teak, afzelia, and movingui) have significantly lower sorption isotherms.

The following moisture levels are recommended for the use of wood:

Furniture, interior decoration	6 – 10 %
Parquet	7 – 11 %
Windows, outer doors	10 – 15 %
Covered, open buildings	12 – 18 %

Adherence to these standard moisture levels during the final working and incorporation of wood reduces later variations in moisture levels and dimensions to a minimum.

In the hygroscopic region (i.e., below fiber saturation; FS in Fig. 12), the dimensions of wood increase (swelling) with increasing moisture content and decrease (shrinkage) with a corresponding decrease in moisture content. Figure 12 B shows the variation with moisture levels in the *degree of swelling* of spruce in the three main directions. Wood can also store liquid water in the pore spaces above the FS level, but no further swelling takes place. Because of the usual surrounding climates, the range of wood moisture content is particularly important between 5 and 20 %. In this range, wood swells and shrinks by ca. 0.2 % in the radial and 0.3 % in the tangential direction on 1 % change in moisture content.

The density of the pure cell wall substance is close to 1.5 g/cm^3 for all wood species. However, the *density* of a piece of timber (mass of wood relative to volume including pore spaces) covers a wide range from ca. 0.17 (balsa) to 1.2 g/cm^3 (pock wood). This range results from the different proportion by volume occupied by the cell wall in a particular species – in balsa ca. 11 %, in pock wood ca. 80 %. The remaining volume is pore space, which is filled with air in dry wood and water in wet wood. Density values quoted refer mostly to standard climate conditions (20 °C and 65 % relative humidity), i.e., a wood moisture level of ca. 12 %, as do the other moisture-dependent properties of wood. Since the density is a measure of the proportion of cell wall in the total wood volume, the following parameters tend to increase with increasing density: the modulus of elasticity and strength, abrasion resistance and hardness, and thermal conductivity.

Because of the structure of wood, all the properties mentioned, such as swelling and shrinkage, are anisotropic (i.e., direction dependent). The advantage of wood as a building material lies in the combination of high *strength* in the fiber direction and low thermal conductivity perpendicular to it. Wood is therefore particularly suitable for load-bearing and heat-insulating roof and wall components. Table 7 lists the average densities, moduli of elasticity, and strengths of a selection of wood species. It shows that

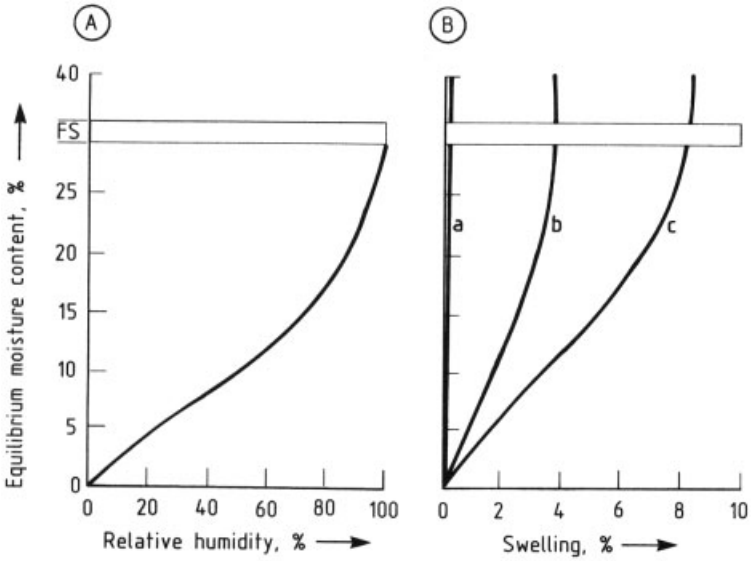

Figure 12. A) Sorption isotherm at 20 °C; B) Relation between degree of swelling and equilibrium moisture content in the three main directions of spruce
a) Longitudinal; b) Radial; c) Tangential

for wood species in general, the compressive strength in the fiber direction is approximately half the value of the tensile strength. However, as the direction of the load deviates from that of the fibers, the tensile strength decreases more sharply than the compressive strength. Wood should therefore not be subjected to tension perpendicular to the fiber direction (cleavage cracks due to shrinkage). However, application of compression perpendicular to the fiber direction is possible and frequently leads only to densification, not breakage.

In each case, the elastic properties and strength in the hygroscopic region decrease with increasing moisture content. However, above fiber saturation (use of wood in marine construction work), these properties are no longer altered by the moisture level in the wood.

For wooden floors, *abrasion resistance* and hardness play an important role. The relation between *Brinell hardness* H_B and density ϱ_N is given by the equation

$$H_B = 76^{\bullet}\varrho_N - 24.5 \left(\mathrm{N/mm^2}\right)$$

for the important density range of 0.5–1.0 g/cm^3.

The following are values for the *thermal conductivity* perpendicular to the fiber direction:

Spruce, pine, and fir	$0.13\,\mathrm{W\,m^{-1}\,K^{-1}}$
Beech and oak	$0.20\,\mathrm{W\,m^{-1}\,K^{-1}}$
Azobe and greenheart	$0.28\,\mathrm{W\,m^{-1}\,K^{-1}}$

2. Wood-Based Materials

2.1. Introduction

Wood has a number of favorable properties and has been a valuable material from time immemorial. It is relatively easy to obtain and can be worked readily. Wood has high strength at relatively low weight and has a decorative effect because of its color and graining. When used properly it often remains functional for hundreds of years even under adverse climatic conditions (e.g., outdoors).

Unfortunately, wood also possesses some less favorable properties. Since the arrangement of the elongated cells that form the bulk of the wood is predominantly parallel to the trunk (see Fig. 5), wood properties are anisotropic. Thus, the strength in the direction of the trunk is much greater than that perpendicular to it. Wood is also a hygroscopic material. Its moisture content is determined by the surrounding climate (see Section 1.4). Variations in the moisture level of wood in the hygroscopic region give rise to

Table 7. Average values for certain properties of a selection of wood species determined for small, defect-free test pieces at 12 % moisture content ∗

Wood species	Density, g/cm^3	Modulus of elasticity, N/mm^2	Strength, N/mm^2		
			Tension	Bending	Compression
Softwoods					
Western red cedar	0.37	8 000	60	54	35
Spruce	0.47	10 000	80	68	40
Pine	0.52	11 000	100	80	45
Larch	0.59	12 000	105	93	48
Hardwoods					
Balsa	0.17	3 500	24	24	10
Abachi	0.40	6 800	79	60	40
Poplar	0.42	8 300	77	70	36
American Mahogany	0.54	9 500	100	80	45
Sipo Mahogany	0.59	11 000	110	100	58
Iroko	0.63	13 000	79	95	55
Oak	0.67	13 000	110	95	52
Beech	0.69	14 000	135	120	60
Dark red meranti	0.71	14 000	130	110	58
Angelique	0.76	14 000	130	120	70
Afzelia	0.79	13 500	120	115	70
Balau	0.94	20 000		142	76
Greenheart	1.00	22 000	220	180	100
Azobe	1.06	17 000	180	180	95

∗ The flexural strength is valid for loads perpendicular to the fiber direction, the tensile and compressive strengths for loads in the fiber direction of the wood.

variations in its dimensions. These variations, in turn, differ depending on whether the anatomical direction is longitudinal, tangential, or radial. This phenomenon is known as warping. Wood is a typical natural material and is therefore inhomogeneous. It has growth irregularities and faults, such as knots, as a result of which the yield of usable material is often low. The lengths and particularly the widths available are limited, thus making the use of wood for certain applications difficult or impossible.

The primary aim of the development of wood-based construction materials is to eliminate or reduce the undesirable properties as much as possible. At the same time, attempts are made to retain the favorable characteristics and make them predominate.

Conversion of solid wood into wood-based materials always follows the same principle. Cordwood is cut into small pieces that are combined with a suitable binder (Fig. 13). Cutting into smaller pieces involves sawing, peeling, chopping, or defibration. Boards, veneers, particles of varying dimensions, and fibers are thus produced. As the size of the piece decreases, the properties of the wood-based panel or construction material differ more and more from those of wood. The wood-based material becomes more homogeneous, and the possibilities for size enlargement in two or three dimensions and molding increase. However, at a comparable density, the strength decreases and the typical wood character is lost. This is exemplified by the changing properties in the series sawn timber, plywood, particle board, and medium-density fiberboard (MDF).

2.2. Laminate Bonding

2.2.1. Laminated Structural Timber (Glulam)

2.2.1.1. Development

The art of building with wood is many centuries old. Unfortunately, the quantity of trunk wood of sufficient dimensions available for making load-bearing building components with the spans desired is continuously decreasing. This consideration gave the German carpenter Otto Hetzer the idea in 1906 of using a completely new method. He cut the cordwood into boards and glued the graded and dried lamellar boards together with adhesive to form new cross-sections (glulam girders, i.e., glued laminated girders, or laminated beams).

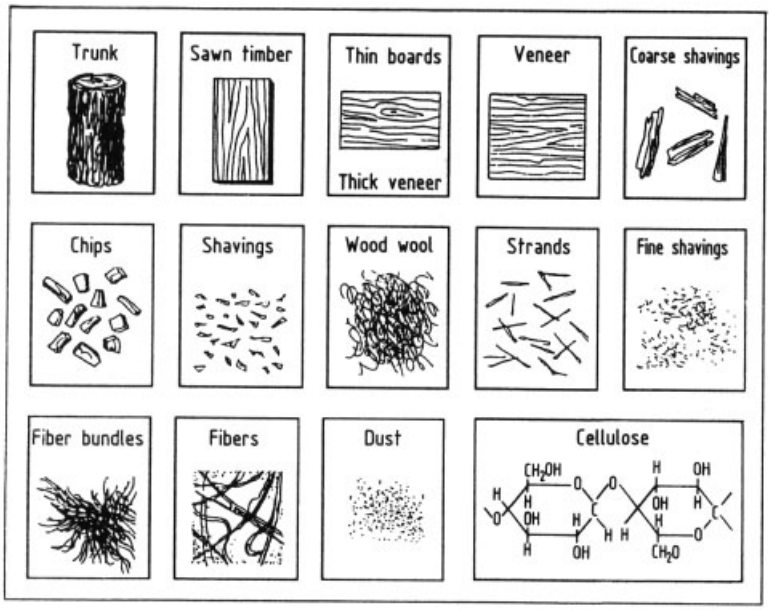

Figure 13. Classification of wood elements (according to MARA 1972 [18])

The actual breakthrough of the process for using glulam girders in building occurred in 1930 with the development of synthetic adhesive resins. Large-scale industrial production opened up new areas of application for this method. Besides being used in tall buildings, glulam girders were used in airplanes (Mosquitoes in England and the Spruce Goose in the United States), sometimes with large wing spreads, and for mine casings because of their antimagnetic properties. Glulam girders are now used mainly for building in central and northern European countries and North America, where wood has traditionally been used widely for this purpose.

Glulam girders were traditionally produced in small, handcrafting workshops. These have been replaced in recent decades by efficient industrial plants, which use machine systems and automatic production lines and can offer quality assurance.

2.2.1.2. Construction and Production

Raw Materials. Laminated wood is made mostly from spruce, which is the easiest wood to process and fulfills the usual requirements concerning strength and durability.

Pine, larch, and Douglas fir are also used in special cases. Hardwoods are generally used only for building components that must support very heavy loads. For some hardwoods, the effect of wood components on the durability of adhesive compounds is not entirely clear.

Adhesives. In a normal indoor climate and for building components protected from weathering, glues based on urea resins are used. For building components that are exposed to strong climatic variations or weathering, weather-resistant adhesives based on phenol – resorcinol – formaldehyde (PRF) are employed. These PRF resins give a dark brown glue line, unlike adhesives based on amino resins. In the 1990s a special glue type for bonding laminated was developed based on using melamine – resorcinol – formaldehyde resins. These bondings are highly resistant to boiling water and give a pale glue line.

Production. Lamellar boards are sawed from cordwood trunks. The boards are predried to ca. 10 % residual moisture, graded, and then bonded in automatic plants to form continuous boards with dovetail joints. The moisture in the boards to be glued should be 2 – 3 % below the expected equilibrium moisture level. Variations in the moisture content of the different boards

should not be greater than 2 %. The broad sides of these joined boards are planed until smooth and uniformly coated with adhesive. In constructing the cross-sections, particular attention is paid to the positions of the annual rings in the boards (Fig. 14). The rifts (vertical) and the flats (horizontal) should be distributed regularly.

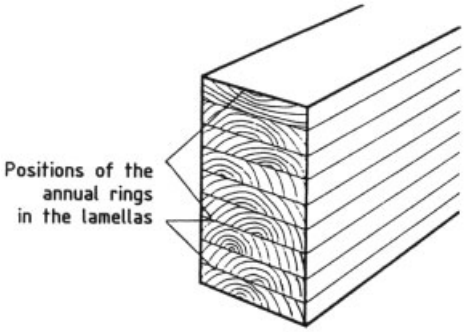

Figure 14. Lamellar construction of a glulam girder cross-section

Depending on how exactly the surfaces fit together, 400 – 600 g of adhesive liquor is applied per square meter. Glued boards are stacked and stored under pressure at room temperature or slightly higher (40 °C) so that the adhesive resin cures after a few hours and a composite beam is formed.

The geometry of the beams and their cross-sections can be varied widely and adapted to requirements of static calculations and aesthetics. Lamination allows bent beams with differing profile shapes and spans of 60 – 100 m to be produced.

Besides individually produced building components, the continuous production of laminated wood as a band was recommended [36]. Such standardization with regard to dimensions is expected to improve the economics of this process.

2.2.1.3. Properties

Laminated wood is clearly superior to squared logs in its performance. By eliminating the natural wood faults through careful sorting of the boards, a more homogeneous construction material is formed with regard to appearance and strength.

The behavior of wooden girders in the event of fire is especially favorable. Through the formation of a charcoal layer on the surface, the center of the girder is protected from further rapid attack by the fire and is thus insulated. Therefore, the desired extent of fire resistance can be determined only by the corresponding dimensioning of the load-bearing beam.

Unlike steel girders, glulam girders have high rigidity. Treatment with preservatives is recommended as a surface protection for building components exposed to weathering to prevent attack by microorganisms.

Testing and Standards. To test glulam girders, European countries have agreed on standards that stipulate classification, strength requirements, and test procedures. Because of the economic importance of glulam beams, bondings based on amino and phenolic resins are continually controlled according to EN 301 and EN 302. EN 302.1 – 4 describes the following tests:

1) Determination of shear strength
2) Determination of resistance to delamination
3) Determination of the effect of temperature and moisture cycles on tensile strength (damage by acid)
4) Determination of the effect of shrinkage on shear stength

The results of these short-term tests are used to predict the long-term behavior of glulam girders. The effects of climate change and of cold and boiling water are determined. The boiling-water test showed that short-term tests do not always reflect the behavior of a gluing in outdoor weathering [37]. Tests according to ASTM D 1101 – 59 and D 905 – 49 show comparable results in the delamination test. Contraction of the adhesives during setting should be as small as possible; otherwise cracks occur in thicker glue lines, which are formed when the sections to be glued fit less exactly.

Tests carried out over many years [38] on the aging of wooden building components gave, for example, unsatisfactory results for brittle pure urea resin adhesives in the case of thick glue lines. Broad application of urea resins in building with wood was made possible only by using suitable extenders and fillers [39].

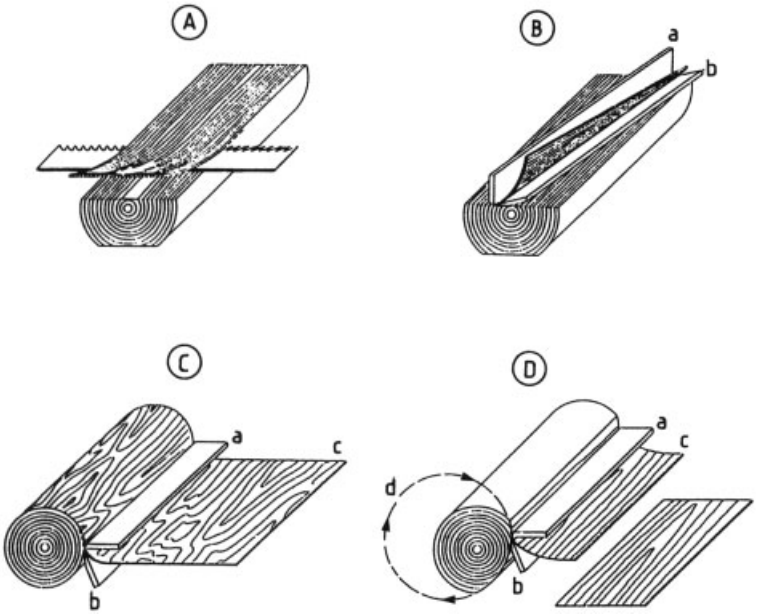

Figure 15. Veneer production methods
A) Sawn veneer; B) Cut veneer; C) Rotary cut veneer; D) Half-round cut veneer
a) Pressure pad; b) Blade; c) Veneer; d) Restraint

2.2.2. Veneers, Plywood, and Derived Products

2.2.2.1. Development

Veneers are thin sheets of wood obtained by sawing, cutting, or peeling. They are among the longest-known wood-based construction materials. The oldest archaeological finds originate from the heyday of ancient Egypt (ca. 3000 B.C.). Everyday and cult objects were decorated with inlays of different types of wood, known as marquetry. Because of their texture and color, veneers were initially used mainly for decorative purposes. However, composite constructions, consisting of a core of plywood with a veneer glued on to it, were soon recognized to warp less than pure wood boards. Since beautifully grained wood was rare and difficult to obtain even in antiquity and in the Middle Ages, veneers also offered the possibility of using wood sparingly [40].

2.2.2.2. Production of Veneers

Veneers are produced from cordwood or from specially divided blocks by using peeling and cutting machines or, less commonly, by sawing (Fig. 15). Many types of wood – e.g., poplar (*Populus* sp.), birch (*Betula* sp.), or Douglas fir (*Pseudotsuga* sp.) – can be peeled or cut either fresh or saturated with moisture through storage in water. Other types of wood – e.g., beech (*Fagus silvatica*), oak (*Quercus* sp.), ash (*Fraxinus* sp.), walnut (*Juglans* sp.), or various tropical high-density deciduous woods – must first be softened by steaming or boiling to alow the wood to be cut more easily.

After peeling, the veneers have a moisture content of 30 – 110 % (based on oven dry weight). This must therefore be reduced by drying to 4 – 12 % for further processing. Special veneer dryers are used through which hot air is blown at 80 – 200 °C.

2.2.2.3. Principal Types of Plywood

By combining veneers of different types of wood, densities, numbers of layers (plies), and different core materials, a large number of materials can be produced with widely differing properties and areas of use [41]. Plywood can be classified according to different criteria, frequently

according to its construction. EN 313 (see Section 2.2.2.6) differentiates among veneer, core, and composite plywood.

Veneer plywood normally consists of at least three layers of wood glued to one another crosswise. The barrier technique (orientation of the grain of adjacent layers at an angle, usually 90°) requires a symmetry around the middle layer with regard to type of wood, grain direction, and thickness of individual layers. By forming a barrier, the contraction in width of layers with longitudinal grain direction depends extensively on the mechanical deformation of transverse layers. At the same time, the strength and deformation properties in the longitudinal and transverse directions of the plies are aligned with one another, thus decreasing the anisotropy.

Plywood boards are thus considerably more dimensionally stable and can be subjected to loads more uniformly than solid wood boards. The number of veneers is usually uneven, and the arrangements are symmetrical. Three-, five-, and seven-fold boards, etc., are produced (Fig. 16).

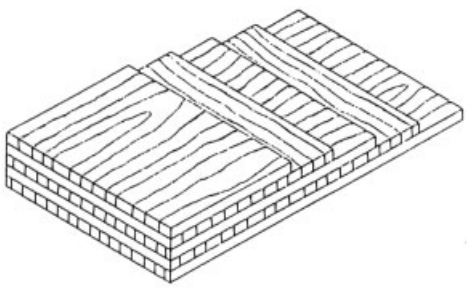

Figure 16. Five-layer veneer plywood

Core Plywood. In core plywood (blockboard), the core consists of battens or strips, usually of light coniferous wood that has been cut from a block of boards or veneers that have been glued together and then blocked on both sides with veneers (Fig. 17). Depending on the core, two types of sandwich board exist: *battenboard* and *laminboard*. Wood core plywood is used for high-grade joinery.

Composite Plywood. Two types of composite plywood exist. In the first type the core consists of a material, usually of low density, that is planked on both sides with veneers. Practically any material can be used for the core (e.g., extruded particle board, paper honeycomb, and for special applications, polyurethane foam). These types of boards are used, for example, for doors and partitions. In the reverse case, the core can be made of plywood and the covering of metal sheeting or glass-fiber-reinforced plastic. The possible combinations are endless. Composite materials with special properties, such as high strength, good insulating capacity, and high abrasion resistance, can be produced for particular applications (e.g., containers).

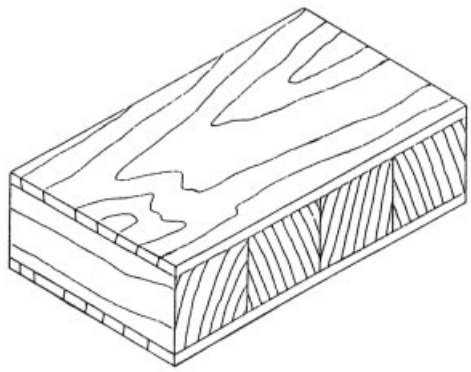

Figure 17. Core plywood

Particular Laminated Materials. *Glulam boards* consist of three or more layers of wood sheets that are glued together crosswise. The thickness of the layers is at least 8 mm. These boards are used as shuttering panels and to an increasing extent as furniture panels.

In *laminated veneer lumber*, the veneers are glued together in the direction of the fibers. This allows the high longitudinal tensile and flexural strengths of the solid wood to be retained. If the veneers used for producing laminated wood are impregnated with synthetic resins – mostly phenol – formaldehyde resins – and compressed under pressure, *densified laminated wood* with a density $>1000\,\mathrm{kg/m^3}$, is obtained. This is used mainly for technical purposes.

Molded plywood is formed when the plywood is molded by using appropriately shaped presses. It is used, for example, for chair backs and seats, furniture, wall paneling, and girder webs.

Wood-based construction materials that are used mainly in building are known in North America as *engineered wood* (see Section 2.2.2.7).

2.2.2.4. Adhesives and Additives

The natural adhesives used initially, such as protein adhesives (gluten, casein, soybean, and blood albumin adhesives), have disadvantages (low resistance to climate variations, sensitivity to attack by microorganisms, long pressing times, and sometimes highly complex processing). Since the 1930s they have been increasingly replaced by synthetic adhesives, and these are nowadays used almost exclusively. They include urea – formaldehyde (UF), melamine – formaldehyde (MF), and phenol – formaldehyde (PF) resins, as well as MUF or MUPF mixed condensates. The use envisaged and the cost determine the choice of adhesive. A detailed description of the chemistry and technology of wood adhesives can be found in [42].

Urea – Formaldehyde Resins. The most important adhesives for gluing wood are urea – formaldehyde resins. The use of UF resins to glue plywood was first described in 1929 [43]. Production of these wood adhesives has since developed enormously [44]. Urea – formaldehyde resins are used for adhesion of wood materials used in the interior and have limited resistance to moisture. They can be readily adapted to operational requirements through the nature and dosage of curing aids and additives. Ammonium salts (e.g., NH_4Cl) are generally used as curing aids. They lower the pH by reacting with formaldehyde. The rate of curing increases with increasing temperature and decreasing pH. The viscosity of the adhesive liquor is adapted to processing conditions by the addition of water, depending on the content of extenders and fillers. The pressing temperature is 80 – 120 °C.

Melamine – formaldehyde resins undergo a curing reaction analogous to that of UF resins. Melamine – formaldehyde resins are characterized by their high water resistance and colorless glue lines.

For cost reasons, however, pure MF resins are rarely employed in plywood production. Instead, MUF or MUPF mixed condensates are used when high weather resistance is required. Depending on the melamine content and the type of wood, adhesions for semiexterior or exterior applications are obtained.

Phenol – formaldehyde resins are generally cured by heating and also by a combination of heating and curing aids. Phenol – formaldehyde resins are less reactive than UF or MF resins. They therefore require higher pressing temperatures (115 – 140 °C). Curing of phenol – formaldehyde resins is more sensitive to wood moisture and to pressing conditions than curing with other adhesives. Their alkali content can have damaging effects, e.g., on subsequent surface treatment of veneered boards.

Poly(vinyl acetate) adhesives are mainly used for bonding and veneer jointing.

Extenders and Fillers. In plywood production, extenders and fillers are used together with the adhesives to avoid glue penetration during hot pressing. *Extenders* are finely ground products of organic origin that are capable of swelling and gelatinizing and thus act as adhesives themselves (e.g., starches and cereal and vegetable flours). They are generally used with UF resins at levels of 15 – 30 % or higher (extender plus water, based on the adhesive resin) to reduce the brittleness of the glue line and resin consumption.

Fillers do not produce their own adhesion. They are finely powdered materials (e.g., nut shell or olive stone flours) or minerals, such as kaolin or pulverized gypsum (lenzin) that can be swelled only slightly or not at all. Fillers are normally used with PF resins.

The addition of extenders and fillers can decrease the resistance of the adhesion to water and microorganisms. Mineral fillers also increase the wear on tools.

2.2.2.5. Production of Veneer Plywood

The main operations involved in the production of veneer plywood are preparation, assembly, application of adhesive, stacking coated veneers (sometimes prepressing), hot pressing,

conditioning, forming (sometimes sanding), and sometimes fault elimination.

The adhesive is generally applied by using two- or four-roll machines, by pouring using an adhesive curtain, or by melting an adhesive film. The coated veneers are stacked either manually or by machine. To make transportation easier and reduce stack height, coated veneer stacks are frequently precompressed in an unheated one-daylight press. Hot pressing occurs in one- or multi-daylight presses similar to those described in Section 2.3.1.5 but usually with a smaller size. Pressing conditions (temperature, time, and pressure) depend on the type of wood, the strength of the plywood, and the type of adhesive. Pressures of 0.6 – 2.5 N/mm^2 and temperatures of 80 – 140 °C are usually applied. After gluing and sometimes climatization, veneer plywood is sanded, cut to standard or ordered measurements, and graded according to quality.

2.2.2.6. Properties, Testing, and Uses

Standards. The properties and quality of the adhesion of plywood are laid down in most countries by standards and trademarks. The following standards are valid in Europe and the United States:

Europe	
Plywood	
Classification and terminology	EN 313 (Parts 1 and 2) 1991
Test methods and properties	EN 314 (Parts 1 and 2) 1991
United States	
Hardwood and decorative plywood	PS 51-71, NBS-1972
Plywood for construction and industrial uses	PS 1-74, NBS-1974
APA grade – trademarks, available from American Plywood Assoc. P.O. Box 117 000, Tacoma, Wash. 98411-0700	

Depending on the type of adhesive and the wood, a distinction is made between plywood for exterior and interior use; further classification is made for the following applications:

1) In dry areas, i.e., a normal indoor climate (interior)
2) In moist areas, i.e., with protection against direct weathering also suitable for outdoor climates (semiexterior)
3) For outdoor use, suitable for long-term weathering (exterior)

The combined tension and shear resistance is generally tested on appropriately shaped test objects. A further criterion for the quality of an adhesion is the nature of the fracture pattern (percentage of fracture in the wood). With good adhesion the break will occur not in the glue line but in the wood. Before the strength of adhesion is tested, the sample is pretreated depending on the type of bonding. Pretreatment normally involves storage under different conditions, e.g., in cold, warm, or boiling water with redrying, or cycle tests.

Properties. Veneer plywood is normally produced in thicknesses of 1 – 12 mm and as multiply boards wich thicknesses of 13 – 50 mm. Wood core plywood is available in thicknesses of 10 – 60 mm. The density of veneer plywood depends on the type of wood and the plywood construction (500 – 800 kg/m^3). Wood core plywood generally has a lower density (450 – 750 kg/m^3).

The *hygroscopic properties* of plywood correspond approximately to those of solid wood. For indoor uses, the equilibrium moisture level is between 8 and 12 %. The water absorption decreases as the number of veneer layers and glue lines increases.

Dimensional changes due to contraction and swelling are low in the surface direction because of the barrier technique. The changes correspond approximately to the values for solid wood in the longitudinal direction (0.005 – 0.01 % for changes in water content of 1 %). However, changes in thickness approach those of solid wood in the radial direction (0.012 – 0.019 % for changes in water content of 1 %) despite the adhesive layers.

The *strength properties* of veneer plywood depend mainly on those of the veneers lying parallel to the main line of stress and on the type of wood. Although in solid wood the modulus of elasticity at right angles to the fiber is only ca. 10 % of that in the fiber direction, the properties of veneer plywood approach isotropy as the number of layers increases. Star plywood has almost the same strength when stressed at any angle. The strength properties of a veneer plywood (DIN 68 705, part 3, 1968) are listed below:

Bulk density (at 12 % moisture content)	500 – 800 kg/m^3
Flexural modulus of elasticity	7000 – 12 000 N/mm^2
In the direction of the	
covering fibers (3 layers)	700 – 1200 N/mm^2
(5 layers or more)	3000 – 7000 N/mm^2

Uses. Most of the plywood produced is used in building. Douglas fir and pine are used predominantly worldwide. Geographically, the consumption of plywood varies widely and depends mainly on local building traditions. Countries such as Canada and the United States are the largest plywood consumers worldwide, with 105 and 77 m^3 per 1000 inhabitants, respectively. In Europe, the Netherlands is the leading consumer, with 28 m^3 per 1000 inhabitants; in Germany, consumption is only 12 m^3, and in Italy only 6 m^3 per 1000 inhabitants [45].

Typical uses for veneer plywood are as construction plywood for wall, floor, and ceiling panels and also as panels for concrete shuttering. Industrial plywood is used to build railway coaches (e.g., for high-speed trains). A large area of use is in containers, boxes, and packagings. Plywood with decorative covering veneers is frequently used in interior furnishings (e.g., wall panels or doors).

2.2.2.7. Engineered Wood

The term engineered wood is used predominantly in North America and refers to wood-based construction materials, made by gluing together small-sized woods, that are used in building. The reason for this development is the increasing scarcity of suitably sized tree trunks.

Particles of small-sized wood and wood residues, such as flakes, strands, or parts of veneers, are combined by oriented agglomeration to give new wood-based construction materials using a method similar to that of particle board production. The properties of these materials are very similar to those of solid wood. Phenol – formaldehyde resins, amino resins (UF/MF), and polymeric methylene diphenylene isocyanate (PMDI) are used to glue the oriented wood pieces, which are mostly Douglas fir and pine.

By using 4-cm-wide and 12-cm-long strands, building materials of the most varying types can be produced in this way. They are known in the building industry as *laminated strand lumber* (LSL). The material is produced in the form of boards with thicknesses up to 14 cm, from which the desired building components can be cut (Fig. 18).

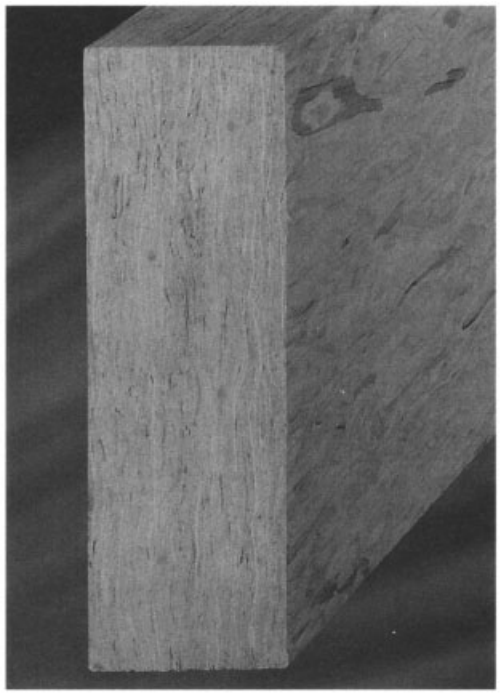

Figure 18. Laminated strand lumber

If strands with a thickness of ca. 3 mm and a length of ca. 250 mm (e.g., from peeled veneers) are glued in a parallel orientation, *parallel strand lumber* (PSL) is obtained. These beams can have a cross-section of 28 × 49 cm and are produced in a batch process in special presses in lengths of 20 m. Parallel strand lumber can be worked with normal tools and can be used like laminated wood (Fig. 19).

Gluing together peeled veneers with the fibers parallel gives *laminated veneer lumber* (LVL). In principle, this corresponds to veneered laminated wood. LVL is produced in lengths of 20 m and thicknesses up to 9 cm. The thickness of the veneers used is 2 – 5 mm (Fig. 20).

Trade Names. These building materials are sold in North America under the commercial names Intrallam (LSL), Parallam (PSL), and Microllam by Trus Joist International Inc. and Mac-Millen Bloedel Ltd.

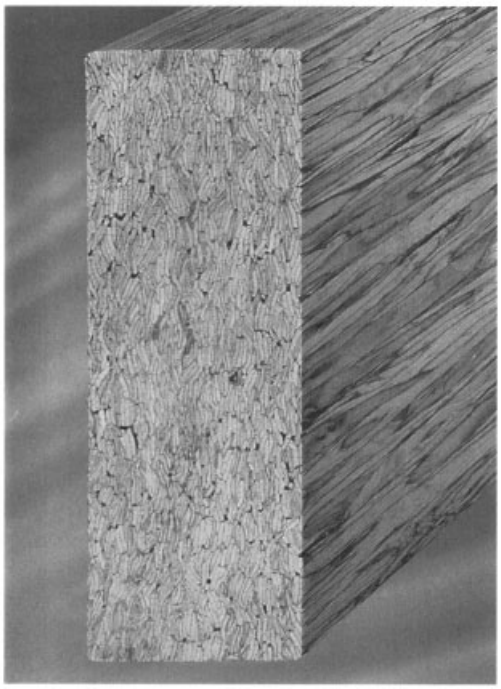

Figure 19. Parallel strand lumber

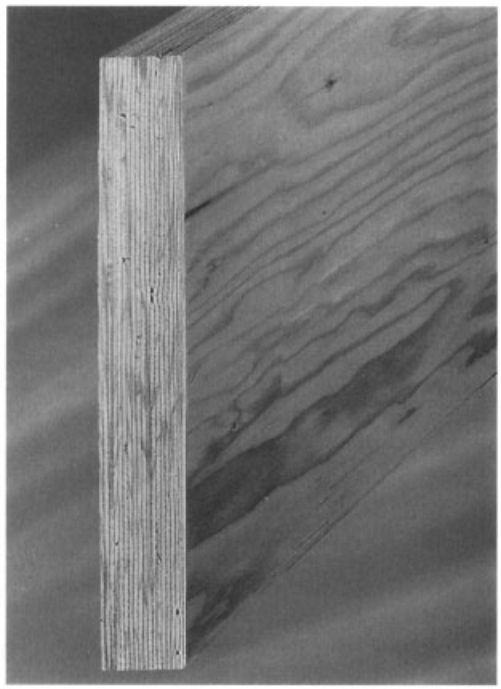

Figure 20. Laminated veneer lumber

2.3. Particle Bonding

To overcome the pronounced anisotropy of solid wood, it must be cut up into small pieces; then the pieces can be oriented to a desired fiber direction. The following wood-based particles have achieved economic importance in the production of wood-based construction materials:

1) Chips for defibration and grinding
2) Shavings from cutting, peeling, and sawing
3) Strands
4) Wafers or flakes
5) Fibers
6) Wood wool

In industrial production, these particles are formed in a certain distribution with regard to dimension and shape. In practice, therefore, they must be sieved or classified. After drying and application of adhesive, the properties of the wood-based construction material can be influenced by targeted orientation of the particles during mat forming and by molding.

2.3.1. Particle Board

2.3.1.1. Development

In 1902, Henry Watson of Indiana had the idea of making particle board from waste wood. Unfortunately, no adhesives available at this time could allow industrial realization of this idea.

Only when the chemical industry made a suitable adhesive system available with the development of amino resins, could the industrial production of particle board be begun in Germany in 1941. After the end of World War II, the development of particle board technology made rapid advances in Germany and subsequently in the whole of Europe. Important contributions came from the Swiss F. Fahrni and the German W. Klauditz. In 1991, particle board was the most important industrially produced wood-based construction material, with an annual worldwide production of ca. 50×10^6 m^3.

2.3.1.2. Types of Particle Board

Flat Pressed Particle Board. In flat pressed particle board, the particles are arranged predominantly parallel to the board surface and bonded. Most of the particle board produced worldwide belongs to this group. The wafer, flake, and strand boards and oriented structural boards (OSB) produced widely in North America are, in principle, flat pressed particle boards. Although *wafer* and *flake boards* are made up predominantly of (decorative) wood particles with large surfaces and a thickness up to 0.02 inch (ca. 0.05 cm), *strand boards* are made from strands with a high ratio of slenderness having a thickness up to 0.2 inch (ca. 0.5 cm) and a length of 3 – 6 inches (ca. 7.6 – 15 cm). The production, transport, gluing, and matforming of these particles require special machines.

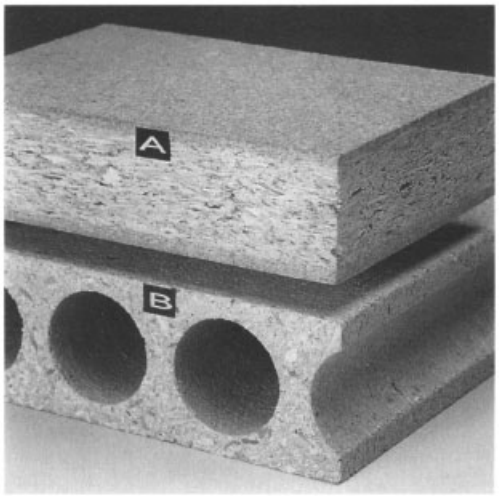

Figure 21. Different texture of flat pressed (A) and extruded particle board (B)

Extruded Particle Board. These boards are produced by extrusion. The process was developed in the 1960s by O. Kreibaum [46] and is still termed Okal process. Particles are arranged predominantly at right angles to the board surface. Short particles or granulated wood are mainly used. Adhesive-coated particles arrive in portions in the heated press, are compressed by using a pressure piston driven by an eccentric flywheel, and are extruded through the hot zone of the press. A continuous band of board is formed as it leaves the press. Heating of the board is accelerated by heated rods that project into the press. The heated rods themselves form tube-shaped hollows in the board. Extruded particle board is generally used only in the planked or veneered state because of the low flexural strength resulting from the production process. The production of profiles constitutes a further variant of the extrusion process. Figure 21 shows the difference in texture between flat pressed and extruded particle board.

Molded Objects. The production of molded particle board was described in 1948 in a Swedish patent [47]. In the same year, work began on the Thermodyne process, in which natural wood constituents are activated as binders at pressing temperatures up to 290 °C and pressures up to 300 bar [48]. The process was not successful economically because of the extreme operating conditions.

The process patented by the Werz company (Werzalite process) in 1956 uses wood particles containing 20 – 30 % amino resins [49]. These are compressed to form a blank and then prehardened. The blank is then coated immediately on all sides with paper impregnated with melamine resin. Operating conditions are very similar to those employed in particle board production. Production of molded objects using the Werzalite or similar processes was successful economically. Tabletops, window sills, paneling, palettes, and other molded objects are produced in this way.

2.3.1.3. Raw Materials

For particle board, low-value wood and industrial waste wood (e.g., remainders, slabs, splinters, and shavings from planing and sawing) are used. In countries with low indigenous wood production, the use of annuals such as cotton stalks, sugarcane bagasse, flax, rice husks, cereal straw, jute, esparto (alfa grass), and similar materials, instead of wood, is possible. These materials give a board that is very similar in properties to that made from wood [50].

In principle, all types of wood are suitable for making particle board. However, coniferous wood is used for most of the production in Europe and North America, which accounts for

75 % of the total worldwide particle board production. A combination of hard and soft types of wood is frequently used; the fact that these types of wood affect the properties of the board because of their differing sorption behavior is taken into account [51].

The time of wood harvest also affects sorption behavior and acidity, which in turn affect the curing of amino resins used to glue the particles. Some deciduous woods, such as oak or chestnut, have a high proportion of tannin and a low pH. Sometimes this impairs the strength produced by the gluing process. Wood bark used to be separated and incinerated or dumped, because the different structure of the bark generally required greater compression of the particles and the use of larger quantities of adhesive resins. Nowadays, however, the use of unpeeled wood has become common, even for the outer layer. Addition of 6 – 15 % bark to middle-layer particles does not affect board quality significantly [52]. Larger quantities of bark impair not only the appearance of the board, but also its strength and sorption properties. Excess bark in particle board factories is incinerated, and the waste gas is used for drying the particles [53].

Flat pressed particle board is now generally produced in several layers. The density of the individual layers and their construction determine the mechanical properties of the board. Light particle board has a density up to 500 kg/m^3. Most of the board produced is, however, in the density range of 600 – 800 kg/m^3. The thickness of the board ranges from 3 to 50 mm.

2.3.1.4. Adhesives and Additives

Amino Resins (UF, MF, and MUF). Urea – formaldehyde adhesive resins are used as 55 – 72 % aqueous solutions or as powders. For gluing particles, the resins now used predominantly have a basic U : F molar ratio of 1:1.0 to 1:1.3 to limit formaldehyde emissions. Catalysts based on ammonium salts or acids are used to accelerate the curing process. The use of resins with increasingly less formaldehyde finally led to a lowering of the cross-link density and thus to impaired quality of adhesion. This was corrected by reinforcing the UF resins with melamine [54]. To adhere to the strict limiting values for formaldehyde emissions in many countries, a special process for low-emission boards was developed by using reactive formaldehyde-rich resins and formaldehyde scavengers [55]. Urea – formaldehyde is used in quantities of 6 – 12 wt % solid resin, based on absolutely dry wood. UF-glued particle board is produced mainly for interior use and for furniture.

Higher-quality adhesions are achieved with melamine – formaldehyde, phenol – formaldehyde, or polyisocyanate resins. Because of their chemical similarity to UF resins, MF resins are mixed or cocondensed with them. Melamine – urea – formaldehyde and phenol-containing MUPF resins with 15 – 25 % melamine and up to 5 % phenol are used for particle board that is exposed to particularly high moisture levels. The dosage of resin is increased to 12 – 15 %, based on the dry wood particles.

Phenol – formaldehyde resins are high-value products, particularly because of their resistance to hydrolysis. They are used as aqueous alkaline solutions with a P : F molar ratio of 1:1.5 to 1:2.5. The alkali content is 3 – 10 %, depending on the degree of condensation of the resin. Phenol – formaldehyde resins are generally less reactive than amino resins. This means that, in practice, longer pressing times must be used for processing PF resins compared with amino resins. Alkaline phenolic resins dye particle board light brown and increase its hygroscopic nature. PF-glued particle board produces no significant formaldehyde emissions. It is used for exterior applications. The usual dosage of resin is 8 – 12 %.

Polymeric Methylene Diphenylene Isocyanate. Since 1980 adhesives based on PMDI have increasingly been used. They consist of a mixture of monomeric, polymeric, and oligomeric diisocyanates, which are capable of reacting with the cellulose, lignin, and moisture in wood. The quality of wood adhesions using PMDI is comparable to that using PF resins. A significant disadvantage of isocyanate adhesives is their tendency to stick to metals. PMDI can therefore not be used without release agents.

Handling PMDI adhesives requires particular care because of their high reactivity toward water and bases. However, they are very economical. Adequate board strengths are achieved with 3 – 5 % adhesive.

Adhesives Based on Natural Materials. Adhesives made from waste sulfite pulping liquor and from wood and bark extracts are known, but have not achieved great economic importance.

Mineral Binders. Special boards for building purposes are also produced using inorganic binders. These include Portland cement, magnesia cement (Sorel cement), and gypsum (see Section 2.3.3).

Additives. For repelling water and delaying swelling during particle board production with synthetic resins, 0.1 – 1.5 % paraffin, based on dry wood, is generally added. Paraffin is mixed with the adhesive liquor as a dispersion or applied to the particles as a finely dispersed melt.

2.3.1.5. Production

Particle Production. The wood to be milled should have a moisture level of ca. 50 % and be free of metallic or mineral substances. Low moisture levels increase the throughput of the size reduction machine, but also increase dust formation and fire hazards.

Chipping machines with very varying mechanisms are used for size reduction. For lengthwood (roundwood) machines with cutting tools fixed in a carrier disk or drum, together with spurs, are preferred. By pressing the wood onto the rotating blade carriers, uniformly thick particles are formed, which can be cut up further in a subsequent mill if desired. The most powerful chipping machines of this type are knife shaft machines with an operating width of 2 m. For making larger particles, knife ring machines are used, which have combined cutting and striking tools. To cut up the particles further, hammer, wing beater, and toothed disk mills are used. The target of the size reduction process is the production of a mixture of particles that are as uniform as possible in shape and thickness. The geometry of the particles affects both the further processing (distribution and dispersion of glue) and the mechanical properties of the board. Since particle board for use in surface coating is produced mainly in three or five layers, the aim of particle preparation is to produce different mixtures of particles. Fine particles for the outer and coarse ones for the middle layer are obtained by air classifying and screening. For particle board used in building, large proportions of fine particles and dust are undesirable.

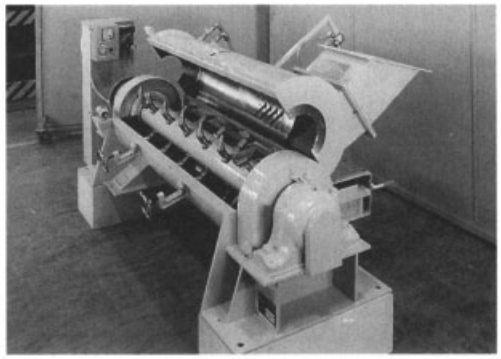

Figure 22. Particle glue-blending machine EK-AB 50 (Loedige, Paderborn, Germany)

Drying. To process the particles further, residual moisture levels of 1 – 3 % are necessary. High moisture levels, together with the water applied with the adhesive resin, generally cause steam cracking in the pressing process. Drying the particles is therefore indispensable.

Drying follows the rules of convection drying. In the first part of the drying process, wood moisture decreases almost linearly. The second part of the drying process is determined by vapor diffusion.

With particles of varying thickness and size, uniform drying is difficult to achieve. Separate drying of the coarse middle-layer and fine outer-layer materials is therefore necessary. To dry the particles the following powerful machines are used: (1) jet dryers, (2) tubular dryers, (3) stream dryers, and (4) drum dryers.

Glue blending of the particles occurs continuously in flow mixers to achieve as uniform a distribution of the adhesive on the surface of the particles as possible. Dried particles are fed to the mixer via a belt weigher. The necessary quantity of adhesive is dosed volumetrically or gravimetrically with electronic control.

In the mixer, adhesive is rubbed on the surface of the particles or applied by using spray jets. The efficiency of the mixer depends on the design of the mixing drum and the conveying

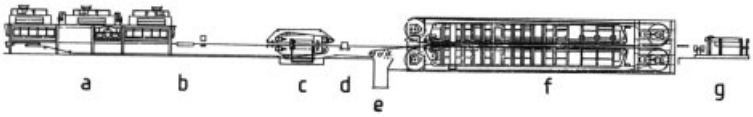

Figure 23. ContiRoll particle board press line (Siempelkamp, Germany)
a) Mat formers; b) Belt weighers; c) Prepress; d) Metal detector; e) Reject nose; f) ContiRoll press; g) Cut-to-size saw

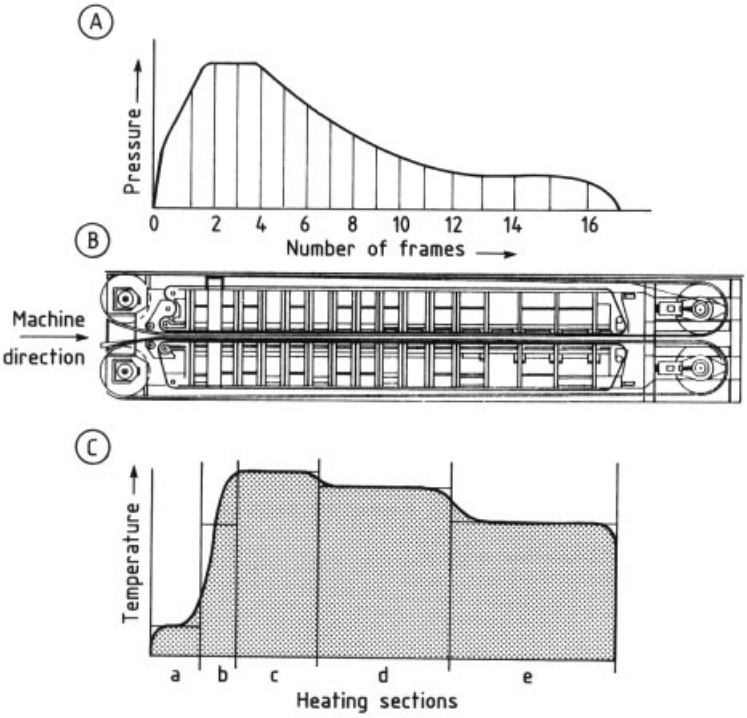

Figure 24. Pressure and temperature control in the ContiRoll press
A) Pressure program; B) Schematic of ContiRoll press; C) Temperature program

and braking tools. Because of the heat of friction produced, the mixing vessel must be cooled with water. Figure 22 shows a high-performance mixer from the Loedige company.

Forming Mats. Glue-blended particles are fed from the storage hopper into the dosing hopper of the forming station by using conveyor screws or belts.

By controlling the volume of the glued chips a uniform flow of particles is produced, which is deposited onto a moving forming belt via a toothed feed roller. The falling particles receive an impulse from a spreading roller, resulting in separation of the coarse and fine particles. Commonly, the mat is formed by using air classifying. In this process, air is blown against a curtain of falling particles so that fine particles are carried further with the air than coarse ones. Here also, classified particles fall onto a conveyor belt below. With two blowers set opposite one another, mats can be formed that have coarse particles in the center and finer particles on the outside. For reducing thickness, the mats are precompressed in a cold press and then passed to a hot press for activation of the adhesive.

Pressing and Pressing Techniques. In hot pressing the blank is compressed to the desired final thickness and the adhesive is irreversibly cured. Both continuously operating and batch presses are used.

Batch Presses. The output of presses for batch processing is limited by the pressing and cycle times. With *one-daylight* presses, high

pressing temperatures (200 – 230 °C) are used to shorten the setting time of the adhesive. *Multi-daylight presses* have up to 20 daylights and are operated at < 200 °C. Automatic charging and emptying of a multi-daylight press require complex machinery. The tolerance in thickness of the boards produced by using the batch press is so great that calibration by sanding is indispensable.

Continuously Operating Presses. In larger particle board plants, continuously operating presses are mainly used. With this technique, boards are produced with significantly smaller variations in thickness than in batch production. Various continuously operating systems exist.

Roller Presses. In roller presses the application of pressure and transportation is effected by a bed of rods. Figure 23 shows a ContiRoll plant of the Siempelkamp company, as a typical example of this type of machine. Pressure and temperature control in the ContiRoll process is illustrated in Figure 24. Figure 25 shows how the pressure is conferred by means of the bed of rods. The operating temperature is 220 – 250 °C. Continuously operating roller presses are usually designed for capacities of more than 400 m^3 of board per day. The production costs are up to 8 % less than those for batch presses because of the lower energy consumption.

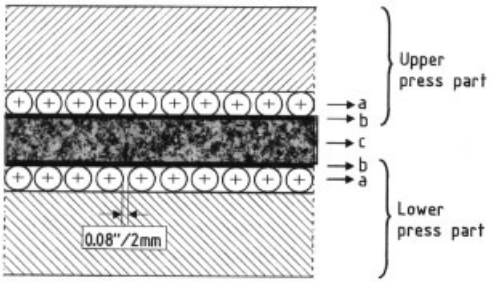

Figure 25. Moving rod system of the ContiRoll press
a) Bed of rolls; b) Steel belt; c) Mat

Gliding Film Presses. In gliding film presses the pressure is transmitted by an oil film (Hydro Dyn process of the Bison company) or by air cushions (Hymmen company).

Calender Presses. In calender presses pressure is applied to the blank by pressing it with steel bands on a heated cylinder (calender). Figure 26 shows a calender press that operates using the Mende process (producer Bison). The system is particularly suitable for producing thin particle board.

Extrusion Presses. In extrusion presses the particles are extruded continuously between heated plates or rods (Okal process, see page 26). Profiles and tubular boards are produced by this method.

Steam Injection Presses. Heating the mats can be accelerated by using steam injection presses. The steam is injected immediately before pressure is applied. The process is particularly suitable for thick boards, which are difficult to heat through. Since 1989, it is being developed further by Siempelkamp.

To *heat the presses* a heat-transfer oil is used predominantly as the energy carrier. Experience with the use of high-frequency energy is extensive. For cost reasons, it is now used only for warming the mats in prepresses.

The *pressing time* necessary is determined by the reactivity of the adhesive system and the pressing temperature. The specific pressure of 20 – 50 bar is decreased in steps as the spring back forces of the compressed mat decrease and the adhesive takes effect (Fig. 24).

Aftertreatment. After hot pressing, particle board is cooled and conditioned by storage for several days. This process leads to further setting of the adhesive so that the internal bonding of the board increases and the thickness swelling values decrease. The board is standardized by sanding.

2.3.1.6. Properties, Testing, and Uses

Properties. Even in glued particle board the hygroscopic properties of the basic material, wood, are noticeable. The state of equilibrium between the moisture in the wood and the relative air humidity causes ad- and desorption of water on climate change, which, in turn, lead to a change in volume. The properties of particle board are determined by the orientation of the particles, which depends on the process used. This effect is most obvious in the case of extruded and flat pressed particle boards (see Fig. 21). Even in flat pressed particle board the longitudinal and thickness-sectional swelling

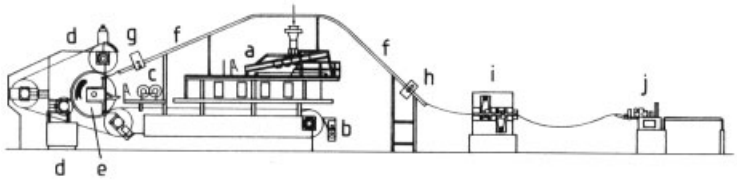

Figure 26. Calender press for thin particle board production (Bison, Germany)
a) Forming station with gamma-ray density control unit; b) Bin for misspreads; c) Paper unwinding device; d) Hydraulic for calender press; e) Calender press; f) Board transfer device; g) Extraction device; h) Braking device; i) Wide belt sander; j) Trimming and sizing saw

differ by a factor of 10. The swelling and shrinkage behavior of particle board is also determined by the intensity of compression of the particles (board density), the nature and quantity of the adhesive, and the type and quality of the wood. Changing properties by altering wood quality is becoming less and less feasible because of the continually increasing use of residual and waste wood.

For mass production of boards in the furniture industry, dimensional stability is extremely important. Assessment criteria are the absorption of moisture and the swelling in moist air and on storage underwater. The sorption test in moist air and determination of the swelling after storage underwater do not always give the same results [56] since the final swelling is not achieved as rapidly in short-term storage underwater because of the delay in water absorption due to high board density or the use of a hydrophobing agent (paraffin).

Figure 27 shows the longitudinal and thickness swelling and the water absorption of flat pressed particle board based on spruce and glued with UF resin on storage in moist air (95 % R.H.).

Mechanical properties are adapted to the use envisaged as much as possible. The strength perpendicular to the board surface (*cross tensile strength*) is an indication of the quality of the bonding. It can be influenced by the nature and quantity of the adhesive, its distribution on the particles, and the compression of the board.

The *flexural strength* (i.e., bending strength) of the board is particularly important for use in building. It can be improved by orientation of the particles (OSB) in composite board. Generally, the outer layer is more highly compressed than the central layer so flexural strength can be influenced by the sandwich effect (low-density core- and high-density toplayers).

The strength properties depend on hygroscopic properties. Figure 28 shows the impairment of flexural and cross tensile strengths by moisture absorption on storage in water.

Testing. Many selective procedures are used to test the chemical, physical, and biological resistance of particle board [57].

Europe is the area with the highest particle board production. Here, efforts have been made to harmonize the standard requirements and tests. In EN 312, parts 1 – 7, the requirements are *classified* according to areas of use. In the United States, a differentiation is made between interior and exterior use in commercial standards CS 236. In the Japanese Industrial Standard JISA 5908, boards are classified according to their flexural strengths.

To assess *formaldehyde emissions* the desiccator test is used predominantly in the United States and Japan, while the perforator and gas analysis tests (EN 129 and EN 717) are used in Europe. Large test chambers are used in different countries to measure formaldehyde equilibrium concentrations, to give a reference value for the short-term emission test. For the corresponding European standard, operating conditions for a large testing chamber have not yet been laid down. In the United States, ASTM E 1333 is followed.

Resistance of the adhesions to *climate rotation* is determined by the short-term aging tests described in ASTM D 1037 and EN 321. By using a multistage cyclic test, involving storage in water, evaporation, action of frost, evaporation, and drying, the adhesions, in particular, are assumed to be exposed to aging. Also, boiling the test object in water for several hours according

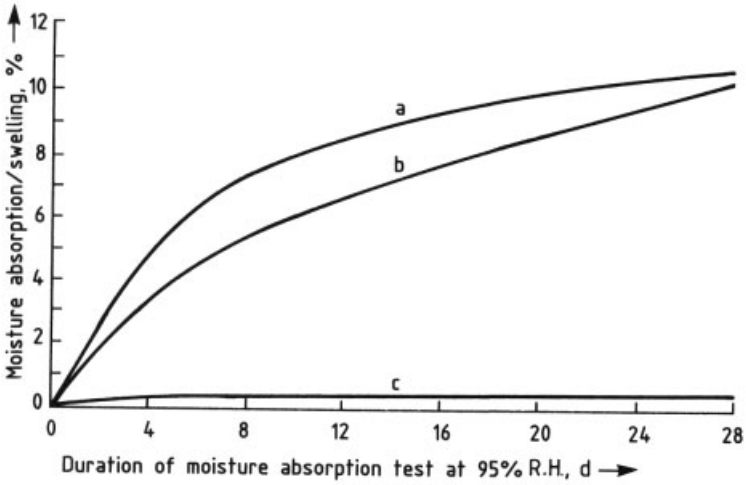

Figure 27. Moisture absorption, thickness swelling, and longitudinal swelling of 18-mm particle board (test object 10×10 cm, according to O. Liiri)
a) Moisture absorption; b) Thickness swelling; c) Longitudinal swelling

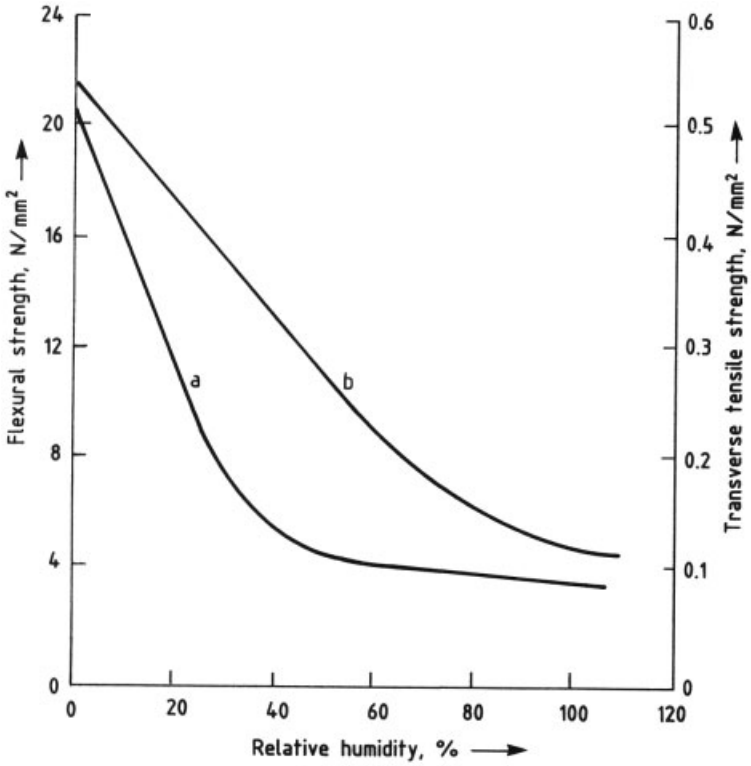

Figure 28. Variation in the flexural strength and transverse tensile strength of 13-mm particle board with moisture absorption on storage in water (according to O. Liiri)
a) Transverse tensile strength; b) Flexural strength

to DIN 68 763 (V 100) gives information on the resistance of adhesions.

Particle board for use in building often has high resistance to *biological decomposition* by microorganisms as well as high-quality adhesion (ASTM D 1413 and EN 335-3).

Widely differing test procedures are used to assess the *flammability* of particle board. Flame propagation, smoke development, and calorific value are determined according to ASTM E 84. In Germany, according to DIN 4102-8, in particular, the length of the test piece that remains undamaged is measured. British Standard BS 476-6 uses an index which is calculated from the temperature of the combustion gases.

2.3.2. Fiberboard

2.3.2.1. Development

Fiberboard consists, like paper, of lignocellulose-containing fibers. Both materials are related with regard to production and construction, and the history of fiberboard is closely associated with that of pulp and paper.

The first fiberboard was *softboard* (an insulating board). It was produced in the United Kingdom at the turn of the century – and somewhat later in the United States – from waste fibers that were unsuitable for paper production because they were too short. *Hardboard* was developed in the United States in the 1920s to utilize sawmill waste.

The production of fibers from chips is similar for all fiberboard production processes. The middle lamellae binding the fibers, which are rich in lignin, are softened by heat and moisture, and the fibers are then separated mechanically.

An important discovery was made by H. Mason, who developed the Mason cannon in 1926. In his process, the chips are first softened in a boiler under high pressure. This is followed by a sudden pressure release accompanied by a loud cracking sound, which indicates breakdown of the wood structure into fibers and bundles of fibers. These are pressed in a hot press to form hardboard without the addition of binders.

The Swede A. Asplund made the pioneering discovery of the *defibrator* in 1931. In the Asplund process chips that have been softened by steam are broken down into fibers by grinding between two disks.

The first hardboard was produced, like paper, by the *wet process* from a thin aqueous fiber suspension. The fiber mat transported to the hot press contained so much water that sieves were necessary for draining. The imprint of the sieve remains on the underside of the finished hardboard. Hence, this type of board is also known as *S1 S* (*smooth one side*).

The sieve imprint was undesirable for many applications. A board was subsequently developed that was smooth on both sides (*S2S, smooth two sides*). In this case, softboard is produced that is then compressed to hardboard between two smooth pressure plates.

Wet processes require enormous quantities of water. Pollutants, such as extractives, hemicelluloses, and lignin, are present in the wastewater. For environmental reasons the *dry process* was developed. Here, the fibers are dried, glue-blended, and pressed as in particle board production. This process gives an S2S board. For reasons of surface quality, gluing is carried out before drying. The *semidry* process lies between the wet and dry processes. Because fiber moisture is still present, sieves must be used and S1S board is produced.

2.3.2.2. Types of Fiberboard

The classification and nomenclature of the different types of fiberboard on the market seem confusing because the names either have a historic origin or represent current types of board or properties [58]. An overview is given in Table 8. Two important criteria used in classification are the production process and the board density.

Wet Process. The following types of board are produced by the wet process (S1S, density $> 800\,kg/m^3$); tempered hardboard with higher strength and water resistance (density $> 960\,kg/m^3$, thickness 2.5 – 5.0 mm); two types of mediumboard: low density ($350 - 560\,kg/m^3$) and high density ($560 - 800\,kg/m^3$); and softboard.

The density of softboard is $< 350\,kg/m^3$ (thickness 8 – 16 mm). Impregnated softboard is treated with bitumen or other additives to increase its water resistance (density up to $400\,kg/m^3$).

Table 8. Classification of types of fiberboard

Process	Board specification	Surface	Density, kg/m^3	Notes
Wet	standard hardboard	S1S	>800	
	tempered hardboard	S1S	>960	higher strength and water resistance
	low-density mediumboard	S1S	350 – 560	
	high-density mediumboard	S1S	560 – 800	
	softboard (insulationboard)		≤ 350	higher water resistance
	impregnated softboard		≤ 400	
Semidry	semidry process hardboard	S1S		
Dry	dry process hardboard	S2S	800 – 1000	adhesive PF
	medium-density fiberboard (MDF)	S2S	600 – 850	adhesive usually UF, occasionally PF or PMDI, thickness <4 mm
	low-density fiberboard	S2S	450 – 600	
	high-density fiberboard (HDF)	S2S	800 – 1000	
	MDF specialities			
	moisture resistant			
	exterior			
	flame and fire retardant			

Dry Process. Dry process hardboard is an S2S board that is normally produced by using PF resins (density 800 – 1100 kg/m^3). Medium-density fiberboard is generally produced by using UF resins (density 600 – 840 kg/m^3). Other types of fiberboard are, e.g., low-density board (450 – 600 kg/m^3); high-density fiberboard (800 – 1000 kg/m^3, thickness <4 mm); and moisture-resistant, exterior, and flame- and fire-retardant fiberboard.

Plasterboard. Besides the classical wood-based fiberboards, a fiberboard bonded with gypsum is becoming increasingly popular for use in interior furnishing for partitions and panels. Plasterboard can be produced from waste paper and gypsum from the desulfurization of exhaust gases. The production process is similar to that used for wood particle board using mineral binders (see Section 2.3.3.2).

2.3.2.3. Raw Materials

Fiberboard can be produced from all lignocellulose-containing materials with industrially utilizable fibers. However, for reasons of availability, wood is almost always used. To minimize variations in the process, only one type or a very limited number of types of wood are employed, the latter as a mixture of constant composition.

Deciduous and coniferous woods differ in their anatomical construction (see Section 1.2). Coniferous wood fibers are generally two to three times as long as those of deciduous wood (Fig. 29). The importance of the properties of the raw materials and fibers for the production process and for board quality is discussed in detail in [59]. Unlike paper, in fiberboard the fiber length is relatively unimportant. However, to a certain extent it determines the orientation of fibers in the board. Short fibers tend to be arranged vertically and longer ones horizontally. Fibers with thin walls nestle together better in the press than those with thick walls. The density of the wood used determines the thickness of the fiber mat. The lower the density, the thicker the mat, and consequently the higher are the degree of compression, the contact between fibers, and the resulting strength of the board. Dry processes are generally less sensitive to the properties of wood than wet ones. In the wet process, deciduous woods give better hardboard and coniferous woods better softboard. Coniferous woods require longer steaming times and, therefore, more energy for fiber production than deciduous woods. Mixtures of deciduous woods are mostly available in groups classified according to density. The lower the density of the deciduous wood, the longer is the digestion time. For cost reasons, efforts are made to use waste wood instead of roundwood as far as possible. Debarked chips from sawmills give good fiberboard.

Sawdust is less suitable because of the irregular particle sizes and short fibers; it requires expensive preparation. Bark is not suitable because of its mineral components and dark color. Of the lignocellulose-containing annuals, only sugar-

cane bagasse is important for softboard production.

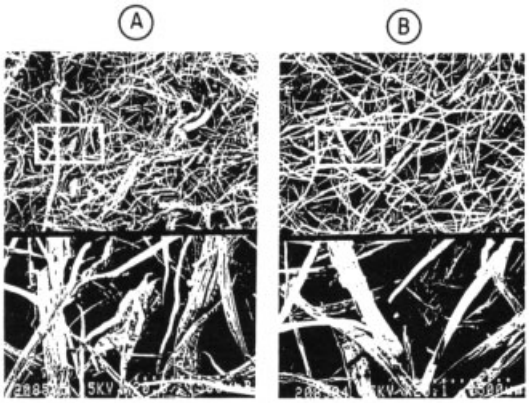

Figure 29. Wood fibers for MDF production
A) Deciduous wood; B) Coniferous wood

2.3.2.4. Binders and Additives

In the production of fiberboard by the wet process, no binders are required in principle because of the binding forces already present in the wood. However, adhesives are necessary for the dry process. Table 9 gives an overview of the main types of fiber bonding.

Table 9. Types of fiber bonding and additives used in fiberboard production, according to [59]

Process	Board type	Main binder	Auxiliary binders
Wet	insulationboard	hydrogen bonds	asphalt, starch
	S1S	lignin	PF, drying oils
	S2S, MDF	lignin	rosin
Dry	S2S	PF	
	MDF	UF, MUF, PF, PMDI	

The main purpose of additives is to improve the mechanical properties and water resistance of fiberboard. In *softboard* the fibers are bonded predominantly by hydrogen bonds. For hydrophobing, molten paraffin (ca. 1 % slack wax based on dry fiber) or asphalt emulsion is added. For *hardboard*, lignin, which is activated in the hot press, is the important binder.

S1S board is produced mainly with the addition of highly condensed, alkaline PF resins (0.5 – 2.0 % based on dry fiber), which are fixed on the fiber by the addition of $Al_2(SO_4)_3$. Paraffin (0.1 – 0.5 % based on dry fiber) and drying oils, such as linseed oil, are also added.

For *S2S board produced by the wet process*, phenolic resins cannot be used because the severe drying conditions would result in premature curing. Resins such as rosin are therefore employed. They are fixed on the fiber by lowering the pH with $Al_2(SO_4)_3$.

S2S dry process hardboard is generally produced by using PF resins (ca. 6 % based on dry fiber). For *MDF produced by the dry process*, UF resins, generally modified with melamine, are the most important binders (10 – 14 % solid resin based on dry fiber). To a small extent, mainly for special boards, MUF, PF, and PMDI are used. As with S2S boards, paraffin is used for hydrophobing.

2.3.2.5. Production Processes

Raw material processing and fiber production in the wet and dry processes are very similar, frequently even identical.

Raw Material Processing. The wood used is predominantly roundwood, also partly debarked chips from sawmills or whole-tree chips. Pieces of bark are generally undesirable. They contain minerals that shorten the service life of refiner disks and processing tools, and impair surface quality.

Roundwood is generally debarked before chip production. This can be carried out by using ring or drum debarkers.

Chips are most frequently produced by using disk chippers or cylinder chopping machines. The cutting disk of the disk chipper, arranged vertically or diagonally, is fitted with 4 – 16 blades, depending on the size. The positions of the blades determine the dimensions of the chips.

Fine chips (< 4-mm sieve mesh) and coarse chips (> 40-mm sieve mesh) are separated from normal chips by sieving on flat or drum sifters. The fine chips are incinerated, and the coarse ones are cut up further in hammer or wing beater mills. The optimum chip size is 20 – 25 mm × 20 mm × 5 mm.

To remove foreign material and mineral components, chips are washed before further processing. This is particularly important with

whole-tree chips. Washing is carried out in specially designed plants with prewashers for separating coarse components and a screw drainer for separating fine mineral particles. The chips are then stored in reinforced concrete or metal silos. Storage acts as a buffer between chip and fiber production and at the same time allows the moisture distribution to become more uniform.

Table 10. Processes for fiber production

Process (company)	Type of defibration	Conditions
Masonite (Masonite)	Mason cannon	1) 30 – 40 s steaming at 24 bar (220 °C) 2) 2 – 3 s steaming at 69 bar (280 °C) 3) explosive pressure release
Asplund (Sunds)	defibrator (one rotating grinding disk)	2 – 5 min steaming at 6 – 12 bar (160 – 195 °C) in preheater
Bauer (Sprout – Bauer)	Bauer refiner (two counterrotating grinding disks)	3 – 10 min steaming at 3 – 8 bar in the Bauer digester

Fiber Production. The aim of defibration is to break up the previously chopped raw material as carefully as possible into individual fibers and bundles of fibers. This is carried out thermomechanically. The lignin-rich middle lamellae that bind the fibers are first softened by heat and steam. The fibers are then separated by mechanical abrasion. They are often broken in the process, giving fiber fragments. The anatomical nature of the fibers and their state after defibration are important for forming the fiber mat and for the properties of the board. This applies particularly to the wet process in which the freeness (draining behavior) and the type of matting are very important.

Thermal treatment of the chips before and during defibration causes part of the hemicelluloses to go into solution. Increasing the temperature and the length of the treatment period leads to increased softening of the interfiber bonding. In the wet process, this improves the natural adhesion between fibers during board production. At the same time, the sugar content of the wastewater and the energy consumption increase, so that a compromise between environmentally friendly and economic process design and board properties is usually aimed at. Fiber production is energy intensive; it consumes more than half of the total process energy. Table 10 gives an overview of important current fiber production processes.

In the *Masonite process*, both the softening of fiber bonds and the separation of fibers are effected by steam. This technique is used exclusively in the United States by the Masonite company (wet process).

Refiners. In all other processes defibration is carried out by using grinding disks. Some defibrators operate at atmospheric and others at elevated pressure. Also, machines exist with only one rotating grinding disk (Sunds) or with two counterrotatory disks (Sprout-Bauer). Refiners operating at atmospheric pressure have the disadvantages of high energy consumption and moderate quality of the fibers produced.

The refiner now used predominantly, which operates under pressure, gives better results. The Asplund defibrator (Sunds) is a prototype. This makes use of the fact that from ca. 170 °C, separation of fibers in the middle lamella is considerably easier and is associated with a significant lowering of the energy consumption and an improvement in fiber quality.

In the defibration process (see Fig. 30), washed chips are fed into the preheater through a conical opening using a screw feeder (c). The preheater (e) is usually a vertical, slightly conical pressure container, which is continuously charged with saturated steam at 150 – 180 °C. The steam heats the chips by condensation. Blowing steam out through the opening is hindered by the plug formed in it by the chips (d).

In 1 – 3 min the chips reach a conveyor screw (g) at the base of the preheater through which they are transferred to the grinding disks of the refiner (h), where defibration takes place under steam pressure. The necessary grinding power is ca. 200 kW per tonne of dry fiber.

The important parameters affecting the quality of the fiber material and the energy consumption are the residence time and vapor pressure in the preheater, throughput quantity, defibrating pressure, distance between grinding disks (0.1 – 0.4 mm), number of revolutions per minute, state of the grinding disk, nature of the raw material, and moisture level.

Defibration is tested by determining the degree of grinding or by fractionating the fiber material. The degree of grinding is a measure of the freeness of the fiber material and determined by means of the dewatering speed of a fiber

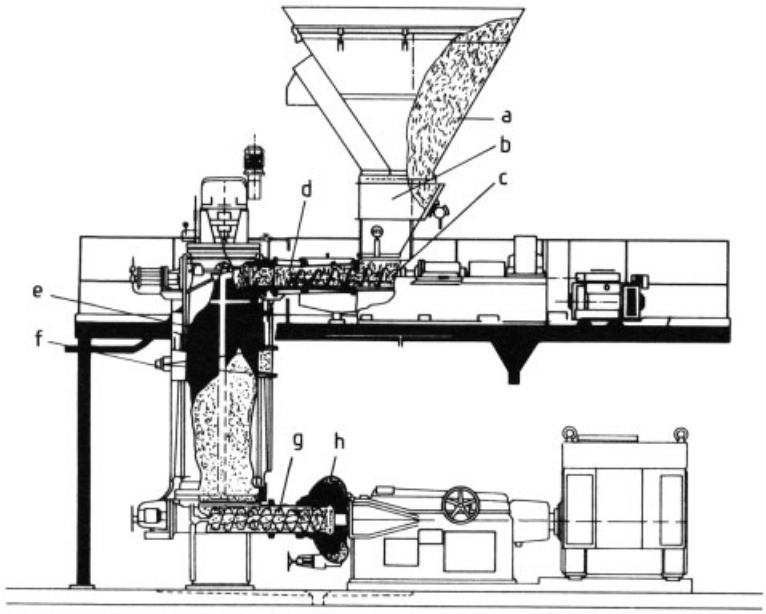

Figure 30. Asplund defibrator continuous pulping plant for fiberboard
a) Chip chute; b) Vibratory; c) Screw feeder; d) Chip plug; e) Preheater; f) Gamma-ray level controller; g) Conveyor screw; h) Asplund defibrator

suspension (standard freeness tester, defibrator-seconds, DS). In the wet process, higher degrees of grinding usually result in an improvement in board properties.

After defibration fibers are blown out through a pressure-regulated exit valve. From this stage onward the wet and dry processes differ and are described separately below.

Wet Process. A fiberboard production plant using the wet process is shown schematically in Figure 31. The fiber material is fed to silos (b) for mixing and adjusting concentration (4 – 6 %) or is further defibrated in disk refiners (f). For hard and semihard fiberboards a degree of grinding of 18 – 30 DS is aimed at, and for porous boards 42 – 70 DS [60].

A system of tanks and conveying devices is used for mixing, storing, and addition of adhesives and additives (see Section 2.3.2.4). A control panel connected to the system regulates concentration, temperature, and pH.

The fiber mat is formed on a sieve by drainage of water from the fiber suspension and subsequent sedimentation of the fibers. Mat formation was previously carried out as a batch process in tanks, but cylinder wet machines and conveyor sieve dewatering machines are predominantly used now. The main component of the latter is a continuously moving sieve. Fiber material is added through a headbox at a concentration of 1 – 2 % and is subsequently drained and compressed by means of a forming table, suction box, and wet pressing section.

To improve the surface properties of the smooth side of the fiberboard, fiber material of higher quality can be applied to the mat on the forming table through a second headbox. After the wet press the thickness of the mat is 20 –25 mm, which gives a final thickness of 3 – 5 mm. The mat is then separated into individual sections corresponding to the pressed board format of the hot press.

In the production of *softboard* or *S2S hardboard* by the wet process, the drained fiber mats are dried in single- or multideck dryers at 150 – 170 °C from moisture levels of 100 – 120 % to 1 – 4 %.

To prevent dried boards from catching fire during subsequent storage, they are cooled immediately after drying by spraying with water and aspiration of ambient air.

S1S board is further dewatered in hydraulic multi-daylight presses by hot pressing using

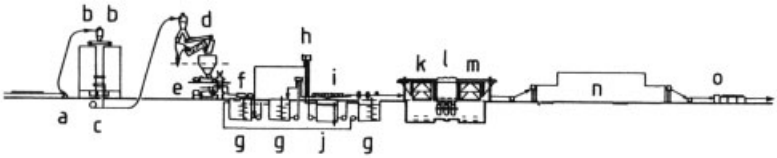

Figure 31. Schematic of fiberboard production by the wet process
a) Chopping machine; b) Silos; c) Chip sorting and disintegration; d) Chip washing; d) Defibrator; f) Refiner; g) Pulp vat; h) Quantity regulator; i) Screen; j) Wastewater vat; k) Pressure feed; l) Hot press; m) Press emptying; n) Thermal curing and climatization; o) Cutting on size

an established pressure – time program, compressed, and strengthened by curing the binders that are naturally present or have been added. For S1S, use of a draining sieve is necessary to remove the large quantities of water. For hardboard the pressing time is generally 2.0 – 3.5 min/mm at a pressing temperature of 180 – 200 °C. A typical pressing diagram consists of three phases, the compression and first high-pressure stage, drying, and the second high-pressure stage.

An improvement in the properties of fiberboard can be effected by thermal aftertreatment in chambers or continuously in tunnels or by impregnating with drying oils (see Section 2.3.2.4). Treatment results in a more uniform moisture distribution in the board as well as better dimensional stability.

Dry Process. The production of MDF is shown schematically in Figure 32. After defibration (g) the fibers pass down a blow line in which glue (see Section 2.3.2.4) is added and then reach the fiber dryer (i). Alternatively, they can be blown directly into the fiber dryer (i) by making use of the vapor pressure present in the refiner. In this case they are subsequently glue-blended in high-speed centrifugal machines, as in particle board production. In almost all plants, glue blending occurs before drying, thus avoiding spots of adhesive on the board surface that impair its quality. Paraffin added to improve hydrophobic properties is generally charged in the molten state to the lower conveyor screw of the refiner.

Tube dryers are now used exclusively for drying the fibers. They are mostly two-stage dryers with tube lengths of 80 – 100 m and diameters of 1.3 – 1.6 m. The air temperature at the entry is 140 – 170 °C. Because of the high air speed (ca. 30 m/s) and the small dimensions of the fibers, the drying time is only a few seconds [61]. The moisture level of the fibers is adjusted to 9 – 11 % by regulating the exit temperature of the dryer.

The fibers reach the forming station (m) via cyclones and fiber hoppers (j). The fiber mat is formed either mechanically by using a snowfall felter or pneumatically by using a pendistor on a vacuum screen conveyor.

The uniformity of distribution of the fiber material over the surface of the screen significantly affects fiberboard properties.

The mat coming out of the forming station has a very low bulk density and contains a large quantity of air, which must be removed before hot pressing. To decrease the size of the press opening, shorten the closing times of the hot press, and improve the transportability of the mat and the surface quality of the finished board, the mat is precompressed in continuously operating belt prepresses. Mat density is thus doubled or trebled.

Hot pressing of the fiber mat occurs either batchwise on multi-daylight presses with preliminary separation into mat sheets or, as is now increasingly common, continuously on double-belt presses or calenders for thin boards (up to 8-mm thickness). The presses are the same as those used for particle board production (see Section 2.3.1.5). To heat the fiber mat more rapidly and shorten the pressing time, the fiber mat is often subjected to high-frequency preheating before entering the hot press. Pressure control in MDF production differs from that in particle board production. Figure 33 gives an example of the pressure program.

The bulk density profile (perpendicular to the board thickness) and thus the properties of the board can be widely varied through the type of compression and the temperature control. High compression in the outer zones and as uniform a density as possible in the inner zones are aimed at (see Section 2.3.2.6).

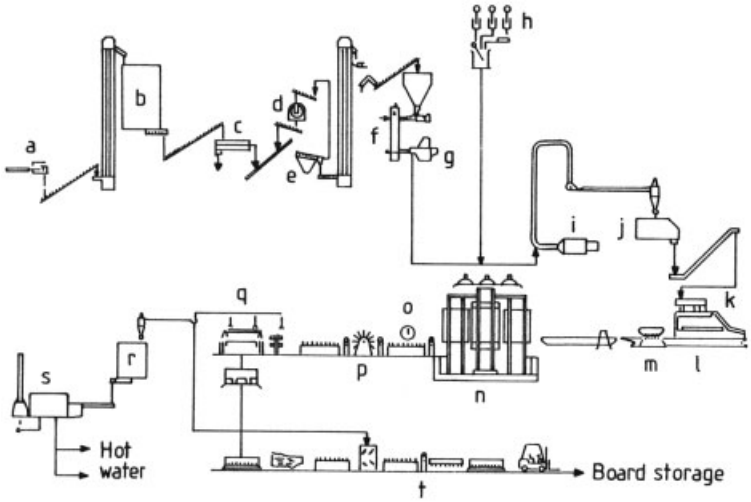

Figure 32. Schematic of MDF production
a) Chopper; b) Chip silo; c) Cleaning; d) Fine chopper; e) Sieve; f) Digester; g) Defibrator; h) Glue preparation; i) Dryer; j) Dry fiber hopper; k) Prepress; l) Spreading machine; m) Forming line; n) Hot press; o) Weigher; p) Cooler; q) Trimming; r) Dust silo; s) Heating boiler; t) Sanding machine

2.3.2.6. Properties, Testing, and Uses

Properties. The main factor affecting almost all properties of fiberboard is its *bulk density*. This determines the flexural strength and modulus of elasticity, and affects the transverse tensile strength and swelling properties. Besides bulk density, the bulk density profile perpendicular to the board thickness is most important in the case of MDF. S1S hardboard produced by the wet process has a highly compressed upper side and a less dense underside (sieve side). S2S boards formed by the wet process have a symmetrical density profile with a minimum in the center of the board and maxima at a certain distance from the board surface. This gives the boards high flexural strength [59].

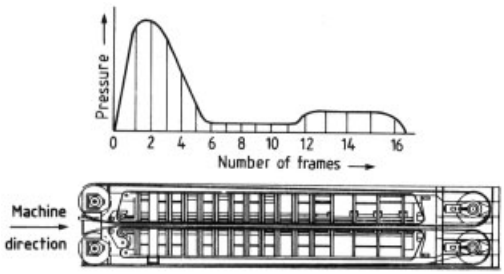

Figure 33. Pressure program for MDF production (Siempelkamp, Germany)

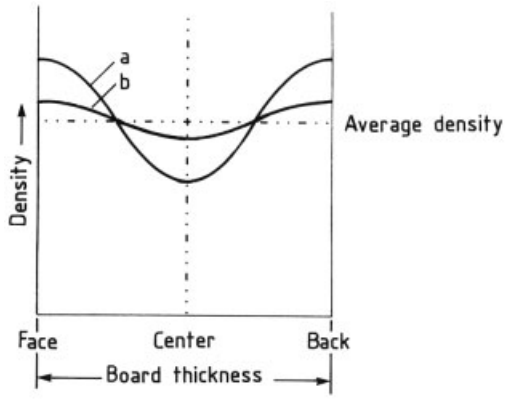

Figure 34. Examples of density profiles of MDF
a) High density contrast; b) Low density contrast at the same average density

Figure 34 shows two different density profiles of MDF board. Highly compressed outer zones with a high density relative to the center, and the resulting high flexural strength and low transverse tensile strength, are achieved by closing the press rapidly. A more uniform density distribution with smaller density differences between the outer and middle layers, lower flexural strength, and higher transverse tensile strength is achieved by closing the press slowly [62]. In practice, a compromise between the two extremes is usually aimed at.

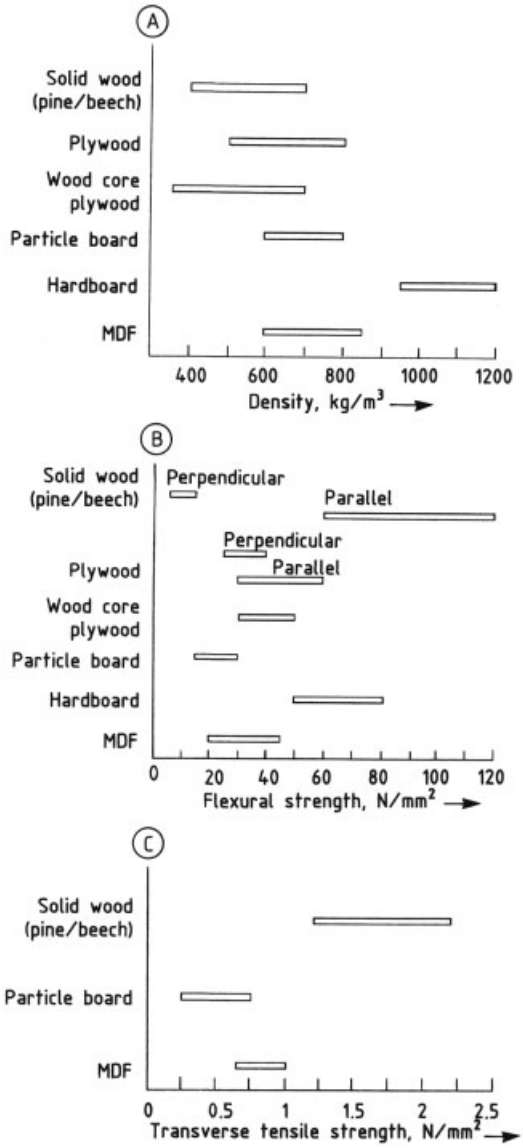

Figure 35. Comparison of density (A), Flexural strength (B) and transverse tensile strength (C) for wood-based construction materials

Quality requirements and test methods for fiberboard are laid down in a large number of international standards. The standards for the United States and the EC are given below.

United States	
Insulationboard	ANSI/AHA A 194.1 – 1985
Basic hardboard	ANSI/AHA A 135.4 – 1982
Prefinished hardboard	ANSI/AHA A 135.5 – 1988
Hardboard siding	ANSI/AHA A 135.6 – 1989
MDF	ANSI/AHA A 208.2 – 1990
EC	
MDF, Industry Standard	Euro MDF Board (EMB) (Gießen – 1990)
Fiberboards (dry process)	EN 622 – 1993

Within the framework of production monitoring and quality assurance, the following fiberboard properties are generally tested:

1) Tensile strength perpendicular to the board surface, either dry or after the wet cyclic test
2) Flexural strength and flexural modulus of elasticity
3) Swelling after 24 h or after the wet cyclic test
4) Formaldehyde emission

The following properties are tested less frequently: screw withdrawal, dimensional stability, sand content, and surface absorption.

A comparison of the strength properties of hardboard and MDF with those of solid wood and particle board is given in Figure 35. The advantages and disadvantages of MDF compared with particle board, veneer plywood, and table board are given in Table 11.

Table 11 shows that MDF does not have any advantages over particle board with a fine covering layer with regard to finishing by veneering, laminating, and liquid coating. However, in all further processing in which the homogeneous fine fiber structure in the interior of the board is utilized, MDF has advantages over the other types of board. The edges can be milled, polished, and coated.

Uses. The main areas of use for fiberboard are as follows:

Insulationboard is used in interior and exterior applications, mostly as parts of walls, and for ceiling paneling, frequently in the form of accoustic panels.

Hardboard produced by the wet process generally consists of thin boards. It is employed mainly in buildings for inner and outer wall paneling. It is also used in the production of furniture or doors or molded for use in motor vehicle interiors.

Table 11. Physicomechanical properties and ease of working of MDF in comparison with other wood-based construction materials [63]

	MDF in comparison with ∗		
	Particle board	Veneer plywood	Wood core plywood
Density	≈	≈	≈
Flexural strength	−	+	+
Transverse tensile strength	−	+	+
Modulus of elasticity	−	+	+
Linear expansion	+	+	+
Screw holding ability			
Surface	≈		+
Edge	−		+
Ease of molding			
Surface	−		−
Edge	−		−
Ease of varnishing			
Surface	≈	≈	≈
Edge	−		−
Ease of laminating	≈	−	−
Ease of veneering	≈	−	−

∗ − less favorable than MDF; ≈ same as MDF; + more favorable than MDF.

MDF. Because of its excellent edge and surface properties and its good workability, MDF is used mainly for furniture production.

2.3.3. Wood-Based Materials with Mineral Binders

Wood-based materials with mineral binders consist of 30 – 90 % of a hardened inorganic binder and 10 – 70 % of wood particles. Because of the high binder content, these materials are heavy and particularly suitable for use in building. The most important mineral binders are cement, gypsum, and magnesia (magnesite) [64].

Mineral-bonded boards are characterized by low thickness swelling through the action of water. Cement- and magnesia-based boards are highly resistant to mold and termites [65]. Cement-bonded boards are used mainly for exterior applications because of their high resistance to weathering. All mineral-bonded wood-based construction materials are classified as building materials of low flammability because of their high binder content.

2.3.3.1. Cement-Bonded Boards

Cement was the first mineral binder used in the industrial production of boards. The originally long hardening times of this hydraulic binder have been shortened greatly by technical measures [66]. The setting time depends not only on the reactivity of a particular cement quality, but also on its compatibility with different types of wood. Wood components, such as free sugars, glucosides, and tannins, delay the hydration of cement and therefore impair the setting process [67]. Cement-bonded wood-based boards are produced predominantly by the wet or semidry processes.

The original production process developed by Elmendorf is still used in different forms [68]. In the process, wood particles, cement, additives, and water are mixed, molded, and compressed. To shorten the hardening times, stacked boards are fed into a heated tunnel under pressure. The boards can generally be removed from their molds after 6 – 8 h and then conditioned for 12 – 18 d in storage until the final strength is reached.

Finished boards consist of ca. 60 wt % cement and additives, ca. 20 wt % wood, and ca. 20 wt % water. The ratio of total water to cement determines board properties such as compressive and flexural strength.

2.3.3.2. Gypsum-Bonded Board

For many years, gypsum has been an indispensable building material. Since the actual strength of pure gypsum is insufficient for the production of board material, it is reinforced with wood fibers, wood wool, or flakes. The suitability of

gypsum as a binder is based on the de- and rehydration processes. The water of crystallization is partially removed from mined or industrially produced gypsum by calcination. In the setting process the water added becomes chemically bonded as water of crystallization. The quantity of water that is necessary for complete rehydration depends on the quality of the gypsum and is generally 14 – 21 %. Excess water is not bonded and must be removed from the system.

Wet Production Process. In the production of plasterboard by the wet process, a large excess of water is used. In a process used by the Siempelkamp company, excess water is first removed on a machine wire web during formation of the moldings and subsequently through drying of the boards [69]. The Knauf company in Germany carries out an interesting variant of the wet process in its production of multilayer boards by prefabricating the different layers separately. A particularly imaginative version involves forming and pressing fiberboard with uncalcined gypsum. The boards are calcined in an autoclave, and water is then added so that the desired strength can be obtained by the formation of gypsum hydrate [70].

Semidry Production Process. The semidry process is particularly suitable for the production of gypsum particle board. Only a small excess of water is used (ca. 30 %) so pourable particles can be processed. By decreasing the proportion of water, the energy balance of the gypsum particle board production is improved because less free water needs to be removed from the board by drying.

Industrially produced boards have a moisture content of 15 % after pressing. They are then dried to an equilibrium moisture level of 2 %. The bulk density of the board is ca. 1150 – 1250 kg/m^3. The pressing times for gypsum are generally shorter than those for cement. Further shortening of the setting time is possible [71].

2.3.3.3. Magnesia-Bonded Boards

Magnesia-based binders have been known as Sorel cement for more than 100 years. By calcination of magnesite or dolomite at ca. 800°C, magnesium carbonate ($MgCO_3$) is converted into magnesium oxide (MgO). The latter undergoes a highly exothermic reaction with water and magnesium chloride or sulfate to give an extremely hard material. This binder system has been used for a long time for the production of wood wool boards (trade name Heraklith). Because of the high solubility of magnesium salts, the water resistance of Sorel cement is not particularly great. However, the highly exothermic nature of the setting reaction is very advantageous. Short hardening and pressing times are thus possible at low pressing temperature (100 °C). For the production of magnesia-bonded particle board, ca. 70 % binder and water and ca. 30 % wood chips are used. The binder consists of 5 mol MgO, 1 mol $MgSO_4 \cdot 7\,H_2O$, and 15 mol water. Normally, the water is applied to the wood chips; magnesium sulfate is added as a concentrated aqueous solution; and then magnesia is mixed in as a powder. At 100 – 150 °C the pressing time is ca. 10 min. The 18-mm-thick boards have a density of 1050 – 1150 kg/m^3, a transverse tensile strength of 0.65 N/mm^2, and a flexural strength of 13 N/mm^2. The thickness swelling of the boards on storage in water is < 2 %. The economic importance of magnesia-bonded board is currently low.

2.4. Surface Treatment

Wood-based panels are used mainly for interior furnishing and furniture production. Thus the appearance and decorative function of these materials are particularly important. Solid wood – both natural wood and veneers – has a very high aesthetic value. Surface treatment is aimed at increasing the decorative value as much as possible and improving properties of application.

Many techniques are known for producing certain effects in the surface treatment of solid wood or veneers. The nature of the surface can be altered mechanically by planing, grinding, polishing, or sanding, and chemically by bleaching, staining, and varnishing. Wood components in the surface are treated with bleaching or staining agents so that the natural color disappears completely or is rendered much more intense.

Materials based on glued wood particles are generally surface treated. Most are veneered, varnished, laquered, or coated with films, sheets,

or laminates. The most important requirement for surface coating is that the surface of the material to be coated should be solid and smooth.

2.4.1. Veneers

The oldest process used for the finishing of wood-based construction materials is the application of veneers.

Today, peeled veneers, in particular, are used mainly for decorating furniture and panels. Chemical stains are often used to enhance natural effects and strengthen color contrasts.

The production of veneers and their processing to give decorative surfaces is described in Section 2.2.2.

2.4.2. Liquid Paints and Coatings

Coating is the most well known type of surface finishing. The coating technique used is determined by the nature of the surface. For glued particle materials such as particle board, MDF, hardboard, wafers, or OSB, which can have very absorbent surfaces, a thick coating is necessary. Besides absorbency, the surface structure of the wood-based construction materials must be taken into account. Here, MDF has advantages over boards made of coarser particles, such as particle board and OSB.

Technological parameters of the carrier board, such as the density of the outer layers, density profile, adhesive content, water absorption of the outer layers, surface hardness, face strength, surface macrostructure, porosity, and sanding quality, have a significant effect on the ease of varnishing. Tests on particle board show that for three-layer boards the outer layer should be made of fine particles so that the surface is as closed and smooth as possible [72]. The quantity of binder in the surface layers should be at least 12 %. The density of the surface layers should be 950 g/cm^3 (max.) and the face strength not less than 1 N/mm^2 [73].

In the industrial coating of wood-based construction materials, attempts are made to compensate for unavoidable deficiencies in the surface quality of the boards by pretreatment with primers or priming films (see Section 2.4.3.1).

On modern coating production lines, finely sanded particle board is cleaned and primed, and a colored undercoat is applied, followed by the actual coating paint. Transparent or pigmented priming or coating systems that can be cured by air or radiation are used. These include:

1) Photochemically curing coating systems (e.g., UV-curing systems),
2) Microwave or electron-beam curing systems, in particular for highly pigmented coatings
3) High-solid coatings with a solid material content of > 80 %
4) Application of prepregs based on unsaturated polyesters
5) Waterborne paints based on water-soluble or dispersed polymer systems

2.4.3. Coating with Sheets, Short-Cycle Films, and Decorative Laminates

Coating with foils, short-cycle films, and decorative laminates is extremely important in industrial finishing. In Western Europe, which has the highest wood-based construction material production worldwide, more than 40 % of the surfaces were coated with these materials in 1992 [74].

2.4.3.1. Coating with Foils

Foils are thin, preformed surfaces that are glued to a wood-based panel carrier. The following types of overlaying exist:

Thermoplastic Foils. The basic material consists of poly(vinyl chloride), polyolefins, or polyester resins, which can be mass dyed, subjected to decorative printing, and embossed on the surface.

Priming foils consist of paper carrier web, which is impregnated by submersion in an aqueous amino resin solution and dried without touching in a suspension type dryer to a residual moisture level of 3 – 7 %; 60 – 120 g/m^2 paper is used. The resin content of the paper is relatively low – 60 % based on the weight of the paper.

Highly penetrating UF resins in aqueous solution are used for impregnation. These undergo extensive acid-catalyzed condensation during the drying process. Priming foils serve as a basic material for further surface treatment (e.g., printing or coating).

Decorative Coating with Foils. Decorative sheeting – an economically important surface material – also consists of paper that is impregnated with a mixture of amino resin and polyacrylate dispersions. For the production of thin sheeting, ca. 40 g/m^2 paper is used, and for thick sheeting, 80 – 120 g/m^2. The papers are usually printed before impregnation. After the paper web has been dried in a lay-on-air dryer, the surface is varnished by using doctor blades. The varnish layer of 10 – 20 g/m^2 is hardened immediately in a drying area. The sheeting is dispatched for further processing in the form of rolls.

The impregnating and drying processes place high demands on the wet tenacity of the paper. The absorption of water by the paper, its longitudinal swelling, and subsequent shrinkage in the drying tunnel are compensated for by electronic control of the web. Figure 36 shows a impregnating machine with an integrated varnish applicator (doctor blade) developed by the Vilts company in Germany.

The temperatures (100 – 180 °C) in individual drying areas can be controlled separately. The quantity of resin (UF resin – acrylate dispersion) applied to the paper is 40 – 70 % of the weight of the base paper. On applying the wood-graining decorative print, silicone-based wetting inhibitors can be added so that defects obtained during subsequent varnishing look like wood pores.

Edge Coating. In the furniture industry, edge coatings are required that fit in with the decorative surface sheeting. For edge sheeting, paper with a weight of 180 g/m^2 or more is used. The material itself is more rigid than that used for coating smooth surfaces, so that nonuniformity in the edges to be coated can be compensated for more easily. Paper-based edge sheeting is produced in various qualities. For less demanding applications, the printed paper is impregnated with an aqueous mixture of UF resin and polyacrylate dispersion. If greater resistance is required, etherified MF resins must be used. On impregnation, the paper absorbs 25 – 60 % of its weight of resin.

The resin is cured by using acid catalysts in drying tunnels heated with hot air. Paper-based edge sheeting is generally coated with varnish.

Application of this thermoset edge sheeting to a carrier board is often carried out with simultaneous forming of the edge material (soft forming). Edges with a purely thermoplastic basis, such as poly(vinyl chloride), polyolefin, or polyurethane, are also produced for these applications.

Sheeting and edge strips are fixed to the wood-based construction material with aqueous amino resins, poly(vinyl acetate) adhesives, melt adhesives, or solvent-containing adhesives. Melamine – formaldehyde resins and melt adhesives are frequently applied to the underside of the sheeting as a dry adhesive layer during sheet production. The sheeting can then be laminated onto the carrier board through the action of pressure and temperature in the press.

2.4.3.2. Decorative Films (Short-Cycle Films)

The application of quick-drying decorative films based on melamine – formaldehyde resins is the economically most important process for coating fiber- and particle board. In Europe in 1993, more than 1.3×10^9 m^2 of wood-based construction material surface was formed by this process.

An important feature of this coating process is the use of low pressure (1.2 – 2.5 N/mm^2) at 140 – 230 °C and shorter pressing time (9 – 40 s). This allows direct coating of pressure-sensitive materials.

The high pressure and long pressing time necessary for conventional processing of resin-impregnated papers for laminate production (see Section 2.4.3.3) would lead to undesirable deformation and damage to the structure of the carrier in coating particle board.

The development of melamine – formaldehyde resins that have a sufficient melt flow at low pressure and can thus form a surface was an important prerequisite for the direct press coating of particle board. The company Th. Goldschmidt succeeded in modifying melamine – formaldehyde resins by incorporating elasticizing agents so that crack-free surfaces could be produced [75].

In principle, polyester resins or prepolymers based on diallyl phthalate are also suitable. These resins allow application of lower pressure (e.g., 1.0 N/mm^2). However, for impregnating, they must be used in solvents such as acetone or methyl ethyl ketone. Also, the surface resistance

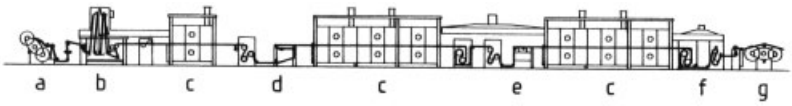

Figure 36. Schematic of an impregnating machine with contact-free drying and integrated coating applicator (Vilts, Germany) a) Paper roll exchange; b) Impregnating station; c) Drying tunnels; d) Applicator; e) Coating station; f) Sheet cooling; g) Sheet rolling and exchange

and hardness of the much cheaper melamine–formaldehyde resins are not achieved when diallyl phthalate and polyester resins are used.

Highly filled papers – either of one color or with decorative printing – with a weight of 60–130 g/m^2 are used mainly for impregnation. To form a closed surface, 100–140 wt % solid melamine resin, based on the weight of paper, is required.

More strongly elasticized MF resins are now generally used for impregnating. In these resins the molar ratio of melamine to formaldehyde is 1:1.5–2. Polyols, *p*-toluenesulfonamide, caprolactam, or sugar, for example, is used as the elasticizing component. Small quantities of wetting agents, mold release agents, and catalysts are also added to the impregnating liquor. The aqueous liquor has a solids content of 48–55 %. The impregnation of paper is usually carried out in two stages to obtain a more uniform application of resin. This is achieved by predrying after the first stage. Applying the necessary quantity of resin to the paper from a 55 % liquor in a single step is difficult.

Two-stage impregnation also offers the possibility of replacing the relatively expensive MF resin partially or completely in the first stage by a more economical UF resin. After the first impregnation step, the papers are dried to a residual moisture level of 10–15 %. They are then coated once more on both sides with an MF-resin-based liquor using doctor blades or gravure rollers. Subsequent drying in a suspended dryer to a residual moisture level of 5.5–7 % gives films that are supplied for coating particle board in the form of rolls or stacked sheets. During drying the paper web is led through a tunnel of drying areas on a cushion of hot air introduced from above and below by a system of jets. Figure 37 shows the schematic arrangement of the jets and the progress of the paper web through a drying tunnel. Figure 38 shows a machine for impregnating and drying decorative films by the two-stage process (lay-on-air dryer).

The amino resin absorbed by the paper undergoes further condensation during drying, but not complete cross-linking. Final curing of the resin takes place only on coating the fiber- or particle board carrier.

Particle board used for coating must satisfy particular requirements. Generally, three-layer particle board is used, which has a highly compressed, fine outer layer with high face strength. The density of such boards is 700 kg/m^3, and the face strength is at least 1.2 N/cm^2. For more highly stressed surfaces, two films are sometimes pressed at the same time, so that possible nonuniformity of the carrier material can be compensated for more effectively.

In the *Unibord process*, decorative films impregnated with melamine resin are placed on the particle mat during actual particle board production so that a coated board can be formed in a single pressing process [76]. The process did not achieve economic success because the indispensable closed-circuit cooling process in the press requires too much energy and prolongs the cycle times of the plant.

In developing new processes for the continuous production of particle board, the idea of direct coating of boards during production is expected to be reinvestigated.

2.4.3.3. Decorative Laminates

Decorative laminates are produced by pressing several layers of paper impregnated with synthetic resin in heated, hydraulic presses at high pressure. The resin melts and hardens to give a composite material in which the cellulose fibers of the paper have a reinforcing and strengthening function. The number of resin-containing paper layers determines the laminate thickness.

The central layer of a laminate is generally formed from several sheets of kraft paper impregnated with phenolic resin. This central layer is coated with an underlay, a decorative, and an

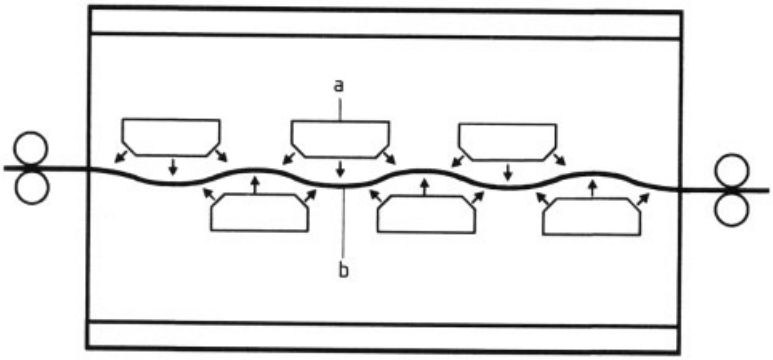

Figure 37. Schematic arrangement of the air jets and the course of paper through a drying tunnel
a) Air jets; b) Paper web

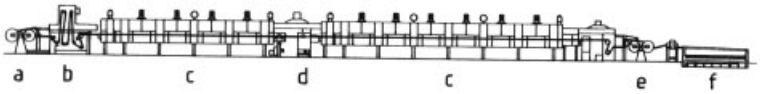

Figure 38. Schematic of a production plant for decorative film according to the two-stage process (Vilts, Germany)
a) Paper roll exchange; b) Impregnating station; c) Drying tunnel; d) Screen roller applicators; e) Paper rolling device; f) Transverse cutter and sheet deposition

overlay paper; these three layers of paper are impregnated with MF resin. The use of MF resin has the advantage that any decorative coloration can be applied, because melamine – formaldehyde resins become absolutely colorless and transparent on pressing. Phenol – formaldehyde resins have a red-brown color and are thus not suitable.

High-Pressure Laminates. Production of high-pressure laminates (HPL) is a batchwise process. The pressure in the hydrualic multidaylight presses is 80 – 120 bar, and the pressing temperature is 140 – 180 °C. After pressing, the undersides of the layered sheets are sanded to facilitate subsequent gluing of these laminates to a carrier board. In the pressing process, several laminates are usually produced in one opening. Plates for modifying the structure allow a wide range of surface forms to be made.

To compensate for tension in the composite material and, in particular, to form a glossy surface, HPL is cooled under pressure to a temperature of ca. 80 °C before it is released from the mold. The pressing cycle times can be 30 min or more, depending on the thickness and number of laminates. Production of HPL is cost intensive because of the long pressing times required.

Continuous pressed laminates (CPL) are produced on double-belt continuous presses (Fig. 39) at pressures up to 50 bar with resin-impregnated papers from a roll.

Unlike the batch process, there is a continuous increase in pressure in the material after passing the entry of the continuous press. For technical reasons CPL can be produced economically only in a thickness range of 0.4 – 1.2 mm.

The covering of wood carrier boards with the prefabricated laminate is effected by gluing under pressure, as in the production of veneers. Decorative laminates are rigid, very stable composite materials. They can also be glued to wood carrier boards with low requirements as regards surface quality. Amino resins, poly(vinyl acetate), or melt adhesives are used for gluing.

Although decorative laminates are regarded as thermoset surface materials, thermoformability can be achieved by employing certain measures. Postforming laminates are obtained by elasticizing the impregnating resin used and incomplete curing (cross-linking) of the resins during high-pressure laminating. Local heating using IR radiation or contact heat permits a single permanent forming even with a small radius of curvature and angles greater than 180°.

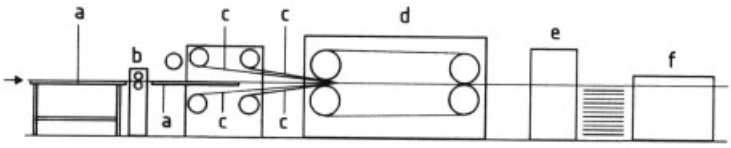

Figure 39. Schematic of a double-belt press for continuous lamination of particle board (Hymmen, Germany)
a) Boards; b) Cleaning; c) Saturated papers; d) Double-belt press; e) Saw; f) Double end tenoner (groove and tongue)

2.4.4. Properties, Testing, and Uses

Coating wood-based construction materials aims at both the production of a decorative effect and the improvement of performance characteristics. The coating protects these materials from chemical and physical attack.

Properties. The following properties are important for assessment of a surface:

1) Mechanical stability: scratch resistance, impact strength, abrasion resistance, and crack resistance
2) Resistance toward physicochemical attack by water and steam (loss of gloss), food and chemicals, weathering, heat, radiation (sunlight), and fire

Test Standards. Besides standards for the assessment of painted surfaces, a number of national and international regulations exist for the classification, testing, and application of thermosets, which are widely used for coating wood-based materials. Classification/specification and determination of properties of HPL are listed in ISO 4586-1 and ISO 4586-2, respectively.

In Europe, particularly in Germany, coating of wood-based materials with short-cycle melamine films is widely used. Special standards were developed for assessment of these surfaces [77].

Uses. The surface coating of wood-based materials has been the major impetus for mass processing these materials. More than half of the fiber- and particle board produced worldwide is currently used in the mass production of furniture and building materials. Their application in wall and ceiling panels has increased considerably. Decorative laminates have revolutionized the entire kitchen furniture industry. Laminates that have been rendered highly abrasion resistant by incorporation of corundum, are being increasingly used as upper surfaces of floors. Plywood and particle board that are laminated with papers impregnated with phenol – formaldehyde resins can be repeatedly used for concrete shuttering. The use of short-cycle films in the processing of particle board moldings and profiles has also proved successful. Here, the coating and the compression molding are applied simultaneously (Werzalite process) [78].

2.5. Environmental and Toxicological Aspects

2.5.1. Wood Dust

In the production and processing of wood-based materials the formation of fine dust particles cannot be avoided. Suspicion has existed for a long time that prolonged occupational contact with wood dust increased the risk of development of adenocarcinomas in the inner nose. According to Acheson workers in the wood industry have a thousandfold greater risk of developing nose cancer than other workers; he based his theory on statistical investigations in the wood industry at High Wycombe (United Kingdom) from 1956 to 1965 [79]. No clear explanation has been found as to why wood dust should be carcinogenic. Natural wood components, materials used in working and processing wood, metabolites of wood fungi, pyrolysis products from processing, or simple mechanical irritation could be responsible for the development of adenocarcinomas [80]. Wood components particularly those of hard deciduous woods, have been suggested to be mutagenically carcinogenic [81]. Besides a large number of mutagenicity tests with wood extracts, epidemiological investigations of nose cancer in various countries, in particular, have led to the formulation of preventive measures. In many countries, adenocarcinomas of the nose have been recognized as occupational diseases of workers in the wood industry.

2.5.2. Emissions

In the production of wood-based materials, processes are used in which the main components, such as wood, adhesives, varnishes, paints, and coatings and other additives, are heated. In these processes, substances are formed or liberated that constitute emissions from the production plant. Emissions from finished materials are also sometimes detected. These are a nuisance and impair the utility of the material.

2.5.2.1. Emissions during Production of Wood-Based Materials

Dust. In large wood processing plants such as fiber- and particle board factories, considerable quantities of dust are produced in cutting and sanding processes. The dust produced from sanding machines is also polluted with adhesives and additives. Removal of dust from waste gases is mandatory. Dust separation is carried out in filters, and the dust is incinerated to produce energy.

Emission of Wood Components. Emissions are also produced during the drying of wood, which depend mainly on the intensity of the drying process. Waste gases with intense odors are liberated from wood along with water vapor. These consist of wood components such as pinenes, terpenes, phenol derivatives, aldehydes, and alcohols. The quantity of these substances can be reduced by mild drying at lower temperature, or they can be eliminated by using gas scrubbers with up to 60 % efficiency.

Adhesives. For the production of glued wood-based construction materials, several million tonnes of adhesive resins are used per year.

Formaldehyde. Condensation products based on amino and phenol – formaldehyde resins are the most important adhesives economically. On hot pressing of glued particles, fibers, or veneers, formaldehyde is emitted. In the waste gas from a 12-daylight press in a particle board factory, 13 kg of formaldehyde per hour was measured for UF adhesives and only 0.6 kg/h for PF adhesives. The quantity of formaldehyde produced in the pressing process depends on the molar ratio of urea to formaldehyde in UF resins, pressing conditions, the moisture level of the molding, and the quantity of catalyst used [82].

Formaldehyde is a very reactive gas with a pungent odor, which can be detected even in very low concentrations. The odor threshold ranges from 0.06 to 1.2 mg per cubic meter of air (0.05 – 1 ppm) and varies from one individual to another.

The hazard potential of formaldehyde to humans is due to irritation of the eyes and mucous membranes of the respiratory passages and to skin allergies (→ Formaldehyde, Chap. 10.).

After two American animal experiments in 1980 showed that inhalation of high concentrations of formaldehyde over a long period can cause nose tumors in rats, extensive research was initiated worldwide into its possible carcinogenicity. The results of these and numerous epidemiological studies are now available. In animal experiments, varying results were obtained, from which carcinogenicity in humans could not be derived. In 1994 the UNO/WHO-IRAC had no information to support the assumption of carcinogenicity. As a precaution, emission and imission values for formaldehyde and maximum workplace concentrations have been officially regulated in almost all industrialized countries.

On using PF resins, a low emission of *phenol* or (hydroxymethyl)phenols must be expected. The more strongly precondensed the PF resin is, the lower are the emissions on hot pressing.

In the processing of PMDI monomeric components (MDI) can be emitted during pressing. The high reactivity of this compound with water usually leads to immediate, nontoxic polyurea formation.

During the drying of resin-impregnated paper for coating wood-based construction materials, formaldehyde, phenols, and sometimes additives are emitted. In modern plants, waste gases from the dryers are usually purified by incineration. Scrubbing these gases would be time consuming and expensive since 15 000 – 20 000 m^3 of waste gas are produced per hour per plant. If melamine – formaldehyde resins are used for impregnation, waste gases from the dryer contain 0.3 – 0.6 g formaldehyde per kilogram of solid resin. Under the same conditions, UF resins emit up to three times this quantity. Lamination of resin-impregnated and dried papers to

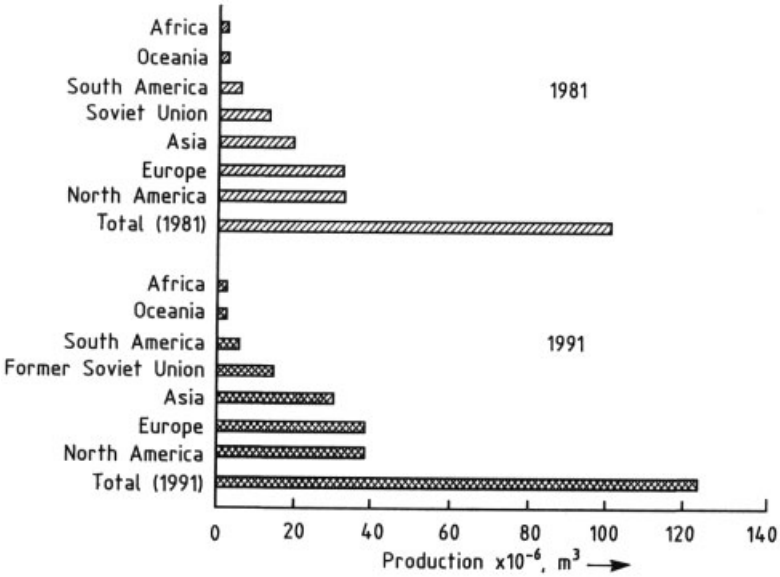

Figure 40. Worldwide production of wood-based panels in 1981 and 1991

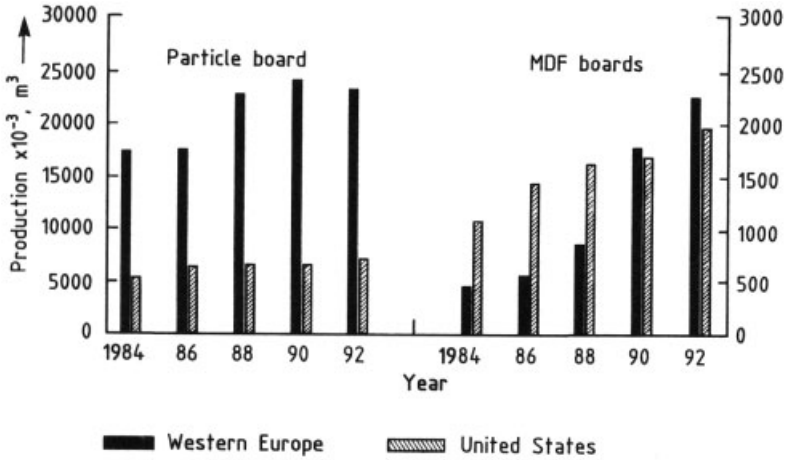

Figure 41. Development of particle board and MDF production in Western Europe and the United States from 1984 to 1992

the wood-based construction material generally results in low formaldehyde emissions.

2.5.2.2. Emissions from Finished Materials

The emission of formaldehyde from wood particle board was first reported in 1962 [83]. The emissions are caused by the adhesive used, whereby formaldehyde is liberated by hydrolysis of the condensation product [44]. In the meantime, a large number of publications on this subject have appeared [84], in which the causes are determined and possibilities for lowering emissions are described [55, 85]. Many methods of measurement have been developed to adhere to the limiting values stipulated by national authorities. In Europe the perforator method (EN 120) and the gas analysis method (EN 717-2) are used as process controls to monitor emissions from particle board.

Through international cooperation, conditions for a chamber test have been worked out. In this test, the formaldehyde concentration in ambient air is measured as an equilibrium con-

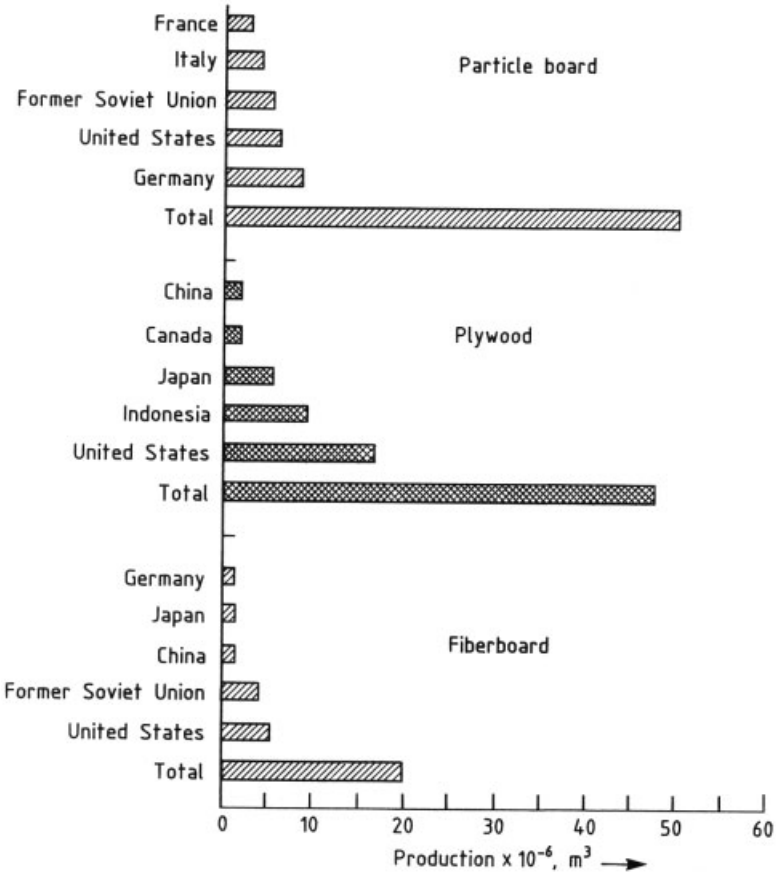

Figure 42. Production of plywood, particle board, and fiberboard by the five most important producer countries in 1991

centration, and the emissions of the material are thus assessed.

In many countries, wood-based construction materials are divided into emission classes E1 – E3 according to their formaldehyde-emitting capacity. These classes represent certain ranges of values determined by the perforator method. For example, for boards in emission class E1, the consumer need not fear any emissions on proper use of the board.

No complaints of emissions from boards bonded with *phenolic resins* have been reported thus far. Values for emissions from *PMDI-bonded boards* have not yet been published.

2.5.3. Disposal

In the production of modern wood-based construction materials, considerable quantities of nonwood products are frequently used. The important materials remaining in the wood are adhesives, varnishes, paints, coating materials, dyes, wood preservatives, and flame retardants.

The waste material formed during production is mostly recycled to the process or incinerated to produce energy. Dumping is used in only a few exceptional cases, when allowed. Wood-based construction materials that are no longer utilizable and those that are at the end of their useful life are combusted. Here, their burning behavior must be taken into account, particularly in the case of nonwood substances. Incomplete combustion favors the emission of hazardous waste gas.

The emission of nitrogen oxides is determined mainly by the nitrogen content of the fuel [86]. For uncoated fiber- or particle board glued with amino resin, the nitrogen content is 3.5 – 5 %. For materials glued with phenolic resin, the

alkali incorporated with the resin inhibits propagation of burning. Wood preservatives must be incinerated under controlled conditions and the ashes disposed of separately. Chloride-containing curing aids for amino resins or coatings based on poly(vinyl chloride) give rise to the hazard of formation of polychlorinated dibenzodioxins and dibenzofurans [87]. By controlled incineration, the behavior is improved and the risk of formation of these hazardous substances is decreased.

2.6. Polymer Wood

Polymer wood consists of solid wood or wood-based materials that have been impregnated with monomers or synthetic resin solutions. The plastic components are then cured thermocatalytically or by radiation.

Plastics are incorporated in the wood texture to give a composite material with improved physical and chemical properties. The increase in strength of the wood can be seen principally in improved hardness and transverse compressive strength. The abrasion resistance of polymer wood is determined by the hardness and chemical structure of the polymer components. The dimensional stability and sorption properties are also improved, the latter being due to a significant reduction in the rate of water absorption. The effect is limited only by the fact that most monomers are deposited in the cell lumens and capillaries, but not in the cell wall itself where water is active.

In the production of polymer wood, attention is paid to the differing impregnation capacities of various types of wood and of sap- and heartwood. The liquid monomers or prepolymer solutions are applied by the vacuum-pressure process to accelerate penetration. The proportion of polymer components (30 – 60 %) determines the properties of the composite material. Direct grafting of the monomer onto the cellulose of wood is desirable [88]. Chemical catalysts, thermal energy, or radiation is used to initiate polymerization. The use of higher temperatures has the disadvantage that considerable proportions of volatile monomers are lost from the system by evaporation. Also, some wood components may inhibit polymerization and cause incomplete conversion of the monomer. The use of high-energy γ-radiation (cobalt-60) is very effective to achieve polymerization in wood [89]. In practice, a radiation dose of ca. 4×10^4 Gy (4 Mrad) is used. The surface hardness can thus be improved by up to 25 % compared with purely thermal polymerization.

The following have been proposed as monomers or prepolymer solutions:

1) Condensation resins (e.g., UF, MF, and PF resins)
2) Polyester and epoxy resins
3) Diisocyanates (e.g., MDI monomer, prepolymer, and polymer)
4) Acrylic esters, acrylonitrile, vinyl chloride, styrene, and maleic anhydride

The degree to which these substances undergo direct chemical bonding with cellulose in wood is not known for many monomers. For example, no urethane groups could be detected when diisocyanates were used, which indicated that reaction with the hydroxyl groups in wood had occurred [90]. The NCO groups can therefore be assumed to have reacted with water in the cell wall to form polyurea.

On impregnating oak with UF resin, a resin absorption of 43 kg/m^3 is achieved, and with sapwood-rich pine, 90 kg/m^3. Water absorption decreased and the compressive strength of the impregnated wood increases with increasing absorption of resin. On prolonged storage in water, the absorption of water increases again probably due to leaching of components of the impregnating agent [91].

Some of the monomers proposed present problems in industrial utilization for toxicological reasons because both the handling of these substances and their quantitative conversion in wood are difficult. A large number of processes for polymer wood production have been worked out and patented [92].

Nevertheless, polymer wood has not yet been widely used industrially. Economic reasons are mainly responsible for the reluctant utilization of this material. Polymer wood production is expensive, and its subsequent working with tools is difficult. Its resistance to weathering is not particularly high. Polymer wood is used for floors (parquet) and also for sports apparatus, tool parts, rifle shafts, and highly stressed industrial hardware.

2.7. Economic Aspects

The worldwide economic development of the wood processing industry was very positive during the decade from 1981 to 1991. The main areas of use for wood-based boards are the furniture and building industries. From 1981 to 1991, worldwide production of wood-based boards increased from 100 to 122×10^6 m^3 or by ca. 22 % [74, 93]. All types of board have contributed to this growth. The highest growth rate, 33.8 %, has been in *fiberboard* production (MDF and hardboard). The production of *particle board* has increased by 27 % to 49.5×10^6 m^3, whereas that of plywood and veneer plywood has achieved a growth rate of 18 %, and light fiberboard (*softboard*) a growth rate of 6.5 %.

Worldwide production and consumption figures for wood products are given in Table 12.

Table 12. Production and consumption of wood products worldwide (in 1000 m^3)

Product	Year	
	1981	1991
Cordwood	2 926 027	3 429 426
Wood-based panels	100 363	122 425
Particle board	39 020	49 459
Plywood	40 329	47 636
Fiberboards (including softboard)	16 517	20 201

The worldwide distribution of the production of wood-based construction materials gives an indication of the economic development of a region and the availability of corresponding raw material resources (Fig. 40).

If the three largest classes of wood-based construction materials – particle board, plywood, and fiberboard – are compared, since the 1960s particle board production has overtaken that of plywood, which was previously the leader in terms of quantity. A similar development is likely to occur with fiberboard (MDF). This can be seen, in particular, from a comparison between MDF production in Western Europe and in the United States from 1984 to 1992 (Fig. 41). In Figure 42 the quantities of particle board, plywood, and fiberboard produced by the five most important producer countries are compared.

Production of glued particle wood-based construction materials has increased markedly. The lack of round timber with large dimensions favors the use of lower-quality wood, mainly for fiber- and particle board. Improvement of the quality of glued particle wood-based construction materials, particularly in industrial processing, has also contributed to their reputation and use.

3. References

General References

1. H.-H. Bosshard: *Holzkunde,* vol. 2: Zur Biologie, Physik und Chemie des Holzes, Birkhäuser-Verlag, Basel 1984.
2. Deutsches Institut für Normung: *Normen über Holz,* DIN-Taschenbuch 31, Beuth-Verlag, Berlin 1992.
3. J. M. Dinnwoodie: *Timber – its Nature and Behaviour,* Van Nostrand Reinhold Co., New York 1981.
4. D. Fengel, G. Wegener: *Wood – Chemistry, Ultrastructure, Reactions,* 4th ed., De Gruyter, Berlin – New York 1989.
5. W. E. Hillis: "Heartwood and Tree Exudates," in T. E. Timell (ed.): *Springer Series in Wood Science,* Springer-Verlag, Berlin 1987.
6. E. A. Wheeler, P. Baas, P. E. Gasson (eds.) in IAWA (International Association of Wood Anatomists): "List of Microscopic Features for Hardwood Identification," *IAWA Bull.* **10** (1989) 219 – 332.
7. F. W. Jane: *The Structure of Wood,* 2nd ed., Adam & Charles Black, London 1970.
8. W. Knigge, H. Schulz: *Grundriß der Forstbenutzung,* Paul Parey, Hamburg 1966.
9. F. Kollmann: *Technologie des Holzes und der Holzwerkstoffe,* 2nd ed., vol. 1, Springer-Verlag, Berlin 1951.
10. R. Mombächer (ed.): *Holz-Lexikon,* 3rd ed., vols. 1 and 2, DRW-Verlag, Stuttgart 1988.
11. P. Niemz: *Physik des Holzes und der Holzwerkstoffe,* DRW-Verlag, Stuttgart 1993.
12. K. V. Sarkanen, C. H. Ludwig: *Lignins, Occurrence, Formation, Structure and Reactions,* Wiley-Interscience, New York 1971.
13. E. Strasburger: *Lehrbuch der Botanik,* 33rd ed., G- Fischer-Verlag, Stuttgart – Jena – New York 1991.
14. R. Trendelenburg, H. Mayer-Wegelin: *Das Holz als Rohstoff,* 2nd ed., Hauser Verlag, München 1955.
15. R. Wagenführ, C. Scheiber: *Holzatlas,* 3rd ed., VEB-Fachbuchverlag, Leipzig 1989.
16. K. Wilson, D. J. B. White: *The Anatomy of Wood its Diversity and Variability,* Stobart & Son, London 1986.

17. H. Willeitner, E. Schwab (eds.): *Holz-Außenverwendung im Hochbau,* Verlagsanstalt Alexander Koch GmbH, Stuttgart 1981, p. 27.
18. M. Paulitsch: *Moderne Holzwerkstoffe,* Springer-Verlag, Berlin 1989, p. 59.

Specific References

19. K. Dreiner et al.: "Holz als umweltfreundlicher Energieträger – Eine Nutzen-Kosten-Untersuchung," *Schriftenr. Bundesmin. Ernähr. Landwirtsch. Forsten,* Reihe A: Angewandte Wissenschaft 432 (1994) 1.
20. L. Ajtay, P. Ketner, P. Durigneud: "Terrestrial Primary Production of Phytomass," in B. Berlin, E. T. Degens, S. Kunze, B. Ketner (eds.): *The Global Carbon Cycle,* J. Wiley and Sons, Chichester 1979, pp. 129 – 181.
21. E. F. Brünig, T. W. Schneider: "Verfügbarkeit forstlicher Rohstoffe," *Mitt. Bundesforschungsanst. Forst Holzwirtsch.* **130** (1980) 13 – 22.
22. FAO: *Yearbook of Forest Products 1961 – 1991,* Rome 1993.
23. Bundesministerium für Ernährung, Landwirtschaft und Forsten: *Nationaler Waldbericht der Bundesrepublik Deutschland,* 1994, pp. 611-617.
24. H. Ollmann: "Holzbilanzen 1991 und 1992 für die Bundesrepublik Deutschland," *Arbeitsbericht des Instituts für Ökonomie 93/3,* Bundesforschungsanstalt für Forst- und Holzwirtschaft, Hamburg, Aug. 1993.
25. FAO: *Yearbook of Forest Products 1945 – 1993,* Rome 1995.
26. R. Wagenführ: *Anatomie des Holzes,* 4th ed., VEB-Fachbuchverlag, Leipzig 1989.
27. J. Bauch: "Variation of Wood Structure due to Secondary Changes," in J. Bauch (ed.): Natural Variations of Wood Properties, *Mitt. Bundesforschungsanst. Forst Holzwirtsch.* **131** (1980) 69 – 97.
28. R. M. Brown, D. Montezinos: "Cellulose Microfibrils: Visualization of Biosynthetic and Orienting Complexes in Association with the Plasma Membrane," *Proc. Natl. Acad. Sci. USA* **73** (1976) 143 –147.
29. W. Herth: "Plant Cell Wall Formation," in A. W. Robards (ed.): *Botanical Microscopy,* Oxford University Press, Oxford 1985, pp. 285 – 310.
30. E. Sjöström: *Wood Chemistry, Fundamentals and Applications,* 2nd ed., Academic Press, San Diego 1993, p. 249.
31. K. Freudenberg in K. Freudenberg, A. C. Neish (eds.): *Constitution and Biosynthesis of Lignin,* Springer-Verlag, Berlin 1968, pp. 47 – 116.
32. H. H. Nimz: "Das Lignin der Buche – Entwurf eines Konstitutionsschemas," *Angew. Chem.* **86** (1974) 336; *Angew. Chem. Int. Ed. Engl.* **13** (1974) 313.
33. T. E. Timell: "Recent Progress in the Chemistry of Wood Hemicelluloses," *Wood Sci. Technol.* **1** (1967) 45 – 70.
34. B. J. Fergus, A. R. Procter, J. A. N. Scott, D. A. J. Goring: "The Distribution of Lignin in Sprucewood as Determined by Ultraviolet Microscopy," *Wood Sci. Technol.* **3** (1969) 117 – 138.
35. J. A. N. Scott, A. R. Procter, B. J. Fergus, D. A. J. Goring in K. V. Sarkanen, C. H. Ludwig (eds.): *Lignin, Occurence Formation, Structure and Reactions,* Wiley-Interscience, New York 1971, p. 52.
36. N. Losberger: "Normbalken und Kundenwunsch," *Holz Zentralbl.* **117** (1991) no. 42, 633.
37. R. Hinterwaldner: "Verleimung von tragenden Holzbauteilen," *Holz Zentralbl.* **116** (1990) no. 30, 445 – 446.
38. K. Egner, H. Kolb: "Untersuchungen zum Alterungsverhalten für tragende Holzbauelemente," *Holz Roh Werkst.* **24** (1966) no. 10, 439 – 442.
39. K. Egner: "Die Leimung tragender Holzbauteile," *Holzbau* 1976, no. 4, 94 – 95.
40. F. Kollmann (ed.): *Furniere, Lagerhölzer und Tischlerplatten,* Springer-Verlag, Berlin 1962.
41. F. Kollmann, E. W. Kuenzi, A. J. Stamm: "Wood Based Materials," *Principles of Wood Science and Technology,* **vol. II,** Springer-Verlag, Berlin 1975, p. 223.
42. A. Pizzi (ed.): *Wood Adhesives,* Marcel Dekker, New York 1983.
43. IG Farbenindustrie (BASF), DE 550 647, 1929 (K. Vierling, M. Schmihing, H. Klingenberg).
44. B. Meyer: *Urea-Formaldehyde Resins,* Advanced Book Program, Addison-Wesley, Reading, Mass. 1979.
45. H. E. Edlund, E. Justuk, *Holz Zentralbl.* **107** (1981) no. 192, 1720 – 1721.
46. O. Kreibaum, *Holz Roh Werkst.* **15** (1957) 1.
47. Ji-Te Aktiebolag, SE 134 584, 1948 (H. Baumann).
48. R. Runkel, J. Jost, DE 841 055, 1948.
49. J. F. Werz KG, DE 1 055 125, 1956 (E. Munk).
50. C. Neumann, Ph. D. Thesis Universität Hamburg 1970.

51. A. A. Moslemi: *Particleboard volume 1, Materials; Particleboard volume 2, Technology,* Southern Illinois University Press, Carbondale – Edwardsville 1974.
52. K. Wishered, J. Wilson, *For. Prod. J.* **29** (1979) no. 2, 35 – 37.
53. H. J. Deppe, K. Ernst: *Taschenbuch Spanplattentechnik,* 3rd ed., DRW-Verlag Weinbrenner KG, Stuttgart 1991.
54. MCN Nederland, EP 0 062 389, 1982 (J. J. Hoetjer). Borden Inc., US 4 536 245, 1985 (D. W. Shiou, E. Smith).
55. VEB Leuna, EP 0 037 878, 1981 (J. Barse et al.). Statens Skogsindustrier, EP 0 027 583, 1980 (A. W. Westling). Fraunhofer-Gesellschaft, DE-OS 2 851 589, 1980 (E. Roffael, L. Mehlhorn).
56. A. Schneider, E. Roffael, H.-A. May, *Holz Roh Werkst.* **40** (1982) 339 – 344.
57. M. Paulitsch: *Methoden der Spanplattenuntersuchung,* Springer Verlag, Berlin 1986.
58. *Asian Pac. For. Ind.* **12** (1991) 41 – 42.
59. O. Suchsland, G. E. Woodson: *Fiberboard Manufacturing Practices in the United States,* Forest Products Research Society, Madison, 1990, p. 263.
60. H. Lampert: *Faserplatten,* VEB Fachbuchverlag, Leipzig 1966, 453 p.
61. G. Gran, *Holztechnologie* **28** (1987) 143 – 146.
62. O. Suchsland, G. E. Woodson: "Effect of Press Cycle Variables on Density Gradient of Medium-density Fiberboard," in T. M. Maloney (ed.): *Proc. 8th Particleboard Symposium, Pullmann,* Washington State University, Pullmann 1974, pp. 375 – 398.
63. Siempelkamp GmbH, *Holz Roh Werkst.* **40** (1982) 77 –80.
64. F. Kollmann: *Technologie des Holzes und der Holzwerkstoffe,* Springer-Verlag, München 1951.
65. J. M. Dinwoodie, B. H. Paxton: "The Long-term Performance of Cement-bonded Wood Particle Board in A. A. Moslemi" (ed.): *Inorganic Bonded Wood and Fiber Composite Materials,* Forest Products Research Society, Madison, pp. 45 – 54.
66. Y. Aoki: "The Manufacture of Heat Hardening Cement Particleboard," in [65], pp. 35 – 44.
67. A. Kasim, M. H. Simatupang, *Holz Roh Werkst.* **47** (1989) 391 – 396.
68. U.S. Secretary of Agriculture, US 4 393 019, 1966 (R. L. Geimer).
69. Siempelkamp, DE 2 751 473, 1977.
70. A. Sattler, K. Lempfer in [47], pp. 19 – 25.
71. M. H. Simatupang, N. Seddig, DE 4 031 935, 1990.
72. C. v. Bismarck, *Holz Roh Werkst.* **37** (1979) 9 – 15.
73. EN 311 "Surface soundness of particleboard," 1992.
74. Fédération Europeén des Syndicats de Fabricants de Panneaux de Particule, FESYP Annual Report 1993, Brussels.
75. Goldschmidt AG, DE 1 954 383, 1969 (R. Mitgan, U. Hohlschmidt).
76. Uniboard AG, DE-OS 1 653 231, 1967 (R. Kunz).
77. DIN 68 861 1 – 8,Verhalten von Möbeloberflächen, 1981. DIN 68 751,Kunststoffbeschichtete dekorative Holzfaserplatten, 1987. DIN 53 799,Prüfung von Spanplatten mit dekorativen Oberflächen auf Aminoplastbasis, 1986. ISO 9352,Determination of resistance to wear by abrasive wheels, 1989.
78. Werz KG, DE 1 055 225, 1956 (E. Munk).
79. E. D. Acheson, R. H. Codwell, E. Hadfield, R. G. Macbeth, *Br. Med. J.* **2** (1968) 587 – 596.
80. H. Wolf et al., *Arbeitsmed. Sozialmed. Präventivmed.* 1986, special issue 7.
81. J. B. Harborne, *Sci. Rep. Nr. 1,* Medical Research Council (MRC) Environmental Epidemiology Unit, Southampton 1982.
82. H. Petersen, W. Reuther, W. Eisele, O. Wittmann, *Holz Roh Werkst.* **30** (1972) 420 – 436; **31** (1973) 463 –469; **32** (1974) 402 – 410.
83. O. Wittmann, *Holz Roh Werkst.* **20** (1962) 221 – 224.
84. H. J. Deppe in K. Aurand, B. Seifert, J. Wegner (eds.): *Luftqualität in Innenräumen,* G.-Fischer-Verlag, Stuttgart 1982.
85. Fraunhofer-Gesellschaft, DE 2 829 021, 1980 (E. Roffael, L. Mehlhorn).
86. T. Nußbaumer, *Holz Roh Werkst.* **49** (1991) 445 – 450.
87. M. Strecker, R. Marutzky, *Holz Roh Werkst.* **52** (1994) 33 – 38.
88. R. Schandy, J. Wendrinsky, E. Proksch, *Holzforschung* **36** (1982) 197 – 206.
89. J. F. Siau, J. A. Meyer, *For. Prod. J.* **16** (1966) no. 8, 47 – 56.
90. A. Burmester, *Holz Roh Werkst.* **28** (1970) no. 5, 183 –186.
91. K. Erler, D. Knospe, *Holz Roh Werkst.* **46** (1988) 327 – 329.

92. Dow Chemical Co., DE 2 707 368, 1977 (W. E. Broxtermann, F. L. Brown). Atlantic Richfield Co., DE 1 953 236, 1969 (B. V. Bell). Mitsui Toatsu Kagaku K. K., DE 2 403 953, 1974 (Y. Tadashi et al.).
93. *FAO Yearbook 1991,* Forest Products, Rome 1993.

Wood Gasification → Gas Production

Wood, Preservation

Ernst-Heinrich Pommer, Limburgerhof, Germany

Thomas Jaetsch, Lanxess Deutschland GmbH, Business Unit Material Protection Products, Krefeld, Germany

1. Introduction 1
2. Wood Preservatives 2
2.1. Creosotes and Tar-Oil-Containing Products 2
2.2. Waterborne Wood Preservatives . 4
2.2.1. CCA and Variations 4
2.2.2. Copper-based (Chromium- and Arsenic-Free) Wood Preservatives . 5
2.2.3. Boron Compounds 6
2.2.4. Quaternary Ammonium Compounds 6
2.2.5. Phenol Derivatives 7
2.3. Solventborne Wood Preservatives 7
2.3.1. Pentachlorophenol 7
2.3.2. Tributyltin Compounds 7
2.3.3. Metal Naphthenates 8
2.3.4. Oxine copper 8
2.4. Organic Active Substances for Wood Preservatives 8
2.4.1. Fungicides 8
2.4.1.1. Triazole Derivatives 8
2.4.1.2. Polychlorinated Aromatic Compounds 9
2.4.1.3. Isothiazolones 9
2.4.1.4. Phenylsulfamides 9
2.4.1.5. Iodine-Containing Compounds . . . 9
2.4.1.6. Thiocyanates 9
2.4.2. Insecticides 10
2.4.2.1. Synthetic pyrethroids 10
2.4.2.2. Neonicotinoids 10
2.4.2.3. Insect Growth Regulators (IGR) . . 10
2.4.2.4. Phosphate Esters 10
2.5. Wood Preservatives for Special Applications 11
2.6. Flame Retardants 11
3. Uses of Wood Preservatives 11
4. Methods for Testing Wood Preservatives 12
5. Approval of Wood Preservatives . 13
6. Analysis 14
7. Processes for Applying Wood Preservatives 14
7.1. Overview 14
7.2. Processes Using High or Low Pressure 15
7.2.1. Pressure Process 15
7.2.2. Low-Pressure Processes 19
7.2.2.1. Low-Pressure Processes for Dry Wood 19
7.2.2.2. Low-Pressure Process for Green Wood (Sap Replacement Process) . 20
7.3. Pressure-Free Processes 20
7.3.1. Long-Term Processes 20
7.3.2. Short-Term Processes 21
7.4. Preservation of Wood-Based Materials 22
7.4.1. Treatment with Preservatives during Production 22
7.4.2. Treatment with Preservatives after Production 22
7.5. Special Processes 22
8. Environmental Aspects 23
8.1. Health Aspects 23
8.2. Soil, Water, and Air 23
8.3. Disposal of Treated Wood 25
9. References 26

1. Introduction

Wood consumption has increased worldwide, accelerated by overfelling and fire clearing. Wood is a renewable raw material, but reforestation measures are not sufficient to compensate for the decrease in forest stands. Therefore, for economic reasons, for the protection of forests, and to lower the consumption of tropical wood, the utility of wood must be prolonged by suitable measures. Wood can undergo physical, thermal, and chemical changes, and can be attacked and even destroyed by fungi, insects, and marine animals. Wood with a moisture level of $<18–20\,\%$ is not attacked by fungi, but can be attacked by insects—in certain climates, often predom-

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_357.pub2

inantly by termites. If constructive wood preservation measures are not possible, or durable and weather-resistant types of wood are not available, chemical preservatives must be used. Depending on the hazard posed, different preservatives are used; these can include flame retardants. Wood preservatives are increasingly subject to registration and approval, particular attention being paid to toxicology and environmental aspects, as well as effectiveness [1]. In many respects, the requirement profiles are comparable to those of plant protection agents. In addition to biological damage, wood can also be exposed to nonbiological damage. These sources of damage and possible preservation measures are listed in Tables 1 2 3.

Table 1. Abiotic damage to wood, according to [2]

Source of damage	Type of damage	Possible protection
Mechanical	deformation, destruction	use of stronger types of wood; combined use with other materials; polymer wood
Thermal (100 °C)	discoloration, decomposition, combustion	fire-retardants; agents producing insulating layers
Chemical (acids, bases)	discoloration, destruction	suitable types of wood; surface treatment
Weathering (UV radiation, water, air)	graying, weathering, swelling, shrinkage, crack formation	resistant types of wood; surface treatment with pigmented, water-repellent paints

2. Wood Preservatives

Wood preservatives must fulfill very different requirements. Depending on the intended use, a limited period of activity (e.g., for the temporary protection of freshly felled or sawn wood) or a prolonged period of activity (e.g., for railway ties, cooling tower wood, poles, and masts) may be expected. Wood preservatives should be chemically stable, should readily penetrate into the wood, and should be resistant to leaching out, evaporation, and light. The environmental pollution during use and processing of preserved wood should be minimal. The disposal of treated wood should not lead to contamination of soil, water, or air. Other important criteria for the choice of a preservative are its corrosion properties and its compatibility with paints and coatings, adhesives, and inorganic or organic building materials. No preservatives have yet been developed that fulfill all of these demands. Considerable advances have been made in the area of water-soluble and solvent-containing preservatives (Section 2.3).

Table 2. Animal pests, according to [3, 4]

Organism causing damage	Damage location
Fresh wood insects	freshly felled logs, freshly sawn timber
Bark beetle (Scolytidae; e.g., *Xyloterus* sp.)	forests, tropics
Heartwood beetle (Platypodidae; e.g., *Platypus* sp.)	logs and sawn timber, tropics
Wood wasps (Siricidae; e.g., *Sirex* sp., *Paururus)*	logs and construction wood, forests
Dry wood insects	
House longhorn beetle (*Hylotrupes bajulus* L.)	pine sapwood (predominantly roof timbering)
Gnawing beetle (Anobiidae; e.g., *Anobium, Ptilinus)*	soft and hardwood (construction wood, furniture, works of art)
Powder post beetles (Lyctidae; e.g., *Lyctus, Minthea)*	starch- and protein-rich hardwood (tropical sawn timber, parquet, furniture, wall paneling)
Boring beetle (Bostrychidae; e.g., *Bostrychus, Heterobostrychus*)	starch-rich hardwood, fresh or dry
Termites (e.g., *Reticulitermes santonenesis* (Feytand)	soft- and hardwood in tropical and subtropical regions
Marine-wood borers	
Shipworm (e.g., *Teredo navalis* L.)	underwater wood (harbor installations, ships' hulls)
Boring isopods (e.g., *Limnoria lignorum* R.)	minimum salt content in water >0.7 %, damage increases with temperature

Wood preservatives can be divided into (1) creosote and creosote–coal-containing preparations; (2) water-soluble inorganic salt preservatives and other water-soluble, predominantly organic compounds; and (3) solvent-based preservatives (solvent-containing preparations); and (4) organic active substances for wood preservatives. The last group consists of fungicides and insecticides which can be usually formulated into modern water- and solvent-based metal-free wood preservatives.

2.1. Creosotes and Tar-Oil-Containing Products

Creosotes in particular coal-tar distillates with *bp* between 200 and 400 °C belong to the most important wood preservatives with a broad spectrum of activity. These tar oil fractions are also

Table 3. Plant pests, according to [5]

Organism causing damage	Damage location	Type of damage	Possible protection
Bacteria	wood with high moisture level (on storage in water, spraying)	partially increased permeability of the wood, excess absorption of preservative, damage to paint	
Soft rot fungi (Ascomycetes; e.g., *Chaetomium globosum* Kunze)	wood with high moisture level (cooling tower wood, wood in contact with soil)	dry rot, initially in a thin layer; microscopic cavity formation	copper-containing salt preservatives, special oil preservatives
Wood-discoloring fungi [e.g., *Aureobasidium pullulans* (de Bary) (Arnaud)]	wood with moisture level > 25 %, fresh roundwood and sawn timber, surfaces exposed to weathering	blue to black discoloration	fungicidal salt preservatives; emulsion concentrates; antifungal primers or varnishes
Brown rot [particularly dry rot fungus (*Serpula lacrymans* S. F. Gray), cellar fungus (*Coniophora puteana* Karst), hite pore fungus (*Androdia sinuosa* Karst), leaflets, Lenzites species]	wood with moisture >20 %	brown coloration, cube-shaped breaks due to cellulose degradation	salt preservatives and oils
White rot [e.g., (*Coriolus versicolor* Quelet)]	wood with moisture >20 %	marblelike discoloration; slightly fibrous, absorbent wood, because of degradation of lignin and cellulose	salt preservatives and oils

used as a mixture with the original coal-tar or mineral oils in some countries (e.g., the United States). For more than 150 years, creosotes have been used to protect railway ties, masts, poles, and wood for cooling towers and for marine timbers. The estimated period of effectiveness is between 20 and 50 years, depending on conditions. A large number of chemical compounds in combination are responsible for long-term biological protection. The most important aromatic hydrocarbons contained are acenaphthene, fluorene, phenanthrene, and anthracene. The toxic values for selected coal-tar oil fractions against wood-destroying fungi are listed in Table 4. Under unfavorable conditions the addition of termiticides is necessary [7]. In coastal waters, boring isopods (*Limnoria tripunctata*) can destroy wood in contact with water if insufficient preservative is used [8].

A large number of specifications and standards for coal-tar oils and their use exist. In Europe the Western European Institute for Wood Preservation (WEI) is in charge of the specification of tar oils. Fractions with a higher boiling range are used for wood in contact with soil or water because of the better long-term protection: middle distillate fractions are suitable for wood subjected to less demanding conditions. The creosote is usually applied to wood by using the pressure process, but fence, tree, and vine posts are also treated by using butt-end treatment.

Benzo[*a*]pyrene (BaP) present in coal-tar–creosote oil is potentially carcinogenic. It is regarded as the key substance from which the content of carcinogenic, mutagenic, and teratogenic polycyclic aromatic hydrocarbons (PAHs) is derived. Polycyclic PAHs that can be present in coal-tar oil are naphthalene, methylnaphthalenes, dimethylnaphthalenes, acenapthene, fluorene, methylfluorenes, phenanthrene, methylphenanthrenes, anthracene, methylanthracenes, fluoranthene, pyrene, and benzo[*a*]pyrene. For chemical formulas, see → Tar and Pitch. Impregnating oils with a significantly reduced content of BaP and other PAHs can now be produced [9].

Because of their strong odor, color, and tendency to sweat, particularly in impregnated pine wood, creosotes can be employed only for outdoor use. The combustibility of wood impregnated with creosote is not increased but actually decreased [10].

In the EU, application of creosote is regulated by the Commission Directive 2001/90/EC which

Table 4. Limiting values for selected coal-tar oil fractions against wood-destroying fungi (in kilograms per cubic meter of wood) according to [6] *

Fraction boiling range, °C	*Coniophora puteana*	*Poria vaporaria*	*Lentinus lepideus* (tar oil tolerant)
200–220	3.9–6.15	6.15–8.0	8.0–12.8
260–280	6.15–8.4	8.8–13.2	20.5–32.2
330–360	7.8–12.0	16.4–20.5	11.4–15.6
Coal-tar oil (creosote)	5.1–6.8	5.55–6.0	24.6–30.0

* The first value gives the largest quantity that still allows attack by fungi; the second, the smallest quantity that prevents attack.

prohibits the use of creosote as a wood preservative except in specified products (< 50 ppm of BaP, and < 3 wt % of water extractable phenols) applied in industrial installations or by professionals. The treated wood must be for industrial use (e. g., railroad ties, utility poles, agricultural stakes without food contact, marinas) only and cannot be used for residential, recreational and leisure facilities where skin contact is possible; also excluded are agricultural structures where food or plants for consumption are grown.

2.2. Waterborne Wood Preservatives

The minimum effective and toxic levels of wood preservatives for various pests are given in Table 5. The most important waterborne wood preservatives are those that are rendered insoluble by reaction of the components with one another in the wood and those that become tightly bonded to the wood fibers, so that they are resistant to being leached out. The first group comprises inorganic salt mixtures that are fixed in the wood by chromium(VI) compounds [alkali-metal chromates and dichromates, and chromium (VI) oxide]. Chromium is responsible for fixation. It is reduced during the fixation process to almost insoluble chromium(III). The second group consists of organic compounds that are applied to wood in the aqueous phase and become bonded to the fibers or are precipitated to form insoluble compounds in the wood. Waterborne preservatives are usually applied by the pressure process. However, tank and butt-end treatment processes can also be used. Certain fluorine and boron compounds, which are suitable wood preservatives, are not fixed and can therefore be used to protect only wood that is not exposed to moisture. Fluorine is partially fixed in the presence of chromium(VI) compounds.

2.2.1. CCA and Variations

For a long time copper–chromium–arsenic salt mixtures (CCA) were the most widely used wood preservatives due to their broad-spectrum activity, lasting permanence in the wood, and excellent cost efficiency. The salt mixtures can have varying compositions. Arsenic(V) compounds used include arsenic pentoxide, arsenic acid, and disodium arsenate. In addition to chromates, the following copper(II) compounds are used: copper sulfate, basic copper carbonate, or copper oxide. Acid fixation of CCA salts occurs in various phases. Chromium(III) arsenate, basic copper arsenate, and chromium(III) hydroxide are formed. The length of the fixing process depends on the moisture level in the wood, pH, and temperature, it can be several days to several weeks [12]. A large number of theories exist about the mechanism of CCA fixing [13].

Variations of CCA such as ammoniacal copper arsenate (ACA), ammoniacal copper–zinc arsenate (ACZA), copper–chrome–boron (CCB), and copper–chrome–fluorine (CCF) had been developed as substitutes for arsenic or chromium containing compounds in order to comply with national legislations or for specific applications. However, none of these CCA variations became universally accepted because of performance or handling disadvantages (e. g., boron and fluorine compounds can be easily leached out of wood) and because more environmentally friendly chromium- and arsenic-free compounds became available in the meantime.

Whilst still being employed in many countries the use of CCA as wood preservative has been restricted or banned in the USA, Europe (EU), and Japan because of environmental and health concerns.

Table 5. Minimum effective and toxic levels for various types of salt preservatives (in kilograms per cubic meter of wood) according to [11]

Type of preservative	Brown rot		Soft rot		Wood longhorn beetle
	Unaged	Leached	Unaged	Leached	Unaged
Boron compounds	1				0.4
Copper sulfate	25–50	> 50	4	ca. 15	1
HF	1				0.3
SiF	1.5				0.3
Cu–Cr–As	6	12	7	15	2
F–Cr–As	2.2	11	> 50	> 60	0.5
Cu–Cr–B	2.5	20	4	10	2
Cu–Cr–F	2.8	12	5	6	0.7
F–Cr	3	14	ca. 50	> 60	0.5

2.2.2. Copper-based (Chromium- and Arsenic-Free) Wood Preservatives

Waterborne copper-based products are the most important class of wood preservatives especially for the hazard classes (HC) 3 (exterior wood above-ground) and 4 (wood in-ground), even though there is a trend to replace them with metal-free (or totally organic) wood preservatives for HC 3. Most widely used products of this class are ammoniacal copper quarternary ammonium compounds (quats) (ACQ) and copper azoles (CuAz) followed by bis(*N*-cyclohexyldiazeniumdioxy)copper (Cu-HDO) (**2**), whereas ammoniacal copper dithiocarbamate (CDDC) [14, 15] and ammoniacal copper citrate (CC) [16] play only very minor roles, because their efficacy spectrum is not broad enough to compete with CuAz or ACQ.

Copper applied in a suitable form provides already on its own a broad spectrum of protection to wood, because it is effective against most wood rotting fungi including soft rot species, as well as insects including termites. However, a number of species are tolerant to copper, the copper tolerant brown rot species are the most relevant of these. Therefore, the effectiveness of copper-based wood preservatives highly depends on the choice of cofungicide which ideally should fully and efficiently cover this efficacy gap of copper [17].

ACQ contains CuO and as cobiocide the quat di-*n*-decyldimethylammonium chloride [*7173-51-5*] (DDAC), or alternatively in a newer version also benzalkonium chloride [*8001-54-5*], in a ratio of 2:1. The addition of ammonia or an amine such as ethanolamine is required for the fixation of copper by formation of the corresponding copper complex. Typical standard retentions of ACQ are 4.0 kg/m^3 for HC 3, and 6.4 kg/m^3 for HC 4 [18]. ACQ treated timber performs very well under different conditions in many situations of use, especially when a water repellent is included in the preservative [19 – 21].

Copper azoles (CuAz) exist in several versions with variations of the azole part and with or without boron as additional booster of the insecticidal efficacy. Examples are copper tebuconazole, copper boron tebuconazole, and copper boron tebuconazole/propiconazole. The azole fungicides such as tebuconazole which are emulsified in copper azoles formulations have proven to be very effective against the copper tolerant brown rot fungi. Therefore, only low concentrations of azole are required to compensate for the copper efficacy gap, i.e., an azole to copper ratio of 1:25 is sufficient. Usually copper is provided as CuO or alkaline copper carbonate and ethanolamine is applied as fixation agent. Specified standard retentions of copper azoles (e. g., copper tebuconazole 1.7 kg/m^3 for HC 3, and 3.3 kg/m^3 for HC 4) [18] lead to excellent performance in many end uses [21 – 23].

N-Cyclohexyldiazenium chelates [24, 25] have been known in Europe as wood preservatives for many years. (Cu-HDO) (**2**) is rendered water soluble by copper–amine complex formation and is a chromium-free wood preservative that can be used in the pressure process [26]. Because of its good fixing properties and effectiveness against wood-destroying basidiomycetes, this preservative is also suitable for wood in continuous contact with moisture. Addition of copper(II) hydroxide carbonate improves the activity against soft rot, and addition of boric acid, which penetrates deeply by diffusion, has proved effective in protecting heart-

wood [27]. The toxic values for Cu-HDO and is 0.8–0.9 kg [27] per cubic meter of wood towards *Coniophora puteana* after leaching out. In the incineration of Cu-HDO-impregnated wood the composition of the flue gases, (CO, CO_2, HCN, and NO_x content) corresponds almost exactly to that of untreated wood [28].

Water-soluble *N-cyclohexyldiazeniumdioxypotassium* [*185843-92-9*] (K-HDO) (**1**) is added to adhesives for the production of wood-based construction materials to hinder attack by basidiomycetes such as *Coniophora puteana* and *Gloeophyllum trabeum* [29].

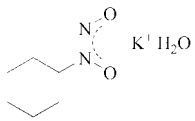

1

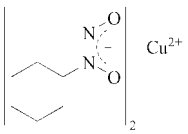

2

2.2.3. Boron Compounds

Boric acid, borax, and other polyborates, such as disodium octaborate tetrahydrate, are themselves wood preservatives and components of wood preservatives and flame retardants, and boron compounds are also suitable for protecting wood against termites [30]. Boron compounds protect wood against brown and white rot, and against the larvae of the house longhorn beetle (*Hylotrupes bajulus*) and the common furniture beetle (*Anobium punctatum*, larvae are called woodworm) [31, 32]. Because of their water solubility and their inability to be fixed, use of these compounds is limited mainly to wood not exposed to rain. All the common processes can be used to apply these preservatives. Because of the good diffusion capacity of boron compounds in moist wood, they are particularly suited to diffusion processes [33].

2.2.4. Quaternary Ammonium Compounds

Quats were introduced in New Zealand in 1978 as wood preservatives to counteract basidiomycetes and soft rot fungi [34, 35]. Extensive investigations on this class of compound have shown that *di-n-decyldimethylammonium chloride* [*7173-51-5*] is the most effective [36 – 40]. The toxic values against basidiomycetes are 1.0–6.2 kg of active substance per cubic meter wood [41]. Di-*n*-decyldimethylammonium chloride is applied in the pressure process or by dip treatment. Like other quats used in wood preservation (e.g., benzalkonium chloride), DDAC is not suitable for wood in contact with ground. In the wood, quats are bonded more strongly to lignin than to cellulose. The tolerance of the cellulose-degrading fungus *Coniophora puteana* to quats is attributed to this [42]. When applied in quantities of 2.0–4.0 kg/m^3, DDAC is also active against soil termites [43]. Apart from DDAC, *benzalkonium chloride* [*8001-54-5*] is also employed. Its average toxic value against fungi is 4.0 kg/m^3 [44, 45]. New developments focus on quats with other anions such as propionates or carbonates to avoid the presence of the corrosive chloride ion [46, 47]. Besides the corrosive properties of the anion, DDAC or benzalkonium chloride treated wood has also the disadvantage of being incompatible with many paints and coatings.

Polymeric betaines, which contain cationic alkylammonium groups and anionic borate groups, are suitable wood preservatives for use in brushing, dipping, and vacuum pressure processes for control of wood-destroying fungi, blue stain fungi, and insects. The commercial product contains *N,N*-didecyl-*N,N*-dipolyethoxyammonium borate/didecylpolyoxethylammonium borate [*214710-34-6*] as the active substance [48]. For basidiomycetes the toxic values after leaching test procedure are 7.6 kg/m^3 (aqueous solution) and 3.5 kg/m^3 (solvent-containing formulation). To prevent the development of house longhorn larvae, the retention of 0.3 kg (aqueous formulation) or 0.17 kg per cubic meter of wood (solvent-containing formulation) is sufficient. For wood that is continually in contact with soil and moisture, a salt was developed for the pressure impregnation process that contains betaine and copper amine [49, 50].

Toxicology. Polymeric betaines are nonmutagenic and nonsensitizing. The LD_{50}, oral, rat,

is between 500 and 2000 mg/kg. Their vapor pressure is very low (not determinable), and no toxic products are formed on incineration of impregnated wood [51].

2.2.5. Phenol Derivatives

Water-soluble *sodium pentachlorophenoxide* [*131-52-2*] is still used in several countries for temporary protection of sawn timber and veneers against discoloring mold and fungi, despite toxicological concerns and the consequent limitations on its use (see Section 2.3.1). It is often used in mixtures with borax [*1303-96-4*] to improve the activity against insects. The preservative is applied either by spraying or by dipping.

2-Phenylphenol [*90-43-7*], biphenyl-2-ol, 2-hydroxybiphenyl, is suitable for treating sawn timber against discoloring fungi and mold. An aqueous alkaline phenoxide solution is applied that reacts with the acidically reacting wood surface to form 2-phenylphenol, which is almost insoluble in water. 2-Phenylphenol has a low toxicity and is approved as a preservative in many countries. The substance is readily biodegradable and environmentally acceptable [52].

2.3. Solventborne Wood Preservatives

Solventborne wood preservatives are formulations comprising nonpolar, organic solvents and at least one biocide. The largest group consists of products with volatile (light) organic solvents (LOSP). They have the advantage that a second coat can be applied to the protected wood, which is not possible for wood treated with heavy-oil-containing preparations. Depending on the purpose envisaged, solvent-borne wood preservatives frequently contain solubilizers, binders, fixing agents, agents that counteract efflorescence of the active substance, pigments, dyes, light stabilizers, and water repellents. Solventborne wood preservatives penetrate wood without altering its structure (dimensions). These preservatives can be applied by brushing, injecting, spraying, dipping, or the double vacuum process with or without pressure.

2.3.1. Pentachlorophenol

Pentachlorophenol (PCP) [*87-86-5*] was a universally accepted wood preservative for many years, because it shows a very broad spectrum of activity against fungi and insects, and is resistant to biodegradation leading to good long-term protection. Today, the use of PCP is restricted or forbidden in many countries not only because of its persistency in the environment but also because of the possible formation of highly toxic polychlorinated dibenzo-*p*-dioxins, particularly octachlorodibenzo-*p*-dioxin (OCDD), on heating PCP or incinerating PCP-treated wood, and the possible presence of dioxins as impurities in technical-grade PCP. Nevertheless, in the USA PCP is still a popular preservative for utility poles where it is applied dissolved in diesel oil by pressure treatment at retention levels of 5–8 kg per cubic meter of wood for pine species [18].

2.3.2. Tributyltin Compounds

Bis(tributyltin) oxide [*56-35-9*] TBTO and other tributyltin compounds such as *bis(tributyltin) naphthenate* [*85409-17-2*] TBTN have been known since the 1950s as effective wood preservatives, in particular against wood-destroying brown and white rot fungi and insects. However, since their insecticidal activity is limited, addition of an insecticide (e. g., permethrin) is required where appropriate. Assessment of the effectiveness against white and soft rot fungi has led to conflicting results [53 – 55]. For toxicological reasons, use of tributyltin compounds has phased out in Europe and Japan. In Australia, New Zealand TBTN is still used in the double vacuum process for impregnating window wood, frames, and wooden doors made of softwood. Differing data exist for the period of activity, ranging from 5 to 30 years [56 – 59]. Various fungi and bacteria are able to detoxify TBTO [60]. TBTN is less volatile than TBTO.

Toxicology. For TBTO the LD_{50}, oral, rat, is 127 mg/kg, and for TBTN 224 mg/kg. Toxic values for the wood-protecting activity of TBTO are between 0.5 and 1.0 kg per cubic meter of wood for *Coniophora puteana* after leaching test procedure.

2.3.3. Metal Naphthenates

Of the metal naphthenates, *copper naphthenate*, in particular, is used against wood-destroying fungi and insects in the United States, and Asia/Pacific Rim. It is also used for impregnating poles in the pressure process [61] at retention levels of up to 2.4 kg Cu per cubic meter of wood [18]. Persistence of the active substance in the wood depends largely on formulation [62, 63]. The green coloration of treated wood can be undesirable. Besides copper naphthenate, zinc naphthenate is also used against wood-destroying fungi.

2.3.4. Oxine copper

Oxine copper [*10380-28-6*], Cu-8, bis(quinolin-8-olato-*O*,*N*) copper, is a copper chelate with 8-hydroxyquinoline. It is dissolved in organic solvents for pressure impregnation of wood for greenhouses, seed containers, transport containers for foods, etc. Cu-8 is only specified for above-ground applications and certain wood species. Water-soluble oxine copper formulations produced by using dodecylbenzenesulfonic acid are employed to suppress the development of discoloring (sapstain) fungi and molds on freshly felled wood [64]. A particular feature of oxine copper is its low toxicity toward warm-blooded animals and its good compatibility with plants.

2.4. Organic Active Substances for Wood Preservatives

Copper-based wood preservatives are inevitable for the protection of wood in ground contact (HC 4) as none of the organic fungicides tested is sufficiently effective against soft rot. However, for the protection of wood above ground (HC 3) modern metal-free wood preservatives have been developed since the 1990s. These are either waterborne or solvent-based wood preservatives with acceptable toxicological and environmental profile, which require usually combinations of two or more organic active substances including an insecticide to cover the full spectrum of biological hazard wood is facing. As most of the organic active substances are either very sparingly or practically nonsoluble in water, they are emulsified in the waterborne formulations.

2.4.1. Fungicides

2.4.1.1. Triazole Derivatives

The triazole fungicides are the most potent organic fungicides used in wood preservation forming the core part of many modern formulations for the treatment of above-ground wood. Azaconazole [*60207-31-0*], 1-[(2-(2,4-dichlorophenyl)-1,3-dioxolan-2-yl)methyl]-1*H*-1,2,4-triazole [65] was the first triazole introduced to wood protection in the mid-1980s. It has been used mainly as anti-sapstain agent in combination with quats for the treatment of sawn timber. In 1989 and 1990 two other triazoles, which have already proved effective in plant protection, have been developed as actives for wood protection with pronounced activity toward wood-destroying fungi. They are propiconazole [*60207-90-1*], 1-[2-(2,4-dichlorophenyl)-4-propyl-1,3-dioxolan-2-ylmethyl]-1*H*-1,2,4-triazole [66] and tebuconazole [*107534-96-3*], α-*tert*-butyl-α-(4-chlorophenylethyl)-1*H*-1,2,4-triazole-1-ethanol [67]. These triazole fungicides are effective against wood-destroying basidiomycetes, and blue stain fungi. The toxic values after evaporative aging of tebuconazole (propiconazole) against the important brown rot test fungi *Coniophora puteana, Gloeophyllum trabeum*, and *Poria placenta* are 0.05 (0.31–0.39) kg/m^3, 0.05 (0.57–0.64) kg/m^3, and 0.15–0.31 (0.10–0.15) kg/m^3, respectively. The combination of tebuconazole and propiconazole (1:1) provides not only a more balanced efficacy spectrum against brown rot than the single components but also is synergistically effective against the white-rot fungus *Coriolus versicolor* [68]. Therefore, the combination of tebuconazole and propiconazole is widely used as anti-basidiomycetes component. In treated wood, both propiconazole and tebuconazole are highly resistant to leaching and evaporation [69]. Another promising triazole for wood protection is cyproconazole [*94361-06-5*], α-(4-chlorophenyl)-α-(1-cyclopropylethyl)-1*H*-1,2,4-triazole-1-ethanol [70], which is very effective against wood rotting basidiomycetes,

however, its water solubility and its vapor pressure are less favorable regarding long-term protection compared to tebuconazole and propiconazole. The *toxicity* of these triazole fungicides toward warm-blooded animals is low.

2.4.1.2. Polychlorinated Aromatic Compounds

Besides pentachlorophenol (Section 2.3.1), chlorothalonil [*1897-45-6*], CTL, 2,4,5,6-tetrachloro-1,3-benzenedicarbonitrile, is used in wood preservatives mainly as cofungicide for the treatment of freshly sawn timber. In the search for alternatives to CCA, CTL was tested for its effects on wood-destroying basidiomycetes and soft rot because of its range of activity, low environmental impact, and low toxicity to mammals [71]. Stakes soaked in oil-borne formulations were almost unaffected by basidiomycetes over a period of several years; the application rate was in the range 6–12 kg CTL/m^3 of pine or eucalyptus wood [72, 73]. Further investigations confirm its potential at these high retention levels although its nonfavorable penetration profile as well as the high depletion rates have to be considered [74, 75].

2.4.1.3. Isothiazolones

Of the *N*-substituted isothiazolones, 4,5-dichloro-2-*n*-octyl-4-isothiazoline-3-one [*64359-81-5*], DCOIT, in particular has been tested for its effect on wood-destroying fungi and soft rot under severe conditions such as the soil burial test and the L-joint test [76 – 78]. Although DCOIT exhibited good performance in these trials, it is not commonly used in wood preservatives.

Toxicology. The LD_{50}, oral, rat, for a formulation containing 25 % a.i. (RH 287) is 1.56 g/kg.

2.4.1.4. Phenylsulfamides

The phenylsulfamides dichlofluanid [*1085-98-9*], *N*-dichlorofluoromethylthio-*N′*,*N′*-dimethyl-*N*-phenylsulfamide, tolylfluanid [*731-27-1*], and *N*-dichlorofluoromethylthio-*N′*,*N′*-dimethyl-*N*-tolylsulfamide [79, 80] are added to solvent-based wood preservatives in quantities of 0.4 to 0.8 % to protect wood against blue stain fungi and mold. The depth of penetration of the active substance into wood depends on the formulation. Both compounds are only stable in acidic aqueous formulations [81], thus, they are not suitable for the majority of the waterborne wood preservatives. Particular properties of these substances are light stability, resistance to leaching out, and involatility.

Toxicology. For both substances the LD_{50}, oral, rat, is 5000 mg/kg.

2.4.1.5. Iodine-Containing Compounds

3-Iodo-3-propynyl butyl carbamate [*55406-53-6*] IPBC, in principle is active against both wood-destroying and discoloring fungi [82], however, it cannot be used for long-term protection due to its easy biodegradability and relatively high leachability [83, 84]. Thus, it is mainly used as anti-mold and anti-blue stain additive in wood preservatives for HC 3 or as antisapstain agent in combination with quats or other fungicides [85, 86]. The good penetrating capacity of the substance is important when it is used in solvent-borne formulations to avoid damage by blue stain fungi [80].

Toxicology. The acute toxicity toward warm-blooded animals is low. The LD_{50}, oral, rat, is 1580 mg/kg.

2.4.1.6. Thiocyanates

2-(Thiocyanomethylthio)benzothiazole [*21564-17-0*], TCMTB, is used mainly for control of discoloring fungi in freshly sawn timber. The period of activity is three to five months [87 – 89]. The effectiveness of TCMTB can be improved significantly by combination with *methylenebis(thiocyanate)* [*6317-18-6*], MBT, which is also effective against discoloring fungi and molds. MBT penetrates rapidly in wood and exhibits high initial activity, while immobile TCMTB remains active on the surface [90, 91]. TCMTB also attacks soft rot [92] and brown and white rot. The toxic values against brown

rot fungi vary [93]. TCMTB is also used to protect wood-based construction materials against fungi.

2.4.2. Insecticides

For toxicological and ecotoxicological reasons, particularly their persistence in the environment, the use of chlorinated hydrocarbons, such as aldrin [*309-00-2*], dieldrin [*60-57-1*], chlordane [*12789-03-6*], and heptachlor [*76-44-8*], has been prohibited in many countries [94].

2.4.2.1. Synthetic pyrethroids

The pyrethroids are the most widely used class of insecticides in wood preservation because of their broad-spectrum activity, high efficacy especially against wood boring beetle, and high resistance to leaching [95]. Most popular are permethrin [*52645-53-1*], 3-phenoxybenzyl (1*RS*,3*RS*; 1*RS*,3*SR*)-3-(2,2-dichlorovinyl)-2,2-dimethylcyclopropanecarboxylate; cypermethrin [*52315-07-8*], (*RS*)-cyano-3-phenoxybenzyl-(1*RS*,3*RS*; 1*RS*,3*RS*)-3-(2,2-dichlorovinyl)-2,2-dimethylcyclopropanecarboxylate; followed by bifenthrin [*82657-04-3*]; 2-methylbiphenyl-3-ylmethyl-(*Z*)-(1*RS*)-*cis*-3-(2-chloro-3,3,3-trifluoroprop-1-enyl)-2,2-dimethylcyclopropanecarboxylate deltamethrin [*52918-63-5*], (*S*)-α-cyano-3-phenoxybenzyl-(1*R*,3*R*)-3-(2,2-dibromovinyl)-2,2-dimethylcyclopropanecarboxylate; and cyflutrin [*68359-37-5*], (*RS*;*SR*)-α-cyano-4-fluoro-3-phenoxybenzyl-(1*RS*,3*RS*; 1*RS*,*SR*)-3-(2,2-dichlorovinyl)-2,2-dimethylcyclopropanecarboxylate which may be used against insects attacking dry wood, in particular house longhorn larvae. Synthetic pyrethroids, take effect through direct contact, inhalation, and feeding. They also repel termites [96]. The protective effectiveness against the common furniture beetle is achieved at low retentions as follows: permethrin 50–80 g/m^3, cypermethrin 20–30 g/m^3, cyfluthrin 8–12 g/m^3, and deltamethrin 1–2 g/m^3 [97]. For the protection against termites two to three times higher retentions are required. The efficacy of the pyrethroids towards insects correlates qualitatively with the acute toxicity which ranges from 1500 mg/kg (LD50, oral, rat) for permethrin to 53 mg/kg (LD50, oral, rat) for bifenthrin [98].

2.4.2.2. Neonicotinoids

Imidacloprid [*138261-41-3*], 1-(6-chloro-3-pyridylmethyl)-*N*-nitroimidazolidin-2-ylideneamine, selectively acts on the insect nervous system as an agonist of the nicotinic acetylcholine receptor resulting in broad-spectrum activity against insects [99]. In contrast to the pyrethroids it is especially effective towards termites. In field trials, imidacloprid has shown to be protective against termites at retentions even lower than 3 g per cubic meter of wood [100]. It is also an effective termiticide for veneer-based products such as plywood [101]. Another neonicotinoid that promises high potential for wood protection is thiacloprid [102].

2.4.2.3. Insect Growth Regulators (IGR)

Insect growth regulators act highly specific on the endocrine and hormone systems of insects inhibiting the further development of the larvae and the eggs, respectively. Fenoxycarb [*72490-01-8*], ethyl [2-(4-phenoxyphenoxy)ethyl]carbamate [103, 104], and flufenoxuron [*101463-69-8*], 1-[4-(2-chloro-α,α, -trifluoro-*p*-tolyloxy)-2-fluorophenyl]-3-(2,6-difluorobenzoyl)urea [105] demonstrate good preventive effectiveness against wood boring beetles, having a favorable toxicological profile at the same time. However, they are not able to protect wood against termite attack due to their mode of action.

2.4.2.4. Phosphate Esters

The phosphate esters belong to an older generation of insecticides which are decreasingly used in wood protection due to both toxicological issues (prohibited in Japan) and the availability of more efficient compounds. Among the phosphate esters, phoxim [*14816-18-3*], *O*,*O*-diethyl-α-cyanobenzylideneaminooxyphosphonothioate, which is effective against the house longhorn and other beetles [79], and chlorpyrifos [*2921-88-2*] [106], *O*,*O*-diethyl *O*-(3,5,6-trichloro-2-pyridyl) phospho-

nothioate, which are used for controlling termites, should be mentioned in particular. However, the period of activity is limited under hot, dry conditions [107 – 110].

2.5. Wood Preservatives for Special Applications

Because of their high toxicity, use of the fumigants hydrogen cyanide or methyl bromide for insect control has greatly receded and is limited to special areas such as art works. They may be used only by trained specialists. The use of methyl bromide is expected to be limited because of possible damage to the ozone layer [111]. Sulfuryl fluoride is replacing it successfully [112].

For combating wood-destroying fungi inside wooden poles (made, for example, from Douglas pine), which have only a small proportion of impregnatable sapwood, trichloromethane; chloropicrin [*76-06-2*]; vapam [*137-42-8*], sodium *N*-methyldithiocarbamate; methyl isothiocyanate [*556-61-6*], MIT; and dazomet [*533-74-4*], 3,5-dimethyltetrahydro-1,3,5(2*H*)-thiadiazine-2-thione are suitable. Their fungicidal activity is effected via the gas phase. These substances are also used in agriculture as soil disinfectants. In wood, they are injected in various formulations into diagonally bored holes, which are then closed. Depending on the type of wood, its moisture level, and the diffusion capacity (distant effect), satisfactory effective activity has been established years after application [113 – 117], for chloropicrin up to nine years [118]. However, vapam is the most frequently used fumigant.

2.6. Flame Retardants

Flame retardants are intended to lower the flammability and combustibility of wood and thus reduce the hazard of destroying it by fire. They may be classified into three types according to their mode of action [119]: charring-promoting, barrier-layer-forming, and insulating-layer-forming flame retardants. All flame retardants are suitable only for interior use.

Charring-promoting flame retardants. consist mainly of impregnating inorganic salts, such as those based on ammonium phosphates, alkali-metal borates, and hydrogencarbonates and sulfates. Wood impregnated with these substances forms noncombustible and heat-insulating charcoal layers when attacked by fire, thermal degradation of the wood substance is thus slowed down. Of the salts that, on heating, form gases that smother the flames, ammonium phosphate is particularly important.

Barrier-Layer-Forming Substances. Substances based on water glass or borates form a thin barrier layer on the wood surface, which is not readily flammable and hinders the entry of oxygen. These substances are applied by brushing. Potassium water glass, with additives such as gypsum, kieselguhr, chalk, barytes, or zinc oxide, reacts with atmospheric carbon dioxide to form hygroscopic potash and a brittle silicic acid layer.

Solid-Phase Char-Forming Substances. The mode of action of these flame retardants is based on the fact that the insulating charcoal layer is not formed in or from the wood but on its surface. Substances are used that swell on heating to form foams, and that char and solidify at flame temperatures of 250–300 °C. Such foam-forming mixtures consist, for example, of phosphates, borax, paraformaldehyde, formaldehyde, and compounds with a high nitrogen content such as urea, guanylurea phosphate, guanidine, and dicyandiamide. Paints and coatings that form insulating layers contain either special plastics that swell on heating, swelling graphite, or mixtures of binders and foam-forming fillers.

3. Uses of Wood Preservatives

Forestry. Particularly in tropical and subtropical regions, susceptible types of wood—after felling and debarking of the logs—must be protected against fungi (mainly discoloring fungi) or insects that attack fresh wood. On prolonged storage or transport, repeated treatment can be necessary. Emulsion concentrates, which can be diluted with water, solvent-borne preparations, or waterborne salt preservative solutions, can be used.

Sawmills. If wood cannot be dried immediately after sawing, it must be protected primarily against discoloring fungi (blue stain fungi) and mold, depending on its susceptibility. The most effective method is short-term dipping. In addition, the spraying process is often used. The same preservatives are used as for the protection of roundwood. After peeling, veneers are sprayed only with preservative. Sawn timber may have to be treated a second time before transportation, for example, by sea, and piles of wood may have to be covered with sheeting for protection. For tropical woods, such as abachi, limba, ilomba, and wenge, that are prone to attack by insects precautionary treatment against attack by lyctus and wood-boring beetles and other insects is necessary.

Structural Engineering. Building lumber, wooden building components, and wood-based construction materials can be attacked and destroyed by fungi and insects, and also by termites in tropical and subtropical regions, depending on the type of wood and the environmental conditions. The wood preservation measures needed depend on the stress and the endangerment of the wood, for example, through continuous contact with soil, continuous or frequent wetting by rain, or use indoors in wet areas. Water-soluble salt preservatives are most common. These are applied to the wood by vacuum, pressure, tank, impregnation, dipping, brushing, injection, spraying tunnel, or flow coating processes. In addition, solvent-containing, low-solvent, and solvent-free formulations are used to protect nonload-bearing building components, such as outer doors and windows. In countries such as Japan, the United States, and Canada, where wood is used widely in building, flame retardants are important, either in the form of water-soluble salts or flame retardant paints or as coatings that form an insulating layer. Particular demands are made on wood preservatives and impregnating processes if, for example, highly stressed poles, railway ties, mine props, marine timber, cooling tower timbering, and posts need to be protected. Apart from a broad spectrum of activity, good penetration capacity and high resistance to leaching are required.

Wood-based materials include particle board, plywood, veneers, laminated board, and fiberboard (→ Wood). Depending on use, protection against wood-destroying fungi and soft rot fungi is necessary and, in some countries, against insects as well. Within the framework of quarantine regulations regarding packaging woods, including plywood boards, the Australian authorities also require precautionary measures against insects (lyctus and termites) [120]. Special concentrates of fungicides, insecticides, or water-soluble salt preservatives are used. These are added to the adhesive liquor or to the particles in the mixer during particle board production. To reduce possible flame propagation, flame-retardant layers or coatings forming insulating layers can be applied.

4. Methods for Testing Wood Preservatives

Wood, characterized by a large number of species of different composition and properties, can be attacked and destroyed by harmful organisms. The hazard depends on the wood's natural durability under the conditions in which it is used. Because of this, methods for testing the efficiency of wood preservatives are particularly important. They should be rapid, reliable, reproducible, and based on practical situations. They must give an indication of the effectiveness of preservation and its duration under changing conditions of use. Preservatives can be tested on wood in contact with nonsterile soil substrates, in the field, or with marine organisms. Another possibility is to use efficient breeds of important wood-destroying insects (e.g., *Hylotrupes bajulus* [121, 122], *Anobium punctatum* [123], and *Lyctus brunneus* [124]), termites [125], and—for testing with fungi—pure cultures of wood-destroying basidiomycetes, soft rot fungi, blue stain fungi, sapstain fungi and molds.

Many institutions in Europe, the United States, Canada, Australia, Japan, and other countries have developed standard test procedures for wood preservatives. These include not only biological tests, but also those concerning the treatment of wood with preservatives, specifications for wood preservatives, analytical procedures, corrosion tests, and tests for flame retardants. As far as possible, the tests should be carried out by using international standard procedures to avoid, for example, lengthy and expensive repetition of tests in the approval of new active substances.

The European standard for wood preservation is particularly important here [126]. It should guarantee uniform assessment of wood preservatives throughout Europe [127 – 129].

In testing not only are organisms relevant to the envisaged area of use incorporated, but test objects treated with wood preservatives are pretreated before the biological tests to simulate later conditions of use. These measures for accelerating the aging of treated wood before biological tests include leaching, evaporation (wind tunnel test), and weathering. Test procedures that are carried out under conditions of use, frequently over a period of several years, are also important. In these tests, proven wood preservatives are used as standards [130].

In Europe, Japan, and the United States, a variety of test procedures have been specified for testing flame retardants for wood and wood-based construction materials [130 – 136].

Internationally, research activities in wood preservation are coordinated by the International Research Group on Wood Preservation (IRG; IRG Secretariat, S-100 44 Stockholm, Sweden). Five working groups are involved with the following aspects of wood preservation:

Section 1 :	Biology
Section 2:	Test Methodology and Assessment
Section 3:	Wood Protection Chemicals
Section 4:	Processes and Properties
Section 5:	Environmental Aspects

5. Approval of Wood Preservatives

European Union. Within the EU, wood preservatives are from 2009 onwards to be registered under the Biocidal Products Directive (BPD; Directive 98/8/EC), which regulates the authorization and placing on the market of biocidal products and active substances used therein. The aim of the BPD is to provide a high level of protection for humans, animals, and the environment, to harmonize the common market for biocides, and to establish a positive list of active substances (Annex I to the BPD) which may be used in biocidal products, e. g., wood preservatives (Product Type 8). The BPD distinguishes between new active substances, which were not used in biocidal products prior to May 14, 2000, and existing active substances, which were used before this deadline and are subject to a review program. The full dossier for existing active substances used in wood preservatives with exhaustive data requirements (especially on toxicological and ecotoxicological properties) had to be submitted by March 2004. The risk-based evaluation and decision on active substances was supposed to take about two years, however, is still ongoing at the time of writing of this article. Once all active substances of a given wood preservative will have been included in Annex I, it will have to be authorized at Member State level. If Annex I entry of a substance is refused, a phase-out period will be defined at that time. Wood preservatives containing active substances, which have not been supported by a full dossier, had to be taken from the market by September 2006. New active substances have to be approved before being placed on the market for biocidal uses. Until the respective active substances will have been listed in Annex I, existing national regulations for wood preservatives will remain in place.

United States. The EPA is responsible for registration of new wood preservatives. A complete and detailed documentation of the chemical, toxicological, and environmental properties of the individual components and the final formulation is required. The EPA is in charge of the Office of Special Pesticides Review (OSPR), which tests wood preservatives already on the market. When a product or component may possibly be carcinogenic, mutagenic, teratogenic, or too toxic to humans and animals, a Rebuttable Presumption Against Reregistration (RPAR) is ordered. Finally, OSHA stipulates health and safety standards for industry (e.g., maximum workplace concentrations of wood preservatives) [118].

Germany. Because of building regulations in Germany, wood preservatives used for load-bearing and reinforcing building components must have a test certificate issued by the German Institute for Construction Engineering (Deutsches Institut für Bautechnik, DIBT). A committee, consisting of wood preservatives experts and representatives of the DIBT, the UBA, and the Federal Institute for Risk Assessment (Bundesinstitut für Risikobewertung, BfR), assesses the biological effectiveness of the preser-

vative, and its effects on health and the environment, on the basis of documents submitted.

6. Analysis

Qualitative and quantitative analysis procedures have been developed, and in many cases standardized, to determine the distribution and presence of preservatives or preservative components in treated wood or wood-based materials. Depending on the treatment process, type of wood, and proportions of sap- and heartwood, the distribution of preservatives in wood can vary widely. Since treated wood is expected to be used for 50 or more years, preservatives might be leached by rain and water, or components may be lost by evaporation (mainly organic substances); in addition, interactions between wood or organisms and the preservative could lead to a lowering of activity. Qualitative chemical analysis is an effective instrument for determining the causes of these losses. Since about the 1980s, analytical procedures have become particularly important as regards disposal of wood treated with preservatives, the testing of soil polluted with wood preservative residues (e.g., in impregnating plants), and the possible discharge of active substances from treated wood into water, soil, and air.

Qualitative Analysis. Colorimetric reactions are frequently used for qualitative detection (and depth of penetration) of water-soluble *inorganic salt preservatives* based on copper salts, fluorides, chromium(III) (fixed form), arsenic salts, boric acid, and borates, [137, 138]. In *creosote impregnation*, besides the characteristic odor, the brown to black coloration of the wood provides a means of optical recognition. Many methods have been proposed for the analysis of coal-tar creosote [139]. *Organic active substances* cannot be detected by a simple color reaction. The Beilstein test [140], frequently used for detecting PCP and lindane, is a nonspecific method of detecting halogen compounds.

Quantitative Analysis. For the quantitative analysis of wood preservatives, sampling is particularly important [141]. *Inorganic elements* are detected after extraction from the wood or in the residue after incineration of the wood. For the determination of metals, atomic absorption spectroscopy [142] and X-ray fluorescence spectroscopy [143, 144], for example, are employed. Colorimetry can be utilized for the determination of boron [145], and for fluorides the ion-selective fluoride electrode. For larger quantities of these elements, classical analytical methods can also be used.

Organic preservatives are generally determined, after extraction from wood, by using various chromatographic methods. Thin layer chromatography with a subsequent color reaction has proved effective for qualitative detection of active substances [146 – 148]. For direct detection and semiquantitative determination of substances such as PCP, dichlofluanid, tebuconazole, propiconazole, and permethrin in wood, GC–MS coupled methods using pyrolysis gas chromatography followed by a mass-selective detector are suitable [149].

7. Processes for Applying Wood Preservatives

7.1. Overview

Most types of wood are subject to attack or damage by wood-destroying fungi, insects, or marine animals if they are exposed to environmental conditions that favor the growth and development of these pests. Some types of wood possess natural durability. However, others (e.g., many of the most frequently used softwoods) are very susceptible. European standard EN 350, part 2, gives the natural durability and ease of impregnation of a selection of wood types that are important in Europe [150]. In other areas (e.g., Japan), similar data have been compiled [151, 152]. In general, wood preservatives are used as a precautionary measure to protect healthy wood. This also applies to flame retardants. The preservative is applied to the wood to as great a depth as possible. An exception is the temporary protection of wood against insects and wood-discoloring fungi after felling, during transport, and in storage areas during air drying.

Many processes for applying preservatives are available. In processes using reduced or excess pressure the preservatives penetrate into the wood because of the pressure differences. The

pressure process and the low-pressure process belong to this category. In pressure-free processes, the capillary forces in wood tissues are responsible for the absorption of preservative solutions. Pressure-free processes are used as either long- or short-term processes. With the latter, only surface protection of the wood is achieved.

Before wood is treated with preservatives, the bark and phloem must be removed. Also, the treatment should occur only after the final machining process. For sap displacement procedures the presence of bark is essential for the process (see Section 7.2.2.2). Because of the different ease of impregnation of various woods, the degree of impregnation aimed at also varies. Heartwood and mature wood are generally much more difficult or nearly impossible to impregnate compared with sapwood [135, 153, 154].

The choice of preservative, impregnating process, impregnating conditions, quantities applied, and minimum efficiency in biological experiments is laid down in EN 599, part 1, within the framework of the European standards for wood preservation [155]. The demands made on wood preservatives, as determined by biological experiments, are described in EN 599, part 2 [156]. With regard to the use of wood preservatives, five possible biological hazard classes are defined in EN 335, parts 1 and 2 [157, 158]. They are listed in Table 6. Wood in hazard classes 4 and 5 generally needs to be treated by using the pressure impregnation process.

Factors affecting the results include the following:

1) The preservative-absorbing capacity and ease of impregnation of a particular type of wood
2) The moisture content of the wood at the time of surface treatment of impregnation
3) The application or impregnation process used
4) The physical properties of the preservative, such as solubility, penetrating capacity, corrosion properties, behavior on temperature change, viscosity, suitability for certain processes, and odor.

Measures for wood attacked by fungi or wood infested by insects are carried out predominantly on wood that has already been used in construction. If the attack is not too severe, particularly inside the wood it can be dealt with by boring holes and using suitable preservatives, for example in poles, beams, or window frames. Badly damaged pieces of wood and construction elements that have been severely weakened in their static properties must be replaced by new ones treated with preservatives.

Depending on the area of use, wood species, its natural durability and ease of impregnation, and the impregnation process and preservative employed, treated wood can be used for 3 to >20 years longer than untreated wood [159]. Thus, properly carried out wood preservation is very economical.

7.2. Processes Using High or Low Pressure

Pressure processes have been used for more than 150 years for impregnating wood with preservatives. The distribution of preservative should be as deep and uniform as possible in the parts of the wood that can be impregnated. Heartwood generally absorbs very little preservative. Different types of wood vary in their ease of impregnation. Treatment using the pressure process is absolutely necessary for wood in constant contact with water or soil. The process is often employed also for treating wood used outdoors in frequent contact with water.

The choice of pressure process depends on the intended use, the nature and quantity of the preservative, and the moisture level of the wood at the time of treatment.

Reduced pressure (vacuum) processes are used for treating dried construction wood, in particular, with solvent-containing preservatives (*double vacuum*). Water-soluble preservatives (*simple vacuum*) are also suitable for this process. Higher penetration depths are achieved than with the dipping process, and less preservative is used than with the pressure process. Vacuum impregnation is not suitable for wood that is in continuous contact with water or soil. Green wood (e.g., poles or roundwood) can be impregnated with salt preservative solutions by sap displacement.

7.2.1. Pressure Process

The basic equipment for a plant for the *full-cell process* consists of an impregnating cylin-

Table 6. Definition of biological hazard classes for solid timber

Hazard class	Situation in service	Description of exposure	Wood moisture content
1	above ground, covered (dry)	permanently dry	permanently 18 %
2	above ground, covered (risk of wetting)	occasionally wet	occasionally 20 %
3	above ground, not covered	frequently wet	frequently 20 %
4	in contact with ground or fresh water	permanently wet	permanently 20 %
5	in salt water	permanently in salt water	permanently 20 %

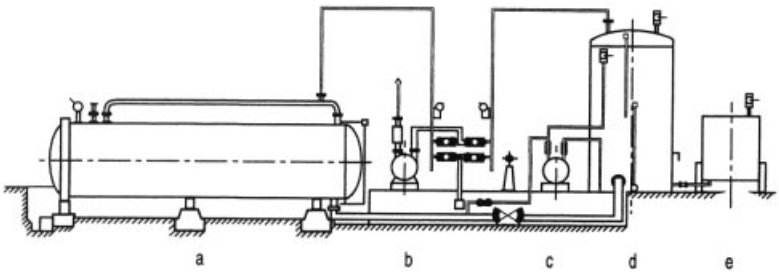

Figure 1. Schematic of a plant for the full-cell process
a) Impregnating cylinder; b) Vacuum pump; c) Pressure pump; d) Storage and measuring vessel; e) Dissolving vessel for water-soluble preservatives

der with one or two seals, one or more storage vessels for various wood preservatives or solutions of different concentrations, a pressure pump, a vacuum pump, measuring and dissolving vessels, and a control device for maintaining impregnation conditions. Plants for the Rüping process also include a compressor and a pressurizable storage vessel.

The pressure process gives the best preservative distribution in wood. Water-soluble salts, creosote, and mineral oils with fungicides or insecticides are used.

Full-Cell Process (Fig. 1). In the full-cell process (saturation impregnation or vacuum–pressure impregnation), air present in the wood is removed with a high initial vacuum, then the impregnation cylinder is filled with the preservative solution while the vacuum is maintained. Finally, after air is admitted to the cylinder, an impregnation pressure is produced and maintained for a defined period, depending on the ease of impregnation of the wood. Generally pressures >0.9 MPa are applied. The pressure is maintained until virtually no more liquid is absorbed by the wood. In some plants this is controlled automatically via an end-phase counter, which determines the quantity of preservative that can still be absorbed by the wood. Application of a short final vacuum should decrease the extent of dripping of the impregnating solution from the wood after it is removed from the cylinder. This process was introduced by Bréant in 1831 for impregnation with salts, and by Bethell in 1838 for impregnation with tar oils. Figure 2 shows an impregnation diagram for the full-cell process. Most plants are now controlled automatically with fixed programs and can produce full impregnation protocols independently.

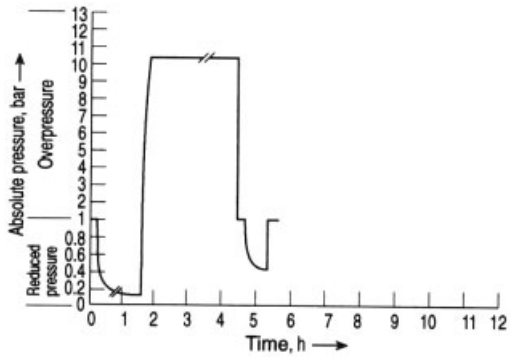

Figure 2. Pressure diagram for the full-cell process

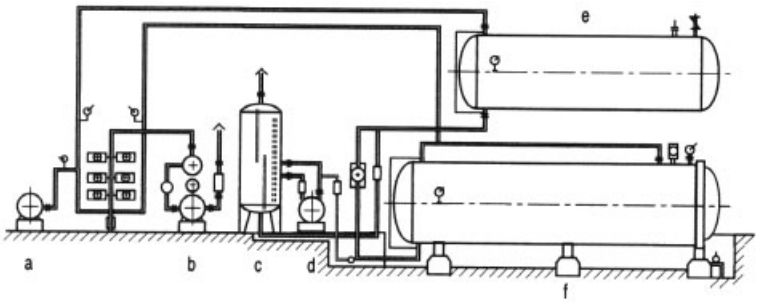

Figure 3. Schematic of a plant for the Rüping process
a) Vacuum pump; b) Air compressor; c) Measuring vessel; d) Pressure pump; e) Storage vessel; f) Impregnating cylinder

The full-cell process is applied mainly for water-soluble wood preservatives. Coal-tar oils and solvent-containing preparations are seldom used because preservative absorption is too high. The absorption of liquid is between 100 L/m^3 (spruce poles) and 500–600 L/m^3 (pine boards or posts), depending on the type of wood, thickness in the case of sawn timber or roundwood, and pretreatment (e.g., perforation of spruce).

Empty-Cell Processes. In 1902 and 1906, respectively, Rüping and Lowry developed processes that allowed absorption of a smaller amount of liquid in the pressure process without impairing impregnation. The processes are known as the empty-cell processes and are used, in particular, for hot creosote oils. Here, only the endangered cell walls are treated with preservative, while the cell cavities remain empty.

Rüping Process. In the Rüping process (see Fig. 3), air present in the cells is first compressed and absorption of liquid is thus prevented. The impregnation cylinder is then filled with preservative while the initial air pressure is maintained and an impregnating pressure (generally >0.9 MPa) is produced and maintained over a defined period. The cylinder is then depressurized to atmospheric pressure. During depressurization, air in the wood, which has been compressed in the prepressure stage and further during impregnation, expands and expels a considerable portion of the preservative from the wood cells. This is then recycled to the storage vessel. Subsequently as high a final vacuum as possible is produced, which results in a second expulsion of preservative. After the cylinder has been aerated to return it to atmospheric pressure, the impregnating solution recovered during the final vacuum is also recycled to the storage vessel. Figure 4 shows the impregnation diagram for the *simple Rüping process*.

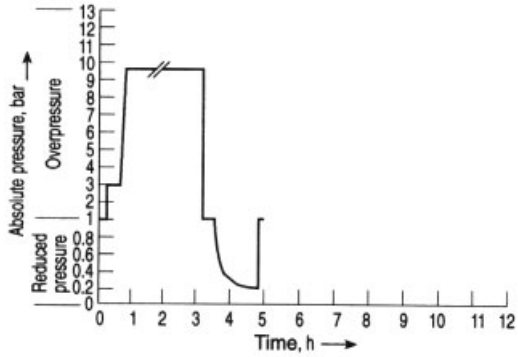

Figure 4. Pressure diagram for the Rüping process

The simple Rüping process is suitable for impregnating softwoods, such as pine, spruce, larch, fir, and Douglas fir, and some hardwoods such as beech. The average quantities of creosote absorbed are, for example, for beech railway ties ca. 175 kg/m^3; for pine railway ties, ca. 60 kg/m^3; and for pine poles or pine, spruce, larch, fir, or Douglas fir posts, ca. 90 kg/m^3. Both tar oils and water-soluble salt mixtures can be used in the Rüping process [160].

Railway ties that are difficult to impregnate can be treated satisfactorily with creosote using the *double Rüping process*, which involves a combination of two simple Rüping impregnations. Improved Rüping processes with *preconditioning* are used only for impregnation with creosote and give a considerably improved tar oil distribution in the wood [161]. The coal-tar creosote is heated to the prescribed impregnation

temperature before the beginning of the pressure period. In a preconditioning phase the impregnating cylinder is filled with hot creosote at 115–120 °C. After cooling the creosote on the wood, the former is heated again to at least 100 °C. This temperature is maintained until the wood is sufficiently prewarmed. The lower viscosity of the creosote achieved by these measures allows more rapid penetration of the preservative and better distribution in areas of the wood that can be impregnated. Both the simple and the double Rüping processes can be improved by preconditioning. In 1986, the improved double Rüping process was simplified by omitting the intermediate vacuum after the first compression period.

Lowry Empty-Cell Process. Unlike the Rüping process, the Lowry empty-cell process does not employ preliminary air compression [162]. After the wood has been charged to the impregnating cylinder, the latter is filled with impregnating liquid at atmospheric pressure. Pressure is then applied, and the length of the compression period can vary depending on the type of wood and its dimensions. As in the Rüping process, impregnation is terminated with as high a final vacuum as possible. However, the quantity of liquid expelled is less than in the Rüping process. Creosote and waterborne salt preservatives are used in this process.

Rotating-Vessel Process. The rotating-vessel process is an extension of the pressure process and was introduced by Poulain in 1938. It is used in some European countries, for example, to doubly protect poles that have already been impregnated with salt preservative solutions, with coal-tar oil in the particularly vulnerable lower sections. Here, the impregnating cylinder, which is charged in a horizontal position, is moved to a vertical position before beginning the impregnation process. Since salt-impregnated poles must be predried for several weeks before rotating-vessel impregnation with creosote, treatment of the lower sections with creosote is often carried out beforehand.

Liquid-Gas Process. The liquid-gas or Cellon process is a further variant of the full-cell process. It was developed by Hudson and is used mainly in the United States [163 – 165]. Active substances, mostly PCP [135], dissolved in liquefied petroleum gas (isobutane—butane) are charged to the evacuated impregnating cylinder, and the impregnating pressure is produced by increasing the temperature of the liquid to 20–30 °C. By gradually lowering the pressure to atmospheric and subsequent evacuation, most of the liquefied petroleum gas is recovered, while the preservative remains in the wood. The process has the advantage that the wood is obtained dry and, particularly for hardwoods, a deeper and more uniform penetration of the active substance is possible. Impregnation requires special safety measures and additional equipment for recovery of liquefied petroleum gas.

Oscillating Pressure Method. To impregnate green cordwood, particularly roundwood that is difficult to treat in the dry state, the oscillating pressure method was developed [166]. Only waterborne preservatives are used. In principle, the process can be carried out in any pressure plant, provided that automatic plant control for cyclic pressure changes is available. An oscillating pressure plant is shown schematically in Figure 5. In this process, vacuum and pressure phases alternate, whereby the phase changeover occurs in less than one second. During the vacuum phases a small quantity of liquid is sucked out of the wood, and in the compression phases the salt solution is pressed in. As the impregnation period continues the pressure phases increase in length. For example, with green spruce, ca. 400 phase changes occur during the ca. 22-h impregnation process. Here the sap still present in the wood is continuously replaced by preservative liquid with an increase in the impregnation of the sapwood. For wood that is difficult to impregnate, the oscillating pressure method gives better impregnation and greater penetration depths than the full-cell method [167].

In a variation of the oscillating pressure method, the usual vacuum phases are omitted [168, 169]. In intervals of minutes, pressure is produced slowly in the cylinder, maintained for a short time, and then reduced to atmospheric again. The length of the cycles is of the order of minutes and remains constant during the impregnation process. The number of cycles is considerably fewer than in the conventional oscillating pressure process. By using the air compressed in the wood, the sap is continually enriched and replaced by salt preservative solution.

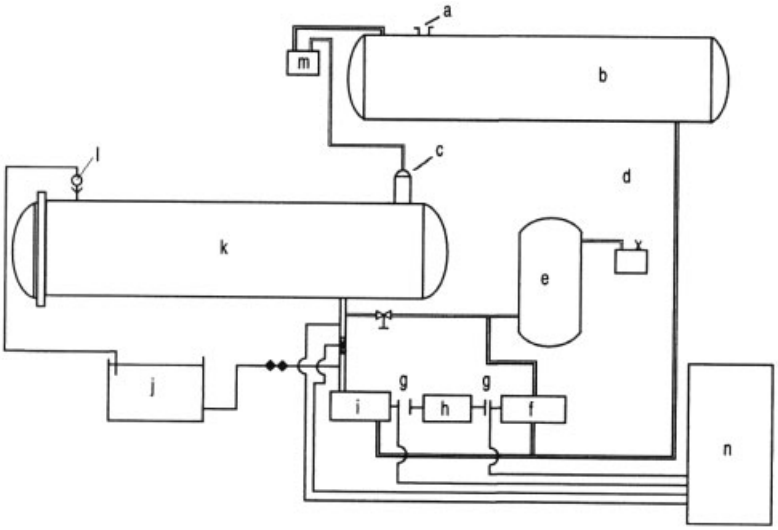

Figure 5. Schematic of an oscillating pressure plant
a) Air trunking; b) Storage container; c) Liquor valve; d) Small compressor; e) Accumulator; f) Pressure pump; g) Magnetic coupling; h) Motor; i) Suction pump; j) Dissolving vessel; k) Impregnating vessel; l) Safety valve; m) Small vacuum pump; n) Control

Supercritical Carbon Dioxide. The use of a supercritical fluid for the impregnation of wood with organic fungicides and insecticides under high pressure ($\geq$ 10 MPa) is a new technology which was commercialized first in 2001 in Denmark [170]. Advantages of this process include short impregnation cycles, avoidance of volatile solvents and water (long drying time; dimensional changes), and enabling the treatment of wood species that are resistant to conventional processes [171]. In addition, the treatment of wood-based composites (e.g., oriented structural board (OSB), medium-density fiberboard (MDF), plywood, laminated veneer lumber (LVL)) offers a large potential for the utilization of the supercritical carbon dioxide technology [172].

7.2.2. Low-Pressure Processes

7.2.2.1. Low-Pressure Processes for Dry Wood

Simple Vacuum Process. In the simple vacuum process, which is no longer important, the wood is charged to the impregnating vessel and as much air as possible is removed from the wood tissue by using a vacuum pump. The impregnation itself is then carried out at atmospheric pressure. Preservative dripping off the wood is collected.

Double Vacuum Process. The double vacuum process allows the economical use of solventborne wood preservatives because excess preservative can be recovered from treated wood and recycled by means of the second vacuum, which is applied at the end of the impregnation process. The first stage is carried out as in the simple impregnation process. For wood that is not readily impregnated (e.g., spruce or Douglas fir), the pressure difference between the initial vacuum and atmospheric pressure is not sufficient to allow the required quantity of preservative to be applied. In this case a modification of the Lowry process can be used, in which the preservative is absorbed under impregnating pressures of 0.2–0.6 MPa. On using this process, depths of penetration are achieved that lie between those of the dipping and the pressure processes.

The double vacuum process is used mainly for wood that must not swell any more after drying. This includes, for example, building components that have been planed or must have an exact size, such as windows, doors, and outside claddings. The advantage of the process that, after evaporation of the solvent, decorative var-

nishes or paints can be applied. Glueing is also possible. Other components, such as hydrophobing agents, can be added to the wood preservatives developed for the vacuum process.

To decrease pollution by solvents, solutions of water-soluble preservatives are now also used in the vacuum process.

7.2.2.2. Low-Pressure Process for Green Wood (Sap Replacement Process)

In the sap replacement process, which was introduced in 1838 by Boucherie, the sap in freshly felled, nondebarked conifers is replaced by a water-soluble salt preservative solution. From this, Gewecke developed the tank suction process in which debarked, green roundwood is placed in an impregnating tank equipped at one or both ends with a suction cap and attached by tubing to a vacuum pump [173]. The sap is sucked from the trunk by a high vacuum and is replaced by the salt preservative solution from the tank, which penetrates through the surface area of the wood [174 – 176]. A variant of this process is the tank pressure suction process developed by Gruber, in which sap is sucked from the top end and replaced by the salt preservative solution in the tank [177]. At the bottom, salt preservative solution at a higher concentration is pumped into the wood from an additional storage container by using a pressure pump.

Despite various improvements [e.g., the cap developed by Bundesanstalt für Materialforschung und -prüfung (BAM cap) [174], sap replacement processes have been superseded by the oscillating pressure process [167] because of shorter impregnating times and more reliable impregnation.

7.3. Pressure-Free Processes

In pressure-free processes, no plants with pressure vessels and pressure or vacuum pumps are necessary for applying the preservative. The wood absorbs preservative solutions through capillary forces and diffusion. Penetration of the preservative can also be promoted by the hydrostatic pressure of the liquid covering the wood in the impregnating container and by temperature differences in the impregnating liquids (hot–cold).

7.3.1. Long-Term Processes

Diffusion Process. In the diffusion process, which is used in only a few countries, peeled or sawn green wood is treated with highly concentrated solutions or pastes of water-soluble salt preservatives. After dipping or brushing on of preservative, the moist wood is piled up, coated with a watertight cover, and stored for several weeks. A good degree of protection is achieved even in wood that is difficult to impregnate, such as spruce or fir, by diffusion of the active substances. Borates are being used increasingly for this process [178 – 180]. At the end of the diffusion process the wood is dried or stored until dry. The disadvantage of this process is that the preservative is not fixed in the wood and can therefore be leached out. In a *double diffusion process* [181, 182], debarked, moist wood is first placed in a waterborne salt preservative solution, frequently copper sulfate, for several days and then impregnated with another salt solution, such as sodium chromate or dichromate, also for several days. Copper sulfate and chromate react to form a water-insoluble precipitate. Treated wood must be dried before use. This process is frequently used where no impregnating plants are available.

Open tank impregnation is still widely used. Here, semidry to dry wood is immersed completely in preservative solutions, which can be diluted with water, for up to several days. The process was introduced in 1832 by Kyan for treating wood with mercuric chloride solutions (kyanization). The use of mercury is now prohibited almost worldwide. Only surface protection of wood is generally achieved. Deep protection is also possible with readily impregnated or moist wood. A disadvantage is that uniform absorption is not guaranteed when salt preservative solutions with several components are used.

Hot–cold open tank impregnation is a variant of tank impregnation. In this process the impregnation period can be shortened and the depth of penetration increased. It is particularly suitable for creosote impregnating oils. Because water-soluble wood preservatives can be applied

at 50 °C (max.), they are not used for this process. The principle underlying the process is the fact that on storing dry to semidry wood in warm impregnating liquid, air present in the wood expands. On subsequent cooling, slight subatmospheric pressure develops in the wood, depending on the temperature difference. Because of this subatmospheric pressure, more impregnating liquid can be absorbed by the wood. Heating and cooling the impregnating liquid can be carried out in either the same or two separate tanks. In using creosote impregnating oil, the preservative and the wood are heated to ca. 80–110 °C and this temperature is maintained for several hours. It is then lowered to ca. 40–50 °C, either by replacing the hot solution with a cold one or by transferring the wood to a tank containing cold solution. The hot–cold open tank impregnation can be modified for the use of heat-sensitive preservatives or those based on flammable solvents. In this case, wood is heated for 1–2 h by, for example, hot water, hot steam, or hot air, and then transferred as rapidly as possible to a tank containing cold preservative solution, where it is cooled.

A variation of the tank and hot–cold open tank impregnation processes is *butt-end treatment*, which is frequently used for preserving fence, fruit tree, and vine posts, and similar objects, in areas where they are in contact with soil. Wooden posts are placed vertically in vessels containing creosote, in which the oil level is 20–30 cm above the envisaged depth of immersion in the soil. The vessels are then heated to >100 °C. Impregnation occurs during the cooling-down phase. Water-soluble or solvent-containing preservatives are used only for impregnating solutions where the temperature is not raised. Depending on its moisture level, the wood can remain from days to weeks in the solution.

7.3.2. Short-Term Processes

Short-term processes are wood treatment methods that are easily carried out and generally do not require expensive equipment. The preservative solution is applied to the wood surface only once or twice. Since preservatives penetrate the wood to a maximum of a few millimeters, only surface protection is achieved. These processes are therefore suitable only for wood that is not exposed to any particular hazard.

The simplest treatment methods are *brushing* and *spraying*. Spraying should be carried out only if precautionary measures are taken with regard to environmental safety and health. Solvent-containing preparations and oily formulations that can be diluted with water are mainly used. The quantity of preservative absorbed essentially depends on the type of wood and the nature of its surface. Two treatments are generally necessary. If water-soluble preservatives are used, wet-in-wet processing is generally employed to achieve greater depths of penetration.

Deluging. A further development of the spraying process is the deluging process, in which dry wood is stored above a receptacle and the preservative solution is allowed to flow over it. Excess solution is collected in the receptacle and pumped back to the storage tank. Since the upper and lower sides of the wood need to be treated, two operations are necessary. The process has the advantage that no spray mist is formed.

Spray Tunnel Plants. The use of spray tunnel plants permits the uniform application of preservatives, while minimizing their waste. Both mobile and stationary plants exist. In these plants, dry wood is passed continuously through a spray chamber. The jets are arranged such that the wood is sprayed uniformly on all sides. Excess preservative solution and that dripping off are collected and can be reused. A device that scrapes off preservative not absorbed is generally placed after the spray tunnel to minimize loss through dripping after the wood has left the tunnel. As with other short-term processes, the absorption of solution and thus the quantity applied depend on the type of wood, the nature of the surface, and sometimes the moisture level in the wood. Higher depths of penetration can be achieved by mechanical pretreatment of the wood, for example, by oblong perforation.

Short-Term Dipping. In short-term dipping, wood is dipped for a short period into a vessel containing a preservative solution, usually via a lifting device. The advantage of this process is that the wood comes into contact with preservative on all sides. Depending on

the type of wood, the way it is worked, and the formulation of the preservative, varying quantities are absorbed. For some preparations, repeated dipping is necessary to achieve sufficient protection. The process is used, in particular, for joinery (e.g., in window production). Both solvent-containing and waterborne preservatives are used.

7.4. Preservation of Wood-Based Materials

Particle board, veneer plywood, laminated plywood, and fiberboard can be attacked and destroyed by fungi. However, the use of chemical preservatives is sensible only if the adhesives used in the production of these wood materials—e.g., alkaline-hardening phenoplastics, aminoplastics reinforced with phenol–resorcinol resins (melamine–urea–phenol–formaldehyde MUPF), or polymeric diphenylmethane diisocyanates (PMDI)—are sufficiently resistant even at high wood moisture levels. In temperate climates, plywood needs to be protected against insects if it is made from tropical woods such as abachi or limba. In the tropics and similarly warm areas, plywood and particle or fiberboard can be damaged by insects, including termites.

7.4.1. Treatment with Preservatives during Production

In *particle board* production, mixing the preservative with the adhesive is the preferred method. Formulations or powders that can be diluted with water, highly concentrated salt solutions, or solvent-borne preparations are mixed with adhesive and other additives and sprayed onto the particles. If the preservative and adhesive cannot be mixed for technical reasons, they can be applied to the particles separately, for example, in the mixer. Also, preservatives in powder form can be mixed with glue-blended particles and dispersed. In the production of *veneer plywood boards*, the preservative is also generally mixed with the adhesive. The quantity of preservative added depends on the construction of the board (i.e., number of veneers and quantity of adhesive applied). Since the preservative must penetrate the veneers from the adhesive seam, use of this process depends on the thickness of the veneers. Veneered laminated boards with thicker veneers can be treated with preservative after production (see Section 7.4.2).

For *laminated plywood* (laminated board, blockboard), strips and face veneers can also be treated by dipping or spraying if penetration of the preservative is sufficient. Otherwise, the strips must be treated by using a pressure process. After impregnated wood has been dried, adhesive is applied and the boards are pressed together.

Treatment of *fiberboard* by adding preservative to the pulp slurry causes problems in the commonly used wet process because the active substance if partly lost when water is pressed out [183].

7.4.2. Treatment with Preservatives after Production

After production, *plywood and veneer plywood* can be impregnated under pressure with salt solutions or solvent-containing preparations. However, the disadvantage is that because of moisture absorption, drying is necessary before processing can be carried out. In another method, trimmed boards, which are still warm from the press, are treated by dipping in an aqueous preservative solution. However, care must be taken so that the boards do not warp.

Fiberboard can be treated after production with aqueous or solvent-containing preservatives by dipping or pouring [184, 185].

7.5. Special Processes

Special processes are used if:

1) After a long period of use the original preservative content needs to be supplemented or increased in building (e.g., beams)
2) The depth of penetration and distribution of the preservative needs to be improved in wood that is not easily impregnated (e.g., spruce or fir)
3) Protection of wood used in building against fungi or insects is required, which cannot be achieved with usual impregnating processes

4) Areas that are particularly at risk (e.g., wood in the below/aboveground zone) need additional protection
5) Only partial wood preservation measures are required (e.g., protection of the ends of joists)

Protection with Bandages. Poles (e.g., for electric and telephone transmission) should exhibit long-term durability. The below/aboveground zone of these poles, which is particularly at risk from attack by fungi (predominantly soft rot fungi), can be protected by bandages. waterborne wood preservatives (e.g., based on fluorine or boron) can be worked into the carriers of these bandages, which are made of foam or other porous materials. The moisture present in roundwood allows deep penetration of preservative. The 50–70 cm-wide bandages are placed around the poles in one piece and generally renewed at intervals of five to ten years. For subsequent treatment or for increased protection in the below/aboveground zone, the injection process (Cobra process) has proved effective [186]. A hollow needle is injected into the wood to a depth of a few centimeters at defined intervals, and a water-soluble salt preservative paste is pressed into the wood through the needle. The preservative is dispersed in the wood by diffusion.

Incising Processes. To improve the depth of penetration and dispersion of preservative in wood that is difficult to impregnate, various incising (perforation) processes are used [187]. The perforations can be made by using rotating or fixed needles or piercing knives. The depth of piercing is generally 10 mm or more [188]. For planed wood, the depth is much less [189 – 191]. Laser processes have also been proposed for perforation [192, 193].

Bored-hole processes are used when building components cannot be impregnated sufficiently by brushing or spraying (e.g., in case of attack by insects) or when beam joints, footings, and beams in contact with walls and soil, or buried palisade wood, are endangered by the action of moisture. The wood is generally bored downward at an angle. The bored holes, with a diameter of 8–15 mm, are filled with a preservative one or more times and plugged. To impregnate as large an area as possible, pressers are frequently used. Here preservative is pressed into the wood by compressed air via a hollow drill. Solid salt cartridges (die-pressed pieces) (e.g., based on boron or fluorine) are used in bore holes for the posttreatment of damp wood or wood that is becoming damp. The active substance is then dispersed by diffusion [194]. To control wood-destroying fungi in window frames, this type of treatment using boron-based cartridges and fluoride injection has proved successful [195 – 197].

Foam Process. By using the foam process, wood can be made to absorb relatively large quantities of preservative in a short time [198]. Water, which contains the dissolved preservative, is foamed after the addition of a surfactant. The foam is applied to the wood and breaks down slowly, whereby the active substances penetrate the wood. The process is used for prevention of and combating attack by pests.

For the use of agents that are active in the gas phase (fumigants) and can be used, in particular, to control wood-destroying fungi inside wooden poles, see Section 2.5.

8. Environmental Aspects

8.1. Health Aspects

Like plant protection agents, wood preservatives are assessed by health and environmental authorities with regard to their acute, subchronic, and chronic toxicity. During production and use of preservatives and their mixtures, direct contact (e.g., via dust, vapors, aerosols, and spray mists) cannot be excluded. Thus, investigations on possible carcinogenicity [e.g., of arsenic and chromium(VI) compounds or BaP in coal-tar oil], mutagenicity, or teratogenicity are particularly important. Knowledge of the maximum concentration at the workplace (MAK, TLV) is also important. Since wood preservatives are also used in inhabited indoor areas, evaluation of the effect on occupants through liberation of active substances from preserved wood is particularly important (see Section 8.2).

8.2. Soil, Water, and Air

Besides investigating the toxicology of wood preservatives, the quantities of active substances

that can get into the soil or water by leaching and into the air by evaporation must also be known. These proportions depend on the application process and the quantity applied to the wood. For organic active substances and coal-tar oil components, investigations on the behavior in soil and on biodegradability (partial or complete) are important. For partial degradation the metabolites should be identified if possible. Other criteria for risk assessment could be adsorption or desorption on soils and sediment, degradation and metabolism in water–sediment systems under anaerobic conditions, and bioaccumulation. With regard to the risk estimation for ground-, drinking, and surface water, knowledge of the percolating and mobility behavior of active substances (in some cases, their metabolites), inorganic salts, and salt mixtures is important. In the case of salts and salt mixtures, boron, fluorine, and the heavy metals arsenic, chromium, and copper play an important role. They can be leached out of treated wood or enter soil because of poor operational hygiene in impregnating plants or during open tank impregnation. In most countries, tests of possible harmful effects on aquatic or terrestrial organisms are required for the ecotoxicological evaluation of wood preservatives. The fish toxicity test [199] can, for example, be carried out according to OECD Test Guideline 203, Testing Guidelines for Chemicals [200]. Depending on use, investigations of the effects of wood preservatives on pets, working, and other animals and on ornamental, useful, and other plants may be necessary.

A five-year observation period of poles impregnated with creosote established that the components that had leaked out had obviously been degraded by microorganisms because they could no longer be detected [201]. *Pseudomonas* strains, in particular, are thought to be involved in biodegradation [202, 203]. For wood that is used in contact with water, a severe effect on the taste of fish and water was established due to the washing out of phenols and other water-soluble compounds. Even traces make fish unpalatable [204].

Wood impregnated with Cu–Cr–As is particularly resistant to leaching and weathering, compared with other chromium-based salt preservative combinations, such as F–Cr–As, Cu–Cr–F, Cu–Cr–B, and Cu–Cr. However, arsenic, copper, and chromium can be leached out in varying quantities. The degree of mobility and the bioavailability of these elements are affected by the type and structure of the soil and the pH [205 – 214]. Fluorine and boron can be leached out relatively easily [215].

Acid rain has little effect on the leaching of copper, chromium, or arsenic, as model experiments on Cu–Cr–As-impregnated wood have shown. At pH 3.0, copper and arsenic are leached out to a greater, and chromium to a lesser, degree [216, 217].

Environmental pollution by preservative components, particularly in impregnating and dipping plants, has to some extent impaired soil and both ground- and drinking water [218 – 220]. Considerable quantities of preservative components [e.g., toxicologically harmful chromium(VI) compounds or arsenic] can enter the environment before they are fixed in the wood through dripping or the action of rain. Salt preservatives can enter soil or wastewater from wood that has been impregnated with water-soluble boron salts or with silico- or hydrogenfluorides, if it is stored unprotected.

To avoid dripping of unfixed preservative solution containing chromate(VI) from impregnated wood, the fixing period can be reduced from 8–30 d to less than 1 h, for example, by using steam at 110 °C [221]. Chromium(III) compounds formed in the fixing process are considered less toxic than chromium(VI) compounds.

Impregnating plants, dipping tanks, and drip collection installations for wood treated with preservatives must be erected on reinforced areas with the appropriate collecting equipment for excess preservative, and stored impregnated wood must be protected from rain for environmental reasons. Impregnating tanks, dipping tanks, spraying tunnels, piping, and pumps must be watertight, and wastewater must be disposed of correctly. For example, no preservative-containing solution should be passed into wastewater channels or sewage plants.

In the European Union and in countries like the United States and Canada, health/environmental authorities require an exhaustive risk assessment of health and environmental impacts not only for the treated wood in-service, but also for production and processing of wood preservatives, for example, in impregnating plants. For certain areas of use, inves-

tigations into the possible absorption of active substances by foods can be required.

No standard test procedures have yet been worked out that stipulate conditions for determination of the concentration of active substances that can be regarded as posing no risk to health indoors.

For risk assessment, the volatility and vapor pressure of the active substance and its acute or subchronic inhalation toxicity, as well as the acute LC_{50} for inhalation of a formulation of this substance, can be determined.

In association with the evaporation of wood preservatives, particularly organic compounds from treated interior surfaces [222], various models for risk assessment with regard to possible health hazards have been developed in Europe [223]. Insecticides or fungicides emitted to the air can be absorbed by humans through inhalation. This can take place over several years.

In long-term tests at 20 °C the concentrations of lindane, endosulfane, dichlofluanid, and pentachlorophenol in the air of living rooms were determined by gas chromatography. The type of wood and associated depth of penetration of preservative, the surface area : room space ratio, and the air exchange rate had a significant effect on the concentration of active substances [224].

Risk assessment can be expressed in terms of the following formula:

$$\text{Acceptable air concentration (AAC)} = \frac{\text{acceptable daily intake (ADI)}}{V}$$

where V is the volume of air that is inhaled by a human of 70 kg weight over 24 h; this is assumed to be 18.5 m^3 [225, 226]. ADI values have a safety margin of 100.

8.3. Disposal of Treated Wood

The disposal of wood treated with preservatives is very important with regard to environmental pollution. For example, the inorganic preservatives arsenic, chromium, copper, and mercury, and the organic preservatives pentachlorophenol, other chlorinated cyclic compounds (e.g., chloronaphthalene, endosulfane, and lindane), and coal-tar oil pose a particular hazard to the environment. In many countries, particularly Germany [227], laws and regulations have been drawn up that must be adhered to in the disposal of treated wood.

At the end of its service life, treated wood becomes waste wood. This category also includes waste pieces produced during production and processing. The choice of disposal method depends on the polluting potential of the preservative and the amount of it in the wood. Color reactions can be used to detect inorganic compounds [137], but generally not organic ones. The Beilstein test, frequently described in the literature for detecting pentachlorophenol and other halogen-containing compounds, is not very reliable [140]. A specific test has been developed for pentachlorophenol itself [227]. In quantitative analyses (see Chap. 6), the sampling procedure before actual analysis is important for obtaining a representative result [141, 228].

Recycling. The recycling of treated waste wood is very limited. Railway ties that are no longer suitable for a main line can still be used for a long period on a branch line, where they are less stressed. Electricity poles can be used as fence posts and railway ties in embankments [229].

For large wooden objects such as poles, railway ties, and girders, the outer layers of wood treated with preservative can be removed and incinerated. The remaining wood, which is largely preservative free, can then be reused.

Disposal. The disposal of treated wood in dumps can be problematic if preservative components are leached out or liberated through degradation of the wood, thus polluting soil, groundwater, or air.

According to the current state of the art, *incineration* is the best form of disposal for wood treated with preservatives. The important prerequisites for this are an incineration technique that allows almost complete combustion and effective dust collection to minimize emissions.

Inorganic preservatives are not destroyed by incineration. Arsenic [230] and fluorine are particularly important since they enter flue gas because of their volatility and are more difficult to collect by dust separation than chromium and copper, which remain in the ash and the filter.

Boron-containing preservatives can lower the melting point of the ash considerably. This can lead to corrosion and slag problems in the incineration chamber [231]. Boric acid itself can evaporate together with the steam produced in incineration.

When wood treated with creosote is mixed with untreated wood in the ratio 1 : 4, under optimum incineration conditions the concentrations of CO, NO_x, and hydrocarbons are in the same range as those found for untreated wood [232].

Organic aromatic substances containing a variety of heteroatoms in the ring can be assumed to be more readily degraded thermally and oxidatively than others, such as pentachlorophenol. The formation of PCDDs and polychlorinated dibenzofurans (PCDFs) during incineration of PCP-treated wood can be kept low by using temperatures >500 °C and a good air supply [231, 233].

A new direction of research is *biological disposal* of wood treated with salt preservatives by the use of selected microorganisms with special physiological properties. The aim of this method is to recover heavy metals such as arsenic, chromium, and sometimes copper. The leached wood can then be disposed of in a dump or composted [234].

Promising experiments have been carried out in which chromate-containing preservatives have been converted into a water-soluble form by the use of inorganic and organic acids and then washed out of the wood [235 – 237].

9. References

1. A. J. Nurmi, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/50033** ,1994.
2. *Ullmann*, 4th ed, **12**, 685.
3. F. Mahlke, E. Troschel: *Handbuch der Holzkonservierung*, 3rd ed., Springer-Verlag, Berlin 1960.
4. J. P. Vité: *Die holzzerstörenden Insekten Mitteleuropas*, Messerschmidt Wissenschaftl. Verlag, Göttingen 1952.
5. W. Bavendamm: *Der Hausschwamm und andere Bauholzpilze*, G. Fischer-Verlag, Stuttgart 1969.
6. B. Schulze, G. Becker, *Holzforschung* **2** (1948) 97–128.
7. G. Becker, *Holz Roh Werkst.* **22** (1964) 51–57.
8. G. Becker, W. Kempf, *Materialprüfung* **2** (1960) 301–307.
9. K. Hillner, G. Streckert, *19. Holzschutz-Tagung 1992*, Rosenheim, Oct. 7–8, 1992, Deutsche Gesellschaft für Holzforschung, München, p. 99.
10. H. B. van Groenou, *Holz Zentralbl.* **106** (1980) no. 48, 708–709.
11. O. Wälchli, *Schweiz. Z. Forstwes.* **119** (1968) 446–464.
12. O. Wälchli, *Schweiz. Arbeitsgemeinschaft SAH-Bull.* **4** (1976) no. 3.
13. D. V. Plackett, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3238**, IRG, Drottning Kristinas vag 37C, S-114 28 Stockholm, May 1983.
14. R. D. Arsenault, J. M. Kressbach, R. F. Fox, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **87** (1991) 17–40.
15. D. K. Stokes, M. H. Freeman, T. L. Woods, R. D. Arsenault, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30045**, 1994.
16. T. C. Anderson et al., *Proposal for the Standardization of Ammoniacal Copper Citrate*, Osmose Wood Preserving, Inc., submitted to AWPA Committee P-4 and the General Preservatives Committee, 1992.
17. M. Connell, J. A. Cornfield, G. R. Williams, Proc. British Wood Preserving and Damp-proofing Assoc. Convention, 1993, 1–18.
18. American Wood Preservers Association Standards, 2002. ISSN 1534-195X.
19. F. Cui, A. R. Zahora, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/00-40168, 2000.
20. F. G. Evans, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/03-30303, 2003.
21. K. Suzuki, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/03-10463, 2003.
22. J. A. Drysdale, M. E. Hedley, E. Loh, L. T. Hong, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/00-30213, 2000.
23. R. N. Wakeling, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/01-20231, 2001.
24. E. Hickmann, E. Hädicke, W. Reuther, *Tetrahedron Lett.* (1979) no. 26, 2457–2460.
25. G. Klebe et al., *Z. Kristallogr.* **211** (1996) 798–803.
26. BASF, DE 2 4 10 603, 1974 (W. Reuther, E.-H. Pommer, P. Raff, R. Göttsche).
27. R. Göttsche, N. Marx, *Holz Roh Werkst.* **47** (1989) 509–513.

28. W. Hettler, S. Breyne, M. Maier, *19. Holzschutz-Tagung 1992*, Rosenheim, Oct 7–8, 1992, Deutsche Gesellschaft für Holzforschung, München p. 217.
29. BASF, Xyligen 25 F, Technical Leaflet, Ludwigshafen 1992.
30. J. A. Drysdale, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30037**, 1994.
31. G. Becker, *Holz Roh Werkst.* **17** (1959) 484–489.
32. J. T. de Jonge, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3400**, Dec. 1986.
33. M. Hamel (ed.): *Proc. First International Conference on Wood Protection with Diffusible Preservatives*, Forest Product Research Society, Madison 1990.
34. J. A. Butcher, J. Drysdale, *Mater. Org.* **12** (1977) 271–277.
35. J. A. Butcher, J. Drysdale, *N. Z. For. Sci* **8** (1978) 403–409.
36. R. J. Tillot, C. R. Coggins, *Annu. Conv. Br. Wood Pres. Assoc.* 1981, 32–46.
37. H. Becker, *Seifen Öle Fette Wachse* **109** (1983) 603–606.
38. H. Becker, *Seifen Öle Fette Wachse* **110** (1984) 15–17.
39. A. F. Preston, P. J. Walcheski, P. A. McKaig, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **83** (1987) 331–348.
40. H. Becker, *Seifen Öle Fette Wachse* **115** (1989) 681–684.
41. D. D. Nicholas, A. F. Preston, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **76** (1980) 13–21.
42. L. Jin, A. F. Preston, *Holzforschung* **45** (1991) 455–459.
43. J. W. Creffield, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/93–30009, May 1993.
44. J. A. Butcher, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/3328, 1985.
45. D. Blow, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/2250, 1986.
46. L. Walker, EP 1 122 044 B1, Lonza AG, 2005.
47. F. Lichtenberg, J. Fritschi, V. Ranft, US 6362370 B1, Lonza AG, 2002.
48. H. Härtner, V. Barth, *19. Holzschutz-Tagung 1992*, Rosenheim, Oct 7–8, 1992, Deutsche Gesellschaft für Holzforschung, München p. 87.
49. V. Barth, H. Härtner, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30014**, 1993.
50. H. Härtner, V. Barth, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30097**, 1996.
51. H. Härtner, V. Barth, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/93–30014**, May 1993.
52. O. Exner, W. Paulus, *III. Enconto Brasileiro em Preservacao de Madeiras*, Sao Paulo, Oct. 25–27 1989.
53. T. Hof, J. G. A. Luijten, *Timber Technol.* **67** (1959) 83–84.
54. H. Becker, *Seifen Öle Fette Wachse* **113** (1987) 773–776.
55. H. Becker, *Seifen Öle Fette Wachse* **114** (1988) 99–100.
56. J. Jermer et al., *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3219**, 1983.
57. M.-L. Edlund, J. Jermer, B. Henningsson, W. Hintze, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3329**, 1985.
58. J. Jermer, M.-L. Edlund, B. Henningsson, W. Hintze, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/2244, 1985.
59. M. Connell, J. A. Cornfield, G. R. Williams, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3573**, 1990.
60. T. E. Dudley-Brendell, D. J. Dickinson, *Int. Res. Group in Wood Preservation*, **Doc. no. IRG/WP/1156**, 1982.
61. E. K. Engdahl, R. T. Baileys, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **88** (1992) 268–287.
62. E. A. Hilditch in R. Thompson (ed.): *The Chemistry of Wood Preservation*, The Royal Society of Chemistry, Cambridge 1991, p. 81.
63. T. P. Schultz, D. D. Nicholas, L. L. Ingram, T. H. Fisher, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30114**, 1996.
64. C. H. Palmer, G. D. Rosebery: *Proposal to Amend Specified AWPA Pressure Treatment Standards to Include Neutral, Solubilized Oxine Copper for Above-ground Use*, Maag Agrochemicals Inc., submitted to AWPA Treatment Committees, 1991.
65. A. Leclercq, *Mater. Org.* **18** (1983) 65–77.
66. A. R. Valcke, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3529**, 1989.
67. O. Exner, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3680**, 1991.
68. H.-W. Buschhaus, A. R. Valcke, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30092**, 1995.
69. B. Wüstenhöfer, H.-W. Wegen, W. Metzner, *Holz Zentralbl.* **119** (1993) no. 58, 984, 988.
70. Sung-Mo Kang, J. J. Morrell, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/03-40259, 2003.

71. T. L. Woods, J. P. Bell, *Am. Wood-Preserv. Assoc.* **86** (1990) 190–196.
72. L. J. Cookson, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30041**, 1994.
73. J. W. Creffield, T. L. Woods, N. Chew, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30101**, 1996.
74. P. E. Laks, K. W. Gutting, R. C. de Groot, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/97-30152, 1997.
75. J. J. Morrell, C. M. Freitag, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/99-30201, 1999.
76. D. Greenly, B. Hegarty, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3495**, 1988.
77. L. Leightley, D. Nicholas, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3612**, 1992.
78. A. F. Preston, K. J. Archer, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/20095**, 1996.
79. W. Metzner, G. Buchwald, S. Cymorek, H. Hinterberger, *Holz Roh Werkst.* **35** (1977) 233–237.
80. A. F. Bravery, D. J. Dickinson, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/2215**, 1984.
81. H.-U. Buschhaus, P. Spetmann, T. Jaetsch, M. Kugler, EP 1 046 337 B1, Bayer Chemicals AG, 2005.
82. J. Hansen, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3295**, 1984.
83. S. R. Cook, J. Sullivan, D. J. Dickinson, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/02-10437, 2002.
84. N. Morsing, B. Lindegaard, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/04-20302, 2004.
85. K. Tsunoda, H. Kumagai, M. Sakurai, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/94-30035, 1994.
86. J. Snow, P. R. S. Cobham, N. Ryan, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/00-30229, 2000.
87. D. V. Plackett, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/3198, 1982.
88. J. A. Drysdale, R. M. Keirle, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3376**, 1986.
89. M. Rose, A. Bedoya, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3402**, 1987.
90. D. A. Lewis, G. R. Williams, R. A. Eaton, *Rec. Annu. Conv. Br. Wood Preserv. Assoc.* 1985, 14.
91. Y. Nomura, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3639**, 1991.
92. W. Verbecke, J. van Acker, M. Stevens, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/2236**, 1989.
93. R. van den Eynde, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3606**, 1990.
94. J. P. La Fage in D. R. Houghton, R. N. Smith, H. O. W. Higgins (eds.): *Biodeterioration 7*, Elsevier Applied Science, London 1987, p. 721.
95. M. M. Serment in D. R. Houghton, R. N. Smith, H. O. W. Higgins (eds.): *Biodeterioration 7*, Elsevier Applied Science, London 1987, p. 762.
96. N.-Y. Su, M. Tamashiro, J. R. Yates, M. I. Haverty, *J. Econ. Entomol.* **75** (1982) 188–193.
97. R. Grüning, R. Pospischil, S. Cymorek, W. Metzner, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/1284, 1986 (Poster).
98. G. Rustenburg, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/95-30076, 1995.
99. M.-Y. Liu, J. E. Casada, *Pestic. Biochem. And Physiol.* **46** (1993) 200–206.
100. A. H. H. Wong, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/05-30389, 2005.
101. J. Doyle, K. Webb, W. R. Rae, A. F. Siraa, J. Malcom-Black, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/03-30325, 2003.
102. T. Jaetsch, R. Bruns, P. Spetmann, WO 2005/063023 A1, Lanxess Deutschland GmbH, 2005.
103. E. Graf, M. Barkhoff, R. Hamberg, H. Büttner, M. Pallaske, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/02-30277, 2002.
104. A. Valcke, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/95-30080, 1995.
105. A. Valcke, M. Pallaske, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/95-30079, 1995.
106. R. D. Fears, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30047**, 1994.
107. B. M. Kard, J. K. Mauldin, S. C. Jones, *Pest Control* **57** (1989) no. 10, 58, 60, 68.
108. J. K. Mauldin, R. H. Beal, *Pest Control* **56** (1987) no. 10, 46–59.
109. M. Lenz, J. A. L. Watson, R. A. Barrett, S. Runko. *Sociobiology* **17** (1990) 9–36.

110. B. M. Kard, J. K. Mauldin, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30043**, 1994.
111. *Int. Pest Control* **35** (1993) no. 6, 149.
112. E. L. Schmidt, B. Kreber, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/00-10343, 2000.
113. N. R. Ruddick, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3253**, May 1983.
114. T. L. Highley, W. E. Eslyn, *Holzforschung* **43** (1989) no. 4, 225–230.
115. T. L. Highley, W. E. Eslyn, *Holzforschung* **45** (1991) no. 3, 223–228.
116. A. R. Zahora, J. J. Morell, *Wood Fiber Sci.* **21** (1989) no. 1, 55–66.
117. A. R. Zahora, J. J. Morell, *Wood Fiber Sci.* **21** (1989) no. 4, 343–353.
118. D. D. Nicholas, R. Cockcroft: "Wood Preservation in the USA," *STU-Inf.* **248** (1982) 64.
119. E. Schwenk, *Dtsch. Farben Z.* **14** (1960) 89–95.
120. R. Paton, *Rec. Annu. Conv. Br. Wood Preserv. Assoc.* 1982, 3–7.
121. G. Technau, W. Behrenz, *Holz Roh Werkst.* **16** (1958) 90.
122. R. W. Berry, *Int. Biodeterior Bull.* **8** (1972) 141.
123. S. Cymorek, *Holz Roh Werkst.* **33** (1975) 239.
124. DIN prEN 20 Wood preservatives; Determination of the Preventive Action Against Lyctus brunneus (Stephens), Beuth-Verlag, Berlin 1990.
125. G. Becker, *Materialprüfung* **12** (1970) 293.
126. European Standard EN 599-1, 1996.
127. F. W. Brooks in R. Thompson (ed.): *The Chemistry of Wood Preservation*, The Royal Chemical Society, Cambridge 1991, p. 245.
128. D. Rudolph, H. Willeitner, *19. Holzschutz-Tagung 1992, Rosenheim, Oct. 7–8, 1992*, Deutsche Gesellschaft für Holzforschung, München, p. 101.
129. H. Willeitner, H.-J. Irmschler, *Holz Zentralbl.* **120** (1994) no. 30, 477, 480.
130. A. F. Bravery in R. Thompson (ed.): *The Chemistry of Wood Preservation*, The Royal Chemical Society, Cambridge 1991, p. 225.
131. JIS A 1301–1975, Method of Fire Test for Wooden Structural parts of Buildings.
132. JIS A 1311–1975, Method of Fire Protecting Test of Fire Door for Building.
133. US ASTM E 84–77 a, Test for Surface Burning Characteristics of Building Materials.
134. US ASTM E 119–78, Methods of Fires Tests of Building Construction and Materials.
135. J. G. Wilkinson: *Industrial Timber Preservation*, Associated Business Press, London 1979.
136. DIN 4102, Brandverhalten von Baustoffen und Bauteilen, Beuth-Verlag, Berlin, Part 1, 1976, Part 2 1977..
137. G. Theden, C. Kottlors, *Mitt. Dtsch. Ges. Holzforsch.* **52** (1965) no. 52, 1–88.
138. H. Becker, *Seifen Öle Fette Wachse* **93** (1967) 1009.
139. H.-J. Petrowitz, *Holz Zentralbl.* **119** (1993) no. 49, 747, 748, 750.
140. H.-J. Petrowitz, *Holz Zentralbl.* **115** (1989) no. 47, 702.
141. C. Kottlors, H.-J. Petrowitz, *Holz Roh Werkst.* **51** (1993) 126–134.
142. A. I. Williams, *Analyst (London)* **97** (1972) 104.
143. A. Visapaa, P. Mansikkamaki, *Par. Puu* **61** (1979) 253.
144. H.-J. Petrowitz, W. Berghoff, *Mater. Org.* **7** (1972) 1.
145. A. I. Williams, *Analyst (London)* **95** (1970) 498.
146. B. G. Henshaw, J. W. Morgan, N. Williams, *J. Chromatogr.* **110** (1975) 37–41.
147. H. Opfermann, *BIT-Fachzeitschrift f. d. Laboratorium Supplement Chromatographie* **4/85** (1985) 56–64.
148. H.-J. Petrowitz, C. Kottlors, *Holz Zentralbl.* **118** (1992) no. 21, 1919–1920.
149. R. Marutzky, *DGFH-Nachrichten* 1993, no. 57, 19.
150. DIN EN 350–2, Durability of Wood and Wood Based Products—Natural Durability of Solid Wood—Part 2: Guide to the Natural Durability Selected Wood Species of Importance in Europe, Oct. 1994, Beuth-Verlag, Berlin.
151. S. Amemya, R. Cockcroft: "Wood Preservation in Japan," *STU-Inf. 272* (1982) 81–85.
152. J. J. Morrell, A. F. Preston, *Proc. Ann. Meet. Am. Wood-Preserv. Assoc. 91st* **81** (1995) 63–81.
153. H. Willeitner in R. v. Halász (ed.): *Holzbau-Taschenbuch*, W. Ernst & Sohn, Berlin 1974, p. 55.
154. G. M. Hunt, G. A. Garrat: *Wood Preservation*, 3rd ed., McGraw-Hill, New York 1967.
155. DIN prEN 599, Durability of Wood and Derived Materials; Performance of Wood Preservatives as Determined by Biological Tests; Part 1: Specifications According to Hazard Class, Apr. 1992, Beuth-Verlag, Berlin.

156. DIN prEN 599, part 2: Classification and Labelling, Apr. 1992, Beuth Verlag, Berlin.
157. DIN EN 335, Durability of Wood and Wood Based Products: Definition of Hazard Classes of Biological Attack; Part 1: General, Sept. 1992, Beuth-Verlag, Berlin.
158. DIN EN 335, Part 2: Application to Solid Wood, Oct. 1992, Beuth Verlag, Berlin.
159. T. L. Highley, T. C. Scheffer, *Int. Res Group on Wood Preservation*, **Doc. no. IRG/WP/93–30015**, May 1993.
160. H. Bellmann, *Holz Roh Werkst.* **32** (1974) 135–143, 340–352, 33 (1975) 100–108.
161. H. Bellmann, *Holz Zentralbl.* **91** (1965) 2436, 2574–2576.
162. O. B. Lowry, US 831 450, 1905.
163. M. S. Hudson, *Proc. Annu. Mect. Am. Wood Preserv. Assoc.* **49** (1953) 146–171.
164. W. T. Henry, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **59** (1963) 68–78.
165. D. S. Belford, *Holz Roh Werkst.* **26** (1968) 197–201.
166. S. T. Henriksson, *Holz Roh Werkst.* **12** (1954) 233–241.
167. O. Wälchli, *J. Inst. Wood Sci.* **5** (1970) no. 2, 42–43.
168. R.-D. Peek, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/3449, 1987.
169. S. Goetsch, R.-D. Peek, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/3668, 1991.
170. O. Hendriksen, 23. Holzschutz-Tagung 2003, Augsburg, March 26–27, 2003, Deutsche Gesellschaft für Holzforschung, München, p. 195.
171. Sung-Mo Kang, J. J. Morrell, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/04-40276, 2004.
172. M. Muin, A. Adachi, K. Tsunoda, Int. Res. Group on Wood Preservation, Doc. no. IRG/WP/01-40199, 2001.
173. Atlasmaste, DE 832 327, 1942 (H. Gewecke).
174. M. Gersonde, *Holz Zentralbl.* **94** (1968) 599–601.
175. M. Gersonde, C. Kottlors, *Mitt. Dtsch. Ges. Holzforsch.* **57** (1970) 77–86.
176. O. Wälchli, *Bull. Schweiz. Elektrotech. Ver.* **63** (1972) 2–7.
177. G. Gruber, AT 176 014, 1951.
178. M. E. Hedley, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **86** (1990) 31.
179. H. M. Barnes, T. L. Amburgey, L. M. Williams, J. J. Morrell, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3542**, 1989.
180. P. Vinden, J. Drysdale, M. Spence, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3632**, 1990.
181. R. H. Baechler, H. G. Roth, *For. Prod. J.* **14** (1964) no. 4, 171–178.
182. D. J. Miller, R. D. Graham, *For. Prod. J.* **20** (1970) no. 6, 43–44.
183. H.-J. Deppe, K. Ernst: *Fortschritte in der Spanplattentechnik*, DRW-Verlag, Stuttgart 1973.
184. H.-J. Deppe, M. Gersonde, *Holz Zentralbl.* **106** (1980) 139–142.
185. M. Gersonde, H.-J. Deppe, *Holz Roh Werkst.* **41** (1983) 323–328.
186. Cobra, DE 827 862, 1950 (H. Peter).
187. P. I. Morris, J. J. Morrell, J. N. R. Ruddick, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/40019**, 1993.
188. M. Anderson, J. J. Morrell, J. E. Winandy, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/40093**, 1997.
189. H. Bellmann, *J. Inst. Wood Sci.* **4** (1963) no. 3, 54–63.
190. J. N. R. Ruddick, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **81** (1985) 148–160.
191. J. N. R. Ruddick, *Holz Roh Werkst.* **44** (1986) 109–113.
192. B. Goodell, F. A. Kamke, J. Liu, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3646**, 1991.
193. J. N. R: Ruddick, *For. Prod. J.* **41** (1991) no. 4, 53–57.
194. T. L. Highley, W. Finney, F. Green III, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/30042**, 1994.
195. J. K. Carey, A. F. Bravery in D. R. Houghton, R. N. Smith, H. O. W. Eggins (eds.): *Biodeterioration 7*, Elsevier Applied Science, London 1988, p. 733.
196. E. M. Schmidt in M. Hamel (ed.): *Proc. First International Conference on Wood Protection with Diffusible Preservatives*, Forest Products Research Society, Madison 1990.
197. M. D. Dietz, E. L. Schmidt, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3432**, May 1987.
198. H.-J. Rafalski, B. Biering: *Holzschutz-Beschäumungsverfahren. Die ökologische Alternative zum Spritzverfahren*, Leipziger Gesellschaft für Bausanierung; Forschungsanstalt für Forst- und Holzwirtschaft Eberswalde, Leipzig 1990.
199. H.-W. Wegen, *Holz Zentralbl.* **114** (1988) 653–654.

200. OECD-Guideline for Testing of Chemicals no. 203: Fish, Acute Toxicity Test, May 1981.
201. H.-J. Petrowitz, *Holz Roh Werkst.* **47** (1989) 414.
202. M. Gersonde, *Holz Roh Werkst.* **31** (1973) 89–96.
203. W. Kerner, *Mater Org*. Beiheft 3 (1975) 319–330.
204. B. Henningsson, *Holz Roh Werkst.* **41** (1983) 471–475.
205. R. D. Arsenault, *Proc. Annu. Meet. Am. Wood Preserv. Assoc.* **71** (1975) 126–147.
206. F. G. Evans, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3122**, 1978.
207. B. Henningsson, B. Carlsson, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3149**, 1984.
208. F. G. Evans, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3433**, 1987.
209. E. B. Jonsson, E. M. A. Nilsson, J. N. R. Ruddick, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3537**, 1989.
210. J. N. R. Ruddick, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3577**, 1990.
211. H. Kipp, H. Willeitner, K. Brandt, A. Müller-Grimm, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3669 and 3669 Add.**, 1991.
212. H. Willeitner, H. Klipp, K. Brandt, *Holz Roh Werkst.* **49** (1991) 140.
213. H. Klipp, K. Brandt, *19. Holzschutz-Tagung 1992, Rosenheim, Oct. 7–8, 1992*, Deutsche Gesellschaft für Holzforschung, München, p. 241.
214. R. J. Orsler, G. E. Holland, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/93–30026**, 1993.
215. M. G. Dietz, E. L. Schmidt, *For. Prod. J.* **38** (1988) no. 5, 9–14.
216. R. J. Murphy, D. J. Dickinson, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3579**, 1990.
217. K. Jae-Jin, K. Gyu-Hyeok, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/93–50004**, 1993.
218. J. Bergholm, *Int. Res. Group on Wood Preservation*, Doc. no. IRG/WP/3571, 1990.
219. R.-D. Peek, H. M. Illner, H. Klipp, *19. Holzschutz-Tagung 1992, Rosenheim, Oct. 7–8, 1992*, Deutsche Gesellschaft für Holzforschung, München, p. 293.
220. R.-D. Peek, H. Willeitner, K. Brandt, *Holz Roh Werkst.* **45** (1987) 59–62.
221. R.-D. Peek, H. Willeitner, *Holz Zentralbl.* **112** (1988) 1590, 1594–1596.
222. A. Piccard, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3458**, 1990.
223. A. Valcke, L. van Leemput, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3565**, 1990.
224. H.-J. Petrowitz, *Holz Roh Werkst.* **44** (1986) 341–345.
225. A. E. Dobbs, N. Williams, *Environ. Pollut. Ser. B* **6** (1983) 271–276.
226. J. van Leemput, V. E. Swysen, W. Meuldermans, J. Heykants, *Chemosphere* **16** (1987) 1281–1288.
227. R. Marutzky, R.-D. Peek, H. Willeitner: *Informationsdienst Holz–Entsorgung von schutzmittelhaltigen Hölzern und Reststoffen*, Deutsche Gesellschaft für Holzforschung, München 1993, p. 35.
228. W. J. Homan, H. Militz, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/50024**, 1994.
229. H. Willeitner, *Holz Zentralbl.* **114** (1990) no. 27, 393, 395–396, 398.
230. C. K. McMahon, P. B. Bush, E. A. Woolson, *For. Prod. J.* **36** (1986) no. 11/12, 45–50.
231. R. Marutzky, *Holz Roh Werkst.* **48** (1990) 19–24.
232. T. Salthammer et al., *Holz Roh Werkst.* **52** (1994) 247–252.
233. T. Salthammer, H. Klipp, R.-D. Peek, R. Marutzky, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/50019**, 1994.
234. I. Stephan, R.-D. Peek, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/3717–92**, 1992.
235. E. A. Pasek, C. R. McIntyre, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/93–50007**, May 1993.
236. I. Stephan, H. H. Nimz, R.-D. Peek, *Int. Res. Group on Wood Preservation*, **Doc. no. IRG/WP/93–50012.**
237. N. Marx, DE 4 1 14 178, 1992.

Wood, Surface Treatment

see also → Wood, Chap. 2.4.

Wilfried Hansemann, Votteler Lackfabrik, Korntal-Münchingen, Federal Republic of Germany

1. Introduction 1
2. Bleaching 3
3. Staining 3
3.1. Stain Systems 3
3.2. Application of Stains 4
4. Coating Systems 4
4.1. Cellulose Nitrate Lacquers 4
4.2. Polyurethane Coatings 5
4.3. Unsaturated Polyester Coatings . . 6
4.4. Waterbased Paints 7
5. Paint Application 8
6. Paint Curing Methods 8
6.1. Heat-Drying Systems 8
6.2. Radiation-Curing Systems 9
6.2.1. IR Irradiation 9
6.2.2. UV Curing 9
6.2.3. Electron-Beam Curing 9

1. Introduction

Being a natural raw material, wood is anisotropic and inhomogeneous; it has a porous cell structure, which can be swelled by water (→ Wood, Chap. 1.4.). Chemically, wood consists of a framework of cellulose, together with lignin (which imparts rigidity to the cell walls and acts as a binder) and hemicelluloses. Dyes, resins, fats, and water are contained in varying quantities (→ Wood, Chap. 1.3.1.). Wood has been used for hundreds of years as a building material and for items of daily use. It is unique in its color and surface structure.

Wood and wood products are sensitive to many external influences. These include animal pests, fungi, discoloration by sunlight, and changes effected by water, atmospheric moisture, and many other substances in daily use. Changes in environmental conditions bring about dimensional changes in wooden objects.

Wood and wooden objects must be protected against these influences by impregnating or coating if their usefulness is to be maintained. Because of the anisotropic nature of wood, surface coatings must have specific properties. The requirements of coatings for wood and wood products for outdoor use differ fundamentally from those for indoor use.

Coatings for Outdoor Use. Outdoors, solid wood is used primarily and generally coated with materials having insulating and impregnating properties. Here, no actual film is formed on the surface. The coatings also contain substances that repel pests (see → Wood, Preservation).

Coatings for Indoor Use. The processing conditions and use of paints and coatings for interior decoration and furniture are completely different. Indoors, solid wood and wood products of the most widely differing varieties are used, e.g., particle board (sheeted, veneered), medium-density fiber (MDF), and other fiber boards, and plywood (→ Wood, Chap. 2.2., → Wood, Chap. 2.3.).

In processing wood for the production of furniture and interior decoration, the original color and surface structure (pore structure), which are characteristic of the type of wood used, are included in the design of the surface. In surface treatment, the coating is either open or closed cell, but some overlap can be found between the two types.

Before surface treatment can be carried out, wood and wood products must be subjected to many preliminary processes, such as shaping (particularly for solid wood articles) and grinding with reducing grades of abrasives. Depending on the use envisaged, furniture must also fulfill certain criteria. For example, kitchen equipment (e.g., work surfaces) requires a higher level of resistance than other furniture. The different requirements determine the choice of coating system.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_385

Coating systems for furniture and interior decoration are compiled in Table 1; common solvents for paints are listed below:

Alcohols
Ethanol [*64-17-5*]
Isopropanol [*67-63-0*]
Butanol [*71-36-3*]
Isomeric butanols

Esters
Ethyl acetate [*141-78-6*]
Isopropyl acetate [*108-21-4*]
Butyl acetate [*123-86-4*]
Isomeric butyl acetates

Aromatic hydrocarbons
Toluene [*108-88-3*]
Xylene [*1330-20-7*]

Aliphatic hydrocarbons
Naphtha fractions with various boiling ranges

Ketones
Acetone [*67-64-1*]
Methyl ethyl ketone [*78-93-3*]

Glycols and derivatives
1-Methoxy-2-propanol [*107-98-2*]
1-Methoxyprop-2-yl acetate [*108-65-6*]
Butyl glycol [*111-76-2*]

Special solvents
Diacetone alcohol [*123-42-2*]
N-Methylpyrrolidone [*872-50-4*]

Reactive solvents
Styrene [*100-42-5*]
Methyl methacrylate [*80-62-6*]
1,6-Hexanediol diacrylate [*13048-33-4*]

The following coating systems exist:

1) *Physically Drying Systems.* The solvent (organic solvent or water) evaporates and a solid polymer film remains on the wood surface. Typical physically drying systems are nitrocellulose (cellulose nitrate) lacquers, stains, and waterbased paints.
2) *Physicochemically Drying Curing Systems.* Here, chemical reactions occur in addition to evaporation of the organic solvent. These reactions impart improved surface properties to the coating film. The chemical reaction is either a polycondensation (with elimination of water and formaldehyde in acid-hardening coating systems) or a polyaddition reaction (e.g., in the case of polyurethane coatings).
3) *Chemically Curing Coating Systems (Polymerizing Systems).* Here, the solvent (a reactive or polyfunctional solvent such as styrene or an acrylic monomer) is copolymerized with the curable system (typically unsaturated polyester resins) to a great extent.

In addition to the type of wood or wood product and the resistance required for the expected use, because of the relatively high proportion of organic solvents contained in the coating system, the corresponding environmental regulations (national, European, or international) must also be taken into account when selecting a paint system. The first priority is the maintenance of air purity (i.e., reducing the quantity of organic solvents) regulated by national laws. Waste avoidance and disposal and storage conditions (as stipulated, e.g., by the German water management law) are also important.

Besides the nature of the coating system and the proportion of organic compounds contained, the application process and its efficiency also determine the quantity of organic solvent emitted. The choice of application system depends, however, on the shape of the workpiece to be painted. The limiting values for residual emissions from the dried or cured coating are being constantly reduced to protect the end user.

Processes for surface treatment of wood products and furniture include preparation (shaping, grinding, etc.), bleaching, bleach-staining, staining, primary coating, intermediate grinding, and application of the top coating.

2. Bleaching

The *dyes* present in wood often have low resistance to sunlight and the wood can become discolored or altered. Three types of change occur. The wood becomes lighter (e.g., jacaranda); it becomes darker (e.g., oak, macoree, Canadian cherry); or its color changes (e.g., European cherry).

In the treatment process, unstable dyes are decomposed by bleaching with oxidants (generally hydrogen peroxide with appropriate additional components) and the color shade of the

wood is then reproduced by using synthetic organic dyes. This bleaching process is carried out primarily on light-colored woods (e.g., maple, birch), various types of cherry, and American walnut. Either the whole or part of the surface is bleached. After applying the oxidizing agents, the wood must be dried at room temperature (20 – 30 °C) for 24 h. In industry the drying time is reduced to 5 – 15 min at ca. 100 – 120 °C. The residual peroxide retained after drying can give rise to discoloration in the case of polyurethane curing components (aromatic polyisocyanates). For certain workpieces, bleaching and staining are carried out simultaneously.

Table 1. Coating systems for furniture and interior decoration

Type of coating	Solvent content, %	Water content, %
Water stains		90 – 95 %
Solvent stains	95	
	70	25 – 30
	30	60 – 70
Nitrocellulose lacquer, colorless	75	
Nitrocellulose lacquer, pigmented	60	
Acid-curing coating, 2-component, colorless	56 – 60	
Acid-curing coating, 1-component, colorless	70 – 75	
PUR [a] coating, colorless	70 – 75	
PUR [a] coating, pigmented	35 – 60	
PUR [a] coating, medium solid, colorless	40 – 50	
UP [b] coating, conventional	35 (15 [c])	
UP [b] coating, UV curing	35 (15 [c])	
UP [c] roller coating, UV curing	0 – 2	
UP [b] paste coating, UV curing	0	
UP [b] spray coating, colorless	65 – 70	
UP [b] spray coating, UV curing	65 – 70	
Waterbased paint, conventionally drying	5 – 7	60 – 65
Waterbased paint, UV curing	2	58 – 60
Waterbased paint, PUR [a] 2 C [d]	9	60

[a] PUR = polyurethane;
[b] UP = unsaturated polyester;
[c] 15 % reactive solvent styrene (emitted);
[d] 2 C = two component.

3. Staining

Staining can be carried out over the whole or part of the wood's surface (to compensate for color defects in the wood or veneer). Staining changes the color of the surface but not the wood structure. The shade always becomes darker compared with bleaching, and the natural grain is enhanced.

The shade of stained wood consists of two main components: the basic color shade of the wood and that of the stain. With certain types of wood – for example, those containing tannins – additional side reactions (e.g., with tannic acid) can occur.

3.1. Stain Systems

Three stain systems exist: water stains (main solvent water), solvent stains (ca. 30 – 95 % organic solvent), and chemical stains (which are generally water thinnable).

Stains always consist of solutions of organic dyes in solvents. Micronized pigments can also be incorporated to achieve better covering of the wood surface. Reactive dyes are used in chemical stains. These dyes react with the wood components such as tannins or resin acids in coniferous woods.

Water Stains. The classical stain system is the water stain, i.e., an aqueous solution of highly lightfast organic dyes. Alkali is sometimes added to achieve better wetting of the wood pores.

Water staining is easy to carry out. It gives very attractive, even surfaces. However, water stains have certain disadvantages:

1) The wood is severely roughened by the water; i.e., wood fibers that have not been completely cut off during grinding protrude again.
2) A relatively long drying period is required.

Roughening of the wood can be avoided by first wetting, drying, and then regrinding it.

Solvent Stains. To speed up processes both in industry and in craft workshops and to avoid roughening, solvent stains were developed. Here, an organic solvent is used instead of water. The dyes, which are highly lightfast, hardly change during normal use. A variety of solvent stains exist, and many have been adapted for use in industrial plants. However, solvent stains contain up to 95 % organic solvent, all of which is emitted.

With solvent stains, equalizing effects are achieved, depending on the additives. This is

particularly important in the case of cross-grain veneers and the cross-cut edges of solid wood parts.

Chemical Stains. Chemical stains produce the color shade as a result of a reaction between stain components. Previously, only *two-component stains* were in use, but *one-component stains* developed in the 1980s are now on the market. The latter react with the wood components themselves. Coniferous wood (softwood) stains are the most well known. Here, the resin acids react with the stain dye. If normal water stains are used, a *negative stain* is formed on the coniferous wood; i.e., earlywood is rendered significantly darker by staining than latewood. This is associated with the suctional capacity of the soft earlywood. The natural appearance of coniferous wood is, however, the reverse; i.e., earlywood is lighter than latewood. To maintain the natural appearance of coniferous wood by staining, *positive stains* are used. These dye latewood significantly darker and earlywood to a much lesser degree. However, in the case of coniferous wood exact periods of application (usually 16 h) must be adhered to. Also, the surface must be brushed. Positive stains can be used with other types of wood, as well (e.g., beech, oak, and cherry). They give very even coloration of the wood surface, without producing the positive effect.

For certain pieces of furniture in which the staining of certain parts of the furniture is to be particularly pronounced (e.g., antiques), *antique or wiping stains* are used.

3.2. Application of Stains

In *craftwork*, stains are applied by using a brush or a sponge and then spread evenly over the entire surface. In *industry*, stains are applied evenly by smooth or cellular rubber rollers. This allows significant reduction in the quantity of stain applied. Distribution over the surface is effected by rapidly moving brushes that operate at an angle over the object.

Combination Staining. To fulfill environmental requirements, in the late 1980s a new type of staining has been developed – combination staining – which is intermediate between water and pure solvent staining. It involves the use of a mixture of organic dyes, water, and organic solvents that can be diluted with water

4. Coating Systems

4.1. Cellulose Nitrate Lacquers (→ Paints and Coatings, Chap. 2.2.1.)

Cellulose nitrate (CN) lacquers [nitrocellulose (NC) lacquers, nitrolacquers] belong to the large group of physically drying coating systems. They have been used in the furniture industry for decades. The lacquers consist of a solution of cellulose nitrate and appropriate plasticizers (phthalates or various resins) in an organic solvent. Nitrocellulose lacquers contain ca. 70 % organic solvent. The solid material content is limited by the cellulose nitrate because the viscosity of the lacquer increases more than proportionally with increasing cellulose nitrate content.

The application of nitrocellulose lacquers is straightforward. In industry, practically any drying time is possible. The lacquers can be applied by spraying, with rollers, and by pouring and dipping.

Many different types of CN lacquers exist. Particular surface effects – from high gloss to deep matt – can be produced easily on furniture.

Cellulose nitrate lacquers intensify the color of the wood considerably and strongly enhance its structure. *Brightener* (*natural wood effect*) *lacquers* can be produced by mixing other cellulose esters (e.g., cellulose acetate butyrate) with cellulose nitrate. Since cellulose esters are physically drying coatings, swelling and dissolving effects as a result of the subsequent action of organic solvents (e.g., esters and ketones) are unavoidable. The resistance toward organic solvents, some of them in daily use, is therefore limited. Cellulose nitrate lacquers also have a limited lightfastness. Discoloration or yellowing of cellulose nitrate lacquers occurs mainly with very light woods such as maple, birch, or ash. However, it can be improved by various photoprotective agents.

4.2. Polyurethane Coatings

Polyurethane (PUR) coatings are systems that are converted into a usable form by the addition

of hardeners (Paints and Coatings, Chap. 2.9.). Unlike urea resins and the acid-curing coatings produced from them, curing occurs here by polyaddition; i.e., no cleavage products, such as formaldehyde, are liberated during curing, not even at a later stage. The hardener and paint must be mixed in a precise ratio, which is determined by the stoichiometry of the chemical reaction. This ratio generally involves whole numbers (e.g., 5 : 1 or 10 : 1).

Polyurethane coatings consist of two components. The first is a hydroxyl-group-containing resin (or resin components) together with appropriate plasticizers, flow-control agents, and accelerators. The second reaction component (hardener) is a polyisocyanate. The building blocks of the polyisocyanates are toluene diisocyanate (TDI) or hexamethylene diisocyanate (HDI). These monomers are unsuitable for use in industry because their vapor pressures are too high. For safe handling, higher molecular mass products with lower vapor pressures are produced. These prepolymers still contain small quantities of monomers (>0.2 %). The *hydroxyl components* are polyester polyols or hydroxyl-group-containing acrylic resins. They cross-link with aliphatic isocyanates to form rapidly drying products with high lightfastness. During formulation, no water or hydroxyl-group-containing solvents (e.g., alcohols) must be present; otherwise, side reactions can occur that impair the resistance of the coating. The reduced amount of paint required and the fact that the paint system is more concentrated compared to nitrocellulose lacquers lead to lower solvent use.

At present, *medium-solid polyurethane coatings* with excellent surface properties can be produced, even for wood.

Uses. Polyurethane coatings are extremely versatile. Colorless and pigmented coating systems have been developed in all matt grades. Natural wood effects can be achieved easily. Even highly elastic coating films can be produced so that plastics (including soft plastics), which are processed together with wood (e.g., in kitchen furniture), can be coated. Polyurethane coating films have an exceptionally high resistance toward usual household chemicals as defined in DIN 68860-68861 and the corresponding EN and ISO standards. They are also unaffected by the components of wood and are often the only means of achieving a satisfactory coating when, for example, wood components interfere with other curing reactions.

Polyurethane coatings with low flammability, according to the requirements of DIN 4102, group B 1 also exist; these are particularly important for public buildings and interior furnishings for hotels, restaurants, and conference centers.

The cured coating films meet the German legal requirements for foods. The films are resistant to perspiration and saliva according to DIN 53 160. A particular feature of polyurethane coating systems is their strongly pronounced insulating and impregnating effect, which is important for precoating MDF boards.

Applying two consecutive coating films seals the wood surface so that in the case of veneered particle board, possible residual emissions of formaldehyde are reduced. These emissions can originate from the particle board itself or can be due to other processes, such as veneering with urea resins, even when E 1 particle board (→ Wood, Chap. 2.5.2.2.) is used. The maximum permissible concentration of formaldehyde emitted from wood products according to GefStoffV is 0.1 ppm. If these types of surface are coated by double application of a polyurethane paint, the formaldehyde concentration, which can increase to >0.1 ppm during veneering with urea resin-containing glues, is reduced to a quarter of the original value. This can, however, be achieved only by using E 1 board and applying special treatment to drilled holes and narrow edges.

4.3. Unsaturated Polyester Coatings

(→ Paints and Coatings, Chap. 2.8.)

Unsaturated polyester (UP) coatings are reactive systems that are purely chemically curing. Cleavage products, such as formaldehyde, are not liberated. The solvents used are polyfunctional reactive solvents; i.e., they serve to adjust processing viscosity and are partially copolymerized with the unsaturated polyester film during polymerization. Compared with other solvent-containing systems, the quantity of solvent contained is not the same as that emitted.

Classical *paraffin-containing polyester* systems, which dry conventionally, are intended for

high-gloss coatings. They are still widely used, e.g., on decorative parts for cars and small objects (e.g., clock cases) with particularly high-gloss effects. Ready-made UP systems are available in the most widely differing colors and matt grades, and are important for industrial surface treatment. Owing to the development of MDF boards that require special priming, pigmented polyester filler systems have been introduced.

Matt polyester coatings for open-pored effects are also still in use. They do contain not only styrene as the solvent, but 60 – 70 % normal organic solvents, as well. To initiate the chemical reaction in any polyester coating system the following are required: an accelerator, normally an organometallic compound (cobalt octanoate), and an organic peroxide solution as the curing agent.

Once mixed, the coating can be applied for only a limited period. Therefore, in industry, only two-component (two-pack) coating systems are generally employed. For flat surfaces (e.g., in the industrial coating of doors and parts for bedroom furniture), hardeners are applied separately either by pouring or with rollers.

Radiation Curing. The chemical reaction can also be initiated *by IR or UV radiation*. Unsaturated polyester resins or special acrylate-modified UP resins can be cured by UV radiation. When using UV-hardening systems, a photoinitiator must always be added. Accelerators based on organometallic compounds and organic peroxides need not be used. The reactivity of the coating system and the application equipment determine to a large extent the number and electrical power of UV lamps required.

Since the hardening components are separated from the actual paint system, the potlife of the system is prolonged and the amounts of special waste are reduced because cured coating residues are not produced.

Acrylate-modified polyester resins are significantly more reactive than unsaturated polyester resins and allow higher rates of feed, fewer UV lamps, and thus less energy input.

Developments of paint raw materials have reached a high level so that a large number of coating systems applied with rollers and spatulas are virtually solvent free. Novel *UV-curing water-thinnable* materials are also in use. They are not comparable to conventional water-thinnable systems. The new systems are generally emulsions with a maximum organic solvent content of 2 – 2.5 %; the main solvent is water. In processing these systems, milling, grinding, and priming must be performed especially carefully. Since water is the main solvent, longer drying times are necessary, generally 5 – 10 min on a flat bed dryer, depending on the quantity. If shaped articles are to be coated, depending on the shape, evaporation times of 15 – 25 min are necessary with a stepwise rise in temperature. If the water is not evaporated entirely, visible white stripes appear immediately after UV curing of the paint, which cannot be removed. Because of the increased use of UV-hardening polyester systems, completely new technologies have been developed in many areas (e.g., the parquet and door industries). An example of a production line for UV-curing coating of smooth, flat parts is given in Figure 1, the quantities of coating applied are listed in Table 2. This technique satisfies environmental requirements (protection of raw material resources) because of the small quantities of coating required. These are usually 10 – 20 g in priming and 10 – 20 g in the top coating.

Table 2. Example of and quantities applied in a production line for roller coating of smooth flat workpieces using a UV curing paint ∗

Type of process	Applied quantity of coating, g/m^2	Quantity of solvent, g/m^2
Priming (water-thinnable, sometimes colored, solvent 2 %)	20	1
Spray drying		
Staining (combination stain, solvent 30 %)	20	6
Spray drying UV roller priming (solvent 2 % max.)	15	1
UV curing zone		
Grinding		
Intermediate staining (only when complete UV curing is used, solvent 90 %)	8	7.2
Spray drying		
UP upper coating (solvent ca. 5 %)	10	0.5
Total quantity applied	93	16.7

∗ These coatings can be produced as complete UV curing coatings. Alternatively, the coating can be applied by pouring or spraying using other systems (e.g., NC or PUR coatings), after priming and grinding.

Figure 1. Production line for UV coating of smooth, flat parts
a) Roller; b) Dryer; c) Smoothing machine; d) UV precuring zone; e) Smooth roller; f) Special system of rollers; g) UV curing zone; h) Grinding machine; i) Spraying machine, casting machine; j) Roller for intermediate stains; k) Roller for top coating

4.4. Waterbased Paints (→ Paints and Coatings, Chap. 3.3.)

Waterbased paints were developed in the 1950s to replace organic paint solvents with water, which is noncombustible and nontoxic. Waterbased paints for the furniture industry are based on proven coating raw materials (acrylate dispersions).

These paint systems have three advantages: (1) the main solvent is water; (2) no environmentally polluting cleavage products are liberated during drying; and (3) the fastness of the coating is excellent.

For certain types of wood and certain color shades, photoprotective agents must be added, particularly for light wood (e.g., special types of maple) and for light stains.

For technical reasons, waterbased paints still contain ca. 5 – 7 % organic solvents. Their constitution is, however, completely different from that of solventborne products. In addition, waterbased paints based on polyurethane, polyurethane – acrylate, or UV-curing coating systems exist.

In *conventionally drying* waterbased paint systems, film formation occurs after evaporation of the water in such a way that the binder components bond to one another and become fixed, thus forming a largely water-insoluble film. For film formation, however, the temperature must not fall below a certain value – the minimum film-forming temperature (16 – 18 °C). Waterborne paint can be applied only at 18 °C, the materials to be coated must be at this temperature.

Wood is swelled by water, and protruding wood fibers cause roughening of the surface. The corresponding preliminary processes must be adapted to compensate for this. A normal water stain can be used only when suitable additives are added. The color intensification achievable with polyurethane and nitrocellulose coatings cannot be obtained with waterborne paints in the case of dark or very bright color shades. It can be achieved, however, with bright mahogany shades by coloring the waterbased paint. When the same coating quantity is applied, waterbased paints have a significantly higher filling power than other paints with virtually the same solid material content. For use in spraying, machines equipped with warming devices have proved particularly useful. Practically no odor is perceptible when the objects are used to furnish existing houses or workplaces. The same applies to parquet coatings. Like polyurethane coatings, waterbased paints are highly insulating, so that the formaldehyde emitted from veneered particle board is far below the equilibrium concentration. A prerequisite is, however, the use of E 1 board and special treatment of drilled holes and narrow edges.

A novel development is the production of *UV-curing water-thinnable* coatings (see Section 4.3). Here, particular precautions must be taken during predrying before actual UV curing.

More recent developments start with water-thinnable *polyurethane coating system*. Here, mixing with the polyisocyanate component, which can have varying viscosities compared to the waterborne paint, must be carried out carefully. Two-component metering devices are recommended, since the potlife is between 2 and 3 h at normal temperature.

5. Paint Application

Although the emission of organic solvents during hardening of a coating depends on the coating materials used (solvent content and quantity applied per square meter), the application process and its efficiency are extremely important. The type of application system used is determined by the production program and the shape and dimensions of the workpiece to be coated. It is also influenced by downstream costs involving, for example, waste production and disposal and, particularly, the overall cost per square meter of coating. Emissions and waste production

are controlled by national and international regulations and must be continuously adapted to the current state of the art. With highly efficient application systems [e.g., dipping, pouring, or rollers (efficiencies 95 – 98 %)], virtually no room exists for improvement.

In *spray systems*, losses due to overspraying are unavoidable and depend on the shape of the coated workpiece. These losses can be minimized by further development of spraying machines with recycling devices, e.g., recovery from the conveyor belt, which allows a significant reduction in the quantity of coating used and thus the cost per square meter at a relatively low investment cost. These recycling devices can be incorporated into existing spray systems. For the use of spraying machines to be reasonably economical, a certain number of workpieces must be coated both simultaneously and subsequently. The amount applied by a spraying machine can be reduced through the use of several spray guns.

6. Paint Curing Methods

After the coating material is applied, the paint must be transformed from the liquid to the solid state as quickly as possible in the absence of dust. Drying must be environmentally friendly, energy saving, and low cost. In industry, paints cannot be dried at room temperature because this is time-consuming and requires too much space. All drying and hardening plants utilize the physical principle that the vapor pressure of a solvent increases with temperature and drying time decreases. At the same time, chemical reactions are accelerated, and curing times thus shortened. The solvent must be removed by suitable means, generally by passage of air, the quantity of which varies with the quantity of solvent. The upper and lower explosion limits of the solvent – air mixtures must be taken into account.

6.1. Heat-Drying Systems

Drying is performed with circulating heat carriers, usually a hot-air stream. Heat-drying systems include:

1) Multi-shelved dryers, aerated left and crosswise
2) Flat track, aerated by a countercurrent
3) Nozzle drying channels with special nozzle systems and high air velocities
4) Tower dryers with different temperature zones, which can be regulated individually

All CN lacquers, PUR coatings, solvent stains, water stains, conventional waterbased paints, and UP coatings can be dried or cured with these heat-drying systems. The construction and mechanical components of these types of drying systems are rather simple. However, a relatively high quantity of heat is required because the air must be heated and then passed over the coated workpiece. This air, together with the evaporated solvents, is generally emitted directly into the atmosphere, particularly in the case of old dryers. Ca. 95 % of the heat energy supplied is thus lost. Simple plants consist of an application device (manually operated spray zone, automatic spraying machine, or pouring machine, a predrying zone, and an evaporation zone). After application, the coated parts are placed in a tray truck and predried in front of an evaporation wall. Part of the solvent is thus emitted into the workplace. The predrying zone is generally followed by temperature-regulated evaporation zones where the temperature rises continuously from 30 to 50 °C.

Drying and hardening by hot air (convection drying) has a relatively poor heat balance. The hot air naturally heats the coated workpiece. A large proportion of the heat energy is used to heat the drying chamber and drying devices, at least in the initial phase. No UV-hardening materials or materials that need to be hardened by electron beams can be cured in convection-drying plants. Steam, hot water, and – at higher temperature – oil are used as the heat medium in these plants.

6.2. Radiation-Curing Systems

The radiation-curing systems IR, UV, and electron-beam curing (EBC) are much more effective. The curing reaction can be accelerated considerably by energy-rich radiation. However, not every type of radiation is suitable for every type of binder. In all radiation-curing systems, energy is not transferred by the ambient air; rather, it interacts directly with the coating

layer, resulting in much more rapid energy transfer (with convection drying the power transfer is between 2 and 5 kW/m^2; with radiation, between 10 and 50 kW/m^2). No startup times, as with boiler houses or heat production devices, are required.

6.2.1. IR Irradiation

Irradiation with IR provides an effective method for accelerated curing. Short- (1.5 – 1.6 μm), medium- (2.5 – 3.0 μm), and long-wavelength (3.5 – 6.0 μm) radiation can be used.

Basically, three physical processes can occur during IR drying: adsorption, reflection, and transmission of radiation. The most efficient process is adsorption, which depends on the thickness of the coating layer and its components. Here, the radiation is converted directly into heat in the coating layer. Medium-wavelength radiation has proved most suitable for wood coating. Generally, twin radiation scources with vapor-deposited reflectors in a quartz casing are used. The radiation sources can be readily built in existing plants. The sources are robust and do not contain any sensitive mechanical components. The physicochemically drying and chemically hardening systems are suited to curing by IR radiation.

6.2.2. UV Curing

UV curing requires special coating systems (see Section 4.3; → Paints and Coatings, Chap. 3.7.). The radiation sources are basically high-pressure mercury lamps with a power output of 80 – 100 W per centimeter arc length. These lamps emit their energy in the form of impulses at a frequency of 100 Hz. Some lamps have been developed more recently with a much higher pulse rate (>100 Hz).

During UV curing, ozone is always produced underneath the radiation source. It is generally removed by means of ventilators that cool the source. UV radiation is generally used for curing stain and paste materials involving colorless and pigmented systems. For pigmented systems the radiation source must be doped accordingly, because not every pigment or color shade can be hardened in the same way and at the same time. The use of UV-curing systems to coat flat workpieces is relatively simple. However, by means of special plants and special arrangement of the radiation source, solid objects and chair frames can now also be irradiated.

6.2.3. Electron-Beam Curing

Unlike UV or IR radiation, which consist of electromagnetic waves, electron beams are not reflected by pigments. Highly accelerated electrons penetrate materials that are impermeable to light, so the system is equally suitable for clear varnishes and pigmented coatings. Electron-beam curing has the shortest drying times. The cost of the plant is relatively high, but energy costs are very low. Special coatings based on unsaturated acrylates must be used. Curing must be carried out in the absence of atmospheric oxygen, i.e., in an inert gas atmosphere (nitrogen). The secondary radiation (X-rays) produced during electron-beam curing requires complex radiation shielding.

Wool

HELMUT ZAHN, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Federal Republic of Germany (Chaps. 1, 2.1, 2.2, 6 – 11, 13)

FRANZ-JOSEF WORTMANN, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Federal Republic of Germany (Chaps. 2.3, 3.1)

GABRIELE WORTMANN, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Federal Republic of Germany (Chap. 8, 10)

KAROLA SCHÄFER, Deutsches Wollforschungsinstitut an der RWTH Aachen, Aachen, Federal Republic of Germany (Chaps. 3.2, 5.4, 5.5, 10, 13)

RAINER HOFFMANN, Hochschule Bremen, Bremen, Federal Republic of Germany (Chaps. 4, 5.1 – 5.4, 12)

ROBERT FINCH, Woolmark Europe, Inc., Ilkley, United Kingdom (Chaps. 11, 12)

1. Introduction 2
1.1. Definition 2
1.2. Historical Aspects 2
1.3. Biology 3
1.4. Chemical Composition 3
2. Structure 6
2.1. Morphology 6
2.2. α- and β-Keratin 7
2.3. Two-Phase Model 9
3. Properties 10
3.1. Physical and Mechanical Properties . 10
3.2. Chemical Reactivity 13
4. Resources and Raw Materials 15
4.1. Sheep Rearing and Breeding 15
4.2. Places of Origin and Types of Raw Wool 16
5. Production Processes 16
5.1. Methods of Obtaining Wool 16
5.2. Raw Wool Scouring 16
5.3. Mechanical Processing 17
5.4. Chemical Processing: Pretreatment . 17
5.5. Dyeing and Finishing 19
6. Environmental Protection 21
7. Quality Specifications 21
8. Testing and Analysis 22
9. Storage and Transportation 23
10. Uses 23
11. Trademarks 23
12. Economic Aspects 24
13. Toxicology and Occupational Health . 26
14. References 27

1. Introduction

Today, wool accounts for only ca. 3 wt % of world textile-fiber production. However, if the value added by dyeing and printing is included, the share in monetary terms is considerably higher. The importance of wool is due to its unique physiological properties as a clothing material and the wide range of techniques for converting it into textiles, including felts and finishing processes.

Some important methods and discoveries in modern biological sciences had their origins in wool research (partition and paper chromatography, the structure of α- and β-keratin, intermediate filaments).

The production of wool and woolen textiles and the trade in these commodities have played an important role in political and economic history.

1.1. Definition

The natural fiber wool, like other types of fine and coarse animal hair and also silks (→ Silk), is an animal fiber [54], and although the term "wool" can be used in conjunction with the names of various animals, e.g., "angora wool", it is here understood to include only the hair of the various breeds of domesticated sheep (*Ovis aries*).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_395

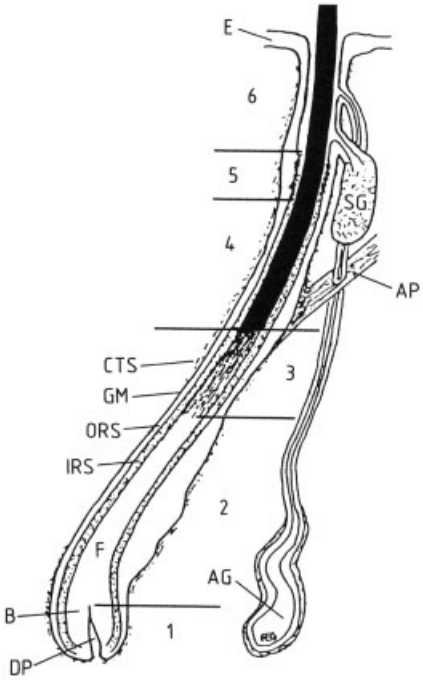

Figure 1. Diagram of fully developed wool follicle (not to scale) [58]
The six zones are: 1) Bulb; 2) Keratogenous zone; 3) Zone of final hardening of the fiber; 4) Zone of inner root sheath degradation; 5) Zone of sloughing of inner and outer root sheath cells; 6) Pilary canal.
AG = aprocine gland; AP = arrector pili; B = bulb; CTS = connective tissue sheath; DP = dermal papilla; E = epidermis; GM = glassy membrane; IRS = inner root sheath; ORS = outer root sheath; SG = sebaceous gland

1.2. Historical Aspects [55–57]

The wool of the sheep was the first textile raw material used by humans for clothing purposes. The development of clothing began when people first dressed themselves in the skins of mammals. Felts, which are more easily shaped, were produced later from hairs plucked or cut from the animal (→ Felts). The development of spinning and weaving to produce fabrics was a still later development.

The first country to process and trade in wool was Babylon (Babylonia = land of wool). The oldest known wool fabrics, dating from the second half of the second millennium B.C., were found in Danish tree coffins.

Sheep rearing reached a high point in the Middle Ages and the Renaissance. The merino sheep, bred in Spain, produced wool of very fine quality, and in the late 1700s Saxon merinos were exported via England and South Africa to Australia, where the climate proved to be very favorable for rearing this breed.

1.3. Biology

Wool fibers and other animal hairs, including human hair, are formed in follicles – sheath-like formations of the epidermis, at the base of which are papillae of mesodermal cells (Fig. 1).

Follicles are dynamic organs in which various processes occur simultaneously: cell division, cell differentiation, upward movement of the keratinocytes, biosynthesis of wool proteins as gene products, self-assembly of the proteins to form supramolecular structures, and finally keratinization (cornification) by formation of cross-linked disulfide and isodipeptide bridges. At the same time, cells of the inner and outer root sheath are catabolized [58].

The various stages of development of the growing wool fiber in the follicle can be divided into six zones (Fig. 1). In the upper part of Zone 1, fine keratin intermediate filaments (KIF) and trichohyalin granules are already forming. Between Zones 1 and 2, filaments aggregate to form macrofibrils, in which keratin-associated proteins (KAP) are deposited. In Zone 2, amorphous protein aggregates, ca. 30 nm in diameter, form in the fiber cuticle cells. By the end of Zone 2 the cuticle cells are hardened, while the final keratinization of the entire wool fiber occurs in Zone 3.

1.4. Chemical Composition [27, 31]

The protein fiber wool consists of carbon, hydrogen, oxygen, nitrogen, and sulfur. The elemental analysis of wool (free of water) is as follows:

Carbon	50.5 wt %
Hydrogen	6.8 wt %
Oxygen	22.0 wt %
Nitrogen	16.5 wt %
Sulfur	3.7 wt %
Ash	0.5 wt %

Except for the sulfur content, this composition is typical for all proteins. The high sulfur

content is due to the high content of cystine, a double amino acid containing two sulfur atoms in a disulfide bond:

$$HOOCCH(NH_2)CH_2S-SCH_2CH(NH_2)COOH$$

The ash contains potassium, sodium, calcium, aluminum, iron, silicon, sulfate, carbonate, phosphorus pentoxide, and chloride. Water-free wool consists mainly (ca. 97 %) of wool proteins, the remainder being made up of ca. 2 % structural lipids (Fig. 2) [59–62], ca. 1 % mineral salts, nucleic acids [63, 64], and carbohydrates.

Industrially scoured wool has only small residual amounts of wool grease (ca. 0.5 %). The structural lipids, in contrast, are components of the cell membrane complex which binds together the cuticle and cortex cells.

A characteristic of the structural lipids of the cuticle is the presence of a covalently bound branched-chain C_{21} fatty acid 18-methyleicosanoic acid [65–67].

Total hydrolysis of the peptide bonds in wool proteins yields 24 amino acids. The data given in Table 1 were obtained by combining data from acidic and from enzymatic hydrolyses of wool.

The amino acids are classified in five groups: "acidic" amino acids, "basic" amino acids, amino acids with hydroxyl groups, sulfur-containing amino acids, and amino acids with no reactive groups in the side chain. The total amount of amino acids with reactive side chain groups is 5395 µmol/g, and of amino acids without reactive side chain groups 3450 µmol/g. The sum of these two figures, after subtraction of the concentration of *C*-terminal amino acids (10 µmol/g), gives a figure of 8835 µmol/g for the concentration of peptide groups in the peptide chains of wool.

Like all proteins, wool contains both cationic and anionic groups, and is therefore amphoteric. The cationic character is due to the protonated side chains of arginine, lysine, and histidine, and to the small number of free amino groups at the ends of the peptide chains. The amino groups of lysine, histidine, the amino end groups, and the thiol groups of cysteine are important sites for the covalent attachment of reactive dyes [68]. Anionic groups are present as dissociated side chains of aspartic and glutamic acids and as carboxyl end groups.

The side chains, which in wool account for a considerable proportion (50 %) of the protein material, interact with each other, thereby stabilizing the peptide by forming links between the chains and rings within a chain:

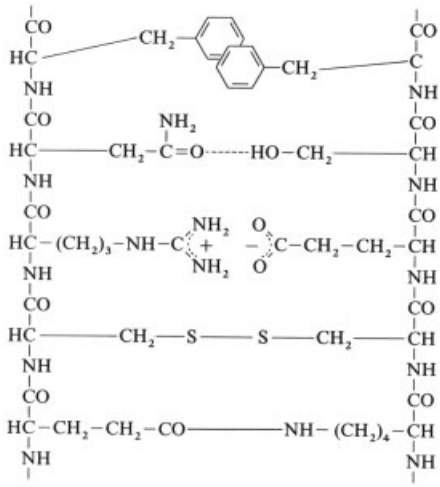

This structural formula is a schematic representation of five such links between segments of two hypothetical peptide chains. From the top downwards it shows interactions between phenyl rings (for the role of aromatic rings as hydrogen bond acceptors see [69]), hydrogen bonds between an asparagine residue and a serine residue, a salt bridge between an arginine residue and a glutamic acid residue [70], a disulfide bridge between two cysteine residues, and an isodipeptide bridge between glutamic acid and lysine [71, 72].

The disulfide bridge plays an important part in stabilizing the wool fiber, leading in particular to its relatively high wet strength, moderate swelling, and insolubility. A second covalent bridge is provided by the isodipeptide N_ε-(γ-glutamyl)lysine, which provides an additional stabilizing effect in the cornified cell envelopes of the cortex and cuticle.

The so-called salt bridge is due to the electrostatic interaction between cationic and anionic side-chain groups. Hydrogen bonds are formed in proteins between

$>NH$ and $-\overset{|}{\underset{|}{C}}-OH$

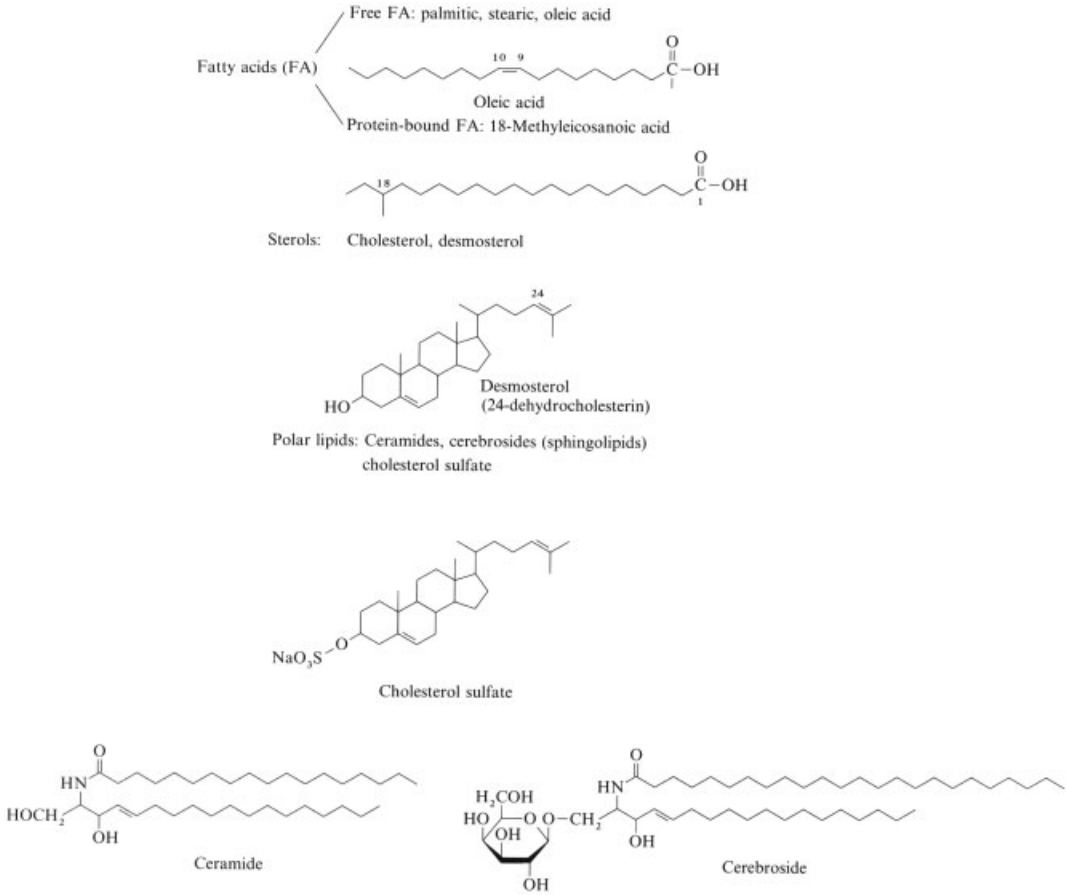

Figure 2. Structural lipids

as donors and

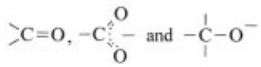

as acceptors, especially in the helical rod domains of the keratin filaments.

Hydrophobic effects stabilize the aggregation of nonpolar side chains, so reducing the area of the interface with water.

The three-dimensional structure of the wool proteins is stabilized both by the hydrophobic effect and by a wide variety of electrostatic interactions between amino acid constituents [73].

Eight *N*-terminal amino acids have been analyzed by means of 2,4-dinitrofluorobenzene: cystine, glycine, threonine, valine, alanine, serine, and glutamic and aspartic acids. Other *N*-terminal amino acids are present in *N*-acetylated form. Thus, the acid hydrolysis of wool liberates 50 µmol/g acetic acid [74].

S-Carboxymethylated soluble wool proteins can be obtained by exhaustive mercaptolysis and carboxymethylation of the resulting SH groups [75–77].

The dissolved *S*-carboxymethylated wool proteins can be separated into the main groups listed in Table 2, and these fractions can be separated into definite proteins, some of which have been sequenced [78]. In two-dimensional polyacrylamide gel electrophoresis [79], the proteins are first separated according to charge, and, in the second dimension, in the presence of sodium dodecylsulfate (SDS), according to molecular mass (Fig. 3). The heterogeneity of this class of proteins can thus be demonstrated by a relatively simple analytical process. The keratins are well separated from the sulfur-rich proteins be-

Table 1. Amino acid composition of fine merino wool ∗ [49]

Group	Name	Side chain	Concentration, μmol/g
"Acidic" amino acids and their ω-amides	aspartic acid	$-CH_2-COOH$	200
	glutamic acid	$-(CH_2)_2-COOH$	600
	asparagine	$-CH_2-CONH_2$	360
	glutamine	$-(CH_2)_2-CONH_2$	450
"Basic amino acids and tryptophan	arginine	$-(CH_2)_3-NHC(=NH)-NH_2$	600
	lysine	$-(CH_2)_4-NH_2$	250
	histidine	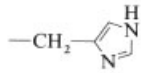	80
	tryptophan	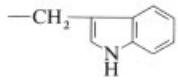	40
Amino acids with hydroxyl groups in the side chain	serine	$-CH_2-OH$	900
	threonine	$-CH(CH_3)-OH$	570
	tyrosine	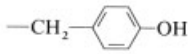	350
Sulfur-containing amino acids	cysteine	$-CH_2-SH$	10
	thiocysteine	$-CH_2-S-SH$	5
	cysteic acid	$-CH_2-SO_3H$	10
	cystine	$-CH_2-S-S-CH_2-$	460
	lanthionine	$-CH_2-S-CH_2-$	5
	methionine	$-(CH_2)_2-S-CH_3$	50
Amino acids without reactive groups in the side chain	glycine	$-H$	760
	alanine	$-CH_3$	470
	valine	$-CH(CH_3)_2$	490
	proline	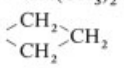	520
	leucine	$-CH_2-CH(CH_3)_2$	680
	isoleucine	$-CH(CH_3)-CH_2-CH_3$	270
	phenylalanine	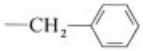	260

cause of their high molecular mass, and the high glycine – tyrosine proteins are nearly at the front of the SDS dimension because their molecular mass is very low. Some of the heterogeneity of the keratins is caused by phosphorylation [80].

The keratins in wool come from the microfibrils, which together with some other protein filaments are classified as intermediate filaments [81–84]. Intermediate filaments form an important filamentary system in the cytoplasm of higher eukaryotic cells. They have a diameter of 8 – 10 nm, and are thus larger than actin filaments and smaller than the microtubuli [41] of the cell.

2. Structure

2.1. Morphology [85–89]

Wool fibers consist of two cell types: cuticle and cortex cells. In coarse fibers, there is also a central medulla (Fig. 4).

The cuticle [87] consists of plate-shaped cells ("scales") that overlap longitudinally and

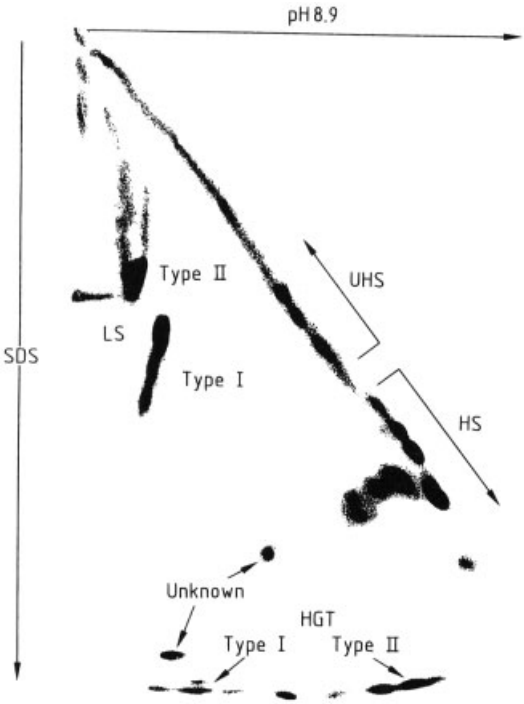

Figure 3. Two-dimensional polyacrylamide gel electrophoretic separation of wool proteins [79].
The proteins were extracted from merino wool and subjected to electrophoresis at pH 8.9 in one direction and in the presence of SDS in the other.
LS = low sulfur; HS = high sulfur; UHS = ultrahigh sulfur; HGT = high glycine – tyrosine

peripherally, with the 1 µm thick edges of the scales pointing in the direction of the tip of the fiber. Each cuticle cell consists of four layers with different cystine and isodipeptide contents: epicuticle, a-layer, exocuticle, and endocuticle. Between overlapping cuticle cells is the cell membrane complex. The epicuticle consists of proteins and lipids, including covalently bound 18-methyleicosanoic acid, (Fig. 2).

The cortex consists of spindle-shaped interdigitated cells [88]. These consist of ortho, para, and more rarely meso cells, all differing in their cystine content and staining behavior with silver salts. A cortex cell of wool contains around 5 – 8 macrofibrils with a diameter of 300 nm at their widest point. Between the macrofibrils are cytoplasmic and nuclear remnants of the keratinocytes. This intermacrofibrillar material swells in water more than the macrofibrils. A macrofibril is a bundle of 500 – 800 keratin intermediate filaments (microfibrils). Keratin-associated proteins (KAP) are intercalated, enveloping individual microfibrils and their aggregates. An individual microfibril consists of protofibrils, which in turn consist of protofilaments (Fig. 5) [89].

Table 2. Proteins of the wool fiber [78]

Type of protein	Number of proteins	Range of molecular mass, kDa
Keratins		
Acidic keratins (Type I)	4	40 – 50
Basic keratins (Type II)	4	56
Keratin associated proteins (KAP)		
High-sulfur proteins	80 – 100	11
		16
		19
		23
Ultrahigh-sulfur proteins		16
High glycine – tyrosine proteins (Type I)	10	6 – 9
High glycine – tyrosine proteins (Type II)	5	6 – 9

The medulla in coarser wool fibers consists of hollow cells with a skeleton of amorphous proteins and fine filaments.

The natural color of brown wool is caused by pigment granula of black to brown eumelanin and yellow to red pheomelanin.

2.2. α- and β-Keratin

The microfibrils are partly crystalline, and on X-ray structural analysis give reflections on the meridian (0.51 nm) and equator (0.98 nm), as well as many long-period reflections on the meridian and some on the equator. Astbury [90] designates the modification that results in this fiber diagram (Fig. 6 A) as α-keratin. Wool that has been stretched by ca. 50 % in water and then steamed gives a new fiber diagram with a meridian reflection at 0.34 nm and equatorial reflections at 0.465 nm and 0.98 nm. This X-ray diffraction pattern is due to the modification known as β-keratin (Fig. 6 B). A new interpretation for the $\alpha \rightarrow \beta$ transition is given by CAO [96] and Kreplak et al. [97].

In α-keratin [91], the peptide chains are twisted like a right-handed screw (α-helix, Fig. 7). The stability of the α-helix is due to intramolecular hydrogen bonds between the >C=O groups of the peptide bonds and adja-

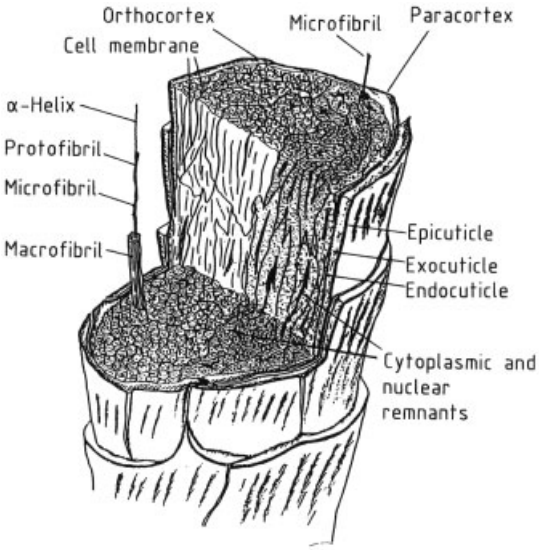

Figure 4. Longitudinal and cross section of a fine merino wool fiber [86]

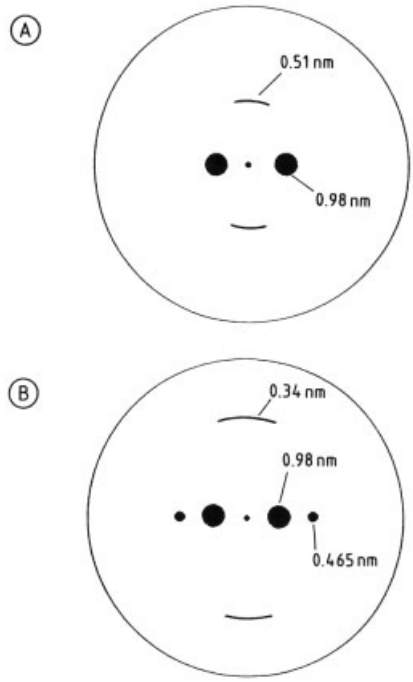

Figure 6. Diagnostic X-ray reflections [9]
A) α-Keratin pattern; B) β-Keratin pattern

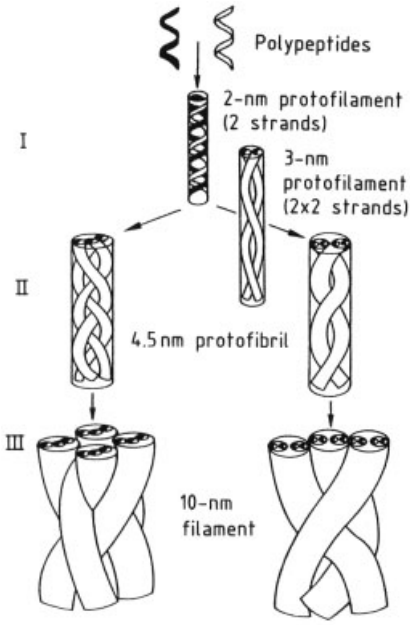

Figure 5. Schematic representation of the three distinct levels of fibrillar organization in keratin filament assembly and architecture [89].
I) The 2 nm protofilament consists of two coiled-coil keratin polypeptides, presumably one acidic (white) and one basic (black) species; II) The 4.5 nm protofibril may be a six-stranded subunit composed of three 2 nm protofilaments (shown on left), or, alternatively, may be an eight-stranded subunit formed by the lateral aggregation of two four-stranded 3 nm protofilament intermediates (shown on right); III) The 10 nm filament may comprise four (left) or three (right) protofibrils, depending upon protofibril substructure. The number of protofibrils (four or three) per 10 nm filament is not unique but just represents one possible polymorphic form.

cent >N−H groups of amino acid units located on the next turn of the α-helix and intrahelical salt linkages.

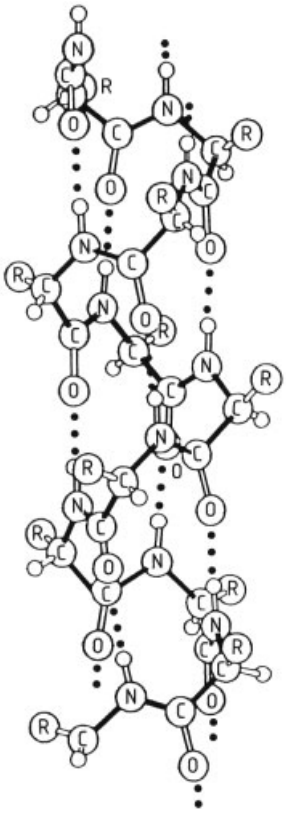

Figure 7. α-Helix structure of α-keratin [91]

The α-helix contains 18 amino acid units in five turns, i.e., 3.6 amino acid units per turn. To give the distance between successive turns of the helix that leads to the observed meridian reflection (0.51 nm), the helical chain must itself be slightly coiled (superhelix, coiled coil). Two superhelices combine to form a left-handed, two-stranded, rope-like assembly in which the superhelices are arranged such that the hydrophobic side chains on the outside of the helix interlink to form a stable "buttonhole" structure. These dimers are the actual physical structural elements of the intermediate filaments (microfibrils), and can be termed "molecular twins."

In the β-keratin structure [92], the peptide chains are considerably stretched, as in silk fibroin (→ Silk, Chap. 2.1.). The chains form so-called pleated sheet structures (Fig. 8). The sheet is formed by cross-linking of the chains by hydrogen bonds between peptide groups of opposing chains. The sheet is folded at the −CHR groups of the chains to such an extent that the meridian reflection of the β-keratin X-ray diffraction pattern at 0.34 nm corresponds with the geometry of the model. The phenomenon of the $\alpha \rightarrow \beta$ transformation has been applied in the new developed OPTIM technology (see Section 5.4).

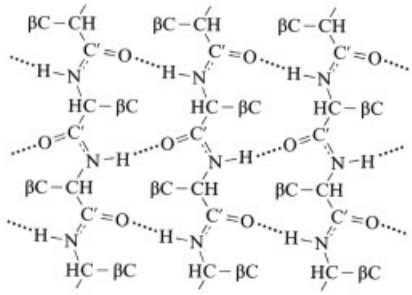

Figure 8. Pleated sheet structure of β-keratin [92]

2.3. Two-Phase Model

The complex morphological and molecular structure of wool (Fig. 4) echoes the construction principle of all biological composite structures and many man-made materials, combining components with different properties in one material so as to maximize suitability for its purpose. Table 3 shows the stepwise differentiation of the morphological structure of the wool fiber into the most important two-phase structures.

With regard to elastic properties, the essential difference between the components is in their degree of order. If only the α-helical central rod-like domains in the microfibrils (intermediate filaments) are regarded as microcrystalline, the crystalline phase accounts for 30 % of the fiber [93]. The other components make up the non-crystalline phase, which includes the cuticle, the cell membrane complex, the intermacrofibrillar matrix, the intermicrofibrillar matrix, and 40 % of the microfibrils, i.e., the nonhelical ends of the keratin molecules or of the dimers.

On stretching the fiber in the linear-viscoelastic region ($\varepsilon < 0.8$ %), the behavior of the crystalline phase (hx) is linearly elastic and

Table 3. Morphology of wool as a combination of two-component structures (see also Fig.)

Composite system	Type ∗	Component 1	Component 2
Wool fiber	r/c	cuticle	cortex
Cortex	f/m	cortex cell	cell membrane complex
Cortex cell	f/m	macrofibrils	intermacrofibrillar matrix
Macrofibril	f/m	microfibrils	intermicrofibrillar matrix

∗ r/c = ring/core structure; f/m = filament in matrix

that of the noncrystalline matrix (nc) is linearly viscoelastic.

The modulus of elasticity of the helices is [94]

$$E_{\mathrm{hx}} = 7.8 - 9.6\,\mathrm{GPa}$$

This value is remarkably close to the modulus of elasticity of crystalline ice, which is also stabilized by hydrogen bonds (ice I, 0 °C: $E = 10$ GPa). This fact emphasizes the ideal arrangement and cooperation of the short α-helical segments in the microfibrillar crystal. The time-dependent modulus of elasticity of the dry (d) and wet (w) noncrystalline phase (nc) are calculated to be

$$E^{\mathrm{d}}_{\mathrm{nc}} = 6.1 - 6.6\,\mathrm{GPa};\quad E^{\mathrm{w}}_{\mathrm{nc}} = 0.5\,\mathrm{GPa}$$
$$E^{\mathrm{d}}_{\mathrm{nc}}/E^{\mathrm{w}}_{\mathrm{nc}} \approx 13$$

The ratio of the moduli of elasticity of the noncrystalline phase is in good agreement with the ratio of the corresponding torsion moduli ($G^{\mathrm{d}}/G^{\mathrm{w}} = 10$) [94]. In a filament – matrix composite, the torsion behavior is mainly a property of the matrix. The small difference between the values for the viscoelastic modulus of the noncrystalline phase in the dry state and the modulus of elasticity of the helical filaments is due to the fact that both components are stabilized mainly by strong hydrogen bonds which are broken under the influence of water in the noncrystalline phase, while they continue to exist in the crystalline region.

For small amounts of longitudinal strain in the linear-viscoelastic region, the relaxation behavior of a wool fiber shows no fine structure ascribable to individual morphological components. This fact can be explained by the low contents of individual components and their chemical and structural similarity. For small deformations, all noncrystalline components behave as constituents of a homogeneous mixture, whereby the effect of each component is proportional to its volume fraction. The elastomechanical behavior is satisfactorily represented by the simple Feughelman two-phase model [95]. The elastic and viscoelastic properties of the fiber largely determine the crease resistance, dimensional stability, drape, and handle of a fabric.

3. Properties [2,6,17,48]

(See also → Fibers, 1. Survey; → Fibers, 6. Testing and Analysis)

3.1. Physical and Mechanical Properties [17]

Wool fibers with no medulla have a density of 1.31 g/cm^3 at 25 °C and 65 % relative humidity. The mean fiber diameter varies between 16 and 40 µm and the mean fiber length between 2.5 and 25 cm, depending on type and origin.

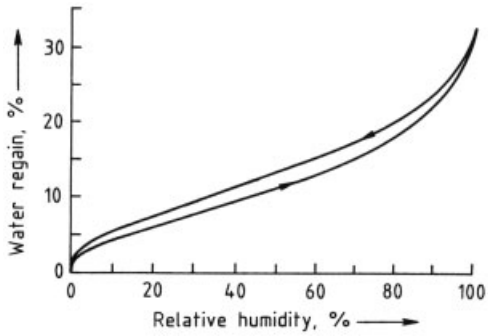

Figure 9. Moisture absorption of wool as a function of relative atmospheric humidity at 25 °C [98]

Water Absorption. Wool is hygroscopic, and the amount of water taken up depends on the relative humidity of the air, temperature, and the history of the wool (Fig. 9) [98, 114, 115]. The adsorption and desorption curves exhibit hysteresis. At average relative humidities, the

difference between the two curves is ca. 2 %. As with all textile fibers, the moisture uptake of wool is accompanied by liberation of heat.

Although wool is hygroscopic, its surface is hydrophobic towards liquid water and is therefore difficult to wet. This apparently contradictory behavior, which is an important factor in the physiological effects of wool as a clothing material, is due to the fact that the interior of the fiber absorbs water vapor, whereas liquid water is repelled by the hydrophobic outer surface of the cuticle.

Water absorption results in swelling, whereby wool shows considerable swelling anisotropy, i.e., longitudinal and radial swelling are different (Fig. 10) [10].

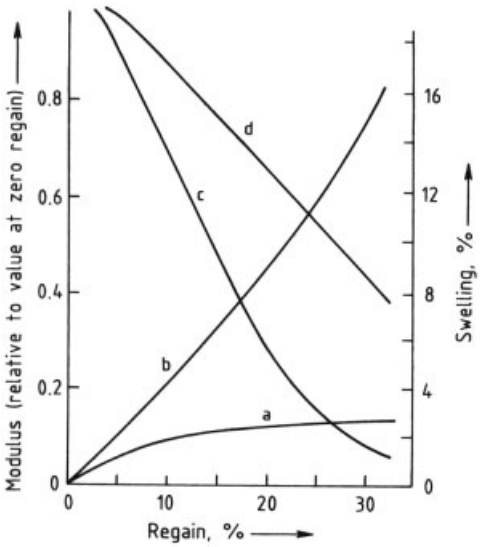

Figure 10. Effect of moisture absorption on swelling and moduli [10]
a) Length swelling; b) Radial swelling; c) Torsional modulus; d) Extensional modulus

An increase in the amount of absorbed moisture from 0 % to 33 % leads to a longitudinal swelling of ca. 2 % and a radial swelling of 16 %. The extent of swelling depends on the pH and composition of the swelling medium.

The degree of fiber swelling is least at the isoionic point (pH 4.9), where there is no excess of ionic groups. Above and below this point, the number of stabilizing salt bridges decreases, and the wool carries excess negative or positive charges which cause electrostatic repulsion of the peptide chains, thereby increasing the swelling [99]. Many physical and mechanical properties of wool depend on its moisture content. Thus, for example, dry undamaged wool has an ultimate tensile strength of ca. 150 – 200 N/mm^2 in its initial state and a wet ultimate tensile strength of only 70 – 80 % of this value. Whereas the tensile strength of wool decreases with increasing moisture content, the fracture strain increases. In the dry state, it is 35 – 55 %, and in the wet state 40 – 60 %.

Stress – Strain Behavior. Figure 11 shows the stress – strain behavior of a wool fiber in standard atmosphere of 65 % relative humidity at 20 °C compared with that in water at 20 °C. The stress – strain curve can be divided into three regions which are affected to different extents by increasing humidity. After decrimping (not shown in Fig. 11), the tension in the fiber increases very rapidly and almost linearly up to a strain of 1 – 2 %. This so-called initial region (A – B) is often incorrectly referred to as the "Hookean region". Above this, the elongation increases rapidly for small increases in stress. This section of the curve (B – C) is known as the yield region. Point C lies between 25 and 30 % elongation. The third region of the stress – strain curve is known as the post-yield region (C – D), and is terminated by rupture of the fiber (D). In this section of the curve, the increase in tension that accompanies a given increase in elongation is larger.

The three characteristic regions of the stress – strain curve are very distinct in water at 20 °C, and the transitions at points B and C are sharp. The slopes in the initial, yield, and post-yield regions are in the approximate ratio 100 : 1 : 10. It has been shown that distinct domains of the keratin filaments undergo the $\alpha \rightarrow \beta$ transformation in the three regions [116].

Noteworthy is that when the wool fiber is stretched in water to point C at room temperature, i.e., by ca. 30 %, there is complete reversibility of mechanical properties if the time in the stretched state does not exceed ca. 1 h. If the fiber is then relaxed and kept in water at room temperature overnight or at ca. 50 °C for 1 h, not only does the fiber return to its original length, but, on repeating the test, gives the same stress – strain curve as in the first cycle.

This high degree of reversibility is not exhibited by any other fiber.

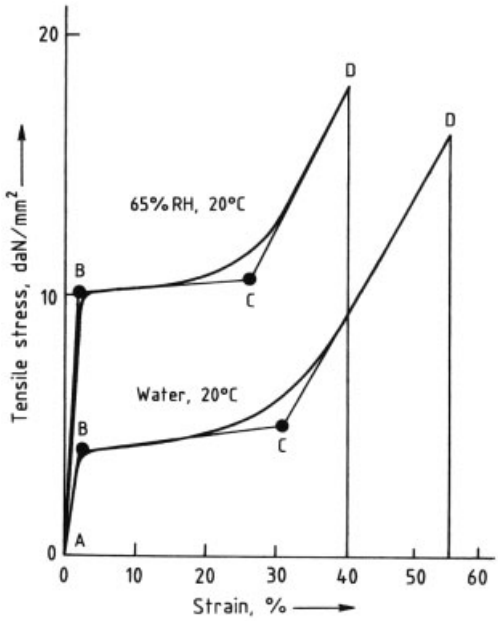

Figure 11. Tensile stress – strain diagrams

In addition to this simple tensile test, tensile elasticity testing can provide further information about the behavior of the fiber. Here the fibers are stressed several times between constant elongation limits or constant tension limits. The test enables the permanent elongation and the elastic elongation to be determined. In this way, the elastic behavior of fibers over the whole range of elongation up to the point of rupture can be determined.

Glass-Transition Temperature. Knowledge of the glass transition temperature T_g of wool [100] as a function of water content is of central importance for understanding its viscoelastic fiber properties, especially under the varying conditions of moisture content and temperature that can exist during the manufacture and wearing of woolen fabrics, and also for an understanding of the aging and tempering processes of these fabrics.

A simple equation due to Fox [101] represents the relationship between the glass-transition temperature of wool and its water content:

$$\frac{1}{T_g} = \frac{w_1}{T_{g1}} + \frac{w_2}{T_{g2}}$$

The mass fraction of each component is represented by w, the indices 1 and 2 denoting dry wool and pure water respectively. By replacing w_1 by (1-w_2), the Fox equation becomes

$$\frac{1}{T_g} = w_2 \left[\frac{1}{T_{g2}} - \frac{1}{T_{g1}} \right] + \frac{1}{T_{g1}}$$

If the reciprocal of the glass-transition temperature of the wool – water system is plotted against the mass fraction of water, a straight line is obtained, confirming the validity of the Fox equation [101]. The glass-transition temperatures of the pure components can be calculated from the slope and the intersection of this line with the axis. The time-dependent recovery of wool fibers after a fixed torsional deformation as a function of temperature and water content has been measured. For dry wool, this gave T_{g_1} = 447 K (174 °C), and for water T_{g_2} = 125 K (−148 °C) [100]. Figure 12 shows the glass-transition temperature of wool as a function of water content in the fiber.

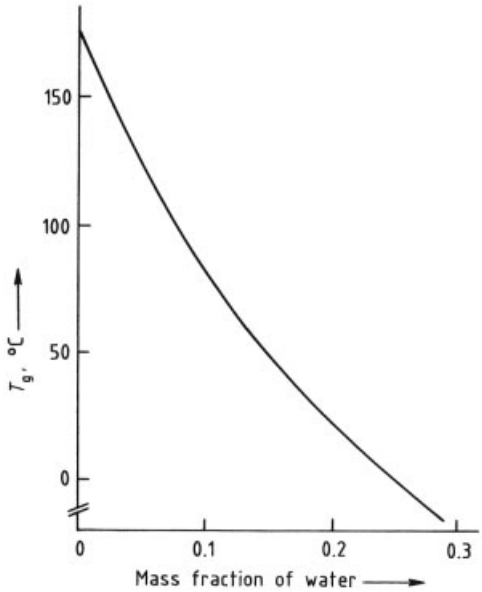

Figure 12. Glass-transition temperature of wool as a function of mass fraction of water in the fiber [100]

The temperature dependence of the physical properties of wool – water systems has been investigated by many authors using a variety of

methods. Wool in water has a T_g of $-5\,°C$, significantly below room temperature. Thus, water in wool eliminates aging and tempering effects because the thermal history of the wool is erased when T_g is exceeded, in analogy with the behavior of other glassy polymers.

3.2. Chemical Reactivity [5,13,27,31]

Dry Heat. Dry wool free of chemicals is stable for hours at 150 °C. Short heating periods, e.g., 185 °C for 30 s, produce tolerable changes in the fibers.

Moist Heat. Depending on the temperature, pH of the liquor, and duration of heating, heating with water can lead to degradation of wool. Heating with water at 130 – 140 °C leads to the disappearance of the α-keratin X-ray diffraction pattern with chemical decomposition. Wool is most stable near the isoionic point (pH 4.9), but chemical degradation occurs even below 100 °C. The cysteine in the wool proteins is decomposed first, liberating hydrogen sulfide which then attacks the cystine in the wool and accelerates the decomposition autocatalytically [102]. This is the cysteine – cystine self-degradation and self-cross-linking mechanism (thiol – disulfide degradation reaction).

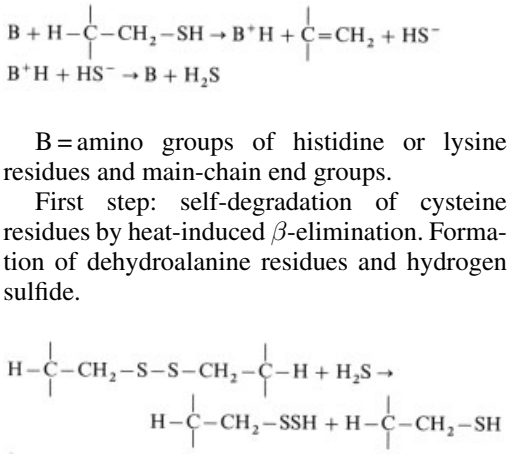

B = amino groups of histidine or lysine residues and main-chain end groups.

First step: self-degradation of cysteine residues by heat-induced β-elimination. Formation of dehydroalanine residues and hydrogen sulfide.

Second step: self-degradation of cystine residues by reduction of disulfide bonds with hydrogen sulfide. Formation of *S*-thiocysteine and cysteine residues.

Third step: Self-cross-linking by reaction of cysteine and lysine residues with dehydroalanine residues. Formation of lanthionine and lysinoalanine cross-links.

Beside cross-links based on sulfur, also further cross-links are formed in wool during heat exposure, e.g., linking via formation of dityrosine [117,118].

Acid Degradation. Mineral acids can degrade wool proteins to an extent that depends on pH, temperature, duration of reaction, and the presence of salts and surfactants. Sensitive sites in the wool proteins include the side chains of asparagine and glutamine, from which ammonia is released, and the peptide bonds formed by serine, threonine, cysteine, aspartic and glutamic acids, and tryptophan. The liquors from acid-degraded wool contain ammonium salts, free amino acids, and peptides ("wool gelatins") [103,104]. The hydrolysis of peptide bonds leads to weight losses and the liberation of end groups in the wool proteins. Carbonizing by heating wool impregnated with sulfuric acid produces an N → O peptidyl shift in the serine and threonine units of the peptide chain. The reaction is reversible if the acid is removed after the process, but if the wool is stored in moist air in the *O*-peptidyl form, the peptide chain is slowly and irreversibly hydrolyzed:

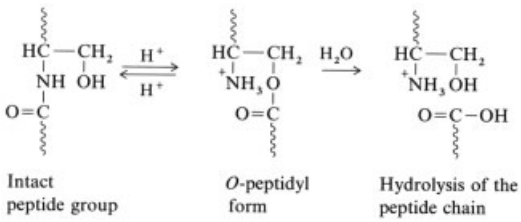

Degradation by Alkali. Wool is sensitive to alkali. In alkaline solution, the alkali binds to

the wool proteins with swelling and simultaneous degradation. Even at low hydroxide concentrations, the reversible binding is accompanied by fiber damage. The alkali binds to the wool so firmly that it cannot be completely removed by washing with water alone. The alkali salts of the wool proteins must first be decomposed by reaction with stronger acids. Wool containing alkali residues undergoes decomposition on heating in a stream of dry air, the most important degradation reactions being the decomposition of cysteine and cystine. Alkalis hydrolyze the primary carbonamide bonds of asparagine and glutamine, cleave peptide bonds in the main chains, and cause racemization of the amino acid residues:

$$\text{Cysteine} \rightarrow \text{dehydroalanine} + \text{hydrogen sulfide}$$

$$\text{Cystine} \rightarrow \text{dehydroalanine} + \text{thiocysteine}$$

$$\text{Dehydroalanine} + \text{cysteine} \rightarrow \text{lanthionine}$$

$$\text{Dehydroalanine} + \text{lysine} \rightarrow \text{lysinoalanine}$$

$$\text{Dehydroalanine} + \text{ammonia} \rightarrow \beta\text{-aminoalanine}$$

$$\text{Glutamine} \rightarrow \text{glutamic acid} + \text{ammonia}$$

$$\text{Asparagine} \rightarrow \text{aspartic acid} + \text{ammonia}$$

$$\text{Isoleucine} \rightarrow \text{alloisoleucine.}$$

Some of the wool proteins become soluble on degradation. The solubility of wool in alkali under standard conditions is used as an indicator of chemical changes in the fiber. Wool that has been damaged by acids, reducing agents, or oxidizing agents becomes sensitized to alkali, and its solubility in a standard solution of alkali increases.

Degradation reactions are also revealed by yellowing of the wool.

The decomposition reactions depend on the type of alkali, the temperature and duration of the reaction, the pH, and added salts and auxiliaries. Cationic surfactants increase the decomposition of cysteine and cystine and formation of lanthionine catalyzed by alkali. Nonionic surfactants have no effect on alkali damage to wool [105].

Reducing Agents. Reducing agents are used in bleaching, chlorine removal, stripping of dyes, discharge printing, production of permanent creases by ironing, and chemical setting. Reducing agents such as sodium mercaptoacetate and mercaptoethanol reduce disulfide groups to thiol groups. A 0.5 M solution of sodium mercaptoacetate dissolves wool above pH 10 in a few hours by reducing the cystine bonds responsible for insolubility:

$$\text{WoCH}_2\text{SSCH}_2\text{Wo} + 2\,\text{HSCH}_2\text{COONa} \rightarrow 2\,\text{WoCH}_2\text{SH} + (\text{SCH}_2\text{COONa})_2$$

where Wo = peptide chains of the wool proteins. Reaction of the thiol groups with iodoacetic acid produces *S*-carboxymethylkeratins:

$$\text{WoCH}_2\text{SH} + \text{ICH}_2\text{COONa} \rightarrow \text{WoCH}_2\text{SCH}_2\text{COONa} + \text{HI}$$

Reversible reduction of cystine bridges without dissolving the fiber can be achieved with thiols if the reaction conditions are sufficiently mild. This is the basis of the cold permanent waving process.

The disulfide bonds in wool can also be reduced by aqueous solutions of sodium sulfide. This reaction is of industrial importance in the production of lime wool, obtained by dehairing the skins of slaughtered sheep.

The sulfitolysis of wool is an important chemical reaction, e.g., in permanent setting. In the reaction with sulfite or hydrogensulfite, 1 mol Bunte salt (*S*-sulfonate anion) and 1 mol cysteine are produced per mole of cystine:

$$\text{WoCH}_2\text{SSCH}_2\text{Wo} + \text{SO}_3^{2-} \rightleftharpoons \text{WoCH}_2\text{S}^- + \text{WoCH}_2\text{SSO}_3^-$$

$$\text{WoCH}_2\text{SSCH}_2\text{Wo} + \text{HSO}_3^- \rightleftharpoons \text{WoCH}_2\text{SH} + \text{WoCH}_2\text{SSO}_3^-$$

Maximum cleavage occurs in the pH range 3 – 6. On rinsing, the Bunte salt and cysteine groups recombine to give wool cystine. Oxidative sulfitolysis converts disulfide into two *S*-sulfonate anions:

$$\text{WoCH}_2\text{SSCH}_2\text{Wo} + 2\,\text{SO}_3^{2-} + [\text{O}] + \text{H}_2\text{O} \rightarrow 2\,\text{WoCH}_2\text{SSO}_3^- + 2\,\text{HO}^-$$

In the reaction of wool with reducing agents, cleavage of the stabilizing disulfide bonds leads

to an increase in solubility. A linear relation exists between alkali solubility and cystine content of oxidatively and reductively bleached wool. At the same time, the enzymatic degradability increases. The dry stength does not always decrease, but the wet strength does. Plasticity increases, which is important for the process of permanent setting by deformation followed by oxidation.

Oxidizing agents attack the disulfide bonds. The reaction proceeds via various cystine oxides (sulfoxides and sulfones) and leads finally to cleavage of the disulfide bond with formation of cysteic acid. Cysteine, methionine, tryptophan, and tyrosine are also attacked. The extent of oxidation depends on the oxidizing agent and its concentration, pH, temperature, and reaction time. Of industrial importance are oxidation reactions with hydrogen peroxide (for bleaching), and with potassium manganate(VII), chlorine, peroxysulfuric acid, and peroxyacetic acid to give antifelting effects.

In the chloramide reaction, peptide bonds are chlorinated:

$$-\mathrm{CONH}- + \mathrm{Cl_2} \rightarrow -\mathrm{CONCl}- + \mathrm{HCl}$$

Photodegradation [106–113]. Wool shows only limited fastness to light. On exposure to daylight wool is bleached after a short time. After prolonged irradiation or irradiation with UV light the wool fiber yellows. Visible blue light (380 – 475 nm) is responsible for photobleaching, while maximum photoyellowing occurs with UV-B radiation (290 – 320 nm). Photodamage results in the formation of a variety of oxidation states of cystine residues, and mercaptans. Phenylalanine, tryptophan, tyrosine, histidine, and proline residues are other targets for photodamage reactions.

Scission of the main protein chain occurs with formation of keto-acyl end groups. Progressive formation of nonspecific protein cross-links (e.g., dityrosine) has also been observed.

Enzymatic Degradation. Under unfavorable conditions, wet wool is attacked by bacteria or fungi, and this can lead to damage to wool products at all stages of manufacture. Such damage can be detected microscopically by the specific color reaction with lactophenol cotton blue and by splitting of the fibers into multicellular fibrillation products, cortex cells, and macrofibrils. These (microbial) degradation reactions are caused by bacteria or fungal proteases. In undamaged wool, mainly the endocuticle and cortical cell nucleus residues are degraded hydrolytically by proteases. Macrofibrils remain intact. Damaged wool is attacked more strongly, especially at weathered fiber ends.

Enzymes are of industrial importance in wool processing, e.g., in the dissolution of skin residues from wool felts by selective degradation, and the removal of skin-scale proteins from washed wool. Enzymatic dehairing of the small hides from sheep and goats is an established process.

The use of enzymes for pretreatment and finishing is described in Sections 5.3 and 5.4.

4. Resources and Raw Materials

4.1. Sheep Rearing and Breeding [44]

Sheep growing is best suited to drier regions of the world. Steppe and savannah pastures are suitable, but tropical regions are not. Sheep have also been bred in countries with intensive agriculture, where the waste products from arable farming are used as feed. Climatic factors such as altitude, precipitation, and air temperature are important for the growth of sheep and their wool. Sheep acclimatize well, and their bodies can adapt to extreme conditions and moderate food supplies.

Depending on the climatic and economic conditions, sheep-farming countries specialize in wool production (Australia and South Africa) or the production of wool and meat (New Zealand, Argentina, and Uruguay). The crossing of English mutton rams and merino yews produces the Corriedale sheep of New Zealand. In Australia, merino rams have been crossed with Lincoln yews. These crossbreds have become established in South America as well as New Zealand, and are good producers of meat and wool. With the introduction of frozen meat technology in 1882, the crossbreds became very

important. The meat is mainly exported to the United Kingdom.

4.2. Places of Origin and Types of Raw Wool

Places of Origin. Australian and New Zealand wool; South African wool (also known as Cape wool) including Karakul and indigenous wool; European wool, e.g., German, English, Scottish, Irish, and French; South American wool, e.g., Argentinian (e.g., Buenos Aires, Bahia Blanca, and Chubut wool), Chilean (e.g., Punta Arenas wool), Uruguayan (e.g., Montevideo wool) and other South American wools from Brazil, Peru, etc.; exotic wools, e.g., from North Africa, Iran, and other Asian countries (China, India, Mongolia, etc.).

Types of Sheep. Merino wool from merino sheep or sheep of the merino type, mean fiber diameter usually $\leq 25\,\mu m$; crossbred wool from crossbred sheep, mean fiber diameter usually $> 25\,\mu m$; comeback wool from sheep obtained by crossbreeding crossbreds with merino sheep; coarse wool from coarse wool sheep, mean fiber diameter usually $> 30\,\mu m$.

5. Production Processes

5.1. Methods of Obtaining Wool [18]

By far the largest proportion of wool is obtained by shearing live sheep.

In addition, biological methods are used to recover wool from the skins of slaughtered animals. The product is known as sweated, pulled, or Mazamet wool, after the French town of that name. In New Zealand, wool recovered from skins is traded under the name slipes. It is obtained by treatment with a mixture of lime and sulfides. Reclaimed wool, obtained by disintegrating used woollen textiles, differs in having considerably shorter fibers and lower quality and purity. In accordance with textile labeling regulations, it is labeled as pure wool.

5.2. Raw Wool Scouring (See also → Textile Auxiliaries, Chap. 4.3.)

Wool obtained by sheep shearing contains varying amounts of impurities which must be removed in the scouring process. The pure wool content of raw wool is variable, the following being typical average values for Australian raw wool:

Pure wool	55 %
Wool grease	10 %
Wool suint	4 %
Soil and vegetable matter	6 – 20 %
Moisture	12 %

As wool is bought and sold on the basis of its pure wool content, the method of determining this has been standardized (IWTO-19-76).

Wool grease is highly soluble in organic solvents such as benzene, hexane, acetone, and perchloroethylene. Consequently, there have been many attempts to use these solvents in wool scouring. However, scouring with water or aqueous alcohol is also necessary to remove the wool suint, skin-flake protein, and soil impurities. Solvent-based processes are, therefore, of minor importance. Solvent scouring systems are the de Smet plant using hexane and isopropyl alcohol as solvents [146], and the Wooltech system using 1,1,2-trichloroethylene as solvent [147, 148]. However, the major wool scouring is performed today by emulsion processes. To evaluate the scouring result, the residual fat content, whiteness, ash content of the wool, and pH of the aqueous extract are determined (IWTO Methods, see Chap. 8).

In scouring machines, vigorous motion gives better washing, but can lead to a higher level of felting and hence to fiber shortening during later stages of the manufacturing process. This problem has been solved satisfactorily by the sieve-drum scouring process. Several other processes have been developed to improve wool scouring. These processes meanwhile all use suction drum bowls in combination with rake systems. The New Zealand Minibowl process and the Australian Siro Scour process have been installed in some industrial plants, but are not yet established in Western Europe.

In wool scouring, ca. 6 – 20 L of wastewater is produced per kilogram wool, depending on the process and raw wool, and even if most of the wool grease is removed, e.g., by floccula-

tion with acid and polyelectrolytes, the remaining wastewater is still highly contaminated. The pollutants still present in the wastewater can be 80 – 90 % degraded in multistage chemical and biological plants. Evaporation and incineration plants are able to destroy > 99 % of the pollutants and recover energy, water, and detergents. This has provided a satisfactory solution to the problem of wastewater purification.

Although it is no longer worth recovering potassium carbonate from the scouring water, recovery of the byproduct wool wax is economically viable (→ Waxes, Chap. 2.3.3.).

Wool-scouring plants are today mainly located in the wool-producing countries, but wool-combing plants (which are normally integrated with wool-scouring plants for cost and quality reasons) are found both in the wool-producing countries (advantage of proximity to the raw material) and in the processing countries (advantage of proximity to the market).

5.3. Mechanical Processing

Total annual wool production, expressed as pure scoured wool, is 1.4×10^6 t in 1998, of which ca. 600 000 t is converted into worsted yarn and 400 000 t into woolen yarn and carpet yarn [149].

To produce worsted yarn, the scoured wool is first disentangled by carding, which removes most of the nonfibrous impurities. The web from the carding machine is gathered to form the so-called sliver. Several slivers are combined on the first drawing frame and stretched, thereby improving uniformity. This is followed by combing, restretching, and final stretching. The back washing process once carried out before final stretching is not now usually used. The combed slivers ("tops") from the final stretching process, which have linear densities of 14 – 24 g/m, are an important commerical form of wool.

The top undergoes further combing passes to give uniformity and can be blended with other types of fiber, e.g., synthetic fibers. The "roving" is then produced by stretching on rubber rollers or on a "flyer", and this is then spun to produce worsted yarn. Ring spinning frames are generally used for this. Rotor spinning frames are now mainly used for short fibers and blends of short fibers. Several developments to increase the production of the ring spinning process, like centrifuge spinning, pot spinning, etc., are being tested in pilot plants. So far, none of these processes is used commercially

In the production of woolen yarn, the slubbing is directly produced by splitting up the carded web, and is then spun on a ring spinning machine. Both worsted and woolen yarns are converted to fabrics by the standard methods also used for other types of fiber (weaving, knitting, woven and tufted carpet manufacture, etc.).

The property of the wool fiber of migrating in various directions within the fiber bundle on mechanical treatment and in the presence of water and auxiliaries, which is due to the frictional difference, is utilized in the production of felt from loose wool (carded web) or woven materials (→ Felts, 2. Felting Process, Chap. 2., → Felts, 3. Pressed Felts, Chap. 3.). Fiber migration is also used to cause fulling of wool fabrics and carpet yarns.

5.4. Chemical Processing: Pretreatment

Carbonization (→ Textile Auxiliaries). The mechanical processes of carding scoured raw wool and of combing prior to spinning worsted yarn remove most of the vegetable matter. However, some of these impurities remain in the wool, so that the next stages of processing, e.g., the conversion of noil to wool yarn, cannot usually be carried out immediately. Noils and woolen fabrics, which sometimes also contain unacceptable amounts of vegetable matter, are therefore carbonized. This name is derived from the fact that the impurities, which consist mainly of cellulose, form dark to black carbonization products after impregnation with sulfuric acid and heating. These can then be removed from the fiber bundle because of their brittleness.

In recent research, the use of enzymes to remove vegetable matter is being investigated [150, 151].

If carbonized material is not dyed while it is still wet and contains acid, the fabric must be protected from the effects of moisture and light, and the storage time must be as short as possible. Also, the dyeing method used must be as mild as possible.

Antifelting Treatments. Most of the antifelting treatments developed for wool modify the surface of the fiber by chemical attack. This affects the scales such that the difference between the "roughness" of the fiber surface in the tip-to-root and root-to-tip directions is reduced or even eliminated. At one time it was believed that antifelting treatments had to remove the surface scales entirely. It is now known that this is not so; a treated fiber will often look no different from an untreated fiber under the microscope [29].

Antifelting finishes work by partially removing, softening, or coating the scale layer, or by preventing movement of the fibers relative to each other by spot welding. Treatment with adhesives can only be used on the finished fabric, whereas the other processes can also be used on tops. They include chlorination and treatment with manganate(VII) or peroxysulfuric acid. This must be carried such that only the scale layer is attacked and not the interior of the fiber. Precise process control gives very good stability to washing accompanied by low losses in weight and strength.

A process widely operated in Germany that can also be used on the top is a combination of relatively mild chlorination followed by treatment with a polyamide – epichlorhydrin resin (e.g., Hercosett 125). This is today the most important process for the production of machine-washable articles of pure wool, and gives products complying with the specifications of the IWS Super Wash mark. The chlorination stage of the process is carried out with hypochlorite, whereas the Kroy process uses gaseous chlorine. In 1994, the total quantity of wool sliver treated worldwide in accordance with the IWS Super Wash specifications was ca. 25 000 t.

The so-called Vantean process, an oxidative process developed in Japan, comprises the following steps:

1) Pretreatment with heavy metal ions (can be omitted)
2) Treatment with hypochlorite in a saturated, strongly acidic salt solution
3) Treatment with sulfite
4) Stabilization with formaldehyde

The process removes the cuticle almost completely, and gives the wool a luster resembling that of much more expensive fine animal hairs. For ecological reasons, work is directed towards the development of a new antifelting process that does not use chlorine or chlorine compounds [152].

The most promising developments so far are treatment with peroxysulfuric acid and manganate(VII), but a commercial process of this type for tops treatment is not yet available.

Research is being carried out into antifelting effects produced by plasma technology [119] or topochemical enzyme processes [120]. Recently, an environmentally friendly procedure for antifelting treatment of wool was developed, involving replacement of the conventional chlorination by a treatment with a low-temperature plasma, followed by application of a new developed polyurethane based polymer [153–156]. The development in plasma technology is far advanced and tests with a pilot plant are running. A commercial process will be available in the next years.

Bleaching of wool is not as important as the bleaching of natural cellulose fibers. For many purposes, the natural cream color of wool is acceptable, e.g., for garments finished in this color or for those dyed in dark shades. For white articles and sometimes for pastel shades, the wool must be bleached, preferably with hydrogen peroxide (see → Textile Auxiliaries).

Naturally pigmented dark wool fibers require special treatment. Although they are present in very small numbers in high-quality wools (a few naturally colored fibers in 10 g wool), they have a detrimental effect on the appearance of garments dyed in pale shades. The colored fibers can be selectively bleached by a process in which the wool is pretreated with an iron salt, carefully washed, and then bleached with hydrogen peroxide [121].

For carpet wools, bleaching during raw wool scouring, simply by adding hydrogen peroxide or reducing agents to the last bowl, is a process of some importance. This practice can lead to increased wool yellowing during dyeing but is commercially attractive because the whiter wool attracts a higher price.

The highest level of whiteness can be obtained by a combination of bleaching with hydrogen peroxide and reductive bleaching with, e.g., Blankit products, which contain sodium

dithionite and sometimes also an optical brightener. However, care has to be taken in this bleaching process as fiber damage can occur.

Bleached and, to an even greater extent, optically brightened wool undergoes more yellowing than unbleached wool on exposure to light. Solubility in alkali is a sensitive indicator of bleaching damage to wool (see page 13). However, it is possible that although the alkali solubility (30 %) may be twice its normal value (12 – 17 %), a decrease in breaking strength of only 2 – 5 % occurs.

New Developments. In the last few years there has been a growing demand for elastic fabrics and easy care properties, not only for sports and leisurewear but for formal wear also, particularly for wool and wool rich trousers (Wool plus Lycra). The disadvantage of elastomeric fibers in the fabric is that these fibers are not dyeable to the fastness standards of the main fibers in the cloth. Some natural stretch can be produced in wool fibers by subtle yarn technology, however, a 20 % stretch requires the application of reducing agents. Sulfite-based textile auxiliaries were developed that provide worsted wool fabrics with a natural stretch [157]. This permanent elasticity of yarns and fabrics is achieved through supercontraction of the wool fiber.

A new technology for reduction of the fiber diameter of wool fibers has been developed by Woolmark Company and CSIRO; both, process and wool fibers have the trade name OPTIM [158, 159]. OPTIM fine (permanently stretched fiber) reduces the diameter of wool by 3 – 4 µm. Another variant, OPTIM max (temporarily stretched fiber) is a fiber which is able to shrink by 20 to 25 % so that a higher bulkiness and softness is achieved. The process combines a stretching and either permanently or temporarily set in stretched configuration ($\alpha \rightarrow \beta$ transformation) so that a new wool fiber with different chemical and physical characteristics is produced [135], due to the $\alpha \rightarrow \beta$ transformation OPTIM fine has a more silk-like character than untreated wool [159].

5.5. Dyeing and Finishing [37, 40, 45, 50]

Dyeing and Printing. Dyeing is the most important wet finishing process for wool. Wool can be dyed in most stages of its manufacture, i.e., as flock, slubbings, yarn, or piece. The dyeing of flock is of importance in the production of melange shades, of woolen yarn consisting of blends with other types of fiber or of large batches of uniform shades. The dyeing of sliver is useful for the production of large batches of completely uniform color. Color discrepancies can usually be leveled by mixing slivers without additional dyeing. Dyeing of wool blends is described in [45, Chap. 9].

The dyeing and printing of wool are treated in detail in → Textile Dyeing, Chap. 5.1., → Textile Printing, Chap. 7.). The most important dye classes for wool dyeing are afterchrome dyes (ca. 25 – 30 %), acid dyes (ca. 25 – 30 %), 1 : 1 metal-complex dyes (ca. 7 %), 1 : 2 metal-complex dyes (ca. 30 %), and reactive dyes (ca. 10 %) [160]. Chrome dyes continue to fulfill a high proportion of wool dyeing capacity, particularly as a cost effective, high-wetfastness system for slubbings and loose wool. Chromium residues (especially Cr(VI)) in the dyed textiles can cause allergies and toxic skin reactions in sensitive individuals. Due to the allergic and toxicological potential of chromium the usage of afterchrome dyes for wool dyeing is strongly decreasing. Furthermore, residual chrome creates a problem in dyehouse wastewater. This has been minimized by the use of modified chrome dyeing techniques [122–124].

Reactive dyes represent an alternative method of achieving the high wetfastness of chrome dyes whilst eliminating the problem of heavy metal contamination of dyehouse, scouring liquors, and dyed textiles [161].

The tenacity of wool is often reduced in stock-dyeing processes. As a result of this, many fibers are broken in the card. Most of the reduction in fiber strength is due to setting of bends or curvatures of the packed fibers by the hot water treatment. Such set fiber bends are nonuniformly strained by tensions and thus resist smaller loads than fibers set in the straight state [129].

In addition to the minimum wool damage exerted due to their optimum application pH in the isoionic region, reactive dyes further enhance wool quality by chemical reaction with thiol groups, H_2S, and histidine side chains, and by acting as antisetting agents [125–128].

Research studies have been performed to dye wool from nonaqueous media, i.e., from super-

critical carbon dioxide [162, 163]. Wool can be dyed from supercritical carbon dioxide with hydrophobic dyes, improved fastness properties are achieved by application of reactive disperse dyes [162, 163].

Setting. According to DE BOOS [130], setting is an essential feature of the finishing of wool fabrics. It is used to stabilize fabrics prior to scouring and dyeing, to provide the required thickness, handle and visual properties to fabrics, to provide some durability to the final finish, and occasionally to minimize dimensional changes in subsequent tailoring and wear. For these reasons, setting is an integral part of good finishing and fabric quality. Cohesive set in wool fibers is thought to involve rearrangement of the hydrogen bonds in the fiber, and is imparted whenever wool is distorted at temperatures above the glass-transition temperature (Fig. 12) of the fiber and cooled while distorted. This set configuration of the fabric is stable as long as the fibers remain below their T_g.

Permanent set in wool requires the rupture of disulfide bonds in addition to the hydrogen bonds in the fiber. Most authors [131, 132] believe that thiol – disulfide exchange reactions at temperatures in excess of 60 °C are required to impart permanent set. Others [131–135] consider the "transition temperature" for the onset of thiol – disulfide exchange reaction as the temperature required for chemical cysteine degradation and H_2S formation (see page 13).

The decisive role of cysteine thiol groups as initiators of the reactions that lead to permanent set is common to both theories. Thus, all industrially proven methods of minimizing dyeing damage due to permanent set are based on the same principle, i.e., blockage or oxidation of the cysteine thiol groups during dyeing [136, 137].

The effectiveness of reactive dyes or other reactive chemicals is due to the fact that they block not only the cysteine thiol groups, which decreases the rate of H_2S formation, but also amino groups of the wool proteins, preventing the catalysis of cysteine degradation reactions.

Shrink-Resistant Finishes. Fabrics or garments can be treated with nonchlorination processes which utilize polymers alone (→ Textile Auxiliaries, Chap. 7.6.5.).

Protection from Moths and Beetles [138–141]. See → Insect Control, Chap. 14.4..

Permethrin-based insect-resist (IR) products account for ca. 90 % of the European market, based on the volume of treated wool. Permethrin-treated wool has poor fastness to moisture and light, so that treatment with these products is recommended mainly for carpets and for storage protection.

Conventional wool-specific protection agents based on Sulcofuron are Mitin FF h.c. (80 % AS, Ciba Specialty Chemicals), and Mitin FF liquid (Ciba Specialty Chemicals). Wool treated with Mitin FF has very good fastness towards moisture and light, and is therefore suitable for articles from which high standards of colorfastness are demanded.

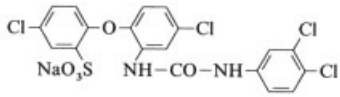

Sulcofuron

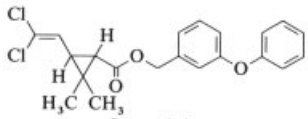

Permethrin

Mothproofing agents are usually applied during the exhaustion dyeing process, i.e., are added to the dyebath and are taken up by the wool with the dye. Other possible methods include application during scouring.

Flameproofing. Wool is less flammable than cellulose and synthetic fibers (see → Textile Auxiliaries, Chap. 7.5.). Its high ignition temperature, high limiting oxygen index, low heat of combustion, and low flame temperature are particularly advantageous and are connected with the chemical and morphological structure of the wool fiber, which has high nitrogen (16 %) and moisture (10 – 14 %) contents.

Negatively charged titanium and zirconium complexes with α-hydroxycarboxylic acids and fluorides can be exhausted on the positively charged wool fiber in acid conditions, the result being a significant improvement in the natural flame resistance of wool that is fast to washing and dry-cleaning. The complexes with carboxylic acids are exhausted on the wool fiber

at the boil, whereas the fluoride complexes can be effectively exhausted at lower temperatures. The fluoride complexes can also be applied by a pad – batch – rinse – dry technique. Titanium complexes are more effective than the zirconium ones, probably because of better penetration of the fiber by the smaller titanium complexes. However, titanium complexes cause yellowing of wool, which increases with light exposure. Zirconium complexes do not affect the shade of wool and are fast to light [142].

Lightfastness Treatment. The effect of light on wool is first to bleach it (photobleaching) and then to cause yellowing (photoyellowing). For interior woolen textiles photobleaching is the main problem because window glass filters short-wave UV light which induces photoyellowing. However, visible light (especially blue light, i.e., light with a wavelength of 400 – 450 nm), which causes a marked photobleaching, passes window glass. A new approach was developed by WRONZ-Institut to prevent photobleaching by application of a sulfonated aromatic compound (Lanalbin APB) [164].

Photobleaching is especially marked when woolen textiles are also thermally stressed (e.g., in automobiles). The lightfastness of wool, including bleached and dyed wool, can be improved by the application of sulfonated UV absorbers, hydroxybenzophenones (Uvinuls) or hydroxybenzotriazoles (Cibafast W) from the dyebath [143–145].

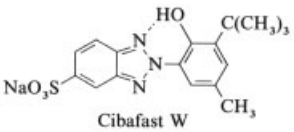

Cibafast W

Hydrophobization and Oleophobization Treatments. If necessary, wool, like other fibers, can be made water- and oil-repellent by treating it with silicones or fluorinated hydrocarbons.

6. Environmental Protection

Like the other organic natural fibers, cotton and silk, wool is an environmentally friendly, renewable textile raw material. Woolen materials are reusable. Wool wastes are biodegraded by bacteria, fungi, and insects (moths, carpet beetles) and can be used as long-acting nitrogen fertilizers [165].

Wool production and wool industry, however, are faced with wastewater limitations due the pesticide residues in the clip [166], the highly polluted wool scouring effluent (see Section 5.2) [167], AOX in the shrink-resist processes, chromium in the dyeing process [168, 169], and the application of insect resist (IR) agents [170]. Dyeing and finishing with minimum environmental impact is crucial to the manufacture of eco-textiles [171, 172].

Modern effluent treatment removes pollutants by ecologically tolerable processes. By strict compliance with dyeing instructions the chromium(VI) content in the effluent can be kept within tolerable limits. Legal and environmental aspects of the textile finishing industry are described in → Textile Dyeing, Chap. 14.; → Textile Auxiliaries, 1.4 Environmental Aspects, Chap. 1.4., → Textile Auxiliaries, 4.4 Ecological Aspects of Fabric Pretreatment , Chap. 4.4.

A detailed and authoritative account of the environmental aspects of wool production and wool processing is afforded by the published proceedings of the two quinquennial wool textile research conferences, held in 1990 and 1995 [42, 52].

7. Quality Specifications [48, 53, 173]

Testing methods are described in → Fibers, 6. Testing and Analysis, 2. Tests of Form and Mass, Chap. 2., → Fibers, 6. Testing and Analysis, 3. Tensile Properties, Chap. 3. Quality of wool or tops implies a statement of the type, origin, and spinnability of the raw wool. An important quality criterion is mean fiber diameter (fineness), which influences the softness of hand and is of paramount importance in spinning. Price is determined mainly by fineness.

Wool fineness is specified by the mean fiber diameter in micrometers. In the past, there were various systems of quality specification, e.g., as used by spinners, merchants, and sheep farmers. The most important systems are the British (numbers ranging from 100's to ca. 36's), French

(numbers in the ranges 150 – 100, I – VI, and C), and German [letters in the range AAAA (4 A) to A, B, etc., up to EE]. Table 4 compares the British system with mean fiber diameter and fineness. Recently British mills have agreed that the existing micron definitions should be extended for superfine wools to 100's – 200's (i.e., 17.0 to 13.5 µm mean fiber diameter), because of increased demand for fine fabrics for traditional tailoring.

Table 4. British quality numbers, mean fiber diameter, and mean fiber fineness [44]

British quality number	Mean fiber diameter, µm	Mean fiber fineness, dtex
90's	≤17	≤3
80's	17 – 18	3 – 3.3
74's	18 – 19	3.3 – 3.7
64's	21	4.5
60/64's	22	5.0
60's	23	5.4
58's	25	6.4
56's	28	8.1
50's	33	11
48's	34	12
46's	37	14

The British grade number gives the number of hanks of 560 yards (512 m) that can be spun from 1 lb (453.6 g) of wool (spinning limit).

For fine yarns, i.e., those with on average ca. 35 fibers in the cross section, wool-fiber-diameter distribution affects spinning performance, as well as yarn evenness and tensile properties. Effective fineness, a parameter derived from yarn-evenness theory, is a better predictor of (fine) physical yarn properties than mean fiber diameter alone [174].

There are two systems for obtaining an objective measurement of hand. In the *Kawabata System*, 17 fabric properties relevant to the hand are determined with four measuring devices. These enable statements to be made about the stiffness, elasticity, and surface condition of the fabric, and objective dimensionless hand indices can be calculated from these if required [175, 176].

In the *FAST System* (Fabric Assurance by Simple Testing), 14 fabric properties are determined using three compact instruments. When these are plotted on a multiaxial graph (the so-called fingerprint), conclusions can be drawn about the processing properties of the fabric during manufacture [177].

8. Testing and Analysis [46, 48]

Chemical methods of testing wool to assess damage at the various stages of manufacture are well established. The use of a combination of methods enables the causes of impairment of mechanical properties to be elucidated with a high degree of certainty. The methods include the determination of foreign substances (wool grease, spinning oils, ash, etc.), amino acids [178], and alkali or urea – bisulfite solubility.

It is not possible to quote mandatory values of mechanical and chemical data for a given wool. It is only possible to collect typical values for undamaged wools [179]. These analytical data vary on wet finishing of the wool in various directions and to various extents. Only in the case of certain standard treatments, such as carbonization, peroxide bleaching, or chlorination, can definite tolerances be given which must not be exceeded after a correctly performed process in order to ensure a defect-free product. Reliable figures are obtained if samples are taken by standardized methods at the various stages of a treatment process and analyzed. Most chemical testing methods for wool have been standardized by the Technology and Standards Committee of the International Wool Textile Organization (IWTO) [180].

The most important criteria for the assessment of changes to wool caused by industrial processing, especially wet finishing, are mechanical properties such as ultimate tensile strength. The dry and wet strength and elongation of yarn bundles are determined (IWTO method 32-1982). The dry values for undamaged wool are in the range 10 – 12 cN/tex, but these can decrease on dyeing. Decreases by up to 15 % of the maximum ultimate tensile strength are acceptable for finishing processes.

Efficient separation methods for amino acids and proteins by chromatography and electrophoresis (see → Amino Acids, Chap. 6.) are applied in wool research for estimation of lanthionine, lysinoalanine, histidinoalanine, cysteic acid in hydrolysates of wool and *S*-carboxymethyl proteins in mercaptolytic digests. One-dimensional electrophoresis allows up to 50 distinct wool proteins to be separated (Fig. 13). Comparison of electrophoretic patterns of digests of dyed and finished wool textiles allow specific statements on the involvement

of morphological components and chemical constituents, such as keratins versus keratin-associated proteins [181–184].

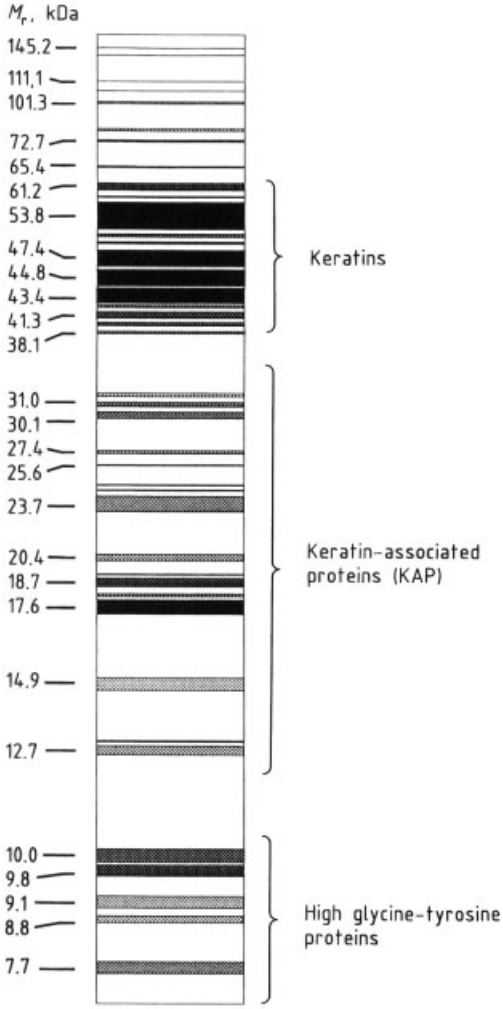

Figure 13. Schematic electrophoretic protein-separation pattern of merino wool [183]

9. Storage and Transportation [185]

Wool must be stored and transported under normal atmospheric conditions. Bales of raw wool, scoured wool, tops, or any other form should be kept dry to avoid bacterial and fungal attack. The packaging material must allow some exchange of air to avoid damp and high-humidity conditions in the bale. The wool pack material should not pose problems in dyeing and finishing.

10. Uses [32]

Merino and crossbred wools are mainly used in clothing, home textiles (carpets, domestic furnishings, and bedding), to some extent in automobile upholstery (most often blended with polyester) and in aircraft seatings [186–188]. According to [189], the wool fiber can now be found in applications as diverse as thermal insulation, weed mats, geotextile products, and as a means to soak up oil spills. Wool has been used for thermal insulation in houses in Europe for many decades and is being increasingly used for both homes and commercial buildings [190–194]. The natural property of wool to effectively absorb specific classes of widespread indoor air pollutants, namely aldehydes, SO_2 and NO_x, has led to new applications in the private and commercial sector [195–197]. Wool is naturally fire resistant, and the increased use is being driven by consumer demand for natural products in their homes. The moisture absorbing and desorbing properties of wool can also contribute to passive heating and cooling effects and, led, furthermore, to the use of wool for high-performance sportswear clothing. Of minor importance are special wool products, especially felts, e.g., for filter materials.

Keratin hydrolysates can be isolated from wool by acid, basic, or enzymatic hydrolysis in combination with a reductive (or sometimes oxidative) treatment. These hydrolysates are used as fiber protective agents during wool dyeing [198], for cosmetics, or for production of new textile materials [199]. L-Cystine can be isolated from hydrolysates of hair or wool (see → Amino Acids, Chap. 3.2.4.).

11. Trademarks

The Woolmark Company was founded in 1937 as the International Wool Secretariat. Today it is known as The Woolmark Company and is the world's leading wool textile marketing organization. Together with its three well known trademarks, it is synonymous with research and

development, innovation, promotion, and quality assurance in pure new wool and wool blend products (cf. http://www.wool.com).

The *Woolmark* symbol was designed by an Italian, Franceso Sarroglia to identify quality products made from pure new wool (Fig. 14). Textiles with this mark must consist of 100 % virgin wool. Admixtures of wool recovered by mild methods from the skins of slaughtered sheep are permitted. Equal status is given to fine animal hair from the alpaca, llama, vicuna, guanaco, camel, angora, rabbit, angora-mohair, cashmere and cashgora goats, and yak, provided that here also it is obtained by a mild process. The Woolmark was launched in 1964, and is now a registered trademark in over 140 countries. It is used by companies licensed by the Woolmark Company in more than 65 countries extending across a wide range of products – clothes, carpets, home furnishings, washing machines and detergents. In 1970, the designation "pure new wool" (Reine Schurwolle) was included in the textile labeling law (TKG) of the Federal Republic of Germany, and became valid in the European Community in 1972. The *Woolmark Blend* symbol (originally called Woolblendmark) was introduced in 1971 to identify quality products made from wool rich blends (products comprising at least 50 % new wool). The *Wool Blend* symbol was introduced in mid-1999 to extend the opportunities offered by The Woolmark Company's blends program; the Wool Blend symbol is used to identify products that contain at least 30 % (but less than 50 %) new wool (Fig. 14).

The same high performance standards apply to all three trademarks. Before a manufacturer can use the Woolmark or any trademarks of the Woolmark Company he needs to apply for a license, and pay the relevant usage fee [200]. He also needs to be able to meet the quality standards that are required under the Woolmark program. Maintaining consistent high quality is a key feature of the Woolmark program, and it is important that all licensees have good in-plant quality assurance procedures in place.

In addition to its three main brands, the Woolmark Company has introduced a series of programs, which highlight selected product attributes. These programs are backed by specifications and clear product descriptions to ensure a clear and globally consistent message. Each program is also identified by a specially designed sub-brand logo; the logos are featured on a range of promotion tickets, including Cool Wool, Merino Extrafine, Pure Merino Wool, Machine Washable Wool, Total Easy Care Wool, Wool plus Lycra [201], Wool Cotton, Sportwool [159]. Further details are available from The Woolmark Company at http://www.woolmarkbrand.com.

12. Economic Aspects

Wool Production. Table 5 gives world wool production figures for the major producing countries during the period of the shearing seasons from 1995/96 to 1999/2000; Figure 15 shows percentage shares of world production. Table 5 shows actual weights of raw wool or lime wool.

Wool production has grown only slowly over recent decades, as grazing land for sheep cannot be increased indefinitely. A decline has even occurred since 1990/91 for reasons of climate, price (lucrative alternatives for farmers), and political economy (economic decline in Eastern Europe).

Sheep rearing and hence wool production exist in almost all parts of the world, although at widely varying levels of intensity. In most countries, home-produced wool is processed and consumed internally, and therefore is not available on the international market. This also applies to important wool-producing countries such as China and the countries of the former Soviet Union. The suppliers of the raw material to the international wool textile industry are those countries that produce and export amounts of wool in excess of their own requirements. These principally comprise Australia, New Zealand, Argentina, Uruguay, and South Africa.

Australia is the world's largest producer and exporter of garment wools (merinos), exporting ca. 60 % of its wool production as raw wool, and thus being the most important supplier for the wool combers of the industrialized countries. New Zealand produces mainly scoured coarse wools that are exported for use in the production of home textiles, especially carpets. Argentina, Uruguay, and South Africa have also built their own wool scouring and combing plants, and export wool either in scoured loose form or as tops (scoured and combed wool). Table 6 lists

Figure 14. Woolmarks for pure new wool and new wool with other fiber

Table 5. World wool production by the major producing countries (in 10^6 kg) (source IWTO).

	1995/96	1996/97	1997/1998	1998/99	1999/2000
Raw wool weight					
Australia	697	725	700	684	671
New Zealand	269	275	266	252	256
Former Soviet Union	238	196	164	140	135
China	277	298	255	277	290
Argentina	81	78	68	70	63
Uruguay	79	85	78	63	58
South Africa	65	59	54	58	53
Others	826	804	811	812	814
World	2532	2520	2396	2356	2340
Merino[a]	1077	1061	992	976	957
Crossbred[b]	626	620	582	557	551
Carpet wools[c]	836	839	822	823	832

[a] Merino wool covers 24.5 µm and finer. [b] Crossbred wool covers 24.6 – 32.5 µm. [c] Carpet wool covers 32.6 µm and coarser.

wool export figures, for the five most important exporting countries for the seasons 1994/95 to 1998/99.

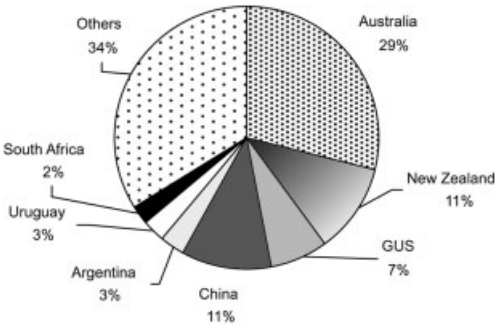

Figure 15. World wool production (1997/98)

Wool Prices. The price differences between the various wool qualities are considerable. The most important factors determining quality and therefore price are the fineness and staple length of the fiber. Not only the mean fineness and length, but also the uniformity of distribution are of great importance. Other quality and price parameters are the content of vegetable matter and pigmented fibers, tensile strength, color, elasticity, crimpiness, and many others.

Since the collapse in early 1991 of the minimum price system which imposed a lower limit on wool prices for 17 years in Australia, New Zealand, and South Africa, prices have been determined by the market forces of supply and demand. Figure 16 shows the development of the average auction price of Australian raw wool, the so-called market indicator (i.e., the weighted average of the qualities) in the sales seasons 1994/95 to 1999/2000. The market indicator during this period fluctuated between 5.1 A$/kg (March 1999) and 8.6 A$/kg (March 1995).

However, the price differences between two given qualities can vary greatly over time with supply and demand. This is illustrated in Figure 16, which also shows the price curve for 19 µm wools. The difference in price between this and the market indicator varied between 0.80 A$/kg (March 1996) and 7.85 A$/kg (September 1994). The reasons for the relatively low prices for fine wools may be an increased supply of "hunger-fine" wools after a drought, and the lack of interest in such low-quality, albeit fine wools. A high price is evidence of a strong,

Table 6. Wool exports (raw wool, scoured wool, and sliver) by the five major exporting countries (in 10^6 kg scoured wool)

	1994/95	1995/96	1996/97	1997/98	1998/99
Australia	591	534	608	556	475
New Zealand	228	196	200	192	171
South Africa	26	29	20	24	25
Argentina	46	32	34	25	24
Uruguay	15	14	17	8	8
Total	906	805	879	805	703

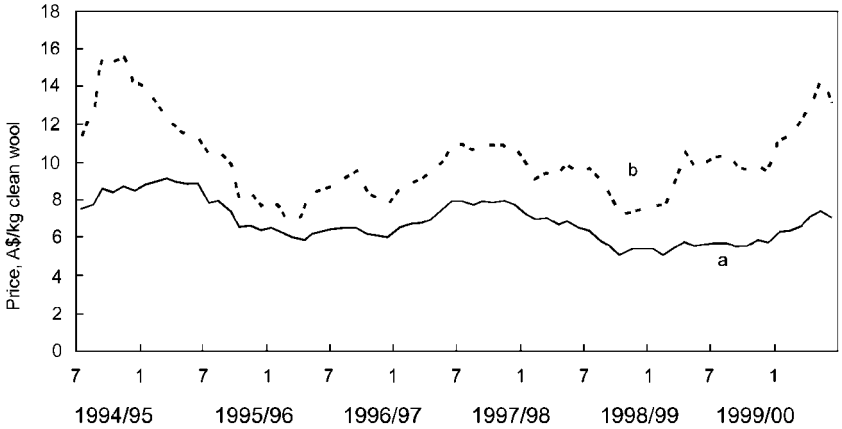

Figure 16. Price development of Australian raw wool a) Average auction price (market indicator); b) 19 µm indicator

fashion-induced demand accompanied by a restricted supply.

Table 7. Consumption of shorn wool by the wool textile industry spinning mills in 1998 (in 10^6 kg scoured wool)

	10^6 kg	%
Italy	263.1	42
Germany	70.5	18
UK	42.1	11
Spain	23.6	6
Other Western Europe	85.4	22
Central and Eastern Europe	110.6	8
Total Europe	*495.2*	*36*
Africa (incl. South Africa)	46.2	3
Middle East	138.0	10
Total +Africa and Middle East	*221.2*	*16*
North America	46.2	3
Central and South America	39.8	3
Total Americas	*86.0*	*6*
India	75.8	53
Other Indian Subcontinent	66.9	47
Total Indian Subcontinent	*142.7*	*10*
China and Hong Kong	280.1	73
Japan	61.3	16
South Korea	34.9	9
Taiwan	8.7	2
Total Far East	*385.0*	*28*
Australia, New Zealand	38.6	3
World	1372	100

Wool Consumption. Table 7 lists world wool consumption. Europe has the highest consumption of wool (36 % of the world total). This underscores the world importance of the European woolen industry. The four EU countries Italy, Germany, United Kingdom, and Spain account for 29 %, the rest of Western Europe 14 %, and Central and Eastern Europe 8 % of total world consumption for spinning mills. In second place is the Far East, with 28 % of the total world industrial demand for wool. For German wool imports, see Table 8.

The comparatively minor importance of the wool textile industry in the United States, i.e., 3 % of the world's consumption of virgin wool, is remarkable.

13. Toxicology and Occupational Health

The wool fiber itself poses no health problems. On the contrary, woolen clothing protects against heat, cold, water, and sun as well as mechanical impact [203].

Wool products that have reached the end of their useful life may be recycled via the woolen

Table 8. German imports of wool and tops (in t) (Source: Comtext, Eurostat)

Import Category	1999 (full year)	Jan – April 1999	Jan – April 2000
Wool, incl. tops and noils (clean basis)	63 302	19 874	18 826
Raw wool (clean basis)	33 800	10 705	10 924
greasy (actual)	33 969	8 873	9 417
scoured (actual)	12 434	4 671	4 520
Animal hair	1 732	819	1 059
Tops, etc.	23 944	9 136	8 378
Noils and waste	5 588	1 938	1 748

process and can be used for production of wool materials of minor quality or technical products. Wool is biodegradable. However, pretreatments or presence of contaminants such as pesticides or textile finishing agents like mothproofing agents, flame retardants or dyes have impact on the ecotoxicology of wool materials.

Prickle [204–208]. No health risks are involved in wearing properly finished and dyed wool clothing, even if it comes in direct contact with the skin. Wool incompatibility is caused neither by toxic components of wool, nor is it a wool allergy; it is a purely mechanical irritation of the skin.

Wool prickle is due to the coarser wool fibers digging into human skin with enough force to excite pain receptors. Fabrics made from wool with no more than 5 % of fibers larger than 30 µm will not be perceived as prickly by most people.

This mechanical skin irritation is not specific to wool, but can be caused by other natural and synthetic fibers if sufficient coarse fiber ends per unit area of skin are present.

Azo Dyes [209–211]. Properly dyed wool articles do not contain such azo dyes that could set free carcinogenic aromatic amines such as benzidine.

Members of the Ecological and Toxicological Association of Dyes and Organic Pigment Manufacturers (ETAD) ceased the manufacture of benzidine-based dyes in the early 1970s due to the inability to ensure their safe handling in dyeing plants. However, because of the particular economic and technical merits of these dyes, they have continued to be manufactured in many countries. Provided the dyed goods have good fastness characteristics, there is not a concern about consumer risk.

Occupational Health. For the ecotoxicological potential of textile dyes, see → Textile Dyeing, Chap. 14.7..

In the processing of wool, problems are caused by exposure of workers to noise and to respirable and alveolar dust [212, 213, 215].

There is a potential human health hazard to shearers and farm workers handling the treated sheep and wools. Users of lanolin products are at lower risk, especially if the commercially available "reduced pesticide" grades of lanolin are used where there is human skin contact. All the existing ectoparasiticide products are evaluated for potential hazard to shearers and farm workers for incorporation into the protocol of the Australian Occupational Health and Safety Commission. As measure of the "safe exposure level" the acceptable daily intake is compared with the potential exposure calculated from the amount of active ingredient present and the amount of wool grease covering the shearer [214]. However, the use of toxic agents for sheep dips, etc., is today discouraged and nonharmful alternatives are recommended. Also operational procedures are now strictly laid down to control and minimize any risks to farm workers, etc.

Acknowledgement. The authors thank Hartwig Höcker (Director of the Deutsches Wollforschungsinstitut an der RWTH Aachen) for checking the manuscript and Klaus Wolf, Hans-Georg Hebecker, Peter Duffield, and Kevin Russel for helpful suggestions.

14. References

1. H. Glafey, D. Krüger, G. Ulrich: "Technologie der Wolle – Chemische Technologie und mechanische Hilfsmittel für die Veredlung von Wolle," in *Technologie der Textilfasern,* vol. VIII, part 3 B, Springer Verlag, Berlin 1938.

2. E. R. Kaswell: *Textile Fibers, Yarns, and Fabrics – A Comparative Survey of their Behavior with Special Reference to Wool,* Reinhold Publ. Co., New York 1953.
3. R. W. Moncrieff: *Wool Shrinkage and its Prevention,* National Trade Press, London 1953.
4. W. H. Ward, H. P. Lundgren, *Adv. Protein Chem.* **IX** (1954) 243 – 297.
5. P. Alexander, R. F. Hudson: *Wool. Its Chemistry and Physics,* Chapman & Hall, London 1954.
6. M. Harris: *Handbook of Textile Fibers,* Wiley-Interscience, New York 1954.
7. W. G. Crewther (ed.), *Proc. Intern. Wool Text. Res. Conf., Australia 1955,* **vol. A – F,** Commonwealth Scientific and Industrial Research Organization, Melbourne 1956.
8. C. E. M. Jones (ed.), 2nd Quinquennial Wool Text. Res. Conf., Harrogate 1960; *J. Text. Inst.* **51** (1960) T 489 – T 1552.
9. E. H. Mercer: *Keratin und Keratinization,* Pergamon, Oxford 1961.
10. W. J. Onions: *Wool – An Introduction to its Properties, Varieties, Uses and Production,* E. Benn, London 1962.
11. P. Alexander, R. F. Hudson in Ch. Earland (ed.): *Wool – Its Chemistry and Physics,* 2nd ed., Chapman & Hall, London 1963.
12. H. Doehner, H. Reumuth (eds.): *Wollkunde,* 2nd ed., P. Parey, Berlin – Hamburg 1964.
13. W. G. Crewther, R. D. B. Fraser, F. G. Lennox, H. Lindley, *Adv. Protein Chem.* **20** (1965) 191 – 346.
14. R. Delerive, F. Maillard (eds.): *Proc. 3rd Int. Wool Text. Res. Conf. Cirtel Paris 1965,* **Sections I – IV,** Institut Textile de France.
15. J. G. Cook: "Natural Fibers," *Handbook of Textile Fibers,* **vol. 1,** Merrow Publ. Co., Watford, Herts 1968.
16. W. G. Crewther (ed.): *Symp. Fibrous Proteins 1967* Butterworths Australia 1968.
17. E. G. Bendit, M. Feughelman in H. F. Mark, N. G. Gaylord, N. M. Bikales (eds.): *Encylopedia of Polymer Science and Technology,* **vol. 8,** Wiley-Interscience, New York 1968, pp. 1 – 44.
18. H. Behrens, H. Doehner, R. Scheelje, R. Wassmuth: *Lehrbuch der Schafzucht,* P. Parey, Hamburg – Berlin 1969.
19. W. von Bergen: *Wool Handbook,* 3rd ed., vol. I, Wiley-Intersience, New York – London 1969.
20. H. Spiebeg (ed.): *British Wool Manual,* 2nd ed., Columbine Press, Buxton 1969.
21. K. R. Makinson in H. F. Mark, N. G. Gaylord, N. M. Bikales (eds.): *Encyclopedia of Polymer Science and Technology,* **vol. 15,** Wiley-Interscience, New York 1971, pp. 41 – 79.
22. L. Rebenfeld, H. H. Ward (eds.): "Proc. 4th Int. Wool Text. Res. Conf. Berkely 1970," Parts I, II, *Applied Polymer Symposia,* no. 18, Interscience Publishers, New York 1971.
23. H. Rath: *Lehrbuch der Textilchemie einschließlich der textilchemischen Technologie,* 3rd ed., Springer Verlag, Berlin 1972.
24. R. D. B. Fraser, T. P. MacRae, G. E. Rogers: *Keratins – Their Composition, Structure and Biosynthesis,* Charles C. Thomas, Springfield, Ill. 1972.
25. J. H. Bradbury, *Adv. Protein Chem.* **27** (1973) 111 –211.
26. K. Ziegler (ed.): *Proc. 5th Int. Wool Text. Res. Conf. Aachen 1975,* **vols I – IV**, Deutsches Wollforschungsinstitut an der RWTH Aachen 1976.
27. R. S. Asquith (ed.): *Chemistry of Natural Protein Fibers,* Plenum Press, New York – London 1977.
28. H. E. Schiecke: *Wolle als textiler Rohstoff,* Fachverlag Schiele & Schön, Berlin 1979.
29. K. R. Makinson: "Shrinkproofing of Wool," in *Fiber Science Series,* vol. 8, Marcel Dekker, New York – Basel 1979.
30. D. P. Veldsman (ed.): Proc. 6th Quinquennial Int. Wool Text. Res. Conf. Pretoria 1980, **vols I – IV**.
31. J. A. Mclaren, B. Milligan: *Wool Science – The Chemical Reactivity of the Wool Fibres,* Science Press, Marrickville, Australia 1981.
32. J. Leeder: *Wool – Nature's Wonder Fibre,* Australasian Textiles Publishers, Belmont/Victoria/Australia 1984.
33. E. Wang, D. Fischman, R. K. H. Liem, T.-T. Sun (eds.): *Intermediate Filaments,* Annals of the New York Academy of Sciences, vol. 455 New York 1985.
34. M. Sakamoto (ed.): *Proc. 7th Int. Wool Text. Res. Conf. Tokyo 1985,* **vols I – IV,** Soc. Fiber Science and Technol. Japan.
35. I. Bearpak, F. W. Marriott, J. Park: *A Practical Introduction to the Dyeing and Finishing of Wool Fabrics,* Society of Dyers and Colourists, Bradford 1986.
36. M. Feughelman in H. F. Mark, N. M. Bikales, C. G. Overberger, G. Menges, J. I. Kroschwitz (eds.): *Encyclopedia of Polymer Science and*

Engineering, **vol. 8,** Wiley-Interscience, New York 1987, pp. 566 – 600.
37. G. Ebner, D. Schelz: *Textilfärberei und Farbstoffe,* Springer Verlag, Berlin 1989.
38. G. E. Rogers, P. J. Reis, K. A. Ward, R. C. Marshall (eds.): *The Biology of Wool and Hair,* Chapman & Hall, London – New York 1989.
39. M. Ried: *Chemie im Kleiderschrank,* Rowohlt Verlag, Reinbeck 1989.
40. M. Peter, H. K. Rouette: *Grundlagen der Textilveredlung, Handbuch der Technologie, Verfahren und Maschinen,* 13th ed., Deutscher Fachverlag, Frankfurt 1989.
41. R. D. Goldman, P. M. Steinert (eds.): *Cellular and Molecular Biology of Intermediate Filaments,* Plenum Press, New York – London 1990.
42. G. H. Crawshaw (ed.): *Proc. 8th Int. Wool Text. Res. Conf. Christchurch New Zealand 1990,* **vols. I – IV,** Wool Res. Org. New Zealand.
43. R. C. Marshall, D. F. G. Orwin, J. M. Gillespie, *Electron Microsc. Rev.* **4** (1991) 47 – 83.
44. H. Zahn, B. Wulfhorst, H. Külter, *Chemiefasern/Textilind.* **41** (1991) 521 – 552.
45. D. M. Lewis (ed.): *Wool Dyeing,* Society of Dyers and Colourists, Bradford 1992.
46. H. Zahn, J. Knott in EEC (ed.): "Chemical Methods for Characterization of Wool at Different Stages of Processing," COMETT-EUROTEX, Guimaraes, Portugal 1992.
47. F.-J. Wortmann: *Thermo- and hydroplastische Eigenschaften von Wollfasern,* Forschungsberichte des Landes Nordrhein-Westfalen, no. 3245, Westdeutscher Verlag 1992.
48. W. Bobeth (ed.): *Textile Faserstoffe – Beschaffenheit und Eigenschaften,* Springer Verlag, Berlin 1993.
49. R. Kellner, F. Lottspeich, H. E. Meyer (eds.): *Microcharacterization of Proteins,* VCH Verlagsgesellschaft, Weinheim 1994.
50. M. Bona: *An Introduction to Wool Fabric Finishing,* Verlag Textilia, Biella, Italy 1994.
51. P. H. Greaves, B. P. Saville: Microscopy of Textile Fibres, Microscopy Handbooks 32, Royal Microscopical Society, BIOS Scientific Publ., Oxford 1995.
52. M. Bona, G. Grupallo: *Proc. 9th Int. Wool Text. Res. Conf. Biella Italy 1995,* **vols. I – V,** Citta degli Studi Biella and Int. Wool Sec.
53. J. P. Laker, F.-J. Wortmann (eds.): *Metrology and Identification of Speciality Animal Fibres,* European Fine Fibre Network, occasional publication no. 4, Maculay Land Use Research Institute, Aberdeen, 1996.
54. ISO 6938-1984: Textiles-Natural Fibres-Generic Names and Definition, International Organization of Standardization, 1984.
55. P.-A. Koch, *Text. Rundsch.* **19** (1964) 189 – 203.
56. W. F. Leggett: *The Story of Wool,* Chem. Publ. Co., Brooklyn N.Y. 1947.
57. N. Hyde, *National Geographic* **173** (1988) 552 – 591.
58. R. E. Chapman in J. Bereiter-Hahn, A. G. Matoltsy, K. S. Richards (eds.): *Biology of the Integument 2,* Springer Verlag, Berlin 1986, p. 298.
59. D. E. Rivett, *Wool Sci. Rev.* **67** (1991) 1 – 25.
60. C. A. Anderson, J. D. Leeder, *Text. Res. J.* **55** (1985) 416 – 421.
61. H. E. Crabtree, P. Nicholls, E. V. Truter, *Proc. Anal. Div. Chem. Soc.* **16** (1979) 235.
62. A. Schwan, J. Herrling, H. Zahn, *Colloid Polym. Sci.* **264** (1986) 171 – 175.
63. J. Kalbe et al., *Biol. Chem. Hoppe – Seyler* **369** (1988) 413 – 416.
64. P. F. Hamlyn, G. Nelson, B. J. McCarthy, *J. Text. Inst.* **83** (1992) 97 – 103.
65. D. J. Evans, J. D. Leeder, J. A. Rippon, D. E. Rivett in [34] vol. I, pp. 135 – 142.
66. P. W. Wertz, D. T. Downing, *Lipids* **23** (1988) 878 – 881.
67. A. Körner, H. Höcker, D. E. Rivett, *Fresenius Z. Anal. Chem.* **344** (1992) 501 – 509.
68. U. Altenhofen, H. Zahn, *ACS Symp. Ser.* **49** (1977) 162 – 175.
69. M. F. Perutz, *Faraday Discuss. Chem. Soc.* **93** (1992) 1 – 11; *Philos. Trans. R. Soc. London A* **345** (1993) 105 – 112.
70. J. B. Speakman, M. Hirst, *Nature (London)* **128** (1931) 1073 – 1074.
71. R. S. Asquith et al., *Biochim. Biophys. Acta* **207** (1970) 342.
72. B. Milligan, L. A. Holt, J. B. Caldwell in [22], part I, pp. 113 – 125.
73. S. K. Burley, G. A. Petsko, *Adv. Prot. Chem.* **39** (1988) 125 – 189.
74. H. Beyer, U. Schenk, *J. Chromatogr.* **39** (1969) 491 –495.
75. D. R. Goddard, L. Michaelis, *J. Biol. Chem.* **106** (1934) 605 – 614; **112** (1935) 361 – 371.
76. W. G. Crewther, R. D. B. Fraser, F. G. Lennox, H. Lindley, *Adv. Protein Chem.* **20** (1965) 191 – 346.
77. J. M. Gillespie: pp. 95 – 128 in [41].
78. L. M. Dowling, L. G. Sparrow, *TIBS* **16** (1991) 115 – 117.

79. R. C. Marshall, *Text. Res. J.* **51** (1981) 106 – 108. G. E. Rogers, E. S. Kuczek, P. J. Mackinnon, R. B. Presland, M. J. Fietz in [38] pp. 69 – 85.
80. B. R. Herbert, A. L. P. Chapman, D. A. Rankin, *Electrophoresis* **17** (1996) 239 – 243.
81. H. Thomas, A. Conrads, K. H. Phan, M. van de Löcht, H. Zahn, *Int. J. Biol. Macromol.* **8** (1986) 258 – 264.
82. M. van de Löcht, *Melliand Textilber.* **68** (1987) 780 –786.
83. A. Conrads, H. Thomas, K.-H. Phan, H. Zahn, H. Höcker, *Naturwissenschaften* **75** (1988) 100 – 101.
84. W. G. Crewther, L. M. Dowling, P. M. Steinert, D. A. D. Parry, *Int. J. Biol. Macromol.* **5** (1983) 267 – 274.
85. J. A. Swift in [27] pp. 81 – 146.
86. M. G. Dobb et al., *J. Text. Inst.* **52** (1961) T 153 –170.
87. K.-H. Phan, H. Thomas, E. Heine in [52], vol. II, pp. 19 – 30.
88. G. E. Rogers, *Ann. N.Y. Acad. Sci.* **83** (1959) 378 –399. G. E. Rogers, *J. Ultrastruct. Res.* **2** (1959 b) 309 –330.
89. R. Eichner, P. Rew, A. Engel, U. Aebi, *Ann. N.Y. Acad. Sci.* **455** (1985) 381 – 402.
90. W. T. Astbury, A. Street, *Phil. Trans. R. London Soc. A* **230** (1931) 75. W. T. Astbury, H. J. Woods, *Phil. Trans. R. London Soc. A* **232** (1933) 333 – 394.
91. L. Pauling, R. B. Corey, H. R. Branson, *Proc. Nat. Acad. Sci. USA* **37** (1951) 205 – 211.
92. L. Pauling, R. B. Corey, *Proc. Nat. Acad. Sci. USA* **39** (1953) 253.
93. F.-J. Wortmann, H. Deutz, *J. Appl. Polym. Sci.* **48** (1993) 137 – 150.
94. F.-J. Wortmann, H. Zahn in: "Domke Festschrift," Lehrstuhl für konstruktive Gestaltung der TH Aachen, Aachen 1982, pp. 157 – 161.
95. M. Feughelman, *Text. Res. J.* **29** (1959) 223 – 228.
96. J. Cao, *J. Mol. Struct.* **553** (2000) 101 – 107.
97. L. Kreplak, J. Doucet, F. Briki, *Biopolymers* **58** (2001) 5269 – 533.
98. J. B. Speakman, C. A. Cooper, *J. Text. Inst.* **27** (1936) T 183 – T 196.
99. E. Elöd, E. Silva, *Z. Phys. Chem. Abt. A* **137** (1928) 141.
100. F.-J. Wortmann, B. J. Rigby, D. G. Phillips, *Text. Res. J.* **54** (1984) 6 – 8.
101. T. G. Fox, *Bull. Am. Phys. Soc.* **1** (1956) 123.
102. H. Zahn in [52], vol. I, pp. 1 – 16.
103. H. Baumann, L. Möchel, *Text. Praxis Int.* **29** (1974) 507 – 510.
104. H. Baumann, Schriftenreihe des Deutschen Wollforschungsinstitutes 85 (1981), 216 – 246.
105. H. Meichelbeck, H. Knittel, *Fette, Seifen, Anstrichm.* **73** (1971) 25 – 29.
106. D. Goddinger, K. Schäfer, H. Höcker, *Wool Technol. Sheep Breed* **42** (1994) 83 – 89.
107. D. Goddinger, K. Schäfer, H. Höcker in [52], vol. II, pp. 92 – 108.
108. G. J. Smith, *J. Photochem. Photobiol. Bull.* **22** (1994) 145 – 149.
109. H. Zahn, J. Knott, *Nova Textil* **13** (1989) 42 – 46.
110. F. G. France, I. L. Weatherall in [52], vol. II, pp. 100 – 108.
111. G. J. Smith, K. R. Markham, W. H. Melhuish, *Photochem. Photobiol.* **60** (1994) 196 – 198.
112. K. Schäfer, *J. Soc. Dyers Colour.* **107** (1991) 206 – 211.
113. C. M. Pande, *J. Soc. Cosmet. Chem.* **45** (1994) 257 – 268.
114. I. C. Watt, *J. Macromol. Sci.-Rev. Macromol. Chem.* **C18** (1980) 169 – 245.
115. F.-J. Wortmann, P. Augustin, C. Popescu, *J. Appl. Polym. Sci.* **79** (2001) 1054 – 1061.
116. F.-J. Wortmann, H. Zahn, *Text. Res. J.* **64** (1994) 737 – 743.
117. K. Schäfer, I. Müllejans, H. Höcker, *Melliand Textilber.* **78** (1997) no. 7/8, 506 – 508, E 108 – E 110.
118. K. Schäfer, I. Müllejans, J. Föhles, H. Höcker, *Melliand Textilber.* **78** (1997) no. 10, 727 – 732, E 164 – E 167.
119. H. Höcker in [52], vol. III, 319 – 324.
120. E. Heine, H. Höcker, *Rev. Progr. Color. Relat. Top.* **25** (1995) 57 – 63.
121. A. Bereck, *Rev. Progr. Color. Relat. Top.* **24** (1994) 17 – 25.
122. A. C. Welham, *J. Soc. Dyers Colour.* **102** (1986) 126 – 131.
123. W. Mosimann, K. Schaffner, *Textilveredlung* **14** (1979) 12 – 15.
124. P. A. Duffield, R. R. D. Holt, *Textilveredlung* **24** (1989) 40 – 45.
125. H. Baumann, *J. Soc. Dyers Colour.* **90** (1974) 125 –129.
126. N. Werkes, *Melliand Textilber.* **70** (1989) 52 – 63.
127. D. M. Lewis, *Melliand Textilber.* **67** (1986) 717 –723; *Proc. 8th Int. Wool Text. Res. Conf.,* **vol. IV,** Christchurch, New Zealand 1990, pp. 1 – 49.
128. K. J. Dodd, C. M. Carr, M. Baird, *J. Soc. Dyers Colour.* **110** (1994) 190 – 194.

129. B. Gullbrandson, *Text. Res. J.* **28** (1958) 965 – 968.
130. A. G. de Boos: "The Conflicting Requirements of Wool Setting for the Finisher and Tailor," Report No. G 61, CSIRO Div. Text. Ind., Belmont/Victoria 1987.
131. A. Robson, M. J. Williams, J. M. Woodhouse, *J. Text. Inst.* **60** (1969) 140 – 151.
132. H. Meichelbeck: "Die Rolle der SH-Gruppen bei der Quervernetzung des Keratins durch Thioätherbildung," Dissertation TH Aachen 1963.
133. H. Zahn in [52], vol. I, pp. 1 – 16.
134. H. K. Rouette, *DWI Reports* **117** (1996) 93 – 112.
135. J. B. Speakman, *J. Soc. Dyers Colour.* **52** (1936) 335 – 346.
136. P. G. Cookson, K. W. Fincher, P. R. Brady, *J. Soc. Dyers Colour.* **107** (1991) 135 – 138.
137. S. M. Smith et al. in [52], vol. III, pp. 113 – 121.
138. I. Hammers-Page in W. Loy (ed): *Taschenbuch für die Textilindustrie 1994,* Fachverlag Schiele und Schön, Berlin, pp. 426 – 436.
139. T. Shaw, M. A. White in M. Lewin, S. B. Sello (eds.): "Chemical Processing of Fibers and Fabrics," *Handbook of Fibre Science and Technology,* **vol. 2, part 2,** Marcel Dekker, New York 1984.
140. R. J. Mayfield, *Text. Prog.* **11** (1982) 1 – 10.
141. K. Wolf, I. Hammers, U. Altenhofen, *Melliand Textilber.* **66** (1985) 596 – 600.
142. L. Benisek, *J. Text. Inst.* **65** (1974) 102 – 108.
143. W. Mosimann et al. in [42], vol. IV, pp. 239 – 249.
144. G. Reinert, *Textilveredlung* **26** (1991) 86 – 92.
145. G. Reinert, E. Schmidt, R. Hilfiker, *Melliand Textilber.* **75** (1994) 606 – 614.
146. G. V. Barker, F. Davin, *Proc. Int. Wool Text. Res. Conf. 8th,* Christchurch, **vol. 3** (1990) 91 – 106.
147. P. Hopkins, *DWI Reports* **119** (1997) 128 – 137.
148. P. Hopkins, *Textile Magazine* **26** (1997) no. 1, 7 – 8.
149. *Wool Fact File 1998/99,* published by Woolmark (Europe) Ltd., Ilkley, UK.
150. E. Heine, A. Ruers, H. Höcker, *DWI Reports* **122** (1999) 466 – 470.
151. E. Heine, A. Ruers, H. Höcker, *DWI Reports* **123** (2000) 475 – 479.
152. B. M. Müller, *Rev. Prog. Colouration* **22** (1992) 14 – 21.
153. H. Thomas, F. Kümmeler, H.-B. Müller, E. Prinz, U. Rott, J. Salge, M. Wolf, T.-J. Zahn, *DWI Reports* **122** (1999) 365 – 370.
154. H. Thomas, M. Hedler, T. Merten, V. Monser, H. Höcker, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) PL-3.
155. B. Jansen, F. Kümmeler, H. Thomas, H.-B. Müller, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) PL-6.
156. H. Höcker, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) PL-7.
157. Anonym., *International Dyer* **185** (2000) no. 4, 21.
158. Anonym., *Wool Record* **157** (1998) (Issue 3644) 21.
159. B. O. Bateup, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) KNL-5.
160. K. Hannemann, H. Flensberg, *DWI Reports* **117** (1996) 287 – 293.
161. K. Hannemann, *Wool Record* **159** (2000) no. 6, 48 – 51.
162. M. Gießmann, K. Schäfer, H. Höcker, *Textilveredlung* **34** (1999) no. 11/12, 12 – 19.
163. A. Giehl, K. Schäfer, H. Höcker, J.-P. Mullender, J. Knott, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) DY-2.
164. G. Carnaby, *Proc. Niches in the World of Textiles, 77th World Conf. of the Textile Inst.,* Tampere, SF, May 22 – 24, 1996, vol. 1 (1997).
165. [32], Chap. 7.
166. G. Mack, *Wool Record* (1995) Dec., 5.
167. R. Hoffmann, G. Timmer, K. Becker in [52], vol. I, pp. 66 – 77.
168. P. A. Duffield, R. R. D. Holt, J. R. Smith, *Melliand Textilber.* **72** (1991) 938 – 942.
169. H. Thomas, K.-H. Phan, H. Höcker Report 17, IWTO Harrogate Meeting, 1995.
170. J. D. Verity, *Wool Record* (1995) Mar., 25.
171. G. Horstmann, *J. Soc. Dyers Colour.* **111** (1995) 182 – 184.
172. H. G. Hebecker: Textil und Umwelt, Textilverband Schweiz, Zürich 1995, pp. 24 – 25.
173. H. J. Henning, G. Mazingue, *Text. Prax. Int.* **35** (1980) 51 – 54, 179 – 185, 298 – 304.
174. G. J. J. B. de Groot, *J. Text. Inst.* **86** (1995) 33 – 44.
175. S. Kawabata, M. Niwa, *J. Text. Inst.* **80** (1989) 19 –50.
176. S. Kawabata, K. Ito, M. Niwa, *J. Text. Inst.* **83** (1992) 361 – 374.
177. C. F. Allan, A. G. de Boos, N. G. Ly, *Melliand Textilber.* **81** (1990) 614 – 618.
178. F.-J. Wortmann, G. Wortmann, H. Zahn, *Text. Res. J.* **65** (1995) 669 – 675.
179. H. Zahn, *Melliand Textilber.* **50** (1969) 807 – 813.

180. International Wool Secretariat (ed.): IWTO Specifications, International Wool Secretariat, Raw Wool Services Department, Valley Drive, Ilkley, Yorkshire, LS29 8PB, England.
181. R. C. Marshall, H. Zahn, G. Blankenburg, *Text. Res. J.* **54** (1984) 126 – 128.
182. H. Zahn, I. Souren, U. Altenhofen, *Melliand Textilber.* **65** (1984) 467 – 471.
183. C. Wilrich, G. Wortmann, F.-J. Wortmann, H. Höcker in [52], vol. II, pp. 289 – 295.
184. P. Seiler, C. Wilrich, G. Wortmann, F.-J. Wortmann, *DWI Reports* **117** (1996) 637 – 648.
185. W. Gocht, *DWI Reports* **117** (1996) 399 – 400.
186. J. D. Leeder in [32], pp. 24 – 31.
187. L. Benisek, P. R. Harnett, M. J. Palin, *Melliand Textilber.* **68** (1987) 878 – 888.
188. M. Ried in [39], pp. 197 – 198.
189. A. Pierlot: "Wool" in J. C. Salomone (ed.): *Polymeric Materials Encyclopedia,* CRC Press, Boca Raton, Fl., in press.
190. G. Fisher, *Tech. Text. Int.* 1993, Feb., 8 – 12.
191. H. Zahn, G. Blankenburg, *Deutsche Schafzucht* **85** (1993) 31 – 33.
192. G. Blankenburg, I. Hammers-Page, *DWI Reports* **114** (1995) 61 – 72.
193. E. Claßen, G. Wortmann, R. Augustin, N. Effertz, F.-J. Wortmann, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) OT-P5.
194. P. Offermann, *DWI Reports* **121** (1998) 247 – 257.
195. H. L. Fraenkel-Conrat, H. S. Olcott, A. Mohammed, *Arch. Biochem.* **24** (1949) 278.
196. S. M. Causer, R. C. McMillan, *Australian Textiles* **14** (1994) 48.
197. G. Wortmann, R. Sweredjuk, G. Zwiener, F. Doppelmayer, F.-J. Wortmann, *DWI Reports* **124** (2001) 378 – 386.
198. K. Schäfer, H. Höcker, *Melliand Textilber.* **77** (1996) 402 – 406.
199. C. Jacques, J. P. Bruggeman, G. Vilarem, *Proc. Int. Wool Text. Res. Conf. 10th,* Aachen, (2000) OT-P7.
200. *Bekleidung und Wäsche* **18** (1989) 14 – 18.
201. A. Kloeden, *Wool Record* (1995) Nov., 6, 51.
202. H. Menkens, *Wool Record* (1995) Dec., 45.
203. G. Reinert, R. Hilfiker, E. Schmidt, F. Fuso, *DWI Reports* **117** (1996) 259 – 273; *Text. Res. J.* **66** (1996) 61 – 70.
204. G. R. S. Naylor, C. J. Veitch, R. J. Mayfield, *Text. Res. J.* **62** (1992) 487 – 493.
205. G. Wortmann, A. Körner, *DWI Reports* **115** (1995) 1 – 135.
206. G. Chalmers, *Wool Record* (1993), July, 41.
207. F. Klaschka, *Melliand Textilber.* **75** (1994) 193 – 202.
208. T. L. Diepgen, B. Salzer, A. Tepe, O. P. Hornstein, *Melliand Textilber.* **76** (1995) 1116 – 1121.
209. E. A. Clarke, D. Steinle, *Rev. Prog. Color. Relat. Top.* **25** (1995) 1 – 5.
210. J. Grütze, F. Hoffmann, U. Sewekow, A. Westerkamp, *Melliand Textilber.* **76** (1995) 875 – 883.
211. R. A. Moll, *Melliand Textilber.* **76** (1995) 993 – 995.
212. W. Beckmann, *ITB Veredlung* **1** (1990) 50 – 59.
213. R. G. Love, M. Muirhead, H. P. R. Collins, C. A. Soutar, *Br. J. Ind. Med.* **48** (1991) 221 – 228.
214. I. M. Russell in [42], vol. III, pp. 165 – 174.
215. V. Schmitz, K. Schäfer, H. Höcker, *Melliand Textilber.* **80** (1999) no. 7/8, 596 – 600.

Xanthan Gum → Polysaccharides

Xanthates

KATHRIN-MARIA ROY, Langenfeld, Federal Republic of Germany

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Production** 4
4.1. **General** 4
4.2. **Production Processes** 5
5. **Quality Specifications** 5
6. **Storage and Transportation** 5
7. **Analysis** 6
8. **Uses** 6
9. **Economic Aspects** 7
10. **Toxicology and Occupational Health** . 7
11. **References** 7

The author would like to thank Hoechst AG for undertaking the literature research for this article

1. Introduction

According to the IUPAC nomenclature, xanthates or xanthogenates (older name in the continental literature) are described as derivatives of *O*-alkyl- or *O*-aryldithiocarbonates (xanthic acids; **1**). The term xanthates refers both to the salts of the general formula **2** and to the corresponding *O*,*S*-diesters **3**.

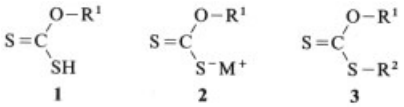

The name xanthogenates (derived from the Greek *xanthos*: yellow) was introduced by ZEISE, who discovered this class of compounds in 1822 [8].

2. Physical Properties

Xanthic Acids. The free xanthic acids (**1**) are oily liquids, that are stable only in the pure state. They are strong acids, the dissociation constants of the C_1-C_5 esters in water are in the range of $1.4-3.4\times10^{-2}$ [9]. Xanthic acids are readily soluble in nonpolar solvents and sparingly soluble in water. Traces of polar substances cause their decomposition into carbon disulfide and the corresponding alcohol. The alcohol formed accelerates the decomposition autocatalytically.

Alkali-Metal Xanthates. In contrast to the free acids, alkali-metal xanthates are relatively stable, yellowish solids of unpleasant odor. The potassium xanthates crystallize without water, whereas the sodium salts form dihydrates upon exposure to atmospheric moisture. The melting points of some industrially important alkali-metal xanthates are given in Table 1. Since the potassium salts melt with decomposition, very different melting points are found in the literature [10].

Alkali-metal xanthates are soluble in water and polar organic solvents such as alcohols, lower ketones, pyridine, and acetonitrile, but sparingly soluble in nonpolar solvents, e.g., ether or petroleum ether. The solubilities of some alkali-metal xanthates in water and alcohols are given in Table 2.

Heavy-Metal Xanthates. The heavy-metal xanthates have lower *mp*s than the corresponding alkali-metal salts and are better soluble in organic solvents, e.g., chloroform, tetrahydrofuran, and carbon disulfide, than in strongly polar ones [11]. The solubilities of heavy-metal *O*-ethyl xanthates in water increase in the order Au^+, Hg^{2+}, Cu^+, Ag^+, Pb^{2+}, Cd^{2+}, Co^{2+}, Ni^{2+}, Zn^{2+}, Fe^{2+} [12].

Aqueous xanthate solutions are relatively stable between pH 8 and 13 with a maximum of stability at pH $\approx$ 10 [13], but they rapidly decompose at pH $<$ 6 and at pH $>$ 13 (see Chap. 3).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_423

Table 1. Melting points of some industrially important xanthates, R – O – CS – S^-M^+

M	R	CAS registry no.	*mp*, °C
K	CH_3	[*2667-20-1*]	182 – 186
	C_2H_5	[*140-89-6*]	225 – 226
	n-C_3H_7	[*2720-67-4*]	233 – 239
	iso-C_3H_7	[*140-92-1*]	278 – 282
	n-C_4H_9	[*871-58-9*]	255 – 256
	iso-C_4H_9	[*13001-46-2*]	260 – 270
Na	C_2H_5	[*140-90-9*]	69 – 69.5
	iso-C_3H_7	[*140-93-2*]	124
	n-C_4H_9	[*141-33-3*]	71.5 – 75
	$(CH_2)_2CH(CH_3)_2$	[*2540-36-5*]	85 – 85.5
	$(CH_2)_9CH_3$		99 – 99.5
	$(CH_2)_{11}CH_3$		106 – 107.2

Table 2. Solubilities of alkali-metal xanthates, R – O – CS – S^-M^+, in water and alcohols

M	R	Solubility, g/100 g solution			
		H_2O (0 °C)	H_2O (35 °C)	R – OH (0 °C)	R – OH (35 °C)
K	*n*-C_3H_7	43.0	58.0	1.9	8.9
	iso-C_3H_7	16.6	37.2		2.0
	n-C_4H_9	32.4	47.9		36.5
	iso-C_4H_9	10.7	47.7	1.6	6.2
	$(CH_2)_2CH(CH_3)_2$	28.4	53.3	2.0	6.5
Na	*n*-C_3H_7	17.6	43.3	10.2	22.5
	iso-C_3H_7	12.1	37.9		19.0
	n-C_4H_9	20.0	76.2		39.2
	iso-C_4H_9	11.2	33.4	1.2	20.5
	$(CH_2)_2CH(CH_3)_2$	24.7	43.5	10.9	15.5

Their stability tends to increase with increasing molecular mass and branching of the alkyl group.

The electron absorption spectra of aqueous xanthate solutions show characteristic bands at 301 and 226 nm with extinction coefficients ε of $\approx$17 500 and 8750, respectively [14]. IR spectra of various xanthates have been published, e.g. [15].

3. Chemical Properties

In the presence of polar substances xanthic acids readily *decompose* into carbon disulfide and the corresponding alcohol, which accelerates the decomposition autocatalytically [16].

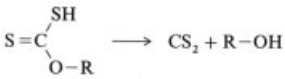

This reaction is also the second step in the decomposition of xanthates in aqueous acidic solution [17]:

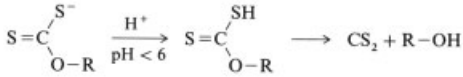

In neutral and alkaline solutions xanthates *hydrolyze* according to the following equation. The trithiocarbonate formed can decompose further into carbon disulfide and hydrogensulfide.

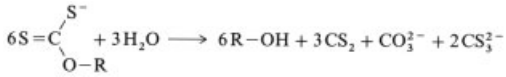

Xanthates are susceptible to *oxidation*, resulting in the formation of dixanthogens, e.g.:

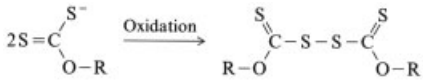

Oxidation can be accomplished by various oxidants [18], e.g., oxygen, chlorine, bromine, iodine, hydrogen peroxide, nitrous acid, and chloramine T. In addition, oxidative coupling is achieved by anodic oxidation [19]. Diisopropyl dixanthogen, used as polymerization regulator,

is produced by oxidation of potassium isopropyl xanthate with persulfate [20].

Condensation of alkali-metal xanthates with sulfur dichlorides gives higher xanthate sulfanes [21].

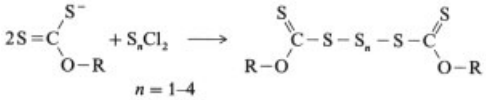

Alkylation of alkali-metal xanthates with alkyl or allyl halides yields the corresponding xanthate esters (*O*,*S*-diesters of dithiocarbonic acid).

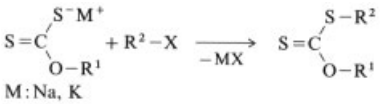

This widely employed synthetic procedure [22] is usually carried out in organic solvents such as ethanol, ethers, ketones, or benzol under reflux, but reaction in DMSO is also possible [23]. In aqueous solution, alkylation is often conducted under phase transfer conditions [24] or in the presence of an emulsifier [25]. Especially in industrial processes, alkylation is carried out as a one-pot reaction without isolation of the corresponding alkali-metal xanthate [26]. Xanthate esters have also been obtained by alkylation of polymer-supported ammonium xanthates [27]. Treatment of lithium enolates with carbon disulfide and subsequent alkylation gives α,β-unsaturated esters [28].

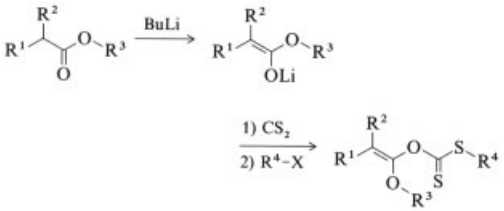

Reaction of alkali-metal xanthates with geminal dihalides affords the corresponding bisxanthates [29].

In contrast to alkylation, *arylation* of alkali-metal xanthates is possible only with aryldiazonium salts [30]. In the Leuckart reaction, the arylesters thus obtained are converted either to thiophenols by hydrolysis [31], or thermally to thioethers.

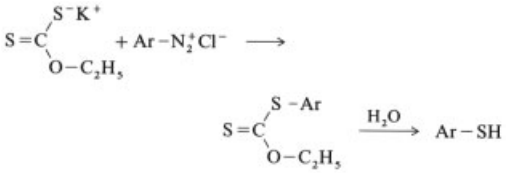

Mild *pyrolysis* of *S*-methylxanthates results in elimination of carbonyl sulfide and methanethiol to give the corresponding alkenes without rearrangement (Chugaev reaction) [22,32]. This reaction is widely used to convert alcohols into alkenes via their *S*-methyl xanthates.

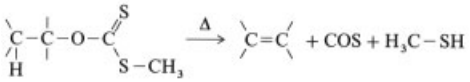

Applications in peptide chemistry [33] and in the terpene field have also been described [34].

Treatment of xanthate esters with tributyltin hydride gives *deoxygenated products* by radical reduction (Barton – McCombie deoxygenation) [35].

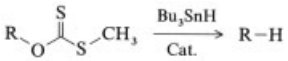

This method is efficiently applied in the deoxygenation of secondary alcohols [36] and of secondary hydroxyl groups of carbohydrates and nucleosides [37].

Other reagents effecting radical deoxygenation of xanthates include various silanes [37,38], dialkyl phosphites [39] and hypophosphorus acid or its salts [40]. Analogously, deoxygenations of primary [41] and tertiary alcohols [42] have been achieved via the corresponding xanthates.

Application of this deoxygenation procedure to bisxanthate esters derived from vicinal diols affords the corresponding alkenes [43].

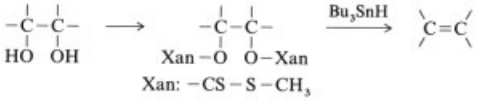

Several interesting *radical cyclizations* involving xanthates have been described. Thus, five- to eight-membered cyclic ethers [44], five- and six-membered *N*-heterocycles [45], and thionolactones [46] have been synthesized by this procedure.

Xanthate esters undergo thermally induced *rearrangement* to give *S*, *S*-dialkyldithiocarbonates [23,47].

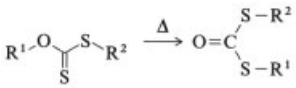

This isomerization can also be effected catalytically with a Lewis acid or trifluoroacetic acid [48]. The isomerization of *O*-allyl xanthate esters occur with simultaneous allyl rearrangement [49].

Alkali-metal xanthates react with alkyl chloroformates to give alkyl xanthogen formates (mixed anhydrosulfides) (**i**), together with dialkyl xanthic anhydrides (**ii**), dialkoxy carbonyl sulfides (**iii**), and dialkyl carbonates (**iv**) as coproducts.

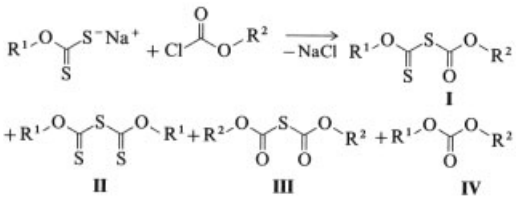

These reaction mixtures are used as collectors in the flotation of copper – molybdenum sulfide ores [50]. Efficient methods for obtaining the xanthogen formates (**i**) as the sole products in nearly quantitative yield have been also described [51].

4. Production

4.1. General

Alkali-metal xanthates are principally synthesized by reaction of an alcohol with sodium or potassium hydroxide and carbon disulfide. This reaction consists of two steps: (1) the formation of the alkali-metal alkoxide; and (2) reaction of the alkoxide with carbon disulfide to give the xanthate.

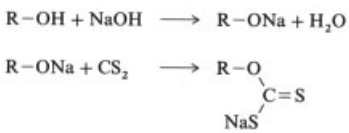

This reaction is applicable to all types of alcohols, including diols, polyols, and carbohydrates, including starch and cellulose. Xanthate salts of primary and secondary alcohols are prepared in aqueous medium, whereas the corresponding derivatives of tertiary alcohols are readily hydrolyzed and are therefore synthesized under anhydrous conditions [52]. Analogous reactions of solid alcohols [52] and of phenol and its derivatives [53] are carried out in aqueous dioxane.

For several purposes (when they are, for example, used as intermediates in the synthesis of xanthate esters) isolation of the alkali-metal xanthates is not necessary. Here, the xanthate is obtained in high yield by converting the alcohol into the corresponding alkoxide with metallic sodium or sodium hydride in an inert solvent and subsequent addition of carbon disulfide. This procedure is also suitable for unsaturated alcohols, e.g., allyl and propargyl alcohol [54].

Another preparation method for alkali-metal xanthates is the reaction of a dialkyl ether with alkali hydroxide and carbon disulfide [55].

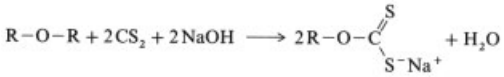

Water-insoluble *heavy-metal xanthates* can be prepared by reaction of dissolved alkali-metal xanthates with heavy-metal salts [5].

Reaction of carbon disulfide with epoxides affords *cyclic dithiocarbonates* (1,3-oxathiolane-2-thiones) [56].

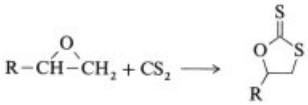

The corresponding 4,5-unsaturated heterocyclic derivatives are obtained by rhodium-catalyzed decomposition of α-diazocarbonyl compounds in the presence of carbon disulfide [57].

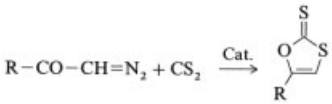

4.2. Production Processes

In the industrial production of alkali-metal xanthates a primary or secondary C_2-C_6 alcohol is reacted first with anhydrous sodium or potassium hydroxide and subsequently with carbon disulfide. To avoid side reactions, e.g., thermal decomposition, the temperature of the reaction mixture is generally maintained below 40 °C by cooling. Since their introduction into flotation in 1925, detailed reaction conditions of the manufacture of xanthates have been evaluated and described in numerous patents. In principle, the reaction can be carried out with or without a solvent.

Production without solvent is carried out in reactors preferentially made from steel that contain cooling and stirring apparatus. The reactors are also equipped with a heating device so that the reaction and the drying of the product can be carried out in the same vessel. Several improvements of the existing methods are reported in the patent literature. Thus, a reactor for the production of alkali-metal alkyl xanthates in one step with economic raw materials utilization and without environmental pollution is described [58]. Another patent reports increased yields and improved quality of xanthates by carrying out the process in a rotor-plate reactor [59]. Potassium butyl xanthate can be produced more efficiently by recycling of the plant mother liquor [60]. Higher reaction yields of potassium amyl xanthate can be obtained by removing part of the water formed with pervaporation membranes [61].

Production in solvents is carried out either in an excess of the corresponding alcohol [62] or in an inert solvent such as petroleum ether [63]. The precipitated or crystallized xanthate is separated by filtration, washed, and dried. The continuous production of xanthates using aqueous alkali-metal hydroxide, an alcohol, and carbon disulfide has also been reported [64]. Moreover, reaction in aqueous solution is used to produce alkali xanthates in the presence of additives, giving dust-free, noncaking, hydrophobic powders [65].

Safety Aspects. The low flash point and ignition temperature and the toxicity of carbon disulfide require appropriate safety measures for the production plants. Dusting of xanthate powders can be prevented by additives [65] or by shaping processes [66].

5. Quality Specifications

The commercially available xanthates from C_2-C_6 alcohols are of technical grade only and usually 90 – 95 % pure. Another publication reports commercial xanthates to be rarely more than 90 % and frequently only 60 – 90 % pure [67]. The content of inorganic byproducts in technical xanthates varies, depending on their exposure to air [68]. Thus, alkali-metal sulfide, sulfate, trithiocarbonate, thiosulfate, or sulfite may be present. In addition, xanthates may contain some residual alcohol or alkali hydroxide. No general specifications for xanthates exist. They can be obtained as powder, granules, flakes, or sticks, and are also available in solution.

6. Storage and Transportation

Xanthates are usually transported and stored in steel drums. Copper, brass, zinc, and similar metals are not suitable. Aqueous xanthate solutions can be transported in vessels made from steel, steel alloys, or plastics, e.g., poly(vinyl chloride) or polyethylene. If the drums are kept cool and dry, solid xanthates are stable for years, but they may ignite spontaneously upon combined action of moisture and warming. Xanthate solutions may form carbon disulfide.

According to GeStoffV xanthates must be labeled as substances that are hazardous to health.

7. Analysis

Xanthates can be analytically determined by different methods [7]. Since many of the published analytical procedures have certain sources of error, the method of assay should be noted together with the specification of the xanthate content.

In general, the xanthate content is determined by a procedure based on decomposition of the xanthate with an excess of acid and subsequent back-titration with a base [3, 69].

Solutions of technical xanthates are usually assayed by *UV spectrometry.* This technique is also used for monitoring the concentration of xanthates in industrial flotation circuits [70]. Furthermore, xanthate solutions can be analyzed *potentiometrically* with silver nitrate [71].

Xanthates of $C_1 - C_5$ alcohols have been separated successfully by *chromatographic methods* such as thin-layer chromatography or HPLC [72].

8. Uses

Cellulose xanthates are the most important xanthates. They are used as starting material in the manufacture of cellophane (→ Films, Chap. 6.1.1.) and viscose fibers (→ Cellulose, Chap. 3.1.).

Flotation Reagents. The principal use for alkali-metal xanthates is as flotation collectors for metal-sulfide ores. Since their introduction in flotation in 1925 many patents and articles dealing with this field of application have been published.

Other applications of xanthates are only of minor importance. A selection of publications on the use of xanthates, drawn mainly from the extensive patent literature, is given below.

Polymerization. Xanthate derivatives are employed in polymerization processes as regulators [73], photoinitiators [74], and for the synthesis of block and graft copolymers [75]. Cyclic dithiocarbonate derivatives are suitable as cross-linking agents [76].

Additives for Lubricating Oils. Various dialkyl xanthates are claimed to be active as additives for lubricating oils [25, 77]. Moreover, xanthate complexes of molybdenum and vanadium are suitable as lubricant additives [78].

Color Photography. Xanthate derivatives are used in the preparation of electrophotographic material [79], as color developer, and photographic couplers [80].

Pesticides and Agricultural Chemicals. Xanthates and their derivatives have been recommended as fungicides [81], acaricides [82], and insecticides [83]. Microbicidal activity is also described for xanthate derivatives of benzeneacetic acid [84] and for quaternary ammonium salts [85] and *N*-methylpyrroles [86] containing dithiocarbonate residues. Halonitroaryl esters of *O*-butylxanthic acid exhibit nematocidal activity [87].

Other xanthate compounds effective as plant growth regulators are derivatives of dipropylene or triethylene glycol [88] and 3-benzothiazoline [89]. Various xanthate derivatives are claimed to be active as herbicides [90].

Pharmaceuticals. Pharmaceutical effects of xanthates include virustatic [91], antihypertensive [92], anthelmintic [93], analgetic and antipyretic [94], and antibiotic activity [95]. Pharmacological uses are also claimed for xanthate derivatives of aminoalkyl glycerols [96]. Potassium and gallium ethyl xanthate are effective antitumor agents [97] and 1,3-benzoxathiol-2-thiones (cyclic dithiocarbonates) can be used for treatment of liver diseases [98]. A number of phenylacyl xanthate esters and xanthogen formates (mixed anhydrosulfides) have been synthesized as potential radiosensitizers [99]. In addition, xanthates have been described as valuable intermediates for the preparation of drugs [100].

Rubber. Xanthates are used as antioxidants and vulcanization accelerators for rubber [101].

Environmental Chemistry. Traces of heavy metals can be removed from wastewaters by precipitation with xanthates [102]. Similarly, silver can be recovered from spent fixing baths using xanthates [103]. Cadmium can be extracted from soil by complexation with long-chain alkyl xanthates [104]. The recovery of xanthates from production mother liquors has been described [105].

Analytics. DNA suitable for polymerase chain reactions is isolated from plants, yeast, and bacteria using potassium or sodium ethyl xanthate [106]. Xanthates are also useful chelating reagents for the spectrophotometric detection of metals [107].

Other Uses. Substituted noble-metal xanthates are used in catalysts systems for removing impurities, e.g., S and N compounds, from hydrocarbons by hydrotreating [108].

Alkoxycarbodithiochloroformates are suitable condensing reagents in peptide synthesis [109].

9. Economic Aspects [110]

Various xanthates are produced by a large number of companies worldwide. The major mining chemical companies producing xanthates in Northern America are American Cyanamid, C-I-L (a wholly owned subsidiary of ICI), Kerley Enterprises, and Minerec Mining Chemicals. The Charles Tennant Co. plans (1993) to build a 12 000 – 14 000 t/a xanthate facility in Montreal, Canada. UCB in Belgium is the principal supplier of xanthates in Western Europe, but also sells to Africa, Canada, and the Philippines. Other producers in Western Europe are Hoechst (Germany), Shell (United Kingdom), and General Quimica (Spain). Robinson Brothers (United Kingdom) produces diisopropylxanthogen disulfide and zinc isopropylxanthate as rubber processing chemicals. Industrias Quimicas (Mexico), owned by Rhône-Poulenc (France), dominates the large Mexican market for xanthates and also sells to other South and Central American countries. Shell Chemicals produces xanthates and xanthogen formates in Chile. Karbochem (Republic of South Africa) manufactures a range of alkali-metal xanthates, marketing them in the Republic of South Africa and other African nations.

In 1989 prices for technical grade xanthates were in the range of $ 1.20 – 3.60 per kilogram for potassium amyl xanthate and $ 1.00 – 2.54 per kilogram for sodium isopropyl xanthate.

10. Toxicology and Occupational Health [111]

Data on the toxicity of xanthates and their derivatives in the literature are rather rare.

The alkali-metal xanthates are moderately toxic by ingestion or inhalation, but poisonous by intravenous route, as indicated in Table 3. In the body, xanthates act similarly to carbon disulfide.

Table 3. Toxicity of potassium xanthates, $R-O-CS-S^-K^+$, to mice

	LD_{50}, mg/kg	
R	Oral	Intravenous
Ethyl	683	199
Isopropyl	583	
Butyl	465	158
Isobutyl	480	158
Isopentyl	480	158

Animal experiments on the *chronic toxicity* of potassium amyl xanthate revealed that liver damage occurred in dogs exposed to an aqueous aerosol of 23 mg/m^3.

Because of their basicity, prolonged contact of alkali-metal xanthates and their concentrated solutions with the skin and especially with the eyes must be avoided. Skin contact may cause dermatitis.

Xanthates emit toxic fumes of SO_x and metal oxide when heated to decomposition. In the preparation and handling of xanthate solutions the possible formation of carbon disulfide requires protection against fire and explosions.

11. References

General References

1. *Ullmann,* 4th ed., **24,** 511.
2. *Houben-Weyl,* **E 4,** 425; **9,** 810.
3. *Kirk-Othmer,* 3rd ed., **24,** 645.
4. A. D. Dunn, W.-D. Rudorf: *Carbon Disulphide in Organic Chemistry,* Ellis Horwood, Chichester 1989, pp. 316 – 339.
5. G. Gattow, W. Behrendt in A. Senning (ed.): "Carbon Sulfides and Their Inorganic and Complex Chemistry," *Topics in Sulfur Chemistry,* **vol. 2,** Georg Thieme, Stuttgart 1977, pp. 119 – 154.
6. E. E. Reid: *Organic Chemistry of Bivalent Sulfur,* **vol. 4,** Chemical Publishing, New York 1962, p. 131.
7. S. R. Rao: *Xanthates & Related Compounds,* Marcel Dekker, New York 1971.

Specific References

8. W. C. Zeise, *Schweiggers J. Chem. Phys.* **35** (1822) 173; **36** (1822) 1.
9. H. Majima, *Sci. Rep. Res. Inst. Tohoku Univ., Ser. A* **13** (1961) 183.

10. F. Drawert, K. H. Reuter, F. Born, *Chem. Ber.* **93** (1960) 3056. I. S. Shupe, *J. Assoc. Off. Agric. Chem.* **25** (1942) 495; *Chem. Abstr.* **36** (1942) 4670.
11. F. Galsbøl, C. E. Schäffer, *Inorg. Synth.* **10** (1967) 42.
12. I. A. Kakovski, *Proc. Int. Congr. Surf. Act. 2nd* 1957, no. 4, 225. A. T. Pilipenko, *Zh. Anal. Khim.* **4** (1949) 227.
13. J. Dyer, L. H. Phifer, *Macromolecules* **2** (1969) 111. R. J. Millican, C. K. Sauers, *J. Org. Chem.* **44** (1979) 1964.
14. M. L. Shankaranarayana, C. C. Patel, *Acta Chem. Scand.* **19** (1965) 1113.
15. L. H. Little, G. W. Poling, J. Leja, *Can. J. Chem.* **39** (1961) 745. F. G. Pearson, R. B. Stasiak, *Appl. Spectrosc.* **12** (1958) 116. A. T. Pilipenko, N. V. Melnikova, *Russ. J. Inorg. Chem. (Engl. Transl.)* **15** (1970) 608.
16. J. Berger, *Acta. Chem. Scand.* **6** (1952) 1564. J. R. Millican, C. K. Sauers, *J. Org. Chem.* **44** (1979) 1664.
17. I. Iwasaki, S. R. B. Cooke, *J. Am. Chem. Soc.* **80** (1958) 285.
18. Bayer, DE 2 533 989, 1975 (W. Behr, H. Königshofen). H. Dautzenberg, B. Philipp, M. Frede, *J. Prakt. Chem.* **312** (1970) 1051. Toyo Soda, JP 61 210 067, 1985 (S. Aman, H. Watanabe, A. Nakanishi, Y. Nonaka). G. Bulmer, F. G. Mann, *J. Chem. Soc.* 1945, 677. E. I. Stout, B. S. Shasha, W. M. Doane, *J. Org. Chem.* **39** (1974) 562.
19. R. Woods, *Aust. J. Chem.* **25** (1972) 2329. Academy of Sciences Kazakh, SU 1 116 035, 1981 (L. G. Ili'chenko et al.). Irkut. Poly., SU 1 244 142, 1984 (S. A. Bogidaev, A. N. Baranov, S. B. Leonov).
20. W. Winkler, W. Gruneberg, DD 118 599, 1966. Orgsintez, Volzhskii, SU 1 518 338, 1987 (Y. P. Polin et al.).
21. F. Fehér, W. Becher, R. Kreutz, F. R. Minz, *Z. Naturforsch. B* **18 B** (1963) 507. W. Knoth, G. Gattow, *Z. Anorg. Allg. Chem.* **552** (1987) 181.
22. H. R. Nace, *Org. React.* **12** (1962) 57.
23. P. Meurling, K. Sjöberg, B. Sjöberg, *Acta Chem. Scand.* **26** (1972) 279. M. B. Diana, M. Marchetti, G. Melloni, *Tetrahedron: Asymmetry* **6** (1995) 1175.
24. I. Degani, R. Fochi, *Synthesis* 1978, 365. I. Degani, R. Fochi, V. Regondi, *Synthesis* 1981, 149.
25. A. S. Nakhmanov, S. I. Krakhmalev, L. I. Nazarova, SU 910 613, 1980; SU 910 614, 1980.
26. A. W. M. Lee, W. H. Chan, H. C. Wong, M. S. Wong, *Synth. Commun.* **19** (1989) 547. Phillips Petroleum, US 5 239 104, 1993 (J. E. Shaw). Inst. Cercetari, Ing. Tehnol., Baia Mare, RO 102 036, 1992 (G. Duda et al.); *Chem. Abstr.* **120** (1994) 191136.
27. A. Le Minor, I. E. Kanjo, D. Villemin, *Polym. Bull. (Berlin)* **21** (1989) 445.
28. S. M. Ali, S. Tanimoto, *J. Org. Chem.* **54** (1989) 5603. S. M. Ali, S. Tanimoto, *Bull. Inst. Chem. Res. Kyoto Univ.* **68** (1990) 199. S. M. Ali, S. Tanimoto, *J. Chem. Soc. Chem. Commun.* 1989, 684.
29. A. Schönberg, E. Frese, W. Knöfel, K. Praefke, *Chem. Ber.* **103** (1970) 938. Establ. Ind. Quim. Oxiquim., Chile, US 5 011 965, 1991 (E. Saphores, A. G. Gleisner, J. C. Vega, W. A. Mardones). Indian Oil, IN 171 122, 1992 (Y. K. Gupta et al.); *Chem. Abstr.* **120** (1994) 216722.
30. D. S. Tarbell, D. K. Fukushima, *Org. Synth. Coll. Vol.* **3** (1955) 809.
31. J. R. Cox, C. L. Gladys, L. Field, D. E. Pearson, *J. Org. Chem.* **25** (1960) 1083.
32. J. March: *Advanced Organic Chemistry,* 4th ed., Wiley-Interscience, New York 1992, p. 1015.
33. J.-R. Dormoy et al., *Angew. Chem. Int. Ed. Engl.* **19** (1980) 742.
34. D. H. R. Barton, *J. Chem. Soc.* 1949, 2174.
35. D. H. R. Barton, S. W. McCombie, *J. Chem. Soc. Perkin Trans. 1* 1975, 1574. D. H. R. Barton, D. Crich, A. Loebberding, S. Z. Zard, *Tetrahedron* **42** (1986) 2329.
36. S. Jacono, J. R. Rasmussen, *Org. Synth.* **64** (1985) 57. K. Nozaki, K. Oshima, K. Utimoto, *Tetrahedron Lett.* **29** (1988) 6125. K. Nozaki, K. Oshima, K. Utimoto, *Bull. Chem. Soc. Jpn.* **63** (1990) 2578.
37. Univ. Georgia Res. Found., WO 92/15 598, 1991 (Y. Chen, C. K. Chu).
38. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Synlett* 1991, 435. J. N. Kirwan, B. P. Roberts, C. R. Willis, *Tetrahedron Lett.* **31** (1990) 5093. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron* **49** (1993) 7193. S. J. Cole, J. N. Kirwan, B. P. Roberts, C. R. Willis, *J. Chem. Soc. Perkin Trans.* **1** 1991, 103.
39. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron Lett.* **33** (1992) 2311.
40. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron Lett.* **33** (1992) 5709.
41. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron Lett.* **32** (1991) 7187. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron* **49** (1993) 2793.

42. D. H. R. Barton, S. I. Parekh, C. L. Tse, *Tetrahedron Lett.* **34** (1993) 2733.
43. A. G. M. Barrett, D. H. R. Barton, R. Bielski, *J. Chem. Soc. Perkin Trans.* **1** 1979, 2378. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *Tetrahedron Lett.* **32** (1991) 2569. D. H. R. Barton, D. O. Jang, J. C. Jaszberenyi, *J. Org. Chem.* **58** (1993) 6838.
44. J. H. Udding, J. P. M. Giesselink, H. Hiemstra, W. N. Speckamp, *J. Org. Chem.* **59** (1994) 6671.
45. J. H. Udding, J. P. M. Giesselink, H. Hiemstra, W. N. Speckamp, *Bull. Soc. Chim. Belg.* **103** (1994) 392. J. Axon et al., *Tetrahedron Lett.* **35** (1994) 1719.
46. M. Yamamoto et al., *J. Chem. Soc. Chem. Commun.* 1989, 1265. S. Iwasa, M. Yamamoto, S. Kohmoto, K. Yamada, *J. Org. Chem.* **56** (1991) 2849.
47. K. G. Rutherford, B. K. Tang, L. K. M. Lam, D. P. C. Fung, *Can. J. Chem.* **50** (1972) 3288. Courtaulds, EP 202 054, 1985 (P. G. Urben).
48. K. Komaki, T. Kawata, K. Harano, T. Taguchi, *Chem. Pharm. Bull.* **26** (1978) 3807. M. W. Fichtner, N. F. Haley, *J. Org. Chem.* **46** (1981) 3141. J. R. Pougny, *J. Carbohydr. Chem.* **5** (1986) 529.
49. K. Harano, T. Taguchi, *Chem. Pharm. Bull.* **20** (1972) 2357. S. N. Suryawanshi, A. Rani, D. S. Bhakuni, *Synth. Commun.* **20** (1990) 625. F. Chatzopoulos-Ouar, G. Descotes, *J. Org. Chem.* **50** (1985) 118. D. Villemin, M. Hachemi, *Synth. Commun.* **25** (1995) 2311. T. Taguchi et al., *Tetrahedron Lett.* **6** (1965) 2717.
50. R. D. G. Crozier, US 4 354 980, US 4 454 051, US 4 605 518, 1981.
51. Pfizer, EP 274 876, 1986 (C. Sklavounos). M. A. Palominos, J. G. Santos, J. C. Vega, M. A. Martinez, *Phosphorus Sulfur* **25** (1985) 91.
52. K. A. Jensen, U. Anthoni, A. Holm, *Acta Chem. Scand.* **23** (1969) 1916.
53. A. K. Kadyrov, L. B. Alavidinov, D. V. Islamova, M. G. Ismatuellaeva, *Dokl. Akad. Nauk. USSR* 1989, 40; *Chem. Abstr.* **112** (1990) 216390.
54. K. Harano, T. Taguchi, *Chem. Pharm. Bull.* **20** (1972) 2357. K. Tomota, M. Nagano, *Chem. Pharm. Bull.* **20** (1972) 2302.
55. Standard Oil, US 2 534 085, 1950 (B. M. Vanderbuilt, J. P. Thorn). J. Sejbl, *Rudy* **31** (1983) 163; *Chem. Abstr.* **99** (1983) 175192.
56. Agency of Industrial Sciences and Technology, JP 63 218 672, 1987 (Y. Taguchi, K. Yanagiya, I. Shibuya, Y. Suhara). N. Kihara, Y. Nakawaki, T. Endo, *J. Org. Chem.* **60** (1995) 473.
57. T. Ibata, H. Nakano, *Chem. Express* **4** (1989) 93. T. Ibata, H. Nakano, *Bull. Chem. Soc. Jpn.* **63** (1990) 3096.
58. Inst. Cercetari, Ing. Tehnologica si Proiectari pentru Minereuri Neferoase, RO 89 624, 1984 (G. Duda et al.); *Chem. Abstr.* **106** (1987) 175787; RO 89 625, 1984 (G. Duda et al.); *Chem. Abstr.* **107** (1987) 25115 g.
59. G. A. Anokhin et al., SU 1 214 662, 1983; *Chem. Abstr.* **105** (1986) 171855 y.
60. Central Ural Copper-Smelting Plant, SU 1 351 925, 1986 (B. N. Driker et al.); *Chem. Abstr.* **108** (1988) 223445 y. Central Ural Copper-Smelting Plant, SU 1 468 899, 1986 (Y. K. Achaev et al.); *Chem. Abstr.* **111** (1989) 96669 j.
61. M. M. Cipriano, V. Geraldes, M. Norberta de Pinho, *J. Membr. Sci.* **62** (1991) 103.
62. Degussa, DE 765 196, 1955 (H. Gunther).
63. Hoechst, DE 1 934 175, 1971 (K. Baessler, G. Folz). D. I. Mendeleev Chem.-Tech. Inst., SU 727 641, 1980 (V. P. Savel'yanova, R. T. Savel'yanova, N. P. Utrobin, T. V. Rudakova).
64. O. Leminger et al., CS 97 929, 1961. P. P. Gnatyuk et al., *Khim. Prom-st. (Moscow)* 1983, 402.
65. Volgograd Polytechn. Inst., SU 1 066 988, 1980 (A. P. Khardin et al.). Volgograd Polytechn. Inst., AU 8 290 043, 1982.
66. Zhuzhon Dressing Chemical Plants, CN 1 074 902, 1992 (L. Xu, D. Chen, C. Lu); *Chem. Abstr.* **120** (1994) 298077. Qixia County Ore Dressing Agent Factory, CN 1 072 354, 1991 (C. H. Zhang, S. Jian); *Chem. Abstr.* **121** (1994) 113856.
67. J. Leja: *Surface Chemistry of Froth Flotation,* Plenum Press, New York 1982, p. 229.
68. S. N. Danilov, N. M. Grad, E. I. Geine, *Zh. Obshch. Khim.* **19** (1949) 826.
69. W. Hirschkind, *Eng. Min. J.* **119** (1925) 968.
70. M. H. Jones, J. T. Woodcock, *Proc. Austr. Inst. Min. Metall.* **238** (1973) 1.
71. C. Du Rietz, *Sven. Kem. Tidskr.* **69** (1957) 310.
72. A. Messina, D. Corradini, *J. Chromatogr.* **207** (1981) 152. J. G. Eckhardt, K. Stetzenbach, M. F. Burke, J. L. Moyers, *J. Chromatogr. Sci.* **16** (1978) 510. R. T. Honeyman, R. R. Schrieke, *Anal. Chim. Acta* **116** (1980) 345.
73. Bayer, EP 53 319, 1980 (G. Arend, H. Königshof, P. Müller, R. Musch). Desoto, US 4 537 667, 1984 (T. E. Bisho, G. Pasternack, J. M. Zimmerman).

74. Tosoh Corp., EP 450 492, 1991 (H. Yamakawa, S. Ozoe). Fuji Photo Film, JP 6 083 079, 1993 (E. Kato, S. Oosawa). A. Ajayaghosh, *Polymer* **36** (1995) 2049.
75. Moscow Textile Inst., SU 821 453, 1979 (G. A. Gabrielyan, M. S. Kislyuk, V. S. Lebedev, R. P. Shul'ga). M. Niwa, N. Higashi, M. Shimizu, T. Matsumoto, *Makromol. Chem.* **189** (1988) 2187.
76. Nippon Zeon, JP 7 145 164, 1993 (K. Watanabe, T. Matsura). Nippon Zeon, JP 7 062 190, 1993 (Y. Takao, K. Watanabe, H. Mori).
77. A. S. Nakhmanovich, SU 996 410, 1981. Inst. Chem. Additives, Acad. Sci. Azerbaidzhan, SU 1 505 931, 1988 (A. B. Kuliev, M. M. Dzhavadov, B. A. Abushova). Inst. Chem. Additives, Acad. Sci, Azerbaidzhan, SU 1 404 504, 1986 (A. M. Levshina, M. A. Agaeva, D. S. Gamidova, I. A. Buniyat-Zade).
78. Exxon, US 4 995 996, 1989 (C. L. Coyle). Exxon, US 4 861 958, 1985 (T. R. Halbert, E. I. Stiefel).
79. Fuji Photo Film, EP 599 660, 1993 (E. Kato). Fuji Photo Film, WO 92/18 907, 1992 (E. Kato, K. Ishii). Yamada, JP 7 076 568, 1993 (F. Kiryama, I. Nakanishi).
80. Konishiroku Photo, JP 57 096 335, 1982.
81. Ciba-Geigy, CH 636 610, 1977 (W. Kunz, A. Hubele, W. Eckhardt). P. C. Joshi, Jr., K. B. Karnatak, M. M. Sah, A. K. Pant, *J. Indian Chem. Soc.* **62** (1985) 164. W. Podkoscielny et al., *Przem. Chem.* **69** (1990) 495; *Chem. Abstr.* **114** (1991) 184935.
82. Begin-Say, FR 2 543 555, 1983 (J. Mentech).
83. Sandoz, EP 169 169, 1985 (C. A. Henrick). Takeda, JP 64 000 063, 1988 (H. Uneme, H. Mitsudera, T. Uekado).
84. BASF, EP 582 902, 1993 (H. Winger, H. Sauter, E. Ammermann, G. Loren).
85. Degussa, EP 386 508 and EP 386 509, 1989 (P. Werle, M. Trageser, S. Weiss).
86. American Cyanamid, EP 545 103, 1992 (D. G. Kuhn, V. Kameswaran).
87. P. Dureja, S. P. Gupta, D. Prasad, *Pesticides* **18** (1984) 43.
88. CPC International, US 4 806 149, 1987 (M. J. Danzig, A. M. Kinnersley).
89. Monsanto, US 4 427 436, 1982 (J. J. D'Amico).
90. Monsanto, US 4 556 415, 1985 (J. E. Franz, K. E. Zwikelmaier). Ciba-Geigy, EP 139 612, 1984 (W. Toepfl).
91. Merz, US 4 602 037, 1983 (A. Scherm, K. Hummel). Merz, BE 891 245, 1981 (K. Hummel, A. Scherm).
92. Boehringer, EP 255 075, 1986 (M. Frigerio et al.). Mitsubishi Kasei, JP 4 021 695, 1990 (F. Takaku et al.).
93. Y. D. Kulkarni, B. Kumar, *J. Indian Chem. Soc.* **60** (1983) 882.
94. Sandoz, DE 3 320 746, 1982 (R. Achini, L. Revesz).
95. Otsuka, WO 84/00 964, 1982 (J. Nogami et al.).
96. Ono Pharmaceutical Co., EP 94 586, 1983 (H. Itoh, T. Okada, T. Miyamoto).
97. D. Khadzhiolov, N. Frank, N. Khadzhiolov, S. Yanev, *Arch. Geschwulstforsch.* **58** (1988) 313. Ajinimoto, JP 6 197 292, 1984 (T. Cho, T. Tsuchiya, T. Yoshihama, T. Shiio).
98. Shinnippon, JP 6 239 856, 1993; JP 6 239 640, 1994 (T. Ueyama et al.).
99. D. Skwarski, H. Sobolewski, *Acta Pol. Pharm.* **49** (1992) 71, 77, 81; *Chem. Abstr.* **120** (1994) 191269, 191270; *Chem. Abstr.* **121** (1994) 8835. D. Skwarski, H. Sobolewski, *Acta Pol. Pharm.* **50** (1993) 213; *Chem. Abstr.* **121** (1994) 108121. J. Tulecki, J. Dudzinska-Usarewicz, P. Radola, *Pol. J. Pharmacol. Pharm.* **37** (1985) 709.
100. Pfizer, EP 131 374, 1984 (E. S. Hamanaka). Sagami Chem. Res. Center, JP 62 132 857, 1985 (M. Matsumoto, N. Watanabe). Leo Pharmaceutical Prod., WO 94/14 766, 1993 (G. Grue-Soerensen, E. R. Ottosen).
101. P. Rajalingam, G. Radhakrishnan, *Polym. Plast. Technol. Eng.* **30** (1991) 405. Asahi Chem. Ind., JP 62 127 330, 1985 (Y. Kitagawa, Y. Hattori). Tashkent Poly., SU 1 065 412, 1982 (A. Safaev, L. Alavidinov, N. B. Niyazoc).
102. VEB Werke Buna, DD 148 044, 1979 (K. D. Rauchstein et al.). R. Braun, R. Rafflenbeul, DE 4 009 589, 1990.
103. M. Nanjo, Y. Itoh, *Tohoku Daigaku Senku Seiren Kenkyusho Iho* **42** (1986) 127; *Chem. Abstr.* **106** (1987) 186253.
104. N. Wolf, D. M. Roundhill, *Polyhedron* **13** (1994) 2801.
105. 'Kazmekhanobr' Inst., SU 1 449 545, 1986 (A. L. Droyaronov et al.); *Chem. Abstr.* **110** (1989) 195647 m.
106. C. E. Williams, P. C. Ronald, *Nucleic Acids Res.* **22** (1994) 1917. Pioneer Hi-Bred Int., EP 493 115, 1991 (A. K. Jhingan).
107. D. Yamamoto, M. Aoyama, *Meiji Daigaku Nogakubu Kenkyu Hokoku* 1985, 69; *Chem. Abstr.* **104** (1986) 198955.
108. Exxon Res. & Eng. US 5 068 375, 1990 (G. H. Singhal).

109. Tadzhik State Univ., SU 1 146 302, 1983 (S. K. Khalikov, M. M. Sobirov, S. V. Alieva, A. A. Akhtyamova); *Chem. Abstr.* **103** (1985) 123921 d. S. K. Khalikov, S. V. Alieva, M. M. Sobirov, *Khim. Prir. Soedin.* 1985, 844; *Chem. Abstr.* **105** (1986) 79339.
110. SRI International (ed.): Chemical Economics Handbook, Directory of Chemical Producers, Europe 1995. E. A. Thiers in SRI International (ed.): Chemical Economics Handbook, Specialty Chemicals, Mining Chemicals-Worldwide, 1989.
111. R. J. Lewis, Sr. (ed.): *Sax's Dangerous Properties of Industrial Materials,* 8th ed., vol. III, Van Nostrand Reinhold, New York 1992.

Xanthene Dyes → Triarylmethane and Diarylmethane Dyes
Xenon → Noble Gases
Xenon Fluorides → Noble Gases
Xenon Oxides → Noble Gases
X-Ray Diffraction → Structure Analysis by Diffraction

Xylenes

Jörg Fabri, Veba AG, Düsseldorf, Federal Republic of Germany

Ulrich Graeser, Veba Öl AG, Gelsenkirchen, Federal Republic of Germany

Thomas A. Simo, Lurgi Öl, Gas, Chemie GmbH, Frankfurt, Federal Republic of Germany

1. Introduction 1
2. Properties 3
3. Occurrence and Raw Materials . . . 6
3.1. Natural Occurrence 6
3.2. Raw Materials for Xylene Production 6
4. Production, Separation, and Further Processing 7
4.1. Production of C_8 Aromatics 7
4.2. Separation and Further Processing 9
4.2.1. Separation of the C_8 Aromatics Fraction 9
4.2.2. Isomerization of the C_8 Fraction . . . 12
4.2.3. Combination of the Technologies in the Aromatics Complex 12
5. Integration into Refinery and Petrochemical Complexes 13
5.1. Backward Integration into Petroleum Refining 13
5.2. Forward Integration into Chemical Processing 13
6. Economic Aspects 13
7. Quality Specifications and Analysis 18
8. Storage, Transport, and Safety . . . 18
9. Environmental Aspects and Toxicology 19
10. References 20

1. Introduction

The benzene homologues of general formula C_8H_{10} are generally known as mixed xylenes. The mixture of isomers with a boiling point range of 135 – 145 °C mainly consists of the three isomeric dimethylbenzenes and ethylbenzene:

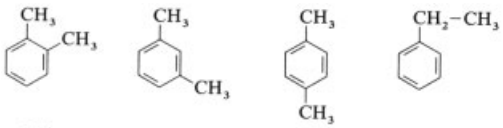

o-Xylene *m*-Xylene *p*-Xylene Ethylbenzene

With the exception of xylene production by disproportionation of toluene, the isomeric xylenes and ethylbenzene are always produced as a mixture in all production processes. However, the relative proportions of the C_8 isomers often differ considerably (see Chap. 3).

Because of their high knock resistance (see Chap. 2), xylenes are well suited to the production of motor fuels. In terms of quantity the production of gasoline exceeds that of BTX aromatics (B = benzene, T = toluene, X = xylenes) quite considerably (in Western Europe in 1995 gasoline production was ca. 150×10^6 t and that of BTX aromatics ca. 13.7×10^6 t, of which ca. 2.7×10^6 t/a was mixed xylenes, 0.65×10^6 t/a *o*-xylene, and 1.4×10^6 t/a *p*-xylene) [1]. The average aromatics content of motor fuels in Western Europe is ca. 38 % [2].

The close association with gasoline production strongly affects the economics of separating xylene mixtures, for example, for use in chemical processes.

In Western Europe the production/use balance in 1995 was as follows [3]:

Production		
from coal	36×10^3 t	(ca. 1.3 %)
from pyrolysis gasoline	300×10^3 t	(ca. 11.1 %)
from reformate	2165×10^3 t	(ca. 80.0 %)
by disproportionation	205×10^3 t	(ca. 7.6 %)
Total production	2706×10^3 t	
Use		
in extraction and chemical processing (incl. imports)	2563×10^3 t	(ca. 88.4 %)
solvents	306×10^3 t	(ca. 10.6 %)
as gasoline component	30×10^3 t	(ca. 1 %)
Total use	2899×10^3 t	
Domestic production	93.3 %	
Net impor	6.7 %	

The structure of world mixed-xylenes production capacity in 1994 was as follows:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_433

Total capacity	23×10^6 t
Reformer	83 %
Disproportionation	10 %
Pyrolysis gasoline	6 %
Coal	1 %

In Western Europe until the year 2000 a significant increase in the proportion of xylene used in gasoline production is predicted, because the growth in production of xylene-containing reformate may exceed its consumption in chemical processes. However, in other regions, such as the United States, the reduction and limitation of the content of aromatics, and thus of xylenes, in gasoline is being discussed (see also Chap. 2 and Section 5.2) [2,4].

The use of xylenes in chemical processes broke down as follows (data for Western Europe 1995) [1]:

p-Xylene	1761×10^3 t (68.7 %)
o-Xylene	650×10^3 t (25.4 %)
Ethylbenzene and miscellaneous	152×10^3 t (5.9 %)
Total	2563×10^3 t

The dominant importance of *p*-xylene can be seen from the relative proportions of the isolated xylene isomers. *p*-Xylene is mainly oxidized to terephthalic acid, which can be esterified to dimethyl terephthalate (precursor for polyesters). *o*-Xylene is oxidized to phthalic anhydride (precursor for plasticizers) and *m*-xylene to isophthalic acid (precursor for polyesters). Ethylbenzene is dehydrogenated to styrene, which is converted to polystyrene and other polymers [5].

History [6]. Xylene was first discovered in crude wood spirit in 1850 by Cahours. The name xylene was derived from the Greek word *xylon* (= wood). Xylene was detected in coal-tar in 1855 by Ritthausen and Church. In 1891 ethylbenzene was found in hard coal-tar by Noelting and Palmar, and Moschner carried out the first synthesis of pure ethylbenzene from ethylbenzenesulfonic acid.

Between 1865 and 1869 Ernst and Fittig found that the "xylene" in coal-tar does not consist of a single compound. They synthesized "xylene" from toluene and named it "methyltoluene." They then established that some properties of "methyltoluene" differed from those of coal-tar xylene, particularly with regard to the nitro compounds. Fittig et al. concluded that in synthetic "methyltoluene" (*p*-xylene) the second methyl group was in a different position from that in coal-tar xylene.

Glinser and Fittig described the production of methyltoluene in 1865. Fittig obtained a hydrocarbon by dry distillation of the calcium salt of mesitoic acid which resembled both known modifications of dimethylbenzene, ("xylene" and "methyltoluene") but was not completely identical to either. The new hydrocarbon was named "isoxylene" (*m*-xylene). This work elucidated the structures of *p*- and *m*-xylene and proved their occurrence in coal-tar.

In a further experiment Fittig identified *o*-xylene as the third isomer, which was discovered in coal-tar by Jacobson in 1877.

m- and *p*-xylene were initially separated on the basis of their different chemical reactivity. *m*-Xylene reacts with concentrated sulfuric acid. *m*-Xylene is then regenerated from the crystalline *m*-xylenesulfonic acid using superheated steam in the presence of sulfuric acid. The fraction unaffected by sulfonation is then reacted with oleum. The sulfonic acid that crystallizes is cleaved to give pure *p*-xylene.

For decades coke-oven tar and benzole were used as raw materials for production of aromatics. Worldwide their production was associated in location and quantity with coal and steel production. The plants for extraction and refining of coke-oven benzole, a mixture of light oil from coke-oven tar and crude benzene, obtained from coke-oven gas by oil scrubbing, were built adjacent to coking plants. These plants were very small compared with those currently in operation. At that time the refining of crude benzole was carried out exclusively by washing with concentrated sulfuric acid, whereby impurities, such as resin-forming compounds, nitrogen and oxygen compounds, and some of the sulfur compounds, were removed as acid tar. However, valuable substances, such as styrene, were also lost in the acid tar. The xylenes were in demand as solvents and were used as raw materials in the chemical and pharmaceutical industries. However, production of the three isomers in pure form was complicated due to their almost identical physical properties. The quantities available were too small for the development of industrial xylene chemistry.

m- and *p*-xylene could be separated from *o*-xylene by distillation and thus *o*-xylene could

be obtained in limited purity. *m*- and *p*-xylene, however, could only be separated laboriously on the basis of their different chemical properties.

With the development of the hydrogenative refining of coke-oven benzole over sulfur-resistant catalysts under pressure, sulfuric acid refining became uneconomical and was replaced by a technically superior process (see → Benzene, Chap. 6.1.). Parallel to this development in process technology, aromatics production became independent of steelworks or coking plants and centralization took place. Large refineries and distillation plants provided a basis for modest chemical processing of the C_8 aromatics in the mid-1950s. Between 1955 and 1960 ca. 30 000 t of xylenes were obtained from a coke production level of $(40-45)\times 10^6$ t/a.

The fourth isomer, ethylbenzene, must be considered separately from the historical development of the isomeric xylenes. Its synthesis from benzene and ethylene in a Friedel – Crafts reaction solved the problem of raw materials at an early stage.

With the discovery of the catalytic dehydrogenation of naphthenes and the catalytic dehydrocyclization of paraffins and isoparaffins on noble metal catalysts and their bifunctional systems (platforming and rheniforming processes), the raw material basis for the development of xylene chemistry on a large scale was secured in Europe from about 1960. A large number of processes have now been developed: liquid – liquid extraction, azeotropic distillation, extractive distillation, isomerization, crystallization, adsorption, and absorption. These not only provided pure products on an industrial scale, but also permitted the separation of *m*- and *p*-xylene. Thus the basis for a large-scale growth of polyester fiber production was created. The residual *m*-xylene could be isomerized or oxidized to isophthalic acid, which had a limited market. The market for *o*-xylene was provided by phthalic anhydride production.

2. Properties [6]

Physical Properties. Technical-grade xylene is a mixtrue of C_8 aromatics, also known as the A_8 fraction, consisting of the three xylene isomers, ethylbenzene, and depending on the origin, varying amounts of nonaromatics which boil in the same range (136 – 145 °C). Table 1 lists some physical data for the aromatic components. Other data can be found in [7, 8]. The relative proportions of these components in a xylene fraction depend on its origin and also which process steps the xylene has already passed through (e.g., hydrogenation).

The vapor pressure curves of the three xylene isomers and ethylbenzene are shown in Figure 1.

The isomeric xylenes and ethylbenzene form azeotropic mixtures with water and numerous organic compounds (Table 2) [9–13].

The absorption properties of xylene are of technical interest because of the significant differences in the solubilities of various gases as a function of temperature (see Table 3).

Chemical Properties. Oxidation of the xylene isomers gives the corresponding aromatic dicarboxylic acids. Phthalic acid is produced industrially from *o*-xylene, isophthalic acid from *m*-xylene, and terephthalic acid from *p*-xylene (see → Phthalic Acid and Derivatives, → Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid).

The oxidation reactions have been established as industrial processes in both the gas and liquid phases. Attempts have been made to introduce the co-oxidation of *p*-xylene with paraformaldehyde (Toray Industries) or acetaldehyde (Eastman Kodak) [5].

Ammonoxidation of *m*- and *p*-xylenes initially gives the corresponding phthalic acid dinitriles, which are important raw materials for the production of isocyanates via reduction to the corresponding xylylenediamines. The dinitriles can also be hydrolyzed to the corresponding phthalic acids. However, this step has only limited industrial and economic importance.

The nitration of *o*- and *m*-xylenes provides a route to xylidines following hydrogenation of the initially formed dimethylnitrobenzene isomers. Xylidines are used as intermediates in dye and rubber additive production, for example (→ Xylidines, Chap. 6.1.).

The capacity of the xylene isomers to undergo isomerization and disproportionation reactions is also exploited industrially (see Chap. 4).

Sulfonation of *m*-xylene and subsequent decomposition of the sulfonic acid derivatives gives 3,5- and 2,4-xylenols, providing starting materials for insecticides, herbicides, etc.

Table 1. Physical data for xylene isomers and ethylbenzene

		o-Xylene	*m*-Xylene	*p*-Xylene	Ethylbenzene
M_r		106.16	106.16	106.16	106.16
bp at 1 bar,	°C	144.4	139.1	138.4	136.2
Critical temperature,	°C	357.1	343.6	342.8	344.0
Critical pressure	bar	35.20	35.47	34.45	37.27
Critical compressibility		0.260	0.270	0.250	0.260
Critical molar volume	L/mol	0.380	0.390	0.370	0.371
	g/cm^3	0.281	0.282	0.290	0.284
mp,	°C	− 25.182	− 47.87	+ 13.26	− 95.00
Surface tension at 15.6 °C,	mN/m	30.70	30.12	28.8	29.5
Dynamic viscosity at 20 °C,	mPa · s	0.809	0.617	0.644	0.6783
Thermal conductivity at 0 °C,	$kJ\,m^{-1}\,h^{-1}\,K^{-1}$	0.544	0.523	0.511	0.494
Enthalpy of evaporation at the *bp* and 1 bar,	kJ/mol	36.89	36.45	36.00	36.05
Lower calorific value,	kJ/g				
Gas		41.25	41.24	41.24	41.34
Liquid		40.84	40.83	40.84	40.94
Density at 1 bar,	g/cm^3				
at 25 °C		0.8760	0.8599	0.8567	0.8624
at 20 °C		0.8802	0.8642	0.8610	0.8670
Refractive index n_D^{20}		1.50449	1.49712	1.49575	1.49588
Heat of fusion,	kJ/mol	13.78	11.56	17.02	9.16
	kJ/kg	129.8	108.9	160.4	86.2
Explosion limits in air,	vol %				
lower		1.0			
upper		5.3 – 7.6			
Ignition temperature,	°C	496 – 502			

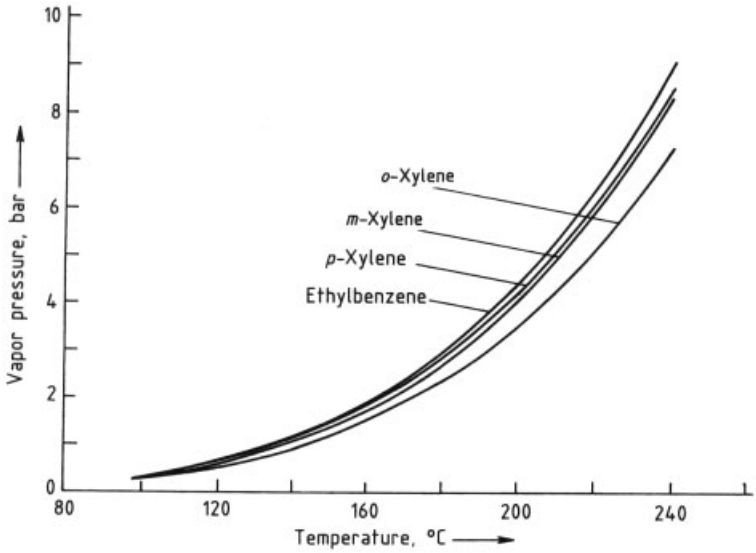

Figure 1. Vapor pressure curves of the xylene isomers and ethylbenzene at 100 – 240 °C

Table 2. Binary azeotropes

Component A	Aromatic	Azeotrope	
		bp, °C	Proportion of A, %
Water	*m*-xylene	94.5	40
	ethylbenzene	33.5 (80 mbar)	33
Methanol	*p*-xylene	64.0	5
Butanol	ethylbenzene	115.85	65
Isobutanol	ethylbenzene	125.7	49
n-Hexanol	*o*-, *m*-, *p*-xylene	138 – 143	18 – 13
Ethylene glycol	*o*-xylene	135.7	7
1,2-Propanediol	*o*-xylene	135.8	10
Glycol	*o*-, *m*-, *p*-xylene and ethylbenzene	139 – 133	14
4-Heptanone	*m*-xylene	139	10
Formic acid	*m*-xylene	94	70
	o-xylene	95.5	74
Acetic acid	*o*-, *m*-, *p*-xylene and ethylbenzene	114 – 116	76 – 66
Propionic acid	*o*-, *m*-, *p*-xylene and ethylbenzene	131 – 135	42 – 28
Phenol	*m*-xylene	133	18
Nonane	*o*-xylene	139	10

Table 3. Temperature dependence of the absorption coefficients of a technical-grade xylene (in m^3/m^3 xylene, STP)

T, °C	H_2S	CO_2	CO	H_2	NH_3	Methane	Ethylene	Propene
0	25	1.53	0.30	0.013	8.0	0.6	4.0	24.0
−10	35	1.62	0.33	0.017	12.4	0.68	4.4	33.0
−20	50	1.71	0.36	0.02	18.6	0.78	4.9	46.5
−30	72	1.85	0.39	0.024	27.7	0.92	5.5	66.0
−40	110	2.01	0.44	0.029	42.2	1.08	6.3	95.0
−50	174	2.21						

Of the chemical properties of ethylbenzene as a component of the A_8 fraction, the most important is catalytic dehydrogenation to styrene, although ethylbenzene used for this purpose mainly comes from other sources (alkylation of benzene with ethylene).

Properties as a Motor Fuel Component. The properties of motor fuels are determined by a large number of quality aspects whose minimum requirements are mostly specified (see → Automotive Fuels). The value which is attributed to xylenes in blending of motor fuels essentially depends on how the xylenes can be combined individually with other available fuel components to produce motor fuel qualities which conform to market requirements. Typically the high knock rating, particularly as a blend component, and the low vapor pressure of xylenes are important factors affecting their use as motor fuel components. Table 4 compares the octane number and vapor pressure values (RVP) of the xylene isomers [14, 15] with the specification of a typical gasoline.

A more detailed account of the overall much more complex interactions with the pool of gasoline components can be found in Section 5.1.

3. Occurrence and Raw Materials

3.1. Natural Occurrence

Practically the only natural source of xylenes is petroleum. The concentration of xylenes varies considerably depending on location and geological age of the crude oil. Table 5 shows a comparison of the C_8 aromatics and the $C_6 - C_8$ aromatics contents in various types of crude oil [16].

Although in a few cases petroleums (e.g., that of South East Asian origin) can have an even higher aromatics content (up to 35 %), the direct isolation of xylenes is not economical.

Table 4. Properties of xylene isomers as motor fuel components compared with finished gasoline

	o-Xylene	*m*-Xylene	*p*-Xylene	Ethylbenzene	Super unleaded (DIN 51 607/EN 228)
RON (pure)	107.4	117.5	116.4	113	
MON (pure)	101.5	110	109	98	
RON (typical blending value)	120	145	146	124	95 ∗
MON (typical blending value)	103	124	127	107	85 ∗
RVP, bar (pure)	0.028	0.025	0.04	0.045	0.9 (Winter) ∗∗ 0.7 (Summer) ∗∗

∗ Minimum values.
∗∗ Maximum values.

Table 5. Aromatics content of various crude oils [15]

	Libya	Louisiana Gulf	West Texas	Venezuela	Nigerian	Iranian
C_8 aromatics, wt %	0.56	0.50	1.10	1.10	1.47	1.05
$C_6 - C_8$ aromatics, wt %	1.0	1.10	1.79	1.85	2.50	1.80

3.2. Raw Materials for Xylene Production

For the industrial extraction of xylenes, the oil fraction or coal used as the raw material is subjected to thermal or catalytic treatment in which aromatics- and xylene-containing fractions are obtained. These conversion processes determine the quantity of xylenes available and supply them in a form that is enriched sufficiently to make isolation and further processing economical.

Raw Materials based on Petroleum.

Reformates. Xylenes are mainly isolated from reformates [17]. Naphtha fractions are used as the raw materials for the reforming process, which is now practically only carried out catalytically. Because of the sulfur sensitivity of the Pt – Re catalysts, the naphtha must first be desulfurized. The composition of the reformates and thus the xylene content depends on how the reforming process is carried out (see Section 4.1) and on the quality of the naphtha fraction used. The composition of the naphtha fraction, in particular its naphthene and aromatics content, significantly affects the composition of the reformate [18].

Naphthenic crude oils give naphtha fractions which are particularly suitable for reforming and aromatics production. However, naphtha fractions produced by hydrocracking (see → Oil Refining, Chap. 3.5.) also generally have high naphthene contents [19] and are therefore suitable starting materials for xylene production by reforming. Besides the composition, the boiling range of the naphtha fraction affects the yield of xylenes (see Table 6). A particularly high yield can be obtained when the naphtha fraction already has a high proportion of the corresponding xylene precursors (e.g., C_8 aromatics or C_8 naphthenes) in a suitable boiling range [20].

Pyrolysis Gasoline. The yield of pyrolysis gasoline and its xylene content is decisively affected by the raw materials used, and by the mode of operation of the plant (see Chap. 4). In steam cracking very different types of raw material can be used, which give rise to widely varying yields and compositions of the pyrolysis gasoline (Table 7) [21].

Only on using higher boiling hydrocarbons feedstocks can appreciable xylene yields and concentrations in pyrolysis gasoline be attained. Therefore, in general, isolation of xylenes is only economical when using these types of raw materials.

Raw Materials based on Coal. The isolation of xylenes from coke-oven tar, coke-oven benzole, or from hydrocarbon mixtures produced by hydrogenation of coal is of minor importance [17] (see also Section 4.1). The effect of the composition of the coal is described in [22]. Processes in which xylene-rich aromatics fractions can be produced from LPG or methanol are described in Chapter 4.

Table 6. Dependence of reformate composition on the boiling range of reformer feedstock (Kuwaitnaphtha) [20]

	Naphtha boiling range, °C		
	60 – 160	107 – 160	90 – 160
Feedstock composition, wt %			
Paraffins	69.6	62.2	64.2
Naphthenes	19.5	21.2	22.2
Aromatics	10.9	16.6	13.6
Aromatics composition of reformate, wt %			
Benzene	9.3	1.6	5.2
Toluene	21.7	19.0	25.1
C_8 aromatics	20.8	34.3	26.2
C_{9+} aromatics	8.8	15.2	11.2

Table 7. Effect of feedstocks on the yields of pyrolysis gasoline and C_8 aromatics from steam cracking (conditions: very high severity with C_2 and C_3 recycle)

Feedstock	Yield of pyrolysis gasoline (C_5 – 200 °C), %	Yield of C_8 aromatics, %	Proportion of C_8 aromatics in pyrolysis gasoline, %
Ethane	1.7		
Propane	6.6		
Butane	7.1	0.4	5.6
Medium-range naphtha	18.7	1.8	9.6
Atmospheric gas oil	18.4	2.2	12.0
Vacuum gas oil	19.3	1.9	9.8

4. Production, Separation, and Further Processing

4.1. Production of C_8 Aromatics

Catalytic Reforming. Besides the nature of the raw materials, process engineering parameters also significantly determine the composition of the reformate and its content of C_8 aromatics [23]. The design of the catalytic reformer (see → Oil Refining, Chap. 3.4.) [24], its mode of operation, and the catalyst used are also very important. The three following types of catalytic reformer are in common use:

1) Semiregenerative reformer, in which the noble metal catalyst is periodically regenerated during a production shutdown (typically every 6 to 12 months)
2) Fully regenerative reformer, in which installation of an additional swing reactor allows the alternating regeneration of one reactor while the other continues to operate
3) Continuously regenerative reformer whose fluidized bed catalyst is continuously regenerated during operation

Operating conditions of the different types of reformer are listed in Table 8 [25].

Since at least six reactions take place in parallel during the reforming process, of which essentially only two lead directly to the formation of aromatics, establishing optimal operation conditions is very important. The reactions which give aromatics (the dehydrogenation of naphthenes and the dehydrocyclization of paraffins) are favored by comparatively high reactor temperatures and low pressures (see Fig. 2). At the same time a simultaneous increase in undesired side reactions (hydrocracking and coke formation) must be reckoned with, so that while increasing the severity leads to an increase in the aromatics content of the reformate, it also decreases reformate yield [26]. Dealkylation reactions are also of specific importance for xylene production. They can lead to decomposition of the xylenes formed, giving toluene, or benzene and methane, particularly at high temperatures.

A further control parameter for xylene production is the space velocity (LHSV). If throughput is lowered to increase the residence time in the reactor, reformates with high aromatics and xylene contents can be produced [25]. The adjustment of the operating parameters to give a maximum aromatics concentration in the reformate is known as aromizing. Within certain limits the quality of the reformate can also be in-

Table 8. Typical operating conditions of various reformer types [25]

Parameter	Semiregenerative reformer	Fully regenerative reformer	Continuously regenerative reformer
Pressure, MPa	1.5 – 2.5	0.7 – 1.5	1
Operating temperature, °C	510 – 540	510 – 540	510 – 540
LHSV ∗, h^{-1}	2 – 3.5	3.5 – 4	1.5 – 4

∗ Liquid hourly space velocity.

fluenced by the choice of catalyst. Besides conventional catalysts, which essentially only contain Pt as the active substances, bimetallic and multimetallic catalyst systems, additionally containing Re, Ge, Ir, or Pb, are used. For aromatics production the selectivity for C_{6+} aromatics is of general importance, while the suppression of dealkylation reactions is particularly important for xylene production. The activity of the catalyst can be increased by regular addition of chlorinated hydrocarbons. Major suppliers of reforming catalysts designed for xylene production include Acreon, Criterion, Exxon, Kataleuna, and UOP [27].

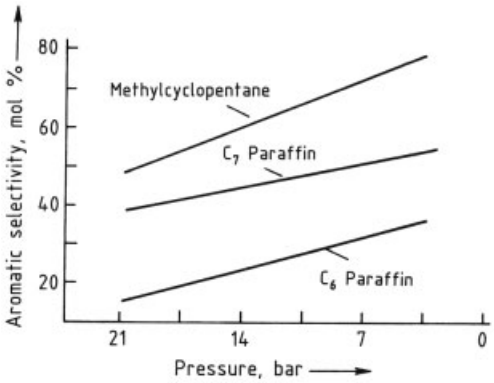

Figure 2. Variation of the selectivity of aromatics production in catalytic reforming with operating pressure

Steam Cracking. In contrast to the reforming process, the xylenes in the pyrolysis gasoline produced by steam cracking are regarded as byproducts of ethylene and propene production. However, their isolation can improve the economics of the process under certain circumstances. Although the operating conditions are not adjusted to maximize the xylene yield in steam cracking, the effect of severity on xylene formation is significant. Increasing the severity – i.e., raising the reaction temperature, lowering the residence time, and reducing the partial pressure by increasing the quantity of steam – significantly decreases the yield of pyrolysis gasoline [28], but increases the concentration of xylene (see Table 9) [29].

Synthesis Processes. In the Cyclar process developed by UOP/BP, propane and butane are converted into aromatics-rich gasoline fractions by cyclization over zeolite catalysts [30, 31]. On using propane the gasoline fraction contains ca. 17.3 wt % C_8 aromatics, and with butane ca. 19.8 wt %. The xylene yield is ca. 15 wt % based on the total starting material. A pilot plant for the Cyclar process has been commissioned at the BP refinery at Grangemouth, Scotland. For economic reasons, however, this process has not been put into industrial-scale operation. The same applies to the Mobil Oil process for synthesizing aromatics from alcohols using ZSM5 catalysts [31].

4.2. Separation and Further Processing

4.2.1. Separation of the C_8 Aromatics Fraction

The reformates used to produce C_8 aromatics can be processed such that C_8 and heavier aromatics are produced in the following approximate proportions [31]:

Benzene	13.2 %
p-Xylene	16.5 %
o-Xylene	8.3 %
C_{9+} aromatics	0.9 %

Besides the 38 % aromatics, 51 – 52 % naphtha fractions and 10 – 11 % gaseous products (hydrogen, fuel gases, etc.) are formed. This optimized result involves the following operations:

1) Toluene is dealkylated to benzene.
2) *m*-Xylene is isomerized to *o*- and *p*-xylenes.
3) Ethylbenzene is transalkylated to xylenes.

Table 9. Influence of severity in steam cracking on the yield and composition of pyrolysis gasoline (naphtha feedstock) [28]

Severity (% Ethylene)	Xylene + ethylbenzene (all products), wt %	Xylene + ethylbenzene in pyrolysis gasoline, wt %	Pyrolysis benzene yield, wt %
Low (24.4)	0.9	3.5	24.9
Medium (28.5)	1.6	7.2	22.6
High (33.4)	1.7	10.7	16.3

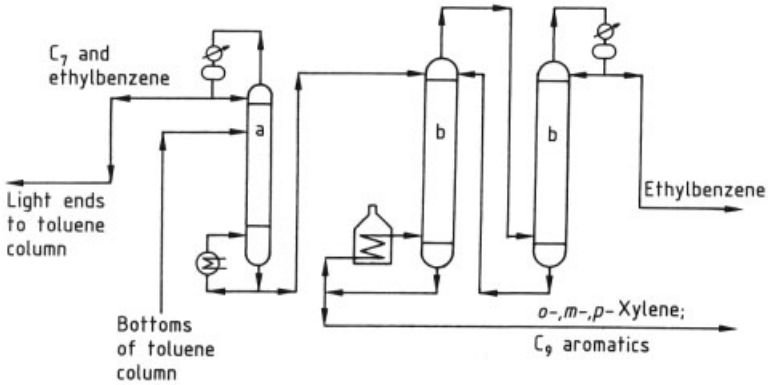

Figure 3. Flow sheet of the ethylbenzene distillation
a) Pre-column; b) Ethylbenzene columns

If the toluene is subjected to transalkylation instead of thermal dealkylation, the yield of xylenes can be further increased at the expense of benzene. The structure of a modern aromatics complex, which produces *o*- and *p*-xylene in the highest possible yield, is based on a combination of a series of these processes.

The ethylbenzene is mainly converted to xylenes because obtaining ethylbenzene from reformate is energy intensive. The main source of ethylbenzene, which is almost exclusively dehydrogenated to styrene, is therefore now the more economical alkylation of benzene with ethylene (see → Ethylbenzene).

Distillative removal of ethylbenzene, which for styrene production must be toluene-free, from reformates can take place after removal of toluene from the sump of the toluene column. Figure 3 shows a flow sheet of the ethylbenzene distillation. Other products obtained by this separation are the xylene and C_9 aromatics. The major separation problem is caused by a boiling point difference of only 2 °C between ethylbenzene and *p*-xylene. The distillative separation of this mixture requires high reflux ratios (1 : 80 to 1 : 120) with 300 – 360 effective plates. A yield of ca. 95 % ethylbenzene with > 99.5 wt % purity is achieved under these conditions [32].

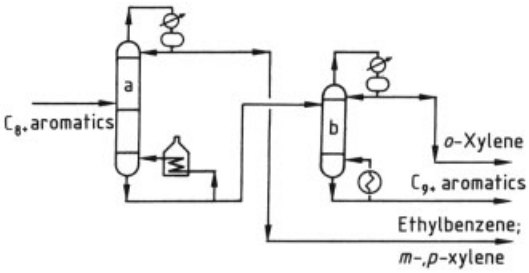

Figure 4. Flow sheet showing separation of *o*-xylene
a) Xylene splitter; b) *o*-Xylene column

The distillative separation of *o*-xylene from the C_{8+} stream is also difficult. Although the boiling point difference is only ca. 5 °C, the required *o*-xylene purity (min. 98 %) can be achieved with 120 – 150 effective plates and a reflux ratio of 1 : 10 to 1 : 15, with a yield of > 95 %. Figure 4 shows a flow sheet for this separation. A further increase in *o*-xylene purity can be achieved by subsequent extractive distillation. Another method of separation is crystallization, mainly used to obtain *p*-xylene. This process utilizes the crystallization behavior of *p*-xylene near the eutectic point of a C_8 aromatics mixture. Depending on the composition of the mixture, this lies between −60 and −68 °C (see Fig. 5).

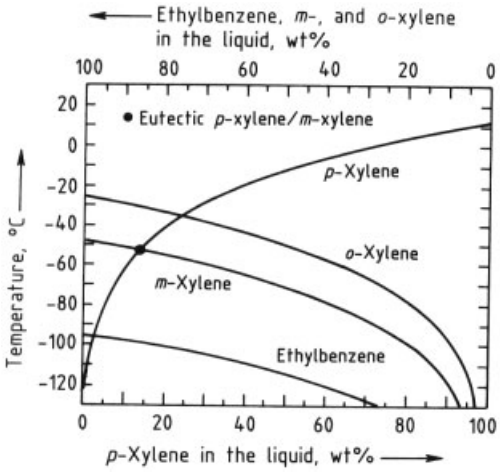

Figure 5. Melting behavior of the C_8 aromatics in their mixtures

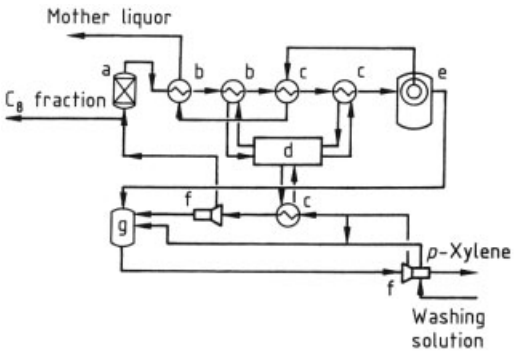

Figure 6. *p*-Xylene crystallization by indirect refrigeration
a) Drier; b) Precooler; c) Scraped-surface crystallizer; d) Refrigeration plant; e) Filter; f) Centrifuge; g) Mixer

The Krupp – Koppers process (Fig. 6) is a typical example of the crystallization process. The actual separation process is a fractional crystallization in which the product of the first separation step in the filter (ca. 65 % of the *p*-xylene from the starting mixture) is mixed with highly concentrated *p*-xylene and is isolated in high purity in a centrifuge. This process and that of Phillips Petroleum [33] operate using indirect refrigeration.

Direct contact refrigeration involves downward flow of the initially cooled liquid in the vertical crystallization vessel. This in turn directly cools the remainder of the contents of the vessel (see Fig. 7).

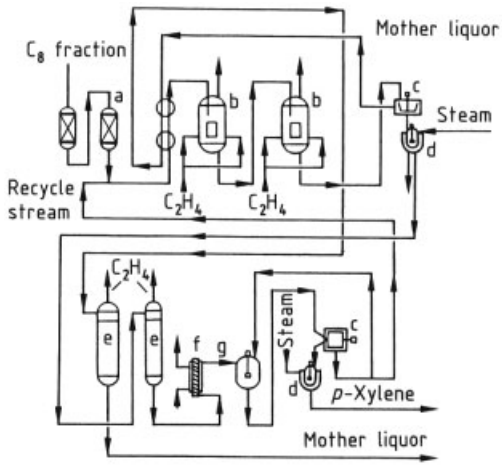

Figure 7. *p*-Xylene crystallization by direct contact refrigeration
a) Drying; b) Crystallization; c) Centrifuge; d) Melting tank; e) Desorber; f) Scraped-surface crystallizer; g) Mixing tank

The cooling agent can be evaporating ethylene (Maruzen Oil Co.) [34]. Chevron Research and Development Corp., Sohio/BP, and Arco Technology use other process designs [35].

In the Parex process (UOP) [36] *p*-xylene is obtained in high purity from the isomer mixture by adsorption on a molecular sieve. The attached finishing column removes lighter components (raffinate consisting of ethylbenzene and a xylene mixture low in *p*-xylene) at the head. The *p*-xylene constitutes the bottom product of this column. Adsorption – desorption on the solid bed is carried out in such a way that a moving bed is simulated (Fig. 8). The rotary valve brings segments of the filled adsorption chamber into contact with the feed and with the desorbent in such a way that part of the molecular sieve adsorbs while another part is desorbed. The rotary valve switches the streams in the direction of flow in the bed. This simulates the movement of the adsorbent in a direction opposite to that of the liquid. The adsorbent exhibits high selectivity towards *p*-xylene compared with the other C_8 aromatics, especially ethylbenzene. Ethylbenzene is therefore one of the most suitable desorbents or eluents; this is kept in circulation in the extract column. The heavy part of the raffinate can also be used as the desorbent. The nonadsorbed C_8 aromatics leave the raffinate column at the head

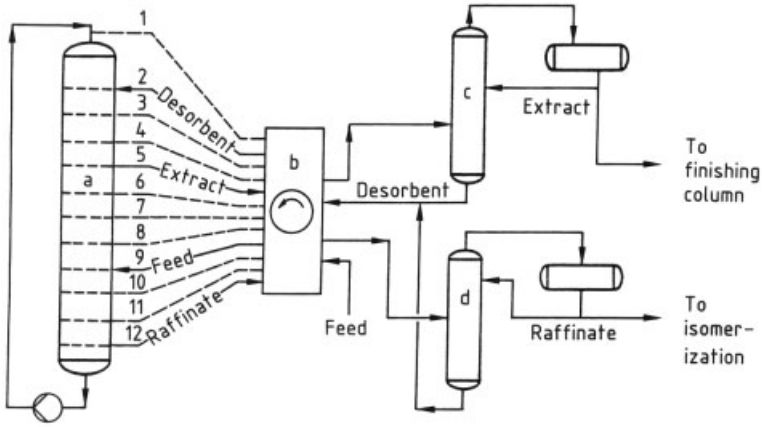

Figure 8. Parex process
a) Adsorbent chamber; b) Rotary valve; c) Extract column; d) Raffinate column

and are then subjected to isomerization (Section 4.2.2).

The Parex process is a variant of the Sorbex group of processes. The first process which used this principle of the simulated moving bed was the Molex process developed in 1964 for the separation of *n*-paraffins from isoparaffins and cyclic hydrocarbons. Other modifications of the Sorbex processes are used to separate olefins and paraffins in $C_3 - C_{18}$ hydrocarbon mixtures (Olex process) and for separating carbohydrates (Sarex process).

4.2.2. Isomerization of the C_8 Fraction

The yield of *o*- and *p*-xylenes can be maximized by using catalytic conversion processes. The Isomar process (Fig. 9) [37] enriches the *p*-xylene in the Parex raffinate by re-establishment of the equilibrium. The acidic metal-containing zeolite catalyst used here also isomerizes ethylbenzene selectively to xylene isomers in their equilibrium ratio [38]. Since hydrogenolysis on the metal components must be guaranteed to maintain the activity of this bifunctional catalyst, the Isomar process is operated with a certain partial pressure of hydrogen in the catalytic reactor. Hydrogen is separated from the effluent from the Isomar reactor and then recycled together with fresh hydrogen. The liquid product is separated in a deheptanizer column. A C_7 stream is obtained at the head, which can be recycled to the starting reformate. The bottom product of this column is a *p*-xylene-enriched stream which, after clay treatment to remove unsaturated compounds is fed to the Parex unit to obtain the newly formed *p*-xylene.

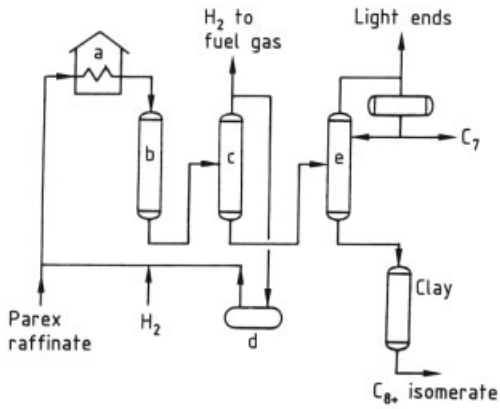

Figure 9. Isomar process
a) Heater; b) Reaction column; c) Hydrogen separation; d) Recycle compressor; e) Deheptanizer column

In the Tatoray process (Fig. 10) toluene and the C_9 (A_9) bottom fractions from the A_9 column (*o*-xylene splitter column) are mainly converted into xylene isomers by disproportionation of the toluene and transalkylation of the C_9 aromatics (→ Toluene, Chap. 3.2.), affording a newly formed $A_6 - A_8$ fraction from A_{9+}. This technology is the product of cooperation between UOP and Toray Industries. Commercial applications have been in operation since 1978. Since then the catalyst has been improved sev-

eral times with regard to activity, thermal stability, and life span [37].

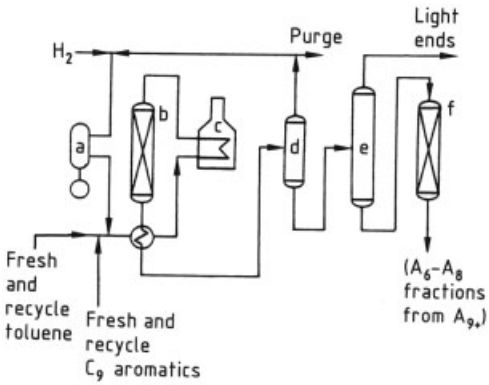

Figure 10. Tatoray process
a) Hydrogen recycle compressor; b) Catalytic reactor; c) Preheater; d) Hydrogen removal; e) Light ends removal; f) Clay tower

The Tatoray process consists of a reactor filled with solid catalyst to which are attached a cooling unit, a gas – liquid separation unit, a unit for distillation of the light components and one for clay treating to remove olefinic impurities. As in the Isomar process a partial pressure of hydrogen is maintained in the reactor. By processing different feed qualities, which can be influenced both by the different recycle rates and a varying charge of fresh reformate fractions, the composition of the product can be controlled.

Other proven disproportionation processes are the LTD process of Mobil Research of Arco Technology. Alkylation of toluene to increase the xylene yield has been developed by Mobil Chemical (Mobil toluene-to-*p*-xylene process = MTPX) and allows *p*-xylene concentrations of ca. 94 % to be achieved [39]. Mobil has announced that it will be using its new MTPX process in two facilities in the United States.

4.2.3. Combination of the Technologies in the Aromatics Complex

The maximization of the yields of *o*- and *p*-xylenes by combining isomerization, transalkylation, disproportionation, and adsorptive extraction of *p*-xylene is exemplified by the UOP technology complex (Fig. 11) [37]. In this complex all processes which are of industrial and commercial importance in this respect are represented. The center of the complex is the Parex process combined with Isomar technology. Tatoray technology is also used for conversion of the toluene and the C_{9+} aromatics fraction (A_9) [40].

Naphtha is used as the feedstock in this complex (cf. Fig. 12). It is obtained by conventional crude oil distillation, sometimes also from an evaporated BTX fraction. The reformer in this case is the UOP – CCR Platformer, which allows particularly high yields of aromatics to be obtained and also provides the hydrogen for subsequent processes, such as the Isomar and Tatoray processes, or for naphtha hydrogenation. The Sulfolane process is a frequently used liquid – liquid extraction, developed by Shell to obtain benzene and toluene from the reformate and from its $C_5 - C_7$-fraction (top product from the deheptanizer; see → Benzene; → Toluene, Chap. 3.). The bottom products from the reformate deheptanizer and from the toluene and post-Isomar deheptanizer columns are fed to the Parex unit. Consequently the products of the complex are *o*- and *p*-xylene, benzene, aromatics-free raffinate, and the C_{9+} aromatics.

5. Integration into Refinery and Petrochemical Complexes

5.1. Backward Integration into Petroleum Refining

Usually both the production and the separation and isomerization of the xylenes are integrated into refinery and petrochemical complexes [41, 42]. There are many advantages to this type of incorporation [43]:

1) Availability of a very favorable raw material base
2) Improved economics, especially in the use of byproduct streams unavoidably produced, such as pyrolysis gasoline or raffinates
3) Use of common infrastructure (energy, steam)
4) Possibilities for further processing of xylenes on site
5) Alternative use as fuel components if economical

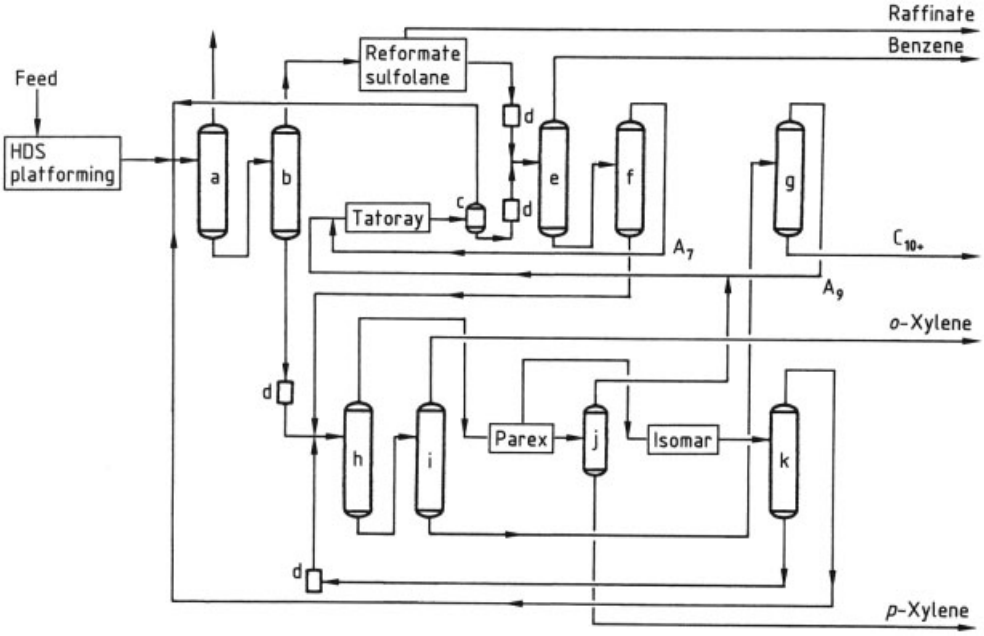

Figure 11. Flow sheet of the UOP aromatics complex
a) Debutanizer; b) Platformate deheptanizer; c) Splitter; d) Clay tower; e) Benzene column; f) Toluene column; g) A_9 column; h) Xylene stripper; i) *o*-Xylene re-run; j) Parex finishing column k) Isomerate deheptanizer

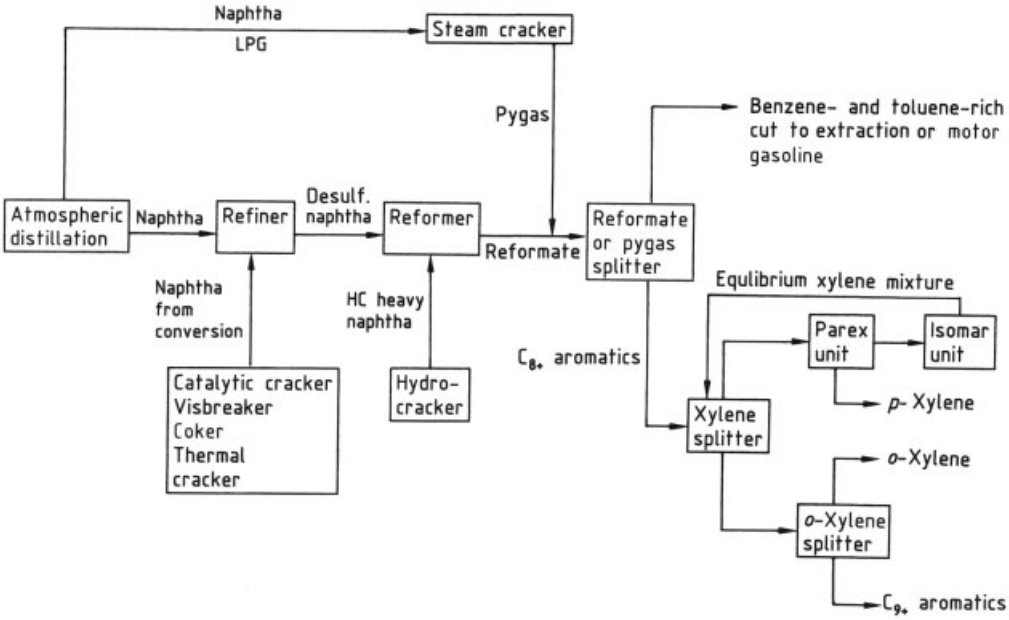

Figure 12. Integration of xylene production into petrochemical refineries
FCCU = Fluid catalytic cracking unit

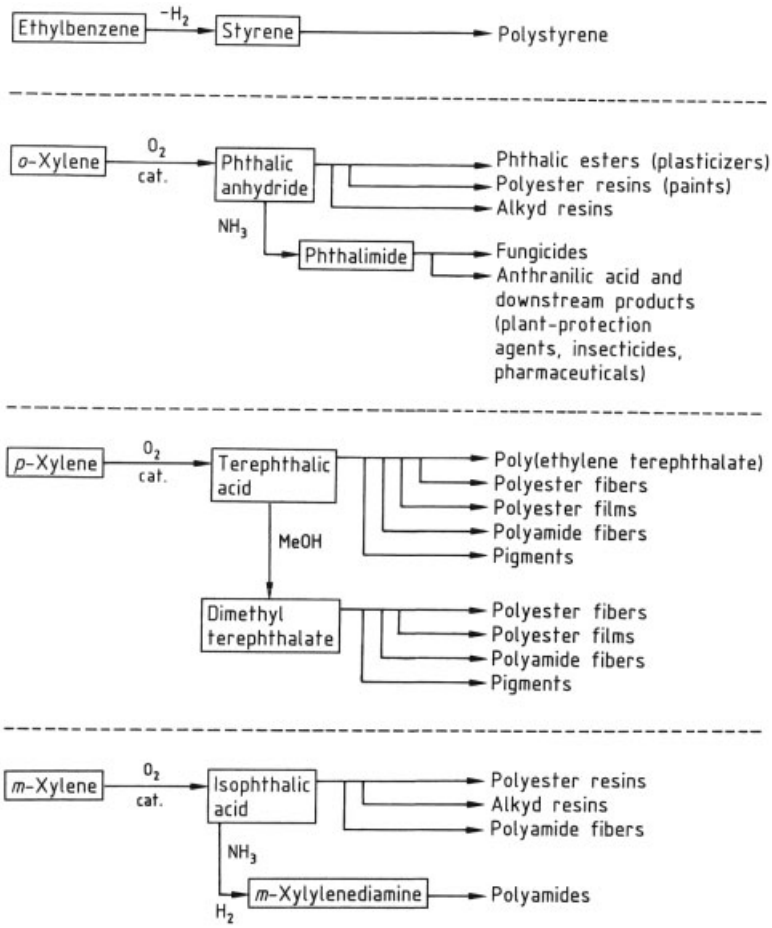

Figure 13. Most important further processing steps and downstream products of xylene isomers and ethylbenzene

Table 10. Western European producers and product capacities for xylene isomers and primary downstream products if produced at the same site (1994, in 10^3 t/a)

Location	Producer	*o*-Xylene	*p*-Xylene	*m*-Xylene	Ethylbenzene from C_8 aromatics	Phthalic anhydride	Terephthalic acid	Dimethyl terephthalate
Geel, Belgium	Amoco			30			250	
Gonfreville, France	Total	90	105					
Schwedt, Germany	PCK	35	25				62	
Gelsenkirchen, Germany	Ruhr Oel	45	105					
Heide, Germany,	RWE/DEA	12			10			
Wesseling, Germany	RWE/DEA	44	110					
Godorf, Germany	Shell	100	75					
Priolo, Italy	Enichem	60	140		70			
Sarroch, Italy	Saras	60	70		50			
Milan, Italy	Sisas			50		45		
Botlek, The Netherlands	EXXON	105	165			50		
Oporto, Portugal	Petrogal	42	115					
Alqeciras, Spain	Cepsa Interquisa	40	34			20	200	70
Wilton, United Kingdom	ICI		330		50		580	

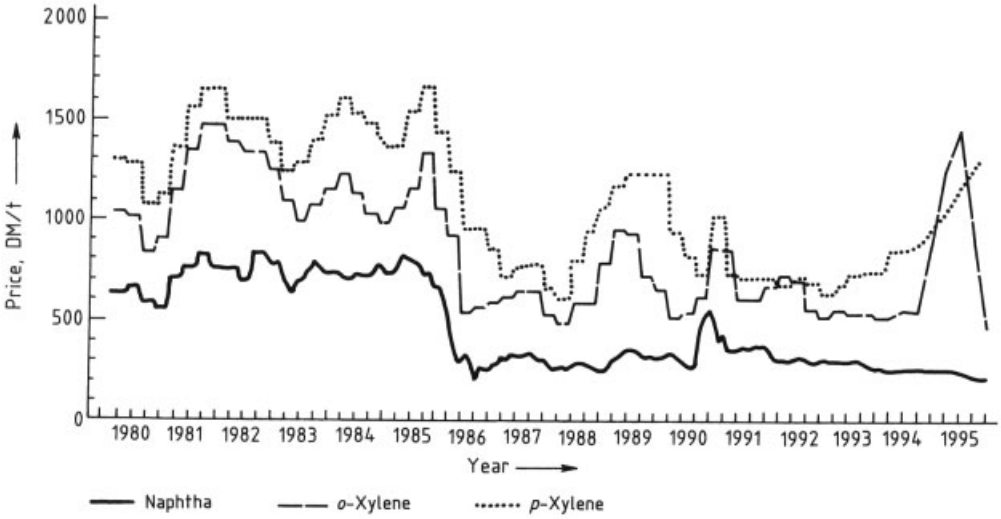

Figure 14. Development of prices of naphtha and xylenes in Western Europe

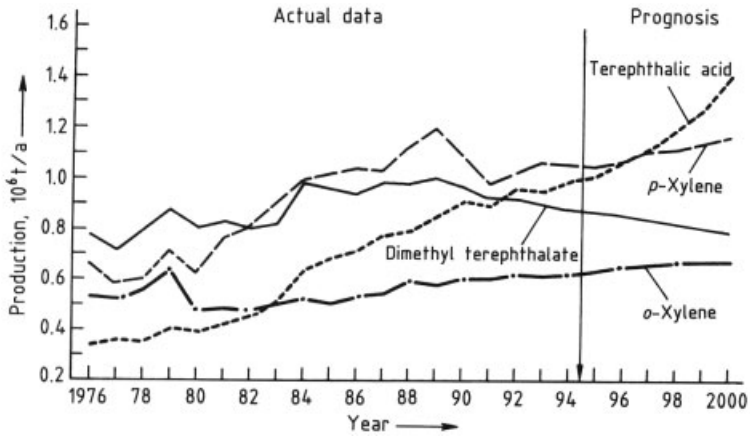

Figure 15. Development of quantities of *o*- and *p*-xylene and primary downstream products produced in Western Europe

The integration of xylene production into the structure of a petrochemical refinery is shown schematically in Figure 12.

The separation of xylenes and other aromatics from reformate or pyrolysis gasoline frequently leaves a gap in the boiling curve of the motor fuels produced in the refinery. This phenomenon is typical of petrochemical refineries and is known as the gap fuel characteristic. To comply with motor fuel specifications it is sometimes necessary to add alternative fuel components in the same boiling range [44]. With regard to a possible future limitations on the total aromatics content of motor fuels, as is being discussed in the United States [4], the isolation of xylenes and other aromatics could, however, also help to fulfill the specification requirements [45].

5.2. Forward Integration into Chemical Processing

Apart from their use as solvents, the individual xylene isomers can be regarded as intermediates which are further processed in a series of reactions to give consumer products [5]. While immediate subsequent synthesis steps, such as oxidation of *o*- and *p*-xylenes to phthalic anhydride and terephthalic acid, can be integrated into petrochemical complexes to some extent,

further processing steps are mainly carried out at purely chemical sites.

The most important further processing steps and downstream products of the various xylene isomers and of ethylbenzene are shown schematically in Figure 13.

6. Economic Aspects

Parameters Affecting Economics. The economics of xylene production and isolation are essentially determined by two parameters [3]:

1) The price difference between the proceeds of xylene production and the cost of the naphtha used (see Fig. 14)
2) The value relative to that when used as a motor fuel component

The first aspect is the most important when targeted xylene production by reforming naphtha without the possibility of the alternative use in motor fuel is considered. However, the second aspect determines the economics of the separation of xylene mixtures, which can originate from reformate or pyrolysis gasoline, if the alternative option of use as a motor fuel component is available. Besides these aspects, other technical parameters, such as the performance of the reformer, the efficiency of the separating and isomerization plants, or the quality of the naphtha, affect the overall economics.

The naphtha/xylene price difference is also controlled by other, overriding factors. While the price of naphtha is closely linked to that of petroleum, the proceeds of xylene production are principally determined by the price and demand for downstream products.

From the historical and expected future development of the quantities of direct downstream products of *p*-xylene produced in Western Europe (see Fig. 15) [3], it may be deduced that conversion to purified terephthalic acid (PTA) will be increasingly favored over dimethyl terephthalate (DMT). On balance this leads to an increased overall demand for *p*-xylene which may essentially be attributed to the increasing demand for poly(ethylene terephthalate) resins for the production of plastic bottles.

In the past the primary downstream product of *o*-xylene, phthalic anhydride, was produced in significant quantities by oxidation of coal-tar naphthalene [3]. Now phthalic anhydride is mainly produced from *o*-xylene so the consumptions of these products are directly associated. They are only increasing slightly because of the relatively low growth rate of plasticizer production.

The main downstream product from ethylbenzene, styrene, is mainly converted to polystyrene and ABS elastomers. Correspondingly, the demand for these products determines the demand for ethylbenzene. Most of the ethylbenzene, however, is not produced within the framework of xylene production, but from benzene and ethylene [46].

The economic importance of *m*-xylene and its primary downstream product, isophthalic acid, is comparatively low. Regional demand for mixed xylenes in 1995 broke down as follows (total demand: 16×10^6 t):

Far East	44 %
North America	28 %
Western Europe	15 %
Eastern Europe	7 %
South America	3 %
Middle East	3 %

Producers. The most important producers and their production capacities for the various xylene isomers in Western Europe are listed in Table 10 [47]. Only in a few cases are additional capacities for conversion to phthalic anhydride or terephthalic acid and dimethyl terephthalate available at the same production site. The largest capacities for mixed xylenes are in North America and the Asia/Pacific region. Production capacities (in 10^3 t) worldwide for mixed xylenes follow [48]:

North America	
United States	
Amoco Chemical	680
Exxon Chemical	625
Chevron Chemical	573
Koch	550
Phillips Chemical	416
Lyondell	314
Fina	311
Sun	270
BP Chemicals	198
Shell Chemical	198
Ashland Chemical	182
Mobil Chemical	152
Coastal	109
South Western Refining	98

Citgo	68
OxyChem	66
Phibro Energy	45
Marathon Oil	43
Uno-Ven	33
Canada	
Kemtec	91
Sunoco	86
Petro-Canada	73
Novacor Chemicals	59 (toluene – xylene mixture)
Shell Canada	32
Imperial Oil	14
South America	
Argentina	
Pasa	80
YPF	80
Brazil	
Petroquimica do Nordeste	250
Petroquimica do Sul	68
Petrobras	25
Petroquimica União	96
Chile	
Corpoven	50
Europe	
Belgium	
Finaneste	35
Germany	
DEA Mineraloel	242
Deutsche Shell	200
Ruhr Oel	175
PK Schwedt	138
Italy	
EniChem	200
Edison Spa	185
The Netherlands	
Exxon Chemical	285
Total	200
Portugal	
Petrogal	185
Spain	
Petresa	216
United Kingdom	
ICI	345
Asia/Pacific	
Japan	
Idemitsu Petrochemical	550
Tonen Chemical	540
Nikko	510
Mitsubishi Oil	420
Showa Shell Sekiyu	370
Nippon Petrochemicals	340
Koa Oil	250
Fuji Oil	190
General Seikyu	190
Maruzen Petrochemical	170
Kashima Oil	120
Nippon Steel Chemical	115
Mitsubishi Kagaku	100
Mitsui Petrochemical	100
Kyushu Aromatics	90
South Korea	
Yukong	500
Honam Oil	365
Ssangyoung Oil	360
Daelim Industrial	95
Singapore	
Singapore Petrochemical	30
Shell Eastern Chemicals	20
Taiwan	
Chinese Petroleum Corp.	750

7. Quality Specifications and Analysis

Specifications for xylenes are generally laid by a particular company, depending on the intended application. Properties relevant to specification are determined by standard procedures [49].

Some important methods of determination for xylene follow:

1) Distillation range (e.g., DIN 51 761, ASTM D 850)
2) Hydrogen sulfide and sulfur dioxide content (e.g., ASTM D 853)
3) Thiol sulfur (e.g., DIN 51 765)
4) Density (e.g., DIN 51 757, ASTM D 891)
5) Flash point (e.g., DIN 51 755, ASTM D 56)
6) Color (e.g., ASTM D 156, ISO 6271)
7) Bromine consumption (e.g., DIN 51 774)
8) Purity/composition (e.g., ASTM D 1016; GLC determination)
9) Residue on evaporation (e.g., EN 5).

Typical specifications for *o*- and *p*-xylenes are given in Tables 11 and 12.

Table 11. Typical product specification for o-xylene

	Method	Value
o-Xylene (min.), wt %	GC	98.0
C_9 and higher aromatics (max.), wt %	GC	0.5
Nonaromatics (max.), wt %	GC	0.5
Density at 15 °C, kg/m^3	DIN 51 757	882 – 885
APHA color (max.)	ISO 6271	10
Copper corrosion (max.)	DIN 51 759	1
Hydrogen sulfide and sulfur dioxide (max.)	ASTM D 853	free of H_2S and SO_2
Acidity	DIN 51 558 T.1	not detectable
Start of boiling (min.), °C	DIN 51 761	143.0
End of boiling (max.), °C	DIN 51 761	145.0
Residue on evaporation (max.), mg/kg	EN 5	20

Table 12. Typical specification for p-xylene

	Method	Value
p-Xylene (min.), wt %	GC	99.4
C_8 aromatics (max.), wt %	GC	0.55
Other aromatics (max.), wt %	GC	0.05
Nonaromatics	GC	0
Density at 15 °C, kg/m^3	DIN 51 757	865 – 866
Copper corrosion (max.)	DIN 51 709	0
Total sulfur (max.), ppm	DIN 51 768	10 – 15
Acidity	DIN 51 558 T.1	not detectable
Distillation range, °C	DIN 51 761	1
mp, °C		13.1

The high purity requirement for *p*-xylene are due to the quality requirements for the production of pure terephthalic acid. Provided that further processing takes place via dimethyl terephthalate, sometimes somewhat lower degrees of purity can be tolerated. For *m*-xylene, which is produced in comparatively small quantities, only purities of 95.4 wt % min. are usually required.

8. Storage, Transport, and Safety

Instructions for ensuring safe storage, transfer, and transport of flammable liquids can be found in the technical rules for flammable liquids (TRbF), drawn up by the German committee for flammable liquids [50, 51]. For xylene the following TRbF numbers are important:

TRbF 001:	General, form, and application of the rules
TRbF 100:	General safety requirements
TRbF 111:	Filling stations, emptying stations
TRbF 112:	Gasoline stations
TRbF 120/121:	Tanks in a fixed location
TRbF 131/1 and 2:	Piping and tubing
TRbF 141:	Tanks on vehicles
TRbF 142:	Tank containers
TRbF 143:	Movable vessels
TRbF 180:	Operating instructions

The flammability limits for xylenes in air are between 1 and 8 %. Because the flash point of xylene mixtures is below 21 °C, they are mostly assigned to hazard class A I. The flash points of pure xylenes (not including ethylbenzene) are, however, > 21 °C so they may be declared as hazard class A II. For storage and transport xylenes must be labeled as readily flammable and harmful [51, 52]. For transport within Germany xylenes are assigned to class 3, cipher 31 c of the GGVS/GGVE regulations (regulations governing transport of hazardous goods by road and rail) [53].

9. Environmental Aspects and Toxicology

Environmental Aspects. In Germany xylenes are assigned to water hazard class 2 (WGK 2) [55]. The solubility of xylenes in water is low (ca. 0.14 g/L). Because of the comparatively low vapor pressure (7 – 8 mbar at 20 °C), the danger of vapor emissions in air is relatively low. However, xylenes can react with other air pollutants to give environmentally damaging products. This applies particularly to the UV-catalyzed photooxidation of xylenes by nitrogen oxides. However, compared with some other hydrocarbons, the reactivity of xylenes is comparatively low, with a reaction rate of ca. 2×10^{-9} min^{-1} [55]. For example, disubstituted internal olefins have a reaction rate of ca. 50×10^{-9} min^{-1}.

Toxicology. Exposure to xylene is possible through inhalation of vapors and resorption through the skin. The MAK value is 100 ppm (ca. 440 mg/m^3) for all three isomers [54].

General Activity. Xylene exhibits acute prenarcotic and narcotic activity. Chronic exposure to xylene leads to disturbance of the CNS (e.g., headaches, sleep disturbance) [56] and damage to the blood picture (dyspepsia). Provided that there is no long-term chronic overexposure, these effects are reversible. Besides developing a certain tolerance, frequent exposure to xylene can also lead to habituation or even addiction (solvent abuse).

Acute Toxicity. The LD_{50}, LC_{50}, and TCLo values for oral administration and inhalation vary widely, depending on the animal investigated and the isomer composition. The following values give an indication of the toxicity of xylene isomer mixtures [54, 57]: LD_{50} (rat, oral) 4300 mg/kg, LDL_0 (rat, i.p.) 2000 mg/kg, TCLo (humans, inhalation) 200 ppm.

A very high exposure to xylene of ca. 10 000 ppm caused by an accident led to lung edema and subsequent death in one person [58]. In other cases severe damage to the CNS, kidneys, and liver were observed.

Irritant Effects. On repeated application xylenes can cause irritation of the respiratory passages and mucous membranes of the eye. Frequent skin contact can lead to blister formation and dermatitis [57, 59–62].

Subchronic and Chronic Toxicity. At a concentration range of 100 to over 1000 ppm xylene in inhaled air, damage to the CNS with disturbance of balance or slowing of reactions is observed. Inhalation of xylene vapors can also cause nausea and headaches [63–65].

A change in the blood picture is also frequently observed. In ca. 10 % of persons who had been exposed to xylene vapors in concentrations of up to ca. 100 ppm for ca. 5 years, a leucocyte level of $< 4500\,\mathrm{mm}^{-3}$ was established [66]. A decrease in the immunobiological activity has occasionally been observed [58].

Carcinogenicity, Mutagenicity, and Embryotoxicity. The assessment of the carcinogenicity of xylenes is not consistent [67–69]. While in one study no indications of carcinogenicity or cocarcinogenicity were found, another investigation indicated that xylenes act as tumor promoters for skin tumors in rats.

None of the three xylene isomers showed mutagenicity in the Ames test [70]. However, slight mutagenicity was detected in a recessive lethal test on *drosophila* [71].

In animal experiments long-term exposure through inhalation of xylenes causes small changes to fetuses [72].

Pharmacokinetics and Metabolism. All three xylene isomers are resorbed in the same way. Various investigations have shown that ca. 60–70 % of the xylene reaching the organism via the lungs is retained [73–76]. This percentage remains approximately constant over the whole exposure period. The ratio of the concentration in the air in the alveoli ($\mathrm{mg/m^3}$) and in the blood (mg/kg) changes with the degree of bodily activity. When the body is at rest the ratio is ca. 15 : 1 and when moving ca. 30–40 : 1 [77]. Xylene can be resorbed at a rate of ca. 2.5 (0.7–4.3) $\mathrm{mg/m^3}$ per minute through intact skin [78].

Xylenes are deposited rapidly in body fat (up to ca. 5 %) and remain there for hours after exposure. The half-life in fat deposits is ca. 0.5–1.0 h [78]. The metabolism of the individual xylene isomers is identical. The main biotransformation pathway initially involves oxidation to methylbenzoic acid, which forms the corresponding methylhippuric acid by conjugation with glycine (A) [74, 79–81]. The methylhippuric acid can be excreted rapidly via the kidneys. Another, less favored biotransformation pathway involves the hydroxylation of the xylene on the aromatic ring, forming xylenols (B) [73].

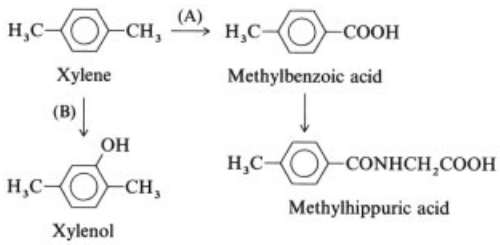

10. References

1. Parpinelli Technon: "West European Petrochemical Industry," Aromatics Report and Tables, Mailand 1995, pp. 18, 31, 38.
2. J. S. McArragher: Paper to the Institute of Petroleum Conference 16 Oct. 1991: "Making Cleaner Fuels in Europe – Their Need and Cost," Institute of Petroleum, London 1991.
3. Chem Systems Ltd.: "Petroleum and Petrochemicals Economics," annual reports 1 and 2, London, Aug. 1995.
4. *Octane Week,* July 19 (1993) 6.
5. K. Weissermel, H. J. Arpe: *Industrielle Organische Chemie,* 3rd ed., Verlag Chemie, Weinheim 1988, pp. 407, 426.
6. *Ullmann,* 4th ed., **24,** 526.
7. C. L. Yaws, *Chem. Eng. (N.Y.)* **82** (1975) no. 15, 113–122; **82** (1975) no. 20, 73–81.
8. *Beilstein,* 3rd ed., suppl., vol. 5, 2nd part (1964) 807–823, 823–845, 845–864, 776–807. W. Berghoff: *Erdölverarbeitung und Petrolchemie,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1968, pp. 80–81. W. L. Nelson: *Petroleum Refinery Engineering,* 4th ed., McGraw-Hill, Toronto 1958. H. J. V. Winkler: *Der Steinkohlenteer und seine Aufarbeitung,* Verlag Glückauf, Essen 1951, pp. 70–71.
9. F. D. Rossini et al.: *Selected Values of Physical and Thermodynamic Properties of Hydrocarbons and Related Compounds,* Carnegie Press, Pittsburgh, PA, 1953.

10. L. H. Horsley, *Anal. Chem.* **19** (1947) no. 8, 508 – 600.
11. L. H. Horsley, *Anal. Chem.* **21** (1949) no. 7, 831 – 873.
12. L. H. Horsley, *Adv. Chem. Ser.* **35** (1962).
13. *Beilstein,* 3rd ed., **suppl., vol. 1, 2nd part** (1964) 811 (o-Xylol), 828 (m-Xylol), 849 (p-Xylol), 782 (Ethylbenzol).
14. K. Owen, T. Coley: *Automotive Fuels Handbook,* Society of Automotive Engineers, Warrendale, PA, 1990, pp. 564 – 565.
15. F. E. Condon, *J. Am. Chem. Soc.* **70** (1948) 1963.
16. H. G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, pp. 60 – 61.
17. Parpinelli Technon: "West European Petrochemical Industry", Aromatics Report and Tables, Mailand 1993, p. 231.
18. J. A. Weiszmann: "The UOP Platforming Process," in R. A. Meyers (ed.): *Handbook of Petroleum Refining Processes,* Chap. 3, McGraw-Hill, London 1986, p. 6.
19. D. G. Tajbl: "UOP HC Unibon Process for Hydrocracking," in R. A. Meyers (ed.): *Handbook of Petroleum Refining Processes,* Chap. 2, Mc-Graw-Hill, London 1986, p. 41.
20. A. Chauvel, G. Lefebre in: *Petrochemical Processes,* vol. 1, Editions Technip IFP, Paris 1989, p. 176.
21. A. Chauvel, G. Lefebre in: *Petrochemical Processes,* vol. 1, Editions Technip IFP, Paris 1989, p. 130.
22. H. G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, pp. 35 – 57.
23. J. A. Weiszmann: "The UOP Platforming Process," in R. A. Meyers (ed.): *Handbook of Petroleum Refining Processes,* Chap. 3, McGraw-Hill, London 1986, pp. 11 – 13.
24. N. L. Gilsdorf, A. P. Furano, R. H. Rachford, R. J. Schmidt, D. L. York: "Platforming Technology for Aromatics Production," in: *Petrochemical Technology: Building Success through Teamwork and Technology,* UOP Technology Conference, Des Plaines, Ill., 1992, pp. 20 – 21.
25. A. Chauvel, G. Lefebre in: *Petrochemical Processes,* vol. 1, Editions Technip IFP, Paris 1989, pp. 170 –172.
26. J. A. Weiszmann: "The UOP Platforming Process," in R. A. Meyers (ed :): *Handbook of Petroleum Refining Processes,* Chap. 3, McGraw-Hill, London 1986, pp. 6 – 9.
27. *Oil Gas J.,* 2nd Oct. (1995) 35 – 37.
28. W. Keim, A. Behr, G. Schmitt in: *Grundlagen der Industriellen Chemie,* Verlag Salle-Sauerländer, Frankfurt 1986, p. 129.
29. A. Chauvel, G. Lefebre in: *Petrochemical Processes,* vol. 1, Editions Technip IFP, Paris 1989, p. 156.
30. C. D. Gosling, G. L. Gray, P. J. Kuchar, L. Sullivan: "Produce BTX from LPG," in: *Building Success through Teamwork and Technology,* UOP Technology Conference, Des Plaines, Ill., 1992, pp. 1 – 17.
31. H. G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, pp. 88 – 91.
32. E. C. Haun, M. W. Golem, P. P. Piotrowski, S. Sapuntzakis: "The Modern Aromatics Complex and Revamp Options for Existing Plants," UOP Technology Conference, Des Plaines, Ill. 1988.
33. J. G. Jenkins, *Oil Gas J.* **63** (1965) Jan. 18, 78 – 86.
34. N. E. Ockerbloom, *Hydrocarbon Process* **50** (1971) July, 113.
35. J. Morrison, *Oil Gas Int.* **10** (1970) no. 12, 67 – 69; *Hydrocarbon Process.* **60** (1981) Nov., 240.
36. Chevron Res., US 3 219 722, 1965. ARCO Techn., *Hydrocarbon Process.* **60** (1981) Nov., 239.
37. J. A. Johnson, A. P. Furfaro, S. H. Hobbs, R. G. Kabza: "Advances in para-Xylene Recovery," UOP Technology Conference, Des Plaines, Ill. 1988.
38. A. K. Aboul-Jheit, *Erdöl Erdgas Kohle* **111** (1995).
39. *Chemical Week,* Aug. 30, 1995, p. 18.
40. P. J. Kuchar, O. Y. Lin, V. Zukauskas, D. Brkic: "Isomar and Tatoray Aromatics Production Flexibility," UOP Technology Conference, Des Plaines, Ill. 1988.
41. L. Grimm, *Erdöl Erdgas* **108** (1992) no. 9, 355.
42. H. Forth, *Oel* **19** (1981) no. 7, 172 – 178.
43. B. Hnat: *Wirtschaftspolitische Studien,* **vol. 87,** Verlag Vandenhoek and Ruprecht, Göttingen – Zürich 1992, pp. 69 – 71.
44. K. Owen, T. Coley: *Automotive Fuels Handbook,* Society of Automotive Engineers Inc., Warrendale, PA, 1990, pp. 144 – 145.
45. UOP: "The Clean Air Act And The Refining Industry," Des Plaines, Ill., 1 Sep. 1991, p. 17.UOP: "Process Solutions for Reformulated Gasoline," Des Plaines, Ill., 1992.

46. Parpinelli Technon: "West European Petrochemical Industry," Aromatics Report and Tables, Mailand 1993, p. 19.
47. Parpinelli Technon: "West European Petrochemical Industry," Aromatics Report and Tables, Mailand 1993, pp. 229 – 311.
48. *Chemical Week,* Oct. 11, 1995, p. 58.
49. H. G. Franck, J. W. Stadelhofer in: *Industrielle Aromatenchemie,* Springer Verlag, Berlin 1987, p. 133.
50. VbF/TRbF, vol. 5 (Regulations), 16. Lfg., June 1989.
51. Kühn, Birett – Merkblätter Gefährliche Arbeitsstoffe, 49th suppl., 6/90 – X03, X06, X07, X08, X19, 1990.
52. BArbBl, no. 9/1990, pp. 65 ff.
53. Unfallmerkblatt für den Straßentransport, MED-Verlagsgesellschaft, Landsberg, edition 5/87.
54. MERCK Sicherheitsdatenbank MS-Safe, 1 Jul. 1991.
55. J. Fabri, A. Reglitzky, M. Voisey, *Erdöl, Ergas* **107** (1991) no. 1, 24 – 29.
56. K. Savolainen, V. Riihimäki, E. Vaheri, M. Linnoila, *Scand. J. Work Environ. Health* **6** (1980) 94.
57. M. Wolf, *Arch. Ind. Health* **14** (1956) 387.
58. R. Morley, D. W. Eccleston, C. P. Douglas, W. E. J. Greville, D. J. Scott, J. Anderson, *Brit. Med. J.* **3** (1970) 442.
59. E. Browning in: *Toxicity and Metabolism of Industrial Solvents,* Elsevier, Amsterdam 1956.
60. H. W. Gerarde: *Toxicology and Biochemistry of Aromatic Hydrocarbons,* Elsevier, Amsterdam 1965.
61. E. Schmidt, *Arch. Gewerbepath. Gewerbehyg.* **15** (1956) 37.
62. W. Matthaus, *Klin. Monatsbl. Augenheilk.* **144** (1964) 713.
63. K. Savolainen, V. Riihimäki, A. M. Seppäläinen, M. Linnoila, *Int. Arch. Occup. Environ. Health* **45** (1980) 105.
64. F. Gamberale, G. Annwall, M. Hultengren, *Scand. J. Work Environ. Health* **4** (1978) 204.
65. K. Savolainen, M. Linnavuo, *Acta Pharmacol. Toxicol.* **44** (1979) 315.
66. V. I. Boiko, *Gig. Tr. Prof. Zabol.* 1970, no. 6, 23.
67. A. Pound, *N. Engl. J. Med.* **67** (1968) 88.
68. A. Pound, *Pathology* **2** (1970) 269.
69. I. Berenblum, *Cancer Res.* **1** (1941) 44.
70. R. P. Bos, R. M. E. Brouns, R. van Doorn, J. L. G. Theuws, P. T. Henderson, *Mutat. Res.* **88** (1981) 273.
71. M. Donner, *Mutat. Res.* **74** (1980) 171.
72. I. Krotov, N. Chebotar, *Gig. Tr. Prof. Zabol.* 1972, no. 16, 40.
73. V. Sedivec, J. Flek, *Int. Arch. Occup. Environ. Health* **37** (1976) 205.
74. I. Astrand, I. Engström, P. Övrum, *Scand. J. Work Environ. Health* **4** (1978) 185.
75. V. Riihimäki, P. Pfäffli, K. Savolainen, K. Pekari, *Scand. J. Work Environ. Health* **5** (1979) 217.
76. W. Senczuk, J. Orlowski, *Br. J. Ind. Med.* **35** (1978) 50.
77. I. Engström, R. Bjurström, *Scand. J. Work Environ. Health* **4** (1978) 195.
78. K. Engström, K. Husman, V. Riihimäki, *Int. Arch. Occup. Environ. Health* **39** (1977) 181.
79. H. Bray, B. Humphris, W. Thorpe, *Biochem. J.* **85** (1949) 241.
80. H. Bray, B. Humphris, W. Thorpe, *Biochem. J.* **87** (1950) 395.
81. G. Bienik, T. Wilczok, *Br. J. Ind. Med.* **38** (1981) 304.

Xylenols → Cresols and Xylenols
Xylidenesulfonic Acids → Benzenesulfonic Acids and Their Derivatives

Xylidines

MICHAEL MEYER, EMS-Dottikon, Dottikon, Switzerland

1. **Introduction** 1
2. **Physical Properties** 1
3. **Chemical Properties** 2
4. **Industrial Production** 2
4.1. **Reduction of Nitroxylenes** 2
4.2. **Amination of Xylenols** 3
4.3. **Other Processes** 3
5. **Quality Specifications, Storage, and Transport** 3
6. **Use** 4
6.1. **Dye Production** 4
6.2. **Plant Protection** 4
6.3. **Production of Pharmaceutical Products, Pharmaceutical Products for Veterinary Medicine, and Vitamins** 5
6.4. **Other Uses** 5
7. **Toxicology** 5
8. **References** 5

1. Introduction

Xylidines are homologues of aniline and toluidine and are therefore very similar in chemical behavior.

Xylidines are used mainly in dye production. Their use in the production of vitamins, pharmaceuticals, and plant-protection agents is also important. Xylidine mixtures (crude xylidines), which were previously mixed with aircraft fuels in large quantities as antiknock agents, are still processed as antioxidants in lubricating oils and rubber products, e.g., car tires.

The xylidines **1**–**6**, $C_8H_{11}N$, molecular mass 121.18, are derived from the three isomeric xylenes:

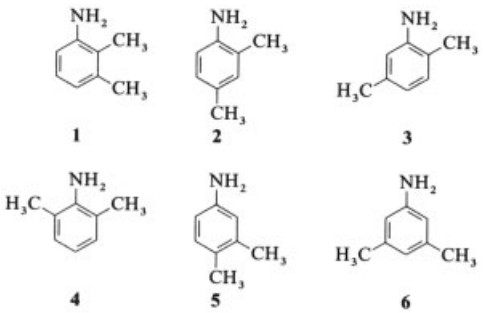

2. Physical Properties

Xylidines are readily soluble in most organic solvents, but are not very soluble in water. At room temperature, xylidines are colorless liquids or congealed melts, each with its characteristic odor. Xylidines become colored within hours in light and air. Physical properties of xylidines are listed in Table 1.

3. Chemical Properties

Like aniline, xylidines are weak bases. The positions of the methyl groups have little influence on the basicity, which increases in the following order in aqueous media: **4** < **3** < **1** < **2** < **6** < **5** [1]. Xylidines form stable salts with strong mineral or organic acids. The differing solubilities of these salts can be used to separate isomers in some cases. However, this method is unimportant in industry [2].

Electrophilic substitution reactions on the aromatic nucleus, e.g., chlorination, sulfonation, nitration, nitrosation, or coupling with diazonium salts, take place very readily because of the high activation by the amino group and the two methyl groups. The positions *ortho* and *para* to the amino group are preferentially attacked. The amino group often needs to be protected. In strongly acidic media, the amino group is protonated and the ring is deactivated. In this case electrophilic substitution is directed to positions *meta* to the amino group [3].

The most important reactions of the amino group are acylation, alkylation, arylation, and diazotization. Xylidines can also be converted into the corresponding isocyanates [4].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_455

Table 1. Physical properties of xylidines

	Other names	CAS registry no.	*mp*, °C	*bp*, °C	d^{20}	n_D^{20}	Flash point, °C
2,3-Xylidine (**1**)	2,3-dimethylaniline, 3-amino-*o*-xylene	[*87-59-2*]	3.5	222.7 – 222.9	0.9931	1.5652	97
2,4-Xylidine (**2**)	2,4-dimethylaniline, 4-amino-*m*-xylene	[*95-68-1*]	−12.5 to −14.3	214.3	0.9763	1.5576	100
2,5-Xylidine (**3**)	2,5-dimethylaniline, 5-amino-*p*-xylene	[*95-78-3*]	14.2 – 15.0	217	0.9755	1.5596	97
2,6-Xylidine (**4**)	2,6-dimethylaniline, 2-amino-*m*-xylene	[*87-62-7*]	11.2	214.8 – 215.1	0.9796	1.5612	98
3,4-Xylidine (**5**)	3,4-dimethylaniline, 4-amino-*o*-xylene	[*95-64-7*]	50 – 51	225	1.0755		97
3,5-Xylidine (**6**)	3,5-dimethylaniline, 5-amino-*m*-xylene	[*108-69-0*]	9.8 – 10.0	218	0.9704	1.5581	103

4. Industrial Production

4.1. Reduction of Nitroxylenes

For all xylidines, apart from 3,5-xylidine and, to some extent, 2,6-xylidine, reduction of nitroxylenes is the most important industrial production route. The corresponding nitroxylenes can be obtained by nitration of the pure xylenes. If mixtures of isomers are obtained, as with *o*- and *m*-xylenes, the nitro compounds can be easily separated by rectification.

Crude xylidines, previously obtained in large quantities by nitration of xylene mixtures (crude xylene) and subsequent reduction, can be further processed to give the pure xylidines, but at great expense [5].

For the reduction of nitroxylenes, processes used in the production of aniline and toluidines can, in principle, be considered. The classical Béchamp reduction with iron and hydrochloric acid [6] has now been widely replaced by catalytic hydrogenation with molecular hydrogen. Usually continuous or batch processes in the liquid phase, generally without solvents, are used [7, 8]. If solvents are used, these are aromatic hydrocarbons, lower aliphatic alcohols, and acetic acid [9, 10]. Palladium on activated charcoal is the most important catalyst. Other possible catalyst systems are platinum or ruthenium on activated charcoal, aluminum oxide, or kieselguhr and Raney nickel [11].

In 1943 and 1944, xylidines were produced in the United States in large-scale industrial plants. Hydrogenation in the gas phase was mostly used [12, 13]. Nickel, nickel – copper, or sulfidic catalysts based on copper, nickel, chromium, molybdenum, or tungsten were used [14, 15].

4.2. Amination of Xylenols

3,5-Xylidine cannot be prepared at all by the nitration of *m*-xylene, as the corresponding 5-nitro-*m*-xylene is formed in $< 1\,\%$ yield . The production of 2-nitro-*m*-xylene, formed in 16 % yield in the nitration of *m*-xylene, depends on the market for 4-nitro-*m*-xylene. 2,6-Xylidine can therefore be produced by the nitration of *m*-xylene only to a limited extent. To produce both these xylidines, processes starting from the generally readily accessible xylenols are more suitable.

In the Halcon process, (also used for the production of aniline) xylenols can be converted to xylidines with excess ammonia at high pressure and temperature on aluminum-oxide-containing

catalysts [16, 17]; a disadvantage is that extensive dealkylations and isomerizations can occur, in particular when xylenols with *ortho* methyl groups or spent catalysts are used [18]. In the Ethyl Corporation process these disadvantages are avoided, if sterically hindered xylenols in the presence of catalytic quantities of a cyclohexanone are converted to xylidines with ammonia using a hydrogen transfer catalyst [19]. In another process developed by the same company, 2,6-xylenol can be aminated under anhydrous conditions via the aluminum salt [20].

4.3. Other Processes

3,4-Xylidine can be obtained by catalytic hydrogenation of 2-chloromethyl-4-nitrotoluene [21]. 3,4- and 3,5-Xylidines can be produced by amination of the corresponding substituted bromoxylenes under high pressure with a copper catalyst [22, 23].

Another process, which can be used to obtain 3,5-xylidine, is the reaction of α,β-unsaturated cyclohexenone oximes with dehydrating agents [24]. Ciba-Geigy describes the condensation of acrolein with the enamine from diethyl ketone and morpholine to give 2,6-dimethylcyclohexenone. The latter can be converted to 2,6-xylidine in a second step with ammonia in the presence of hydrogen-transfer catalysts using a method similar to that of Ethyl Corporation [25].

5. Quality Specifications, Storage, and Transport

Technically pure xylidines are rectified products, generally of purity $> 99\,\%$. The content can be determined simply by titration, GLC [26], or HPLC [27]. Other isomers and side products with hydrogenated aromatic nuclei can be present as impurities. To identify other impurities, e.g., hydroxylamines, azo and azoxy compounds, or phenols, TLC methods are suitable, particularly when these substances are present in trace amounts. The residual moisture is $< 0.1\,\%$ and is determined titrimetrically by the Karl Fischer method. In the case of crude xylidines, the isomer distribution should usually be specified. This can vary within wide limits, depending on the supplier.

The freshly distilled product is pale yellow. On storage over longer periods, particularly in the presence of light and air, and depending on the other storage conditions, the product increasingly becomes red with no detectable change in the xylidine content. The Hazen color index provides a measure of the coloration. The coloration can be suppressed for a certain period by addition of various stabilizers in quantities of $0.05 - 0.5\,\%$. (For the mode of action of structurally different stabilizers see [28].) The following are effective stabilizers: hydrazine hydrate [29], aliphatic polyamines [30], aromatic biguanides [31], alkylphenols [32], and sulfur-containing compounds [33].

Iron containers are used for storage and transport. Table 2 gives the hazard classes and registrations of the individual xylidines [34].

6. Use

6.1. Dye Production

A large number of azo dyes are derived from 2,4- and 2,5-xylidines, whereby the xylidines can function both as diazo and coupling components. Instead of pure 2,4-xylidines, crude xylidines are frequently used to produce red and orange acid dyes.

Yellow and orange solvent dyes and some red acid and direct dyes are produced from 2,4-xylidine. An example of an azo dye based on 2,5-xylidine is Solvent Red 26 (C.I. 26 120).

Of the three theoretically possible 2,4-xylidinesulfonic acids, two are important as diazo components in the production of azo dyes. 4-Amino-1,3-dimethylbenzene-6-sulfonic acid is formed by sulfonation of 2,4-xylidine with oleum [35], and is preferentially used for the production of acid azo dyes. Baking sulfonation of 2,4-xylidine gives 4-amino-1,3-dimethyl-5-sulfonic acid [35], which is further processed to give Acid Orange 9 (C.I. 17 925) and Direct Violet 7 (C.I. 27 855). The reaction of xylidines with diketene [36, 37] is an important first step in the production of pigment dyes. An interesting example of this class of compound is the 2,4-xylidide of acetoacetic acid, used principally as the coupling component for Pigment Yellow 13 (C.I. 21100).

Table 2. Hazard classes and registrations of xylidenes

	EINECS ∗ number	TSCA status	RID/ADR number	UN number	RCRA ∗∗
2,3-Xylidine	201-755-0	listed	6.1/12 b	UN 1711	
2,4-Xylidine	202-437-4	listed	6.1/12 b	UN 1711	
2,5-Xylidine	202-451-0	listed	6.1/12 b	UN 1711	8 EHQ-0990-1080 S
2,6-Xylidine	201-758-7	listed	6.1/12 b	UN 1711	
3,4-Xylidine	202-437-4	listed	6.1/12 b	UN 1711	
3,5-Xylidine	203-607-0	listed	6.1/12 b	UN 1711	

∗ European Inventory of Existing Chemical Substances.
∗∗ Resource Conservation and Recovery Act.

Table 3. Toxicological data for xylidines

	Microsomal mutagenicity on *Salmonella typhimurium*, µmol/plate	LD_{50}, rat oral, mg/kg	References
2,3-Xylidine	5	836	[66]
2,4-Xylidine	1	467	[67]
2,5-Xylidine	10	1297	[66]
2,6-Xylidine	50	707 [b]	[67]
3,4-Xylidine	5	812	[66,69]
3,5-Xylidine	1 [a]	707	[68,69]

[a] mg/plate.
[b] Mouse.

2,5-Xylidine is also used to produce triphenylmethane dyes, such as Solvent Yellow 30 (C.I. 21240), Solvent Red 22 (C.I. 21250), and Acid Red 65 (C.I. 24830).

3,5-Xylidines are currently almost exclusively used to produce the perylene pigment dye Pigment Red 149 (C.I. 71137). Only a few areas of use are known for 2,3-xylidine; e.g., in the form of the amide of 2-hydroxy-3-naphthoic acid, it is a coupling component for azo dyes used in printing inks.

The condensation product of bromamine acid with 2,6-xylidine is an intermediate in the production of acidic blue dyes for wool [38] and reactive dyes.

6.2. Plant Protection

2,3-Xylidine is a component of the herbicide xylachlor (Combat) [39]. The insecticide and acaricide amitraz is obtained from 2,4-xylidine.

5-Acetamido-2,4-dimethylaniline, which can be obtained by nitration and subsequent catalytic hydrogenation of *N*-acetyl-2,4-xylidine, is an intermediate in the production of the herbicide and growth regulator methafluoridamid, mefluidide (Embark, Vistar) [40]. Prowl (Pendimethalin, Penoxalin) is also used as a herbicide and growth regulator [41,42]; it is obtained from 3,4-xylidine by alkylation and dinitration.

2,6-Xylidine is an interesting precursor for a large number of plant-protection agents. Derivatives of *N*-chloroacetyl-2,6-xylidine are of interest as herbicides, growth regulators, and fungicides. Commercial products based on 2,6-xylidine are the fungicides metalaxyl (Ridomil) [43] and furalaxyl (Fongarid) [44].

6.3. Production of Pharmaceutical Products, Pharmaceutical Products for Veterinary Medicine, and Vitamins

The antiphlogistic, antirheumatic, and analgesic mefenamic acid is produced from 2,3-xylidines [45].

2,4-Xylidine is the starting material for the active substance cymiazole (Tifatol, Besuntol), used in veterinary medicine as an ectoparasiticide for combatting ticks [46,47]. 2,6-Xylidine is a component of a range of similarly structured local anesthetics, such as lidocaine, bupivacaine, mepivacaine, and etidocaine [48,49]. 2,6-Xylidine is also used in a range of other medicinal products, e.g., the diuretic xipamide [50], the amidinourea derivative lidamidine hydrochloride [51], which combats diarrhoea, the arylurea derivative xilobam, which has a tranquil-

izing effect on the central nervous system [52], and the coronary vasodilator lidoflazine [53]. The active substance xylazine hydrochloride (Rompun) is also produced from 2,6-xylidine [54], and is used in veterinary medicine as a narcotic and sedative.

3,4-Xylidine is used in the synthesis of riboflavin (vitamin B_2) [55]. It is also a precursor in the biosynthesis of vitamin B_{12} [56].

6.4. Other Uses

The use of xylidines as antiknock agents for gasoline [57] and as antioxidants for lubricating oils and rubber products [58, 59] has already been mentioned. Xylidines are still added to gasoline as coloring agents; 2,5-xylidine is used as an additive for liquid crystal mixtures [60]. The use of xylidines as corrosion inhibitors in metallic materials has also been described [61]. Xylidines can be used as absorption agents for sulfur dioxide in desulfurizing exhaust gas [62].

The reaction of formaldehyde with xylidines is analogous to that with aniline, and gives diaminotetramethyldiphenylmethanes [63] and, under hydrogenation conditions, diaminotetramethyldicyclohexylmethanes [64]. These products are used as cross-linking agents and hardeners for epoxy resins, or can be further processed to polyurethanes.

7. Toxicology

Xylidines are readily resorbed by the skin and mucous membranes. Their ability to bind hemoglobin is known, in good agreement with the established maximum methemoglobin level.

Detailed test have been carried out with 3,4-xylidine and 2,4-xylidine [65]. Very high xylidine concentrations attack the central nervous system directly. Xylidines also lead to induction of microsomal enzyme systems in the liver. In animal experiments, toxic effects on the lungs and kidneys have also been established.

All the isomers of xylidine are identified as mutagenic, 2,4-, 2,5-, and 2,6-xylidines have also been found to cause tumors. The LD_{50} (oral) for rats is 0.1 – 1 g. The MAK value has been set at 5 mg/kg for any single xylidine. Table 3 gives an overview of the toxicological data described.

There are reports of tests for carcinogenesis on 2,4-, 2,5-, and 2,6-xylidines, which have been found to cause tumors. High doses of 2,6-xylidine, which has been identified as a component of tobacco smoke, leads to adenomas and carcinomas in the nasal cavities in rats. Mesenchymal tumor and rhabdomysarcoma have been observed as malignant carcinomas. Subcutaneous tumors and neoplastic lumps in the liver have also been found. In rats 2,6-xylidine is classed as definitely carcinogenic [70].

8. References

1. R. N. Beale, *J. Chem. Soc.* 1954, 4494.
2. G. T. Morgan, W. J. Hickinbottom, *J. Chem. Soc. Ind. London Trans. Commun.* **45** (1926) 221.
3. J. Bergman, P. Sand, *Tetrahedron* **46** (1990) 6085.
4. Sumitomo Chemical, JP 58 08, 052 [83 08, 052], 1983.
5. Anglo-Iranian Oil, GB 599 252, 1944.
6. BIOS Final Report, **1146** (1947) 65.
7. Clayton Aniline, EP 2308, 1978.
8. ICI, GB 982 902, 1962.
9. Teijin, JP 70 490, 1966.
10. Mitsui Toatsu Chem., JP 7 0 31 652, 1967.
11. Texaco, US 3 832 401, 1971.
12. Shell, US 2 458 214, 1944.
13. Standard Oil, GB 599 252, 1944.
14. J. Werner, *Ind. Eng. Chem.* **41** (1949) 1841.
15. Standard Oil, US 2 455 713, 1945.
16. General Aniline Works, US 2 013 873, 1934.
17. Bayer, DE-OS 3 023 487, 1980.
18. Rütgerswerke, DE-OS 2 516 316, 1975.
19. Ethyl Corp., DE-OS 2 208 827, 1972.
20. Ethyl Corp., DE-AS 1 933 636, 1969.
21. Y. P. Sobolev, V. M. Berezivskii, SU 118 129, 1959.
22. American Home Products, US 2 347 652, 1941.
23. V. I. Rogovik, I. I. Gulianyuk, SU 287 958, 1969.
24. Hoechst, DE-OS 2 654 852, 1978.
25. Ciba-Geigy, US 4 188 341, 1975.
26. A. Ono, *Chromatographia* **14** (1981) 39, 692.
27. F. K. Chow, E. Grushka, *Anal. Chem.* **49** (1977) 1756.
28. J. F. Kung, Jr., W. C. Howell, Jr., C. E. Starr, Jr., *Ind. Eng. Chem.* **40** (1948) 1534.
29. Du Pont, DE-AS 1 518 107, 1965; GB 1 047 607, 1965.

30. Standard Oil, US 2 447 615, 1944.
31. Standard Oil, US 2 469 745, 1946.
32. Standard Oil, US 2 510 849, 1946.
33. Du Pont, US 2 686 809, 1952.
34. Sigma-Aldrich Library of Regulation and Safety Data, vol. I, Sigma-Aldrich Corp., Milwaukee, Wis., 1993.
35. A. Courtin, H. R. v. Tobel, P. Doswald, *Helv. Chim. Acta* **61** (1978) 3079.
36. Hoechst, DE-OS 2 749 327, 1977; DE-OS 2 519 036, 1975.
37. Ciba-Geigy, EP 4611, 1979.
38. Geigy, DE-AS 1 148 341, 1959.
39. Amer. Cyanamid, DE-OS 2 632 437, 1976.
40. Minnesota Mining, DE-OS 2 406 475, 1974.
41. Amer. Cyanamid, DE-OS 2 429 958, 1973.
42. Amer. Cyanamid, DE-OS 2 232 263, 1971.
43. Ciba-Geigy, DE-OS 2 515 091, 1975.
44. Ciba-Geigy, DE-OS 2 513 732, 1975.
45. Parke Davis, DE-AS 1 186 073, 1963.
46. Ciba-Geigy, DE-OS 2 619 724, 1976.
47. Bayer, DE-OS 2 531 606, 1975.
48. S. M. Waraszkiewicz, W. O. Foye, H. J. Adams, B. H. Takman, *J. Med. Chem.* **19** (1976) 541.
49. J. Marcus, A. Bardeanu, Z. Theodorescu, *Pharmazie* **14** (1959) 322.
50. Beiersdorf, DE-OS 1 270 544, 1965.
51. William H. Rorer Inc., DE-OS 2 433 837, 1973.
52. Mc Neil Laboratories, DE-OS 2 313 257, 1971.
53. Janssen, NL 6 507 312, 1965.
54. Bayer, DE-AS 1 173 475, 1962.
55. BASF, DE-OS 3 615 834, 1986.
56. M. H. Kay, C. Radford, W. L. Alworth, *J. Chem. Soc. Chem. Commun.* **78** (1978) 22.
57. J. E. Brown, F. X. Markley, H. Shapiro, *Ind. Eng. Chem.* **47** (1955) 2141.
58. Du Pont, FR 1 553 317, 1967.
59. Goodyear, US 4 269 763, 1980.
60. Bayer, DE-OS 3 536 267, 1985.
61. M. L. J. Bernard, DE-AS 1 084 544, 1958.
62. Metallgesellschaft, DE-AS 1 025 400, 1956.
63. ICI, GB 852 651, 1958.
64. Bayer, DE-OS 2 945 614, 1979.
65. G. Birnes, H. G. Neumann, *Arch. Toxicol.* **62** (1988) 110.
66. D. Zimmer, J. Mazurek, G. Petzold, B. K. Bhuyan, *Muta. Res.* **77** (1980) 317.
67. D. Zimmer, J. Mazurek, G. Petzold, B. K. Bhuyan, *J. Agric. Food Chem.* **34** (1986) 157.
68. E. Zeiger et al., *Environ. Mol. Mutagen.* **11** (1988) 1.
69. J. M. Mac Ewen, E. H. Vernot, *Aerosp. Med. Res. Lab. Tech. Rep. AMRL-TRUS* **72** (1972) 72–62.
70. M. Kornreich, C. A. Montgomery, *Nat. Toxicol. Program Tech. Rep. Ser.* **278** (1990).

Xylitol → Sugar Alcohols

Yeasts

Albert K. Athnasios, Fleischmanns Yeast, Division of Burns Philp Food Company, Fenton, Missouri 63026-2837, United States

1. Introduction 1
2. The Yeast Cell 2
3. Food and Feed Yeast 3
3.1. Composition 4
3.2. Use of Yeast as a Major Protein Source 5
3.3. Production of Biomass 6
3.3.1. Baker's Yeast 6
3.3.2. Active Dry Yeast 9
3.3.3. *Candida utilis* (Torula) Yeast 10
3.3.4. Dairy Yeasts 11
3.3.5. Brewer's Yeast 11
3.3.6. Wine Yeasts 11
3.3.7. Distiller's Yeasts 12
4. Yeast-Derived Products 12
4.1. Flavor Products 12
4.2. Nutritional Yeast 12
4.3. Colorants Derived from Yeast 12
4.4. Yeast-Derived Enzymes 13
4.5. Pharmaceutical and Cosmetic Products 14
5. References 15

1. Introduction [1]

The first records on the use of yeast date back to 6000 B.C. Egypt, where a type of acid beer called "boozah" was produced.

The production of wine, beer, and leavened bread probably developed over the next few thousand years. In Roman times, beer was a common drink in Spain and Germany. The British welcomed the production of beer by the Romans and adopted it as their national drink.

The people of South Asia have for thousands of years prepared and stored food by fermentation methods, e.g., rice wine, sugar cane wine, pickles, fish sauces, and various soy foods.

Many other foods, such as cultured dairy products, fermented sausages, etc., also contain microorganisms, including yeast. So, microbial cells have long been consumed by humans. Kaffir beer, for example, which is a traditional beverage in Africa, is still produced and is almost a complete diet due to its balance of yeast and grain. The nature of yeast long remained unknown until Pasteur established beyond doubt the role of yeast in the fermentation of sugar to alcohol, and that yeast is a living organism.

The technology of microbial cell production on a commerical scale for food and feed has developed only within the past 100 years. In ancient times people recovered top fermenting yeasts of the genus Saccharomyces from the production of fermented beverages such as beer, and used them as leavening agents in baked goods. Preparation of compressed brewer's yeast began in the late 1700s in Great Britain, The Netherlands, and Germany.

By 1868 Fleischmann had founded the compressed yeast industry in the United States and concurrent improvements in the biochemistry of fermentation and technical engineering greatly aided the development of the modern yeast factory. Continuous aeration and separation by centrifugation were developed for baker's yeast production.

Modern technology for microbial cell production dates from World War I. Large-scale fermenters with effective aeration were developed and new continuous processes were used. New raw materials (molasses, sulfite liquor, hydrocarbons, whey, starchy wastes from potato processing, date syrup, etc.) were used. The number of yeast strains and species grown on different raw materials also increased.

In the fermented foods and beverages industries, the growth and human consumption of microbes started in Germany in about 1910 with the brewer's yeast which was separated from beer, washed, and dried. During and following World War II *Candida utilis* yeast was grown on wood sugar containing sulfite liquors. Although the main usage of microbial biomass is at the present in feeds, good possibilities exist for the increase of food applications.

Four species of yeasts are presently used in foods: *Saccharomyces cerevisiae*, *Sac-*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_461

charomyces uvarum (formerly *Saccharomyces carlsbergensis*), *Candida utilis* (*Torula utilis*), and *Kluyveromyces fragilis* (formerly *Saccharomyces fragilis*). *Saccharomyces cerevisiae* and *Saccharomyces uvarum*, the yeast of the brewing industry, are byproducts of beer production. Other yeasts are grown especially for food and feed use and are often called primary yeasts. *Torula utilis* is grown in sulfite liquor, and *Kluyveromyces fragilis* on cheese whey. Yeast may also be grown on hydrocarbons; however, this was found to be infeasible due to economic and safety reasons.

Detailed food analysis shows that yeast is a rich source of protein, essential minerals, and one of the richest sources of B group vitamins.

Although fermentation and leavening are the most common use of yeast in food production, the application of yeast and yeast derivatives for nutritional fortification (as protein and vitamin sources) of processed foods and also for refining their flavor profile is growing.

Industrially important yeasts include baker's yeast, brewer's yeast, wine yeasts, and distiller's yeasts. Yeast-derived products are numerous including flavors and flavor enhancers, nutritional yeast, colorants, and a broad spectrum of products of pharmaceutical and cosmetic value. From different yeasts, several enzymes and biologically active compounds are produced on a commercial scale such as alcohol dehydrogenase; hexokinase; L-lactate dehydrogenase; glucose-6-phosphate dehydrogenase; glyceraldehyde-3-phosphate dehydrogenase; inorganic phosphates; coenzyme-A; oxidized and reduced diphosphopyridine nucleotides; and mono-, di-, and triphosphates of adenine, cytidine, guanosine, and uridine.

It is difficult to precisely define yeast on the basis of general morphological and physiological considerations. For example, most yeast cells are colorless and transparent, while some produce carotenoid pigments (e.g., Rhodotorula). Bud formation is widespread, but is not common to all yeasts. Cell shape, in most cases, is round to oval, but several species of some genera may produce a true mycellium.

The taxonomy of yeasts is in a continual state of flux. As new microorganisms are discovered, increasing knowledge about them results in the yeasts being shifted within subgroups. However, such changes are slow to be put into practice, except by taxonomists.

Frequently, species were named on the basis of fermentation they were associated with, e.g., *Saccharomyces vinii* or *Saccharomyces* sake. The morphological and physiological properties are often very similar, and separate names have not been justified taxonomically. For various industrial processes, special characteristics of production strains are of great importance; therefore, strain differentiation may be more valuable than species classification.

2. The Yeast Cell

Electron micrographs of thin sections of yeast cells show the existence of a cell wall, a plasma membrane, nucleus, mitochondria, and vacuoles.

The cell wall makes up 15 % to 20 % of the dry weight of the cell. The proportional weight of the cell wall usually decreases during the growth phase and increases again in the stationary phase. The main cell wall components are the polysaccharides glucan and mannan, chitin, protein, and lipids. Acid hydrolysis of the cell wall results in the formation of glucose, mannose, glucosamine, amino acids, phosphates, and lipids. The chemical composition of the cell wall of *Saccharomyces cerevisiae* shows great variability, with carbohydrates varying from 60 to 91 %, proteins from 6 to 13 %, and lipids from 2 to 5.8 %. The two main polysaccharides are present in about equal amounts. Chitin is a polymer of β-(1 – 4)-linked *N*-acetylglucosamine.

The cell-wall protein contains all the common amino acids, with high proportions of glutamic and aspartic acids. In contrast to the whole-cell protein amino acid composition, cell-wall protein has a high content of sulfur amino acids which, especially in younger cells, contain SH groups. In older cells the protein chains are linked by disulfide bonds. The structural protein of the cell wall is mainly bound to polysaccharides to form a complex structure in which glucosamine acts as the connecting link between the polysaccharides and proteins.

The lipid content of the cell wall varies both with species of yeast and growth conditions. Published data show great variations, probably due to incomplete removal of the lipid-rich

plasma membrane. Some of the lipid content is firmly bound to the cell wall and may play a role in maintaining its orderly structure. $C_{16:1}$, $C_{18:1}$, and $C_{16:0}$ acids are the dominant fatty acids of the cell-wall lipids. The plasma membrane lipids have a higher content of $C_{18:1}$ acids; in addition the phosphatidylinositol and phosphatidylserine components of phospholipids are enriched compared to the cell.

The nuclei of yeast are surrounded by an envelope which is characterized by the presence of many pores.

Mitochondria are abundant in all fungi. Their most morphological characteristic is the presence of inner and outer membranes. The size, shape, number, and composition of mitochondria vary widely under different conditions of growth.

Vacuoles are important subcellular organelles in yeasts. Both vegetative and reproductive cells contain vacuoles that vary significantly in size. The vacuoles contain a variety of hydrolytic enzymes such as proteases, and occasionally they contain degrading intracellular substances.

3. Food and Feed Yeast [2]

Yeast-fermented foods constitute a significant percentage of the human diet. Nutritionally, the most important yeast fermentations are baked goods and beer. Yeasts also have some importance as foods and feeds.

About $(70-80)\times10^6$ kg of baker's yeast (dry basis) are produced annually in the United States. Most of this is used in the production of baked goods as compressed or active dry yeast. About $(2.3-4.6)\times10^6$ kg is used for direct consumption as dried nutritional yeast. In the brewing industry 2.3 – 3.6 kg of yeast (dry basis) become available as excess yeast per barrel of beer. Thus, the total supply is ca. 55×10^6 kg/a. Most of this yeast is used in feed, usually by co-drying it with spent grains, another byproduct of the brewing industry. But a considerable percentage, as much as 14×10^6 kg, is used in the production of yeast extracts or sold as debittered brewer's yeast for nutritional purposes.

In some fermentation industries the yeast cannot be recovered as such because the fermented grape must or the fermented grain mash contains insoluble particles that make separation of the yeast impractical. In the wine industry, the yeast is discarded with the lees. In the production of distilled beverages the yeast is co-dried with the distiller's solubles or the spent grain and used as feed. Active dry yeast is also used as a supplement to ruminant feeds and has been reported to increase feed efficiency in cattle.

Yeast flavor may be pleasant for some people or indifferent or less than desired to others. It is not used because of its flavor but because of its nutritional contribution of protein, vitamin B_1, and minerals. Yeasts are sold for human consumption as inactive dried yeasts. They are killed by pasteurization or during drum or spray drying. The dried powders or flakes are usually consumed by blending them with fruit juices or by sprinkling over cereals. In the health food industry yeasts are promoted on the basis of unidentified nutritional factors. Such claims have to be treated with considerable caution because of their promotional aspects. However, two such previously undefined nutritional factors that were discovered in yeasts in the 1950s are selenium and chromium, which are now recognized as essential trace elements.

3.1. Composition

The gross composition of yeasts is quite variable. The percentage of proteins and nucleic acids depend on the substrate and growth rate. Yeast composition is outlined here on the basis of *Saccharomyces cerevisiae*. The other two nutritional yeasts, namely *Candida utilis* and *Kluyveromyces marxianus*, have somewhat similar composition.

Protein. The protein of yeasts is usually determined by Kjeldahl nitrogen determination and multiplication of the nitrogen content by 6.25 ($N\times6.25$). The total nitrogen includes the nitrogen of the nucleic acids and, therefore, this procedure overestimates the protein content. Protein calculated in this manner should be designated crude protein. A factor of $N\times5.5$ is more appropriate to estimate pure protein.

General use of the 6.25 factor has led to underestimation of the nutritional value of yeast proteins expressed by the protein efficiency ratio (PER) because animal feeding tests were carried out at an assumed protein value higher than

the true value. Much of the literature uses a PER value of 1.8 (cf. casein PER of 2.5).

Total yeast proteins are generally higher in lysine and deficient in sulfur amino acids, methionine, cysteine, and cystine as compared to cell-wall protein amino acid composition (see Chap. 2). The amino acid composition of yeasts resembles that of oil-seed proteins, particularly soy protein.

Vitamins. Most yeast species are unable to synthesize one or more vitamins and are dependent on external sources.

Yeast contains predominantly the vitamins of the B complex, and is thus an excellent source of B vitamins for human and animal nutrition.

Inactive dry yeast is frequently used as a vitamin supplement rather than as a source of protein in food formulations. Some dry yeast products are even fortified with vitamins such as B_1, B_2, and niacin to meet certain special requirements of vitamin tablet manufacturers. Although the daily requirements for most of the B vitamins are not fully met by the recommended yeast intake, their contribution to meeting the B complex vitamin requirements in the human diet is substantial. However, yeast does not provide vitamin C and fat-soluble vitamins such as A, E, K, and D. Nevertheless, dry yeast can function as a valuable vitamin source when used in combination with other food ingredients, as is generally the case with human diets.

Baker's yeast contains 7 – 10 % ergosterol, mainly in the membranes. It serves as the precursor for vitamin D_2. This conversion occurs when ergosterol is irradiated with ultraviolet light. However, this process is not economical since synthetic vitamin D_2 is less expensive. Baker's yeast does not contain 7-dehydrocholesterol, the precursor of vitamin D_3.

Minerals. Yeasts take up substantial quantities of macro- and micronutrients from the surrounding growth medium. The mineral-ash content amounts to ca. 8 % on a dry solid basis. Although the concentrations of potassium and phosphorus in dry yeast are higher than those of other elements such as Ca, Mg, and S, the mineral contributions made by an allowable daily serving of 20 g of dry yeast is minor considering the high mineral content in the bulk of an average human diet.

However, nutritional studies have shown that certain diets are deficient in some important trace elements. Supplementation of these diets with dry yeast has partially or wholly alleviated such dietary problems. Such inadequacies in the diet were later determined to be due to deficiencies of trace elements such as chromium, selenium, and molybdenum. Brewer's and baker's yeasts contain these elements in trace levels. Specially produced yeast products are now commercially available with higher levels of these micronutrients.

Chromium. In 1955 researchers reported that rats fed with certain nutritionally deficient diets showed impaired tolerance to blood glucose. The conditions were then reversed by supplementing the feed with brewer's yeast. It was later suggested that the active component responsible for reversing the glucose intolerance was an organic complex rich in chromium.

Diabetes, which is caused by glucose intolerance, can manifest itself in two ways. In juvenile diabetics, the pancreas fails to secrete insulin into the bloodstream. Those who are affected by another type of glucose intolerance begin to show symptoms in midlife. These patients secrete insulin into the bloodstream but cannot control their blood sugar level. It is possible to correct this problem by supplementing the diet with brewer's yeast. Findings have indicated the importance of another component besides insulin for the proper control of the blood sugar level. This active component, which is a trivalent chromium complex, is now referred to as glucose tolerance factor (GTF). Patients in the second category have GTF early in life, but tend to lose it with age, causing increased vulnerability to diabetes. It is suggested that GTF functions as a cofactor for insulin, thereby enhancing the binding of insulin to receptive sites on the membranes of insulin-sensitive tissues.

Other investigations showed a reduction in the level of cholesterol and triglycerides in the blood of humans supplementing their diets with brewer's yeast rich in GTF. However, further feeding studies are necessary to confirm the results before yeast rich in GTF could be recommended as a possible treatment for lowering blood sugar or curing lipid disorders.

Selenium has been recognized as an essential trace element for both human and animal nutrition. Recent surveys in some parts of Fin-

land show that people with blood selenium less than 0.04 µg/L are three times more prone to heart attacks than their counterparts with normal selenium levels. Also, in China a fatal heart disease in children known as Kaishan's disease is presently treated by fortifying the diet with sodium selenite.

The importance of selenium in animal nutrition is also well documented. A common disease in lambs, white muscle disease, is due to selenium deficiency. It can be cured by adding 0.05 ppm Se to the diet.

A diet containing Se at the levels approaching toxicity is effective in lowering the incidence of some types of cancer in animals. Research findings strongly suggest that certain types of cancer can be prevented by maintaining a proper level of Se in the diet.

Early feeding studies conducted with selenite or selenate salts as sources of Se showed that the uptake of Se was quite poor. In contrast, organically bound Se, such as selenomethionine found in some natural foods, is incorporated several times faster into the body tissue than inorganic Se.

Organically bound Se is now recognized as nutritionally important because of its ability to prevent some vitamin E deficiency disorders, at least in laboratory animals fed diets partially reduced in nucleic acids.

Reports, particularly from Finland, indicate that Se deficiencies can be corrected by supplementation with Se-rich yeasts.

There are several factors that restrict Se uptake by yeast under conventional batch propagation conditions primarily because of toxic effects of Se salts on yeast growth. However, a new procedure for the propagation of food-grade Se-rich yeast has been developed, based on the concept that, under conditions of sulfur deficiency, sulfur could be replaced by selenium in yeast. The growth medium is fed incrementally so that the Se concentration never reaches toxic levels. Under these conditions the yeast assimilates Se as a reaction to sulfur deficiency. Nutritional yeasts with an intracellular Se concentration of 1000 ppm are commercially available in the United States as dietary supplements.

Lipids. The lipid content of yeasts can vary between 4 and 7 % (dry basis). About 1 % can be extracted directly by solvents. The determination of total lipids requires acid hydrolysis prior to solvent extraction. Major constituents are fatty acid glycerides with a predominance of palmitic and oleic acids, sterols, and lipolipids.

Carbohydrates. Total carbohydrates account for about 30 – 35 % of the yeast cell (dry basis). They consist mainly of carbohydrate storage compounds such as glycogen, the disaccharide trehalose, and the structural materials of the cell wall: the glucans and mannans. Level of fiber in whole baker's yeast cells is about 18 % (dry weight basis).

3.2. Use of Yeast as a Major Protein Source

Candida utilis yeast was used as an important human diet supplement in Germany during World War II. Sulfite waste liquor, a byproduct of the paper pulp industry, and wood hydrolysates were used as raw materials. The large-scale production of yeasts on hydrocarbon substrates was found to be infeasible for economic and safety reasons.

For use as a major source of protein by humans the presence of nucleic acid is a serious obstacle; nucleic acid nitrogen accounts for about 10 – 15 % of the total nitrogen of yeast cells. The intake of nucleic acid leads to elevated blood plasma levels of uric acid and may cause gout. Sources of nucleic acids in the diet are meat, particularly organ meat such as liver. Beer and possibly other fermented beverages also contain nucleic acids.

Search for methods is underway to reduce the level of nucleic acids or to remove them completely, since present methods are not economical. There is no need, however, for the removal of nucleic acids from yeast biomass for use in feed.

Yeast biomass in its inactive dried form is used widely as feed supplement. It is used in poultry rations and pig starter feeds. The use of brewer's and distiller's byproduct yeast in feed has been demonstrated. In countries that lack cheap sources of oil-seed meals (mainly soybean meal), *Candida utilis* biomass is used extensively as a protein supplement. Live yeast cells in the form of active dried yeast or yeast culture are also used in the feed industry.

Yeast culture is produced by combining slurries of baker's yeast with cereal feed grains. The mash is incubated to permit yeast multiplication and fermentation. It is then dried at temperatures that preserve yeast viability.

3.3. Production of Biomass [2]

3.3.1. Baker's Yeast

The production of baker's yeast exclusively for the production of bread doughs dates from the late 1800s. It started with the production on grain mashes, which were later replaced with the least expensive source of fermentable sugar, molasses.

During the last century the major advances in the production of baker's yeast have been the following:

1) Aeration: The stimulating effect of aeration on yeast growth was well known in the late 1800s, and continuous aeration of mashes was used in Britain in 1886.
2) Fed-batch process: The use of incremental feeding was introduced by Danish and German scientists between 1910 and 1920. This process is used today because it is the only practical method that permits production of yeast biomass without simultaneous production of large quantities of ethanol.
3) Molasses: Before 1920 the mash consisted of corn, malt, and malt sprouts. During the 1920s and 1930s, the grains were slowly replaced with molasses as carbon and energy source for yeast growth.
4) Active dry yeast: Progress was made in the late 1930s in the drying of compressed yeast (CY) to the more stable active dry yeast (ADY). Continued progress in the drying methods has led to ADY replacing CY in several bakery applications, particularly in institutional baking and in pizzerias.
5) Automation: During the past decades there has been a rapid shift to automatic control of the fermentation process.

Cane or beet molasses supply sugar as a carbon and energy source as well as organic nitrogen, minerals, sulfur, vitamins, and trace elements. The liquid growth medium must be supplemented with additional nitrogen (ammonia or ammonium salts) and phosphate, minerals (calcium and magnesium), vitamins, and trace elements.

Yeast is grown in large fermenters by the fed-batch process. The fermenters are equipped with cooling coils, and with means of vigorous aeration to maintain highly aerobic growth. In the fermenter liquid, 4 – 6 % of yeast solids can be produced.

Postfermentation processing begins with centrifuging to produce a concentrate (yeast cream) with 18 – 20 % solids. The yeast cream is washed and pressed or filtered to a semi-solid yeast mass of 30 % solids. The filter cake is packaged as a crumbly mass in bags or extruded in blocks that are wax wrapped. It is cooled and shipped refrigerated to bakeries. Some bakeries accept shipments of refrigerated yeast cream.

ADY is produced from the yeast cake by extrusion in the form of thin strands. These are dried on a continuous belt or in air-lift dryers. ADY with a solid content of 92 – 96 % is shipped in vacuum-packed or nitrogen-flushed pouches. It does not require refrigerated shipment or storage.

Until the early 1970s, two strains of *Saccharomyces cerevisiae* were used in the United States. The strain used for the production of baker's compressed yeast was grown to a nitrogen level of 8.2 – 8.8 % (dry weight basis). The ADY strain No. 7752 in the American Type Culture Collection was grown to a nitrogen level rarely exceeding 7 %. The lower nitrogen level favored the stability of the yeast, but resulted in a somewhat lower fermentation activity in comparison with compressed yeast on an equivalent solid basis.

A need for additional strains has arisen due to the following requirements of the baking industry: a dry yeast strain with improved fermentation activity, a yeast strain with exceptional fermentation activity in high-sugar dough, and yeast with improved performance in yeast-leavened frozen dough. Some new strains have now become available, but acceptance by the baking industry has been slow, partly because bakers do not like to deal with several strains in their bakeries, and partly because the production and distribution of several strains by yeast companies are more costly. Nevertheless, a trend towards more yeast strains for specialty applications can be expected.

Strains of baker's yeast for industrial production are preserved as pure cultures, usually on agar slants in test tubes. They are frequently transferred by well-known pure-culture methods. Commercial fermentations are started from such slants. Industrial strains of baker's yeast are quite stable.

Baker's yeast strains are available from public and private culture collections, and they can also be isolated from commercial baker's yeast since this yeast enters the market as a viable product. There are several patents protecting specific strains. Enforcement of such patent rights depends, of course, on the ability to identify the patented strain with certainty, which is often difficult.

Raw Materials.

Carbon and Energy Sources. Historically the major source of carbon for yeast growth was maltose derived from malt-converted grain mashes. Today, glucose, fructose, sucrose and a portion of raffinose from molasses serve universally as the carbon source. The discussion here is restricted to the growth of *Saccharomyces cerevisiae* on molasses and a brief indication of potential sugar sources from starches and whey.

Cane molasses is generally used in subtropical areas where cane is grown or in places to which it can be readily shipped by sea. In northern climates where sugar beet is extensively grown, and in areas distant from marine transportation, beet molasses is the major constituent of the mash. The price of molasses is largely determined by its value as a feed. In fermentation industry one of its major uses is the production of baker's yeast. At present, molasses is the cheapest source of fermentable sugars.

In the United States, mash bills contain at least 20 % cane molasses to supply enough biotin for yeast growth. Beet molasses is relatively easy to clarify, whereas cane molasses contains colloidal substances that make clarification more difficult. Molasses is deficient in nitrogen, phosphorus, and magnesium. These and some trace minerals and vitamins have to be supplied by enrichment of molasses or by separate addition to the fed-batch fermentation.

Beet and cane molasses are shipped at 80 – 85 Brix. They have a sugar content that varies between 50 and 55 wt %. Molasses can be stored well in large tanks often holding several thousand tonnes. It can be pumped at temperatures above 20 °C. Molasses contains some insoluble materials and some precipitation takes place on heating. Therefore, molasses is clarified. Cane molasses cannot be filtered well. It is clarified by addition of phosphates and sterilization followed by centrifugation. Beet molasses can be clarified in the same manner or by filtration. Prior to clarification, molasses is diluted with water to 30 – 40 Brix, which facilitates the process.

Some molasses is not satisfactory for the production of yeast. This may be due to the lack of or to insufficient concentrations of certain nutrients or to the presence of yeast-growth inhibitors, which may include fatty acids, nitrites, various insecticides or herbicides used in the field, or bactericides used in the sugar factory.

Beet molasses contains 1 – 1.5 wt % raffinose on dry solid basis. Baker's yeast strains assimilate only the fructose moiety of the raffinose molecule, which accounts for one-third of the trisaccharide content.

Nitrogen. Beet or cane molasses contains low concentrations of nitrogenous materials. Betaine nitrogen in beet molasses is not assimilated. A minor fraction of the low amino acid content of molasses is assimilated. In the production of baker's yeast, nitrogen is supplied in a readily assimilable form as ammonia, ammonium phosphate, ammonium sulfate, or urea.

Feeding of nitrogen in the form of ammonium sulfate or phosphate leads to liberation of sulfuric acid or phosphoric acid, which must be neutralized to maintain the desired pH range. When urea is used as a nitrogen source, fermentation requires higher levels of biotin.

Baker's yeast contains 6 – 9 % nitrogen. High nitrogen levels in the yeast result in higher activity but lower stability, both for compressed and active dry yeasts.

Minerals. Due to the low level of phosphorus in molasses, additional P should be added to the growth medium. The required concentration of phosphorus expressed as P_2O_5 should be one-third of the nitrogen concentration; thus 2 – 3 % P is supplied in the form of phosphoric acid or its salts such as ammonium phosphate. Molasses contains sufficient potassium, calcium, and sulfur, but is low in magnesium, which must be added to the growth medium.

Vitamins. Baker's yeast strains require biotin for growth. Cane molasses supplies ample

amounts of this vitamin, usually 0.5 – 0.8 ppm. Beet molasses contains only traces of biotin. Therefore, at least 20 % cane molasses is used in beet/cane blends for yeast production. For fermentation of carbohydrate sources that contain little or no biotin (e.g., corn syrup and date concentrate) the synthetic vitamin can be added at a rate of 60 – 100 µg per 100 g of yeast solids grown.

Molasses contains sufficient amounts of other vitamins including pantothenate, thiamin, pyridoxine, and nicotinic acid. Baker's yeast can synthesize thiamin from thiazole and pyrimidine compounds, and there is no absolute need for thiamin. However, concentrations above 100 µg/g of yeast solids in American commercial baker's yeast indicate that thiamin has been added to the growth medium. Such high levels of thiamin do not improve growth rate or yield, but have a small but demonstrable effect on yeast fermentation activity.

Alternative Carbon Substrates. Molasses is the least expensive source of sugar for growing baker's yeast, but its price occasionally fluctuates and may approach that of corn-derived sugars. Therefore, other carbohydrate sources have been investigated, including hydrolyzed lactose from cheese whey, and hydrolyzed starch. For growth of baker's yeast the hydrolysis of starch by commercial amylases would be more cost effective.

Fermentation Activators and Inhibitors. Activators are the nutrients required for the successful completion of the fermentation process. Activators include P, Mg, K, Na, Mn, and S. Only the trace elements (Cu, Fe, and Zn) are essential for yeast metabolism. They become integral part of the proteins as activators of enzymes and function as cofactors in biochemical reactions within the cell.

Fermentation inhibitors include heavy metals at very low concentrations. SO_2 also inhibits yeast growth, but there is good adaptation even at concentrations of 80 – 100 ppm. Molasses contains variable amounts of nitrate that can be reduced to nitrite by bacterial action during growth. Yield losses due to nitrite inhibition have been reported for nitrite concentrations of 0.001 – 0.004 %.

Contamination. Baker's yeast always contains some contaminating organisms since it is grown in fermenters that cannot be kept completely sterile.

Specifications for baker's yeast require absence of Salmonella and other pathogens, and restrict the number of coliform organisms to less than 1000 per gram and the count of *Escherichia coli* to less than 100 per gram. Baker's yeast plants follow Good Manufacturing Practice (GMP) rules of the Food and Drug Administration, and great efforts are made to keep the number of contaminants to a minimum.

Molasses is generally sterilized in bulk or in continuous heat exchangers. The air used to aerate the fermenter is passed through depth filters or membrane filters, or is heated. Fermentation tanks are thoroughly cleaned by hot alkaline solutions or detergents, usually with clean-in-place (CIP) equipment, and if necessary by scrubbing.

All pipes, pumps, valves, and other auxiliary equipment must be thoroughly sanitized. It is also important to consider potential contamination from other accessory equipment such as air pipes, leaking cooling coils, probes for pH meters and other instruments and sampling devices. Particular attention has to be paid to ducts for exit air to prevent condensate from dripping back into the fermenter. Similar attention must be paid to further processing of the grown yeast, which includes the centrifuges, presses, filters, and all packaging operations. For active dry yeast it includes the dryers. Yeast is an excellent growth medium for bacterial contaminants, and deposits of wet or dry yeasts on equipment become a ready source of infection.

The overwhelming proportion of contaminants in CY or ADY are lactic acid-producing organisms which almost always belong to the heterofermentive lactic acid-producing organisms of the genus Leuconostoc or to the homofermentive bacteria of the genus Lactobacillus. Some coliform organisms and occasionally a few *Escherichia coli* can be found. Active dry yeast made by drying compressed yeast contains the same kinds of microorganisms, but the total count generally decreases during drying and upon subsequent storage of the dried yeast. Specifications for rope spores require an upper limit of 200 per gram ADY.

Occasionally contamination with Oidium or other molds has been reported. Various species of contaminating yeast have also been reported,

most of them falling into the genera Candida and Torulopsis.

3.3.2. Active Dry Yeast

Baker's yeast can be dried in its vegetative state to a product having less than 8 % moisture without substantial loss of viability, but success of the drying method requires mastery of the various requirements of the methodology.

In addition to the household market, ADY is now also used extensively by retail bakers, institutional users, pizza chains, and in all areas in which it is impossible or too expensive to distribute refrigerated CY. The latter has not been replaced in major wholesale bakeries.

Until a few years ago almost all the ADY was produced with strain No. 7752 of the American Type Culture Collection, which gives a better yield on a sugar basis. Its cells are larger than those of strains used for the production of compressed yeast. There is some loss of solids during the early stages of drying due to increased respiration. This strain can be dried in continuous tunnel (chamber) dryers with minimal loss of fermentation activity.

Efforts to improve the properties of dried yeasts have been concerned with questions of yeast composition, strains, and improved methods of drying. With regard to composition, it was found that the trehalose concentration should be 12 % or higher (solid basis). It was also found that compressed yeast used for the production of ADY must be thoroughly washed to reduce the concentration of water-soluble substances in the extracellular aqueous solutions. Retention of residual salt, resulting from acid treatment is particularly harmful during drying.

For a given strain, the fermentation activity is considerably higher in yeast grown to a higher nitrogen level, but such yeasts are also more sensitive to drying. Hence, the search for better methods of drying assumed particular importance. Methods are sought that would permit drying at a lower temperature, in a shorter period of time, and with much smaller particles of press cake to improve the uniformity of moisture removal. Some success has been achieved with a fluid-bed drying method, a drying time of less than two hours, and a yeast strain with a nitrogen content of more than 8 %. The resulting ADY has a smaller particle size and more regular shape than ADY dried in tunnel dryers. This type of dry yeast is called instant active dry yeast (IADY) because it does not require separate rehydration before use.

Numerous patents have claimed certain strains of *Saccharomyces cerevisiae* as particularly useful in production of ADY, often in conjunction with definite processing conditions. Such strains are produced by classical hybridization, protoplast fusion, or mutation.

Many compounds added to the yeast cream or to the press cake are claimed to produce a more active ADY or IADY. Methyl cellulose or carboxymethyl cellulose have been suggested as suitable "swelling" agents at 1 – 2 % of the yeast solids. Emulsifiers such as monoglycerides, soya lecithin, glycerol polyesters, and sorbitan esters will also help maintain the yeast activity. In practice, sorbitan esters are widely used at levels of 0.2 – 1 %. The added emulsifier facilitates processing and lightens the color of ADY or IADY, but the main reason for use of these compounds is the reduction of leached solids when ADY is rehydrated.

Active dry yeast loses activity upon storage when exposed to atmospheric oxygen. Storage stability can be increased by addition of an antioxidant such as butylated hydroxyanisole (BHA) at about 0.1 % based on yeast solids to yeast cream before pressing or filtration. The antioxidant is dispersed with the aid of an emulsifier, preferably a sorbitan ester. The treatment is effective for ADYs with a moisture content of < 6 %. The storage stability of such "protected" ADY is considerably enhanced. However, it does not match the extended stability that can be achieved by complete exclusion of air during storage.

For use in baking ADY must be rehydrated in water at 30 – 40°C. Rehydration at lower water temperature lessens the fermentation activity of ADY. The material leached from the yeast cells, as a result of rehydration, has a slackening effect on doughs due to the extraction of reducing compounds such as glutathione or cysteine. In many dough systems slackening is not desired. However, for bucky dough made from very strong flours, slackening may be desired, but it is better achieved in a controlled fashion by the addition of proteolytic enzymes or cysteine to the dough.

IADY need not be rehydrated in water, but may be mixed in with other dry dough ingredients. Its activity is better upon direct incorporation than with separate rehydration, possibly due to the difference in moisture content between ADY and IADY.

The storage stability of ADY is decreased by higher temperature, higher moisture content of the yeast, and by the presence of air in the package. ADY loses about 7 % of its fermentation activity per month if improperly stored without a protective inert atmosphere at ambient temperature. For storage under nitrogen, or when vacuum packed, the loss is about 1 % per month and generally less than 10 % per year.

Accelerated storage tests under the effect of various atmospheres and temperature approaching 55°C have been tried. Such accelerated tests are useful for purposes of quality control. They do not provide reliable information on commercial conditions unless they have been correlated with long-term storage at normal temperatures.

IADY is always distributed in hermetically sealed pouches in vacuo or flushed with an inert gas. ADY absorbs CO_2 readily. When used with large packages, a sufficient vacuum is created to give the appearance and performance of a vacuum packaging.

The packages may be nitrogen-flushed tin or aluminum foil pouches with low permeability to water vapor and gaseous oxygen. The aluminum foil, which must be free from pin holes, is laminated to a heat-sealable plastic film such as saran or pliofilm.

Once the package has been opened the ADY should be used quickly. In particular, it should not be exposed to an atmosphere with high humidity.

Inactive dried nutritional yeast is produced from a yeast cream of about 18 % solids by pasteurization followed by drum drying or spray drying. The spray-dried product has very fine particles. The drum-dried product that exits the drums as a thin sheet is available as small flakes or ground to a powder.

Spent excess brewer's yeast is recovered in breweries by sedimentation or centrifugation. It contains 12 – 16 % total solids, of which approximately two-thirds are yeast solids and one-third are beer solids. The slurry must be kept refrigerated before use because even at 4°C respiration lowers the carbohydrate content. This yeast may be pasteurized and dried as such to a rather bitter dried brewer's yeast, or the yeast may be separated from the beer by decanter or nozzle centrifuges and partially debittered. It is also pasteurized and dried. Distiller's yeast is not recovered as a separate product because of the presence of other insolubles in the still residues.

3.3.3. *Candida utilis* (Torula) Yeast

Torula is a common commercial trade name for dried nutritional and feed yeasts. The major substrate for its production in the United States are sulfite waste liquors of the paper pulp process, and in Europe and former Eastern Bloc countries they are wood hydrolysates or sulfite waste. *Candida utilis* is grown on wood sugars because it assimilates both pentoses and hexoses, and its growth requires the addition of nitrogen sources, phosphate, and potassium to the medium. In contrast to *S. cerevisiae*, the Candida yeast does not require the addition of biotin.

*C. utilis*is produced by continuous homogeneous fermentation, which can be run for months. The yield of yeast is generally 50 % based on fermentable pentoses plus hexoses. The yeast is recovered by centrifugation. After minimal washing with water *C. utilis* can be sold as feed yeast, which contains a considerable amount of lignosulfonic acids. For use as human food, the yeast must be centrifuged at least twice with extensive washing. Dried yeast is obtained by drum or spray drying.

3.3.4. Dairy Yeasts

Large amounts of cheese whey are available as a byproduct of cheese production. The whey may be dried and sold as food or as feed ingredients, but supply by far surpasses demand. Thus, there exists a problem of waste disposal and there have been numerous attempts to use whey as a fermentation substrate for the production of ethanol and biomass.

Dairy yeasts are lactose-assimilating and lactose-fermenting yeasts; the preferred species for the production of biomass is *Kluyveromyces marxianus.*

3.3.5. Brewer's Yeast

Yeast plays an important role in the brewing process not only by converting fermentable sugars in the wort to ethanol and carbon dioxide but also by producing a variety of volatile and nonvolatile constituents, mostly in trace amounts, that contribute to the whole flavor of the beer.

Two species of the genus Saccharomyces are often considered important for brewing depending on the type of the brewing process.

Most beers are produced by the bottom fermentation process in which strains of the yeast *S. uvarum* are used. These yeasts have a tendency to remain suspended for a limited time in the fermentation medium while the fermentation is active, after which they settle to the bottom of the fermentation vessel. This strain is predominantly used in Germany and North America for the production of lager beer.

The other type of beer, referred to as ale, is produced by a top fermenting yeast such as *Saccharomyces cerevisiae*. The strains tend to rise to the top of the fermenting medium and are generally skimmed off the surface of the fermenting liquid after the fermentation is completed. Ales are popular in England and Ireland.

Malt extract or beer wort is one of the oldest classical media for growing yeast cultures. It contains nitrogen, mineral salts, and vitamins in amounts adequate for maintaining yeast growth, and fermentable carbohydrates to supply carbon and energy for biosynthesis. There is also an absolute requirement for molecular oxygen. The yeast can, however, exhibit some growth under completely anaerobic conditions.

3.3.6. Wine Yeasts

Fermentations of fruits and fruit juices occur naturally and spontaneously without human intervention. Ripe berries can undergo spontaneous fermentation in the field if they are overripe or damaged.

These fermentations are initiated by yeasts found on the grape skin. Yeasts survive from year to year in the intestines of bees and wasps and they are readily transferred during the crushing season by fruit flies (*Drosophila melanogaster*) during the transport of grapes to the winery and within the winery. Once the crushing season is underway, the major source of yeasts is likely to be found on winery equipment such as crushers, destemmers, and tanks.

The terms brewer's yeast and baker's yeast are restricted to certain specific strains of *Saccharomyces cerevisiae*. These yeasts are almost always produced from selected pure strains of this species. In contrast, wine yeasts are strains of various species participating in must fermentation.

In a broad sense, any yeast species occurring on grapes and participating in must fermentation, no matter how transient, may be called a natural yeast. In a narrow sense, strains grown on pure culture in winery fermentations are wine yeasts. Such strains generally belong to the species *Saccharomyces cerevisiae*.

A large number of yeast cells found on grapes belong to the genera Kloeckera, Hanseniaspora, and to a lesser extent to the genera Metschnikowia and Candida.

There is a considerable increase in cell counts as the grapes are transported and the must is prepared in the winery. The fermentation is initiated by weakly fermenting yeast, mainly *Kloeckera apiculata*. The early fermenters are replaced by more alcohol-tolerant yeasts, and during the main part of the fermentation and toward the end the true wine yeasts *S. cerevisiae* dominate.

3.3.7. Distiller's Yeasts

Distiller's yeasts include strains of *S. cerevisiae*, except for natural fermentation, in which any number of species may be involved. Yeasts suitable for grain fermentations can be baker's compressed yeast, distiller's active dry yeast, or strains grown in the distillery from selected pure-culture slants.

The fermentation of the various raw materials is best understood by categorizing them into the fermentation of cereal mashes, which contain all the constituents of cereal (whisky and sake), and the fermentation of juices and liquid extracts, which contain all carbohydrates in the form of fermentable sugars (fruit juices, molasses, and fully hydrolyzed starches). Cereal mashes are characterized by the concurrent enzymatic formation of sugars, and their fermentation, at least in the latter stages of the process.

4. Yeast-Derived Products

4.1. Flavor Products

Yeast extracts are commercially available as a powder or paste and have been used extensively by the food industry as flavoring agents. They are concentrates of the soluble fraction of yeast, generally made following autolysis, which is essentially a degradation process carried out by activating the yeast's own degradation enzymes to solubilize the within-cell components.

Yeast extracts can also be made by hydrolysis with external reagents or enzymes capable of releasing the cell contents. Yeast extracts have received wide acceptance in the food industry as natural flavors. They are also cost-effective compared to other synthetic food flavors, on the basis of equivalent flavor intensity.

They are used extensively in the fermentation industry and as a complex nutritional source for the preparation of microbiological growth media for plate assays. The production of yeast extracts worldwide is ca. 32 000 t.

Surplus yeast from the brewing industry is relatively cheap and is utilized in the production of extracts to meet the needs of the food and fermentation industries.

Extracts made from primary-grown yeast such as *Saccharomyces cerevisiae* (baker's), *Candida utilis*, or *Kluyveromyces marianus* have flavor profiles somewhat different from those of brewer's yeast. The latter is likely to have undesirable flavor characteristics as a result of a carryover of hop resins and beer solids.

Autolysate is the total content of yeast following autolysis. The product is in the form of a concentrated paste or dried without separation of the particulate matter. Consequently, autolysates are less intense in flavor than extracts on an equal weight basis because of the dilution effect of the cell-wall material.

4.2. Nutritional Yeast

Yeast is generally recognized as an excellent source of proteins, B vitamins, fiber, and many other micronutrients. However, excessive intake of yeast has caused serious human health problems which were later identified as being due to excessive consumption of nucleic acid present in yeast. Decrease of nucleic acid content is often carried out following the separation of yeast proteins in crude form from the whole cell. Yeast is ruptured using several methods including high-pressure homogenization, colloid mill, sonic desintegration, freeze – thaw treatment, use of lytic enzymes, or a combination of these methods for maximum recovery. Several methods to lower the level of nucleic acid have been described in the literature.

4.3. Colorants Derived from Yeast

In the last decade or so, there has been considerable interest in developing food colorants from natural products. Although different types of natural pigments are widely distributed in the plant and animal kingdoms, a variety of carotenoid pigments of various pink and yellow shades are also produced by a number of yeast species. Yeast offers considerable advantages over other natural sources in that it is inexpensive to grow on an industrial scale. Yeast has now begun to emerge as a cost-effective source for the production of natural food colorants.

The best-known genera among the pigmented yeasts are Rhodotorula, Rhodospondium, Cryptococcus, Sporidiobolus, and Sporobolomyces. These yeasts are characterized by their ability to produce pigments such as β-carotene, α-carotene, torulene, torularhodin, plectanisxanthin, and 2-hydroxyplectanisxanthin.

Carotenoids produced by different Rhodotorula species contain primarily the two oxygenated compounds torulene and torularhodin, and the ratio of the two major pigments determines the intensity of the shades of red in different species.

Phaffia rhodozyma is a basidiomycetous yeast that differs from other pigmented yeasts in producing astaxanthin, which is widely distributed in the animal kingdom. Much of the interest in astaxanthin is due to the fact that it is responsible for the characteristic color in the plumage of birds such as flamingos, in the exoskeleton of marine invertebrates such as lobsters, crabs, and shrimps, and in fish such as trout and salmon, whose pink color is due to the storage of the pigment in their flesh. Lack of the characteristic pink color in pen-reared salmonids renders them less acceptable. Astax-

anthin can restore the pink color in seafood, resulting in better marketability.

4.4. Yeast-Derived Enzymes

Invertase is an enzyme capable of splitting the sucrose molecule into glucose and fructose, a noncrystallizable mixture known as invert sugar. It is sweeter than sucrose on an equivalent weight basis due to the higher sweetening power of fructose. As a result, invert sugar is generally regarded as an important ingredient in the food and confectionery industries.

Baker's yeast is rich in invertase and is the main source of commercially available invertase.

There are two important applications of invertase in the food industry. The first is in the manufacturing of confectionaries, where invertase is used mainly in the production of soft-centered candies. The use of invertase provides the plastic consistency to the center and prevents sucrose from crystallizing during storage. The enzymatic hydrolysis that brings about the creamy consistency of the center generally takes place after casting and covering the hard center with chocolate. The second important application is in the inversion of syrups for the production of high-test molasses, used extensively in the fermentation industry. High-test molasses, if not inverted, would soon tend to crystallize due to high concentration of sucrose. Inversion also raises the osmotic pressure and results in a microbiologically stable concentrate with a consistency that is easy to handle.

The most common method of extracting and concentrating invertase from yeast is by autolysis and to recover the enzyme from the filtered autolysate by precipitation or concentration. The precipitating agents commonly used are ethanol, acetone, and ammonium sulfate. The precipitate thus formed is dissolved in water, filtered, and treated with activated carbon to remove coloring matter, followed by concentration under vacuum, or by ultrafiltration.

Invertase is commercially available as a concentrate stabilized with glycerol.

Lactase catalyzes the hydrolysis of lactose (milk sugar) into its principal components, glucose and galactose. This enzyme is widely used in the food industry for the treatment of frozen whole-milk concentrates to prevent crystallization during storage, and most importantly to make milk-based products acceptable for those who cannot tolerate lactose.

Lactase is not present in baker's yeast, but it is present in a number of other common yeasts.

Melibiase hydrolyzes the disaccharide melibiose to glucose and galactose. Consequently, there is likely to be a potential market for this enzyme, specifically to improve the production yield in the beet sugar industry.

In the production of beet sugar, raffinose, a trisaccharide (galactose – glucose – fructose), is known to hamper the crystallization process and thereby reduce extract yields. This results in a high concentration of raffinose and sucrose in the final molasses. Melibiase splits raffinose into galactose and sucrose. This treatment not only increases the recovery of sucrose in the syrup but also minimizes the raffinose content, therefore enhancing the beet sugar crystallization process.

A potential application for melibiose is in the baker's yeast industry. Although the predominant sugar in both beet and cane molasses is sucrose, beet molasses contains 1 – 1.5 % raffinose (solid basis). Only one-third of this trisaccharide is utilized by baker's yeast due to the presence of invertase, which converts raffinose into melibiose and fructose. The latter is assimilated by the yeast. For complete utilization of raffinose molecule the enzyme melibiase should either be added to the wort, wort should pass through a column of beads containing immobilized melibiase, or the baker's yeast strain must be genetically altered to produce melibiase.

Other enzymes produced commercially by yeasts include alcohol dehydrogenase, hexokinase, L-lactate dehydrogenase, glucose-6-phosphate dehydrogenase, phenylalanine ammonia lyase, and glyceraldehyde-3-phosphate dehydrogenase.

4.5. Pharmaceutical and Cosmetic Products

Skin Respiratory Factor. A yeast-derived product with wound-healing properties is commercially available for the production of hemorrhoidal preparations and is often referred to as skin respiratory factor (SRF) of live yeast cell derivative (LYCD). SRF prepared from yeast is known to exhibit certain biological properties

capable of enhancing the oxygen uptake rate, thereby stimulating wound healing. Its effectiveness in skin care has made it a commercially attractive ingredient for the cosmetic and pharmaceutical industries.

Glycan. The term glycan is used broadly to refer to commercial preparations of the insoluble yeast cell wall fraction. It is generally recovered as a byproduct during the production of soluble extracts from baker's or brewer's yeast. Yeast cell wall is made up of glucan, mannan, and chitin polymers that are nutritionally nonfunctional in the human digestive tract. The three components account for more than 80 % of the dry matter of the glycan fraction, and about 20 % is noncarbohydrate material, including protein, nucleic acid, lipids, and ash.

A desirable property of glycan is its ability to function as a noncaloric food thickener. It provides a fat-like mouthfeel, thus permitting its use as a fat replacer in the production of low-calorie salad dressings, cheese analogs, and frozen desserts such as ice cream.

Its high water-holding capacity has made it a useful ingredient in the formulation of sausages, hot dogs and certain other meat-based products.

More recent research has begun to show certain unique properties in glycan that are likely to permit its use as a remedy for certain human disorders.

Recent studies have demonstrated the beneficial effect of glycan in improving host resistance by enhancing both humoral and cellular immunity to certain malignant tumors and viral infections. The therapeutic effect of glycan is primarily due to stimulation of the reticuloendothelial system (RES), which in turn produces increased amounts of macrophages that play a key role in the body's natural immune system. The macrophages absorb and destroy invading particles by phagocytosis.

Coenzyme A Synthesizing Protein Complex. Research is presently underway to develop methods for the detection of cancer at an early stage, long before any symptoms become apparent in the patient. One such diagnostic techniques involves the early detection of β-protein in the bloodstream, which is an early indication of the body's reaction to cancer.

A biologically active compound referred to as coenzyme A synthesizing protein complex (COA-SPC), derived from baker's yeast, has found application as one of the components used for detecting β-protein in human blood.

The biologically active COA-SPC complex has been identified in the particulate fraction following the lysis of baker's yeast.

Genetically Engineered Products from Yeast. With the technological advances that have already been made in genetic engineering, substantial effort has now gone into the development of methodology to manipulate and modify the yeast genome. Yeast has become one of the most frequently exploited microorganisms, capable of producing valuable products of commercial importance (see also → Genetic Engineering, Chap. 4.2.1., → Genetic Engineering, Chap. 4.2.2.).

Despite the exceedingly long lag time and high expense of putting new products through the entire approval process, at least a few biotechnology companies have been able to bring certain genetically engineered yeast-derived products into the final stages of marketing. More genetically engineered products are in human clinical trials and are being evaluated for final approval.

At present, the list of commercial products includes only three well-known products: human insulin, interferon, and hepatitis-B vaccine.

5. References

1. A. Haláz, Radomir Lásztity: *Use of Yeast Biomass in Food Production,* CRC Press, Boca Raton, FL, 1991.
2. G. Reed, T. W. Nagodawithana: *Yeast Technology,* 2nd ed., AVI, New York 1991.

Ytterbium → Rare Earth Elements
Yttrium → Rare Earth Elements

Zeolites

ECKEHART ROLAND, Degussa AG, ZN Wolfgang, Hanau, Federal Republicof Germany

PETER KLEINSCHMIT, Degussa AG, ZN Wolfgang, Hanau, Federal Republicof Germany

1. Introduction 1
2. Nomenclature 2
3. Structure, Composition, and Properties of Zeolites 3
3.1. Framework Structure and Positions of Nonframework Cations 3
3.2. Chemical and Physical Properties . 6
4. Characterization 7
5. Natural Zeolites 8
5.1. Occurrence, Mining, and Processing 8
5.2. Commercial Applications and Importance 9
6. Production of Synthetic Zeolites . . 9
6.1. Zeolite Synthesis: Routes and Raw Materials 10
6.1.1. Synthesis from Gels 10
6.1.2. Synthesis from Clay Minerals 11
6.2. Examples of Industrial Zeolite Syntheses 12
6.3. Modification and Formulation . . . 15
6.3.1. Ion Exchange 15
6.3.2. Dealumination 17
6.3.3. Other Methods of Zeolite Modification 18
6.3.4. Shaping 18
6.4. Environmental Aspects 18
7. Applications of Synthetic Zeolites . 19
7.1. Ion Exchange: Zeolites as Detergent Builders 19
7.2. Adsorption 20
7.3. Catalysis 23
7.4. Other Uses 26
8. Economic Aspects of Synthetic Zeolites 26
9. Toxicology 27
10. References 28

1. Introduction

The exact definition of the term "zeolite" is still the subject of discussion [1–4]. A version in general use is due to BARRER [1] and BRECK [5], which states that zeolites [*1318-02-1*] are crystalline, hydrated aluminosilicates with a framework structure. Their three-dimensional, polyanionic networks are constructed of SiO_4 and AlO_4 tetrahedra linked through oxygen atoms. Depending on the structure type, they contain regular channels or interlinked voids whose aperture diameters are in the micropore range. These pores contain water molecules and the cations necessary to balance the negative charge of the framework. The cations, which are mobile and can be exchanged, are mainly alkali metal or alkaline earth metal ions.

Formula **1** accords with these characteristics. The expression enclosed in the square brackets shows the composition of the anionic framework in the crystallographic unit cell. M represents the nonframework metal cation, and n is its charge. Formula **2** occurs frequently in the literature and, unlike Formula **1**, simply represents the overall chemical composition.

$$M_{x/n}[(AlO_2)_x(SiO_2)_y] \cdot z\,H_2O \quad \mathbf{1}$$

$$M_{2/n}O \cdot Al_2O_3 \cdot y'\,SiO_2 \cdot z'\,H_2O \quad \mathbf{2}$$

Many synthetic zeolites also occur naturally as minerals. The observation that a certain mineral began to bubble on strong heating led CRONSTEDT in 1756 to give the name "zeolites" to such materials from the Greek *zeo*: to boil and *lithos*: stone [6], [7]. However, zeolites only became of industrial importance in the 1950s, when synthetic examples became available on an industrial scale [8]. Since then, world consumption of synthetic zeolites in the fields of ion exchange, adsorption, and catalysis has grown continuously. The total for 1993 exceeded 10^6 t, mainly a consequence of the use of the zeolite NaA as a detergent builder. Zeolites are also important as adsorbents for separating and purifying substances and as catalysts in a number of important processes in the chemical and petrochemical industries.

In addition, ca. 300 000 t/a natural zeolites are sold worldwide. Their main uses are as ion ex-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_475

changers for water treatment, in soil improvement, in animal feeds, in cat litter, and in pozzolanic cement.

The wide range of commerical applications has stimulated intensive investigation of the chemistry and physics of the zeolites [5], [9–19]. Current research is increasingly concerned with zeolite-like molecular sieves, which have expanded the field of zeolites enormously during the last two decades. In these new materials, the aluminum and silicon of the classical zeolites are partially or completely replaced by other elements [20], [21]. The most important groups of these molecular sieves are:

1) Modifications of silica ("zeosils," see → Silica, Chap. 8.2.1.)
2) Metallosilicates with tri- and tetravalent heteroatoms in place of aluminum as framework components, e.g., boro- and gallosilicates [22], ferrisilicates [23], and titanosilicates [24]
3) Alumino- and gallophosphates and compounds derived from them [25–27]

These compounds generally have framework structures analogous to those of the zeolites. Some new structures have been discovered, too. However, with a few exceptions, these compounds are still of rather limited economic importance, and are discussed here only briefly.

2. Nomenclature

The naming of zeolites in the literature seldom follows a scientific system. If natural analogs exist, the synthetic zeolites are often named after the minerals (e.g., faujasite, mordenite, ferrierite, and offretite). Alternatively, the names given by the discoverers are used. In the 1950s and early 1960s, new synthetic zeolites were named using the letters of the Latin and Greek alphabets. Examples include the Linde molecular sieves A, L, X, and Y of Union Carbide and the zeolites beta, rho, theta-1, and omega. Acronyms that indicate the industrial or university laboratory that first synthesized a given material were then increasingly used, supplemented by a consecutive numbering system. Names such as ZSM-5 (Mobil Oil), Nu-13 (ICI), LZ-210 (Union Carbide/UOP), CSZ-3 (Grace), and ECR-1 (Exxon) have gradually become generally accepted.

Zeolites with identical framework structures can therefore appear in the literature under different names. Some order has been imposed by the introduction of a classification system in the *Atlas of Zeolite Structure Types* [28], which indicates which different materials have the same structure type. The structure types themselves are denoted by the combination of three capital letters. For example, the type LTA is found in zeolite A, FAU in the faujasites X and Y, and MFI in the ZSM-5 family of zeolites.

CAS numbers are available for only a few zeolites. Moreover, as these are rarely used, they are of little benefit in literature searches.

3. Structure, Composition, and Properties of Zeolites

3.1. Framework Structure and Positions of Nonframework Cations

The third edition of the *Atlas of Zeolite Structure Types* [28], published in 1992, describes 85 different structures, including both classical zeolites and zeolite-like molecular sieves. The majority of these structures can be represented by Formulas **1** and **2**, but a few have been found to occur only as alumino- or gallophosphates.

Zeolites are classified structurally as tectosilicates [3]. The basic elements of their frameworks are TO_4 tetrahedra (T = Si, Al) linked through oxygen atoms. The description and classification of the topology of zeolites is based on the concept of larger units known as secondary building units (SBUs) [28]. These can consist, for example, of simple rings and prisms of various sizes. By combining such basic units the known zeolite frameworks can be constructed. Alternative concepts are based on more complex chains or layer structures known as structural subunits (SSUs) [29].

The most readily visualized structural units are polyhedra such as the truncated octahedra known as sodalite cages or β-cages, found in the structure of the industrially important zeolites A, X, and Y. Figure 1 shows the sodalite unit and the framework structures of the zeolites. In this representation, the tetrahedrally coordinated silicon and aluminum atoms are symbolized by

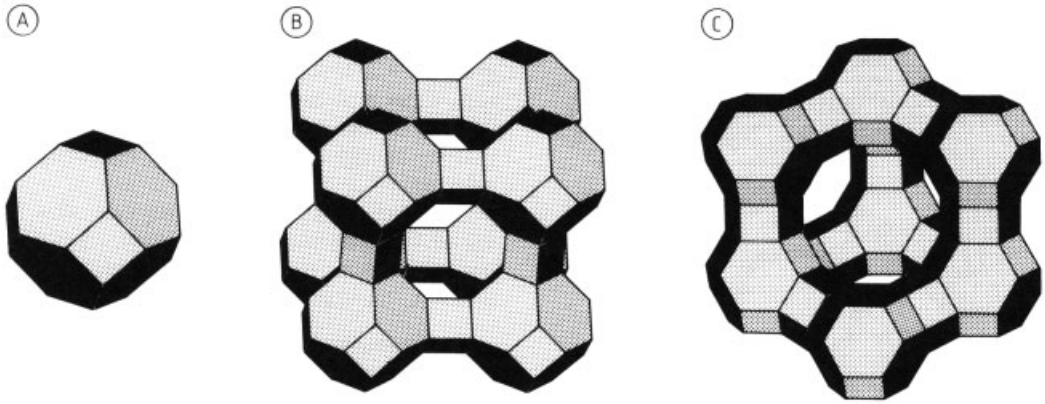

Figure 1. A) Sodalite cage or β-cage; B) Framework structure of zeolite A (structure type LTA) with α-cage; C) Framework structure of faujasite (structure type FAU; zeolites X, Y) with supercage

vertices, and the oxygen atoms are located between these.

The principal characteristic of all the structures is the well-defined system of regular cavities or channels which is essential for application of zeolites as molecular sieves. In zeolite A (type LTA), the β-cages are linked together via cubic units to form a larger cavity with the shape of a truncated cubo-octahedron. These so-called α-cages (diameter 1.14 nm) are interconnected by windows formed by rings consisting of 8 T atoms, giving a three-dimensional pore system. If the β-cages are linked together by hexagonal prisms, this results in the faujasite structure (type FAU). In this case also, the linking leads to a larger cage. This is known as a supercage because of its large diameter (ca. 1.20 nm), and has windows formed by 12 T atoms.

Whereas the cavities in the LTA and FAU structure types are characteristically in the form of interlinked cages, the MOR type (mordenite) and the MFI type (e.g., ZSM-5) contain channel systems, as shown in Figure 2. The MFI structure is easier to visualize if only the cavity structure is represented, and not the framework structure. Linear parallel ten-ring channels are linked together by zig-zag shaped continuous pores with ten-ring apertures perpendicular to the channels, resulting in a three-dimensional cavity system. In contrast, the pore system of mordenite is two-dimensional, the 12-ring channels being in effect only one-dimensional.

The various zeolite structures differ not only in the type and dimensionality of their pore systems, but also in the size of the pore apertures. Narrow-pore, medium-pore, and wide-pore zeolites have different pore apertures formed by rings of 8, 10, or 12 T atoms, with corresponding crystallographic diameters of 0.35 – 0.45, 0.45 – 0.60, and 0.60 – 0.80 nm [5], [30]. In the zeolites themselves, pore apertures larger than these have not so far been observed, although they do occur in the alumino- and gallophosphates, in which structures containing 14-, 18-, and 20-ring apertures are known, with diameters ranging from 0.80 to 1.45 nm [31], [32]. The pore apertures of the industrially important synthetic zeolites are listed in Table 1, together with other characteristic data for these materials [5], [30], [33]. It can also be seen from the table that the apertures can deviate to varying extents from circular. For example, the cross-section of the zig-zag pores in ZSM-5 is distinctly elliptical.

The regular nature of the pores and their apertures, whose dimensions are of the same order of magnitude as molecular diameters, enables the zeolites to function as molecular sieves. This is the outstanding property of zeolites that gives them their value as selective adsorbents for separating substances and as shape-selective catalysts (see Sections 7.2 and 7.3). Depending on the zeolite type and its pore system, molecules can penetrate into the cavity system or be excluded from it (Fig. 3).

The concepts of kinetic molecular diameter and effective pore size take account of the fact that the adsorptive molecules and the zeolite framework are not rigid, but undergo thermal vi-

brations. Therefore, the selectivity of the exclusion of molecules whose diameters are similar to that of the zeolite apertures determined crystallographically depends on temperature. Thus, the crystallographic diameter of the 6-ring aperture of the β-cage, as in type A, X, and Y zeolites, is given as 0.22 nm. However, at room temperature, water molecules with a kinetic diameter of 0.26 nm can readily penetrate into the β-cage, so that this should be regarded as the effective pore aperture.

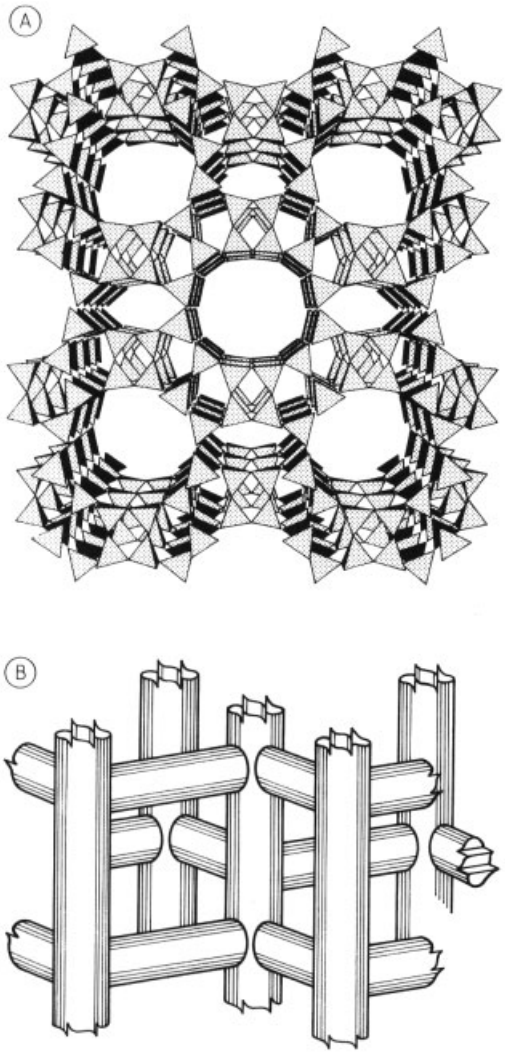

Figure 2. A) Framework structure of mordenite (structure type MOR) represented as linked SiO_4/AlO_4 tetrahedra; B) Hollow tube representation of ZSM-5 (structure type MFI)

The apparent pore diameter can also be varied by using cations of different sizes. Figure 3 illustrates this for the example of zeolites KA, NaA, and CaA.

Nonframework Cations. The cations needed to balance the negative charge on the zeolite framework occupy definite positions in the cavities and channels, where they are surrounded by partially negatively charged oxygen atoms from the rings of the framework. The framework – cation bond thus has a mainly ionic character. The preference of particular cations for particular available sites can be due to energetic, steric, and coordinative considerations [34–36]. In the presence of adsorptive molecules, a redistribution of the cations is possible. The nature and distribution of the cations in the framework are of importance when zeolites are used as adsorbents and catalysts. Any change can lead to differences in both apparent pore diameter and charge distribution.

Positions and distributions of cations in A-type and faujasite-type zeolites have been researched most thoroughly [35], [36]. In the latter type, the most important positions are the centers of the hexagonal prisms, in which octahedral coordination is possible, and positions close to the centers of the hexagonal surfaces in the supercage and β-cages (see Figure 1 C). The pattern of occupation of available sites depends on the particular cations and on whether the zeolite is anhydrous or hydrated.

A special situation arises if the zeolite is converted into the so-called hydrogen form in which the framework charge is formally balanced by H^+. In this case, acidic OH groups are bonded to the framework (see Section 7.3).

3.2. Chemical and Physical Properties

Most of the chemical and physical properties of the zeolites, and hence their areas of use, are essentially determined by the aluminum content of their frameworks. In the literature, this is usually expressed by the Si/Al or SiO_2/Al_2O_3 ratio (y/x in Formula **1** and $y' \approx 2\,y/x$ in Formula **2**).

According to the Loewenstein Rule, Al – O – Al groups cannot occur in crystalline aluminosilicates [37]. Hence, an SiO_2/Al_2O_3 ratio of < 2 is impossible for zeolites. As can be

Table 1. Structures and compositions of some industrially important synthetic zeolites

Zeolite	Structure type [28]	Dimenisonality of pore system	Window	Pore apertures, nm	SiO_2/Al_2O_3	Typical composition of unit cell
A	LTA	3	8-ring	0.41	2.0 – 6.8	$Na_{12}[(AlO_2)_{12}(SiO_2)_{12}] \cdot 27\,H_2O$
P	GIS	3	8-ring	0.31 × 0.45 0.28 × 0.48	2.0 – 5.0	$Na_6[(AlO_2)_6(SiO_2)_{10}] \cdot 12\,H_2O$
ZSM-5	MFI	3	10-ring 10-ring	0.53 × 0.56 0.51 × 0.55	25 – ∞	$(Na,TPA)_3[(AlO_2)_3(SiO_2)_{93}] \cdot 16\,H_2O$ *
X	FAU	3	12-ring	0.74	2.0 – 3.0	$Na_{86}[(AlO_2)_{86}(SiO_2)_{106}] \cdot 264\,H_2O$
Y	FAU	3	12-ring	0.74	3.0 – 6.0	$Na_{56}[(AlO_2)_{56}(SiO_2)_{136}] \cdot 250\,H_2O$
Mordenite	MOR	2	12-ring 8-ring	0.65 × 0.70 0.26 × 0.57	9.0 – 35	$Na_8[(AlO_2)_8(SiO_2)_{40}] \cdot 24\,H_2O$

* TPA = Tetrapropylammonium.

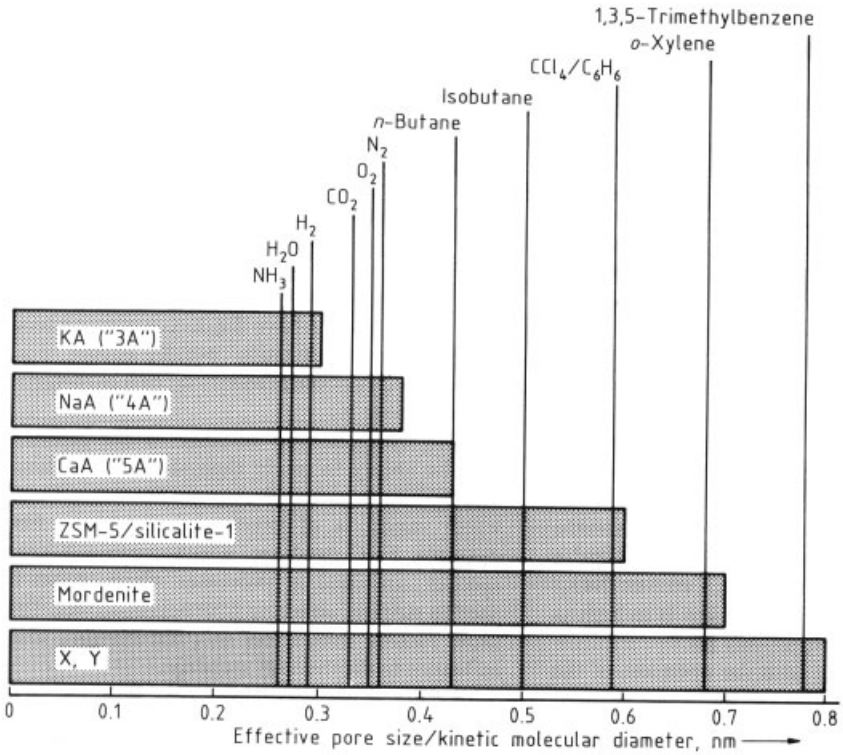

Figure 3. Correlation between the effective pore size of industrially important zeolites and the kinetic molecular diameter of selected compounds (room temperature)

seen for the synthetic zeolites in Table 1, each structure type exhibits a phase breadth with respect to the SiO_2/Al_2O_3 ratio. A distinction can be made between low-silica ($SiO_2/Al_2O_3 < 4$), intermediate-silica ($4 < SiO_2/Al_2O_3 < 20$), and high-silica ($20 < SiO_2/Al_2O_3 < 200$) zeolites, and zeosils, in which aluminum is present only as a trace impurity (→ Silica, Chap. 8.2.1.) [38]. In some intermediate-silica zeolites, a higher SiO_2/Al_2O_3 ratio can be produced by post-synthetic dealumination (see Section 6.3.2).

The aluminum atoms may not be uniformly distributed over the cross section of the zeolite crystal. More often, and especially with synthetic high-silica zeolites, a zoning effect has been observed [39], [40], which is important in catalytic applications.

Physical Properties. Synthetic zeolites containing alkali metal and alkaline earth metal cations are normally colorless crystalline powders. Colors only occur if the zeolites contain transition metals as impurities or if they have been modified by ion exchange with these elements. As a rule, the primary particles of industrially synthesized zeolites have an average size of 0.1 – 15 µm, although crystals with diameters of up to 5 mm can be produced under special conditions [41]. The single crystals often form larger aggregates. Like slurries of many other finely divided substances with surface-active groups, suspensions of zeolites can exhibit thixotropy or rheopexy, which is relevant in production processes.

On heating, hydrated zeolites release water. Many zeolites can be almost completely freed of adsorbed water without major alteration of their crystal structure by calcination at 400 – 500 °C, leaving solids with large free pore volumes and interior surfaces whose extent depends on the framework density of the particular structure type [28] and on the number and nature of the cations. The cavities are then available for the adsorption of guest molecules. The micropore volume of, e.g., H-ZSM-5 ($SiO_2/Al_2O_3 = 75$) is 0.18 cm^3/g [42], and that of the NaA and NaX zeolites ($SiO_2/ Al_2O_3 = 2.5$), including the β-cages, is as high as 0.29 or 0.36 cm^3/g, respectively [5].

Densities also vary with the openness of the framework structure and the nature of the cations, generally lying in the range 1.9 – 2.3 g/cm^3. If the zeolite contains heavy cations such as Ba^{2+}, densities can be higher.

The surface selectivity of the zeolites as adsorbents depends on the SiO_2/Al_2O_3 ratio. Aluminum-rich zeolites preferentially adsorb strongly polar molecules, and are therefore widely used as drying agents (see Section 7.2). Increasing silicon content leads to increasingly hydrophobic character, the transition from hydrophilic to hydrophobic behavior occurring at a SiO_2/Al_2O_3 ratio of ca. 20 [38].

Due to the mobility of the cations in their cavities and channels, zeolites exhibit ionic conductivity [5, pp. 397 – 410], which depends on the diameter of the channels, the nature and concentration of the cations, and the water content. Control of these parameters and a wide range of other techniques enable the conductivity properties of zeolites to be modified to make them suitable for future use as solid-state electrolytes, membranes in ion-selective electrodes, and components of cathode materials in batteries [43], [44].

Chemical Properties. In spite of their open structure, the zeolites have remarkably good thermal stability, which varies with the structure type, but more so with the SiO_2/Al_2O_3 ratio and the nature of the cations. The effect of the SiO_2/Al_2O_3 ratio is illustrated by the example of faujasite. In a low-silica NaX zeolite ($SiO_2/Al_2O_3 = 2.5$), structural collapse was observed above 660 °C, while a structurally equivalent NaY zeolite ($SiO_2/Al_2O_3 = 5.0$) became amorphous above 700 °C [5, p. 495]. Highly dealuminated Y type zeolites, however, are structurally stable at >1000 °C [45]. The effect of the cation can be seen in zeolite A, the Ca form losing its structure above 700 °C, and the Na form above 660 °C [5, p. 495]. Above the decomposition temperature, recrystallization usually occurs, e.g., forming β-cristobalite (zeolite A) or a carnegieite phase (zeolite X).

The hydrothermal stability of the zeolites, i.e., the ability to retain their structure without undergoing a phase change in the presence of water vapor, is of practical importance. For example, FCC catalysts, which contain Y-type zeolites, are used in processes where steam is used to strip off adhering hydrocarbons [46]. The same factors that affect thermal stability also influence hydrothermal stability.

Strong acids decompose low-silica zeolites such as NaA and NaX by dissolving the aluminum atoms out of the framework, with consequent breakdown of the crystal structure [47]. With increasing SiO_2/Al_2O_3 ratio the zeolites become less sensitive to dealumination by acids (see Section 6.3.2); high-silica zeolites are structurally stable even in strong mineral acids [45].

Aqueous alkalis lead to phase transitions in high-aluminum zeolites. Thus, sodium hydroxide solution converts NaA zeolite into hydroxysodalite, which must be taken into account in production (see Section 6.2). Very high silica zeolites are dissolved by strong alkalis [48].

Suspensions of high-aluminum zeolites in which the cation is Na^+ have a pH of 9 – 12 due to partial hydrolysis.

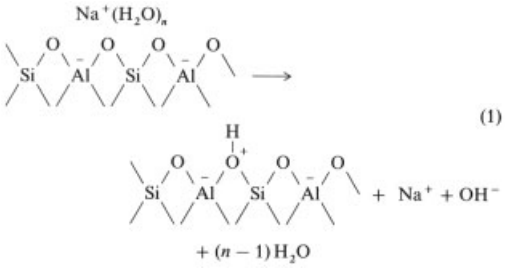

(1)

The ion-exchange capacity of zeolites decreases with increasing SiO_2/Al_2O_3 ratio (see Sections 6.3.1 and 7.1). The catalytic properties of the hydrogen form, in which H^+ ions formally replace the metal ions, are discussed in Section 7.3.

4. Characterization

In principle, all methods used for investigating solids can be used for zeolites [49]. Moreover, the availability of modern physicochemical methods has significantly promoted the development of synthetic zeolites with tailored properties.

The size and shape of the zeolite crystals or aggregates must be investigated for most applications. This is carried out by microscopy, generally with a scanning electron microscope because of the small particle size. The particle size distribution can be rapidly determined by laser scattering or the Coulter counter method on zeolite suspensions.

The most important technique used to determine structure type and phase purity is X-ray diffraction (XRD). For known structures, a classification can be carried out by using the *Collection of Simulated XRD Powder Patterns for Zeolites* [50]. When structures of new materials are determined, the crystals available are seldom large enough for single-crystal X-ray structure analysis. Powder diffraction methods are therefore usual, sometimes with use of synchroton radiation [51]. The Rietveld method of profile analysis of the diffraction pattern has been developed into a very powerful tool [52]. XRD can only be used to estimate the fraction of crystalline material in a mixture with amorphous impurities in a zeolite sample by comparison with a standard material. This so-called crystallinity is usually determined by comparing the intensities of selected reflections [53].

The thermal stability of zeolites can be determined by XRD or differential thermal analysis.

Zeolites can be analyzed for elements by classical wet chemical methods, but modern physical methods are usually preferred, i.e., X-ray fluorescence analysis (XRF), atomic absorption spectrometry (AA), and inductively coupled plasma spectrometry (ICP). X-ray photoelectron spectroscopy (XPS) is particularly useful for obtaining information about the composition of the outer atomic layers of a zeolite crystal.

The determination of the SiO_2/Al_2O_3 ratio of zeolites is of importance in many of their applications [54]. While the SiO_2/Al_2O_3 ratio of the bulk sample can easily be analyzed by the above-mentioned methods, more detailed information can be obtained by high-resolution magic angle spinning NMR spectroscopy. Thus ^{29}Si MAS-NMR spectroscopy is very suitable for determining the SiO_2/Al_2O_3 ratio of the zeolite framework, especially if the sample also contains aluminum that is not part of the framework. ^{29}Si and ^{27}Al MAS-NMR spectroscopy also gives important information on the short-range structural environment of the silicon and aluminum atoms [55–57]. Thus, the five groups $Si(OSi)_{4-n}(OAl)_n$ ($n = 0-4$) give signals with different chemical shifts in the ^{29}Si MAS-NMR spectrum. In the ^{27}Al MAS-NMR spectrum, aluminum atoms in different states of coordination give separate signals. In zeolites of the faujasite type, the SiO_2/Al_2O_3 ratio can also be estimated from measurements of the crystallographic unit cell [58] and from the position of the framework-vibration bands in the IR spectrum [59].

Information about the characteristics of the pore system can be obtained from adsorption measurements. The size of the apertures can be estimated by testing their accessibility to adsorptives of various kinetic diameters. However, determination of the inner surface area and the micropore volume from the adsorption and desorption isotherms of nitrogen is problematic [60]. ^{129}Xe MAS-NMR spectroscopy is a relatively new method of investigating the cavity systems in large-pore zeolites, the adsorbed xenon atoms acting as probes [61]. Test catalytic reactions, the outcome of which depends on the shape selectivity of the zeolite (see Section 7.3), can also be used to characterize pore systems. Depending

on the reaction, the characteristic value obtained is known as the constraint index, the modified constraint index, or the spaciousness index [62].

In the development of zeolite catalysts, it is also important to characterize their acidity properties, i.e., the nature, number, and strength of the acid sites [49], [63]. Methods include temperature-programmed desorption (TPD) of adsorbed bases such as ammonia and pyridine, ^{1}H MAS-NMR spectroscopy [64], IR spectroscopy in the region of O – H stretching frequencies [65], and measurement of the catalytic activity of the zeolites in standard reactions [63].

In addition to these general methods of investigation, zeolites are also characterized in industry by parameters that relate to a particular application. For example, where zeolites are used as detergent builders, these can include calcium binding power (determined by titration) and whiteness [66]. The suitability of a zeolite as a catalyst is determined by testing it in the reaction concerned [67].

5. Natural Zeolites

5.1. Occurrence, Mining, and Processing

Zeolite minerals occur in many rocks on the earth's surface and on the sea bed. They were first discovered in crevices and cavities in basalt flows, where they were formed by the hydrothermal decomposition of the magmatic rock. It was only in the mid-1900s that zeolites were found to occur in enormous quantities as sedimentary rock. These zeolitic tuffs were formed during various geological periods, mainly by the action of water on volcanic ash. Zeolites were also formed from poorly crystalline clays, biogenic silica, feldspars, and feldspathoids. Moreover, the less stable types of zeolite formed during the initial crystallization changed into more stable types, e.g., phillipsite or clinoptilolite in analcime.

The formation of zeolites was influenced by many specific local conditions that had to be favorable: temperature (usually in the range 25 – 50 °C), pressure, pH, type and concentration of salts, and availability and reactivity of silicon and aluminum sources. Suitable conditions arise in various geological or hydrological systems, which can be open or closed [17–19], [68]. Thus, zeolites occur in deposits formed from lakes with high salt or alkali contents in arid and semiarid regions. Here, dissolved sodium carbonate/sodium hydrogencarbonate can cause the pH to reach ca. 9.5. Zeolites have also been formed in alkaline soils on the earth's surface and in ocean sediments as a result of the action of percolating water, hydrothermal activity, and burial diagenesis. Depending on the system, a characteristic series of deposits varying in composition, thickness, and extent is formed. The zeolite-containing layers can be several hundred meters in depth, and can extend over areas of hundreds of square kilometers.

Of the 40 zeolitic minerals discovered in nature, clinoptilolite [*12173-10-3*] and the structurally related heulandite, mordenite [*12173-98-7*], phillipsite, chabazite [*12251-32-0*], erionite, analcime, and laumontite occur in the largest quantities. However, only the first six of these materials are of commercial importance at present. Information on the structures and composition ranges of these zeolites is given in Table 2.

Natural zeolites are mined in the western United States, Cuba, Japan, Slovakia, Hungary, Serbia, Bulgaria, Italy, Germany, and the CIS states, mainly by the open-cast method. Depending on the deposit, zeolites can contain other minerals as impurities, e.g., montmorillonite, cristobalite, quartz, biotite, and potassium feldspar. Mixtures of different zeolites often occur. Zeolitic minerals are processed by crushing, grinding, and classification. If they are to be used as fillers in the paper industry, wet chemical bleaching is also necessary [69].

5.2. Commercial Applications and Importance

The consumption of clinoptilolite, which has a wide range of uses, far exceeds that of any other natural zeolite [19], [70], [71]. For example, it is used in agriculture as an animal feed additive and for soil improvement (mainly in Japan). The latter application depends on two effects: it increases the pH of the soil, and, being an ion exchange material, it stores ammonium and potassium ions, thereby prolonging the activity of fertilizers. The ability to take up ammonium

Table 2. Structures and compositions of some important natural zeolites

Zeolite	Structure type [28] (dimensionality of pore system)	Window	SiO_2/ $[Al_2O_3 + Fe_2O_3]$	Most common cations	Typical composition of unit cell
Clinoptilolite	HEU (3)	8-/10-/8-ring	8.0 – 10.2	K > Na	$(Na,K)_6[(AlO_2)_6(SiO_2)_{30}] \cdot 24\,H_2O$
Heulandite	HEU (3)	8-/10-/8-ring	5.8 – 8.0	Ca, Na	$Ca_4[(AlO_2)_8(SiO_2)_{28}] \cdot 24\,H_2O$
Mordenite	MOR (2)	12-/8-ring	8.6 – 10.6	Na > K	$Na_8[(AlO_2)_8(SiO_2)_{40}] \cdot 24\,H_2O$
Phillipsite	PHI (3)	8-/8-/8-ring	2.6 – 6.8	K, Na, Ca	$(Ca,K_2,Na_2)_3[(AlO_2)_6(SiO_2)_{10}] \cdot 12\,H_2O$
Chabazite	CHA (3)	8-ring	3.4 – 7.6	Ca, Na, K	$Ca_6[(AlO_2)_{12}(SiO_2)_{24}] \cdot 40\,H_2O$
Erionite	ERI (3)	8-ring	6.0 – 7.0	K, Mg, Ca, Na	$(Ca,Mg,Na_2,K_2)_{4.5}[(AlO_2)_9(SiO_2)_{27}] \cdot 27\,H_2O$

ions is also the reason for the use of clinoptilolite and phillipsite in fish farming and cat-litter production.

If high purity, constant composition, and high capacity are not essential, natural zeolites can compete with their synthetic counterparts in a number of industrial applications. Examples include the drying of gases or liquids with clinoptilolite tuffs, and the selective adsorption of CO_2 and SO_2 by clinoptilolite or mordenite. In these applications, the stability of these zeolites towards acids is advantageous. Clinoptilolite is also sometimes used for decontaminating radioactive wastewater because of its selectivity for ^{90}Sr and ^{137}Cs ions [19], [70–72]. Natural zeolites are also used as fillers in paper and plastics, and as constituents of pozzalanic cements and lightweight building blocks.

The quantities of natural zeolites mined worldwide and the total consumption can be estimated only approximately. A figure of ca. 300 000 t/a is realistic. Japan consumes the major proportion of this, i.e., currently 130 000 t/a, while consumption in western Europe and the United States is comparatively low, namely 25 000 t/a and 15 000 t/a, respectively.

6. Production of Synthetic Zeolites

The methods used in the industrial production of solids are always influenced by the specific requirements associated with the eventual use of such materials. As this also applies to zeolites, it is normally not enough to synthesize a single-phase microcrystalline zeolite powder of definite chemical composition. Other properties are also important, such as crystallite size and morphology. Moreover, for many applications, the as-synthesized zeolite must be chemically modified by processes such as ion exchange or dealumination. Finally, for most applications in adsorption and catalysis, shaped products are required. These usually contain binders, which can impart additional properties to the product. The manufacturing of a zeolite product therefore includes much more than the actual zeolite synthesis, and consists of a well-defined series of individual steps aimed at providing a specific spectrum of tailored properties.

6.1. Zeolite Synthesis: Routes and Raw Materials

Synthetic zeolites are produced under conditions similar to those that led to formation of natural zeolites, i.e., by the mineralizing effects of water and OH^- ions on reactive sources of silica and alumina in the presence of cations. However, the time required for formation of zeolites is shortened by using considerably higher temperatures and pH values. For example, high-silica zeolites are synthesized under hydrothermal conditions at temperatures exceeding 100 °C and at high pressure [73], while low-silica zeolites are usually crystallized at 70 – 100 °C [5], [11]. The pH is usually in the range 10 – 14. Reactive precursors for zeolite formation are mainly silica – alumina hydrogels produced in situ, although clay minerals are also used.

The composition of the reaction mixture is of vital significance for the synthesis process. This is usually expressed as the oxide ratio $a\,M_{2/n}O : Al_2O_3 : b\,SiO_2 : c\,R : d\,H_2O$, in

which M represents cations with charge n, and R is an organic template compound. Such compositions describe a synthesis field and lead to a specific zeolite type. Synthesis fields are often represented in a triangular diagram of the system $M_{2/n}O - Al_2O_3 - SiO_2$ (Fig. 4).

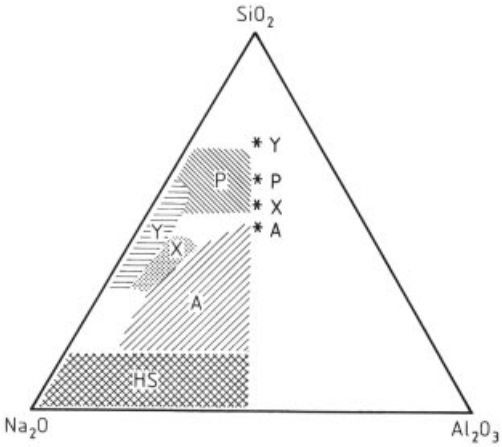

Figure 4. Synthesis fields
$a\,Na_2O : b\,Al_2O_3 : c\,SiO_2 : (90-100)\,H_2O$ for various zeolites (HS = hydroxysodalite) at a crystallization temperature of 100 °C
Points marked with asterisks indicate typical compositions of the products.

6.1.1. Synthesis from Gels

In the synthesis of zeolites from hydrogels, the composition of the reaction mixture is not the only important factor. The selection of the raw materials, the way in which the reactants are mixed to precipitate the gel, and its aging influence the course and rate of the subsequent zeolite crystallization and properties of the product, such as crystal morphology.

Various silica and alumina sources can be used as raw materials. In industrial production, the cost is an important consideration apart from the fundamental suitability of the materials.

The main silica sources used for the large-scale production of zeolites are sodium silicate solutions and precipitated amorphous silicas. Silica-containing waste products from large-scale industrial processes, e.g., blast furnace slags or fly ash, have the disadvantage that they are usually contaminated with heavy metals which to some extent enter the zeolites produced. For the synthesis of expensive specialty zeolites, pyrogenic silica, colloidal silica, and esters of silicic acid such as tetraethoxysilane are used. The most common alumina source is sodium aluminate.

A further reactant is also necessary that provides OH^- ions. Sodium hydroxide solution is used except for those few cases in which the presence of Na^+ ions prevents the formation of the desired zeolite type. Alternatively, the function of the OH^- ion as a mineralizer can be performed by the fluoride ion, as has been demonstrated in a number of zeolite syntheses [74].

The synthesis of high-silica specialty zeolites often has to be carried out in the presence of organic templates. These can be alkylammonium salts, alkylammonium hydroxides, amines, alcohols, or ethers [75], [76]. The function of these additives in zeolite syntheses is not yet fully understood. A purely template effect can be excluded in syntheses in which completely different agents lead to the same structure type [76]. Often the template molecules or ions simply seem to act as space fillers. The metal cations (usually Na^+) in the reaction mixtures can also be assumed to exhibit a structure-directing effect.

The formation of zeolites from the hydrogel that is formed by bringing together the above components is a very complex process, all details of which are not completely understood. However, the following concept is generally accepted [11], [77–80]: The gel partially dissolves during the aging process and especially when the temperature is raised to initiate zeolite crystallization. The liquid phase present with the gel becomes supersaturated, first with monomeric silicate, and then with oligomeric silicate species and aluminosilicate building blocks. The state of supersaturation ends with the formation of seeds of a metastable zeolitic phase. These then grow to form crystals.

Factors affecting the rate of zeolite formation include the reactivity of the gel, the pH, and the temperature. Increase in pH and temperature leads to increase in the rate of formation of zeolite crystals. However, the resulting zeolitic phases are metastable in their mother liquors, and the zeolite initially formed may undergo a transformation into an undesired thermodynamically more stable phase. A typical example is the formation of hydroxysodalite from NaA ze-

olite. The rate of synthesis of the zeolite can also be controlled by addition of seeds [11], [81].

6.1.2. Synthesis from Clay Minerals

Clays can also be suitable raw materials, especially for the production of the zeolites NaA, NaX, and NaY [5], [11]. The iron content of these natural materials should be as low as possible. The most important clays for zeolite production are kaolin and metakaolin, the latter being obtained by calcining kaolin at 500 – 700 °C. On further heating, spinel is formed at >980 °C and mullite at >1050 °C, with formation of silica-rich phases.

$$\underset{\text{Kaolin}}{2\,Al_2Si_2O_5(OH)_4} \rightarrow \underset{\text{Metakaolin}}{2\,Al_2Si_2O_7} + 4\,H_2O \qquad (2)$$

To produce zeolite NaA, sodium hydroxide is also required, and to produce silica-rich faujasites, further SiO_2 equivalents (e.g., sodium silicate).

The synthesis of zeolites from kaolin or metakaolin is of industrial interest because its allows production of binder-free extrudates and granules (see Section 6.3.4). These shaped materials consist entirely of active substance, which is an advantage in adsorption technology. Various methods for achieving this have been described. For example [82], [83], kaolin or metakaolin is first shaped, and SiO_2 and seed crystals are added, depending on the zeolite. The zeolite is then crystallized while in the preformed shape. Alternatively, metakaolin is used as the binder as the zeolite is formed, and the binder component is then converted into zeolite by hydrothermal treatment with sodium hydroxide solution [84], [85].

In-situ production of FCC catalysts from kaolin is of major importance. The process, developed by Engelhard, has many variations [86–88]. The products are NaY zeolites in a matrix that is enriched in Al_2O_3 compared with the raw material kaolin. In one variation, microgranules containing >40 % type Y zeolite are obtained [88]. For this, a slurry containing kaolin tempered at ca. 980 °C, water-containing kaolin, sodium silicate, and zeolite seed crystals is first spray dried, and the microgranules formed are calcined at ca. 750 °C, converting the watercontaining kaolin into metakaolin. The matrix-bonded NaY zeolite is then formed by treatment with sodium hydroxide and sodium silicate at 98 °C. This is followed by the process that generates the active catalyst, e.g., ion exchange or ultrastabilization (see Sections 6.3.1 and 6.3.2). Thus, in contrast to other processes, the zeolite component of the catalyst is not synthesized separately.

6.2. Examples of Industrial Zeolite Syntheses

The processes and potential problems in the industrial production of zeolites [5], [89–91] are illustrated here by the syntheses of the commerically important zeolites NaA, NaY, and Na-ZSM-5.

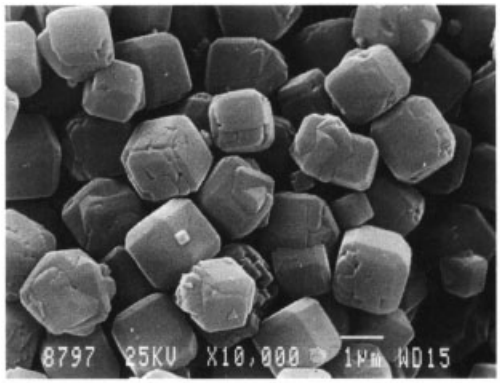

Figure 5. Electron micrograph showing crystals of detergent NaA zeolite

Zeolite NaA [*68989-22-0*] is mainly used directly as a detergent builder, but is also a starting material for the production of a wide range of adsorbents (see Sections 7.1 and 7.2). An NaA zeolite detergent builder should have the following properties: high ion-exchange capacity (>160 mg CaO/g dry weight at 20 °C) [66] with the ability to bind ions (especially Ca^{2+}) rapidly, good dispersibility, low tendency of suspensions to settle out, low abrasive properties, and good whiteness. This means that the zeolite must consist of a single phase, as the presence of impurities such as hydroxysodalite, which forms readily along with NaA, reduces the ability of the product to bind Ca^{2+} ions. Furthermore, the average particle size must be <4 μm and the zeolite

crystals must not have sharp corners or edges. A scanning electron micrograph of a materials of this type is shown in Figure 5.

In the Degussa [92], Henkel [93], [94], and PQ processes, such a zeolite is synthesized by reacting sodium silicate containing various amounts of Na_2O and SiO_2 with sodium aluminate solutions. Clays are less often used, as the presence of impurities, especially iron, can impart an undesirable color to the product. A special variation is the Mizusawa process in which an acid-treated clay is used [95].

In the production of zeolite NaA, the stoichiometry of the reaction mixture can in principle be varied over a relatively wide range. However, efficient production requires optimization of the composition from the point of view of product quality, yield, and reaction time. A possible stoichiometry is $3.4\,Na_2O : Al_2O_3 : 1.8\,SiO_2 : 90\,H_2O$ [93]. In this example, an excess of aluminum compared with the product stoichiometry ($SiO_2/Al_2O_3 = 2.0$) is used.

Figure 6 is a flow diagram of a process for the production of this zeolite [96]. Gel precipitation and aging are usually carried out at 50 – 70 °C. Gel precipitation is generally a batch process, but it can also be carried out continuously [94]. However, the subsequent crystallization at 80 – 90 °C is a batch operation in stirred vessels, often $>100\,m^3$ in volume. Under these conditions, the zeolite is present in crystalline form after ca. 1 h. The suspension formed contains 120 – 200 kg zeolite (dry basis)/m^3. The product is recovered and washed on a filter press, rotary filter, or band filter. The filter cake is slurried in water and, after the addition of stabilizers to prevent rapid sedimentation, is usually supplied to the detergent manufacturer in this form. The products are obtained in powder form by spray drying the zeolite suspension or by directly drying the zeolite filter cake.

Both the mother liquor from zeolite crystallization and the wash water are alkaline as they contain residual sodium hydroxide. If an excess of Al_2O_3 is used, sodium aluminate is also present. In this case, the filtrate is concentrated by evaporation and recycled.

The particle size, a critical parameter when the product is used as a detergent builder, is influenced by the Na_2O/H_2O ratio in the reaction mixture. High basicity favors the formation of small particles [97]. The sequence of addition of the various components and the rate of addition during gel preparation also exert an influence. Other factors include temperature and method of stirring, especially during gel formation, and use of seed crystals [89].

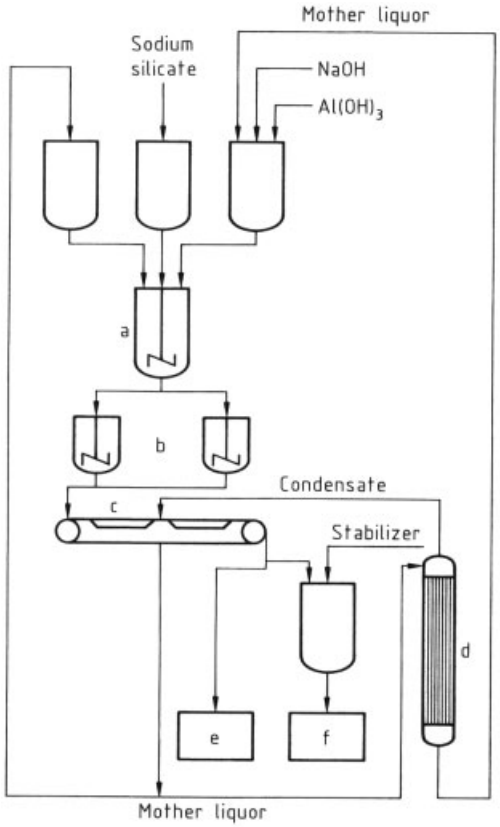

Figure 6. Process for the preparation of NaA zeolite [96] a) Silica – alumina hydrogel precipitation; b) Zeolite crystallization; c) Filtration; d) Mother liquor evaporation; e) Zeolite spray drying; f) Stabilized suspension of NaA zeolite

Zeolite NaY is the basis for the zeolite components of FCC and hydrocracking catalysts (see Section 7.2). This raw material must have high phase purity and good crystallinity. The SiO_2/Al_2O_3 ratio should be as high as possible, as type Y zeolites modified for catalytic purposes must have good thermal and hydrothermal stability.

Commercially available NaY zeolites have an SiO_2/Al_2O_3 ratio of 5.0 – 5.6. In principle, materials with higher SiO_2 content can be produced by direct synthesis, but the

necessary reduction of the Na_2O/SiO_2 and Na_2O/H_2O ratios in the reaction mixture lead to a drastic increase in crystallization time, so that such syntheses are not economic [98], [99]. Reaction mixtures with composition (2.5 – 3.5) Na_2O : Al_2O_3 : (8.0 – 10.0) SiO_2 : (120 – 180) H_2O are usually used for industrial production of NaY zeolite. The various synthetic and mineral raw materials are described in Sections 6.1.1 and 6.1.2.

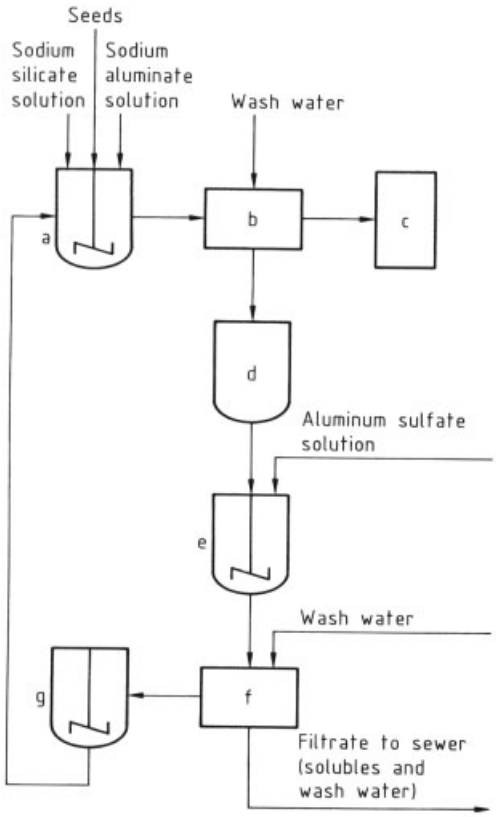

Figure 7. Process for the production of NaY zeolite [103] a) Zeolite synthesis (gel precipitation, aging, crystallization); b) Zeolite filtration; c) Zeolite storage; d) Mother liquor and wash water storage; e) Silica – alumina hydrogel formation; f) Silica – alumina hydrogel filtration and washing; g) Silica – alumina hydrogel slurry storage

In early syntheses of NaY zeolite, the gel initially precipitated had to be aged for one or more days at room temperature. Today, seeding mixtures are generally added instead, often in colloidal form [100], [101]. They have compositions in the range (10 – 16) Na_2O : Al_2O_3 : (10 – 16) SiO_2 : (180 – 320) H_2O and strongly influence the rate of crystallization.

The temperature during crystallization is 85 – 100 °C. Depending on this and on the composition and nature of the gel, a well-crystallized NaY zeolite can be obtained in 12 – 24 h. In some cases, vigorous stirring can lead to the formation of gmelinite and phillipsite as foreign phases. Ifthe heating time is too long, this can cause transformation into more thermodynamically stable zeolites of the P or cancrinite type [102].

Figure 7 is a schematic flow diagram of a process for the production of NaY zeolite [103]. Characteristic of this process is the work-up of the mother liquor, which in the case of the synthesis of NaY zeolite is a dilute solution of sodium silicate.

ZSM-5 [*79982-98-2*]. As the SiO_2/Al_2O_3 ratio of ZSM-5 can vary over a wide range (see Table 1), ZSM-5, first synthesized by Mobil [104], is not a single material, but a family of zeolites of the MFI structure type [28]. The synthesis of these compounds differs from that of NaA and NaY zeolites mainly in that it is carried out at 120 – 180 °C in pressure vessels. ZSM-5 can be crystallized with or without template compounds [40], [73], [104–106]. Tetrapropylammonium salts are mainly used as template, but hexamethylenediamine and a series of other compounds are also suitable [75], [76]. However, the ranges of SiO_2/Al_2O_3 ratios achievable in the product by varying the initial stoichiometry by the two methods are not identical. Thus, without a structure-directing agent, SiO_2/Al_2O_3 ratios between ca. 25 and 100 can be obtained, while with template compounds a significantly wider range can be covered, i.e., from ca. 30 to the aluminum-free form (silicalite-1). Moreover, depending on the raw materials used and the exact synthesis conditions, the crystal morphology and distribution of the aluminum atoms over the crystal cross section can vary greatly [40], [73]. These parameters are of major importance in the use of ZSM-5 zeolites in catalysts.

The composition 17.1 $(TPA)_2O$: Na_2O : Al_2O_3 : 27.7 SiO_2 : 453 H_2O is a typical example of a synthesis mixture for which tetrapropylammonium bromide (TPABr) is used as the template. After 5.5 d at 125 °C, this gives ZSM-5

with a SiO_2/Al_2O_3 ratio of ca. 45 [104]. In contrast, the synthesis mixture 11.2 $Na_2O : Al_2O_3 : 70\,SiO_2 : 3213\,H_2O$, without a template and using colloidal silica as the source of silica, gives a zeolite with an SiO_2/Al_2O_3 ratio of 44 after 24 h at 190 °C [105].

The crystal suspension obtained is worked up as described above for NaA and NaY zeolite. If an organic template compound is used, it must be calcined from the ZSM-5 pores by heating to ca. 650 °C in air.

6.3. Modification and Formulation

The hydrothermal synthesis step described in Sections 6.1 and 6.2 establishes the structure type of the zeolites and the size and shape of their crystals. Only detergent zeolites are in their final form at this stage. For use as a catalyst or adsorbent, their properties must be fine tuned. This is mainly performed by ion exchange and dealumination. In addition the properties of zeolites are modified by forming them into shapes appropriate for the flow conditions under which they will be used.

6.3.1. Ion Exchange

In the production of zeolitic adsorbents, the most important techniques for increasing or decreasing the apparent pore diameter ("pore-size engineering") are those involving ion exchange [5]. An important example is the exchange of sodium ions in zeolite NaA with calcium or potassium ions. After partial replacement of ca. 30 % of the Na^+ ions by Ca^{2+}, a sudden "dilation" of the pore openings from ca. 0.38 nm (4A type [*70955-01-0*]) to 0.43 nm (5A type [*69912-79-4*], [*68989-20-8*]) is observed. Conversely, the replacement of ca. 25 % of the Na^+ ions by K^+ causes "narrowing" of the windows to ca. 0.30 nm (3A type [*68989-21-9*]).

Ion exchange is also important in the synthesis of zeolite catalysts. The aim is usually to produce Brønsted acid sites, as zeolites act as solid acids in catalysis (see Section 7.3). For certain intermediate-silica and high-silica zeolites (e.g., mordenite), the necessary ion exchange can be achieved by treatment with mineral acids. However, this method also involves the risk of dealuminating the zeolite framework (see Section 6.3.2), and many high-aluminum species such as NaY zeolite lose their crystal structure, at least partially, at higher acid concentrations [47]. Therefore, an indirect route via an ion exchange with ammonium salt solutions must be followed. This produces the "ammonium form", which is calcined at ca. 400 °C to liberate ammonia and give the hydrogen form. In the case of zeolite Y, an acid form is also obtained by incorporating polyvalent cations, especially of rare earth metals, by ion exchange (see Section 7.3).

In the production of bifunctional zeolite catalysts that combine acidity with hydrogen-transfer capability, the inclusion of metals such as palladium and platinum is important. The first step is ion exchange with $[Pd(NH_3)_4]^{2+}$ or $[Pt(NH_3)_4]^{2+}$. The chemistry of Pd^{2+} and Pt^{2+} in zeolitic pores has been thoroughly investigated, especially in the case of zeolite Y. Calcination in the presence of oxygen to destroy the ligands followed by a heat treatment in a hydrogen atmosphere leads to formation of metal clusters of various sizes, locations, and states of dispersion, depending on the exact conditions [107], [108].

In the most widely used ion-exchange process, zeolites are treated with aqueous solutions that contain the cations to be introduced. Solid-state ion exchange in which the zeolites, usually in the ammonium or hydrogen form, are heated with crystalline salts, is as yet rarely used in industry [109], [110]. Ion exchange also occurs on impregnating zeolites with solutions of metal salts.

The equilibria and kinetics of ion exchange with zeolites in aqueous suspension are complex subjects in which several factors interact [5], [72], [111].

The theoretical maximum ion-exchange capacity of a zeolite is determined by its SiO_2/Al_2O_3 ratio. The true or maximum attainable ion-exchange capacity to be expected in a given case is influenced by ion-sieve and volume effects. For example, the pore apertures of zeolite X (crystallographic diameter 0.74 nm) do not permit penetration of $[N(C_2H_5)_4]^+$ ions (diameter 0.80 nm) into the cavity system. Metal cations with large, strongly bonded hydration shells such as Li^+ and Mg^{2+} cause similar problems. Nevertheless, slow ion exchange is possible in these cases, for which the cation has to

"cast off" the coordinated water molecules, at least partially. However, even in the case of wide pore apertures cations that are large in the unhydrated state still give low rates of ion exchange. The space requirements of these cations in the cavities and the need to balance the charge of the framework are then in opposition to each other.

In addition, a sieve effect within the zeolite framework can occur if the cations being exchanged are located in small and inaccessible cages. An important example of this is provided by the faujasites, in which the Na^+ ions located in the sodalite cages cannot be replaced by larger cations under mild conditions. Hence, the maximum achievable degree of exchange with rare earth cations for zeolite NaY at 25 °C is only ca. 70 % [112]. This exchange is of importance in the production of FCC catalysts (Section 7.3). It can only be carried out to completion at 180 °C under pressure. However, a different method is usually used: After achieving the degree of exchange attainable under mild conditions, the zeolite is calcined at ca. 350 °C. This causes the Na^+ ions to migrate out of the sodalite cages into the supercages where they are replaceable, and a further treatment of the zeolites with rare earth salt solutions enables exchange to be completed. The same method can be used for ammonium ion exchange [113].

In general, the ion exchange reaction for two cations A and B with charges n_A and n_B is represented by Equation (3). The symbols (z) and (s) indicate whether the cation is in the zeolite phase or in aqueous solution. The position of the equilibrium is a measure of the ion-exchange selectivity. The selectivity for given cations is determined by the electric fields of the coordinating anionic locations in the zeolite framework. Thus, for each zeolite type there is a specific selectivity series with the cations arranged in order of increasing exchangeability [5], [114].

$$n_A B^{n_{B^+}}\,(z) + n_B A^{n_{A^+}}\,(s) \rightleftharpoons n_A B^{n_{B^+}}\,(s) + n_B A^{n_{A^+}}\,(z) \qquad (3)$$

The selectivity of an ion-exchange process involving a given cation pair is illustrated by isotherms in which the equivalent fraction of the entering cation A in solution A_s is plotted against the equivalent fraction for the same cation in the zeolitic phase A_z [5], [111], [115]. The isotherms describe the situation after equilibrium has been reached. Figure 8 shows examples of isotherms for several cations in the ion-exchange reaction on the zeolite NaA (B = Na^+). For A = Ag^+, Ca^{2+} there is selectivity in favor of the added ions, whereas if A = Li^+ the equilibrium of Equation (3) always lies on the left-hand side. For A = K^+, the selectivity varies with the extent of ion exchange.

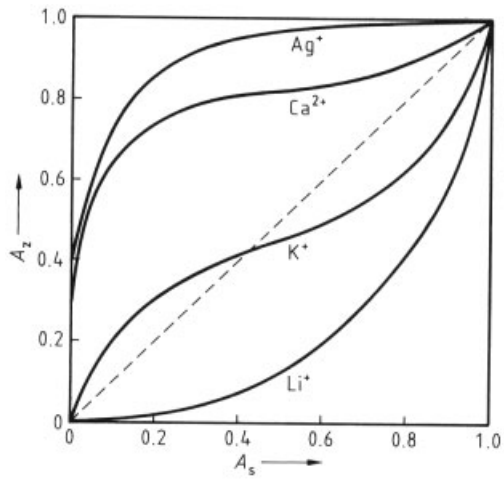

Figure 8. Ion-exchange isotherms for Li^+, K^+, Ag^+, and Ca^+ on zeolite NaA at 25 °C
Total concentration in aqueous phase: 0.1 N

Another important parameter is the pH of the exchange solution. At low pH values, H_3O^+ ions compete with the metal cations for the exchange locations. The pH must also be considered when ion exchange takes place with highly charged transition metal cations such as Fe^{3+} or Cu^{2+}, as an acidic medium is necessary here to prevent precipitation of metal hydroxides. However, slurries of low-silica zeolites such as A or X are alkaline, and these materials are unstable at low pH (see Section 3.2). These problems can often be overcome by the use of complex cations (e.g., ammine complexes).

In industry, there are many methods of carrying out an ion-exchange reaction in aqueous solution. The simplest is to suspend the zeolite in the solution under the appropriate conditions for ion exchange (amount and concentration of the salt, temperature, and pH). This is followed by filtration and washing of the filter cake. A more elegant technique is to combine these steps in a continuous process, whereby ion exchange and

washing are carried out on a band filter, for example.

A process of this kind often needs to be repeated to obtain the desired extent of ion exchange. In these cases, a process with countercurrent flow of the components is more efficient and economic [116]. The solid and the salt solution are fed in opposite directions, so that fresh unreacted zeolite comes into contact with a solution that has already been used to treat partly ion exchanged material and is therefore partially depleted in the cations to be introduced.

6.3.2. Dealumination

Dealumination increases the SiO_2/Al_2O_3 ratio of zeolites. The term is usually understood to mean the removal of aluminum from the zeolite framework, even where the overall composition of the material is unaltered or only slightly altered because the aluminum removed from the framework remains in the channels and cavities.

Dealumination is mainly carried out on zeolites with an intermediate SiO_2/Al_2O_3 ratio (see Section 3.2) that cannot be directly produced in a high-silica form and have sufficient stability to survive the sometimes drastic treatment without appreciable loss of crystallinity. The prototype for the development of dealumination methods was zeolite Y, as both its hydrothermal stability can be considerably increased and its catalytic activity for use in FCC catalysts can be modified by dealumination [45], [86], [117]. Mordenite is also often modified by dealumination [118], [119]. Dealumination is rarely used for zeolites such as ZSM-5, which can be obtained directly as a high-silica molecular sieve by hydrothermal synthesis. The catalytic properties of high-silica materials synthesized by dealumination are expected to differ from those of zeolites with the same SiO_2/Al_2O_3 ratio produced directly, as the dealumination of the structure can lead to the production of crystal defects, creation of a secondary pore system, and formation of nonframework aluminum.

A range of chemical methods [45], [117] are available for dealumination of zeolites, some of which are practiced on an industrial scale.

The simplest method is to extract aluminum with a mineral acid. This is often carried out with mordenite, which is usually produced in the Na form with an SiO_2/Al_2O_3 ratio of 10 – 12 [118], [119]. Depending on the reaction conditions, a high-silica material with a low cation content is obtained. In the case of zeolite Y, obtainable with an SiO_2/Al_2O_3 ratio of up to 6.0 by direct synthesis, this method, which causes defects, leads to considerable destruction of the framework structure. For this zeolite, the process is only suitable when used in conjunction with preliminary dealumination by another method in which not only aluminum is dissolved out of the framework but also some isomorphic replacement of aluminum by silicon takes place.

Of these methods, the steaming of Y zeolites in the NH_4^+ form (residual content of Na_2O <3 – 4 %) is of the greatest importance. Industrially, this is carried out at 600 – 800 °C in rotary kilns [46], [86], [87]. This causes aluminum atoms to migrate from the framework into the cavities, forming nonframework species. In principle, this process is a high-temperature hydrogenolysis of Si – O – Al bonds. The process causes limited damage to the structure, which can be detected by the formation of mesopores. The resulting holes are partially filled by silicon atoms from amorphous SiO_2 originating from destroyed crystal domains. The resulting zeolites have a framework structure with an increased SiO_2/Al_2O_3 ratio and very good thermal and hydrothermal stability. Today, they are of major importance as components of FCC catalysts, and are known as ultrastable Y zeolites (USY zeolites, see Section 7.3). Additional stabilization is obtained by subsequent exchange of the residual Na^+ ions with NH_4^+ ions. Nonframework aluminum species can be removed by washing with solutions of acids or alkalis [45].

In a Degussa process [120], [121], NaY zeolite is dealuminated by complete isomorphic replacement of aluminum atoms by silicon atoms using gaseous silicon tetrachloride at 250 – 500 °C. The reaction product $NaAlCl_4$ is removed from the cavities by washing with water. Depending on the temperature used, low-defect Y zeolites with various degrees of dealumination can be obtained. In particular this method offers a direct route to materials with SiO_2/Al_2O_3 ratios of 200 or more. These are extremely hydrophobic and are used for waste-air and exhaust gas purification (see Section 7.2). In contrast to zeolite Y, it is very difficult to

dealuminate mordenite with silicon tetrachloride [118].

Another isomorphic replacement technique is to react ammonium-exchanged Y zeolites with an aqueous solution of ammonium hexafluorosilicate [122], [123], producing thermally and hydrothermally stable Y zeolites with SiO_2/Al_2O_3 ratios of up to ca. 20. These are mainly used in FCC catalysts. Equations (4) and (5) represent the two methods, idealized for a Y zeolite with a SiO_2/Al_2O_3 ratio of 5.0.

$$Na_{55}[(AlO_2)_{55}(SiO_2)_{137}] + y\,SiCl_4 \rightarrow Na_{55-y}[(AlO_2)_{55-y}(SiO_2)_{137+y}] + y\,NaAlCl_4 \quad (4)$$

$$(NH_4,Na)_{55}[(AlO_2)_{55}(SiO_2)_{137}] + y\,(NH_4)_2SiF_6 \rightarrow (NH_4,Na)_{55-y}[(AlO_2)_{55-y}(SiO_2)_{137+y}] + y\,(NH_4)_3AlF_6) \quad (5)$$

Other dealumination processes [45], [117], involving, e.g., the use of F_2, $CrCl_3$, or H_4EDTA, are not of industrial importance.

6.3.3. Other Methods of Zeolite Modification

Other less industrially important methods are known in addition to those described in Sections 6.3.1 and 6.3.2. The silanization of zeolites with mono- and disilane and the chemisorption of diborane have been thoroughly investigated. These lead to a narrowing of the pore system and to changes in the selectivity of the interior surfaces [124]. Reaction with tetraalkoxysilanes followed by calcination in the presence of oxygen only reduces the size of the pore apertures [125].

In an alternative method to ion exchange, transition metals can be introduced into the pore systems of zeolites by means of organometallic complexes [126]. In particular, socalled "ship-in-a-bottle" phthalocyanine complexes can be synthesized directly in the cavities [127].

6.3.4. Shaping

Various industrial operations for shaping other powdered materials [128–130] are also used for shaping zeolites for use as catalysts and adsorbents. These include extrusion, and granulation by the granulating pan, fluidized-bed, and spray methods. The choice of method depends on the size and geometry most suitable for a particular application. Shaping is usually followed by drying and calcination.

To give the shaped bodies adequate mechanical stability, binders must be added to the zeolite powders. These can consist of clays such as kaolins and bentonites or synthetic materials such as amorphous silica, alumina, and silica – alumina gels. Substances that act as sources of silica, e.g., silica sols or, more rarely, esters of silicic acid, are also used in the shaping process. Binder-free shaped zeolite bodies in which the binders are transformed into zeolitic phases are a special case (see Section 6.1.2).

The selection of the type and quantity of binder must take account of the fact that the matrix in which the zeolite crystals are embedded may perform a large number of other functions in addition to its primary function, and that these may have an important effect on the property profile of the zeolitic catalysts and adsorbents. Thus, the matrix may "dilute" the activity of the zeolite, but may actually contribute to the activity of the catalyst. It can also act as a sink for sodium or vanadium ions, increasing thermal and hydrothermal stability in the case of FCC catalysts [86].

When NaA zeolite powder is granulated with water glass, the latter adheres to the pores to some extent [91], [131], leading to a reduction in the apparent pore size to ca. 0.30 nm. This effect is essential if the granulate is to be used as a drying agent in insulating glass (double glazing) (see Section 7.2).

The macroporosity of the shaped body can be influenced by the addition of certain organic compounds such as melamine. These are burnt off in a subsequent calcination stage [130].

6.4. Environmental Aspects

In the industrial-scale production of zeolites such as NaA and NaY, substances present in the mother liquors are recycled to the synthesis process (production-integrated environmental protection). The use of template compounds in the production of specially zeolites presents greater problems. The tetraalkylammonium salts and organic amines which are often used are very difficult to destroy by oxidation in the mother liquor.

There are therefore ecological as well as economic reasons to minimize the amounts of these compounds used in the synthesis.

The same is true of the metal salts used in ion-exchange processes. Here, judicious operation can be of benefit in a multistage process [116]. Dilute solutions of ammonium salts, e.g., those produced during the conversion of zeolites into the catalytically active hydrogen form, can be treated in various ways [132].

7. Applications of Synthetic Zeolites

7.1. Ion Exchange: Zeolites as Detergent Builders

The most important use of synthetic zeolites based on their ion-exchange properties is as builders in detergents. This is the largest single application of zeolites (see Chap. 8). The removal of undesired ions from wastewater is mainly carried out with the comparatively cheap natural zeolites (see Section 5.2). The treatment of wastewater from the electroplating industry with powdered NaA zeolite is an exception [133].

Pentasodium tripolyphosphate (STPP) was long regarded as the best builder for detergents on account of its excellent sequestering powers for calcium and magnesium ions and its outstanding ability to disperse soil particles. However, an intensive search for alternatives began in the early 1970s. This was driven by the discovery that a high proportion of the phosphate pollution in stationary or slow-moving natural surface waters in industrialized countries was due to the use of phosphates in detergents. In 1975, this proportion reached an average of 40 % in the Federal Republic of Germany. High phosphate levels in lakes and rivers constitute an ecological problem, leading to overfertilization, eutrophication, and excessive growth of algae, whose eventual death leads to oxygen deficiency, endangering aquatic organisms. For this reason, restrictions on the phosphate content of detergents and cleaning agents were introduced in many countries [66]. The search for a replacement lead to the development of zeolite NaA as a water-insoluble builder. It was important that the zeolite should have a small average particle size of ca. 3.5 µm and a narrow particle size distribution, to prevent incrustations in the textile fiber. More recently, the zeolite NaP has also been used as a builder component in detergents [134], [135].

The main function of zeolite NaA used as a detergent builder is to bind ions that cause hardness [93], [96], [136–138]. The ion exchange capacity with respect to Ca^{2+} ions is ca. 160 mg CaO/g (dry basis) at 20 °C, which is comparable with that of STPP, and ca. 190 mg CaO/g (dry basis) at 90 °C, exceeding that of STPP [66]. However, the rate of reaction with Ca^{2+} ions is lower than that of the soluble builders. The corresponding relative rates are even lower in the case of Mg^{2+} ions on account of the ion-sieve effect (Fig. 9). This preference of zeolite NaA for Ca^{2+} ions is not a problem, as tap water contains considerably lower concentrations of Mg^{2+} ions than of Ca^{2+} ions.

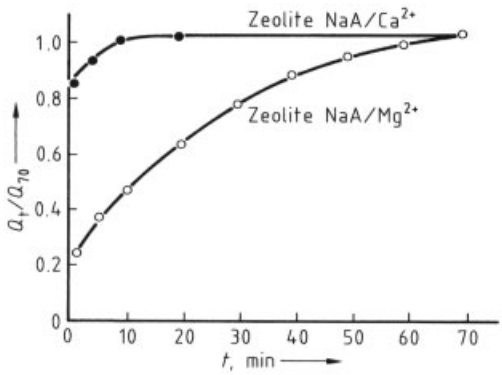

Figure 9. Ion exchange by zeolite NaA as a function of time [137]
$[Ca^{2+}]$, $[Mg^{2+}] = 0\,5.36 \times 10^{-3}$ mol/L; NaA zeolite: 1 g/L; temperature: 25 °C; Q_t: amount of Ca^{2+}/Mg^{2+} exchanged after time t; Q_{70}: amount of Ca^{2+}/Mg^{2+} exchanged after 70 min

Especially at high temperatures, the comparatively low rate of ion exchange in washing liquor, an alkaline medium, causes calcium carbonate to be precipitated before Ca^{2+} ions are taken up by the zeolite. For this reason, substoichiometric amounts of water-soluble cobuilders are added to the zeolite. In phosphate-free formulations, these generally consist of polycarboxylates in combination with sodium carbonate, and alternatively nitrilotriacetic acid, phosphonates, and citric acid. The Ca^{2+} ions in the washing liquor are initially bound to the

cobuilder, which transports them to the surface of the zeolite crystals where they are exchanged with Na^+ ions. Thus, the cobuilder acts as a carrier, and the zeolite as a means of storing Ca^{2+} ions, without which the polycarboxylates would be precipitated as calcium salts [137]. Furthermore, the cobuilder adheres to crystallization seeds of insoluble calcium salts, thereby preventing their growth (threshold effect). In Europe, phosphatefree heavy duty detergents contain 20 – 35 % zeolite NaA, 3 – 8 % polycarboxylates, and 5 – 20 % sodium carbonate as builder system [139]. In soft water areas such as Japan, where washing is carried out at lower temperatures (15 – 30 °C) than in Europe (30 – 95 °C), a cobuilder is not required.

The use of zeolite NaA also has positive side effects. For example, the zeolite adsorbs heavy metal ions present in trace amounts (e.g., Cu^{2+}) which impair the bleaching effect of perborate and percarbonate. Also, laundry discoloration due to adsorption of dissolved dyes, and graying due to retention of soil particles are minimized by heterocoagulation.

Extensive investigations have been carried out into the behavior of zeolite NaA in public wastewater systems and on the effects of its use on waterborne organisms and plant growth [93], [140]. No indications of harmful effects were found. A more recent study has shown that zeolite NaA in water undergoing treatment in sewage works retains the Ca^{2+} and phosphate ions present, converting them into an insoluble, relatively stable calcium aluminum silicate phosphate, and can thereby contribute to phosphate immobilization [141].

7.2. Adsorption

Along with silica gels, alumina, activated carbons, carbon molecular sieves, and organic resins zeolites are important industrial adsorbents with a wide range of uses [142–144]. Their diversity is due to the fact that zeolites not only have a high adsorption capacity but also show great selectivity, enabling substances to be separated or purified. Three effects contribute to the selectivity of a zeolite, and these can be influenced by the synthesis methods used to tailor it to requirements:

1) Zeolites are characterized by regular pores with definite aperture sizes, so that substances whose molecules are of different sizes can be separated (molecular-sieve effect, see Section 3.1). This selectivity can be influenced by post-synthetic operations, as explained in Section 6.3.1 using the example of ion exchange with zeolite NaA. The effect on the adsorption of various molecules of increasing the apparent pore size by transformation from 4A type to 5A type zeolite by calcium exchange is shown in Figure 10. Conversely, the pores can be narrowed to give 3A type zeolite by exchange with potassium salts, by steam treatment [5], or by appropriate choice of binder in the shaping process (Section 6.3.4).
2) The molecular-sieve effect described under (1) is a limiting case of kinetic selectivity, i.e., selection according to differences in the rates of diffusion of molecules into the zeolite pores.
3) The adsorption equilibrium is determined by electrostatic forces including dipole and quadrupole interactions, and also by dispersion forces. In high-alumina zeolites of types A and X (NaX is often denoted as 13X [*63231-69-6*] in adsorption technology) in which the negative framework charge is balanced by the appropriate number of cations, electrostatic interactions predominate. This favors adsorption by these zeolites of substances with a large dipole moment (e.g., H_2O and NH_3) or quadrupole moment (e.g., N_2). The interaction becomes stronger as the ratio of charge to radius of the cation concerned increases. In contrast, adsorption on high-silica zeolites is due only to van der Waals forces. The bonding affinity of adsorptive molecules is approximately proportional to their polarizability and therefore to their molecular mass. This is the reason for the hydrophobic surface of zeolites of this type (see Section 3.2). The interactions are especially favorable in this case if there is close contact between the pore walls and the adsorbed molecule.

Zeolites are used in adsorption technology both in gas and liquid phases. Processes can be classified according to the method of regenerating the loaded adsorbent and the method of

making the contact between the fluid and the adsorbent [142], [143]. In practice, pressure swing [145], [146] and temperature swing adsorption (PSA and TSA methods) are preferred. More rarely, desorption is carried out with a purge gas or displacement auxiliary at constant temperature and pressure (purge swing).

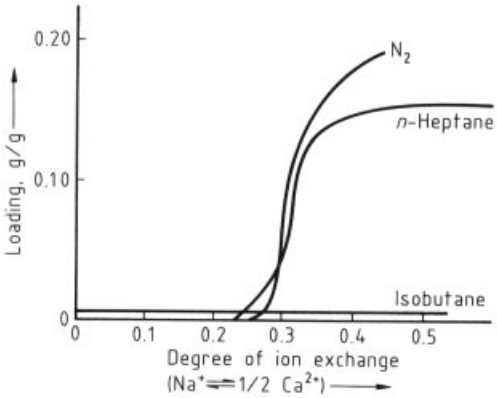

Figure 10. Effect of calcium ion exchange on zeolite NaA [5]
N_2 adsorption: 2.0 kPa, −196 °C; *n*-heptane adsorption; 6.0 kPa, 25 °C; isobutane adsorption: 53.3 kPa, 25 °C

Processes can also be divided into purification processes in which an undesired low-concentration component of a mixture of substances is adsorbed, and separation processes in which the components of the mixture are present in comparable amounts. The adsorption of water from gases and liquids is in fact a purification process, but, in view of its great economic importance, is discussed separately.

Static and Dynamic Adsorption of Water. Dehydrated ("activated") zeolites of types A and X are widely used in industry as drying agents in closed or open systems, i.e., for the static or dynamic adsorption of water. Their use is favored if small amounts of water have to be removed from gases or liquids. The adsorption isotherms in Figure 11 show that, unlike silica gel, zeolite NaA is able, for example, to reduce the dew point temperature of the gas to extremely low values. However, as water is very strongly adsorbed and is therefore difficult to desorb, zeolites cannot be considered for coarse drying in regenerative processes.

In drying operations, measures must often be taken to prevent larger molecules from being adsorbed and possibly undergoing catalytic reactions. This can be achieved using 3A zeolite whose pore diameter has been reduced to ca. 0.30 nm, e.g., in the drying of alkenes.

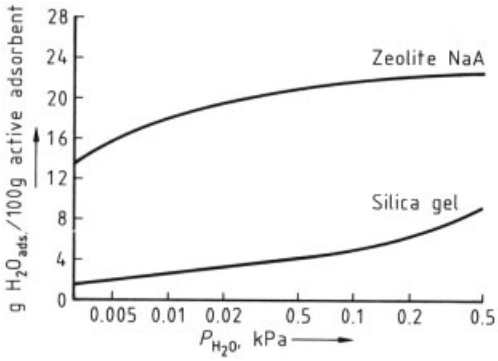

Figure 11. Adsorption isotherms for H_2O at 25 °C on zeolite NaA and silica gel at low partial pressures

A similar situation exists for the removal of atmospheric moisture from the panes of insulating glass [147], an important example of static adsorption. Here, the coadsorption of nitrogen and oxygen or insulating gases such as argon and SF_6 is undesired. Other examples of static adsorption of water by type A zeolites are the drying of coolant cycles and removal of moisture from solvent-free, noncellular polyurethane systems, in which reaction of the isocyanate component with water, forming CO_2 bubbles, must be avoided [148]. In this case, the zeolite powders remain in the cast resins, or sealing and coating compounds.

Dynamic water-adsorption processes include drying of air, natural gas, cracked gas, alkenes, and organic solvents using TSA and PSA methods with type A zeolites as the sorbents [5], [143], [145], [149], [150].

The large amount of heat liberated during the adsorption of water by high-aluminum zeolites could in future be used for zeolite-based heat storage and heat pumps. If, in another arrangement, a water reservoir were brought into contact with an activated low-silica zeolite via the gas phase, the evaporation, accelerated by the presence of the zeolite, would cause a large drop in temperature. This effect could be used for cooling [151].

Purification of Air and Gases. Zeolites of types A and X are well established in purifica-

Table 3. Commercial purification processes with synthetic zeolitic adsorbents (excluding drying)

Components removed	Crude gas	Zeolites	Refs.
CO_2	natural gas, air	4A, 5A, X	[144]
H_2S, organic sulfur compounds	natural gas, hydrogen	5A	[142]
SO_x, NO_x	waste air from H_2SO_4/HNO_3 plants	not published	[154]
Solvents, odoriferous substances	industrial waste air	dealuminated Y zeolites, silicalite-1	[152], [155]
Dioxins	waste air from incinerations	dealuminated Y zeolites	[156]
Hg	natural gas, cracked gas, hydrogen, naphtha	modified Y zeolites	[157]
	waste gas from incinerators		[153]

tion processes (Table 3). They are used to remove carbon dioxide from gas streams [144], and also to remove H_2S and organic sulfur compounds, mainly from natural gas. Problems can arise in H_2S adsorption if CO_2 is also present, as this reacts with H_2S on the zeolite to form COS [142]. As the activity of calcium-exchanged zeolite A is less than that of zeolite NaA in this reaction, 5A sieves are mainly used in this process.

The removal of solvents, dioxins, and elemental mercury from moist waste-air and exhaust-gas streams by hydrophobic, high-silica zeolites such as dealuminated Y zeolite is a relatively new field [152], [153]. Compared with other adsorbents, these materials have the advantage of high thermal capacity, nonflammability, low catalytic activity, and pronounced hydrophobic properties. Regeneration is possible even if the compounds adsorbed have high boiling points or can form polymers as the zeolite absorbent can be treated at high temperatures. Moreover, they are structurally stable in the presence of acidic components, such as SO_2 and HCl in incineration off-gases.

Separation of Substances. Zeolitic adsorbents of the 5A type are used in a number of important industrial processes for the separation of mixtures of substances (Table 4). They are alternatives to carbon molecular sieves and activated carbons for the recovery of hydrogen from reformer, refinery, and coke-oven gases in various PSA processes [146], [150], [158], [159]. Here, the separation effect is based on the fact that the gaseous components listed in Table 4 are adsorbed better than hydrogen.

The production of oxygen from air by vacuum pressure swing adsorption (VPSA) processes (→ Oxygen, Chap. 4.2.) represents a growing market for zeolites. Type 5A molecular sieves are mainly used, but the use of specially optimized lithium-containing X zeolites is increasing [146], [150], [160–162]. The separation of O_2 from N_2 is possible with these materials because the nitrogen molecule, which has a quadrupole moment, is more strongly bound than the oxygen molecule. The purity of the oxygen produced can exceed 95 %, but argon cannot be separated. VPSA processes are more economic than the cryogenic production of oxygen for flexible small to medium-sized plants [161].

The O_2/N_2 separation on 4A zeolite to produce nitrogen is based on the kinetically preferred adsorption of the smaller O_2 molecules. This process is hardly used on an industrial scale, but zeolites can be useful for oxygen removal in the production of ozone [146].

The separation of *n*-alkanes from iso-alkanes on 5A zeolites is of importance in the mineral oil industry. For example, PSA processes have been developed by UOP (Isosiv), Texaco (TSF), Leuna (Parex-Leuna), ELF (N-Iself), Exxon, and BP [142], [146]. In these processes, regeneration is sometimes performed by displacement media such as NH_3 or short chain *n*-alkanes. The Molex process of UOP is a liquid-phase process designed on the principle of simulated countercurrent chromatography (Sorbex technology) [163], [164]. All these processes are based on the fact that *n*-alkanes are adsorbed by 5A zeolites, while iso-alkanes are too bulky to enter the pore system (see Fig. 10) and therefore break through immediately.

Sorbex technology is also used for separating xylene isomers (→ Xylenes, Chap. 4.2.2.), alkenes from alkanes, and fructose from glucose (Parex, Olex, and Sarex processes of UOP) [163]. These separations, which take place on modified zeolites of the faujasite type, are due to the selectivity of cation – adsorpt interactions.

Table 4. Some important commercial separation processes with zeolitic adsorbents

Process	Mixture	Phase ∗	Technology	Zeolites	Refs.
H_2 production by removal of other gases (e.g., N_2, CH_4, CO, SO_2, H_2O, O_2, CH_4, C_2H_6)	reformer, refinery, and coke-oven gases	G	PSA	5A	[146], [150], [158], [159]
O_2 production	air	G	VPSA ∗∗	5A, modified X zeolites	[150], [160–162]
Ozone production	O_3/O_2	G	PSA	not published	[146]
Separation of *n*- from iso-alkanes	C_6-C_{10} distillate	G	VPSA ∗∗	5A	[142]
	$C_{10}-C_{18}$ kerosene	G	purge swing	5A	
	C_5-C_6, $C_{10}-C_{15}$, $C_{10}-C_{23}$ hydrocarbons	L	Sorbex (Molex)	5A	[163]
Separation of xylene isomers	C_8 aromatics	L	Sorbex (Parex)	Sr/BaX	[163]
Separation of alkenes from alkanes	C_8-C_{18} hydrocarbons	L	Sorbex (Olex)	Ca/SrX	[163]
Fructose/glucose	molasses, hydrolyzed starch	L	Sorbex (Sarex)	CaY	[163]

∗ G = gas phase, L = liquid phase.
∗∗ Vacuum pressure-swing adsorption.

As well as the industrial-scale processes mentioned, there is a growing list of potential separation processes based on zeolitic adsorbents, some of which are at the pilot plant stage.

7.3. Catalysis

Apart from the very alumina rich types such as A and X, which are unstable in acidic media, zeolites can be transformed into the hydrogen form [47]. In this form, they are solid acids which can show both Brønsted and Lewis acidity [165–168]. Brønsted acid sites are protons attached to framework oxygen atoms.

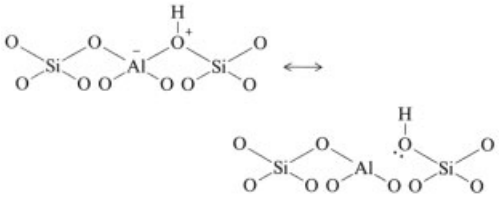

The number and strength of these sites depends on the SiO_2/Al_2O_3 ratio in the framework. Decreasing framework aluminum content is associated with a decrease in the number of acidic OH groups, but generally the strength of the remaining sites initially increases [168], [169]. Structurally related zeolite-like materials in which aluminum is isomorphously replaced by gallium or boron have weaker Brønsted acidity.

Methods used industrially to produce Brønsted centers include:

1) Careful ion exchange of intermediate-silica and high-silica zeolites (e.g., mordenite) with mineral acids (Eq. 6)
2) Removal of ammonia from the ammonium form at ca. 400 °C (Eq. 7) or the decomposition of tetraalkylammonium cations
3) Ion exchange with polyvalent cations, usually of rare earth (RE) metals, followed by a dehydration (Eq. 8)
4) Reduction of noble metal cations with hydrogen (Eq. 9)

$$Na^+[O{-}Z] + H^+ \rightarrow [{}^+H{}^-O{-}Z] + Na^+ \quad (6)$$

$$NH_4^+[{}^-O{-}Z] \rightarrow [{}^+H{}^-O{-}Z] + NH_3 \quad (7)$$

$$[RE(H_2O)_n]^{3+}[({}^-O{-})_3Z] \rightarrow [RE(OH)_2]^+[{}^-O{-}Z({-}O{}^-H^+)_2] + (n\text{-}2)\,H_2O \quad (8)$$

$$Pt^{2+}[({}^-O{-})_2Z] + H_2 \rightarrow Pt + [({}^+H{}^-O{-})_2Z] \quad (9)$$

$[({}^-O{-})_nZ]$ = zeolite framework

Calcination of the hydrogen form above ca. 400 °C causes decomposition of the acidic hydroxyl groups with liberation of water. This dehydroxylation leads to formation of Lewis acid sites. As dehydroxylation takes place at similar temperatures to those used in deammoniation (Eq. 7), both types of acidic sites may be formed at the same time. The question

whether these consist of three-coordinate Al or Si sites within the framework structure or of AlO^+ species outside the framework structure continues to be the subject of debate [168].

Base catalysis by modified zeolites has not yet been used in industry. Suitable zeolites are ion exchanged with rubidium and cesium ions or contain sodium clusters or cesium oxide as active species in their pores [170]. Bifunctional catalysts in which the acidity of the zeolites is combined with the hydrogen-transferring function of noble metals are important in large-scale industrial processes such as hydrocracking or C_5/C_6 alkane hydroisomerization (see below).

A unique feature of the use of zeolites as catalysts is that the reactions can be shape-selective [166], [171], [172]. Shape selectivity in catalysis is closely related to the molecular-sieve effect in adsorption. Thus, it can occur if the pores of the catalyst have similar dimensions to those of the molecules or transition states taking part in the reaction:

1) In reactant shape selectivity, only reactants that can penetrate into the pore system by virtue of their slender shape and which can thus diffuse to the catalytic sites of the interior surfaces take part in the reaction
2) If the more bulky of two possible products is not formed, this is known as product shape selectivity
3) Restricted transition state shape selectivity occurs if the reaction pathway in the zeolite is directed by spatial requirements of the possible transition states

In (1) and (2), the selectivity is ultimately a result of diffusion limitation.

Shape selectivity has so far mainly been observed in reactions in which the catalyst is ZSM-5. However, the phenomenon is not limited to the use of this material, as shown by shape-selective reactions on erionite, ferrierite, mordenite, and other zeolites and zeolite-like materials.

Zeolite catalysts are used in four fields of technology:

Oil Refining. Zeolite-containing catalysts play a central role in several oil-refining processes (Table 5) [173], [174]. The most important is fluid catalytic cracking (FCC), which converts vacuum distillates and residues into gaseous alkenes, gasolines, and diesel fuel. Y type zeolites have been used as active components in FCC catalysts since 1964 [46], [86], [175]. These materials constitute 5 – 40 % of the catalyst, which is spray granulated to give a mean particle size of 60 µm.

Table 5. Zeolite catalysts in oil refining

Oil fraction	Zeolite-catalyzed process	Zeolites
Naphtha	C_5/C_6 alkane isomerization	Pt/H-mordenite, H-ZSM-5
	reforming of C_6/C_7 alkanes	Pt/K-L, Pt/BaK-L
	M-forming	H-ZSM-5
Gas oil, lubricating oil	dewaxing/hydrodewaxing	H-ZSM-5, Pt/H-mordenite
	hydrogenation of aromatics/hydrodecyclization	noble metal – zeolite
Vacuum distillates, residues	FCC	RE-HY, USY, H-ZSM-5
	hydrocracking	Pd-, Pt-, Ni/Mo-, Ni/W-USY
	residue hydroprocessing	specially modified USY zeolites

The zeolite catalysts have higher activity and give higher gasoline yields and less coke formation than the amorphous alumina – silica and high-alumina catalysts formerly used. Today, mainly USY zeolites (see Section 6.3.2) are chosen for FCC catalysts. The extensive use of USY zeolites has reduced the consumption of RE-HY zeolites [*83984-84-3*] in this application. Hybrids of the two types are also used. The trend towards the use of USY zeolites, which began in the early 1980s, was a result of the prohibition of tetraalkyllead compounds as antiknock agents, necessitating an increase in the octane number of gasoline. The disadvantage of RE-HY zeolites is their high activity with respect to hydrogen transfer between alkenes and naphthenes, which leads to the formation of aromatics and alkanes. The latter lower the research octane number (RON) and motor octane number (MON) of the gasoline. The use of the less active USY types avoids this effect, which has been explained in terms of the greater distance between the active sites. Moreover, the USY types have higher hydrothermal stability and give a further reduction in coke formation and an increased gasoline yield.

There has been much discussion of the use of ZSM-5 as an additional catalyst, mainly in combination with RE-HY catalysts. ZSM-5 is

generally used in the form of a separate spray-granulated material with a binder content of ca. 75 %. The proportion of this zeolite in the total catalyst formulation is 0.5 – 3.0 %. The original aim was "octane boosting" by shape-selective cracking of unbranched hydrocarbons. Addition of ZSM-5 led to an increase in RON but does not always affect MON. However, another interesting effect is the increase in the proportion of C_3/C_4 alkenes, mainly propene, in the product spectrum. These alkenes can be used, for example, in the alkylation of isobutane, and can thus contribute to the increase in the octane number of the gasoline pool. However, the use of ZSM-5 does cause a slight reduction in the gasoline yield on cracking. Besides ZSM-5, other zeolites, such as beta, have been tested for possible use in the FCC process [46].

Residues from the distillation of crude oil contain high molecular mass carbon compounds and are contaminated with nickel, vanadium, sulfur, and nitrogen compounds. When these residues are cracked – a process that will increase in importance – basically the same zeolites are used as those for the catalytic cracking of gas oils [86], [176]. The main differences are in the catalyst matrix, which must have large pores and be catalytically active to precrack large molecules. The matrix must also contain additives to bond with the metals and chemically reduce SO_x, and promoters of CO oxidation.

In the hydrocracking process, an alternative to catalytic cracking, low-grade, high-boiling gas oils are converted into fuels, mainly kerosene and diesel. In this process, which is flexible with respect to the product spectrum, the majority of the catalysts used today are bifunctional, consisting of USY zeolites doped with palladium, platinum (ca. 0.5 %), or combinations of base metals such as nickel – molybdenum and nickel – tungsten [177]. The zeolite content of the catalysts is 5 – 80 %. To protect the hydrogenation components, conventional HDS/HDN (hydrodesulfurization/hydrodenitrogenation) catalysts for the removal of organic sulfur and nitrogen compounds are used prior to the hydrocracking catalysts. Specially modified Y zeolites, e.g., Fe-USY impregnated with Co/Mo, seem to be particularly suitable for the hydrocracking of residues [178].

To supply the current demand for fuels low in aromatics, new processes in which straight-run middle distillates are subjected to hydrogenation have been considered. Noble metal – zeolite catalysts that are stable towards sulfur and nitrogen compounds are used to hydrogenate or hydrodecyclize the aromatics [179].

Catalytic dewaxing is a technology for shape-selective cracking of *n*-alkanes in gas and lubricating oils to reduce the pour point of waxlike raw materials to temperatures below the usual limit of −5 °C. Catalysts for dewaxing under hydrogenating conditions include H-ZSM-5 and Pt/H-mordenite [166], [180].

Another important process in oil refining is the isomerization of C_5/C_6 *n*-alkanes with Pt/H-mordenite or H-ZSM-5. Shell has combined this with the adsorptive separation of *n*- and iso-alkanes (UOP Isosiv process) to give the TIP process (total isomerization process) [173].

The zeolites Pt/K-L and Pt/BaK-L may be used in the future for dehydrocyclization, especially of C_6/C_7 alkanes, as conventional reforming catalysts (e.g., Pt/Cl – Al_2O_3) show poor selectivity in this range [173], [181]. Problems still arise from sensitivity of the new catalysts towards poisoning by sulfur. M-forming is operated as a post-reforming process with H-ZSM-5 as catalyst. This consists of shape-selective cracking of *n*-alkanes with simultaneous alkylation of benzene and toluene by the short-chain alkenes produced [166], [182].

Conversion of Gases. Both methanol from synthesis gas and LPG (liquefied petroleum gas) are raw materials for mineral oils and petrochemical products. Shape-selective reactions, catalyzed mainly by the zeolite ZSM-5, make these conversions possible.

The MTG (methanol to gasoline) process developed by Mobil [183], [184] and used in New Zealand to produce gasoline from methanol aroused much interest in the past. Investigations into the mechanism have shown that the methanol is first converted by the catalyst H-ZSM-5 into the intermediate dimethyl ether by splitting off water, and then into short-chain alkenes. These form oligomers which undergo conversion into alkanes/cycloalkanes and aromatics by hydrogen transfer and cyclization. However, more recently, processes to obtain gasoline di-

rectly from synthesis gas with the aid of zeolite catalysts are being developed.

In the MTO (methanol to olefins) process, which is closely related chemically to the MTG process, the object is to obtain C_2-C_4 alkenes from methanol. Possible catalysts for this process include ZSM-5, narrow-pore zeolites of the erionite or chabazite type, and structurally modified silicoaluminophosphates (SAPO-17, SAPO-34, and SAPO-44) [185]. The short-chain alkenes obtained can be converted into middle distillates in zeolite-catalyzed processes such as MOGD (Mobil olefin to gasoline and distillates) or SPGK (Shell poly gasoline and kero) [166], [173].

The CYCLAR process, recently developed jointly by UOP and BP, uses a Ga-ZSM-5 catalyst to synthesize aromatics from LPG (C_3/C_4 alkanes), which is available in abundance [186], [187].

Petrochemistry and Synthesis of Organic Intermediates. The prospect of producing petrochemicals and organic intermediates in shapeselective catalytic processes has led to intensive research [188–190]. This work was to some extent also motivated by the fact that some large-scale processes are increasingly under pressure for environmental reasons, e.g., processes using acid catalysts such as H_2SO_4, HF, H_3PO_4 on silica, and $AlCl_3$. Unlike these acids, zeolite catalysts are noncorrosive.

Typical zeolite-catalyzed reactions which have led to industrial processes include the alkylation of aromatics with alkenes, the disproportionation of alkyl- and chlorobenzenes, the isomerization, oligomerization, and hydration of alkenes, and the formation of amines from alcohols and ammonia. Relatively new reactions include oxidations using hydrogen peroxide as the oxidizing agent and the molecular sieve titanium silicalite-1 as the catalyst, e.g., the hydroxylation of aromatics, the ammoxidation of ketones, and the epoxidation of alkenes [24], [191].

A selection of processes which have been operated on an industrial scale are listed in Table 6 [166], [192]. Of these, the production of ethylbenzene by the reaction of ethene with benzene (Mobil – Badger process) and the isomerization of xylenes are the most important processes [166].

Environmental Protection. Advances have been made in the use of zeolitic catalysts for selective catalytic reduction of NO_x to nitrogen in the exhaust gases of lean-burn and diesel engines [201–204]. Metal-modified zeolites, especially Cu-ZSM-5 materials, have been intensively investigated. Suitable reducing agents include hydrocarbons, oxygen-containing, water-miscible organic compounds, and CO and H_2. However, the durability of the new catalyst under working conditions is still insufficient for broad application. Therefore, for the moment only stationary systems and mobile systems with diesel engines, where the heat and water-vapor stress is comparatively low, will be equipped with these catalysts [204]. Special zeolites have also been used for the reduction of NO_x with ammonia in commercial applications [205].

7.4. Other Uses

Zeolites can act as carriers or reservoirs for reactive substances. For example, zeolites containing silver, copper, or zinc are used commercially as antibacterial powders applied in man-made fibers for textiles for domestic or personal use [206], [207].

8. Economic Aspects of Synthetic Zeolites

Since synthetic zeolites have become available in commercial quantities, world consumption of these materials has greatly increased. Starting from a few hundred tonnes in the late 1950s, this became 60×10^3 t/a in 1975, 460×10^3 t/a in 1987 [91], and ca. 1.2×10^6 t/a in 1994. An overview of the important producers of zeolites for various application fields is given in Table 7.

Although the main use of zeolites was initially adsorption, detergent builders now represent the largest application quantitatively, 1 050 000 t/a or almost 90 % of zeolites produced in 1994 having been used for this purpose. Of this, ca. 950 000 t/a was zeolite NaA, and ca. 100 000 t/a NaP, a material which has only more recently been used in detergent formulations. In Japan and Western Europe, zeolites have increasingly replaced phosphates as

Table 6. Zeolite catalysts in petrochemical processes and processes for the production of organic intermediates

Process	Zeolites	Process developer	Refs.
Production of ethylbenzene from benzene and ethylene	H-ZSM-5	Mobil-Badger	[166]
Production of cumene from benzene and propene (liquid phase)	HY	Chemical Research & Licensing Co.	[193]
Production of *p*-ethyltoluene from toluene and ethylene	Modified ZSM-5	Mobil	[166]
Isomerization of xylene	Pt/H-ZSM-5 Pt/H-Mordenite	Mobil and others	[166]
Disproportionation of toluene to produce xylene	H-Mordenite ZSM-5	Mobil UOP/Toray	[166]
Production of *m*-chlorotoluene, *m*-di-chlorobenzene and 2,6-dichlorotoluene by isomerization processes	H-ZSM-5 H-Mordenite	Toray	[194]
Amination of methanol by ammonia to give dimethylamine as the principal product	Na/H-Mordenite	Nitto	[195], [196]
Hydration of cyclohexene to cyclohexanol	H-ZSM-5	Asahi	[196], [197]
Production of pyridine from acetaldehyde, formaldehyde, and ammonia	H-ZSM-5	Nepera	[198]
High-purity isobutene from MTBE	B-ZSM-5	Snamprogetti	[199]
Cyclohexanone oxime by ammoximation of cyclohexanone	Titanium silicalite-1	Enichem	[24], [191], [200]
Hydroxylation of phenol to hydroquinone/catechol	Titanium silicalite-1	Enichem	[24], [191]

detergent builders since the late 1970s. In the United States, this development was interrupted in 1983, but since 1987 detergents with zeolite-containing formulations have been increasingly used there also. In 1994, ca. 55 % of all detergent zeolites were consumed in Western Europe, 30 % in the United States, and 15 % in the Far East (mainly Japan, Taiwan, and Korea).

Compared with the enormous consumption of detergent zeolites, the use of zeolites for catalytic purposes seems small, the world total for 1994 being 115 000 t/a. However, when applying monetary criteria, it must be remembered that the value addition achieved by the use of catalysts greatly exceeds the cost of the catalysts themselves [192]. Of the zeolites used in catalysts, type Y zeolites occupy a dominant position (ca. 105 000 t/a), and of this, 100 000 t/a is a constituent of FCC catalysts alone. This figure is based on an average content of 25 % Y zeolite in the catalyst. Approximately 5000 t/a is used in catalysts for hydrocracking and other special processes. The consumption of ZSM-5 zeolites for the purposes mentioned in Section 7.3 is ca. 3000 t/a, and of synthetic mordenite 600 t/a. Other zeolites are quantitatively of minor significance.

The zeolites used in adsorption are mainly types A and X. The total zeolite consumption for the diverse individual uses (see Section 7.2) was estimated to be ca. 50 000 t/a in 1994. Over half the catalysts and adsorbents that contain zeolites are produced and used in the United States. Western Europe and Eastern Asia follow some way behind.

Table 7. Synthetic-zeolite producers

Application	Zeolites	Producers
Detergent builders	NaA	Degussa AG (DE), Enimont (IT), Ethyl Corporation (US), Henkel KGaA (DE), Mizusawa Industrial Chemicals (JP), The PQ Corporation (US), Soprolit SA (BE), Zeochem SA (BE)
	NaP	Crosfield Chemicals (GB)
Catalysis	Y zeolites	Akzo Chemicals BV (NL), Catalysts & Chemical Ind. Corp. (JP), Engelhard Corporation (US), W.R. Grace & Co. (US), The PQ Corporation/Zeolyst Enterprises (US), UOP (US)
	ZSM-5, mordenite, and other specialty zeolites	ALSI-Penta Zeolithe GmbH (DE), CU Chemie Uetikon AG (CH), Degussa AG (DE), Eka Nobel AB (SE), Engelhard Corporation (US), Mobil Corporation (US), Südchemie (DE), Tosoh Corporation (JP), UOP (US)
Adsorption	A, X, specialty zeolites	Bayer AG (DE), Ceca SA (FR), Degussa AG (DE), W.R. Grace & Co. (US), CU Chemie Uetikon AG (CH), UOP (US), Tosoh Corporation (JP), Zeochem Company (US)

9. Toxicology

Epidemiological [208] and experimental investigations [209], [210] have shown that natural erionite, as found in Cappadocia (Turkey), can cause formation of mesotheliomas and bronchial carcinomas. The oncogenic activity of these minerals is explained in a similar way to that of asbestos, which has a fibrous crystal morphology. This crystal form is extremely rare in zeolites.

Natural erionite is so far the only example of a zeolite with toxic effects. A review can be found in [211]. Zeolite NaA, which is used as a detergent builder, was found to have negligible acute, subacute, chronic, or inhalative toxicity in comprehensive tests [66], [93], [212]. Furthermore, regular medical checkups of workers exposed to zeolite NaA for a period of years in production areas gave no indication whatever of fibrogenic or silicogenic effects [66]. In Germany, the MAK is 6 mg/m^3, i.e., the general dust threshold [93], [213].

10. References

1. R. M. Barrer, *Pure Appl. Chem.* **51** (1979) 1091 – 1100.
2. L. V. C. Rees, *Nature (London)* **296** (1982) 491 – 492.
3. F. Liebau, *Zeolites* **3** (1983) 191 – 193.
4. J. V. Smith, *Zeolites* **4** (1984) 309 – 310.
5. D. W. Breck: *Zeolite Molecular Sieves,* Wiley, New York 1974.
6. A. F. Cronstedt, *Svenska Vetenskaps Akademiens Handlingar Stockholm* **17** (1756) 120 – 123.
7. J. L. Schlenker, G. H. Kühl in R. von Ballmoss, J. B. Higgins, M. M. J. Treacy (eds.): *Proceedings from the Ninth International Zeolite Conference,* Butterworth-Heinemann, Boston 1993, p. 3.
8. R. M. Milton in M. L. Occelli, H. E. Robson (eds.): "Zeolite Synthesis," *ACS Symp. Ser.* **398** (1989) 1.
9. J. A. Rabo (ed.): "Zeolite Chemistry and Catalysis," *ACS Monogr.* **171** (1976).
10. R. M. Barrer: *Zeolite and Clay Minerals as Sorbents and Molecular Sieves,* Academic Press, London 1978.
11. R. M. Barrer: *Hydrothermal Chemistry of Zeolites,* Academic Press, London 1982.
12. A. Dyer: *An Introduction to Zeolite Molecular Sieves,* Wiley, New York 1988.
13. R. Szostak: *Molecular Sieves,* Van Nostrand Reinhold, New York 1989.
14. H. van Bekkum, E. M. Flanigen, J. C. Jansen (eds.) "Introduction to Zeolite Sciene and Practice," *Stud. Surf. Sci. Catal.* **58** (1991).
15. E. G. Derouane, F. Lemos, C. Naccache, F. R. Ribeiro (eds.): "Zeolite Microporous Solids: Synthesis, Structure, and Reactivity," *NATO ASI Ser. C.* **352** (1992).
16. J. C. Jansen, M. Stöcker, H. G. Karge, J. Weitkamp (eds.): "Advanced Zeolite Science and Applications," *Stud. Surf. Sci. Catal.* **85** (1994).
17. L. B. Sand, F. A. Mumpton (eds.): *Natural Zeolites,* Pergamon Press, New York 1978.
18. G. Gottardi, E. Galli: *Natural Zeolites,* Springer Verlag, Berlin 1985.
19. G. V. Tsitsishvili, T. G. Andronikashvili, G. N. Kirov, L. D. Filizova: *Natural Zeolites,* Ellis Horwood, New York 1992.
20. J. A. Rabo in [15], p. 531.
21. H. Kessler in R. von Ballmoos, J. B. Higgins, M. M. J. Treacy (eds.): *Proceedings from the Ninth International Zeolite Conference,* Butterworth-Heinemann, Boston 1993, p. 73.
22. N. A. Kutz in W. H. Flank, T. E. Whyte, Jr. (eds.): *"Perspectives in Molecular Sieve Science," ACS Symp. Ser.* **368** (1988) 532.
23. P. Ratnasamy, R. Kumar, *Catal. Today* **9** (1991) 329 – 416.
24. G. Bellussi, M. S. Rigutto in [16], p. 177.
25. E. M. Flanigen, R. L. Patton, S. T. Wilson in P. J. Grobet, W. J. Mortier, E. F. Vansant, G. Schulz-Ekloff (eds.): "Innovation in Zeolite Materials Science," *Stud. Surf. Sci. Catal.* **37** (1988) 13.
26. S. T. Wilson in [14], p. 137.
27. J. A. Martens, P. A. Jacobs in [16], p. 653.
28. W. M. Meier, D. H. Olson: *Atlas of Zeolite Structure Types,* 3rd revised ed., Butterworth-Heinemann, London 1992.
29. H. van Koningsfeld in [14], p. 35.
30. M. E. Davis, *Acc. Chem. Res.* **26** (1993) 111 – 115.
31. M. E. Davis, *Catal. Today* **19** (1994) 1 – 6.
32. J. B. Higgins, *Catal. Today* **19** (1994) 7 – 26.
33. R. M. Barrer in D. M. Bibby, C. D. Chang, R. F. Howe, S. Yurchak (eds.): "Methane Conversion," *Stud. Surf. Sci. Catal.* **36** (1988) 307.
34. W. J. Mortier: *Compilation of Extra Framework Sites in Zeolites,* Butterworth, London 1982.

35. J. V. Smith in [9], p. 1.
36. H. Kosslick, A. Roethe, K. P. Roethe, *Chem. Tech. (Leipzig)* **46** (1994) 101 – 107; **47** (1995) 61 – 67.
37. W. Loewenstein, *Am. Mineral.* **39** (1954) 92 – 96.
38. E. M. Flanigen, *Pure Appl. Chem.* **52** (1980) 2191 –2211.
39. J. Dwyer, F. R. Fitch, G. Qin, J. C. Vickerman, *J. Phys. Chem.* **86** (1982) 4574 – 4578.
40. A. Tissler et al. in H. G. Karge, J. Weitkamp (eds.): "Zeolites as Catalysts, Sorbents and Detergent Builders," *Stud. Surf. Sci. Catal.* **46** (1989) 399.
41. A. Kupermann et al., *Nature (London)* **265** (1993) 239 – 242.
42. D. H. Olsen, W. O. Haag, R. M. Lago, *J. Catal.* **61** (1980) 390 – 396.
43. G. A. Ozin, A. Kuperman, A. Stein, *Angew. Chem.* **101** (1989) 373 – 390; *Angew. Chem. Int. Ed. Engl.* **28** (1989) 359 – 376.
44. D. R. Rolison in [16], p. 544.
45. J. Scherzer in T. E. Whyte, Jr., R. A. Dalla Betta, E. G. Derouane, R. T. K. Baker (eds.): "Catalytic Materials," *ACS Symp. Ser.* **248** (1984) 157.
46. J. Biswas, I. E. Maxwell, *Appl. Catal.* **63** (1990) 197 –258.
47. C. V. McDaniel, P. K. Maher in [9], p. 285.
48. M. Briend-Faure et al., *Appl. Catal.* **38** (1988) 71 – 87.
49. J. H. C. van Hooff, J. W. Roelofsen in [14], p. 241.
50. R. von Ballmoos, J. B. Higgins: *Collection of Stimulated XRD Powder Patterns for Zeolites,* 2nd revised ed., Butterworth-Heinemann, London 1990.
51. J. M. Thomas, D. E. W. Vaughan, *J. Phys. Chem. Solids* **50** (1989) 449 – 467.
52. L. B. McCusker in J. Weitkamp, H. G. Kage, H. Pfeifer, W. Hölderich (eds.): "Zeolites and Related Microporous Materials: State of the Art 1994," *Stud. Surf. Sci. Catal.* **84** (1994) 341.
53. ASTM D 3906-91.
54. D. R. Corbin, B. F. Burgess, Jr., A. J. Vega, R. D. Farlee, *Anal. Chem.* **59** (1987) 2722 – 2728.
55. G. Engelhardt, D. Michel: *High-Resolution Solid-State NMR of Silicates and Zeolites,* Wiley, Chichester 1987.
56. J. Klinowski, *Chem. Rev.* **91** (1991) 1459 – 1479.
57. C. A. Fyfe et al., *Chem. Rev.* **91** (1991) 1525 – 1543.
58. ASTM D 3942-85.
59. H. Fichtner-Schmittler, U. Lohse, H. Miessner, H. E. Maneck, *Z. Phys. Chem. (Leipzig)* **271** (1990) 69 – 79.
60. J. Seifert, G. Emig, *Chem. Ing. Tech.* **59** (1987) 475 –484.
61. J. Fraissard, T. Ito, *Zeolites* **8** (1988) 350 – 361.
62. S. Ernst, R. Kumar, M. Neuber, J. Weitkamp in K. K. Unger, J. Rouquero, K. S. W. Sing, H. Kral (eds.): "Characterization of Porous Solids," *Stud. Surf. Sci. Catal.* **39** (1988) 531.
63. A. Humphries, D. H. Harris, P. O'Connor in [175], p. 41.
64. H. Pfeifer in: *NMR Basic Principles and Progress,* vol. 31, Solid-State NMR II: Inorganic Matter, Springer Verlag, Berlin 1994, p. 31.
65. J. W. Ward in [9], p. 118.
66. Degussa, Technical Bulletin Pigments, 4th ed., no. 71, Aug. 1993.
67. J. Weitkamp in P. J. Grobet, W. J. Mortier, E. F. Vansant, G. Schulz-Ekloff (eds.): "Innovation in Zeolite Materials Science," *Stud. Surf. Sci. Catal.* **37** (1988) 515.
68. R. L. Hay in Y. Murakami, A. Iijima, J. W. Ward (eds.): "New Developments in Zeolite Science and Technology," *Stud. Surf. Sci. Catal.* **28** (1986) 35.
69. K. Torii in [17], p. 441.
70. F. A. Mumpton in [17], p. 3.
71. A. Dyer, *Chem. Ind. (London)* 1984, 241 – 245.
72. J. D. Sherman in F. R. Ribeiro, A. E. Rodrigues, L. R. Rollmann, C. Naccache (eds.): "Zeolites: Science and Technology," *NATO ASI Ser. E* **80** (1984) 583.
73. P. A. Jacobs, J. A. Martens: "Synthesis of High-Silica Aluminosilicate Zeolites," *Stud. Surf. Sci. Catal.* **33** (1987).
74. J. L. Guth et al. in M. L. Occelli, H. E. Robson (eds.): "Zeolite Synthesis," *ACS Symp. Ser.* **398** (1989) 176.
75. B. M. Lok, T. R. Cannan, C. A. Messina, *Zeolites* **3** (1983) 282 – 291.
76. G. P. Gilson in [15], p. 19.
77. R. M. Barrer in M. L. Occelli, H. E. Robson (eds.): "Zeolite Synthesis," *ACS Symp. Ser.* **398** (1989) 11.
78. J. C. Jansen in [14], p. 77.
79. E. J. P. Feijen, J. A. Martens, P. A. Jacobs in J. Weitkamp, H. G. Karge, H. Pfeifer, W. Hölderich (eds.): "Zeolites and Related Microporous Materials: State of the Art 1994," *Stud. Surf. Sci. Catal.* **84** (1994) 3.

80. A. V. McCormick, A. T. Bell, *Catal. Rev. Sci. Eng.* **31** (1989) 97 – 127.
81. S. Gonthier, R. W. Thompson in [16], p. 43.
82. Union Carbide Corp., US 3 119 659, 1960 (L. R. L. Taggart, G. L. Ribaud).
83. Toyo Soda Manufacturing Co., EP 9 226 195, 1986 (M. Harada, H. Miyazaki).
84. W. R. Grace & Co., DE 2 707 313, 1977 (C. W. Chi, G. H. Hoffman).
85. W. R. Grace & Co., US 3 906 076, 1973 (J. A. Goytisolo, D. D. Chi, H. Lee).
86. J. Scherzer: "Octane-Enhancing Zeolitic FCC Catalysts," *Chem. Ind. (Dekker)* **42** (1990); *Catal. Rev. Sci. Eng.* **31** (1989) 215 – 354.
87. G. M. Woltermann, J. S. Magee, S. D. Griffith in [175], p. 105.
88. Engelhard Corp., US 4 493 902, 1983 (S. M. Brown, V. A. Durante, W. J. Reagan, B. K. Speronello).
89. C. W. Roberts in R. P. Townsend (ed.): *The Properties and Applications of Zeolites,* The Chemical Society, London 1980, p. 103.
90. D. E. W. Vaughan, *Chem. Eng. Prog.* **84** (1988) 25 – 31.
91. E. Roland in H. G. Karge, J. Weitkamp (eds.): "Zeolites as Catalysts, Sorbents and Detergent Builders," *Stud. Surf. Sci. Catal.* **46** (1989) 645.
92. Degussa, DE 2 660 722, DE 2 660 725, 1976 (H. Strack, W. Roebke, D. Kneitel, E. Parr).
93. P. Christophliemk, P. Gerike, M. Potokar in P. Hutzinger, N. T. de Oude (eds.): *The Handbook of Environmental Chemistry,* **vol. 3 F (Detergents),** Springer Verlag, Berlin 1992, p. 205.
94. Henkel, EP 0 037 018, 1981 (F. J. Carduck, B. W. Peters, O. Koch, R. Vogler, P. Cristophliemk, W. Wüst).
95. I. Yamane, T. Nakazawa in Y. Murakami, A. Iijima, J. W. Ward (eds.): "New Developments in Zeolite Science and Technology," *Stud. Surf. Sci. Catal.* **28** (1986) 991.
96. G. Jacobi, A. Löhr: *Detergents and Textile Washing,* VCH Verlagsgesellschaft, Weinheim 1987.
97. M. Hadan, F. Fischer, *Cryst. Res. Technol.* **27** (1992) 343 – 350.
98. H. Robson in M. L. Occelli, H. E. Robson (eds.): "Zeolite Synthesis," *ACS Symp. Ser.* **398** (1989) 436.
99. S. Ernst, J. Weitkamp, *Catal. Today* **19** (1994) 27 – 60.
100. S. Kasahara, K. Itabashi, K. Igawa in Y. Murakami, A. Iijima, J. W. Ward (eds.): "New Developments in Zeolite Science and Technology," *Stud. Surf. Sci. Catal.* **28** (1986) 185.
101. H. Lechert, P. Staelin, M. Wrobel, U. Schimmel in J. Weitkamp, H. G. Karge, H. Pfeifer, W. Hölderich (eds.): "Zeolites and Related Microporous Materials: State of the Art 1994," *Stud. Surf. Sci. Catal.* **84** (1994) 147.
102. J. Janiga, *Zeolites* **10** (1990) 38 – 43.
103. W. R. Grace & Co., US 4 164 551, 1979 (C. H. Elliott, Jr.).
104. Mobil Oil Corp., US 3 702 886, 1969 (R. J. Argauer, G. R. Landolt).
105. F. Y. Dai, M. Suzuki, H. Takahashi, Y. Saito in M. L. Occelli, H. E. Robson (eds.): "Zeolite Synthesis," *ACS Symp. Ser.* **398** (1989) 244.
106. VAW Aluminium, EP 0 402 801, 1990 (R. Thome, A. Tissler).
107. P. Gallezot, *Catal. Rev. Sci. Eng.* **20** (1979) 121 – 154.
108. W. M. H. Sachtler, Z. Zhang, *Adv. Catal.* **39** (1993) 129 – 220.
109. H. G. Karge, H. K. Beyer in P. A. Jacobs, N. I. Jaeger, L. Kubelková, B. Wichterlová (eds.): "Zeolite Chemistry and Catalysis," *Stud. Surf. Sci. Catal.* **69** (1991) 43.
110. Degussa, DE 4 304 821, 1993 (J. Weitkamp, S. Ernst, T. Bock, T. Kromminga, A. Kiss, P. Kleinschmit).
111. P. Townsend in [14], p. 359.
112. D. Keir, E. F. T. Lee, L. V. C. Rees, *Zeolites* **8** (1988) 228 – 231.
113. T. L. M. Maesen et al., *J. Chem. Soc. Faraday Trans.* **87** (1991) 787 – 790.
114. J. D. Sherman in J. D. Sherman (ed.): "Adsorption and Exchange Separations," *AIChE Symp. Ser.* **74** (1978) no. 179, 98.
115. A. Dyer, H. Enamy, R. P. Townsend, *Sep. Sci. Technol.* **16** (1981) 173 – 183.
116. H. Sherry, *Zeolites* **13** (1993) 377 – 383.
117. R. Szostak in [14], p. 153.
118. H. G. Karge, J. Weitkamp, *Chem. Ing. Tech.* **58** (1986) 946 – 959.
119. M. M. Olken, J. M. Garces in R. von Ballmoos, J. B. Higgins, M. M. J. Treacy (eds.): *Proceedings from the Ninth International Zeolite Conference,* Butterworth-Heinemann, Boston 1993, p. 559.
120. J. A. Martens, P. J. Grobet, P. A. Jacobs in G. Poncelet, P. A. Jacobs, P. Grange, B. Delmon (eds.): "Preparation of Catalysts V," *Stud. Surf. Sci. Catal.* **63** (1991) 355.
121. Degussa, EP 0 072 397, 1982 (H. Strack, P. Kleinschmit).

122. Union Carbide Corp., US 4 503 023, 1981 (D. W. Breck, G. W. Skeels).
123. G. Garralón, V. Fornés, A. Corma, *Zeolites* **8** (1988) 268 – 272.
124. E. F. Vansant: *Pore Size Engineering in Zeolites,* Wiley, New York 1990.
125. M. Niwa, S. Kato, T. Hattori, Y. Murakami, *J. Chem. Soc. Faraday Trans. 1* **80** (1984) 3135 – 3145.
126. J. H. Lunsford, *Rev. Inorg. Chem.* **9** (1987) 1 – 35.
127. G. Meyer, D. Wöhrle, M. Mohl, G. Schulz-Ekloff, *Zeolites* **4** (1984) 30 – 34.
128. C. E. Capes: *Particle Size Enlargement,* Elsevier, Amsterdam 1980.
129. A. B. Stiles: "Catalyst Manufacture: Laboratory and Commercial Preparations," *Chem. Ind. (Dekker)* **14** (1983).
130. U. Hammon, M. Kotter, *Chem. Ing. Tech.* **56** (1984) 455 – 463.
131. Degussa, DE 3 132 379, 1981 (H. Strack, E. Parr, P. Kleinschmit).
132. R. Rautenbach, O. Machhammer, *Chem. Ing. Tech.* **60** (1988) 23 – 31.
133. E. P. Hertzenberg in D. Olson, A. Bisio (eds.): *Proceedings of the Sixth International Zeolite Conference,* Butterworths, Guildford 1984, p. 975.
134. Unilever, EP 0 384 070, 1989 (G. T. Brown, T. J. Osinga, M. J. Parkington, A. T. Steel).
135. Joseph Crosfield & Sons, EP 0 565 364, 1993 (A. Araya).
136. L. V. C. Rees in R. P. Townsend (ed.): *Properties and Applications of Zeolites,* The Chemical Society, London 1980, p. 218.
137. M. J. Schwuger, M. Liphard in H. G. Karge, J. Weitkamp (eds.): "Zeolites as Catalysts, Sorbents and Detergent Builders," *Stud. Surf. Sci. Catal.* **46** (1989) 673.
138. R. P. Denkewicz, Jr., A. G. Monino, D. E. Russ, H. S. Sherry, *JAOCS J. Am. Oil Chem. Soc.* **72** (1995) 31 – 35.
139. H. Upadek, P. Krings, *Seifen Öle Fette Wachse* **117** (1991) 554 – 558.
140. Umweltbundesamt (ed.): *Materialien 4/79: Die Prüfung des Umweltverhaltens von Natrium-Aluminium-Silikat Zeolith A als Phosphatersatzstoff in Wasch- und Reinigungsmitteln,* Erich Schmidt Verlag, Berlin 1979.
141. P. Kuhm, W. Lortz in J. Weitkamp, H. G. Karge, H. Pfeifer, W. Hölderlich (eds.): "Zeolites and Related Microporous Materials: State of the Art 1992," *Stud. Surf. Sci. Catal.* **84** (1994) 2303.
142. D. M. Ruthven: *Principles of Adsorption and Adsorption Processes,* Wiley, New York 1984.
143. D. M. Ruthven, *Chem. Eng. Progr.* **84** (1988) 42 – 50.
144. H. Knoll, *Chem. Tech. (Leipzig)* **26** (1974) 391 – 396.
145. E. Richter, *Erdöl Kohle Erdgas Petrochem.* **40** (1987) 394 – 399, 432 – 438.
146. H. Knoll, U. Harms, W. Höse, H. Fürtig, *Chem. Tech. (Leipzig)* **42** (1990) 149 – 154.
147. R. Wolter, *Glaswelt* **36** (1983) 323 – 324.
148. Bayer, EP 0 310 916, 1988 (L. Puppe, M. Mengel).
149. A. L. Weiner, *Chem. Eng. (N.Y.)* **81** (1974) no. 19, 92 – 101.
150. W. Kast: *Adsorption aus der Gasphase,* VCH Verlagsgesellschaft, Weinheim 1988.
151. J. Schwarz, *Luft Kältetech.* **27** (1991) no. 3, 127 – 132.
152. W. Otten, E. Gail, T. Frey, *Chem. Ing. Tech.* **64** (1992) 915 – 925.
153. O. Petzoldt, G. Poetsch, *WLB Wasser Luft Boden* **38** (1994) nos. 7 – 8, 44 – 47.
154. W. C. Miller, *Chem. Eng. (N.Y.)* **80** (1973) no. 18, 62 – 63.
155. Union Carbide Corp., US 4 795 482, 1987 (A. J. Gioffre, B. K. Marcus).
156. Metallgesellschaft, DE 4 128 106, 1991 (P. Cornel, H. W. Bräuer, D. Rolke).
157. UOP, US 4 874 525, 1989 (J. Markovs).
158. D. Werner, *Chem. Ing. Tech.* **53** (1981) 73 – 81.
159. A. M. Watson, *Hydrocarbon Process* **62** (1983) no. 3, 91 – 95.
160. U. Harms et al., *Chem. Tech. (Leipzig)* **41** (1989) 183 – 187.
161. G. Reiss in H. G. Karge, J. Weitkamp (eds.): "Zeolite as Catalysts, Sorbents and Detergent Builders," *Stud. Surf. Sci. Catal.* **46** (1989) 607.
162. Air Products and Chemicals, US 5 152 813, 1991 (C. G. Coe, J. F. Kirner, R. Pierantozzi, T. R. White).
163. J. A. Johnson, A. R. Oroskar in H. G. Karge, J. Weitkamp (eds.): "Zeolites as Catalysts, Sorbents and Detergent Builders," *Stud. Surf. Sci. Catal.* **46** (1989) 451.
164. P. Deckert, W. Arlt, *Chem. Ing. Tech.* **66** (1994) 1334 – 1340.
165. W. O. Haag, N. Y. Chen in L. L. Hegedus (ed.): *Catalyst Design – Progress & Perspectives,* Wiley, New York 1987, p. 163.
166. N. Y. Chen, W. E. Garwood, F. G. Dwyer: "Shape Selective Catalysis in Industrial Applications," *Chem. Ind. (Dekker)* **36** (1989).

167. K. Tanabe, M. Misono, Y. Ono, H. Hattori: "New Solid Acids and Bases," *Stud. Surf. Sci. Catal.* **51** (1989) 142.
168. J. Dwyer in [14], pp. 303, 321
169. D. Barthomeuf in B. Imelik et al. (eds.): "Catalysis in Zeolites," *Stud. Surf. Sci. Catal.* **5** (1980) 55 – 64.
170. C. B. Dartt, M. E. Davis, *Catal. Today* **19** (1994) 151 – 186.
171. S. M. Csicsery, *Pure Appl. Chem.* **58** (1986) 841 – 856.
172. J. Weitkamp, S. Ernst, H. Dauns, E. Gallei, *Chem. Ing. Tech.* **58** (1986) 623 – 632.
173. I. E. Maxwell, W. H. J. Stork in [14], p. 571.
174. A. Corma in [15], p. 373.
175. J. S. Magee, M. M. Mitchell, Jr. (eds.): "Fluid Catalytic Cracking: Science and Technology," *Stud. Surf. Sci. Catal.* **76** (1993).
176. M. M. Mitchell, Jr., J. F. Hoffman, H. F. Moore in [175], p. 293.
177. J. W. Ward, *Fuel Process. Technol.* **35** (1993) 55 – 85.
178. S. Hidaka et al. in Y. Murakami, A. Iijima, J. W. Ward (eds.): "New Developments in Zeolite Science and Technology," *Stud. Surf. Sci. Catal.* **28** (1986) 329.
179. J. P. van den Berg, J. P. Lucien, G. Germaine, G. L. B. Thielemans, *Fuel Process. Technol.* **35** (1993) 119 – 136.
180. K. W. Smith, W. C. Starr, N. Y. Chen, *Oil Gas J.* **78** (1980) no. 21, 75 – 84.
181. E. Iglesia, J. E. Baumgartner in L. Guczi, F. Solymosi, P. Tétényi (eds.): "New Frontiers in Catalysis," *Stud. Surf. Sci. Catal.* **75** (1993) 993.
182. N. Y. Chen, W. E. Garwood, R. H. Heck, *Ind. Eng. Chem. Res.* **26** (1987) 706 – 711.
183. C. D. Chang, *Catal. Rev. Sci. Eng.* **25** (1983) 1 – 118.
184. C. D. Chang, *Catal. Today* **13** (1992) 103 – 111.
185. G. F. Froment, W. H. J. Dehertog, A. J. Marchi in J. J. Spivey (ed.): *Catalysis,* **vol. 9,** The Royal Society of Chemistry, Cambridge 1992, p. 1.
186. M. Guisnet, N. S. Gnep, F. Alario, *Appl. Catal.* **A 89** (1992) 1 – 30.
187. P. R. Pujadó, J. A. Rabó, G. J. Antos, S. A. Gembicki, *Catal. Today* **13** (1992) 113 – 141.
188. P. Venuto, *Microporous Materials* **2** (1994) 297 – 411.
189. W. F. Hölderich, H. van Bekkum in [14], p. 631.
190. W. Hölderich, M. Hesse, F. Näumann, *Angew. Chem.* **100** (1988) 232 – 251; *Angew. Chem. Int. Ed. Eng.* **27** (1988) 226 – 246.
191. B. Notari in T. Inui, S. Namba, T. Tatsumi (eds.): "Chemistry of Microporous Crystals," *Stud. Surf. Sci. Catal.* **60** (1991) 343.
192. W. F. Hölderlich in L. Guczi, F. Solymosi, P. Tétényi (eds.): "New Frontiers in Catalysis," *Stud. Surf. Sci. Catal.* **75** (1993) 127.
193. Chemical Research & Licensing Co., US 4 849 569, 1988 (L. A. Smith, Jr.).
194. Toray Industries, EP 0 278 729, 1987 (K. Iwayama, Y. Magatani, K. Tada).
195. Nitto Kagaku Kogyo Kabushiki Kaisha, EP 125 616, 1984 (Y. Ashina, T. Fujita, M. Fukatsu, J. Yagi).
196. M. Misono, N. Nojiri, *Appl. Catal.* **64** (1990) 1 – 30.
197. Asahi Kasei Kogyo, DE 3 441 072, 1984 (O. Mitsui, K. Yohei).
198. Nepera, EP 0 131 887, 1984 (D. Feitler, W. Schimming, H. Wetstein).
199. Snamprogetti, US 4 656 016, 1984 (M. Taramasso, G. Manara, V. Fattore, B. Notari).
200. A. Zecchina et al. in L. Guczi, F. Solymosi, P. Té-tényi (eds.): "New Frontiers in Catalysis," *Stud. Surf. Sci. Catal.* **75** (1993) 719.
201. M. Iwamoto, H. Yahiro, *Catal. Today* **22** (1994) 5 – 18.
202. S. Matsumoto et al., *Catal. Today* **22** (1994) 127 –146.
203. T. Tabata, M. Kokitsu, O. Okada, *Catal. Today* **22** (1994) 147 – 169.
204. M. Shelef, *Chem. Rev.* **95** (1995) 209 – 225.
205. J. W. Byrne, J. M. Chen, B. K. Speronello, *Catal. Today* **13** (1992) 33 – 42.
206. Kanebo, EP 0 103 214, 1983 (H. Zenji, O. Hideo, H. Shigetaka, N. Saburo, I. Shunja, T. Kenichi).
207. Shinagawa Fuel Co., EP 0 270 129, 1987 (R. Niira, T. Yamamoto, M. Uchida).
208. I. Baris et al., *Int. J. Cancer* **39** (1987) 10 – 17.
209. J. W. Wagner, J. W. Skidmore, R. J. Hill, D. M. Griffiths, *Br. J. Cancer* **51** (1985) 727 – 730.
210. K. T. Kelsey, E. Yano, H. L. Liber, J. B. Little, *Br. J. Cancer* **54** (1986) 107 – 114.
211. J. A. Thomas, B. Ballantyne, *J. Am. Coll. Toxicol.* **11** (1992) 259 – 273.
212. C. Gloxhuber et al., *Food Chem. Toxicol.* **21** (1983) 209 – 220.
213. DFG: *Maximale Arbeitsplatzkonzentrationen und biologische Arbeitsstofftoleranzwerte,* VCH Verlagsgesellschaft, Weinheim 1988.

Ziegler Processes

ARNO BEHR, Universität Dortmund, Dortmund, Federal Republic of Germany

1. **Introduction** 1
2. **Ziegler Direct Process** 1
3. **Ziegler Olefin Synthesis** 2
4. **Ziegler Alcohol Synthesis** 2
5. **Ziegler Catalysts** 3
6. **References** 4

1. Introduction

KARL ZIEGLER (1898 to 1973) was a Professor at the Universities of Heidelberg, Halle, and Aachen. From 1943 to 1969 he directed the Max Planck Institut für Kohlenforschung in Mülheim/Ruhr, Germany, where he led many research projects in the fields of organometallic chemistry and metal-catalyzed synthesis. His work was crowned by the award of the Nobel Prize for Chemistry, which he received jointly with GIULIO NATTA in 1963.

The research of ZIEGLER led to many industrial processes, most of which have already been described under other keywords. This article gives a brief overview of these processes, emphasizing the features they have in common. They include the Ziegler direct process for the production of aluminum triorganyls (Chap. 2), the industrial synthesis of Ziegler olefins (Chap. 3) and Ziegler alcohols (Chap. 4), and the versatile uses of Ziegler catalysts in the production of oligomers or polymers (Chap. 5).

2. Ziegler Direct Process

The Ziegler direct process is an important method of synthesizing aluminum trialkyls AlR_3 and dialkylaluminum hydrides R_2AlH (→ Aluminum Compounds, Organic, Chap. 1.). It is a loop process in which two reactions, known as hydrogenation and hydroalumination, proceed in parallel. In the hydrogenation step, aluminum reacts with hydrogen in the presence of AlR_3 to form dialkylaluminum hydrides. In the hydroalumination step, the aluminum hydride undergoes addition of an olefin. Taking triethylaluminum as an example, these reactions proceed as follows:

Hydrogenation:

$$Al + 1.5\,H_2 + 2\,AlEt_3 \xrightarrow[10-20\ MPa]{80-160\,^\circ C} 3\,Et_2AlH$$

Hydroalumination:

$$3\,Et_2AlH + 3\,CH_2{=}CH_2 \xrightarrow[0.1-1\ MPa]{80-110\,^\circ C} 3\,AlEt_3$$

Overall reaction:
$Al + 1.5\,H_2 + 3\,CH_2{=}CH_2 \longrightarrow AlEt_3$

The overall reaction is thus the formation of aluminum trialkyls from activated aluminum powder, hydrogen, and α-olefins.

Variation of the aluminum/olefin ratio enables the reaction to be directed towards the formation of trialkyls or towards dialkyl hydrides. The alkyl group can be varied. Apart from ethyl compounds, *n*-butyl and isobutyl compounds in particular are produced on an industrial scale. The industrial synthesis is continuous, and can be carried out in one or two stages. The ready availability of triethylaluminum afforded by the Ziegler direct process was an important precondition for the development of the stoichiometric and catalytic reactions of $AlEt_3$ described in the following Sections.

3. Ziegler Olefin Synthesis

In the early 1950s, ZIEGLER discovered the "growth reaction," in which the repeated insertion of ethylene into aluminum – ethyl bonds produces long-chain aluminum trialkyls, from which long-chain α-olefins were obtained in a subsequent displacement reaction [2]:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_505

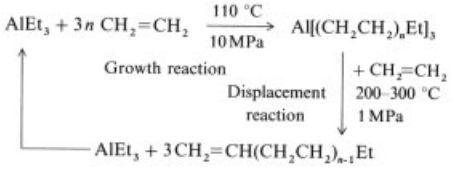

This process has been realized on an industrial scale by Gulf Oil and Ethyl Corporation. The two processes are briefly discussed below. A detailed treatment can be found in → Hydrocarbons.

In the *Gulf Oil Process*, only catalytic quantities of triethylaluminum are used. The growth and displacement reactions take place in the same reactor. This process yields a relatively broad molecular mass distribution of linear $C_4 - C_{30}$ olefins. These products ("Gulftenes") are produced in a Gulf Oil plant at Cedar Bajou, Texas. Mitsubishi uses the same process in a plant at Mizushima, Japan.

In the *Ethyl process*, the Ziegler olefin synthesis is carried out in two successive steps. The first step, which uses catalytic quantities of triethylaluminum produces a broad distribution of olefins, as in the Gulf Oil Process. The short-chain olefins are fractionally distilled, and are then reacted in a second stage with stoichiometric amounts of aluminum trialkyls. Ethyl Corporation operates a plant based on this process in Pasadena, Texas.

4. Ziegler Alcohol Synthesis [3], [4]

In the production of alcohols by the Ziegler route, the aluminum trialkyls formed are oxidized by atmospheric oxygen to form aluminum trialkoxides, from which the long-chain alcohols are then obtained by hydrolysis (→ Fatty Alcohols, Chap. 2.4.1.):

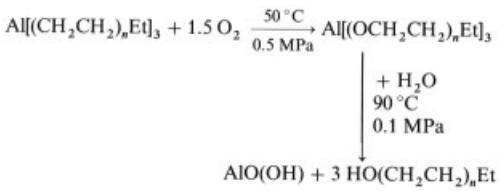

Alfol Process. The first industrial realization of the Ziegler alcohol synthesis was by Conoco. In 1962, this company built an "Alfol" plant (now operated by Vista Chemical) at Lake Charles, United States. A similar plant was then constructed in 1964 by Condea-Chemie in Brunsbüttel, Germany. The products of particular interest are the alcohols of chain lengths C_{12}, C_{14}, and C_{16}, which can be sulfated to form fatty alcohol sulfates with good biodegradability. However, as in the Gulf Oil Process for Ziegler olefins, the Alfol Process also produces a mixture of products with a broad Poisson distribution. The coproduct aluminum oxides (→ Aluminum Oxide, Chap. 6.2.3.) have important industrial applications as starting materials for catalysts and ceramics.

Epal Process. The Epal process is a development of the Ethyl Corporation, and is based on the principle of the Ethyl olefin process (see Chap. 3). Growth reactions are followed by fractional distillation, enabling an alcohol mixture with a considerably narrower distribution of homologues to be obtained, e.g., in the desired $C_{12} - C_{16}$ range. The hydrolysis after the oxidation step is carried out with sulfuric acid, so that high-purity aluminum sulfate is obtained as a byproduct.

The Ethyl Corporation plant in Pasadena, Texas can produce alcohols as well as olefins. However, this flexibility requires a very complex plant, so that capital and operating costs are high.

5. Ziegler Catalysts

Catalyst Systems. The terms Ziegler catalysts and Ziegler – Natta catalysts are used to denote the general range of catalyst systems formed from the following building blocks:

1) An organometallic compound of a metal of groups 1, 2, or 13 (e.g., an aluminum trialkyl)
2) A compound of a transition metal from groups 4 – 8 (e.g., titanium tetrachloride)

These two components are mixed in an inert solvent and form polynuclear complexes. Even today, the structure and function of the various Ziegler catalysts are the subject of controversy in the literature. In a fundamental experiment with bis(cyclopentadienyl)titanium dichloride and diethylaluminum chloride it was

possible to isolate a binuclear complex with a vacant coordination site on the titanium:

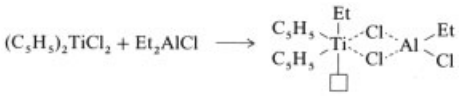

It is assumed that unsaturated molecules such as ethylene are bound to this coordination site and then inserted into the titanium – ethyl bond. This insertion produces another vacant coordination site, and the process can begin again. If this chain mechanism is broken, oligomers are formed, but if the insertion reaction is repeated indefinitely, long-chain polymers are obtained.

Oligomerization with Ziegler Catalysts. The first oligomerization with a Ziegler catalyst system was discovered in 1952 by HOLZKAMP at the Max Planck Institut für Kohlenforschung [5], [6]. A growth reaction with aluminum alkyls and ethylene (see Chap. 3) was carried out with addition of nickel salts. The Ni – Al catalyst system led to exclusive formation of ethylene dimers (i.e., butenes) instead of a broad distribution of α-olefins. This "nickel effect" marked the birth of the Ziegler catalysts.

Ziegler catalysts based on titanium, zirconium, chromium, tungsten, cobalt, and palladium compounds can also be used for the oligomerization of ethylene [7]. In most cases, butenes are the main products. Higher olefins can be obtained by suitable choice of catalyst and reaction conditions.

Both propene and butene can be oligomerized with Ziegler catalysts. In the Dimersol process of the Institut Français de Pétrole (IFP) [8–10], propene and butene are dimerized or codimerized. The products are used as octane enhancers in gasoline or in the production of plasticizers (→ Hydrocarbons).

1,3-Dienes such as butadiene can also be oligomerized with Ziegler catalysts. Ziegler catalysts based on nickel or titanium dimerize butadiene to vinylcyclohexene and 1,5-cyclooctadiene, or trimerize it, forming 1,5,9-cyclododecatriene. The course of this reaction can be controlled by the addition of ligands [5], [6].

Polymerizations with Ziegler Catalysts [11–16]. By far the most important application of Ziegler catalysts is in the synthesis of polyolefins. The first catalysts for the polymerization of ethylene were based on β-titanium trichloride, which was produced in situ from $TiCl_4$ and aluminum ethyl chlorides. Since 1969, $MgCl_2$-supported Ziegler catalysts, which possess considerably greater activity, have been used (→ Polyolefins, Chap. 1.3.2.2., → Polyolefins, Chap. 2.1.).

In March 1954, in the Politecnico di Milano, NATTA first used the Ziegler catalyst $TiCl_4/AlEt_3$ with propene. Surprisingly, it was found that this reaction is stereoselective, and that crystalline isotactic polypropylene is formed. The first polypropylene plant was commissioned as early as 1957 by Montecatini in Ferrara.

A further important development in Ziegler catalysts was the use of metallocenes activated by methylaluminoxane (MAO) [17–21]. The hypothesis that the catalytically active species is an ion pair consisting of a metallocene cation and a stabilized methylaluminoxane anion is currently under discussion. These catalyst systems were discovered in the 1970s by KAMINSKY and SINN at the University of Hamburg, and very soon extremely high activities (e.g., 40 000 kg polyethene per gram metal per hour) were being observed.

Ziegler catalysts based on metallocene/MAO led to a new breakthrough in the synthesis of polypropylene in 1985. The use of bridged bis(indenyl)zirconocenes, developed by BRINTZINGER at the University of Constance, enabled isotactic polypropylene to be produced with extremely high catalyst activities and stereoselectivities.

Work is in progress on many other developments worldwide, and many polymerization plants will be put into operation based on new catalyst systems in the coming years.

6. References

General References

1. G. Fink, R. Mülhaupt, H. H. Brintzinger (eds.): *Ziegler Catalysts,* Springer, Berlin 1995.

Specific References

2. K. Ziegler et al., *Angew. Chem.* **64** (1952) 323; **67** (1955) 542.

3. P. H. Washecheck in E. J. Wickson (ed.): "Monohydric Alcohols," *ACS Symp. Ser.* **159** (1981) 87.
4. H. D. Komp, H. P. Kubersky: *Fatty Alcohols,* 1st ed., Henkel KGaA, Düsseldorf 1982, p. 49.
5. G. Wilke, *Angew. Chem.* **100** (1988) 190; *Angew. Chem. Int. Ed. Engl.* **27** (1988) 185.
6. W. Keim, *Angew. Chem.* **102** (1990) 251 – 260.
7. W. Keim, A. Behr, M. Röper in G. Wilkinson (ed.): *Comprehensive Organometallic Chemistry,* **vol. 8,** Pergamon Press, Oxford 1982, p. 371.
8. J. Leonard, J. F. Gaillard, *Hydrocarbon Process.* 1981, 99 – 100.
9. Y. Chauvin, A. Hennico, G. Leger, J. L. Nocca, *Erdöl Erdgas Kohle* **106** (1990) 309 – 315.
10. A. Convers, D. Commereuc, B. Torck, *Rev. Inst. Fr. Pet.* **49** (1994) no. 5, 437 – 451.
11. E. J. Vandenberg, J. C. Salamone (eds.): "Catalysis in Polymer Synthesis," *ACS Symp. Ser.* **496** (1992) 1.
12. U. Zucchini, *Makromol. Chem. Macromol. Symp.* **66** (1993) 25 – 42.
13. K. Soga, *Makromol. Chem. Macromol. Symp.* **66** (1993) 43 – 54.
14. M. Ystenes, *Makromol. Chem. Macromol. Symp.* **66** (1993) 71 – 82.
15. G. Guerra et al., *Makromol. Chem. Macromol. Symp.* **69** (1993) 237 – 246.
16. T. Simonazzi, U. Giannini, *Gazz. Chim. Ital.* **124** (1994) 533 – 541 and references cited therein.
17. K. B. Sinclair, R. B. Wilson, *Chem. Ind. (London)* 1994, Nov. 7, 857 – 862.
18. H. Sinn, W. Kaminsky, *Adv. Organomet. Chem.* **18** (1980) 99.
19. W. Kaminsky, *Catal. Today* **20** (1994) 257 – 271.
20. W. Kaminsky, *Angew. Makromol. Chem.* **223** (1994) 101 – 120.
21. P. C. Möhring, N. J. Coville, *J. Organometal. Chem.* **479** (1994) 1 – 29.

Ziegler-Natta Catalysts → Polyolefins

Zinc

Günter G. Graf, Freiberg/Sachsen, Federal Republic of Germany

1. History 1
2. Properties 3
3. Starting Materials for Zinc Production 3
3.1. Zinc Minerals 3
3.2. Zinc Reserves 4
3.3. Secondary Raw Materials 5
4. Pretreatment of Raw Materials for the Production of Zinc Metal 5
4.1. Roasting 6
4.2. Pyrometallurgical Concentration Processes 7
5. Production of Zinc Metal 8
5.1. Pyrometallurgical Zinc Production by the Retort Process 8
5.2. Zinc Production in the Shaft Furnace 10
5.3. Refining of Crude Zinc 11
5.4. Hydrometallurgical Zinc Production 12
5.5. Treatment of Leaching Residues . . . 15
5.6. Liquor Purification 15
5.7. Electrolytic Production of Zinc 18
5.8. New Developments in Zinc Production 19
6. Uses and Economic Aspects 19
6.1. Galvanizing 19
6.2. Zinc as a Construction Material 20
6.3. Pressure Die Casting with Zinc 21
6.4. Zinc as a Chemical 21
6.5. Economic Aspects 22
7. References 22

1. History [1], [2], [4], [6–8]

Before 1500 A.D. the Chinese and Indians knew of zinc as a metal in its own right, and although it is not known precisely when zinc was first produced, coins containing 99 % zinc were cast during the Ming Dynasty (1368 – 1644). In a book written in 1637, the Chinese describe zinc production by controlled smelting in stacked, hermetically sealed pots. In India, remains of zinc smelting operations found in Sawai Madhopur (Rajasthan) confirm that zinc was produced prior to the 1300s. Although the Romans, the Greeks, and the entire western world knew how to produce brass, which was used as early as 8000 B.C. to make jewelry, metallic zinc was unknown. At that time, brass was produced by smelting copper ores or metallic copper with the zinc ore smithsonite. This method lasted until the 1800s. Brass smelting works were developed in the Middle Ages where localized deposits of smithsonite occurred in various places, e.g., near Lüttich and Aachen, and in Carinthia.

The efforts of the alchemists to convert base metals into gold led them to become interested in smithsonite, the ore that gave copper a golden color.

The first written sources to mention the word "zinc" appeared in the 1500s. However, the derivation is not entirely clear. The most likely assumption is that it is connected with the German word *Zinke* (the tooth of a comb), itself derived from *Zacke* (jagged peak), from the jagged appearance of the ore or of deposits in the furnace. A second possibility is that the word is derived from the Persian word *seng* meaning stone. Georgius Agricola (1494 – 1555), in his important *De re metallica*, describes how a material found in large quantities in the eastern Alps was called "zinc" by the local people. He gave the name "conterfey" to the white substance that was deposited at the front of the smelting furnaces during the treatment of Rammelsberg ores. This often consisted of metallic zinc, although he did not recognize it as such.

The metallurgical properties of zinc were first unambiguously described in Europe when zinc arrived there from India and China after the establishment of trade relations with the Far East.

In 1595, Libavius investigated a sample of this East Indian zinc imported from Holland and named "calaem." However, he confused it with silver and did not realize that it was the same material as that known as "cadmia" which was scraped off the walls of the furnace when Rammelsberg ore was smelted in the Harz Mountains to obtain lead and silver.

In 1617, Löhneyss also described this process and clearly recognized this metal as "zinc" or the "conterfey" already described by Agri-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_509

COLA, used to imitate gold. KUNKEL (1700) and STAHL (1718) considered that smithsonite added a metallic component to the copper during brass production. In 1721, HENKEL produced zinc from smithsonite in Upper Silesia using a process that he kept secret. In 1746, in Berlin, MARGGRAF produced zinc by distilling smithsonite in sealed vessels. He described his production principle in detail, thereby establishing the basic theory of zinc production. The first industrial production of zinc was carried out in England, where it was produced on a large scale as early as 1720, probably based on knowledge of operating methods used in East Asia. CHAMPION patented a process adapted from brass production to produce zinc by distillation and built the first zinc smelting works in Bristol in 1743, with a capacity of >200 t/a zinc. He also patented the process for producing zinc from sulfidic ores after preliminary roasting. The English zinc industry is concentrated in Bristol and Swansea even today.

The German zinc industry was founded by RUBERG. The first zinc smelting works, which used the horizontal retort process developed by him, was built in Wessola in Upper Silesia in 1798. The raw material initially used was zinc galmei, a byproduct of lead and silver production. Later, it became possible to produce zinc directly from smithsonite, an easily smelted ore. This was shortly followed by the use of zinc blende, which had first to be converted into the oxidic form by roasting. After this development, other smelting works were soon erected in Silesia near the deposits, in the areas around Lüttich, Belgium and Aachen, Germany and in the Rhineland and Ruhr regions.

In the United States, the rich ore deposits led to rapid growth in zinc production in 1840, so that, by 1907, Germany, which had for long been the world's leading producer of zinc, was left behind.

The ores were often highly intergrown, and ore processing by sorting and flotation methods was started in the early 1920s.

At this time, the electrolytic production of high-purity zinc from sulfuric acid solutions was developed by the Consolidated Mining and Smelting Co. in Trail (Canada) and by the Anaconda Copper Co. in Anaconda, Montana (United States).

An important development in zinc production was achieved in 1930 at the New Jersey Co. in Palmerton (United States). Replacing the individual horizontal retorts by a single vertical retort enabled continuous operation and also yielded relatively pure zinc (ca. 99.5 %).

Statistical data on world production of zinc have been available since 1871. At this time, total production was 121 000 t/a, of which 58 000 t/a was produced in Germany and 45 000 t/a in Belgium. Total production at the turn of the century was 472 000 t/a and in 1913, 10^6 t/a. The earliest use for zinc was to produce brass for the manufacture of jewelry, domestic utensils, and coins.

Zinc castings were first produced on an industrial scale in the 1800s by GEISS (1805 – 1875). Backed by the building contractor SCHINKEL, the zinc foundry established by GEISS soon became of major importance.

The good resistance of zinc towards atmospheric corrosion soon led to its use in sheet steel production. The possibility of rolling zinc at 100 – 150 °C was discovered as early as 1805, and the first rolling mill was built in Belgium in 1812 by DONY. More such mills were built in Silesia from 1821 onwards.

Hot-dip galvanizing, the oldest anticorrosion process, was introduced in 1836 in France. This became possible on an industrial scale only after the development of effective processes for cleaning iron and steel surfaces. At first, only small workpieces were zinc coated. Continuous hot-dip galvanizing of semifinished products and wire came later.

2. Properties [1], [2], [7–11]

Physical Properties. Some physical properties of zinc are listed below. (see below)

The remarkably low boiling point of 906 °C (STP) is important in pyrometallurgical zinc production processes. Zinc can be worked at temperatures as low as 120 °C, and its good deformability is the basis for some of its uses.

In the electrochemical series, zinc is more negative than iron. The use of zinc in sacrificial anodes is based on this fact. The important use of zinc to coat iron to protect it from corrosion is based on the formation of passivating protective coatings of basic zinc carbonate by reaction with the atmosphere.

Relative atomic mass	65.39
Crystal structure	hexagonal
Density	
at 20 °C	7.14 g/cm^3
at *mp*	6.83 g/cm^3
Density of liquid zinc	
at *mp*	6.65 g/cm^3
at 800 °C	6.25 g/cm^3
mp	419.5 °C
bp	906 °C
Heat of fusion	7387 J/mol
Heat of vaporization	114.8 kJ/mol
Thermal conductivity	
solid	
at 0 °C	116 W m^{-1} K^{-1}
at *mp*	96 W m^{-1} K^{-1}
liquid	
at *mp*	60.7 W m^{-1} K^{-1}
at 750 °C	56.5 W m^{-1} K^{-1}
Specific heat capacity	
at 0 °C	0.381 J g^{-1} K^{-1}
at 20 °C	0.385 J g^{-1} K^{-1}
at 100 °C	0.398 J g^{-1} K^{-1}
at 300 °C	0.420 J g^{-1} K^{-1}
Electrical conductivity	
at 0 °C	18.1 mΩ^{-1} mm^{-1}
at 25 °C	16.82 mΩ^{-1} mm^{-1}
at 100 °C	12.17 mΩ^{-1} mm^{-1}
Superconductivity critical temperature	0.905 K

Chemical Properties. Zinc has an oxidation state of 2 in all its compounds. It forms complexes with ammonia, amines, and cyanide and halide ions. Zinc dissolves in mineral acids, usually with evolution of hydrogen, but in nitric acid with evolution of NO_x. Zinc is used as a strong reducing agent in chemical processes, mainly in the form of powder or granules. Zinc is resistant to air because of the protective coating formed. It is also resistant to halogens, but is rapidly corroded by HCl gas. Because of its high surface area, zinc dust is much more reactive and can even be pyrophoric, e.g., reacting vigorously at elevated temperatures with the elements oxygen, chlorine, and sulfur.

Zinc chloride and sulfate are very soluble in water, while the oxide, carbonate, phosphate, silicate, and organic complexes are insoluble or sparingly soluble.

3. Starting Materials for Zinc Production [1], [2], [4], [12–15]

3.1. Zinc Minerals

With an average concentration in the earth's crust of 65 g/t (0.0065 %), zinc is the 24th element in order of abundance. It occurs only in the chemically combined state. Like copper and lead, zinc is a strongly chalcophilic element. It usually occurs as the sulfide, mostly deposited from hydrothermal solutions and simic and sialic magmas. Sedimentary deposits are of less importance. A series of deposits in carbonatic strata shows the characteristics of true marine sedimentary deposits. However, in these cases it can be assumed that the metal was introduced by hydrohermal transport from simic magmas. Zinc deposits, classified according to their origin, are listed here in order of their economic importance:

1) Simic vulcanogenic-sedimentary deposits that sometimes occur as massive ore bodies and sometimes as impregnations, e.g., Kidd Creek (Canada), Mt. Isa and Broken Hill (Australia), Anvil Mine (Canada), Bleiberg (Austria), Reocin (Spain), and Rammelsberg and Meggen (Germany). Ores from these deposits represent 60 % of total world reserves.
2) Simic-hydrothermal or marine sedimentary deposits, e.g., Kipishi (Zaire), Kabwe (Zambia), Tsumed (South-West Africa), and New Zinc Belt and Old and New Lead Belt (United States). Approximately 30 % of all ore is produced from these deposits.
3) Sialic lodes (impregnation and replacement deposits), e.g., Bad Grund (Germany); western U.S. states such as Idaho, Colorado, and Utah; and Tepca (Jugoslavia). Only 10 % of world ore production is from these deposits.

Cadmium, silver, lead, and copper are important byproducts of zinc ore extraction. Isomorphous components of zinc blende such as cadmium, gallium, germanium, indium, and thallium do not form deposits of their own. Some of these metals are recovered during zinc ore processing.

By far the most important zinc mineral today, and probably also the only primary ore, is zinc blende. The other minerals listed here are of only local importance. Native zinc is practically nonexistent.

Zinc blende, sphalerite, ZnS, theoretical composition 67.09 % Zn, 32.91 % S, is light yellow to black in color, usually black to brown depending on the iron content, crystallizes in the regular tetrahedral system, and has a Mohs hardness of 3 – 4 and a density of 3.9 – 4.1 g/cm^3.

The most important impurity is FeS, which is always present at concentrations between 0.3 and >20 %. The grades with the highest iron contents are marmatite and christophite. Sulfides of Pb, Cd, Mn, and Cu are often present. Zinc blende also often contains small amounts of As, Sn, Bi, Co, Hg, In, Tl, Ga, and Ge, and nearly always contains Ag and often Au.

Wurzite is the hexagonal modification of ZnS. It often shows a conchoidal or radiating crystalline structure and is therefore also known as conchoidal or radiating blende. Its density is 3.98 g/cm^3.

Smithsonite, zinc spar, or calamine, $ZnCO_3$, has the theoretical composition 52.14 % Zn, 35.10 % CO_2. Some of the ZnO can be replaced by FeO, MnO, CaO, or MgO, so that the ZnO content can be as low as 20 %. Other typical impurities include SiO_2, Fe_2O_3, and Al_2O_3.

The Mohs hardness is 5 and the density 4.30 – 4.45 g/cm^3. The color is yellowish or reddish to brown, rarely colorless. Smithsonite crystallizes in the trigonal scalenohedral form and is usually amorphous or in amorphous mixtures with other carbonates. Minerals with higher impurity contents include manganese zinc spar (up to 15 % $MnCO_3$), iron zinc spar (up to 53 % $FeCO_3$), copper zinc spar (up to 6 % $CuCO_3$), zinc lead spar (up to 1 % PbO), and cadmium zinc spar (up to 7 % $CdCO_3$).

Hemimorphite, $Zn_2SiO_4 \cdot H_2O$, has the theoretical composition 54.30 % Zn, 24.94 % SiO_2, and 7.4 % water of crystallization of which 50 % is evolved on heating to 500 °C, the remainder at much higher temperatures.

The color is yellow to brown, more rarely colorless, green, or blue. It forms rhombic pyramidal crystals, but usually occurs in rough, massive, botryoidal, or earthy forms. Its Mohs hardness is 4 – 5 and its density 3.4 – 3.5 g/cm^3. It is often a component of "calamine" ores.

Willemite, Zn_2SiO_4, has the theoretical composition 58.68 % Zn, 26.95 % SiO_2. Its color can be white, yellow to reddish, greenish, or rarely blue. It forms rhombohedral crystals, has a Mohs hardness of 5 – 6 and a density of 4.0 – 4.2 g/cm^3, and can contain FeO and MnO as impurities.

Franklinite, $(Zn, Fe, Mn)O \cdot (Fe_2, Mn_2)O_3$ has the theoretical composition 21 % ZnO, 59 % Fe_2O_3, 11 % FeO, and 8 % Mn_2O_3, and can contain SiO_2, Al_2O_3, and MgO as impurities. Its color is black, it is weakly magnetic, and it forms cubic crystals, with Mohs hardness 6.0 – 6.5, and density 5.0 – 5.1 g/cm^3.

Other zinc minerals of minor metallurgical importance include zinc bloom or hydrozincite, $ZnCO_3 \cdot n\,Zn(OH)_2$; red zinc ore or zincite, ZnO; and troostite or bainite, $(Zn, Mn)_2SiO_4$.

3.2. Zinc Reserves

The proven zinc mineral reserves worldwide are estimated at 185.3×10^6 t zinc. Potential reserves amount to 118.1×10^6 t. The undoubtedly considerable Chinese deposits are not included in this estimate, as no concrete information on these is available.

The proven and potential reserves are listed in Table 1. Canada, with 37.7×10^6 t (20.3 %), has the largest reserves of any country, over 90 % of which are in stratified vulcanogenic-sedimentary deposits. North and South America account for over 45.4 % of known deposits. Western Europe has 16.7 % of world zinc reserves, of which the most important are in Ireland, Sweden, and Yugoslavia (Kosovo). Australia accounts for over 9.5 % of world reserves, the largest deposits being in Mt. Isa (Queensland) and Broken Hill (New South Wales).

Table 1. The zinc reserves of the world [2], [3], [12], [13]

Country	Proven reserves		Potential reserves	
	10^6 t	%	10^6 t	%
Canada	37 560	20.3	6 620	5.6
United States	31 800	17.1	32 200	27.3
Mexico	3 600	1.9	1 000	0.8
Peru	7 150	3.8	1 500	1.3
Brazil	1 790	1.0	480	0.4
Rest of Latin America	2 360	1.3	2 820	2.4
Africa	5 930	3.2	9 580	8.1
Germany	1 850	1.0	1 500	1.3
Sweden	4 000	2.2	2 000	1.7
Finland	2 000	1.1	1 000	0.8
Ireland	8 820	4.8	350	0.3
Spain	5 060	2.7	1 200	1.0
Italy	2 200	1.2	1 100	0.9
Yugoslavia (Kosovo)	3 500	1.9	2 000	1.7
Rest of Western Europe	3 420	1.8	2 250	1.9
Australia	17 560	9.5	27 950	23.7
Iran	3 700	2.0	1 800	1.5
India	2 650	1.4	3 500	3.0
Japan	9 150	5.0	6 100	5.2
Rest of Asia	1 160	0.6	3 100	2.6
GUS and Eastern Europe	30 000	16.2	10 000	8.5
Total world reserves	185 260	100.0	118 050	100.0

The most important deposits in the GUS, accounting for 13.2 % of the total, are in the central and eastern Kazak, Uzbek, and Kirghiz. The reserves in Poland (Upper Silesia) are also important.

3.3. Secondary Raw Materials [16], [17]

Typical secondary zinc raw materials include:

1) Electric steel plant dusts containing 20 – 30 % zinc (low grade) or 30 – 40 % zinc (high grade)
2) Liquor residues from zinc electrolysis containing 25 – 30 % zinc
3) Lead shaft furnace slags containing 10 – 15 % zinc
4) Residues from thermal zinc production containing 8 – 12 % zinc

Thermal enrichment of such products in the Waelz kiln process gives a Waelz oxide containing 50 – 58 % zinc and 8 – 12 % lead. The lead content can be reduced by calcination. Other sources of secondary raw materials include zinc dross, which is produced in galvanizing plants and has a relatively high zinc content (50 – 85 %); zinc-containing sludges (10 – 65 % zinc), spent catalysts (50 – 60 % zinc), and dusts from the brass industry (60 – 75 % zinc).

The increasing problems in disposing of zinc-containing materials are resulting in the continual extension of the range of zinc-containing secondary raw materials.

4. Pretreatment of Raw Materials for the Production of Zinc Metal [1–4], [10], [18–23]

Pyrometallurgical and electrolytic zinc production require an oxidic raw material. Sulfidic concentrates must be prepared for metal extraction by roasting, which removes the sulfur as SO_2 and converts ZnS to ZnO, but does not remove gangue material.

Low-zinc oxidic ores that are not suitable for mechanical beneficiation (e.g., by flotation), slag flue dusts, and other secondary raw materials are subjected to a thermal concentration stage, which usually consists of either slag fuming or roasting in a Waelz kiln.

4.1. Roasting

Both electrolytic and pyrometallurgical zinc production require oxidic raw materials. Sulfidic raw materials must therefore be converted to the oxidic form (known as calcine) before leaching or reduction.

Whereas early accounts of zinc metallurgy listed a large number of roasting processes, fluidized-bed and sinter roasting have now almost completely replaced older processes such as multiple-hearth and flash roasting. An essential requirement of any roasting process is that it must be possible to recover and use the SO_2 from the waste gas.

The various zinc production processes require feed materials with fundamentally different physical properties. A fine-grained calcine is necessary for leaching with dilute sulfuric acid. Flotation concentrates meet this requirement if oxidized by roasting in a fluidized bed. On the other hand, thermal processes require a porous feed material in lump form, as provided by sinter roasting [18].

Fluidized-Bed Roasting. Given a processed concentrate of suitable chemical constitution and particle size, the fluidized-bed reactor allows a calcine suitable for the leaching process to be produced. The parameters to be optimized are bed height, gas velocity, residence time, and temperature. In the fluidized-bed reactor, the reaction conditions are easily controlled and stabilized. The theoretical relationships, as shown in the phase diagram Zn – Fe – S – O (Fig. 1), have been thoroughly investigated [19], [22], [23]. Fluidized-bed roasters generally operate at conditions in which ZnO and Fe_2O_3 are stable. Developments in electrolytic processing technology have enabled the presence of sulfates and ferrites in the calcine to be tolerated.

Fluidized-bed furnaces are often operated in two stages, in the first of which the roasted product contains 6 – 8 % S. This overflows while hot into a second fluidized bed and is roasted with an excess of air to a residual sulfur content of <1.5 % [3, p. 662].

Sinter Roasting. The first industrial application of sinter roasting was in the roasting of lead ore in sintering pans. The crucial development in

processing technology came from the sintering of iron ores, the results from which are transferable to the sinter roasting of sulfidic concentrates [21]. It is important that the temperature of the mixture should reach the sintering point. This is determined by the fuel content of the mixture (i.e., in single-stage sinter roasting, by the ratio of concentrate to roasted recycled material), the particle size distribution of the mixture, and the flow resistance (controllable by means of the particle size), which in turn determines the air pressure difference both in the induced-draft and pressure sintering processes. The moisture content of the mixture is an important factor here. It is important that all parameters affecting the process should be kept as constant as possible. Automatic control is not yet widely used.

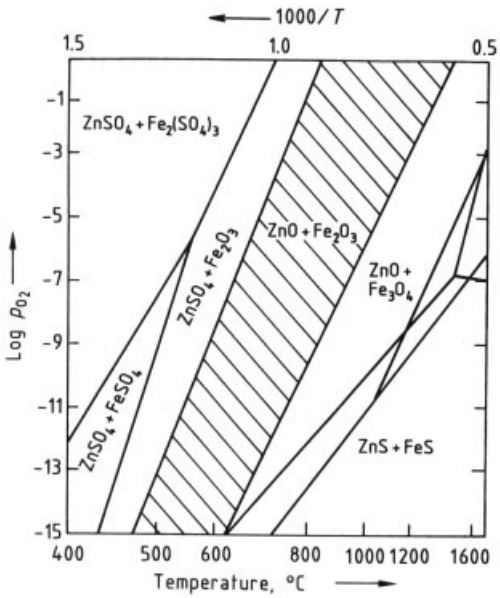

Figure 1. Phase diagram of the system Zn – Fe – S – O [23] The conditions that exist in most fluidized-bed roasters are indicated by the hatched area

In contrast to fluidized-bed roasting, where maximum reaction temperatures are in the range 900 – 950 °C and no volatilization occurs, maximum temperatures in the sintering zone reach 1450 °C, so that the cadmium and lead are volatilized to a considerable extent. For this reason, it is necessary to limit the maximum temperature in the sintering zone.

In principle, the temperatures in the reaction zone during sinter roasting are sufficient to ensure homogeneous sintering not only of concentrates but also of added flue dusts, dross, sludge, and other zinc-containing materials. An important advantage of sinter roasting is thus that it enables these materials to be prepared for shaft furnace treatment (Section 5.2) continuously and by the shortest route. The important characteristic parameters for sinter roasting, fluidized-bed roasting, and roasting in a multiple hearth furnace are compared in Table 2 [3, p. 662].

Table 2. Characteristic values for sinter and fluidized bed roasting compared with multiple hearth roasting , p. 662[3]

	Multiple hearth roaster	Sintering belt	Fluidized-bed roaster
Roasting capacity, $t\,S\,m^{-2}\,d^{-1}$	0.08	0.7	7.5
Throughput, t/d	60	500	800
Grain size, mm	1.5	5	3
Dust content, %	13 – 15	2 – 3	50
SO_2 in exhaust gas, %	4 – 5	10	10
Steam output, t/t ore	0.4		1.0
Floor area, $m^3\,t\,d$	7	10	3

4.2. Pyrometallurgical Concentration Processes

The aim of pyrometallurgical concentration processes is usually the removal of zinc from flue dusts, slags, and low-grade, zinc-containing raw materials or raw materials of complex composition.

The Waelz kiln process was originally used for the treatment of low-grade oxidic ores to obtain oxide suitable for smelting [24]. It is now mainly used as an enrichment process to treat secondary raw materials from zinc metallurgy, e.g., slags, flue dusts, and sludges.

In the Waelz kiln process, the zinc and lead from the raw materials appear in the oxide product, together with much of the sulfur, chlorine, and fluorine, both from the raw materials and from the carbonaceous reductant. The zinc metal is usually obtained from the oxide product by a thermal process. For this, the oxide can be compacted by sintering or forming it into briquettes. If oxide with a low lead content is required, it must be calcined, i.e., reheated to 1100 – 1350 °C without addition of reducing agents, to volatilize the Pb as PbO. The

Waelz kiln process is carried out in a slightly inclined, slowly rotating, lined Waelz or rotary kiln through which air is passed at a controlled rate so that the mixture of raw material and coke breeze does not fuse together. The residues are discharged in a sintered but not molten state. Air from the furnace atmosphere diffuses to a limited extent into the reaction mixture which is carried up the walls of the rotating furnace, and a strongly reducing CO/CO_2 atmosphere is formed. This causes reduction not only of ZnO but also of zinc ferrites and silicates, and reoxidation to ZnO occurs after discharge. The following reactions take place in the rotating furnace:
Reaction in the solid charge:

$$ZnO + C \rightarrow Zn(g) + CO$$

Reaction in the gas space:

$$Zn(g) + CO + O_2 \rightarrow ZnO + CO_2$$

The heat liberated is due only to the overall reaction, i.e., the combustion of C to form CO_2, as the reduction of ZnO to Zn is immediately reversed in the gas space above the charge. The zinc oxide leaving the furnace space is collected in bag filters. As the sulfur and halogens in the waste gases are bound to the zinc oxide, there are no environmental problems. The iron compounds in the charge are reduced to wuestite and to some extent to iron metal. In practical operation, the composition of the reaction mixture must be correctly formulated and an excess of coke breeze must be present to prevent the iron-containing components from fusing and forming incrustations in the furnace. The temperature regime of the process can be controlled by means of the filling height, residence time, and air flow rate in the upper part of the furnace. Overheating can cause incrustations to form and must be prevented [24–27].

Slag Fuming. In contrast to the Waelz kiln method, with which a range of raw materials can be treated, the slag fuming process is only used to remove zinc from molten slags from lead and copper smelting [28]. The zinc oxide so obtained can be used for zinc production. The molten slag is blown with reducing gases, so producing zinc vapor, which is burnt in the gas space above the molten slag by secondary air. The resulting high-grade oxide is cooled in a waste heat boiler in an EGR and collected [3, p. 65]. The composition of the reaction products from slag fuming in Port Pirie (Australia) is given in Table 3.

Table 3. Composition (in wt %) of reaction products from a slag fuming plant in Port Pirie, Australia

	Zn	Pb	CaO	FeO	SiO_2	F	Cl	S
Slag before Zn removal	18.3	2.6	14.7	25.6	21.0	0.1		1.7
Slag after Zn removal	2.6	0.03						
Oxide	66.0	12.5				0.25		0.2

Oxide Calcination. The oxides obtained from Waelz kilns or slag fuming plants can be compacted in preparation for further processing by sintering or briquetting, without pretreatment. However, for electrolytic processing, for pigment use, or for production of other zinc compounds, removal of harmful elements such as Pb, Cl, and F is required. This is carried out by calcination in rotary kilns at 1100 – 1150 °C. Not only are the harmful impurities removed, but the material is also densified [29], [30].

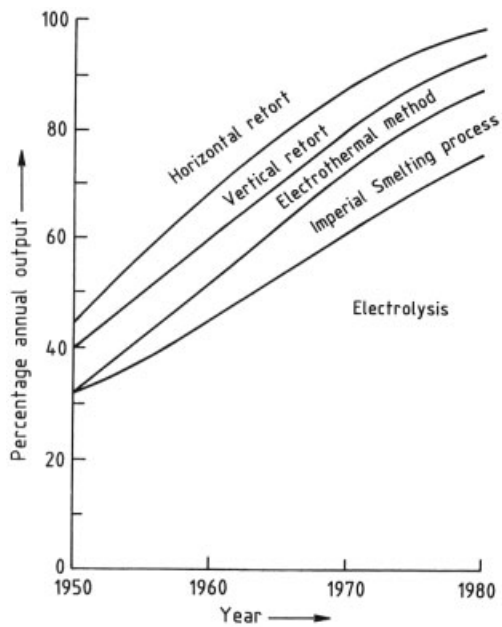

Figure 2. Annual outputs of the various zinc production processes expressed as a percentage of total annual output (idealized curves) [3, p. 664]

5. Production of Zinc Metal [1–5]

There are five industrial processes for the production of zinc: one electrolytic and four pyrometallurgical. Their relative importance during recent decades is illustrated in Figure 2 [3, p. 664]. Although the horizontal retort was the most important zinc production process at the turn of the century and still accounted for ca. 50 % of total production capacity in 1950, the process is no longer of practical importance. The question of whether the energy-intensive electrolytic zinc production process can retain or even extend its lead depends on the development of the energy supply situation.

5.1. Pyrometallurgical Zinc Production by the Retort Process

Zinc oxide, produced by roasting, is reduced by carbon according to the equation

$$ZnO + C \rightleftharpoons Zn + CO \quad \Delta H = 353\ kJ$$

The reaction is so strongly endothermic that 5.4 MJ of heat is required to produce 1 kg of zinc. The reaction starts at ca. 1100 °C and is complete at 1300 °C, and the zinc (*bp* 920 °C) is therefore produced as a vapor. Zinc production by the retort process hence involves not only reduction of the oxide but also condensation of the zinc from the vapor to the liquid state. Care must be taken to prevent reoxidation, and the zincvapor must not come into contact with the combustion gases. The reduction must therefore be carried out either in vessels with exterior heating (retorts) or by the electrothermal method. Horizontal, batch-operated retorts with circular or elliptical cross section can be used, or furnaces containing vertical, continuously operated retorts with a rectangular cross section. The quality of the furnace vessels is of great importance for the process, especially their stability at high temperature, resistance to large temperature fluctuations, good thermal conductivity, impermeability to reaction gases, and resistance towards chemical attack by the charge materials.

Reduction in Horizontal Retorts. Zinc has been produced in horizontal retorts since ca. 1800 in Belgium and Upper Silesia from smithsonite and later from roasted zinc blende. Retorts vary in their shape and arrangement in the furnace and in the fitting of receivers, in which the zinc vapor produced in the retort condenses as a liquid. Detachable devices made of sheet metal (so-called allongen) are fitted to the condensers to collect zinc escaping in the form of dust.

The retort furnaces are equipped with regenerators or recuperators which utilize the hot waste gases to preheat the combustion air and sometimes the fuel gas. The retorts used to be charged by hand, an extremely labor-intensive operation: one worker smelted ca. 0.5 t ore per shift. Moreover, the yield was poor, only 60 % of the zinc being obtained as molten castable metal, 10 % remaining unreacted in the ash produced, and the rest reporting to the dross and dust in equal amounts. Although efforts were made to mechanize the process, especially the charging and discharging of the retorts, the main problem, i.e., the poor efficiency (ca. 20 %), could not be solved. The horizontal retort is no longer an industrially important zinc production process.

Reduction in Vertical Retorts. The poor efficiency, batch operation, and above all labor intensity of zinc production in horizontal retorts provided the incentive to develop an improved process. The fact that the horizontal retort was inconsistent with the natural downward movement of the charge and upward movement of the zinc vapor suggested that a vertical arrangement would be better. Zinc reduction furnaces were therefore constructed with vertical retorts, but only after overcoming major difficulties. The system is known as the New Jersey process. The furnace, illustrated in Figure 3, consists of a retort 13.5 m high, 1.83 m long, and 0.3 m wide, constructed of silicon carbide. The retort is externally heated over a vertical length of 8 m (starting at 1.5 m from the lower end) to a temperature of 1250 – 1300 °C. Producer gas is fed from above into the heating chambers, and the exhaust gas phases out of the furnace at the bottom. The reaction air is fed through the sides of the furnace along its whole length. Above the heated part of the furnace is an unheated part 4 m in length which acts as a reflux condenser for the lead formed and is known as the lead collector. In the original design, the gas mixture passing out of the retort at 850 °C was cooled

to 500 °C by a plate cooler, which also condensed and collected the zinc. This was tapped off from time to time. The zinc dust formed was collected by means of a water spray. The process was much improved by replacing the plate cooler by a spray condenser. A graphite disc, partially immersed in the pool of molten metal and spinning at ca. 400 rpm, atomizes the metal so that the waste gases from the retort passes through a metal fog. The metal is collected at 500 °C in an exterior cooler. This prevents formation of dross, and the amount of zinc dust formed is reduced to 3 % of the zinc product.

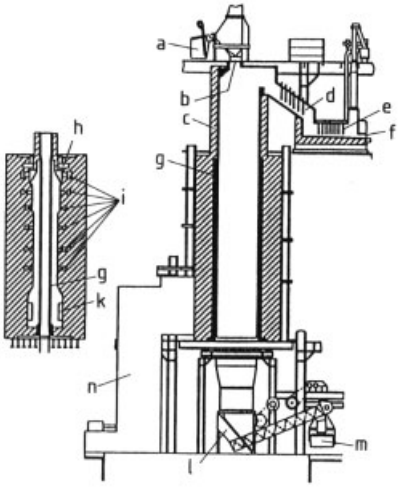

Figure 3. Schematic diagram of vertical retort [3, p. 666] New Jersey retort with plate condenser for zinc recovery- The left hand and right hand sections are rotated 90° with respect to each other.a) Charging bucket; b) Cone; c) Lead collector; d) Inclined channel; e) Condenser; f) Tap hole; g) Retort; h) Fuel gas inlet; i) Air inlets; k) Exhaust gas exit openings; l) Ash removal equipment; m) Ash bucket; n) Recuperator

The charge material, containing 60 – 70 % calcine and 30 – 40 % coal, is formed into briquettes, which are then coked in a coke oven. The coked briquettes are charged into the top of the retort. After extraction of the zinc from the briquettes, they are discharged from the bottom of the retort through a water lute. The zinc-containing gases discharged from the top of the retort are cooled to ca. 500 °C in a condenser, in which the molten zinc collects at the bottom. The tail gases produced, which contain ca. 85 % carbon monoxide, are scrubbed and recycled to the producer gas stream. Furnaces with vertical retorts are much more thermally efficient than those with horizontal retorts, and they have the further advantages of providing higher capacity, higher zinc purity, and longer working life of the retorts, and, above all, do not involve heavy manual labor or batch operation.

5.2. Zinc Production in the Shaft Furnace [31]

Prolonged and extensive development work in the 1950s, aimed at utilizing the improved thermal efficiency and higher reaction temperatures of the coke shaft furnace, led to the lead – zinc shaft furnace of the Imperial Smelting Corp. Thus, a process became available that could be used with raw materials of complex composition, such as mixed ores and secondary raw materials containing impurities that would cause severe problems in electrolytic zinc production.

In the shaft furnace, the heat for the reduction is provided by the oxidation of C to CO_2 and CO. Under the reaction conditions normally used in a shaft furnace, the zinc vapor formed would be reoxidized to ZnO by the CO_2 in the furnace gas. Intensive research on the equilibria enabled conditions to be found under which this reoxidation does not occur.

In normal shaft furnace operation, the best possible heat utilization is achieved by using countercurrent flow to heat the charge material. The furnaces are relatively tall, and furnace gas exit temperatures of ca. 200 °C are aimed for. In accordance with the Boudouard equilibrium, this results in higher CO_2 contents in the furnace gas. The zinc shaft furnace requires gas temperatures of 1000 °C. In order not to lower the furnace gas exit temperature by the addition of the cold charge material, the coke is heated to >800 °C and the calcine is added as hot sinter.

The shaft furnace obtains its heat of reaction from the combustion of coke in the reaction zone. This leads to large amounts of nitrogen in the furnace gas, lowering the partial pressure of the zinc. Therefore, the heat required is generated to a large extent outside the furnace. Thus, the combustion air is preheated to 700 – 950 °C, enabling the amount of furnace gas to be much reduced. The CO-rich furnace gas is cooled as it

passes from the furnace to the condenser. To prevent reoxidation of zinc to ZnO, a small amount of atmospheric oxygen can be added to avoid a temperature drop. This process step must be carefully controlled.

The zinc collects in a spray condenser. In contrast to the vertical retort process, lead is used as the condensing liquid instead of zinc. Dissolving zinc in lead greatly reduces its activity. As the solubility of zinc in lead is strongly temperature dependent, the zinc can be precipitated by cooling. In practice, a temperature of ca. 550 °C is maintained in the spray condenser. The metal flowing out of the condenser is cooled to 440 °C, causing the zinc to separate from the lead and float on the surface. It then flows over a weir and is collected. As the process does not take place under equilibrium conditions, the recovery of 1 t zinc does not require the theoretically calculated amount of lead (60 t), but ca. 400 t. The zinc is either cast as commercial grade ingots (98.5 % Zn) or refined to higher grades.

The first full-scale industrial furnace went into operation in Swansea in 1960. The cross-sectional area of the shaft was 17.2 m^2, and this remains the standard size even today.

On licensing its ISO process, Imperial Smelting required that licensees should compare their experiences of the process, and this has led to standardization of the major zinc-producing plants and to improvements in the plant performance. For example, the annual output of a furnace with a cross section of 17.2 m^2 has been increased from 35 000 t zinc and 17 000 t lead in 1963 to 80 000 t zinc and 40 000 t lead in 1990. At the same time the zinc yield was increased from 91.7 to 95.5 %, and the lead yield from 94 to 95 %. The Pb : Zn ratio in the charge material was increased from 0.49 to 0.6, further enhancing the main advantage of the process, i.e., its ability to treat ores of complex composition. In addition to Zn and Pb, many ores also contain Cu. Most of this is reduced to the metal and dissolved by the lead. Copper contents below 4 % give few problems, as the copper separates as dross on cooling the lead and can be removed in the form of a solid crust. Higher concentrations can give problems due to the limited solubility of copper in lead. Special techniques are then required. A schematic diagram of a lead – zinc shaft furnace is given in Figure 4.

5.3. Refining of Crude Zinc

Zinc produced hydrometallurgically (see Section 5.4) does not need to be refined. However, crude zinc produced by pyrometallurgy contains 0.7 – 3 % Pb, up to 0.2 % Fe, up to 0.3 % Cd, and sometimes small amounts of As, depending on the production method and the composition of the ore. The zinc used to produce pressure casting alloys should have a purity of 99.99 – 99.995 %, and galvanizing processes require a purity of 98.5 – 99.95 %, depending on the process used. Commercial grades of zinc cannot be produced simply by remelting and liquation or by treatment with fluxes. Simply cooling zinc melts until they approach the eutectics in the systems Zn – Pb or Zn – Fe does not produce the theoretical compositions 0.8 % Pb and 0.02 % Fe, respectively. The concentrations obtained in practice are 1.2 % Pb and 0.03 % Fe.

Hence, these processes are only suitable for producing zinc for the next stage of purification by distillation or for producing zinc dust. The refining of commercial zinc produced by a pyrometallurgical process (today usually the Imperial Smelting Shaft Furnace process) is carried out by distillation in high-temperature distillation columns constructed of ceramic-bonded SiC elements. The process was developed by the New Jersey Zinc Co., Palmerton (United States), and the first plant went into operation in 1932.

The distillation process makes use of the comparatively low boiling point of zinc (906 °C). However, as not only impurities with a high boiling point such as lead, copper, tin, and iron must be removed, but also cadmium, which boils at 762 °C, the refining process must be carried out in two stages. A plant based on the New Jersey process is illustrated in Figure 5. Molten crude zinc is fed into the top third of a distillation column, the lower part being heated to ca. 1100 °C. The upper part acts as a reflux cooler for the less volatile constituents, so that only Zn and Cd pass into the condenser. Because of the relatively small difference in boiling point between zinc and cadmium and the low cadmium content, a further very careful fractional distillation must be carried out. The column for this is heated to ca. 950 °C, and the Cd which distills off is collected as a dust with a high zinc content. Depending on the composition of the initial

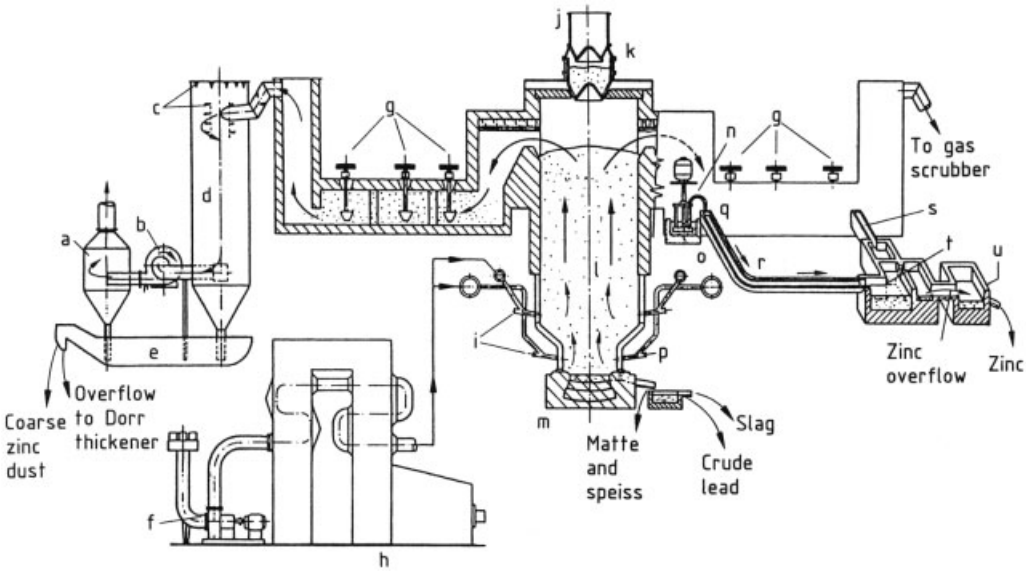

Figure 4. Zinc – lead shaft furnace
a) Moisture trap; b) Mechanical washer; c) Water jets; d) Gas scrubbing tower; e) Water seal for zinc dust removal; f) Blower; g) Stirrers; h) Air preheater; i) Jets; j) Feeding vessel for hot coke and calcine; k) Cone; l) Furnace; m) Hearth; n) Lead pump; o) Pump sump; p) Water jacket; q) Condenser; r) Water-cooled trough; s) Lead recycle trough; t) Separating wall; u) Zinc bath

materials and the quality of the process control, the zinc produced can have a purity of 99.996 % and the concentration of the undesired elements Pb, Cd, and Fe can be <0.001 %.

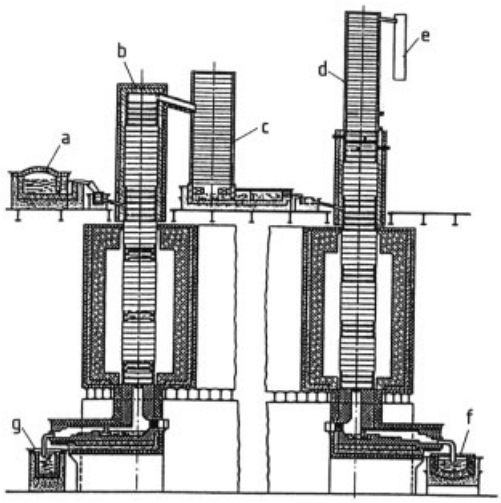

Figure 5. Zinc refining plant using the New Jersey process [3, Vol. 1a, p. 116, Fig. 34]
a) Melting furnace; b) Lead column; c) Condenser; d) Cadmium column; e) Cadmium canister; f) High purity zinc furnace; g) Washed inc furnace

The "washing" zinc produced in the lead column contains Pb, Fe, Cu, and sometimes also In and Tl. The washing zinc is liquated, giving hard zinc and very impure lead. The hard zinc is recycled to the process, but the lead can be treated separately. In practice, this very impure lead is often added to crude lead from a smelter before it is refined.

The working life of the distillation columns is 2 – 3 years, and the energy consumption for 1 t of purified zinc is 6.5 – 7.3 GJ.

5.4. Hydrometallurgical Zinc Production [1–5], [10], [13]

Zinc can be deposited electrolytically both from aqueous solutions and molten salts. However, only aqueous electrolytes (i.e., solutions of zinc sulfate in dilute sulfuric acid) have so far been used in industry. Other proposed processes such as the electrolytic production of zinc dust, amalgam electrolysis, and sulfate – chloride electrolysis have not attained industrial importance. Although a full-scale version of the electrolytic refining process, in which the impure zinc forms the anode, has been developed, it is not used, the

reasons being mainly economic. The electrolytic production (electrowinning) of zinc from solutions of zinc sulfate in sulfuric acid has undergone unparalleled growth in recent decades. Approximately 80 % of total world production comes from this process today, although a number of problems still need to be solved, principally in the preparation of pure electrolyte solutions.

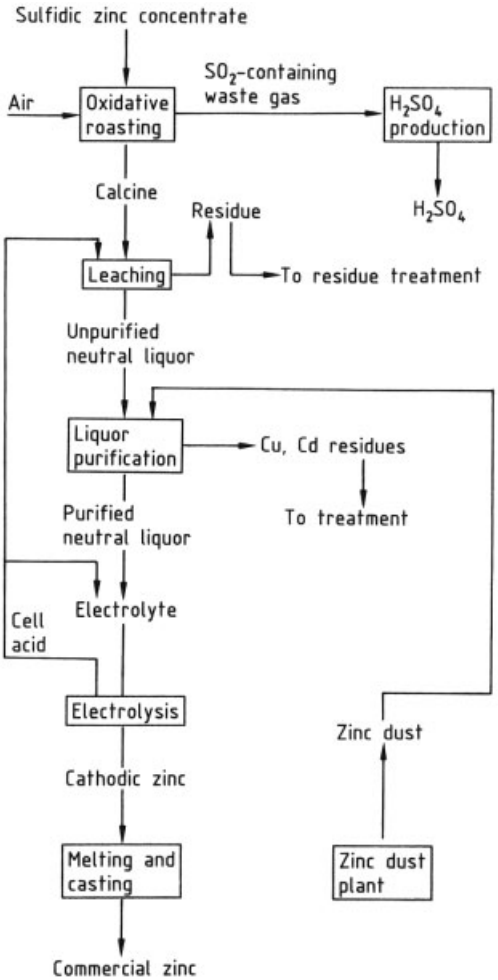

Figure 6. Flow diagram of the electrolytic zinc production process [2, p. 602]

Like the thermal processes of zinc production, the industrial electrolysis of zinc uses oxidic starting materials. The most important natural raw material, zinc blende, therefore still needs to be roasted to be converted to oxide. To ensure rapid dissolution of the calcine, the roasting processes used should therefore conserve the fine grained structure of the concentrate obtained by flotation. Commercial zinc oxide can usually be used directly and need only be treated if it contains impurities that have a harmful effect on the electrolytic process. Roasted oxide must be ground to improve its solubility. Using typical starting materials, the electrolytic process of zinc production consists of the following stages (Fig. 6):

1) Roasting
2) Leaching
3) Liquor purification
4) Electrolysis
5) Melting and casting.

The main problem in leaching and liquor purification is zinc – iron separation. In earlier processes, leaching was carried out such that a large excess of acid was avoided. Thus, the calcine and acid (exhausted electrolyte) were added simultaneously to the leaching tanks, the amount of acid being exactly that required to dissolve the zinc oxide. High yields of zinc could only be obtained with concentrates containing ≤ 3 % Fe. Iron in higher concentrations was dissolved to some extent. The solid residues contained ca. 20 % Zn and 30 % Fe. However, as a result of the roasting process, part of the zinc in the calcine is in the form of zinc ferrite ($ZnFe_2O_4$), which can only be dissolved at higher acid concentrations (50 – 150 g H_2SO_4/L). Maximum yields of zinc and also of other valuable impurities such as copper and cadmium require the use of higher acid concentrations in the leaching process. Hence, almost complete dissolution of the iron content of the zinc blende cannot be avoided. As the iron interferes with the electrolytic process even at low concentrations, it must be precipitated from the zinc sulfate solution obtained.

Precipitation of Iron Hydroxide. Precipitation of iron as its hydroxide has now been largely superseded owing to the poor zinc yield obtained (85 – 88 %). The process is usually carried out in two stages.

The readily dissolved zinc oxide fraction is brought into solution by "neutral leaching" in a mixture of cell acid (150 – 200 g H_2SO_4/L) and the acidic solution (ca. 50 g H_2SO_4/L) produced during iron precipitation. At this stage of the process, it should be ensured that impurities such as

iron, arsenic, antimony, and germanium are retained in the solid residues. A solid – liquid separation is carried out, giving an impure neutral liquor which is fed to the purification stage and a zinc-rich residue which is fed to the second, acid leaching stage, for which hot (ca. 95 °C) cell acid is used. A further solid – liquid separation produces a solid residue containing up to 5 % zinc and any lead and silver present (which can be processed in a lead smelter) and a zinc solution containing ca. 80 g/L zinc together with the iron and other impurities, such as arsenic, antimony, germanium, nickel, and cobalt.

After oxidation of Fe(II) to Fe(III) in the solution obtained by "neutral leaching," the iron is precipitated as its hydroxide by adding further calcine. The precipitation starts at pH 2.6 and continues up to pH 5.0. This gives solutions with residual iron contents of 10 – 20 mg/L. Because precipitation of iron causes arsenic, germanium, and zinc to be adsorptively coprecipitated, and the calcine used for neutralization cannot be subjected to hot acid leaching, this process gives unsatisfactory zinc yields. Higher zinc yields are only possible if the leaching residues can be more readily filtered. Three methods of achieving this are available, i.e., precipitation of the iron as jarosite, goethite, or hematite.

Iron Removal by the Jarosite Process. In the jarosite process, an Fe(III) compound of the type $X[Fe_3(SO_4)_2(OH)_6]$, where X represents H_3O^+, Na^+, K^+, or NH_4^+, is precipitated by adding alkali metal or ammonium ions. These compounds correspond to the mineral jarosite. The process was independently developed between 1960 and 1965 in various zinc smelting plants arount the world. It makes use of the fact that jarosite precipitates have the following advantages over iron hydroxide:

1) They are almost insoluble in sulfuric acid
2) They have a high iron content, so that the amount of precipitate formed is low
3) They have good filtering properties

However, jarosite precipitates do not adsorb the harmful impurities as effectively, so that the purification effect of the process is inferior.

The additional cations needed for the precipitation are added to the liquor in the form of sulfates. In practice, ammonium sulfate, which forms an ammonium jarosite, is usually used.

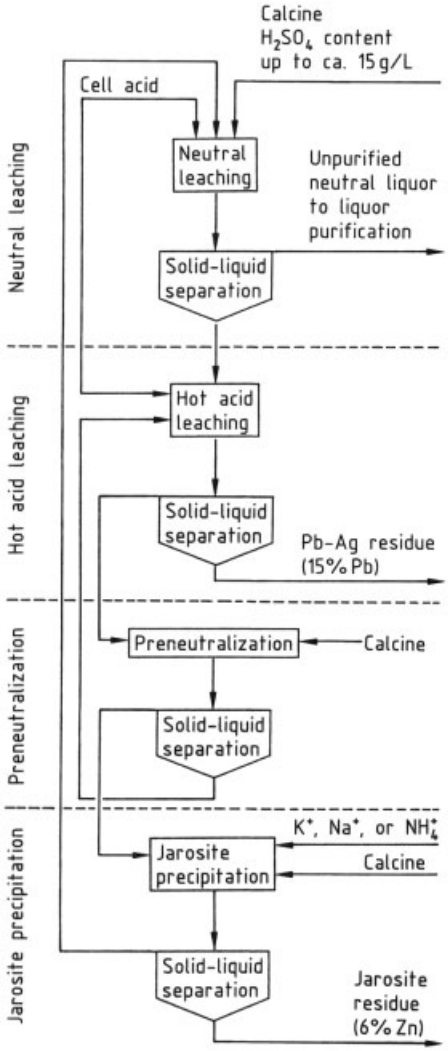

Figure 7. Flow diagram of the jarosite process with preneutralization [3, p. 678]

A flow diagram of the process is given in Figure 7.

Precipitation of jarosite begins at pH <1 and is complete at pH 1.5. Jarosite formation and precipitation being a temperature-dependent process, two alternative industrial processes exist, the first operating at temperatures up to 180 °C and with relatively high acid concentrations (60 – 90 g H_2SO_4/L) giving a high reaction rate, and the second at ca. 95 °C and considerably lower acid concentrations (5 – 10 g/L). The lower temperature method is more widely used, mainly because of the lower equipment costs.

Jarosite theoretically contains 37 % iron and 13 % sulfur as sulfate. In practice, these values are not reached, as the calcine added for pH control is not completely dissolved and hence adds impurities to the solid residue. For this reason, additional leaching processes are carried out, giving zinc contents of 4 – 6 % in the residue. Zinc yields of 96 – 98 % can be achieved by the jarosite process. As the jarosite contains sulfate, it helps regulate the sulfate balance of the plant.

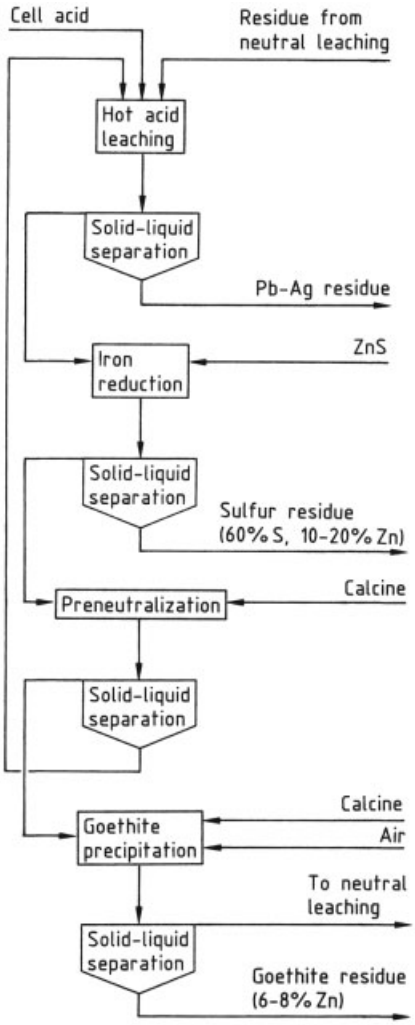

Figure 8. Flow diagram of the goethite process [3, p. 679]

Worldwide, the jarosite process has superseded the precipitation of iron hydroxide. However, problems are caused by the fact that the jarosite contains small amounts of soluble zinc and so must be deposited in special waste disposal sites. New processes aimed at further increasing the zinc yield have therefore been developed.

Iron Removal by the Goethite Process. As the flow diagram (Fig. 8) shows, the first stages of the goethite process are the same as in the jarosite process. The solution from the first hot leach is then treated with zinc concentrate. The zinc dissolves, reducing the iron to the divalent state, and elemental sulfur is precipitated, so that compounds other than goethite (e.g., jarosite) are not precipitated in the final precipitation process. The excess zinc concentrate and the elemental sulfur are recycled to the roasting process. Oxygen is added in a controlled manner, precipitating goethite, FeO(OH), at pH 2 – 3.5 and 70 – 90 °C.

The advantage of the goethite process is that the amount of residue that must be disposed of is considerably less than in the jarosite process. Whereas the zinc yield is comparable to the jarosite process, the yield of copper is lower (80 % instead of 90 %) and that of silver is higher (96 % instead of 90 %). As the goethite produced, which contains 40 – 45 % iron, always contains small amounts of basic sulfates, it cannot be used in blast furnaces.

Iron Removal by the Hematite Process. The hematite process (Fig. 9) was developed to enable iron-containing residues from zinc production to be disposed of at moderate cost and without ecological problems. It differs from the other processes in that the residues are subjected to reductive leaching in which the reducing agent is an excess of zinc concentrate. The iron is reduced with formation of elemental sulfur. The insoluble solids and the liquor are heated in autoclaves, causing the sulfur to melt and wet the excess of zinc blende, forming small pellets. These are cooled, classified, and fed to the roasting plant. A lead – silver residue which separates from the solution is sent to a lead smelting plant. The liquor is preneutralized with calcine, and the resulting precipitate is recycled to the hot reductive leaching stage. The copper is precipitated from the zinc – iron solution by addition of zinc (cementation), and the solution is then neutralized with calcium carbonate. This is followed by the most important stage of the process, pressure

treatment in an autoclave at 200 °C at an oxygen pressure of 10 – 15 bar. Under these conditions, iron is precipitated as Fe_2O_3 (hematite), which can be used in iron metallurgy. The zinc solution is recycled to the neutral leaching stage. The gypsum produced is dewatered and can then be used in the building materials industry. The yields of zinc and other valuable metals are comparable to those in the jarosite and goethite processes. As the iron can be produced in the form of a useful product, no landfill space is required. However, the costly and complex technology is restricting more widespread use of the process.

5.5. Treatment of Leaching Residues

Residues from calcine leaching are primarily subjected to pyrometallurgical treatment. In some plants of older design, the residues from the neutral leaching stage are treated in a Waelz kiln. This removes the lead, and the resulting sinter oxide is ground and recycled to the leaching process. In a development of the classical Waelz kiln process, vaporization is carried out in a Waelz kiln fitted with an external burner. A rotary flame-fired furnace of the Dörschel type can also be used in a batch process.

A combination of flotation with pyrometallurgical and hydrometallurgical processes (sulfating roasting followed by leaching) enables the valuable metals to be recovered from the leaching residues.

5.6. Liquor Purification

Impurities still present in the unpurified neutral liquors from the iron precipitation stage can lead to:

1) Lower current efficiency
2) The presence of impurities in the cathodically deposited zinc
3) Adverse effects on the anode and the cathode

If the critical impurity concentrations are exceeded and several impurities are present simultaneously, the electrolytic process may no longer be possible. The concentration limits of the harmful impurities cannot be precisely specified, but the following guidelines can be used (mg/L):

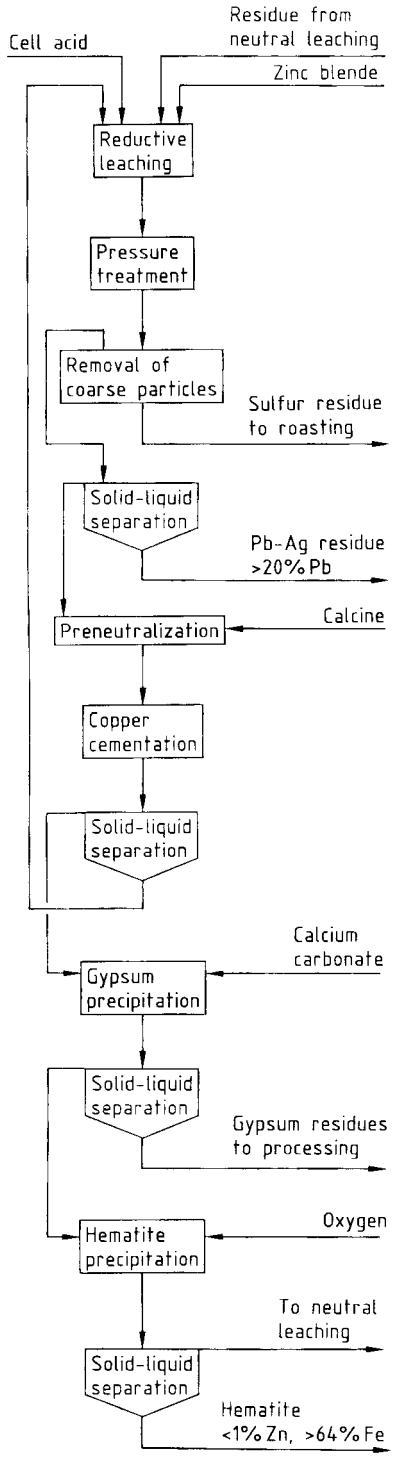

Figure 9. Flow diagram of the hematite process [3, p. 680]

As	0.01 – 1.
Sb	0.05 – 0.1
Ge	0.002 – 0.05
Ni	0.05 – 3.0
Co	0.1 – 1.0
Fe	20 – 30.

In the formerly used iron hydroxide precipitation process, the adsorptive effect of the hydroxide precipitate caused most of the arsenic, antimony, and germanium to be removed along with the iron, so reducing their concentrations below the acceptable limits.

However, when the residues from the neutral leaching are subjected to hot acid leaching (mainly in the interest of higher zinc yields), the neutral liquors contain much higher levels of impurities. As these are not removed either in the iron precipitation or the leaching stage, careful purification of the liquor is necessary. In principle, there are the following possibilities:

1) Chemical precipitation
2) Electrochemical deposition
3) Ion exchange (solvent extraction)
4) Cementation with zinc

For high quality requirements, a combination of methods can be used. Chemical precipitation is mainly used for nickel and cobalt. For example, nickel can be precipitated by dimethylglyoxime and cobalt by α-nitroso-β-naphthol. However, as these reagents are expensive, their use can only be considered in exceptional cases for liquors with very low levels of impurities.

Processes involving the use of fluidized-bed electrolysis to remove impurities more noble than zinc and the use of solid and liquid ion exchangers are still at the development stage.

In most zinc electrolysis plants, the liquor is purified by cementation of harmful elements on zinc dust in a continuous, multistage process. An inert coating can form on the zinc particles or they may be chemically dissolved. Therefore, not all the zinc dust added can take part in the cementation process and a large excess of zinc dust (5 – 10 fold) is necessary to ensure effective purification.

A typical process used in many plants is the hot – cold liquor purification process (Fig. 10), which has two successive precipitation stages.

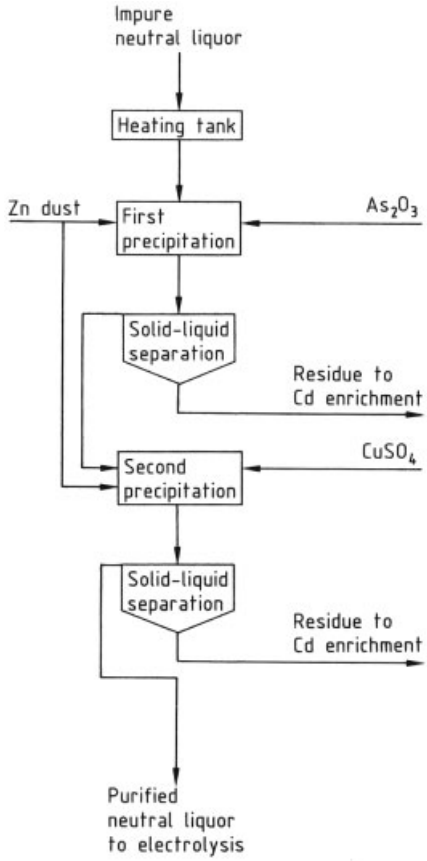

Figure 10. Hot-cold liquor purification [2, p. 607]

In the first stage, copper, nickel, and cobalt are precipitated by adding zinc dust in the presence of an added activator, usually As_2O_3. In the second stage, copper sulfate is the activator, and cadmium is removed. In many cases, a third fine purification stage is included.

The conventional hot – cold liquor purification process has a number of disadvantages:

1) The necessity of adding copper if too little copper is present in the starting materials
2) The possibility of releasing highly toxic arsine
3) The presence of the impurities arsenic and antimony in the precipitated copper
4) The possibility that the precipitated metals may redissolve

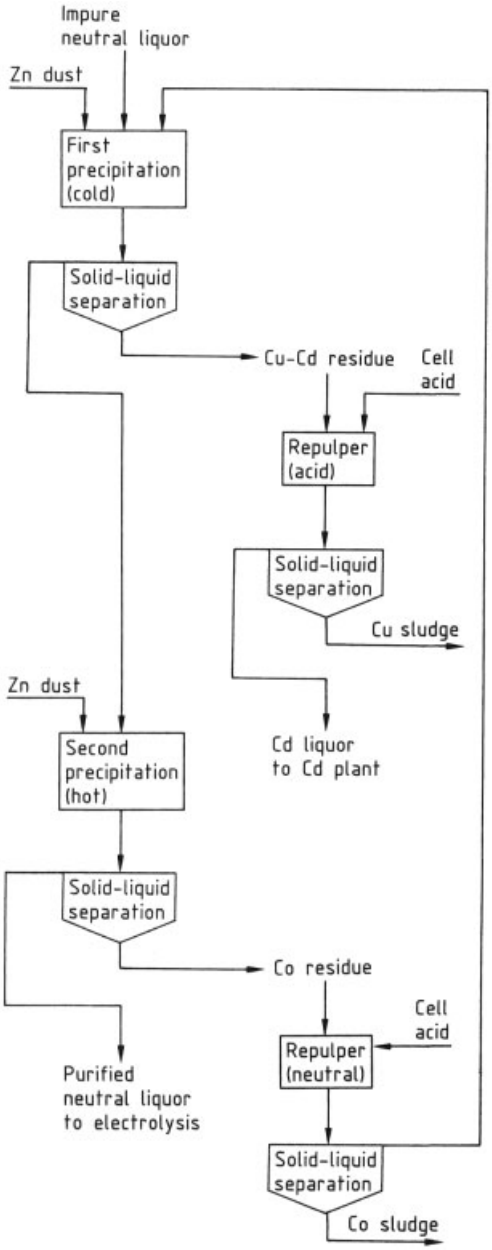

Figure 11. Reversed liquor purification [2, p. 608]

These disadvantages have led to the development of the alternative reversed liquor purification process (Fig. 11). Here, it is not necessary to add copper to remove cobalt and other impurities. The low concentrations of arsenic are sufficient to act as activator, the costly copper additions are avoided, and relatively pure cementates are obtained in both precipitation stages. This process can also be operated continuously, whereby the addition of zinc dust is controlled by measuring the electrode potential, which is a function of the copper and cadmium concentrations in the liquor.

5.7. Electrolytic Production of Zinc

[1–5], [10], [13]

Principles of Zinc Electrolysis. As the standard electrode potential of Zn^{2+} is -0.763 V, a solution of zinc sulfate should not deposit zinc on electrolysis: hydrogen should be liberated instead. Zinc deposition is only possible because of the hydrogen overvoltage at the zinc electrode, which causes a voltage shift of magnitude approximately equivalent to the standard electrode potential of zinc. The overvoltage is influenced by a number of factors. For example, the effect of current density is expressed by the Tafel equation:

$$\eta = a + b \log i$$

It decreases with increasing temperature, but, as the conductivity decreases with decreasing temperature, the energy consumption increases. With these effects in mind, two typical methods of performing zinc electrolysis were used in industry for a long time. Either the electrolytes were slightly acidic and current densities were ca. 325 A/m^2 or they were strongly acidic and current densities were ca. 850 A/m^2. The process parameters widely used today give good current yields in zinc electrolysis using current densities in the range 400 – 600 A/m^2 and a process temperature not exceeding 40 °C.

However, the hydrogen overvoltage is also influenced by the condition of the surface of the electrodes. Smooth surfaces lead to high overvoltages. For this reason, the cathode plate must be stripped every other day in practical operation. Zinc electrolysis is extremely sensitive to impurities in the electrolyte. These fall into several groups:

1) Impurities that are more electronegative than zinc, e.g., potassium, sodium, calcium, magnesium, aluminum, and manganese. These impurities do not directly interfere with the electrolytic process. If concentrations are

too high, the viscosity of the electrolyte increases and the diffusion processes in the boundary layers are disturbed. A limit of 60 g/L is used as a guideline figure.

2) Impurities whose electrode potential lies between those of Zn^{2+} and H^+. In the case of lead, cadmium, thallium, and tin, the hydrogen overvoltage is higher than that of zinc. These metals are deposited and form impurities in the zinc. In the case of iron, cobalt, and nickel, the overvoltage is lower than that of zinc. These are also deposited electrolytically but can be redissolved by the cell acid. Redox effects can lead to reduction in current yield. For this reason, the maximum tolerable concentrations are very low (see Section 5.6).
3) Impurities that are more electropositive than zinc. These include copper, arsenic, antimony, indium, tellurium, and germanium. These lead to a reduction in the overvoltage. The deposition of copper, nickel, cobalt, arsenic, and antimony can have a strongly detrimental effect on current yield due to the formation of local elements on the cathode, leading to dissolution of the zinc. For this reason, the maximum tolerable concentrations of these impurities are also very low, and the zinc electrolytes must therefore be rigorously purified.

The Practice of Zinc Electrolyis. Zinc electrolysis is a complex system in which positive and negative factors, especially as they affect current yield, energy consumption, required purity of the liquor, and mechanization, must be balanced against each other.

The purification of the liquor is now so well controlled that lower current densities can be used, enabling the electrolysis time between cathode changes to be extended from one to three days. Also, the cathode stripping operation, which involved heavy manual work, has been improved by the development of mechanical stripping equipment, and this has enabled cathodes with increased surface areas to be used.

In general, the electrolytic cells are concrete baths cast in one piece and lined with polyethylene. Lead linings are no longer used in modern electrolysis plants. The cells are arranged in rows and connected in series. The electrolyte supply to each cell is usually individually controlled.

The cathode plates consist of sheet aluminum (99.5 %), 5 – 7 mm thick, supported on carrier rods and provided with welded or plated copper contacts. The cathodes are polished at intervals of 4 – 6 weeks to prevent excessive adhesion of the cathode deposits. The latter problem can also be controlled by suspending the cathode plates in the bath and briefly running them as anodes.

Anode materials are usually Pb – Ag alloys with a silver content of 0.25 – 0.75 %. The anodes are usually made by casting the metal around a central copper carrier rod. Anodes that contain Sr and/or Ca and Ti in addition to Ag have higher strength and improved corrosion resistance.

5.8. New Developments in Zinc Production [2], [32]

Pressure Leaching of Sulfidic Zinc Concentrates. There have been many attempts to avoid the roasting process in the treatment of sulfidic zinc concentrates. For example, a process of pressure leaching of sulfidic concentrates has been developed by Sherritt Gordon Mines, Fort Saskatchewan, Alberta, Canada, and is used on an industrial scale.

The process can be represented by the equation:

$$ZnS + H_2SO_4 + 1/2\,O_2 \rightarrow ZnSO_4 + H_2O + S$$

which consists of the partial reactions:

$$ZnS + Fe_2(SO_4)_3 \rightarrow ZnSO_4 + 2\,FeSO_4 + S$$

$$2\,FeSO_4 + H_2SO_4 + 1/2\,O_2 \rightarrow Fe_2(SO_4)_3 + H_2O$$

Dissolved iron originates either from marmatitic blende (Zn, Fe)S or pyrrhotite Fe_7S_8, but not from pyrites FeS_2, which does not dissolve under these liquor conditions. This iron promotes the oxygen transfer reaction that converts the sulfide to sulfate.

The process is carried out in two stages, as shown schematically in Figure 12. The zinc concentrate is size reduced by wet grinding to 90 % <44 µm and oxidatively leached in two stages at 150 °C at an oxygen pressure of 1 bar. Exhausted electrolyte (cell acid) is used as solvent. In the

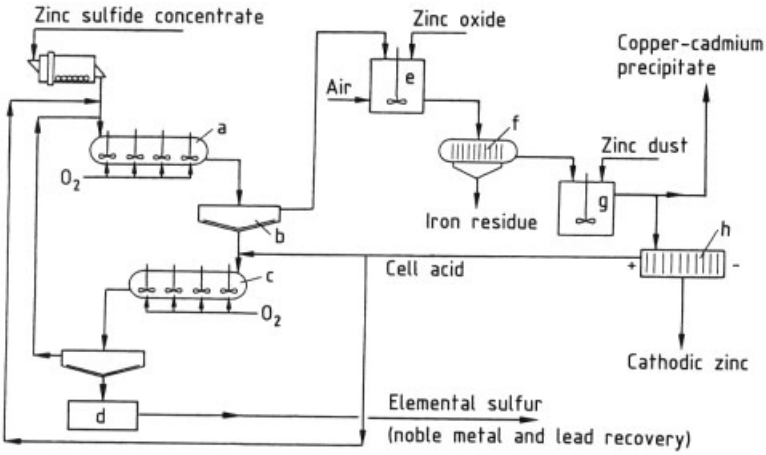

Figure 12. Pressure leaching of zinc concentrate [2, p. 610]
a) Pressure leaching, stage 1; b) Thickener; c) Pressure leaching, stage 2; d) Washing; e) Iron precipitation; f) Filtration; g) Cleaning tank; h) Electrolysis

first stage, 80 % of the zinc present is dissolved, and in the second stage a further 18 %.

The overflow from the thickener in the first stage, at a pH of 1.5 – 2.5, is treated with zinc oxide to neutralize the residual iron that was not precipitated in the autoclave, and the liquor so obtained is filtered and purified in the usual way. It is then fed to the zinc electrolysis stage.

The residue from the first stage is pumped together with cell acid into the autoclave of the second stage. After solid – liquid separation, the solution from the second stage is recycled to the first autoclave. The residue is washed in countercurrent flow to remove elemental sulfur if necessary, and the residual iron oxide is treated to recover lead and noble metals or disposed of, depending on its composition.

The process is used in two Canadian zinc smelting plants, one in Trail (Cominco), and the other in Timmins (Texas Gulf) in Ontario.

Direct Leaching of Oxidic Zinc Ores [2]. Some new methods of directly leaching oxidic ores (smithsonite, willemite, etc.) have been developed.

A willemite containing 40 % zinc has been directly leached at the zinc smelting plant at Trés Marias of the Companhia Mineira de Metais in Brazil. Smithsonite with a lower zinc content (<30 %) is first enriched by treatment in a Waelz kiln, and the oxide obtained is then leached.

6. Uses and Economic Aspects

Two properties of zinc are of special relevance to its applications. Firstly, its standard electrode potential is considerably more positive than that of the construction material iron. Zinc coatings therefore provide very good corrosion protection for steel structures. Secondly, zinc has a relatively low melting point, which makes it suitable for use in pressure die casting of complex components.

6.1. Galvanizing [3], [33]

The ever-increasing need to make the best use of the world's raw material reserves means that corrosion prevention as a method of preserving the usefulness of commercial metals continues to increase in importance. Moreover, as the atmosphere becomes increasingly corrosive, standards of corrosion protection must continue to rise, especially in steel construction.

A zinc coating on steel ensures lasting protection against the corrosive effects of the weather by the following mechanisms:

1) The effect on zinc of the carbon dioxide and moisture in the air is to form protective coatings with a zinc carbonate basis, resulting in a low corrosion rate

2) Because of its electrochemical property of providing cathodic protection, zinc actively prevents the occurrence of corrosion in the region of small defects, damaged areas of coating, and cut edges of sheet

In the industrialized countries, almost 50 % of the zinc consumed is used for corrosion protection.

Commercial zinc (98.5 – 99.5 %) is usually used for hot-dip galvanizing in production and jobbing plants, and high-purity zinc (99.95 %) for the continuous galvanizing of steel strip or wire. The metallurgical processes used to produce the less pure grades of zinc mean that lead is the main impurity present. This is an advantage, as the presence of 1 % lead reduces the surface tension of zinc by ca. 40 % compared with that of the high-grade product. The components being galvanized are therefore more effectively wetted by the molten zinc.

During the process of immersion in the zinc bath, which takes only a few minutes, the component is heated to ca. 450 °C. Diffusion processes cause layers of iron – zinc alloys to form on the steel surface, giving very good adhesion of the zinc coating.

Other galvanizing processes have been developed in addition to the hot-dip process. In electrolytic galvanizing, a d.c.current causes the deposition of lustrous coating, 5 – 25 µm thick, from acidic, cyanide, or alkaline cyanide-free electrolytes. However, these only give protection against mildly corrosive media. In general, additional protection is provided by chromate passivation.

In the spray galvanizing process, zinc wire is melted in a fuel gas/oxygen flame or an electric arc at the jet of a spray pistol and is propelled by compressed air from a distance of 15 cm, giving a coating ca. 0.1 mm thick. The resulting purely mechanical adhesion of the zinc is inferior to that produced by alloy formation at the metal interface in hot-dip galvanizing.

6.2. Zinc as a Construction Material [2], [34]

Although zinc has fairly good strength properties, it was for a long time not used as a material of construction because of its very poor creep resistance. Low-alloy zinc grades based on Zn – Ti – Cu, which have very good ductility and creep resistance, were developed only ca. 40 years ago. This material can be rolled to form zinc plate or sheet which can be used in the manufacture of roof drainage components or for covering buildings. A wide range of shapes has been developed, enabling aesthetic appeal to be combined with protection of buildings from the rigors of nature.

6.3. Pressure Die Casting with Zinc [2], [35]

Pressure die casting is another major area of use and is dominated by alloys of the type ZnAl4Cu1 and modifications thereof.

Some of the advantages of zinc for pressure die casting are:

1) Production runs can be very large (60 – 1000 components), depending on the size and complexity of the components
2) The dies are extremely durable, enabling up to 10^6 components to be produced from one die
3) The low viscosities and low melting points of the alloys ensure a high definition product
4) Dimensional accuracy is ensured as shrinkage of the zinc alloys is low and easily controlled
5) Lustrous or matt surfaces can be obtained with various metallic coatings and effects
6) Costs are low due to large production runs
7) The products are durable because of their good mechanical and physical properties
8) The energy requirements of manufacture are low

Approximately 50 % of all zinc pressure die castings go to the automobile industry.

6.4. Zinc as a Chemical [2], [36]

Zinc-based chemicals, including zinc dust, account for ca. 12 – 15 % of world zinc consumption. Zinc oxide is quantitatively by far the most important zinc-based chemical product, followed by zinc dust, zinc sulfate, and zinc chloride in order of importance and quantity. All other zinc compounds are of minor importance.

World consumption figures for zinc oxide and zinc dust have remained constant for many years. However, the demand for zinc sulfate and chloride is decreasing. The rate of growth of the consumption of zinc thiocarbonate and zinc stearate matches that of rubber production, their main area of use. Zinc oxide, the compound most in demand, can be produced by various processes. Its purity and quality depend on the production method used. In the indirect (so-called French) process, pure zinc is heated and oxidized in a current of air. The very pure zinc oxide formed is collected as a powder in settling chambers where it separates into different particle sizes. This material is usually known as zinc white. In the direct (so-called American) process, the raw materials used are zinc ores or zinc byproducts, which usually contain lead. A carbonaceous material is heated with the raw material, reducing it with formation of zinc vapor. This is oxidized with air and separated into particle size fractions.

The starting materials for wet chemical processes are solutions of purified zinc. Zinc carbonate or hydroxide is precipitated and is then filtered, washed, dried, and calcined at ca. 800 °C.

The various types of zinc oxide include an industrial grade, which is used as an intermediate product in the production of other chemicals; pigment grades, which include almost all the lead-containing grades; precipitated zinc oxides which have no pigment properties; and special grades such as the extremely finely-divided zinc oxide for photocopying paper.

The main consumer of zinc oxide is the rubber industry, in which it is used as a vulcanization activator and sometimes also as a filler.

Zinc oxide as a pigment in aqueous latex paints has lost a significant share of the market. Its use in agriculture as an additive to fertilizers to treat zinc deficiency in soils is of lesser importance. Zinc oxide is a component of formulations in the glass, enamel, and ceramic industries, where it influences the melting point and also the optical and elastic properties, color, and luster of the glaze. It is also a component of face powder, lipsticks, and creams used in the cosmetics industry. It is also used as an additive to lubricants, adhesives, drying agents, and delustering agents, and as a catalyst in methanol synthesis.

Table 4. Development of zinc consumption [37]

	1981	1985	1990
Germany	405.7	408.8	484.0
Belgium	151.8	169.1	177.6
France	330.0	246.9	284.0
United Kingdom	181.3	189.3	189.0
Italy	236.0	218.0	278.0
The Netherlands	45.2	51.1	71.7
Greece	18.8	15.2	20.3
Portugal	19.0	8.3	16.0
Spain	91.1	95.3	118.7
Finland	24.4	26.0	29.1
Yugoslavia	87.8	105.6	108.9
Norway	20.2	21.0	15.9
Austria	26.9	31.9	39.0
Sweden	38.0	31.5	39.5
Switzerland	25.2	26.3	22.3
Denmark	16.1	12.2	13.0
Other Europe	1.4	1.3	1.4
Europe	**1719,1**	**1657.8**	**1908.4**
Hong Kong	22.6	21.5	17.3
India	95.3	134.0	135.0
Indonesia	51.4	50.6	53.2
Japan	752.3	780.1	814.3
Malaysia	14.5	14.6	22.6
Pakistan	9.7	11.6	25.0
Philippines	16.1	13.0	33.1
Singapore	11.8	6.4	13.0
S. Korea	68.0	119.7	227.4
Taiwan	38.4	49.4	79.4
Thailand	33.3	39.6	66.0
Turkey	12.3	49.8	47.7
Other Asia	13.3	38.2	52.8
Asia	**1139.0**	**1328.5**	**1586.6**
Egypt	2.7	7.0	10.0
Algeria	11.0	11.0	15.0
Ivory Coast	4.0	4.6	2.0
Kenia	7.2	6.7	8.2
Morocco	3.0	2.7	3.0
Nigeria	12.7	7.9	6.0
Rep. of South Africa	84.1	84.1	86.5
Tanzania	3.2	3.5	2.0
Other Africa	8.6	9.3	18.3
Africa	**136.5**	**136.8**	**150.0**
Argentina	30.9	25.8	19.1
Brazil	155.9	132.4	129.7
Canada	140.0	156.5	126.1
Columbia	14.6	18.3	8.6
Mexico	88.9	99.2	110.1
Peru	23.2	41.2	71.0
Venezuela	20.2	15.4	10.6
United States	878.6	961.4	996.8
Other America	28.0	22.7	14.1
America	**1380.3**	**1472.9**	**1486.5**
Australia	95.4	82.1	82.2
New Zealand	20.7	24.7	20.0
Other Oceania	0.1	0.1	0.1
Australia and Oceania	**116.2**	**106.9**	**102.3**

Table 4. (Continued)

	1981	1985	1990
Western Countries	**4491.1**	**4702.9**	**5234.7**
Bulgaria	75.0	70.0	46.6
German Democratic Republic	86.2	69.7	45.5
Poland	178.3	157.2	109.5
Rumania	49.9	35.3	12.8
Czechoslovakia	67.0	58.0	51.8
Soviet Union	1030.0	1000.0	920.0
Hungary	25.0	24.5	12.0
China	259.0	349.0	500.0
Others	28.0	32.0	45.0
Cuba	1.5	1.8	1.2
Eastern Countries	**1799.9**	**1797.5**	**1744.4**
World total	**6291.0**	**6500.4**	**6979.1**

Zinc dust is produced by a similar process to that used for zinc oxide, but in a reducing atmosphere. There are two grades: pigment zinc dust and zinc dust for chemical use. Zinc dust is used as a pigment for anticorrosion paints and coatings whose primary function is to form an impermeable surface. The action of the environment leads to the formation of zinc carbonate and basic zinc compounds between the zinc particles, reinforcing this effect. These coatings also give cathodic protection, which is especially useful if there is slight damage to the coating layer.

6.5. Economic Aspects

The development of world zinc consumption in the important industrial countries and regions of the world is illustrated by the figures for the years 1980, 1985, and 1990, which are listed in Table 4.

7. References

1. V. Tafel: *Lehrbuch der Metallhüttenkunde,* 2nd ed., **vol. 2,** Hirzel-Verlag, Leipzig 1953, pp. 357 – 655.
2. *Ullmann,* 4th ed., **24,** 594 – 626.
3. F. Pawlek: *Metallhüttenkunde,* De Gruyter, Berlin – New York 1983, pp. 654 – 689.
4. Z. Horvath: *A Zinc Kohàszata,* Akadèmiai Kiadò, Budapest 1961.
5. N. N. Muratsch: *Handbuch des Metallhüttenmanns,* Verlag Technik, Berlin 1954, pp. 511 – 566.
6. J. Torka, R. Mewes: *Die Wunder der Technik,* Paul Oestergaard, Berlin.
7. *Meyers Neues Lexikon,* 2nd ed., **vol. 15,** Bibliographisches Institut, Leipzig 1977, pp. 444 – 445.
8. P. Klemm: *Der Weg aus der Wildnis,* vol. 3, Kinderbuchverlag, Berlin 1962.
9. *Gmelin,* System-no. 32, 1969, 21 – 26, 56 – 61.
10. R. Zimmermann, K. Günther: *Metallurgie und Werkstofftechnik,* **vol. 1,** pp. 216 ff; vol. 2, p. 117, Deutscher Verlag für Grundstoffindustrie, Leipzig 1977.
11. *Römpp,* 9th ed., **vol. 6,** Thieme Verlag, Stuttgart 1992, pp. 5136 – 5139.
12. Metallgesellschaft, *Metallstatistik 1980 – 1990,* vol. 78, Frankfurt 1990.
13. *Zink, der vielseitige Werkstoff,* vol. 22, Mitteilungen aus den Arbeitsbereichen der Metallgesellschaft, Frankfurt/Main 1980.
14. *Brockhaus Taschenbuch der Geologie,* Brockhaus Verlag, Leipzig 1955.
15. S. Jankovic: *Wirtschaftsgeologie der Erze,* Springer Verlag, Wien – New York 1974, Specific References.
16. R. Kola, *Erzmetall* **35** (1982) 130 – 137.
17. H. Maczek, R. Kola, *J. Met.* **32** (1980) 53 – 58.
18. *Sintering Symposium,* Port Pirie, South Australia, Sept. 1958.
19. N. B. Gray, M. R. Harvey, G. M. Willis: "Physical Chemistry of Process Metallurgy," *Inst. Min. Metall. Sect.* (1974) 19.
20. Y. Fukunaka et al., *Metall. Trans. B* **7B** (1960) 307.
21. F. Cappel, H. Wendeborn: *Sintern von Eisenerzen,* Stahleisen, Düsseldorf 1973.
22. T. R. Ingrahm, H. H. Kelogg, *Trans. Am. Inst. Min. Metall. Eng.* **227** (1963) 1419.
23. J. Krüger, R. Püllenberg, *Erzmetall* **33** (1980) 70; **34** (1981) 380.
24. F. Johannsen, *Z. Erzbergbau Metallhüttenwes.* **1** (1948) 235; *Erzmetall* **27** (1974) 313.
25. Ch. Friedrich et al., *Neue Hütte* **16** (1971) 457.
26. H. Maczek, R. Kola, *J. Met.* **32** (1980) 53.
27. G. Kossek et al., *Stahl Eisen* **96** (1976) 482.
28. J. Krüger, *Freiberg. Forschungsh. B* **B228** (1981) 47.
29. O. Knacke, W. Neumann, *Erzmetall* **9** (1956) 261.
30. O. H. Schütze, *Erzmetall* **3** (1950) 69.
31. G. Hofmann: *Industrieöfen,* Deutscher Verlag für Grundstoffindustrie, Leipzig 1970.
32. H. Veltmann, D. R. Weir, *Erzmetall* **35** (1982) 76 – 77.

33. W. Thiele in: *Zink – der vielseitige Werkstoff,* Mitteilung aus den Arbeitsbereichen der Metallgesellschaft, vol. 22, Frankfurt 1980, pp. 67 – 73.
34. W. Burggraef in: *Zink – der vielseitige Werkstoff,* Mitteilung aus den Arbeitsbereichen der Metallgesellschaft, vol. 22, Frankfurt 1980, pp. 81 – 82.
35. J. Johnen in: *Zink – der vielseitige Werkstoff,* Mitteilungen aus den Arbeitsbereichen der Metallgesellschaft, vol. 22, Frankfurt 1980, pp. 83 – 84.
36. H. Kunze in: *Zink – der vielseitige Werkstoff,* Mitteilungen aus den Arbeitsbereichen der Metallgesellschaft, vol. 22, Frankfurt 1980, pp. 88 – 90.
37. H. J. Holtmeier, M. Kuhn in: *Zink – der vielseitige Werkstoff,* Mitteilungen aus den Arbeitsbereichen der Metallgesellschaft, vol. 22, Frankfurt 1980, pp. 95 – 97.

Zinc Alloys

Copper – Zinc Alloys → Copper Alloys

HEINRICH J. JOHNEN, Kaarst, Federal Republic of Germany

1. **Introduction** 1
2. **Casting Alloys** 2
3. **Ductile Alloys** 5
4. **Zinc Alloys for Special Applications** . . 6
5. **Economic Aspects** 6
6. **References** 6

Based on the article Zink-Legierungen in Ullmann's, 4th ed., written by MANFRED PÖTZSCHKE

1. Introduction

Zinc recrystallizes just above room temperature and has low creep resistance. It is therefore only suitable as a construction material when alloyed. The alloying elements, mainly Al, Cu, Ti, and Mg, cause grain refinement, mixed crystal formation, or precipitation hardening, thereby considerably improving the mechanical properties.

The properties of zinc alloys are largely determined by the hexagonal structure of zinc (*c/a* ratio 1.856). The physical and mechanical properties and plastic deformation behaviour vary with the orientation with respect to the principal hexagonal axis (anisotropy). There can also be marked texturization due to casting and deformation, such that characteristic values depend on the direction of the deformation. No significant hardening occurs on cold working, as recovery and relaxation take place rapidly. Recrystallization by seed formation begins at temperatures in the range 100 – 300 °C and above, depending on the alloy type.

The industrial zinc alloys combine outstanding metal forming properties with a well-balanced range of application properties. Casting alloys can be machined and ductile alloys can be formed using standard methods in each case.

Corrosion Behavior. Zinc alloys can be coated by the usual methods of electroplating, vapor deposition, or painting. Corrosion resistance can be improved by chromating or phosphating. The alloys are amphoteric, dissolving in acids and strong alkalis. They are therefore widely used under mildly corrosive conditions, e.g., in water or solutions whose pH is in the range 6 – 12.5, and in the atmosphere. Strongly adherent insoluble (basic) coatings of hydrozincite are formed which inhibit further attack. Depending on the aggressiveness of the atmosphere, the annual corrosion rate can be between 1 – 4 µm (rural air) and 4 – 13 µm (industrial air) [2], [3].

Zinc undergoes galvanic corrosion by ions of more noble metals (e.g., copper), and contact corrosion can lead to pitting, although in structures in which zinc alloys are in contact with aluminum, iron, or stainless steel components, contact corrosion does not occur as the difference in electrode potential is too small.

If foreign elements originating from the zinc ore, e.g., Pb, Cd, Sn, In, or Tl, are present in Zn – Al alloys, there is a tendency to intergranular corrosion. This can be limited to some extent by adding Mg. This tendency is the reason why zinc alloys only became widely used when it became possible to produce high-grade zinc (99.99 – 99.995 %, DIN 1706) on an industrial scale in the mid-1930 s. Stress corrosion of zinc alloys is not observed.

Monographs are available on the corrosion behavior of zinc exposed to the atmosphere [4], in water [5], and in the presence of chemicals [6] and bitumen [7].

Methods of measuring the corrosion of zinc alloys are given in the following standards. (see below)

Also, tests are carried out in atmospheres saturated with water vapor and in autoclaves to simulate the behavior of castings exposed for long periods. Both tests also give information about casting defects in the surface.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_531

DIN 50 017	Corrosion in damp-heat atmospheres
DIN 50 018	Corrosion by condensates: changeable climates with SO_2-containing atmospheres
DIN 50 021	Corrosion by mist sprays of various sodium chloride solutions
DIN 50 956	Tests of porosity of coatings on zinc and zinc alloys: copper sulfate method
DIN 50 958	Corrosion testing of chromated articles: modified Corrodkote method

Material Properties. Testing of zinc alloys is regulated not only by national and international standards, but also by the industry standards of the European zinc producers, e.g., for the following tests:

1) Tensile testing of sheet and strip under uniaxial stress
2) Tensile testing of pressure die cast test pieces (DIN 50 148)
3) Vickers or Brinell hardness testing (DIN 50 351)
4) Flat bending fatigue test on sheet, strip, and pressure die cast test pieces (DIN 50 100)
5) Notched bar (Charpy) impact test (DIN 50 116, July 1982)
6) Folding behavior

The material tests also measure variations in some characteristic values with testing rate, an effect typical of zinc alloys. For example, tensile strength increases with increasing strain rate, while elongation decreases.

A comprehensive metallographic atlas of zinc and zinc alloys has been produced by the European zinc manufacturers [8].

2. Casting Alloys

ZnAl(Cu) Alloys. Standard specifications for casting alloys are given in DIN 1743 (April 1978) Sheet 1 (ingots) and Sheet 2 (castings). The main alloying elements are Al, Cu, and Mg, and the zinc has a purity of 99.99 – 99.995 % (DIN 1706). The solubility of Al in Zn in the solid state is only 0.2 wt % (see Zn – Al phase diagram, Fig. 1), and that of Cu 2.7 wt %. The decreasing solubility of both alloying elements with decreasing temperature causes segregation processes and hence grain refinement and a considerable improvement in the mechanical properties, especially tensile strength, hardness, and impact strength. Aluminum lowers the solubility of Fe, so that molten Zn can be melted and kept molten in iron vessels. The Mg addition (up to 0.06 wt %) inhibits intercrystalline corrosion and compensates for the harmful effect of companion elements (DIN 1743). It also increases the hardness.

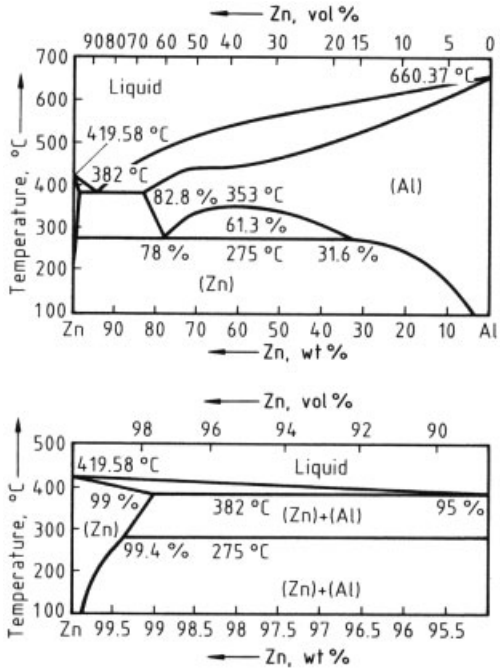

Figure 1. Zn – Al phase diagram

Typical of ZnAl and ZnAlCu alloys is the eutectoid structural transformation in the two-component system (beta decomposition of the Al-rich mixed crystal) or phase transformation of the Γ-phase in the three-component system. Both phenomena cause dimensional changes and deterioration of mechanical properties (aging), but do not affect the technical application properties.

Table 1 lists the ZnAl(Cu) alloys, some specified in DIN 1743, with their designations, material properties, compositions, and typical applications. For pressure die casting (indicated by the letters GD), alloys containing ca. 4 wt % Al and ca. 1 wt % Cu (Z 410) and Cu-free alloys containing ca. 4 wt % Al (Z 400) are used. In Germany, the alloy Z 410 accounts for 90 % of total consumption, whereas ca. 90 % of U.S.

Table 1. Zinc – aluminum (– copper) casting alloys

Designation	Material properties					Alloying		Impurity limits,		Density,	Recommended use	Commercial
	Tensile strength, N/mm^2	0.2 % offset yield strength, N/mm^2	Fracture strain, %	Brinell hardness	Flexural fatigue strength, ** N/mm^2	constituents, wt %		wt %		kg/dm^3 (approx.)		designation
GD-ZnAl4Cu1 *	280 – 350	220 – 250	2 – 5	85 – 105	70 – 100	Al	3.5 – 4.3	Fe	0.05	6.7	master alloy pressure die casting	Z 410
						Cu	0.4 – 1.1	Ni	0.02			
						Mg	0.020 – 0.06	Pb + Cd	0.009			
						Zn	remainder	Sn	0.002			
GD-ZnAl4 *	250 – 300	200 – 230	3 – 6	70 – 90	60 – 80	Al	3.5 – 4.3	Cu	0.1	6.6	pressure die casting	Z 400
						Mg	0.020 – 0.06	Fe	0.05			
						Zn	Ni	0.02				
						remainder						
								Pb + Cd	0.009			
								Sn	0.002			
G-ZnAl4Cu3 *	220 – 260	170 – 200	0.5 – 2	90 – 100		Al	3.5 – 4.3	Fe	0.075	6.8		Z 430
GK-ZnAl4Cu3 *	240 – 280	200 – 230	1 – 3	100 – 110		Cu	2.5 – 3.2	Pb + Cd	0.009	6.8	all types of pressure and gravity die casting and sand casting, sheet metal press tools, molds for injection, blow, and deep draw molding of plastics	
						Mg	0.03 – 0.06	Sn	0.002			
						Zn	remainder					
G-ZnAl6Cu1 *	180 – 230	150 – 180	1 – 3	80 – 90		Al	5.6 – 6.0	Fe	0.075	6.5		Z 610
GK-ZnAl6Cu1 *	220 – 260	170 – 200	1.5 – 3	80 – 90		Cu	1.2 – 1.6	Mg	0.005	6.5	technically difficult sand and gravity die casting, strand casting	
						Zn	remainder	Pb + Cd	0.009			
								Sn	0.002			
GK-ZnAl8Cu1	220 – 255	206	1 – 2	85 – 90		Al	8.0 – 8.8	Fe	0.10	6.3	thin-walled precision casting	ZA 8
						Cu	0.8 – 1.3	Pb	0.004			
						Mg	0.015 – 0.03	Cd	0.003			
						Zn	remainder	Sn	0.002			
G-ZnAl12Cu1	275 – 310	206	1 – 3	105 – 125		Al	10.5 – 11.5	Fe	0.075	6.0	thin-walled precision casting	ZA 12
						Cu	0.5 – 1.25	Pb	0.004			
						Mg	0.015 – 0.03	Cd	0.003			
						Zn	remainder	Sn	0.002			
G-ZnAl27Cu2	400 – 440	365	3 – 6	110 – 120		Al	25 – 28	Fe	0.10	5.0	thin-walled precision casting	ZA 27

Table 1. (Continued)

Designation	Material properties					Alloying		Impurity limits,		Density,	Recommended use	Commercial
	Tensile strength, N/mm^2	0.2 % offset yield strength, N/mm^2	Fracture strain, %	Brinell hardness	Flexural fatigue strength,** N/mm^2	constituents, wt %		wt %		kg/dm^3 (approx.)		designation
(as cast)						Cu	2.0 – 2.5	Pb	0.004			
Heat treated 3 h, 320 °C	310 – 325	255	8 – 11	90 – 100		Mg	0.01 – 0.02	Cd	0.003			
						Zn	remainder	Sn	0.002			
ZnAl30Cu5	200 – 500	200 – 250		140 – 150						4.8	plain bearings	

* In accordance with DIN 1743, Sheet 2.
** After 20×10^6 load reversals.

consumption is of Cu-free alloy. The standard alloys for sand casting (indicated by S) or gravity die casting (indicated by GK) include those containing 4 wt % Al and 3 wt % Cu (Z 430) and the quasi-eutectic alloy containing 6 wt % Al and 1 wt % Cu (Z 610). Other alloys containing 8, 12, and 27 wt % Al (ZA 8, 12, and 27) have been developed as casting and forging alloys for special applications. There is also an alloy containing 30 wt % Al and 5 wt % Cu for use in plain bearings.

Uses and Casting Methods (see also → Foundry Technology). High-purity or fine zinc (FZn) alloys are mainly used in automatic pressure-die casting plants, in which a measured quantity of molten alloy at 390 – 420 °C (the casting temperature) is rapidly injected at high pressure into a steel die with a temperature of 160 – 200 °C. The cast component so formed is dimensionally very accurate and requires very little treatment to remove sprue and flash. The cast components have an excellent surface finish which can be further improved electrolytically, chemically, or by coating with organic materials. In Germany, ca. 10^{10} cast components are produced annually. The smallest components, with wall thicknesses of 0.3 – 1.0 mm, can be produced at rates of 500 – 5 000 shots per hour, and medium sized components with wall thicknesses of 1.0 – 2.0 mm at 100 – 400 shots per hour. (Note that dies with several cavities, i.e., multiple dies, can produce several components per shot.) Large components with wall thicknesses of 1.2 – 3.0 mm can be produced at the rate of 60 – 150 components per hour. The steel dies have a working life of 500 000 – 5 000 000 shots. Cavities and recesses in the components can be cast directly.

Pressure die casting is an economical process which enables complex components to be produced. The small wall thicknesses cannot be produced by any other casting process or casting alloy. Even so, wall thicknesses of >50 mm can also be obtained.

The dies are hardened permanent molds of high-grade hot-forming tool steel. Gating technology in pressure die casting is based on the latest findings of fluid mechanics. Molten-metal flow rates can reach 80 m/s.

Zinc pressure die castings have a wide range of uses as components in the automobile indus-

try, bicycle and motor car wheels, machinery and equipment, vending machines, office machines, domestic appliances, precision equipment, electrical technology, and a wide range of everyday utensils.

Sand and gravity die casting of zinc alloys is preferable to pressure die casting if the number of castings required does not justify manufacture of dies. In sand casting, the patterns can be made of wood, gypsum, metal, or plastics. Casting temperatures are 420 – 480 °C for Z 430, and 390 – 420 °C for the quasi-eutectic alloy Z 610. The shell or Croning process can also be used.

Gray iron molds are suitable for gravity die casting. The castings obtained have a smoother surface than those produced by sand casting. Casting temperatures are the same as those for sand casting. Mold temperatures are usually in the region of 200 °C. Sheet metal pressing tools are manufactured by zinc casting. These can be accurately cast, minimizing fettling costs. The self-lubricating properties of zinc are utilized in sheet metal pressing.

3. Ductile Alloys

ZnCuTi Alloys. Pack-rolled ductile zinc alloys based on commercial zinc (purity 98.5 – 99.5 % Zn, the rest mainly Pb) were produced until the mid-1960 s. This was followed by rapid developments that led to the production of rolled strip ZnCuTi materials (titanium zinc) in which the principal component was high-purity (99.995 %) zinc. Titanium zinc is a zinc alloy in accordance with DIN 17 770, containing 0.1 – 0.8 wt % Cu, 0.08 – 0.15 wt %Ti, and 0.002 – 0.01 wt % Al. It is dispersion-hardened "in situ", the dispersoids being the intermetallic compounds $TiZn_{15}$ and ε-Zn – Cu. A combination of alloying, rolling, and heat treatment leads to materials with relatively high creep strength (creep limit $\geq$50 N/mm^2 for 1 %/a elongation), good mechanical properties (tensile strength $\geq$150 N/mm^2, 0.2 % offset yield stress $\geq$100 N/mm^2), high ductility (fracture strain $\geq$40 %) and good working properties (e.g., solderability). They are mainly used in the building industry for roof drainage systems, flashing, and non-load-bearing profiles. Using a special technique, titanium zinc can be deep drawn like brass or steel.

Other Ductile Alloys. Etched plates of FZn – Al – Mg alloys (ca. 0.01 wt % Al, ca. 0.05 wt % Mg) are used for graphic letterpress printing, FZn-Pb-Cd alloys (0.03 – 1.0 wt % Pb, 0.03 – 1.0 wt % Cd) for the negative electrodes of primary cells, and FZn – Mn alloys (up to 1.0 wt % Mn) for stampings.

During World War II, zinc hot pressing alloys and zinc conductors were used as substitutes for other nonferrous metals, and high-Al alloys were sometimes used as low-temperature alloys.

Superplastic ductile Zn – Al alloys containing, e.g., 4 or 22 wt % Al (trade name Zilôn) have become of some industrial importance. Titanium zinc has superplastic forming properties under certain conditions.

Forging alloys (Zn – Al alloys containing 12 or 27 wt % Al) are used in special applications.

4. Zinc Alloys for Special Applications

Zinc performs an extremely important function in preventing the corrosion of steel, and a series of special zinc alloys (mainly based on commercial zinc, DIN 1706) have been developed for hot-dip galvanizing or for use in cathodic corrosion protection:

1) Zinc containing 0.05 – 0.3 % Pb and 0.15 – 0.35 % Al (Sendzimir galvanizing)
2) Zinc alloy containing 55 % Al and 1.0 % Si (Galvalume, Zincalume)
3) Zinc alloy containing 5 % Al and 0.1 % SE (Galfan)

Master alloys for hot-dip galvanizing:

1) Zinc alloy containing 4.5 % or 9 % Al
2) Zinc containing 0.05 % Al, 0.008 % Mg, 0.15 % Sn (Polygalva for galvanizing Si-containing steel)
3) Zinc having 1 % Pb and reduced Cd content (0.05 wt % Cd) for pipes for potable water (DIN 2444)

Anodes for cathodic corrosion protection: Zinc with "activating" additives, e.g., Al, Cd.

5. Economic Aspects

The pattern of consumption of zinc in the western world and in Germany in 1991 is given in the following:

Castings	117 600 t
Brass production	114 000 t
Galvanizing	92 800 t
Semifinished products	95 000 t
Chemical products (including dust and powder)	23 100 t
Others	7 400 t

Western World consumption was 5.36×10^6 t, and total World production was 6.89×10^6 t.

6. References

General References

1. Zinkberatung e.V. (eds.): *Zink-Taschenbuch,* Metall-Verlag, Berlin-Heidelberg 1981. Zinkberatung e.V. (eds.): *Gießen mit Zink,* Düsseldorf 1982. Metallgesellschaft: Mitteilungen aus den Arbeitsbereichen, issue 22: Zink, der vielseitige Werkstoff, Frankfurt 1980. G. Wassermann, P. Winchierz: *Das Metall-Laboratorium der Metallgesellschaft AG, 1918 – 1981 (Chronik und Bibliographie)* , Metallgesellschaft, Frankfurt 1981. W. Heine, U. Zwicker, *Z. Metallkd.* **53** (1962) 380 – 388. F. E. Goodwin, A. L. Ponikvar: *Engineering Properties of Zink Alloys,* ILZRO, 1988. Zinkberatung e.V. (eds.): Technische Mitteilung Zink-Druckguß, Schriftenreihe V 3.25: Optimierte Legierungs- und Gießtechnik für Konstruktionen in Zink-Druckguß, Feb. 1984; V. 3.34: Untersuchungen an den Zinkdruckguß-Legierungen Z 400, Z 410, Z 430, Z A 8, Aug. 1990. Düsseldorf.

Specific References

2. H.-J. Böttcher, *Werkst. Ihre Veredl.* **4** (1982) 109 – 113.
3. Beratungsstelle für Stahlverwendung (eds.): Merkblatt 400 ”Korrosionsverhalten von feuerverzinktem Stahl,“ 4th ed., Düsseldorf 1983.
4. G. Schikorr in Zinkberatung e.V. (eds.): *Korrosionsverhalten von Zink,* **vol. 1**: Verhalten von Zink an der Atmosphäre, Metall-Verlag, Berlin 1962.
5. W. Wiederholt in Zinkberatung e.V. (eds.): *Korrosionsverhalten von Zink,* **vol. 2**: Verhalten von Zink in Wässern, Metall-Verlag, Berlin 1965.
6. W. Wiederholt in Zinkberatung e.V. (eds.): *Korrosionsverhalten von Zink,* **vol. 3**: Verhalten von Zink gegen Chemikalien, Metall-Verlag, Berlin 1976.
7. C. A. Witt, D. Haller in Zinkberatung e.V. (eds.): *Korrosionsverhalten von Zink,* **vol. 4**: Verhalten von Zink in Verbindung mit bituminösen Baustoffen, Metall-Verlag, Berlin 1976.C. A. Witt in Zinkberatung e.V. (eds.): *Korrosionsverhalten von Zink,* **vol. 4-II**: Verhalten von Zink in Verbindung mit Bitumen – Korrosion und Prüfung, Metall-Verlag, Berlin 1980.
8. Centre de Recherches Métallurgiques (eds.): *Europäische Zinkerzeuger: Metallographischer Atlas für Zink und Zink-Legierungen,* Liege, Belgium, 1981.

Zinc Compounds

Zinc Borates → Boric Oxide, Boric Acid, and Borates; Zinc Cyanide → Cyano Compounds, Inorganic; Zinc Oxide → Pigments, Inorganic; Zinc Sulfide → Luminescent Materials

DIETER M. M. ROHE, Grillo-Werke AG, Duisburg, Federal Republic of Germany (Chaps. 1, 2)

HANS UWE WOLF, Universität Ulm, Federal Republic of Germany (Chap. 3)

1. Zinc Halides 1
1.1. Zinc Fluoride 1
1.2. Zinc Chloride 1
1.3. Zinc Bromide 2
1.4. Zinc Iodide 3
2. Zinc Sulfate 3
3. Toxicology and Occupational Health . 3
4. References 6

1. Zinc Halides

1.1. Zinc Fluoride

Zinc fluoride [*7783-49-5*], ZnF_2, M_r 103.37 forms monoclinic tetragonal crystals with the rutile structure [1], *mp* 872 °C [2], ϱ_{25} 4.95 g/cm^3. The solubility of $ZnF_2 \cdot 4\,H_2O$ at 20 °C is only 1.62 g ZnF_2 in 1000 mL solution, but it is very soluble in aqueous ammonia. ZnF_2 reacts with sodium hydroxide solution to form flocculant precipitates of basic zinc fluorides.

Synthesis. $ZnF_2 \cdot 4\,H_2O$ is formed when zinc carbonate or zinc oxide reacts with aqueuos hydrofluoric acid. The water of crystallization can be driven off by heating to >200 °C and/or in a vacuum.

Analysis. Calcined commercial zinc fluoride has the typical analysis 61 – 63 % Zn, and 33 – 36 % F, with the following maximum impurity levels: Fe 500 ppm, Pb 100 ppm, and SO_4 100 ppm.

Uses. Zinc fluoride is used as an additive to electrolytic galvanizing baths, in glazes and enamels for porcelain, and special types of glass with a high refractive index. It is also used as a flux for welding and soldering and as a fluorinating agent in organic syntheses.

1.2. Zinc Chloride

Zinc chloride [*7646-85-7*], $ZnCl_2$, M_r 136.29 forms hexagonal-rhombic crystals, usually in the form of a white powder, *mp* 283 °C, *bp* 732 °C, ϱ 2.91 g/cm^3 at 25 °C, which is very soluble in water (432 g $ZnCl_2$ in 100 g water at 25 °C, 615 g $ZnCl_2$ in 100 g water at 100 °C). It is also soluble in alcohols, ether, acetone, ethyl acetate, glycerine, pyridine, amines, and nitriles.

Zinc chloride is strongly hygroscopic and forms hydrates with 1 – 4 mol water. It has a strong affinity for water, and causes many organic compounds to undergo condensation reactions. Considerable heat is evolved when $ZnCl_2$ dissolves in water.

Zinc chloride reacts with NH_4Cl to form the compounds $(NH_4)_2[ZnCl_4]$, $(NH_4)_3[ZnCl_5]$, and $(NH_4)_4[ZnCl_6]$. Analogous compounds are formed with amine hydrochlorides. Ammonia forms complexes with the composition $[Zn(NH_3)_xCl_2]$ ($x = 2, 4, 6$). Amines give analogous compounds. In water, zinc chloride behaves as an acid, forming aquoacids with the compositions $H[ZnCl_2OH]$ and $H_2[ZnCl_2(OH)_2]$. With increasing dilution, the aquoacid anions take up water and lose Cl^-, eventually being converted into the aquocation complex $[Zn(H_2O)_6]^{2+}$ and losing the strongly acid reaction.

Commercial zinc chloride often contains basic zinc salt which leads to cloudiness in aqueous media.

Production. High-purity zinc chloride is formed from zinc and HCl gas at 700 °C [3]. Spectroscopically pure zinc chloride can be obtained by decomposing zinc amalgam with HCl [4]. Quality specifications are listed in Table 1.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_537

Table 1. Quality specifications for zinc chloride

	Analytical-grade $ZnCl_2$ (ACS, ISO)	High-purity $ZnCl_2$ (DAB, PhEur, BP, JP, USP)
Assay (complexometric)	≥98 %	98 – 100.5 %
pH (10 % aq.)		4.6 – 5.5
SO_4	≤0.002 %	<0.01 %
Total N	≤0.001	<0.01 %
As		<0.0002 %
Ca	≤0.001 %	<0.01 %
Fe	≤0.0005 %	<0.001 %
K	≤0.001 %	0.001 %
NH_4	≤0.001 %	0.02 %
Na	≤0.001 %	
Oxychloride (as ZnO)	≦1.2 %	in accordance with test

Technical-grade zinc chloride is usually produced by leaching zinc oxides or zinc-containing waste such as zinc dross or ammonium chloride slags (unavoidable byproducts of hot-dip galvanizing) with hydrochloric acid. Raw materials containing a high percentage of NH_3 must be subjected to an additional treatment stage, i.e., passing chlorine gas into the aqueous suspension to oxidize the ammonia to nitrogen.

The zinc-containing raw materials are dissolved in concentrated hydrochloric acid. The crude liquor contains varying amounts of the elements Fe, Mn, Pb, Cu, and Cd. Fe^{2+} is oxidized to Fe^{3+} with H_2O_2 or $KMnO_4$, and $Fe(OH)_3$ is precipitated at pH 4 – 5. Also, the Mn is precipitated as MnO_2 by $KMnO_4$. The remaining elements Cu, Pb, and Cd are removed by treatment with zinc dust (cementation). The solution is then evaporated to dryness. Aqueous solutions are sometimes subjected to further processing.

Uses of zinc chloride:

1) Hydrochlorination of ethanol to monochloroethane

$$C_2H_5OH + HCl \rightarrow C_2H_5Cl + H_2O$$

2) Catalytic hydrolysis of benzotrichloride to benzoyl chloride
3) Catalyst in the production of ammonium methylene-bis-dithiocarbamate from hydrogen peroxide, aqueous ammonia, carbon disulfide, and ethylenediamine
4) Production of benzyl chloride from benzene, formaldehyde, and $HCl/ZnCl_2$
5) Filling material for batteries, emulsion breaker in petrochemical processes, deodorizing agent
6) Production of zinc compounds, e.g., zinc cyanide from potassium cyanide and a solution of zinc chloride (sometimes zinc sulfate)
7) Production of zinc soaps from alkali metal soaps and zinc chloride

1.3. Zinc Bromide

Zinc bromide [*7699-45-8*], $ZnBr_2$, M_r 225.21, forms colorless rhombic crystals, *mp* 394 °C, *bp* 650 °C, ϱ 4.29 g/cm^3, which are strongly hygroscopic. The dihydrate, $ZnBr_2 \cdot 2\,H_2O$, M_r 261.24, melts at 35 – 47 °C with loss of water of crystallization. It is obtained by dissolving an aqueous suspension of ZnO or metallic zinc in aqueous HBr. Anhydrous $ZnBr_2$ is formed from its elements at ca. 600 °C or by thermal dehydration of the dihydrate in a stream of dry CO_2. Its solubility in water at 25 °C is 820 g/L.

Uses of zinc bromide:

1) A mild Lewis acid for alkylation reactions [5]
2) A catalyst for intramolecular reactions [6]
3) An activating agent in the production of activated carbon
4) An additive to photographic gelatins
5) An electrolyte for zinc bromide batteries

High-purity $ZnBr_2$ contains ≥98 % $ZnBr_2$, with the following maximum impurity contents: Ba 10 ppm, Ca 100 ppm, Fe 10 ppm, K 100 ppm, Mg 100 ppm, and Na 100 ppm.

1.4. Zinc Iodide

Zinc iodide [*10139-47-6*], ZnI_2, M_r 319.22, forms white to yellowish cubic crystals, *mp* 446 °C, *bp* 624 °C, ϱ_{25} 4.73 g/cm^3. It is produced by adding metallic zinc to constant-boiling aqueous hydriodic acid until the yellow color disappears, by neutralizing aqueous HI with ZnO in the presence of a catalyst (precipitated silver), or by the reaction of I_2 vapor with molten zinc at 600 °C in a stream of N_2

[7]. Its solubility in water at 20 °C is 4500 g/L (hygroscopic).

Zinc iodide is used in medicine as a topical antiseptic astringent and in the preparation of iodide – zinc – starch solution.

2. Zinc Sulfate

Zinc sulfate [*7733-02-0*], $ZnSO_4$, M_r 161.44, ϱ 4.33 g/cm^3, forms rhombic crystals. On heating to >650 °C, zinc sulfate undergoes considerable decomposition with formation of basic sulfates. Above 800 °C, it decomposes completely into ZnO and SO_3.

$ZnSO_4$ is very soluble in water, and forms hydrates with 1, 2, 4, 6, and 7 molecules of water of crystallization. The hepta-, hexa-, and monohydrates are produced on an industrial scale.

Zinc sulfate heptahydrate [*7746-20-0*], $ZnSO_4 \cdot 7\,H_2O$, M_r 287.55, forms white, lustrous, columnar, rhombic crystals. It is transformed into the hexahydrate at 39 °C. It occurs in nature as goslarite. Mixed crystals are formed with the sulfates of Fe, Mg, Mn, Co, and Ni. It is very soluble in glycerine, but only sparingly soluble in other organic solvents.

Zinc sulfate is obtained by leaching zinc-containing materials with sulfuric acid in an aqueous medium. Suitable materials include zinc dross, slags, sweepings, oxides, sludges, metallic zinc, and zinc-containing flue dusts [8]. Used sulfuric acid can be utilized. It should be noted that arsine can be formed during the treatment of metallic dross and waste. Iron, manganese, and heavy metals are removed by the same process as that used to treat electrolyte liquors.

The purified solution of zinc sulfate is evaporated until its density reaches ca. 1.4 at 80 °C. On cooling, crystals of zinc sulfate heptahydrate form and are centrifuged off. The crystals obtained are sometimes dried at ca. 30 °C to prevent efflorescence.

Pharmaceutical-grade $ZnSO_4 \cdot 7\,H_2O$ can be produced from sulfuric acid and pharmaceutical-grade zinc oxide. The impurity levels in the various grades of zinc sulfate are listed in Table 2.

Zinc sulfate monohydrate [*7446-19-7*], $ZnSO_4 \cdot H_2O$, is produced by thermal dehydration or by dehydrating with 95 % ethanol. It crystallizes from a zinc sulfate mother liquor at temperatures above 70 °C [9]. It can be dried in a rotary, spray, or fluidized-bed driers.

Uses of zinc sulfate can be divided into direct and indirect uses.

Direct uses:

1) Precipitating baths for viscose manufacture
2) Electrolytes for galvanizing baths
3) Additive (trace element) in fertilizers and animal feeds
4) In medicine, as an emetic, astringent, or disinfectant
5) Wood preservative, additive for paper bleaching, flocculant
6) Water treatment

Indirect uses:

1) Starting material for the production of other zinc compounds such as the fungicides Zineb and Mancozeb (→ Fungicides, Agricultural, Chap. 6.3.), zinc soaps (→ Metallic Soaps, Chap. 5.10.), and the antidandruff agent zinc pyrithione
2) Production of lithopone and zinc sulfide pigments
3) Production of zinc cyanamide from potassium cyanamide
4) Production of zinc phosphides

3. Toxicology and Occupational Health

Zinc metal and most of its compounds show very low toxicity compared to most other heavy metals. The toxicity of zinc salts varies to a certain extent, being dependent primarily on the toxicity of the anionic part of the compound. For example, the high toxicity and carcinogenicity of zinc chromate, $ZnCrO_4$, is not due to Zn^{2+}, but to the anionic CrO_4^{2-} component.

Intoxications with zinc and its compounds are comparatively rare. They can occur by ingestion of food contaminated with zinc leached from food containers or by inhalation of zinc or zinc oxide dust under occupational conditions (see below).

On the other hand, zinc is one of the most important essential elements for humans and all

Table 2. Commercial grades of zinc sulfate

	Analytical-grade $ZnSO_4 \cdot 7\,H_2O$	High-purity $ZnSO_4 \cdot 7\,H_2O$ (DAB, USP, PhEur, FCC)	Pure $ZnSO_4 \cdot 7\,H_2O$
Assay	≤99.5 %	99.5 %	98 %
pH (5 %, 20 °C)	4.4 – 5.6	4.4 – 5.6	
As	0.5 ppm	2 ppm	10 ppm
Ca	10 ppm	50 ppm	
Cd	2 ppm	2 ppm	30 ppm
Cu	5 ppm	1 ppm	10 ppm
Fe	5 ppm	10 ppm	30 ppm
K	10 ppm	100 ppm	
Mg	10 ppm	50 ppm	
Mn	2 ppm	10 ppm	50 ppm
Na	50 ppm	1000 ppm	
Pb	5 ppm	5 ppm	50 ppm
Chloride	5 ppm	100 ppm	4000 ppm
Hg		1 ppm	
Se		5 ppm	

forms of plant and animal life. It is necessary for growth, skin integrity and function, testicular maturation, immunocompetence, wound healing, and for a variety of metabolic processes including carbohydrate, lipid, protein and nucleic acid synthesis or degradation. It is an essential, coenzyme-like component for the function of more than 70 metallo-enzymes, including alcohol dehydrogenase, carbonic anhydrase, and carboxypeptidase [10]. Physiological amounts of zinc have been shown to decrease the toxicity of other heavy metal ions such as Cd, Hg, Pb, and Sn [11].

Biological significance, biochemistry, toxicokinetics, human and animal toxicity, levels of tolerance, detoxification, evaluation of health risks to humans, and ecotoxicity of zinc and its compounds have been reviewed [11–14].

Toxicokinetics. As an essential element, zinc is absorbed in an active transport process, regulated by dietary zinc status and occurring primarily in the ileum [15]. Zinc is bound to proteins to a large extent. Excretion occurs predominantly via the feces, and to a minor extent via urine and sweat. Excretion in the urine is up to 12.2 mmol/d; higher values indicate zinc intoxication [16].

Zinc is the most prevalent metal ion in human tissues other than blood. The body of a 70-kg man contains ca. 2.3 g of zinc [17], 64 % of which is found in muscle and 28 % in bone [18]. The highest zinc concentrations are found in reproductive organs, especially the prostate gland (87 µg/g wt weight), whereas the whole-body average is 33 µg/g wet weight [17]. Serum contains 10.7 – 22.9 mmol/L, plasma 17.4 ± 1.8 mmol/L, erythrocytes 184 – 198 mmol/L, and hair 3.30 ± 1.33 mmol/g [16].

Acute Toxicity. In contrast to numerous other heavy metals, zinc shows low acute toxicity. Zinc overload of the organism is rather unlikely and occurs almost exclusively by ingestion when food or drinks, especially acidic ones, are prepared or stored in galvanized containers [19], [20].

Symptoms of acute intoxication are gastrointestinal distress, diarrhea, nausea, and abdominal pain. Vomiting occurs almost universally. In elemental-zinc overdose, symptoms such as lethargy, slight ataxia, and difficulty in writing also appear. Abnormal laboratory parameters include increase of pancreatic enzyme activities, hypokalemia secondary to acute pancreatitis, and microhematuria [21], [22].

Zinc chloride, $ZnCl_2$, is caustic to the gastrointestinal tract, occasionally leading to hematemesis, beside the symptoms mentioned above. Furthermore, it is extremely detrimental to the lungs. An episode of pulmonary exposure to $ZnCl_2$ smoke resulted in ten fatalities due to pulmonary toxicity. Most survivers showed lacrimation, rhinitis, dyspnea, stridor, and retrosternal chest pain. In two autopsies, severe inflammation of the upper respiratory tract mucosa and pulmonary edema, but no changes in liver and kidneys, could be detected [23].

Zinc or zinc oxide fume is responsible for a syndrome which may occur in welding, galvanizing, or smelting operations, called "metal fume fever". Symptoms such as rapid breathing, dyspnea, cough, fever, shivering, sweating, chest and leg pain, myalgias, fatigue, metallic taste, salivation, thirst, and leukocytosis appear within 4 – 6 h after exposure and disappear after 24 – 48 h [24], [25]. The X-ray of the chest is normally clear. A rapid tolerance (tachyphylaxis, "quick immunity" [26]) develops in workers within 48 h, but may the lost over the weekend ("monday morning fever"). Metal fume fever can also be caused by exposure to fumes of aluminum, antimony, copper, iron, magnesium, manganese, and nickel.

Zinc phosphide, Zn_3P_2, is extremely toxic because of the formation of phosphine, PH_3, on reaction with water or acids.

Chronic Toxicity. The development of a toxicosis has been reported in the case of a continuous uptake of drinking water contaminated with 40 ppm Zn from a galvanized pipe [27]. Dialysis patients are at risk of an elevation of the plasma concentration of zinc (and other heavy metal ions) [28], [29]. The symptoms are similar to those of acute toxicity, but are less pronounced.

Genotoxicity. Zinc apparently does not show any mutagenicity or carcinogenicity [11]. In human lymphocytes, 0.1 – 1000 µmol/L of zinc acetate, alone or in combination with cadmium and lead acetates, failed to induce chromatid-type aberrations (deletions, acentric fragments) and gaps [30]. The carcinogenic effect of zinc chromate, $ZnCrO_4$, is due to the chromate anion, not to the zinc cation.

Reproductive Toxicity. There are no indications for reproductive effects, especially for teratogenicity, of zinc [11].

Immunotoxicity. There are no reports concerning immunotoxic effects of zinc. However, zinc chromate, $ZnCrO_4$, shows strong sensitizing effects and therefore is able to cause obstinate eczema under occupational conditions [31].

Toxicological Data. The lethal dose of $ZnCl_2$ in humans is 3 – 5 g. Since zinc sulfate, $ZnSO_4$, is less caustic than the chloride, it has lower toxicity [32].

Treatment. Decontamination of the gastrointestinal tract after oral uptake of zinc compounds by syrup of ipecac, lavage, or cathartics is mostly unnecessary, since patients usually vomit sufficiently. For normalization of zinc levels, $CaNa_2EDTA$ is the agent of choice [21], [23]. After oral uptake of $ZnCl_2$, milk may be administered to decrease absorption of the metal [34].

After contact of the skin with $ZnCl_2$, immediate removal is necessary using soap and plenty of water.

After contact of the eyes with $ZnCl_2$, adequate measures are rinsing with plenty of water, use of Isogutt eye drops, and contacting an ophthalmologist as soon as possible [34], [35].

In cases of $ZnCl_2$ fume inhalation, cortisone preparations should be applied immediately (e.g., by inhalation of Auxiloson) to avoid development of lung edema [35]. In severe cases clinical observation for at least 24 h is necessary.

Occupational Health. The MAK value of zinc oxide, ZnO, is 5 mg ZnO/m^3, measured as total fine dust [36]. The same TLV value has been set in Hungary, Japan, Poland, Sweden, and by the WHO, while it is 10 mg ZnO/m^3 in Bulgaria, Romania, and Pennsylvania (United States); it is 15 mg ZnO/m^3 in Finland. The United States has a TLV value of 1 mg $ZnCl_2$/m^3, and Massachusetts a 30-min limit of 0.2 mg/m^3 zinc chromate, $ZnCrO_4$ [17].

Zinc chromate, $ZnCrO_4$ is classified in group A 1 (proven human carcinogens) of Appendix III of the MAK list [36].

4. References

1. W. Zachariasen, *Skr. Akad. Oslo* **1** (1926) 7.
2. *Handbook of Chemistry and Physics,* 54th ed., CRC Press, Cleveland, Ohio 1973/1974, B-154, B-155.
3. O. Hönigschmid, M. von Mack, *Z. Anorg. Allg. Chem.* **246** (1941) 363.
4. R. T. Hamilton, J. A. von Butler, *J. Chem. Soc. (London)* 1932, 2283.
5. I. Paterson, *Tetrahedron Lett.* 1979, 1519.

6. L. F. Tietze, K. Biefuss, *Angew. Chem. Int. Ed. Engl.* **24** (1985) 1042.
7. M. Sato, M. Yosiyana, *Nippon Kagaku Kaishi* **60** (1939) 918.
8. F. Ensslin, *Met. Erz* **38** (1941) 196.
9. *Gmelin,* 8th ed., Zinc, p. 942.
10. B. L. Vallee, *BioFactors* **1** (1988) no. 1, 31 – 36.
11. A. Léonhard, G. B. Gerber, F. Léonhard, *Mutat. Res.* **168** (1986) 343 – 353.
12. R. L. Bertholf: "Zinc," in H. G. Seiler, H. Sigel, A. Sigel (eds.): Handbook on Toxicity of Inorganic Compounds, Marcel Dekker, New York 1988, pp. 787 – 800.
13. A. B. Abdel-Mageed, F. W. Oehme, *Vet. Hum. Toxicol.* **32** (1990) no. 1, 34 – 39.
14. R. S. Bedwal, N. Nair, R. S. Mathur, *Trace Elem. Med.* **8** (1991) no. 2, 89 – 100.
15. D. L. Antonson, A. J. Barak, J. A. Vanderhoof, *J. Nutr.* **109** (1979) 142 – 147.
16. N. S. Tietz in: *Clinical Guides to Laboratory Tests,* W. B. Saunders, Philadelphia 1983, pp. 516 – 519.
17. National Research Council, Subcommittee on Zinc: *Zinc,* University Park Press, Baltimore 1979, pp. 249 –271.
18. H. Tipton, M. J. Cook, *Health Phys.* **9** (1963) 103 –145.
19. G. R. Callendar, C. J. Genzkow, *Mil. Surg.* **80** (1937) 67 – 72.
20. M. A. Brown et al., *Arch. Environ. Health* **8** (1964) 657 – 660.
21. S. J. Chobasanian, *Ann. Emerg. Med.* **10** (1981) 91 – 93.
22. J. V. Murphy, *J. Am. Med. Assoc.* **212** (1970) 2119 –2120.
23. E. H. Evans, *Lancet* **2** (1945) 368 – 370.
24. M. J. Ellenhorn, D. G. Barceloux: *Medical Toxicology – Diagnosis and Treatment of Human Poisoning,* Elsevier, New York 1988, pp. 1064 – 1065.
25. H. E. Stokinger: "The Metals. Zinc," in G. D. Clayton, F. E. Clayton (eds.): *Patty's Industrial Hygiene and Toxicology,* 3rd revised ed.,**vol. 2 A**, John Wiley & Sons, New York 1981.
26. C. P. McCord, *Ind. Med. Surg.* **29** (1960) 101 – 107.
27. G. Laurence, *Br. Med. J.* **1** (1958) 582 – 587.
28. J. Savory, R. L. Bertholf, M. R. Wills, *Clin. Endocrinol. Metab.* **14** (1985) 681 – 702.
29. E. D. M. Gallery, J. Blomfield, S. R. Dixon, *Br. Med. J.* **4** (1972) 331 – 333.
30. K. Gasiorek, M. Bauchinger, *Environ. Mutagen.* **3** (1981) 513 – 518.
31. L. C. Robis, *New Engl. J. Med.* **260** (1969) 1091 – 1098.
32. W. Wirth, Chr. Gloxhuber: *Toxikologie,* 5th ed., Georg Thieme Verlag, Stuttgart 1994, pp. 126 – 128.
33. J. L. Potter, *Ann. Emerg. Med.* **10** (1981) 267 – 269.
34. R. Kühn, K. Birett: *Merkblätter Gefährliche Arbeitsstoffe,* 4th ed., ecomed verlag, Landsberg/Lech 1979, Datenblatt Z-004.
35. S. Risi, H. U. Wolf: "Zink," in H. U. Wolf (ed.): *HAGERs Handbuch der Pharmazeutischen Praxis,***vol. 3**, Gifte, Springer Verlag, Berlin 1992, pp. 1257 –1259.
36. Deutsche Forschungsgemeinschaft (ed.): *MAK- und BAT-Werte-Liste 1995,* Maximale Arbeitsplatzkonzentrationen und Biologische Arbeitsstofftoleranzwerte, Mitteilungen der Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, Mitteilung 31, VCH-Verlagsgesellschaft mbH, Weinheim 1995.

Zinc White → Pigments, Inorganic

Zirconium and Zirconium Compounds

RALPH NIELSEN, Teledyne Wah Chang, Albany, Oregon 97321, United States

1. Zirconium 2
1.1. Introduction 2
1.2. Properties 2
1.3. Occurrence, Ores, and Mining . . . 4
1.4. Production 5
1.4.1. Ore Cracking 5
1.4.2. Hafnium Separation 7
1.4.3. Reduction to Metal 9
1.4.4. Refining 10
1.5. Fabrication 11
1.6. Hazards in Handling Zirconium Metal 12
1.7. Grades and Specifications 13
1.8. Uses 13
1.9. Economic Aspects 14
2. Zirconium Compounds 15
2.1. Zirconium Silicate 15
2.2. Zirconium Oxide 16
2.3. Hydrous Zirconium Oxide 17
2.4. Zirconium Chlorides 17
2.5. Zirconium Hydroxide Chlorides . . 18
2.6. Zirconium Bromides 18
2.7. Zirconium Iodides 19
2.8. Zirconium Fluorides 19
2.9. Zirconium Sulfates 19
2.10. Zirconium Carbonates 20
2.11. Zirconium Hydride 20
2.12. Zirconium Carbide 21
2.13. Zirconium Nitrides 21
2.14. Zirconium Borides 22
2.15. Zirconium Tetrahydridoborate . . . 22
2.16. Zirconium Tungstate 22
2.17. Zirconium Phosphate 22
2.18. Zirconium Nitrates 23
2.19. Zirconium Carboxylates 23
2.20. Zirconium Alkoxides 23
3. Analysis 23
4. Toxicology 24
5. Storage and Transportation 24
6. References 25

1. Zirconium

1.1. Introduction

On analyzing the gem mineral known as jargon from Sri Lanka, KLAPROTH in 1789 found it contained an oxide which he could not identify and which he later called " zirconerde". In 1824 BERZELIUS prepared the first zirconium metal, an impure black powder, by heating potassium hexafluorozirconate and potassium in a closed pot. LELY and HAMBURGER prepared the first relatively pure zirconium in 1904 by reduction of zirconium tetrachloride with sodium in a bomb, producing malleable, corrosion-resistant zirconium pellets. But it was a hundred years after BERZELIUS before VAN ARKEL, DE BOER, and FAST [1], [2] developed the iodide process which produced the first massive zirconium metal that could be cold worked and which exhibited good ductility at room temperature.

Early uses for zirconium metal were in pyrophoric devices such as fuses, ammunition primers, fireworks, photoflash powder, and flares. Zirconium powder was used extensively as a gettering material in vacuum tubes. By 1946 the consumption of zirconium for these purposes was about 5 t/a. Less than 50 kg of ductile iodide metal was sold, at ca. $ 700 per kilogram.

W. J. KROLL carried out the first inert-atmosphere magnesium reduction of zirconium tetrachloride in his Luxembourg laboratory in 1938 [3]. In 1944 the U.S. Bureau of Mines started a project to make ductile zirconium economically. KROLL was assigned to take over direction of the project in January, 1945. By 1947, a pilot plant was producing 30 kg of zirconium sponge per week. Concurrently, researchers at Massachusetts Institute of Technology and at Oak Ridge National Laboratory were evaluating the physical and atomic properties of metals as potential uranium fuel cladding materials for nuclear power stations. In 1948 hafnium-free zirconium was selected as most promising. By 1949 zirconium had been chosen as the structural material for the fuel core of the submarine thermal reactor, the land-based reactor prototype, and during 1949 – 1950 a satisfactory hafnium separation process was developed at Oak Ridge. By

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.a28_543

the end of 1950, 3000 kg of hafnium-free zirconium metal strip had been produced, and the zirconium metal producing industry was launched as an integral part of the beginning nuclear power industry.

1.2. Properties

Physical Properties. Zirconium [*7440-67-7*] is a lustrous, strong, ductile metal similar in appearance to stainless steel. Pure zirconium metal has three solid phases: ω-Zr, which is stable below 200 K at ambient pressure and up to 1000 K at 6 GPa [4]; α-Zr, which is stable from 200 K to 1125 K; and β-Zr, which is the stable form between 1125 K and the melting point. These transition temperatures change with the addition of α- or β-stabilizing alloying elements. The important physical properties are presented in the following:

Relative atomic mass	91.224
Atomic radius	15.90 nm (in metal lattice)
Ionic radius (Zr^{4+})	7.5 nm
Electronegativity	1.22
Standard potential M/MO_2	1.53 V
Melting point	$1852 \pm 2\,^{\circ}C$
Boiling point	$3850\,^{\circ}C$
Crystal structure	
ω-Zr	hexagonal open
α-Zr	hexagonal dense, a = 32.3 mm, c = 51.5 nm
β-Zr	body-centered cubic, a = 36.1 nm
Transformation temperatures	
$\omega \rightarrow \alpha$	$-73\,^{\circ}C$
$\alpha \rightarrow \beta$	$862 \pm 5\,^{\circ}C$
Heat of fusion	2.30×10^4 J/mol
Heat of evaporation	5.96×10^5 J/mol
Electrical resistivity	$3.89 \times 10^{-5}\ \Omega \cdot cm$
Thermal conductivity	
at 25 °C	$21.1\ W\,m^{-1}\,K^{-1}$
at 100 °C	$20.4\ W\,m^{-1}\,K^{-1}$
at 300 °C	$18.7\ W\,m^{-1}\,K^{-1}$
Specific heat	
at 25 °C	$0.285\ J\,g^{-1}\,kg^{-1}$
at 865 °C	$0.335\ J\,g^{-1}\,kg^{-1}$
Thermal expansion coefficient	
α, bulk, at 25 °C	$5.89 \times 10^{-6}\ K^{-1}$
α-Zr, parallel to the C axis	$6.4 \times 10^{-6}\ K^{-1}$
α-Zr perpendicular to C axis	$5.6 \times 10^{-6}\ K^{-1}$
β-Zr	$9.7 \times 10^{-6}\ K^{-1}$
Density	
α-Zr	6.50 g/cm
β-Zr	6.05 g/cm
Effective cross section for thermal neutrons	$1.9 \times 10^{-29}\ m^2$ (0.19 barns)

Mechanical Properties. Although zirconium is a high-melting metal, its mechanical properties are similar to those of much lower melting metals: its elastic modulus is quite low, and its strength diminishes rapidly with increasing temperature. The mechanical properties of zirconium are strongly dependent on purity, especially the oxygen and nitrogen content, the amount of cold work, and the crystallographic texture. Forged or rolled zirconium shows marked anisotropy in mechanical properties in both the as-formed and annealed conditions. Pure zirconium is ductile at liquid-nitrogen temperatures but the metal is subject to hydrogen embrittlement whenever the hydrogen content sufficiently exceeds the hydrogen solubility at the operating temperature. Tables 1 and 2 gives typical values of the more important mechanical properties of zirconium and of Zircaloy 2 and Zircaloy 4 (see also Section 1.9).

Chemical Properties. Zirconium is a very reactive metal that, in air or aqueous solution, immediately develops a surface oxide film. This stable, adherent film is the basis for zirconium's corrosion resistance. In most media, zirconium is more resistant than titanium or stainless steel. Its acid resistance approaches that of tantalum, but unlike tantalum, zirconium is also resistant to caustic media.

Fluoride ions bond strongly to zirconium, and even a trace of fluoride in most media will drastically reduce its corrosion resistance.

The corrosion resistance in sulfuric acid is excellent at temperatures up to 150 °C and acid concentrations up to 70 %, except that welds and adjacent heat-affected zones are susceptible to corrosive attack above 60 % concentration. This can be alleviated by annealing. Ferric, cupric, and nitrate ion impurities in sulfuric acid decrease the corrosion resistance at acid concentrations above 65 %. Even at room temperature zirconium is rapidly attacked at concentrations above 75 %.

The corrosion resistance in hydrochloric acid is excellent at temperatures up to 130 °C and concentrations up to 37 %. However, even small amounts of ferric or cupric ions will lead to severe pitting and stress cracking.

The corrosion resistance in nitric acid is excellent at all concentrations up to 90 % and tem-

Table 1. Mechanical properties of zirconium

Property	20 °C	200 °C	300 °C
Tensile strength, MPa	300 – 450	220 – 250	150 – 180
Elastic limit (0.2 %), MPa	200 – 300	110 – 130	75 – 85
Elongation, %	25 – 35	50 – 60	45 – 55
Modulus of elasticity, MPa	9.5×10^4	8.3×10^4	7.2×10^4
Brinell hardness	90 – 130		
Shear modulus	36 500		
Poisson's ratio	0.35		

Table 2. Mechanical properties of Zircaloy 2 and Zircaloy 4

Property	Annealed			Recrystallized		
	20 °C	200 °C	400 °C	20 °C	200 °C	400 °C
Tensile strength, MPa	785	440	380	520	290	190
0.2 % yield strength, MPa	600	360	300	370	150	130
Ultimate elongation (50 mm), %	17	20	20	34	45	55

peratures up to 200 °C. In concentrations above 65 %, stress corrosion cracking may occur if high tensile stresses are present [5].

The corrosion resistance in phosphoric acid is excellent in all concentrations up to 65 °C, and, at concentrations below 40 %, up to 185 °C.

Zirconium is rapidly attacked by hydrofluoric acid, even at concentrations below 0.1 %.

The corrosion resistance in sodium hydroxide and potassium hydroxide is excellent at all concentrations, up to the boiling temperature. Zirconium is resistant to molten sodium hydroxide up to 1000 °C.

Zirconium is very resistant to most organic compounds at all concentrations and temperatures. But, when air or moisture are not available to reform the surface oxide film, zirconium is attacked by anhydrous chlorinated organics at elevated temperatures and is etched by bromine or iodine dissolved in anhydrous organics. Stress corrosion cracking may also occur.

Zirconium reacts with most gases at relatively low temperatures.

Zirconium forms a visible interference color film in air starting above 200 °C. Above 400 °C the adherent oxide film is black. Above 540 °C a loose gray-white oxide forms, the oxide layer becomes thicker as oxygen diffuses into the underlying metal, and the metal can become embrittled after prolonged exposure. The maximum continuous operating temperature of zirconium in air is ca. 450 °C. Short-term hot working such as forging, rolling, and extrusion are conducted at 550 – 1000 °C. Afterwards, the surface oxide and embrittled metal layer are mechanically and chemically removed. While solid zirconium metal is oxidized slowly in air, the oxidation is exothermic, and for high-surface-area forms, such as powder or sponge metal, the large heat release may cause ignition (see Section 1.8).

Oxygen-containing gases, such as carbon dioxide, carbon monoxide, sulfur dioxide, steam, and nitrogen oxides, oxidize zirconium somewhat slower than air. The maximum continuous operating temperature is about 400 °C.

Zirconium reacts more slowly with nitrogen than with oxygen because the reaction is less exothermic.

Zirconium readily absorbs hydrogen above 300 °C. The rate of absorption varies inversely with the thickness of the surface oxide film. Hydrogen solubility increases with temperature. On cooling, when the solubility is exceeded, zirconium hydride platelets precipitate, embrittling the metal. The reaction is reversible: hydrogen is removed on heating the metal above 600 °C in a high vacuum.

Zirconium's oxide film protects it from dry chlorine gas at room temperature, but any impact or abrasion that exposes a fresh metal surface may initiate an exothermic reaction producing zirconium tetrachloride vapor.

More information on the chemical behavior of zirconium metal can be found in [6].

1.3. Occurrence, Ores, and Mining

Zirconium is the ninth most abundant metal in the earth's crust, with an estimated concentration of 0.016 – 0.025 wt % [7], about one twentieth that of titanium but more plentiful than nickel, copper, or zinc. Zirconium occurs only in fully oxidized form, never as free metal.

There are over 40 known zirconium minerals which can be grouped:

1) Zirconium orthosilicates: Zircon and its metamict varieties
2) Zirconium dioxide: baddeleyite and its altered varieties
3) Zirconsilicates with sodium, calcium, iron, and other elements: eudialyte, eucolite, gittinsite and others
4) Zirconium carbonates with sodium, calcium, and other elements: weloganite and others
5) Others, including zirconolite ($CaZrTi_2O_7$)

Zircon is the predominant commercial zirconium mineral, but the minerals baddeleyite and eudialyte are also being used. All are obtained as byproducts of other mineral recovery operations.

Zircon [*1490-68-2*], $ZrSiO_4$, occurs as an accessory mineral in silica-rich igneous rocks, particularly granite, nepheline syenite, and pegmatite, and also in metamorphic and sedimentary rocks. Zircon is rarely found in rocks in economically minable concentration. Weathering and erosion of these rocks frees the zircon grains, and the combined action of rivers, seas, and wind concentrate the heavier minerals by natural gravitation processes in placer deposits, deltas, and ocean beaches. As an ore, zircon is recovered from unconsolidated sands in beach deposits.

Large heavy mineral sands deposits are being extensively mined in Florida, West Australia, South Africa, India, Russia and Kazakhstan, with smaller operations in Sierra Leone, Sri Lanka, Madagascar, Malaysia, Brazil, Indonesia, Thailand, Ukraine, and Vietnam. In Canada, some zircon is obtained from the processing of Athabasca tar sands. Undeveloped heavy mineral beach sand reserves containing zircon exist in New Jersey, the Carolinas, Georgia, Tennessee, Colorado, New Mexico, Oregon, Utah, Wyoming, Egypt, Malawi, Senegal, and Tanzania.

These heavy mineral sands deposits are mined to obtain the titanium minerals rutile, leucoxene, and ilmenite to supply the titanium metal and titanium oxide pigment industries. Depending on the deposit's composition and current market pricing, zircon and other minerals are viable byproducts. Australia and South Africa are major zircon exporters.

Mining of heavy mineral sands involves first removing a light sand overburden, followed by removal of the heavy sands layers using elevating scrapers and bulldozers, or by flooding the excavation and using dredges in the pond. The heavy minerals are concentrated by gravity separation, and the quartz, light minerals, fines, clay, and silt are returned to the back end of the excavation. Since the heavy mineral sands are only 4 – 7 % of the deposit, the site can be rehabilitated close to the original elevation, topsoil replaced, and native vegetation replanted. The concentrate is hauled to a ore-dressing site where the grains may be scrubbed to remove surface coatings, dried, and separated into individual mineral components: rutile, leucoxene, ilmenite, zircon, xenotime, monazite, staurolite, garnet, kyanite, and sillimanite. These separation steps utilize differences in the specific gravity, induced magnetism, and electrical conductivity of the individual minerals grains to gradually recover each mineral in good purity.

The various mining operations produce a range of zircon products from high grades with very low levels of impurities to lower grades where the impurities vary with the nature of the orebody and the type of separation process used. These various grades supply a range of applications: foundry sands, refractories, abrasives, opacifiers, zirconium chemicals, zirconium metal, and welding rod coatings. A typical analysis of zircon for the metal industry is 66.6 % $(Zr + Hf)O_2$, 0.2 % Al_2O_3, 0.15 % TiO_2, 0.1 % Fe_2O_3, 0.1 % P_2O_5, 0.025 % U, 0.020 % Th.

Baddeleyite [*12036-23-6*], ZrO_2, has been found in Brazil, South Africa, Sri Lanka, and Russia. Brazilian baddeleyite often occurs mixed with zircon. The mixture is known as caldasite and usually contains 65 – 80 % zirconium oxide. South African baddeleyite is a byprod-

uct of mining a volcanic orebody for copper and phosphate fertilizer. The baddeleyite concentrate contains 96 % zirconium oxide. Russian baddeleyite from the Kola Peninsula is a byproduct of open pit iron ore mining.

Eudialyte [*12173-26-1*], $(Ca, Na_2)_6(Zr, Fe)(Si_3O_9)_2(OH, Cl)_2$ has been found in Greenland, Norway, Brazil, Australia, Transvaal, and New Mexico. Some eudialytes contain 12 – 15 % zirconium oxide, yttrium, and small amounts of niobium and rare earth metals.

Hafnium Content. All known zirconium minerals contain hafnium, usually in the range of 1.5 – 2.5 % Hf/Zr + Hf, although higher hafnium contents have been found. There is a tendency for higher-hafnium ores to be also higher in uranium and thorium content. Hafnon, hafnium orthosilicate, with >95 % Hf/Zr + Hf was found associated with a tantalum ore in Mozambique.

1.4. Production

Processing steps for the production of zirconium metal from zircon sand are shown in Figure 1.

1.4.1. Ore Cracking

Zircon is a refractory mineral whose decomposition requires the use of high temperature and aggressive chemicals. Selection of a particular procedure depends on the subsequent product use or processing.

Caustic Fusion. Caustic fusion of zircon with a slight excess of sodium hydroxide at 650 °C (or sodium carbonate at 1000 °C) is the usual zircon decomposition process [8], [9]:

$$ZrSiO_4 + 4\,NaOH \rightarrow Na_2ZrO_3 + Na_2SiO_3 + 2\,H_2O$$

The cooled reaction mass is crushed and slurried in water. Water dissolves the sodium silicate and hydrolyzes the sodium zirconate to soluble sodium hydroxide and insoluble hydrous zirconia. The hydrous zirconia, recovered by filtration, can be fired to oxide, dissolved in mineral acid for further conversion to aqueous zirconium compounds, or dissolved in mineral acid for feeding to a hafnium extraction process.

A variant of the caustic fusion process uses less caustic to produce an acid-soluble sodium zirconium silicate

$$ZrSiO_4 + Na_2CO_3 \xrightarrow{1000\,^\circ C} Na_2ZrSiO_5 + CO_2$$

The frit from this reaction is added to strong acid to yield a solution containing sodium and zirconium salts and silica gel:

$$Na_2ZrSiO_5 + 4\,HCl \rightarrow 2\,NaCl + ZrOCl_2 + SiO_2 + 2\,H_2O$$

Chlorination. Fluidized-bed carbochlorination of milled zircon sand is the zircon decomposition process used by American and French zirconium metal producers. Chlorine is the fluidizing gas. The reaction is endothermic and supplemental energy is supplied, usually by induction heating of the interior graphite wall of the chlorinator.

$$ZrSiO_4 + 4\,C + 4\,Cl_2 \xrightarrow{1100\,^\circ C} ZrCl_4 + SiCl_4 + 4\,CO$$

The product gases are cooled to 200 °C in a primary condenser that collects zirconium (and hafnium) tetrachloride as a powder. The remaining gas mixture is cooled to −20 °C in a secondary condenser, collecting silicon tetrachloride as a liquid which is subsequently purified and used to make fumed silica, fused quartz, and fused-quartz optical fiber.

Thermal Dissociation. Thermal dissociation of zircon in an arc plasma forms zirconium oxide in droplets of liquid silica. Rapid quenching, to minimize recombination, produces intimately mixed crystals of zirconium oxide in a bead of amorphous silicon oxide [10].

$$ZrSiO_4 \xrightarrow{1800\,^\circ C} \text{liquid} \xrightarrow{\text{Quench}} ZrO_2 + SiO_2$$

The mixture can be leached with sulfuric acid to yield a zirconium sulfate solution and insoluble silica, or leached with sodium hydroxide to yield a sodium silicate solution and insoluble zirconium oxide.

Fluorosilicate Fusion. The fusion of zircon and potassium hexafluorosilicate produces potassium hexafluorozirconate [11].

$$ZrSiO_4 + K_2SiF_6 \xrightarrow{700\,^\circ C} K_2ZrF_6 + 2\,SiO_2$$

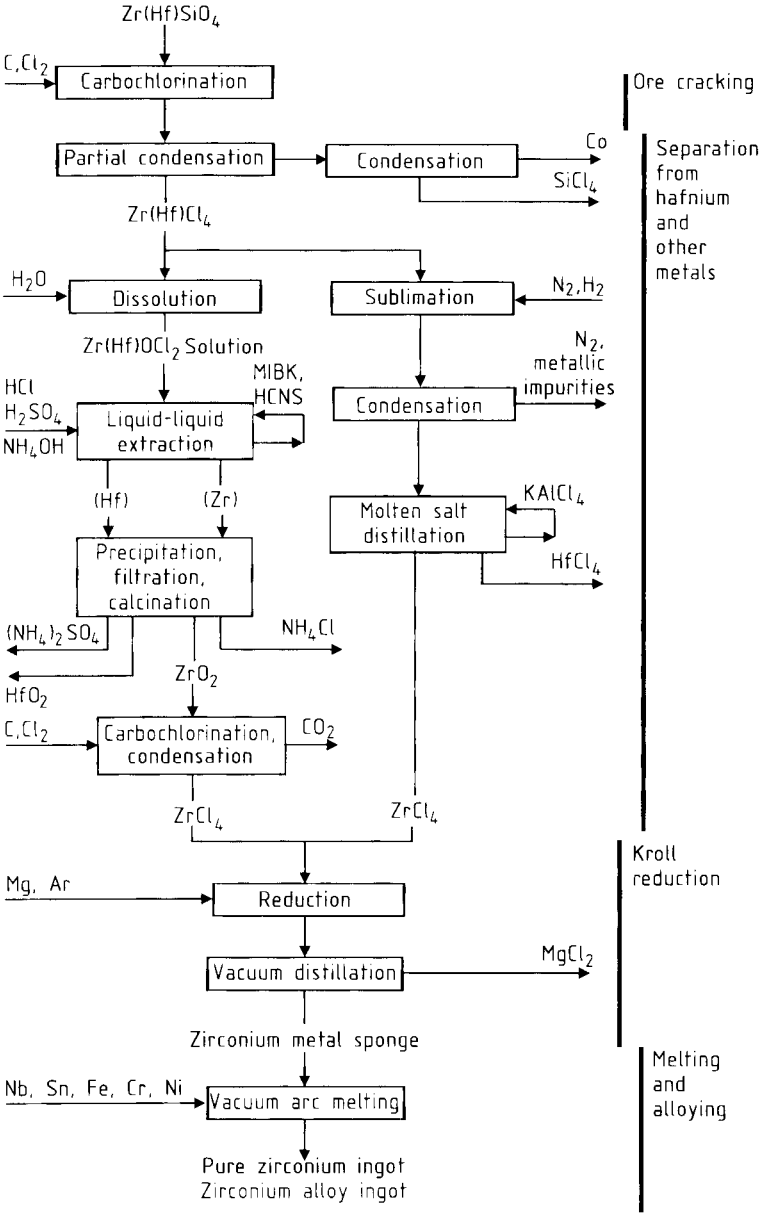

Figure 1. Flow diagram for production of zirconium

The fused mass is crushed, then the fluoride salt is dissolved with acidified hot water. The hot solution is filtered to remove silica and cooled to crystallize potassium hexafluorozirconate.

Lime Fusion. Fusion of zircon with limestone or dolomite produces calcium zirconate and calcium or magnesium silicate:

$$2\,ZrSiO_4 + 5\,CaCO_3 \rightarrow 2\,CaZrO_3 + (CaO)_3(SiO_2)_2 + CO_2$$

On cooling, the fused mass disintegrates into a very fine calcium silicate powder and coarse calcium zirconate crystals, which can be recovered by mechanical means. The acid-soluble calcium zirconate can be converted into zirconium salts or zirconium oxide.

Carbiding. An intimate mix of zircon sand and coke is continually fed into an open-top electric arc furnace. Insufficient carbon is used for complete conversion to carbide, so that silicon monoxide is vaporized at the arc temperature [12]:

$$ZrSiO_4 + 3\,C \xrightarrow{2500\,^\circ C} Zr(C,N,O) + SiO + 3\,CO$$

A crude zirconium carbonitride ingot grows under the electrode, surrounded by unreacted mix, which acts as insulation for the steel furnace shell. When the furnace is full, it is moved to a cooling location. When cool, the ingot is separated from unreacted charge, which is recycled. The product is a dense block, having a silver to golden-yellow fracture surface, depending on nitrogen content. The block is broken and crushed to −75 mm chunks, which are subsequently chlorinated to zirconium tetrachloride. If the hot ingot is immediately removed from the furnace shell and unreacted charge, the ingot is oxidized to zirconium oxide containing ca. 5 % silica.

High-Temperature Fusion. Direct electric arc smelting of baddeleyite or zircon is used to make fused zirconia or fused alumina – zirconia – silica (AZS). The silica is partially vaporized and recovered as fumed silica.

Baddeleyite is converted to zirconium sulfate by direct dissolution in concentrated sulfuric acid at ca. 200 °C. The recovery is low. Better recoveries are obtained by first heating the baddeleyite with lime to yield calcium zirconate, which reacts with sulfuric acid to give zirconium sulfate and gypsum.

1.4.2. Hafnium Separation

Zirconium and hafnium are very similar chemically and metallurgically, and for most uses of zirconium their separation is unnecessary. Their separation is conducted only to produce hafnium-free zirconium metal for the nuclear power industry. Of the many methods developed since the discovery of hafnium, four are used industrially: fluoride salt crystallization, methyl isobutyl ketone extraction, tributyl phosphate extraction, and extractive distillation, in order of their development. In addition, ion exchange has been used for low-throughput, low-investment operations.

Fluoride Salt Crystallization. The first separation of zirconium and hafnium by repeated crystallization of potassium hexafluorozirconate from hot aqueous solution is credited to Von Hevesy [13]. The solution is acidified to minimize oxide – fluoride salt formation. At each step the salt crystals are depleted in hafnium. This multistep recrystallization procedure has been used for many years in the Ukraine [14].

Methyl isobutyl ketone extraction, developed at Oak Ridge National Laboratory and used by both American producers, is based on the preferred extraction of a hafnium dihydroxide thiocyanate complex from hydrochloric acid solution by methyl isobutyl ketone [15].

Zirconium – hafnium tetrachloride, produced by carbochlorination of zircon, is dissolved in water, with hydrolysis to form zirconium – hafnium dihydroxide chloride in a hydrochloric acid solution. This solution is contacted with methyl isobutyl ketone to extract iron as $HFeCl_4$ into the organic solvent. Then ammonium thiocyanate is added to the dihydroxide chloride solution, providing a mixed dihydroxide chloride/dihydroxide thiocyanate solution of zirconium and hafnium. This solution is countercurrently contacted with a methyl isobutyl ketone/thiocyanic acid solution to preferentially extract hafnium dihydroxide thiocyanate into the organic phase. This is conducted in a series of mixer-settlers or a series of packed or internally agitated columns. Hafnium is recovered from the ketone solution by scrubbing with dilute sulfuric acid.

Zirconium is recovered from the hydrochloric acid solution by heating the solution above 90 °C, adding precisely two moles of sulfuric acid to each five moles of zirconium, and then carefully raising the pH to 1.2 – 1.5 with dilute ammonium hydroxide. This precipitates a granular zirconium basic sulfate, $Zr_5O_8(SO_4)_2 \cdot x\,H_2O$ which is easily filtered and washed to remove aluminum and uranium impurities. The sulfate filter cake is repulped with dilute ammonium hydroxide to convert it to zirconium

hydrous oxide, which is fired to give pure zirconium oxide [16].

In this extraction process, methyl isobutyl ketone and thiocyanate are recovered and reused, but considerable quantities of hydrochloric acid, sulfuric acid, and ammonia are consumed. Zirconium produced typically contains 35 – 90 mg Hf/kg and the hafnium contains 200 – 2000 mg Zr/kg, but any degree of separation can be attained, without further chemical consumption, by providing fewer or more contacting stages during separation.

The zirconium oxide and hafnium oxide are each chlorinated in fluidized beds to give their tetrachlorides. Chlorination of the oxides is faster than that of the silicates, and can be efficiently conducted at a lower temperature.

$$ZrO_2 + C + 2\,Cl_2 \xrightarrow{900\,^\circ C} ZrCl_4 + CO_2$$

The zirconium tetrachloride is then purified by sublimation at 350 – 400 °C in a nitrogen atmosphere containing 1 – 5 % hydrogen. The hot gases pass through a filter to remove entrained particles before being cooled to condense the zirconium tetrachloride. This sublimation reduces the levels of oxide, iron, phosphorus, and aluminum.

Tributyl phosphate extraction, developed concurrently in Britain, France, and the United States, was used commercially in the United States and is being used in India to obtain zirconium for the nuclear industry [17]. Hydrous zirconium – hafnium oxide from caustic fusion of zircon is dissolved in nitric acid. The nitric acid solution is countercurrently contacted with an organic solution of tributyl phosphate in kerosine. A zirconium tributyl phosphate complex is preferentially extracted into the kerosine, leaving hafnium and most metallic impurities in the aqueous phase. The zirconium is stripped from the kerosine solution with dilute sulfuric acid solution, precipitated, and fired to pure zirconium oxide.

Extractive Distillation. Hafnium tetrachloride is slightly more volatile than zirconium tetrachloride, and separation by fractional distillation would be feasible if the tetrachlorides could be handled as liquids. However, the tetrachlorides are either solid or gaseous, depending on temperature, unless kept under pressure while being heated to the triple point of zirconium tetrachloride (435 °C, 2.0 MPa). However a distillation process was developed in which the liquid phase was molten potassium chloroaluminate, in which zirconium and hafnium tetrachloride are soluble without forming compounds that could impede the separation or the recovery of separated chloride [18].

The zirconium – hafnium tetrachloride from carbochlorination of zircon is first purified by sublimation. The purified tetrachloride is revaporized and the vapor is introduced into the distillation column, above the midpoint. The potassium chloroaluminate liquid, equilibrated with hafnium-rich tetrachloride, is fed into the top tray of the distillation column, which is at 350 °C. As the liquid cascades down the column, it is gradually stripped of the more volatile hafnium tetrachloride while acquiring zirconium tetrachloride from the rising vapor. The zirconium tetrachloride in the liquid melt reaching the bottom of the column contains <50 mg Hf/kg Zr.

In the boiler at the column bottom some of the zirconium tetrachloride is vaporized to provide the rising vapor in the column. From the boiler the liquid melt goes to a short, hotter column where it is stripped with nitrogen to remove the product zirconium tetrachloride.

The stripped liquid melt is returned to an absorber – condenser above the distillation column where it is equilibrated with hafnium-rich tetrachloride vapor, then returned to the top of the distillation column. Unabsorbed hafnium-rich tetrachloride vapor ($HfCl_4$ content 30 – 50 %) is led to a condenser. The hafnium-rich tetrachloride is accumulated until it can be reprocessed in the distillation column in a separate campaign to recover pure hafnium tetrachloride (<1 % $ZrCl_4$).

Control of the potassium to aluminum ratio in the melt is crucial. Excess potassium chloride preferentially bonds with hafnium tetrachloride, reducing its effective vapor pressure and thereby decreasing the separation efficiency. With a deficiency of potassium chloride, aluminum chloride is lost into the nitrogen stream in the stripper, contaminating the zirconium tetrachloride product.

This extractive distillation process has a higher yield than the liquid – liquid extraction

processes, does not consume large quantities of chemicals, and eliminates the rechlorination step. It is very capital intensive.

Other Processes. Although zirconium is not easily reduced except by very strong reducing agents, it can be reduced more easily than hafnium. This has led to several separation procedures. The most significant [19] involved both molten metal (zinc base) and molten halide salts for good contact between the phases and for the close approach to equilibrium needed for good separation factors.

1.4.3. Reduction to Metal

Oxygen strengthens zirconium metal, decreasing its ductility and therefore its formability as a metal. To produce ductile zirconium metal, the zirconium compound to be reduced and the reducing agent should be as free of oxygen (and nitrogen and carbon) as possible. Anhydrous zirconium tetrachloride or tetrafluoride are usually used.

Kroll Process. Ductile zirconium metal is produced by reduction of pure zirconium tetrachloride with molten magnesium under an inert gas (helium or argon). The reduction retort is a long, vertical cylinder composed of two sections. The lower third is a stainless-steel-lined crucible which holds the magnesium ingots. The upper two-thirds has a central annular pipe slightly shorter than the retort wall. The space between the wall and the central pipe is filled with zirconium tetrachloride powder, then the two sections are welded together and a lid is bolted to the top. The assembled retort is warmed to 200 °C, evacuated and filled with argon several times to remove the last traces of air. The retort is lowered into a furnace with heating zones for each retort section. The lower section is then heated to 850 °C to melt the magnesium, and argon is bled from the warming retort to maintain atmospheric pressure (the retort is not a pressure vessel). Zirconium tetrachloride vapor passes down through the central pipe, which communicates with the lower section, and is reduced by the molten magnesium. At the end of the reduction, the lower crucible contains a thick suspension, or mud, of tiny zirconium metal beads in liquid magnesium, under a layer of liquid magnesium chloride. The retort is cooled and unloaded. The stainless steel liner is separated from the reduction mass and the layer of magnesium chloride is mechanically removed from the zirconium – magnesium metallic regulus. Several reduction reguli are stacked and loaded into a distillation furnace for removal of magnesium and residual magnesium chloride. The distillation is conducted in a vacuum (<1 Pa), with the charge located in the upper end of the retort. The bell furnace around the top of the retort is heated in programmed steps, reaching 980 °C in ca. 30 h. As the charge heats, the magnesium chloride melts and drains away, and the magnesium metal slowly evaporates, condensing on the water-cooled retort wall. As the magnesium evaporates, the zirconium beads come into closer contact and begin to sinter together. The final sintered porous mass is known as zirconium sponge. The heating schedule is critical: If the magnesium receives heat faster than heat is removed by magnesium transpiration at the surface of the reguli, the reguli become soft and flow down out of the heated zone. After the retort is cooled and the sponge conditioned by slow, controlled exposure to air, the sponge regulus is removed from the retort. The regulus is broken into chunks by using a hydraulic chisel. The chunks are sorted into quality grades depending on color and relative location in the regulus. The chunks are crushed to less than 2 cm, and screened to remove -20 mesh fines, which are higher in impurities.

Other Reduction Processes. Ductile zirconium has been produced commercially in a two-step sodium reduction analogous to a titanium reduction process [20]. Zirconium tetrachloride vapor and liquid sodium are both slowly fed into an argon-filled stirred reactor containing granular sodium chloride to obtain a $ZrCl_2 - NaCl$ mixture. This first step is very exothermic and the feed rate is controlled to allow excess heat to escape. The mixture is then transferred to a reactor by means of a screw conveyor. The reactor, already loaded with sodium, is sealed and heated to 800 °C:

$$ZrCl_2 + 2\,Na \rightarrow Zr + 2\,NaCl$$

The cooled mass was mechanically broken out of the reactor and leached with water to remove the sodium chloride from the zirconium. However, it was necessary to drip-melt the zirconium, forming splat-cooled chunklets, to remove the last traces of salt.

The Russian practice is to reduce potassium hexafluorozirconate with calcium metal in a sealed bomb. In Canada, zirconium tetrafluoride was reduced with calcium metal in a sealed bomb.

Zirconium oxide can be reduced by calcium or magnesium. Finely divided metal powder is recovered by leaching with cold hydrochloric acid. The powder has 0.3 – 0.5 % oxygen, and would not be ductile or melleable if melted into ingot. The powder particles are 1 – 8 µm in diameter and are very pyrophoric.

Electrolysis. Electrowinning and electrorefining of zirconium have been conducted but are not in commercial use. A pilot electrolysis system tested for zirconium was used commercially for hafnium [21]. A summary of the difficulties in developing an electrolysis process is given in [22].

1.4.4. Refining

From 1925 to 1945 the van Arkel and de Boer iodide process was the only consistent method for obtaining pure ductile zirconium metal. Now Kroll zirconium is used for most applications and iodide zirconium is used only when the highest purity is needed. In the iodide process, iodine vapor is reacted with Kroll zirconium sponge or calcium-reduced zirconium metal powder to produce zirconium tetraiodide [23], [24]. The tetraiodide vapor diffuses to a heated filament, usually zirconium wire, where the iodide is thermally dissociated, depositing zirconium and releasing iodine to repeat the cycle:

$$2\,I_2 + Zr \xrightarrow{300\,^\circ C} ZrI_4 \xrightarrow{1400\,^\circ C} Zr + 2\,I_2$$

The reaction is conducted in an Inconel vessel at an absolute pressure of 40 Pa. The wire filament diameter grows as zirconium is deposited on the wire. Bars up to 40 – 50 mm in diameter can be grown.

Electron beam melting is not usually used to purify zirconium because the metal's vapor pressure at its melting point is higher than that of most impurities. Iron and aluminum levels can be reduced by electron beam melting. The metal vapor pressure results in considerable evaporation loss in the high vacuum utilized in electron beam melting.

1.5. Fabrication

The procedures for melting zirconium sponge and for conversion of ingots to mill products are similar to those developed by the titanium industry, with the added requirement that high-neutron-cross-section materials such as boron be rigorously excluded.

Melting. Zirconium sponge is pressed into compacts, cylindrical or hexagonal discs, which then are stacked and surface-welded together in an electron beam furnace to form an electrode. The discs may be pure sponge, or may contain alloying ingredients added in the midst of each disc during pressing, or may be recycled clean turnings of the same intended composition. The welded electrode is then melted in a vacuum arc furnace. An electric arc is struck between the electrode and a starter pad of zirconium chips in the bottom of a water-cooled cylindrical copper crucible. The electrode is not cooled and it slowly melts off in the arc, producing a molten pool in the cooled copper crucible. The pool edge freezes on contact with the cool crucible wall so that the remaining pool is separated from the crucible wall. As melting continues, the electrode is consumed and the ingot grows in the crucible, cooling and solidifying from the bottom and sides as the pool moves upward. The outer skin contains impurities that vaporize during melting, primarily magnesium and magnesium chloride, and any possible copper pickup from sliding the ingot out of the crucible. The first-melt ingot is surface machined to remove the outer skin. Since the molten pool was, at any one instant, only a small portion of the ingot, distribution of alloying ingredients in the ingot is not homogeneous after one melt. Two machined first-melt ingots are formed into a second electrode, first-melt bottom now at the top, which

is again melted in a larger diameter copper crucible. A third melt assures complete homogeneity and dissolution of any inclusions. 750 mm diameter ingots of up to 9 t are produced.

Forming. Ingots are surface machined then heated to over 1000 °C and hot forged to 50 – 150 mm thick slabs, or to 200 – 400 mm diameter rounds. At the conclusion of hot working, the surface oxide scale is removed by sandblasting and abrasive-wheel grinding. The underlying dark layer of oxygen-diffused metal is removed by pickling in dilute hydrofluoric – nitric acid solution [25].

Strip and sheet are formed by cold working the slabs in conventional rolling mills, including producing foil in a Sendzimir mill. The allowable cold deformation before annealing is 20 – 70 %, depending on alloy toughness and the type of deformation process. Intermediate and final anneals are conducted in vacuum or inert-atmosphere furnaces.

For round products, the 200 – 400 mm rounds are cut to length, hot rotary forged to bar, cleaned, and swaged or drawn to rod or wire for welding use.

For hollow round products, short lengths of the rounds are drilled or trepanned to produce extrusion billets which are warm extruded to heavy wall tubeshells. Prior to extrusion the billets are coated with proprietary layers of glass, graphite, or copper. These coatings minimize oxidation of the surface and function as a lubricant during extrusion, but must be removed before subsequent cold working.

The extruded tubeshells are tube-reduced in pilger mills, with intermediate vacuum anneals, to produce thin-wall tube lengths. Zircaloy 2 or 4 (see Section 1.9) tubes are used as cladding for uranium oxide pellets for light-water nuclear fuel assemblies. This is the major use of zirconium metal.

Powder Metallurgy. Malleable zirconium and high-zirconium alloy powders are obtained by the hydride – dehydride process. Zirconium is hydrided by heating to 800 °C and cooling in a pure hydrogen atmosphere. The product is brittle and easily ground to powder in an argon atmosphere chamber. The hydrogen is then removed by degassing in a vacuum above 650 °C. The metal powder is screened to obtain the desired particle-size distribution. The hydride powder can also be directly pressed and sintered, allowing for hydrogen outgassing.

Welding. Zirconium can be welded in argon or helium atmospheres. If welding is not conducted in an inert-gas chamber, the welding fixtures or jigs must be so arranged that argon or helium gas flooding is sufficient to keep air completely away from the welding zone. Care must be taken to avoid oxygen or hydrogen pickup as these affect the ductility and corrosion behavior of the heat-affected zone. Electron beam welding is used for high-quality welds in repetitive situations such as nuclear fuel assembly.

While zirconium can be welded to titanium, hafnium, niobium or tantalum, joining of zirconium with other metals (i.e., copper, iron, nickel) by fusion welding is not practical because of the formation of brittle intermetallic phases. To join these metals, methods must be used that avoid formation of liquid phases but use mechanical interlocking: friction welding or explosive bonding, which is used for cladding of zirconium lining onto ferrous alloys. Welding of zirconium to titanium, niobium, or tantalum destroys the corrosion resistance.

1.6. Hazards in Handling Zirconium Metal

Solid zirconium is not hazardous, but any high-surface-area form of zirconium metal (powder, dust, chips, foil, sponge) is hazardous because of its easy ignition in air and its exothermic reaction with many other elements and compounds.

Zirconium chemical equipment, if used under conditions not recommended, such as sulfuric acid above 80 % concentration or strong hydrochloric acid with ferric ion contamination, may suffer intergranular attack. This creates a highly pyrophoric surface layer, usually black, that is ignited easily when dry. The combination of γ-zirconium hydride and metallic particles has been suggested to be responsible for the pyrophoricity. Careful treatment in hot air or steam is recommended to stabilize this surface [26].

The ignition temperature of zirconium powder varies with particle size, method of production, and prior treatment. A <40 mm pow-

der prepared in inert atmosphere will immediately ignite upon exposure to air, unless it is first slowly exposed such as by gradually back-filling the airlock of a glove box, allowing the heat of surface oxidation to dissipate.

Zirconium powder produced wet, such as by acid leaching of a reduction mass, is usually safe if the water content is >50 wt %. This prevents localized temperature excursions. The powder should not be allowed to settle and compact for long periods of time, especially when freshly prepared. Hydrogen is still slowly generated, and accumulated gas bubbles in a compact powder mass can insulate some grains from the water, allowing localized heating to ignition.

Drying of wet powder is most hazardous. While wet powder is harder to ignite than dry powder, it burns much more vigorously because of the available oxygen from the water and is more explosive because of the generation of hydrogen and steam. The most hazardous water content is 5 – 10 %. Powder is dried as a thin layer on warm-water-heated trays in a vacuum drier.

Machining of zirconium generates fine chips which can be ignited easily. Hot chips are the usual ignition source. The best metal removal procedures consistent with keeping chip temperatures down are: using slow cutting speeds with heavy feeds, and using an ample flood of water-soluble cutting oil as a coolant. Accumulation of chips should be avoided. Chips should be removed as generated and stored in water-filled drums kept outside in an isolated area. Coarse chips, if not contaminated, possibly may be chopped, cleaned, and recycled. Fine chips should be disposed of by burning in an incinerator or an open pit, feeding the fire in small increments of 1 – 2 kg.

Many unexpected flash fires have been caused by improper handling of machining chips resulting in more than one incident of fatal flash burns. Workers involved in handling, storing, cleaning, or disposal of machining chips or other zirconium fines should wear heat-reflectant clothing.

Zirconium fires can be controlled only by cutting off the supply of oxygen and other oxidants such as water, carbon dioxide, carbon tetrachloride, iron oxide, and limestone. Small fires can be smothered with dry silica sand or dry table salt, or with argon or helium if the zirconium is enclosed in a container. Do not use water. Large fires are best left alone and allowed to burn out, keeping the surroundings from igniting.

Safe handling procedures for zirconium, including machining, grinding, welding, and descaling, are discussed in [27].

Solutions that corrode zirconium metal may react violently or explosively with zirconium powder. A solution of 83 % concentrated H_2SO_4 and 17 % $KHSO_4$ was added dropwise to 2 g of zirconium powder. Upon adding about 2 mL of the acid mixture, a violent explosion occurred.

1.7. Grades and Specifications

There are two parallel grading systems for zirconium: industrial (intended primarily for use as a construction material for equipment and piping in the chemical industry) and nuclear (intended for use as a construction material for fuel cladding and internal parts for nuclear power reactors). Some of these grades have little present usage.

The industrial grades are listed,showing first the ASTM designation, then any alternative designation:

R 60 701 (Zr 701)	an unalloyed zirconium of low hardness and containing low levels of the impurities iron and chromium
R 60 702 (Zr 702)	the usual unalloyed zirconium specified for general corrosion-resistant service
R 60 703 (Zr 703)	the least pure, available only as sponge metal, used primarily for alloying with other metals
R 60 704 (Zr 704)	an alloy containing tin and iron and used in chemical process applications where the higher strength of the alloy is needed and the corrosion resistance is sufficient
R 60 705 (Zr 705)	an alloy containing 2.5 % niobium used where high strength is needed
R 60 706 (Zr 706)	a softer version of Zr 705 used specifically for severe forming applications

All of these grades may contain the usual 2 % hafnium, or may be furnished hafnium-free. This does not affect the quality or usefulness in industrial applications.

Zirconium for use in nuclear service is supplied to meet individual company specifications (which usually use ASTM nuclear grades and specifications as a starting point). The nuclear grades are listed, showing first the ASTM designation, then any alternative designation:

R 60 001	unalloyed zirconium with a long list of controlled impurities to make it satisfactory for nuclear service. Primarily iodide zirconium for the first reactor core
R 60 802	(Zircaloy 2) an alloy containing 1.5 Sn, 0.15 Fe, 0.1 Cr, and 0.05 Ni used primarily in pressurized water reactors
R 60 804	(Zircaloy 4) an alloy containing 1.5 Sn, 0.2 Fe, 0.1 Cr used primarily in boiling water reactors
R 60 901	(Zr 2.5 Nb, 705) an alloy containing 2.5 Nb, used primarily for pressure tubes in the Canadian CANDU reactors

All of these grades are hafnium free, i.e., <100 mg Hf/kg Zr.

U.S. and Canadian specifications for zirconium products are listed in Table 3.

1.8. Uses

Zirconium is produced almost exclusively as the hafnium-free metal. Most of this metal is used in zirconium alloys containing low levels of tin or niobium, for structural parts in the core of water-moderated nuclear reactors. For this use, zirconium has several desirable attributes: it is ductile, i.e., it can be formed. It has good strength up to 450 °C, i.e., it does not deform at reactor core temperature. Its alloys have hotwater and steam corrosion resistance, i.e., it lasts a long time in normal use. It has low thermal neutron cross section, i.e., neutrons are not absorbed, shutting down the nuclear reaction. Its ore is readily available.

Zirconium is used for building chemical process equipment for those applications where its corrosion resistance is needed. In hot sulfuric acid up to 65 % concentration, zirconium is used in facilities which produce hydrogen peroxide, acrylic films and fiber, methyl methacrylate, and butyl alcohol. Zirconium is used in the cooler condenser on a nitric acid absorption column. The operation conditions are 200 °C and 1035 MPa. One condenser constructed with zirconium tubes and zirconium/304 L stainless steel explosion-bonded tubesheets contains over 18 km of zirconium tubing and has been in service since the mid 1970s. Zirconium is used in contact with ammonium carbamate in urea production, in production of acetic and formic acid, and in many hydrochloric acid environments. In these applications zirconium's corrosion resistance is excellent and the long life of the equipment has justified the use of zirconium.

Zirconium foil, 0.002 mm, is used as the ignition-flash material in photographic flash bulbs, just as zirconium powder used to be one ingredient in the old open "flash pans" of earlier photographers, but this usage is fading because of builtin electronic flashes in newer camers. Zirconium powder is still used in pyrophoric applications.

Zirconium and zirconium alloys with aluminum, iron, titanium, or vanadium are used for gettering in vacuum tubes [28], inert gases, and ultra-high-purity environments for the semiconductor industry [29]. Heated zirconium absorbs, traces of oxygen, nitrogen, carbon monoxide, carbon dioxide, and water irreversibly. Hydrogen is reversibly adsorbed. The adsorbed materials diffuse into the bulk of the getter alloy, providing fresh surface for renewed adsorption. For ultra-pure inert gases, getters are used at the point of use to remove contaminants picked up from storage tanks or piping systems. Getters are used in gettering pumps for improving the quality of vacuum in ultra-high-vacuum systems.

Zirconium in the forms of clean, chopped machining chips, crushed sponge or magnesium-zirconium reguli turnings are often used in place of master alloys for zirconium additions to steel melts, super alloys, and nonferrous alloys.

1.9. Economic Aspects

Ores and Minerals. The four major producing countries for the world's current estimated annual production of 10^6 t of zirconium minerals are Australia 50 %, Republic of South Africa 25 %, United States 10 %, and the Ukraine 6 %. Australia and South Africa are the major exporting countries. Europe, Japan, China and the United States are major importers.

There are no zircon mines. Zircon is recovered as a byproduct of the extraction of the titanium-containing minerals rutile, ilmenite, and leucoxene. The producers of these minerals adjust their output to follow paint-market demand for titanium oxide pigment. Therefore, the supply and the demand for zircon are rarely

Table 3. Specifications for zirconium and zirconium alloy products

Zirconium product	Industrial		Nuclear		
	ASTM	ASME	ASTM	DOE	AECL
Sponge	B 494		B 349		
Ingot	B 495		B 350	M10-1 T	MET 52
Forgings	B 493			M 2-9 T	
Bar (rod) and wire	B 550	SB 550	B 351	M 7-9 T	MET 62
Strip, sheet, and plate	B 551	SB 551	B 352	M 5-6 T	MET 59
Tubes, seamless and welded	B 523	SB 523	B 353	M 3-8 T	MET 92
Tubes, nuclear fuel cladding			B 811		MET 56
Pipe, seamless and welded	B 658	SB 658			
Welding fittings	B 653				
Bare welding rods		SFA 2.54∗			
Castings	B 752				

∗ Also American Welding Society spec. A 2.54.

in phase and zircon prices have wide swings. During 1975 – 1995, zircon prices ranged from $ 83 per tonne to over $ 1100 per tonne in late 1989. The price during 1990 – 1995 averaged $ 275 – 330 per tonne, varying with quality.

Consumption of zirconium ores continues its gradual growth. In 1975, world production of zircon was 455 000 t and by 1995 it had essentially doubled. Currently, the growth in demand for zirconium materials is being led by its use as a opacifier in glazed ceramic tile manufacture.

Metals. Less than 4 % of the zirconium ores are used in the manufacture of zirconium metals. Since the incident at Three Mile Island, the meltdown at Chernobyl, and the ending of the cold war, the demand for commercial nuclear power and for military naval propulsion systems has diminished. It is estimated that in 1994 total world production was ca. 3600 t, compared to ca. 7300 t in 1982.

The United States has the largest production capacity with two producers: Teledyne Wah Chang in Albany, Oregon (nameplate capacity ca. 2700 t/a) and Western Zirconium in Ogden, Utah (nameplate capacity ca. 1360 t/a). In France, Compagnie Europeene du Zirconium (CEZUS), owned jointly by Pechiney, Framatome and Cogema, has a capacity of ca. 1800 t/a. The capacity of the Prinieprovsky plant in the Ukraine is estimated to be 2700 t/a of calcium-reduced reguli, which is crystal-bar refined at Glasnov, Russia. The Indian plant in Hyderabad has a nameplate capacity of 270 t; production is consumed internally. Japan is not presently producing zirconium sponge.

Prices in 1994 were $ 20 – 28 per kilogram of sponge and $ 45 – 110 per kilogram for cold-rolled sheet, strip, or foil.

2. Zirconium Compounds

In anhydrous halide-salt melts zirconium may exhibit valences of 4, 3, 2 or 1, but in aqueous solution its oxidation state is +4. Zirconium compounds have coordination numbers of 6, 7, or 8. The colorless ion is hydrolyzed in aqueous solution. Because of the ion's high charge and small radius, zirconium has a great tendency to hydrolyze and to form polymers. It is believed that most zirconium in aqueous solution is present as a tetramer $[Zr_4(OH)_4]^{8+}$ [30]. Therefore, $ZrOCl_2 \cdot 8\,H_2O$, by formula, is actually $[Zr_4(OH)_4 \cdot 7\,H_2O]_4Cl_8$. Hydrolysis and further polymerization of these tetramer units occurs with time, by heating, or with decreasing acidity. True equilibrium may take days or weeks. The following anions are listed with respect to their ability to form complexes with zirconium: $I^- \approx Br < ClO_4^- < Cl^- < NO_3^- < SO_4^- < F^- \approx C_2O_4^{2-} \approx PO_4^{3-} < CO_3^{2-} < OH^-$.

2.1. Zirconium Silicate

Zirconium silicate, $ZrSiO_4$, occurs in nature as the mineral zircon. It can be made by heating an intimate mix of zirconia and silica to just above 1500 °C. The crystals are tetragonal with the zirconia and silica linked through shared oxygen atoms, forming edge-sharing alternating ZrO_8 triangular dodecahedra and SiO_4 tetrahedra [31]. Other, more complex silicates are also found in nature (see Section 1.5). Zirconium silicate is resistant to acids, aqua regia and cold alkali solutions. It is readily attacked by sodium oxide, by sodium hydroxide at 600 °C, and by reaction with hot alkaline compounds (see Section 1.4.1). Above 800 °C zircon is reduced by carbon, the basis for the carbochlorination process.

Uses. Zircon's mechanical strength, chemical stability and high melting point make it a desirable refractory. Total world production of zircon was $(770-960)\times10^3$ t/a for 1990 – 1995. Of this ca. 35 – 40 % is used in ceramic glazes and enamels, ca. 30 – 35 % in refractories, 15 – 20 % in foundry use, 8 – 12 % in abrasives, and 8 – 12 % in other uses including chemicals, metals and alloys, and glass constituents.

In the largest use, decorative ceramics, milled or micronized zircon is used as an opacifier for tile glazes and porcelain enamels. Zircon's high refractive index gives the glaze a white, opaque appearance.

Zircon is a popular refractory in the glass and steel industries. The glass industry uses fused zircon and AZS (alumina – zirconia – silica) refractories (→Refractory Ceramics, Chap. 7.1.4.) which have a high corrosion resistance to molten glass. The AZS refractories are made by electric furnace fusion of alumina and zircon. The steel industry uses zircon in ladles and continuous casting nozzles. Refractory applications were once the leading markets for zircon but they decreased significantly following the zircon shortage in the late 1980s. Other less-expensive refractories such as alumina spinels were substituted.

Zircon is used in foundries as a basic mold material and as a facing on other mold materials. The high thermal conductivity of zircon gives higher cooling rates than other mold materials. Zircon's high melting point, low coefficient of thermal expansion, and chemical stability make it the preferred mold facing, particularly in precision casting.

Alloyed AZ (alumina – zirconia) abrasives (→Abrasives, Chap. 3.2.) are made from 60 – 90 % alumina – zirconia grains produced by fusing alumina with baddeleyite, or zirconia derived from zircon, in electric arc furnaces. The addition of zirconia toughens the alumina, reducing its brittleness. The abrasives have particular application in the grinding of steel and steel alloys because of high strength, hardness, and surface crystal sharpness, resulting in fast grinding and long wheel life.

Zircon is added to glass for television tubes because of its X-ray absorption properties and its ability to toughen the glass.

2.2. Zirconium Oxide

Zirconium oxide [*1314-23-4*] occurs naturally as the mineral baddeleyite and is produced by calcining of other zirconium compounds. Three phases are stable at atmospheric pressure: cubic above 2370 °C, tetragonal above 1170 °C, and monoclinic below 1000 °C. Transformation of the monoclinic phase to the tetragonal phase begins at ca. 1050 °C and is complete at ca. 1170 °C. The transformation is accompanied by a volume shrinkage of 3 – 5 %. The transformation has a thermal hysteresis: on gradual cooling the tetragonal phase is stable to ca. 1000 °C and only finishes conversion to monoclinic at ca. 800 °C. With rapid quenching the tetragonal phase is metastable to room temperature.

The high-temperature cubic phases can be stabilized down to room temperature by the addition of magnesia, calcia, yttria, or rare earth oxides, whose presence creates vacancies in the zirconium anion lattice. These stabilized zirconium oxides contain from 3 % yttria to 8 % calcia. They have the same crystal structure from room temperature to melting, avoiding the catastrophic mechanical failure suffered by pure zirconia ceramic parts on transforming from the tetragonal to monoclinic phase while cooling.

Later, Garvie [32] discovered that certain partially stabilized zirconias could be thermally cycled to precipitate metastable tetragonal zirconia within the grains of cubic zirconia. These

materials had higher toughness than fully stabilized zirconias. The increased toughness of partially stabilized zirconia (PSZ) is the result of stress-induced martensitic transformation of the metastable tetragonal grains to the monoclinic form in the stress field of a propagating crack. This led to development of an entirely tetragonal zirconia, known as tetragonal zirconia polycrystal (TZP). The evolution of these materials and their growing application in structural ceramics can be followed in [33–38]. An introduction to zirconia ceramics is given in [39].

Zirconia also has a high-pressure (10 GPa) orthorhombic structure which can be stabilized to atmospheric pressure by addition of >12 mol % of niobia or tantala or a mixture thereof [40]. No transformation toughening was found [41].

Stable lower oxides of zirconium are not known, although oxygen dissolved in the zirconium metal lattice has led to the identification of some superstructures as $ZrO_{0.3}$ [42]. Zirconium monoxide, ZrO, has been observed in mass spectrographic measurements [43]. The black oxide surface on zirconium metal after exposure to water at 350 °C is a slightly substoichiometric $ZrO_{1.95}$.

Zirconia is very resistant to acids and alkalis, but slowly dissolves in concentrated hydrofluoric acid or hot concentrated sulfuric acid. Zirconia is resistant to many fluxes, molten glasses, or melts, silicate, phosphate, or borate, but is attacked by fluoride or alkaline melts. Zirconium oxide and alkaline oxides or caustics can be fired together to form solid-solution oxides, zirconate compounds, or a mixture of both.

Zirconium oxide is reduced by carbon, beginning at ca. 600 °C. With excess carbon, in a vacuum furnace, the reaction proceeds rapidly to zirconium carbide at 1500 °C. In the presence of chlorine the carbon reduction gives zirconium tetrachloride and carbon dioxide.

Uses. Zirconium oxide is used to coat the surface of titanium oxide pigment particles for some grades of exterior service paints. The coating minimizes ultraviolet excitation of the titania, which would interact with the organic paint binder, shortening the service life. Zirconia has been used in place of titania as the pigment of white camouflage paints for use in snow environments because zirconia more closely simulates snow in the infrared and microwave spectra. Over 300 t of single-crystal cubic zirconia are grown each year to provide low-cost gems for the jewelry trade [44]. Stabilized zirconia is being used in everyday applications such as fishing rod ferrules, knives, unbreakable shirt buttons, and golf putter heads.

While stabilized zirconias are insulators at room temperature, at elevated temperatures the vacancies in the anion lattice allow O^{2-} ions to diffuse and the zirconia becomes a solid electrolyte with applications in oxygen sensors and high-temerature fuel cells. Some stabilized zirconias can be used as resistors or susceptors. Inductively heated yttria-stabilized zirconia cylinders are used as heat sources to melt quartz boules for the drawing of quartz optical fibres.

Zirconia is a constituent of lead zirconate titanate (PZT) used in piezoelectric ceramics (→Ceramics, Electronic) for applications as gas furnace and barbeque igniters, microphone and phonograph crystals, ultrasonic transducers for medical ultrasound imaging, for agitation in cleaning tanks, and for underwater sonar. With the further addition of lanthanum (PLZT), ferroelectric optically active transparent ceramics have become available (→Ceramics, Electronic) [45].

2.3. Hydrous Zirconium Oxide

Neutralization of an aqueous zirconium solution causes hydrolysis and condensation of the zirconium cations, forming a white gel called zirconium hydroxide but generally considered to be a hydrous oxide. The freshly precipitated amorphous gel is generally considered to be a network of tetramer units $[Zr_4(OH)_8 \cdot 16\,H_2O]^{8+}$ linked through hydroxyl groups; it is easily dissolved by dilute or weak acids. As the gel ages some of the hydroxyl links are converted to oxo links, and the gel's reactivity decreases. Upon heating to 500 °C the gel yields monoclinic or metastable tetragonal oxide, or a mixture of these. The specific structure obtained has been attributed to the precursor zirconium salt [46], pH of the liquid in final contact with the condensed gel [47], electronegativity and pH [48], slow or rapid condensation from sol to gel, ageing, the presence during gelling of sulfate, sodium, potassium, or ammonium ions, and particle size [49].

Hydrous zirconium oxide in controlled particle size is prepared by hydrolysis of sodium zirconate [50].

Hydrous zirconium oxide sol is used in sol – gel processes to produce stabilized zirconias in spherical or fibrous form, and to form an oxidation-protective film on stainless steels.

2.4. Zirconium Chlorides

The fully chlorinated form of zirconium is zirconium tetrachloride. The tetrachloride is normally produced by carbochlorination of zircon or zirconia [51] (see page 6) or by reacting zirconium metal with chlorine or hydrogen chloride. Chlorinated hydrocarbons, particularly carbon tetrachloride, chloroform, or hexachloroethane, are useful laboratory chlorinating agents. Zirconium tetrachloride can be formed by reacting ferrozirconium or silicozirconium with iron(II) chloride at elevated temperature [52].

Zirconium tetrachloride vapor is a tetrahedral monomer which crystallizes below 331 °C in a AB_4 type structure with $ZrCl_6$ octahedra coupled to each other by two edges to form a zig-zag chain. Each zirconium has four bridging chlorine ligands and two terminal chlorines which are mutually *cis* [53]. Some physical properties of zirconium tetrachloride are shown in Table 4.

Zirconium tetrachloride reacts rapidly with water, vapor or liquid, to exchange two chlorines for oxygen. Any handling of zirconium tetrachloride must be in moisture-free atmospheres to avoid degradation. Zirconium tetrachloride powder fed into water hydrolyzes to form a solution of zirconium hydroxide chloride and free hydrochloric acid. Zirconium tetrachloride vapor reacts with steam to yield zirconium oxide, and reacts with water to yield either an oxide slurry or hydroxide chloride solution.

Zirconium tetrachloride is used to form zirconium nitride coatings, to react electrochemically to form zirconia in high-temperature fuel cells, to react with alcohols to form alkoxides, and to produce zirconium organometallic compounds. Zirconium tetrachloride is reduced by molten alkali and alkaline earth metals, yielding zirconium metal.

Zirconium tetrachloride reacts reversibly with ammonium chloride, alkali metal chlorides, and alkaline earth chlorides to form hexachlorozirconates [54], [55]. Zirconium tetrachloride forms addition compounds with aluminum and iron trichlorides and phosphorus pentachloride, but $FeZrCl_6$ dissociates to $FeCl_2$ and $ZrCl_4$ above 350 °C.

Lower valent zirconium chlorides including zirconium trichloride [*10241-03-9*], dichloride [*13762-26-0*], monochloride [*14989-34-5*] and zirconium chloride cluster phases Zr_6Zl_{12} and Zr_6Cl_{15} are produced by reduction of zirconium tetrachloride with zirconium metal or with zirconium monochloride in sealed tantalum containers. The reactions are slow, usually requiring weeks at 300 to 800 °C [56–58]. Zirconium monochloride reacts reversibly with hydrogen at 340 °C in a sealed tube to form $ZrClH_x$ ($x \leq 1$) [59].

2.5. Zirconium Hydroxide Chlorides

Zirconium hydroxide dichloride [*22196-48-1*], empirically $Zr(OH)_2Cl_2 \cdot 7\,H_2O$, has been known by several names: zirconyl chloride, zirconium oxychloride, zirconium oxide dichloride, all commonly identified as $ZrOCl_2 \cdot 8\,H_2O$. Structurally the hydroxide dichloride is really $[Zr_4(OH)_8 \cdot 16\,H_2O]Cl_8 \cdot 12\,H_2O$ [60].

Zirconium hydroxide dichloride solution is formed by dissolving hydrous zirconium oxide or zirconium carbonate in hot hydrochloric acid, or by hydrolysis of zirconium tetrachloride in water. The solubility in hot water is quite high and the solubility is decreased by cooling and by adding hydrochloric acid. Excellent recoveries are obtained by crystallization ending with a 6 N hydrochloric acid mother liquor. This crystallization procedure also leaves most impurities dissolved in the mother liquor. Crystallization may be repeated to obtain very pure crystals.

Zirconium hydroxide dichloride crystals dissolve in their hydration water on gentle heating, and readily lose the hydrates outside the tetramer unit. With continued heating both water and hydrogen chloride are evolved until zirconium oxide remains [61], [62].

Zirconium hydroxide monochloride $[Zr_4(OH)_{12} \cdot 16\,H_2O]Cl_4$ is formed in solution by reacting equimolar portions of zirconium hydroxide dichloride and hydrous zirconium oxide or zirconium basic carbonate, or by the action of

Table 4. Physical properties of zirconium tetrahalides

Property	ZrF_4	$ZrCl_4$	$ZrBr_4$	ZrI_4
Color	white	white	white	orange-yellow
Density, kg/m^3	4430	2800	3980	4850
Sublimation temp (101.3 kPa)	903	331	357	431
Melting point, K	923	437	450	499
Critical temperature, K	932	503.5	532	686
Critical pressure, MPa		5.7	4.3	4.1
Crystal structure	monoclinic	monoclinic	cubic	cubic
CAS registry no.	[*7783-64-4*]	[*10026-11-6*]	[*13777-25-8*]	[*13986-26-0*]

hydrogen peroxide upon zirconium hydroxide dichloride [63].

Anhydrous zirconium oxide chlorides have been prepared: the reaction of zirconium dioxide with zirconium monochloride at 980 °C produced $ZrClO_y$ ($y < 0.42$) [64]. The reaction of zirconium tetrachloride powder in carbon tetrachloride with dichlorine oxide at −30 °C yielded $ZrOCl_2$, which dissociated to $ZrCl_4$ and ZrO_2 above 200 °C [65].

2.6. Zirconium Bromides

Zirconium tetrabromide, $ZrBr_4$, can be prepared by heating zirconium metal or carbide above 400 °C in a stream of bromine, or similarly a fine mixture of zirconium oxide and carbon at 900 °C. Its behavior is similar to zirconium tetrachloride but, because of its higher cost, the tetrabromide is of little commercial interest. Some physical properties are shown in Table 4.

Zirconium tribromide [*24621-18-9*], zirconium dibromide [*24621-17-8*], zirconium monobromide [*31483-18-8*], and zirconium monobromide hydride [*60921-40-6*] can be made by the same procedures used for their chloride analogues [66].

2.7. Zirconium Iodides

Zirconium tetraiodide, ZrI_4, is important because of its use in the van Arkel – de Boer refining process (see Section 1.4.4). The tetraiodide cannot be produced by heating fine zirconium oxide and carbon in iodine vapor, but is produced by heating zirconium metal to 300 °C in iodine. A temperature of ca. 1000 °C is required for a reasonably rapid reaction of iodine with zirconium carbide, carbonitride, or nitride. Zirconium tetraiodide is oxidized by dry air at about 200 °C. Some physical properties are shown in Table 4.

Lower-valence iodides have been made by reduction of zirconium tetraiodide with zirconium in sealed tantalum tubes or with zirconium or aluminum in aluminum iodide solvent [67]. Compounds formed include zirconium triiodide [*13779-87-8*], zirconium diiodide [*15513-85-6*] (both α monoclinic and β orthorhombic form), zirconium clusters Zr_6I_{12} [*66908-75-6*], and zirconium monoiodide [*14728-76-8*] [68].

2.8. Zirconium Fluorides

There are several methods for the production of zirconium tetrafluoride, ZrF_4. Anhydrous zirconium tetrafluoride is produced in good purity by mixing hydrogen fluoride gas and zirconium tetrachloride vapor at 350 °C [69]. The hydrofluorination of zirconium oxide proceeds rapidly at 25 °C, followed by a purifying sublimation in dry hydrogen fluoride at 825 °C [70]. Zirconium tetrafluoride monohydrate is precipitated by adding strong hydrofluoric acid to a concentrated zirconium nitrate/nitric acid solution. The filtered crystals are washed with nitric acid, dried, and dehydrated with hydrogen fluoride at 450 °C [71]. Zirconium metal or zirconium hydride can be fluorinated with hydrogen fluoride but unless the reaction is conducted above 800 °C, the tetrafluoride products coats the metal, impeding the reaction. Zirconium oxides mixed with ammonium hydrogen difluoride and heated to 200 °C yield ammonium heptafluorozirconate [*17250-81-6*], $(NH_4)_3ZrF_7$, which on heating to 500 °C gives off ammonium flu-

oride. Some physical properties of zirconium tetrafluoride are shown in Table 4.

Zirconium tetrafluoride is hydrolysed in water to a hydroxide fluoride $Zr_4(OH)_6F_{10} \cdot 3\,H_2O$ with limited solubility. With increasing hydrofluoric acid concentration, $ZrF_4 \cdot 3\,H_2O$ crystallizes at 30 – 35 wt % HF, $HZrF_5 \cdot 4\,H_2O$ at 30 – 35 wt % HF, and $H_2ZrF_6 \cdot 2\,H_2O$ at 40 – 60 wt % HF. In strong nitric or sulfuric acid containing excess hydrofluoric acid, the monohydrate crystallizes. Heating these hydrates produces $ZrOF_2$ at 300 °C.

Very high purity anhydrous zirconium tetrafluoride is a major constituent of some fluoride glasses, particularly ZBLAN, being developed because their transparency for near UV to mid IR (0.3 – 6 µm) was projected to be better than that of quartz optical fiber.

Potassium hexafluorozirconate production by fluorosilicate fusion is described in Section 1.6. This compound has been used to introduce zirconium into molten aluminum and magnesium as a grain refiner.

As mentioned in Section 1.7, zirconium metal is customarily pickled in dilute hydrofluoric acid/nitric acid solution. To minimize fluoride waste discharge the pickling solution is regenerated by adding sodium fluoride to precipitate sodium fluorozirconates. The particular salt precipitated varies with the sodium fluoride concentration: $NaZrF_5 \cdot H_2O$ is precipitated when the sodium fluoride concentration is less than 0.21 %, and Na_2ZrF_6 is precipitated if the sodium fluoride concentration is between 0.21 and 0.4 %. At higher sodium fluoride concentration $Na_5Zr_2F_{13}$ and Na_3ZrF_7 precipitate.

Zirconium tetrafluoride forms many double fluoride compounds in different proportions with alkali fluorides, alkaline earth fluorides, and others as diverse as stannous hexafluorozirconate and zirconium hexafluorogermanate, both proposed for caries prophylaxis.

2.9. Zirconium Sulfates

The sulfates of zirconium are classified as anionic, normal, or basic, depending on the sulfate to zirconium ratio. For anionic complexes this ratio is >2. Normal sulfates have a ratio of exactly two, and basic sulfates have less than two sulfates per zirconium. Basic sulfates made by hydrolysis of neutral sulfates usually have a ratio >1, while basic sulfates precipitated by adding sulfate ions to hot dilute solutions of zirconium hydroxide chloride have a ratio <1, usually 0.4 – 0.7.

Zirconium sulfate solutions can be generated by the action of strong sulfuric acid on hydrous zirconium oxide, zirconium oxide, zirconium carbonitride, zirconium metal, or basic zirconium sulfate. Normal zirconium sulfate [*7446-31-3*], $Zr(SO_4)_2 \cdot 4\,H_2O$, also known as zirconium orthosulfate, can be crystallized from the above solution, or from a zirconium hydroxide chloride solution by adding sulfuric acid up to a concentration of 45 % and allowing the solution to cool below 60 °C. A better purification is obtained during the crystallization if the solution contains some chloride ion.

Ammonium zirconium sulfates can be prepared by heating zirconium oxide with ammonium sulfate at 400 °C [72] or by precipitation from a zirconium sulfate solution [73]. A crude sodium zirconium sulfate can be prepared by adding sulfuric acid to sodium zirconate. Structure and composition of many sulfate variants are discussed in [30], [74], [75].

Basic zirconium sulfates containing less than one sulfate per zirconium are precipitated from hot acidic solutions of zirconium hydroxide chloride by adding 0.4 – 0.6 mol of sulfate per zirconium. The stoichiometric formula is $Zr_5O_{(10-x)}(SO_4)_x$ (x=1.8 – 3.5). The most common basic sulfate is $Zr_5O_7(SO_4)_3 \cdot n\,H_2O$.

Normal zirconium sulfate is used for leather tanning, as a catalyst, to coat titanium oxide pigment powder, and for conversion to other zirconium chemicals. The basic sulfate is converted into other zirconium chemicals, many retaining the particle size developed during the controlled precipitation of the basic sulfate.

2.10. Zirconium Carbonates

Basic zirconium carbonate [*37356-18-6*] is produced by adding basic zirconium sulfate to a sodium carbonate solution, and filtering the insoluble basic carbonate. The empirical formula is $ZrO_2 \cdot CO_2 \cdot x\,H_2O$ but the zirconium to carbonate ratio varies considerably.

The basic carbonate is reacted with organic acids to make soluble zirconium salts, such as the citrate, oxalate, and acetate, free of chloride and sulfate ions. The carbonate is used to make a zirconium – aluminum – glycine complex, an active ingredient of antiperspirant formulations.

Zirconium basic carbonate is added to potassium carbonate or ammonium carbonate solutions to make water-soluble double carbonates. At room temperature, these solutions hydrolyze so slowly that they are considered to be stable for a month, but on heating carbon dioxide evolves and insoluble hydrous zirconium oxide forms. Ammonium zirconium carbonate [*22829-17-0*], empirically $(NH_4)_3[Zr_2O(OH)_3(CO_3)_3]$, is used extensively in bonding starch onto coated paper [76], and in controlling pitch deposition in pulp and papermaking systems [77].

2.11. Zirconium Hydride

Zirconium hydride [*7704-99-6*], ZrH_2, is prepared by the reversible absorption of hydrogen into heated zirconium, usually sponge or ingot slices. The metal is heated above 600 °C in an inert atmosphere or vacuum to dissolve the surface oxide film, then hydrogen is introduced. At 600 °C, hydrogen dissolves into the α-phase metal up to 6 atom %, when β-Zr appears. (Hydrogen is a β-phase stabilizer, lowering the transformation temperature from 863 °C to 550 °C.) At 40 atom % the solubility in β-Zr is exceeded and δ-phase hydride appears. At 60 – 64 atom % only δ-phase exists. Above 64 at. % the tetragonal ε-phase is stable [78]. To obtain fully hydrided ZrH_2 the charge is cooled to 300 °C in full hydrogen atmosphere.

The reduction of zirconium oxide with calcium hydride yields a fine (2 – 5 µm particles) hydride powder for use in pyrotechnics such as fuses or fireworks:

$$ZrO_2 + 2\,CaH_2 \rightarrow ZrH_2 + 2\,CaO + H_2$$

Extreme care must be used in dissolving the calcium oxide to recover the hydride powder.

Hydriding of zirconium sponge or ingot is the first step in producing coarse zirconium powder for powder metallurgy applications. Hydrided metal begins to lose its ductility above 40 atom % hydrogen, and above 60 atom % the hydrided metal is quite brittle and suitable for crushing, which is conducted in an inert atmosphere. During hydriding, the metal density decreases with increasing hydrogen content, going from 6.5 to 5.6 when fully hydrided.

Zirconium hydride loses hydrogen upon heating. For complete dehydriding the powder must be heated under vacuum (e.g., 700 °C until the vacuum falls to 10^{-3} Pa). Dehydriding is faster at higher temperatures but the powder begins to sinter.

Zirconium hydride contains about the same number of hydrogen atoms per volume as room-temperature water, and some alloys such as ZrNi and $ZrCr_2$ can hold even more [79]. Zirconium hydride has been used in the nuclear industry as a moderator for thermal neutrons in compact, high-temperature reactors. Zirconium hydride is of interest for hydrogen storage and nickel-base hydrogen battery applications.

Zirconium hydride is inert to air and water at room temperature. It ignites in air at 340 °C or lower, depending on powder size. Zirconium hydride is resistant to corrosion by carbon dioxide up to 315 °C.

Zirconium hydride heated with carbon yields the hexagonal zirconium carbohydride, $ZrC \cdot ZrH_{1.46}$ [80]. Zirconium monochloride reversibly absorbs hydrogen up to ZrClH [81].

2.12. Zirconium Carbide

Zirconium carbide [*12020-14-3*], ZrC, is a hard, brittle metallic carbide. It is an electrical conductor and does not hydrolyze in water. It is an interstitial carbide with the carbon atoms occupying the octahedral interstices of the zirconium lattice, at 2000 °C, carbon is soluble in liquid zirconium up to 38.5 atom %. Zirconium carbide is a single phase within the concentration range of 38.5 – 49.4 atom % carbon with a congruent composition of 44 atom % carbon at the melting point of 3445 °C [82], [83].

Zirconium carbide can be produced from zircon by reduction with coke in electric arc furnaces [84]. An intimate mix of fine zirconium oxide and carbon black heated in vacuum or hydrogen in an induction-heated graphite crucible at 1600 °C yields a porous, light zirconium carbide. Milling and refiring above 2200 °C gives

a dense powder. Zirconium tetrachloride and methane in a hydrogen atmosphere are heated to 1200 °C to deposit zirconium carbide films on tool bits. Pure zirconium metal powder and fine graphite powder react exothermically, once ignited, to form zirconium carbide, usually to form a composite with an excess of one reagent.

2.13. Zirconium Nitrides

Zirconium nitride [*25658-42-8*], ZrN, is a golden-colored, hard, brittle metallic nitride. It is thermally and chemically stable, and has low resistivity. As with the carbide it is an interstitial compound with a NaCl cubic structure. At 2000 °C, nitrogen is soluble in liquid zirconium up to 39 atom % nitrogen. Zirconium nitride is a single phase within the concentration range of 39 – 50 atom % nitrogen [85], [86]. While 3253 K is the accepted melting point of zirconium nitride, the melting temperature of 3970 K as measured in a positive pressure of nitrogen [87] is compatible with a thermodynamic study of the Zr – N system.

Zirconium nitride has been made by heating zirconium or zirconium hydride powder in nitrogen at 1000 °C, and by heating a zirconium oxide/carbon mixture in a nitrogen atmosphere. The reaction of zirconium tetrachloride with ammonia or nitrogen and hydrogen has been used to produce nitride powder and to deposit coherent films on substrates at 1000 – 1200 °C. Very limited quantities of nitride have been made by the very exothermic reaction of lithium nitride and zirconium tetrachloride powders [88].

Metastable Zr_3N_4 has been made by dual ion beam deposition [89] and by reacting zirconium tetraiodide with ammonia at 500 °C and heating the product to Zr_3N_4 at 700 °C [90]. On further heating to 1000 °C ZrN is formed. Thin films of Zr_3N_4 have been grown by passing ammonia and tetrakis(diethylamido)zirconium over substrates at 200 – 400 °C [91].

Zirconium nitride has been used as a wear-resistant coating on steel drill bits and cemented-carbide tool bits, as decorative coating, as a protective coating on steel vessels handling molten metal, and as solar energy collector surface film.

2.14. Zirconium Borides

While three zirconium borides, ZrB, ZrB_2 and ZrB_{12}, are found in the zirconium – boron system, only the diboride is chemically and thermally stable. Zirconium diboride [*12045-64-6*] is a gray refractory solid with a hexagonal crystal structure and a very high melting point of 3245 °C. Technical-quality zirconium diboride can be prepared by reacting zircon, boron oxide, and carbon in a submerged-electrode arc furnace [92]. Purer diboride can be produced by co-reduction of zirconium tetrachloride and boron trichloride with hydrogen or aluminum in a chloride bath. Zirconium diboride films can be produced by subliming zirconium tetrahydridoborate at 25 °C in a vacuum and passing the vapor over a substrate heated to 250 °C.

Zirconium diboride has high hardness, good oxidation resistance, and excellent thermal-shock resistance. It has been used as a diffusion barrier in semiconductors, as a container for molten metals, and as a burnable absorber in nuclear reactor cores.

2.15. Zirconium Tetrahydridoborate

Zirconium tetrahydridoborate [*12370-59-1*], $Zr(BH_4)_4$, is prepared by reaction of zirconium tetrachloride with lithium tetrahydridoborate in diethyl ether, followed by double distillation at 20 °C [93]. It has a tetrahedral structure with three hydrogen atoms bridging the zirconium atom and each boron atom. Zirconium tetrahydridoborate is one of the most volatile zirconium compounds known: *mp* 29 °C; *bp* 118 °C; vapor pressure 2 kPa at 25 °C. Its decomposition near 250 °C forms the basis for a low-temperature method of depositing zirconium boride films. Zirconium tetrahydridoborate catalyzes the polymerization of unsaturated hydrocarbons. It reacts with aliphatic alcohols to form zirconium and boron alkoxides [94]. Zirconium tetrahydroborate inflames on contact with dry air.

2.16. Zirconium Tungstate

Zirconium tungstate is precipitated as a white gel by simultaneous addition of dilute solutions

of sodium tungstate and zirconium hydroxide dichloride to vigorously stirred water. After acidification the slurry is refluxed. The dried and fired product is ZrW_2O_8 [95]. Other studies indicate that ZrW_2O_8 formed by the reaction of a mixture of ZrO_2 and WO_3 which is held between 1100 and 1400 °C for at least 24 h melts incongruently at 1257 °C and dissociates below 1105 °C [96]. ZrW_2O_8 is an unusual compound which contracts equally in each dimension when it is heated over the temperature range 1 – 1050 K [97].

2.17. Zirconium Phosphate

Insoluble, amorphous zirconium phosphate is obtained when a zirconium salt is added to an excess of phosphoric acid. If the original salt was solid rather than in solution, the resulting phosphate will have a similar physical size and shape: i.e., beads, granules, fibers. Refluxing the precipitate in strong phosphoric acid gives a crystalline product of constant composition, $Zr(HPO_4)_2 \cdot H_2O$. Zirconium bis(monohydrogenphosphate) with one water of hydration is known as α-zirconium phosphate. Below 80 °C a second water of hydration is added to form γ-zirconium phosphate. On heating α-zirconium phosphate, the water is lost between 300 and 650 °C, leaving β-zirconium phosphate. Conversion to zirconium pyrophosphate, ZrP_2O_7, occurs between 800 and 920 °C [98].

Zirconium bis(monohydrogenphosphate) has ion exchange properties and, because of its structure and bonding between layers, can also act as an intercalation compound [99]. Zirconium phosphate shows excellent catalytic activity for the selective condensation of acetone to mesityl oxide, and when palladium is deposited on the phosphate surface the combination is an effective catalyst for the direct synthesis of methyl isobutyl ketone from acetone and hydrogen [100].

The addition of univalent ions Li, Na, K and compensating trivalent ions In, Y, Eu into ZrP_2O_7 produces a solid electrolyte [101]. Ultralow thermal expansion ceramics have been developed in the $Na_2O - ZrO_2 - P_2O_5 - SiO_2$ system [102], while the addition of vanadium to sodium zirconium pyrophosphate: $Na_2PrP_{2-x}V_xO_7$ yields a ceramic which shrinks when heated above 60 °C [103].

2.18. Zirconium Nitrates

Zirconium hydroxide dinitrate, [*20213-65-4*], $Zr(OH)_2(NO_3)_2$, is formed by dissolving hydrous zirconium oxide in nitric acid or by dissolving zirconium hydroxide dichloride in nitric acid and distilling off the chlorine.

Zirconium tetranitrate [*12372-57-5*], $Zr(NO_3)_4 \cdot 5\,H_2O$, can be precipitated from strong nitric acid at low temperature, less than 15 °C [104].

2.19. Zirconium Carboxylates

Zirconium hydroxide carboxylates form upon adding carboxylate salts to a solution of zirconium dihydroxide chloride, or carboxylic acids to zirconium basic carbonate. The general formula is $Zr(OH)_{4-n}(OOR)_n$, where n is 1 – 3 usually.

When zirconium basic carbonate paste is slowly stirred into a mineral spirit solution of octanoic acid, a solution of zirconium dihydroxide octanoate is produced which is of use in oil-based paints as a dryer catalyst.

Short-chain carboxylates may be soluble, i.e., zirconium dihydroxide diacetate. Longer aliphatic chain carboxylates (zirconium soaps) are water insoluble [105].

2.20. Zirconium Alkoxides

Zirconium alkoxides are prepared by adding zirconium tetrachloride to anhydrous alcohols [106]:

$$ZrCl_4 + C_3H_7OH \rightarrow ZrCl_2(OC_3H_7)_2 \cdot C_3H_7OH + 2\,HCl$$

If there is no steric hindrance, all four chlorines can be exchanged if ammonia is added to combine with the hydrogen chloride.

Alkoxides of different alcohols can be formed by alcohol exchange. This can be facilitated by removal of the more volatile alcohol, or a lower-boiling azeotrope with an inert solvent such as benzene. The conversion of zirconium

tetramethoxide by refluxing with tertiary butanol proceeds only to zirconium methoxide tri-*tert*-butoxide because of steric factors. The following order was found in the interchange of alkoxy groups:

$$MeO > EtO > Pr^{i}O > Bu^{t}O$$

The reaction of *tert*-butanol with zirconium tetrahydridoborate yielded a double alkoxide complex of zirconium *tert*-butoxide with bis(*tert*-butoxy) borane [94].

Zirconium alkoxides hydrolyze quite easily. This provides a route to high-purity, high-surface-area zirconium oxide.

3. Analysis

Analysis for zirconium in ore, soil, vegetation in compounds, and as a minor alloying constituent involves bringing the zirconium into solution, then precipitating with mandelic acid [107] and firing the precipitate to oxide. For the determination of zirconium in air samples, plasma emission spectroscopy is used.

Methods to determine the hafnium associated with the zirconium include atomic absorption spectroscopy, emission spectroscopy, mass spectroscopy, X-ray spectroscopy, and neutron activation.

Most impurities in zirconium and zirconium compounds are determined by emission spectroscopy, with arc-spark [108] or plasma excitation [109], [110]. Atomic absorption is also used to determine a variety of impurities [111].

Carbon and sulfur in zirconium are measured by combustion followed by chromatographic or IR determination of the oxides [112]. Hydrogen is determined by thermal conductivity after hot-vacuum extraction, or by fusion of zirconium with a transition metal in an inert atmosphere and subsequent separation by gas chromatography [113], [114].

Simultaneous determination of oxygen and nitrogen is accomplished by fusion of zirconium with a transition metal in the presence of carbon in an inert atmosphere and subsequent separation by gas chromatography [115], [116]. Nitrogen can be determined by the Kjeldahl technique [112].

Phosphorus may be determined by visible spectrophotometry using molybdenum blue [112] but can be determined more rapidly by phosphine evolution and flame emission spectroscopy.

Chloride at low level is determined indirectly by precipitation with silver nitrate and determination of the silver by X-ray or atomic absorption spectroscopy. Fluoride and higher levels of chloride can be measured by selective ion electrode techniques [117], [118].

4. Toxicology

Zirconium and its compounds are generally of low toxicity. However, the strongly acidic compounds such as zirconium tetrabromide, zirconium tetrachloride, zirconium tetrafluoride, zirconium tetraiodide, zirconium tetranitrate and zirconium sulfate are strongly corrosive to eyes, lungs, skin, mouth, esophagus, and stomach. Exposure to these compounds should be avoided, not because they contain zirconium, but because of the acidic nature of their hydrolysis products generated upon contact with animal or plant tissue.

Inhalation of zirconium materials in the workplace has not resulted in significant toxicity in 40 years of considerable exposure if reasonable care was used. One case of pulmonary fibrosis, believed to have been induced by heavy and continued exposure to zirconium oxide dust while preparing slurries for lens polishing, developed after a 15-year latency period [119]. Severe respiratory tract irritation from inhaling zirconium acetoacetonate disappeared after exposure ceased [120]. In 1978 the FDA banned the production of antiperspirant aerosols containing zirconium because of concern about possible lung granulomas.

Inhalation laboratory tests with insoluble zirconium oxide varied. Some showed no effect, some caused fibrosis. Russian tests with the insoluble compounds zirconium boride [121] and zirconium hydride [122] both showed fibrogenic action in the lungs of albino rats. Aerosols of soluble zirconium dihydroxide chloride and zirconium dihydroxide nitrate each caused intense irritation of upper breathing passages similar to the action of strong acid. The animals died with 30 – 40 min. Sodium zirconium lactate caused

lung granulomatosis in rabbits [123], and it is indicated that a complex of zirconium chlorhydrate and aluminum chlorhydrate also induces granulomatous lung changes [124].

Antiperspirants and poison ivy remedies containing zirconium have caused skin granulomas as a delayed hypersensitivity reaction in a few users [125].

Toxicity tests on laboratory animals showed that the acute toxicity of inorganic zirconium salts is very low when administered orally. Administered intraperitoneally, both organic and inorganic compounds were 2 – 20 times more toxic.

An excellent summation of zirconium toxicology through 1975 is available [126]. Toxicity data for some zirconium compounds is shown in Table 5.

Table 5. Toxicity data (rat) for some zirconium compounds [127]

Compound	Oral LD_{50}, mg/kg	Inhalation LCLo mg/m^3
Dihydroxide chloride	3500	500 for 30 min
Dihydroxide nitrate	2290	500 for 30 min
Sulfate	3500	
Tetrachloride	1688	3124 ppm HCl for 1 hour
Acetate	4100	

The OSHA exposure limit for zirconium and zirconium compounds as airborne particulate is 5 mg/m^3 TWA and 10 mg/m^3 STEL, as zirconium. In Germany, the MAK is 5 mg/m^3 as zirconium. In Russia, MAK is 6 mg/m^3 as zirconium.

5. Storage and Transportation

High-surface-area zirconium metal such as sponge or powder is a combustible solid which should be kept dry and stored away from sources of ignition and oxidizers.

The Department of Transportation has classified all dry zirconium powder as "spontaneously combustible", hazard class 4.2, with identification number UN 2008. The powders are subdivided into packing group I, II, or III, as evaluated by their burning rate in a standardized testing procedure. Group I is for very fine zirconium powder which is a pyrophoric material. Group I powder may not be shipped by air.

Some zirconium powder, usually fuse powder, is recovered wet and shipped wet. This powder is classified as a flammable solid, hazard class 4.1, UN 1358.

Zirconium hydride is classified as a flammable solid, hazard class 4.1, UN 1437.

Zirconium nitrate is classified as an oxidizer, hazard class 5.1, UN 2728.

Zirconium sulfate is classified as corrosive, hazard class 8, NA 9163.

Zirconium tetrachloride is classified as corrosive, hazard class 8, UN 2503.

6. References

1. A. E. van Arkel, J. H. de Boer, *Z. Anorg. Allg. Chem.* **148** (1925) 345 – 350.
2. J. H. De Boer, J. D. Fast, *Z. Anorg. Allg. Chem.* **153** (1926) 18.
3. W. J. Kroll, *J. Franklin Inst.* **260** (1955) 169 – 192.
4. O. Botstein, A. Rabinkin, M. Talianker, *Scr. Metall.* **15** (1981) 151 – 155.
5. T. L. Yau in C. S. Young, J. C. Durham (eds): *Industrial Applications of Titanium and Zirconium: Fourth Conference STP 917.* American Society for Testing and Materials. Philadelphia 1986, pp. 57 – 68.
6. T. L. Yau, R. T. Webster in: *Metals Handbook,* 9th ed.,**vol. 13**, Corrosion, ASM International, Metals Park, Ohio, 1987, pp. 707 – 721.
7. B. Mason: *Principles of Geochemistry,* 3rd ed., Wiley & Sons, New York 1966, p. 45.
8. H. L. Gilbert, C. Q. Morrison, A. Jones, A. W. Henderson, *Bur. Mines Invest.* **5091** (1954).
9. H. S. Choi, *Can. Min. Metall, Bull. Trans.* **67** (1965) 65 – 70.
10. P. H. Wilks, *Chem. Eng. Prog.* **68** (1974).
11. Kawecki Chemical, US 2 653 855, 1953 (H. C. Kawecki).
12. A. J. Hudson, A. C. Haskell, Jr., *Electr. Furn. Conf. Proc.* 1958, 211 – 220.
13. G. von Hevesy, *Chem. Rev.* **2** (1925) 1.
14. N. P. Sajin, E. A. Pepelyaeva, *Proc. Int. Conf. Peaceful Uses At. Energy. 8th* 1956, 559 – 562.
15. J. M. Googin in F. R. Bruce, J. M. Fletcher, H. H. Hyman (eds.): *Progress in Nuclear Energy Series 3,* **Process Chemistry, vol. 2,** Pergamon Press, Oxford 1958, pp. 194 – 209.
16. R. H. Nielsen, R. L. Govro, *Bur. Mines, Invest.* **5214,** (1956).

17. N. P. H. Padmanabhan, T. Sreenivas, N. K. Rao, *High Temp. Mater. Process* **9** (1990) no. 2 – 4, 217 –247.
18. L. Moulin, P. Thouvenin, P. Brun in D. G. Franklin, R. B. Adamson (eds.): "Zirconium in the Nuclear Industry: Sixth International Symposium," *ASTM Spec. Tech. Publ.* **824** (1984) 37 – 44.
19. J. A. Megy, H. Freun, *Metall. Trans. B* **10 B** (1979) 413 – 421.
20. F. W. Starratt, *J. Met.* **II** (1959) 441 – 443.
21. A. P. Lamaze, D. Charquet in K. C. Liddell, D. R. Sadoway, R. G. Bautista (eds.): *Refractory Metals: Extraction, Processing and Applications,* The Minerals, Metals & Materials Society, Warrendale, PA, 1990 pp. 231 – 253.
22. S. N. Flengas, G. J. Kipouros, P. Tumidajski, *Met. Maters. Processes* **2** (1990) no. 3, 151 – 177.
23. Z. M. Shapiro in B. Lustman, F. Kerze, Jr. (eds.): *The Metallurgy of Zirconium,* McGraw-Hill, New York 1955, pp. 135 – 215.
24. R. F. Rolston: *Iodide Metals and Metal Iodides.* The Electrochemical Society, New York 1961.
25. ASTM Specification B614:Descaling and Cleaning Zirconium and Zirconium Alloy Surfaces.
26. T. L. Yau in R. T. Webster, C. S. Young (eds.): "Industrial Application of Titanium and Zirconium: Third Conference," *ASTM Spec. Tech. Publ.* **830** (1984) 124.
27. J. H. Schemel: "ASTM Manual on Zirconium and Hafnium," *ASTM Spec. Tech. Publ.* (1977).
28. R. I. Jaffee in Z. M. Shapiro in B. Lustman, F. Kerze, Jr. (eds.): *The Metallurgy of Zirconium,* McGraw-Hill, New York 1955, pp. 26 – 29.
29. D. A. Lorimer, E. J. Baker, M. Succi, D. K. Weber, *Solid State Technol.* **33** (1990) 77 – 79.
30. A. Clearfield, *Rev. Pure Appl. Chem.* **14** (1964) 91 – 108.
31. J. A. Speer in P. A. Ribbe (ed.): *Orthosilicates,* 2nd ed., Mineralogical Society of America, Washington D. C. 1982.
32. R. C. Garvie, R. H. Hannink, R. T. Pascoe, *Nature (London)* **258** (1975) 703 – 704.
33. R. C. Garvie in A. E. Alper (ed.): *High Temperature Oxides,* part II. Academic Press, New York 1970, pp. 117 – 166.
34. A. H. Heuer, L. W. Hobbs (eds.): *Science and Technology of Zirconia,* American Ceramic Society, Columbus, Ohio, 1981.
35. N. Claussen, M. Ruhle, A. H. Heuer (eds.): *Science and Technology of Zirconia II.* American Ceramic Society, Columbus, Ohio, 1984.
36. S. Somiya, N. Yamamotom, H. Yanagida (eds.): *Science and Technology of Zirconia III,* American Ceramic Society, Westerville, Ohio, 1988.
37. S. Meriani, C. Palmonari (eds.): *Zirconia '88 –Advances in Zirconia Science and Technology.* Elsevier Appl. Sci., London 1989.
38. S. P. S. Badwal, M. J. Bannister, R. H. J. Hannink (eds.): *Science and Technology V,* Technomic Publishing Co., Lancaster, Pa 1993.
39. R. Stevens: *Zirconia and Zirconia Ceramics,* 2nd ed., Magnesium Elektron, Manchester, UK, 1986.
40. A. Pissenberger, G. Gritzner, *J. Mater. Sci. Lett* **14** (1995) 1580 – 1582.
41. G. Gritzner, C. Puchner, J. Dusza, *J. Eur. Ceram. Soc.* **15** (1995) 45 – 49.
42. Y. Sugizaki et al., *J. Phys. Soc. Japan,* **54** (1985) 2543 –2551.
43. R. J. Ackerman, E. G. Rauh, C. A. Alexander, *High Temp. Sci.* **7** (1975) 304 – 316.
44. J. F. Menckus, *J. Crystl. Growth* **128** (1993) 13 – 14.
45. G. H. Haertling, C. E. Land. *Am. Ceram. Soc. Bull.* **49** (1970) 411.
46. R. Srinivason, B. H. Davis, *Catal. Lett.* **14** (1992) 165 – 170.
47. G. T. Mamott et al. *J. Mater. Sci.* **26** (1991) 4054 –4061.
48. M. Henry, J. P. Jolivet, J. Livage, *Struct. Bonding (Berlin)* **77** (1992) 153 – 206.
49. R. C. Garvie, *J. Phys. Chem.* **69** (1968) 1238 – 1243.
50. Societe Europeen des Produits Refractaires, US 5 149 510, 1992 (J. Recasens, D. Urffer, P. Ferlanda).
51. W. W. Stephens, H. L. Gilbert, *J. Met.* **194** (1952) 733 – 737.
52. Electro Metallurgical Co., US 2 433 253, 1948 (W. J. Kroll, F. E. Bacon).
53. B. Krebs, *Angew. Chem. Int. Ed. Engl.* **8** (1969) 146 –147.
54. M. Ohashi, S. Yamaaka, Y. Morimoto, M. Hattori, *Bull. Chem. Soc. Jpn.* **60** (1987) 2387 – 2390.
55. S. N. Flengas, P. Pint, *Can. Met.* **8** (1969) 151 –166.
56. A. W. Strauss, J. D. Corbett, *Inorg. Chem.* **9** (1970) 1373 – 1376.
57. D. G. Adolphson, J. D. Corbett, *Inorg. Chem.* **15** (1976) 1820 – 1823.

58. J. D. Corbett, *Pure Appl. Chem.* **56** (1984) 1527 – 1543.
59. H. S. Marek, J. D. Corbett, R. L. Daake, *J. Less Common Metals* **89** (1983) 243 – 249.
60. A. Clearfiled, P. A. Vaughan, ACTA *Crystallogr.* **9** (1956) 555 – 558.
61. K. I. Arsenin, L. A. Malinko, I. A. Sheka, I. Ya Pishcai, Russ. *J. Inorg. Chem. (Engl. Transl.)* **35** (1990) 1327 – 1331.
62. M. D. Atherton, H. Sutcliffe, *J. Less Common Met.* **138** (1988) 63 – 70.
63. Teledyne, WO 94/12435, 1993 (J. A. Sommers).
64. L. M. Seaverson, J. D. Corbett, *Inorg. Chem.* **22** (1983) 3202 – 3210.
65. K. Dehnicke, J. Weidlein, *Angew. Chem. Int. Ed. Engl.* **5** (1966) 1041.
66. R. L. Daake, J. D. Corbett, *Inorg. Chem.* **16** (1977) 2029 – 2033.
67. E. M. Larsen, J. S. Wrazel, L. G. Hoard, *Inorg. Chem.* **21** (1982) 2619 – 2624.
68. D. H. Guthrie, J. D. Corbett, *Inorg. Chem.* **21** (1982) 3290 – 3295.
69. Teledyne, WO 89/1087, 1989 (J. A. Sommers).
70. M. Robinson, K. C. Fuller, *Mater. Res. Bull.* **22** (1987) 1725 – 1732.
71. W. J. S. Craigen, E. G. Joe, G. M. Ritchey, *Can. Met. Q.* **9** (1970) 485 – 492.
72. National Lead, US 2 525 474, 1950 (W. B. Blumenthal).
73. ICI Australia, WO 86/06362, 1986 (K. Ngian, A. J. Hartshoren, D. H. Jenkins).
74. I. J. Bear, W. G. Mumme, *Rev. Pure Appl. Chem.* **21** (1971) 189 – 211.
75. P. J. Squattrito, P. R. Rudolf, A. Clearfield, *Inorg. Chem.* **25** (1987) 4240 – 4244.
76. Hopton Technologies, US 5 472 485, 1995 (V. E. Pandian, C. V. Calcar, R. W. Wolff).
77. Nalco Chemical, US 5 230 774, 1993 (C. S. Greer, N. P. James).
78. T. B. Massalski (ed.): *Binary Alloy Phase Diagrams,* 2nd ed., ASM International, Materials Park, Ohio, 1990, pp. 2078 – 2080.
79. W. M. Mueller, J. P. Blackledge, G. G. Libowitz: *Metal Hydrides,* Academic Press, New York 1968, pp. 241 – 321.
80. J. Rexer, D. T. Peterson in: *Nuclear Metallurgy, International Symposium on Compounds of Interest in Nucelar Reactor Technology,* vol. X, AIME, New York 1964, p. 327.
81. A. W. Struss, J. P. Corbett, *Inorg. Chem.* **16** (1977) 360.
82. R. V. Sara, *J. Am. Ceram. Soc.* **48** (1965) 243 – 247.
83. E. Rudy: *Compendium of Phase Diagram Data,* Air Force Materials Laboratory, Wrigth-Patterson Air Force Base, Ohio, 1969.
84. Norton Co., US 3 161 470, 1958 (J. J. Scott).
85. T. Ogawa, *J. Alloys and Compounds* **203** (1994) 221 –227.
86. W. E. Wang, D. R. Olander, *J. Alloys and Compounds* **224** (1995) 153 – 158.
87. M. A. Eronyan, R. G. Avarbe, *Inorg. Mater. (Engl. Transl.)* **10** (1974) 1850.
88. E. G. Gillan, R. B. Kaner, *Inorg. Chem.* **33** (1994) 5693 – 5700.
89. B. O. Johansson, H. T. G. Hentzell, J. M. E. Harper, J. J. Cuomo, *J. Mater. Res.* **1** (1986) 442 – 451.
90. R. Juza, A. Rabenau, I. Nitschke, *Z. Anorg. Allgem. Chem.* **332** (1964) 1 – 4.
91. R. Fix, R. G. Gordon, D. M. Hoffman, *Chem. Mater.* **3** (1991) 1138 – 1148.
92. J. C. McMullen, W. D. McKee Jr., *Am. Ceram. Soc. Bull.* **44** (1965) 448 – 491.
93. B. D. James, B. E. Smith, *Synth. React. Inorg. Met. Org. Chem.* **4** (1974) 461 – 465.
94. L. A. Petrova, A. P. Borisov, V. D. Makhaev, *Sov. J. Coord. Chem. (Engl. Transl.)* **18** (1992) 425 – 428.
95. A. Clearfield, R. H. Blessing, *J. Inorg. Nucl. Chem.* **36** (1974) 1174 – 1176.
96. L. L. Y. Chang, M. G. Scroger, B. Phillips, *J. Am. Ceram. Soc.* **50** (1967) 211 – 215.
97. T. A. Mary et al., Science **272** (1996) 90 – 92.
98. A. Clearfield, J. A. Stynes, *J. Inorg. Nucl. Chem.* **26** (1964) 117 – 129.
99. U. Costantino in A. Clearfield (ed.): *Inorganic Ion Exchange Materials,* CRC Press, Boca Raton 1982.
100. Y. Watanabe, Y. Matsumura, Y. Izumi, Y. Mizutani, *Bull. Chem. Soc. Jpn.* **47** (1974) 2922 – 2925.
101. R. Sacks, Y. Avigal, E. Banks, *J. Electrochem. Soc.* **129** (1982) 726 – 729.
102. J. Alamo, R. Roy, *J. Am. Ceram. Soc.* **67** (1984) C 78 – 82.
103. V. Korthuis et al. *Chem. Mater.* **7** (1995) 412 – 417.
104. W. B. Blumenthal: *The Chemical Behavior of Zirconium,* D. Van Nostrand, Princeton, 1958, p. 286.
105. W. B. Blumenthal: *The Chemical Behavior of Zirconium,* D. Van Nostrand, Princeton 1958, p. 319.
106. D. C. Bradley, R. C. Mehrota, D. P. Gaur: *Metal Alkoxides,* Academic Press, London 1978, pp. 10 – 41.
107. R. B. Hahn, E. S. Baginski, *Anal. Chim. Acta* **14** (1965) 45 – 47.

108. R. Brayer, R. O'Connell, A. Powell, R. H. Gale, *Appl. Spectrosc.* **15** (1961) 10 – 13.
109. G. L. Beck, O. T. Farmer, *J. Anal. Spectrom.* **3** (1988) 771 – 773.
110. G. L. Beck, J. P. Fraley, Teledyne Wah Chang procedure ASP-OES-1 Rev. 0, 1988.
111. J. Schlewitz, M. Shields, *At. Absorpt. Newsl.* **10** (1971) 39 – 43.
112. R. VanSanten, J. Schlewitz, G. Beck, R. Walsh in F. D. Snell, L. S. Ettre (eds.): *Encyclopedia of Industrial Chemical Analysis,* **vol. 14,** Wiley-Interscience, New York 1971, pp. 103 – 148.
113. R. K. McGeary in: *A Symposium on Zirconium and Zirconium Alloys,* American Society for Metals, Cleveland 1953, pp. 168 – 175.
114. W. G. Guldner, *Talanta* **8** (1961) 191 – 202.
115. W. G. Smiley, *Anal. Chem.* **27** (1955) 1098 –1102.
116. P. Ebling, G. W. Howard, *Anal. Chem.* **32** (1960) 1610 – 1613.
117. J. Surak, D. Fisher, C. Burros, L. Bate, *Anal. Chem.* **32** (1960) 117 – 119.
118. J. Lingane, *Anal. Chem.* **39** (1967) 881.
119. T. Bartter et al. *Arch. Intern. Med.* **151** (1991) 1197 –1201.
120. M. Y. Longley et al.: *A Toxic Hazard Study of Selected Missile Propellants,* Technical Documentary Report no. ARML-TDR-64-28 Biomedical Lab., Aerospace Med. Res. Lab. Wright-Patterson Air Force Base, Ohio, 1964.
121. I. T. Brakhova: *Environmental Hazards of Metals. Toxicity of Powdered Metals and Compounds,* translated from the Russian (original pub. 1971) by S. L. Slep, Consultants Bureau, New York, NY 1975.
122. I. T. Brakhova, G. A. Shkupko, *Gig. Sanit.* **37** (1972) 36 – 39.
123. J. T. Prior, G. A. Cronk, D. D. Ziegler, *Arch. Environ. Health* **1** (1960) 297 – 300.
124. "Aerosol Drug and Cosmetic Products Containing Zirconium. Proposed Determination," *Fed. Regist.* **40** (1975) 24327 – 24344.
125. W. B. Shelley, H. J. Jurley, Jr., in M. Sauter (ed.): *Immunological Diseases,* 2nd ed., Little, Brown & Co., Boston 1971, pp. 722 – 734.
126. I. C. Smith, B. L. Carson: *Trace Metals in the Environment,* **vol. 3,** Zirconium, Ann Arbor Science Publ., Ann Arbor 1978, pp. 173 – 371.
127. U. S. Registry of Toxic Effects of Chemical Substances.

Zone Melting → Crystal Growth

Mathematics in Chemical Engineering

Bruce A. Finlayson, Department of Chemical Engineering, University of Washington, Seattle, Washington, United States (Chap. 1, 2, 3, 4, 5, 6, 7, 8, 9, 11 and 12)

Lorenz T. Biegler, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (Chap. 10)

Ignacio E. Grossmann, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States (Chap. 10)

1. Solution of Equations 3
1.1. Matrix Properties 3
1.2. Linear Algebraic Equations 5
1.3. Nonlinear Algebraic Equations . . . 8
1.4. Linear Difference Equations 10
1.5. Eigenvalues 10
2. Approximation and Integration . . . 11
2.1. Introduction 11
2.2. Global Polynomial Approximation . 11
2.3. Piecewise Approximation 13
2.4. Quadrature 16
2.5. Least Squares 18
2.6. Fourier Transforms of Discrete Data 19
2.7. Two-Dimensional Interpolation and Quadrature 20
3. Complex Variables 20
3.1. Introduction to the Complex Plane . 20
3.2. Elementary Functions 21
3.3. Analytic Functions of a Complex Variable 23
3.4. Integration in the Complex Plane . . 24
3.5. Other Results 26
4. Integral Transforms 27
4.1. Fourier Transforms 27
4.2. Laplace Transforms 31
4.3. Solution of Partial Differential Equations by Using Transforms . . . 35
5. Vector Analysis 38
6. Ordinary Differential Equations as Initial Value Problems 47
6.1. Solution by Quadrature 48
6.2. Explicit Methods 48
6.3. Implicit Methods 52
6.4. Stiffness 53
6.5. Differential – Algebraic Systems . . . 54
6.6. Computer Software 55
6.7. Stability, Bifurcations, Limit Cycles 56
6.8. Sensitivity Analysis 58
6.9. Molecular Dynamics 59
7. Ordinary Differential Equations as Boundary Value Problems 59
7.1. Solution by Quadrature 60
7.2. Initial Value Methods 60
7.3. Finite Difference Method 61
7.4. Orthogonal Collocation 63
7.5. Orthogonal Collocation on Finite Elements 67
7.6. Galerkin Finite Element Method . . 69
7.7. Cubic B-Splines 71
7.8. Adaptive Mesh Strategies 71
7.9. Comparison 71
7.10. Singular Problems and Infinite Domains 72
8. Partial Differential Equations 73
8.1. Classification of Equations 73
8.2. Hyperbolic Equations 75
8.3. Parabolic Equations in One Dimension 77
8.4. Elliptic Equations 82
8.5. Parabolic Equations in Two or Three Dimensions 84
8.6. Special Methods for Fluid Mechanics 84
8.7. Computer Software 86
9. Integral Equations 87
9.1. Classification 87
9.2. Numerical Methods for Volterra Equations of the Second Kind 89
9.3. Numerical Methods for Fredholm, Urysohn, and Hammerstein Equations of the Second Kind 89
9.4. Numerical Methods for Eigenvalue Problems 90
9.5. Green's Functions 90
9.6. Boundary Integral Equations and Boundary Element Method 92
10. Optimization 93
10.1. Introduction 93
10.2. Gradient-Based Nonlinear Programming 94
10.3. Optimization Methods without Derivatives 103
10.4. Global Optimization 104
10.5. Mixed Integer Programming 108
10.6. Dynamic Optimization 119

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_01.pub2

10.7. Development of Optimization Models 122
11. Probability and Statistics 123
11.1. Concepts 123
11.2. Sampling and Statistical Decisions . 127
11.3. Error Analysis in Experiments 130
11.4. Factorial Design of Experiments and Analysis of Variance 131
12. Multivariable Calculus Applied to Thermodynamics 133
12.1. State Functions 133
12.2. Applications to Thermodynamics . . 134
12.3. Partial Derivatives of All Thermodynamic Functions 135
13. References 136

Symbols

Variables

a	1, 2, or 3 for planar, cylindrical, spherical geometry
a_i	acceleration of i-th particle in molecular dynamics
A	cross-sectional area of reactor; Helmholtz free energy in thermodynamics
Bi	Biot number
Bim	Biot number for mass
c	concentration
C_p	heat capacity at constant pressure
C_s	heat capacity of solid
C_v	heat capacity at constant volume
Co	Courant number
D	diffusion coefficient
Da	Damköhler number
D_e	effective diffusivity in porous catalyst
E	efficiency of tray in distillation column
F	molar flow rate into a chemical reactor
G	Gibbs free energy in thermodynamics
h_p	heat transfer coefficient
H	enthalpy in thermodynamics
J	mass flux
k	thermal conductivity; reaction rate constant
k_e	effective thermal conductivity of porous catalyst
k_g	mass transfer coefficient
K	chemical equilibrium constant
L	thickness of slab; liquid flow rate in distillation column; length of pipe for flow in pipe
m_i	mass of i-th particle in molecular dynamics
M	holdup on tray of distillation column
n	power-law exponent in viscosity formula for polymers
p	pressure
Pe	Peclet number
q	heat flux
Q	volumetric flow rate; rate of heat generation for heat transfer problems
r	radial position in cylinder
$\boldsymbol{r}_i$	position of i-th particle in molecular dynamics
R	radius of reactor or catalyst pellet
Re	Reynolds number
S	entropy in thermodynamics
Sh	Sherwood number
t	time
T	temperature
$\boldsymbol{u}$	velocity
U	internal energyy in thermodynamics
$\boldsymbol{v}_i$	velocity of i-th particle in molecular dynamics
V	volume of chemical reactor; vapor flow rate in distillation column; potential energy in molecular dynamics; specific volume in thermodynamics
x	position
z	position from inlet of reactor

Greek symbols

a	thermal diffusivity
δ	Kronecker delta
Δ	sampling rate
ε	porosity of catalyst pellet
η	viscosity in fluid flow; effectiveness factor for reaction in a catalyst pellet
η_0	zero-shear rate viscosity
ϕ	Thiele modulus
φ	void fraction of packed bed
λ	time constant in polymer flow
ρ	density
ρ_s	density of solid
τ	shear stress
μ	viscosity

Special symbols

$\vert$	subject to
:	mapping. For example, $h: R^n \to R^m$, states that functions h map real numbers into m real numbers. There are m functions h written in terms of n variables
$\in$	member of
$\to$	maps into

1. Solution of Equations

Mathematical models of chemical engineering systems can take many forms: they can be sets of algebraic equations, differential equations, and/or integral equations. Mass and energy balances of chemical processes typically lead to large sets of *algebraic* equations:

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

Mass balances of stirred tank reactors may lead to ordinary *differential* equations:

$$\frac{\mathrm{d}\,y}{\mathrm{d}\,t} = f\ [y\,(t)]$$

Radiative heat transfer may lead to *integral equations:*

$$y\,(x) = g\,(x) + \lambda \int_0^1 K\,(x, s)\, f\ (s)\,\mathrm{d}\,s$$

Even when the model is a differential equation or integral equation, the most basic step in the algorithm is the solution of sets of algebraic equations. The solution of sets of algebraic equations is the focus of Chapter 1.

A single linear equation is easy to solve for either x or y:

$$y = ax + b$$

If the equation is nonlinear,

$$f\,(x) = 0$$

it may be more difficult to find the x satisfying this equation. These problems are compounded when there are more unknowns, leading to simultaneous equations. If the unknowns appear in a linear fashion, then an important consideration is the structure of the matrix representing the equations; special methods are presented here for special structures. They are useful because they increase the speed of solution. If the unknowns appear in a nonlinear fashion, the problem is much more difficult. Iterative techniques must be used (i.e., make a guess of the solution and try to improve the guess). An important question then is whether such an iterative scheme converges. Other important types of equations are linear difference equations and eigenvalue problems, which are also discussed.

1.1. Matrix Properties

A matrix is a set of real or complex numbers arranged in a rectangular array.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

The numbers a_{ij} are the elements of the matrix $\boldsymbol{A}$, or $(\mathrm{A})_{ij} = a_{ij}$. The transpose of $\boldsymbol{A}$ is $(\boldsymbol{A}^T) = a_{ji}$.

The determinant of a square matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

If the i-th row and j-th column are deleted, a new determinant is formed having $n-1$ rows and columns. This determinant is called the minor of a_{ij} denoted as M_{ij}. The cofactor A'_{ij} of the element a_{ij} is the signed minor of a_{ij} determined by

$$A'_{ij} = (-1)^{i+j} M_{ij}$$

The value of $|\boldsymbol{A}|$ is given by

$$|\boldsymbol{A}| = \sum_{j=1}^{n} a_{ij} A'_{ij} \text{ or } \sum_{i=1}^{n} a_{ij} A'_{ij}$$

where the elements a_{ij} must be taken from a single row or column of $\boldsymbol{A}$.

If all determinants formed by striking out whole rows or whole columns of order greater than r are zero, but there is at least one determinant of order r which is not zero, the matrix has rank r.

The value of a determinant is not changed if the rows and columns are interchanged. If the elements of one row (or one column) of a determinant are all zero, the value of the determinant is zero. If the elements of one row or column are multiplied by the same constant, the determinant is the previous value times that constant. If two adjacent rows (or columns) are interchanged, the value of the new determinant is the negative of the value of the original determinant. If two rows (or columns) are identical, the determinant is zero. The value of a determinant is not changed if one row (or column) is multiplied by a constant and added to another row (or column).

A matrix is *symmetric* if

$$a_{ij}=a_{ji}$$

and it is *positive definite* if

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j\geq 0$$

for all $\boldsymbol{x}$ and the equality holds only if $\boldsymbol{x}=0$.

If the elements of $\boldsymbol{A}$ are complex numbers, $\boldsymbol{A}^*$ denotes the complex conjugate in which $(A^*)_{ij}=a^*_{ij}$. If $\boldsymbol{A}=\boldsymbol{A}^*$ the matrix is Hermitian.

The *inverse of a matrix* can also be used to solve sets of linear equations. The inverse is a matrix such that when $\boldsymbol{A}$ is multiplied by its inverse the result is the identity matrix, a matrix with 1.0 along the main diagonal and zero elsewhere.

$$\boldsymbol{A}\boldsymbol{A}^{-1}=\boldsymbol{I}$$

If $\boldsymbol{A}^T=\boldsymbol{A}^{-1}$ the matrix is orthogonal.

Matrices are added and subtracted element by element.

$$\boldsymbol{A}+\boldsymbol{B} \text{ is } a_{ij}+b_{ij}$$

Two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are equal if $a_{ij}=b_{ij}$. Special relations are

$$(\boldsymbol{AB})^{-1}=\boldsymbol{B}^{-1}\boldsymbol{A}^{-1},\quad (\boldsymbol{AB})^T=\boldsymbol{B}^T\boldsymbol{A}^T$$
$$(\boldsymbol{A}^{-1})^T=(\boldsymbol{A}^T)^{-1},\quad (\boldsymbol{ABC})^{-1}=\boldsymbol{C}^{-1}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$$

A diagonal matrix is zero except for elements along the diagonal.

$$a_{ij}=\begin{cases} a_{ii}, & i=j \\ 0, & i\neq j \end{cases}$$

A tridiagonal matrix is zero except for elements along the diagonal and one element to the right and left of the diagonal.

$$a_{ij}=\begin{cases} 0 \text{ if } j<i-1 \\ a_{ij} \text{ otherwise} \\ 0 \text{ if } j>i+1 \end{cases}$$

Block diagonal and pentadiagonal matrices also arise, especially when solving partial differential equations in two- and three-dimensions.

QR Factorization of a Matrix. If A is an $m\times n$ matrix with $m\geq n$, there exists an $m\times m$ unitary matrix $\boldsymbol{Q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\ldots\boldsymbol{q}_m]$ and an $m\times n$ right-triangular matrix $\boldsymbol{R}$ such that $\boldsymbol{A}=\boldsymbol{QR}$. The QR factorization is frequently used in the actual computations when the other transformations are unstable.

Singular Value Decomposition. If $\boldsymbol{A}$ is an $m\times n$ matrix with $m\geq n$ and rank $k\leq n$, consider the two following matrices.

$$\boldsymbol{A}\boldsymbol{A}^* \text{ and } \boldsymbol{A}^*\boldsymbol{A}$$

An $m\times m$ unitary matrix $\boldsymbol{U}$ is formed from the eigenvectors $\boldsymbol{u}_i$ of the first matrix.

$$\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\ldots\boldsymbol{u}_m]$$

An $n\times n$ unitary matrix $\boldsymbol{V}$ is formed from the eigenvectors $\boldsymbol{v}_i$ of the second matrix.

$$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\ldots,\boldsymbol{v}_n]$$

Then the matrix $\boldsymbol{A}$ can be decomposed into

$$\boldsymbol{A}=\boldsymbol{U}\Sigma\boldsymbol{V}^*$$

where Σ is a $k\times k$ diagonal matrix with diagonal elements $d_{ii}=\sigma_i>0$ for $1\leq i\leq k$. The eigenvalues of $\Sigma^*\Sigma$ are σ_i^2. The vectors $\boldsymbol{u}_i$ for $k+1\leq i\leq m$ and $\boldsymbol{v}_i$ for $k+1\leq i\leq n$ are eigenvectors associated with the eigenvalue zero; the eigenvalues for $1\leq i\leq k$ are σ_i^2. The values of σ_i are called the singular values of the matrix $\boldsymbol{A}$. If $\boldsymbol{A}$ is real, then $\boldsymbol{U}$ and $\boldsymbol{V}$ are real, and hence orthogonal matrices. The value of the singular value decomposition comes when a process is represented by a linear transformation and the elements of $\boldsymbol{A}$, a_{ij}, are the contribution to an output i for a particular variable as input variable j. The input may be the size of a disturbance, and the output is the gain [1]. If the rank is less than n, not all the variables are independent and they cannot all be controlled. Furthermore, if the singular values are widely separated, the process is sensitive to small changes in the elements of the matrix and the process will be difficult to control.

1.2. Linear Algebraic Equations

Consider the $n \times n$ linear system

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n &= f_1 \\ a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n &= f_2 \\ &\ldots \\ a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n &= f_n \end{aligned}$$

In this equation $a_{11}, \ldots, a_{nn}$ are known parameters, $f_1, \ldots, f_n$ are known, and the unknowns are $x_1, \ldots, x_n$. The values of all unknowns that satisfy every equation must be found. This set of equations can be represented as follows:

$$\sum_{j=1}^{n} a_{ij}\,x_j = f_j \text{ or } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$$

The most efficient method for solving a set of linear algebraic equations is to perform a lower – upper (LU) decomposition of the corresponding matrix $\boldsymbol{A}$. This decomposition is essentially a Gaussian elimination, arranged for maximum efficiency [2, 3].

The LU decomposition writes the matrix as

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$$

The $\boldsymbol{U}$ is upper triangular; it has zero elements below the main diagonal and possibly nonzero values along the main diagonal and above it (see Fig. 1). The $\boldsymbol{L}$ is lower triangular. It has the value 1 in each element of the main diagonal, nonzero values below the diagonal, and zero values above the diagonal (see Fig. 1). The original problem can be solved in two steps:

$$\boldsymbol{L}\boldsymbol{y} = \boldsymbol{f}, \boldsymbol{U}\boldsymbol{x} = \boldsymbol{y} \text{ solves } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{L}\boldsymbol{U}\boldsymbol{x} = \boldsymbol{f}$$

Each of these steps is straightforward because the matrices are upper triangular or lower triangular.

When f is changed, the last steps can be done without recomputing the LU decomposition. Thus, multiple right-hand sides can be computed efficiently. The number of multiplications and divisions necessary to solve for m right-hand sides is:

$$\text{Operation count } = \frac{1}{3}n^3 - \frac{1}{3}n + m\,n^2$$

The *determinant* is given by the product of the diagonal elements of $\boldsymbol{U}$. This should be calculated as the LU decomposition is performed.

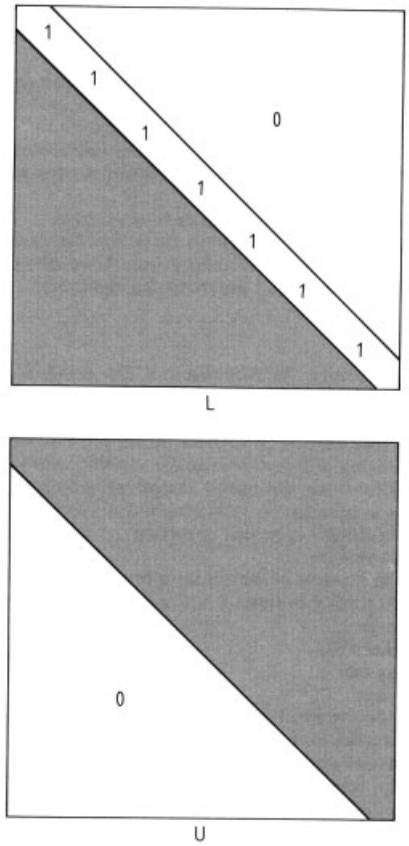

Figure 1. Structure of $\boldsymbol{L}$ and $\boldsymbol{U}$ matrices

If the value of the determinant is a very large or very small number, it can be divided or multiplied by 10 to retain accuracy in the computer; the scale factor is then accumulated separately. The condition number κ can be defined in terms of the singular value decomposition as the ratio of the largest d_{ii} to the smallest d_{ii} (see above). It can also be expressed in terms of the norm of the matrix:

$$\kappa\,(A) = \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|$$

where the norm is defined as

$$\|\boldsymbol{A}\| \equiv \sup_{x \neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_k \sum_{j=1}^{n} |a_{jk}|$$

If this number is infinite, the set of equations is singular. If the number is too large, the matrix is said to be ill-conditioned. Calculation of

the condition number can be lengthy so another criterion is also useful. Compute the ratio of the largest to the smallest pivot and make judgments on the ill-conditioning based on that.

When a matrix is ill-conditioned the LU decomposition must be performed by using pivoting (or the singular value decomposition described above). With pivoting, the order of the elimination is rearranged. At each stage, one looks for the largest element (in magnitude); the next stages if the elimination are on the row and column containing that largest element. The largest element can be obtained from only the diagonal entries (partial pivoting) or from all the remaining entries. If the matrix is nonsingular, Gaussian elimination (or LU decomposition) could fail if a zero value were to occur along the diagonal and were to be a pivot. With full pivoting, however, the Gaussian elimination (or LU decomposition) cannot fail because the matrix is nonsingular.

The *Cholesky decomposition* can be used for real, symmetric, positive definite matrices. This algorithm saves on storage (divide by about 2) and reduces the number of multiplications (divide by 2), but adds n square roots.

The linear equations are solved by

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{f}$$

Generally, the inverse is not used in this way because it requires three times more operations than solving with an LU decomposition. However, if an inverse is desired, it is calculated most efficiently by using the LU decomposition and then solving

$$\boldsymbol{A}\boldsymbol{x}^{(i)} = \boldsymbol{b}^{(i)}$$

$$b_j^{(i)} = \begin{cases} 0 & j \neq i \\ 1 & j = i \end{cases}$$

Then set

$$\boldsymbol{A}^{-1} = \left(\boldsymbol{x}^{(1)} | \boldsymbol{x}^{(2)} | \boldsymbol{x}^{(3)} | \cdots | \boldsymbol{x}^{(n)}\right)$$

Solutions of Special Matrices. Special matrices can be handled even more efficiently. A *tridiagonal matrix* is one with nonzero entries along the main diagonal, and one diagonal above and below the main one (see Fig. 2). The corresponding set of equations can then be written as

$$a_i x_{i-1} + b_i x_i + c_i x_{i+1} = d_i$$

The LU decomposition algorithm for solving this set is

$b'_1 = b_1$
for $k = 2, n$ do
$a'_k = \frac{a_k}{b'_{k-1}}$, $b'_k = b_k - \frac{a_k}{b'_{k-1}} c_{k-1}$
enddo
$d'_1 = d_1$
for $k = 2, n$ do
$d'_k = d_k - a'_k d'_{k-1}$
enddo
$x_n = d'_n / b'_n$
for $k = n-1, 1$do
$x_k = \frac{d'_k - c_k x_{k+1}}{b'_k}$
enddo

The number of multiplications and divisions for a problem with n unknowns and m right-hand sides is

$$\text{Operation count} = 2(n-1) + m(3n-2)$$

If

$$|b_i| > |a_i| + |c_i|$$

no pivoting is necessary. For solving two-point boundary value problems and partial differential equations this is often the case.

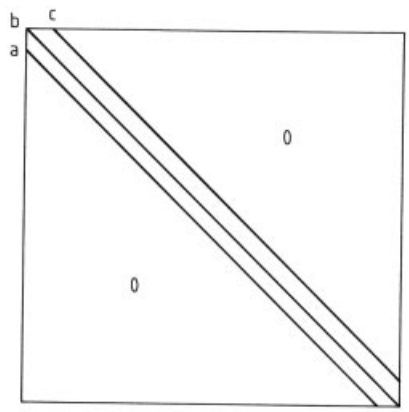

Figure 2. Structure of tridiagonal matrices

Sparse matrices are ones in which the majority of elements are zero. If the zero entries occur in special patterns, efficient techniques can be used to exploit the structure, as was done above for tridiagonal matrices, block tridiagonal

matrices, arrow matrices, etc. These structures typically arise from numerical analysis applied to solve differential equations. Other problems, such as modeling chemical processes, lead to sparse matrices but without such a neatly defined structure—just a lot of zeros in the matrix. For matrices such as these, special techniques must be employed: efficient codes are available [4]. These codes usually employ a symbolic factorization, which must be repeated only once for each structure of the matrix. Then an LU factorization is performed, followed by a solution step using the triangular matrices. The symbolic factorization step has significant overhead, but this is rendered small and insignificant if matrices with exactly the same structure are to be used over and over [5].

The efficiency of a technique for solving sets of linear equations obviously depends greatly on the arrangement of the equations and unknowns because an efficient arrangement can reduce the bandwidth, for example. Techniques for renumbering the equations and unknowns arising from elliptic partial differential equations are available for finite difference methods [6] and for finite element methods [7].

Solutions with Iterative Methods. Sets of linear equations can also be solved by using iterative methods; these methods have a rich historical background. Some of them are discussed in Chapter 8 and include Jacobi, Gauss – Seidel, and overrelaxation methods. As the speed of computers increases, direct methods become preferable for the general case, but for large three-dimensional problems iterative methods are often used.

The *conjugate gradient method* is an iterative method that can solve a set of *n* linear equations in *n* iterations. The method primarily requires multiplying a matrix by a vector, which can be done very efficiently on parallel computers: for sparse matrices this is a viable method. The original method was devised by Hestenes and Stiefel [8]; however, more recent implementations use a *preconditioned* conjugate gradient method because it converges faster, provided a good "preconditioner" can be found. The system of *n* linear equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$$

where $\boldsymbol{A}$ is symmetric and positive definite, is to be solved. A preconditioning matrix $\boldsymbol{M}$ is defined in such a way that the problem

$$\boldsymbol{M}\boldsymbol{t} = \boldsymbol{r}$$

is easy to solve exactly ($\boldsymbol{M}$ might be diagonal, for example). Then the preconditioned conjugate gradient method is

Guess $\boldsymbol{x}_0$

Calculate $\boldsymbol{r}_0 = \boldsymbol{f} - \boldsymbol{A}\boldsymbol{x}_0$

Solve $\boldsymbol{M}\,\boldsymbol{t}_0 = \boldsymbol{r}_0$, and set $\boldsymbol{p}_0 = \boldsymbol{t}_0$

for $k = 1, n$ (or until convergence)

$a_k = \frac{\boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{t}_k}{\boldsymbol{p}_k^{\mathrm{T}}\boldsymbol{A}\boldsymbol{p}_k}$

$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + a_k\boldsymbol{p}_k$

$\boldsymbol{r}_{k+1} = \boldsymbol{r}_k - \boldsymbol{a}_k\boldsymbol{A}\boldsymbol{p}_k$

Solve $\boldsymbol{M}t_{k+1} = \boldsymbol{r}_{k+1}$

$b_k = \frac{\boldsymbol{r}_{k+1}^{\mathrm{T}}\boldsymbol{t}_{k+1}}{\boldsymbol{r}_k^{\mathrm{T}}\boldsymbol{t}_k}$

$\boldsymbol{p}_{k+1} = \boldsymbol{t}_{k+1} + b_k\boldsymbol{p}_k$

test for convergence

enddo

Note that the entire algorithm involves only matrix multiplications. The generalized minimal residual method (GMRES) is an iterative method that can be used for nonsymmetric systems and is based on a modified Gram – Schmidt orthonormalization. Additional information, including software for a variety of methods, is available [9 – 13].

In dimensional analysis if the dimensions of each physical variable P_j (there are *n* of them) are expressed in terms of fundamental measurement units m_j (such as time, length, mass; there are *m* of them):

$$[P_j] = m_1^{\alpha_{1j}} m_2^{\alpha_{2j}} \cdots m_m^{\alpha_{mj}}$$

then a matrix can be formed from the α_{ij}. If the rank of this matrix is *r*, $n - r$ independent dimensionless groups govern that phenomenon. In chemical reaction engineering the chemical reaction stoichiometry can be written as

$$\sum_{i=1}^{n} \alpha_{ij}\, C_i = 0,\ j = 1, 2, \ldots, m$$

where there are *n* species and *m* reactions. Then if a matrix is formed from the coefficients α_{ij}, which is an $n \times m$ matrix, and the rank of the matrix is *r*, there are *r* independent chemical reactions. The other $n - r$ reactions can be deduced from those *r* reactions.

1.3. Nonlinear Algebraic Equations

Consider a single nonlinear equation in one unknown,

$$f(x) = 0$$

In Microsoft Excel, roots are found by using Goal Seek or Solver. Assign one cell to be x, put the equation for $f(x)$ in another cell, and let Goal Seek or Solver find the value of x making the equation cell zero. In MATLAB, the process is similar except that a function (m-file) is defined and the command fzero('f',x0) provides the solution x, starting from the initial guess x_0.

Iterative methods applied to single equations include the *successive substitution* method

$$x^{k+1} = x^k + \beta f\left(x^k\right) \equiv g\left(x^k\right)$$

and the Newton – Raphson method.

$$x^{k+1} = x^k - \frac{f\left(x^k\right)}{\mathrm{d}f/\mathrm{d}x\left(x^k\right)}$$

The former method converges if the derivative of $g(x)$ is bounded [3]. The latter method

$$\left|\frac{\mathrm{d}g}{\mathrm{d}x}(x)\right| \leq \mu \text{ for } |x-\alpha| < h$$

is based on a Taylor series of the equation about the k-th iterate:

$$f\left(x^{k+1}\right) = f\left(x^k\right) + \frac{\mathrm{d}f}{\mathrm{d}x}\Big|_{x^k}\left(x^{k+1}-x^k\right) + \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\Big|_{x^k}\frac{1}{2}\left(x^{k+1}-x^k\right)^2 + \cdots$$

The second and higher-order terms are neglected and $f\left(x^{k+1}\right) = 0$ to obtain the method.

$$\left|\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x^0} > 0,\ \left|x^1 - x^0\right| = \left|\frac{f\left(x^0\right)}{\mathrm{d}f/\mathrm{d}x\left(x^0\right)}\right|_{x^0} \leq b, \quad \text{and } \left|\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right| \leq c$$

Convergence of the Newton–Raphson method depends on the properties of the first and second derivative of $f(x)$ [3, 14]. In practice the method may not converge unless the initial guess is good, or it may converge for some parameters and not others. Unfortunately, when the method is nonconvergent the results look as though a mistake occurred in the computer programming; distinguishing between these situations is difficult, so careful programming and testing are required. If the method converges the difference between successive iterates is something like 0.1, 0.01, 0.0001, 10^{-8}. The error (when it is known) goes the same way; the method is said to be quadratically convergent when it converges. If the derivative is difficult to calculate a numerical approximation may be used.

$$\frac{\mathrm{d}f}{\mathrm{d}x}\Big|_{x^k} = \frac{f\left(x^k+\varepsilon\right) - f\left(x^k\right)}{\varepsilon}$$

In the *secant method* the same formula is used as for the Newton – Raphson method, except that the derivative is approximated by using the values from the last two iterates:

$$\frac{\mathrm{d}f}{\mathrm{d}x}\Big|_{x^k} = \frac{f\left(x^k\right) - f\left(x^{k-1}\right)}{x^k - x^{k-1}}$$

This is equivalent to drawing a straight line through the last two iterate values on a plot of $f(x)$ versus x. The Newton – Raphson method is equivalent to drawing a straight line tangent to the curve at the last x. In the *method of false position* (or regula falsi), the secant method is used to obtain x^{k+1}, but the previous value is taken as either x^{k-1} or x^k. The choice is made so that the function evaluated for that choice has the opposite sign to $f\left(x^{k+1}\right)$. This method is slower than the secant method, but it is more robust and keeps the root between two points at all times. In all these methods, appropriate strategies are required for bounds on the function or when $\mathrm{d}f/\mathrm{d}x = 0$. *Brent's method* combines bracketing, bisection, and an inverse quadratic interpolation to provide a method that is fast and guaranteed to converge, if the root can be bracketed initially [15, p. 251].

In the *method of bisection*, if a root lies between x_1 and x_2 because $f(x_1) < 0$ and $f(x_2) > 0$, then the function is evaluated at the center, $x_c = 0.5(x_1 + x_2)$. If $f(x_c) > 0$, the root lies between x_1 and x_c. If $f(x_c) < 0$, the root lies between x_c and x_2. The process is then repeated. If $f(x_c) = 0$, the root is x_c. If $f(x_1) > 0$ and $f(x_2) > 0$, more than one root may exist between x_1 and x_2 (or no roots).

For systems of equations the Newton – Raphson method is widely used, especially for equations arising from the solution of differential equations.

$$f_i\left(\{x_j\}\right) = 0,\ 1 \leq i, j \leq n,$$
$$\text{where } \{x_j\} = (x_1, x_2, \ldots, x_n) = \boldsymbol{x}$$

Then, an expansion in several variables occurs:

$$\boldsymbol{f}_i\left(\boldsymbol{x}^{k+1}\right)=f_i\left(\boldsymbol{x}^k\right)+\sum_{j=1}^{n}\frac{\partial f_i}{\partial x_j}|_{\boldsymbol{x}^k}\left(x_j^{k+1}-x_j^k\right)+\cdots$$

The Jacobian matrix is defined as

$$\boldsymbol{J}_{ij}^k=\left.\frac{\partial f_i}{\partial x_j}\right|_{\boldsymbol{x}^k}$$

and the Newton – Raphson method is

$$\sum_{j=1}^{n}\boldsymbol{J}_{ij}^k\left(x^{k+1}-x^k\right)=-f_i\left(\boldsymbol{x}^k\right)$$

For convergence, the norm of the inverse of the Jacobian must be bounded, the norm of the function evaluated at the initial guess must be bounded, and the second derivative must be bounded [14, p. 115], [3, p. 12].

A review of the usefulness of solution methods for nonlinear equations is available [16]. This review concludes that the Newton – Raphson method may not be the most efficient. Broyden's method approximates the inverse to the Jacobian and is a good all-purpose method, but a good initial approximation to the Jacobian matrix is required. Furthermore, the rate of convergence deteriorates for large problems, for which the Newton – Raphson method is better. Brown's method [16] is very attractive, whereas Brent's is not worth the extra storage and computation. For large systems of equations, efficient software is available [11 – 13].

Homotopy methods can be used to ensure finding the solution when the problem is especially complicated. Suppose an attempt is made to solve $f(\boldsymbol{x}) = 0$, and it fails; however, $g(\boldsymbol{x}) = 0$ can be solved easily, where $g(\boldsymbol{x})$ is some function, perhaps a simplification of $f(\boldsymbol{x})$. Then, the two functions can be embedded in a homotopy by taking

$$\boldsymbol{h}(\boldsymbol{x},t)=t\,f(\boldsymbol{x})+(1-t)\,g(x)$$

In this equation, $\boldsymbol{h}$ can be a $n\times n$ matrix for problems involving n variables; then x is a vector of length n. Then $h(x, t) = 0$ can be solved for $t = 0$ and t gradually changes until at $t = 1$, $\boldsymbol{h}(\boldsymbol{x}, 1) = f(\boldsymbol{x})$. If the Jacobian of $\boldsymbol{h}$ with respect to x is nonsingular on the homotopy path (as t varies), the method is guaranteed to work. In classical methods, the interval from $t = 0$ to $t = 1$ is broken up into N subdivisions. Set $\Delta t = 1/N$ and solve for $t = 0$, which is easy by the choice of $g(x)$. Then set $t = \Delta t$ and use the Newton – Raphson method to solve for x. Since the initial guess is presumably pretty good, this has a high chance of being successful. That solution is then used as the initial guess for $t = 2\,\Delta t$ and the process is repeated by moving stepwise to $t = 1$. If the Newton – Raphson method does not converge, then Δt must be reduced and a new solution attempted.

Another way of using homotopy is to create an ordinary differential equation by differentiating the homotopy equation along the path (where $\boldsymbol{h} = 0$).

$$\frac{\mathrm{d}\boldsymbol{h}\left[\boldsymbol{x}(t),t\right]}{\mathrm{d}t}=\frac{\partial\boldsymbol{h}}{\partial\boldsymbol{x}}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}+\frac{\partial\boldsymbol{h}}{\partial t}=0$$

This can be expressed as an ordinary differential equation for $x(t)$:

$$\frac{\partial\boldsymbol{h}}{\partial\boldsymbol{x}}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}=-\frac{\partial\boldsymbol{h}}{\partial t}$$

If Euler's method is used to solve this equation, a value x^0 is used, and $\mathrm{d}x/\mathrm{d}t$ from the above equation is solved for. Then

$$\boldsymbol{x}^{1,0}=\boldsymbol{x}^0+\Delta t\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$$

is used as the initial guess and the homotopy equation is solved for $\boldsymbol{x}^1$.

$$\frac{\partial\boldsymbol{h}}{\partial x}\left(\boldsymbol{x}^{1,k+1}-\boldsymbol{x}^{1,k}\right)=-\boldsymbol{h}\left(\boldsymbol{x}^{1,k},t\right)$$

Then t is increased by Δt and the process is repeated.

In arc-length parameterization, both x and t are considered parameterized by a parameter s, which is thought of as the arc length along a curve. Then the homotopy equation is written along with the arc-length equation.

$$\frac{\partial\boldsymbol{h}}{\partial\boldsymbol{x}}\frac{\mathrm{d}x}{\mathrm{d}\boldsymbol{s}}+\frac{\partial\boldsymbol{h}}{\partial t}\frac{\mathrm{d}t}{\mathrm{d}\boldsymbol{s}}=0$$

$$\frac{\mathrm{d}\boldsymbol{x}^{\mathrm{T}}}{\mathrm{d}\boldsymbol{s}}\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}s}+\left(\frac{\mathrm{d}t}{\mathrm{d}\boldsymbol{s}}\right)^2=1$$

The initial conditions are

$$\boldsymbol{x}(0)=\boldsymbol{x}^0$$
$$t(0)=0$$

The advantage of this approach is that it works even when the Jacobian of $\boldsymbol{h}$ becomes singular because the full matrix is rarely singular. Illustrations applied to chemical engineering are available [17]. Software to perform these computations is available (called LOCA) [18].

1.4. Linear Difference Equations

Difference equations arise in chemical engineering from staged operations, such as distillation or extraction, as well as from differential equations modeling adsorption and chemical reactors. The value of a variable in the n-th stage is noted by a subscript n. For example, if $y_{n,i}$ denotes the mole fraction of the i-th species in the vapor phase on the n-th stage of a distillation column, $x_{n,i}$ is the corresponding liquid mole fraction, R the reflux ratio (ratio of liquid returned to the column to product removed from the condenser), and $K_{n,i}$ the equilibrium constant, then the mass balances about the top of the column give

$$y_{n+1,i} = \frac{R}{R+1}x_{n,i}+\frac{1}{R+1}x_{0,i}$$

and the equilibrium equation gives

$$y_{n,i} = K_{n,i}x_{n,i}$$

If these are combined,

$$K_{n+1,i}x_{n+1,i} = \frac{R}{R+1}x_{n,i}+\frac{1}{R+1}x_{0,i}$$

is obtained, which is a linear difference equation. This particular problem is quite complicated, and the interested reader is referred to [19, Chap. 6]. However, the form of the difference equation is clear. Several examples are given here for solving difference equations. More complete information is available in [20].

An equation in the form

$$x_{n+1}-x_n=f_{n+1}$$

can be solved by

$$x_n = \sum_{i=1}^{n} f_i$$

Usually, difference equations are solved analytically only for linear problems. When the coefficients are constant and the equation is linear and homogeneous, a trial solution of the form

$$x_n = \varphi^n$$

is attempted; φ is raised to the power n. For example, the difference equation

$$cx_{n-1}+bx_n+ax_{n+1} = 0$$

coupled with the trial solution would lead to the equation

$$a\varphi^2+b\varphi+c = 0$$

This gives

$$\varphi_{1,2} = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$$

and the solution to the difference equation is

$$x_n = A\varphi_1^n+B\varphi_2^n$$

where A and B are constants that must be specified by boundary conditions of some kind.

When the equation is nonhomogeneous, the solution is represented by the sum of a particular solution and a general solution to the homogeneous equation.

$$x_n = x_{n,P}+x_{n,H}$$

The general solution is the one found for the homogeneous equation, and the particular solution is any solution to the nonhomogeneous difference equation. This can be found by methods analogous to those used to solve differential equations: the method of undetermined coefficients and the method of variation of parameters. The last method applies to equations with variable coefficients, too. For a problem such as

$$x_{n+1}-f_nx_n = 0$$
$$x_0 = c$$

the general solution is

$$x_n=c\prod_{i=1}^{n} f_{i-1}$$

This can then be used in the method of variation of parameters to solve the equation

$$x_{n+1}-f_nx_n = g_n$$

1.5. Eigenvalues

The $n\times n$ matrix $\boldsymbol{A}$ has n eigenvalues λ_i, $i = 1, \ldots, n$, which satisfy

$$\det(\boldsymbol{A}-\lambda_i\boldsymbol{I}) = 0$$

If this equation is expanded, it can be represented as

$$P_n(\lambda) = (-\lambda)^n+a_1(-\lambda)^{n-1}+a_2(-\lambda)^{n-2}+\cdots +a_{n-1}(-\lambda)+a_n = 0$$

If the matrix $\boldsymbol{A}$ has real entries then a_i are real numbers, and the eigenvalues either are real

numbers or occur in pairs as complex numbers with their complex conjugates (for definition of complex numbers, see Chap. 3). The Hamilton – Cayley theorem [19, p. 127] states that the matrix $\boldsymbol{A}$ satisfies its own characteristic equation.

$$P_n(\boldsymbol{A}) = (-\boldsymbol{A})^n + a_1(-\boldsymbol{A})^{n-1} + a_2(-\boldsymbol{A})^{n-2} + \cdots + a_{n-1}(-\boldsymbol{A}) + a_n \boldsymbol{I} = 0$$

A laborious way to find the eigenvalues of a matrix is to solve the n-th order polynomial for the λ_i—far too time consuming. Instead the matrix is transformed into another form whose eigenvalues are easier to find. In the *Givens method* and the *Housholder method* the matrix is transformed into the tridiagonal form; then, in a fixed number of calculations the eigenvalues can be found [15]. The Givens method requires $4\,n^3/3$ operations to transform a real symmetric matrix to tridiagonal form, whereas the Householder method requires half that number [14]. Once the tridiagonal form is found, a Sturm sequence is applied to determine the eigenvalues. These methods are especially useful when only a few eigenvalues of the matrix are desired.

If all the eigenvalues are needed, the $\boldsymbol{Q}\,\boldsymbol{R}$ algorithm is preferred [21].

The eigenvalues of a certain tridiagonal matrix can be found analytically. If $\boldsymbol{A}$ is a tridiagonal matrix with

$$a_{ii} = p, a_{i,i+1} = q, a_{i+1,i} = r, qr > 0$$

then the eigenvalues of $\boldsymbol{A}$ are [22]

$$\lambda_i = p + 2(qr)^{1/2}\cos\frac{i\pi}{n+1} \quad i = 1,2,\ldots,n$$

This result is useful when finite difference methods are applied to the diffusion equation.

2. Approximation and Integration

2.1. Introduction

Two types of problems arise frequently:

1) A *function* is known exactly at a set of points and an interpolating function is desired. The interpolant may be exact at the set of points, or it may be a "best fit" in some sense. Alternatively it may be desired to represent a function in some other way.
2) *Experimental data* must be fit with a mathematical model. The data have experimental error, so some uncertainty exists. The parameters in the model as well as the uncertainty in the determination of those parameters is desired.

These problems are addressed in this chapter. Section 2.2 gives the properties of polynomials defined over the whole domain and Section 2.3 of polynomials defined on segments of the domain. In Section 2.4, quadrature methods are given for evaluating an integral. Least-squares methods for parameter estimation for both linear and nonlinear models are given in Sections 2.5. Fourier transforms to represent discrete data are described in Section 2.7. The chapter closes with extensions to two-dimensional representations.

2.2. Global Polynomial Approximation

A *global* polynomial $P_m(x)$ is defined over the entire region of space

$$P_m(x) = \sum_{j=0}^{m} c_j x^j$$

This polynomial is of degree m (highest power is x^m) and order $m + 1$ ($m + 1$ parameters $\{c_j\}$). If a set of $m + 1$ points is given,

$$y_1 = f(x_1), y_2 = f(x_2), \ldots, y_{m+1} = f(x_{m+1})$$

then Lagrange's formula yields a polynomial of degree m that goes through the $m + 1$ points:

$$P_m(x) = \frac{(x-x_2)(x-x_3)\cdots(x-x_{m+1})}{(x_1-x_2)(x_1-x_3)\cdots(x_1-x_{m+1})} y_1 + \frac{(x-x_1)(x-x_3)\cdots(x-x_{m+1})}{(x_2-x_1)(x_2-x_3)\cdots(x_2-x_{m+1})} y_2 + \cdots + \frac{(x-x_1)(x-x_2)\cdots(x-x_m)}{(x_{m+1}-x_1)(x_{m+1}-x_2)\cdots(x_{m+1}-x_m)} y_{m+1}$$

Note that each coefficient of y_j is a polynomial of degree m that vanishes at the points $\{x_j\}$ (except for one value of j) and takes the value of 1.0 at that point, i.e.,

$$P_m(x_j) = y_j \quad j = 1, 2, \ldots, m+1$$

If the function $f(x)$ is known, the error in the approximation is [23]

$$|\text{error}(x)| \le \frac{|x_{m+1} - x_1|^{m+1}}{(m+1)!} \max_{x_1 \le x \le x_{m+1}} \left| f^{(m+1)}(x) \right|$$

The evaluation of $P_m(x)$ at a point other than the defining points can be made with Neville's algorithm [15]. Let P_1 be the value at x of the unique function passing through the point (x_1, y_1); i.e., $P_1 = y_1$. Let P_{12} be the value at x of the unique polynomial passing through the points x_1 and x_2. Likewise, $P_{ijk\ldots r}$ is the unique polynomial passing through the points $x_i, x_j, x_k, \ldots, x_r$. The following scheme is used:

$$\begin{array}{llllll} x_1 & y_1 = P_1 & & & \\ & & P_{12} & & \\ x_2 & y_2 = P_2 & & P_{123} & \\ & & P_{23} & & P_{1234} \\ x_3 & y_3 = P_3 & & P_{234} & \\ & & P_{34} & & \\ x_4 & y_4 = P_4 & & & \end{array}$$

These entries are defined by using

$$P_{i(i+1)\cdots(i+m)} = \frac{(x-x_{i+m})P_{i(i+1)\cdots(i+m-1)} + (x_i - x)P_{(i+1)(i+2)\cdots(i+m)}}{x_i - x_{i+m}}$$

Consider P_{1234}: the terms on the right-hand side of the equation involve P_{123} and P_{234}. The "parents," P_{123} and P_{234}, already agree at points 2 and 3. Here $i = 1$, $m = 3$; thus, the parents agree at $x_{i+1}, \ldots, x_{i+m-1}$ already. The formula makes $P_{i(i+1)\ldots(i+m)}$ agree with the function at the additional points x_{i+m} and x_i. Thus, $P_{i(i+1)\ldots(i+m)}$ agrees with the function at all the points $\{x_i, x_{i+1}, \ldots, x_{i+m}\}$.

Orthogonal Polynomials. Another form of the polynomials is obtained by defining them so that they are orthogonal. It is required that $P_m(x)$ be orthogonal to $P_k(x)$ for all $k = 0, \ldots, m-1$.

$$\int_a^b W(x) P_k(x) P_m(x)\,\mathrm{d}x = 0$$

$$k = 0,1,2,\ldots,m-1$$

The orthogonality includes a nonnegative weight function, $W(x) \geq 0$ for all $a \leq x \leq b$. This procedure specifies the set of polynomials to within multiplicative constants, which can be set either by requiring the leading coefficient to be one or by requiring the norm to be one.

$$\int_a^b W(x) P_m^2(x)\,\mathrm{d}x = 1$$

The polynomial $P_m(x)$ has m roots in the closed interval a to b.

The polynomial

$$p(x) = c_0 P_0(x) + c_1 P_1(x) + \cdots c_m P_m(x)$$

minimizes

$$I = \int_a^b W(x)\,[f(x) - p(x)]^2\,\mathrm{d}x$$

when

$$c_j = \frac{\int_a^b W(x)\, f(x) P_j(x)\,\mathrm{d}x}{W_j},$$

$$W_j = \int_a^b W(x) P_j^2(x)\,\mathrm{d}x$$

Note that each c_j is independent of m, the number of terms retained in the series. The minimum value of I is

$$I_{\min} = \int_a^b W(x)\, f^2(x)\,\mathrm{d}x - \sum_{j=0}^{m} W_j c_j^2$$

Such functions are useful for continuous data, i.e., when $f(x)$ is known for all x.

Typical orthogonal polynomials are given in Table 1. Chebyshev polynomials are used in spectral methods (see Chap. 8). The last two rows of Table 1 are widely used in the orthogonal collocation method in chemical engineering.

Table 1. Orthogonal polynomials [15, 23]

a	b	$W(x)$	Name	Recursion relation
-1	1	1	Legendre	$(i+1)P_{i+1} = (2i+1)xP_i - iP_{i-1}$
-1	1	$\frac{1}{\sqrt{1-x^2}}$	Chebyshev	$T_{i+1} = 2xT_i - T_{i-1}$
0	1	$x^{q-1}(1-x)^{p-q}$	Jacobi (p, q)	
$-\infty$	∞	e^{-x^2}	Hermite	$H_{i+1} = 2xH_i - 2iH_{i-1}$
0	∞	$x^c\,\mathrm{e}^{-x}$	Laguerre (c)	$(i+1)L_{i+1}^c = (-x+2i+c+1)L_i^c - (i+c)L_{i-1}^c$
0	1	1	shifted Legendre	
0	1	1	shifted Legendre, function of x^2	

The last entry (the shifted Legendre polynomial as a function of x^2) is defined by

$$\int_0^1 W\left(x^2\right) P_k\left(x^2\right) P_m\left(x^2\right) x^{a-1} \mathrm{d}x = 0$$
$$k = 0,1,\ldots,m-1$$

where $a = 1$ is for planar, $a = 2$ for cylindrical, and $a = 3$ for spherical geometry. These functions are useful if the solution can be proved to be an even function of x.

Rational Polynomials. Rational polynomials are ratios of polynomials. A rational polynomial $R_{i(i+1)\ldots(i+m)}$ passing through $m + 1$ points

$$y_i = f\left(x_i\right),\ i = 1,\ldots,m+1$$

is

$$R_{i(i+1)\cdots(i+m)} = \frac{P_\mu(x)}{Q_\nu(x)} = \frac{p_0+p_1x+\cdots+p_\mu x^\mu}{q_0+q_1x+\cdots+q_\nu x^\nu},$$
$$m+1 = \mu+\nu+1$$

An alternative condition is to make the rational polynomial agree with the first $m + 1$ terms in the power series, giving a Padé approximation, i.e.,

$$\frac{\mathrm{d}^k R_{i(i+1)\cdots(i+m)}}{\mathrm{d}x^k} = \frac{\mathrm{d}^k f(x)}{\mathrm{d}x^k} \quad k = 0,\ldots,m$$

The Bulirsch – Stoer recursion algorithm can be used to evaluate the polynomial:

$$R_{i(i+1)\cdots(i+m)} = R_{(i+1)\cdots(i+m)}$$
$$+\frac{R_{(i+1)\cdots(i+m)}-R_{i(i+1)\cdots(i+m-1)}}{\mathrm{Den}}$$
$$\mathrm{Den} = \left(\frac{x-x_i}{x-x_{i+m}}\right)$$
$$\left(1-\frac{R_{(i+1)\cdots(i+m)}R_{i(i+1)\cdots(i+m-1)}}{R_{(i+1)\cdots(i+m)}-R_{(i+1)\cdots(i+m-1)}}\right)-1$$

Rational polynomials are useful for approximating functions with poles and singularities, which occur in Laplace transforms (see Section 4.2).

Fourier series are discussed in Section 4.1. Representation by sums of exponentials is also possible [24].

In summary, for discrete data, Legendre polynomials and rational polynomials are used. For continuous data a variety of orthogonal polynomials and rational polynomials are used. When the number of conditions (discrete data points) exceeds the number of parameters, then see Section 2.5.

2.3. Piecewise Approximation

Piecewise approximations can be developed from difference formulas [3]. Consider a case in which the data points are equally spaced

$$x_{n+1}-x_n = \Delta x$$
$$y_n = y\left(x_n\right)$$

forward differences are defined by

$$\Delta y_n = y_{n+1}-y_n$$
$$\Delta^2 y_n = \Delta y_{n+1}-\Delta y_n = y_{n+2}-2y_{n+1}+y_n$$

Then, a new variable is defined

$$\alpha = \frac{x_\alpha - x_0}{\Delta x}$$

and the finite interpolation formula through the points $y_0, y_1, \ldots, y_n$ is written as follows:

$$y_\alpha = y_0+\alpha\Delta y_0+\frac{\alpha(\alpha-1)}{2!}\Delta^2 y_0+\cdots+ \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}\Delta^n y_0 \tag{1}$$

Keeping only the first two terms gives a straight line through $(x_0, y_0) - (x_1, y_1)$; keeping the first three terms gives a quadratic function of position going through those points plus (x_2, y_2). The value $\alpha = 0$ gives $x = x_0$; $\alpha = 1$ gives $x = x_1$, etc.

Backward differences are defined by

$$\nabla y_n = y_n - y_{n-1}$$
$$\nabla^2 y_n = \nabla y_n - \nabla y_{n-1} = y_n - 2y_{n-1} + y_{n-2}$$

The interpolation polynomial of order n through the points $y_0, y_{-1}, y_{-2}, \ldots$ is

$$y_\alpha = y_0+\alpha\nabla y_0+\frac{\alpha(\alpha+1)}{2!}\nabla^2 y_0+\cdots+ \frac{\alpha(\alpha+1)\cdots(\alpha+n-1)}{n!}\nabla^n y_0$$

The value $\alpha = 0$ gives $x = x_0$; $\alpha = -1$ gives $x = x_{-1}$. Alternatively, the interpolation polynomial of order n through the points $y_1, y_0, y_{-1}, \ldots$ is

$$y_\alpha = y_1+(\alpha-1)\nabla y_1+\frac{\alpha(\alpha-1)}{2!}\nabla^2 y_1+\cdots+ \frac{(\alpha-1)\alpha(\alpha+1)\cdots(\alpha+n-2)}{n!}\nabla^n y_1$$

Now $\alpha = 1$ gives $x = x_1$; $\alpha = 0$ gives $x = x_0$.

The *finite element method* can be used for piecewise approximations [3]. In the finite element method the domain $a \le x \le b$ is divided

into elements as shown in Figure 3. Each function $N_i(x)$ is zero at all nodes except x_i; $N_i(x_i) = 1$. Thus, the approximation is

$$y(x) = \sum_{i=1}^{NT} c_i N_i(x) = \sum_{i=1}^{NT} y(x_i) N_i(x)$$

where $c_i = y(x_i)$. For convenience, the trial functions are defined within an element by using new coordinates:

$$u = \frac{x - x_i}{\Delta x_i}$$

The Δx_i need not be the same from element to element. The trial functions are defined as $N_i(x)$ (Fig. 3 A) in the global coordinate system and $N_I(u)$ (Fig. 3 B) in the local coordinate system (which also requires specification of the element). For $x_i < x < x_{i+1}$

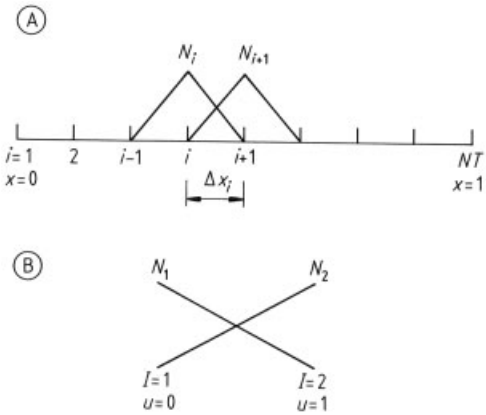

Figure 3. Galerkin finite element method – linear functions
A) Global numbering system; B) Local numbering system

$$y(x) = \sum_{i=1}^{NT} c_i N_i(x) = c_i N_i(x) + c_{i+1} N_{i+1}(x)$$

because all the other trial functions are zero there. Thus

$$y(x) = c_i N_{I=1}(u) + c_{i+1} N_{I=2}(u),$$

$$x_i < x < x_{i+1},\ 0 < u < 1$$

Then

$$N_{I=1} = 1 - u,\ N_{I=2} = u$$

and the expansion is rewritten as

$$y(x) = \sum_{I=1}^{2} c_I^e N_I(u) \tag{2}$$

x in e-th element and $c_i = c_I^e$ within the element e. Thus, given a set of points (x_i, y_i), a finite element approximation can be made to go through them.

Quadratic approximations can also be used within the element (see Fig. 4). Now the trial functions are

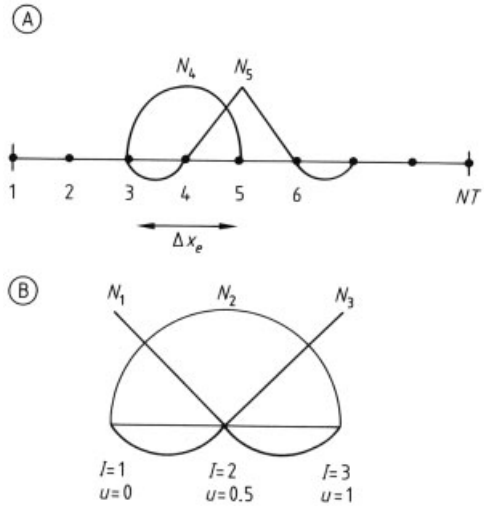

Figure 4. Finite elements approximation – quadratic elements
A) Global numbering system; B) Local numbering system

$$\begin{aligned} N_{I=1} &= 2(u-1)\left(u-\tfrac{1}{2}\right) \\ N_{I=2} &= 4u(1-u) \\ N_{I=3} &= 2u\left(u-\tfrac{1}{2}\right) \end{aligned} \tag{3}$$

The approximation going through an odd number of points (x_i, y_i) is then

$$y(x) = \sum_{I=1}^{3} c_I^e N_I(u) \quad x \text{ in } e\text{-th element}$$

with $c_I^e = y(x_i)$, $i = (e-1)2 + I$
in the e-th element

Hermite cubic polynomials can also be used; these are continuous and have continuous first derivatives at the element boundaries [3].

Splines. Splines are functions that match given values at the points $x_1, \ldots, x_{NT}$, shown in Figure 5, and have continuous derivatives up to some order at the knots, or the points $x_2, \ldots, x_{NT-1}$. *Cubic splines* are most common. In this case the function is represented by a cubic polynomial within each interval and has continuous first and second derivatives at the knots.

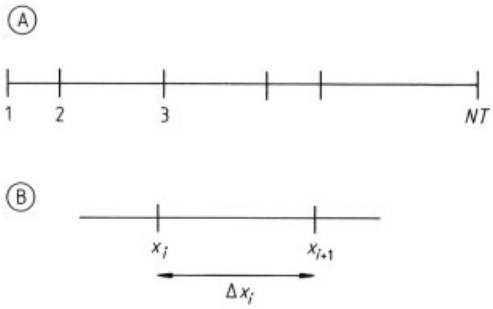

Figure 5. Finite elements for cubic splines
A) Notation for spline knots. B) Notation for one element

Consider the points shown in Figure 5 A. The notation for each interval is shown in Figure 5 B. Within each interval the function is represented as a cubic polynomial.

$$C_i(x) = a_{0i} + a_{1i}x + a_{2i}x^2 + a_{3i}x^3$$

The interpolating function takes on specified values at the knots.

$$C_{i-1}(x_i) = C_i(x_i) = f(x_i)$$

Given the set of values $\{x_i, f(x_i)\}$, the objective is to pass a smooth curve through those points, and the curve should have continuous first and second derivatives at the knots.

$$C'_{i-1}(x_i) = C'_i(x_i)$$

$$C''_{i-1}(x_i) = C''_i(x_i)$$

The formulas for the cubic spline are derived as follows for one region. Since the function is a cubic function the third derivative is constant and the second derivative is linear in x. This is written as

$$C''_i(x) = C''_i(x_i) + \left[C''_i(x_{i+1}) - C''_i(x_i)\right]\frac{x - x_i}{\Delta x_i}$$

and integrated once to give

$$C'_i(x) = C'_i(x_i) + C''_i(x_i)(x - x_i) + \left[C''_i(x_{i+1}) - C''_i(x_i)\right]\frac{(x - x_i)^2}{2\Delta x_i}$$

and once more to give

$$C_i(x) = C_i(x_i) + C'_i(x_i)(x - x_i) + C''_i(x_i)\frac{(x - x_i)^2}{2} + \left[C''_i(x_{i+1}) - C''_i(x_i)\right]\frac{(x - x_i)^3}{6\Delta x_i}$$

Now

$$y_i = C_i(x_i), y'_i = C'_i(x_i), y''_i = C''_i(x_i)$$

is defined so that

$$C_i(x) = y_i + y'_i(x - x_i) + \tfrac{1}{2}y''_i(x - x_i)^2 + \frac{1}{6\Delta x_i}\left(y''_{i+1} - y''_i\right)(x - x_i)^3$$

A number of algebraic steps make the interpolation easy. These formulas are written for the i-th element as well as the $i - 1$-th element. Then the continuity conditions are applied for the first and second derivatives, and the values y'_i and y'_{i-1} are eliminated [15]. The result is

$$y''_{i-1}\Delta x_{i-1} + y''_i\, 2(\Delta x_{i-1} + \Delta x_i) + y''_{i+1}\Delta x_i = -6\left(\frac{y_i - y_{i-1}}{\Delta x_{i-1}} - \frac{y_{i+1} - y_i}{\Delta x_i}\right)$$

This is a tridiagonal system for the set of $\{y''_i\}$ in terms of the set of $\{y_i\}$. Since the continuity conditions apply only for $i = 2, \ldots, NT - 1$, only $NT - 2$ conditions exist for the NT values of y''_i. Two additional conditions are needed, and these are usually taken as the value of the second derivative at each end of the domain, y''_1, y''_{NT}. If these values are zero, the *natural cubic splines* are obtained; they can also be set to achieve some other purpose, such as making the first derivative match some desired condition at the two ends. With these values taken as zero, in the natural cubic spline, an $NT - 2$ system of tridiagonal equations exists, which is easily solved. Once the second derivatives are known at each of the knots, the first derivatives are given by

$$y'_i = \frac{y_{i+1} - y_i}{\Delta x_i} - y''_i\frac{\Delta x_i}{3} - y''_{i+1}\frac{\Delta x_i}{6}$$

The function itself is then known within each element.

Orthogonal Collocation on Finite Elements. In the method of orthogonal collocation on finite elements the solution is expanded in a polynomial of order $NP = NCOL + 2$ within each element [3]. The choice $NCOL = 1$ corresponds to using quadratic polynomials, whereas $NCOL = 2$ gives cubic polynomials. The notation is shown in Figure 6. Set the function to a known value at the two endpoints

$$y_1 = y(x_1)$$

$$y_{NT} = y(x_{NT})$$

and then at the $NCOL$ interior points to each element

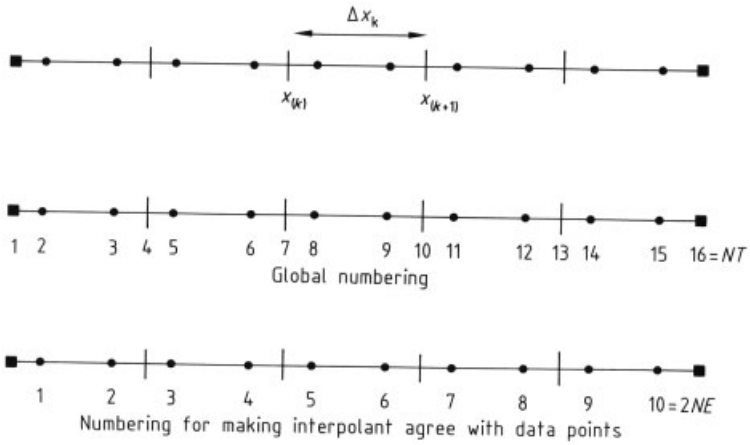

Figure 6. Notation for orthogonal collocation on finite elements
• Residual condition; ■ Boundary conditions; | Element boundary, continuity
NE = total no. of elements.
NT = (*NCOL* + 1) *NE* + 1

$$y_i^e = y_i = y(x_i), i = (NCOL+1)e+I$$

The actual points x_i are taken as the roots of the orthogonal polynomial.

$$P_{NCOL}(u) = 0 \text{ gives } u_1, u_2, \ldots, u_{NCOL}$$

and then

$$x_i = x_{(e)} + \Delta x_e u_I \equiv x_I^e$$

The first derivatives must be continuous at the element boundaries:

$$\frac{\mathrm{d}y}{\mathrm{d}x}\Big|_{x=x_{(2)-}} = \frac{\mathrm{d}y}{\mathrm{d}x}\Big|_{x=x_{(2)+}}$$

Within each element the interpolation is a polynomial of degree *NCOL* + 1. Overall the function is continuous with continuous first derivatives. With the choice *NCOL* = 2, the same approximation is achieved as with Hermite cubic polynomials.

2.4. Quadrature

To calculate the value of an integral, the function can be approximated by using each of the methods described in Section 2.3. Using the first three terms in Equation 1 gives

$$\int_{x_0}^{x_0+h} y(x)\,\mathrm{d}x = \int_0^1 y_\alpha h \mathrm{d}\alpha$$
$$= \frac{h}{2}(y_0+y_1) - \frac{1}{12}h^3 y_0''(\xi), x_0 \le \xi \le x_0+h$$

This corresponds to passing a straight line through the points (x_0, y_0), (x_1, y_1) and integrating under the interpolant. For equally spaced points at $a = x_0$, $a + \Delta x = x_1$, $a + 2\,\Delta x = x_2$, . . . , $a + N\,\Delta x = x_N$, $a + (N + 1)\,\Delta x = b = x_{n+1}$, the trapezoid rule is obtained.

Trapezoid Rule.

$$\int_a^b y(x)\,\mathrm{d}x = \frac{h}{2}(y_0+2y_1+2y_2+\cdots+2y_N$$
$$+y_{N+1}) + O(h^3)$$

The first five terms in Equation 1 are retained and integrated over two intervals.

$$\int_{x_0}^{x_0+2h} y(x)\,\mathrm{d}x = \int_0^2 y_\alpha h \mathrm{d}\alpha = \frac{h}{3}(y_0+4y_1+y_2)$$
$$-\frac{h^5}{90} y_0^{(\mathrm{IV})}(\xi), x_0 \le \xi \le x_0+2h$$

This corresponds to passing a quadratic function through three points and integrating. For an even number of intervals and an odd number of points, 2 *N* + 1, with $a = x_0$, $a + \Delta x = x_1$, $a + 2\,\Delta x = x_2$, . . . , $a + 2\,N\,\Delta x = b$, Simpson's rule is obtained.

Simpson's Rule.

$$\int_a^b y(x)\,\mathrm{d}x = \frac{h}{3}(y_0+4y_1+2y_2+4y_3+2y_4$$
$$+\cdots+2y_{2N-1}+4y_{2N}+y_{2N+1}) + O(h^5)$$

Within each pair of intervals the interpolant is continuous with continuous derivatives, but only the function is continuous from one pair to another.

If the *finite element representation* is used (Eq. 2), the integral is

$$\int_{x_i}^{x_{i+1}} y(x)\,\mathrm{d}x = \int_0^1 \sum_{I=1}^{2} c_I^e N_I(u)(x_{i+1}-x_i)\,\mathrm{d}u$$
$$= \Delta x_i \sum_{I=1}^{2} c_I^e \int_0^1 N_I(u)\,\mathrm{d}u = \Delta x_i \left(c_1^e \tfrac{1}{2} + c_2^e \tfrac{1}{2}\right)$$
$$= \tfrac{\Delta x_i}{2}(y_i + y_{i+1})$$

Since $c_1^e = y_i$ and $c_2^e = y_{i+1}$, the result is the same as the trapezoid rule. These formulas can be added together to give *linear elements*:

$$\int_a^b y(x)\,\mathrm{d}x = \sum\nolimits_e \frac{\Delta x_e}{2}(y_1^e + y_2^e)$$

If the *quadratic* expansion is used (Eq. 3), the endpoints of the element are x_i and x_{i+2}, and x_{i+1} is the midpoint, here assumed to be equally spaced between the ends of the element:

$$\int_{x_i}^{x_{i+2}} y(x)\,\mathrm{d}x = \int_0^1 \sum_{I=1}^{3} c_I^e N_I(u)(x_{i+2}-x_i)\,\mathrm{d}u$$
$$= \Delta x_i \sum_{I=1}^{3} c_I^e \int_0^1 N_I(u)\,\mathrm{d}u$$
$$= \Delta x_e \left(c_1^e \tfrac{1}{6} + c_2^e \tfrac{2}{3} + c_3^e \tfrac{1}{6}\right)$$

For many elements, with different Δx^e, *quadratic elements:*

$$\int_a^b y(x) = \sum\nolimits_e \frac{\Delta x_e}{6}(y_1^e + 4y_2^e + y_3^e)$$

If the element sizes are all the same this gives Simpson's rule.

For cubic splines the quadrature rule within one element is

$$\int_{x_i}^{x_{i+1}} C_i(x)\,\mathrm{d}x = \tfrac{1}{2}\Delta x_i (y_i + y_{i+1}) - \tfrac{1}{24}\Delta x_i^3 \left(y_i'' + y_{i+1}''\right)$$

For the entire interval the quadrature formula is

$$\int_{x_1}^{x_{NT}} y(x)\,\mathrm{d}x = \tfrac{1}{2}\sum_{i=1}^{NT-1} \Delta x_i (y_i + y_{i+1}) - \tfrac{1}{24}\sum_{i=1}^{NT-1} \Delta x_i^3 (y_i'' + y_{i+1}'')$$

with $y_1'' = 0$, $y''_{NT} = 0$ for natural cubic splines.

When orthogonal polynomials are used, as in Equation 1, the m roots to $P_m(x) = 0$ are chosen as quadrature points and called points $\{x_j\}$. Then the quadrature is *Gaussian:*

$$\int_0^1 y(x)\,\mathrm{d}x = \sum_{j=1}^{m} W_j y(x_j)$$

The quadrature is exact when y is a polynomial of degree $2m - 1$ in x. The m weights and m Gauss points result in $2m$ parameters, chosen to exactly represent a polynomial of degree $2m - 1$, which has $2m$ parameters. The Gauss points and weights are given in Table 2. The weights can be defined with $W(x)$ in the integrand as well.

Table 2. Gaussian quadrature points and weights *

N	x_i	W_i
1	0.5000000000	0.6666666667
2	0.2113248654	0.5000000000
	0.7886751346	0.5000000000
3	0.1127016654	0.2777777778
	0.5000000000	0.4444444445
	0.8872983346	0.2777777778
4	0.0694318442	0.1739274226
	0.3300094783	0.3260725774
	0.6699905218	0.3260725774
	0.9305681558	0.1739274226
5	0.0469100771	0.1184634425
	0.2307653450	0.2393143353
	0.5000000000	0.2844444444
	0.7692346551	0.2393143353
	0.9530899230	0.1184634425

* For a given N the quadrature points $x_2, x_3, \ldots, x_{NP-1}$ are given above. $x_1 = 0$, $x_{NP} = 1$. For $N = 1$, $W_1 = W_3 = 1/6$ and for $N \geq 2$, $W_1 = W_{NP} = 0$.

For orthogonal collocation on finite elements the quadrature formula is

$$\int_0^1 y(x)\,\mathrm{d}x = \sum\nolimits_e \Delta x_e \sum_{j=1}^{NP} W_j y(x_j^e)$$

Each special polynomial has its own quadrature formula. For example, Gauss – Legendre polynomials give the quadrature formula

$$\int_0^\infty e^{-x} y(x)\,\mathrm{d}x = \sum_{i=1}^{n} W_i y(x_i)$$

(points and weights are available in mathematical tables) [23].

For Gauss – Hermite polynomials the quadrature formula is

$$\int_{-\infty}^{\infty} e^{-x^2} y(x)\,\mathrm{d}x = \sum_{i=1}^{n} W_i y(x_i)$$

(points and weights are available in mathematical tables) [23].

Romberg's method uses extrapolation techniques to improve the answer [15]. If I_1 is the value of the integral obtained by using interval size $h = \Delta x$, I_2 the value of I obtained by using interval size $h/2$, and I_0 the true value of I, then the error in a method is approximately h^m, or

$$I_1 \approx I_0 + ch^m$$

$$I_2 \approx I_0 + c\left(\frac{h}{2}\right)^m$$

Replacing the $\approx$ by an equality (an approximation) and solving for c and I_0 give

$$I_0 = \frac{2^m I_2 - I_1}{2^m - 1}$$

This process can also be used to obtain $I_1, I_2, \ldots$, by halving h each time, calculating new estimates from each pair, and calling them $J_1, J_2, \ldots$ (i.e., in the formula above, I_0 is replaced with J_1). The formulas are reapplied for each pair of J's to obtain $K_1, K_2, \ldots$. The process continues until the required tolerance is obtained.

I_1	I_2	I_3	I_4
	J_1	J_2	J_3
		K_1	K_2
			L_1

Romberg's method is most useful for a low-order method (small m) because significant improvement is then possible.

When the integrand has singularities, a variety of techniques can be tried. The integral may be divided into one part that can be integrated analytically near the singularity and another part that is integrated numerically. Sometimes a change of argument allows analytical integration. Series expansion might be helpful, too. When the domain is infinite, Gauss – Legendre or Gauss – Hermite quadrature can be used. Also a transformation can be made [15]. For example, let $u = 1/x$ and then

$$\int_a^b f(x)\,\mathrm{d}x = \int_{1/b}^{1/a} \frac{1}{u^2} f\left(\frac{1}{u}\right)\mathrm{d}u \quad a, b > 0$$

2.5. Least Squares

When fitting experimental data to a mathematical model, it is necessary to recognize that the experimental measurements contain error; the goal is to find the set of parameters in the model that best represents the experimental data. Reference [23] gives a complete treatment relating the least-squares procedure to maximum likelihood.

In a least-squares parameter estimation, it is desired to find parameters that minimize the sum of squares of the deviation between the experimental data and the theoretical equation

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i; a_1, a_2, \ldots, a_M)}{\sigma_i}\right]^2$$

where y_i is the i-th experimental data point for the value x_i, $y(x_i; a_1, a_2, \ldots, a_M)$ the theoretical equation at x_i, σ_i the standard deviation of the i-th measurement, and the parameters $\{a_1, a_2, \ldots, a_M\}$ are to be determined to minimize χ^2. The simplification is made here that the standard deviations are all the same. Thus, we minimize the variance of the curve fit.

$$\sigma^2 = \sum_{i=1}^{N} \frac{[y_i - y(x_i; a_1, a_2, \ldots, a_M)]^2}{N}$$

Linear Least Squares. When the model is a straight line, one is minimizing

$$\chi^2 = \sum_{i=1}^{N} [y_i - a - bx_i]^2$$

The linear correlation coefficient r is defined by

$$r = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}}$$

and

$$\chi^2 = (1 - r^2) \sum_{i=1}^{N} [y_i - \overline{y}]^2$$

where $\overline{y}$ is the average of y_i values. Values of r near 1 indicate a positive correlation; r near -1 means a negative correlation, and r near zero means no correlation. These parameters are easily found by using standard programs, such as Microsoft Excel.

Polynomial Regression. In polynomial regression, one expands the function in a polynomial in x.

$$y(x) = \sum_{j=1}^{M} a_j x^{j-1}$$

The parameters are easily determined using computer software. In Microsoft Excel, the data is put into columns A and B and the graph is created as for a linear curve fit. Then add a trendline and choose the degree of polynomial desired.

Multiple Regression. In multiple regression, any set of functions can be used, not just polynomials.

$$y(x) = \sum_{j=1}^{M} a_j f_j(x)$$

where the set of functions $\{f_j(x)\}$ is known and specified. Note that the unknown parameters $\{a_j\}$ enter the equation linearly. In this case, the spreadsheet can be expanded to have a column for x, and then successive columns for $f_j(x)$. In Microsoft Excel, choose Regression under Tools/Data Analysis, and complete the form. In addition to the actual correlation, one gets the expected variance of the unknowns, which allows one to assess how accurately they were determined.

Nonlinear Regression. In nonlinear regression, the same procedure is followed except that an optimization routine must be used to find the minimum χ^2. See Chapter 10.

2.6. Fourier Transforms of Discrete Data [15]

Suppose a signal $y(t)$ is sampled at equal intervals

$$y_n = y(n\Delta), n = \ldots, -2, -1, 0, 1, 2, \ldots$$

$\Delta =$ sampling rate

(e.g., number of samples per second)

The Fourier transform and inverse transform are

$$Y(\omega) = \int_{-\infty}^{\infty} y(t)\, e^{\mathrm{i}\omega t} \mathrm{d}t$$

$$y(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} Y(\omega)\, e^{-\mathrm{i}\omega t} \mathrm{d}\omega$$

(For definition of i, see Chap. 3.) The Nyquist critical frequency or critical angular frequency is

$$f_c = \frac{1}{2\Delta}, \omega_c = \frac{\pi}{\Delta}$$

If a function $y(t)$ is bandwidth limited to frequencies smaller than f_c, i.e.,

$$Y(\omega) = 0 \text{ for } \omega > \omega_c$$

then the function is completely determined by its samples y_n. Thus, the entire information content of a signal can be recorded by sampling at a rate $\Delta^{-1} = 2f_c$. If the function is not bandwidth limited, then aliasing occurs. Once a sample rate Δ is chosen, information corresponding to frequencies greater than f_c is simply aliased into that range. The way to detect this in a Fourier transform is to see if the transform approaches zero at $\pm f_c$; if not, aliasing has occurred and a higher sampling rate is needed.

Next, for N samples, where N is even

$$y_k = y(t_k),\ t_k = k\Delta,\ k = 0, 1, 2, \ldots, N-1$$

and the sampling rate is Δ; with only N values $\{y_k\}$ the complete Fourier transform $Y(\omega)$ cannot be determined. Calculate the value $Y(\omega_n)$ at the discrete points

$$\omega_n = \frac{2\pi n}{N\Delta}, n = -\frac{N}{2}, \ldots, 0, \ldots, \frac{N}{2}$$

$$Y_n = \sum_{k=0}^{N-1} y_k \mathrm{e}^{2\pi \mathrm{i} kn/N}$$

$$Y(\omega_n) = \Delta Y_n$$

The discrete inverse Fourier transform is

$$y_k = \frac{1}{N} \sum_{k=0}^{N-1} Y_n \mathrm{e}^{-2\pi \mathrm{i} kn/N}$$

The *fast Fourier transform* (*FFT*) is used to calculate the Fourier transform as well as the inverse Fourier transform. A discrete Fourier transform of length N can be written as the sum of two discrete Fourier transforms, each of length $N/2$, and each of these transforms is separated into two halves, each half as long. This continues until only one component is left. For this reason, N is taken as a power of 2, $N = 2^p$.

The vector $\{y_j\}$ is filled with zeroes, if need be, to make $N = 2^p$ for some p. For the computer program, see [15, p. 381]. The standard

Fourier transform takes N^2 operations to calculate, whereas the fast Fourier transform takes only $N \log_2 N$. For large N the difference is significant; at $N = 100$ it is a factor of 15, but for $N = 1000$ it is a factor of 100.

The discrete Fourier transform can also be used for *differentiating* a function; this is used in the spectral method for solving differential equations. Consider a grid of equidistant points:

$$x_n = n\Delta x, n = 0,1,2,\ldots, 2N-1, \Delta x = \frac{L}{2N}$$

the solution is known at each of these grid points $\{Y(x_n)\}$. First, the Fourier transform is taken:

$$y_k = \frac{1}{2N}\sum_{n=0}^{2N-1} Y(x_n)\,\mathrm{e}^{-2\mathrm{i}k\pi x_n/L}$$

The inverse transformation is

$$Y(x) = \frac{1}{L}\sum_{k=-N}^{N} y_k \mathrm{e}^{2\mathrm{i}k\pi x/L}$$

which is differentiated to obtain

$$\frac{\mathrm{d}Y}{\mathrm{d}x} = \frac{1}{L}\sum_{k=-N}^{N} y_k \frac{2\pi\mathrm{i}k}{L}\mathrm{e}^{2\mathrm{i}k\pi x/L}$$

Thus, at the grid points

$$\left.\frac{\mathrm{d}Y}{\mathrm{d}x}\right|_n = \frac{1}{L}\sum_{k=-N}^{N} y_k \frac{2\pi\mathrm{i}k}{L}\mathrm{e}^{2\mathrm{i}k\pi x_n/L}$$

The process works as follows. From the solution at all grid points the Fourier transform is obtained by using FFT $\{y_k\}$. This is multiplied by $2\pi\,\mathrm{i}\,k/L$ to obtain the Fourier transform of the derivative:

$$y'_k = y_k \frac{2\pi\mathrm{i}k}{L}$$

The inverse Fourier transform is then taken by using FFT, to give the value of the derivative at each of the grid points:

$$\left.\frac{\mathrm{d}Y}{\mathrm{d}x}\right|_n = \frac{1}{L}\sum_{k=-N}^{N} y'_k \mathrm{e}^{2\mathrm{i}k\pi x_n/L}$$

Any nonlinear term can be treated in the same way: evaluate it in real space at N points and take the Fourier transform. After processing using this transform to get the transform of a new function, take the inverse transform to obtain the new function at N points. This is what is done in direct numerical simulation of turbulence (DNS).

2.7. Two-Dimensional Interpolation and Quadrature

Bicubic splines can be used to interpolate a set of values on a regular array, $f(x_i, y_j)$. Suppose NX points occur in the x direction and NY points occur in the y direction. Press et al. [15] suggest computing NY different cubic splines of size NX along lines of constant y, for example, and storing the derivative information. To obtain the value of f at some point x, y, evaluate each of these splines for that x. Then do one spline of size NY in the y direction, doing both the determination and the evaluation.

Multidimensional integrals can also be broken down into one-dimensional integrals. For example,

$$\int_a^b \int_{f_1(x)}^{f_2(x)} z(x,y)\,\mathrm{d}x\mathrm{d}y = \int_a^b G(x)\,\mathrm{d}x;$$

$$G(x) = \int_{f_1(x)}^{f_2(x)} z(x,y)\,\mathrm{d}x$$

3. Complex Variables [25 – 31]

3.1. Introduction to the Complex Plane

A complex number is an ordered pair of real numbers, x and y, that is written as

$$z = x + \mathrm{i}y$$

The variable i is the imaginary unit which has the property

$$\mathrm{i}^2 = -1$$

The real and imaginary parts of a complex number are often referred to:

$$\mathrm{Re}\,(z) = x, \mathrm{Im}\,(z) = y$$

A complex number can also be represented graphically in the complex plane, where the real part of the complex number is the abscissa and the imaginary part of the complex number is the ordinate (see Fig. 7).

Another representation of a complex number is the *polar form*, where r is the magnitude and θ is the argument.

$$r = |x + \mathrm{i}\,y| = \sqrt{x^2 + y^2},\ \theta = \arg(x + \mathrm{i}\,y)$$

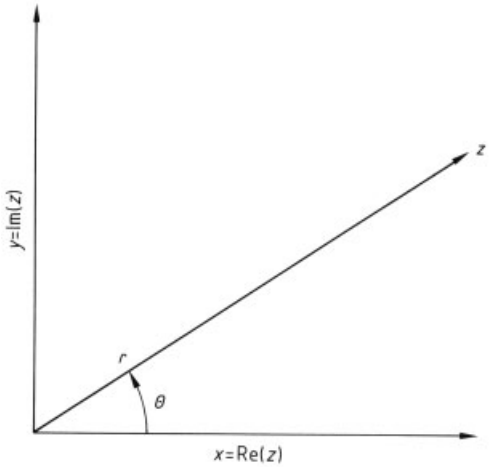

Figure 7. The complex plane

Write

$$z = x + \mathrm{i}\, y = r\,(\cos\theta + \mathrm{i}\sin\theta)$$

so that

$$x = r\cos\theta, y = r\sin\theta$$

and

$$\theta = \arctan\frac{y}{x}$$

Since the arctangent repeats itself in multiples of π rather than $2\,\pi$, the argument must be defined carefully. For example, the θ given above could also be the argument of $-\,(x + \mathrm{i}\, y)$. The function $z = \cos\theta + \mathrm{i}\sin\theta$ obeys $|z| = |\cos\theta + \mathrm{i}\sin\theta| = 1$.

The rules of equality, addition, and multiplication are

$$z_1 = x_1 + \mathrm{i}\, y_1, z_2 = x_2 + \mathrm{i}\, y_2$$

Equality :
$z_1 = z_2$ if and only if $x_1 = x_2$ and $y_1 = y_2$
Addition :
$z_1 + z_2 = (x_1 + x_2) + \mathrm{i}\,(y_1 + y_2)$
Multiplication :
$z_1 z_2 = (x_1 x_2 - y_1 y_2) + \mathrm{i}\,(x_1 y_2 + x_2 y_1)$

The last rule can be remembered by using the standard rules for multiplication, keeping the imaginary parts separate, and using $\mathrm{i}^2 = -\,1$. In the complex plane, addition is illustrated in Figure 8. In polar form, multiplication is

$$z_1 z_2 = r_1 r_2\,[\cos(\theta_1 + \theta_2) + \mathrm{i}\sin(\theta_1 + \theta_2)]$$

The magnitude of $z_1 + z_2$ is bounded by

$$|z_1 \pm z_2| \le |z_1| + |z_2| \text{ and } |z_1| - |z_2| \le |z_1 \pm z_2|$$

as can be seen in Figure 8. The magnitude and arguments in multiplication obey

$$|z_1 z_2| = |z_1|\,|z_2|\,, \arg(z_1 z_2) = \arg z_1 + \arg z_2$$

The *complex conjugate* is $z^* = x - \mathrm{i}\, y$ when $z = x + \mathrm{i}\, y$ and $|z^*| = |z|$, $\arg z^* = -\arg z$

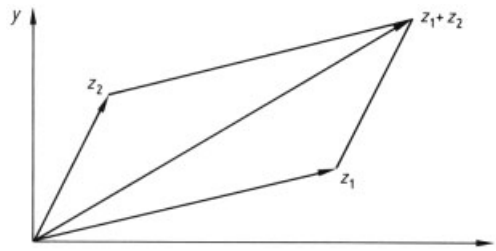

Figure 8. Addition in the complex plane

For complex conjugates then

$$z^*\, z = |z|^2$$

The reciprocal is

$$\frac{1}{z} = \frac{z^*}{|z|^2} = \frac{1}{r}\,(\cos\theta - \mathrm{i}\sin\theta)\,, \arg\left(\frac{1}{z}\right) = -\arg z$$

Then

$$\frac{z_1}{z_2} = \frac{x_1 + \mathrm{i}\, y_1}{x_2 + \mathrm{i}\, y_2} = (x_1 + \mathrm{i}\, y_1)\,\frac{x_2 - \mathrm{i}\, y_2}{x_2^2 + y_2^2}$$
$$= \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + \mathrm{i}\,\frac{x_2 y_1 - x_1 y_2}{x_2^2 + y_2^2}$$

and

$$\frac{z_1}{z_2} = \frac{r_1}{r_2}\,[\cos(\theta_1 - \theta_2) + \mathrm{i}\sin(\theta_1 - \theta_2)]$$

3.2. Elementary Functions

Properties of elementary functions of complex variables are discussed here [32]. When the polar form is used, the argument must be specified because the same physical angle can be achieved with arguments differing by $2\,\pi$. A complex number taken to a real power obeys

$$u = z^n,\ |z^n| = |z|^n, \arg\,(z^n) = n\arg z\,(\mathrm{mod}\,2\pi)$$

$$u = z^n = r^n\,(\cos n\theta + \mathrm{i}\sin n\theta)$$

Roots of a complex number are complicated by careful accounting of the argument

$z = w^{1/n}$ with $w = R(\cos\Theta + \mathrm{i}\sin\Theta)$, $0 \le \Theta \le 2\pi$

then

$$z_k = R^{1/n}\{\cos\left[\frac{\Theta}{n} + (k-1)\frac{2\pi}{n}\right]$$

$$+\mathrm{i}\sin\left[\frac{\Theta}{n} + (k-1)\frac{2\pi}{n}\right]\}$$

such that

$(z_k)^n = w$ for every k

$$z = r(\cos\theta + \mathrm{i}\sin\theta)$$
$$r^n = R, n\theta = \Theta \pmod{2\pi}$$

The exponential function is

$$e^z = \mathrm{e}^x(\cos y + \mathrm{i}\sin y)$$

Thus,

$$z = r(\cos\theta + \mathrm{i}\sin\theta)$$

can be written

$$z = r\mathrm{e}^{\mathrm{i}\theta}$$

and

$$|\mathrm{e}^z| = \mathrm{e}^x, \arg \mathrm{e}^z = y \pmod{2\pi}$$

The exponential obeys

$\mathrm{e}^z \neq 0$ for every finite z

and is periodic with period 2π:

$$e^{z+2\pi\,\mathrm{i}} = \mathrm{e}^z$$

Trigonometric functions can be defined by using

$e^{iy} = \cos y + \mathrm{i}\sin y$, and $\mathrm{e}^{-\mathrm{i}y} = \cos y - \mathrm{i}\sin y$

Thus,

$$\cos y = \frac{\mathrm{e}^{\mathrm{i}\,y} + \mathrm{e}^{-\mathrm{i}\,y}}{2} = \cosh \mathrm{i}\,y$$

$$\sin y = \frac{\mathrm{e}^{\mathrm{i}\,y} - \mathrm{e}^{-\mathrm{i}\,y}}{2\mathrm{i}} = -\mathrm{i}\sinh \mathrm{i}\,y$$

The second equation follows from the definitions

$$\cosh z \equiv \frac{\mathrm{e}^z + \mathrm{e}^{-z}}{2},\ \sinh z \equiv \frac{\mathrm{e}^z - \mathrm{e}^{-z}}{2}$$

The remaining hyperbolic functions are

$$\tanh z \equiv \frac{\sinh z}{\cosh z}, \coth z \equiv \frac{1}{\tanh z}$$

$$\mathrm{sech}\, z \equiv \frac{1}{\cosh z}, \mathrm{csch}\, z \equiv \frac{1}{\sinh z}$$

The circular functions with complex arguments are defined

$$\cos z = \frac{\mathrm{e}^{\mathrm{i}\,z} + \mathrm{e}^{-\mathrm{i}\,z}}{2}, \sin z = \frac{\mathrm{e}^{\mathrm{i}\,z} - \mathrm{e}^{-\mathrm{i}\,z}}{2},$$

$$\tan z = \frac{\sin z}{\cos z}$$

and satisfy

$$\sin(-z) = -\sin z, \cos(-z) = \cos z$$

$$\sin(\mathrm{i}\,z) = \mathrm{i}\sinh z, \cos(\mathrm{i}\,z) = -\cosh z$$

All trigonometric identities for real, circular functions with real arguments can be extended without change to complex functions of complex arguments. For example,

$$\sin^2 z + \cos^2 z = 1,$$

$$\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \cos z_1 \sin z_2$$

The same is true of hyperbolic functions. The absolute boundaries of $\sin z$ and $\cos z$ are not bounded for all z.

Trigonometric identities can be defined by using

$$e^{i\theta} = \cos\theta + \mathrm{i}\sin\theta$$

For example,

$$\mathrm{e}^{\mathrm{i}(\alpha+\beta)} = \cos(\alpha+\beta) + \mathrm{i}\sin(\alpha+\beta)$$

$$= \mathrm{e}^{\mathrm{i}\alpha}\mathrm{e}^{\mathrm{i}\beta} = (\cos\alpha + \mathrm{i}\sin\alpha)$$
$$(\cos\beta + \mathrm{i}\sin\beta)$$
$$= \cos\alpha\cos\beta - \sin\alpha\sin\beta$$
$$+\mathrm{i}\ (\cos\alpha\sin\beta + \cos\beta\sin\alpha)$$

Equating real and imaginary parts gives

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

$$\sin(\alpha+\beta) = \cos\alpha\sin\beta + \cos\beta\sin\alpha$$

The logarithm is defined as

$$\ln z = \ln|z| + \mathrm{i}\arg z$$

and the various determinations differ by multiples of 2π i. Then,

$$\mathrm{e}^{\ln z} = z$$

$$\ln(\mathrm{e}^z) - z \equiv 0 \pmod{2\pi\mathrm{i}}$$

Also,

$$\ln(z_1 z_2) - \ln z_1 - \ln z_2 \equiv 0 \pmod{2\pi\mathrm{i}}$$

is always true, but

$$\ln(z_1 z_2) = \ln z_1 + \ln z_2$$

holds only for some determinations of the logarithms. The principal determination of the argument can be defined as $-\pi < \arg \le \pi$.

3.3. Analytic Functions of a Complex Variable

Let $f(z)$ be a single-valued continuous function of z in a domain D. The function $f(z)$ is differentiable at the point z_0 in D if

$$\lim_{h\to 0}\frac{f(z_0+h)-f(z_0)}{h}$$

exists as a finite (complex) number and is independent of the direction in which h tends to zero. The limit is called the derivative, $f'(z_0)$. The derivative now can be calculated with h approaching zero in the complex plane, i.e., anywhere in a circular region about z_0. The function $f(z)$ is differentiable in D if it is differentiable at all points of D; then $f(z)$ is said to be an analytic function of z in D. Also, $f(z)$ is analytic at z_0 if it is analytic in some ε neighborhood of z_0. The word analytic is sometimes replaced by holomorphic or regular.

The *Cauchy – Riemann equations* can be used to decide if a function is analytic. Set

$$f(z) = f(x+\mathrm{i}\,y) = u(x,y) + \mathrm{i}\,v(x,y)$$

Theorem [30, p. 51]. Suppose that $f(z)$ is defined and continuous in some neighborhood of $z = z_0$. A necessary condition for the existence of $f'(z_0)$ is that $u(x, y)$ and $v(x, y)$ have first-order partials and that the Cauchy – Riemann conditions (see below) hold.

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \text{ and } \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \text{ at } z_0$$

Theorem [30, p. 61]. The function $f(z)$ is analytic in a domain D if and only if u and v are continuously differentiable and satisfy the Cauchy – Riemann conditions there.

If $f_1(z)$ and $f_2(z)$ are analytic in domain D, then $\alpha_1 f_1(z) + \alpha_2 f_2(z)$ is analytic in D for any (complex) constants α_1, α_2.

$$f_1(z) + f_2(z) \text{ is analytic in D}$$

$$f_1(z)/f_2(z) \text{ is analytic in D except where } f_2(z) = 0$$

An analytic function of an analytic function is analytic. If $f(z)$ is analytic, $f'(z) \neq 0$ in D, $f(z_1) \neq f(z_2)$ for $z_1 \neq z_2$, then the inverse function $g(w)$ is also analytic and

$$g'(w) = \frac{1}{f'(z)} \text{ where } w = f(z),$$

$$g(w) = g[f(z)] = z$$

Analyticity implies continuity but the converse is not true: $z^* = x - \mathrm{i}\,y$ is continuous but, because the Cauchy – Riemann conditions are not satisfied, it is not analytic. An entire function is one that is analytic for all finite values of z. Every polynomial is an entire function. Because the polynomials are analytic, a ratio of polynomials is analytic except when the denominator vanishes. The function $f(z) = |z^2|$ is continuous for all z but satisfies the Cauchy – Riemann conditions only at $z = 0$. Hence, $f'(z)$ exists only at the origin, and $|z|^2$ is nowhere analytic. The function $f(z) = 1/z$ is analytic except at $z = 0$. Its derivative is $-1/z^2$, where $z \neq 0$. If $\ln z = \ln|z| + \mathrm{i}\arg z$ in the cut domain $-\pi < \arg z \leq \pi$, then $f(z) = 1/\ln z$ is analytic in the same cut domain, except at $z = 1$, where $\log z = 0$. Because e^z is analytic and $\pm\,\mathrm{i}\,z$ are analytic, $\mathrm{e}^{\pm \mathrm{i} z}$ is analytic and linear combinations are analytic. Thus, the sine and cosine and hyperbolic sine and cosine are analytic. The other functions are analytic except when the denominator vanishes.

The derivatives of the elementary functions are

$$\frac{\mathrm{d}}{\mathrm{d}z}\mathrm{e}^z = \mathrm{e}^z,\ \frac{\mathrm{d}}{\mathrm{d}z}z^n = n\,z^{n-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}(\ln z) = \frac{1}{z},\ \frac{\mathrm{d}}{\mathrm{d}z}\sin z = \cos z,$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\cos z = -\sin z$$

In addition,

$$\frac{\mathrm{d}}{\mathrm{d}z}(f\,g) = f\frac{\mathrm{d}g}{\mathrm{d}z} + g\frac{\mathrm{d}f}{\mathrm{d}z}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}f[g(z)] = \frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}z}$$

$$\frac{\mathrm{d}}{\mathrm{d}z}\sin w = \cos w\frac{\mathrm{d}w}{\mathrm{d}z},\ \frac{\mathrm{d}}{\mathrm{d}z}\cos w = -\sin w\frac{\mathrm{d}w}{\mathrm{d}z}$$

Define $z^a = \mathrm{e}^{a\ln z}$ for complex constant a. If the determination is $-\pi < \arg z \leq \pi$, then z^a is analytic on the complex plane with a cut on the negative real axis. If a is an integer n, then $\mathrm{e}^{2\pi i n} = 1$ and z^n has the same limits approaching the cut from either side. The function can be made continuous across the cut and the function is analytic there, too. If $a = 1/n$ where n is an integer, then

$$z^{1/n} = \mathrm{e}^{(\ln z)/n} = |z|^{1/n}\mathrm{e}^{\mathrm{i}(\arg z)/n}$$

So $w = z^{1/n}$ has n values, depending on the choice of argument.

Laplace Equation. If $f(z)$ is analytic, where

$$f(z) = u(x,y) + \mathrm{i}\, v(x,y)$$

the Cauchy – Riemann equations are satisfied. Differentiating the Cauchy – Riemann equations gives the Laplace equation:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 u}{\partial y^2} \text{ or}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Similarly,

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

Thus, general solutions to the Laplace equation can be obtained from analytic functions [30, p. 60]. For example,

$$\ln \frac{1}{|z - z_0|}$$

is analytic so that a solution to the Laplace equation is

$$\ln \left[(x-a)^2 + (y-b)^2\right]^{-1/2}$$

A solution to the Laplace equation is called a harmonic function. A function is harmonic if, and only if, it is the real part of an analytic function. The imaginary part is also harmonic. Given any harmonic function u, a conjugate harmonic function v can be constructed such that $f = u + \mathrm{i}\, v$ is locally analytic [30, p. 290].

Maximum Principle. If $f(z)$ is analytic in a domain D and continuous in the set consisting of D and its boundary C, and if $|f(z)| \leq M$ on C, then $|f(z)| < M$ in D unless $f(z)$ is a constant [30, p. 134].

3.4. Integration in the Complex Plane

Let C be a rectifiable curve in the complex plane

$$C: z = z(t), 0 \leq t \leq 1$$

where $z(t)$ is a continuous function of bounded variation; C is oriented such that $z_1 = z(t_1)$ precedes the point $z_2 = z(t_2)$ on C if and only if $t_1 < t_2$. Define

$$\int_C f(z)\,\mathrm{d}z = \int_0^1 f[z(t)]\,\mathrm{d}z(t)$$

The integral is linear with respect to the integrand:

$$\int_C [\alpha_1 f_1(z) + \alpha_2 f_2(z)]\,\mathrm{d}z$$
$$= \alpha_1 \int_C f_1(z)\,\mathrm{d}z + \alpha_2 \int_C f_2(z)\,\mathrm{d}z$$

The integral is additive with respect to the path. Let curve C_2 begin where curve C_1 ends and $C_1 + C_2$ be the path of C_1 followed by C_2. Then,

$$\int_{C_1+C_2} f(z)\,\mathrm{d}z = \int_{C_1} f(z)\,\mathrm{d}z + \int_{C_2} f(z)\,\mathrm{d}z$$

Reversing the orientation of the path replaces the integral by its negative:

$$\int_{-C} f(z)\,\mathrm{d}z = -\int_C f(z)\,\mathrm{d}z$$

If the path of integration consists of a finite number of arcs along which $z(t)$ has a continuous derivative, then

$$\int_C f(z)\,\mathrm{d}z = \int_0^1 f[z(t)]\,z'(t)\,\mathrm{d}t$$

Also if $s(t)$ is the arc length on C and $l(C)$ is the length of C

$$\left|\int_C f(z)\,\mathrm{d}z\right| \leq \max_{z \in C} |f(z)|\, l(C)$$

and

$$\left|\int_C f(z)\,\mathrm{d}z\right| \leq \int_C |f(z)|\,|\mathrm{d}z| = \int_0^1 |f[z(t)]|\,\mathrm{d}s(t)$$

Cauchy's Theorem [25, 30, p. 111]. Suppose $f(z)$ is an analytic function in a domain D and C is a simple, closed, rectifiable curve in D such that $f(z)$ is analytic inside and on C. Then

$$\oint_C f(z)\,\mathrm{d}z = 0 \tag{4}$$

If D is simply connected, then Equation 4 holds for every simple, closed, rectifiable curve C in D. If D is simply connected and if a and b are any two points in D, then

$$\int_a^b f(z)\,\mathrm{d}z$$

is independent of the rectifiable path joining a and b in D.

Cauchy's Integral. If C is a closed contour such that $f(z)$ is analytic inside and on C, z_0 is a point inside C, and z traverses C in the counterclockwise direction,

$$f(z_0) = \frac{1}{2\pi \mathrm{i}} \oint_C \frac{f(z)}{z-z_0} \mathrm{d}z$$

$$f'(z_0) = \frac{1}{2\pi \mathrm{i}} \oint_C \frac{f(z)}{(z-z_0)^2} \mathrm{d}z$$

Under further restriction on the domain [30, p. 127],

$$f^{(m)}(z_0) = \frac{m!}{2\pi \mathrm{i}} \oint_C \frac{f(z)}{(z-z_0)^{m+1}} \mathrm{d}z$$

Power Series. If $f(z)$ is analytic interior to a circle $|z - z_0| < r_0$, then at each point inside the circle the series

$$f(z) = f(z_0) + \sum_{n=1}^{\infty} \frac{f^{(n)}(z_0)}{n!} (z-z_0)^n$$

converges to $f(z)$. This result follows from Cauchy's integral. As an example, e^z is an entire function (analytic everywhere) so that the MacLaurin series

$$\mathrm{e}^z = 1 + \sum_{n=1}^{\infty} \frac{z^n}{n!}$$

represents the function for all z.

Another result of Cauchy's integral formula is that if $f(z)$ is analytic in an annulus R, $r_1 < |z - z_0| < r_2$, it is represented in R by the Laurent series

$$f(z) = \sum_{n=-\infty}^{\infty} A_n (z-z_0)^n, r_1 < |z-z_0| \leq r_2$$

where

$$A_n = \frac{1}{2\pi \mathrm{i}} \int_C \frac{f(z)}{(z-z_0)^{n+1}} \mathrm{d}z,$$

$$n = 0, \pm 1, \pm 2, \ldots,$$

and C is a closed curve counterclockwise in R.

Singular Points and Residues [33, p. 159, 30, p. 180]. If a function in analytic in every neighborhood of z_0, but not at z_0 itself, then z_0 is called an isolated singular point of the function. About an isolated singular point, the function can be represented by a Laurent series.

$$f(z) = \cdots + \frac{A_{-2}}{(z-z_0)^2} + \frac{A_{-1}}{z-z_0} + A_0 + A_1 (z-z_0) + \cdots 0 < |z-z_0| \leq r_0 \tag{5}$$

In particular,

$$A_{-1} = \frac{1}{2\pi \mathrm{i}} \oint_C f(z) \mathrm{d}z$$

where the curve C is a closed, counterclockwise curve containing z_0 and is within the neighborhood where $f(z)$ is analytic. The complex number A_{-1} is the residue of $f(z)$ at the isolated singular point z_0; $2\pi \mathrm{i} A_{-1}$ is the value of the integral in the positive direction around a path containing no other singular points.

If $f(z)$ is defined and analytic in the exterior $|z - z_0| > \mathrm{R}$ of a circle, and if

$$v(\zeta) = f\left(z_0 + \frac{1}{\zeta}\right) \text{ obtained by } \zeta = \frac{1}{z-z_0}$$

has a removable singularity at $\zeta = 0$, $f(z)$ is analytic at infinity. It can then be represented by a Laurent series with nonpositive powers of $z - z_0$.

If C is a closed curve within which and on which $f(z)$ is analytic except for a finite number of singular points $z_1, z_2, \ldots, z_n$ interior to the region bounded by C, then the residue theorem states

$$\oint_C f(z) \mathrm{d}z = 2\pi \mathrm{i} (\varrho_1 + \varrho_2 + \cdots \varrho_n)$$

where ϱ_n denotes the residue of $f(z)$ at z_n.

The series of negative powers in Equation 5 is called the principal part of $f(z)$. If the principal part has an infinite number of nonvanishing terms, the point z_0 is an essential singularity. If $A_{-m} \neq 0$, $A_{-n} = 0$ for all $m < n$, then z_0 is called a pole of order m. It is a simple pole if $m = 1$. In such a case,

$$f(z) = \frac{A_{-1}}{z-z_0} + \sum_{n=0}^{\infty} A_n (z-z_0)^n$$

If a function is not analytic at z_0 but can be made so by assigning a suitable value, then z_0 is a removable singular point.

When $f(z)$ has a pole of order m at z_0,

$$\varphi(z) = (z-z_0)^m f(z), 0 < |z-z_0| < r_0$$

has a removable singularity at z_0. If $\varphi(z_0) = A_{-m}$, then $\varphi(z)$ is analytic at z_0. For a simple pole,

$$A_{-1} = \varphi(z_0) = \lim_{z \to z_0} (z-z_0) f(z)$$

Also $|f(z)| \to \infty$ as $z \to z_0$ when z_0 is a pole. Let the function $p(z)$ and $q(z)$ be analytic at z_0, where $p(z_0) \neq 0$. Then

$$f(z) = \frac{p(z)}{q(z)}$$

has a simple pole at z_0 if, and only if, $q(z_0) = 0$ and $q'(z_0) \neq 0$. The residue of $f(z)$ at the simple pole is

$$A_{-1} = \frac{p(z_0)}{q'(z_0)}$$

If $q^{(i-1)}(z_0) = 0$, i = 1, . . . , m, then z_0 is a pole of $f(z)$ of order m.

Branch [33, p. 163]. A branch of a multiple-valued function $f(z)$ is a single-valued function that is analytic in some region and whose value at each point there coincides with the value of $f(z)$ at the point. A branch cut is a boundary that is needed to define the branch in the greatest possible region. The function $f(z)$ is singular along a branch cut, and the singularity is not isolated. For example,

$$z^{1/2} = f_1(z) = \sqrt{r}\left(\cos\frac{\theta}{2} + \mathrm{i}\sin\frac{\theta}{2}\right)$$

$$-\pi < \theta < \pi,\ r > 0$$

is double valued along the negative real axis. The function tends to $\sqrt{r}\mathrm{i}$ when $\theta \to \pi$ and to $-\sqrt{r}\mathrm{i}$ when $\theta \to -\pi$; the function has no limit as $z \to -r$ $(r > 0)$. The ray $\theta = \pi$ is a branch cut.

Analytic Continuation [33, p. 165]. If $f_1(z)$ is analytic in a domain D_1 and domain D contains D_1, then an analytic function $f(z)$ may exist that equals $f_1(z)$ in D_1. This function is the analytic continuation of $f_1(z)$ onto D, and it is unique. For example,

$$f_1(z) = \sum_{n=0}^{\infty} z^n,\ |z| < 1$$

is analytic in the domain $D_1 : |z| < 1$. The series diverges for other z. Yet the function is the MacLaurin series in the domain

$$f_1(z) = \frac{1}{1-z},\ |z| < 1$$

Thus,

$$f_1(z) = \frac{1}{1-z}$$

is the analytic continuation onto the entire z plane except for $z = 1$.

An extension of the Cauchy integral formula is useful with Laplace transforms. Let the curve C be a straight line parallel to the imaginary axis and z_0 be any point to the right of that (see Fig. 9). A function $f(z)$ is of order z^k as $|z| \to \infty$ if positive numbers M and r_0 exist such that

$$|z^{-k} f(z)| < M \text{ when } |z| > r_0, \text{ i.e.,}$$

$$|f(z)| < M|z|^k \text{ for } |z| \text{ sufficiently large}$$

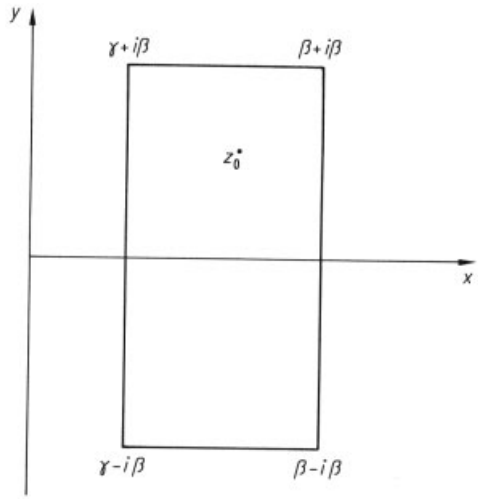

Figure 9. Integration in the complex plane

Theorem [33, p. 167]. Let $f(z)$ be analytic when $R(z) \geq \gamma$ and $O(z^{-k})$ as $|z| \to \infty$ in that half-plane, where γ and k are real constants and $k > 0$. Then for any z_0 such that $R(z_0) > \gamma$

$$f(z_0) = -\frac{1}{2\pi\mathrm{i}} \lim_{\beta\to\infty} \int_{\gamma-\mathrm{i}\beta}^{\gamma+\mathrm{i}\beta} \frac{f(z)}{z-z_0}\mathrm{d}z,$$

i.e., integration takes place along the line $x = \gamma$.

3.5. Other Results

Theorem [32, p. 84]. Let $P(z)$ be a polynomial of degree n having the zeroes $z_1, z_2, \ldots, z_n$ and let π be the least convex polygon containing the zeroes. Then $P'(z)$ cannot vanish anywhere in the exterior of π.

If a polynomial has real coefficients, the roots are either real or form pairs of complex conjugate numbers.

The radius of convergence R of the Taylor series of $f(z)$ about z_0 is equal to the distance from z_0 to the nearest singularity of $f(z)$.

Conformal Mapping. Let $u(x, y)$ be a harmonic function. Introduce the coordinate transformation

$$x = \hat{x}(\xi, \eta), \; y = \hat{y}(\xi, \eta)$$

It is desired that

$$U(\xi, \eta) = u[\hat{x}(\xi, \eta), \hat{y}(\xi, \eta)]$$

be a harmonic function of ξ and η.

Theorem [30, p. 284]. The transformation

$$z = f(\zeta) \tag{6}$$

takes all harmonic functions of x and y into harmonic functions of ξ and η if and only if either $f(\zeta)$ or $f^*(\zeta)$ is an analytic function of $\zeta = \xi + \mathrm{i}\,\eta$.

Equation 6 is a restriction on the transformation which ensures that

$$\text{if } \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \text{ then } \frac{\partial^2 U}{\partial \zeta^2} + \frac{\partial^2 U}{\partial \eta^2} = 0$$

Such a mapping with $f(\zeta)$ analytic and $f'(\zeta) \neq 0$ is a conformal mapping.

If Laplace's equation is to be solved in the region exterior to a closed curve, then the point at infinity is in the domain D. For flow in a long channel (governed by the Laplace equation) the inlet and outlet are at infinity. In both cases the transformation

$$\zeta = \frac{a\,z + b}{z - z_0}$$

takes z_0 into infinity and hence maps D into a bounded domain D*.

4. Integral Transforms [34 – 39]

4.1. Fourier Transforms

Fourier Series [40]**.** Let $f(x)$ be a function that is periodic on $-\pi < x < \pi$. It can be expanded in a Fourier series

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos n\,x + b_n \sin n\,x)$$

where

$$a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\,\mathrm{d}x, \; a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos n\,x\,\mathrm{d}x,$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin n\,x\,\mathrm{d}x$$

The values $\{a_n\}$ and $\{b_n\}$ are called the finite cosine and sine transform of f, respectively. Because

$$\cos n\,x = \frac{1}{2}\left(\mathrm{e}^{\mathrm{i}nx} + \mathrm{e}^{-\mathrm{i}nx}\right)$$

$$\text{and } \sin n\,x = \frac{1}{2\mathrm{i}}\left(\mathrm{e}^{\mathrm{i}nx} - \mathrm{e}^{-\mathrm{i}nx}\right)$$

the Fourier series can be written as

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \mathrm{e}^{-\mathrm{i}nx}$$

where

$$c_n = \begin{cases} \frac{1}{2}(a_n + \mathrm{i}\,b_n) & \text{for} \quad n \geq 0 \\ \frac{1}{2}(a_{-n} - \mathrm{i}\,b_{-n}) & \text{for} \quad n < 0 \end{cases}$$

and

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\,\mathrm{e}^{\mathrm{i}nx}\,\mathrm{d}x$$

If f is real

$$c_{-n} = c_n^*.$$

If f is continuous and piecewise continuously differentiable

$$f'(x) = \sum_{-\infty}^{\infty} (-\mathrm{i}\,n)\,c_n \mathrm{e}^{-\mathrm{i}nx}$$

If f is twice continuously differentiable

$$f''(x) = \sum_{-\infty}^{\infty} \left(-n^2\right) c_n \mathrm{e}^{-\mathrm{i}nx}$$

Inversion. The Fourier series can be used to solve linear partial differential equations with constant coefficients. For example, in the problem

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$$

$$T(x, 0) = f(x)$$

$$T(-\pi, t) = T(\pi, t)$$

Let

$$T = \sum_{-\infty}^{\infty} c_n(t)\,\mathrm{e}^{-\mathrm{i}nx}$$

Then,

$$\sum_{-\infty}^{\infty} \frac{\mathrm{d}c_n}{\mathrm{d}t}\,\mathrm{e}^{-\mathrm{i}nx} = \sum_{-\infty}^{\infty} c_n(t)\left(-n^2\right)\mathrm{e}^{-\mathrm{i}nx}$$

Thus, $c_n(t)$ satisfies

$$\frac{\mathrm{d}c_n}{\mathrm{d}t} = -n^2\,c_n, \text{ or } c_n = c_n(0)\,\mathrm{e}^{-n^2 t}$$

Let $c_n(0)$ be the Fourier coefficients of the initial conditions:

$$f(x) = \sum_{-\infty}^{\infty} c_n(0)\,\mathrm{e}^{-\mathrm{i}nx}$$

The formal solution to the problem is

$$T = \sum_{-\infty}^{\infty} c_n(0)\,\mathrm{e}^{-n^2 t}\mathrm{e}^{-\mathrm{i}nx}$$

Fourier Transform [40]**.** When the function $f(x)$ is defined on the entire real line, the Fourier transform is defined as

$$\boldsymbol{F}[f] \equiv \hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{\mathrm{i}\omega x}\mathrm{d}x$$

This integral converges if

$$\int_{-\infty}^{\infty} |f(x)|\,\mathrm{d}x$$

does. The inverse transformation is

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)\,\mathrm{e}^{-\mathrm{i}\omega x}\mathrm{d}\omega$$

If $f(x)$ is continuous and piecewise continuously differentiable,

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{\mathrm{i}\omega x}\mathrm{d}x$$

converges for each ω, and

$$\lim_{x\to\pm\infty} f(x) = 0$$

then

$$\boldsymbol{F}\left[\frac{\mathrm{d}f}{\mathrm{d}x}\right] = -\mathrm{i}\omega\boldsymbol{F}[f]$$

If f is real $F[f(-\omega)] = F[f(\omega)^*]$. The real part is an even function of ω and the imaginary part is an odd function of ω.

A function $f(x)$ is *absolutely integrable* if the improper integral

$$\int_{-\infty}^{\infty} |f(x)|\,\mathrm{d}x$$

has a finite value. Then the improper integral

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$$

converges. The function is *square integrable* if

$$\int_{-\infty}^{\infty} |f(x)|^2\mathrm{d}x$$

has a finite value. If $f(x)$ and $g(x)$ are square integrable, the product $f(x)\,g(x)$ is absolutely integrable and satisfies the Schwarz inequality:

$$\left|\int_{-\infty}^{\infty} f(x)\,g(x)\,\mathrm{d}x\right|^2 \le \int_{-\infty}^{\infty} |f(x)|^2\mathrm{d}x \int_{-\infty}^{\infty} |g(x)|^2\mathrm{d}x$$

The triangle inequality is also satisfied:

$$\left\{\int_{-\infty}^{\infty} |f+g|^2\mathrm{d}x\right\}^{1/2} \le \left\{\int_{-\infty}^{\infty} |f|^2\mathrm{d}x\right\}^{1/2}\left\{\int_{-\infty}^{\infty} |g|^2\mathrm{d}x\right\}^{1/2}$$

A sequence of square integrable functions $f_n(x)$ converges in the mean to a square integrable function $f(x)$ if

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} |f(x) - f_n(x)|^2\mathrm{d}x = 0$$

The sequence also satisfies the Cauchy criterion

$$\lim_{\substack{n\to\infty\\ m\to\infty}}\int_{-\infty}^{\infty} |f_n - f_m|^2\mathrm{d}x = 0$$

Theorem [40, p. 307]. If a sequence of square integrable functions $f_n(x)$ converges to a function $f(x)$ uniformly on every finite interval $a \le x \le b$, and if it satisfies Cauchy's criterion, then $f(x)$ is square integrable and $f_n(x)$ converges to $f(x)$ in the mean.

Theorem (Riesz – Fischer) [40, p. 308]. To every sequence of square integrable functions $f_n(x)$ that satisfy Cauchy's criterion, there corresponds a square integrable function $f(x)$ such that $f_n(x)$ converges to $f(x)$ in the mean. Thus, the limit in the mean of a sequence of functions is defined to within a null function.

Square integrable functions satisfy the Parseval equation.

$$\int_{-\infty}^{\infty} \left|\hat{f}(\omega)\right|^2 \mathrm{d}\omega = 2\pi \int_{-\infty}^{\infty} |f(x)|^2 \mathrm{d}x$$

This is also the total power in a signal, which can be computed in either the time or the frequency domain. Also

$$\int_{-\infty}^{\infty} \hat{f}(\omega)\, \hat{g}*(\omega)\, \mathrm{d}\omega = 2\pi \int_{-\infty}^{\infty} f(x)\, g*(x)\, \mathrm{d}x$$

Fourier transforms can be used to solve differential equations too. Then it is necessary to find the inverse transformation. If $f(x)$ is square integrable, the Fourier transform of its Fourier transform is $2\pi f(-x)$, or

$$f(x) = \boldsymbol{F}\left[\hat{f}(\omega)\right] = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)\, \mathrm{e}^{-\mathrm{i}\omega x}\mathrm{d}\omega$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\, \mathrm{e}^{\mathrm{i}\omega x}\mathrm{d}x e^{-\mathrm{i}\omega x}\mathrm{d}\omega$$

$$f(x) = \frac{1}{2\pi}\boldsymbol{F}\left[\boldsymbol{F} f(-x)\right] \text{ or } f(-x) = \frac{1}{2\pi}\boldsymbol{F}\left[\boldsymbol{F}[f]\right]$$

Properties of Fourier Transforms [40, p. 324], [15].

$$\boldsymbol{F}\left[\frac{\mathrm{d}f}{\mathrm{d}x}\right] = -\mathrm{i}\,\omega\,\boldsymbol{F}[f] = \mathrm{i}\,\omega\hat{f}$$

$$\boldsymbol{F}[\mathrm{i}\,x\,f(x)] = \frac{\mathrm{d}}{\mathrm{d}\omega}\boldsymbol{F}[f] = \frac{\mathrm{d}}{\mathrm{d}\omega}\hat{f}$$

$$\boldsymbol{F}[f(a\,x-b)] = \frac{1}{|a|}\mathrm{e}^{\mathrm{i}\omega b/a}\hat{f}\left(\frac{\omega}{a}\right)$$

$$\boldsymbol{F}\left[\mathrm{e}^{\mathrm{i}cx} f(x)\right] = \hat{f}(\omega+c)$$

$$\boldsymbol{F}[\cos\omega_0 x\, f(x)] = \frac{1}{2}\left[\hat{f}(\omega+\omega_0) + \hat{f}(\omega-\omega_0)\right]$$

$$\boldsymbol{F}[\sin\omega_0 x\, f(x)] = \frac{1}{2\mathrm{i}}\left[\hat{f}(\omega+\omega_0) - \hat{f}(\omega-\omega_0)\right]$$

$$\boldsymbol{F}\left[\mathrm{e}^{-\mathrm{i}\omega_0 x} f(x)\right] = \hat{f}(\omega-\omega_0)$$

If $f(x)$ is real, then $f(-\omega) = \hat{f}*(\omega)$. If $f(x)$ is imaginary, then $\hat{f}(-\omega) = -\hat{f}*(\omega)$. If $f(x)$ is even, then $\hat{f}(\omega)$ is even. If $f(x)$ is odd, then $\hat{f}(\omega)$ is odd. If $f(x)$ is real and even, then $\hat{f}(\omega)$ is real and even. If $f(x)$ is real and odd, then $\hat{f}(\omega)$ is imaginary and odd. If $f(x)$ is imaginary and even, then $\hat{f}(\omega)$ is imaginary and even. If $f(x)$ is imaginary and odd, then $\hat{f}(\omega)$ is real and odd.

Convolution [40, p. 326].

$$f*h(x_0) \equiv \int_{-\infty}^{\infty} f(x_0-x)\, h(x)\, \mathrm{d}x$$

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}\omega x_0}\hat{f}(-\omega)\,\hat{h}(\omega)\,\mathrm{d}\omega$$

Theorem. The product

$$\hat{f}(\omega)\,\hat{h}(\omega)$$

is the Fourier transform of the convolution product $f*h$. The convolution permits finding inverse transformations when solving differential equations. To solve

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$$

$$T(x,0) = f(x), -\infty<x<\infty$$

T bounded

take the Fourier transform

$$\frac{\mathrm{d}\hat{T}}{\mathrm{d}t} + \omega^2\hat{T} = 0$$

$$\hat{T}(\omega,0) = \hat{f}(\omega)$$

The solution is

$$\hat{T}(\omega,t) = \hat{f}(\omega)\,\mathrm{e}^{-\omega^2 t}$$

The inverse transformation is

$$T(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\omega x}\,\hat{f}(\omega)\,\mathrm{e}^{-\omega^2 t}\mathrm{d}\omega$$

Because

$$\mathrm{e}^{-\omega^2 t} = \boldsymbol{F}\left[\frac{1}{\sqrt{4\pi t}}\,\mathrm{e}^{-x^2/4t}\right]$$

the convolution integral can be used to write the solution as

$$T(x,t) = \frac{1}{\sqrt{4\pi t}}\int_{-\infty}^{\infty} f(y)\,\mathrm{e}^{-(x-y)^2/4t}\mathrm{d}y$$

Finite Fourier Sine and Cosine Transform [41]. In analogy with finite Fourier transforms (on $-\pi$ to π) and Fourier transforms (on $-\infty$ to $+\infty$), finite Fourier sine and cosine transforms (0 to π) and Fourier sine and cosine transforms (on 0 to $+\infty$) can be defined.

The finite Fourier sine and cosine transforms are

$$f_s(n) = \boldsymbol{F}_s^n[f] = \frac{2}{\pi}\int_0^\pi f(x)\sin n x \mathrm{d}x,$$

$$n = 1, 2, \ldots,$$

$$f_c(n) = \boldsymbol{F}_c^n[f] = \frac{2}{\pi}\int_0^\pi f(x)\cos n x \mathrm{d}x$$

$$n = 0, 1, 2, \ldots$$

$$f(x) = \sum_{n=1}^{\infty} f_s(n)\sin n x,$$

$$f(x) = \frac{1}{2} f_c(0) + \sum_{n=1}^{\infty} f_c(n)\cos n x$$

They obey the operational properties

$$\boldsymbol{F}_s^n\left[\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right] = -n^2 \boldsymbol{F}_s^n[f] + \frac{2n}{\pi}[f(0) - (-1)^n f(\pi)]$$

f, f' are continuous, f'' is piecewise continuous on $0 \le x \le \pi$.

$$\left.\begin{aligned} &f_s(n)\cos nk \\ &= \boldsymbol{F}_s^n[\tfrac{1}{2} f_1(x-k) + \tfrac{1}{2} f_1(x+k)] \\ &f_s(n)(-1)^{n+1} = \boldsymbol{F}_s^n[f(\pi - x)] \end{aligned}\right\} \begin{aligned} &f \text{ is piecewise} \\ &\text{continuous on} \\ &0 \le x \le \pi \end{aligned}$$

and f_1 is the extension of f, k is a constant

$$\left.\begin{aligned} &f_1(x) = f(x) \quad 0 < x < \pi \\ &f_1(-x) = -f_1(x) \\ &f_1(x + 2\pi) = f_1(x) \end{aligned}\right\} -\infty < x < \infty$$

Also,

$$\boldsymbol{F}_c^n\left[\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right] = -n^2 \boldsymbol{F}_c^n[f] - \frac{2}{\pi}\frac{\mathrm{d}f}{\mathrm{d}x}(0) + (-1)^n \frac{2}{\pi}\frac{\mathrm{d}f}{\mathrm{d}x}(\pi)$$

$$f_c(n)\cos nk = \boldsymbol{F}_c^n\left[\tfrac{1}{2} f_2(x-k) + \tfrac{1}{2} f_2(x+k)\right]$$

$$f_c(n)(-1)^n = \boldsymbol{F}_c^n[f(\pi - x)]$$

and f_2 is the extension of f.

$$\left.\begin{aligned} &f_2(x) = f(x) \quad 0 < x < \pi \\ &f_2(-x) = f_2(x) \\ &f_2(x + 2\pi) = f_2(x) \end{aligned}\right\} -\infty < x < \infty$$

Table 3. Finite sine transforms [41]

$f_s(n) = \int_0^\pi F(x)\sin n x \mathrm{d}x \ (n = 1, 2, \ldots)$	$F(x) \ (0<x<\pi)$
$(-1)^{n+1} f_s(n)$	$F(\pi - x)$
$\frac{1}{n}$	$\frac{(\pi - x)}{\pi}$
$\frac{(-1)^{n+1}}{n}$	$\frac{x}{\pi}$
$\frac{1-(-1)^n}{n}$	1
$\frac{\pi}{n^2}\sin n c \ (0<c<\pi)$	$\begin{cases} (\pi - c)x & (x \le c) \\ c(\pi - x) & (x \ge c) \end{cases}$
$\frac{\pi}{n}\cos n c \ (0<c<\pi)$	$\begin{cases} -x & (x<c) \\ \pi - x & (x) \end{cases}$
$\frac{\pi^2(-1)^{n-1}}{n} - \frac{2[1-(-1)^n]}{n^3}$	x^2
$\pi(-1)^n\left(\frac{6}{n^3} - \frac{\pi^2}{n}\right)$	x^3
$\frac{n}{n^2+c^2}[1-(-1)^n e^{c\pi}]$	e^{cx}
$\frac{n}{n^2+c^2}$	$\frac{\sinh c(\pi - x)}{\sinh c\pi}$
$\frac{n}{n^2-k^2} \ (\lvert k\rvert \ne 0, 1, 2, \ldots)$	$\frac{\sin k(\pi - x)}{\sin k\pi}$
$0 \ (n \ne m); f_s(m) = \frac{\pi}{2}$	$\sin m x \ (m = 1, 2, \ldots)$
$\frac{n}{n^2-k^2}[1-(-1)^n \cos k x]$	$\cos k x \ (\lvert k\rvert \ne 1, 2, \ldots)$
$\frac{n}{n^2-m^2}[1-(-1)^{n+m}]$, $(n \ne m); f_s(m) = 0$	$\cos m x \ (m = 1, 2, \ldots)$

The material is reproduced with permission of McGrawHill, Inc.

Also,

$$\boldsymbol{F}_s^n\left[\frac{\mathrm{d}f}{\mathrm{d}x}\right] = -n\boldsymbol{F}_c^n[f]$$

$$\boldsymbol{F}_c^n\left[\frac{\mathrm{d}f}{\mathrm{d}x}\right] = n\boldsymbol{F}_s^n[f] - \frac{2}{\pi} f(0) + (-1)^n \frac{2}{\pi} f(\pi)$$

When two functions $F(x)$ and $G(x)$ are defined on the interval $-2\pi < x < 2\pi$, the function

$$F(x) * G(x) = \int_{-\pi}^{\pi} f(x-y)\, g(y)\, \mathrm{d}y$$

is the convolution on $-\pi < x < \pi$. If F and G are both even or both odd, the convolution is even; it is odd if one function is even and the other odd. If F and G are piecewise continuous on $0 \le x \le \pi$, then

$$f_s(n)\, g_s(n) = \boldsymbol{F}_c^n\left[-\tfrac{1}{2} F_1(x) * G_1(x)\right]$$

$$f_s(n)\, g_c(n) = \boldsymbol{F}_s^n\left[\tfrac{1}{2} F_1(x) * G_2(x)\right]$$

$$f_c(n)\, g_c(n) = \boldsymbol{F}_c^n\left[\tfrac{1}{2} F_2(x) * G_2(x)\right]$$

where F_1 and G_1 are odd extensions of F and G, respectively, and F_2 and G_2 are even extensions of F and G, respectively. Finite sine and cosine transforms are listed in Tables 3 and 4.

Table 4. Finite cosine transforms [41]

$f_c(n) = \int_0^\pi F(x)\cos nx\,\mathrm{d}x \; (n = 0,1,\ldots)$	$F(x) \;\; (0<x<\pi)$
$(-1)^n f_c(n)$	$F(\pi-x)$
0 when n=1,2, . . . ; $f_c(0)=\pi$	1
$\frac{2}{n}\sin nc$; $f_c(0) = 2c-\pi$	$\begin{cases} 1 & (0<x<c) \\ -1 & (c<x<\pi) \end{cases}$
$-\frac{1-(-1)^n}{n^2}$; $f_c(0) = \frac{\pi^2}{2}$	x
$\frac{(-1)^n}{n^2}$; $f_c(0) = \frac{\pi^2}{6}$	$\frac{x^2}{2\pi}$
$\frac{1}{n^2}$; $f_c(0) = 0$	$\frac{(\pi-x)^2}{2\pi} - \frac{\pi}{6}$
$\frac{(-1)^n e^{cx}-1}{n^2+c^2}$	$\frac{1}{c}e^{cx}$
$\frac{1}{n^2+c^2}$	$\frac{\cosh c(\pi-x)}{c\sinh c\pi}$
$\frac{(-1)^n\cos k\pi-1}{n^2-k^2}$ $(\lvert k\rvert \neq 0,1,\ldots)$	$\frac{1}{k}\sin kx$
$\frac{(-1)^{n+m}-1}{n^2-m^2}$; $f_c(m) = 0$ $(m = 1,\ldots)$	$\frac{1}{m}\sin mx$
$\frac{1}{n^2-k^2}$ $(\lvert k\rvert \neq 0,1,\ldots)$	$-\frac{\cos k(\pi-x)}{k\sin kx}$
0 $(n\neq m)$; $f_c(m) = \frac{\pi}{2}$ $(m = 1,2,\ldots)$	$\cos mx$

The material is reproduced with permission of McGrawHill, Inc.

On the semi-infinite domain, $0 < x < \infty$, the Fourier sine and cosine transforms are

$$\boldsymbol{F}_{\mathrm{s}}^{\omega}[f] \equiv \int_0^\infty f(x)\sin\omega x\,\mathrm{d}x,$$

$$\boldsymbol{F}_{\mathrm{c}}^{\omega}[f] \equiv \int_0^\infty f(x)\cos\omega x\,\mathrm{d}x \text{ and}$$

$$f(x) = \frac{2}{\pi}\boldsymbol{F}_{\mathrm{s}}^{\omega\prime}\left[\boldsymbol{F}_{\mathrm{s}}^{\omega}[f]\right],\; f(x) = \frac{2}{\pi}\boldsymbol{F}_{\mathrm{c}}^{\omega\prime}\left[\boldsymbol{F}_{\mathrm{c}}^{\omega}[f]\right]$$

The sine transform is an odd function of ω, whereas the cosine function is an even function of ω. Also,

$$\boldsymbol{F}_{\mathrm{s}}^{\omega}\left[\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right] = f(0)\,\omega - \omega^2\boldsymbol{F}_{\mathrm{s}}^{\omega}[f]$$

$$\boldsymbol{F}_{\mathrm{c}}^{\omega}\left[\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right] = -\frac{\mathrm{d}f}{\mathrm{d}x}(0) - \omega^2\boldsymbol{F}_{\mathrm{c}}^{\omega}[f]$$

provided $f(x)$ and $f'(x) \to 0$ as $x \to \infty$. Thus, the sine transform is useful when $f(0)$ is known and the cosine transform is useful when $f'(0)$ is known.

Hsu and Dranoff [42] solved a chemical engineering problem by applying finite Fourier transforms and then using the fast Fourier transform (see Chap. 2).

4.2. Laplace Transforms

Consider a function $F(t)$ defined for $t > 0$. The Laplace transform of $F(t)$ is [41]

$$\boldsymbol{L}[F] = f(s) = \int_0^\infty e^{-st}F(t)\,\mathrm{d}t$$

The Laplace transformation is linear, that is,

$$\boldsymbol{L}[F+G] = \boldsymbol{L}[F] + \boldsymbol{L}[G]$$

Thus, the techniques described herein can be applied only to linear problems. Generally, the assumptions made below are that $F(t)$ is at least piecewise continuous, that it is continuous in each finite interval within $0 < t < \infty$, and that it may take a jump between intervals. It is also of exponential order, meaning $e^{-\alpha t}\lvert F(t)\rvert$ is bounded for all $t > T$, for some finite T.

The unit step function is

$$S_k(t) = \begin{cases} 0 & 0\le t<k \\ 1 & t>k \end{cases}$$

and its Laplace transform is

$$\boldsymbol{L}[S_k(t)] = \frac{e^{-ks}}{s}$$

In particular, if $k = 0$ then

$$\boldsymbol{L}[1] = \frac{1}{s}$$

The Laplace transforms of the first and second derivatives of $F(t)$ are

$$\boldsymbol{L}\left[\frac{\mathrm{d}F}{\mathrm{d}t}\right] = s\,f(s) - F(0)$$

$$\boldsymbol{L}\left[\frac{\mathrm{d}^2F}{\mathrm{d}t^2}\right] = s^2 f(s) - s\,F(0) - \frac{\mathrm{d}F}{\mathrm{d}t}(0)$$

More generally,

$$\boldsymbol{L}\left[\frac{\mathrm{d}^nF}{\mathrm{d}t^n}\right] = s^n f(s) - s^{n-1}F(0) - s^{n-2}\frac{\mathrm{d}F}{\mathrm{d}t}(0) - \cdots - \frac{\mathrm{d}^{n-1}F}{\mathrm{d}t^{n-1}}(0)$$

The inverse Laplace transformation is

$$F(t) = \boldsymbol{L}^{-1}[f(s)] \text{ where } f(s) = \boldsymbol{L}[F]$$

The inverse Laplace transformation is not unique because functions that are identical except for

isolated points have the same Laplace transform. They are unique to within a null function. Thus, if

$$\boldsymbol{L}[F_1] = f(s) \text{ and } \boldsymbol{L}[F_2] = f(s)$$

it must be that

$$F_2 = F_1 + N(t)$$

$$\text{where} \int_0^T (N(t))\,\mathrm{d}t = 0 \text{ for every } T$$

Laplace transforms can be inverted by using Table 5, but knowledge of several rules is helpful.

Table 5. Laplace transforms (see [23] for a more complete list)

$L[F]$	$F(t)$
$\frac{1}{s}$	1
$\frac{1}{s^2}$	t
$\frac{1}{s^n}$	$\frac{t^{n-1}}{(n-1)!}$
$\frac{1}{\sqrt{s}}$	$\frac{1}{\sqrt{\pi t}}$
$s^{-3/2}$	$2\sqrt{t/\pi}$
$\frac{\Gamma(k)}{s^k}$ $(k>0)$	t^{k-1}
$\frac{1}{s-a}$	e^{at}
$\frac{1}{(s-a)^n}$ $(n = 1,2,\ldots)$	$\frac{1}{(n-1)!}t^{n-1}e^{at}$
$\frac{\Gamma(k)}{(s-a)^k}$ $(k>0)$	$t^{k-1}\,e^{at}$
$\frac{1}{(s-a)(s-b)}$	$\frac{1}{a-b}\left(e^{at}-e^{bt}\right)$
$\frac{s}{(s-a)(s-b)}$	$\frac{1}{a-b}\left(ae^{at}-be^{bt}\right)$
$\frac{1}{s^2+a^2}$	$\frac{1}{a}\sin a\,t$
$\frac{s}{s^2+a^2}$	$\cos a\,t$
$\frac{1}{s^2-a^2}$	$\frac{1}{a}\sinh a\,t$
$\frac{s}{s^2-a^2}$	$\cosh a\,t$
$\frac{s}{(s^2+a^2)^2}$	$\frac{t}{2a}\sin a\,t$
$\frac{s^2-a^2}{(s^2+a^2)^2}$	$t\cos a\,t$
$\frac{1}{(s-a)^2+b^2}$	$\frac{1}{b}e^{a\,t}\sin b\,t$
$\frac{s-a}{(s-a)^2+b^2}$	$e^{at}\cos b\,t$

Substitution.

$$f(s-a) = \boldsymbol{L}\left[e^{at}\,F(t)\right]$$

This can be used with polynomials. Suppose

$$f(s) = \frac{1}{s} + \frac{1}{s+3} = \frac{2s+3}{s(s+3)}$$

Because

$$\boldsymbol{L}[1] = \frac{1}{s}$$

Then

$$F(t) = 1 + e^{-3t},\ t \geq 0$$

More generally, translation gives the following.

Translation.

$$f(a\,s-b) = f\left[a\left(s-\frac{b}{a}\right)\right] = \boldsymbol{L}\left[\frac{1}{a}\,e^{bt/a}\,F\left(\frac{t}{a}\right)\right],$$

$$a>0$$

The step function

$$S(t) = \begin{cases} 0 & 0 \leq t < \frac{1}{h} \\ 1 & \frac{1}{h} \leq t < \frac{2}{h} \\ 2 & \frac{2}{h} \leq t < \frac{3}{h} \end{cases}$$

has the Laplace transform

$$\boldsymbol{L}[S(t)] = \frac{1}{s}\frac{1}{1-e^{-hs}}$$

The Dirac delta function $\delta(t-t_0)$ (see Equation 116 in → Mathematical Modeling) has the property

$$\int_0^\infty \delta(t-t_0)\,F(t)\,\mathrm{d}t = F(t_0)$$

Its Laplace transform is

$$\boldsymbol{L}[\delta(t-t_0)] = e^{-st_0},\ t_0 \geq 0,\ s>0$$

The square wave function illustrated in Figure 10 has Laplace transform

$$\boldsymbol{L}[F_c(t)] = \frac{1}{s}\tanh\frac{c\,s}{2}$$

The triangular wave function illustrated in Figure 11 has Laplace transform

$$L[T_c(t)] = \frac{1}{s^2}\tanh\frac{c\,s}{2}$$

Other Laplace transforms are listed in Table 5.

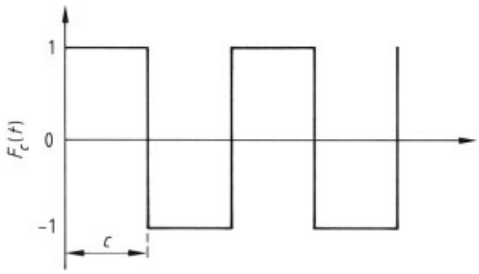

Figure 10. Square wave function

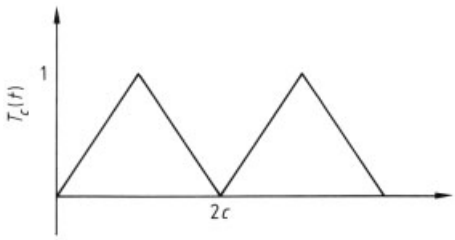

Figure 11. Triangular wave function

Convolution properties are also satisfied:

$$F\ (t)\ *G\ (t) = \int_0^t F\ (\tau)\ G\ (t-\tau)\,\mathrm{d}\tau$$

and

$$f\,(s)\,g\,(s) = \boldsymbol{L}\,[F\,(t)\ *G\,(t)]$$

Derivatives of Laplace Transforms. The Laplace integrals $\boldsymbol{L}\,[F\,(t)]$, $\boldsymbol{L}\,[t\,F\,(t)]$, $\boldsymbol{L}\,[t^2\,F\,(t)]$, . . . are uniformly convergent for $s_1 \geq \alpha$ and

$$\lim_{s\to\infty} f\ (s) = 0,\ \lim_{s\to\infty} \boldsymbol{L}\,[t^n\ F\ (t)] = 0, n = 1, 2,\ldots$$

and

$$\frac{\mathrm{d}^n f}{\mathrm{d}s^n} = \boldsymbol{L}\,[(-t)^n\ F\ (t)]$$

Integration of Laplace Transforms.

$$\int_s^\infty f\ (\xi)\,\mathrm{d}\xi = \boldsymbol{L}\left[\frac{F\ (t)}{t}\right]$$

If $F\ (t)$ is a periodic function, $F\ (t) = F\ (t + a)$, then

$$f\ (s) = \tfrac{1}{1-\mathrm{e}^{-as}}\int_0^a \mathrm{e}^{-st} F\ (t)\,\mathrm{d}t,$$

where $F\ (t) = F\ (t+a)$

Partial Fractions [43]**.** Suppose $q\ (s)$ has m factors

$$q\ (s) = (s-a_1)\,(s-a_2)\cdots(s-a_m)$$

All the factors are linear, none are repeated, and the a_n are all distinct. If $p\ (s)$ has a smaller degree than $q\ (s)$, the Heaviside expansion can be used to evaluate the inverse transformation:

$$\boldsymbol{L}^{-1}\left[\frac{p\ (s)}{q\ (s)}\right] = \sum_{i=1}^{m} \frac{p\ (a_i)}{q'\ (a_i)}\,\mathrm{e}^{a_i t}$$

If the factor $(s - a)$ is repeated m times, then

$$f\ (s) = \tfrac{p\,(s)}{q\,(s)} = \tfrac{A_m}{(s-a)^m} + \tfrac{A_{m-1}}{(s-a)^{m-1}} + \cdots + \tfrac{A_1}{s-a} + h\ (s)$$

where

$$\varphi\ (s) \equiv \tfrac{(s-a)^m\ p\,(s)}{q\,(s)}$$

$$A_m = \varphi\ (a)\,,\ A_k = \tfrac{1}{(m-k)!}\,\tfrac{\mathrm{d}^{m-k}\ \varphi\,(s)}{\mathrm{d}s^{m-k}}|_a,$$

$$k = 1,\ldots,m-1$$

The term $h\ (s)$ denotes the sum of partial fractions not under consideration. The inverse transformation is then

$$F\ (t) = \mathrm{e}^{at}\left(A_m\frac{t^{m-1}}{(m-1)\,!} + A_{m-1}\frac{t^{m-2}}{(m-2)\,!} + \cdots + A_2\frac{t}{1!} + A_1\right) + H\ (t)$$

The term in $F\ (t)$ corresponding to

$s-a$ in $q\ (s)$ is $\varphi\ (a)\ \mathrm{e}^{a\,t}$

$(s-a)^2$ in $q\ (s)$ is $[\varphi'\ (a) + \varphi\ (a)\ t]\,\mathrm{e}^{a\,t}$

$(s-a)^3$ in $q\ (s)$ is $\frac{1}{2}\left[\varphi''\,(a) + 2\varphi'\ (a)\ t + \varphi\ (a)\ t^2\right]\mathrm{e}^{a\,t}$

For example, let

$$f\ (s) = \frac{1}{(s-2)\ (s-1)^2}$$

For the factor $s - 2$,

$$\varphi\ (s) = \frac{1}{(s-1)^2},\ \varphi\ (2) = 1$$

For the factor $(s - 1)^2$,

$$\varphi\ (s) = \tfrac{1}{s-2}, \varphi'\ (s) = -\tfrac{1}{(s-2)^2}$$

$$\varphi\ (1) = -1,\ \varphi'\,(1) = -1$$

The inverse Laplace transform is then

$$F\,(t) = \mathrm{e}^{2t} + [-1-t]\ \mathrm{e}^{t}$$

Quadratic Factors. Let $p(s)$ and $q(s)$ have real coefficients, and $q(s)$ have the factor

$$(s-a)^2+b^2, b>0$$

where a and b are real numbers. Then define $\varphi(s)$ and $h(s)$ and real constants A and B such that

$$f(s)=\frac{p(s)}{q(s)}=\frac{\varphi(s)}{(s-a)^2+b^2}$$
$$=\frac{A\,s+B}{(s-a)^2+b^2}+h(s)$$

Let φ_1 and φ_2 be the real and imaginary parts of the complex number $\varphi(a+\mathrm{i}\,b)$.

$$\varphi(a+\mathrm{i}\,b)\equiv\varphi_1+\mathrm{i}\,\varphi_2$$

Then

$$f(s)=\frac{1}{b}\frac{(s-a)\varphi_2+b\varphi_1}{(s-a)^2+b^2}+h(s)$$

$$F(t)=\frac{1}{b}\,\mathrm{e}^{at}\,(\varphi_2\cos b\,t+\varphi_1\sin b\,t)+H(t)$$

To solve ordinary differential equations by using these results:

$$Y''(t)-2Y'(t)+Y(t)=\mathrm{e}^{2t}$$

$$Y(0)=0, Y'(0)=0$$

Taking Laplace transforms

$$\boldsymbol{L}\left[Y''(t)\right]-2\boldsymbol{L}\left[Y'(t)\right]+\boldsymbol{L}\left[Y(t)\right]=\frac{1}{s-2}$$

using the rules

$$s^2\,y(s)-s\,Y(0)-Y'(0)-2\left[s\,y(s)-Y(0)\right]+y(s)=\frac{1}{s-2}$$

and combining terms

$$\left(s^2-2\,s+1\right)y(s)=\frac{1}{s-2}$$

$$y(s)=\frac{1}{(s-2)(s-1)^2}$$

lead to

$$Y(t)=\mathrm{e}^{2t}-(1+t)\,\mathrm{e}^{t}$$

To solve an integral equation:

$$Y(t)=a+2\int_0^t Y(\tau)\cos(t-\tau)\,\mathrm{d}\tau$$

it is written as

$$Y(t)=a+Y(t)*\cos t$$

Then the Laplace transform is used to obtain

$$y(s)=\frac{a}{s}+2\,y(s)\frac{s}{s^2+1}$$

$$\text{or } y(s)=\frac{a\left(s^2+1\right)}{s(s-1)^2}$$

Taking the inverse transformation gives

$$Y(t)=s\left[1+2t\mathrm{e}^{t}\right]$$

Next, let the variable s in the Laplace transform be complex. $F(t)$ is still a real-valued function of the positive real variable t. The properties given above are still valid for s complex, but additional techniques are available for evaluating the integrals. The real-valued function is O [exp (x_0t)]:

$$|F(t)|<M\mathrm{e}^{x_0t},\ z_0=x_0+\mathrm{i}\,y_0$$

The Laplace transform

$$f(s)=\int_0^\infty \mathrm{e}^{-st}\,F(t)\,\mathrm{d}t$$

is an analytic function of s in the half-plane $x>x_0$ and is absolutely convergent there; it is uniformly convergent on $x\geq x_1>x_0$.

$$\frac{\mathrm{d}^n f}{\mathrm{d}s^n}=\boldsymbol{L}\left[(-t)^n\,F(t)\right]\quad n=1,2,\ldots,x>x_0$$

$$\text{and } f^*(s)=f(s^*)$$

The functions $|f(s)|$ and $|x\,f(s)|$ are bounded in the half-plane $x\geq x_1>x_0$ and $f(s)\to 0$ as $|y|\to\infty$ for each fixed x. Thus,

$$|f(x+\mathrm{i}\,y)|<M,\ |x\,f(x+\mathrm{i}\,y)|<M,$$

$$x\geq x_1>x_0$$

$$\lim_{y\to\pm\infty} f(x+\mathrm{i}\,y)=0,\ x>x_0$$

If $F(t)$ is continuous, $F'(t)$ is piecewise continuous, and both functions are O [exp (x_0t)], then $|f(s)|$ is $O(1/s)$ in each half-plane $x\geq x_1>x_0$.

$$|s\,f(s)|<M$$

If $F(t)$ and $F'(t)$ are continuous, $F''(t)$ is piecewise continuous, and all three functions are O [exp (x_0t)], then

$$\left|s^2f(s)-s\,F(0)\right|<M, x\geq x_1>x_0$$

The additional constraint $F(0)=0$ is necessary and sufficient for $|f(s)|$ to be $O(1/s^2)$.

Inversion Integral [41]. Cauchy's integral formula for $f(s)$ analytic and $O(s^{-k})$ in a half-plane $x\geq y$, $k>0$, is

$$f(s)=\frac{1}{2\,\pi\mathrm{i}}\lim_{\beta\to\infty}\int_{\gamma-\mathrm{i}\beta}^{\gamma+\mathrm{i}\beta}\frac{f(z)}{s-z}\mathrm{d}z,\ \mathrm{Re}(s)>\gamma$$

Applying the inverse Laplace transformation on either side of this equation gives

$$F\ (t) = \frac{1}{2\,\pi \mathrm{i}} \lim_{\beta\to\infty} \int_{\gamma-\mathrm{i}\beta}^{\gamma+\mathrm{i}\beta} \mathrm{e}^{zt}\, f\ (z)\, \mathrm{d}z$$

If $F\ (t)$ is of order $O\ [\exp\ (x_0 t)]$ and $F\ (t)$ and $F'\ (t)$ are piecewise continuous, the inversion integral exists. At any point t_0, where $F\ (t)$ is discontinuous, the inversion integral represents the mean value

$$F\ (t_0) = \lim_{\varepsilon\to\infty} \frac{1}{2}\, [F\ (t_0+\varepsilon) + F\,(t_0-\varepsilon)]$$

When $t = 0$ the inversion integral represents 0.5 $F\ (O+)$ and when $t < 0$, it has the value zero.

If $f\ (s)$ is a function of the complex variable s that is analytic and of order $O\ (s^{-k-m})$ on $R\ (s) \geq x_0$, where $k > 1$ and m is a positive integer, then the inversion integral converges to $F\ (t)$ and

$$\frac{\mathrm{d}^n F}{\mathrm{d}t^n} = \frac{1}{2\,\pi \mathrm{i}} \lim_{\beta\to\infty} \int_{\gamma-\mathrm{i}\beta}^{\gamma+\mathrm{i}\beta} \mathrm{e}^{zt} z^n\, f\ (z)\, \mathrm{d}z,$$

$$n = 1,2,\ldots, m$$

Also $F\ (t)$ and its n derivatives are continuous functions of t of order $O\ [\exp\ (x_0 t)]$ and they vanish at $t = 0$.

$$F\,(0) = F'\,(0) = \cdots F^{(m)}\ (0) = 0$$

Series of Residues [41]. Let $f\ (s)$ be an analytic function except for a set of isolated singular points. An isolated singular point is one for which $f\ (z)$ is analytic for $0 < |z - z_0| < \varrho$ but z_0 is a singularity of $f\ (z)$. An isolated singular point is either a pole, a removable singularity, or an essential singularity. If $f\ (z)$ is not defined in the neighborhood of z_0 but can be made analytic at z_0 simply by defining it at some additional points, then z_0 is a *removable singularity*. The function $f\ (z)$ has a *pole* of order $k \geq 1$ at z_0 if $(z - z_0)^k\, f\ (z)$ has a removable singularity at z_0 whereas $(z - z_0)^{k-1}\, f\ (z)$ has an unremovable isolated singularity at z_0. Any isolated singularity that is not a pole or a removable singularity is an *essential* singularity.

Let the function $f\ (z)$ be analytic except for the isolated singular point $s_1, s_2, \ldots, s_n$. Let $\varrho_n\ (t)$ be the residue of $\mathrm{e}^{zt}\, f\ (z)$ at $z = s_n$ (for definition of residue, see Section 3.4). Then

$$F\ (t) = \sum_{n=1}^{\infty} \varrho_n\ (t)$$

When s_n is a simple pole

$$\varrho_n\ (t) = \lim_{z\to s_n} (z-s_n)\, \mathrm{e}^{zt}\, f\ (z)$$

$$= \mathrm{e}^{s_n t} \lim_{z\to s_n} (z-s_n)\ f\ (z)$$

When

$$f\ (z) = \frac{p\ (z)}{q\ (z)}$$

where $p\ (z)$ and $q\ (z)$ are analytic at $z = s_n$, $p\ (s_n) \neq 0$, then

$$\varrho_n\ (t) = \frac{p\ (s_n)}{q'\ (s_n)}\, \mathrm{e}^{s_n t}$$

If s_n is a removable pole of $f\ (s)$, of order m, then

$$\varphi_n\ (z) = (z-s_n)^m\, f\,(z)$$

is analytic at s_n and the residue is

$$\varrho_n\ (t) = \frac{\Phi_n\ (s_n)}{(m-1)!} \text{ where } \Phi_n\ (z) = \frac{\partial^{m-1}}{\partial z^{m-1}}\, \left[\varphi_n\ (z)\, \mathrm{e}^{zt}\right]$$

An important inversion integral is when

$$f\ (s) = \frac{1}{s} \exp\left(-s^{1/2}\right)$$

The inverse transform is

$$F\ (t) = 1 - \operatorname{erf}\left(\frac{1}{2\sqrt{t}}\right) = \operatorname{erfc}\left(\frac{1}{2\sqrt{t}}\right)$$

where erf is the error function and erfc the complementary error function.

4.3. Solution of Partial Differential Equations by Using Transforms

A common problem facing chemical engineers is to solve the heat conduction equation or diffusion equation

$$\varrho\, C_p \frac{\partial T}{\partial t} = k\, \frac{\partial^2 T}{\partial x^2} \text{ or } \frac{\partial c}{\partial t} = D\, \frac{\partial^2 c}{\partial x^2}$$

The equations can be solved on an infinite domain $-\infty < x < \infty$, a semi-infinite domain $0 \leq x < \infty$, or a finite domain $0 \leq x \leq L$. At a boundary, the conditions can be a fixed temperature $T\ (0, t) = T_0$ (boundary condition of the first kind, or Dirichlet condition), or a fixed flux $-k \frac{\partial T}{\partial x}\ (0,t) = q_0$ (boundary condition of the second kind, or Neumann condition), or a

combination $-k\frac{\partial T}{\partial x}(0,t) = h[T(0,t)-T_0]$ (boundary condition of the third kind, or Robin condition).

The functions T_0 and q_0 can be functions of time. All properties are constant (ϱ, C_p, k, D, h), so that the problem remains linear. Solutions are presented on all domains with various boundary conditions for the heat conduction problem.

$$\frac{\partial T}{\partial t} = \alpha\frac{\partial^2 T}{\partial x^2},\ \alpha = \frac{k}{\varrho C_p}$$

Problem 1. Infinite domain, on $-\infty < x < \infty$.

$T(x,0) = f(x)$, initial conditions

$T(x,t)$ bounded

Solution is via Fourier transforms

$$\hat{T}(\omega,t) = \int_{-\infty}^{\infty} T(x,t)\,\mathrm{e}^{\mathrm{i}\omega x}\,\mathrm{d}x$$

Applied to the differential equation

$$\boldsymbol{F}\left[\frac{\partial^2 T}{\partial x^2}\right] = -\omega^2\,\alpha\boldsymbol{F}[T]$$

$$\frac{\partial\hat{T}}{\partial t}+\omega^2\,\alpha\hat{T} = 0, \hat{T}(\omega,0) = \hat{f}(\omega)$$

By solving

$$\hat{T}(\omega,t) = \hat{f}(\omega)\,\mathrm{e}^{-\omega^2\alpha t}$$

the inverse transformation gives [40, p. 328], [44, p. 58]

$$T(x,t) = \frac{1}{2\pi}\lim_{L\to\infty}\int_{-L}^{L}\mathrm{e}^{-\mathrm{i}\omega x}\,\hat{f}(\omega)\,\mathrm{e}^{-\omega^2\alpha t}\mathrm{d}\omega$$

Another solution is via Laplace transforms; take the Laplace transform of the original differential equation.

$$s\,t(s,x)-f(x) = \alpha\frac{\partial^2 t}{\partial x^2}$$

This equation can be solved with Fourier transforms [40, p. 355]

$$t(s,x) = \frac{1}{2\sqrt{s\alpha}}\int_{-\infty}^{\infty}\mathrm{e}\left\{-\sqrt{\frac{s}{\alpha}}\,|x-y|\right\}f(y)\,\mathrm{d}y$$

The inverse transformation is [40, p. 357], [44, p. 53]

$$T(x,t) = \frac{1}{2\sqrt{\pi\,\alpha\,t}}\int_{-\infty}^{\infty}\mathrm{e}^{-(x-y)^2/4\alpha t}\,f(y)\,\mathrm{d}y$$

Problem 2. Semi-infinite domain, boundary condition of the first kind, on $0 \le x \le \infty$

$T(x,0) = T_0 =$ constant

$T(0,t) = T_1 =$ constant

The solution is

$$T(x,t) = T_0+[T_1-T_0]\left[1-\mathrm{erf}\left(x/\sqrt{4\alpha\,t}\right)\right]$$

$$\text{or } T(x,t) = T_0+(T_1-T_0)\ \mathrm{erfc}\left(x/\sqrt{4\alpha\,t}\right)$$

Problem 3. Semi-infinite domain, boundary condition of the first kind, on $0 \le x < \infty$

$T(x,0) = f(x)$

$T(0,t) = g(t)$

The solution is written as

$$T(x,t) = T_1(x,t)+T_2(x,t)$$

where

$$T_1(x,0) = f(x),\ T_2(x,0) = 0$$

$$T_1(0,t) = 0,\ T_2(0,t) = g(t)$$

Then T_1 is solved by taking the sine transform

$$U_1 = F_s^{\omega}[T_1]$$

$$\frac{\partial U_1}{\partial t} = -\omega^2\,\alpha\,U_1$$

$$U_1(\omega,0) = F_s^{\omega}[f]$$

Thus,

$$U_1(\omega,t) = F_s^{\omega}[f]\,\mathrm{e}^{-\omega^2\alpha t}$$

and [40, p. 322]

$$T_1(x,t) = \frac{2}{\pi}\int_0^{\infty}F_s^{\omega}[f]\,\mathrm{e}^{-\omega^2\alpha t}\sin\omega x\,\mathrm{d}\omega$$

Solve for T_2 by taking the sine transform

$$U_2 = F_s^{\omega}[T_2]$$

$$\frac{\partial U_2}{\partial t} = -\omega^2\,\alpha\,U_2+\alpha\,g(t)\,\omega$$

$$U_2(\omega,0) = 0$$

Thus,

$$U_2(\omega,t)=\int_0^t \mathrm{e}^{-\omega^2\alpha(t-\tau)}\alpha\omega\, g\,(\tau)\,\mathrm{d}\tau$$

and [40, p. 435]

$$T_2(x,t)=\frac{2\,\alpha}{\pi}\int_0^\infty \omega\sin\omega x\int_0^t \mathrm{e}^{-\omega^2\,\alpha(t-\tau)}\,g\,(\tau)\,\mathrm{d}\tau\mathrm{d}\omega$$

The solution for T_1 can also be obtained by Laplace transforms.

$$t_1=\boldsymbol{L}\,[T_1]$$

Applying this to the differential equation

$$st_1-f\,(x)=\alpha\,\frac{\partial^2 t_1}{\partial x^2},\ t_1\,(0,s)=0$$

and solving gives

$$t_1=\frac{1}{\sqrt{s\,\alpha}}\int_0^x \mathrm{e}^{-\sqrt{s/\alpha}(x'-x)}\,f\,(x')\,\mathrm{d}x'$$

and the inverse transformation is [40, p. 437], [44, p. 59]

$$T_1\,(x,t)= \tag{7}$$
$$\frac{1}{\sqrt{4\,\pi\,\alpha\,t}}\int_0^\infty\left[\mathrm{e}^{-(x-\xi)^2/4\alpha t}-\mathrm{e}^{-(x+\xi)^2/4\alpha t}\right]f\,(\xi)\,\mathrm{d}\xi$$

Problem 4. Semi-infinite domain, boundary conditions of the second kind, on $0\le x<\infty$.

$$T\,(x,0)=0$$

$$-k\frac{\partial T}{\partial x}\,(0,t)=q_0=\text{constant}$$

Take the Laplace transform

$$t\,(x,s)=L\,[T\,(x,t)]$$

$$s\,t=\alpha\,\tfrac{\partial^2 t}{\partial x^2}$$

$$-k\,\tfrac{\partial t}{\partial x}=\tfrac{q_0}{s}$$

The solution is

$$t\,(x,s)=\frac{q_0\sqrt{\alpha}}{k\,s^{3/2}}\mathrm{e}^{-x\sqrt{s/\alpha}}$$

The inverse transformation is [41, p. 131], [44, p. 75]

$$T\,(x,t)=\frac{q_0}{k}\left[2\sqrt{\frac{\alpha\,t}{\pi}}\,\mathrm{e}^{-x^2/4\alpha t}-x\,\mathrm{erfc}\left(\frac{x}{\sqrt{4\,\alpha\,t}}\right)\right]$$

Problem 5. Semi-infinite domain, boundary conditions of the third kind, on $0\le x<\infty$

$$T\,(x,0)=f\,(x)$$

$$k\frac{\partial T}{\partial x}\,(0,t)=h\;T\;(0,t)$$

Take the Laplace transform

$$s\,t-f\;(x)=\alpha\,\tfrac{\partial^2 t}{\partial x^2}$$

$$k\,\tfrac{\partial t}{\partial x}\,(0,s)=h\,t\;(0,s)$$

The solution is

$$t\;(x,s)=\int_0^\infty f\;(\xi)\;g\;(x,\xi,s)\,\mathrm{d}\xi$$

where [41, p. 227]

$$2\sqrt{s/\alpha}\,g\;(x,\xi,s)=\exp\left(-\,|x-\xi|\,\sqrt{s/\alpha}\right)$$
$$+\tfrac{\sqrt{s}-h\sqrt{\alpha}/k}{\sqrt{s}+h\sqrt{\alpha}/k}\exp\left[-\,(x+\xi)\;\sqrt{s/\alpha}\right]$$

One form of the inverse transformation is [41, p. 228]

$$T\;(x,t)=\tfrac{2}{\pi}\int_0^\infty \mathrm{e}^{-\beta^2\alpha t}\cos\;[\beta\,x-\mu\;(\beta)]\int_0^\infty f\,(\xi)$$
$$\cos\;[\beta\,\xi-\mu\;(\beta)]\;\mathrm{d}\xi\mathrm{d}\beta$$

$$\mu\;(\beta)=\arg\;(\beta+i\,h/k)$$

Another form of the inverse transformation when $f=T_0=$ constant is [41, p. 231], [44, p. 71]

$$T\;(x,t)=T_0\left[\mathrm{erf}\left(\frac{x}{\sqrt{4\,\alpha\,t}}\right)+\mathrm{e}^{hx/k}\,\mathrm{e}^{h^2\,\alpha t/k^2}\right.$$
$$\left.\mathrm{erfc}\left(\frac{h\sqrt{\alpha t}}{k}+\frac{x}{\sqrt{4\,\alpha\,t}}\right)\right]$$

Problem 6. Finite domain, boundary condition of the first kind

$$T\,(x,0)=T_0=\text{constant}$$

$$T\,(0,t)=T\;(L,t)=0$$

Take the Laplace transform

$$s\,t\,(x,s)-T_0=\alpha\,\frac{\partial^2 t}{\partial x^2}$$

$$t\,(0,s)=t\,(L,s)=0$$

The solution is

$$t\;(x,s)=-\frac{T_0}{s}\frac{\sinh\sqrt{\frac{s}{\alpha}}\,x}{\sinh\sqrt{\frac{s}{\alpha}}\,L}-\frac{T_0}{s}\frac{\sinh\sqrt{\frac{s}{\alpha}}\,(L-x)}{\sinh\sqrt{\frac{s}{\alpha}}\,L}+\frac{T_0}{s}$$

The inverse transformation is [41, p. 220], [44, p. 96]

$$T\ (x,t) = \frac{2}{\pi}\, T_0 \sum\nolimits_{n=1,3,5,\ldots} \frac{2}{n}\, \mathrm{e}^{-n^2\pi^2\alpha t/L^2} \sin\frac{n\,\pi\,x}{L}$$

or (depending on the inversion technique) [40, pp. 362, 438]

$$T\ (x,t) = \frac{T_0}{\sqrt{4\,\pi\,\alpha\,t}} \int_0^L \sum_{n=-\infty}^{\infty} \left[\mathrm{e}^{-[(x-\xi)+2nL]^2/4\alpha t} - \mathrm{e}^{-[(x+\xi)+2nL]^2/4\alpha t}\right] \mathrm{d}\xi$$

Problem 7. Finite domain, boundary condition of the first kind

$$T\,(x,0) = 0$$

$$T\,(0,t) = 0$$

$$T\,(L,0) = T_0 = \text{constant}$$

Take the Laplace transform

$$s\,t\ (x,s) = \alpha\ \tfrac{\partial^2 t}{\partial x^2}$$

$$t\ (0,s) = 0, t\ (L,s) = \tfrac{T_0}{s}$$

The solution is

$$t\ (x,s) = \frac{T_0}{s}\,\frac{\sinh\frac{x}{L}\sqrt{s/\alpha}}{\sinh\sqrt{s/\alpha}}$$

and the inverse transformation is [41, p. 201], [44, p. 313]

$$T\ (x,t) = T_0\left[\frac{x}{L} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n}{n}\,\mathrm{e}^{-n^2\pi^2\alpha t/L^2}\sin\frac{n\,\pi\,x}{L}\right]$$

An alternate transformation is [41, p. 139], [44, p. 310]

$$T\ (x,t) = T_0 \sum_{n=0}^{\infty}\left[\mathrm{erf}\left(\frac{(2\,n+1)\ L+x}{\sqrt{4\,\alpha\,t}}\right) - \mathrm{erf}\left(\frac{(2\,n+1)\ L-x}{\sqrt{4\,\alpha\,t}}\right)\right]$$

Problem 8. Finite domain, boundary condition of the second kind

$$T\,(x,0) = T_0$$

$$\frac{\partial T}{\partial x}\,(0,t) = 0, T\ (L,t) = 0$$

Take the Laplace transform

$$s\,t\ (x,s) - T_0 = \alpha\,\tfrac{\partial^2 t}{\partial x^2}$$

$$\tfrac{\partial t}{\partial x}\,(0,s) = 0, t\ (L,s) = 0$$

The solution is

$$t\ (x,s) = \frac{T_0}{s}\left[1 - \frac{\cosh x\,\sqrt{s/\alpha}}{\cosh L\,\sqrt{s/\alpha}}\right]$$

Its inverse is [41, p. 138]

$$T\ (x,t) = T_0 - T_0 \sum_{n=0}^{\infty}(-1)^n\left[\mathrm{erfc}\left(\frac{(2n+1)\ L-x}{\sqrt{4\,\alpha\,t}}\right) + \mathrm{erfc}\left(\frac{(2n+1)\ L+x}{\sqrt{4\,\alpha\,t}}\right)\right]$$

5. Vector Analysis

Notation. A *scalar* is a quantity having magnitude but no direction (e.g., mass, length, time, temperature, and concentration). A *vector* is a quantity having both magnitude and direction (e.g., displacement, velocity, acceleration, force). A second-order *dyadic* has magnitude and two directions associated with it, as defined precisely below. The most common examples are the stress dyadic (or tensor) and the velocity gradient (in fluid flow). Vectors are printed in boldface type and identified in this chapter by lower-case Latin letters. Second-order dyadics are printed in boldface type and are identified in this chapter by capital or Greek letters. Higher order dyadics are not discussed here. Dyadics can be formed from the components of tensors including their directions, and some of the identities for dyadics are more easily proved by using tensor analysis, which is not presented here (see also, → Transport Phenomena, Chap. 1.1.4). Vectors are also first-order dyadics.

Vectors. Two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are equal if they have the same magnitude and direction. If they have the same magnitude but the opposite direction, then $\boldsymbol{u} = -\boldsymbol{v}$. The sum of two vectors is identified geometrically by placing the vector $\boldsymbol{v}$ at the end of the vector $\boldsymbol{u}$, as shown in Figure 12. The product of a scalar m and a vector $\boldsymbol{u}$ is a vector $m\,\boldsymbol{u}$ in the same direction as $\boldsymbol{u}$ but with a magnitude that equals the magnitude of $\boldsymbol{u}$ times the scalar m. Vectors obey commutative and associative laws.

$\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$	Commutative law for addition
$\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$	Associative law for addition
$m\,\boldsymbol{u} = \boldsymbol{u}\,m$	Commutative law for scalar multiplication
$m\,(n\,\boldsymbol{u}) = (m\,n)\,\boldsymbol{u}$	Associative law for scalar multiplication
$(m + n)\,\boldsymbol{u} = m\,\boldsymbol{u} + n\,\boldsymbol{u}$	Distributive law
$m\,(\boldsymbol{u} + \boldsymbol{v}) = m\,\boldsymbol{u} + m\,\boldsymbol{v}$	Distributive law

The same laws are obeyed by dyadics, as well.

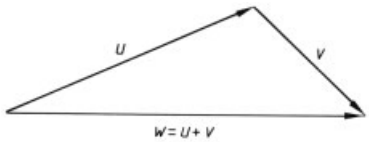

Figure 12. Addition of vectors

A unit vector is a vector with magnitude 1.0 and some direction. If a vector has some magnitude (i.e., not zero magnitude), a unit vector $\boldsymbol{e}_u$ can be formed by

$$\boldsymbol{e}_u = \frac{\boldsymbol{u}}{|\boldsymbol{u}|}$$

The original vector can be represented by the product of the magnitude and the unit vector.

$$\boldsymbol{u} = |\boldsymbol{u}|\,\boldsymbol{e}_u$$

In a cartesian coordinate system the three principle, orthogonal directions are customarily represented by *unit vectors*, such as $\{\boldsymbol{e}_x, \boldsymbol{e}_y, \boldsymbol{e}_z\}$ or $\{\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}\}$. Here, the first notation is used (see Fig. 13). The coordinate system is right handed; that is, if a right-threaded screw rotated from the x to the y direction, it would advance in the z direction. A vector can be represented in terms of its components in these directions, as illustrated in Figure 14. The vector is then written as

$$\boldsymbol{u} = u_x\boldsymbol{e}_x + u_y\boldsymbol{e}_y + u_z\boldsymbol{e}_z$$

The magnitude is

$$|\boldsymbol{u}| = \sqrt{u_x^2 + u_y^2 + u_z^2}$$

The position vector is

$$\boldsymbol{r} = x\boldsymbol{e}_x + y\boldsymbol{e}_y + z\boldsymbol{e}_z$$

with magnitude

$$|\boldsymbol{r}| = \sqrt{x^2 + y^2 + z^2}$$

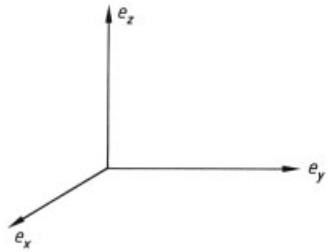

Figure 13. Cartesian coordinate system

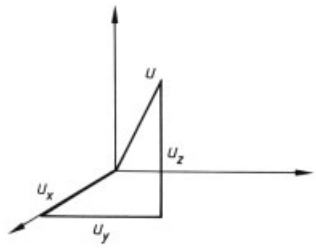

Figure 14. Vector components

Dyadics. The dyadic $\boldsymbol{A}$ is written in component form in cartesian coordinates as

$$\boldsymbol{A} = A_{xx}\boldsymbol{e}_x\boldsymbol{e}_x + A_{xy}\boldsymbol{e}_x\boldsymbol{e}_y + A_{xz}\boldsymbol{e}_x\boldsymbol{e}_z$$

$$+ A_{yx}\boldsymbol{e}_y\boldsymbol{e}_x + A_{yy}\boldsymbol{e}_y\boldsymbol{e}_y + A_{yz}\boldsymbol{e}_y\boldsymbol{e}_z$$

$$+ A_{zx}\boldsymbol{e}_z\boldsymbol{e}_x + A_{zy}\boldsymbol{e}_z\boldsymbol{e}_y + A_{zz}\boldsymbol{e}_z\boldsymbol{e}_z$$

Quantities such as $\boldsymbol{e}_x\boldsymbol{e}_y$ are called *unit dyadics*. They are second-order dyadics and have two directions associated with them, $\boldsymbol{e}_x$ and $\boldsymbol{e}_y$; the order of the pair is important. The components $A_{xx}, \ldots, A_{zz}$ are the components of the tensor A_{ij} which here is a 3×3 matrix of numbers that are transformed in a certain way when variables undergo a linear transformation. The $y\,x$ momentum flux can be defined as the flux of x momentum across an area with unit normal in the y direction (→ Fluid Mechanics, Chap. 2.2; introduction to → Transport Phenomena). Since two directions are involved, a second-order dyadic (or tensor) is needed to represent it, and because the y momentum across an area with unit normal in the x direction may not be the same thing, the order of the indices must be kept straight. The dyadic $\boldsymbol{A}$ is said to be symmetric if

$$\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}$$

Here, the indices i and j can take the values x, y, or z; sometimes (x, 1), (y, 2), (x, 3) are identified and the indices take the values 1, 2, or 3. The dyadic $\boldsymbol{A}$ is said to be antisymmetric if

$$\boldsymbol{A}_{ij} = -\boldsymbol{A}_{ji}$$

The transpose of $\boldsymbol{A}$ is

$$\boldsymbol{A}_{ij}^{\mathrm{T}} = \boldsymbol{A}_{ji}$$

Any dyadic can be represented as the sum of a symmetric portion and an antisymmetric portion.

$$\boldsymbol{A}_{ij} = \boldsymbol{B}_{ij} + \boldsymbol{C}_{ij}, \boldsymbol{B}_{ij} \equiv \tfrac{1}{2}\left(\boldsymbol{A}_{ij} + \boldsymbol{A}_{ji}\right),$$

$$\boldsymbol{C}_{ij} \equiv \tfrac{1}{2}\left(\boldsymbol{A}_{ij} - \boldsymbol{A}_{ji}\right)$$

An ordered pair of vectors is a second-order dyadic.

$$\boldsymbol{u}\,\boldsymbol{v} = \sum_i \sum_j \boldsymbol{e}_{\mathrm{i}} \boldsymbol{e}_{\mathrm{j}} u_i\, v_j$$

The transpose of this is

$$(\boldsymbol{u}\,\boldsymbol{v})^{\mathrm{T}} = \boldsymbol{v}\,\boldsymbol{u}$$

but

$$\boldsymbol{u}\,\boldsymbol{v} \neq \boldsymbol{v}\,\boldsymbol{u}$$

The Kronecker delta is defined as

$$\delta_{ij} = \left\{ \begin{array}{l} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{array} \right\}$$

and the unit dyadic is defined as

$$\delta = \sum_i \sum_j \boldsymbol{e}_i \boldsymbol{e}_j\, \delta_{ij}$$

Operations. The *dot* or *scalar product* of two vectors is defined as

$$\boldsymbol{u} \cdot \boldsymbol{v} = |\boldsymbol{u}|\,|\boldsymbol{v}| \cos\theta,\ 0 \leq \theta \leq \pi$$

where θ is the angle between $\boldsymbol{u}$ *and* $\boldsymbol{v}$. The scalar product of two vectors is a scalar, not a vector. It is the magnitude of $\boldsymbol{u}$ multiplied by the projection of $\boldsymbol{v}$ on $\boldsymbol{u}$, or vice versa. The scalar product of $\boldsymbol{u}$ with itself is just the square of the magnitude of $\boldsymbol{u}$.

$$\boldsymbol{u} \cdot \boldsymbol{u} = \left|\boldsymbol{u}^2\right| = u^2$$

The following laws are valid for scalar products

$\boldsymbol{u} \cdot \boldsymbol{v} = \boldsymbol{v} \cdot \boldsymbol{u}$	Commutative law for scalar products
$\boldsymbol{u} \cdot (\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{u} \cdot \boldsymbol{v} + \boldsymbol{u} \cdot \boldsymbol{w}$	Distributive law for scalar products
$\boldsymbol{e}_x \cdot \boldsymbol{e}_x = \boldsymbol{e}_y \cdot \boldsymbol{e}_y = \boldsymbol{e}_z \cdot \boldsymbol{e}_z = 1$	
$\boldsymbol{e}_x \cdot \boldsymbol{e}_y = \boldsymbol{e}_x \cdot \boldsymbol{e}_z = \boldsymbol{e}_y \cdot \boldsymbol{e}_z = 0$	

If the two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are written in component notation, the scalar product is

$$\boldsymbol{u} \cdot \boldsymbol{v} = u_x\, v_x + u_y\, v_y + u_z\, v_z$$

If $\boldsymbol{u} \cdot \boldsymbol{v} = 0$ and $\boldsymbol{u}$ and $\boldsymbol{v}$ are not null vectors, then $\boldsymbol{u}$ and $\boldsymbol{v}$ are perpendicular to each other and $\theta = \pi/2$.

The *single dot product* of two dyadics is

$$\boldsymbol{A} \cdot \boldsymbol{B} = \sum_i \sum_j \boldsymbol{e}_i \boldsymbol{e}_j \left(\sum_k A_{ik}\, B_{kj}\right)$$

The *double dot product* of two dyadics is

$$\boldsymbol{A} : \boldsymbol{B} = \sum_i \sum_j A_{ij}\, B_{ji}$$

Because the dyadics may not be symmetric, the order of indices and which indices are summed are important. The order is made clearer when the dyadics are made from vectors.

$$(\boldsymbol{u}\,\boldsymbol{v}) \cdot (\boldsymbol{w}\,\boldsymbol{x}) = \boldsymbol{u}\,(\boldsymbol{v} \cdot \boldsymbol{w})\,\boldsymbol{x} = \boldsymbol{u}\,\boldsymbol{x}\,(\boldsymbol{v} \cdot \boldsymbol{w})$$

$$(\boldsymbol{u}\,\boldsymbol{v}) : (\boldsymbol{w}\,\boldsymbol{x}) = (\boldsymbol{u} \cdot \boldsymbol{x})\,(\boldsymbol{v} \cdot \boldsymbol{w})$$

The dot product of a dyadic and a vector is

$$\boldsymbol{A} \cdot \boldsymbol{u} = \sum_i \boldsymbol{e}_i \left(\sum_j A_{ij}\, u_j\right)$$

The *cross* or *vector product* is defined by

$$\boldsymbol{c} = \boldsymbol{u} \times \boldsymbol{v} = \boldsymbol{a}\,|u|\,|v| \sin\theta,\ 0 \leq \theta \leq \pi$$

where $\boldsymbol{a}$ is a unit vector in the direction of $\boldsymbol{u} \times \boldsymbol{v}$. The direction of $\boldsymbol{c}$ is perpendicular to the plane of $\boldsymbol{u}$ and $\boldsymbol{v}$ such that $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{c}$ form a right-handed system. If $\boldsymbol{u} = \boldsymbol{v}$, or $\boldsymbol{u}$ is parallel to $\boldsymbol{v}$, then $\theta = 0$ and $\boldsymbol{u} \times \boldsymbol{v} = 0$. The following laws are valid for cross products.

$\boldsymbol{u} \times \boldsymbol{v} = -\boldsymbol{v} \times \boldsymbol{u}$	Commutative law fails for vector product
$\boldsymbol{u} \times (\boldsymbol{v} \times \boldsymbol{w}) \neq (\boldsymbol{u} \times \boldsymbol{v}) \times \boldsymbol{w}$	Associative law fails for vector product
$\boldsymbol{u} \times (\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{u} \times \boldsymbol{v} + \boldsymbol{u} \times \boldsymbol{w}$	Distributive law for vector product
$\boldsymbol{e}_x \times \boldsymbol{e}_x = \boldsymbol{e}_y \times \boldsymbol{e}_y = \boldsymbol{e}_z \times \boldsymbol{e}_z = 0$	
$\boldsymbol{e}_x \times \boldsymbol{e}_y = \boldsymbol{e}_z,\ \boldsymbol{e}_y \times \boldsymbol{e}_z = \boldsymbol{e}_x,\ \boldsymbol{e}_z \times \boldsymbol{e}_x = \boldsymbol{e}_y$	

$$\boldsymbol{u} \times \boldsymbol{v} = \det \begin{bmatrix} \boldsymbol{e}_x & \boldsymbol{e}_y & \boldsymbol{e}_z \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{bmatrix}$$

$$= \boldsymbol{e}_x\,(u_y\, v_z - v_y\, u_z) + e_y\,(u_z\, v_z - u_x\, v_z)$$

$$+ \boldsymbol{e}_z\,(u_x\, v_y - u_y\, v_x)$$

This can also be written as

$$\boldsymbol{u}\times\boldsymbol{v} = \sum_i \sum_j \varepsilon_{kij}\, u_i\, v_j \boldsymbol{e}_k$$

where

$$\varepsilon_{ijk} = \begin{cases} 1 & \text{if } i,j,k \text{ is an even permutation of } 123 \\ -1 & \text{if } i,j,k \text{ is an odd permutation of } 123 \\ 0 & \text{if any two of } i,j,k \text{ are equal} \end{cases}$$

Thus ε_{123} = 1, ε_{132} =- 1, ε_{312} = 1, ε_{112} = 0, for example.

The magnitude of $\boldsymbol{u}\times\boldsymbol{v}$ is the same as the area of a parallelogram with sides $\boldsymbol{u}$ and $\boldsymbol{v}$. If $\boldsymbol{u}\times\boldsymbol{v}$ = $\boldsymbol{0}$ and $\boldsymbol{u}$ *and* $\boldsymbol{v}$ are not null vectors, then $\boldsymbol{u}$ *and* $\boldsymbol{v}$ are parallel. Certain triple products are useful.

$$(\boldsymbol{u}\cdot\boldsymbol{v})\,\boldsymbol{w} \neq \boldsymbol{u}\,(\boldsymbol{v}\cdot\boldsymbol{w})$$

$$\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w}) = \boldsymbol{v}\cdot(\boldsymbol{w}\times\boldsymbol{u}) = \boldsymbol{w}\cdot(\boldsymbol{u}\times\boldsymbol{v})$$

$$\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w}) = (\boldsymbol{u}\cdot\boldsymbol{w})\,v - (\boldsymbol{u}\cdot\boldsymbol{v})\,\boldsymbol{w}$$

$$(\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w} = (\boldsymbol{u}\cdot\boldsymbol{w})\,v - (\boldsymbol{v}\cdot\boldsymbol{w})\,\boldsymbol{u}$$

The cross product of a dyadic and a vector is defined as

$$\boldsymbol{A}\times\boldsymbol{u} = \sum_i \sum_j \boldsymbol{e}_i \boldsymbol{e}_j \left(\sum_k \sum_l \varepsilon_{klj}\, A_{ik}\, u_l\right)$$

The magnitude of a dyadic is

$$|\boldsymbol{A}| = A = \sqrt{\frac{1}{2}\,(\boldsymbol{A}{:}\boldsymbol{A}^{\mathrm{T}})} = \sqrt{\frac{1}{2}\sum_i \sum_j A_{ij}^2}$$

There are three *invariants* of a dyadic. They are called invariants because they take the same value in any coordinate system and are thus an intrinsic property of the dyadic. They are the trace of $\boldsymbol{A}, \boldsymbol{A}^2, \boldsymbol{A}^3$ [45].

$$\mathrm{I} = \mathrm{tr}\boldsymbol{A} = \sum_i A_{ii}$$

$$\mathrm{II} = \mathrm{tr}\boldsymbol{A}^2 = \sum_i \sum_j A_{ij}\, A_{ji}$$

$$\mathrm{III} = \mathrm{tr}\boldsymbol{A}^3 = \sum_i \sum_j \sum_k A_{ij}\, A_{jk}\, A_{ki}$$

The invariants can also be expressed as

$$\mathrm{I}_1 = \mathrm{I}$$

$$\mathrm{I}_2 = \tfrac{1}{2}\,\left(\mathrm{I}^2 - \mathrm{II}\right)$$

$$\mathrm{I}_3 = \tfrac{1}{6}\,\left(\mathrm{I}^3 - 3\,\mathrm{I}\cdot\mathrm{II} + 2\,\mathrm{III}\right) = \det \boldsymbol{A}$$

Invariants of two dyadics are available [46]. Because a second-order dyadic has nine components, the *characteristic equation*

$$\det\,(\lambda\,\boldsymbol{\delta} - \boldsymbol{A}) = 0$$

can be formed where λ is an eigenvalue. This expression is

$$\lambda^3 - I_1\lambda^2 + I_2\lambda - I_3 = 0$$

An important theorem of Hamilton and Cayley [47] is that a second-order dyadic satisfies its own characteristic equation.

$$\boldsymbol{A}^3 - I_1\boldsymbol{A}^2 + I_2\boldsymbol{A} - I_3\boldsymbol{\delta} = 0 \tag{7}$$

Thus $\boldsymbol{A}^3$ can be expressed in terms of $\boldsymbol{\delta}$, $\boldsymbol{A}$, and $\boldsymbol{A}^2$. Similarly, higher powers of $\boldsymbol{A}$ can be expressed in terms of $\boldsymbol{\delta}$, $\boldsymbol{A}$, and $\boldsymbol{A}^2$. Decomposition of a dyadic into a symmetric and an antisymmetric part was shown above. The antisymmetric part has zero trace. The symmetric part can be decomposed into a part with a trace (the isotropic part) and a part with zero trace (the deviatoric part).

$$\boldsymbol{A} = \underbrace{1/3\,\boldsymbol{\delta}\boldsymbol{\delta}:\boldsymbol{A}}_{\text{Isotropic}} + \underbrace{1/2\,[\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}} - 2/3\,\boldsymbol{\delta}\boldsymbol{\delta}:\boldsymbol{A}]}_{\text{Deviatoric}} + \underbrace{1/2\,[\boldsymbol{A} - \boldsymbol{A}^{\mathrm{T}}]}_{\text{Antisymmetric}}$$

Differentiation. The derivative of a vector is defined in the same way as the derivative of a scalar. Suppose the vector $\boldsymbol{u}$ depends on t. Then

$$\frac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} = \lim_{\Delta t\to 0} \frac{\boldsymbol{u}\,(t+\Delta t) - \boldsymbol{u}\,(t)}{\Delta t}$$

If the vector is the position vector $\boldsymbol{r}\,(t)$, then the difference expression is a vector in the direction of $\Delta\,\boldsymbol{r}$ (see Fig. 15). The derivative

$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = \lim_{\Delta t\to 0} \frac{\Delta\boldsymbol{r}}{\Delta t} = \lim_{\Delta t\to 0} \frac{\boldsymbol{r}\,(t+\Delta t) - \boldsymbol{r}\,(t)}{\Delta t}$$

is the velocity. The derivative operation obeys the following laws.

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\boldsymbol{u}+\boldsymbol{v}) = \tfrac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t} + \tfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}$$

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\boldsymbol{u}\cdot\boldsymbol{v}) = \tfrac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}\cdot\boldsymbol{v} + \boldsymbol{u}\cdot\tfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}$$

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\boldsymbol{u}\times\boldsymbol{v}) = \tfrac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}\times\boldsymbol{v} + \boldsymbol{u}\times\tfrac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}$$

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\,(\varphi\boldsymbol{u}) = \tfrac{\mathrm{d}\varphi}{\mathrm{d}t}\boldsymbol{u} + \varphi\tfrac{\mathrm{d}\boldsymbol{u}}{\mathrm{d}t}$$

If the vector $\boldsymbol{u}$ depends on more than one variable, such as x, y and z, partial derivatives are defined in the usual way. For example,

$$\text{if } \boldsymbol{u}\,(x, y, z)\text{, then}$$

$$\frac{\partial\boldsymbol{u}}{\partial x} = \lim_{\Delta t\to 0} \frac{\boldsymbol{u}(x+\Delta x,\, y,\, z) - \boldsymbol{u}(x,\, y,\, z)}{\Delta x}$$

Rules for differentiation of scalar and vector products are

$$\frac{\partial}{\partial x}(\boldsymbol{u}\cdot\boldsymbol{v}) = \frac{\partial \boldsymbol{u}}{\partial x}\cdot\boldsymbol{v}+\boldsymbol{u}\cdot\frac{\partial \boldsymbol{v}}{\partial x}$$

$$\frac{\partial}{\partial x}(\boldsymbol{u}\times\boldsymbol{v}) = \frac{\partial \boldsymbol{u}}{\partial x}\times\boldsymbol{v}+\boldsymbol{u}\times\frac{\partial \boldsymbol{v}}{\partial x}$$

Differentials of vectors are

$$\mathrm{d}\boldsymbol{u} = \mathrm{d}u_x\boldsymbol{e}_x+\mathrm{d}u_y\boldsymbol{e}_y+\mathrm{d}u_z\boldsymbol{e}_z$$

$$\mathrm{d}(\boldsymbol{u}\cdot\boldsymbol{v}) = \mathrm{d}\boldsymbol{u}\cdot\boldsymbol{v}+\boldsymbol{u}\cdot\mathrm{d}\boldsymbol{v}$$

$$\mathrm{d}(\boldsymbol{u}\times\boldsymbol{v}) = \mathrm{d}\boldsymbol{u}\times\boldsymbol{v}+\boldsymbol{u}\times\mathrm{d}\boldsymbol{v}$$

$$\mathrm{d}\boldsymbol{u} = \frac{\partial \boldsymbol{u}}{\partial x}\mathrm{d}x+\frac{\partial \boldsymbol{u}}{\partial y}\mathrm{d}y+\frac{\partial \boldsymbol{u}}{\partial z}\mathrm{d}z$$

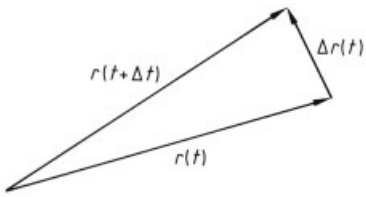

Figure 15. Vector differentiation

If a curve is given by $\boldsymbol{r}(t)$, the length of the curve is [43]

$$L = \int_a^b \sqrt{\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}\cdot\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t}}\,\mathrm{d}t$$

The arc-length function can also be defined:

$$s(t) = \int_a^t \sqrt{\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t^*}\cdot\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t^*}}\,\mathrm{d}t^*$$

This gives

$$\left(\frac{\mathrm{d}s}{\mathrm{d}t}\right)^2 = \frac{\mathrm{d}r}{\mathrm{d}t}\cdot\frac{\mathrm{d}r}{\mathrm{d}t} = \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right)^2+\left(\frac{\mathrm{d}z}{\mathrm{d}t}\right)^2$$

Because

$$d\boldsymbol{r} = \mathrm{d}x\boldsymbol{e}_x+\mathrm{d}y\boldsymbol{e}_y+\mathrm{d}z\boldsymbol{e}_z$$

then

$$ds^2 = \mathrm{d}\boldsymbol{r}\cdot\mathrm{d}\boldsymbol{r} = \mathrm{d}x^2+\mathrm{d}y^2+\mathrm{d}z^2$$

The derivative d$\boldsymbol{r}$/dt is tangent to the curve in the direction of motion

$$\boldsymbol{u} = \frac{\frac{\mathrm{d}r}{\mathrm{d}t}}{\left|\frac{\mathrm{d}r}{\mathrm{d}t}\right|}$$

Also,

$$\boldsymbol{u} = \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s}$$

Differential Operators (see also, → Transport Phenomena, Chap. 1.1.4). The vector differential operator (del operator) $\boldsymbol{\nabla}$ is defined in cartesian coordinates by

$$\boldsymbol{\nabla} = \boldsymbol{e}_x\frac{\partial}{\partial x}+\boldsymbol{e}_y\frac{\partial}{\partial y}+\boldsymbol{e}_z\frac{\partial}{\partial z}$$

The *gradient* of a scalar function is defined

$$\boldsymbol{\nabla}\varphi = \boldsymbol{e}_x\frac{\partial\varphi}{\partial x}+\boldsymbol{e}_y\frac{\partial\varphi}{\partial y}+\boldsymbol{e}_z\frac{\partial\varphi}{\partial z}$$

and is a vector. If φ is height or elevation, the gradient is a vector pointing in the uphill direction. The steeper the hill, the larger is the magnitude of the gradient.

The *divergence* of a vector is defined by

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = \frac{\partial u_x}{\partial x}+\frac{\partial u_y}{\partial y}+\frac{\partial u_z}{\partial z}$$

and is a scalar. For a volume element ΔV, the net outflow of a vector $\boldsymbol{u}$ over the surface of the element is

$$\int_{\Delta S}\boldsymbol{u}\cdot\boldsymbol{n}\mathrm{d}S$$

This is related to the divergence by [48, p. 411]

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = \lim_{\Delta V\to 0}\frac{1}{\Delta V}\int_{\Delta S}\boldsymbol{u}\cdot\boldsymbol{n}\mathrm{d}S$$

Thus, the divergence is the net outflow per unit volume.

The *curl* of a vector is defined by

$$\boldsymbol{\nabla}\times\boldsymbol{u} = \left(\boldsymbol{e}_x\frac{\partial}{\partial x}+\boldsymbol{e}_y\frac{\partial}{\partial y}+\boldsymbol{e}_z\frac{\partial}{\partial z}\right)$$

$$\times(\boldsymbol{e}_x u_x+\boldsymbol{e}_y u_y+\boldsymbol{e}_z u_z)$$

$$= \boldsymbol{e}_x\left(\frac{\partial u_z}{\partial y}-\frac{\partial u_y}{\partial z}\right)+\boldsymbol{e}_y\left(\frac{\partial u_x}{\partial z}-\frac{\partial u_z}{\partial x}\right)$$

$$+\boldsymbol{e}_z\left(\frac{\partial u_y}{\partial x}-\frac{\partial u_x}{\partial y}\right)$$

and is a vector. It is related to the integral

$$\int_C \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{s} = \int_C u_s\mathrm{d}s$$

which is called the circulation of $\boldsymbol{u}$ around path C. This integral depends on the vector and the contour C, in general. If the circulation does not depend on the contour C, the vector is said to be irrotational; if it does, it is rotational. The relationship with the curl is [48, p. 419]

$$\boldsymbol{n}\cdot(\boldsymbol{\nabla}\times\boldsymbol{u}) = \lim_{\Delta\to 0}\frac{1}{\Delta S}\int_C \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{s}$$

Thus, the normal component of the curl equals the net circulation per unit area enclosed by the contour C.

The gradient, divergence, and curl obey a distributive law but not a commutative or associative law.

$$\boldsymbol{\nabla}(\varphi+\psi) = \boldsymbol{\nabla}\varphi+\boldsymbol{\nabla}\psi$$

$$\boldsymbol{\nabla}\cdot(\boldsymbol{u}+\boldsymbol{v}) = \boldsymbol{\nabla}\cdot\boldsymbol{u}+\boldsymbol{\nabla}\cdot\boldsymbol{v}$$

$$\boldsymbol{\nabla}\times(\boldsymbol{u}+\boldsymbol{v}) = \boldsymbol{\nabla}\times\boldsymbol{u}+\boldsymbol{\nabla}\times\boldsymbol{v}$$

$$\boldsymbol{\nabla}\cdot\varphi\neq\varphi\boldsymbol{\nabla}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u}\neq\boldsymbol{u}\cdot\boldsymbol{\nabla}$$

Useful formulas are [49]

$$\boldsymbol{\nabla}\cdot(\varphi\boldsymbol{u}) = \boldsymbol{\nabla}\varphi\cdot\boldsymbol{u}+\varphi\boldsymbol{\nabla}\cdot\boldsymbol{u}$$

$$\boldsymbol{\nabla}\times(\varphi\boldsymbol{u}) = \boldsymbol{\nabla}\varphi\times\boldsymbol{u}+\varphi\boldsymbol{\nabla}\times\boldsymbol{u}$$

$$\boldsymbol{\nabla}\cdot(\boldsymbol{u}\times\boldsymbol{v}) = \boldsymbol{v}\cdot(\boldsymbol{\nabla}\times\boldsymbol{u})-\boldsymbol{u}\cdot(\boldsymbol{\nabla}\times\boldsymbol{v})$$

$$\boldsymbol{\nabla}\times(\boldsymbol{u}\times\boldsymbol{v}) = \boldsymbol{v}\cdot\boldsymbol{\nabla}\boldsymbol{u}-\boldsymbol{v}(\boldsymbol{\nabla}\cdot\boldsymbol{u})-\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{v}+\boldsymbol{u}(\boldsymbol{\nabla}\cdot\boldsymbol{v})$$

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{u}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{u})-\nabla^2\boldsymbol{u}$$

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\varphi) = \nabla^2\varphi = \frac{\partial^2\varphi}{\partial x^2}+\frac{\partial^2\varphi}{\partial y^2}+\frac{\partial^2\varphi}{\partial z^2}, \text{ where } \nabla^2$$

is called the Laplacian operator. $\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\varphi) = \boldsymbol{0}$. The curl of the gradient of φ is zero.

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\times\boldsymbol{u}) = 0$$

The divergence of the curl of $\boldsymbol{u}$ is zero. Formulas useful in fluid mechanics are

$$\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}\boldsymbol{v})^{\mathrm{T}} = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{v})$$

$$\boldsymbol{\nabla}\cdot(\tau\cdot\boldsymbol{v}) = \boldsymbol{v}\cdot(\boldsymbol{\nabla}\cdot\tau)+\tau:\boldsymbol{\nabla}\boldsymbol{v}$$

$$\boldsymbol{v}\cdot\boldsymbol{\nabla}\boldsymbol{v} = \tfrac{1}{2}\boldsymbol{\nabla}(\boldsymbol{v}\cdot\boldsymbol{v})-\boldsymbol{v}\times(\boldsymbol{\nabla}\times\boldsymbol{v})$$

If a coordinate system is transformed by a rotation and translation, the coordinates in the new system (denoted by primes) are given by

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

Any function that has the same value in all coordinate systems is an invariant. The gradient of an invariant scalar field is invariant; the same is true for the divergence and curl of invariant vectors fields.

The gradient of a vector field is required in fluid mechanics because the velocity gradient is used. It is defined as

$$\boldsymbol{\nabla}\boldsymbol{v} = \sum_i\sum_j \boldsymbol{e}_i\boldsymbol{e}_j\frac{\partial v_j}{\partial x_i} \text{ and}$$

$$(\boldsymbol{\nabla}\boldsymbol{v})^{\mathrm{T}} = \sum_i\sum_j \boldsymbol{e}_i\boldsymbol{e}_j\frac{\partial v_i}{\partial x_j}$$

The divergence of dyadics is defined

$$\boldsymbol{\nabla}\cdot\boldsymbol{\tau} = \sum_i \boldsymbol{e}_i\left(\sum_j\frac{\partial\tau_{ji}}{\partial x_j}\right) \text{ and}$$

$$\boldsymbol{\nabla}\cdot(\varphi\boldsymbol{u}\,\boldsymbol{v}) = \sum_i \boldsymbol{e}_i\left[\sum_j\frac{\partial}{\partial x_j}(\varphi u_j\,v_i)\right]$$

where $\boldsymbol{\tau}$ is any second-order dyadic.

Useful relations involving dyadics are

$$(\varphi\,\boldsymbol{\delta}:\boldsymbol{\nabla}\boldsymbol{v}) = \varphi(\boldsymbol{\nabla}\cdot\boldsymbol{v})$$

$$\boldsymbol{\nabla}\cdot(\varphi\,\boldsymbol{\delta}) = \boldsymbol{\nabla}\varphi$$

$$\boldsymbol{\nabla}\cdot(\varphi\boldsymbol{\tau}) = \boldsymbol{\nabla}\varphi\cdot\boldsymbol{\tau}+\varphi\boldsymbol{\nabla}\cdot\boldsymbol{\tau}$$

$$\boldsymbol{n}\,\boldsymbol{t}:\boldsymbol{\tau} = \boldsymbol{t}\cdot\boldsymbol{\tau}\cdot\boldsymbol{n} = \boldsymbol{\tau}:\boldsymbol{n}\,\boldsymbol{t}$$

A surface can be represented in the form

$$f\,(x,y,z) = c = \text{ constant}$$

The normal to the surface is given by

$$\boldsymbol{n} = \frac{\boldsymbol{\nabla}f}{|\boldsymbol{\nabla}f|}$$

provided the gradient is not zero. Operations can be performed entirely within the surface. Define

$$\boldsymbol{\delta}_{\mathrm{II}}\equiv\boldsymbol{\delta}-\boldsymbol{n}\,\boldsymbol{n}, \boldsymbol{\nabla}_{\mathrm{II}}\equiv\boldsymbol{\delta}_{\mathrm{II}}\cdot\boldsymbol{\nabla}, \frac{\partial}{\partial n}\equiv\boldsymbol{n}\cdot\boldsymbol{\nabla}$$

$$\boldsymbol{v}_{\mathrm{II}}\equiv\boldsymbol{\delta}_{\mathrm{II}}\cdot\boldsymbol{v}, v_n\equiv\boldsymbol{n}\cdot\boldsymbol{v}$$

Then a vector and del operator can be decomposed into

$$\boldsymbol{v} = \boldsymbol{v}_{\mathrm{II}}+\boldsymbol{n}v_n, \boldsymbol{\nabla} = \boldsymbol{\nabla}_{\mathrm{II}}+\boldsymbol{n}\frac{\partial}{\partial n}$$

The velocity gradient can be decomposed into

$$\boldsymbol{\nabla}\boldsymbol{v} = \boldsymbol{\nabla}_{\mathrm{II}}\boldsymbol{v}_{\mathrm{II}}+(\boldsymbol{\nabla}_{\mathrm{II}}\boldsymbol{n})\,v_n+\boldsymbol{n}\boldsymbol{\nabla}_{\mathrm{II}}v_n+\boldsymbol{n}\frac{\partial\boldsymbol{v}_{\mathrm{II}}}{\partial n}+\boldsymbol{n}\,\boldsymbol{n}\frac{\partial v_n}{\partial n}$$

The surface gradient of the normal is the negative of the curvature dyadic of the surface.

$$\boldsymbol{\nabla}_{\mathrm{II}}\boldsymbol{n} = -\boldsymbol{B}$$

The surface divergence is then

$$\boldsymbol{\nabla}_{\mathrm{II}}\cdot\boldsymbol{v} = \boldsymbol{\delta}_{\mathrm{II}}:\boldsymbol{\nabla}\boldsymbol{v} = \boldsymbol{\nabla}_{\mathrm{II}}\cdot\boldsymbol{v}_{\mathrm{II}}-2\,H\,v_n$$

where H is the mean curvature.

$$H = \frac{1}{2}\boldsymbol{\delta}_{\mathrm{II}} : \boldsymbol{B}$$

The surface curl can be a scalar

$$\boldsymbol{\nabla}_{\mathrm{II}} \times \boldsymbol{v} = -\boldsymbol{\varepsilon}_{\mathrm{II}} : \boldsymbol{\nabla}\boldsymbol{v} = -\boldsymbol{\varepsilon}_{\mathrm{II}} : \boldsymbol{\nabla}_{\mathrm{II}}\boldsymbol{v}_{\mathrm{II}} = -\boldsymbol{n} \cdot (\boldsymbol{\nabla} \times \boldsymbol{v}),$$

$$\boldsymbol{\varepsilon}_{\mathrm{II}} = \boldsymbol{n} \cdot \boldsymbol{\varepsilon}$$

or a vector

$$\boldsymbol{\nabla}_{\mathrm{II}} \times \boldsymbol{v} \equiv \boldsymbol{\nabla}_{\mathrm{II}} \times \boldsymbol{v}_{\mathrm{II}} = \boldsymbol{n}\boldsymbol{\nabla}_{\mathrm{II}} \times \boldsymbol{v} = \boldsymbol{n}\,\boldsymbol{n} \cdot \boldsymbol{\nabla} \times v$$

Vector Integration [48, pp. 206 – 212]. If $\boldsymbol{u}$ is a vector, then its integral is also a vector.

$$\int \boldsymbol{u}(t)\,\mathrm{d}t = \boldsymbol{e}_x \int u_x(t)\,\mathrm{d}t + \boldsymbol{e}_y \int u_y(t)\,\mathrm{d}t + \boldsymbol{e}_z \int u_z(t)\,\mathrm{d}t$$

If the vector $\boldsymbol{u}$ is the derivative of another vector, then

$$\boldsymbol{u} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}, \int \boldsymbol{u}(t)\,\mathrm{d}t = \int \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}\,\mathrm{d}t = \boldsymbol{v} + \text{constant}$$

If $\boldsymbol{r}(t)$ is a position vector that defines a curve C, the line integral is defined by

$$\int_C \boldsymbol{u} \cdot \mathrm{d}\boldsymbol{r} = \int_C (u_x \mathrm{d}x + u_y \mathrm{d}y + u_z \mathrm{d}z)$$

Theorems about this line integral can be written in various forms.

Theorem [43]. If the functions appearing in the line integral are continuous in a domain D, then the line integral is independent of the path C if and only if the line integral is zero on every simple closed path in D.

Theorem [43]. If $\boldsymbol{u} = \boldsymbol{\nabla}\varphi$ where φ is single-valued and has continuous derivatives in D, then the line integral is independent of the path C and the line integral is zero for any closed curve in D.

Theorem [43]. If f, g, and h are continuous functions of x, y, and z, and have continuous first derivatives in a simply connected domain D, then the line integral

$$\int_C (f\,\mathrm{d}x + g\mathrm{d}y + h\mathrm{d}z)$$

is independent of the path if and only if

$$\frac{\partial h}{\partial y} = \frac{\partial g}{\partial z}, \frac{\partial f}{\partial z} = \frac{\partial h}{\partial x}, \frac{\partial g}{\partial x} = \frac{\partial f}{\partial y}$$

or if f, g, and h are regarded as the x, y, and z components of a vector $\boldsymbol{v}$:

$$\boldsymbol{\nabla} \times \boldsymbol{v} = 0$$

Consequently, the line integral is independent of the path (and the value is zero for a closed contour) if the three components in it are regarded as the three components of a vector and the vector is derivable from a potential (or zero curl). The conditions for a vector to be derivable from a potential are just those in the third theorem. In two dimensions this reduces to the more usual theorem.

Theorem [48, p. 207]. If M and N are continuous functions of x and y that have continuous first partial derivatives in a simply connected domain D, then the necessary and sufficient condition for the line integral

$$\int_C (M\mathrm{d}x + N\mathrm{d}y)$$

to be zero around every closed curve C in D is

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$$

If a vector is integrated over a surface with incremental area $\mathrm{d}S$ and normal to the surface $\boldsymbol{n}$, then the surface integral can be written as

$$\int\int_S \boldsymbol{u} \cdot \mathrm{d}\boldsymbol{S} = \int\int_S \boldsymbol{u} \cdot \boldsymbol{n}\mathrm{d}S$$

If $\boldsymbol{u}$ is the velocity then this integral represents the flow rate past the surface S.

Divergence Theorem [48, 49]. If V is a volume bounded by a closed surface S and $\boldsymbol{u}$ is a vector function of position with continuous derivatives, then

$$\int_V \boldsymbol{\nabla} \cdot \boldsymbol{u}\mathrm{d}V = \int_S \boldsymbol{n} \cdot \boldsymbol{u}\mathrm{d}S = \int_S \boldsymbol{u} \cdot \boldsymbol{n}\mathrm{d}S = \int_S \boldsymbol{u} \cdot \mathrm{d}\boldsymbol{S}$$

where $\boldsymbol{n}$ is the normal pointing outward to S. The normal can be written as

$$\boldsymbol{n} = \boldsymbol{e}_x \cos(x, n) + \boldsymbol{e}_y \cos(y, n) + \boldsymbol{e}_z \cos(z, n)$$

where, for example, cos (x, n) is the cosine of the angle between the normal n and the x axis. Then the divergence theorem in component form is

$$\int_V \left(\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z}\right) \mathrm{d}x\mathrm{d}y\mathrm{d}z =$$

$$\int_S [u_x \cos(x, n) + u_y \cos(y, n) + u_z \cos(z, n)]\,\mathrm{d}S$$

If the divergence theorem is written for an incremental volume

$$\nabla\cdot\boldsymbol{u} = \lim_{\Delta V\to 0}\frac{1}{\Delta V}\int_{\Delta S} u_n \mathrm{d}S$$

the divergence of a vector can be called the integral of that quantity over the area of a closed volume, divided by the volume. If the vector represents the flow of energy and the divergence is positive at a point *P*, then either a source of energy is present at *P* or energy is leaving the region around *P* so that its temperature is decreasing. If the vector represents the flow of mass and the divergence is positive at a point *P*, then either a source of mass exists at *P* or the density is decreasing at the point *P*. For an incompressible fluid the divergence is zero and the rate at which fluid is introduced into a volume must equal the rate at which it is removed.

Various theorems follow from the divergence theorem.

Theorem. If φ is a solution to Laplace's equation

$$\nabla^2\varphi = 0$$

in a domain D, and the second partial derivatives of φ are continuous in D, then the integral of the normal derivative of φ over any piecewise smooth closed orientable surface *S* in D is zero. Suppose $\boldsymbol{u} = \varphi\,\nabla\psi$ satisfies the conditions of the divergence theorem: then *Green's theorem* results from use of the divergence theorem [49].

$$\int_V \left(\varphi\nabla^2\psi + \nabla\varphi\cdot\nabla\psi\right)\mathrm{d}V = \int_S \varphi\frac{\partial\psi}{\partial n}\mathrm{d}S$$

and

$$\int_V \left(\varphi\nabla^2\psi - \psi\nabla^2\varphi\right)\mathrm{d}V = \int_S \left(\varphi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\varphi}{\partial n}\right)\mathrm{d}S$$

Also if φ satisfies the conditions of the theorem and is zero on *S* then φ is zero throughout D. If two functions φ and ψ both satisfy the Laplace equation in domain D, and both take the same values on the bounding curve *C*, then $\varphi = \psi$; i.e., the solution to the Laplace equation is unique.

The divergence theorem for dyadics is

$$\int_V \nabla\cdot\boldsymbol{\tau}\mathrm{d}V = \int_S \boldsymbol{n}\cdot\boldsymbol{\tau}\,\mathrm{d}S$$

Stokes Theorem [48, 49]. Stokes theorem says that if *S* is a surface bounded by a closed, nonintersecting curve *C*, and if $\boldsymbol{u}$ has continuous derivatives then

$$\oint_C \boldsymbol{u}\cdot\mathrm{d}\boldsymbol{r} = \iint_S (\nabla\times\boldsymbol{u})\cdot\boldsymbol{n}\mathrm{d}S = \iint_S (\nabla\times\boldsymbol{u})\cdot\mathrm{d}\boldsymbol{S}$$

The integral around the curve is followed in the counterclockwise direction. In component notation, this is

$$\oint_C \left[u_x\cos\,(x,s) + u_y\cos\,(y,s) + u_z\cos\,(z,s)\right]\mathrm{d}s =$$
$$\iint_S \left[\left(\frac{\partial u_z}{\partial y} - \frac{\partial u_y}{\partial z}\right)\cos(x,n)\right.$$
$$+\left(\frac{\partial u_x}{\partial z} - \frac{\partial u_z}{\partial x}\right)\cos\,(y,n)$$
$$\left.+\left(\frac{\partial u_y}{\partial x} - \frac{\partial u_x}{\partial y}\right)\cos(z,n)\right]\mathrm{d}S$$

Applied in two dimensions, this results in Green's theorem in the plane:

$$\oint_C (M\mathrm{d}x + N\mathrm{d}y) = \iint_S \left(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\right)\mathrm{d}x\mathrm{d}y$$

The formula for dyadics is

$$\iint_S \boldsymbol{n}\cdot(\nabla\times\boldsymbol{\tau})\,\mathrm{d}S = \oint_C \boldsymbol{\tau}^{\mathrm{T}}\cdot\mathrm{d}\boldsymbol{r}$$

Representation. Two theorems give information about how to represent vectors that obey certain properties.

Theorem [48*, p. 422].* The necessary and sufficient condition that the curl of a vector vanish identically is that the vector be the gradient of some function.

Theorem [48*, p. 423].* The necessary and sufficient condition that the divergence of a vector vanish identically is that the vector is the curl of some other vector.

Leibniz Formula. In fluid mechanics and transport phenomena, an important result is the derivative of an integral whose limits of integration are moving. Suppose the region *V* (*t*) is moving with velocity $\boldsymbol{v}_s$. Then Leibniz's rule holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}\iiint_{V(t)} \varphi\mathrm{d}V = \iiint_{V(t)} \frac{\partial\varphi}{\partial t}\mathrm{d}V + \iint_S \varphi\boldsymbol{v}_s\cdot\boldsymbol{n}\mathrm{d}S$$

Curvilinear Coordinates. Many of the relations given above are proved most easily by using tensor analysis rather than dyadics. Once proven, however, the relations are perfectly general in any coordinate system. Displayed here are the specific results for cylindrical and spherical geometries. Results are available for a few other geometries: parabolic cylindrical, paraboloidal, elliptic cylindrical, prolate spheroidal, oblate spheroidal, ellipsoidal, and bipolar coordinates [45, 50].

For *cylindrical coordinates*, the geometry is shown in Figure 16. The coordinates are related to cartesian coordinates by

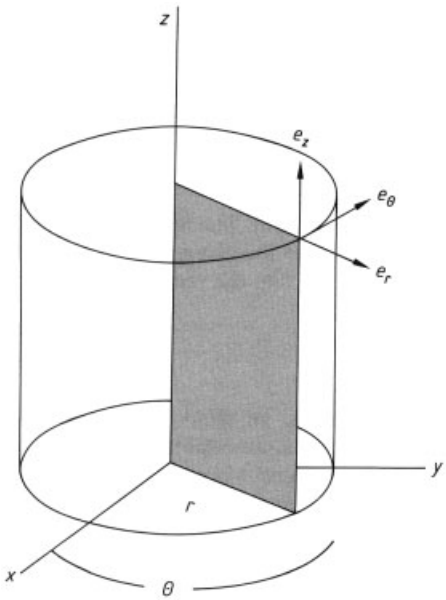

Figure 16. Cylindrical coordinate system

$x = r\cos\theta$	$r = \sqrt{x^2+y^2}$
$y = r\sin\theta$	$\theta = \arctan\left(\frac{y}{x}\right)$
$z = z$	$z = z$

The unit vectors are related by

$\boldsymbol{e}_r = \cos\theta\boldsymbol{e}_x + \sin\theta\boldsymbol{e}_y$	$\boldsymbol{e}_x = \cos\theta\boldsymbol{e}_r - \sin\theta\boldsymbol{e}_\theta$
$\boldsymbol{e}_\theta = -\sin\theta\boldsymbol{e}_x + \cos\theta\boldsymbol{e}_y$	$\boldsymbol{e}_y = \sin\theta\boldsymbol{e}_r + \cos\theta\boldsymbol{e}_\theta$
$\boldsymbol{e}_z = \boldsymbol{e}_z$	$\boldsymbol{e}_z = \boldsymbol{e}_z$

Derivatives of the unit vectors are

$$\mathrm{d}\boldsymbol{e}_\theta = -\boldsymbol{e}_r\mathrm{d}\theta, \mathrm{d}\boldsymbol{e}_r = \boldsymbol{e}_\theta\mathrm{d}\theta, \mathrm{d}\boldsymbol{e}_z = 0$$

Differential operators are given by [45]

$$\boldsymbol{\nabla} = \boldsymbol{e}_r\frac{\partial}{\partial r} + \frac{\boldsymbol{e}_\theta}{r}\frac{\partial}{\partial\theta} + \boldsymbol{e}_z\frac{\partial}{\partial z},$$

$$\boldsymbol{\nabla}\varphi = \boldsymbol{e}_r\frac{\partial\varphi}{\partial r} + \frac{\boldsymbol{e}_\theta}{r}\frac{\partial\varphi}{\partial\theta} + \boldsymbol{e}_z\frac{\partial\varphi}{\partial z}$$

$$\boldsymbol{\nabla}^2\varphi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\varphi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\varphi}{\partial\theta^2} + \frac{\partial^2\varphi}{\partial z^2}$$

$$= \frac{\partial^2\varphi}{\partial r^2} + \frac{1}{r}\frac{\partial\varphi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\varphi}{\partial\theta^2} + \frac{\partial^2\varphi}{\partial z^2}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{v} = \frac{1}{r}\frac{\partial}{\partial r}(r\,v_r) + \frac{1}{r}\frac{\partial v_\theta}{\partial\theta} + \frac{\partial v_z}{\partial z}$$

$$\boldsymbol{\nabla}\times\boldsymbol{v} = \boldsymbol{e}_r\left(\frac{1}{r}\frac{\partial v_z}{\partial\theta} - \frac{\partial v_\theta}{\partial z}\right) + \boldsymbol{e}_\theta\left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)$$

$$+\boldsymbol{e}_z\left[\frac{1}{r}\frac{\partial}{\partial r}(r\,v_\theta) - \frac{1}{r}\frac{\partial v_r}{\partial\theta}\right]$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{\tau} = \boldsymbol{e}_r\left[\frac{1}{r}\frac{\partial}{\partial r}(r\tau_{rr}) + \frac{1}{r}\frac{\partial\tau_{\theta r}}{\partial\theta} + \frac{\partial\tau_{zr}}{\partial z} - \frac{\tau_{\theta\theta}}{r}\right] +$$

$$+\boldsymbol{e}_\theta\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\tau_{r\theta}\right) + \frac{1}{r}\frac{\partial\tau_{\theta\theta}}{\partial\theta} + \frac{\partial\tau_{z\theta}}{\partial z} + \frac{\tau_{\theta r}-\tau_{r\theta}}{r}\right] +$$

$$+\boldsymbol{e}_z\left[\frac{1}{r}\frac{\partial}{\partial r}(r\tau_{rz}) + \frac{1}{r}\frac{\partial\tau_{\theta z}}{\partial\theta} + \frac{\partial\tau_{zz}}{\partial z}\right]$$

$$\boldsymbol{\nabla}\boldsymbol{v} = \boldsymbol{e}_r\boldsymbol{e}_r\frac{\partial v_r}{\partial r} + \boldsymbol{e}_r\boldsymbol{e}_\theta\frac{\partial v_\theta}{\partial r} + \boldsymbol{e}_r\boldsymbol{e}_z\frac{\partial v_z}{\partial r} +$$

$$+\boldsymbol{e}_\theta\boldsymbol{e}_r\left(\frac{1}{r}\frac{\partial v_r}{\partial\theta} - \frac{v_\theta}{r}\right) + \boldsymbol{e}_\theta\boldsymbol{e}_\theta\left(\frac{1}{r}\frac{\partial v_\theta}{\partial\theta} + \frac{v_r}{r}\right) +$$

$$+\boldsymbol{e}_\theta\boldsymbol{e}_z\frac{1}{r}\frac{\partial v_z}{\partial\theta} + \boldsymbol{e}_z\boldsymbol{e}_r\frac{\partial v_r}{\partial z} + \boldsymbol{e}_z\boldsymbol{e}_\theta\frac{\partial v_\theta}{\partial z} + \boldsymbol{e}_z\boldsymbol{e}_z\frac{\partial v_z}{\partial z}$$

$$\nabla^2\boldsymbol{v} = \boldsymbol{e}_r\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(r\,v_r)\right) + \frac{1}{r^2}\frac{\partial^2 v_r}{\partial\theta^2} + \frac{\partial^2 v_r}{\partial z^2}\right.$$

$$\left.-\frac{2}{r^2}\frac{\partial v_\theta}{\partial\theta}\right] +$$

$$+\boldsymbol{e}_\theta\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(r\,v_\theta)\right) + \frac{1}{r^2}\frac{\partial^2 v_\theta}{\partial\theta^2} + \frac{\partial^2 v_\theta}{\partial z^2} + \frac{2}{r^2}\frac{\partial v_r}{\partial\theta}\right] +$$

$$+\boldsymbol{e}_z\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_z}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 v_z}{\partial\theta^2} + \frac{\partial^2 v_z}{\partial z^2}\right]$$

For *spherical coordinates*, the geometry is shown in Figure 17. The coordinates are related to cartesian coordinates by

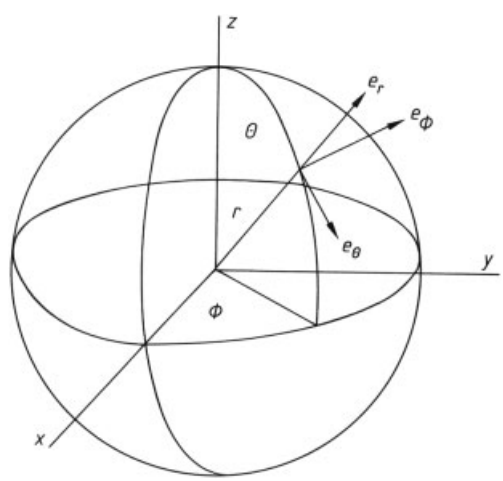

Figure 17. Spherical coordinate system

$x = r\sin\theta\cos\varphi \qquad r = \sqrt{x^2+y^2+z^2}$

$y = r\sin\theta\sin\varphi \qquad \theta = \arctan\left(\sqrt{x^2+y^2}/z\right)$

$z = r\cos\theta \qquad \varphi = \arctan\left(\frac{y}{x}\right)$

The unit vectors are related by

$$\boldsymbol{e}_r = \sin\theta\cos\varphi\boldsymbol{e}_x + \sin\theta\sin\varphi\boldsymbol{e}_y + \cos\theta\boldsymbol{e}_z$$
$$\boldsymbol{e}_\theta = \cos\theta\cos\varphi\boldsymbol{e}_x + \cos\theta\sin\varphi\boldsymbol{e}_y - \sin\theta\boldsymbol{e}_z$$
$$\boldsymbol{e}_\phi = -\sin\varphi\boldsymbol{e}_x + \cos\varphi\boldsymbol{e}_y$$
$$\boldsymbol{e}_x = \sin\theta\cos\varphi\boldsymbol{e}_r + \cos\theta\cos\varphi\boldsymbol{e}_\theta - \sin\varphi\boldsymbol{e}_\phi$$
$$\boldsymbol{e}_y = \sin\theta\sin\varphi\boldsymbol{e}_r + \cos\theta\sin\varphi\boldsymbol{e}_\theta + \cos\varphi\boldsymbol{e}_\phi$$
$$\boldsymbol{e}_z = \cos\theta\boldsymbol{e}_r - \sin\theta\boldsymbol{e}_\theta$$

Derivatives of the unit vectors are

$$\frac{\partial\boldsymbol{e}_r}{\partial\theta} = \boldsymbol{e}_\theta,\ \frac{\partial\boldsymbol{e}_\theta}{\partial\theta} = -\boldsymbol{e}_r$$
$$\frac{\partial\boldsymbol{e}_r}{\partial\varphi} = e_\phi\sin\theta,\ \frac{\partial\boldsymbol{e}_\theta}{\partial\varphi} = \boldsymbol{e}_\phi\cos\theta,$$
$$\frac{\partial\boldsymbol{e}_\phi}{\partial\varphi} = -\boldsymbol{e}_r\sin\theta - \boldsymbol{e}_\theta\cos\theta$$

Others 0

Differential operators are given by [45]

$$\boldsymbol{\nabla} = \boldsymbol{e}_r\frac{\partial}{\partial r} + \boldsymbol{e}_\theta\frac{1}{r}\frac{\partial}{\partial\theta} + \boldsymbol{e}_\phi\frac{1}{r\sin\theta}\frac{\partial}{\partial\varphi},$$

$$\boldsymbol{\nabla}\psi = \boldsymbol{e}_r\frac{\partial\psi}{\partial r} + \boldsymbol{e}_\theta\frac{1}{r}\frac{\partial\psi}{\partial\theta} + \boldsymbol{e}_\phi\frac{1}{r\sin\theta}\frac{\partial\psi}{\partial\varphi}$$

$$\boldsymbol{\nabla}^2\psi = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + $$
$$+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\varphi^2}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{v} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 v_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(v_\theta\sin\theta\right) + $$
$$+\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\varphi}$$

$$\boldsymbol{\nabla}\times\boldsymbol{v} = \boldsymbol{e}_r\left[\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(v_\phi\sin\theta\right) - \frac{1}{r\sin\theta}\frac{\partial v_\theta}{\partial\varphi}\right] + $$
$$+\boldsymbol{e}_\theta\left[\frac{1}{r\sin\theta}\frac{\partial v_r}{\partial\varphi} - \frac{1}{r}\frac{\partial}{\partial r}\left(r\,v_\phi\right)\right] + $$
$$+\boldsymbol{e}_\phi\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\,v_\theta\right) - \frac{1}{r}\frac{\partial v_r}{\partial\theta}\right]$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{\tau} = \boldsymbol{e}_r\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\tau_{rr}\right)\right.$$
$$\left.+\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\tau_{\theta r}\sin\theta\right) + \frac{1}{r\sin\theta}\frac{\partial\tau_{\phi r}}{\partial\varphi} - \frac{\tau_{\theta\theta}+\tau_{\phi\phi}}{r}\right] + $$
$$+\boldsymbol{e}_\theta\left[\frac{1}{r^3}\frac{\partial}{\partial r}\left(r^3\tau_{r\theta}\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\tau_{\theta\theta}\sin\theta\right) + \right.$$
$$\left.\frac{1}{r\sin\theta}\frac{\partial\tau_{\phi\theta}}{\partial\varphi} + \frac{\tau_{\theta r}-\tau_{r\theta}-\tau_{\phi\phi}\cot\theta}{r}\right] + $$
$$+\boldsymbol{e}_\phi\left[\frac{1}{r^3}\frac{\partial}{\partial r}\left(r^3\tau_{r\phi}\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(\tau_{\theta\phi}\sin\theta\right) + \right.$$
$$\left.\frac{1}{r\sin\theta}\frac{\partial\tau_{\phi\phi}}{\partial\varphi} + \frac{\tau_{\phi r}-\tau_{r\phi}+\tau_{\phi\theta}\cot\theta}{r}\right]$$

$$\boldsymbol{\nabla}\boldsymbol{v} = \boldsymbol{e}_r\boldsymbol{e}_r\frac{\partial v_r}{\partial r} + \boldsymbol{e}_r\boldsymbol{e}_\theta\frac{\partial v_\theta}{\partial r} + \boldsymbol{e}_r\boldsymbol{e}_\phi\frac{\partial v_\phi}{\partial r} + $$
$$+\boldsymbol{e}_\theta\boldsymbol{e}_r\left(\frac{1}{r}\frac{\partial v_r}{\partial\theta} - \frac{v_\theta}{r}\right) + \boldsymbol{e}_\theta\boldsymbol{e}_\theta\left(\frac{1}{r}\frac{\partial v_\theta}{\partial\theta} + \frac{v_r}{r}\right) + $$
$$+\boldsymbol{e}_\theta\boldsymbol{e}_\phi\frac{1}{r}\frac{\partial v_\phi}{\partial\theta} + \boldsymbol{e}_\phi\boldsymbol{e}_r\left(\frac{1}{r\sin\theta}\frac{\partial v_r}{\partial\varphi} - \frac{v_\phi}{r}\right) + $$
$$+\boldsymbol{e}_\phi\boldsymbol{e}_\theta\left(\frac{1}{r\sin\theta}\frac{\partial v_\theta}{\partial\varphi} - \frac{v_\phi}{r}\cot\theta\right) + $$
$$+\boldsymbol{e}_\phi\boldsymbol{e}_\phi\left(\frac{1}{r\sin\theta}\frac{\partial v_\phi}{\partial\varphi} + \frac{v_r}{r} + \frac{v_\theta}{r}\cot\theta\right)$$

$$\nabla^2\boldsymbol{v} = \boldsymbol{e}_r\left[\frac{\partial}{\partial r}\left(\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\,v_r\right)\right)\right.$$
$$+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial v_r}{\partial\theta}\right)$$
$$\left.+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 v_r}{\partial\varphi^2} - \frac{2}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(v_\theta\sin\theta\right) - \frac{2}{r^2\sin\theta}\frac{\partial v_\phi}{\partial\varphi}\right]$$
$$+\boldsymbol{e}_\theta\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial v_\theta}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial\theta}\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(v_\theta\sin\theta\right)\right)\right.$$
$$\left.+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 v_\theta}{\partial\varphi^2} + \frac{2}{r^2}\frac{\partial v_r}{\partial\theta} - \frac{2}{r^2}\frac{\cot\theta}{\sin\theta}\frac{\partial v_\phi}{\partial\varphi}\right]$$
$$+\boldsymbol{e}_\phi\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial v_\phi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial\theta}\left(\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(v_\phi\sin\theta\right)\right)\right.$$
$$\left.+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 v_\phi}{\partial\varphi^2} + \frac{2}{r^2\sin\theta}\frac{\partial v_r}{\partial\varphi} + \frac{2}{r^2}\frac{\cot\theta}{\sin\theta}\frac{\partial v_\theta}{\partial\varphi}\right]$$

6. Ordinary Differential Equations as Initial Value Problems

A differential equation for a function that depends on only one variable (often time) is called an ordinary differential equation. The general solution to the differential equation includes many possibilities; the boundary or initial conditions are required to specify which of those are desired. If all conditions are at one point, the problem is an initial value problem and can be integrated from that point on. If some of the conditions are available at one point and others at another point, the ordinary differential equations become two-point boundary value problems, which are treated in Chapter 7. Initial value problems as ordinary differential equations arise in control of lumped-parameter models, transient models of stirred tank reactors, polymerization reactions and plug-flow reactors, and generally in models where no spatial gradients occur in the unknowns.

6.1. Solution by Quadrature

When only one equation exists, even if it is nonlinear, solving it by quadrature may be possible. For

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\ (y)$$

$$y\ (0) = y_0$$

the problem can be separated

$$\frac{\mathrm{d}y}{f\ (y)} = \mathrm{d}t$$

and integrated:

$$\int_{y_0}^{y} \frac{\mathrm{d}y'}{f\ (y')} = \int_0^t \mathrm{d}t = t$$

If the quadrature can be performed analytically then the exact solution has been found.

For example, consider the kinetics problem with a second-order reaction.

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -kc^2, c\ (0) = c_0$$

To find the function of the concentration versus time, the variables can be separated and integrated.

$$\frac{\mathrm{d}c}{c^2} = -k\mathrm{d}t,$$

$$-\frac{1}{c} = -kt+D$$

Application of the initial conditions gives the solution:

$$\frac{1}{c} = kt+\frac{1}{c_0}$$

For other ordinary differential equations an integrating factor is useful. Consider the problem governing a stirred tank with entering fluid having concentration c_{in} and flow rate F, as shown in Figure 18. The flow rate out is also F and the volume of the tank is V. If the tank is completely mixed, the concentration in the tank is c and the concentration of the fluid leaving the tank is also c. The differential equation is then

$$V\frac{\mathrm{d}c}{\mathrm{d}t} = F\ (c_{\mathrm{in}}-c)\ ,\ c\ (0) = c_0$$

Upon rearrangement,

$$\frac{\mathrm{d}c}{\mathrm{d}t}+\frac{F}{V}c = \frac{F}{V}c_{\mathrm{in}}$$

is obtained. An integrating factor is used to solve this equation. The integrating factor is a function that can be used to turn the left-hand side into an exact differential and can be found by using Fréchet differentials [51]. In this case,

$$\exp\left(\frac{F\,t}{V}\right)\left[\frac{\mathrm{d}c}{\mathrm{d}t}+\frac{F}{V}c\right] = \frac{\mathrm{d}}{\mathrm{d}t}\left[\exp\left(\frac{F\,t}{V}\right)c\right]$$

Thus, the differential equation can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\exp\left(\frac{F\,t}{V}\right)c\right] = \exp\left(\frac{F\,t}{V}\right)\left[\frac{F}{V}c_{\mathrm{in}}\right]$$

This can be integrated once to give

$$\exp\left(\frac{F\,t}{V}\right)c = c\ (0)+\frac{F}{V}\int_0^t \exp\left(\frac{F\,t'}{V}\right)c_{\mathrm{in}}\ (t')\ \mathrm{d}t'$$

or

$$c\ (t) = \exp\left(-\frac{F\,t}{V}\right)c_0$$

$$+\frac{F}{V}\int_0^t \exp\left(-\frac{F\,(t-t')}{V}\right)c_{\mathrm{in}}\ (t')\,\mathrm{d}t'$$

If the integral on the right-hand side can be calculated, the solution can be obtained analytically. If not, the numerical methods described in the next sections can be used. Laplace transforms can also be attempted. However, an analytic solution is so useful that quadrature and an integrating factor should be tried before resorting to numerical solution.

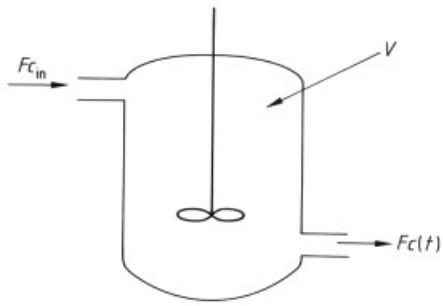

Figure 18. Stirred tank

6.2. Explicit Methods

Consider the ordinary differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\ (y)$$

Multiple equations that are still initial value problems can be handled by using the same techniques discussed here. A higher order differential equation

$$y^{(n)}+F\left(y^{(n-1)}, y^{(n-2)},\ldots, y', y\right) = 0$$

with initial conditions

$$G_i\ \left(y^{(n-1)}\,(0)\,, y^{(n-2)}\ (0)\,,\ldots,y'\ (0)\,, y\ (0)\right) = 0$$

$$i = 1,\ldots,n$$

can be converted into a set of first-order equations. By using

$$y_i \equiv y^{(i-1)} = \frac{\mathrm{d}^{(i-1)}\,y}{\mathrm{d}t^{(i-1)}} = \frac{\mathrm{d}}{\mathrm{d}t}\,y^{(i-2)} = \frac{\mathrm{d}y_{i-1}}{\mathrm{d}t}$$

the higher order equation can be written as a set of first-order equations:

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = y_2$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = y_3$$

$$\frac{\mathrm{d}y_3}{\mathrm{d}t} = y_4$$

$$\ldots$$

$$\frac{\mathrm{d}y_n}{\mathrm{d}t} = -\,F\,(y_{n-1}, y_{n-2},\ldots,y_2,\, y_1)$$

The initial conditions would have to be specified for variables $y_1(0), \ldots, y_n\ (0)$, or equivalently $y\ (0), \ldots, y^{(n-1)}\ (0)$. The set of equations is then written as

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{f}\,(\boldsymbol{y}, t)$$

All the methods in this chapter are described for a single equation; the methods apply to the multiple equations as well. Taking the single equation in the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\ (y)$$

multiplying by dt, and integrating once yields

$$\int_{t_n}^{t_{n+1}} \frac{\mathrm{d}y}{\mathrm{d}t'}\,\mathrm{d}t' = \sum_{t_n}^{t_{n+1}} f\ \left(y\ (t')\right)\mathrm{d}t'$$

This is

$$y^{n+1} = y^n + \int_{t_n}^{t_{n+1}} \frac{\mathrm{d}y}{\mathrm{d}t'}\,\mathrm{d}t'$$

The last substitution gives a basis for the various methods. Different interpolation schemes for $y\ (t)$ provide different integration schemes; using low-order interpolation gives low-order integration schemes [3].

Euler's method is first order

$$y^{n+1} = y^n + \Delta t f\,(y^n)$$

Adams – Bashforth Methods. The *second-order* Adams – Bashforth method is

$$y^{n+1} = y^n + \frac{\Delta\,t}{2}\left[3\,f\ (y^n)\ - f\ \left(y^{n-1}\right)\right]$$

The *fourth-order* Adams – Bashforth method is

$$y^{n+1} = y^n + \frac{\Delta\,t}{24}\left[55\,f\ (y^n)\ - 59\,f\ \left(y^{n-1}\right)\right.$$
$$\left. +37\,f\ \left(y^{n\,-\,2}\right)\ - 9\,f\ \left(y^{n-3}\right)\right]$$

Notice that the higher order explicit methods require knowing the solution (or the right-hand side) evaluated at times in the past. Because these were calculated to get to the current time, this presents no problem except for starting the evaluation. Then, Euler's method may have to be used with a very small step size for several steps to generate starting values at a succession of time points. The error terms, order of the method, function evaluations per step, and stability limitations are listed in Table 6. The advantage of the fourth-order Adams – Bashforth method is that it uses only one function evaluation per step and yet achieves high-order accuracy. The disadvantage is the necessity of using another method to start.

Runge – Kutta Methods. Runge – Kutta methods are explicit methods that use several function evaluations for each time step. The general form of the methods is

$$y^{n+1} = y^n + \sum_{i=1}^{v} w_i\,k_i$$

with

$$k_i = \Delta t\,f\left(t^n + c_i\Delta t, y^n + \sum_{j=1}^{i-1} a_{ij}\,k_j\right)$$

Runge – Kutta methods traditionally have been writen for $f\ (t,\ y)$ and that is done here, too. If these equations are expanded and compared with a Taylor series, restrictions can be placed on the parameters of the method to make it first order, second order, etc. Even so, additional parameters can be chosen. A *second-order* Runge – Kutta method is

$$y^{n+1} = y^n + \frac{\Delta t}{2}\left[f^n + f\ \left(t^n + \Delta t, y^n + \Delta t\,f^{\,n}\right)\right]$$

The midpoint scheme is another second-order Runge – Kutta method:

Table 6. Properties of integration methods for ordinary differential equations

Method	Error term	Order	Function evaluations per step	Stability limit, $\lambda\ \Delta t \leq$
Explicit methods				
Euler	$\frac{h^2}{2}y''$	1	1	2.0
Second-order Adams – Bashforth	$\frac{5}{12}h^3y'''$	2	1	
Fourth-order Adams – Bashforth	$\frac{251}{720}h^5y^{(5)}$	4	1	0.3
Second-order Runge – Kutta (midpoint)		2	2	2.0
Runge – Kutta – Gill		4	4	2.8
Runge – Kutta – Feldberg	$y^{n+1}-z^{n+1}$	5	6	3.0
Predictor – corrector methods				
Second-order Runge – Kutta		2	2	2.0
Adams, fourth-order		2	2	1.3
Implicit methods, stability limit ∞				
Backward Euler		1	many, iterative	∞ *
Trapezoid rule	$-\frac{1}{12}h^3y'''$	2	many, iterative	2 *
Fourth-order Adams – Moulton		4	many, iterative	3 *

* Oscillation limit, $\lambda\ \Delta t \leq$.

$$y^{n+1} = y^n + \Delta t\ f\ \left(t^n + \frac{\Delta t}{2}, y^n + \frac{\Delta t}{2}\ f^n\right)$$

A popular *fourth-order* method is the Runge – Kutta – Gill method with the formulas

$$k_1 = \Delta t\ f\ (t^n, y^n)$$

$$k_2 = \Delta t\ f\ \left(t^n + \tfrac{\Delta t}{2}, y^n + \tfrac{k_1}{2}\right)$$

$$k_3 = \Delta t\ f\ \left(t^n + \tfrac{\Delta t}{2}, y^n + a\,k_1 + b\,k_2\right)$$

$$k_4 = \Delta t\ f\ (t^n + \Delta t, y^n + c\,k_2 + d\,k_3)$$

$$y^{n+1} = y^n + \tfrac{1}{6}\,(k_1 + k_4) + \tfrac{1}{3}\,(b\,k_2 + d\,k_3)$$

$$a = \tfrac{\sqrt{2}-1}{2}, b = \tfrac{2-\sqrt{2}}{2},$$

$$c = -\tfrac{\sqrt{2}}{2}, d = 1 + \tfrac{\sqrt{2}}{2}$$

Another fourth-order Runge – Kutta method is given by the Runge – Kutta – Feldberg formulas [52]; although the method is fourth-order, it achieves fifth-order accuracy. The popular integration package RKF 45 is based on this method.

$$k_1 = \Delta t\ f\ (t^n, y^n)$$

$$k_2 = \Delta t\ f\ \left(t^n + \tfrac{\Delta t}{4}, y^n + \tfrac{k_1}{4}\right)$$

$$k_3 = \Delta t\ f\ \left(t^n + \tfrac{3}{8}\Delta t, y^n + \tfrac{3}{32}k_1 + \tfrac{9}{32}k_2\right)$$

$$k_4 = \Delta t f\ \left(t^n + \tfrac{12}{13}\Delta t, y^n + \tfrac{1932}{2197}k_1 - \tfrac{7200}{2197}k_2 + \tfrac{7296}{2197}k_3\right)$$

$$k_5 = \Delta t f\ \left(t^n + \Delta t, y^n + \tfrac{439}{216}k_1 - 8\,k_2 + \tfrac{3680}{513}k_3 - \tfrac{845}{4104}k_4\right)$$

$$k_6 = \Delta t\ f\ \left(t^n + \tfrac{\Delta t}{2}, y^n - \tfrac{8}{27}k_1 + 2k_2 - \tfrac{3544}{2565}k_3 + \tfrac{1859}{4104}k_4 - \tfrac{11}{40}k_5\right)$$

$$y^{n+1} = y^n + \tfrac{25}{216}k_1 + \tfrac{1408}{2565}k_3 + \tfrac{2197}{4104}k_4 - \tfrac{1}{5}k_5$$

$$z^{n+1} = y^n + \tfrac{16}{135}k_1 + \tfrac{6656}{12\,825}k_3 + \tfrac{28\,561}{56\,430}k_4 - \tfrac{9}{50}k_5 + \tfrac{2}{55}k_6$$

The value of $y^{n+1} - z^{n+1}$ is an estimate of the error in y^{n+1} and can be used in step-size control schemes.

Generally, a high-order method should be used to achieve high accuracy. The Runge – Kutta – Gill method is popular because it is high order and does not require a starting method (as does the fourth-order Adams – Bashforth method). However, it requires four function evaluations per time step, or four times as many as the Adams – Bashforth method. For problems in which the function evaluations are a significant portion of the calculation time this might be important. Given the speed of computers and the widespread availability of desktop computers, the efficiency of a method is most important only for very large problems that are going to be solved many times. For other problems the most important criterion for choosing a method is probably the time the user spends setting up the problem.

The stability of an integration method is best estimated by determining the rational polynomial corresponding to the method. Apply this method to the equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\lambda y, y(0) = 1$$

and determine the formula for r_{mn}:

$$y^{k+1} = r_{mn}(\lambda \Delta t) y^k$$

The rational polynomial is defined as

$$r_{mn}(z) = \frac{p_n(z)}{q_m(z)} \approx e^{-z}$$

and is an approximation to exp $(-z)$, called a Padé approximation. The stability limits are the largest positive z for which

$$|r_{mn}(z)| \leq 1$$

The method is A acceptable if the inequality holds for $Re\, z > 0$. It is $A(0)$ acceptable if the inequality holds for z real, $z > 0$ [53]. The method will not induce oscillations about the true solution provided

$$r_{mn}(z) > 0$$

A method is L acceptable if it is A acceptable and

$$\lim_{z \to \infty} r_{mn}(z) = 0$$

For example, Euler's method gives

$$y^{n+1} = y^n - \lambda \Delta t y^n \text{ or } y^{n+1} = (1 - \lambda \Delta t) y^n$$

$$\text{or } r_{mn} = 1 - \lambda \Delta t$$

The stability limit is then

$$\lambda \Delta t \leq 2$$

The Euler method will not oscillate provided

$$\lambda \Delta t \leq 1$$

The stability limits listed in Table 6 are obtained in this fashion. The limit for the Euler method is 2.0; for the Runge – Kutta – Gill method it is 2.785; for the Runge – Kutta – Feldberg method it is 3.020. The rational polynomials for the various explicit methods are illustrated in Figure 19. As can be seen, the methods approximate the exact solution well as $\lambda \Delta t$ approaches zero, and the higher order methods give a better approximation at high values of $\lambda \Delta t$.

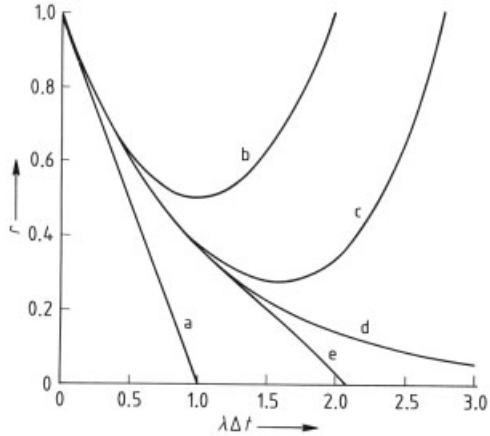

Figure 19. Rational approximations for explicit methods a) Euler; b) Runge – Kutta – 2; c) Runge – Kutta – Gill; d) Exact curve; e) Runge – Kutta – Feldberg

In solving sets of equations

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{f}, \boldsymbol{y}(0) = y_0$$

all the eigenvalues of the matrix $\boldsymbol{A}$ must be examined. Finlayson [3] and Amundson [54, p. 197 – 199] both show how to transform these equations into an orthogonal form so that each equation becomes one equation in one unknown, for which single equation analysis applies. For linear problems the eigenvalues do not change, so the stability and oscillation limits must be satisfied for every eigenvalue of the matrix $\boldsymbol{A}$. When solving nonlinear problems the equations are linearized about the solution at the local time, and the analysis applies for small changes in time, after which a new analysis about the new solution must be made. Thus, for nonlinear problems the eigenvalues keep changing.

Richardson extrapolation can be used to improve the accuracy of a method. Step forward one step Δt with a p-th order method. Then redo the problem, this time stepping forward from the same initial point but in two steps of length $\Delta t/2$, thus ending at the same point. Call the solution of the one-step calculation y_1 and the solution of the two-step calculation y_2. Then an improved solution at the new time is given by

$$y = \frac{2^p y_2 - y_1}{2^p - 1}$$

This gives a good estimate provided Δt is small enough that the method is truly convergent with

order p. This process can also be repeated in the same way Romberg's method was used for quadrature (see Section 2.4).

The accuracy of a numerical calculation depends on the step size used, and this is chosen automatically by efficient codes. For example, in the Euler method the local truncation error LTE is

$$\text{LTE} = \frac{\Delta t^2}{2} y_n''$$

Yet the second derivative can be evaluated by using the difference formulas as

$$y_n'' = \nabla(\Delta t\, y_n') = \Delta t\, (y_n' - y_{n-1}') =$$

$$\Delta t\, (f_n - f_{n-1})$$

Thus, by monitoring the difference between the right-hand side from one time step to another, an estimate of the truncation error is obtained. This error can be reduced by reducing Δt. If the user specifies a criterion for the largest local error estimate, then Δt is reduced to meet that criterion. Also, Δt is increased to as large a value as possible, because this shortens computation time. If the local truncation error has been achieved (and estimated) by using a step size Δt_1

$$\text{LTE} = c\, \Delta t_1^p$$

and the desired error is ε, to be achieved using a step size Δt_2

$$\varepsilon = c\, \Delta t_2^p$$

then the next step size Δt_2 is taken from

$$\frac{\text{LTE}}{\varepsilon} = \left(\frac{\Delta t_1}{\Delta t_2}\right)^p$$

Generally, things should not be changed too often or too drastically. Thus one may choose not to increase Δt by more than a factor (such as 2) or to increase Δt more than once every so many steps (such as 5) [55]. In the most sophisticated codes the alternative exists to change the order of the method as well. In this case, the truncation error of the orders one higher and one lower than the current one are estimated, and a choice is made depending on the expected step size and work.

6.3. Implicit Methods

By using different interpolation formulas, involving y^{n+1}, implicit integration methods can be derived. Implicit methods result in a nonlinear equation to be solved for y^{n+1} so that iterative methods must be used. The backward Euler method is a first-order method:

$$y^{n+1} = y^n + \Delta t\, f\, (y^{n+1})$$

The trapezoid rule (see Section 2.4) is a second-order method:

$$y^{n+1} = y^n + \frac{\Delta t}{2} \left[f\, (y^n) + f\, (y^{n+1}) \right]$$

When the trapezoid rule is used with the finite difference method for solving partial differential equations it is called the Crank – Nicolson method. Adams methods exist as well, and the fourth-order Adams – Moulton method is

$$y^{n+1} = y^n + \frac{\Delta t}{24} \left[9 f\, (y^{n+1}) + 19 f\, (y^n) - 5 f\, (y^{n-1}) + f\, (y^{n-2}) \right]$$

The properties of these methods are given in Table 6. The implicit methods are stable for any step size but do require the solution of a set of nonlinear equations, which must be solved iteratively. An application to dynamic distillation problems is given in [56].

All these methods can be written in the form

$$y^{n+1} = \sum_{i=1}^{k} \alpha_i\, y^{n+1-i} + \Delta t \sum_{i=0}^{k} \beta_i\, f\, (y^{n+1-i})$$

or

$$y^{n+1} = \Delta t\, \beta_0\, f\, (y^{n+1}) + w^n$$

where w^n represents known information. This equation (or set of equations for more than one differential equation) can be solved by using successive substitution:

$$y^{n+1,\, k+1} = \Delta t\, \beta_0 f \left(y^{n+1,\, k} \right) + w^n$$

Here, the superscript k refers to an iteration counter. The successive substitution method is guaranteed to converge, provided the first derivative of the function is bounded and a small enough time step is chosen. Thus, if it has not converged within a few iterations, Δt can be reduced and the iterations begun again. The *Newton – Raphson method* (see Section 1.2) can also be used.

In many computer codes, iteration is allowed to proceed only a fixed number of times (e.g.,

three) before Δt is reduced. Because a good history of the function is available from previous time steps, a good initial guess is usually possible.

The best software packages for stiff equations (see Section 6.4) use Gear's backward difference formulas. The formulas of various orders are [57].

$$1: y^{n+1} = y^n + \Delta t\, f\left(y^{n+1}\right)$$

$$2: y^{n+1} = \tfrac{4}{3}\, y^n - \tfrac{1}{3}\, y^{n-1} + \tfrac{2}{3}\, \Delta t\, f\left(y^{n+1}\right)$$

$$3: y^{n+1} = \tfrac{18}{11}\, y^n - \tfrac{9}{11}\, y^{n-1} + \tfrac{2}{11}\, y^{n-2} + \tfrac{6}{11}\, \Delta t\, f\left(y^{n+1}\right)$$

$$4: y^{n+1} = \tfrac{48}{25}\, y^n - \tfrac{36}{25}\, y^{n-1} + \tfrac{16}{25}\, y^{n-2} - \tfrac{3}{25}\, y^{n-3} + \tfrac{12}{25}\, \Delta t\, f\left(y^{n+1}\right)$$

$$5: y^{n+1} = \tfrac{300}{137}\, y^n - \tfrac{300}{137}\, y^{n-1} + \tfrac{200}{137}\, y^{n-2} - \tfrac{75}{137}\, y^{n-3} + \tfrac{12}{137}\, y^{n-4} + \tfrac{60}{137}\, \Delta t\, f\left(y^{n+1}\right)$$

The stability properties of these methods are determined in the same way as explicit methods. They are always expected to be stable, no matter what the value of Δt is, and this is confirmed in Figure 20.

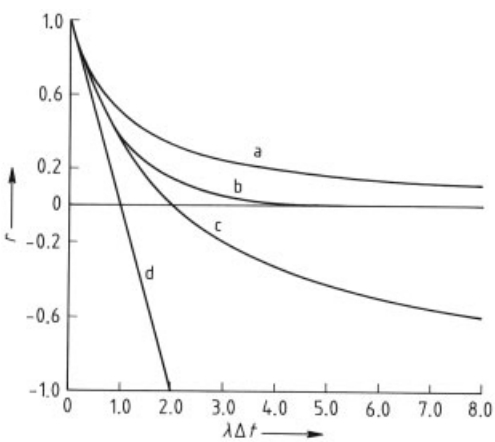

Figure 20. Rational approximations for implicit methods a) Backward Euler; b) Exact curve; c) Trapezoid; d) Euler

Predictor – corrector methods can be employed in which an explicit method is used to predict the value of y^{n+1}. This value is then used in an implicit method to evaluate $f\,(\,y^{n+1})$.

6.4. Stiffness

Why is it desirable to use implicit methods that lead to sets of algebraic equations that must be solved iteratively whereas explicit methods lead to a direct calculation? The reason lies in the stability limits; to understand their impact, the concept of stiffness is necessary. When modeling a physical situation, the time constants governing different phenomena should be examined. Consider flow through a packed bed, as illustrated in Figure 21.

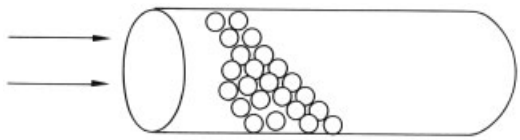

Figure 21. Flow through packed bed

The superficial velocity u is given by

$$u = \frac{Q}{A\,\varphi}$$

where Q is the volumetric flow rate, A is the cross-sectional area, and φ is the void fraction. A time constant for flow through the device is then

$$t_{\text{flow}} = \frac{L}{u} = \frac{\varphi\, A\, L}{Q}$$

where L is the length of the packed bed. If a chemical reaction occurs, with a reaction rate given by

$$\frac{\text{Moles}}{\text{Volume time}} = -k\,c$$

where k is the rate constant (time^{-1}) and c is the concentration (moles/volume), the characteristic time for the reaction is

$$t_{\text{rxn}} = \frac{1}{k}$$

If diffusion occurs inside the catalyst, the time constant is

$$t_{\text{internal diffusion}} = \frac{\varepsilon\, R^2}{D_{\text{e}}}$$

where ε is the porosity of the catalyst, R is the catalyst radius, and D_{e} is the effective diffusion

coefficient inside the catalyst. The time constant for heat transfer is

$$t_{\text{internal heat transfer}} = \frac{R^2}{\alpha} = \frac{\varrho_{\mathrm{s}} C_{\mathrm{s}} R^2}{k_{\mathrm{e}}}$$

where ϱ_{s} is the catalyst density, C_{s} is the catalyst heat capacity per unit mass, k_{e} is the effective thermal conductivity of the catalyst, and α is the thermal diffusivity. The time constants for diffusion of mass and heat through a boundary layer surrounding the catalyst are

$$t_{\text{external diffusion}} = \frac{R}{k_{\mathrm{g}}}$$

$$t_{\text{external heat transfer}} = \frac{\varrho_{\mathrm{s}} C_{\mathrm{s}} R}{h_{\mathrm{p}}}$$

where k_{g} and h_{p} are the mass-transfer and heat-transfer coefficients, respectively. The importance of examining these time constants comes from realization that their orders of magnitude differ greatly. For example, in the model of an automobile catalytic converter [58] the time constant for internal diffusion was 0.3 s, for internal heat transfer 21 s, and for flow through the device 0.003 s. Flow through the device is so fast that it might as well be instantaneous. Thus, the time derivatives could be dropped from the mass balance equations for the flow, leading to a set of differential-algebraic equations (see below). If the original equations had to be solved, the eigenvalues would be roughly proportional to the inverse of the time constants. The time interval over which to integrate would be a small number (e.g., five) multiplied by the longest time constant. Yet the explicit stability limitation applies to all the eigenvalues, and the largest eigenvalue would determine the largest permissible time step. Here, $\lambda = 1/0.003\ \mathrm{s}^{-1}$. Very small time steps would have to be used, e.g., $\Delta t \leq 2 \times 0.003$ s, but a long integration would be required to reach steady state. Such problems are termed stiff, and implicit methods are very useful for them. In that case the stable time constant is not of any interest, because any time step is stable. What is of interest is the largest step for which a solution can be found. If a time step larger than the smallest time constant is used, then any phenomena represented by that smallest time constant will be overlooked—at least transients in it will be smeared over. However, the method will still be stable. Thus, if the very rapid transients of part of the model are not of interest, they can be ignored and an implicit method used [59].

The idea of stiffness is best explained by considering a system of linear equations:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{A}\,\boldsymbol{y}$$

Let λ_i be the eigenvalues of the matrix $\boldsymbol{A}$. This system can be converted into a system of n equations, each of them having only one unknown; the eigenvalues of the new system are the same as the eigenvalues of the original system [3, pp. 39 – 42], [54, pp. 197 – 199]. Then the stiffness ratio SR is defined as [53, p. 32]

$$\mathrm{SR} = \frac{\max_i |\mathrm{Re}\,(\lambda_i)|}{\min_i |\mathrm{Re}\,(\lambda_i)|}$$

SR = 20 is not stiff, SR = 10^3 is stiff, and SR = 10^6 is very stiff. If the problem is nonlinear, the solution is expanded about the current state:

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = f_i\left[\boldsymbol{y}\left(t^n\right)\right] + \sum_{j=1}^{n} \frac{\partial f_i}{\partial y_j}\left[y_j - y_j\left(t^n\right)\right]$$

The question of stiffness then depends on the eigenvalue of the Jacobian at the current time. Consequently, for nonlinear problems the problem can be stiff during one time period and not stiff during another. Packages have been developed for problems such as these. Although the chemical engineer may not actually calculate the eigenvalues, knowing that they determine the stability and accuracy of the numerical scheme, as well as the step size employed, is useful.

6.5. Differential – Algebraic Systems

Sometimes models involve ordinary differential equations subject to some algebraic constraints. For example, the equations governing one equilibrium stage (as in a distillation column) are

$$M\frac{\mathrm{d}x^n}{\mathrm{d}t} = V^{n+1}\,y^{n+1} - L^n\,x^n - V^n\,y^n + L^{n-1}\,x^{n-1}$$

$$x^{n-1} - x^n = E^n\left(x^{n-1} - x^{*,n}\right)$$

$$\sum_{i=1}^{N} x_i = 1$$

where x and y are the mole fractions in the liquid and vapor, respectively; L and V are liquid and vapor flow rates, respectively; M is the holdup; and the superscript n is the stage number. The efficiency is E, and the concentration in equilibrium with the vapor is x^*. The first equation is

an ordinary differential equation for the mass of one component on the stage, whereas the third equation represents a constraint that the mass fractions add to one. As a second example, the following kinetics problem can be considered:

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = f\ (c_1, c_2)$$

$$\frac{\mathrm{d}c_2}{\mathrm{d}t} = k_1\, c_1 - k_2\, c_2^2$$

The first equation could be the equation for a stirred tank reactor, for example. Suppose both k_1 and k_2 are large. The problem is then stiff, but the second equation could be taken at equilibrium. If

$$c_1 \rightleftarrows 2c_2$$

The equilibrium condition is then

$$\frac{c_2^2}{c_1} = \frac{k_1}{k_2} \equiv K$$

Under these conditions the problem becomes

$$\frac{\mathrm{d}c_1}{\mathrm{d}t} = f\ (c_1, c_2)$$

$$0 = k_l c_1 - k_2 c_2^2$$

Thus, a differential-algebraic system of equations is obtained. In this case, the second equation can be solved and substituted into the first to obtain differential equations, but in the general case that is not possible.

Differential-algebraic equations can be written in the general notation

$$\boldsymbol{F}\left(t, \boldsymbol{y}, \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t}\right) = 0$$

or the variables and equations may be separated according to whether they come primarily from differential [$y(t)$] or algebraic equations [$x(t)$]:

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{f}\ (t, \boldsymbol{y}, \boldsymbol{x}), \boldsymbol{g}\ (t, \boldsymbol{y}, \boldsymbol{x}) = 0$$

Another form is not strictly a differential-algebraic set of equations, but the same principles apply; this form arises frequently when the Galerkin finite element is applied:

$$\boldsymbol{A}\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = \boldsymbol{f}\ (\boldsymbol{y})$$

The computer program DASSL [60, 61] can solve such problems. They can also be solved by writing the differential equation as

$$\frac{d\boldsymbol{y}}{dt} = \boldsymbol{A}^{-1}\boldsymbol{f}\ (\boldsymbol{y})$$

When $\boldsymbol{A}$ is independent of $\boldsymbol{y}$, the inverse (from LU decompositions) need be computed only once.

In actuality, higher order backward-difference Gear methods are used in the computer program DASSL [60, 61].

Differential-algebraic systems are more complicated than differential systems because the solution may not always be defined. Pontelides et al. [62] introduced the term "index" to identify possible problems. The index is defined as the minimum number of times the equations must be differentiated with respect to time to convert the system to a set ofordinary differential equations. These higher derivatives may not exist, and the process places limits on which variables can be given initial values. Sometimes the initial values must be constrained by the algebraic equations [62]. For a differential-algebraic system modeling a distillation tower, the index depends on the specification of pressure for the column [62]. Several chemical engineering examples of differential-algebraic systems and a solution for one involving two-phase flow are given in [63].

6.6. Computer Software

Efficient software packages are widely available for solving ordinary differential equations as initial value problems. In each of the packages the user specifies the differential equation to be solved and a desired error criterion. The package then integrates in time and adjusts the step size to achieve the error criterion, within the limitations imposed by stability.

A popular explicit Runge – Kutta package is RKF 45. An estimate of the truncation error at each step is available. Then the step size can be reduced until this estimate is below the user-specified tolerance. The method is thus automatic, and the user is assured of the results. Note, however, that the tolerance is set on the local truncation error, namely, from one step to another, whereas the user is generally interested in the global trunction error, i.e., the error after several steps. The global error is generally made smaller by making the tolerance smaller, but the absolute accuracy is not the same as the tolerance. If the problem is stiff, then very small step sizes are used and the computation becomes

very lengthy. The RKF 45 code discovers this and returns control to the user with a message indicating the problem is too hard to solve with RKF 45.

A popular implicit package is LSODE, a version of Gear's method [57] written by Alan Hindmarsh at Lawrence Livermore Laboratory. In this package, the user specifies the differential equation to be solved and the tolerance desired. Now the method is implicit and, therefore, stable for any step size. The accuracy may not be acceptable, however, and sets of nonlinear equations must be solved. Thus, in practice, the step size is limited but not nearly so much as in the Runge – Kutta methods. In these packages both the step size and the order of the method are adjusted by the package itself. Suppose a k-th order method is being used. The truncation error is determined by the $(k + 1)$-th order derivative. This is estimated by using difference formulas and the values of the right-hand sides at previous times. An estimate is also made for the k-th and $(k + 2)$-th derivative. Then, the errors in a $(k - 1)$-th order method, a k-th order method, and a $(k + 1)$-th order method can be estimated. Furthermore, the step size required to satisfy the tolerance with each of these methods can be determined. Then the method and step size for the next step that achieves the biggest step can be chosen, with appropriate adjustments due to the different work required for each order. The package generally starts with a very small step size and a first-order method—the backward Euler method. Then it integrates along, adjusting the order up (and later down) depending on the error estimates. The user is thus assured that the local truncation error meets the tolerance. A further difficulty arises because the set of nonlinear equations must be solved. Usually a good guess of the solution is available, because the solution is evolving in time and past history can be extrapolated. Thus, the Newton – Raphson method will usually converge. The package protects itself, though, by only doing a few (i.e., three) iterations. If convergence is not reached within these iterations, the step size is reduced and the calculation is redone for that time step. The convergence theorem for the Newton – Raphson method (Chap. 1) indicates that the method will converge if the step size is small enough. Thus, the method is guaranteed to work. Further economies are possible. The Jacobian needed in the Newton – Raphson method can be fixed over several time steps. Then if the iteration does not converge, the Jacobian can be reevaluated at the current time step. If the iteration still does not converge, then the step size is reduced and a new Jacobian is evaluated. The successive substitution method can also be used—wh ich is even faster, except that it may not converge. However, it too will converge if the time step is small enough.

The Runge – Kutta methods give extremely good accuracy, especially when the step size is kept small for stability reasons. If the problem is stiff, though, backward difference implicit methods must be used. Many chemical reactor problems are stiff, necessitating the use of implicit methods. In the MATLAB suite of ODE solvers, ode45 uses a revision of the RKF45 program, while the ode15s program uses an improved backward difference method. Ref. [64] gives details of the programs in MATLAB. Fortunately, many packages are available. On the NIST web page, http://gams.nist.gov/ choose "problem decision tree", and then "differential and integral equations" to find packages which can be downloaded. On the Netlib web site, http://www.netlib.org/, choose "ode" to find packages which can be downloaded. Using Microsoft Excel to solve ordinary differential equations is cumbersome, except for the simplest problems.

6.7. Stability, Bifurcations, Limit Cycles

In this section, bifurcation theory is discussed in a general way. Some aspects of this subject involve the solution of nonlinear equations; other aspects involve the integration of ordinary differential equations; applications include chaos and fractals as well as unusual operation of some chemical engineering equipment. An excellent introduction to the subject and details needed to apply the methods are given in [65]. For more details of the algorithms described below and a concise survey with some chemical engineering examples, see [66] and [67]. Bifurcation results are closely connected with stability of the steady states, which is essentially a transient phenomenon.

Consider the problem

$$\frac{\partial u}{\partial t} = F\ (u, \lambda)$$

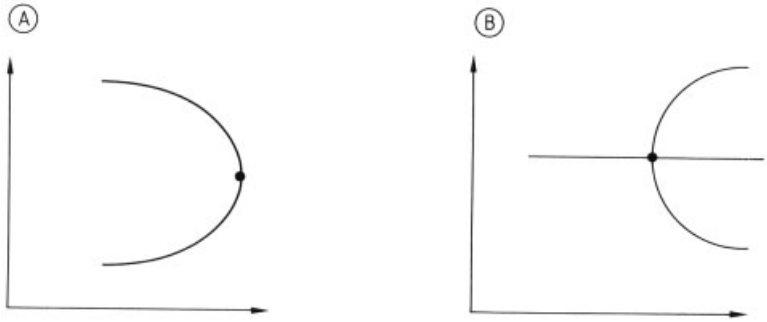

Figure 22. Limit points and bifurcation – limit points
A) Limit point (or turning point); B) Bifurcation-limit point (or singular turning point or bifurcation point)

The variable u can be a vector, which makes F a vector, too. Here, F represents a set of equations that can be solved for the steady state:

$$F\ (u, \lambda) = 0$$

If the Newton – Raphson method is applied,

$$F_u^s\ \delta\ u^s = -F\ (u^s, \lambda)$$

$$u^{s+1} = u^s + \delta u^s$$

is obtained, where

$$F_u^s = \frac{\partial F}{\partial u}\ (u^s)$$

is the Jacobian. Look at some property of the solution, perhaps the value at a certain point or the maximum value or an integral of the solution. This property is plotted versus the parameter λ; typical plots are shown in Figure 22. At the point shown in Figure 22 A, the determinant of the Jacobian is zero:

$$\det F_u = 0$$

For the limit point,

$$\frac{\partial F}{\partial \lambda} \neq 0$$

whereas for the bifurcation-limit point

$$\frac{\partial F}{\partial \lambda} = 0$$

The stability of the steady solutions is also of interest. Suppose a steady solution u_{ss}; the function u is written as the sum of the known steady state and a perturbation u':

$$u = u_{\mathrm{ss}} + u'$$

This expression is substituted into the original equation and linearized about the steady-state value:

$$\frac{\partial u_{\mathrm{ss}}}{\partial t} + \frac{\partial u'}{\partial t} = F\ (u_{\mathrm{ss}} + u', \lambda)$$

$$\approx F\ (u_{\mathrm{ss}}, \lambda) + \frac{\partial F}{\partial u}|_{u_{\mathrm{ss}}} u' + \cdots$$

The result is

$$\frac{\partial u'}{\partial t} = F_u^{\mathrm{ss}}\ u'$$

A solution of the form

$$u'\ (x, t) = \mathrm{e}^{\sigma t}\ X\ (x)$$

gives

$$\sigma \mathrm{e}^{\sigma t}\ X\ = F_u^{\mathrm{ss}} \mathrm{e}^{\sigma t} X$$

The exponential term can be factored out and

$$(F_u^{\mathrm{ss}} - \sigma\ \delta)\ X = 0$$

A solution exists for X if and only if

$$\det\ |F_u^{\mathrm{ss}} - \sigma\ \delta| = 0$$

The σ are the eigenvalues of the Jacobian. Now clearly if Re $(\sigma) > 0$ then u' grows with time, and the steady solution u_{ss} is said to be unstable to small disturbances. If Im $(\sigma) = 0$ it is called stationary instability, and the disturbance would grow monotonically, as indicated in Figure 23 A. If Im $(\sigma) \neq 0$ then the disturbance grows in an oscillatory fashion, as shown in Figure 23 B, and is called oscillatory instability. The case in which Re $(\sigma) = 0$ is the dividing point between stability and instability. If Re $(\sigma) = 0$ and Im $(\sigma) = 0$—the point governing the onset of stationary instability—then $\sigma = 0$. However, this means that $\sigma = 0$ is an eigenvalue of the Jacobian, and the determinant of the Jacobian is zero. Thus, the points at which the determinant of the Jacobian is zero (for limit points

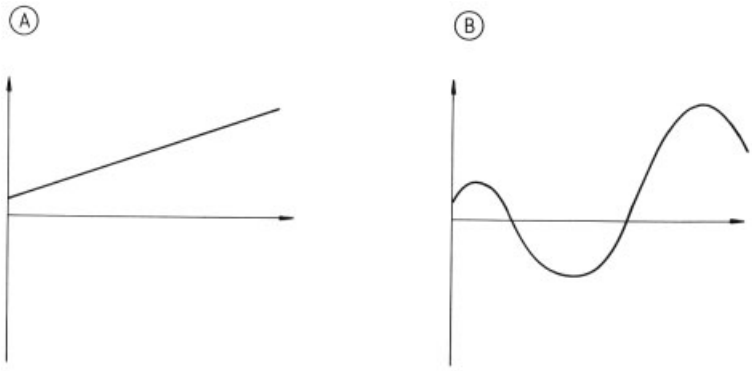

Figure 23. Stationary and oscillatory instability
A) Stationary instability; B) Oscillatory instability

and bifurcation-limit points) are the points governing the onset of stationary instability. When Re (σ) = 0 but Im (σ) $\neq$ 0, which is the onset of oscillatory instability, an even number of eigenvalues pass from the left-hand complex plane to the right-hand complex plane. The eigenvalues are complex conjugates of each other (a result of the original equations being real, with no complex numbers), and this is called a Hopf bifurcation. Numerical methods to study Hopf bifurcation are very computationally intensive and are not discussed here [65].

To return to the problem of solving for the steady-state solution: near the limit point or bifurcation-limit point two solutions exist that are very close to each other. In solving sets of equations with thousands of unknowns, the difficulties in convergence are obvious. For some dependent variables the approximation may be converging to one solution, whereas for another set of dependent variables it may be converging to the other solution; or the two solutions may all be mixed up. Thus, solution is difficult near a bifurcation point, and special methods are required. These methods are discussed in [66].

The first approach is to use *natural continuation* (also known as Euler – Newton continuation). Suppose a solution exists for some parameter λ. Call the value of the parameter λ_0 and the corresponding solution u_0. Then

$$F\ (u_0\ , \lambda_0) = 0$$

Also, compute u_λ as the solution to

$$F_u^{\mathrm{ss}}\ u_\lambda = -F_\lambda$$

at this point [λ_0, u_0]. Then predict the starting guess for another λ using

$$u^0 = u_0 + u_\lambda\ (\lambda - \lambda_0)$$

and apply Newton – Raphson with this initial guess and the new value of λ. This will be a much better guess of the new solution than just u_0 by itself.

Even this method has difficulties, however. Near a limit point the determinant of the Jacobian may be zero and the Newton method may fail. Perhaps no solutions exist at all for the chosen parameter λ near a limit point. Also, the ability to switch from one solution path to another at a bifurcation-limit point is necessary. Thus, other methods are needed as well: arc-length continuation and pseudo-arc-length continuation [66]. These are described in Chapter 1.

6.8. Sensitivity Analysis

Often, when solving differential equations, the solution as well as the sensitivity of the solution to the value of a parameter must be known. Such information is useful in doing parameter estimation (to find the best set of parameters for a model) and in deciding whether a parameter needs to be measured accurately. The differential equation for $y\ (t,\ \alpha)$ where α is a parameter, is

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f\ (y, \alpha)\ ,\ y\ (0) = y_0$$

If this equation is differentiated with respect to α, then because y is a function of t and α

$$\frac{\partial}{\partial \alpha}\left(\frac{\mathrm{d}y}{\mathrm{d}t}\right) = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \alpha} + \frac{\partial f}{\partial \alpha}$$

Exchanging the order of differentiation in the first term leads to the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial y}{\partial \alpha}\right) = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \alpha}+\frac{\partial f}{\partial \alpha}$$

The initial conditions on $\partial y/\partial\alpha$ are obtained by differentiating the initial conditions

$$\frac{\partial}{\partial \alpha}\left[y\ (0,\alpha) = y_0\right], \text{or } \frac{\partial y}{\partial \alpha}\ (0) = 0$$

Next, let

$$y_1 = y, y_2 = \frac{\partial y}{\partial \alpha}$$

and solve the set of ordinary differential equations

$$\frac{\mathrm{d}y_1}{\mathrm{d}t} = f\ (y_1,\alpha) \qquad y_1\ (0) = y_0$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t} = \frac{\partial f}{\partial y}\ (y_1,\alpha)\ y_2+\frac{\partial f}{\partial \alpha} \qquad y_2\ (0) = 0$$

Thus, the solution $y\ (t,\ \alpha)$ and the derivative with respect to α are obtained. To project the impact of α, the solution for $\alpha = \alpha_1$ can be used:

$$y\ (t,\alpha) = y_1\ (t,\alpha_1)+\frac{\partial y}{\partial \alpha}\ (t,\alpha_1)\ (\alpha-\alpha_1)+\cdots$$

$$= y_1\ (t,\alpha_1)+y_2\ (t,\alpha_1)\ (\alpha-\alpha_1)+\cdots$$

This is a convenient way to determine the sensitivity of the solution to parameters in the problem.

6.9. Molecular Dynamics

(→ Molecular Dynamics Simulations)
Special integration methods have been developed for molecular dynamics calculations due to the structure of the equations. A very large number of equations are to be integrated, with the following form based on molecular interactions between molecules

$$m_i\frac{d^2\boldsymbol{r}_i}{dt^2}=\boldsymbol{F}_i\ (\{\boldsymbol{r}\}),\boldsymbol{F}_i\ (\{\boldsymbol{r}\})=-\boldsymbol{\nabla}V$$

where m_i is the mass of the i-th particle, $\boldsymbol{r}_i$ is the position of the i-th particle, $\boldsymbol{F}_i$ is the force acting on the i-th particle, and V is the potential energy that depends upon the location of all the particles (but not their velocities). Since the major part of the calculation is in the evaluation of the forces, or potentials, a method must be used that minimizes the number of times the forces are calculated to move from one time to another time. Rewrite this equation in the form of an acceleration.

$$\frac{d^2\boldsymbol{r}_i}{dt^2}=\frac{1}{m_i}\boldsymbol{F}_i\ (\{\boldsymbol{r}\})\equiv\boldsymbol{a}_i$$

In the Verlot method, this equation is written using central finite differences (Eq. 12). Note that the accelerations do not depend upon the velocities.

$$\boldsymbol{r}_i\ (t+\Delta t)=2\boldsymbol{r}_i\ (t)-\boldsymbol{r}_i\ (t-\Delta t)+\boldsymbol{a}_i\ (t)\ \Delta t^2$$

The calculations are straightforward, and no explicit velocity is needed. The storage requirement is modest, and the precision is modest (it is a second-order method). Note that one must start the calculation with values of $\{r\}$ at time t and $t-\Delta t$.

In the Velocity Verlot method, an equation is written for the velocity, too.

$$\frac{d\boldsymbol{v}_i}{dt}=\boldsymbol{a}_i$$

The trapezoid rule (see page 16) is applied to obtain

$$\boldsymbol{v}_i\ (t+\Delta t)=\boldsymbol{v}_i\ (t)+\frac{1}{2}\left[\boldsymbol{a}_i\ (t)+\boldsymbol{a}_i\ (t+\Delta t)\right]\Delta t$$

The position of the particles is expanded in a Taylor series.

$$\boldsymbol{r}_i\ (t+\Delta t)=\boldsymbol{r}_i\ (t)+\boldsymbol{v}_i\Delta t+\frac{1}{2}\boldsymbol{a}_i\ (t)\ \Delta t^2$$

Beginning with values of $\{\boldsymbol{r}\}$ and $\{\boldsymbol{v}\}$ at time zero, one calculates the new positions and then the new velocities. This method is second order in Δt, too. For additional details, see [68 – 72].

7. Ordinary Differential Equations as Boundary Value Problems

Diffusion problems in one dimension lead to boundary value problems. The boundary conditions are applied at two different spatial locations: at one side the concentration may be fixed and at the other side the flux may be fixed. Because the conditions are specified at two different locations the problems are not initial value in character. To begin at one position and integrate directly is impossible because at least one of the conditions is specified somewhere else and not enough conditions are available to begin the

calculation. Thus, methods have been developed especially for boundary value problems. Examples include heat and mass transfer in a slab, reaction – diffusion problems in a porous catalyst, reactor with axial dispersion, packed beds, and countercurrent heat transfer.

7.1. Solution by Quadrature

When only one equation exists, even if it is nonlinear, it may possibly be solved by quadrature. For

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(y)$$

$$y(0) = y_0$$

the problem can be separated

$$\frac{\mathrm{d}y}{f(y)} = \mathrm{d}t$$

and integrated

$$\int_{y_0}^{y} \frac{dy'}{f(y')} = \int_0^t \mathrm{d}t = t$$

If the quadrature can be performed analytically, the exact solution has been found.

As an example, consider the flow of a non-Newtonian fluid in a pipe, as illustrated in Figure 24. The governing differential equation is [73]

$$\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}(r\tau) = -\frac{\Delta p}{L}$$

where r is the radial position from the center of the pipe, τ is the shear stress, Δp is the pressure drop along the pipe, and L is the length over which the pressure drop occurs. The variables are separated once

$$\mathrm{d}(r\tau) = -\frac{\Delta p}{L} r\mathrm{d}r$$

and then integrated to give

$$r\tau = -\frac{\Delta p}{L}\frac{r^2}{2} + c_1$$

Proceeding further requires choosing a constitutive relation relating the shear stress and the velocity gradient as well as a condition specifying the constant. For a Newtonian fluid

$$\tau = -\eta\frac{\mathrm{d}v}{\mathrm{d}r}$$

where v is the velocity and η the viscosity. Then the variables can be separated again and the result integrated to give

$$-\eta v = -\frac{\Delta p}{L}\frac{r^2}{4} + c_1 \ln r + c_2$$

Now the two unknowns must be specified from the boundary conditions. This problem is a two-point boundary value problem because one of the conditions is usually specified at $r = 0$ and the other at $r = R$, the tube radius. However, the technique of separating variables and integrating works quite well.

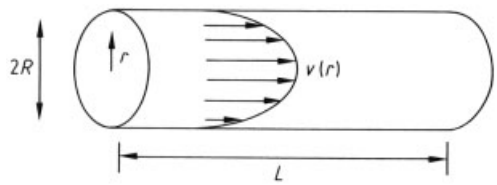

Figure 24. Flow in pipe

When the fluid is non-Newtonian, it may not be possible to do the second step analytically. For example, for the Bird – Carreau fluid [74, p. 171], stress and velocity are related by

$$\tau = \frac{\eta_0}{\left[1+\lambda\left(\frac{\mathrm{d}v}{\mathrm{d}r}\right)^2\right]^{(1-n)/2}}$$

where η_0 is the viscosity at $v = 0$ and λ the time constant.

Putting this value into the equation for stress as a function of r gives

$$\frac{\eta_0}{\left[1+\lambda\left(\frac{\mathrm{d}v}{\mathrm{d}r}\right)^2\right]^{(1-n)/2}} = -\frac{\Delta p}{L}\frac{r}{2} + \frac{c_1}{r}$$

This equation cannot be solved analytically for $\mathrm{d}v/\mathrm{d}r$, except for special values of n. For problems such as this, numerical methods must be used.

7.2. Initial Value Methods

An initial value method is one that utilizes the techniques for initial value problems but allows for an iterative calculation to satisfy all the boundary conditions. Suppose the nonlinear boundary value problem

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} = f\left(x, y, \frac{\mathrm{d}y}{\mathrm{d}x}\right)$$

with the boundary conditions

$$a_0\, y\,(0) - a_1 \tfrac{\mathrm{d}y}{\mathrm{d}x}\,(0) = \alpha,\ a_{\mathrm{i}} \geq 0$$

$$b_0\, y\,(1) - b_1 \tfrac{\mathrm{d}y}{\mathrm{d}x}\,(1) = \beta,\ b_{\mathrm{i}} \geq 0$$

Convert this second-order equation into two first-order equations along with the boundary conditions written to include a parameter s.

$$\tfrac{\mathrm{d}u}{\mathrm{d}x} = v$$

$$\tfrac{\mathrm{d}v}{\mathrm{d}x} = f\,(x, u, v)$$

$$u\,(0) = a_1 s - c_1 \alpha$$

$$v\,(0) = a_0 s - c_0 \alpha$$

The parameters c_0 and c_1 are specified by the analyst such that

$$a_1 c_0 - a_0 c_1 = 1$$

This ensures that the first boundary condition is satisfied for any value of parameter s. If the proper value for s is known, u (0) and u' (0) can be evaluated and the equation integrated as an initial value problem. The parameter s should be chosen iteratively so that the last boundary condition is satisfied.

The model for a *chemical reactor with axial diffusion* is

$$\tfrac{1}{Pe}\tfrac{\mathrm{d}c^2}{\mathrm{d}z^2} - \tfrac{\mathrm{d}c}{\mathrm{d}z} = DaR\,(c)$$

$$-\tfrac{1}{Pe}\tfrac{\mathrm{d}c}{\mathrm{d}z}\,(0) + c\,(0) = c_{\mathrm{in}},\ \tfrac{\mathrm{d}c}{\mathrm{d}z}\,(1) = 0$$

where Pe is the Péclet number and Da the Damköhler number.

The boundary conditions are due to Danckwerts [75] and to Wehner and Wilhelm [76]. This problem can be treated by using initial value methods also, but the method is highly sensitive to the choice of the parameter s, as outlined above. Starting at $z = 0$ and making small changes in s will cause large changes in the solution at the exit, and the boundary condition at the exit may be impossible to satisfy. By starting at $z = 1$, however, and integrating backwards, the process works and an iterative scheme converges in many cases [77]. However, if the problem is extremely nonlinear the iterations may not converge. In such cases, the methods for boundary value problems described below must be used.

Packages to solve boundary value problems are available on the internet. On the NIST web page, http://gams.nist.gov/ choose "problem decision tree", and then "differential and integral equations", then "ordinary differential equations", "multipoint boundary value problems". On the Netlib web site, http://www.netlib.org/, search on "boundary value problem". Any spreadsheet that has an iteration capability can be used with the finite difference method. Some packages for partial differential equations also have a capability for solving one-dimensional boundary value problems [e.g., Comsol Multiphysics (formerly FEMLAB)].

7.3. Finite Difference Method

To apply the finite difference method, we first spread grid points through the domain. Figure 25 shows a uniform mesh of n points (nonuniform meshes are also possible). The unknown, here $c\,(x)$, at a grid point x_i is assigned the symbol $c_i = c\,(x_i)$. The finite difference method can be derived easily by using a Taylor expansion of the solution about this point.

$$\begin{aligned} c_{i+1} &= c_i + \left.\tfrac{\mathrm{d}c}{\mathrm{d}x}\right|_i \Delta x + \left.\tfrac{\mathrm{d}^2 c}{\mathrm{d}x^2}\right|_i \tfrac{\Delta x^2}{2} + \ldots \\ c_{i-1} &= c_i - \left.\tfrac{\mathrm{d}c}{\mathrm{d}x}\right|_i \Delta x + \left.\tfrac{\mathrm{d}^2 c}{\mathrm{d}x^2}\right|_i \tfrac{\Delta x^2}{2} - \ldots \end{aligned} \tag{8}$$

These formulas can be rearranged and divided by Δx to give

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i = \frac{c_{i+1} - c_i}{\Delta x} - \left.\frac{\mathrm{d}^2 c}{\mathrm{d}x^2}\right|_i \frac{\Delta x}{2} + \ldots \tag{9}$$

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i = \frac{c_i - c_{i-1}}{\Delta x} - \left.\frac{\mathrm{d}^2 c}{\mathrm{d}x^2}\right|_i \frac{\Delta x}{2} + \ldots \tag{10}$$

which are representations of the first derivative. Alternatively the two equations can be subtracted from each other, rearranged and divided by Δx to give

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i = \frac{c_{i+1} - c_{i-1}}{2\Delta x} - \left.\frac{\mathrm{d}^2 c}{\mathrm{d}x^3}\right|_i \frac{\Delta x^2}{3!} \tag{11}$$

If the terms multiplied by Δx or Δx^2 are neglected, three representations of the first derivative are possible. In comparison with the Taylor series, the truncation error in the first two expressions is proportional to Δx, and the methods are said to be first order. The truncation error in the last expression is proportional to Δx^2, and the method is said to be second order. Usually, the

last equation is chosen to ensure the best accuracy.

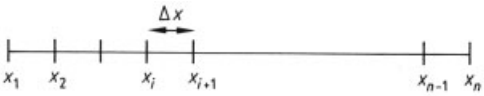

Figure 25. Finite difference mesh; Δx uniform

The finite difference representation of the second derivative can be obtained by adding the two expressions in Equation 8. Rearrangement and division by Δx^2 give

$$\left.\frac{\mathrm{d}^2c}{\mathrm{d}x^2}\right|_i = \frac{c_{i+1}-2_{ci}+c_{i-1}}{\Delta x^2} - \left.\frac{\mathrm{d}^4c}{\mathrm{d}x^4}\right|_i \frac{\Delta x^2}{4!} + \ldots \tag{12}$$

The truncation error is proportional to Δx^2.

To see how to solve a differential equation, consider the equation for convection, diffusion, and reaction in a tubular reactor:

$$\frac{1}{Pe}\frac{\mathrm{d}^2c}{\mathrm{d}x^2} - \frac{\mathrm{d}c}{\mathrm{d}x} = Da\,R\,(c)$$

To evaluate the differential equation at the i-th grid point, the finite difference representations of the first and second derivatives can be used to give

$$\frac{1}{Pe}\frac{c_{i+1}-2c_i+c_{i-1}}{\Delta x^2} - \frac{c_{i+1}-c_{i-1}}{2\Delta x} = Da\,R \tag{13}$$

This equation is written for $i = 2$ to $n - 1$ (i.e., the internal points). The equations would then be coupled but would involve the values of c_1 and c_n, as well. These are determined from the boundary conditions.

If the boundary condition involves a derivative, the finite difference representation of it must be carefully selected; here, three possibilities can be written. Consider a derivative needed at the point $i = 1$. First, Equation 9 could be used to write

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_1 = \frac{c_2-c_1}{\Delta x} \tag{14}$$

Then a second-order expression is obtained that is one-sided. The Taylor series for the point c_{i+2} is written:

$$c_{i+2} = c_i + \left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i 2\Delta x + \left.\frac{\mathrm{d}^2c}{\mathrm{d}x^2}\right|_i \frac{4\Delta x^2}{2!}$$
$$+ \left.\frac{\mathrm{d}^3c}{\mathrm{d}x^3}\right|_i \frac{8\Delta x^3}{3!} + \ldots$$

Four times Equation 8 minus this equation, with rearrangement, gives

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i = \frac{-3c_i+4c_{i+1}-c_{i+2}}{2\Delta x} + O\left(\Delta x^2\right)$$

Thus, for the first derivative at point $i = 1$

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_i = \frac{-3c_i+4c_2-c_3}{2\Delta x} \tag{15}$$

This one-sided difference expression uses only the points already introduced into the domain. The third alternative is to add a false point, outside the domain, as $c_0 = c\ (x = -\Delta x)$. Then the centered first derivative, Equation 11, can be used:

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_1 = \frac{c_2-c_0}{2\Delta x}$$

Because this equation introduces a new variable, another equation is required. This is obtained by also writing the differential equation (Eq. 13), for $i = 1$.

The same approach can be taken at the other end. As a boundary condition, any of three choices can be used:

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_n = \frac{c_n-c_{n-1}}{\Delta x}$$

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_n = \frac{c_{n-2}-4c_{n-1}+3c_n}{2\Delta x}$$

$$\left.\frac{\mathrm{d}c}{\mathrm{d}x}\right|_n = \frac{c_{n+1}-c_{n-1}}{2\Delta x}$$

The last two are of order Δx^2 and the last one would require writing the differential equation (Eq. 13) for $i = n$, too.

Generally, the first-order expression for the boundary condition is not used because the error in the solution would decrease only as Δx, and the higher truncation error of the differential equation (Δx^2) would be lost. For this problem the boundary conditions are

$$-\frac{1}{Pe}\frac{\mathrm{d}c}{\mathrm{d}x}(0) + c(0) = c_{\mathrm{in}}$$

$$\frac{\mathrm{d}c}{\mathrm{d}x}(1) = 0$$

Thus, the three formulations would give first order in Δx

$$-\frac{1}{Pe}\frac{c_2-c_1}{\Delta x} + c_1 = c_{\mathrm{in}}$$

$$\frac{c_n-c_{n-1}}{\Delta x} = 0$$

plus Equation 13 at points $i = 2$ through $n - 1$; second order in Δx, by using a three-point one-sided derivative

$$-\frac{1}{Pe}\frac{-3c_1+4c_2-c_3}{2\Delta x}+c_1=c_{\text{in}}$$

$$\frac{c_{n-2}-4c_{n-1}+3c_n}{2\Delta x}=0$$

plus Equation 13 at points $i = 2$ through $n - 1$; second order in Δx, by using a false boundary point

$$-\frac{1}{Pe}\frac{c_2-c_0}{2\Delta x}+c_1=c_{\text{in}}$$

$$\frac{c_{n+1}-c_{n-1}}{2\Delta x}=0$$

plus Equation 13 at points $i = 1$ through n.

The sets of equations can be solved by using the Newton – Raphson method, as outlined in Section 1.2.

Frequently, the transport coefficients (e.g., diffusion coefficient D or thermal conductivity) depend on the dependent variable (concentration or temperature, respectively). Then the differential equation might look like

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(D(c)\frac{\mathrm{d}c}{\mathrm{d}x}\right)=0$$

This could be written as

$$-\frac{\mathrm{d}J}{\mathrm{d}x}=0 \tag{16}$$

in terms of the mass flux J, where the mass flux is given by

$$J=-D(c)\frac{\mathrm{d}c}{\mathrm{d}x}$$

Because the coefficient depends on c the equations are more complicated. A finite difference method can be written in terms of the fluxes at the midpoints, $i + 1/2$. Thus,

$$-\frac{J_{i+1/2}-J_{i-1/2}}{\Delta x}=0$$

Then the constitutive equation for the mass flux can be written as

$$J_{i+1/2}=-D\left(c_{i+1/2}\right)\frac{c_{i+1}-c_i}{\Delta x}$$

If these are combined,

$$\frac{D\left(c_{i+1/2}\right)\left(c_{i+1}-c_i\right)-D\left(c_{i-1/2}\right)\left(c_i-c_{i-1}\right)}{\Delta x^2}=0$$

This represents a set of nonlinear algebraic equations that can be solved with the Newton – Raphson method. However, in this case a viable iterative strategy is to evaluate the transport coefficients at the last value and then solve

$$\frac{D\left(c_{i+1/2}^k\right)\left(c_{i+1}^{k+1}-c_i^{k+1}\right)-D\left(c_{i-1/2}^k\right)\left(c_i^{k+1}-c_{i-1}^{k+1}\right)}{\Delta x^2}=0$$

The advantage of this approach is that it is easier to program than a full Newton – Raphson method. If the transport coefficients do not vary radically, the method converges. If the method does not converge, use of the full Newton – Raphson method may be necessary.

Three ways are commonly used to evaluate the transport coefficient at the midpoint. The first one employs the transport coefficient evaluated at the average value of the solutions on either side:

$$D\left(c_{i+1/2}\right)\approx D\left[\frac{1}{2}\left(c_{i+1}+c_i\right)\right]$$

The second approach uses the average of the transport coefficients on either side:

$$D\left(c_{i+1/2}\right)\approx\frac{1}{2}\left[D\left(c_{i+1}\right)+D\left(c_i\right)\right] \tag{17}$$

The truncation error of these approaches is also Δx^2 [78, Chap. 14], [3, p. 215]. The third approach employs an "upstream" transport coefficient.

$$D\left(C_{i+1/2}\right)\approx D\left(c_{i+1}\right),\ \text{when } D\left(c_{i+1}\right)>D\left(c_i\right)$$

$$D\left(c_{i+1/2}\right)\approx\left(c_i\right),\ \text{when } D\left(c_{i+1}\right)<D\left(c_i\right)$$

This approach is used when the transport coefficients vary over several orders of magnitude and the "upstream" direction is defined as the one in which the transport coefficient is larger. The truncation error of this approach is only Δx [78, Chap. 14] , [3, p. 253], but this approach is useful if the numerical solutions show unrealistic oscillations [3, 78].

Rigorous error bounds for linear ordinary differential equations solved with the finite difference method are dicussed by Isaacson and Keller [79, p. 431].

7.4. Orthogonal Collocation

The orthogonal collocation method has found widespread application in chemical engineering, particularly for chemical reaction engineering. In the collocation method [3], the dependent variable is expanded in a series.

$$y(x)=\sum_{i=1}^{N+2}a_iy_i(x) \tag{18}$$

Suppose the differential equation is

$$N[y] = 0$$

Then the expansion is put into the differential equation to form the residual:

$$\text{Residual} = N\left[\sum_{i=1}^{N+2} a_i y_i(x)\right]$$

In the collocation method, the residual is set to zero at a set of points called collocation points:

$$N\left[\sum_{i=1}^{N+2} a_i y_i(x_j)\right] = 0,\ j = 2,\ldots,N+1$$

This provides N equations; two more equations come from the boundary conditions, giving $N + 2$ equations for $N + 2$ unknowns. This procedure is especially useful when the expansion is in a series of orthogonal polynomials, and when the collocation points are the roots to an orthogonal polynomial, as first used by Lanczos [80, 81]. A major improvement was the proposal by Villadsen and Stewart [82] that the entire solution process be done in terms of the solution at the collocation points rather than the coefficients in the expansion. Thus, Equation 18 would be evaluated at the collocation points

$$y(x_j) = \sum_{i=1}^{N+2} a_i y_i(x_j),\ j = 1,\ldots,N+2$$

and solved for the coefficients in terms of the solution at the collocation points:

$$a_i = \sum_{j=1}^{N+2} [y_i(x_j)]^{-1}\, y(x_j),\ i = 1,\ldots,N+2$$

Furthermore, if Equation 18 is differentiated once and evaluated at all collocation points, the first derivative can be written in terms of the values at the collocation points:

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x_j) = \sum_{i=1}^{N+2} a_i \frac{\mathrm{d}y_i}{\mathrm{d}x}(x_j),\ j = 1,\ldots,N+2$$

This can be expressed as

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x_j) = \sum_{i,k=1}^{N+2} [y_i(x_k)]^{-1}\, y(x_k)\, \frac{\mathrm{d}y_i}{\mathrm{d}x}(x_j),\ j = 1,\ldots,N+2$$

or shortened to

$$\frac{\mathrm{d}y}{\mathrm{d}x}(x_j) = \sum_{k=1}^{N+2} A_{jk}\, y(x_k),$$

$$A_{jk} = \sum_{i=1}^{N+2} [y_i(x_k)]^{-1} \frac{\mathrm{d}y_i}{\mathrm{d}x}(x_j)$$

Similar steps can be applied to the second derivative to obtain

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}(x_j) = \sum_{k=1}^{N+2} B_{jk}\, y(x_k),$$

$$B_{jk} = \sum_{i=1}^{N+2} [y_i(x_k)]^{-1} \frac{\mathrm{d}^2y_i}{\mathrm{d}x^2}(x_j)$$

This method is next applied to the differential equation for reaction in a tubular reactor, after the equation has been made nondimensional so that the dimensionless length is 1.0.

$$\begin{aligned} &\frac{1}{Pe}\frac{\mathrm{d}^2c}{\mathrm{d}x^2} - \frac{\mathrm{d}c}{\mathrm{d}x} = Da\, R(c), \\ &-\frac{\mathrm{d}c}{\mathrm{d}x}(0) = Pe\,[c(0) - c_{\text{in}}],\ \frac{\mathrm{d}c}{\mathrm{d}x}(1) = 0 \end{aligned} \tag{19}$$

The differential equation at the collocation points is

$$\frac{1}{Pe}\sum_{k=1}^{N+2} B_{jk}\, c(x_k) - \sum_{k=1}^{N+2} A_{jk}\, c(x_k) = Da\, R(c_j) \tag{20}$$

and the two boundary conditions are

$$\begin{aligned} &-\sum_{k=1}^{N+2} A_{lk}\, c(x_k) = Pe\,(c_1 - c_{\text{in}}), \\ &\sum_{k=1}^{N+2} A_{N+2,k}\, c(x_k) = 0 \end{aligned} \tag{21}$$

Note that 1 is the first collocation point ($x = 0$) and $N + 2$ is the last one ($x = 1$). To apply the method, the matrices A_{ij} and B_{ij} must be found and the set of algebraic equations solved, perhaps with the Newton – Raphson method. If orthogonal polynomials are used and the collocation points are the roots to one of the orthogonal polynomials, the orthogonal collocation method results.

In the orthogonal collocation method the solution is expanded in a series involving orthogonal polynomials, where the polynomials $P_{i-1}(x)$ are defined in Section 2.2.

$$\begin{aligned} y &= a + bx + x(1-x)\sum_{i=1}^{N} a_i P_{i-1}(x) \\ &= \sum_{i=1}^{N+2} b_i P_{i-1}(x) \end{aligned} \tag{22}$$

which is also

$$y = \sum_{i=1}^{N+2} d_i x^{i-1}$$

Figure 26. Orthogonal collocation points

The collocation points are shown in Figure 26. There are N interior points plus one at each end, and the domain is always transformed to lie on 0 to 1. To define the matrices A_{ij} and B_{ij} this expression is evaluated at the collocation points; it is also differentiated and the result is evaluated at the collocation points.

$$y\left(x_j\right)=\sum_{i=1}^{N+2} d_i x_j^{i-1}$$

$$\frac{\mathrm{d}y}{\mathrm{d}x}\left(x_j\right)=\sum_{i=1}^{N+2} d_i\left(i-1\right) x_j^{i-2}$$

$$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\left(x_j\right)=\sum_{i=1}^{N+2} d_i\left(i-1\right)\left(i-2\right) x_j^{i-3}$$

These formulas are put in matrix notation, where $\boldsymbol{Q}$, $\boldsymbol{C}$, and $\boldsymbol{D}$ are $N+2$ by $N+2$ matrices.

$$\boldsymbol{y}=\boldsymbol{Q}\boldsymbol{d}, \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x}=\boldsymbol{C}\boldsymbol{d}, \frac{\mathrm{d}^2\boldsymbol{y}}{\mathrm{d}x^2}=\boldsymbol{D}\boldsymbol{d}$$

$$Q_{ji}=x_j^{i-1},\ C_{ji}=\left(i-1\right) x_j^{i-2},$$

$$D_{ji}=\left(i-1\right)\left(i-2\right) x_j^{i-3}$$

In solving the first equation for $\boldsymbol{d}$, the first and second derivatives can be written as

$$\boldsymbol{d}=\boldsymbol{Q}^{-1}\boldsymbol{y},\ \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x}=\boldsymbol{C}\boldsymbol{Q}^{-1}\boldsymbol{y}=\boldsymbol{A}\boldsymbol{y},$$
$$\frac{\mathrm{d}^2\boldsymbol{y}}{\mathrm{d}x^2}=\boldsymbol{D}\boldsymbol{Q}^{-1}\boldsymbol{y}=\boldsymbol{B}\boldsymbol{y} \tag{23}$$

Thus the derivative at any collocation point can be determined in terms of the solution at the collocation points. The same property is enjoyed by the finite difference method (and the finite element method described below), and this property accounts for some of the popularity of the orthogonal collocation method. In applying the method to Equation 19, the same result is obtained; Equations 20 and 21, with the matrices defined in Equation 23. To find the solution at a point that is not a collocation point, Equation 22 is used; once the solution is known at all collocation points, $\boldsymbol{d}$ can be found; and once $\boldsymbol{d}$ is known, the solution for any x can be found.

To use the orthogonal collocation method, the matrices are required. They can be calculated as shown above for small N ($N<8$) and by using more rigorous techniques, for higher N (see Chap. 2). However, having the matrices listed explicitly for $N=1$ and 2 is useful; this is shown in Table 7.

For some reaction diffusion problems, the solution can be an even function of x. For example, for the problem

$$\frac{\mathrm{d}^2c}{\mathrm{d}x^2}=kc,\ \frac{\mathrm{d}c}{\mathrm{d}x}\left(0\right)=0,\ c\left(1\right)=1 \tag{24}$$

the solution can be proved to involve only even powers of x. In such cases, an orthogonal collocation method, which takes this feature into

Table 7. Matrices for orthogonal collocation

$N=1, a=0.1666666667$

$$x_j=\begin{pmatrix}0\\0.5\\1\end{pmatrix}, W_j=\begin{pmatrix}a\\4\,a\\a\end{pmatrix}, A_{ji}=\begin{pmatrix}-3&4&-1\\-1&0&1\\1&-4&3\end{pmatrix}, B_{ji}=\begin{pmatrix}4&-8&4\\4&-8&4\\4&-8&4\end{pmatrix}, Q_{ji}^{-1}=\begin{pmatrix}1&0&0\\-3&4&-1\\2&-4&2\end{pmatrix}$$

$N=2$

$$x_j=\begin{pmatrix}0\\e\\1-e\\1\end{pmatrix}, W_j=\begin{pmatrix}0\\0.5\\0.5\\0\end{pmatrix}, Q_{ji}^{-1}=\begin{pmatrix}1&0&0&0\\-7&7+b&-1-b&1\\12&-18-f&12+f&-6\\-6&10+d&-10-d&6\end{pmatrix}$$

$$A_{ji}=\begin{pmatrix}-7&7+b&-1-b&1\\-1-a&a&a&1-a\\-1+a&-a&-a&1+a\\-1&1+b&-7-b&7\end{pmatrix}, B_{ji}=\begin{pmatrix}24&-37-c&25+c&-12\\16+d&-24&12&-4-d\\-4-d&12&-24&16+d\\-12&25+c&-37-c&24\end{pmatrix}$$

where $a=1.732050808$, $b=1.196152423$, $c=0.17691454$, $d=0.392304846$
$e=0.2113248654$, $f=0.58845727$

account, is convenient. This can easily be done by using expansions that only involve even powers of x. Thus, the expansion

$$y\left(x^2\right) = y\left(1\right) + \left(1-x^2\right)\sum_{i=1}^{N} a_i P_{i-1}\left(x^2\right)$$

is equivalent to

$$y\left(x^2\right) = \sum_{i=1}^{N+1} b_i P_{i-1}\left(x^2\right) = \sum_{i=1}^{N+1} d_i x^{2i-2}$$

The polynomials are defined to be orthogonal with the weighting function $W\left(x^2\right)$.

$$\int_0^1 W\left(x^2\right) P_k\left(x^2\right) P_m\left(x^2\right) x^{a-1}\mathrm{d}x = 0 \qquad k \le m-1 \tag{25}$$

where the power on x^{a-1} defines the geometry as planar or Cartesian ($a = 1$), cylindrical ($a = 2$), and spherical ($a = 3$). An analogous development is used to obtain the $(N + 1)\times(N + 1)$ matrices

$$y\left(x_j\right) = \sum_{i=1}^{N+1} d_i x_j^{2i-2}$$

$$\frac{\mathrm{d}y}{\mathrm{d}x}\left(x_j\right) = \sum_{i=1}^{N+1} d_i\left(2i-2\right) x_j^{2i-3}$$

$$\nabla^2 y\left(x_i\right) = \sum_{i=1}^{N+1} d_i \nabla^2\left(x^{2i-2}\right)\Big|_{x_j}$$

$$\boldsymbol{y} = \boldsymbol{Q}\boldsymbol{d}, \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x} = \boldsymbol{C}\boldsymbol{d}, \nabla^2\boldsymbol{y} = \boldsymbol{D}\boldsymbol{d}$$

$$Q_{ji} = x_j^{2i-2},\ C_{ji} = \left(2i-2\right) x_j^{2i-3},$$

$$D_{ji} = \nabla^2\left(x^{2i-2}\right)|_{x_j}$$

$$\boldsymbol{d} = \boldsymbol{Q}^{-1}\boldsymbol{y}, \frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}x} = \boldsymbol{C}\boldsymbol{Q}^{-1}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y},$$

$$\nabla^2\boldsymbol{y} = \boldsymbol{D}\boldsymbol{Q}^{-1}\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y}$$

In addition, the quadrature formula is

$$\boldsymbol{W}\boldsymbol{Q} = \boldsymbol{f},\ \boldsymbol{W} = \boldsymbol{f}\boldsymbol{Q}^{-1}$$

where

$$\int_0^1 x^{2i-2} x^{a-1}\mathrm{d}x = \sum_{j=1}^{N+1} W_j x_j^{2i-2} = \frac{1}{2i-2+a} \equiv f_i$$

As an example, for the problem

$$\frac{1}{x^{a-1}}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{a-1}\frac{\mathrm{d}c}{\mathrm{d}x}\right) = \varphi^2 R\left(c\right)$$

$$\frac{\mathrm{d}c}{\mathrm{d}x}\left(0\right) = 0,\ c\left(1\right) = 1$$

orthogonal collocation is applied at the interior points

$$\sum_{i=1}^{N+1} B_{ji} c_i = \varphi^2 R\left(c_j\right),\ j = 1,\ldots,N$$

and the boundary condition solved for is

$$c_{N+1} = 1$$

The boundary condition at $x = 0$ is satisfied automatically by the trial function. After the solution has been obtained, the effectiveness factor η is obtained by calculating

$$\eta \equiv \frac{\int_0^1 R\left[c\left(x\right)\right] x^{a-1}\mathrm{d}x}{\int_0^1 R\left[c\left(1\right)\right] x^{a-1}\mathrm{d}x} = \frac{\sum_{i=1}^{N+1} W_j R\left(c_j\right)}{\sum_{i=1}^{N+1} W_j R\left(1\right)}$$

Note that the effectiveness factor is the average reaction rate divided by the reaction rate evaluated at the external conditions. Error bounds have been given for linear problems [83, p. 356]. For planar geometry the error is

$$\text{Error in } \eta = \frac{\varphi^{2(2N+1)}}{\left(2N+1\right)!\left(2N+2\right)!}$$

This method is very accurate for small N (and small φ^2); note that for finite difference methods the error goes as $1/N^2$, which does not decrease as rapidly with N. If the solution is desired at the center (a frequent situation because the center concentration can be the most extreme one), it is given by

$$c\left(0\right) = d_1 \sum_{i=1}^{N+1}\left[Q^{-1}\right]_{1i} y_i$$

The collocation points are listed in Table 8. For small N the results are usually more accurate when the weighting function in Equation 25 is $1 - x^2$. The matrices for $N = 1$ and $N = 2$ are given in Table 9 for the three geometries. Computer programs to generate matrices and a program to solve reaction diffusion problems, OCRXN, are available [3, p. 325, p. 331].

Orthogonal collocation can be applied to distillation problems. Stewart et al. [84, 85] developed a method using Hahn polynomials that retains the discrete nature of a plate-to-plate distillation column. Other work treats problems with multiple liquid phases [86]. Some of the applications to chemical engineering can be found in [87 – 90].

Table 8. Collocation points for orthogonal collocation with symmetric polynomials and W=1

	Geometry		
N	Planar	Cylindrical	Spherical
1	0.5773502692	0.7071067812	0.7745966692
2	0.3399810436	0.4597008434	0.5384693101
	0.8611363116	0.8880738340	0.9061793459
3	0.2386191861	0.3357106870	0.4058451514
	0.6612093865	0.7071067812	0.7415311856
	0.9324695142	0.9419651451	0.9491079123
4	0.1834346425	0.2634992300	0.3242534234
	0.5255324099	0.5744645143	0.6133714327
	0.7966664774	0.8185294874	0.8360311073
	0.9602898565	0.9646596062	0.9681602395
5	0.1488743390	0.2165873427	0.2695431560
	0.4333953941	0.4803804169	0.5190961292
	0.6794095683	0.7071067812	0.7301520056
	0.8650633667	0.8770602346	0.8870625998
	0.9739065285	0.9762632447	0.9782286581

7.5. Orthogonal Collocation on Finite Elements

In the method of orthogonal collocation on finite elements, the domain is first divided into elements, and then within each element orthogonal collocation is applied. Figure 27 shows the domain being divided into *NE* elements, with *NCOL* interior collocation points within each element, and *NP* = *NCOL* + 2 total points per element, giving *NT* = *NE* * (*NCOL* + 1) + 1 total number of points. Within each element a local coordinate is defined

$$u = \frac{x - x_{(k)}}{\Delta x_k},\ \Delta x_k = x_{(k+1)} - x_{(k)}$$

The reaction – diffusion equation is written as

$$\frac{1}{x^{a-1}} \frac{\mathrm{d}}{\mathrm{d}x} \left(x^{a-1} \frac{\mathrm{d}c}{\mathrm{d}x} \right) = \frac{\mathrm{d}^2 c}{\mathrm{d}x^2} + \frac{a-1}{x} \frac{\mathrm{d}c}{\mathrm{d}x} = \varphi^2 R(c)$$

and transformed to give

$$\frac{1}{\Delta x_k^2} \frac{\mathrm{d}^2}{\mathrm{d}u^2} + \frac{a-1}{x_{(k)} + u \Delta x_k} \frac{1}{\Delta x_k} \frac{\mathrm{d}c}{\mathrm{d}u} = \varphi^2 R(c)$$

The boundary conditions are typically

$$\frac{\mathrm{d}c}{\mathrm{d}x}(0) = 0,\ -\frac{\mathrm{d}c}{\mathrm{d}x}(1) = Bi_m \left[c(1) - c_B \right]$$

where Bi_m is the Biot number for mass transfer. These become

$$\frac{1}{\Delta x_1} \frac{\mathrm{d}c}{\mathrm{d}u}(u = 0) = 0,$$

in the first element;

$$-\frac{1}{\Delta x_{N\ E}} \frac{\mathrm{d}c}{\mathrm{d}u}(u = 1) = Bi_m \left[c(u = 1) - c_B \right],$$

in the last element. The orthogonal collocation method is applied at each interior collocation point.

$$\frac{1}{\Delta x_k^2} \sum_{J=1}^{N\ P} B_{I\ J} c_J + \frac{a-1}{x_{(k)} + u_I \Delta x_k} \frac{1}{\Delta x_k} \sum_{J=1}^{N\ P} A_{I\ J} c_J = \\ = \varphi^2 R(c_J),\ I = 2, \ldots, N\ P - 1$$

The local points $i = 2, \ldots, NP - 1$ represent the interior collocation points. Continuity of the function and the first derivative between elements is achieved by taking

$$\frac{1}{\Delta x_{k-1}} \sum_{J=1}^{N\ P} A_{N\ P,J} c_J \Big|_{\text{element } k-1} \\ = \frac{1}{\Delta x_k} \sum_{J=1}^{N\ P} A_{1,J} c_J \Big|_{\text{element } k}$$

at the points between elements. Naturally, the computer code has only one symbol for the solution at a point shared between elements, but the derivative condition must be imposed. Finally, the boundary conditions at $x = 0$ and $x = 1$ are applied:

$$\frac{1}{\Delta x_k} \sum_{J=1}^{N\ P} A_{1,J} c_J = 0,$$

in the first element;

$$-\frac{1}{\Delta x_{N\ E}} \sum_{J=1}^{N\ P} A_{N\ P,J} c_J = Bi_m \left[c_{NP} - c_B \right],$$

in the last element.

These equations can be assembled into an overall matrix problem

$$\boldsymbol{A}\ \boldsymbol{A}\boldsymbol{c} = \boldsymbol{f}$$

Table 9. Matrices for orthogonal collocation with symmetric polynomials and $W=1-x^2$

Planar geometry, $a = 1$

$N = 1$

$$x_j = \begin{pmatrix} 0.447214 \\ 1.000000 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -1.118034 & 1.118034 \\ -2.500000 & 2.500000 \end{pmatrix}$$

$$W_j = \begin{pmatrix} 0.833333 \\ 0.166667 \end{pmatrix}, \quad B_{ji} = \begin{pmatrix} -2.5 & 2.5 \\ -2.5 & 2.5 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.25 & -0.25 \\ -1.25 & 1.25 \end{pmatrix}$$

$N = 2$

$$x_j = \begin{pmatrix} 0.285232 \\ 0.765055 \\ 1.000000 \end{pmatrix}, \quad W_j = \begin{pmatrix} 0.554858 \\ 0.378475 \\ 0.066667 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -1.752962 & 2.507614 & -0.754652 \\ -1.370599 & -0.653547 & 2.024146 \\ 1.791503 & -8.791503 & 7.000000 \end{pmatrix}$$

$$B_{ji} = \begin{pmatrix} -4.73987 & 5.67713 & -0.93725 \\ 8.32288 & -23.26013 & 14.93725 \\ 19.07189 & -47.07190 & 28.00000 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.26430 & -0.38930 & 0.125 \\ -3.42435 & 5.17435 & -1.750 \\ 2.16005 & -4.78505 & 2.625 \end{pmatrix}$$

Cylindrical geometry, $a = 2$

$N = 1$

$$x_j = \begin{pmatrix} 0.577350 \\ 1.000000 \end{pmatrix}, \quad W_j = \begin{pmatrix} 0.375 \\ 0.125 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -1.732051 & 1.732051 \\ -3.000000 & 3.000000 \end{pmatrix}$$

$$B_{ji} = \begin{pmatrix} -6 & 6 \\ -6 & 6 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.5 & -0.5 \\ -1.5 & 1.5 \end{pmatrix}$$

$N = 2$

$$x_j = \begin{pmatrix} 0.39377 \\ 0.80309 \\ 1.000000 \end{pmatrix}, \quad W_j = \begin{pmatrix} 0.18820 \\ 0.25624 \\ 0.05555 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -2.53958 & 3.82562 & -1.28603 \\ -1.37768 & -1.24519 & 2.62287 \\ 1.71548 & -9.71548 & 8.00000 \end{pmatrix}$$

$$B_{ji} = \begin{pmatrix} -9.90238 & 12.29966 & -2.39728 \\ 9.03367 & -32.76429 & 23.73061 \\ 22.7575 & -65.42415 & 42.66667 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.58808 & -0.89141 & 0.33333 \\ -3.97389 & 6.64056 & -2.66667 \\ 2.41582 & -5.74914 & 3.33333 \end{pmatrix}$$

Spherical geometry, $a = 3$

$N = 1$

$$x_j = \begin{pmatrix} 0.654654 \\ 1.000000 \end{pmatrix}, \quad W_j = \begin{pmatrix} 0.233333 \\ 0.100000 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -2.291288 & 2.291288 \\ -3.500000 & 3.500000 \end{pmatrix}$$

$$B_{ji} = \begin{pmatrix} -10.5 & 10.5 \\ -10.5 & 10.5 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.75 & -0.75 \\ -1.75 & 1.75 \end{pmatrix}$$

$N = 2$

$$x_j = \begin{pmatrix} 0.46885 \\ 0.83022 \\ 1.00000 \end{pmatrix}, \quad W_j = \begin{pmatrix} 0.09491 \\ 0.19081 \\ 0.04762 \end{pmatrix}, \quad A_{ji} = \begin{pmatrix} -3.19933 & 5.01517 & -1.81584 \\ -1.40870 & -1.80674 & 3.21544 \\ 1.69677 & -10.69677 & 9.00000 \end{pmatrix}$$

$$B_{ji} = \begin{pmatrix} -15.66996 & 20.03488 & -4.36492 \\ 9.96512 & -44.33004 & 34.36492 \\ 26.93285 & -86.93229 & 60.00000 \end{pmatrix}, \quad Q_{ji}^{-1} = \begin{pmatrix} 1.88193 & -1.50693 & 0.625 \\ -4.61225 & 8.36225 & -3.750 \\ 2.73032 & -6.85532 & 4.125 \end{pmatrix}$$

The form of these equations is special and is discussed by FINLAYSON [3, p. 116], who also gives the computer code to solve linear equations arising in such problems. Reaction – diffusion problems are solved by the program OCFERXN [3, p. 337]. See also the program COLSYS described below.

The error bounds of DEBOOR [91] give the following results for second-order problems solved with cubic trial functions on finite elements with

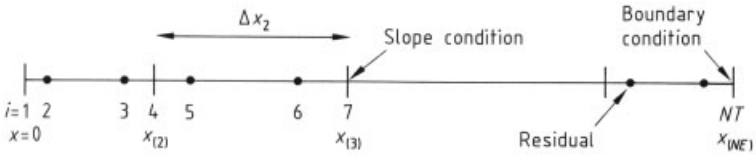

Figure 27. Grid for orthogonal collocation on finite elements

continuous first derivatives. The error at all positions is bounded by

$$\left\| \frac{\mathrm{d}^i}{\mathrm{d}x^i} \left(y - y_{\text{exact}}\right) \right\|_\infty \leq \text{constant} |\Delta x|^2$$

The error at the collocation points is more accurate, giving what is known as superconvergence.

$$\left| \frac{\mathrm{d}^i}{\mathrm{d}x^i} \left(y - y_{\text{exact}}\right) \right|_{\text{collocation points}} \leq \text{constant} |\Delta x|^4$$

7.6. Galerkin Finite Element Method

In the finite element method the domain is divided into elements and an expansion is made for the solution on each finite element. In the Galerkin finite element method an additional idea is introduced: the Galerkin method is used to solve the equation. The Galerkin method is explained before the finite element basis set is introduced.

To solve the problem

$$\frac{1}{x^{a-1}} \frac{\mathrm{d}}{\mathrm{d}x} \left(x^{a-1} \frac{\mathrm{d}c}{\mathrm{d}x} \right) = \varphi^2 R(c)$$

$$\frac{\mathrm{d}c}{\mathrm{d}x}(0) = 0, \; -\frac{\mathrm{d}c}{\mathrm{d}x}(1) = Bi_m \left[c(1) - c_B \right]$$

the unknown solution is expanded in a series of known functions $\{b_i(x)\}$, with unknown coefficients $\{a_i\}$.

$$c(x) = \sum_{i=1}^{NT} a_i b_i(x)$$

The series (the trial solution) is inserted into the differential equation to obtain the residual:

$$\text{Residual} = \sum_{i=1}^{NT} a_i \frac{1}{x^{a-1}} \frac{\mathrm{d}}{\mathrm{d}x} \left(x^{a-1} \frac{\mathrm{d}b_i}{\mathrm{d}x} \right) - \varphi^2 R \left[\sum_{i=1}^{NT} a_i b_i(x) \right]$$

The residual is then made orthogonal to the set of basis functions.

$$\int_0^1 b_j(x) \left\{ \sum_{i=1}^{NT} a_i \frac{1}{x^{a-1}} \frac{\mathrm{d}}{\mathrm{d}x} \left(x^{a-1} \frac{\mathrm{d}b_i}{\mathrm{d}x} \right) - \varphi^2 R \left[\sum_{i=1}^{NT} a_i b_i(x) \right] \right\} x^{a-1} \mathrm{d}x = 0$$

$$j = 1, \ldots, NT$$

This process makes the method a Galerkin method. The basis for the orthogonality condition is that a function that is made orthogonal to each member of a complete set is then zero. The residual is being made orthogonal, and if the basis functions are complete, and an infinite number of them are used, then the residual is zero. Once the residual is zero the problem is solved. It is necessary also to allow for the boundary conditions. This is done by integrating the first term of Equation 26 by parts and then inserting the boundary conditions:

$$\begin{aligned} &\int_0^1 b_j(x) \frac{1}{x^{a-1}} \frac{\mathrm{d}}{\mathrm{d}x} \left(x^{a-1} \frac{\mathrm{d}b_i}{\mathrm{d}x} \right) x^{a-1} \mathrm{d}x = \\ &\int_0^1 \frac{\mathrm{d}}{\mathrm{d}x} \left[b_j(x) x^{a-1} \frac{\mathrm{d}b_i}{\mathrm{d}x} \right] \mathrm{d}x - \int_0^1 \frac{\mathrm{d}b_j}{\mathrm{d}x} \frac{\mathrm{d}b_i}{\mathrm{d}x} x^{a-1} \mathrm{d}x \\ &= \left[b_j(x) x^{a-1} \frac{\mathrm{d}b_i}{\mathrm{d}x} \right]_0^1 - \int_0^1 \frac{\mathrm{d}b_j}{\mathrm{d}x} \frac{\mathrm{d}b_i}{\mathrm{d}x} x^{a-1} \mathrm{d}x \\ &= -\int_0^1 \frac{\mathrm{d}b_j}{\mathrm{d}x} \frac{\mathrm{d}b_i}{\mathrm{d}x} x^{a-1} \mathrm{d}x - Bi_m b_j(1) \left[b_i(1) - c_B \right] \end{aligned} \tag{27}$$

Combining this with Equation 26 gives

$$\begin{aligned} &-\sum_{i=1}^{NT} \int_0^1 \frac{\mathrm{d}b_j}{\mathrm{d}x} \frac{\mathrm{d}b_i}{\mathrm{d}x} x^{a-1} \mathrm{d}x a_i \\ &- Bi_m b_j(1) \left[\sum_{i=1}^{NT} a_i b_i(1) - c_B \right] \\ &= \varphi^2 \int_0^1 b_j(x) \left[\sum_{i=1}^{NT} a_i b_i(x) \right] x^{a-1} \mathrm{d}x \\ &j = 1, \ldots, NT \end{aligned} \tag{28}$$

This equation defines the Galerkin method, and a solution that satisfies this equation (for all $j = 1, \ldots, \infty$) is called a weak solution. For an approximate solution the equation is written once for each member of the trial function, $j = 1, \ldots, NT$. If the boundary condition is

$$c(1) = c_B$$

then the boundary condition is used (instead of Eq. 28) for $j = NT$,

$$\sum_{i=1}^{NT} a_i b_i(1) = c_B$$

The Galerkin finite element method results when the Galerkin method is combined with a finite element trial function. Both linear and quadratic finite element approximations are described in Chapter 2. The trial functions $b_i(x)$ are then generally written as $N_i(x)$.

$$c(x) = \sum_{i=1}^{NT} c_i N_i(x)$$

Each $N_i(x)$ takes the value 1 at the point x_i and zero at all other grid points (Chap. 2). Thus c_i are the nodal values, $c(x_i) = c_i$. The first derivative must be transformed to the local coordinate system, $u = 0$ to 1 when x goes from x_i to $x_i + \Delta x$.

$$\frac{\mathrm{d}N_j}{\mathrm{d}x} = \frac{1}{\Delta x_e}\frac{\mathrm{d}N_J}{\mathrm{d}u},\ \mathrm{d}x = \Delta x_e \mathrm{d}u$$

in the e-th element. Then the Galerkin method is

$$\begin{aligned}
&-\sum\nolimits_e \frac{1}{\Delta x_e}\sum_{I=1}^{NP}\int_0^1 \frac{\mathrm{d}N_J}{\mathrm{d}u}\frac{\mathrm{d}N_I}{\mathrm{d}u}(x_e+u\Delta x_e)^{a-1}\mathrm{d}u c_I^e \\
&-Bi_m\sum\nolimits_e N_J(1)\left[\sum_{I=1}^{NP} c_I^e N_I(1) - c_1\right] \\
&= \varphi^2\sum\nolimits_e \Delta x_e \int_0^1 N_J(u)\, R\left[\sum_{I=1}^{NP} c_I^e N_I(u)\right] \\
&(x_e+u\Delta x_e)^{a-1}\mathrm{d}u
\end{aligned} \tag{29}$$

The element integrals are defined as

$$B_{JI}^e = -\frac{1}{\Delta x_e}\int_0^1 \frac{\mathrm{d}N_J}{\mathrm{d}u}\frac{\mathrm{d}N_I}{\mathrm{d}u}(x_e+u\Delta x_e)^{a-1}\mathrm{d}u,$$

$$F_J^e = \varphi^2\Delta x_e\int_0^1 N_J(u)\, R\left[\sum_{I=1}^{NP} c_I^e N_I(u)\right]$$
$$(x_e+u\Delta x_e)^{a-1}\mathrm{d}u$$

whereas the boundary element integrals are

$$BB_{JI}^e = -Bi_m N_J(1) N_I(1),$$

$$FF_J^e = -Bi_m N_J(1) c_1$$

Then the entire method can be written in the compact notation

$$\sum\nolimits_e B_{JI}^e c_I^e + \sum\nolimits_e BB_{JI}^e c_I^e = \sum\nolimits_e F_J^e + \sum\nolimits_e FF_J^e$$

The matrices for various terms are given in Table 10. This equation can also be written in the form

$$\boldsymbol{AAc} = \boldsymbol{f}$$

where the matrix $\boldsymbol{AA}$ is sparse. If linear elements are used the matrix is tridiagonal. If quadratic elements are used the matrix is pentadiagonal. Naturally the linear algebra is most efficiently carried out if the sparse structure is taken into account. Once the solution is found the solution at any point can be recovered from

$$c^e(u) = c_{I=1}^e(1-u) + c_{I=2}^e u$$

for linear elements

$$c^e(u) = c_{I=1}^e 2(u-1)\left(u-\tfrac{1}{2}\right)$$
$$+c_{I=2}^e 4u(1-u) + c_{I=3}^e 2u\left(u-\tfrac{1}{2}\right)$$

for quadratic elements

Table 10. Element matrices for Galerkin method

Linear shape functions	*Quadratic shape functions*
$N_1 = 1-u,\ N_2 = u,\ \frac{\mathrm{d}N_1}{\mathrm{d}u} = -1,\ \frac{\mathrm{d}N_2}{\mathrm{d}u} = 1$	$N_1 = 2(u-1)\left(u-\frac{1}{2}\right),\ N_2 = 4u(1-u),\ N_3 = 2u\left(u-\frac{1}{2}\right),$ $\frac{\mathrm{d}N_1}{\mathrm{d}u} = 4u-3,\ \frac{\mathrm{d}N_2}{\mathrm{d}u} = 4-8u,\ \frac{\mathrm{d}N_3}{\mathrm{d}u} = 4u-1$
$\int_0^1 \frac{\mathrm{d}N_J}{\mathrm{d}u}\frac{\mathrm{d}N_I}{\mathrm{d}u}\mathrm{d}u = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},\ \int_0^1 N_J\frac{\mathrm{d}N_I}{\mathrm{d}u}\mathrm{d}u = \begin{pmatrix} -\frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$	$\int_0^1 \frac{\mathrm{d}N_J}{\mathrm{d}u}\frac{\mathrm{d}N_I}{\mathrm{d}u}\mathrm{d}u = \frac{1}{3}\begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix},\ \int_0^1 N_J\frac{\mathrm{d}N_I}{\mathrm{d}u}\mathrm{d}u = \frac{1}{6}\begin{pmatrix} -3 & 4 & -1 \\ -4 & 0 & 4 \\ 1 & -4 & 3 \end{pmatrix}$
$\int_0^1 N_J N_I\,\mathrm{d}u = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{3} \end{pmatrix},\ \int_0^1 N_J\,\mathrm{d}u = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix},\ \int_0^1 N_J u\,\mathrm{d}u = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{3} \end{pmatrix}$	$\int_0^1 N_J N_I\,\mathrm{d}u = \frac{1}{30}\begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix},\ \int_0^1 N_J\,\mathrm{d}u = \frac{1}{6}\begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix},\ \int_0^1 N_J u\,\mathrm{d}u = \frac{1}{6}\begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}$

Because the integrals in Equation 28 may be complicated, they are usually formed by using Gaussian quadrature. If *NG* Gauss points are used, a typical term would be

$$\int_0^1 N_J(u)\, R\left[\sum_{I=1}^{NP} c_I^e N_I(u)\right](x_e+u\Delta x_e)^{a-1}\mathrm{d}u$$
$$= \sum_{k=1}^{NG} W_k N_J(u_k)\, R\left[\sum_{I=1}^{NP} c_I^e N_I(u_k)\right]$$
$$(x_e+u_k\Delta x_e)^{a-1}$$

For an application of the finite element method in fluid mechanics, see → Fluid Mechanics, Chap. 5.2.

7.7. Cubic B-Splines

Cubic B-splines have cubic approximations within each element, but first and second derivatives continuous between elements. The functions are the same ones discussed in Chapter 2, and they can be used to solve differential equations, too. See Sincovec [92].

7.8. Adaptive Mesh Strategies

In many two-point boundary value problems, the difficulty in the problem is the formation of a boundary layer region, or a region in which the solution changes very dramatically. In such cases small mesh spacing should be used there, either with the finite difference method or the finite element method. If the region is known a priori, small mesh spacings can be assumed at the boundary layer. If the region is not known though, other techniques must be used. These techniques are known as adaptive mesh techniques. The general strategy is to estimate the error, which depends on the grid size and derivatives of the solution, and refine the mesh where the error is large.

The adaptive mesh strategy was employed by Ascher et al. [93] and by Russell and Christiansen [94]. For a second-order differential equation and cubic trial functions on finite elements, the error in the *i*-th element is given by

$$\|\text{Error}\|_i = c\Delta x_i^4 \|u^{(4)}\|_i$$

Because cubic elements do not have a nonzero fourth derivative, the third derivative in adjacent elements is used [3, p. 166]:

$$a_i = \frac{1}{\Delta x_i^3}\frac{\mathrm{d}^3 c^i}{\mathrm{d}u^3},\ a_{i+1} = \frac{1}{\Delta x_{i+1}^3}\frac{\mathrm{d}^3 c^{i+1}}{\mathrm{d}u^3}$$

$$\|u^{(4)}\|_i \approx \frac{1}{2}\left[\frac{a_i - a_{i-1}}{\frac{1}{2}(x_{i+1}-x_{i-1})} + \frac{a_{i+1}-a_i}{\frac{1}{2}(x_{i+2}-x_i)}\right]$$

Element sizes are then chosen so that the following error bounds are satisfied

$$C\Delta x_i^4 \|u^{(4)}\|_i \le \varepsilon \text{ for all } i$$

These features are built into the code COLSYS (http://www.netlib.org/ode/).

The error expected from a method one order higher and one order lower can also be defined. Then a decision about whether to increase or decrease the order of the method can be made by taking into account the relative work of the different orders. This provides a method of adjusting both the mesh spacing (Δx, sometimes called h) and the degree of polynomial (p). Such methods are called $h-p$ methods.

7.9. Comparison

What method should be used for any given problem? Obviously the error decreases with some power of Δx, and the power is higher for the higher order methods, which suggests that the error is less. For example, with linear elements the error is

$$y(\Delta x) = y_{\text{exact}} + c_2\Delta x^2$$

for small enough (and uniform) Δx. A computer code should be run for varying Δx to confirm this. For quadratic elements, the error is

$$y(\Delta x) = y_{\text{exact}} + c_3\Delta x^3$$

If orthogonal collocation on finite elements is used with cubic polynomials, then

$$y(\Delta x) = y_{\text{exact}} + c_4\Delta x^4$$

However, the global methods, not using finite elements, converge even faster [95], for example,

$$y(N) = y_{\text{exact}} + c_N\left(\frac{1}{NCOL}\right)^{NCOL}$$

Yet the workload of the methods is also different. These considerations are discussed in [3]. Here, only sweeping generalizations are given.

If the problem has a relatively smooth solution, then the orthogonal collocation method

is preferred. It gives a very accurate solution, and N can be quite small so the work is small. If the problem has a steep front in it, the finite difference method or finite element method is indicated, and adaptive mesh techniques should probably be employed. Consider the reaction – diffusion problem: as the Thiele modulus φ increases from a small value with no diffusion limitations to a large value with significant diffusion limitations, the solution changes as shown in Figure 28. The orthogonal collocation method is initially the method of choice. For intermediate values of φ, $N = 3 - 6$ must be used, but orthogonal collocation still works well (for η down to approximately 0.01). For large φ, use of the finite difference method, the finite element method, or an asymptotic expansion for large φ is better. The decision depends entirely on the type of solution that is obtained. For steep fronts the finite difference method and finite element method with adaptive mesh are indicated.

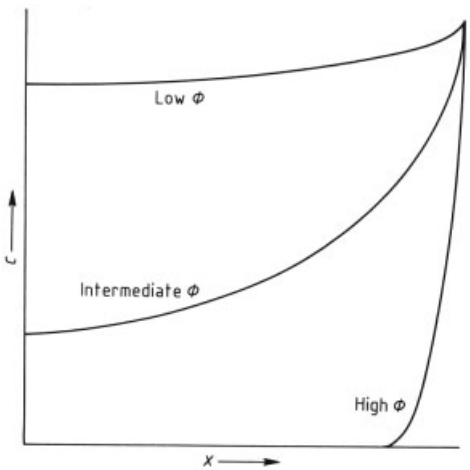

Figure 28. Concentration solution for different values of Thiele modulus

7.10. Singular Problems and Infinite Domains

If the solution being sought has a singularity, a good numerical solution may be hard to find. Sometimes even the location of the singularity may not be known [96, pp. 230 – 238]. One method of solving such problems is to refine the mesh near the singularity, by relying on the better approximation due to a smaller Δx. Another approach is to incorporate the singular trial function into the approximation. Thus, if the solution approaches $f(x)$ as x goes to zero, and $f(x)$ becomes infinite, an approximation may be taken as

$$y(x) = f(x) + \sum_{i=1}^{N} a_i y_i(x)$$

This function is substituted into the differential equation, which is solved for a_i. Essentially, a new differential equation is being solved for a new variable:

$$u(x) \equiv y(x) - f(x)$$

The differential equation is more complicated but has a better solution near the singularity (see [97, pp. 189 – 192], [98, p. 611]).

Sometimes the domain is infinite. Boundary layer flow past a flat plate is governed by the Blasius equation for stream function [99, p. 117].

$$2\frac{\mathrm{d}^3}{\mathrm{d}\eta^3} + f\frac{\mathrm{d}^2}{\mathrm{d}\eta^2} = 0$$

$$f = \frac{\mathrm{d}f}{\mathrm{d}\eta} = 0 \text{ at } \eta = 0$$

$$\frac{\mathrm{d}f}{\mathrm{d}\eta} = 1 \text{ at } \eta \to \infty$$

Because one boundary is at infinity using a mesh with a constant size is difficult! One approach is to transform the domain. For example, let

$$z = \mathrm{e}^{-\eta}$$

Then $\eta = 0$ becomes $z = 1$ and $\eta = \infty$ becomes $z = 0$. The derivatives are

$$\frac{\mathrm{d}z}{\mathrm{d}\eta} = -\mathrm{e}^{-\eta} = -z, \ \frac{\mathrm{d}^2 z}{\mathrm{d}\eta^2} = \mathrm{e}^{-\eta} = z$$

$$\frac{\mathrm{d}f}{\mathrm{d}\eta} = \frac{\mathrm{d}f}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}\eta} = -z\frac{\mathrm{d}f}{\mathrm{d}z}$$

$$\frac{\mathrm{d}^2 f}{\mathrm{d}\eta^2} = \frac{\mathrm{d}^2 f}{\mathrm{d}z^2}\left(\frac{\mathrm{d}z}{\mathrm{d}\eta}\right)^2 + \frac{\mathrm{d}f}{\mathrm{d}z}\frac{\mathrm{d}^2 z}{\mathrm{d}\eta^2} = z^2\frac{\mathrm{d}^2 f}{\mathrm{d}z^2} + z\frac{\mathrm{d}f}{\mathrm{d}z}$$

The Blasius equation becomes

$$2\left[-z^3\frac{\mathrm{d}^3 f}{\mathrm{d}z^3} - 3z^2\frac{\mathrm{d}^2 f}{\mathrm{d}z^2} - z\frac{\mathrm{d}f}{\mathrm{d}z}\right]$$

$$+ f\left[z^2\frac{\mathrm{d}^2}{\mathrm{d}z^2} + z\frac{\mathrm{d}f}{\mathrm{d}z}\right] = 0 \qquad \text{for } 0 \le z \le 1.$$

The differential equation now has variable coefficients, but these are no more difficult to handle than the original nonlinearities.

Another approach is to use a variable mesh, perhaps with the same transformation. For example, use $z = e^{-\eta}$ and a constant mesh size in z. Then with 101 points distributed uniformly from $z = 0$ to $z = 1$, the following are the nodal points:

$$z = 0.,0.01,0.02,\ldots,0.99,1.0$$

$$\eta = \infty,4.605,3.912,\ldots,0.010,0$$

$$\Delta\eta = \infty,0.639,\ldots,0.01$$

Still another approach is to solve on a finite mesh in which the last point is far enough away that its location does not influence the solution. A location that is far enough away must be found by trial and error.

8. Partial Differential Equations

Partial differential equations are differential equations in which the dependent variable is a function of two or more independent variables. These can be time and one space dimension, or time and two or more space dimensions, or two or more space dimensions alone. Problems involving time are generally either hyperbolic or parabolic, whereas those involving spatial dimensions only are often elliptic. Because the methods applied to each type of equation are very different, the equation must first be classified as to its type. Then the special methods applicable to each type of equation are described. For a discussion of all methods, see [100 – 103]; for a discussion oriented more toward chemical engineering applications, see [104]. Examples of hyperbolic and parabolic equations include chemical reactors with radial dispersion, pressure-swing adsorption, dispersion of an effluent, and boundary value problems with transient terms added (heat transfer, mass transfer, chemical reaction). Examples of elliptic problems include heat transfer and mass transfer in two and three spatial dimensions and steady fluid flow.

8.1. Classification of Equations

A set of differential equations may be hyperbolic, elliptic, or parabolic, or it may be of mixed type. The type may change for different parameters or in different regions of the flow. This can happen in the case of nonlinear problems; an example is a compressible flow problem with both subsonic and supersonic regions. *Characteristic curves* are curves along which a discontinuity can propagate. For a given set of equations, it is necessary to determine if characteristics exist or not, because that determines whether the equations are hyperbolic, elliptic, or parabolic.

Linear Problems For linear problems, the theory summarized by Joseph et al. [105] can be used.

$$\frac{\partial}{\partial t},\frac{\partial}{\partial x_i},\ldots,\frac{\partial}{\partial x_n}$$

is replaced with the Fourier variables

$$i\xi_0,i\xi_1,\ldots,i\xi_n$$

If the m-th order differential equation is

$$P = \sum\nolimits_{|\alpha|=m} a_\alpha \partial^\alpha + \sum\nolimits_{|\alpha|<m} b_\alpha \partial^\alpha$$

where

$$\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_n)\,,\ |\alpha| = \sum_{i=0}^{n} \alpha_i$$

$$\partial^\alpha = \frac{\partial^{|\alpha|}}{\partial t^{\alpha_0}\partial x_1^{\alpha_1}\ldots\partial x_n^{\alpha_n}}$$

the characteristic equation for P is defined as

$$\begin{aligned} &\textstyle\sum_{|\alpha|=m} a_\alpha \sigma^\alpha = 0,\ \sigma = (\sigma_0,\sigma_1,\ldots,\sigma_n) \\ &\sigma^\alpha = \sigma_0^{\alpha_0}\sigma_1^{\alpha_1}\ldots\sigma_n^{\alpha_n} \end{aligned} \tag{30}$$

where σ represents coordinates. Thus only the highest derivatives are used to determine the type. The surface is defined by this equation plus a normalization condition:

$$\sum_{k=0}^{n} \sigma_k^2 = 1$$

The shape of the surface defined by Equation 30 is also related to the type: elliptic equations give rise to ellipses; parabolic equations give rise to parabolas; and hyperbolic equations give rise to hyperbolas.

$$\frac{\sigma_1^2}{a^2} + \frac{\sigma_2^2}{b^2} = 1,\ \text{Ellipse}$$

$$\sigma_0 = a\sigma_1^2,\ \text{Parabola}$$

$$\sigma_0^2 - a\sigma_1^2 = 0,\ \text{Hyperbola}$$

If Equation 30 has no nontrivial real zeroes then the equation is called elliptic. If all the roots are

real and distinct (excluding zero) then the operator is hyperbolic.

This formalism is applied to three basic types of equations. First consider the equation arising from steady diffusion in two dimensions:

$$\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}=0$$

This gives

$$-\xi_1^2-\xi_2^2=-\left(\xi_2^1+\xi_2^2\right)=0$$

Thus,

$$\sigma_1^2+\sigma_2^2=1 \text{ (normalization)}$$

$$\sigma_1^2+\sigma_2^2=0 \text{ (equation)}$$

These cannot both be satisfied so the problem is elliptic. When the equation is

$$\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x^2}=0$$

then

$$-\xi_0^2+\xi_1^2=0$$

Now real ξ_0 can be solved and the equation is hyperbolic

$$\sigma_0^2+\sigma_1^2=1 \text{ (normalization)}$$

$$-\sigma_0^2+\sigma_1^2=0 \text{ (equation)}$$

When the equation is

$$\frac{\partial c}{\partial t}=D\left(\frac{\partial^2 c}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}\right)$$

then

$$\sigma_0^2+\sigma_1^2+\sigma_2^2=1 \text{ (normalization)}$$

$$\sigma_1^2+\sigma_2^2=0 \text{ (equation)}$$

thus we get

$$\sigma_0^2=1 \text{ (for normalization)}$$

and the characteristic surfaces are hyperplanes with t = constant. This is a parabolic case.

Consider next the telegrapher's equation:

$$\frac{\partial T}{\partial t}+\beta\frac{\partial^2 T}{\partial t^2}=\frac{\partial^2 T}{\partial x^2}$$

Replacing the derivatives with the Fourier variables gives

$$i\xi_0-\beta\xi_0^2+\xi_1^2=0$$

The equation is thus second order and the type is determined by

$$-\beta\sigma_0^2+\sigma_1^2=0$$

The normalization condition

$$\sigma_0^2+\sigma_1^2=1$$

is required. Combining these gives

$$1-(1+\beta)\,\sigma_0^2=0$$

The roots are real and the equation is hyperbolic. When $\beta = 0$

$$\xi_1^2=0$$

and the equation is parabolic.

First-order quasi-linear problems are written in the form

$$\sum_{l=0}^{n}\boldsymbol{A}_1\frac{\partial \boldsymbol{u}}{\partial x_1}=\boldsymbol{f},\ \boldsymbol{x}=(t,x_1\ldots,x_n) \qquad (31)$$
$$\boldsymbol{u}=(u_1,u_2,\ldots,u_k)$$

The matrix entries $\boldsymbol{A}_1$ is a $k\times k$ matrix whose entries depend on $\boldsymbol{u}$ but not on derivatives of $\boldsymbol{u}$. Equation 31 is hyperbolic if

$$\mathrm{A}=\mathrm{A}_\mu$$

is nonsingular and for any choice of real λ_1, l = 0, . . . , n, $l \neq \mu$ the roots α_k of

$$\det\left(\alpha\boldsymbol{A}-\sum_{\substack{l=0\\ l\neq\mu}}^{n}\lambda_1\boldsymbol{A}_1\right)=0$$

are real. If the roots are complex the equation is elliptic; if some roots are real and some are complex the equation is of mixed type.

Apply these ideas to the *advection equation*

$$\frac{\partial u}{\partial t}+F(u)\frac{\partial u}{\partial x}=0$$

Thus,

$$\det(\alpha\boldsymbol{A}_0-\lambda_1\boldsymbol{A}_1)=0 \text{ or } \det(\alpha\boldsymbol{A}_1-\lambda_0\boldsymbol{A}_0)=0$$

In this case,

$$n=1,\ \boldsymbol{A}_0=1,\ \boldsymbol{A}_1=F(u)$$

Using the first of the above equations gives

$$\det(\alpha-\lambda_1F(u))=0, \text{or } \alpha=\lambda_1F(u)$$

Thus, the roots are real and the equation is hyperbolic.

The final example is the heat conduction problem written as

$$\varrho C_p \frac{\partial T}{\partial t} = -\frac{\partial q}{\partial x},\ q = -k\frac{\partial T}{\partial x}$$

In this formulation the constitutive equation for heat flux is separated out; the resulting set of equations is first order and written as

$$\varrho C_p \tfrac{\partial T}{\partial t} + \tfrac{\partial q}{\partial x} = 0$$

$$k\tfrac{\partial T}{\partial x} = -q$$

In matrix notation this is

$$\begin{bmatrix} \varrho C_p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial T}{\partial t} \\ \frac{\partial q}{\partial t} \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ k & 0 \end{bmatrix} \begin{bmatrix} \frac{\partial T}{\partial x} \\ \frac{\partial q}{\partial t} \end{bmatrix} = \begin{bmatrix} 0 \\ -q \end{bmatrix}$$

This compares with

$$\boldsymbol{A}_0 \frac{\partial \boldsymbol{u}}{\partial x_0} + \boldsymbol{A}_1 \frac{\partial \boldsymbol{u}}{\partial x_1} = \boldsymbol{f}$$

In this case $\boldsymbol{A}_0$ is singular whereas $\boldsymbol{A}_1$ is nonsingular. Thus,

$$\det\left(\alpha \boldsymbol{A}_1 - \lambda_0 \boldsymbol{A}_0\right) = 0$$

is considered for any real λ_0. This gives

$$\begin{vmatrix} -\varrho C_p \lambda_0 & \alpha \\ k\alpha & 0 \end{vmatrix} = 0$$

or

$$\alpha^2 k = 0$$

Thus the α is real, but zero, and the equation is parabolic.

8.2. Hyperbolic Equations

The most common situation yielding hyperbolic equations involves unsteady phenomena with convection. A prototype equation is

$$\frac{\partial c}{\partial t} + \frac{\partial F(c)}{\partial x} = 0$$

Depending on the interpretation of c and $F(c)$, this can represent accumulation of mass and convection. With $F(c) = u\,c$, where u is the velocity, the equation represents a mass balance on concentration. If diffusive phenomenon are important, the equation is changed to

$$\frac{\partial c}{\partial t} + \frac{\partial F(c)}{\partial x} = D\frac{\partial^2 c}{\partial x^2} \tag{32}$$

where D is a diffusion coefficient. Special cases are the *convective diffusive equation*

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} = D\frac{\partial^2 c}{\partial x^2} \tag{33}$$

and *Burgers viscosity equation*

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \tag{34}$$

where u is the velocity and ν is the kinematic viscosity. This is a prototype equation for the Navier – Stokes equations (→ Fluid Mechanics). For *adsorption phenomena* [106, p. 202],

$$\varphi\frac{\partial c}{\partial t} + \varphi u\frac{\partial c}{\partial x} + (1-\varphi)\frac{\mathrm{d}f}{\mathrm{d}c}\frac{\partial c}{\partial t} = 0 \tag{35}$$

where φ is the void fraction and $f(c)$ gives the equilibrium relation between the concentrations in the fluid and in the solid phase. In these examples, if the diffusion coefficient D or the kinematic viscosity ν is zero, the equations are hyperbolic. If D and ν are small, the phenomenon may be essentially hyperbolic even though the equations are parabolic. Thus the numerical methods for hyperbolic equations may be useful even for parabolic equations.

Equations for several methods are given here, as taken from [107]. If the convective term is treated with a centered difference expression the solution exhibits oscillations from node to node, and these vanish only if a very fine grid is used. The simplest way to avoid the oscillations with a hyperbolic equation is to use upstream derivatives. If the flow is from left to right, this would give the following for Equations (40):

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} + \frac{F(c_i) - F(c_{i-1})}{\Delta x} = D\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$$

for Equation 34:

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} + u_i\frac{u_i - u_{i-1}}{\Delta x} = \nu\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2}$$

and for Equation 35:

$$\varphi\frac{\mathrm{d}c_i}{\mathrm{d}t} + \varphi u_i\frac{c_i - c_{i-1}}{\Delta x} + (1-\varphi)\frac{\mathrm{d}f}{\mathrm{d}c}\Big|_i\frac{\mathrm{d}c_i}{\mathrm{d}t} = 0$$

If the flow were from right to left, then the formula would be

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} + \frac{F(c_{i+1}) - F(c_i)}{\Delta x} = D\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2}$$

If the flow could be in either direction, a local determination must be made at each node i and the appropriate formula used. The effect of using upstream derivatives is to add artificial or numerical diffusion to the model. This can be

ascertained by taking the finite difference form of the convective diffusion equation

$$\frac{\mathrm{d}c_i}{\mathrm{d}t}+u\frac{c_i-c_{i-1}}{\Delta x}=D\frac{c_{i+1}-2c_i+c_{i-1}}{\Delta x^2}$$

and rearranging

$$\frac{\mathrm{d}c_i}{\mathrm{d}t}+u\frac{c_{i+1}-c_{i-1}}{2\Delta x}$$

$$=\left(D+\frac{u\Delta x}{2}\right)\frac{c_{i+1}-2c_i+c_{i-1}}{\Delta x^2}$$

Thus the diffusion coefficient has been changed from

$$D \text{ to } D+\frac{u\Delta x}{2}$$

Expressed in terms of the cell Peclet number, $Pe_\Delta = u\Delta x/D$, this is D is changed to $D[1+Pe_\Delta/2]$

The cell Peclet number should always be calculated as a guide to the calculations. Using a large cell Peclet number and upstream derivatives leads to excessive (and artificial) smoothing of the solution profiles.

Another method often used for hyperbolic equations is the *MacCormack method*. This method has two steps; it is written here for Equation 33.

$$c_i^{*n+1}=c_i^n-\frac{u\Delta t}{\Delta x}\left(c_{i+1}^n-c_i^n\right)$$

$$+\frac{\Delta t D}{\Delta x^2}\left(c_{i+1}^n-2c_i^n+c_{i-1}^n\right)$$

$$c_i^{n+1}=\frac{1}{2}\left(c_i^n+c_i^{*n+1}\right)-\frac{u\Delta t}{2\Delta x}\left(c_i^{*n+1}-c_{i-1}^{*n+1}\right)$$

$$+\frac{\Delta t D}{2\Delta x^2}\left(c_{i+1}^{*n+1}-2c_i^{*n+1}+c_{i-1}^{*n+1}\right)$$

The concentration profile is steeper for the MacCormack method than for the upstream derivatives, but oscillations can still be present. The flux-corrected transport method can be added to the MacCormack method. A solution is obtained both with the upstream algorithm and the MacCormack method; then they are combined to add just enough diffusion to eliminate the oscillations without smoothing the solution too much. The algorithm is complicated and lengthy but well worth the effort [107 – 109].

If finite element methods are used, an explicit Taylor – Galerkin method is appropriate. For the convective diffusion equation the method is

$$\frac{1}{6}\left(c_{i+1}^{n+1}-c_{i+1}^n\right)+\frac{2}{3}\left(c_i^{n+1}-c_i^n\right)+\frac{1}{6}\left(c_{i-1}^{n+1}-c_{i-1}^n\right)$$

$$=-\frac{u\Delta t}{2\Delta x}\left(c_{i+1}^n-c_{i-1}^n\right)+\left(\frac{\Delta t D}{\Delta x^2}+\frac{u^2\Delta t^2}{2\Delta x^2}\right)\left(c_{i+1}^n\right.$$

$$\left.-2c_i^n+c_{i-1}^n\right)$$

Leaving out the $u^2\ \Delta t^2$ terms gives the Galerkin method. Replacing the left-hand side with

$$c_i^{n+1}-c_i^n$$

gives the Taylor finite difference method, and dropping the $u^2\ \Delta t^2$ terms in that gives the centered finite difference method. This method might require a small time step if reaction phenomena are important. Then the implicit Galerkin method (without the Taylor terms) is appropriate

A stability diagram for the explicit methods applied to the convective diffusion equation is shown in Figure 29. Notice that all the methods require

$$Co=\frac{u\Delta t}{\Delta x}\leq 1$$

where *Co* is the Courant number. How much *Co* should be less than one depends on the method and on $r = D\ \Delta t/\Delta x^2$, as given in Figure 29. The MacCormack method with flux correction requires a smaller time step than the MacCormack method alone (curve a), and the implicit Galerkin method (curve e) is stable for all values of *Co* and *r* shown in Figure 29 (as well as even larger values).

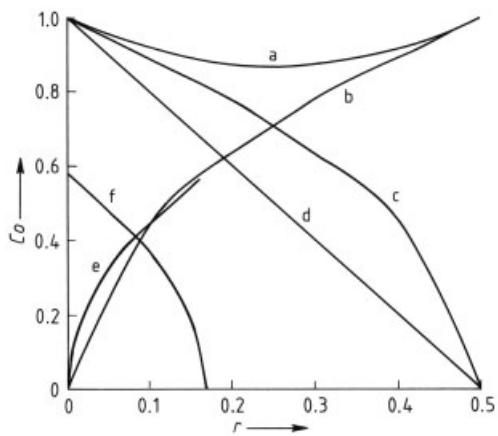

Figure 29. Stability diagram for convective diffusion equation (stable below curve)
a) MacCormack; b) Centered finite difference; c) Taylor finite difference; d) Upstream; e) Galerkin; f) Taylor – Galerkin

Each of these methods tries to avoid oscillations that would disappear if the mesh were

fine enough. For the steady convective diffusion equation these oscillations do not occur provided

$$\frac{u\Delta x}{2D} = \frac{Pe_{\Delta}}{2} <= 1 \tag{36}$$

For large u, Δx must be small to meet this condition. An alternative is to use a small Δx in regions where the solution changes drastically. Because these regions change in time, the elements or grid points must move. The criteria to move the grid points can be quite complicated, and typical methods are reviewed in [107]. The criteria include moving the mesh in a known way (when the movement is known a priori), moving the mesh to keep some property (e.g., first- or second-derivative measures) uniform over the domain, using a Galerkin or weighted residual criterion to move the mesh, and Euler – Lagrange methods which move part of the solution exactly by convection and then add on some diffusion after that.

The final illustration is for adsorption in a packed bed, or chromatography. Equation 35 can be solved when the adsorption phenomenon is governed by a Langmuir isotherm.

$$f(c) = \frac{\alpha c}{1+Kc}$$

Similar numerical considerations apply and similar methods are available [110 – 112].

8.3. Parabolic Equations in One Dimension

In this section several methods are applied to parabolic equations in one dimension: separation of variables, combination of variables, finite difference method, finite element method, and the orthogonal collocation method. Separation of variables is successful for linear problems, whereas the other methods work for linear or nonlinear problems. The finite difference, the finite element, and the orthogonal collocation methods are numerical, whereas the separation or combination of variables can lead to analytical solutions.

Analytical Solutions. Consider the diffusion equation

$$\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2}$$

with boundary and initial conditions

$$c(x,0) = 0$$

$$c(0,t) = 1,\ c(L,t) = 0$$

A solution of the form

$$c(x,t) = T(t)\,X(x)$$

is attempted and substituted into the equation, with the terms separated to give

$$\frac{1}{DT}\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{1}{X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2}$$

One side of this equation is a function of x alone, whereas the other side is a function of t alone. Thus, both sides must be a constant. Otherwise, if x is changed one side changes, but the other cannot because it depends on t. Call the constant $-\lambda$ and write the separate equations

$$\frac{\mathrm{d}T}{\mathrm{d}t} - \lambda DT,\ \frac{\mathrm{d}^2X}{\mathrm{d}x^2} - \lambda X$$

The first equation is solved easily

$$T(t) = T(0)\,\mathrm{e}^{-\lambda Dt}$$

and the second equation is written in the form

$$\frac{\mathrm{d}^2X}{\mathrm{d}x^2} + \lambda X = 0$$

Next consider the boundary conditions. If they are written as

$$c(L,t) = 1 = T(t)\,X(L)$$

$$c(0,t) = 0 = T(t)\,X(0)$$

the boundary conditions are difficult to satisfy because they are not homogeneous, i.e. with a zero right-hand side. Thus, the problem must be transformed to make the boundary conditions homogeneous. The solution is written as the sum of two functions, one of which satisfies the nonhomogeneous boundary conditions, whereas the other satisfies the homogeneous boundary conditions.

$$c(x,t) = f(x) + u(x,t)$$

$$u(0,t) = 0$$

$$u(L,t) = 0$$

Thus, $f(0) = 1$ and $f(L) = 0$ are necessary. Now the combined function satisfies the boundary conditions. In this case the function $f(x)$ can be taken as

$$f(x) = L - x$$

The equation for u is found by substituting for c in the original equation and noting that the $f(x)$ drops out for this case; it need not disappear in the general case:

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}$$

The boundary conditions for u are

$$u(0,t) = 0$$

$$u(L,t) = 0$$

The initial conditions for u are found from the initial condition

$$u(x,0) = c(x,0) - f(x) = \frac{x}{L} - 1$$

Separation of variables is now applied to this equation by writing

$$u(x,t) = T(t)\,X(x)$$

The same equation for $T(t)$ and $X(x)$ is obtained, but with $X(0) = X(L) = 0$.

$$\frac{d^2X}{dx^2} + \lambda X = 0$$

$$X(0) = X(L) = 0$$

Next $X(x)$ is solved for. The equation is an eigenvalue problem. The general solution is obtained by using e^{mx} and finding that $m^2 + \lambda = 0$; thus $m = \pm i\sqrt{\lambda}$. The exponential term

$$e^{\pm i\sqrt{\lambda}x}$$

is written in terms of sines and cosines, so that the general solution is

$$X = B\cos\sqrt{\lambda}x + E\sin\sqrt{\lambda}x$$

The boundary conditions are

$$X(L) = B\cos\sqrt{\lambda}L + E\sin\sqrt{\lambda}L = 0$$

$$X(0) = B = 0$$

If $B = 0$, then $E \neq 0$ is required to have any solution at all. Thus, λ must satisfy

$$\sin\sqrt{\lambda}L = 0$$

This is true for certain values of λ, called eigenvalues or characteristic values. Here, they are

$$\lambda_n = n^2\pi^2/L^2$$

Each eigenvalue has a corresponding eigenfunction

$$X_n(x) = E\sin n\pi x/L$$

The composite solution is then

$$X_n(x)\,T_n(t) = E\,A\sin\frac{n\pi x}{L}\,e^{-\lambda_n Dt}$$

This function satisfies the boundary conditions and differential equation but not the initial condition. To make the function satisfy the initial condition, several of these solutions are added up, each with a different eigenfunction, and $E\,A$ is replaced by A_n.

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L}\,e^{-n^2\pi^2 Dt/L^2}$$

The constants A_n are chosen by making $u(x, t)$ satisfy the initial condition.

$$u(x,0) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L} = \frac{x}{L} - 1$$

The residual $R(x)$ is defined as the error in the initial condition:

$$R(x) = \frac{x}{L} - 1 - \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L}$$

Next, the Galerkin method is applied, and the residual is made orthogonal to a complete set of functions, which are the eigenfunctions.

$$\int_0^L \left(\frac{x}{L} - 1\right)\sin\frac{m\pi x}{L}\,dx$$

$$= \sum_{n=1}^{\infty} A_n \int_0^L \sin\frac{m\pi x}{L}\sin\frac{n\pi x}{L}\,dx = \frac{A_m}{2}$$

The Galerkin criterion for finding A_n is the same as the least-squares criterion [3, p. 183]. The solution is then

$$c(x,t) = 1 - \frac{x}{L} + \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L}\,e^{-n^2\pi^2 Dt/L^2}$$

This is an "exact" solution to the linear problem. It can be evaluated to any desired accuracy by taking more and more terms, but if a finite number of terms are used, some error always occurs. For large times a single term is adequate, whereas for small times many terms are needed. For small times the Laplace transform method is also useful, because it leads to solutions that converge with fewer terms. For small times, the method of combination of variables may be used

as well. For nonlinear problems, the method of separation of variables fails and one of the other methods must be used.

The method of combination of variables is useful, particularly when the problem is posed in a semi-infinite domain. Here, only one example is provided; more detail is given in [3, 113, 114]. The method is applied here to the nonlinear problem

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left[D(c)\frac{\partial c}{\partial x}\right] = D(c)\frac{\partial^2 c}{\partial x^2} + \frac{\mathrm{d}\,D(c)}{\mathrm{d}c}\left(\frac{\partial c}{\partial x}\right)^2$$

with boundary and initial conditions

$$c(x,0) = 0$$

$$c(0,t) = 1,\ c(\infty,t) = 0$$

The transformation combines two variables into one

$$c(x,t) = f(\eta) \text{ where } \eta = \frac{x}{\sqrt{4D_0 t}}$$

The use of the 4 and D_0 makes the analysis below simpler. The equation for c (x, t) is transformed into an equation for f (η)

$$\frac{\partial c}{\partial t} = \frac{\mathrm{d}f}{\mathrm{d}\eta}\frac{\partial \eta}{\partial t},\ \frac{\partial c}{\partial x} = \frac{\mathrm{d}f}{\mathrm{d}\eta}\frac{\partial \eta}{\partial x}$$

$$\frac{\partial^2 c}{\partial x^2} = \frac{\mathrm{d}^2 f}{\mathrm{d}\eta^2}\left(\frac{\partial \eta}{\partial x}\right)^2 + \frac{\mathrm{d}f}{\mathrm{d}\eta}\frac{\partial^2 \eta}{\partial x^2}$$

$$\frac{\partial \eta}{\partial t} = -\frac{x/2}{\sqrt{4D_0 t^3}},\ \frac{\partial \eta}{\partial x} = \frac{1}{\sqrt{4D_0 t}},\ \frac{\partial^2 \eta}{\partial x^2} = 0$$

The result is

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\left[K(c)\frac{\mathrm{d}f}{\mathrm{d}\eta}\right] + 2\eta\frac{\mathrm{d}f}{\mathrm{d}\eta} = 0$$

$$K(\mathrm{c}) = D(c)/D_0$$

The boundary conditions must also combine. In this case the variable η is infinite when either x is infinite or t is zero. Note that the boundary conditions on c (x, t) are both zero at those points. Thus, the boundary conditions can be combined to give

$$f(\infty) = 0$$

The other boundary condition is for $x = 0$ or $\eta = 0$,

$$f(0) = 1$$

Thus, an ordinary differential equation must be solved rather than a partial differential equation. When the diffusivity is constant the solution is the well-known complementary error function:

$$c(x,t) = 1 - \mathrm{erf}\,\eta = \mathrm{erfc}\,\eta$$

$$\mathrm{erf}\,\eta = \frac{\int_0^{\eta} \mathrm{e}^{-\xi^2}\mathrm{d}\xi}{\int_0^{\infty} \mathrm{e}^{-\xi^2}\mathrm{d}\xi}$$

This is a tabulated function [23].

Numerical Methods. Numerical methods are applicable to both linear and nonlinear problems on finite and semi-infinite domains. The *finite difference method* is applied by using the method of lines [115]. In this method the same equations are used for the spatial variations of the function, but the function at a grid point can vary with time. Thus the linear diffusion problem is written as

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = D\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} \tag{37}$$

This can be written in the general form

$$\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{A}c$$

This set of ordinary differential equations can be solved by using any of the standard methods. The stability of explicit schemes is deduced from the theory presented in Chapter 6. The equations are written as

$$\frac{\mathrm{d}c_i}{\mathrm{d}t} = D\frac{c_{i+1} - 2c_i + c_{i-1}}{\Delta x^2} = \frac{D}{\Delta x^2}\sum_{j=1}^{n+1} B_{ij}c_j$$

where the matrix $\boldsymbol{B}$ is tridiagonal. The stability of the integration of these equations is governed by the largest eigenvalue of $\boldsymbol{B}$. If Euler's method is used for integration,

$$\Delta t\frac{D}{\Delta x^2} \le \frac{2}{|\lambda|_{\max}}$$

The largest eigenvalue of $\boldsymbol{B}$ is bounded by the Gerschgorin theorem [14, p. 135].

$$|\lambda|_{\max} \le \max_{2<j<n}\sum_{i=2}^{n}|B_{ji}| = 4$$

This gives the well-known stability limit

$$\Delta t\frac{D}{\Delta x^2} \le \frac{1}{2}$$

If other methods are used to integrate in time, then the stability limit changes according to the method. It is interesting to note that the eigenvalues of Equation 37 range from $D\,\pi^2/L^2$ (smallest) to $4\,D/\Delta x^2$ (largest), depending on

the boundary conditions. Thus the problem becomes stiff as Δx approaches zero [3, p. 263].

Implicit methods can also be used. Write a finite difference form for the time derivative and average the right-hand sides, evaluated at the old and new times:

$$\frac{c_i^{n+1}-c_i^n}{\Delta t} = D\,(1-\theta)\,\frac{c_{i+1}^n-2c_i^n+c_{i-1}^n}{\Delta x^2}$$
$$+D\theta\frac{c_{i+1}^{n+1}-2c_i^{n+1}+c_{i-1}^{n+1}}{\Delta x^2}$$

Now the equations are of the form

$$-\frac{D\Delta t\theta}{\Delta x^2}c_{i+1}^{n+1}+\left[1+2\frac{D\Delta t\theta}{\Delta x^2}\right]c_i^{n+1}-\frac{D\Delta t\theta}{\Delta x^2}c_{i-1}^{n+1}$$
$$=c_i^n+\frac{D\Delta t(1-\theta)}{\Delta x^2}\left(c_{i+1}^n-2c_i^n+c_{i-1}^n\right)$$

and require solving a set of simultaneous equations, which have a tridiagonal structure. Using $\theta = 0$ gives the Euler method (as above); $\theta = 0.5$ gives the Crank – Nicolson method; $\theta = 1$ gives the backward Euler method. The stability limit is given by

$$\frac{D\Delta t}{\Delta x^2}\leq\frac{0.5}{1-2\theta}$$

whereas the oscillation limit is given by

$$\frac{D\Delta t}{\Delta x^2}\leq\frac{0.25}{1-\theta}$$

If a time step is chosen between the oscillation limit and stability limit, the solution will oscillate around the exact solution, but the oscillations remain bounded. For further discussion, see [3, p. 218].

Finite volume methods are utilized extensively in computational fluid dynamics. In this method, a mass balance is made over a cell accounting for the change in what is in the cell and the flow in and out. Figure 30 illustrates the geometry of the *i*-th cell. A mass balance made on this cell (with area A perpendicular to the paper) is

$$A\Delta x(c_i^{n+1}-c_i^n) = \Delta tA(J_{j-1/2}-J_{i+1/2})$$

where J is the flux due to convection and diffusion, positive in the $+x$ direction.

$$J = uc-D\frac{\partial c}{\partial x},\, J_{i-1/2} = u_{i-1/2}c_{i-1/2}-D\frac{c_i-c_{i-1/2}}{\Delta x}$$

The concentration at the edge of the cell is taken as

$$c_{i-1/2} = \frac{1}{2}(c_i+c_{i-1})$$

Rearrangement for the case when the velocity u is the same for all nodes gives

$$\frac{c_i^{n+1}-c_i^n}{\Delta t}+\frac{u(c_{i+1}-c_{i-1})}{2\Delta x} = \frac{D}{\Delta x^2}(c_{i+1}-2c_i+c_{i-1})$$

This is the same equation as obtained using the finite difference method. This is not always true, and the finite volume equations are easy to derive. In two- and three-dimensions, the mesh need not be rectangular, as long as it is possible to compute the velocity normal to an edge of the cell. The finite volume method is useful for applications involving filling, such as injection molding, when only part of the cell is filled with fluid. Such applications do involve some approximations, since the interface is not tracked precisely, but they are useful engineering approximations.

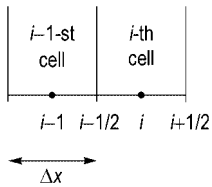

Figure 30.

The *finite element method* is handled in a similar fashion, as an extension of two-point boundary value problems by letting the solution at the nodes depend on time. For the diffusion equation the finite element method gives

$$\sum\nolimits_e\sum\nolimits_I C_{JI}^e\frac{\mathrm{d}c_I^e}{\mathrm{d}t} = \sum\nolimits_e\sum\nolimits_I B_{JI}^e\,c_I^e$$

with the mass matrix defined by

$$C_{JI}^e = \Delta x_e\int_0^1 N_J\,(u)\,N_I\,(u)\,\mathrm{d}u$$

This set of equations can be written in matrix form

$$\boldsymbol{CC}\frac{\mathrm{d}c}{\mathrm{d}t} = \boldsymbol{AA}c$$

Now the matrix $\boldsymbol{C}\,\boldsymbol{C}$ is not diagonal, so that a set of equations must be solved for each time step, even when the right-hand side is evaluated explicitly. This is not as time-consuming as it seems, however. The explicit scheme is written as

$$CC_{ji}\frac{c_i^{n+1}-c_i^n}{\Delta t}=AA_{ji}c_i^n$$

and rearranged to give

$$CC_{ji}\left(c_i^{n+1}-c_i^n\right)=\Delta tAA_{ji}c_i^n \text{ or}$$

$$\boldsymbol{CC}\left(\boldsymbol{c}^{n+1}-\boldsymbol{c}^n\right)=\Delta t\boldsymbol{AAc}$$

This is solved with an $L\,U$ decomposition (see Section 1.1) that retains the structure of the mass matrix $\boldsymbol{C}\,\boldsymbol{C}$. Thus,

$$\boldsymbol{CC}=\boldsymbol{LU}$$

At each step, calculate

$$\boldsymbol{c}^{n+1}-\boldsymbol{c}^n=\Delta t\boldsymbol{U}^{-1}\boldsymbol{L}^{-1}\boldsymbol{AAc}^n$$

This is quick and easy to do because the inverse of $\boldsymbol{L}$ and $\boldsymbol{U}$ are simple. Thus the problem is reduced to solving one full matrix problem and then evaluating the solution for multiple right-hand sides. For implicit methods the same approach can be used, and the LU decomposition remains fixed until the time step is changed.

The *method of orthogonal collocation* uses a similar extension: the same polynomial of x is used but now the coefficients depend on time.

$$\left.\frac{\partial c}{\partial t}\right|_{x_j}=\frac{\mathrm{d}c\left(x_j,t\right)}{\mathrm{d}t}=\frac{\mathrm{d}c_j}{\mathrm{d}t}$$

Thus, for diffusion problems

$$\frac{\mathrm{d}c_j}{\mathrm{d}t}=\sum_{i=1}^{N+2}B_{ji}c_i,\ j=2,\ldots,N+1$$

This can be integrated by using the standard methods for ordinary differential equations as initial value problems. Stability limits for explicit methods are available [3, p. 204].

The method of orthogonal collocation on finite elements can also be used, and details are provided elsewhere [3, pp. 228 – 230].

The maximum eigenvalue for all the methods is given by

$$|\lambda|_{\max}=\frac{LB}{\Delta x^2} \tag{38}$$

where the values of LB are as follows:

Finite difference	4
Galerkin, linear elements, lumped	4
Galerkin, linear elements	12
Galerkin, quadratic elements	60
Orthogonal collocation on finite elements, cubic	36

Spectral methods employ the discrete Fourier transform (see 2 and Chebyshev polynomials on rectangular domains [116]).

In the Chebyshev collocation method, $N+1$ collocation points are used

$$x_j=\cos\frac{\pi j}{N},\ j=0,1,\ldots,N$$

As an example, consider the equation

$$\frac{\partial u}{\partial t}+f\left(u\right)\frac{\partial u}{\partial x}=0$$

An explicit method in time can be used

$$\frac{u^{n+1}-u^n}{\Delta t}+f\left(u^n\right)\left.\frac{\partial u}{\partial x}\right|^n=0$$

and evaluated at each collocation point

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+f\left(u_j^n\right)\left.\frac{\partial u}{\partial x}\right|_j^n=0$$

The trial function is taken as

$$u_j\left(t\right)=\sum_{p=0}^{N}a_p\left(t\right)\cos\frac{\pi pj}{N},\ u_j^n=u_j\left(t^n\right) \tag{39}$$

Assume that the values u_j^n exist at some time. Then invert Equation 39 using the fast Fourier transform to obtain $\{a_p\}$ for $p=0,1,\ldots,N$; then calculate S_p

$$S_p=S_{p+2}+\left(p+1\right)a_{p+1},\ 0\le p\le N-1$$

$$S_N=0,\ S_{N+1}=0$$

and finally

$$a_p^{(1)}=\frac{2S_p}{c_p}$$

Thus, the first derivative is given by

$$\left.\frac{\partial u}{\partial x}\right|_j=\sum_{p=0}^{N}a_p^{(1)}\left(t\right)\cos\frac{\pi pj}{N}$$

This is evaluated at the set of collocation points by using the fast Fourier transform again. Once the function and the derivative are known at each collocation point the solution can be advanced forward to the $n+1$-th time level.

The advantage of the spectral method is that it is very fast and can be adapted quite well to parallel computers. It is, however, restricted in the geometries that can be handled.

8.4. Elliptic Equations

Elliptic equations can be solved with both finite difference and finite element methods. One-dimensional elliptic problems are two-point boundary value problems and are covered in Chapter 7. Two-dimensional elliptic problems are often solved with direct methods, and iterative methods are usually used for three-dimensional problems. Thus, two aspects must be considered: how the equations are discretized to form sets of algebraic equations and how the algebraic equations are then solved.

The prototype elliptic problem is steady-state heat conduction or diffusion,

$$k\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right)=Q$$

possibly with a heat generation term per unit volume, Q. The boundary conditions can be

Dirichlet or 1st kind:	$T = T_1$ on boundary S_1
Neumann or 2nd kind:	$k\frac{\partial T}{\partial n}=q_2$ on boundary S_2
Robin, mixed, or 3rd kind:	$-k\frac{\partial T}{\partial n}=h\left(T-T_3\right)$ on boundary S_3

Illustrations are given for constant physical properties k, h, while T_1, q_2, T_3 are known functions on the boundary and Q is a known function of position. For clarity, only a two-dimensional problem is illustrated. The finite difference formulation is given by using the following nomenclature

$$T_{i,j}=T\left(i\Delta x,j\Delta y\right)$$

The finite difference formulation is then

$$\frac{T_{i+1,j}-2T_{i,j}+T_{i-1,j}}{\Delta x^2}+\frac{T_{i,j+1}-2T_{i,j}+T_{i,j-1}}{\Delta y^2}=Q_{i,j}/k \tag{40}$$

$$T_{i,j}=T_1 \text{ for } i,j \text{ on boundary } S_1$$

$$k\frac{\partial T}{\partial n}\Big|_{i,j}=q_2 \text{ for } i,j \text{ on boundary } S_2$$

$$-k\frac{\partial T}{\partial n}\Big|_{i,j}=h\left(T_{i,j}-T_3\right) \text{ for } i,j \text{ on boundary } S_3$$

If the boundary is parallel to a coordinate axis the boundary slope is evaluated as in Chapter 7, by using either a one-sided, centered difference or a false boundary. If the boundary is more irregular and not parallel to a coordinate line, more complicated expressions are needed and the finite element method may be the better method.

Equation 40 is rewritten in the form

$$2\left(1+\frac{\Delta x^2}{\Delta y^2}\right)T_{i,j}=T_{i+1,j}+T_{i-1,j}+\frac{\Delta x^2}{\Delta y^2}\left(T_{i,j+1}+T_{i,j-1}\right)-\Delta x^2\frac{Q_{i,j}}{k}$$

And this is converted to the Gauss–Seidel iterative method.

$$2\left(1+\frac{\Delta x^2}{\Delta y^2}\right)T_{i,j}^{s+1}=T_{i+1,j}^{s}+T_{i-1,j}^{s+1}+\frac{\Delta x^2}{\Delta y^2}\left(T_{i,j+1}^{s}+T_{i,j-1}^{s+1}\right)-\Delta x^2\frac{Q_{i,j}}{k}$$

Calculations proceed by setting a low i, computing from low to high j, then increasing i and repeating the procedure. The relaxation method uses

$$2\left(1+\frac{\Delta x^2}{\Delta y^2}\right)T_{i,j}^{*}=T_{i+1,j}^{s}+T_{i-1,j}^{s+1}+\frac{\Delta x^2}{\Delta y^2}\left(T_{i,j+1}^{s}-T_{i,j-1}^{s+1}\right)-\Delta x^2\frac{Q_{i,j}}{k}$$

$$T_{i,j}^{s+1}=T_{i,j}^{s}+\beta\left(T_{i,j}^{*}-T_{i,j}^{s}\right)$$

If $\beta = 1$, this is the Gauss – Seidel method. If $\beta > 1$, it is overrelaxation; if $\beta < 1$, it is underrelaxation. The value of β may be chosen empirically, $0 < \beta < 2$, but it can be selected theoretically for simple problems like this [117, p. 100], [3, p. 282]. In particular, the optimal value of the iteration parameter is given by

$$-\ln\left(\beta_{\text{opt}}-1\right)\approx R$$

and the error (in solving the algebraic equation) is decreased by the factor $(1 - R)^N$ for every N iterations. For the heat conduction problem and Dirichlet boundary conditions,

$$R=\frac{\pi^2}{2n^2}$$

(when there are n points in both x and y directions). For Neumann boundary conditions, the value is

$$R=\frac{\pi^2}{2n^2}\frac{1}{1+\max\left[\Delta x^2/\Delta y^2,\Delta y^2/\Delta x^2\right]}$$

Iterative methods can also be based on lines (for 2D problems) or planes (for 3D problems).

Preconditioned conjugate gradient methods have been developed (see Chap. 1). In these methods a series of matrix multiplications are done iteration by iteration; and the steps lend

themselves to the efficiency available in parallel computers. In the multigrid method the problem is solved on several grids, each more refined than the previous one. In iterating between the solutions on different grids, one converges to the solution of the algebraic equations. A chemical engineering application is given in [118]. Software for a variety of these methods is available, as described below.

The Galerkin finite element method (FEM) is useful for solving elliptic problems and is particularly effective when the domain or geometry is irregular [119 – 125]. As an example, cover the domain with triangles and define a trial function on each triangle. The trial function takes the value 1.0 at one corner and 0.0 at the other corners, and is linear in between (see Fig. 31). These trial functions on each triangle are pieced together to give a trial function on the whole domain. For the heat conduction problem the method gives [3]

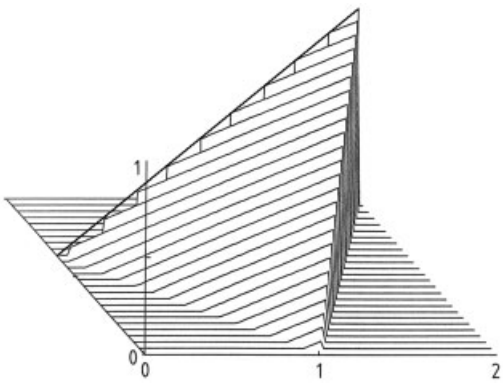

Figure 31. Finite elements trial function: linear polynomials on triangles

$$\sum_e \sum_J A^e_{IJ} T^e_J = \sum_e \sum_J F^e_I \tag{41}$$

where

$$A^e_{IJ} = -\int k \nabla N_I \cdot \nabla N_J \mathrm{d}A - \int_{C_3} h_3 N_I N_J \mathrm{d}C$$

$$F^e_I = \int N_I Q \mathrm{d}A + \int_{C_2} N_I q_2 \mathrm{d}C - \int_{C_3} N_I h_3 T_3 \mathrm{d}C$$

Also, a necessary condition is that

$$T_i = T_1 \text{ on } C_1$$

In these equations I and J refer to the nodes of the triangle forming element e and the summation is made over all elements. These equations represent a large set of linear equations, which are solved using matrix techniques (Chap. 1).

If the problem is nonlinear, e.g., with k or Q a function of temperature, the equations must be solved iteratively. The integrals are given for a triangle with nodes I, J, and K in counterclockwise order. Within an element,

$$T = N_I(x,y)\,T_I + N_J(x,y)\,T_J + N_K(x,y)\,T_K$$

$$N_I = \frac{a_I + b_I\,x + c_I\,y}{2\Delta}$$

$$a_I = x_J\,y_K - x_K\,y_J$$

$$b_I = y_I - y_K$$

$$c_I = x_K - x_J$$

plus permutation on I, K, J

$$2\Delta = \det \begin{bmatrix} 1 & x_I & y_I \\ 1 & x_J & y_J \\ 1 & x_K & y_K \end{bmatrix} = 2\,(\text{area of triangle})$$

$$a_I + a_J + a_K = 1$$

$$b_I + b_J + b_K = 0$$

$$c_I + c_J + c_K = 0$$

$$A^e_{IJ} = -\frac{k}{4\Delta}\,(b_I\,b_J + c_I\,c_J)$$

$$F^e_{IJ} = \frac{Q}{2}\,(a_I + b_I\,\overline{x} + c_I\,\overline{y}) = \frac{QD}{3}$$

$$\overline{x} = \frac{x_I + x_J + x_K}{3},\ \overline{y} = \frac{y_I + y_J + y_K}{3}$$

$$a_I + b_I\,\overline{x} + c_I\,\overline{y} = \frac{2}{3}\Delta$$

The trial functions in the finite element method are not limited to linear ones. Quadratic functions, and even higher order functions, are frequently used. The same considerations hold as for boundary value problems: the higher order trial functions converge faster but require more work. It is possible to refine both the mesh (h) and power of polynomial in the trial function (p) in an hp method. Some problems have constraints on some of the variables. If singularities exist in the solution, it is possible to include them in the basis functions and solve for the difference between the total solution and the singular function [126 – 129].

When applying the Galerkin finite element method, one must choose both the shape of the element and the basis functions. In two dimensions, triangular elements are usually used because it is easy to cover complicated geometries and refine parts of the mesh. However, rectangular elements sometimes have advantages, par-

ticularly when some parts of the solution do not change very much and the elements can be long. In three dimensions the same considerations apply: tetrahedral elements are frequently used, but brick elements are also possible. While linear elements (in both two and three dimensions) are usually used, higher accuracy can be obtained by using quadratic or cubic basis functions within the element. The reason is that all methods converge according to the mesh size to some power, and the power is larger when higher order elements are used. If the solution is discontinuous, or has discontinuous first derivatives, then the lowest order basis functions are used because the convergence is limited by the properties of the solution, not the finite element approximation.

One nice feature of the finite element method is the use of natural boundary conditions. In this problem the natural boundary conditions are the Neumann or Robin conditions. When using Equation 41, the problem can be solved on a domain that is shorter than needed to reach some limiting condition (such as at an outflow boundary). The externally applied flux is still applied at the shorter domain, and the solution inside the truncated domain is still valid. Examples are given in [107] and [131]. The effect of this is to allow solutions in domains that are smaller, thus saving computation time and permitting the solution in semi-infinite domains.

8.5. Parabolic Equations in Two or Three Dimensions

Computations become much more lengthy with two or more spatial dimensions, for example, the unsteady heat conduction equation

$$\varrho C_p \frac{\partial T}{\partial t} = k\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}\right)-Q$$

or the unsteady diffusion equation

$$\frac{\partial c}{\partial t} = D\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 c}{\partial y^2}\right)-R(c)$$

In the finite difference method an explicit technique would evaluate the right-hand side at the n-th time level:

$$\rho C_p \frac{T_{i,j}^{n+1}-T_{i,j}^{n}}{\Delta t}$$
$$= \frac{k}{\Delta x^2}\left(T_{i+1,j}^{n}-2T_{i-1,j}^{n}+T_{i-1,j}^{n}\right)$$
$$+\frac{k}{\Delta y^2}\left(T_{i,j+1}^{n}-2T_{i,j}^{n}+T_{i,j-1}^{n}\right)-Q$$

When $Q = 0$ and $\Delta x = \Delta y$, the time step is limited by

$$\Delta t \leq \frac{\Delta x^2 \varrho C_p}{4k} \text{ or } \frac{\Delta x^2}{4D}$$

These time steps are smaller than for one-dimensional problems. For three dimensions, the limit is

$$\Delta t \leq \frac{\Delta x^2}{6D}$$

To avoid such small time steps, which must be smaller when Δx decreases, an implicit method could be used. This leads to large sparse matrices, rather than convenient tridiagonal matrices.

8.6. Special Methods for Fluid Mechanics

The method of operator splitting is also useful when different terms in the equation are best evaluated by using different methods or as a technique for reducing a larger problem to a series of smaller problems. Here the method is illustrated by using the Navier – Stokes equations. In vector notation the equations are

$$\varrho\frac{\partial \boldsymbol{u}}{\partial t}+\varrho\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u} = \varrho\boldsymbol{f}-\boldsymbol{\nabla}p+\mu\nabla^2\boldsymbol{u}$$

The equation is solved in the following steps

$$\varrho\frac{\boldsymbol{u}^*-\boldsymbol{u}^n}{\Delta t} = -\varrho\boldsymbol{u}^n\cdot\boldsymbol{\nabla}\boldsymbol{u}^n+\varrho\boldsymbol{f}+\mu\nabla^2\boldsymbol{u}^n$$

$$\nabla^2 p^{n+1} = \frac{1}{\Delta t}\boldsymbol{\nabla}\cdot\boldsymbol{u}^*$$

$$\varrho\frac{\boldsymbol{u}^{n+1}-\boldsymbol{u}^*}{\Delta t} = -\boldsymbol{\nabla}p$$

This can be done by using the finite difference [132, p. 162] or the finite element method [133 – 135].

In fluid flow problems solved with the finite element method, the basis functions for pressure and velocity are often different. This is required by the LBB condition (named after Ladyshenskaya, Brezzi, and Babuska) [134, 135]. Sometimes a discontinuous basis function is used for pressure to meet this condition. Other times a

penalty term is added, or the quadrature is done using a small number of quadrature points. Thus, one has to be careful how to apply the finite element method to the Navier – Stokes equations. Fortunately, software exists that has taken this into account.

Level Set Methods. Multiphase problems are complicated because the terms in the equations depend on which phase exists at a particular point, and the phase boundary may move or be unknown. It is desirable to compute on a fixed grid, and the level set formulation allows this. Consider a curved line in a two-dimensional problem or a curved surface in a three-dimensional problem. One defines a level set function ϕ, which is the signed distance function giving the distance from some point to the closest point on the line or surface. It defined to be negative on one side of the interface and positive on the other. Then the curve

$$\phi(x,y,z) = 0$$

represents the location of the interface. For example, in flow problems the level set function is defined as the solution to

$$\frac{\partial \phi}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla} \phi = 0$$

The physics governing the velocity of the interface must be defined, and this equation is solved along with the other equations representing the problem [130 – 137].

Lattice Boltzmann Methods. Another way to solve fluid flow problems is based on a molecular viewpoint and is called the Lattice Boltzmann method [138 – 141]. The treatment here follows [142]. A lattice is defined, and one solves the following equation for $f_i\,(x,t)$, the probability of finding a molecule at the point x with speed c_i.

$$\frac{\partial f_i}{\partial t} + c_i \cdot \boldsymbol{\nabla} f_i = -\frac{f_i - f_i^{\mathrm{eq}}}{\tau}$$

The right-hand side represents a single time relaxation for molecular collisions, and τ is related to the kinematic viscosity. By means of a simple stepping algorithm, the computational algorithm is

$$f_i(\mathbf{x} + \mathbf{c}_i \Delta t, t + \Delta t) = f_i(\mathbf{x}, t) - \frac{\Delta t}{\tau}(f_i - f_i^{\mathrm{eq}})$$

Consider the lattice shown in Figure 32. The various velocities are

$$c_1 = (1,0), c_3 = (0,1), c_5 = (-1,0), c_7 = (0,-1)$$

$$c_2 = (1,1), c_4 = (-1,1), c_6 = (-1,-1), c_8 = (1,-1)$$

$$c_0 = (0,0)$$

The density, velocity, and shear stress (for some k) are given by

$$\rho = \sum\nolimits_i f_i(\boldsymbol{x},t), \rho \mathbf{u} = \sum\nolimits_i \boldsymbol{c}_i f_i(\boldsymbol{x},t),$$

$$\tau = k \sum\nolimits_i \boldsymbol{c}_i \boldsymbol{c}_i \left[f_i(\boldsymbol{x},t) - f_i^{\mathrm{eq}}(\boldsymbol{x},t) \right]$$

For this formulation, the kinematic viscosity and speed of sound are given by

$$\nu = \frac{1}{3}(\tau - \frac{1}{2}), c_s = \frac{1}{3}$$

The equilibrium distribution is

$$f_i^{\mathrm{eq}} = \rho w_i \left[1 + \frac{\boldsymbol{c}_i \cdot \mathbf{u}}{c_s^2} + \frac{(\boldsymbol{c}_i \cdot \mathbf{u})^2}{2c_s^4} - \frac{\mathbf{u} \cdot \mathbf{u}}{2c_s^2} \right]$$

where the weighting functions are

$$w_0 = \frac{4}{9}, w_1 = w_3 = w_5 = w_7 = \frac{1}{9},$$
$$w_2 = w_4 = w_6 = w_8 = \frac{1}{36}$$

With these conditions, the solution for velocity is a solution of the Navier—Stokes equation. These equations lead to a large computational problem, but it can be solved by parallel processing on multiple computers.

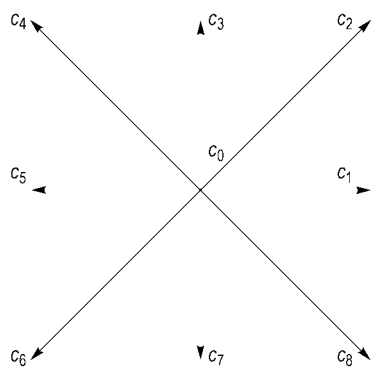

Figure 32.

8.7. Computer Software

A variety of general-purpose computer programs are available commercially. Mathematica (http://www.wolfram.com/), Maple (http://www.maplesoft.com/) and Mathcad (http://www.mathcad.com/) all have the capability of doing symbolic manipulation so that algebraic solutions can be obtained. For example, Mathematica can solve some ordinary and partial differential equations analytically; Maple can make simple graphs and do linear algebra and simple computations, and Mathcad can do simple calculations. In this section, examples are given for the use of Matlab (http://www.mathworks.com/), which is a package of numerical analysis tools, some of which are accessed by simple commands, and some of which are accessed by writing programs in C. Spreadsheets can also be used to solve simple problems. A popular program used in chemical engineering education is Polymath (http://www.polymath-software.com/), which can numerically solve sets of linear or nonlinear equations, ordinary differential equations as initial value problems, and perform data analysis and regression.

The mathematical methods used to solve partial differential equations are described in more detail in [143 – 148]. Since many computer programs are available without cost, consider the following decision points. The first decision is whether to use an approximate, engineering flow model, developed from correlations, or to solve the partial differential equations that govern the problem. Correlations are quick and easy to apply, but they may not be appropriate to your problem, or give the needed detail. When using a computer package to solve partial differential equations, the first task is always to generate a mesh covering the problem domain. This is not a trivial task, and special methods have been developed to permit importation of a geometry from a computer-aided design program. Then, the mesh must be created automatically. If the boundary is irregular, the finite element method is especially well-suited, although special embedding techniques can be used in finite difference methods (which are designed to be solved on rectangular meshes). Another capability to consider is the ability to track free surfaces that move during the computation. This phenomenon introduces the same complexity that occurs in problems with a large Peclet number, with the added difficulty that the free surface moves between mesh points, and improper representation can lead to unphysical oscillations. The method used to solve the equations is important, and both explicit and implicit methods (as described above) can be used. Implicit methods may introduce unacceptable extra diffusion, so the engineer needs to examine the solution carefully. The methods used to smooth unphysical oscillations from node to node are also important, and the engineer needs to verify that the added diffusion or smoothing does not give inaccurate solutions. Since current-day problems are mostly nonlinear, convergence is always an issue since the problems are solved iteratively. Robust programs provide several methods for convergence, each of which is best in some circumstance or other. It is wise to have a program that includes many iterative methods. If the iterative solver is not very robust, the only recourse to solving a steady-state problem may be to integrate the time-dependent problem to steady state. The solution time may be long, and the final result may be further from convergence than would be the case if a robust iterative solver were used.

A variety of computer programs is available on the internet, some of them free. First consider general-purpose programs. On the NIST web page, http://gams.nist.gov/ choose "problem decision tree", and then "differential and integral equations", then "partial differential equations". The programs are organized by type of problem (elliptic, parabolic, and hyperbolic) and by the number of spatial dimensions (one or more than one). On the Netlib web site, http://www.netlib.org/, search on "partial differential equation". The website: http://software.sandia.gov has a variety of programs available. Lau [141, 145] provides many programs in C++ (also see http://www.nr.com/). The multiphysics program Comsol Multiphysics (formerly FEMLAB) also solves many standard equations arising in Mathematical Physics.

Computational fluid dynamics (CFD) programs are more specialized, and most of them have been designed to solve sets of equations that are appropriate to specific industries. They can then include approximations and correlations for some features that would be difficult to solve for directly. Four widely used major packages are

Fluent (http://www.fluent.com/), CFX (now part of ANSYS), Comsol Multiphysics (formerly FEMLAB) (http://www.comsol.com/), and ANSYS (http://www.ansys.com/). Of these, Comsol Multiphysics is particularly useful because it has a convenient graphical user interface, permits easy mesh generation and refinement (including adaptive mesh refinement), allows the user to add in phenomena and additional equations easily, permits solution by continuation methods (thus enhancing convergence), and has extensive graphical output capabilities. Other packages are also available (see http://cfd-online.com/), and these may contain features and correlations specific to the engineer's industry. One important point to note is that for turbulent flow, all the programs contain approximations, using the k-epsilon models of turbulence, or large eddy simulations; the direct numerical simulation of turbulence is too slow to apply it to very big problems, although it does give insight (independent of any approximations) that is useful for interpreting turbulent phenomena. Thus, the method used to include those turbulent correlations is important, and the method also may affect convergence or accuracy.

9. Integral Equations [149 – 155]

If the dependent variable appears under an integral sign an equation is called an integral equation; if derivatives of the dependent variable appear elsewhere in the equation it is called an integrodifferential equation. This chapter describes the various classes of equations, gives information concerning Green's functions, and presents numerical methods for solving integral equations.

9.1. Classification

Volterra integral equations have an integral with a variable limit, whereas Fredholm integral equations have a fixed limit. Volterra equations are usually associated with initial value or evolutionary problems, whereas Fredholm equations are analogous to boundary value problems. The terms in the integral can be unbounded, but still yield bounded integrals, and these equations are said to be weakly singular. A *Volterra equation of the second kind* is

$$y(t) = g(t) + \lambda \int_a^t K(t,s) y(s)\,\mathrm{d}s \tag{42}$$

whereas a *Volterra equation of the first kind* is

$$y(t) = \lambda \int_a^t K(t,s)\, y(s)\,\mathrm{d}s$$

Equations of the first kind are very sensitive to solution errors so that they present severe numerical problems.

An example of a problem giving rise to a Volterra equation of the second kind is the following heat conduction problem:

$$\varrho C_p \frac{\partial T}{\partial t} = k \frac{\partial^2 T}{\partial x^2},\ 0 \le x < \infty,\ t > 0$$

$$T(x,0) = 0,\ \frac{\partial T}{\partial x}(0,t) = -g(t),$$

$$\lim_{x\to\infty} T(x,t) = 0,\ \lim_{x\to\infty} \frac{\partial T}{\partial x} = 0$$

If this is solved by using Fourier transforms the solution is

$$T(x,t) = \frac{1}{\sqrt{\pi}} \int_0^t g(s) \frac{1}{\sqrt{t-s}} \mathrm{e}^{-x^2/4(t-s)}\mathrm{d}s$$

Suppose the problem is generalized so that the boundary condition is one involving the solution T, which might occur with a radiation boundary condition or heat-transfer coefficient. Then the boundary condition is written as

$$\frac{\partial T}{\partial x} = -G(T,t),\ x = 0,\ t > 0$$

The solution to this problem is

$$T(x,t) = \frac{1}{\sqrt{\pi}} \int_0^t G(T(0,s),s) \frac{1}{\sqrt{t-s}}\, \mathrm{e}^{-x^2/4(t-s)}\mathrm{d}s$$

If $T^*(t)$ is used to represent $T(0, t)$, then

$$T^*(t) = \frac{1}{\sqrt{\pi}} \int_0^t G(\boldsymbol{T}(s),s) \frac{1}{\sqrt{t-s}}\mathrm{d}s$$

Thus the behavior of the solution at the boundary is governed by an integral equation. Nagel and Kluge [156] use a similar approach to solve for adsorption in a porous catalyst.

The existence and uniqueness of the solution can be proved [151, p. 30, 32].

Sometimes the kernel is of the form

$$K(t,s) = K(t-s)$$

Equations of this form are called convolution equations and can be solved by taking the Laplace transform. For the integral equation

$$Y(t) = G(t) + \lambda \int_0^t K(t-\tau)\, Y(\tau)\, \mathrm{d}\tau$$

$$K(t) * Y(t) \equiv \int_0^t K(t-\tau)\, Y(\tau)\, \mathrm{d}\tau$$

the Laplace transform is

$$y(s) = g(s) + k(s)\, y(s)$$

$$k(s)\, y(s) = \boldsymbol{L}[K(t) * Y(t)]$$

Solving this for $y(s)$ gives

$$y(s) = \frac{g(s)}{1-k(s)}$$

If the inverse transform can be found, the integral equation is solved.

A *Fredholm equation of the second kind* is

$$y(x) = g(x) + \lambda \int_a^b K(x,s)\, y(s)\, \mathrm{d}s \tag{43}$$

whereas a *Fredholm equation of the first kind* is

$$\int_a^b K(x,s)\, y(s)\, \mathrm{d}s = g(x)$$

The limits of integration are fixed, and these problems are analogous to boundary value problems. An eigenvalue problem is a homogeneous equation of the second kind.

$$y(x) = \lambda \int_a^b K(x,s)\, y(s)\, \mathrm{d}s \tag{44}$$

Solutions to this problem occur only for specific values of λ, the eigenvalues. Usually the Fredholm equation of the second or first kind is solved for values of λ different from these, which are called regular values.

Nonlinear Volterra equations arise naturally from initial value problems. For the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = F(t, y(t))$$

both sides can be integrated from 0 to t to obtain

$$y(t) = y(0) + \int_0^t F(s, y(s))\, \mathrm{d}s$$

which is a nonlinear Volterra equation. The general nonlinear Volterra equation is

$$y(t) = g(t) + \int_0^t K(t,s,y(s))\, \mathrm{d}s \tag{45}$$

Theorem [151, *p. 55].* If $g(t)$ is continuous, the kernel $K(t, s, y)$ is continuous in all variables and satisfies a Lipschitz condition

$$|K(t,s,y) - K(t,s,z)| \le L\, |y-z|$$

then the nonlinear Volterra equation has a unique continuous solution.

A successive substitution method for its solution is

$$y_{n+1}(t) = g(t) + \int_0^t K[t,s,y_n(s)]\, \mathrm{d}s$$

Nonlinear Fredholm equations have special names. The equation

$$f(x) = \int_0^1 K[x,y,f(y)]\, \mathrm{d}y$$

is called the Urysohn equation [150 p. 208]. The special equation

$$f(x) = \int_0^1 K[x,y]\, F[y,f(y)]\, \mathrm{d}y$$

is called the Hammerstein equation [150, p. 209]. Iterative methods can be used to solve these equations, and these methods are closely tied to fixed point problems. A fixed point problem is

$$x = F(x)$$

and a successive substitution method is

$$x_{n+1} = F(x_n)$$

Local convergence theorems prove the process convergent if the solution is close enough to the answer, whereas global convergence theorems are valid for any initial guess [150, p. 229 – 231]. The successive substitution method for nonlinear Fredholm equations is

$$y_{n+1}(x) = \int_0^1 K[x,s,y_n(s)]\, \mathrm{d}s$$

Typical conditions for convergence include that the function satisfies a Lipschitz condition.

9.2. Numerical Methods for Volterra Equations of the Second Kind

Volterra equations of the second kind are analogous to initial value problems. An initial value problem can be written as a Volterra equation of the second kind, although not all Volterra equations can be written as initial value problems [151, p. 7]. Here the general nonlinear Volterra equation of the second kind is treated (Eq. 45). The simplest numerical method involves replacing the integral by a quadrature using the trapezoid rule.

$$y_n \equiv y(t_n) = g(t_n) + \Delta t \left\{ \frac{1}{2} K(t_n, t_0, y_0) + \sum_{i=1}^{n-1} K(t_n, t_i, y_i) + \frac{1}{2} K(t_n, t_n, y_n) \right\}$$

This equation is a nonlinear algebraic equation for y_n. Since y_0 is known it can be applied to solve for y_1, y_2, ... in succession. For a single integral equation, at each step one must solve a single nonlinear algebraic equation for y_n. Typically, the error in the solution to the integral equation is proportional to Δt^{μ}, and the power μ is the same as the power in the quadrature error [151, p. 97].

The stability of the method [151, p. 111] can be examined by considering the equation

$$y(t) = 1 - \lambda \int_0^t y(s)\,\mathrm{d}s$$

whose solution is

$$y(t) = \mathrm{e}^{-\lambda t}$$

Since the integral equation can be differentiated to obtain the initial value problem

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\lambda y,\ y(0) = 1$$

the stability results are identical to those for initial value methods. In particular, using the trapezoid rule for integral equations is identical to using this rule for initial value problems. The method is A-stable.

Higher order integration methods can also be used [151, p. 114, 124]. When the kernel is infinite at certain points, i.e., when the problem has a weak singularity, see [151, p. 71, 151].

9.3. Numerical Methods for Fredholm, Urysohn, and Hammerstein Equations of the Second Kind

Whereas Volterra equations could be solved from one position to the next, like initial value differential equations, Fredholm equations must be solved over the entire domain, like boundary value differential equations. Thus, large sets of equations will be solved and the notation is designed to emphasize that.

The methods are also based on quadrature formulas. For the integral

$$I(\varphi) = \int_a^b \varphi(y)\,\mathrm{d}y$$

a quadrature formula is written:

$$I(\varphi) = \sum_{i=0}^{n} w_i \varphi(y_i)$$

Then the integral Fredholm equation can be rewritten as

$$f(x) - \lambda \sum_{i=0}^{n} w_i K(x, y_i) f(y_i) = g(x), \quad a \le x \le b \tag{46}$$

If this equation is evaluated at the points $x = y_j$,

$$f(y_j) - \lambda \sum_{i=0}^{n} w_i K(y_j, y_i) f(y_i) = g(y_i)$$

is obtained, which is a set of linear equations to be solved for $\{f(y_j)\}$. The solution at any point is then given by Equation 46.

A common type of integral equation has a singular kernel along $x = y$. This can be transformed to a less severe singularity by writing

$$\int_a^b K(x,y) f(y)\,\mathrm{d}y = \int_a^b K(x,y)\,[f(y) - f(x)]\,\mathrm{d}y + \int_a^b K(x,y) f(x)\,\mathrm{d}y = \int_a^b K(x,y)\,[f(y) - f(x)]\,\mathrm{d}y + f(x) H(x)$$

where

$$H(x) = \int_a^b K(x,y) f(x)\,\mathrm{d}y$$

is a known function. The integral equation is then replaced by

$$f(x) = g(x) + \sum_{i=0}^{n} w_i K(x,y_i)\,[f(y_i) - f(x)] + f(x)\,H(x)$$

Collocation methods can be applied as well [149, p. 396]. To solve integral Equation 43 expand f in the function

$$f = \sum_{i=0}^{n} a_i \varphi_i(x)$$

Substitute f into the equation to form the residual

$$\sum_{i=0}^{n} a_i \varphi_i(x) - \lambda \sum_{i=0}^{n} a_i \int_a^b K(x,y)\,\varphi_i(y)\,\mathrm{d}y = g(x)$$

Evaluate the residual at the collocation points

$$\sum_{i=0}^{n} a_i \varphi_i(x_j) - \lambda \sum_{i=0}^{n} a_i \int_a^b K(x_j,y)\,\varphi_i(y)\,\mathrm{d}y = g(x_j)$$

The expansion can be in piecewise polynomials, leading to a collocation finite element method, or global polynomials, leading to a global approximation. If orthogonal polynomials are used then the quadratures can make use of the accurate Gaussian quadrature points to calculate the integrals. Galerkin methods are also possible [149, p. 406]. MILLS et al. [157] consider reaction – diffusion problems and say the choice of technique cannot be made in general because it is highly dependent on the kernel.

When the integral equation is nonlinear, iterative methods must be used to solve it. Convergence proofs are available, based on Banach's contractive mapping principle. Consider the Urysohn equation, with $g(x) = 0$ without loss of generality:

$$f(x) = \int_a^b F[x,y,f(y)]\,\mathrm{d}y$$

The kernel satisfies the Lipschitz condition

$$\max_{a \le x, y \le b} |F[x,y,f(y)] - F[x,z,f(z)]| \le K\,|y-z|$$

Theorem [150, p. 214]. If the constant K is $<$ 1 and certain other conditions hold, the successive substitution method

$$f_{n+1}(x) = \int_a^b F[x,y,f_n(y)]\,\mathrm{d}y, n = 0,1\ldots$$

converges to the solution of the integral equations.

9.4. Numerical Methods for Eigenvalue Problems

Eigenvalue problems are treated similarly to Fredholm equations, except that the final equation is a matrix eigenvalue problem instead of a set of simultaneous equations. For example,

$$\sum_{i=1}^{n} w_i K(y_i,y_i)\,f(y_i) = \lambda f(y_j),$$
$$i = 0,1,\ldots,n$$

leads to the matrix eigenvalue problem

$$\boldsymbol{K}\,\boldsymbol{D}\,\boldsymbol{f} = \lambda \boldsymbol{f}$$

Where $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = w_i$.

9.5. Green's Functions [158 – 160]

Integral equations can arise from the formulation of a problem by using Green's function. For example, the equation governing heat conduction with a variable heat generation rate is represented in differential forms as

$$\frac{\mathrm{d}^2 T}{\mathrm{d}x^2} = \frac{Q(x)}{k},\ T(0) = T(1) = 0$$

In integral form the same problem is [149, pp. 57 – 60]

$$T(x) = \frac{1}{k}\int_0^1 G(x,y)\,Q(y)\,\mathrm{d}y$$

$$G(x,y) = \begin{cases} -x(1-y) & x \le y \\ -y(1-x) & y \le x \end{cases}$$

Green's functions for typical operators are given below.

For the Poisson equation with solution decaying to zero at infinity

$$\nabla^2 \psi = -4\pi\varrho$$

the formulation as an integral equation is

$$\psi(\boldsymbol{r}) = \int_V \rho(\boldsymbol{r}_0)\,G(\boldsymbol{r},\boldsymbol{r}_0)\,\mathrm{d}V_0$$

where Green's function is [50, p. 891]

$G(\boldsymbol{r},\boldsymbol{r}_0) = \frac{1}{r}$ in three dimensions

$= -2\ln r$ in two dimensions

where $r = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}$ in three dimensions

and $r = \sqrt{(x-x_0)^2 + (y-x_0)^2}$ in two dimensions

For the problem

$$\frac{\partial u}{\partial t} = D\nabla^2 u,\ u = 0 \text{ on } S,$$

with a point source at x_0, y_0, z_0

Green's function is [44, p. 355]

$$u = \frac{1}{8[\pi D(t-\tau)]^{3/2}}$$

$$\mathrm{e}^{-\left[(x-x_0)^2+(y-y_0)^2+(z-z_0)^2\right]/4\,D(t-\tau)}$$

When the problem is

$$\frac{\partial c}{\partial t} = D\nabla^2 c$$

$$c = f\,(x,y,z) \text{ in } S \text{ at } t = 0$$

$$c = \varphi\,(x,y,z) \text{ on } S, t>0$$

the solution can be represented as [44, p. 353]

$$c = \int\int\int (u)_{\tau=0} f\,(x,y,z)\,\mathrm{d}x\mathrm{d}y\mathrm{d}z$$

$$+D\int_0^t\int\int\int \varphi\,(x,y,z,\tau)\,\frac{\partial u}{\partial n}\mathrm{d}S\mathrm{d}t$$

When the problem is two dimensional,

$$u = \frac{1}{\sqrt{4\pi D(t-\tau)}}\mathrm{e}^{-\left[(x-x_0)^2+(y-y_0)^2\right]/4D(t-\tau)}$$

$$c = \int\int (u)_{\tau=0} f\,(x,y)\,\mathrm{d}x\mathrm{d}y$$

$$+D\int_0^t\int \varphi\,(x,y,\tau)\,\frac{\partial u}{\partial n}\mathrm{d}C\mathrm{d}t$$

For the following differential equation and boundary conditions

$$\frac{1}{x^{a-1}}\frac{\mathrm{d}}{\mathrm{d}x}\left(x^{\mathrm{a}-1}\frac{\mathrm{d}c}{\mathrm{d}x}\right) = f\,[x,c\,(x)]\,,$$

$$\frac{\mathrm{d}c}{\mathrm{d}x}\,(0) = 0,\ \frac{2}{Sh}\frac{\mathrm{d}c}{\mathrm{d}x}\,(1) + c\,(1) = g$$

where *Sh* is the Sherwood number, the problem can be written as a Hammerstein integral equation:

$$c\,(x) = g - \int_0^1 G\,(x,y,Sh)\,f\,[y,c\,(y)]\,y^{a-1}\mathrm{d}y$$

Green's function for the differential operators are [163]

$a = 1$

$$G\,(x,y,Sh) = \begin{cases} 1+\frac{2}{Sh}-x, & y\le x \\ 1+\frac{2}{Sh}-y, & x<y \end{cases}$$

$a = 2$

$$G\,(x,y,Sh) = \begin{cases} \frac{2}{Sh}-\ln x, & y\le x \\ \frac{2}{Sh}-\ln y, & x<y \end{cases}$$

$a = 3$

$$G\,(x,y,Sh) = \begin{cases} \frac{2}{Sh}+\frac{1}{x}-1, & y\le x \\ \frac{2}{Sh}+\frac{1}{y}-1, & x<y \end{cases}$$

Green's functions for the reaction diffusion problem were used to provide computable error bounds by Ferguson and Finlayson [163].

If Green's function has the form

$$K\,(x,y) = \begin{cases} u\,(x)\,v\,(y) & 0\le y\le x \\ u\,(y)\,v\,(x) & x\le y\le 1 \end{cases}$$

the problem

$$f\,(x) = \int_0^1 K\,(x,y)\,F\,[y,f\,(y)]\,\mathrm{d}y$$

may be written as

$$f\,(x) - \int_0^x [u\,(x)\,v\,(y) - u\,(y)\,v\,(x)]\cdot$$

$$f\,[y,f\,(y)]\,\mathrm{d}y = \alpha v\,(x)$$

where

$$\alpha = \int_0^1 u\,(y)\,F\,[y,f\,(y)]\,\mathrm{d}y$$

Thus, the problem ends up as one directly formulated as a fixed point problem:

$$f = \Phi\,(f)$$

When the problem is the diffusion – reaction one, the form is

$$c\,(x) = g - \int_0^x [u\,(x)\,v\,(y) - u\,(y)\,v\,(x)]$$

$$f\,[y,c\,(y)]\,y^{a-1}\mathrm{d}y - \alpha v\,(x)$$

$$\alpha = \int_0^1 u\,(y)\,f\,[y,c\,(y)]\,y^{a-1}\mathrm{d}y$$

Dixit and Taularidis [164] solved problems involving Fischer – Tropsch synthesis reactions in a catalyst pellet using a similar method.

9.6. Boundary Integral Equations and Boundary Element Method

The boundary element method utilizes Green's theorem and integral equations. Here, the method is described briefly for the following boundary value problem in two or three dimensions

$$\nabla^2\varphi = 0,\ \varphi = f_1 \text{ on } S_1,\ \frac{\partial\varphi}{\partial n} = f_2 \text{ on } S_2$$

Green's theorem (see page 44) says that for any functions sufficiently smooth

$$\int_V \left(\varphi\nabla^2\psi - \psi\nabla^2\varphi\right) \mathrm{d}V = \int_S \left(\varphi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S$$

Suppose the function ψ satisfies the equation

$$\nabla^2\psi = 0$$

In two and three dimensions, such a function is

$$\psi = \ln r,\ r = \sqrt{(x-x_0)^2 + (y-y_0)^2}$$

in two dimensions

$$\psi = \tfrac{1}{r},\ r = \sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}$$

in three dimensions

where $\{x_0, y_0\}$ or $\{x_0, y_0, z_0\}$ is a point in the domain. The solution φ also satisfies

$$\nabla^2\varphi = 0$$

so that

$$\int_S \left(\varphi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S = 0$$

Consider the two-dimensional case. Since the function ψ is singular at a point, the integrals must be carefully evaluated. For the region shown in Figure 33, the domain is $S = S_1 + S_2$; a small circle of radius r_0 is placed around the point P at x_0, y_0. Then the full integral is

$$\int_S \left(\varphi\frac{\partial \ln r}{\partial n} - \ln r\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S + \int_{\theta=0}^{\theta=2\pi} \left(\varphi\frac{\partial \ln r_0}{\partial n} - \ln r_0\frac{\partial\varphi}{\partial n}\right) r_0 \mathrm{d}\theta = 0$$

As r_0 approaches 0,

$$\lim_{r_0\to 0} r_0 \ln r_0 = 0$$

and

$$\lim_{r_0\to 0} \int_{\theta=0}^{\theta=2\pi} \varphi\frac{\partial \ln r_0}{\partial n} r_0 \mathrm{d}\theta = -\varphi(P)\, 2\pi$$

Thus for an internal point,

$$\varphi(P) = \frac{1}{2\pi}\int_S \left(\varphi\frac{\partial \ln r}{\partial n} - \ln r\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S \tag{47}$$

If P is on the boundary, the result is [165, p. 464]

$$\varphi(P) = \frac{1}{\pi}\int_S \left(\varphi\frac{\partial \ln r}{\partial n} - \ln r\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S$$

Putting in the boundary conditions gives

$$\pi\varphi(P) = \int_{S_1} \left(f_1\frac{\partial \ln r}{\partial n} - \ln r\frac{\partial\varphi}{\partial n}\right) \mathrm{d}S + \int_{S_2} \left(\varphi\frac{\partial \ln r}{\partial n} - f_2 \ln r\right) \mathrm{d}S \tag{48}$$

This is an integral equation for φ on the boundary. Note that the order is one less than the original differential equation. However, the integral equation leads to matrices that are dense rather than banded or sparse, so some of the advantage of lower dimension is lost. Once this integral equation (Eq. 48) is solved to find φ on the boundary, it can be substituted in Equation 47 to find φ anywhere in the domain.

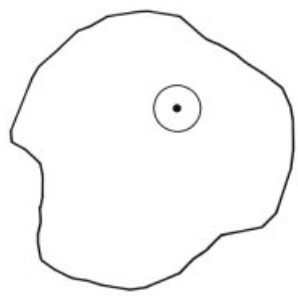

Figure 33. Domain with singularity at P

In the boundary finite element method, both the function and its normal derivative along the boundary are approximated.

$$\varphi = \sum_{j=1}^{N} \varphi_j N_j(\xi),\ \frac{\partial\varphi}{\partial n} = \sum_{j=1}^{N} \left(\frac{\partial\varphi}{\partial n}\right)_j N_j(\xi)$$

One choice of trial functions can be the piecewise constant functions shown in Figure 34. The integral equation then becomes

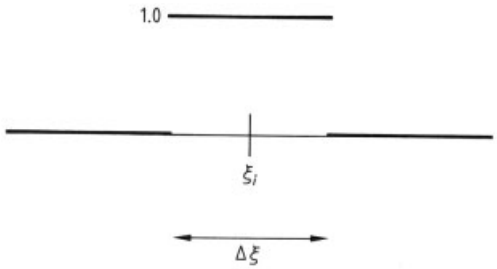

Figure 34. Trial function on boundary for boundary finite element method

$$\pi\varphi_i \sum_{j=1}^{N} \left[\varphi_j \int_{s_j} \frac{\partial \ln r_i}{\partial n} \mathrm{d}S - \left(\frac{\partial \varphi_j}{\partial n} \right) \int_{s_j} \ln r_i \right] \mathrm{d}s$$

The function φ_j is of course known along s_1, whereas the derivative $\partial\varphi_j/\partial n$ is known along s_2. This set of equations is then solved for φ_i and $\partial\varphi_i/\partial n$ along the boundary. This constitutes the boundary integral method applied to the Laplace equation.

If the problem is Poisson's equation

$$\nabla^2\varphi = g\,(x,y)$$

Green's theorem gives

$$\int_S \left(\varphi \frac{\partial \ln r}{\partial n} - \ln r \frac{\partial \varphi}{\partial n} \right) \mathrm{d}S + \int_A g \ln r \,\mathrm{d}A = 0$$

Thus, for an internal point,

$$2\pi\varphi\,(P) = \int_S \left(\varphi \frac{\partial \ln r}{\partial r} - \ln r \frac{\partial \varphi}{\partial n} \right) \mathrm{d}S + \int_A g \ln r \mathrm{d}A \tag{49}$$

and for a boundary point,

$$\pi\varphi\,(P) = \int_S \left(\varphi \frac{\partial \ln r}{\partial n} - \ln r \frac{\partial \varphi}{\partial n} \right) \mathrm{d}S + \int_A g \ln r \mathrm{d}A \tag{50}$$

If the region is nonhomogeneous this method can be used [165, p. 475], and it has been applied to heat conduction by Hsieh and Shang [166]. The finite element method can be applied in one region and the boundary finite element method in another region, with appropriate matching conditions [165, p. 478]. If the problem is nonlinear, then it is more difficult. For example, consider an equation such as Poisson's in which the function depends on the solution as well

$$\nabla^2\varphi = g\,(x,y,\varphi)$$

Then the integral appearing in Equation 50 must be evaluated over the entire domain, and the solution in the interior is given by Equation 49. For further applications, see [167] and [168].

10. Optimization

We provide a survey of systematic methods for a broad variety of optimization problems. The survey begins with a general classification of mathematical optimization problems involving continuous and discrete (or integer) variables. This is followed by a review of solution methods of the major types of optimization problems for continuous and discrete variable optimization, particularly nonlinear and mixed-integer nonlinear programming. In addition, we discuss direct search methods that do not require derivative information as well as global optimization methods. We also review extensions of these methods for the optimization of systems described by differential and algebraic equations.

10.1. Introduction

Optimization is a key enabling tool for decision making in chemical engineering [306]. It has evolved from a methodology of academic interest into a technology that continues to make significant impact in engineering research and practice. Optimization algorithms form the core tools for a) experimental design, parameter estimation, model development, and statistical analysis; b) process synthesis analysis, design, and retrofit; c) model predictive control and real-time optimization; and d) planning, scheduling, and the integration of process operations into the supply chain [307, 308].

As shown in Figure 35, optimization problems that arise in chemical engineering can be classified in terms of continuous and discrete variables. For the former, nonlinear programming (NLP) problems form the most general case, and widely applied specializations include linear programming (LP) and quadratic programming (QP). An important distinction for NLP is whether the optimization problem is convex or nonconvex. The latter NLP problem may

have multiple local optima, and an important question is whether a global solution is required for the NLP. Another important distinction is whether the problem is assumed to be differentiable or not.

Mixed-integer problems also include discrete variables. These can be written as mixed-integer nonlinear programs (MINLP) or as mixed-integer linear programs (MILP) if all variables appear linearly in the constraint and objective functions. For the latter an important case occurs when all the variables are integer; this gives rise to an integer programming (IP) problem. IPs can be further classified into many special problems (e.g., assignment, traveling salesman, etc.), which are not shown in Figure 35. Similarly, the MINLP problem also gives rise to special problem classes, although here the main distinction is whether its relaxation is convex or nonconvex.

The ingredients of formulating optimization problems include a mathematical model of the system, an objective function that quantifies a criterion to be extremized, variables that can serve as decisions, and, optionally, inequality constraints on the system. When represented in algebraic form, the general formulation of discrete/continuous optimization problems can be written as the following mixed-integer optimization problem:

$$\begin{aligned} \text{Min} \quad & f(x,y) \\ \text{s.t.} \quad & h(x,y) = 0 \\ & g(x,y) \leq 0 \\ & x \in \Re^n, y \in \{0,1\}^t \end{aligned} \tag{51}$$

where $f(x, y)$ is the objective function (e.g., cost, energy consumption, etc.), $h(x, y) = 0$ are the equations that describe the performance of the system (e.g., material balances, production rates), the inequality constraints $g(x, y) \leq 0$ can define process specifications or constraints for feasible plans and schedules, and s.t. denotes subject to. Note that the operator Max $f(x)$ is equivalent to Min $-f(x)$. We define the real n-vector x to represent the continuous variables while the t-vector y represents the discrete variables, which, without loss of generality, are often restricted to take 0–1 values to define logical or discrete decisions, such as assignment of equipment and sequencing of tasks. (These variables can also be formulated to take on other integer values as well.) Problem 51 corresponds to a mixed-integer nonlinear program (MINLP) when any of the functions involved are nonlinear. If all functions are linear it corresponds to a mixed-integer linear program (MILP). If there are no 0–1 variables, then problem 51 reduces to a nonlinear program 52 or linear program 65 depending on whether or not the functions are linear.

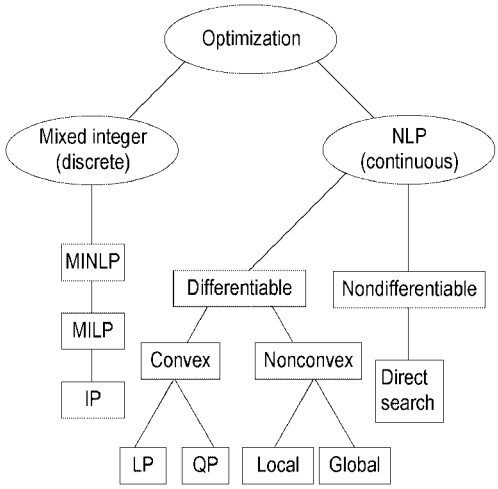

Figure 35. Classes of optimization problems and algorithms

We first start with continuous variable optimization and consider in the next section the solution of NLPs with differentiable objective and constraint functions. If only local solutions are required for the NLP, then very efficient large-scale methods can be considered. This is followed by methods that are not based on local optimality criteria; we consider direct search optimization methods that do not require derivatives, as well as deterministic global optimization methods. Following this, we consider the solution of mixed-integer problems and outline the main characteristics of algorithms for their solution. Finally, we conclude with a discussion of optimization modeling software and its implementation in engineering models.

10.2. Gradient-Based Nonlinear Programming

For continuous variable optimization we consider problem 51 without discrete variables y.

The general NLP problem 52 is presented below:

$$\begin{aligned} &\text{Min} && f(x) \\ &\text{s.t.} && h(x) = 0 \\ & && g(x) \leq 0 \end{aligned} \tag{52}$$

and we assume that the functions $f(x)$, $h(x)$, and $g(x)$ have continuous first and second derivatives. A key characteristic of problem 52 is whether the problem is convex or not, i.e., whether it has a convex objective function and a convex feasible region. A function $\phi(x)$ of x in some domain X is convex if and only if for all points $x_1, x_2 \in X$:

$$\phi[\alpha x_1 + (1-\alpha) x_2] \leq \alpha\phi(x_1) + (1-\alpha)\phi(x_2) \tag{53}$$

holds for all $\alpha \in (0, 1)$. If $\phi(x)$ is differentiable, then an equivalent definition is:

$$\phi(x_1) + \nabla\phi(x_1)^T (x - x_1) \leq \phi(x) \tag{54}$$

Strict convexity requires that the inequalities 53 and 54 be strict. Convex feasible regions require $g(x)$ to be a convex function and $h(x)$ to be linear. If 52 is a convex problem, than any local solution is guaranteed to be a global solution to 52. Moreover, if the objective function is strictly convex, then this solution x^* is unique. On the other hand, nonconvex problems may have multiple local solutions, i.e., feasible solutions that minimize the objective function within some neighborhood about the solution.

We first consider methods that find only local solutions to nonconvex problems, as more difficult (and expensive) search procedures are required to find a global solution. Local methods are currently very efficient and have been developed to deal with very large NLPs. Moreover, by considering the structure of convex NLPs (including LPs and QPs), even more powerful methods can be applied. To study these methods, we first consider conditions for local optimality.

Local Optimality Conditions – A Kinematic Interpretation. Local optimality conditions are generally derived from gradient information from the objective and constraint functions. The proof follows by identifying a local minimum point that has no feasible descent direction. Invoking a theorem of the alternative (e.g., Farkas – Lemma) leads to the celebrated Karush – Kuhn – Tucker (KKT) conditions [169]. Instead of a formal development of these conditions, we present here a more intuitive, kinematic illustration. Consider the contour plot of the objective function $f(x)$ given in Figure 36 as a smooth valley in space of the variables x_1 and x_2. For the contour plot of this unconstrained problem, Min $f(x)$, consider a ball rolling in this valley to the lowest point of $f(x)$, denoted by x^*. This point is at least a local minimum and is defined by a point with zero gradient and at least nonnegative curvature in all (nonzero) directions p. We use the first derivative (gradient) vector $\nabla f(x)$ and second derivative (Hessian) matrix $\nabla_{xx} f(x)$ to state the necessary first- and second-order conditions for unconstrained optimality:

$$\nabla_x f(x^*) = 0 \quad p^T \nabla_{xx} f(x^*) p \geq 0 \quad \text{for all } p \neq 0 \tag{55}$$

These necessary conditions for local optimality can be strengthened to sufficient conditions by making the inequality in the relations 55 strict (i.e., positive curvature in all directions). Equivalently, the sufficient (necessary) curvature conditions can be stated as: $\nabla_{xx} f(x^*)$ has all positive (nonnegative) eigenvalues and is therefore defined as a positive (semi-)definite matrix.

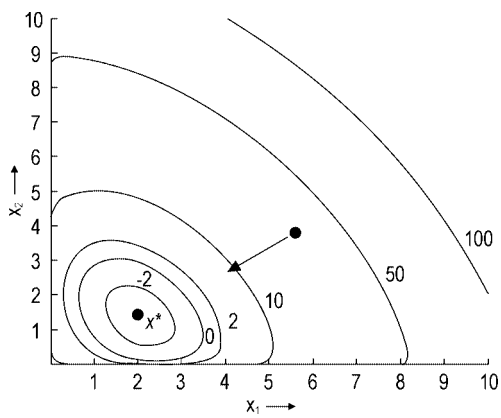

Figure 36. Unconstrained minimum

Now consider the imposition of inequality $[g(x) \leq 0]$ and equality constraints $[h(x) = 0]$ in Figure 37. Continuing the kinematic interpretation, the inequality constraints $g(x) \leq 0$ act as "fences" in the valley, and equality constraints

$h(x) = 0$ as "rails". Consider now a ball, constrained on a rail and within fences, to roll to its lowest point. This stationary point occurs when the normal forces exerted by the fences $[-\nabla g(x^*)]$ and rails $[-\nabla h(x^*)]$ on the ball are balanced by the force of "gravity" $[-\nabla f(x^*)]$. This condition can be stated by the following *KKT necessary conditions* for constrained optimality:

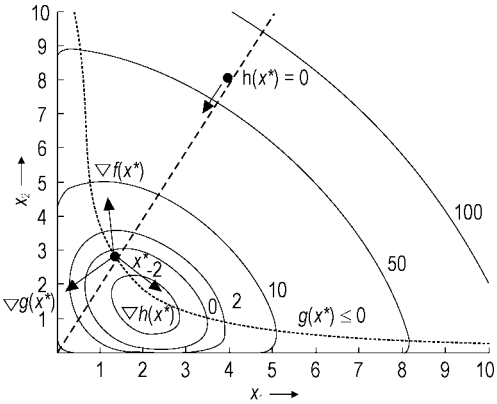

Figure 37. Constrained minimum

Stationarity Condition: It is convenient to define the Lagrangian function $L(x, \lambda, \nu) = f(x) + g(x)^T\lambda + h(x)^T\nu$ along with "weights" or multipliers λ and ν for the constraints. These multipliers are also known as "dual variables" and "shadow prices". The stationarity condition (balance of forces acting on the ball) is then given by:

$$\nabla L(x,\lambda,\nu) = \nabla f(x) + \nabla h(x)\lambda + \nabla g(x)\nu = 0 \tag{56}$$

Feasibility: Both inequality and equality constraints must be satisfied (ball must lie on the rail and within the fences):

$$h(x) = 0, \quad g(x) \leq 0 \tag{57}$$

Complementarity: Inequality constraints are either strictly satisfied (active) or inactive, in which case they are irrelevant to the solution. In the latter case the corresponding KKT multiplier must be zero. This is written as:

$$\nu^T g(x) = 0, \quad \nu \geq 0 \tag{58}$$

Constraint Qualification: For a local optimum to satisfy the KKT conditions, an additional regularity condition or constraint qualification (CQ) is required. The KKT conditions are derived from gradient information, and the CQ can be viewed as a condition on the relative influence of constraint curvature. In fact, linearly constrained problems with nonempty feasible regions require no constraint qualification. On the other hand, as seen in Figure 38, the problem Min x_1, s.t. $x_2 \geq 0$, $(x_1)^3 \geq x_2$ has a minimum point at the origin but does not satisfy the KKT conditions because it does not satisfy a CQ.

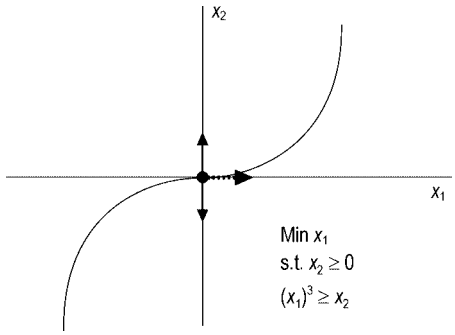

Figure 38. Failure of KKT conditions at constrained minimum (note linear dependence of constraint gradients)

CQs be defined in several ways. For instance, the linear independence constraint qualification (LICQ) requires that the active constraints at x^* be linearly independent, i.e., the matrix $[\nabla h(x^*) \,|\, \nabla g_A(x^*)]$ is full column rank, where g_A is the vector of inequality constraints with elements that satisfy $g_{A,i}(x^*) = 0$. With LICQ, the KKT multipliers (λ, ν) are guaranteed to be unique at the optimal solution. The weaker Mangasarian – Fromovitz constraint qualification (MFCQ) requires only that $\nabla h(x^*)$ have full column rank and that a direction p exist that satisfies $\nabla h(x^*)^T p = 0$ and $\nabla g_A(x^*)^T p > 0$. With MFCQ, the KKT multipliers (λ, ν) are guaranteed to be bounded (but not necessarily unique) at the optimal solution. Additional discussion can be found in [169].

Second Order Conditions: As with unconstrained optimization, nonnegative (positive) curvature is necessary (sufficient) in all of the allowable (i.e., constrained) nonzero directions p. This condition can be stated in several ways. A typical necessary second-order condition requires a point x^* that satisfies LICQ and first-order conditions 56–58 with multipliers (λ, ν) to satisfy the additional conditions given by:

$$p^T \nabla_{xx} L(x^*,\lambda,\nu)p \geq 0 \quad \text{for all}$$
$$p \neq 0, \nabla h(x^*)^T p = 0,$$
$$\nabla g_A(x^*)^T p = 0 \tag{59}$$

The corresponding sufficient conditions require that the inequality in 59 be strict. Note that for the example in Figure 36, the allowable directions p span the entire space for x while in Figure 37, there are *no* allowable directions p.

Example: To illustrate the KKT conditions, consider the following unconstrained NLP:

$$\begin{aligned} &\text{Min } (x_1)^2 - 4x_1 + 3/2(x_2)^2 \\ &-7x_2 + x_1x_2 + 9 - \ln x_1 - \ln x_2 \end{aligned} \tag{60}$$

corresponding to the contour plot in Figure 36. The optimal solution can be found by solving for the first order conditions 54:

$$\nabla f(x) = \begin{bmatrix} 2x_1 - 4 + x_2 - 1/x_1 \\ 3x_2 - 7 + x_1 - 1/x_2 \end{bmatrix} = 0$$
$$\Rightarrow x^* = \begin{bmatrix} 1.3475 \\ 2.0470 \end{bmatrix} \tag{61}$$

and $f(x^*) = -2.8742$. Checking the second-order conditions leads to:

$$\nabla_{xx} f(x*) = \begin{bmatrix} 2 + 1/(x_1^*)^2 & 1 \\ 1 & 3 + 1/(x_2^*)^2 \end{bmatrix} \Rightarrow$$
$$\nabla_{xx} f(x*) = \begin{bmatrix} 2.5507 & 1 \\ 1 & 3.2387 \end{bmatrix} \text{(positive definite)} \tag{62}$$

Now consider the constrained NLP:

$$\begin{aligned} \text{Min} \quad & (x_1)^2 - 4x_1 + 3/2(x_1)^2 - 7x_2 + x_1x_2 \\ & + 9 - \ln x_1 - \ln x_2 \\ \text{s.t.} \quad & 4 - x_1x_2 \leq 0 \\ & 2x_1 - x_2 = 0 \end{aligned} \tag{63}$$

that corresponds to the plot in Figure 37. The optimal solution can be found by applying the first-order KKT conditions 56–58:

$$\nabla L(x,\lambda,\nu) = \nabla f(x) + \nabla h(x)\lambda + \nabla g(x)\nu =$$
$$\begin{bmatrix} 2x_1 - 4 + x_2 - 1/x_1 \\ 3x_2 - 7 + x_1 - 1/x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ -1 \end{bmatrix} \lambda + \begin{bmatrix} -x_2 \\ -x_1 \end{bmatrix} \nu = 0$$
$$g(x) = 4 - x_1x_2 \leq 0, \quad h(x) = 2x_1 - x_2 = 0$$
$$g(x)\nu = (4 - x_1x_2)\nu, \quad \nu \geq 0$$
$$\Downarrow$$
$$x* = \begin{bmatrix} 1.4142 \\ 2.8284 \end{bmatrix}, \lambda* = 1.036, \nu* = 1.068$$

and $f(x^*) = -1.8421$. Checking the second-order conditions 59 leads to:

$$\nabla_{xx} L(x*,\lambda*,\nu*) = \nabla_{xx}[f(x*) + h(x*)\lambda* + g(x*)\nu*] =$$
$$\begin{bmatrix} 2 + 1/(x_1)^2 & 1 - \nu \\ 1 - \nu & 3 + 1/(x_2)^2 \end{bmatrix} = \begin{bmatrix} 2.5 & 0.068 \\ 0.068 & 3.125 \end{bmatrix}$$
$$[\nabla h(x*) | \nabla g_A(x*)]$$
$$p = \begin{bmatrix} 2 & -2.8284 \\ -1 & -1.4142 \end{bmatrix} p = 0, p \neq 0$$

Note that LICQ is satisfied. Moreover, because $[\nabla h(x^*) | \nabla g_A(x^*)]$ is square and nonsingular, there are no nonzero vectors p that satisfy the allowable directions. Hence, the sufficient second-order conditions ($p^T \nabla_{xx} L(x^*,\lambda^*,\nu^*) > 0$, for all allowable p) are *vacuously satisfied* for this problem.

Convex Cases of NLP. Linear programs and quadratic programs are special cases of problem 52 that allow for more efficient solution, based on application of KKT conditions 56–59. Because these are convex problems, any locally optimal solution is a global solution. In particular, if the objective and constraint functions in problem 52 are linear then the following linear program (LP):

$$\begin{aligned} \text{Min} \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & Cx \leq d \end{aligned} \tag{65}$$

can be solved in a finite number of steps, and the optimal solution lies at a vertex of the polyhedron described by the linear constraints. This is shown in Figure 39, and in so-called primal degenerate cases, multiple vertices can be alternate optimal solutions with the same values of the objective function. The standard method to solve problem 65 is the simplex method, developed in the late 1940s [170], although, starting from KARMARKAR's discovery in 1984, interior point methods have become quite advanced and competitive for large scale problems [171]. The simplex method proceeds by moving successively from vertex to vertex with improved objective function values. Methods to solve problem 65 are well implemented and widely used,

especially in planning and logistical applications. They also form the basis for MILP methods (see below). Currently, state-of-the-art LP solvers can handle millions of variables and constraints and the application of further decomposition methods leads to the solution of problems that are two or three orders of magnitude larger than this [172, 173]. Also, the interior point method is described below from the perspective of more general NLPs.

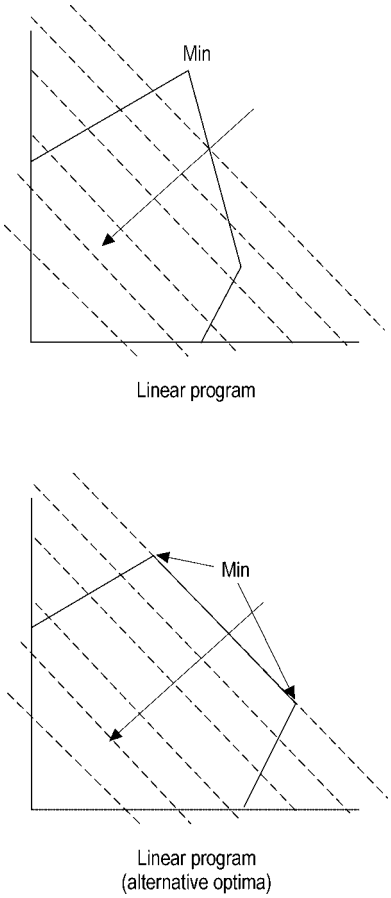

Figure 39. Contour plots of linear programs

Quadratic programs (QP) represent a slight modification of problem 65 and can be stated as:

$$\begin{array}{ll} \text{Min} & c^T x + \frac{1}{2} x^T Q x \\ \text{s.t.} & Ax = b \\ & x \leq d \end{array} \tag{66}$$

If the matrix Q is positive semidefinite (positive definite), when projected into the null space of the active constraints, then problem 66 is (strictly) convex and the QP is a global (and unique) minimum. Otherwise, local solutions exist for problem 66, and more extensive global optimization methods are needed to obtain the global solution. Like LPs, convex QPs can be solved in a finite number of steps. However, as seen in Figure 40, these optimal solutions can lie on a vertex, on a constraint boundary, or in the interior. A number of active set strategies have been created that solve the KKT conditions of the QP and incorporate efficient updates of active constraints. Popular QP methods include null space algorithms, range space methods, and Schur complement methods. As with LPs, QP problems can also be solved with interior point methods [171].

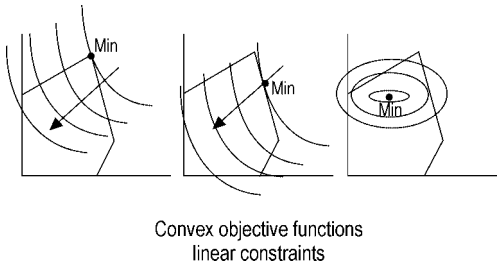

Figure 40. Contour plots of convex quadratic programs

Solving the General NLP Problem. Solution techniques for problem 52 deal with satisfaction of the KKT conditions 56–59. Many NLP solvers are based on successive quadratic programming (SQP) as it allows the construction of a number of NLP algorithms based on the Newton – Raphson method for equation solving. SQP solvers have been shown to require the fewest function evaluations to solve NLPs [174] and they can be tailored to a broad range of process engineering problems with different structure.

The SQP strategy applies the equivalent of a Newton step to the KKT conditions of the nonlinear programming problem, and this leads to a fast rate of convergence. By adding slack variables s the first-order KKT conditions can be rewritten as:

$$\nabla f(x)+\nabla h(x)\lambda+\nabla g(x)\nu=0 \tag{67a}$$

$$h(x)=0 \tag{67b}$$

$$g(x)+s=0 \tag{67c}$$

$$SVe=0 \tag{67d}$$

$$(s,\nu)\geq 0 \tag{67e}$$

where $e=[1,\ 1,\ \ldots,\ 1]^T$, $S=\mathrm{diag}(s)$ and $V=\mathrm{diag}(\nu)$. SQP methods find solutions that satisfy Equations 67a, 67b, 67c, 67d and 67e by generating Newton-like search directions at iteration k. However, equations 67d and active bounds 67e are dependent and serve to make the KKT system ill-conditioned near the solution. SQP algorithms treat these conditions in two ways. In the *active set strategy*, discrete decisions are made regarding the active constraint set, $i\in I=\{i|g_i(x^*)=0\}$, and Equation 67d is replaced by $s_i=0$, $i\in I$, and $v_i=0$, $i\notin I$. Determining the active set is a combinatorial problem, and a straightforward way to determine an estimate of the active set [and also satisfy 67e] is to formulate, at a point x^k, and solve the following QP at iteration k:

$$\begin{array}{ll} \text{Min} & \nabla f(x^k)^T p+\frac{1}{2}p^T\nabla_{xx}L(x^k,\lambda^k,\nu^k)\,p \\ \text{s.t.} & h(x^k)+\nabla h(x^k)^T p=0 \\ & g(x^k)+\nabla g(x^k)^T p+s=0,\ s\geq 0 \end{array} \tag{68}$$

The KKT conditions of 68 are given by:

$$\nabla f\left(x^k\right)+\nabla^2 L\left(x^k,\lambda^k,\nu^k\right)p$$
$$+\nabla h\left(x^k\right)\lambda+\nabla g\left(x^k\right)\nu=0 \tag{69a}$$

$$h\left(x^k\right)+\nabla h\left(x^k\right)^T p=0 \tag{69b}$$

$$g\left(x^k\right)+\nabla g\left(x^k\right)^T p+s=0 \tag{69c}$$

$$SVe=0 \tag{69d}$$

$$(s,\nu)\geq 0 \tag{69e}$$

where the Hessian of the Lagrange function $\nabla_{xx}L(x,\lambda,\nu)=\nabla_{xx}\left[f(x)+h(x)^T\lambda+g(x)^T\nu\right]$ is calculated directly or through a quasi-Newtonian approximation (created by differences of gradient vectors). It is easy to show that 69a–69c correspond to a Newton–Raphson step for 67a–67c applied at iteration k. Also, selection of the active set is now handled at the QP level by satisfying the conditions 69d, 69e. To evaluate and change candidate active sets, QP algorithms apply inexpensive matrix-updating strategies to the KKT matrix associated with the QP 68. Details of this approach can be found in [175, 176].

As alternatives that avoid the combinatorial problem of selecting the active set, interior point (or barrier) methods modify the NLP problem 52 to form problem 70

$$\begin{array}{ll} \text{Min} & f\left(x^k\right)-\mu\sum_i \ln s_i \\ \text{s.t.} & h\left(x^k\right)=0 \\ & g\left(x^k\right)+s=0 \end{array} \tag{70}$$

where the solution to this problem has $s>0$ for the penalty parameter $\mu>0$, and decreasing μ to zero leads to solution of problem 52. The KKT conditions for this problem can be written as Equation 71

$$\begin{array}{l} \nabla f(x*)+\nabla h(x*)\lambda+\nabla g(x*)\nu=0 \\ h(x*)=0 \\ g(x*)+s=0 \\ SVe=\mu e \end{array} \tag{71}$$

and at iteration k the Newton steps to solve 71 are given by:

$$\begin{bmatrix} \nabla_{xx}L(x_k,\lambda_k,\nu_k) & & \nabla h(x_k) & \nabla g(x_k) \\ & S_k^{-1}V_k & & I \\ \nabla h(x_k)^T & & & \\ \nabla g(x_k)^T & I & & \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta s \\ \Delta\lambda \\ \Delta\nu \end{bmatrix} = -\begin{bmatrix} \nabla_x L(x_k,\lambda_k,\nu_k) \\ \nu_k-S_k^{-1}\mu e \\ h(x_k) \\ g(x_k)+s_k \end{bmatrix} \tag{72}$$

A detailed description of this algorithm, called IPOPT, can be found in [177].

Both active set and interior point methods have clear trade-offs. Interior point methods may require more iterations to solve problem 70 for various values of μ, while active set methods require the solution of the more expensive QP subproblem 68. Thus, if there are few inequality constraints or an active set is known (say from a good starting guess, or a known QP solution from a previous iteration) then solving problem 68 is not expensive and the active set method is favored. On the other hand, for problems with

many inequality constraints, interior point methods are often faster as they avoid the combinatorial problem of selecting the active set. This is especially the case for large-scale problems and when a large number of bounds are active. Examples that demonstrate the performance of these approaches include the solution of model predictive control (MPC) problems [178 – 180] and the solution of large optimal control problems using barrier NLP solvers. For instance, IPOPT allows the solution of problems with more than 10^6 variables and up to 50 000 degrees of freedom [181, 182].

Other Gradient-Based NLP Solvers. In addition to SQP methods, a number of NLP solvers have been developed and adapted for large-scale problems. Generally, these methods require more function evaluations than SQP methods, but they perform very well when interfaced to optimization modeling platforms, where function evaluations are cheap. All of these can be derived from the perspective of applying Newton steps to portions of the KKT conditions.

LANCELOT [183] is based on the solution of bound constrained subproblems. Here an augmented Lagrangian is formed from problem 52 and subproblem 73 is solved.

$$\begin{aligned} &\text{Min } f(x)+\lambda^T h(x) \\ &\quad +\nu(g(x)+s)+\tfrac{1}{2}\rho\|g(x),g(x)+s\|^2 \\ &\text{s.t.} \quad s\geq 0 \end{aligned} \tag{73}$$

The above subproblem can be solved very efficiently for fixed values of the multipliers λ and v and penalty parameter ρ. Here a gradient-projection, trust-region method is applied. Once subproblem 73 is solved, the multipliers and penalty parameter are updated in an outer loop and the cycle repeats until the KKT conditions for problem 52 are satisfied. LANCELOT works best when exact second derivatives are available. This promotes a fast convergence rate in solving each subproblem and allows a bound constrained trust-region method to exploit directions of negative curvature in the Hessian matrix.

Reduced gradient methods are active set strategies that rely on partitioning the variables and solving Equations 67a, 67b, 67c, 67d and 67e in a nested manner. Without loss of generality, problem 52 can be rewritten as problem 74.

$$\begin{aligned} &\text{Min } f(z) \\ &\text{s.t. } c(z)=0 \\ &a\leq z\leq b \end{aligned} \tag{74}$$

Variables are partitioned as nonbasic variables (those fixed to their bounds), basic variables (those that can be solved from the equality constraints), and superbasic variables (those remaining variables between bounds that serve to drive the optimization); this leads to $z^T=\left[z_N^T,z_S^T,z_B^T\right]$. This partition is derived from local information and may change over the course of the optimization iterations. The corresponding KKT conditions can be written as Equations 75a, 75b, 75c, 75d and 75e

$$\nabla_{\mathrm{N}} f(z)+\nabla_{\mathrm{N}} c(z)\gamma=\beta_{\mathrm{a}}-\beta_{\mathrm{b}} \tag{75a}$$

$$\nabla_{\mathrm{S}} f(z)+\nabla_{\mathrm{S}} c(z)\gamma=0 \tag{75b}$$

$$\nabla_{\mathrm{B}} f(z)+\nabla_{\mathrm{B}} c(z)\gamma=0 \tag{75c}$$

$$c(z)=0 \tag{75d}$$

$$z_{\mathrm{N},j}=a_j \text{ or } b_j,\ \beta_{\mathrm{a},j}\geq 0,\beta_{\mathrm{b},j}=0 \text{ or } \beta_{\mathrm{b},j}\geq 0,\beta_{\mathrm{a},j}=0 \tag{75e}$$

where λ and β are the KKT multipliers for the equality and bound constraints, respectively, and 75e replaces the complementarity conditions 58. Reduced gradient methods work by nesting equations 75b, 75d within 75a, 75c. At iteration k, for fixed values of z_{N}^k and z_{S}^k, we can solve for z_{B} using 75d and for λ using 75b. Moreover, linearization of these equations leads to constrained derivatives or *reduced gradients* (Eq. 76)

$$\frac{df}{dz_{\mathrm{s}}}=\nabla f_{\mathrm{s}}-\nabla c_{\mathrm{s}}(\nabla c_{\mathrm{B}})^{-1}\nabla f_{\mathrm{B}} \tag{76}$$

which indicate how $f(z)$ (and z_{B}) change with respect to z_{S} and z_{N}. The algorithm then proceeds by updating z_{S} using reduced gradients in a Newton-type iteration to solve equation 75c. Following this, bound multipliers β are calculated from 75a. Over the course of the iterations, if the variables z_{B} or z_{S} exceed their bounds or if some bound multipliers β become negative, then the variable partition needs to be changed and the equations 75a, 75b, 75c, 75d and 75e are reconstructed. These reduced gradient methods are embodied in the popular GRG2, CONOPT, and SOLVER codes [173].

The SOLVER code has been incorporated into Microsoft Excel. CONOPT [184] is an efficient and widely used code in several optimization modeling environments.

MINOS [185] is a well-implemented package that offers a variation on reduced gradient strategies. At iteration *k,* equation 75d is replaced by its linearization (Eq. 77)

$$c(z_{\mathrm{N}}^k, z_{\mathrm{S}}^k, z_{\mathrm{B}}^k) + \nabla_{\mathrm{B}} c\left(z^k\right)^T (z_{\mathrm{B}} - z_{\mathrm{B}}^k) + \nabla_{\mathrm{S}} c\left(z^k\right)^T (z_{\mathrm{S}} - z_{\mathrm{S}}^k) = 0 \quad (77)$$

and (75a–75c, 75e) are solved with Equation 77 as a subproblem using concepts from the reduced gradient method. At the solution of this subproblem, the constraints 75d are relinearized and the cycle repeats until the KKT conditions of 75a, 75b, 75c, 75d and 75e are satisfied. The augmented Lagrangian function 73 is used to penalize movement away from the feasible region. For problems with few degrees of freedom, the resulting approach leads to an extremely efficient method even for very large problems. MINOS has been interfaced to a number of modeling systems and enjoys widespread use. It performs especially well on large problems with few nonlinear constraints. However, on highly nonlinear problems it is usually less reliable than other reduced gradient methods.

Algorithmic Details for NLP Methods. All of the above NLP methods incorporate concepts from the Newton – Raphson Method for equation solving. Essential features of these methods are a) providing accurate derivative information to solve for the KKT conditions, b) stabilization strategies to promote convergence of the Newton-like method from poor starting points, and c) regularization of the Jacobian matrix in Newton's method (the so-called KKT matrix) if it becomes singular or ill-conditioned.

a) *Providing first and second derivatives:* The KKT conditions require first derivatives to define stationary points, so accurate first derivatives are essential to determine locally optimal solutions for differentiable NLPs. Moreover, Newton–Raphson methods that are applied to the KKT conditions, as well as the task of checking second-order KKT conditions, necessarily require information on second derivatives. (Note that second-order conditions are not checked by methods that do not use second derivatives). With the recent development of automatic differentiation tools, many modeling and simulation platforms can provide exact first and second derivatives for optimization. When second derivatives are available for the objective or constraint functions, they can be used directly in LANCELOT, SQP and, less efficiently, in reduced gradient methods. Otherwise, for problems with few superbasic variables, reduced gradient methods and reduced space variants of SQP can be applied. Referring to problem 74 with *n* variables and *m* equalities, we can write the QP step from problem 68 as Equation 78.

$$\begin{aligned} \text{Min} \quad & \nabla f\left(z^k\right)^T p + {}^1\!/_2\, p^T \nabla_{xx} L\left(z^k, \gamma^k\right) p \\ \text{s.t.} \quad & c\left(z^k\right) + \nabla c\left(z^k\right)^T p = 0 \\ & a \le z^k + p \le b \end{aligned} \quad (78)$$

Defining the search direction as $p = Z_k p_Z + Y_k p_Y$, where $\nabla c(x^k)^T Z_k = 0$ and $[Y_k | Z_k]$ is a nonsingular $n \times n$ matrix, allows us to form the following reduced QP subproblem (with $n-m$ variables)

$$\begin{aligned} \text{Min} \quad & \left[Z_k^T \nabla f\left(z^k\right) + w_k\right]^T p_Z + \tfrac{1}{2} p z^T B_k p_Z \\ \text{s.t.} \quad & a \le z^k + Z_k p_Z + Y_k p_Y \le b \end{aligned} \quad (79)$$

where $p_Y = -[\nabla c(z^k)^T Y_k]^{-1} c(z^k)$. Good choices of Y_k and Z_k, which allow sparsity to be exploited in $\nabla c(z^k)$, are: $Y_k^T = [0|I]$ and $Z_k^T = \left[I | -\nabla_{\mathrm{N,S}} c\left(z^k\right) \nabla_B c\left(z^k\right)^{-1}\right]$. Here we define the reduced Hessian $B_k = Z_z^T \nabla_{zz} L\left(z^k, \gamma^k\right) Z_k$ and $w_k = Z_k^T \nabla_{zz} L\left(z^k, \gamma^k\right) Y_k p_Y$. In the absence of second-derivative information, B_k can be approximated using positive definite quasi-Newton approximations [175]. Also, for the interior point method, a similar reduced space decomposition can be applied to the Newton step given in 72.

Finally, for problems with least-squares functions, as in data reconciliation, parameter estimation, and model predictive control, one can often assume that the values of the objective function and its gradient at the solution are vanishingly small. Under these conditions, one can show that the multipliers (λ, ν) also vanish and $\nabla_{xx} L(x^*, \lambda, \nu)$ can be substituted

by $\nabla_{xx} f(x^*)$. This *Gauss – Newton approximation* has been shown to be very efficient for the solution of least-squares problems [175].

b) *Line-search and trust-region methods* are used to promote convergence from poor starting points. These are commonly used with the search directions calculated from NLP subproblems such as problem 68. In a *trust-region approach*, the constraint, $\|p\| \leq \Delta$ is added and the iteration step is taken if there is sufficient reduction of some merit function (e.g., the objective function weighted with some measure of the constraint violations). The size of the trust region Δ is adjusted based on the agreement of the reduction of the actual merit function compared to its predicted reduction from the subproblem [183]. Such methods have strong global convergence properties and are especially appropriate for ill-conditioned NLPs. This approach has been applied in the KNITRO code [186]. *Line-search methods* can be more efficient on problems with reasonably good starting points and well-conditioned subproblems, as in real-time optimization. Typically, once a search direction is calculated from problem 68, or other related subproblem, a step size $\alpha \in (0, 1]$ is chosen so that $x^k + \alpha p$ leads to a sufficient decrease of a merit function. As a recent alternative, a novel *filter-stabilization* strategy (for both line-search and trust-region approaches) has been developed based on a bicriterion minimization, with the objective function and constraint infeasibility as competing objectives [187]. This method often leads to better performance than those based on merit functions.

c) *Regularization of the KKT matrix for the NLP subproblem* (e.g., in Equation 72) is essential for good performance of general purpose algorithms. For instance, to obtain a unique solution to Eqn. 72, active constraint gradients must be full rank, and the Hessian matrix, when projected into the null space of the active constraint gradients, must be positive definite. These properties may not hold far from the solution, and corrections to the Hessian in SQP may be necessary [176]. Regularization methods ensure that subproblems like Eqns. 68 and 72 remain well-conditioned; they include addition of positive constants to the diagonal of the Hessian matrix to ensure its positive definiteness, judicious selection of active constraint gradients to ensure that they are linearly independent, and scaling the subproblem to reduce the propagation of numerical errors. Often these strategies are heuristics built into particular NLP codes. While quite effective, most of these heuristics do not provide convergence guarantees for general NLPs.

Table 11 summarizes the characteristics of a collection of widely used NLP codes. Much more information on widely available codes can also be found on the NEOS server (www-neos.mcs.anl.gov) and the NEOS Software Guide.

Table 11. Representative NLP solvers

Method	Algorithm type	Stabilization	Second-order information
CONOPT [184]	reduced gradient	line search	exact and quasi-Newton
GRG2 [173]	reduced gradient	line search	quasi-Newton
IPOPT [177]	SQP, barrier	line search	exact
KNITRO [186]	SQP, barrier	trust region	exact and quasi-Newton
LANCELOT [183]	aug mented Lagrangian, bound constrained	trust region	exact and quasi-Newton
LOQO [188]	SQP, barrier	line search	exact
MINOS [185]	reduced gradient, augmented Lagrangian	line search	quasi-Newton
NPSOL [189]	SQP, Active set	line search	quasi-Newton
SNOPT [190]	reduced space SQP, active set	line search	quasi-Newton
SOCS [191]	SQP, active set	line search	exact
SOLVER [173]	reduced gradient	line search	quasi-Newton
SRQP [192]	reduced space SQP, active set	line search	quasi-Newton

10.3. Optimization Methods without Derivatives

A broad class of optimization strategies does not require derivative information. These methods have the advantage of easy implementation and little prior knowledge of the optimization problem. In particular, such methods are well suited for "quick and dirty" optimization studies that explore the scope of optimization for new problems, prior to investing effort for more sophisticated modeling and solution strategies. Most of these methods are derived from heuristics that naturally spawn numerous variations. As a result, a very broad literature describes these methods. Here we discuss only a few important trends in this area.

Classical Direct Search Methods. Developed in the 1960s and 1970s, these methods include one-at-a-time search and methods based on experimental designs [193]. At that time, these direct search methods were the most popular optimization methods in chemical engineering. Methods that fall into this class include the pattern search [194], the conjugate direction method [195], simplex and complex searches [196], and the adaptive random search methods [197 – 199]. All of these methods require only objective function values for unconstrained minimization. Associated with these methods are numerous studies on a wide range of process problems. Moreover, many of these methods include heuristics that prevent premature termination (e.g., directional flexibility in the complex search as well as random restarts and direction generation). To illustrate these methods, Figure 41 illustrates the performance of a pattern search method as well as a random search method on an unconstrained problem.

Simulated Annealing. This strategy is related to random search methods and derives from a class of heuristics with analogies to the motion of molecules in the cooling and solidification of metals [200]. Here a "temperature" parameter θ can be raised or lowered to influence the probability of accepting points that do not improve the objective function. The method starts with a base point x and objective value $f(x)$. The next point x' is chosen at random from a distribution. If $f(x') < f(x)$, the move is accepted with x' as the new point. Otherwise, x' is accepted with probability $p(\theta,x',x)$. Options include the Metropolis distribution, $p(\theta,x,x') = \exp[-\{f(x')-f(x)\}/\theta]$, and the Glauber distribution, $p(\theta, x, x') = \exp[-\{f(x')-f(x)\}/(1 + \exp[-\{f(x')-f(x)\}]/\theta)]$. The θ parameter is then reduced and the method continues until no further progress is made.

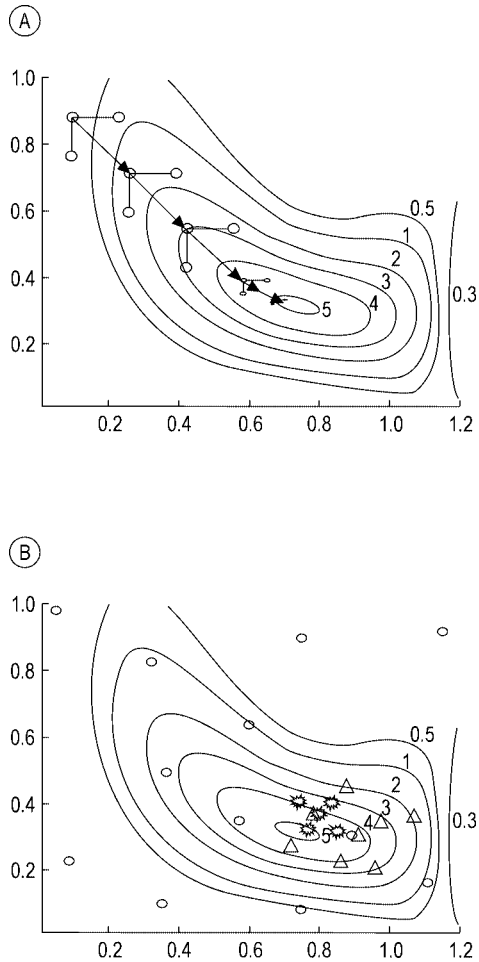

Figure 41. Examples of optimization methods without derivatives
A) Pattern search method; B) Random search method
Circles: 1st phase; Triangles: 2nd phase; Stars: 3rd phase

Genetic Algorithms. This approach, first proposed in [201], is based on the analogy of improving a population of solutions through modifying their gene pool. It also has similar performance characteristics as random search methods and simulated annealing. Two forms of ge-

netic modification, crossover or mutation, are used and the elements of the optimization vector $\boldsymbol{x}$ are represented as binary strings. Crossover deals with random swapping of vector elements (among parents with highest objective function values or other rankings of population) or any linear combinations of two parents. Mutation deals with the addition of a random variable to elements of the vector. Genetic algorithms (GAs) have seen widespread use in process engineering and a number of codes are available. A related GA algorithm is described in [173].

Derivative-Free Optimization (DFO). In the past decade, the availability of parallel computers and faster computing hardware and the need to incorporate complex simulation models within optimization studies have led a number of optimization researchers to reconsider classical direct search approaches. In particular, Dennis and Torczon [202] developed a multidimensional search algorithm that extends the simplex approach [196]. They note that the Nelder–Mead algorithm fails as the number of variables increases, even for very simple problems. To overcome this, their multidimensional pattern-search approach combines reflection, expansion, and contraction steps that act as line search algorithms for a number of linear independent search directions. This approach is easily adapted to parallel computation and the method can be tailored to the number of processors available. Moreover, this approach converges to locally optimal solutions for unconstrained problems and exhibits an unexpected performance synergy when multiple processors are used. The work of Dennis and Torczon [202] has spawned considerable research on analysis and code development for DFO methods. Moreover, Conn et al. [203] construct a multivariable DFO algorithm that uses a surrogate model for the objective function within a trust-region method. Here points are sampled to obtain a well-defined quadratic interpolation model, and descent conditions from trust-region methods enforce convergence properties. A number of trust-region methods that rely on this approach are reviewed in [203]. Moreover, a number of DFO codes have been developed that lead to black-box optimization implementations for large, complex simulation models. These include the DAKOTA package at Sandia National Lab [204], http://endo.sandia.gov/DAKOTA/software.html and FOCUS developed at Boeing Corporation [205].

Direct search methods are easy to apply to a wide variety of problem types and optimization models. Moreover, because their termination criteria are not based on gradient information and stationary points, they are more likely to favor the search for global rather than locally optimal solutions. These methods can also be adapted easily to include integer variables. However, rigorous convergence properties to globally optimal solutions have not yet been discovered. Also, these methods are best suited for unconstrained problems or for problems with simple bounds. Otherwise, they may have difficulties with constraints, as the only options open for handling constraints are equality constraint elimination and addition of penalty functions for inequality constraints. Both approaches can be unreliable and may lead to failure of the optimization algorithm. Finally, the performance of direct search methods scales poorly (and often exponentially) with the number of decision variables. While performance can be improved with the use of parallel computing, these methods are rarely applied to problems with more than a few dozen decision variables.

10.4. Global Optimization

Deterministic optimization methods are available for nonconvex nonlinear programming problems of the form problem 52 that guarantee convergence to the global optimum. More specifically, one can show under mild conditions that they converge to an ε distance to the global optimum on a finite number of steps. These methods are generally more expensive than local NLP methods, and they require the exploitation of the structure of the nonlinear program.

Global optimization of nonconvex programs has received increased attention due to their practical importance. Most of the deterministic global optimization algorithms are based on spatial branch-and-bound algorithm [206], which divides the feasible region of continuous variables and compares lower bound and upper bound for fathoming each subregion. The one that contains the optimal solution is found by eliminating subregions that are proved not to contain the optimal solution.

For nonconvex NLP problems, Quesada and Grossmann [207] proposed a spatial branch-and-bound algorithm for concave separable, linear fractional, and bilinear programs using linear and nonlinear underestimating functions [208]. For nonconvex MINLP, Ryoo and Sahinidis [209] and later Tawarmalani and Sahinidis [210] developed BARON, which branches on the continuous and discrete variables with bounds reduction method. Adjiman et al. [211, 212] proposed the SMIN-αBB and GMIN-αBB algorithms for twice-differentiable nonconvex MINLPs. Using a valid convex underestimation of general functions as well as for special functions, Adjiman et al. [213] developed the αBB method, which branches on both the continuous and discrete variables according to specific options. The branch-and-contract method [214] has bilinear, linear fractional, and concave separable functions in the continuous variables and binary variables, uses bound contraction, and applies the outer-approximation (OA) algorithm at each node of the tree. Smith and Pantelides [215] proposed a reformulation method combined with a spatial branch-and-bound algorithm for nonconvex MINLP and NLP.

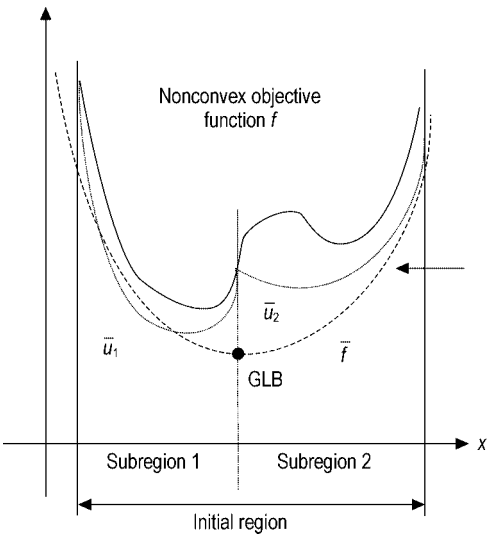

Figure 42. Convex underestimator for nonconvex function

Because in global optimization one cannot exploit optimality conditions like the KKT conditions for a local optimum, these methods work by first partitioning the problem domain (i.e., containing the feasible region) into subregions (see Fig. 42). Upper bounds on the objective function are computed over all subregions of the problem. In addition, lower bounds can be derived from convex underestimators of the objective function and constraints for each subregion. The algorithm then proceeds to eliminate all subregions that have lower bounds that are greater than the least upper bound. After this, the remaining regions are further partitioned to create new subregions and the cycle continues until the upper and lower bounds converge. Below we illustrate the specific steps of the algorithm for nonlinear programs that involve bilinear, linear fractional, and concave separable terms [207, 214].

Nonconvex NLP with Bilinear, Linear Fractional, and Concave Separable Terms. Consider the following specific nonconvex NLP problem,

$$\underset{x}{\text{Min}} \quad f(x) = \sum_{(i,j)\in BL_0} a_{ij}\, x_i x_j + \sum_{(i,j)\in LF_0} b_{ij} \frac{x_i}{x_j} + \sum_{i\in C_0} g_i(x_i) + h(x)$$

subject to

$$f_k(x) = \sum_{(i,j)\in BL_k} a_{ijk} x_i x_j + \sum_{(i,j)\in LF_k} b_{ijk} \frac{x_i}{x_j} + \sum_{i\in C_k} g_{i,k}(x_i) + h_k(x) \le 0 \quad k\in K$$

$$x\in S\cap \Omega_0 \subset R^n$$

where a_{ij}, a_{ijk}, b_{ij}, b_{ijk} are scalars with $i\in I=\{1,2,\cdots,n\}$, $j\in J=\{1,2,\cdots,n\}$, and $k\in K=\{1,2,\cdots,m\}$. BL_0, BL_k, LF_0, LF_k are (i, j)-index sets, with $i\neq j$, that define the bilinear and linear fractional terms present in the problem. The functions $h(x)$, $h_k(x)$ are convex, and twice continuously differentiable. C_0 and C_k are index sets for the univariate twice continuously differentiable concave functions $g_i(x_i)$, $g_{i,k}(x_i)$. The set $S\subset R^n$ is convex, and $\omega_0 \supset R^n$ is an n-dimensional hyperrectangle defined in terms of the initial variable bounds $x^{\mathrm{L,in}}$ and $x^{\mathrm{U,in}}$:

$$\Omega_0 = \{x\in R^n : 0 \le x^{\mathrm{L,in}} \le x \le x^{\mathrm{U,in}},\ x_j^{\mathrm{L,in}} > 0 \text{ if } (i,j)\in LF_0 \cup LF_k,\quad i\in I, j\in J, k\in K\}$$

The feasible region of problem 80 is denoted by D. Note that a nonlinear equality constraint of the form $f_k(x) = 0$ can be accommodated in

problem 51 through the representation by the inequalities $f_k(x) \leq 0$ and $-f_k(x) \leq 0$, provided $h_k(x)$ is separable.

To obtain a lower bound LB (Ω) for the global minimum of problem 80 over $D \cap \Omega$, where $\Omega = \{x \in R^n : x^{\mathrm{L}} \leq x \leq x^{\mathrm{U}}\} \subseteq \Omega_0$, the following problem is proposed:

$$\underset{(x,y,z)}{\text{Min}} \quad \hat{f}(x,y,z) = \sum_{(i,j)\in BL_0} a_{ij}\, y_{ij}$$
$$+ \sum_{(i,j)\in LF_0} b_{ij} z_{ij} + \sum_{i\in C_0} \hat{g}_i(x_i) + h(x)$$

subject to

$$\hat{f}_k(x,y,z) = \sum_{(i,j)\in BL_k} a_{ijk} y_{ij} + \sum_{(i,j)\in LF_k} b_{ijk} z_{ij}$$
$$+ \sum_{i\in C_k} \hat{g}_{i,k}(x_i) + h_k(x) \leq 0 \quad k \in K$$
$$(x,y,z) \in T(\Omega) \subset R^n \times R^{n_1} \times R^{n_2}$$
$$x \in S \cap \Omega \subset R^n,\ y \in R_+^{n_1},\ z \in R_+^{n_2},$$

where the functions and sets are defined as follows:

a) $\hat{g}_i(x_i)$ and $\hat{g}_{i,k}(x_i)$ are the convex envelopes for the univariate functions over the domain $x_i \in [x_i^L, x_i^U]$ [216]:

$$\hat{g}_i(x_i) = g_i\left(x_i^{\mathrm{L}}\right) + \left(\frac{g_i\left(x_i^{\mathrm{U}}\right) - g_i\left(x_i^L\right)}{x_i^{\mathrm{U}} - x_i^{\mathrm{L}}}\right)\left(x_i - x_i^{\mathrm{L}}\right) \leq g_i(x_i) \tag{80}$$

$$\hat{g}_{i,k}(x_i) = g_{i,k}\left(x_i^{\mathrm{L}}\right) + \left(\frac{g_{i,k}\left(x_i^{\mathrm{U}}\right) - g_{i,k}\left(x_i^L\right)}{x_i^{\mathrm{U}} - x_i^{\mathrm{L}}}\right)\left(x_i - x_i^{\mathrm{L}}\right) \leq g_{i,k}(x_i) \tag{81}$$

where $\hat{g}_i(x_i) = g_i(x_i)$ at $x_i = x_i^{\mathrm{L}}$, and $x_i = x_i^{\mathrm{U}}$; likewise, $\hat{g}_{i,k}(x_i) = g_{i,k}(x_i)$ at $x_i = x_i^{\mathrm{L}}$, and $x_i = x_i^{\mathrm{U}}$.

b) $y = \{y_{ij}\}$ is a vector of additional variables for relaxing the bilinear terms in 80, and is used in the following inequalities which determine the convex and concave envelopes of bilinear terms:

$$\begin{aligned} y_{ij} &\geq x_j^{\mathrm{L}} x_i + x_i^{\mathrm{L}} x_j - x_i^{\mathrm{L}} x_j^{\mathrm{L}} && (i,j) \in BL^+ \\ y_{ij} &\geq x_j^{\mathrm{U}} x_i + x_i^{\mathrm{U}} x_j - x_i^{\mathrm{U}} x_j^{\mathrm{U}} && (i,j) \in BL^+ \end{aligned} \tag{82}$$

$$\begin{aligned} y_{ij} &\leq x_j^{\mathrm{L}} x_i + x_i^{\mathrm{U}} x_j - x_i^{\mathrm{U}} x_j^{\mathrm{L}} && (i,j) \in BL^- \\ y_{ij} &\leq x_j^{\mathrm{U}} x_i + x_i^{\mathrm{L}} x_j - x_i^{\mathrm{L}} x_j^{\mathrm{U}} && (i,j) \in BL^- \end{aligned} \tag{83}$$

where

$$BL^+ = \{(i,j) : (i,j) \in BL_0 \cup BL_k,\ a_{ij} > 0 \text{ or } a_{ijk} > 0,\ k \in K\}$$

$$BL^- = \{(i,j) : (i,j) \in BL_0 \cup BL_k,\ a_{ij} < 0 \text{ or } a_{ijk} < 0,\ k \in K\}$$

The inequalities 82 were first derived by McCormick [208], and along with the inequalities 83 theoretically characterized by Al-Khayyal and Falk [217, 218].

c) $z = \{z_{ij}\}$ is a vector of additional variables for relaxing the linear fractional terms in problem 80; these variables are used in the following inequalities:

$$\begin{aligned} z_{ij} &\geq \frac{x_i}{x_j^{\mathrm{L}}} + x_i^{\mathrm{U}}\left(\frac{1}{x_j} - \frac{1}{x_j^{\mathrm{L}}}\right) && (i,j) \in LF^+ \\ z_{ij} &\geq \frac{x_i}{x_j^{\mathrm{U}}} + x_i^{L}\left(\frac{1}{x_j} - \frac{1}{x_j^{\mathrm{U}}}\right) && (i,j) \in LF^+ \end{aligned} \tag{84}$$

$$z_{ij} \geq \frac{1}{x_j}\left(\frac{x_i + \sqrt{x_i^{\mathrm{L}} x_i^{\mathrm{U}}}}{\sqrt{x_i^{\mathrm{L}}} + \sqrt{x_i^{\mathrm{U}}}}\right)^2 \quad (i,j) \in LF^+ \tag{85}$$

$$\begin{aligned} z_{ij} &\leq \frac{1}{x_j^{\mathrm{L}} x_j^{\mathrm{U}}}\left(x_j^{\mathrm{U}} x_i - x_i^{\mathrm{L}} x_j + x_i^{\mathrm{L}} x_j^{\mathrm{L}}\right) && (i,j) \in LF^- \\ z_{ij} &\leq \frac{1}{x_j^{\mathrm{L}} x_j^{\mathrm{U}}}\left(x_j^{\mathrm{L}} x_i - x_i^{\mathrm{U}} x_j + x_i^{\mathrm{U}} x_j^{\mathrm{U}}\right) && (i,j) \in LF^- \end{aligned} \tag{86}$$

where

$$LF^+ = \{(i,j) : (i,j) \in LF_0 \cup LF_k,\ b_{ij} > 0 \text{ or } b_{ijk} > 0,\ k \in K\}$$

$$LF^- = \{(i,j) : (i,j) \in LF_0 \cup LF_k,\ b_{ij} < 0 \text{ or } b_{ijk} < 0,\ k \in K\}$$

The inequalities 84 and 85, 86 are convex underestimators due to Quesada and Grossmann [207, 219] and Zamora and Grossmann [214], respectively.

d) $T(\Omega) = \{(x,y,z) \in R^n \times R^{n_1} \times R^{n_2} :$ 80– 85 are satisfied with $x^{\mathrm{L}}, x^{\mathrm{U}}$ as in $\Omega\}$. The feasible region, and the solution of problem 52 are denoted by $M(\Omega)$, and $(\hat{x}, \hat{y}, \hat{z})_{\Omega}$, respectively. We define the *approximation gap* $\varepsilon(\Omega)$ at a branch-and-bound node as

$$\varepsilon(\Omega) = \begin{cases} \infty & \text{if OUB} = \infty \\ -\text{LB}(\Omega) & \text{if OUB} = 0 \\ \frac{(\text{OUB} - \text{LB}(\Omega))}{|\text{OUB}|} & \text{otherwise} \end{cases} \tag{87}$$

where the *overall upper bound* (OUB) is the value of $f(x)$ at the best available feasible point $x \in D$; if no feasible point is available, then OUB$=\infty$.

Note that the underestimating problem 80 is a linear program if $LF^{+}=\varnothing$. During the execution of the spatial branch-and-bound algorithm, problem (NBNC) is solved initially over $M(\Omega_0)$ (root node of the branch-and-bound tree). If a better approximation is required, $M(\Omega_0)$ is refined by partitioning Ω_0 into two smaller hyperrectangles Ω_{01} and Ω_{02}, and two children nodes are created with relaxed feasible regions given by $M(\Omega_{01})$ and $M(\Omega_{02})$. Problem 80 might be regarded as a basic underestimating program for the general problem 80. In some cases, however, it is possible to develop additional convex estimators that might strengthen the underestimating problem. See, for instance, the projections proposed by Quesada and Grossmann [207], the reformulation–linearization technique by Sherali and Alameddine [220], and the reformulation–convexification approach by Sherali and Tuncbilek [221].

The Set of Branching Variables. A set of branching variables, characterized by the index set BV (Ω) defined below, is determined by considering the optimal solution $(\hat{x},\hat{y},\hat{z})_{\Omega}$ of the underestimating problem:

$$\begin{aligned}&\mathrm{BV}(\Omega)=\{i,j\colon |\hat{y}_{ij}-\hat{x}_i\hat{x}_j|=\zeta_l \text{ or } |\hat{z}_{ij}-\hat{x}_i/\hat{x}_j|=\zeta_l\\ &\text{or } g_i(\hat{x}_i)-\hat{g}_i(\hat{x}_i)=\zeta_l \text{ or } g_{i,k}(\hat{x}_i)-\hat{g}_{i,k}(\hat{x}_i)=\zeta_l,\\ &\text{for } i\in I, j\in J, k\in K, l\in L\}\end{aligned} \tag{88}$$

where, for a prespecified number l_n, $L=\{1,2,\ldots,l_n\}$ and ζ_1 is the magnitude of the largest approximation error for a nonconvex term in problem 80 evaluated at $(\hat{x},\hat{y},\hat{z})_{\Omega}$:

$$\begin{aligned}&\xi_1=\mathrm{Max}\ [|\hat{y}_{ij}-\hat{x}_i\hat{x}_j|,\ |\hat{z}_{ij}-\hat{x}_i/\hat{x}_j|,\ g_i(\hat{x}_i)-\hat{g}_i(\hat{x}_i),\\ &g_{i,k}(\hat{x}_i)-\hat{g}_{i,k}(\hat{x}_i)]\\ &i\in I, j\in J, k\in K\end{aligned}$$

Similarly, we define $\xi_l<\xi_{l-1}$ with $l\in L\setminus\{1\}$ as the l-th largest magnitude for an approximation error; for instance, $\xi_2<\xi_1$ is the second largest magnitude for an approximation error. Note that in some cases it might be convenient to introduce weights in the determination of BV (Ω) in order to scale differences in the approximation errors or to induce preferential branching schemes. This might be particularly useful in applications where specific information can be exploited by imposing an order of precedence on the set of complicating variables.

This basic concept in spatial branch-and-bound for global optimization is as follows. *Bounds* are related to the calculation of upper and lower bounds. For the former, any feasible point or, preferably, a locally optimal point in the subregion, can be used. For the lower bound, convex relaxations of the objective and constraint function are derived such as in problem 80. The *refining* step deals with the construction of partitions in the domain and further partitioning them during the search process. Finally, the *selection* step decides on the order of exploring the open subregions. Thus, the feasible region and the objective function are replaced by convex envelopes to form relaxed problems. Solving these convex relaxed problems leads to global solutions that are lower bounds to the NLP in the particular subregion. Finally, we see that gradient-based NLP solvers play an important role in global optimization algorithms, as they often yield the lower and upper bounds for the subregions. The following *spatial branch-and-bound* global optimization algorithm can therefore be given by the following steps:

0. Initialize algorithm: calculate upper bound by obtaining a local solution to problem 80. Calculate a lower bound solving problem 80 over the entire (relaxed) feasible region Ω_0.

For iteration k with a set of partitions $\Omega_{k,j}$ and bounds in each subregion OLB_j and OUB_j:

1) *Bound:* Define best upper bound: OUB = Min_j OUB_j and delete (fathom) all subregions j with lower bounds $\mathrm{OLB}_j \geq \mathrm{OUB}-\varepsilon$, stop.
2) *Refine:* Divide the remaining active subregions into partitiions $\Omega_{k,j1}$ and $\Omega_{k,j2}$. (Several branching rules are available for this step.)
3) *Select:* Solve the convex NLP 80 in the new partitions to obtain OLB_{j1} and OLB_{j2}. Delete partition if no feasible solution.
4) *Update:* Obtain upper bounds, OUB_{j1} and OUB_{j2} to new partitions if present. Set $k=k+1$, update partition sets and go to step 1.

Example: To illustrate the spatial branch-and-bound algorithm, consider the global solution of:

$$\begin{aligned}&\mathrm{Min}\ f(x)=5/2\,x^4-20\,x^3+55\,x_2-57\,x\\ &\text{s.t. } 0.5\leq x\leq 2.5\end{aligned} \tag{89}$$

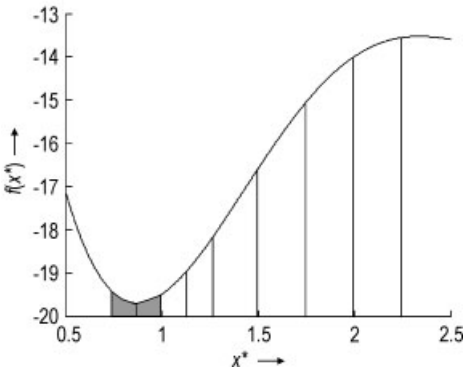

Figure 43. Global optimization example with partitions

As seen in Figure 43, this problem has local solutions at $x^* = 2.5$ and at $x^* = 0.8749$. The latter is also the global solution with $f(x^*) = -19.7$. To find the global solution we note that all but the $-20\ x^3$ term in problem 89 are convex, so we replace this term by a new variable and a linear underestimator within a particular subregion, i.e.:

$$\begin{aligned} \text{Min} \quad & f_L\,(x) = 5/2\,x^4 - 20\,w + 55\,x_2 - 57\,x \\ \text{s.t.} \quad & x_1 \le x \le x_u \\ & w = (x_l)^3\,\frac{(x_u - x)}{(x_u - x_l)} + (x_u)^3\,\frac{(x - x_l)}{(x_u - x_l)} \end{aligned} \tag{90}$$

In Figure 43 we also propose subregions that are created by simple bisection partitioning rules, and we use a "loose" bounding tolerance of $\varepsilon = 0.2$. In each partition the lower bound, f_{L} is determined by problem 90 and the upper bound f_{U} is determined by the local solution of the original problem in the subregion. Figure 44 shows the progress of the spatial branch-and-bound algorithm as the partitions are refined and the bounds are updated. In Figure 43, note the definitions of the partitions for the nodes, and the sequence numbers in each node that show the order in which the partitions are processed. The grayed partitions correspond to the deleted subregions and at termination of the algorithm we see that $f_{\mathrm{L}j} \ge f_{\mathrm{U}} - \varepsilon$ (i.e., $-19.85 \ge -19.7 - 0.2$), with the gray subregions in Figure 43 still active. Further partitioning in these subregions will allow the lower and upper bounds to converge to a tighter tolerance.

A number of improvements can be made to the bounding, refinement, and selection strategies in the algorithm that accelerate the convergence of this method. A comprehensive discussion of all of these options can be found in [222 – 224]. Also, a number of efficient global optimization codes have recently been developed, including αBB, BARON, LGO, and OQNLP. An interesting numerical comparison of these and other codes can be found in [225].

10.5. Mixed Integer Programming

Mixed integer programming deals with both discrete and continuous decision variables. For simplicity in the presentation we consider the most common case where the discrete decisions are binary variables, i.e., $y_i = 0$ or 1, and we consider the mixed integer problem 51. Unlike local optimization methods, there are no optimality conditions, like the KKT conditions, that can be applied directly.

Mixed Integer Linear Programming. If the objective and constraint functions are all linear in problem 51, and we assume 0–1 binary variables for the discrete variables, then this gives rise to a mixed integer linear programming (MILP) problem given by Equation 91.

$$\begin{aligned} \text{Min} \quad & Z = a^T x + c^T y \\ \text{s.t.} \quad & Ax + By \le b \\ & x \ge 0,\ y \in \{0,1\}^t \end{aligned} \tag{91}$$

As is well known, the (MILP) problem is NP-hard. Nevertheless, an interesting theoretical result is that it is possible to transform it into an LP with the convexification procedures proposed by Lovacz and Schrijver [226], Sherali and Adams [227], and Balas et al. [228]. These procedures consist of sequentially lifting the original relaxed $x - y$ space into higher dimension and projecting it back to the original space so as to yield, after a finite number of steps, the integer convex hull. Since the transformations have exponential complexity, they are only of theoretical interest, although they can be used as a basis for deriving cutting planes (e.g. lift and project method by [228]).

As for the solution of problem (MILP), it should be noted that this problem becomes an LP problem when the binary variables are relaxed as continuous variables $0 \le y \le 1$. The most common solution algorithms for problem

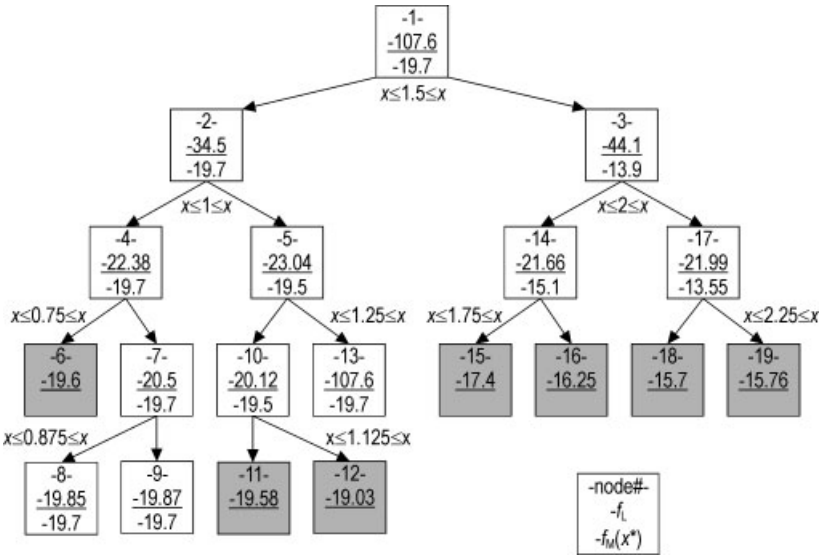

Figure 44. Spatial branch-and-bound sequence for global optimization example

(MILP) are LP-based branch-and-bound methods, which are enumeration methods that solve LP subproblems at each node of the search tree. This technique was initially conceived by LAND and DOIG [229], BALAS [230], and later formalized by DAKIN, [231]. Cutting-plane techniques, which were initially proposed by GOMORY [232], and consist of successively generating valid inequalities that are added to the relaxed LP, have received renewed interest through the works of CROWDER et al. [233], VAN ROY and WOLSEY [234], and especially the lift-and-project method of BALAS et al. [228]. A recent review of branch-and-cut methods can be found in [235]. Finally, Benders decomposition [236] is another technique for solving MILPs in which the problem is successively decomposed into LP subproblems for fixed 0–1 and a master problem for updating the binary variables.

LP-Based Branch and Bound Method. We briefly outline in this section the basic ideas behind the branch-and-bound method for solving MILP problems. Note that if we relax the t binary variables by the inequalities $0 \leq y \leq 1$ then 91 becomes a linear program with a (global) solution that is a lower bound to the MILP 91. There are specific MILP classes in which the LP relaxation of 91 has the same solution as the MILP. Among these problems is the well-known *assignment problem*. Other MILPs that can be solved with efficient special-purpose methods are the *knapsack* problem, the *set-covering* and *set-partitioning* problems, and the *traveling salesman* problem. See [237] for a detailed treatment of these problems.

The branch-and-bound algorithm for solving MILP problems [231] is similar to the spatial branch-and-bound method of the previous section that explores the search space. As seen in Figure 45, binary variables are successively fixed to define the search tree and a number of bounding properties are exploited in order to fathom nodes in to avoid exhaustive enumeration of all the nodes in the tree.

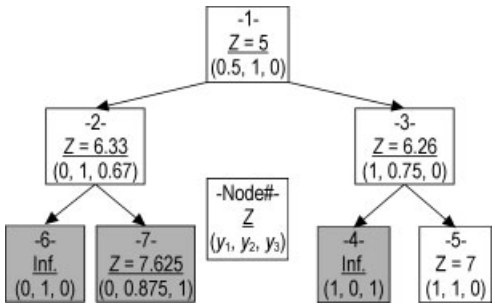

Figure 45. Branch and bound sequence for MILP example

The basic idea in the search is as follows. The top, or root node, in the tree is the solution to the linear programming relaxation of 91. If all the

y variables take on 0–1 values, the MILP problem is solved, and no further search is required. If at least one of the binary variables yields a fractional value, the solution of the LP relaxation yields a lower bound to problem 91. The search then consists of branching on that node by fixing a particular binary variable to 0 or 1, and the corresponding restricted LP relaxations are solved that in turn yield a lower bound for any of their descendant nodes. In particular, the following properties are exploited in the branch-and-bound search:

- Any node (initial, intermediate, leaf node) that leads to feasible LP solution corresponds to a valid upper bound to the solution of the MILP problem 91.
- Any intermediate node with an infeasible LP solution has infeasible leaf nodes and can be fathomed (i.e., all remaining children of this node can be eliminated).
- If the LP solution at an intermediate node is not less than an existing integer solution, then the node can be fathomed.

These properties lead to pruning of the nodes in the search tree. Branching then continues in the tree until the upper and lower bounds converge.

The basic concepts outlined above lead to a branch-and-bound algorithm with the following features. LP solutions at intermediate nodes are relatively easy to calculate since they can be effectively updated with the dual simplex method. The selection of binary variables for branching, known as the *branching rule,* is based on a number of different possible criteria; for instance, choosing the fractional variable closest to 0.5, or the one involving the largest of the smallest pseudocosts for each fractional variable. *Branching strategies* to navigate the tree take a number of forms. More common *depth-first strategies* expand the most recent node to a leaf node or infeasible node and then backtrack to other branches in the tree. These strategies are simple to program and require little storage of past nodes. On the other hand, *breadth-first strategies* expand all the nodes at each level of the tree, select the node with the lowest objective function, and then proceed until the leaf nodes are reached. Here, more storage is required, but generally fewer nodes are evaluated than in depth-first search. In practice, a combination of both strategies is commonly used: branch on the dichotomy 0–1 at each node (i.e., like breadth-first), but expand as in depth-first. Additional description of these strategies can be found in [237].

Example: To illustrate the branch-and-bound approach, we consider the MILP:

$$\begin{aligned} \text{Min} \quad & Z = x + y_1 + 2y_2 + 3y_3 \\ \text{s.t.} \quad & -x + 3y_1 + y_2 + 2y_3 \le 0 \\ & -4y_1 - 8y_2 - 3y_3 \le -10 \\ & x \ge 0, y_1, y_2, y_3 = \{0,1\} \end{aligned} \tag{92}$$

The solution to problem 92 is given by $x = 4$, $y_1 = 1$, $y_2 = 1$, $y_3 = 0$, and $Z = 7$. Here we use a depth-first strategy and branch on the variables closest to zero or one. Figure 45 shows the progress of the branch-and-bound algorithm as the binary variables are selected and the bounds are updated. The sequence numbers for each node in Figure 45 show the order in which they are processed. The grayed partitions correspond to the deleted nodes and at termination of the algorithm we see that $Z = 7$ and an integer solution is obtained at an intermediate node where coincidentally $y_3 = 0$.

Mixed integer linear programming (MILP) methods and codes have been available and applied to many practical problems for more than twenty years (e.g., [237]. The LP-based branch-and-bound method [231] has been implemented in powerful codes such as OSL, CPLEX, and XPRESS. Recent trends in MILP include the development of branch-and-price [238] and branch-and-cut methods such as the lift-and-project method [228] in which cutting planes are generated as part of the branch-and-bound enumeration. See also [235] for a recent review on MILP. A description of several MILP solvers can also be found in the NEOS Software Guide.

Mixed Integer Nonlinear Programming The most basic form of an MINLP problem when represented in algebraic form is as follows:

$$\begin{aligned} & \min z = f(x,y) \\ & s.t.\ g_j(x,y) \le 0 \quad j \in J \qquad \text{(P1)} \\ & x \in X,\ y \in Y \end{aligned} \tag{93}$$

where $f(\cdot)$, $g(\cdot)$ are *convex, differentiable* functions, J is the index set of inequalities, and x and y are the continuous and discrete

variables, respectively. The set X is commonly assumed to be a convex compact set, e.g., $X=\{x|x\in\mathbf{R}^n, Dx\leq d, x^L\leq x\leq x^U\}$; the discrete set Y corresponds to a polyhedral set of integer points, $Y=\{y|x\in Z^m, Ay\leq a\}$, which in most applications is restricted to 0–1 values, $y\in\{0,1\}^m$. In most applications of interest the objective and constraint functions $f(\cdot)$, $g(\cdot)$ are linear in y (e.g., fixed cost charges and mixed-logic constraints): $f(x,y)=c^T y+r(x)$, $g(x,y)=By+h(x)$.

Methods that have addressed the solution of problem 93 include the branch-and-bound method (BB) [239 – 243], generalized Benders decomposition (GBD) [244], outer-approximation (OA) [245 – 247], LP/NLP based branch-and-bound [248], and extended cutting plane method (ECP) [249].

There are three basic *NLP subproblems* that can be considered for problem 93:

a) NLP relaxation

$$\begin{aligned} &\text{Min } Z_{LB}^k = f(x,y) \\ &s.t.\ g_j(x.y)\leq 0\ j\in J \\ &\quad x\in X, y\in Y_R \qquad \text{(NLP1)} \\ &\quad y_i\leq\alpha_i^k\ i\in I_{FL}^k \\ &\quad y_i\geq\beta_i^k\ i\in I_{FU}^k \end{aligned} \tag{94}$$

where Y_{R} is the continuous relaxation of the set Y, and $I_{\mathrm{FL}}^k, I_{\mathrm{FU}}^k$ are index subsets of the integer variables y_i, $i\in I$ which are restricted to lower and upper bounds α_i^k, β_i^k at the k-th step of a branch-and-bound enumeration procedure. Note that $\alpha_i^k=\lfloor y_i^l\rfloor, \beta_i^k=\lceil y_i^m\rceil, l<k, m<k$, where y_i^l, y_i^m are noninteger values at a previous step, and $\lfloor\cdot\rfloor$, $\lceil\cdot\rceil$ are the floor and ceiling functions, respectively.

Also note that if $I_{\mathrm{FU}}^k=I_{\mathrm{FL}}^k=\varnothing\ (k=0)$, problem 94 corresponds to the continuous NLP relaxation of problem 93. Except for few and special cases, the solution to this problem yields in general a noninteger vector for the discrete variables. Problem 94 also corresponds to the k-th step in a branch-and-bound search. The optimal objective function $Z_{\mathrm{LB}}^{\mathrm{o}}$ provides an absolute lower bound to problem 93; for $m\geq k$, the bound is only valid for $I_{\mathrm{FL}}^k\subset I_{\mathrm{FL}}^m$, $I_{\mathrm{FU}}^k\subset I_{\mathrm{FL}}^m$.

b) NLP subproblem for fixed y^k:

$$\begin{aligned} &\text{Min}\quad Z_{\mathrm{U}}^k = f(x,y^k) \\ &\text{s.t.}\quad g_j(x,y^k)\leq 0\quad j\in J \\ &\qquad x\in X \end{aligned} \tag{95}$$

which yields an upper bound Z_{U}^k to problem 93 provided problem 95 has a feasible solution. When this is not the case, we consider the next subproblem:

c) Feasibility subproblem for fixed y^{k}:

$$\begin{aligned} &\text{Min } u \\ &\text{s.t. } g_j(x,y^k)\leq u\ j\in J \\ &\quad x\in X, u\in R^1 \end{aligned} \tag{96}$$

which can be interpreted as the minimization of the infinity-norm measure of infeasibility of the corresponding NLP subproblem. Note that for an infeasible subproblem the solution of problem 96 yields a strictly positive value of the scalar variable u.

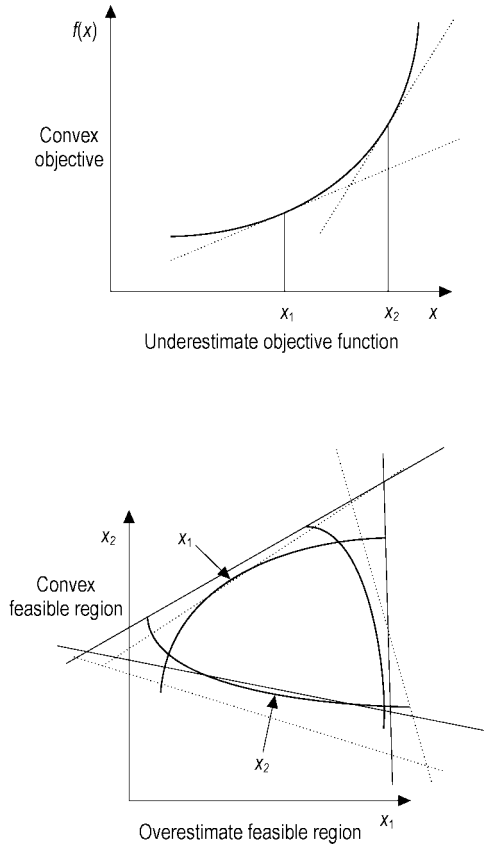

Figure 46. Geometrical interpretation of linearizations in master problem 97

The convexity of the nonlinear functions is exploited by replacing them with supporting hyperplanes, which are generally, but not necessarily, derived at the solution of the NLP subproblems. In particular, the new values y^K (or (x^K, y^K) are obtained from a *cutting-plane MILP problem* that is based on the K points (x^k, y^k), $k = 1 \ldots K$, generated at the K previous steps:

$$\begin{array}{l} \text{Min } Z_{\mathrm{L}}^K = \alpha \\ \text{st} \quad \left.\begin{array}{l} \alpha \geq f(x^k, y^k) \\ \quad + \nabla f(x^k, y^k)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \\ g_j(x^k, y^k) \\ \quad + \nabla g_j(x^k, y^k)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \leq 0 \; j \in J^k \end{array}\right\} k = 1, \ldots K \\ \qquad x \in X, \; y \in Y \end{array} \tag{97}$$

where $J^k \supseteq J$. When only a subset of linearizations is included, these commonly correspond to violated constraints in problem 93. Alternatively, it is possible to include all linearizations in problem 97. The solution of 97 yields a valid lower bound Z_{L}^k to problem 93. This bound is nondecreasing with the number of linearization points K. Note that since the functions $f(x,y)$ and $g(x,y)$ are convex, the linearizations in problem 97 correspond to outer approximations of the nonlinear feasible region in problem 93. A geometrical interpretation is shown in Figure 46, where it can be seen that the convex objective function is being underestimated, and the convex feasible region overestimated with these linearizations.

Algorithms. The different methods can be classified according to their use of the subproblems 94–96, and the specific specialization of the MILP problem 97, as seen in Figure 47. In the GBD and OA methods (case b), as well in the LP/NLP-based branch-and-bound mehod (case d), problem 96 is solved if infeasible subproblems are found. Each of the methods is explained next in terms of the basic subproblems.

Branch and Bound. While the earlier work in branch and bound (BB) was aimed at linear problems [231] this method can also be applied to nonlinear problems [239 – 243]. The BB method starts by solving first the continuous NLP relaxation. If all discrete variables take integer values the search is stopped. Otherwise, a tree search is performed in the space of the integer variables y_i, $i \in I$. These are successively fixed at the corresponding nodes of the tree, giving rise to relaxed NLP subproblems of the form (NLP1) which yield lower bounds for the subproblems in the descendant nodes. Fathoming of nodes occurs when the lower bound exceeds the current upper bound, when the subproblem is infeasible, or when all integer variables y_i take on discrete values. The last-named condition yields an upper bound to the original problem.

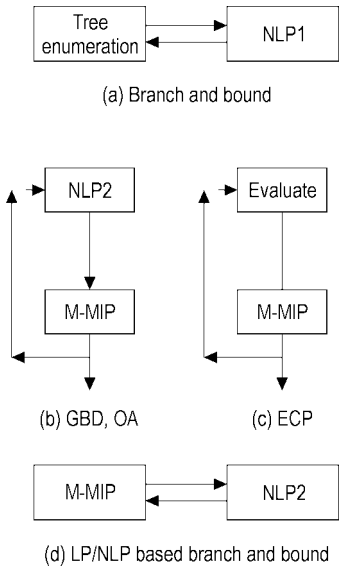

Figure 47. Major steps in the different algorithms

The BB method is generally only attractive if the NLP subproblems are relatively inexpensive to solve, or when only few of them need to be solved. This could be either because of the low dimensionality of the discrete variables, or because the integrality gap of the continuous NLP relaxation of 93 is small.

Outer Approximation [245 – 247]**.** The OA method arises when NLP subproblems 95 and MILP master problems 97 with $J^k = J$ are solved successively in a cycle of iterations to generate the points (x^k, y^k). Since the master problem 97 requires the solution of all feasible discrete variables y^k, the following MILP

relaxation is considered, assuming that the solution of K different NLP subproblems ($K = |\mathrm{KFS} \cup \mathrm{KIS}|$), where KFS is a set of solutions from problem 95, and KIS set of solutions from problem 96 is available:

$$\begin{aligned} &\text{Min } Z_{\mathrm{L}}^{K} = \alpha \\ &\text{st} \quad \left.\begin{aligned} &\alpha \geq f(x^k, y^k) \\ &\quad + \nabla f(x^k, y^k)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \\ &g_j(x^k, y^k) \\ &\quad + \nabla g_j(x^k, y^k)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \leq 0 \quad j \in J \end{aligned}\right\} k = 1, \ldots K \\ &\qquad x \in X, y \in Y \end{aligned} \tag{98}$$

Given the assumption on convexity of the functions $f(x,y)$ and $g(x,y)$, it can be proved that the solution of problem 98 Z_{L}^{k} corresponds to a lower bound of the solution of problem 93. Note that this property can be verified in Figure 46. Also, since function linearizations are accumulated as iterations proceed, the master problems 98 yield a nondecreasing sequence of lower bounds $Z_{\mathrm{L}}^{1} \ldots \leq Z_{\mathrm{L}}^{k} \ldots Z_{\mathrm{L}}^{k}$ since linearizations are accumulated as iterations k proceed.

The OA algorithm as proposed by Duran and Grossmann consists of performing a cycle of major iterations, $k = 1,..K$, in which problem 95 is solved for the corresponding y^K, and the relaxed MILP master problem 98 is updated and solved with the corresponding function linearizations at the point (x^k, y^k) for which the corresponding subproblem NLP2 is solved. If feasible, the solution to that problem is used to construct the first MILP master problem; otherwise a feasibility problem 96 is solved to generate the corresponding continuous point [247]. The initial MILP master problem 98 then generates a new vector of discrete variables. The subproblems 95 yield an upper bound that is used to define the best current solution, $\mathrm{UB}^k = \min_k \left\{ Z_{\mathrm{U}}^{k} \right\}$. The cycle of iterations is continued until this upper bound and the lower bound of the relaxed master problem Z_{L}^{k} are within a specified tolerance. One way to avoid solving the feasibility problem 96 in the OA algorithm when the discrete variables in problem 93 are 0–1, is to introduce the following integer cut whose objective is to make infeasible the choice of the previous 0–1 values generated at the K previous iterations [245]:

$$\sum_{i \in B^k} y_i - \sum_{i \in N^k} y_i \leq \left| B^k \right| - 1 \quad k = 1, \ldots K \tag{99}$$

where $B^k = \left\{ i | y_i^k = 1 \right\}$, $N^k = \left\{ i | y_i^k = 0 \right\}$, $k = 1, \ldots K$. This cut becomes very weak as the dimensionality of the 0–1 variables increases. However, it has the useful feature of ensuring that new 0–1 values are generated at each major iteration. In this way the algorithm will not return to a previous integer point when convergence is achieved. Using the above integer cut the termination takes place as soon as $Z_{\mathrm{L}}^{K} \geq \mathrm{UB}^K$.

The OA algorithm trivially converges in one iteration if $f(x,y)$ and $g(x,y)$ are linear. This property simply follows from the fact that if $f(x,y)$ and $g(x,y)$ are linear in x and y the MILP master problem 98 is identical to the original problem 93. It is also important to note that the MILP master problem need not be solved to optimality.

Generalized Benders Decomposition (GBD) [244]**.** The GBD method [250] is similar to the outer-approximation method. The difference arises in the definition of the MILP master problem 97. In the GBD method only active inequalities are considered $J^k = \{ j | \, g_j(x^k, y^k) = 0 \}$ and the set $x \in X$ is disregarded. In particular, consider an outer-approximation given at a given point (x^k, y^k)

$$\begin{aligned} &\alpha \geq f\left(x^k, y^k\right) + \nabla f\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \\ &g\left(x^k, y^k\right) + \nabla g\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \leq 0 \end{aligned} \tag{100}$$

where for a fixed y^k the point x^k corresponds to the optimal solution to problem 95. Making use of the Karush – Kuhn – Tucker conditions and eliminating the continuous variables x, the inequalities in 100 can be reduced as follows [248]:

$$\begin{aligned} &\alpha \geq f\left(x^k, y^k\right) + \nabla_y f\left(x^k, y^k\right)^T \left(y - y^k\right) + \\ &\left(\mu^k\right)^T \left[g\left(x^k, y^k\right) + \nabla_y g\left(x^k, y^k\right)^T \left(y - y^k\right) \right] \end{aligned} \tag{101}$$

which is the Lagrangian cut projected in the y-space. This can be interpreted as a surrogate constraint of the equations in 100, because it is obtained as a linear combination of these.

For the case when there is no feasible solution to problem 95, then if the point x^k is obtained from the feasibility subproblem (NLPF), the following feasibility cut projected in y can be obtained using a similar procedure.

$$\left(\lambda^k\right)^T \left[g\left(x^k,y^k\right)+\nabla_y g\left(x^k,y^k\right)^T\left(y-y^k\right)\right] \leq 0 \tag{102}$$

In this way, problem 97 reduces to a problem projected in the y-space:

$$\begin{aligned}
&\text{Min} \quad Z_{\mathrm{L}}^K=\alpha \\
&\text{st}\ \alpha \geq \quad f\left(x^k,y^k\right)+\nabla_y f\left(x^k,y^k\right)^T\left(y-y^k\right)+ \\
&\left(\mu^k\right)^T\left[g\left(x^k,y^k\right)+\nabla_y g\left(x^k,y^k\right)^T\left(y-y^k\right)\right] \\
&\qquad k\in KFS \\
&\left(\lambda^k\right)^T\left[g\left(x^k,y^k\right)+\nabla_y g\left(x^k,y^k\right)^T\left(y-y^k\right)\right] \leq 0 \\
&\qquad k\in KIS \\
&x\in X, \alpha\in R^1
\end{aligned}$$

where KFS is the set of feasible subproblems 95, and KIS the set of infeasible subproblems whose solution is given by problem 96. Also $|\mathrm{KFS} \cup \mathrm{KIS}| = K$. Since master problem 103 can be derived from master problem 98, in the context of problem 93, GBD can be regarded as a particular case of the outer-approximation algorithm. In fact one can prove that given the same set of K subproblems, the lower bound predicted by the relaxed master problem 98 is greater than or equal to that predicted by the relaxed master problem 103 [245]. This proof follows from the fact that the Lagrangian and feasibility cuts 101 and 102 are surrogates of the outer-approximations 100. Given the fact that the lower bounds of GBD are generally weaker, this method commonly requires a larger number of cycles or major iterations. As the number of 0–1 variables increases this difference becomes more pronounced. This is to be expected since only one new cut is generated per iteration. Therefore, user-supplied constraints must often be added to the master problem to strengthen the bounds. Also, it is sometimes possible to generate multiple cuts from the solution of an NLP subproblem in order to strengthen the lower bound [251]. As for the OA algorithm, the trade-off is that while it generally predicts stronger lower bounds than GBD, the computational cost for solving the master problem (M-OA) is greater, since the number of constraints added per iteration is equal to the number of nonlinear constraints plus the nonlinear objective.

If problem 93 has zero integrality gap, the GBD algorithm converges in one iteration once the optimal (x^*, y^*) is found [252]. This property implies that the only case in which one can expect the GBD method to terminate in one iteration is that in which the initial discrete vector is the optimum, and when the objective value of the NLP relaxation of problem 93 is the same as the objective of the optimal mixed-integer solution.

Extended Cutting Plane (ECP) [249]**.** The ECP method, which is an extension of Kelley's cutting-plane algorithm for convex NLP [253], does not rely on the use of NLP subproblems and algorithms. It relies only on the iterative solution of problem 97 by successively adding a linearization of the most violated constraint at the predicted point (x^k, y^k):

$$J^k=\{\hat{j}\in\arg\left\{\max_{j\in J} g_j\left(x^k,y^k\right)\right\}\}$$

Convergence is achieved when the maximum constraint violation lies within the specified tolerance. The optimal objective value of problem 97 yields a nondecreasing sequence of lower bounds. It is of course also possible to either add to problem 97 linearizatons of all the violated constraints in the set J^k, or linearizations of all the nonlinear constraints $j \in J$. In the ECP method the objective must be defined as a linear function, which can easily be accomplished by introducing a new variable to transfer nonlinearities in the objective as an inequality.

Note that since the discrete and continuous variables are converged simultaneously, the ECP method may require a large number of iterations. However, this method shares with the OA method Property 2 for the limiting case when all the functions are linear.

LP/NLP-Based Branch and Bound [248]**.** This method is similar in spirit to a branch-and-cut method, and avoids the complete solution of the MILP master problem (M-OA) at each major iteration. The method starts by solving an initial NLP subproblem, which is linearized as in (M-OA). The basic idea consists then of performing an LP-based branch-and-bound method for (M-OA) in which NLP subproblems 95 are solved at those nodes in which feasible integer solutions

are found. By updating the representation of the master problem in the current open nodes of the tree with the addition of the corresponding linearizations, the need to restart the tree search is avoided.

This method can also be applied to the GBD and ECP methods. The LP/NLP method commonly reduces quite significantly the number of nodes to be enumerated. The trade-off, however, is that the number of NLP subproblems may increase. Computational experience has indicated that often the number of NLP subproblems remains unchanged. Therefore, this method is better suited for problems in which the bottleneck corresponds to the solution of the MILP master problem. Leyffer [254] has reported substantial savings with this method.

Example: Consider the following MINLP problem, whose objective function and constraints contain nonlinear convex terms

$$\begin{aligned} &\text{Min } Z = y_1 + 1.5y_2 + 0.5y_3 + x_1^2 + x_2^2 \\ &s.t. \quad (x_1-2)^2 - x_2 \le 0 \\ &\qquad x_1 - 2y_1 \ge 0 \\ &\qquad x_1 - x_2 - 4(1-y_2) \ge 0 \\ &\qquad x_2 - y_2 \ge 0 \\ &\qquad x_1 + x_2 \ge 3y_3 \\ &\qquad y_1 + y_2 + y_3 \ge 1 \\ &\qquad 0 \le x_1 \le 4, 0 \le x_2 \le 4 \\ &\qquad y_1, y_2, y_3 = 0,1 \end{aligned} \tag{103}$$

The optimal solution of this problem is given by $y_1 = 0$, $y_2 = 1$, $y_3 = 0$, $x_1 = 1$, $x_2 = 1$, $Z = 3.5$. Figure 48 shows the progress of the iterations with the OA and GBD methods, while the table lists the number of NLP and MILP subproblems that are solved with each of the methods. For the case of the MILP problems the total number of LPs solved at the branch-and-bound nodes are also reported.

Extensions of MINLP Methods. Extensions of the methods described above include a quadratic approximation to (RM-OAF) [247] using an approximation of the Hessian matrix. The quadratic approximations can help to reduce the number of major iterations, since an improved representation of the continuous space is obtained. This, however, comes at the price of having to solve an MIQP instead of an MILP at each iteration.

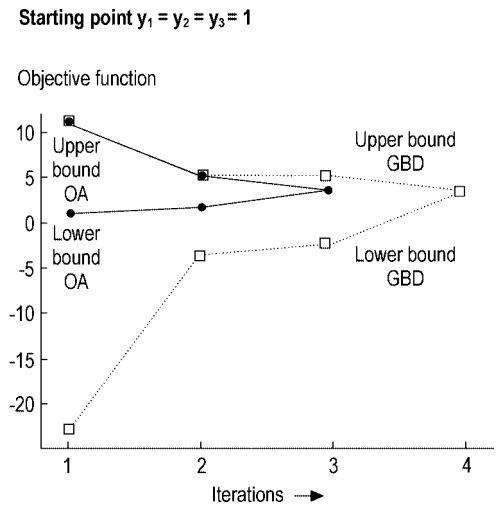

Figure 48. Progress of iterations with OA and GBD methods, and number of subproblems for the BB, OA, GBD, and ECP methods.

The master problem 98 can involve a rather large number of constraints, due to the accumulation of linearizations. One option is to keep only the last linearization point, but this can lead to nonconvergence even in convex problems, since then the monotonic increase of the lower bound is not guaranteed. As shown [248], linear approximations to the nonlinear objective and constraints can be aggregated with an MILP master problem that is a hybrid of the GBD and OA methods.

For the case when linear equalities of the form $h(x, y) = 0$ are added to problem 93 there is no major difficulty, since these are invariant to the linearization points. If the equations are nonlinear, however, there are two difficulties. First, it is not possible to enforce the linearized equalities at K points. Second, the nonlinear equations may generally introduce nonconvexities, unless they relax as convex inequalities [255]. Kocis and Grossmann [256] proposed an equality relaxation strategy in which the nonlinear equalities are replaced by the inequalities

$$T^k \nabla h\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \le 0 \tag{104}$$

where $T^k = \{t_{ii}^k\}$, and $t_{ii}^k = \text{sign}\left(\lambda_i^k\right)$ sign in which λ_i^k is the multiplier associated to the equation $h_i(x, y) = 0$. Note that if these equations relax as the inequalities $h(x, y) \leq 0$ for all y and $h(x, y)$ is convex, this is a rigorous procedure. Otherwise, nonvalid supports may be generated. Also, in the master problem 103 of GBD, no special provision is required to handle equations, since these are simply included in the Lagrangian cuts. However, similar difficulties as in OA arise if the equations do not relax as convex inequalities.

When $f(x,y)$ and $g(x,y)$ are nonconvex in problem 93, or when nonlinear equalities $h(x, y) = 0$ are present, two difficulties arise. First, the NLP subproblems 94 – 96 may not have a unique local optimum solution. Second, the master problem (M-MIP) and its variants (e.g., M-MIPF, M-GBD, M-MIQP) do not guarantee a valid lower bound Z_{L}^K or a valid bounding representation with which the global optimum may be cut off.

Rigorous global optimization approaches for addressing nonconvexities in MINLP problems can be developed when special structures are assumed in the continuous terms (e.g. bilinear, linear fractional, concave separable). Specifically, the idea is to use convex envelopes or underestimators to formulate lower-bounding convex MINLP problems. These are then combined with global optimization techniques for continuous variables [206, 207, 209, 214, 216, 222, 257], which usually take the form of spatial branch-and-bound methods. The lower bounding MINLP problem has the general form,

$$\begin{aligned} &\text{Min } Z = \overline{f}\,(x,y) \\ &\text{s.t. } \overline{g}_j\,(x,y) \leq 0 \; j \in J \\ &\quad x \in X,\, y \in Y \end{aligned} \tag{105}$$

where $\overline{f}, \overline{g}$ are valid convex underestimators such that $\overline{f}\,(x,y) \leq f\,(x,y)$ and the inequalities $\overline{g}\,(x,y) \leq 0$ are satisfied if $g\,(x,y) \leq 0$. A typical example of convex underestimators are the convex envelopes for bilinear terms [208].

Examples of global optimization methods for MINLP problems include the branch-and-reduce method [209, 210], the α-BB method [212], the reformulation/spatial branch-and-bound search method [258], the branch-and-cut method [259], and the disjunctive branch-and-bound method [260]. All these methods rely on a branch-and-bound procedure. The difference lies in how to perform the branching on the discrete and continuous variables. Some methods perform the spatial tree enumeration on both the discrete and continuous variables of problem 105. Other methods perform a spatial branch and bound on the continuous variables and solve the corresponding MINLP problem 105 at each node using any of the methods reviewed above. Finally, other methods branch on the discrete variables of problem 105, and switch to a spatial branch and bound on nodes where a feasible value for the discrete variables is found. The methods also rely on procedures for tightening the lower and upper bounds of the variables, since these have a great effect on the quality of the underestimators. Since the tree searches are not finite (except for ε convergence), these methods can be computationally expensive. However, their major advantage is that they can rigorously find the global optimum. Specific cases of nonconvex MINLP problems have been handled. An example is the work of Pörn and Westerlund [261], who addressed the solution of MINLP problems with pseudoconvex objective function and convex inequalities through an extension of the ECP method.

The other option for handling nonconvexities is to apply a heuristic strategy to try to reduce as much as possible the effect of nonconvexities. While not being rigorous, this requires much less computational effort. We describe here an approach for reducing the effect of nonconvexities at the level of the MILP master problem.

Viswanathan and Grossmann [262] proposed to introduce slacks in the MILP master problem to reduce the likelihood of cutting off feasible solutions. This master problem (augmented penalty/equality relaxation) has the form:

$$\begin{aligned} &\min Z^K = \alpha + \sum_{k=1}^{K} w_p^k\, p^k + w_q^k\, q^k \\ &\left.\begin{aligned} \text{s.t.}\quad & \alpha \geq f\left(x^k, y^k\right) \\ & \quad + \nabla f\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \\ & T^k \nabla h\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \leq p^k \\ & g\left(x^k, y^k\right) + \nabla g\left(x^k, y^k\right)^T \begin{bmatrix} x - x^k \\ y - y^k \end{bmatrix} \\ & \quad \leq q^k \end{aligned}\right\} k = 1, \ldots K \end{aligned} \tag{106}$$

$$\sum_{i\in B^k} y_i - \sum_{i\in N^k} y_i \leq \left|B^k\right| - 1 \; k=1,\ldots K$$
$$x\in X, y\in Y, \alpha\in R^1, p^k, q^k \geq 0$$

where w_p^k, w_q^k are weights that are chosen sufficiently large (e.g., 1000 times the magnitude of the Lagrange multiplier). Note that if the functions are convex then the MILP master problem 106 predicts rigorous lower bounds to problem 93 since all the slacks are set to zero.

Computer Codes for MINLP. Computer codes for solving MINLP problems include the following. The program DICOPT [262] is an MINLP solver that is available in the modeling system GAMS [263]. The code is based on the master problem 106 and the NLP subproblems 95. This code also uses relaxed 94 to generate the first linearization for the above master problem, with which the user need not specify an initial integer value. Also, since bounding properties of problem 106 cannot be guaranteed, the search for nonconvex problems is terminated when there is no further improvement in the feasible NLP subproblems. This is a heuristic that works reasonably well in many problems. Codes that implement the branch-and-bound method using subproblems 94 include the code MINLP_BB, which is based on an SQP algorithm [243] and is available in AMPL, the code BARON [264], which also implements global optimization capabilities, and the code SBB, which is available in GAMS [263]. The code a-ECP implements the extended cutting-plane method [249], including the extension by Pörn and Westerlund [261]. Finally, the code MINOPT [265] also implements the OA and GBD methods, and applies them to mixed-integer dynamic optimization problems. It is difficult to derive general conclusions on the efficiency and reliability of all these codes and their corresponding methods, since no systematic comparison has been made. However, one might anticipate that branch-and-bound codes are likely to perform better if the relaxation of the MINLP is tight. Decomposition methods based on OA are likely to perform better if the NLP subproblems are relatively expensive to solve, while GBD can perform with some efficiency if the MINLP is tight and there are many discrete variables. ECP methods tend to perform well on mostly linear problems.

Logic-Based Optimization. Given difficulties in the modeling and solution of mixed integer problems, the following major approaches based on logic-based techniques have emerged: generalized disjunctive programming 107 [266], mixed-logic linear programming (MLLP) [267], and constraint programming (CP) [268]. The motivations for this logic-based modeling has been to facilitate the modeling, reduce the combinatorial search effort, and improve the handling of nonlinearities. In this section we mostly concentrate on generalized disjunctive programming and provide a brief reference to constraint programming. A general review of logic-based optimization can be found in [269, 309].

Generalized disjunctive programming in 107 [266] is an extension of disjunctive programming [270] that provides an alternative way of modeling MILP and MINLP problems. The general formulation 107 is as follows:

$$\begin{aligned}
&\text{Min } Z = \sum_{k\in K} c_k + f(x) \\
&s.t. \quad g(x) \leq 0 \\
&\bigvee_{j\in J_k} \begin{bmatrix} Y_{jk} \\ h_{jk}(x) \leq 0 \\ c_k = \gamma_{jk} \end{bmatrix}, k\in K \\
&\Omega(Y) = True \\
&x\in R^n, c\in R^m, Y\in\{true,\ false\}^m
\end{aligned} \tag{107}$$

where Y_{jk} are the Boolean variables that decide whether a term j in a disjunction $k \in K$ is true or false, and x are continuous variables. The objective function involves the term $f(x)$ for the continuous variables and the charges c_k that depend on the discrete choices in each disjunction $k \in K$. The constraints $g(x) \leq 0$ hold regardless of the discrete choice, and $h_{jk}(x) \leq 0$ are conditional constraints that hold when Y_{jk} is true in the j-th term of the k-th disjunction. The cost variables c_k correspond to the fixed charges, and are equal to γ_{jk} if the Boolean variable Y_{jk} is true. $\Omega(Y)$ are logical relations for the Boolean variables expressed as propositional logic.

Problem 107 can be reformulated as an MINLP problem by replacing the Boolean variables by binary variables y_{jk},

$$\begin{aligned}
&\text{Min } Z=\sum_{k\in K}\sum_{j\in J_k}\gamma_{jk}y_{jk}+f(x)\\
&\text{s.t.}\quad g(x)\le 0\\
&h_{jk}(x)\le M_{jk}\left(1-y_{jk}\right),\ j\in J_k,\ k\in K \qquad \text{(BM)}\\
&\sum_{j\in J_k}y_{jk}=1,\ k\in K\\
&Ay\le a\\
&0\le x\le x^U,\ y_{jk}\in\{0,1\},\ j\in J_k,\ k\in K
\end{aligned} \tag{108}$$

where the disjunctions are replaced by "Big-M" constraints which involve a parameter M_{jk} and binary variables y_{jk}. The propositional logic statements $\Omega(Y)=$ True are replaced by the linear constraints $Ay\le a$ [271] and [272]. Here we assume that x is a nonnegative variable with finite upper bound x^{U}. An important issue in model 108 is how to specify a valid value for the Big-M parameter M_{jk}. If the value is too small, then feasible points may be cut off. If M_{jk} is too large, then the continuous relaxation might be too loose and yield poor lower bounds. Therefore, finding the smallest valid value for M_{jk} is desired. For linear constraints, one can use the upper and lower bound of the variable x to calculate the maximum value of each constraint, which then can be used to calculate a valid value of M_{jk}. For nonlinear constraints one can in principle maximize each constraint over the feasible region, which is a nontrivial calculation.

Lee and Grossmann [273] have derived the *convex hull relaxation* of problem 107. The basic idea is as follows. Consider a disjunction $k\in K$ that has convex constraints

$$\begin{aligned}
&\bigvee_{j\in J_k}\begin{bmatrix} Y_{jk}\\ h_{jk}(x)\le 0\\ c=\gamma_{jk}\end{bmatrix}\\
&0\le x\le x^U, c\ge 0
\end{aligned} \tag{109}$$

where $h_{jk}(x)$ are assumed to be convex and bounded over x. The convex hull relaxation of disjunction 109 [242] is given as follows:

$$\begin{aligned}
&x=\sum_{j\in J_k}v^{jk},\quad c=\sum_{j\in J}\lambda_{jk}\gamma_{jk}\\
&0\le v^{jk}\le\lambda_{jk}x^{\mathrm{U}}_{jk},\ j\in J_k\\
&\sum_{j\in J_k}\lambda_{jk}=1,\ 0\le\lambda_{jk}\le 1,\ j\in J_k \qquad \text{(CH)}\\
&\lambda_{jk}h_{jk}\left(v^{jk}/\lambda_{jk}\right)\le 0,\ j\in J_k\\
&x,c,v^{jk}\ge 0,\ j\in J_k
\end{aligned} \tag{110}$$

where v^{jk} are disaggregated variables that are assigned to each term of the disjunction $k\in K$, and λ_{jk} are the weight factors that determine the feasibility of the disjunctive term. Note that when λ_{jk} is 1, then the j-th term in the k-th disjunction is enforced and the other terms are ignored. The constraints $\lambda_{jk}h_{jk}\left(v^{jk}/\lambda_{jk}\right)$ are convex if $h_{jk}(x)$ is convex [274, p. 160]. A formal proof can be found in [242]. Note that the convex hull 110 reduces to the result by Balas [275] if the constraints are linear. Based on the convex hull relaxation 110, Lee and Grossmann [273] proposed the following convex relaxation program of problem 107.

$$\begin{aligned}
&\text{Min } Z^L=\sum_{k\in K}\sum_{j\in J_k}\gamma_{jk}\lambda_{jk}+f(x)\\
&\text{s.t.}\quad g(x)\le 0\\
&x=\sum_{j\in J_k}v^{jk},\ \sum_{j\in J_k}\lambda_{jk}=1,\ k\in K \qquad \text{(CRP)}\\
&0\le x,\ v^{jk}\le x^{\mathrm{U}},\ 0\le\lambda_{jk}\le 1,\ j\in J_k,\ k\in K\\
&\lambda_{jk}h_{jk}\left(v^{jk}/\lambda_{jk}\right)\le 0,\ j\in J_k,\ k\in K\\
&A\lambda\le a\\
&0\le x,\ v^{jk}\le x^{\mathrm{U}},\ 0\le\lambda_{jk}\le 1,\ j\in J_k,\ k\in K
\end{aligned} \tag{111}$$

where x^{U} is a valid upper bound for x and v. Note that the number of constraints and variables increases in problem 111 compared with problem 107. Problem 111 has a unique optimal solution and it yields a valid lower bound to the optimal solution of problem 107 [273]. Grossmann and Lee [276] proved that problem 111 has the useful property that the lower bound is greater than or equal to the lower bound predicted from the relaxation of problem 108.

Further description of algorithms for disjunctive programming can be found in [277].

Constraint Programming. Constraint programming (CP) [268, 269] is a relatively new modeling and solution paradigm that was originally developed to solve feasibility problems, but it has been extended to solve optimization problems as well. Constraint programming is very expressive, as continuous, integer, and Boolean variables are permitted and, moreover, variables can be indexed by other variables. Constraints can be expressed in algebraic form (e.g., $h(x)\le 0$), as disjunctions (e.g., $[A_1x\le b_1]\vee A_2x\le b_2]$), or as conditional logic statements (e.g., If $g(x)\le 0$ then $r(x)\le 0$). In addition, the language can support special implicit functions such as the all-different $(x_1, x_2, \ldots x_n)$ constraint for

assigning different values to the integer variables $x_1, x_2, \ldots x_n$. The language consists of C++ procedures, although the recent trend has been to provide higher level languages such as OPL. Other commercial CP software packages include ILOG Solver [278], CHIP [279], and ECLiPSe [280].

10.6. Dynamic Optimization

Interest in dynamic simulation and optimization of chemical processes has increased significantly during the last two decades. Chemical processes are modeled dynamically using differential-algebraic equations (DAEs), consisting of differential equations that describe the dynamic behavior of the system, such as mass and energy balances, and algebraic equations that ensure physical and thermodynamic relations. Typical applications include control and scheduling of batch processes; startup, upset, shut-down, and transient analysis; safety studies; and the evaluation of control schemes. We state a general differential-algebraic optimization problem 112 as follows:

$$\begin{aligned}
&\text{Min} && \Phi\left(z\left(t_{\mathrm{f}}\right); y\left(t_{\mathrm{f}}\right); u\left(t_{\mathrm{f}}\right); t_{\mathrm{f}}; p\right) \\
&\text{s.t.} && F\left(dz/dt; z(t); u(t); t; p\right)=0,\ z(0)=z_0 \\
& && G_{\mathrm{s}}\left[z\left(t_{\mathrm{s}}\right); y\left(t_{\mathrm{s}}\right); u\left(t_{\mathrm{s}}\right); t_{\mathrm{s}}; p\right]=0 \\
& && z^{\mathrm{L}} \le z(t) \le x^{\mathrm{U}} \\
& && y^{\mathrm{L}} \le y(t) \le y^{\mathrm{U}} \\
& && u^{\mathrm{L}} \le u(t) \le y^{\mathrm{U}} \\
& && p^{\mathrm{L}} \le p \le p^{\mathrm{U}} \\
& && t_{\mathrm{f}}^{t} \le t_{\mathrm{f}} \le t_{\mathrm{f}}^{U}
\end{aligned} \tag{112}$$

where Φ is a scalar objective function at final time t_{f}, and F are DAE constraints, G_{s} additional point conditions at times t_{s}, $z(t)$ differential state profile vectors, $y(t)$ algebraic state profile vectors, $u(t)$ control state profile vectors, and p is a time-independent parameter vector.

We assume, without loss of generality, that the index of the DAE system is one, consistent initial conditions are available, and the objective function is in the above Mayer form. Otherwise, it is easy to reformulate problems to this form. Problem 112 can be solved either by the variational approach or by applying some level of discretization that converts the original continuous time problem into a discrete problem. Early solution strategies, known as indirect methods, were focused on solving the classical variational conditions for optimality. On the other hand, methods that discretize the original continuous time formulation can be divided into two categories, according to the level of discretization. Here we distinguish between the methods that discretize only the control profiles (partial discretization) and those that discretize the state and control profiles (full discretization). Basically, the partially discretized problem can be solved either by dynamic programming or by applying a nonlinear programming (NLP) strategy (direct-sequential). A basic characteristic of these methods is that a feasible solution of the DAE system, for given control values, is obtained by integration at every iteration of the NLP solver. The main advantage of these approaches is that, for the NLP solver, they generate smaller discrete problems than full discretization methods.

Methods that fully discretize the continuous time problem also apply NLP strategies to solve the discrete system and are known as direct-simultaneous methods. These methods can use different NLP and discretization techniques, but the basic characteristic is that they solve the DAE system only once, at the optimum. In addition, they have better stability properties than partial discretization methods, especially in the presence of unstable dynamic modes. On the other hand, the discretized optimization problem is larger and requires large-scale NLP solvers, such as SOCS, CONOPT, or IPOPT.

With this classification we take into account the degree of discretization used by the different methods. Below we briefly present the description of the variational methods, followed by methods that partially discretize the dynamic optimization problem, and finally we consider full discretization methods for problem 112.

Variational Methods. These methods are based on the solution of the first-order necessary conditions for optimality that are obtained from Pontryagin's maximum principle [281, 282]. If we consider a version of problem 112 without bounds, the optimality conditions are formulated as a set of DAEs:

$$\frac{\partial F(z,y,u,p,t)}{\partial z'}\lambda' = \frac{\partial H}{\partial z} = \frac{\partial F(z,y,u,p,t)}{\partial z}\lambda \tag{113a}$$

$$F(z,y,u,p,t)=0 \tag{113b}$$

$$G_f(z,y,u,p,t_f)=0 \tag{113c}$$

$$G_s(z,y,u,p,t_s)=0 \tag{113d}$$

$$\frac{\partial H}{\partial y}=\frac{\partial F(z,y,u,p,t)}{\partial y}\lambda=0 \tag{113e}$$

$$\frac{\partial H}{\partial u}=\frac{\partial F(z,y,u,p,t)}{\partial u}\lambda=0 \tag{113f}$$

$$\int_0^{t_f}\frac{\partial F(z,y,u,p,t)}{\partial p}\lambda\, dt=0 \tag{113g}$$

where the Hamiltonian H is a scalar function of the form $H(t)=F(z, y, u, p, y)^T\lambda(t)$ and $\lambda(t)$ is a vector of adjoint variables. Boundary and jump conditions for the adjoint variables are given by:

$$\begin{aligned}&\frac{\partial F}{\partial z'}\lambda(t_f)+\frac{\partial \Phi}{\partial z}+\frac{\partial G_f}{\partial z}v_f=0\\&\frac{\partial F}{\partial z'}\lambda\left(t_s^-\right)+\frac{\partial G_s}{\partial z}v_s=\frac{\partial F}{\partial z'}\lambda\left(t_s^+\right)\end{aligned} \tag{114}$$

where v_f, v_s are the multipliers associated with the final time and point constraints, respectively. The most expensive step lies in obtaining a solution to this boundary value problem. Normally, the state variables are given as initial conditions, and the adjoint variables as final conditions. This formulation leads to boundary value problems (BVPs) that can be solved by a number of standard methods including single shooting, invariant embedding, multiple shooting, or some discretization method such as collocation on finite elements or finite differences. Also the point conditions lead to an additional calculation loop to determine the multipliers v_f and v_s. On the other hand, when bound constraints are considered, the above conditions are augmented with additional multipliers and associated complementarity conditions. Solving the resulting system leads to a combinatorial problem that is prohibitively expensive except for small problems.

Partial Discretization. With partial discretization methods (also called sequential methods or control vector parametrization), only the control variables are discretized. Given the initial conditions and a given set of control parameters, the DAE system is solved with a differential algebraic equation solver at each iteration. This produces the value of the objective function, which is used by a nonlinear programming solver to find the optimal parameters in the control parametrization ν. The sequential method is reliable when the system contains only stable modes. If this is not the case, finding a feasible solution for a given set of control parameters can be very difficult. The time horizon is divided into time stages and at each stage the control variables are represented with a piecewise constant, a piecewise linear, or a polynomial approximation [283, 284]. A common practice is to represent the controls as a set of Lagrange interpolation polynomials.

For the NLP solver, gradients of the objective and constraint functions with respect to the control parameters can be calculated with the sensitivity equations of the DAE system, given by:

$$\begin{aligned}&\frac{\partial F}{\partial z'}^T s_k\prime+\frac{\partial F}{\partial z}^T s_k+\frac{\partial F}{\partial y}^T w_k+\frac{\partial F}{\partial q_k}^T=0,\\&\qquad s_k(0)=\frac{\partial z(0)}{\partial q_k}\quad k=1,\ldots N_q\end{aligned} \tag{115}$$

where $s_k(t)=\frac{\partial z(t)}{\partial q_k}$, $w_k(t)=\frac{\partial y(t)}{\partial q_k}$, and $q^T=[p^T,\nu^T]$. As can be inferred from Equation 115, the cost of obtaining these sensitivities is directly proportional to N_q, the number of decision variables in the NLP. Alternately, gradients can be obtained by integration of the adjoint Equations 113a, 113e, 113g [282, 285, 286] at a cost independent of the number of input variables and proportional to the number of constraints in the NLP.

Methods that are based on this approach cannot treat directly the bounds on state variables, because the state variables are not included in the nonlinear programming problem. Instead, most of the techniques for dealing with inequality path constraints rely on defining a measure of the constraint violation over the entire horizon, and then penalizing it in the objective function, or forcing it directly to zero through an end-point constraint [287]. Other techniques approximate the constraint satisfaction (constraint aggregation methods) by introducing an exact penalty function [286, 288] or a Kreisselmeier–Steinhauser function [288] into the problem.

Finally, initial value solvers that handle path constraints directly have been developed [284].

The main idea is to use an algorithm for constrained dynamic simulation, so that any admissible combination of the control parameters produces an initial value problem that is feasible with respect to the path constraints. The algorithm proceeds by detecting activation and deactivation of the constraints during the solution, and solving the resulting high-index DAE system and their related sensitivities.

Full Discretization. Full discretization methods explicitly discretize all the variables of the DAE system and generate a large-scale nonlinear programming problem that is usually solved with a successive quadratic programming (SQP) algorithm. These methods follow a simultaneous approach (or infeasible path approach); that is, the DAE system is not solved at each iteration; it is only solved at the optimum point. Because of the size of the problem, special decomposition strategies are used to solve the NLP efficiently. Despite this characteristic, the simultaneous approach has advantages for problems with state variable (or path) constraints and for systems where instabilities occur for a range of inputs. In addition, the simultaneous approach can avoid intermediate solutions that may not exist, are difficult to obtain, or require excessive computational effort. There are mainly two different approaches to discretize the state variables explicitly, multiple shooting [289, 290] and collocation on finite elements [181, 191, 291].

With *multiple shooting*, time is discretized into P stages and control variables are parametrized using a finite set of control parameters in each stage, as with partial discretization. The DAE system is solved on each stage, $i = 1,\ldots P$ and the values of the state variables $z(t_i)$ are chosen as additional unknowns. In this way a set of relaxed, decoupled initial value problems (IVP) is obtained:

$$\begin{aligned} &F\left(dz/dt; z\left(t\right); y\left(t\right); \nu_i; p\right) = 0, \\ &t \in \left[t_{i-1}, t_i\right], z\left(t_{i-1}\right) = z_i \\ &z_{i+1} - z\left(t_i; z_i; \nu_i; p\right) = 0, i = 1, \ldots P-1 \end{aligned} \tag{116}$$

Note that continuity among stages is treated through equality constraints, so that the final solution satisfies the DAE system. With this approach, inequality constraints for states and controls can be imposed directly at the grid points, but path constraints for the states may not be satisfied between grid points. This problem can be avoided by applying penalty techniques to enforce feasibility, like the ones used in the sequential methods.

The resulting NLP is solved using SQP-type methods, as described above. At each SQP iteration, the DAEs are integrated in each stage and objective and constraint gradients with respect to p, z_i, and ν_i are obtained using sensitivity equations, as in problem 115. Compared to sequential methods, the NLP contains many more variables, but efficient decompositions have been proposed [290] and many of these calculations can be performed in parallel.

In *collocation methods*, the continuous time problem is transformed into an NLP by approximating the profiles as a family of polynomials on finite elements. Various polynomial representations are used in the literature, including Lagrange interpolation polynomials for the differential and algebraic profiles [291]. In [191] a Hermite–Simpson collocation form is used, while Cuthrell and Biegler [292] and Tanartkit and Biegler [293] use a monomial basis for the differential profiles. All of these representations stem from implicit Runge–Kutta formulae, and the monomial representation is recommended because of smaller condition numbers and smaller rounding errors. Control and algebraic profiles, on the other hand, are approximated using Lagrange polynomials.

Discretizations of problem 112 using collocation formulations lead to the largest NLP problems, but these can be solved efficiently using large-scale NLP solvers such as IPOPT and by exploiting the structure of the collocation equations. Biegler et al. [181] provide a review of dynamic optimization methods using simultaneous methods. These methods offer a number of advantages for challenging dynamic optimization problems, which include:

- Control variables can be discretized at the same level of accuracy as the differential and algebraic state variables. The KKT conditions of the discretized problem can be shown to be consistent with the variational conditions of problem 112. Finite elements allow for discontinuities in control profiles.
- Collocation formulations allow problems with unstable modes to be handled in an

efficient and well-conditioned manner. The NLP formulation inherits stability properties of boundary value solvers. Moreover, an elementwise decomposition has been developed that pins down unstable modes in problem 112.

- Collocation formulations have been proposed with moving finite elements. This allows the placement of elements both for accurate breakpoint locations of control profiles as well as accurate DAE solutions.

Dynamic optimization by collocation methods has been used for a wide variety of process applications including batch process optimization, batch distillation, crystallization, dynamic data reconciliation and parameter estimation, nonlinear model predictive control, polymer grade transitions and process changeovers, and reactor design and synthesis. A review of this approach can be found in [294].

10.7. Development of Optimization Models

The most important aspect of a successful optimization study is the formulation of the optimization model. These models must reflect the real-world problem so that meaningful optimization results are obtained, and they also must satisfy the properties of the problem class. For instance, NLPs addressed by gradient-based methods require functions that are defined in the variable domain and have bounded and continuous first and second derivatives. In mixed integer problems, proper formulations are also needed to yield good lower bounds for efficient search. With increased understanding of optimization methods and the development of efficient and reliable optimization codes, optimization practitioners now focus on the *formulation* of optimization models that are realistic, well-posed, and inexpensive to solve. Finally, convergence properties of NLP, MILP, and MINLP solvers require accurate first (and often second) derivatives from the optimization model. If these contain numerical errors (say, through finite difference approximations) then performance of these solvers can deteriorate considerably. As a result of these characteristics, modeling platforms are essential for the formulation task. These are classified into two broad areas: optimization modeling platforms and simulation platforms with optimization.

Optimization modeling platforms provide general purpose interfaces for optimization algorithms and remove the need for the user to interface to the solver directly. These platforms allow the general formulation for all problem classes discussed above with direct interfaces to state of the art optimization codes. Three representative platforms are GAMS (General Algebraic Modeling Systems), AMPL (A Mathematical Programming Language), and AIMMS (Advanced Integrated Multidimensional Modeling Software). All three require problem-model input via a declarative modeling language and provide exact gradient and Hessian information through automatic differentiation strategies. Although possible, these platforms were not designed to handle externally added procedural models. As a result, these platforms are best applied on optimization models that can be developed entirely within their modeling framework. Nevertheless, these platforms are widely used for large-scale research and industrial applications. In addition, the MATLAB platform allows the flexible formulation of optimization models as well, although it currently has only limited capabilities for automatic differentiation and limited optimization solvers. More information on these and other modeling platforms can be found on the NEOS server www-neos.mcs.anl.gov

Simulation platforms with optimization are often dedicated, application-specific modeling tools to which optimization solvers have been interfaced. These lead to very useful optimization studies, but because they were not originally designed for optimization models, they need to be used with some caution. In particular, most of these platforms do not provide exact derivatives to the optimization solver; often they are approximated through finite difference. In addition, the models themselves are constructed and calculated through numerical procedures, instead of through an open declarative language. Examples of these include widely used process simulators such as Aspen/Plus, PRO/II, and Hysys. More recent platforms such as Aspen Custom Modeler and gPROMS include declarative models and exact derivatives.

For optimization tools linked to procedural models, reliable and efficient automatic differ-

entiation tools are available that link to models written, say, in FORTRAN and C, and calculate exact first (and often second) derivatives. Examples of these include ADIFOR, ADOL-C, GRESS, Odyssee, and PADRE. When used with care, these can be applied to existing procedural models and, when linked to modern NLP and MINLP algorithms, can lead to powerful optimization capabilities. More information on these and other automatic differentiation tools can be found on: http://www-unix.mcs.anl.gov/autodiff/AD_Tools/.

Finally, the availability of automatic differentiation and related sensitivity tools for differential equation models allows for considerable flexibility in the formulation of optimization models. In [295] a seven-level modeling hierarchy is proposed that matches optimization algorithms with models that range from completely open (fully declarative model) to fully closed (entirely procedural without sensitivities). At the lowest, fully procedural level, only derivative-free optimization methods are applied, while the highest, declarative level allows the application of an efficient large-scale solver that uses first and second derivatives. Depending on the modeling level, optimization solver performance can vary by several orders of magnitude.

11. Probability and Statistics [15, 296 – 303]

The models treated thus far have been deterministic, that is, if the parameters are known the outcome is determined. In many situations, all the factors cannot be controlled and the outcome may vary randomly about some average value. Then a range of outcomes has a certain probability of occurring, and statistical methods must be used. This is especially true in quality control of production processes and experimental measurements. This chapter presents standard statistical concepts, sampling theory and statistical decisions, and factorial design of experiments or analysis of variances. Multivariant linear and nonlinear regression is treated in Chapter 2.

11.1. Concepts

Suppose N values of a variable y, called $y_1, y_2, \ldots, y_N$, might represent N measurements of the same quantity. The *arithmetic mean* $E(y)$ is

$$E(y) = \frac{\sum_{i=1}^{N} y_i}{N}$$

The *median* is the middle value (or average of the two middle values) when the set of numbers is arranged in increasing (or decreasing) order. The *geometric mean* $\bar{y}_G$ is

$$\overline{y}_G = (y_1 y_2 \ldots y_N)^{1/N}$$

The *root-mean-square* or quadratic mean is

$$\text{Root-mean-square} = \sqrt{E(y^2)} = \sqrt{\sum_{i=1}^{N} y_i^2 / N}$$

The *range* of a set of numbers is the difference between the largest and the smallest members in the set. The *mean deviation* is the mean of the deviation from the mean.

$$\text{Mean-deviation} = \frac{\sum_{i=1}^{N} |y_i - E(y)|}{N}$$

The *variance* is

$$var(y) = \sigma^2 = \frac{\sum_{i=1}^{N} (y_i - E(y))^2}{N}$$

and the *standard deviation* σ is the square root of the variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y_i - E(y))^2}{N}}$$

If the set of numbers $\{y_i\}$ is a small sample from a larger set, then the *sample average*

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

is used in calculating the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

and the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}}$$

The value $n-1$ is used in the denominator because the deviations from the sample average must total zero:

$$\sum_{i=1}^{n}(y_i-\overline{y})=0$$

Thus, knowing $n-1$ values of $y_i-\bar{y}$ and the fact that there are n values automatically gives the n-th value. Thus, only $n-1$ degrees of freedom ν exist. This occurs because the unknown mean $E(y)$ is replaced by the sample mean y derived from the data.

If data are taken consecutively, running totals can be kept to permit calculation of the mean and variance without retaining all the data:

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}y_1^2-2\overline{y}\sum_{i=1}^{n}y_i+(\overline{y})^2$$

$$\overline{y}=\sum_{i=1}^{n}y_i/n$$

Thus,

$$n,\sum_{i=1}^{n}y_1^2,\ \text{and}\sum_{i=1}^{n}y_i$$

are retained, and the mean and variance are computed when needed.

Repeated observations that differ because of experimental error often vary about some central value in a roughly symmetrical distribution in which small deviations occur more frequently than large deviations. In plotting the number of times a discrete event occurs, a typical curve is obtained, which is shown in Figure 49. Then the probability p of an event (score) occurring can be thought of as the ratio of the number of times it was observed divided by the total number of events. A continuous representation of this probability density function is given by the *normal distribution*

$$p(y)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-[y-E(y)]^2/2\sigma^2}. \tag{117}$$

This is called a normal probability distribution function. It is important because many results are insensitive to deviations from a normal distribution. Also, the central limit theorem says that if an overall error is a linear combination of component errors, then the distribution of errors tends to be normal as the number of components increases, almost regardless of the distribution of the component errors (i.e., they need not be normally distributed). Naturally, several sources of error must be present and one error cannot predominate (unless it is normally distributed). The normal distribution function is calculated easily; of more value are integrals of the function, which are given in Table 12; the region of interest is illustrated in Figure 50.

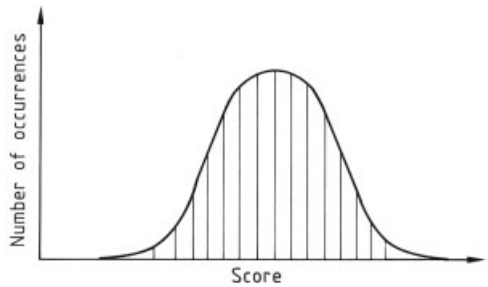

Figure 49. Frequency of occurrence of different scores

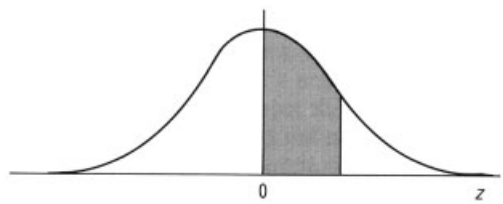

Figure 50. Area under normal curve

For a small sample, the variance can only be estimated with the sample variance s^2. Thus, the normal distribution cannot be used because σ is not known. In such cases *Student's t-distribution*, shown in Figure 51 [303, p. 70], is used:

$$p(y)=\frac{y_0}{\left(1+\frac{t^2}{n-1}\right)^{n/2}},\ t=\frac{\overline{y}-E(y)}{s/\sqrt{n}}$$

and y_0 is chosen such that the area under the curve is one. The number $\nu=n-1$ is the degrees of freedom, and as ν increases, Student's t-distribution approaches the normal distribution. The normal distribution is adequate (rather than the t-distribution) when $\nu>15$, except for the tails of the curve which require larger ν. Integrals of the t-distribution are given in Table 13, the region of interest is shown in Figure 52.

Table 12. Area under normal curve *

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-z^2/2} dz$$

z	$F(z)''$	z	$F(z)''$
0.0	0.0000	1.5	0.4332
0.1	0.0398	1.6	0.4452
0.2	0.0793	1.7	0.4554
0.3	0.1179	1.8	0.4641
0.4	0.1554	1.9	0.4713
0.5	0.1915	2.0	0.4772
0.6	0.2257	2.1	0.4821
0.7	0.2580	2.2	0.4861
0.8	0.2881	2.3	0.4893
0.9	0.3159	2.4	0.4918
1.0	0.3413	2.5	0.4938
1.1	0.3643	2.7	0.4965
1.2	0.3849	3.0	0.4987
1.3	0.4032	4.0	0.499968
1.4	0.4192	5.0	0.4999997

* Table gives the probability F that a random variable will fall in the shaded region of Figure 50. For a more complete table (in slightly different form), see [23, Table 26.1]. This table is obtained in Microsoft Excel with the function NORMDIST(z,0,1,1)-0.5.

Table 13. Percentage points of area under Students t-distribution *

ν	α=0.10	α=0.05	α=0.01	α=0.001
1	6.314	12.706	63.657	636.619
2	2.920	4.303	9.925	31.598
3	2.353	3.182	5.841	12.941
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.859
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
15	1.753	2.131	2.947	4.073
20	1.725	2.086	2.845	3.850
25	1.708	2.060	2.787	3.725
30	1.697	2.042	2.750	3.646
∞	1.645	1.960	2.576	3.291

* Table gives t values such that a random variable will fall in the shaded region of Figure 52 with probability α. For a one-sided test the confidence limits are obtained for $\alpha/2$. For a more complet table (in slightly different form), see [23, Table 26.10]. This table is obtained in Microsoft Excel with the function TINV(α,ν).

Other probability distribution functions are useful. Any distribution function must satisfy the following conditions:

$$0 \leq F(x) \leq 1$$

$$F(-\infty) = 0, F(+\infty) = 1$$

$$F(x) \leq F(y) \quad \text{when } x \leq y$$

The *probability density function* is

$$p(x) = \frac{dF(x)}{dx}$$

where

$$dF = p dx$$

is the probability of x being between x and $x + dx$. The probability density function satisfies

$$p(x) \geq 0$$

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

The *Bernoulli distribution* applies when the outcome can take only two values, such as heads or tails, or 0 or 1. The probability distribution function is

$$p(x = k) = p^k (1-p)^{1-k}, \; k = 0 \text{ or } 1$$

and the mean of a function $g(x)$ depending on x is

$$E[g(x)] = g(1)\, p + g(0)\,(1-p)$$

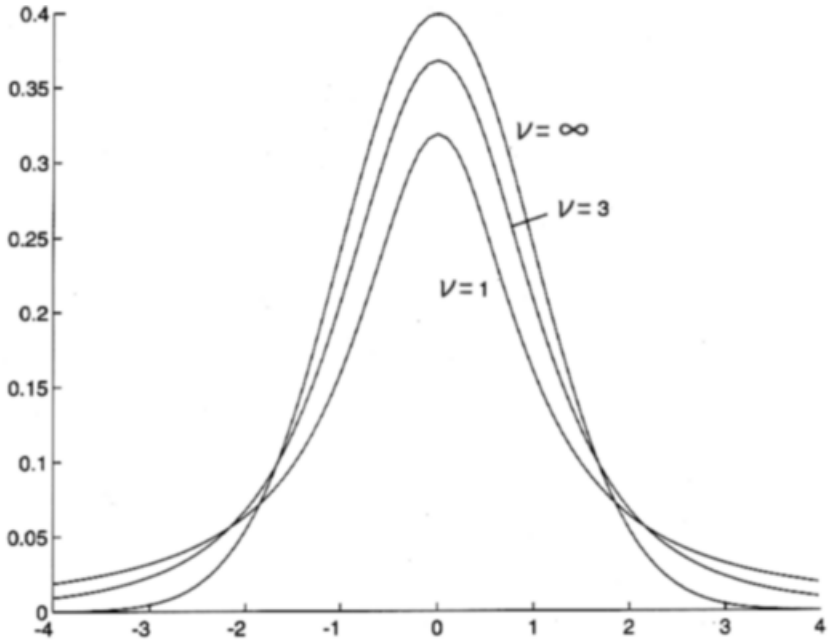

Figure 51. Student's *t*-distribution.
For explanation of ν see text

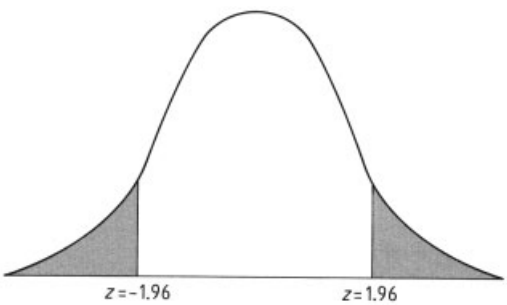

Figure 52. Percentage points of area under Student's *t*-distribution

The *binomial distribution function* applies when there are n trials of a Bernoulli event; it gives the probability of k occurrences of the event, which occurs with probability p on each trial

$$p\,(x=k) = \frac{n!}{k\,(n-k)\,!} p^k (1-p)^{n-k}$$

The mean and variance are

$$E\,(x) = np$$

$$var\,(x) = np\,(1-p)$$

The *hypergeometric distribution function* applies when there are N objects, of which M are of one kind and $N - M$ are of another kind. Then the objects are drawn one by one, without replacing the last draw. If the last draw had been replaced the distribution would be the binomial distribution. If x is the number of objects of type M drawn in a sample of size n, then the probability of $x = k$ is

$$p\,(x=k) =$$

$$\frac{M!(N-M)!n!(N-n)!}{k!(M-k)!(n-k)!(N-M-n+k)!N!}$$

The mean and variance are

$$E\,(x) = \frac{nM}{N}$$

$$var\,(x) = np\,(1-p)\,\frac{N-n}{N-1}$$

The *Poisson distribution* is

$$p\,(x=k) = \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}$$

with a parameter λ. The mean and variance are

$$E\,(x) = \lambda$$

$$var\,(x) = \lambda$$

The simplest continuous distribution is the uniform distribution. The probability density function is

$$p = \begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & x<a,\ x>b \end{cases}$$

and the probability distribution function is

$$F(x) = \begin{cases} 0 & x<a \\ \frac{x-a}{b-a} & a<x<b \\ 1 & b<x \end{cases}$$

The mean and variance are

$$E(x) = \frac{a+b}{2}$$

$$var(x) = \frac{(b-a)^2}{12}$$

The *normal distribution* is given by Equation 117 with variance σ^2.

The *log normal probability density* function is

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log x-\mu)^2}{2\sigma^2}\right]$$

and the mean and variance are [305, p. 89]

$$E(x) = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

$$var(x) = \exp\left(\sigma^2-1\right)\exp\left(2\mu+\sigma^2\right)$$

11.2. Sampling and Statistical Decisions

Two variables can be statistically dependent or independent. For example, the height and diameter of all distillation towers are statistically independent, because the distribution of diameters of all columns 10 m high is different from that of columns 30 m high. If y_B is the diameter and y_A the height, the distribution is written as

$$p(y_B|y_A = \text{constant}), \text{ or here}$$

$$p(y_B|y_A = 10) \neq p(y_B|y_A = 30)$$

A third variable y_C, could be the age of the operator on the third shift. This variable is probably unrelated to the diameter of the column, and for the distribution of ages is

$$p(y_C|y_A) = p(y_C)$$

Thus, variables y_A and y_C are distributed independently. The joint distribution for two variables is

$$p(y_A,y_B) = p(y_A)\,p(y_B|y_A)$$

if they are statistically dependent, and

$$p(y_A,y_B) = p(y_A)\,p(y_B)$$

if they are statistically independent. If a set of variables y_A, y_B, . . . is independent and identically distributed,

$$p(y_A,y_B,\ldots) = p(y_A)\,p(y_B)\ldots$$

Conditional probabilities are used in hazard analysis of chemical plants.

A measure of the linear dependence between variables is given by the *covariance*

$$Cov(y_A,y_B) = E\{[y_A-E(y_A)]\,[(y_B-E(y_B))]\}$$

$$= \frac{\sum_{i=1}^{N}[y_{Ai}-E(y_A)][y_{Bi}-E(y_B)]}{N}$$

The correlation coefficient ϱ is

$$\varrho(y_A,y_B) = \frac{Cov(y_A,y_B)}{\sigma_A\sigma_B}$$

If y_A and y_B are independent, then $Cov(y_A, y_B) = 0$. If y_A tends to increase when y_B decreases then $Cov(y_A, y_B) < 0$. The sample correlation coefficient is [15, p. 484]

$$r(y_A,y_B) = \frac{\sum_{i=1}^{n}(y_{Ai}-\overline{y}_A)(y_{Bi}-\overline{y}_B)}{(n-1)\,s_A s_B}$$

If measurements are for independent, identically distributed observations, the errors are independent and uncorrelated. Then $\overline{y}$ varies about $E(y)$ with variance σ^2/n, where n is the number of observations in $\overline{y}$. Thus if something is measured several times today and every day, and the measurements have the same distribution, the variance of the means decreases with the number of samples in each day's measurement n. Of course, other factors (weather, weekends) may cause the observations on different days to be distributed nonidentically.

Suppose Y, which is the sum or difference of two variables, is of interest:

$$Y = y_A \pm y_B$$

Then the mean value of Y is

$$E(Y) = E(y_A) \pm E(y_B)$$

and the variance of Y is

$$\sigma^2(Y) = \sigma^2(y_A) + \sigma^2(y_B)$$

More generally, consider the random variables $y_1, y_2, \ldots$ with means $E(y_1), E(y_2), \ldots$ and variances $\sigma^2(y_1), \sigma^2(y_2), \ldots$ and correlation coefficients ϱ_{ij}. The variable

$$Y = \alpha_1 y_1 + \alpha_2 y_2 + \ldots$$

has a mean

$$E(Y) = \alpha_1 E(y_1) + \alpha_2 E(y_2) + \ldots$$

and variance [303, p. 87]

$$\sigma^2(Y) = \sum_{i=1}^{n} \alpha_i^2 \sigma^2(y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \alpha_i \alpha_j \sigma(y_i)\,\sigma(y_j)\,\varrho_{ij}$$

or

$$\sigma^2(Y) = \sum_{i=1}^{n} \alpha_i^2 \sigma^2(y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \alpha_i \alpha_j Cov(y_i, y_j) \tag{118}$$

If the variables are uncorrelated and have the same variance, then

$$\sigma^2(Y) = \left(\sum_{i=1}^{n} \alpha_i^2\right)\sigma^2$$

This fact can be used to obtain more accurate cost estimates for the purchased cost of a chemical plant than is true for any one piece of equipment. Suppose the plant is composed of a number of heat exchangers, pumps, towers, etc., and that the cost estimate of each device is $\pm$ 40 % of its cost (the sample standard deviation is 20 % of its cost). In this case the α_i are the numbers of each type of unit. Under special conditions, such as equal numbers of all types of units and comparable cost, the standard deviation of the plant costs is

$$\sigma(Y) = \frac{\sigma}{\sqrt{n}}$$

and is then $\pm(40/\sqrt{n})$ %. Thus the standard deviation of the cost for the entire plant is the standard deviation of each piece of equipment divided by the square root of the number of units. Under less restrictive conditions the actual numbers change according to the above equations, but the principle is the same.

Suppose modifications are introduced into the manufacturing process. To determine if the modification causes a significant change, the mean of some property could be measured before and after the change; if these differ, does it mean the process modification caused it, or could the change have happened by chance? This is a statistical decision. A hypothesis H_0 is defined; if it is true, action *A* must be taken. The reverse hypothesis is H_1; if this is true, action *B* must be taken. A correct decision is made if action *A* is taken when H_0 is true or action *B* is taken when H_1 is true. Taking action *B* when H_0 is true is called a type I error, whereas taking action *A* when H_1 is true is called a type II error.

The following test of hypothesis or test of significance must be defined to determine if the hypothesis is true. The level of significance is the maximum probability that an error would be accepted in the decision (i.e., rejecting the hypothesis when it is actually true). Common levels of significance are 0.05 and 0.01, and the test of significance can be either one or two sided. If a sampled distribution is normal, then the probability that the z score

$$z = \frac{y - \overline{y}}{s_y}$$

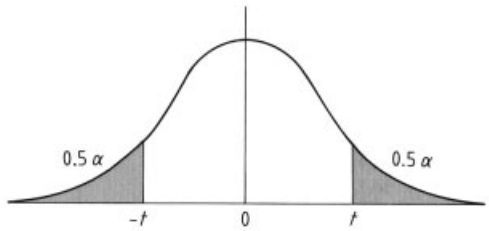

Figure 53. Two-sided statistical decision

is in the unshaded region is 0.95. Because a two-sided test is desired, $F = 0.95/2 = 0.475$. The value given in Table 12 for $F = 0.475$ is $z = 1.96$. If the test was one-sided, at the 5 % level of significance, 0.95 = 0.5 (for negative z) + F (for positive z). Thus, $F = 0.45$ or $z = 1.645$. In the two-sided test (see Fig. 53), if a single sample is chosen and $z < -1.96$ or $z > 1.96$, then this could happen with probability 0.05 if the hypothesis were true. This z would be significantly different from the expected value (based on the chosen level of significance) and the tendency would be to reject the hypothesis. If the value of z was between -1.96 and 1.96, the hypothesis would be accepted.

The same type of decisions can be made for other distributions. Consider Student's *t*-distri-

bution. At a 95 % level of confidence, with ν = 10 degrees of freedom, the t values are $\pm$ 2.228. Thus, the sample mean would be expected to be between

$$\overline{y} \pm t_c \frac{s}{\sqrt{n}}$$

with 95 % confidence. If the mean were outside this interval, the hypothesis would be rejected.

The *chi-square distribution* is useful for examining the variance or standard deviation. The statistic is defined as

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$= \frac{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\ldots+(y_n-\overline{y})^2}{\sigma^2}$$

and the chi-square distribution is

$$p(y) = y_0 \chi^{\nu-2} \mathrm{e}^{-\chi^2/2}$$

$\nu = n - 1$ is the number of degrees of freedom and y_0 is chosen so that the integral of p (y) over all y is 1. The probability of a deviation larger than χ^2 is given in Table 14; the area in question, in Figure 54. For example, for 10 degrees of freedom and a 95 % confidence level, the critical values of χ^2 are 0.025 and 0.975. Then

$$\frac{s\sqrt{n}}{\chi_{0.975}} < \sigma < \frac{s\sqrt{n}}{\chi_{0.025}}$$

or

$$\frac{s\sqrt{n}}{20.5} < \sigma < \frac{s\sqrt{n}}{3.25}$$

with 95 % confidence.

Tests are available to decide if two distributions that have the same variance have different means [15, p. 465]. Let one distribution be called x, with N_1 samples, and the other be called y, with N_2 samples. First, compute the standard error of the difference of the means:

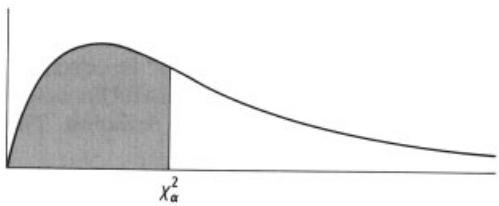

Figure 54. Percentage points of area under chi-squared distribution with ν degrees of freedom

$$s_D = \sqrt{\frac{\sum_{i=1}^{N_1}(x_i-\overline{x})^2+\sum_{i=1}^{N_2}(y_i-\overline{y})^2}{N_1+N_2-2}\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}$$

Next, compute the value of t

$$t = \frac{\overline{x}-\overline{y}}{s_D}$$

and evaluate the significance of t using Student's t-distribution for $N_1 + N_2 - 2$ degrees of freedom.

If the samples have different variances, the relevant statistic for the t-test is

$$t = \frac{\overline{x}-\overline{y}}{\sqrt{var(x)/N_1+var(y)/N_2}}$$

The number of degrees of freedom is now taken approximately as

$$\nu = \frac{\left(\frac{var(x)}{N_1}+\frac{var(y)}{N_2}\right)^2}{\frac{[var(x)/N_1]^2}{N_1-1}+\frac{[var(y)/N_2]^2}{N_2-1}}$$

There is also an F-test to decide if two distributions have significantly different variances. In this case, the ratio of variances is calculated:

$$F = \frac{var(x)}{var(y)}$$

where the variance of x is assumed to be larger. Then, a table of values is used to determine the significance of the ratio. The table [23, Table 26.9] is derived from the formula [15, p. 169]

$$Q(F|\nu_1,\nu_2) = I_{\nu_2/(\nu_2+\nu_1 F)}\left(\frac{\nu_2}{2},\frac{\nu_1}{2}\right)$$

where the right-hand side is an incomplete beta function. The F table is given by the Microsoft Excel function FINV(fraction, ν_x, ν_y), where fraction is the fractional value (≤ 1) representing the upper percentage and ν_x and ν_y are the degrees of freedom of the numerator and denominator, respectively.

Example. For two sample variances with 8 degrees of freedom each, what limits will bracket their ratio with a midarea probability of 90 %? FINV(0.95,8,8) = 3.44. The 0.95 is used to get both sides to toal 10 %. Then

$$P[1/3.44 \leq \mathrm{var}(x)/\mathrm{var}(y) \leq 3.44] = 0.90.$$

Table 14. Percentage points of area under chi-square distribution with ν degrees of freedom *

ν	α=0.995	α=0.99	α=0.975	α=0.95	α=0.5	α=0.05	α=0.025	α=0.01	α=0.005
1	7.88	6.63	5.02	3.84	0.455	0.0039	0.0010	0.0002	0.00004
2	10.6	9.21	7.38	5.99	1.39	0.103	0.0506	0.0201	0.0100
3	12.8	11.3	9.35	7.81	2.37	0.352	0.216	0.115	0.072
4	14.9	13.3	11.1	9.49	3.36	0.711	0.484	0.297	0.207
5	16.7	15.1	12.8	11.1	4.35	1.15	0.831	0.554	0.412
6	18.5	16.8	14.4	12.6	5.35	1.64	1.24	0.872	0.676
7	20.3	18.5	16.0	14.1	6.35	2.17	1.69	1.24	0.989
8	22.0	20.1	17.5	15.5	7.34	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	8.34	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	9.34	3.94	3.25	2.56	2.16
12	28.3	26.2	23.3	21.0	11.3	5.23	4.40	3.57	3.07
15	32.8	30.6	27.5	25.0	14.3	7.26	6.26	5.23	4.60
17	35.7	33.4	30.2	27.6	16.3	8.67	7.56	6.41	5.70
20	40.0	37.6	34.2	31.4	19.3	10.9	9.59	8.26	7.43
25	46.9	44.3	40.6	37.7	24.3	14.6	13.1	11.5	10.5
30	53.7	50.9	47.0	43.8	29.3	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	39.3	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	49.3	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	59.3	43.2	40.5	37.5	35.5
70	104.2	100.4	95.0	90.5	69.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	79.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	89.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	99.3	77.9	74.2	70.1	67.3

* Table value is χ^2_α; $\chi^2 < \chi^2_\alpha$ with probability α. For a more complete table (in slightly different form), see [23, Table 26.8]. The Microsoft Excel function CHIINV(1-α,ν) gives the table value.

11.3. Error Analysis in Experiments

Suppose a measurement of several quantities is made and a formula or mathematical model is used to deduce some property of interest. For example, to measure the thermal conductivity of a solid k, the heat flux q, the thickness of the sample d, and the temperature difference across the sample ΔT must be measured. Each measurement has some error. The heat flux q may be the rate of electrical heat input $\dot{Q}$ divided by the area A, and both quantities are measured to some tolerance. The thickness of the sample is measured with some accuracy, and the temperatures are probably measured with a thermocouple, to some accuracy. These measurements are combined, however, to obtain the thermal conductivity, and the error in the thermal conductivity must be determined. The formula is

$$k = \frac{d}{A\Delta T}\,\dot{Q}$$

If each measured quantity has some variance, what is the variance in the thermal conductivity?

Suppose a model for Y depends on various measurable quantities, $y_1, y_2, \ldots$ Suppose several measurements are made of $y_1, y_2, \ldots$ under seemingly identical conditions and several different values are obtained, with means $E\,(y_1)$, $E\,(y_2)$, . . . and variances $\sigma_1^2, \sigma_2^2, \ldots$ Next suppose the errors are small and independent of one another. Then a change in Y is related to changes in y_i by

$$\mathrm{d}Y = \frac{\partial Y}{\partial y_1}\mathrm{d}y_1 + \frac{\partial Y}{\partial y_2}\mathrm{d}y_2 + \ldots$$

If the changes are indeed small, the partial derivatives are constant among all the samples. Then the expected value of the change is

$$E\,(\mathrm{d}Y) = \sum_{i=1}^{N}\left(\frac{\partial Y}{\partial y_i}\right)E\,(\mathrm{d}y_i)$$

Naturally $E\,(\mathrm{d}y_i) = 0$ by definition so that $E\,(\mathrm{d}Y) = 0$, too. However, since the errors are independent of each other and the partial derivatives are assumed constant because the errors are small, the variances are given by Equation 119 [296, p. 550]

$$\sigma^2\,(\mathrm{d}Y) = \sum_{i=1}^{N}\left(\frac{\partial Y}{\partial y_i}\right)^2\sigma_i^2 \tag{119}$$

Thus, the variance of the desired quantity Y can be found. This gives an independent estimate of the errors in measuring the quantity Y from the errors in measuring each variable it depends upon.

11.4. Factorial Design of Experiments and Analysis of Variance

Statistically designed experiments consider, of course, the effect of primary variables, but they also consider the effect of extraneous variables, the interactions among variables, and a measure of the random error. Primary variables are those whose effect must be determined. These variables can be quantitative or qualitative. Quantitative variables are ones that may be fit to a model to determine the model parameters. Curve fitting of this type is discused in Chapter 2. Qualitative variables are ones whose effect needs to be known; no attempt is made to quantify that effect other than to assign possible errors or magnitudes. Qualitative variables can be further subdivided into type I variables, whose effect is determined directly, and type II variables, which contribute to performance variability, and whose effect is averaged out. For example, in studying the effect of several catalysts on yield in a chemical reactor, each different type of catalyst would be a type I variable, because its effect should be known. However, each time the catalyst is prepared, the results are slightly different, because of random variations; thus, several batches may exist of what purports to be the same catalyst. The variability between batches is a type II variable. Because the ultimate use will require using different batches, the overall effect including that variation should be known, because knowing the results from one batch of one catalyst precisely might not be representative of the results obtained from all batches of the same catalyst. A randomized block design, incomplete block design, or Latin square design, for example, all keep the effect of experimental error in the blocked variables from influencing the effect of the primary variables. Other uncontrolled variables are accounted for by introducing randomization in parts of the experimental design. To study all variables and their interaction requires a factorial design, involving all possible combinations of each variable, or a fractional factorial design, involving only a selected set. Statistical techniques are then used to determine the important variables, the important interactions and the error in estimating these effects. The discussion here is a brief overview of [303].

If only two methods exist for preparing some product, to see which treatment is best, the sampling analysis discussed in Section 11.2 can be used to deduce if the means of the two treatments differ significantly. With more treatments, the analysis is more detailed. Suppose the experimental results are arranged as shown in Table 15, i.e., several measurements for each treatment. The objective is to see if the treatments differ significantly from each other, that is, whether their means are different. The samples are assumed to have the same variance. The hypothesis is that the treatments are all the same, and the null hypothesis is that they are different. Deducing the statistical validity of the hypothesis is done by an analysis of variance.

Table 15. Estimating the effect of four treatments

Treatment	1	2	3	4
	–	–	–	–
	–	–	–	–
	–	–	–	–
		–	–	–
			–	–
				–
Treatment average, $\overline{y}_t$	–	–	–	–
Grand average, $\overline{y}$		–		

The data for $k = 4$ treatments are arranged in Table 15. Each treatment has n_t experiments, and the outcome of the i-th experiment with treatment t is called y_{ti}. The treatment average is

$$\overline{y}_t = \frac{\sum_{i=1}^{n_t} y_{ti}}{n_t}$$

and the grand average is

$$\overline{y} = \frac{\sum_{t=1}^{k} n_t \overline{y}_t}{N}, \; N = \sum_{t=1}^{k} n_t$$

Next, the sum of squares of deviations is computed from the average within the t-th treatment

$$S_t = \sum_{i=1}^{n_t} (y_{ti} - \overline{y}_t)^2$$

Since each treatment has n_t experiments, the number of degrees of freedom is $n_t - 1$. Then the sample variances are

$$s_t^2 = \frac{S_t}{n_t - 1}$$

The within-treatment sum of squares is

$$S_R = \sum_{t=1}^{k} S_t$$

and the within-treatment sample variance is

$$s_R^2 = \frac{S_R}{N-k}$$

Now, if no difference exists between treatments, a second estimate of σ^2 could be obtained by calculating the variation of the treatment averages about the grand average. Thus, the between-treatment mean square is computed:

$$s_T^2 = \frac{S_T}{k-1},\ S_T = \sum_{t=1}^{k} n_t(\overline{y}_t - \overline{y})^2$$

Basically the test for whether the hypothesis is true or not hinges on a comparison between the within-treatment estimate s_R^2 (with $\nu_R = N - k$ degrees of freedom) and the between-treatment estimate s_T^2 (with $\nu_T = k - 1$ degrees of freedom). The test is made based on the F distribution for ν_R and ν_T degrees of freedom [23, Table 26.9], [303, p. 636].

Table 16. Block design with four treatments and five blocks

Treatment	1	2	3	4	Block average
Block 1	–	–	–	–	–
Block 2	–	–	–	–	–
Block 3	–	–	–	–	–
Block 4	–	–	–	–	–
Block 5	–	–	–	–	–
Treatment average	–	–	–	–	grand average

Next consider the case in which *randomized blocking* is used to eliminate the effect of some variable whose effect is of no interest, such as the batch-to-batch variation of the catalysts in the chemical reactor example. With k treatments and n experiments in each treatment, the results from $n\ k$ experiments can be arranged as shown in Table 16; within each block, various treatments are applied in a random order. The block average, the treatment average, and the grand average are computed as before. The following quantities are also computed for the analysis of variance table:

Name	Formula	Degrees of freedom
Average	$S_A = nk\overline{y}^2$	1
Blocks	$S_B = k\sum_{i=1}^{n}(\overline{y}_t - \overline{y})^2$	n-1
Treatments	$S_T = n\sum_{t=1}^{k}(\overline{y}_t - \overline{y})^2$	k-1
Residuals	$S_R = \sum_{t=1}^{k}\sum_{i=1}^{n}(y_{ti} - \overline{y}_i - \overline{y}_t + \overline{y})^2$	(n-1)(k-1)
Total	$S = \sum_{t=1}^{k}\sum_{i=1}^{n} y_{ti}^2$	$N = n\ k$

The key test is again a statistical one, based on the value of

s_T^2/s_R^2, where $s_T^2 = \frac{S_T}{k-1}$

and $s_R^2 = \frac{S_R}{(n-1)(k-1)}$

and the F distribution for ν_R and ν_T degrees of freedom [303, p. 636]. The assumption behind the analysis is that the variations are linear [303, p. 218]. Ways to test this assumption as well as transformations to make if it is not true are provided in [303], where an example is given of how the observations are broken down into a grand average, a block deviation, a treatment deviation, and a residual. For two-way factorial design, in which the second variable is a real one rather than one you would like to block out, see [303, p. 228].

Table 17. Two-level factorial design with three variables

Run	Variable 1	Variable 2	Variable 3
1	–	–	–
2	+	–	–
3	–	+	–
4	+	+	–
5	–	–	+
6	+	–	+
7	–	+	+
8	+	+	+

To measure the effects of variables on a single outcome, a *factorial design* is appropriate. In a two-level factorial design, each variable is considered at two levels only, a high and low value, often designated as a + and a −. The two-level factorial design is useful for indicating trends and showing interactions; it is also the basis for a fractional factorial design. As an example, consider a 2^3 factorial design, with 3 variables and 2 levels for each. The experiments are indicated in Table 17. The main effects are calculated by determining the difference between results from

all high values of a variable and all low values of a variable; the result is divided by the number of experiments at each level. For example, for the first variable, calculate

$$\text{Effect of variable 1} = [(y_2+y_4+y_6+y_8) - [(y_1+y_3+y_5+y_7)]]/4$$

Note that all observations are being used to supply information on each of the main effects and each effect is determined with the precision of a fourfold replicated difference. The advantage of a one-at-a-time experiment is the gain in precision if the variables are additive and the measure of nonadditivity if it occurs [303, p. 313].

Interaction effects between variables 1 and 2 are obtained by comparing the difference between the results obtained with the high and low value of 1 at the low value of 2 with the difference between the results obtained with the high and low value 1 at the high value of 2. The 12-interaction is

$$12\,\text{interaction} = [(y_4-y_3+y_8-y_7) - [(y_2-y_1+y_6-y_5)]]/2$$

The key step is to determine the errors associated with the effect of each variable and each interaction so that the significance can be determined. Thus, standard errors need to be assigned. This can be done by repeating the experiments, but it can also be done by using higher order interactions (such as 123 interactions in a 2^4 factorial design). These are assumed negligible in their effect on the mean but can be used to estimate the standard error [303, pp. 319 – 328]. Then calculated effects that are large compared to the standard error are considered important, whereas those that are small compared to the standard error are considered due to random variations and are unimportant.

In a fractional factorial design, only part of the possible experiments is performed. With k variables, a factorial design requires 2^k experiments. When k is large, the number of experiments can be large; for $k = 5$, $2^5 = 32$. For k this large, Box et al. [296, p. 235] do a fractional factorial design. In the fractional factorial design with $k = 5$, only 8 experiments are chosen. Cropley [298] gives an example of how to combine heuristics and statistical arguments in application to kinetics mechanisms in chemical engineering.

12. Multivariable Calculus Applied to Thermodynamics

Many of the functional relationships required in thermodynamics are direct applications of the rules of multivariable calculus. In this short chapter, those rules are reviewed in the context of the needs of thermodynamics. These ideas were expounded in one of the classic books on chemical engineering thermodynamics [299].

12.1. State Functions

State functions depend only on the state of the system, not on its past history or how one got there. If z is a function of two variables x and y, then $z\,(x, y)$ is a state function, because z is known once x and y are specified. The differential of z is

$$\mathrm{d}z = M\mathrm{d}x + N\mathrm{d}y$$

The line integral

$$\int_C (M\mathrm{d}x + N\mathrm{d}y)$$

is independent of the path in $x - y$ space if and only if

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \tag{120}$$

Because the total differential can be written as

$$\mathrm{d}z = \left(\frac{\partial z}{\partial x}\right)_y \mathrm{d}x + \left(\frac{\partial z}{\partial y}\right)_x \mathrm{d}y \tag{121}$$

for path independence

$$\frac{\partial}{\partial y}\left(\frac{\partial z}{\partial x}\right)_y = \frac{\partial}{\partial x}\left(\frac{\partial z}{\partial y}\right)_x$$

or

$$\frac{\partial^2 z}{\partial y \partial x} = \frac{\partial^2 z}{\partial x \partial y} \tag{122}$$

is needed.

Various relationships can be derived from Equation 121. If z is constant,

$$\left[0 = \left(\frac{\partial z}{\partial x}\right)_y \mathrm{d}x + \left(\frac{\partial z}{\partial y}\right)_x \mathrm{d}y\right]_z$$

Rearrangement gives

$$\left(\frac{\partial z}{\partial x}\right)_y = -\left(\frac{\partial y}{\partial x}\right)_z \left(\frac{\partial z}{\partial y}\right)_x = -\frac{(\partial y/\partial x)_z}{(\partial y/\partial z)_x} \tag{123}$$

Alternatively, if Equation 121 is divided by dy while some other variable w is held constant,

$$\left(\frac{\partial z}{\partial y}\right)_w = \left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial x}{\partial y}\right)_w + \left(\frac{\partial z}{\partial y}\right)_x \tag{124}$$

Dividing both the numerator and the denominator of a partial derivative by dw while holding a variable y constant yields

$$\left(\frac{\partial z}{\partial x}\right)_y = \frac{(\partial z/\partial w)_y}{(\partial x/\partial w)_y} = \left(\frac{\partial z}{\partial w}\right)_y \left(\frac{\partial w}{\partial x}\right)_y \tag{125}$$

In thermodynamics the state functions include the internal energy U, the enthalpy H, and the Helmholtz and Gibbs free energies A and G, respectively, which are defined as follows:

$$H = U + pV$$

$$A = U - TS$$

$$G = H - TS = U + pV - TS = A + pV$$

where S is the entropy, T the absolute temperature, p the pressure, and V the volume. These are also state functions, in that the entropy is specified once two variables (e.g., T and p) are specified. Likewise V is specified once T and p are specified, and so forth.

12.2. Applications to Thermodynamics

All of the following applications are for closed systems with constant mass. If a process is reversible and only $p - V$ work is done, one form of the first law states that changes in the internal energy are given by the following expression

$$dU = T\mathrm{d}S - p\mathrm{d}V \tag{126}$$

If the internal energy is considered a function of S and V, then

$$\mathrm{d}U = \left(\frac{\partial U}{\partial S}\right)_V \mathrm{d}S + \left(\frac{\partial U}{\partial V}\right)_S \mathrm{d}V$$

This is the equivalent of Equation 121 and

$$T = \left(\frac{\partial U}{\partial S}\right)_V, \; p = -\left(\frac{\partial U}{\partial V}\right)_S$$

Because the internal energy is a state function, Equation 122 is required:

$$\frac{\partial^2 U}{\partial V \partial S} = \frac{\partial^2 U}{\partial S \partial V}$$

which here is

$$\left(\frac{\partial T}{\partial V}\right)_S = -\left(\frac{\partial p}{\partial S}\right)_V \tag{127}$$

This is one of the Maxwell relations and is merely an expression of Equation 122.

The differentials of the other energies are

$$dH = T\,\mathrm{d}S + V\,\mathrm{d}p \tag{128}$$

$$dA = -S\mathrm{d}T - p\mathrm{d}V \tag{129}$$

$$dG = -S\,\mathrm{d}T + V\,\mathrm{d}p \tag{130}$$

From these differentials, other Maxwell relations can be derived in a similar fashion by applying Equation 122.

$$\left(\frac{\partial T}{\partial p}\right)_S = \left(\frac{\partial V}{\partial S}\right)_p \tag{131}$$

$$\left(\frac{\partial S}{\partial V}\right)_T = \left(\frac{\partial p}{\partial T}\right)_V \tag{132}$$

$$\left(\frac{\partial S}{\partial p}\right)_T = -\left(\frac{\partial V}{\partial T}\right)_p \tag{133}$$

The heat capacity at constant pressure is defined as

$$C_p = \left(\frac{\partial H}{\partial T}\right)_p$$

If entropy and enthalpy are taken as functions of T and p, the total differentials are

$$\mathrm{d}S = \left(\frac{\partial S}{\partial T}\right)_p \mathrm{d}T + \left(\frac{\partial S}{\partial p}\right)_T \mathrm{d}p \tag{134}$$

$$\begin{aligned}\mathrm{d}H &= \left(\frac{\partial H}{\partial T}\right)_p \mathrm{d}T + \left(\frac{\partial H}{\partial p}\right)_T \mathrm{d}p \\ &= C_p \mathrm{d}T + \left(\frac{\partial H}{\partial p}\right)_T \mathrm{d}p\end{aligned} \tag{135}$$

If the pressure is constant,

$$\mathrm{d}S = \left(\frac{\partial S}{\partial T}\right)_p \mathrm{d}T \text{ and } \mathrm{d}H = C_p \mathrm{d}T$$

When enthalpy is considered a function of S and p, the total differential is

$$\mathrm{d}H = T\,\mathrm{d}S + V\,\mathrm{d}p$$

When the pressure is constant, this is

$$\mathrm{d}H = T\mathrm{d}S$$

Thus, at constant pressure

$$\mathrm{d}H = C_p \mathrm{d}T = T\mathrm{d}S = T\left(\frac{\partial S}{\partial T}\right)_p \mathrm{d}T$$

which gives

$$\left(\frac{\partial S}{\partial T}\right)_p = \frac{C_p}{T}$$

When p is not constant, using the last Maxwell relation gives

$$\mathrm{d}S = \frac{C_p}{T}\mathrm{d}T - \left(\frac{\partial V}{\partial T}\right)_p \mathrm{d}p \tag{136}$$

Then the total differential for H is

$$\mathrm{d}H = T\mathrm{d}S + V\mathrm{d}p = C_p\mathrm{d}T - T\left(\frac{\partial V}{\partial T}\right)_p \mathrm{d}p + V\mathrm{d}p$$

Rearranging this, when H (T, p), yields

$$\mathrm{d}H = C_p\mathrm{d}T + \left[V - T\left(\frac{\partial V}{\partial T}\right)_p\right]\mathrm{d}p \tag{137}$$

This equation can be used to evaluate enthalpy differences by using information on the equation of state and the heat capacity:

$$\begin{aligned} H(T_2,p_2) - H(T_1,p_1) &= \int_{T_1}^{T_2} C_p(T,p_1)\,\mathrm{d}T \\ &+ \int_{p_1}^{p_2}\left[V - T\left(\frac{\partial V}{\partial T}\right)_p\right]|_{T_2,p}\mathrm{d}p \end{aligned} \tag{138}$$

The same manipulations can be done for internal energy:

$$\left(\frac{\partial S}{\partial T}\right)_V = \frac{C_v}{T} \tag{139}$$

$$\mathrm{d}S = -\left[\frac{(\partial V/\partial T)_p}{(\partial V/\partial p)_T}\right]\mathrm{d}V + \frac{C_v}{T}\mathrm{d}T \tag{140}$$

$$\mathrm{d}U = C_v\mathrm{d}T - \left[p + T\frac{(\partial V/\partial T)_p}{(\partial V/\partial p)_T}\right]\mathrm{d}V$$

12.3. Partial Derivatives of All Thermodynamic Functions

The various partial derivatives of the thermodynamic functions can be classified into six groups. In the general formulas below, the variables U, H, A, G, or S are denoted by Greek letters, whereas the variables V, T, or p are denoted by Latin letters.

Type 1 (3 possibilities plus reciprocals).

General: $\left(\frac{\partial a}{\partial b}\right)_c$, Specific: $\left(\frac{\partial p}{\partial T}\right)_V$

Equation 123 yields

$$\left(\frac{\partial p}{\partial T}\right)_V = -\left(\frac{\partial V}{\partial T}\right)_p \left(\frac{\partial p}{\partial V}\right)_T \tag{141}$$

$$= -\frac{(\partial V/\partial T)_p}{(\partial V/\partial p)_T} \tag{142}$$

This relates all three partial derivatives of this type.

Type 2 (30 possibilities plus reciprocals).

General: $\left(\frac{\partial \alpha}{\partial b}\right)_c$, Specific: $\left(\frac{\partial G}{\partial T}\right)_V$

Using Equation 130 gives

$$\left(\frac{\partial G}{\partial T}\right)_V = -S + V\left(\frac{\partial p}{\partial T}\right)_V$$

Using the other equations for U, H, A, or S gives the other possibilities.

Type 3 (15 possibilities plus reciprocals).

General: $\left(\frac{\partial a}{\partial b}\right)_\alpha$, Specific: $\left(\frac{\partial V}{\partial T}\right)_S$

First the derivative is expanded by using Equation 123, which is called expansion without introducing a new variable:

$$\left(\frac{\partial V}{\partial T}\right)_S = -\left(\frac{\partial S}{\partial T}\right)_V \left(\frac{\partial V}{\partial S}\right)_T = -\frac{(\partial S/\partial T)_V}{(\partial S/\partial V)_T}$$

Then the numerator and denominator are evaluated as type 2 derivatives, or by using Equations (99) and (100):

$$\left(\frac{\partial V}{\partial T}\right)_S = -\frac{C_v/T}{-(\partial V/\partial T)_p(\partial p/\partial V)_T} \tag{143}$$

$$= \frac{C_v}{T}\frac{\left(\frac{\partial V}{\partial p}\right)_T}{\left(\frac{\partial V}{\partial T}\right)_p} \tag{144}$$

These derivatives are important for reversible, adiabatic processes (e.g., in an ideal turbine or compressor) because the entropy is constant. Similar derivatives can be obtained for isenthalpic processes, such as a pressure reduction at a valve. In that case, the Joule – Thomson coefficient is obtained for constant H:

$$\left(\frac{\partial T}{\partial p}\right)_H = \frac{1}{C_p}\left[-V + T\left(\frac{\partial V}{\partial T}\right)_p\right]$$

Type 4 (30 possibilities plus reciprocals).

General: $\left(\frac{\partial\alpha}{\partial\beta}\right)_c$, Specific: $\left(\frac{\partial G}{\partial A}\right)_p$

Now, expand through the introduction of a new variable using Equation 125:

$$\left(\frac{\partial G}{\partial A}\right)_p = \left(\frac{\partial G}{\partial T}\right)_p \left(\frac{\partial T}{\partial A}\right)_p = \frac{(\partial G/\partial T)_p}{(\partial A/\partial T)_p}$$

This operation has created two type 2 derivatives. Substitution yields

$$\left(\frac{\partial G}{\partial A}\right)_p = \frac{S}{S+p(\partial V/\partial T)_p}$$

Type 5 (60 possibilities plus reciprocals).

General: $\left(\frac{\partial\alpha}{\partial b}\right)_\beta$, Specific: $\left(\frac{\partial G}{\partial p}\right)_A$

Starting from Equation 130 for dG gives

$$\left(\frac{\partial G}{\partial p}\right)_A = -S\left(\frac{\partial T}{\partial p}\right)_A + V$$

The derivative is a type 3 derivative and can be evaluated by using Equation 123.

$$\left(\frac{\partial G}{\partial p}\right)_A = S\frac{(\partial A/\partial p)_T}{(\partial A/\partial T)_p} + V$$

The two type 2 derivatives are then evaluated:

$$\left(\frac{\partial G}{\partial p}\right)_A = \frac{S_p(\partial V/\partial p)_T}{S+p(\partial V/\partial T)_p} + V$$

These derivatives are also of interest for free expansions or isentropic changes.

Type 6 (30 possibilities plus reciprocals).

General: $\left(\frac{\partial\alpha}{\partial\beta}\right)_\gamma$, Specific: $\left(\frac{\partial G}{\partial A}\right)_H$

Equation 125 is used to obtain two type 5 derivatives.

$$\left(\frac{\partial G}{\partial A}\right)_H = \frac{(\partial G/\partial T)_H}{(\partial A/\partial T)_H}$$

These can then be evaluated by using the procedures for type 5 derivatives.

The difference in molar heat capacities ($C_p - C_v$) can be derived in similar fashion. Using Equation 139 for C_v yields

$$C_v = T\left(\frac{\partial S}{\partial T}\right)_V$$

To evaluate the derivative, Equation 124 is used to express dS in terms of p and T:

$$\left(\frac{\partial S}{\partial T}\right)_V = -\left(\frac{\partial V}{\partial T}\right)_p \left(\frac{\partial p}{\partial T}\right)_V + \frac{C_p}{T}$$

Substitution for $(\partial p/\partial T)_V$ and rearrangement give

$$C_p - C_v = T\left(\frac{\partial V}{\partial T}\right)_p \left(\frac{\partial p}{\partial T}\right)_V = -T\left(\frac{\partial V}{\partial T}\right)_p^2 \left(\frac{\partial p}{\partial V}\right)_T \tag{145}$$

Use of this equation permits the rearrangement of Equation 143 into

$$\left(\frac{\partial V}{\partial T}\right)_S = \frac{(\partial V/\partial T)_p^2 + \frac{C_p}{T}(\partial V/\partial p)_T}{(\partial V/\partial T)_p}$$

The ratio of heat capacities is

$$\frac{C_p}{C_v} = \frac{T(\partial S/\partial T)_p}{T(\partial S/\partial T)_V}$$

Expansion by using Equation 123 gives

$$\frac{C_p}{C_v} = \frac{-(\partial p/\partial T)_S(\partial S/\partial p)_T}{-(\partial V/\partial T)_S(\partial S/\partial V)_T}$$

and the ratios are then

$$\frac{C_p}{C_v} = \left(\frac{\partial p}{\partial V}\right)_S \left(\frac{\partial V}{\partial p}\right)_T$$

Using Equation 123 gives

$$\frac{C_p}{C_v} = -\left(\frac{\partial p}{\partial V}\right)_S \left(\frac{\partial T}{\partial p}\right)_V \left(\frac{\partial V}{\partial T}\right)_p$$

Entropy is a variable in at least one of the partial derivatives.

13. References

Specific References

1. D. E. Seborg, T. F. Edgar, D. A. Mellichamp: *Process Dynamics and Control*, 2nd ed., John Wiley & Sons, New York 2004.
2. G. Forsyth, C. B. Moler: *Computer Solution of Lineart Algebraic Systems*, Prentice-Hall, Englewood Cliffs 1967.
3. B. A. Finlayson: *Nonlinear Analysis in Chemical Engineering*, McGraw-Hill, New York 1980; reprinted, Ravenna Park, Seattle 2003.

4. S. C. Eisenstat, M. H. Schultz, A. H. Sherman: "Algorithms and Data Structures for Sparse Symmetric Gaussian Elimination," *SIAM J. Sci. Stat. Comput.* **2** (1981) 225 – 237.
5. I. S. Duff: *Direct Methods for Sparse Matrices*, Charendon Press, Oxford 1986.
6. H. S. Price, K. H. Coats, "Direct Methods in Reservoir Simulation," *Soc. Pet. Eng. J.* **14** (1974) 295 – 308.
7. A. Bykat: "A Note on an Element Re-Ordering Scheme," *Int. J. Num. Methods Egn.* **11** (1977) 194 –198.
8. M. R. Hesteness, E. Stiefel: "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bur. Stand* **29** (1952) 409 – 439.
9. Y. Saad: *Iterative Methods for Sparse Linear Systems*, 2nd ed., Soc. Ind. Appl. Math., Philadelphia 2003.
10. Y. Saad, M. Schultz: "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems." *SIAM J. Sci. Statist. Comput.* **7** (1986) 856–869.
11. http://mathworld.wolfram.com/GeneralizedMinimalResidualMethod.html.
12. http://www.netlib.org/linalg/html_templates/Templates.html.
13. http://software.sandia.gov/.
14. E. Isaacson, H. B. Keller, *Analysis of Numerical Methods*, J. Wiley and Sons, New York 1966.
15. W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling: *Numerical Recipes*, Cambridge University Press, Cambridge 1986.
16. R. W. H. Sargent: "A Review of Methods for Solving Non-linear Algebraic Equations," in R. S. H. Mah, W. D. Seider (eds.): *Foundations of Computer-Aided Chemical Process Design*, American Institute of Chemical Engineers, New York 1981.
17. J. D. Seader: "Computer Modeling of Chemical Processes," *AIChE Monogr. Ser.* **81** (1985) no. 15.
18. http://software.sandia.gov/trilinos/packages/nox/loca_user.html
19. N. R. Amundson: *Mathematical Methods in Chemical Engineering*, Prentice-Hall, Englewood Cliffs, N.J. 1966.
20. R. H. Perry, D. W Green: *Perry's Chemical Engineers' Handbook*, 7th ed., McGraw-Hill, New York 1997.
21. D. S. Watkins: "Understanding the QR Algorithm," *SIAM Rev.* **24** (1982) 427 – 440.
22. G. F. Carey, K. Sepehrnoori: "Gershgorin Theory for Stiffness and Stability of Evolution Systems and Convection-Diffusion," *Comp. Meth. Appl. Mech.* **22** (1980) 23 – 48.
23. M. Abranowitz, I. A. Stegun: *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C. 1972.
24. J. C. Daubisse: "Some Results about Approximation of Functions of One or Two Variables by Sums of Exponentials, " *Int. J. Num. Meth. Eng.* **23** (1986) 1959 – 1967.
25. O. C. McGehee: *An Introduction to Complex Analysis*, John Wiley & Sons, New York 2000.
26. H. A. Priestley: *Introduction to complex analysis*, Oxford University Press, New York 2003.
27. Y. K. Kwok: *Applied complex variables for scientists and engineers,* Cambridge University Press, New York 2002.
28. N. Asmar, G. C. Jones: *Applied complex analysis with partial differential equations*, Prentice Hall, Upper Saddle River, NJ, 2002.
29. M. J. Ablowitz, A. S. Fokas: *Complex variables: Introduction and applications*, Cambridge University Press, New York 2003.
30. J. W. Brown, R. V. Churchill: *Complex variables and applications*, 6th ed., McGraw-Hill, New York 1996; 7th ed. 2003.
31. W. Kaplan: *Advanced calculus*, 5th ed., Addison-Wesley, Redwood City, Calif., 2003.
32. E. Hille: *Analytic Function Theory*, Ginn and Co., Boston 1959.
33. R. V. Churchill: *Operational Mathematics*, McGraw-Hill, New York 1958.
34. J. W. Brown, R. V. Churchill: *Fourier Series and Boundary Value Problems*, 6th ed., McGraw-Hill, New York 2000.
35. R. V. Churchill: *Operational Mathematics*, 3rd ed., McGraw-Hill, New York 1972.
36. B. Davies: *Integral Transforms and Their Applications*, 3rd ed., Springer, Heidelberg 2002.
37. D. G. Duffy: *Transform Methods for Solving Partial Differential Equations*, Chapman & Hall/CRC, New York 2004.
38. A. Varma, M. Morbidelli: *Mathematical Methods in Chemical Engineering*, Oxford, New York 1997.
39. Bateman, H., *Tables of Integral Transforms, vol. I*, McGraw-Hill, New York 1954.
40. H. F. Weinberger: *A First Course in Partial Differential Equations*, Blaisdell, Waltham, Mass. 1965.
41. R. V. Churchill: *Operational Mathematics*, McGraw-Hill, New York 1958.

42. J. T. Hsu, J. S. Dranoff: "Numerical Inversion of Certain Laplace Transforms by the Direct Application of Fast Fourier Transform (FFT) Algorithm," *Comput. Chem. Eng.* **11** (1987) 101 – 110.
43. E. Kreyzig: *Advanced Engineering Mathematics*, 9th ed., John Wiley & Sons, New York 2006.
44. H. S. Carslaw, J. C. Jaeger: *Conduction of Heat in Solids*, 2nd ed., Clarendon Press, Oxford – London 1959.
45. R. B. Bird, R. C. Armstrong, O. Hassager: *Dynamics of Polymeric Liquids*, 2nd ed., Appendix A, Wiley-Interscience, New York 1987.
46. R. S. Rivlin, *J. Rat. Mech. Anal.* **4** (1955) 681 – 702.
47. N. R. Amundson: *Mathematical Methods in Chemical Engineering; Matrices and Their Application*, Prentice-Hall, Englewood Cliffs, N.J. 1966.
48. I. S. Sokolnikoff, E. S. Sokolnikoff: *Higher Mathematics for Engineers and Physicists*, McGraw-Hill, New York 1941.
49. R. C. Wrede, M. R. Spiegel: *Schaum's outline of theory and problems of advanced calculus*, 2nd ed, McGraw Hill, New York 2006.
50. P. M. Morse, H. Feshbach: *Methods of Theoretical Physics*, McGraw-Hill, New York 1953.
51. B. A. Finlayson: *The Method of Weighted Residuals and Variational Principles*, Academic Press, New York 1972.
52. G. Forsythe, M. Malcolm, C. Moler: *Computer Methods for Mathematical Computation*, Prentice-Hall, Englewood Cliffs, N.J. 1977.
53. J. D. Lambert: *Computational Methods in Ordinary Differential Equations*, J. Wiley and Sons, New York 1973.
54. N. R. Amundson: *Mathematical Methods in Chemical Engineering*, Prentice-Hall, Englewood Cliffs, N.J. 1966.
55. J. R. Rice: *Numerical Methods, Software, and Analysis*, McGraw-Hill, New York 1983.
56. M. B. Bogacki, K. Alejski, J. Szymanowski: "The Fast Method of the Solution of Reacting Distillation Problem," *Comput. Chem. Eng.* **13** (1989) 1081 – 1085.
57. C. W. Gear: *Numerical Initial-Value Problems in Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J. 1971.
58. N. B. Ferguson, B. A. Finlayson: "Transient Modeling of a Catalytic Converter to Reduce Nitric Oxide in Automobile Exhaust," *AIChE J* **20** (1974) 539 – 550.
59. W. F. Ramirez: *Computational Methods for Process Simulations*, 2nd ed., Butterworth-Heinemann, Boston 1997.
60. U. M. Ascher, L. R. Petzold: *Computer methods for ordinary differential equations and differential-algebraic equations*, SIAM, Philadelphia 1998.
61. K. E. Brenan, S. L. Campbell, L. R. Petzold: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Elsevier, Amsterdam 1989.
62. C. C. Pontelides, D. Gritsis, K. R. Morison, R. W. H. Sargent: "The Mathematical Modelling of Transient Systems Using Differential-Algebraic Equations," *Comput. Chem. Eng.* **12** (1988) 449 – 454.
63. G. A. Byrne, P. R. Ponzi: "Differential-Algebraic Systems, Their Applications and Solutions," *Comput. Chem. Eng.* **12** (1988) 377 – 382.
64. L. F. Shampine, M. W. Reichelt: The MATLAB ODE Suite, *SIAM J. Sci. Comp.* **18** (1997) 122.
65. M. Kubicek, M. Marek: *Computational Methods in Bifurcation Theory and Dissipative Structures*, Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1983.
66. T. F. C. Chan, H. B. Keller: "Arc-Length Continuation and Multi-Grid Techniques for Nonlinear Elliptic Eigenvalue Problems," *SIAM J. Sci. Stat. Comput.* **3** (1982) 173 – 194.
67. M. F. Doherty, J. M. Ottino: "Chaos in Deterministic Systems: Strange Attractors, Turbulence and Applications in Chemical Engineering," *Chem. Eng. Sci.* **43** (1988) 139 – 183.
68. M. P. Allen, D. J. Tildesley: *Computer Simulation of Liquids*, Clarendon Press, Oxford 1989.
69. D. Frenkel, B. Smit: *Understanding Molecular Simulation*, Academic Press, San Diego 2002.
70. J. M. Haile: *Molecular Dynamics Simulation*, John Wiley & Sons, New York 1992.
71. A. R. Leach: *Molecular Modelling: Principles and Applications*, Prentice Hall, Englewood Cliffs, NJ, 2001.
72. T. Schlick: *Molecular Modeling and Simulations*, Springer, New York 2002.
73. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena*, 2nd ed., John Wiley & Sons, New York 2002.
74. R. B. Bird, R. C. Armstrong, O. Hassager: *Dynamics of Polymeric Liquids*, 2nd ed., Wiley-Interscience, New York 1987.

75. P. V. Danckwerts: "Continuous Flow Systems," *Chem. Eng. Sci.* **2** (1953) 1 – 13.
76. J. F. Wehner, R. Wilhelm: "Boundary Conditions of Flow Reactor," *Chem. Eng. Sci.* **6** (1956) 89 – 93.
77. V. Hlavácek, H. Hofmann: "Modeling of Chemical Reactors-XVI-Steady-State Axial Heat and Mass Transfer in Tubular Reactors. An Analysis of the Uniqueness of Solutions," *Chem. Eng. Sci.* **25** (1970) 173 – 185.
78. B. A. Finlayson: *Numerical Methods for Problems with Moving Fronts*, Ravenna Park Publishing Inc., Seattle 1990.
79. E. Isaacson, H. B. Keller: *Analysis of Numerical Methods*, J. Wiley and Sons, New York 1966.
80. C. Lanczos: "Trigonometric Interpolation of Empirical and Analytical Functions," *J. Math. Phys. (Cambridge Mass.) 17* (1938) 123 – 199.
81. C. Lanczos: *Applied Analysis*, Prentice-Hall, Englewood Cliffs, N.J. 1956.
82. J. Villadsen, W. E. Stewart: "Solution of Boundary-Value Problems by Orthogonal Collocation," *Chem. Eng. Sci.* **22** (1967) 1483 – 1501.
83. M. L. Michelsen, J. Villadsen: "Polynomial Solution of Differential Equations" pp. 341 – 368 in R. S. H. Mah, W. D. Seider (eds.): *Foundations of Computer-Aided Chemical Process Design*, Engineering Foundation, New York 1981.
84. W. E. Stewart, K. L. Levien, M. Morari: "Collocation Methods in Distillation," in A. W. Westerberg, H. H. Chien (eds.): *Proceedings of the Second Int. Conf. on Foundations of Computer-Aided Process Design*, Computer Aids for Chemical Engineering Education (CACHE), Austin Texas, 1984, pp. 535 – 569.
85. W. E. Stewart, K. L. Levien, M. Morari: "Simulation of Fractionation by Orthogonal Collocation," *Chem. Eng. Sci.* **40** (1985) 409 – 421.
86. C. L. E. Swartz, W. E. Stewart: "Finite-Element Steady State Simulation of Multiphase Distillation," *AIChE J.* **33** (1987) 1977 – 1985.
87. K. Alhumaizi, R. Henda, M. Soliman: "Numerical Analysis of a reaction-diffusion-convection system," *Comp. Chem. Eng.* **27** (2003) 579–594.
88. E. F. Costa, P. L. C. Lage, E. C. Biscaia, Jr.: "On the numerical solution and optimization of styrene polymerization in tubular reactors," *Comp. Chem. Eng.* **27** (2003) 1591–1604.
89. J. Wang, R. G. Anthony, A. Akgerman: "Mathematical simulations of the performance of trickle bed and slurry reactors for methanol synthesis," *Comp. Chem. Eng.* **29** (2005) 2474–2484.
90. V. K. C. Lee, J. F. Porter, G. McKay, A. P. Mathews: "Application of solid-phase concentration-dependent HSDM to the acid dye adsorption system," *AIChE J.* **51** (2005) 323–332.
91. C. deBoor, B. Swartz: "Collocation at Gaussian Points," *SIAM J. Num. Anal.* **10** (1973) 582 – 606.
92. R. F. Sincovec: "On the Solution of the Equations Arising From Collocation With Cubic B-Splines," *Math. Comp.* **26** (1972) 893 – 895.
93. U. Ascher, J. Christiansen, R. D. Russell: "A Collocation Solver for Mixed-Order Systems of Boundary-Value Problems," *Math. Comp.* **33** (1979) 659 – 679.
94. R. D. Russell, J. Christiansen: "Adaptive Mesh Selection Strategies for Solving Boundary-Value Problems," *SIAM J. Num. Anal.* **15** (1978) 59 – 80.
95. P. G. Ciarlet, M. H. Schultz, R. S. Varga: "Nonlinear Boundary-Value Problems I. One Dimensional Problem," *Num. Math.* **9** (1967) 394 – 430.
96. W. F. Ames: *Numerical Methods for Partial Differential Equations*, 2nd ed., Academic Press, New York 1977.
97. J. F. Botha, G. F. Pinder: *Fundamental Concepts in The Numerical Solution of Differential Equations*, Wiley-Interscience, New York 1983.
98. W. H. Press, B. P. Flanner, S. A. Teukolsky, W. T. Vetterling: *Numerical Recipes*, Cambridge Univ. Press, Cambridge 1986.
99. H. Schlichting, *Boundary Layer Theory*, 4th ed. McGraw-Hill, New York 1960.
100. R. Aris, N. R. Amundson: *Mathematical Methods in Chemical Engineering, vol. 2, First-Order Partial Differential Equations with Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
101. R. Courant, D. Hilbert: *Methods of Mathematical Physics, vol. I and II*, Intersicence, New York 1953, 1962.
102. P. M. Morse, H. Feshbach: *Methods of Theoretical Physics, vol. I and II*, McGraw-Hill, New York 1953.
103. A. D. Polyanin: *Handbook of Linear Partial Differential Equations for Engineers and*

Scientists, Chapman and Hall/CRC, Boca Raton, FL 2002.
104. D. Ramkrishna, N. R. Amundson: *Linear Operator Methods in Chemical Engineering with Applications to Transport and Chemical Reaction Systems*, Prentice Hall, Englewood Cliffs, NJ, 1985.
105. D. D. Joseph, M. Renardy, J. C. Saut: "Hyperbolicity and Change of Type in the Flow of Viscoelastic Fluids," *Arch. Rational Mech. Anal.* **87** (1985) 213 – 251.
106. H. K. Rhee, R. Aris, N. R. Amundson: *First-Order Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, N.J. 1986.
107. B. A. Finlayson: *Numerical Methods for Problems with Moving Fronts*, Ravenna Park Publishing Inc., Seattle 1990.
108. D. L. Book: *Finite-Difference Techniques for Vectorized Fluid Dynamics Calculations*, Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1981.
109. G. A. Sod: *Numerical Methods in Fluid Dynamics*, Cambridge University Press, Cambridge 1985.
110. G. H. Xiu, J. L. Soares, P. Li, A. E. Rodriques: "Simulation of five-step one-bed sorption-endanced reaction process," *AIChE J.* **48** (2002) 2817–2832.
111. A. Malek, S. Farooq: "Study of a six-bed pressure swing adsorption process," *AIChE J.* **43** (1997) 2509–2523.
112. R. J. LeVeque: *Numerical Methods for Conservation Laws*, Birkhäuser, Basel 1992.
113. W. F. Ames: "Recent Developments in the Nonlinear Equations of Transport Processes," *Ind. Eng. Chem. Fundam.* **8** (1969) 522 – 536.
114. W. F. Ames: *Nonlinear Partial Differential Equations in Engineering*, Academic Press, New York 1965.
115. W. E. Schiesser: *The Numerical Method of Lines*, Academic Press, San Diego 1991.
116. D. Gottlieb, S. A. Orszag: *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, PA 1977.
117. D. W. Peaceman: *Fundamentals of Numerical Reservoir Simulation*, Elsevier, Amsterdam 1977.
118. G. Juncu, R. Mihail: "Multigrid Solution of the Diffusion-Convection-Reaction Equations which Describe the Mass and/or Heat Transfer from or to a Spherical Particle," *Comput. Chem. Eng.* **13** (1989) 259 – 270.
119. G. R. Buchanan: *Schaum's Outline of Finite Element Analysis*," McGraw-Hill, New York 1995.
120. M. D. Gunzburger: *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, San Diego 1989.
121. H. Kardestuncer, D. H. Norrie: *Finite Element Handbook*, McGraw-Hill, New York 1987.
122. J. N. Reddy, D. K. Gartling: *The Finite Element Method in Heat Transfer and Fluid Dynamics*, 2nd ed., CRC Press, Boca Raton, FL 2000.
123. O. C. Zienkiewicz, R. L. Taylor, J. Z. Zhu: *The Finite Element Method: Its Basis & Fundamentals*, vol. **1**, 6th ed., Elsevier Butterworth-Heinemann, Burlington, MA 2005.
124. O. C. Zienkiewicz, R. L. Taylor: *The Finite Element Method*, Solid and Structural Mechanics, vol. **2**, 5th ed., Butterworth-Heinemann, Burlington, MA 2000.
125. O. C. Zienkiewicz, R. L. Taylor: *The Finite Element Method*, Fluid Dynamics, vol. **3**, 5th ed., Butterworth-Heinemann, Burlington, MA 2000.
126. Z. C. Li: *Combined Methods for Elliptic Equations with Singularities, Interfaces and Infinities*, Kluwer Academic Publishers, Boston, MA, 1998.
127. G. J. Fix, S. Gulati, G. I. Wakoff: "On the use of singular functions with finite element approximations," *J. Comp. Phys.* **13** (1973) 209–238.
128. N. M. Wigley: "On a method to subtract off a singularity at a corner for the Dirichlet or Neumann problem," *Math. Comp.* **23** (1968) 395–401.
129. H. Y. Hu, Z. C. Li, A. H. D. Cheng: "Radial Basis Collocation Methods for Elliptic Boundary Value Problems," *Comp. Math. Applic.* **50** (2005) 289–320.
130. S. J. Osher, R. P. Fedkiw:, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York 2002.
131. M. W. Chang, B. A. Finlayson: "On the Proper Boundary Condition for the Thermal Entry Problem," *Int. J. Num. Methods Eng.* **15** (1980) 935 – 942.
132. R. Peyret, T. D. Taylor: *Computational Methods for Fluid Flow*, Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1983.
133. P. M. Gresho, S. T. Chan, C. Upson, R. L. Lee: "A Modified Finite Element Method for Solving the Time-Dependent, Incompressible Navier – Stokes Equations," *Int. J. Num.*

Method. Fluids (1984) "Part 1. Theory": 557 – 589; "Part 2: Applications": 619 – 640.
134. P. M. Gresho, R. L. Sani: *Incompressible Flow and the Finite Element Method*, Advection-Diffusion, vol. **1**, John Wiley & Sons, New York 1998.
135. P. M. Gresho, R. L. Sani: *Incompressible Flow and the Finite Element Method* , Isothermal Laminar Flow, vol. **2**, John Wiley & Sons, New York 1998.
136. J. A. Sethian, *Level Set Methods*, Cambridge University Press, Cambridge 1996.
137. C. C. Lin, H. Lee, T. Lee, L. J. Weber: "A level set characteristic Galerkin finite element method for free surface flows," *Int. J. Num. Methods Fluids* **49** (2005) 521–547.
138. S. Succi: *The Lattice Boltzmann Equation for Fluid Dynamics and Beyond*, Oxford University Press, Oxford 2001.
139. M. C. Sukop, D. T. Thorne, Jr.: *Lattice Boltzmann Modeling: An Introduction for Geoscientists and Engineers*, Springer, New York 2006.
140. S. Chen, S. G. D. Doolen: "Lattice Boltzmann method for fluid flows," *Annu. Rev. Fluid Mech.* **30** (2001) 329–364.
141. H. T. Lau: *Numerical Library in C for Scientists and Engineers*, CRC Press, Boca Raton, FL 1994.
142. Y. Y. Al-Jaymany, G. Brenner, P. O. Brum: "Comparative study of lattice-Boltzmann and finite volume methods for the simulation of laminar flow through a 4:1 contraction," *Int. J. Num. Methods Fluids* **46** (2004) 903–920.
143. R. L. Burden, R. L J. D. Faires, A. C. Reynolds: *Numerical Analysis*, 8th ed., Brooks Cole, Pacific Grove, CA 2005.
144. S. C. Chapra, R. P. Canal: *Numerical Methods for Engineers*," 5th ed., McGraw-Hill, New York 2006.
145. H. T. Lau: *Numerical Library in Java for Scientists and Engineers*, CRC Press, Boca Raton, FL 2004.
146. K. W. Morton, D. F. Mayers: *Numerical Solution of Partial Differential Equations*, Cambridge University Press, New York 1994.
147. A. Quarteroni, A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer, Heidelberg 1997.
148. F. Scheid: "Schaum's Outline of Numerical Analysis," 2nd ed., McGraw-Hill, New York 1989.
149. C. T. H. Baker: *The Numerical Treatment of Integral Equations*, Clarendon Press, Oxford 1977.
150. L. M. Delves, J. Walsh (eds.): *Numerical Solution of Integral Equations*, Clarendon Press, Oxford 1974.
151. P. Linz: *Analytical and Numerical Methods for Volterra Equations*, SIAM Publications, Philadelphia 1985.
152. M. A. Golberg (ed.): *Numerical Solution of Integral Equations*, Plenum Press, New York 1990.
153. C. Corduneanu: *Integral Equations and Applications*, Cambridge Univ. Press, Cambridge 1991.
154. R. Kress: *Linear Integral Equations*, 2nd ed. Springer, Heidelberg 1999.
155. P. K. Kythe P. Purl: *Computational Methods for Linear Integral Equations*, Birkhäuser, Basel 2002.
156. G. Nagel, G. Kluge: "Non-Isothermal Multicomponent Adsorption Processes and Their Numerical Treatment by Means of Integro-Differential Equations," *Comput. Chem. Eng.* **13** (1989) 1025 – 1030.
157. P. L. Mills, S. Lai, M. P. Duduković, P. A. Ramachandran: "A Numerical Study of Approximation Methods for Solution of Linear and Nonlinear Diffusion-Reaction Equations with Discontinuous Boundary Conditions," *Comput. Chem. Eng.* **12** (1988) 37 – 53.
158. D. Duffy: *Green's Functions with Applications*, Chapman and Hall/CRC, New York 2001.
159. I. Statgold: *Green's Functions and Boundary Value Problems*, 2nd ed., Interscience, New York 1997.
160. B. Davies: *Integral Transforms and Their Applications*, 3rd ed., Springer, New York 2002.
161. J. M. Bownds, "Theory and performance of a subroutine for solving Volterra integral equations," *J. Comput.* **28** 317–332 (1982).
162. J. G. Blom, H. Brunner, "Algorithm 689; discretized collocation and iterated collocation for nonlinear Volterra integral equations of the second kind," *ACM Trans. Math. Software* **17** (1991) 167–177.
163. N. B. Ferguson, B. A. Finlayson: "Error Bounds for Approximate Solutions to Nonlinear Ordinary Differential Equations," *AIChE J.* **18** (1972) 1053 – 1059.
164. R. S. Dixit, L. L. Taularidis: "Integral Method of Analysis of Fischer – Tropsch Synthesis Reactions in a Catalyst Pellet," *Chem. Eng. Sci.* **37** (1982) 539 – 544.

165. L. Lapidus, G. F. Pinder: *Numerical Solution of Partial Differential Equations in Science and Engineering*, Wiley-Interscience, New York 1982.
166. C. K. Hsieh, H. Shang: "A Boundary Condition Dissection Method for the Solution of Boundary-Value Heat Conduction Problems with Position-Dependent Convective Coefficients," *Num. Heat Trans. Part B: Fund.* **16** (1989) 245 – 255.
167. C. A. Brebbia, J. Dominguez: *Boundary Elements – An Introductory Course*, 2nd ed., Computational Mechanics Publications, Southhamtpon 1992.
168. J. Mackerle, C. A. Brebbia (eds.): *Boundary Element Reference Book*, Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1988.
169. O. L. Mangasarian: *Nonlinear Programming*, McGraw-Hill, New York 1969.
170. G. B. Dantzig: *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J, 1963.
171. S. J. Wright: *Primal-Dual Interior Point Methods*, SIAM, Philadelphia 1996.
172. F. Hillier, G. J. Lieberman: *Introduction to Operations Research*, Holden-Day, San Francisco, 1974.
173. T. F. Edgar, D. M. Himmelblau, L. S. Lasdon: *Optimization of Chemical Processes*, McGraw-Hill Inc., New York 2002.
174. K. Schittkowski: *More Test Examples for Nonlinear Programming Codes, Lecture notes in economics and mathematical systems no. 282*, Springer-Verlag, Berlin 1987.
175. J. Nocedal, S. J. Wright: *Numerical Optimization*, Springer, New York 1999.
176. R. Fletcher: *Practical Optimization*, John Wiley & Sons, Ltd., Chichester, UK 1987
177. A. Wächter, L. T. Biegler: "On the Implementation of an Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming," **106** *Mathematical Programming* (2006) no. 1, 25 – 57.
178. C. V. Rao, J. B. Rawlings S. Wright: On the Application of Interior Point Methods to Model Predictive Control. *J. Optim. Theory Appl.* **99** (1998) 723.
179. J. Albuquerque, V. Gopal, G. Staus, L. T. Biegler, B. E. Ydstie: "Interior point SQP Strategies for Large-scale Structured Process Optimization Problems," *Comp. Chem. Eng.* **23** (1997) 283.
180. T. Jockenhoevel, L. T. Biegle, A. Wächter: "Dynamic Optimization of the Tennessee Eastman Process Using the OptControlCentre," *Comp. Chem. Eng.* **27** (2003) no. 11, 1513–1531.
181. L. T. Biegler, A. M. Cervantes, A. Wächter: "Advances in Simultaneous Strategies for Dynamic Process Optimization," *Chemical Engineering Science* **57** (2002) no. 4, 575–593.
182. C. D. Laird, L. T. Biegler, B. van Bloemen Waanders, R. A. Bartlett: "Time Dependent Contaminant Source Determination for Municipal Water Networks Using Large Scale Optimization," *ASCE Journal of Water Resource Management and Planning* **131** (2005) no. 2, 125.
183. A. R. Conn, N. Gould, P. Toint: *Trust Region Methods*, SIAM, Philadelphia 2000.
184. A. Drud: "CONOPT – A Large Scale GRG Code," *ORSA Journal on Computing* **6** (1994) 207–216.
185. B. A. Murtagh, M. A. Saunders: MINOS 5.1 User's Guide, Technical Report SOL 83-20R, Stanford University, Palo Alto 1987.
186. R. H. Byrd, M. E. Hribar, J. Nocedal: An Interior Point Algorithm for Large Scale Nonlinear Programming, *SIAM J. Opt.* **9** (1999) no. 4, 877.
187. R. Fletcher, N. I. M. Gould, S. Leyffer, Ph. L. Toint, A. Waechter: "Global Convergence of a Trust-region (SQP)-filter Algorithms for General Nonlinear Programming," *SIAM J. Opt.* **13** (2002) no. 3, 635–659.
188. R. J. Vanderbei, D. F. Shanno: An Interior Point Algorithm for Non-convex Nonlinear Programming. Technical Report SOR-97-21, CEOR, Princeton University, Princeton, NJ, 1997.
189. P. E. Gill, W. Murray, M. Wright: *Practical Optimization*, Academic Press, New York 1981.
190. P. E. Gill, W. Murray, M. A. Saunders: *User's guide for SNOPT: A FORTRAN Package for Large-scale Nonlinear Programming, Technical report, Department of Mathematics*, University of California, San Diego 1998.
191. J. T. Betts: *Practical Methods for Optimal Control Using Nonlinear Programming, Advances in Design and Control 3*, SIAM, Philadelphia 2001.
192. gPROMS User's Guide, PSE Ltd., London 2002.
193. G. E. P. Box: "Evolutionary Operation: A Method for Increasing Industrial Productivity," *Applied Statistics* **6** (1957) 81–101.
194. R. Hooke, T. A. Jeeves: "Direct Search Solution of Numerical and Statistical Problems," *J. ACM* **8** (1961) 212.

195. M. J. D. Powell: "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives, *Comput. J.* **7** (1964) 155.
196. J. A. Nelder, R. Mead: "A Simplex Method for Function Minimization," *Computer Journal* **7** (1965) 308.
197. R. Luus, T. H. I. Jaakola: "Direct Search for Complex Systems," *AIChE J* **19** (1973) 645–646.
198. R. Goulcher, J. C. Long: "The Solution of Steady State Chemical Engineering Optimization Problems using a Random Search Algorithm," *Comp. Chem. Eng.* **2** (1978) 23.
199. J. R. Banga, W. D. Seider: "Global Optimization of Chemical Processes using Stochastic Algorithms," C. Floudas and P. Pardalos (eds.): *State of the Art in Global Optimization*, Kluwer, Dordrecht 1996, p. 563.
200. P. J. M. van Laarhoven, E. H. L. Aarts: *Simulated Annealing: Theory and Applications*, Reidel Publishing, Dordrecht 1987.
201. J. H. Holland: *Adaptations in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor 1975.
202. J. E. Dennis, V. Torczon:, "Direct Search Methods on Parallel Machines," *SIAM J. Opt.* **1** (1991) 448.
203. A. R. Conn, K. Scheinberg, P. Toint: "Recent Progress in Unconstrained Nonlinear Optimization without Derivatives," *Mathemtatical Programming, Series B*, **79** (1997) no. 3, 397.
204. M. Eldred: "DAKOTA: A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis," 2002, http://endo.sandia.gov/DAKOTA/software.html
205. A. J. Booker et al.: "A Rigorous Framework for Optimization of Expensive Functions by Surrogates," CRPC Technical Report 98739, Rice University, Huston TX, February 1998.
206. R. Horst, P. M. Tuy: "Global Optimization: Deterministic Approaches," 3rd ed., Springer Verlag, Berlin 1996.
207. I. Quesada, I. E. Grossmann: "A Global Optimization Algorithm for Linear Fractional and Bilinear Programs," *Journal of Global Optimization* **6** (1995) no. 1, 39–76.
208. G. P. McCormick: "Computability of Global Solutions to Factorable Nonconvex Programs: Part I – Convex Underestimating Problems," *Mathematical Programming* **10** (1976) 147–175.
209. H. S. Ryoo, N. V. Sahinidis: "Global Optimization of Nonconvex NLPs and MINLPs with Applications in Process Design," *Comp. Chem. Eng.* **19** (1995) no. 5, 551–566.
210. M. Tawarmalani, N. V. Sahinidis: "Global Optimization of Mixed Integer Nonlinear Programs: A Theoretical and Computational Study," *Mathematical Programming* **99** (2004) no. 3, 563–591.
211. C. S. Adjiman., I.P. Androulakis, C.A. Floudas: "Global Optimization of MINLP Problems in Process Synthesis and Design. *Comp. Chem. Eng.*, **21** (1997) Suppl., S445–S450.
212. C.S. Adjiman, I.P. Androulakis and C.A. Floudas, "Global Optimization of Mixed-Integer Nonlinear Problems," *AIChE Journal*, **46** (2000) no. 9, 1769–1797.
213. C. S. Adjiman, I.P. Androulakis, C.D. Maranas, C. A. Floudas: "A Global Optimization Method, αBB, for Process Design," *Comp. Chem. Eng.*, **20** (1996) Suppl., S419–S424.
214. J. M. Zamora, I. E. Grossmann: "A Branch and Contract Algorithm for Problems with Concave Univariate, Bilinear and Linear Fractional Terms," *Journal of Gobal Optimization* **14** (1999) no. 3, 217–249.
215. E. M. B. Smith, C. C. Pantelides: "Global Optimization of Nonconvex NLPs and MINLPs with Applications in Process Design," *Comp. Chem. Eng.*, **21** (1997) no. 1001, S791–S796.
216. J. E. Falk, R. M. Soland: "An Algorithm for Separable Nonconvex Programming Problems," *Management Science* **15** (1969) 550–569.
217. F. A. Al-Khayyal, J. E. Falk: "Jointly Constrained Biconvex Programming," *Mathematics of Operations Research* **8** (1983) 273–286.
218. F. A. Al-Khayyal: "Jointly Constrained Bilinear Programs and Related Problems: An Overview," *Computers and Mathematics with Applications*, **19** (1990) 53–62.
219. I. Quesada, I. E. Grossmann: "A Global Optimization Algorithm for Heat Exchanger Networks," *Ind. Eng. Chem. Res.* **32** (1993) 487–499.
220. H. D. Sherali, A. Alameddine: "A New Reformulation-Linearization Technique for

Bilinear Programming Problems," *Journal of Global Optimization* **2** (1992) 379–410.
221. H. D. Sherali, C. H. Tuncbilek: "A Reformulation-Convexification Approach for Solving Nonconvex Quadratic Programming Problems," *Journal of Global Optimization* **7** (1995) 1–31.
222. C. A. Floudas: *Deterministic Global Optimization: Theory, Methods and Applications*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.
223. M. Tawarmalani, N. Sahinidis: *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*, Kluwer Academic Publishers, Dordrecht 2002.
224. R. Horst, H. Tuy: *Global optimization: Deterministic approaches*. Springer-Verlag, Berlin 1993.
225. A. Neumaier, O. Shcherbina, W. Huyer, T. Vinko: "A Comparison of Complete Global Optimization Solvers," *Math. Programming B* **103** (2005) 335–356.
226. L. Lovász A. Schrijver: "Cones of Matrices and Set-functions and 0-1 Optimization," *SIAM J. Opt.* **1** (1991) 166–190.
227. H. D. Sherali, W.P. Adams: "A Hierarchy of Relaxations Between the Continuous and Convex Hull Representations for Zero-One Programming Problems," *SIAM J. Discrete Math.* **3** (1990) no. 3, 411–430.
228. E. Balas, S. Ceria, G. Cornuejols: "A Lift-and-Project Cutting Plane Algorithm for Mixed 0-1 Programs," *Mathematical Programming* **58** (1993) 295–324.
229. A. H. Land, A.G. Doig: "An Automatic Method for Solving Discrete Programming Problems," *Econometrica* **28** (1960) 497–520.
230. E. Balas: "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," *Operations Research* **13** (1965) 517–546.
231. R. J. Dakin: "A Tree Search Algorithm for Mixed-Integer Programming Problems", *Computer Journal* **8** (1965) 250–255.
232. R. E. Gomory: "Outline of an Algorithm for Integer Solutions to Linear Programs," *Bulletin of the American Mathematics Society* **64** (1958) 275–278.
233. H. P. Crowder, E. L. Johnson, M. W. Padberg: "Solving Large-Scale Zero-One Linear Programming Problems," *Operations Research* **31** (1983) 803–834.
234. T. J. Van Roy, L. A. Wolsey: "Valid Inequalities for Mixed 0-1 Programs," *Discrete Applied Mathematics* **14** (1986) 199–213.
235. E. L. Johnson, G. L. Nemhauser, M. W. P. Savelsbergh: "Progress in Linear Programming Based Branch-and-Bound Algorithms: Exposition," *INFORMS Journal of Computing* **12** (2000) 2–23.
236. J. F. Benders: "Partitioning Procedures for Solving Mixed-variables Programming Problems," *Numeri. Math.* **4** (1962) 238–252.
237. G. L. Nemhauser, L. A. Wolsey: *Integer and Combinatorial Optimization*, Wiley-Interscience, New York 1988.
238. C. Barnhart, E. L. Johnson, G. L. Nemhauser, M. W. P. Savelsbergh, P. H. Vance: "Branch-and-price: Column Generation for Solving Huge Integer Programs," *Operations Research* **46** (1998) 316–329.
239. O. K. Gupta, V. Ravindran: "Branch and Bound Experiments in Convex Nonlinear Integer Programming," *Management Science* **31** (1985) no. 12, 1533–1546.
240. S. Nabar, L. Schrage: "Modeling and Solving Nonlinear Integer Programming Problems," Presented at Annual AIChE Meeting, Chicago 1991.
241. B. Borchers, J.E. Mitchell: "An Improved Branch and Bound Algorithm for Mixed Integer Nonlinear Programming," *Computers and Operations Research* **21** (1994) 359–367.
242. R. Stubbs, S. Mehrotra: "A Branch-and-Cut Method for 0-1 Mixed Convex Programming," *Mathematical Programming* **86** (1999) no. 3, 515–532.
243. S. Leyffer: "Integrating SQP and Branch and Bound for Mixed Integer Noninear Programming," *Computational Optimization and Applications* **18** (2001) 295–309.
244. A. M. Geoffrion: "Generalized Benders Decomposition," *Journal of Optimization Theory and Applications* **10** (1972) no. 4, 237–260.
245. M. A. Duran, I.E. Grossmann: "An Outer-Approximation Algorithm for a Class of Mixed-integer Nonlinear Programs," *Math Programming* **36** (1986) 307.
246. X. Yuan, S. Zhang, L. Piboleau, S. Domenech : "Une Methode d'optimisation Nonlineare en Variables Mixtes pour la Conception de Procedes," *RAIRO* **22** (1988) 331.
247. R. Fletcher, S. Leyffer: "Solving Mixed Integer Nonlinear Programs by Outer Approximation," *Math Programming* **66** (1974) 327.

248. I. Quesada, I.E. Grossmann: "An LP/NLP Based Branch and Bound Algorithm for Convex MINLP Optimization Problems," *Comp. Chem. Eng.* **16** (1992) 937–947.
249. T. Westerlund, F. Pettersson: "A Cutting Plane Method for Solving Convex MINLP Problems," *Comp. Chem. Eng.* **19** (1995) S131–S136.
250. O. E. Flippo, A. H. G. R. Kan: "Decomposition in General Mathematical Programming," *Mathematical Programming* **60** (1993) 361–382.
251. T. L. Magnanti, R. T. Wong: "Acclerated Benders Decomposition: Algorithm Enhancement and Model Selection Criteria," *Operations Research* **29** (1981) 464–484.
252. N. V. Sahinidis, I. E. Grossmann: "Convergence Properties of Generalized Benders Decomposition," *Comp. Chem. Eng.* **15** (1991) 481.
253. J. E. Kelley Jr.: "The Cutting-Plane Method for Solving Convex Programs," *J. SIAM* **8** (1960) 703–712.
254. S. Leyffer: "Deterministic Methods for Mixed-Integer Nonlinear Programming," Ph.D. thesis, Department of Mathematics and Computer Science, University of Dundee, Dundee 1993.
255. M. S. Bazaraa, H. D. Sherali, C. M. Shetty: *Nonlinear Programming*, John Wiley & Sons, Inc., New York 1994.
256. G. R. Kocis, I. E. Grossmann: "Relaxation Strategy for the Structural Optimization of Process Flowsheets," *Ind. Eng. Chem. Res.* **26** (1987) 1869.
257. I. E. Grossmann: "Mixed-Integer Optimization Techniques for Algorithmic Process Synthesis", *Advances in Chemical Engineering*, vol. **23**, Process Synthesis, Academic Press, London 1996, pp. 171–246.
258. E. M. B. Smith, C. C. Pantelides: "A Symbolic Reformulation/Spatial Branch and Bound Algorithm for the Global Optimization of Nonconvex MINLPs," *Comp. Chem. Eng.* **23** (1999) 457–478.
259. P. Kesavan P. P. I. Barton: "Generalized Branch-and-cut Framework for Mixed-integer Nonlinear Optimization Problems," *Comp. Chem. Eng.* **24** (2000) 1361–1366.
260. S. Lee I. E. Grossmann: "A Global Optimization Algorithm for Nonconvex Generalized Disjunctive Programming and Applications to Process Systems," *Comp. Chem. Eng.* **25** (2001) 1675–1697.
261. R. Pörn, T. Westerlund: "A Cutting Plane Method for Minimizing Pseudo-convex Functions in the Mixed-integer Case," *Comp. Chem. Eng.* **24** (2000) 2655–2665.
262. J. Viswanathan, I. E. Grossmann: "A Combined Penalty Function and Outer-Approximation Method for MINLP Optimization," *Comp. Chem. Eng.* **14** (1990) 769.
263. A. Brooke, D. Kendrick, A. Meeraus, R. Raman: GAMS – A User's Guide, www.gams.com, 1998.
264. N. V. A. Sahinidis, A. Baron: "A General Purpose Global Optimization Software Package," *Journal of Global Optimization* **8** (1996) no.2, 201–205.
265. C. A. Schweiger, C. A. Floudas: "Process Synthesis, Design and Control: A Mixed Integer Optimal Control Framework," Proceedings of DYCOPS-5 on Dynamics and Control of Process Systems, Corfu, Greece 1998, pp. 189–194.
266. R. Raman, I. E. Grossmann: "Modelling and Computational Techniques for Logic Based Integer Programming," *Comp. Chem. Eng.* **18** (1994) no.7, 563.
267. J. N. Hooker, M. A. Osorio: "Mixed Logical.Linear Programming," *Discrete Applied Mathematics* **96-97** (1994) 395–442.
268. P. V. Hentenryck: *Constraint Satisfaction in Logic Programming*, MIT Press, Cambridge, MA, 1989.
269. J. N. Hooker: *Logic-Based Methods for Optimization: Combining Optimization and Constraint Satisfaction*, John Wiley & Sons, New York 2000.
270. E. Balas: "Disjunctive Programming," *Annals of Discrete Mathematics* **5** (1979) 3–51.
271. H. P. Williams: *Mathematical Building in Mathematical Programming*, John Wiley, Chichester 1985.
272. R. Raman, I. E. Grossmann: "Relation between MILP Modelling and Logical Inference for Chemical Process Synthesis," *Comp. Chem. Eng.* **15** (1991) no. 2, 73.
273. S. Lee, I. E. Grossmann: "New Algorithms for Nonlinear Generalized Disjunctive Programming," *Comp. Chem. Eng.* **24** (2000) no. 9-10, 2125–2141.
274. J. Hiriart-Urruty, C. Lemaréchal: *Convex Analysis and Minimization Algorithms*, Springer-Verlag, Berlin, New York 1993.
275. E. Balas: "Disjunctive Programming and a Hierarchy of Relaxations for Discrete

Optimization Problems," *SIAM J. Alg. Disc. Meth.* **6** (1985) 466–486.
276. I. E. Grossmann, S. Lee: "Generalized Disjunctive Programming: Nonlinear Convex Hull Relaxation and Algorithms," *Computational Optimization and Applications* **26** (2003) 83–100.
277. S. Lee, I. E. Grossmann: "Logic-based Modeling and Solution of Nonlinear Discrete/Continuous Optimization Problems," *Annals of Operations Research* **139** 2005 267–288.
278. ILOG, Gentilly Cedex, France 1999, www.ilog.com,
279. M. Dincbas, P. Van Hentenryck, H. Simonis, A. Aggoun, T. Graf, F. Berthier: The Constraint Logic Programming Language CHIP, FGCS-88: Proceedings of International Conference on Fifth Generation Computer Systems, Tokyo, 693–702.
280. M. Wallace, S. Novello, J. Schimpf: "ECLiPSe: a Platform for Constraint Logic Programming," *ICL Systems Journal* **12** (1997) no.1, 159–200.
281. V. Pontryagin, V. Boltyanskii, R. Gamkrelidge, E. Mishchenko: *The Mathematical Theory of Optimal Processes*, Interscience Publishers Inc., New York, NY, 1962.
282. A. E. Bryson, Y. C. Ho: "Applied Optimal Control: Optimization, Estimation, and Control," Ginn and Company, Waltham, MA, 1969.
283. V. Vassiliadis, PhD Thesis, Imperial College, University of London 1993.
284. W. F. Feehery, P. I. Barton: "Dynamic Optimization with State Variable Path Constraints," *Comp. Chem. Eng.*, **22** (1998) 1241–1256.
285. L. Hasdorff:. *Gradient Optimization and Nonlinear Control*, Wiley-Interscience, New York, NY, 1976.
286. R. W. H. Sargent, G. R. Sullivan: "Development of Feed Changeover Policies for Refinery Distillation Units," *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) 113–124.
287. V. Vassiliadis, R. W. H. Sargent, C. Pantelides: "Solution of a Class of Multistage Dynamic Optimization Problems," *I & EC Research* **33** (1994) 2123.
288. K. F. Bloss, L. T. Biegler, W. E. Schiesser: "Dynamic Process Optimization through Adjoint Formulations and Constraint Aggregation,". *Ind. Eng. Chem. Res.* **38** (1999) 421–432.
289. H. G. Bock, K. J. Plitt: A Multiple Shooting Algorithm for Direct Solution of Optimal Control Problems, 9th IFAC World Congress, Budapest 1984.
290. D. B. Leineweber, H. G. Bock, J. P. Schlöder, J. V. Gallitzendörfer, A. Schäfer, P. Jansohn: "A Boundary Value Problem Approach to the Optimization of Chemical Processes Described by DAE Models," *Computers & Chemical Engineering*, April 1997. (IWR-Preprint 97-14, Universität Heidelberg, March 1997.
291. J. E. Cuthrell, L. T. Biegler: "On the Optimization of Differential- algebraic Process Systems," *AIChE Journal* **33** (1987) 1257–1270.
292. A. Cervantes, L. T. Biegler: "Large-scale DAE Optimization Using Simultaneous Nonlinear Programming Formulations. *AIChE Journal* **44** (1998) 1038.
293. P. Tanartkit, L. T. Biegler: "Stable Decomposition for Dynamic Optimization," *Ind. Eng. Chem. Res.* **34** (1995) 1253–1266.
294. S. Kameswaran, L. T. Biegler: "Simultaneous Dynamic Optimization Strategies: Recent Advances and Challenges," *Chemical Process Control – 7*, to appear 2006.
295. B. van Bloemen Waanders, R. Bartlett, K. Long, P. Boggs, A. Salinger: "Large Scale Non-Linear Programming for PDE Constrained Optimization," Sandia Technical Report SAND2002-3198, October 2002.
296. G. P. Box, J. S. Hunter, W. G. Hunter: *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed., John Wiley & Sons, New York 2005.
297. D. C. Baird: *Experimentation: An Introduction to Measurement Theory and Experiment Design*, 3rd ed., Prentice Hall, Engelwood Cliffs, NJ, 1995.
298. J. B. Cropley: "Heuristic Approach to Comples Kinetics," *ACS Symp. Ser.* **65** (1978) 292 – 302.
299. S. Lipschutz, J. J. Schiller, Jr: *Schaum's Outline of Theory and Problems of Introduction to Probability and Statistics*, McGraw-Hill, New York 1988.
300. D. S. Moore, G. P. McCabe: *Introduction to the Practice of Statistics*," 4th ed., Freeman, New York 2003.
301. D. C. Montgomery, G. C. Runger: *Applied Statistics and Probability for Engineers*, 3 rd ed., John Wiley & Sons, New York 2002.
302. D. C. Montgomery, G. C. Runger, N. F. Hubele: *Engineering Statistics*, 3rd ed., John Wiley & Sons, New York 2004.

303. G. E. P. Box, W. G. Hunter, J. S. Hunter: *Statistics for Experimenters*, John Wiley & Sons, New York 1978.
304. B. W. Lindgren: *Statistical Theory*, Macmillan, New York 1962.
305. O. A. Hougen, K. M. Watson, R. A. Ragatz: *Chemical Process Principles*, 2nd ed., part II, "Thermodynamics," J. Wiley & Sons, New York 1959.
306. L. T. Biegler, I. E. Grossmann, A. W. Westerberg: *Systematic Methods for Chemical Process Design*, Prentice-Hall Englewood Cliffs, NJ 1997.
307. I. E. Grossmann (ed.), "Global Optimization in Engineering Design", Kluwer, Dordrecht 1996.
308. J. Kallrath, "Mixed Integer Optimization in the Chemical Process Industry: Experience, Potential and Future," *Trans. I .Chem E.*, **78** (2000) Part A, 809–822.
309. J. N. Hooker: *Logic-Based Methods for Optimization*, John Wiley & Sons, New York 1999.

Mathematical Modeling

HENNING BOCKHORN, Technische Hochschule, Darmstadt, Federal Republic of Germany

1. Introduction 5
1.1. Terminology 5
1.2. Application Areas of Mathematical Modeling in Industrial Chemistry and Chemical Engineering 5
1.3. Limitations of Mathematical Models 6
2. Construction and Classification of Mathematical Models 7
2.1. Construction of Mathematical Models 7
2.2. Classification of Mathematical Models 8
3. Empirical Models 11
3.1. Linear Empirical Models 12
3.1.1. Linear Empirical Models with One Variable 12
3.1.1.1. Parameter Estimation (Linear Regression) 12
3.1.1.2. Assessment of Estimated Parameter Values 13
3.1.1.3. Sensitivity Analysis 14
3.1.1.4. Example of a Linear Model 15
3.1.1.5. Concluding Remarks 17
3.1.2. Linear Empirical Models with Several Variables 17
3.1.2.1. Parameter Estimation 18
3.1.2.2. Assessment of Estimated Parameter Values 19
3.1.2.3. Concluding Remarks 20
3.2. Nonlinear Empirical Models . . 20
3.2.1. Parameter Estimation (Nonlinear Regression) 20
3.2.1.1. Transformation of the Model into Linear Form 21
3.2.1.2. Direct Search Methods 22
3.2.1.3. Gradient Methods 23
3.2.2. Assessment of Estimated Parameter Values 25
3.2.3. Example of a Nonlinear Model . 26
3.2.4. Concluding Remarks 27
3.3. Further Calculation Methods and Model Types 27
3.3.1. Further Methods of Parameter Estimation 27
3.3.2. Other Types of Models 28
3.3.2.1. Models with Constraints on the Parameters 28
3.3.2.2. Models Based on Differential Equations 29
4. Models Based on Transport Equations for Probability Density Functions 30
4.1. Terminology 31
4.1.1. One-Dimensional Distribution and Probability Density Functions 31
4.1.2. Multidimensional Distribution and Probability Density Functions 34
4.1.3. Conditional Probability Density Functions 35
4.2. Transport Equations for Single-Point Probability Density Functions 35
4.2.1. General form of the Transport Equations for Probability Density Functions 36
4.2.2. Limitations of Single-Point Probability Density Functions 39
4.3. Examples of Calculating Probability Density Functions 40
4.3.1. Solutions for Deterministic Systems 41
4.3.1.1. Age (Residence Time) Distributions in Chemical Reactors 41
4.3.1.2. Size Distribution in Continuously Operating Crystallizers 45
4.3.2. Solutions for Statistical Systems . 45
4.3.2.1. Closure of the Transport Equation for Probability Density Functions 46
4.3.2.2. Solution Methods of the Transport Equation for Probability Density Functions 49
4.3.2.3. Example: Combustion of Propane in a Turbulent Diffusion Flame . 50
5. Models Based on Physicochemical Principles (Transport Phenomena) 52
5.1. Application of the Principle of Conservation of Momentum . . 53
5.1.1. Laminar Tube Flow 53
5.1.2. Turbulent, Nonreactive Free Jets . 54
5.1.2.1. Models for Reynolds Stresses . . 55

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_02

5.1.2.2. Solution Method of the Resulting System of Partial Differential Equations 57
5.1.2.3. Example: Turbulent Flow of Nitrogen in Air 60
5.2. Applications of the Principle of Conservation of Enthalpy 60
5.2.1. Heat Conduction 61
5.2.2. Heat Transfer 62
5.2.2.1. Exact Solution for a Boundary Layer Problem 62
5.2.2.2. General Principles of Modeling Heat Transfer 64
5.3. Applications of the Law of Conservation of Mass 65
5.3.1. Mass Transfer without Chemical Reaction 66
5.3.1.1. Exact Solution for a Boundary Layer Problem 66
5.3.1.2. General Principles for Modeling Mass Transfer 67
5.3.1.3. Examples of Simultaneous Mass and Heat Transfer: Dynamic Models 68
5.3.1.4. Examples of Simultaneous Mass and Heat Transfer: Static Models 69
5.3.2. Mass Transfer with Chemical Reactions 73
5.3.3. Chemical Reactions in the Homogeneous Phase 75
5.3.3.1. Isothermal Reactors with Frictionless Flow, Constant Density, and Reactions Without Volume Changes 76
5.3.3.1.1. Stability of Isothermal Reactors . 79
5.3.3.1.2. Sensitivity Analysis 84
5.3.3.2. Nonisothermal Reactors 85
5.3.3.2.1. Heterogeneous Catalytic Reactions 85
5.3.3.2.2. Stability Analysis of Nonisothermal Reactors 92
5.3.3.2.3. Use on Statistical Processes . . . 97
6. References 102

Symbols

a general coefficient, temperature conductivity
$\boldsymbol{a}$ general matrix
A chemical component
A general coefficient, source term in momentum conservation equation
b general coefficient, general model parameter (estimate)
$\boldsymbol{b}$ general model parameter in vector notation (estimate)
B chemical component
B general coefficient
$\boldsymbol{B}$ general matrix
Bi Biot number
Bo Bodenstein number
c concentration, general coefficient
$\boldsymbol{c}$ general matrix
c_p specific heat capacity
C chemical component
C model parameter in turbulence models
C_1, C_2, C_3, C_4 integration constants
D chemical component
D diameter, diffusion coefficient
D_{R} tube diameter
$Da_{\mathrm{I}}, Da_{\mathrm{II}}$ Damköhler number (of 1st and 2nd kind)
E statistical event
$\boldsymbol{E}$ statistical error in vector notation
E_{a} activation energy
f dimensionless flow function, mixture fraction (normalized mass fraction of an element in a mixture)
F cross-sectional area, feed flow of a liquid, value of the F-distribution
F distribution function of a statistical variable
g acceleration due to gravity
G mass flow of gas
h enthalpy
$\boldsymbol{h}$ step size in vector notation
Δh_{v} specific latent heat of vaporization
H Heaviside step function, height
Ha Hatta number
HTU height of one transfer unit
$\boldsymbol{I}$ unit matrix
I_u, I_Φ, I_p class of integrals
j total flux of a scalar quantity
J flux of scalar quantities per unit area
$\boldsymbol{J}$ Jacobi matrix
k general parameter, rate coefficient, total heat-transfer coefficient, turbulence energy
K "adsorption" coefficient
K_p class of integrals

L length, likelihood function, mass flow of liquid, number of grid nodes in the x-direction, size, transport operator
Le Lewis number
m deterministic moment, mass fraction, number of occurrences of an event
M number of grid nodes in the y- or r-direction
M_i molecular mass
n number of data sets, number of units
Nu Nusselt number
NTU number of transfer units
O order of magnitude
p general parameter, pressure, statistical weight
$\boldsymbol{p}$ statistical weight in vector notation
P chemical component
P probability density function of a statistical variable
Pe Péclet number
Pr Prandtl number
q general parameter, number of independent variables in multivariable models, volumetric flow rate
Q general function
r general parameter, radial distance, rate of reaction
R gas constant, radius, ratio of enthalpies of vaporization
Re Reynolds number
s specific surface area, general parameter, estimate for standard deviation
S sink flow of a liquid, source term in conservation equation, surface area
s_{r}^2 mean square deviation about regression
$s^2{}_{\varepsilon}$ mean square deviation of data set
$s^2{}_{\bar{Y}_i}$ mean square deviation about $\bar{Y}_i$
Sc Schmidt number
Sh Sherwood number
t time, value of student t-variable
T temperature
u velocity
$\boldsymbol{u}$ velocity in vector notation, i.e., $\boldsymbol{u} = (u_1, u_2, u_3)$; $\boldsymbol{u} = (u, v, w)$
U velocity in a turbulent (statistical) flow
$\boldsymbol{U}$ velocity in a turbulent (statistical) flow in vector notation, i.e., $\boldsymbol{U} = (U_1, U_2, U_3)$; $\boldsymbol{U} = (\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$
V volume
w residence time distribution, velocity
W residence time distribution function, probability, velocity in a turbulent (statistical) flow
x distance, general independent variable
$\boldsymbol{x}$ distance in vector notation, i.e., $\boldsymbol{x} = (x_1, x_2, x_3)$; $\boldsymbol{x} = (x, y, z)$, general independent variable in vector notation
X partial derivatives of expected values with respect to model parameters in matrix notation
y general dependent variable
$\boldsymbol{y}$ general dependent variable in vector notation
Y general statistical variable
$\boldsymbol{Y}$ general statistical variable in vector notation

Greek Symbols

α age, general coefficient, heat-transfer coefficient, normalized mass fraction, significance level
β general coefficient, general model parameter, mass-transfer coefficient, normalized mass fraction
$\boldsymbol{\beta}$ general model parameter in vector notation
β_{r} acceleration factor
β_{H} Prater number
γ Arrhenius number, correlation coefficient, scaling factor
γ_{r} reciprocal Arrhenius number
δ Dirac delta function
δ film thickness, Kronecker symbol
ε dissipation rate, intergranular volume, normalized rate coefficient, statistical error
ζ phase shift
η effectiveness factor, normalized distance, similarity coordinate
$\boldsymbol{\eta}$ expected values in vector notation
θ general variable, normalized age, normalized residence time, normalized temperature
Θ general statistical variable, source term in scalar quantities equation
κ permeability, normalized rate coefficient
λ eigenvalues, Lagrangian multipliers, mean free path, scaling factor, thermal conductivity, wavelength
μ dynamic viscosity, normalized mass fraction

μ_{t} apparent turbulent viscosity
μ_n nth moment of a probability density function
ν frequency, kinematic viscosity, stoichiometric coefficient
ξ general variable
Ξ general statistical variable
π normalized mass fraction
ϱ density, scaling factor
σ standard deviation
σ_φ turbulent Schmidt/Prandtl numbers
τ normalized time, residence time, turbulent time scale, viscous stress
τ_{N} cooling time
φ general variable, normalized mass fraction
Φ general variable, general statistical variable
$\boldsymbol{\Phi}$ general variable and general statistical variable in vector notation, sum of squares of deviations
Φ_{K} Thiele modulus
ψ general variable
Ψ general variable, general statistical variable, stream function
$\boldsymbol{\Psi}$ general variable and general statistical variable in vector notation
Ω general function, general operator

Subscripts

a, b reference states a and b
A ambient, chemical component
ad adiabatic
B chemical component
Bf bifurcation
c central
C cooling
Cat catalyst
diff diffusion
e east, end, expansion
eff effective
ext external
G gas phase
h enthalpy
i component of a vector, chemical component
int internal
j component of a vector, chemical component
k component of a vector, chemical component, contraction, turbulence energy, summation according to the Einstein summation convention
l number of elements in a difference equation
L liquid phase
L total length
m mean
m mass
n north
o upper limit
p pressure, pressure-related
Ph phase boundary
qs quasi-steady-state
r reaction, reflected
ref reference
s south
S surface
t turbulent
T temperature
tot total
u lower limit
u velocity
w west
W wall
Wi inner wall
Wo outer wall
x local
α a component
ε dissipation rate
φ variable
ω chemical component, summation according to the Einstein summation convention
0 initial, reference state

Superscripts

$*$ transformed or normalized variable or parameter
$'$ scaled variable
$\sim$ weighted or normalized variable
$-$ mean
$\hat{}$ calculated model response
A, B, C, D chemical component
eq equilibrium
f fluctuating
G gas phase
h enthalpy
L liquid phase
m mass
(n) number of iterations
T transpose of a vector or matrix

1. Introduction

1.1. Terminology

Models are used in almost every branch of science and technology. It is difficult to encompass the many meanings of the term "model" with one single definition. The classical definition is "An object M (an object, physical or ideal system, or process) is a model if analogies exist between M and another object O that permit conclusions to be made about O" [1]. This definition describes the representation of the object O and presupposes its isolation from the whole of reality. Thus a model represents reality or a part of it. Since only certain conclusions are required, the model is reduced to parts or individual aspects of reality. The restriction to analogies between model M and object O results in limited conformity of their function, structure, and behavior. The representation of the object O by the model M may have limited accuracy and the model M can be based on a different scale than the object O. The model is satisfactory as long as important variables and phenomena are correctly represented for the specific context or investigation.

An adequate model according to the above definition can rarely explain reality in the form of a theory. Information that has been lost in the representation cannot be retrieved from the model. Attempts at theoretically interpreting the object O with the help of model M must therefore be discussed according to the aspect, scale, and accuracy of the representation.

Analogies between model M and object O can be established in the form of mathematical equations. A mathematical model therefore represents a set of algebraic and/or differential and/or integral equations which are used to describe the behavior of the object O. The terms variables and parameters are used in this article in accordance with their definitions given in mathematical relationships, and not as defined in process control, see for example [2].

1.2. Application Areas of Mathematical Modeling in Industrial Chemistry and Chemical Engineering

The two major tasks facing engineers and scientists in industrial chemistry are (1) the operation and optimization of existing processes and (2) the design of new or improved ones. In this article the term process means a sequence of operations or treatments and transformations performed on specific materials.

For the first task engineers and scientists must obtain exact information about the particular process under consideration. Qualitative aspects and criteria must be quantified. The basic variables and parameters of the relationships which describe the individual parts of the process must be worked out. The individual parts of the process must be combined. Mathematical models are very important in this respect.

For the second task mathematical models help in applying existing processes to new or modified plants and in the definition of safer, more economically viable operating conditions. Data for the construction of new plants cannot be obtained from an operating process by running it to its technical limits; this entails a high degree of risk. In contrast, a mathematical model of a process is easy to manipulate. Unusual operating conditions can easily be simulated. The plant can even be modeled under hazardous conditions to define the limits of operating parameters or risk areas.

The main applications of mathematical models in industrial chemistry are summarized as follows:

1) *Simple Experimentation.* By means of mathematical models, existing processes can be investigated more quickly, economically, and thoroughly than in operating industrial plants. Mathematical models can simulate the plant in time lapse or slow motion. The results from the model are therefore available more quickly or have a higher time resolution.
2) *Repetition of Experiments.* Simulation of processes with mathematical models allows the influence of variations in variables or parameters to be investigated. Errors can easily be implanted into or removed from the model. In a running industrial process, this type of empirical sensitivity analysis has the disadvantage that it requires a high degree of technical effort and entails a high degree of risk.
3) *Sensitivity Analysis.* With a mathematical model, sensitivity analysis does not have to

be carried out on an empirical basis by repetition of tests with systematic changes of the variables and parameters. Sensitivity analysis as a mathematical tool allows calculation of the gradients of dependent variables with respect to the model parameters.

4) *Control and Operation.* Mathematical models can be used as an economical way of estimating the stability of industrial processes or subsystems as a prerequisite for effective control or operation. Again, use of the industrial process for direct experimentation must be ruled out because of the high risk involved.
5) *Optimization.* Description of a complex system in terms of a mathematical model allows optimization objectives for the process to be developed. The optimization objectives can easily be adjusted in accordance with changing requirements. Optimum values of operating variables or parameters can also be determined easily.
6) *Extrapolation.* A suitable mathematical model can be used to test extreme operating conditions that are not possible or practical in the running process. Surfaces of operating variables and parameters can then be constructed and the optimum conditions for running the plant can be extrapolated. Extrapolability must, however, always be discussed in the light of the limits of the model and the part of the object O being modeled.

1.3. Limitations of Mathematical Models

Mathematical models have some significant limitations. The first is the type, quantity, and accuracy of the available data. The success and results of mathematical models depend largely on the information about the object O being modeled. A mathematical model cannot be better than the physical or chemical data on which it is based. In many cases available data are not sufficient and engineers and scientists have to obtain sufficient data for a model by experimental analysis. The alternative is to simplify the model adequately. The two following examples illustrate the associated problems.

The number of theoretical plates necessary for solving a separation problem (e.g., the number of theoretical plates of a rectification column) can usually be calculated more accurately than the effectiveness factor of a separation plate in a real piece of equipment. The latter value depends on the fluid dynamics of the system and on the operating conditions of the column. However, the inaccuracy in the estimate of the effectiveness factor strongly affects the commercial viability of a separation process with a large number of theoretical plates.

Another example concerns the kinetic data for industrial processes that are mainly derived from laboratory experiments. On a laboratory scale, side reactions may be much less noticeable than in an industrial plant. Lack of information about side reactions can lead to rate data that are unsuitable for industrial process design. A further problem is the impurities of the reactants used in an industrial plant; impurities can lead to changes in the reaction rate that cannot be predicted with the pure chemicals used in laboratory experiments.

The accuracy required for the individual parameters of the model depends on the sensitivity with which the results of the model respond to changes in these parameters (see Sections 3.1.1.3 and 5.3). Those parameters which exert the greatest influence on the results of the model must be determined with the greatest accuracy.

A second restriction in the use of mathematical models lies in the mathematical tools that are available for solving the resulting system of equations. Complex structures can often be formulated mathematically. However, the mathematical and numerical methods required to solve the resulting system of equations are either not yet developed or are beyond the capacity of available computers. The rapid development of supercomputers in recent years has enabled increasingly complex problems to be solved.

A further restriction concerns the interpretation of the results of the model and the extrapolability of the representation. This limitation is illustrated in the simple example in Figure 1. The solid curve shows the dynamics of a physicochemical quantity. The object could be, for example, an intermediate in consecutive chemical reactions, φ represents its concentration:

$$\varphi = \frac{k_1}{k_2 - k_1}\psi_0 \left[\exp\left(-k_1 t\right) - \exp\left(-k_2 t\right)\right] \qquad (1)$$

where ψ_0 is the initial concentration of starting compound from which φ is formed, t is the time, and k_1 and k_2 are the rate coefficients of two consecutive first-order reactions, $k_1 \neq k_2$. The available data cover the range $0 < t < t_1$ and are shown in Figure 1 as squares. An obvious empirical model based on these data is $\varphi = kt$ and is given by the dashed line. In fact the object is represented by this model in the range $k_1 t \ll 1$ and $k_2 t \ll 1$. If the exponential functions in Equation (1) are developed for the above condition, it follows that:

$$\varphi \approx \frac{k_1}{k_2 - k_1} \psi_0 \left[(k_2 - k_1)\, t \right] = k_1 \psi_0 t \tag{2}$$

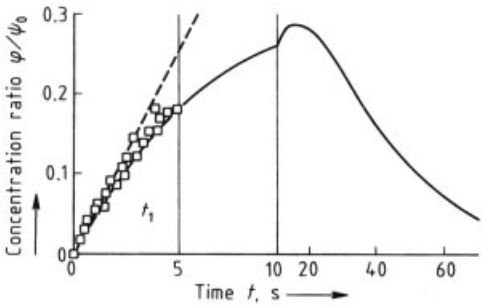

Figure 1. Relationship between a model and object with limited data

It can be seen that for the range covered by experimental data ($0 < t < t_1$), an analogy exists between the model and the object and the phenomenon of interest is also adequately reproduced within this region. The representation has a coarser scale, i.e., $\varphi = kt$ instead of $\varphi = k_1\, \psi_0 t$. However, the model does not explain the physicochemical characteristics of the object – an intermediate in a system of consecutive chemical reactions. Finally, extrapolation of this model beyond the range covered by the available data leads to a false representation.

2. Construction and Classification of Mathematical Models

2.1. Construction of Mathematical Models

Based on the terminology used in Section 1.1, the construction of a mathematical model consists of setting up a consistent set of mathematical relationships. In the ideal case these would be identical to the relationships between the process variables. Often, only certain phenomena of particular interest can be incorporated in the model due to the complexity of the real process.

The strategy of analysis used for developing mathematical models generally contains the following steps (see also [3–10]):

1) Formulation of the problem and compilation of objectives and decision criteria
2) Investigation of the process and its classification with the aim of splitting it up into subsystems (process elements)
3) Determination of the relationships between the subsystems
4) Analysis of the variables and relationships between the variables for the individual process elements
5) Setting up of mathematical relationships with variables and parameters; data acquisition
6) Investigation of the representation of the process by the model; comparison of the simulation with real process data
7) Installation of the model; interpretation and collation of the results

For further evaluation of the model the above steps can be complemented by:

1) Sensitivity analysis of the model: identification of the parameters with strong and weak influences on the response of the model
2) Model simplification

If the above steps are compiled in a flow diagram, then steps 4 – 9 must be repeated, possibly several times, until the interpretation of the results of the model still makes sense within the framework of the formulated objectives and expected solutions to the problem.

Industrial processes are very complex. The success of mathematical models in describing them often depends on whether they can be subdivided into process elements. The entire process can then be built up from the individual elements if the relationships between them have been worked out. Subdivision into process elements is not necessarily based on the physical principles of the entire process. For example, a nonideal tubular reactor can be simulated by a

series of ideal stirred tank reactors although such units may not exist in the industrial process. A packed absorption column can be represented by a number of theoretical plates even though mass transfer occurs continuously and does not result in stepwise attainment of phase equilibrium.

By correct treatment and adjustment of the process elements, modeling strives to represent the entire process as accurately as possible on the basis of simple, established principles of the individual elements.

Industrial processes are not only very complex, they can also consist of a number of qualitatively different process elements. A chemical process may, for example, encompass process elements for the preparation of the reactants, the chemical reactor, heat exchangers, and finally process elements for separation and possibly processing of the product. The different qualities of the process elements require different mathematical models depending on the physical or physicochemical principles on which they are based. An industrial heat exchanger is represented by a different type of model from that required for a rectification column, the latter has a different mathematical description from a chemical reactor, and so on. Due to the underlying physicochemical principles, mathematical modeling and the construction of mathematical models employ a wide variety of mathematical methods and methods of solution. For example, the design of distillation columns with models based on theoretical plates leads to a system of linear equations which is often solved graphically. The representation of a nonideal reactor by a series of ideal stirred tank reactors leads to a system of algebraic equations which can be highly nonlinear depending on the chemical reaction under consideration. Finally, a reactor model for a homogeneous, turbulent-flow reactor based on the modeled Navier–Stokes equations leads to a system of coupled partial differential equations that can be solved by various methods.

Mathematical modeling is used to formulate widely differing physical and physicochemical phenomena: transfer of heat, mass, and momentum, as well as chemical reactions in homogeneous and heterogeneous systems. Mathematical modeling is thus used in the design of mass-transfer operations, calculation of heat exchangers, chemical reaction engineering, and finally process control. A wide range of methods are used to formulate models and to solve the resulting systems of equations.

For this reason further formalization of the steps for building mathematical models and a further division into individual steps will be avoided. The following section examines the classification of models based on the physical background of the model, the type of system of equations, and the necessary methods for their solution.

2.2. Classification of Mathematical Models

Mathematical models can be classified according to their physical background, the type of systems of equations, and the corresponding methods of solution. Alternatives are discussed in [3], [8–12]. Figure 2 shows three classes of models with different physical backgrounds: models based on physicochemical principles, probability density function (pdf) models, and empirical models.

The first category comprises models based mainly on the mathematical formulation of transport phenomena. Use of this principle requires that the process can be subdivided into process elements that can be described by the laws governing the transport of mass, momentum, and energy, i.e., their conservation principles.

Such models are subdivided into deterministic and statistical models. Deterministic models or model elements have a determined value or set of values for each variable or parameter for any given set of conditions. In contrast, the variables and parameters used in statistical models (or elements of statistical models) are statistical quantities. They can only be given with a certain probability or in terms of moments of probability density functions. If, for example, the probability density function $P(Y)$ holds for the statistical variable Y, then $P(Y)\,\mathrm{d}Y$ is the probability for the variable lying in the range $\mathrm{d}Y$ around Y. Normally the full statistical information is not necessary for technical purposes; therefore statistical variables are often described by moments of the probability density functions. These moments are defined as

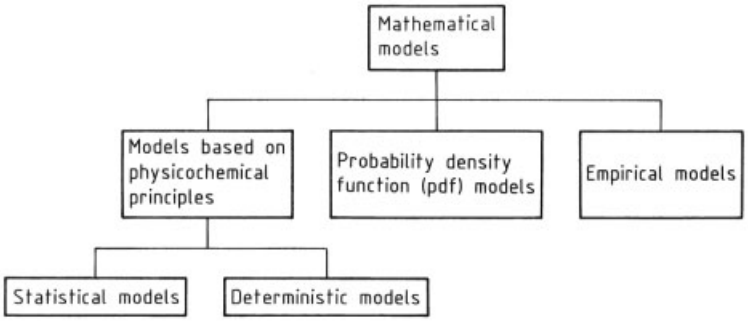

Figure 2. Classification of mathematical models according to their physical principles

$$\mu_n \equiv \int_{-\infty}^{+\infty} Y^n P(Y)\,\mathrm{d}Y$$

For example the first moment

$$\mu_1 \equiv \int_{-\infty}^{+\infty} Y P(Y)\,\mathrm{d}Y$$

is the mean value of the statistical variable Y, the second moment about the mean value is its variance Var (Y):

$$\mu_2' \equiv \int_{-\infty}^{+\infty} (Y-\mu_1)^2 P(Y)\,\mathrm{d}Y, \text{ etc.}$$

A statistical mathematical model is the mathematical description of a process, e.g., in terms of the moments defined above. Figure 3 shows some possibilities for statistical processes. Normally a deterministic process leads to deterministic results (Fig. 3 A). If a statistical error is added to the process results (Fig. 3 B), it leads to a statistical result. The same result is obtained if the deterministic process becomes unstable (Fig. 3 C); a simple example for this is buoyancy-driven turbulence. If the input process variables are statistical variables then the results of the process are also statistical variables (Fig. 3 D). Figure 3 shows that the macroscopic phenomena in statistical processes change but that the physicochemical background of the process and hence of the process model remains the same. The principles of the conservation of mass, momentum, and energy are also valid for statistical variables; the mathematical overhead for handling these cases is, however, more complicated. The transport equations used to describe a statistical process by means of statistical moments contain, for example, a number of statistical hypotheses and model assumptions for their closure but the character of the model equations does not change. For these reasons statistical models are classified as shown in Figure 3. Methods for representing statistical processes will be discussed in Chapter 5.

A further subdivision of models based on physicochemical principles according to the type of model equations is shown in Figure 4. The complexity of the method of solution decreases from right to left. The division is based on the most commonly used types of models but is by no means definitive. Thus, multidimensional models can be formulated in the form of algebraic relationships; models which can be described by algebraic relationships do not have to be steady-state models. Steady-state models describe processes in which the accumulation terms (the changes in the variables over time) disappear. Nonsteady-state models include changes in the variables over time. In models with constant variables, the properties and state are not functions of space so that the system is homogeneous. Description with locally distributed variables considers local changes in the dependent variables. (The nomenclature of process analysis does not make the distinction between variables and parameters introduced in Section 1.1; the latter two types of models are then classified as models with constant or concentrated parameters or as models with distributed parameters [10], [13–15].) Figure 4 also shows that the description of processes by differential equations and difference

equations is frequently equivalent. Differential equations must normally be solved by numerical methods. In order to achieve this they are transformed into difference equations. These methods, which are really mathematical tools, have their physical analogy in the description of processes with models based on finite changes (see Sections 4.3.1 and 5.3.3.1).

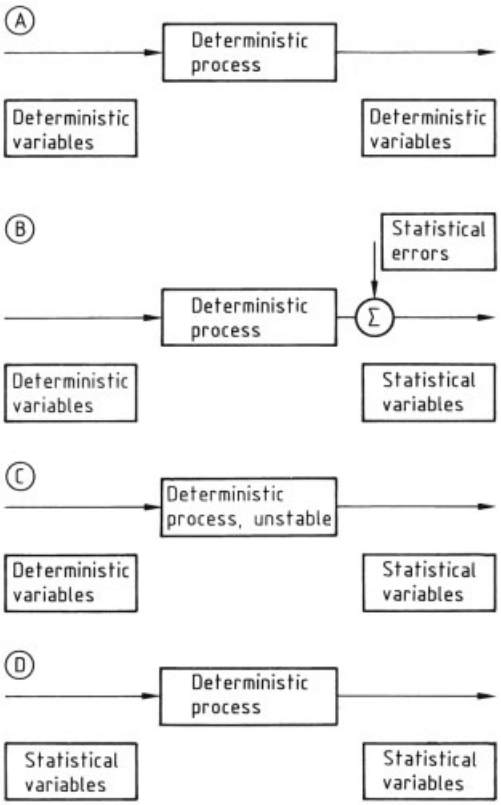

Figure 3. Statistical and deterministic processes

Models based on transport equations for probability density functions generally give their results in the form of functionals $P(\Phi_1, \ldots, \Phi_n)$. These are defined in such a way that the probability of finding the dependent variables $(\Phi_1, \ldots, \Phi_n)$ in the small range $\mathrm{d}\Phi_1, \ldots, \mathrm{d}\Phi_n$ around the functions $\Phi_1(\boldsymbol{x}, t), \ldots, \Phi_n(\boldsymbol{x}, t)$ is $P(\Phi_1, \ldots, \Phi_n)\,\mathrm{d}\Phi_1, \ldots, \mathrm{d}\Phi_n$. Probability density functions $P(\Phi_1, \ldots, \Phi_n; \boldsymbol{x}, t)$ that are defined analogously for given values of the independent variables $\boldsymbol{x}$, t are more common. They give the probability that the dependent variables at a fixed distance $\boldsymbol{x} = (x_1, x_2, x_3)$ and time t lie in the range $\mathrm{d}\Phi_1, \ldots, \mathrm{d}\Phi_n$ around the values $\Phi_1, \ldots, \Phi_n$ as $P(\Phi_1, \ldots, \Phi_n; \boldsymbol{x}, t)\,\mathrm{d}\Phi_1, \ldots, \mathrm{d}\Phi_n$. These models provide the complete statistical information for statistical processes or give distribution functions of particular process variables. They are closely related to the description of statistical processes by the statistical moments discussed above, the difference is that the transport equations for the probability density functions are formulated.

Classical examples for models based on transport equations for probability density functions are to be found in statistical mechanics and kinetic gas theory. This concept was introduced in industrial systems to describe flow and mixing in nonideally mixed reactors [16]. One-dimensional models which represent the "macromixing" in a chemical reactor in terms of age (residence time) distributions are often sufficient to allow estimation of its behavior. Other simple pdf models are used in other areas, for example to model the crystal size distribution in crystallization, the activity distribution of catalyst pellets, or the age and size distribution of microbiological cultures [17]. This concept has been used to describe reacting turbulent flow [18], [19].

Many industrial processes cannot be described with pdf models or models based on physicochemical principles because of their complexity. In these cases empirical models are used. In such cases the response of the process to variation of one or more process variables is known and the model correlates the process results. The simplest example of this is the fitting of a polynomial to experimental data. Another example in process control is the description of a process response in the form of transfer functions in the time or frequency domain. Empirical models are statistical (Fig. 3 B) because the necessary data have to be obtained experimentally and contain statistical errors. Empirical models have only limited value for describing processes or process elements, e.g., if predictions are to be made outside the range covered by experimental data or if a theoretical approach is to be verified. In spite of this empirical mathematical models are still preferred in many cases.

Discussion on mathematical modeling in the following sections is based on the classification

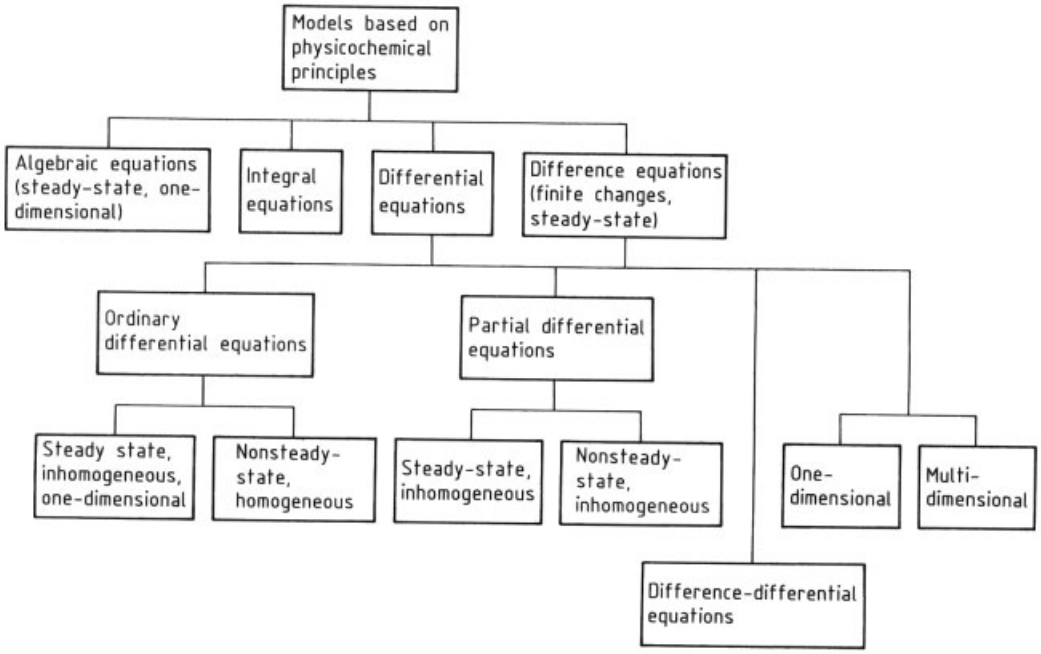

Figure 4. Classification of deterministic models based on physicochemical principles according to the type of resulting equations

of mathematical models shown in Figures 3 and 4. Empirical models will be discussed first, followed by models based on the transport equations for probability density functions, and finally models based on physicochemical principles (transport phenomena). Presentation of the important principles of mathematical modeling will be given priority over a complete description.

Nonstandard numerical and mathematical methods required for solving the resulting set of equations for the developed mathematical models are discussed in more depth. More detailed classifications of industrial processes or process variables can be found in the literature [2], [10], [13–15]. Relevant aspects of such classifications will be discussed where necessary.

Mathematical modeling of physical or physicochemical phenomena or industrial processes is a concept that is applied to many other disciplines. Aspects of mathematical modeling related to other topics of engineering are described elsewhere in the Encyclopedia, e.g., → Transport Phenomena; → Fluid Mechanics; → Process Control Engineering.

3. Empirical Models

The fundamental variables of a process are known, for example, from a series of experiments. If, however, the complexity of the process prevents the formulation of a model on the basis of physicochemical principles, the process can be described by empirical models. The fitting of a polynomial or a similar function to a set of measured values is the simplest empirical model. An empirical model represents a general relationship:

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) \tag{3}$$

where $\boldsymbol{y}$ is the vector of the dependent variables (model responses, process responses), $\boldsymbol{x}$ is the vector of the independent variables (model variables, process variables), and $\boldsymbol{\beta}$ is the vector of the parameters from which the functional relationship is constructed. This relationship is empirical and must be found on the basis of existing experimental data. In the following sections empirical models will be introduced through several examples.

3.1. Linear Empirical Models

If Equation (3) can be represented by the relation

$$\boldsymbol{y} = \boldsymbol{\beta x} \tag{4}$$

the model is linear. Equation (4) is linear with respect to both its parameters and its independent variables. This implies that the first derivatives with respect to the components of the parameter vector and the vector of the independent variables disappear. The simplest relationship of type (4) is a linear model with one dependent variable:

$$y = \beta_0 + \beta_1 x \tag{5}$$

where the estimates b_0 and b_1 of the parameters β_0 and β_1 have to be determined from experimental data. This will be outlined briefly in the following. Only the essential steps of the calculation will be presented. Detailed derivations and proofs are given in [2], [4], [5], [7], [20].

3.1.1. Linear Empirical Models with One Variable

The determination of the estimates of the parameters β_0 and β_1 will be discussed in detail for a linear model with one dependent variable. This model is based on the following assumptions:

1) The dependent variable is a random variable so that the model in Equation (5) can be given in the following form:

$$\langle \bar{Y}_i \rangle = \bar{Y}_i - \varepsilon_i = \beta_0' + \beta_1 (x_i - \bar{x}) \tag{6}$$

   where $\bar{Y}_i$ is the mean value of the dependent variables of n data sets with repeated measurements at x_i; ε_i is the random error in $\bar{Y}_i$. Hence $(\bar{Y}_i - \langle \bar{Y}_i \rangle) = \varepsilon_i$ has an expected value of zero and a constant variance. The angular brackets denote the expected values or mathematical expectation.
2) The variance of Y_i is also constant, has a normal distribution, and is equal to the variance of ε_i.
3) The observations of Y are statistically independent so that the random errors ε_i are also statistically independent.
4) The variable x is not a random variable.

A process based on these assumptions is shown in Figure 5 and is discussed as an example in Section 3.1.1.4.

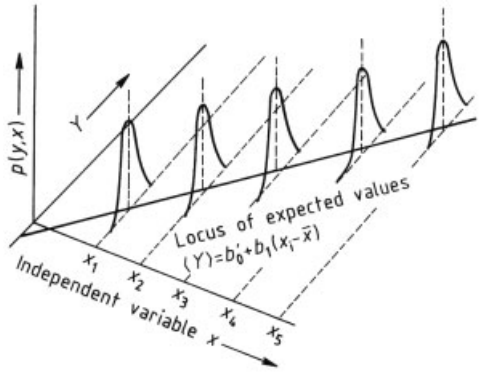

Figure 5. Representation of the statistical process described in Section 3.1.1

3.1.1.1. Parameter Estimation (Linear Regression)

Estimates b_0' and b_1 of the parameters β_0' and β_1 are found by minimizing the sum of the squares of the deviations between the observed values Y_i and the expected values of $\bar{Y}_i$, $\langle \bar{Y}_i \rangle$, i.e., the variance

$$\sum_{i=1}^{n} \left(\bar{Y}_i - \langle \bar{Y}_i \rangle \right)^2 \stackrel{!}{=} \text{Min} \tag{7}$$

This is known as least squares estimation.

The problem can also be posed in other ways. For the overdetermined system of equations for β_0' and β_1

$$\bar{Y}_1 - \beta_0' - \beta_1 (x_1 - \bar{x}) = 0 \tag{8a}$$

$$\bar{Y}_2 - \beta_0' - \beta_1 (x_2 - \bar{x}) = 0 \tag{8b}$$

$$\vdots$$

$$\bar{Y}_n - \beta_0' - \beta_1 (x_n - \bar{x}) = 0 \tag{8n}$$

the estimated values b_0 and b_1 have to be found that reduce the residuals on the left-hand sides to as close to zero as possible.

The problem is the minimization of the sum:

$$\boldsymbol{\Phi} = \sum_{i=1}^{n} p_i \left(\bar{Y}_i - \langle \bar{Y}_i \rangle\right)^2$$
$$= \sum_{i=1}^{n} p_i \left[\bar{Y}_i - \beta_0' - \beta_1 (x_i - \bar{x})\right]^2 \tag{9}$$

where p_i are statistical weights determined by the number of measurements made for each data set, i.e., the number of measured values of the dependent variables at each value of the independent variable x. If Equation (9) is partially differentiated with respect to the parameters β_0' and β_1 and the derivatives set to zero, rearrangement gives the normal equations

$$\sum_{i=1}^{n} p_i \bar{Y}_i = b_0' \sum_{i=1}^{n} p_i + b_1 \sum_{i=1}^{n} p_i (x_i - \bar{x}) \tag{10a}$$

and

$$\sum_{i=1}^{n} p_i \bar{Y}_i (x_i - \bar{x}) = b_0' \sum_{i=1}^{n} p_i (x_i - \bar{x})$$
$$+ b_1 \sum_{i=1}^{n} p_i (x_i - \bar{x})^2 \tag{10b}$$

in which the parameters β_0 and β_1 are replaced by their estimated values b_0' and b_1. Since

$$\sum_{i=1}^{n} p_i (x_i - \bar{x}) \equiv 0 \tag{11}$$

the coupling between Equations (10 a) and (10 b) disappears and they can be solved separately for b_0' and b_1:

$$b_0' = \frac{\sum_{i=1}^{n} p_i \bar{Y}_i}{\sum_{i=1}^{n} p_i} = \bar{Y} \tag{12a}$$

$$b_1 = \frac{\sum_{i=1}^{n} p_i \bar{Y}_i (x_i - \bar{x})}{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} \tag{12b}$$

By forming the second partial derivatives, it can easily be shown that Equations (10 a) and (10 b) actually represent a minimum of the sum $\boldsymbol{\Phi}$. The simple uncoupled form only arises using the scaled form given in Equation (6).

3.1.1.2. Assessment of Estimated Parameter Values

Although the estimation of the parameter values is important, the quality of the empirical model is of particular interest (see points 6 and 7, Section 2.1). The first assessment of the empirical model is the test of the hypothesis $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1 (x_i - \bar{x})$. For this the mean square deviations

$$s_r^2 = \frac{1}{n-2} \sum_{i=1}^{n} p_i \left(\bar{Y}_i - \hat{Y}_i\right)^2 \tag{13}$$

and

$$s_\varepsilon^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{p_i} \left(Y_{ij} - \bar{Y}_i\right)^2}{\sum_{i=1}^{n} p_i - n} \tag{14}$$

are formed. Equation (13) represents the mean square of the residuals, i.e., the mean square deviations from the regression line. Equation (14) represents the mean square deviations within the data sets; it is a measure of the random error which occurs in the experiments at the points x_i.

The hypothesis $\langle Y_i \rangle = \beta_0' + \beta_1 (x_i - \bar{x})$ is confirmed if the mean squares of the residuals are significantly smaller than the mean deviations within the data sets. The hypothesis can be verified by means of the F-test [4], [5], [20–23]. If, for a preset confidence number $(1-\alpha)$, for example, 0.95

$$\frac{s_r^2}{s_\varepsilon^2} > F_{(1-\alpha)}^{(n-2),\left(\sum_i p_i - n\right)} \tag{15}$$

then the mean squares of the residuals are significantly greater than the mean square deviations within one data set, and the hypothesis is not valid. In Equation (15) $F_{(1-\alpha)}^{(n-2),\left(\sum_i p_i - n\right)}$ is the value of the F-distribution for the confidence number $(1-\alpha)$ and the degrees of freedom $(n-2)$, $\left(\sum_i p_i - n\right)$. In the positive case both s_r^2 and s_ε^2 are good estimates for $\sigma_{\bar{Y}_i}^2$, the variance in $\bar{Y}_i$. If the hypothesis is valid, the variance of $\bar{Y}_i$ is required for the calculation of the variances of the estimated values of the model parameters and their confidence intervals. The variances as given by their definitions are

$$\mathrm{Var}\left(b_0'\right) = \left\langle \left(b_0' - \beta_0'\right) \right\rangle = \frac{\sigma^2_{\bar{Y}_i}}{\sum\limits_i p_i} \tag{16}$$

(cf. Eq. 12 a) and

$$\mathrm{Var}\left(b_1\right) = \left\langle \left(b_1 - \beta_1\right)^2 \right\rangle = \frac{\sigma^2_{\bar{Y}_i}}{\sum\limits_i p_i \left(x_i - \bar{x}\right)^2} \tag{17}$$

(cf. Eq. 12 b). The variance of the model response $\hat{Y}_i$ calculated from the linear model is given by

$$\mathrm{Var}\left(\hat{Y}_i\right) = \mathrm{Var}\left(b_0'\right) + \left(x_i - \bar{x}\right)^2 \mathrm{Var}\left(b_1\right)$$

$$= \sigma^2_{\bar{Y}_i} \left[\frac{1}{\sum\limits_i p_i} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\limits_i p_i \left(x_i - \bar{x}\right)^2} \right] \tag{18}$$

since b_0' and b_1 are statistically independent. Equation (18) shows that the variance of Y_i is dependent on x and shows a minimum at $\bar{x}$. A first assessment of the linear model defined by the estimated parameter values b_0' and b_1 is achieved by means of the calculated variances.

In addition to the variances, which are a measure of the scatter of the estimated values, the confidence limits of the estimated values are used and can be calculated from the variances for further assessment. Since b_0' and b_1 are linear combinations of Y_i (cf. Eq. 12), they also have a normal distribution about β_0' and β_1 respectively because of the assumed normal distribution of Y_i. Under these conditions the Student t-variable can be formed [4], [5], [21–23]:

$$t = \frac{b_0' - \beta_0'}{s_{b_0'}} = \frac{b_0' - \beta_0'}{\sigma_{\bar{Y}_i} / \left(\sum\limits_i p_i\right)^{1/2}} \tag{19}$$

This has a t-distribution with $\left(\sum\limits_i p_i - 2\right)$ degrees of freedom. Estimates of the confidence interval can be obtained using the t-distribution in the following form:

$$b_0' - t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{b_0'} \le \beta_0' \le b_0' + t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{b_0'} \tag{20}$$

With known $\sigma^2 \bar{\mathrm{Y}}_\mathrm{i}$ the interval in which the expected value β_0' lies is given. Values of the t-distribution can be found in statistics books [4], [5], [21–23].

An analogous procedure is adopted for β_1. The Student t-variable is then

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1 - \beta_1}{\sigma^2_{\bar{Y}_i} / \left[\sum\limits_i p_i \left(x_i - \bar{x}\right)^2\right]^{1/2}} \tag{21}$$

This quantity also has a t-distribution with $\sum\limits_i p_i - 2$ degrees of freedom. As before, the confidence interval for β_1 becomes:

$$b_1 - t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{b_1} \le \beta_1 \le b_1 + t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{b_1} \tag{22}$$

Confidence intervals for the expected values $\langle \bar{Y}_i \rangle$ are defined in a similar fashion. Since the $\bar{Y}_i$ values are assumed to have a normal distribution around $\langle \bar{Y}_i \rangle$, the t-variable can be expressed in the form

$$t = \frac{\bar{Y}_i - \langle \bar{Y}_i \rangle}{s_{\bar{Y}_i}} = \frac{\bar{Y}_i - \langle \bar{Y}_i \rangle}{\sigma_{\bar{Y}_i} / \left[\frac{1}{\sum\limits_i p_i} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\limits_i p_i \left(x_i - \bar{x}\right)^2}\right]^{1/2}} \tag{23}$$

which again has a t-distribution with $\left(\sum\limits_i p_i - 2\right)$ degrees of freedom. In analogy with Equations (20) and (22) the confidence interval for $\langle \bar{Y}_i \rangle$ is as follows:

$$\hat{Y}_i - t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{\bar{Y}_i} \le \langle \bar{Y}_i \rangle \le \hat{Y}_i + t_{(1-\alpha/2)}^{\left(\sum\limits_i p_i - 2\right)} s_{\bar{Y}_i} \tag{24}$$

Further discussion and illustration of the method of parameter estimation and assessment are given in Section 3.1.1.4.

3.1.1.3. Sensitivity Analysis

In addition to the analysis of the variances of β_0', β_1, and $\bar{Y}_i$, the sensitivity of the linear empirical model $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1 (x_i - \bar{x})$ can help in interpreting the results obtained from the model as well as in the design of experiments for establishing an empirical model. The sensitivity of the model may be described in various ways. In order to select experiments and methods of measurement for verification of the fundamental variables of the process, it is useful to have a measure of the

sensitivity of the sum of the squares of the deviations and of the estimated parameter values with respect to Y_i. Sensitivity can be regarded as the relative change in the sum of squares of the deviations $\partial \boldsymbol{\Phi}_{\min} / \Phi_{\min}$ or the relative change in the estimated parameter values $\partial b_i / b_i$ for a relative change $\partial \bar{Y}_i / \bar{Y}_i$ in $\bar{Y}_i$. Since

$$\boldsymbol{\Phi}_{\min} = \sum_{i=1}^{n} p_i \left(\bar{Y}_i - \hat{Y}_i \right)^2 \tag{25}$$

(cf. Section 3.1.1.1), the sensitivity of $\boldsymbol{\Phi}_{\min}$ with respect to $\bar{Y}_i$ is:

$$\frac{\partial \boldsymbol{\Phi}_{\min}}{\boldsymbol{\Phi}_{\min}} / \frac{\partial \bar{Y}_i}{\bar{Y}_i} = \frac{2 \bar{Y}_i p_i \left(\bar{Y}_i - \hat{Y}_i \right)}{\sum\limits_{i=1}^{n} p_i \left(\bar{Y}_i - \hat{Y}_i \right)^2} \tag{26}$$

Hence, relative change in a value $\bar{Y}_i$ causes a change in the minimum of the sum of squares, which increases as the deviation between $\bar{Y}_i$ and the model solution $\hat{Y}_i$ becomes larger and as the minimum becomes smaller.

The sensitivity of the estimated parameter values b'_0 and b_1 to changes in $\bar{Y}_i$ can be obtained simply and directly from Equations (12 a) and (12 b):

$$\frac{\partial b'_0}{b'_0} / \frac{\partial \bar{Y}_i}{\bar{Y}_i} = \frac{1}{\sum\limits_i p_i} \frac{p_i \bar{Y}_i}{b'_0} \tag{27}$$

and

$$\frac{\partial b_1}{b_1} / \frac{\partial \bar{Y}_i}{\bar{Y}_i} = \frac{1}{(x_i - \bar{x})^2} \frac{p_i \bar{Y}_i (x_i - \bar{x})}{b_1} \tag{28}$$

The sensitivity obtained from Equations (27) and (28) is equally simple to interpret. The more reliable the measured value is, i.e., the larger p_i becomes in comparison to $\sum\limits_i p_i$, the more rapidly b'_0 changes with $\bar{Y}_i$. The same applies to b_1, here the interval $(x_i - \bar{x})$ at which the values are measured has an additional weight.

Finally, an estimate of the sensitivity of the model with respect to its parameters is a sensible measure to evaluate the experimental effort required in the model assessment. This sensitivity is defined by the relative change in the expected value $\langle \bar{Y}_i \rangle$, $\partial \langle \bar{Y}_i \rangle / \langle \bar{Y}_i \rangle$, caused by a change of $\partial \beta_i / \beta_i$ in the model parameter. For the linear model described here formulation of the problem is trivial since the solutions

$$\frac{\partial \langle \bar{Y}_i \rangle}{\langle \bar{Y}_i \rangle} / \frac{\partial \beta'_0}{\beta'_0} = \frac{\beta'_0}{\langle \bar{Y}_i \rangle} \tag{29}$$

and

$$\frac{\partial \langle \bar{Y}_i \rangle}{\langle \bar{Y}_i \rangle} / \frac{\partial \beta_1}{\beta_1} = \frac{\beta_1}{\langle \bar{Y}_i \rangle} (x_i - \bar{x}) \tag{30}$$

are obvious and can be derived directly from Equation (6). For more complex models there are other more subtle relationships for deciding the degree of complexity of the experiments required for the model development. This type of sensitivity analysis will be discussed in detail later (Section 5.3.3.1).

3.1.1.4. Example of a Linear Model

The determination of the model parameters will now be illustrated using a simple example of a linear model with one dependent variable, as defined in Section 3.1.1.1.

As an example we will consider the correlation of the mean heat-transfer coefficients for hot air at the wall of a pipe with turbulent flow (Table 1). The gas velocity (indicated by x in Table 1) is varied and can be precisely adjusted so that it can be regarded as a deterministic variable. The measured values (the mean heat-transfer coefficients denoted by Y) are statistical quantities which satisfy the assumptions presented in Section 3.1.1. Table 1 also contains all conversions necessary for linear regression.

According to Equations (12 a) and (12 b), the estimated values for the parameters of the linear model $\langle \bar{Y}_i \rangle = \beta'_0 + \beta_1 (x_i - \bar{x})$ are $b'_0 = 15.002$ and $b = 1.282$. To test the hypothesis, the mean square deviations $s^2{}_r$ and s^2_ε are first obtained from Equations (13) and (14): $s^2_{\mathrm{r}} = 7.749/3 = 2.583$ and $s^2_\varepsilon = 7.613/9 = 0.846$. The ratio of the mean square deviations is then 3.05. The associated F-value [4] is $F\,{}^{3.9}_{0.95} = 3.86$, so that the hypothesis is valid for a significance level of 0.05. To obtain an estimated value $s^2_{\bar{Y}_i}$ for $\sigma^2_{\bar{Y}_i}$ it is better to use the combined square deviations from s^2_{r} and s^2_ε:

$$s^2_{\bar{Y}_i} = \frac{\sum\limits_{i=1}^{n} \sum\limits_{j=1}^{p_i} \left(Y_{ij} - \hat{Y}_i \right)^2}{\sum\limits_{i=1}^{n} p_i - 2} = \frac{7.613 + 7.746}{9 + 3} = 1.28$$

Table 1. Example of linear regression, model with one independent variable

Data set	p_i	x_i	Y	$\bar{Y}_i$	$p_i(x_i-\bar{x})\bar{Y}_i$	$p_i(x_i-\bar{x})^2$	$(Y-\bar{Y}_i)^2$	$(\bar{Y}_i-\hat{Y}_i)^2$	$(x_i-\bar{x})$
1	3	8.00	8.08	9.09	−136.35	75.00	1.020	0.248	−5.00
			9.95				0.739		
			9.26				0.029		
2	3	10.50	11.21	12.18	− 91.35	18.75	0.941	0.146	−2.50
			12.36				0.032		
			12.98				0.640		
3	2	13.00	14.12	15.04	0.00	0.00	0.846	0.001	0.00
			15.96				0.846		
4	2	15.50	16.80	17.71	88.55	12.50	0.828	0.247	2.50
			18.62				0.828		
5	4	18.00	19.60	20.18	403.60	100.00	0.336	1.517	5.00
			20.08				0.010		
			20.90				0.518		
			20.16				0.004		
$\sum n=5$	$\sum_i p_i=14$	$\bar{x}=13.00$	$\bar{Y}=15.002$		$\sum_i p_i\bar{Y}_i(x_i-\bar{x})=264.45$	$\sum_i p_i(x_i-\bar{x})^2=206.25$	$\sum_i(Y-\bar{Y}_i)^2=7.613$	$\sum_i p_i\left(\bar{Y}_i-\hat{Y}_i\right)^2=7.746$	

The variances of the estimated values of the model parameters can be obtained using Equations (16) and (17):

$$\mathrm{Var}\left(b_0'\right)=\frac{1.28}{14}=9.14\times10^{-2}$$

$$\mathrm{Var}\left(b_1\right)=\frac{1.28}{206.25}=6.20\times10^{-3}$$

and finally the x_i-dependent variance of the model responses $\hat{Y}_i$ are given by Equation (18).

$$\mathrm{Var}\left(\hat{Y}_i\right)=1.28\left[\frac{1}{14}+\frac{1}{206.25}\left(x_i-\bar{x}\right)^2\right]$$
$$=9.14\times10^{-2}+6.20\times10^{-3}\left(x_i-\bar{x}\right)^2$$

The individual confidence intervals of the expected values of these three quantities can be calculated for a significance level of 0.05 using Equations (20), (22), and (24) ($t_{0.975}^{\left(\sum_i p_i-2\right)}=2.179$, [4]):

$$b_0'-0.658\leq\beta_0'\leq b_0'+0.658$$

and

$$b_1-0.171\leq\beta_1\leq b_1+0.171$$

The individual confidence intervals for $\langle\bar{Y}_i\rangle$ are also dependent on x_i because of the dependence of the variance on x_i:

$$\hat{Y}_i-\left[1.99\times10^{-1}+1.35\times10^{-2}\left(x_i-\bar{x}\right)^2\right]^{1/2}$$
$$\leq\langle\bar{Y}_i\rangle\leq\hat{Y}_i+\left[1.99\times10^{-1}\right.$$
$$\left.+1.35\times10^{-2}\left(x_i-\bar{x}\right)^2\right]^{1/2}$$

The results of the regression analysis are illustrated in Figure 6: the regression line, the confidence intervals of the parameters, and the confidence intervals for the expected values of the model responses are shown. In spite of the verification of the hypothesis, the results of the linear regression are only moderately satisfactory due to the relatively large scatter of the measurements. Sensitivity analysis emphasizes this. For example, for data set number 5 from Table 1

$$\frac{\partial\boldsymbol{\Phi}_{\min}}{\boldsymbol{\Phi}_{\min}}\Big/\frac{\partial\bar{Y}_5}{\bar{Y}_5}=\frac{2\times20.28\times4\times1.231}{7.746}=25.66$$

A change in the measured value $\bar{Y}_5$ of 1 % causes a change of 25.55 % in $\boldsymbol{\Phi}_{\min}$. Similarly high sensitivities arise with

$$\frac{\partial b_0'}{b_0'} / \frac{\partial \bar{Y}_5}{\bar{Y}_5} = \frac{1}{14} \cdot \frac{4 \times 20.18}{15.002} = 0.384$$

and

$$\frac{\partial b_1}{b_1} / \frac{\partial \bar{Y}_5}{\bar{Y}_5} = \frac{1}{206.25} \cdot \frac{4 \times 20.18 \times 5}{1.282} = 1.526$$

so that the experimenter concerned with this task would be well advised to repeat the experiments with a more precise measurement technique. In fact other experiments for this example give the relationship

$$Nu = \frac{\xi/8\,(Re - 1000)\,Pr}{1 + 12.7\sqrt{\xi/8}\,(Pr^{2/3} - 1)} \left[1 + \left(\frac{D_R}{L}\right)^{2/3}\right]$$

where $\xi = (1.82 \log Re - 1.64)^{-2}$ [24], [25]: this relationship does not produce a linear dependency $Nu = f\,(Re)$ and hence a linear model appears to be inadequate.

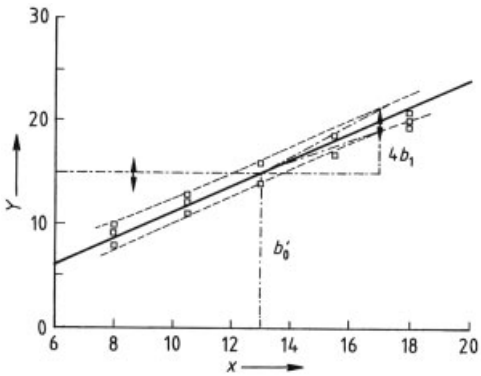

Figure 6. Results of linear regression for the example discussed in Section 3.1.1.4
Arrows indicate the confidence intervals of the parameters, square symbols denote measured values, and the dashed lines the confidence intervals of the expected values.

3.1.1.5. Concluding Remarks

The assumptions stated in Section 3.1.1 are not always appropriate; for example, the random error of the measurements may be a function of the independent variables. In addition, the random errors may be statistically dependent on each other and, in a given process, both the dependent and the independent variables can be statistical quantities. Estimation of the model parameters by linear regression in such cases does not differ in principle from that presented so far: more complex calculations may be required and further hypotheses or statistical rules may be needed. For discussion of these special cases, the reader is referred to the literature [4], [5], [7], [20].

Standardized computer programs are available for linear regression by least squares estimations for models with one dependent variable [26]. The solution of linear equations required for linear regression is also available through program libraries for numerical problems or for problems in linear algebra. Most of these libraries are readily available [26–32], so that special-purpose computer programs can easily be assembled for particular requirements.

3.1.2. Linear Empirical Models with Several Variables

In the case of linear empirical models with several independent variables, similar problems to those in Section 3.1.1 arise. Thus in the following, only the generalized notation will be discussed.

The linear relationship for a model with several independent variables is given by

$$y = \boldsymbol{\beta x} \tag{31}$$

and in extended notation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q \tag{32}$$

so that the problem is extended to q variables and $(q+1)$ parameters. The same assumptions hold as in Section 3.1.1, so that the model given by Equation (31) can be written in the form

$$\langle \bar{Y}_i \rangle = \bar{Y}_i - \varepsilon_i = \beta_0' + \sum_{j=1}^{q} \beta_j\,(x_{ij} - \bar{x}_j) \tag{33}$$

The $(q+1)$ model parameters are calculated from the measured values defined at the n data points in the same way as in Section 3.1.1.

In matrix notation Equation (33) reads

$$\langle \bar{\boldsymbol{Y}} \rangle = \bar{\boldsymbol{Y}} - \boldsymbol{E} = \boldsymbol{\beta x}' \tag{34}$$

with the abbreviations

$$\langle \bar{\boldsymbol{Y}} \rangle = \begin{pmatrix} \langle \bar{Y}_i \rangle \\ \langle \bar{Y}_2 \rangle \\ \vdots \\ \langle \bar{Y}_n \rangle \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0' \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix},$$

and

$$\boldsymbol{x}' = \begin{pmatrix} 1 & (x_{11} - \bar{x}_1) & \cdots & (x_{1q} - \bar{x}_q) \\ 1 & (x_{21} - \bar{x}_1) & \cdots & (x_{2q} - \bar{x}_q) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_{n1} - \bar{x}_1) & \cdots & (x_{nq} - \bar{x}_q) \end{pmatrix} \boldsymbol{p}$$

The first step is the same as for models with only one independent variable – the estimation of model parameters $\boldsymbol{\beta}$.

3.1.2.1. Parameter Estimation

As in Section 3.1.1.1, the estimates of the parameters $\boldsymbol{\beta}$ are found by minimizing the sum of squares of the deviations between the observed values $\bar{Y}_i$ and the expected values $\langle \bar{Y}_i \rangle$ (i.e., the sum of the variances):

$$\left(\bar{\boldsymbol{Y}} - \langle \bar{\boldsymbol{Y}} \rangle\right)^T \boldsymbol{p} \left(\bar{\boldsymbol{Y}} - \langle \bar{\boldsymbol{Y}} \rangle\right) \stackrel{!}{=} \mathrm{Min} \tag{35}$$

By using Equation (34), the sum of squares of the deviations can be formulated as

$$\boldsymbol{\Phi} = \left(\bar{\boldsymbol{Y}} - \boldsymbol{\beta}\boldsymbol{x}'\right)^T \boldsymbol{p} \left(\bar{\boldsymbol{Y}} - \boldsymbol{\beta}\boldsymbol{x}'\right) \tag{36}$$

which has to be differentiated with respect to all the elements of vector $\boldsymbol{\beta}$ and then set to zero. In Equation (36) $\boldsymbol{p}$ once again represents the number of repeated measurements or, in general terms, the statistical weights. In matrix notation

$$\boldsymbol{p} = \begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{pmatrix}$$

Differentiation yields

$$\frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \boldsymbol{\Phi}}{\partial \beta_0} \\ \frac{\partial \boldsymbol{\Phi}}{\partial \beta_1} \\ \vdots \\ \frac{\partial \boldsymbol{\Phi}}{\partial \beta_q} \end{pmatrix} = \frac{2\left[\partial \left(\bar{\boldsymbol{Y}} - \boldsymbol{\beta}\boldsymbol{x}'\right)^T\right] \boldsymbol{p} \left(\bar{\boldsymbol{Y}} - \boldsymbol{\beta}\boldsymbol{x}'\right)}{\partial \boldsymbol{\beta}} \tag{37}$$

so that for a minimum to occur

$$\left(\boldsymbol{x}'\right)^T \boldsymbol{p} \left(\bar{\boldsymbol{Y}} - \boldsymbol{\beta}\boldsymbol{x}'\right) = 0 \tag{38}$$

Further differentiation shows that Equation (38) actually represents a minimum. After rearrangement the normal equation

$$\left(\boldsymbol{x}'\right)^T \boldsymbol{p}\boldsymbol{x}'\boldsymbol{b} = \left(\boldsymbol{x}'\right)^T \boldsymbol{p}\bar{\boldsymbol{Y}} \tag{39}$$

is obtained.

In Equation (39) the parameters β_i are replaced by their estimated values. Equation (39) is the matrix equivalent of Equations (9 a, b). Solution of Equation (39) for the parameters gives

$$\boldsymbol{b} = \left[\left(\boldsymbol{x}'\right)^T \boldsymbol{p}\boldsymbol{x}'\right]^{-1} \left(\boldsymbol{x}'\right)^T \boldsymbol{p}\bar{\boldsymbol{Y}} \tag{40}$$

Inversion of the matrix $(\boldsymbol{x}')^{\mathrm{T}}\boldsymbol{p}\,\boldsymbol{x} \equiv \boldsymbol{a}$ is required for the solution. A condition for this is that the determinant is not equal to zero, or in other words that matrix $\boldsymbol{a}$ is nonsingular. Numerical problems related to ill-conditioned matrices $\boldsymbol{a}$ can be avoided by suitably scaling or normalizing the independent variables or by suitable experimental design. This can be illustrated by a simple example. From measurements at

$$\boldsymbol{x} = \begin{pmatrix} 9.5 \\ 10.0 \\ 10.5 \end{pmatrix},$$

a relationship of the form $y = \beta_0 + \beta_1 x$ will be investigated. Estimation of the model parameters involves inversion of the matrix

$$a \equiv \boldsymbol{x}^T \boldsymbol{p}\boldsymbol{x} = \begin{pmatrix} 3.0 & 30.0 \\ 30.0 & 300.5 \end{pmatrix}$$

As can be easily verified,

$$\begin{aligned} \mathrm{Det}\,(\boldsymbol{a}) &= 3.0 \times 300.5 - 30.0 \times 30.0 \\ &= 901.5 - 900.0 = 1.5 \end{aligned}$$

However, because of the spread of the measured values, the matrix is ill-conditioned for inversion since Det $(\boldsymbol{a}) = 0$ if only the first two digits are significant. If it is scaled instead and we use $y = \beta_0' + \beta_1(x - \bar{x})$ then

$$\boldsymbol{a} \equiv \left(\boldsymbol{x}'\right)^T \boldsymbol{p}\boldsymbol{x}' = \begin{pmatrix} 3.0 & 0 \\ 0 & 0.5 \end{pmatrix}$$

$$\text{or Det } (\boldsymbol{a}) = 3.0 \times 0.5 - 0 = 1.5$$

The result is the same. The problem of subtraction of large numbers which arises from the design of the experiments is thereby avoided. As can be seen from the numbers given above, the problem of handling differences between large numbers is aggravated if the empirical model is in the form of a polynomial, i.e.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

with $x_1 = x, \quad x_2 = x^2$ etc.

Numerical problems with the inversion of $\boldsymbol{a}$ are also avoided if the main diagonal of $\boldsymbol{a}$ can be occupied predominantly through proper normalization of the independent variables. This technique can be demonstrated using the simple example of measured values at

$$\boldsymbol{x} = \begin{pmatrix} 5.0 & 10.0 \\ 5.0 & 20.0 \\ 10.0 & 10.0 \\ 10.0 & 20.0 \end{pmatrix}$$

If the variables are normalized in the form

$$\tilde{x}_{i1} = \frac{x_{i1} - 7.5}{2.5}$$

and $\tilde{x}_{i2} = \dfrac{x_{i2} - 15.0}{5.0}$ then

$$\tilde{\boldsymbol{x}}' = \begin{pmatrix} 1.0 & -1.0 & -1.0 \\ 1.0 & -1.0 & 1.0 \\ 1.0 & 1.0 & -1.0 \\ 1.0 & 1.0 & 1.0 \end{pmatrix}$$

With $\boldsymbol{a} \equiv (\tilde{\boldsymbol{x}}')^{\mathrm{T}} \boldsymbol{p} \tilde{\boldsymbol{x}}'$ this leads to the occupation of the main diagonal only with

$$\boldsymbol{a} = \begin{pmatrix} 4.0 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 4.0 \end{pmatrix},$$

so that inversion does not cause any problems. The variables normalized in this form are orthogonal since $\sum x_0 x_{i1} = \sum x_0 x_{i2} = \sum x_{i1} x_{i2} = 0$. Ill-conditioning of $\boldsymbol{a}$ can therefore be avoided by planning the experiment in a manner which allows simple orthogonalization of the dependent variables.

A brief geometric explanation of the normal equations will now be given. Figure 7 shows the vector $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ in the space of measured values. Since the relationship $y = \beta_0 + \beta_1(x - \bar{x})$ is linear in its parameters, a plane surface for the calculated model solutions that depends on the estimated values b_0' and b_1 results.

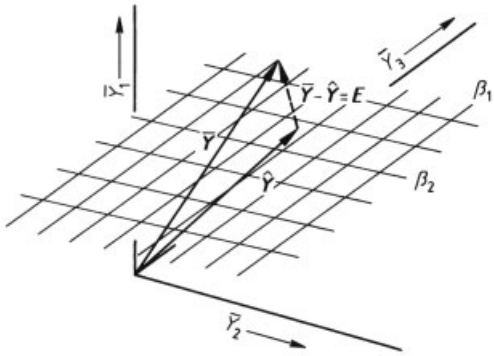

Figure 7. Geometrical illustration of the normal equations for a linear model

The set of estimated values b_0' and b_1 that gives the shortest distance between the end points of the vectors $\bar{\boldsymbol{Y}}$ and $\hat{\boldsymbol{Y}}$ now has to be found. Obviously the shortest distance $\bar{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}$ is given by the normal from the end of $\bar{\boldsymbol{Y}}$ onto the surface of the calculated model solution. The normal Equations (10) and (39) fix $\boldsymbol{b}$ such that the vector of the residuals $\bar{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}$ passes through the end point of $\bar{\boldsymbol{Y}}$ and is perpendicular to the surface of values calculated from the model.

3.1.2.2. Assessment of Estimated Parameter Values

The approach used to assess the estimated parameter values is the same as that in Section 3.1.1.2, and so will only be briefly discussed. To test the hypothesis $\langle \bar{\boldsymbol{Y}} \rangle = \beta \boldsymbol{x}'$, the mean square deviations from the regression line must be obtained. If $\bar{\boldsymbol{Y}} - \langle \bar{\boldsymbol{Y}} \rangle$ is represented by $\boldsymbol{E}$ then

$$s_{\mathrm{r}}^2 = \frac{(\boldsymbol{E}^T \boldsymbol{p} \boldsymbol{E})}{n - q - 1} \tag{41}$$

where s_{r}^2 is the sum of squares of the residuals divided by the number of degrees of freedom. The number of degrees of freedom is the number of data points minus the number of conditions constraining the minimization, in this case the number of parameters. The mean square deviations for the individual data points are calculated from Equation (19), so that the F-test can be conducted using the ratio $s_{\mathrm{r}}^2 / s_{\varepsilon}^2$. The variances of the

parameters and the model solutions can be calculated by means of the estimated values $s^2_{\bar{Y}_i}$ of $\sigma^2_{\bar{Y}_i}$ which are obtained from s^2_{r} and/or s^2_{ε} in the case of a positive F-test. The covariances of the parameters are

$$\mathrm{Cov}\,(\boldsymbol{b}) = \langle\, (\boldsymbol{b}-\boldsymbol{\beta})^T\,(\boldsymbol{b}-\boldsymbol{\beta})\,\rangle = \sigma^2_{\bar{Y}_i}\boldsymbol{a}^{-1} \tag{42}$$

If $\boldsymbol{a}^{-1}$ is denoted as $\boldsymbol{c}$, then the variances of the estimated parameter values can be obtained from the diagonal elements of $\boldsymbol{c}$ with estimated values $s^2_{\bar{Y}_i}$ for $\sigma^2_{\bar{Y}_i}$:

$$s^2_{b_k} = s^2_{\bar{Y}_i}\, c_{kk} \tag{43}$$

With

$$\begin{aligned}\mathrm{Var}\left(\hat{Y}_i\right) &= \mathrm{Var}\left(\boldsymbol{x}'\boldsymbol{b}\right) = \boldsymbol{x}'\,\mathrm{Var}\,(\boldsymbol{b})\,\left(\boldsymbol{x}'\right)^T \\ &= \sigma^2_{\bar{Y}_i}\,\boldsymbol{x}'\boldsymbol{c}\,\left(\boldsymbol{x}'\right)^T\end{aligned} \tag{44}$$

and $s^2_{\bar{Y}_i}$ for $\sigma^2_{\bar{Y}_i}$, the following is obtained:

$$s_{\hat{Y}_i} = s_{\bar{Y}_i}\sqrt{\boldsymbol{x}'\boldsymbol{c}\,(\boldsymbol{x}')^T} \tag{45}$$

so that, as in Section 3.1.1.2 (Eqs. 20, 22, and 24), the confidence intervals for $\boldsymbol{b}$ and $\hat{Y}_i$ are obtained.

More detailed methods for the assessment of the estimated parameter values for models with several independent variables are treated in [4], [5], [20].

3.1.2.3. Concluding Remarks

The sensitivity analysis for empirical models with several independent variables will not be discussed. The relationships presented in Section 3.1.1.3 can be used, and so need not be repeated. The basic rules of linear regression were illustrated using an example with one independent variable, a further example for a model with several independent variables will not be given. Cases which do not comply with the assumptions presented here are described in the literature; for example, processes with random errors which are not normally distributed or with random errors which depend on x are discussed in [4], [7]. Standardized computer programs are available for the treatment of empirical linear models with several independent variables or for multiple linear or polynomial regression [26], [33]. Programs can be easily constructed to fit users' requirements with the aid of modules from extensive program libraries for matrix and linear algebra [27–32].

3.2. Nonlinear Empirical Models

In nonlinear models the relationship from Equation (3) cannot be represented by Equation (4). The derivatives of the model equation with respect to the parameters are themselves functions of the parameters. The nonlinear model $\boldsymbol{y} = f\,(\boldsymbol{x}, \boldsymbol{\beta})$ is once again based on several assumptions:

1) The dependent variable is a random variable for which n data sets of measured values $\bar{Y}_i$ are available. The model can therefore be represented in the form

$$\langle \bar{\boldsymbol{Y}} \rangle = \bar{\boldsymbol{Y}} - \boldsymbol{E} = f\,(\boldsymbol{x}, \boldsymbol{\beta}) \tag{46}$$

   where $\boldsymbol{E}$ once again represents the random error in $\bar{Y}_i$, so that $\bar{\boldsymbol{Y}} - \langle \bar{\boldsymbol{Y}} \rangle = \boldsymbol{E}$. The expected value of the random error is zero and its variance is constant (independent of x).
2) The random variable Y has a normal distribution about $\bar{Y}_i$ and has the same variance as ε_i.
3) Both the random variable and the random error are statistically independent.
4) The variables $\boldsymbol{x}$ are not random variables.

The empirical nonlinear statistical model can be represented analogously to that in Figure 5, whereby the relationship is nonlinear and is characterized by several independent variables. As with the linear model, the first task consists of estimating the model parameters using the experimental process data.

3.2.1. Parameter Estimation (Nonlinear Regression)

Estimation of the model parameters $\boldsymbol{\beta}$ involves the now familiar task of defining the estimated values $\boldsymbol{b}$ such that the sum of squares of the deviations between the values $\bar{Y}_i$ observed at the n data points and the estimated values $\langle \bar{Y}_i \rangle$ is a minimum. [In the following, the equations are expressed in a simplified form, without repeated

measurements, i.e., with unit statistical weights $\boldsymbol{p}$, Det $(\boldsymbol{p}) = 1$]:

$$\left(\bar{\boldsymbol{Y}} - \langle\bar{\boldsymbol{Y}}\rangle\right)^T \left(\bar{\boldsymbol{Y}} - \langle\bar{\boldsymbol{Y}}\rangle\right) \stackrel{!}{=} \text{Min} \tag{47}$$

A geometric representation of this task is shown in Figure 8. The end point of the vector $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ in the space of measured values should be linked by the shortest possible distance with the end point of the vector $\hat{\boldsymbol{Y}}$. This is once again given by the normal to the surface of model responses in the parameter space, which in this case is curved because of the nonlinearity of the relationship $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\beta})$.

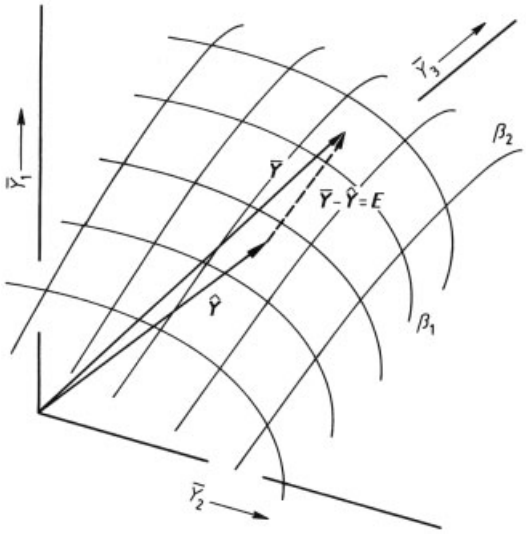

Figure 8. Geometrical illustration of the normal equation for a nonlinear model

Partial differentiation of the sum of squares of the deviations

$$\boldsymbol{\Phi} = \left[\bar{\boldsymbol{Y}} - f(x, \boldsymbol{\beta})\right]^T \left[\bar{\boldsymbol{Y}} - f(x, \boldsymbol{\beta})\right] \tag{48}$$

with respect to the parameters and then setting the derivatives to zero produces a system of nonlinear normal equations. This is the main difference from linear regression in which a system of linear equations in $\boldsymbol{\beta}$ is obtained for the normal equations. This can be easily verified by analyzing the sums of squares (Eqs. 9, 36, and 48).

Nonlinear regression is therefore the solution of a system of nonlinear equations for $\boldsymbol{\beta}$.

3.2.1.1. Transformation of the Model into Linear Form

The simplest method of estimating parameters in nonlinear regression is transformation of the model equations into linear form to avoid the problem of solving the nonlinear equations described above. Certain types of nonlinear models can be transformed into linear form but this may involve several problems; this will now be briefly illustrated.

The relationship

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \tag{49}$$

e.g., the rate of a complex chemical reaction, can be transformed into linear form by taking logarithms:

$$\log y = \log\beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 \tag{50}$$

The transformed model is therefore

$$\log\bar{Y}_i - \varepsilon_i = \log\beta_0 + \beta_1 \log x_{i1} + \beta_2 \log x_{i2} \tag{51}$$

The previously presented methods of regression analysis for testing hypotheses and calculating the confidence intervals of the parameters or model solutions presume that the additive errors ε_i and random variables have normal distributions. For the nonlinear model of Equation (49), the additive error must be multiplicative:

$$\bar{Y}_i = \beta_0 x_{i1}^{\beta_1} x_{i2}^{\beta_2} \varepsilon_i' \tag{52}$$

and, like the random variable Y_i, must have a log-normal distribution to satisfy the above-mentioned presumptions. This is not always the case and must be experimentally verified.

Another example is the relationship of the form

$$y = \frac{\beta_1 \beta_2 x}{(1 + \beta_2 x)^2} \tag{53}$$

which represents, for example, the rate of a heterogeneous catalytic reaction (Hougen – Watson – Langmuir – Hinshelwood kinetics). Transformation into a linear model gives

$$\left(\frac{x}{y}\right)^{1/2} = \frac{1}{(\beta_1\beta_2)^{1/2}} + \frac{\beta_2}{(\beta_1\beta_2)^{1/2}} x \tag{54}$$

The additive statistical error ε_i in the linear model for the estimation of the parameters

(Eq. 54) is not identical to the random error of the nonlinear model (Eq. 53). Inverse transformation of Equation (54) gives:

$$y = \frac{\beta_1\beta_2 x}{(1+\beta_2 x)^2} + \frac{\beta_1\beta_2}{(1+\beta_2 x)^2}\left(y\varepsilon'^2 - 2\sqrt{yx}\varepsilon'\right) \qquad (55)$$

which contains an additive random error that is dependent on x, y, β_1, and β_2.

3.2.1.2. Direct Search Methods

Another way of circumventing the problem of solving nonlinear equations is to use direct search methods. For this the normal equations are not derived from the sum of squares of the deviations (Eq. 48), instead the minimum of the function $\boldsymbol{\Phi}$ is sought directly. Vectorial search methods [34], Simplex methods [35], and Simplex methods with variable Simplex geometry [36] are the most commonly used direct search methods. The structure of these methods is presented schematically for a model with two parameters in Figure 9.

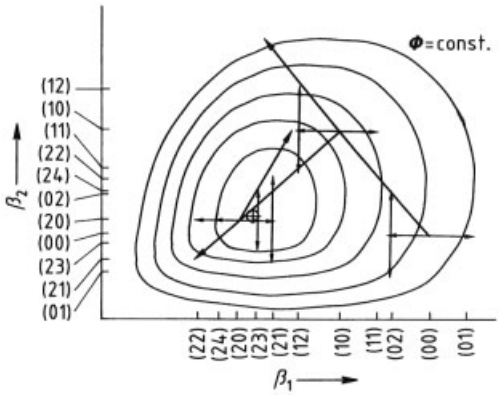

Figure 9. Representation of direct search methods
The first digits in the parentheses indicate the number of iteration. The second digits denote the number of successive variation of β_i. For example, (02) on the β_1 axis means the initial guess of β_1 and the second variation

Vectorial search methods start from estimated values $b_i^{(0)}$ for the parameters β_i. The direction of the search vector is established from successive calculations of $\boldsymbol{\Phi}$ by varying the estimated values b_i. The direction of the vector leading to a reduction in $\boldsymbol{\Phi}$ is used to calculate improved estimates of b_i. If there is no further reduction in $\boldsymbol{\Phi}$ along the search vector, the direction is redefined and the minimum of $\boldsymbol{\Phi}$ is located by reducing the step size (Fig. 9).

Simplex methods are more effective, especially with large numbers of parameters. For a twoparameter problem (Fig. 10), a regular Simplex is an equilateral triangle which, starting from estimated values $b_i^{(0)}$, is placed at a point on the surface of $\boldsymbol{\Phi}$ in the parameter space. During the search, $\boldsymbol{\Phi}$ is evaluated at each of the three vertices, and the triangle is reflected about the side opposite the largest value of $\boldsymbol{\Phi}$. If there is no further reduction in $\boldsymbol{\Phi}$ after reflection, then the Simplex is reduced in size to allow more accurate localization of the minimum.

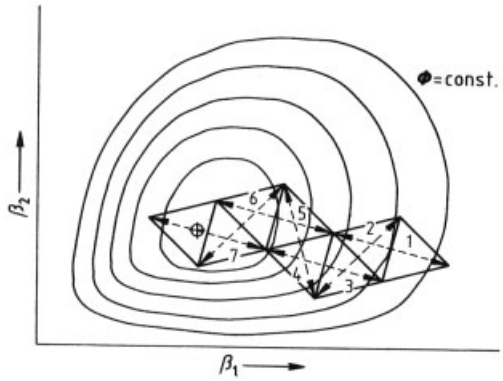

Figure 10. Representation of the Simplex method
The numbers represent the number of reflection

Simplexes with variable geometry are obtained by expanding after successful reflection or contracting after unsuccessful reflection. An example is given in Figure 11 for a twoparameter problem. The reflection algorithm is

$$b_{i\mathrm{r}} = (1+\gamma_\mathrm{r})\, b_{i\mathrm{c}} - \gamma_\mathrm{r} b_{i\max}$$

where $b_{i\mathrm{r}}$, $b_{i\,\max}$, and $b_{i\mathrm{c}}$ are the coordinates of the reflected Simplex point, of the Simplex point with the largest sum of squares of deviations, and of the midpoint of the side lying opposite it, respectively; γ_r is the reflection factor. The expansion algorithm is

$$b_{i\mathrm{e}} = \gamma_\mathrm{e} b_{i\mathrm{r}} + (1-\gamma_\mathrm{e})\, b_{i\mathrm{c}}$$

where $b_{i\mathrm{e}}$ are the coordinates of the expanded Simplex point and γ_e is the expansion factor. After unsuccessful expansion or reflection, contraction is achieved with the algorithm

$$b_{i\mathrm{k}} = (1-\gamma_\mathrm{k})\, b_{i\mathrm{c}} + \gamma_\mathrm{k} b_{i\max}$$

where b_{ik} and γ_k have analogous meanings to those in reflection or expansion.

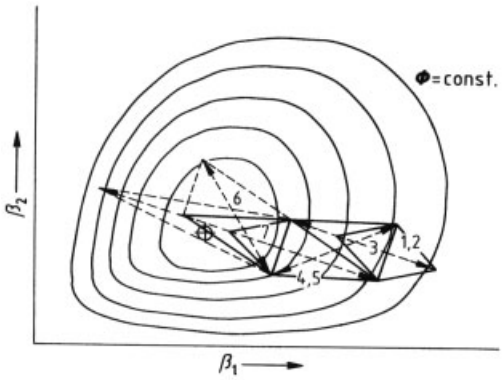

Figure 11. Representation of the Simplex method with variable geometry
The numbers represent the number of reflection, expansion, or contraction

3.2.1.3. Gradient Methods

Discussion of the methods for solving nonlinear systems of equations arising from the minimum conditions for $\boldsymbol{\Phi}$ will be limited to two methods. Further methods are described in the literature on nonlinear optimization [5], [7], [20], [37–39] (see also → Mathematics in Chemical Engineering, Chap. 10.).

Through use of the abbreviation

$$\langle \bar{\boldsymbol{Y}} \rangle \equiv \boldsymbol{\eta} = f(\boldsymbol{x}, \boldsymbol{\beta}) \tag{56}$$

the sum of the squares of the deviations (Eq. 48) can be written as

$$\boldsymbol{\Phi} = \left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}\right)^T \left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}\right) \tag{57}$$

The condition for the minimum with respect to the parameters $\boldsymbol{\beta}$ then becomes:

$$\frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial \boldsymbol{\Phi}}{\partial \beta_1} \\ \frac{\partial \boldsymbol{\Phi}}{\partial \beta_2} \\ \vdots \\ \frac{\partial \boldsymbol{\Phi}}{\partial \beta_q} \end{pmatrix} = \frac{2\left[\partial \left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}\right)^T\right]\left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}\right)}{\partial \boldsymbol{\beta}} = 0 \tag{58}$$

The first method discussed here is based on the Newton – Raphson method for the solution of nonlinear systems of equations. Here the nonlinear Equation (58) is expanded around an estimated solution into a Taylor series, which is truncated after the first term. For Equation (58) this leads to the expansion of $\boldsymbol{\eta}$ about $\boldsymbol{\eta}^{(0)}$ at $\boldsymbol{b}^{(0)}$, hence

$$\boldsymbol{\eta} = \boldsymbol{\eta}^{(0)} + \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}\right)^{(0)} \Delta \boldsymbol{b}^{(0)} \tag{59}$$

with $\Delta \boldsymbol{b}^{(0)} = \boldsymbol{\beta} - \boldsymbol{b}^{(0)}$. A system of linear equations for the new variable $\Delta \boldsymbol{b}^{(0)}$ then arises, solution of which leads to an improved estimated value for the parameters:

$$\boldsymbol{b}^{(1)} = \boldsymbol{b}^{(0)} + \Delta \boldsymbol{b}^{(0)} \tag{60}$$

If this procedure is applied to Equation (58), substitution of Equation (59) into Equation (58) gives

$$\frac{2\left[\partial\left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}^{(0)} - \left(\frac{\partial \boldsymbol{\eta}}{\boldsymbol{\beta}}\right)^{(0)} \Delta \boldsymbol{b}^{(0)}\right)^T\right]\left[\left(\frac{\partial \boldsymbol{\eta}}{\boldsymbol{\beta}}\right)^{(0)} \Delta \boldsymbol{b}^{(0)}\right]}{\partial \Delta \boldsymbol{b}^{(0)}} = 0 \tag{61}$$

This gives the following system of equations for $\Delta \boldsymbol{b}^{(0)}$:

$$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{(0)} \boldsymbol{B}^{(0)} = \left(\boldsymbol{X}^T \boldsymbol{E}\right)^{(0)} \tag{62}$$

which employs the abbreviations

$$\boldsymbol{X} \equiv \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}}, \quad \boldsymbol{E} \equiv \left(\bar{\boldsymbol{Y}} - \boldsymbol{\eta}\right), \text{ and } \boldsymbol{B} \equiv \Delta \boldsymbol{b}$$

Equation (62) is similar to Equation (39). In Equation (62) $\boldsymbol{B}^{(0)}$ is the dependent variable, whereas Equation (39) represents a system of equations for the parameters themselves. (Since there are no repeated measurements according to the assumptions in Section 3.2.1 the statistical weights for all measured points are equal, $\boldsymbol{p}$ is replaced by the unit matrix). The solution of Equation (62) for $\boldsymbol{B}^{(0)}$ gives

$$\boldsymbol{B}^{(0)} = \left[\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{(0)}\right]^{-1} \left(\boldsymbol{X}^T \boldsymbol{E}\right)^{(0)} \tag{63}$$

By using Equation (60) an improved set of parameters is obtained and the method is iterated until a previously defined convergence condition of $\boldsymbol{B}^{(n)}$ or $\boldsymbol{\Phi}$ is satisfied.

With the second method for the estimation of parameters from nonlinear empirical models, the sum of squares itself is linearized. Expansion of Equation (48) around $\boldsymbol{b}^{(0)}$ into a Taylor series truncated after the first term gives

$$\boldsymbol{\Phi} \approx (\boldsymbol{\Phi})^{(0)} + \sum_{j=1}^{q} \left(\frac{\partial \boldsymbol{\Phi}}{\beta_j}\right)^{(0)} \left(\beta_j - b_j^{(0)}\right) \tag{64}$$

The direction of search used to calculate improved estimated values from the relation

$$\boldsymbol{b}^{(n+1)} = \boldsymbol{b}^{(n)} + \boldsymbol{h}^T \Delta \boldsymbol{b}^{(n)} \tag{65}$$

($\boldsymbol{h}$ is a vector of the step size) can be defined by the gradient of $\boldsymbol{\Phi}$; grad $\boldsymbol{\Phi}$ is a vector that is normal to the surface of $\boldsymbol{\Phi}$ in the parameter space and indicates the direction of the steepest ascent in $\boldsymbol{\Phi}$ at the point $\boldsymbol{b}$. Conversely, the vector $-$ grad $\boldsymbol{\Phi}$ shows the direction of the steepest descent of $\boldsymbol{\Phi}$. For parameter estimation the components of the vectors at $\boldsymbol{b}^{(0)}$

$$-\mathrm{grad}\boldsymbol{\Phi} = -\left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_1}\right)^{(0)} \delta\beta_1 - \left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_2}\right)^{(0)} \delta\beta_2 - \ldots - \left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_q}\right)^{(0)} \delta\beta_q \tag{66}$$

are calculated. For this the vector $-$grad $\boldsymbol{\Phi}$, normalized by its absolute value is used:

$$\frac{-\mathrm{grad}\boldsymbol{\Phi}}{\| -\mathrm{grad}\boldsymbol{\Phi}\|} = \frac{-\left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_1}\right)^{(0)} \delta\beta_1 - \left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_2}\right)^{(0)} \delta\beta_2 - \ldots - \left(\frac{\partial \boldsymbol{\Phi}}{\partial \beta_q}\right)^{(0)} \delta\beta_q}{\sqrt{\left(-\frac{\partial \boldsymbol{\Phi}}{\partial \beta_1}\right)^{(0)2} + \left(-\frac{\partial \boldsymbol{\Phi}}{\partial \beta_2}\right)^{(0)2} + \ldots + \left(-\frac{\partial \boldsymbol{\Phi}}{\partial \beta_q}\right)^{(0)2}}} \tag{67}$$

In Equations (66) and (67), $\delta\beta_i$ denotes the unit vectors in the direction of β_i. The components of the normalized vector (Eq. 67) produce the term $\Delta\boldsymbol{b}$ in Equation (64) from which the iteration formula (Eq. 65) can proceed.

Both methods (solution of the system of nonlinear normal equations and the method of steepest descent) can, under certain conditions of the sum of squares $\boldsymbol{\Phi}$, lead to unfavorable trajectories over the surface of $\boldsymbol{\Phi}$ in the parameter space [40]. To deal with this, according to MARQUARDT a diagonal matrix is added to the coefficient matrix ($\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$) in Equation (62):

$$\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)\boldsymbol{B} = \boldsymbol{X}^T\boldsymbol{E} \quad \lambda \geq 0 \tag{68}$$

where $\boldsymbol{I}$ is the unit matrix and λ is a scaling factor. When $\lambda = 0$ Equation (62) is obtained. As $\lambda \to \infty$, then $\boldsymbol{B} \approx \boldsymbol{X}^{\mathrm{T}}\boldsymbol{E}/\lambda$ which, with the identity $\lambda = \frac{\|-\mathrm{grad}\boldsymbol{\Phi}^{(n)}\|}{\boldsymbol{h}^{(n)}}$ leads to the method of steepest descent for calculating $\boldsymbol{B}$. By fitting λ using Marquardt's method, the essential characteristics of both methods can be combined, [4], [7], [40].

Numerical Problems Connected with Gradient Methods. Iterative solution of Equation (62) or the method of steepest descent can give rise to problems, one of them being the initial guesses $\boldsymbol{b}^{(0)}$. Under certain conditions, the selection of $\boldsymbol{b}^{(0)}$ leads to a neighboring minimum in $\boldsymbol{\Phi}$, or is so unfavorable that iteration does not lead to convergence. The first of these problems requires a more precise investigation of the sum of the squares of the deviations or repetition of the calculation with various initial values for $\boldsymbol{b}^{(0)}$. In order to avoid divergence of the iterative method (especially with poorly chosen $\boldsymbol{b}^{(0)}$), only the direction of the vector $\boldsymbol{B}^{(0)}$ should be used; new estimated values for $\boldsymbol{\beta}$ are then calculated using Equation (65) with a reduced step size in place of Equation (60). In Equation (65), $\boldsymbol{h}$ has the character of a user-selected damping factor, which can be adjusted during each iteration in accordance with the evolution of $\boldsymbol{\Phi}$. With favorable conditioning of the nonlinear equation system, this factor can also be used to accelerate convergence.

Another problem lies in the mathematical form of the model. For example, for a model consistent with Equation (53) it is easy to show that $\boldsymbol{\eta}$, as well as the partial derivatives $\partial\boldsymbol{\eta}/\partial\beta_1$ and $\partial\boldsymbol{\eta}/\partial\beta_2$, are unbounded if $\beta_2 x = -1$. The only way to circumvent this problem is to constrain the range of the estimated values of the parameters on the basis of the physicochemical characteristics of the process.

The sum of the squares of the deviations can contain terms of differing orders of magnitude, which lead to very different gradients in the direction of the individual parameters. In such cases scaling of the parameters is always advisable, whereby the individual components in Equation (66) are brought to the same order of magnitude by linear transformation.

Transformations of the independent variables are necessary if interactions exist between indi-

vidual parameters. This is the case, for example, with a simple Arrhenius type of rate coefficient of a chemical reaction

$$k = k_0 \mathrm{e}^{-E_\mathrm{a}/RT}$$

where E_a is the activation energy, R the universal gas constant, and T the temperature.

If the exponential function is expanded so that

$$k \approx k_0 - k_0 \frac{E_\mathrm{a}}{RT} + \frac{1}{2} k_0 \left(\frac{E_\mathrm{a}}{RT}\right)^2 - \cdots$$

it becomes clear that k_0 has a multiplicative effect on the other parameter E_a. Minima of the sum of squares can thus be found for various combinations of k_0 and E_a that do not make sense in physical terms. In such cases either the range of allowable estimated values of the parameters is restricted on the basis of the physicochemical process characteristics, or the matrix $(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})$ is brought into a form in which the diagonal elements dominate through transformation of the independent variables. The smaller the off-diagonal elements of the matrix $(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})$ are compared to the diagonal elements, the smaller the interaction between the parameters becomes. This can be demonstrated for the above example of the simple Arrhenius type of rate coefficient in which T is the independent variable. Since

$$\frac{\partial k}{\partial k_0} = \mathrm{e}^{-\frac{E_\mathrm{a}}{RT}} = \frac{k}{k_0} \text{ and } \frac{\partial k}{\partial (E_\mathrm{a}/R)} = -\frac{k_0}{T}\mathrm{e}^{\frac{-E_\mathrm{a}}{RT}} = \frac{k}{T}$$

$$\text{then} \quad \left(\boldsymbol{X}^T\boldsymbol{X}\right) = \begin{pmatrix} \frac{1}{k_0^2}\sum i k_i^2 & -\frac{1}{k_0}\sum i \frac{k_i^2}{T_i} \\ -\frac{1}{k_0}\sum i \frac{k_i^2}{T_i} & \sum i \left(\frac{k_i^2}{T_i}\right)^2 \end{pmatrix}$$

Therefore

$$\mathrm{Det}\left(\boldsymbol{X}^T\boldsymbol{X}\right) = \frac{1}{k_0^2}\left[\sum_i k_i^2 \sum_i \left(\frac{k_i^2}{T_i}\right)^2 - \left(\sum_i \left(\frac{k_i^2}{T_i}\right)\right)^2\right]$$

which can become singular under certain conditions, since all terms in the square brackets are positive for all regions of the measured k_i and T_i.

If the transformation $T^* = (T - \bar{T})/T$ is used, we obtain $k = k_0^* \mathrm{e}^{E^*T^*}$ with $k_0^* = k_0/\mathrm{e}^{E^*}$ and $E^* = E_a/R\bar{T}$. For the transformed model

$$\frac{\partial k}{\partial k_0^*} = \mathrm{e}^{E^*T^*} = \frac{k}{k_0^*}$$

and

$$\frac{\partial k}{\partial E^*} = -k_0^* T^* \mathrm{e}^{E^*T^*} = kT^*$$

Hence

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right) = \begin{pmatrix} \frac{1}{k_0^{*2}}\sum i k_i^2 & -\frac{1}{k_0^*}\sum i k_i^2 T_i^* \\ -\frac{1}{k_0^*}\sum i k_i^2 T_i^* & \sum i \left(k_i T_i^*\right)^2 \end{pmatrix}$$

Since T_i* can be positive or negative, the diagonal elements of the matrix dominate, and in

$$\mathrm{Det}\left(\boldsymbol{X}^T\boldsymbol{X}\right) = \frac{1}{k_0^{*2}}\left[\sum_i k_i^2 \sum_i \left(k_i T_i^*\right)^2 - \left(\sum_i k_i^2 T_i^*\right)^2\right]$$

the first term in the square brackets always dominates. The danger of the matrix $(\boldsymbol{X}^\mathrm{T}\boldsymbol{X})$ being singular (which would prevent solution of Eq. 63) is thus overcome.

3.2.2. Assessment of Estimated Parameter Values

Several methods have been developed to assess the estimated parameter values from nonlinear empirical models [20], [41], [42]. A method which is similar to the approach used in Sections 3.1.1.2 and 3.1.2.2 will be presented here.

The nonlinear model is linearized around the estimated values $\boldsymbol{b}$ in the parameter space, i.e., is expanded into a Taylor series truncated after the first term. The approximate variances and covariances of the parameters are then given by (cf. Eqs. 42 and 59):

$$\mathrm{Cov}\,(\boldsymbol{b}) \approx \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \sigma_{\bar{Y}_i}^2 \tag{69}$$

Each element of $\boldsymbol{X}$ is calculated either analytically or numerically from the value of $\boldsymbol{b}$ indicated by the minimum. If the hypothesis is valid, s_r^2 can once again be used as an estimated value for $\sigma_{\bar{Y}_i}^2$ (Eq. 41). If repeated measurements are available, the hypothesis $\langle \bar{Y}_i \rangle = f(x, \beta)$ can be tested beforehand. The approximate variances

of the estimated values of the parameters are obtained from the diagonal elements of $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$:

$$s_{bk}^2 \approx s_{\hat{Y}_i}^2 \cdot c_{kk} \tag{70}$$

using the abbreviation $\boldsymbol{c} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. The approximate confidence intervals can be estimated from this as described in Section 3.1.1.2.

The variances of the model solutions are obtained from the linearized model according to

$$\begin{aligned} s_{\hat{Y}_i}^2 &\approx \sum_{j=1}^{q} \left(\frac{\partial \hat{Y}}{\partial b_j}\right) s_{b_j}^2 \\ &+ \sum_{i=1}^{q} \sum_{j=1, i\neq j}^{q} \left(\frac{\partial \hat{Y}}{\partial b_i}\right) \left(\frac{\partial \hat{Y}}{\partial b_j}\right) \mathrm{Cov}\,(b_i b_j) \\ &\approx s_{\hat{Y}_i} \sum_{i=1}^{q} \sum_{j=1, i\neq j}^{q} \left(\frac{\partial \hat{Y}}{\partial b_i}\right) \left(\frac{\partial \hat{Y}}{\partial b_j}\right) c_{ij} \end{aligned} \tag{71}$$

so that the approximate confidence intervals of the expected values can also be estimated.

3.2.3. Example of a Nonlinear Model

The estimation of the parameters of a nonlinear empirical model will be demonstrated by further discussion of the process presented in Section 3.1.1.4. Because of the relatively large scatter of the measured values within the individual data sets, the result of the linear regression is merely tolerable in spite of a positive test of the hypothesis $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1(x_i - \bar{x})$. The assessment of the estimated parameters gives large individual confidence intervals for the parameters and the expected values. Furthermore, the sum of squares and the estimated parameter values show a high sensitivity to the measured values $\bar{Y}_i$.

The same process will now be tested with a second approximation using a model of the form

$$\langle \bar{Y}_i \rangle = \beta_0 x_i^{\beta_1} \tag{72}$$

The data for nonlinear regression are given in Table 1. The Newton – Raphson method is used to solve the system of nonlinear equations (Eq. 58). The results from the application of the linear model $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1(x_i - \bar{x})$ are used as an initial guess for b_0 and b_1, hence $b_0^{(0)} = 1.282$, $b_1^{(0)} = 1$. The step size $h_i^{(n)}$ (Eq. 65) is adjusted automatically according to the improvement in $\boldsymbol{\Phi}$ with each iteration. Figure 12 shows the estimated values $b_0^{(n)}$ and $b_1^{(n)}$ and the sum of squares of the deviations $\boldsymbol{\Phi}^{(n)}$ plotted against the number of iterations n. The method is regarded as convergent when $\boldsymbol{\Phi}^{(n-1)} - \boldsymbol{\Phi}^{(n)} \leq 10^{-4}$.

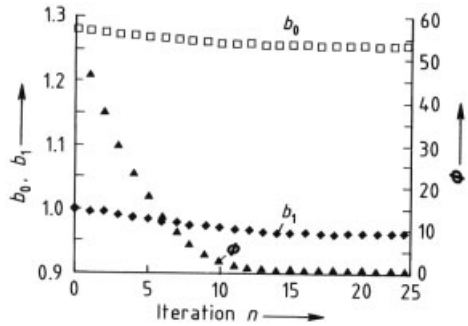

Figure 12. Results of nonlinear regression for the example discussed in Section 3.2.3

For the region of the experimentally adjusted gas velocities, the calculated results $b_0 = 1.257$ and $b_1 = 0.9626$ produce a smaller minimum in $\boldsymbol{\Phi}$, $\boldsymbol{\Phi}_{\min} = 0.332$, than the linear regression with $\boldsymbol{\Phi}_{\min} = 7.746$ (Fig. 6). The improved quality of the regression is obvious from the value of the mean square deviations about the regression function ($s_{\mathrm{r}}^2 = 0.1105$) so that the F-test is fulfilled to a higher level of significance. If the combined squared deviations s_{r}^2 and s_{ε}^2 are used to estimate $\sigma_{\hat{Y}_i}^2$, $s_{\hat{Y}_i}^2 = 0.996$ is obtained; the improvement seen in Figure 12 also results in a smaller variance in the estimated parameter values b_0 and b_1 and in the model solution $\bar{Y}_i$. From Equations (70) and (71) the model linearized in the minimum of $\boldsymbol{\Phi}$ gives

$$s_{b_0}^2 \approx 0.996 \times 5.21 \times 10^{-2} = 5.03 \times 10^{-2}$$

$$s_{b_1}^2 \approx 0.996 \times 4.51 \times 10^{-3} = 4.3 \times 10^{-3}$$

$$s_{Y_i}^2 \approx 0.996 \left[\left(x_{\mathrm{i}}^{0.9626}\right)^2 \left(5.21 \times 10^{-2} - 3.6 \ln x_i + 7.13 \times 10^{-3} (\ln x_i)^2\right)\right]$$

The relation $\langle \bar{Y}_i \rangle = \beta_0 x_i^{\beta_1}$ with $b_1 \leq 1$ reflects the true dependency better (see Section 3.1.1.4).

3.2.4. Concluding Remarks

The treatment of nonlinear empirical models is conducted as in Sections 3.1.1 and 3.1.2 using a relatively simple model to obtain the simplest formulation possible. The principles of nonlinear regression and some of the resulting problems can be clarified using this simple model. For cases in which the assumptions listed in the introduction to Section 3.2 do not hold, more advanced literature should be consulted, [4], [5], [7], [20]. Many computer programs are available for routine problems of parameter estimation with nonlinear empirical models (e.g., for gradient methods [27], for Marquardt's method [43], or for direct search methods [44]). The individual steps of the Newton – Raphson method (see Sections 3.1.1.5 and 3.1.2.3) can be obtained as modules from readily available program libraries [26–32].

3.3. Further Calculation Methods and Model Types

The discussion on the quantitive definition of empirical models (Sections 3.1 and 3.2) referred to the minimization of the sum of squares of the deviations as well as to simple models based on Equations (5) and (46). The section on empirical models will now be concluded with a short discussion of other methods of parameter estimation for empirical models and other types of models.

3.3.1. Further Methods of Parameter Estimation

Further methods of defining the parameters of empirical models include the maximum likelihood method.

A statistical empirical model can be described with the use of the probability density function $P(x, y)$ (Section 2.2 and Fig. 5). If a model for the process can be represented in the form $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1(x_i - \bar{x})$ (Section 3.1.1), the probability density function is defined by the parameters β_0', β_1, and $\sigma^2_{\bar{Y}_i}$, hence $P(\bar{Y}, x; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i})$. The problem of estimating the parameters now consists of defining β_0', β_1, and $\sigma^2_{\bar{Y}_i}$ such that the likelihood function becomes a maximum. With one observation the likelihood function L for the parameters is the probability density function in which the variables $\bar{Y}$ and x are viewed as parameters and the parameters β_0', β_1, and $\sigma^2_{\bar{Y}_i}$ are viewed as variables.

$$L\left(\beta_0', \beta_1, \sigma^2_{\bar{Y}_i}; \bar{Y}_1, x_1\right) = P\left(\bar{Y}_1, x_1; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right) \quad (73)$$

With n statistically independent observations the likelihood function is the product of the individual functions:

$$\begin{aligned} &L\left(\beta_0', \beta_1, \sigma^2_{\bar{Y}_i}; \bar{Y}_1, x_1 \ldots \bar{Y}_n, \bar{x}_n\right) \\ &= \prod_{i=1}^{n} L\left(\beta_0', \beta_1, \sigma^2_{\bar{Y}_i}; \bar{Y}_i, x_i\right) \\ &= P\left(\bar{Y}_1, x_1; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right) \cdot P\left(\bar{Y}_2, x_2; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right) \\ &\quad \ldots P\left(\bar{Y}_n, x_n; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right) \end{aligned} \quad (74)$$

To define β_0', β_1, and $\sigma^2_{\bar{Y}_i}$, L must now be maximized. The logarithms of Equation (74) are taken giving the following relations for the parameters:

$$\frac{\partial \ln L}{\partial \beta_0'} = \frac{\partial \sum_{i=1}^{n} P\left(\bar{Y}_i, x_i; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right)}{\partial \beta_0'} = 0 \quad (75a)$$

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{\partial \sum_{i=1}^{n} P\left(\bar{Y}_i, x_i; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right)}{\partial \beta_1} = 0 \quad (75b)$$

$$\frac{\partial \ln L}{\partial \sigma^2_{\bar{Y}_i}} = \frac{\partial \sum_{i=1}^{n} P\left(\bar{Y}_i, x_i; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right)}{\partial \sigma^2_{\bar{Y}_i}} = 0 \quad (75c)$$

For the linear problem $\langle \bar{Y}_i \rangle = \beta_0' + \beta_1(x_i - \bar{x})$ where the random variable Y is assumed to have a normal distribution, it can easily be shown that Equations (9) for defining β_0' and β are obtained from Equation (75) since

$$\begin{aligned} &P\left(\bar{Y}_i, x_i; \beta_0', \beta_1, \sigma^2_{\bar{Y}_i}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2_{\bar{Y}_i}}} \cdot \exp\left[-\frac{1}{2\sigma^2_{\bar{Y}_i}} \left(\bar{Y}_i - \langle \bar{Y}_i \rangle\right)^2 p_i\right] \end{aligned} \quad (76)$$

Detailed discussion of this method can be found in [4], [22].

Further search algorithms are discussed in more advanced literature on optimization, → Mathematics in Chemical Engineering, Chap. 10., [5], [7], [38], [39].

3.3.2. Other Types of Models

In addition to the models described by Equations (5) and (46), empirical models in which underlying conditions constrain the parameters are frequently encountered in industrial chemistry. Furthermore the functional relationship may not be in the simple form of algebraic equations (Eqs. 5, 32, and 46) but in the form of differential equations or transfer functions.

3.3.2.1. Models with Constraints on the Parameters

Constraints on parameters are found in many models used in industrial chemistry. A simple example is the reaction rate of a heterogeneously catalyzed reaction according to Equation (53), which only makes sense when

$$\beta_i \geq 0 \tag{77a}$$

Equally common are constraints of the type

$$\beta_i \geq k_i, \text{ where } k_i \text{ is positive} \tag{77b}$$

$$0 \leq \beta_i \leq 1 \tag{77c}$$

$$k_i^* \leq \beta_i \leq k_i \tag{77d}$$

Explicit constraints of this type can easily be removed with appropriate transformations. For example, the general transformation rule for Equation (77 b) is

$$\beta_i = k_i + e^{\beta_i^*} \tag{78a}$$

in which the unknown β_i^* is not affected by any constraints. For Equation (77 d) the general transformation rule

$$\beta_i = k_i^* + (k_i - k_i^*) \sin^2 \beta_i^* \tag{78b}$$

can be used, whereas for Equation (77 a)

$$\beta_i = \beta_i^{*2} \text{ or } \beta_i = e^{\beta_i^*} \tag{78c,d}$$

are suitable transformations.

The general constraint

$$g_i(\boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{\beta}) \geq 0, \quad i = 1, \ldots q \tag{79}$$

for which transformations are not applicable, can be added as a penalty function to the sum of squares of the deviations:

$$\boldsymbol{\Phi} + \sum_{i=1}^{q} \lambda_i (g_i)^{\varrho} = \boldsymbol{\Phi}^* \stackrel{!}{=} \text{Min} \tag{80}$$

For constraints given in Equation (77 b) the penalty function is $g_i = 0$ for $b_j \geq k_j$ and $g_i = (b_j - k_j)^2$ for $b_j \leq k_j$. Here the scaling factors λ and ϱ are set to 1 and 2, respectively.

Similarly constraints in the form

$$g_i(\boldsymbol{Y}, \boldsymbol{x}, \boldsymbol{\beta}) = 0, \quad i = 1, \ldots q \tag{81}$$

can be handled in the form of penalty functions which are added to the sum of squares of the deviations. They disappear when the associated conditions (Eq. 81) are satisfied. Alternatively they are weighted according to the magnitude of the deviation and added to $\boldsymbol{\Phi}$.

A more rigorous method for the treatment of associated constraints in the form of equations is based on their geometrical interpretation. In Figure 13 the sum of squares of the deviations for a two-parameter problem is shown in the parameter space. A constraint of the form

$$\beta_1 + a\beta_2 = c \tag{82}$$

can be represented as a straight line which intersects the lines of constant $\boldsymbol{\Phi}$. The minimum of $\boldsymbol{\Phi}$ is now searched for along this straight line, so that the two-dimensional problem is reduced to a one-dimensional one. This can be easily demonstrated for a linear model of the form according to Equation (5) with the constraint given by Equation (82) because one parameter is directly eliminated by substitution.

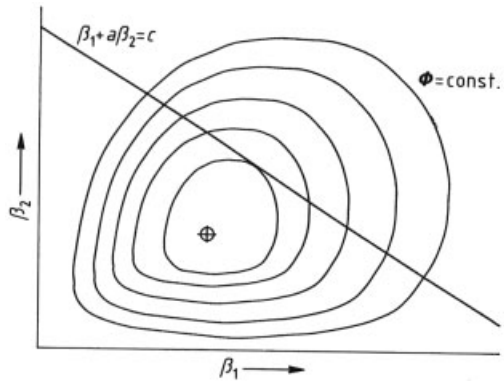

Figure 13. Geometric representation of least squares estimations with constraints on the parameters

In problems involving many parameters or nonlinear models, reduction of a multidimensional problem with one or more constraints by substitution is either impossible or very difficult. In such cases the method of Lagrangian multipliers is used. Figure 13 illustrates that for the constrained minimum, the slopes $\partial\beta_1/\partial\beta_2$ of the two curves $\boldsymbol{\Phi}=\text{const.}$ and $\beta_1+a\,\beta_2-c=0$ are equal. In general these conditions can be expressed as

$$\frac{\partial\boldsymbol{\Phi}/\partial\beta_1}{\partial\boldsymbol{\Phi}/\partial\beta_2}-\frac{\partial g/\partial\beta_1}{\partial g/\partial\beta_2}=0 \tag{83}$$

or as the gradient ratio

$$\frac{\partial\boldsymbol{\Phi}/\partial\beta_1}{\partial g/\partial\beta_1}=\frac{\partial\boldsymbol{\Phi}/\partial\beta_2}{\partial g/\partial\beta_2}=\lambda \tag{84}$$

From Equation (84) the system of equations

$$\frac{\partial\boldsymbol{\Phi}}{\partial\beta_1}-\lambda\frac{\partial g}{\partial\beta_1}=0 \tag{85a}$$

and

$$\frac{\partial\boldsymbol{\Phi}}{\partial\beta_2}-\lambda\frac{\partial g}{\partial\beta_2}=0 \tag{85b}$$

is obtained. Equations (85 a), (85 b), and the constraint given by Equation (82) may be derived from the extended sum of squares

$$\boldsymbol{\Phi}^*=\boldsymbol{\Phi}+\lambda\left(\beta_1+a\beta_2-c\right) \tag{86}$$

by partial differentiation with respect to the parameters and the Lagrangian multipliers λ. In general, therefore, when the associated constraints are in the form of Equation (81), the minimum of the extended sum of squares

$$\boldsymbol{\Phi}+\sum_{i=1}^{q}\lambda_i g_i\left(\boldsymbol{Y},\boldsymbol{x},\boldsymbol{\beta}\right)=\boldsymbol{\Phi}^*\overset{!}{=}\text{Min} \tag{87}$$

must be found by varying the parameters $\boldsymbol{\beta}$ and the Lagrange multipliers λ_i. This can be carried out using the methods discussed previously (see Section 3.2.1).

3.3.2.2. Models Based on Differential Equations

Chemical processes often give rise to models in the form of differential equations. Examples are models of reactors based on transport equations or the rates of chemical reactions. For differential equations that are simple to integrate the parameters can be determined according to the principles presented in Section 3.2, for example, as with the common first-order differential equation

$$\frac{\mathrm{d}y\left(t\right)}{\mathrm{d}t}=\alpha y\left(t\right)+x\left(t\right),\quad\text{where}\quad y_{t=0}=y_0 \tag{88}$$

The general solution of Equation (88) is [45]

$$y\left(t\right)=y_0\mathrm{e}^{\alpha t}+\int_0^t x\left(\tau\right)\mathrm{e}^{\alpha(t-\tau)}\mathrm{d}\tau \tag{89}$$

If discrete measured values $Y(t_i)$ are available, the condition for the minimum with respect to α and y_0 is given by

$$\boldsymbol{\Phi}=\sum_{i=1}^{n}\left[Y\left(t_i\right)-y_0\mathrm{e}^{\alpha t_i}+\frac{x_0}{\alpha}\left(1-\mathrm{e}^{-\alpha t_i}\right)\right]^2 \tag{90}$$

This leads to the system of nonlinear equations

$$\sum_{i=1}^{n}\left[Y\left(t_i\right)-\hat{y}_0\mathrm{e}^{\hat{\alpha}t_i}+\frac{x_0}{\hat{\alpha}}\left(1-\mathrm{e}^{-\hat{\alpha}t_i}\right)\right]\mathrm{e}^{\hat{\alpha}t_i}=0 \tag{91a}$$

and

$$\sum_{i=1}^{n}\left[Y\left(t_i\right)-\hat{y}_0\mathrm{e}^{\hat{\alpha}t_i}+\frac{x_0}{\hat{\alpha}}\left(1-\mathrm{e}^{-\hat{\alpha}t_i}\right)\right]\cdot\left[t_i\hat{y}_0\mathrm{e}^{\hat{\alpha}t_i}-\frac{x_0}{\hat{\alpha}^2}\left(\mathrm{e}^{\hat{\alpha}t_i}\left(1-\hat{\alpha}t_i\right)-1\right)\right]=0 \tag{91b}$$

which can be solved for $\hat{y}_0$ and $\hat{\alpha}$ by an appropriate method.

In the majority of cases, the differential equations of the model are more complex than Equation (88), so that no analytical solution is available; numerical solution of the differential equations is required to evaluate $\langle Y_i\rangle$. New problems then arise depending on the type of differential equation. Especially with systems of coupled partial differential equations used to model chemical reactors on the basis of transport equations (see Chap. 5), the solution of the differential equations is the real problem. In such cases other methods are often used to determine the model parameters. Changing certain process variables produces a characteristic behavior in some model solutions which is studied in suitably designed experiments.

An example is the spread of the concentration profile of a tracer added in the form of a

Dirac function into reactors with axial and/or radial dispersion. By measurement of the concentration at two points in the reactor, the Péclet number *Pe* can be derived in a relatively simple manner. Further examples are discussed in [4], [46]. A feature of these methods is that an analytical or numerical solution of the differential equations is often not necessary because the model parameters are comparatively simple functions of the deterministic moments of the measured variables. Deterministic moments are defined analogously to the moments of probability density functions of statistical variables (Section 2.2) and are related to the distribution functions of the deterministic variables. If, for example, $w(t)$ is the residence time distribution for a chemical reactor, then the first two deterministic moments m_1 and m_2' are defined as

$$m_1 = \bar{t} = \int_0^\infty t \cdot w(t)\,\mathrm{d}t$$

$$m_2' = \int_0^\infty (t - m_1)^2\, w(t)\,\mathrm{d}t$$

An example is the determination of the Péclet number for a chemical reactor with back mixing on the basis of measurements of the residence-time distribution [46], [47].

Finally, it should be noted that for processes which can be described in the form of systems of common linear differential equations analogous to Equation (88), i.e.,

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} + \boldsymbol{\beta}\boldsymbol{y} - \boldsymbol{x}(t)\,, \quad \boldsymbol{y}_{t=0} = 0 \tag{92}$$

the parameters can be determined with the Laplace transforms of the transfer functions. The transfer functions can be given in the physical time coordinates as well as in the coordinates of the Laplace transformation. The methods are discussed in detail in [4], [9], [13–15].

4. Models Based on Transport Equations for Probability Density Functions

Probability density functions and the probabilities which can be calculated from them are the most adequate way of describing statistical variables. Examples of statistical variables can be found at the molecular level in the molecular velocity resulting from random thermal movements; the corresponding probability density function is the Maxwell – Boltzmann velocity distribution [48]. The use of this concept in kinetic gas theory and statistical mechanics has been referred to in Section 2.2.

Statistical variables can also be found at the macroscopic level. In the example discussed in Chapter 3, statistical process variables with presumed distribution or probability density functions were examined and their expected values, variances, and functional relationships to other variables were determined.

Another example is the jet stirred fluid-phase reactor with turbulent flow presented schematically in Figure 14. The surrounding fluid entrains the jet by exchange of momentum. The intake volume of the jet is greater than the volume of the surrounding fluid. This leads to recirculation of material from further downstream and mixing of the reactor contents. The fluid in the jet exhibits turbulent flow. The velocity of the fluid particles is so high that they can no longer be held in regular streamlines by the viscous forces, and irregular vortices are formed. Vortices of different sizes with different velocities and different macroscopic properties pass a fixed observer. Therefore, the fixed observer measures statistical, time-dependent variations in the physical quantity Φ.

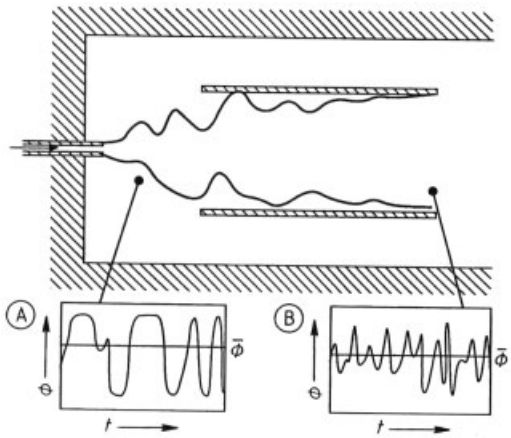

Figure 14. Example of a process with statistical variables: a jet stirred reactor with turbulent flow

In Figure 14 this is shown at two positions in the jet. At the edge of the nozzle (A) either the

properties of the jet or the surroundings are measured. Further downstream (B) the macroscopic properties are balanced due to turbulent mixing. All physical parameters are thus characterized by position- and time-dependent distribution or probability density functions. The objective of models based on transport equations for probability density functions is the calculation of functions $P(\Phi_1, \ldots, \Phi_n)$ in which $\Phi_1, \ldots, \Phi_n$ is the vector of the dependent statistical variables. This has to be specified for each problem. Models based on transport equations for probability density functions generally give a differential equation for $P(\Phi_1, \ldots, \Phi_n)$.

4.1. Terminology

4.1.1. One-Dimensional Distribution and Probability Density Functions

The example shown in Figure 14 will be used to explain terminology. The temperature of the fluid is treated as a macroscopic physical quantity. Since the temperature varies between two limits $T_b > T > T_a$, it is better to use a normalized variable so that the statistical variable Φ is defined as

$$\Phi = \frac{T - T_b}{T_a - T_b} \tag{93}$$

The sample space for Φ (see Fig. 15) is a line on which only values between 0 and 1 can be realized. An event is defined as $E \equiv \varphi_u \leq \Phi \leq \varphi_o$. The probability of the occurrence of a particular event is between 0 and 1. When $W(E) = 0$, $\varphi_u \leq \Phi \leq \Phi_o$ never occurs and when $W(E) = 1$, $\varphi_u < \Phi < \varphi_o$ always occurs. Probability defined in this way does not predict E for the next experiment but instead the fraction of occurrences of E over a large number of experiments. Since each event is linked to a defined area of the sample space, the probability of an event is the probability that a random variable lies in this area. The probability of an event can be defined from the distribution function of the random variables:

$$F(\Phi) \equiv W(\Phi \leq \varphi) \tag{94}$$

Since the normalized temperatures defined in Equation (93) can only have values between $0 \leq \Phi \leq 1$, then

$$W(\Phi \leq 0) = F(0) = 0 \quad \text{and} \tag{95a}$$

$$W(\Phi \leq 1) = F(1) = 1 \tag{95b}$$

Any event, i.e., any interval $E \equiv \varphi_u \leq \Phi \leq \varphi_o$ in the sample space can be obtained by subtraction (see Fig. 15):

$$E = (\Phi \leq \varphi_o) - (\Phi \leq \varphi_u) \tag{96}$$

Thus

$$W(E) = W(\varphi_u \leq \Phi \leq \varphi_o) = W(\Phi \leq \varphi_o) - W(\Phi \leq \varphi_u) = F(\varphi_o) - F(\varphi_u) \tag{97}$$

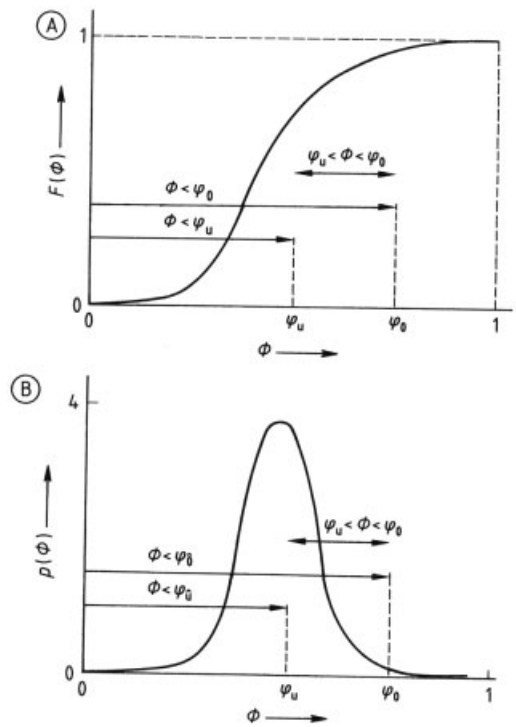

Figure 15. Representation of the distribution function (A) and the probability density function (B) of a statistical variable
The φ axis represents the sample space for the statistical variable Φ. The values φ_o and φ_u confine a finite region within the sample space.

The fundamental characteristics of the distribution function are easily derived by similar reasoning. Since in the general case $W(\Phi \leq -\infty) = F(-\infty) = 0$ and $W(\Phi \leq +\infty) = F(+\infty) = 1$, then

$$W(E) = F(\varphi_o) - F(\varphi_u) \geq 0 \tag{98}$$

and

$$F(\varphi_o) \geq F(\varphi_u) \quad \text{for} \quad \varphi_o \geq \varphi_u \tag{99}$$

The distribution function $F(\Phi)$ is therefore a nondecaying function which grows from 0 at $\varphi = -\infty$ to 1 at $\varphi = +\infty$ (Fig. 15 A).

The probability density function of the random variable Φ is the derivative of the distribution function with respect to the random variable (Fig. 15 B).

$$P(\Phi) = \frac{d}{d\Phi}[F(\Phi)] \tag{100}$$

Using this result then

$$\int_{\varphi_u}^{\varphi_o} P(\Phi)\, d\Phi = F(\varphi_o) - F(\varphi_u)$$
$$= W(\varphi_u \leq \Phi \leq \varphi_o) \tag{101}$$

According to Equation (101) the probability that a random variable falls in a particular area is equal to the integral of the probability density function over that area. For an infinitely small area $d\varphi$ this then becomes

$$W(\varphi \leq \Phi \leq \varphi + d\varphi) = P(\Phi)\, d\Phi \tag{102}$$

(cf. Section 2.2). The fundamental characteristics of $P(\Phi)$ can be derived from Equation (100). Since $F(\Phi)$ is a nondecaying function this gives

$$P(\Phi) \geq 0 \tag{103}$$

Further

$$\int_{-\infty}^{+\infty} P(\Phi)\, d\Phi = 1 \tag{104}$$

since according to Equation (101) the probability of the occurrence of the random variable in the area from $-\infty$ to $+\infty$ in the sample space must be 1. The third fundamental characteristic of the probability density function follows from the monotonic approximation of the distribution function to 0 or 1 when $-\Phi- \to \infty$. Thus

$$\frac{d}{d\Phi}[F(\Phi)]_{-\infty} = P(\Phi)_{-\infty}$$
$$= \frac{d}{d\Phi}[F(\Phi)]_{+\infty} = P(\Phi)_{+\infty} = 0 \tag{105}$$

For discontinuous distribution functions with statistical variables which can only take discrete values (see Chap. 3) the above relationships can be given with additional derivation steps [18], [19], [22]. Continuity of the distribution function is not a necessary condition for the above relationships.

The relationships for transformations of distribution functions and probability density functions are easy to see from Figure 15 and Equations (98) – (100). For example if the relationship $\psi = f(\varphi)$ holds for a given transformation of Φ, then the events $\Phi \leq \varphi$ and $\Psi \leq \psi = f(\varphi)$ are identical. From this $F(\Phi) = F(\Psi)$ and according to Equation (100)

$$P(\Psi) = \frac{d}{d\Psi}F(\Psi) = \frac{d}{d\Phi}F(\Phi) \cdot \frac{d\Phi}{d\Psi} =$$

$$P(\Phi) \cdot \frac{d\Phi}{d\Psi} \tag{106}$$

Figure 14 shows that the distribution functions and probability density functions can be dependent on position and are thus not always homogeneous. For the area at the edge of the jet close to the nozzle exit an almost bimodal probability density function is produced due to intermittencies of the jet and the surrounding fluid. Further downstream in the middle of the turbulent jet these intermittencies no longer affect the form of the probability density function. Thus probability density functions hold for a given position at a given time; they will be written as $P(\Phi; x, t)$ from now on.

For engineering applications the probability density functions themselves are not very illustrative. Statistical variables are usually described in terms of expected values (see Section 2.2 and Chap. 3). These are easier to handle but have a lower information content. The expected values of statistical variables (synonyms are mean values, mathematical expectations, and first moment) are given by

$$\langle \Phi(\boldsymbol{x}, t) \rangle = \int_{-\infty}^{+\infty} \Phi P(\Phi; \boldsymbol{x}, t)\, d\Phi \tag{107}$$

For statistical steady-state processes or for $d\langle\Phi\rangle/dt \ll d\Phi/dt$ the equivalent form of the averaging is

$$\langle \Phi(\boldsymbol{x},t) \rangle = \frac{1}{\Delta\tau} \int_{t}^{t+\Delta\tau} \Phi(\boldsymbol{x},t)\,\mathrm{d}t \tag{108}$$

For this a continuous function of Φ versus time is necessary as indicated in the plot of Φ against time given in Figure 14.

Functions of statistical variables are themselves statistical values, so that the operations related to Equation (107) for deriving expected values of functions can be used

$$\langle Q(\Phi) \rangle = \int_{-\infty}^{+\infty} Q(\Phi)\, P(\Phi;\boldsymbol{x},t)\,\mathrm{d}\Phi \tag{109}$$

The definitions of Equations (107) and (109) respectively are analogously applicable to higher moments

$$\mu_n \equiv \langle \Phi(\boldsymbol{x},t)^n \rangle = \int_{-\infty}^{+\infty} \Phi^n P(\Phi;\boldsymbol{x},t)\,\mathrm{d}\Phi \tag{110}$$

For example, the second moment around the mean (the variance) is also a characteristic of statistical quantities which is often used in engineering applications (see Chap. 3):

$$\begin{aligned}\mathrm{Var}(\Phi) &\equiv \mu_2' \equiv \langle (\Phi - \langle\Phi\rangle)^2 \rangle \\ &= \int_{-\infty}^{+\infty} (\Phi - \langle\Phi\rangle)^2 P(\Phi;\boldsymbol{x},t)\,\mathrm{d}\Phi\end{aligned} \tag{111}$$

All of these quantities can be determined from the probability density function. However, the probability density function can only be given approximately from a finite number of moments which are relatively easy to measure [49].

If a statistical quantity is split up into the expected value and a fluctuating component Φ^{f}

$$\Phi = \langle\Phi\rangle + \Phi^{\mathrm{f}} \tag{112}$$

then from Equation (111)

$$\mathrm{Var}(\Phi) = \langle \left(\Phi^{\mathrm{f}}\right)^2 \rangle \quad \text{and} \quad \langle \Phi^{\mathrm{f}} \rangle = 0 \tag{113a,b}$$

Rules for measuring the statistical moments of statistical steady-state processes are derived from Equation (108). The continuous function $\Phi(\boldsymbol{x}, t)$ is necessary for this. However measurements of statistical variables are mostly available in the form of discrete values rather than in continuous form. Therefore all values referred to here must be derived from discrete data. This has been done for the empirical models discussed in Chapter 3 where discrete values of the statistical variables were available.

For n independent repetitions of an experiment under constant conditions the ensemble mean value

$$\langle \Phi \rangle_n = \frac{1}{n} \sum_{i=1}^{n} \varphi_i \tag{114}$$

gives a good approximation of $\langle\Phi\rangle$ when $n \to \infty$.

In the examples discussed in Chapter 3 the number of occurrences is so small that the mean values from the individual data sets do not represent the expected values but are statistical values. Similarly to Equation (114) the means of functions $Q(\Phi)$ or higher moments are estimated, c.f. Chapter 3 where the sums of squares of the deviations were used as estimates for the variances. The probability density function is measurable through the number of occurrences of φ_i in a given area $\Delta\varphi$ of the sample space. If this number is denoted as $m(\Phi, \Delta\Phi)$, the normalized density of occurrences in the sample space provides a good estimate for the probability density function:

$$P_n(\Phi) = \frac{m(\Phi, \Delta\Phi)}{n\Delta\Phi} \approx P(\Phi;\boldsymbol{x},t) \quad \text{for} \quad n \to \infty \tag{115a}$$

The discrete form of the probability density function can also be expressed as

$$P_n(\Phi) = \frac{1}{n} \sum_{i=1}^{n} \delta(\Phi - \varphi_i) \tag{115b}$$

Here $\delta(a-b)$ is the impulse or Dirac delta function which is defined as the derivative of the Heaviside or step function:

$$\delta(a-b) = \frac{\mathrm{d}H(a-b)}{\mathrm{d}a} \tag{116}$$

with $H(a-b) = 0$ for $a \le b$ and $H(a-b) = 1$ for $a > b$.

According to Equation (115 b) the discrete form of the probability density function in the sample space for Φ is the number impulses of discrete occurrences normalized with respect to the total count. The discrete form of the distribution function

$$F_n(\Phi) = \frac{1}{n}\sum_{i=1}^{n} H(\Phi - \varphi_i) \tag{115c}$$

becomes a stepped curve as shown in Figure 16.

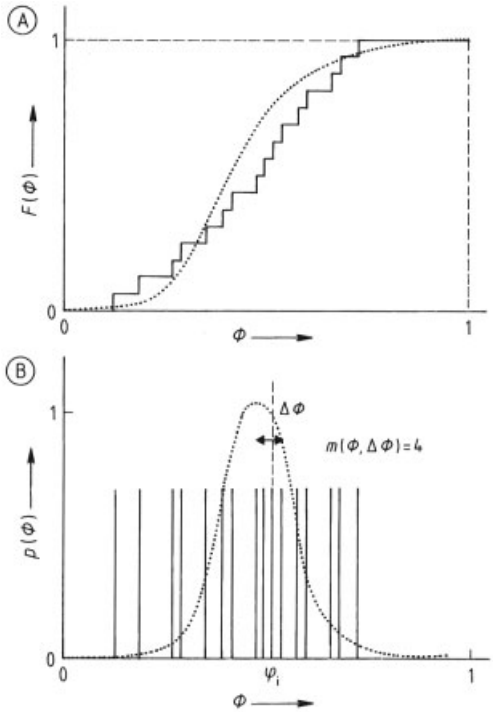

Figure 16. Discrete forms of the distribution function (A) and the probability density function (B)
The continuous functions are shown as dotted lines.

4.1.2. Multidimensional Distribution and Probability Density Functions

For the example shown in Figure 14 the normalized temperature alone is clearly not sufficient to describe the state at a point $\boldsymbol{x}$ and time t. Other variables such as the axial velocity U can be used which is expressed in normalized form as

$$\Psi = \frac{U - U_b}{U_a - U_b} \tag{117}$$

For the statistical variable Ψ the same definitions hold as previously for Φ so that

$$W(\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi) = P(\Phi; \boldsymbol{x}, t)\,\mathrm{d}\Phi$$

and

$$W(\psi \leq \Psi \leq \psi + \mathrm{d}\psi) = P(\Psi; \boldsymbol{x}, t)\,\mathrm{d}\Psi$$

However, this information is not sufficient. More information is required about the probability of the joint events

$$E_\varphi = (\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi)$$

and

$$E_\psi = (\psi \leq \Psi \leq \psi + \mathrm{d}\psi)$$

This information can be derived from the joint distribution function $F(\Phi, \Psi; \boldsymbol{x}, t)$ which has characteristics analogous to those of the simple distribution function:

$$0 \leq F(\Phi, \Psi) \leq 1, \quad \text{nondecaying}, \tag{118a}$$

$$F(-\infty, \Psi) = F(\Phi, -\infty) = 0, \tag{118b}$$

$$F(\infty, \Psi) = F(\Psi), \quad \text{and} \tag{118c}$$

$$F(\Phi, \infty) = F(\Phi) \tag{118d}$$

The joint probability density function becomes

$$P(\Phi, \Psi; \boldsymbol{x}, t) = \frac{\partial^2}{\partial\Phi\partial\Psi}\left[F(\Phi, \Psi; \boldsymbol{x}, t)\right] \tag{119}$$

which possesses the same fundamental characteristics as the simple probability density function:

$$P(\Phi, \Psi; \boldsymbol{x}, t) \geq 0 \tag{120a}$$

$$\int\int P(\Phi, \Psi; \boldsymbol{x}, t)\,\mathrm{d}\Psi\mathrm{d}\Phi = 1 \tag{120b}$$

The simple probability density function or marginal distributions (the reduced information) can be derived from the joint probability density function:

$$\int P(\Phi, \Psi; \boldsymbol{x}, t)\,\mathrm{d}\Psi = P(\Phi; \boldsymbol{x}, t) \tag{121a}$$

$$\int P(\Phi, \Psi; \boldsymbol{x}, t)\,\mathrm{d}\Phi = P(\Psi; \boldsymbol{x}, t) \tag{121b}$$

Conversely the joint probability density functions cannot normally be derived from the marginal distributions. Arguments analogous to those above may be used for the functions $Q(\Phi, \Psi)$.

Covariances are another quality of joint probability density functions:

$$\mathrm{Cov}\,(\Phi,\Psi) = \left\langle \Phi^{\mathrm{f}}\Psi^{\mathrm{f}} \right\rangle$$
$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left(\Phi - \langle\Phi\rangle\right)\left(\Psi - \langle\Psi\rangle\right) \cdot P\,(\Phi,\Psi;\boldsymbol{x},t)\,\mathrm{d}\Phi\mathrm{d}\Psi \tag{122}$$

They are mostly given in normalized form as correlation coefficients:

$$\gamma = \frac{\mathrm{Cov}\,(\Phi,\Psi)}{\left[\mathrm{Var}\,(\Phi)\,\mathrm{Var}\,(\Psi)\right]^{1/2}} \tag{123}$$

With the help of Equation (122) it is easy to show that $-1 \leq \gamma \leq 1$. If a linear relationship of the form $\psi = \alpha + \beta\,\varphi$ exists, then $|\gamma| = 1$.

For joint multidimensional probability density functions, transformation rules are also valid. Using the same reasoning as with one-dimensional probability density functions and assuming that $\xi = \xi\,(\varphi, \psi)$ and $\theta = \theta\,(\varphi, \psi)$

$$P\,(\Xi,\Theta) = P\,(\Phi,\Psi) \cdot |\boldsymbol{J}|^{-1} \tag{124}$$

where $\boldsymbol{J}$ is the Jacobi matrix of the system of transformation equations.

4.1.3. Conditional Probability Density Functions

Conditional probability is a prerequisite for the derivation of transport equations for probabiliby density functions. It denotes the probability of an event, for example $\Psi \leq \psi$, subject to the condition that the event $\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi$ occurs. It is given by the probability for a general event divided by the probability of the condition:

$$W\,[(\Psi \leq \psi)\,|\,(\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi)] = \frac{W\,(\Psi \leq \psi)\,(\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi)}{W\,(\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi)} \tag{125}$$

Consequently

$$W\,[(\Psi \leq \psi)\,|\,(\varphi \leq \Phi \leq \varphi + \mathrm{d}\varphi)] = \frac{\int_{-\infty}^{\psi}\int_{\varphi}^{\varphi+\mathrm{d}\varphi} P\,(\Phi,\Psi)\,\mathrm{d}\Phi\mathrm{d}\Psi}{\int_{\varphi}^{\varphi+\mathrm{d}\varphi} P\,(\Phi)\,\mathrm{d}\Phi} \tag{126}$$

The conditional distribution function is the limit of the conditional probability for $\mathrm{d}\varphi \to 0$ so that

$$F\,(\Psi|\varphi) = P\,[(\Psi \leq \psi)\,|\,(\Phi = \varphi)] = \int_{-\infty}^{\psi} \frac{P\,(\Psi,\Phi)\,\mathrm{d}\Psi}{P\,(\Phi)} \tag{127}$$

The usual definition of probability density function then gives the following:

$$P\,(\Psi|\varphi) = \frac{\mathrm{d}}{\mathrm{d}\Psi}F\,(\Psi|\varphi) = \frac{P\,(\Psi,\Phi)}{P\,(\Phi)} \tag{128}$$

It is easy to show that the conditional distribution functions and probability density functions exhibit the same characteristics as unconditional ones. By using the same operations as above, the conditional expected value of a function $Q\,(\Psi, \Phi)$ can be derived

$$\left\langle Q\,(\Psi,\Phi)\,|\Phi = \varphi\right\rangle = \int_{-\infty}^{+\infty} Q\,(\Psi,\Phi)\,P\,(\Psi|\varphi)\,\mathrm{d}\Psi \tag{129}$$

From Equation (128) it can be seen that the joint probability density function $P\,(\Psi, \Phi)$ can be obtained from the conditional probability density function $P\,(\Psi|\varphi)$ and the marginal distribution $P\,(\Phi)$. This important result is used in the determination of expected values and leads to a further important result:

$$\left\langle Q\,(\Psi,\Phi)\right\rangle = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Q\,(\Psi,\Phi)\,P\,(\Psi,\Phi)\,\mathrm{d}\Psi\mathrm{d}\Phi$$
$$= \int_{-\infty}^{+\infty} P\,(\Phi)\left\{\int_{-\infty}^{+\infty} Q\,(\Psi,\Phi)\,P\,(\Psi|\varphi)\,\mathrm{d}\Psi\right\}\mathrm{d}\Phi$$
$$= \int_{-\infty}^{+\infty} P\,(\Phi) < Q\,(\Psi,\Phi)\,|\Phi = \varphi > \mathrm{d}\Phi \tag{130}$$

4.2. Transport Equations for Single-Point Probability Density Functions

Before going into some examples, the following sections will briefly discuss transport equations for single-point probability density functions starting from the general conservation equations. Only the important relationships will be described. A more detailed discussion of the general conservation equations is given in [48].

The state of a reacting fluid mixture (e.g., the contents of the reactor shown in Fig. 14) is accurately defined by the velocity $\boldsymbol{U}(\boldsymbol{x},t)$, the pressure $p(\boldsymbol{x},t)$, the mass fraction of the chemical species present in the mixture $m_i(\boldsymbol{x},t)$, $i=1,\ldots,N$, and the specific enthalpy $h(\boldsymbol{x},t)$. A complete statistical single-point description of the mixture is given by the joint probability density function $P(\boldsymbol{U},\Phi;\boldsymbol{x},t)$ where $\boldsymbol{U}=(U_1,U_2,U_3)$ and $\boldsymbol{\Phi}$ is a vector of length $N+2$ which includes the rest of the scalar quantities. Thus the required function is a surface in the $(N+3+2)$-dimensional Euclidean space.

The relationship between the instantaneous values of the variables stated above is given by the laws of conservation of total mass, momentum, and mass of the individual chemical species, and by the enthalpy. These can be given in the form [48], [50], [51]:

Accumulation + Change due to convective transport = Change due to molecular transport + Source

This leads to

$$\frac{\partial \varrho}{\partial t}+\frac{\partial}{\partial x_k}(\varrho U_k)=0 \tag{131}$$

for the total mass,

$$\varrho\frac{DU_i}{Dt}=\frac{\partial \tau_{ik}}{\partial x_k}-\frac{\partial p}{\partial x_i}+\varrho g_i,\quad i=1,2,3 \tag{132}$$

for the component of the momentum,

$$\varrho\frac{Dm_i}{Dt}=-\frac{\partial J_k^{m_i}}{\partial x_k}+\varrho S_i,\quad i=1,\ldots,N-1 \tag{133}$$

for the mass of the individual chemical species, and finally

$$\varrho\frac{Dh}{Dt}=-\frac{\partial J_k^{h}}{\partial x_k}+\varrho S_h \tag{134}$$

for the enthalpy.

In Equations (132) – (134) the operator D denotes the substantial derivative:

$$\frac{D}{Dt}=\frac{\partial}{\partial t}-U_k\frac{\partial}{\partial x_k} \tag{135}$$

The Einstein summation convention is also used, for example:

$$\frac{\partial}{\partial x_k}(\varrho U_k)\equiv\frac{\partial}{x_1}(\varrho U_1)+\frac{\partial}{x_2}(\varrho U_2)+\frac{\partial}{x_3}(\varrho U_3)$$

The formulation of Equations (132) – (134) implies the continuity equation; ϱ is the density; m_i is the mass fraction of component i; τ_{ij} is the viscous stress tensor; $\varrho\, g_i$ is the specific mass force in the x_i direction (e.g., $\varrho\, g_1$, where g is the acceleration due to gravity); S_i is the reaction rate and S_h the rate of enthalpy conversion produced by pressure changes, viscosity effects etc.; J^{m} is the mass flux of the chemical species and J^{h} is the enthalpy flux, based on molecular transport mechanisms, like τ_{ij} they are functions of local physical quantities and their gradients.

Equations (131) – (134) can be closed by a thermal equation of state

$$\varrho=\varrho(m_i,h,p) \tag{136}$$

The expressions for the sources S_i and S_h are also generally functions of the scalar quantities:

$$S_i=S_i(m_i,h,p) \tag{137a}$$

$$S_h=S_h\left(m_i,h,p,\frac{\partial p}{\partial t}\right) \tag{137b}$$

The assumptions and conditions included in this system of equations are discussed fully in [48], [50–52], see also Chapter 5. The system of Equations (131) – (137) is deterministic, this means that a determined solution exists for each set of determined boundary conditions. The solution of this system of equations for deterministic variables is illustrated with some examples in Chapter 5. Problems arise for statistical variables [52] which can only be solved for simple cases [53], [54] or by introducing far-reaching assumptions (see Chap. 5). The system of conservation equations serves as a starting point for the derivation of the transport equations for the probability density functions which are used to overcome the principle problems for solving this set of equations for statistical variables.

4.2.1. General form of the Transport Equations for Probability Density Functions

Equations (131) – (137) show that the state of a chemically reacting system at a position $\boldsymbol{x}=(x_1,x_2,x_3)$ and time t is determined by the velocity vector $\boldsymbol{U}=(U_1,U_2,U_3)$ in the sample space of velocity (Section 3.1.2.1 and Fig. 7) and the vector of scalar quantities $\boldsymbol{\Phi}=(m_1,\ldots,m_N,h)$

of length σ in the sample space of scalar quantities. $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ are vectors of random variables, i.e., random vectors. This formulation holds for low Mach numbers.

The state of the reacting system represents a measured point in the three-dimensional velocity space of the vector $\boldsymbol{V}$ and likewise a measured point in the Euclidean σ-dimensional space of the vector $\boldsymbol{\Phi}$.

For a homogeneous mixture at rest with constant enthalpy, Equation (133) gives

$$\frac{\mathrm{d}m_i}{\mathrm{d}t} = S_i \tag{138}$$

A chemical reaction can be represented in the composition space as a trajectory between two measured points. A system with two chemical species is shown in Figure 17. The reaction rate is the velocity in the composition space, and the reaction rates of the individual chemical species are given by the tangents to the trajectory in the direction of each species. Since the state of the system can be described by a vector with its end point in the plane of the required probability density function, changes of state due to chemical reactions are therefore trajectories in this plane. Chemical reactions that constitute the source terms in the conservation equations of the chemical species (Eq. 133) are contained in the transport equations for the probability density functions merely as transport in the composition space. Similarly, forces which are the source terms in the momentum conservation equations are expressed in the transport equations for probability density functions as transport in the velocity space. These are the main differences between the simple conservation equations and the transport equations for the probability density functions. The latter do not include forces or chemical reactions as sources.

Both $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ are necessary to complete the information. The complete information at position $\boldsymbol{x} = (x_1, x_2, x_3)$ and time t is represented by a measured point in Euclidean $(\sigma + 3)$-dimensional space. Consequently, the objective of the complete description is a $(\sigma + 3)$-dimensional distribution function $F(\boldsymbol{U}, \Phi; \boldsymbol{x}, t)$ or the corresponding probability density function:

$$\frac{\partial^{\sigma+3}}{\partial U_1 \partial U_2 \ldots \partial \Phi_{n+1}} \left[F\left(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t\right)\right] = P\left(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t\right) \tag{139}$$

The characteristics described in Sections 4.1.1, 4.1.2, 4.1.3 apply to these functions. In the terminology of probabilities these functions predict the fraction of occurrences of discrete values $\boldsymbol{u}$ and $\boldsymbol{\varphi}$ for a large number of occurrences.

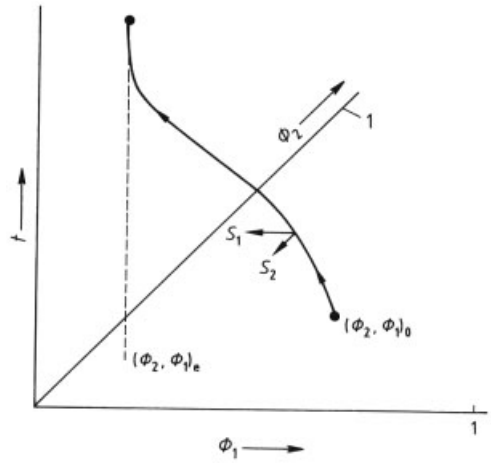

Figure 17. Chemical reactions for a system with two chemical species represented as transport in the sample space of the concentration (Φ_1, Φ_2) of the species
S_1 and S_2 are the partial derivatives of the trajectory in the Φ_1 and Φ_2 direction, the subscripts 0 and e denote initial and final states, respectively.

A transport equation for the joint velocity – composition probability density function can be derived with help of Equations (132) – (134) and the terminology described in Section 4.1. This method is one possibility for deriving the transport equation for probability density functions. This will be used in preference to other methods [18], [55–58] because of its simplicity.

Equations (132) and (133) are rewritten as Equations (140) and (141), respectively:

$$\frac{DU_i}{Dt} = A_i \tag{140}$$

$$\frac{D\Phi_\alpha}{Dt} = \Theta_\alpha \tag{141}$$

where

$$\varrho A_i\left(\boldsymbol{x}, t\right) = \frac{\partial \tau_{ik}}{\partial x_k} - \frac{\partial p}{\partial x_i} + \varrho g_i \quad \text{and} \tag{142a}$$

$$\varrho \Theta_\alpha\left(\boldsymbol{x}, t\right) = -\frac{\partial J_k^\alpha}{\partial x_k} + \varrho S_\alpha \tag{142b}$$

For any function $Q(\boldsymbol{U}, \boldsymbol{\Phi})$ averaging according to Equation (109) gives

$$\left\langle \varrho \frac{DQ\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{Dt}\right\rangle = \frac{\partial}{\partial t}\int\int \varrho\left(\Phi\right)Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)P\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} + \frac{\partial}{\partial x_k}\int\int \varrho\left(\boldsymbol{\Phi}\right)U_k Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)P\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} = \int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\left\{\varrho\left(\boldsymbol{\Phi}\right)\frac{\partial P}{\partial t} + \varrho\left(\boldsymbol{\Phi}\right) U_k\frac{\partial P}{\partial x_k}\right\}\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{143}$$

In Equation (143) and in the following discussions, P will be used to represent $P\left(\boldsymbol{U},\Phi;\boldsymbol{x},t\right)$. In addition, $\mathrm{d}\boldsymbol{U}$ will be used to describe the vector $\mathrm{d}U_1, \mathrm{d}U_2, \mathrm{d}U_3$ and $\mathrm{d}\boldsymbol{\Phi}$ for $\mathrm{d}\Phi_1,\ldots,\mathrm{d}\Phi_{N+1}$. The integral sign in Equation (143) will also be interpreted in this way. The continuity equation (Eq. 131) is used to derive the expression $\left\langle \varrho\frac{DQ(\boldsymbol{U},\boldsymbol{\Phi})}{Dt}\right\rangle$ in Equation (143).

A second independent expression for $\left\langle \varrho\frac{DQ(\boldsymbol{U},\boldsymbol{\Phi})}{Dt}\right\rangle$ is obtained if $Q\left(\boldsymbol{U},\Phi\right)$ is differentiated with respect to $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ (the Einstein summation convention is used for ω in the same way as for k in Eqs. 131 – 135).

$$\frac{DQ\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{Dt} = \frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}\cdot\frac{DU_k}{Dt} + \frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial \Phi_\omega}\cdot\frac{D\Phi_\omega}{Dt} \tag{144}$$

If Equations (140) and (141) are used, the substantial derivatives can be substituted with the respective right-hand sides. Thus,

$$\left\langle \varrho\frac{DQ\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{Dt}\right\rangle = \left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}A_k\right\rangle + \left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial \Phi_\omega}\Theta_\omega\right\rangle \tag{145}$$

The terms on the right-hand side of Equation (145) can be rewritten as conditional expected values by using Equation (130). For the first term of the right-hand side of Equation (145) this gives

$$\left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}A_k\right\rangle = \int\int \left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{146}$$

Since the differential quotient $\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)/\partial U_k$ is a known function of $\boldsymbol{u}$ and $\boldsymbol{\varphi}$ for a given $\boldsymbol{u}$ and $\boldsymbol{\varphi}$ then from Equation (146)

$$\left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}A_k\right\rangle = \int\int \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{147}$$

If Equation (147) is partially integrated, then

$$\left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial U_k}A_k\right\rangle = I_U - \int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\frac{\partial}{\partial U_k}\left(\varrho\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right)\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{148}$$

where

$$I_U = \int\int \frac{\partial}{\partial U_k}\left[\varrho\left(\boldsymbol{\Phi}\right)Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right]\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi}$$

An analogous procedure for the second term on the right-hand side of Equation (145) gives

$$\left\langle \varrho\frac{\partial Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{\partial \Phi_\omega}\Theta_\omega\right\rangle = I_{\boldsymbol{\Phi}} - \int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\frac{\partial}{\partial \Phi_\omega}\left(\varrho\left\langle \Theta_\omega|\boldsymbol{u},\varphi\right\rangle P\right)\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{149}$$

where

$$I_{\boldsymbol{\Phi}} = \int\int \frac{\partial}{\partial \Phi_\omega}\left[\varrho\left(\boldsymbol{\Phi}\right)Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\left\langle \Theta_\omega|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right]\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi}$$

It can be shown that the integrals I_U and I_Φ disappear for a wide class of functions $Q\left(\boldsymbol{U},\Phi\right)$. The general condition for this is that the functions are bounded and continuous [19]. With this a second independent expression for $\left\langle \varrho\frac{DQ(\boldsymbol{U},\boldsymbol{\Phi})}{Dt}\right\rangle$ is obtained from Equation (145):

$$\left\langle \varrho\frac{DQ\left(\boldsymbol{U},\boldsymbol{\Phi}\right)}{Dt}\right\rangle = -\int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\frac{\partial}{\partial U_k}\left(\varrho\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right)\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} - \int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\frac{\partial}{\partial \Phi_\omega}\left(\varrho\left\langle \Theta_\omega|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right)\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} \tag{150}$$

Subtraction of Equation (150) from Equation (143) gives

$$\int\int Q\left(\boldsymbol{U},\boldsymbol{\Phi}\right)\left\{\varrho\left(\boldsymbol{\Phi}\right)\frac{\partial P}{\partial t} + \varrho\left(\boldsymbol{\Phi}\right)U_k\frac{\partial P}{\partial x_k} + \frac{\partial}{\partial U_k}\left[\varrho\left(\boldsymbol{\Phi}\right)\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle \mathrm{P}\right] + \frac{\partial}{\partial \Phi_\omega}\left[\varrho\left(\boldsymbol{\Phi}\right)\left\langle \Theta_\omega|\boldsymbol{u},\varphi\right\rangle P\right]\right\}\mathrm{d}\boldsymbol{U}\mathrm{d}\boldsymbol{\Phi} = 0 \tag{151}$$

All the terms in the curved brackets are independent of $Q(\boldsymbol{U},\Phi)$. If $Q(\boldsymbol{U},\Phi)$ is the type of function where the integrals I_u and I_Φ disappear, then a sufficient condition to fulfil Equation (151) is that the sums in the curved brackets are zero. With this the transport equation for the joint probability density function $P(\boldsymbol{U},\Phi;\boldsymbol{x},t)$ is

$$\varrho(\boldsymbol{\Phi})\frac{\partial P}{\partial t}+\varrho(\boldsymbol{\Phi})U_k\frac{\partial P}{\partial x_k} = -\frac{\partial}{\partial U_k}\left[\varrho(\boldsymbol{\Phi})\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\rangle P\right]-\frac{\partial}{\partial \Phi_\omega}\cdot\left[\varrho(\boldsymbol{\Phi})\langle\Theta_\omega|\boldsymbol{u},\boldsymbol{\varphi}\rangle P\right] \tag{152}$$

For further discussion of this transport equation, A_i and Θ_α are replaced by their respective definition functions:

$$\varrho(\boldsymbol{\Phi})\langle\Theta_\alpha|\boldsymbol{u},\boldsymbol{\varphi}\rangle=-\langle\frac{\partial J_k^\alpha}{\partial x_k}+\varrho(\boldsymbol{\Phi})S_\alpha(\boldsymbol{\Phi})|\boldsymbol{u},\ vec\varphi\rangle = -\langle\frac{\partial J_k^\alpha}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\rangle+\varrho(\boldsymbol{\Phi})S_\alpha(\boldsymbol{\Phi}) \tag{153}$$

Using the decomposition $p^{\mathrm{f}}\equiv p-\langle p\rangle$ gives

$$\varrho(\boldsymbol{\Phi})\langle A_i|\boldsymbol{u},\boldsymbol{\varphi}\rangle =\langle\frac{\partial\tau_{ik}}{\partial x_k}-\frac{\partial\langle p\rangle}{\partial x_i}-\frac{\partial p^{\mathrm{f}}}{\partial x_i}+\varrho g_i|\boldsymbol{u},\boldsymbol{\varphi}\rangle = \langle\frac{\partial\tau_{ik}}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\rangle-\frac{\partial\langle p\rangle}{\partial x_i}-\langle\frac{\partial p^{\mathrm{f}}}{\partial x_i}|\boldsymbol{u},\boldsymbol{\varphi}\rangle +\langle\varrho g_i|\boldsymbol{u},\boldsymbol{\varphi}\rangle \tag{154}$$

Substitution of these expressions into Equation (152) gives

$$\varrho(\boldsymbol{\Phi})\frac{\partial P}{\partial t}+\varrho(\boldsymbol{\Phi})U_k\frac{\partial P}{\partial x_k}+\left[\varrho(\boldsymbol{\Phi})g_k-\frac{\partial\langle p\rangle}{\partial x_k}\right]\frac{\partial P}{\partial U_k}+\frac{\partial}{\partial\Phi_\omega}\left[\varrho(\boldsymbol{\Phi})S_\omega(\boldsymbol{\Phi})P\right] = -\frac{\partial}{\partial U_k}\left[\langle\frac{\partial\tau_{lk}}{\partial x_l}+\frac{\partial p^{\mathrm{f}}}{\partial x_l}|\boldsymbol{u},\boldsymbol{\varphi}\rangle P\right] +\frac{\partial}{\partial\Phi_\omega}\left[\langle\frac{\partial J_k^\omega}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\rangle P\right] \tag{155}$$

This is the general form of the transport equation for the probability density functions in the combined velocity – composition space (in the first term of the right-hand side k and l must be summed according to the Einstein summation convention). Equation (155) can be solved if the conditional expected values on the right-hand side are known. All physical processes represented by the terms on the left-hand side are described exactly in the context of the method of description given here. These processes are transport in physical space (first and second terms), transport in the velocity space due to gravitational forces and average pressure gradient (third term), and transport in the composition space by chemical reactions (fourth term). All of these processes can be described without any model assumptions.

The expressions on the right-hand side of Equation (155) describe transport in the velocity space due to momentum transfer at the molecular level as well as transport in the composition space due to molecular diffusion. For a turbulent flow the decay of turbulent fluctuations is caused by these mechanisms. Viscous forces are responsible for the dissipation of turbulent (statistical) fluctuations in velocity. Molecular diffusion is responsible for the dissipation of turbulent fluctuations in scalar quantities.

Before Equation (155) can be solved, these terms must be defined or approximated using models. The following sections will describe with some examples how this can be done.

4.2.2. Limitations of Single-Point Probability Density Functions

Solution of Equation (155) gives the joint probability density function $P(\boldsymbol{U},\Phi;\boldsymbol{x},t)$ of the random vectors $\boldsymbol{U}$ and $\boldsymbol{\Phi}$ at position $\boldsymbol{x}=(x_1,x_2,x_3)$ and time t. This result contains the complete statistical information at a defined position for a given time. Even this information includes only a partial view of the statistical process being considered. The functionals $P(\boldsymbol{U},\Phi)$ described in Section 2.2 or multipoint probability density functions, such as the two-point probability density function $P(\boldsymbol{U}_1,\Phi_1;\boldsymbol{x}_1,t_1,\boldsymbol{U}_2,\ \Phi_2;\boldsymbol{x}_2,t_2)$ which can be derived from them, provide more comprehensive information. The meaning of these multipoint probability density functions can be illustrated by the example shown in Figure 14. The structure of the vortices in the turbulent flow of the jet stirred reactor is dependent on position and time. The "lifetime" and "tracks" of the vortices being transported in the flow are crucial for the chemical reaction. They give an indication of the degree of mixing and the "segregation" of the flow. (Segregation denotes the state of a fluid in which individual fluid parcels

maintain their identities when passing through the reactor and do not exchange properties with the surrounding fluid [24], [46].) Second-order moments (the correlation functions) are used to describe mixing and segregation:

$$R_{\boldsymbol{\Phi},\boldsymbol{\Psi}} = \left\langle \left(\boldsymbol{\Phi}(\boldsymbol{x}_1,t_1) - \left\langle \boldsymbol{\Phi}(\boldsymbol{x}_1,t_1)\right\rangle\right)\left(\boldsymbol{\Psi}(\boldsymbol{x}_2,t_2) - \left\langle \boldsymbol{\Psi}(\boldsymbol{x}_2,t_2)\right\rangle\right)\right\rangle$$
$$= \int\int \left(\boldsymbol{\Phi}(\boldsymbol{x}_1,t_1) - \left\langle \boldsymbol{\Phi}(\boldsymbol{x}_1,t_1)\right\rangle\right)\left(\boldsymbol{\Psi}(\boldsymbol{x}_2,t_2) - \left\langle \boldsymbol{\Psi}(\boldsymbol{x}_2,t_2)\right\rangle\right) \cdot P\left(\boldsymbol{U}_1,\boldsymbol{\Phi}_1;\boldsymbol{x}_1,t_1,\boldsymbol{U}_2,\boldsymbol{\Phi}_2;\boldsymbol{x}_2,t_2\right)\mathrm{d}\boldsymbol{\Phi}\mathrm{d}\boldsymbol{\Psi}$$

For $\boldsymbol{\Phi} = \boldsymbol{\Psi}$ the autocorrelation functions are obtained which provide data about the vortex structure for a physical property of the fluid e.g., the velocity. Time and length scales for the turbulent exchange of the considered property can be derived from the autocorrelation functions [49], [59–61]; these play an important role in the modeling of turbulent processes.

The dimensionality of the transport equations for multipoint probability density functions is proportional to the number of points (Eq. 155). The problem soon becomes unmanageable, even with a small number of scalar quantities and points.

Single-point probability density functions can be derived from probability density functionals or multipoint probability density functions:

$$P(\boldsymbol{\Phi}_1;\boldsymbol{x},t) = \int_{-\infty}^{+\infty} P\left(\boldsymbol{\Phi}_1;\boldsymbol{x}_1,t_1,\boldsymbol{\Phi}_2;\boldsymbol{x}_2,t_2\right)\mathrm{d}\boldsymbol{\Phi}_2 \tag{157}$$

Single-point probability density functions only give limited information about the field of random variables. They give complete statistical information at a given site and time but not joint information about two sites and/or two times. For the example shown in Figure 14 the single-point functions therefore contain no information about the length or time scales of the statistical (turbulent) fluctuations.

This deficiency can be illustrated by a simple example. Suppose that a turbulent fluctuation of a physical variable is in the form of a sine wave of frequency ν and wavelength λ. The statistical character is due to a statistical phase shift ζ which is evenly distributed between 0 and 1. The amplitude of the physical variable is given by

$$\Phi(\boldsymbol{x},t) = \frac{1}{2} + \frac{1}{2}\sin\left(2\pi\nu t + x_1\lambda + \zeta\right) \tag{158}$$

If the amplitude is measured sufficiently often for a given $\boldsymbol{x}, t$ then the probability density function can be obtained from the density of measured points in the sample space (Eq. 115 a). The density of the measured points is given by

$$\left(\frac{\Delta\Phi}{\Delta\zeta}\right)_{\boldsymbol{x},t} \approx \left(\frac{\partial\Phi}{\partial\zeta}\right)_{\boldsymbol{x},t} = \frac{1}{\pi}\left[1 - (2\Phi - 1)^2\right]^{-1/2}$$

The single-point probability density function is thus independent of the frequency ν and the wavelength λ. In addition, information about values such as $\langle(\partial\Phi/\partial x)^2\rangle$ cannot be derived from the single-point probability density function. This value is, however, dependent on the wavelength, $\langle(\partial\Phi/\partial x)^2\rangle = 1/2(\pi^2/\lambda^2)$, compare Equation (158). The measured single-point probability density function gives no information about the spatial and time propagation of the disturbances.

In spite of these limitations, single-point probability density functions are improvements over the handling of statistical variables with the moments of probability density functions.

4.3. Examples of Calculating Probability Density Functions

For further discussions Equation (155) is rewritten as

$$\varrho(\Phi)\frac{\partial P(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t)}{\partial t} + \varrho(\Phi)U_k\frac{\partial P(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t)}{\partial x_k} + \frac{\partial}{\partial U_k}\left[\varrho(\boldsymbol{\Phi})\left\langle A_k|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t)\right] + \frac{\partial}{\partial\Phi_\omega}\left[\varrho(\boldsymbol{\Phi})\left\langle\Theta_\omega|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t)\right] = 0 \tag{159}$$

where A_i and Θ_α are given by Equations (142 a, b). Equation (159) is deterministic, i.e., determined solutions exist for given initial and boundary conditions. The conditional expected values $\langle A_i — \boldsymbol{u},\boldsymbol{\varphi}\rangle$ and $\langle\Theta_\alpha — \boldsymbol{u},\boldsymbol{\varphi}\rangle$ are not statistical values. In the general case they are unknown functions of the variables $\boldsymbol{u}$ and $\boldsymbol{\varphi}$. For the general case (i.e., for fields of statistical variables) model assumptions must be introduced so that

Equation (155) or (159) can be solved. For deterministic systems Equations (155) and (159) can be treated without further assumptions because $\langle A_i — \boldsymbol{u}, \boldsymbol{\varphi}\rangle = A_i\,(\boldsymbol{u}, \boldsymbol{\varphi})$ and $\langle \Theta_\alpha — \boldsymbol{u}, \boldsymbol{\varphi}\rangle = \Theta_\alpha\,(\boldsymbol{u}, \boldsymbol{\varphi})$, respectively.

4.3.1. Solutions for Deterministic Systems

4.3.1.1. Age (Residence Time) Distributions in Chemical Reactors

The individual types of chemical reactors are normally treated as deterministic systems although under industrial conditions flow is not laminar. This simplification is justified if the turbulent time scales for the reactions occurring in the reactor lie several orders of magnitude above that for the turbulent mass or momentum exchange. The reactor contents have then reached the state of molecular mixing, without noticeable chemical conversion having taken place. Distribution functions or probability density functions may be given for the residence time, for example, and can thus indicate the probability for the residence time (the age α) of a volume element in an interval $\mathrm{d}\alpha$ around α; residence time distribution functions are easily derived from Equation (159).

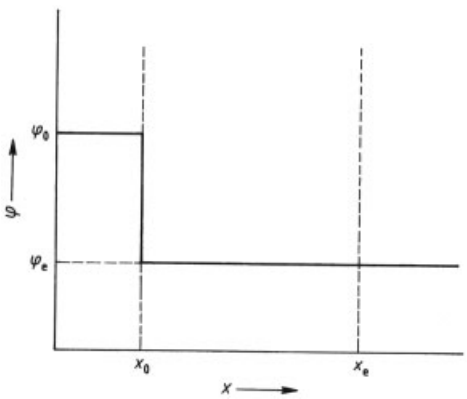

Figure 18. Mathematical definition of an ideally mixed reactor
The subscripts 0 and e denote inlet and outlet, respectively.

Ideally Mixed Stirred-Tank Reactor. In the mathematical sense an ideally mixed reactor is a one-dimensional control volume with homogeneous internal properties. The upstream characteristics are transported into the control volume while the characteristics at the downstream boundary surface relate to the internal properties (Fig. 18). Viscosity, pressure, and gravitational forces and other transport processes (with the exception of convection) are ignored.

For expediency Equation (159) is integrated with respect to the volume:

$$\int_V \varrho\,(\boldsymbol{\Phi})\,\frac{\partial P}{\partial t}\mathrm{d}V + \int_V \varrho\,(\boldsymbol{\Phi})\,u\,\frac{\partial P}{\partial x}\mathrm{d}V + \int_V \frac{\partial}{\partial \Phi_\omega}\left[\varrho\,(\boldsymbol{\Phi})\,S_\omega\,(u, \boldsymbol{\Phi})\,P\right]\mathrm{d}V = 0 \tag{160}$$

Using the Gauss – Ostrogradski integrals, Equation (160) can be rewritten as

$$\frac{\partial}{\partial t}\int_V \varrho\,(\boldsymbol{\Phi})\,P\mathrm{d}V + \oint_F \varrho\,(\boldsymbol{\Phi})\,uP\mathrm{d}F + \int_V \frac{\partial}{\partial \Phi_\omega}\left[\varrho\,(\boldsymbol{\Phi})\,S_\omega\,(u, \boldsymbol{\Phi})\,P\right]\mathrm{d}V = 0 \tag{161}$$

where integration has to be carried out for the surfaces perpendicular to the direction of the flow.

If the age (residence time) distribution is considered, the variables u, Φ in $P\,(u, \Phi)$ and $S\,(u, \Phi)$ must be substituted by the age α. Since intervals in the age space are time intervals, the velocity in the age space is equal to $\partial\alpha/\partial t = S\,(\alpha) = 1$. With this result and for constant density the following expression can be derived from Equation (161):

$$\frac{\partial}{\partial t}P\,(\alpha)\,\mathrm{d}V + uP\,(\alpha)\,\mathrm{d}z\mathrm{d}y - uP_0\,(\alpha)\,\mathrm{d}z\mathrm{d}y + \frac{\partial}{\partial \alpha}P\,(\alpha)\,\mathrm{d}V = 0 \tag{162}$$

In the steady state, Equation (162) is easy to solve since the first term on the left-hand side is zero:

$$\frac{\partial}{\partial \alpha}\left[P\,(\alpha)\right] = \frac{1}{\tau}\left[P_0\,(\alpha) - P\,(\alpha)\right] \tag{163a}$$

where $\tau = u\,\mathrm{d}z\,\mathrm{d}y/\mathrm{d}V$ is the mean hydrodynamic residence time. With the boundary condition $P_0(\alpha) = \delta\,(\alpha - 0)$ and the condition $\int_0^{+\infty} P\,(\alpha)\,\mathrm{d}\alpha = 1$ (cf. Eq. 104) then

$$P\,(\alpha) = \frac{1}{\tau}\exp\left(-\frac{\alpha}{\tau}\right) \tag{164}$$

$\delta(\alpha - 0)$ denotes the impulse function or Dirac delta function (Section 4.1.1).

$$\int_0^{+\infty} \alpha\delta(\alpha - b)\, g(\alpha)\, \mathrm{d}\alpha = a g(b)\,; \quad b \geq 0$$

is used to integrate Equation (163 a). Equation (164) is the familiar residence time distribution function for ideally mixed reactors (denoted as $w(t)$ in Section 3.3.2.2). If it is transformed to normalized age ($\theta = \alpha/\tau$) then

$$P(\theta) \equiv w(\theta) = P(\alpha)\frac{\partial\alpha}{\partial\theta} = \frac{1}{\tau}\exp\left(-\frac{\alpha}{\tau}\right)\tau$$
$$= \exp(-\theta) \tag{165}$$

In Equation (165) the distinction between age α and time t is no longer maintained. The integral over the residence time distribution

$$W(t) = \int_0^{+\infty} w(t)\,\mathrm{d}t \quad \text{or}$$
$$W(\theta) = \int_0^{+\infty} w(\theta)\,\mathrm{d}\theta$$

is denoted as distribution function. Figure 19 shows the residence time distribution and the distribution function for an ideal stirred-tank reactor.

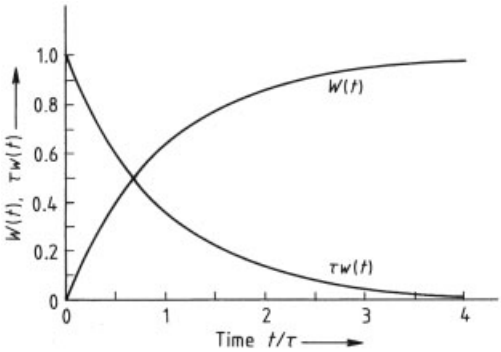

Figure 19. Residence time distribution w (t) and distribution function W (t) for the ideal stirred-tank reactor

Alternative boundary conditions lead to alternative solutions of Equation (163 a). For example, if n ideally mixed reactors are linked in a cascade, then $P_{n-1}(\alpha) = P_{0n}(\alpha)$ etc. With $P_{01}(\alpha) = \delta(\alpha - 0)$ and $\tau_1 = \tau_2 = \cdots \tau_n = \tau/n$, the solution of Equation (163 a) then becomes

$$w_n(t) = \frac{1}{(n-1)!} n^n \frac{t^{(n-1)}}{\tau^n} \exp\left(-n\frac{t}{\tau}\right) \tag{166}$$

Figure 20 shows residence time distribution functions and distribution functions for a cascade of n ideally mixed reactors for various values of n. It is easy to prove from Equation (166) that simple relationships exist between the deterministic moments

$$m_1 = \int_0^{+\infty} t w(t)\,\mathrm{d}t = \bar{t} \quad \text{and}$$
$$m_2' = \int_0^{+\infty} (t - m_1)^2\, w(t)\,\mathrm{d}t$$

(see Section 3.3.2.2) and the number of elements in the cascade. Thus

$$m_1 = n\tau_i = \tau \quad \text{and} \quad m_2' = n\tau_i^2 = \frac{\tau^2}{n} \tag{167a,b}$$

Given that a reactor can be represented by a cascade of ideal stirred-tank reactors, the number of elements required to represent the reactor can easily be determined with Equation (167) from the first two deterministic moments of the measured residence time distribution, namely from the mean residence time and the variance.

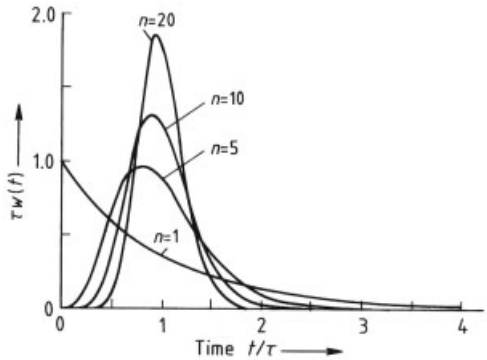

Figure 20. Residence time distribution for a cascade of n ideal stirred-tank reactors

For the nonsteady case, Equation (162) becomes:

$$\frac{\partial}{\partial t}[P(\alpha;t)] + \frac{\partial}{\partial\alpha}[P(\alpha;t)] =$$
$$\frac{1}{\tau}[P_0(\alpha;t) - P(\alpha;t)] \tag{163b}$$

The general solution for Equation (163 b) can easily be found by using Laplace or Fourier

transforms and applying the concept of generalized functions [62], [63]:

$$P(\alpha;t) = \exp\left(-\frac{t}{\tau}\right) P_0(\alpha - t) + \frac{1}{\tau}\exp\left(-\frac{\alpha}{\tau}\right)[H(\alpha) - H(\alpha - t)] \quad (168)$$

The solution holds for the initial conditions $P(\alpha;t)_{t=0} = P(\alpha)$ and the same boundary conditions as above; $H(\alpha)$ denotes the Heaviside or step function (cf. Eq. 116). It can easily be shown that Equation (168) becomes Equation (165) for $t \to \infty$.

Ideally Nonmixed Reactor (Ideal Tubular Reactor). An ideally nonmixed reactor is a onedimensional flow system with inhomogeneous internal properties. Analogously to the ideally mixed reactor, in this one-dimensional model viscosity, pressure, gravitational forces, and other transport processes (excluding convection) are neglected. The changes in the internal properties occur as a result of physical or chemical processes.

The age (residence time) distribution of this model is obtained from Equation (159):

$$\frac{\partial}{\partial t}[P(\alpha;x,t)] + u\frac{\partial}{\partial x}[P(\alpha;x,t)] + \frac{\partial}{\partial \alpha}[P(\alpha;x,t)] = 0 \quad (169a)$$

In the steady state, Equation (169 a) becomes

$$u\frac{\partial}{\partial x}[P(\alpha;x)] + \frac{\partial}{\partial \alpha}[P(\alpha;x)] = 0 \quad (169b)$$

The partial differential Equation (169 b) can be solved in a number of ways. The calculation domain (the tubular reactor) can be divided into a number of finite control volumes, for each of which Equation (169 b) holds. The individual control volumes have the same characteristics as the ideally mixed reactor. The same operations as discussed above can be carried out for each of the finite control volumes and an equation of type (163 b) is obtained for each control volume. Physically this means that an ideally nonmixed reactor is represented by a cascade with an infinite number of elements. Numerically this means that the partial differential Equation (169 b) is solved with an upwind difference scheme. The solution for the boundary condition $P_0(\alpha) = \delta(\alpha - 0)$ is Equation (166) with $n \to \infty$, i.e., an impulse function at $t/n\,\tau_i = 1$.

Equation (169 b) can also be solved analytically, the general solution is

$$P(\alpha;x) = \Omega\left(\alpha - \frac{1}{u}x\right) \quad (170)$$

where Ω is any differentiable function [64]. Equation (170) clarifies the behavior of ideally nonmixed reactors. A fluid flowing into the reactor with a given age distribution retains the form of the age distribution. The age of the individual fluid elements increases uniformly by the time necessary to pass the distance x, $t = x/u$. For example, at the reactor inlet (i.e., $x = 0$)

$$\Omega(\alpha - 0) = \delta(\alpha - 0)$$

at the reactor outlet (i.e., $x = L$)

$$P(\alpha;x) = \delta\left(\alpha - \frac{L}{u}\right) = \delta(\alpha - \tau)$$

This is identical to Equation (166) for $n \to \infty$.

If an ideally mixed reactor with a mean residence time τ_1 and an ideally nonmixed reactor with a mean residence time τ_2 are connected together, then $\Omega(\alpha - 0) = H(\alpha - 0)\exp(-\alpha/\tau_1)$, and from Equation (170)

$$P(\alpha) = H(\alpha - \tau_2)\cdot\frac{1}{\tau_1}\exp\left[-\frac{1}{\tau_1}(\alpha - \tau_2)\right]$$

It is seen that the age distribution of the ideally mixed reactor is shifted by the mean residence time τ_2 of the nonmixed reactor. Figure 21 shows the residence time distribution and the distribution function for this case.

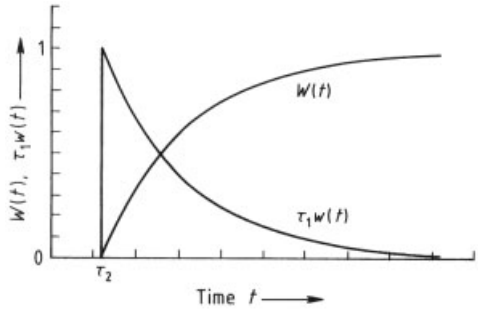

Figure 21. Residence time distribution w (t) and distribution function W (t) for an ideal stirred-tank reactor connected in series with an ideal tubular reactor

Reactors with Finite Mixing. Reactors with finite mixing can be thought of as the transport mechanisms discussed in page 41 combined with another transport mechanism. Usually this is treated in the form of a "diffusion" mechanism in which the relevant diffusion coefficient has the character of a model parameter (dispersion model).

Diffusion is expressed in the transport equation for the probability density function as an additional transport term in the sample space of scalar values (see Eq. 155, last term on the right-hand side and Eq. 159). The basic treatment of these terms is discussed in [19], [65–67], see also Section 4.3.2.1. Here, a heuristic method will be used to find a transport equation for the age distribution in reactors where diffusion is superimposed on convective transport.

The problem consists of deriving a term for the additional transport in the age space for a superimposed diffusion process. In chemical reactors with finite mixing, fluid parcels are convected by turbulent motion, i.e., this process consists of the diffusion of fluid elements which can be handled in the same way as molecular diffusion.

The fluid parcels experience a displacement due to the "diffusion" in addition to the displacement resulting from convective transport with velocity u. Analogous to molecular motion, the probability density function for the displacement by diffusion is a normal distribution [68]:

$$P(x)_{\text{diff}} = \frac{1}{2\sqrt{\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{171}$$

where D is the diffusion coefficient. The probability density function for the displacement x of the fluid elements due to diffusion can be rewritten as the probability density function for the age with the help of the transformation rule given by Equation (106) and $\alpha = x/u$:

$$P(\alpha)_{\text{diff}} = P(x)_{\text{diff}} \cdot \frac{\mathrm{d}x}{\mathrm{d}\alpha} = \frac{u}{2\sqrt{\pi D \alpha}} \cdot \exp\left(-\frac{x^2}{4D\alpha}\right) \tag{172}$$

Equation (172) represents the age distribution for a diffusion process that is analogous to molecular diffusion. The term required for the transport equation is the differential of this distribution in the age space, i.e., $\frac{\partial}{\partial \alpha}\left[P(\alpha)_{\text{diff}}\right]$. Equation (169 b) is now supplemented by the term $\frac{\partial}{\partial \alpha}\left[P(\alpha)_{\text{diff}}\right]$ so that the transport equation for the age distribution in the steady-state case takes the form

$$\frac{\partial}{\partial \alpha}\left[P(\alpha;x)\right] + u\frac{\partial}{\partial x}\left[P(\alpha;x)\right] + D\frac{\partial^2}{\partial x^2}\left[P(\alpha)\right] = 0 \tag{173}$$

Equation (173) can be solved analytically for simple boundary conditions. For the boundary condition

$$\frac{\partial}{\partial x}\left[P(\alpha;x)\right] = 0 \quad \text{for} \quad x = 0 \quad \text{and} \quad x = L$$

this gives

$$P(\alpha) = \frac{1}{2}\sqrt{\frac{Pe}{\pi \tau \alpha}} \exp\left(-\frac{(1-\alpha/\tau)^2\, Pe}{4\alpha/\tau}\right) \tag{174}$$

where Pe is the axial Péclet number, $Pe = u\,L/D$ (Section 3.3.2.2), it is often called the Bodenstein number, Bo.

Figure 22 shows the residence time distributions (Eq. 174) for different Bodenstein numbers. For large Bodenstein numbers the figure resembles the distributions for the cascade of ideally mixed reactors (Fig. 20). This similarity is inevitable, because the cascade of ideally mixed reactors represents the numerical solution of the transport Equation (169 b) as $n \to \infty$ and Equation (173) will be identical to this for $D \to 0$. The deterministic moments m_1 and m_2' can be calculated from Equation (174):

$$m_1 = \left(1 + \frac{2}{Bo}\right)\tau \quad \text{and}$$

$$m_2' = \left(\frac{2}{Bo} + \frac{8}{Bo^2}\right)\tau^2 \tag{175a,b}$$

For $Bo \to \infty$, the first moment corresponds to the mean hydrodynamic residence time (Eq. 167 a). If Equation (175 b) is compared with Equation (167 b) on the basis of the similarity of the solutions of the age distribution for large Bodenstein numbers, then the relationship

$$\frac{1}{n} = \frac{2}{Bo} + \frac{8}{Bo^2}$$

can be derived. For large Bodenstein numbers (i.e., little axial dispersion), $n \approx Bo/2$.

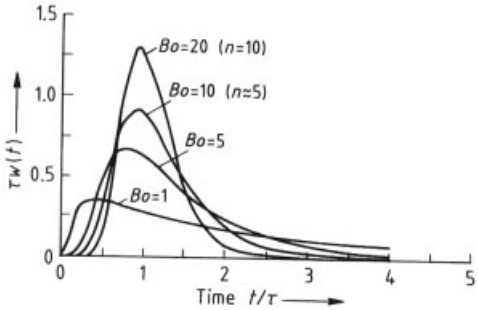

Figure 22. Residence time distribution $\tau\, w$ (t) for a reactor with axial mixing
Bo = Bodenstein number; *n* = number of elements of a cascade of ideally mixed reactors to represent the reactor with the corresponding Bodenstein number

This discussion shows that the definition or experimental determination of model parameters can be relatively simple even for complex models. In the above case the formal diffusion coefficient, the axial mixing coefficient, or the Bodenstein number can be simply determined from the second deterministic moment of the age (residence time) distribution. Furthermore, complex processes may be built up from a number of simple elements. In the case described above, a reactor with finite axial mixing is constructed from a cascade of ideally mixed reactors; a simple relationship exists between the number of elements and the model parameters of this complex one-dimensional process. Further examples for the representation of complex processes by simple elements are found in [4], [11], [12], [24], [46], and text books on chemical reaction engineering, see also Chapter 5.

4.3.1.2. Size Distribution in Continuously Operating Crystallizers

The model used for the representation of a continuously operating crystallizer is an ideally mixed reactor. The supersaturated solution is fed into the crystallizer and the substrate grows on the existing crystals. Using the operations discussed in page 41 this gives the transport equation

$$\frac{\partial}{\partial \Phi}\left[S\left(\Phi\right) P\left(\Phi\right)\right] = \frac{1}{\tau}\left[P_0\left(\Phi\right) - P\left(\Phi\right)\right] \tag{176}$$

The property under consideration is the crystal size L. The rate of crystal growth can be described empirically by $S\,(L) = S\,(L, c)$, i.e., it is a function of the crystal size and the concentration of the substrate c. Often $S\,(L) = S_0(c)L^{\mathrm{b}}$ is used so that Equation (176) becomes

$$\frac{\partial}{\partial L}\left[S_0\left(c\right) L^b P\left(L\right)\right] = \frac{1}{\tau}\left[P_0\left(L\right) - P\left(L\right)\right] \tag{177}$$

If the feed does not contain solids, $P_0(L) = \delta\,(L - 0)$. Assuming that the substrate concentration is approximately constant, the solution of Equation (177) is

$$P\left(L\right) = C_1 L^{-b} \exp\left[-\frac{L^{1-b}}{S_0\left(c\right)\tau\left(1-b\right)}\right] \tag{178}$$

where C_1 is an integration constant. For $b = 0$ (McCabe ΔL law) Equation (178) becomes

$$P\left(L\right) = P\left(L_0\right) \exp\left(-\frac{L}{S_0\left(c\right)\tau}\right) \tag{179}$$

where C_1 is substituted by the initial condition $P\,(L_0)$, i.e., the number density of the nuclei in the crystallizer. Equation (179) represents the number density distribution of crystal sizes for this simple case ($b = 0$).

4.3.2. Solutions for Statistical Systems

The discussion in Section 4.3.1 was confined to deterministic systems for which either the physical quantities are not subject to turbulent (statistical) fluctuations or for which the time scales of the individual physicochemical processes are of vastly different orders of magnitude. In the ideally mixed reactor the time scale for the chemical reactions is so large in comparison with that of the turbulent mass exchange that the reactor contents can be considered as being mixed at the molecular level without having reacted. The point of this discussion was to bring the transport equations for the probability density functions (Eqs. 155, 159) into a simple form which is closed without any model assumptions and is also easy to solve analytically in simple cases.

These simplifications do not apply to statistical systems, such as the chemical reactions in a turbulent single-phase flow shown in Figure 14. The transport equations for the probability density function in the combined velocity – composition space have advantages

for describing such processes because, in contrast to the simple equations for turbulent reacting flows (see Section 5.1.2), the most important processes (convection and chemical reaction) can be handled without model assumptions. Assumptions must be introduced for the dissipation of turbulent velocity fluctuations or fluctuations of scalar quantities by molecular mixing. These are manifested in the transport equations for probability density functions as transport in the velocity or composition space.

A method of solving the transport equations must also be found. Analytical solutions can only be obtained for very simple flows, simple models, or deterministic systems under simple conditions (Section 4.3.1). In general, Equation (155) must be solved numerically with the relevant terms on the right-hand side being modeled. Methods using finite differences (see Section 4.3.1.1) can be ruled out for this since $P(\boldsymbol{U}, \Phi; \boldsymbol{x}, t)$ is a function of $(\sigma+6)$ independent variables and time. If, for example, a case was investigated which required 20 finite volumes in the direction of each of these variables, then the number of grid points for a statistical two-dimensional flow with only two chemical species under isothermal conditions would be $20^7 \approx 10^9$. Other methods must therefore be used.

4.3.2.1. Closure of the Transport Equation for Probability Density Functions

The equation for the derivation of the joint probability density function for a turbulent reacting flow is given by

$$\varrho(\boldsymbol{\Phi})\frac{\partial P}{\partial t} + \varrho(\boldsymbol{\Phi})U_k\frac{\partial P}{\partial x_k} + \left[\varrho(\boldsymbol{\Phi})g_k - \frac{\partial\langle p\rangle}{\partial x_k}\right]\frac{\partial P}{\partial U_k} + \frac{\partial}{\partial\Phi_\omega}\left[(\varrho(\boldsymbol{\Phi})S_\omega(\boldsymbol{\Phi})P)\right] = -\frac{\partial}{\partial U_k}\left[\left\langle\frac{\partial\tau_{lk}}{\partial x_l} + \frac{\partial p^{\mathrm{f}}}{\partial x_l}|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right] + \frac{\partial}{\partial\Phi_\omega}\left[\left\langle\frac{\partial J_k^\omega}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right] \tag{155}$$

All the terms on the right-hand side of the equation (convection in physical space, transport in the velocity space due to gravitational forces, forces due to the gradient of the average pressure, and transport in the composition space due to chemical reactions) are closed and do not require any model assumptions.

The term $\varrho(\boldsymbol{\Phi})$ is a known function (e.g., a form of thermal equation of state, Eq. 136) and $S_\alpha(\boldsymbol{\Phi})$ is given by Equation (137 a). For example, for a bimolecular reaction $\mathrm{A}+\mathrm{B}\rightarrow\mathrm{C}+\mathrm{D}$ this relationship is $\mathrm{S}(\mathrm{A}) = r_\mathrm{A} = -k\,m_\mathrm{A}\,m_\mathrm{B}$ where the rate coefficient k can often be represented by three parameters in the form $k = k_0\,T^\alpha\exp(-E_\mathrm{a}/RT)$. If the equations for the reaction rate and the rate coefficients are not considered to be a model, then the representation of the reaction rates in Equation (155) is free of models.

No explicit equation for the mean pressure is contained in the system of Equations (131) – (137). The momentum conservation equations (Eq. 132) contain only the pressure gradient in each of the directions under consideration. Equations (131) – (137) assume a known pressure field which is given implicitly by the conservation equations for the components of momentum and the continuity equation. The mean pressure can be determined from a Poisson equation which is derived from the conservation equations for the momentum components and the continuity equation. Equation (132) can be rewritten as

$$\frac{\partial\varrho U_i}{\partial t} + \frac{\partial}{\partial x_k}(\varrho U_i U_k) = -\frac{\partial\langle p\rangle}{\partial x_i} + \varrho A_i' \tag{180}$$

with

$$A_i' \equiv \frac{\partial\tau_{ik}}{x_k} - \frac{\partial p^{\mathrm{f}}}{\partial x_i} + \varrho g_i$$

If Equation (180) is averaged and differentiated with respect to x_k and the continuity equation differentiated with respect to t

$$\frac{\partial}{\partial x_k}\frac{\partial}{\partial t}\left(\langle\varrho U_k\rangle\right) = -\frac{\partial^2\langle\varrho\rangle}{\partial t^2} \tag{181}$$

is subtracted, then the Poisson equation for the mean pressure is obtained which has to be solved simultaneously to the equation for the probability density function:

$$\frac{\partial^2\langle p\rangle}{\partial x_k\partial x_k} = \frac{\partial^2\langle\varrho\rangle}{\partial t^2} - \frac{\partial^2\langle\varrho U_k U_k\rangle}{\partial x_k x_k} + \frac{\partial\langle\varrho A_k'\rangle}{\partial x_k} \tag{182}$$

In the next step the model assumptions for the conditional expected values on the right-hand

side of Equation (155) must be discussed, in particular for

$$\frac{\partial}{\partial U_k}\left[\left\langle \frac{\partial \tau_{ik}}{\partial x_k} + \frac{\partial p^{\mathrm{f}}}{\partial x_k} | \boldsymbol{u}, \boldsymbol{\varphi} \right\rangle P\right] \text{ and}$$
$$\frac{\partial}{\partial \Phi_\alpha}\left[\left\langle \frac{\partial J_k^\alpha}{\partial x_k} | \boldsymbol{u}, \boldsymbol{\varphi} \right\rangle P\right]$$

Modeling of these expected conditional values is the main problem in developing mathematical models for chemically reacting turbulent flows on the basis of transport equations for probability density functions. The principles used in these methods will be discussed here, for further details see [18], [19], [70], [71], [79], [80].

The probability density functions in the velocity space and the composition space will be treated separately. They can be calculated by integrating Equation (155) over the velocity space and composition space or by using the methods shown in Section 4.2.1 for the momentum components or the scalar quantities. For further simplification a turbulent system is considered with constant density, homogeneous turbulence, and no gradients of the mean velocity.

A uniform fluid flow which flows through a turbulence-generating grid represents a good approximation to these simplified assumptions. If the coordinate system moves with the flow at the mean flow velocity, then no relative convection occurs. The probability density function becomes

$$\frac{\partial P(\boldsymbol{U};t)}{\partial t} = -\frac{\partial}{\partial U_k}\left[\left\langle \left(\mu \frac{\partial^2 U_k}{\partial x_k^2} - \frac{\partial p^{\mathrm{f}}}{\partial x_k}\right) | \boldsymbol{u} \right\rangle P(\boldsymbol{U};t)\right] \quad (183)$$

Equation (183) represents the time development of the probability density function (μ is the dynamic viscosity). The only term to be modeled is

$$\frac{\partial}{\partial U_k}\left[\left\langle \mu \frac{\partial^2 U_k}{\partial x_k^2} - \frac{\partial p^{\mathrm{f}}}{\partial x_k} | \boldsymbol{u} \right\rangle P(\boldsymbol{U};t)\right]$$

Experimental observations in a turbulent flow of this type show that the turbulent velocity fluctuations decay with time. Further, if the eddy dissipation model outlined in 4 is used, it can be shown that the turbulent momentum exchange is an isotropic diffusion process of the turbulent eddies [49] that is rate determining for the dissipation of the velocity fluctuations at the molecular level. Similarly to the diffusion process examined in page 43, this model gives a normal distribution for the velocity (a three-dimensional distribution in this case).

A normal distribution is precisely defined by the mean value and the variances of the quantities concerned. Since the normal distribution being discussed is three dimensional, the variances and covariances of the velocities are important and, using Equations (112) and (113), can be expressed as $\langle U^{\mathrm{f}}{}_i U^{\mathrm{f}}{}_j \rangle$ (Reynolds stresses). The problem now is to find models for the Reynolds stresses and their decay over time which will guarantee the required normal distribution as the solution for Equation (183).

A frequently used model for the Reynolds stresses (see also Section 5.1.2.1) which satisfies these conditions, is that developed by Lumley [72] and Launder [73] and modified by Rotta [49], [74]:

$$\frac{\partial}{\partial t}\langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle = \left[-\langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle - C_{\langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle}\left(\langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle - \frac{2}{3}k\right)\delta_{ij}\right]/\tau \quad (184)$$

where δ_{ij} is the Kronecker delta, $\delta_{ij} = 1$ for $i = j$, $\delta_{ij} = 0$ for $i \neq j$; k is the specific kinetic energy of the turbulent velocity fluctuations, $k = 1/2\,\langle U_k^{\mathrm{f}} \mathrm{U}_k^{\mathrm{f}} \rangle$; and τ is the turbulent time scale, which has to be defined from model assumptions.

The first term on the right-hand side of Equation (184) causes exponential decay of the Reynolds stresses. The second term is responsible for the isotropy of the "diffusion" of the kinetic energy of the turbulent fluctuations.

The second problem for the solution of Equation (183) is the modeling of the conditional expected values

$$\left\langle \left(\frac{\partial p^{\mathrm{f}}}{\partial x_i}\right) | \boldsymbol{u} \right\rangle$$

A Poisson equation for the pressure variations can be derived from the conservation equations for the momentum components and the continuity equation in a similar manner to the mean pressure:

$$\frac{\partial^2 p^{\mathrm{f}}}{\partial x_k \partial x_k} = -2\varrho \frac{\partial \langle U_k \rangle}{\partial x_l} \cdot \frac{\partial U_l^{\mathrm{f}}}{\partial x_k} - \varrho \frac{\partial^2 U_k^{\mathrm{f}} U_l^{\mathrm{f}}}{\partial x_k \partial x_l} \quad (185)$$

Thus the pressure fluctuations have two sources. The first is the interactions of the turbulent fluctuations themselves (second term of the right-hand side) and the interactions of the turbulent fluctuations with the gradients of the mean velocities. The second term is described by the equation for the Reynolds stressses (Eq. 184); the first term ("rapid pressure") must be modeled, this can also be carried out with the help of the Reynolds stresses [19], [66].

The third problem is the description of the conditional expected values

$$\frac{\partial}{\partial \Phi_\alpha}\left[\left\langle \frac{\partial J_k^\alpha}{\partial x_k} | \boldsymbol{u}, \boldsymbol{\varphi} \right\rangle P\right]$$

These terms can be modeled by using the simplifications described above. For the same homogeneous system at constant density with isotropic turbulence and a moving coordinate system, the change in the probability density function over time for a passive scalar quantity, e.g., the mass fraction of a nonreacting chemical species is given by:

$$\frac{\partial P(\Phi;t)}{\partial t} = \frac{\partial}{\partial \Phi}\left[\left\langle \frac{\partial J_k^\Phi}{\partial x_k} | \varphi \right\rangle P(\Phi;t)\right]$$
$$= \frac{\partial}{\partial \Phi}\left[\left\langle \frac{D}{\varrho}\frac{\partial^2 \Phi}{\partial x_k^2} | \varphi \right\rangle P(\Phi;t)\right] \tag{186}$$

where D is the diffusion coefficient. Equation (186) is obtained for the nonreacting scalar Φ from the general transport equation by integrating over the velocity space or by using the method given in Section 4.2.1. As before with the dissipation of velocity fluctuations, the eddy transport model allows the turbulent exchange of mass to be considered as "eddy diffusion". This "diffusion" is rate determining for the dissipation of the turbulent fluctuations of the scalar at the molecular level. The term

$$\frac{\partial}{\partial \Phi}\left[\left\langle \frac{D}{\varrho}\frac{\partial^2 \Phi}{\partial x_k^2} | \varphi \right\rangle P(\Phi;t)\right]$$

must be modeled in such a way that the solution of Equation (186) is a normal distribution which decays over time. Thus, the problem consists of modeling the second moment $\langle \Phi^{f2} \rangle$ (the variance of the scalar quantities).

A simple deterministic model [75] which satisfies the above requirements is

$$\left\langle \left(\frac{D}{\varrho}\frac{\partial^2 \Phi}{\partial x_k^2}\right) | \varphi \right\rangle = -\frac{1}{2} C_\Phi \left(\Phi - \langle \Phi \rangle\right)\frac{1}{\tau} \tag{187}$$

in which τ is the turbulent time scale $\tau / C_\Phi = \tau_\Phi$. Since diffusion transport is in the composition space, Equation (187) can be interpreted as transport velocity in the composition space which is proportional to the deviation from the mean value. However, model Equation (187) does not yield a normal distribution from all initial conditions. Instead, the initial conditions must contain information about the form of the distribution [19] as in the applications described in [57], [76], [77]. A general equation can be derived with a statistical particle interaction model, for details refer to [19].

The variances of Φ, $\langle \Phi^{f2} \rangle$, according to Equation (187) are described by an equation similar to Equation (184):

$$\frac{\partial}{\partial t}\langle \Phi^{f2} \rangle = -C_{\langle \Phi^{f2} \rangle}\langle \Phi^{f2} \rangle / \tau \tag{188}$$

For the combined velocity – composition space of the homogeneous system with isotropic turbulence the same line of reasoning can be applied. The joint probability density function $P(\boldsymbol{U}, \Phi; t)$ is normally distributed so that it is defined by the mean values and the second moments $\langle U_i^f U_j^f \rangle$, $\langle \Phi_\alpha^f \Phi_\beta^f \rangle$, and $\langle U_i^f \Phi_\alpha^f \rangle$. The latter quantities (the Reynolds fluxes) are usually defined by a model equation analogous to Equations (188) and (184):

$$\frac{\partial}{\partial t}\langle U_i^f \Phi_\alpha^f \rangle = -C_{\langle U_i^f \Phi_\alpha^f \rangle}\langle U_i^f \Phi_\alpha^f \rangle / \tau \tag{189}$$

which does not change the character of the solution for the joint probability density function. In the combined velocity – composition space, the diffusion of the Reynolds stresses $\langle U_i^f U_j^f \rangle$, the fluctuations of scalar quantities $\langle \Phi_\alpha^{f2} \rangle$, and the Reynolds fluxes $\langle U_i^f \Phi_\alpha^f \rangle$ are all superimposed on one another. The remaining problem is then the modeling of the turbulent time scales for these "diffusion" processes and the treatment of the "rapid pressure". Some ideas for modeling these are discussed in Sections 5.1 and 5.3.

Under industrial conditions, chemical reactions in turbulent flows occur in nonhomogeneous systems that usually involve gradients of the mean velocities and nonisotropic turbulence.

The extension of the simple model described above to such cases is still being developed [18],

[19]. The general principle is that the main properties of the model for the second moments are retained. The additional effects are treated as being superimposed on the undisturbed "diffusion". Often "hybrid" models are also used where the expected values of the velocities are computed by other methods (cf. Section 5.1.2.3) and only the joint probability density function for the composition space is evaluated (cf. Section 4.3.2.3).

4.3.2.2. Solution Methods of the Transport Equation for Probability Density Functions

Equation (155) (Section 4.2.1)

$$\varrho(\boldsymbol{\Phi})\frac{\partial P}{\partial t}+\varrho(\boldsymbol{\Phi})U_k\frac{\partial P}{\partial x_k}+\left[\varrho(\boldsymbol{\Phi})g_k-\frac{\partial\langle p\rangle}{\partial x_k}\right]\frac{\partial P}{\partial U_k}+\frac{\partial}{\partial\Phi_\omega}\left[(\varrho(\boldsymbol{\Phi})S_\omega(\boldsymbol{\Phi})P)\right]=-\frac{\partial}{\partial U_k}\left[\left\langle\left(\frac{\partial\tau_{lk}}{\partial x_l}+\frac{\partial p^{\mathrm{f}}}{\partial x_l}\right)|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right]+\frac{\partial}{\partial\Phi_\omega}\left[\left\langle\frac{\partial J_k^\omega}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\right] \tag{155}$$

is a partial differential equation in $(\sigma+6)$ independent variables and time. Finite difference methods for solving this equation fail due to the large number of variables if drastic simplifications cannot be applied. Successful methods of solution have been based on Monte Carlo simulations of the discrete form of the probability density functions in the velocity – composition space [19]. These methods will be described briefly before an example is discussed.

Equation (155) can be rewritten as

$$\frac{\partial P}{\partial t}=(\Omega_1+\Omega_2+\Omega_3)P \tag{190}$$

in which Ω_1, Ω_2, and Ω_3 are operators that are obtained by comparison with Equation (155):

$$\Omega_1\equiv-g_k\frac{\partial}{\partial U_k}-S_\omega\frac{\partial}{\partial\Phi_\omega}-\frac{\partial S_\omega}{\partial\Phi_\omega}I \tag{191a}$$

$$\Omega_2\equiv[\ldots] \tag{191b}$$

$$\Omega_3\equiv U_k\frac{\partial}{\partial x_k}+\frac{1}{\varrho(\boldsymbol{\Phi})}\frac{\partial\langle p\rangle}{\partial x_k}\frac{\partial}{\partial U_k} \tag{191c}$$

In Equation (190) P denotes $P(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t)$ and in Equation (191 a) I is the identity operator. The terms represented by dots in Equation (191 b) symbolize the models used for each of the "diffusion" processes, i.e., the types of models to describe the conditional expected values on the right-hand side of Equation (155). Equation (190) indicates that all three operators influence the probability density function simultaneously. However, for a solution the operators are separated.

A first-order approximation $P^{(1)}(t)$ for $P(t)$ starting from given initial conditions $P(t_0)$ is estimated by calculating the separate effects of the three operators in the following way:

$$P_1^{(1)}(t)=(I+\Delta t\Omega_1)P^{(1)}(t) \tag{192a}$$

$$P_2^{(1)}(t)=(I+\Delta t\Omega_2)P_1^{(1)}(t) \tag{192b}$$

$$P^{(1)}(t+\Delta t)=(I+\Delta t\Omega_3)P_2^{(1)}(t) \tag{192c}$$

If the fractional changes $P_1^{(1)}(t)$ and $P_2^{(1)}(t)$ are eliminated backwards, then $P^{(1)}(t+\Delta t)$ can be expressed by $P^{(1)}(t)$:

$$\begin{aligned}P^{(1)}(t+\Delta t)&=(I+\Delta t\Omega_3)(I+\Delta t\Omega_2)(I+\Delta t\Omega_1)P^{(1)}(t)\\&=P^{(1)}(t)+\Delta t(\Omega_1+\Omega_2+\Omega_3)P^{(1)}(t)+O(\Delta t^2)\end{aligned} \tag{193a}$$

On the other hand if $P^{(1)}(t)$ is developed as a Taylor series, then

$$P^{(1)}(t+\Delta t)=P^{(1)}(t)+\frac{\partial P^{(1)}(t)}{\partial t}\Delta t+\frac{\partial^2P^{(1)}(t)}{\partial t^2}\frac{\Delta t^2}{2!}+\ldots=P^{(1)}(t)+\frac{\partial P^{(1)}(t)}{\partial t}\Delta t+O(\Delta t^2) \tag{193b}$$

If Equations (193 a) and (193 b) are set equal to each other and divided by Δt then

$$\frac{\partial P^{(1)}(t)}{\partial t}=(\Omega_1+\Omega_2+\Omega_3)P^{(1)}(t)+O(\Delta t^2) \tag{194}$$

Equation (194) is identical to Equation (190) apart from the development errors of the order $O(\Delta t^2)$. Thus this procedure of separating the operators gives a first-order approximation for $P(t)$.

The Monte Carlo simulation of the development of the probability density function uses the discrete form of the probability density function with n statistical fluid particles:

$$P_n\left(\boldsymbol{U},\boldsymbol{\Phi};\boldsymbol{x},t\right) = \frac{1}{n}\sum_{i=1}^{n}\delta\left(\boldsymbol{U}-\boldsymbol{u}_i\right)\delta\left(\boldsymbol{\Phi}-\boldsymbol{\varphi}_i\right)\delta\left(\boldsymbol{x}-\boldsymbol{x}_i\right) \qquad (195)$$

(cf. Eq. 115 b, Section 4.1.1). The discrete form of the probability density function converges with $\sqrt{n}$ towards $P(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t)$. The numerical problem is to determine the movement of n statistical fluid particles in the physical, velocity, and composition spaces consistent with the operators Ω_1, Ω_2, and Ω_3. Stated in another way: starting from a given initial distribution $P_n(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t=0)$, n fluid particles with the properties $\boldsymbol{U}_i(t_0)$, $\boldsymbol{\Phi}_i(t_0)$, and $\boldsymbol{x}_i(t_0)$ are chosen at random. The fractional changes of P_n are then calculated which lead to $P_n(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t=t+\Delta t)$ with the particle properties $\boldsymbol{U}_i(t=t_0+\Delta t)$, $\boldsymbol{\Phi}_i(t=t_0+\Delta t)$, and $\boldsymbol{x}_i(t=t_0+\Delta t)$. This gives an algorithm for calculating the development of $P_n(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t)$ over time by using forward differences with respect to time. The movement equations consistent with the operators Ω_1, Ω_2, and Ω_3 are for the third fractional change Ω_3

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = -\frac{1}{\varrho\left(\boldsymbol{\Phi}(t)\right)}\left[\frac{\partial\langle p\rangle}{\partial x_k}\right]_{x(t)}, \quad \frac{\mathrm{d}\Phi_\omega}{\mathrm{d}t} = 0, \quad \text{and}$$
$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{U}(t) \qquad (196a,b,c)$$

According to Equation (191 a), for the operator Ω_1 and thus for the first fractional change:

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = g_k + [\ldots], \quad \frac{\mathrm{d}\Phi_\omega}{\mathrm{d}t} = S_\omega\left[\boldsymbol{\Phi}(t)\right], \quad \text{and}$$
$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = 0 \qquad (197a,b,c)$$

In Equation (197 a) the contents of the square brackets indicated by dots must be substituted by the model being used for describing the dissipation of the turbulent velocity fluctuations. The second operator only causes a fractional change in the composition space so that the movement equations for the second fractional change are

$$\frac{\mathrm{d}\boldsymbol{U}}{\mathrm{d}t} = 0, \quad \frac{\mathrm{d}\Phi_\omega}{\mathrm{d}t} = [\ldots], \quad \text{and} \quad \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = 0 \qquad (198a,b,c,)$$

The movement of the statistical fluid particles in the composition space (Eq. 198 b) must again be obtained from the model being used.

After i time steps, each with three fractional changes $P_n(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t=t+i\,\Delta t)$ is obtained in discrete form and the remaining task is to calculate the mean values, variances, covariances, and all values of interest from the discrete probability density function. This may be achieved by the averaging procedure according to Equation (114) and the analogous rules for higher moments. However, for discrete probability density functions, special interpolation routines are recommended for calculating the moments [19].

4.3.2.3. Example: Combustion of Propane in a Turbulent Diffusion Flame

The solution of Equation (155) for turbulent reacting flows gives the joint probability density function $P(\boldsymbol{U}, \boldsymbol{\Phi}; \boldsymbol{x}, t)$ in discrete form according to the methods outlined in Section 4.3.2.2. All the information required for engineering applications can be obtained from this complete statistical description.

To solve Equation (155) a number of model parameters must be determined (see Eqs. 184, 188, or 189 for simple systems) as well as the turbulent time scale τ. Equation (155) contains $(\sigma+6)$ independent variables so a finite difference method of solution is ruled out. Even if the method of solution shown in Section 4.3.2.2 overcomes these problems, a reduction in the number of variables by a simplification of Equation (155) is worthwhile. In most applications this is achieved by solving Equation (155) for the composition space only [57], [70], [71], [78–80]. The result of this is that, besides the turbulent time scale τ, the expected values of the velocities must also be calculated by other models (Section 5.1.2).

The combustion of propane with air in a turbulent diffusion flame will be described as an example. The propane flows from a simple jet assembly with a Reynolds number of 40 000 related to the nozzle outlet. The gas is mixed with the surrounding air and burns in a turbulent diffusion flame. This example may resemble an industrial flue gas flare.

The numerical calculation is confined to the probability density function in the composition space [70]. Integration of Equation (155) over the velocity space according to Equation (121) gives

$$\frac{\partial\left[\varrho\left(\boldsymbol{\Phi}\right)P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right]}{\partial t}+\frac{\partial}{\partial x_k}\left[\varrho\left(\boldsymbol{\Phi}\right)\left\langle U_k|\boldsymbol{\varphi}\right\rangle\right.$$

$$\cdot P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)]+\frac{\partial}{\partial\Phi_\omega}\left[\left(\varrho\left(\boldsymbol{\Phi}\right)S_\omega\left(\boldsymbol{\Phi}\right)P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right)\right]=$$

$$\left|\frac{\partial}{\partial\Phi_\omega}\left[\left\langle\frac{\partial J_k^\omega}{\partial x_k}|\boldsymbol{u},\boldsymbol{\varphi}\right\rangle P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right]\right] \tag{199}$$

By using density-weighted variables (indicated by a tilde) [81]

$$\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)=\frac{1}{\langle\varrho\rangle}\int\varrho\left(\boldsymbol{\Phi}\right)P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\mathrm{d}\varrho \tag{200}$$

and decomposing the velocities according to Equation (112)

$$U_i|\varphi=\langle\tilde{U}_i\rangle+\langle\tilde{U}_i^{\mathrm{f}}|\boldsymbol{\varphi}\rangle \tag{201}$$

the following is obtained:

$$\langle\varrho\rangle\frac{\partial\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)}{\partial t}+\frac{\partial}{\partial x_k}\left[\langle\varrho\rangle\langle\tilde{U}_k\rangle\right.$$

$$\cdot P\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)]+\frac{\partial}{\partial\Phi_\omega}\left[\langle\varrho\rangle S_\omega\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right]$$

$$=\frac{\partial}{\partial\Phi_\omega}\left[\left\langle\frac{1}{\varrho}\frac{\partial J_k^\omega}{\partial x_k}|\boldsymbol{\varphi}\right\rangle\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right]$$

$$-\frac{\partial}{\partial x_k}\left[\langle\tilde{U}_k^{\mathrm{f}}|\boldsymbol{\varphi}\rangle\langle\varrho\rangle\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)\right] \tag{202}$$

The problem in using Equation (202) is the unknown velocity $\langle\boldsymbol{U}\rangle$ and the modeling of the terms for the dissipation of turbulent fluctuations of the scalar quantities at the molecular level (first term on the right-hand side of Eq. 202) and transport in the physical space by turbulent velocity fluctuations (second term on the right-hand side of Eq. 202). For the latter term a gradient transport approach is used [70] that describes the transport of higher moments analogously to diffusion. The background of such equations is briefly treated in Section 5.1. In the special case of propane combustion:

$$\langle\tilde{U}_k^{\mathrm{f}}|\varphi\rangle\langle\varrho\rangle\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right)$$

$$\approx\ -C_S\langle\varrho\rangle\frac{\tilde{k}}{\tilde{\varepsilon}}\langle U_k^{\mathrm{f}}U_l^{\mathrm{f}}\rangle\frac{\partial}{\partial x_l}\tilde{P}\left(\boldsymbol{\Phi};\boldsymbol{x},t\right) \tag{203}$$

where C_{s} is a model parameter; $\tilde{k}$ is the turbulence energy (see Section 4.3.2.1); $\tilde{\varepsilon}=-\mathrm{d}\tilde{k}/\mathrm{d}t$ such that $\tilde{k}/\tilde{\varepsilon}\equiv\tau$ the turbulent time scale; and finally $\langle U_i^{\mathrm{f}}\mathrm{U}_j^{\mathrm{f}}\rangle$ are the Reynolds stresses. The last three quantities are unknown and must be provided before the solution of Equation (202). In [70] this occurs together with the calculation of the expected velocity values using a "second-order closure model". In this procedure the conservation equations are averaged and the correlations which arise from averaging (the Reynold stresses and Reynolds fluxes) are modeled directly (see also Section 5.1). The remaining problem is the modeling of the dissipation of the turbulent fluctuations of the scalar quantities at the molecular level, this is taken care of with information about the form of the distribution using the method described in Section 4.3.2.1 [57], [71], [76], [78].

The combustion of propane with air in a diffusion flame, comprises ca. 300 elementary reactions between ca. 45 chemical species [82]. The composition space is therefore approximately 45-dimensional and numeric simulation using the Monte Carlo method is prohibitive due to the high dimensionality of the problem. The reaction mechanism for the combustion of propane may be reduced to [70]

$$\mathrm{C_3H_8+3/2\,O_2\rightarrow 3\,CO+4\,H_2} \tag{i}$$

$$\mathrm{CO+OH\rightleftharpoons CO_2+H} \tag{ii}$$

and the recombination reactions of the $\mathrm{H_2-O_2}$ system

$$\mathrm{H+O_2+M\rightleftharpoons HO_2+M} \tag{iii}$$

$$\mathrm{H+H+M\rightleftharpoons H_2+M} \tag{iv}$$

$$\mathrm{H+OH+M\rightleftharpoons H_2O+M} \tag{v}$$

$$\mathrm{H+O+M\rightleftharpoons OH+M} \tag{vi}$$

where M denotes a third body collision partner.

The chemical system is thus described by four scalar variables; the normalized elemental mass fraction of carbon $\Phi_1\equiv f$, the mixture fraction; the total molar concentration $\Phi_2\equiv c_{\mathrm{tot}}$; the molar concentration of propane $\Phi_3\equiv c_{\mathrm{C_3H_8}}$; and the molar concentration of carbon monoxide $\Phi_4\equiv c_{\mathrm{CO}}$. This manipulation reduces the number of variables in Equation (202) to four. The reaction rates of these four variables are derived from the reduced mechanism:

$$S(\Phi_1) = 0$$
$$S(\Phi_2) = 4\frac{1}{2}r_{(\mathrm{i})} - r_{(\mathrm{ii})} - r_{(\mathrm{iv})} - r_{(\mathrm{v})} - r_{(\mathrm{vi})}$$
$$S(\Phi_3) = -r_{(\mathrm{i})}$$
$$S(\Phi_4) = 3r_{(\mathrm{i})} - r_{(\mathrm{iv})}$$

Figures 23 and 24 show the results of the numeric simulation according to [70]. The results given in Figure 23 are the joint probability density function of f at a distance of 10 jet diameters from the fuel nozzle. There is a peak near the beginning of the coordinates which means that in the region near the fuel nozzle, the air from the surrounding fluid entrains the flow. Figure 24 A gives radial profiles of the mean values of the propane mole fraction and the mixture fraction at a distance of 40 mm from the burner nozzle. These results are important for setting up an exhaust gas flare and can be derived from the general probability density function. There is relatively good correlation between the calculated and measured values. The correlation between the combustion products (CO and CO_2) is not as good (Fig. 24 B). This is due to the models contained in the transport equation for the joint probability density function and the chemical model which contains drastic assumptions about the concentrations of radicals formed during propane combustion (Eqs. i – vi).

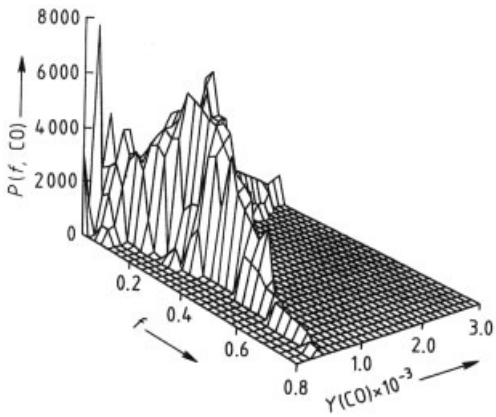

Figure 23. Joint probability density function for the normalized elemental mass fraction (mixture fraction) and mass fraction of CO according to [70] for a propane air diffusion flame at 10 nozzle diameters above the burner nozzle
f = mixture fraction

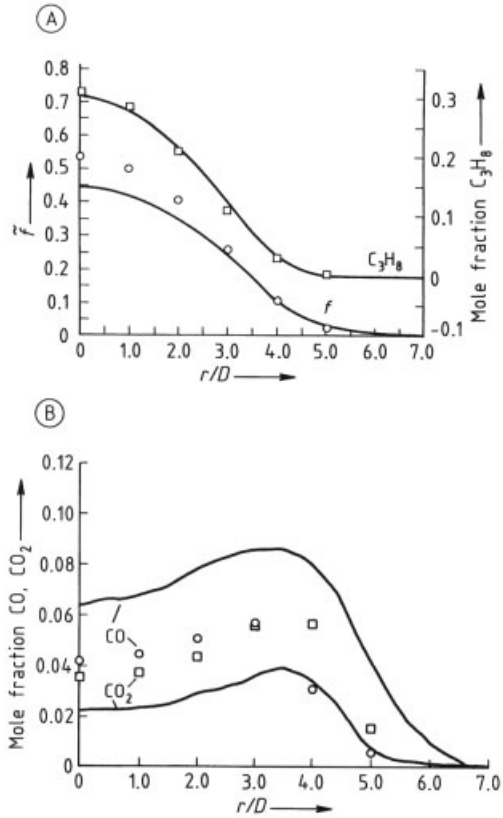

Figure 24. Simulation of a propane air diffusion flame 40 mm above the burner nozzle according to [70]
A) Mean values for the normalized elemental mass fraction (mixture fraction) and the mole fraction of propane; B) Mean values for the mole fraction of carbon monoxide and carbon dioxide
The open squares and circles denote measured values, the curves denote calculated values; r = radial distance; D = nozzle diameter.

5. Models Based on Physicochemical Principles (Transport Phenomena)

See also → Transport Phenomena and → Fluid Mechanics.

Models based on physicochemical principles apply the principles of the transport of momentum, mass, and energy to chemical engineering processes. The representation of processes by this type of model consists of the formulation of the conservation equations for the total mass, the mass of the individual chemical species, components of momentum, and the enthalpy for the system or subsystem under consideration. The

conservation equations can be given in the form shown in Section 4.2:

Accumulation + Change due to convective transport = Change due to molecular transport + Source

The required number, combination, and formulation of the conservation equations and any other physicochemical relationships necessary for closing the set of equations depend on the kind of process and the physical boundary conditions for the system under consideration.

In the following sections, applications of the principles of conservation of mass, momentum, and energy will be discussed. Wherever possible the classification according to Figure 4 will be used.

5.1. Application of the Principle of Conservation of Momentum

(The notation used in this chapter is the same as that employed in Chapter 4.)

The detailed notation for the equation for the conservation of momentum (Eq. 132) is

$$\varrho \frac{\partial u_i}{\partial t} + \varrho u_k \frac{\partial u_i}{\partial x_k} = \frac{\partial \tau_{ik}}{\partial x_k} - \frac{\partial p}{\partial x_i} + \varrho g_i \quad i = 1, 2, 3 \tag{204}$$

Physically this equation represents the change in momentum flux in a differential control volume caused by volume-specific forces: viscous stress (the first term on the right-hand side of Eq. 204), pressure (the second term on the right-hand side), and body forces, in this case the gravitational forces (the third term on the right-hand side). The assumptions made in the formulation of Equation (204) are discussed in [48], [50], [51].

For Newtonian fluids a gradient expression relates the viscous stresses to the flow velocities. In accordance with Newton's law of friction the shear stress tensor component τ_{ij} can be written as

$$\tau_{ij} = \mu \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} \mu \frac{\partial u_k}{\partial x_k} \delta_{ij} \tag{205}$$

where δ_{ij} is the Kronecker delta: $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. The assumptions used in the formulation of Equation (205) are discussed in [48], [50], [51]. Except for the pressure gradients, gravitational forces, density ϱ, and the dynamic viscosity of the material μ, Equation (204) is closed. In the case of gases, for example, μ can be determined from the kinetic theory of gases; other methods of calculating the dynamic viscosity are given in [48], [50]. For systems with constant density (e.g., incompressible fluids) and a known pressure field, Equation (204) is sufficient for calculating the velocity field. This will be demonstrated for the simple example of laminar flow in a tube.

5.1.1. Laminar Tube Flow

Laminar tube flow is illustrated in Figure 25. The flow is assumed to be fully developed and steady state, the accumulation term and the gradients in the direction of flow therefore disappear. There are no components of velocity transverse to the direction of flow. The problem is thus reduced to a one-dimensional problem with one dependent variable.

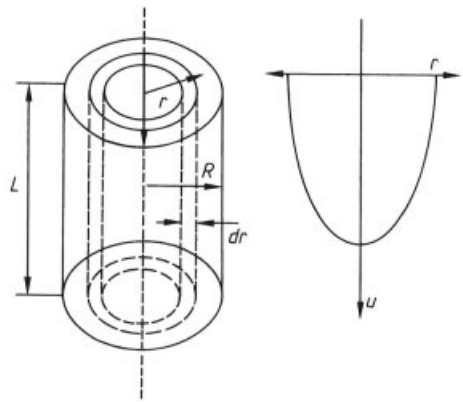

Figure 25. Laminar tube flow and parabolic velocity profile L = length; R = radius; r = radial distance; u = velocity in the axial direction

With the definitions specified in Figure 25 and the appropriate formulation in cylindrical coordinates, Equation (204) can easily be arranged as follows

$$0 = \frac{1}{r} \frac{\partial}{\partial r} (r \tau_{rz}) - \frac{\partial p}{\partial z} + \varrho g_z \tag{206}$$

If the pressure gradient in Equation (206) is replaced by the pressure difference over length L, which is taken as known, Equation (207) is obtained:

$$\frac{\partial}{\partial r}(r\tau_{rz}) = -\left(\frac{p_0 - p_L}{L} + \varrho g\right) r \tag{207}$$

Integration of Equation (207) gives

$$r\tau_{rz} = -\left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{r^2}{2} + C_1 \tag{208}$$

where C_1 is an integration constant. If the term from Newton's law of friction is substituted for the laminar viscous shear stress τ_{rz}, then Equation (208) becomes

$$\frac{\mathrm{d}u}{\mathrm{d}r} = -\left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{r}{2\mu} + \frac{C_1}{\mu r} \tag{209}$$

Integration of Equation (209) results in

$$u = -\left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{r^2}{4\mu} + \frac{C_1}{\mu}\ln r + C_2 \tag{210}$$

Since for $r = 0$ the velocity is finite, C_1 must be zero. The constant C_2 can be calculated from a further boundary condition, $u = 0$ for $r = R$:

$$C_2 = \left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{R^2}{4\mu}$$

Equation (210) thus takes the form

$$u = -\left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{R^2}{4\mu}\left[1 - \left(\frac{r}{R}\right)^2\right] \tag{211}$$

This is the well-known parabolic velocity profile for laminar tube flow (Hagen – Poiseuille flow). Further integration of Equation (211) with the relationship $\mathrm{d}q = u\,2\,\pi\,r\,\mathrm{d}r$ provides the Hagen – Poiseuille relationship for the volumetric flow rate q in laminar tube flow

$$q = \left(\frac{p_0 - p_L}{L} + \varrho g\right)\frac{\pi R^4}{8\mu} \tag{212}$$

Apart from the assumptions and simplifications associated with the momentum balance itself, the example of laminar flow does not contain any troublesome modeling hypotheses. This is not the case for turbulent flow.

5.1.2. Turbulent, Nonreactive Free Jets

Turbulent flows were described in Chapter 4 as flows with high kinetic energy so that the damping viscous forces can no longer hold the fluid particles on controlled paths due to their high inertial forces. This leads to the development of irregular statistical fluctuations. Turbulent flows are therefore basically unsteady and three-dimensional. This does not, however, represent the principal problem in mathematical modeling. As indicated in Section 2.2, compliance with physicochemical laws does not lose its validity even in statistical processes, such as turbulent flows. Direct numerical simulation of such flows in the nonsteady state with threedimensionally formulated conservation equations is possible in principle.

However, systematic difficulties arise which can only be solved for simple cases [49], [53], [54]. These problems include the small measurements of characteristic turbulent length scales (Kolmogorov length) and the extreme sensitivity of the nonsteady, three-dimensional conservation equations to changes in the boundary conditions. Other options for describing statistical processes are the use of probability density functions (Sections 4.2 and 4.3.2) or the moments of probability density functions. Although the latter do not provide complete statistical information, time- or ensemble-averaged values are often adequate for engineering purposes.

The most important features of models of statistical processes in the form of conservation equations for the moments of probability density functions will be outlined below for turbulent nonreactive flow. The aim is to derive transport equations for the mean values of the velocities starting from the conservation Equation (204) for the compounds of momentum which form a system of differential equations for the instantaneous values of the velocity components.

The example used is based on a jet stirred reactor, similar to that depicted in Figure 14.

The first step is easily carried out by using the decomposition Equation (112) and is demonstrated on a convective term in Equation (204). Equation (204) is first converted into

$$\frac{\partial}{\partial t}(\varrho U_i) + \frac{\partial}{\partial x_k}(\varrho U_k U_i) = \frac{\partial \tau_{ik}}{\partial x_k} - \frac{\partial p}{\partial x_i} + \varrho g_i \quad i = 1, 2, 3 \tag{213}$$

using the continuity Equation (131). From the convective term in one direction and Equation (112) the following is obtained:

$$\begin{aligned}\varrho U_i U_j &= \left(\langle\varrho\rangle + \varrho^{\mathrm{f}}\right)\left(\langle U_i\rangle + U_i^{\mathrm{f}}\right)\left(\langle U_j\rangle + U_j^{\mathrm{f}}\right)\\ &= \langle\varrho\rangle\langle U_i\rangle\langle U_j\rangle + \varrho\langle U_i\rangle U_j^{\mathrm{f}}\\ &\quad + \langle\varrho\rangle U_i^{\mathrm{f}}\langle U_j\rangle + \langle\varrho\rangle U_i^{\mathrm{f}} U_j^{\mathrm{f}}\\ &\quad + \varrho^{\mathrm{f}}\langle U_i\rangle\langle U_j\rangle + \varrho^{\mathrm{f}}\langle U_i\rangle U_j^{\mathrm{f}}\\ &\quad + \varrho^{\mathrm{f}} U_i^{\mathrm{f}}\langle U_j\rangle + \varrho^{\mathrm{f}} U_i^{\mathrm{f}} U_j^{\mathrm{f}}\end{aligned} \tag{214}$$

If Equation (214) is averaged, then by taking Equation (113 b) into account we get

$$\begin{aligned}\langle\varrho U_i U_j\rangle &= \langle\varrho\rangle\langle U_i\rangle\langle U_j\rangle + \langle\varrho\rangle\langle U_i^{\mathrm{f}} U_j^{\mathrm{f}}\rangle\\ &\quad + \langle U_i\rangle\langle\varrho^{\mathrm{f}} U_j^{\mathrm{f}}\rangle + \langle U_j\rangle\langle\varrho^{\mathrm{f}} U_i^{\mathrm{f}}\rangle\\ &\quad + \langle\varrho^{\mathrm{f}} U_i^{\mathrm{f}} U_j^{\mathrm{f}}\rangle\end{aligned} \tag{215}$$

This shows the basic problem in the derivation of conservation equations for the expected values: in addition to the expected values, the convective terms contain the covariances of the velocity components and density–velocity covariances. These terms must be modeled or replaced by assumptions. This is the major difference to dealing with the problem in terms of probability density functions, where convection is not modeled in the transport equations (cf. Section 4.2).

If Equation (215) is substituted into Equation (213), taking the average of the other terms and neglecting fluctuations in density, then

$$\begin{aligned}&\frac{\partial}{\partial t}\left(\langle\varrho\rangle\langle U_i\rangle\right) + \frac{\partial}{\partial x_k}\left(\langle\varrho\rangle\langle U_k\rangle\langle U_i\rangle\right)\\ &= \frac{\partial}{\partial x_k}\left(\langle\tau_{ik}\rangle - \langle\varrho\rangle\langle U_i^{\mathrm{f}} U_k^{\mathrm{f}}\rangle\right)\\ &\quad - \frac{\partial\langle p\rangle}{\partial x_i} + \langle\varrho\rangle\langle g_i\rangle, \quad i=1,2,3\end{aligned} \tag{216}$$

Neglection of the density fluctuations in Equation (216) represents an important modeling assumption. Further model formation concerns the terms $\langle\varrho\rangle\,\langle U_i^{\mathrm{f}} \mathrm{U}_j^{\mathrm{f}}\rangle$ which have the units of stresses and are thus called Reynolds stresses.

5.1.2.1. Models for Reynolds Stresses

The modeling of Reynolds stresses has already been referred to in Section 4.3.2.1. Discussion is continued here for the closure of the averaged transport equations for the momentum components. More detailed information is given in [49], [51], [52], [83–86].

The most commonly used model in engineering is based on the analogy between laminar and turbulent transfer of momentum (Chap. 4), established by Boussinesq [87] and expanded by Prandtl [88], [89]. Analogous to the gradient expression for laminar viscous shear stresses (Eq. 205) the apparent turbulent viscous stresses (Reynolds stresses) are coupled with the gradients of the mean velocity.

A tensor for the turbulent shear stresses is obtained:

$$\tau_{t_{ij}} = \mu_{\mathrm{t}}\left(\frac{\partial\langle U_i\rangle}{\partial x_j} + \frac{\partial\langle U_j\rangle}{\partial x_i}\right) - \frac{2}{3}\mu_{\mathrm{t}}\frac{\partial\langle U_k\rangle}{\partial x_k}\,\delta_{ij} \tag{217}$$

in which δ_{ij} is the Kronecker delta and μ_{t} is the apparent turbulent viscosity.

The problem is now to determine the apparent turbulent viscosity, which is not (as in the laminar case) a propensity of the material, but a characteristic generated by the flow. When determining this characteristic the analogy between molecular and turbulent transfer of momentum is retained.

In the laminar case the viscosity can be given as

$$\mu = \frac{1}{2}\varrho\lambda\bar{u} \tag{218}$$

from the kinetic gas theory, where $\bar{u}$ is the mean velocity and λ the mean free path of the molecules. For turbulent transfer of momentum the same argument leads to

$$\mu_{\mathrm{t}} \sim \bar{\varrho} L^* u^* \tag{219}$$

where $L*$ is the analogous term for the mean free path of the molecules and $u*$ the analogue for the mean velocity of the molecules. This corresponds with Prandtl's theory of mixing lengths [60]. In the turbulent case, not molecules but turbulent eddies are the elements which transfer momentum. Both of the quantities $L*$ and $u*$ representing the eddy viscosity hypothesis must therefore be determined by plausible arguments [84]. The characteristic velocity for the turbulent eddies is proportional to the fluctuations of the velocities $u* \sim (2/3\,\langle U_k^{\mathrm{f}} \mathrm{U}_k^{\mathrm{f}}\rangle)^{1/2}$ and the characteristic length is the distance which the turbulent eddies travel perpendicularly to the direction of flow, before they lose their identity through dissipation at a molecular level.

This can be expressed as $L* \sim u* \tau$, where τ is the turbulent time scale; using the approximation $\tau = \langle k \rangle / \langle \varepsilon \rangle$ (Section 4.3.2.3) gives $L* \sim \langle U_k^{\mathrm{f}} \mathrm{U}_k^{\mathrm{f}} \rangle^{3/2} / \langle \varepsilon \rangle$. The eddy viscosity hypothesis is thus

$$\mu_{\mathrm{t}} = C_\mu \langle \varrho \rangle \frac{\langle k \rangle^2}{\langle \varepsilon \rangle} \tag{220}$$

where C_μ is a modeling parameter. As a result of measurements on turbulent boundary layers its value is given as $C_\mu = 0.09$ [60].

Many of the problems involved in applying the eddy viscosity hypothesis (e.g., the physical requirements for gradient transport approaches) are discussed in the literature [84–86]. A mathematical model of this simplicity entails a series of other problems which are physically awkward to deal with. For the model with gradient transport approach and the eddy viscosity hypothesis ($k-\varepsilon$ turbulence model) these problems lie in the determination of the turbulence energy $\langle k \rangle$ and its dissipation rate $\langle \varepsilon \rangle$ and require the derivation of corresponding transport equations [84].

The description of turbulent flows in terms of the mean values of the velocities thus requires the calculation of second moments, in this case of the variances $\langle U_i^{\mathrm{f}} \mathrm{U}_j^{\mathrm{f}} \rangle$, from which the turbulence energy and its dissipation rate are derived. An equation for the turbulence energy may be derived by multiplying Equation (213) formulated for the i-direction by U_j^{f} and formulated for the j-direction by U_i^{f}. Addition of the two equations followed by use of Equation (112) for the components of the velocity, rearrangement of differential quotients, consideration of the continuity equation, and averaging finally give

$$\begin{aligned}
&\frac{\partial}{\partial t}\left(\langle \varrho \rangle \langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle\right) + \frac{\partial}{\partial x_k}\left(\langle \varrho \rangle \langle U_k \rangle \langle U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle\right) \\
&= -\frac{\partial}{\partial x_k}\left(\langle \varrho \rangle \langle U_k^{\mathrm{f}} U_i^{\mathrm{f}} U_j^{\mathrm{f}} \rangle\right) \\
&\quad - \langle \varrho \rangle \langle U_i^{\mathrm{f}} U_k^{\mathrm{f}} \rangle \frac{\partial \langle U_j \rangle}{\partial x_k} \\
&\quad - \langle \varrho \rangle \langle U_j^{\mathrm{f}} U_k^{\mathrm{f}} \rangle \frac{\partial \langle U_i \rangle}{\partial x_k} \\
&+ \langle U_i^{\mathrm{f}} \frac{\partial \tau_{jk}}{\partial x_j} \rangle + \langle U_j^{\mathrm{f}} \frac{\partial \tau_{ik}}{\partial x_i} \rangle \\
&\quad - \langle U_i^{\mathrm{f}} \frac{\partial p}{\partial x_j} \rangle - \langle U_j^{\mathrm{f}} \frac{\partial p}{\partial x_i} \rangle
\end{aligned} \tag{221}$$

gravitational forces being ignored. Summation over the diagonal elements from Equation (221) gives half the trace of the tensor for the Reynolds stresses $\langle U_l^{\mathrm{f}} \mathrm{U}_l^{\mathrm{f}} \rangle$, $\langle k \rangle = 1/2 \langle U_l^{\mathrm{f}} U_l^{\mathrm{f}} \rangle$, and hence the equation for turbulence energy

$$\begin{aligned}
&\frac{\partial}{\partial t}\left(\langle \varrho \rangle \langle k \rangle\right) + \frac{\partial}{\partial x_k}\left(\langle \varrho \rangle \langle U_k \rangle \langle k \rangle\right) \\
&= -\frac{\partial}{\partial x_k}\left(\langle \varrho \rangle \langle U_k^{\mathrm{f}} k \rangle\right) - \langle \varrho \rangle \langle U_l^{\mathrm{f}} U_k^{\mathrm{f}} \rangle \frac{\partial \langle U_l \rangle}{\partial x_k} \\
&+ \langle U_l^{\mathrm{f}} \frac{\partial \tau_{lk}}{\partial x_l} \rangle - \langle U_l^{\mathrm{f}} \frac{\partial p}{\partial x_l} \rangle
\end{aligned} \tag{222}$$

The description of statistical flows using mean values (first moments) additionally requires the calculation of second moments, in this case those of the velocity variances in the form of turbulence energy. Transport equations can be simply derived for this (see Eq. 222), but they contain terms with moments of the next higher order so that this procedure never leads to a closed equation system. The first expression on the right-hand side of Equation (222), for example, contains moments of the third order. To close the transport equations at a certain order, model assumptions have to be introduced for the higher order moments. These are often based on the same assumptions as the simple gradient transport assumptions and the eddy viscosity hypothesis. The physical basis and the larger number of model parameters arising in Reynolds stress models are the reasons that further development of these [83], [90], [91] is only gradually gaining acceptance in engineering.

The generally accepted form of the transport equation for turbulence energy (Eq. 223) is obtained from Equation (222) by using the definition for turbulent shear stresses, gradient transport approaches for third-order moments, the dissipation hypothesis [49], and other model assumptions [92]:

$$\begin{aligned}
&\frac{\partial}{\partial t}\left(\langle \varrho \rangle \langle k \rangle\right) + \frac{\partial}{\partial x_k}\left(\langle \varrho \rangle \langle U_k \rangle \langle k \rangle\right) \\
&= \frac{\partial}{\partial x_k}\left[\frac{\mu_{\mathrm{t}}}{\sigma_{k\mathrm{eff}}} \frac{\partial \langle k \rangle}{\partial x_k}\right] + G_k - \langle \varrho \rangle \langle \varepsilon \rangle
\end{aligned} \tag{223}$$

with the abbreviation

$$\begin{aligned}
G_k = \mu_{\mathrm{t}} &\left[\left(\frac{\partial \langle U_l \rangle}{\partial x_k} + \frac{\partial \langle U_k \rangle}{\partial x_l}\right)\right. \\
&\left. - \frac{2}{3}\left(\frac{\partial \langle U_k \rangle}{\partial x_k} + \langle \varrho \rangle \langle k \rangle\right) \delta_{ik}\right] \frac{\partial \langle U_k \rangle}{\partial x_k}
\end{aligned} \tag{224}$$

The individual steps in model formation for Equation (123) are described in [48], [51], [86], [92] and Section 4.3.2.1. For the simplifying homogeneous cases described in Section 4.3.2.1, Equation (223) becomes Equation (184) where $\langle\varepsilon\rangle = \langle k\rangle/\tau$.

For complete definition one more equation is needed for the dissipation rate $\langle\varepsilon\rangle = \mathrm{d}\langle k\rangle/\mathrm{d}t$ which occurs both in Equation (222) and also in the eddy viscosity hypothesis (Eq. 220). The derivation takes a similar path to that for the equation for turbulence energy and is described in [49], [92]. The accepted form is

$$\frac{\partial}{\partial t}\left(\langle\varrho\rangle\langle\varepsilon\rangle\right) + \frac{\partial}{\partial x_k}\left(\langle\varrho\rangle\langle U_k\rangle\langle\varepsilon\rangle\right) = \frac{\partial}{\partial x_k}\left[\frac{\mu_\mathrm{t}}{\sigma_{\varepsilon\mathrm{eff}}}\frac{\partial\langle\varepsilon\rangle}{\partial x_k}\right] + \left[C_{\varepsilon 1}G_k - 2C_{\varepsilon 2}\langle\varrho\rangle\langle\varepsilon\rangle\right]\frac{\langle\varepsilon\rangle}{\langle k\rangle} \tag{225}$$

In both Equations (223) and (225) $\sigma_{\varphi\ \mathrm{eff}}$ is the turbulent Schmidt number, which compares the mixing length for the velocity fluctuations to the mixing length for the Reynolds stresses or the dissipation rate respectively; $C_{\varepsilon 1}$ and $C_{\varepsilon 2}$ are further model parameters. From the solution of Equation (225), the turbulent time scale can be determined; this is a precondition for the closure of the models discussed in Sections 4.3.2.1 and 4.3.2.3.

Describing a nonreactive, isothermal turbulent flow with the model discussed above in terms of the expected velocity values leads to the averaged conservation equations of momentum (Eq. 216), as well as the equations for turbulence energy (Eq. 223) and the dissipation rate (Eq. 225). This system of simultaneous partial differential equations has to be solved by a suitable numerical procedure (see Section 5.1.2.2).

5.1.2.2. Solution Method of the Resulting System of Partial Differential Equations

The partial, simultaneous nonlinear differential Equations (216), (223), and (225) can be put into the form

$$\mathrm{div}\boldsymbol{J}_\varphi = \langle S_\varphi\rangle$$
$$\langle\Phi\rangle = \langle U\rangle, \langle V\rangle, \langle W\rangle, \langle k\rangle, \langle\varepsilon\rangle \tag{226}$$

for the steady-state problem being considered here. In Equation (226) $\boldsymbol{J}_\varphi$ is

$$\boldsymbol{J}_\varphi = \langle\varrho\rangle\langle\Phi\rangle\langle U_k\rangle - \frac{\mu_\mathrm{t}}{\sigma_{\varphi\mathrm{eff}}}\frac{\partial\langle\Phi\rangle}{\partial x_k} \tag{227}$$

and $\langle S_\varphi\rangle$ accomodates all other expressions not included by Equation (227). For the momentum equation $\langle S_\varphi\rangle$ also contains specific terms for the turbulent shear stress tensor from Equation (217) and the right-hand sides of the equation for $\langle k\rangle$ and $\langle\varepsilon\rangle$ can be seen from Equations (223) and (225).

The laminar transport of momentum is generally ignored compared with the turbulent shear stresses. To develop a numerical solution procedure, Equation (226) is integrated over a control volume (page 41). Using the Gauss–Ostrogradski integration principle this gives

$$\int_V \mathrm{div}\boldsymbol{J}_\varphi = \oint_F J_{\varphi k} = \int_V \langle S_\varphi\rangle \tag{228}$$

The integration has to be carried out over the surfaces of the control volume which are perpendicular to the components $J_{\varphi i}$, so that for a cartesian system of coordinates the relationship

$$\Delta J_{\varphi k}\mathrm{d}x_i\mathrm{d}x_{j,i,j\neq k} = \int_V \langle S_\varphi\rangle$$
$$\langle\Phi\rangle = \langle U\rangle, \langle V\rangle, \langle W\rangle, \langle k\rangle, \langle\varepsilon\rangle \tag{229}$$

results, in which the operator Δ denotes the difference at two opposite surfaces of the control volume. Equation (229) is the basis for setting up difference equations which are substituted for the continuous differential operators in Equations (216), (223), and (225) to allow numerical integration.

The conservation equations for the momentum components contain the simple pressure gradients as sources; a reduction in pressure in the direction of the ith coordinate thus leads to an increase of momentum in this direction. Equations (216), (223), and (225) do not, however, contain a pressure term so that the pressure $\langle p\rangle$ or the pressure gradients $\partial\langle p\rangle/\partial x_i$ cannot be directly calculated from them. The same problem occurs in closing the transport equation for the probability density functions (Section 4.3.2.1). Solution of these equations assumes a known pressure field, which is implicitly given through the conservation equations for the momentum

components and the continuity equation. If an estimate for the pressure field is assumed for a numerical solution, then the continuity equation (Eq. 131) is usually not complied with. Consequently the numerical solution in an iterative procedure may get further and further away from physically meaningful results because the equation system does not contain correction factors to ensure compliance with the continuity equation.

In the numerical simulation of turbulent steady flow or when using the Navier–Stokes equations, this problem is frequently solved by replacing the primitive variables $\langle \boldsymbol{U} \rangle$ and $\langle p \rangle$ with stream functions and vorticity [94], [95] through the von Mises transformation [93]. The stream function is defined by the von Mises transformation such that the continuity equation is complied with (see also Section 5.2.2.1). In the procedure outlined below another possibility is suggested, in which the pressure correction is obtained from the continuity equation and a simplified form of the momentum balance (the SIMPLE algorithm) [96], [97].

For the example considered here of a turbulent free jet the formulation in cylinder coordinates is used as in Section 5.1.1. Due to the rotational symmetry Equation (226) can be expressed as

$$L(\Phi) = \langle S_\varphi \rangle, \quad \Phi = \langle U \rangle, \langle V \rangle, \langle k \rangle, \langle \varepsilon \rangle \tag{230}$$

The turbulent transport operator $L(\Phi)$ is given by

$$L(\Phi) \equiv \frac{\partial\left(\langle\varrho\rangle\langle U\rangle\langle\Phi\rangle\right)}{\partial x} + \frac{1}{r}\frac{\partial\left(r\langle\varrho\rangle\langle V\rangle\langle\Phi\rangle\right)}{\partial r} - \frac{\partial}{\partial x}\left(\frac{\mu_t}{\sigma_{\varphi\mathrm{eff}}}\frac{\partial\langle\Phi\rangle}{\partial x}\right) - \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\mu_t}{\sigma_{\varphi\mathrm{eff}}}\frac{\partial\langle\Phi\rangle}{\partial r}\right)$$

$$\langle\Phi\rangle = \langle U\rangle, \langle V\rangle, \langle k\rangle, \langle\varepsilon\rangle \tag{231}$$

Since an axially symmetrical system is invariant with respect to rotation around the axis of symmetry, then a two-dimensional grid with L nodal points in the x-direction and M nodal points in the r-direction is used for replacing the differential operators in Equation (231) by differences (Fig. 26). The spatial structure of the grid arises from rotation about the axis of symmetry, and the control volumes are defined by revolving the grid through the unit angle. The control volume is indicated in Figure 26 by a shaded area, the relevant flows, for example $J_{\varphi 1_w}$, can be expressed as

$$J_{\varphi 1_w} = -\Big(\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle U\rangle\right)_w, 0\right\}\langle\Phi\rangle_{l-1} - \mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle U\rangle\right)_w, 0\right\}\langle\Phi\rangle_l - \frac{\mu_t}{\sigma_{\varphi\mathrm{eff}}}\frac{\langle\Phi\rangle_l - \langle\Phi\rangle_{l-1}}{x_l - x_{l-1}}\Big) F_w \tag{232}$$

The formulation given in Equation (232) contains an "up-wind" difference scheme [Max $\{+(\langle\varrho\rangle\langle U\rangle)_w, 0\}$ $\langle\Phi\rangle_{l-1}$ − Max $\{-(\langle\varrho\rangle \cdot \langle U\rangle)_w, 0\}\langle\Phi\rangle_l$] for the convective terms (analogous to the cascade of ideally mixed stirred reactors discussed in Section 4.3.1.1) and a central difference scheme for the diffusive terms. As with the model for ideally mixed reactors, the up-wind difference scheme is based on the assumption that the upstream value of a physical quantity is transported over the control surfaces.

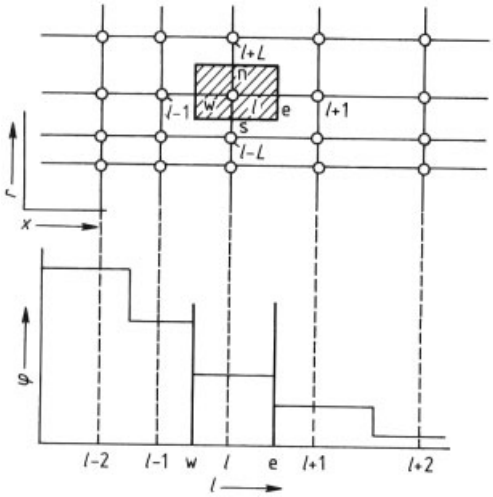

Figure 26. Diagram illustrating the "up-wind" difference scheme for elliptical differential equations
l = number of grid point; L = number of grid points in n direction; n = north; e = east; s = south; w = west; φ = any variable

The profiles of the variables brought into discrete form in the x-direction in accordance with the "up-wind" difference scheme are also shown in Figure 26. Problems associated with this difference scheme (e.g., "numerical diffusion") and measures to limit them are described in [97–101].

The grid with L nodal points in the x-direction and M nodal points in the r-direction and the

difference scheme can be used to develop the difference equations from Equation (229); the fluxes through the remaining surfaces of the control volume are defined analogously to Equation (232). Summation of the fluxes gives

$$
\begin{aligned}
&-\Big(\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{w}},0\right\}\langle\Phi\rangle_{l-1} \\
&-\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{w}},0\right\}\langle\Phi\rangle_{l} \\
&-\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{\langle\Phi\rangle_{l}-\langle\Phi\rangle_{l-1}}{x_l-x_{l-1}}\Big)F_{\mathrm{w}} \\
&+\Big(\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{e}},0\right\}\langle\Phi\rangle_{l} \\
&-\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{e}},0\right\}\langle\Phi\rangle_{l+1} \\
&-\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{\langle\Phi\rangle_{l+1}-\langle\Phi\rangle_{l}}{x_{l+1}-x_l}\Big)F_{\mathrm{e}} \\
&-\Big(\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{s}},0\right\}\langle\Phi\rangle_{l-L} \\
&-\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{s}},0\right\}\langle\Phi\rangle_{l} \\
&-\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{\langle\Phi\rangle_{l}-\langle\Phi\rangle_{l-L}}{x_l-x_{l-L}}\Big)F_{\mathrm{s}} \\
&+\Big(\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{n}},0\right\}\langle\Phi\rangle_{l} \\
&-\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{n}},0\right\}\langle\Phi\rangle_{l+L} \\
&-\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{\langle\Phi\rangle_{l+L}-\langle\Phi\rangle_{l}}{x_{l+L}-x_l}\Big)F_{\mathrm{n}}=\langle S\rangle_{\varphi l}
\end{aligned}
\tag{233}
$$

In Equation (233) $\langle S\rangle_{\varphi l}$ represents the source terms from Equation (230) integrated over the control volume. If the continuity equation in the corresponding discrete form is taken into account

$$
\begin{aligned}
&\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{w}}F_{\mathrm{w}}-\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{e}}F_{\mathrm{e}}+\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{s}}F_{\mathrm{s}} \\
&\quad-\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{n}}F_{\mathrm{n}}=0
\end{aligned}
\tag{234}
$$

then Equation (233), after addition of Equation (234) multiplied by $\langle\Phi\rangle_l$, can be transformed into

$$
\begin{aligned}
&-a_{l-L}\langle\Phi\rangle_{l-L}-a_{l-1}\langle\Phi\rangle_{l-1}+b_l\langle\Phi\rangle_l \\
&\quad-a_{l+1}\langle\Phi\rangle_{l+1}-a_{l+L}\langle\Phi\rangle_{l+L}-\langle S\rangle_{\varphi l}=0
\end{aligned}
\tag{235}
$$

In accordance with Equation (233) the coefficients for Equation (235) are

$$
a_{l-L}=\Big[\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{s}},0\right\}+\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{1}{x_l-x_{l-L}}\Big]F_{\mathrm{s}} \tag{236a}
$$

$$
a_{l-1}=\Big[\mathrm{Max}\left\{+\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{w}},0\right\}+\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{1}{x_l-x_{l-1}}\Big]F_{\mathrm{w}} \tag{236b}
$$

$$
a_{l+1}=\Big[\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle U\rangle\right)_{\mathrm{e}},0\right\}+\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{1}{x_{l+1}-x_l}\Big]F_{\mathrm{e}} \tag{236c}
$$

$$
a_{l+L}=\Big[\mathrm{Max}\left\{-\left(\langle\varrho\rangle\langle V\rangle\right)_{\mathrm{n}},0\right\}+\frac{\mu_{\mathrm{t}}}{\sigma_{\varphi\mathrm{eff}}}\frac{1}{x_{l+L}-x_l}\Big]F_{\mathrm{n}} \tag{236d}
$$

$$
b_l=a_{l-L}+a_{l-1}+a_{l+1}+a_{l+L} \tag{236e}
$$

Difference equations corresponding to Equation (235) can be formulated for the $(L-2)$, $(M-2)$ inner grid points of the calculation domain for all variables $\langle\Phi\rangle=\langle U\rangle$, $\langle V\rangle$, $\langle k\rangle$, $\langle\varepsilon\rangle$. Similar relationships are derived for the grid points lying on the boundaries of the calculation domain. The resulting system of equations is nonlinear and coupled via $\langle S\rangle_{\varphi l}$ and the coefficients a and b, which are not constant but depend on the components of the velocity and turbulent viscosity. Because the numerical procedure for solving Equation (235) is an iterative one these dependencies are not resolved.

An analogous equation for pressure corrections can be obtained with the procedure suggested in [96], [97]:

$$
\begin{aligned}
&-a_{pl-L}\Delta\langle p\rangle_{l-L}-a_{pl-1}\Delta\langle p\rangle_{l-1}+b_{pl}\Delta\langle p\rangle_l \\
&-a_{pl+1}\Delta\langle p\rangle_{l+1}-a_{pl+L}\Delta\langle p\rangle_{l+L}-Rs_l=0
\end{aligned}
\tag{237}
$$

where $\Delta\langle p\rangle$ represents the corrections for the given pressure field, and Rs_l the "continuity error", which is obtained from the residuals of the continuity equation with the calculated velocity field. The coefficients of Equation (237) are obtained from a simplified form of the momentum balance [96], [97]. To avoid numerical problems

which originate from the structure of the averaged momentum equations, Equations (235) and (237) are best solved with a "staggered grid" arrangement for the velocities [98].

Equations (235) and (237) constitute a system of nonlinear equations with 5 $(L-2)$ $(M-2)$ unknowns for the example considered. The values of the variables at the boundaries of the calculation domain are given by plausible boundary conditions. The system matrix for this system of nonlinear equations has a block pentadiagonal structure due to the elliptical formulation of the problem (Eq. 231), and the difference scheme used. The system of equations can now be solved, for example with the Newton – Raphson procedure outlined in Section 3.2.1.3 or other methods.

The model for the Reynolds stresses discussed in Section 5.1.2.1 as well as the numerical procedure given in this section are examples from a rapidly developing field of engineering. Other models for the Reynolds stresses represent the phenomena observed in turbulent flows better than the $k-\varepsilon$ turbulence model [52], [84–86], [90–92]. Difference schemes other than the combination of "up-wind" and central differences are also used for partial elliptical differential equations and other solution procedures [101–104] are suggested.

5.1.2.3. Example: Turbulent Flow of Nitrogen in Air

In the example discussed here nitrogen is discharged at room temperature from a tube of 10 mm diameter with an exit velocity of 85 m/s into air which is almost at rest (0.3 m/s) (Fig. 14). The system of nonlinear Equations (235) and (237) is solved for a tensor product grid by an alternating direction implicit (ADI) procedure [101]. In this procedure the block structure is first resolved by solving the system of equations one after the other for the individual variables. The estimated values from the previous iteration are used for the variables contained in the couplings. The system is converted into a quasi-one-dimensional one by including the elements of the two outer diagonals in the right-hand sides.

The system of equations is solved alternately along a line of the grid in Figure 26 in the x- or r-direction. The coefficients situated on the neighboring lines are calculated with the estimated values from the previous iteration.

To increase the velocity of convergence, linear relaxation is used (Section 3.2.1.3) and the source terms are linearized in the form $\langle S_\varphi \rangle = \langle S_{\mathrm{lin}\varphi} \rangle + \langle S_{\mathrm{prop}\varphi} \rangle \ \langle \Phi \rangle$. The latter measure is particularly useful for stiff differential equation systems, for which the source terms are highly nonlinear and coupled. The modeling parameters [91] used for calculation are $C_\mu = 0.09$, $C_{\varepsilon 1} = 1.4$, $C_{\varepsilon 2} = 0.925$, $\sigma_{k\ \mathrm{eff}} = 1.0$, and $\sigma_{\varepsilon\ \mathrm{eff}} = 1.3$. Figure 27 shows the predicted values for the mean velocity in the x-direction for the turbulent flow of nitrogen at two distances from the nozzle outlet in comparison with the similarity solutions from [105]. The latter give the flow velocity as

$$\langle U \rangle_m = \langle U \rangle_0 / \left(0.16 \frac{x}{D_{\mathrm{R}}} - 1.5 \right) \tag{238}$$

and

$$\langle U \rangle = \langle U \rangle_m \exp \left[-C_u \left(\frac{r}{x} \right)^2 \right] \tag{239}$$

where $C_u = 90$.

5.2. Applications of the Principle of Conservation of Enthalpy

The conservation equation of enthalpy (Eq. 134) is written in detail as

$$\varrho \frac{\partial h}{\partial t} + \varrho u_k \frac{\partial h}{\partial x_k} = -\frac{\partial J_k^h}{\partial x_k} + \varrho S_h \tag{240a}$$

In accordance with Equation (240 a) enthalpy transport occurs due to convection and molecular transport. The term S_h contains all the mechanisms for conversion of enthalpy (e.g., volumetric work due to a change in pressure, work against physical forces or radiation).

Fourier's simple gradient transport approach can be used to represent molecular transport:

$$J_i^h = -\lambda \frac{\partial T}{\partial x_i} \tag{241}$$

where λ is the thermal conductivity. If a Lewis number, $Le = Pr/Sc$, of one is assumed and other transport effects are neglected, the simple form of the conservation of energy equation is obtained

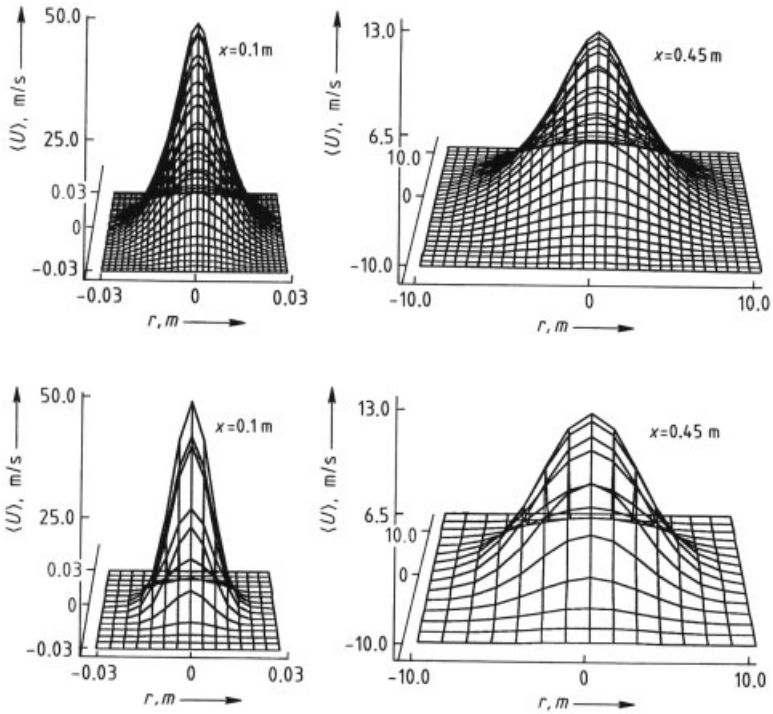

Figure 27. Profiles of the expected values for velocities in the x-direction of a turbulent inert flow from a free turbulent jet The upper diagrams show numerical simulations, the lower diagrams show similarity solutions. The radial profiles at two distances from the nozzle are given.

$$\varrho\frac{\partial h}{\partial t} + \varrho u_k \frac{\partial h}{\partial x_k} = \frac{\partial}{\partial x_k}\lambda\frac{\partial T}{\partial x_k} + \varrho S_h \tag{240b}$$

Detailed discussion of the assumptions and simplifications for this formulation are found in [48], [50], [51].

The conservation equation of enthalpy also contains the density and the components of the velocity. The enthalpy and momentum equations are thus coupled and have to be solved simultaneously. To avoid this a simple heat conduction problem is discussed first.

5.2.1. Heat Conduction

A simple heat conduction problem is shown in Figure 28. A pipe consisting of two different materials (pipe and insulation) has an internal temperature T_{Wi}, the external surface has a temperature of T_{Wo}. If only molecular conduction is considered, Equation (240 a) is reduced to

$$-\frac{1}{r}\frac{\mathrm{d}\left(rJ_r^h\right)}{\mathrm{d}r}=0 \tag{242}$$

Integration gives

$$-rJ_r^h=C_1 \tag{243a}$$

Applying Equation (241) gives

$$r\lambda\frac{\mathrm{d}T}{\mathrm{d}r} = C_1 \tag{243b}$$

which after integration results in the relationship

$$T = \frac{C_1}{\lambda}\ln r + C_2 \tag{244a}$$

for the temperature profile. For the inner pipe surface, the boundary condition $T = T_{\mathrm{Wi}}$ applies for $r = r_{\mathrm{Wi}}$. Thus $C_2 = T_{\mathrm{Wi}} - (C_1/\lambda_{\mathrm{i}})\ln r_{\mathrm{Wi}}$ and Equation (244 a) becomes

$$T = T_{\mathrm{Wi}} - \frac{C_1}{\lambda_i} - \ln\frac{r}{r_{\mathrm{Wi}}} \tag{244b}$$

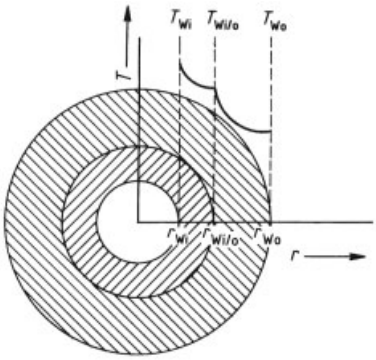

Figure 28. Arrangement and temperature profile for a simple heat conduction problem
The subscripts Wi and Wo denote inner wall and outer wall, respectively.

Correspondingly for the outer pipe surface the boundary condition is

$$T_{\mathrm{Wi/o}} = T_{\mathrm{Wi}} - \frac{C_1}{\lambda_i} \ln \frac{r_{\mathrm{Wi/o}}}{r_{\mathrm{Wi}}}$$

and

$$T = T_{\mathrm{Wi/o}} - \frac{C_1}{\lambda_{\mathrm{o}}} \ln \frac{r}{r_{\mathrm{Wi/o}}} \tag{244c}$$

In Equations (244 a – c), C_1 is given by the total heat flux (Eq. 243 a).

5.2.2. Heat Transfer

5.2.2.1. Exact Solution for a Boundary Layer Problem

An exact solution for the heat transfer from a fluid to a solid wall is possible for laminar flow over a flat plate (Fig. 29). A fluid flows over a flat plate of infinitely small thickness and of length L at a velocity u_∞. Due to the viscous stresses at the wall the fluid particles slow down and a boundary layer of increasing thickness forms along the plate in which the velocity increases from zero to u_∞. This gives the boundary conditions $u = 0$ for $y = 0$ and $u_{\delta_u}(x) = u_\infty$. For the sake of simplicity, constant pressure is assumed, $p_{\delta_u}(x) = p_\infty =$ const. The flow is steady-state and the plate extends infinitely in one direction so that the problem can be formulated twodimensionally. The boundary conditions for temperature are $T = T_{\mathrm{W}}$ for $y = 0$ and $T_{\delta_T}(x) = T_\infty$. An equation for the temperature can be obtained from Equation (240 b) by neglecting sources such as the conversion of mechanical energy or viscous dissipation and by using the equation for the definition of enthalpy $\mathrm{d}h = c_p \mathrm{d}T$ and the usual boundary layer approximations [59], [60]:

$$c_p \varrho u \frac{\partial T}{\partial x} + c_p \varrho v \frac{\partial T}{\partial y} = \lambda \frac{\partial^2 T}{\partial y^2} \tag{245a}$$

At constant density and specific heat

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = a \frac{\partial^2 T}{\partial y^2} \tag{245b}$$

where $a = \lambda / \varrho\, c_p$.

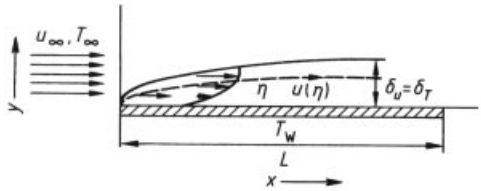

Figure 29. Development of a velocity (δ_u) and temperature (δ_T) boundary layer for laminar incompressible flow over a flat plate
$Pr = 1$; u_∞ = velocity in the x-direction outside the boundary layer; T_∞ = temperature outside the boundary layer; η = similarity coordinate; T_{W} = wall temperature; L = plate length

The differential equation for the temperature still contains the unknown velocities u and v. Equations for these variables are obtained from the continuity equation (Eq. 131) and the momentum equation (Eq. 204) in the x-direction

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{246}$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \nu \frac{\partial^2 u}{\partial y^2} \tag{247}$$

where $\nu = \mu / \varrho$ is the kinematic viscosity. For $Pr = \nu / a = 1$ and hence for $\nu = a$, the enthalpy equation is identical with the momentum equation, so that the solution of the latter produces a solution to the enthalpy equation. Furthermore, the coupling of the momentum and enthalpy equations are removed by these assumptions.

The general expression for temperature is

$$\frac{T}{T_\infty} = A \frac{u}{u_\infty} + B$$

$$\text{where} \quad A = \frac{T_\infty - T_{\mathrm{W}}}{T_\infty} \quad \text{and} \quad B = \frac{T_{\mathrm{W}}}{T_\infty} \tag{248}$$

The heat flux to the wall is then

$$J_{\mathrm{W}}^{h} = -\lambda \left(\frac{\partial T}{\partial y}\right)_{\mathrm{W}} = -\lambda \frac{T_\infty - T_{\mathrm{W}}}{u_\infty} \left(\frac{\partial u}{\partial y}\right)_{\mathrm{W}} \quad (249)$$

Determination of the heat flux to the wall thus depends on determination of the velocity gradient at the wall, which is a function of x and can be obtained from the solution of Equations (246) and (247).

For the solution, the stream function Ψ is used to eliminate the velocities u and v [59], [60]:

$$u = \frac{\partial \Psi}{\partial y} \quad \text{and} \quad v = -\frac{\partial \Psi}{\partial x} \quad (250\text{a,b})$$

The stream function invariably satisfies the continuity equation, as can be seen by suitable differentiation. With this transformation, Equation (247) becomes a partial differental equation of the third order

$$\Psi_y \Psi_{yx} - \Psi_x \Psi_{yy} = \nu \Psi_{yyy} \quad (251)$$

in which Ψ_x and the appropriate symbols for each of the differential operators are used.

In the boundary layer formed in laminar flow locations of constant velocity u lie on a parabola $\eta \sim y/\sqrt{x}$ (Fig. 29). If the equation system is therefore transformed to this similarity coordinate η, then Equations (247) and (251) can be reduced to ordinary differential equations. The exact transformation is

$$\eta\,(x, y) = y\sqrt{\frac{u_\infty}{\nu x}} \quad (252)$$

in which u_∞/ν represents a scale factor that makes η dimensionless. As u is only a function of η, the general expression $u/u_\infty = g\,(\eta)$ can be used for making the stream function dimensionless:

$$\Psi\,(x, y) = \int u \mathrm{d}y = u_\infty \int g\,(\eta)\,\mathrm{d}\eta$$
$$= u_\infty \sqrt{\frac{u_\infty}{\nu}} \int g\,(\eta)\,\mathrm{d}\eta \quad (253)$$

where $\int g\,(\eta)\,\mathrm{d}\eta \equiv f\,(\eta)$ is called the dimensionless stream function. Equations (250 a, b), (252), and (253) provide all the transformations and conversions that are needed to convert the partial differential Equation (251) into an ordinary differential equation of dimensionless quantities.

$$\Psi_y = u_\infty f_\eta, \quad \Psi_x = \frac{1}{2}\sqrt{\frac{u_\infty}{\nu x}}\,[\eta f_\eta - f\,(\eta)]\,,$$
$$\Psi_{yx} = -\frac{1}{2}\frac{u_\infty}{x}\eta f_{\eta\eta},$$
$$\Psi_{yy} = u_\infty \sqrt{\frac{u_\infty}{\nu x}} f_{\eta\eta},$$
$$\Psi_{yyy} = u_\infty \frac{u_\infty}{\nu x} f_{\eta\eta\eta} \quad (254\text{a,b,c,d,e})$$

In Equations (254) f_η and the corresponding symbols denote the derivatives of f with respect to η. The ordinary differential equation

$$f_{\eta\eta\eta} + \frac{1}{2} f \cdot f_{\eta\eta} = 0 \quad (255)$$

is thus obtained from Equation (251) which has to be solved for the boundary conditions $f\,(\eta) = 0$, $f_\eta = 0$ for $y = 0$ or $\eta = 0$ and also for $f_\eta = 1$ for $y \to \infty$ or $\eta \to \infty$.

Equation (255) can be solved numerically but not analytically. The essential quantities for the boundary layer problem, $f\,(\eta) = \int (u/u_\infty)\,\mathrm{d}\eta$, $f_\eta = u/u_\infty$, and $f_{\eta\eta} = u^{-1}{}_\infty\,(\partial u/\partial \eta)$, can be found in the literature, tabulated as a function of η [60], [106]. According to [106], $u/u_\infty = 0.99$ for $\eta = 5.0$ and $f_{\eta\eta} = 0.332$ for $\eta = 0$.

The boundary layer thickness is calculated from the first of these two values and is defined here as the distance δ_u at which the velocity reaches 99 % of the flow velocity outside the boundary layer. It then follows from Equation (252) that

$$\delta_u = 5\sqrt{\frac{\nu x}{u_\infty}} \quad \text{or} \quad \frac{\delta_u}{x} = \frac{5}{\sqrt{Re_x}} \quad (256\text{a,b})$$

The essential information provided by Equation (256) is that the development of the boundary layer thickness for laminar flow over a flat plate is proportional to $\sqrt{x}$ and $1/\sqrt{u_\infty}$, and inversely proportional to the square root of the Reynolds number $Re_x = u_\infty x/\nu$. The solutions of the momentum and enthalpy equations are similar because $Pr = 1$; as a result $\delta_u = \delta_T$. A detailed discussion of further implications of the solution of the boundary layer equations is given in [59], [60].

Using the value

$$0.332 = f_{\eta\eta} = \frac{1}{u_\infty}\frac{\partial u}{\partial \eta} = \frac{1}{u_\infty}\frac{\partial u}{\partial y}\frac{\partial y}{\partial \eta}$$
$$= \frac{1}{u_\infty}\frac{1}{\sqrt{u_\infty/\nu x}}\frac{\partial u}{\partial y}$$

the following relation is obtained:

$$\left(\frac{\partial u}{\partial y}\right)_{\mathrm{W}} = 0.332 u_\infty \sqrt{\frac{u_\infty}{\nu x}} \tag{257}$$

Substitution of this term in Equation (249) gives

$$J_{\mathrm{W}}^h = -0.332\lambda\sqrt{\frac{u_\infty}{\nu x}}\,(T_\infty - T_{\mathrm{W}}) \tag{258}$$

If the definition for the heat-transfer coefficient $J_{\mathrm{W}}^h \equiv \alpha\,\Delta T$ where $\Delta T = (T_{\mathrm{W}} - T_\infty)$ is now substituted in Equation (258), then

$$\alpha = 0.332\lambda\sqrt{\frac{u_\infty}{\nu x}} = 0.332\lambda\frac{1}{x}\sqrt{Re_x} \tag{259}$$

Rearrangement of this equation in dimensionless groups using $Nu_x = \alpha\, x/\lambda$ yields the solution for the heat-transfer problem for laminar flow over a flat plate:

$$\frac{Nu_x}{\sqrt{Re_x}} = 0.332 \tag{260}$$

The physical background of Equations (259) and (260) is that the local heat-transfer coefficient, like the boundary layer thickness, is inversely proportional to the distance in the x-direction and the velocity of flow. This is easy to understand because the temperature and velocity gradients level out at the wall with the formation of the boundary layer.

In technical problems the local heat-transfer coefficient is often not as important as its mean value α_m from the beginning of the plate to a distance x or the total length L. Equation (259) can be written as $\alpha = \text{const.}\ x^{-1/2}$:

$$\alpha_m = \frac{1}{L}\int_0^L \alpha\,(x)\,\mathrm{d}x = \frac{\text{const.}}{L}\int_0^L \frac{\mathrm{d}x}{\sqrt{x}}$$

$$= 2\frac{\text{const.}}{L}\sqrt{x}\Big|_0^L = 2\alpha_L \tag{261a}$$

The mean heat-transfer coefficient for a plate with length L is thus double the local heat-transfer coefficient at the distance L. This can also be formulated with Equation (260):

$$Nu_m = 0.664\sqrt{Re_L} = 2Nu_L \tag{261b}$$

The analysis described here only applies under the highly simplified assumptions of incompressible flow and constant density and also $Pr = \nu/a = 1$. Nevertheless this allows the general principles of modeling for heat transfer to be described.

5.2.2.2. General Principles of Modeling Heat Transfer

Under industrial conditions the simplifications introduced in the previous section become void. The consequence of $Pr \neq 1$ and variable density is that Equations (245), (213), and (131) are coupled through the density. Moreover, the solutions for the momentum balances and the enthalpy balances are no longer identical. The dimensions of the temperature boundary layer δ_T and the hydrodynamic boundary layer δ_u differ. In addition, flow is mostly turbulent sometimes with very complicated geometries.

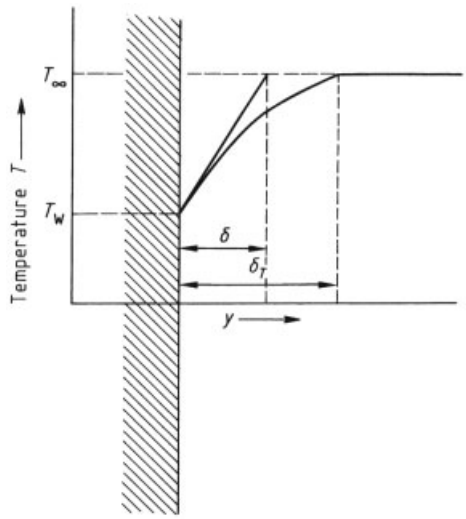

Figure 30. Representation of a general model for heat transfer
δ = film thickness; δ_T = thickness of temperature boundary layer; T_{W} = wall temperature; T_∞ = temperature outside the boundary layer.

As the heat transfer under such conditions can still be defined as

$$J_{\mathrm{W}}^h = -\lambda\left(\frac{\partial T}{\partial y}\right)_{\mathrm{W}} \equiv \alpha\Delta T \tag{262}$$

or from the appropriate mean value over a surface, the problem is thus to determine the temperature gradients at the wall. Heat transfer at the wall is primarily a heat conduction problem. The temperature profile in the vicinity of the wall

is however determined by simultaneous transfer of momentum, heat, and mass. Analytical or numerical solutions for the system of Equations (131), (213), and (245) will only succeed in a few cases with a justifiable amount of effort [25], [59], [107], [108], see also the solution of the momentum equations for the turbulent inert jet of gas in Section 5.1.2.2.

In most engineering applications semi-empirical methods are therefore used which assume a linearized temperature profile at the wall. Figure 30 shows the temperature profile for laminar flow over a plate.

If the temperature is developed at the wall, $y = 0$ in a Taylor series that is truncated after the first member, then

$$T = T_{\mathrm{W}} + \left(\frac{\partial T}{\partial y}\right)_{\mathrm{W}} \mathrm{d}y \text{ or } T_\infty = T_{\mathrm{W}} + \left(\frac{\partial T}{\partial y}\right)_{\mathrm{W}} \delta \tag{263}$$

If this is substituted into Equation (262) then

$$\lambda \frac{T_\infty - T_{\mathrm{W}}}{\delta} = \alpha\,(T_\infty - T_{\mathrm{W}}) \quad \text{or} \quad \alpha = \frac{\lambda}{\delta} \tag{264a,b}$$

Here δ is the film thickness produced by linearization of the temperature profile as shown in Figure 30. This is not identical with the boundary layer thickness δ_u or δ_T. As can be seen from the solutions for the laminar boundary layer

$$\left(\frac{\partial T}{\partial y}\right)_{\mathrm{W}} = 0.332\,(T_\infty - T_{\mathrm{W}})\,\sqrt{Re_x}/x$$

whereas

$$\left(\frac{\partial T}{\partial y}\right)^*_{\mathrm{W}} = \frac{1}{5}\,(T_\infty - T_{\mathrm{W}})\,\sqrt{Re_x}/x$$

when the thickness δ_T of the temperature boundary layer is used for the difference quotients. This latter quantity is thus only ca. 60 % of the true temperature gradient. The determination of heat transfer thus depends on determination of the film thickness as in Equation (264). This problem has been solved experimentally in most technical applications, where the mean heat-transfer number depending on the various variables is represented in the form $Nu = f\,(Re, Pr, L/D\ldots)$, see also Equations (260) and (261).

These relationships for industrial heat exchangers are not treated here. They are described in [25], see also → Heat Exchange. An example of such a relationship for the transfer of heat for turbulent flow was discussed in Sections 3.1.1.4 and 3.2.3.

With the establishment of the mean heat-transfer coefficients α and the relevant temperature difference, heat transfer (Eq. 262) is clearly defined. Arrangements such as that shown in Figure 28 can be calculated completely with regard to temperature profiles or heat losses. Summation of the temperature differences for the individual "phases" of the arrangement allows heat transfer to be related to an "overall resistance"

$$\frac{1}{k} = \frac{1}{\alpha_{\mathrm{Wi}}} + \sum \frac{1}{\lambda_{\mathrm{i}}/\,(r_{\mathrm{Wo}} - r_{\mathrm{Wi}})} + \frac{1}{\alpha_{\mathrm{Wo}}}$$

and to the total temperature difference. The equation for heat transfer analogous to Equation (262) will then be used depending on the statement of the question. The calculation procedure will be shown in an analogous problem of mass transfer in Section 5.3.1.3.

5.3. Applications of the Law of Conservation of Mass

The law of conservation of mass is formulated through the continuity equation (Eq. 131) and the conservation equations for the mass of the individual chemical species, m_i

$$\varrho \frac{\partial m_i}{\partial t} + \varrho u_k \frac{\partial m_i}{\partial x_k} = -\frac{\partial J_k^{m_i}}{\partial x_k} + \varrho S_i$$
$$i = 1, \ldots, N-1 \tag{265a,b}$$

A further equation results from the condition $\sum_1^N m_i = 1$. Analogous to the formulations for the conservation of momentum and enthalpy, the transport of chemical species results from convection and molecular transport. The term S_i contains sources or sinks based on all possible forms of transformation, for example due to chemical reactions or due to phase changes in multiphase systems. If molecular transport is considered to be adequately described by Fick's law (gradient transport approach)

$$J_j^{m_i} = -\,\varrho D_i \frac{\partial m_i}{\partial x_j} \tag{241a}$$

then a simple form of the conservation equations is obtained that is analogous to Equation (240 b)

$$\varrho \frac{\partial m_i}{\partial t} + \varrho u_k \frac{\partial m_i}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\varrho D_i \frac{\partial m_i}{\partial x_k}\right] + \varrho S_i$$
$$i = 1, \ldots, N-1 \tag{265b}$$

In Equation (265 b) thermal and pressure diffusion are neglected. Furthermore in general $S_i = S_i\,(m_i,\, h,\, p)$ (Eq. 137 a) is to be set in general for the transformation term so that Equation (265 b) is coupled with the conservation equation of momentum through the density and velocities and coupled with the conservation equation of enthalpy through the transformation term. The coupled system of equations given in Section 5.2 is thus expanded by one equation.

5.3.1. Mass Transfer without Chemical Reaction

5.3.1.1. Exact Solution for a Boundary Layer Problem

The flat plate in Figure 29 is replaced by a water surface (Fig. 31) from which the water evaporates into the higher temperature air flow. Under the same assumptions as in Section 5.2.2.1 and with the boundary layer approximations for Equation (265 b), the following system of equations is obtained for this steady-state problem:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{246}$$

$$u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = \nu\frac{\partial^2 u}{\partial y^2}, \quad \nu = \frac{\mu}{\varrho} \tag{247}$$

$$u\frac{\partial T}{\partial x} + v\frac{\partial T}{\partial y} = a\frac{\partial^2 T}{\partial y^2}, \quad a = \frac{\lambda}{\varrho c_p} \tag{245b}$$

$$u\frac{\partial m}{\partial x} + v\frac{\partial m}{\partial y} = D\frac{\partial^2 m}{\partial y^2} \tag{266}$$

where $m = m_{\mathrm{H_2O}}$. The equation system is to be solved for the boundary conditions $u = 0$, $v = v_{\mathrm{W}}$, $T = T_{\mathrm{W}}$, and also $m = m_{\mathrm{W}}$ for $y = 0$ and $u = u_\infty$, and $m = m_\infty$ for $y \to \infty$.

Since the inclusion of the conservation equation of mass has not altered the nature of the equations, it is again logical to transform the equations into the similarity coordinates

$$\eta\,(x, y) = y\sqrt{\frac{u_\infty}{\nu x}} = \frac{y}{x}\sqrt{Re_x}$$

(see Eq. 252), and to express them in the dimensionless streams function $f\,(\eta) = \int (u/u_\infty)\,\mathrm{d}\eta$.

Due to the different boundary conditions for temperature and mass fraction and in analogy with $f_\eta = u/u_\infty$, the term $\theta = (T - T_{\mathrm{W}})/(T_\infty - T_{\mathrm{W}})$ is used for the temperature and $\varphi = (m - m_{\mathrm{W}})/(m_\infty - m_{\mathrm{W}})$ for the mass fraction. With the aid of the relationships given in Section 5.2.2.1, these transformations yield the equation system

$$f_{\eta\eta\eta} + \frac{1}{2} f f_{\eta\eta} = 0 \tag{255}$$

$$\theta_{\eta\eta} + \frac{1}{2} Pr f \theta_\eta = 0 \tag{267}$$

$$\varphi_{\eta\eta} + \frac{1}{2} Sc f \varphi_\eta = 0 \quad \text{where} \quad Sc = \nu/D \tag{268}$$

that has to be solved simultaneously.

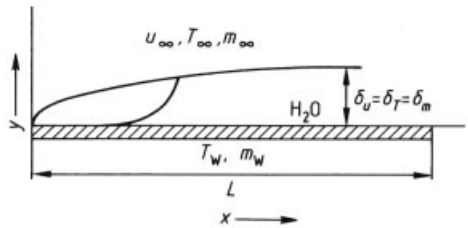

Figure 31. Development of a velocity (δ_u), temperature (δ_T), and concentration (δ_m) boundary layer for laminar incompressible flow over a flat water surface
$Pr = Sc = 1$
The subscripts ∞ and W indicate values outside the boundary layer and at the wall, respectively.

The boundary conditions with the transformed variables are $f_\eta = 0$, $\theta = 0$, $\varphi = 0$ and also $(v_W/u_\infty)\sqrt{Re_x} = \text{const.}$ for $\eta = 0$ and $f_\eta = 1$, $\theta = 1$, $\varphi = 1$ for $\eta \to \infty$.

The system of Equations (255), (267), (268) cannot be solved analytically, but for $Le = Pr/Sc = 1$ all profiles are identical and thus result in identical values for δ_u, δ_T, and δ_m. All transport mechanisms exhibit the same development of the boundary layer. For $Pr = Sc = 1$ and $Pr = Sc = 0.7$ and also some values of $(v_W/u_\infty)\sqrt{Re_x}$, Figure 32 shows the profiles $f_\eta = u/u_\infty$, θ, and φ [59]. The curve for $Pr = Sc = 1$ and $(v_W/u_\infty)\sqrt{Re_x} = 0$ corresponds to the profile discussed in the previous sections and is called the Blasius profile. The value of $\eta = 5$ for $\theta = \varphi = u/u_\infty \approx 0.99$ can be read off from the curve for $(v_{\mathrm{W}}/u_\infty)\sqrt{Re_x} = 0$. Figure 32 shows that, depending on

the magnitude of the "blow out parameter" $(v_W/u_\infty)\sqrt{Re_x}$, the profiles of the flat plate will be either blown off or sucked on. For many technical applications this is taken advantage of for stabilization or removal of boundary layers.

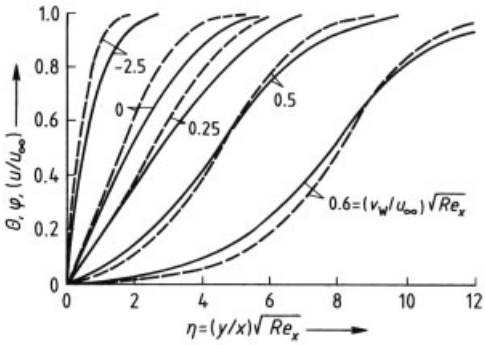

Figure 32. Concentration and temperature distribution with mass transfer for laminar flow over a flat surface
The curves for $Pr = Sc = 1$ indicated by the dashed lines also give the velocity profiles. The curves for $Pr = Sc = 0.7$ indicated by the solid lines show temperature and concentration distributions.
$\theta = (T - T_W)/(T_\infty - T_W)$; $\varphi = (m - m_W)/(m_\infty - m_W)$; η= similarity coordinate $= (y/x)\, Re_x\sqrt{Re_x}$; v = velocity in y-direction; u = velocity in x-direction; Re_x = Reynolds number defined with u_∞ and axial distance x; the subscripts W and ∞ indicate values at the wall and outside the boundary layer, respectively.

In the case of the evaporation example shown in Figure 31, $(v_W/u_\infty)\sqrt{Re_x}$ can be neglected so that the mass transfer, which is to be defined analogously to the heat transfer, is written in the form

$$J_W^m = -\varrho D\left(\frac{\partial m}{\partial y}\right)_W = -\varrho D\,(m_\infty - m_W)\sqrt{\frac{u_\infty}{\nu x}}\left(\frac{\partial \varphi}{\partial \eta}\right)_W \tag{269}$$

If the Blasius solution $f_{\eta\eta} = \varphi_\eta = \theta_\eta = 0.332$ is used for $Pr = Sc = 1$ and, analogous to the heat-transfer coefficient, a mass-transfer coefficient is defined as $J_W^m = \varrho\,\beta\,\Delta m$ where $\Delta m = (m_W - m_\infty)$, it follows that

$$\beta = 0.332D\sqrt{\frac{u_\infty}{\nu x}} = 0.332D\frac{1}{x}\sqrt{Re_x} \tag{270}$$

In order to complete the analogy with heat transfer the Sherwood number $Sh_x = \beta\, x/D$ should be substituted. Thus for $Pr = Sc = 1$

$$\frac{Sh_x}{\sqrt{Re_x}} = 0.332 \tag{271}$$

Further treatment of mass-transfer coefficients – for example averaging over the length – follows the scheme given in Section 5.2.2.1 and need not be repeated here.

5.3.1.2. General Principles for Modeling Mass Transfer

As mentioned in Section 5.2.2.2 for heat transfer, only in very few cases do the Prandtl or Schmidt numbers have the value one; in addition the geometric or hydrodynamic assumptions listed in Section 5.3.1.1 seldom apply. Even if the system of Equations (204), (240 b), and (265 b) can be solved under such conditions, the resulting solutions deviate from the complete analogy.

For engineering applications mass transfer is therefore also treated by semi-empirical models which assume a linearization of the profile of the concentrations at the phase boundary. Because of the analogy with heat transfer (Fig. 30) discussion of this topic can be kept short. If the mass fraction profile is linearized at the phase boundary

$$m = m_W + \left(\frac{\partial m}{\partial y}\right)_W \mathrm{d}y \quad \text{or} \quad m_\infty = m_W + \left(\frac{\partial m}{\partial y}\right)_W \delta \tag{272}$$

the equation defining the mass-transfer coefficient results in

$$D\frac{m_\infty - m_W}{\delta} = \beta\,(m_\infty - m_W) \quad \text{or} \quad \beta = \frac{D}{\delta} \tag{273a,b}$$

where δ again represents the film thickness produced by linearization of the mass fraction profile. Determination of the mass-transfer coefficient thus depends on the determination of the film thickness, which is carried out experimentally for many engineering systems. The results for the mean mass-transfer coefficients in the form of the relationships $Sh = f\,(Re, Sc, L/D)$ are described in [25], [59], [109]. For evaluation of industrial mass transfer equipment there remains the problem, analogous to heat transfer, of the addition of the "individual resistances" for mass transfer and reference to the total concentration difference.

5.3.1.3. Examples of Simultaneous Mass and Heat Transfer: Dynamic Models

The system used to illustrate the general principles of the modeling of simultaneous heat and mass transfer (Sections 5.2.2.2 and 5.3.1.2) is based on the boundary layer example shown in Figure 31. Figure 33 shows an evaporation cooler which is used in an air-conditioning device. A liquid (water) trickles in a thin film over suitable packings and comes into contact with a gas (air) flowing in the same direction. The liquid evaporates and is thus cooled, so that heat flow occurs from gas to liquid thus lowering the gas temperature.

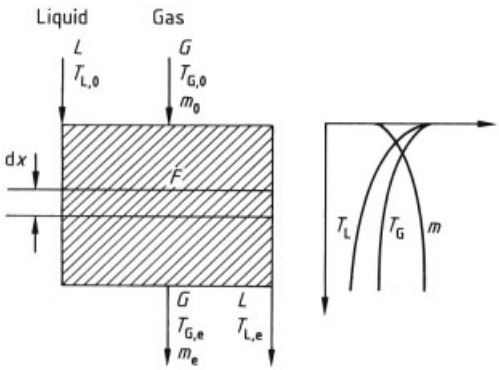

Figure 33. Example of simultaneous mass and heat transfer in an evaporation cooler
L = mass flow of liquid; G = mass flow of gas; F = cross sectional area; x = axial distance
The subscripts L, G, 0, and e denote liquid, gas, initial, and final, respectively.

The final steady-state temperature of the gas must be determined for a given area, $S = s\,F\,H$, where s is the specific transfer area. All terms are defined in Figure 33. The thickness of the layer of liquid flowing over the packings is assumed to be so small that the temperature in the liquid phase is uniform. The mean heat- and mass-transfer coefficients for this problem are known. The temperature changes lie in a range which allows the material values to be regarded as constants. The functions for the Nusselt and Sherwood numbers generally contain terms which are functions of the properties of the material. The assumption of approximately constant material properties avoids recalculation of the example with improved Nusselt and Sherwood numbers. The mass transfer is so small compared with the mass flow of gas and liquid (G and L, respectively) that the latter can likewise be regarded as nearly constant. The required quantities only change in the x-direction so that the problem can be formulated one-dimensionally.

The heat balance for the gas phase gives the temperature change in the gas based on the heat transfer to the liquid

$$\frac{\mathrm{d}T_\mathrm{G}}{\mathrm{d}x} = -\alpha \frac{sF}{Gc_{p\mathrm{G}}}\,(T_\mathrm{G} - T_\mathrm{L}) \tag{274}$$

where c_p is the specific heat. The mass transfer to the gas phase gives the change in mass fraction in the gas phase

$$\frac{\mathrm{d}m}{\mathrm{d}x} = \beta \frac{\varrho sF}{G}\,(m_\mathrm{Ph} - m) \tag{275}$$

where β is the mass-transfer coefficient and m_Ph the mass fraction of liquid in the phase boundary. The temperature change in the liquid phase is finally obtained from the heat balance for the liquid phase

$$\frac{\mathrm{d}T_\mathrm{L}}{\mathrm{d}x} = \alpha \frac{sF}{L\,c_{p\mathrm{L}}}\,(T_\mathrm{G} - T_\mathrm{L}) - \beta \frac{\Delta h_\mathrm{v}\,\varrho sF}{L\,c_{p\mathrm{L}}}\,(m_\mathrm{Ph} - m) \tag{276}$$

where Δh_v is the specific latent heat of vaporization of the liquid.

Equations (274) – (276) represent a system of simultaneous ordinary differential equations for T_G, T_L, and m. In order to close these equations a relationship has to be found for m_Ph. This can be obtained from the Clausius–Clapeyron equation for vapor pressure $\ln p = -A'/T + B'$. Since the changes are assumed to be very small, a linearized expression can be used in the temperature interval being considered, so that $m_\mathrm{Ph} = A\,T_\mathrm{L} + B$ or $\mathrm{d}m_\mathrm{Ph}/\mathrm{d}x = A\,(\mathrm{d}T_\mathrm{L}/\mathrm{d}x)$. By substituting the transformations $\theta = T_\mathrm{G} - T_\mathrm{L}$ and $\varphi = m_\mathrm{Ph} - m$, the number of equations can be reduced by one:

$$\frac{\mathrm{d}\theta}{\mathrm{d}x} = -\alpha \frac{sF}{Gc_{p\mathrm{G}}\,L\,c_{p\mathrm{L}}}\,(Gc_{p\mathrm{G}} + L\,c_{p\mathrm{L}})\,\theta + \beta \frac{\Delta h_\mathrm{v}\,\varrho sF}{L\,c_{p\mathrm{L}}}\,\varphi \tag{277a}$$

$$\frac{\mathrm{d}\varphi}{\mathrm{d}x} = \alpha \frac{AsF}{L\,c_{p\mathrm{L}}}\,\theta + \left[\beta \frac{A\Delta h_\mathrm{v}\,\varrho sF}{L\,c_{p\mathrm{L}}} + \beta \frac{\varrho sF}{G}\right]\varphi \tag{277b}$$

Equations (277 a) and (277 b) can also be abbreviated as

$$\theta_x = a_1\theta + b_1\varphi \tag{278a}$$

$$\varphi_x = a_2\theta - b_2\varphi \tag{278b}$$

where θ_x and φ_x are the derivatives of θ and φ with respect to x. Differentiation of Equation (278 a) and elimination of φ_x and φ give the ordinary homogeneous second-order differential equation

$$\theta_{xx} + (a_1 + b_2)\,\theta_x + (a_1 b_2 - a_2 b_1)\,\theta = 0 \tag{279}$$

to which the system of coupled equations is now reduced. Since $(a_1 + b_2)^2 > 4\,(a_1\,b_2 - a_2\,b_1)$, the characteristic equation of differential Equation (279) $\lambda^2 + (a_1 + b_2)\,\lambda + (a_1\,b_2 - a_2\,b_1) = 0$ has two real roots. The general solution for Equation (279) is thus

$$\theta = C_1 \mathrm{e}^{\lambda_1 x} + C_2 \mathrm{e}^{\lambda_2 x} \tag{280a}$$

The integration constants C_1 and C_2 are determined from the boundary conditions, which are simplified here as $T_{\mathrm{L0}} = T_{\mathrm{G0}}$, since water and air are both assumed to have the same (room) temperature. This leads to $\theta = 0$ and $\theta_x = b_1\,\varphi_0$ for $x = 0$. These conditions give $C_1 = -C_2$ and $\lambda_1 C_1 + \lambda_2 C_2 = b_1\,\varphi_0$. Thus the solution of Equation (279) is

$$\theta = \frac{b_1\varphi_0}{\lambda_1 - \lambda_2}\left[\mathrm{e}^{\lambda_1 x} - \mathrm{e}^{\lambda_2 x}\right] \tag{280b}$$

where φ_0 is to be determined from the known initial temperature of the water and the initial mass fraction of water in the air; λ_1 and λ_2 have to be calculated from the characteristic equation. Equation (280 b) shows that cooling is initiated entirely by mass transfer into the gas phase. The gas temperature can now be determined with the known solution for θ by integration of Equation (274):

$$T_{\mathrm{G}} = T_{\mathrm{G0}} - a_3 \frac{b_1\varphi_0}{(\lambda_1 - \lambda_2)}\left[\frac{1}{\lambda_1}\mathrm{e}^{\lambda_1 x} - \frac{1}{\lambda_2}\mathrm{e}^{\lambda_2 x}\right] \tag{281}$$

where $a_3 = \alpha s F / G c_{p\mathrm{G}}$

5.3.1.4. Examples of Simultaneous Mass and Heat Transfer: Static Models

In the examples outlined in Section 5.3.1.3, the changes in temperature and mass fractions were calculated from the rates of transport of enthalpy and mass between the individual phases. The concept used in this model is shown in Figure 34 A for the mass transfer of a component from the gas phase into a counterflowing liquid phase. At the phase boundary the mass fraction profile shows a jump caused by the relevant phase equilibrium which may be expressed by the relationship e.g., $m_{\mathrm{PhG}} = A m_{\mathrm{PhL}}$. In contrast to the cases discussed in Sections 5.3.1.1 and 5.3.1.3 the component being transferred is dissolved in the liquid phase. The mass fraction differences between the bulk phases and the phase boundary (Fig. 34 A) can be expressed as

$$(m_{\mathrm{G}} - m_{\mathrm{PhG}}) = \frac{j^m}{\varrho_{\mathrm{G}}\beta_{\mathrm{G}} s F \mathrm{d}x} \tag{282a}$$

and

$$(m_{\mathrm{PhL}} - m_{\mathrm{L}}) = \frac{j^m}{\varrho_{\mathrm{L}}\beta_{\mathrm{L}} s F \mathrm{d}x} \tag{282b}$$

where j^{m} is the mass flux of the transferred component. If Equations (282 a) and (282 b) are added and the relationship for the phase equilibrium is used to eliminate m_{PhL}, then the following equation is obtained for the steady-state case, taking into consideration the mass balance for the gas phase:

$$j^m = \frac{sF}{\frac{1}{\varrho_{\mathrm{G}}\beta_{\mathrm{G}}} + \frac{A}{\varrho_{\mathrm{L}}\beta_{\mathrm{L}}}}\,(m_{\mathrm{G}} - A m_{\mathrm{L}})\,\mathrm{d}x = -\mathrm{d}m_{\mathrm{G}} G \tag{283}$$

Rearrangement gives

$$-\frac{\mathrm{d}m_{\mathrm{G}}}{(m_{\mathrm{G}} - A m_{\mathrm{L}})} = \frac{sF}{\left(\frac{1}{\varrho_{\mathrm{G}}\beta_{\mathrm{G}}} + \frac{A}{\varrho_{\mathrm{L}}\beta_{\mathrm{L}}}\right) G}\mathrm{d}x \tag{284}$$

and after integration

$$\int_{m_{\mathrm{G}e}}^{m_{\mathrm{G0}}} \frac{\mathrm{d}m_{\mathrm{G}}}{(m_{\mathrm{G}} - A m_{\mathrm{L}})} = \frac{sF}{\left(\frac{1}{\varrho_{\mathrm{G}}\beta_{\mathrm{G}}} + \frac{A}{\varrho_{\mathrm{L}}\beta_{\mathrm{L}}}\right) G}\int_0^H \mathrm{d}x \tag{285a}$$

which can be abbreviated as

$$\mathrm{NTU} = \frac{H}{\mathrm{HTU}} \tag{285b}$$

where NTU denotes the number of transfer units and HTU the height of one transfer unit.

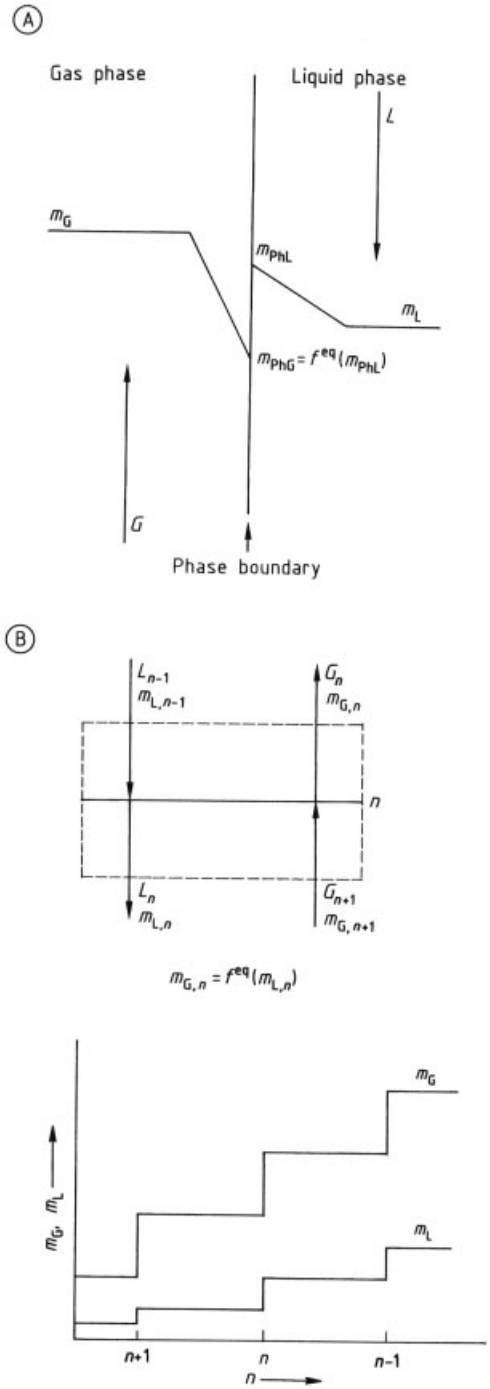

Figure 34. Representation of the dynamic (A) and static (B) models for mass transfer
L = mass flow of liquid; G = mass flow of gas
The subscripts L, G, Ph, and n denote liquid, gas, phase boundary, and the number of theoretical plates, respectively. The superscript eq denotes equilibrium.

The notation and assumptions for the derivation of Equation (285 b) are the same as those in Section 5.3.1.3. The same procedure as indicated in Section 5.2.2.2 for the heat transfer leads to Equation (285 b). A similar relationship is often used for total heat transfer in heat exchangers [108].

Equation (285 b) is used to determine the height of a separation column with a known specific area and cross-sectional area or conversely to determine the concentration difference for a given height of the separation column. The quantity Am_L contained in the integral N T U still has to be expressed in terms of m_G and the boundary conditions $m_{G,0}$ and $m_{L,0}$ or $m_{L,e}$. All quantities in the H T U expression are known.

A completely different model for mass and heat transfer in mass separation processes is shown in Figure 34 B. Here the mass transfer device consists of a series of separation stages, in which phase equilibrium occurs between the counterflowing phases. Mass transfer is assumed to occur infinitely quickly and the separation stage is homogeneous with regard to the variables of state. The mass balance in the steady-state case for the components transferred is then

$$Lm_{L,n-1}+Gm_{G,n+1}=Lm_{L,n}+Gm_{G,n} \tag{286}$$

Since phase equilibrium should exist in the separation stage, then $m_{G,n}$ can be expressed by $m_{L,n}$. Due to the modeling assumptions, the mass balance does not contain any transport terms. If the relationship for the phase equilibrium is expressed in the form $m_G = A\,m_L + B$, Equation (286) becomes

$$Lm_{L,n-1}+G\left(Am_{L,n+1}+B\right) = Lm_{L,n}+G\left(Am_{L,n}+B\right) \tag{287a}$$

or

$$Lm_{L,n-1}-(L+AG)\,m_{L,n}+AGm_{L,n+1} = 0 \tag{287b}$$

For a separation column with N separation stages (Fig. 35) for given mass fractions $m_{G,0}$ and $m_{L,0}$, the relevant linear equation system is

$$-(L+AG)\,m_{L,1}+AGm_{L,2}=-Lm_{L,0} \tag{288,1}$$

⋮

$$Lm_{\mathrm{L},n-1} - (L + AG)\, m_{\mathrm{L},n} + AGm_{\mathrm{L},n+1} = 0 \tag{288,n}$$

$$\vdots$$

$$Lm_{\mathrm{L},N-1} - (L + AG)\, m_{\mathrm{L},N} = -G\left(m_{\mathrm{G},0} - B\right) \tag{288,N}$$

In matrix form this can be written as

$$\begin{pmatrix} -(L+AG) & AG & \dots \\ L & -(L+AG)\,AG & \dots \\ \vdots & \ddots \quad \ddots & \vdots \\ \dots & L & -(L+AG) \end{pmatrix} \times \begin{pmatrix} m_{\mathrm{L},1} \\ \vdots \\ m_{\mathrm{L},N} \end{pmatrix} = \begin{pmatrix} -Lm_{\mathrm{L},0} \\ \vdots \\ -G\left(m_{\mathrm{G},0} - B\right) \end{pmatrix}$$

or

$$\boldsymbol{Bm} = \boldsymbol{c} \tag{288b}$$

The solution of this equation with one of the methods discussed in Section 3.1.2 is

$$\boldsymbol{m} = \boldsymbol{B}^{-1}\boldsymbol{c} \tag{289}$$

and provides the mass fraction profile in discrete form for the component being transferred to the liquid phase over the entire length of the column. Using the appropriate relationships for phase equilibrium the corresponding profile can easily be determined for the gas phase. The system of equations could equally well be derived for the mass fractions in the gas phase.

The above problem can easily be solved in another way. Transformation of Equation (286) gives

$$m_{\mathrm{L},1} - m_{\mathrm{L},0} = \frac{G}{L}\left(m_{\mathrm{G},2} - m_{\mathrm{G},1}\right) \quad \text{or} \tag{290a}$$

$$\Delta m_{\mathrm{L}} = \frac{G}{L}\Delta m_{\mathrm{G}} \tag{290b}$$

for the first separation stage in the column shown in Figure 35. The geometric interpretation of this equation is to find the point with coordinates ($m_{\mathrm{L},1}$, $m_{\mathrm{G},2}$) on a straight line of slope (L/G) starting from the point with the coordinates ($m_{\mathrm{L},0}$, $m_{\mathrm{G},1}$). As the first step, a point ($m_{\mathrm{L},1}$, $m_{\mathrm{G},1}$) has to be found that satisfies the phase equilibrium relationship $m_{\mathrm{G}} = Am_{\mathrm{L}} + B$. This is done by inserting $m_{\mathrm{G},1}$ into the phase equilibrium relationship as shown by the horizontal line in Figure 36. Working from this point, point ($m_{\mathrm{L},1}$, $m_{\mathrm{G},2}$) can be found by inserting $m_{\mathrm{L},1}$ into the mass balance equation (Eq. 290 a); this is shown in Figure 36 by the vertical line. In this way Equation (288) can easily be solved graphically by completing the construction for all separation stages step by step. The graphical solution method is the basis for the McCabe – Thiele method and similar procedures for calculating mass separation equipment (→ Distillation and Rectification, Chap. 4.1.1.).

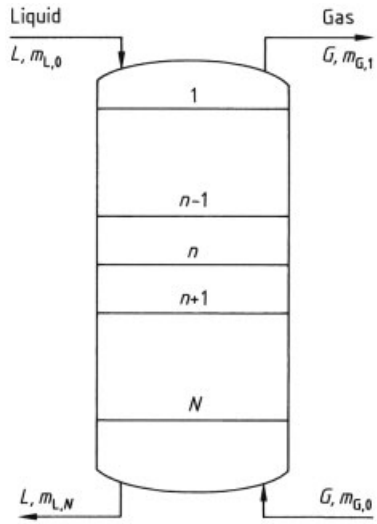

Figure 35. Schematic representation of a counterflow separation column
For explanation of symbols see Figure 34.

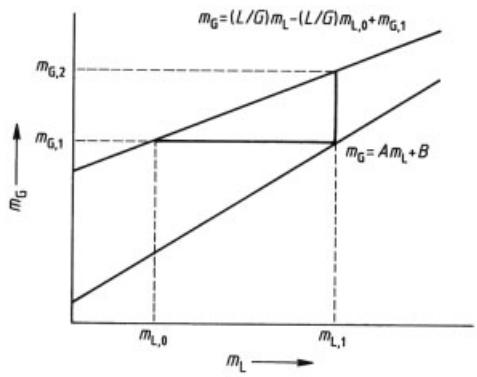

Figure 36. Graphical solution of the equation system (288) for a counterflow separation column: McCabe – Thiele diagram
A and B are coefficients of the relation for thermodynamic phase equilibrium $m_{\mathrm{G}} = f^{\mathrm{eq}}(m_{\mathrm{L}})$. For explanation of other symbols see Figure 34.

Static models can also be used for simultaneous heat and mass transfer in mass separation processes. This will be described briefly for the separation stage of the rectification column shown in Figure 37. Phase equilibrium is assumed in the separation stage and the separation stage is regarded as being homogeneous with respect to the variables of state. Phase equilibria relationships exist both for the mass fractions and the enthalpies. With these preconditions, balances for the total mass, the mass of the individual components of the mixture, and the total enthalpy can be given.

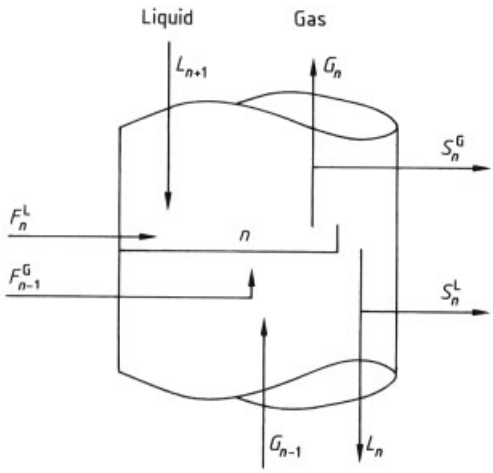

Figure 37. Mass flows in a separation stage of a rectification column
F = liquid or gaseous feed; S = liquid or gaseous sink
The superscripts L and G denote liquid and gas phase, respectively. For explanation of other symbols see Figure 34.

For the sake of simplicity the mixture is taken to be a binary mixture of components A and B. The balances are considered for the steady-state case. (The mass and enthalpy balances can, however, also be formulated without difficulty for a multicomponent mixture or for the nonsteady state, for details see [3], [8], [9].) The equation for the total mass is (notation is defined in Fig. 37)

$$L_{n+1}+G_{n-1}+F_n^{\mathrm{L}}+F_{n-1}^{\mathrm{G}} = L_n + G_n + S_n^{\mathrm{G}}+S_n^{\mathrm{L}} \tag{291}$$

Analogously, the equation for the mass of a component is

$$L_{n+1}m_{\mathrm{L},n+1}+G_{n-1}m_{\mathrm{G},n-1}+F_n^{\mathrm{L}}m_{\mathrm{L},n} + F_{n-1}^{\mathrm{G}}m_{\mathrm{G},n-1} = L_n m_{\mathrm{L},n}+G_n m_{\mathrm{G},n}+S_n^{\mathrm{G}}m_{\mathrm{G},n} + S_n^{\mathrm{L}}m_{\mathrm{L},n} \tag{292}$$

and finally that for the total enthalpy is

$$h_{\mathrm{L},n+1}+h_{\mathrm{G},n-1}+h_{\mathrm{L},F_n^{\mathrm{L}}}+h_{\mathrm{G},F_{n-1}^{\mathrm{G}}} = h_{\mathrm{L},n}+h_{\mathrm{G},n}+h_{\mathrm{G},S_n^{\mathrm{G}}}+h_{\mathrm{L},S_n^{\mathrm{L}}} \tag{293}$$

For a rectification column of N separation stages, for example, the mass fraction in the liquid phase, the mass flows of the liquid phase, and the temperature have to be calculated. To do this Equations (291) – (293) are simplified by setting all flows to and from the external surroundings to zero (see also Fig. 37). This simplification does not change the character of the resulting equations.

The expansions of the equations for inputs or outputs at the individual separation stages, and likewise the formulations for the condensor at the top of the column and the reboiler at the bottom of the column, offer no major difficulties. To reduce Equations (291) – (293) to a system of equations in L, m_{L}, and T_{L} further relationships are needed to eliminate m_{G} and G. Relationships for the phase equilibrium are used for this: $m_{\mathrm{G}}=f^{\mathrm{eq}}(m_{\mathrm{L}})$, and $G_n/G_{n-1}=R_n$.

Here f^{eq} is the relation for phase equilibrium and R_n is the ratio of the evaporation enthalpies $R_n=\Delta h_{\mathrm{v},\,n}/\Delta h_{\mathrm{v},\,n-1}$. If mixing effects are neglected the enthalpies are additive, so that for example $h_{\mathrm{L}}=(h_{\mathrm{L}}^{\mathrm{A}}-h_{\mathrm{L}}^{\mathrm{B}})\,m_{\mathrm{L}}+h_{\mathrm{L}}^{\mathrm{B}}$. Finally, neglecting the mixing effects, $p_{\mathrm{tot}}=[p^{\mathrm{A}}(T)-p^{\mathrm{B}}(T)]\,m_{\mathrm{L}}+p^{\mathrm{B}}(T)$ is used to obtain a correlation between the temperature and the composition of the liquid phase. Using $h=h_0+c_p(T-T_0)$, the enthalpies can finally be eliminated. The resulting system of equations is

$$\left[p^{\mathrm{A}}(T_n)-p^{\mathrm{B}}(T_n)\right]m_{\mathrm{L},n}+p^{\mathrm{B}}(T_n)=p_{\mathrm{tot}} \tag{294a}$$

$$\left[m_{\mathrm{L},n}-\left(\frac{1}{1+R_n}\right)f^{\mathrm{eq}}(m_{\mathrm{L},n}) + \left(\frac{R_n}{1-R_n}\right)f^{\mathrm{eq}}(m_{\mathrm{L},n-1})\right]L_n + \left[\left(\frac{1}{1+R_n}\right)f^{\mathrm{eq}}(m_{\mathrm{L},n})-\left(\frac{R_n}{1-R_n}\right)f^{\mathrm{eq}}(m_{\mathrm{L},n-1}) - m_{\mathrm{L},n-1}\right]L_{n+1}=0 \tag{294b}$$

$$\left[\left(\frac{R_n}{1-R_n}\right)(L_{n+1}-L_n)\right]$$
$$\cdot\left\{\left[\left(h_{\mathrm{L}}^{\mathrm{A}}+\Delta h_{\mathrm{v}}^{\mathrm{A}}\right)-\left(h_{\mathrm{L}}^{\mathrm{B}}+\Delta h_{\mathrm{v}}^{\mathrm{B}}\right)\right]m_{\mathrm{L},n-1}\right.$$
$$\left.+\left(h_{\mathrm{L}}^{\mathrm{B}}+\Delta h_{\mathrm{v}}^{\mathrm{B}}\right)\right\}-L_n\left[\left(h_{\mathrm{L}}^{\mathrm{A}}-h_{\mathrm{L}}^{\mathrm{B}}\right)m_{\mathrm{L},n}+h_{\mathrm{L}}^{\mathrm{B}}\right]$$
$$-\left[\left(\frac{1}{1-R_n}\right)(L_{n+1}-L_n)\right]$$
$$\cdot\left\{\left[\left(h_{\mathrm{L}}^{\mathrm{A}}+\Delta h_{\mathrm{v}}^{\mathrm{A}}\right)-\left(h_{\mathrm{L}}^{\mathrm{B}}+\Delta h_{\mathrm{v}}^{\mathrm{B}}\right)\right]m_{\mathrm{L},n}\right.$$
$$\left.+\left(h_{\mathrm{L}}^{\mathrm{B}}+\Delta h_{\mathrm{v}}^{\mathrm{B}}\right)\right\}$$
$$+L_{n+1}\left[\left(h_{\mathrm{L}}^{\mathrm{A}}-h_{\mathrm{L}}^{\mathrm{B}}\right)m_{\mathrm{L},n+1}+h_{\mathrm{L}}^{\mathrm{B}}\right]=0 \qquad (294c)$$

In Equations (294) the relationships for $m_{\mathrm{G}}=f^{\mathrm{eq}}(m_{\mathrm{L}}), p^i(T), h_{\mathrm{L}}^i(T)$, and R_n are not explicitly formulated. In this form Equation (294) should be adequate to point to the similarity to Equation (288). A system of N equations is obtained whose system matrix exhibits a distorted tridiagonal structure. Each element consists of a block with 3×3 elements. In contrast to Equation (288) the sytem is nonlinear and coupled.

Consideration of output and input at the individual separation stages or at the top and bottom of the column does not alter the structure of this system of equations; further terms are simply added to the right-hand side and the dimensions of the blocks of the system matrix are increased on account of consideration of a multicomponent mixture. Equation (294) is easy to solve because of the conditioning of the system matrix. Equation (294 a) contains entries only in the main diagonals, and Equation (294 b) on two diagonals. In most applications the system is not therefore solved simultaneously, but is solved successively starting with Equation (294 a) and predefined boundary conditions. Examples of computer programs for the solution of these problems with multicomponent formulations and nonsteady-state conditions are found in [8], [9].

5.3.2. Mass Transfer with Chemical Reactions

In the dynamic and static models for calculating simultaneous mass and heat transfer, chemical reactions have not yet been considered. In many industrial mass separation processes, mass transfer is associated with chemical reactions. The effect of chemical reactions on mass transfer is now discussed using a simple example (Fig. 38) based on the simple boundary layer problem given in Sections 5.2.2.1 and 5.3.1.1. The plate is replaced by a soluble component A which passes into the fluid phase where it reacts with a second component B. The reaction rate is first order with respect to the individual components, the reaction rate is thus

$$\frac{\mathrm{d}m_{\mathrm{A}}}{\mathrm{d}t}=-k'(T)\,m_{\mathrm{A}}m_{\mathrm{B}} \qquad (295)$$

The component B is assumed to be present in excess such that $m_{\mathrm{B}}\approx$ const. If $m=m_{\mathrm{A}}$, then Equation (295) can be simplified as

$$\frac{\mathrm{d}m}{\mathrm{d}t}=-k(T,m_{\mathrm{B}})\,m \qquad (296)$$

Using the same assumptions as in Sections 5.2.2.1 and 5.3.1.1 (the boundary layer approximations and negligible reaction enthalpy), the following system of equations is obtained

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \qquad (246)$$

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\nu\frac{\partial^2 u}{\partial y^2}, \qquad (247)$$

$$u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}=a\frac{\partial^2 T}{\partial y^2} \qquad (245b)$$

From the addition of a source term S_{A} to Equation (266) and with $S_A=\mathrm{d}m/\mathrm{d}t$:

$$u\frac{\partial m}{\partial x}+v\frac{\partial m}{\partial y}=D\frac{\partial^2 m}{\partial y^2}-k(T,m_{\mathrm{B}})\,m \qquad (297)$$

Due to the simplification used in the formulation of Equation (296), the conservation equations for the chemical species are no longer coupled.

Equation system (246), (247), (245 b), and (297) cannot be solved analytically. Numerical solutions for this problem will not be discussed here. However, also for $Pr=Sc=1$ the similarity of the equations is lost due to the source term $-k(T,m_{\mathrm{B}})\,m$ in Equation (297) originating from the chemical reaction. Thus the boundary layer δ_m develops differently from δ_u and δ_T.

Since the source term in Equation (297) is negative, a larger gradient of the mass fraction of A occurs and this leads to a thinner boundary layer δ_m in comparison with δ_u, δ_T. From the definition of the mass-transfer number (see Section 5.3.1.1) it can be seen that the local mass-transfer coefficient increases due to the larger gradients. Mass transfer is accelerated by the chemical reaction.

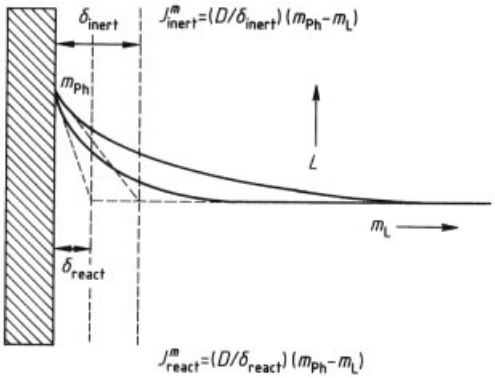

Figure 38. Acceleration of mass transfer by chemical reactions
δ = film thickness; J^m = mass fluxes; D = diffusion coefficient
For explanation of other symbols see Figure 34.

For a more thorough discussion of this acceleration effect, the boundary layer problem is further simplified by considering the situation at a large distance from the edge of the plate. Since $\delta_u \sim \sqrt{x}$, the gradients of u, T, and m in the x-direction become extremely small, and v approaches zero. Thus for $x \to \infty$ Equation (297) is simplified to

$$D\frac{\mathrm{d}^2 m}{\mathrm{d}y^2} - k\,(T, m_\mathrm{B})\,m \approx 0 \tag{298}$$

If y is transformed into $\eta = y/\delta_m$, then

$$D\frac{\mathrm{d}^2 m}{\mathrm{d}\eta^2} - Ha^2 m = 0 \tag{299}$$

where

$$Ha = \delta_m\sqrt{\frac{k}{D}} = \frac{\delta_m}{D}\sqrt{Dk} = \frac{1}{\beta}\sqrt{Dk}$$

is the Hatta number. The boundary conditions given in Figure 38 are $m = m_\mathrm{Ph}$ for $\eta = 0$ and $m = m_\mathrm{L}$ for $\eta = 1$. The characteristic equation of the ordinary homogeneous differential Equation (299), $\lambda^2 - Ha^2 = 0$ has the solutions $\lambda_1 = Ha$ and $\lambda_2 = -Ha$ and thus the general solution for Equation (299) is

$$m = C_1 \mathrm{e}^{Ha\eta} + C_2 \mathrm{e}^{-Ha\eta} \tag{300}$$

From the above boundary conditions

$$C_1 = \frac{1}{\mathrm{e}^{Ha} - \mathrm{e}^{-Ha}}\left[m_\mathrm{L} - m_\mathrm{Ph}\mathrm{e}^{-Ha}\right]$$

and

$$C_2 = -C_1 = \frac{1}{\mathrm{e}^{Ha} - \mathrm{e}^{-Ha}}\left[m_\mathrm{Ph}\mathrm{e}^{-Ha} - m_\mathrm{L}\right]$$

so that the solution becomes

$$m = \frac{1}{\sinh Ha}\left[m_\mathrm{Ph}\sinh\left(Ha\,(1-\eta)\right)\right.$$
$$\left. + m_\mathrm{L}\sinh\left(Ha\eta\right)\right] \tag{301}$$

Mass transfer is defined as usual by the condition

$$J^m_\mathrm{W} = -\varrho D\left(\frac{\mathrm{d}m}{\mathrm{d}y}\right)_{y=0} = -\varrho D\frac{1}{\delta_m}\left(\frac{\mathrm{d}m}{\mathrm{d}\eta}\right)_{\eta=0}$$

From Equation (301) this condition gives

$$(J^m_\mathrm{W})_\mathrm{react} = \varrho\frac{D}{\delta_m}\frac{Ha}{\sinh Ha}\left[m_\mathrm{Ph}\cosh Ha - m_\mathrm{L}\right] \tag{302}$$

If the mass-transfer rates from Equation (302) are compared with the condition without reaction (indicated by the subscript inert)

$$(J^m_\mathrm{W})_\mathrm{inert} = \varrho\frac{D}{\delta_m}\left(m_\mathrm{Ph} - m_\mathrm{L}\right) \tag{303}$$

the accelerating effect of the chemical reaction is immediately obvious. The ratio of Equations (302) and (303) gives the "acceleration factor" β_r, i.e., the factor by which mass transfer is accelerated as a result of the chemical reaction:

$$\beta_\mathrm{r} = \frac{Ha}{\tanh Ha}\frac{m_\mathrm{Ph}}{m_\mathrm{Ph} - m_\mathrm{L}}\left[1 - \frac{m_\mathrm{L}}{m_\mathrm{Ph}\cosh Ha}\right] \tag{304a}$$

For $m_\mathrm{L} = 0$

$$\beta_\mathrm{r} = \frac{Ha}{\tanh Ha} \tag{304b}$$

which in turn, for $Ha > 3$, gives

$$\beta_\mathrm{r} = Ha \tag{304c}$$

The acceleration of mass transfer is therefore proportional to the square root of the rate coefficient of the reaction rate. This only applies,

however, if the assumptions made in deriving the equation for the expression of the reaction rate (Eq. 296) are fulfilled. For very fast reactions, i.e., for very high Hatta numbers, this is no longer the case.

The concentration of component B is then no longer "unlimited". On the contrary, transport of B into the boundary layer becomes rate-determining. The simultaneous solution for the conservation equations for A and B produces a result which is given in Figure 39 [110]. For $Ha >> 3$

$$\beta_{\mathrm{r}} = 1 + \frac{\nu_{\mathrm{A}} m_{\mathrm{BL}} D_{\mathrm{B}}}{\nu_{\mathrm{B}} m_{\mathrm{APh}} D_{\mathrm{A}}}$$

The accelerating effect of the reaction is controlled by the limited availability of B.

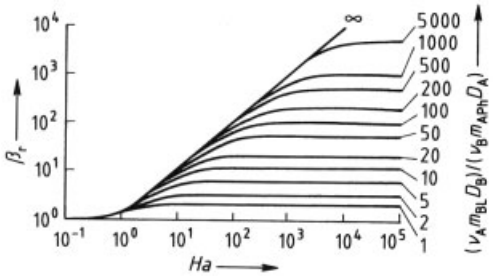

Figure 39. Acceleration factor β_{r} as a function of the Hatta number *Ha* for various stoichiometric characteristic values [110]
ν = stoichiometric coefficient; D = diffusion coefficient
The subscripts A and B denote components, L and Ph denote liquid and phase boundary, respectively.

Introduction of the acceleration factor β_{r} means that the effect of the chemical reactions is incorporated into the mass transfer; the chemical reactions are no longer coupled with the mass transfer. Further aspects of mass transfer with chemical reactions are discussed in Section 5.3.3.1.

5.3.3. Chemical Reactions in the Homogeneous Phase

Many chemical reactors are single phase. The term single phase should be understood here to mean that mass transfer from one phase to the other is not rate determining. Many heterogeneous systems also come within the term "single phase", for example heterogeneous catalytic reactions, which are dealt with as quasi-single phase. The mathematical treatment of single-phase reactors differs from the problems discussed in the earlier chapters due to the boundary conditions and the models for the processes occurring at and in the phase boundary layers. Definition of the state of the reactive mixture in chemically reactive flows with the fields $\boldsymbol{u}(\boldsymbol{x}, t)$, $p(\boldsymbol{x}, t)$, $m_i(\boldsymbol{x}, t)$, $i = 1, \ldots, N$ and $h(\boldsymbol{x}, t)$ requires the solution of the conservation equations for the total mass, the components of the momentum, the mass of the individual chemical species, and the enthalpy. For closure of the system of equations other fundamental equations may have to be incorporated. With the simplifications introduced in the earlier sections for the molecular transport processes this system of equations can be given in the form

$$\frac{\partial \varrho}{\partial t} + \frac{\partial}{\partial x_k}(\varrho u_k) = 0 \tag{131}$$

for the total mass,

$$\varrho \frac{\partial u_i}{\partial t} + \varrho u_k \frac{\partial u_i}{\partial x_k} = \frac{\partial \tau_{ik}}{\partial x_k} - \frac{\partial p}{\partial x_i} + \varrho g_i \qquad i = 1, 2, 3 \tag{204}$$

for the components of the momentum,

$$\varrho \frac{\partial h}{\partial t} + \varrho u_k \frac{\partial h}{\partial x_k} = \frac{\partial}{\partial x_k} \lambda \frac{\partial T}{\partial x_k} + \varrho S_h \tag{240b}$$

for the enthalpy, and

$$\varrho \frac{\partial m_i}{\partial t} + \varrho u_k \frac{\partial m_i}{\partial x_k} = \frac{\partial}{\partial x_k}\left[\varrho D_i \frac{\partial m_i}{\partial x_k}\right] + \varrho S_i \qquad i = 1, \ldots, N-1 \tag{265b}$$

for the chemical species. Since the system of Equations (131), (204), (240 b), and (265 b) is coupled, simultaneous solution is necessary. This is usually done numerically.

Before discussing a numerical procedure, some fundamental characteristics of the modeling of reactors will be shown by a simplified version of the system of Equations (131), (204), (240 b), and (265 b). The first simplification is the assumption of constant density, isothermal conditions, chemical reactions that do not involve volume changes, and frictionless flow.

5.3.3.1. Isothermal Reactors with Frictionless Flow, Constant Density, and Reactions Without Volume Changes

Under the above conditions it is easy to show that the equation system

$$\varrho\frac{\partial m_i}{\partial t}+\varrho u\frac{\partial m_i}{\partial x}=\frac{\partial}{\partial x}\left[\varrho D_i\frac{\partial m_i}{\partial x}\right]+\varrho S_i \quad i=1,\ldots,N-1 \tag{305}$$

is adequate for the definition of the problem provided that pressure effects and physical forces are neglected and the assumption is made that $v=w=0$ (one-dimensional formulation).

If only two chemical species A and B are considered which react in a first-order reaction $A+B\rightarrow$ products, Equation (305) for the steady-state case and with $S_A=-k\,m_A$ is

$$\varrho u\frac{dm}{dx}-\varrho D\frac{d^2m}{dx^2}+\varrho km=0 \tag{306}$$

In Equation (306) $m=m_A$, and thus also $S_A=S=-k\,m$. Analytical solutions can be found for this equation for various boundary conditions. If the equation is, for example, normalized after dividing by the density ϱ in the form $\eta=x/L$, then

$$\frac{d^2m}{d\eta^2}-Pe\frac{dm}{d\eta}-Da_I Pem=0$$

Here $Pe=u\,L/D$, and Da_I is the Damköhler number of the first kind, $Da_I=k\,\tau$, where $\tau=L/u$ is the mean hydrodynamic residence time. For $Da_I<Pe$, the characteristic equation of the above differential equation has two real roots

$$\lambda_{1,2}=\frac{1}{2}Pe\pm\sqrt{\frac{Pe^2}{4}+Da_I Pe}$$

The general solution is thus

$$m=C_1e^{\lambda_1 x}+C_2e^{\lambda_2 x}$$

where C_1 and C_2 are found from the boundary conditions $m=m_0$ for $x=0$ and $dm/dx=0$ for $x=L$ as

$$C_1=-m_0\frac{\lambda_2e^{\lambda_2 L}}{\lambda_1e^{\lambda_1 L}-\lambda_2e^{\lambda_2 L}} \quad \text{and}$$

$$C_2=m_0-C_1=m_0\left[1+\frac{\lambda_2e^{\lambda_2 L}}{\lambda_1e^{\lambda_1 L}-\lambda_2e^{\lambda_2 L}}\right]$$

Here however the numerical procedure analogous to Section 5.1.2.2 is discussed in order to compare various concepts of mathematical modeling.

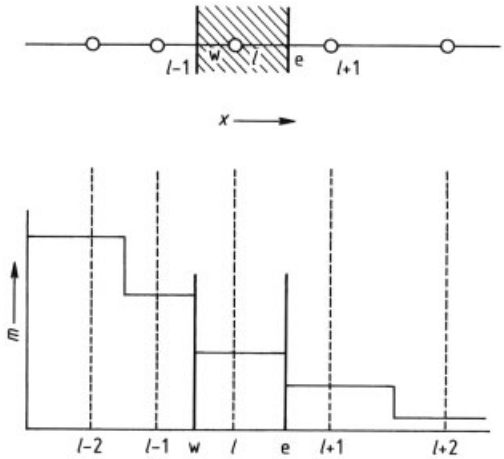

Figure 40. "Up-wind" difference scheme for a second-order differential equation
For explanation of symbols see Figure 26.

To convert the differential operators into differences, Equation (306) is subjected to the same manipulations as Equation (226) and brought into discrete form with an "up-wind" difference scheme. For a one-dimensional problem a line along the x-direction of the grid in Figure 26 is adequate (Fig. 40). In Figure 40 the profile for m is also shown. When the operations carried out in Section 5.1.2.2 are applied to Equation (306) with the grid given in Figure 40, then the system of equations

$$-a_{l-1}m_{l-1}+b_l m_l-a_{l+1}m_{l+1}-\varrho S\Delta V_l=0 \tag{307}$$

is obtained. The coefficients of this tridiagonal system of equations, with the notation from Figure 40 are given by

$$a_{l-1}=\left[+(\varrho u)_w+\varrho D\frac{1}{x_l-x_{l-1}}\right]F_w \tag{308a}$$

$$a_{l+1}=\left[+\varrho D\frac{1}{x_{l+1}-x_l}\right]F_e \tag{308b}$$

$$b_l=a_{l-1}+a_{l+1} \tag{308c}$$

Under the conditions given above $u>0$. Assuming that $(u\,F)=\Delta V_l/\tau_l$ where τ_l is the hydrodynamic residence time over a distance l, and Pe_l is a local Péclet number

$Pe_l = u\,(x_l - x_{l-1})/D$, the coefficients can be converted into

$$a_{l-1} = \frac{\varrho}{\tau_l}\Delta V_l + \frac{\varrho}{\tau_l Pe_l}\Delta V_l \tag{308d}$$

$$a_{l+1} = \frac{\varrho}{\tau_l Pe_{l+1}}\Delta V_l \tag{308e}$$

$$b_l = \frac{\varrho}{\tau_l}\Delta V_l + \frac{\varrho}{\tau_l Pe_l}\Delta V_l + \frac{\varrho}{\tau_l Pe_{l+1}}\Delta V_l \tag{308f}$$

From Equations (307) and (308 d – f) the following tridiagonal equation system

$$\frac{\varrho}{\tau_l}\left[1+\frac{1}{Pe_l}\right] m_{l-1} - \frac{\varrho}{\tau_l}\left[1+\frac{1}{Pe_l}+\frac{1}{Pe_{l+1}}\right] m_l + \frac{\varrho}{\tau_l}\frac{1}{Pe_{l+1}} m_{l+1} + \varrho S\,(m_l) = 0 \tag{309}$$

is obtained. If the case of convectively dominated flows is considered where $Pe \to \infty$, Equation (309) becomes

$$\frac{\varrho}{\tau_l}\,(m_{l-1} - m_l) + \varrho S\,(m_l) = 0 \tag{310}$$

This is the mass balance for a cascade of ideally mixed reactors. This analysis again shows that reactors with very large Bodenstein or Péclet numbers (ideally unmixed reactors) can be represented by a cascade of ideally mixed reactors (see Fig. 41 and Section 4.3.1.1).

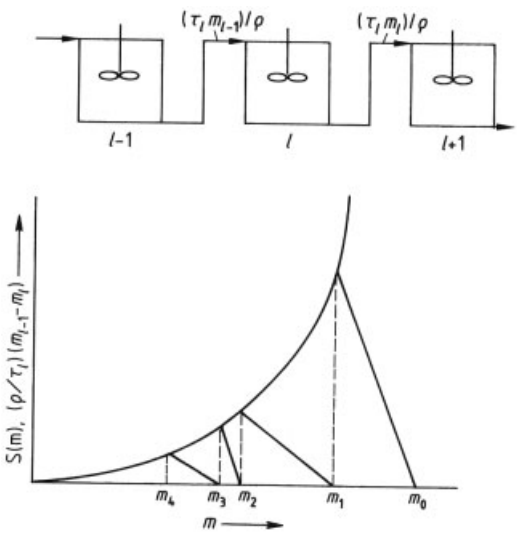

Figure 41. Cascade of ideally mixed reactors equivalent to an ideal tubular reactor
τ = residence time; ϱ = density; l = number of reactor.

Numerically this physical model finds its equivalent in the discrete form of differential Equation (306) using an "up-wind" difference scheme for the convectively dominated case. Equation (310) can easily be solved either by elimination or graphically by plotting both terms of the equation against m_l. The graphical solution is shown in Figure 41 for a nonequidistant grid, i.e., for ideally mixed reactors with different hydrodynamic residence times. It is important that there is no feedback from elements further downstream.

If the flow becomes predominantly diffusive, then Pe is very small and Equation (309) is written as

$$\frac{\varrho}{\tau_l}\frac{1}{Pe_l} m_{l-1} - \frac{\varrho}{\tau_l}\left[\frac{1}{Pe_l}+\frac{1}{Pe_{l+1}}\right] m_l + \frac{\varrho}{\tau_l}\frac{1}{Pe_{l+1}} m_{l+1} + \varrho S\,(m_l) = 0 \tag{311}$$

Equation (311) has the same structure as Equation (310) but the feed for the ideally mixed reactor with the number l consists of a flow with the composition m_{l-1} and a flow with the composition m_{l+1} (Fig. 42). There is therefore a feedback from the elements further downstream, and the volumetric mass flows are increased by the factor $1/Pe_l$ etc., as compared with the convective mass flow $\varrho\, m_l/\tau_l$. Algorithms for the solution of the tridiagonal Equation (311) on the basis of LU decomposition of the system matrix can be found in [29–32].

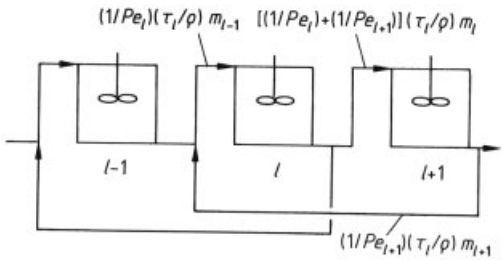

Figure 42. Cascade of ideally mixed reactors with feedback, equivalent to a tubular reactor with strong back-mixing
Pe = Péclet number. For explanation of other symbols see Figure 41.

The general case where $1/Pe_l = O\,(1)$ has already been formulated with Equation (309).

Feedback dependent on the local Péclet number exists from the element $l+1$ and the volumetric mass flows are increased as against the convective flows by the factor $[1+(1/Pe_l)]$ etc. (Fig. 43). For solution, the modeling parameter Pe is required, which contains the molecular diffusion coefficients or experimentally determined effective mixing coefficients (see page 43).

In the foregoing analysis real reactors were described by numerical solution of differential Equation (306). This is equivalent to a representation by elements of ideally mixed reactors. Alternative structures of matrices of elements of ideal reactors can be derived by physical analysis of complex reactors [4], [11], [12], [24], [111], [112]. An example which is of some importance in chemical engineering will be described here.

Example of Combinations of Elements of Ideal Reactors: the Loop Reactor. The layout of an ideal plug flow reactor with recirculation (loop reactor) is shown in Figure 44, part of the product flow is added to the feed. For analysis the same assumptions apply as in Section 5.3.3.1.

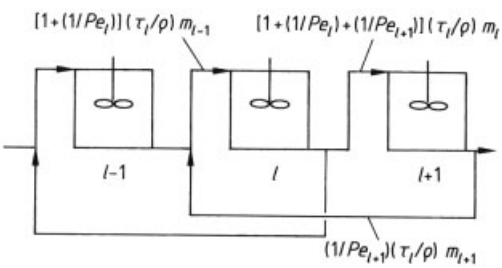

Figure 43. Cascade of ideally mixed reactors with feedback, equivalent to a tubular reactor with back-mixing, general case
Pe = Péclet number. For explanation of symbols see Figure 41.

With the definitions given in Figure 44, the mass fraction m_e at the outlet of the reactor can be obtained from Equation (310)

$$m_e = m_0' \left[\frac{1}{(1+k\tau_l)^n} \right] \tag{312}$$

For the sake of simplicity the ideal plug flow reactor in n equidistant sections so that $\tau_l = \tau = \tau_{tot}/n$. Thus from Equation (312)

$$m_e = m_0' \left[\frac{1}{\left(1 + k\tau_{tot} + \frac{n-1}{n2!}(k\tau_{tot})^2 + \frac{(n-1)(n-2)}{n^2 3!}(k\tau_{tot})^3 + \ldots\right)} \right] \tag{313}$$

For $n \gg 1$ and $k\,\tau_{tot} \ll 1$

$$m_e \approx m_0' \left[\frac{1}{\left(1 + k\tau_{tot} + \frac{1}{2!}(k\tau_{tot})^2 + \frac{1}{3!}(k\tau_{tot})^3 + \ldots\right)} \right] = m_0' e^{-k\tau_{tot}} \tag{314}$$

where m_0' and τ_{tot} are unknown and depend on the magnitude of the recirculated volume q_r.

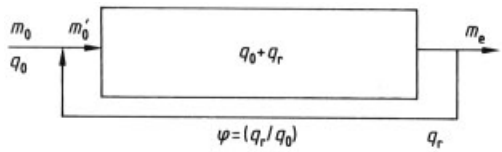

Figure 44. Schematic representation of a loop reactor
q = flow rate; m = mass fraction
The subscripts 0, r, and e denote inlet, loop (recirculated), and final, respectively; the superscript (′) denotes conditions after mixing of inlet and backflow.

The term m_0' results from the combined mass fractions of the flows entering the plug flow reactor and τ_{tot} from the volumes of the flow reactor and the total volumetric flow. If $\varphi = q_r/q_0$ (the recirculation ratio), then

$$m_0' = m_0 \frac{1}{1+\varphi} + m_e \frac{\varphi}{1+\varphi} \tag{315a}$$

and

$$\tau_{tot} = \tau_0 \frac{1}{1+\varphi} \tag{315b}$$

where τ_0 is the hydrodynamic residence time without recirculation. Substitution into Equation (314) gives

$$m_e = m_0 \frac{e^{-k\tau_0/(1+\varphi)}}{1+\varphi\left(1 + e^{-k\tau_0/(1+\varphi)}\right)} \tag{316}$$

For $\varphi \to 0$ the loop reactor takes on the character of the ideal plug flow reactor and $m_e = m_0\, e^{-k\tau_0}$. For $k\,\tau_0 = Da_I = 0.1$, $m_e/m_0 \approx 90\,\%$. For $\varphi \to \infty$, $e^{-k\tau_0/\varphi} = 1 - (k\,\tau_0/\varphi)$ and $m_e = m_0/(1 + k\,\tau_0)$. The loop reactor is similar to an ideally stirred reactor with correspondingly lower conversions.

The adjustable recirculation ratio of the loop reactor is fully exploited in important engineering applications. When measuring kinetic data in laboratory reactors the lowest possible conversion is required in the reactor so that the measured reaction rates can be assigned to specific temperatures and concentrations. Thus $Da_I \ll 1$. On the other hand the concentration differences measured for evaluating the reaction rates should be as large as possible to minimize statistical errors. As may be calculated from Equations (314) and (316), for $Da_I = 0.1$ and $\varphi = 10$ the conversion in the plug flow reactor is $1 - (m_e/m'_0) \approx 1\,\%$. The measured difference in the mass fractions is, however, $m_e/m_0 \approx 90\,\%$.

For reactions with positive reaction order relative to the reactants, the reaction rate falls off as the concentration decreases. Recirculation of the product lowers the concentration in the feed due to dilution. There is however a class of reactions in which the reaction rate first increases with decreasing concentration and then decreases after passing through a maximum (e.g., in heterogeneous catalysis, enzymatic catalysis, and autocatalysis). Recirculation has a positive effect on conversion in these cases.

5.3.3.1.1. Stability of Isothermal Reactors

Steady-State Cases. The importance of the stability of chemical reactors for safety and economic reasons is obvious. The description of stability analysis given here refers to the description of processes in physical space, location $\boldsymbol{x} = (x_1, x_2, x_3)$ and time t. Stability analysis and theories often refer to the description of chemical processes in the form of transfer functions in the space of Laplace-transformed variables or to the frequency behavior. Detailed discussion of these relationships can be found in [3], [9], [113–116].

To illustrate stability analysis in isothermal, steady-state reactors one element of the ideal plug flow reactor will be used (Fig. 40).

This can be regarded as an ideally mixed reactor where the essentials of stability analysis can be shown. Three types of chemical reactions will be considered:

1) Normal reactions with declining reaction rate
   $A + B \to$ products
   where $dm_A/dt = -\,k_2 m_A\, m_B$ and $m_B = m_{B_0} - (m_{A_0} - m_A)$
2) Autocatalysis
   $A + 2\,B \to 3\,B$
   where $dm_A/dt = -\,k_3\, m_A\, m_B^2$ and $m_B = m_{B_0} + (m_{A_0} - m_A)$
3) Self poisoning
   $A \to B$
   where $dm_A/dt = -\,k_1\, m_A/(1 + K\, m_A)^2$ (Section 3.2.1.1, Eq. 53).

Along with the assumptions from Section 5.3.3.1, equal molar masses of A and B are also assumed. Using the abbreviations $\alpha = m_A/m_{A_0}$ and $\beta_0 = m_{B_0}/m_{A_0}$ and Equation (310), the following equations are obtained for the three cases

$$\frac{1-\alpha}{\tau k_2 m_{A_0}} = \alpha^2 + (\beta_0 - 1)\,\alpha \tag{317a}$$

$$\frac{1-\alpha}{\tau k_3 m_{A_0}^2} = \alpha\,[(\beta_0 + 1) - \alpha]^2 \tag{317b}$$

$$\frac{1-\alpha}{\tau k_1} = \alpha \left[\frac{1}{1 + \alpha m_{A_0} K\,(2 + \alpha m_{A_0} K)}\right] \tag{317c}$$

where $1 - \alpha$ is the conversion. The term $\tau\, k_2 m_{A_0}$, $\tau\, k_3 m_{A_0}^2$, or $\tau\, k_1$ can be regarded as the ratio of the hydrodynamic residence time to a characteristic chemical time and is denoted as the Damköhler number of the first kind Da_I. Equations (317) can conveniently be solved graphically by plotting both sides against $1 - \alpha$. Figures 45, 46, 47 show such plots for the three cases above. For the normal case (Fig. 45 A), decreasing reaction rate with decreasing reactant concentration, an unambiguous steady-state solution exists for each Damköhler number. With increasing Damköhler number (i.e., increasing hydrodynamic residence time), the steady-state solutions give higher conversions. All steady-state solutions, which are plotted against the Damköhler number in Figure 45 B, are stable. If a small disturbance occurs to the left or right on the conversion axis in the steady-state solution, opposing changes are produced in the right-hand side (RHS) or left-hand side (LHS) respectively of Equation (317 a).

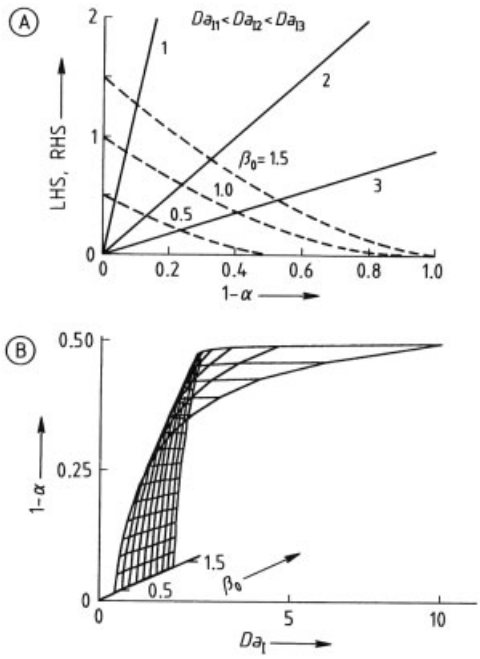

Figure 45. Mono-steady-state conversion points for "normal" chemical reactions (A) and bifurcation diagram (B) for isothermal conditions in a well-stirred reactor
LHS = left-hand side (Eq. 317 a); RHS = right-hand side (Eq. 317 a); $1-\alpha$ = conversion; Da_{I} = Damköhler number of first kind; $\beta_0 = m_{\mathrm{B}_0}/m_{\mathrm{A}_0}$

The "convective term" on the LHS decreases (increases), whilst the "reaction term" on the RHS increases (decreases). A lesser (greater) availability of A due to convection therefore counterbalances a greater (lesser) consumption of A due to reaction, so that the conversion is shifted to the right or to the left. The general condition for stability can be given in the terminology of Equation (317) as

$$\frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,[\mathrm{LHS}] > \frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,[\mathrm{RHS}] \quad \text{at} \quad \mathrm{LHS} = \mathrm{RHS} \tag{318a}$$

For Equation (317 a), $\frac{\mathrm{d}}{\mathrm{d}(1-\alpha)}\,[\mathrm{LHS}] = 1/Da_{\mathrm{I}}$ and $\frac{\mathrm{d}}{\mathrm{d}(1-\alpha)}\,[\mathrm{RHS}] = 2\,(1-\alpha) - \beta - 1$; this satisfies the condition for stability for all possible values of $1-\alpha$ based on the inlet condition β_0.

A completely different picture results for autocatalysis (Fig. 46 A) or self poisoning (Fig. 47 A). In these cases multiple steady-state solutions occur for a particular range of Damköhler numbers. Some of the solutions do not comply with the conditions for stability (Eq. 318). For Equation (317 b) $\frac{\mathrm{d}}{\mathrm{d}(1-\alpha)}\,[\mathrm{RHS}] = [\beta_0 + (1-\alpha)]\,[2 - 3\,(1-\alpha) - \beta_0]$, so that the stability criterion is not met for small values of $(1-\alpha)$.

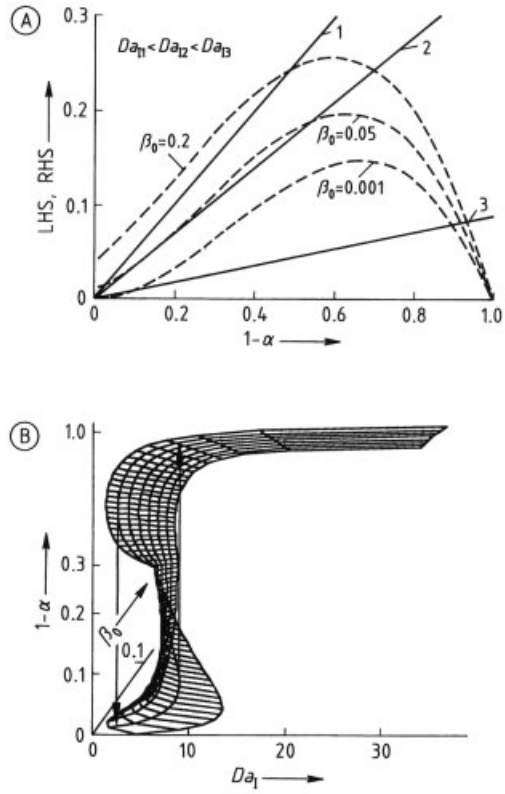

Figure 46. Multiple steady-state conversion points for autocatalytic chemical reactions (A) and bifurcation diagram (B) for isothermal conditions in a well-stirred reactor
LHS = left-hand side (Eq. 317 b); RHS = right-hand side (Eq. 317 b)
For explanation of other symbols see Figure 45.

In this range a small disturbance of $(1-\alpha)$ to the right leads to a larger increase in the "reaction term" than in the availability due to convection; conversion therefore increases. The reaction system thus always heads for the nearest adjacent stable operating point. In the bifurcation diagrams shown in Figures 46 B and 47 B, these unstable steady-state solutions lie on an S-shaped branch. With a continuous increase in the Damköhler number (residence time) however the system cannot pass through this S-shaped branch, but jumps from a condition with low conversion to one with high conversion. The bifurcation diagrams show that in both cases with

multiple steady-state solutions the Damköhler number for "igniting" the reaction is higher than that for "extinguishing" it. Thus hysteresis occurs. The jumping points are given by the condition for the coalescence of an unstable with a stable steady-state solution:

$$\frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,[\mathrm{LHS}] = \frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,[\mathrm{RHS}]$$
$$\text{at}\quad \mathrm{LHS} = \mathrm{RHS} \tag{318b}$$

In the case of autocatalysis this condition occurs for

$$\frac{8}{Da_{\mathrm{I}}} = 1 + 20\beta_0 - 8\beta_0^2 \pm (1-4\beta_0)\sqrt{(1-8\beta_0)}$$

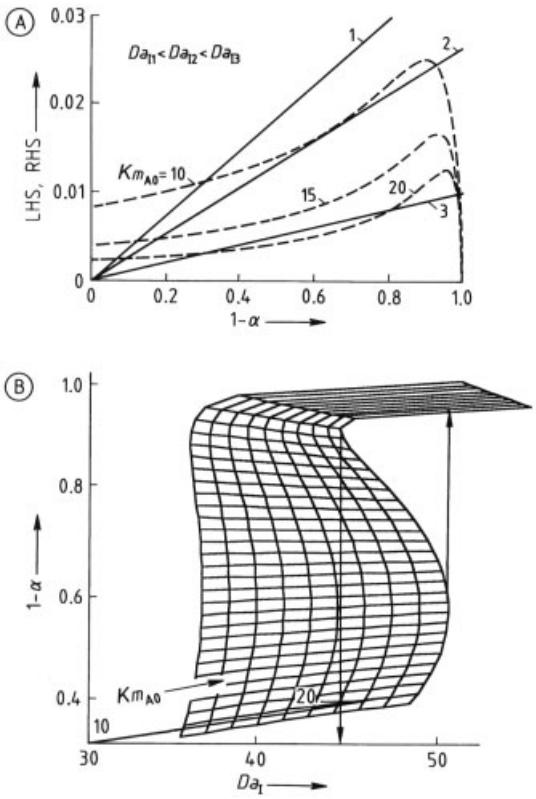

Figure 47. Multiple steady-state conversion points for chemical reactions with self poisoning (A) and bifurcation diagram (B) for isothermal conditions in a well-stirred reactor
K = adsorption coefficient; m_{A_0} = inlet mass fraction of component A; LHS = left-hand side (Eq. 317 c), RHS = right-hand side (Eq. 317 c)
For explanation of other symbols see Figure 45.

The positions of the two jumping points move closer together and, with an increase in the inlet concentration of the catalyst β_0, the S-shaped curve unfolds. Increasing the concentration of the catalyst at the reactor inlet results in stable reactor behavior, shown by the disappearance of multiple steady-state solutions.

An analogous picture results for self poisoning. The appropriate conditions can easily be calculated from Equation (317 c) and the stability condition. Here the parameter controlling the nonambiguity of the solutions is Km_{A_0} which can be regarded as a measure of the poisoning. A decrease in this quantity (lower pressure or lower inlet concentration of A in heterogeneous catalytic reactions) leads to a range of conversions that exhibits unambiguous, stable, steady-state solutions (compare Figs. 45 B, 46 B, 47 B).

Nonsteady-State Cases. The stability analysis of steady-state, isothermal, ideally mixed reactors has shown that the phenomenon of multiple steady states with hysteresis always occurs when there is feedback. A feedback system will now be considered for the stability analysis of nonsteady-state reactions. The chemical reaction is formulated as

$$\mathrm{P} \xrightarrow{k_0} \mathrm{A} \xrightarrow{k_u} \mathrm{B} \xrightarrow{k_2} \mathrm{C}, \qquad \mathrm{A} \xrightarrow{k_1} \mathrm{B}$$

where the step $\mathrm{A} \rightarrow \mathrm{B}$ is a noncatalytic first-order reaction and a cubic (third-order) reaction catalyzed with B. The feedback mechanism in this reaction is clear: the more B is formed, the faster A is consumed. Since the supply of A from P remains limited, at high concentrations of catalyst B the reaction $\mathrm{A} \rightarrow \mathrm{B}$ breaks down until adequate A is again supplied.

Analysis will be performed on a batch reactor so that the system of Equations (305) can be written in the form

$$\frac{\mathrm{d}m_{\mathrm{P}}}{\mathrm{d}t} = -k_0 m_{\mathrm{P}} \tag{319a}$$

$$\frac{\mathrm{d}m_{\mathrm{A}}}{\mathrm{d}t} = k_0 m_{\mathrm{P}} - k_1 m_{\mathrm{A}} m_{\mathrm{B}}^2 - k_{\mathrm{u}} m_{\mathrm{A}} \tag{319b}$$

$$\frac{\mathrm{d}m_{\mathrm{B}}}{\mathrm{d}t} = k_1 m_{\mathrm{A}} m_{\mathrm{B}}^2 + k_{\mathrm{u}} m_{\mathrm{A}} - k_2 m_{\mathrm{B}} \tag{319c}$$

A condition for m_{C} results from the mass balance of the system. The quantity $(k_2/k_1)^{1/2}$ is

used to normalize the mass fraction, and the normalization of the time is achieved through $(1/k_2)$. The system of equations can then be written as

$$\frac{d\pi}{d\tau} = -\varepsilon\pi \tag{320a}$$

where $\pi = m_P(k_2/k_1)^{1/2}$, $\varepsilon = (k_0/k_2)$, and $\tau = t/k_2$

$$\frac{d\alpha}{d\tau} = \varepsilon\pi - \alpha\beta^2 - \kappa\alpha \tag{320b}$$

where $\alpha = m_A(k_2/k_1)^{1/2}$, $\beta = m_B(k_2k_1)^{1/2}$, and $\kappa = (k_u/k_2)$, and finally

$$\frac{d\beta}{d\tau} = \alpha\beta^2 + \kappa\alpha - \beta \tag{320c}$$

The initial conditions are $\pi = \pi_0$, $\alpha = \beta = 0$ for $\tau = 0$. Thus Equation (320 a) can easily be solved. The result

$$\pi = \pi_0 e^{-\varepsilon\tau} \tag{321}$$

is used in the system of Equations (320) which thus become a system of coupled first-order differential equations

$$\frac{d\alpha}{d\tau} = \mu - \alpha\beta^2 - \kappa\alpha = f(\alpha, \beta) \tag{322a}$$

$$\frac{d\beta}{d\tau} = \alpha\beta^2 + \kappa\alpha - \beta = g(\alpha, \beta) \tag{322b}$$

where the abbreviation μ denotes $\mu = \varepsilon\,\pi_0\,e^{-\varepsilon\tau} = \mu_0 e^{-\varepsilon\tau}$. For $k_0 \ll k_1$, k_u, k_2 the principle of quasi-steady state is applicable for the mass fractions α and β. Thus Equations (322 a, b) are solved without integration; approximate solutions for α and β can be obtained from the conditions

$$\frac{d\alpha}{d\tau} = \frac{d\beta}{d\tau} = 0$$

From

$$\mu - \alpha\beta^2 - \kappa\alpha = 0 \tag{323a}$$

and

$$\alpha\beta^2 + \kappa\alpha - \beta = 0 \tag{323b}$$

the quasi-steady-state mass fractions α_{qs}, β_{qs} are given by

$$\alpha_{qs} = \mu/\left(\mu^2 + \kappa\right) \text{ and } \beta_{qs} = \mu \tag{324a,b}$$

For the stability analysis α and β are regarded as being subject to small disturbances $\Delta\alpha$ and $\Delta\beta$, respectively. The time evolution of these disturbances are then considered. The perturbed quantities thus become

$$\alpha = \alpha_0 + \Delta\alpha,\ f(\alpha,\beta) = f(\alpha,\beta)_0 + \left(\frac{\partial f(\alpha,\beta)}{\partial\alpha}\right)_0 \Delta\alpha + \left(\frac{\partial f(\alpha,\beta)}{\partial\beta}\right)_0 \Delta\beta + \ldots \tag{325a}$$

and

$$\beta = \beta_0 + \Delta\beta,\ g(\alpha,\beta) = g(\alpha,\beta)_0 + \left(\frac{\partial g(\alpha,\beta)}{\partial\alpha}\right)_0 \Delta\alpha + \left(\frac{\partial g(\alpha,\beta)}{\partial\beta}\right)_0 \Delta\beta + \ldots \tag{325b}$$

The functions $f(\alpha, \beta)$ and $g(\alpha, \beta)$ in Equations (325) are developed in Taylor series at α_0 and β_0 that are truncated after the first term. This is a reasonable approximation for small disturbances. If Equation (325) is substituted into Equation (322), the time evolution of the disturbances can be expressed as

$$\frac{\partial\alpha_0}{\partial\tau} + \frac{\partial\Delta\alpha}{\partial\tau} = f(\alpha,\beta)_0 + \left(\frac{\partial f(\alpha,\beta)}{\partial\alpha}\right)_0 \Delta\alpha + \left(\frac{\partial f(\alpha,\beta)}{\partial\beta}\right)_0 \Delta\beta \tag{326a}$$

and

$$\frac{\partial\beta_0}{\partial\tau} + \frac{\partial\Delta\beta}{\partial\tau} = g(\alpha,\beta)_0 + \left(\frac{\partial g(\alpha,\beta)}{\partial\alpha}\right)_0 \Delta\alpha + \left(\frac{\partial g(\alpha,\beta)}{\partial\beta}\right)_0 \Delta\beta \tag{326b}$$

If the quasi-steady-state solutions are considered, then Equation (326) becomes the system of coupled ordinary differential equations

$$\begin{pmatrix} \frac{d\Delta\alpha}{d\tau} \\ \frac{d\Delta\beta}{d\tau} \end{pmatrix} - \boldsymbol{J}_{qs} \begin{pmatrix} \Delta\alpha \\ \Delta\beta \end{pmatrix} = 0 \tag{327a}$$

where $\boldsymbol{J}_{qs}$ is the Jacobi matrix of Equation (323)

$$\boldsymbol{J}_{qs} = \begin{pmatrix} \frac{\partial f(\alpha,\beta)}{\partial\alpha} & \frac{\partial f(\alpha,\beta)}{\partial\beta} \\ \frac{\partial g(\alpha,\beta)}{\partial\alpha} & \frac{\partial g(\alpha,\beta)}{\partial\beta} \end{pmatrix}_{qs} \tag{327b}$$

Equation (327) can easily be solved analytically after being transformed into a second-order differential equation (see Section 5.3.1.3). The general solution is

$$\Delta\alpha(\tau) = C_1 e^{\lambda_1\tau} + C_2 e^{\lambda_2\tau} \tag{328a}$$

$$\Delta\beta(\tau) = C_3 e^{\lambda_1 \tau} + C_4 e^{\lambda_2 \tau} \tag{328b}$$

where the eigenvalues λ_1 and λ_2 are determined from the characteristic equation

$$\lambda^2 - \mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})\lambda + \mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = 0 \tag{329}$$

This expression has already been used in Sections 5.3.1.3, 5.3.2, and 5.3.3.1. The trace of the Jacobi matrix is the sum of the diagonal elements

$$\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) = \left(\frac{\partial f(\alpha,\beta)}{\partial \alpha}\right)_{\mathrm{qs}} + \left(\frac{\partial g(\alpha,\beta)}{\partial \beta}\right)_{\mathrm{qs}}$$

and its determinant is

$$\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = \left(\frac{\partial f(\alpha,\beta)}{\partial \alpha}\right)_{\mathrm{qs}} \left(\frac{\partial g(\alpha,\beta)}{\partial \beta}\right)_{\mathrm{qs}} - \left(\frac{\partial g(\alpha,\beta)}{\partial \alpha}\right)_{\mathrm{qs}} \left(\frac{\partial f(\alpha,\beta)}{\partial \beta}\right)_{\mathrm{qs}}$$

so that the solution of the characteristic equation is

$$\lambda_{1,2} = \frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) \pm \left\{\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})^2 - 4\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})\right\}^{1/2}\right] \tag{330}$$

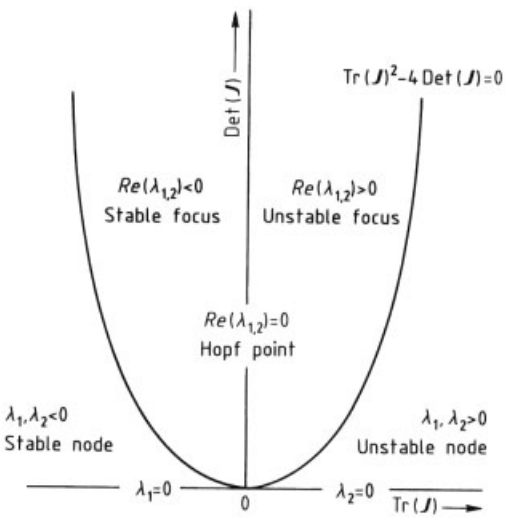

Figure 48. Stability diagram and nature of the steady-state solutions for the example in Section 5.3.3.1
$\lambda_{1,2}$ = eigenvalues of Equations (328).

The time evolution of the disturbances $\Delta\alpha$ and $\Delta\beta$ is affected by the eigenvalues of $\lambda_{1,2}$, which depend on the values of $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})$ and $\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})$. The different domains of the solutions of Equations (327) are shown in Figure 48 in a stability diagram.

To demonstrate some of the possible cases the simplifying assumption is made that the first-order noncatalytic reaction (A → B) proceeds very slowly, i.e., $\kappa \ll 1$. Equations (323) and (324) can then be further simplified

$$\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) = (1-\mu^2) \tag{331a}$$

$$\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = \mu^2 \tag{331b}$$

$$\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})^2 - 4\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = \mu^4 - 6\mu^2 + 1 \tag{331c}$$

As μ is a function of time, it is to be expected that during the course of the reaction with the depletion of the reservoir P, the discriminants, Equation (331 c), as well as $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})$ and $\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})$ will change their values and signs, so that $\lambda_{1,2}$ take all possible combinations

1) Initial condition, $\mu > 1+\sqrt{2}$
   Initially only P is present so μ is very large. Then $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})$ is negative, $\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})$ is positive, and the discriminant is also positive. Thus $\lambda_{1,2}$ become both real and negative. Small disturbances $\Delta\alpha$ and $\Delta\beta$ decay monotonically, and the quasi-steady-state mass fractions represent a stable node of the system. The terms node (and focus) are derived from the phase diagrams of $\Delta\alpha$ and $\Delta\beta$, see [114]. In Figure 49 this range is shown in the plot of α_{qs} and β_{qs} against μ.
2) Progressive reaction, $1 < \mu < 1+\sqrt{2}$
   In this range of conversion of P, $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})$ is negative and $\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})$ positive. The discriminant has now changed its sign however. Due to this $\lambda_{1,2}$ become conjugate complex with a negative real part. In this case the disturbances $\Delta\alpha$ and $\Delta\beta$ behave as damped oscillations and the quasi-steady-state solutions represent a stable focus of the system.
3) Further progress of the reaction, $\sqrt{2} - 1 < \mu < 1$
   In this concentration range of P, $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})$ is positive and the discriminant negative. Here too $\lambda_{1,2}$ are conjugate complex but with a positive real part. The disturbances $\Delta\alpha$ and $\Delta\beta$ increase in the form of divergent oscillations. The quasi-steady-state solutions are an unstable focus of the system.

4) Depletion of P, $0<\mu<\sqrt{2}-1$
   Towards the end of the reaction the trace of the Jacobi matrix is always positive; the discriminant has changed its sign and is also positive again. Now $\lambda_{1,2}$ are both real and positive so that the disturbances grow exponentially. The state of the system is described as an unstable node.

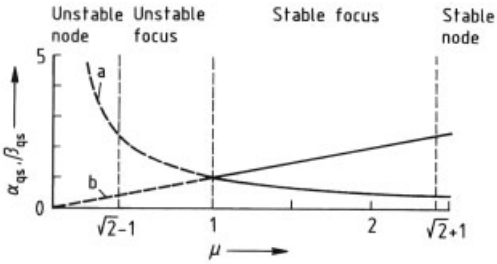

Figure 49. Quasi-steady-state solutions (a) $\alpha_{qs}(\mu)$ and (b) $\beta_{qs}(\mu)$ for an isothermal nonsteady-state system with cubic autocatalysis
For further explanation of symbols see text.

Between the conditions (2) and (3), a state exists at which $\mu=1$. At this point $\mathrm{Tr}(\boldsymbol{J}_{qs})=0$ and $\mathrm{Det}(\boldsymbol{J}_{qs})>0$ so that $\lambda_{1,2}$ become imaginary. Here the disturbances $\Delta\alpha$ and $\Delta\beta$ behave as undamped sinusoidal oscillations and the system forms stable oscillations around the quasi-steady-state solutions. This point is designated the Hopf bifurcation point.

For the simplified system without the noncatalytic reaction, the trace of the Jacobi matrix is given by Equation (331 a). Thus the condition for the change of sign, i.e., $\mathrm{Tr}(\boldsymbol{J}_{qs})=0$ and $d\mathrm{Tr}(\boldsymbol{J}_{qs})/d\mu \neq 0$, only occurs when $\mu=1$. If the noncatalytic reaction is also considered, then

$$\mathrm{Tr}(\boldsymbol{J}_{qs}) = \frac{\mu^4-(1-2\kappa)\mu^2+\kappa(1+\kappa)}{\mu^2+\kappa}$$

(see also Eq. 323). Two Hopf bifurcation points are then obtained from

$$(\mu_{1,2Bf})^2 = \frac{1}{2}\left[(1-2\kappa)\pm(1-8\kappa)^{1/2}\right]$$

For $\kappa \ll 1$ the upper equation may be expanded into a series and the two bifurcation points are

$$\mu_{1Bf} = 1-\frac{3}{2}\kappa+\ldots \quad \text{and}$$

$$\mu_{2Bf} = \kappa^{1/2}(1+2\kappa)+\ldots$$

Since $\mathrm{Det}(\boldsymbol{J}_{qs})$ is always positive, then with consumption of P the system passes through two Hopf bifurcation points between which solutions for $\lambda_{1,2}$ occur with a predominantly imaginary part. In this range instabilities can develop in the form of stable oscillations, the amplitude and frequency of which alter with the decreasing concentration of P [113].

In chemical engineering practice instabilities can only be avoided in systems with the properties quoted above for the range $\mu>1$. Details of the Hopf bifurcation analysis and the detailed calculation of the oscillation properties (frequency, amplitude, and stability) are given in [113–116], see also → Mathematics in Chemical Engineering, Chap. 6.7.

5.3.3.1.2. Sensitivity Analysis

In Section 3.1.1.3 sensitivity analysis was described as the change in the dependent variables of a model with a change of its modeling parameters. If the dependent variables $\boldsymbol{\Phi}$ are developed as a Taylor series at a specific solution $\boldsymbol{\Phi}_0$, the change is given by

$$\boldsymbol{\Phi}=\boldsymbol{\Phi}_0+\left(\frac{\partial\boldsymbol{\Phi}}{\partial\boldsymbol{\beta}}\right)_0\Delta\boldsymbol{\beta}+\left(\frac{\partial^2\boldsymbol{\Phi}}{\partial\boldsymbol{\beta}^2}\right)_0\frac{\Delta\boldsymbol{\beta}^2}{2!}+\ldots \qquad (332)$$

where $\boldsymbol{\beta}$ is the vector of the parameters.

The simplest measure of the system's sensitivity are the gradients $(\partial\boldsymbol{\Phi}/\partial\boldsymbol{\beta})$ with which the change in the dependent variables is calculated with the aid of a Taylor series truncated after the first term. The gradients $(\partial\boldsymbol{\Phi}/\partial\boldsymbol{\beta})$ are denoted as a first-order sensitivity coefficients, the curve and the higher derivatives of Equation (332) as sensitivity coefficients of the second and correspondingly higher orders.

If the solution $\boldsymbol{\Phi}$ is available in explicit form (e.g., Eq. 323), calculation of the sensitivity coefficients is simple. For very many models of chemical engineering processes the differential equations which constitute the model cannot, however, be solved analytically or with the help of simple assumptions. The sensitivity coefficients must then be determined from the differential equations or from the numerical solutions

[117], [118]. The latter method will now be described.

Numerical solutions provide the dependent variables not in the form of continuous functions but in discrete form (discrete values), calculation of the gradients is therefore not trivial. The numerical solution of a differential equation is (e.g., for the example in Section 5.3.3.1) given implicitly through Equation (309) in the form of a system of equations.

$$F(\boldsymbol{\Phi}) = 0 \tag{333}$$

The first-order sensitivity coefficients are obtained from Equation (333) by the differentiation rule

$$\frac{\partial F(\boldsymbol{\Phi})}{\partial(\boldsymbol{\Phi})}\frac{\partial(\boldsymbol{\Phi})}{\partial\boldsymbol{\beta}} + \frac{\partial F(\boldsymbol{\Phi})}{\partial\boldsymbol{\beta}} = 0 \tag{334}$$

Thus the gradients are given by

$$\frac{\partial(\boldsymbol{\Phi})}{\partial\boldsymbol{\beta}} = -\boldsymbol{J}^{-1}\frac{\partial F(\boldsymbol{\Phi})}{\partial\boldsymbol{\beta}} \tag{335}$$

In the example from Section 5.3.3.1 the corresponding relations are easily verified: the Jacobi matrix is given by the coefficients a_{l-1}, $-[b_l - \varrho(\partial S(m_l)/\partial m_l)]$, and a_{l+1} and is identical to the coefficient matrix for the equation system (Eq. 309). For $\beta = k$ the derivatives with respect to the parameters are simply $\varrho(\partial S(m_l)/\partial k)$. For a first-order reaction $\varrho(\partial S(m_l)/\partial k) = -\varrho m_l$. If we simplify in Equation (309) $Pe_l = Pe_{l+1} = Pe$ then

$$\frac{\partial F}{\partial Pe} = -\frac{\varrho}{\tau_l Pe^2}(m_{l-1} - 2m_l + m_{l+1})$$

so for the sensitivity coefficients the equation system (336) has to be solved. Equation system (336) for the gradients $\partial(\boldsymbol{\Phi})/\partial\beta$ has the same structure as the original equation system (Eq. 309) and can consequently be solved by similar methods.

$$\begin{pmatrix} \frac{\partial m_1}{\partial k} & \frac{\partial m_1}{\partial Pe} \\ \vdots & \vdots \\ \frac{\partial m_n}{\partial k} & \frac{\partial m_n}{\partial Pe} \end{pmatrix} = -\begin{pmatrix} -(b_1+\varrho k) & a_2 & & & \\ & a_1 & -(b_2+\varrho k) & a_3 & \\ & & & \ddots & \\ & & & a_{n-1} & -(b_n+\varrho k) \end{pmatrix}^{-1} \cdot \begin{pmatrix} -\varrho m_1 & -\frac{\varrho}{\tau_1 Pe^2}(m_0 - 2m_1 + m_2) \\ \vdots & \vdots \\ -\varrho m_n & -\frac{\varrho}{\tau_n Pe^2}(m_{n-1} - 2m_n) \end{pmatrix} \tag{336}$$

For the calculation and discussion of the sensitivity coefficients of higher order, reference should be made to the literature [117], [118]. An example of calculating first-order sensitivity coefficients for a complex model follows in Section 5.3.3.2.

5.3.3.2. Nonisothermal Reactors

In nonisothermal reactors the density does not remain constant. The couplings in the system of Equations (131), (204), (240 b), and (265 b) therefore cannot be resolved and the equations must be solved simultaneously.

5.3.3.2.1. Heterogeneous Catalytic Reactions

Example: Heterogeneous Catalytic Dehydration of Ethylbenzene to Styrene. The heterogeneous catalytic dehydration of ethylbenzene to styrene is outlined in Figure 50. Ethylbenzene is preheated and mixed with superheated steam. The mixture is passed through a cylindrical reactor packed with an iron oxide catalyst. Dehydration is endothermic. The steam, used as a heat carrier, causes some secondary reactions. A suitable reaction scheme is [11], [119], [120]:

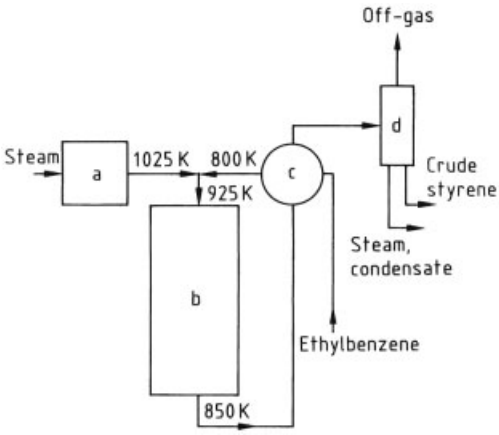

Figure 50. Schematic of a pilot plant for catalytic dehydration of ethylbenzene
a) Superheater; b) Catalytic reactor; c) Heat exchanger; d) Separator.

$$C_6H_5{-}C_2H_5(\mathbf{1}) \rightarrow C_6H_5{-}C_2H_3(\mathbf{2}) + H_2(\mathbf{3}) \tag{i}$$

$$C_6H_5{-}C_2H_5 \rightarrow C_6H_6(\mathbf{4}) + C_2H_4(\mathbf{7}) \tag{ii}$$

$$C_6H_5{-}C_2H_5 + H_2 \rightarrow C_6H_5{-}CH_3(\mathbf{5}) + CH_4(\mathbf{6}) \tag{iii}$$

$$1/2C_2H_4 + H_2O(\mathbf{10}) \rightarrow CO(\mathbf{8}) + 2H_2 \tag{iv}$$

$$CH_4 + H_2O \rightarrow CO + 3H_2 \tag{v}$$

$$CO + H_2O \rightarrow CO_2(\mathbf{9}) + H_2 \tag{vi}$$

The formulation of this set of reactions is necessary because they all have an appreciable reaction enthalpy so that conversion into the various byproducts also affects the temperature profile. Furthermore the reaction rate of reaction (i) is given by $r_{(i)} = k_1(m_1 - m_2 m_3/K)$ [8], [108] which indicates poisoning due to the products styrene and hydrogen, and thus the mass fraction m_1 is coupled with m_2, m_3.

The appropriate formulation of Equations (131), (204), (240 b), and (265 b) is in cylindrical coordinates for this axially symmetrical problem. The continuity equation in this form is

$$\frac{\partial \varrho}{\partial t} + \varrho \frac{\partial}{\partial x}(u) + \varrho \frac{1}{r}\frac{\partial}{\partial r}(rv) = 0 \tag{337a}$$

If we use the normalized variables $u* = u/u_0$, $v* = v/u_0$, $\varrho* = \varrho/\varrho_0$, $t* = t/\tau$, $x* = x/L$, and $r* = r/R$, then from Equation (337 a) it follows that

$$\frac{\varrho_0}{\tau}\frac{\partial \varrho^*}{\partial t^*} + \frac{\varrho_0 u_0}{L}\varrho^* \frac{\partial}{\partial x^*}(u^*) + \frac{\varrho_0 u_0}{R}\varrho^* \frac{1}{r^*}\frac{\partial}{\partial r^*}(r^* v^*) = 0 \tag{337b}$$

This can be transformed into

$$\frac{\partial \varrho^*}{\partial t^*} + \varrho^* \frac{\partial}{\partial x^*}(u^*) + \varrho^* \frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^* v^* \frac{R}{L}\right) = 0 \tag{337c}$$

Since $u*$, $x*$, and $r*$ are all of the order of magnitude $O(1)$, then $v*(L/R) = O(1)$ or $v* \approx (R/L)$, and thus for $(R/L) \ll 1$ the third term in Equation (337 c) can be neglected. In dimensional quantities the continuity equation for the steady-state case thus has the form

$$\varrho \frac{\mathrm{d}u}{\mathrm{d}x} = 0 \tag{337d}$$

The balance of momentum (Eq. 204) with the above result in cylindrical coordinates is

$$\varrho \frac{\partial u}{\partial t} + \varrho u \frac{\partial u}{\partial x} = \left[\frac{\partial \tau_{xx}}{\partial x} + \frac{1}{r}\frac{\partial (r\tau_{rx})}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta x}}{\partial \theta}\right] - \frac{\partial p}{\partial x} \tag{338a}$$

For flows through packed catalyst beds the viscous stress term can be expressed by the permeability κ according to Darcy's Law [8], [121]. The term in the square brackets in Equation (338 a) then becomes

$$\left[\frac{\partial \tau_{xx}}{\partial x} + \frac{1}{r}\frac{\partial (r\tau_{rx})}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta x}}{\partial \theta}\right] = \frac{\mu u}{\kappa}$$

If the same normalized variables are used as in Equation (337) together with the normalized pressure $p* = p/p_0$, Equation (338 a) becomes

$$\frac{R}{\tau u_0}\frac{\partial u^*}{\partial t^*} + \frac{R}{L}u^* \frac{\partial u^*}{\partial x^*} = -\frac{\mu R}{\kappa u_0 \varrho_0}\frac{u^*}{\varrho^*} - \frac{R p_0}{L\,\varrho_0 u_0^2}\frac{1}{\varrho^*}\frac{\partial p^*}{\partial x^*} \tag{338b}$$

Estimates of the orders of magnitude of the factors in front of the differential quotients in Equation (338 b) under pilot plant conditions ($\tau \approx 1$ s, $u_0 \approx 0.2$ m/s, $R/L \approx 5\times10^{-2}$, $\varrho_0 \approx 0.5$ kg/m^3, $p_0 \approx 1.5\times10^5$ Pa, $\mu \approx 10^{-5}$ kg m^{-1} s^{-1}, and finally $\kappa \approx 1.5\times10^{-8}$ m^2) leads to

$$\frac{R}{L} < \frac{R}{\tau u_0} < 1 < \frac{\mu R}{\kappa u_0 \varrho_0} \ll \frac{R p_0}{L\,\varrho_0 u_0^2}$$

Thus, for the given conditions, the pressure term predominates in the steady-state balance of momentum which in dimensional quantities simplifies to

$$\frac{\mathrm{d}p}{\mathrm{d}x} \approx 0 \quad \text{or} \quad p \approx \text{const.} \tag{338c}$$

Consequently the continuity equation and the momentum equation are no longer coupled with the other transport equations and pressure p and velocity u are simple to estimate.

Using the definition equation for enthalpy, the conservation equation for the enthalpy (Eq. 240 b) can be transformed into a differential equation for temperature (see Section 5.2.2.1). For a multicomponent system the conservation equations for the mass of the chemical species must be used for this. The result, also in cylindrical coordinates, is

$$\varrho c_p \frac{\partial T}{\partial t} + \varrho c_p u \frac{\partial T}{\partial x} + \lambda \left[\frac{\partial^2 T}{\partial x^2} + \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T}{\partial r} \right) \right] = \sum r_i \left(-\Delta h_{r_i} \right) + S_\mathrm{C} \tag{339a}$$

Equation (339 a) contains the assumptions introduced in Section 5.2. In addition since $p \approx$ const., S_C only includes a term for heat transfer from the gas phase to the catalyst. Due to the transformation of the conservation equation of enthalpy into a differential equation for temperature, the term $\sum r_i \left(-\Delta h_{r_i} \right)$ occurs which describes the release of heat due to chemical reactions. As in Equations (337) and (338), all quantities are normalized to the conditions at the inlet of the reactor. For Equation (339 a) this results in

$$\begin{aligned} &\frac{L}{\tau u_0} \varrho^* c_p^* \frac{\partial T^*}{\partial t^*} + \varrho^* c_p^* u^* \frac{\partial T^*}{\partial x^*} \\ &= \frac{\lambda_0}{L\, \varrho_0 c_{p0} u_0} \lambda^* \frac{\partial^2 T^*}{\partial x^{*2}} \\ &+ \frac{\lambda_0 L}{R^2 \varrho_0 c_{p0} u_0} \frac{\lambda^*}{r^*} \frac{\partial}{\partial r^*} \left(r^* \frac{\partial T^*}{\partial r^*} \right) \\ &+ \frac{L\, \alpha s}{\varepsilon \varrho_0 c_{p0} u_0} \left(T^*_\mathrm{Cat} - T^* \right) \\ &+ \frac{L}{u_0 T_0 \varrho_0 c_{p0}} \sum r_i \left(-\Delta h_{r_i} \right) \end{aligned} \tag{339b}$$

where α is the heat-transfer coefficient, s the specific surface area, ε the intergranular volume, and T_Cat the catalyst temperature. Estimation of the factors preceding the differential quotients in Equation (339 b) ($c_{p0} \approx 2.5\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$, $\lambda_0 \approx 4 \times 10^{-5}\,\mathrm{kJ\,m^{-1}\,K^{-1}\,s^{-1}}$, $\alpha \approx 4 \times 10^{-2}\,\mathrm{kW \cdot m^{-2}\,K^{-1}}$, $\varepsilon \approx 0.36$, and $s \approx 1.5 \times 10^3\,\mathrm{m^{-1}}$) results in

$$\frac{\lambda_0 L}{R^2 \varrho_0 c_{p0} u_0} < \frac{\lambda_0}{L\, \varrho_0 c_{p0} u_0} \ll \frac{L}{\tau u_0} \approx 1 \ll \frac{L\, \alpha s}{\varepsilon \varrho_0 c_{p0} u_0}$$

The fundamental mechanisms for energy transport are therefore expressed in the one-dimensional form of the temperature equation, which in steady-state formulation with dimensional quantities becomes

$$\varrho c_p u \frac{\mathrm{d}T}{\mathrm{d}x} = \frac{\alpha s}{\varepsilon} \left(T_\mathrm{Cat} - T \right) + \sum r_i \left(-\Delta h_{r_i} \right) \tag{339c}$$

The heat-transfer term in Equation (339 c) couples the temperature of the gas phase T with the catalyst temperature T_Cat. This coupling requires an equation for determining the catalyst temperature, which is derived from the heat balance for the catalyst bed:

$$\varrho_\mathrm{Cat} c_{p\mathrm{Cat}} \frac{\partial T_\mathrm{Cat}}{\partial t} = \lambda_\mathrm{eff} \frac{\partial^2 T_\mathrm{Cat}}{\partial x^2} + \frac{\lambda_\mathrm{eff}}{r} \frac{\partial}{\partial r} \left(r \frac{\partial T_\mathrm{Cat}}{\partial r} \right) - \left(\frac{1}{1-\varepsilon} \right) \alpha s \left(T_\mathrm{Cat} - T \right) \tag{340a}$$

Normalizing the equation in similar fashion to Equations (337), (338), or (339) results in

$$\begin{aligned} &\varrho_\mathrm{Cat} c_{p\mathrm{Cat}} \left(\frac{T_0}{\tau} \right) \frac{\partial T^*_\mathrm{Cat}}{\partial t^*} = \frac{\lambda_\mathrm{eff} T_\mathrm{Cat0}}{L^2} \frac{\partial^2 T^*_\mathrm{Cat}}{\partial x^{*2}} \\ &+ \lambda_\mathrm{eff} \frac{T_\mathrm{Cat0}}{R^2} \frac{1}{r^*} \frac{\partial}{\partial r^*} \left(r^* \frac{\partial T^*_\mathrm{Cat}}{\partial r^*} \right) \\ &- \left(\frac{1}{1-\varepsilon} \right) \alpha s T_\mathrm{Cat0} \left(T^*_\mathrm{Cat} - T^* \right) \end{aligned} \tag{340b}$$

Estimation of the orders of magnitude of the factors preceding the differential quotients gives

$$\frac{\lambda_\mathrm{eff} T_\mathrm{Cat0}}{L^2} \ll \lambda_\mathrm{eff} \frac{T_\mathrm{Cat0}}{R^2}$$

and thus the fundamental mechanism for the transport of enthalpy in the catalyst bed in the steady state is represented by an equation which is one-dimensional in the radial direction

$$0 = \frac{\lambda_\mathrm{eff}}{r} \frac{\mathrm{d}}{\mathrm{d}r} \left(r \frac{\mathrm{d}T_\mathrm{Cat}}{\mathrm{d}r} \right) - \left(\frac{1}{1-\varepsilon} \right) \alpha s \left(T_\mathrm{Cat} - T \right) \tag{340c}$$

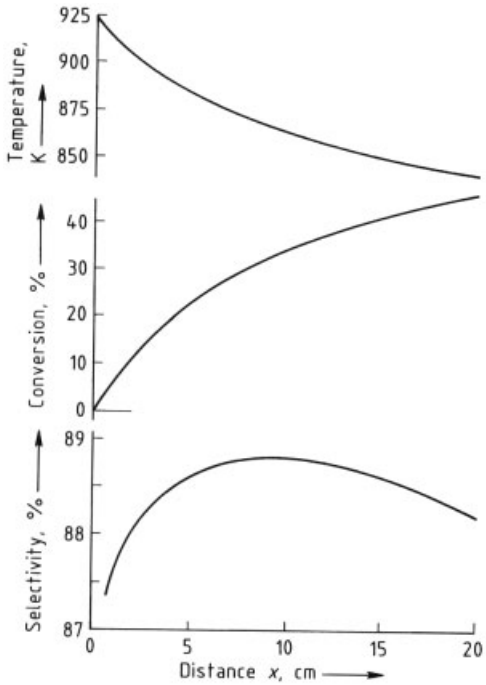

Figure 51. Calculated temperature, conversion, and selectivity profiles for the catalytic dehydration of ethylbenzene [119]

The remaining problem is the expression of the conservation equations for the chemical species by means of similar estimates of orders of magnitudes. The conservation equation for the chemical species with the simplified continuity equation (Eq. 337 d)

$$\varrho\frac{\partial m_i}{\partial t}+\varrho u\frac{\partial m_i}{\partial x}=\frac{\partial}{\partial x}\left[\varrho D_i\frac{\partial m_i}{\partial x}\right]+\frac{1}{r}\frac{\partial}{\partial r}\left[r\varrho D_i\frac{\partial m_i}{\partial r}\right]+\varrho S_i \quad i=1,\ldots,8 \tag{341a}$$

is first formulated with the variables normalized to the state at the reactor inlet.

$$\frac{\varrho_0 m_{i0}}{\tau}\varrho^*\frac{\partial m_i^*}{\partial t^*}+\frac{\varrho_0 m_{i0}u_0}{L}\varrho^* u^*\frac{\partial m_i^*}{\partial x^*}=\frac{\varrho_0 D_{i0}m_{i0}}{L^2}\varrho^* D_i^*\frac{\partial^2 m_i^*}{\partial x^{*2}}+\frac{\varrho_0 D_{i0}m_{i0}}{R^2}\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\varrho^* D_i^*\frac{\partial m_i^*}{\partial r^*}\right)+\varrho_0\varrho^* S_i,\quad i=1,\ldots,8 \tag{341b}$$

With $D_{i0}\approx 10^{-3}\,\mathrm{m^2\,s^{-1}}$

$$\frac{\varrho_0 D_{i0}m_{i0}}{L^2}\ll\frac{\varrho_0 D_{i0}m_{i0}}{R^2}\ll\frac{\varrho_0 m_{i0}u_0}{L}$$

and for the pilot plant conditions assumed here, obviously convective transport in the axial direction is dominant. The following one-dimensional equation results for the steady-state case

$$\varrho u\frac{\mathrm{d}m_i}{\mathrm{d}x}=\varrho S_i=\varrho r_i,\quad i=1,\ldots,8 \tag{341c}$$

The above estimates have resulted in a reduction of the model for the fixed bed catalyst reactor: convective transport of mass in the axial direction is the main transport mechanism for the gas phase, whilst for the catalyst phase conductive transport of heat in the radial direction along with heat transfer predominates. Consequently the couplings in Equations (131), (204), (240 b), and (265 b) are largely reduced.

To solve the system of Equations (339 c), (340 c), and (341 c), Equation (340 c) for the catalyst temperature is transformed into normalized variables with the transformations

$$r^*=\frac{r}{R},\quad \theta=\frac{T_{\mathrm{Cat}}-T}{T_{\mathrm{W}}-t}$$

(T_{W} is the wall temperature) and

$$\beta^2=\frac{R^2\alpha s}{\lambda_{\mathrm{eff}}\,(1-\varepsilon)}$$

This gives

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}r^{*2}}+\frac{1}{r^*}\frac{\mathrm{d}\theta}{\mathrm{d}r^*}-\beta^2\theta=0 \tag{342}$$

with the boundary condition $(\mathrm{d}\theta/\mathrm{d}r*)=0$ at $r*=0$. This is the modified Bessel differential equation with variable coefficients, the general solution of which is given by [122]:

$$\theta=C_1 I_0\,(\beta r^*)+C_2 K_0\,(\beta r^*) \tag{343}$$

where

$$I_p\,(\beta r^*)=\sum_{n=0}^{\infty}\left(\frac{\beta r^*}{2}\right)^{2n+p}\frac{1}{n!\,(n+p)\,!}\quad\text{and}$$

$$K_p\,(\beta r^*)=I_p\,(\beta r^*)\int\frac{\mathrm{d}\,(\beta r^*)}{\beta r^* I_p^2\,(\beta r^*)}$$

From the above boundary condition and the properties of the Bessel functions $\mathrm{d}I_0/\mathrm{d}\,(\beta r*)=I_{-1}\,(\beta r*)$ and $\mathrm{d}K_0/\mathrm{d}\,(\beta r*)=K_{-1}(\beta r*)$,

then C_2 is zero. With a further boundary condition $\theta = 1$ for $r* = 1$, it follows that $C_1 = 1/I_0(\beta)$, so the general solution changes to

$$\theta = \frac{I_0(\beta r^*)}{I_0(\beta)} \quad \text{or} \quad (T_C - T) = \frac{I_0(\beta r^*)}{I_0(\beta)}(T_W - T) \tag{344}$$

Equations (339 c) and (341 c) can only be solved numerically. Equation (339 c) has to be transformed into a difference equation in the axial direction. With the solution of Equation (344) for the catalyst temperature the cross-section-averaged heat flow to the catalyst bed can be calculated for each control volume. For this Equation (339 c) with solution Equation (344) is integrated over the cross section of the reactor:

$$0 = -\int_0^1 \varepsilon 2\pi r^* \varrho c_p u \frac{\mathrm{d}T}{\mathrm{d}x} \mathrm{d}r^* + \int_0^1 \varepsilon 2\pi r^* \frac{\alpha s}{\varepsilon}(T_C - T)\,\mathrm{d}r^* + \int_0^1 2\pi r^* \sum r_i(-\Delta h_{r_i})\,\mathrm{d}r^* \tag{345}$$

The solution of the integral $\int_0^1 (T_C - T)\, r^* \mathrm{d}r^*$ also results from the general properties of the Bessel functions

$$\int_0^1 (T_C - T)\, r^* \mathrm{d}r^* = \frac{R^{3/2}}{\sqrt{\beta}} \frac{I_1(\beta)}{I_0(\beta)}(T_W - T) \tag{346}$$

This expression can be inserted into Equation (339 c) which is then solved numerically and simultaneously with Equation (341 c).

Numerical solutions for pilot plant conditions with integration routines from [28] are taken from [119] and shown in Figure 51. Kinetic data for reactions (i)–(vi) (see above) are given in [8], [11], [119]. A drop in temperature of ca. 75 K and a continuous increase in conversion occur over the length of the reactor. At higher conversions the reactions forming the byproducts gain greater importance.

The estimates used to obtain the model in the form of Equations (339 c), (340 c), and (341 c) are undoubtedly drastic. They should therefore be regarded as an example of how a two-dimensional problem (a reactor model with axial and radial dispersion) can be reduced to a one-dimensional problem and of how the system of model differential equations is decoupled. Nevertheless the model successfully reproduces experimental temperature and concentration profiles measured in a pilot plant [11]. Another special feature of the above model is the treatment of the heterogeneous catalytic reactions. Mass transport processes in the catalyst phase were assumed to have no effect on the reaction rates. The heterogeneous system is thereby regarded as being quasi-homogeneous and all couplings between mass fractions in the gas phase and catalyst phase are neglected. This assumption is not always true. The effect of transport processes on the reaction rates and the repercussion on models for reactors with heterogeneous catalysis are discussed below.

Some Particulars of Heterogeneous Catalytic Reactions. The effect of transport processes on reaction rates in heterogeneous catalysis can be treated analogously to the acceleration of mass transport by chemical reactions described in Section 5.3.2 (Fig. 38). The core of the liquid phase is replaced by a porous catalyst sphere that is one-dimensional in spherical coordinates (Fig. 52). Furthermore if a simple reaction A → products is considered, the heterogeneous reaction rate of which is assumed to follow a formal first-order rate law $r = -k' m_A = -k' m$, where k' is a rate coefficient related to the surface area of the catalyst.

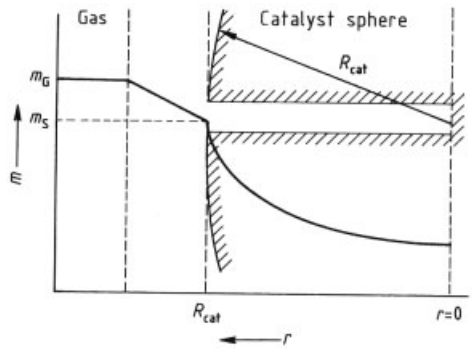

Figure 52. The effect of mass transport in heterogeneous catalytic reactions
m = mass fraction; r = radial distance; R_{cat} = radius of catalyst sphere
The subscripts G and S denote gas phase and surface of catalyst, respectively.

The steady-state mass flux to the surface of the catalyst (see also Section 5.3.1) is given by

$$j^m = \varrho\beta s\,(m_G - m_S) \tag{347}$$

This flow is compensated for by the consumption of component A by the chemical reaction. If an "effective" reaction rate $r_S = -k'\,s\,m_S$ with the mass fraction at the surface of the catalyst is formulated, then the mass fraction m_S for steady-state conditions is given by

$$m_S = \frac{m_G}{\left(1+\frac{k}{\beta s}\right)} \tag{348}$$

where $k = k'\,s$. In Equation (347) m_S can now be eliminated so that

$$j^m = \varrho\beta s m_G\left(1-\left(1+\frac{k}{\beta s}\right)^{-1}\right) \tag{349}$$

Equation (349) also gives the effective reaction rate for steady-state conditions. As in Section 5.3.2, Equation (349) illustrates the acceleration of mass transport with an increasing ratio $Da_{II} = k/\beta\,s$, i.e., with increasing reaction rate where Da_{II} is the Damköhler number of the second kind. The degree of external utilization (i.e., effectiveness) of the catalyst η_{ext} is defined as the ratio of the surface reaction rate r_S to the reaction rate for $Da_{II} \to 0$, i.e., $m_S \to m_G$. It thus follows that (cf. Eq. 348)

$$\eta_{ext} = \frac{k m_G}{k m_G\left(1+\frac{k}{\beta s}\right)} = \frac{1}{1+Da_{II}} \tag{350}$$

For nonisothermal conditions the temperature dependence of the reaction rate must be considered. If Da_{II} is defined with the conditions in the gas phase and constant density is assumed, it follows from Equation (350) that

$$\eta_{ext} = \frac{k\,(T_S)}{k\,(T_G)}\left(1+\frac{k\,(T_S)}{k\,(T_G)}Da_{II}\right)^{-1} \tag{351}$$

Given the temperature dependence of the rate coefficient in the form of an Arrhenius equation $k = k_0 e^{-E_a/RT}$, the temperature at the surface T_S must be eliminated from Equation (351) to calculate η_{ext}. The temperature at the surface T_S may be calculated in a similar way as m_S. In the steady-state case

$$j^h = \alpha s\,(T_S - T_G) \tag{352}$$

The heat transferred to the surroundings is released by chemical reactions occurring at the catalyst. Thus from the steady-state condition $j^h = r_S(-\Delta h_r)$

$$\frac{T_G}{T_S} = 1+\left[\frac{\varrho\,(-\Delta h_r)\,m_G}{\frac{\lambda}{D_{eff}}LeT_G}\,\frac{\eta_{ext}k\,(T_G)}{\beta s}\right]^{-1} \tag{353}$$

To derive Equation (353) full analogy for heat and mass transfer, i.e., $Pr = Le \cdot Sc$, is assumed. Using the abbreviations $\gamma = E_a/R\,T_G$ (Arrhenius number), and

$$\beta_H = \frac{\varrho\,(-\Delta h_r)\,m_G}{\frac{\lambda}{D_{eff}}LeT_G} \quad \text{(Prater number)}$$

Equations (351) and (353) are combined to give

$$\frac{\eta_{ext}}{\exp\left[-\gamma\,(1+\beta_H\eta_{ext}Da_{II})^{-1}\right]} = \left\{1+Da_{II}\exp\left[-\gamma\,(1+\beta_H\eta_{ext}Da_{II})^{-1}\right]\right\}^{-1} \tag{354}$$

η_{ext} can also take values greater than 1 in the nonisothermal case if $\beta_H > 0$. In this case the reaction at the surface of the catalyst occurs at a higher temperature than the gas-phase temperature T_G. Since all quantities in Equation (354) are related to T_G, then values of $\eta_{ext} > 1$ are also consistent with the definition of η_{ext}.

Conditions inside the catalyst can be illustrated in a similar fashion. For spherical geometry (Fig. 52) diffusive mass flux inside the catalyst is given by

$$j^m = -\varrho D_{eff}4\pi r^2\frac{dm}{dr} \tag{355}$$

Changes in the diffusive mass flux are caused by chemical reactions at the internal surface of the catalyst, so that under steady-state conditions the mass balance is

$$\frac{dj^m}{dr} = \varrho k' 4\pi r^2 m \tag{356}$$

Using Equation (355)

$$\frac{d^2\varphi}{d\eta^2}+\frac{2}{\eta}\frac{d\varphi}{d\eta} - \Phi_K^2\varphi = 0 \tag{357}$$

is obtained where the normalized quantities $\varphi = m/m_S$ and $\eta = r/R_{Cat}$ and the abbreviation $\Phi_K = R_{Cat}\sqrt{k's/D_{eff}}$ are used (R_{Cat} is the radius of the catalyst sphere). The solution of

Equation (357) with the boundary conditions $\varphi = 1$ for $\eta = 1$ and $d\varphi/d\eta = 0$ for $\eta = 0$ is

$$\varphi = \frac{\sinh(\Phi_K \eta)}{\eta \sinh(\Phi_K)} \tag{358}$$

Completely analogous to the external mass transfer, a degree of internal utilization of the catalyst η_{int} (internal effectiveness factor) can be defined as the ratio of the mean reaction rate in the catalyst,

$$r_{eff} = \int_0^{R_{Cat}} k' s m 4\pi r^2 dr$$

and reaction rate for $m \to m_S$. Integration of Equation (358) gives

$$\eta_{int} = \frac{3}{\Phi_K}\left(\tanh\Phi_K - \frac{1}{\Phi_K}\right)^{-1} \tag{359}$$

For $\Phi_K > 3$, $\eta_{int} \approx 3/\Phi_K$, for $\Phi_K < 0.3$, $\eta_{int} \approx 1$. The result is similar to the acceleration of mass transfer by chemical reactions, cf. Section 3.3.2. The calculation of η_{int} for other catalyst geometries including more realistic pore structures and other classes of reaction are described in [24], [123].

For nonisothermal conditions the temperature dependence of the reaction rate must once again be taken into consideration. If the simple Arrhenius expression is substituted for k' in Equation (356), then assuming constant density and neglecting the temperature dependence of the effective diffusion coefficient, the result is

$$\frac{d^2 m}{dr^2} + \frac{2}{r}\frac{dm}{dr} = \frac{k_0'}{D_{eff}} s \exp\left(-\frac{E_a}{RT}\right) m \tag{360}$$

Inside the catalyst heat is transported by conduction. The conductive heat flow is given by

$$j^h = -\lambda 4\pi r^2 \frac{dT}{dr} \tag{361}$$

Changes in heat flow are caused by heat release due to chemical reactions. Under steady-state conditions

$$\frac{dj^h}{dr} = \varrho k' s 4\pi r^2 m (-\Delta h_r) \tag{362}$$

From Equations (361) and (362) a differential equation for the temperature is obtained

$$\frac{d^2 T}{dr^2} + \frac{2}{r}\frac{dT}{dr} = \frac{\varrho(-\Delta h_r) k_0' s \exp\left(-\frac{E_a}{RT}\right)}{D_{eff}\frac{\lambda_{eff}}{D_{eff}}} m \tag{363}$$

Subtracting Equation (363) from Equation (360) gives

$$\frac{d^2 m}{dr^2} + \frac{2}{r}\frac{dm}{dr} - \frac{\frac{\lambda_{eff}}{D_{eff}}}{\varrho(-\Delta h_r)}\left[\frac{d^2 T}{dr^2} + \frac{2}{r}\frac{dT}{dr}\right] = 0 \tag{364}$$

which, after integrating twice with the boundary conditions $m = m_S$ and $T = T_S$ for $r = R_{Cat}$, and $dm/dr = 0$, and $dT/dr = 0$ for $r = 0$, leads to the result

$$T - T_S = \frac{\varrho D_{eff}(-\Delta h_r)}{\lambda_{eff}}(m_S - m) \tag{365}$$

With this the temperature can now be eliminated from Equation (360). If the Prater number, Arrhenius number, and Thiele modulus Φ_K are defined with the conditions at the outer surface of the catalyst, i.e.,

$$\beta_H = \frac{\varrho(-\Delta h_r) m_S}{\frac{\lambda}{D_{eff}} T_S}, \quad \gamma = \frac{E_a}{RT_S}, \quad \text{and}$$

$$\Phi_K^2 = R_{Cat}^2 \frac{k_0' s}{D_{eff}} \exp\left(-\frac{E_a}{RT_S}\right)$$

Equation (364) becomes

$$\frac{d^2\varphi}{d\eta^2} + \frac{2}{\eta}\frac{d\varphi}{d\eta} - \Phi_K^2 \varphi \exp\left[\gamma\beta_H \frac{(1-\varphi)}{1+\beta_H(1-\varphi)}\right] = 0 \tag{366}$$

Equation (366) cannot be solved analytically, numerical solutions are given in [124]. If the internal effectiveness factor η_{int} is calculated from the numerical solutions [124] the results shown in Figure 53 are obtained. For $\gamma = 20$, multiple solutions exist for η_{int} at high values of β_H and low values of the Thiele modulus. The middle solutions are unstable so that ignition and extinction of the exothermic chemical reaction in the catalyst can occur for certain values of Thiele moduli. As all parameters are related to the state at the outer surface of the catalyst, values of $\eta_{int} > 1$ are consistent with the definition of the internal effectiveness factor. Due to the exothermic reaction the temperature inside the catalyst can be higher than temperature on the outer surface.

By introducing an effective reaction rate $r_{eff} = \eta_{eff} r_G$, the rate of heterogeneous catalytic reaction is related to the conditions in the gas

phase and is thus no longer coupled with the internal transport in the catalyst. This permits the description of reactors with heterogeneous catalysis as quasi-single-phase systems. The effectiveness factors η_{int} or η_{ext} can be calculated from the variables of state at the surface of the catalyst or in the gas phase, respectively. For the isothermal case and spherical geometry

$$\eta_{\mathrm{eff}} = \left(\frac{1}{\eta_{\mathrm{int}}} + \frac{\Phi_{\mathrm{K}}^2}{3Bi} \right)^{-1}$$

where the Biot number $Bi = \beta R_{\mathrm{Cat}}/D_{\mathrm{eff}}$.

Nonisothermal conditions and other geometries are discussed in [24], [123], [124].

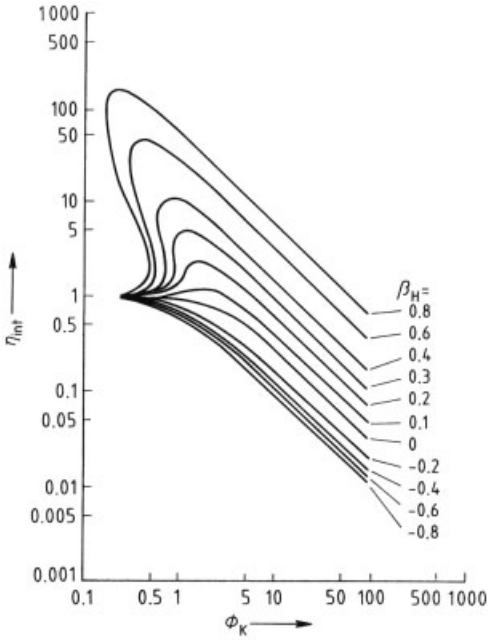

Figure 53. Effectiveness factor of a catalyst as a function of the Thiele modulus Φ_{K} and the Prater number β_{H} for spherical catalyst geometry (Arrhenius number $\gamma = 20$) [124]
γ = Arrhenius number; η_{int} = internal effectiveness factor.

The discussion in this section was based on first-order reactions with no changes in the mole number. Very few reactions in heterogeneous catalysis can be described with these expressions, higher order reactions are much more frequent. Furthermore the catalysts may be "poisoned" by adsorption of the reactants or products (see page 85) yielding nonintegral orders of reaction. Inhibition by mass transport influences the order of the catalytic reaction. In the extreme case $r_{\mathrm{eff}} = \beta s (m_{\mathrm{G}} - m_{\mathrm{S}})$. This rate expression is of the first order with respect to the mass fraction of A. Inhibition by mass transport shifts the order of reaction in the direction of first order. Selectivity in parallel or secondary reactions will be similarly affected.

The temperature dependence of the reaction rate is also influenced by inhibition by mass transport. If the expression $r_{\mathrm{eff}} = -\eta_{\mathrm{eff}} k m_{\mathrm{G}} = -k* m_{\mathrm{G}}$ is used for a first-order reaction, and the temperature dependence is given as usual by $k* = k_0^* \mathrm{e}^{-E_{\mathrm{a}}^*/RT}$, then

$$\ln k^* = \ln k_0^* - \frac{E_{\mathrm{a}}^*}{RT} \tag{367}$$

For $\eta_{\mathrm{eff}} = 1$ there is chemical control, $k_0^* = k_0$ and $E_{\mathrm{a}}^* = E_{\mathrm{a}}$. If diffusion in the catalyst pores is rate controlling, $\eta \approx 3/\Phi_{\mathrm{K}}$. Then $\ln k* = \ln k_0 + \ln 3 - \ln \Phi_{\mathrm{K}} - E_{\mathrm{a}}/RT$. Substitution of $\Phi_{\mathrm{K}} = R_{\mathrm{Cat}} \sqrt{k/D_{\mathrm{eff}}}$ gives

$$\ln k^* = \frac{1}{2} \ln k_0 + \ln 3 + \frac{1}{2} \ln D_{\mathrm{eff}} - \frac{1}{2} \frac{E_{\mathrm{a}}}{RT}$$

thus $k_0^* = 3\sqrt{k_0 D_{\mathrm{eff}}}$ and $E_{\mathrm{a}}^* = \frac{1}{2} E_{\mathrm{a}}$. If the temperature dependence of the effective diffusion coefficients is neglected, the activation energy for pore diffusion-controlled reaction is about half that for chemical control. If control is exclusively due to mass transfer, then $r_{\mathrm{eff}} \approx -\beta s m_{\mathrm{G}}$ and $k* = \beta s \sim D_{\mathrm{eff}}$. Taking into account the temperature dependence of D_{eff}, then $\ln k^* = \ln k_0^* + \frac{3}{2} \ln T$ which leads to an activation energy E_{a}^* of ca. 5 kJ/mol for this case. Thus the experimental determination of the effective formal activation energy for a heterogeneous catalytic reaction clarifies the extent of control by different transport processes. For further discussion of these relationships see [24], [123], [124].

In addition to the formulation of conservation equations for reactor models for heterogeneous catalytic reactions, combinations of simple reactor elements are frequently used for modeling these systems. For further information, see [12], [125]. Other complicated reactors (e.g., fluidized bed reactors) for heterogeneous catalytic reactions can be similarly dealt with [24], [123].

5.3.3.2.2. Stability Analysis of Nonisothermal Reactors

Steady-State Cases. The stability analysis of isothermal reactors discussed in Section

5.3.3.1 demonstrated the close connection of the phenomenon of multiple steady states in steady-state systems and the instability of nonsteady-state systems to a feedback mechanism. For the isothermal cases a chemical feedback mechanism was discussed in the form of autocatalysis and self-poisoning. Another feedback mechanism is the heat release from a chemical reaction. This will be discussed for a simple first-order reaction, $A \rightarrow B + \text{heat}$, the temperature dependence of the rate coefficient is of the Arrhenius form, $k = k_0 e^{-E^a/RT}$. With increasing conversion of A heat is released. The released heat increases the temperature of the reaction mixture and the reaction rate. This in turn increases the rate of heat release. The reaction rate is an exponential function of the temperature. Therefore the feedback mechanism is nonlinear in temperature.

Stability analysis for nonisothermal steady-state reactors will again be carried out for a well-stirred reactor (one element of a reactor cascade of ideally mixed reactors which is used to represent a flow reactor with convectively dominated flow). In analogy with Equation (310), the mass balance for A gives

$$\frac{\varrho}{\tau}(m_0 - m) - \varrho m k_0 e^{-E_a/RT} = 0 \tag{368}$$

From the enthalpy balance and transformation of Equation (240 b) into an equation for temperature, the following equation is obtained:

$$\frac{c_p \varrho}{\tau}(T_0 - T) + (-\Delta h_r)\,\varrho m k_0 e^{-E_a/RT} - \alpha s (T - T_A) = 0 \tag{369}$$

where the specific surface area s is the effective surface for heat transfer to the surroundings divided by the total volume of the ideally mixed reactor. The source term in Equation (240 b) is thus represented by a heat-transfer term. As in Section 5.3.3.1, Equations (368) and (369) are transformed into normalized variables: $\alpha = m_A/m_{A_0}$, $\theta = (T - T_0)E_a/RT_0^2$, $\theta_C = (T_A - T_0)E_a/RT_0^2$, and $Da_I = \tau k_0$. The Damköhler number (the ratio of the mean residence time to a characteristic chemical time, the time required until $m/m_0 = e^{-1}$) is defined with the inlet conditions. If the adiabatic temperature difference $\Delta T_{ad} = (-\Delta h_r)\,m_0/c_p$ is also defined with E_a/RT_0^2, then $\theta_{ad} = (-\Delta h_r)\,m_0 E_a/c_p RT_0^2$. Finally a "cooling time"

$$\tau_N = \frac{c_p \varrho}{\alpha s t_{chem}} = \frac{c_p \varrho k_0}{\alpha s}$$

can be defined from the heat transfer. A reciprocal Arrhenius number $\gamma_r = RT_A/E_a$ is also introduced. With these abbreviations Equations (368) and (369) become

$$\frac{1-\alpha}{Da_I} - \alpha \exp\left(\frac{\theta}{1+\gamma_r\theta}\right) = 0 \tag{370a}$$

$$\theta_{ad}\alpha \exp\left(\frac{\theta}{1+\gamma_r\theta}\right) - \left(\frac{1}{Da_I} + \frac{1}{\tau_N}\right)\theta + \frac{\theta_C}{\tau_N} = 0 \tag{370b}$$

Equations (370 a, b) contain the variables α and θ as a function of six parameters. To reduce the number of parameters, the case of very high activation energy, $\gamma_r \ll 1$, $T_A = T_0$, and $\theta_C = 0$ is considered. With these assumptions the following simplified equations are obtained

$$\frac{1-\alpha}{Da_I} - \alpha e^{\theta} = 0 \tag{371a}$$

$$\theta_{ad}\alpha e^{\theta} - \left(\frac{1}{Da_I} + \frac{1}{\tau_N}\right)\theta = 0 \tag{371b}$$

In order to remain as close as possible to the notation used in Section 5.3.3.2, θ in Equation (371 a) is replaced by Equation (371 b) so that

$$\frac{1-\alpha}{Da_I} = \alpha \exp\left[\frac{(1-\alpha)\,\theta_{ad}}{1+\frac{Da_I}{\tau_N}}\right] \tag{372a}$$

The temperature increase can easily be attributed to α from Equations (371 a, b):

$$\theta = \left[\frac{(1-\alpha)\,\theta_{ad}}{1+\frac{Da_I}{\tau_N}}\right] \tag{372b}$$

Equation (372 a) can be solved graphically by plotting both sides against the progress of the reaction $1 - \alpha$ (conversion). In order to further reduce the number of parameters, ideal adiabatic conditions are assumed, i.e., $\tau_N \rightarrow \infty$. Thus the left-hand side (LHS) of Equation (372 a) is only dependent on Da_I, whilst the right-hand side (RHS) is only dependent on the adiabatic temperature increase. The solutions for the case discussed here (Fig. 54) are qualitatively similar to the solutions for the isothermal case with self poisoning (see page 79). Here too for a given residence time (Damköhler number) multiple steady states occur for particular areas of the

adiabatic temperature increase θ_{ad}. Using the same argumentation as in Section 5.3.3.2, the stability condition Equation (318 a)

$$\frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,(\mathrm{LHS}) > \frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,(\mathrm{RHS}) \quad \text{at} \quad \mathrm{LHS} = \mathrm{RHS} \tag{318a}$$

is not complied with for the middle one of the multiple solutions. In this state a small perturbation of the conversion causes an increase or decrease in the "convective term" (LHS). The "reaction term" (RHS) however reacts in the same way so that a greater (smaller) availability of A due to convection is followed by a greater (smaller) consumption due to reaction. The system therefore tends towards one of the external stable steady-state solutions. As can be seen from Figure 54 A, the shape of the curve depends on the adiabatic temperature increase θ_{ad}. Figure 54 B shows that as θ_{ad} becomes smaller the S-shaped profile vanishes and the surface of solutions unfolds. Figure 54 B also demonstrates that the unstable branch of the solutions is not passed. The system jumps from low conversion to high conversion with increasing residence time Da_{I}. On igniting and extinguishing the reaction, hysteresis occurs and the jumping points can be calculated from Equation (318 b)

$$\frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,(\mathrm{LHS}) = \frac{\mathrm{d}}{\mathrm{d}\,(1-\alpha)}\,(\mathrm{RHS}) \quad \text{at} \quad \mathrm{LHS} = \mathrm{RHS} \tag{318b}$$

For the example discussed here the jumping points are given by

$$(1-\alpha) = \frac{1}{2}\left[1 \pm \sqrt{\left(1-\frac{4}{\theta_{\mathrm{ad}}}\right)}\right] \tag{373}$$

Since θ_{ad} is always positive for exothermic reactions, then for the point at which the discriminant in Equation (373) disappears, $\theta_{\mathrm{ad}} = 4$. If $\theta_{\mathrm{ad}} > 4$, multiple steady states are observed.

If the simplification $\gamma_{\mathrm{r}} \to 0$ is discarded, the following equation is obtained for the steady-state solutions

$$\frac{1-\alpha}{Da_{\mathrm{I}}} = \alpha \exp\left[\frac{(1-\alpha)\,\theta_{\mathrm{ad}}}{1+\gamma_{\mathrm{r}}\theta_{\mathrm{ad}}\,(1-\alpha)}\right] \tag{374a}$$

The right-hand side of Equation (374 a) now depends on two parameters. In general it can be written in the form

$$F\,(x, y; p, q, r, s, \ldots) = 0 \tag{374b}$$

where x is the variable for describing the steady state, y the bifurcation parameter which controls the multiple steady state, and $p, q, r, s, \ldots$ are additional parameters which produce folding or unfolding of the surface of x in space y and $p, q, r, s, \ldots$

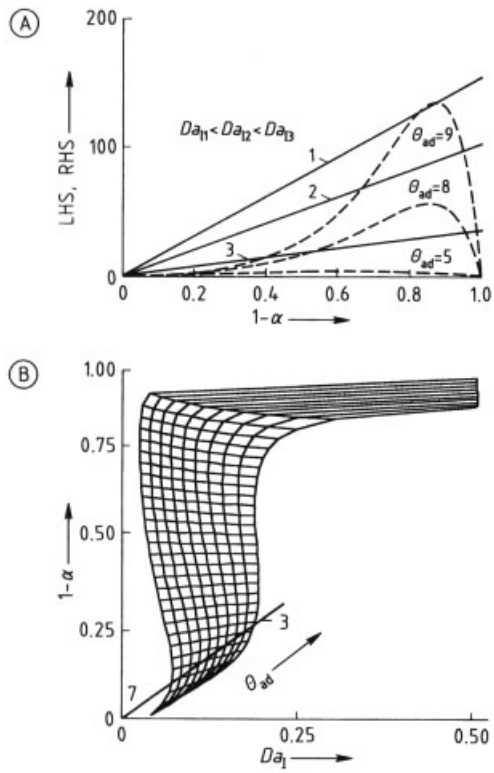

Figure 54. Multiple steady-state conversion points for "normal" exothermic chemical reactions (A) and bifurcation diagram (B) for nonisothermal conditions in a well-stirred reactor
θ_{ad} = adiabatic temperature increase; for explanation of other symbols see Figure 45.

One or more steady-state solutions exist for the condition given by Equation (374 b). The steady-state solution is stable for

$$F\,(x, y; p, q, r, s, \ldots) = 0 \text{ and } F_x > 0 \tag{374c}$$

For the unfolding of the hysteresis loop the following applies

$$F\,(x, y; p, q, r, s, \ldots) = 0, \quad F_x > 0, \quad \text{and} \quad F_{xx} = 0 \tag{374d}$$

and "islands" finally occur in the bifurcation diagram which grow into "mushrooms" when

$$F(x, y; p, q, r, s, \ldots) = 0, \quad F_x = 0, \quad F_y = 0,$$
$$F_{xy} \neq 0, \quad F_{xx} \neq 0, \quad \text{and} \quad F_{yy} \neq 0 \tag{374e}$$

Depending on the process under consideration these conditions apply for various parameter combinations. For Equation (374 a) the condition for unfolding occurs when $\theta_{\mathrm{ad}}(4\gamma_{\mathrm{r}} - 1) + 4 = 0$. In the system of Equations (374), F_x, F_y, etc. denote the derivatives according to the variable in the index.

Nonsteady-State Cases. As in the stability analysis of isothermal reactors, the stability of nonsteady-state nonisothermal reactors is investigated for a consecutive reaction $\mathrm{P} \rightarrow \mathrm{A} \rightarrow \mathrm{B}$ where A is produced from a large reservoir of P. The second reaction $\mathrm{A} \rightarrow \mathrm{B}$ is exothermic and its rate coefficient k_1 is temperature dependent. For the reaction sequence $\mathrm{P} \rightarrow \mathrm{A} \rightarrow \mathrm{B}$ where $k_1 = k_{01}\, \mathrm{e}^{-\mathrm{E_a}/\mathrm{RT}}$, the mass and enthalpy balances for a batchwise operating ideally mixed reactor are given by

$$\frac{\mathrm{d}m_{\mathrm{P}}}{\mathrm{d}t} = -k_0 m_{\mathrm{P}} \tag{375}$$

$$\frac{\mathrm{d}m_{\mathrm{A}}}{\mathrm{d}t} = k_0 m_{\mathrm{P}} - k_1 m_{\mathrm{A}} \tag{376a}$$

$$\varrho c_p \frac{\mathrm{d}T}{\mathrm{d}t} = (-\Delta h_{\mathrm{r}})\, k_1 m_{\mathrm{A}} - \alpha s\, (T - T_{\mathrm{A}}) \tag{376b}$$

(see page 81). The feedback in this system results from the temperature dependence of the second reaction rate which increases the reaction rate with increasing conversion. As previously, normalized variables are used: $\alpha = m_{\mathrm{A}}/m_{\mathrm{ref}}$, $\pi = m_{\mathrm{P}}/m_{\mathrm{ref}}$, $\theta = (T - T_{\mathrm{A}})\, E_{\mathrm{a}}/RT_{\mathrm{A}}^2$ and $\gamma_{\mathrm{r}} = RT_{\mathrm{A}}/E_{\mathrm{a}}$. For the temperature-dependent reaction rate coefficient it thus follows that

$$k_1 = k_{01}\, \mathrm{e}^{-E_{\mathrm{a}}/RT_{\mathrm{A}}} \exp\left[\theta/\left(1 + \gamma_{\mathrm{r}}\theta\right)\right]$$

The cooling time is not related to a characteristic chemical time $\tau_{\mathrm{N}} = \varrho\, c_p/\alpha\, s$, so that a normalized time can be given as $\tau = t/\tau_{\mathrm{N}}$. Thus the rate coefficients k_0 and k_1 are normalized so that $\varepsilon = k_0 \tau_{\mathrm{N}}$ and $\kappa = \tau_{\mathrm{N}} k_{01}\, \mathrm{e}^{-\mathrm{E_a}/\mathrm{RT_A}}$. Finally m_{ref} is established from

$$m_{\mathrm{ref}} = \frac{\alpha s R T_{\mathrm{A}}^2}{\varrho E_{\mathrm{a}}\, (-\Delta h_{\mathrm{r}})\, k_{01}\, \mathrm{e}^{-E_{\mathrm{a}}/RT_{\mathrm{A}}}}$$

There is no feedback to reservoir P thus Equation (375) can be integrated separately:

$$m_{\mathrm{P}} = m_{0\mathrm{P}}\, \mathrm{e}^{-k_0 t} \quad \text{or} \quad \pi = \pi_0 \mathrm{e}^{-\varepsilon\tau} \tag{377}$$

Using this Equations (376 a) and (377 b) give

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\tau} = \varepsilon \pi_0 \mathrm{e}^{-\varepsilon\tau} - \kappa\alpha \exp\left[\frac{\theta}{(1 + \gamma_{\mathrm{r}}\theta)}\right]$$
$$= \mu - \kappa\alpha \exp\left[\frac{\theta}{(1 + \gamma_{\mathrm{r}}\theta)}\right] \tag{378a}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = \alpha \exp\left[\frac{\theta}{(1 + \gamma_{\mathrm{r}}\theta)}\right] - \theta \tag{378b}$$

where $\mu = \varepsilon\, \pi_0 \mathrm{e}^{-\varepsilon\tau} = \mu_0 \mathrm{e}^{-\varepsilon\tau}$. For the sake of simplicity the case of high activation energy (i.e., $\gamma_{\mathrm{r}} \ll 1$) will be investigated. Equations (378 a, b) are thus again simplified into

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\tau} = \mu - \kappa\alpha \mathrm{e}^{\theta} \tag{379a}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}\tau} = \alpha \mathrm{e}^{\theta} - \theta \tag{379b}$$

For very small values of ε (i.e., for a very much quicker time scale for the changes in α and θ than for the consumption of P) the principle of quasi-steady-state can again be used, and hence Equations (379 a, b) do not have to be integrated. The quasi-steady-state solutions for α and θ result from $\mathrm{d}\alpha/\mathrm{d}\tau = 0$ and $\mathrm{d}\theta/\mathrm{d}\tau = 0$, hence

$$\mu - \kappa\alpha \mathrm{e}^{\theta} = 0 \tag{380a}$$

$$\alpha \mathrm{e}^{\theta} - \theta = 0 \tag{380b}$$

giving

$$\theta_{\mathrm{qs}} = \frac{\mu}{\kappa} \tag{381a}$$

$$\alpha_{\mathrm{qs}} = \frac{\mu}{\kappa} \mathrm{e}^{-\mu/\kappa} \tag{381b}$$

The approximate solutions α_{qs} and θ_{qs} are a function of the ratio μ/κ, and also of time since μ is a function of time. With these operations and simplifications the quasi-steady-state solutions α_{qs} and θ_{qs} in Equation (381) have the same form as for the isothermal case (Eq. 324). The stability analysis which now follows is carried out in the same way as for the isothermal case (Section 5.3.3.1).

The development of Equation (380 a) into a Taylor series around the quasi-steady-state solutions provides a system of coupled differential

equations for the disturbances $\Delta\alpha$ and $\Delta\theta$, the general solution of which is given by

$$\Delta\alpha(\tau) = C_1 e^{\lambda_1 \tau} + C_2 e^{\lambda_2 \tau} \tag{382a}$$

$$\Delta\theta(\tau) = C_3 e^{\lambda_1 \tau} + C_4 e^{\lambda_2 \tau} \tag{382b}$$

The eigenvalues λ_1, λ_2 once again determine the characteristics of the time evolution of the disturbances. They are given by the solutions to the characteristic Equation (329)

$$\lambda^2 - \mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})\lambda + \mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = 0 \tag{329}$$

i.e.,

$$\lambda_{1,2} = \frac{1}{2}\left[\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) \pm \left\{\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})^2 - 4\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})\right\}^{1/2}\right] \tag{330}$$

In this case the Jacobi matrix takes the form

$$\boldsymbol{J}_{\mathrm{qs}} = \begin{pmatrix} \frac{\partial f(\alpha,\theta)}{\partial \alpha} & \frac{\partial f(\alpha,\theta)}{\partial \theta} \\ \frac{\partial g(\alpha,\theta)}{\partial \alpha} & \frac{\partial g(\alpha,\theta)}{\partial \theta} \end{pmatrix}_{\mathrm{qs}} \tag{383}$$

The three quantities which decide the nature of the solutions are the discriminant of the characteristic equation $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}})^2 - 4\,\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}})$, the trace of the Jacobi matrix

$$\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) = \left(\frac{\mu}{\kappa}\right) - 1 - \kappa e^{\mu/\kappa} \tag{384a}$$

and also their determinant

$$\mathrm{Det}(\boldsymbol{J}_{\mathrm{qs}}) = \kappa e^{\mu/\kappa} \tag{384b}$$

As μ/κ is always positive, not all the combinations of λ_1, λ_2 given in the stability diagram (Fig. 48) will be passed through with increasing conversion of P.

Two states will now be discussed that are of practical importance. The first case is the transition of the solutions from a node of the system into a focus. The condition for this is that λ_1, λ_2 become conjugate complex, the discriminant thus passes through zero. If μ in Equations (384 a, b) is replaced by Equation (381 a), the result is the following equation for the discriminant

$$\kappa^2 e^{2\theta_{\mathrm{qs}}} - 2\kappa(1+\theta_{\mathrm{qs}})e^{\theta_{\mathrm{qs}}} + (\theta_{\mathrm{qs}} - 1)^2 = 0 \tag{385}$$

For each positive value of θ_{qs}, there are thus two values of κ

$$\kappa_{1,2} = \left(\sqrt{\theta_{\mathrm{qs}}} \pm 1\right)^2 e^{-\theta_{\mathrm{qs}}} \tag{386a}$$

which satisfy the above condition for the discriminant to be zero. The appropriate values for μ are

$$\mu_{1,2} = \kappa_{1,2}\theta_{\mathrm{qs}} \tag{386b}$$

The stability diagram is best displayed in the form given in Figure 55 with the two parameters μ and κ. Two closed loops are obtained, the outer of which corresponds to the larger root from Equation (386 a) and the inner to the smaller. The area outside the outer curve represents stable nodes, stable or unstable foci lie between the two curves, and the area within the inner curve gives unstable nodes.

The second case is the loss of stability of the system given by the condition $\mathrm{Tr}(\boldsymbol{J}_{\mathrm{qs}}) = 0$. This is also a condition for Hopf bifurcation. If μ in Equation (384 a) is replaced by Equation (381 a), this condition is given as

$$\kappa = (\theta_{\mathrm{qs}} - 1)e^{-\theta_{\mathrm{qs}}} \quad \text{and} \quad \mu = \kappa\theta_{\mathrm{qs}} \tag{387a,b}$$

The pertinent curve in the stability diagram is a closed loop which lies between the two loops for the transition from nodes to foci. The region within this curve denotes unstable foci, see the enlarged section of Figure 55 in the area of the origin. The maximum of this curve is at e^{-2}.

Stability analysis for nonisothermal nonsteady-state batchwise operating models exhibits the following important results: for each combination of the experimental conditions μ and κ there is an unambiguous solution for the quasi-steady-state concentration of the intermediate α_{qs} and the normalized temperature increase θ_{qs}. If the normalized reaction rate coefficient is $\kappa > e^{-2}$, then the quasi-steady-state solutions are always stable. For $\kappa \approx 1.78$ transition from stable nodes to stable foci occurs. For $\kappa < e^{-2}$ the quasi-steady-state solutions are unstable. At the location of the Hopf bifurcations two solutions occur (Fig. 55). Between the two bifurcation points for α_{qs} and θ_{qs} as functions of the conversion (i.e., as functions of μ), oscillations of varying frequency and amplitude occur.

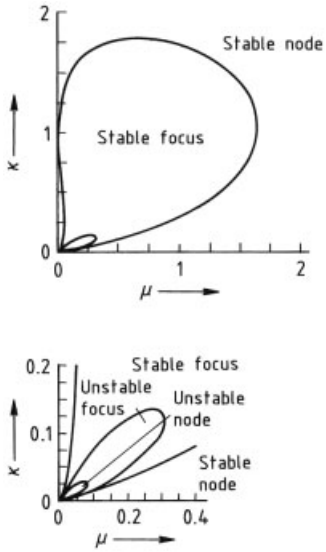

Figure 55. Stability diagram for exothermic consecutive reactions for quasi-steady-state nonisothermal conditions [113]
For explanation of symbols see Figure 49 and text.

For engineering purposes it is important that, for the simplified case discussed here, stability can be controlled through the parameter κ. For $\kappa > e^{-2}$ the system remains stable. If the properties of the chemical reaction are given by $k_1 = k_{01}\, e^{-E_a/RT}$, stability can only be achieved by altering the cooling time by means of increased heat transfer to the surroundings. Detailed discussions on the stability of nonisothermal reactors can be found in [113], [114].

5.3.3.2.3. Use on Statistical Processes

The description of statistical processes by averaging the conservation equations of momentum is discussed in Section 5.1.2 for isothermal inert systems. The averaged equations however contain terms with higher moments which must be modeled; in Equation (216) these are the Reynolds stresses $\langle \varrho \rangle \langle U_i^f U_j^f \rangle$.

A model for the Reynolds stresses which shows them in analogy to the laminar viscous stresses as an isotropic tensor of turbulent viscous stresses

$$\tau_{t_{ij}} = \mu_t \left(\frac{\partial \langle U_i \rangle}{\partial x_j} + \frac{\partial \langle U_j \rangle}{\partial x_i} \right) - \frac{2}{3} \mu_t \frac{\partial \langle U_k \rangle}{\partial x_k} \partial_{ij}$$

is given in Section 5.1.2.1. The averaged equations then take the form

$$\frac{\partial}{\partial t} \left(\langle \varrho \rangle \langle U_i \rangle \right) + \frac{\partial}{\partial x_k} \left(\langle \varrho \rangle \langle U_k \rangle \langle U_i \rangle \right) = \frac{\partial}{\partial x_k} \left(\langle \tau_{t_{ik}} \rangle \right) - \frac{\partial \langle p \rangle}{\partial x_i} + \langle \varrho \rangle \langle g_i \rangle, \quad i = 1, 2, 3 \tag{388}$$

which is a transport equation for the first moment of the velocities. The accompanying problem of describing the turbulent viscosity was solved using the eddy viscosity hypothesis, $\mu_t = C_\mu \langle \varrho \rangle \left(\langle k \rangle^2 / \langle \varepsilon \rangle \right)$.

As well as the solution of Equation (388), the model requires the solution of the equations for the turbulence energy, $\langle k \rangle$ (Eq. 222), and its dissipation rate, $\langle \varepsilon \rangle$ (Eq. 225), thereby defining a turbulent time scale according to $\tau = \langle k \rangle / \langle \varepsilon \rangle$. A numerical solution that takes into account the special nature of the averaged conservation equations of momentum is demonstrated in Section 5.1.2.2.

For nonisothermal reacting systems this concept has to be expanded to the equations for the conservation of enthalpy (Eq. 134 or 240 a) and for the conservation of mass of the chemical species (Eq. 133 or 265 a). Averaging Equations (240 a) and (265 a) using the same assumptions as in Section 5.1.2.1 results in the equations

$$\frac{\partial}{\partial t} \left(\langle \varrho \rangle \langle h \rangle \right) + \frac{\partial}{\partial x_k} \left(\langle \varrho \rangle \langle U_k \rangle \langle h \rangle \right) = - \frac{\partial}{\partial x_k} \left(\langle J^h \rangle + \langle \varrho \rangle \langle h^f U_k^f \rangle \right) + \langle \varrho S_h \rangle \tag{389}$$

and

$$\frac{\partial}{\partial t} \left(\langle \varrho \rangle \langle m_i \rangle \right) + \frac{\partial}{\partial x_k} \left(\langle \varrho \rangle \langle U_k \rangle \langle m_i \rangle \right) = - \frac{\partial}{\partial x_k} \left(\langle J^{m_i} \rangle + \langle \varrho \rangle \langle m_i^f U_k^f \rangle \right) + \langle \varrho S_i \rangle \quad i = 1, \ldots, N-1 \tag{390}$$

The problem now lies in modeling the Reynolds fluxes $\langle \varrho \rangle \langle h^f U_j^f \rangle$ and $\langle \varrho \rangle \langle m_i^f U_j^f \rangle$, and also the expected values of the source term $\langle \varrho S_i \rangle$ and $\langle \varrho S_h \rangle$.

Modeling the Reynolds Fluxes. The problem of modeling the Reynolds fluxes $\langle \varrho \rangle \langle h^f U_j^f \rangle$ and $\langle \varrho \rangle \langle m_i^f U_j^f \rangle$ can be solved analogously to the modeling of the Reynolds stresses. As with the transport equations for

the Reynolds stresses (see Section 5.1.2.1, Eq. 221), the analogous transport equation for the Reynolds fluxes is not closed however. The same reasons as were discussed in Section 5.1.2.1 for the Reynolds stresses, lie behind the fact that transport equations for Reynolds fluxes are usually not solved in engineering but are modeled directly.

The direct modeling of Reynolds fluxes is based on the Boussinesq hypothesis and assumes that there is a similarity between the turbulent transport of momentum, and enthalpy, and mass. This principle has already been used in Sections 5.2.2.1 and 5.3.1.1 for deterministic systems. The turbulent flux of a scalar quantity is thus related to the gradients of the expected values of the scalar. The proportionality number is a turbulent transport coefficient which is obtained from the turbulent viscosity:

$$-\langle\varrho\rangle\langle\Phi^{\mathrm{f}}U_k^{\mathrm{f}}\rangle = \frac{\mu_t}{\sigma_\varphi}\frac{\partial\langle\Phi\rangle}{\partial x_k} \tag{391}$$

This expression has already been used in the equation for turbulence energy (Eq. 223); σ_φ is a turbulent Prandtl or Schmidt number, which relates the turbulent transfer of the scalar in question and the turbulent transfer of momentum. With the help of the eddy viscosity hypothesis, σ_φ can be interpreted as the ratio of the mixing length of the velocity fluctuations to the mixing length of the fluctuations in the scalar.

With this hypothesis and neglecting the proportion of laminar flux to turbulent flux, Equations (389) and (390) become

$$\frac{\partial}{\partial t}\left(\langle\varrho\rangle\langle h\rangle\right) + \frac{\partial}{\partial x_k}\left(\langle\varrho\rangle\langle U_k\rangle\langle h\rangle\right) = \frac{\partial}{\partial x_k}\frac{\mu_t}{\sigma_h}\frac{\partial\langle h\rangle}{\partial x_k} + \langle\varrho S_h\rangle \tag{392}$$

and

$$\frac{\partial}{\partial t}\left(\langle\varrho\rangle\langle m_i\rangle\right) + \frac{\partial}{\partial x_k}\left(\langle\varrho\rangle\langle U_k\rangle\langle m_i\rangle\right) = \frac{\partial}{\partial x_k}\frac{\mu_t}{\sigma_m}\frac{\partial\langle m_i\rangle}{\partial x_k} + \langle\varrho S_i\rangle \quad i = 1,\ldots,N-1 \tag{393}$$

The remaining problem is now the modeling of the terms $\langle\varrho S_h\rangle$ and $\langle\varrho S_i\rangle$. If the change in enthalpy due to mechanical work and radiation are neglected in Equation (392), then the term $\langle\varrho S_h\rangle$ vanishes. A modeling expression then only needs to be found for $\langle\varrho S_i\rangle$.

Modeling of the Mean Reaction Rates. The discussion of models for the mean reaction rates in turbulent (statistical) flows is of similar importance in engineering to the modeling of Reynolds stresses and Reynolds fluxes. Reviews are given in [126–128].

The reaction rates are generally a function of mass fractions and temperature (Eq. 137 a). The expected values $\langle\varrho S_i\rangle$ are obtained from Equation (109) since the averaging rule (Eq. 107) is applicable to functions. If changes in density are neglected, then Equation (137 a) gives

$$\langle S_i\rangle = \int_{T_u}^{T_0}\int_0^1 \cdots \int_0^1 P\left(m_l,\ldots,m_o,T\right) \cdot S_i\left(m_l,\ldots,m_o,T\right)\mathrm{d}m_l\ldots\mathrm{d}m_o\mathrm{d}T \tag{394}$$

In Equation (394) $P\left(m_l,\ldots,m_o,T\right)$ is a joint probability density function of temperature and the mass fractions of the chemical species that are contained in the expression for the reaction rates. $S_i\left(m_l,\ldots,m_o,T\right)$ are usually known functions and hence to determine $\langle S_i\rangle$, the joint probability density functions have to be established. The example shown in Figure 14 demonstrates that a priori determinations for this are difficult; there are intermittencies in the region at the jet boundary so that $P\left(m_l,\ldots,m_o,T\right)$ becomes "bimodal". The bimodal structure is not dominant further downstream and within the turbulent jet. An acceptable assumption for the form of $P\left(m_l,\ldots,m_o,T\right)$ is a multidimensional normal distribution, which must be suitably clipped due to the restricted domain of definition of temperature and mass fractions. Although this form does not cover the intermittencies at the boundary of the turbulent jet, it is a good approximation for the inside of the turbulent jet and further downstream. For many applications the deficiencies in representation by a multidimensional normal distribution is of no consequence because the mean reaction rates in the areas with strong intermittencies are mostly small compared with those inside the turbulent jet. If the approximation of $P\left(m_l,\ldots,m_o,T\right)$ by means of a multidimensional normal distribution is adequate, the form of $P\left(m_l,\ldots,m_o,T\right)$ must be established quantitatively. As in the one-dimensional case, this is achieved through the expected values

and the second moments. Consequently in order to determine $P(m_l, \ldots, m_o, T)$ calculation of $\langle m_l \rangle, \ldots, \langle m_o \rangle$, $\langle T \rangle$, $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$, $i, j = l, o$, and $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$, $i = l, o$ is necessary.

Equation (394) can be presented in the form

$$\langle S_i \rangle = S_i \left(\langle m_l \rangle, \ldots, \langle m_o \rangle, \langle T \rangle \right) \cdot F \left(\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle, \langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle, i, j = l, o \right) \tag{395}$$

in which $F(\langle m_i^{\mathrm{f}} m_i^{\mathrm{f}} \rangle, \langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle, i, j = l, o)$ are correction functions for reproducing the effect of turbulent fluctuations in temperature and mass fractions in the averaging procedure of weighting with the joint probability density functions.

These correction functions can be given as polynomials of the second moments $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$, $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$ (or more precisely of the turbulence intensities and correlation coefficients) and of the activation energies for the reactions in question [129], [130]. The coefficients of these polynomials can be obtained by applying the averaging procedure (Eq. 394) for a number of predefined values of $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$ $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$ and E_{a}; $\langle S_i \rangle$ is then represented as a function of these quantities by an empirical polynomial expression [129]. Thus integration of Equation (394) is unnecessary for the numerical solution of the resulting system of equations.

Calculation of the expected values $\langle S_i \rangle$ of the reaction rates from Equation (395) assumes knowledge of the second moments $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$, $i, j = 1, \ldots, N$ and $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$, $i = 1, \ldots, N$. In this case too closure of the equation system at the level of the expected values necessitates calculation of the second moments. The relevant transport equations can be derived by similar considerations and modeling assumptions used for deriving the equation for other second moments, see Section 5.1.2.1 [52], [126–128], [130].

The numerical work involved in calculating a statistical reacting nonisothermal flow is considerable, despite the relatively simple structure of the modeling components. In addition to the solution of the balances of momentum, direct modeling of Reynolds stresses requires the solution of two further equations for turbulence energy and the turbulent time scale. The Reynolds fluxes are modeled on the same basis as the Reynolds stresses, so that apart from the enthalpy balance (Eq. 392) and the transport equations for the mass of the chemical species (Eq. 393), no other equations originate from this part of the model. Modeling of the mean chemical reaction rates through the simple expression (Eq. 394) with presumed shape joint probability density functions requires the solution of equations for the variances and covariances $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$, $i, j = 1, \ldots, N$, and $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$, $i = 1, \ldots, N$. In addition the density has to be calculated from the averaged form of the thermal equation of state

$$\langle \varrho \rangle = \frac{\langle p \rangle}{R} \sum_{i=1}^{N} \frac{m_i}{\langle T \rangle \langle m_i \rangle + \langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle} \tag{396}$$

which also contains the second moments $\langle m_i^{\mathrm{f}} T^{\mathrm{f}} \rangle$, $i = 1, \ldots, N$. Equation (396) can easily be derived from the ideal gas equation. The temperature must be determined from the expected value of the enthalpy of the mixture through the definition of the enthalpy:

$$\langle h \rangle = \sum_{i=1}^{N} \langle m_i h_i \rangle = \sum_{i=1}^{N} \langle m_i \left(h_{0i} + \int c_{pi} \mathrm{d}T \right) \rangle \tag{397}$$

Equation (397) is not solved here for the temperature, for further details see [129–131].

Example: Combustion of Hydrogen in a Turbulent Diffusion Flame. The simulation of a turbulent nonisothermal reacting flow by the system of equations outlined above will be demonstrated briefly on an example. A circular jet of hydrogen with a Reynolds number of 12 000 relative to the nozzle outlet is injected into virtually stationary air and is burned in a horizontal combustion chamber. In the experiments described in [132–134], velocities, temperatures, and concentrations of stable chemical species and of OH radicals are measured downstream of the burner nozzle. The system is axially symmetrical.

The numerical solution contains the solutions of Equations (223), (225), (388), (392), (393) and also of the equations for the variances and covariances $\langle m_i^{\mathrm{f}} m_j^{\mathrm{f}} \rangle$, $i, j = 1, \ldots, N$ and $\langle m^{\mathrm{f}}{}_i T^{\mathrm{f}} \rangle$, $i = 1, \ldots, N$ in axially symmetrical formulation. The numerical procedure is the same as that outlined for the isothermal case in Section 5.1.2.3. The chemical reactions are described by 44 elementary reactions between

eleven chemical species: H_2, O_2, H_2O, N_2, H, O, OH, HO_2, H_2O_2 and also N and NO [135]. The reaction rates of each of these components consists of the summation of reaction rates of all the elementary reactions j in which the component in question i occurs, i.e., $r_i = \sum_{j=1}^{R} r_{ij}$. The rate coefficients of the reactions are expressed in the form $k_j = k_{0j}\, T^{\alpha j}\, e^{-E_a/RT}$ and are thus not freely selectable modeling parameters. For details of the reaction rate coefficients of the H_2–O_2 system see [130], [135]. The modeling parameters used are the coefficients $C_\mu = 0.09$, $C_{\varepsilon 1} = 1.4$, $C_{\varepsilon 2} = 0.925$, $\sigma_{k\,\mathrm{eff}} = 1.0$, $\sigma_{\varepsilon\,\mathrm{eff}} = 1.3$; the turbulent Prandtl numbers in Equations (392) and (393); and also the equations for the second moments which are defined as $\sigma_\varphi = 0.7$ [91].

Figure 56 shows the result of the simulations from the model compared with the measured results given in [132–134]. The correspondence between the measured results and the model is satisfactory. In particular the penetration of oxygen to the axis of the turbulent jet close to the nozzle is well predicted. Along with the calculation of the OH radical concentrations, this is a particular potential of the model that is based on an elementary reaction scheme for describing the chemical reactions for the H_2–O_2–N_2 system.

Another aspect of the model presented here can be shown with the results of sensitivity analysis. The first-order sensitivity coefficients are calculated as described in Section 5.3.3.1.2. Due to the elliptical form of the conservation equations and the difference form which follows from this (see Section 5.1.2.2, Equation 232), the resulting system of equations has a block pentadiagonal structure and is solved by an ADI procedure.

Figure 57 A shows the results in the form of the gradients of the axial velocity $\langle U \rangle$ with respect to the model parameter C_μ, which controls the turbulent viscosity. The sensitivity coefficients are given in relative form $(\partial\langle U\rangle/\langle U\rangle)/(\partial\langle C_\mu\rangle/C_\mu)$ as in Equations (29) and (30). The sensitivity with respect to a change in the parameter C_μ is particularly large in areas with high velocity gradients. The sensitivity analysis thus reflects the physical background and the limitations of the model used: since the turbulent stresses are described by an approach which relates them to the gradients of the mean velocities, large sensitivities inevitably result in response to changes in the apparent turbulent viscosity in areas with high velocity gradients.

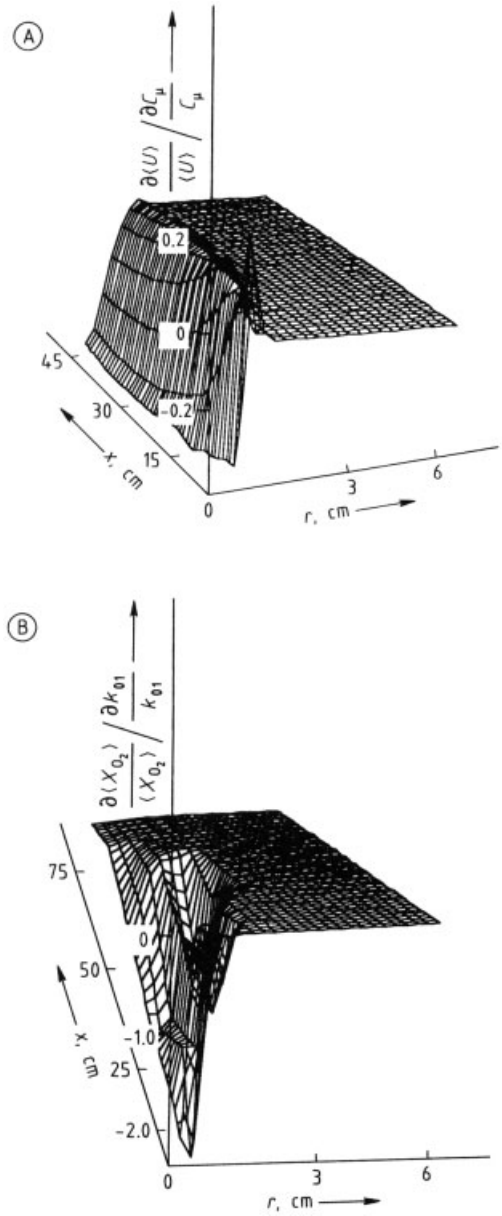

Figure 57. Relative sensitivity coefficient for a H_2–O_2 diffusion flame
A) Coefficients for the axial velocity $\langle U \rangle$ with respect to the modeling parameter C_μ; B) Coefficients for the mole fraction of oxygen $\langle X_{O_2} \rangle$ with respect to the rate coefficient k_{01} x = axial distance; r = radial distance.

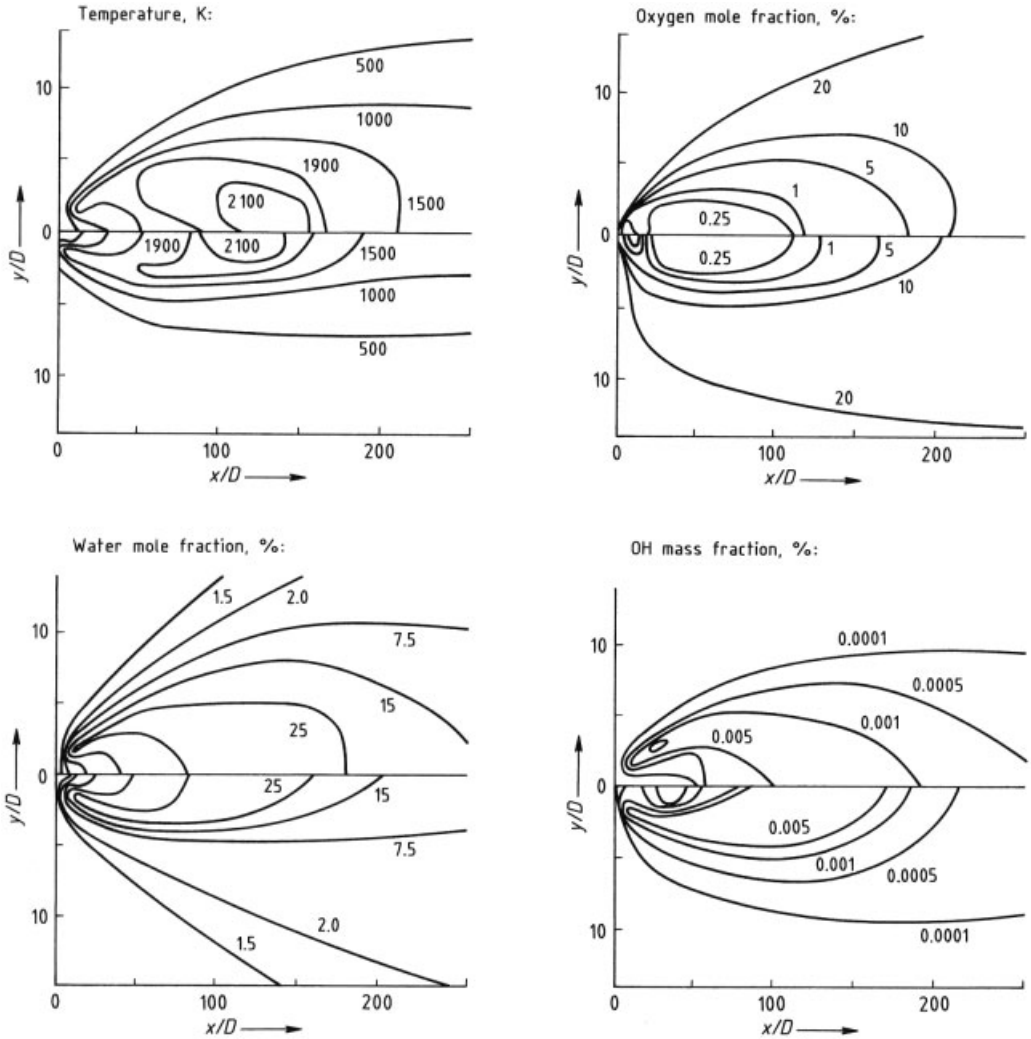

Figure 56. Measured (upper section) and calculated (lower section) profiles for temperature, mole fraction of oxygen and water, and mass fraction of OH radicals for a turbulent hydrogen – air diffusion flame [130], [131]
x = axial distance; y = radial distance; D = nozzle diameter

Figure 57 B shows the gradients of the molar fractions for oxygen with respect to the rate coefficient for the reaction $O_2 + H \rightarrow OH + O$, this is one of the most important oxygen-consuming reactions. Here to, the sensitivity coefficients are given in relative form $(\partial\langle X_{O_2}\rangle/\langle X_{O_2}\rangle)/(\partial k_{01}/k_{01})$. Sensitivity is only obvious in the main reaction zones of the turbulent hydrogen flame. Thus, the molar fraction of oxygen only reacts sensitively to the change in these rate coefficients in the main reaction zone of the flame.

The result of averaging the reaction rates with Equations (394) and (395) can be interpreted as the change in the rate coefficients due to the correction functions. The unsatisfactory representation of the joint probability density function $P(m_l, \ldots, m_o, T)$ by a multidimensional normal distribution used to calculate these correction functions is therefore no longer of importance at the edge of the jet for the prediction of the oxygen molar fraction. This is due to the low reaction rates for the consumption of oxygen in this region.

For further discussion of sensitivity analysis, especially for discussion of the possibility of reduction of the chemical models, see [131].

6. References

1. G. Klaus: *Wörterbuch der Kybernetik,* Fischer Verlag, Frankfurt 1969.
2. A. I. Bojarinow, W. W. Kafarow: *Optimierungsmethoden in der chemischen Technologie,* Verlag Chemie, Weinheim 1972.
3. D. M. Himmelblau, K. B. Bischoff: *Process Analysis and Simulation: Deterministic Systems,* J. Wiley and Sons, New York 1968.
4. D. M. Himmelblau: *Process Analysis by Statistical Methods,* J. Wiley and Sons, New York 1970.
5. U. Hoffmann, H. Hofmann: *Einführung in die Optimierung mit Beispielen aus dem Chemie-Ingenieur-Wesen,* Verlag Chemie, Weinheim 1971.
6. D. M. Himmelblau in A. Bisio, R. Kabel (eds.): *Scaleup of Chemical Processes,* J. Wiley and Sons, New York 1985.
7. T. E. Edgar, D. M. Himmelblau: *Optimization of Chemical Processes,* McGraw-Hill, New York 1988.
8. F. W. Ramirez: *Computational Methods for Process Simulation,* Butterworth Publishers, Stoneham 1989.
9. W. F. Luyben: *Process Modeling, Simulation and Control for Chemical Engineers,* McGraw-Hill, New York 1990.
10. *Ullmann,* 4th ed., **4,** 451.
11. G. Emig, *Ber. Bunsenges. Phys. Chem.* **90** (1986) 986.
12. H. Hofmann: "Future Trends in Chemical Engineering Modelling," *Proceedings of the 18th EFCE Congress,* Apr. 26 – 30, 1987, p. 579.
13. W. W. Kafarow: *Kybernetische Methoden in der Chemie und chemischen Technologie,* Verlag Chemie, Weinheim 1971.
14. L. A. Gould: *Chemical Process Control: Theory and Applications,* Addison-Wesley Publishing Company, Reading 1969.
15. P. S. Buckley: *Techniques of Process Control,* Robert E. Krieger Publishing Company, Huntington 1979.
16. P. V. Danckwerts, *Chem. Eng. Sci.* **2** (1953) 1.
17. A. D. Randolph, *Can. J. Chem. Eng.* **42** (1964) 280.
18. E. E. O'Brien in P. A. Libby, F. A. Williams (eds.): *Turbulent Reacting Flows,* Springer Verlag, Berlin-Heidelberg 1980, p. 191.
19. S. B. Pope, *Prog. Energy Combust. Sci.* **11** (1985) 119.
20. G. Emig, U. Hoffmann, H. Hofmann: *DECHEMA-Kurs: Planung und Auswertung von Versuchen zur Erstellung mathematischer Modelle,* **parts I and II,** DECHEMA, Frankfurt-Main 1974.
21. E. Kreyszig: *Statistische Methoden und ihre Anwendung,* Verlag Vandenhoeck und Ruprecht, Göttingen 1972.
22. M. Fisz: *Wahrscheinlichkeitsrechnung und mathematische Statistik,* VEB Deutscher Verlag der Wissenschaften, Berlin 1980.
23. A. Lindner: *Statistische Methoden,* Birkhäuser, Stuttgart 1964.
24. M. Baerns, H. Hofmann, A. Renken: *Chemische Reaktionstechnik, Lehrbuch der Technischen Chemie 1,* Thieme Verlag, Stuttgart 1987.
25. VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen: "Berechnungsblätter für den Wärmeübergang," *VDI-Wärmeatlas,* VDI-Verlag, Düsseldorf 1984.
26. IBM – Scientific Subroutine Package – System 360, IBM Corporation, White Plains 1972.
27. NAG Fortran Library-Mark 8, Numerical Algorithms Group Inc., Downers Grove 1981.
28. IMSL Library Reference Manual, IMSL Inc., Houston 1980.
29. W. H. Press, B. B. Flannery, S. A. Teukolsky, W. T. Vetterling: *Numerical Recipes. The Art of Scientific Computing (Fortran Version),* Cambridge University Press, Cambridge 1989 (also available in Pascal and C).
30. W. H. Press, B. B. Flannery, S. A. Teukolsky, W. T. Vetterling: *Numerical Recipes. Example Book (Fortran),* Cambridge University Press, Cambridge 1985 (also available in Pascal and C).
31. G. Engeln-Müllges, F. Reutter: *Numerische Mathematik für Ingenieure,* B.-I. Wissenschaftsverlag, Mannheim 1987.
32. G. Engeln-Müllges, F. Reutter: *Formelsammlung zur Numerischen Mathematik mit Standard-Fortran 77-Programmen,* B.-I. Wissenschaftsverlag, Mannheim 1988.
33. J. Hildenrath et al.: "OMNITAB, A Computer Program for Statistical and Numerical Analysis," *Handbook 101,* Nat. Bur. of Standards, U.S. Government Printing Office, Washington, D.C., 1968.

34. R. Hooke, T. A. Jeeves, *J. Assoc. Compt. Mach.* **8** (1961) 212.
35. N. Spendley, G. R. Hext, F. R. Himsworth, *Technometrics* **4** (1962) 441.
36. J. A. Nelder, R. Mead, *Comput. J.* **7** (1965) 308.
37. H. H. Rosenbrock, C. Storey: *Computational Techniques for Chemical Engineers,* Pergamon Press, Oxford 1966.
38. C. S. Beightler, D. T. Phillips, D. G. Wilde: *Foundations of Optimization,* Prentice Hall, Englewood Cliffs, N.Y. 1979.
39. P. E. Gill, W. Murray, M. H. Wright: *Practical Optimization,* Academic Press, New York 1981.
40. D. W. Marquardt, *J. Soc. Ind. Appl. Math.* **11** (1963) 431.
41. H. O. Hartley, A. Booker, *Ann. Math. Stat.* **36** (1965) 638.
42. E. J. Williams, *J. Royal Stat. Soc. B* **24** (1962) 125.
43. J. J. More in G. A. Watson (ed.): *Numerical Analysis, Lecture Notes in Mathematics,* **vol. 630,** Springer Verlag, Berlin 1977, p. 105.
44. Z. E. Beisinger, S. Bell: H2 SAND MIN, Sandia Corporation.
45. I. N. Bronstein, K. H. Semendjajew: *Taschenbuch der Mathematik,* Verlag Harry Deutsch, Thun 1983.
46. O. Levenspiel: *Chemical Reaction Engineering,* J. Wiley and Sons, New York 1972.
47. O. Levenspiel, W. K. Smith, *Chem. Eng. Sci.* **6** (1975) 227.
48. J. O. Hirschfelder, C. F. Curtiss, R. B. Bird: *Molecular Theory of Gases and Liquids,* J. Wiley and Sons, New York 1954.
49. J. C. Rotta: *Turbulente Strömungen,* B. G. Teubner, Stuttgart 1972.
50. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley and Sons, New York 1960.
51. F. A. Williams: *Combustion Theory,* The Benjamin/Cummings Publishing Company, Menlo Park, Calif. 1985.
52. P. A. Libby, F. A. Wiliams in P. A. Libby, F. A. Williams (eds.): *Turbulent Reacting Flows,* Springer Verlag, Berlin – Heidelberg 1980, p. 1.
53. P. Givi, *Prog. Energy Combust. Sci.* **15** (1989) 121.
54. M. Lesieur: *Turbulence in Fluids,* Martinus Nijhoff Publishers, Dordrecht 1987.
55. T. S. Lundgreen, *Phys. Fluids* **12** (1969) 485.
56. S. B. Pope, *Combust. Flame* **27** (1976) 294.
57. J. Janicka, W. Kolbe, W. Kollmann in C. T. Crowe, W. L. Grosshandler (eds.): *Proc. of the 1978 Heat Transf. Fluid Mech. Inst.,* Standford University Press, Stanford 1978, p. 296.
58. S. B. Pope, *Philos. Trans. R. Soc. Lond. Ser. A* **291** (1979) 529.
59. M. Jischa: *Konvektiver Impuls-, Wärme- und Stoffaustausch,* Vieweg und Sohn, Braunschweig 1982.
60. H. Schlichting: *Grenzschichttheorie,* Verlag G. Braun, Karlsruhe 1982.
61. N. Peters in E. S. Oran, J. P. Boris (eds.): *Approaches to Combustion Modeling,*
62. M. J. Lighthill: *Introduction to Fourier Analysis and Generalized Functions,* Cambridge University Press, Cambridge 1958.
63. H. M. Hulburt, S. Katz, *Chem. Eng. Sci.* **19** (1964) 555.
64. E. Kamke: *Differentialgleichungen, Lösungsmethoden und Lösungen II,* B. G. Teubner Verlag, Stuttgart 1979.
65. P. Langevin, *C. R. Acad. Sci. Paris* **146** (1908) 530.
66. S. B. Pope, *Phys. Fluids* **24** (1981) 588.
67. L. Arnold: *Stochastic Differential Equations: Theory and Applications,* J. Wiley and Sons, New York 1974.
68. W. Jost: *Diffusion in Solids, Liquids and Gases,* Academic Press, New York 1952.
69. O. Levenspiel, K. B. Bischoff: *Adv. Chem. Eng.* **4** (1963) 95.
70. J.-Y. Chen, W. Kollmann: *Twenty-Second Symposium (International) on Combustion,* The Combustion Institute, Pittsburgh 1989, p. 645.
71. J.-Y. Chen, W. Kollmann, *Combust. Flame* **79** (1990) 75.
72. J. L. Lumley, *Adv. Appl. Mech.* **18** (1978) 123.
73. B. E. Launder, C. J. Reece, W. Rodi, *J. Fluid. Mech.* **68** (1975) 573.
74. J. C. Rotta, *Z. Phys.* **129** (1951) 547.
75. C. Dopazo, *Phys. Fluids* **18** (1975) 397.
76. J. Janicka, W. Kolbe, W. Kollman, *J. Non-Equilib. Thermodyn.* **4** (1978) 47.
77. C. Dopazo, *Phys. Fluids* **22** (1979) 20.
78. J. Janicka, W. Kollman in L. J. S. Bradbury et al. (eds.): *Turbulent Shear Flows 4,* Selected Papers from the Fourth International Symposium on Turbulent Shear Flows, Springer Verlag, Berlin – Heidelberg 1985, p. 73.
79. S. B. Pope, *Combust. Sci. Technol.* **25** (1981) 159.

80. T. V. Nguyen, S. B. Pope, *Combust. Sci. Technol.* **42** (1984) 13.
81. A. Favre in Society for Industrial and Applied Mathematics (ed.): *Problems of Hydrodynamics and Continuous Mechanics,* Philadelphia 1969, p. 231.
82. G. Stahl, J. Warnatz, B. Rogg in AIAA (ed.): *Progr. Aeronaut. Astronaut.,* **vol. 113,** Washington 1988, p. 195.
83. P. Bradshaw in P. Bradshaw (ed.): *Turbulence,* Springer Verlag, Berlin – Heidelberg 1978, p. 1.
84. B. E. Launder, D. B. Spalding: *Lectures in Mathematical Models of Turbulence,* Academic Press, London – New York 1972.
85. K. K. Kuo: *Principles of Combustion,* J. Wiley and Sons, New York 1986.
86. J. O. Hinze: *Turbulence,* McGraw-Hill, New York 1975.
87. J. Boussinesq, *Mém. prés. Acad. Sci.* **XXIII, 46,** Paris1877.
88. L. Prandtl, *Z. Angew. Math. Mech.* **22** (1942) 241.
89. L. Prandtl, *Nachr. Akad. Wiss. Göttingen, Math. Phys. Kl. 2* (1945) 6 – 19.
90. J. Janicka: *Twenty-First Symposium (International) on Combustion,* The Combustion Institute, Pittsburgh 1988, p. 1409.
91. S. M. Correa, W. Shyy, *Prog. Energy Combust. Sci.* **13** (1987) 249.
92. H. Tennekes, J. L. Lumley: *A First Course in Turbulence,* MIT-Press, Boston 1972.
93. W. F. Ames: *Nonlinear Partial Differential Equations in Engineering,* Academic Press, New York – London 1965.
94. A. D. Gosman et al.: *Heat and Mass Transfer in Recirculating Flows,* Academic Press, London 1969.
95. M. D. Smooke, A. A. Turnbull, R. E. Mitchell, D. E. Keyes in C.-M. Brauner, C. Schmidt-Lainé (eds.): *Mathematical Modeling in Combustion and Related Topics,* Martinus Nijhoff Publishers, Dordrecht 1988, p. 261.
96. D. B. Spalding: *Basic Equations of Fluid Mechanics and Heat and Mass Transfer,* Imperial College London, Mechanical Engineering Department, Report HTS/76/6, London 1976.
97. S. V. Patankar: *Numerical Heat Transfer and Fluid Flow,* McGraw-Hill, New York 1980.
98. F. H. Harlow, J. E. Welch, *Phys. Fluids* **8** (1965) 2182.
99. A. O. Demurren, *Comput. Fluids* **13** (1985) 411.
100. P. J. Roach: *Computational Fluid Dynamics,* Hermosa Publishers, Albuquerque 1976.
101. E. S. Oran, J. P. Boris: *Numerical Simulation of Reactive Flow,* Elsevier Science Publishing Company, New York 1987.
102. C. Hirsch: *Numerical Computation of Internal and External Flows,* vol. 1, J. Wiley and Sons, New York 1988.
103. C. Hirsch: *Numerical Computation of Internal and External Flows,* vol. 2, J. Wiley and Sons, New York 1990.
104. A. J. Baker: *Finite Element Computational Fluid Mechanics,* McGraw-Hill, New York 1985.
105. J. M. Beer, N. A. Chigier: *Combustion Aerodynamics,* Applied Science Publishers Ltd., London 1972.
106. L. Horvarth, *Proc. R. Soc. London Ser. A* **164** (1938) 547.
107. E. R. G. Eckert, R. M. Drake: *Analysis of Heat and Mass Transfer,* McGraw-Hill, New York 1972.
108. M. N. Özisik: *Basic Heat Transfer,* McGraw-Hill, New York 1985.
109. H. Brauer: *Stoffaustausch einschließlich chemischer Reaktionen,* Verlag Sauerländer, Aarau 1971.
110. P. Trambouze in A. E. Rodriguez, J. M. Calo, N. H. S. Sweed (eds.): *Multiphase Chemical Reactors, Fundamentals* **(vol. I),** Sijthoff and Noordhoff, Alphen 1981.
111. J. A. van der Vusse, *Chem. Eng. Sci.* **17** (1962) 507.
112. L. Cloutier, *Can. J. Chem. Eng.* **37** (1959) 105.
113. P. Gray, S. K. Scott: *Chemical Oscillations and Instabilities: Nonlinear Chemical Kinetics,* Clarendon Press, Oxford 1990.
114. G. Iooss, D. D. Joseph: *Elementary Stability and Bifurcation Theory,* Springer Verlag, New York 1980.
115. J. H. Merkin, D. J. Needham, S. K. Scott, *Proc. R. Soc. London Ser. A* **406** (1986) 299.
116. B. F. Gray, M. J. Roberts, *Proc. R. Soc. London Ser. A* **416** (1988) 391.
117. P. M. Frank: *Introduction to System Sensitivity Theory,* Academic Press, New York 1978.
118. T. P. Coffee, J. M. Heimerl, *Combust. Flame* **50** (1983) 323.
119. D. E. Clough, W. F. Ramirez, *AIChE J.* **22** (1976) no. 4, 1097.
120. J. C. P. Sheel, C. M. Crowe, *Can. J. Chem. Eng.* **47** (1969) 183.
121. R. A. Greenkorn: *Flow Phenomena in Porous Media,* Marcel Dekker, New York 1983.

122. E. Kreyszig: *Advanced Engineering Mathematics,* J. Wiley and Sons, New York 1988.
123. G. F. Fromment, K. B. Bishoff: *Chemical Reactor Analysis and Design,* J. Wiley and Sons, New York 1990.
124. P. B. Weisz, J. S. Hicks, *Chem. Eng. Sci.* **17** (1962) 265.
125. J. Ganoulis, F. Durst: "Finite Elements in Water Resources," *Proceedings VI Int. Conf.,* Lisbon 1986, p. 655.
126. S. N. B. Murthy (ed.): *Turbulent Mixing in Nonreactive and Reactive Flows,* Plenum Press, New York – London 1975.
127. R. W. Bilger in P. A. Libby, F. A. Williams (eds.): *Turbulent Reacting Flows,* Springer Verlag, Berlin – Heidelberg 1980, p. 65.
128. R. Borghi, *Prog. Energy Combust. Sci.* **14** (1988) 245.
129. H. Bockhorn in C.-M. Brauner, C. Schmidt-Lainé (eds.): *Mathematical Modeling in Combustion and Related Topics,* Martinus Nijhoff Publishers, Dordrecht 1988, p. 411.
130. H. Bockhorn: *Twenty-Second Symposium (International) on Combustion,* The Combustion Institute, Pittsburgh 1988, p. 665.
131. H. Bockhorn: *Twenty-Third Symposium (International) on Combustion,* The Combustion Institute, Pittsburgh.
132. M. C. Drake et al.: *Twentieth Symposium (International) on Combustion,* The Combustion Institute, Pittsburgh 1984, p. 327.
133. M. C. Drake, R. W. Pitz, M. Lapp: *AIAA Pap.* **84-0544** (1984).
134. G. M. Faeth, G. S. Samuelson, *Prog. Energy Combust. Sci.* **12** (1986) 305.
135. J. Warnatz in W. C. Gardiner, Jr. (ed.): *Combustion Chemistry,* Springer Verlag, New York 1984, p. 196.

mathematics → mathematics in chemical engineering

Scale-Up in Chemical Engineering

MARKO ZLOKARNIK, Graz, Austria

1. Introduction 2
2. Dimensional Analysis 3
2.1. The Fundamental Principle 3
2.2. What is a Dimension? 3
2.3. What is a Physical Quantity? 3
2.4. Base and Derived Quantities, Dimensional Constants 3
2.5. Dimensional Systems 4
2.6. Dimensional Homogeneity of a Physical Content 4
2.7. The Pi Theorem 6
3. The Determination of a Pi Set by Matrix Transformation 6
3.1. The Establishment of a Relevance List of a Problem 6
3.2. Determination of the Characteristic Geometric Parameter 6
3.3. Constructing and Solving of the Dimensional Matrix 7
3.4. Determination of the Process Characteristics 7
4. Fundamentals of the Theory of Models and Scale-Up 8
4.1. Theory of Models 8
4.2. Model Experiments and Scale-Up . 8
5. Further Procedures to Establish a Relevance List 9
5.1. Consideration of the Acceleration due to Gravity *g* 9
5.2. Introduction of Intermediate Quantities 10
5.3. Scale-up Procedure at Unavailability of Model Material Systems 11
5.4. Partial Similarity 12
6. Short Summary of the Essentials of the Dimensional Analysis and Scale-Up 13
6.1. Advantages of Dimensional Analysis 13
6.2. Area of Applicability of Dimensional Analysis 14
6.3. Experimental Methods for Scale-Up 14
7. Treatment of Variable Physical Properties by Dimensional Analysis 15
7.1. Dimensionless Representation of the Material Function $\mu(T)$ 15
7.2. Dimensionless Representation of the Material Function $\rho(T)$ 16
7.3. Pi Set for Variable Physical Properties 16
7.4. Material Function in Non-Newtonian Liquids 18
7.4.1. Pseudoplastic Flow Behavior 18
7.4.2. Viscoelastic Flow Behavior 19
7.5. Pi Space in Processes with Non-Newtonian Fluids 21
8. Optimization of Process Conditions by Combining Process Characteristics 21
9. References 24

List of important, named pi numbers:

Name	Symbol	Pi number	Remarks
A Mechanical process engineering			
Archimedes	*Ar*	$\frac{g\Delta\varrho l^3}{\varrho\nu^2}$	$\equiv(\Delta\rho/\rho)\,Ga$
Bond	*Bd*	$\frac{\varrho g l^2}{\sigma}$	$\equiv We/Fr$
Brinkman	*Br*	$\frac{v^2\mu}{Jk\Delta T}$	J – Joule's heat equivalent
Deborah	*De*	λn	λ – relaxation time
Euler	*Eu*	$\frac{\Delta p}{\varrho v^2}$	
Froude	*Fr*	$\frac{v^2}{lg}$	
Froude extended	*Fr*∗	$\frac{v^2\varrho}{lg\Delta\varrho}$	$\equiv Fr\,(\rho/\Delta\rho)$
Galilei	*Ga*	$\frac{gl^3}{\nu^2}$	$\equiv Re^2/Fr$
Laplace	*La*	$\frac{\Delta p d}{\sigma}$	$\equiv Eu\,We$
Mach	*Ma*	v/v_s	v_s – velocity of sound
Newton	*Ne*	$\frac{F}{\varrho v^2 l^2}$	F – force
	Ne	$\frac{P}{\varrho v^3 l^2}$	P – power
Ohnesorge	*Oh*	$\frac{\mu}{(\varrho\sigma l)^{1/2}}$	$\equiv We^{1/2}/Re$
Reynolds	*Re*	$\frac{vl}{\nu}$	$\nu\equiv\mu/\rho$
Stokes	*Sto*	$\frac{\varrho_p d_p^2 v}{l\mu}$	$\equiv Re_p\,(d_p/D)$
Strouhal	*Sr*	lf/v	f – frequency
Viscosity ratio	*Vis*	μ_w/μ	w – wall
Weber	*We*	$\frac{\varrho v^2 l}{\sigma}$	
Weissenberg	*Wi*	N_1/τ	N_1 – 1. normal sress

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_03

Name	Symbol	Pi number	Remarks
B Thermal process engineering (heat transfer)			
Fourier	Fo	at/l^2	a – thermal diffusivity
Grashof	Gr	$\frac{\beta \Delta T g l^3}{\nu^2}$	$\equiv \beta \Delta T \mathrm{Ga}$
Nusselt	Nu	hl/k	h – heat transfer coefficient k – heat conductivity
Péclet	Pe	$\mathrm{v}l/a$	$\equiv Re\,Pr$
Prandtl	Pr	ν/a	$a \equiv k/(\rho\, C_p)$
Rayleigh	Ra	$\frac{\beta \Delta T g l^3}{a\,\nu}$	$\equiv Gr\,Pr$
Stanton	St	$\frac{h}{\mathrm{v}\,\varrho\, C_p}$	$\equiv \frac{Nu}{Re\,Pr}$
C Thermal process engineering (mass transfer)			
Bodenstein	Bo	$\mathrm{v}l/D_{\mathrm{ax}}$	D_{ax} – dispersion coefficient
Lewis	Le	a/D	$\equiv Sc/Pr$
Schmidt	Sc	ν/D	D – mass diffusivity
Sherwood	Sh	kl/D	k – mass transfer coefficient
Stanton	St	k/v	$\equiv \frac{Sh}{Re\,Sc}$
D Chemical reaction engineering			
Arrhenius	Arr	$\frac{E}{RT}$	E – energy of activation
Damköhler	Da	$\frac{c\,\Delta H_{\mathrm{R}}}{\varrho\, C_{\mathrm{p}}\, T_0}$	
	Da_{I}	$k_1\,\tau$	k_1 – reaction rate constant τ – mean residence time
	Da_{II}	$k_1\,L^2/D$	$\equiv Da_{\mathrm{I}}\,Bo$
	Da_{III}	$k_1\,\tau\left(\frac{c\,\Delta H_{\mathrm{R}}}{\varrho\, C_{\mathrm{p}}\, T_0}\right)$	$\equiv Da_{\mathrm{I}}\,Da$
	Da_{IV}	$\frac{k_1\,c\,\Delta H_{\mathrm{R}}\,l^2}{k\,T_0}$	$\equiv Da_{\mathrm{I}}\,Pe\,Da$
Hatta	Hat_1	$(k_1 D)^{1/2}/k_{\mathrm{L}}$	1st order reaction
	Hat_2	$(k_2\,\mathrm{c}_2\,D)^{1/2}/k_{\mathrm{L}}$	2nd order reaction

1. Introduction

A chemical engineer is generally concerned with the industrial implementation of processes in which chemical or microbiological conversion of material takes place in conjunction with the transfer of mass, heat, and momentum. These processes are scale dependent, that is, they behave differently on a small scale (in laboratories or pilot plants) and a large scale (in production). They include heterogeneous chemical reactions and most unit operations. Understandably, chemical engineers have always wanted to find ways of simulating these processes in models to gain insights that will assist them in designing new industrial plants. Occasionally, they are faced with the same problem for another reason: an industrial facility already exists but does not function properly, if at all, and suitable measures have to be carried out to discover the cause of the difficulties and provide a solution

Irrespective of whether the model involved represents a "scale-up" or a "scale-down", important questions always apply:

1) How small can the model be? Is one model sufficient or should tests be carried out in models of different sizes?
2) When must or when can physical properties differ? When must the measurements be carried out on the model with the original system of materials?
3) Which rules govern the adaptation of the process parameters in the model measurements to those of the full-scale plant?
4) Is it possible to achieve complete similarity between the processes in the model and those in its full-scale counterpart? If not: how should one proceed?

These questions touch on the fundamentals of the theory of models, which are based on dimensional analysis. Although they have been used in the field of fluid dynamics and heat transfer for more than a century – cars, aircrafts, vessels, and heat exchangers were scaled up according to these principles – these methods have gained only a modest acceptance in chemical engineering. University graduates are usually not skilled enough to deal with such problems at all. On the other hand, there is no motivation for this type of research at universities, since as a rule, they are not confronted with scale-up tasks and are not equipped with the necessary apparatus on the bench-scale. All this gives a totally wrong impression that these methods are, at most, of marginal importance in practical chemical engineering, otherwise they would have been taught and dealt with in greater depth.

2. Dimensional Analysis

2.1. The Fundamental Principle

The dimensional analysis is based upon the recognition that a mathematical formulation of a physico-technological problem can be of general validity only when the process equation is dimensionally homogeneous.

2.2. What is a Dimension?

A dimension is a purely qualitative description of a perception of a physical entity or a natural appearance. A length can be experienced as a height, a depth, a breadth. A mass presents itself as a light or heavy body, the time as a short moment or a long period. The dimension of a length is L, the dimension of a mass is M, and so on.

2.3. What is a Physical Quantity?

Unlike a dimension, a physical quantity represents a quantitative description of a physical quality (e.g., a mass of 5 kg). A physical quantity consists of a measuring unit and a numerical value. The measuring unit of length can be a meter, a foot, a cubit, a yardstick, a nautical mile, a light-year, and so on. Measuring units of energy are, for example, joules, calories, or electronvolts. It is therefore necessary to establish the measuring units in an appropriate measuring system.

2.4. Base and Derived Quantities, Dimensional Constants

A distinction is made between base or primary quantities and the secondary quantities which are derived from them (derived quantities). Base quantities are based on standards and are quantified by comparison with them. According to the Système International d' Unités (SI), mass is classified as a primary quantity instead of force, which was used until the 1960s.

Secondary quantities are derived from primary ones according to physical laws, e.g., velocity = length/time: $[\nu] = \mathrm{L/T}$. Its coherent measuring unit is m/s. Coherence of the measuring units means that the secondary quantities have to have only such measuring units which correspond with per definitionem fixed primary ones and therefore present themselves as power products of these primary quantities. Giving the velocity in mph (miles per hour) would contradict this.

If a secondary unit has been established by a physical law, it can happen that it contradicts another one. Example: According to Newton's second law of motion, the force F is expressed as a product of mass m and acceleration a: $F = m \cdot a$, having the measuring unit of $\mathrm{kg\,m/s^2} \equiv \mathrm{N}$. According to Newton's law of gravitation, the force is defined by $F \propto m_1 \cdot m_2/r^2$, thus leading to a completely different measuring unit ($\mathrm{kg^2/m^2}$). To remedy this, the gravitational constant g – a dimensional constant – had to be introduced to ensure the dimensional homogeneity of the latter equation: $F = g\, m_1 \cdot m_2/r^2$.

Another example is the universal gas constant R, the introduction of which ensures that in the perfect gas equation of state $p \cdot V = n \cdot R \cdot T$ the secondary unit for work $W = p \cdot V$ $[\mathrm{M\,L^2\,T^{-2}}]$ is not offended.

Another class of derived quantities is represented by the coefficients in transfer equations for momentum, mass, and heat. These coefficients were also established by the respective physical laws – therefore they are often called "defined quantities" – and can only be determined via measurement of their constituents.

2.5. Dimensional Systems

A dimensional system consists of all the primary and secondary dimensions and corresponding measuring units. The currently used International System of Dimensions (Système International d'Unités; SI) is based on seven base dimensions. They are presented in Table 1 together with their corresponding base units. For some of them a few explanatory remarks may be necessary:

Table 1. Base quantities, their dimensions and units according to SI

Base quantity	Base dimension	SI base unit	
Length	L	m	meter
Mass	M	kg	kilogram
Time	T	s	second
Thermodynamic temperature	Θ	K	kelvin
Amount of substance	N	mol	mole
Electric current	I	A	ampere
Luminous intensity	J	cd	candela

Temperature expresses the thermal level of a system, not its energetic contents. (A fivefold mass of a matter has a fivefold thermal energy at the same temperature!) The thermal energy of a system can indeed be converted into the mechanical energy (by the Joule's mechanical

equivalent of heat, J). Moles are amount of matter and must not be confused with the quantity of mass. The molecules react as individual entities regardless of their mass: 1 mol hydrogen (molar mass: 2 g/mol) reacts with 1 mol chlorine (molar mass: 71 g/mol) to produce 2 mol hydrochloric acid, HCl.

Table 2 shows the most important secondary quantities and their dimensions, Table 3 refers to some very frequently used secondary units which have been named after famous researchers.

Table 2. Commonly used secondary quantities and their respective dimensions according to the SI units currently used in mechanical and thermal problems

Physical entity	Dimension
Area	L^2
Volume	L^3
Angular velocity, Shear rate, Frequency	T^{-1}
Velocity	$L\,T^{-1}$
Acceleration	$L\,T^{-2}$
Kinematic viscosity, Diffusivity, Thermal conductivity	$L^2\,T^{-1}$
Density	$M\,L^{-3}$
Surface tension	$M\,T^{-2}$
Dynamic viscosity	$M\,L^{-1}\,T^{-1}$
Momentum	$M\,L\,T^{-1}$
Force	$M\,L\,T^{-2}$
Pressure, tension	$M\,L^{-1}\,T^{-2}$
Angular momentum	$M\,L^2\,T^{-1}$
Mechanical energy, work, torque	$M\,L^2\,T^{-2}$
Power	$M\,L^2\,T^{-3}$
Specific heat	$L^2\,T^{-2}\,\Theta^{-1}$
Heat conductivity	$M\,L\,T^{-3}\,\Theta^{-1}$
Heat transfer coefficient	$M\,T^{-3}\,\Theta^{-1}$

Table 3. Important secondary measuring units in mechanical engineering, named after famous researchers

Secondary quantity	Dimension	Measuring unit		Name
Force	$M\,L\,T^{-2}$	$kg\,m\,s^{-2}$	$\equiv$ N	Newton
Pressure	$M\,L^{-1}\,T^{-2}$	$kg\,m^{-1}\,s^{-2}$	$\equiv$ Pa	Pascal
Energy	$M\,L^2\,T^{-2}$	$kg\,m^2\,s^{-2}$	$\equiv$ J	Joule
Power	$M\,L^2\,T^{-3}$	$kg\,m^2\,s^{-3}$	$\equiv$ W	Watt

2.6. Dimensional Homogeneity of a Physical Content

The aim of the dimensional analysis is to check whether the physical content of the case under examination can be formulated in a dimensionally homogeneous manner. The procedure necessary to accomplish this consists of two parts:

1) All physical parameters necessary to describe the problem are listed. This "relevance list" of the problem consists of the quantity in question and all the parameters that influence it. In each case only one target quantity must be considered; it is the only dependent variable. On the other hand, all the influencing parameters must be primarily independent of each other.
2) The dimensional homogeneity of the physical content is checked by transferring it in a dimensionless form. (Note: A physical content which can be transformed in dimensionless expressions is dimensionally homogeneous.)

The information given thus far can be made clear by a famous example (see Fig. 1).

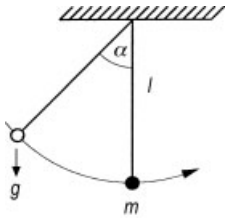

Figure 1. Sketch of the pendulum

Example 1: What Determines the Period of Oscillation of a Pendulum? We first sketch a pendulum and write down all the quantities which could be involved in this question. It may be assumed that the period of oscillation of a pendulum, t, depends on the length and mass of the pendulum, the gravitational acceleration, and the amplitude of swing:

Our aim is to express t as a function of l, m, g, and α:

$$t = f(l, m, g, \alpha)$$

Can this dependence be dimensionally homogeneous? No! The first thing that now becomes clear is that the base unit for mass M occurs only in the mass, m, itself. Changing this base unit (e.g., from kilograms to pounds) would change the numerical value of the function. This is unacceptable. Either our list should include a further variable containing M, or mass is not a relevant

variable. If we assume the latter, the relationship is reduced to:

$t = f(l, g, \alpha)$

Both l and g incorporate the base unit of length. When combined as a ratio (l/g) they become dimensionless with regard to L and thus independent of changes in the basic unit of length:

$t = f(l/g, \alpha)$

Since the angle α has no dimension, we are left with the dimension T on the left-hand side of the equation and T^2 on the right. To remedy this, we will have to write $\sqrt{l/g}$. This expression will keep its dimension T only if it remains unchanged, so we have to take it out of the function f and we obtain finally:

$$t = \sqrt{l/g}\ f(\alpha) \text{ or } t\sqrt{g/l} = f(\alpha)$$

The dependency between four dimensional quantities, containing two base dimensions (L and T) in their dimensions is reduced to a $4 - 2 = 2$ parametric relationship between dimensionless expressions ("numbers")!

This equation is the only statement that dimensional analysis can offer in this case. It is not capable of producing information on the form of f. The integration of Newton's equation of motion for small amplitudes leads to $f = 2\pi$ and is independent of α. The relationship can now be expressed as:

$$t\sqrt{g/l} = 2\pi$$

The elegant solution of this example should not tempt the reader to believe that dimensional analysis can solve every problem solely by a theoretical consideration. To treat this example by dimensional analysis, the acceleration due to gravity ($g = 9.81$ m/s^2) had to be known, this being calculated from the gravitational law. Sir Isaak Newton derived it from Kepler's laws of planet movements.

From Example 1 we also learn that a transformation of a physical dependence from a dimensional into a dimensionless form is automatically accompanied by an essential compression of the statement: The set of the dimensionless numbers is smaller than the set of the quantities contained in them, but it describes the problem equally comprehensively. In the Example 1 the dependence between four dimensional parameters is reduced to a dependency between only two dimensionless numbers. This is the proof of the so-called pi theorem (pi after Π, the sign used for products).

2.7. The Pi Theorem

The pi theorem states:

Every physical relationship between n physical quantities can be reduced to a relationship between $m = n - r$ mutually independent dimensionless groups, whereby r stands for the rank of the dimensional matrix, made up of the physical quantities in question and generally equal to the number of the base quantities contained in them.

(The pi theorem is often associated with the name of E. Buckingham, because he introduced this term in 1914, but the proof of it was accomplished in the course of a mathematical analysis of partial differential equations by A. Federmann already in 1911, see [1, p. 7].)

3. The Determination of a Pi Set by Matrix Transformation

3.1. The Establishment of a Relevance List of a Problem

As a rule, more than two dimensionless numbers will be necessary to describe a physico-technological problem and therefore they cannot be derived by the method described above. In this case the easy and transparent matrix calculation introduced by J. Pawlowski [2] is increasingly used. The calculation will be demonstrated by the following example. It treats an important problem in industrial chemistry and biotechnology because the mixing vessel belongs to the most important production equipment.

Example 2 : Which is the Pi Set for the Mixing Power of a Stirrer? In this case, too, we first sketch the apparatus (see Fig. 2). Then we facilitate our work by proceeding systematically:

1) Target quantity: mixing power P
2) Parameters
   a) geometrical: stirrer diameter d

b) physical properties:
   fluid density ρ
   kinematic viscosity ν
c) process related:
   stirrer speed n

We obtain the following relevance list:

$$\{P;\ d;\ \rho;\ \nu;\ n\} \tag{1}$$

3.2. Determination of the Characteristic Geometric Parameter

Obviously, we could name all the geometric parameters indicated in the sketch. They were all the geometric parameters of the stirrer and of the vessel, especially its diameter D and the liquid height H. In case of complex geometry such a procedure would deflect from the problem. It is therefore advisable to introduce only one characteristic geometric parameter, knowing that all the others can be transformed into dimensionless geometric numbers by division by this one.

In the case of Example 2, we introduced the stirrer diameter as the characteristic geometric parameter. This is reasonable. One can imagine how the mixing power would react to an increase of the vessel diameter D: It is obvious that from a certain D on there would be no influence of it, but a small change of the stirrer diameter d would always have an impact.

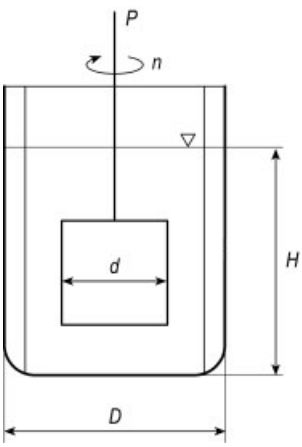

Figure 2. Sketch of the mixing vessel

3.3. Constructing and Solving of the Dimensional Matrix

In transforming the relevance list, Equation (1), of the above five physical quantities into a dimensional matrix, the following should be kept in mind in order to minimize the calculations required:

1) The dimensional matrix consists of a square core matrix and a residual matrix.
2) The rows of the matrix are formed of base dimensions, contained in the dimensions of the quantities, and they will determine the rank r of the matrix. The columns of the matrix are presented by the physical quantities or parameters.
3) Quantities of the square core matrix may eventually appear in all of the dimensionless numbers as "fillers", whereas each element of the residual matrix will appear only in one dimensionless number. For this reason the residual matrix should be loaded with essential variables like the target quantity and the most important physical properties and process-related parameters.
4) By the (extremely easy!) matrix transformation, the core matrix is transformed into a matrix of unity: The main diagonal consists only of ones and the remaining elements are all zero. One should therefore arrange the quantities in the core matrix in a way to facilitate this procedure.
5) After the generation of the matrix of unity, the dimensionless numbers are created as follows: Each element of the residual matrix forms the numerator of a fraction, while its denominator consists of the fillers from the matrix of unity with the exponents indicated in the residual matrix.

In Example 2 the dimensional matrix reads:

		ρ	d	n	P	ν
Mass	M	1	0	0	1	0
Length	L	−3	1	0	2	2
Time	T	0	0	−1	−3	−1
			core matrix		residual matrix	

Only one linear transformation is necessary to transform −3 in L-row/ρ-column into zero. The subsequent multiplication of the T-row by −1 transfers −1 to 1:

	ρ	d	n	P	ν
M	**1**	0	0	1	0
3M+L	0	**1**	0	5	2
−T	0	0	**1**	3	1
		unity matrix		residual matrix	

The residual matrix contains only two parameters, therefore only two Π numbers result:

$$\Pi_1 \equiv \frac{\mathrm{P}}{r^1\mathrm{n}^3\mathrm{d}^5} = \frac{\mathrm{P}}{r\mathrm{n}^3\mathrm{d}^5} \equiv Ne \ \text{(Newton number)}$$

$$\Pi_2 \equiv \frac{\nu}{r^0\mathrm{n}^1\mathrm{d}^2} = \frac{\nu}{\mathrm{nd}^2} \equiv Re^{-1}$$

(Reynolds number)

The interdependence of five dimensional quantities of the relevance list, Equation (1), reduces to a set of only $5-3=2$ dimensionless numbers:

$$\{Ne, Re\} \text{ or } f(Ne, Re) = 0 \tag{2}$$

thus again confirming the pi theorem.

3.4. Determination of the Process Characteristics

The functional dependence, Equation (2), is the maximum that dimensional analysis can offer here. It cannot provide any information about the form of the function f. This can be accomplished solely by experiments. But such experiments can now be conducted far more easily because it is the Reynolds number that must be varied and this can be accomplished by altering the stirrer speed n, or the characteristic length (here d), or the kinematic viscosity ν. By choosing appropriate model fluids, the viscosity can be very easily altered by several orders of magnitude. This way measurements were conducted to determine the *power characteristic* $Ne(Re)$ of a blade stirrer of fixed geometry in a vessel.

From Figure 3 it follows that:

1) In the range $Re < 20$ the proportionality $Ne \propto Re^{-1}$ is found, thus resulting in the expression $Ne\,Re \equiv P/(\mu n^2 d^3) = \text{const}$. Density is irrelevant here, we are dealing with the *laminar* flow region.
2) In the range $Re > 50$ (vessel with baffles) or $Re > 5 \times 10^4$ (unbaffled vessel) a constant Newton number $Ne \equiv P/(\rho n^3 d^5)$ is found. In this case, viscosity is irrelevant, we are dealing with a *turbulent* flow region.
3) Understandably, the baffles do not influence the power characteristics within the laminar flow region. However, their influence is extremely strong at $Re > 5 \times 10^4$. Here, the installation of baffles under otherwise unchanged operating conditions increases the power consumption of the stirrer by a factor of 20!
4) When liquid is stirred in an unbaffled vessel, it begins to rotate and a vortex is formed. Gravitational acceleration g, and hence the Froude number $Fr \equiv n^2\,d/g$, plays no role under these circumstances as determined experimentally (Fig. 3). This is confirmed by the points on the lower $Ne(Re)$ curve where the same Re value was set for fluids with different viscosities. This was only possible by a proportional alteration of the stirrer speed. At Re = idem (= identical value), Fr clearly was not idem, but this had no influence on Ne, g is therefore irrelevant!

4. Fundamentals of the Theory of Models and Scale-Up

4.1. Theory of Models

The results in Figure 3 could have also been acquired by changing the stirrer diameter: It does not matter, by which means a relevant number is changed. This gives clear evidence that the representation of a physico-technological problem in a dimensionless form is independent of scale ("scale-invariant"), and this presents the basis for a reliable scale-up:

> Two processes may be considered completely similar if they take place in similar geometrical space and if all the dimensionless numbers necessary to describe them have the same numerical value (Π_i = identical or idem).

Clearly, the scale-up of a desired process condition from a model to industrial scale can be accomplished reliably only if the problem was formulated and dealt with according to the dimensional analysis.

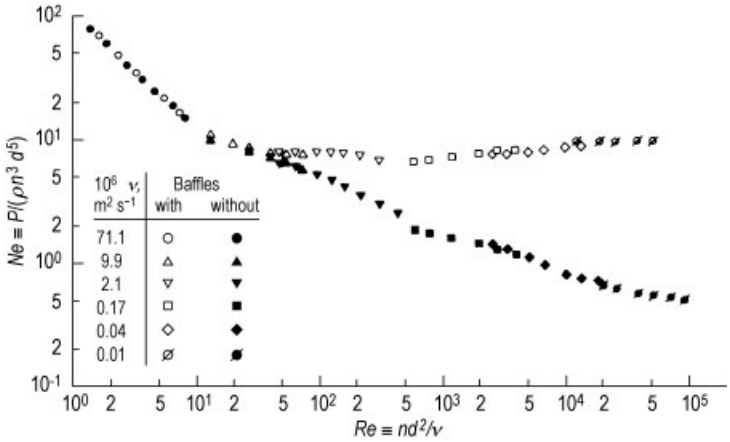

Figure 3. Power characteristics of a leaf stirrer of a given geometry and given installation conditions [3]

4.2. Model Experiments and Scale-Up

In Example 2 process characteristics were evaluated, that presented a comprehensive description of the process. This often expensive and time-consuming method is certainly not necessary if a given process condition only has to be scaled up from the model to the industrial plant (or vice versa). With Example 2, and assuming that the $Ne(Re)$ characteristics like that in Figure 3 is not known, the task is to predict the power consumption of a leaf stirrer of $d = 2$ m, installed in a baffled vessel of $D = 4$ m ($D/d = 2$) and rotating with $n = 10\ \mathrm{min}^{-1}$. The kinematic viscosity of the liquid is $\nu = 1 \times 10^{-4}\ \mathrm{m}^2/\mathrm{s}$.

It only has to be known – and this is essential – that the hydrodynamics in this case are governed solely by the Reynolds number and that the process is described by an unknown dependence $Ne(Re)$. Then Re of the industrial plant can be calculated:

$$Re \equiv n\,d^2/\nu = 6.7 \times 10^3$$

Let us assume that we have a geometrically similar laboratory device of $D = 0.2$ m and $d = 0.1$ m and that the stirrer speed can be arbitrarily chosen. Which rotational speed must be chosen to obtain $Re =$ idem using water ($\nu = 1 \times 10^{-6}\ \mathrm{m}^2/\mathrm{s}$) as model liquid? The answer is

$$n\,d^2/\nu = 6.7 \times 10^3 \rightarrow n = 40\ \mathrm{min}^{-1}$$

Under these conditions the stirrer power must be measured and the power number $Ne \equiv P/(\rho\, n^3\, d^5)$ calculated. We find $Ne = 8.0$. Because $Re =$ idem results in $Ne =$ idem, the power consumption P_T of the industrial stirrer can be obtained:

$$Ne = \mathrm{idem} \rightarrow Ne_\mathrm{T} = Ne_\mathrm{M} \rightarrow \left(\frac{P}{\varrho n^3 d^5}\right)_\mathrm{T} = \left(\frac{P}{\varrho n^3 d^5}\right)_\mathrm{M}$$

From $Ne = 8.0$ found in laboratory measurement, it follows for the leaf stirrer having $d = 2$ m and stirrer speed $n = 10\ \mathrm{min}^{-1}$:

$$P = Ne\varrho n^3 n^5 = 8.0 \times 1 \times 10^3 \mathrm{kg/m^3} \times \left(10/60\mathrm{s}^{-1}\right)^3 \times 2^5 \mathrm{m}^5 = 1185\mathrm{W} \cong 1.2\mathrm{kW}$$

We realize that in scale-up, knowledge of the functional dependency $f(\Pi_\mathrm{i}) = 0$ is not necessary. All we need is to know which pi space describes the process.

5. Further Procedures to Establish a Relevance List

5.1. Consideration of the Acceleration due to Gravity *g*

If a natural or universal physical constant has an impact on the process, it has to be incorporated into the relevant list, whether it will be altered or not. In this context the greatest mistakes are made with regard to the gravitational constant g.

This is all the more surprising in view of the fact that the relevance of this quantity is easy enough to recognize if one asks the following question: Would the process function differently if it took place on the moon instead of on earth? If the answer to this question is affirmative, g is a relevant variable.

The gravitational acceleration g can be effective solely in connection with the density as gravity $g\rho$. When inertial forces play a role, the density ρ has to be listed additionally. Thus it follows that:

1) In cases involving the ballistic movement of bodies, the formation of vortices in stirring, the bow wave of a ship, the movement of a pendulum and other processes affected by the earth's gravity, the relevance list comprises $g\rho$ and ρ.
2) Creeping flow in a gravitational field is governed by the gravity $g\rho$ alone.
3) In heterogeneous physical systems with density differences (sedimentation or buoyancy), the difference in gravity $g\Delta\rho$ and ρ play a decisive role.

In Example 3 the problem is discussed which belongs to the mentioned second class.

Example 3: Which Pi Number Describes the Creeping Flow of a very Viscous Liquid in a Gravitational Field (see Fig. 4)**?**

Physical quantity	Symbol	Dimension
Flow velocity	v	$\mathrm{L\,T^{-1}}$
Pipe diameter	d	L
Dynamic viscosity	μ	$\mathrm{M\,L^{-1}\,T^{-1}}$
Liquid weight	ρg	$\mathrm{M\,L^{-2}\,T^{-2}}$

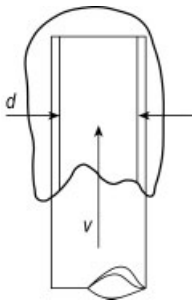

Figure 4. Sketch of creeping flow of a liquid in a gravitational field

Here, the density has to be incorporated into the relevance list in spite of the fact that the liquid motion is creeping flow, because without the fluid density the acceleration due to gravity would not work. Therefore, the liquid density has to be combined with the gravitational acceleration as liquid weight, $g\rho$.

In this case, the dimensional analysis is very easy:

1) Elimination of the base dimension [M] leads to $\mu/\rho g$ [LT]
2) Elimination of the base dimension [T] leads to $\mathrm{v}\mu/\rho g$ [L^2]
3) Elimination of the base dimension [L] leads to $\mathrm{v}\mu/(\rho g\, d^2)$ [–]

The resulting pi number, which governs the process, can be interpreted as the quotient of the Reynolds number *Re* and the Froude number *Fr*:

$$\frac{\mathrm{Fr}}{\mathrm{Re}} = \frac{\mathrm{v}\mu}{\mathrm{g}rd^2} \text{ with } Re \equiv \frac{\mathrm{v}dr}{\mu} \text{ and } Fr \equiv \frac{\mathrm{v}^2}{dg} \qquad (3)$$

5.2. Introduction of Intermediate Quantities

Relevance lists of some problems contain a whole host of parameters. This makes the elaboration of the process characteristics a difficult endeavor. In some cases a closer look at a problem (or previous experience) facilitates a reduction of the number of physical quantities in the relevance list. This is the case when some relevant variables affect the process by way of a so-called intermediate quantity. Assuming that this intermediate quantity can be measured experimentally, it should be included in the problem relevance list, if this facilitates the removal of more than one quantity from the list. If, e.g., the target quantity, y, depends on five parameters

$$y = f_1(x_1, x_2, x_3, x_4, x_5)$$

and three of them can be replaced by an intermediate quantity z

$$z = f_2(x_2, x_3, x_4)$$

then the introduction of z reduces the functional dependence by two parameters to

$$y = f_3(x_1, z, x_5)$$

In the following, a few examples are presented.

A universally known intermediate quantity is the flow velocity, v, in pipes or the so-called superficial gas velocity, v_G, in stirring vessels at gassing or in bubble columns:

$$v = \frac{\text{volume throughput}\, q}{\text{cross-sectional area}\, S} \propto \frac{q}{D^2} \quad S = \pi D^2/4$$

Its introduction into the relevance list replaces two quantities: throughput, q, and diameter, D.

A settling process (sedimentation) depends on the parameters of the disperse phase. They can be substituted by the sedimentation velocity v_s:

$$v_s = f(d_p, g\Delta\rho, \varphi_p, \mu)$$

where d_p is the mean particle diameter, $g\Delta\rho$ the weight difference between solid and liquid phase, φ_π the volume portion of the solid phase, and μ the dynamic viscosity of the liquid.

In some cases, e.g., in the flotation process, many influencing parameters can be replaced by only one or two intermediate quantities. Then, it is legitimate to speak of intermediate quantities as of "lumped parameters". For details see [4, Example 31].

A tableting process depends on the physical parameters of the powder, as for instance mean particle diameter d_p, bulk density (porosity), and flow behavior (lubricating ability). The flow behavior depends also on humidity. These physical properties can be replaced by the lumped parameter powder compressibility κ [5].

The advantage gained by the introduction of intermediate quantities is demonstrated by an elegant example.

Example 4: Mixing Characteristics for Liquid Mixtures with Density and Viscosity Differences. The mixing time Θ necessary to achieve a molecular homogeneity of a liquid mixture – normally measured by decolorization methods – depends in material systems *without* differences in density and viscosity on only four parameters: stirrer diameter d, density ρ, kinematic viscosity ν, and stirrer speed n:

$$\{\Theta;\, d;\, \rho;\, \nu;\, n\} \tag{4}$$

This five-parametric dimensional space leads to the two-parametric mixing time characteristic:

$$n\Theta = f(Re) \text{ where } Re \equiv n\, d^2/\nu \tag{5}$$

In material systems *with* differences in density and viscosity, the relevance list, Equation (4), enlarges by the physical properties of the second mixing component, by the volume ratio of both phases $\varphi = V_2/V_1$ and, because of the density differences, inevitably by the gravity difference $g\Delta\rho$, to nine parameters:

$$\{\Theta;\, d;\, \rho_1;\, \nu_1;\, \rho_2;\, \nu_2;\, \varphi;\, g\Delta\rho;\, n\} \tag{6}$$

This results in a mixing time characteristic incorporating six numbers:

$$n\Theta = f(Re, Ar, \rho_2/\rho_1, \nu_2/\nu_1, \varphi) \tag{7}$$

where $Ar \equiv g\Delta\rho\, d^3/(\rho_1\, \nu_1^2)$ is the Archimedes number.

Meticulous visual observation of this mixing process (the slow disappearance of the Schlieren patterns as result of the disappearance of density differences), reveals that macromixing is quickly accomplished as compared with micromixing. The time-consuming micromixing process already takes place in a material system which can be fully described by the physical properties of the mixture:

$$\nu^* = f(\nu_1, \nu_2, \varphi) \text{ and } \rho^* = f(\rho_1, \rho_2, \varphi)$$

By introducing the intermediate quantities ν^* and ρ^*, the relevance list (Eq. 6), can be reduced by three parameters to

$$\{\Theta;\, d;\, \rho^*;\, \nu^*;\, g\Delta\rho;\, n\} \tag{8}$$

to obtain a mixing characteristic composed of only three numbers:

$$n\Theta = f(Re, Ar) \tag{9}$$

(In this case, Re and Ar have to be formed by ρ^* and ν^*!)

The process characteristics of a cross-beam stirrer was established in this pi space by evaluation of corresponding measurements [3, pp. 110 – 111], in two differently sized mixing vessels ($D = 0.3$ and 0.6 m), using different liquid mixtures ($\Delta\rho/\rho^* = 0.01 - 0.29$ and $\nu_2/\nu_1 = 1 - 5300$). The characteristics read:

$$\sqrt{nJ} = 51.6 Re^{-1} \left(Ar^{1/3} + 3 \right)$$
$$Re = 10^1 - 10^5; Ar = 10^2 - 10^{11}$$

This example clearly shows the big advantages achieved by the introduction of intermediate quantities.

5.3. Scale-up Procedure at Unavailability of Model Material Systems

When model material systems are not available (e.g., with non-Newtonian fluids) or the relevant material parameters are unknown (e.g., with foams, slimes, and sludges), model measurements must be carried out in models of various sizes.

The unavailability of the model material systems can sometimes limit the application of the dimensional analysis. In such cases it is, of course, absolutely wrong to speak of "limits of the dimensional analysis".

The following example shows how design and scale-up data can be obtained by model measurements with the same material system in differently sized laboratory devices.

Example 5: Scale-up of Mechanical Foam Breakers. To obtain reliable information on dimensioning and then scale-up of a given type of mechanical foam breaker (foam centrifuge, see sketch in Fig. 5), we wish to ascertain the mode of performing and to evaluate the model measurements of the device. Preliminary experiments have shown that for each foam emergence – proportional to the gas throughput q_G – for each foam breaker of diameter d a minimum rotational speed n_{min} exists that is necessary to control foam formation. The dynamic properties of the foam (density and viscosity, elasticity of the foam lamella, etc.) cannot be named or measured. We have to content ourselves by listing them wholesale as material properties S_i. In our model experiments we will of course be able to replace S_i by the known type of surfactant (foamer) and its concentration c_f, expressed in ppm.

In discerning the process parameters, we realize that the gravitational acceleration g has no impact on the foam breaking *within* the foam centrifuge: The centrifugal acceleration $n^2 d$ exceeds the gravitational one (g) by far! However, the water content of the foam entering the centrifuge depends very much on the gravitational acceleration: On the moon the water drainage would be by far less effective! In contrast to the dimensional analysis presented elsewhere [6] we are well advised to add g to the relevance list:

$$\{n_{min}; d; \text{type of foamer}, c_f; q_G, g\} \tag{10}$$

For the sake of simplicity, in the following n_{min} is replaced by n and q_G by q. For each type of foamer we obtain the following pi space:

$$\left\{ \frac{nd^3}{q}, \frac{q^2}{d^5 g}, c_f \right\} \text{ or, abbreviated } \{Y, Fr, c_f\} \tag{11}$$

To prove this pi space, measurements must be made in model equipment of different sizes, to produce a reliable process characteristics. Figure 5 gives results for a commercial foamer (Mersolat H, Bayer, Germany), the pi space of Equation (11) is seen to be fully satisfied.

The straight line in Figure 5 corresponds to the analytic expression

$$Y = Fr^{-0.4} c_f^{0.32} \tag{12}$$

which can be reduced after remodeling to

$$n\,d = \text{const}\, q^{0.2} g^{0.4} f(c_f) \tag{13}$$

Here, the foam breaker will be scaled-up according to its tip speed $u = \pi n d$ in model experiments, which will also moderately depend on the foam yield (q).

In all other foamers examined [6], the relationshipe $Y \sim Fr^{-0.45}$ was found. If the correlation

$$Y \sim Fr^{-0.5} f(c_f)$$

proves to be true, then it can be deduced to

$$n^2\, d/g = c_f^b$$

In this case the centrifugal acceleration ($n^2 d$) would present the scale-up criterion and would depend only on the foamer concentration and not on foam yield (q).

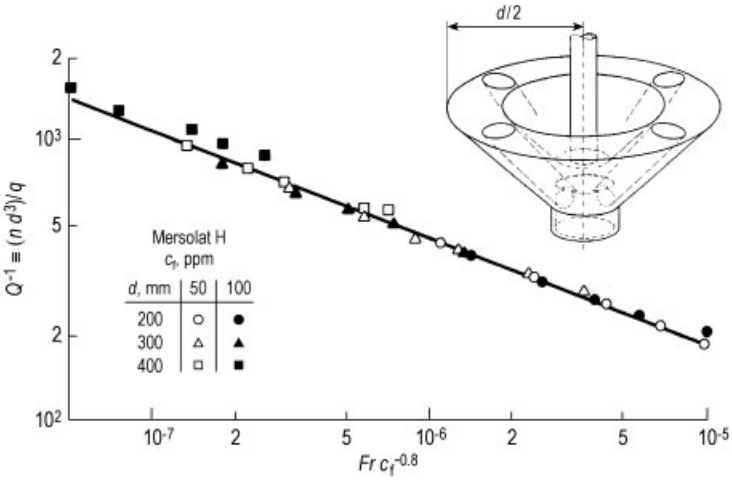

Figure 5. Process characteristics of a foam centrifuge (sketch) for a given foaming agent (Mersolat H) in two concentrations (c_f in ppm) [6]

5.4. Partial Similarity

In the scale-up from a model (subscript M) to the industrial scale (subscript T), not only must the geometric similarity be ensured, but also all dimensionless numbers describing the problem retain the same numerical values (Π_i = idem). This means, for example, that in the scale-up of boats or ships the dimensionless numbers governing the hydrodynamics

$$Fr \equiv \frac{v^2}{Lg} \text{ and } Re \equiv \frac{vL}{\nu}$$

must retain their numerical values: $Fr_T = Fr_M$ and $Re_T = Re_M$.

It can easily be shown that this requirement cannot be fulfilled here!

Since the gravitational acceleration g cannot be varied on earth, the Froude number of the model can be adjusted to that of the full-scale vessel only by its velocity v_M. Subsequently, Re = idem can be achieved only by the adjustment of the viscosity of the model fluid. If the model size is only 10 % of the full size (scale factor $\mu = L_T/L_M = 10$), Fr = idem is achieved in the model at $v_M = 0.32\ v_T$. To fulfill Re = idem, for the kinematic viscosity of the model fluid, ν_M, it follows:

$$\frac{\nu_M}{\nu_T} = \frac{v_M}{v_T}\frac{L_M}{L_T} = 0.32 \times 0.1 = 0.032$$

No liquid exists, whose viscosity would be only 3 % of that of water!

Some requirements concerning physical properties of model materials cannot be implemented. In such cases only a partial similarity can be realized. For this, essentially only two procedures are available (for details see [4, Examples 9, 10 and 41]). One consists of a well-planned experimental strategy, in which the process is divided into parts, which are then investigated separately under conditions of complete similarity. This approach was first applied by William Froude (1810 – 1879) in his efforts to scale-up the drag resistance of the ship's hull.

The second approach consists in deliberately abandoning certain similarity criteria and checking the effect on the entire process. This technique was used by Gerhard Damköhler (1908 – 1944) in his attempts to treat a chemical reaction in a catalytic fixed-bed reactor by means of dimensional analysis. Here the problem of a simultaneous mass and heat transfer arises, two processes that obey to completely different fundamental principles.

It is seldom realized that many "rules of thumb" utilized for scale-up of different types of equipment are represented by quantities which fulfill only a partial similarity. Examples are the volume-related mixing power P/V (widely used for scaling up mixing vessels) and the superficial velocity v_G (normally used for scale-up of bubble columns).

The volume-related mixing power P/V presents an adequate scale-up criterion only in

liquid-liquid dispersion processes and can be deduced from the pertinent process characteristics $d_p/d \sim We^{-0.6}$ (d_p, droplet diameter; *We*, Weber number). In the most common mixing operation, the homogenization of miscible liquids, where macro- and backmixing are required, this criterion fails completely [3].

Similarly, the superficial velocity v_G of the gas throughput as an intensity quantity is a reliable scale-up criterion only in mass transfer in gas-liquid systems in bubble columns. In mixing operations in bubble columns, requiring the whole liquid content being backmixed (e.g., in homogenization), this criterion completely loses its validity [4].

Thus, a particular scale-up criterion that is valid in a given type of apparatus for a particular process is not necessarily applicable to other processes occurring in the same device.

6. Short Summary of the Essentials of the Dimensional Analysis and Scale-Up

6.1. Advantages of Dimensional Analysis

The advantages made possible by correct and timely use of dimensional analysis are as follows:

1) *Reduction of the number of parameters required to define the problem.* The Π theorem states that a physical problem can always be described in dimensionless terms. Thus the number of dimension*less* groups that fully describe it is much smaller than the number of dimensional physical quantities and is generally equal to the number of physical quantities minus the number of base dimensions contained in their dimensions.
2) *Reliable scale-up of the desired operating conditions from the model to the full-scale plant.* According to the theory of models, two processes may be considered to be similar if they take place under geometrically similar conditions and all dimensionless numbers which describe the process have the same numerical value.
3) *A deeper insight into the physical nature of the process.* When experimental data are presented in a dimensionless form, distinct physical states can be isolated from one another (e.g., turbulent or laminar flow regions), and the effect of individual physical variables can be identified.
4) *Flexibility in the choice of parameters and their reliable extrapolation within the range covered by the dimensionless numbers.* These advantages become clear when considering the Reynolds number, $Re \equiv v\,l/\nu$, which can be varied by altering the characteristic velocity v, or a characteristic length *l*, or the kinematic viscosity ν. By choosing appropriate model fluids, the viscosity can be very easily altered by several orders of magnitude. Once the effect of the Reynolds number is known, extrapolation of both v and *l* is allowed within the examined range of *Re*.

6.2. Area of Applicability of Dimensional Analysis

The application of dimensional analysis is indeed heavily dependent on the available knowledge. The following five steps (see Fig. 6) can be outlined:

1) The physics of the basic phenomenon is unknown.
   → Dimensional analysis cannot be applied.
2) Enough is known about the physics of the basic phenomenon to compile a first, tentative relevance list.
   → The resultant pi set is unreliable.
3) All the relevant physical variables describing the problem are known.
   → The application of dimensional analysis is unproblematic.
4) The problem can be expressed in terms of a mathematical equation.
   → A closer insight into the pi relationship is feasible and may facilitate a reduction of the set of dimensionless numbers.
5) A mathematical solution of the problem exists.
   → The application of dimensional analysis is superfluous.

It must, of course, be said that approaching a problem from the point of view of dimensional analysis also remains useful even if all the variables relevant to the problem are not yet known:

The timely application of dimensional analysis may often lead to the discovery of forgotten variables or the exclusion of artifacts.

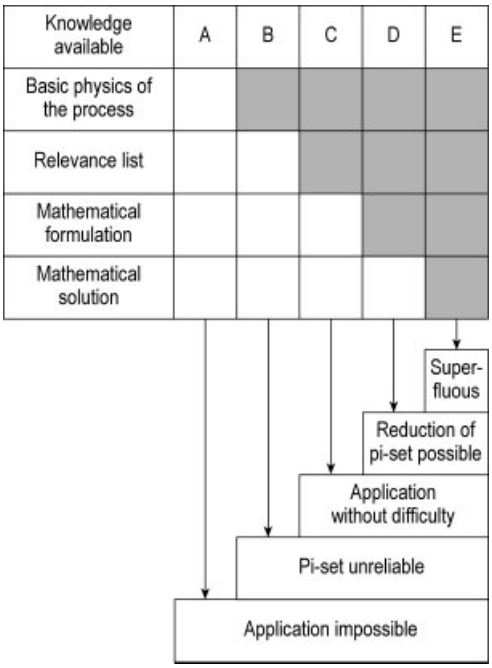

Figure 6. Applicability of dimensional analysis, as dependent on the knowledge available; after *J. Pawlowski*

6.3. Experimental Methods for Scale-Up

In Chapter 1 a number of questions were posed which are often asked in connection with model experiments.

How small can a model be? The size of a model depends on the scale factor L_T/L_M, and on the experimental precision of measurement. Where $L_T/L_M = 10$, $a \pm 10\,\%$ margin of error may already be excessive. A larger scale for the model will therefore have to be chosen to reduce the error.

Is one model scale sufficient or should tests be carried out in models of different sizes? One model scale is sufficient if the relevant numerical values of the dimensionless numbers necessary to describe the problem (the so-called "process point" in the pi space describing the operational condition of the technical plant) can be adjusted by choosing the appropriate process parameters or physical properties of the model material system. If this is not possible, the process characteristics must be determined in models of different sizes, or the process point must be extrapolated from experiments in technical plants of different sizes.

When must model experiments be carried out exclusively with the original material system? Where the material model system is unavailable (e.g., in the case of non-Newtonian fluids) or where the relevant physical properties are unknown (e.g., foams, sludges, slimes), the model experiments must be carried out with the original material system. In this case measurements must be performed in models of various sizes (cf. Example 5).

7. Treatment of Variable Physical Properties by Dimensional Analysis

When using dimensional analysis to tackle engineering problems, it is generally assumed that the physical properties of the material system remain unaltered in the course of the process. Relationships such as the heat transfer characteristic of a mixing vessel $Nu = f\ (Re,\ Pr)$ are valid for any material system with Newtonian viscosity and for any constant process temperature, i.e., for *constant* physical properties.

However, constancy of physical properties cannot be assumed in every physical process. A temperature *field* may well generate a viscosity *field* or even a density *field* in the material system treated. In non-Newtonian (pseudoplastic or viscoelastic) liquids, a shear rate can also produce a viscosity *field*.

Although most physical properties (e.g., viscosity, density, heat conductivity and capacity, surface tension) must be regarded as variable, it is particularly the value of viscosity that can be varied by many orders of magnitude under certain process conditions. Besides this, it shows the highest temperature coefficients.

Why is this topic given so much space in a treatise on dimensional analysis? An explanation is easy to give: A complete similarity requires a geometrical, material, and process-related similarity, but the material similarity often cannot be easily obtained.

A problem arises, e.g., when model (laboratory, bench-scale) measurements are to be per-

formed in a so-called "cold model", but the industrial plant operates at high temperatures (petrochemicals; $T \approx 800-1000\,°C$). How can we ascertain that the laboratory model system behaves hydrodynamically similarly to that in the industrial plant? Here, different temperature dependence of physical properties (viscosity, density) may cause problems.

A problem also arises when laboratory measurements are to be performed with cheap and easy to handle model fluids in order to gain information about the scaling-up of an apparatus for treatment of cell cultures in biotechnology (mammal and plant cells, aerobic cultures, yeasts), the rheological behavior of which is very complex (non-Newtonian: pseudoplastic and viscoelastic). Which model system may we choose?

The answer is clear and unambiguous: We may choose any model material system whose dimensionless material function in question is similar to that of the original material system. In this chapter the necessary procedures to obtain this information will be shown.

7.1. Dimensionless Representation of the Material Function $\mu(T)$

Similar behavior of a certain physical property common to different material systems can only be visualized by a *dimensionless representation* of the material function of that property. It is furthermore desirable to formulate this function as uniformly as possible.

This can be achieved by the "standard representation" of the material function $\mu(T)$ in which a standardizing transformation is defined in such a way that the expression produced

$$\mu/\mu_0 = f\,\{-\gamma_0\,(T - T_0)\} \tag{14}$$

meets the requirement $f\,(0) = f'(0) = 1$.

$$\gamma_0 \equiv \left(\frac{1}{\mu}\frac{\partial\mu}{\partial T}\right)_{T_0}$$

is the temperature coefficient of viscosity and

$$\mu_0 \equiv \mu(T_0)$$

where T_0 is any reference temperature.

To achieve this standard representation, the original dependence $\mu(T)$ had to be enlarged by three additional parameters to $\mu(T, \mu_0, T_0, \gamma_0)$.

Figures 7A and 7 B depicts the dramatic effect of this standard transformation. They show that all liquids presented in this diagram behave similarly to one another with respect to $\mu(T)$. This is the more surprising in view of the fact that their viscosities cover a range of five orders of magnitude and also their temperature coefficients of the viscosity γ_0 differ widely.

T_0 stands for the mean process temperature. In fact, any temperature could have been taken in the case of these substances as the reference temperature in the range covered by experiments. The presented evaluation of $\mu(T)$ is independent of the reference point T_0, it is *invariant* of it.

Engineers prefer the representation

$$\mu/\mu_0 = \exp\,[-\gamma_0\,(T - T_0)] \tag{15}$$

which is a special case of Equation (14) – f being an e function – , although this is not the best possible approximation (represented by the dotted line in Fig. 7B).

A slightly better approximation of the material function $\mu(T)$ is provided by the well known Arrhenius relationship:

$$\mu/\mu_0 = \exp\left\{-\frac{E_0}{RT_0}\left(\frac{T_0}{T} - 1\right)\right\} \tag{16}$$

see [4, Chap. 8].

Because of the relation between μ/μ_0 and $\gamma_0\Delta T$ in Equation (15), the engineering literature prefers the incorporation of the quotient μ/μ_0, e.g., μ_w/μ (w = wall), instead of $\gamma_0\Delta T$ in the heat transfer characteristics. Therefore they normally read

$$Nu = f\,(Re, Pr, \mu_w/\mu) \tag{17}$$

Were the function $\mu(T)$ not independent (invariant) of the reference point T_0, an additional number, $\Delta T/T_0$, had necessarily be incorporated in Equation (17) to fully describe the $\mu(T)$ behavior.

In order to describe the *process* by dimensional analysis, it is advisable to formulate the reference temperature T_0 in a process-related manner, using a characteristic, possibly mean process temperature as reference. A class of functions exists, however, which are independent (invariant) of the reference point T_0. The

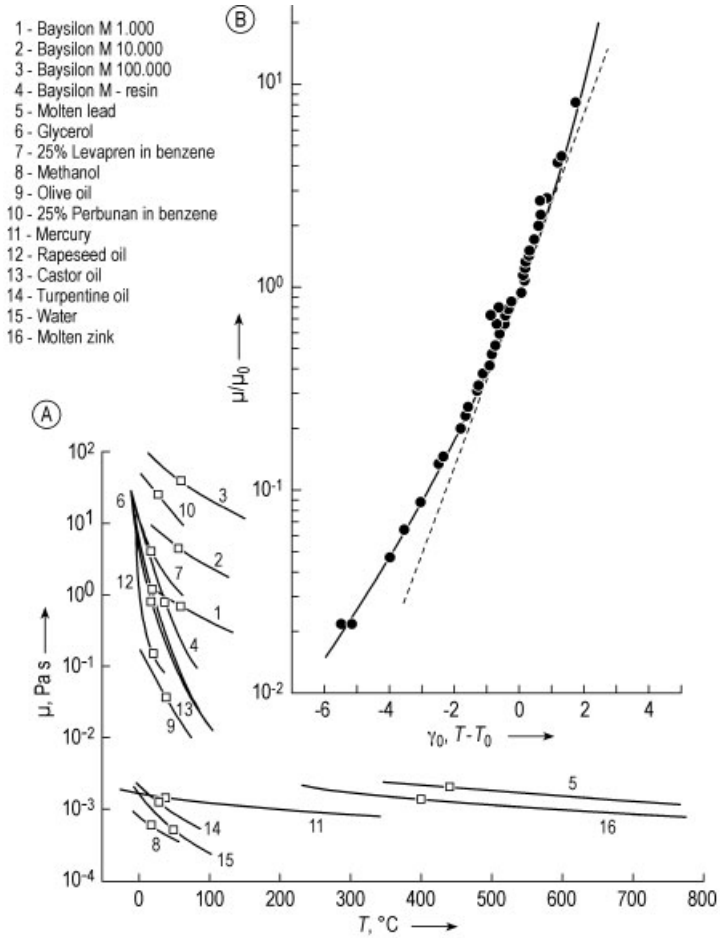

Figure 7. Temperature dependence of the viscosity (**A**) and its standard representation (**B**) which proves to be invariant with respect to the reference point T_0.

solid line in Figure 7B shows a function of this type. (Note: Functions given in Eqs. 14, 15, and 16 are special cases of invariant functions.) For details, see [4, 7].

7.2. Dimensionless Representation of the Material Function $\rho(T)$

The dependence $\rho(T)$ for some organic fluids and water is represented in Figure 8, whereas Figure 9 shows the standard transformation of this dependence. The temperature coefficient of the density, β_0, is given by the expression

$$\beta_0 \equiv \left(\frac{1}{\varrho_0}\frac{\partial \varrho}{\partial T}\right)_{T_0} \tag{18}$$

As shown in Figure 9 the standard transformation of $\rho(T)$ is invariant of the reference point T_0 only for the organic fluids considered, but not for water (see the signs o and •). This indicates that water cannot be taken as a model fluid in the so-called "cold model", in order to obtain a reliable information about the behavior of organic fluids at high temperatures in an industrial plant. Only in the range close to the reference point, $\beta_0 \Delta T \approx 0$, water behaves similarly to these organic fluids with respect to $\rho(T)$.

7.3. Pi Set for Variable Physical Properties

The type of the dimensionless representation of the material function affects the extended pi set

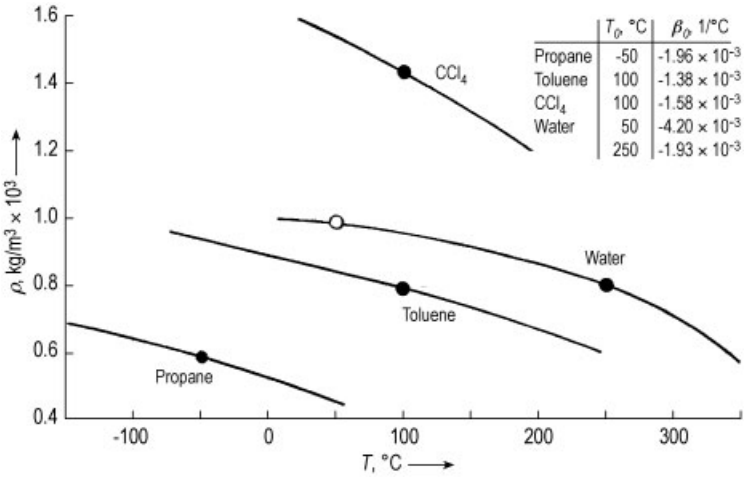

Figure 8. Dependence $\rho(T)$ for some organic fluids and water

within which the process equation is formulated. In case of the standard representation of $\mu(T)$, the relevance list must include two additional parameters γ_0, T_0. This leads to two additional dimensionless numbers in the process characteristics. With regard to the heat transfer characteristics of a mixing vessel, it followed that

$$Nu = f\,(Re_0, Pr_0, \gamma_0 \Delta T, \Delta T/T_0) \tag{19}$$

where the subscript $_0$ in Re and Pr denotes that these two dimensionless numbers are to be formed with μ_0 (which is the numerical value of μ at T_0).

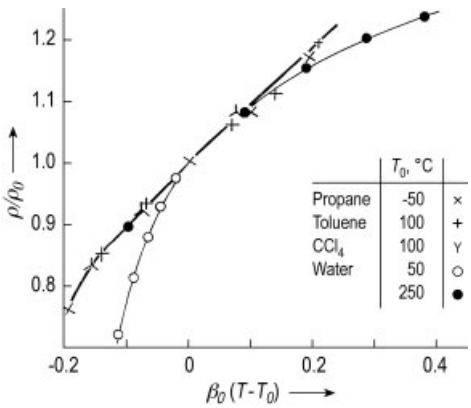

Figure 9. Standard transformation of the dependence $\rho(T)$ in Figure 8, which proves that for water this dependence is not invariant with respect to the reference point T_0

Considering that the standard transformation of the material function can be expressed invariantly with regard to the reference temperature T_0, see Fig. 7B), the relevance list is extended by only one additional parameter γ_0. This, in turn, leads to only one additional dimensionless number. For the above problem it then follows that

$$Nu = f\,(Re_0, Pr_0, \gamma_0 \Delta T) \text{ or } Nu = f\,(Re_0, Pr_0, \mu_w/\mu_0) \tag{20}$$

7.4. Material Function in Non-Newtonian Liquids

In Newtonian liquids the shear stress τ(Pa) is proportional to the shear rate $\dot{\gamma}(\mathrm{s}^{-1})$

$$\tau = \mu\,\dot{\gamma} \rightarrow \mu = \tau/\dot{\gamma} \tag{21}$$

and the proportionality constant is the dynamic viscosity μ (→ Fluid Mechanics, Chap. 3.1.).

In non-Newtonian liquids μ actually depends on the effective shear rate $\dot{\gamma}$ and occasionally on its history as well. Such liquids are subdivided accordingly to their flow behavior into three classes (→ Fluid Mechanics, Chap. 4.):

1) The viscosity does not depend upon the duration of the shear
2) The viscosity depends upon the duration of the shear
3) The liquid partly behaves like a solid body

These interrelations are recorded in the rheological constitutive equations.

7.4.1. Pseudoplastic Flow Behavior (→ Fluid Mechanics, Chap. 4.1.)

The pseudoplastic or *shear-thinning* fluids represent the most important group within the first class of materials. Under shear stress aggregates of a liquid – solid or liquid – liquid (e.g., dyes) dispersion disintegrate into single particles, which then orientate themselves in the direction of flow. For instance, entangled chains of macromolecules of a polymer solution or melt are stretched, spherical erythrocytes of blood are elongated. In these cases viscosity is degraded by shear.

In most pseudoplastic liquids, Newtonian flow behavior is observed at sufficiently low and at high shear rates $\dot{\gamma}$, see Figure 10. Viscosity approaches a constant value with low shear rates, which is called the zero-shear viscosity, μ_0, and its constant value at very high shear rates is called the infinite shear viscosity, μ_∞.

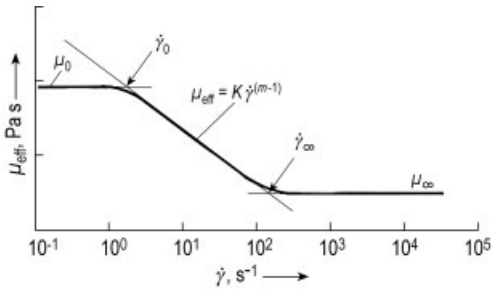

Figure 10. Typical course of μ_{eff} $(\dot{\gamma})$ in pseudoplastic fluids

In a dimensional-analytical discussion of the rheological constitutive equations, PAWLOWSKI [7] furnished the proof that the rank of the dimensional matrix is always two. By a convenient association of the quantities contained in them, two dimensional material parameters can be produced (see Fig. 10):

H – a characteristic viscosity constant, e.g., μ_0, and (22)

Θ – a characteristic time constant, e.g., $1/\dot{\gamma}_0$ or $1/\dot{\gamma}_\infty$ (23)

The relevance list of a process in which a rheological material participates is therefore extended by only two dimensional quantities. All the other material parameters can be transferred into dimensionless pi numbers Π_{rheol}.

Thus the rheological constitutive equation reads

$$\mu/H = f\,(\dot{\gamma}\Theta, \Pi_{\mathrm{rheol}}) \rightarrow \mu/\mu_0 = f\,(\dot{\gamma}/\dot{\gamma}_0, \Pi_{\mathrm{rheol}}) \qquad (24)$$

If the viscosity curve is plotted in a log – log scale and the transition range between μ_0 and μ_∞ is represented by a straight line with the slope $m < 1$ (see Fig. 10), this is called an Ostwald – de Waele (power law) fluid (→ Fluid Mechanics, Chap. 4.1.2.)

$$\mu_{\mathrm{eff}} = K\dot{\gamma}^{(\mathrm{m}-1)} \qquad (25)$$

where μ_{eff} is the effective viscosity, K the consistency index, and m the flow index.

Shear rates normally appearing in mixing vessels are in the range of $\dot{\gamma} = 50 - 500\,\mathrm{s}^{-1}$, therefore many liquids behave like Ostwald – de Waele fluids. This explains why the power law is so often used to describe rheological behavior.

Many years ago, PAWLOWSKI [2, p. 124] pointed out that Equation (25) injures the principle of consistency of physical quantities, because the dimension of $[K] = \mathrm{M\,L^{-1}\,T}^{m-2}$ depends on the power m. If the viscosity is temperature-dependent this entails that in a temperature field, for example, the quantity K changes its dimension from point to point and therefore neither *grad K* nor K/K_0 can be formed.

The flow index m is already a dimensionless parameter of the set Π_{rheol}. With the aid of parameters H and Θ (Eqs. 22, 23), the power law can be transformed into a dimensionless form. The rheological constitutive equation of an Ostwald – de Waele fluid reads:

$$\frac{\mu_{\mathrm{eff}}}{H} = \frac{K}{H\,J^{m-1}}\,(\dot{\gamma}\,J)^{m-1} \rightarrow$$
$$\frac{\mu_{\mathrm{eff}}}{\mu_0} = \frac{K}{\mu_0\,(1/\dot{\gamma})^{m-1}}\,(\dot{\gamma}/\dot{\gamma}_0)^{m-1} \qquad (26)$$

According to Equation (24), the material function of pseudoplastic fluids ($\Pi_{\mathrm{rheol}} = m$), whose viscosity obeys the power law of Ostwald – de Waele, can be represented in the pi space as follows:

$$\{\mu_{\mathrm{eff}}/\mu_0, \dot{\gamma}/\dot{\gamma}_0, m\}$$

HENZLER correlated the viscosity behavior of aqueous carboxymethyl cellulose (CMC) and xanthan solutions in this pi space with good success, as can be seen in Figure 11 [9]. The fitting line corresponds to the process equation

$$\mu_{\mathrm{eff}}/\mu_0 = (1 + (\dot{\gamma}/\dot{\gamma}_0)^{2(1-\mathrm{m})})^{-1/2} \qquad (27)$$

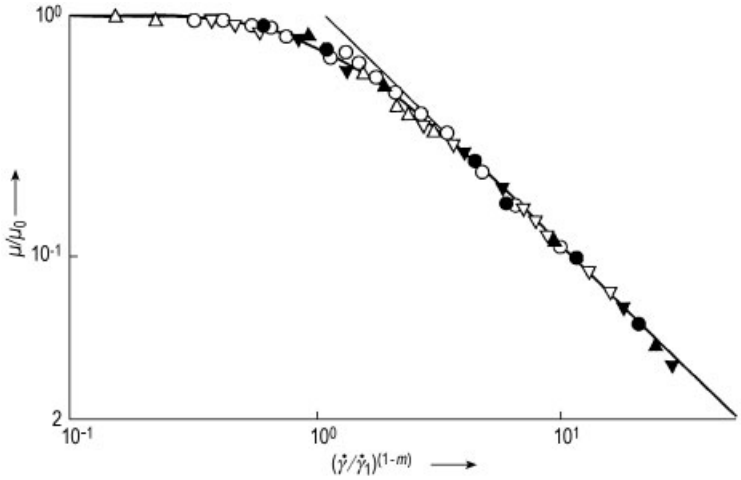

	Medium	c, g/l	μ_0, Pa s	μ_0/K, s^{1-m}	m
○	CMC CK 10 000	10	0.25		0.6
△		15	0.7	0.5	0.53
▽		20	2.5		0.45
●	Xanthan Rhodigel	0.5	0.041	0.78	0.58
▲		1	0.114	0.86	0.5
▼		2	0.6	1.3	0.35

Figure 11. Dimensionless, standardized material function of some pseudoplastic fluids [9]; for the meaning of $\dot{\gamma}_0$, see Figure 10

7.4.2. Viscoelastic Flow Behavior (→ Fluid Mechanics, Chap. 4.3.)

The viscoelastic fluids represent the third class of non-Newtonian fluids. Many liquids also possess elastic properties in addition to viscous properties. This means that the distortion work resulting from a stress is not completely irreversibly converted into frictional heat, but is stored partly elastically and reversibly. In this sense, viscoelastic fluids are similar to solid bodies. The liquid strains give way to the mechanical shear stress as do elastic bonds by contracting. This is shown in shear experiments as a restoring force acting against the shear force which, at the sudden ending of the effect of force, moves back the plate to a certain extent. (The procedure is fully described under → Fluid Mechanics, Chap. 4.)

In viscoelastic fluids at steady-state laminar flow, besides shear stress $\tau = \sigma_{21} = \mu\,\dot{\gamma}$, normal stresses are observed in all three directions:

$$\begin{array}{l} \text{in the direction of flow} \\ \sigma_{11} + p \\ \text{perpendicular to the direction of flow} \\ \sigma_{22} + p \text{ and } \sigma_{33} + p \end{array} \tag{28}$$

The isotropic pressure, p, can be eliminated by the formation of normal stress differences:

$$1.\text{normal stress difference } N_1 = \sigma_{11} - \sigma_{22} \tag{29}$$

$$2.\text{normal stress difference } N_2 = \sigma_{22} - \sigma_{33} \tag{30}$$

N_2 values are always lower than N_1 values, see, e.g., [9]. Therefore, for many processes considering only N_1 will suffice. The normal stress differences are independent of the direction of flow and, in laminar flow (low $\dot{\gamma}$) are proportional to $\dot{\gamma}^2$. In following $\mu = \tau/\dot{\gamma}$ for a Newtonian fluid, normal stress coefficients $\Psi_1 \equiv N_1/\dot{\gamma}^2$ and $\Psi_2 \equiv N_2/\dot{\gamma}^2$ are occasionally used. Their dependence on the shear rate $\Psi(\dot{\gamma})$ describes the non-linear viscoelastic behavior of the fluid.

For a correct dimensional-analytical representation of the viscoelastic behavior of a fluid, the ratio of normal stress to shear stress is used. The so-called Weissenberg number *Wi* is defined as

$$Wi_1 \equiv N_1/\tau \tag{31}$$

In Figure 12 Wi_1 values as a function of $\dot{\gamma}$ are presented for aqueous CMC and xanthan solutions.

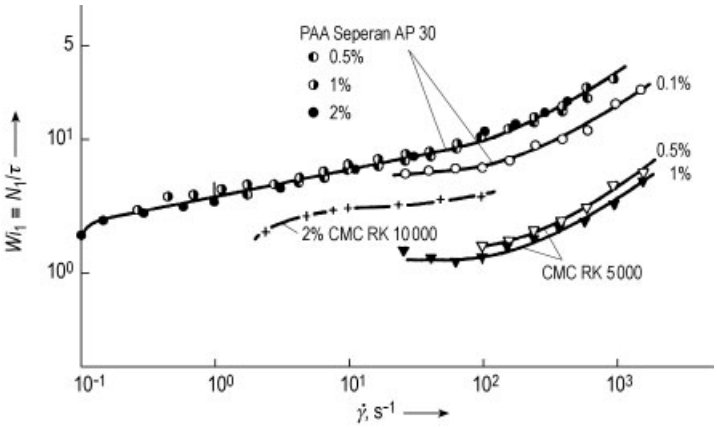

Figure 12. (Wi_1, $\dot{\gamma}$) dependence for CMC and polyacrylamide (PAA) solutions [9]

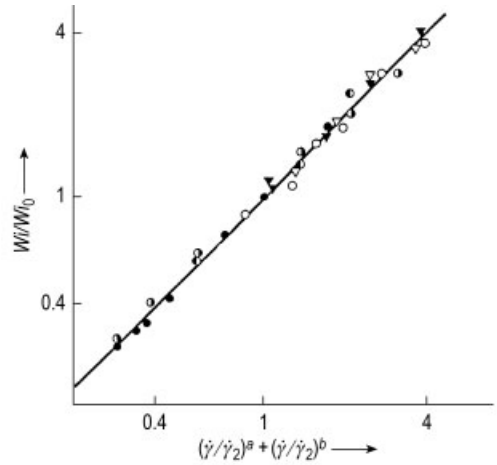

	Mixture	conc, %	Wi_0	$\dot{\gamma}_2$, s^{-1}	$\dot{\gamma}_3$, s^{-1}	a	b	$\dot{\gamma}$, s^{-1}
○	PAA Seperan AP 30	0.1	6.4	760	630	0.06	1.15	25-1600
◐		0.5						
◑		1	10	250	590	0.18		0.3-1000
●		2						
▽	CMC RK 5000	0.5	1.6		670	0		40-1600
▼		1	1.2					

Figure 13. Dimensionless standardized material function of some viscoelastic fluids [9]

Frequently, a characteristic relaxation time, λ, is used to describe viscoelastic behavior. It is a measure for the time needed to transform the reversibly-elastically stored energy into friction heat:

$$\lambda \equiv N_1/(2\,\tau\,\dot{\gamma}) = Wi_1/(2\,\dot{\gamma}) \tag{32}$$

If, according to the concept of METZNER and OTTO, $\dot{\gamma}$ is replaced by the stirrer speed n, then Equation (32) can also be written as

$$\mathrm{De} \equiv \lambda\, n \stackrel{\triangle}{=} \lambda\,\dot{\gamma} = Wi_1/2 \tag{33}$$

The product $n\,\lambda$ is named Deborah number, *De* (→ Rheology, Chap. 1.2.2.).

Figure 13 shows the dimensionless, standardized material function of two viscoelastic fluids, whose dependences Wi_1 ($\dot{\gamma}$) were given in Figure 10. The fitting line corresponds to the process equation

$$Wi_{\mathrm{eff}}/Wi_0 = (\dot{\gamma}/\dot{\gamma}_2)^{\mathrm{a}} + (\dot{\gamma}/\dot{\gamma}_3)^{\mathrm{b}} \tag{34}$$

(Exponents a and b have different values depending on the respective substance; see Fig. 13).

7.5. Pi Space in Processes with Non-Newtonian Fluids

As mentioned in Section 7.4.1, the transition from Newtonian to non-Newtonian fluids has the following consequences with respect to the enlargement of the pi space:

1) All pi numbers of the Newtonian case also appear in the non-Newtonian case, whereby μ is replaced by the quantity H (usually μ_0) with the same dimension (see Eq. 22).
2) An additional pi number is included which contains Q (usually $1/\dot{\gamma}_0$).
3) The pure material numbers are increased by Π_{rheol}

This will be demonstrated on the heat-transfer characteristics of a smooth straight pipe. Case "a" is valid for the temperature independent viscosity and "b" for the temperature dependent viscosity (Eq. 35):

	Newtonian fluid	non-Newtonian fluid
a	Nu, Re, Pr	Nu , $Re_{\mathrm{H}}, Pr_{\mathrm{H}}, \nu\Theta/L, \Pi_{\mathrm{rheol}}$
b	$Nu, Re_0, Pr_0, \gamma_0\Delta T$	$Nu, Re_{\mathrm{Ho}}, Pr_{\mathrm{Ho}}, \nu\Theta_0/L, \gamma_{\mathrm{Ho}}\Delta T, \gamma_{\mathrm{Q}}/\gamma_{\mathrm{Q}}, \Pi_{\mathrm{rheol}}$

In case (b) μ_{w}/μ and $\gamma_{\mathrm{Ho}}\Delta T$, respectively, as well as γ_Θ/γ_H have to be added ($\gamma_\Theta \equiv \delta \ln_\Theta/\delta T$).

In addition, it can happen that in the non-Newtonian case completely new phenomena take place (e.g., shaft climbing by a viscoelastic fluid against the acceleration due to gravity, the so-called Weissenberg effect, → Fluid Mechanics, Chap. 4.3.), calling for additional parameters (in this case g).

8. Optimization of Process Conditions by Combining Process Characteristics

In this chapter, one example is presented to demonstrate the possibility of optimizing a process with respect to the desirable objective by the appropriate combination of process characteristics.

Example 6: Process Characteristics of a Self-Aspirating Hollow Stirrer and the Determination of its Optimum Process Conditions. As a result of their form, hollow stirrers utilize the suction generated behind their edges (Bernoulli effect) to suck in gas from the head space above the liquid. As "rotating ejectors", they are stirrers and gas pumps in one and are therefore particularly suitable for laboratory use (especially in high pressure autoclaves) because they achieve intensive gas-liquid contacting via internal gas recycling without a separate gas pump [3]. A particularly effective type of this stirrer, the pipe stirrer, is depicted in Figure 14.

In operating a hollow stirrer, both target quantities – self-aspirated gas throughput, q, and stirrer power, P, – adapt themselves simultaneously. Both target quantities depend on the following parameters:

Geometry:	Stirrer diameter	d
Physical properties:	density	ρ
	viscosity	μ
Process parameters:	stirrer speed	n
	gravitational constant	g

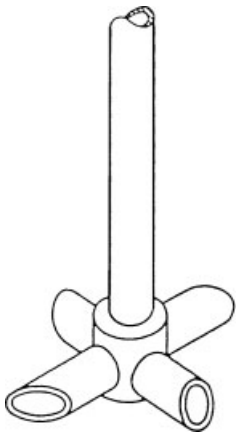

Figure 14. Sketch of the pipe stirrer

Thus we obtain two separate relevance lists:

$$\{q; d, \rho, \mu; n, g\} \tag{36}$$

$$\{P; d, \rho, \mu; n, g\} \tag{37}$$

The first relevance list (36) leads to the following matrix

	ρ	d	n	q	μ	g
M	1	0	0	0	1	0
L	−3	1	0	3	−1	1
T	0	0	−1	−1	−1	−2
M	**1**	0	0	0	1	0
3M+L	0	**1**	0	3	2	1
−T	0	0	**1**	1	1	2

which results in the following three pi numbers:

$$\Pi_1 \equiv \frac{q}{nd^3} \equiv Q \; \Pi_2 \equiv \frac{\mu}{rd^2 n} \equiv Re^{-1}$$
$$\Pi_3 \equiv \frac{g}{dn^2} \equiv Fr^{-1} \tag{38}$$

Π_1 is named the gas throughput number, Q, Π_2 is the inverse Reynolds number, Re, and Π_3 is the inverse Froude number, Fr. The gas throughput characteristics of a hollow stirrer then reads:

$$\frac{q}{nd^3} = f_1\left(\frac{nd^2 r}{\mu}, \frac{n^2 d}{\mu}\right) \rightarrow Q = f_1\left(Re, Fr\right) \tag{39}$$

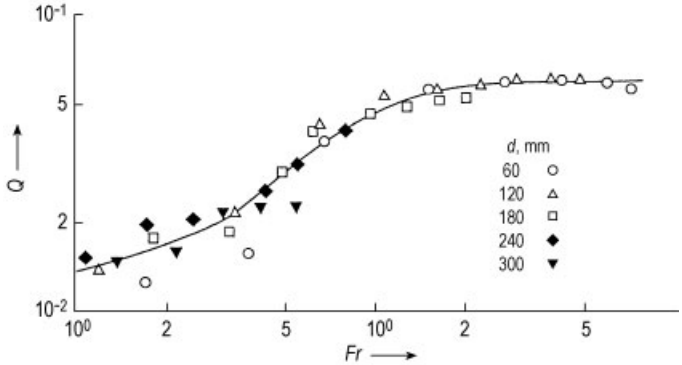

Figure 15. Gas throughput characteristic of a three-edged stirrer [3]

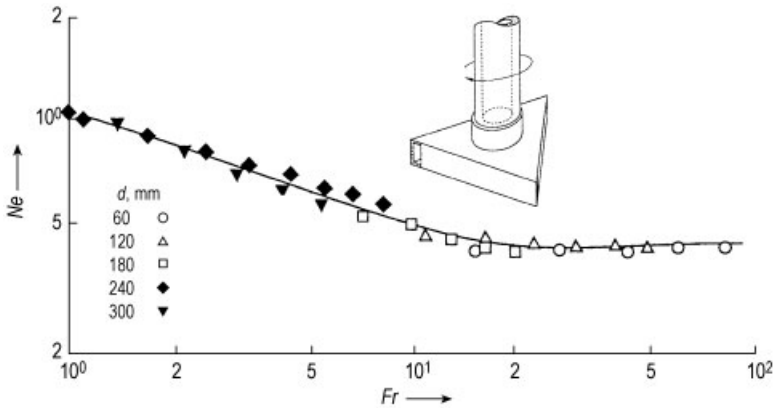

Figure 16. Power characteristic of a three-edged stirrer [3]

The power characteristics, obtained from the relevance list (Eq. 37) by a similar dimensional matrix containing P instead of q leads to:

$$\frac{P}{rn^3d^5} = f_2\left(\frac{nd^2r}{\mu}, \frac{n^2d}{\mu}\right) \rightarrow Ne = f_2\left(Re, Fr\right) \tag{40}$$

The pi number containing P is termed the Newton number, Ne.

Model experiments with another type of hollow stirrer (three-edged stirrer, see sketch in Fig. 16) were performed in water-air under defined experimental conditions and the scale factor $\mu = d_{\mathrm{T}}/d_{\mathrm{M}}$ was changed in the range of $\mu = 1:2:3:4:5$. The results demonstrate that the Reynolds number is irrelevant in the turbulent flow range ($Re > 10^4$) and the process is exclusively governed by the Froude number. Both process characteristics can therefore be represented as $Q = f_1(Fr)$ and $Ne = f_2(Fr)$ (see Figs. 15 and 16). These results impressively demonstrate to which extent information can be compressed by dimensional analysis.

These process characteristics present a reliable basis for the scale-up of this hollow stirrer under the given geometric conditions. But they also allow a further optimization of this process, for example with respect to the process conditions under which this stirrer will achieve a given gas throughput with the minimum power, $P/q = \min$. These conditions can be easily found by combining the above characteristics in Figures 15 and 16 in such a way that a dimensionless expression for P/q is produced. This is the pi number combination

$$\frac{NeFr}{Q} \equiv \frac{\boldsymbol{P}}{\boldsymbol{q}drg} = f_3\left(Fr\right) \tag{41}$$

This new pi number is, as are its constituents, also a function of the Froude number. This dependence is shown in Figure 17. It can be seen that two sections exist:

$\mathrm{Fr} \leq 10 : Ne\,Fr/Q = \mathrm{const} \rightarrow P/q \propto d$

Here, P/q diminishes directly proportionally to scale (d).

$Fr > 10 : Ne\,Fr/Q \propto Fr \rightarrow P/q \propto d^2$

In this range, the hollow stirrer is less effective, because the power per unit gas throughput increases with the square of the scale (d^2).

Under these circumstances, which can be described by "small is beautiful", it can be clearly shown that hollow stirrers are not suitable for sucking in large amounts of gas on a full-scale. In this case the gas throughput and the stirrer power have to be separated by choosing an appropriate stirrer (Rushton turbine is a good choice!) and providing the gas from below the stirrer by a blower.

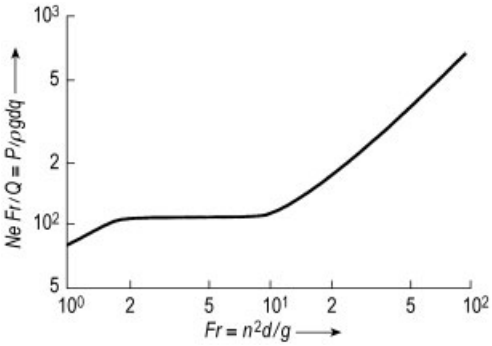

Figure 17. Dimensionless interdependence of P/q and d under different process conditions (Fr)

In transport-limited reactions in gas – liquid systems, mass transfer is usually dimensioned according to $P/V = \mathrm{idem}$ and $\mathrm{v} = q/S = \mathrm{idem}$; ($S = \pi D^2/4$). In scaling up, these conditions also speak in favor of decoupling the gas supply and stirrer speed, because processes with two mutually independent process parameters can be more easily optimized than those having only one process parameter.

However, there are many chemical reactions in the gas – liquid system in which the gas throughput does not play a role because microkinetics is rate determining. For such cases, the hollow stirrers, due to their dual role as stirrers and gas conveyers, are especially suited, particularly in high pressure chemical engineering. For details see [3, Section 4.12.1].

9. References

1. M. Zlokarnik: *Dimensional Analysis and Scale-up in Chemical Engineering,* Springer-Verlag Berlin, Heidelberg, New York 1991.
2. J. Pawlowski: *Die Ähnlichkeitstheorie in der physikalisch-technischen Forschung,* Springer-Verlag, Berlin-Heidelberg-New York 1971.
3. M. Zlokarnik: *Rührtechnik – Theorie und Praxis,* Springer-Verlag, Berlin-Heidelberg-New York 1999. M. Zlokarnik: *Stirring – Theory and Practice,* Wiley-VCH, Weinheim 2001.
4. M. Zlokarnik: *Scale-up – Modellübertragung in der Verfahrenstechnik,* Wiley-VCH, Weinheim 2000. M. Zlokarnik: *Scale-up in Chemical Engineering,* Wiley-VCH, Weinheim 2002.
5. M. Levin, M. Zlokarnik, in *Pharmaceutical Process Scale-Up,* Marcel Dekker, Inc., New York 2002.
6. M. Zlokarnik, *Ger. Chem. Eng.* **9** (1986) 314 – 320.
7. J. Pawlowski, *AIChE J.* **15** (1969) 303 – 305.
8. J. Pawlowski: Veränderliche Stoffgrößen in der Ähnlichkeitstheorie, Salle + Sauerländer, Aarau -Frankfurt/M. 1991.
9. H.-J. Henzler, *Chem.-Ing.-Tech.* **60** (1988) 1 – 8.

Transport Phenomena

Raymond W. Flumerfelt, Department of Chemical Engineering, Texas A & M University, College Station, Texas 77843, United States

Charles J. Glover, Department of Chemical Engineering, Texas A & M University, College Station, Texas 77843, United States

1. Foundations 8
1.1. Mathematical Preliminaries 9
1.1.1. Coordinate Systems 9
1.1.2. Vector and Tensor Operations 9
1.1.3. The Jacobian 10
1.1.4. Calculus of Vectors and Tensors . . . 11
1.1.5. Divergence Theorem 12
1.1.6. Kinematic Relations 12
1.1.7. Partial and Total Derivatives 13
1.1.8. Relation Between Different Time Derivatives 13
1.2. Basic Equations for Compositionally Homogeneous Systems 13
1.2.1. The Reynolds Transport Theorem . . 13
1.2.2. Conservation of Total Mass 14
1.2.3. Conservation of Linear Momentum . 14
1.2.4. Condition of Local Stress Equilibrium 15
1.2.5. Stress Tensor 15
1.2.6. Equation of Motion 15
1.2.7. Conservation of Angular Momentum 15
1.2.8. Mechanical Energy Accounting Equation 16
1.2.9. Conservation of Total Energy 16
1.2.10. Thermal Energy Accounting Equation 16
1.2.11. Forms of the Governing Equations . 17
1.2.12. Entropy Inequality 17
1.2.13. Linear Transport Fluxes and Relations 17
1.2.14. Transport Properties from MolecularTheories 18
1.2.15. Non-Newtonian Fluids 19
1.3. Summary of Basic Equations . . . 20
1.4. Boundary Conditions 21
1.5. Solution Philosophy 22
1.6. Dimensionless Equations of Change 22
2. Transport in Compositionally Homogeneous Systems 24
2.1. Heat Conduction in Solids 24
2.1.1. Heat Conduction Equations 24
2.1.2. Initial and Boundary Conditions . . . 25
2.1.3. Steady, One-Dimensional Conduction 26
2.1.4. Convective Boundary Conditions . . 27
2.1.5. Composite Systems 28
2.1.6. Steady Conduction with Heat Generation and in Multidimensions 28
2.1.7. Transient Heat Conduction – Lumped Capacity Systems 29
2.1.8. Transient Heat Conduction – More General Solutions 29
2.2. Steady, One-Dimensional Flows . . 32
2.2.1. Generalized Couette Flow 33
2.2.2. One-Dimensional Poiseuille Flows . 34
2.2.3. Flow in Channels of Arbitrary Cross Section 35
2.2.4. Poiseuille Flow of Non-Newtonian Fluids 36
2.2.5. Two-Phase Concentric Flow in a Tube – Segregated Flow 38
2.3. Multidimensional Momentum Transfer 39
2.3.1. Two-Dimensional Flows – Stream Function Equations 39
2.3.2. Creeping Flow Around a Sphere and Other Bodies of Revolution 40
2.3.3. Flow in Channels with Varying Cross Sections – Lubrication Analysis . . . 43
2.4. Coupled Momentum and Energy Transfer 46
2.4.1. Heat Transfer in Laminar Tube Flow 47
2.4.2. Momentum and Heat Transfer in Laminar Boundary Layers 49
2.4.3. Free Convection on a Vertical Plate . 53
2.5. Turbulent Momentum and Energy Transfer 54
2.5.1. Physical Characteristics 54
2.5.2. Time-Smoothed Momentum and Energy Balances for Turbulent Flow . . 57
2.5.3. Mixing Length Theories 58
2.5.4. Turbulent Heat Transfer in Tubes – Reynolds, Prandtl, and von Karman Analogies 59
2.5.5. Turbulent Momentum and Heat Transferon a Flat Plate 61
2.6. Summary of Heat-Transfer Relations 62

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_04

3. Transport in Multicomponent Systems 62
3.1. Diffusive Mass Transfer in Binary Systems 62
3.1.1. Species Mass Balances 62
3.1.2. Species Diffusion Fluxes 66
3.1.3. Fick's First Law of Diffusion 66
3.1.4. Special Cases of Diffusion 67
3.1.5. Diffusivities of Gases and Liquids . 68
3.1.6. Theoretical Foundations of Steady-State Measurement of Diffusion . . . 69
3.1.7. Diffusion with Homogeneous Chemical Reaction 72
3.1.8. Diffusion with Heterogeneous Reaction 73
3.1.9. Perspective 75
3.2. Convective Mass Transport 75
3.2.1. Gas Absorption in a Falling Film with Reaction 75
3.2.2. Mass Transfer in Laminar and Turbulent Boundary Layers 78
3.3. Mass Transfer Across Interfaces . 79
3.3.1. Mass-Transfer Coefficients 80
3.3.2. Functional Forms of Mass-Transfer Coefficient Relations 82
3.3.3. Relations for Low Mass-Transfer Rates 83
3.3.4. Relations for High Mass-Transfer Rates 84
4. Macroscopic Systems 86
4.1. Conservation of Total Mass 86
4.2. Conservation of Linear Momentum 87
4.3. Conservation of Angular Momentum 88
4.4. An Accounting of Mechanical Energy 88
4.5. Conservation of Total Energy . . . 89
4.6. An Accounting of Thermal Energy 90
4.7. The Second Law of Thermodynamics 90
4.8. Tabular Summary of the Macroscopic Equations 92
4.9. Some Examples 92
4.9.1. Pressure Drops and Temperature Changes for Closed Channel Flow . 92
4.9.2. Compressible Flow in a Tube 93
4.9.3. Heterogeneous Reaction in a Fluidized-Bed Reactor 94
5. References 96

Symbols

a, b	major and minor axes of spheroids
$\boldsymbol{a}$	acceleration
$\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$	arbitrary vectors
A	area
$\boldsymbol{A}$	area vector (magnitude and direction)
$\boldsymbol{A}, \boldsymbol{B}$	arbitrary second-order tensors
A_{c}	cross-sectional area
A_{n}	numerical parameters defined by Equation (2.174)
A_{w}	wall surface area
A_{ξ}	area perpendicular to ξ-direction
$\hat{A}$	Helmholtz free energy per unit mass
ATSB	A through stagnant B diffusion situation
b	flow geometry height
$b*$	dimensionless flow channel height, defined below Equation (2.174)
b_0, b_L	flow channel vertical dimensions at $x = 0$ and $x = L$, Figure 13
B	applied load, Figure 13
Bi	Biot number; Equation (2.19) and definition below Equation (2.27)
Br	Brinkman number, Equation (1.118a)
c	molar concentration, moles per volume
c_p	constant pressure heat capacity per unit mass
c_v	constant volume heat capacity per unit mass
$c_{\mathrm{A}}, c_{\mathrm{B}}, c_{\mathrm{i}}$	moles of A, B, i in a mixture per unit volume of mixture
c_{A}^*	dimensionless molar concentration of species A
$\bar{c}_{\mathrm{A}}$	time-averaged value of c_{A}
c_{A0}	molar concentration at interface
$c_{\mathrm{A}\infty}$	molar concentration of A in free stream, or in bulk fluid in contact with surface
C	constant
C_{D}	drag coefficient defined by Equation (2.155)
$C_{\mathrm{D}x}$	local drag coefficient defined in Equation (2.235)
$C_{\mathrm{D}L}$	overall drag coefficient defined by Equation (2.236)
C_k	coefficients in Equation (2.49)
$C_1, C_2, \hat{C}_1, \hat{C}_2$	integrating constants
d	diameter

d/dt	time derivative seen by an observer moving at arbitrary velocity; the total time derivative of a function
D/Dt	time derivative seen by observer maintaining constant material coordinates (following the motion of an element, Eq. 1.50)
D_{AB}	diffusivity of species A in B
D°_{AB}	diffusivity of infinitely dilute A in solvent B
Da_{δ}	Damköhler number based on film thickness, Equation (3.100)
Da_L	Damköhler number based on film length
e	height of wall roughness (protuberances)
e_{ijk}	permutation symbol, Equation (1.12)
E^2	differential operator, Equation (2.137)
$E(k)$	complete elliptic integral with argument k
ECD	equimolar counter diffusion situation
f	in Section 2.4.1, function of r defined by Equation (2.197); in Section 2.3.2, function of r defined by Equations (2.133) and (2.134); in Section 2.3.3, a function defined below Equation (2.162); friction factor defined by Equation (2.79)
F	a function of time (1.56)
$\boldsymbol{F}$	force, a vector
$\boldsymbol{F}_{\mathrm{B}}$	buoyancy force
$\boldsymbol{F}_{\mathrm{D}}$	drag force
$\boldsymbol{F}_{\mathrm{ext}}$	external force
$F_{\mathrm{in}}, F_{\mathrm{out}}$	volumetric flow rate entering and leaving a system
F_x, F_y, F_z	force components
F^*_x, F^*_y, F^*_z	dimensionless force components
$\mathfrak{F}$	in Chapter 1, an arbitrary scalar, vector, or tensor function of time and position; in Section 2.5.2, any linear time or spatial derivative operator
Fr	Froude number, Equation (1.114)
$\hat{F}$	conversion of mechanical energy to thermal energy by viscous dissipation, per unit mass (fluid friction losses)
$\boldsymbol{g}$	body force per unit mass or gravitational acceleration
$\boldsymbol{g}_{\mathrm{i}}, \boldsymbol{g}_{\mathrm{j}}, \boldsymbol{g}_{\mathrm{k}}$	i, j, k base vectors in a curvilinear coordinate system
g_x, g_y, g_z	components of gravitational force per unit mass, in x, y, z directions
G	scalar function in Equation (2.125)
Gr_x	local Grashoff number, Equation (2.273)
Gr_L	Grashoff number associated with surface of length L
$\hat{G}$	Gibbs free energy per unit mass
h	heat-transfer coefficient, Equation (2.6)
h_k	heat-transfer coefficient associated with surface at position x_k
h_{lm}	log mean heat-transfer coefficient, Equation (2.206)
$h_{x,\mathrm{lam}}$	local heat-transfer coefficient, laminar boundary layer
$h_{x,}{}^{\mathrm{turb}}$	local heat-transfer coefficient, turbulent boundary layer
h_L	overall heat-transfer coefficient over surface of length L, Equation (2.254)
$\bar{h}$	average heat-transfer coefficient over surface A
$\hat{H}$	enthalpy per unit mass
$\boldsymbol{i}_{\mathrm{i}}, \boldsymbol{i}_{\mathrm{j}}, \boldsymbol{i}_{\mathrm{k}}$	base vectors in a rectangular cartesian coordinate system
$\boldsymbol{i}_{\mathrm{x}}, \boldsymbol{i}_{\mathrm{y}}, \boldsymbol{i}_{\mathrm{z}}$	unit base vectors in x, y, z directions
$\boldsymbol{i}_{\mathrm{r}}, \boldsymbol{i}_{\theta}, \boldsymbol{i}_{\mathrm{z}}$	unit base vectors in cylindrical coordinates
$\boldsymbol{i}_{\mathrm{r}}, \boldsymbol{i}_{\theta}, \boldsymbol{i}_{\phi}$	unit base vectors in spherical coordinates
$\boldsymbol{I}$	identity tensor
$\mathrm{I}_{\Delta}, \mathrm{II}_{\Delta}, \mathrm{III}_{\Delta}$	first, second, and third scalar invariants of second-order tensor Δ, defined by Equations (1.107) – (1.109)
I_{u}	intensity of turbulence, Equation (2.280)
I_{ϕ}	a measure of the fluctuating part of ϕ in a turbulent flow, Equation (2.279)
$\boldsymbol{j}_{\mathrm{A}}$	mass diffusion flux of species A relative to $\boldsymbol{u}$, Equation (3.16)
$\tilde{\boldsymbol{j}}_{\mathrm{A}}$	mass diffusion flux of species A relative to $\tilde{\boldsymbol{u}}$, Equation (3.19)
J	Jacobian relating differential volume elements in two coordinate systems and defined by Equations (1.23) and (1.24)
$\mathrm{J}_n(\,)$	Bessel functions of the first kind with argument ( ) and $n = 0, 1, \ldots$
$\boldsymbol{J}_{\mathrm{A}}$	molar diffusion flux of species A, relative to $\boldsymbol{u}$, Equation (3.18)
$\tilde{\boldsymbol{J}}_{\mathrm{A}}$	molar diffusion flux of species A, relative to $\tilde{\boldsymbol{u}}$, Equation (3.14)

k	arbitrary constant, Equation (1.4); in Section 2.2.3, defined below Equation (2.94); in Section 2.5.3, proportionality constant in Equation (2.299); reaction rate constant, Equation (4.29)
k'	reaction rate constant, Equation (3.98)
k'_{y}	drift factor-corrected gas-phase mass-transfer coefficient
k_c	single-phase mass-transfer coefficient based on molar concentrations
$k_{\mathrm{c}L}$	average mass-transfer coefficient defined by Equation (3.112)
k_{G}	gas-phase mass-transfer coefficient based on gas-phase partial pressures, Equation (3.132)
k_x	liquid-phase mass-transfer coefficient based on liquid-phase mole fractions
k_y	gas-phase mass-transfer coefficient based on gas-phase mole fractions
k_1	first-order reaction rate constant
K	pressure gradient defined by Equation (2.68); overall reaction rate constant, Equation (4.31)
$K*$	dimensionless pressure gradient, defined by Equation (2.73)
K_c	overall two-phase mass-transfer coefficient based on liquid-phase molar concentrations
K_{G}	overall two-phase mass-transfer coefficient based on gas-phase partial pressures, Equation (3.134)
K_x	overall two-phase mass-transfer coefficient based on liquid-phase mole fractions
K_y	overall two-phase mass-transfer coefficient based on gas-phase mole fractions
$\hat{KE}$	kinetic energy per unit mass
l	characteristic length, Equation (2.82); in Section 2.5.3, the Prandtl mixing length
$\boldsymbol{L}$	angular momentum
L	characteristic length for a process or flow, Chapter 1; length of body, surface, or flow channel
m, n	material parameters for power-law non-Newtonian fluid, Equation (2.97)
n	coordinate in direction normal to surface, Equation (2.6)
$\boldsymbol{n}$	outwardly directed unit normal on surface enclosing volume V
$\boldsymbol{N}_{\mathrm{A}}$	molar flux (vector) of species A
$N_{\mathrm{A}y}$	mole flux of A in y-direction
$\bar{N}_{\mathrm{A}y}$	time-averaged value of $N_{\mathrm{A}y}$
Nu_{AB}	mass-transfer Nusselt number, $Nu_{\mathrm{AB}} = k_x d/c\, D_{\mathrm{AB}}$
Nu_d	Nusselt number based on diameter, Equation (2.202)
$Nu_{d,\mathrm{lm}}$	log mean Nusselt number, Equations (2.205) and (2.206)
Nu_L	overall Nusselt number associated with surface of length L, Equation (2.255)
Nu_x	local heat-transfer Nusselt number, Equation (2.251)
O ( )	order of magnitude of ( )
$\boldsymbol{p}$	linear momentum
p	pressure, the isotropic stress
p_0	reference pressure
$p_{\mathrm{A}}, p_{\mathrm{B}}$	partial pressure of species A, B
p_{H}	hydrostatic pressure, Equation (2.63)
$\bar{p}$	time-averaged pressure
P	dynamic pressure, Equation (2.62)
P_0	reference dynamic pressure
ΔP	pressure drop across flow channel in direction of flow
Pe	Péclet number, Equation (1.117)
Pe_L	mass-transfer Péclet number based on film length
Pe_δ	mass-transfer Péclet number based on film thickness, Equation (3.83)
Pr	Prandtl number ($\equiv \nu/\alpha$)
$\hat{PE}$	potential energy per unit mass
q	a dimensionless parameter defined by Equation (3.103)
$\boldsymbol{q}$	heat flux vector (heat flow rate per unit area)
$\bar{\boldsymbol{q}}$	time-averaged heat flux
q_n	heat flux component normal to surface, Equation (2.6)
q_r, q_θ, q_z	heat flux components in cylindrical coordinates
q_w	wall heat flux
q_x, q_y, q_z	heat flux components in x-, y-, and z-directions
$q_y^{(t)}$	turbulent heat flux component in y-direction, Equation (2.301)
q_ξ	heat flux in ξ-direction
Q	energy transferred as heat, Chapter 4; volumetric flow rate, Chapter 2

Q, $\hat{Q}$ volumetric flow rates in inner and outer regions, respectively, Section 2.2.5
$Q*$ defined by Equation (2.110)
$\hat{Q}$ heat transfer per unit mass
$\dot{Q}$ heat transfer rate
r molecular separation, Lennard – Jones function
$\boldsymbol{r}$ position vector
$r*$ dimensionless radial position
r_{A}, r_{B}, r_i mass rate of production of species A, B, i per unit volume due to reaction ($r_{\mathrm{A}} > 0$ for generation of A; $r_{\mathrm{A}} < 0$ for consumption of A)
r_{c} radial position of interface between inner and outer phases
r_c^* dimensionless radial position, defined below Equation (2.107)
r_{H} hydraulic radius defined below Equation (2.96)
r_{Y} radial position of yield surface
r_0, r_1 radii associated with cylinder or spherical surfaces
R ideal gas constant; in Section 2.3.3, the flow channel radius
R_{A}, R_{B}, R_i molar rate of production of species A, B, i per unit volume of mixture due to reaction ($R_{\mathrm{A}} > 0$ for production of A; $R_{\mathrm{A}} < 0$ for consumption of A)
R_{cond} conductive resistance, Equation (2.14)
$R_{\mathrm{conv},i}$ convective heat-transfer resistance, Equation (2.24)
Re Reynolds number, Equations (1.113) and (2.81)
Re_{cr} critical Reynolds number for laminar to turbulent transition
Re_d Reynolds number based on diameter d, defined below Equation (2.127)
Re_L Reynolds number based on length L
Re_x local Reynolds number, Equation (2.231)
Re_δ Reynolds number based on length scale δ
r, θ, z coordinates in cylindrical coordinate system
r, θ, ϕ coordinates in spherical coordinate system
s defined below Equation (2.98)
$\boldsymbol{s}$ displacement vector
S control volume surface
Sc Schmidt number ($\equiv \nu/\mathrm{D}_{\mathrm{AB}}$)
Sh Sherwood number ($\equiv k_x\, d/cD_{\mathrm{AB}}$)
Sh_L Sherwood number based on length L, Equation (3.95)
Sh_x local Sherwood number defined below Equation (3.110)
St Stanton number, Equation (2.311)
$\hat{\boldsymbol{S}}$ entropy per unit mass
$\dot{\boldsymbol{S}}_{\mathrm{gen}}$ entropy generation rate
$\dot{\boldsymbol{S}}_{\mathrm{in}}$ rate of entropy addition
t time
t_{exp} diffusion penetration theory exposure time, Equations (3.136) and (3.137)
$\boldsymbol{t}_{(n)}$ traction vector acting on surface A, Equation (1.67)
t_0 in Section 2.5.1, time over which quantities are averaged; elsewhere, a reference time
$t*$ dimensionless time, defined below Equation (2.27)
tr trace operation of a second-order tensor, Equation (1.8)
T absolute temperature
$\boldsymbol{T}$ total stress tensor, a symmetric second-order tensor
$T*$ dimensionless temperature, Equation (2.34)
T_{b} bulk average temperature, Equation (2.195)
T_{f} film temperature [$\equiv (T_{\mathrm{w}} + T_\infty)/2$] in Section 2.5.5 and Table 7; bulk fluid temperature
$T_{\mathrm{f}k}$ temperature of bulk fluid next to surface at x_k
T_{i} initial temperature
T_k temperature at position x_k
T_0 characteristic reference temperature for a process
T_{w} wall temperature
T_{w0}, T_{wL} wall temperatures at positions $z = 0$ and $z = L$
T_x, T_y, T_z, T_r solutions defined below Equations(2.54) – (2.56)
T_ξ time-averaged temperature at $y = \xi$
T_∞ temperature in free stream
T_0, T_1, T_2 temperatures at positions x_0, x_1, and x_2
Δ_{T} ratio of momentum and thermal boundary layer thicknesses ($\equiv \delta_{\mathrm{m}}/\delta_{\mathrm{T}}$)
ΔT_{lm} log mean temperature difference, Equation (2.207)

$\bar{T}$ average spatial temperature in a region, Section 2.1.7; time-averaged temperature, Section 2.5.2

$\boldsymbol{u}$, u velocity vector (momentum per unit mass) of magnitude u; in a mixture: a mass average velocity, Equation (3.8)

$\boldsymbol{u}'$ fluctuating part of velocity in turbulent field

u'_x, u'_y, u'_z fluctuating velocity components

u_{avg} average velocity

u_c characteristic velocity, Equation (2.268)

$\boldsymbol{u}_i$ velocity of species i in a mixture

$u_{interface}$ fluid velocity at a gas–liquid interface

u_{max} maximum velocity

u_x, u_y, u_z components of velocity in rectangular cartesian system

u_0 characteristic velocity, Equation (2.159)

u_∞ free stream velocity; see Figures 10, 11, 14, 15, 16

u_* friction velocity, Equation (2.286)

$u*$ dimensionless velocity

u^+ dimensionless velocity, Equation (2.284)

$\tilde{\boldsymbol{u}}$ mixture molar average velocity, Equation (3.12)

$\tilde{u}$ time-averaged velocity in turbulent field defined by Equation (2.277)

$\bar{u}_x$, $\bar{u}_y$, $\bar{u}_z$ time-averaged velocity components

u_z, $\hat{u}_z$ velocity components in inner and outer regions, respectively, Section 2.2.5

$\bar{u}_\xi$ time-averaged velocity at $y = \xi$

U characteristic velocity for a process or flow; velocity of moving surface, Figure 5; velocity of lower plate, Figure 13

$\hat{U}$ internal energy per unit mass

V volume

V_A molar volume of species A, volume per mole

V_0 material volume in the reference coordinates

$\hat{V}$ volume per unit mass ($\equiv 1/\varrho$)

$\boldsymbol{w}$ arbitrary velocity of an observer, Equation (1.55)

w flow geometry width

w_A mass fraction of species A in a mixture: $w_A = \varrho_A/\varrho$

W film width (in x direction), Figure 25; work, Equation (4.8)

$\mathfrak{W}_A$ rate of mass transfer, moles/time

$\dot{W}$ work rate (power)

$\hat{W}_{es}$ work done on the fluid at the entrance and exit points by extra stress, per unit mass

$\dot{W}_{np}$ nonpressure work rate, Section 4.5

$\hat{W}_s$ shaft work per unit mass

$\boldsymbol{x}$, x^i, x^j, x^k i, j, k components of a position vector $\boldsymbol{x}$ in a curvilinear coordinate system

x_A mole fraction of species A in a mixture: $x_A = c_A/c$

x_{Ae} mole fraction of species A in the liquid phase required for equilibrium with a given bulk gas-phase mole fraction, Figure 29

x_{Ai}, x_{Ab} mole fraction of species A at the interface in the liquid phase and in the bulk liquid phase, Figure 29

x_{cr} position on surface where boundary layer becomes turbulent, Figure 14

$x*$ dimensionless coordinate position

x_0, y_0 principal axes of elliptic cross section; defined above Equation (2.92)

x_0, y_0, z_0 linear half-thickness associated with rectangular parallelepiped or rectangular bar, Section 2.1.8; reference position, Section 2.2

Δx_k thickness of kth slab ($x_k - x_{k-1}$)

X_i, X_j, X_k material coordinates (y_i, y_j, y_k at a reference time) of an element of mass

y_A mole fraction of species A in a gas mixture

y_{Ae} mole fraction of species A in the gas phase required for equilibrium with a given bulk liquid-phase mole fraction, Figure 29

y_{Ai}, y_{Ab} mole fraction of species A at the interface in the gas phase and in the bulk gas phase, Figure 29

y_i, y_j, y_k i, j, k components of a vector $\boldsymbol{y}$ or position vector in a rectangular cartesian coordinate system

y' y coordinate measured from channel centerline, defined above Equation (2.76)

$y*$ dimensionless y coordinate

y^+ dimensionless distance from wall, Equation (2.285)

z_e thermal entry length defined above Equation (2.204)

z_1, z_2 elevation at points 1 and 2 in a flow conduit

$z*$ dimensionless z coordinate

Greek Symbols

α thermal diffusivity ($\equiv \lambda/\varrho\, c_{\mathrm{p}}$)

α_{KE} factor representing deviation from a flat velocity profile, Equation (4.6)

α_{mom} factor representing deviation from a flat velocity profile, Equation (4.4)

β dimensionless quantity defined by Equation (2.43); isobaric coefficient of thermal expansion, Equation (2.185)

β_n eigenvalues given by Equations (2.48), (2.51), or (2.53) with $n = 1$, 2,...

β_∞ coefficient of thermal expansion at T_∞

γ defined below Equation (3.69)

$\dot{\gamma}$ shear rate, Equation (1.110c)

Δ rate of deformation tensor, Equation (1.105)

δ film thickness

δ, δ_{ij} Kronecker delta defined by Equations (1.10) and (1.11)

δ_{c} boundary layer thickness for mass transfer analogous to Equations (2.211) and (2.237)

δ_{m} momentum boundary layer thickness, Equation (2.211)

δ_{T} thermal boundary layer thickness, Equation (2.237)

ε fractional penetration of a gas into a film, used for nondimensionalizing the penetration, Equation (3.81a); smallness parameter, Equation (2.158); in Section 2.5.3, the eddy diffusivity; in Section 3.2.1, a smallness parameter defined by Equation (3.84); Lennard–Jones potential function energy parameter, Chapter 1

$\varepsilon_{\mathrm{A}}, \varepsilon_{\mathrm{B}}, \varepsilon_{\mathrm{AB}}$ Lennard–Jones potential function energy parameters for species A and B and for the mixture, $\varepsilon_{\mathrm{AB}} = (\varepsilon_{\mathrm{A}}\varepsilon_{\mathrm{B}})^{1/2}$

$\varepsilon_{\mathrm{c}}, \varepsilon_{\mathrm{m}}, \varepsilon_{\mathrm{T}}$ eddy mass, momentum, and thermal diffusivities, Equations (3.113), (2.296), and (2.297)

$\varepsilon_{\mathfrak{M}}$ eddy diffusivity associated with transfer of quantity $\mathfrak{M}$ by turbulent fluctuations

ε_{s} surface shear viscosity

ξ dependent variable, Equation (2.35)

η viscosity; catalyst effectiveness factor, Equation (4.28)

η, $\hat{\eta}$ in Section 2.2.5, viscosities of core fluid and outer fluid, respectively; in Section 2.3.1, viscosities of continuous and drop phases, respectively

η_{B} viscosity of solvent B, Equation (3.29)

η_{p} plastic viscosity in Bingham plastic model

$\eta_\infty, \eta_{\mathrm{w}}$ viscosities evaluated at bulk and wall temperatures, respectively

θ angle subtended by two vectors, Equation (1.5)

κ heat capacity ratio, c_p/c_v; equal to b_L/b_0 in Section 2.3.3

Λ interference parameter associated with hindered settling (see Eq. 2.157)

λ thermal conductivity

$\bar{\lambda}$ average thermal conductivity similar to Equation (2.10) with T replacing T_1

$\bar{\lambda}_1$ average thermal conductivity, Equation (2.10)

$\bar{\lambda}_k$ average thermal conductivity, Equation (2.10) with T_k replacing T_1

ν kinematic viscosity ($\equiv \mu/\varrho$)

ξ coordinate direction; dummy integration variable; y position associated with edge of sublayer in turbulent flow

ϱ density, mass per unit volume

$\varrho_{\mathrm{A}}, \varrho_{\mathrm{B}}, \varrho_i$ mass per unit volume of A, B, i in a mixture

ϱ_{s} density of sphere

ϱ_∞ density at T_∞

σ hard-sphere diameter, Equation (1.99); Lennard–Jones potential function separation parameter, Equation (1.100)

$\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}, \sigma_{\mathrm{AB}}$ Lennard–Jones hard-sphere diameters for species A and B and for the mixture, $\sigma_{\mathrm{AB}} = (\sigma_{\mathrm{A}} + \sigma_{\mathrm{B}})/2$

$\boldsymbol{\tau}$ extra stress second-order tensor, Equation (1.84)

τ_{ij} stress components

$\tau_{rz}, \hat{\tau}_{rz}$ stress components in inner and outer regions, respectively, Section 2.2.5

τ_{w} magnitude of wall shear stress

τ_{Y} yield stress in Bingham plastic model

$\tau_{yx}^{(t)}$ turbulent stress component, Equation (2.300)

τ_0 shear stress at $y = 0$

Φ arbitrary flux vector or tensor, Equation (1.39); in Equations (2.289) and (2.290), represents a scalar, vector, or tensor function; rate of heat production and rate of radiation energy absorption, per unit volume

Φ_v rate of conversion of mechanical energy to thermal energy by viscous dissipation, per unit mass (Eq. 1.97) for a constant-density Newtonian material

Φ_0, Φ_1, Φ_2 coefficients for the general viscous fluid, Equation (1.106)

ϕ function described by Equations (2.90) and (2.91); association factor, Equation (3.29); fractional flow, Equation (2.109)

$\phi(y*, z*)$ in Section 3.2.1, the c_A solution from Equation (3.91)

χ ratio of concentric cylinder inner and outer radii

ψ stream function defined by Equation (2.125); in Section 2.4.1, function defined by Equation (2.210); in Equations (2.289) and (2.290), represents a scalar, vector, or tensor function

Ω_D collision integral correction factor for intermolecular interactions, Equation (3.28)

Ω_v collision integral correction factor for intermolecular interactions, Equations (1.101) and (1.102)

$\partial/\partial t$ time derivative seen by an observer maintaining constant spatial coordinates (position)

Others

∇ Del operator for spatial differentiation, defined by Equation (1.29) for rectangular cartesian coordinate systems

∇^2 Laplacian operator, $\nabla^2 = \nabla \cdot \nabla$, Equation (1.38)

$\sim$ order of magnitude estimate

1. Foundations

The objective of this article is to provide the foundations for understanding and analyzing phenomena that occur in continua as a result of driving forces in momentum, mass, and heat transfer. The descriptions can be concerned with continua on either differential or macroscopic scales; the results can be used for process analysis or for design to improve the efficiency of existing processes and to develop new processes.

The analysis of transport processes relies on a small set of fundamental laws. These laws, together with specific information and data about the properties and behavior of the process materials and about the constitutive relations between driving forces and their resulting fluxes, are used to provide a mathematical description of the behavior of the process or phenomenon. This mathematical description, together with process constraints such as geometry and boundary conditions, defines the mathematical problem. Its solution, then, is approached by using a variety of mathematical and engineering strategies for solutions to algebraic, ordinary, and partial differential equations.

The governing laws that establish the behavior of continua are the conservation laws for mass, energy, and linear and angular momentum (and their subset accounting equations for species mass, mechanical energy, and thermal energy) plus the second law of thermodynamics. These laws require a knowledge of the properties and behavior of materials in the form of (1) volumetric (pressure – volume – temperature) property relations and data; (2) thermodynamic property relations and data (such as internal energy, enthalpy, and entropy); and (3) transport flux relations and data (for heat, mass, and momentum fluxes in terms of their driving forces: temperature gradient, mole fraction gradient, and velocity gradient).

The foundations establish conventions and identities that are used throughout this article: (1) kinematic relationships used in describing the motion of continua, (2) development of the conservation laws and their corollary equations related to mechanical and thermal energy balances, (3) linear transport relations for heat and momentum fluxes, and (4) discussion of how this material is combined in a methodology establishing quantitative models for describing the behavior of continua. Additionally, the molecular foundations for transport properties of materials are discussed. Theories for these properties, used to quantify the relations between the

transport fluxes and their driving forces, work reasonably well for gases to calculate properties such as thermal conductivity, viscosity, and diffusivity. However, for more dense phases in which intermolecular interactions play a significant role the theories serve primarily as a guiding framework for empirical relations. Finally, a brief discussion of non-Newtonian fluids and their constitutive relations is presented.

After development of the basic equations, boundary conditions and a solution philosophy for attacking problems are discussed briefly. Also, the equations of change in dimensionless form are presented as the basis for dimensionless group correlations and analogies between different modes of transport.

Chapter 2 is limited to compositionally homogeneous systems, whereas transport in multicomponent systems is discussed in Chapter 3.

1.1. Mathematical Preliminaries

1.1.1. Coordinate Systems

The study of transport processes is concerned with scalars, which are represented by a single number (their magnitude), and with vectors and tensors, which have directions in space as well as magnitude associated with them.

To describe vectors and tensors a coordinate system is defined for representing positions in three-dimensional space (direction and magnitude). A variety of systems are possible and can be used. Each coordinate system has defined base vectors, and all other vectors can be described as linear combinations of these base vectors. Any three noncoplanar vectors can be used as base vectors. However, orthogonal vectors (mutually perpendicular) that form a right-handed system are normally used. Base vectors are cartesian if they are of unit length. If they are both orthogonal and cartesian, then they are orthonormal. Base vectors are rectangular, cartesian if they are orthonormal and everywhere the same (in both magnitude and direction). In general, the base vectors for a coordinate system are not independent of position. For example, the base vectors for a cylindrical or spherical system vary with position (in a cylindrical system, the radius vector changes direction as the angular coordinate varies and the z coordinate is held constant). A coordinate system is referred to as curvilinear if it is not rectangular.

1.1.2. Vector and Tensor Operations

Useful references for vector and tensor notation, and their operation are given in [1] and [2]. The relations presented below summarize operations for *rectangular cartesian systems* for the purpose of setting notation and conventions.

Accordingly, any vector $\boldsymbol{a}$ may be written as a linear combination of its components and base vectors in the following way:

$$\boldsymbol{a} = a_i \boldsymbol{i}_i \equiv \sum_{i=1}^{3} a_i \boldsymbol{i}_i \tag{1.1}$$

A second-order tensor $\boldsymbol{A}$ is written in terms of its base vectors and components in the following way:

$$\boldsymbol{A} = A_{ij} \boldsymbol{i}_i \boldsymbol{i}_j \tag{1.2}$$

Here summation occurs over both indices i and j. The addition of two vectors is a vector that is obtained by adding corresponding components:

$$\boldsymbol{a} + \boldsymbol{b} = \sum_{i=1}^{3} (a_i + b_i)\, \boldsymbol{i}_i \tag{1.3}$$

Multiplication of a vector or tensor by a scalar k is performed by multiplying each component by that scalar:

$$k\boldsymbol{a} = k\,(a_i \boldsymbol{i}_i) = (k a_i)\, \boldsymbol{i}_i \tag{1.4}$$

The *dot product* of two vectors (which is a scalar) by definition (→ Mathematics in Chemical Engineering) is the product of their magnitudes and the cosine of their subtended angle, which implies a sum of the products of the corresponding components (for an orthonormal system):

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}|\,|\boldsymbol{b}| \cos\theta = a_i\, b_i \tag{1.5}$$

The dot product of a vector and tensor is calculated by a similar sum over adjacent indices according to

$$\begin{aligned} \boldsymbol{a} \cdot \boldsymbol{B} &= a_i\, B_{ij} \boldsymbol{i}_j \\ \boldsymbol{B} \cdot \boldsymbol{a} &= B_{ij}\, a_j \boldsymbol{i}_i \end{aligned} \tag{1.6}$$

The components thus obtained are analogous to the elements of a vector obtained from the matrix multiplication of a vector and a square matrix. A different result is obtained for pre-versus post-multiplication. The dot product of two second-order tensors is a second-order tensor given by

$$\boldsymbol{A}\cdot\boldsymbol{B} = A_{ij}\,B_{jk}\boldsymbol{i}_i\boldsymbol{i}_k \tag{1.7}$$

The *trace* of a second-order tensor is a scalar given by the sum of its diagonal terms

$$\mathrm{tr}\,(\boldsymbol{A}) = A_{ii} \tag{1.8}$$

and the trace of the dot product of two second-order tensors is a double sum performed in the following way:

$$\mathrm{tr}\,(\boldsymbol{A}\cdot\boldsymbol{B}) = A_{ij}\,B_{ji} \tag{1.9}$$

The identity second-order tensor is defined such that

$$\delta = \delta_{ij}\,\boldsymbol{i}_i\,\boldsymbol{i}_j \tag{1.10}$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases} \tag{1.11}$$

This identity tensor, either pre- or post-dotted with a vector, returns the original vector. Likewise, when it is dotted with a second-order tensor, the original second-order tensor is obtained.

The *permutation symbol* e_{ijk} is defined such that

$$e_{ijk} = \begin{cases} 1 & \text{for an even permutation of } 123 \\ 0 & \text{when any two indices are repeated} \\ -1 & \text{for an odd permutation of } 123 \end{cases} \tag{1.12}$$

The *cross product* of two vectors is then defined by

$$\boldsymbol{a}\times\boldsymbol{b} = a_i\,b_j\,e_{ijk}\boldsymbol{i}_k \tag{1.13}$$

and the cross product of a vector and a second-order tensor is similarly given by

$$\boldsymbol{a}\times\boldsymbol{B} = a_i\,B_{jm}\,e_{ijk}\boldsymbol{i}_k\boldsymbol{i}_m \tag{1.14}$$

Two convenient identities between the permutation symbol and the Kronecker delta are

$$e_{ijk}\,e_{mnk} = (\delta_{im}\,\delta_{jn} - \delta_{in}\,\delta_{jm}) \tag{1.15}$$

and

$$e_{ijk}\,e_{mjk} = 2\,\delta_{im} \tag{1.16}$$

The *dyadic* or *tensor product* of two vectors is a second-order tensor whose dot product with a vector is given by

$$(\boldsymbol{a}\boldsymbol{b})\cdot\boldsymbol{c} = \boldsymbol{a}\,(\boldsymbol{b}\cdot\boldsymbol{c}) \tag{1.17}$$

In terms of index notation the tensor product $\boldsymbol{a}\,\boldsymbol{b}$ is written

$$\boldsymbol{a}\boldsymbol{b} = a_i\,b_j\boldsymbol{i}_i\boldsymbol{j}_j \tag{1.18}$$

The *transpose* $\boldsymbol{A}^{\mathrm{T}}$ of a second-order tensor is defined so that the ij component of $\boldsymbol{A}$ is the ji component of the second-order tensor $\boldsymbol{A}^{\mathrm{T}}$. Accordingly,

$$\boldsymbol{A}^{\mathrm{T}} = A_{ij}\boldsymbol{i}_j\boldsymbol{i}_i \tag{1.19}$$

1.1.3. The Jacobian

The Jacobian is the ratio of differential volume elements represented in two coordinate systems. It has special value when one of the coordinate systems involves material coordinates and the other involves the spatial coordinate system discussed below in kinematics. In rectangular cartesian coordinates, the differential volume element can be written as

$$\mathrm{d}V = \mathrm{d}y_1\mathrm{d}y_2\mathrm{d}y_3 \tag{1.20}$$

which can be expressed equivalently in terms of the triple scalar product:

$$\mathrm{d}V = \mathrm{d}y_1\boldsymbol{i}_1\cdot(\mathrm{d}y_2\boldsymbol{i}_2\times\mathrm{d}y_3\boldsymbol{i}_3)$$

A position vector in space can be represented in terms of a rectangular coordinate system or in terms of an alternate system according to

$$\boldsymbol{r} = y_i\boldsymbol{i}_i = x_i\boldsymbol{g}_i \tag{1.21}$$

where x_i are the components of the position vector in the alternate system with base vectors $\boldsymbol{g}_i$. Each of the cartesian coordinates can be expressed as a function of the other system coordinates, $y_i = y_i\,(x_1, x_2, x_3)$. Then, the differential volume element can be expressed in terms of the

triple scalar product of three vectors written in terms of the derivatives with respect to each of the new coordinate system's three coordinates:

$$\mathrm{d}V = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}_1}\mathrm{d}x_1 \cdot \left(\frac{\partial \boldsymbol{y}}{\partial x_2}\mathrm{d}x_2 \times \frac{\partial \boldsymbol{y}}{\partial x_3}\mathrm{d}x_3\right) \tag{1.22}$$

or

$$\mathrm{d}y_1\mathrm{d}y_2\mathrm{d}y_3 = J\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 \tag{1.23}$$

and the Jacobian is the determinant of the partial derivatives relating changes in coordinates between the two systems:

$$J \equiv \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \frac{\partial y_3}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_3}{\partial x_2} \\ \frac{\partial y_1}{\partial x_3} & \frac{\partial y_2}{\partial x_3} & \frac{\partial y_3}{\partial x_3} \end{vmatrix} \tag{1.24}$$

The Jacobian only exists if there is a unique transformation between the two coordinate systems.

1.1.4. Calculus of Vectors and Tensors

In transport phenomena generally, two kinds of vector and tensor derivatives exist with respect to time and with respect to position.

Derivatives with Respect to Time. The derivative of a vector that is a function of time is a vector; a derivative of a second-order tensor with respect to time is a second-order tensor. Derivatives are obtained straightforwardly by differentiating each of their components with respect to time:

$$\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \frac{\mathrm{d}a_i}{\mathrm{d}t}\boldsymbol{i}_i \tag{1.25}$$

$$\frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} = \frac{\mathrm{d}A_{ij}}{\mathrm{d}t}\boldsymbol{i}_i\boldsymbol{i}_j \tag{1.26}$$

Differentiating products of vectors is similar to differentiating products of scalars:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{a}\cdot\boldsymbol{b}) = \frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t}\cdot\boldsymbol{b} + \boldsymbol{a}\cdot\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}t} \tag{1.27}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{a}\times\boldsymbol{b}) = \frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t}\times\boldsymbol{b} + \boldsymbol{a}\times\frac{\mathrm{d}\boldsymbol{b}}{\mathrm{d}t} \tag{1.28}$$

Derivatives with Respect to Position. The *del operator* ∇ is defined to differentiate quantities that are functions of position. In rectangular cartesian coordinate systems the del operator is

$$\nabla \equiv \boldsymbol{i}_i\frac{\partial}{\partial y_i} \tag{1.29}$$

The del operator can operate in dyadic, dot, or cross product forms on scalars, vectors, or tensors. As a dyadic operation, *grad* (φ) is

$$\nabla\varphi = \frac{\partial\varphi}{\partial y_i}\boldsymbol{i}_i \tag{1.30}$$

and grad ($\boldsymbol{a}$)

$$\nabla\boldsymbol{a} = \frac{\partial a_j}{\partial y_i}\boldsymbol{i}_i\boldsymbol{i}_j \tag{1.31}$$

For grad ($\boldsymbol{a}$), note the order of the index corresponding to the del operator with respect to that corresponding to the vector. The opposite order is sometimes used in the literature, and care must be taken to follow a consistent convention throughout a given work.

The *dot product* of the del operator with a vector $\boldsymbol{a}$ or tensor $\boldsymbol{A}$ is given by

$$\nabla\cdot\boldsymbol{a} = \frac{\partial a_i}{\partial y_i} \tag{1.32}$$

$$\nabla\cdot\boldsymbol{A} = \frac{\partial A_{ij}}{\partial x_i}\boldsymbol{i}_j \tag{1.33}$$

and is called the *divergence* of the vector or tensor. Again, note the order of the indices corresponding to the del operator and the second-order tensor. The divergence of a vector is a scalar, so the order is immaterial. The del operator can also be crossed with a vector or tensor, and these operations are given by

$$\nabla\times\boldsymbol{a} = \frac{\partial a_j}{\partial y_i}e_{ijk}\boldsymbol{i}_k \tag{1.34}$$

$$\nabla\times\boldsymbol{A} = \frac{\partial}{\partial y_i}A_{jm}e_{ijk}\boldsymbol{i}_k\boldsymbol{i}_m \tag{1.35}$$

and called the *curl* of the vector or tensor. Here, only one definition with respect to the order of the indices seems to exist in the literature.

Because of its dual vector – operator role, the del operator may be dotted with one entity and operate on (differentiate) another. For example, in the thermal energy equation in Section 1.2 (Eq. 1.96), the term

$$\boldsymbol{u}\cdot\nabla F \tag{1.36}$$

appears, in which the del operator is pre-dotted with the velocity vector $\boldsymbol{u}$ but operates (by differentiation) on the temperature (see also Eq. 1.53):

$$\boldsymbol{u}\cdot\nabla T = u_i \frac{\partial T}{\partial y_i} \tag{1.37}$$

The *Laplacian* is another operator, a scalar operator, defined by $\nabla \cdot \nabla$. For rectangular cartesian systems it is given by

$$\nabla^2 \varphi = \nabla \cdot \nabla \varphi = \frac{\partial^2 \varphi}{\partial y_i \, \partial y_i} \tag{1.38}$$

1.1.5. Divergence Theorem

The divergence theorem is extremely important in deriving and understanding the basic transport relations. This theorem equates an integral over the entire volume of a region of space enclosed by a closed surface to an integral over that surface:

$$\int_V \nabla \cdot \boldsymbol{\Phi} \mathrm{d}V = \int_A \boldsymbol{n} \cdot \boldsymbol{\Phi} \, \mathrm{d}A \tag{1.39}$$

In this equation, $\boldsymbol{\Phi}$ can be a vector or a second-order tensor. If it is an isotropic second-order tensor ($p\boldsymbol{I}$, for example), then a form involving a scalar function is obtained:

$$\int_V \nabla p \mathrm{d}V = \int_A p\boldsymbol{n} \, \mathrm{d}A \tag{1.40}$$

In transport phenomena, $\boldsymbol{\Phi}$ represents the flux of an extensive property and the surface integral represents the rate at which the property (as written here using the outward normal vector $\boldsymbol{n}$) is leaving the volume element across the surface. With the del operator defined as above, the surface integral must be written with the surface outward normal vector pre-dotting the quantity $\boldsymbol{\Phi}$.

1.1.6. Kinematic Relations

A fundamental concept of kinematics, the description of motion without reference to its cause, is that of material coordinates. Given a fluid or material undergoing motion (translation, rotation, deformation), each particle of that material has a current position $\boldsymbol{y}$

$$\boldsymbol{y} = y_i \boldsymbol{i}_i \tag{1.41}$$

Also, each element of the material can be identified with the position it had at some reference time ($t = 0$, for example):

$$\boldsymbol{X} = X_i \boldsymbol{i}_i \tag{1.42}$$

The *spatial coordinates* y_i are used to identify a position in space, and the *material coordinates* X_i are used to identify elements of the material. Furthermore, the position in space can be defined in terms of the element of the material that resides there (identified by its position at time $t = 0$, its reference position) at time t:

$$\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{X}, t) \tag{1.43}$$

Inversely, the material coordinates $\boldsymbol{X}$ can be defined in terms of the spatial coordinates and time t; each element of mass can be defined in terms of its location at time t

$$\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{y}, t) \tag{1.44}$$

This inversion depends on $0 < J < \infty$ where J is the Jacobian.

Every quantity expressed in terms of spatial coordinates and time can just as well be expressed in terms of material coordinates and time. For example, for a flowing fluid the density, velocity, etc., vary with position

$$\varrho = \varrho(\boldsymbol{y}, t) \tag{1.45}$$

$$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{y}, t) \tag{1.46}$$

or, equivalently, vary from material element to material element:

$$\varrho = \varrho(\boldsymbol{X}, t) \tag{1.47}$$

$$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{X}, t) \tag{1.48}$$

1.1.7. Partial and Total Derivatives

The partial derivative with respect to time [of a function $\mathfrak{F}(\boldsymbol{x}, t)$], if position is held constant, is written as

$$\frac{\partial \mathfrak{F}(\boldsymbol{x}, t)}{\partial t} \equiv \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{x}} \tag{1.49}$$

and represents changes in the variable of interest with time at a fixed point in space.

The partial derivative with respect to time [of a function $\mathfrak{F}(\boldsymbol{X}, t)$], if the material coordinates are kept constant, is called the *material derivative* or *substantial derivative*, and represents changes observed while following the motion of the material:

$$\frac{\mathrm{D}\,\mathfrak{F}(\boldsymbol{X}, t)}{\mathrm{D}t} \equiv \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{X}} \tag{1.50}$$

Note that if $\mathfrak{F}$ is the spatial coordinate of an element of fluid, then $\mathrm{D}\,\mathfrak{F}/\mathrm{D}\,t = \mathrm{D}\,\boldsymbol{x}/\mathrm{D}\,t$ is the change in position with time of a material element, i.e., the fluid velocity:

$$\boldsymbol{u} \equiv \left(\frac{\partial \boldsymbol{x}}{\partial t}\right)_{\boldsymbol{X}} = \frac{\mathrm{D}\boldsymbol{x}}{\mathrm{D}\,t} \tag{1.51}$$

The *total derivative* represents changes that would be observed while moving around at an arbitrary velocity $\boldsymbol{w}$; it is represented by $\mathrm{d}\mathfrak{F}/\mathrm{d}t$. This velocity $\boldsymbol{w}$ then represents the total derivative of the position vector:

$$\boldsymbol{w} = \mathrm{d}\boldsymbol{x}/\mathrm{d}t$$

1.1.8. Relation Between Different Time Derivatives

A property $\mathfrak{F}$ written in terms of $\boldsymbol{x}$ and t has the total differential

$$\mathrm{d}\mathfrak{F} = \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{x}} \mathrm{d}t + \mathrm{d}x_i \left(\frac{\partial \mathfrak{F}}{\partial x_i}\right)_t \tag{1.52}$$

from which the relation between the substantial and partial derivatives and $\boldsymbol{u}$ may be determined:

$$\left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{X}} = \frac{\mathrm{D}\,\mathfrak{F}}{\mathrm{D}t} = \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{x}} + \boldsymbol{u}\cdot\nabla\mathfrak{F} \tag{1.53}$$

In an analogous way, the total derivative is related to the partial and substantial derivatives and to $\boldsymbol{w}$ by

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}t} = \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{x}} + \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\cdot\nabla\mathfrak{F} \tag{1.54}$$

or

$$\frac{\mathrm{d}\mathfrak{F}}{\mathrm{d}t} = \left(\frac{\partial \mathfrak{F}}{\partial t}\right)_{\boldsymbol{x}} + \boldsymbol{w}\cdot\nabla\mathfrak{F} \tag{1.55}$$

A physical interpretation of the three derivatives follows [3]. An observer interested in determining the changes in the concentration of fish in a river c_f can do so while staying at a fixed position in the river $\partial c_f/\partial t$, while drifting in a canoe at the velocity of the river $\mathrm{D}\,c_f/\mathrm{D}\,t$, or while moving around in a motorboat with velocity $\boldsymbol{w}\,(\mathrm{d}c_f/\mathrm{d}t)$.

1.2. Basic Equations for Compositionally Homogeneous Systems [3–7]

1.2.1. The Reynolds Transport Theorem

In deriving the conservation equations of mass, energy, and momentum for a continuum, a volume (system) that always encloses the same elements of mass should be defined. This volume must then move and deform with the fluid, which is convenient because no convected input or output term need be considered. This concept, and the substantial derivative and Jacobian from kinematics, are very useful in deriving the desired conservation equations.

If $\mathfrak{F}(\boldsymbol{x}, t)$ is a function of position and time (scalar, vector, or tensor), representing the amount of the desired property per unit volume, and $V(t)$ is the closed volume moving with the fluid, then

$$F(t) \equiv \int_V \mathfrak{F}(\boldsymbol{x}, t)\,\mathrm{d}V \tag{1.56}$$

represents the total amount of the desired property contained within the system $\boldsymbol{V}(t)$. For the conservation equation, then, the accumulation rate of the property within the system is required $\mathrm{D}\,F(t)/\mathrm{D}\,t$. Here the substantial derivative must be used because it is the derivative "following the motion." Then

$$\frac{\mathrm{D}F}{\mathrm{D}t} = \frac{\mathrm{D}}{\mathrm{D}t}\int_V \mathfrak{F}(\boldsymbol{x}, t)\,\mathrm{d}V \tag{1.57}$$

However, the derivative cannot be taken across the integral in this form because the volume is a function of time. This difficulty is circumvented by transforming the system to the reference coordinates, differentiating, and then transforming back. In the reference coordinates, the volume element is a constant. Now,

$$\mathrm{d}V = J\mathrm{d}V_0 \tag{1.58}$$

So

$$\frac{\mathrm{D}F}{\mathrm{D}t} = \frac{\mathrm{D}}{\mathrm{D}t}\int_{V_0} \mathfrak{F}[\boldsymbol{x}(\boldsymbol{X}, t), t]\, J\,\mathrm{d}V_0$$

$$= \int_{V_0}\left(\frac{\mathrm{D}\mathfrak{F}}{\mathrm{D}t}J + \mathfrak{F}\frac{\mathrm{D}J}{\mathrm{D}t}\right)\mathrm{d}V_0 \tag{1.59}$$

$$\frac{\mathrm{D}J}{\mathrm{D}t} = J\nabla\cdot\boldsymbol{u} \text{ so}$$

$$\frac{\mathrm{D}F}{\mathrm{D}t} = \int_{V_0}\left(\frac{\mathrm{D}\mathfrak{F}}{\mathrm{D}t} + \mathfrak{F}\nabla\cdot\boldsymbol{u}\right) J\,\mathrm{d}V_0 \tag{1.60}$$

Finally, converting back to spatial coordinates gives

$$\frac{\mathrm{D}F}{\mathrm{D}t} = \int_V \left(\frac{\mathrm{D}\mathfrak{F}}{\mathrm{D}t} + \mathfrak{F}\nabla\cdot\boldsymbol{u}\right)\mathrm{d}V \tag{1.61}$$

which is known as the *Reynolds transport theorem* (RTT).

1.2.2. Conservation of Total Mass

In the absence of nuclear conversions, total mass is conserved; no generation or consumption occurs. (The term balance is commonly used to apply to both conservation and accounting equations. In this section, and in Chapter 4 on macroscopic systems, the term "accounting" is used for properties that may be generated or consumed and the term "conservation" for properties that may not be generated or consumed.) Now, for the volume that moves and deforms with the flowing fluid, as described above, no mass enters or leaves the system, so the mass conservation equation is

$$\frac{\mathrm{D}}{\mathrm{D}t}\int_V \varrho\,\mathrm{d}V = 0 \tag{1.62}$$

By the Reynolds transport theorem then,

$$\int_V \left(\frac{\mathrm{D}\varrho}{\mathrm{D}t} + \varrho\nabla\cdot\boldsymbol{u}\right)\mathrm{d}V = 0 \tag{1.63}$$

The volume chosen as the system is completely arbitrary, so this result implies

$$\frac{\mathrm{D}\varrho}{\mathrm{D}t} = -\varrho\nabla\cdot\boldsymbol{u} \tag{1.64}$$

or, in terms of the partial derivative,

$$\frac{\partial\varrho}{\partial t} = -\nabla\cdot\varrho\boldsymbol{u} \tag{1.65}$$

This equation (in either form) is called the continuity equation (see also, → Fluid Mechanics, Chap. 2.1.). Note that if $\mathfrak{F}$ in Equation (1.61) is proportional to ϱ (i.e., $\mathfrak{F} = \varrho G$), then by continuity

$$\frac{\mathrm{D}}{\mathrm{D}t}\int_V \varrho G\,\mathrm{d}V = \int_V \varrho\frac{\mathrm{D}G}{\mathrm{D}t}\mathrm{d}V \tag{1.66}$$

1.2.3. Conservation of Linear Momentum

An equation for the conservation of momentum or energy for a continuum must be able to express the forces that are exerted on an element in the continuum. These forces are of two kinds:

1) *Body forces* act upon the entire volume (reach inside the volume and act upon each element) and may be gravitational, electromagnetic, etc.
2) *Contact* (*or internal*) *forces* act at the surface of the element of the continuum.

Cauchy's stress principle states that if $\boldsymbol{t}_{(n)}$ is the force per unit area (*stress vector* or *traction*) at a point on the surface of a system volume (surface orientation indicated by $\boldsymbol{n}$), then $\boldsymbol{t}_{(n)}$ is a function of $\boldsymbol{x}$, t, and the orientation of the surface. The stress vector at a point $\boldsymbol{x}$ at a particular time is not sufficient; the surface upon which the stress vector acts must also be defined.

Note that $\boldsymbol{t}_{(n)}$ is a vector force per area (the stress vector or traction vector) and that $\boldsymbol{t}_{(n)}\,\mathrm{d}A$

is the force acting from outside the surface. For a volume (with a surface) the total force acting on this volume due to these contact forces is the integral over the entire surface A:

$$\int_A \boldsymbol{t}_{(\boldsymbol{n})} \mathrm{d}A \tag{1.67}$$

Body forces can be summed by a volume integral ($\boldsymbol{g}$ is force per unit mass)

$$\int_V \varrho \boldsymbol{g} \mathrm{d}V \tag{1.68}$$

and the conservation of linear momentum (for a system that moves with the material) can be expressed as

$$\frac{\mathrm{D}}{\mathrm{D}t} \int_V \varrho \boldsymbol{u} \mathrm{d}V = \int_V \varrho \boldsymbol{g} \mathrm{d}V + \int_A \boldsymbol{t}_{(\boldsymbol{n})} \mathrm{d}A \tag{1.69}$$

This is known as Euler's first law.

1.2.4. Condition of Local Stress Equilibrium

As a direct consequence of the conservation of linear momentum, the stress vector (traction) acting at a point on a surface is matched by an opposite and equal reaction force on the surroundings. That is, if momentum is conserved, the transfer of momentum between the system and its surroundings due to forces is mutual (*condition of local stress equilibrium*).

1.2.5. Stress Tensor

The condition of local stress equilibrium leads to an important decoupling of stress from the surface orientation. Consideration of the stresses acting on the surfaces of a tetrahedron shows that

$$\boldsymbol{t}_{(\boldsymbol{n})} = \boldsymbol{n} \cdot \boldsymbol{T} \tag{1.70}$$

where $\boldsymbol{T}$ is the second-order *stress tensor* (→ Fluid Mechanics, Chap. 3.1.), which may be a function of position and time but not of surface orientation.

The state of stress in a continuum at a point is the totality of all possible pairs of traction vectors $\boldsymbol{t}_{(\boldsymbol{n})}$ and surface orientations $\boldsymbol{n}$ (i.e., an infinite number). However, the nine-component stress tensor can be used to obtain $\boldsymbol{t}_{(\boldsymbol{n})}$ if $\boldsymbol{n}$ is given.

1.2.6. Equation of Motion

Now,

$$\frac{\mathrm{D}}{\mathrm{D}t} \int_V \varrho \boldsymbol{u} \mathrm{d}V = \int_V \varrho \boldsymbol{g} \mathrm{d}V + \int_A \boldsymbol{n} \cdot \boldsymbol{T} \mathrm{d}A \tag{1.71}$$

or, by the divergence theorem and Reynolds transport theorem,

$$\int_V \varrho \frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} \mathrm{d}V = \int_V \varrho g \mathrm{d}V + \int_V \nabla \cdot \boldsymbol{T} \mathrm{d}V \tag{1.72}$$

so (again, V is arbitrary)

$$\varrho \frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} = \varrho g + \nabla \cdot \boldsymbol{T} \tag{1.73}$$

or

$$\varrho \boldsymbol{a} = \varrho \boldsymbol{g} + \nabla \cdot \boldsymbol{T} \tag{1.74}$$

where $\boldsymbol{a} = \mathrm{D}\boldsymbol{u}/\mathrm{D}t$, the acceleration. This is known as Cauchy's first law, the conservation of linear momentum, or simply the *equation of motion* for a continuum. For further details, see → Fluid Mechanics, Chap. 2.2.

1.2.7. Conservation of Angular Momentum

If angular momentum is conserved locally in a continuum (the usual situation),

$$\frac{\mathrm{D}}{\mathrm{D}t} \int_V \varrho \left(\boldsymbol{u} \times \boldsymbol{x}\right) \mathrm{d}V = \int_V \varrho \left(\boldsymbol{g} \times \boldsymbol{x}\right) \mathrm{d}V + \int_A \left(\boldsymbol{t}_{(\boldsymbol{n})} \times \boldsymbol{x}\right) \mathrm{d}A \tag{1.75}$$

which is Euler's second law. By writing $\boldsymbol{t}_{(\boldsymbol{n})}$ in terms of $\boldsymbol{T}$ (Eq. 1.70) and using the divergence theorem, Reynold's transport theorem, and vector – tensor identities, this statement of the conservation of angular momentum reduces to the result that the stress tensor is symmetric. That is, a necessary and sufficient condition for satisfying the conservation of angular momentum is:

$$\boldsymbol{T} = \boldsymbol{T}^{\mathrm{T}} \tag{1.76}$$

However, this is restricted to fluids that have no couple stresses (i.e., torques on the fluid are the result only of forces) [5], [6].

Also, if $\boldsymbol{T}$ is a symmetric second-order tensor (nondiagonal, in general), a rotated coordinate system exists in which $\boldsymbol{T}$ is diagonal. In this system, the coordinate axes are called the principal axes of stress and the three normal stresses (eigenvalues) are called the principal stresses. In mechanics, analysis of the principal axes of stress provides the basis for Mohr's stress circle technique [6], [7].

1.2.8. Mechanical Energy Accounting Equation

Now, because of the conservation of linear momentum (Eq. 1.74),

$$(\varrho \boldsymbol{a} - \varrho \boldsymbol{g} - \nabla \cdot \boldsymbol{T}) \cdot \boldsymbol{u} = 0 \tag{1.77}$$

or

$$\varrho \frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}\boldsymbol{u}\cdot\boldsymbol{u}\right) - \varrho \boldsymbol{g}\cdot\boldsymbol{u} - (\nabla\cdot\boldsymbol{T})\cdot\boldsymbol{u} = 0$$

$$\left[\frac{\mathrm{D}\,(\boldsymbol{u}\cdot\boldsymbol{u})}{\mathrm{D}t} = \frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t}\cdot\boldsymbol{u} + u\cdot\frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} = \boldsymbol{a}\cdot\boldsymbol{u} + \boldsymbol{u}\cdot\boldsymbol{a} = 2\boldsymbol{a}\cdot\boldsymbol{u}\right] \tag{1.78}$$

Now, as an identity, for $\boldsymbol{T}$ symmetric

$$\nabla\cdot(\boldsymbol{T}\cdot\boldsymbol{u}) = (\nabla\cdot\boldsymbol{T})\cdot\boldsymbol{u} + \mathrm{tr}\,(\boldsymbol{T}\cdot\nabla\boldsymbol{u}) \tag{1.79}$$

so that

$$\varrho\frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}\boldsymbol{u}\cdot\boldsymbol{u}\right) = \varrho\boldsymbol{g}\cdot\boldsymbol{u} + \nabla\cdot(\boldsymbol{T}\cdot\boldsymbol{u}) - \mathrm{tr}\,(\boldsymbol{T}\cdot\nabla\boldsymbol{u}) \tag{1.80}$$

This scalar equation is a rate accounting equation of mechanical energy in the form of kinetic energy (the left-hand side), work done by body forces ($\varrho\,\boldsymbol{g}\cdot\boldsymbol{u}$), work done by traction forces [$\nabla\cdot(\boldsymbol{T}\cdot\boldsymbol{u})$], and the conversion of mechanical energy to thermal energy due to traction through both fluid compression and viscous dissipation [$\mathrm{tr}\,(\boldsymbol{T}\cdot\nabla\boldsymbol{u})$], all per unit volume. Obviously, this result is not independent of the equation of motion and, in fact, as a scalar equation must contain less information than the vector equation of motion.

1.2.9. Conservation of Total Energy

If $\hat{U}$ is the internal energy per unit mass, then a total energy balance is

$$\frac{\mathrm{D}}{\mathrm{D}t}\int_V \varrho\left(\frac{1}{2}\boldsymbol{u}^2 + \hat{U}\right)\mathrm{d}V = \int_V \varrho\boldsymbol{g}\cdot\boldsymbol{u}\,\mathrm{d}V + \int_A \boldsymbol{t}_{(\boldsymbol{n})}\cdot\boldsymbol{u}\,\mathrm{d}A - \int_A \boldsymbol{n}\cdot\boldsymbol{q}\,\mathrm{d}A + \int_V \Phi\,\mathrm{d}V \tag{1.81}$$

where $\boldsymbol{q}$ is the heat flux at the system boundary. The left side of this equation represents changes in kinetic and internal energy, and the terms on the right side have the following meaning:

$\int_V \varrho g\cdot u\mathrm{d}V$	rate at which work is done by body forces,
$\int_A t_{(n)}\cdot u\mathrm{d}A$	rate at which work is done due to traction,
$-\int_A n\cdot q\mathrm{d}A$	rate of energy transfer due to $\boldsymbol{q}$, and
$\int_V \Phi\mathrm{d}V$	rate of radiant energy absorption.

This last term may also be used to represent heat generation (conversion of energy from other forms to thermal energy) by mechanisms such as electrical heating and chemical or nuclear reactions instead of accounting for these conversions through changes in, e.g., changes in internal energy from one form to another within the material volume. From Equation (1.81) the continuum statement of total energy conservation is

$$\varrho\frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}\boldsymbol{u}^2 + \hat{U}\right) = \varrho\boldsymbol{g}\cdot\boldsymbol{u} + \nabla\cdot(\boldsymbol{T}\cdot\boldsymbol{u}) - \nabla\cdot\boldsymbol{q} + \Phi \tag{1.82}$$

1.2.10. Thermal Energy Accounting Equation

Subtracting the mechanical energy accounting equation from the total energy conservation equation gives an accounting of thermal energy:

$$\varrho\frac{\mathrm{D}\hat{U}}{\mathrm{D}t} = -\nabla\cdot\boldsymbol{q} + \mathrm{tr}\,(\boldsymbol{T}\cdot\nabla\boldsymbol{u}) + \Phi \tag{1.83}$$

In this result, $\mathrm{tr}\,(\boldsymbol{T}\cdot\nabla\boldsymbol{u})$ represents the conversion of mechanical energy to thermal energy. Equation (1.83) is not an independent equation but has the advantage of not involving mechanical energy terms. It can be used in place of the total energy equation.

1.2.11. Forms of the Governing Equations

Many possible forms of the continuity, motion, and energy equations exist, depending, for example, on whether partial or substantial derivatives are used and whether heat capacities (c_p or c_v) are used instead of internal energy. A convenient table for compositionally homogeneous continua summarizing these forms appears in [3]. Further summary of the governing equations is given in Section 1.4.

In these forms, the stress tensor is commonly separated into *isotropic stress* $p\boldsymbol{I}$ and *extra stress* $\boldsymbol{\tau}$ (→ Fluid Mechanics, Chap. 3.1.). Following the convention of BIRD,

$$\boldsymbol{T} = -\boldsymbol{\tau} - p\boldsymbol{I} \tag{1.84}$$

where $p \equiv -\operatorname{tr}(\boldsymbol{T})/3$. As a result $\nabla \cdot \boldsymbol{T}$ becomes $-\nabla \cdot \boldsymbol{\tau} - \nabla p$ and $\operatorname{tr}(\boldsymbol{T} \cdot \nabla \boldsymbol{u})$ becomes $-\operatorname{tr}(\boldsymbol{\tau} \cdot \nabla \boldsymbol{u}) - p\nabla \cdot \boldsymbol{u}$.

1.2.12. Entropy Inequality

If $\boldsymbol{T} = -\boldsymbol{\tau} - p\boldsymbol{I}$ for a compressible fluid, then the thermal energy equation is

$$\varrho \frac{\mathrm{D}\hat{U}}{\mathrm{D}t} = -\nabla \cdot \boldsymbol{q} - p\nabla \cdot \boldsymbol{u} - \operatorname{tr}(\tau \cdot \nabla \boldsymbol{u}) + \Phi \tag{1.85}$$

Now if $\hat{U} = \hat{U}(\hat{S}, \hat{V})$, where $\hat{S}$ is the entropy per unit mass and $\hat{V}$ the volume per unit mass, then

$$\mathrm{d}\hat{U} = T\mathrm{d}\hat{S} - p\mathrm{d}\hat{V} \tag{1.86}$$

or by the equation of continuity

$$\varrho \frac{\mathrm{D}\hat{U}}{\mathrm{D}t} = \varrho T \frac{\mathrm{D}\hat{S}}{\mathrm{D}t} - \varrho p \frac{\mathrm{D}\hat{V}}{\mathrm{D}t} = \varrho T \frac{\mathrm{D}\hat{S}}{\mathrm{D}t} - p\nabla \cdot \boldsymbol{u} \tag{1.87}$$

This gives for the thermal energy equation

$$\varrho T \frac{\mathrm{D}\hat{S}}{\mathrm{D}t} = -\nabla \cdot \boldsymbol{q} - \operatorname{tr}(\boldsymbol{\tau} \cdot \nabla \boldsymbol{u}) + \Phi \tag{1.88}$$

or

$$\varrho \frac{\mathrm{D}\hat{S}}{\mathrm{D}t} = -\nabla \cdot \left(\frac{\boldsymbol{q}}{T}\right) - \boldsymbol{q} \cdot \frac{\nabla T}{T^2} - \frac{1}{T}\operatorname{tr}(\boldsymbol{\tau} \cdot \nabla \boldsymbol{u}) + \frac{\Phi}{T} \tag{1.89}$$

Now, from experience, $\boldsymbol{q}$ is known to be in the opposite direction of ∇T, so $-\boldsymbol{q} \cdot \nabla T \geq 0$. Furthermore, and also from experience, $-\operatorname{tr}(\boldsymbol{\tau} \cdot \nabla \boldsymbol{u}) \geq 0$. Consequently, an inequality results that must hold for all continuum processes:

$$\varrho \frac{\mathrm{D}\hat{S}}{\mathrm{D}t} + \nabla \cdot \left(\frac{\boldsymbol{q}}{T}\right) - \frac{\Phi}{T} \geq 0 \tag{1.90}$$

Alternatively, $-\boldsymbol{q} \cdot \nabla T \geq 0$ and $-\operatorname{tr}(\boldsymbol{\tau} \cdot \nabla v) \geq 0$ may be viewed as constraints on allowable constitutive equations for $\boldsymbol{q}$ and $\boldsymbol{T}$.

1.2.13. Linear Transport Fluxes and Relations

The above transport equations describe the conservation of mass, linear momentum, and total energy, and include terms representing heat and momentum fluxes. The flux relations that are normally adopted follow the observation that the flux is proportional to a gradient driving force; i.e., the heat flux is proportional to the temperature gradient and the momentum flux is proportional to the velocity gradient (actually, the symmetric part of the velocity gradient tensor).

Fourier's Law. For *heat conduction*,

$$\boldsymbol{q} = -\lambda \nabla T \tag{1.91}$$

where λ is the thermal conductivity and, in general, may be a function of location in the material and temperature. (To be more general, the heat flux can be written in terms of an anisotropic thermal conductivity tensor to allow for directionality in the thermal conductivity.) In rectangular cartesian coordinates the components of the heat flux are

$$q_x = -\lambda \left(\frac{\partial T}{\partial x}\right)_{y,z}$$

$$q_y = -\lambda \left(\frac{\partial T}{\partial y}\right)_{x,z}$$

$$q_z = -\lambda \left(\frac{\partial T}{\partial z}\right)_{x,y} \tag{1.92}$$

Momentum Flux Relations. For viscous materials the most common momentum flux relation is Newton's law of viscosity, which for an incompressible material is written in terms of the *rate of deformation* tensor $\boldsymbol{\Delta}$ as

$$\boldsymbol{\tau} = -\eta \boldsymbol{\Delta} \tag{1.93}$$

where η is the viscosity. For Newtonian materials the viscosity is constant at constant temperature (i.e., it is independent of shear rate), resulting in this linear flux relation. For one-dimensional shear flow with only one velocity gradient component, this result becomes

$$\tau_{yx} = -\eta \frac{\mathrm{d}u_x}{\mathrm{d}y} \tag{1.94}$$

These linear relations can be used together with the conservation equations to model the behavior of continua. For a Newtonian fluid of constant viscosity and density (or for isochoric flow for which $\nabla \cdot \boldsymbol{u} = 0$), the equation of motion becomes the Navier–Stokes equation (for further details, see → Fluid Mechanics, Chap. 3.2., → Fluid Mechanics, Chap. 3.3.)

$$\varrho \left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} \right) = \varrho \boldsymbol{g} - \nabla p + \eta \nabla^2 \boldsymbol{u} \tag{1.95}$$

Also, for a Newtonian fluid of constant density, thermal conductivity, and viscosity, the thermal energy equation becomes

$$\varrho c_p \left(\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T \right) = \lambda \nabla^2 T + \Phi_\mathrm{v} \tag{1.96}$$

where the distinction between constant pressure and constant volume heat capacities vanishes for true incompressibility. In this result, Φ_v accounts for the conversion of mechanical energy to thermal energy through viscous dissipation and is given by

$$\Phi_\mathrm{v} = \eta \left\{ \mathrm{tr}\, [\nabla \boldsymbol{u} \cdot \nabla \boldsymbol{u}] + \mathrm{tr} \left[(\nabla \boldsymbol{u})^\mathrm{T} \cdot \nabla \boldsymbol{u} \right] \right\} \tag{1.97}$$

Further simplifications are often appropriate, such as for a solid material whose velocity terms in the thermal energy equation become insignificant, giving

$$\varrho c_p \left(\frac{\partial T}{\partial t} \right) = \lambda \nabla^2 T \tag{1.98}$$

Understanding material behavior with respect to thermal conductivity and viscosity is an important factor in understanding transport phenomena. Theoretical foundations for calculating these properties for low-density gases and non-Newtonian stress relations are given in the following paragraphs.

1.2.14. Transport Properties from Molecular Theories

The calculation of transport properties by using molecular transport theories with parameters estimated from experimental data is quite reliable for dilute (low-density) gases. This is a direct result of the fact that at low densities, intermolecular collisions and forces are minimized. Details of the theories are available in the original references or in summaries such as [8]. In this section, theories for viscosity and thermal conductivity are considered. Section 3.1.5 discusses theories for diffusivities.

Viscosity. Based on two-body, hard-sphere collisions the viscosity of a gas (in micropascal seconds, i.e., $10^{-6}\,\mathrm{Pa \cdot s}$) theoretically is given by

$$\eta = \frac{\frac{5}{16} (\pi M_r R T)^{1/2}}{\pi \sigma^2} = 26.69 \frac{\sqrt{M_r T}}{\sigma^2} \tag{1.99}$$

where T is in kelvin and the hard-sphere diameter σ is in ångströms (i.e., 10^{-10} m); M_r is the molecular mass of the gas.

To adjust for non-hard-sphere interactions, Chapman and Enskog introduced the use of spherically symmetric molecular potential functions [9], [10]. The most commonly used function is the Lennard–Jones 12–6 potential

$$\psi(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \tag{1.100}$$

from which can be calculated the interaction forces between molecules as a result of their separation r and model parameters σ and ε. These forces are then used to adjust the viscosity calculations through the dimensionless collision integral Ω_v, which is a function of kT/ε where k is the Boltzmann constant. Tabulations of this quantity are readily found in the literature, and a convenient and accurate computational form for Ω_v is given by Neufeld et al. [8], [11]. The value of Ω_v is unity for the ideal hard sphere; thus, deviations are a measure of the departure from this ideal model. The Chapman–Enskog relation for viscosity, then, is

$$\eta = \frac{\frac{5}{16}\,(\pi\, M_r\, RT)^{1/2}}{(\pi\,\sigma^2)\;\Omega_\mathrm{v}} = 26.69\,\frac{\sqrt{M_r\, T}}{\sigma^2\,\Omega_\mathrm{v}} \tag{1.101}$$

where again η is in 1 micropascal seconds, T in kelvin, and σ in ångströms.

Other potential functions have been used. The Stockmayer potential adjusts for polarity and should be used instead of the Lennard – Jones potential for polar molecules. Calculation of the Stockmayer potential by using the analytical correction to the Lennard – Jones potential provided by BROKAW is recommended [8], [12].

These potential energy functions rely on molecular parameters that must be determined from data. In practice, tabulations of these parameters, which are available for individual materials, are derived from experimental data for viscosity by using the appropriate potential function with the Chapman – Enskog theory. Using the parameters, then, along with the potential function (collision integral) to calculate viscosity, amounts to adjusting the measured values to the desired temperature by using the square-root temperature dependence of the molecular theory. As long as the form of the theory is correct, the calculations should be very good. In fact, when tabulated potential function parameters are used for a dilute gas, calculated viscosity values are correct to within ca. 1 %. Methods also exist for estimating the parameters; these too give quite reasonable viscosity values, with errors from 1 to 3 % [8].

Thermal Conductivity. The molecular theory for thermal conductivity is complicated by the fact that internal energy is stored in molecules in internal vibrational and rotational modes as well as in the translational mode, unlike viscosity (momentum) which is manifest only in translation. Consequently, although the basic molecular theory parallels that for viscosity, an additional term for these internal energy modes is necessary for polyatomic molecules. For monatomic gases (which do not have the internal modes) the Chapman – Enskog thermal conductivity relation is

$$\lambda = \frac{\frac{25}{32}\,(\pi\, M_r\, RT)^{1/2}\; c_v}{M_r\,(\pi\,\sigma^2)\;\Omega_\mathrm{v}} = 0.0833\,\frac{\sqrt{T/M_r}}{\sigma^2\,\Omega_\mathrm{v}} \tag{1.102}$$

where T is expressed in K, σ in Å, and λ in $\mathrm{J\,m^{-1}\,s^{-1}\,K^{-1}}$. In this relation a value of $c_v = 3\,k/2$ (monatomic ideal gas) was used. By using the viscosity relation (Eq. 1.101), the thermal conductivity can be expressed in terms of the viscosity and constant volume heat capacity as

$$\frac{\lambda\, M_r}{\eta\, c_v} = 2.5 \tag{1.103}$$

This result works quite well for monatomic gases. For polyatomic gases it must be adjusted to account for the additional degrees of freedom. The Eucken equation

$$\frac{\lambda\, M_r}{\eta} = c_v + 1.872 \times 10^4 \tag{1.104}$$

is a simple, yet reasonably accurate method for this correction. Other methods are available and are discussed and evaluated by REID et al. [8].

1.2.15. Non-Newtonian Fluids (see also → Fluid Mechanics, Chap. 4.)

In the momentum and energy balances, $\boldsymbol{q}$ and $\boldsymbol{\tau}$ represent heat and momentum fluxes. Before problems can be solved, constitutive relations must be selected for these fluxes in terms of appropriate driving forces (e.g., $\nabla \boldsymbol{T}$ or $\nabla \boldsymbol{u}$).

The specific relation for $\boldsymbol{\tau}$ that is appropriate for a particular material is the subject of rheology. Specific rules dictate the form of constitutive equations or set constraints on them. Useful references on non-Newtonian rheology are [13–15].

The rheological constitutive equation that is applicable in a given circumstance can be viewed as depending on either the type of material or on the specific flow situation. That is, the material may behave in a certain way because it behaves according to that model in all circumstances (i.e., it is that kind of material) or because all materials behave in that way for a particular flow situation.

From a materials viewpoint, behavior is normally classified as viscous, elastic, or viscoelastic. Complicating this classification by material type is the presence of history or memory effects (→ Fluid Mechanics, Chap. 4.1.4.).

One very common type of constant stretch history flow is *viscometric flow.* The importance of viscometric flows with respect to the characterization of materials is that all materials,

whether they are viscous, elastic, or viscoelastic, can be characterized by the same set of viscometric functions. The functions are referred to as the shear stress (or, equivalently, the viscosity) and the first and second normal stress differences (for further details, see → Fluid Mechanics, Chap. 4.3.1.).

Other kinds of flows are *elongational, slow, and small deformation oscillatory flows*. The last are periodic and yield a set of dynamic viscoelastic functions such as the dynamic elasticity and the dynamic viscosity (also referred to as the storage and loss moduli, respectively).

A Newtonian fluid has a constant viscosity independent of shear rate (at constant temperature). Other materials may show a decrease (pseudoplastic) or an increase (dilatant) in viscosity with increasing shear rate or exhibit a yield stress (e.g., the Bingham plastic). Complicating this picture is the fact that time-dependent behavior may also be observed and such fluids are referred to as thixotropic (decrease in viscosity with time at fixed shear stress) or rheopectic (increase).

In addition to shear stress, normal stresses may be generated by steady shear flow. These are stresses that are normal to the flow direction and are additional to the isotropic stress (pressure). The generation of normal stresses can have some rather dramatic effects on the flow of fluids and may have to be considered in process design. Experimentally, the only materials that have been observed to exhibit normal stresses are viscoelastic, such as polymer melts and polymer solutions.

More quantitatively, a viscous fluid is one whose stress tensor is a function only of the rate of deformation tensor $\boldsymbol{\Delta}$ (a function of strain rate and not of strain), where

$$\boldsymbol{\Delta} \equiv \nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^{\mathrm{T}} \tag{1.105}$$

A completely general form for a purely viscous fluid is

$$\boldsymbol{T} = \Phi_0 \boldsymbol{I} + \Phi_1 \boldsymbol{\Delta} + \Phi_2 \boldsymbol{\Delta}^2 \tag{1.106}$$

where the coefficients Φ_0, Φ_1, and Φ_2 are functions of the three scalar invariants of $\boldsymbol{\Delta}$ ($\mathrm{I}_{\boldsymbol{\Delta}}$, $\mathrm{II}_{\boldsymbol{\Delta}}$, and $\mathrm{III}_{\boldsymbol{\Delta}}$) and are defined in the following way:

$$\mathrm{I}_{\boldsymbol{\Delta}} = \mathrm{tr}\,(\boldsymbol{\Delta}) \tag{1.107}$$

$$\mathrm{II}_{\boldsymbol{\Delta}} = \frac{1}{2}\left[(\mathrm{tr}\,(\boldsymbol{\Delta}))^2 - \mathrm{tr}\,(\boldsymbol{\Delta}^2)\right] \tag{1.108}$$

$$\mathrm{III}_{\boldsymbol{\Delta}} = \det \boldsymbol{\Delta} \tag{1.109}$$

Note that $\mathrm{I}_{\boldsymbol{\Delta}} = \mathrm{tr}\,(\boldsymbol{\Delta}) = 2\,\nabla \cdot \boldsymbol{u}$, which is zero for an incompressible fluid or for isochoric flow. For incompressible flow then, the general viscous equation reduces to

$$\boldsymbol{T} = -p\boldsymbol{I} + \Phi_1 \boldsymbol{\Delta} + \Phi_2 \boldsymbol{\Delta}^2 \tag{1.110}$$

which is called the Reiner – Rivlin equation. In this case, Φ_1 and Φ_2 are functions only of $\mathrm{II}_{\boldsymbol{\Delta}}$ and $\mathrm{III}_{\boldsymbol{\Delta}}$ because $\mathrm{I}_{\boldsymbol{\Delta}} = 0$. Apparently, actual fluids of this type have not been documented (i.e., fluids with a nonzero second-order term and which are purely viscous have not been observed). Nevertheless, it may be a useful model because the second-order term provides a mechanism for generating normal stress differences in steady shear flow. However, because normal stress differences are not generally observed with viscous fluids, the second-order term is usually dropped ($\Phi_2 = 0$), and the model then reduces to the form generally used for incompressible Newtonian and non-Newtonian (non-constant-viscosity) fluids:

$$\boldsymbol{T} = -p\boldsymbol{I} - \eta \boldsymbol{\Delta} \tag{1.110a}$$

or, in terms of the extra stress $\boldsymbol{\tau}$,

$$\boldsymbol{\tau} = -\eta \boldsymbol{\Delta} \tag{1.110b}$$

where $\eta = \eta\,(\mathrm{II}_{\boldsymbol{\Delta}}, \mathrm{III}_{\boldsymbol{\Delta}})$ is the non-Newtonian viscosity. For shear flows, $[u_1 = u_1(x_2)]$, $\mathrm{III}_{\boldsymbol{\Delta}}$ is zero so that η is only a function of the second invariant which can be related to the shear rate $\dot{\gamma}$ by

$$\dot{\gamma} \equiv |\mathrm{d}u_1/\mathrm{d}x_2| = \left(\frac{1}{2}|\mathrm{II}_{\boldsymbol{\Delta}}|\right)^{1/2} \tag{1.110c}$$

Equation (1.110 b) is commonly used with $\eta = \eta\,(\dot{\gamma})$ for other kinds of flow (nonshear), even though to do so is not well-founded theoretically.

1.3. Summary of Basic Equations

A handful of conservation laws govern physical behavior and, as such, provide the basics of transport phenomena: total mass, linear momentum, angular momentum, and total energy. In

Table 1. Summary of the basic transport relations

Equation name	Conservation law	Equation	Equation number
Continuity	total mass	$\frac{D\varrho}{\mathrm{D}t} = -\varrho\nabla\cdot u$	(1.64)
	total mass	$\frac{\partial\varrho}{\partial t} = -\nabla\cdot(\varrho u)$	(1.65)
Motion	linear		
	momentum	$\varrho\boldsymbol{a} = \varrho\boldsymbol{g} + \nabla\cdot\boldsymbol{T}$	(1.74)
	linear momentum	$\varrho\left(\frac{\partial u}{\partial t} + u\cdot\nabla u\right) = \varrho g + \nabla\cdot T$	
	linear momentum	$\varrho\left(\frac{\partial u}{\partial t} + u\cdot\nabla u\right) = \varrho g - \nabla\cdot\tau - \nabla p$	
	angular momentum	$\boldsymbol{T} = \boldsymbol{T}^{\mathrm{T}}$	(1.76)
	total energy	$\varrho\frac{\mathrm{D}}{\mathrm{D}t}\left(\frac{1}{2}u^2 + \hat{U}\right) = \varrho g\cdot u + \nabla\cdot(T\cdot u) - \nabla\cdot q + \Phi$	(1.82)
Thermal energy accounting		$\varrho\frac{\mathrm{D}\hat{u}}{\mathrm{D}t} = -\nabla\cdot q + \mathrm{tr}\,(T\cdot\nabla u) + \Phi$	(1.83)
		$\varrho c_v\frac{\mathrm{D}T}{\mathrm{D}t} = -\nabla\cdot q + T\left(\frac{\partial p}{\partial T}\right)_\varrho(\nabla\cdot u) - \mathrm{tr}\,(\tau\cdot\nabla u) + \Phi$	
		$\varrho c_p\frac{\mathrm{D}T}{\mathrm{D}t} = -\nabla\cdot q + \left(\frac{\partial\ln\hat{V}}{\partial\ln T}\right)_p\frac{\mathrm{D}p}{\mathrm{D}t} - \mathrm{tr}\,(\tau\cdot\nabla u) + \Phi$	
Entropy inequality		$\varrho\frac{\mathrm{D}\hat{S}}{\mathrm{D}t} + \nabla\cdot\left(\frac{q}{T}\right) - \frac{\varrho\Phi}{T} \geq 0$	(1.90)

addition, subsets of energy give useful accounting relations, and the entropy inequality (second law of thermodynamics) provides constraints on the conservation laws. These relations are summarized in Table 1 along with some forms not presented in the text.

1.4. Boundary Conditions

The transport conservation relations (mass, total energy – or an accounting of thermal energy – and linear and angular momentum), along with the second law of thermodynamics and relations that describe material behavior, are not completely adequate for describing transport processes. With the addition of appropriate initial and boundary conditions the transport partial differential equations can be solved to obtain future behavior of the process over the required region of space. Specific examples appear in later sections. The principal types of conditions that are encountered and applied are outlined briefly in this section.

The conditions apply to the dependent variables of the transport equations (such as temperature or velocity), their derivatives, or other aspects of their behavior. The following are the types of conditions encountered:

1) A dependent variable is known throughout the region of space of interest at time zero; values at subsequent times result from the behavior of the system as destined through transport principles. For example, the temperature of a fluid over a region of space may be initially set to a prescribed value or function of position. Likewise, the velocity of the fluid may be known initially over the region of interest (e.g., if the fluid initially is stagnant, then $\boldsymbol{u} = 0$ at $t = 0$).
2) The dependent variable at a boundary of the region of space of interest is known; it may be fixed at a constant value over time or it may be a known variable as a function of time. For example, in fluid mechanics a fluid in contact with a solid usually meets the no-slip condition: the velocity of the fluid at this solid boundary is equal to the velocity of the solid surface. Similarly, the temperature at a boundary could be a known value or function of time.
3) The transport flux at a boundary of the region of interest is known as a function of time. This is not a condition on a dependent variable directly; however, it is an indirect condition. For example, in Fourier's law the heat flux is proportional to the temperature gradient; thus, specifying the heat flux at the boundary places a condition on the behav-

ior of the temperature at the boundary. If the component of the heat flux normal to the surface is known, a condition on $\boldsymbol{n} \cdot \boldsymbol{q}$ is established. For momentum transport, a known traction (stress vector) at the boundary is stated in terms of $\boldsymbol{n} \cdot \boldsymbol{\tau}$. As an example, at gas – liquid interfaces the stresses supported by the gas phase are very low (low viscosity), the momentum flux (traction) due to the extra stress at the interface is therefore zero: $\boldsymbol{n} \cdot \boldsymbol{\tau} = \boldsymbol{0}$.

4) The continuity of properties across the surface can also serve as a boundary condition. For example, at a surface (boundary) velocity, temperature, traction, and heat flux are normally continuous. The continuity of heat transfer across the interface holds as long as no thermal energy accumulates at the surface of interest. This, however, does not imply that temperature gradients are continuous; thermal conductivities may change at a boundary. Likewise, the continuity of traction is necessitated by the condition of local stress equilibrium (i.e., by the conservation of linear momentum).
5) The stress in fluids is finite, that is, fluids cannot sustain stresses without deforming so as to reduce those stresses. This condition is applicable, for example, in pipe flow problems where, if such a condition is not imposed a priori, solutions to the equation of motion could yield an infinite stress at the center of the pipe.

1.5. Solution Philosophy

We now have a total of five equations and one inequality. The equations are (1) continuity (mass conservation), (2) linear momentum conservation (three equations, one for each spatial direction), and (3) total energy conservation. Additionally, the conservation of angular momentum constrains $\boldsymbol{T}$ to be symmetric, and the entropy inequality places further constraints on $\boldsymbol{q}$ and $\boldsymbol{T}$. Also, linear momentum conservation can be used to obtain an accounting of mechanical energy (not an independent equation), which can be employed with total energy conservation to give an accounting of thermal energy that can be used in place of the total energy conservation law. The above equations have six independent variables: ϱ, u_x, u_y, u_z (or x, y, and z), p, and T. [The stress $\boldsymbol{T}$ and heat flux $\boldsymbol{q}$ are expressed in terms of position derivatives and temperature gradients, respectively. Furthermore, the thermodynamic properties ($\hat{U}$, $\hat{H}$, and $\hat{S}$) are expressed as functions of T and pressure (or ϱ) through heat capacity and $p - \varrho - T$ data or equation of state.] The five equations plus one volumetric equation of state [$p = p(\varrho, T)$] along with appropriate boundary or initial conditions dictate the subsequent behavior of the system.

If temperature is constant, the number of variables is reduced by one, as is the number of equations (energy conservation reduces to the mechanical energy accounting equation, which is not an independent equation).

If density is constant, the number of variables and equations (volumetric equation of state) both decrease by one.

1.6. Dimensionless Equations of Change

(→ Scale-Up in the Chemical Industry)

As discussed above, for a given flow situation involving momentum and heat transfer, a scalar equation of continuity, a vector equation of motion, and a scalar equation of energy (either total energy or a thermal energy accounting equation) are available. These laws govern the behavior of the process at hand. In these equations, each term has a specific physical meaning or origin important to the current situation. Consequently, each term carries with it dimensional characteristics that result from those physical meanings.

Frequently, from both computational and conceptual perspectives, equations are usefully recast into a dimensionless form by means of dimensions or parameters that are characteristic of the process. These dimensionless forms have several advantages. First, grouping the characteristic parameters or dimensions into a reduced number of parameters allows parametric calculations for a reduced number of cases. Second, the dimensionless equations allow comparison of processes of different size, which otherwise are the same; the dimensionless equations clarify the adjustments that must be made between the two processes to have them behave the same dynamically. In this way, a small-scale process can

give an accurate prediction of a large-scale process for design purposes. Third, correlations of experimental data can be made very efficiently by using dimensionless groups and correlations. Correlations developed for one transport phenomenon can even be used for another one when the situations (process, apparatus, and boundary conditions) are analogous (e.g., heat-transfer results can be used to predict mass-transfer behavior). This is the very important *Reynolds analogy*, which rests upon the similarity of conservation laws (momentum, mass, and energy) and flux relations (Fourier's law, Newton's law of viscosity, and Fick's first law of diffusion) for certain situations. Finally, dimensionless forms of the transport laws have considerable value in understanding the relative importance of the various contributions to process behavior (e.g., convective versus conductive heat transfer). This order-of-magnitude scaling is discussed further in Chapter 2.

The objective of dimensional analysis using the governing equations, then, is to convert the equation to a dimensionless form, thereby extracting dimensionless groups. First, the variables of the equations are expressed in terms of dimensionless quantities (herein denoted by $*$) by dividing each of them by a characteristic parameter of the same dimensions. For example, a temperature-dependent variable would be divided by a reference characteristic temperature or, as is more frequently used, a temperature difference ($T_1 - T_0$); velocity, by a velocity U that is characteristic of the process; variables involving distance, such as a derivative with respect to a distance coordinate, by a characteristic length L; time, either by a characteristic process time (e.g., t_c, a relaxation time) or by the ratio of the characteristic distance to the characteristic velocity L/U; mass, by a characteristic mass or, if the material density is constant, by the density (M/L^3); and pressure, by ϱU^2. Then, when all the variables have been nondimensionalized by appropriate constants, the resulting coefficients of the various terms can be grouped into a smaller number of coefficients. When this is carried out, some terms in a given equation will consist solely of the nondimensional variables of the problem, and other terms will consist of quantities involving nondimensional variables multiplied by nondimensional groups of the constants and characteristic parameters of the equation.

The selection of specific constants for the nondimensionalizing process is not necessarily unique. In comparing two processes of geometrical similarity and comparable boundary conditions, however, the same physical dimension must be selected for the two processes (e.g., the diameter of a process tank for both situations). Also, for order-of-magnitude scaling, the characteristic dimensions chosen must represent the physical phenomena in a quantitatively meaningful way.

Nondimensionalizing the equation of motion for an incompressible, Newtonian material provides a basis for comparing the relative importance of viscous forces, inertial forces, body forces, etc. The Navier – Stokes equation (Eq. 1.95),

$$\varrho\left(\frac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = \varrho\boldsymbol{g} - \nabla p + \eta\nabla^2\boldsymbol{u} \tag{1.111}$$

can be nondimensionalized by the three characteristic dimensions L, U, and ϱ (time is nondimensionalized by L/U) to give

$$\frac{\partial \boldsymbol{u}^*}{\partial t^*} + \boldsymbol{u}^*\cdot\nabla\boldsymbol{u}^* = \left(\frac{L\,g}{U^2}\right)\left(\frac{\boldsymbol{g}^*}{g}\right) - \nabla^* p^* + \left(\frac{\eta}{L\,U\,\varrho}\right)\nabla^{*2}\boldsymbol{u}^* \tag{1.112}$$

where in Equation (1.111) the $\varrho\boldsymbol{u}\cdot\nabla\boldsymbol{u}$ term represents inertial forces and the viscosity term represents viscous forces. Body forces such as gravity are accounted for by $\varrho\boldsymbol{g}$. In obtaining the dimensionless form (Eq. 1.112) the coefficient of the inertial term has been set to unity by division so that the remaining coefficients represent a ratio to the inertial forces. For example, the coefficient of the viscous term in Equation (1.112) is a dimensionless group representing the ratio of the viscous force to the inertial force characteristic quantities. The inverse of this group is the *Reynolds number*

$$Re = \frac{L\,U\,\varrho}{\eta} \tag{1.113}$$

The coefficient of the body force term represents the ratio of the body forces to the inertial forces and is the inverse of the *Froude number*:

$$Fr = \frac{U^2}{g\,L} \tag{1.114}$$

These dimensionless groups represent the relative importance of these terms in the equation of motion. A large Reynolds number signifies large inertial forces compared to viscous forces, and a large Froude number signifies a greater importance of inertial forces relative to the body forces (e.g., forced convection versus natural convection).

The equation of energy can be nondimensionalized in a similar way to obtain analogous dimensionless groups. In dimensional form the equation of thermal energy for an incompressible Newtonian fluid with Fourier's law of heat conduction is

$$\varrho c_p \left(\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \nabla T \right) = \lambda \nabla^2 T + \Phi_{\mathrm{v}} \tag{1.115}$$

where Φ_{v} is the viscous dissipation defined by Equation (1.97). This energy equation can be nondimensionalized to give

$$\frac{\partial T^*}{\partial t^*} + \boldsymbol{u}^* \cdot \nabla^* T^* = \frac{1}{Pe} \nabla^{*2} T^* + \frac{Br}{Pe} \Phi_{\mathrm{v}}^* \tag{1.116}$$

In Equation (1.115) the $\boldsymbol{u} \cdot \nabla T$ term is the transfer of heat as a result of fluid convection, $\lambda \nabla^2 T$ represents heat transfer by conduction, and Φ_{v} is the irreversible conversion of mechanical energy to thermal energy by viscous dissipation. When these terms are nondimensionalized and the coefficient of the convection heat-transfer term is made equal to unity, the inverse of the coefficient of the conduction term represents the relative importance of convention to conduction and is termed the *Péclet number*:

$$Pe = \frac{L\, U\, \varrho\, c_p}{\lambda} \tag{1.117}$$

It may also be viewed as the product of the Reynolds number and another dimensionless group termed the *Prandtl number*, given by

$$Pr = \frac{c_p\, \eta}{\lambda} \tag{1.118}$$

Evidently, the Prandtl number may be thought of as a conversion factor relating momentum transport (the ratio of transport by convection to that by molecular processes, i.e., stresses) to heat transport (the ratio of transport by convection to that by molecular processes, i.e., conduction) and expressed alternatively as the ratio of momentum diffusivity (kinematic viscosity) ν (= μ/ϱ) to thermal diffusivity, α (= $\lambda/\varrho\, c_p$). The coefficient of the viscous dissipation term represents the relative importance of viscous dissipation to convection and is normally expressed as the ratio of the Brinkman to the Péclet number, where the *Brinkman number* is

$$Br = \frac{\eta U^2}{\lambda T_0} \tag{1.118a}$$

and T_0 is a reference temperature or temperature difference.

Use of dimensionless forms of the governing equations provides an excellent basis for comparing the same process at different scales. If two processes are the same, that is, if they accomplish the same task in geometrically similar apparatuses and with comparable boundary conditions, then they are mathematically identical with respect to these governing equations. For further details, see → Scale-Up in the Chemical Industry.

2. Transport in Compositionally Homogeneous Systems

2.1. Heat Conduction in Solids

Heat transfer by conduction arises from spatial variations of temperature in a body. If the body is a stationary, opaque solid, conduction is the only mechanism for heat transfer. In this section the basic equations for heat conduction are summarized and specific results are presented that illustrate the effects of geometry, thermal properties, initial state, and boundary conditions. Good general references for this material are [16, Chaps. 2 – 4], [17, Chaps. 9 – 11], [18, Chaps. 1 – 5], and [19].

2.1.1. Heat Conduction Equations

The general energy balance was given in Equation (1.83). For heat conduction in solids it is simplified in several ways: In the case of stationary materials where $\boldsymbol{u} = \boldsymbol{0}$, only the accumulation, conduction, and generation terms are retained. Also, the continuity equation (Eq. 1.65) implies that the density ϱ is time-invariant. Further, local equilibrium is assumed, and the specific internal energy $\hat{U}$ is represented in terms of the density and temperature. Then,

$$\frac{\partial}{\partial t}\hat{U}(\varrho, T) = c_v \frac{\partial T}{\partial t} \tag{2.1}$$

For solids, the specific heat capacity at constant volume $c_v \equiv (\partial \hat{U}/\partial T)_\varrho$ is approximately equal to the specific heat capacity at constant pressure $c_p = (\partial \hat{H}/\partial T)_p$. The latter quantity is usually measured and found in data tables.

With these simplifications, the energy balance (Eq. 1.83) reduces to

$$\varrho c_p \frac{\partial T}{\partial t} = -\nabla \cdot \boldsymbol{q} + \boldsymbol{\Phi} \tag{2.2}$$

This equation is the differential form of the heat conduction equation. The term on the left is the accumulation term, and the terms on the right are the conductive flux and thermal energy generation and radiation terms, respectively.

To use the above equations to determine temperature fields and heat-transfer rates, the heat flux vector $\boldsymbol{q}$ must be related to the temperature gradient through Fourier's law, i.e.,

$$\boldsymbol{q} = -\lambda \nabla T \tag{2.3}$$

which is written here for isotropic materials. In general, the thermal conductivity λ depends on temperature, pressure, and composition. For homogeneous stationary solids, only the temperature dependence is important.

Substituting Fourier's law into the heat conduction equation gives

$$\varrho c_p \frac{\partial T}{\partial t} = \nabla \cdot (\lambda \nabla T) + \boldsymbol{\Phi} \tag{2.4}$$

which, with the specifications of $\lambda(T)$, $c_p(T)$, and $\boldsymbol{\Phi}$, as well as the necessary boundary and initial conditions, can be used to determine the temperature field. Once the temperature field is known, the directional heat-transfer rates can be determined from Fourier's law.

When a temperature-dependent thermal conductivity is used, the conduction equation is nonlinear and advanced analytical or numerical methods are generally required for solution [19, pp. 10 – 12]. For many materials, the dependence of λ on T is weak and λ can be taken as constant. Under these conditions, the heat conduction equation reduces to the nonhomogeneous form of the classical heat conduction (or diffusion) equation:

$$\frac{\partial T}{\partial t} = \alpha \nabla^2 T + \frac{\Phi}{\varrho c_p} \tag{2.5}$$

The quantity $\alpha \equiv \lambda / \varrho c_p$ is the *thermal diffusivity,* which plays the same role in heat transfer that mass diffusivity plays in mass transfer; it also has the same units (m^2/s).

Typical values of the thermal properties of various solid materials are given in Table 2. Properties can vary widely depending on the material, composition, and structural form.

Table 2. Thermal properties of various solid materials∗ at $\theta = 20\,^\circ C$

Material	λ, $W\,m^{-1}\,K^{-1}$	c_p, $J\,kg^{-1}\,K^{-1}$	ϱ, kg/m^3	$\alpha \times 10^6$, m^2/s
Aluminum	237	905	2 707	96.7
Copper	398	384	8 954	115.7
Cast iron (4 % C)	52	420	7 272	17.0
Stainless steel (304)	13.8	400	8 000	4.0
Lead	35	130	11 373	23.4
Silver	427	236	10 524	171.9
Common red brick	0.69	840	1 600	0.51
Concrete	1	800	2 100	0.6
Glass	0.76	800	2 690	0.35
Plaster	0.48	800	1 400	0.43
Wood, pine	0.1	2 500	600	0.07
Asbestos	0.2		580	
Glass wool	0.038		24	
Corkboard	0.043		160	

∗ Adapted from [16, p. 543] and [21, p. 592]. The values for construction materials can vary significantly depending on the specific composition and packing of the material.

2.1.2. Initial and Boundary Conditions

The initial condition required in the solution of the conduction equation involves the specification of temperature as a function of position at some initial time, say $t = 0$.

The boundary conditions can take several forms:

1) The temperature on the bounding surface of the solid can be specified as a function of position and time. Although this is the most convenient boundary condition from a solution standpoint, surface temperatures are very difficult to measure. More commonly, the temperature of the fluid in contact with the surface is known (see condition 3 below).
2) The heat flux normal to the bounding surface is specified as a function of position and time. A special case arises with a perfectly

insulated surface where the normal heat flux component is zero.

3) For a fluid – solid boundary, the normal component of the conductive heat flux on the solid side is equated to the "convective" flux away from the surface on the fluid side. This is expressed as

$$\text{On } A: \quad q_n = -\lambda \frac{\partial T}{\partial n} = h\,(T - T_{\mathrm{f}}) \tag{2.6}$$

where the conductive flux has been represented by Fourier's law and the "convective" flux by Newton's law of cooling. The latter relation represents the "convective" flux in terms of the temperature difference between the surface and the fluid (T_{f}) and the quantity h, which is called the *heat-transfer coefficient*. Actually, the flux on the fluid side arises from complex, coupled, conduction, and convection mechanisms that depend on the flow field in the neighborhood of the surface. Such effects are not represented explicitly in this relation but are embedded in the heat-transfer coefficient h, which is an effective thermal conductance that depends on fluid properties, geometry, and fluid motion. Large values of h are associated with efficient fluid mixing and efficient convective – conductive heat transfer in the neighborhood of the surface (e.g., turbulent transport in high-conductivity liquids). Small values of h are associated with nearly stagnant fluids of low conductivity. Typical values of h for various physical situations are given in Table 3.

Table 3. Order of magnitude of heat-transfer coefficients

Physical situation	h, W m^{-2} K^{-1}
Free convection, air	4 – 50
Forced convection, air	10 – 300
Forced convection, oil	60 – 1800
Forced convection, water	250 – 15 000
Boiling water	300 – 100 000
Condensing steam	4000 – 160 000

2.1.3. Steady, One-Dimensional Conduction

In many problems, the temperature field is time-invariant and depends on only one spatial variable. Consider first the plane slab case illustrated

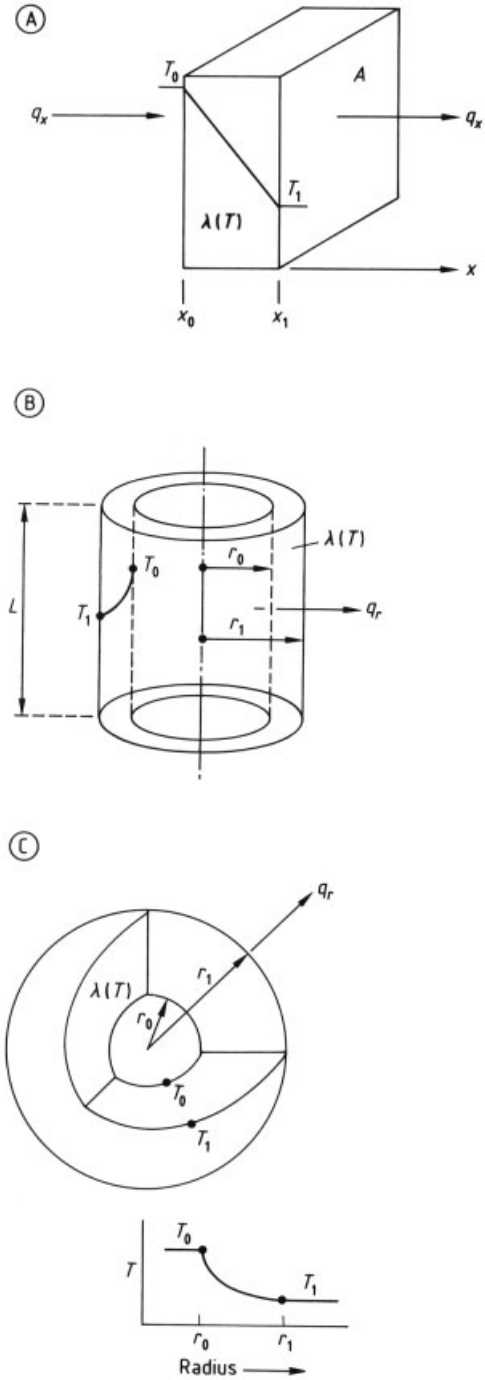

Figure 1. Steady one-dimensional heat conduction in various geometries
A) Conduction in a plane wall; B) Radial conduction in a cylindrical shell; C) Radial conduction in a spherical shell
Conductive resistance R_{cond} obtained from Equation (2.14); heat-transfer rate $\dot{Q}$ from Equation (2.13).

in Figure 1 A where no heat is generated, conduction is in the x direction, and the temperature dependence is $T(x)$. In this case, Equation (2.2) reduces to

$$\frac{\mathrm{d}q_x}{\mathrm{d}x} = 0 \tag{2.7}$$

or

$$q_x = \frac{\dot{Q}}{A} = \text{constant} \tag{2.8}$$

where $\dot{Q}$ is the heat-transfer rate in joules per second across the slab face of area A. If the temperatures T_0 and T_1 are specified at x_0 and x_1, and Fourier's law is used for the heat flux, the heat-transfer rate $\dot{Q}$ can be obtained by direct integration:

$$\dot{Q} = \frac{\overline{\lambda}_1 A (T_0 - T_1)}{\Delta x_1} \tag{2.9}$$

where $\Delta x_1 \equiv x_1 - x_0$ and $\overline{\lambda}_1$ is the average thermal conductivity defined by

$$\overline{\lambda}_1 \equiv \frac{1}{(T_1 - T_0)} \int_{T_0}^{T_1} \lambda \mathrm{d}T \tag{2.10}$$

The temperature field is obtained similarly except that the range of integration is from x_0 to any position x:

$$T = T_0 - \frac{\dot{Q}}{\overline{\lambda} A}(x - x_0) \tag{2.11}$$

where $\overline{\lambda} = \lambda(T)$ is the average conductivity in the range from T_0 to T and is defined by an equation similar to Equation (2.10) except that T_1 is replaced everywhere by T. If Equation (2.9) is substituted into Equation (2.11), then

$$T = T_0 - \frac{\overline{\lambda}_1 (T_0 - T_1)}{\overline{\lambda} \Delta x_1}(x - x_0) \tag{2.12}$$

If the dependence of λ on T is sufficiently weak and $\overline{\lambda} \approx \overline{\lambda}_1$, a linear temperature profile results. If λ is a stronger function of temperature, nonlinear profiles develop.

These results can be extended to more general geometries with a few modifications. If ξ is the direction of the nonzero heat flux component, the heat-transfer rate can be expressed as

$$\dot{Q} = q_\xi A_\xi = \frac{(T_0 - T_1)}{R_{\text{cond}}} \tag{2.13}$$

with the conductive resistance R_{cond} given by

$$R_{\text{cond}} = \frac{1}{\overline{\lambda}_1} \int_{\xi_0}^{\xi_1} \frac{\mathrm{d}\xi}{A_\xi} \tag{2.14}$$

Here A_ξ is the cross-sectional area perpendicular to the ξ direction. For radial conduction in cylindrical shells (see Fig. 1 B), ξ corresponds to r and A_ξ is $2\pi rL$; for radial conduction in spherical shells (see Fig. 1 C), ξ corresponds to r and A_ξ is $4\pi r^2$. The temperature profile is obtained in the same way as for the plane slab:

$$T = T_0 - \frac{\dot{Q}}{\overline{\lambda}(T)} \int_{\xi_0}^{\xi} \frac{\mathrm{d}\xi}{A_\xi} \tag{2.15}$$

If the dependence of thermal conductivity on temperature is sufficiently weak, where $\lambda(T) \approx \lambda(T_1)$, using Equations 2.13 and 2.14 gives

$$T = T_0 - \frac{(T_0 - T_1)}{\int_{\xi_0}^{\xi_1} \frac{\mathrm{d}\xi}{A_\xi}} \int_{\xi_0}^{\xi} \frac{\mathrm{d}\xi}{A_\xi} \tag{2.16}$$

For the cylindrical and spherical geometries illustrated in Figure 1, the conductive resistances and the heat-transfer rates are given by:

$$\text{(Cylindrical)} \quad R_{\text{cond}} = \frac{\ln(r_1, r_0)}{2\pi L \overline{\lambda}_1} \tag{2.16a}$$

$$\dot{Q} = \frac{2\pi \overline{\lambda}_1 (T_0 - T_1)}{\ln(r_1/r_0)} \tag{2.16b}$$

$$\text{(Spherical)} \quad R_{\text{cond}} = \frac{(r_1 - r_0)}{4\pi r_0 r_1 \overline{\lambda}_1} \tag{2.16c}$$

$$\dot{Q} = \frac{4\pi r_0 r_1 \overline{\lambda}_1 (T_0 - T_1)}{(r_1 - r_0)} \tag{2.16d}$$

2.1.4. Convective Boundary Conditions

As noted previously, the temperature of the fluid in contact with a surface is more commonly known than the surface temperature. Consider again the plane slab case in Figure 1 A, but now assume that the left face is in contact with a fluid of temperature T_{f0}, then from Newton's law of cooling,

$$T_{f0} - T_0 = \frac{q_x}{h_0} = \frac{\dot{Q}}{h_0 A} \tag{2.17}$$

Eliminating T_0 between Equations (2.9) and (2.17) gives

$$\dot{Q} = \frac{(T_{f0} - T_1)}{\left(\frac{1}{h_0 A} + \frac{\Delta x_1}{\bar{\lambda}_1 A}\right)} \tag{2.18}$$

where $\Delta x_1 = x_1 - x_0$. This equation gives the total heat-transfer rate in terms of the temperature driving force $T_{f0} - T_1$ and the series of convective and conductive resistances. The ratio of the conductive ($\Delta x_1/\bar{\lambda}_1 A$) and the convective resistances ($1/h_0 A$) is the dimensionless Biot number:

$$Bi \equiv \frac{h_0 \Delta x_1}{\bar{\lambda}_1} \tag{2.19}$$

When the Biot number is large, conduction is the controlling heat-transfer mechanism and $T_0 \approx T_{f0}$. When the Biot number is small, convection is controlling and $T_0 \approx T_1$.

2.1.5. Composite Systems

The results of Sections 2.1.1, 2.1.2, 2.1.3, and 2.1.4 can be extended to composite or multilayered systems by simply applying equations similar to Equation (2.9) across each solid layer and Newton's law of cooling at each fluid – solid boundary. For the two-layer slab shown in Figure 2

$$\dot{Q} = \frac{\bar{\lambda}_1 (T_0 - T_1)}{\Delta x_1} = \frac{\bar{\lambda}_2 (T_1 - T_2)}{\Delta x_2}$$
$$= \alpha_0 (T_{f0} - T_0) = \alpha_2 (T_2 - T_{f2}) \tag{2.20}$$

where $\Delta x_k = x_k - x_{k-1}$. Using these equations to eliminate T_0, T_1, and T_2 yields

$$\dot{Q} = \frac{(T_{f0} - T_{f2})}{\left(\frac{1}{\alpha_0 A} + \frac{\Delta x_1}{\bar{\lambda} A} + \frac{\Delta x_2}{\bar{\lambda}_2 A} + \frac{1}{\alpha_2 A}\right)} \tag{2.21}$$

or

$$\dot{Q} = \frac{(T_{f0} - T_{f2})}{\sum R_{conv} + \sum R_{cond}} \tag{2.22}$$

Generalizing to N layers gives

$$\dot{Q} = \frac{(T_{f0} - T_{fN})}{\left[\frac{1}{\alpha_0 A} + \sum_{i=1}^{N} \left(\frac{\Delta x_i}{\bar{\lambda}_i A}\right) + \frac{1}{\alpha_N A}\right]} \tag{2.23}$$

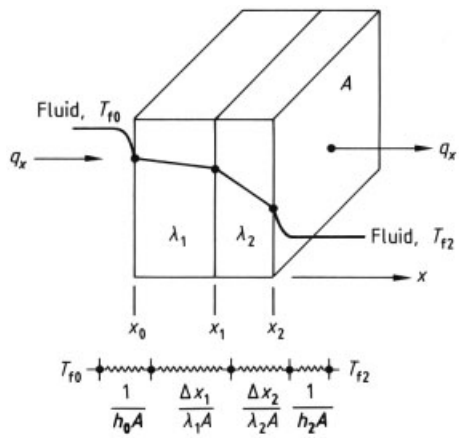

Figure 2. Conduction through a composite two-layer slab
Left face is exposed to a fluid at temperature T_{f0}; right face is exposed to a fluid of temperature T_{f2}. The overall driving force for the heat transfer and the conductive resistances are shown in the equivalent electronic circuit.

Similar expressions can be written for composite systems in other geometries (cylinders, spheres, etc.). In these cases, the heat-transfer rate is given by Equation (2.22), with the $R_{cond,i}$ contributions for each layer obtained from Equation (2.14) and the $R_{conv,i}$ contributions from

$$R_{conv,i} = \frac{1}{h_i A_i} \tag{2.24}$$

where h_i is the heat-transfer coefficient at the "ith" fluid – solid surface and A_i the corresponding surface area.

2.1.6. Steady Conduction with Heat Generation and in Multidimensions

Heat is commonly generated in solids by electrical heating as well as by chemical and nuclear reactions. If heat generation is uniform and conduction is one-dimensional, the temperature field can be determined by direct integration. For more complex problems involving multidimensional conduction and nonuniform generation, advanced mathematical methods are required. These are outside the scope of this article but are discussed in [19].

2.1.7. Transient Heat Conduction – Lumped Capacity Systems

When a solid body with an initial temperature T_i is exposed to a fluid of temperature T_f, the temperature field in the body changes with time to decrease the magnitude of the temperature difference. Under the conditions of small Biot numbers, the determination of the resulting transient temperature change can be analyzed by using a lumped parameter approach. In particular, if Newton's law of cooling describes the conditions at the fluid – solid boundary, an order-of-magnitude estimate (symbolized by "$\sim$") gives

$$\Delta T \sim \left(\frac{h\,l}{\lambda}\right)(T_0 - T_f) \tag{2.25}$$

where l is the characteristic length scale of the body and ΔT is the approximate temperature change occurring over this length scale. Clearly, when $Bi \equiv (h\,l/\lambda) \ll 1$, the spatial temperature change ΔT in the body is much smaller than the temperature changes $T_0 - T_f$ across the fluid layer adjacent to the bounding face. Under such conditions, the surface temperature of the body (T_0) can be assumed to be approximately equal to the average body temperature ($\bar{T}$). The energy balance then becomes

$$\varrho c_p V \frac{d\bar{T}}{dt} = -\bar{h} A \left(\bar{T} - T_f\right) \tag{2.26}$$

where $\bar{h}$ and $\bar{T}$ represent average values over the surface and volume V of the body, respectively. If the material properties are taken to be constant, this equation can be integrated to give

$$\bar{T} = T_f + (T_i - T_f)\exp\{-Bi\,t^*\} \tag{2.27}$$

where, in this case, the characteristic length is $l \equiv V/A$, the Biot number is $Bi \equiv l\,h/\lambda$, and the dimensionless time is $t* \equiv \alpha\,t/l^2$. The corresponding rate of heat transfer to the fluid is given by

$$\dot{Q} = \bar{h} A (T_i - T_f)\exp\{-Bi\,t^*\} \tag{2.28}$$

The significance of these results for $Bi \ll 1$ is that the rate of heat transfer given by Equation (2.28) represents an upper bound; i.e., the cooling or heating rates found at higher Biot numbers will be slower and the characteristic time for temperature accommodation of the body with the fluid will be longer.

2.1.8. Transient Heat Conduction – More General Solutions

At higher Biot numbers, the spatial variations of temperature cannot be neglected and the problem is mathematically more complex than that given in the previous section. In this section, solutions for the semi-infinite body, the plane slab, the infinite cylinder, and the sphere are summarized. Mathematical details are presented in [19, pp. 50 – 73 (semi-infinite solid), pp. 119 – 127 (plane slab), pp. 201 – 203 (infinite cylinder), and pp. 237 – 238 (sphere)].

Consider first the transient heating (or cooling) of a body bounded by the plane $x = 0$ and extending to infinity in the positive x direction. The body is initially at temperature T_i and its face at $x = 0$ is held at temperature T_f. If constant properties are assumed, the governing equations for $T(x, t)$ are

$$\frac{\partial T}{\partial t} = \alpha \frac{\partial^2 T}{\partial x^2} \tag{2.29}$$

$$T(x, 0) = T_i, \quad x \geq 0 \tag{2.30}$$

$$T(0, t) = T_f, \quad t > 0 \tag{2.31}$$

$$T(\infty, t) = T_i, \quad t \geq 0 \tag{2.32}$$

These equations suggest a solution of the form

$$T - T_f = f(x, t, \alpha, T_i - T_f) \tag{2.33}$$

If dimensional analysis is used, the dimensionless temperature can be expressed in terms of a single independent variable, i.e.,

$$T^* \equiv \frac{T - T_f}{T_i - T_f} = T^*(\zeta) \tag{2.34}$$

where

$$\zeta = \frac{x}{\sqrt{4\alpha t}} \tag{2.35}$$

When Equation (2.34) is substituted into Equations (2.29) – (2.32), the problem is reduced to the solution of an ordinary differential equation

$$\frac{d^2 T^*}{d\zeta^2} + 2\zeta \frac{dT^*}{d\zeta} = 0 \tag{2.36}$$

subject to the boundary conditions

$$T^*(0) = 0 \tag{2.37}$$

$$T^*(\infty) = 1 \tag{2.38}$$

Integrating Equation (2.36) first for $dT*/d\zeta$ and then again for $T*(\zeta)$ and using the boundary conditions to determine the two integrating constants give

$$T^* = \frac{2}{\sqrt{\pi}} \int_0^{\zeta} e^{-\xi^2} d\xi \equiv \text{erf}(\zeta) \tag{2.39}$$

The middle term is called the error function [erf (ζ)]. Tabulated values can be found in various sources [19, pp. 485–487]. Of particular interest here is the fact that the error function (as well as $T*$) takes on a value of approximately 0.99 when its argument ζ takes on a value of 1.8. Physically, this implies that the temperature change ($T - T_f$) reaches 99 % of the maximum change ($T_i - T_f$) within a thermal boundary layer defined by

$$\delta_T = 1.8\sqrt{4\alpha t} = 3.6\sqrt{\alpha t} \tag{2.40}$$

This is illustrated in Figure 3 where the time evolution of the temperature profile is plotted for the transient heating of a semi-infinite body.

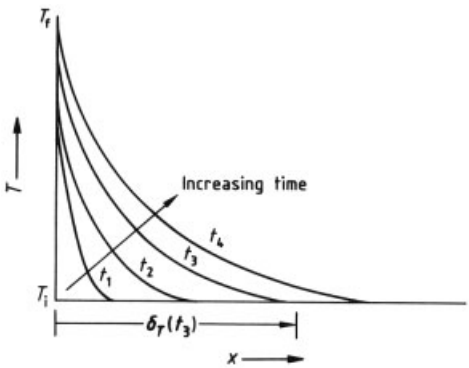

Figure 3. Time evolution of the temperature profile during heating of a semi-infinite slab ($x > 0$)
The x position where $T = T_f - 0.99\,(T_f - T_i)$ defines the thermal boundary layer δ_T which according to Equation (2.40) increases in proportion to $\alpha t\sqrt{\alpha t}$.

If the temperature boundary condition at $x = 0$ (Eq. 2.31) is replaced by a convective flux condition, i.e.,

$$\lambda \frac{\partial T}{\partial x}(0, t) = h\,[T(0, t) - T_f] \tag{2.41}$$

the solution takes the form

$$T^* = \text{erf}(\zeta) + \exp\left(\zeta\beta + \frac{1}{4}\beta^2\right) \cdot \left[\text{erfc}\left(\zeta + \frac{1}{2}\beta\right)\right] \tag{2.42}$$

where

$$\beta \equiv \frac{2\sqrt{\alpha t}\,h}{\lambda} \tag{2.43}$$

and erfc (y) = 1 − erf (y) is the complementary error function.

Although these results have been developed in the context of a semi-infinite body, they are also valid for a body of finite thickness (e.g., a slab of thickness $2l$) for all times when $\delta_T \le l$. For times longer than this, the boundary layer exceeds the half thickness and a semi-infinite solution is no longer valid. Then solutions for finite bodies must be used. Solutions of the latter type are now presented for the plane slab, the infinite cylinder, and the sphere.

Consider one-dimensional heat conduction in a plane slab bounded by the surfaces at $x = \pm l$. For constant thermal properties the governing energy equation is Equation (2.29); the initial and boundary conditions take the form

$$T(x, 0) = T_i, \quad -l \le x \le l \tag{2.44}$$

$$\frac{\partial T}{\partial x}(0, t) = 0, \quad \text{all } t \tag{2.45}$$

$$-\lambda \frac{\partial T}{\partial x}(l, t) = h\,[T(l, t) - T_f], \quad t > 0 \tag{2.46}$$

The second equation arises from the fact that the midplane ($x = 0$) is a plane of symmetry and the flux across it is zero.

The solution to this problem in terms of the dimensionless position $x* \equiv x/l$ and the dimensionless time $t* \equiv \alpha\, t/l^2$ is obtained as a Fourier series:

$$T^* \equiv \frac{T - T_f}{T_i - T_f} = 2\sum_{n=1}^{\infty} \frac{\sin\beta_n \cos(\beta_n x^*)}{\beta_n + \sin\beta_n \cos\beta_n} \exp\{-\beta_n^2 t^*\} \tag{2.47}$$

where the eigenvalues β_n are obtained from

$$\beta_n \tan\beta_n = Bi \tag{2.48}$$

For short times the convergence of Equation (2.47) is slow and many terms must be taken before an accurate solution can be obtained. However, for $t* > 0.25$, only the first term of this equation is needed to obtain results with errors less than 1 %. Even in these cases the solution is not convenient because β_1 must be determined from an implicit relation (Eq. 2.48). Fortunately, the roots of the latter equation can be accurately fitted to an expression of the form

$$\beta_1 = \sum_{k=1}^{5} C_k \left(\log Bi\right)^{k-1} \quad \text{for} \quad 0.02 \leq Bi \leq 8 \quad (2.49)$$

For the plane slab case, the coefficients C_k are given in the first column of Table 4. When this equation is used with the first term of Equation (2.47), temperatures can be predicted to within 5 % and, in most cases, depending on the Biot number, to within 1 %. When $Bi > 8$, tabulated values of β_n can be used [19, p. 491]. Such approaches are preferred to the time – temperature charts used by many investigators [22–25] because these charts are often difficult to read and to interpolate accurately.

Table 4. Coefficients for Equation (2.49)

Coefficient	Plane slab	Cylinder	Sphere
C_1	0.859777	1.256701	1.573305
C_2	0.702996	1.112492	1.450377
C_3	0.050636	0.148517	0.251563
C_4	−0.144345	−0.260841	−0.306763
C_5	−0.044631	−0.111339	−0.139485

Similar types of solutions are available for the infinite cylinder (radius r_0) and the sphere (also radius r_0). Here again, the body is assumed to be at an initial temperature T_i and at $t = 0$ is immersed in a fluid at temperature T_f. For the cylinder, the solution in terms of the first- and second-order Bessel functions of the first kind is

$$T^* = 2 \sum_{n=1}^{\infty} \frac{J_1\left(\beta_n\right) J_0\left(\beta_n r^*\right)}{\beta_n \left[J_0^2\left(\beta_n\right) + J_1^2\left(\beta_n\right)\right]} \exp\left\{-\beta_n^2 t^*\right\} \quad (2.50)$$

where $T*$ is defined as in Equation (2.47), $r* \equiv r/r_0$ and $t* \equiv \alpha\, t/r_0^2$. The eigenvalues β_n are obtained from

$$\beta_n \frac{J_1\left(\beta_n\right)}{J_0\left(\beta_n\right)} = Bi \quad (2.51)$$

As before, for $t* > 0.25$, only the first term of Equation (2.50) is important. Also, the first root of Equation (2.51) can be obtained from Equation (2.49) with the coefficients C_k given in the second column of Table 4.

For a sphere, the solution is

$$T^* = \frac{2\,Bi}{r^*} \sum_{n=1}^{\infty} \frac{\sin\beta_n \sin\left(\beta_n r^*\right)}{\beta_n\left[\beta_n - \sin\beta_n \cos\beta_n\right]} \exp\left\{-\beta_n^2 t^*\right\} \quad (2.52)$$

with the eigenvalues given by

$$\beta_n \cot\beta_n = 1 - Bi \quad (2.53)$$

The one-term solution is again valid for $t* > 0.25$, and the first eigenvalue can be obtained from Equation (2.49) by using the coefficients in the third column of Table 4.

The solutions of these one-dimensional unsteady-state problems can be used in simple ways to obtain solutions to multidimensional, unsteady-state problems. These are now summarized for several cases. Consider first the case of an *infinite rectangular bar* bounded by the surfaces at $x = \pm x_0$ and $y = \pm y_0$. The solution at any position (x, y) at time t is given by

$$T\left(x, y, t\right) = T_x\left(x, t\right) T_y\left(y, t\right) \quad (2.54)$$

Here $T_x(x, t)$ is obtained from Equation (2.47) with x_0 replacing l as the length scale. Similarly, $T_y(y, t)$ is obtained from Equation (2.47) with y replacing x in that equation and y_0 replacing l.

For a *rectangular parallelepiped* with bounding surfaces at $x = \pm x_0$, $y = \pm y_0$, and $z = \pm z_0$, the solution is given by

$$T\left(x, y, z, t\right) = T_x\left(x, t\right) T_y\left(y, t\right) T_z\left(z, t\right) \quad (2.55)$$

The T_x and T_y solutions are obtained in the manner just described for the infinite rectangular bar. The T_z solution is obtained from Equation (2.47) with z replacing x in that equation and z_0 replacing l.

Finally, for a *finite cylinder* bounded by the surfaces $r = r_0$ and $z = \pm l$, the solution is

$$T\left(r, z, t\right) = T_r\left(r, t\right) T_z\left(z, t\right) \quad (2.56)$$

Here T_r is obtained from Equation (2.50) and T_z from Equation (2.47) (z replaces x).

Examples of the use of Equations (2.54) – (2.56) are given in many texts [26, pp. 307 – 310], [27, pp. 174 – 177]. The approach suggested here is different in that the one-dimensional solutions are obtained from one-term analytical solutions instead of time – temperature charts.

2.2. Steady, One-Dimensional Flows

(→ Fluid Mechanics, Chap. 3.3.1.)

As in the sections above where heat transfer arose from the existence of temperature gradients, momentum transfer arises from fluid-deforming velocity gradients. In particular, for the one-dimensional unidirectional flow

$$\boldsymbol{u} \equiv (u_x, u_y, u_z) = [0, 0, u_z(y)] \tag{2.57}$$

the resulting momentum transfer involves z-momentum being transferred in the y-direction. This is illustrated in Figure 4 where faster moving fluid layers transfer momentum to adjacent slower moving layers. The nonzero stress components associated with this velocity field can be interpreted as momentum fluxes [17, Chap. 2], and if the fluid is Newtonian, these components can be expressed as

$$\tau_{yz} = \tau_{zy} = -\eta \frac{\mathrm{d}u_z}{\mathrm{d}y} \tag{2.58}$$

which is analogous to Fourier's law for one-dimensional heat conduction in the y-direction.

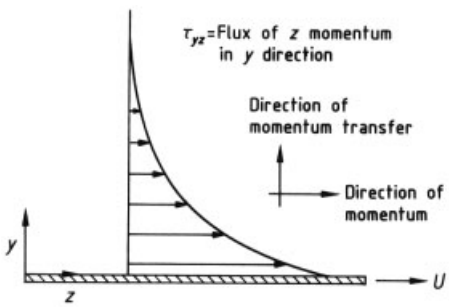

Figure 4. Flow resulting from the movement of a surface In this case, z-momentum is being transferred from layer to layer in the y-direction. The stress component τ_{yz} can be interpreted as a momentum flux (z momentum transferred in the y direction).

The flow above is a special case of the general class of unidirectional flows. Unidirectional flows commonly arise with internal flows in straight channels where the cross-sectional area is invariant in the flow direction. Under such conditions, the velocity u_z does not vary in the z-direction and the most general form of the velocity field is $\boldsymbol{u} = [0, 0, u_z(x, y)]$. The latter two-dimensional problem arises with systems of arbitrary cross section. For simpler systems such as flow between parallel plates or flow in cylindrical tubes, the problem becomes one-dimensional and u_z is dependent on only one spatial variable.

The driving force for unidirectional flows can be pressure gradients, gravity, or moving boundaries. The pressure gradient and gravity effects appear directly in the equation of motion and can be combined into an overall driving force term ∇P:

$$\nabla P = \nabla p - \varrho \boldsymbol{g} \tag{2.59}$$

To achieve a unidirectional flow without moving boundaries, ∇P must be nonzero and the flow direction must be coincident with the directions of ∇P. For flow in the z-direction where $P(z)$, the only nonzero component of Equation (2.59) is

$$\frac{\mathrm{d}p}{\mathrm{d}z} = \frac{\mathrm{d}P}{\mathrm{d}z} - \varrho g_z \tag{2.60}$$

where g_z is the component of gravity in the z-direction. Integration of this equation along the flow channel from $z = z_0$ to any point z gives

$$P - P_0 = p - p_0 - \varrho g_z (z - z_0) \tag{2.61}$$

That is, the driving force for flow arises from variations of pressure and potential energy along the flow direction. The quantity P is sometimes referred to as the *dynamic pressure* because it can be written as the difference between the pressure existing during flow and that arising under no flow or hydrostatic conditions:

$$P = p - p_{\mathrm{H}} \tag{2.62}$$

Here, p_H is the *hydrostatic pressure* which is obtained from

$$\nabla p_{\mathrm{H}} - \varrho \boldsymbol{g} = \boldsymbol{0} \tag{2.63}$$

In Equation (2.61) the hydrostatic pressure part is $p_0 + \varrho g_z (z - z_0)$.

Another characteristic of steady unidirectional flows is that these are noninertial flows. Specifically, a fluid element in this flow does not accelerate along its path, and the nonlinear inertial term $\boldsymbol{u} \cdot \boldsymbol{\nabla u}$ in the equation of motion is zero. The z-component of the equation of motion then takes the form

$$(\boldsymbol{\nabla} \cdot \boldsymbol{\tau}) \cdot i_z + \frac{\mathrm{d}P}{\mathrm{d}z} = 0 \tag{2.64}$$

where i_z is the unit base vector in the z direction. If the fluid is an incompressible Newtonian fluid, this equation can be expressed as

$$\eta \nabla^2 u_z - \frac{\mathrm{d}P}{\mathrm{d}z} = 0 \tag{2.65}$$

With respect to Newton's laws of motion, this equation can be interpreted as a balance of forces per unit volume on any fluid element in the flow; specifically, the driving force for the flow is $(-\,\mathrm{d}P/\mathrm{d}z)$ and the viscous resisting force is $\eta\ \nabla^2 u_z$.

The only material parameter that arises in the treatment of steady, unidirectional flows is the viscosity η. In more general flows, the density ϱ also arises. In Table 5, typical values of density and viscosity for a number of fluids are presented. Values for the heat capacity c_p, the thermal conductivity λ, and the Prandtl number $Pr \equiv \nu/\alpha$ are also provided. All of these quantities arise later in the treatment of convective heat transfer.

2.2.1. Generalized Couette Flow

A flow pattern arising in many different applications (polymer extrusion, screw pumping and expression, journal-bearing lubrication, etc.) is the unidirectional plane parallel flow illustrated in Figure 5. The flow is driven by both the movement of the upper plate at velocity U and the application of a pressure gradient $(P_0 - P_L)/L$ in the direction of flow. The flow geometry is assumed to be of thickness b in the y-direction, width w in the x-direction, and length L in the z-direction. The thickness-to-width ratio (b/w) and the thickness-to-length ratio (b/L) are assumed to be sufficiently small that edge effects and entrance or exit effects can be neglected, and the velocity field can be represented by $\boldsymbol{u} = (u_x, u_y, u_z) = [0, 0, u_z\,(y)]$.

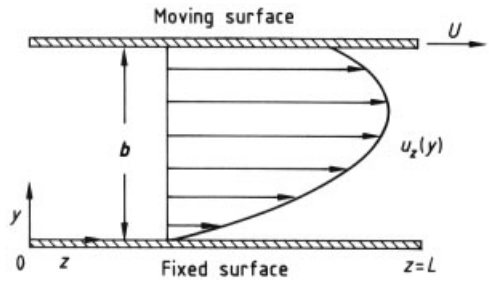

Figure 5. Generalized Couette flow between parallel surfaces
The upper surface moves with velocity U; the lower surface is stationary. The nature of the flow depends on the magnitude and sign of the applied pressure gradient $\mathrm{d}P/\mathrm{d}z$. In the case above, $\mathrm{D}P/\mathrm{D}z > 0$.

The nonzero stress components for this flow are τ_{yz} and τ_{zy}, and Equation (2.64) reduces to

$$\frac{\mathrm{d}\tau_{yz}}{\mathrm{d}y} + \frac{\mathrm{d}P}{\mathrm{d}z} = 0 \tag{2.66}$$

Since the first term on the left-hand side is at most a function of y and the second term at most a function of z, this equation is satisfied only if both terms are equal to a constant. Integration then gives

$$\tau_{yz} = K y + \tau_0 \tag{2.67}$$

$$K \equiv \frac{(P_0 - P_L)}{L} \tag{2.68}$$

where τ_0 is the shear stress at $y = 0$. At this point, the quantity τ_0 is unknown. It is determined by using the velocity boundary conditions.

If the fluid is an incompressible Newtonian fluid, Equation (2.58) can be combined with Equation (2.67) and integrated to give the velocity profile

$$u_z = \frac{b^2 \Delta P}{2\eta L}\left(\frac{y}{b}\right)\left(1 - \frac{y}{b}\right) + U\left(\frac{y}{b}\right) \tag{2.69}$$

where the quantity τ_0 and the additional constant that arises from the integration are evaluated by using the boundary conditions

$$u_z\,(0) = 0 \tag{2.70}$$

$$u_z\,(b) = U \tag{2.71}$$

In dimensionless form, the velocity is

$$\frac{u_z}{U} = K^* \left(\frac{y}{b}\right)\left(1 - \frac{y}{b}\right) + \frac{y}{b} \tag{2.72}$$

Table 5. Thermophysical properties of various gases and liquids∗

Fluid	θ, °C	ϱ, kg/m^3	η, Pa s	λ, W m^{-1} K^{-1}	c_p, J kg^{-1} K^{-1}	Pr
Gases (101.3 kPa)						
Air	27	1.183	1.853×10^{-5}	0.02614	1.003×10^{3}	0.711
Carbon dioxide	27	1 .7973	1.4958×10^{-5}	0.016572	0.871×10^{3}	0.770
Helium	−18	0.1906	1.817×10^{-5}	0.1357	5.200×10^{3}	0.70
Hydrogen	27	0.08185	0.8963×10^{-5}	0.182	14.314×10^{3}	0.706
Nitrogen	27	1.1233	1.784×10^{-5}	0.0259	1.0408×10^{3}	0.715
Oxygen	27	1.3007	2.063×10^{-5}	0.02676	0.09203×10^{3}	0.709
Liquids (saturated)						
Dichlorodifluoromethane	27	1305	0.2545×10^{-3}	0.0690	980	3.62
Glycerol	20	1261	1412×10^{-3}	0.285	2350	11 630
Mercury	27	13 611	1.633×10^{-3}	8.34	139.1	0.027
Light machine oil	16	907	145.1×10^{-3}			
Water	27	996.6	0.8232×10^{-3}	0.6084	4177	5.65

∗ Adapted from [21, pp. 595 – 601].

where

$$K^* = \frac{b^2 \Delta P}{2\eta U L} \tag{2.73}$$

is the dimensionless pressure gradient.

Depending on the magnitude and sign of $K*$, the velocity profile can take various forms (see Fig. 6). When $K*=0$, the profile is linear and corresponds to simple Couette flow:

$$u_z = U\left(\frac{y}{b}\right) \tag{2.74}$$

When $K*>0$, the pressure gradient and the moving upper boundary both contribute to flow in the positive z direction, and if $K*$ is sufficiently large a maximum occurs in the velocity profile (see the $K*\geq 2$ cases in Fig. 6). When $K*<0$, the pressure gradient causes a reverse flow. If the magnitude of this adverse gradient is sufficiently large, the fluid moves in the positive z-direction in the upper part of the flow channel (due to motion of the upper boundary) and in the negative z-direction in the lower part of the channel (see the $K*=-4$ case in Fig. 6).

The volumetric flow rate Q and the average velocity u_{avg} associated with this flow are given by

$$Q = u_{\mathrm{avg}}\, b w = \int_0^b u_z\, w \mathrm{d}y = \frac{b^3 w \Delta P}{12\eta L} + \frac{1}{2} b w U \tag{2.75}$$

The first term on the right represents the flow contribution arising from the pressure gradient and the second that from the moving boundary. When $K*=-3$, the volumetric flow rate associated with the moving boundary is balanced by the flow rate arising from the adverse pressure gradient and $Q=0$.

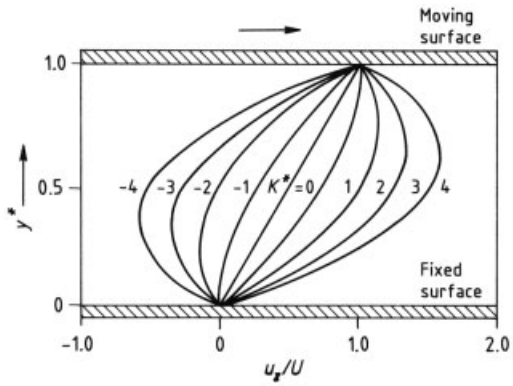

Figure 6. Dimensionless velocity profiles for generalized Couette flow

If the fluid is non-Newtonian, significantly more complex flow equations are obtained for generalized Couette flow. For the case of power-law, viscous non-Newtonian behavior, see [28], [29] (see also → Fluid Mechanics, Chap. 4.1.2.).

2.2.2. One-Dimensional Poiseuille Flows (→ Fluid Mechanics)

When $U=0$, the flow illustrated in Figure 5 becomes a pressure-driven Poiseuille flow. Under such conditions, the velocity profile becomes parabolic with a plane of symmetry at $y=b/2$. Specifically, in terms of a y coordi-

nate measured from the channel centerline, i.e., $y' = y - b/2$, Equation (2.69) takes the form

$$u_z = \left(\frac{b^2 \Delta P}{8\eta L}\right)\left[1 - 4\left(\frac{y'}{b}\right)^2\right] \tag{2.76}$$

and the volumetric flow rate reduces to a Hagen – Poiseuille relation where Q is proportional to $(P_0 - P_L)/L$, i.e.,

$$Q = u_{\mathrm{avg}} b w = \frac{b^3 w \Delta P}{12\eta L} \tag{2.77}$$

In the case of one-dimensional Poiseuille flows in channels with circular and concentric annulus cross-sectional geometries, analogous results can be obtained . In all cases the volumetric flow rate is proportional to the pressure drop (for further details, see → Fluid Mechanics, Chap. 3.3.1.).

An alternate representation of flow rate – pressure drop results for unidirectional flows is in terms of the friction factor relation. In particular, the wall shear stress $\tau_{\mathrm{w}} \equiv -\eta \left.\frac{\mathrm{d}u_z}{\mathrm{d}y'}\right|_{\mathrm{w}}$ is assumed to be proportional to the kinetic energy per unit volume expressed in terms of the average velocity u_{avg}:

$$\tau_{\mathrm{w}} = f\left(\frac{1}{2}\varrho u_{\mathrm{avg}}^2\right) \tag{2.78}$$

Here, the constant of proportionality is the friction factor f. Since the pressure force must be balanced by the drag force on the walls

$$\Delta P A_{\mathrm{c}} = \tau_{\mathrm{w}} A_{\mathrm{w}} = f A_{\mathrm{w}}\left(\frac{1}{2}\varrho u_{\mathrm{avg}}^2\right) \tag{2.79}$$

Here A_{c} is the cross-sectional area and A_{w} is the wall surface area. The friction factor can then be expressed as

$$f Re = C \equiv \frac{l^2 (\Delta P/L)}{2\eta u_{\mathrm{avg}}} \tag{2.80}$$

where the Reynolds number Re and the characteristic length l are defined by

$$Re = \frac{\varrho u_{\mathrm{avg}} l}{\eta} \tag{2.81}$$

and

$$l = \frac{4 A_{\mathrm{c}} L}{A_{\mathrm{w}}} \tag{2.82}$$

For plane Poiseuille flow where $l = 2b$, Equation (2.77) can be used in Equation (2.80) to obtain

$$f Re = C = 24 \tag{2.83}$$

where $Re \equiv 2\varrho u_{\mathrm{avg}} b/\eta$.

Another flow that arises in many industrial applications is Poiseuille flow in a tube. If the tube has radius r_1 and length L, the stress and velocity fields for incompressible Newtonian fluids are

$$\tau_{rz} = \tau_{zr} = \frac{1}{2}\left(\frac{\Delta P}{L}\right) r \tag{2.84}$$

$$u_z = \frac{r_1^2}{4\eta}\left(\frac{\Delta P}{L}\right)\left[1 - \left(\frac{r}{r_1}\right)^2\right] \tag{2.85}$$

where r is the radial coordinate. The volumetric flow rate is then

$$Q = u_{\mathrm{avg}} \pi r_1^2 = 2\pi \int_0^{r_1} u_z r \mathrm{d}r = \frac{\pi r_1^4}{8\eta}\left(\frac{\Delta P}{L}\right) \tag{2.86}$$

If the latter result is used in Equation (2.80), then

$$f Re = 16 \tag{2.87}$$

where the characteristic length l is $2r_1$, and the Reynolds number is defined by $Re \equiv 2\varrho u_{\mathrm{avg}} r_1/\eta$.

Similar results can be obtained for axial flow in the annular region between concentric cylinders of radii r_0 and r_1 where $r_1 > r_0$. In this case, the characteristic length is $l = 2(r_1 - r_0)$, and the value of C in Equation (2.80) varies with the ratio $\chi \equiv r_0/r_1$. In the limit of small values of r_0/r_1, C takes on the value of 16; it approaches an upper limit of 24 as r_0/r_1 approaches 1.

2.2.3. Flow in Channels of Arbitrary Cross Section

For flows in channels with more complex cross sections (e.g., flow in rectangular, elliptic, or eccentric annular channels) or flow in channels with irregular cross sections, the velocity u_z is a function of two spatial variables, i.e.,

$$u_z = u_z(x, y) \tag{2.88}$$

For incompressible Newtonian fluids, Equation (2.65) is applicable, and the solution to this equation is given by

$$u_z = \frac{1}{4\eta}\left(x^2 + y^2\right)\left(\frac{\mathrm{d}P}{\mathrm{d}z}\right) + \varphi\left(x, y\right) \tag{2.89}$$

where the function $\varphi(x, y)$ is a solution of Laplace's equation

$$\frac{\partial^2\varphi}{\partial x^2} + \frac{\partial^2\varphi}{\partial y^2} = 0 \tag{2.90}$$

If a no-slip velocity condition is assumed (i.e., $u_{z,\mathrm{S}} = 0$), then

$$\varphi = \varphi_{\mathrm{S}} = \left[-\frac{1}{4\eta}\left(x^2 + y^2\right)\left(\frac{\mathrm{d}P}{\mathrm{d}z}\right)\right]_{\mathrm{S}} \tag{2.91}$$

on the bounding surfaces.

For a channel with an *elliptic cross section* defined by $(x/x_0)^2 + (y/y_0)^2 = 1$, the velocity is given by [30, p. 38]

$$u_z = \frac{x_0^2 y_0^2 \Delta P}{2\eta\left(x_0^2 + y_0^2\right) L}\left[1 - \left(\frac{x}{x_0}\right)^2 - \left(\frac{y}{y_0}\right)^2\right] \tag{2.92}$$

and the volumetric flow rate by

$$Q = u_{\mathrm{avg}}\pi x_0 y_0 = \frac{\pi x_0^3 y_0^3 \Delta P}{4\eta\left(x_0^2 + y_0^2\right) L} \tag{2.93}$$

In terms of the friction factor, this result can be written

$$f\,Re = \frac{2\pi^2}{[E\,(k)]^2}\left[1 + \left(\frac{y_0}{x_0}\right)^2\right] = \left\{\text{function of } \frac{y_0}{x_0}\right\} \tag{2.94}$$

where $E\,(k)$ is the complete elliptic integral with argument $k = [1 - (y_0/x_0)^2]$ [31, p. 509]. When $y_0/x_0 \to 0$, $fRe \to 19.74$, and when $y_0/x_0 = 1$, $fRe = 16$ (the cylindrical tube result).

For flow in channels with *square and equilateral-triangle cross sections* with sides a, the flow rate – pressure drop relations are given by

$$\text{Square}: \quad f\,Re = 14.2 \tag{2.95}$$

$$\text{Triangular}: \quad f\,Re = 13.3 \tag{2.96}$$

Results for other cross sections (rectangular and eccentric annuli) are given in [30, pp. 33 – 39]. For even more complex and less regular cross-sectional geometries, estimates (sometimes crude) can be obtained by using a hydraulic radius approach where an effective radius is defined for the channel, i.e., $r_{\mathrm{H}} \equiv l/2 = 2\,A_{\mathrm{c}}\,L/A_{\mathrm{w}}$, and C in Equation (2.80) is taken to be 16, the value corresponding to Poiseuille flow in a tube. However, this approach is approximate and can lead to sizable errors if the actual C for the geometry differs significantly from 16. For flow in channels with elliptic cross sections, the maximum error would be ca. 19 %; for square channels, the error would be 12 %.

All of these results are for laminar flow. When the Reynolds number reaches sufficiently high values, the flow becomes unstable and turbulent flow arises. For flow in circular tubes, turbulent flow is generally observed when $Re \equiv 2\,\varrho\,u_{\mathrm{avg}}\,r_1/\eta > 2100$. For other geometries, the transition values are different and must be determined in each individual case.

2.2.4. Poiseuille Flow of Non-Newtonian Fluids (→ Fluid Mechanics, Chap. 4.2.)

The flow characteristics of non-Newtonian fluids in tubes and planar channels can be quite different from those observed with Newtonian fluids. This is illustrated by considering the steady flow of a power-law, non-Newtonian fluid in a tube of radius r_0. As in the Newtonian case, the only nonzero stress components are $\tau_{rz} = \tau_{zr}$, and these are described by Equation (2.84). Because $\mathrm{d}u_z/\mathrm{d}r < 0$ for this flow, the shear stress can be written as

$$\tau_{rz} = m\left(-\frac{\mathrm{d}u_z}{\mathrm{d}r}\right)^n \tag{2.97}$$

Combining this equation with Equation (2.84), integrating, and using the no-slip boundary condition $u_z = 0$ at the tube wall give

$$u_z = \left(\frac{\tau_w}{m}\right)^s \frac{r_1}{s+1}\left[1 - \left(\frac{r}{r_1}\right)^{s+1}\right] \tag{2.98}$$

where $s \equiv 1/n$, and τ_{w} is the wall shear stress given by

$$\tau_w \equiv |\tau_{rz}|_{r=r_1} = \frac{1}{2}\left(\frac{\Delta P}{L}\right) r_1 \tag{2.99}$$

Because the maximum velocity occurs at $r = 0$, Equation (2.98) can also be written as

$$u_z = u_{\max}\left[1 - \left(\frac{r}{r_1}\right)^{s+1}\right] \tag{2.100}$$

with $u_{\max}$ given by

$$u_{\max} = \left(\frac{\tau_w}{m}\right)^s \frac{r_1}{(s+1)} \tag{2.101}$$

For $s = 1$ (or $n = 1$), the behavior is Newtonian and the profile is parabolic. For pseudoplastic fluids where $s > 1$ (or $n < 1$), the profile becomes more blunted and pluglike as s takes on larger and larger values (see Fig. 7). For dilatant fluids where $s < 1$ (or $n > 1$), the velocity profile becomes more and more pointed as s decreases.

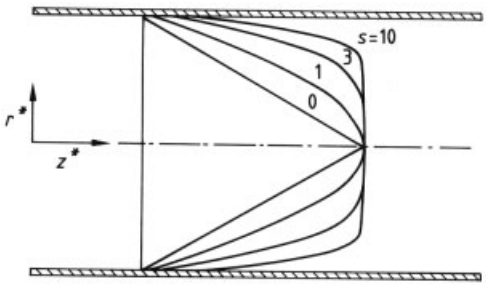

Figure 7. Velocity profiles for the flow of a power-law fluid in a tube
The parameter s is the reciprocal of the power-law slope parameter n. When $s = 1$, Newtonian behavior is observed. When $s > 0$, the behavior is dilatant.

Comparison of the average velocity with the maximum velocity gives

$$\frac{u_{\text{avg}}}{u_{\max}} = \frac{s+1}{s+3} \tag{2.102}$$

For Newtonian fluids ($s = 1$), this reduces to $u_{\text{avg}}/u_{\max} = 1/2$. For pseudoplastic fluids ($s > 1$), this ratio increases with s and approaches the limit of 1 (plug flow) as s gets very large ($n \to 0$). For dilatant fluids, $u_{\text{avg}}/u_{\max}$ ranges from 1/2 to 1/3 as s decreases (or n increases). These results are illustrated in Figure 7.

The flow rate – pressure drop expression for this flow is obtained directly from $Q = \pi r_1^2 u_{\text{avg}}$. For values of s different from 1, this relation is nonlinear, with $\mathrm{d}(\Delta P/L)/\mathrm{d}Q$ increasing with Q for dilatant fluids and $\mathrm{d}(\Delta P/L)/\mathrm{d}Q$ decreasing with Q for pseudoplastic fluids.

In the case of a Bingham plastic with a yield stress τ_Y and a plastic viscosity η_p, the nature of the flow depends on the radial position of the yield surface, i.e., the position $r = r_Y$ where $-\tau_{rz}- = \tau_Y$. From Equation (2.84)

$$r_Y = \frac{\tau_Y r_1}{\tau_w} = \frac{2\tau_Y}{(\Delta P/L)} \tag{2.103}$$

At radial positions outside the yield surface ($r > r_Y$), the magnitude of the shear stress exceeds the yield stress and a shear flow [$u_z = u_z(r)$] results. At radial positions inside the yield surface ($r < r_Y$), the magnitude of the shear stress is less than τ_Y and a uniform plug flow occurs with velocity $u_z(r_Y)$. As $(\Delta P/L)$ decreases, the radial position of the yield surface moves closer and closer to the tube wall. At a value of $(\Delta P/L) = 2\tau_Y/r_1$, the position of the yield surface coincides with the wall and flow ceases. Hence, application of a finite pressure gradient does not ensure flow in the case of materials with plastic behavior.

Flow rate results for Newtonian fluids, power-law fluids, and Bingham plastics may be summarized as follows:

Flow geometry: Circular tube radius r_1, length L

Newtonian fluid (viscosity η)

$$Q = \frac{\pi r_1^4}{8\eta}\frac{\Delta P}{L}$$

Power-law fluid (parameters m, n)

$$Q = \frac{\pi r_1^3}{(s+3)}\left(\frac{\Delta P r_1}{2Lm}\right)^s, \quad s \equiv \frac{1}{n}$$

Bingham plastic (parameters τ_Y, η_p)

$$Q = \frac{\pi r_1^3 \tau_w}{4\eta_p}\left[1 - \frac{4}{3}\left(\frac{\tau_y}{\tau_w}\right) + \frac{1}{3}\left(\frac{\tau_y}{\tau_w}\right)^4\right],$$
$$\text{when } \tau_w > \tau_y$$

where

$$\tau_w \equiv \frac{\Delta P r_1}{2L}$$
$$Q = 0, \quad \text{when } \tau_w \le \tau_y$$

Expressions for other types of non-Newtonian fluids can be found in various sources. These developments are described and the results summarized for various non-Newtonian fluid models [32, Chap. 5], [33–35]. Further details are given in → Fluid Mechanics, Chap. 4.2.1.

2.2.5. Two-Phase Concentric Flow in a Tube – Segregated Flow

An interesting example of viscous, one-dimensional laminar flow in a tube takes place when two immiscible fluids are injected into a tube of radius r_1 such that one fluid occupies the core region $r < r_c$ and the other fluid the outer region $r > r_c$ (see Fig. 8). The core fluid has viscosity η and the outer fluid viscosity $\hat{\eta}$.

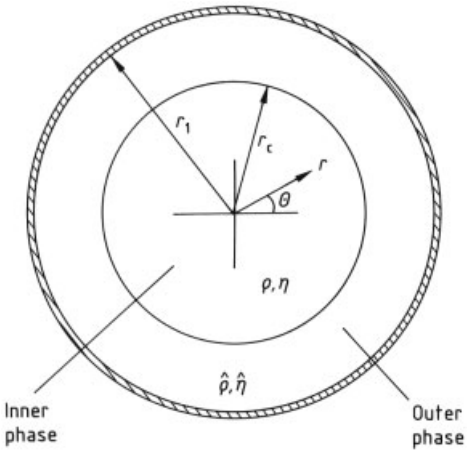

Figure 8. Segregated two-phase flow in a tube

Such problems are often encountered in pipelines where the flow of a viscous fluid is facilitated by the injection of a second low-viscosity immiscible phase. The interest here is in the flow rate enhancement that results from the injection of this second fluid.

The analysis of flow in the separate fluid regions is similar to that for Poiseuille flow in cylindrical tubes and annuli. In particular, for Newtonian fluids, Equations (2.64) and (2.65) can be integrated to give the stress and velocity fields for the inner fluid region as

$$\tau_{rz} = \frac{1}{2}\left(\frac{\Delta P}{L}\right) r + \frac{C_1}{r} \tag{2.104}$$

$$u_z = -\frac{1}{4\eta}\left(\frac{\Delta P}{L}\right) r^2 - \frac{C_1}{\eta}\ln r + C_2 \tag{2.105}$$

where C_1 and C_2 are the integrating constants. Identical expressions are obtained for the outer fluid region except that τ_{rz}, u_z, η, C_1, and C_2 are replaced by $\hat{\tau}_{rz}$, $\hat{u}_z$, $\hat{\eta}$, $\hat{C}_1$, and $\hat{C}_2$. It writing these expressions, the axial pressure distributions in the two phases are assumed to be equal.

The integrating constants can be evaluated by using the conditions (1) that the stress τ_{rz} *is* finite at $r = 0$, (2) that the velocity u_z is zero at the tube wall, and (3) that the stress and velocity are continuous at the boundary $r = r_c$ between the phases [$\tau_{rz}(r_c) = \hat{\tau}_{rz}(r_c)$ and $u_z(r_c) = \hat{u}_z(r_c)$]. Integrating the respective velocity profiles over the inner and outer fluid regions according to Equation (2.86) gives the corresponding volumetric flow rates

$$Q = \frac{\pi}{8\hat{\eta}}\left(\frac{\Delta P}{L}\right) r_1^4 \left[\frac{\hat{\eta}}{\eta} r_c^{*4} + 2\left(1 - r_c^{*2}\right) r_c^{*2}\right] \tag{2.106}$$

$$\hat{Q} = \frac{\pi}{8\hat{\eta}}\left(\frac{\Delta P}{L}\right) r_1^4 \left(1 - r_c^{*2}\right)^2 \tag{2.107}$$

Here, $r_c* \equiv r_c/r_1$ is the dimensionless radial position of the boundary between the inner and outer fluid regions.

Eliminating $\Delta P/L$ between these equations gives an implicit expression for r_c*:

$$(1-\varphi)\left(1 - r_c^{*2}\right)^2 - \varphi r_c^{*2}\left[\frac{\hat{\eta}}{\eta} r_c^{*2} + 2\left(1 - r_c^{*2}\right)\right] = 0 \tag{2.108}$$

where

$$\varphi = \frac{\hat{Q}}{\hat{Q} + Q} \tag{2.109}$$

is the fractional flow of the outer fluid. Specification of φ and the viscosity ratio $\hat{\eta}/\eta$ then allows the determination of r_c* by using simple numerical root finding methods. The flow rate enhancement for the inner fluid is

$$Q^* \equiv \frac{Q}{(Q)_{r_c^*=1}} = r_c^{*4} + 2\left(\frac{\eta}{\hat{\eta}}\right)\left(1 - r_c^{*2}\right) r_c^{*2} \tag{2.110}$$

Here $(Q)_{r_c^*=1}$ is the volumetric flow rate when the inner fluid occupies the entire flow region (no outer fluid present).

For a specified viscosity ratio $\hat{\eta}/\eta$, Equations (2.108) – (2.110) can be used to determine the flow rate enhancement in terms of the fractional flow of the outer fluid and the viscosity ratio. The numerical results for several values of viscosity ratio are given in Figure 9. In particular, for the case in which the outer fluid is 100 times less viscous than the inner fluid ($\hat{\eta}/\eta = 0.01$) and the

fractional flow is 10 %, the flow rate of the more viscous inner fluid is enhanced by a factor of 30. For higher fractional flows and lower viscosity ratios, even greater enhancement factors are obtained. Clearly, when such segregated flows are possible, significant increases in throughput can be achieved per unit of applied pressure drop. In practice, the application of such approaches is limited only by the ability to create and maintain a stable outer concentric region of the lower viscosity fluid.

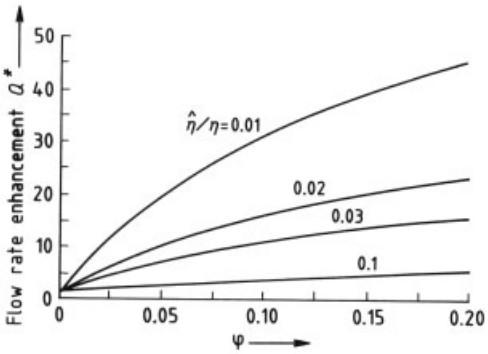

Figure 9. Flow rate enhancement associated with the injection of an immiscible low-viscosity fluid into a transport line

2.3. Multidimensional Momentum Transfer

2.3.1. Two-Dimensional Flows – Stream Function Equations

The flow examples considered to this point have involved only one nonvanishing velocity component. More complex flows are now considered, particularly two-dimensional incompressible flows of Newtonian fluids having two nonvanishing velocity components that depend on two spatial variables and time. The governing equations are the continuity equation and the two nonvanishing components of the equation of motion. Rather than solving these equations directly, the equations are more conveniently transformed into a single equation in terms of a stream function ψ (→ Fluid Mechanics, Chap. 2.1.). Because the stream function is defined in terms of the nonvanishing velocity components, solution of this equation then provides the velocity field. Physically, lines of constant ψ represent streamlines of the flow.

To illustrate this more explicitly, consider the flow

$$u_x = u_x(x, y, t), \; u_y = u_y(x, y, t), \; u_z = 0 \quad (2.111)$$

where the governing equations of continuity and motion are

$$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} = 0 \quad (2.112)$$

and

$$\varrho\left(\frac{\partial u_x}{\partial t} + u_x\frac{\partial u_x}{\partial x} + u_y\frac{\partial u_x}{\partial y}\right) = -\frac{\partial P}{\partial x} + \eta\left(\frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2}\right) \quad (2.113)$$

$$\varrho\left(\frac{\partial u_y}{\partial t} + u_x\frac{\partial u_y}{\partial x} + u_y\frac{\partial u_y}{\partial y}\right) = -\frac{\partial P}{\partial y} + \eta\left(\frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2}\right) \quad (2.114)$$

If the stream function ψ is defined in terms of the velocity components by

$$u_x = -\frac{\partial \psi}{\partial y} \quad (2.115)$$

$$u_y = \frac{\partial \psi}{\partial x} \quad (2.116)$$

the continuity equation (Eq. 2.112) is satisfied identically. To obtain the desired stream function equation, the motion equations are combined in the following way. First operate on Equation (2.113) by $\partial/\partial y$ and Equation (2.114) by $\partial/\partial x$; then subtract the resulting equations to eliminate the pressure terms; finally, replace u_x and u_y with Equations (2.115) and (2.116). After simplification, the resulting equation is

$$\frac{\partial\left(\nabla^2\psi\right)}{\partial t} + \frac{\partial\left(\psi, \nabla^2\psi\right)}{\partial(x, y)} = \nu\nabla^4\psi \quad (2.117)$$

where $\nu \equiv \eta/\varrho$ and

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}, \; \nabla^4 \equiv \nabla^2\nabla^2 \quad (2.118)$$

Also,

$$\frac{\partial(f, g)}{\partial(x, y)} \equiv \begin{vmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{vmatrix} \quad (2.119)$$

where a determinant operation is implied by the straight brackets.

Solution of Equation (2.117) gives the stream function $\psi(x, y, t)$. The velocity components are then obtained from Equations (2.115) and (2.116), and the pressure P is obtained from these velocity components and Equations (2.113) and (2.114).

A similar approach can be used in other two-dimensional flows, but the velocity – stream function relations and the stream function equation may take different forms depending on flow characteristics. For example, for axisymmetric flow around a sphere (see Fig. 10) where

$$u_r = u_r(r, \theta, t), \; u_\theta = u_\theta(r, \theta, t), \; u_\varphi = 0 \quad (2.120)$$

the velocity – stream function relations are

$$u_r = \frac{1}{r^2 \sin\theta} \frac{\partial \psi}{\partial \theta} \quad (2.121)$$

$$u_\theta = -\frac{1}{r \sin\theta} \frac{\partial \psi}{\partial r} \quad (2.122)$$

and the stream function equation is

$$\frac{\partial\left(E^2\psi\right)}{\partial t} - \frac{1}{r^2 \sin\theta} \frac{\partial\left(\psi, E^2\psi\right)}{\partial(r, \theta)} + \frac{2E^2\psi}{r^2 \sin^2\theta} \cdot \left(\frac{\partial \psi}{\partial r}\cos\theta - \frac{1}{r}\frac{\partial \psi}{\partial \theta}\sin\theta\right) = \nu E^2\left(E^2\psi\right) \quad (2.123)$$

where E^2 is an operator defined by

$$E^2 \equiv \frac{\partial^2}{\partial r^2} + \frac{\sin\theta}{r^2}\frac{\partial}{\partial \theta}\left(\frac{1}{\sin\theta}\frac{\partial}{\partial \theta}\right) \quad (2.124)$$

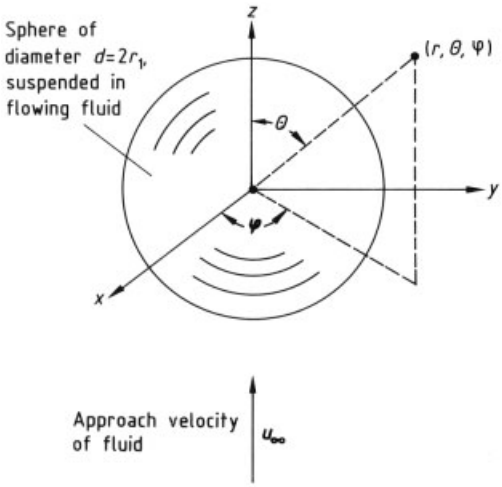

Figure 10. Axisymmetric flow past a sphere

Similar relations for other flows are given in [17, p. 131], [36, pp. 114 – 115]. Specification of the velocity in terms of the stream function ψ is given by

$$\boldsymbol{u} = \boldsymbol{\nabla}\psi \times \boldsymbol{\nabla}G \quad (2.125)$$

where G is a scalar function. This form satisfies the continuity equation ($\nabla \cdot \boldsymbol{u} = 0$) for incompressible flow [20, pp. 227 – 228]. For axisymmetric flow, the scalar function G is simply the coordinate ξ corresponding to the azimuthal angle about the axis of symmetry. For axisymmetric flow around a sphere, $G = \varphi$ and the resulting components of Equation (2.125) are the same as those of Equations (2.121) and (2.122). For two-dimensional axisymmetric flow in a cylindrical geometry with $u_\theta = 0$, $G = \theta$ is used, and the velocity – stream function relations take the forms

$$u_r = -\frac{1}{r}\frac{\partial \psi}{\partial z}, \quad u_z = \frac{1}{r}\frac{\partial \psi}{\partial r} \quad (2.126)$$

If Equation (2.123) is put in dimensionless form by scaling t, r, and ψ using d/u_∞, d, and $u_\infty d^2$, respectively, where d is the sphere diameter and u_∞ the free stream velocity, then

$$Re_{\mathrm{d}}\{\text{left} - \text{hand side terms}\} = E^{*4}\psi^* \quad (2.127)$$

where $Re_{\mathrm{d}} \equiv \varrho u_\infty d/\eta$ is the Reynolds number. Since all terms except Re_{d} are scaled to order 1 by the selections of d and u_∞, when $Re_{\mathrm{d}} \ll 1$ the left-hand side can be set to zero and a *creeping flow* solution can be obtained. The latter is generally associated with small spheres, viscous fluids, and low velocities. When $Re_{\mathrm{d}} \gg 1$, the right-hand side can be set to zero and a nonviscous flow solution can be obtained.

2.3.2. Creeping Flow Around a Sphere and Other Bodies of Revolution

Consider the steady, axisymmetrical flow of an incompressible Newtonian fluid past a sphere as illustrated in Figure 10. If the coordinate system is attached to the sphere, this representation is valid either for flow past a stationary sphere as illustrated or for the steady fall of a sphere (at terminal velocity u_∞) in a fluid at rest. The velocity and pressure fields under creeping flow

conditions ($Re_{\mathrm{d}} \ll 1$) are of particular interest, along with the drag force associated with viscous and pressure effects.

Under creeping flow conditions, Equation (2.123) reduces to

$$E^4 \psi = 0 \tag{2.128}$$

with boundary conditions

$$\text{at} \quad r = r_1, \quad u_r = \frac{1}{r^2 \sin\theta} \frac{\partial \psi}{\partial \theta} = 0 \tag{2.129}$$

$$\text{at} \quad r = r_1, \quad u_\theta = -\frac{1}{r \sin\theta} \frac{\partial \psi}{\partial r} = 0 \tag{2.130}$$

$$\text{as} \quad r \to \infty, \quad u_z = \frac{u_r}{\cos\theta} = -\frac{u_\theta}{\sin\theta} \to u_\infty \tag{2.131}$$

By using Equations (2.121) and (2.122), the latter condition can be written as

$$r \to \infty, \quad \psi \to \frac{1}{2} u_\infty r^2 \sin^2\theta \tag{2.132}$$

Because Equation (2.132) is valid at all θ, the θ dependence at smaller r values might be expected to be the same. This suggests a solution to Equation (2.128) of the form

$$\psi = f(r) \sin^2\theta \tag{2.133}$$

Substituting this expression into Equation (2.128) and using Equations (2.129), (2.130), and (2.132) give

$$f = -\frac{1}{4} u_\infty \left(\frac{r_1^3}{r} + 3 r_1 r - 2 r^2 \right) \tag{2.134}$$

Then, from Equations (2.133), (2.121), and (2.122),

$$u_r = -u_\infty \left[1 - \frac{3}{2} \left(\frac{r_1}{r} \right) + \frac{1}{2} \left(\frac{r_1}{r} \right)^3 \right] \cos\theta \tag{2.135}$$

$$u_\theta = -u_\infty \left[1 - \frac{3}{4} \left(\frac{r_1}{r} \right) - \frac{1}{4} \left(\frac{r_1}{r} \right)^3 \right] \sin\theta \tag{2.136}$$

The dynamic pressure P is obtained from a line integration between the position $r \to \infty$ and $\theta = \pi/2$ (where $P = 0$) and any position r, θ [where $P = P(r, \theta)$], i.e.,

$$P = \int_0^{P(r,\theta)} \mathrm{d}P = \int_\infty^r \left(\frac{\partial P}{\partial r} \right)_{\theta=\pi/2} \mathrm{d}r + \int_{\pi/2}^{\theta} \left(\frac{\partial P}{\partial \theta} \right)_r \mathrm{d}\theta \tag{2.137}$$

If $(\partial P/\partial r)$ and $(\partial P/\partial \theta)$ are obtained from the r and θ components of the equation of motion (creeping flow conditions) by using Equations (2.135) and (2.136), then

$$P = -\frac{3 \eta u_\infty}{2 r_1} \left(\frac{r_1}{r} \right)^2 \cos\theta \tag{2.138}$$

Also, the hydrostatic pressure is

$$p_{\mathrm{H}} = p_0 - \varrho g z \tag{2.139}$$

where p_0 is the pressure at $r \to \infty$ and $\theta = \pi/2$. Combining the dynamic and hydrostatic results (Eq. 2.62) gives the pressure field

$$p = p_0 - \varrho g z - \frac{3 \eta u_\infty}{2 r_1} \left(\frac{r_1}{r} \right)^2 \cos\theta \tag{2.140}$$

The drag force of the fluid on the sphere is then obtained from

$$\boldsymbol{F} = \int_S \boldsymbol{n} \cdot (-p \boldsymbol{I} - \boldsymbol{\tau}) \, \mathrm{d}S \tag{2.141}$$

where $\boldsymbol{n} = \boldsymbol{i}_r$ is the outwardly directed unit normal to the sphere surface. Expanding Equation (2.141) and finding the component in the z-direction give

$$F_z = -\int_0^{2\pi} \int_0^{\pi} [(p + \tau_{rr}) \cos\theta - \tau_{r\theta} \sin\theta]_{r_1} \, r_1^2 \sin^2\theta \, \mathrm{d}\theta \mathrm{d}\varphi \tag{2.142}$$

Then, by using Equations (2.135) and (2.136), the stress components and the pressure are evaluated at r_1 with the results

$$\tau_{rr}\Big|_{r_1} = -2\eta \frac{\partial u_r}{\partial r}\Big|_{r_1} = 0 \tag{2.143}$$

$$\tau_{r\theta}\Big|_{r_1} = -\eta \left[r \frac{\partial}{\partial r} \left(\frac{u_r}{r} \right) + \frac{1}{r} \frac{\partial u_r}{\partial \theta} \right]_{r_1} = \frac{3 \eta u_\infty}{r_1} \sin\theta \tag{2.144}$$

and

$$p\Big|_{r_1} = p_0 - \varrho g r_1 \cos\theta - \frac{3 \eta u_\infty}{2 r_1} \cos\theta \tag{2.145}$$

Thus,

$$F_z = \underset{\substack{\text{Buoyancy}\\ \text{from } p_H}}{4/3\,\pi\, r_1^3\, \varrho\, g} + \underset{\substack{\text{Form drag}\\ \text{from } P}}{2\,\pi\,\eta\, r_1\, u_\infty} + \underset{\substack{\text{Viscous drag}\\ \text{from } \tau_{r\theta}}}{4\,\pi\,\eta\, r_1\, u_\infty} \tag{2.146}$$

The first term arises from hydrostatic pressure effects around the sphere (pressure is higher as $\theta \to \pi$ causing a net upward force); the second term, from dynamic pressure effects (again from higher pressure on the sphere bottom); and the third, from viscous effects associated with $\tau_{r\theta}$ on the sphere surface. If this result is represented in terms of a buoyancy force F_B (static effect) and a drag force F_D (dynamic effect), i.e.,

$$F_z = F_B + F_D \tag{2.147}$$

then F_B is just the first term of Equation (2.146) and F_D the sum of the other two terms, i.e.,

$$F_D = 6\,\pi\,\eta\, r_1\, u_\infty \tag{2.148}$$

This equation is known as *Stokes law* and indicates that the drag force on a solid sphere increases in direct proportion to viscosity, radius, and terminal (or free stream) velocity.

If a sphere of density ϱ_s settles under its own weight in a fluid, the terminal velocity can be determined by equating the sum of F_B and F_D to the gravitational force on the sphere ($4/3\,\pi\, r_1^3\, \varrho_s\, g$). The result is

$$u_\infty = \frac{2\,(\varrho_s - \varrho)\, r_1^2\, g}{9\,\eta} \tag{2.149}$$

This result indicates that for particle settling in viscous fluids, the terminal velocity is most sensitive to variations in particle size.

Finally, the drag force results for several variations of this problem are noted. In particular, for creeping flow past a *fluid sphere* (or droplet) of viscosity $\hat{\eta}$ [30, pp. 127 – 129]

$$F_D = 6\,\pi\,\eta\, r_1\, u_\infty \left(\frac{3\,\hat{\eta} + 2\,\eta}{3\,\hat{\eta} + 3\,\eta}\right) \tag{2.150}$$

For a *very viscous droplet* or *solid sphere* where $\hat{\eta} \gg \eta$, this result reduces to Stokes law. For *gas bubbles* where $\hat{\eta} \ll \eta$

$$F_D = 4\,\pi\,\eta\, r_1\, u_\infty \tag{2.151}$$

In this case, the drag is two-thirds that of a rigid sphere.

If *surface-active agents* are present that give rise to a surface shear viscosity ε_s, the drag force for a fluid sphere is [30, pp. 129]:

$$F_D = 6\,\pi\,\eta\, r_1\, u_\infty \left(\frac{(\varepsilon_s/r_1) + 3\,\hat{\eta} + 2\,\eta}{(\varepsilon_s/r_1) + 3\,\hat{\eta} + 3\,\eta}\right) \tag{2.152}$$

The important point here is that if r_1 is sufficiently small, surface viscosity effects dominate to give rigid sphere results, even when $\hat{\eta}$ is small (gas bubble case). Small bubbles often behave as rigid spheres when contaminant amounts of surface-active agents are present.

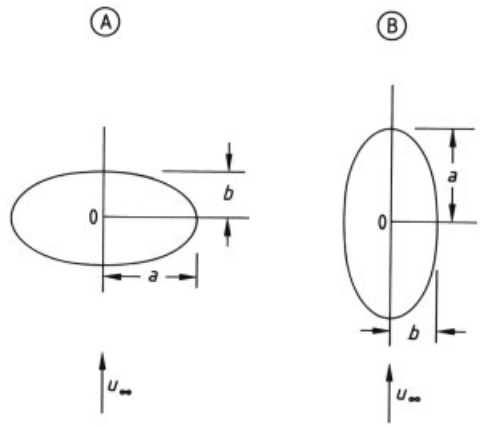

Figure 11. Drag force for creeping flow around oblate and prolate spheroids
A) Oblate spheroid, drag force $F_D = 6\,\pi\,\eta a\, u_\infty K$; B) Prolate spheroid, drag force $F_D = 6\,\pi\,\eta\, b\, u_\infty K$

b/a	K (oblate)	*K* (prolate)
0.0	0.8488	∞
0.1	0.8525	2.6471
0.5	0.9053	1.2039
1.0	1.0000	1.0000

For *nonspherical* solid bodies, somewhat different drag force results are obtained. For oblate and prolate spheroids, the results are summarized in Figure 11. Note that the drag force for the prolate spheroid is particularly sensitive to the ratio of the minor and major axes. In the limit as *b/a* goes to zero, the oblate spheroid reduces to a circular flat disk with the drag force given by

$$F_D = 5.093\,\pi\,\eta\, a\, u_\infty \tag{2.153}$$

where *a* is the radius of the flat disk. In the limit as *b/a* goes to zero for the prolate spheroid, the

results for an elongated rod of radius b and length $2a$ are

$$F_{\mathrm{D}} = \frac{4\pi\eta a u_\infty}{\ln(a/b) + 0.1932} \tag{2.154}$$

Additional drag force results for creeping flow around bodies are given in [30].

An alternate way of representing drag force results around bodies is in terms of a drag coefficient C_{D} (→ Fluid Mechanics, Chap. 3.3.3.):

$$F_{\mathrm{D}} = C_{\mathrm{D}}\,(\text{Characteristic area of body})\cdot (\text{Characteristic kinetic energy of fluid per unit volume}) \tag{2.155}$$

For flow past a sphere

$$F_{\mathrm{D}} = C_{\mathrm{D}}\left(\frac{\pi}{4}\mathrm{d}^2\right)\left(\frac{1}{2}\varrho u_\infty\right)^2 \tag{2.156}$$

$$C_{\mathrm{D}} = \frac{24}{Re_{\mathrm{d}}} \tag{2.156a}$$

For the settling of multiple bodies, interaction effects can be quite important, even over fairly large separation distances. In the case of two identical spheres of diameter d settling either parallel or perpendicular to their line of centers, the drag coefficient on each sphere is

$$C_{\mathrm{D}} = \frac{24\Lambda}{Re_{\mathrm{d}}} \tag{2.157}$$

where Λ is an interference parameter which is a function of the ratio l/d of the distance between the particles and the particle diameter. When $l/d \to \infty$, $\Lambda \to 1$. For two spheres settling side by side, Λ is found to be < 1 for finite values of l/d. As a result, two interfering spheres fall faster than isolated spheres.

Even though two spheres may fall faster than an isolated sphere, a suspension of many spheres usually experiences a higher drag than the isolated sphere case. As a result, when a suspension settles in a beaker, a sharp line is observed separating the clear fluid from the particle-laden fluid. This hindered settling phenomenon arises because particles in the low-concentration regions near the top of the settling suspension tend to overtake the more slowly settling particles in the underlying high-concentration regions.

2.3.3. Flow in Channels with Varying Cross Sections – Lubrication Analysis

In many problems the flow geometry is sufficiently complex that exact solutions of the motion equations are not possible, and approximate or numerical methods are required to obtain the solutions desired. In this section an approximate analysis is described, called lubrication theory. (The name lubrication theory arose from early applications of this approximate analysis to thin-film lubrication problems). This method is applicable when the change in flow channel geometry in the direction of flow is weak.

Consider steady two-dimensional flow $\boldsymbol{u} = [u_x(x, y),\ u_y(x, y), 0]$ in a channel with a slowly changing cross-sectional area as shown in Figure 12. The flow channel gap b varies with x, but the magnitude of this change at any x is small, i.e.,

$$\varepsilon \equiv \frac{\mathrm{d}b}{\mathrm{d}x} \ll 1 \tag{2.158}$$

The equation of continuity and the x and y components of the equation of motion are the same as Equations (2.112) – (2.114), but without the time derivative terms. To scale these equations and determine the most dominant terms, the velocity component u_x must first be scaled to the characteristic average velocity u_0 in the channel:

$$u_x \sim u_{\mathrm{avg}}(x) = \frac{Q}{b(x)\,w} \sim \frac{Q}{b_0 w} \equiv u_0 \tag{2.159}$$

Here Q is the volumetric flow rate, w the channel width, b_0 the channel gap at $x = 0$, and the symbol $\sim$ indicates an order-of-magnitude estimate. From this result,

$$\frac{\partial u_x}{\partial x} \sim \frac{Q}{b^2 w}\frac{\mathrm{d}b}{\mathrm{d}x} \sim \frac{u_0}{b_0}\varepsilon \tag{2.160}$$

and from the equation of continuity,

$$\frac{\partial u_y}{\partial y} = -\frac{\partial u_x}{\partial x} \sim \frac{u_0}{b_0}\varepsilon \tag{2.161}$$

or

$$u_y \sim u_0\varepsilon,\quad \frac{\partial f}{\partial x} \sim \frac{\varepsilon}{b_0} f,\quad \frac{\partial f}{\partial y} \sim \frac{1}{b_0} f \tag{2.162}$$

where f may be u_x or u_y, or higher derivatives of these variables.

Based upon these scalings, the magnitudes of the various terms in the x component of the equation of motion can now be compared:

$$\underbrace{\varrho\left(u_x\frac{\partial u_x}{\partial x}+u_y\frac{\partial u_x}{\partial y}\right)}_{O\left(\frac{\varrho u_0^2}{b_0}\varepsilon\right)}=-\frac{\partial P}{\partial x}+\underbrace{\eta\frac{\partial^2 u_x}{\partial x^2}}_{O\left(\frac{\eta u_0}{b_0^2}\varepsilon^2\right)}+\underbrace{\eta\frac{\partial^2 u_x}{\partial y^2}}_{O\left(\frac{\eta u_0}{b_0^2}\right)} \tag{2.163}$$

The inertial term will be negligible compared to the y derivative viscous term if

$$\frac{\varrho u_0^2}{b_0}\varepsilon \ll \frac{\eta u_0}{b_0^2} \tag{2.164}$$

$$\frac{\varrho u_0 b_0}{\eta}\varepsilon \ll 1 \tag{2.164a}$$

Similarly, the x-derivative viscous term will be negligible compared to the y-derivative viscous term if

$$\frac{\eta u_0}{b_0^2}\varepsilon^2 \ll \frac{\eta u_0}{b_0^2} \tag{2.165}$$

or

$$\varepsilon^2 \ll 1 \tag{2.165a}$$

Equation (2.163) then reduces to

$$0 = -\frac{\partial P}{\partial x}+\eta\frac{\partial^2 u_x}{\partial y^2} \tag{2.166}$$

In a similar way, the y-component of the equation of motion can be simplified to

$$0 = -\frac{\partial P}{\partial y}+\eta\frac{\partial^2 u_y}{\partial y^2} \tag{2.167}$$

From these equations,

$$\frac{\partial P}{\partial y} \sim \frac{\eta u_0}{b_0^2}\varepsilon \sim \frac{\partial P}{\partial x}\varepsilon \tag{2.168}$$

which indicates that when $\varepsilon \ll 1$, pressure changes in the y-direction are negligible compared to those in the x-direction. Hence, changes of P in the y-direction can be neglected and the assumption made that $P = P(x)$.

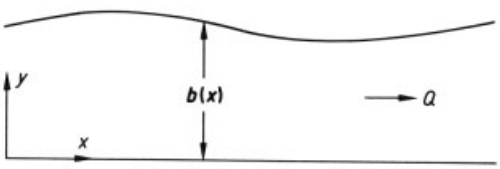

Figure 12. Flow in a two-dimensional channel with varying cross section
Flow can arise from an applied pressure gradient or from moving wall(s).

The solution of Equation (2.166) with the boundary conditions

$$\text{at } y = 0, \quad u_x = U \tag{2.169}$$

$$\text{at } y = h(x), \quad u_x = 0 \tag{2.170}$$

is

$$u_x = U\left(1-\frac{y}{b}\right)-\frac{b^2}{2\eta}\left(\frac{\mathrm{d}P}{\mathrm{d}x}\right)\left[\frac{y}{b}-\left(\frac{y}{b}\right)^2\right] \tag{2.171}$$

with the volumetric flow rate being

$$Q = \int_0^b u_x w \mathrm{d}y = \frac{1}{2}Ubw-\frac{b^3 w}{12\eta}\left(\frac{\mathrm{d}P}{\mathrm{d}x}\right) \tag{2.172}$$

If this equation is solved for ($\mathrm{d}P/\mathrm{d}x$) and then integrated, the pressure drop between $x = 0$ and $x = L$ is

$$\Delta P \equiv \frac{12 A_3 \eta Q L}{w h_0^3}-\frac{6 A_2 \eta U L}{h_0^2} \tag{2.173}$$

Here A_2 and A_3 are numerical parameters that depend on the geometry of the flow channel [$h(x)$]; they are given by

$$A_n = \int_0^1 \frac{\mathrm{d}x^*}{h^{*n}}, \quad n = 2, 3 \tag{2.174}$$

where $x* \equiv x/L$ and $b* \equiv b/b_0$. Equation (2.173) is the pressure drop – flow rate relation for flow in a two-dimensional channel with a weakly varying cross section.

Similar relations can be obtained for axisymmetric flows where the cross section varies along the axis of flow [39]. This includes flows in sinusoidal tubes, and converging or diverging ducts. Lubrication theory is applicable in such problems as long as $\varepsilon \equiv \mathrm{d}R/\mathrm{d}z$ is sufficiently small, where $R(Z)$ is the local channel radius. Good engineering approximations are often possible even for values of ε as high as 0.2.

To conclude this section an application of lubrication theory is presented which involves the analysis of slipper block performance. Consider the slipper block shown in Figure 13, which could represent a wiping block or ring on a moving piston. In some applications, the surface moves and the block is stationary (see Fig. 13);

in others, the block moves and the surface is stationary. The thickness of the film gap between the block and the moving surface, as well as the drag force on the block, depends on the applied load B, the speed U, the viscosity of the film fluid η, the width w, the length L, and the variation of the gap thickness with x. Lubrication theory is used to develop the appropriate relations between these variables.

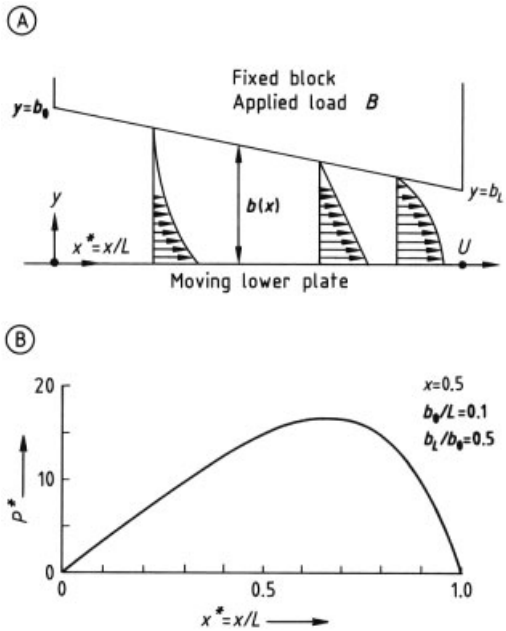

Figure 13. Lubrication flow between block and moving surface
A) Geometry; B) Pressure distribution

Begin by using Equation (2.172):

$$\frac{\mathrm{d}P}{\mathrm{d}x} = \frac{12\eta}{b^3}\left(\frac{Ub}{2} - \frac{Q}{w}\right) \tag{2.175}$$

In this problem, Q is not known and must be determined from known pressure conditions. In particular, at $x=0$ and $x=L$ the pressure is assumed to be zero. Integration of Equation (2.175) between these points then yields

$$\int_0^L \mathrm{d}P = 0 = \frac{12\eta L}{h_0^2}\left(\frac{1}{2}A_2 U - \frac{A_3 Q}{h_0 w}\right) \tag{2.176}$$

where the A_n are defined by Equation (2.174). The volumetric flow rate is then

$$Q = \frac{A_2 w b_0 U}{2A_3} \tag{2.177}$$

If the gap geometry is described by the linear relation

$$b = b_0 + (b_L - b_0)\frac{x}{L} \tag{2.178}$$

the dimensionless pressure distribution can be obtained from Equations (2.175) and (2.177) with the result

$$P^* \equiv \frac{(P - P_0)L}{6\eta U} = \frac{(b^* - 1)(b^* - \kappa)}{(b_0/L)^2(\kappa^2 - 1)} \tag{2.179}$$

where $\kappa \equiv b_L/b_0$. As depicted in Figure 13, the pressure is positive in the gap and exhibits a maximum toward the narrow gap end. The magnitude of the pressure increases with U and η, and decreases with (b_0/L). The pressure also causes a *vertical force* F_y on the *block*, which is obtained by integrating P over the block length. The result is

$$F_y^* \equiv \frac{F_y}{6\eta U w} = \frac{1}{(b_0/L)^2(\kappa - 1)^2}\left[\frac{2(\kappa - 1)}{\kappa + 1} - \ln\kappa\right] \tag{2.180}$$

For a given b_0/L, the maximum vertical force is obtained for a block geometry where $\kappa = 2.19$. Also, because the vertical pressure force must be equal to the load B, the resulting thickness $b*_0$ is

$$b_0^* \equiv \frac{b_0}{L} = \left\{\frac{6\eta U w}{B(\kappa - 1)^2}\left[\frac{2(\kappa - 1)}{\kappa + 1} - \ln\kappa\right]\right\}^{1/2} \tag{2.181}$$

Here, for a given κ, the gap thickness increases with increasing U and η, and decreases with increasing load B, and each case has a square-root dependence.

The drag force F_x on the *moving plate is* obtained by integrating τ_{yx} over this surface:

$$F_x^* \equiv \frac{F_x}{6\eta w U} = \frac{1}{(b_0/L)(\kappa - 1)}\left[\frac{(\kappa - 1)}{\kappa + 1} - \frac{2}{3}\ln\kappa\right] \tag{2.182}$$

Finally, if Equation (2.181) is used to eliminate b_0/L,

$$F_x = \left(\frac{2}{3}\eta U w B\right)^{1/2}\left\{\frac{3\left(\frac{\kappa - 1}{\kappa + 1}\right) - 2\ln\kappa}{2\left(\frac{\kappa - 1}{\kappa + 1}\right) - \ln\kappa}\right\} \tag{2.183}$$

which shows that the drag force is proportional to the square root of the applied load. This half-power dependence when a lubricating film is present is in contrast to the first-power dependence generally assumed between sliding solid surfaces with no lubricant. Also, the effective coefficient of friction for lubricated surfaces is orders of magnitude smaller than that associated with direct – contact solid surfaces (→ Lubricants and Related Products, Chap. 2.).

2.4. Coupled Momentum and Energy Transfer

This discussion has been concerned with examples that involve only heat or momentum transfer, but not both. In this section, examples are considered in which heat and momentum are transferred simultaneously and, specifically, heat is being transferred in a flowing fluid. When such heat transfer results directly from fluid movement, the transfer is termed *heat convection.* If the flow arises from applied pressure gradients or moving boundaries, the heat transfer is called *forced convection*. When the flow arises from buoyancy effects associated with density variations in nonisothermal fluids, it is called *free or natural convection.*

In analyzing heat transfer with flow, the energy balance is expressed in the form [16, pp. 215 – 216]:

$$\varrho c_p \left(\frac{\partial T}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla} T \right) = \boldsymbol{\nabla} \cdot (\lambda \boldsymbol{\nabla} T) + T \beta \left(\frac{\partial p}{\partial t} + \boldsymbol{u} \cdot \boldsymbol{\nabla} p \right) + \mathrm{tr}\,(\boldsymbol{\tau} \cdot \boldsymbol{\nabla} \boldsymbol{u}) + \Phi \qquad (2.184)$$

which is equivalent to Equation (1.85) with Fourier's law (Eq. 1.91) used for the conductive heat flux $\boldsymbol{q}$; β is the coefficient of thermal expansion defined by

$$\beta \equiv -\frac{1}{\varrho} \left(\frac{\partial \varrho}{\partial T} \right)_p \qquad (2.185)$$

In Equation (2.184) the second term on the left, which involves velocity, accounts for convective heat transfer, and the terms on the right are the conductive, reversible work (fluid expansion – contraction), viscous dissipation, and thermal generation terms, respectively.

Consider now the case of steady flow of a Newtonian fluid in a heated tube where the temperature of the fluid varies in the radial and axial directions because of heat exchange with the walls, viscous heat generation, and reversible work effects. If free convection, generation, and radiation effects are neglected, and the velocity field is assumed to be $\boldsymbol{u} = [0, 0, u_z\,(r)]$, Equation (2.184) takes the form

$$\varrho c_p u_z \frac{\partial T}{\partial z} = \boldsymbol{\nabla} \cdot (\lambda \boldsymbol{\nabla} T) + T \beta u_z \frac{\mathrm{d}p}{\mathrm{d}z} + 2\eta \left(\frac{\mathrm{d}u_z}{\mathrm{d}r} \right)^2 \qquad (2.186)$$

The contribution of *viscous dissipation* to the temperature gradient in the z-direction can be estimated from

$$\left(\frac{\partial T}{\partial z} \right)_{\text{viscous dissipation}} = \frac{2\eta\,(\mathrm{d}u_z/\mathrm{d}r)^2}{\varrho c_p u_z} \sim O\left(Re_{\mathrm{d}} \frac{\nu^2}{c_p d^3} \right) \qquad (2.187)$$

where $Re_{\mathrm{d}} \equiv (u_{\mathrm{avg}}\, d/\nu)$ is the Reynolds number; $\nu \equiv \eta/\varrho$, the momentum diffusivity, d the tube diameter; and u_{avg} the average velocity in the tube. For air at room temperature ($\nu \approx 2 \times 10^{-5}\,\mathrm{m^2/s}$, $c_p \approx 10^3\,\mathrm{J\,kg^{-1}\,K^{-1}}$) flowing in a 0.025-m tube at a Reynolds number $Re_{\mathrm{d}} = 10^3$, the temperature gradient due to viscous effects is on the order of 3×10^{-5} K/m. For water under the same conditions ($\nu \approx 8 \times 10^{-7}\,\mathrm{m^2/s}$, $c_p \approx 4 \times 10^3\,\mathrm{J\,kg^{-1}\,K^{-1}}$), the gradient is on the order of 1×10^{-8} K/m; for an oil with properties $\nu \approx 9 \times 10^{-4}\,\mathrm{m^2/s}$, $c_p \sim 2 \times 10^3\,\mathrm{J\,kg^{-1}\,K^{-1}}$, the gradient is ca. 3×10^{-2} K/m. In most heat-exchange applications, such effects are negligible compared to the conduction contributions (first term on right-hand side, Eq. 2.186) which can be on the order of several kelvin per meter. Therefore, viscous dissipation effects are generally neglected in most heat-exchange applications. However, viscous dissipation can be important in lubrication problems where viscous oils are being sheared at high rates between moving surfaces.

If a similar order-of-magnitude analysis is carried out for the contribution of the *reversible work term* to the axial temperature gradient:

$$\left(\frac{\partial T}{\partial z}\right)_{\text{rev. work}} = \frac{T\,\beta\,(\mathrm{d}p/\mathrm{d}z)}{\varrho\,c_p} \sim O\left(32\,Re_{\mathrm{d}}\,\frac{T\,\beta\,\nu^2}{c_p\,d^3}\right) \tag{2.188}$$

Here, the pressure gradient is related to velocity through the Hagen – Poiseuille equation (Eq. 2.86). Comparing Equations (2.187) and (2.188) gives

$$\left(\frac{\partial T}{\partial z}\right)_{\text{rev. work}} \sim \pm 32\,T\,\beta\left(\frac{\partial T}{\partial z}\right)_{\text{viscous dissipation}} \tag{2.189}$$

where the sign depends on whether $\mathrm{d}p/\mathrm{d}z$ in Equation (2.188) is positive or negative. When $\mathrm{d}p/\mathrm{d}z < 0$, the reversible work results in a decrease in temperature in the direction of flow; in contrast, the viscous dissipation causes an increase in temperature. For air at room temperature where $\beta = -1/T$ (ideal gas), the temperature gradient contribution associated with reversible work effects is on the order of 30 times that associated with viscous dissipation effects. For liquids where $\beta \sim 10^{-3}\,\mathrm{K}^{-1}$, the reversible work contribution to the temperature gradient is on the order of 10 times that arising from viscous dissipation. Still, in most applications, these effects are small compared to conduction or convection effects and can be neglected.

The other effect that is not commonly considered in convection analyses is the *temperature dependence of thermal conductivity.* For most fluids, the magnitude of $(1/\lambda)\,(\partial\lambda/\partial T)$ is on the order of $10^{-3}\,\mathrm{K}^{-1}$ and

$$\boldsymbol{\nabla}\cdot(\lambda\boldsymbol{\nabla}T) = \lambda\left[\nabla^2 T + \frac{1}{\lambda}\left(\frac{\partial\lambda}{\partial T}\right)|\boldsymbol{\nabla}T|^2\right] \sim \lambda\nabla^2 T \tag{2.190}$$

Hence, if viscous dissipation and reversible work effects are neglected, the heat generation term is zero, and the conductivity is assumed to be temperature-independent, the energy balance (Eq. 2.184) reduces to

$$\frac{\partial T}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}T = a\nabla^2 T \tag{2.191}$$

where $\alpha \equiv \lambda/\varrho\,c_p$ is the thermal diffusivity. This is the equation used in the forced convection examples considered here.

2.4.1. Heat Transfer in Laminar Tube Flow

Consider the heat exchange that occurs when a fluid is passed through a heated or cooled tube under steady laminar flow conditions. If a Newtonian fluid, constant physical properties, negligible axial conduction, and fully developed velocity and temperature profiles are assumed, the energy balance can be written as

$$u_z\frac{\partial T}{\partial z} = \alpha\,\frac{1}{r}\frac{\partial}{\partial r}\left(r\,\frac{\partial T}{\partial r}\right) \tag{2.192}$$

where

$$u_z = \frac{1}{2}\,u_{\mathrm{avg}}\left[1 - \left(\frac{r}{r_1}\right)^2\right] \tag{2.193}$$

with r_1 being the tube radius and u_{avg} the average velocity. Depending on the boundary conditions, various solutions are possible. When the heat flux at the wall q_{w} is constant, a particularly simple solution is possible. In particular, if Equation (2.192) is multiplied by r and integrated with respect to r, then

$$\frac{\mathrm{d}T_b}{\mathrm{d}z} = \frac{2\,\alpha\left(\frac{\partial T}{\partial r}\right)_{r_1}}{r_1\,u_{\mathrm{avg}}} = \frac{2\,\alpha\,q_{\mathrm{w}}}{r_1\,u_{\mathrm{avg}}\,\lambda} \tag{2.194}$$

where T_{b} is the *bulk average temperature* defined by

$$T_b = \frac{1}{\pi\,r_1^2\,u_{\mathrm{avg}}}\int_0^{r_1} u_z\,T\,2\,\pi\,r\mathrm{d}r \tag{2.195}$$

The bulk average temperature is simply the "cup-mixing temperature" that would be measured if the tube were chopped off at z and the fluid issuing forth were collected and thoroughly mixed in a container.

Equation (2.194) implies that T_{b} varies linearly with z. The achievement of a fully developed temperature profile requires that

$$\frac{\partial}{\partial z}\left(\frac{T_{\mathrm{w}} - T}{T_{\mathrm{w}} - T_{\mathrm{b}}}\right) = 0 \tag{2.196}$$

or

$$\frac{T_{\mathrm{w}} - T}{T_{\mathrm{w}} - T_{\mathrm{b}}} = f\,(r) \tag{2.197}$$

From Newton's law of cooling,

$$h = -\frac{\lambda\left(\frac{\partial T}{\partial r}\right)_{r_1}}{(T_\mathrm{w} - T_\mathrm{b})} = \lambda\left[\frac{\partial}{\partial r}\left(\frac{T_\mathrm{w} - T}{T_\mathrm{w} - T_\mathrm{b}}\right)\right]_{r_1}$$
$$= \lambda f\,(r_1) = \text{const.} \tag{2.198}$$

or

$$T_\mathrm{w} - T_\mathrm{b} = \frac{q_\mathrm{w}}{h} = \text{const.} \tag{2.199}$$

By using this equation with Equations (2.194) and (2.197),

$$\frac{\mathrm{d}T_\mathrm{w}}{\mathrm{d}z} = \frac{\mathrm{d}T_\mathrm{b}}{\mathrm{d}z} = \frac{\partial T}{\partial z} = \frac{\mathrm{d}T}{\mathrm{d}z} = \frac{2\,q_\mathrm{w}\alpha}{r_1\,u_\mathrm{avg}\lambda} \tag{2.200}$$

Substituting this result into Equation (2.192) and integrating over r give

$$T_\mathrm{w} - T_\mathrm{b} = \frac{11}{24}\frac{q_\mathrm{w} r_1}{\lambda} = \frac{11}{48}\frac{q_\mathrm{w} d}{\lambda} \tag{2.201}$$

The latter result indicates that when the fluid is heated (or cooled) under *constant wall heat flux* conditions, the temperature difference between the wall and the bulk fluid is constant.

If Equations (2.201) and (2.199) are combined,

$$Nu_d \equiv \frac{h\,d}{\lambda} = 4.364 \quad (q_\mathrm{w} = \text{const., large}\,z) \tag{2.202}$$

where Nu_d is the dimensionless Nusselt number based upon the tube diameter d. This result provides a simple expression for estimating h under constant wall heat flux conditions and in zones (large z) where the profiles are fully developed. For the case of *constant wall temperature,* the solution is somewhat more involved [38]; however, Nu_d is still constant but equal to a different value

$$Nu_d = 3.658 \quad (T_\mathrm{w} = \text{const., large}\,z) \tag{2.203}$$

Use of this result is limited to z values greater than the thermal entry length z_e which is the position from the tube entrance where Equation (2.196) is approximately satisfied, say $\partial(T_\mathrm{w} - T)/\partial z \sim 0.01$. For laminar flow, this entry length can be estimated from

$$\frac{z_\mathrm{e}}{d} \approx 0.05\,Re_d\,Pr \tag{2.204}$$

In particular, for water at room temperature ($Pr \equiv \nu/\alpha \sim 10$) flowing at a Reynolds number of 10^3, the entry length would be 500 diameters; for oils with $Pr \sim 10^4$, the entry lengths are very long and fully developed thermal profiles are seldom achieved. In the latter cases, more complete solutions must be used to describe heat transfer in the developing regions, and these are considerably more complicated than the fully developed solutions presented above [38–42]. The classical Graetz solution [38], which involves the solution of Equation (2.192) for parabolic and flat velocity profiles and constant wall temperatures are the best known among these solutions.

A useful empirical expression is given by SIEDER and TATE [43]; it is an empirical modification of the Graetz solution, (parabolic velocity) i.e.,

$$Nu_{d,\mathrm{lm}} \equiv \frac{h_\mathrm{lm} d}{\lambda} =$$
$$1.86\,Re_d^{1/3}\,Pr^{1/3}\,(L/d)^{-1/3}\,(\eta_\mathrm{bm}/\eta_\mathrm{w})^{0.14} \tag{2.205}$$

Here, the physical properties of the fluid are evaluated at the mean bulk fluid temperature, $T_\mathrm{bm} \equiv (T_\mathrm{b0} + T_\mathrm{b}L)/2$, where T_b0 and $T_\mathrm{b}L$ are the entering and exit bulk temperatures for a tube section of length L. The viscosity ratio $(\eta_\mathrm{bm}/\eta_\mathrm{w})$ is the ratio of the viscosities evaluated at the mean bulk temperature and at the mean wall temperature; the quantity h_lm is the log-mean film heat-transfer coefficient defined by

$$h_\mathrm{lm} = \frac{\dot{Q}}{\lambda L\,d\Delta T_\mathrm{lm}} = -\frac{\int_0^L (q_r)_{r_1}\,\mathrm{d}z}{L\Delta T_\mathrm{lm}} \tag{2.206}$$

where

$$\Delta T_\mathrm{lm} \equiv \left[\frac{(\mathrm{T}_\mathrm{w}L - T_\mathrm{b}L) - (T_\mathrm{w0} - T_\mathrm{b0})}{\ln\left(\frac{T_\mathrm{w}L - T_\mathrm{b}L}{T_\mathrm{w0} - T_\mathrm{b0}}\right)}\right] \tag{2.207}$$

is the log-mean temperature difference. The heat-transfer rate $\dot{Q}$ across the wall can be expressed in terms of the difference between the convected energy input and output to the heated (or cooled) section by

$$\dot{Q} = \frac{\pi d^2}{4}\,u_\mathrm{avg}\left[(\varrho c_p T)_L - (\varrho c_p T)_0\right] \tag{2.208}$$

For large values of L/d, Equation (2.205) predicts that the log-mean Nusselt number goes to zero, which is not consistent with results for the fully developed regions. WHITAKER [16,

pp. 324 – 325] suggests the use of the following expression, which is based on the work of HAUSEN [42]:

$$Nu_{d,\mathrm{lm}} = 3.66 + \frac{0.0745\,\psi^3}{1 + 0.04\,\psi^2} \tag{2.209}$$

where

$$\psi \equiv (Re\,Pr)^{1/3}\,(L/d)^{-1/3}\,(\eta_{\mathrm{bm}}/\eta_{\mathrm{w}})^{0.14} \tag{2.210}$$

This result is accurate to ca. $\pm 10\,\%$ for situations in which the wall temperature is constant or nearly constant. In the case of constant wall heat flux, variations of T_{w} can be large; however, the local Nusselt numbers are only ca. 20 % higher than those for constant-temperature cases. As a result, Equation (2.209) is used even for the case of constant heat flux and surface temperature variations are ignored.

When the entering fluid temperature and the wall temperature are known, the exit temperature T_L and the overall heat transfer rate $\dot{Q}$ can be calculated by using Equations (2.206) – (2.210). Since the evaluation of the properties requires a knowledge of T_L, an iterative procedure is required in which T_L is assumed and h_{lm} is calculated from Equation (2.209). Then this value of the heat-transfer coefficient is used with Equations (2.207) and (2.208) to determine T_L and $\dot{Q}$. The new value of T_L is then employed to update the value of h_{lm}, and the procedure is repeated until convergence is obtained.

2.4.2. Momentum and Heat Transfer in Laminar Boundary Layers

A characteristic of real fluids moving around solid bodies is the adherence of the fluid to the solid surface. This no-slip condition causes fluid-deforming velocity gradients that tend to be large in the neighborhood of the body and to diminish far away from it. If the Reynolds number based on the characteristic length of the body is sufficiently high, the resulting velocity gradient is restricted largely to a thin region in the neighborhood of the boundary. Under such conditions, a boundary layer approach is possible where the flow field is divided into two regions: (1) a thin boundary layer in which both viscous and inertial effects are important, and (2) an external region in which viscous effects (and fluid-deforming velocity gradients) are negligible. When applicable, such an approach simplifies the solution considerably, and approximate analytical solutions are possible in many cases.

Momentum Transfer in a Laminar Boundary Layer. To illustrate this approach, consider first the momentum transfer associated with flow past the surface shown in Figure 14. In this case, the flow is two dimensional where $\boldsymbol{u} = [u_x\,(x, y), u_y\,(x, y), 0]$, with the x-coordinate measured along the surface and the y-coordinate measured normal to the surface. As the fluid moves over the surface with velocity $\boldsymbol{u}$, a boundary layer is formed that encompasses most of the velocity changes resulting from interactions of the fluid with the surface. Although many different ways of defining the boundary layer thickness exist [44], in this article thickness is defined as the y-position at which the velocity is 99 % of the mainstream velocity u_∞, i.e.,

$$\delta_{\mathrm{m}}\,(x) \equiv \{\text{The } y\text{−position at each } x \text{ where } u_x = 0.99\,u_\infty\} \tag{2.211}$$

The subscript m indicates the boundary layer thickness for momentum transfer; boundary layer thicknesses for heat- and mass-transfer processes are defined later and denoted by δ_T and δ_c, respectively.

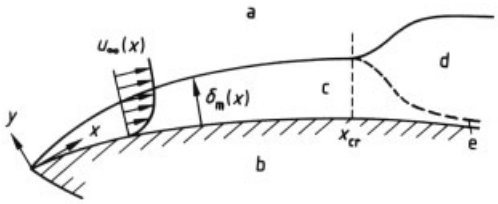

Figure 14. Momentum boundary layer on a surface
a) Fluid; b) Solid body; c) Laminar boundary layer; d) Turbulent boundary layer; e) Laminar sublayer

As illustrated in Figure 14, the boundary layer thickness increases with distance x along the surface. For a certain distance $x < x_{\mathrm{cr}}$, flow in the boundary layer is laminar. At the position x_{cr}, which can be estimated from

$$Re_{\mathrm{cr}} \equiv \frac{u_\infty\,x_{\mathrm{cr}}}{\nu} \approx 2 \times 10^5 \tag{2.212}$$

inertial effects become sufficiently large compared to viscous effects to destabilize the flow and except for a very thin laminar sublayer next

to the surface of the body, the flow exhibits sporadic vortex-like instabilities that eventually grow into fully developed turbulence. Physically, this is observed in flow visualization when the boundary layer thickness increases rather abruptly as shown in Figure 14 (in graphically depicting a boundary layer as in Fig. 14, the thickness is exaggerated and not to scale). Generally, x_{cr} may be on the order of meters and δ_{m} on the order of millimeters.

The velocity field in the laminar portion of the boundary layer, and the associated drag force on the surface are now determined. If the radius of curvature of the surface is sufficiently large compared to the boundary layer thickness, the flow can be represented as that along a flat plate as shown in Figure 15. If the scalings

$$u_x \sim u_\infty \tag{2.213}$$

$$\frac{\partial u_x}{\partial x} \sim \frac{u_\infty}{L} \tag{2.214}$$

are used, the equation of continuity

$$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} = 0 \tag{2.215}$$

yields

$$\frac{\partial u_y}{\partial y} \sim \frac{u_\infty}{L} \tag{2.216}$$

$$u_y \sim u_\infty \frac{\delta_{\mathrm{m}}}{L} \tag{2.217}$$

By using these scalings, the x- and y-components of the equation of motion can be simplified. In particular, if $\delta_{\mathrm{m}}/L \ll 1$, the x component takes the form [44, p. 131]

$$u_x \frac{\partial u_x}{\partial x} + u_y \frac{\partial u_x}{\partial y} = -\frac{1}{\varrho}\frac{\mathrm{d}p}{\mathrm{d}x} + \nu \frac{\partial^2 u_x}{\partial y^2} + g_x \tag{2.218}$$

The y-component simply reduces to $\partial p/\partial y = 0$, which indicates that the pressure is independent of y and hence is simply the pressure $p(x)$ outside the boundary layer. If the external flow is assumed to be nonviscous the pressure field can be determined from:

$$p = -\frac{1}{2}\varrho u_\infty^2 - \varrho g_x x + \text{constant} \tag{2.219}$$

using this equation in Equation (2.218) gives

$$u_x \frac{\partial u_x}{\partial x} + u_y \frac{\partial u_x}{\partial y} = u_\infty \frac{\mathrm{d}u_\infty}{\mathrm{d}x} + \nu \frac{\partial^2 u_x}{\partial y^2} \tag{2.220}$$

In principle, if u_∞ is known (obtained from the solution of the nonviscous equations for the flow geometry [45]), Equations (2.215) and (2.220) can be solved for the velocity components u_x and u_y. Although these equations can be solved numerically [44], [46], here the focus is on the use of approximate integral methods.

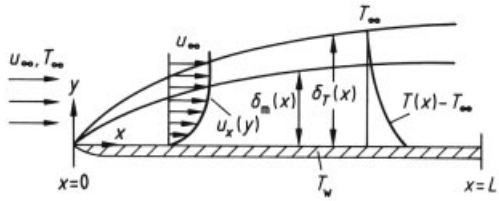

Figure 15. Laminar momentum and thermal boundary layers on a flat plate

In particular, note that

$$u_y \frac{\partial u_x}{\partial y} = \frac{\partial}{\partial y}(u_y u_x) - \frac{\partial u_y}{\partial y} u_x \tag{2.221}$$

If $(\partial u_y/\partial y)$ is replaced with $-(\partial u_x/\partial x)$ by using the equation of continuity, Equation (2.220) can then be written as

$$\frac{\partial}{\partial x}(u_x^2) + \frac{\partial}{\partial y}(u_y u_x) = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}x}(u_\infty^2) + \nu \frac{\partial^2 u_x}{\partial y^2} \tag{2.222}$$

When this equation is integrated over y from $y = 0$ to $y = \infty$, and the conditions

$$u_x = 0, \quad u_y = 0 \quad \text{at } y = 0 \tag{2.223}$$

$$u_x = u_\infty, \quad \frac{\partial u_x}{\partial y} = 0 \quad \text{at } y \geq \delta_{\mathrm{m}} \tag{2.224}$$

along with (from Eq. 2.215)

$$u_y(x, \infty) = -\int_0^{\delta_{\mathrm{m}}} \frac{\partial u_x}{\partial x} \mathrm{d}y \tag{2.225}$$

are used, the result is

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{\delta_{\mathrm{m}}} u_x(u_\infty - u_x)\,\mathrm{d}y + \frac{\mathrm{d}u_\infty}{\mathrm{d}x}\int_0^{\delta_{\mathrm{m}}} (u_\infty - u_x)\,\mathrm{d}y = \frac{\tau_{\mathrm{w}}}{\varrho} \tag{2.226}$$

where τ_{w} is the wall shear stress given by

$$\tau_{\mathrm{w}} = \eta \left(\frac{\partial u_x}{\partial y} \right)_{y=0} \tag{2.227}$$

Equation (2.226) is the *von Karman momentum integral equation.* For a given external flow $u_\infty(x)$, its solution requires a knowledge of the velocity u_x as a function of y. Even with crude approximations for $u_x(y)$, reasonably accurate estimates of the drag force can be obtained.

As an example, consider the case of a flat plate in a free stream flow where u_∞ = constant. If a cubic relation is assumed for $u_x(y)$, i.e.,

$$u_x = a + by + cy^2 + dy^3 \tag{2.228}$$

and the boundary conditions of Equations (2.223) and (2.224) are used, along with $(\partial^2 u_x/\partial y^2) = 0$ at $y = \delta_{\mathrm{m}}$ to evaluate the constants, then

$$u_x = u_\infty \left[\frac{3}{2} \frac{y}{\delta_{\mathrm{m}}} - \frac{1}{2} \left(\frac{y}{\delta_{\mathrm{m}}} \right)^3 \right] \tag{2.229}$$

If this assumed profile is substituted in the momentum integral balance (Eq. 2.226) and the necessary integrations are performed, the following differential equation for boundary layer thickness is obtained:

$$\frac{\mathrm{d}\delta_{\mathrm{m}}}{\mathrm{d}x} = \frac{140\,\nu}{13\,u_\infty} \frac{1}{\delta_{\mathrm{m}}} \tag{2.230}$$

Integration with the condition $\delta_{\mathrm{m}} = 0$ at $y = 0$ gives the boundary layer thickness

$$\frac{\delta_{\mathrm{m}}}{x} = \frac{4.64}{Re_x^{1/2}} \tag{2.231}$$

where $Re_x \equiv u_\infty\, x/\nu$ is the local Reynolds number. The corresponding drag coefficient is given by

$$C_{\mathrm{D}x} \equiv \frac{\tau_{\mathrm{w}}}{\frac{1}{2}\varrho u_\infty^2} = \frac{2}{\varrho u_\infty^2} \frac{3\eta u_\infty}{2\delta_{\mathrm{m}}} = \frac{0.646}{Re_x^{1/2}} \tag{2.232}$$

where the subscript x on C_{D} indicates the coefficient is a local value that varies with x. The drag force per unit width is then

$$\frac{F_{\mathrm{D}}}{w} = \int_0^L \tau_{\mathrm{w}}\,\mathrm{d}x = C_{\mathrm{D}L}\, L \left(\frac{1}{2} \varrho u_\infty^2 \right) \tag{2.233}$$

where the overall film coefficient $C_{\mathrm{D}L}$ is given by

$$C_{\mathrm{D}L} \equiv \frac{1}{L} \int_0^L C_{\mathrm{D}x}\,\mathrm{d}x = \frac{1.292}{Re_L^{1/2}} \tag{2.234}$$

and $Re_L \equiv u_\infty\, L/\nu$.

The *exact solution* for a laminar boundary layer on a flat plate with a uniform free stream velocity was obtained by Blasius [44]. The corresponding local and overall drag coefficient results are

$$C_{\mathrm{D}x} = \frac{0.664}{Re_x^{1/2}} \tag{2.235}$$

$$C_{\mathrm{D}L} = \frac{1.328}{Re_L^{1/2}} \tag{2.236}$$

The approximate results of Equations (2.232) and (2.234) compare favorably with the exact results. This demonstrates the utility of the integral method. Even more accurate results could have been obtained if higher order approximations (beyond the cubic) were used for the velocity profile. In general, such integral approaches are valuable tools in engineering analysis and provide excellent approximations with comparable ease relative to the numerical methods required to obtain exact results.

Heat Transfer in a Laminar Boundary Layer. Just as the momentum boundary layer thickness was defined as the region in which most of the fluid-deforming velocity gradients occur, a thermal boundary layer thickness δ_T can be defined as that region next to a heated (or cooled) surface in which most of the temperature gradients occur. Specifically, for heat transfer from a flat plate at temperature T_{w} to a flowing fluid with free stream temperature T_∞, the thermal boundary layer is defined as

$$\delta_t \equiv \{\text{The } y\text{-position where } T - T_{\mathrm{w}} = 0.99\,(T_\infty - T_{\mathrm{w}})\} \tag{2.237}$$

In general, as shown in Figure 15, the thermal boundary layer thickness is not equal to the momentum boundary layer thickness. However, the relation between the thermal and momentum boundary layers in laminar flow can be approximated by [47]

$$\frac{\delta_{\mathrm{m}}}{\delta_T} \equiv \Delta_T = \text{const.} \tag{2.238}$$

Equation (2.238) implies that the dependence of δ_T on x is the same as that of δ_m on x.

For heat transfer in a laminar boundary layer on a flat plate where $\delta_T/L \ll 1$, the energy balance (Eq. 2.191) can be simplified to

$$u_x \frac{\partial T}{\partial x} + u_y \frac{\partial T}{\partial y} = \alpha \frac{\partial^2 T}{\partial y^2} \tag{2.239}$$

By integrating over y from 0 to ∞ and using the equation of continuity to replace u_y (the steps involved are analogous to Eqs. 2.221 – 2.226), the energy balance can be put into the integral form

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_0^{\delta_T} u_x (T - T_\infty)\, \mathrm{d}y = \frac{q_w}{\varrho c_p} \tag{2.240}$$

where q_w is the wall heat flux given by

$$q_w = -\lambda \frac{\partial T}{\partial y}\Big|_{y=0} \tag{2.241}$$

Equation (2.240) is analogous to the momentum integral equation (Eq. 2.226). To obtain a solution to this integral energy balance, a temperature profile must be assumed that satisfies the conditions

$$\text{at} \quad y = 0, \quad T = T_w \tag{2.242}$$

$$\text{at} \quad y = \delta_T, \quad T = T_\infty \tag{2.243}$$

$$\text{at} \quad y = \delta_T, \quad \frac{\partial^n T}{\partial y^n} = 0 \quad \text{for} \quad n = 1, 2, \ldots \tag{2.244}$$

If a cubic power relation is assumed for the temperature profile

$$T = a + by + cy^2 + dy^3 \tag{2.245}$$

the boundary conditions of Equations (2.242) – (2.244) can be used to determine the coefficients, with the result

$$T = T_w + (T_\infty - T_w) \left[\frac{3}{2} \left(\frac{y}{\delta_T} \right) - \frac{1}{2} \left(\frac{y}{\delta_T} \right)^3 \right] \tag{2.246}$$

When this profile is substituted into the integral form of the energy balance, Equation (2.229) is used for u_x, and Equation (2.238) is used for δ_m, then

$$\delta_T \frac{\mathrm{d}\delta_T}{\mathrm{d}x} = \frac{10\,\alpha}{u_\infty} \left(\frac{14\,\Delta_T^3}{14\,\Delta_T^2 - 1} \right) \tag{2.247}$$

where $\Delta_T \equiv \delta_m/\delta_T$ has been assumed to be constant. By integrating this equation with the condition $\delta_T = 0$ at $y = 0$, the thermal boundary layer thickness is found to be

$$\frac{\delta_T}{x} = \frac{4.47\,\Delta_T^{3/2}}{Re_x^{1/2}\, Pr^{1/2}} \left(\frac{14}{14\,\Delta_T^2 - 1} \right)^{1/2} \tag{2.248}$$

where $Pr = \nu/\alpha$ is the Prandtl number, the ratio of the momentum and thermal diffusivities. Then, from Equations (2.231), (2.238), and (2.248),

$$\Delta_T \equiv \frac{\delta_m}{\delta_T} \left[\frac{14}{13} \left(1 - \frac{1}{14\,\Delta_T^2} \right) \right]^{1/3} Pr^{1/3}$$

$$\approx Pr^{1/3} \tag{2.249}$$

and

$$\frac{\delta_T}{x} = 4.47\, Re_x^{-1/2}\, Pr^{1/3} \tag{2.250}$$

where it has been assumed

$$14\,\Delta_T^2 \gg 1$$

Based upon these results, the momentum boundary layer thickness increases relative to the thermal boundary layer as the momentum diffusivity ν increases relative to the thermal diffusivity a (or Pr increases). For gases, the Prandtl number is on the order of 1 and the boundary layer thickness δ_m and δ_T have similar magnitudes. For liquids, the Prandtl number can be much higher than 1 and δ_m can be larger than δ_T. For liquid metals, $Pr \ll 1$ and the momentum boundary layer thickness is much smaller than the thermal boundary layer thickness.

Thus, from Equations (2.241), (2.246), and Newton's law of cooling,

$$Nu_x \equiv \frac{h_x \chi}{\lambda} = 0.36\, Re_x^{1/2}\, Pr^{1/3} \tag{2.251}$$

where Nu_x is the dimensionless *local Nusselt number*. This result is ca. 8 % higher than the exact result from numerical solutions, which is given by

$$Nu_x = 0.332\, Re_x^{1/2}\, Pr^{1/3} \tag{2.252}$$

Heat transfer from the plate can then be determined from

$$\dot{Q} = \int_0^L h\,(T_\mathrm{w} - T_\infty)\,w\mathrm{d}x = h_L\,w\,L\,(T_\mathrm{w} - T_\infty) \tag{2.253}$$

where

$$h_L \equiv \frac{1}{L}\int_0^L h\mathrm{d}x \tag{2.254}$$

Using Equation (2.252) yields

$$Nu_L \equiv \frac{h_L\,L}{\lambda} = 0.664\,Re_L^{1/2}\,Pr^{1/3} \tag{2.255}$$

where $Re_L \equiv (u_\infty\,L/\nu)$. Hence, if the properties are determined from the average film temperature, $1/2\,(T_\mathrm{w}+T_\infty)$, the total heat-transfer rate from the plate can be determined from Equation (2.253), with h_L estimated from Equation (2.255).

For the case of *constant heat flux* at the plate surface, the local Nusselt number is given by

$$Nu_x = 0.453\,Re_x^{1/2}\,Pr^{1/3} \tag{2.256}$$

In this case, wall temperature varies with position x on the plate according to

$$T_\mathrm{w} = T_\infty + \frac{q_\mathrm{w}}{h_x} \tag{2.257}$$

Although the above relations have been derived for flat surfaces, they are also useful as approximations for curved surfaces if the characteristic radius of curvature for the surface is much larger than the boundary layer thickness.

2.4.3. Free Convection on a Vertical Plate

When a stagnant fluid is in contact with a heated surface, the fluid layers near the surface are at higher temperatures than the fluid layers further removed from the surface. Because of their higher temperatures, the layers near the surface have lower densities and will rise relative to the stagnant bulk fluid. Such free convection processes can be analyzed in much the same way as forced convection on a flat plate was analyzed by using boundary layer methods.

The velocity and temperature field inside the momentum and thermal boundary layers are illustrated in Figure 16. The boundary layer equations, Equations (2.218) and (2.239), describe the behavior of these variables. Outside the boundary layer, the pressure gradient is simply the hydrostatic gradient

$$\frac{\mathrm{d}p}{\mathrm{d}x} = -\varrho_\infty\,g \tag{2.258}$$

where ϱ_∞ is the density of the fluid at T_∞. Inside the thermal boundary layer, the density can be related to the temperature by

$$\varrho = \varrho_\infty + \left(\frac{\partial\varrho}{\partial T}\right)_{T_\infty}(T - T_\infty)$$
$$= \varrho_\infty\,[1 - \beta_\infty\,(T - T_\infty)] \tag{2.259}$$

Substituting Equations (2.258) and (2.259) into Equation (2.218) gives the momentum equation for free convection, i.e.,

$$u_x\frac{\partial u_x}{\partial x} + u_y\frac{\partial u_x}{\partial y} = \nu\frac{\partial^2 u_x}{\partial y^2} + g\beta_\infty\,(T - T_\infty) \tag{2.260}$$

In integral form, this equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{\delta_\mathrm{m}} u_x^2\mathrm{d}y = -\nu\frac{\partial u_x}{\partial y}\Big|_{y=0} +$$
$$g\beta_\infty\int_0^{\delta_T}(T - T_\infty)\,\mathrm{d}y \tag{2.261}$$

The integral form of the energy balance is the same as Equation (2.240).

In selecting appropriate velocity and temperature expressions to use in these equations, the following boundary conditions are applicable:

$$\text{at}\quad y = 0,\quad u_x = 0 \tag{2.262}$$

$$T = T_\mathrm{w} \tag{2.263}$$

$$\text{at}\quad y = \delta_\mathrm{m},\quad u_x = 0 \tag{2.264}$$

$$\frac{\partial^n u_x}{\partial y^n} = 0,\quad n = 1, 2, \ldots \tag{2.265}$$

$$\text{at}\quad y = \delta_\mathrm{T},\quad T = T_\infty \tag{2.266}$$

$$\frac{\partial^n T}{\partial y^n} = 0,\quad n = 1, 2, \ldots \tag{2.267}$$

In particular, if a cubic expression is used for $u_x\,(y)$ and a quadratic expression for $T\,(y)$, the resulting profiles that are consistent with the boundary conditions are

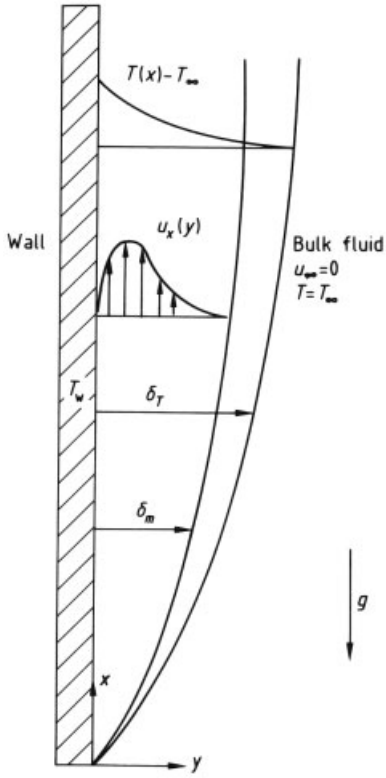

Figure 16. Free convection from a vertical heated plate, development of momentum, and thermal boundary layers

$$u_x = u_c(x)\,\frac{y}{\delta_m}\left(1 - \frac{y}{\delta_m}\right)^2 \tag{2.268}$$

$$T = T_\infty + (T_w - T_\infty)\left(1 - \frac{y}{\delta_T}\right)^2 \tag{2.269}$$

where u_c is a characteristic velocity that varies with x. If $u_c = A\,x^m$ and $\delta_m = \delta_T = B\,x^n$, and these relations are substituted into Equations (2.261) and (2.240), the boundary layer thickness δ_m and the local and overall heat-transfer coefficients can be determined [21, pp. 371–376]. The results are

$$\frac{\delta}{x} = 3.936\left(\frac{0.952 + Pr}{Pr^2}\right)^{1/4}\frac{1}{Gr_x^{1/4}} \tag{2.270}$$

$$Nu_x \equiv \frac{h_x\,x}{\lambda} = 0.508\,(Pr\,Gr_x)^{1/4}\left(\frac{Pr}{0.952 + Pr}\right)^{1/4} \tag{2.271}$$

and

$$Nu_L \equiv \frac{h_L\,L}{\lambda} = 0.678\,(Pr\,Gr_L)^{1/4}\left(\frac{Pr}{0.952 + Pr}\right)^{1/4} \tag{2.272}$$

Here Gr_x is the local Grashof number given by

$$Gr_x = \frac{g\beta_\infty\,(T_w - T_\infty)\,x^3}{\nu} \tag{2.273}$$

and $Pr\,Gr_x$ is the Rayleigh number given by

$$Ra_x \equiv Pr\,Gr_x = \frac{g\beta\,(T_w - T_\infty)\,x^3}{\nu\alpha} \tag{2.273a}$$

The assumption of $\delta_m = \delta_T$ would seem to limit the results to Prandtl numbers on the order of 1; however, because the effect of the velocity profile on the temperature profile is small as Pr increases, the result is also useful at higher Prandtl numbers.

Empirically, CHURCHILL and CHU [47] have found that free convection data on vertical surfaces correlate quite accurately with

$$Nu_L = 0.68 + 0.67\,(Pr\,Gr_L)^{1/4} \cdot \left[1 + \left(\frac{0.492}{Pr}\right)^{9/16}\right]^{-4/9} \tag{2.274}$$

In using this correlation, the thermal properties should be evaluated at the mean boundary layer temperature, except for β_∞ which should be evaluated at T_∞ if the fluid is a gas. Equation (2.272) is quite accurate at $Pr \approx 1$ or greater; however, it is not valid at low Prandtl numbers (liquid metals), and Equation (2.274) must be used instead.

2.5. Turbulent Momentum and Energy Transfer

2.5.1. Physical Characteristics

Much flow of industrial importance is turbulent flow. Momentum and energy transport in turbulent flow are characterized by time-varying velocity and temperature fields. In particular, for turbulent flow in a tube under constant flow rate conditions, the axial velocity component at a specific position in the flow appears as in Figure 17 A with the velocity fluctuating around a time-averaged value. The same fluctuating behavior is also observed for the other velocity

components. Also, if heat is convected between the wall and the fluid, the temperature at each spatial position in the flow also fluctuates.

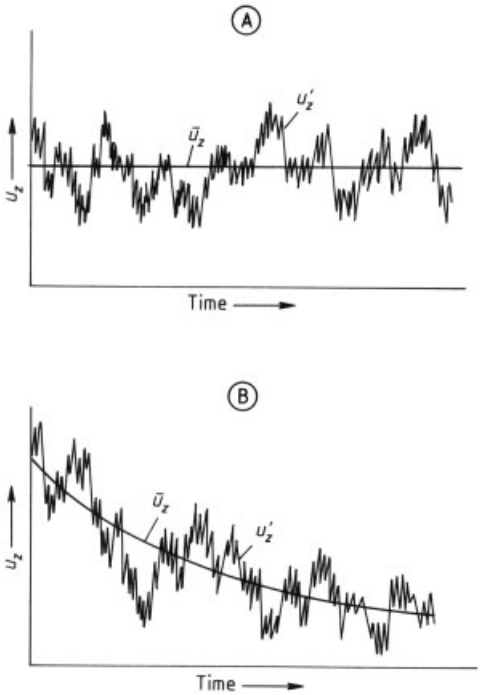

Figure 17. Illustration of velocity fluctuations in turbulent flow
The velocity consists of a time-average part ($\bar{u}z$) and a fluctuating part ($u'z$).
A) Steady turbulent flow, $\bar{u}_z$ constant with time; B) Unsteady turbulent flow, $\bar{u}_z$ varying with time

In describing this behavior, the velocity and temperature variables are conveniently separated into time-averaged values plus fluctuating parts, i.e.,

$$\boldsymbol{u} = \bar{\boldsymbol{u}} + \boldsymbol{u}' \tag{2.275}$$

and

$$T = \bar{T} + T' \tag{2.276}$$

where $\bar{\boldsymbol{u}}$ and $\bar{T}$ are the time-averaged values and $\boldsymbol{u}'$ and T' are the fluctuating parts. The time-averaged value at time t of a variable φ is defined by

$$\bar{\varphi} \equiv \frac{1}{t_0} \int_t^{t+t_0} \varphi \mathrm{d}t \tag{2.277}$$

where φ can correspond to $\boldsymbol{u}$, T, p, or any other dependent variable. The time t_0 over which the average is taken is long compared to the reciprocal frequency of the fluctuations, but short compared to the time of macroscopic changes in pressure drop, wall temperature, and other externally controlled variables.

Although turbulence is inherently unsteady, steady conditions in turbulent flow can be considered if

$$\frac{\partial \bar{\varphi}}{\partial t} = 0 \tag{2.278}$$

In Figure 17 A, steady turbulent flow is illustrated; in Figure 17 B, unsteady turbulent flow. The latter might arise in a flow field if the applied pressure drop decreases with time.

To obtain a measure of the relative magnitude of the fluctuating part of a turbulent flow variable, the quantity

$$I_\varphi \equiv \frac{\sqrt{(\varphi')^2}}{\bar{\varphi}} \tag{2.279}$$

can be defined. In particular, for a shear flow [$\bar{u}_x = \bar{u}_x(y)$] the intensity of turbulence is defined by

$$I_u \equiv \frac{\sqrt{(u_x')^2 + (u_y')^2 + (u_z')^2}}{\bar{u}} \tag{2.280}$$

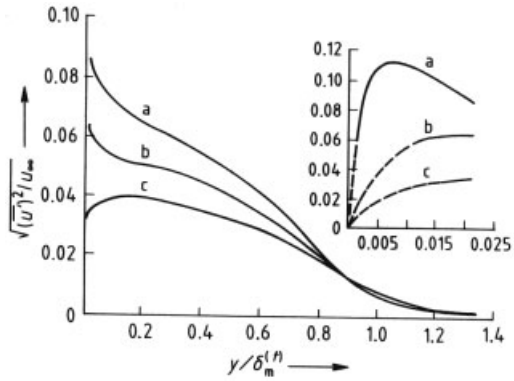

Figure 18. Distribution of intensities of the different velocity components in a turbulent boundary layer

$$\mathrm{a} = \frac{\left(\overline{u_x'}\right)^2 \sqrt{(u_x')^2 + (u_y')^2 + (u_z')^2}}{u_\infty},$$

$$\mathrm{b} = \frac{\left(\overline{u_z'}\right)^2 \sqrt{(u_x')^2 + (u_y')^2 + (u_z')^2}}{u_\infty},$$

$$\mathrm{c} = \frac{\left(\overline{u_y'}\right)^2 \sqrt{(u_x')^2 + (u_y')^2 + (u_z')^2}}{u_\infty}$$

In Figure 18, the intensities of individual velocity components are shown for a turbulent boundary layer on a flat plate as observed by KLEBANOFF [16, p. 275], [49]. As can be seen, the largest component intensity is that associated with the contribution in the direction of flow (u_x), with a maximum value of 0.11 which occurs very close to the wall ($y/\delta_{\mathrm{m}}^{(t)} \approx 0.007$). The smallest component intensity is that associated with u_y, the velocity component normal to the wall, where a value of 0.04 is observed at $y/\delta_{\mathrm{m}}^{(t)} = 0.15$. Although the fluctuating intensities may appear small, these quantities are important in the heat- and mass-transfer rates observed in turbulent flow.

In Figure 19, the normalized velocity profile $\bar{u}_x/u_\infty$ for a turbulent boundary layer on a flat plate (b) is compared with that for a laminar boundary layer (a), both corresponding to a local Reynolds number Re_x of 10^6. The turbulent boundary layer thickness is nearly five times that of the laminar boundary layer, and in turbulent flow the velocity profile is more uniform with a higher ratio of average velocity to maximum velocity. Also, the wall velocity gradient and the corresponding wall friction drag are higher in the turbulent case. The latter might suggest that a laminar boundary layer would be more desirable in minimizing frictional drag on surfaces; however, in most cases a turbulent boundary layer is preferred because it is more effective in resisting boundary layer separation and increased form drag.

For turbulent flow in tubes and along flat surfaces, various flow regions can be identified. At distances sufficiently close to the wall, the turbulent fluctuations diminish and a *laminar sublayer* is observed. In this region, momentum or energy transfer occurs largely from conventional molecular and laminar convection mechanisms. At distances sufficiently remote from the wall, a fully developed *turbulent core* is present where the momentum- and energy-transfer mechanisms are largely a result of convective mixing of the turbulent eddies. In the *buffer zone* between the laminar sublayer and the turbulent core, the flow is neither laminar nor fully developed turbulent flow. In this region, both types of molecular and turbulent transport mechanisms are present.

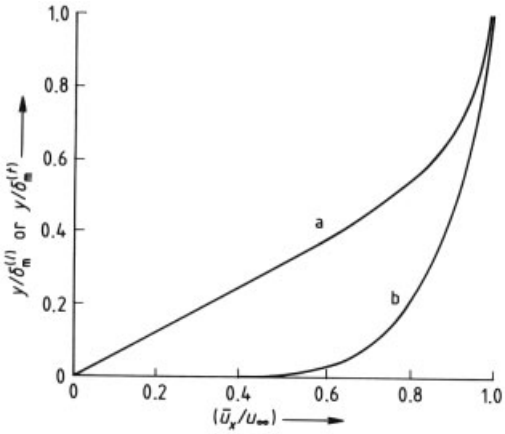

Figure 19. Dimensionless laminar and turbulent velocity profiles for flow on a flat plate at a local Reynolds number $Re_x = 10^6$
Here, $\delta_{\mathrm{m}}^{(l)}$ and $\delta_{\mathrm{m}}^{(t)}$ denote the momentum boundary layer thickness for laminar (a) and turbulent flow, (b) respectively; $\delta_{\mathrm{m}}^{(t)}/\delta_{\mathrm{m}}^{(l)} \approx 4.8$.

The velocity distribution across these three regions can be described by the semiempirical relations
Laminar sublayer:

$$u^+ = y^+ \quad \text{for} \quad y^+ < 5 \tag{2.281}$$

Buffer zone:

$$u^+ = -3.05 + 5\ln y^+ \quad \text{for} \quad 5 \le y^+ < 30 \tag{2.282}$$

Turbulent core:

$$u^+ = 5.5 + 2.5\ln y^+ \quad \text{for} \quad 30 \le y^+ \tag{2.283}$$

Here, u^+ and y^+ are defined by

$$u^+ \equiv u_x/u_* \tag{2.284}$$

$$y^+ \equiv y u_*/\nu, \tag{2.285}$$

and u_* is the friction velocity defined by

$$u_* \equiv \sqrt{\tau_{\mathrm{w}}/\varrho} \tag{2.286}$$

where τ_{w} is the wall shear stress. The velocity distribution represented by Equations (2.281) – (2.283) is commonly referred to as the *universal velocity distribution.*

Another form of the velocity distribution that is commonly used to describe turbulent flow in tubes is

$$\frac{\bar{u}_x}{u_{\max}} = \left(\frac{y}{r_1}\right)^{1/n} \tag{2.287}$$

where y is the distance measured from the wall and r_1 is the tube radius. NIKURADSE [50] obtained excellent agreement with experimental observations when n was allowed to vary with Reynolds number (see Table 6). For Reynolds numbers of ca. 10^5, a value of $n=7$ is commonly used. Note that Equation (2.287) gives an infinite velocity gradient at the tube wall. As a result, drag force values cannot be predicted directly from velocity gradients calculations at the wall.

Table 6. Characteristics of velocity profile for turbulent flow in a tube [50]

Re_d	n in Equation (2.287)	$u_{\mathrm{avg}}/u_{\mathrm{max}}$
4×10^3	6	0.791
1.1×10^5	7	0.817
3.24×10^6	10	0.865

Although Equation (2.281)–(2.283) and (2.287) were developed originally for tube flow, they are also valid for turbulent boundary layers on flat plates. In this case, $u_{\max}$ in these equations is replaced by u_∞ and r_1 by $\delta_{\mathrm{m}}^{(t)}$. Also, because $\tau_{\mathrm{w}}=f\varrho u_{\mathrm{avg}}^2/2$ for tubes, an expression for $u_{\mathrm{avg}}/u_{\max}$ must be used to apply the results to flat plates. As illustrated in Table 6, this ratio for tube flows varies with Reynolds number. For Reynolds numbers of ca. 10^5, a value of $u_{\mathrm{avg}}/u_{\max}\approx 0.8$ is commonly used.

2.5.2. Time-Smoothed Momentum and Energy Balances for Turbulent Flow

The equations of continuity and motion given previously are perfectly valid for describing turbulent flow; however, because of the unsteady, three-dimensional fluctuations of velocity and other dependent variables, the analysis requires simultaneous solution of the full three-dimensional equations of motion along with the equation of continuity. Even if appropriate boundary and initial conditions could be specified, such solutions are beyond current mathematical capabilities. A less ambitious approach is to focus on the time-smoothed quantities and to describe these variables as a function of position and time by using time-smoothed forms of the equations of continuity, motion, and energy.

The specific forms of the *continuity* and *motion equations* can be obtained by time-smoothing Equations (1.65) and (1.73) term by term. Before carrying out these operations, Equation (1.73) can be put in the equivalent form

$$\frac{\partial}{\partial t}(\varrho\boldsymbol{u}) + \boldsymbol{\nabla}\cdot(\varrho\boldsymbol{u}\boldsymbol{u}) = -\boldsymbol{\nabla}p - \boldsymbol{\nabla}\cdot\boldsymbol{\tau} + \varrho\boldsymbol{g} \quad (2.288)$$

where the equation of continuity has been used on the left side of Equation (1.73) and Equation (1.84) on the right side. The key time-smoothing identities are

$$\overline{\mathfrak{F}\varphi} = \mathfrak{F}\overline{\varphi} \quad (2.289)$$

$$\overline{\mathfrak{F}\varphi\psi} = \mathfrak{F}\overline{\varphi\psi} + \mathfrak{F}\overline{\varphi'\psi'} \quad (2.290)$$

where $\mathfrak{F}$ is any linear time or spatial derivative operator (e.g., $\partial/\partial t$, ∇, $\nabla\cdot$) and φ and ψ represent scalar, vector, or tensor functions, or combinations thereof. By using these identities, the following time-smoothed equations are obtained (incompressible fluids):

$$\boldsymbol{\nabla}\cdot\bar{\boldsymbol{u}} = 0 \quad (2.291)$$

$$\varrho\left(\frac{\partial\bar{\boldsymbol{u}}}{\partial t} + \bar{\boldsymbol{u}}\cdot\boldsymbol{\nabla}\bar{\boldsymbol{u}}\right) = -\boldsymbol{\nabla}\bar{p} - \boldsymbol{\nabla}\cdot \left(\bar{\boldsymbol{\tau}} - \varrho\overline{\boldsymbol{u}'\boldsymbol{u}'}\right) + \varrho\boldsymbol{g} \quad (2.292)$$

If this equation is compared with the form of Equation (2.288) for incompressible fluids, i.e.,

$$\varrho\left(\frac{\partial\boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right) = -\boldsymbol{\nabla}p - \boldsymbol{\nabla}\cdot\boldsymbol{\tau} + \varrho\boldsymbol{g} \quad (2.293)$$

the time-smoothed equation is seen to have the same form, except that $\bar{\boldsymbol{u}}$ and $\bar{p}$ replace $\boldsymbol{u}$ and p, and $\boldsymbol{\tau}^{(\mathrm{t})} \equiv \bar{\tau} - \varrho\overline{\boldsymbol{u}'\boldsymbol{u}'}$ replaces $\boldsymbol{\tau}$. The components of $\varrho\overline{\boldsymbol{u}'\boldsymbol{u}'}$ are called the *Reynolds stresses.* Physically, these represent the convective momentum flux components associated with the fluctuating part of the velocity.

One could expect that just as $\boldsymbol{\tau}$ depends on the deformation rate, the Reynolds stresses might be expressible in terms of the mean velocity and its spatial derivatives. Unfortunately, this does not appear to be possible, and when attempted, the coefficients associated with the mean velocity terms and its derivatives are spatially dependent and specific to each flow field.

In general, a proper description of the $\varrho\overline{\boldsymbol{u}'\boldsymbol{u}'}$ contributions requires a more complete statistical analysis of turbulence. Although progress is being made in this area, an approach that can be

used in engineering analyses is not yet available. As a result, only forms of the Reynolds stresses that are valid for specific flows can be obtained. This is the approach taken here in restricting attention to flow in tubes and along flat surfaces.

Finally, note that the time-smoothed form of the *energy balance* Equation (2.191) is

$$\varrho c_p \left(\frac{\partial \bar{T}}{\partial t} + \bar{\boldsymbol{u}} \cdot \boldsymbol{\nabla} \bar{T} \right) = -\boldsymbol{\nabla} \cdot \left(\bar{\boldsymbol{q}} + \varrho c_p \overline{\boldsymbol{u}' T'} \right) \tag{2.294}$$

where $\bar{\boldsymbol{q}} = -\lambda \boldsymbol{\nabla} \bar{T}$. This equation is of the same form as Equation (2.199), except that $\boldsymbol{q}^{(t)} \equiv \bar{\boldsymbol{q}} + \varrho c_p \overline{\boldsymbol{u}' T'}$ replaces $\boldsymbol{q} = -\lambda \boldsymbol{\nabla} T$. Also, this equation neglects viscous heat dissipation, reversible work, and heat generation. The quantity $\varrho c_p \overline{\boldsymbol{u}' T'}$ is the heat flux associated with turbulent fluctuations. As with Reynolds stresses, $\varrho c_p \overline{\boldsymbol{u}' T'}$ must be expressed in terms of the time-smoothed quantities and their derivatives before solutions to Equation (2.294) can be obtained. Similar to Reynolds stresses, these relations are not universal and the coefficients in such expressions are spatially dependent and specific to each flow.

2.5.3. Mixing Length Theories

One of the common approaches used in relating $\varrho \overline{\boldsymbol{u}' \boldsymbol{u}'}$ and $\varrho c_p \overline{\boldsymbol{u}' T'}$ to time-smoothed quantities is through a mixing length interpretation of turbulent transport. In particular, this approach assumes a mechanism for turbulent transport which involves irregular movement of discrete fluid elements in the flow in a manner analogous to the kinetic behavior of molecules in a gas.

For the case of unidirectional turbulent flow along a surface as illustrated in Figure 20, where the time-smoothed entity $\mathfrak{M}$ (momentum, energy, mass, etc.) varies with distance y from the surface, the flux arising from the turbulent fluctuations is given by

$$\text{Flux of } \mathfrak{M} \text{ in } y \text{ direction} = -\varepsilon_{\mathfrak{M}} \frac{\mathrm{d}\bar{\mathfrak{M}}}{\mathrm{d}y} \tag{2.295}$$

Here $\varepsilon_{\mathfrak{M}}$ is the *eddy diffusivity* associated with the transfer of $\mathfrak{M}$ by the turbulent fluctuations. Basically, the fluctuations give rise to the "diffusion" of discrete fluid elements that produce a flux of $\mathfrak{M}$ down the gradient of decreasing $\bar{\mathfrak{M}}$.

For the exchange of x-momentum between different y layers moving at different velocities $\bar{u}_x$:

$$\varrho \overline{u_y' u_x'} = \varepsilon_{\mathrm{m}} \frac{\mathrm{d}}{\mathrm{d}y} (\varrho \bar{u}_x) = \varrho \varepsilon_{\mathrm{m}} \frac{\mathrm{d}\bar{u}_x}{\mathrm{d}y} \tag{2.296}$$

where ε_{m} is the *eddy momentum diffusivity*.

Similarly, for the heat flux between different y layers at different temperatures $\bar{T}$:

$$\varrho c_p \overline{u_y' T'} = \varepsilon_T \frac{\mathrm{d}}{\mathrm{d}y} (\varrho c_p \bar{T}) = \varrho c_p \varepsilon_T \frac{\mathrm{d}\bar{T}}{\mathrm{d}y} \tag{2.297}$$

where ε_T is the *eddy thermal diffusivity.*

From largely physical arguments, PRANDTL [46], [51] was able to relate ε_{m} and ε_T to the motion of discrete fluid elements between layers with different mean velocities to obtain

$$\varepsilon_{\mathrm{m}} = \varepsilon_T \equiv \varepsilon = l^2 \left| \frac{\mathrm{d}\bar{u}_x}{\mathrm{d}y} \right| \tag{2.298}$$

where l is the *Prandtl mixing length*. The mixing length is a measure of the distance (in the y-direction) that the discrete fluid elements move before they accommodate to the mean velocity of the surroundings. This quantity is somewhat analogous to the mean free path in the kinetic theory of gases. For one-dimensional flow in tubes and along surfaces (e.g., Fig. 20), the mixing length has been empirically found to increase with distance from the wall. Specifically, the relation

$$l = ky \tag{2.299}$$

appears to represent experimental observations with the constant k being ca. 0.4.

In summary, for turbulent transport in shear flows where $\bar{u}_x = \bar{u}_x(y)$, the total x-momentum flux in the y-direction can be given by

$$\tau_{yx}^{(t)} = -\varrho (\nu + \varepsilon) \frac{\mathrm{d}\bar{u}_x}{\mathrm{d}y} \tag{2.300}$$

and the total heat flux $q_y^{(t)}$ in the y-direction by

$$q_y^{(t)} = -\varrho c_p (\alpha + \varepsilon) \frac{\mathrm{d}\bar{T}}{\mathrm{d}y} \tag{2.301}$$

where ν and α are the molecular momentum and thermal diffusivities, respectively, and ε is the eddy diffusivity. At positions close to the wall in the laminar sublayer, the fluctuations have negligible importance, and $\nu \gg \varepsilon$ and $\alpha \gg \varepsilon$;

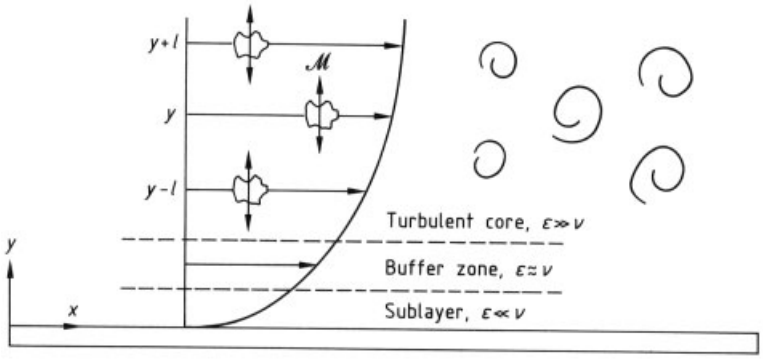

Figure 20. Turbulent transport in unidirectional flow along a surface

whereas in the bulk turbulent field away from the wall, $\varepsilon \gg \nu$ and $\varepsilon \gg \alpha$ for high-intensity turbulent flow. In the buffer region between the laminar sublayer and the turbulent core, the molecular and eddy diffusivities are generally of the same order of magnitude.

2.5.4. Turbulent Heat Transfer in Tubes – Reynolds, Prandtl, and von Karman Analogies

Results are now presented for estimating heat-transfer coefficients for convective heat transfer in tubes under turbulent flow conditions. The heat transfer is assumed to arise from temperature differences between the heated (or cooled) bounding surface and the bulk fluid. The analysis follows that of PRANDTL [51] and is aimed at developing an expression involving the heat-transfer coefficient h, the friction factor f, the average velocity u_{avg}, and the properties of the fluid.

The key assumptions are (1) the turbulent flow field is divided into two zones, a laminar sublayer that extends to $y^+ = 5$ and a turbulent core that occupies the region beyond this point; (2) the velocity profile in the laminar sublayer is given by Equation (2.281); and (3) the ratio $\bar{q}_y/\bar{\tau}_{yx}$ of the heat and momentum fluxes is constant in both the sublayer and the turbulent core. The last assumption implies that the temperature and velocity profiles are functionally similar.

In the *laminar sublayer* where turbulent eddy transport is negligible,

$$\frac{\bar{q}_y}{\bar{\tau}_{yx}} = \frac{-\varrho c_p \alpha \frac{\mathrm{d}\bar{T}}{\mathrm{d}y}}{-\varrho \nu \frac{\mathrm{d}\bar{u}_x}{\mathrm{d}y}} = \frac{c_p \alpha}{\nu} \frac{\mathrm{d}\bar{T}}{\mathrm{d}\bar{u}_x} \tag{2.302}$$

or, because $\bar{q}_y/\bar{\tau}_{yx}$ is assumed to be constant,

$$\frac{\bar{q}_y}{\bar{\tau}_{yx}} = \frac{q_{\mathrm{w}}}{(-\tau_{\mathrm{w}})} = \frac{c_p}{Pr} \frac{\mathrm{d}\bar{T}}{\mathrm{d}\bar{u}_x} \tag{2.303}$$

Integration of this equation over the laminar sublayer gives

$$\bar{T}_\xi - T_{\mathrm{w}} = -\frac{Pr}{c_p}\left(\frac{q_{\mathrm{w}}}{\tau_{\mathrm{w}}}\right)\bar{u}_\xi \tag{2.304}$$

where $Pr = \nu/\alpha$. Here the edge of the sublayer is denoted by $y = \xi$, and $\bar{T}_\xi$ and $\bar{u}_\xi$ are the temperature and velocity at this position.

In the *turbulent core* the eddy momentum and energy transport dominate the molecular transport. As a result,

$$\frac{\bar{q}_y}{\bar{\tau}_{yx}} = -\frac{q_{\mathrm{w}}}{\tau_{\mathrm{w}}} \approx \frac{-\varrho c_p \varepsilon_T \frac{\mathrm{d}\bar{T}}{\mathrm{d}y}}{-\varrho \varepsilon_{\mathrm{m}} \frac{\mathrm{d}\bar{u}_x}{\mathrm{d}y}} \tag{2.305}$$

or

$$-\frac{q_{\mathrm{w}}}{\tau_{\mathrm{w}}} = c_p \frac{\mathrm{d}\bar{T}}{\mathrm{d}\bar{u}_x} \tag{2.306}$$

where $\varepsilon_T = \varepsilon_{\mathrm{m}} = \varepsilon$ is assumed as in the Prandtl mixing length theory. Integrating this equation from $y = \xi$, where $\bar{u}_x = \bar{u}_\xi$ and $\bar{T} = \bar{T}_\xi$, to the bulk of the fluid, where $\bar{u}_x = u_{\mathrm{avg}}$ and $\bar{T} = T_\infty$, gives

$$T_\infty - T_\xi = -\frac{1}{c_p}\left(\frac{q_{\mathrm{w}}}{\tau_{\mathrm{w}}}\right)\left(\bar{u}_{\mathrm{avg}} - \bar{u}_\xi\right) \tag{2.307}$$

Adding Equations (2.304) and (2.307) and using

$$q_{\mathrm{w}} = h\,(T_{\mathrm{w}} - T_\infty) \tag{2.308}$$

$$\tau_{\mathrm{w}} = f\,\frac{1}{2}\,\varrho\,u_{\mathrm{avg}}^2 \tag{2.309}$$

and,

$$\bar{u}_\xi = 5\,\sqrt{\tau_{\mathrm{w}}/\varrho} = 5\,u_{\mathrm{avg}}\,\sqrt{f/2}, \tag{2.310}$$

give

$$St \equiv \frac{Nu_d}{Re_d\,Pr} = \frac{h}{\varrho\,c_p\,u_{\mathrm{avg}}} = \frac{f/2}{1+5\,\sqrt{f/2}\,(Pr-1)} \tag{2.311}$$

where *St* is the Stanton number. This equation is called the *Prandtl analogy* and relates the heat-transfer coefficient to the friction factor (or drag coefficient). The quantities Nu_d and Re_d are the Nusselt and Reynolds numbers based on the tube diameter d and the average velocity, i.e., $h\,d/\lambda$ and $u_{\mathrm{avg}}\,d/\nu$, respectively.

For Prandtl numbers of 1 (applicable for many gases), Equation (2.311) reduces to

$$\frac{h}{\varrho\,c_p\,u_{\mathrm{avg}}} = \frac{f}{2} \tag{2.312}$$

which is the *Reynolds analogy.* This result can be obtained directly from the above analysis by neglecting the sublayer and allowing the turbulent core to occupy the entire flow region. Hence, $u_\xi = 0$ and $T_\xi = T_{\mathrm{w}}$ in Equation (2.307), and the Reynolds analogy follows directly from Equations (2.308) and (2.309).

Von Karman [52, pp. 223–227] conducted a more complete analysis including all three zones in the turbulent field (i.e., laminar sublayer, buffer zone, and turbulent core). The final result was

$$St \equiv \frac{h}{\varrho\,c_p\,u_{\mathrm{avg}}} = \frac{f/2}{1+5\,\sqrt{f/2}\,\left\{Pr-1+\ln\left[1+\frac{5}{6}\,(Pr-1)\right]\right\}} \tag{2.313}$$

This analogy includes the Reynolds and Prandtl analogies as special cases.

To use the above analogies to estimate heat-transfer coefficients, an expression is needed for the friction factor f in terms of the Reynolds number and appropriate roughness parameters.

For smooth tubes, a useful empirical expression for f in the range $4\times10^3 \le Re_d \le 10^5$ is the Blasius formula [53] (→ Fluid Mechanics, Chap. 3.3.1.):

$$f \equiv \frac{\tau_{\mathrm{w}}}{\frac{1}{2}\,\varrho\,u_{\mathrm{avg}}^2} = \frac{0.0791}{Re_d^{1/4}} \tag{2.314}$$

At higher Reynolds numbers ($> 10^5$), the universal velocity profile can be used to develop an expression for the friction factor. In particular, if the turbulent core is assumed to extend to the wall, Equation (2.283) can be integrated over the tube cross section to obtain the average velocity:

$$u_{\mathrm{avg}} = u_*\left[1.75+2.5\ln\left(\frac{u_*\,r_1}{\nu}\right)\right] \tag{2.315}$$

Note that

$$u_* \equiv \sqrt{\tau_{\mathrm{w}}/\varrho} = u_{\mathrm{avg}}\,\sqrt{f/2} \tag{2.316}$$

thus, Equation (2.315) can be simplified to the von Karman friction factor relation, i.e.,

$$\frac{1}{\sqrt{f}} = 4.07\log\left(Re\,\sqrt{f}\right) - 0.601 \tag{2.317}$$

This result is quite close to the empirical result of Nikuradse [50] given by

$$\frac{1}{\sqrt{f}} = 4.0\log\left(Re\,\sqrt{f}\right) - 0.40 \tag{2.318}$$

Both of these results are for smooth tubes and are generally used at elevated Reynolds numbers ($> 10^5$).

For *rough tubes,* the effects of relative roughness *e/d* ($e \approx$ height of wall protuberances) must be included. For Reynolds numbers above the value $Re\,\sqrt{f} = 0.01\,d/e$, the Colebrook equation can be used to determine f [54]:

$$\frac{1}{\sqrt{f}} = -4\log\left(\frac{4.67}{Re\,\sqrt{f}} + \frac{e}{d}\right) + 2.28 \tag{2.319}$$

The Colebrook equation reduces to the Nikuradse result when $e/d = 0$.

The degree of roughness of a tube is not always easy to determine accurately. As a result, pressure drops predicted by substituting the above results into Equation (2.80) can often involve considerable error because of uncertainties in estimating *e/d.* This is particularly true for tubing susceptible to scaling and corrosion and tubing that has been in use for some time.

In addition to the above analogies, another result can be obtained from the boundary layer results of Equations (2.235) and (2.255). These latter equations can be combined to give

$$St\,Pr^{2/3} = f/2 \tag{2.320}$$

where f has been substituted for $C_{\mathrm{D}x}$. This equation is identical with the Reynolds analogy (Eq. 2.312) when $Pr = 1$. Based on this agreement at $Pr = 1$, COLBURN suggested the use of Equation (2.320) at higher Prandtl numbers even for turbulent flow (Colbum analogy). Although it is strictly valid only for laminar flow, this simple relation has been found to be a useful approximation for turbulent flow applications particularly, where there is no form drag and where $0.5 < Pr < 50$.

The Colburn analogy can be put in another form:

$$j_{\mathrm{H}} = \frac{F}{2} \tag{2.320a}$$

where

$$j_{\mathrm{H}} \equiv St\,Pr^{2/3} \tag{2.320b}$$

is the *Colburn j factor* for heat transfer. The mass transfer analogy of j_{H} will be discussed later.

2.5.5. Turbulent Momentum and Heat Transferon a Flat Plate

For a turbulent boundary layer on a flat plate, the wall shear stress can be related to the free stream velocity u_∞ and the turbulent boundary layer thickness $\delta_{\mathrm{m}}^{(t)}$ by the Blasius expression:

$$\tau_{\mathrm{w}} = 0.0225\,\varrho\,u_\infty^2 \left(\frac{\nu}{u_\infty\,\delta_{\mathrm{m}}^{(t)}}\right)^{1/4} \tag{2.321}$$

This equation follows from Equation (2.314) if $d = 2\,r_1 \to 2\,\delta_{\mathrm{m}}^{(t)}$, $u_{\max} \to \mathrm{u}_\infty$, and $u_{\mathrm{avg}}/u_{\max} \approx 0.8$. It is valid for flat plate Reynolds numbers up to 10^7. In terms of the drag coefficient $C_{\mathrm{D}x}$, this equation can be written as

$$C_{\mathrm{D}x} \equiv \frac{\tau_{\mathrm{w}}}{\frac{1}{2}\varrho\,u_\infty^2} = \frac{0.0450}{Re_x^{1/4}} \left(\frac{x}{\delta_{\mathrm{m}}^{(t)}}\right)^{1/4} \tag{2.322}$$

To use Equations (2.321) and (2.322), the turbulent boundary layer thickness $\delta_{\mathrm{m}}^{(t)}$ must be determined. If Equation (2.287) is used with $n = 7$, $r_1 \to \delta_{\mathrm{m}}^{(t)}$, and $u_{\max} \to \mathrm{u}_\infty$ in the time-smoothed momentum integral balance (the same as Eq. 2.226, but with the dependent variables replaced by time-smoothed quantities),

$$\frac{\delta_{\mathrm{m}}^{(t)}}{x} = \frac{0.371}{Re_x^{1/5}} \tag{2.323}$$

and

$$C_{\mathrm{D}x} = \frac{0.0577}{Re_x^{1/5}} \tag{2.324}$$

which are valid up to a local Reynolds number of 10^7. In determining $\delta_{\mathrm{m}}^{(t)}$, the turbulent boundary layer has been assumed to extend to the leading edge of the plate ($\delta_{\mathrm{m}}^{(t)} = 0$ at $x = 0$). Obviously, this can lead to considerable error, particularly because the laminar boundary layer that exists up to $x_{\mathrm{cr}} = 2 \times 10^5\ (\nu/u_0)$ has a square-root dependence on x compared to the 4/5 power dependence that results from Equation (2.323). Regardless, if Equation (2.324) is used in the Colburn analogy (with $C_{\mathrm{D}x}$ replacing f and St_x replacing St)

$$Nu_x = 0.0289\,Re_x^{4/5}\,Pr^{1/3} \tag{2.325}$$

is obtained. Given the restrictions on Equation (2.322), this equation is limited to $Re_x < 10^7$.

Based on the experimental work of ZHUKAUSKAS and AMBRAZYAVICHYUS [55], WHITAKER [16, pp. 333 – 336] suggests the use of

$$Nu_x = 0.029\,Re_x^{4/5}\,Pr^{0.43} \left(\frac{\eta_\infty}{\eta_{\mathrm{w}}}\right)^{1/4} \tag{2.326}$$

instead of Equation (2.325), where all properties are evaluated at the free stream temperature T_∞. This equation is similar to Equation (2.325), except for the power dependence on the Prandtl number and the addition of a temperature-dependent viscosity effect. Remember that the origin of the 1/3 power dependence on Pr in Equation (2.325) was from laminar boundary theory. In extending this result to turbulent boundary layers, some modifications would be expected. The 1/4 power dependence on $(\eta_\infty/\eta_{\mathrm{w}})$ is different from that for tube flow where a 0.14 dependence was found. Although such differences are justified on the basis of available data, relatively small experimental errors could cause these differences. As a result, some uncertainty must be associated with the power dependences on Pr and $(\eta_\infty/\eta_{\mathrm{w}})$.

For a plate of length L, the overall heat-transfer coefficient can be determined from

$$h_L = \frac{1}{L}\left\{\int_0^{x_{\rm cr}} h_{x,\rm lam}\,{\rm d}x + \int_{x_{\rm cr}}^{L} h_{x,\rm turb}\,{\rm d}x\right\} \tag{2.327}$$

Using a modified form of Equation (2.252) for $\alpha_{x,\rm lam}$ [Eq. 2.252 with $(\eta_\infty/\eta_{\rm w})^{1/4}$ as a multiplier], and Equation (2.326) for $h_{x,}{}^{\rm turb}$ gives

$$Nu_L = 0.0361\,Re_L^{4/5}\,Pr^{0.43}\left(\frac{\eta_\infty}{\eta_{\rm w}}\right)^{1/4} \cdot\left\{1 - 17\,400\,Re_L^{-4/5}\left(1 - \frac{0.473}{Pr^{0.1}}\right)\right\} \tag{2.328}$$

This equation assumes that the transition Reynolds number is given by $Re_{\rm cr} \equiv (u_\infty\,x_{\rm cr}/\nu) = 2\times10^5$. For sufficiently large values of Re_L where $Re_{\rm cr}/Re_L \ll 1$, Equation (2.328) can be simplified to

$$Nu_L = 0.0361\,Re_L^{4/5}\,Pr^{0.43}\left(\frac{\eta_\infty}{\eta_{\rm w}}\right)^{1/4} \tag{2.329}$$

As with Equation (2.326), all properties in Equations (2.328) and (2.329) are evaluated at the free stream conditions. These equations are valid for $Re_L < 10^7$, Pr between 0.7 and 380, and viscosity ratios between 0.26 and 3.5.

WHITAKER found that the use of Equation (2.328) versus Equation (2.329) depends on the degree of turbulence in the free stream. When the latter is high, the transition Reynolds number is less than 2×10^5 and Equation (2.329) tends to be valid. When the free stream turbulence is low, Equation (2.328) represents experimental observations more accurately. In addition to this effect of free stream turbulence, the roughness of the plate also has a significant effect on the transition Reynolds number. In particular, very smooth plates can exhibit transition values considerably above 2×10^5. Because such effects are difficult to characterize in many applications, some degree of uncertainty is associated with the use of Equations (2.328) or (2.329).

If the fluid properties are evaluated at the film temperature $T_{\rm f} \equiv (T_{\rm w} + T_\infty)/2$ instead of T_∞, the $(\eta_\infty/\eta_{\rm w})$ dependence in Equations (2.328) and (2.329) can be dropped. Such differences are noted in Table 7.

To calculate the overall drag coefficient on a plate of length L that extends into the turbulent boundary layer region, we use

$$C_{\rm D}L = \frac{1}{L}\left\{\int_0^{x_{\rm cr}} C_{{\rm D}x,\rm lam}\,{\rm d}x + \int_{x_{\rm cr}}^{L} C_{{\rm D}x,\rm turb}\,{\rm d}x\right\} \tag{2.330}$$

Using Equations (2.235) and (2.324) and assuming that $Re_{\rm cr} = 2\times10^5$ give

$$C_{\rm D}L = 0.072\,Re_L^{-1/5}\left(1 - 9\,200\,Re_L^{-4/5}\right) \tag{2.331}$$

This equation is valid for $Re_L < 10^7$. At higher Reynolds numbers, the result of PRANDTL can be used [56]:

$$C_{\rm D}L = \frac{0.455}{[\log Re_L]^{2.58}} \tag{2.332}$$

which is valid for Re_L from 10^6 to 10^9.

2.6. Summary of Heat-Transfer Relations

In Table 7, heat-transfer and friction (or drag) coefficient results are summarized for various geometries. They include the flat plate and tube flow results presented here, as well as results for heat transfer around cylinders and spheres. In addition, free convection results are provided for several common situations.

3. Transport in Multicomponent Systems

Multicomponent transport processes require consideration of the fact that to the extent mixture compositions are not uniform, diffusion (the motion of species of the mixture relative to each other) will be nonzero. In binary systems, diffusion processes can be analogous to heat transfer. Multicomponent systems are considerably more complex, however. Useful references on diffusion and general transport references are given in [63–67]. An extensive review of methods for obtaining diffusion coefficients appears in [68].

3.1. Diffusive Mass Transfer in Binary Systems

3.1.1. Species Mass Balances

In considerations of single-component (compositionally homogeneous) systems the equation of continuity (conservation of total mass)

Table 7. Summary of heat-transfer and friction coefficient results (all properties evaluated at $T_f \equiv (T_w - T_\infty)/2$ unless otherwise noted)

Situation	Equation	Restrictions	Equation number () or reference [ ]
Flow over Flat Plates			
	Heat-transfer coefficients		
Laminar, local	$Nu_x = 0.332\, Re_x{}^{1/2}\, Pr^{1/3}$	constant wall temperature; $Re_x < 2\times10^5$, $0.6 < Pr < 50$	(2.252)
Laminar, local	$Nu_x = 0.453\, Re_x{}^{1/2}\, Pr^{1/3}$	constant wall heat flux; $Re_x < 2\times10^5$, $0.6 < Pr < 50$	(2.256)
Laminar, overall	$Nu_L = 0.664\, Re_L{}^{1/2}\, Pr^{1/3}$	constant wall temperature; $Re_L < 2\times10^5$, $0.6 < Pr < 50$	(2.255)
Turbulent, local	$Nu_x = 0.029\, Re_x{}^{4/5}\, Pr^{0.43}\, (\eta_\infty/\eta_w)^{1/4}$	constant wall temperature; $2\times10^5 < Re_x < 10^7$, $0.7 < Pr < 380$; properties evaluated at T_∞	(2.326)
Turbulent, overall	$Nu_L = 0.0361\, Re_L^{4/5}\, Pr^{0.43}\, (\eta_\infty/\eta_w)^{1/4} \times \left[1 - 17\,400 Re_L^{-4/5}\left(1 - \frac{0.473}{Pr^{0.1}}\right)\right]$	constant wall temperature; $2\times10^5 < Re_L < 10^7$, $0.7 \le Pr \le 380$; relatively low free stream turbulence; properties evaluated at T_∞	(2.328)
Turbulent, overall	$Nu_L = 0.0361\, Re_L{}^{4/5}\, Pr^{0.43}\, (\eta_\infty/\eta_w)^{1/4}$	same as above except relatively high free stream turbulence	(2.329)
	Drag coefficients		
Laminar, local	$C_{Dx} = 0.664\, Re_x{}^{-1/2}$	$Re_x < 2\times10^5$	(2.235)
Laminar, overall	$C_{DL} = 1.328\, Re_L{}^{-1/2}$	$Re_L < 2\times10^5$	(2.236)
Turbulent, local	$C_{Dx} = 0.0577\, Re_x{}^{-1/5}$	$2\times10^5 \le Re_x < 10^7$	(2.324)
Turbulent, overall	$C_{DL} = 0.072\, Re_L{}^{-1/5}\, (1 - 9200\, Re_L{}^{-4/5})$	$2\times10^5 \le Re_L < 10^7$	(2.331)
Flow in Tubes			
(Properties evaluated at average bulk mean temperature between $z = 0$ and $z = L$ unless otherwise noted)			
	Heat-transfer coefficients		
Fully developed laminar	$Nu_d = 3.658$	T_w = const.; $Re_d < 2100$; $z > 0.05\, d\, Re_d Pr$	(2.203)
Fully developed laminar	$Nu_d = 4.364$	q_w = const.; $Re_d < 2400$; $z > 0.05\, d\, Re_d Pr$	(2.202)
Developing laminar	$Nu_{d,\,lm} = 1.86\, Re_d{}^{1/3}\, Pr^{1/3}\, (L/d)^{-1/3}\, (\eta_\infty/\eta_w)^{0.14}$	T_w = const.; $Re_d < 2400$; intermediate-length tubes	(2.205)
Developing laminar	$Nu_{d,lm} = 3.66 + \frac{0.0745\,\psi^3}{1 + 0.04\,\psi^2}$ where $\psi \equiv (Re_d Pr)^{1/3}\, (L/d)^{1/3}\, (\eta_{bm}/\eta_w)^{0.14}$	$Re_d < 2400$	(2.209) (2.210)
Turbulent	$St \equiv \frac{Nu_d}{Re_d\, Pr} = \frac{f/2}{1 + 5\sqrt{f/2}\left\{Pr - 1 + \ln\left[1 + \frac{5}{6}(Pr - 1)\right]\right\}}$	$Re_d > 2400$	(2.313)
Turbulent	$Nu_{d,lm} = 0.015\, Re_d{}^{0.83}\, Pr^{0.42}\, (\eta_{bm}/\eta_w)^{0.14}$	$Re_d > 4000$; $0.4 < Pr < 600$	[57]
	Friction coefficients		
Fully developed laminar	$f = 16/Re_d$	$Re_d < 2100$; $z > 0.05\, Re_d d$	(2.87)
Turbulent	$f = 0.0791\, Re_d{}^{-1/4}$	$2100 \le Re_d \le 10^5$; smooth tubes	(2.314)
Turbulent	$\frac{1}{\sqrt{F}} = -4\log\left(\frac{4.67}{Re_d\sqrt{f}} + \frac{e}{d}\right) + 2.28$	$Re_d > 10^5$; smooth ($e = 0$) and rough tubes	(2.319)

Table 7. (Continued)

Situation	Equation	Restrictions	Equation number () or reference [ ]
Flow Around Bodies			
Flow normal to cylinder	$Nu_d = 0.3 + \frac{0.62\, Re_d^{1/2}\, Pr^{1/3}}{\left[1+(0.4/\mathrm{Pr})^{2/3}\right]^{3/4}} \cdot \left[1 + \left(\frac{Re_d}{28200}\right)^{5/8}\right]^{4/5}$	$10^2 < Re_d < 4\times10^6$; $Re_d \equiv u_\infty\, d/\nu$; properties evaluated at film temperature	[58]
Flow past a sphere	$Nu_d = 2 + (0.4\, Re_d{}^{1/2} + 0.06\, Re_d{}^{2/3}) \cdot Pr^{0.4}\, (\eta_\infty/\eta_\mathrm{w})^{1/4}$	$3.5 < Re_d < 8\times10^4$; $0.7 < Pr < 380$; properties evaluated at T_∞	[59]
Free Convection from Various Geometries			
Vertical surface, overall	$Nu_L = 0.68 + \frac{0.67\, Ra_L^{1/4}}{\left[1+\left(\frac{0.492}{Pr}\right)^{9/16}\right]^{4/9}}$	$Ra_L < 10^9$	[60]
Vertical surface, overall	$Nu_L = \left\{0.825 + \frac{0.387\, Ra_L^{1/6}}{\left[1+(0.492/Pr)^{9/16}\right]^{8/27}}\right\}^2$	T_w=const.; $10^{-1} < Ra_L < 10^{12}$	[60]
Vertical surface, laminar, local	$Nu_x = 0.60\, Ra_x{}^{1/5}$	q_w=const.; $10^5 < Gr_x\, Nu_x < 10^{11}$	[61]
Vertical surface, laminar, local	$Nu_x = 0.17\, Ra_x{}^{1/4}$	q_w=const.; $2\times10^{13} < Gr_x\, Nu_x\, Pr < 10^{16}$	[61]
Horizontal cylinder, laminar, overall	$Nu_d = 0.36 + \frac{0.518\, Ra_d^{1/4}}{\left[1+(0.559/Pr)^{9/16}\right]^{4/9}}$	T_w=const.;1 $0^{-6} \le Ra_d \le 10^9$	[61]
Horizontal cylinder, overall	$Nu_d = \left\{0.60 + 0.387\left[\frac{Ra_d}{\left[1+(0.559/Pr)^{9/16}\right]^{16/9}}\right]^{1/6}\right\}^2$	$10^{-5} \le Ra_d \le 10^{12}$	[61]
Horizontal plate, down ward facing hot plates or upward facing cold plates; overall	$Nu_L = 0.58\, Ra_L{}^{1/3}$	T_w=const.; $10^6 < Ra_L < 10^{11}$	[62]
Horizontal plate, upward facing hot plates	$Nu_L = 0.13\, Ra_L{}^{1/3}$	T_w=const.; $Ra_L < 2\times10^8$	[62]
or downward facing cold plates; overall	$Nu_L = 0.16\, Ra_L^{1/3}$	T_w=const.; $2\times10^8 \le Ra_L < 10^{11}$	[62]

$$\frac{\partial \varrho}{\partial t} = -\nabla\cdot(\varrho \boldsymbol{u}) \tag{3.1}$$

plays a vital role in establishing relationships between components of velocity. In this form, $(\partial\varrho/\partial t)_x$ physically represents the accumulation rate per unit volume of mass; the expression $(-\nabla\cdot\varrho\boldsymbol{u})$ represents the net input rate per unit volume to the volume fixed in space as a result of the mass flux $\varrho\boldsymbol{u}$.

An analogous result can be obtained for an individual species (e.g., species A), however, since individual chemical species are not conserved (except for the elements), the net production of A due to reaction must be considered. Accordingly, an equation accounting for the amount of species A is given by

$$\frac{\partial \varrho_\mathrm{A}}{\partial t} = -\nabla\cdot(\varrho_\mathrm{A}\boldsymbol{u}_\mathrm{A}) + r_\mathrm{A} \tag{3.2}$$

where r_A is the rate of production of species A per unit volume in the mixture, ϱ_A is the mass of species A per unit volume of the mixture, and $\boldsymbol{u}_\mathrm{A}$ is the average velocity of the molecules of species A in the mixture.

This equation can also be expressed in terms of molar concentration by dividing by molecular mass to give

$$\frac{\partial c_\mathrm{A}}{\partial t} = -\nabla\cdot(c_\mathrm{A}\boldsymbol{u}_\mathrm{A}) + R_\mathrm{A} \tag{3.3}$$

where c_{A} is the number of moles per volume of species A in the mixture and R_{A} is the molar rate of production of species A by reaction per unit volume of the mixture. The velocity in this equation is the same as in the preceding equation.

An equation of this sort in either mass or molar form can be written for each species in the mixture. When all the equations for all the species of an n-component mixture are added together, the following result is obtained:

$$\sum_{i=1}^{n} \frac{\partial \varrho_i}{\partial t} = -\sum_{i=1}^{n} \nabla \cdot (\varrho_i \boldsymbol{u}_i) + \sum_{i=1}^{n} r_i \tag{3.4}$$

which, with interchange of the summations and derivatives, is

$$\frac{\partial \sum\limits_{i=1}^{n} \varrho_i}{\partial t} = -\nabla \cdot \left(\sum_{i=1}^{n} \varrho_i \boldsymbol{u}_i \right) + \sum_{i=1}^{n} r_i \tag{3.5}$$

Now, by definition, the sum of the individual species' densities (mass per unit volume of the mixture) is the total mass per unit volume of the mixture, which is customarily defined as ϱ to be consistent with the notation for compositionally homogeneous systems. Furthermore, the sum of all the individual species production rates by reaction must be zero, when expressed in mass units, because total mass is conserved (i.e., no net generation of total mass occurs). Consequently, the equation for the total mixture becomes

$$\frac{\partial \varrho}{\partial t} = -\nabla \cdot \left(\sum \varrho_i \boldsymbol{u}_i \right) \tag{3.6}$$

Because of the similarity of this equation to the total mass conservation equation, the sum of the individual species fluxes is commonly defined as the product of the total mass density times an average velocity for the mixture. In fact, this becomes a defining relation for what is referred to as the mixture *mass average velocity* $\boldsymbol{u}$

$$\varrho \boldsymbol{u} \equiv \sum_{i=1}^{n} \varrho_i \boldsymbol{u}_i \tag{3.7}$$

or

$$\boldsymbol{u} \equiv \frac{\sum\limits_{i=1}^{n} \varrho_i \boldsymbol{u}_i}{\sum\limits_{i=1}^{n} \varrho_i} \tag{3.8}$$

Thus, in a mixture, different species may move at different velocities and the mixture velocity is defined as a weighted average of these individual velocities.

A parallel set of equations exists for molar concentrations. Accordingly, the sum of the molar concentrations of the individual species gives

$$\sum_{i=1}^{n} \frac{\partial c_i}{\partial t} = -\sum_{i=1}^{n} \nabla \cdot (c_i \boldsymbol{u}_i) + \sum_{i=1}^{n} R_i \tag{3.9}$$

Again the summation and derivatives can be interchanged and the sum of the individual species concentrations (number of moles of species i per unit volume of the mixture) gives the total number of moles of the mixture per unit volume, customarily denoted as c. Accordingly, an accounting equation for the total number of moles in the mixture is

$$\frac{\partial c}{\partial t} = -\nabla \cdot \sum c_i \boldsymbol{u}_i + \sum R_i \tag{3.10}$$

In terms of moles, however, the net number of moles produced by reaction is not necessarily zero; the total number of moles due to reaction is not conserved. Consequently, this total molar production rate must be included in the total molar accounting equation for the mixture. However, for convenience, a mixture average velocity is defined such that the summation of the individual species fluxes represents a mixture average flux in terms of the total molar concentration and the mixture *molar average velocity*. Accordingly,

$$c \tilde{\boldsymbol{u}} \equiv \sum_{i=1}^{n} c_i \boldsymbol{u}_i \tag{3.11}$$

or, equivalently,

$$\tilde{\boldsymbol{u}} \equiv \frac{\sum\limits_{i=1}^{n} c_i \boldsymbol{u}_i}{\sum\limits_{i=1}^{n} c_i} \tag{3.12}$$

This molar average velocity is not the same as the mass average velocity (Eqs. 3.7 and 3.8) because the individual species are not necessarily of the same molecular mass. The definition of a molar average velocity for the mixture allows

the molar accounting equation to be written in the form

$$\frac{\partial c}{\partial t} = -\nabla \cdot (c\tilde{\boldsymbol{u}}) + \sum_{i=1}^{n} R_i \tag{3.13}$$

Again, the net number of moles produced by reaction is not necessarily zero as it was for mass, and as a result, this term must remain in the equation. Mass and molar accounting equations are thus obtained for the individual species, along with statements of the conservation of total mass for the mixture. Now diffusion must be defined for the individual species and methods for calculating diffusion fluxes must be introduced.

3.1.2. Species Diffusion Fluxes

The preceding discussion recognized that each species may be characterized by its own velocity which, in general, is different from the average velocity of the mixture. Furthermore, the mixture average velocity may be defined in terms of mass or in terms of number of moles, and the two velocities in general are different. As a result, each species has a relative motion with respect to the mixture average motion.

This difference in fluxes between each species relative to the mixture average is referred to as diffusion. Because two mixture average velocities were defined, one based on number of moles and one based on mass, two different diffusion fluxes exist for each species, one with respect to the mixture molar average velocity and one with respect to its mass average velocity.

Accordingly, the molar diffusion flux $\tilde{\boldsymbol{J}}_{\mathrm{A}}$ of species A can be defined as the difference between the total flux of species A and that due to the mixture motion on the average [63]:

$$\tilde{\boldsymbol{J}}_{\mathrm{A}} \equiv c_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} - x_{\mathrm{A}}\,(c\tilde{\boldsymbol{u}}) = c_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} - c_{\mathrm{A}}\tilde{\boldsymbol{u}} \tag{3.14}$$

where x_{A} is the mole fraction of species A. Using this definition of a diffusion flux in the accounting equation for species A gives

$$\frac{\partial c_{\mathrm{A}}}{\partial t} = -\nabla \cdot (c_{\mathrm{A}}\tilde{\boldsymbol{u}}) - \nabla \cdot \tilde{\boldsymbol{J}}_{\mathrm{A}} + R_{\mathrm{A}} \tag{3.15}$$

Here, the accumulation rate of the number of moles of A results from the molar flux of A due to the average flow of the mixture, plus the molar diffusion flux of species A above and beyond this mixture molar average flow, plus the molar rate of production of A per unit volume.

Similarly, a diffusion flux can be defined in terms of mass fluxes, i.e., the difference between the total mass flux of species A and the mass flux of species A that results from the mixture moving at its mass average velocity:

$$\boldsymbol{j}_{\mathrm{A}} \equiv \varrho_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} - \varrho_{\mathrm{A}}\boldsymbol{u} \tag{3.16}$$

This mass diffusion flux used in the mass accounting equation for species A gives

$$\frac{\partial \varrho_{\mathrm{A}}}{\partial t} = -\nabla \cdot (\varrho_{\mathrm{A}}\boldsymbol{u}) - \nabla \cdot \boldsymbol{j}_{\mathrm{A}} + r_{\mathrm{A}} \tag{3.17}$$

Again the accumulation rate of species A at a point in space results from the mass flux of species A due to the bulk flow of the fluid (the mass average velocity of the mixture) plus the diffusion flux of species A plus the mass rate of production (per unit volume) of A by reaction.

In addition to the molar diffusion flux of A and the mass diffusion flux of A as defined above, a molar diffusion flux of A could be defined relative to the mass average velocity according to the equation

$$\boldsymbol{J}_{\mathrm{A}} \equiv c_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} - c_{\mathrm{A}}\boldsymbol{u} \tag{3.18}$$

or a mass diffusion flux of A could be defined relative to the molar average velocity:

$$\tilde{\boldsymbol{j}}_{\mathrm{A}} \equiv \varrho_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} - \varrho_{\mathrm{A}}\tilde{\boldsymbol{u}} \tag{3.19}$$

Neither of these is as commonly used as the first two, so they are not considered further here. The point of this discussion is that when talking about diffusion and diffusion coefficients the velocity used to characterize the mixture must be specified. If the molar average velocity is used, one value for diffusion is appropriate; if the mass average velocity is used, another (different) diffusion is described.

3.1.3. Fick's First Law of Diffusion

Until now the transport equations presented in this section have been truly general for multicomponent (as opposed to binary) systems. The species accounting equations and the mixture mass conservation equations are correct, independent of whether binary or multicomponent

systems are involved. Likewise, the definitions of diffusion flux are still valid for multicomponent systems.

However, at this point a relationship is defined between diffusion driving forces and species fluxes. The relationship to be discussed is restricted to binary systems and is known as Fick's first law of diffusion. This law is a linear transport equation relating mass transfer or diffusion driving forces to the diffusion flux. Accordingly, in a binary system the driving force for diffusion (i.e., the reason different species move with different velocities) is the compositional inhomogeneity (the existence of mole fraction gradients). (Pressure and temperature gradients may also induce diffusion [63]. These effects are normally small and are not considered further here.) In the presence of such gradients, species within the mixture move in such a way as to remove these mole fraction gradients, a motion that occurs spontaneously (implying an increase in entropy of the mixture).

In accordance with a linear relation between this driving force and diffusion flux, *Fick's first law of diffusion* is written as

$$\tilde{\boldsymbol{J}} = -cD_{\mathrm{AB}}\nabla x_{\mathrm{A}} \tag{3.20}$$

In this equation, ∇x_{A} is the driving force, and the product of the total mixture molar concentration and the diffusitivity is the proportionality factor required to convert this driving force to a molar flux. The minus sign indicates that the diffusion flux is in the direction opposite to the direction of increasing mole fraction. The diffusivity is a parameter that describes how easily species A moves through the rest of the mixture, species B – consequently, the notation D_{AB}.

Similarly, the mass diffusion flux of species A can be written in terms of a mass fraction gradient with the proportionality factor being the product of the mixture mass density and the same diffusion coefficient as for the molar diffusion flux. A priori, the same diffusivity might not be expected to appear in both the molar and mass diffusion flux equations. However, this must be the case for binary systems. Furthermore $D_{\mathrm{AB}} = D_{\mathrm{BA}}$ in a binary system, because $x_{\mathrm{A}} = 1 - x_{\mathrm{B}}$ and $\nabla x_{\mathrm{A}} = -\nabla x_{\mathrm{B}}$, and the sum of the two species diffusion fluxes must be zero. Consequently, for binary mixtures a single diffusivity needs to be reported for any given mixture composition of A and B.

3.1.4. Special Cases of Diffusion

Using the relations for diffusion molar and mass fluxes (Fick's first law, Section 3.1.3) and the accounting equations presented in Section 3.1.2 yields the general results that account for species A in either moles or mass. Accordingly, in molar units

$$\frac{\partial c_{\mathrm{A}}}{\partial t} = -\nabla\cdot(c_{\mathrm{A}}\tilde{\boldsymbol{u}}) + \nabla\cdot(cD_{\mathrm{AB}}\nabla x_{\mathrm{A}}) + R_{\mathrm{A}} \tag{3.21}$$

or in mass units

$$\frac{\partial \varrho_{\mathrm{A}}}{\partial t} = -\nabla\cdot(\varrho_{\mathrm{A}}\boldsymbol{u}) + \nabla\cdot(\varrho D_{\mathrm{AB}}\nabla w_{\mathrm{A}}) + r_{\mathrm{A}} \tag{3.22}$$

where w_{A} is the mass fraction of A in the mixture. These equations are quite general within the restrictions of Fick's first law. They do not assume constancy of total molar concentration or of diffusivities.

Three special cases are of particular interest [63].

Case 1. If both the mixture mass density ϱ and D_{AB} are constant, then the following result is obtained:

$$\frac{\partial \varrho_{\mathrm{A}}}{\partial t} = -\boldsymbol{u}\cdot\nabla\varrho_{\mathrm{A}} + \varrho D_{\mathrm{AB}}\nabla^2 w_{\mathrm{A}} + r_{\mathrm{A}} \tag{3.23}$$

which may be divided by molecular mass to yield a molar accounting equation

$$\frac{\partial c_{\mathrm{A}}}{\partial t} = -\boldsymbol{u}\cdot\nabla c_{\mathrm{A}} + cD_{\mathrm{AB}}\nabla^2 x_{\mathrm{A}} + R_{\mathrm{A}} \tag{3.24}$$

This result, which assumes that the divergence of the velocity is zero (the mixture mass density is constant), is appropriate for dilute liquid solutions at constant temperature and pressure. This particular form of the equation is also expressed in terms of mixed quantities; the concentration is given in terms of molar concentration, but the mixture average velocity is the mass average velocity.

Case 2. If the mixture total molar concentration c and D_{AB} are constant, the mixture continuity equation gives

$$\nabla\cdot\tilde{\boldsymbol{u}} = \frac{R_{\mathrm{A}} + R_{\mathrm{B}}}{c} \tag{3.25}$$

and the accounting equation for species A becomes

$$\frac{\partial c_{\mathrm{A}}}{\partial t} = -\tilde{\boldsymbol{u}}\cdot\nabla c_{\mathrm{A}} + c D_{\mathrm{AB}} \nabla^2 x_{\mathrm{A}} + R_{\mathrm{A}}\left(1 - x_{\mathrm{A}}\right) - x_{\mathrm{A}} R_{\mathrm{B}} \tag{3.26}$$

This result is appropriate for low-density gases at constant temperature and pressure.

Case 3. Furthermore, if no chemical reaction occurs, if (for liquids) the mixture mass density is constant, and if the mixture mass average velocity is zero or (for gases) the mixture molar concentration is constant and the mixture molar average velocity is zero (i.e., no bulk motion of the fluid occurs), then Equations (3.25) and (3.26) reduce to the form

$$\frac{\partial c_{\mathrm{A}}}{\partial t} = D_{\mathrm{AB}} \nabla^2 c_{\mathrm{A}} \tag{3.27}$$

This result is known as *Fick's second law of diffusion* and states that the accumulation of species A at a point in space results solely from the diffusion flux of species A; no convective transfer occurs, nor does any generation or consumption as a result of chemical reactions. This result is also called the diffusion equation and is identical to the partial differential equations (1) that are obtained for heat conduction in solids of constant heat capacity, thermal conductivity, and density; and (2) that result from the equation of motion for a fluid of constant density and viscosity near a wall suddenly set in motion. This type of agreement between different modes of transport provides the basis for analogous correlations between certain heat-, mass-, and momentum-transfer problems. That is, the same mathematical equations can govern these three different modes of transport. Then, with analogous geometrical systems and boundary conditions, exactly the same mathematical problem exists, resulting in indentical mathematical solutions. Consequently, the results of many diffusion problems are available in the heat-transfer literature, e.g., [67], [69].

3.1.5. Diffusivities of Gases and Liquids

Gases. The diffusivities of gases can be estimated quite well by using molecular theory (similar to viscosity). Diffusion calculations are not complicated by multiple internal energy modes as in the case of thermal conductivity, although the presence of two different molecules must be taken into account. Accordingly, the Chapman – Enskog equation for diffusivity (in square centimeters per second) in a binary mixture of gases is

$$D_{\mathrm{AB}} = 1.883 \times 10^{-3} T^{3/2} \cdot \frac{\left[\left(M_{r,\mathrm{A}} + M_{r,\mathrm{B}}\right)/M_{r,\mathrm{A}} M_{r,\mathrm{B}}\right]^{1/2}}{p \sigma_{\mathrm{AB}}^2 \Omega_{\mathrm{D}}} \tag{3.28}$$

where p is in bar, σ_{AB} is in ångströms, and T is in kelvin. This result incorporates a molecular potential energy function such as the Lennard – Jones 12 – 6 function, which may be used for mixtures of nonpolar species. For a binary mixture the characteristic separation distance σ_{AB} is an arithmetic average of the values for the two pure species. Also, the collision integral Ω_{D} is a function of $kT/\varepsilon_{\mathrm{AB}}$, where $\varepsilon_{\mathrm{AB}}$ is the geometric mean of the values for the two species. As for viscosity, values of the collision integral are tabulated or may be calculated by using a relation given by NEUFELD et al. [70] and presented by REID et al. [68]. However, the values for calculating diffusivities are different from those used for viscosity. For binary mixtures with at least one polar species, a different potential function may be used (along with its own values for the molecular parameters) to calculate a different collision integral. Alternatively, the Lennard – Jones collision integral may be modified by an additive contribution due to molecular polarity as manifested by the dipole moment [71].

Diffusivities for low-density gases can also be estimated by using empirical relations built around the basic form of the Chapman – Enskog relation. The relation of FULLER et al. [72] is applicable to both polar and nonpolar molecules and is similar to a group contribution method wherein the different structural units of a molecule are assigned specific diffusion volume increments. The total diffusion volume for each molecule is calculated and used to determine the binary diffusivity. Finally, corresponding states methods also are available for estimating gas diffusivities.

Liquids. Theories for estimating liquid diffusivities are not nearly as successful as for gases because of the much smaller separation distances between molecules, which result in multibody interactions and magnified interaction forces. The Wilke – Chang method is probably the most widely used relation for estimating the diffusivity (in square centimeters per second) of a solute A at infinite dilution in a solvent B [73]:

$$D^0_{\mathrm{AB}} = 7.4 \times 10^{-11} \frac{\left(\varphi M_{r,\mathrm{B}}\right)^{1/2} T}{\eta_{\mathrm{B}} V_{\mathrm{A}}^{0.6}} \tag{3.29}$$

Accordingly, $M_{\mathrm{r,B}}$ is the molecular mass of the solvent; η_{B} the solvent viscosity, in pascal seconds, and D^0_{AB} the diffusivity when the solute is at infinite dilution in the solvent. Also, an association factor φ accounts for the degree of association exhibited by the solvent. It is 1.0 for nonassociated solvents and increases with association: 1.5 for ethanol, 1.9 for methanol, and 2.6 for water. Other methods are available and are summarized and compared in [68]; see also → Estimation of Physical Properties, Chap. 6.3.2.

3.1.6. Theoretical Foundations of Steady-State Measurement of Diffusion

To understand the techniques for measurement of diffusion, two situations are addressed with respect to gas-phase diffusion that occur as idealized limiting cases. The first is equimolar counterdiffusion (ECD) in which the two species of a binary mixture have exactly equal and opposite molar fluxes. The second is the diffusion of one species A through a stagnant film of a second species B (ATSB); the flux of one species is not zero whereas the other is exactly zero. Both situations are considered to be occurring at steady state. Since they are gas-phase diffusion problems at constant pressure the equations are given in molar form. In each case analysis involves using a continuity equation for each species (Eq. 3.3), a diffusion flux equation giving the total flux of a species as the sum of that due to the mixture molar average flux plus that due to diffusion (Eq. 3.14), and finally Fick's first law of diffusion (Eq. 3.20).

Equimolar Counterdiffusion. The term equimolar counterdiffusion means that the two species in the binary mixture diffuse at exactly the same molar flow rates per unit area but in opposite directions. A diffusion cell schematic that approximates ECD is depicted in Figure 21.

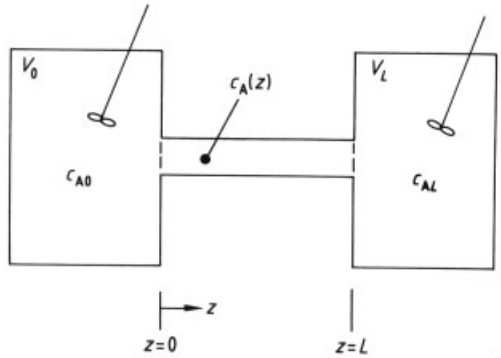

Figure 21. Schematic of an equimolar counterdiffusion (ECD) cell

This is a one-dimensional problem for which the mixture molar average velocity (Eq. 3.11) is zero; hence, the flux of either species occurs only because of its diffusion. If $\boldsymbol{N}_{\mathrm{A}}$ is defined as the total molar flux of species A, from Equation (3.14)

$$\boldsymbol{N}_{\mathrm{A}} = c_{\mathrm{A}} \boldsymbol{u}_{\mathrm{A}} = c_{\mathrm{A}} \tilde{\boldsymbol{u}} + \tilde{\boldsymbol{J}}_{\mathrm{A}} = \tilde{\boldsymbol{J}}_{\mathrm{A}} \tag{3.30}$$

or in one dimension using Fick's first law

$$N_{\mathrm{A}z} = -cD_{\mathrm{AB}} \frac{\mathrm{d}y_{\mathrm{A}}}{\mathrm{d}z} \tag{3.31}$$

Now, an accounting of species A (the continuity equation, Eq. 3.3) with no reactions and at steady state gives

$$\nabla \cdot \boldsymbol{N}_{\mathrm{A}} = 0 \tag{3.32}$$

which in one-dimensional form is

$$\frac{\mathrm{d}N_{\mathrm{A}z}}{\mathrm{d}z} = 0 \tag{3.33}$$

and from which the molar flux of species A (and similarly of species B) is seen to be invariant with respect to position in the direction of flow.

Consequently, Equation (3.31) can be integrated directly (for constant total concentration and diffusivity) to give a linear relation for the concentration of species A (and hence also of species B) with respect to position in the direction of flow

$$y_A(z) = y_{A0} + \left(\frac{y_{AL} - y_{A0}}{L}\right) z \tag{3.34}$$

where y_{A0} and y_{AL} are the mole fractions of A at $z = 0$ and $z = L$, respectively. The molar flux in terms of diffusivity and concentrations at the ends of the diffusion tube is

$$N_{Az} = \frac{-cD_{AB}}{L}(y_{AL} - y_{A0}) \tag{3.35}$$

This result then can be used to measure diffusivities if the tube-end concentrations are known as a function of time. The above derivation assumes a steady state, i.e., the two end-cell concentrations do not change with time. The result can still be used if the changes in concentration are slow compared to the establishment of the concentration profiles in the diffusion cell. By using this pseudosteady-state assumption the changes in concentration at the two cell ends can be related to the flux through the cell by writing a mass accounting balance for species A in each cell. Accordingly,

$$\frac{d}{dt}(c_{A0}V_0) = N_{Az}A_c = -\frac{A_c D_{AB}}{L}(c_{AL} - c_{A0}) \tag{3.36}$$

where A_c is the cross-sectional area through which diffusion occurs and

$$\frac{d}{dt}(c_{AL}V_L) = -N_{Az}A_c = \frac{A_c D_{AB}}{L}(c_{AL} - c_{A0}) \tag{3.37}$$

from which

$$\frac{d}{dt}(c_{A0} - c_{AL}) = -\frac{A_c D_{AB}}{L}(c_{AL} - c_{A0})\left[\frac{1}{V_0} + \frac{1}{V_L}\right] \tag{3.38}$$

so that

$$[c_{AL}(t) - c_{A0}(t)] = [c_{AL}(t_0) - c_{A0}(t_0)] \cdot \exp\left\{-\frac{A_c D_{AB}}{L}\left(\frac{1}{V_0} + \frac{1}{V_L}\right)(t - t_0)\right\} \tag{3.39}$$

Measuring the concentration difference between cells as a function of time gives D_{AB}. The diffusion cell itself can consist of a glass frit or even filter paper. In the latter case the cell length is ill-defined because of tortuosity and the coefficient $A_c(1/V_0 + 1/V_L)/L$ is determined by calibration with a material of known diffusivity [65].

ATSB Diffusion. For the diffusion of species A through a stagnant film of B (ATSB), the situation is again taken to be constant total pressure and steady state. Diffusion occurs through a tube of uniform diameter and length L. In this case, however, the motion of one of the species is constrained to be zero as in the evaporation of a liquid A from a tube open at one end (Fig. 22). The second gas-phase component is not soluble in A (or at least A is saturated with B), so its flux at the interface of liquid A with the gas phase is zero. The liquid evaporates at the interface and diffuses to the open end of the tube where its concentration is maintained at a reduced level by a steady flow of fresh gas phase across the open end of the tube.

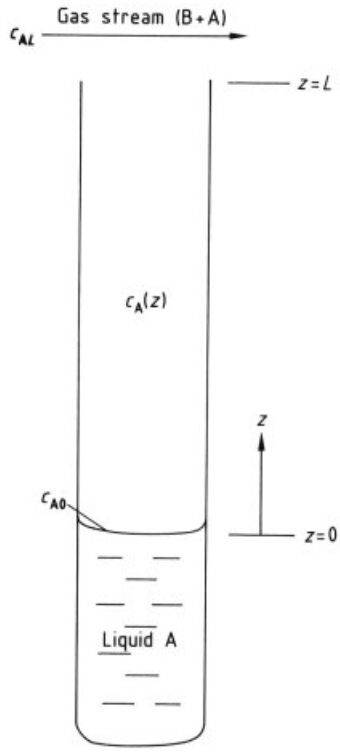

Figure 22. Schematic of "A through stagnant B" (ATSB) diffusion process

For steady state with no reactions in the gas phase, the continuity of species A in the gas phase requires (as in the equimolar counterdiffusion problem) that

$$0 = -\nabla \cdot \boldsymbol{N}_{\mathrm{A}} = -\frac{\mathrm{d}N_{\mathrm{A}z}}{\mathrm{d}z} \tag{3.40}$$

Again, the molar flux of A along the length of the tube is constant. Likewise, a similar result is obtained for species B. Because liquid A is saturated with B, the flux of species B into the liquid at the interface must be zero. Then, because this flux is constant along the entire length of the tube, it must be zero everywhere; species B is stagnant. Now, by writing $\boldsymbol{N}_{\mathrm{A}}(=c_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}})$ and $\boldsymbol{N}_{\mathrm{B}}(=c_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}})$ as the molar fluxes of species A and B, respectively, the combined molar flux for the mixture is given by (see Eq. 3.11)

$$c\tilde{\boldsymbol{u}} = c_{\mathrm{A}}\boldsymbol{u}_{\mathrm{A}} + c_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}} = \boldsymbol{N}_{\mathrm{A}} + \boldsymbol{N}_{\mathrm{B}} \tag{3.41}$$

which, combined with Equation (3.14), yields

$$\boldsymbol{N}_{\mathrm{A}} = y_{\mathrm{A}}\left(\boldsymbol{N}_{\mathrm{A}} + \boldsymbol{N}_{\mathrm{B}}\right) + \tilde{\boldsymbol{J}}_{\mathrm{A}} \tag{3.42}$$

i.e., the total flux of A is that fraction due to the total mixture average flow plus the flow due to the diffusion of species A. With species B stagnant, $\boldsymbol{N}_{\mathrm{B}}=0$ and

$$\boldsymbol{N}_{\mathrm{A}} = \frac{\tilde{\boldsymbol{J}}_{\mathrm{A}}}{(1 - y_{\mathrm{A}})} \tag{3.43}$$

which in one-dimensional form is

$$N_{\mathrm{A}z}\mathrm{d}z = -cD_{\mathrm{AB}}\frac{\mathrm{d}y_{\mathrm{A}}}{(1 - y_{\mathrm{A}})} \tag{3.44}$$

Because of the constancy of N_{Az} along the length of the tube, this result can be integrated directly to give either a result for the molar flux of A (in terms of tube length, total pressure, diffusivity, and concentrations at the end of the tube)

$$N_{\mathrm{A}z} = \frac{cD_{\mathrm{AB}}}{L}\ln\left(\frac{1 - y_{\mathrm{A}L}}{1 - y_{\mathrm{A}0}}\right) \tag{3.45}$$

or a result for the concentration profile of species A along the length of the tube

$$y_{\mathrm{A}}(z) = 1 - (1 - y_{\mathrm{A}0})\left(\frac{1 - y_{\mathrm{A}L}}{1 - y_{\mathrm{A}0}}\right)^{z/L} \tag{3.46}$$

Note that, contrary to equimolar counterdiffusion, the concentration profile is nonlinear even though the flux of species A is still constant along the tube length.

Comparison of ATSB and ECD Results. Comparing the ATSB and ECD results shows that for the same mixture concentration c, diffusivity D_{AB}, and driving force for diffusion $\mathrm{d}y_{\mathrm{A}}/\mathrm{d}z$, the flux for ATSB is greater than that for ECD due to the fact that the mole fraction of species A is less than unity and hence the driving force is divided by $(1 - y_{\mathrm{A}})$. This enhancement is perhaps best understood by looking at Equation (3.42) for the total molar flux of A in terms of the bulk flow of the mixture and the diffusion flux. The bulk flow of the mixture is not zero for ATSB, whereas it is zero for ECD. Enhancement occurs in the flux of species A out of the tube as a result of this nonzero bulk flow. Note also that even though B is stagnant, its diffusion flux, through Fick's law, is nonzero because its mole fraction gradient is nonzero. With species B, however, the diffusion flux down the tube toward the liquid in the bottom of the tube is exactly counterbalanced by its flow up the tube because of the bulk flow of the gas mixture. The net effect is that species B is stagnant ($N_{\mathrm{B}}=0$) even though it is in fact diffusing.

In principle, this experimental situation provides another method for measuring gas-phase diffusivities. In terms of changes in the amount of evaporating liquid in the tube,

$$N_{\mathrm{A}z}(t) = \frac{\varrho}{M_{\mathrm{r}}}\frac{\mathrm{d}L}{\mathrm{d}t} \tag{3.47}$$

so that Equation (3.45) becomes

$$\frac{\varrho}{M_{\mathrm{r}}}\frac{\mathrm{d}L}{\mathrm{d}t} = \frac{cD_{\mathrm{AB}}}{L}\ln\left(\frac{1 - y_{\mathrm{A}L}}{1 - y_{\mathrm{A}0}}\right) \tag{3.48}$$

which can be integrated to give

$$L^2 = \frac{2cD_{\mathrm{AB}}M_r}{\varrho}\left[\ln\left(\frac{1 - y_{\mathrm{A}L}}{1 - y_{\mathrm{A}0}}\right)\right](t - t_0) + L_0^2 \tag{3.49}$$

Measuring L as a function of time (by weight) can then be used to determine D_{AB}. This result, like the ECD method, rests upon the pseudosteady-state assumption, in this case that the concentration profile and the flux adjust rapidly to changes in L. Further complications in this technique involve maintaining constant temperature in the presence of evaporation and convection due to differences in the molecular mass of the two gases [63].

3.1.7. Diffusion with Homogeneous Chemical Reaction

A situation is now considered that involves chemical reaction in addition to diffusion: the transfer of a component from a gas phase to a liquid film followed by diffusion away from the gas – liquid interface and reaction within the film. At a distance δ from the gas – liquid interface, the film meets an impermeable solid boundary (Fig. 23). As a boundary condition it is assumed that diffusion of the gas to the liquid film is not limiting so that the concentration of the gaseous material in the liquid film at the gas – liquid interface is that established by equilibrium with the bulk gas-phase concentration. Species A then diffuses away from the gas – liquid interface in an attempt to increase its concentration within the liquid film. However, the chemical reaction serves to deplete the concentration and a nonlinear steady-state concentration profile is achieved. At the solid – liquid interface, the concentration profile must be flat; because of the impermeability of the solid surface, the flux at that surface is zero.

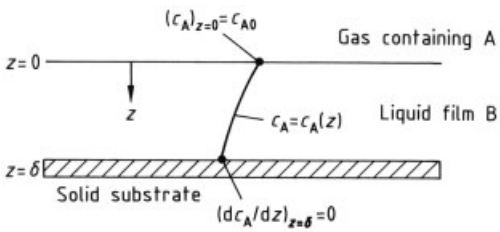

Figure 23. Schematic of diffusion and chemical reaction in a liquid film

The objective of the analysis is to obtain a relation for the molar flow rate of A within the film and the concentration profile of A across the film. For reactions of species A with the liquid (species B), consider a situation with the reaction given by

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} \tag{3.50}$$

which is characterized by a rate constant k_1 such that the rate of production of species A is given by the kinetic first-order equation

$$R_\mathrm{A} = -k_1 c_\mathrm{A}$$
$$(R_\mathrm{A} < 0 \text{ means that A is being consumed}) \tag{3.51}$$

This reaction occurs homogeneously throughout the liquid film in the sense that it takes place everywhere in the film that species A and B exist together. Reaction rate varies with position because of variation of the concentration of A with position z through the film.

As usual, we begin with the equation of continuity for A which is

$$\frac{\partial c_\mathrm{A}}{\partial t} = -\nabla \cdot \boldsymbol{N}_\mathrm{A} + R_\mathrm{A} \tag{3.52}$$

For the steady-state condition and for the rate of production of A given by chemical kinetics, this equation becomes

$$\frac{\mathrm{d}N_{\mathrm{A}z}}{\mathrm{d}z} = -k_1 c_\mathrm{A}(z) \tag{3.53}$$

This situation differs in a fundamental way from the previously described situations in that the flux of A in the diffusion direction is not constant because of the depletion of A by the chemical reaction. The concentration of A is a function of position z so that the result cannot be integrated to obtain an immediate conclusion about the flux of A. Instead, the flux equation of A is used in terms of bulk flow in the solution and diffusion:

$$\boldsymbol{N}_\mathrm{A} = y_\mathrm{A}(\boldsymbol{N}_\mathrm{A} + \boldsymbol{N}_\mathrm{B} + \boldsymbol{N}_\mathrm{C}) - cD_\mathrm{AB}\nabla y_\mathrm{A} \tag{3.54}$$

Because species A reacts as it diffuses into the thin film, a reasonable assumption (depending on the reaction kinetics versus the diffusion rate) may be that the concentration of A in the liquid film is everywhere small so that the bulk flow (convection) term that contributes to the total flux of species A can be neglected relative to the diffusion flux term. In this case and for one dimension, the equation becomes

$$N_{\mathrm{A}z} = -cD_\mathrm{AB}\frac{\mathrm{d}y_\mathrm{A}}{\mathrm{d}z} \tag{3.55}$$

Combining this with the continuity equation (Eq. 3.53) gives a second-order ordinary differential equation for the concentration of species A

$$\frac{\mathrm{d}^2 y_\mathrm{A}}{\mathrm{d}z^2} - \frac{k_1}{D_\mathrm{AB}} y_\mathrm{A} = 0 \tag{3.56}$$

This equation is subject to the boundary conditions at the gas – liquid film interface where the concentration in the liquid film is established by thermodynamic equilibrium ($y_\mathrm{A} = y_{\mathrm{A}0}$) and at the solid surface ($z = \delta$) where $N_{\mathrm{A}z} = 0$ and

$dy_A/dz = 0$. Imposing these conditions allows integration of the equation to give

$$y_A(z) = y_{A0}\left[\cosh\left(z\sqrt{k_1/D_{AB}}\right) - \tanh\left(\delta\sqrt{k_1/D_{AB}}\right)\sinh\left(z\sqrt{k_1/D_{AB}}\right)\right] \quad (3.57)$$

The flux of A is a function of position and is given by

$$N_{Az}(z) = c y_{A0}\sqrt{k_1 D_{AB}}\cdot\left[-\sinh\left(z\sqrt{k_1/D_{AB}}\right) + \tanh\left(\delta\sqrt{k_1/D_{AB}}\right)\cosh\left(z\sqrt{k_1/D_{AB}}\right)\right] \quad (3.58)$$

Note that the concentration profile through the liquid film is exponential in form through the cosh and sinh functions and that the degree of decay depends on the ratio (relative magnitude) of the reaction rate to the diffusion rate. The larger the reaction rate, the more steeply the profile falls as the distance into the film away from the gas interface increases. Conversely, the faster the diffusion (i.e., the higher the diffusion rate relative to the reaction rate), the more slowly the (steady-state) concentration falls or decreases with increasing distance from the interface.

A similar problem for an infinitely thick liquid film is approached in a similar way except that instead of the no-flux boundary condition at the solid surface the condition $c_A \to 0$ as $z \to \infty$ is used. Such a solution is appropriate even for liquid films of finite thickness if the reaction rate is large enough compared to the diffusion rate so that the diffusing species A in fact never "sees" the solid surface, i.e., it is depleted by reaction before it reaches the solid surface.

3.1.8. Diffusion with Heterogeneous Reaction

The next problem addresses diffusion through a gas film to a surface at which a reaction occurs and from which the reaction product must diffuse back through the gas phase. This is the situation in many catalytic conversion reactions where a catalyst pellet may be considered to be surrounded by a thin stagnant film through which a species must diffuse. Because the film is stagnant, diffusion rather than reaction kinetics may be the controlling phenomenon.

The situation is depicted schematically in Figure 24 for a spherical catalyst particle of radius R surrounded by a boundary layer gaseous film of thickness δ. Reactant A diffuses through the film to the surface where the reaction $A + A \to A_2$ is assumed to occur very quickly compared to the diffusion processes. As a result, at the catalyst surface the concentration of reacting species is assumed to be zero; the reacting species is depleted by reaction as quickly as it is delivered to the surface by diffusion. This is a binary system with A and A_2 as the two components. Because the reaction occurs at the solid surface rather than in the gas phase, it is not relevant to the continuity equation of species A within the gas phase. The role of the reaction stoichiometry and kinetics is to provide a boundary condition on species fluxes at the catalyst surface.

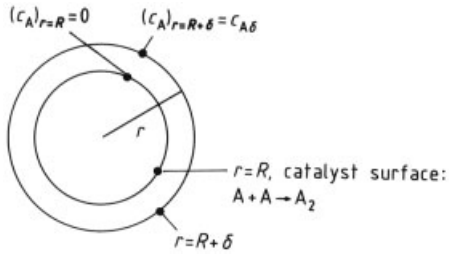

Figure 24. Schematic of diffusion and heterogeneous chemical reaction at a catalyst surface

As for the preceding problems, the objective is to establish quantitative relations for the flux of reactant A through the stagnant thin film and for the concentration profile of A across the film. As before, the continuity equation for reacting species A is employed:

$$\frac{\partial c_A}{\partial t} = -\nabla\cdot\boldsymbol{N}_A + R_A \quad (3.59)$$

For steady state with no homogeneous reaction, this gives

$$\nabla\cdot\boldsymbol{N}_A = 0 \quad (3.60)$$

which in spherical coordinates is

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 N_{Ar}\right) = 0 \quad (3.61)$$

which is integrated to give

$$r^2 N_{Ar} = \text{constant} \quad (3.62)$$

In this spherical geometry, the flux of a species is not constant but is proportional to $1/r^2$. Now because at any position r the spherical surface area is proportional to r^2, the molar flow rate of A rather than the flux is constant, independent of position, according to

$$r^2 N_{\mathrm{A}r} = \text{constant} = \frac{r^2 N_{\mathrm{A}}}{4\pi r^2} = \frac{N_{\mathrm{A}}}{4\pi} \tag{3.63}$$

Here N_{A} is the number of moles of species A per unit time passing through the entire spherical shell at radius r. This molar flow rate is a constant, independent of radial position. As before, the result for the molar flux of species A is written in terms of the convective bulk flow and diffusion:

$$\boldsymbol{N}_{\mathrm{A}} = y_{\mathrm{A}} = \left(\boldsymbol{N}_{\mathrm{A}} + \boldsymbol{N}_{\mathrm{A}_2}\right) - cD_{\mathrm{AB}} \nabla y_{\mathrm{A}} \tag{3.64}$$

which, in one-dimensional form, is

$$N_{\mathrm{A}r} = y_{\mathrm{A}} \left(N_{\mathrm{A}r} + N_{\mathrm{A}_2 r}\right) - cD_{\mathrm{AB}} \frac{\mathrm{d}y_{\mathrm{A}}}{\mathrm{d}r} \tag{3.65}$$

Because of the reaction stoichiometry, the molar flow rate of species A can be related to that of product A_2 according to

$$N_{\mathrm{A}} = -2N_{\mathrm{A}_2} \tag{3.66}$$

These are molar flow rates integrated over the entire surface of a spherical shell and, as shown above, because of continuity they are independent of radial position. The fluxes of each of the species can be written in terms of these molar flow rates so that Equation (3.65) becomes

$$\frac{N_{\mathrm{A}}}{r^2} = y_{\mathrm{A}} \left(\frac{N_{\mathrm{A}}}{r^2} + \frac{N_{\mathrm{A}_2}}{r^2}\right) - 4\pi c D_{\mathrm{AB}} \frac{\mathrm{d}y_{\mathrm{A}}}{\mathrm{d}r} \tag{3.67}$$

which, because of the relation between the molar flow rates of A and A_2, gives

$$N_{\mathrm{A}} \frac{\mathrm{d}r}{r^2} = \frac{4\pi c D_{\mathrm{AB}}\, \mathrm{d}y_{\mathrm{A}}}{\left(1 - \frac{1}{2} y_{\mathrm{A}}\right)} \tag{3.68}$$

As the molar flow rate is a constant, this result can be integrated directly to give a relation for the concentration profile through the film. In doing so, the boundary conditions are that the concentration of reactant A is zero at the surface of the pellet because it is instantaneously depleted by the reaction, and the concentration at the outer extent of the film (i.e., at $r = R + \delta$) is a known value (the concentration in the bulk gas phase). The resulting concentration profile is

$$y_{\mathrm{A}}(r) = 2\left[1 - \left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right)^{\left(\frac{\gamma}{R+\gamma}\right)\left(\frac{R+\delta}{\delta}\right)}\right] \tag{3.69}$$

where $\gamma \equiv r - R$ and

$$N_{\mathrm{A}} = 8\pi c D_{\mathrm{AB}} \frac{R(R+\delta)}{\delta} \ln\left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right) \tag{3.70}$$

Now this result is written in terms of an unknown film thickness δ. It can be evaluated, however, for two limiting cases to obtain limiting or boundary concentration profiles and molar flow rates. If gas flow across the pellet is slow or essentially stagnant, then the film thickness goes to infinity and the results obtained are

$$y_{\mathrm{A}}(r) = 2\left[1 - \left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right)^{\left(1 - \frac{R}{r}\right)}\right] \tag{3.71}$$

and

$$N_{\mathrm{A}} = 8\pi c D_{\mathrm{AB}} R \ln\left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right) \tag{3.72}$$

Alternatively, if the gas velocity is high enough, the stagnant film becomes very thin ($\delta \ll R$) and results obtained are

$$y_{\mathrm{A}}(r) = 2\left[1 - \left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right)^{\left(\frac{\gamma}{\delta}\right)}\right] \tag{3.73}$$

and

$$N_{\mathrm{A}} = 8\pi c D_{\mathrm{AB}} \frac{R^2}{\delta} \ln\left(1 - \frac{1}{2} y_{\mathrm{A}\delta}\right) \tag{3.74}$$

In these latter equations, the film thickness remains as a parameter, whereas in the stagnant film situation the velocity profile and molar flow rate can be calculated directly; as the gas film becomes very thick, the molar flow rate and concentration profile become independent of the film thickness. At the other extreme, for a very thin film, the molar flow rate becomes inversely proportional to the film thickness; a doubling of film thickness halves the molar flow rate.

3.1.9. Perspective

The aforementioned results are useful in the context of this article; not so much for the specific problems addressed as for the types of approaches that are made to solving diffusion problems. In these problems, steady-state situations have involved either thin films or infinitely thick films, and slowly changing processes have been observed which can be treated in a pseudosteady-state manner by first solving the steady-state problem and then integrating over time to obtain the results as a function of time. Flat and spherical geometries have also been observed, although curved geometries, mathematically, look exactly like flat geometries when the film thickness is small in relation to the radius of curvature. Therefore, flat geometry calculations take on an importance above and beyond their actual application to specific physical problems. Finally, both homogeneous and heterogeneous chemical reactions have been addressed; homogeneous chemical reactions occur throughout the phase through which the diffusion or motion of molecules is occurring, whereas heterogeneous reactions occur at a boundary or surface associated with the diffusion process. As a result, homogeneous reaction kinetics appear directly in the continuity equations for the reacting species within the diffusion phase, whereas heterogeneous reaction kinetics play a role in the fluxes that occur at interfaces or boundaries of the diffusion region.

3.2. Convective Mass Transport

In the previous examples, motion of the individual species resulted from the diffusive fluxes that occur because of concentration gradients. In this section, examples in which species transport arises from bulk flow (or convection), as well as diffusion, are considered.

3.2.1. Gas Absorption in a Falling Film with Reaction

A problem encountered in many industrial scrubbers, fixed-bed gas – liquid reactors, and other mass-transfer devices, is the exchange of a gas species A between a gas mixture and a liquid film that moves on a solid surface. A simple example of such a situation is shown in Figure 25, where the gas species A is exchanged between a bulk gas phase and a falling liquid film of component B. This problem is analyzed first as steady-state *simple absorption with no chemical reaction*. Then, this solution is used to analyze the more general case in which *simultaneous absorption and reaction* occur.

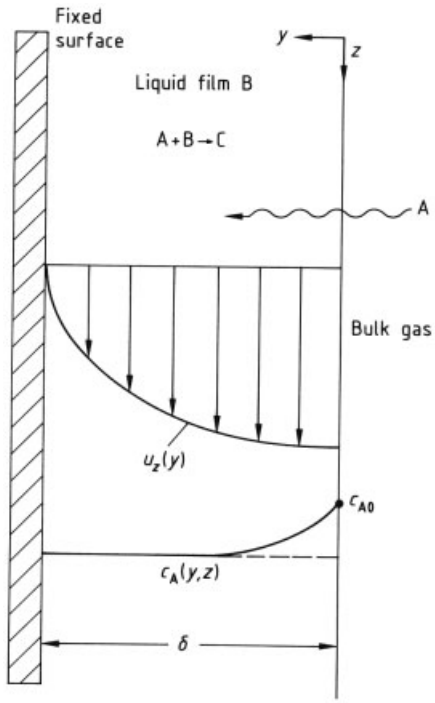

Figure 25. Absorption and reaction in a falling liquid film

In the analysis, A is assumed to be only slightly soluble in Newtonian liquid B, so that the viscosity and density of the liquid film are not changed appreciably. Under such conditions, the film velocity profile is given by (end effects neglected)

$$u_z = u_{\max}\left[1 - \left(\frac{y}{\delta}\right)^2\right] \tag{3.75}$$

The concentration of A in the film varies with both y and z, and the species mass balance (Eq. 3.24) can be written as

$$u_z \frac{\partial c_A}{\partial z} = D_{AB}\left(\frac{\partial^2 c_A}{\partial y^2} + \frac{\partial^2 c_A}{\partial z^2}\right) \tag{3.76}$$

with the boundary conditions

$$\text{at} \quad z = 0, \quad c_{\mathrm{A}} = 0 \tag{3.77}$$

$$\text{at} \quad z \to \infty, \quad c_{\mathrm{A}} \to c_{\mathrm{A0}} \tag{3.78}$$

$$\text{at} \quad y = 0, \quad c_{\mathrm{A}} = c_{\mathrm{A0}} \tag{3.79}$$

$$\text{at} \quad y = \delta, \quad \frac{\partial c_{\mathrm{A}}}{\partial y} = 0 \tag{3.80}$$

The first boundary condition expresses the fact that pure B enters at $z = 0$; the second, that the film is saturated with A after a sufficiently long exposure distance. The third condition gives the concentration c_{A0} at the gas – liquid boundary. Generally, unless significant resistance to mass transfer occurs on the gas side of the interface, the concentration c_{A0} is the concentration in the liquid that is in equilibrium with the gas. The fourth condition simply indicates that no mass flux exists normal to the solid wall; hence, the concentration gradient $\partial c_{\mathrm{A}}/\partial y$ is zero at the wall.

By assuming that over some film length $z \sim \delta$ the gas penetrates only a fraction of the distance into the film (e.g., $y \sim \varepsilon\,\delta$), the y and z variables can be scaled as

$$y \equiv \varepsilon\,\delta\,y^* \tag{3.81a}$$

$$z \equiv \delta\,z^* \tag{3.81b}$$

where $y*$ and $z*$ are dimensionless quantities of order 1. Further, if c_{A} and u_z are scaled by

$$c_{\mathrm{A}} = c_{\mathrm{A0}}\,c_{\mathrm{A}}^* \tag{3.81c}$$

and

$$u_z = u_{\max} u_z^* \tag{3.81d}$$

Equation (3.76) can be written as

$$\left(1 - \varepsilon^2\,y^{*2}\right)\frac{\partial c_{\mathrm{A}}^*}{\partial z^*} = \frac{1}{Pe_\delta}\left(\frac{1}{\varepsilon^2}\frac{\partial^2 c_{\mathrm{A}}^*}{\partial y^{*2}} + \frac{\partial^2 c_{\mathrm{A}}^*}{\partial z^{*2}}\right) \tag{3.82}$$

where

$$Pe_\delta \equiv \frac{u_{\max}\delta}{D_{\mathrm{AB}}} \tag{3.83}$$

is the mass-transfer Péclet number.

If convection is important, the left-hand side of Equation (3.82) must be of the same order as the largest diffusion term on the right-hand side (the y diffusion term). In particular, this requires that

$$\varepsilon \equiv \frac{1}{Pe_\delta^{1/2}} \tag{3.84}$$

Equation (3.82) can then be written as

$$\left(1 - \frac{1}{Pe_\delta}\,y^{*2}\right)\frac{\partial c_{\mathrm{A}}^*}{\partial z^*} = \frac{\partial^2 c_{\mathrm{A}}^*}{\partial y^{*2}} + \frac{1}{Pe_\delta}\frac{\partial^2 c_{\mathrm{A}}^*}{\partial z^{*2}} \tag{3.85}$$

Equation (3.84) indicates that short penetration distances ($\varepsilon \ll 1$) are associated with large Péclet numbers. For the specific case in which $u_{\max} \approx 0.1$ cm/s, $\delta \approx 0.1$ cm, and $D_{\mathrm{AB}} \approx 10^{-5}$ cm^2/s, the value of Pe_δ is 10^3 (or $\varepsilon \approx 0.03$). Such high Péclet number, short penetration distance conditions are typical of those encountered in many gas – liquid contact processes. Under such conditions, Equation (3.82) can be reduced to

$$\frac{\partial c_{\mathrm{A}}^*}{\partial z^*} = \frac{\partial^2 c_{\mathrm{A}}^*}{\partial y^{*2}} \tag{3.86}$$

with the boundary conditions

$$\text{at} \quad z^* = 0, \quad c_{\mathrm{A}}^* = 0 \tag{3.87}$$

$$\text{at} \quad y^* = 0, \quad c_{\mathrm{A}}^* = 1 \tag{3.88}$$

$$\text{at} \quad y^* \to \infty, \quad c_{\mathrm{A}}^* \to 0 \tag{3.89}$$

The solution of Equation (3.86) subject to these boundary conditions is [63, p. 539]

$$c_{\mathrm{A}}^* = \frac{c_{\mathrm{A}}}{c_{\mathrm{A0}}} = 1 - \frac{2}{\sqrt{\pi}}\int_0^{\frac{y}{\sqrt{4D_{\mathrm{AB}}z/u_{\max}}}} \mathrm{e}^{-\zeta^2}\,\mathrm{d}\zeta$$

$$= 1 - \mathrm{erf}\left(\frac{y}{\sqrt{4\mathrm{D}_{\mathrm{AB}}z/u_{\max}}}\right) \tag{3.90}$$

or

$$\frac{c_{\mathrm{A}}}{c_{\mathrm{A0}}} = \mathrm{erfc}\left(\frac{y}{\sqrt{4\mathrm{D}_{\mathrm{AB}}z/u_{\max}}}\right) \tag{3.91}$$

where erf and erfc are the error and complementary error functions, respectively. For a contact film of length L and width W, the rate of mass transfer (number of moles of A per unit time) to the film is given by

$$\mathfrak{W}_{\mathrm{A}} = \tag{3.92}$$

$$\int_0^L N_{\mathrm{A}y}\big|_{y=0}\,W\mathrm{d}z = \int_0^L \left(-D_{\mathrm{AB}}\frac{\partial c_{\mathrm{A}}}{\partial y}\right)\Big|_{y=0} W\mathrm{d}z$$

or, if Equation (3.91) is used,

$$\mathfrak{W}_{\mathrm{A}} = L\,W\,c_{\mathrm{A0}}\sqrt{\frac{4\,D_{\mathrm{AB}}u_{\max}}{\pi\,L}} \tag{3.93}$$

If the mass-transfer coefficient k_c is defined by

$$\mathfrak{W}_{\mathrm{A}} = k_c\,L\,W\,c_{\mathrm{A0}} \tag{3.94}$$

then

$$Sh_L \equiv \frac{k_c\,L}{D_{\mathrm{AB}}} = \frac{2}{\sqrt{\pi}}\,Re_L^{1/2}\,Sc^{1/2} = \frac{2}{\sqrt{\pi}}\,Pe_L^{1/2} \tag{3.95}$$

where Sh_L is the Sherwood number, and Re_L and Sc are the Reynolds and Schmidt numbers, respectively, which are defined by

$$Re_L \equiv \frac{u_{\max}L}{\nu}, \quad Sc \equiv \frac{\nu}{D_{\mathrm{AB}}} \tag{3.96}$$

Hence, for short penetration distances ε (or short contact times or high Péclet numbers), the dimensionless form of the mass-transfer coefficient (Sh_L) varies as the square root of the product of Re_L and Sc.

If the *chemical reaction*

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} \tag{3.97}$$

in the film is now considered and a sufficient excess of B is assumed so that a pseudo-firstorder reaction expression is valid, i.e.,

$$R_{\mathrm{A}} = k'\,c_{\mathrm{A}}c_{\mathrm{B}} \approx k'\,c_{\mathrm{B0}}c_{\mathrm{A}} = k_1\,c_{\mathrm{A}} \tag{3.98}$$

the mass balance takes the form

$$\frac{\partial c_{\mathrm{A}}^*}{\partial z^*} = \frac{\partial^2 c_{\mathrm{A}}^*}{\partial y^{*2}} - \frac{Da_\delta\,c_{\mathrm{A}}^*}{Re_\delta\,Sc} \tag{3.99}$$

where Da_δ is the Damköhler number defined by

$$Da_\delta \equiv \frac{k_1\,\delta^2}{D_{\mathrm{AB}}} \tag{3.100}$$

Here again, in writing Equation (3.99), the penetration distance is assumed to be small (i.e., $Pe_\delta \equiv Re_\delta\,Sc$ is large). Also, the boundary conditions are assumed to be the same as given by Equations (3.87) – (3.89).

The solution to this problem can be expressed in terms of the solution for the nonreacting case (Eq. 3.91) by [69, pp. 32 – 33]

$$c_{\mathrm{A}}^* = \frac{c_{\mathrm{A}}}{c_{\mathrm{A0}}} = \varphi\,(y^*, z^*)\exp\left(\frac{Da_\delta}{Re_\delta\,Sc}z^*\right) + \frac{Da_\delta}{Re_\delta\,Sc}\int_0^{z^*}\varphi\,(y^*, \zeta)\exp\left(\frac{Da_\delta}{Re_\delta\,Sc}\zeta\right)\mathrm{d}\zeta \tag{3.101}$$

where $\varphi\,(y*, z*)$ is the c_{A} solution from Equation (3.91). Using Equation (3.92) to determine the total gas exchange over a surface of length L and width W gives

$$\mathfrak{W}_{\mathrm{A}} = W\,c_{\mathrm{A0}}\frac{u_{\max}}{\sqrt{k_1/D_{\mathrm{AB}}}} \cdot \left[\left(\frac{1}{2}+q\right)\mathrm{erf}\,\sqrt{q} + \sqrt{\frac{q}{\pi}}\cdot\exp\,(-q)\right] \tag{3.102}$$

where

$$q \equiv \frac{Da_L}{Re_L\,Sc} \tag{3.103}$$

In terms of $Sh_L \equiv k_c\,L/D_{\mathrm{AB}}$, this result can be written as

$$Sh_L = Re_L^{1/2}\,Sc^{1/2} \cdot \left[\frac{(1+2q)}{2\sqrt{q}}\,\mathrm{erf}\,\sqrt{q} + \frac{1}{\sqrt{\pi}}\exp\,(-q)\right] \tag{3.104}$$

If $q \rightarrow 0$ (no chemical reaction), this equation reduces to Equation (3.95); if $q \rightarrow \infty$,

$$Sh_L = Da_L^{1/2} = (q\,Re_L\,Sc)^{1/2} \tag{3.105}$$

In Figure 26, the results are given for a full range of q values. As q increases (higher reaction rates), mass transfer is significantly enhanced (i.e., higher mass-transfer coefficients) compared to the nonreacting case ($q = 0$). Many scrubbers and other absorption processes utilize absorption with reaction to obtain the most efficient removal of dilute gases from bulk gas streams (→ Absorption).

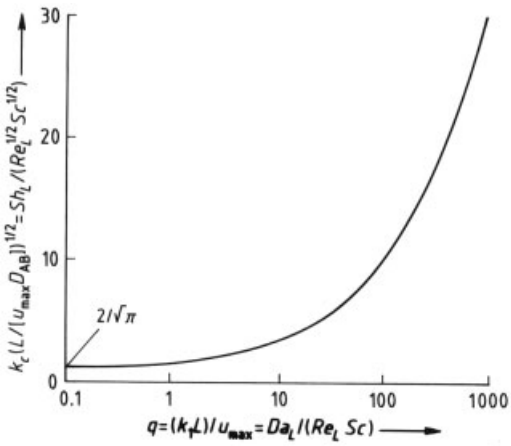

Figure 26. Enhancement of mass transfer with chemical reaction in a falling liquid film

3.2.2. Mass Transfer in Laminar and Turbulent Boundary Layers

Consider the convective mass transfer of species A from a flat or nearly flat surface into a free stream of species B. Assume that concentration changes of A are restricted to a concentration boundary layer of thickness δ_c. Conceptually, this is analogous to the problem illustrated in Figure 15, except that δ_c replaces δ_T, $c_{A\infty}$ replaces T_∞, and c_{A0} replaces T_w.

For this problem where $c_A = c_A(x, y)$, the species mass balance takes the form

$$u_x \frac{\partial c_A}{\partial x} + u_y \frac{\partial c_A}{\partial y} = D_{AB} \frac{\partial^2 c_A}{\partial y^2} \tag{3.106}$$

where the x diffusion term $[D_{AB}(\partial^2 c_A/\partial x^2)]$ has been neglected compared to the y diffusion term (right-hand side of above equation). This equation is analogous to Equation (2.239) for heat transfer in thermal boundary layers and can be integrated over the boundary layer thickness to obtain

$$\frac{d}{dx}\int_0^{\delta_c} u_x (c_A - c_{A\infty})\, dy = N_{Ay}|_{y=0}$$
$$= -D_{AB} \frac{\partial c_A}{\partial y}\Big|_{y=0} \tag{3.107}$$

The solution of this equation for $\delta_c(x)$ follows in the same way δ_T was obtained from the integral form of the energy balance (Eq. 2.240). If a cubic expression is used for c_A in terms of y and the constants a, b, c, and d are evaluated with the boundary conditions

$$\text{at} \quad y = 0, \quad c_A = c_{A0} \tag{3.108}$$

$$\text{at} \quad y = \delta_c, \quad c_A = c_{A\infty} \tag{3.109}$$

$$\text{at} \quad y = \delta_c, \quad \frac{\partial^n c_A}{\partial y^n} = 0, \quad n = 1, 2, \ldots \tag{3.110}$$

the same results can be obtained as for Equations (2.250) and (2.251), with δ_c replacing δ_T, $Sc \equiv \nu/D_{AB}$ replacing Pr, and $Sh_x \equiv k_c x/D_{AB}$ replacing Nu_x. The Blasius solutions (Eqs. 2.252 and 2.255) also have the same forms for the mass-transfer results if the same replacements are made. The mass-transfer rate (number of moles A per unit time) from the surface is then

$$\mathfrak{W}_A = \int_0^L k_c (c_{A0} - c_{A\infty})\, W dx$$
$$= k_{cL} W L (c_{A0} - c_{A\infty}) \tag{3.111}$$

where

$$k_{cL} \equiv \frac{1}{L}\int_0^L k_c dx \tag{3.112}$$

is the average mass-transfer coefficient over the plate length L.

These results are valid for the laminar part of the boundary layer. In the turbulent part ($x \geq x_{cr} = 2\times 10^5\, \nu/u_\infty$), the molar flux in the y-direction can be written as

$$\bar{N}_{Ay} = -(D_{AB} + \varepsilon_c)\frac{d\bar{c}_A}{dy} \tag{3.113}$$

where $\bar{N}_{Ay}$ and $\bar{c}_A$ are time-smoothed quantities and ε_c is the eddy mass diffusivity. If a Prandtl mixing length expression is assumed for ε_c, then

$$\varepsilon_c = \varepsilon_m = \varepsilon_T = \varepsilon = l^2 \left|\frac{d\bar{u}_x}{dy}\right| \tag{3.114}$$

with the mixing length l given by Equation (2.299). Here again, as with energy transfer in turbulent flow, $D_{AB} \gg \varepsilon$ for the laminar sublayer next to the wall, $D_{AB} \approx \varepsilon$ in the buffer zone, and $D_{AB} \ll \varepsilon$ in the turbulent core.

Carrying out an analysis analogous to that of Section 2.5.4 gives the mass-transfer forms of the various analogies:

Reynolds analogy:

$$\frac{Sh_x}{Re_x Sc} = \frac{k_c}{u_\infty} = C_{Dx}/2 \tag{3.115}$$

Prandtl analogy:

$$\frac{Sh_x}{Re_x Sc} = \frac{C_{Dx}/2}{1 + 5\sqrt{C_{Dx}/2}\,(Sc - 1)} \tag{3.116}$$

Von Karman analogy:

$$\frac{Sh_x}{Re_x Sc} = \frac{C_{Dx}/2}{1 + 5\sqrt{C_{Dx}/2}\,\{Sc - 1 + \ln[(1 + 5Sc)/6]\}} \tag{3.117}$$

Further, the Colburn analogy can be written as

$$\frac{Sh_x}{Re_x Sc} = \frac{k_c}{u_\infty} = \frac{C_{Dx}/2}{Sc^{2/3}} \tag{3.118}$$

The last result is obtained directly from laminar boundary layer results; however, as noted in Section 2.5.4 it can be used as a first approximation to the Reynolds analogy for turbulent transport at Schmidt numbers different from 1. All of these analogies can be used for turbulent mass transport in tubes if u_∞ is replaced by u_{avg}.

Equation (3.118) can also be written as

$$j_{\mathrm{D}} = \frac{C_{\mathrm{D}x}}{2} \tag{3.118a}$$

where

$$j_{\mathrm{D}} \equiv \frac{k_{\mathrm{c}} Sc^{2/3}}{2u_\infty} \tag{3.118b}$$

is the j factor for mass transfer. When this result is combined with Equation (2.320 a) with $C_{\mathrm{D}x}$ replacing f, we obtain

$$j_{\mathrm{D}} = j_{\mathrm{H}} = \frac{C_{\mathrm{D}x}}{2} \tag{3.118c}$$

This result is called the Chilton – Colburn analogy and connects the mass transfer, heat transfer, and momentum transfer processes in a given flow geometry. Although Equation (3.118 c) is obtained directly from analysis of transport on laminar boundary layers, it can be used as an approximation for other geometries and for turbulent flows as long as there is no form drag. If form drag is present, the first part of Equation (3.118 c), i.e., $j_{\mathrm{D}}=j_{\mathrm{H}}$, can still be used to relate mass transfer coefficients to heat transfer coefficients if $0.6 < Sc < 2500$ and $0.6 < Pr < 100$.

Finally, analogous to Equation (2.325),

$$Sh_x = 0.0289\, Re_x^{4/5}\, Sc^{1/3} \tag{3.119}$$

for mass transfer in turbulent boundary layers. Combining this with the mass-transfer equivalent of Equation (2.252) for the laminar part of the boundary layer, i.e.,

$$Sh_x = 0.332\, Re_x^{1/2}\, Sc^{1/3} \tag{3.120}$$

yields

$$Sh_L = 0.0361\, Re_L^{4/5}\, Sc^{1/3} \left(1 - 9170\, Re_L^{-4/5}\right) \tag{3.121}$$

for the overall mass-transfer coefficient relation over a plate of length L where $L \geq x_{\mathrm{cr}}$. Equation (3.121) is the mass-transfer equivalent of Equation (2.328).

3.3. Mass Transfer Across Interfaces

In the situations described in the preceding sections, the rate of mass transfer within a single phase is calculated by using fundamental transport relations (mass balances and diffusion flux relations) and transport properties (diffusivities). Such calculations are possible for well-defined geometries and flow situations of sufficient simplicity that can be modeled and calculated a priori and without empiricism.

This not always the case, however, because a great many more situations occur in which the flow of the fluid phases or a complex geometry prevents exact modeling or calculation. An example is forced convection transfer across an interface. In this case, the situation is described by using *mass-transfer coefficients* instead of diffusivities. These coefficients play a role similar to diffusivities in that they describe the transport rate of mass that occurs primarily because of molecular motion, but they also allow for other effects such as the enhancement of mass transfer because of forced or natural convection. Once these coefficients have been determined experimentally for a number of flow and mass-transfer situations, and correlated by dimensionless groups, the coefficients for analogous situations can be estimated and used for process design.

The calculations are complicated, however, by the dependence of mass-transfer coefficients on mass flux. For example, in Section 3.1.6 the diffusion of species A through a stagnant film B was enhanced, above and beyond pure diffusion, by the net bulk flow of the mixture. Similarly, mass-transfer coefficients that apply to low mass-transfer rates are affected by the species mass fluxes and should be adjusted accordingly for the most accurate design calculations at high mass-transfer rates.

Figure 27 shows how mass-transfer coefficients that are calculated for low mass-transfer rates by using diffusion coefficients or dimensionless group correlations can be adjusted for high mass-transfer rates. Then the coefficients for mass transfer through films on both sides of an interface can be combined (a process requiring thermodynamic equilibrium between the two phases) to give an overall coefficient for transfer between phases.

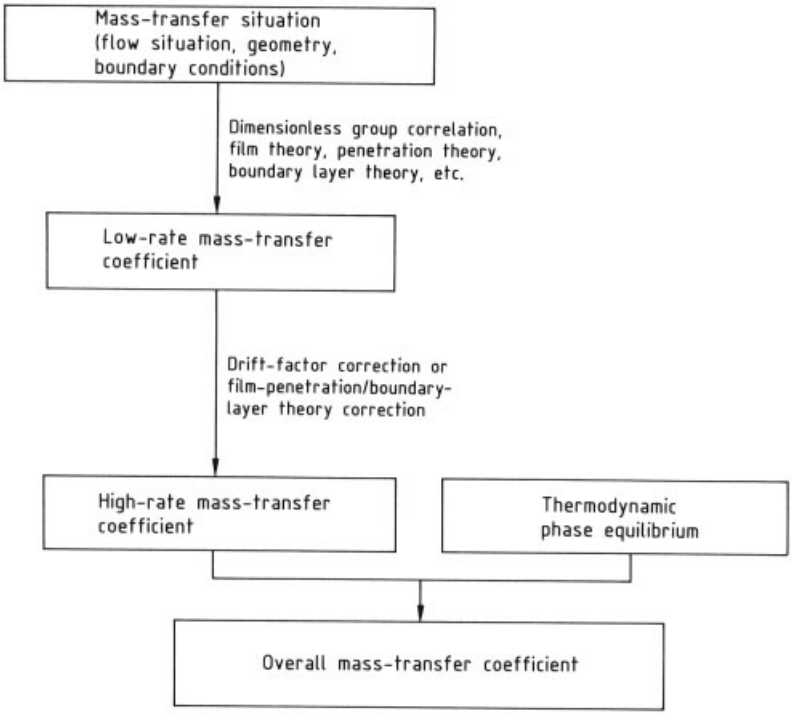

Figure 27. Representation of the calculation of overall mass-transfer coefficients showing corrections of the single-phase coefficients for high mass-transfer rates

The commonly used definitions of mass-transfer coefficients for each side of an interface and for the overall coefficients calculated from them are given in Section 3.3.1; their functional forms based upon theoretical considerations for low mass-transfer rates or equimolar counterdiffusion are presented in Section 3.3.2; and correlations for calculating mass-transfer coefficients for low mass-transfer rates and the adjustments that may be made for high mass-transfer rates are considered in Sections 3.3.3 and 3.3.4, respectively. General references providing more detailed discussion of mass-transfer coefficient definitions and calculations include texts in the areas of transport phenomena [63], [74] and chemical engineering unit operations [75–79].

3.3.1. Mass-Transfer Coefficients

Mass transfer occurs because of an imbalance of concentrations, a departure from equilibrium. This imbalance provides a driving force for mass transfer. Uniformity of composition is the equilibrium state in a single phase; if mole fractions are not uniform, then a nonequilibrium condition exists and diffusion occurs until uniformity is reached. Two-phase thermodynamic equilibrium is the equilibrium state across an interface; to the extent that the two phases on opposite sides of an interface are not in equilibrium, mass transfer tends to occur in such a way as to move the system toward equilibrium.

The degree of departure from equilibrium directly affects the rate of mass transfer. In a single phase, the degree of departure from equilibrium is represented by the mole fraction (or mass fraction) gradient, and Fick's first law of diffusion (the most commonly used flux relation) models the diffusion flux as proportional to this driving force. The proportionality factor defines the diffusivity.

For mass transfer across interfaces or films, an analogous relationship is normally used to define mass-transfer coefficients. The mass-transfer flux of a species at an interface is modeled as proportional to the driving force (concentra-

tion difference) which exists for that transfer, through a thin film next to the interface. This situation is depicted in Figure 28. At the interface the two phases are normally assumed to be in thermodynamic equilibrium. Away from the interface, however, the bulk concentrations of the two phases are not necessarily at equilibrium with each other, and possible concentration or mole fraction profiles are shown as a function of distance from the interface. The majority of the concentration change is modeled to occur over a laminar film region near the interface. The actual concentrations and film depths are not known, however, which makes the definitions quite empirical and dependent on parameters such as fluid flow and turbulence. In Figure 28, concentration profiles are shown in both phases and, for simplicity, one phase is called a gas phase and the other a liquid phase, although this is not a limitation or constraint on the situation. The discussion could just as well be for two liquid phases or for a fluid and a solid phase. The model also normally assumes that concentrations at the interface are at steady state; flux to the interface through one phase equals that away from the interface through the other.

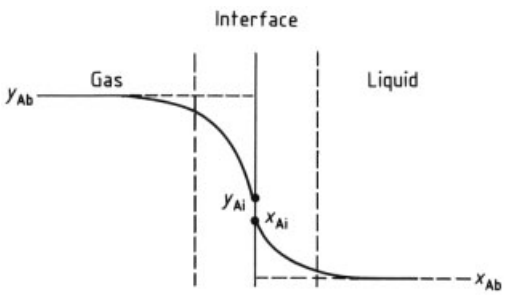

Figure 28. Hypothetical concentration profiles across the gas – liquid interface region with the transfer of A from the gas to the liquid
y_{Ab} = mole fraction of A in bulk gas phase; y_{Ai} = gas-phase mole fraction of A at interface; x_{Ai} = liquid-phase mole fraction of A at interface; x_{Ab} = mole fraction of A in bulk liquid phase

Mass-transfer coefficients, then, are defined for each of the two phases. The definition of a *liquid-phase mass-transfer coefficient* (based on a liquid-phase mole fraction driving force) is

$$\text{Flux of A} = k_x\,(x_{\mathrm{Ai}} - x_{\mathrm{Ab}}) \tag{3.122}$$

Likewise, the defining relation for the *gas-phase mass-transfer coefficient* for species A based on the gas-phase mole fractions is

$$\text{Flux of A} = k_y\,(y_{\mathrm{Ab}} - y_{\mathrm{Ai}}) \tag{3.123}$$

In each of these equations, a departure from equilibrium exists that represents the extent to which the interface mole fraction (x_{Ai} or y_{Ai}) differs from that in the bulk fluid (x_{Ab} or y_{Ab}) of the same phase.

Whereas the above relations define mass-transfer coefficients for a driving force within a single phase at an interface, interphase mass-transfer coefficients are also defined according to concentration or mole fraction differences that exist across the two phases, where the average or bulk concentrations are used for each phase. In this case the mass-transfer coefficients K_x and K_y are defined according to the relations

$$\text{Flux of A} = K_x\,(x_{\mathrm{Ae}} - x_{\mathrm{Ab}}) \tag{3.124}$$

$$\text{Flux of A} = K_y\,(y_{\mathrm{Ab}} - y_{\mathrm{Ae}}) \tag{3.125}$$

and are called *overall mass-transfer coefficients.* They describe the flux in terms of mole fractions in the bulk phases.

Here, instead of defining a driving force that exists within one phase or the other, a driving force that spans the two phases is defined. The mole fractions and driving forces are shown relative to a typical interfacial equilibrium curve in Figure 29. For a mass-transfer coefficient based on liquid-phase mole fractions, the driving force that is used is the difference between the actual mole fraction of A in the bulk liquid phase (x_{Ab}) and the mole fraction of A that would exist (x_{Ae}) if the liquid phase were in equilibrium with the mole fraction of A in the bulk gas phase. Likewise, in terms of gas-phase concentrations, mass transfer of A occurs to the extent that the bulk gas-phase mole fraction (y_{Ab}) differs from the value that would exist (y_{Ae}) if the gas phase were in equilibrium with the actual bulk liquid-phase mole fraction.

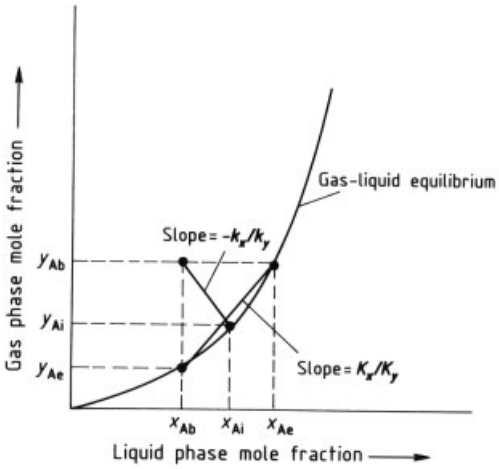

Figure 29. Relationships among interface, bulk, and equilibrium concentrations used in mass-transfer rate equations

The slopes of lines that represent the *ratios of mass-transfer coefficients* are also shown in Figure 29. If species A does not accumulate at the interface, the liquid- and gas-phase relationships for flux in terms of mass-transfer coefficients must be equal. Accordingly,

$$k_x\,(x_{\mathrm{Ai}} - x_{\mathrm{Ab}}) = k_y\,(y_{\mathrm{Ab}} - y_{\mathrm{Ai}}) \tag{3.126}$$

which gives

$$-\frac{k_y}{k_x} = \frac{x_{\mathrm{Ab}} - x_{\mathrm{Ai}}}{y_{\mathrm{Ab}} - y_{\mathrm{Ai}}} \tag{3.127}$$

and the ratio of the interphase mass-transfer coefficients is the slope of a tieline connecting the point with composition coordinates equal to the liquid- and gas-phase bulk concentrations to a point with coordinates equal to the equilibrium interface liquid- and gas-phase concentrations. Similarly, a ratio can be obtained for the overall transfer coefficients:

$$\frac{K_x}{K_y} = \frac{y_{\mathrm{Ab}} - y_{\mathrm{Ae}}}{x_{\mathrm{Ae}} - x_{\mathrm{Ab}}} \tag{3.128}$$

In the limit of small driving forces or for a linear isotherm this ratio is the slope m of a tangent to the equilibrium curve in the concentration region of interest.

From the definition of the mass-transfer coefficients and for a locally linear isotherm (slope = m),

$$\frac{1}{K_x} = \frac{1}{k_x} + \frac{1}{m\,k_y} \tag{3.129}$$

and

$$\frac{1}{K_y} = \frac{1}{k_y} + \frac{m}{k_x} \tag{3.130}$$

Hence, the overall or combined resistance to mass transfer through the two phases ($1/K_x$ or $1/K_y$) is equal to the sum of the resistances through each of the phases individually. Before summing, however, one of the individual phase coefficients must be scaled by using the (local) slope of the equilibrium curve in order to be consistent with the resistance offered by the other mass-transfer coefficient. Note that if $k_x/m \gg k_y$, then the gas-phase mass transfer is limiting and $K_y \approx k_y$.

Dimensions of Mass-Transfer Coefficients. Because the flux of A is the number of moles of A per time per (cross-sectional) area, the mass-transfer coefficients as defined by these relations must also have the dimensions of number of moles per time per area. Other definitions using different driving force concentration units are employed, however, and the dimensions of the mass-transfer coefficient vary accordingly. For example, number of moles per volume is frequently used for liquid-phase concentrations and partial pressure for gas-phase concentration. In these situations, mass-transfer coefficients may be defined according to

$$\text{Flux of A} = k_{\mathrm{c}}\,(c_{\mathrm{Ai}} - c_{\mathrm{Ab}}) \tag{3.131}$$

$$\text{Flux of A} = k_{\mathrm{G}}\,(p_{\mathrm{Ab}} - p_{\mathrm{Ai}}) \tag{3.132}$$

$$\text{Flux of A} = K_{\mathrm{c}}\,(c_{\mathrm{Ae}} - c_{\mathrm{Ab}}) \tag{3.133}$$

$$\text{Flux of A} = K_{\mathrm{G}}\,(p_{\mathrm{Ab}} - p_{\mathrm{Ae}}) \tag{3.134}$$

Here, k_{c} and K_{c} have the dimensions of volume per time per area (length per time), and k_{G} and K_{G} have the dimension of number of moles per time per area per unit pressure.

3.3.2. Functional Forms of Mass-Transfer Coefficient Relations

In the preceding discussion, the mass flux through a film is assumed to be proportional

to the driving force (the concentration difference) across a film, and this proportionality factor defines the mass-transfer coefficient. Combining this definition with theoretical models for mass transfer in terms of diffusion coefficients allows calculation of the dependence of mass-transfer coefficients on diffusivity. These are only estimates, however, because the situations are too complex for exact modeling, which is why mass-transfer coefficients were defined in the first place. A complete discussion of several models appears in [75].

Two-Film Theory. A particularly simple but useful model for steady-state transfer between two phases, which is frequently used to describe mass-transfer coefficients in terms of diffusion coefficients, is the two-film theory of WHITMAN [80]. The concentration profile is taken to be flat (independent of position) in the bulk fluid and then to vary linearly across a thin film of thickness δ up to the interface (Fig. 30). For steady-state mass transfer at low mass-transfer rates the results of this model are the same as for equimolar counterdiffusion (see Section 3.1.6); the molar flux is directly proportional to the diffusivity, and the mass-transfer coefficient, in terms of film thickness and diffusivity, is

$$k_x = \frac{c D_{\mathrm{Ab}}}{\delta} \tag{3.135}$$

Hence, this film theory predicts a mass-transfer coefficient that is directly proportional to diffusivity through the film.

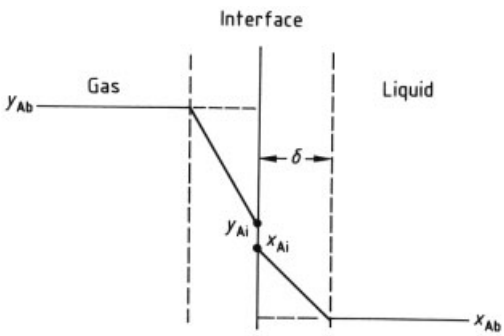

Figure 30. Simplified concentration profiles for the two-film theory

Other Theories. Other theories relating the coefficient to diffusion may also be appropriate. Examples are the *penetration theory* [81] and its adaptations [82], [83]. The penetration theory is applicable to a species that is being transferred from a gas to a liquid phase; it gives an estimate of the mass-transfer coefficient in the liquid phase based upon a nonsteady-state situation. The liquid phase is assumed to be a falling film and the velocity profile of the falling film at the gas – liquid interface is essentially flat, i.e., the velocity of the falling liquid is nearly independent of penetration depth from the gas into the liquid as long as the penetration depth is low. This is true if the time that the gas is in contact with the liquid is short. In this situation, any one packet of fluid is in contact with the gas for a fixed amount of exposure time; therefore, the diffusion of a species from the gas to the liquid – from the viewpoint of this packet of fluid – is a nonsteady-state process as opposed to the steady-state diffusion of the film theory. This different type of model gives rise to a mass-transfer coefficient that is proportional to the square root of diffusivity:

$$k_x = 2c\sqrt{\frac{D_{\mathrm{AB}}}{\pi t_{\mathrm{exp}}}} \tag{3.136}$$

Here, t_{exp} is the exposure time of a given packet or element of fluid to the gas stream and k_x is an average value over this time. Exposure time can be calculated in terms of the fluid velocity at the gas film interface and the distance of contact of the gas and liquid L:

$$t_{\mathrm{exp}} = \frac{L}{u_{\mathrm{interface}}} \tag{3.137}$$

A third theory, actually a combination of both, is the *film-penetration theory* [84]. For some flow situations, the film theory result is recovered, whereas for others, the penetration theory's $D^{1/2}$ dependence holds, suggesting that different dependencies might be expected in different flow situations.

3.3.3. Relations for Low Mass-Transfer Rates

By using a somewhat different approach, dimensionless group correlations can be postulated to exist for mass-transfer coefficients. For example, the Sherwood number ($Sh_{\mathrm{AB}} = k_x\, d/c\, D_{\mathrm{AB}}$)

can be expressed in terms of a function of the Reynolds number ($Re = d\,u\,\varrho/\eta$), Schmidt number ($Sc = \eta/\varrho\,D_{\mathrm{AB}}$), and L/d [63]:

$$Sh_{\mathrm{AB}} = f\ (Re,\, Sc,\, L/d) \tag{3.138}$$

The specific functional form depends on the actual flow situation.

Furthermore, as discussed in Section 3.2, analogous correlations exist for heat, mass, and momentum transfer because of similarities in the conservation equations that govern all these situations and in the corresponding flux relations. Consequently, correlations exist for mass-transfer coefficients that are very similar to those for heat or momentum transfer (friction loss) in analogous situations of geometry and boundary conditions. Results for some specific flow and mass-transfer situations are summarized in Table 8. Note that in these relations, the dependence of the mass-transfer coefficient on diffusivity ranges from about $D_{\mathrm{AB}}^{0.6}$ to D_{AB}, corresponding to the $D_{\mathrm{AB}}^{0.5}$ to D_{AB} predicted by the penetration and film theories, respectively.

3.3.4. Relations for High Mass-Transfer Rates

In Section 3.1.6, two important limiting cases of binary diffusion are considered for a one-dimensional diffusion process. In the case of equimolar counterdiffusion the flux of each species is the result of diffusion only and is calculated by using Fick's first law of diffusion; for each species the flux is directly proportional to the mole fraction gradient of that species. For the case of one species diffusing through a stagnant film of the other, however, the diffusion of each species is the result of the combination of Fick's law of diffusion and a bulk flow that is induced by the diffusion process. For the species that was defined to be stagnant, the diffusion flux exactly balances its flux due to the mixture bulk flow. For the other species, however, the total flux is that due to Fick's law of diffusion enhanced by the bulk flow, which results in the Fick's law flux being divided by $(1 - x_{\mathrm{A}})$.

In the latter case, if the nonstagnant species is present at low molar concentration (infinite dilution or approaching infinite dilution), then obviously the same result for the flux of this species is obtained as for the equimolar counterdiffusion situation but for a different reason. The other species is still stagnant, and the bulk flow is not reduced by the low concentration but rather by transport of the nonstagnant species that occurs because of reduced bulk flow.

These two limiting cases (ECD and ATSB) provide a basis for understanding the corrections to mass-transfer coefficients that are made for high mass-transfer rates. The mass-transfer coefficients described in the correlations of Table 8 are defined for equimolar counterdiffusion-like situations; that is, they apply to true equimolar counterdiffusion situations or to low transfer rate calculations. Evidently, then, if they are to be used for ATSB situations, they must be corrected in a way similar to the comparison between equimolar counterdiffusion and ATSB calculations; the presence of a stagnant film and high mass-transfer rates provides an induced bulk flow that enhances the mass-transfer coefficient. If the low mass-transfer coefficient is used in a high mass-transfer stagnant film situation, the mass-transfer rates may be significantly underestimated.

A common correction to the low mass-transfer rate – ECD coefficients is to introduce a drift factor $\overline{(1 - y_{\mathrm{A}})}_L$ analogous to the adjustment of ECD to ATSB. A common, although not unique, notation is to distinguish between these two mass-transfer coefficients with a prime ($'$); however, the notation may be switched with k' indicating ECD rather than ATSB as is done here:

$$k'_y = \frac{k_y}{\overline{(1 - y_{\mathrm{A}})}_L} \tag{3.139}$$

where $\overline{(1 - y_{\mathrm{Ai}})}_L$ is the nonlinear, log-mean average of $(1 - y_{\mathrm{Ai}})$ and $(1 - y_{\mathrm{Ab}})$:

$$\overline{(1 - y_{\mathrm{A}})}_L = \frac{(1 - y_{\mathrm{Ai}}) - (1 - y_{\mathrm{Ab}})}{\ln \frac{(1 - y_{\mathrm{Ai}})}{(1 - y_{\mathrm{A}})}} \tag{3.140}$$

This correction is commonly used for ATSB-like situations and appears in design calculations such as packed gas-absorption towers where one species in the gas phase is absorbed and the others are not.

Instead of a simple ATSB correction to the mass-transfer coefficient, however, less restrictive theories for flux ratios may be used to make the adjustment. These are (1) the film theory, (2) the penetration theory, and (3) the boundary

Table 8. Mass-transfer coefficients [a]

Situation	Equation	Restrictions	Equation (), reference []
Transfer from a Cylindrical Tube Wall to a Fluid			
[Tube diameter = d, tube length = L, average fluid velocity = u_{avg}, $Sh = k_c d/D_{AB}$, $Re = d u \varrho/\mu$, $Sc = \eta/(\varrho D_{AB})$]			
Coefficient at distance L downstream with fully developed laminar velocity profile, constant wall concentration	$Sh = 3.66 + \dfrac{0.0668\,[(d/L)\,Re_d\,Sc]}{1 + 0.04\,[(d/L)\,Re_d\,Sc]^{2/3}}$	$Re < 2000$	[3.23]
Same as above but with constant wall mass flux	$Sh = 4.36 + \dfrac{0.023\,[(d/L)\,Re_d\,Sc]}{1 + 0.0012\,[(d/L)\,Re_d\,Sc]}$	$Re < 2000$	[3.23]
Turbulent liquid flow, $L/d > 6$, transfer of solid from the wall [b]	$Sh = 0.023\,Re^{0.83}\,Sc^{1/3}$	$2000 < Re < 70\,000$ $0.6 < Sc < 3000$	[3.24]
Turbulent gas flow, $L/d > 6$, transfer from liquid film at the wall (wetted-wall column) [b]	$Sh = 0.023\,Re^{0.83}\,Sc^{0.44}$	$2000 < Re < 35\,000$ $0.6 < Sc < 2.5$	[3.25]
Turbulent liquid flow, high Re, high Sc numbers	$Sh = 0.0096\,Re^{0.913}\,Sc^{0.346}$	$10\,000 < Re < 100\,000$ $430 < Sc < 100\,000$	[3.26]
Transfer from a Flat Plate Immersed in a Fluid Parallel to the Flow			
[$Sh_x = k_c x/D_{AB}$, $Sh_L = k_{CL} L/D_{AB}$, $Re_x = \varrho u_\infty x/\eta$, $Re_L = \varrho u_\infty L/\eta$, $Sc = \eta/(\varrho D_{AB})$, u_∞ = fluid velocity far from the plate]			
Laminar flow, coefficient at distance x from the plate leading edge	$Sh_x = 0.332\,Re_x^{0.5}\,Sc^{1/3}$	$Re_x < 2 \times 10^5$ $0.6 < Sc < 2500$	(3.120), [3.27]
Laminar flow, coefficient averaged over the plate from the leading edge to distance L	$Sh_L = 0.664\,Re_L^{0.5}\,Sc^{1/3}$	$Re_L < 2 \times 10^5$ $0.6 < Sc < 2500$	[3.17], [3.27]
Turbulent flow	$Sh_L = 0.036\,Re_L^{0.8}\,Sc^{1/3}$	$2 \times 10^5 < Re_L < 10^7$ $0.6 < Sc < 2500$	[3.27]
Transfer from Single Cylinders Perpendicular to the Flow			
[Cylinder diameter = d, $Sh = k_c d/D_{AB}$, $Re = d u_\infty \varrho/\eta$, $Sc = \eta/(\varrho D_{AB})$]			
Flow of gases	$Sh = 0.281\,Re^{0.6}\,Sc^{0.44}$	$400 < Re < 25\,000$ $0.6 < Sc < 2.6$	[3.16], [3.28]
Transfer in Packed or Fluidized Beds			
[Particle diameter = d_p, superficial velocity = u', void fraction = ε, $Re = d_p u' \varrho/\eta$, $Re' = d_p u' \varrho/\eta(1-\varepsilon)$, $Sc = \eta/(\varrho D_{AB})$, $j_D = k_c Sc^{2/3}/u'$]			
Packed or fluidized beds; preliminary design only unless gases for $Re < 10$	$Sh = (0.765\,Re^{0.18} + 0.365\,Re^{0.614})\dfrac{Sc^{1/3}}{\varepsilon}$	$0.01 < Re < 15\,000$	[3.33]
Gas or liquid fluidized bed with spherical particles	$Sh = \left[0.01 + \dfrac{0.863}{Re^{0.58} - 0.483}\right]\dfrac{Re\,Sc^{1/3}}{\varepsilon}$	$10 < Re < 2000$	[3.34]
Packed or fluidized bed, gases or liquids, spherical or regular cylindrical pellets [d]	$j_D = 5.7\,Re'^{-0.78}$	$0 < Re' < 30$ $0.6 < Sc < 1400$	[3.35]
	$j_D = 177\,Re'^{-0.44}$	$30 < Re' < 10\,000$ $0.6 < Sc < 1400$	[3.35]

[a] Unless otherwise noted, the coefficients are average values and solute concentrations are constant at the interface. [b] The difference between mass transfer from a liquid film and transfer from a solid is attributed to ripples in the liquid film. [c] Also reported are correction factors for cylinders [for $L/d = 1$, $\varepsilon j_D/(\varepsilon j_D)_{sphere} = 0.79$] and cubes [$\varepsilon j_D/(\varepsilon j_D)_{cube} = 0.71$] where $j_D = k_c Sc^{2/3}/u'$. [d] Correction factors for irregularly shaped particles—Raschig rings, Berl saddles, and flakes—are also reported.

layer theory. Each of these theories provides a means of calculating a correction factor to the mass-transfer coefficient.

The situation of ATSB diffusion is actually a specific limiting case of the more complete film theory summarized by BIRD et al. [63]. They use the film theory to adjust EDC mass-transfer coefficients for the enhancements which result from diffusion-induced bulk flow. The model allows for relative fluxes of the two species (A and B) that differ from the stagnant B situation. For a given flux ratio, adjustments to the low mass-transfer rate coefficient can be calculated (or, equivalently, the ECD mass-transfer coefficient), which would be calculated or estimated from the correlations of the previous section. Again, ATSB situation falls within this less restrictive film theory.

The penetration theory of HIGBIE [81] has also been used to extend the low mass-transfer rate calculations to high mass-transfer rates due to diffusion-induced enhancement. The results obtained are presented in a similar way to those for the film theory, and correction factors are given for a boundary layer theory approach [63]. Qualitatively, the corrections calculated by these three theories are very similar although the quantitative results are somewhat different. The penetration and boundary layer theories predict an even larger effect of mass transfer on the coefficients than does the film theory.

4. Macroscopic Systems

The development of conservation and accounting equations for macroscopic systems parallels the development of those for differential systems that led to the partial differential equations of Table 1. Each law or statement is written for a single extensive property (a property that can be counted, for which the whole is equal to the sum of the parts) for a specified region (system) inside a closed-surface boundary and for a specified time period (either finite or differential) (Fig. 31). In the case of macroscopic (finite-size) systems, either process rate equations are obtained, which are ordinary differential equations, or for a finite time period, an algebraic equation is obtained. For conserved properties, the generation (or consumption) is always zero.

The conservation and accounting equations and their development are summarized below for compositionally homogeneous systems. The purpose is to provide an overall picture relating conservation laws, accounting equations, *U, H, A, G,* and second law equations of thermodynamics, which provide the basics for quantitatively understanding transport phenomena (and other areas of engineering as well) on a macroscopic scale, rather than to give handbook-type details and examples. General references offering further discussion of the development and application of macroscopic equations include texts in the areas of transport phenomena [98–100], thermodynamics [101], [102], chemical engineering unit operations [103], and fluid mechanics [104].

4.1. Conservation of Total Mass

Total mass is conserved in the absence of nuclear conversions or relativistic effects. This means that for the system and its surroundings together the total amount of mass is a constant. Equivalently, if the amount of mass contained within the system changes, then the surroundings must change to an equal but opposite extent; there is a one-for-one exchange of mass between the system and its surroundings as mass moves across the system boundary. Allowing for multiple flow streams into and out of the system, an equation for the conservation of total mass contained within the system is

$$\sum_{\text{entering streams}} (\varrho u_{\text{avg}} A)_{\text{in}} - \sum_{\text{exiting streams}} (\varrho u_{\text{avg}} A)_{\text{out}} = \frac{\mathrm{d}}{\mathrm{d}t} (\varrho V)_{\text{system}} \tag{4.1}$$

At each area A where mass exchanges occur between the system and surroundings, the fluid density and velocity components normal to the boundary surface u are appropriately averaged values. Likewise, the density of the system is averaged over the system volume V_{system} [98], [100], [105].

In addition to a conservation equation for total mass, accounting statements for individual species within the system can be written. If the system contains n species that together comprise

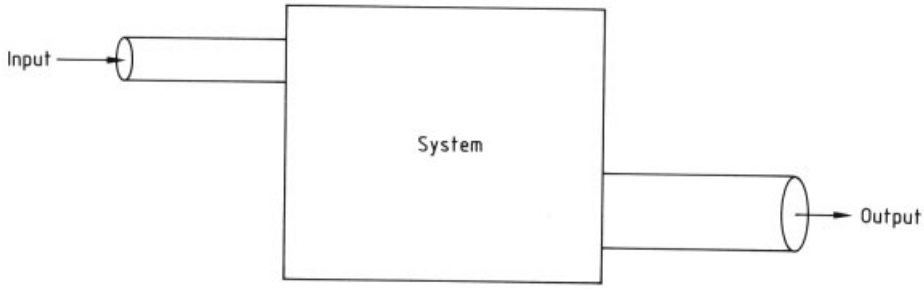

$$\left\{\begin{matrix}\text{Extensive property}\\ \text{entering the system}\\ \text{during the time period}\end{matrix}\right\} - \left\{\begin{matrix}\text{Extensive property}\\ \text{leaving the system}\\ \text{during the time period}\end{matrix}\right\} + \left\{\begin{matrix}\text{Extensive property}\\ \text{generated within the system}\\ \text{during the time period}\end{matrix}\right\} - \left\{\begin{matrix}\text{Extensive property}\\ \text{consumed within the system}\\ \text{during the time period}\end{matrix}\right\}$$

$$= \left\{\begin{matrix}\text{Accumulation of extensive}\\ \text{property within the system}\\ \text{during the time period}\end{matrix}\right\}$$

$$\equiv \left\{\begin{matrix}\text{Extensive property}\\ \text{contained within the system}\\ \text{at the end of the time period}\end{matrix}\right\} - \left\{\begin{matrix}\textbf{Extensive property contained}\\ \textbf{within the system at the}\\ \textbf{beginning of the time period}\end{matrix}\right\}$$

Figure 31. The concept of accounting for an extensive property – the cornerstone of transport phenomena
Input and output represent an exact exchange with the surroundings; a conserved property has no generation or consumption term.

the total mass (and could be molecular, ionic, or other species), then n accounting equations exist, one for each species. Because species are not necessarily conserved (except for the chemical elements), these accounting equations must allow for any generation or consumption of individual species through chemical reactions. Accordingly, for each of the n species within the system,

$$\sum_{\text{entering streams}} (\varrho_i\, u_{\text{avg},i} A)_{\text{in}} - \sum_{\text{exiting streams}} (\varrho_i\, u_{\text{avg},i} A)_{\text{out}} + r_i\, V_{\text{sys}} = \frac{\mathrm{d}}{\mathrm{d}t}(\varrho_i\, V)_{\text{system}} \tag{4.2}$$

The densities and velocities are average values, as discussed above, and r_i is an average generation rate of species i per unit volume of the system. An equivalent accounting can be done in molar concentration units (see, e.g., Section 4.9.3).

Of these $n+1$ equations, however, only n are independent inasmuch as the total mass must be the sum of the individual species mass (because total mass is conserved, the sum of all of the generation and consumption terms in the n species equations must be zero; no net generation or consumption of mass occurs). Note that the total mass of each species entering and leaving the system is obtained by summing over all entering and leaving streams.

4.2. Conservation of Linear Momentum

Linear momentum is conserved. Given a system, the amount of linear momentum contained therein may change only as a result of exchange of momentum between the system and its surroundings; linear momentum is not generated or consumed. Linear momentum may be exchanged as the result of mass entering or leaving the system and as the result of forces acting on the system by its surroundings. The rate of momentum exchange as mass enters or leaves the system is the momentum per unit mass $\boldsymbol{u}$ multiplied by the rate at which that mass enters or leaves the system. External forces act upon the system either by contact with the system's surfaces or as body forces, such as gravity, that

act individually on each piece of mass within the system. The momentum of the system then changes according to

$$\sum \left[\frac{\left(u^2\right)_{\text{avg}}}{u_{\text{avg}}} \varrho u_{\text{avg}} \boldsymbol{A} \right]_{\text{in}} - \sum \left[\frac{\left(u^2\right)_{\text{avg}}}{u_{\text{avg}}} \varrho u_{\text{avg}} \boldsymbol{A} \right]_{\text{out}} + \sum \boldsymbol{F}_{\text{external}} = \frac{\mathrm{d}\boldsymbol{p}_{\text{system}}}{\mathrm{d}t} \tag{4.3}$$

The $(u^2)_{\text{avg}}$ term arises because the momentum flow rate, calculated as the momentum per unit mass $\boldsymbol{u}$ times the mass flow rate $\varrho\, u\, \mathrm{d}\boldsymbol{A}$, is integrated (averaged) over the cross section [98, Chap. 7], [99, Chap. 7]. If the velocity profile across the area is not flat, then $(u^2)_{\text{avg}} \neq (u_{\text{avg}})^2$.

If $\alpha_{\text{mom}} \equiv (u_{\text{avg}})^2/(u^2)_{\text{avg}}$, where α_{mom} is a measure of the extent to which the velocity profile is not flat, this equation becomes

$$\sum \left[\frac{\varrho \left(u_{\text{avg}}\right)^2}{\alpha_{\text{mom}}} \boldsymbol{A} \right]_{\text{in}} - \sum \left[\frac{\varrho \left(u_{\text{avg}}\right)^2}{\alpha_{\text{mom}}} \boldsymbol{A} \right]_{\text{out}} + \sum \boldsymbol{F}_{\text{external}} = \frac{\mathrm{d}\boldsymbol{p}_{\text{system}}}{\mathrm{d}t} \tag{4.4}$$

For fully developed laminar flow of a Newtonian fluid, α_{mom} is 3/4; for turbulent flow, it is nearly unity [106, p. 252]. Also $\boldsymbol{p}_{\text{system}}$ is integrated over the entire macroscopic system.

4.3. Conservation of Angular Momentum

Angular momentum is also a conserved property. The angular momentum of a system can change only by exchange with its surroundings. A piece of mass possesses angular momentum that is equal to the cross product of the position vector of that piece of mass and its momentum. Consequently, angular momentum enters and leaves with mass. Also, angular momentum enters the system as a result of external forces and is equal to the cross product of the position vector to the point of application of the force with that force:

$$\boldsymbol{r}_{\text{in}} \times \boldsymbol{u}_{\text{in}} \left(\varrho u_{\text{avg}} A\right)_{\text{in}} - \boldsymbol{r}_{\text{out}} \times \boldsymbol{u}_{\text{out}} \left(\varrho u_{\text{avg}} A\right)_{\text{out}} + \sum \left(\boldsymbol{r} \times \boldsymbol{F}_{\text{ext}}\right) = \frac{\mathrm{d}\boldsymbol{L}_{\text{system}}}{\mathrm{d}t} \tag{4.5}$$

where $\boldsymbol{L}_{\text{system}}$ is the total angular momentum of the system. Again, the entering and leaving momenta are actually appropriate integral averages across the flow cross section, and $\boldsymbol{L}_{\text{system}}$ is integrated over the entire volume.

The conservation of angular momentum is not considered further here. It is useful, for example, for the analysis of torques on pumps and compressors for design purposes [100].

4.4. An Accounting of Mechanical Energy

Mechanical energy accounting can be made for finite-sized systems and can be obtained by integrating the continuum conservation law of linear momentum. It cannot be obtained from a macroscopic linear momentum law and consequently is an independent equation providing additional insight into flow situations. In this sense, the set of macroscopic equations differs from the continuum equations.

The mechanical energy equation is an accounting equation, as opposed to a conservation equation, because mechanical energy is not conserved. Conversion of mechanical energy to thermal or other forms of energy may occur so that the amount of mechanical energy in an isolated system is not necessarily constant. Nevertheless, such an accounting is extremely useful for certain problems, e.g., the design and sizing of piping systems and flow networks.

Derivation of the mechanical energy accounting equation is complex and lengthy [107]. The result is stated here for steady-state flow situations with one entrance and one exit. Mechanical energy is accounted for by the following steady-state equation (expressed as energy per unit mass of flow; the symbol ˆ over a quantity implies "per unit mass"):

$$g\left(z_1 - z_2\right) + \frac{1}{2}\left[\frac{\left(u_{1,\text{avg}}\right)^2}{\alpha_{KE,1}} - \frac{\left(u_{2,\text{avg}}\right)^2}{\alpha_{KE,2}}\right] - \int_1^2 \frac{\mathrm{d}p}{\varrho} + \hat{W}_{\text{s}} + \hat{W}_{\text{es}} - \hat{F} = 0 \tag{4.6}$$

Here $\alpha_{KE} \equiv (u_{\text{avg}})^3/(u^3)_{\text{avg}}$ is another measure of the flatness of the velocity profile, this time with respect to kinetic energy. For fully developed laminar flow of a Newtonian fluid, α_{KE}

is 1/2; for turbulent flow it is nearly unity [106, p. 252]. This result holds regardless of heat transfer or other thermal effects that may occur.

Equation (4.6) is referred to as the *extended Bernoulli equation* (→ Fluid Mechanics, Chap. 3.5.) and through $\hat{F}$ allows for the irreversible or dissipative conversion of mechanical energy to thermal energy in a manner analogous to that of Equation (1.80), the mechanical energy equation for a continuum. These losses are of the type encountered in flow through pipes where friction losses result in pressure drop in a straight run of horizontal pipe. The potential energy gz and kinetic energy $\boldsymbol{u}^2/2\,\alpha_{\mathrm{KE}}$ terms are evaluated (as averages over the cross section) at the entrance (1) and exit (2) points of the system. The shaft work $\hat{\boldsymbol{W}}_{\mathrm{s}}$ and the flow work due to the fluid extra stresses $\hat{\boldsymbol{W}}_{\mathrm{es}}$ act across the system boundary, whereas the friction losses occur both within the system (from viscous stresses inside the pipe) and across the system boundary (from viscous stresses at the pipe wall). The pressure term $\mathrm{d}p/\varrho$ must be integrated along a path through the process from the entrance to the exit, not necessarily an easy task for a compressible flow. If the flow is incompressible, then the extended Bernoulli equation becomes

$$g\,(z_1 - z_2) + \frac{1}{2}\left[\frac{(u_{1,\mathrm{avg}})^2}{\alpha_{KE,1}} - \frac{(u_{2,\mathrm{avg}})^2}{\alpha_{KE,2}}\right] + \frac{(p_1 - p_2)}{\varrho} + \hat{W}_{\mathrm{s}} + \hat{W}_{\mathrm{es}} - \hat{F} = 0 \tag{4.7}$$

for which the pressure terms also use entrance and exit values only. Mechanical energy in the form of (1) upstream kinetic energy, (2) potential energy, (3) isotropic and extra stress flow work, and (4) shaft work is either redistributed downstream among these forms of mechanical energy, converted to work, or lost as mechanical energy by conversion to thermal energy.

4.5. Conservation of Total Energy

Total energy is conserved. Mass possesses energy in the form of kinetic, potential, and internal energy. When mass enters a system, it brings with it energy in these forms, and the rate at which it does so is calculated as the product of the specific energy in each of these forms (i.e., the energy per unit mass) and the rate of mass flow into the system.

Heat Q and work W exist at the system boundary but are not energy possessed by mass. Heat enters the system by either conduction or radiation because of a temperature difference across the system boundary. Work is energy transferred when a force acts through a distance upon elements of the system at the boundary. More properly, work may be expressed as

$$W = \int_{s_1}^{s_2} \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{s} \tag{4.8}$$

The dot product of the force $\boldsymbol{F}$ with the displacement $\mathrm{d}\boldsymbol{s}$ provides that the component of force in the direction of displacement produces work, whereas that which is normal to the displacement does not. The convention here is that both heat and work are positive when energy is added to the system.

With energy existing within the system as kinetic, potential, and internal energy, the conservation of total energy for a macroscopic system is expressed in rate forms as

$$\sum\left[\left(\hat{U}+\frac{u^2}{2\alpha_{KE}}+gz\right)(\varrho u_{\mathrm{avg}}A)\right]_{\mathrm{in}} - \sum\left[\left(\hat{U}+\frac{u^2}{2\alpha_{KE}}+gz\right)(\varrho u_{\mathrm{avg}}A)\right]_{\mathrm{out}} + \dot{W} + \dot{Q} = \frac{\mathrm{d}}{\mathrm{d}t}\left[\left(\hat{U}+\widehat{KE}+\widehat{PE}\right)m\right]_{\mathrm{sys}}$$

The dots over W and Q imply work rate and heat-transfer rate, respectively, rather than a derivative rate of change of work or heat. Again, properties of the mass entering and leaving the system are averaged over the flow cross section.

Usually, the form of the equation is changed by separating the work required to move mass into and out of the system against the pressure (isotropic stress) at the boundary (expressed per unit mass p/ϱ) from all other work of the process, the "nonpressure" work $\dot{W}_{\mathrm{np}}$. (The term p/ϱ is commonly referred to as "flow work". However, because it represents only the flow work due to isotropic stress and does not include that due to extra stress, it is referred here to as pressure work and all other forms as "nonpressure" work.) The statement is rewritten by combining the p/ϱ terms with the other inlet and outlet flow stream terms and replacing the total work term with $\dot{W}_{\mathrm{np}}$. The p/ϱ terms are combined with the

internal energy to give the enthalpy of the flow streams ($\hat{H} \equiv \hat{U} + p/\varrho$) and the resulting equation is

$$\sum \left[\left(\hat{H} + \frac{u^2}{2\alpha_{KE}} + gz \right) (\varrho u_{\text{avg}} A) \right]_{\text{in}} - \sum \left[\left(\hat{H} + \frac{u^2}{2\alpha_{KE}} + gz \right) (\varrho u_{\text{avg}} A) \right]_{\text{out}} + \dot{W}_{\text{np}} + \dot{Q} = \frac{\mathrm{d}}{\mathrm{d}t} \left[\left(\hat{U} + \widehat{KE} + \widehat{PE} \right) m \right]_{\text{sys}} \tag{4.10}$$

The "nonpressure" work includes work associated with volume changes of the system acting against external forces (a nonsteady-state term), all shaft work (compressors, pumps, turbines, electric motors or generators, etc.), other forms of energy (electrical, magnetic, etc.), plus – if important – any flow work due to extra (nonisotropic) stresses at the system boundaries. The last term is frequently neglected.

4.6. An Accounting of Thermal Energy

By subtracting the steady-state mechanical energy equation for one inlet (point 1) and one outlet (point 2) flow stream (Eq. 4.6) from the corresponding result for total energy conservation (Eq. 4.10), an accounting of thermal energy per unit mass is obtained which, although not independent from the total and mechanical energy equations, can be a useful way of viewing thermal effects:

$$\hat{H}_2 - \hat{H}_1 = \int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} + \hat{Q} + \hat{F} \tag{4.11}$$

This result is true regardless of any changes in elevation, pressure, density, or flow velocity or the amount of shaft work, as long as a steady state exists through a single conduit. Furthermore, for a reversible situation ($\hat{F} = 0$), this result is readily identified as the thermodynamic relation $\mathrm{d}\hat{H} = T\,\mathrm{d}\hat{S} + \mathrm{d}p/\varrho$, which is integrated to follow an element of the fluid through the process along the flow path in a sequence of quasi-equilibrium steps. For irreversible processes this thermodynamic relation still holds, but $\hat{Q}$ and $\hat{F}$ together correspond to the integral of $T\,\mathrm{d}\hat{S}$; both heat transfer to the system and irreversibilities result in an increase in entropy (see Section 4.7).

These thermodynamic relations for $\mathrm{d}\hat{U}$ (and, equivalently, for $\mathrm{d}\hat{H}$, $\mathrm{d}\hat{A}$, and $\mathrm{d}\hat{G}$) may be viewed as thermal energy accounting equations, expressed in terms of thermodynamic state functions rather than the path functions $\hat{Q}$ and $\hat{F}$ of Equation (4.11). On the one hand, they represent relations between thermodynamic state variables regardless of the path or process used to bring about change; on the other hand, they represent thermal energy conversion for a specific process.

Both statements have intrinsic value in their own right. Given a change in enthalpy that occurs in a process and is calculated from such a thermal energy equation, the thermodynamic relations can be used to express this change in terms of changes in observable properties such as heat capacity, temperature, pressure, and volume. HOUGEN et al. [101, p. 508] recognize this separation of mechanical and thermal energy, and say that these state functions of a body do not depend upon "its external position or motion relative to other bodies." This result is considered further in Section 4.9.1.

The approach outlined in Sections 4.2, 4.3, 4.4, 4.5, and 4.6 for deriving the macroscopic momentum, mechanical energy, total energy, and thermal energy equations is not the most common. More frequently, the extended Bernoulli equation is hypothesized by considering total energy for adiabatic, incompressible situations or by applying the thermodynamic realtions for $\mathrm{d}\hat{U}$ or $\mathrm{d}\hat{H}$ (e.g., in the form of Eq. 4.11) to total energy (Eq. 4.10) to obtain Equation (4.6) or (4.7). The approach given here is preferred because it is more direct and parallels the traditional approach used in continua (Chap. 1). This preference has also been expressed by WHITAKER [105].

4.7. The Second Law of Thermodynamics

The second law of thermodynamics places constraints on processes that can occur in macroscopic systems. As discussed for continua, the thermodynamic property of entropy quantifies the fact that certain processes are observed to occur in only one direction. In the context of a macroscopic law, the entropy contained within a macroscopic system changes through exchange

with the surroundings and also through generation within the system as a result of irreversibilities. On the one hand, accounting for entropy is similar to accounting for other nonconserved properties because the entropy of the system may change as a result of generation within the system, as well as exchange with the surroundings. Unlike other accounting statements (e.g., mechanical energy, electrical energy, individual chemical species), however, entropy cannot be consumed within the system; it can only be generated. Also, unlike other accounting relations, an independent relation exists for the surroundings; the change in entropy of the surroundings due to this process arises because of exchange with the system and also as a result of generation of entropy within the surroundings due to the process. Again, this generation of entropy must be positive; it can never be negative. Consequently, when the two entropy accounting expressions for the system and surroundings are summed, the total entropy change of the system plus surroundings is necessarily positive (or, in the limit of a reversible process, equal to zero).

These statements are embodied in an accounting of entropy for a macroscopic system

$$\sum\left[\left(\hat{S}\right)\varrho u_{\text{avg}}A\right]_{\text{in}}-\sum\left[\left(\hat{S}\right)\varrho u_{\text{avg}}A\right]_{\text{out}}+\sum\left(\frac{\dot{Q}}{T}\right)_{\text{system}}+\left(\dot{S}_{\text{gen}}\right)_{\text{system}}=\frac{\mathrm{d}}{\mathrm{d}t}\left[\left(\hat{S}\right)m\right]_{\text{system}} \tag{4.12}$$

and its surroundings

$$\sum\left[\left(\hat{S}\right)\varrho u_{\text{avg}}A\right]_{\text{out}}-\sum\left[\left(\hat{S}\right)\varrho u_{\text{avg}}A\right]_{\text{in}}+\sum\left(\frac{\dot{Q}}{T}\right)_{\text{surr}}+\left(\dot{S}_{\text{gen}}\right)_{\text{surr}}=\frac{\mathrm{d}}{\mathrm{d}t}\left[\left(\hat{S}\right)m\right]_{\text{surr}} \tag{4.13}$$

and for the system and surroundings combined

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\left(\hat{S}m\right)_{\text{system}}+\left(\hat{S}m\right)_{\text{surr}}\right]\geq 0 \tag{4.14}$$

In these equations, the exchange of entropy between the system and surroundings is the result of mass (possessing entropy per unit mass) transferring between the system and surroundings (values of $\hat{S}$ are averaged over the flow cross section).

Heat transfer is also a one-for-one-exchange between the system and its surroundings, provided the temperatures of the system and the surroundings at the boundary ($T_{\text{system, bound}}$ and $T_{\text{surr, bound}}$, respectively) are the same, this is normally the case (however, see the next paragraph concerning entropy generation). In this case, the rate of entropy added to the system ($\dot{S}_{\text{in, system}}$) due to a rate of heat transfer from the surroundings to the system $\dot{Q}$ is

$$\dot{S}_{\text{in, system}}=\frac{\dot{Q}}{T_{\text{system, bound}}} \tag{4.15}$$

and the rate of entropy transfer to the surroundings ($\dot{S}_{\text{in, surr}}$) due to the same heat transfer is

$$\dot{S}_{\text{in, surr}}=-\frac{\dot{Q}}{T_{\text{surr, bound}}} \tag{4.16}$$

If the temperatures of the system and surroundings at the boundary are the same, then these two terms cancel exactly and a true exchange of entropy occurs between the system and surroundings.

The generation of entropy in the system and surroundings is frequently difficult or impossible to quantify (although not always) but conceptually is the result of finite driving forces for mass, energy, or momentum transfer. Consequently, if temperature, mole fraction, or velocity gradients within the system or surroundings are dissipated as a result of heat, mass, or momentum transfer within the system, then these dissipation processes cause a movement toward a homogeneous condition which necessarily results in an increase in entropy. Likewise, if heat or momentum transfer at the system boundary results from finite differences in temperature or forces, then an entropy increase is associated with the heat or momentum transfer due to subsequent temperature or momentum equilibriation within the system and surroundings. For the system and surroundings of homogeneous but different temperatures (T_{system} and T_{surr}), the net total increase in entropy of the universe resulting from the two-step heat transfer and subsequent equilibration process is calculated according to

$$\frac{Q}{T_{\text{system}}}-\frac{Q}{T_{\text{surr}}}\geq 0 \tag{4.17}$$

as though a single-step process of transfer to the system (whose boundary temperature

is T_{system}) and from the surroundings (with boundary temperature T_{surr}) has occurred.

Again, because entropy is not conserved, an accounting of entropy for either the system or the surroundings alone is not an adequate statement of the second law of thermodynamics. Furthermore, the entropy of a material is a function of state and consequently can be calculated (with respect to an arbitrarily defined reference value) in terms of heat capacities, equations of state, and energy changes associated with phase changes. Consequently, the entropy exchange associated with mass crossing the system boundary is well defined, and the change of entropy of the system, given its composition and change in state, is also well defined. As mentioned previously, however, the generation of entropy as a result of irreversibilities is not always as well defined. In some cases, such as the diffusive mixing of two ideal gases at equal pressure and temperature, the entropy increase can be calculated from theoretical considerations. In other cases, however, such as the rapid expansion of a gas against an external force, the entropy increase due to irreversibilities cannot be calculated because the expansion is associated with a turbulent velocity profile that is unpredictable and, because of velocity gradients, ultimately results in an increase in entropy.

4.8. Tabular Summary of the Macroscopic Equations

The macroscopic equations given in Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, and 4.7 are summarized in Table 9. Also given are the number of independent scalar equations that are represented in each case and any limitations on the result.

4.9. Some Examples

4.9.1. Pressure Drops and Temperature Changes for Closed Channel Flow

Perhaps the most basic problem of macroscopic fluid mechanics is to calculate the pressure drop or pump requirement (or turbine output) associated with fluid flow through a pipe, tube, or other closed conduit. The basic design problem specifies a given amount of material that is to flow from one location to another at a given design flow rate. This requires either sufficient upstream kinetic, potential, and pressure energy or the addition of energy through a pump. Calculation of the required upstream energy or pump work is done by accounting for mechanical energy through the extended Bernoulli equation (Eq. 4.6 for a compressible fluid and Eq. 4.7 for an incompressible fluid). For example, to move water to an elevated storage tank, a pump must be supplied to overcome the change in potential energy and any frictional losses of mechanical energy due to the flow of water through the pipe. Kinetic energy changes may also be a factor, although these are usually relatively small.

One of the primary factors in such flow calculations is determining this mechanical energy loss term. This is not a calculation that can be made solely on theoretical grounds. Instead, it is handled through a dimensionless group correlation for a friction factor f from which the mechanical energy losses for flow at average velocity u_{avg} through a straight section of horizontal pipe of length L and diameter D are calculated according to

$$\hat{F} = 4f\left(\frac{L}{D}\right)\left(\frac{u_{\text{avg}}^2}{2}\right) \tag{4.18}$$

The dimensionless group correlation typically appears in graphic form (the original chart of MOODY [108] or its modifications) or in analytical form, convenient for computer use (given by CHURCHILL [109]). [Alternative definitions for the friction factor in the literature differ from each other by a factor of 2 or 4. Care must be taken that the dimensionless group correlation for f versus Re and the friction factor definition (e.g., Eq. 4.18) employed are consistent with each other.]

The friction factor depends on the Reynolds number of the flow through the pipe and, for turbulent flow, the roughness of the wall. Additional losses occur in valves, fittings, and pipe bends; relations for calculating these losses also are found in standard texts on fluid flow.

The effect of mechanical energy losses on the temperature of the fluid is normally quite small for liquids, although in some instances it may have to be taken into account. The thermal energy equation provides a basis for calculating this effect as discussed in Section 4.6. For

Table 9. Summary of the macroscopic equations

Equation name	Equation number	Number of independent macroscopic equations	Comments
Total mass	(4.1)	1	
Species mass (n species)	(4.2)	$n-1$	
Linear momentum	(4.4)	3∗	
Angular momentum	(4.5)	3∗	
Mechanical energy (compressible)	(4.6)	1∗∗	steady state, one flow stream
Mechanical energy (incompressible)	(4.7)	1∗∗	steady state, one flow stream
Total energy	(4.10)	1	
Thermal energy	(4.11)	0	steady state, one flow stream
Entropy, system	(4.12)	1	
Entropy, surroundings	(4.13)	1	
Total entropy (second law of thermodynamics)	(4.14)	0	an inequality

∗ One equation for each spatial dimension, e.g., for a two-dimensional situation, only two nontrivial equations exist. ∗∗ Only one of these two equations can be used in any one situation.

a compressible, single-phase fluid the change in enthalpy is written in terms of heat capacity and thermodynamic $p-\varrho-T$ properties by using standard thermodynamic arguments:

$$\hat{H}_2 - \hat{H}_1 = \int_{T_1}^{T_2} c_p \left(\partial T\right)_{p_1} + \int_{p_1}^{p_2} (1 - T\beta) \frac{(\partial p)_{T_2}}{\varrho} \tag{4.19}$$

where the temperature integral is at constant pressure, the pressure integral is at constant temperature, and β is the isobaric coefficient of thermal expansion $[\beta \equiv -1/\varrho\,(\partial\varrho/\partial T)_p]$. Therefore, by combining Equations (4.11) and (4.19), in the absence of heat transfer and for a temperature-independent heat capacity c_p, the net temperature change is obtained:

$$T_2 - T_1 = \frac{1}{c_p}\left[\hat{F} + \int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} - \int_{p_1}^{p_2} (1 - T\beta) \frac{(\partial p)_T}{\varrho}\right] \tag{4.20}$$

where $\hat{F}$ is given by Equation (4.18). If the fluid is truly incompressible, then $\beta = 0$, the two pressure integrals cancel exactly and the temperature increases in direct proportion to frictional losses. For compressible fluids, however, this heating is counteracted by the expansion of the fluid; if $p_2 < p_1$, then

$$\int_{p_1}^{p_2} \beta T/\varrho\,(\partial p)_T < 0$$

which, if large enough, can result in net cooling.

As an example, for the flow of oil through the Trans-Alaska pipeline between pumping stations, the appropriate parameters for this problem are $D = 1.19\,\mathrm{m}$, $L = 107\,\mathrm{km}$, $u_{\mathrm{avg}} A = 3.18\times10^5\,\mathrm{m^3/d}$, $\beta = 8.17\times10^{-4}\,\mathrm{K^{-1}}$, $f = 0.00335$, and $c_p = 1.8\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ [110], and the temperature rise is approximately

$$T_2 - T_1 \approx \frac{1}{c_p}\left[4f\left(\frac{L}{D}\right)\left(\frac{u_{\mathrm{avg}}^2}{2}\right)(1 - \beta T_1)\right] = 2.7\,\mathrm{K} \tag{4.21}$$

For comparison, neglecting the fluid compressibility results in a temperature increase of 3.7 K. Mechanical energy conversion to thermal energy in pumps, when present, also contributes to temperature increases.

Although the temperature increase can be fairly small for flow of a fluid over a short distance, it can accumulate and become appreciable if distances are long enough as in a transcontinental pipeline or a continuously circulating flow loop. Temperature changes accompanying flow can also contribute to errors in viscosity measurement.

4.9.2. Compressible Flow in a Tube

If the flow through a closed channel or tube is compressible, then accurate design and flow calculations can be considerably more complicated. The compressible form of the extended Bernoulli equation must be used (again this is

for steady-state flow situations), and the friction factor is calculated in the same way as for incompressible flow by using the friction factor correlations. However, the pressure term is shown as an integral from the entrance of the flow channel to the exit. The difficulty of this calculation is that the density must be known as a function of pressure through the pipe and along the actual flow path of the fluid through the pipe. Knowing the conditions at the entrance and exit is not enough; conditions must be known from one end to the other, along the path taken by the fluid. Normally, these are unknown because of the flow complexity.

To handle this situation, idealized approximations to the flow are calculated as limiting or extreme cases. For example, if the flow of an ideal gas ($p_1 M_r = RT_1 \varrho_1$, where R is the ideal gas constant and M_r the molecular mass) is adiabatic and reversible (no heat transfer occurs to or from the gas, and mechanical energy losses are negligible – *isentropic flow*), the pressure and density are related by

$$\frac{p}{\varrho^\kappa} = \text{const.} = \frac{p_1}{\varrho_1^\kappa} \tag{4.22}$$

where κ is the ratio of the heat capacities, c_p/c_v, so that

$$-\int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} = \frac{RT_1\,\kappa}{(\kappa-1)}\left[1-\left(\frac{p_2}{p_1}\right)^{(\kappa-1)/\kappa}\right] \tag{4.23}$$

A second situation is *isothermal flow* of an ideal gas for which

$$\frac{p}{\varrho} = \frac{RT}{M_r} = \text{constant} \tag{4.24}$$

and

$$-\int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} = \frac{RT}{M_r}\ln\frac{p_1}{p_2} \tag{4.25}$$

For the flow of a nonideal gas, the equation of state for that gas can be used instead of the ideal gas equation of state to obtain corresponding integrals.

For adiabatic, nonreversible flow, COULSON and RICHARDSON [111] give an approach for calculating the pressure integral and conclude that "the rate of flow of gas under adiabatic conditions is never more than 20 % greater than that obtained for the same pressure difference with isothermal conditions." Furthermore, they point out that the difference becomes less for long pipes and that "for pipes of length at least one-thousand diameters the difference does not exceed about 5 %." Thus, isothermal flow and adiabatic reversible flow are commonly assumed to bracket the actual observed situation, and for long pipelines, little difference exists between these two extremes.

One other aspect of compressible flow should be pointed out. For large pressure drops and high flow rates, a limit to the flow rate may be obtained even though the pressure drop is allowed to increase by lowering the downstream pressure. This limit is established by the sonic velocity of the gas in the pipe. If the gas reaches sonic velocity, any further reduction of the downstream pressure at constant pipe cross-sectional area results in no additional increase in velocity. The sonic velocity of the gas at the downstream end, based on thermodynamic principles and in terms of the density ϱ, pressure p, and specific entropy $\hat{S}$, is

$$u_{\text{sonic}} = \left[\left(\frac{\partial p}{\partial \varrho}\right)_{\hat{S}}\right]^{1/2} \tag{4.26}$$

4.9.3. Heterogeneous Reaction in a Fluidized-Bed Reactor

Consider now the reaction of component A to product B carried out in a well-mixed fluidized-bed reactor (Fig. 32). Such a conversion might represent the catalytic cracking of heavy hydrocarbons in a refinery process. The reactor stream is fed to the fluidized bed as a gas with sufficient velocity to maintain the catalyst particles in a fluidized state. Once inside the reactor, reacting species A is converted to product B at the catalyst surface. However, before this reaction can occur, the reactant must first migrate to the surface of the catalyst pellet. Then, after reaction at the surface, the product B must migrate to the bulk fluid to be removed from the reactor in the exit flow stream. What complicates this situation is that the catalyst itself may be porous, so that not only does the reactant contact the exterior surface of the catalyst, but part of the reactant species also diffuses into the micropores

before reacting. Consequently, some reactant is converted at a reduced concentration inside the pores and, therefore, at a lower rate than if it were converted at the concentration existing at the macroscopic exterior surface of the catalyst. Calculation of the effective reaction rate requires considering each of these mass-transfer and reaction rate processes.

To illustrate the concept, the macroscopic material balance of reacting species A is considered. Under the assumption of a well-mixed fluidized bed, the gas-phase concentration is the same everywhere in the reactor and equal to the concentration in the exit flow stream. Of course, this model is an approximation because concentration gradients exist near the catalyst pellets due to the heterogeneous reaction. The model assumes that enough reaction locations of sufficiently small size exist for them to be treated like a homogeneous reaction occurring uniformly throughout the system volume. An accounting of species A within the fluidized bed in terms of its concentration in the feed stream c_{Ain}, the concentration in the bulk fluid in the bed and outlet flow stream c_{A}, and the feed and product volumetric flow rates (F_{in} and F_{out}, respectively) is

$$c_{\mathrm{Ain}} F_{\mathrm{in}} - c_{\mathrm{A}} F_{\mathrm{out}} = r_{\mathrm{A}} V \tag{4.27}$$

The amount of A that enters the reactor at concentration c_{Ain} either leaves at concentration c_{A} or is converted to products at an effective reaction rate r_{A} per unit volume.

Although complicated by r_{A}, this result provides a clear approach to reactor design. Once an expression for this reaction rate is obtained based on transport and kinetic factors, the problem can be completed to obtain the conversion in the reactor. Then economic and design calculations can be made for the reactor in terms of fractional conversion, space time, and reactor volume and flow rates [112], [113]. The key to reactor design is to determine the overall reaction rate for a given situation. The effect of transport on this effective rate is as follows.

The transport of species from the bulk of the fluid in the fluidized bed to the catalyst surface can be estimated by using mass-transfer coefficient correlations for fluidized beds. A j factor correlation has been presented by CHU et al. [114] that is applicable to gas – solid and liquid – solid systems and for both fluidized-bed and fixed-bed reactors; this is presented in Table 8. Graphs of these correlations are given in [113], [114].

For porous catalysts, determination of the reaction rate at the catalyst, although primarily a problem in reaction kinetics, can also be a problem in transport phenomena because of diffusion in the catalyst pores. The chemical kinetics problem must be determined from experimental laboratory reactors. LIN provides extensive literature references to studies of a range of industrial reactions [113]. The transport problem is normally addressed by the use of an *effectiveness factor* η. The effectiveness factor adjusts the reaction rate to account for the fact that within the pores of a catalyst the reaction occurs at a reduced rate compared to the pellet's exterior surface. At the catalyst surface, the reaction proceeds at a rate that is dictated by the chemical kinetics and by the concentration of the reacting species at the surface. However, as the reactant becomes depleted with greater depth into the pores, the reaction rate decreases accordingly; the effectiveness of the catalyst surface within the pores is lower than that at the surface.

Calculation of an average (averaged over the entire catalyst surface, whether deep inside pores or not) effectiveness factor is based on combined diffusion and reaction inside the catalyst pores and was addressed originally by THIELE [115] and ARIS [116] and subsequently in many references on reaction and transport [98], [112], [117]. Such analysis leads to calculation of an average effectiveness factor for a catalyst pellet as a function of a dimensionless variable called the *Thiele modulus*. This parameter depends on the relative rates of the kinetic reaction to diffusion. If the diffusion rate into the pores is very high compared to the reaction constant, then the fact that some of the catalyst surface is available only in pores makes little difference and the effectiveness factor is close to unity. However, if the kinetic reaction rate constant is large compared to the diffusion rate, then the concentrations, and hence reaction rates, within the pores are reduced considerably, giving a substantially lower effectiveness factor.

By using the effectiveness factor η to adjust for the reaction rate at the catalyst, the reaction rate (per unit volume of catalyst particle) for a reaction that is first-order irreversible in A with rate constant k is given by

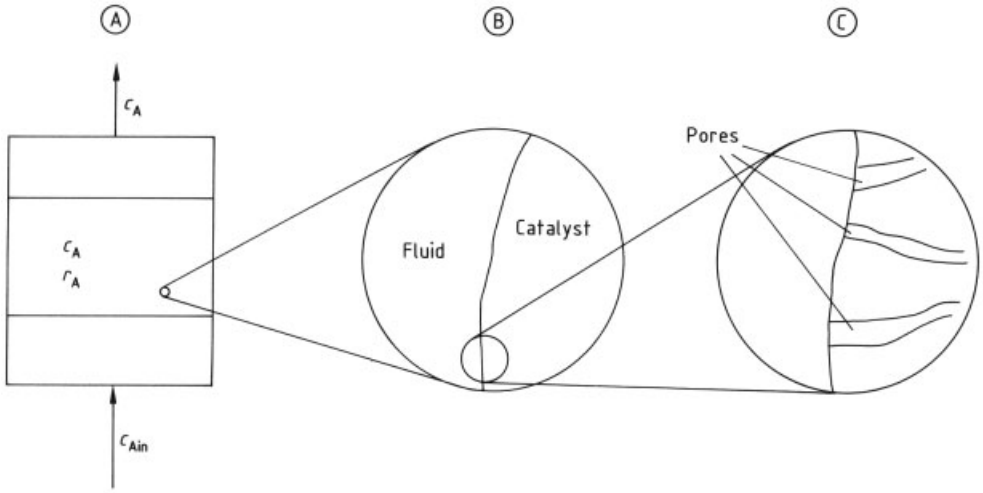

Figure 32. Modeling of the fluidized-bed reactor: macro-scale for the reactor accounting of mass (A), mesoscale for the mass-transfer rate (B), and microscale for definition and calculation of the effectiveness factor (C)
A) Macroscale: fluidized-bed region is modeled as well mixed and uniform in temperature, composition, and reaction rate; B) Mesoscale: mass transfer occurs between the bulk fluid and the catalyst surface because of concentration differences; C) Microscale: diffusion within the catalyst pores reduces the effective reaction rate

$$r_A = \eta k c_{Ai} \tag{4.28}$$

If a steady state exists at the surface of the catalyst (i.e., no accumulation of materials occurs at the catalyst surface), the transport rate of materials to the pellet equals the reaction rate at the pellet. For species A then (by using the mass-transfer coefficients of Section 3.3),

$$k_G S_p (c_A - c_{Ai}) = \eta k V_p c_{Ai} \tag{4.29}$$

where the external pellet surface area S_p and the pellet volume V_p are required because k_G is per surface area and k is per pellet volume. For a true first-order kinetic reaction, the rate constant is a function of temperature but not of species concentrations. If the reaction is not first order, an additional functionality depending on the interface concentration exists. This equation can be solved for the interface concentration in terms of the rate constants and the bulk fluid concentrations, which can then be used to express the overall reaction rate of species A in terms of the bulk phase concentration

$$r_A = K c_A \tag{4.30}$$

where

$$K = \frac{1}{\frac{1}{\eta k (V_p/S_p)} + \frac{1}{k_G}} \tag{4.31}$$

This shows that both the mass-transfer coefficient to the catalyst and the effectiveness factor – rate constant product act as resistances to the conversion of the reactant to the product. When the resistances are added in series, a combined rate constant is obtained.

This result leads to the notion of transport-limited versus kinetic-limited reactions. If the mass-transfer coefficient is small enough, mass transfer is the rate limiting step and reaction kinetics play no role in the overall conversion rate. Likewise, if the kinetic rate constant – effectiveness factor product is small enough (compared to the mass-transfer coefficient), it becomes the limiting part of the rate process and the mass-transfer coefficient plays no role in establishing the reaction rate. If both rate constants are of the same order, then the overall rate coefficient depends on both the kinetic and the mass-transfer rate factors.

5. References

1. R. Aris: *Vectors, Tensors, and the Basic Equations of Fluid Mechanics,* Prentice Hall, Englewood Cliffs, N.J., 1962.

2. L. Brand: *Vector and Tensor Analysis,* J. Wiley & Sons, New York 1957.
3. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
4. S. Whitaker: *Fundamental Principles of Heat Transfer,* Pergamon Press, New York 1977.
5. J. C. Slattery: *Momentum, Energy and Mass Transfer in Continua,* 2nd ed., R. E. Krieger Publishing Co, Huntington, N.Y. 1981.
6. L. E. Malvern: *Introduction to the Mechanics of a Continuous Medium,* Prentice Hall, Englewood Cliffs, N.J., 1969.
7. G. E. Mase: *Theory and Problems of Continuous Mechanics,* McGraw-Hill, New York 1970.
8. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
9. S. Chapman, T. G. Cowling: *The Mathematical Theory of Nonuniform Gases,* Cambridge University Press, New York 1939.
10. J. O. Hirschfelder, C. F. Curtiss, R. B. Bird: *Molecular Theory of Gases and Liquids,* Wiley Interscience, New York 1954.
11. P. D. Neufeld, A. R. Janzen, R. A. Aziz, *J. Chem. Phys.* **57** (1972) 1100.
12. R. S. Brokaw, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 240.
13. C. Truesdell: *The Elements of Continuum Mechanics,* Springer Verlag, New York 1966.
14. R. Darby: *Viscoelastic Fluids,* Marcel Dekker, New York 1976.
15. W. R. Schowalter: *Mechanics of Non-Newtonian Fluids,* Pergamon Press, New York 1978.
16. S. Whitaker: *Fundamental Principles of Heat Transfer,* Pergamon Press, New York 1977.
17. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
18. A. J. Chapman: *Heat Transfer,* Macmillan Publ. Co., New York 1960.
19. H. S. Carslaw, J. C. Jaeger: *Conduction of Heat in Solids,* 2nd ed., Oxford University Press, London 1959.
20. L. Brand: *Vector and Tensor Analysis,* J. Wiley & Sons, New York 1947.
21. J. H. Lienhard: *A Heat Transfer Textbook,* Prentice Hall, Englewood Cliffs, N.J., 1987.
22. H. P. Gurney, J. Lurie, *Ind. Eng. Chem.* **15** (1923) 1170.
23. H. C. Groeber, *VDI-Z.* **69** (1925) 705.
24. A. Shack: *Industrial Heat Transfer,* J. Wiley & Sons, New York 1933.
25. M. P. Heisler, *Trans. ASME* **69** (1947) 227 – 236.
26. J. R. Welty, C. E. Wicks, R. E. Wilson: *Fundamentals of Momentum, Heat and Mass Transfer,* J. Wiley & Sons, New York 1976.
27. L. E. Sissom, D. R. Pitts: *Elements of Transport Phenomena,* McGraw-Hill, New York 1972.
28. F. W. Kroesser, S. Middleman, *Polym. Eng. Sci.* **5** (1965) 1.
29. R. W. Flumerfelt, M. W. Pierick, S. L. Cooper, R. B. Bird, *Ind. Eng. Chem. Fundam.* **8** (1969) 354.
30. J. Happel, H. Brenner: *Low Reynolds Number Hydrodynamics,* Prentice Hall, Englewood Cliffs, N.J., 1965.
31. S. M. Selby: *Standard Mathematical Tables,* 16th ed., The Chemical Rubber Company, Cleveland 1968.
32. R. B. Bird, R. C. Armstrong, O. Hassager: *Dynamics of Polymeric Liquids,* vol. 1, *Fluid Mechanics,* J. Wiley & Sons, New York 1977.
33. A. B. Metzner, J. C. Reed, *AIChE J.* **1** (1955) 434.
34. A. H. P. Skelland: *Non-Newtonian Fluid and Heat Transfer,* J. Wiley & Sons, New York 1967.
35. D. W. Dodge, A. B. Metzner, *AIChE J.* **5** (1959) 189.
36. S. Goldstein: *Modern Developments in Fluid Dynamics,* Oxford University Press, London 1938.
37. S. Prasad: *Non-Newtonian Flow Through Constricted Geometries,* M. S. Thesis, University of Houston, Houston, TX, 1978.
38. T. B. Drew: *Trans. Am. Inst. Chem. Eng.* **26** (1931) 26.
39. H. L. Langhaar, *J. Appl. Mech.* **64** (1942) A-55.
40. W. M. Kays, *Trans. ASME* **77** (1955) 1265.
41. J. R. Sellars, M. Tribus, J. S. Klein, *Trans. ASME* **78** (1956) 441.
42. E. N. Sieder, G. E. Tale, *Ind. Eng. Chem.* **28** (1936) 1429.
43. H. Hausen, *Verfahrenstechnik (Berlin)* **4** (1943) 91.
44. H. Schlichting: *Boundary Layer Theory,* 6th ed., McGraw-Hill, New York 1968.
45. L. M. Milne-Thomson: *Theoretical Hydrodynamics,* 5th ed., Macmillan, New York 1968.
46. J. Schetz: *Foundations of Boundary Layer Theory,* Prentice-Hall, Englewood Cliffs, N.J., 1984.

47. E. Pohlhausen, *Z. Angew. Math. Mech.* **1** (1921) 115.
48. S. W. Churchill, H. H. S. Chu, *Int. J. Heat Mass Transfer* **18** (1975) 1323.
49. P. S. Klebanoff: "Characteristics of Turbulence on a Boundary Layer with Zero Pressure Gradient," NACA Report 1247, 1955.
50. J. Nikuradse, *Ing. Arch.* **1** (1930) 150; *VDI-Forschungsh.* **361** (1933) 1; *Pet. Eng.* **11** (1940) 164; **11** (1940) 75; **11** (1940) 124; **11** (1940) 38; **11** (1940) 83.
51. L. Prandtl, *Z. Angew. Math. Mech.* **5** (1925) 136.
52. J. M. Kay, R. W. Nedderman: *An Introduction to Fluid Mechanics and Heat Transfer,* 3rd ed., Cambridge University Press, London 1974.
53. H. Blasius, *VDI-Forschungsh.* **131** (1913).
54. C. F. Colebrook, *J. Inst. Civ. Eng. (London)* **133** (1938 – 1939).
55. A. A. Zhukauskas, A. B. Ambrazyavichyus: NACA Report 909, 1949.
56. L. Prandtl: " Über den Reibungswiderstand strömender Luft," Reports of the Aerol. Versuchsanst. Göttingen, 3rd Series, 1927; see also: "Zur turbulenten Strömung in Rohren und Längsplatten," Reports of the Aerol. Versuchsanst. Göttingen, 4th Series, 1931.
57. W. L. Friend, A. B. Metzner, *AIChE J.* **4** (1958) 393.
58. S. W. Churchill, M. Bernstein, *J. Heat Transfer* **99** (1977) 300.
59. S. Whitaker, *AIChE J.* **18** (1972) 361.
60. S. W. Churchill, H. H. S. Chu, *Int. J. Heat Mass Transfer* **18** (1975) 1323.
61. J. P. Holman: *Heat Transfer,* 6th ed., McGraw-Hill, New York 1986.
62. T. Fujii, H. Imura, *Int. J. Heat Mass Transfer* **15** (1972) 755.
63. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
64. J. C. Slattery: *Momentum, Energy and Mass Transfer in Continua,* 2nd ed., R. E. Krieger Publishing Co, Huntington, N.Y. 1981.
65. E. L. Cussler: *Diffusion Mass Transfer in Fluid Systems,* Cambridge University Press, New York 1984.
66. E. L. Cussler: *Multicomponent Diffusion,* Elsevier Scientific Publishing Company, New York 1976.
67. J. Crank: *The Mathematics of Diffusion,* Clarendon Press, Oxford 1956.
68. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
69. H. S. Carslaw, J. C. Jaeger: *Conduction of Heat in Solids,* Oxford University Press, London 1959.
70. P. D. Neufeld, A. R. Janzen, R. A. Aziz, *J. Chem. Phys.* **57** (1972) 1100.
71. R. S. Brokaw, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 240.
72. E. N. Fuller, P. D. Schettler, J. C. Giddings, *Ind. Eng. Chem.* **58** (1966) 18.
73. C. R. Wilke, P. Chang, *AIChE J.* **1** (1955) 264.
74. J. R. Welty, C. E. Wicks, R. E. Wilson: *Fundamentals of Momentum, Heat, and Mass Transfer,* 2nd ed., J. Wiley & Sons, New York 1976.
75. J. M. Coulson, J. F. Richardson: *Chemical Engineering,* 3rd ed., vol. 1, Pergamon Press, Oxford 1977.
76. A. S. Foust et al.: *Principles of Unit Operations,* 2nd ed., J. Wiley & Sons, New York 1980.
77. W. L. McCabe, J. C. Smith, P. Harriott: *Unit Operations of Chemical Engineering,* 4th ed., McGraw-Hill, New York 1985.
78. R. E. Treybal: *Mass-Transfer Operations,* 3rd ed., McGraw-Hill, New York 1980.
79. C. O. Bennett, J. E. Myers: *Momentum, Heat, and Mass Transfer,* 2nd ed., McGraw-Hill, New York 1974.
80. W. G. Whitman, *Chem. Metall. Eng.* **29** (1923) 147.
81. R. Higbie, *Trans. Am. Inst. Chem. Eng.* **31** (1935) 365.
82. P. V. Danckwerts, *Ind. Eng. Chem.* **43** (1951) 1460.
83. P. Harriott, *Chem. Eng. Sci.* **17** (1962) 149.
84. H. L. Toor, J. M. Marchello, *AIChE J.* **4** (1958) 97.
85. A. L. Hines, R. N. Maddox: *Mass Transfer, Fundamentals and Applications,* Prentice Hall, Englewood Cliffs, N.J., 1985.
86. W. H. Linton, T. K. Sherwood, *Chem. Eng. Prog.* **46** (1950) 258.
87. E. R. Gilliland, T. K. Sherwood, *Ind. Eng. Chem.* **26** (1934) 516.
88. P. Harriott, R. M. Hamilton, *Chem. Eng. Sci.* **20** (1965) 1073.
89. M. J. Christian, S. P. Kezibs, *AIChE J.* **5** (1959) 61.
90. C. H. Bedingfield, I. B. Drew, *Ind. Eng. Chem.* **42** (1950) 1164.
91. R. L. Steinberger, R. E. Treybal, *AIChE J.* **6** (1960) 227.
92. F. H. Garner, R. D. Sucking, *AIChE J.* **4** (1958) 114.

93. E. J. Wilson, C. J. Geankoplis, *Ind. Eng. Chem. Fundam.* **5** (1966) 9.
94. A. Sen Gupta, G. Thodos, *AIChE J.* **9** (1963) 751.
95. P. N. Dwivedi, S. N. Upadhyay, *Ind. Eng. Chem. Process Des. Dev.* **16** (1977) 157.
96. A. Sen Gupta, G. Thodos, *AIChE J.* **8** (1962) 608.
97. J. C. Chu, J. Kalil, W. A. Wetteroth, *Chem. Eng. Prog.* **49** (1953) 141.
98. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
99. R. S. Brodkey, H. C. Hershey: *Transport Phenomena, A Unified Approach,* McGraw-Hill, New York 1988.
100. J. R. Welty, C. E. Wicks, R. E. Wilson: *Fundamentals of Momentum, Heat, and Mass Transfer,* 2nd ed., J. Wiley & Sons, New York 1969.
101. O. A. Hougen, K. M. Watson, R. A. Ragatz: *Chemical Process Principles,* part II, *Thermodynamics,* 2nd ed., J. Wiley & Sons, New York 1959.
102. J. M. Smith, H. C. Van Ness: *Introduction to Chemical Engineering Thermodynamics,* 3rd ed., McGraw-Hill, New York 1975.
103. W. L. McCabe, J. C. Smith, P. Harriott: *Unit Operations of Chemical Engineering,* 4th ed., McGraw-Hill, New York 1985.
104. F. M. White: *Fluid Mechanics,* 2nd ed., McGraw-Hill, New York 1986.
105. S. Whitaker in N. A. Peppas (ed.): *One Hundred Years of Chemical Engineering,* Kluwer Academic Publishers, Dordrecht 1989, pp. 47 – 109.
106. J. C. Slattery: *Momentum, Energy, and Mass Transfer in Continua,* 2nd ed., R. E. Krieger Publishing Co, Huntington, N.Y. 1981.
107. R. B. Bird, *Chem. Eng. Sci.* **6** (1957) 123.
108. L. W. Moody, *Trans. ASME* **66** (1944) 672.
109. S. W. Churchill, *Chem. Eng. (N.Y.)* **87** (1977) Nov. 7, 91.
110. P. R. Hooker, W. E. Brigham, *JPT J. Pet. Technol.* **30** (1978) 747.
111. J. M. Coulson, J. F. Richardson: *Chemical Engineering,* 3rd ed., vol. 1, Pergamon Press, Oxford 1977.
112. O. Levenspiel: *Chemical Reaction Engineering,* J. Wiley & Sons, New York 1962.
113. K.-H. Lin in R. H. Perry, C. H. Chilton (eds.): *Chemical Engineer's Handbook,* 5th ed., section 4, McGraw-Hill, New York 1973.
114. J. C. Chu, J. Kalil, W. A. Wetteroth, *Chem. Eng. Prog.* **49** (1953) 141.
115. E. W. Thiele, *Ind. Eng. Chem.* **31** (1939) 916.
116. R. Aris, *Chem. Eng. Sci.* **6** (1957) 262.
117. C. D. Holland, R. G. Anthony: *Fundamentals of Chemical Reaction Engineering,* Prentice Hall, Englewood Cliffs, N.J., 1979.

Fluid Mechanics

David V. Boger, Department of Chemical Engineering, The University of Melbourne, Parkville, Victoria 3052, Australia

Y. Leong Yeow, Department of Chemical Engineering, The University of Melbourne, Parkville, Victoria 3052, Australia

1. Introduction 3
2. Basic Equations of Fluid Mechanics 5
2.1. Continuity Equation 5
2.2. Cauchy Equations of Motion 6
2.3. Energy Transport Equation and Bernoulli Equation 8
2.4. Constitutive Equations and Classification of Fluids 9
3. Newtonian Fluids 10
3.1. Deviatoric Stress and Viscosity . . . 10
3.2. Navier – Stokes Equations 11
3.3. Applications of Navier – Stokes Equations 12
3.3.1. Flow in Pipes 12
3.3.2. Concentric Cylinder Flow 14
3.3.3. Creeping Flow Past a Sphere 16
3.4. Some Other Important Flows 18
3.4.1. Flow Through Granular Beds 18
3.4.2. Fluidization 19
3.4.3. Gas – Liquid Flow 20
3.5. Mechanical Energy Balance for Macroscopic Systems 22
3.5.1. Fully Developed Tube Flows 23
3.5.2. Accelerating and Decelerating Flows 23
3.5.3. The Orifice Plate and Similar Flow Rate Measurement Devices 25
4. Non-Newtonian Fluids 26
4.1. Classification of Fluids According to Viscosity Behavior 27
4.1.1. Bingham Type Behavior 27
4.1.2. Shear-Thinning (Pseudoplastic) Fluids 28
4.1.3. Shear-Thickening (Dilatant) Fluids . . 29
4.1.4. Fluids with Time-Dependent Viscosity (Thixotropic Fluids) 29
4.2. Fully Developed Tube Flow 30
4.2.1. Volumetric Flow Rate – Pressure Drop Relationship 31
4.2.2. Generalized Treatment 31
4.2.3. Velocity Distribution 32
4.2.4. Friction Factor – Reynolds Number Relationships 33
4.3. Viscoelastic Fluid Mechanics 34
4.3.1. Steady Shear Behavior of Viscoelastic Fluids 36
4.3.2. Behavior of Viscoelastic Fluids in Oscillatory Shear Flow 37
4.3.3. Examples of Constitutive Equations for Viscoelastic Fluids 38
4.3.4. Extensional Behavior of Viscoelastic Fluids 40
4.3.5. Accelerating and Decelerating Flows of Viscoelastic Fluids 42
5. Numerical Methods in Fluid Mechanics 46
5.1. Finite Difference Method 46
5.2. Finite Element Method 48
5.3. General Remarks 51
6. References 51

Symbols:

A	cross-sectional area, surface area, amplitude of oscillation
C	loss coefficient
C_{d}	drag coefficient, discharge coefficient
d	distance
D	diameter
D_{e}	extrudate diameter
D_{hy}	hydraulic diameter
D_{p}	particle diameter
D_{v}	vane diameter
e	surface roughness, end correction
$\boldsymbol{e}$	rate of strain tensor
$e_{\mathrm{xx}}, e_{\mathrm{xy}}, e_{\mathrm{ij}}$	components of rate of strain tensor
Ev	frictional loss
Ev_{ex}	losses in valves, fittings, etc.
Ev_{fd}	losses in fully developed flows
f	friction factor
$\boldsymbol{F}$	force vector
F_{d}	drag force
F_i	component of force vector
G	volumetric flow rate in gas phase
G'	storage modulus
G''	loss modulus

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_05

h elevation
H gap between plates
H_{v} vane height
$\boldsymbol{i}$ unit vector in x direction
$\boldsymbol{I}$ identity matrix
$\boldsymbol{j}$ unit vector in y direction
J_{G} superficial gas phase velocity
J_{L} superficial liquid phase velocity
k thermal conductivity, parameter in the power-law equation
$\boldsymbol{k}$ unit vector in z direction
K' intercept in the log – log plot of τ_{w} versus $8\,V/D$
L length, volumetric flow rate in liquid phase
L_{e} equivalent length
L_0 initial length
n flow-behavior index of power law equation
$\boldsymbol{n}$ unit outward pointing normal
n' slope of log – log plot of τ_{w} versus $8\,V/D$
N_1 first normal stress
p isotropic pressure
$\boldsymbol{q}$ heat flux vector
Q volumetric flow rate
r radial coordinate
r_0 plug radius
$R_{\mathrm{c}}, R_{\mathrm{x}}, R_{\mathrm{y}}$ residuals
R, R_1, R_2 radius
Re Reynolds number
Re' generalized Reynolds number for time-independent fluid
S path joining two points
t time
T temperature
$\boldsymbol{T}$ stress tensor
$T_{ij}, T_{xx}, T_{xy}, T_{xz}$ stress components
$\boldsymbol{u}$ velocity vector
u_{f} fluidizing velocity
$u_i, u_x, u_y, u_r, u_{\mathrm{t}}, u_0$ velocity components
U, U_0, U_∞ velocity
V average velocity, volume
W mass flow rate
W_{s} shaft work
$\boldsymbol{x}$ position vector
x_i components of position vector
z elevation
α parameter in Meter model
γ shear strain
γ_0 amplitude of sinusoidal shear strain
$\dot{\gamma}$ shear rate
$\dot{\gamma}_0$ amplitude of sinusoidal shear rate
$\Gamma, \Gamma_{\mathrm{m}}$ torque
δ_{ij} Kronecker delta
ε porosity
$\dot{\varepsilon}_0$ constant extensional rate
η viscosity
η_{el} extensional viscosity
η_{pl} plastic viscosity
η_0 viscosity parameter of Meter, Maxwell, and Oldroyd-B models
η' dynamic viscosity
η'' component of complex viscosity
θ azimuthal coordinate
λ bulk viscosity, time constant
λ_{el} extensional viscosity
λ_1 relaxation time
λ_2 retardation time
ϱ density
ϱ_{p} particle density
$\boldsymbol{\tau}$ deviatoric stress tensor
τ_{ij} component of deviatoric stress tensor
$\tau_{\mathrm{rz}}, \tau_{\mathrm{r}\theta}$ shear stress
τ_{w} wall shear stress
τ_y yield stress
$\tau_{1/2}$ parameter of Ellis and Meter model
Φ interpolation function
ψ_1 first normal stress coefficient
ψ_2 second normal stress coefficient
Ψ stream function
$\omega_1, \omega_2, \omega$ angular velocity, vorticity
Ω angular velocity

1. Introduction

The traditional approach for writing an article or a text book on fluid mechanics has been to deal with the Newtonian fluid and the resulting mechanics associated with such materials. Thus the basic tools (in a sequence of events such as those illustrated in Figure 1) are treated in detail to illustrate how to solve a Newtonian fluid mechanics problem. The laws of conservation of mass (continuity equation) and conservation of momentum (Cauchy momentum equations or equations of motion) are established for a general system and in multidimensional form. From these equations it then becomes clear that the solution to any fluid mechanics problem is not possible unless a relationship is developed between the stress tensor and rate of deformation tensor, or, between the stresses and velocity gradients in the system. Such an equation is termed

the constitutive equation. For Newtonian fluids the relationship between the stress tensor and the rate of deformation is linear, the proportionality constant is the viscosity. Substitution of this relationship into the Cauchy momentum equations leads to the Navier – Stokes equations. Thus the Navier – Stokes equations, in conjunction with the viscosity, form the basis for the solution of classical fluid mechanics problems. Much of the effort in classical fluid mechanics is devoted to establishing methods of solution of the Navier – Stokes equations. These methods can be analytical in nature and in recent times, of course, are numerical. While one whole section of this article is devoted to numerical methods because of their importance in modern fluid mechanics, no attention is directed towards classical analytical methods such as those used in the solution of boundary layer and slow flow (creeping flow) problems. The emphasis is directed towards the industrial chemist, process scientist, and chemical engineer who encounter a vast range of materials in today's industries where the Newtonian fluid is only one small subset of a vast array of fluid behavior which can be observed. Figure 2 illustrates the range of behavior that can be encountered in terms of the viscosity of fluids and in terms of whether materials are classified as being inelastic or viscoelastic fluids. It is clear from examination of Figure 2 that a vast range of behavior in terms of the viscosity of materials can be and is observed. An industrial chemist or process engineer will encounter low molecular mass liquids and gases (Newtonian materials) but will also be dealing with polymer solutions, polymer melts, rubbers, mineral suspensions, food products, pharmaceuticals, and energy products such as coal oil and coal water fuels, etc. In such systems the viscosity can be constant, but more likely will vary with shear rate or rate of flow. Viscosity can be a function of shear rate and time of shear; or it can be a function of shear rate, time, and thermal history and each of these viscosity behaviors can be observed in inelastic and/or viscoelastic fluids. In addition many materials will not flow until a certain stress is exceeded (e.g., toothpaste). This stress, termed yield stress, can be present in both inelastic and viscoelastic fluids. Therefore, it is important to be aware of the vast range of material behavior present in industry.

In Chapter 2 the basic equations for fluid mechanics – the continuity equation, the Cauchy equations of motion, and the energy transport and Bernoulli equations – are established. Chapter 3 deals with Newtonian fluids and Newtonian fluid mechanics where the Navier – Stokes equations are developed. In addition, the mechanical energy balance is established for macroscopic systems in order that pipe flow systems can be designed. The emphasis in the article is directed toward pipe flows because the process engineer is more concerned with flow in conduits than in other geometries. Little emphasis is placed on the details associated with turbulence because process engineers are primarily concerned with turbulent flow in pipes of Newtonian fluids and are not concerned with design of aerofoil structures, etc.

Chapter 4 concentrates on non-Newtonian fluids. In this chapter fluids are classified according to their viscosity behavior and are examined in some depth in fully developed tube flow. Viscoelastic fluid mechanics are also summarized – the steady shear and dynamic shear behavior of viscoelastic fluids are reviewed and examples of simple constitutive equations are given. The article is completed with a brief examination of numerical methods in fluid mechanics.

The emphasis of this article is shifted toward non-Newtonian fluid mechanics, the sequence of events and tools of importance are illustrated in Figure 3. Here the basic conservation of mass (continuity equation) and the conservation of momentum (Cauchy momentum equations) remain the equations of fundamental importance, but, the relevant constitutive equation in most cases is still unknown. Thus the Navier – Stokes equivalent set of equations in many cases is not known and because any number of constitutive equations are available it is not entirely clear what basic flow property information is required for the most general viscoelastic material. It is only now (1990) that fluid mechanics problems are solved for the first time for viscoelastic fluids. The solutions evolving now for such problems will be extremely important in the future. It is also of considerable importance to keep abreast of the numerical software packages which have been developed in this area particularly in regard to the processing of polymers [1], [2].

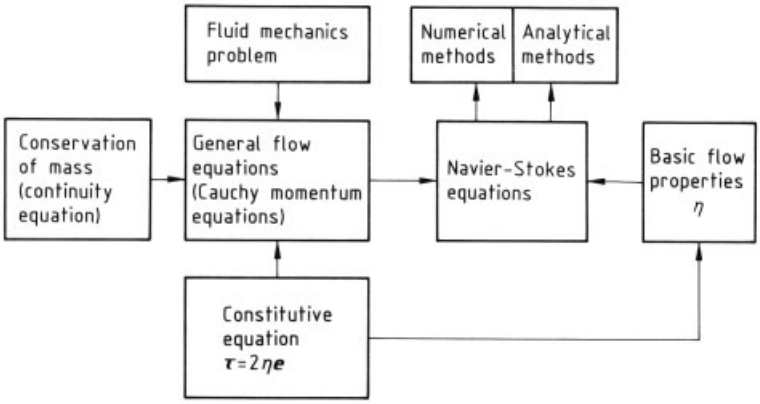

Figure 1. Newtonian fluid mechanics

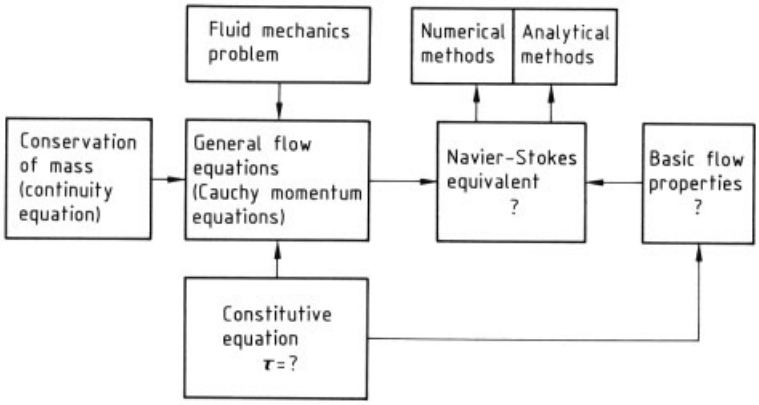

Figure 2. Schematic classification of fluid behavior

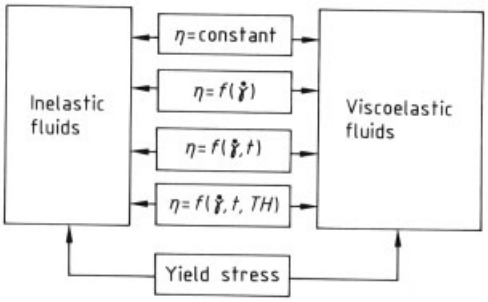

Figure 3. Non-Newtonian fluid mechanics
TH = Thermal history

2. Basic Equations of Fluid Mechanics

The basic equations of fluid mechanics can be derived from a small number of fundamental physical laws such as conservation of mass, of momentum (Newton's second law), and of energy. Although very general statements of these laws can be written down (applicable to all substances, solids as well as fluids), in fluid mechanics these laws are formulated in terms of physical variables such as velocity $\boldsymbol{u}$, pressure p, temperature T, and density ϱ. The resulting statements may be less general but they are more directly applicable in analyzing fluid mechanical problems.

2.1. Continuity Equation

The requirement that mass be conserved at every point in a flowing fluid imposes certain restrictions on the velocity $\boldsymbol{u}$ and density ϱ. Consider an arbitrary region in space, of volume V, bounded by surface A, through which the fluid flows (see Fig. 4). The rate of *increase in mass* within this region is given by the volume integral

$$\int_V \frac{\partial \varrho}{\partial t} \mathrm{d}V \tag{1}$$

The rate at which fluid is *leaving the region* is given by the surface integral of the outward pointing normal velocity multiplied by the fluid's density taken over the surface A

$$\int_A \varrho \boldsymbol{u} \cdot \boldsymbol{n} \mathrm{d}A = \int_V \nabla \cdot (\varrho \boldsymbol{u}) \, \mathrm{d}V \tag{2}$$

where $\boldsymbol{n}$ is the unit outward pointing normal at any point on A. The divergence theorem has been used to convert the surface integral into a volume integral. Mass conservation requires the sum of the rate of accumulation in V and the rate of outflow from V to be zero:

$$\int_V \left[\frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \boldsymbol{u}) \right] \mathrm{d}V = 0 \tag{3}$$

Since V is arbitrary, the integrand itself must vanish at every point in space:

$$\frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \boldsymbol{u}) = 0 \tag{4}$$

Equation (4) is usually referred to as the *continuity equation*. It is a statement of mass conservation in terms of fluid velocity and density. Expanding the second term and regrouping, the equation can be written as

$$\frac{\partial \varrho}{\partial t} + \boldsymbol{u} \cdot \nabla \varrho + \varrho \nabla \cdot \boldsymbol{u} = 0 \tag{5}$$

or

$$\frac{\mathrm{D}\varrho}{\mathrm{D}t} + \varrho \nabla \cdot \boldsymbol{u} = 0 \tag{6}$$

where $\mathrm{D}/\mathrm{D}t$ is used to represent the operator $\frac{\partial}{\partial t} + \boldsymbol{u} \cdot \nabla$.

It is known as the substantial derivative operator. By following the change in the density $\varrho(\boldsymbol{x}, t)$ of a fluid element occupying position $\boldsymbol{x}$ at time t and its density $\varrho(\boldsymbol{x} + \delta \boldsymbol{x}, \mathrm{t} + \delta t)$ at a later time $t + \delta t$ when it is occupying position $\boldsymbol{x} + \delta \boldsymbol{x}$, it can be seen that $\mathrm{D}\varrho/\mathrm{D}t$ is the rate of change in density experienced by a fluid element as it moves about in space. $\mathrm{D}\varrho/\mathrm{D}t$ should not be confused with $\partial \varrho/\partial t$ which is the rate of change in density at a fixed point in space. The substantial derivative operator can also be applied to the velocity $\boldsymbol{u}$, temperature T, or other properties associated with a fluid element with a similar physical interpretation of the result.

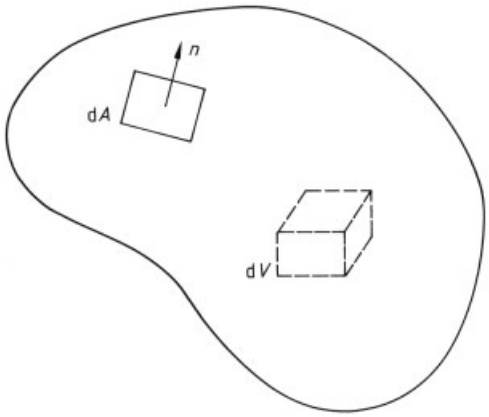

Figure 4. Fluids flowing through an arbitrary region in space of volume V and surface area A

A number of special cases of the continuity equation are frequently encountered. For example, under steady-state conditions the equation reduces to

$$\nabla \cdot (\varrho \boldsymbol{u}) = 0 \tag{7}$$

Much of the subsequent analyses will be restricted to fluids that are incompressible and homogeneous. For such fluids the continuity equation takes on a particularly simple form

$$\nabla \cdot \boldsymbol{u} = 0 \tag{8}$$

In many practical problems the fluid flow can be regarded as approximately two-dimensional. For such problems, Equation 8 (in cartesian coordinates) takes on the form

$$\frac{\partial u_x(x,y)}{\partial x} + \frac{\partial u_y(x,y)}{\partial y} = 0 \tag{9}$$

Equation (9) and, hence, mass conservation are identically satisfied if the velocity components are given by

$$u_x = \frac{\partial \Psi\,(x,y)}{\partial y} \text{ and } u_y = -\frac{\partial \Psi\,(x,y)}{\partial x} \tag{10}$$

where Ψ (x, y), the *stream function*, is any scalar function of x and y. A simple physical interpretation can be given to Ψ. Consider a line of constant Ψ in the $x - y$ plane, along such a line

$$\delta\Psi = \frac{\partial \Psi}{\partial x}\delta x + \frac{\partial \Psi}{\partial y}\delta y = 0 \tag{11}$$

i.e.,

$$\begin{aligned} &-u_y \delta x + u_x \delta y = 0 \\ &\text{or } \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{u_y}{u_x} \end{aligned} \tag{12}$$

Since dy/dx is the local slope of the line, Equation (12) shows that the velocity vector is everywhere tangential to lines of constant Ψ. Such lines are referred to as *streamlines*. They give the instantaneous flow pattern of the fluid. By integrating along any path S joining streamlines where $\Psi = \Psi_{\mathrm{A}}$ and $\Psi = \Psi_{\mathrm{B}}$, it can be seen that

$$\Psi_{\mathrm{B}} - \Psi_{\mathrm{A}} = \int_S \frac{\mathrm{d}\Psi}{\mathrm{d}S}\mathrm{d}S = \int_{u_n} u_{\mathrm{n}}\,\mathrm{d}S \tag{13}$$

i.e., volumetric flow crossing S; u_{n} is the velocity component normal to S. Equation (13) shows that the difference in stream function between two streamlines is equal to the volumetric flow rate across any line joining the two streamlines.

The continuity equation and the stream function take on different forms when written in non-cartesian coordinates. Listing of the equations of fluid mechanics in a number of commonly encountered coordinates can be found in most standard textbooks on this subject, e.g., in [3]. A partial list is given in Table 1.

2.2. Cauchy Equations of Motion

Before deriving the equations of motion for fluids, it is necessary to examine the *forces* acting on a fluid element as it moves about in space. Consider again the arbitrary volume V in space (see Fig. 4). One of the external forces acting on the fluid inside V is its weight which is given by the volume integral

$$\int_V \varrho \boldsymbol{g}\,\mathrm{d}V \tag{14}$$

where g is the acceleration due to gravity. Another external force is that exerted by the fluid outside V. This force acts across the closed surface A and can be expressed in terms of the stress tensor $\boldsymbol{T}$ on the surface. In cartesian coordinates this stress tensor has nine components T_{xx}, T_{xy}, T_{xz}, T_{yx}, $T_{yy} \ldots T_{zz}$, or more concisely T_{ij}; $i, j = 1, 2$ or 3. x_1, x_2, x_3 are identified with x, y and z, respectively. T_{ij} is the force per unit area in direction j on a surface with normal in the i direction. The force $\delta\boldsymbol{F}$ acting on an area dA which has a normal $\boldsymbol{n}$ is given by

$$\delta\boldsymbol{F} = \boldsymbol{T}\cdot\boldsymbol{n}\mathrm{d}A \quad \text{or} \quad \delta F_i = \sum_{j=1}^{3} T_{ji} n_j \mathrm{d}A \tag{15}$$

Equation (15) is often referred to as the *Cauchy stress principle*. For the fluids commonly encountered, the stress tensor can be taken to be symmetric, i.e., $T_{ij} = T_{ji}$ for all combinations of i and j. Thus, the external force acting through the closed surface A is given by the surface integral

$$\int_A \boldsymbol{T}\cdot\boldsymbol{n}\mathrm{d}A \tag{16}$$

which is equivalent to

$$\int_V \nabla\cdot\boldsymbol{T}\mathrm{d}V \tag{17}$$

where the divergence theorem is again used to convert the surface integral to a volume integral. The i component of the vector $(\nabla\cdot\boldsymbol{T})$ is

$$(\nabla\cdot\boldsymbol{T})_i = \sum_{j=1}^{3} \frac{\partial T_{ij}}{\partial x_j} \tag{18}$$

To obtain the equations of motion in terms of $\boldsymbol{T}$, $\boldsymbol{u}$, and $(\varrho\, g)$, it is convenient to start with a statement of conservation of momentum which asserts that: the rate of change of momentum inside the volume V and the net rate at which momentum is being convected out of the volume is equal to the sum of the external forces acting on the volume. The rate of accumulation is given by

$$\int_V \frac{\partial\,(\varrho\boldsymbol{u})}{\partial t}\mathrm{d}V \tag{19}$$

Table 1. Continuity and Navier–Stokes equations for incompressible homogeneous fluids in cartesian, cylindrical, and spherical coordinates

Cartesian	Cylindrical	Spherical
	Continuity equation	
$\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} + \frac{\partial u_z}{\partial z} = 0$	$\frac{\partial (r u_r)}{r \partial r} + \frac{1}{r}\left(\frac{\partial u_\theta}{\partial \theta}\right) + \frac{\partial u_z}{\partial z} = 0$	$\frac{1}{r^2}\frac{\partial (r^2 u_r)}{\partial r} + \frac{1}{r \sin\theta}\frac{\partial (u_\theta \sin\theta)}{\partial \theta} + \frac{1}{r \sin\theta}\frac{\partial u_\varphi}{\partial \varphi} = 0$
	Navier–Stokes equation	
$\varrho\left(\frac{\partial u_x}{\partial t} + u_x \frac{\partial u_x}{\partial x} + u_y \frac{\partial u_x}{\partial y} + u_z \frac{\partial u_x}{\partial z}\right) = \frac{\partial p}{\partial x} + \eta\left(\frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2} + \frac{\partial^2 u_x}{\partial z^2}\right)$	$\varrho\left(\frac{\partial u_r}{\partial t} + u_r \frac{\partial u_r}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_r}{\partial \theta} - \frac{u_\theta^2}{r} + u_z \frac{\partial u_r}{\partial z}\right) = -\frac{\partial p}{\partial r} + \eta\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(r u_r)\right) + \frac{1}{r^2}\frac{\partial^2 u_r}{\partial \theta^2} - \frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta} + \frac{\partial^2 u_r}{\partial z^2}\right]$	$\varrho\left(\frac{\partial u_r}{\partial t} + u_r \frac{\partial u_r}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_r}{\partial \theta} + \frac{u_\varphi}{r \sin\theta}\frac{\partial u_r}{\partial \varphi} - \frac{u\theta^2 + u_\varphi^2}{r}\right) = -\frac{\partial p}{\partial r} + \eta\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u_r}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_r}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u_r}{\partial \varphi^2} - \frac{2}{r^2} u_r - \frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta} - \frac{2}{r^2} u_\theta \cot\theta - \frac{2}{r^2 \sin\theta}\frac{\partial u_\varphi}{\partial \varphi}\right]$
$\varrho\left(\frac{\partial u_y}{\partial t} + u_x \frac{\partial u_y}{\partial x} + u_y \frac{\partial u_y}{\partial y} + u_z \frac{\partial u_y}{\partial z}\right) = -\frac{\partial p}{\partial y} + \eta\left(\frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2} + \frac{\partial^2 u_y}{\partial z^2}\right)$	$\varrho\left(\frac{\partial u_\theta}{\partial t} + u_r \frac{\partial u_\theta}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{u_r u_\theta}{r} + u_z \frac{\partial u_\theta}{\partial z}\right) = -\frac{1}{r}\frac{\partial p}{\partial \theta} + \eta\left[\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial}{\partial r}(r u_\theta)\right) + \frac{1}{r^2}\frac{\partial^2 u_\theta}{\partial \theta^2} + \frac{2}{r^2}\frac{\partial u_r}{\partial \theta} + \frac{\partial^2 u_\theta}{\partial z^2}\right]$	$\varrho\left(\frac{\partial u_\theta}{\partial t} + u_r \frac{\partial u_\theta}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_\theta}{\partial \theta} + \frac{u_\varphi}{r \sin\theta}\frac{\partial u_\theta}{\partial \varphi} + \frac{u_r u_\theta}{r} - \frac{u_\varphi^2 \cot\theta}{r}\right) = -\frac{1}{r}\frac{\partial p}{\partial \theta} + \eta\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u_\theta}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_\theta}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u_\theta}{\partial \varphi^2} + \frac{2}{r^2}\frac{\partial u_r}{\partial \theta} - \frac{u_\theta}{r^2 \sin^2\theta} - \frac{2\cos\theta}{r^2 \sin^2\theta}\frac{\partial u_\varphi}{\partial \varphi}\right]$
$\varrho\left(\frac{\partial u_z}{\partial t} + u_x \frac{\partial u_z}{\partial x} + u_y \frac{\partial u_z}{\partial y} + u_z \frac{\partial u_z}{\partial z}\right) = -\frac{\partial p}{\partial z} + \eta\left(\frac{\partial^2 u_z}{\partial x^2} + \frac{\partial^2 u_z}{\partial y^2} + \frac{\partial^2 u_z}{\partial z^2}\right)$	$\varrho\left(\frac{\partial u_z}{\partial t} + u_r \frac{\partial u_z}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_z}{\partial \theta} + u_z \frac{\partial u_z}{\partial z}\right) = -\frac{\partial p}{\partial z} + \eta\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial u_z}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u_z}{\partial \theta^2} + \frac{\partial^2 u_z}{\partial z^2}\right]$	$p\left(\frac{\partial u_\varphi}{\partial t} + u_r \frac{\partial u_\varphi}{\partial r} + \frac{u_\theta}{r}\frac{\partial u_\varphi}{\partial \theta} + \frac{u_\varphi}{r \sin\theta}\frac{\partial u'_\varphi}{\partial \varphi} + \frac{u_\varphi u_r}{r} + \frac{u_\theta u_\varphi}{r}\cot\theta\right) = -\frac{1}{r \sin\theta}\frac{\partial p}{\partial \varphi} + \eta\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u_\varphi}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_\varphi}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u_\varphi}{\partial \varphi^2} - \frac{u_\varphi}{r^2 \sin^2\theta} + \frac{2}{r^2 \sin\theta}\frac{\partial u_r}{\partial \varphi} + \frac{2\cos\theta}{r^2 \sin^2\theta}\frac{\partial u_\theta}{\partial \varphi}\right]$

The rate at which momentum is being convected out is given by the surface integral

$$\int_A \varrho \boldsymbol{u}\,(\boldsymbol{u}\cdot\boldsymbol{n})\,\mathrm{d}A = \int_V \nabla\cdot(\varrho\boldsymbol{u}\boldsymbol{u})\,\mathrm{d}V \tag{20}$$

where $\varrho\,\boldsymbol{u}$ is the momentum per unit volume and $\boldsymbol{u}\cdot\boldsymbol{n}$ is the volumetric flux leaving A. Conservation of momentum in V then requires

$$\int_V \frac{\partial(\varrho\boldsymbol{u})}{\partial t}\mathrm{d}V + \int_V \nabla\cdot(\varrho\boldsymbol{u}\boldsymbol{u})\,\mathrm{d}V = \int_V \varrho\boldsymbol{g}\mathrm{d}V + \int_V \nabla\cdot\boldsymbol{T}\mathrm{d}V \tag{21}$$

Since V is arbitrary, the integrands must be equal

$$\frac{\partial(\varrho\boldsymbol{u})}{\partial t} + \nabla\cdot(\varrho\boldsymbol{u}\boldsymbol{u}) = \varrho\boldsymbol{g} + \nabla\cdot\boldsymbol{T} \tag{22}$$

Upon expanding the terms on the left-hand side and regrouping, Equation (22) becomes

$$\boldsymbol{u}\left[\frac{\partial\varrho}{\partial t} + \nabla\cdot(\varrho\boldsymbol{u})\right] + \varrho\left[\frac{\partial\boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right] = \varrho\boldsymbol{g} + \nabla\cdot\boldsymbol{T} \tag{23}$$

From the continuity equation (Eq. 4) the first group of terms on the left-hand side is zero. The second group of terms can be identified as the product of ϱ and the substantial derivative of velocity. Therefore

$$\varrho\frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} = \varrho\boldsymbol{g} + \nabla\cdot\boldsymbol{T} \tag{24}$$

This is the equation of motion for fluids. It is often referred to as the *Cauchy momentum equation* or equation of motion. The left-hand side is referred to as the inertia term. The first term on the right-hand side is the force per unit volume acting on a fluid element arising from gravitational body force. The second term is the force per unit volume acting on the fluid element as a result of the spatial variation of the stress tensor. The form of the Cauchy equations of motion in different coordinate systems can be found in [3].

2.3. Energy Transport Equation and Bernoulli Equation

In flow problems where a substantial variation in temperature occurs, an additional equation is needed to relate the local temperature T of the fluid to variables such as velocity, velocity gradient, and temperature gradient. Such an equation takes the general form

$$\varrho C_V\frac{\mathrm{D}T}{\mathrm{D}t} = -\nabla\cdot\boldsymbol{q} - T\left(\frac{\partial\varrho}{\partial T}\right)_V\nabla\cdot\boldsymbol{u} + \text{dissipation terms}$$

where $\boldsymbol{q}$ is the heat flux and C_V is the heat capacity at constant volume. This equation simplifies to

$$\varrho C_p\frac{\mathrm{D}T}{\mathrm{D}t} = k\nabla^2 T + \text{dissipation terms} \tag{25}$$

for an incompressible fluid with a heat capacity, C_p, at constant pressure and when the temperature dependence of the thermal conductivity k can be neglected. In arriving at Equation (25), Fourier's expression for heat conduction ($\boldsymbol{q} = -k\nabla T$) was used to express $\boldsymbol{q}$ in terms of temperature gradient. Equation (25), the energy transport equation, is a statement of conservation of energy and is a form of the First Law of Thermodynamics. It equates the rate of energy transfer to a fluid element by thermal conduction and the rate of dissipation/conversion of kinetic energy into thermal energy to the accumulation of thermal energy within the element. The dissipation terms arise as a result of the deformation suffered by the fluid element. The derivation of the energy transport equation and the exact form of the dissipation terms are given in [3]. In the following sections only isothermal flows will be discussed. In such flows the dissipation terms are negligible and the energy transport equation is satisfied. This greatly simplifies the analysis.

In many flow fields of practical interest, a simple statement of conservation of kinetic and potential energy can be derived. This is obtained by forming the dot product between the velocity vector $\boldsymbol{u}$, and the equations of motion (24), and integrating the result between any two points on the same streamline. The resulting equation takes the form

$$\frac{1}{2}u_1^2 + gh_1 + \frac{p_1}{\varrho} = \frac{1}{2}u_1^2 + gh_2 + \frac{p_2}{\varrho} + \text{losses} \tag{26}$$

This is one form of the well-known *Bernoulli equation; h, u,* and p are the vertical elevation, the velocity, and the isotropic pressure of the fluid respectively. Subscript 1 and 2 denote where these variables are to be evaluated

on the streamline. Losses refer to the conversion of mechanical (kinetic and potential) energy into internal (thermal) energy between the two points. In deriving this equation, the fluid is taken to be incompressible and the flow to be steady. Relaxation of these assumptions will result in a more general form of the Bernoulli equation [1]. The Bernoulli equation finds many practical applications, especially in flows where the losses are small or can be estimated with reasonable accuracy. In many applications the spatial average value of the variables at point 1 and point 2 are used instead of local point values, resulting in further simplification (see Section 3.5).

2.4. Constitutive Equations and Classification of Fluids

The equations derived from the conservation laws do not contain any information about the mechanical properties of fluids. An applied force or stress acting on fluids with different properties will result in different flow patterns. Conversely, the same flow pattern will induce different stresses in different fluids. The diverse mechanical properties of fluids are described by equations that relate the local stress tensor to the local flow kinematics. Such equations are known as *rheological constitutive equations.*

One of the most important kinematic variables that appear in rheological constitutive equations is the *rate of strain tensor* $\boldsymbol{e}$. This is defined as

$$\boldsymbol{e}=\frac{1}{2}\left[\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{T}}\right] \tag{27}$$

or in cartesian coordinates

$$e_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right) \tag{28}$$

The rate of strain tensor is a symmetric tensor. Simple physical interpretations can be given to the components of $\boldsymbol{e}$. For example, e_{11} or e_{xx} is the rate at which a fluid element is being stretched in the x direction, e_{12} or e_{xy} is the rate at which the fluid element is being sheared in the $x-y$ plane. The form of the rate of strain and other kinematic tensors in different coordinate systems can be found in [4]. It should be mentioned that some authors, particularly those working on non-Newtonian fluids, define the rate of strain tensor without the factor of 1/2 as shown in Equation (27). This difference in the definition can lead to considerable confusion and should carefully noted.

A convenient way of classifying fluids is according to the form of their constitutive equations. For a large class of fluids, the stress tensor $\boldsymbol{T}$ is a linear function of the rate of strain tensor $\boldsymbol{e}$. Such fluids are referred to as Newtonian fluids. If $\boldsymbol{T}$ is a non-linear function of $\boldsymbol{e}$ the fluids are said to be non-Newtonian. In chemically more complex fluids, the stress is not only a function of $\boldsymbol{e}$, it also depends on the deformation or the entire history of deformation suffered by the fluid. Such fluids may also exhibit solidlike elastic behavior and are known as viscoelastic fluids. (Non-Newtonian and viscoelastic fluids are treated in greater detail in Chap. 4).

3. Newtonian Fluids

3.1. Deviatoric Stress and Viscosity

In a stagnant fluid, the stress at any point is the hydrostatic pressure. This is an isotropic compressive stress, i.e., the stress has the same magnitude in any direction. Such a stress can be represented by

$$\boldsymbol{T}=-p\boldsymbol{I} \tag{29}$$

or in component form

$$T_{ij}=-p\delta_{ij} \tag{30}$$

where $\boldsymbol{I}$ is the identity tensor

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and δ_{ij} is the Kronecker delta

$$\begin{aligned} \delta_{ij} &= 1 \; i=j \\ \delta_{ij} &= 0 \; i\neq j \end{aligned} \tag{31}$$

Unless otherwise stated, i and j take on values 1, 2, or 3 in all the equations in this article. The shear components (when i is not equal to j) of the isotropic stress are identically zero and the three normal components take on the same value

p, the magnitude of the pressure; p is a function of the vertical position in the fluid. It is customary to have a negative sign before p so that for hydrostatic compressive stress, the numerical value associated with p is positive. All these are in agreement with the well known nature of hydrostatic pressure.

For a fluid in motion the stress will, in general, no longer be isotropic. At any point the stress is a combination of the hydrostatic pressure and the stresses arising from the deformation experienced by the fluid. It is convenient to decompose the stress tensor T_{ij} into an isotropic part $-p\,\delta_{ij}$ and a nonisotropic part τ_{ij},

$$T_{ij} = -p\delta_{ij} + \tau_{ij} \tag{32}$$

τ_{ij}, usually referred to as the deviatoric stress tensor, is a consequence of the deformation. For a fluid element that has not experienced any deformation, the deviatoric stress vanishes. In a stagnant fluid the deviatoric stress is therefore zero and the resulting isotropic stress can be identified with the hydrostatic pressure. However, because of fluid motion, the isotropic component of the stress tensor will, in general, be different from the hydrostatic pressure.

The nature of the relationship between deviatoric stress and deformation depends on the rheological constitutive equation of the fluid. The constitutive equation completely describes the deformation arising from a specified deviatoric stress and, conversely, gives the stresses needed to produce a specified deformation. For a large class of incompressible fluids the deviatoric stress $\boldsymbol{\tau}$ is directly proportional to the rate of strain tensor $\boldsymbol{e}$:

$$\boldsymbol{\tau} = 2\eta\boldsymbol{e} \text{ or } \tau_{ij} = 2\eta e_{ij} \tag{33}$$

where η the constant of proportionality, is a property of the fluid and is known as the viscosity. Equation (33) in the *Newtonian constitutive equation*. It is an example of a rheological constitutive equation used to relate kinematic tensors, such as the rate of strain tensor, to the deviatoric stress. Most commonly encountered low molecular mass liquids and gases follow the Newtonian constitutive equation, they are called Newtonian fluids. Viscosity is the most important flow property of fluids. It is a strong function of temperature. The SI unit of viscosity is pascal – second. It should be pointed out that when dealing with compressible fluids, such as gases, the Newtonian constitutive equation must be modified to allow for compressibility. This is done by introducing a bulk viscosity λ into the Newtonian constitutive equation

$$\boldsymbol{\tau} = \lambda\nabla\cdot\boldsymbol{u}\boldsymbol{I} + 2\eta\boldsymbol{e} \tag{34}$$

The bulk viscosity term is zero for incompressible fluids. This article is concerned almost exclusively with incompressible fluids, thus the bulk viscosity term will not appear in all subsequent equations and discussions. For further discussion on compressible flow and bulk viscosity see [5].

Specialized instruments known as viscometers have been developed for measuring viscosity. The basic principles of such instruments can be explained by examining the rate of strain and the deviatoric stress tensors in a fluid contained between two parallel plates, shown in Figure 5. The lower plate is held stationary while the upper is moving parallel to itself, with a constant velocity u_0. Let F denote the force required to maintain the motion of the upper plate, A the area of the plates, and d the spacing between the plates. In this flow, the velocity of the fluid varies linearly from zero at the lower plate to u_0 at the upper plate. With reference to the coordinates shown, it can be seen that the only nonzero components of the rate of strain tensor are

$$e_{xy} = e_{yx} = \frac{1}{2}\frac{\mathrm{d}u_x}{\mathrm{d}y} = \frac{1}{2}\frac{u_0}{d} \tag{35}$$

where u_0/d is the velocity gradient across the plate spacing. This velocity gradient is often referred to as the *shear rate* and denoted by $\dot{\gamma}$. In this simple flow field (the steady shear flow) for Newtonian fluids, the only nonzero components of the deviatoric stress are

$$\tau_{xy} = \tau_{yx} = \frac{F}{A} \tag{36}$$

By definition, the viscosity is given by

$$\eta = \frac{\tau_{xy}}{2e_{xy}} = \frac{F}{A}\Big/\left(\frac{u_0}{d}\right) \tag{37}$$

In a viscometer F and u_0 are measured, and A and d are obtained from the dimensions of the instrument. From these data the viscosity can be calculated. A detailed description of viscometric techniques is given in [6–8].

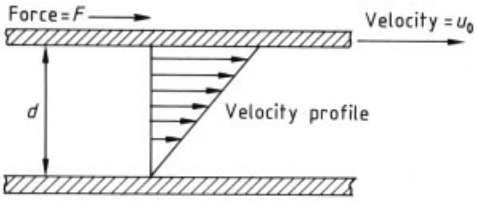

Figure 5. Shearing motion between two parallel plates

Typical values of viscosity (in Pa s at 20 °C or as indicated) for several fluids are given in the following list. These values show the large variation of viscosity that can be encountered in fluid mechanical problems as well as the variation of viscosity with temperature.

Steam (373 K)	1.28×10^{-5}
Carbon dioxide	1.46×10^{-5}
Air (273 K)	1.71×10^{-5}
(293 K)	1.81×10^{-5}
(373 K)	2.18×10^{-5}
(473 K)	2.58×10^{-5}
Pentane	2.34×10^{-4}
Hexane	3.26×10^{-4}
Heptane	4.09×10^{-4}
Octane	5.45×10^{-4}
Benzene	6.47×10^{-4}
Mercury	1.55×10^{-3}
Water (273 K)	1.79×10^{-3}
(283 K)	1.30×10^{-3}
(293 K)	1.00×10^{-3}
(303 K)	7.98×10^{-4}
Olive oil	8.40×10^{-2}
Castor oil	9.86×10^{-2}
Glycerol (273 K)	1.21×10^{1}
(293 K)	1.49
(298 K)	0.942
(303 K)	0.622
Low- density polyethylene at zero shear rate	
(385 K)	2.65×10^{5}
(403 K)	1.05×10^{5}
(443 K)	2.38×10^{4}
(463 K)	1.40×10^{4}

3.2. Navier – Stokes Equations

When the Newtonian constitutive equation is substituted into the Cauchy equations of motion (Eq. 24), the following equation is obtained:

$$\varrho\frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t}=-\nabla p+\eta\nabla^2\boldsymbol{u}+\varrho\boldsymbol{g} \tag{38}$$

or in cartesian component form

$$\varrho\frac{\partial u_i}{\partial t}+\sum_{j=1}^{3}\varrho u_j\frac{\partial u_i}{\partial x_j}=-\frac{\partial p}{\partial x_i}+\eta\sum_{j=1}^{3}\frac{\partial^2 u_i}{\partial x_j^2}+\varrho g_i \tag{39}$$

The first term on the right-hand side is the isotropic pressure term and the second term is the viscous term. This equation together with the continuity equation forms a set of four equations which can be solved to give the three unknown velocity components and the isotropic pressure for any flow of a Newtonian fluid. Equation (39), known as the Navier – Stokes equation, describes the motion of Newtonian fluids. The equivalent forms in some of the more commonly encountered coordinate systems are listed in Table 1. Viscosity is a strong function of temperature in nonisothermal flow, where the variation of viscosity with temperature cannot be ignored, the Navier – Stokes equations must be modified. Thus, it now takes the form

$$\varrho\frac{\partial u_i}{\partial t}+\sum_{j=1}^{3}\varrho u_j\frac{\partial u_i}{\partial x_j}=-\frac{\partial p}{\partial x_i}+\sum_{j=1}^{3}\frac{\partial}{\partial x_j}\left(\eta\left(T\right)\frac{\partial u_i}{\partial x_j}\right)+\varrho g_i \tag{40}$$

This equation must now be solved simultaneously with the continuity equation and the energy transport equation for the unknowns – the three velocity components, the isotropic pressure, and the temperature. It is assumed that the variation of viscosity with temperature is a known experimentally measured quantity. Variable viscosity adds to the difficulty of solving the equations describing the flow. The complexity of these equations is such that they often do not have simple analytical solutions and computers are used to generate approximate numerical answers.

3.3. Applications of Navier – Stokes Equations

3.3.1. Flow in Pipes

Steady flow of Newtonian fluids in pipes is probably the most commonly encountered flow field. It is also one of the small number of flow fields for which the Navier – Stokes equations have a

simple analytical solution. In this flow the only nonvanishing velocity component is that along the axis of the pipe. It is convenient, in this case, to write down the equations of motion in cylindrical coordinates. The axial velocity u_z is a function of the radial coordinate only and is independent of the other coordinates and time. Examination of the Navier – Stokes equations in cylindrical coordinates in Table 1 shows that the only nontrivial component is

$$0 = -\frac{\Delta p}{L} + \eta \left[\frac{1}{r} \frac{\mathrm{d}}{\mathrm{d}r} \left(r \frac{\mathrm{d}u_z}{\mathrm{d}r} \right) \right] \tag{41}$$

The driving force of this flow is the applied pressure gradient $\Delta p/L$; the effect of gravity has been left out in this equation. Equation (41) can be integrated to give

$$u_z = \frac{1}{4\eta} \frac{\Delta p}{L} r^2 + A \ln r + B \tag{42}$$

where A and B are the constants of integration that can be determined by considering the boundary conditions. The velocity along the axis of the tube is finite; this then requires A to be zero. At the pipe wall $r = R$, it is assumed that the fluid adheres to the wall, i.e., $u_z = 0$. This is the well known no-slip boundary condition that is normally imposed at the interface between a solid wall and a flowing fluid. This boundary condition is satisfied for

$$B = -\frac{1}{4\eta} \frac{\Delta p}{L} R^2$$

Thus the resulting velocity profile in the pipe is given by

$$u_z = \frac{R^2}{4\eta} \left(-\frac{\Delta p}{L} \right) \left(1 - \frac{r^2}{R^2} \right) \tag{43}$$

From Equation (43) the volumetric flow rate Q through the pipe can be obtained by integrating

$$Q = \int_0^R 2\pi r u_z \mathrm{d}r = \frac{\pi R^4}{8\eta} \left(-\frac{\Delta p}{L} \right) \tag{44}$$

The average velocity V in the pipe is given by

$$V = \frac{R^2}{8\eta} \left(-\frac{\Delta p}{L} \right) \tag{45}$$

In terms of V, the velocity profile in the pipe takes the simple form

$$u_z = 2V \left(1 - \frac{r^2}{R^2} \right) \tag{46}$$

Thus the velocity profile is a parabola and the maximum velocity, which occurs at the centre of the pipe, is twice the average velocity. For a given pipeline, the driving force needed to attain a specified average velocity is given by

$$\frac{\Delta p}{L} = -\frac{8\eta V}{R^2} \tag{47}$$

Equation (45) is the *Hagen – Poiseuille equation*. It can be put in a dimensionless form

$$f = -\frac{\frac{D\Delta p}{4L}}{\frac{\varrho V^2}{2}} = \frac{16}{\frac{\varrho V D}{\eta}} = \frac{16}{Re} \tag{48}$$

where D is the diameter of the pipe, $Re = \varrho V D/\eta$, and f is as defined by this equation; f is the friction factor and Re is the Reynolds number. Both these quantities are dimensionless. Such dimensionless numbers are widely encountered in fluid mechanical analysis. Equation (48) shows that f is a function of Re only. It will be shown later that shear stress at the pipe wall is given by $D\,\Delta p/4\,L$, while shear rate at the pipe wall is $8\,V/D$. Thus the ratio of the two measurable quantities $\Delta p\,D/4\,L$ and $8\,V/D$ gives the viscosity of the fluid. Measurement in pipe flow can be used for viscosity measurements.

For flow in pipes with $Re < 2100$, the simple relationship between f and Re in Equation (48) has been confirmed by experimental measurements. Such flows are known as laminar flows. As the Reynolds number is increased above 2100, the flow gradually loses it steady nature and becomes more and more chaotic. The expressions derived for the velocity profile, volumetric flow rate, and friction factor are no longer valid. The chaotic fluid motion associated with flows at high Reynolds number ($Re > 4000$) is known as turbulent flow. Turbulent flow is characterized by rapid and apparently random fluctuations in fluid velocity and pressure. The same definition of f and Re given in Equation (48) can be used in turbulent flow. However, the relationship between these two dimensionless numbers in turbulent flow can not be determined analytically and a number of empirical and semiempirical correlations have been developed. In the range of $4000 < Re < 10^5$, experimental data of f versus Re closely follow the empirical *Blasius equation*

$$f = 0.079 Re^{-1/4} \tag{49}$$

Another equation that can be used to describe the relationship between f and Re is the *von Karman – Nikuradse equation*

$$\frac{1}{\sqrt{f}} = 4.0 \log\left(Re\sqrt{f}\right) - 0.4 \tag{50}$$

Equation (50) is valid for $Re \geq 4000$ [5]. The constants 4.0 and 0.4 are determined from experiment. Analysis of turbulent flow requires special techniques and will not be discussed in this article. For further details see [9].

Equations (48) – (50) which relate the friction factor to the Reynolds number were derived for pipes with a smooth inner surface. In commercial pipes surface irregularities are invariably present – depend on the material of construction and the manufacturing process. A quantitative measure of this surface irregularity is the *relative roughness* of the pipe which is defined as the ratio of the average surface roughness e to the pipe diameter D. The relative roughness has to be included as an additional factor in the correlations for friction factor. For most commercial pipes the relative roughness is much smaller than unity and surface irregularities have no noticeable effect on pressure drop as long as the flow remains *laminar*. Thus Equation (48) remains valid for rough pipes. In *turbulent flow,* the situation is quite different. The friction factor is altered significantly even for small *e/D*. Empirical equations have been developed to relate the friction factor to the relative roughness and the Reynolds number. One example is the *Colebrook equation*:

$$\frac{1}{\sqrt{f}} = -4.0 \log\left(\frac{e}{D} + \frac{4.67}{Re\sqrt{f}}\right) + 2.28 \tag{51}$$

A disadvantage is that for a given Re and e/D, Equation (51) must be solved iteratively for f. Another difficulty in using the Colebrook equation is the large uncertainty associated with the relative roughness.

The friction factor is important in the sizing of pipelines. It is customary to plot f against Re in the form of a log – log plot. An example of such a plot is shown in Figure 6 (*Moody chart*). The relative roughness of the pipe is included in this plot as an independent parameter.

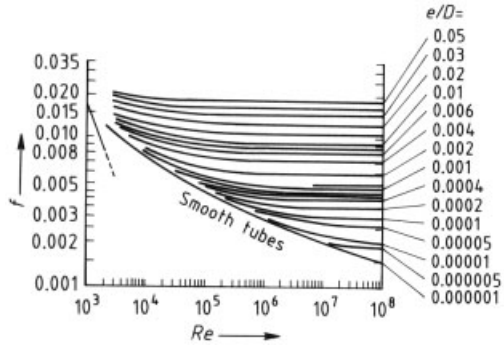

Figure 6. Friction factor for flow in circular pipes at different Reynolds number and for different relative pipe roughness (*e/D*) [1]

3.3.2. Concentric Cylinder Flow

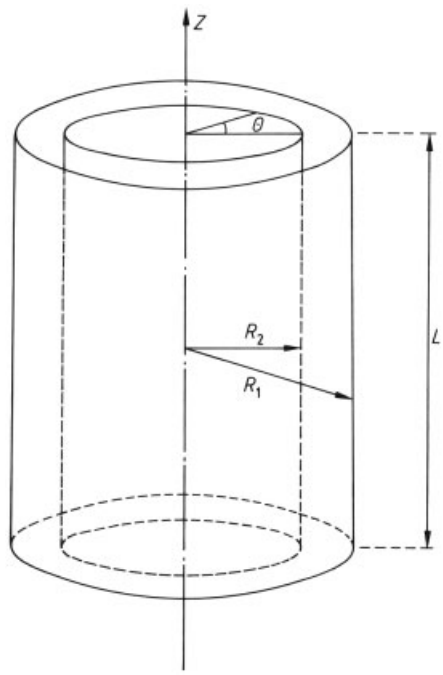

Figure 7. Flow between two concentric cylinders

Another exact solution of the Navier – Stokes equation is the one that describes the velocity profile and stresses generated in a fluid contained between the annular gap of two rotating concentric cylinders (see Fig. 7). This flow is of great practical interest because it is the basis of a large class of commercial viscometers. It is also of

considerable theoretical importance; it is one of the few flows for which the transition (via a series of progressively more complicated flows) of the simple laminar solution to turbulent flow has been carefully observed and analyzed theoretically.

To simplify the analysis, the length of the cylinders are assumed to be long compared to their radii so that the end effects can be ignored. There is no applied pressure gradient in this flow. Fluid motion is brought about by the rotation of the cylinders. Let ω_1 and ω_2 denote the angular velocity of the outer and inner cylinders respectively. The corresponding linear velocity of the cylinders are $\omega_1 R_1$ and $\omega_2 R_2$, where R_1 and R_2 are the respective radii. The only nonvanishing velocity component, in cylindrical coordinates, is the azimuthal component u_θ. The Navier–Stokes equations reduces to

$$-\varrho\left(\frac{u_\theta^2}{r}\right)=-\frac{\mathrm{d}p}{\mathrm{d}r} \tag{52}$$

$$0=\eta\frac{\mathrm{d}}{\mathrm{d}r}\left[\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}\left(ru_\theta\right)\right] \tag{53}$$

Equation (53) can be integrated to give u_θ in terms of the radial coordinate and the boundary conditions on the inner and outer cylinder. Using the newly obtained velocity, Equation (52) can in turn be integrated to give the pressure. The resulting expressions for u_θ and p are

$$u_\theta=\frac{1}{R_2^2-R_1^2}\left[r\left(R_2^2\omega_2-R_1^2\omega_1\right)+\frac{R_1^2R_2^2}{r}\left(\omega_1-\omega_2\right)\right]$$

$$p=p_2+\frac{\varrho}{R_2^2-R_1^2}\left[\frac{\left(R_2^2\omega_2-R_1^2\omega_1\right)^2\left(r^2-R_2^2\right)}{2}+\right.$$

$$2R_1^2R_2^2\left(\omega_1-\omega_2\right)\left(R_2^2\omega_2-R_1^2\omega_1\right)\ln\frac{r}{R_2}+$$

$$\left.R_1^4R_2^4\left(\omega_1-\omega_2\right)^2\left(\frac{1}{r^2}-\frac{1}{R_2^2}\right)\right]$$

where p_2 is the pressure at the inner cylinder wall. It can easily be verified that the no-slip boundary condition is satisfied at the outer and inner cylindrical walls.

In most commercial concentric cylinder viscometers, the outer cylinder is held fixed while the inner one is rotated at a steady angular velocity by an applied torque Γ. For this special case, the velocity in the annular gap is given by

$$u_\theta=\frac{1}{R_2^2-R_1^2}\left(rR_2^2\omega_2-\frac{R_1^2R_2^2\omega_2}{r}\right) \tag{54}$$

The corresponding shear rate and shear stress on the inner wall are

$$\dot{\gamma}_{r\theta_2}=\frac{2\omega_2}{1-\left(R_2/R_1\right)^2} \tag{55}$$

$$\tau_{r\theta_2}=\frac{\Gamma}{2\pi R_2^2L} \tag{56}$$

where L is the length of the cylinder. In viscometry, these two expressions are combined to yield an explicit expression for the viscosity of the fluid:

$$\eta=\frac{\tau_{r\theta_2}}{\dot{\gamma}_{r\theta_2}}=\frac{\Gamma}{L}\frac{1}{4\pi\omega_2}\left(\frac{1}{R_2^2}-\frac{1}{R_1^2}\right) \tag{57}$$

With Equation (57) the viscosity of the fluid can be calculated from the applied torque per unit height of the cylinders Γ/L and the measured angular velocity ω_2. For Newtonian fluids, in principle, only a single reading of the torque together with the corresponding angular velocity is needed to determine the viscosity. For non-Newtonian fluids, the situation is more complex and a series of data points are needed to determine the viscosity as a function of shear rate [6].

According to the solution of the Navier–Stokes equations, the streamlines should form a family of circles concentric with the cylinders. At low Reynolds number this is indeed the observed flow pattern. However, as the Reynolds number is increased (either by increasing the angular velocity or decreasing the viscosity) this simple flow loses its stability and gives way to a new steady flow in which the streamlines are spirals whose axes are concentric with the cylinders. A sketch of the new flow pattern, at sufficiently large *Re*, is shown in Figure 8. The vortices formed by the fluid are known as Taylor vortices. The transition between the two steady flows can be observed by plotting the torque Γ against the shear rate $\dot{\gamma}$ (see Fig. 9). The abrupt change in slope of the $\Gamma-\dot{\gamma}$ plot can be used to locate the transition point accurately. In using Equation (57) to obtain the viscosity from concentric viscometer measurements, it is important to ensure that all the data points are taken prior to the onset of transition flow. At even higher Reynolds number, the Taylor vortices in turn become unstable and are replaced by a series of progressively more complicated flows, which ultimately lead to turbulent flow. The detailed

step by step mapping of the laminar-to-turbulent transition has been investigated experimentally and analyzed theoretically with excellent agreement between the two.

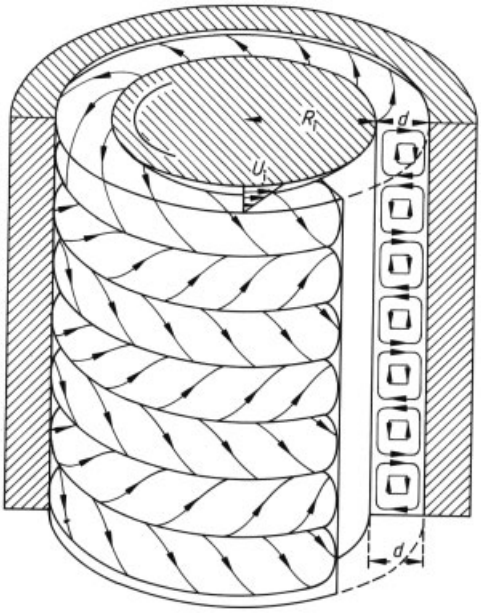

Figure 8. Taylor vortices [5] (reproduced with permission of McGraw-Hill)

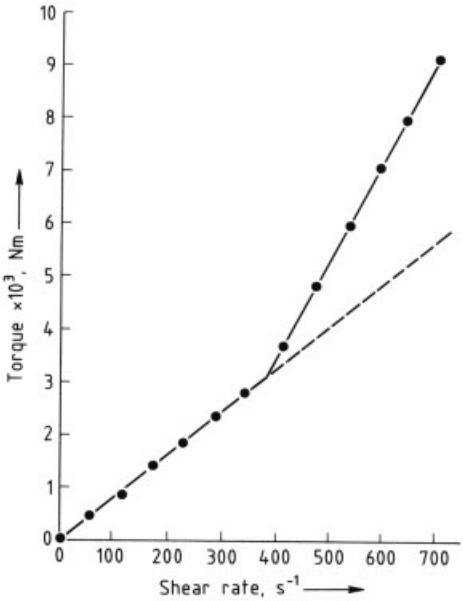

Figure 9. Change in slope in the torque – shear rate plot for flow between concentric cylinders indicating the onset of Taylor vortices [1]

In making viscosity measurements, it is common practice to use more than one viscometer so that the experimental data covers as wide a range of shear rate as possible. This step is particularly important if the possibility exists that the fluid may not be Newtonian (i.e., it has a viscosity that is a function of shear rate). Figure 10 shows a log – log plot of shear stress against shear rate obtained using four different viscometers. The data at low shear rates were obtained using a concentric cylinder viscometer. Those at higher shear rates were obtained from the volumetric flow rate and pressure gradient data from three different capillary tubes. (Experimental techniques associated with capillary viscometer are described in Section 4.2). The fact that all the data points, from four different viscometers, fall on the same straight line with slope equal to unity means that the fluid is Newtonian over the shear rates covered and the viscosity can be read off from the intercept of the straight line with the vertical line at $\dot{\gamma} = 1.0$; η is, in this case, ca. 0.91 Pa s.

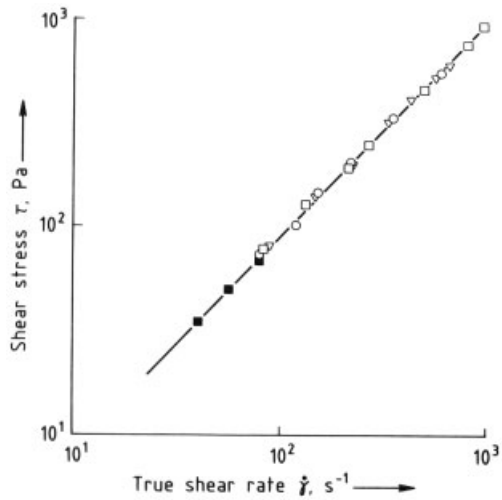

Figure 10. Log – log plot of shear stress versus shear rate for a Newtonian fluid (glycerol at 21.6 °C) obtained using four different viscometers
□ $D = 5.573$ mm, $L = 300$ mm, $L/D = 53.8$; ▽ $D = 5.569$ mm, $L = 477$ mm, $L/D = 85.7$; ○ $D = 5.595$ mm, $L = 700$ mm; $L/D = 125.1$; ■ concentric cylinder viscometer

3.3.3. Creeping Flow Past a Sphere

Steady flow of a Newtonian fluid past a sphere has been the subject of numerous investigations.

This flow is of great practical importance in its own right, but also because it illustrates the steps involved in the evaluation of the drag force exerted by a flowing fluid on an object. It is also of considerable theoretical interest because of the part it played in the development of some of the modern mathematical techniques, such as singular perturbation and matched asymptotic analyses, used in solving flow problems.

The mathematical analysis is greatly simplified if the inertia terms in the Navier–Stokes equations are assumed to be small compared to the viscous terms and thus are ignored. This assumption is valid if the fluid velocity or density is small, its viscosity large, or if the sphere diameter D is small. These conditions are equivalent to requiring the Reynolds number $Re = \varrho U_0 D/\eta$, to be small; U_0 is the undisturbed velocity of the fluid at a large distance from the sphere. Flows with low Reynolds number are often referred to as *inertialess* or *creeping flows.*

In spherical coordinates, the equations for creeping flow past a sphere are

$$0 = -\frac{\partial p}{\partial r} + \eta \left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u_r}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_r}{\partial \theta}\right) - \frac{2u_r}{r^2} - \frac{2}{r^2}\frac{\partial u_\theta}{\partial \theta} - \frac{2}{r^2} u_\theta \cot\theta \right]$$

$$0 = -\frac{1}{r}\frac{\partial p}{\partial \theta} + \eta \left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u_\theta}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u_\theta}{\partial \theta}\right) + \frac{2}{r^2}\frac{\partial u_r}{\partial \theta} - \frac{u_\theta}{r^2 \sin^2\theta} \right] \quad (58)$$

It is convenient to locate the center of the sphere at the origin of the spherical coordinates and to have the undisturbed upstream velocity of the fluid parallel to the $\theta = 0$ line (see Fig. 11). In this coordinate system, u_r and u_θ are the only two nonvanishing velocity components. The Newtonian fluid is assumed to adhere to the surface of the sphere (the no-slip boundary condition), i.e.,

$$u_r = u_\theta = 0 \text{ at } r = D/2 \quad (59)$$

Far away from the sphere, $\boldsymbol{u}$ approaches the uniform velocity U_0, i.e.,

$$u_r = U_0 \cos\theta,\ u_\theta = -U_0 \sin\theta \quad \text{as} \quad r \to \infty \quad (60)$$

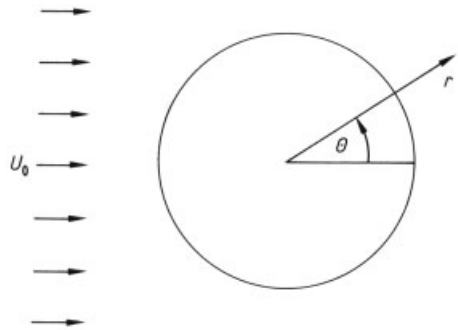

Figure 11. Definitive sketch for flow past a sphere

In addition, the velocity components have to satisfy the continuity equation

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 u_r\right) + \frac{1}{r \sin\theta}\frac{\partial}{\partial \theta}\left(u_\theta \sin\theta\right) \quad (61)$$

Equations (58) and (61) can be solved to give the following expressions for the velocity components and pressure

$$u_r = U_0 \left[1 - \frac{3}{4}\frac{D}{r} + \frac{1}{2}\left(\frac{D}{2r}\right)^3 \right] \cos\theta$$

$$u_\theta = -U_0 \left[1 - \frac{3}{8}\frac{D}{r} - \frac{1}{4}\left(\frac{D}{2r}\right)^3 \right] \sin\theta$$

$$p = p_0 - \left(\frac{3\eta U_0}{D}\right)\left(\frac{D}{2r}\right)^2 \cos\theta \quad (62)$$

where p_0 is the uniform pressure far from the sphere. For mathematical details leading to these expressions see [1].

To calculate the *drag force* on the sphere, the isotropic pressure and deviatoric stress on the surface of the sphere must be evaluated. These are given, respectively, by

$$p = p_0 - \frac{3\eta U_0}{D}\cos\theta \quad (63)$$

$$\tau_{r\theta} = \eta \left[r \frac{\partial}{\partial r}\left(\frac{u_\theta}{r}\right) + \frac{1}{r}\frac{\partial u_r}{\partial \theta} \right]_{r=D/2} = -\frac{3\eta U_0}{D}\sin\theta$$

The other components of the deviatoric stress are zero on the surface. The drag force F_d is obtained by integrating the resulting Cauchy stress over the entire surface of the sphere. From the symmetry of the problem, it is clear that the drag force is in the same direction as the undisturbed velocity:

$$F_\mathrm{d} = \int_{\varphi=0}^{\pi}\int_{\theta=0}^{2\pi} \left[\left(-p_0 + \frac{3\eta U_0}{D}\cos\theta\right)\cos\theta + \frac{3\eta U_0}{D}\sin^2\theta \right] R^2 \sin\theta \,\mathrm{d}\theta\,\mathrm{d}\varphi = 3\pi\eta D U_0$$

This result is often referred to as *Stokes law*. One of the practical applications of Stokes law is the *falling ball viscometer*. Here the terminal velocity of a small sphere of known diameter that falls through a fluid is measured. When the sphere has attained its terminal velocity, the drag force on the sphere is exactly balanced by it weight, i.e.:

$$F_{\mathrm{d}} = \frac{4}{3}\pi\,(D/2)^3\,\Delta\varrho g = 3\pi\eta D U_0 \tag{64}$$

where $\Delta\,\varrho$ is the difference between the density of the solid sphere and the fluid. This then yields a simple expression which allows the fluid viscosity to be calculated from measurable quantities such as the densities of the fluid and the sphere, and the terminal velocity:

$$\eta = \frac{1}{18}\,\frac{D^2\Delta\varrho g}{U_0} \tag{65}$$

Usually the drag force F_{d} is expressed in terms of a dimensionless *drag coefficient* C_{d} obtained by dividing F_{d} by $\frac{1}{2}\varrho U_0^2$

and by the cross-sectional area of the sphere normal to the velocity U_0; thus the drag coefficient is

$$C_{\mathrm{d}} = \frac{3\pi\eta D U_0}{\frac{1}{2}\varrho U_0^2 \pi D^2/4} = \frac{24\eta}{\varrho U_0 D} \tag{66}$$

In terms of C_{d} and the Reynolds number $Re = \varrho\, U_0 D/\eta$, Stokes law takes the form

$$C_{\mathrm{d}} = 24/Re \tag{67}$$

This result was obtained for *Re* much smaller than unity: in this case the error incurred is acceptable for most practical purposes. For larger *Re*, the inertia terms can no longer be ignored and additional terms must be included in the expression for C_{d}. The result now takes the form of an infinite series in *Re*, the leading terms are

$$C_{\mathrm{d}} = \frac{24}{Re}\left[1 + \frac{3}{16}Re + \frac{9}{160}\,(\ln Re)\,Re^2 + \ldots\right] \tag{68}$$

The validity of this series extends to *Re* of the order of unity. The mathematical analysis leading to this series is described in [2].

Expressions (67) and (68) for C_{d} are based on the assumption that the flow remains steady and laminar and *Re* is small. As *Re* is increased, the flow pattern around the sphere becomes more and more complicated. (See the numerically generated flow pattern around a cylinder shown in Figure 47) The laminar flow eventually gives way to turbulent flow, and C_{d} is expected to change dramatically with *Re*. Experimentally measured drag coefficients as a function of *Re* are shown in Figure 12. The validity of Equations (67) and (68) is restricted to *Re* less than ca. unity with the range of validity of Equation (68) slightly larger than that of Equation (67). For *Re* in the range 1 to 10^3, C_{d} can be approximated by

$$C_{\mathrm{d}} \approx 18 Re^{-0.6} \tag{69}$$

And for $10^3 \leq Re \leq 2\times10^5$, C_{d} is approximately constant

$$C_{\mathrm{d}} \approx 0.44 \tag{70}$$

For a detailed description of the flow phenomena at increasing *Re*, including the sudden decrease in C_{d} at $Re \approx 2\times10^5$, see [1].

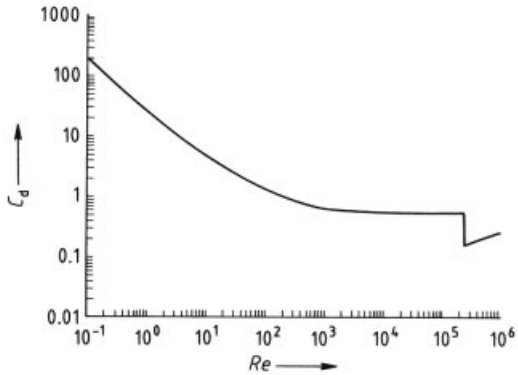

Figure 12. Log – log plot of drag coefficient on a sphere as a function of the Reynolds number

3.4. Some Other Important Flows

3.4.1. Flow Through Granular Beds

The flow of fluids through granular beds occurs widely in industry and in nature, some examples are flow of water down a filter bed, seepage of oil through porous underground soil structure, and flow of chemicals in a packed bed reactor. The key variable in the study of flow through a granular bed is the volumetric flow rate per unit cross-sectional area of the bed, *Q/A*. This flow rate depends on the density ϱ and viscosity η of the fluid, the average size (and shape) of the

particles D_p that make up the bed, the applied pressure gradient $\Delta p/L$ across the bed, and the porosity of the bed ε. The relationship between these variables can be expressed in the general form

$$\Delta p/L = f\,(Q/A, \varrho, \eta, \varepsilon, D_p)$$

The exact form of the function in this equation is expected to be complex and cannot be determined from fundamental considerations only. It must therefore be determined by experimental measurements. A clearer understanding of the influence of each of these variables can be achieved by presenting the experimental results in a dimensionless form. For the present problem, the relevant dimensionless variables are an appropriately defined friction factor f and an appropriately defined Reynolds number:

$$f = \frac{D_p \varepsilon^3}{\varrho U_\infty^2 (1-\varepsilon)} \frac{\Delta p}{L} \tag{71}$$

$$Re = \frac{\varrho D_p U_\infty}{(1-\varepsilon)\,\eta} \tag{72}$$

U_∞ is the superficial velocity and is defined by $U_\infty = Q/A$. From the accumulated data in the literature, it can be shown that f and Re are reasonably correlated by the following semi- empirical expressions:

$$\begin{aligned} f &= 150/Re \text{ for } Re<10 \\ f &= 1.75 + \frac{150}{Re} \text{ for } 10<Re<1000 \\ f &= 1.75 \text{ for } 1000<Re \end{aligned} \tag{73}$$

This set of expressions for f is known as the *Ergun relationship*. At low Re, corresponding to inertialess laminar flow through the bed, the friction factor is proportional to the reciprocal of Re, as is the case of laminar flow in a circular pipe. This appears to be a general property of laminar flow of Newtonian fluids. The Ergun relationship was based on experimental data – pressure drop in granular beds of spherical particles. For nonspherical particles, D_p must be replaced by the hydraulic diameter, which is defined by $D_{hy} = 4\times$ volume of particle/surface area of particle wetted by fluid. For a granular bed of particles which deviate greatly from sphericity, a shape factor may have to be introduced.

3.4.2. Fluidization

Fluidization is a common operation in the process industry in which a fluid stream, usually gas, is used to ensure thorough mixing of a bed of solid particles. The gas stream is passed vertically upward through the solid bed. The pressure drop across the solid bed (or equivalently the upward drag force on the particles) increases with the gas velocity. At a sufficiently high gas velocity the drag force becomes equal to the weight of the individual particles and the particles enter a weightless state. Further increase in gas velocity is not accompanied by an increase in pressure drop (see Fig. 13), instead the bed begins to expand. The bed of solid particles now acquires many of the properties normally associated with a fluid, such as the ability to flow freely; it is said to be fluidized. The weightless particles move rapidly and randomly throughout the expanded bed, thereby ensuring thorough mixing. This is exploited, for example, in a fluidized-bed reactor where the solid particles may be the catalyst on which an exothermic chemical reaction takes place. The random motion in the fluidized state results in good particle – particle and particle – wall contact and hence in an efficient removal of the heat of reaction.

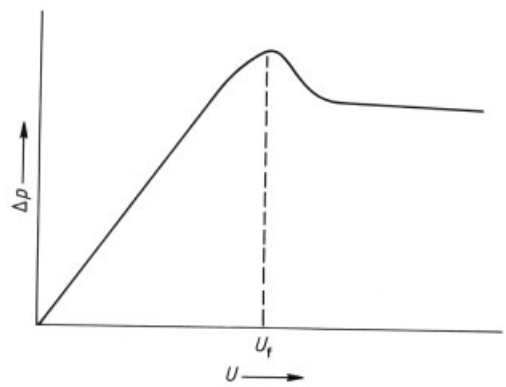

Figure 13. Variation of pressure drop Δp with flow rate u in a granular bed showing the leveling of pressure drop after the onset of fluidization (u_f)

One of the key variables in the operation of a fluidized bed is the minimum superficial gas velocity u_f needed to maintain the particles in the fluidized state. This minimum velocity can be estimated using the Ergun equation (73) for the pressure drop Δp in a packed bed. The force

exerted by the gas stream on a bed of solid particles of cross-sectional area A is given by $A\,\Delta p$. At the point of fluidization, this is equal to the weight of the bed, i.e.

$$A\Delta p = (\varrho_{\mathrm{p}} - \varrho)\, g A H\, (1-\varepsilon) \tag{74}$$

where H is the height of the bed and ε is its porosity; $(\varrho_{\mathrm{p}} - \varrho)$ is the difference between the density of the solid particles and the gas; and Δp is given by the Ergun equation. Since, in most fluidized beds, the particle size is small and the Reynolds number is often < 10, the friction factor in the Ergun equation can be approximated by $150/Re$. Substituting this result into Equation (74) gives the following expression for u_{f}

$$u_{\mathrm{f}} = \frac{(\varrho_{\mathrm{p}} - \varrho)\, g D_{\mathrm{p}}^2 \varepsilon^3}{150\eta\,(1-\varepsilon)} \tag{75}$$

The difficulty with using Equation (75) lies in the uncertainty of estimating the value of ε, the porosity of the bed at the point of fluidization. In the absence of experimental data, $\varepsilon^3/(1-\varepsilon)$ can be approximated by 0.091. For a more accurate estimation of the minimum fluidization velocity, it is also necessary to take into account the nonspherical shape of the particles. A large number of semi-empirical correlations exist in the literature on fluidization which take into account these and additional factors [10].

3.4.3. Gas – Liquid Flow

Many industrial operations require the handling of fluid streams composed of two distinct phases. The two phases can be a gas and a liquid, two immiscible liquids, or a suspension of fine solid particles in a gas or liquid. The flow behavior of a two-phase stream is considerably more complicated than that of a single-phase fluid. Investigation of two-phase flow is an active area of research with an extensive literature. This section provides only a brief qualitative description of gas – liquid flow in tubes, probably the most frequently encountered two-phase flow.

Flow in Vertical Tubes. One of the most distinctive features of gas – liquid flow is the large number of flow patterns that can be observed. The flow patterns in a pipe for example, depend on the velocity of the two phases, their physical properties, pipe diameter, and the geometrical arrangement of the pipe. Some of the flow patterns that can be observed in a gas – liquid stream flowing up a vertical tube are shown in Figure 14 together with the terms used to describe the various flow regimes.

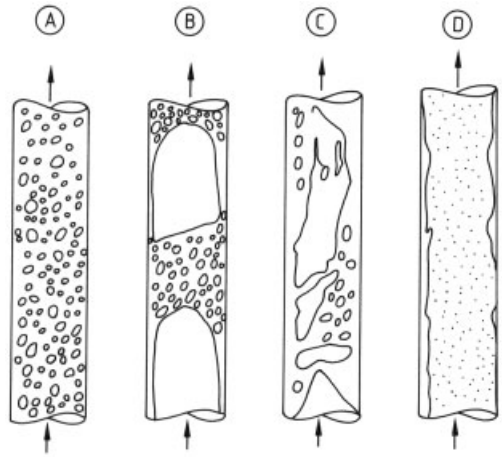

Figure 14. Flow patterns for gas – liquid flow in vertical pipes [12]
A) Bubble flow; B) Slug flow; C) Churn flow; D) Annular flow

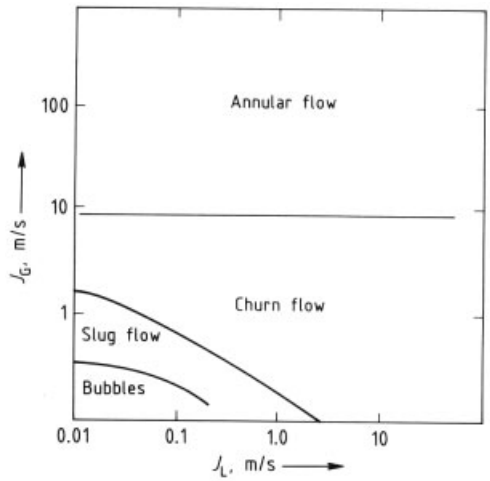

Figure 15. Map of different flow regimes for gas – liquid flow in vertical pipes [11]

In the *bubble flow regime* (Fig. 14A), observed at very low gas flow rates, the gas phase exists in the form of many small bubbles distributed in the continuous liquid phase. As the gas rate is increased, the bubbles grow and coalesce to form large slugs. The transition from bubble flow to *slug flow* (Fig. 14B) is gradual.

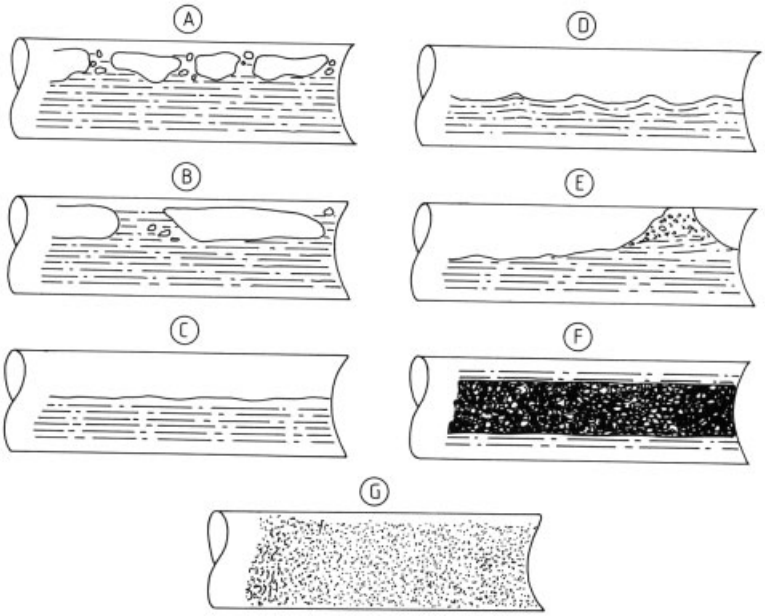

Figure 16. Flow patterns for gas – liquid flow in horizontal pipes [13]
A) Bubble flow; B) Plug flow; C) Stratified flow; D) Wavy flow; E) Slug flow; F) Annular flow; G) Spray

The slug diameter is of the order of the tube diameter. Most of the gas phase now resides within the slugs with only a small fraction dispersed as small bubbles in the liquid bridges separating the slugs. A thin liquid film remains between the slug and the tube wall. When the gas rate is increased further this film suddenly becomes unstable and *churn flow* (Fig. 14C) takes over. Here the flow is highly chaotic and the slugs eventually lose their identity. At yet higher gas rates the flow becomes more steady. The gas phase now becomes the continuous phase occupying the central core of the tube with an *annular*, usually wavy, *liquid film* (Fig. 14D) lining the tube wall. There is also a fine mist of liquid droplets entrained in the continuous gas phase.

One of the objectives of gas – liquid flow experimentation is to determine the conditions for the transition from one flow regime to another. Experimental observations are often summarized in the form of a *flow regime map*. Figure 15 is a typical map for gas – liquid flow in a vertical tube. The horizontal and the vertical axes are the superficial liquid velocity J_{L} and gas phase velocity J_{G}, respectively. These are defined by

$$J_{\mathrm{L}} = L/\left(\pi D^2/4\right)$$

$$J_{\mathrm{G}} = G/\left(\pi D^2/4\right)$$

where L is the volumetric flow rate of the liquid phase and G is that of the gas phase; D is the tube diameter. The boundaries between the various flow regimes are far from well defined. They also depend on the physical properties of the gas and liquid, and on the diameter of the tube. Like most flow regime maps reported in the literature, Figure 15 is based on data collected for air – water systems at normal temperature and pressure in a tube 25 mm in diameter. Using this map to determine the flow pattern of two-phase systems with vastly different properties or in much larger or smaller tubes is not recommended.

Flow in Horizontal Tubes. As expected, the observed flow patterns of a gas – liquid stream in a horizontal tube differ significantly from that in a vertical tube. In the horizontal tube the two phases are more likely to segregate because of their density difference. Figure 16 is the horizontal analogue of Figure 14. Similar terms are used to describe the observed flow patterns at increasing gas flow rates. At low gas flow rates, the gas phase exists as *bubbles* (Fig. 16A) which

move along the upper part of the tube at approximately the same speed as the liquid phase. Increasing gas flow will gradually lead to *plug flow* (Fig. 16B), where alternate plugs of liquid and gas move along the upper part of the tube. In *stratified* (Fig. 16C) and in *wavy flow* (Fig. 16D) the two phases occupy different parts of the tube. In stratified flow the gas – liquid interface is relatively smooth while in wavy flow the higher gas velocity causes the interface to become wavy. When the interface becomes grossly distorted, the gas phase becomes separated into *slugs* (Fig. 16E) which move through the tube at a higher speed than the liquid phase. In *annular flow* (Fig. 16F), the liquid flows as a thin film lining the wall of the tube and the gas flows at a higher speed in the central core. The liquid film is thicker at the bottom half of the tube and is generally not smooth. A part of the liquid phase exists as droplets entrained in the gas stream. At very high gas rates, the liquid film completely disintegrates and the liquid phase is carried as a *fine spray* (Fig. 16G) by the high speed gas stream.

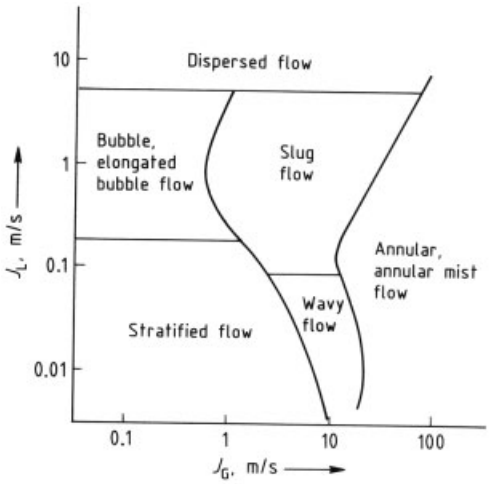

Figure 17. Map of different flow regimes for gas – liquid flow in horizontal pipes [14]

The regime map for gas – liquid flow in a horizontal tube is shown in Figure 17. The vertical axis is the superficial liquid velocity and the horizontal axis is the gas velocity. Figure 17 was constructed from data for air – water mixtures in tubes with diameter < 100 mm.

Apart from mapping the flow regimes in gas – liquid flows, considerable effort has been directed at solving the equations of motion for two-phase flow and the development of correlations that can be used in design calculations. Details of these can be found in [11–15].

3.5. Mechanical Energy Balance for Macroscopic Systems

The solution of pipe network flow problems for a macroscopic system such as that illustrated in Figure 18 is often of extreme importance. Once again the three basic conservation principles are required but this time expressed for a system of finite size. At steady state the statements of conservation of mass and energy for the control volume shown in Figure 18 are

$$\varrho_1 \langle u_1 \rangle A_1 = \varrho_2 \langle u_2 \rangle A_2 = W \tag{76}$$

and

$$\frac{1}{2}\Delta\frac{\langle \boldsymbol{u}^3 \rangle}{\langle \boldsymbol{u} \rangle} + \boldsymbol{g}\Delta z + \int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} + Ev + W_\mathrm{s} = 0 \tag{77}$$

where W is the mass flow rate, ϱ the fluid density, A the cross-section of the flow, $\boldsymbol{u}$ the flow velocity, z the height above some arbitrary datum, p the pressure, g the acceleration due to gravity, W_s the shaft work, i.e., pump work, Ev the friction loss, and $\langle\rangle$ is used to denote average value. Each of the terms in Equation (77) has the unit of energy per unit mass; Δ refers to the outlet minus the inlet conditions. Since the velocity $\boldsymbol{u}$ across any flow cross-section is not in general uniform, an average velocity V defined by

$$V = \langle \boldsymbol{u} \rangle = \frac{1}{A}\int_A \boldsymbol{u}\,\mathrm{d}A \tag{78}$$

and for tube flow by

$$V = \langle \boldsymbol{u} \rangle = \frac{1}{\pi R^2}\int_0^R 2\pi r \boldsymbol{u}\,\mathrm{d}r \tag{79}$$

must be used in the mass and energy balances. For turbulent flow prior knowledge of the velocity distribution at a particular cross-section is of little consequence because the velocity is essentially uniform across the cross-section, i.e.,

$$\langle \boldsymbol{u} \rangle = V \text{ and } \frac{\langle \boldsymbol{u}^3 \rangle}{\langle \boldsymbol{u} \rangle} = \frac{V^3}{V} = V^2 \tag{80}$$

and for an incompressible fluid

$$\int_{p_1}^{p_2} \frac{\mathrm{d}p}{\varrho} = \frac{1}{\varrho}(p_2 - p_1) \tag{81}$$

Thus, for *turbulent flow* of an incompressible fluid Equation (77) reduces to:

$$\Delta\left(\frac{V^2}{2}\right) + \boldsymbol{g}\Delta z + \frac{p_2 - p_1}{\varrho} + Ev + W_\mathrm{s} = 0 \tag{82}$$

Equation (82) is sometimes referred to as the *mechanical energy balance.* If in addition no shaft work W_s crosses the system boundary, then

$$\Delta\left(\frac{V^2}{2}\right) + \boldsymbol{g}\Delta z + \frac{p_2 - p_1}{\varrho} + Ev = 0 \tag{83}$$

this is identical to Equation (26), the Bernoulli equation. Both the mechanical energy balance and the Bernoulli equation are not generally applicable to all flow situations.

For isothermal flow of incompressible fluids (*laminar* and *turbulent flow*, Eq. 76)

$$\varrho_1 \langle u_1 \rangle A_1 = \varrho_2 \langle u_2 \rangle A_2 = W$$

and

$$\frac{1}{2}\Delta\frac{\langle \boldsymbol{u}^3 \rangle}{\langle \boldsymbol{u} \rangle} + \boldsymbol{g}\Delta z + \frac{p_2 - p_1}{\varrho} + Ev + W_\mathrm{s} = 0 \tag{84}$$

The problem is to evaluate the terms in Equation (84) so that the performance of an existing flow system can be evaluated or a new flow system can be designed [16].

The major unknown in Equation (84) is the friction loss term Ev. The friction loss term is conveniently split into a contribution due to fully developed flows in long, straight sections of the pipe and an excess contribution due to losses in valves, bends, fittings, expansions, contractions, etc:

$$Ev = Ev_\mathrm{fd} + Ev_\mathrm{ex}$$

Ev_ex is determined from experimental observation whereas Ev_fd can in some instances be determined from fundamental principles. The evaluation of Ev_fd for tube flows is considered in Section 3.5.1.

3.5.1. Fully Developed Tube Flows

Consider the control volume shown in Figure 19 for fully developed tube flow. Application of Equation (84) to this control volume yields

$$Ev_\mathrm{fd} = \frac{p_1 - p_2}{\varrho} \tag{85}$$

That is, the energy loss per unit mass is simply the pressure drop divided by the fluid density. Application of a *force balance* to the same control volume yields

$$\pi R^2 (p_1 - p_2) = \tau_\mathrm{w} 2\pi R L \tag{86}$$

or

$$\tau_\mathrm{w} = \frac{R(p_1 - p_2)}{2L} = \frac{D(p_1 - p_2)}{4L} \tag{87}$$

and at any radius r

$$\tau_{rz} = \frac{r(p_1 - p_2)}{2L} \tag{88}$$

Elimination of $(p_1 - p_2)$ from Equations (87) and (88) leads to

$$\tau_{rz} = \frac{r}{R}\tau_\mathrm{w} \tag{89}$$

Equation (89) indicates a linear distribution of shear stress in a tube from zero at the centerline to τ_w at the wall. This relationship is valid for Newtonian and non-Newtonian fluids and for laminar and turbulent flows.

Frictional losses for fully developed flows are expressed in terms of the friction factor (Eq. 48). Combination of Equations (85), (87), and (48) leads to the *Fanning equation*:

$$Ev_\mathrm{fd} = \frac{2fV^2L}{D} \tag{90}$$

which is the basic equation for defining the frictional losses for fully developed flow in pipes. For Newtonian fluids f is given by Equation (48) for laminar flow and by Equations (49), (50), and (51) for turbulent pipe flow.

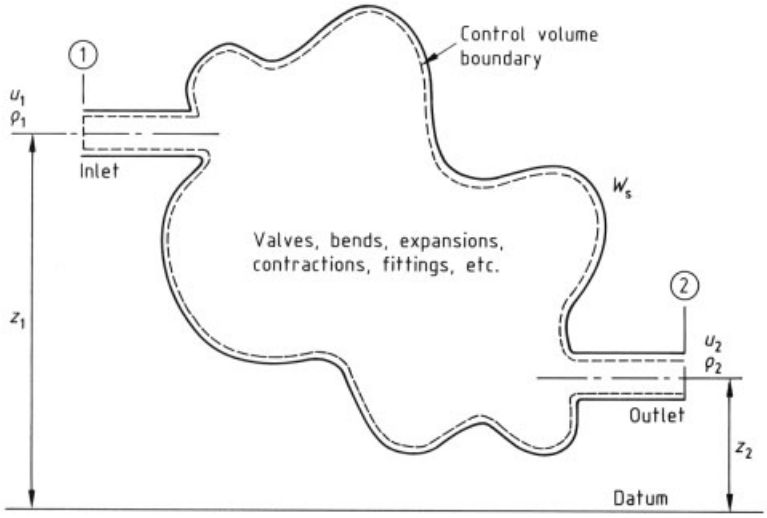

Figure 18. Generalized representation of a control volume for a flow system

3.5.2. Accelerating and Decelerating Flows

In any flow system (in addition to fully developed flows in long straight pipes) it is essential that energy losses which occur in the accelerating and decelerating flow in valves, bends, expansions, contractions, and fittings can be estimated. For turbulent flow these losses must be evaluated from available experimental data, whereas numerical simulation results are now becoming available for some laminar flows.

As an example consider the flow through a sudden contraction in pipe diameter as is illustrated in Figure 20. Application of the mechanical energy balance (Eq. 84) between stations 1 and 2 yields

$$\frac{p_1 - p_2}{\varrho} = \frac{1}{2}\Delta\frac{\langle u_z^3\rangle}{\langle u_z\rangle} + Ev_{1,2} \tag{91}$$

Suppose the flow is turbulent, then

$$\frac{1}{2}\Delta\frac{\langle u_z^3\rangle}{\langle u_z\rangle} = \frac{1}{2}\Delta V^2 = \frac{V_2^2 - V_1^2}{2} \tag{92}$$

For incompressible flow, conservation of mass (Eq. 76) yields

$$V_2 = V_1\left(\frac{D_1}{D_2}\right)^2 \tag{93}$$

and then

$$\frac{1}{2}\Delta V^2 = \frac{V_2^2}{2}\left[1 - \left(\frac{D_2}{D_1}\right)^4\right] = \frac{V_2^2}{2}, \quad \text{if } D_1 >> D_2 \tag{94}$$

Also

$$Ev_{1,2} = Ev_{\mathrm{fd}_{1,2}} + Ev_{\mathrm{ex}_{1,2}} \tag{95}$$

Then

$$Ev_{\mathrm{ex}_{1,2}} = \left(\frac{p_1 - p_2}{\varrho} - \frac{V_2^2}{2}\right) - Ev_{\mathrm{fd}_{1,2}} \tag{96}$$

Now $Ev_{\mathrm{fd}_{1,2}}$ will be known at a fixed flow rate, as will V_2. Hence measurement of $p_1 - p_2$ is sufficient to define $Ev_{\mathrm{ex}_{1,2}}$. It is found that the resulting measurements for most situations can be generalized in the form

$$\frac{Ev_{\mathrm{ex}}}{\frac{V^2}{2}} = C \tag{97}$$

where C is the dimensionless *loss coefficient*. Loss coefficients are tabulated in many standard handbooks such as in [17]. A great deal of confusion can arise in the use of loss coefficients. Sometimes C is referred to as the *number of velocity heads*. This number times $V^2/2$ yields the excess energy loss for that particular fitting. If loss coefficients are used to estimate losses in valve and fittings, etc., then the total frictional loss for a flow system is simply the sum of the individual components:

$$Ev_{\text{total}} = \sum \frac{2fV^2L}{D} + \sum \frac{CV^2}{2} \tag{98}$$

Another and perhaps more convenient way to determine $E\,v_{\text{total}}$ for a flow system is to use the Fanning equation in the form

$$Ev_{\text{total}} = \sum \frac{2fV^2L}{D} + \sum \frac{2fV^2L_{\text{e}}}{D} \tag{99}$$

where L is the length of straight pipe of diameter D between fittings and L_{e} the approximate length of straight pipe of diameter D which would have the same energy loss as the fitting. Equivalent length data for various fittings are available in standard references, for example in [18]. A sufficient amount of data for C and L_{e} have been tabulated for good design of turbulent pipe flow systems. Very little data are available for laminar flows and almost no results are available for non-Newtonian fluids in both laminar and turbulent flows.

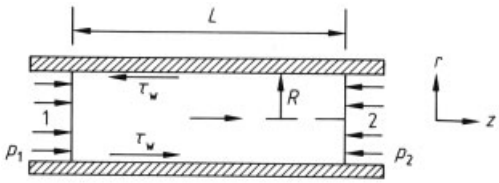

Figure 19. Control volume for fully developed tube flow

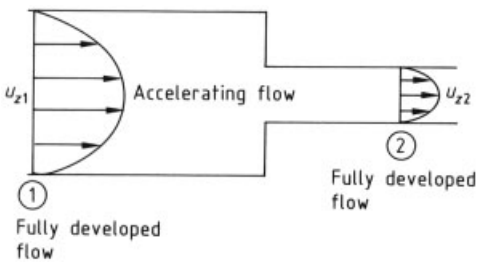

Figure 20. Schematic diagram for flow through a sudden contraction

3.5.3. The Orifice Plate and Similar Flow Rate Measurement Devices

In many chemical plants a simple device known as an orifice plate, (see Fig. 21) is used to measure the volumetric flow rate of fluids in turbulent pipe flow. In its simplest form the orifice plate is a disk with usually (but not always) a circular hole through which the fluid flows. Consider station 1 upstream of the plate and station 2 at or just downstream of the plate. Station 1 is sufficiently far upstream so that the flow there is undisturbed by the presence of the orifice plate. Neglecting losses due to fluid friction, the average pressure and velocities between these two stations are related by Equation (83). In addition, from a mass balance

$$u_1A_1 = u_2A_2 \tag{100}$$

Combining these two equations leads to

$$Q = u_1A_1 = A_1\sqrt{\frac{2\Delta p}{\varrho\left(\frac{A_1^2}{A_2^2}-1\right)}} \tag{101}$$

Equation (101) defines the volumetric flow rate Q, in the pipe in terms of the measured pressure difference between stations 1 and 2. The losses between 1 and 2 are not taken into consideration. Losses can be accounted for by the incorporation of an empirical correction factor known as the discharge coefficient C_{d} into the equation for volumetric flow rate

$$Q = C_{\text{d}}A_1\sqrt{\frac{2\Delta p}{\varrho\left(\frac{A_1^2}{A_2^2}-1\right)}} \tag{102}$$

C_{d} is usually in the range 0.59 – 0.65. Its exact value depends on the geometry of the orifice plate, such as the size and shape of the hole and the location of the pressure measurement points. It also depends, to a lesser extent, on the Reynolds number in the pipe. C_{d} can either be determined by direct calibration or, in the case of an orifice plate of standard design, from tables and correlations published by organisations such as the ISO.

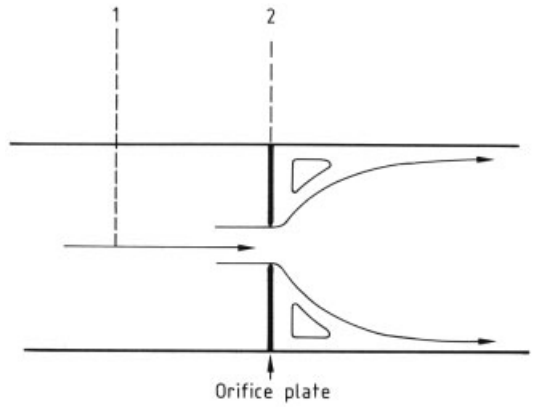

Figure 21. Flow through an orifice plate

A number of other flow measurement devices based on the Bernoulli equation with an empirical correction for losses are also in common use. Examples are the *pitot tube* and the *venturi meter*. In each case the volumetric flow rate is obtained from the measured pressure difference using an equation similar to that for the orifice plate. The value of C_d must be determined for each device.

4. Non-Newtonian Fluids [4], [19]

For fluids such as polymer solutions, molten plastics, emulsions, suspensions and many others, the deviatoric stress tensor may not be directly proportional to the rate of strain tensor and the simple Newtonian constitutive equation does not describe the flow behavior of these fluids. For such fluids, the viscosity, defined as the ratio of shear stress to shear rate, is now a function of the rate of strain tensor $\boldsymbol{e}$:

$$\boldsymbol{\tau} = 2\eta(\boldsymbol{e})\,\boldsymbol{e} \tag{103}$$

Equation (103) is the generalized Newtonian constitutive equation. The viscosity of some non-Newtonian fluids can change by several orders of magnitude as a result of shearing. Such behavior has serious practical consequences for the handling of these fluids. Apart from their variable viscosity, these complex fluids often exhibit other flow behavior – such as stress relaxation, normal stresses, and stresses that depend on strain history – which are not observed in Newtonian fluids and also cannot be accounted for by the generalized Newtonian constitutive equation. Some non-Newtonian fluids exhibit some of the characteristics of an elastic solid and thus are referred to as *viscoelastic fluids*. The simple imaginary experiment shown in Figure 22 helps to visualize the difference between inelastic and viscoelastic fluids. Imagine a fluid confined between two flat plates where a force is applied to the upper plate, resulting in the displacement of a fluid particle just below the upper plate through a distance dx. Also imagine a similar experiment for an elastic solid. The work done in displacing the fluid or solid particle a distance dx is then F dx. If an inelastic fluid is confined between the flat plates, on removing the force F, the fluid particle at point A will remain at point A (Fig. 22A). All of the work done on an inelastic fluid is dissipated as heat and cannot be recovered. After removing the force acting on the ideally elastic or Hookean solid, the solid particle at point A will recoil to

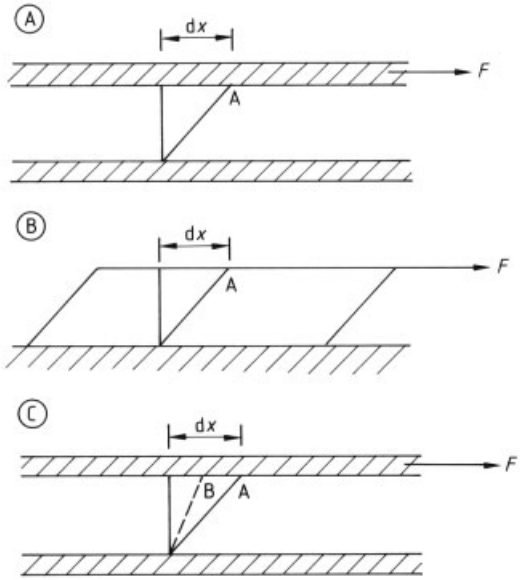

Figure 22. Inelastic and viscoelastic fluids
A) Viscous, inelastic liquid, work dissipated as heat; B) Elastic solid, work recoverable; C) Viscoelastic fluid, part of work recoverable

its original position (Fig. 22B). All of the work done on an elastic solid is recoverable. For a viscoelastic fluid partial elastic recovery is observed. The fluid particle at point A recovers to an intermediate position B (Fig. 22C). A portion of the work done on a viscoelastic fluid is recoverable and the remainder is dissipated as heat. In fact, the amount of recovery depends on the rate at which the displacement dx is imposed. In general, increasing the displacement rate increases the recovery. An inelastic fluid can be described sufficiently by measuring its viscosity and by determining the variation of viscosity with shear rate, time of shear, and, in some instances, thermal history. The viscosity is a steady state concept and is also applicable to viscoelastic fluids without any modification. In addition to the viscosity, however, there are other fundamental properties which must be measured to adequately characterize a viscoelastic fluid. The distinction between inelastic and viscoelastic fluids is clearly one of time scales. If man-

ifestation of recoverable elastic deformation is observed on a time scale of interest, then the material is viscoelastic. In practical terms molecular relaxation times of $< 10^{-3}$ s can probably be ignored in all cases. Viscoelastic fluids will be dealt with in Section 4.3.

4.1. Classification of Fluids According to Viscosity Behavior

In Section 3.1 it has been shown that in one-dimensional flows the shear rate $\dot{\gamma}$ is equal to the velocity gradient for that flow. Since the viscosity is measured in one-dimensional flows (viscometric flows) the fluid behavior is classified according to the observed shear stress – shear rate behavior in one dimension. A graph of the measured shear stress as a function of shear rate is called the *flow curve*.

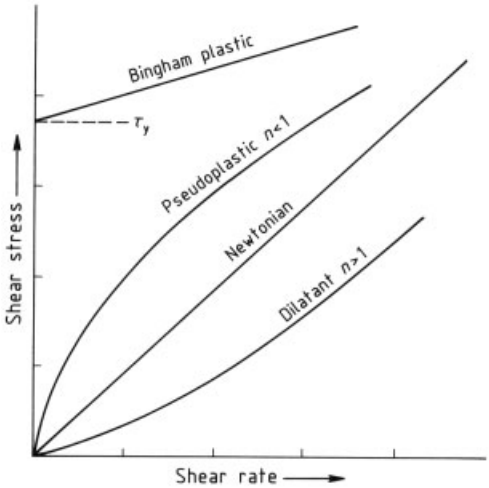

Figure 23. Typical flow curves

4.1.1. Bingham Type Behavior

Bingham type material exhibits a *yield stress*. That is, the shear stress τ must exceed a certain yield value τ_y before the fluid deforms and flows. Classical Bingham plastic behavior is illustrated in Figure 23 and can be described by

$$\tau = \left(\frac{\tau_y}{|\dot{\gamma}|} + \eta_{pl}\right)\dot{\gamma} \text{ for } \tau \geq \tau_y \tag{104a}$$

i.e.,

$$\eta = \frac{\tau_y}{|\dot{\gamma}|} + \eta_{pl} \text{ for } \tau \geq \tau_y \tag{104b}$$

The stress is undetermined when $|\tau| < \tau_y$. The term $\dot{\gamma}/|\dot{\gamma}|$ is necessary to ensure that the stress takes on the same sign as the shear rate and η_{pl} is the *plastic viscosity.* The true viscosity η approaches η_{pl} only when $\eta_{pl}|\dot{\gamma}| >> \tau_y$. In contrast the viscosity for the Bingham and any yield stress material becomes unbounded as $|\dot{\gamma}| \to 0$.

When the shear stress – shear rate function is nonlinear the data are often fit by a *Herschel – Buckley model*:

$$\tau = \left(\frac{\tau_y}{|\dot{\gamma}|} + k\dot{\gamma}^{n-1}\right)\dot{\gamma} \text{ for } |\tau| \geq \tau_y \tag{105a}$$

$$\eta = \frac{\tau_y}{|\dot{\gamma}|} + k|\dot{\gamma}|^{n-1} \text{ for } |\tau| \geq \tau_y \tag{105b}$$

The parameters *k* and *n* are obtained by plotting $\tau - \tau_y$ versus $\dot{\gamma}$ on logarithmic coordinates. Another equation often used to describe yield stress materials is the *Casson equation*

$$\tau = \left(\frac{\tau_y^{1/2}}{|\dot{\gamma}|^{1/2}} + \eta_{pl}^{1/2}\right)^2 \dot{\gamma} \text{ for } |\tau| \geq \tau_y \tag{106a}$$

$$\eta = \left(\frac{\tau_y^{1/2}}{|\dot{\gamma}|^{1/2}} + \eta_{pl}^{1/2}\right)^2 \text{ for } |\tau| \geq \tau_y \tag{106b}$$

Typical shear stress – shear rate data for a meat extract concentrate and for a tomato soup concentrate are presented in Figures 24 and 25, respectively. The meat extract behaves like a Bingham solid while the tomato soup concentrate exhibits Herschel – Buckley behavior.

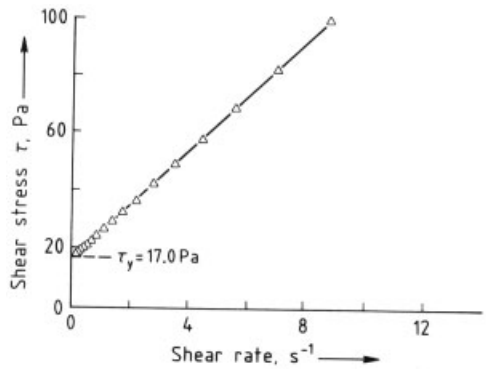

Figure 24. Shear stress – shear rate data (flow curve) for a meat extract

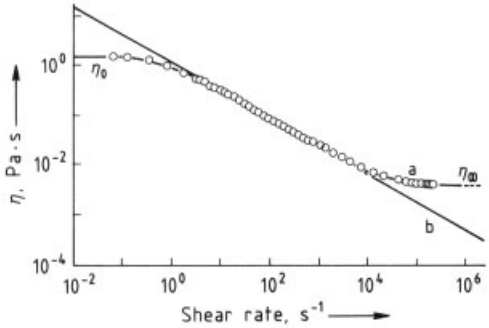

Figure 26. Viscosity as a function of shear rate for a polyacrylamide solution

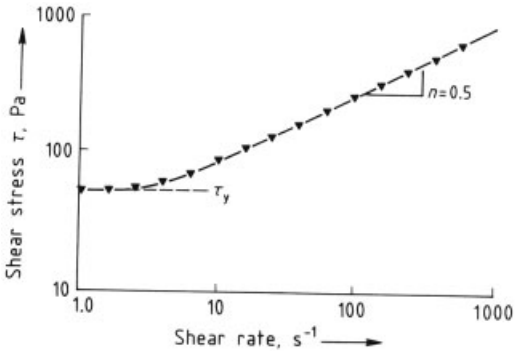

Figure 25. Shear stress – shear rate data (flow curve) for a tomato soup concentrate

The yield stress is associated with a three-dimensional structure that must deform elastically before flow can occur. It is commonly observed in fine particle suspension systems, especially at high particle concentrations. Common examples are toothpaste, paint, oil well drilling muds, foams, and many food and pharmaceutical products.

4.1.2. Shear-Thinning (Pseudoplastic) Fluids

Solutions and melts of flexible macromolecules and many suspensions exhibit pseudoplastic shear stress – shear rate behavior such as that illustrated in Figure 23 and shown specifically for a polymer solution in Figure 26. Here the ratio $\tau/\dot{\gamma}$ (i.e., η) is a decreasing function of $\dot{\gamma}$. Such a fluid has a shear-thinning viscosity and is referred to as pseudoplastic. The decrease of viscosity with increasing rate of deformation can usually be attributed to the breakdown of a structure at the colloidal or molecular level. Macromolecules will become more aligned and hence less entangled and less resistant to deformation as the deformation rate is increased.

The viscosity of pseudoplastic fluids is typically plotted versus shear rate on logarithmic coordinates because the viscosity for such fluids can change by orders of magnitude over several decades of shear rate. The data for the polyacrylamide solution shown in Figure 26 are typical for a macromolecular fluid. The viscosity approaches a constant value at low shear rates called the *zero-shear viscosity* η_0 and a constant value in the limit of very high shear rates called the *infinite shear viscosity* η_∞. Experimental measurements in both regions can be difficult, particularly in the infinite shear region. Thus data are often only available between the two limits. Between the two limits the shear stress (or viscosity) is often linear with shear rate over several decades on logarithmic coordinates. In this restricted region

$$\log\tau = \log k + n\log\dot{\gamma} \tag{107a}$$

$$\log\eta = \log k + (n-1)\log\dot{\gamma} \tag{107b}$$

where n is the slope in the linear region of the plot ($n = 0.41$ for the data in Fig. 26). Thus

$$\tau = k|\dot{\gamma}|^{n-1}\dot{\gamma} \tag{108a}$$

$$\eta = k|\dot{\gamma}|^{n-1} \tag{108b}$$

Equation (108) is known as the *power law* and sometimes as the Ostwald – de Waele model. This law is empirical and must fail for both high and low shear rates. The factor k is very temperature sensitive (and concentration sensitive for suspensions), but n is typically insensitive to temperature (and concentration) changes.

Many empirical and semi-empirical equations have been proposed to represent the viscosity data for shear-thinning fluids. The *Meter model*:

$$\eta = \eta_\infty + \frac{\eta_0 - \eta_\infty}{1 + \left|\frac{\tau}{\tau_{1/2}}\right|^{\alpha-1}} \tag{109}$$

has four parameters, η_0, η_∞, $\tau_{1/2}$, and α with the following asymptotic behavior

$$\frac{\tau}{\tau_{1/2}} \to 0,\ \eta \approx \eta_0 \tag{109a}$$

$$\frac{\tau}{\tau_{1/2}} \to \infty,\ \eta \approx \eta_\infty \tag{109b}$$

and for

$$1 << \left| \frac{\tau}{\tau_{1/2}} \right|^{\alpha-1} << \frac{\eta_0}{\eta_\infty}$$

then

$$\eta = \frac{\eta_0 - \eta_\infty}{\tau_{1/2}^{1-\alpha}} |\tau|^{1-\alpha} \tag{109c}$$

The Meter model exhibits power-law behavior in the intermediate region, with $\alpha = 1/n$; $\tau_{1/2}$ is the stress at which the viscosity is midway between the values of η_0 and η_∞. If η_∞ is taken to be zero, which is often a good approximation for molten polymers, then $\tau_{1/2}$ is the stress at which the viscosity drops to one-half the zero-shear value. In this case Equation (109) is known as the *Ellis model*. Line (a) through the data in Figure 26 is the fit of the Meter model to the data with $\eta_0 = 1.43$ Pa s, $\eta_\infty = 0.004$ Pa s, $\tau_{1/2} = 1.21$ Pa, and $\alpha = 2.43$. The power-law fit is line (b) with $n = 0.41$ ($= 1/\alpha$). The Ellis model follows the Meter model at low shear rates (incorporating the zero-shear viscosity behavior) and the power-law at higher shear rates.

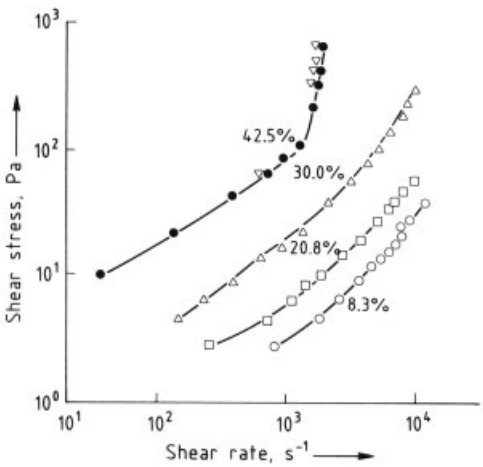

Figure 27. Shear stress – shear rate for a titanium dioxide suspension [20]

4.1.3. Shear-Thickening (Dilatant) Fluids

Some suspensions with a high concentration of solids and with a relatively uniform particle-size distribution have a viscosity that increases with shear rate. An example for a titanium dioxide suspension is shown in Figure 27 where at the highest volume fraction of solids (42.5 %) the viscosity increases with shear rate beyond a critical shear rate. The increase in stress observed at this point is related to a shear-induced structure. The phenomena is easily illustrated in the kitchen by mixing a concentrated suspension of corn starch and water. Data for shear-thickening liquids are usually fit by equations such as those for shear-thinning fluids but with $\eta_\infty = 0$ and $n > 1$ (or $\alpha < 1$). Dramatic equipment failures have been documented in the processing industries where a dilatant fluid response was not anticipated.

4.1.4. Fluids with Time-Dependent Viscosity (Thixotropic Fluids)

A special nomenclature is used in cases when the viscosity varies with the time of shear and the shear rate. A material whose apparent viscosity *decreases* with the time of shear is said to be *thixotropic* and one whose viscosity *increases* with time of shear is said to be *rheopectic*. The behavior of these materials depends not only on the time of shear but also on the past shear and thermal history. Rheopectic fluids are encountered very rarely and hence the discussion in this section will be confined to thixotropic fluids.

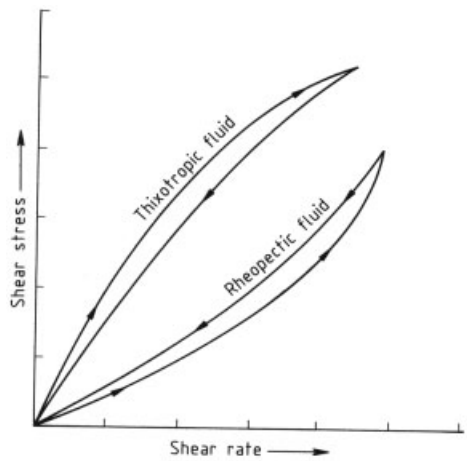

Figure 28. Time dependent fluids

If the flow curve is measured in a single experiment in which the shear rate is steadily increased at a constant rate from zero to some maximum value and then decreased at the same rate

to zero shear rate, a form of a hysteresis loop as is shown in Figure 28 is observed. The position of this loop, its shape, and the area within the loop depends on the rate at which the shear rate is increased and decreased as well as the past thermal and shear history of the material. This type of measurement has little meaning for the solution of practical flow problems. What is required is a series of experimentally determined shear stress – shear rate – time curves which can then be converted to a series of flow curves with time of shear as the parameter. Because of the dependence of these curves on shear and thermal history it is important that the curves represent the behavior of a sample of material which has experienced a history approximately the same as that anticipated for the fluid whose flow behavior is to be analyzed.

The difference between thixotropy and pseudoplasticity is thought to be the time element in structural breakdown, which is finite and measurable for the thixotropic fluid and in contrast very small and undetectable for the pseudoplastic fluid. A practical example of time- dependent behavior is the waxy crude oils which in addition also exhibit a yield stress (see Fig. 29). It should be noted that a fluid can show time-dependent and viscoelastic behavior (for viscoelastic behavior see Section 4.3). Some liquid crystal polymers appear to be rheopectic and viscoelastic.

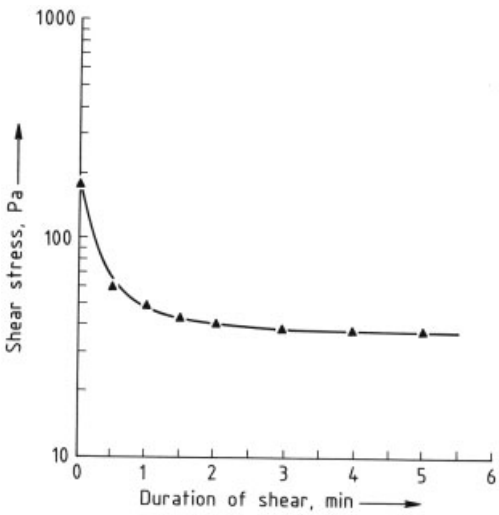

Figure 29. Shear stress versus time of shear for an Australian waxy crude oil at 10 °C and a shear rate of 458 s^{-1}

4.2. Fully Developed Tube Flow

The flow in tubes (pipes) of circular cross section is the flow field of the greatest interest. For non-Newtonian fluids this is not only important because of processing applications but also because of the extreme importance of tube flow in the measurement of viscosity. Both laminar and turbulent flows will be considered. However, because of the sticky, gooey, or highly viscous nature of non-Newtonian fluids, most applications are in laminar flow. For instance a typical Reynolds number in polymer processing might be 10^{-3}. The energy required to reach turbulent flow for many non-Newtonian fluids is simply prohibitive. The treatment here is confined to fully developed tube flows, i.e., flow in long straight tubes where the velocity profile is not a function of axial position. Important results required from the discussion of Newtonian fluids include the definition of the friction factor (cf. Eq. 48)

$$f=\frac{\tau_{\mathrm{w}}}{\frac{\varrho V^2}{2}}$$

the general results for the shear stress in tube flow, i.e., Equations (88) and (87)

$$\tau_{rz}=\frac{r\Delta p}{2L}$$

$$\tau_{\mathrm{w}}=\frac{R\Delta p}{2L}=\frac{D\Delta p}{4L}$$

and (Eq. 89)

$$\tau_{rz}=\frac{r}{R}\tau_{\mathrm{w}}$$

and Equation (44) which defines the volumetric flow rate

$$Q=\int_0^R 2\pi r u_z \mathrm{d}r$$

The results obtained for fully developed tube flow will be valid for time-independent inelastic and viscoelastic fluids. Elastic effects are important in accelerating and decelerating flows and in time- dependent flow fields only.

4.2.1. Volumetric Flow Rate – Pressure Drop Relationship

Equation (44) can be used to obtain a general relationship between Q and τ_w for fully developed tube flows. Integration by parts yields

$$Q=\pi\left[u_z r^2-\int r^2 \mathrm{d}u_z\right]_0^R \tag{110}$$

At the tube wall ($r=R$) $u_z=0$, then

$$Q=-\pi\int_0^R r^2 \mathrm{d}u_z \tag{111}$$

The shear rate in fully developed tube flow is $\mathrm{d}u_z/\mathrm{d}r$ and is a unique function of the shear stress τ_{rz}:

$$\frac{\mathrm{d}u_z}{\mathrm{d}r}=-f(\tau_{rz}) \tag{112}$$

The minus sign is introduced because $\mathrm{d}u_z/\mathrm{d}r$ is negative (u_z decreases from a maximum value at $r=0$ to zero at $r=R$). Then

$$du_z=-f(\tau_{rz})\,\mathrm{d}r \tag{113}$$

Eliminating r from Equations (111) and (113) with Equation (89) and then substituting for $\mathrm{d}u_z$ in Equation (111) with Equation (113) yields

$$\frac{Q}{\pi R^3}=\frac{1}{\tau_w^3}\int_0^{\tau_w}\tau_{rz}^2 f(\tau_{rz})\,\mathrm{d}\tau_{rz} \tag{114}$$

Equation (114) is an important result which can be used in two ways: (1) *integrated* directly for a particular fluid model such as those represented by Equations (105), (106), (108), and (109) or (2) *differentiated* which leads to the definition of the wall shear rate for any fluid.

Consider the integration of Equation (114) for a power-law fluid. For tube flow of a power-law fluid

$$\tau_{rz}=-k\left(\frac{\mathrm{d}u_z}{\mathrm{d}r}\right)^n \tag{115}$$

and

$$f(\tau_{rz})=-\frac{\mathrm{d}u_z}{\mathrm{d}r}=\left(\frac{\tau_{rz}}{k}\right)^{1/n} \tag{116}$$

Substitution for $f(\tau_{rz})$ in Equation (114) and integration leads to

$$\frac{Q}{\pi R^3}=\frac{n}{(3n+1)}\left(\frac{\tau_w}{k}\right)^{1/n}=\frac{n}{(3n+1)}\left(\frac{R\Delta p}{2Lk}\right)^{1/n} \tag{117}$$

For $n=1$ Equation (117) reduces to the Hagen – Poiseuille equation derived in Section 3.3.1. The pressure drop for a pseudoplastic fluid is relatively insensitive to changes in the flow rate as compared to a Newtonian fluid. Thus flow rate measuring devices like the orifice plate that rely on the measurement of differential pressure are not very suitable for shear-thinning fluids, particularly if the flow behavior index n is low.

Equation (114) can be integrated for any of the semi-empirical equations which have been proposed to represent shear stress – shear rate data. The volumetric flow rate – wall shear stress results obtained by this integration for a number of fluids may be summarized as follows ($\tau_w=R\,\Delta p/2\,L$):

Newtonian fluid	$\frac{Q}{\pi R^3}=\frac{1}{4}\frac{\tau_w}{\eta}$
Bingham plastic	$\frac{Q}{\pi R^3}=\frac{\tau_w}{4\eta_{pl}}\left[1-\frac{4}{3}\frac{\tau_y}{\tau_w}+\frac{1}{3}\left(\frac{\tau_y}{\tau_w}\right)^4\right]$
Power-law	$\frac{Q}{\pi R^3}=\frac{n}{3n+1}\left(\frac{\tau_w}{k}\right)^{1/n}$
Ellis fluid	$\frac{Q}{\pi R^3}=\frac{1}{4\eta_0}\tau_w+\frac{\tau_w^{\alpha}}{(\alpha+3)\eta_0\tau_{1/2}^{\alpha-1}}$

4.2.2. Generalized Treatment

If Equation (114) is multiplied by τ_w^3 and differentiated with respect to τ_w

$$-\left.\frac{\mathrm{d}u_z}{\mathrm{d}r}\right|_{\tau_w}=f(\tau_w)=\frac{1}{\pi R^3}\left[3Q+\Delta p\frac{\mathrm{d}Q}{\mathrm{d}\Delta p}\right] \tag{118}$$

where $f(\tau_w)$ is the wall shear rate $\dot{\gamma}_w$. In arriving at the above, Equation (87) was used to replace τ_w by the more directly measurable quantity Δp. Equation (118) can be rearranged to

$$\dot{\gamma}_w=\left(\frac{3n'+1}{4n'}\right)\left(\frac{8V}{D}\right) \tag{119}$$

where

$$n'=\frac{\mathrm{d}\ln\tau_w}{\mathrm{d}\ln\frac{8V}{D}} \tag{120}$$

and V is the average velocity in the tube $\left(V=\frac{Q}{\pi R^2}\right)$.

Equation (118) and the rearranged form (Eq. 119) is the *Rabinowitsch – Mooney equation* which enables the calculation of the wall

shear rate from the measurable quantities, R, Q or V, and Δp. Thus measurements in fully developed tube or capillary flow can be used to determine the shear stress (Eq. 87) and the shear rate (Eq. 119 and 120) at the tube wall. Since the shear stress – shear rate curve is a function of the material and not the apparatus used, a plot of τ_w versus $\dot{\gamma}_w$ is one way in which a flow curve can be determined for any time-independent fluid.

The parameter n' is the slope of the log – log plot of τ_w versus 8 V/D and must not be confused with the flow-behavior index n in the power law. The flow-behavior index n is constant over the range of shear rates for which it is defined, whereas n' can vary with 8 V/D. Since n' in Equation (120) represents the slope of a line, the line may be represented by an equation of the form

$$ln\tau_w = n' \ln 8V/D + \ln K' \tag{121}$$

If n' is not constant, but varies with τ_w then Equation (121) represents the tangent to the curve at a given point. On the other hand, if n' is constant, Equation (121) represents the actual relationship between τ_w and 8 V/D. When n' is constant, it is useful to write Equation (121) in the form

$$\tau_w = K' \left(\frac{8V}{D}\right)^{n'} \tag{122}$$

Comparing Equation (122) to Equation (117) shows that for a power-law fluid, $n = n'$ and

$$k = K' \left(\frac{4n}{3n+1}\right)^{n} \tag{122a}$$

It is not necessary to compute the slope of the experimentally determined log – log plot of τ_w versus 8 V/D if calculations for laminar tube flow are to be made. The plot can be used directly for design or process evaluation, provided that process conditions are within the range of the experimental data. In fact, a single experimental determination of τ_w in a convenient model at a given value of V/D enables the prediction of the pressure drop in a larger pipe of the same value of V/D.

The terms n' and K' need only be determined when the τ_w versus 8 V/D data are to be used for other laminar flow geometries or for turbulent flow. It is important to realize that n' and K', as determined from the log – log plot, are only valid for the shear rate or the 8 V/D range in which they were determined and cannot be blindly used for design or process evaluation if the process shear rates are out of this range.

4.2.3. Velocity Distribution

For fully developed laminar flow, the velocity distribution in a tube for any time-independent fluid is obtained by integrating the velocity gradient from the tube wall $r = R$ to any radial position $r = r$, assuming no slip at the wall:

$$u_z(r) = \int_r^R \left(\frac{-\mathrm{d}u_z}{\mathrm{d}r}\right) \mathrm{d}r \tag{123a}$$

or

$$u_z(r) = \frac{R}{\tau_w} \int_{\tau_{rz}}^{\tau_w} f(\tau_{rz})\, \mathrm{d}\tau_{rz} \tag{123b}$$

For a power-law fluid where $f(\tau_{rz}) = (\tau_{rz}/k)^{1/n}$ and remembering that $\tau_{rz} = \frac{r}{R}\tau_w$ and $\tau_{rz} = r\,\Delta p/2\,L$ yields

$$u_z(r) = \frac{n+1}{R^n}\left(\frac{\Delta p}{2kL}\right)^{\frac{1}{n}}\left(\frac{n}{n+1}\right)\left[1-\left(\frac{r}{R}\right)^{\frac{n+1}{n}}\right] \tag{124}$$

when Equation (123 b) is integrated. In terms of the average velocity

$$V = \frac{2\pi}{\pi R^2}\int_0^R r u_z \mathrm{d}r = R^{\frac{(n+1)}{n}}\left(\frac{\Delta p}{2kL}\right)^{\frac{1}{n}}\left(\frac{n}{3n+1}\right) \tag{125}$$

and hence

$$u_z(r) = \left(\frac{3n+1}{n+1}\right) V \left[1-\left(\frac{r}{R}\right)^{\frac{n+1}{n}}\right] \tag{126}$$

Equation (126) defines the velocity distribution for *power-law fluids* in fully developed tube flow. The velocity gradient at the tube wall is

$$\left.\frac{\mathrm{d}u_z}{\mathrm{d}r}\right|_{r=R} = \dot{\gamma}_w = \frac{3n+1}{4n}\left(\frac{8V}{D}\right) \tag{127}$$

As expected Equation (127) is identical to Equation (119) when $n = n'$ but more importantly for a *Newtonian fluid* ($n' = n = 1$)

$$\dot{\gamma}_{\mathrm{w}}=\frac{8V}{D} \tag{128}$$

Thus 8 *V/D* is the wall shear rate for a Newtonian fluid in laminar tube flow.

For a *Bingham material* similar calculations yield the following velocity distribution:

$$u_z(r)=\frac{R^2}{4\eta_{\mathrm{pl}}}\left(\frac{\Delta p}{L}\right)\left(1-\frac{r_0^2}{R}\right) \quad 0\le r\le r_0 \tag{129a}$$

and

$$u_z(r)=\frac{R^2}{4\eta_{\mathrm{pl}}}\frac{\Delta p}{L}\left[1-\left(\frac{r}{R}\right)^2\right]-\frac{\tau_{\mathrm{y}}R}{\eta_{\mathrm{pl}}}\left[1-\left(\frac{r}{R}\right)\right]$$
$$r_0\le r\le R \tag{129b}$$

where $r_0=2L\tau_{\mathrm{y}}/\Delta p$ and η_{pl} is the plastic viscosity of the fluid. Equations (129 a) and (129 b) show that there is a region of radius r_0, in the center of the pipe, where the fluid moves as a solid plug with a uniform velocity. The volumetric flow rate in a pipe for a Bingham material is given in Section 4.2.1, where τ_{y} is the yield stress and τ_{w} is the wall shear stress. This equation is only valid when τ_{w} is greater than τ_{y}. Flow in the pipe will not occur if τ_{w} is less than τ_{y}. Thus

$$\frac{\Delta p}{L}>\frac{2\tau_{\mathrm{y}}}{R} \tag{130}$$

is necessary for the initiation of flow in a pipe for yield stress materials.

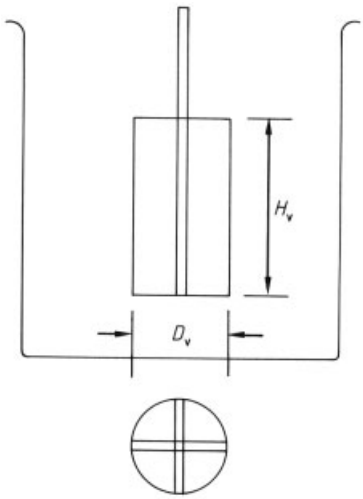

Figure 30. The vane used for yield stress measurement D_{v} = vane diameter; H_{v} = vane height

The yield stress of a material can be determined by extrapolation of the basic shear stress – shear rate data to zero shear rate. However, in view of the importance of the yield stress special techniques have been developed for its measurement. In one such technique, a vane is inserted into the fluid (see Fig. 30). The minimum torque Γ_{m} required to initiate a rotation motion in the vane is recorded. The yield stress is then related to this torque by

$$\Gamma_{\mathrm{m}}=\frac{\pi D_{\mathrm{v}}^3}{2}\left(\frac{H_{\mathrm{v}}}{D_{\mathrm{v}}}+\frac{1}{3}\right)\tau_{\mathrm{y}} \tag{131}$$

In deriving Equation (131) it is assumed that the yielding surface is the cylindrical surface and the two end circular disks traced out by the vane. For short vanes, end effects may become important and must be taken into consideration [21].

4.2.4. Friction Factor – Reynolds Number Relationships

General procedures applicable to all types of fluids are of course desirable because of the difficulty in classifying a fluid as one type or another, and because of the tendency of some fluids to change from one type to another with changing shear rate, temperature, or composition. At present generalizations are established only for time-independent fluids.

Laminar Flow. For laminar flow of Newtonian fluids in circular pipes, the friction factor satisfies the relation (see Section 3.3.1 and Equation (48))

$$f=\frac{16}{Re}$$

with the Reynolds number defined as $Re=(D\,V\,\varrho)/\eta$. For the laminar flow of any time-independent fluid in a tube, the wall shear stress is given by

$$\tau_{\mathrm{w}}=K'\left(\frac{8V}{D}\right)^{n'} \tag{132}$$

where, in general, K' and n' are not constant, but depend on 8 *V/D*. When Equation (132) is substituted into the defining equation for the friction factor one obtains

$$f = \frac{16}{Re'} \tag{133}$$

where

$$Re' = \frac{D^{n'} V^{2-n'}}{K' 8^{n'-1}} \tag{134}$$

which is a generalized Reynolds number valid for all time-independent fluids. In the special case where $n' = 1$ and $K' = \eta$, Equation (134) reduces to the Reynolds number for a Newtonian fluid. For a power-law fluid $n' = n$ and Equation (122 a) is valid.

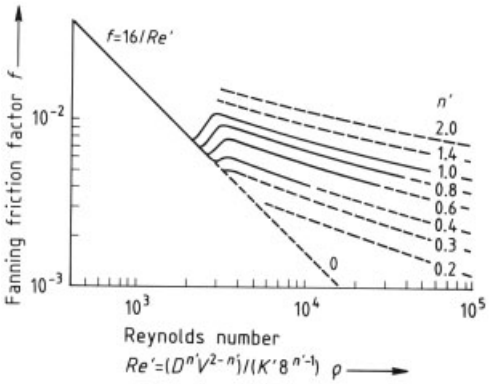

Figure 31. Friction factor design chart for non-Newtonian fluids
—— Experimental regions; – – – extrapolated regions

Transition Flow. So far only the laminar flow of non-Newtonian fluids in tubes has been considered and the important question, when is the flow laminar, has been avoided. For Newtonian fluids, the question is answered by calculating the Reynolds number and comparing it to the empirically determined value of 2100 below which the flow is laminar. A similar criterion has been sought for non-Newtonian fluids by METZNER and REED [22] and DODGE and METZNER [23]. METZNER and REED first considered that stable laminar flow in tubes usually ends when the friction factor becomes < 0.008, which corresponded to a generalized Reynolds number of 2100. In the later work by DODGE and METZNER, which led to the generalized f–Re chart shown in Figure 31, laminar flow was found to end at generalized Reynolds numbers which increase slightly as n' decreases. At n' values of 1.0, 0.726, and 0.38, the transition region appeared to begin at Re' of about 2100, 2700 and 3100, respectively. Others have examined the flow transition question for non-Newtonian fluids but the conservative conclusion remains:

$$Re' \leq 2100 \tag{135}$$

has been accepted and is used to establish whether the flow is laminar or turbulent for a non-Newtonian fluid.

Turbulent Flow. The friction factor chart for time-independent inelastic fluids is reproduced in Figure 31. The friction factor is defined by

$$\sqrt{\frac{1}{f}} = \frac{4.0}{(n')^{0.75}} \log\left[Re' f^{1-n'/2}\right] - \frac{0.4}{n'^{1.2}} \tag{136}$$

Equation (136) reduces to the von Karman–Nikuradse equation (50) for Newtonian fluids when $n' = 1$. The solid lines in Figure 31 are based on experimental results, and the dashed lines are extrapolations made using Equation (136). The constants in Equation (136) were evaluated from experimental data. If the fluid does not obey the power law with an n' and K' which vary with 8 V/D, then a trial-and-error procedure is required to calculate f, τ_w, or Δp. First a value of Δp is assumed, then the approximate values of n' and K' are found from the laminar flow curve (τ_w versus 8 V/D) at the assumed τ_w. From these parameters, Re' is calculated and f is obtained from Figure 31. If the pressure drop calculated from the friction factor does not agree with the original assumed value, another iteration is carried out. The process converges quickly.

4.3. Viscoelastic Fluid Mechanics

Materials such as molten plastics, solutions of synthetic macromolecules and many biological fluids – apart from possessing the properties normally associated with purely viscous fluids – also exhibit a range of behavior which is normally associated with elastic solids. For this reason such materials have been given the name viscoelastic fluids. In a viscoelastic fluid the stress generated by deformation is no longer a simple function of the instantaneous rate of strain. It also depends, as in solids, on the strain and, for

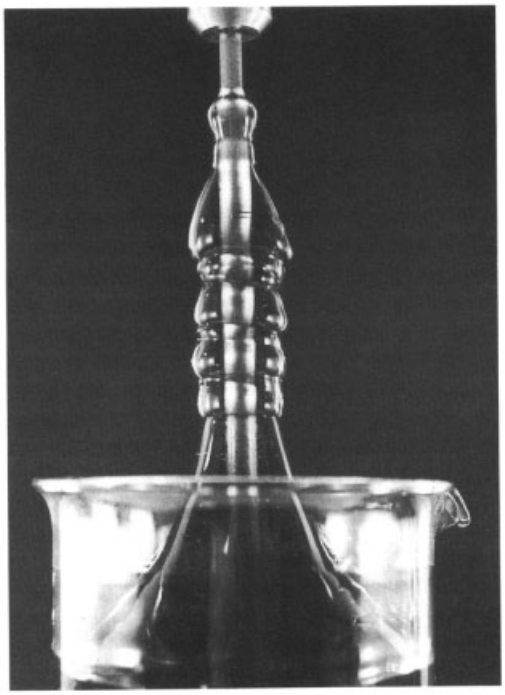

Figure 32. Rod-climbing by a viscoelastic fluid

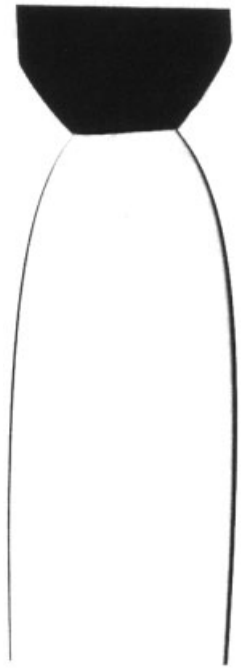

Figure 33. Extrudate swell by a viscoelastic fluid

large strains, on the strain history experienced by the fluid. As discussed earlier, in a purely viscous fluid, all the energy required to produce the deformation is dissipated as heat while in a perfectly elastic solid this energy is stored and is completely recovered when the forces acting on the solid are removed. In a viscoelastic fluid, this energy is partly dissipated and partly stored, the ratio of dissipation to storage depends on the fluid, the nature of the deformation, and the history of the deformation. Thus, depending on the flow, a viscoelastic fluid may exhibit, in varying degrees, the elastic properties of solids and the viscous properties of fluids.

It is therefore not surprising that some of the flow behaviors of viscoelastic fluids have no parallel in purely viscous fluids. A number of these viscoelastic phenomena are reproduced in Figures 32, 33, 34, and 35. Figure 32 shows a rod slowly rotating in a beaker filled with a viscoelastic fluid. The rotation causes the viscoelastic fluid to climb up the rod. Under similar conditions, a purely viscous fluid is undisturbed by the slow rotation and its surface remains essentially flat. Rod climbing by a viscoelastic fluid is often referred to as the *Weissenberg effect*. When a viscoelastic fluid is extruded from a capillary (see Fig. 33) the resulting fluid stream may swell up to several times the capillary diameter. This phenomenon is known as *extrudate swell*. Newtonian fluids also exhibit extrudate swell at small Reynolds number, but the increase in diameter is only 13 %. Large extrudate swell is observed in polymer extrusion and clearly has to be taken into account in the design of extrusion shaping dies. Figure 34 shows a *tubeless syphon*, where a viscoelastic fluid is being sucked vertically upwards into a tube. The suction tube can be lifted to a large height (well over four to five tube diameters in the case shown) above the viscoelastic fluid without the syphoning action being interrupted. In comparison, for a purely viscous fluid, syphoning ceases as soon as the tube is lifted out of the fluid. The tubeless syphon illustrates the large tensile stress that can be sustained in a viscoelastic fluid. Figure 35 is a comparison of the streamline patterns formed by a Newtonian fluid and a viscoelastic fluid as they flow through a pipe contraction. Figure 35A shows a Newtonian fluid; Figures 35B – D show a viscoelastic

fluid at increasing flow rate. For the Newtonian fluid, two small recirculating vortices can be observed at the contraction plane. The size of the Newtonian vortex, shown at low Reynolds number, is not very sensitive to changes in flow rate. For viscoelastic fluids, the situation is quite different. At low flow rates the viscoelastic vortices are approximately of the same size as those for the Newtonian fluid, but they rapidly increase in size, up to four times the size of Newtonian vortex, as the flow rate is increased. The flow then becomes unstable and the vortices begin to pulsate, growing and decreasing in size in some erratic manner. Although the streamline pattern of the Newtonian fluid can be obtained by solving the Navier – Stokes equations, attempts to understand the viscoelastic flow patterns have so far been, at most, only partially successful [24]. The most likely source of the problem here is the present inability to describe, with the necessary degree of accuracy, the complex rheological behavior of the viscoelastic fluid when it undergoes shearing and stretching deformation as it flows through the contraction.

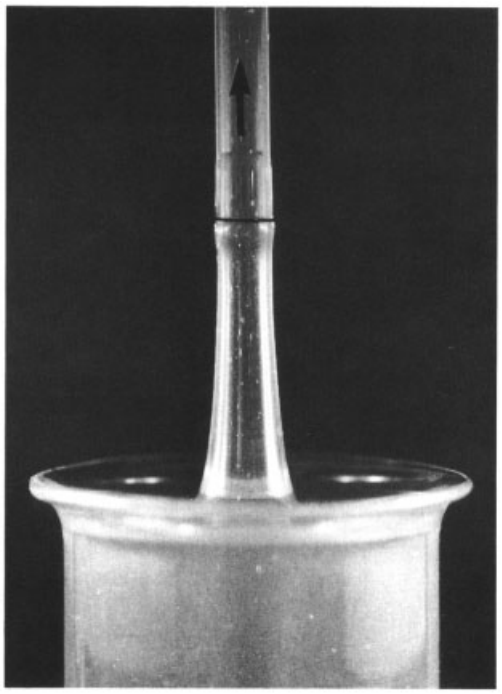

Figure 34. Viscoelastic fluid being sucked up (tubeless syphon)

Some of the more fundamental aspects of viscoelastic fluid mechanics will be discussed in the next section. The treatment is, however, of a more descriptive nature.

4.3.1. Steady Shear Behavior of Viscoelastic Fluids

Under steady shear the shear stress T_{xy} generated in a viscoelastic fluid is related to the shear rate by

$$T_{xy} = \eta(\dot{\gamma})\,\dot{\gamma}$$

where η is the shear-rate dependent viscosity of the fluid. Figure 26 is the plot of the viscosity of a solution of polyacrylamide, a typical viscoelastic fluid. Like most non-Newtonian fluids, most viscoelastic fluids exhibit shear thinning, and, over a range of shear rates, behave like a power-law fluid. However, if the normal components of the stress tensor in the x direction, T_{xx}, and in the y direction, T_{yy}, are measured, it will be found that they are, in general, not equal. This is not observed in Newtonian and generalized Newtonian fluids. The difference in T_{xx} and T_{yy} is referred to as the *first normal stress difference* N_1. Similarly, it will be found that T_{yy} and T_{zz} are unequal, and the difference $T_{yy} - T_{zz}$ is known as the *second normal stress difference*. The first and second normal stress difference are related to the shear rate by

$$\begin{aligned} T_{xx} - T_{yy} &= \psi_1(\dot{\gamma})\,\dot{\gamma}^2 \\ T_{yy} - T_{zz} &= \psi_2(\dot{\gamma})\,\dot{\gamma}^2 \end{aligned}$$

The functions ψ_1 and ψ_2 are known as the *first* and *second normal stress coefficients*, respectively. They are a function of the shear rate $\dot{\gamma}$: η, ψ_1 and ψ_2 are known as the viscometric functions of the fluid which together describe the response of the viscoelastic fluid under steady shear. ψ_1 is measured in special viscometers, which measure the normal stress, as well as the shear stress, exerted by the test fluid when it is being sheared [8]. Figure 36 is a plot of the first normal stress difference exhibited by a polyacrylamide solution. For comparison, the shear stress is also included in Figure 36. It can be seen that the normal stress is very much larger than the shear stress which means that this viscoelastic fluid can be expected to behave quite differently from a purely viscous shear-thinning

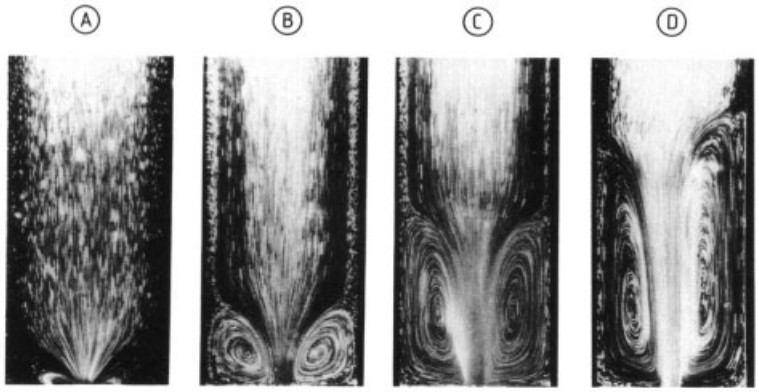

Figure 35. Comparison of the flow patterns exhibited by a Newtonian fluid and by a viscoelastic fluid
A) Newtonian fluid; B) – D) Viscoelastic fluid at increasing flow rates

fluid with similar shear stress versus shear rate behavior. The second normal stress difference is usually much smaller than the first normal stress difference and is very difficult to measure accurately. It is not normally reported in the literature.

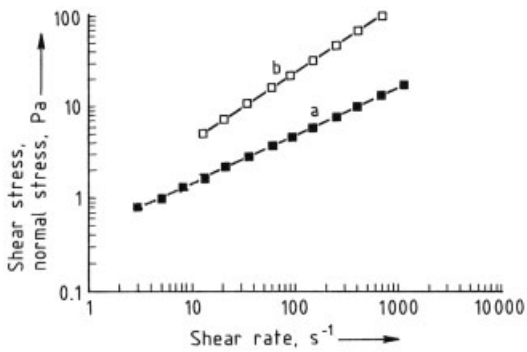

Figure 36. Normal stress versus shear rate for a viscoelastic fluid
a) Shear stress τ; b) First normal stress difference N_1

In laboratory investigations of viscoelasticity, it is very useful to be able to isolate the effects of shear thinning and fluid inertia from that of fluid elasticity so as to concentrate on the latter. For this purpose, a class of special test fluids have been developed. They are commonly referred to as *Boger fluids* or ideal elastic fluids. Typically, they are prepared by dissolving a high molecular mass solute in a highly viscous Newtonian solvent. The resulting solutions are characterized by their near constant viscosity, over a wide range of shear rates, and they exhibit all the elastic properties normally associated with viscoelastic fluids. Boger fluids are used extensively in the comparison of experimentally observed elastic response of fluids with that predicted by viscoelastic constitutive equations. They play an important role in the development and selection of such constitutive equations.

4.3.2. Behavior of Viscoelastic Fluids in Oscillatory Shear Flow

In steady shear motion the viscoelastic fluid is subjected to a constant rate of strain which does not reveal the dependence of the current stress on the strain history. Viscoelastic behavior under a time-dependent rate of strain is frequently studied experimentally by putting the fluid through an oscillatory shear flow, such as that generated when the fluid is sheared between two parallel plates: one of the plates is held stationary and the other oscillates sinusoidally with a frequency ω. The resulting shear rate and shear stress both have the same frequency ω, but there is a phase difference between them. If the oscillatory shear strain is represented by

$$\gamma = \gamma_0 \sin(\omega t) \tag{137}$$

where γ_0 is the amplitude of the sinusoidal shear strain, then the shear rate experienced by the fluid is given by

$$\dot{\gamma} = \gamma_0 \omega \cos(\omega t) = \dot{\gamma}_0 \cos(\omega t) \tag{138}$$

$\dot{\gamma}_0 (= \omega\, \gamma_0)$ is used to denote the amplitude of the sinusoidal shear rate.

It is customary to express the resulting shear stress in the form

$$T_{xy}(\omega) = \gamma_0 \left[G'(\omega)\sin\omega t + G''(\omega)\cos\omega t\right] \quad (139)$$

The terms associated with $G'(\omega)$ and $G''(\omega)$ are the in-phase and the out-phase component of the shear stress. $G'(\omega)$ is usually referred to as the *storage modulus* of the fluid, it is related to the storage of energy. $G''(\omega)$ is the *loss modulus* of the fluid and is connected with energy dissipation. As indicated above, $G'(\omega)$ and $G''(\omega)$ are functions of ω, and for small amplitude oscillatory shear, they are independent of γ_0. For purely viscous fluids $G'(\omega)$ is zero.

$G'(\omega)$ and $G''(\omega)$ are material properties of a viscoelastic fluid. Two other material functions, related to $G'(\omega)$ and $G''(\omega)$ are also in common use. These are the complex viscosity of the viscoelastic fluid, defined by

$$\eta'(\omega) = G''(\omega)/\omega \quad (140a)$$

and

$$\eta''(\omega) = G'(\omega)/\omega \quad (140b)$$

The function $\eta'(\omega)$ is called the *dynamic viscosity* of the fluid; $\eta''(\omega)$ does not seem to have a generally accepted name. For a Newtonian fluid $\eta'(\omega)$ is equal to the viscosity of the fluid and $\eta''(\omega)$ is zero. It is evident that a large number of material functions are needed just to describe the behavior of viscoelastic fluids in shear. At low shear rates and low angular frequencies, the material functions for steady-shear and oscillatory shear are related to one another. The dynamic viscosity at very low frequencies is related to the steady shear viscosity at very low shear rates by

$$\lim_{\omega\to 0} \eta'(\omega) = \lim_{\dot\gamma\to 0} \eta(\dot\gamma) \quad (141)$$

Furthermore, the storage modulus at very low frequencies is related to the first normal stress coefficient $\psi_1(\dot\gamma)$ at very low shear rates by

$$\lim_{\omega\to 0}\left(\frac{2G'(\omega)}{\omega^2}\right) = \lim_{\dot\gamma\to 0}\left(\frac{N_1(\dot\gamma)}{\dot\gamma^2}\right) = \lim_{\dot\gamma\to 0}\psi_1(\dot\gamma) \quad (142)$$

These limiting relationships are very useful for checking the consistency of experimental data on viscoelastic fluids. A number of other empirical relationships have been developed to relate the steady and oscillatory data. An example of these is the Cox – Merz rule which relates $\eta(\dot\gamma)$ to $\eta'(\omega)$ and $\eta''(\omega)$:

$$\eta(\dot\gamma) = \left[\eta'(\omega)^2 + \eta''(\omega)\right]^{1/2}\Big|_{\omega=\dot\gamma} \quad (143)$$

The Cox – Merz rule is often used to extend the steady shear viscosity data to large shear rates where only oscillatory measurements have been carried out. Further details and the limitations of this and other similar relationships are given in [4].

$G'(\omega)$ and $G''(\omega)$ or equivalently $\eta'(\omega)$ and $\eta''(\omega)$ are measured with special viscometers in which one of the shearing surfaces can be driven sinusoidally [8]. A typical set of oscillatory data is shown in Figure 37. These data were extracted from a comprehensive study carried out on polyethylene [25]. Oscillatory measurements are now routinely carried out by polymer manufacturers as a means of characterizing their products.

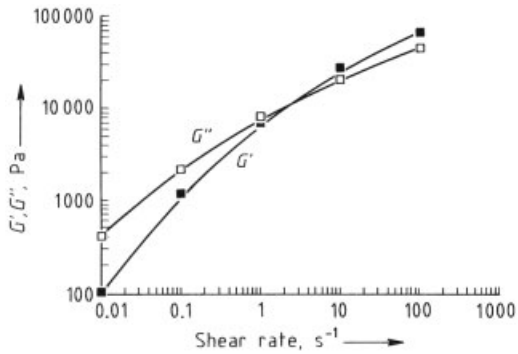

Figure 37. Storage and loss modulus of a low density polyethylene [25]
■ Storage modulus G'; □ Loss modulus G''

4.3.3. Examples of Constitutive Equations for Viscoelastic Fluids

Numerous constitutive equations have been proposed to describe the diverse relationship, observed in viscoelastic fluids, between the stresses and the various kinematic quantities used to measure the strain and the rate of strain experienced by the fluid. One of the earliest viscoelastic constitutive equations is the linear *Maxwell equation*. In shear flow, the stress T_{xy} and the shear rate $\dot\gamma$ are related by a constitutive equation of the form

$$\lambda \frac{\mathrm{d}T_{xy}}{\mathrm{d}t} + T_{xy} = \eta_0 \dot{\gamma} \tag{144}$$

λ and η_0 are the two material parameters of the linear Maxwell equation where λ is the time constant of the fluid and η_0 is its viscosity. Under steady shear, the stress is not a function of time, hence the first term on the left-hand side vanishes and the linear Maxwell equation reduces to the Newtonian constitutive equation where the stress is related to the shear rate via the viscosity η_0. However, the introduction of the time constant λ means that, in general, the stress is no longer a unique function of the instantaneous shear rate. For example, during the start up of the shear flow or in oscillatory shear test, the stress is out of phase with the instantaneous shear rate.

In particular, in oscillatory shear test the *oscillation* of the top plate is given by $A \sin \omega t$. The *velocity* of the top plate is

$$u_0 = A\omega \cos \omega t$$

For most viscoelastic fluids of practical interest, experimental conditions are such that the inertia terms in the Cauchy equations of motion, Equation (24), are small and can be ignored. As a result, the velocity variation across the gap separating the plates is linear and the velocity profile is given by

$$u\,(y,t) = \frac{A\omega y}{H} \cos\,(\omega t) \tag{145}$$

where H is the gap between plates. The shear rate is given by

$$\dot{\gamma}\,(t) = \frac{A\omega}{H} \cos\,(\omega t) = \gamma_0 \omega \cos\,(\omega t) \tag{146}$$

where $\gamma_0 = A/H$ is used to denote the amplitude of the shear strain. According to the linear Maxwell constitutive equation, the stress exerted by the fluid on the oscillating top plate is given by

$$\lambda \frac{\mathrm{d}T_{xy}}{\mathrm{d}t} + T_{xy} = \gamma_0 \omega \cos\,(\omega t) \tag{147}$$

This equation can be solved to give

$$T_{xy} = \gamma_0 \left(\frac{\eta_0 \omega \cos \omega t}{1+\lambda^2 \omega^2} + \frac{\eta_0 \lambda \omega^2}{1+\lambda^2 \omega^2} \sin \omega t \right) + \text{transient terms} \tag{148}$$

The transient terms decay away exponentially. As oscillatory measurements are taken at large t, the transient terms can be ignored. The resulting sinusoidally fluctuating stress can be written in the form

$$T_{yx} = \gamma_0 \left(\frac{\eta_0 \omega}{1+\lambda^2 \omega^2} \cos \omega t + \frac{\eta_0 \lambda \omega^2}{1+\lambda^2 \omega^2} \sin \omega t \right) \tag{149}$$

From the definition of $G'\,(\omega)$ and $G''\,(\omega)$ and of $\eta'\,(\omega)$ and $\eta''\,(\omega)$ it can then be seen that for a linear Maxwell fluid

$$G'\,(\omega) = \frac{\eta_0 \lambda \omega^2}{1+\lambda^2 \omega^2} \quad G''\,(\omega) = \frac{\eta_0 \omega}{1+\lambda^2 \omega^2} \tag{150a}$$

$$\eta'\,(\omega) = \frac{\eta_0}{1+\lambda^2 \omega^2} \quad \eta''\,(\omega) = \frac{\eta_0 \lambda \omega}{1+\lambda^2 \omega^2} \tag{150b}$$

By choosing the appropriate value for the viscosity η_0 and time constant λ, these functions can provide an adequate description of the observed behavior of real viscoelastic fluids. The linear Maxwell equation does not satisfy all the requirements of modern continuum mechanics for a theoretically sound constitutive equation. Numerous modifications of the Maxwell equations can be found in the literature. Some of the modifications were made so as to satisfy the requirements of continuum mechanics, others resulted from improved understanding of the molecular structure and dynamics of the polymer constituents of viscoelastic fluids [4], [26].

A constitutive equation that has been very popular is the *Oldroyd-B equation*. It relates the deviatoric stress tensor $\boldsymbol{\tau}$ to the strain tensor $\boldsymbol{\gamma}$ by an equation of the form

$$\lambda_1 \boldsymbol{\tau}_{(1)} + \boldsymbol{\tau} = \eta_0 \left(\boldsymbol{\gamma}_{(1)} + \lambda_2 \boldsymbol{\gamma}_{(2)} \right) \tag{151}$$

where λ_1 and λ_2 are the relaxation and retardation time of the Oldroyd-B fluid, respectively and η_0 is, as before, the viscosity of the fluid. Subscripts (1) and (2) are used to denote the first and second convected derivatives of the stress and the strain tensor. The definition of the convected and other derivatives can be found in [4]. The introduction of these derivatives is required by the theory of constitutive equations in continuum mechanics. The derivative $\boldsymbol{\gamma}_{(1)}$ is equal to the rate of strain tensor $\boldsymbol{e}$. The Oldroyd-B equation satisfies all the requirements of continuum mechanics. In terms of the three material parameters, λ_1, λ_2, and η_0 the viscometric functions (i.e., the steady shear properties) and the oscillatory behavior of the Oldroyd-B equation can be shown to be

$$\eta=\eta_0 \quad \psi_1= 2\eta_0\,(\lambda_1 - \lambda_2) \quad \psi_2= 0 \tag{152a}$$

$$G'(\omega)=\frac{\eta_0\,(\lambda_1 - \lambda_2)\,\omega^2}{1+(\lambda_1\omega)^2}$$

$$G''(\omega)=\frac{\eta_0\omega\,\left(1+\lambda_1\lambda_2\omega^2\right)}{1+(\lambda_1\omega)^2} \tag{152b}$$

$$\eta'(\omega)=\frac{\eta_0\,\left(1+\lambda_1\lambda_2\omega^2\right)}{1+(\lambda_1\omega)^2}$$

$$\eta''(\omega)=\frac{\eta_0\,(\lambda_1 - \lambda_2)\,\omega}{1+(\lambda_1\omega)^2} \tag{152c}$$

As with the linear Maxwell equation, the parameters of the Oldroyd-B equations are chosen so that these functions give an approximate description of the measured viscometric and oscillatory properties of real fluids. It is interesting to note that both linear Maxwell and the Oldroyd-B equations give a constant viscosity. This means that these equations are particularly suitable for describing the rheological behavior of viscoelastic fluids that are not shear-thinning. In Figure 38 the measured $G'(\omega)$ and $\eta'(\omega)$ for a silicone oil are plotted against ω. This fluid can be regarded as a non-shear-thinning viscoelastic fluid (Boger fluid). The expression for these two quantities according to the Maxwell and the Oldroyd-B equations are also plotted on the figure. Standard procedures have been applied to extract the fluid parameters of these two constitutive equations [27]. The Oldroyd-B equation provides a better description of the rheological behavior of this viscoelastic fluid than the simpler Maxwell equation. For comparison, the steady shear data for this silicone oil are also included in Figure 38. The experimental data confirm the limiting relationship between N_1 and $2\,G'$ and between η and η' at low $\dot{\gamma}$ and ω.

The linear Maxwell and the Oldroyd-B equations are just two examples of the large number of viscoelastic constitutive equations in the literature. The construction of constitutive equations is guided by the principles of continuum mechanics and by the understanding of the dynamical properties of the macromolecules and the solvent that make up the viscoelastic fluids. For more details on these and related topics see [28], [29].

One of the recent developments in viscoelastic fluid mechanics is the use of Boger fluids to gain a physical understanding of constitutive equations. With these ideal test fluids good agreement between experimental observations and theoretical predictions could be observed for the first time. It has also been possible to relate the parameters in the constitutive equations to the physical properties of the constituents of Boger fluids.

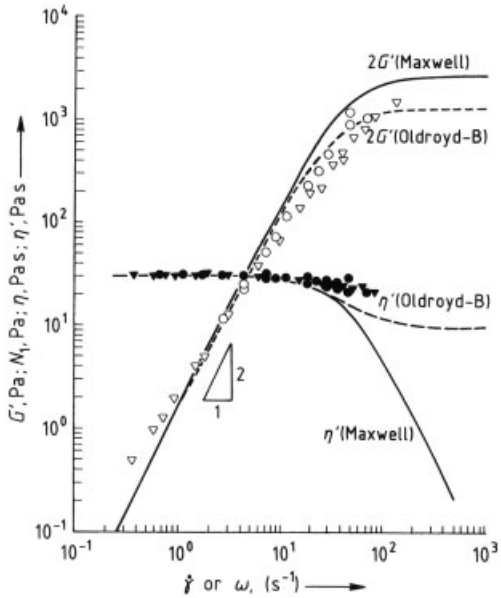

Figure 38. Comparison of the observed normal stress and storage modulus of a silicone oil against that predicted by the Maxwell and the Oldroyd-B constitutive equations, fluid: silicone oil
○ N_1, Pa; ▽ $2\,G'$, Pa; ● η, Pa s; ▼ η', Pa s

4.3.4. Extensional Behavior of Viscoelastic Fluids

In many of the flow fields encountered in industrial processes the fluid undergoes a stretching deformation. A specific example is the deformation suffered by a fluid element when it is forced to flow through a channel or tube of gradually or abruptly changing cross-sectional area. This kind of deformation is usually referred to as extensional deformation. In extensional deformation the macromolecules in the viscoelastic fluid are being stretched and may generate large tensile stresses as a result of this. The behavior of a viscoelastic fluid in extensional deformation can be quite different from its behavior in shear deformation.

An idealized flow field that can be used to investigate the extensional behavior of a viscoelas-

tic fluid is shown in Figure 39. The tensile stress T_{zz} applied to stretch the specimen of test fluid, in the form of a circular cylinder, is measured. The applied stress T_{zz} is varied so that the axial velocity gradient is uniform within the fluid sample and remains constant in time, i.e.,

$$e_{zz}=\frac{\mathrm{d}u_z}{\mathrm{d}z}=\dot{\varepsilon}_0=\frac{1}{L}\frac{\mathrm{d}L}{\mathrm{d}t} \tag{153}$$

where $\dot{\varepsilon}_0$ is the constant extensional rate, which can be identified with the ratio of the instantaneous rate of increase in length of the fluid sample $\mathrm{d}L(t)/\mathrm{d}t$ to its instantaneous length $L(t)$. It can be assumed that the azimuthal velocity component in the test specimen is zero. The continuity equation for an incompressible fluid then requires the velocity components in the axial and radial directions to be given by

$$\begin{aligned} u_z &= \dot{\varepsilon}_0 z \\ u_r &= -\frac{1}{2}\dot{\varepsilon}_0 r \\ u_\theta &= 0 \end{aligned} \tag{154}$$

The rate of strain tensor, in cylindrical coordinates, is given by

$$\boldsymbol{e}=\begin{vmatrix} \dot{\varepsilon}_0 & 0 & 0 \\ 0 & -\frac{1}{2}\dot{\varepsilon}_0 & 0 \\ 0 & 0 & -\frac{1}{2}\dot{\varepsilon}_0 \end{vmatrix} \tag{155}$$

A characteristic of extensional flow is the vanishing off-diagonal elements in the rate of strain tensor. This idealized constant-rate fluid stretching experiment, known as the steady axisymmetric extensional flow, is the extensional equivalent of steady shear flow. It is a difficult experiment to carry out and consequently reliable extensional properties of fluids are difficult to obtain. One of the difficulties becomes apparent when Equation (153) is integrated to give

$$L=L_0\mathrm{e}^{\dot{\varepsilon}_0 t} \tag{156}$$

where L_0 is the initial length of the cylindrical fluid test specimen. The exponential increase in length of the test specimen makes it very difficult to sustain measurement for a period long enough for the effects of initial conditions to be ignored. Specialized rheometers have been designed to overcome this and other related problems [8]. However, good elongational flow measurements are rare.

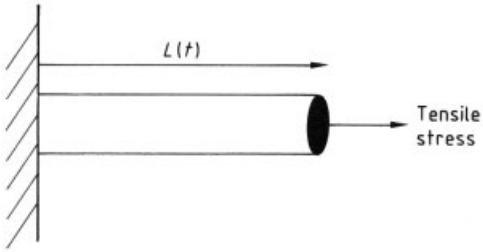

Figure 39. Schematic diagram of extensional flow

An important extensional property of a viscoelastic fluid is its *extensional viscosity* η_{el}. In terms of the idealized extensional measurement described above, η_{el} is defined by

$$\eta_{\mathrm{el}}=\frac{T_{zz}-T_{rr}}{\dot{\varepsilon}_0} \tag{157}$$

The difference between the normal stress in the axial direction T_{zz} and that in the radial direction T_{rr} gives the deviatoric tensile stress arising from the extensional deformation. The definition of extensional viscosity is analogous to the definition of shear viscosity. Like its shear counterpart, the extensional viscosity is in general a function of the extensional rate $\dot{\varepsilon}_0$.

The extensional viscosity predicted by different constitutive equations can be found in the literature [30]. The extensional viscosity of a Newtonian fluid with constant viscosity η_0 is relatively simple to obtain. The total normal stress in the axial and the radial directions are given by

$$\begin{aligned} T_{zz} &= -p+2\eta\frac{\mathrm{d}u_z}{\mathrm{d}z} = -p+2\eta_0\dot{\varepsilon}_0 \\ T_{rr} &= -p+2\eta\frac{\mathrm{d}u_r}{\mathrm{d}r} = -p-\eta_0\dot{\varepsilon}_0 \end{aligned}$$

Hence

$$\eta_{\mathrm{el}}=\frac{T_{zz}-T_{rr}}{\dot{\varepsilon}_0}=3\eta_0 \tag{158}$$

According to this result, the extensional viscosity of a Newtonian fluid is three times its shear viscosity. This result has been verified experimentally by Trouton in 1906 and extensional viscosity is often referred to as *Trouton viscosity* [31]. The extensional viscosity of viscoelastic fluids is, in general, much larger than three times it shear viscosity.

The extensional viscosity obtained for the Oldroyd-B equation can be shown to be

$$\eta_{\mathrm{el}}(\dot{\varepsilon}_0)=3\eta_0\left[\frac{(1-\lambda_2\dot{\varepsilon}_0)(1+2\lambda_1\dot{\varepsilon}_0)}{(1+\lambda_1\dot{\varepsilon}_0)(1-2\lambda_1\dot{\varepsilon}_0)}\right] \tag{159}$$

At low extensional rate, the Newtonian result of $3\,\eta_0$ is again obtained. As the extensional rate is increased, the extensional viscosity increases rapidly. This large extensional viscosity is in agreement with the observed extensional behavior of viscoelastic fluids. However, according to the Oldroyd-B equation, the extensional viscosity grows without bound as the extensional rate approaches $\lambda_1/2$. This singularity in the relationship between extensional rate and extensional viscosity is physically unrealistic and is an indication that the Oldroyd-B equation must be modified. Comparison of the measured extensional viscosity with that predicted by a constitutive equation provides an additional test of the validity of the constitutive equation.

It is clear that fluid elasticity has greatly increased the complexity of fluid motion. There are a large number of flow phenomena which are only observed in viscoelastic fluids. Many of these have practical implications for industrial processes. For further details on the mechanics of viscoelastic fluids see [4], [19], [26].

4.3.5. Accelerating and Decelerating Flows of Viscoelastic Fluids

Fluid elasticity does not affect the energy requirements for fully-developed/laminar tube flow. However, in accelerating and decelerating flows, such as in the entrance and exit of a tube, the influence of fluid elasticity becomes quite pronounced. Entrance and exit effects are conveniently discussed with reference to the capillary rheometer, an important instrument for the measurement of fundamental flow properties of fluids.

One essential feature of the capillary viscometer is that the wall shear stress can be directly determined from the measured fully-developed flow pressure drop (see Eq. 87). However, for a laboratory-scale capillary viscometer it is more practical to measure the overall pressure drop from the upstream fluid reservoir to the exit of the tube, rather than the pressure drop in the tube itself. Schematics of a pressure-driven and a ram-driven capillary rheometer are shown in Figure 40. In the *pressure-driven* instrument, the independent variable is the shear stress, whereas in the *ram-driven instrument* the independent variable is the shear rate. Shear rates of $> 10^3\ \mathrm{s}^{-1}$ are easily obtained in the capillary rheometer. This is one of its main advantages over conventional rotation instruments.

Assuming that the flow is laminar and that the fluid is a time-independent inelastic or viscoelastic fluid, corrections to the measured pressure drop ($p_{\mathrm{gas}} - p_{\mathrm{atm}}$) for pressure-driven rheometers and ($p_{\mathrm{app}} - p_{\mathrm{atm}}$) for ram-driven rheometers may have to be made due to the following effects:

1) Head of fluid above the tube exit
2) Kinetic energy effects
3) Tube entrance and exit losses
4) Weight of filament after exit (this will be ignored here)

$$p_{\mathrm{gas}} - p_{\mathrm{atm}} = \Delta p_{\mathrm{fd}_1} + \Delta p_{\mathrm{en}} + \Delta p_{\mathrm{fd}_2} + \Delta p_{\mathrm{ex}} + \Delta KE + \Delta PE \tag{160}$$

Equation (160) is a mechanical energy balance written between surface 0 and 2 as shown in Figure 40: Δp_{fd_1} and Δp_{fd_2} are the fully-developed flow pressure drops in the reservoir and tube, respectively; Δp_{en} is the entry loss over and above the fully-developed flow loss for the flow into the tube; and Δp_{ex} is the exit loss over and above the fully-developed flow loss for flow out of the tube to the atmosphere; ΔKE and ΔPE are the kinetic and potential energy losses, respectively. Kinetic and potential energy effects can normally be neglected for polymer melts and for concentrated solutions and suspensions. In addition, the reservoir diameter is usually much greater than that of the downstream tube, so that the fully-developed pressure drop in the upstream tube or reservoir can also be neglected. Thus, Equation (160) becomes

$$p_{\mathrm{gas}} - p_{\mathrm{atm}} = \Delta p_{\mathrm{en}} + \Delta p_{\mathrm{fd}_2} + \Delta p_{\mathrm{ex}} \tag{161}$$

where

$$\Delta p_{\mathrm{fd}_2} = \frac{2\tau_{\mathrm{w}} L}{R}$$

If high L/R capillary tubes are used for the pressure drop – flow rate measurements ($L/R \geq 200$), fundamental shear stress – shear rate data can be determined directly with the capillary rheometer because

$$\frac{2\tau_{\mathrm{w}} L}{R} >> \Delta p_{\mathrm{en}} + \Delta p_{\mathrm{ex}} \tag{162}$$

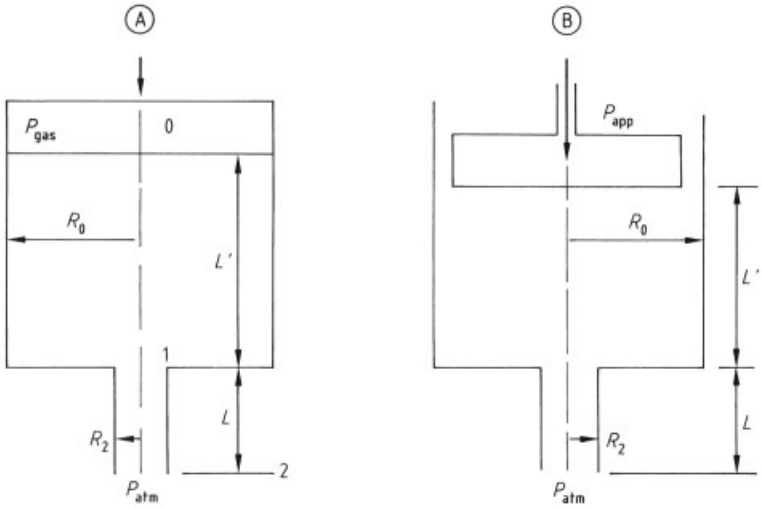

Figure 40. Schematic diagram of (A) pressure and (B) ram driven capillary rheometers

Therefore

$$\tau_{\mathrm{w}} = \frac{R\,(p_{\mathrm{gas}} - p_{\mathrm{atm}})}{2L} \tag{163}$$

i.e., the wall shear stress can be directly determined from the measured pressure drop, and the wall shear rate is specified by the Rabinowitsch – Mooney equation

$$\dot{\gamma}_{\mathrm{w}} = \left(\frac{3n'+1}{4n'}\right)\left(\frac{8V}{D}\right) \tag{164}$$

where

$$n' = \frac{\mathrm{d}\ln\tau_{\mathrm{w}}}{\mathrm{d}\ln\frac{8V}{D}} \tag{165}$$

However, for many materials in tubes with a high *L/R* ratio, the pressures required to obtain shear rates of interest are prohibitive, and methods have been derived to correct for entry and exit effects when low *L/R* capillary tubes are used. The method most commonly used is that first suggested by Bagley in 1957 [32]:

$$\Delta p_{\mathrm{en}} + \Delta p_{\mathrm{ex}} = \frac{2\tau_{\mathrm{w}} L_{\mathrm{e}}}{R} = 2\tau_{\mathrm{w}} e \tag{166}$$

where e is the dimensionless extra length of tube which defines the exit and entry losses in excess of the fully-developed flow losses.

$$p_{\mathrm{gas}} - p_{\mathrm{atm}} = 2\tau_{\mathrm{w}}\,(L/R + e) \tag{167}$$

and

$$\tau_{\mathrm{w}} = \frac{p_{\mathrm{gas}} - p_{\mathrm{atm}}}{2\,(L/R + e)} \tag{168}$$

The end correction e is determined by first plotting $(p_{\mathrm{gas}} - p_{\mathrm{atm}})/(2\,L/R)$ versus 8 *V/D* on log – log coordinates. Different lines or curves such as those illustrated in Figure 41 will be obtained for different *L/R* tubes. From such a graph $p_{\mathrm{gas}} - p_{\mathrm{atm}}$ can be determined as a function of *L/R* for various values of 8 *V/D*. A linear plot of $p_{\mathrm{gas}} - p_{\mathrm{atm}}$ versus *L/R* is then made for various values of 8 *V/D*. The end correction e is the intercept on the abscissa of this plot (see Fig. 42). The end correction is then a known function of 8 *V/D* and hence the corrected value of τ_{w} can be computed as a function of 8 *V/D* from Equation (168) whereas the true shear rate can be computed with the aid of Equation (164) and the slope of a log – log plot of τ_{w} versus 8 *V/D* (Eq. 165).

The capillary rheometer is an important instrument for the measurement of the shear stress as a function of shear rate, particularly for molten polymers where measurements can be made at processing shear rates. It is also an instrument for determining quantitative measurement of fluid elasticity. For instance, the end correction is strongly influenced by the elasticity of the fluid. Figure 43 shows end correction results for molten polymers while Table 2 lists the end

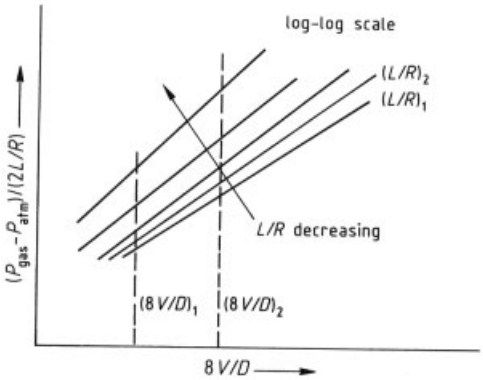

Figure 41. First plot of capillary rheometer data for determination of the end effect

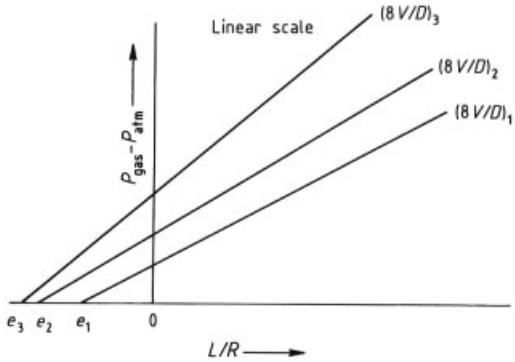

Figure 42. Determination of the end correction

correction for inelastic power-law fluids as determined from the numerical solution of the tube entry and exit flow problem [34].

Table 2. The end correction for inelastic power-law fluids

Power-law index, n	$\Delta p_{en}/2\tau_w$	$\Delta p_{ex}/2\tau_w$	e
1.0	0.588	0.246	0.834
0.5	1.34	0.26	1.60
0.3	1.76	0.28	2.04
0.167	2.33	0.59	2.92

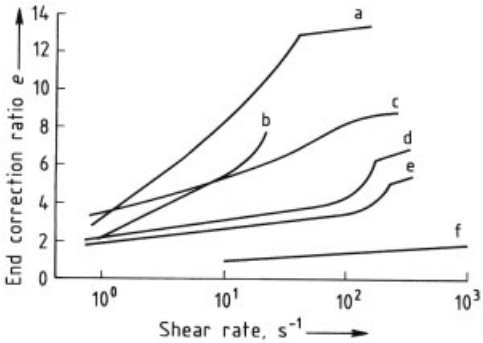

Figure 43. Typical values of the end correction as a function of shear rate [33]
a) Polystyrene; b) Poly(methyl methacrylate); C) Low-density polyethylene MFI 2.0; d) High- density polyethylene MFI 0.25; e) High- density polyethylene; f) Polyacetal

In the absence of fluid elasticity, the end correction increases as a result of the shear-thinning characteristics of the fluid, but not to the extent of the values observed for many commercial polymers which in general are shear-thinning and also highly elastic. Even higher end correction values than those illustrated for molten polymers in Figure 43 have been observed for concentrated polymer solutions. The end correction is used to differentiate between different polymers and in fact to distinguish between different grades of the same polymer (see curves d and e in Fig. 43).

Significantly different behavior is indeed observed in inlet and exit flows of inelastic and viscoelastic fluids. Figure 44 shows streamline photographs obtained for an inelastic Newtonian fluid (Fig. 44A) and for an elastic fluid which shows no shear thinning (Fig. 44B). Both fluids have identical viscosities. Flow is from left to right in the photographs and represents flow from the reservoir *into* the tube of a capillary rheometer. For the inelastic Newtonian fluid a small secondary flow is present in the corner of the reservoir. This cell remains essentially constant in size with increasing flow rate and ultimately disappears for Reynolds numbers $>$ ca. 0.1 when the fluid inertia starts to become important. For the viscoelastic fluid of the same viscosity (2000 Pa s) the secondary flow is much larger. The size of the secondary flow grows with increasing flow rate for Reynolds numbers < 0.1 and continues to grow until the flow becomes unstable. For a molten polymer being extruded through a die, the flow instability results in a distorted extrudate. The flow phenomena, called melt fracture, represents an upper limit on the rate at which a molten polymer can be

extruded. Similar flow instabilities are not observed in tubular inlet flows for inelastic fluids.

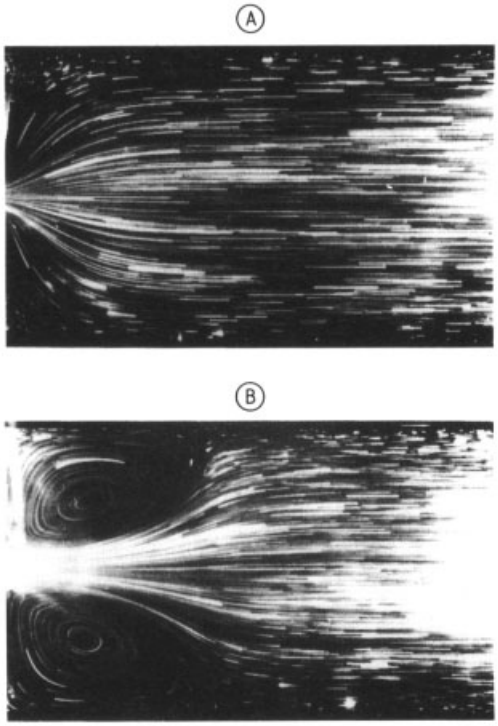

Figure 44. Comparison of entry flow patterns for a Newtonian and a non-shear thinning elastic fluid with the same viscosity
Reservoir to tube diameter ratio is 7.67
A) Newtonian; B) Non-shear thinning elastic

For a viscoelastic fluid the *exit flow* from a capillary tube also differs significantly from that of an inelastic fluid. For fully-developed flow of a viscoelastic fluid in a tube, a tension along the streamlines associated with the deviatoric normal stresses is present. When the fluid passes through the exit of the tube to the atmosphere, it will relax the tension along the streamlines by contracting in the longitudinal direction. For an incompressible fluid, this results in a lateral expansion of the fluid. This relaxation phenomena results in extrudate swell (see Fig. 33), where the diameter of the extrudate D_e is significantly greater than the internal tube diameter. For large tube length to diameter ratios the extrudate swell can be estimated as follows [35]:

$$\frac{D_e}{D} = 0.1 + \left[1 + \frac{1}{2}\left(\frac{\tau_{11} - \tau_{22}}{2\tau_{12}}\right)_w^2\right]^{1/6}$$

where the subscript w indicates that the stresses are to be evaluated at the wall shear rate for fully-developed flow. Die swell ratios of 2 or more are not unusual in the processing of molten polymers. Since the die swell depends not only on the particular polymer but also on the operating conditions such as temperature and flow rate, the industrial problems related to extrudate swell are particularly complex and challenging.

Fluid elasticity does not effect the energy requirements for fully-developed *laminar* tube flow, that is

$$f_{\text{viscoelastic}} = f_{\text{inelastic}} \tag{169}$$

The equivalent conclusion, however, is not applicable for viscoelastic fluids in *turbulent flow*. Here considerable reduction in the friction factor below the expected inelastic value is observed. This *drag reduction phenomenon*, first observed by Toms in 1948, has received considerable attention in the literature because of its possible commercial significance. Parts per million of certain polymers dissolved in water can reduce the friction factor considerably; drag reduction by as much as 90 % has been observed. Figure 45 shows some friction factor – Reynolds number data for a very dilute aqueous polymer solution. The data agrees with the laminar flow prediction but deviates from the solvent line characterization for turbulent flow for Reynolds numbers $> 10^4$.

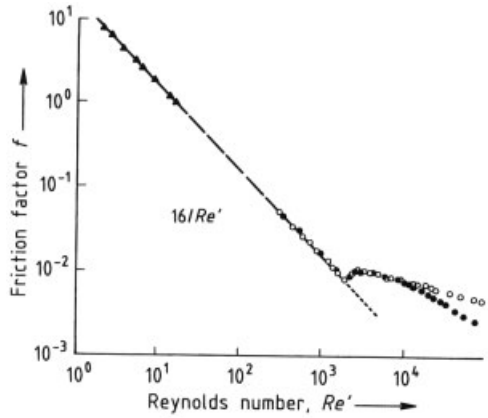

Figure 45. Friction factor – Reynolds number data for Boger fluids, distilled water, and a solution of 296 ppm by weight of polyethylene oxide in distilled water in an 8.46 mm-diameter pipe [36]. (Reproduced by permission of the American Institute of Chemical Engineers © 1975 AIChE)
▲ Boger fluid; ○ Distilled water; ● 296 ppm by weight polyethylene oxide

5. Numerical Methods in Fluid Mechanics

Many of the flows of practical importance have irregular geometry and their boundary conditions are often mathematically difficult to handle. Furthermore, the fluid properties may vary rapidly with flow and thermal conditions. These complications mean that the set of equations that describe the fluid motion are unlikely to have an analytic solution. Even in those rare cases where an analytic solution can be found, it may be in a form, that is not convenient for practical use, e.g., as a slowly converging infinite series. For such flows, it is necessary to obtain an approximate numerical solution to the governing equations. Numerical methods have always been used, but their development has been greatly accelerated due to the availability of modern digital computers that perform the repetitive calculations. Computational fluid mechanics is now a well established subject capable of producing highly accurate predictions of the fluid motion at great speed. The ever increasing computing power combined with the graphical capabilities of modern computers has established numerical solution of fluid mechanical problems as an extremely effective way of generating and presenting information about fluid flows. In this short section, the principles of two general numerical methods for solving the steady-state Navier–Stokes equations and the continuity equation in two dimensions will be described. Further details are given in [37–41].

5.1. Finite Difference Method

(→ Mathematics in Chemical Engineering, Chap. 7.3.)

In the finite difference method, the partial derivatives in the governing equations are approximated by finite differences. To do this, the flow field of interest is divided into a regular grid as shown in Figure 46.

At any grid point (i, j), the first derivative $\partial u_x/\partial x$ in the continuity equation can be approximated by

$$\left(\frac{\partial u_x}{\partial x}\right)_{i,j} = \frac{u_{x_{i+1,j}} - u_{x_{i-1,j}}}{2\Delta x} \tag{170}$$

Subscripts are used to indicate the grid point at which the variable is to be evaluated. The error incurred in this approximation is dependent on the size of the grid Δx. It approaches zero as Δx approaches zero. A similar approximation can be obtained for $\partial u_y/\partial y$. It is common practice to make Δy equal to Δx. The finite difference approximation of the incompressible continuity equation at the grid point (i, j) is

$$u_{x_{i+1,j}} + u_{y_{i,j+1}} - u_{x_{i-1,j}} - u_{y_{i,j-1}} = 0 \tag{171}$$

This is an algebraic equation for the unknowns $u_{x_{i+1,j}}$, $u_{x_{i-1,j}}$ etc. A similar equation can be written down for each of the grid points, $1 \le i \le N_i$ and $1 \le j \le N_j$, where N_i and N_j are the number of grid points in the x and y directions, respectively.

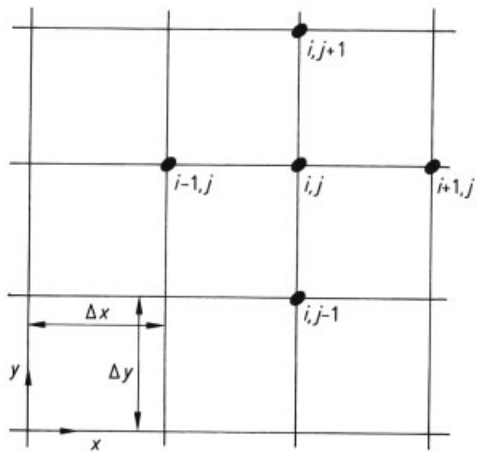

Figure 46. Finite difference grid

At the grid point (i, j), $\partial^2 u_x/\partial x^2$, a typical second derivative, can be approximated by the finite difference:

$$\left(\frac{\partial^2 u_x}{\partial x^2}\right)_{i,j} = \frac{u_{x_{i+1,j}} + u_{x_{i-1,j}} - 2u_{x_{i,j}}}{\Delta x^2} \tag{172}$$

Replacing all the partial derivatives by their finite difference approximations, the following finite difference equivalent of the x-component of the Navier–Stokes equation is obtained:

$$\varrho\left[u_{x_{i,j}}\left(\frac{u_{x_{i+1,j}}-u_{x_{i-1,j}}}{2\Delta x}\right)+\right.$$

$$\left.u_{y_{i,j}}\left(\frac{u_{x_{i,j+1}}-u_{x_{i,j-1}}}{2\Delta y}\right)\right]$$

$$=\frac{p_{i+1,j}-p_{i-1,j}}{2\Delta x}+\eta\left(\frac{u_{x_{i+1,j}}+u_{x_{i-1,j}}-2u_{x_{i,j}}}{\Delta x^2}\right)$$

$$+\eta\left(\frac{u_{x_{i,j+1}}+u_{x_{i,j-1}}-2u_{x_{i,j}}}{\Delta y^2}\right) \tag{173}$$

A similar finite difference equation can be written for the y- component of the Navier – Stokes equation. The finite differences have converted the differential equations describing the fluid flow to a set of simultaneous algebraic equations for the unknown velocity components and pressure at the grid points. After the incorporation of the boundary conditions, they can, in principle, be solved for the unknown variables. However, because of the complexity of the Navier – Stokes equations, the success of the finite difference method depends critically on how the finite difference equations are handled. They often have to be completely reorganized and rewritten in various forms that are numerically and computationally more manageable. For example, in this description of the finite difference method, the physical variables u_x, u_y and p appear explicitly. The finite difference method can be reformulated in terms of the stream function Ψ, or a combination of stream function and the vorticity ω. For two-dimensional flow ω is defined as the z component of curl of the velocity vector:

$$\omega=-\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x} \tag{174}$$

In two-dimensional flows, the stream function or stream function-vorticity formulations are, in fact, used in preference to the velocity – pressure formulation. Further details, particularly regarding the choice of numerical schemes for handling the resulting equations, can be found in the references listed.

In this introductory description of the finite difference method, the partial derivatives are approximated by central differences. Other finite difference approximation schemes can be applied. Irrespective of the finite difference scheme employed (because of the inertia terms on the left-hand side of the Navier – Stokes equations) the resulting algebraic equations are nonlinear in the unknown velocity components. These equations are not easy to solve directly. In one of the numerical procedures developed to deal with the nonlinear terms, the typical inertia term u_x $\partial u_x/\partial x$ is approximated by

$$u_x\frac{\partial u_x}{\partial x}\approx u_{x_{i,j}}^{(n-1)}\left(\frac{u_{x_{i+1,j}}^{(n)}-u_{x_{i-1,j}}^{(n)}}{2\Delta x}\right) \tag{175}$$

Here the superscript n is used to denote the values of the unknowns at the nth iteration. At the nth iteration the values of the $(n-1)$th iteration are already known. This way the nonlinear terms are approximated as linear ones and the finite difference equations become a set of linear algebraic equations for which many computationally efficient methods of solution are available. The calculation is terminated when the difference between two successive iterations is less than some preassigned small tolerance. To start the iterative procedure, it is necessary to supply the zeroth iteration of the unknown variables $u_{i,j}^{(0)}$, etc. In the absence of any information, these zeroth iteration values can be taken to be zeros. Convergence of this straightforward iterative scheme is not guaranteed and even if it converges, it may require a large number of iterations. Convergence, for example, becomes slower and slower as the Reynolds number of the flow is increased and may fail completely when the Reynolds number becomes too large. Many numerical procedures have been developed to improve the convergence performance. Iterative procedures have also been developed to deal with the nonlinearity introduced by variable fluid properties.

Figure 47 shows the streamlines of the flow of a Newtonian fluid past a cylinder at increasing Reynolds number obtained by finite difference computation using the stream function – vorticity formulation [42]. Numerical convergence at a Reynolds number as high as 100 was achieved by a specially developed iteration scheme. The effects of increasing Reynolds number show up clearly in the increasing size of the recirculating wake behind the cylinder. These numerical results are in excellent agreement with the available experimental data.

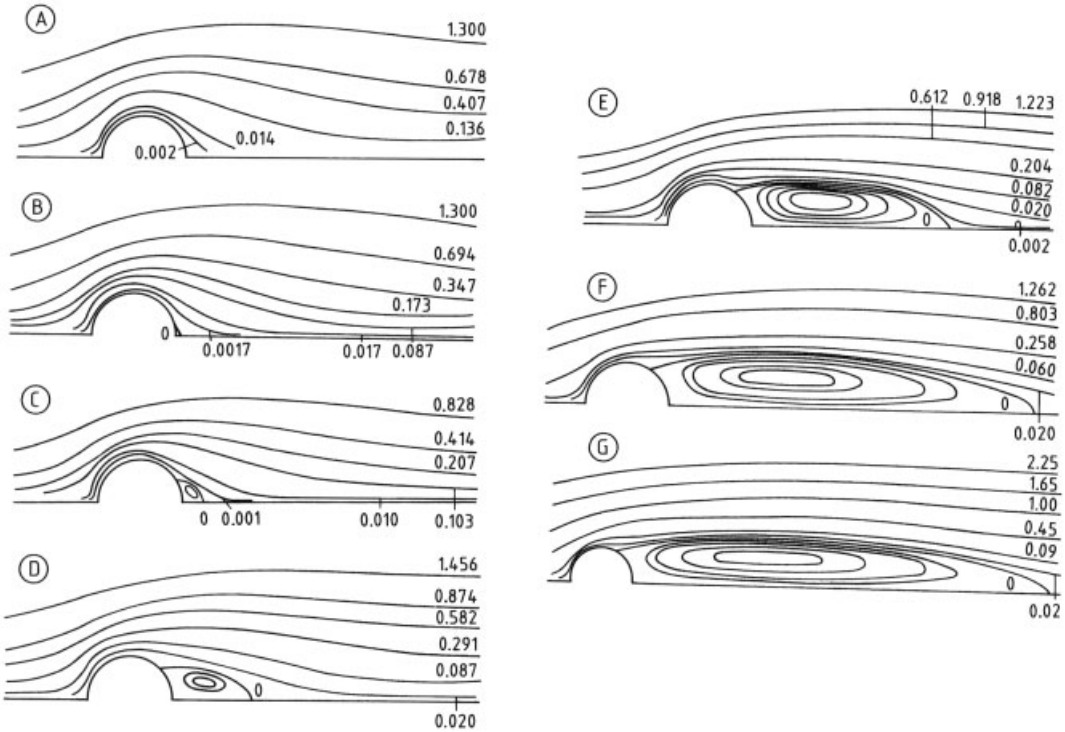

Figure 47. Streamlines around a cylinder at increasing Reynolds number obtained by finite difference computation [42] (reproduced with permission of Cambridge University Press)
A) $Re = 5$; B) $Re = 7$; C) $Re = 10$; D) $Re = 20$; E) $Re = 40$; F) $Re = 70$; G) $Re = 100$.

5.2. Finite Element Method

(→ Mathematics in Chemical Engineering, Chap. 7.5., → Mathematics in Chemical Engineering, Chap. 7.6.)

The finite element technique was originally developed by structural engineers for calculating the stresses and strains in structures of complex shapes. It has been generalized and developed as a numerical technique for solving differential equations, particularly partial differential equations. As with the finite difference method, the differential equations for the unknown variables are converted into a set of algebraic equations which can then be solved for the unknown variables at discrete points. The application of finite element computation to fluid mechanics started in the early 1960s and has since been greatly refined. A brief description of the finite element technique based on the Galerkin approach will be presented here. The mathematical principles and numerical details of the *Galerkin finite element* method, particularly the handling of boundary conditions, can be found in [37–41].

In finite element approximation, the flow field of interest is again subdivided by a grid into a large number of small connected domains of different shapes and sizes known as elements. (Following the general practice in finite element computation, the grid points, referred to as the nodal points, are identified by a single subscript *i* instead of the double subscripts *i, j* in finite difference.) The unknown variables, u_x, u_y , and p are approximated by series of the form

$$u_x\,(x,y) = \sum_{i=1}^{N} u_{xi}\Phi_i\,(x,y)$$

$$u_y\,(x,y) = \sum_{i=1}^{N} u_{yi}\Phi_i\,(x,y) \qquad (176)$$

$$p\,(x,y) = \sum_{i=1}^{N'} p_i\Psi_i\,(x,y)$$

where N (and N', see below) is the total number of nodal points in the grid; u_{xi}, u_{yi}, and p_i are unknown numerical coefficients to be determined; $\Phi_i(x, y)$, for $i = 1$ to N, and $\Psi_i(x, y)$ for $i = 1$ to N' are two sets of known functions defined for each of the nodal points. They are usually referred to as the interpolation functions. While one is free to choose the form of these functions, $\Phi_i(x, y)$ and $\Psi_i(x, y)$ are usually restricted to simple functions such as low-order polynomials. In many finite element computations $\Phi_i(x, y)$ are taken to be quadratic functions of x and y and $\Psi_i(x, y)$ to be linear functions of x and y. Both $\Phi_i(x, y)$ and $\Psi_i(x, y)$ take the value of unity at the nodal point i and decrease to zero at all the neighboring nodes surrounding node i. These functions are also defined to be identically zero beyond these neighboring nodes, i.e., Φ_i and Ψ_i are nonzero only in those elements which have node i as one of its nodal points. The quadratic interpolating function $\Phi_i(x, y)$ is also defined to be zero at the midpoints (the mid-side nodes) of each side of the elements that have node i as one of its nodes. The total number of nodes N for the quadratic interpolation functions includes the mid-side nodes. Thus N', the total number of nodes for the linear interpolation function, is $< N$. In order that the series representation approaches the true solution as N and N' become large, the interpolation functions have to satisfy a number of mathematical requirements. For example, all the interpolation functions have to be continuous across the common boundary of two adjacent elements.

Typical examples of the interpolating functions for quadrilateral and triangular elements are shown in Figure 48. In the finite element method, elements of different sizes and shapes can be mixed relatively easily. This greatly simplifies the discretization of irregularly shaped flow fields and is the major advantage of the finite element method over the finite difference method. When the series representations (Eq. 176) are substituted in the equations of motion and the continuity equation, the following are obtained:

$$\varrho\left[\sum_{i=1}^{N} u_{xi}\Phi_i\left(\sum_{i=1}^{N} u_{xi}\frac{\partial\Phi_i}{\partial x}\right)+\sum_{i=1}^{N} u_{yi}\Phi_i\left(\sum_{i=1}^{N} u_{xi}\frac{\partial\Phi_i}{\partial y}\right)\right]+\sum_{i=1}^{N'} p_i\frac{\partial\Psi_i}{\partial x}-\eta\left[\sum_{i=1}^{N} u_{xi}\frac{\partial^2\Phi_i}{\partial x^2}+\sum_{i=1}^{N} u_{xi}\frac{\partial^2\Phi_i}{\partial y^2}\right]=R_x(x,y) \tag{177a}$$

$$\varrho\left[\sum_{i=1}^{N} u_{xi}\Phi_i\left(\sum_{i=1}^{N} u_{yi}\frac{\partial\Phi_i}{\partial x}\right)+\sum_{i=1}^{N} u_{yi}\Phi_i\left(\sum_{i=1}^{N} u_{yi}\frac{\partial\Phi_i}{\partial y}\right)\right]+\sum_{i=1}^{N'} p_i\frac{\partial\Psi_i}{\partial y}-\eta\left[\sum_{i=1}^{N} u_{yi}\frac{\partial^2\Phi_i}{\partial x^2}+\sum_{i=1}^{N} u_{yi}\frac{\partial^2\Phi_i}{\partial y^2}\right]=R_y(x,y) \tag{177b}$$

$$\sum_{i=1}^{N} u_{xi}\frac{\partial\Phi_i}{\partial x}+\sum_{i=1}^{N} u_{yi}\frac{\partial\Phi_i}{\partial y}=R_c(x,y) \tag{177c}$$

In general the series representations do not satisfy the governing equations and leave behind residuals $R_x(x, y)$, $R_y(x, y)$, and $R_c(x, y)$. These residuals are functions of the unknown coefficients u_{xi}, u_{yi}, p_i, as well as x and y. An approximate solution is obtained by finding the set of the unknown coefficients that minimize, in some sense, these residuals. In the case of Galerkin finite element method, this is done by requiring the $R_x(x, y)$ and $R_y(x, y)$ to be orthogonal to all the $\Phi_i(x, y)$ and $R_c(x, y)$ to be orthogonal to all the $\Psi_i(x, y)$ over the region of interest, i.e.,

$$\int_A R_x(x,y)\,\Phi_k(x,y)\,\mathrm{d}A=0 \quad k=1\,\mathrm{to}\,N \tag{178a}$$

$$\int_A R_y(x,y)\,\Phi_k(x,y)\,\mathrm{d}A=0 \quad k=1\,\mathrm{to}\,N \tag{178b}$$

$$\int_A R_c(x,y)\,\Psi_k(x,y)\,\mathrm{d}A=0 \quad k=1\,\mathrm{to}\,N' \tag{178c}$$

A is the area occupied by the flow field. After performing the integration, these become a set of algebraic equations which can be solved for the unknown coefficients u_{xi}, u_{yi}, p_i. It will be found that the number of independent algebraic equations, after taking into consideration the boundary conditions, is exactly equal to the number of unknown numerical coefficients in the series representation. From the definition

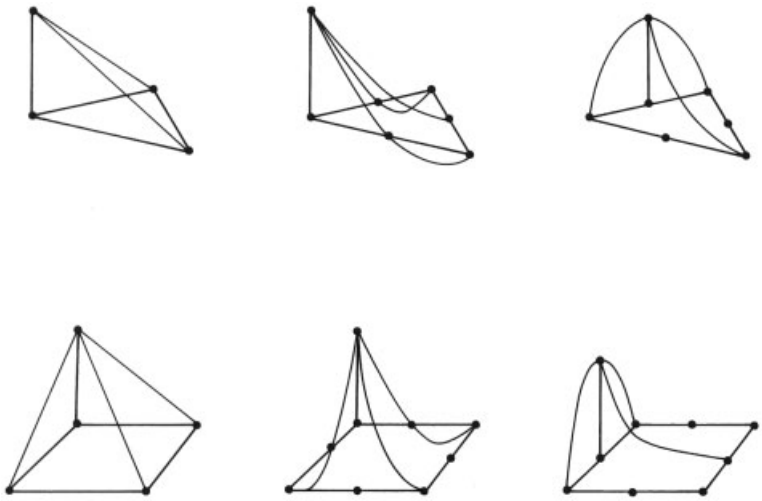

Figure 48. Typical finite element linear and quadratic interpolation functions

of the interpolation functions, it is clear that at node j, $x = x_j$ and $y = y_j$, the only nonzero terms in the series representation are $\Phi_j\,(x_j, y_j)$ and $\Psi_j\,(x_j, y_j)$ where they take on the value of unity. Hence

$$u_x\,(x_j,y_j) = u_{xj}\Phi_j\,(x_j,y_j) = u_{xj} \tag{179a}$$
$$u_y\,(x_j,y_j) = u_{yj}\Phi_j\,(x_j,y_j) = u_{yj} \tag{179b}$$
$$p\,(x_j,y_j) = p_j\Psi_j\,(x_j,y_j) = p_j \tag{179c}$$

Thus the numerical coefficients in the series representation are also the values of the unknown velocity components and pressure at the nodal points.

A considerable amount of algebraic manipulation and numerical integration must be carried out to set up and solve these algebraic equations. Some of these steps require considerable ingenuity in computer programming. In a well designed finite element computer package most of these are taken care of by the computer requiring a minimal amount of human intervention. Because of the inertia terms and variable fluid properties, the algebraic equations are again nonlinear in the unknown coefficients. A procedure similar to that outlined above for handling the set of nonlinear algebraic equations arising from the finite difference method can also be applied here, the details for implementing the iterative procedure may, of course, be quite different.

In the finite element method u_x, u_y, and p appear explicitly in the computation. It is again possible to formulate a finite element scheme where the unknown variable is the stream function or a combination of stream function and vorticity. In the case of viscoelastic fluids, it is also common practice to have the deviatoric stress components τ_{xy}, τ_{xx}, and τ_{yy} as well as u_x, u_y, and p appearing explicitly in the finite element computation scheme.

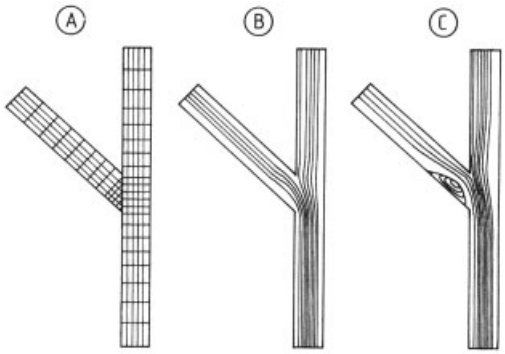

Figure 49. Finite element mesh and streamlines in a branching channel at increasing Reynolds number obtained by finite element computation
A) Finite element mesh; B) Streamlines for $Re = 10$; C) Streamlines for $Re = 100$

Figure 49 shows the streamlines obtained by finite element computation for the flow in a branching channel. The finite element grid employed is shown in Figure 49A. The use of a small number of triangular elements together with rectangular elements has made the discretization of the irregularly shaped flow field a

relatively simple task. A Newtonian fluid, traveling vertically upwards, enters the channel as a single stream and leaves as two separate streams, generally of unequal size. The effects of increasing Reynolds number, based on upstream channel width, now show up as a recirculating zone in the branch channel. At low Reynolds number, the flow is divided equally between the two branches (Fig. 49B). As the Reynolds number is increased (Fig. 49C), the flow through the straight channel increases at the expense of the branch channel.

5.3. General Remarks

Both the finite difference method and the finite element method are now routinely used to solve fluid flow problems. A detailed discussion of the relative merits of these two methods would not serve a very useful purpose and is certainly out of place here. It is, however, generally agreed that the principles of the finite difference method are easier to understand and simpler to implement on computers. In contrast the finite element method can cope with irregular flow fields very efficiently, better than the finite difference method.

The development of a computer program to solve a nontrivial flow problem numerically is a major undertaking that requires expertise not only in fluid mechanics, but also in numerical analysis, computer programming, and organization. It is not a task that can normally be carried out by a single person working in isolation. A number of commercial software packages specially designed for solving fluid flow problems are now available. Some of these packages are very well designed and simple to use. Most of them incorporate special procedures for dealing with highly nonlinear flow problems. They usually have built-in computer graphics for efficient presentation of the large amount of numerical results generated by these packages. Intelligent use of these computing tools coupled with a sound understanding of the physics and mathematics of fluid mechanics has lead rapidly to advances in the solution of flow problems that would otherwise be intractable.

6. References

1. M. M. Denn: *Process Fluid Mechanics,* Prentice-Hall, Englewood Cliffs 1980.
2. S. W. Churchill: *Viscous Flow the Practical Use of Theory,* Butterworths, Boston 1988.
3. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
4. R. B. Bird, R. C. Armstrong, O. Hassager: *Dynamics of Polymeric Liquids,* 2nd ed., vol. 1, J. Wiley & Sons, New York 1987.
5. H. Schlichting: *Boundary-Layer Theory,* 7th ed., McGraw-Hill, New York 1977.
6. K. Walters: *Rheometry,* Chapman and Hall, London 1975.
7. R. W. Whorlow: *Rheological Techniques,* Ellis Horwood, London 1980.
8. J. M. Dealy: *Rheometers for Molten Plastics,* Van Nostrand, New York 1982.
9. J. O. Hinze: *Turbulence,* 2nd ed., McGraw-Hill, New York 1975.
10. D. Kunii, O. Levenspiel: *Fluidization Engineering,* R. E. Krieger Pub. Co., Huntington 1977.
11. J. M. Kay, R. M. Nedderman: *Fluid Mechanics and Transfer Processes,* Cambridge University Press, Cambridge 1985.
12. Y. Taitel, A. E. Dukler, *AIChE J.* **26** (1980) 345.
13. G. E. Alves, *Chem. Eng. Progr.* **50** (1954) 449.
14. J. M. Mandhane, G. H. Gregory, K. Aziz, *Int. J. Multiphase Flow* **1** (1974) 537.
15. P. Griffith: "Two-Phase Flow," in W. M. Rohsenow, J. P. Hartnett, E. N. Ganic (eds.): *Handbook of Heat Transfer Fundamentals,* 2nd ed., McGraw-Hill, New York 1985.
16. F. A. Holland: *Fluid Flow for Chemical Engineers,* E. Arnold, London 1973.
17. R. H. Perry, D. W. Green: *Perry's Chemical Engineers Handbook,* 6th ed., McGraw-Hill, New York 1984.
18. A. S. Foust et al.: *Principles of Unit Operations,* 2nd ed., J. Wiley & Sons, New York 1980.
19. H. A. Barnes, J. F. Hutton, K. Walters: *An Introduction to Rheology,* Elsevier, Amsterdam 1989.
20. A. B. Metzner, M. Whitlock, *Trans. Soc. Rheol.* **2** (1958) 239.
21. Q. D. Nguyen, D. V. Boger, *J. Rheol. (N.Y.)* **29** (1985) 335.
22. A. B. Metzner, J. C. Reed, *AIChE J.* **1** (1957) 434.
23. D. W. Dodge, A. B. Metzner, *AIChE J.* **5** (1955) 189.

24. D. V. Boger: "Viscoelastic Flows Through Contractions," in J. L. Lumley, M. Van Dyke, H. L. Reed (eds.): *Annual Review of Fluid Mechanics,* vol. 19, Ann. Reviews Inc., Palo Alto 1987.
25. J. Meissner, *Pure Appl. Chem.* **42** (1975) 553.
26. R. B. Bird, C. F. Curtiss, R. C. Armstrong, O. Hassager: *Dynamics of Polymeric Liquids,* 2nd ed., vol. 2, J. Wiley & Sons, New York 1987.
27. G. Prilutski, R. K. Gupta, T. Sridhar, M. E. Ryan, *J. Non-Newtonian Fluid Mech.* **12** (1983) 233.
28. M. Doi, S. F. Edwards: *The Theory of Polymer Dynamics,* Oxford University Press, Oxford 1986.
29. R. L. Larson: *Constitutive Equations for Polymer Melts and Solutions,* Butterworths, Boston 1988.
30. C. J. S. Petrie: *Elongational Flows,* Pitman, London 1979.
31. F. T. Trouton, *Proc. R. Soc. London A* **77** (1906) 426.
32. E. B. Bagley, *J. Appl. Phys.* **28** (1957) 624.
33. J. A. Brydson: *Flow Properties of Polymer Melts,* Butterworth, London 1970.
34. D. V. Boger, R. K. Gupta, R. I. Tanner, *J. Non-Newtonian Fluid Mech.* **4** (1978) 239.
35. R. I. Tanner, *J. Polym. Sci. Part A-2* **8** (1970) 2067.
36. P. S. Virk, *AIChE J.* **21** (1975). 625.
37. T. J. Chung: *Finite Element Analysis in Fluid Dynamics,* McGraw-Hill, New York 1978.
38. M. M. Gupta: "Numerical Methods for Viscous Flow Problems," in A. S. Mujumdar, R. A. Mashelkar (eds.): *Advances in Transport Processes,* vol. 1, Wiley, New York 1980.
39. A. J. Baker: *Finite Element Computational Fluid Mechanics,* McGraw-Hill, New York 1983.
40. O. C. Zienkiewicz, K. Morgan: *Finite Elements and Approximation,* Wiley & Sons, London 1983.
41. M. J. Crochet, A. R. Davies, K. Walters: *Numerical Simulation of Non-Newtonian Flow,* Elsevier, Amsterdam 1984.
42. S. C. R. Dennis, G. Z. Chang, *J. Fluid Mech.* **42** (1970) 471.

Fluorene → Hydrocarbons

Estimation of Physical Properties

Ulfert Onken, Fachbereich Chemietechnik, Universität Dortmund, Dortmund, Federal Republic of Germany

Kai Fischer, Fachbereich Technische Chemie, Universität Oldenburg, Oldenburg, Federal Republic of Germany

Jürgen Rarey, Fachbereich Technische Chemie, Universität Oldenburg, Oldenburg, Federal Republic of Germany

1. Introduction 3
1.1. Scope 3
1.2. Data Sources 3
1.3. Theoretical and Empirical Methods 4
2. Molecular and Macroscopic Properties 5
2.1. Intermolecular Forces 5
2.2. Theorem of Corresponding States 7
2.3. Estimation of Critical Properties 8
2.4. Other Characteristic Constants of Pure Compounds 10
3. Thermal and Caloric Properties of Single-Phase Systems 11
3.1. Pressure – Volume – Temperature Behavior of Gases and Liquids . . 11
3.1.1. Pure Gases 11
3.1.2. Pure Liquids 13
3.1.3. Mixtures 15
3.1.3.1. Gas Mixtures 15
3.1.3.2. Liquid Mixtures 17
3.2. Caloric Properties 17
3.2.1. Heat Capacity of Ideal Gases 18
3.2.2. Heat Capacity of Pure Liquids . . . 18
3.2.3. Heat Capacities of Liquid Mixtures 20
4. Phase Equilibria 20
4.1. Pure Compounds 20
4.1.1. Vapor Pressure, Boiling Point, and Melting Point 20
4.1.2. Enthalpy of Vaporization 22
4.2. Mixtures 24
4.2.1. Vapor – Liquid Equilibria with Liquid-Phase Activity Coefficients 25
4.2.1.1. Basic Considerations 25
4.2.1.2. Data Correlation: Binary and Multicomponent Mixtures of Nonelectrolytes 25
4.2.1.3. Prediction of Equilibrium Data . . . 26
4.2.2. Vapor – Liquid and Gas – Liquid Equilibria with Equations of State . 30
4.2.2.1. Basic Considerations 30
4.2.2.2. Data Correlation 30
4.2.2.3. Prediction of Vapor-Liquid and Gas-Liquid Equilibria 31
4.2.3. Liquid – Liquid Equilibria 31
4.2.4. Solid – Liquid Equilibria 32
5. Thermodynamic Data of Chemical Reactions 33
5.1. Definitions 33
5.1.1. Reaction Enthalpy 33
5.1.2. Gibbs Energy of Reaction and Chemical Equilibrium 33
5.2. Group Contribution Methods for the Estimation of Enthalpies and Gibbs Energies of Formation . . . 35
5.3. Applications 37
5.3.1. Reaction Enthalpy 37
5.3.2. Chemical Equilibrium 38
6. Transport Properties of Pure Compounds and Mixtures 40
6.1. Viscosity 40
6.1.1. Viscosity of Pure Gases 40
6.1.2. Viscosity of Gas Mixtures 42
6.1.3. Viscosity of Pure Liquids 42
6.1.4. Viscosity of Liquid Mixtures 43
6.2. Thermal Conductivity 45
6.2.1. Thermal Conductivity of Gases at Low Pressure 45
6.2.2. Thermal Conductivity of Gases at High Pressure 47
6.2.3. Thermal Conductivity of Gas Mixtures 49
6.2.4. Thermal Conductivity of Liquids . . 49
6.2.5. Thermal Conductivity of Liquid Mixtures 51
6.3. Diffusion Coefficients 51
6.3.1. Diffusion Coefficients of Gases at Low and Moderate Pressures 52
6.3.2. Diffusion Coefficients of Liquids . 53
7. Surface Tension of Liquids 56
7.1. Surface Tension of Pure Liquids . 56
7.2. Surface Tension of Liquid Mixtures 58
8. References 59

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_06

Symbols

a	activity
B, B'	second virial coefficient, m^3/mol or Pa
c	molar heat capacity, J mol^{-1} K^{-1}
C, C'	third virial coefficient, $(m^3/kmol)^2$ or Pa^2
D	diffusion coefficient, m^2/s
E	energy, J; Eucken factor
f	fugacity, Pa or atm
G	molar Gibbs energy (free enthalpy), J/mol
ΔG_f	Gibbs energy of formation, J/mol
ΔG_f^0	Gibbs energy of formation at standard pressure, J/mol
ΔG_r	Gibbs energy of reaction, J/mol
H	molar enthalpy, J/mol
ΔH_f	enthalpy of formation, J/mol
ΔH_f^0	enthalpy of formation at standard pressure, J/mol
ΔH_r	enthalpy of reaction, J/mol
ΔH^{lv}	enthalpy of vaporization, J/mol
$\Delta H^{lv}_{T_b}$	enthalpy of vaporization at boiling temperature, J/mol
ΔH^{sv}	enthalpy of sublimation, J/mol
ΔH^{trs}	enthalpy of phase transition, J/mol
I	increment
k	Boltzmann constant, J/K
K	equilibrium constant
K_p	equilibrium constant expressed in partial pressure
M	molecular mass, g/mol
n	number of moles
p	pressure, Pa (or bar)
P	parachor, $m^3\,kg^{0.25}\,s^{-0.5}\,mol^{-1}$
Q	group interaction parameter
r	distance, m; number of components of a mixture
R	universal gas constant, J mol^{-1} K^{-1}
S	entropy, J/K
T	temperature, K, °C
V	molar volume, m^3/mol
w	weight fraction
x	mole fraction in liquid phase
y	mole fraction in gas phase
z	compressibility factor; coordination number
α	relative volatility
β	compressibility, Pa^{-1}
ε	Lennard – Jones energy, J
γ	activity coefficient
η	dynamic viscosity, Pa · s
λ	thermal conductivity, W m^{-1} K^{-1}
μ	chemical potential, J/mol; dipole moment, Debye
ν	kinematic viscosity, m^2/s; number of groups in a molecule; stoichiometric coefficient
ϱ	density, kg/m^3
σ	Lennard – Jones length, 10^{-10} m; surface tension, N/m
ω	acentric factor
Δ	difference
χ	Stiel polar factor
Γ	group activity coefficient
φ	fugacity coefficient
Π	Poynting correction
Ω	collision integral

Subscripts

b	at normal boiling point (101.325 kPa)
c	critical property
D	diffusion
id	ideal state (gas, mixture)
i, j, k	components
m	mixture; at melting point
p	constant pressure
r	reduced quantity; reaction
sat	state of saturation
T	constant temperature
v	constant volume
0	pure component

Superscripts

C	combinatorial
E	excess
g	gas
int	internal
l	liquid
s	solid
ref	reference
R	residual
trans	translational
trs	transition
v	vapor
vap	vaporization
0	reference-state value, atmospheric pressure
∞	infinite dilution
cal	calculated value
exp	measured value

1. Introduction

1.1. Scope

The design of chemical processes requires a knowledge of the values of the properties of the individual chemical compounds and mixtures to be handled, as well as data on the chemical reaction systems involved in the process. Important types of relevant data are given in Table 1; other data may, however, also be required (e.g., for optical or electrical properties).

Table 1. Data for chemical processes

Type of compounds, material, system	Type of data
Individual chemical compounds (starting and end products, intermediates)	physical properties [a]
	1) thermal state properties [a], e.g., density, vapor pressure 2) caloric properties [a], e.g., heat capacity, latent heats 3) transport properties [a], chemical properties [b] (safety, toxicity, environmental hazards)
Mixtures, nonreacting	physical properties [a] phase equilibria [a]
Mixtures, reacting	chemical equilibrium [a] reaction enthalpy [a] reaction velocity [a, b] catalytic activity [a, b]
Materials of construction	mechanical properties [b] corrosion [b]

[a] Data required for basic design of processes.
[b] Data estimation not possible.

Although all the classes of data listed in Table 1 are relevant for the design and operation of chemical processes, only some of them are essential. Thus basic design of a process unit for liquid – liquid extraction requires data on both the selectivity and capacity of the solvent, whereas data on toxicity and inflammability are not necessary, although they are important for plant operation. The present review of estimation methods covers only those groups of data that are essential for process design. Reaction velocity data cannot, however, be included, despite the fact that they are very important. On the basis of present knowledge reaction velocities cannot be predicted with an acceptable degree of reliability and this situation is not expected to change in the near future.

Before estimating data, data tables and other sources should be consulted for measured values (see Section 1.2). If no such data can be found (as is often the case), experimental determination of missing data can be considered. For several properties (e.g., density or vapor pressure) such experiments can be performed without much expense. However, for those physical properties whose measurement requires appreciable experimental effort (e.g., phase equilibria), information from data sources is rather scarce.

Another, even more crucial problem in data retrieval is the fact that multicomponent mixtures have to be handled in most chemical processes. Since the concentrations in these mixtures are important variables that may vary both locally and with time, it is practically impossible to provide all experimental data for the required concentration ranges.

These considerations show that methods for data estimation and prediction are necessary tools for the design of chemical processes. This is true not only for planning production plants, but also for process development. Many property data are not available, especially in processes for new products. Experimental determination is often impossible due to shortage of time. In these cases qualified data estimation is necessary. Other areas in which data estimation and prediction are widely used are feasibility and comparative process studies. These activities must often be performed within a short time and with a limited financial budget.

1.2. Data Sources

The first step in obtaining the data needed for a design problem is a search for experimental data; such data can generally be considered to be more reliable than predicted data. Moreover, most methods of data estimation require knowledge of certain property data for the respective compounds (e.g., boiling point, density, critical temperature, and critical pressure). Besides, prediction of mixture properties is more reliable when it is based on qualified experimental data for the pure components.

Available data sources range from handbooks for everyday work (D'Ans – Lax [73], Handbook of Chemistry and Physics [74]) to the comprehensive tables of Landolt – Börnstein [75]

and the numerous specialized data collections, e.g., the Dechema Data Series [76] and other data books (see General References [10] to [71]).

An efficient and rapidly developing way of obtaining data for physical and chemical properties and equlibria is data retrieval with computerized data banks (e.g., [77–79]). Prediction methods for several types of data (e.g., phase equilibria) have been incorporated into these data banks.

1.3. Theoretical and Empirical Methods

Methods for estimating property data [1–9] may differ in nature. They may be based on laws of physics and physical chemistry, or they may employ purely empirical rules; most methods can be classified somewhere between these two extremes.

The enthalpy of vaporization, for example, can be determined with a thermodynamic equation, i.e., the Clausius – Clapeyron equation for the dependence of vapor pressure p^{lv} on the absolute temperature T

$$\frac{\mathrm{d}p^{\mathrm{lv}}}{\mathrm{d}T}=\frac{\Delta H^{\mathrm{lv}}}{T\left(V^{\mathrm{v}}-V^{\mathrm{l}}\right)} \tag{1.1}$$

where ΔH^{lv} is the enthalpy of vaporization, V^{v} is the volume of the vapor phase, and V^{l} the volume of the liquid phase. Equation (1.1) is a rigorous relation; when it is used to determine the vaporization enthalpy, vapor pressure data at several temperatures must be known as well as the densities of the vapor and liquid phases at the saturated vapor pressure p^{lv}. At low vapor pressures ($\leq$ ca. 0.1 MPa) V^{l} is small compared to V^{v} and may therefore be neglected. Furthermore, the vapor phase may be assumed to show ideal behavior. With these simplifications Equation (1.1) can be expressed as:

$$\frac{\mathrm{d}\ln p^{\mathrm{lv}}}{\mathrm{d}\left(1/T\right)}=-\frac{\Delta H^{\mathrm{lv}}}{R} \tag{1.2}$$

where R is the gas constant. For vaporization at 0.1 MPa the deviations in ΔH^{lv} caused by neglecting V^{l} are usually $<1\,\%$ and $\leq$ ca. 5 % by assuming ideal gas behavior for the vapor phase. Real gas behavior may be accounted for by the second virial coefficient B' (see Section 3.1, Eq. 3.13):

$$p^{\mathrm{lv}}V^{\mathrm{v}}=RT\left(1+B'p^{\mathrm{lv}}\right) \tag{1.3}$$

Substituting this into Equation (1.2) gives:

$$\frac{\mathrm{d}\ln p^{\mathrm{lv}}}{\mathrm{d}\left(1/T\right)}=-\frac{\Delta H^{\mathrm{lv}}}{R\left(1+B'p^{\mathrm{lv}}\right)} \tag{1.4}$$

In contrast to this exact procedure, enthalpies of vaporization $\Delta H^{\mathrm{lv}}_{T_{\mathrm{b}}}$ at the boiling point T_{b} can be estimated roughly with the simple rule of Pictet and Trouton which states that the ratio of $\Delta H^{\mathrm{lv}}_{T_{\mathrm{b}}}$ to T_{b} should have the same value for all liquids. As can be seen from Table 2, this simple rule is true for many compounds:

$$\Delta H^{\mathrm{lv}}_{T_{\mathrm{b}}}/T_{\mathrm{b}}=\text{const.}\approx 88\,\mathrm{J\,K^{-1}\,mol^{-1}} \tag{1.5}$$

This behavior is presumably related to intermolecular forces. A possible explanation is that the energy necessary for vaporizing a liquid is mainly used to separate the liquid molecules from each other. The strength of intermolecular forces should thus be more or less independent of the type of compound and molecular structure.

Table 2. Rule of Pictet and Trouton

Substance	$\Delta H^{\mathrm{lv}}_{T_{\mathrm{b}}}$, kJ/mol	T_{b}, K	$\Delta H^{\mathrm{lv}}_{T_{\mathrm{b}}}/T_{\mathrm{b}}$, $\mathrm{J\,mol^{-1}\,K^{-1}}$
He	0.8	4.2	19.8
Ne	1.8	27.1	65.0
Ar	6.5	87.3	74.5
H_2	9.0	20.4	44.3
N_2	5.6	77.4	72.1
CH_4	8.2	111.6	73.3
HCl	16.2	188.1	85.8
CS_2	26.8	319.3	83.8
CCl_4	30.0	349.9	85.8
n-C_6H_{14}	29.0	341.9	84.7
n-C_8H_{18}	34.6	398.5	86.8
C_6H_6	30.8	353.3	87.2
$C_2H_5OC_2H_5$	26.6	307.8	86.4
C_2H_5Cl	24.6	285.3	86.2
NH_3	23.4	239.7	97.4
H_2O	40.7	373.2	109.0
C_2H_5OH	38.7	351.6	110.1
$C_6H_5NH_2$	45.1	457.5	98.6
CH_3COOH	23.7	391.7	60.5
Hg	59.1	629.8	93.9
Na	89.0	1163.0	76.6
Zn	114.7	1180.0	97.2
Pb	179.5	2025.0	88.6
Ni	374.3	2915.0	113.8

Associating liquids obviously do not obey the rule of Pictet and Trouton. The carboxylic acids

(formic and acetic acid) show negative deviations for Trouton's constant and form dimers in both the liquid and the vapor phases. This means that in the vaporization process the effective molar mass is higher than that of the monomeric acid, resulting in a higher value of $\Delta H^{\mathrm{lv}}_{T_\mathrm{b}}$ and consequently of Trouton's constant. For the associating liquids in Table 2 showing positive deviations from Equation (1.5) (i.e., ammonia, water, methanol, ethanol) association is limited to the liquid phase; additional energy is required to vaporize these liquids because the bonds responsible for association have to be broken.

For liquified gases with very low boiling points (e.g., helium, hydrogen) Trouton's constant is much lower than $88\,\mathrm{J\,mol^{-1}\,K^{-1}}$, whereas some of the higher boiling metals show a tendency towards higher values for this parameter. Evidently at the reference temperature T_b of Equation (1.5) conditions are not really comparable. Several modifications of Equation (1.5) have therefore been proposed of which only that proposed by Kistiakowsky is mentioned here, because it does not require additional data [80]:

$$\frac{\Delta H^{\mathrm{lv}}_{T_\mathrm{b}}}{T_\mathrm{b}} = 36.6 + R \ln T_\mathrm{b} \left[\mathrm{J\,kmol^{-1}K^{-1}}\right] \tag{1.6}$$

Example. Vaporization enthalpy of diethyl ether

Boiling point [81]:	$T_\mathrm{b} = 34.60\,^\circ\mathrm{C}$ (307.75 K)
Vapor pressure [81]	at 30.00 °C (303.15 K): $p^{\mathrm{lv}} = 0.08629\,\mathrm{MPa}$
	at 40.00 °C (313.15 K): $p^{\mathrm{lv}} = 0.12280\,\mathrm{MPa}$

Enthalpy of vaporization at boiling point ($\Delta H^{\mathrm{lv}}_{T_\mathrm{b}}$, kJ/mol):

Calculated value	
Clausius – Clapeyron equation:	
simplified version (Eq. 1.2)	27.86
with second virial coefficient ($B' = -0.43\,\mathrm{MPa^{-1}}$ [82]), (Eq. 1.4)	26.65
Pictet – Trouton rule (Eq. 1.5)	27.10
Equation (1.6)	25.90
Experimental value [83]	26.70

The above example demonstrates that results from theoretical equations will be more reliable than estimations based on empirical rules provided that the theoretical relationships have not been grossly simplified. The increase in accuracy obtained by using the more exact Equation (1.4) instead of the simplified Equation (1.2) was only possible by the input of additional data. Such data are often not available, and a suboptimum result has to suffice. As with the rule of Pictet and Trouton, pitfalls may often be encountered when many other empirical and semi-empirical methods are applied to compounds with special or extreme molecular properties, for example, those that form hydrogen bonds (e.g., water, hydrogen fluoride, trichloromethane, alcohols, carboxylic acids) or that have a very low molecular mass (hydrogen, helium).

2. Molecular and Macroscopic Properties

2.1. Intermolecular Forces

Physical macroscopic properties of substances are determined by molecular properties and interactions. Molecular properties may be classified as internal or external. *Internal properties* (e.g., spectroscopic properties and chemical reactivity) are directly determined by the chemical structure of the molecule, i.e., composition, type of chemical bond, and steric structure. *External properties* are mainly relevant to molecular interaction (e.g., molecular cross section, dipole moment, polarizability) and are of more general character although they are also related to chemical structure. A strict distinction between internal and external molecular properties cannot be made. For example, optical spectra with distinct narrow lines in the gas phase may show line broadening when the spectrum is recorded from the liquid phase. This is obviously due to molecular interaction.

Thermal behavior of matter (i.e., the dependence of density on temperature and pressure, phase equilibria etc.) has its origin in interaction between molecules, caused by forces of attraction and repulsion.

Intermolecular forces include:

1) Electrostatic forces between ions, permanent dipoles and multipoles (e.g., quadrupoles), as well as ions and dipoles
2) Induction forces, e.g., dipoles caused by the influence of the electric field of permanent dipoles on the electronic cloud of nonpolar molecules (displacement of electrons)

3) Dispersion forces (so-called Van der Waals forces), i.e., forces caused by the interactions of fluctuating dipoles of nonpolar molecules
4) Chemical forces leading to the formation of associates and complexes

Molecular interactions are functions of the spatial positions of molecules relative to each other. These positions are determined by the potential energy E of the particles, which is the sum of the repulsive and attractive forces, E_{rep} and E_{att}, respectively. E_{rep} affects E positively while E_{att} contributes negatively:

$$E = E_{\mathrm{rep}} - E_{\mathrm{att}} \tag{2.1}$$

In the simple case of the interaction between two spherically symmetric molecules, E is a function of the distance r between the molecules only. For the general case of nonspherical molecules E is dependent on the distance and the orientation of the molecules. The contribution of the attractive forces E_{att} to the potential energy is in general a function of the reciprocal of the sixth power of the intermolecular distance r (except in the case of ions and chemical forces):

$$E_{\mathrm{att}} \sim \frac{1}{r^6} \tag{2.2}$$

Consequently, Equation (2.2) is valid for all types of "normal" nonpolar and polar species, but not for compounds or mixtures showing association or dissociation.

Little is known about the characteristics of repulsive forces. They are, however, obviously of stronger influence at very small distances than attractive forces. A commonly used relation between E_{rep} and the distance r is

$$E \sim \frac{1}{r^{12}} \tag{2.3}$$

The Lennard–Jones (LJ) 12–6 potential is obtained from Equations (2.1)–(2.3) and describes the dependence of $E(r)$ on the intermolecular distance

$$E(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{2.4}$$

where ε is the LJ energy and σ the LJ length; these parameters are characteristic for a given species of molecules. As shown in Figure 1, ε is the minimum of the potential energy at a distance r_0. The LJ length σ determines the distance between two molecules at which E is zero. The two characteristic parameters of the LJ potential and other two-parameter models can, in principle, be related not only to molecular data, but also to macroscopic properties (e.g., critical temperature and critical pressure). Hence it should be possible to determine the molecular parameters ε and σ from macroscopic quantities. However, in practice such calculations using statistical mechanics are rather cumbersome. Therefore empirical relations have been proposed, for example by Stiel and Thodos [84]:

$$\varepsilon = 65.3 k T_{\mathrm{c}} z_{\mathrm{c}}^{18/5} \tag{2.5}$$

$$\sigma = 1.866 \times 10^{-3} V_{\mathrm{c}}^{1/3} z_{\mathrm{c}}^{-6/5} \left[10^{-10}\,\mathrm{m}\right] \tag{2.6}$$

where k is the Boltzmann constant, T_{c} the critical temperature, z_{c} the critical compressibility factor, and V_{c} the critical molar volume.

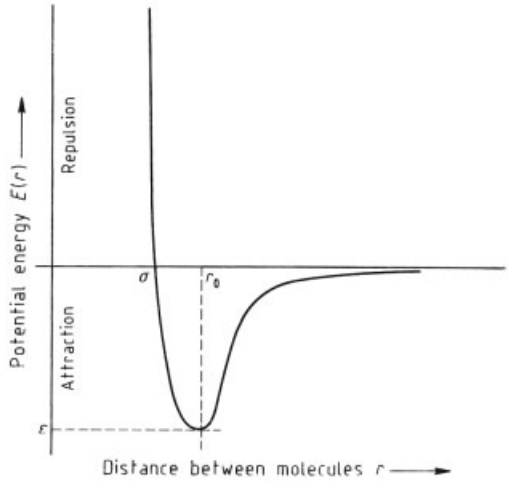

Figure 1. The Lennard–Jones 12-6 potential energy of spherical molecules

Equations (2.5) and (2.6) demonstrate the analogy between the intermolecular potential and the theorem of corresponding states as a general concept for the pVT behavior of fluid phases (see Section 2.2). As a consequence this molecular approach using a two-parameter potential function is also called the microscopic corresponding states theorem. The application of both concepts is not limited to spherical molecules; they may also be used for aspherical molecules with low to moderate polarity.

Among the various two-parameter potential functions [85] the LJ potential is preferred because of its simplicity. Thus Equation (2.4) is used for the estimation of virial coefficients and transport properties (e.g., viscosity, diffusion coefficients). In order to account for effects of polar structures, a third parameter may be introduced, as in the Stockmayer potential [86]. The Stockmayer potential is an extension of the LJ potential and takes the dipole moment into consideration.

2.2. Theorem of Corresponding States

Van der Waals was the first to successfully quantify the finding that the pressure volume temperature (pVT) behavior in the gas and liquid states of many compounds is similar. He proposed an equation of state containing two parameters a and b specific for the compound considered:

$$\left(p+\frac{a}{V^2}\right)(V-b)=RT \tag{2.7}$$

where p is the pressure, V the molar volume, R the gas constant, and T the absolute temperature. This equation (which is discussed in more detail in Section 3.1) describes not only the nonideal gas phase, but also the liquid-phase compressibility and the region of coexisting liquid and vapor phases including the critical point. Since the van der Waals constants a and b are related to the critical properties (denoted here by the subscript c), the introduction of dimensionless reduced variables T_r, p_r, and V_r defined by

$$T_r=\frac{T}{T_c};\ p_r=\frac{p}{p_c};\ V_r=\frac{V}{V_c} \tag{2.8}$$

yields a universal "reduced" equation of state which is in principle valid for all substances. This procedure is justified when all compounds show the same pVT behavior described in reduced variables. Figure 2 shows a pressure – density – temperature diagram in reduced variables. This theorem of corresponding states for fluid phases is conveniently expressed using the compressibility factor z defined by (cf. Section 3.1.1)

$$z=\frac{pV}{RT} \tag{2.9}$$

Hence the following universal function is obtained:

$$z=f\,(T_r,p_r) \tag{2.10}$$

Equation (2.10) implicitly contains the characteristics of the respective chemical compound because T_r and p_r are defined by the critical parameters T_c and p_c. This approach employs only two characteristic parameters to describe the pVT behavior of a compound. Hence satisfactory agreement between theory and experiment cannot be expected for all types of substances. However, for a large number of "normal" compounds (e.g., weakly polar, nonassociating substances, see Section 1.3) reasonable quantitative representation of pVT behavior can be achieved by applying methods based upon the corresponding states principle with two parameters [87]. Calculations of gas-phase properties are generally more reliable than those of the liquid state. The region close to the critical point cannot be represented with sufficient accuracy. For compounds with a very low boiling point (e.g., hydrogen, nitrogen), modified critical parameters must be used for the definitions of the reduced variables [88]. In order to apply corresponding states methods to polar compounds a third characteristic parameter is necessary, e.g., the acentric factor or the polar factor of Stiel (Section 2.4).

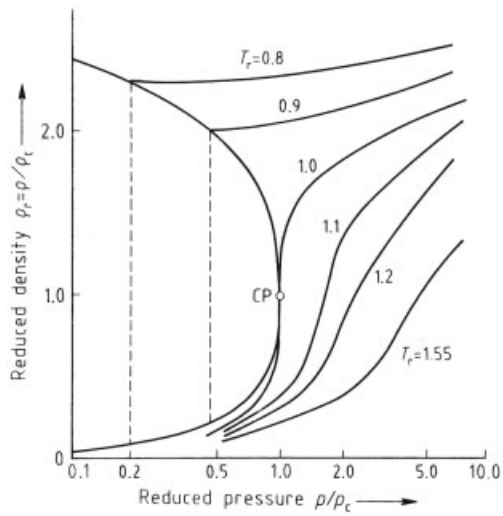

Figure 2. Pressure – density diagram of a pure substance

2.3. Estimation of Critical Properties

In view of the benefits of the principle of corresponding states much attention has been devoted to the critical properties of pure substances. Knowledge of these properties is important for many methods of estimating pure component properties.

Since measurement of critical volume, pressure, and temperature is difficult and tedious, estimation methods are useful tools for providing information about the critical constants. Most estimation techniques employ the principle of additivity which assumes that a physical property can be obtained from the sum of the contributions of the atoms, functional atomic groups, and bonds of the molecule considered. These group contribution or incremental methods are widely used for the prediction of physical properties.

Critical Temperature. According to the empirical Guldberg rule the ratio of the normal boiling point T_b to the critical temperature T_c is equal to 2/3:

$$T_b/T_c = 2/3 \tag{2.11}$$

Based on the more general form of this rule

$$T_c = \text{const.} T_b \tag{2.12}$$

various methods have been proposed for predicting critical temperatures [88], [89], e.g., the widely used procedure of LYDERSEN [90] (for equations and increments, see [89]). With efficient statistical methods and new experimental data, JOBACK [91], [92] developed an improved version of Lydersen's method, which is given by:

$$T_c = \frac{T_b}{0.584 + 0.965 \sum s_i I_{T_c} - (\sum s_i I_{T_c})^2} \tag{2.13}$$

where s_i is the number of structural groups of type i and I_{T_c} the Joback increment of functional groups as given in Table 3. The information needed for the estimation of the critical temperature comprises the normal boiling point and the tabulated group increments.

Critical Pressure. For the prediction of critical pressure p_c JOBACK [91], [92] used an equation containing the number of atoms N in the molecule and the increments I_{p_c} for the functional groups (Table 3):

$$p_c = \frac{0.1}{(0.113 + 0.0032N - \sum s_i I_{p_c})^2} \text{ [MPa]} \tag{2.14}$$

Critical Volume. The difficulties encountered in the experimental determination of the critical molar volume V_c lead to difficulties in the accurate prediction of this quantity too. When Joback's method is used [91], [92], only increments are needed:

$$V_c = 17.5 + \sum s_i I_{V_c} \tag{2.15}$$

where V_c is the critical volume [cm^3/mol] and I_{V_c} are Joback's increments (Table 3).

Example. Calculation of critical properties of chlorobenzene, C_6H_5Cl, with Joback's method and for comparison the results obtained with Lydersen's method.

Functional groups: 5 =C− (ring), 1 =C< (ring), 1 −Cl; T_b= 404.9 K [93]; superscripts: exp = experimental value, cal = calculated value

Critical temperature:	Deviation
T_c^{exp} = 632.4 K	$T_c^{cal} - T_c^{exp}$
Joback (Eq. 2.13): $\sum s_i I_{T_c}$ = 0.0658	
	− 2.9 K
T_c^{cal} = 404.9/(0.584 + 0.0635 − 0.00433) = 629.5 K	
Lydersen T_c^{cal} = 629.6 K	− 2.8 K
Critical pressure:	
p_c^{exp} = 4.52 MPa	$(p_c^{cal} - p_c^{exp})/p_c^{exp}$
Joback (Eq. 2.14): $\sum s_i I_{p_c}$ = 0.0014	
	− 1.7 %
p_c^{cal} = 0.1/[0.113 + (0.032 × 12) − 0.0014] = 4.44 MPa	
Lydersen p_c^{cal} = 4.55 MPa	+ 0.6 %
Critical volume:	
V_c^{exp} = 308 cm^3/mol	$(V_c^{cal} - V_c^{exp})/V_c^{exp}$
Joback (Eq. 2.15): $\sum s_i I_{V_c}$ = 295.0	
V_c^{cal} = 17.5 + 295.0 = 312.5 cm^3/mol	+ 1.5 %
Lydersen V_c^{cal} = 310.0 cm^3/mol	+ 0.6 %

Discussion. Table 4 gives an overview of several methods for the estimation of critical properties together with average uncertainties of the predicted values. The uncertainty in the prediction of the critical pressure is considerably higher than that for the prediction of critical temperature; for the critical volume it is often still greater. Higher errors are also observed for molecules of larger size or complex structure because relevant experimental data are scarce.

Table 3. Increments used in the estimation of critical temperature T_c, pressure p_c, volume V_c, normal boiling point T_b, and melting point T_m (Joback's method [91], [92])

Type of increment	I_{T_c}	I_{p^c}	I_{V^c}	I_{T^b}	I_{T^m}
Nonring increments					
$-CH_3$	0.0141	−0.0012	65	23.58	−5.10
$>CH_2$	0.0189	0	56	22.88	11.27
>CH−	0.0164	0.0020	41	21.74	12.64
>C<	0.0067	0.0043	27	18.25	46.43
$=CH_2$	0.0113	−0.0028	56	18.18	−4.32
=CH−	0.0129	−0.0006	46	24.96	8.73
=C<	0.0117	0.0011	38	24.14	11.14
=C=	0.0026	0.0028	36	26.15	17.78
≡CH	0.0027	−0.0008	46	9.20	−11.18
≡C−	0.0020	0.0016	37	27.38	64.32
Ring increments					
$-CH_2-$	0.0100	0.0025	48	27.15	7.75
>CH−	0.0122	0.0004	38	21.78	19.88
>C<	0.0042	0.0061	27	21.32	60.15
=CH−	0.0082	0.0011	41	26.73	8.13
=C<	0.0143	0.0008	32	31.01	37.02
Halogen increments					
−F	0.0111	−0.0057	27	−0.03	−15.78
−Cl	0.0105	−0.0049	58	38.13	13.55
−Br	0.0133	0.0057	71	66.86	43.43
−I	0.0068	−0.0034	97	93.84	41.69
Oxygen increments					
−OH (alcohol)	0.0741	0.0112	28	92.88	44.45
−OH (phenol)	0.0240	0.0184	−25	76.34	82.83
−O− (nonring)	0.0168	0.0015	18	22.42	22.23
−O− (ring)	0.0098	0.0048	13	31.22	23.05
>C=O (nonring)	0.0380	0.0031	62	76.75	61.20
>C=O (ring)	0.0284	0.0028	55	94.97	75.97
O=CH− (aldehyde)	0.0379	0.0030	82	72.24	36.90
−COOH (acid)	0.0791	0.0077	89	169.09	155.50
−COO− (ester)	0.0481	0.0005	82	81.10	53.60
=O (except as above)	0.0143	0.0101	36	−10.50	2.08
Nitrogen increments					
$-NH_2$	0.0243	0.0109	38	73.23	66.89
>NH (nonring)	0.0295	0.0077	35	50.17	52.66
>NH (ring)	0.0130	0.0114	29	52.82	101.51
>N− (nonring)	0.0169	0.0074	9	11.74	48.84
−N= (nonring)	0.0255	−0.0099		74.60	
−N= (ring)	0.0085	0.0076	34	57.55	68.40
−CN	0.0496	−0.0101	91	125.66	59.89
$-NO_2$	0.0437	0.0064	91	152.54	127.24
Sulfur increments					
−SH	0.0031	0.0084	63	63.56	20.09
−S− (nonring)	0.0119	0.0049	54	68.78	34.40
−S− (ring)	0.0019	0.0051	38	52.10	79.93

Comparison of the methods listed in Table 4 shows the considerable improvement of LYDERSEN's method by JOBACK. Somewhat more accurate but more complicated than JOBACK's method is the method of AMBROSE [94], [95]. With FEDOR's method [96] only critical temperatures can be predicted. Normal boiling point data are not needed, but the results are less reliable than those obtained by the methods of JOBACK and AMBROSE.

When the properties of molecules with strongly interacting groups are estimated, the application of classical group contribution methods leads to systematic errors. Moreover, these methods cannot distinguish between certain isomers, e.g., *o*-, *m*-, and *p*-derivatives of benzene. To avoid these shortcomings CONSTANTINOU and GANI [97] proposed a new incremental method which does not require any information other than the molecular structure of the com-

Table 4. Selected methods for the prediction of critical data of organic compounds

Method	Applicability	Information required	Error
Lydersen	T_c, p_c, V_c with increments for 41 functional groups; >Si< increments for T_c and p_c; –B< increment for T_c	T_b for T_c	T_c: 1.5 %
1955 [90]		M for p_c	p_c: 10 % V_c: 4 %
Joback 1984 [91]	T_c, p_c, V_c with increments for 40 functional groups	T_b for T_c	T_c: 0.8 % p_c: 6 % V_c: 2.5 %
Ambrose	T_c (52), p_c (52), V_c (35) with increments for functional groups (number of functional groups given in parentheses)	T_b for T_c	T_c: 0.7 %
1978 [94] 1979 [95]		M for p_c	p_c: 5 % V_c: 3 %
Fedors 1982 [96]	T_c with increments for 48 functional groups	T_b not required	T_c: 5 %
Constantinou, Gani	T_c, p_c, V_c with increments for 78 functional groups and for structural effects (e.g., proximity, isomerism, ring structure)	T_b not required	T_c: 0.9 %
1994 [97]			p_c: 3 % V_c: 2 %

pound. To allow for secondary structural effects, so-called second-order groups are introduced in addition to the 78 first-order groups. Including the increments of second-order groups reduces the error in estimated critical data by a factor of about 1.5 on average.

Comprehensive compilations of critical data have been published for organic compounds [98], [99] and inorganic compounds [100].

2.4. Other Characteristic Constants of Pure Compounds

The Acentric Factor ω. In order to extend corresponding states methods to compounds and mixtures consisting of nonspherical and polar molecules, the Pitzer acentric factor ω was defined as a third parameter:

$$\omega = -\log\left(\frac{p^{lv}}{p_c}\right)_{T_r=0.7} - 1 \tag{2.16}$$

where p^{lv} is the vapor pressure of the pure component. The Pitzer acentric factor is intended to give a measure of nonspherity or acentricity of a molecule's potential force field. For an ideal spherical molecule it should be zero. Acentric factors for some compounds are given below [89]:

Compound	ω	Compound	ω
Methane	0.008	Water	0.344
Nonane	0.445	Ammonia	0.250
Cyclohexane	0.212	Hydrogen	− 0.216
Benzene	0.212	Phenol	0.438
Methanol	0.556	Aniline	0.384
1-Octanol	0.587		

Knowing the critical temperature and pressure of a substance and its vapor pressure p^{lv} (i.e., with the Antoine coefficients *A, B,* and *C*, Eq. 4.5), the acentric factor can be calculated as follows:

$$\omega = \log\left[\frac{p_c}{\exp\left(A - \frac{B}{0.7T_c + C}\right)}\right] - 1 \tag{2.17}$$

If only the normal boiling point T_b is known the Equation of Edmister [101] gives a reasonable estimate (p_c in atmospheres):

$$\omega = \frac{3}{7}\left(\frac{\log p_c}{T_c/T_b - 1}\right) - 1 \tag{2.18}$$

The average uncertainty of this approximation is 5 % [102].

Example. Calculation of the acentric factor for benzoic acid
T_c = 752 K; p_c = 45.6 bar; T_b = 523 K; ω^{exp} = 0.620 [88]; Antoine constants [89] A = 10.5432, B = 4190.70, C = −125.2
Substituting into Equation (2.17)

$$\omega^{cal} = \log\left[\frac{p_c}{\exp\left(10.5432 - \frac{4190.7}{(0.7\times 752) - 125.2}\right)}\right] - 1$$
$$= 0.617$$

Relative deviation = 0.6 %.
With Equation (2.18)

$$\omega^{\mathrm{cal}}=\frac{3}{7}\left(\frac{1}{(752/523)-1}\right)\log\left(\frac{45.6}{1.013}\right)-1=0.618$$

Relative deviation = 0.3 %.

An incremental method for the estimation of the acentric factor has been developed by Constantinou et al. [103]. It is based on the same structural groups as the procedure of Constantinou and Gani [97] for the prediction of critical properties. No experimental data are needed for this method. The average error is 3 %; however, for 31 of the 78 structural groups (e.g., nitrogen and sulfur groups), increments are not available.

The Stiel Polar Factor χ. This factor relates the actual reduced vapor pressure to the value calculated by the Pitzer equation (Eq. 2.16) at the reduced temperature $T_{\mathrm{r}}=0.6$:

$$\chi=\log\left(\frac{p^{\mathrm{lv(exp)}}}{p^{\mathrm{lv(Pitzer)}}}\right)_{T_{\mathrm{r}}=0.6} \tag{2.19}$$

It is therefore also a measure of polarity and is related to the acentric factor by

$$\chi=\log\left(\frac{p^{\mathrm{lv}}}{p_{\mathrm{c}}}\right)_{T_{\mathrm{r}}=0.6}+1.7\omega+1.552 \tag{2.20}$$

Example. Calculation of the Stiel polar factor for benzoic acid p^{lv} $(T_{\mathrm{r}}=0.6)=0.099$ bar (Antoine Eq. 4.5 [88])
Substituting into Equation (2.20):

$$\chi=\log\left(\frac{0.099}{45.6}\right)+1.7\times 0.620+1.552=-0.0573$$

3. Thermal and Caloric Properties of Single-Phase Systems

3.1. Pressure – Volume – Temperature Behavior of Gases and Liquids

3.1.1. Pure Gases

The pressure – volume – temperature (pVT) behavior can be calculated with equations of state. In this review emphasis is placed upon basic principles and easy-to-handle methods. More detailed information is given in [104–106].

For n moles of an ideal gas consisting of rigid molecules of zero volume and zero interaction potential energy the following equation holds

$$pV=RT \tag{3.1}$$

where V is the molar volume.

The ideal gas law is a universal limiting law valid for all substances when the density ϱ approaches 0. Equation (3.1) can usually be used with sufficient accuracy for gases at low pressure $p<<p_{\mathrm{c}}$. The deviation of the behavior of real substances with respect to Equation (3.1) is usually expressed by the compressibility factor z

$$z=\frac{pV}{RT} \tag{3.2}$$

The dependence of z on pressure is plotted in Figure 3 for a real fluid which obeys the principle of corresponding states. Below the critical pressure z is less than unity, i.e., the attractive forces are dominant. Above T_{c} and p_{c}, z exceeds unity due to increasing influence of the repulsive forces.

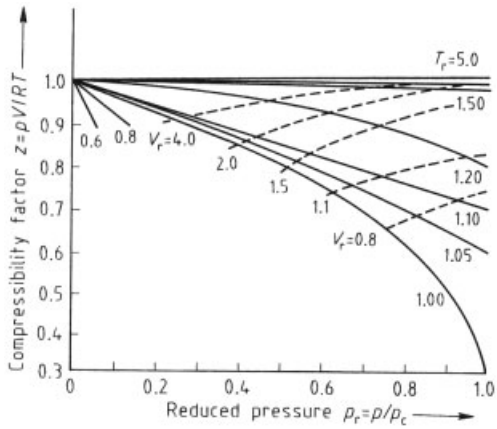

Figure 3. Generalized compressibility chart [107]

In order to describe real fluid behavior an expression must be found of the form

$$z=f\,(T,p)\ \text{ or }\ z=f'\,(T,V) \tag{3.3}$$

that is capable of describing pVT behavior for a wide variety of substances over a wide range of pressure and temperature. The two following

fundamental boundary conditions must be satisfied:

$$\left(\frac{\partial p}{\partial V}\right)_{T_c} = 0 \tag{3.4}$$

$$\left(\frac{\partial^2 p}{\partial V^2}\right)_{T_c} = 0 \tag{3.5}$$

Cubic Equations of State. In his pioneering work of 1873 VAN DER WAALS introduced his famous equation of state which is the basic form for many other cubic (in volume) expressions:

$$\left(p+\frac{a}{V^2}\right)(V-b)=RT \tag{3.6}$$

The substance-specific constants a and b can be determined using conditions (3.4) and (3.5):

$$a=\frac{27}{64}\frac{(RT_c)^2}{p_c} \tag{3.7}$$

$$b=\frac{1}{8}\frac{RT_c}{p_c} \tag{3.8}$$

One of the most widely used two-parameter equations was suggested in 1949 by REDLICH and KWONG (RK):

$$p=\frac{RT}{V-b}-\frac{a}{T^{1/2}V(V+b)} \tag{3.9}$$

Since the parameters a and b are substance-specific constants, the equation is completely determined by the critical properties. The RK equation and its subsequent modifications are more accurate than the Van der Waals equation. One very succesful modification of the attractive term in the RK equation was recommended by SOAVE (RKS) [108]:

$$p=\frac{RT}{V-b}-\frac{a_c\left[1+M(\omega)\left(1-T_r^{1/2}\right)\right]^2}{V(V+b)} \tag{3.10}$$

where

$$M(\omega)=\left(0.48+1.574\omega-0.176\omega^2\right) \tag{3.10a}$$

$$a_c=0.42748\,(RT_c)^2/p_c \tag{3.10b}$$

$$b=0.08664\,(RT_c)/p_c \tag{3.10c}$$

Another widely accepted modification of the RK equation was proposed by PENG and ROBINSON (PR) [109].

The pVT behavior can be predicted with a generalized equation when information on the critical properties and the acentric factor ω is available. These equations are particularly useful for hydrocarbons. According to [110] the average error for this class of substances is 1 – 2 %. Deviations increase in the vicinity of the critical point and for polar substances.

Virial Equation of State. Statistical mechanics yield an equation of state in which the compressibility factor z is given as a polynomial in reciprocal molar volume:

$$\frac{pV}{RT}=z=1+\frac{B}{V}+\frac{C}{V^2}+\ldots \tag{3.11}$$

In this so-called virial equation the virial coefficients B, C, ... are functions of temperature only and can be related to molecular interaction parameters. Another form of the virial equation uses pressure as the independent parameter:

$$\frac{pV}{RT}=z=1+B'p+C'p^2+\ldots \tag{3.12}$$

The two forms of the virial Equations (3.11) and (3.12) can be truncated after the second term if they are applied at low pressure, i.e., at reduced densities < 0.5. Then the coefficients B and B' are related by:

$$B'=\frac{B}{RT} \tag{3.13}$$

Use of the virial equation is very convenient because of its simple form. Experimental data on second and, in part, third virial coefficients of pure gases and mixtures are given in [111].

Calculation of Second Virial Coefficients. Many calculation schemes for the prediction of second virial coefficients have been published, e.g., [112–114]. The frequently used method of Tsonopoulos [113] will be described here briefly. The equations can be used for most nonpolar or slightly polar substances. Real fluid behavior is calculated using Pitzer's approach

$$z=1+B\frac{p_c}{RT_c}\frac{p_r}{T_r}=z^{(0)}+\omega z^{(1)} \tag{3.14}$$

$$B\frac{p_c}{RT_c}=b^{(0)}+\omega b^{(1)} \tag{3.15}$$

with

$$b^{(0)} = 0.1445 - \frac{0.33}{T_r} - \frac{0.1385}{T_r^2} - \frac{0.0121}{T_r^3} - \frac{6.07 \times 10^{-4}}{T_r^8} \quad (3.15a)$$

$$b^{(1)} = 0.0637 + \frac{0.331}{T_r^2} - \frac{0.423}{T_r^3} - \frac{0.008}{T_r^8} \quad (3.15b)$$

For strongly polar compounds TSONOPOULOS suggested the following modification of Equation (3.15)

$$B\frac{p_c}{RT_c} = b^{(0)} + \omega b^{(1)} + \frac{c}{T_r^6} - \frac{d}{T_r^8} \quad (3.16)$$

For several substances (e.g., ketones, aldehydes, ethers, nitriles) TSONOPOULOS gives recommendations for the calculation of the parameter c

$$c = -2.113 \times 10^{-4}\mu_r - 3.885 \times 10^{-21}\mu_r^8 \quad (3.17)$$

where

$$\mu_r = \frac{10^6 \mu_p^2 p_c}{T_c^2} \quad (3.18)$$

(p_c in MPa, T in K, and the dipole moment μ_p in Debye). The parameter d vanishes for substances that do not exhibit hydrogen bonding. For other compounds it has to be determined by experiment.

Example. Calculation of the second virial coefficient of acetone, C_3H_6O, at 310 K with Tsonopoulos's method
$M = 58.08$ g/mol, $T_c = 508.1$ K, $p_c = 4.7$ MPa, $\mu = 2.9$ Debye, $\omega = 0.304$
With $T_r = 0.610$ substitution into Equations (3.15 a) and (3.15 b) gives $b^{(0)} = -0.8537$, $b^{(1)} = -1.3276$.
Substituting into Equation (3.16) and neglecting the third and fourth terms yields:
$B_{nonpolar} = -1.1303\ m^3/kmol$
With $\mu_r = 153.107$ from Equation (3.18), $c = -33.525 \times 10^{-3}$ from Equation (3.17), and substituting into Equation (3.16):
$B_{polar} = -1.715\ m^3/kmol$
Experimental: $B = -1.730\ m^3/kmol$ [111]
Relative deviation = 0.9 %

Other Equations of State. The representation of pVT behavior of polar gases and liquids with two-parameter cubic equations of state is usually unsatisfactory. Accuracy can be improved significantly by applying equations of the Benedict – Webb – Rubin (BWR) type:

$$p = \frac{RT}{V} + \left(B_0RT - A_0 - \frac{C_0}{T^2}\right)\frac{1}{V^2} + \frac{(bRT - a)}{V^3} + \frac{a\alpha}{V^6} + \frac{c}{T^2V^3}\left(1 + \beta/V^2\right)\exp\left(-\beta/V^2\right) \quad (3.19)$$

where A_0, B_0, C_0, a, b, c, α, and β are substance-specific constants. An advantageous modification is the generalized form suggested by LEE and KESLER [115] and its extension by PLÖCKER et al. [116] (LKP equation). For this type of equation of state the compressibility factor of a real fluid is related to the properties of a simple fluid ($\omega = 0$), e.g., to those of n-octane as a reference fluid [105].

A BWR extension with 20 substance-specific parameters and remarkable accuracy was suggested in the 1970s by BENDER [117]. In 1987 POLT reported these constants for 51 substances [118]. Figure 4 shows the $p\,\varrho T$ behavior of toluene calculated with the Bender equation.

3.1.2. Pure Liquids

Because of their importance in industrial processes, density data are available for most organic liquids. Usually at least one experimental value for ambient conditions can be found. It is therefore more important to gain information about the temperature dependence of liquid densities rather than to estimate liquid densities from other properties. Nevertheless, some methods have been published for the estimation of liquid densities, e.g., [119].

Liquid Density at Normal Boiling Temperature. SCHROEDER suggested a very simple and convenient method for the estimation of liquid molar volumes at the normal boiling point [120] (see also [106, p. 52]). He proposed an additive rule for the prediction of molar volumes. The number of atoms of each type of element in the molecule (i.e., carbon, hydrogen, oxygen, and nitrogen, etc.) must be counted, multiplied by a weighting factor, and summed up. Additional increments are then added to account for the molecular structure (i.e., double bonds) as shown in Table 5.

The method gives surprisingly good estimates of molar volume with an average error of ca. 3 %. Exceptions are substances that exhibit

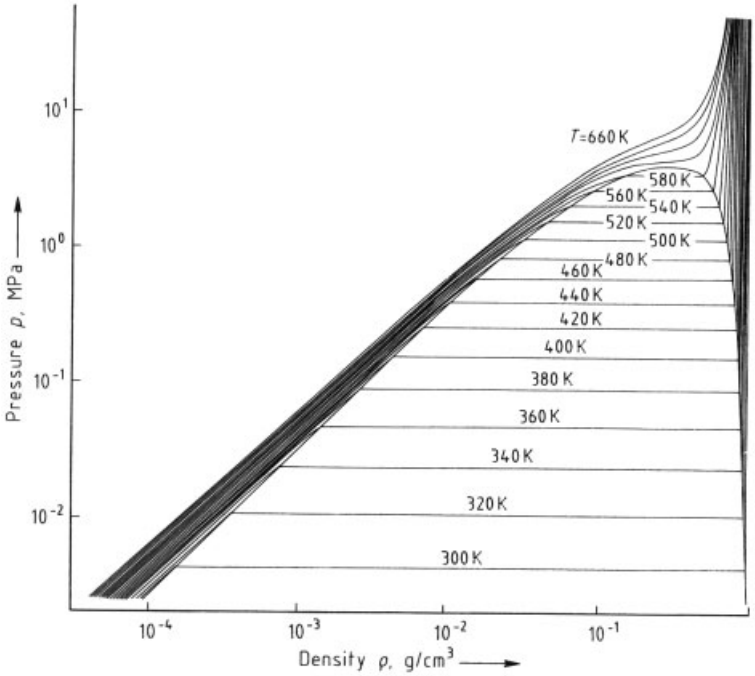

Figure 4. Pressure – density behavior of toluene calculated with the Bender equation of state [118]

Table 5. Increments for the prediction of molar volumes of pure liquids at the normal boiling point according to Schroeder's method [120]

Element	Increment, m^3/kmol	Bond	Increment, m^3/kmol
Carbon	0.007	double bond (C=C)	0.007
Hydrogen	0.007	triple bond (C≡C)	0.014
Oxygen	0.007	ring systems	0.007
Nitrogen	0.007	(including naphthalene, anthracene)	
Bromine	0.0315		
Chlorine	0.0245		
Fluorine	0.0105		
Iodine	0.0385		
Sulfur	0.0210		

hydrogen bonding or a high degree of association.

Example. Calculation of liquid molar volume V_b^l of ethyl acetate, $CH_3COOCH_2CH_3$, at the normal boiling point (101.325 kPa) using Schroeder's method

Groups:	4 C + 2 O + 8 H
Calculated:	$V_b^l = 4\times0.007 + 2\times0.007 + 8\times0.007 = 0.098\ m^3/kmol$
Experimental [121]:	$V_b^l = 0.0978\ m^3/kmol$
Relative deviation = 0.2 %	

More reliable results for the molar volume at normal boiling temperature can be obtained with the group contribution method of Constantinou et al. [122]. The procedure is the same as for the authors' method for estimating the acentric factor (see Section 2.4, [103]). For 17 of the 78 structural groups the values of the increments are not given. The average error in the results is 0.9 %.

More Sophisticated Estimation Methods. Many analytical methods have been published, they are mostly based upon the principle of corresponding states. Consequently, standard density values are necessary for which the critical density can be used. A single experimental value can often be used for very accurate extrapolations with respect to temperature.

The Gunn – Yamada expression is convenient and very accurate (without any additional adjusted parameters) for predicting liquid molar volumes V^l at the state of saturation [123]:

$$V^l/V_{sc}=V_r^{(0)}\,(1-\omega I) \tag{3.20}$$

where I and $V_r^{(0)}$ are generalized functions and V_{sc} is a scaling parameter.
For $0.2 < T_r < 1.0$:

$$I = 0.29607 - 0.09045T_r - 0.04842T_r^2 \tag{3.21}$$

For $0.2 < T_r < 0.8$:

$$V_r^{(0)} = 0.33593 - 0.33953T_r + 1.51941T_r^2 - 2.02512T_r^3 + 1.11422T_r^4 \tag{3.22}$$

For $0.8 < T_r < 1.0$

$$V_r^{(0)} = 1.0 + 1.3\,(1-T_r)^{1/2}\log(1-T_r) - 0.50879\,(1-T_r) - 0.91534\,(1-T_r)^2 \tag{3.23}$$

In terms of $V_{0.6}$ (the saturated liquid molar volume at the reduced temperature $T_r = 0.6$), V_{sc} is defined as

$$V_{sc} = \frac{V_{0.6}}{(0.3862 - 0.0866\omega)} \tag{3.24}$$

If $V_{0.6}$ is not available, V_{sc} can be estimated as follows (p_c in MPa):

$$V_{sc} = RT_c\,(0.2920 - 0.0967\omega)\,\frac{0.10132}{p_c} \tag{3.25}$$

Usually, one experimental value of V^l under ambient conditions ($T_r \approx 0.6$) is available and can be applied as a reference value V^{ref} at T_r^{ref}. In these cases the following form of the method of Gunn and Yamada is useful:

$$\frac{V^l}{V^{ref}} = \frac{V_r^{(0)}\,(T_r)}{V_r^{(0)}\,(T_r^{ref})}\cdot\frac{[1-\omega I\,(T_r)]}{[1-\omega I\,(T_r^{ref})]} \tag{3.26}$$

Example. Calculation of the saturated liquid molar volume V^l of *n*-propylamine, $CH_3CH_2CH_2NH_2$, at 293.15 K using the method of Gunn and Yamada.
$M = 59.112$ g/mol, $T_c = 497$ K, $\omega = 0.303$, $p_c = 4.81$ MPa [121]

Calculated:	
From Equation (3.22)	$V_r^{(0)} = 0.3839$
From Equation (3.25)	$V_{sc} = 0.2287\ m^3/kmol$
With Equation (3.21)	$I = 0.2259$
Substituting in Equation (3.20)	$V^l = 0.08178\ m^3/kmol$
Experimental [121]:	$V^l = 0.0824\ m^3/kmol$
Relative deviation: 0.75 %	

The more recently published method of Hankinson and Thomson achieves higher accuracy for liquid density [124]. This technique employs specific constants for various groups of substances as well as three substance-specific parameters that have been tabulated for more than 400 compounds (e.g., in [106], [124]) or can be estimated. An overview of estimation methods is given in Table 6.

Pressure Dependence of Liquid Density. With the isothermal compressibility β_T

$$\beta_T = -\frac{1}{V}\left(\frac{\partial V}{\partial p}\right)_T \tag{3.27}$$

and the relation [127]

$$\frac{\beta^{1/7}}{V} = \frac{\beta_{sat}^{1/7}}{V_{sat}} \tag{3.28}$$

the following approximation holds:

$$V^l/V_{sat}^l = [1 + 9\beta_{sat}\,(p-p_0)]^{-1/9} \tag{3.29}$$

where β_{sat} is estimated by the relation

$$\beta_{sat} = \frac{V_c}{RT_c}\cdot\left(1 - 0.89\omega^{1/2}\right)\cdot\exp\left(6.9547 - 76.2853T_r + 191.3060T_r^2 - 203.5472T_r^3 + 82.7631T_r^4\right) \tag{3.30}$$

This method holds for the range of reduced temperature $0.35 < T_r < 0.98$ and for reduced pressure of $p_r < 10$ with an average uncertainty of 3 %.

3.1.3. Mixtures

3.1.3.1. Gas Mixtures

The constants used in the methods for the calculation of pVT behavior in gas mixtures must be related to the relevant parameters of the pure components and to the composition of the mixture. Additional correlation parameters may be necessary. The objective of this section is to provide some information on mixing rules.

Table 6. Selected methods for the determination of saturated liquid molar volumes of pure organic compounds

Method	Range of application		Information required	Error, %
	Substances	Temperature		
Yen, Woods [125]	nonpolar and polar	$0.3 < T_r < 1$	ϱ_c, T_c, z_c, z	1
Gunn, Yamada [123]	nonpolar and weakly polar	$0.3 < T_r < 0.99$	$T_c, \omega, \varrho^{ref}$	$(T - T^{ref})/100$
Hankinson, Thomson [124]		$0.25 < T_r < 1.0$	T_c [b], V_c [b], ω_c [b], increments	< 2
Bhirud [126]	nonpolar (polar)		$T_c, \omega, \varrho^{ref}$ [a]	1

[a] Not absolutely necessary.
[b] Tabulated in [106], [124]; can also be calculated.

Virial Equation. Statistical thermodynamics yield the following rigorous relation for the second virial coefficient B_m of a mixture of r components with a quadratic dependency on the molar fractions y_i and y_j:

$$B_m = \sum_i^r \sum_j^r y_i y_j B_{ij} \tag{3.31}$$

The virial coefficient $B_{ii} \equiv B_i$ of the pure component i can be calculated as described in Section 3.1.1. For the cross virial coefficients of the mixture B_{ij} similar methods can be employed if the critical properties of the mixture and the acentric factor ω_{ij} are known.

PRAUSNITZ discusses simple rules that yield sufficiently accurate results for practical applications [128]:

$$T_{cij} = (T_{ci} T_{cj})^{0.5} (1 - k_{ij}) \tag{3.32}$$

where k_{ij} is a specific parameter that has to be adjusted to experimental data.

$$V_{cij} = \left(\frac{V_{ci}^{1/3} + V_{cj}^{1/3}}{2} \right)^3 \tag{3.33}$$

$$z_{cij} = \frac{z_{ci} + z_{cj}}{2} \tag{3.34}$$

The following relations can also be readily derived:

$$p_{cij} = \frac{z_{cij} R T_{cij}}{V_{cij}} \tag{3.35}$$

$$\omega_{ij} = \frac{\omega_i + \omega_j}{2} \tag{3.36}$$

Cubic Equations of State. For cubic equations of state with two parameters (e.g., Eqs. 3.6, 3.9, 3.10), the quadratic mixing rule is usually applied for the constant a. Analogous to Equation (3.31) one obtains for a mixture of r components:

$$a_m = \sum_i^r \sum_j^r y_i y_j a_{ij} \tag{3.37}$$

where the cross coefficients a_{ij} are given by the geometric mean of the pure component parameters a_{ii} and a_{jj} corrected by a binary parameter k_{ij} :

$$a_{ij} = (a_{ii} a_{jj})^{0.5} (1 - k_{ij}) \tag{3.38}$$

For the parameter b, a linear dependence on composition is generally assumed:

$$b_m = \sum_i^r y_i b_i \tag{3.39}$$

Pseudocritical Properties. Since many methods take advantage of the principle of corresponding states using reduced quantities for temperature, pressure, volume, and critical compressibility, knowledge of these properties in mixtures is important. If a mixture with a fixed composition is assumed to behave like a pure fluid, so-called pseudocritical constants can be introduced that serve as scaling factors.

Many attempts to calculate pseudocritical properties have been published. Some basic methods are given below (the subscript m denotes mixture):

1) Kay's rule for T_{cm} [129]

$$T_{\mathrm{cm}} = \sum y_i T_{\mathrm{c}i} \tag{3.40}$$

Deviations from results obtained by using more complicated methods are less than 2 % [110].

2) p_{cm} according to Prausnitz and Gunn [130]

$$p_{\mathrm{cm}} = RT_{\mathrm{cm}} \frac{\sum (y_i z_{\mathrm{c}i})}{\sum y_i V_i^{\mathrm{g}}} \tag{3.41}$$

3) Joffe's rule for the acentric factor ω_{m} of the mixture [131]

$$\omega_{\mathrm{m}} = \sum y_i \omega_i \tag{3.42}$$

Algebraic expressions of varying complexity have been suggested for the mixing rules for each equation of state applicable to mixtures. They are summarized and discussed extensively in [106].

3.1.3.2. Liquid Mixtures

Molar volumes $V_{\mathrm{m}}^{\mathrm{l}}$ of mixtures comprising nonreacting components which do not differ greatly in structure can be calculated for low and moderate pressure ($p < p_{\mathrm{c}}$) by Amagat's rule:

$$V_{\mathrm{m}}^{\mathrm{l}} = \sum x_i V_{0i}^{\mathrm{l}} \tag{3.43}$$

where x_i is the mole fraction of component i. Deviations for real mixtures do not generally exceed ± 1 % of $V_{\mathrm{m}}^{\mathrm{l}}$.

3.2. Caloric Properties

Enthalpy is the basic thermodynamic quantity used in energy balances for chemical processes. According to rigorous thermodynamics, the difference in molar enthalpy (H) between two states (denoted by subscripts 1 and 2) can be calculated via

$$H(T_2,p_2) - H(T_1,p_1) = \int_{T_1}^{T_2} c_p(T,p)\,\mathrm{d}T + \int_{p_1}^{p_2} \left(\frac{\partial H}{\partial p}\right) \mathrm{d}p \tag{3.44}$$

where c_p is the molar heat capacity at constant pressure. Since heat capacity data for high pressures are scarce and the variation in H between two states is independent of the path of integration, the difference in H is preferably determined via

$$H_2 - H_1 = \left[H\left(T_1,p^0\right) - H(T_1,p_1)\right] + \int_{T_1}^{T_2} c_p^0(T)\,\mathrm{d}T - \left[H\left(T_2,p^0\right) - H(T_2,p_2)\right] \tag{3.45}$$

where the superscript 0 indicates the reference state shown by an arrow in Figure 5. As a consequence, the temperature dependence of enthalpy is calculated at zero pressure, i.e., in the ideal gas state. The dependence on pressure is given by the so-called departure functions (e.g., $H - H^0$, $S - S^0$) that can be determined with an appropriate equation of state. For the heat capacity the rigorous thermodynamic equation is

$$c_p - c_p^0 = -T \int_{p^0}^{p} \left(\frac{\partial^2 V}{\partial T^2}\right) \mathrm{d}p \tag{3.46}$$

Expressions for many equations of state as well as extensive tabulations are given in [106], [110], [127].

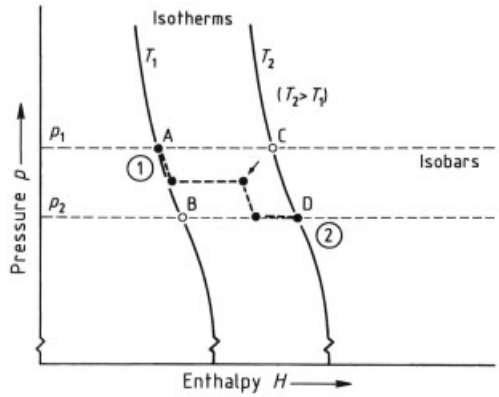

Figure 5. Integration paths for enthalpy calculation between states 1 (T_1, p_1) and 2 (T_2, p_2)
Arrow denotes the reference value referred to in the text

3.2.1. Heat Capacity of Ideal Gases

Classical kinetic theory leads to the simple relation $c_p^0 = 1/2\,R$ for each translational degree of freedom, a contribution of $1/2\,R$ for each axis of rotation, and of R for each vibrational degree of freedom. Neglecting vibrational degrees of freedom, a primitive estimate for c_p^0 applicable to small molecules can be obtained from the relation

$$c_p^0 = R\,(1+f/2) \tag{3.47}$$

where f is the sum of translational and rotational degrees of freedom of the molecule.

Example. Consider molecular nitrogen N_2 with five degrees of freedom (three translational, two rotational). At moderate temperatures (ca. 300 K) Equation (3.47) yields $c_p = 29.1\ \mathrm{J\,mol^{-1}\,K^{-1}}$ whereas the IUPAC compilation [132] for N_2 reports $c_p = 29.2\ \mathrm{J\,mol^{-1}\,K^{-1}}$ at 300 K and 0.1 MPa.

This approximation is not of sufficient accuracy for practical engineering purposes over a wide range of temperatures. Hence empirical estimation methods are commonly accepted for the calculation of the heat capacity of ideal gases.

Group Contribution Methods. Joback's method [133], [134] is a simple additive group contribution method based on a polynomial in temperature using the same structural groups as in his method for the estimation of critical properties (Section 2.3). For a molecule with r different types of groups, c_p^0 can be calculated by

$$\begin{aligned} c_p^0 = &\left(\sum^r s_j I_\mathrm{a} - 37.93\right) + \left(\sum^r s_j I_\mathrm{b} + 0.210\right) T \\ &+ \left(\sum^r s_j I_\mathrm{c} - 3.91\times10^{-4}\right) T^2 \\ &+ \left(\sum^r s_j I_\mathrm{d} + 2.06\times10^{-7}\right) T^3 \end{aligned} \tag{3.48}$$

where s_j is the number of the group j and the increments I_a, I_b, I_c, and I_d, are component-specific parameters and are tabulated for various groups not only for the ideal-gas heat capacity but also for the enthalpy of formation ΔH^0_f298 and the Gibbs enthalpy of formation ΔG^0_f298 (see Chap. 5, Table 14). The average error in c_p^0 is usually less than 2 % [106].

Example. Heat capacity c_p^0 of diisopropyl ether, $(CH_3)_2CHOCH(CH_3)_2$, $M = 102.177$ g/mol.

Group	s_j	I_a	I_b	I_c	I_d
$-CH_3$	4	19.5	-8.08×10^{-3}	1.53×10^{-4}	-9.67×10^{-8}
$>CH-$	2	−23.0	0.204	-2.65×10^{-4}	1.20×10^{-7}
$-O-$	1	25.5	-6.32×10^{-2}	1.11×10^{-4}	-5.48×10^{-8}

Substitution into Equation (3.48):
$c_p^0\ (800\ \mathrm{K}) = 313.1\ \mathrm{J\,mol^{-1}\,K^{-1}}$
Experimental value:
$c_p^0 = 311.5\ \mathrm{J\,mol^{-1}\,K^{-1}}$ at 800 K [106]
Relative deviation: 0.5 %

Benson's method [135] yields accurate values not only for c_p^0 but also for ΔH^0_f298 and S^0_{298}. Many functional groups have been incorporated into this method. Its strength lies in the fact that the neighborhood of each group is considered. For special classes of molecules and isomers additional corrections may be applied (e.g., for rings with heteroatoms).

Other methods for the estimation of molar heat capacities are listed in Table 7.

Heat Capacities of Ideal Gas Mixtures. The heat capacity of a mixture $c_{p\mathrm{m}}^0$ in the ideal gas state is given by the simple mixing rule

$$c^0_{p\mathrm{m}} = \sum x_i c^0_{pi} \tag{3.49}$$

3.2.2. Heat Capacity of Pure Liquids

Several approaches have been proposed for determining the heat capacity of liquids, i.e., theoretical methods, corresponding states, Watson's thermodynamic cycles, or group contribution methods. These methods are discussed extensively in [106].

A successful group contribution method was proposed by Chueh and Swanson [141]. It can be used to calculate the heat capacities of liquids at room temperature (293.15 K) for many organic compounds. The average uncertainty is 2 – 3 %.

Table 7. Selected incremental methods for the calculation of ideal gas heat capacities

Method	Range of application		Error, %
	Substances	*T*, K	
Dobratz 1941 [136]	no triple bonds		
Rihani, Doraiswamy 1965 [137]	no triple bonds		2 – 3
Benson 1968, 1969 [135]		280 – 1100	1
Thinh 1976 [138]	hydrocarbons only	200 – 1500	1 – 2
Yoneda 1979 [139]	hydrocarbons only	280 – 1100	< 2
Joback 1984 [133], [134]		280 – 1100	1 – 2
Harrison, Seaton 1988 [140]		280 – 1500	3

Table 8. Increments of structural groups for the prediction of liquid heat capacities (Missenard's method) [142]

Structural group	Temperature *T*, K					
	248	273	298	323	348	373
$-H$	12.5	13.4	14.6	15.5	16.7	18.8
$-CH_3$	38.5	40.0	41.6	43.5	45.8	48.3
$>CH_2$	27.2	27.6	28.2	29.1	29.9	31.0
$>CH-$	20.9	23.8	24.9	25.7	26.6	28.0
$-\overset{\vert}{\underset{\vert}{C}}-$	8.4	8.4	8.4	8.4	8.4	
$-C\equiv C-$	46.0	46.0	46.0	46.0		
$-O-$	28.9	29.3	29.7	30.1	30.5	31.0
$-CO-$ (ketone)	41.8	42.7	43.5	44.4	45.2	46.0
$-OH$	27.2	33.5	43.9	52.3	61.7	71.1
$-COO-$ (ester)	56.5	57.7	59.0	61.1	63.2	64.9
$-COOH$	71.1	74.1	78.7	83.7	90.0	94.1
$-NH_2$	58.6	58.6	62.8	66.9		
$-NH-$	51.0	51.0	51.0			
$-\overset{\vert}{N}-$	8.4	8.4	8.4			
$-CN$	56.1	56.5	56.9			
$-NO_2$	64.4	64.9	65.7	66.9	68.2	
$-NH-NH-$	79.5	79.5	79.5			
C_6H_5- (phenyl)	108.8	113.0	117.2	123.4	129.7	136.0
$C_{10}H_7-$ (naphthyl)	179.9	184.1	188.3	196.6	205.0	213.0
$-F$	24.3	24.3	25.1	25.9	27.0	28.2
$-Cl$	28.9	29.3	29.7	30.1	30.8	31.4
$-Br$	35.1	35.6	36.0	36.4	37.2	38.1
$-I$	39.3	39.7	40.4	41.0		
$-S-$	37.2	37.7	38.5	39.3		

Missenard published an incremental method for the prediction of liquid heat capacities in the temperature range 250 – 373 K [142]. This method is limited, however, to reduced temperatures $T_r < 0.75$. It employs the equation

$$c_p^l(T) = \sum s_i I_i \tag{3.50}$$

where s_i is the the number of group i and I_i the increment of group i (values are given in Table 8).

A more refined incremental method for the estimation of heat capacities of organic liquids has been developed by Ruzicka and Domalski [143]. With a detailed definition of 29 hydrocarbon and 85 heteroatomic groups and 36 structural corrections, this method predicts heat capacities as a function of temperature from the melting point up to the normal boiling point. The average error in results from these temperature functions is 2 – 6 %; above the boiling point, errors may be 1 – 2 % higher.

Several estimation methods based upon the principle of corresponding states have been discussed by Bondi [144]. Particularly at higher temperatures these methods are superior to the group contribution methods. They require, however, additional information (see also Table 9). Bondi [144] (see also [106]) proposed a modification of Rowlinson's method:

$$c_p^l - c_p^0 = R\cdot\left[1.45 + \frac{0.45}{1-T_r}\right] + R\cdot 0.25\omega\left[17.11 + 25.2\frac{(1-T_r)^{1/3}}{T_r} + \frac{1.742}{1-T_r}\right] \tag{3.51}$$

where c_p^0 is the heat capacity in the ideal gas state.

Example. Calculation of heat capacity of liquid nitrobenzene, $C_6H_5NO_2$, at 303.2 K using the method of Missenard [142]

M = 123.112 g/mol

Calculated:

	$I_{C_6H_5}$	I_{NO_2}
298 K	117.2	65.7
323 K	123.4	66.9
303 K	118.4	65.9

Table 9. Selected methods for the prediction of heat capacities of liquids

Method	Range of application		Information required	Error, %	Type of method
	Substances	Temperature			
Bondi – Rowlinson 1966 [144]	not recommended for polar substances	$0.4 < T_r < 0.96$	c_p^0, T_c, ω	5	corresponding states
Chueh, Swanson 1973 [141]		293 K		< 3	group contribution
Missenard 1965 [142]	no double bonds	$248 < T < 373$ K, but $T_r < 0.75$		5	group contribution
Ruzicka, Domalski 1993 [143]		$T_m < T < T_b$		2 – 6	group contribution

Substitution into Equation (3.50): c_p^l =184.3 J mol^{-1} K^{-1}
Experimental: c_p^l = 177.4 J mol^{-1} K^{-1} [121]
Relative deviation: 3.9 %

3.2.3. Heat Capacities of Liquid Mixtures

The heat capacity of liquid mixtures can usually be calculated as the arithmetic mean value of the individual components because deviations from observed values do not exceed 5 % if the components are similar in structure and do not exhibit hydrogen bonding:

$$c_{pm}^l = \sum x_i c_{p_i}^l \tag{3.52}$$

A comprehensive study of the heat capacity behavior of liquid mixtures is given in [146].

4. Phase Equilibria

4.1. Pure Compounds

4.1.1. Vapor Pressure, Boiling Point, and Melting Point

Vapor Pressure Correlation Functions. Many correlation functions have been published which are related to the Clausius – Clapeyron equation:

$$\frac{dp^{lv}}{dT} = \frac{\Delta H^{lv}}{T\left(V^v - V^l\right)} \tag{4.1}$$

where ΔH^{lv} is the enthalpy of vaporization, and V^l and V^v are the molar volumes of the liquid and vapor phase, respectively. With $\Delta H^{lv} \approx$ const., $V^v = RT/p$, and $V^v >> V^l$ (see Section 1.3), integration yields

$$\ln\frac{p_1^{lv}}{p_2^{lv}} = \frac{\Delta H^{lv}}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right) \tag{4.2}$$

where R is the gas constant. This relation leads to the simple *August equation* with two adjustable constants

$$\ln p^{lv} = A + \frac{B}{T} \tag{4.3}$$

Integrating Equation (4.1) using a linear function in temperature for the enthalpy of vaporization gives the *Rankine – Kirchhoff equation*

$$\log p^{lv} = A + \frac{B}{T} + C\log T \tag{4.4}$$

Figure 6 shows a typical vapor pressure versus temperature curve exhibiting an inflection point. Two-parameter equations such as Equation (4.3) are only able to represent a linear relationship between $\ln p^{lv}$ and $1/T$. Three-parameter equations, such as the frequently used *Antoine equation*,

$$\log p^{lv} = A - \frac{B}{T+C} \tag{4.5}$$

can also describe positive or negative curvatures. The description of the vapor pressure of a pure component, within or close to the experimental uncertainty over a wide range of temperature, is possible using the so-called *Frost – Kalkwarf equation*:

$$\ln p^{lv} = A + \frac{B}{T} + C\ln T + \frac{Dp^{lv}}{T^2} \tag{4.6}$$

which is implicit in pressure. This relation, along with a number of other correlation forms (e.g., Wagner, Riedel, and Riedel – Plank – Miller equations), achieves the best performance for practical purposes [147] (not considering polynomial functions such as the Chebyshev polynomial series proposed by AMBROSE [148]). Experimental vapor pressure data are given in [149].

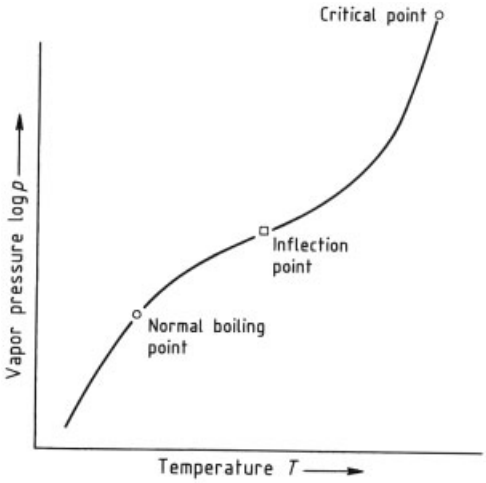

Figure 6. Vapor pressure curve of a pure liquid

Estimation Methods. Methods for estimating the vapor pressures of pure substances are of low accuracy compared to the quality of measurement. For several homologous series (e.g., straight-chain alkanes) p^{lv} can, however, be predicted with considerable accuracy.

For reasonable predictions of vapor pressure over a wider range of temperature the normal boiling point must be known from experiments. Equation (4.3) provides a rough estimate if p^{lv} data for two temperatures, e.g., the boiling point T_{b} and the critical point (p_{c} and T_{c}), are known. From critical temperatures and pressures (which can be estimated with the methods discussed in Chap. 2), generalized functions such as the Lee – Kesler method [150] or the Gomez-Nieto – Thodos equation [151] yield vapor pressures with uncertainties of a few percent. The latter method is based upon the following equation:

$$\ln p_{\mathrm{r}}^{\mathrm{lv}} = \beta\left(T_{\mathrm{r}}^{-m} - 1\right) + \delta\left(T_{\mathrm{r}}^{7} - 1\right) \tag{4.7}$$

with relations for the parameters β, m, and δ as follows:

1) Nonpolar organic and inorganic substances

$$\beta = -4.267 - \frac{221.79}{h^{2.5}\exp\left(0.0384h^{2.5}\right)} + \frac{3.8126}{\exp\left(2272.44h^{-3}\right)} + H \tag{4.8}$$

where $H = 0$ with exceptions for H_2 ($H = 0.19904$), He ($H = 0.41815$), and Ne ($H = 0.02319$).

$$m = 0.78425\exp\left(0.089315h\right) - 8.5217\exp\left(-0.74826h\right) \tag{4.9}$$

$$\delta = \frac{-\ln\left(p_{\mathrm{c}}/1.01325\right) + \left(1 - T_{\mathrm{br}}^{-m}\right)\beta}{T_{\mathrm{br}}^{7} - 1} \tag{4.10}$$

h in Equations (4.8) and (4.9) is defined as

$$h = \frac{\ln\left(p_{\mathrm{c}}/1.01325\right)}{\left(1/T_{\mathrm{br}}\right) - 1} \tag{4.11}$$

2) Polar compounds (except alcohols and water)

$$\delta = 0.08594\exp\left(7.462 \times 10^{-4} T_{\mathrm{c}}\right) \tag{4.12}$$

$$m = 0.466 T_{\mathrm{c}}^{0.166} \tag{4.13}$$

$$\beta = \frac{\delta\left(1 - T_{\mathrm{br}}^{7}\right) + \ln\left(p_{\mathrm{c}}/1.01325\right)}{\left(1 - T_{\mathrm{br}}^{-m}\right)} \tag{4.14}$$

3) Alcohols and water

$$\delta = \frac{2.464}{M\exp\left(9.8 \times 10^{-6} M T_{\mathrm{c}}\right)} \tag{4.15}$$

$$m = 0.0052 M^{0.29} T_{\mathrm{c}}^{0.72} \tag{4.16}$$

β from Equation (4.14)

Example. Calculation of vapor pressure of benzene, C_6H_6, with the Gomez method
$T_b = 353.2$ K, $T_c = 562.2$ K, $p_c = 4.89$ MPa [152]
Substituting into Equations (4.8) – (4.11):
$h = 6.5513, m = 1.3446, \delta = 0.1530, \beta = -4.2955$
Experimental vapor pressure from [153]

T, K	p^{lv}, kPa		Rel. deviation,
	exp.	cal.	cal − exp, %
288	7.784	8.01	3.0
330	46.63	46.92	0.6
400	352.4	351.7	− 0.2

Vapor pressures of pure liquids can also be predicted with the aid of a cubic equation of state via an iterative calculation. The method offers the advantage that other properties such as densities at saturation and heats of vaporization can also be obtained.

Boiling Point and Melting Point. Several methods have been developed for the prediction of the normal boiling point T_b, i.e., the boiling temperature at a pressure of 101.325 kPa. Some are discussed in [147]. A convenient method for organic compounds proposed by Ogata and Tsuchida [154] is based on the simple equation

$$T_b = py + q + 273.15 \ \text{[K]} \tag{4.17}$$

The parameter y is related to the alkyl groups, and p and q are characteristic constants for other functional groups. This method is limited to compounds with one functional group besides the alkyl group. The authors report an accuracy of 2 to 3 K for 91 % of more than 600 substances investigated.

For the estimation of both boiling and melting temperatures of pure compounds Joback [155], [156] developed an incremental method in which he employs the same structural groups as in his method for critical data estimation:
normal boiling point

$$T_b = 198.0 + \sum I_{T_b} \ \text{[K]} \tag{4.18}$$

melting point

$$T_m = 122.0 + \sum I_{T_m} \ \text{[K]} \tag{4.19}$$

Increments I_{T_b} and I_{T_m} are given in Table 3.

Boiling points of 438 organic compounds calculated with this method showed an average error of 12.9 K [152]. This appears to be rather large, but considering that no experimental information is required, the T_b data are acceptable as approximate values (cf. example below). This is not, however, true for the estimation of melting points. Here a test of Joback's method on 388 organic compounds yielded an average error of 23 K; moreover, deviations of more than 100 K may occur, e.g., for benzene 171 K (calc) and 279 K (exp), for cyclohexane 168.5 K (calc) and 280 K (exp).

The reason for these discrepancies is the large variety of intermolecular structures in the crystalline solid state. In fact, almost all compounds form an individual crystalline solid state (and often more than one such state) as a separate phase. This means that specific information is required for describing solid phases whereas more general relations are sufficient for the treatment of liquid or gas phases. Consequently for reliable prediction of the temperature of solid – liquid phase changes specific information for the solid concerned is needed.

Example. Calculation of normal boiling point of ethyl acrylate, $CH_2{=}CHCOOCH_2CH_3$. Experimental: $T_b = 373.0$ K [152]

1) Ogata – Tsuchida method (Eq. 4.17)

	y	p	q
$CH_2{=}CH{-}COOR$		0.918	29.0
CH_3CH_2	77.1		

$T_b = 372.9$ K
Deviation: − 0.07 K

2) Joback's method (Eq. 4.18)

$-CH_3$	$>CH_2$	$=CH-$	$=CH_2$	$-COO-$
23.58	22.88	24.96	18.18	81.10

$T_b = 368.7$ K
Deviation: − 4.3 K

A new and more highly refined incremental method for predicting T_b and T_m has been developed by Constantinou and Gani [157] (see also Section 2.4) with average errors of 5.4 K for T_b and 14 K for T_m.

4.1.2. Enthalpy of Vaporization

Determination from Vapor Pressure Data. As has been shown in Section 1.3 the vaporiza-

tion enthalpy of a pure liquid can be determined from vapor pressure data with the rigorous equation of Clausius and Clapeyron (Eq. 1.1). Its approximate version (Eq. 1.2) yields

$$\Delta H^{\mathrm{lv}} = RT^2 \frac{\mathrm{d}\ln p^{\mathrm{lv}}}{\mathrm{d}T} \tag{4.20}$$

With the Antoine Equation (4.5) for the differential quotient $\mathrm{d}\ln p^{\mathrm{lv}}/\mathrm{d}T$ the following relation is obtained:

$$\Delta H^{\mathrm{lv}} = RT \frac{B}{(T+C)^2} \tag{4.21}$$

Equation (4.21) can be readily used to calculate ΔH^{lv} with the aid of tabulated Antoine constants [149]. For moderate to high pressures ($p^{\mathrm{lv}} > 0.1$ MPa) the simplifying assumptions leading to Equation (1.2) are not justified. This means that the molar volumes of the liquid and the vapor must be taken into account. In Equation (1.4) this has been done only for the vapor phase.

A convenient way to consider both vapor and liquid molar volume is described by Haggenmacher [158]. He employs the difference in the compressibility factors of the vapor and the liquid:

$$\Delta z^{\mathrm{lv}} = z^{\mathrm{v}} - z^{\mathrm{l}} = \frac{p}{RT}\left(V^{\mathrm{v}} - V^{\mathrm{l}}\right) \tag{4.22}$$

which can be approximated by the relation

$$\Delta z^{\mathrm{lv}} = \left(1 - \frac{p_{\mathrm{r}}^{\mathrm{lv}}}{T_{\mathrm{r}}^3}\right)^{1/2} \tag{4.23}$$

Use of Equations (4.22) and (4.23) allows determination of $V^{\mathrm{v}} - V^{\mathrm{l}}$. Introduction of this term into Equation (4.1) improves the calculation of the enthalpy of vaporization except in the region close to the critical point. Correction factors for acetone and trichloromethane are shown in Figure 7.

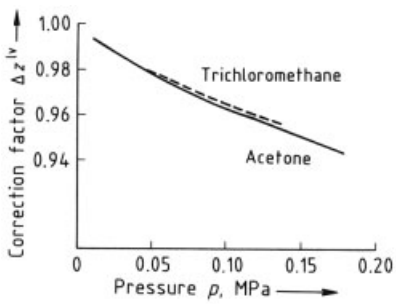

Figure 7. Influence of the Haggenmacher correction on the prediction of the heats of vaporization

Other Methods. Vapor pressure data are often unavailable or of insufficient accuracy for the application of the Clausius–Clapeyron equation. Other estimation methods then have to be used for the estimation of the vaporization enthalpies. Two of these are described in Section 1.3 (Eqs. 1.5 and 1.6). Vetere [159] suggested the following set of correlations (ΔH^{lv}, J/mol; M, g/mol; T_{b}, K) [147]:

1) Hydrocarbons

$$\frac{\Delta H_{T_{\mathrm{b}}}^{\mathrm{lv}}}{RT_{\mathrm{b}}} = 7.0054 + 1.647\log M + \frac{0.7806\left[T_{\mathrm{b}} - (263M)^{0.581}\right]^{-1.037}}{M} \tag{4.24}$$

2) Alcohols, organic acids, methylamine

$$\frac{\Delta H_{T_{\mathrm{b}}}^{\mathrm{lv}}}{RT_{\mathrm{b}}} = 9.7643 + 1.5748\log T_{\mathrm{b}} - \frac{3.1018T_{\mathrm{b}}}{M} + \frac{0.0176375T_{\mathrm{b}}^2}{M} - \frac{2.57131\times10^{-5}T_{\mathrm{b}}^3}{M} \tag{4.25}$$

3) Esters

$$\frac{\Delta H_{T_{\mathrm{b}}}^{\mathrm{lv}}}{RT_{\mathrm{b}}} = 5.501 + 1.901\log T_{\mathrm{b}} + \frac{0.04852T_{\mathrm{b}}}{M} + \frac{0.53689\times10^{-3}T_{\mathrm{b}}^2}{M} - \frac{0.6977\times10^{-6}T_{\mathrm{b}}^3}{M} \tag{4.26}$$

4) Other polar compounds

$$\frac{\Delta H_{T_{\mathrm{b}}}^{\mathrm{lv}}}{RT_{\mathrm{b}}} = 5.3405 + 1.8453\log T_{\mathrm{b}} + \frac{0.047109T_{\mathrm{b}}}{M} + \frac{0.52125\times10^{-3}T_{\mathrm{b}}^2}{M} - \frac{0.67738\times10^{-6}T_{\mathrm{b}}^3}{M} \tag{4.27}$$

Corresponding states methods (e.g., Pitzer's correlation) should also be mentioned. These methods are discussed in [160].

Table 10. Selected methods for the prediction of the enthalpy of vaporization ΔH^{lv} of pure substances

Method	Range of application		Information required *	Error, %
	Substances	Temperature		
Riedel 1954 [162]		$T = T_b$	T_b, T_c, p_c	2
Chen 1965 [163]		$T = T_b$	T_b, T_c, p_c	2
Vetere 1973 [159]		$T = T_b$	$p^{lv}(T), T_c, p_c$	2 **
Antoine equation, (Haggenmacher 1946 [158])		$0.4 < T_r < 0.8$	$p^{lv}(T), T_c, p_c, \varrho^l$	for $T_r > 0.5$ with inclusion of Eq. (4.23)
Wagner 1973 [164]		$T^{vls} < T < 0.95\,T_c$	$p^{lv}(T), (T_c, p_c)$	1 **
Kobayashi 1984 [165]	nonpolar, weakly polar	$T_r < 0.98$	T_c, ω	1.5
Watson 1943 [166]	nonpolar, weakly polar	$0.5 < T_r < 0.75$	$T_c, \Delta H^{lv}_{ref}$	$0.05\,(T - T^{ref})$
Lee – Kesler – Pitzer 1975 [150]	nonpolar, weakly polar	$0.5 < T_r < 0.98$	T_c, p_c, ω	1.5

* In addition to the corresponding parameters and increments.
** Depends on the accuracy of vapor pressure data.

A summary of estimation methods for ΔH^{lv} appears in Table 10. Figure 8 shows a comparison of different estimation methods. Experimental data can be found in [161].

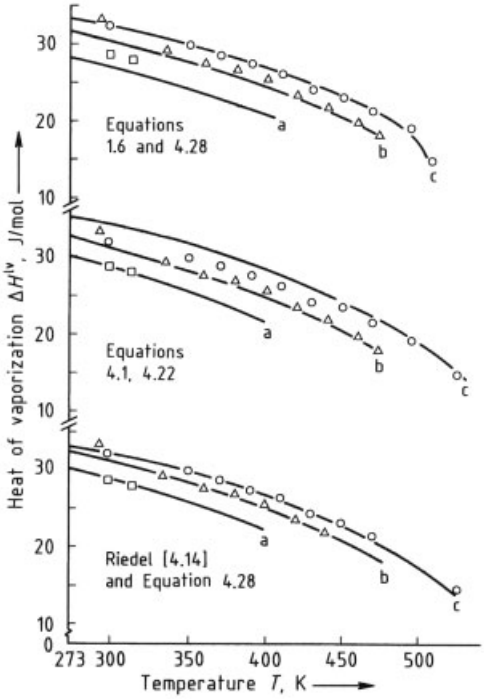

Figure 8. Comparison of estimation methods for the prediction of heat of vaporization of chloromethanes
a) Dichloromethane; b) Trichloromethane; c) Tetrachloromethane
Open symbols denote experimental data, lines denote predicted values

Temperature Dependence. The temperature dependence of the vapor pressure can be represented satisfactorily using Watson's equation [166]:

$$\Delta H^{lv}_T = \Delta H^{lv}_{T_b} \left[\frac{1 - T_r}{1 - T_{br}} \right]^n \tag{4.28}$$

A value of 0.38 is recommended for the exponent n, but if other values of ΔH^{lv} are available this equation can be used efficiently for the correlation of the temperature dependence of ΔH^{lv} by adjusting the exponent n.

4.2. Mixtures

Calculation of phase equilibria is based on the thermodynamic conditions for equilibrium, i.e., equality of temperature T, pressure p, and fugacities f_i of each component i in the coexisting phases. This means that for the mathematical description of phase equilibria, expressions representing fugacities as a function of composition are required. Two types of formulations are used which differ in the state of reference. With the ideal gas as state of reference, the fugacity f_i of component i in a mixture is given by

$$f_i = x_i \varphi_i P \tag{4.29}$$

where x_i is the concentration of i (usually as mole fraction) and φ_i the fugacity coefficient (with $\varphi_i = 1$ for ideal gases).

With an arbitrary state of reference, which is very often the pure component i at its vapor pressure, the following relation is defined:

$$f_i = x_i \gamma_i f_i^0 \tag{4.30}$$

with γ_i as activity coefficient of i in the mixture and f_i^0 as fugacity of i at the state of reference.

4.2.1. Vapor – Liquid Equilibria with Liquid-Phase Activity Coefficients

4.2.1.1. Basic Considerations

For the correlation and prediction of vapor – liquid equilibria at low to moderate pressure (< ca. 1.0 MPa) Equation (4.30) is usually applied to the liquid phase. The vapor phase is either considered to be ideal or a correction for real gas behavior is applied (e.g., by estimated second virial coefficients). The reference state in the liquid phase for each component i in the mixture is the pure liquid i with a vapor pressure p_{0i}^{lv} at temperature T. With this reference state Equation (4.30) for component i in a liquid mixture is given by

$$f_i^{\mathrm{l}} = x_i \gamma_i p_{0i}^{\mathrm{lv}} \varphi_{0i} \exp\left(\frac{1}{RT}\int_{p_{0i}}^{p} V_{0i}^{\mathrm{l}}\,\mathrm{d}p\right) \tag{4.31}$$

which includes two corrections, i.e., the fugacity coefficient φ_{0i} for the pure liquid at pressure p_{0i} and the exponential term called the Poynting factor which corrects for the difference of fugacity between p_{0i}^{lv} and system pressure p. If the liquid volume V_{0i}^{l} is assumed to be constant (independent of pressure), the Poynting factor Π_i is simply given by

$$\Pi_i = \exp\left(\frac{V_{0i}^{\mathrm{l}}\left(p - p_{0i}^{\mathrm{lv}}\right)}{RT}\right) \tag{4.32}$$

Equality of fugacities in the vapor and liquid phases yields with Equations (4.29) and (4.31) the following general relation for vapor – liquid equilibria:

$$y_i \varphi_i p = x_i \gamma_i p_{0i}^{\mathrm{lv}} \varphi_{0i} \Pi_i \tag{4.33}$$

with y_i and x_i as the mole fractions of i in the vapor and liquid, respectively.

At pressures up to several hundred kilopascal (several bar) the fugacity coefficients φ_i and φ_{0i} and the Poynting factor Π_i are usually not significantly different ($< 2.3 \times 10^{-3}$) from unity. This means that a simplified version of Equation (4.33) can be used for the majority of vapor – liquid equilibria without much loss in accuracy:

$$y_i p = x_i \gamma_i p_{0i}^{\mathrm{lv}} \tag{4.34}$$

With Equation (4.34) the vapor phase is considered ideal and all nonideality effects are attributed to the liquid-phase activity coefficient γ_i. For $\gamma_i = 1$, Equation (4.34) transforms into Raoult's law for the partial pressure p_i in ideal mixtures:

$$p_i = y_i p = x_i p_{0i}^{\mathrm{lv}} \tag{4.35}$$

The activity coefficients γ_i in liquid mixtures are directly related to an important thermodynamic quantity characterizing nonideal mixing behavior, i.e., the molar excess Gibbs energy of mixing ΔG^{E} which is defined as the difference in the molar Gibbs energy of mixing between real and ideal mixtures:

$$\Delta G^{\mathrm{E}} = \Delta G - \Delta G^{\mathrm{id}} \tag{4.36}$$

This quantity is related to the activity coefficients γ_i of all components i in a mixture by the following equation:

$$\Delta G^{\mathrm{E}} = RT \sum x_i \ln \gamma_i \tag{4.37}$$

Conversely, the activity coefficients γ_i in a mixture can be obtained from ΔG^{E} when this quantity is known as a function of composition (i.e., of all x_i) according to:

$$RT \ln \gamma_i = \left(\frac{\partial n \cdot \Delta G^{\mathrm{E}}}{\partial n_i}\right)_{T,p,n_j} \tag{4.38}$$

where n is the total number of moles and n_i the number of moles of component i.

4.2.1.2. Data Correlation: Binary and Multicomponent Mixtures of Nonelectrolytes

Numerous correlation methods have been published. They include flexible polynomials (e.g., Margules, van Laar, Redlich – Kister series) and other more sophisticated semitheoretical models, especially those based on the concept of local composition [Wilson, Non-Random

Table 11. Model equations for calculating the molar excess Gibbs energy of mixing ΔG^{E} in liquid mixtures

Model	Parameter	Equation
Margules-3 (binary systems)	A B	$\ln\gamma_1 = x_2^2[A + 2(B - A)x_1]$ $\ln\gamma_2 = x_1^2[B + 2(A - B)x_2]$
Wilson [167]	$\Lambda_{ij}, \Lambda_{ji}$ ($\Lambda_{ii} = \Lambda_{jj} = 1$)	$\frac{\Delta G^{E}}{RT} = -\sum_i^r x_i \ln\left(\sum_i^r x_j \Lambda_{ij}\right)$ $ln\gamma_i = -\ln\left(\sum_i^r x_j \Lambda_{ij}\right) + 1 - \sum_k^r \frac{x_k \Lambda_{ki}}{\sum_j^r x_j \Lambda_{kj}}$ with $\Lambda_{ij} = \frac{V_{0j}^{l}}{V_{0i}^{l}} \exp\left(-\frac{\lambda_{ij} - \lambda_{ii}}{RT}\right)$
NRTL [168]	$\tau_{ij}, \tau_{ji}, \alpha_{ij} = \alpha_{ji}$	$\frac{\Delta G^{E}}{RT} = -\sum_i^r x_i \frac{\sum_j^r \tau_{ji} G_{ji} x_j}{\sum_k^r G_{ki} x_k}$ $\ln\gamma_i = \frac{\sum_i^r \tau_{ji} G_{ji} x_j}{\sum_k^r G_{ki} x_k} + \sum_j^r \frac{x_j G_{ij}}{\sum_k^r G_{kj} x_k}\left(\tau_{ij} - \frac{\sum_k^r x_k \tau_{kj} G_{kj}}{\sum_k^r G_{kj} x_k}\right)$ with $\tau_{ij} = \frac{g - giiij}{RT}$ and $G_{ij} = \exp\{-\alpha_{ij}\tau_{ij}\}$
UNIQUAC [169]	τ_{ij}, τ_{ji}	$\frac{\Delta G^{E}}{RT} = \sum_i x_i \ln S_i + \frac{z}{2}\sum_i q_i x_i \ln\frac{A_i}{S_i} - \sum_i q_i x_i \ln \sum_i x_j A_j \tau_{ji}$ $\ln\gamma_i = \ln\gamma_i^{C} + \ln\gamma_i^{R}$ $\ln\gamma_i^{C} = 1 - S_i + \ln S_i - \frac{z}{2} q_i \left(1 + \frac{S_i}{A_i} + \ln\frac{S_i}{A_i}\right)$ $\ln\gamma_i^{R} = q_i\left(1 - \ln\frac{\sum_j q_j x_j \tau_{ji}}{\sum_j q_j x_j} - \sum_j \frac{q_j x_j \tau_{ij}}{\sum_k q_k x_k \tau_{kj}}\right)$ with $\tau_{ij} = \exp\frac{-\Delta u_{ij}}{T}$, $\Delta u_{ii} = \Delta u_{jj} = 0$, $z = 10$ S_i, A_i, see Equations (4.42) and (4.43)

Two-Liquid (NRTL), and UNIversal QUAsi-Chemical (UNIQUAC) equations].

Relations for some of these model equations are given in Table 11. If reliable experimental data are available, equations with two parameters per binary system are usually suitable for correlating and extrapolating available data with respect to composition and, to a limited degree, to temperature. Numerous papers have been published comparing different ΔG^{E} models. But no essential advantages for any of these correlation methods could be found with respect to their ability to fit data.

Another important aspect has to be considered. The majority of published vapor – liquid equilibrium data are concerned with binary systems, only a few with ternary systems, and very few with systems of four and more components. In practical chemical engineering problems, however, systems with at least five or more components often have to be treated. Here the models based on the concept of local composition show the advantage that only binary parameters are required for the representation of multicomponent systems. The local composition concept was introduced by Wilson [167]; the essential feature of this idea is that it considers only binary interactions between neighboring molecules in the liquid mixture.

Other models using the local composition concept are the NRTL [168] and UNIQUAC [169] equations. In contrast to the Wilson equation these models are able to describe the coexistence of two liquid phases, i.e., liquid – liquid phase splitting.

4.2.1.3. Prediction of Equilibrium Data

Solution-of-Groups Concept. Measurement of vapor – liquid equilibria is difficult, time-consuming, and costly. Thus reliable and widely applicable estimation methods are of utmost interest. Numerous attempts to derive methods for the prediction of the properties of nonelectrolyte liquid mixtures using only pure component information (e.g., dipole moment or parachor) have failed. In contrast, methods based on structural groups have turned out to be quite successful. The starting point for these concepts is the assumption that a mixture is com-

posed of structural groups and not of molecules. Consequently interactions between structural groups are considered instead of interactions between molecules. The large variety of molecular mixtures resulting from the huge number of compounds is reduced to a limited number of combinations of structural groups (e.g., $-CH_3$, $>CH_2$, $-OH$, $-CHO$, $>C{=}O$). This "solution-of-groups" concept was originally introduced to phase equilibrium calculations by REDLICH et al. [167] and WILSON and DEAL [171]. The first method of practical importance based on this concept was the Analytical Solution Of Groups (ASOG) method of DERR and DEAL [172].

ASOG Method. This method employs an athermal Flory – Huggins term to consider size effects and the Wilson equation for the interactions of the structural groups. Therefore the activity coefficient γ_i is split into two parts:

$$\ln\gamma_i = \ln\gamma_i^{\mathrm{C}} + \ln\gamma_i^{\mathrm{R}} \tag{4.39}$$

where the combinatorial part $\ln\gamma_i^{\mathrm{C}}$ considers size effects and the residual part $\ln\gamma_i^{\mathrm{R}}$ the group interactions. Equation (4.39) is the basic relationship of the solution-of-groups concept. The residual part $\ln\gamma_i^{\mathrm{R}}$ is related to the pure liquid *i* according to

$$\ln\gamma_i^{\mathrm{R}} = \sum_k \nu_k^i \left[\ln\Gamma_k - \ln\Gamma_k^{(i)}\right] \tag{4.40}$$

where ν_k^{i} is the number of groups of type *k* in component *i*; Γ_k and $\Gamma_k^{(\mathrm{i})}$ are the corresponding group activity coefficients in the mixture and in the pure liquid, respectively. They are calculated from interaction parameters that have been determined from reliable vapor – liquid equilibrium data. The latest version of ASOG gives interaction parameters for 43 structural groups [173].

UNIFAC Method. More recently FREDENSLUND et al. [174] derived the UNIversal Functional group Activity Coefficients (UNIFAC) method. This group contribution model applies the UNIQUAC equation [169] to the solution-of-groups concept (Eq. 4.39). Since this model is most widely accepted and applied, it is described here in more detail. The combinatorial part contains pure component parameters only and is the Staverman expression taken from the UNIQUAC model:

$$\ln\gamma_i^{\mathrm{C}} = 1 - S_i + \ln S_i + \frac{z}{2} q_i \left(1 - \frac{S_i}{A_i} + \ln\frac{S_i}{A_i}\right) \tag{4.41}$$

where x_i is the mole fraction. The volume fraction S_i and surface fraction A_i are defined as follows:

$$S_i = \frac{r_i}{\sum_j r_j x_j} \tag{4.42}$$

$$A_i = \frac{q_i}{\sum_j q_j x_j} \tag{4.43}$$

where the parameters r_i and q_i represent the relative volume and relative surface area of the component *i; z* is the coordination number and is usually fixed as $z = 10$. These size parameters have been tabulated for many groups [175] or can be calculated [176].

The residual part considers group – group interactions via the group activity coefficient Γ_i following the UNIQUAC model for the main groups *m* and *n*:

$$\ln\Gamma_k = Q_k \left[1 - \ln\left(\sum \Theta_m \Psi_{mk}\right) - \sum_m \frac{\Theta_m \Psi_{km}}{\sum_n \Theta_n \Psi_{nm}}\right] \tag{4.44}$$

with the area fraction Θ_m of group *m* defined by

$$\Theta_m = \frac{Q_m X_m}{\sum Q_n X_n} \tag{4.45}$$

where X_m is the group mole fraction and $Q_m <$ the group surface area. The group energy term Ψ_{nm} for the interaction between the groups *m* and *n* is defined by

$$\Psi_{nm} = \exp\left(\frac{-a_{mn}}{T}\right) \tag{4.46}$$

At present, interaction parameters a_{mn} are known for 50 main groups, some of which comprise subgroups (e.g., main group CH_2 with subgroups C, CH, CH_2, and CH_3). Several smaller molecules (e.g., water, methanol) and some molecules with special structures (e.g., pyridine, *N*-methylpyrrolidone) have been defined as main groups.

The interaction parameters have been determined from reliable, thermodynamically consistent experimental vapor – liquid equilibrium

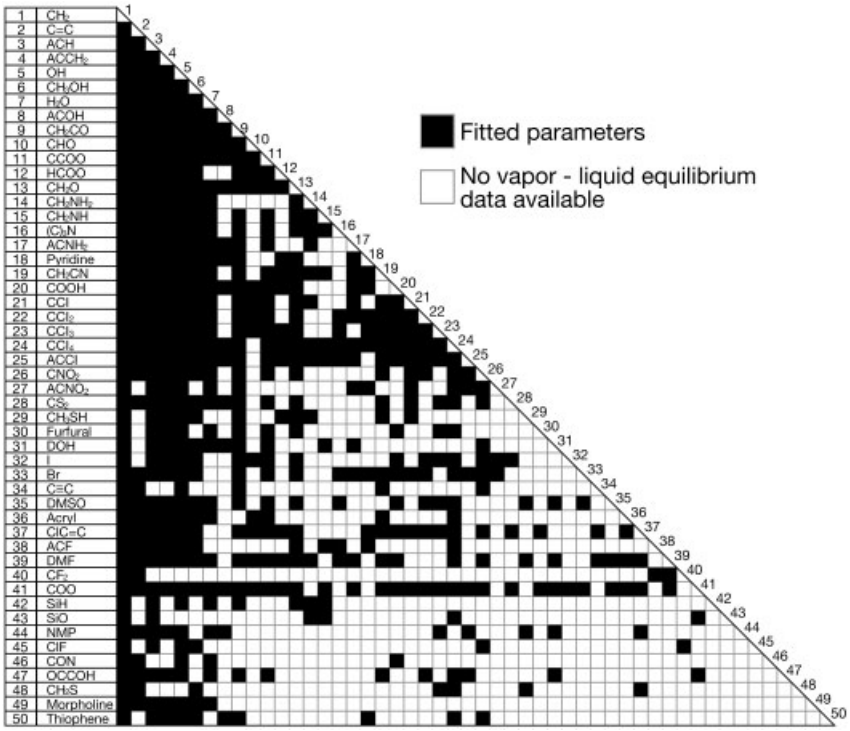

Figure 9. Status (1991) of the UNIFAC parameter matrix [178]
Abbreviations: A = aromatic; $(C)_3N$ = tertiary amine; DMF = dimethylformamide; DMSO = dimethyl sulfoxide; DOH = 1,2-ethanediol; NMP = *N*-methylpyrrolidone

data [175], [177], [178]. The present state can be seen in Figure 9.

The predictive ability of the UNIFAC method is quite remarkable. Provided accurate pure-component vapor pressures are known, errors in vapor phase composition are usually ca. 1 mol%. Typical results of UNIFAC predictions are displayed in Figure 10, which demonstrates the predictive power of the UNIFAC method. In cases where only little or unreliable primary data are available for the determination of group interaction parameters (e.g., for combinations with aldehydes) serious deviations may be expected.

This original UNIFAC method has additional limitations:

1) Limited temperature dependence leading to unsatisfactory prediction of heat of mixing data
2) In many cases isomers cannot be distinguished from one another
3) Neighboring effects (i.e., interaction between "strong" functional groups within the same molecule) are not considered.

Much effort has been and still is devoted to overcoming these shortcomings, e.g., [179]. Gmehling et al. [180], [181] have included experimental data for both heats of mixing and activity coefficients at infinite dilution into their data base from which they determined temperature-dependent group interaction parameters. They showed that this modified UNIFAC method can be used to predict temperature-dependent activity coefficients at both bulk composition and high dilution and also mixing enthalpies with satisfactory accuracy.

A different approach in applying group contributions for the prediction of activity coefficients was proposed by Kehaian et al. [182]. Their method is based on Barker's lattice theory; interaction parameters have only been published for a limited number of group combinations.

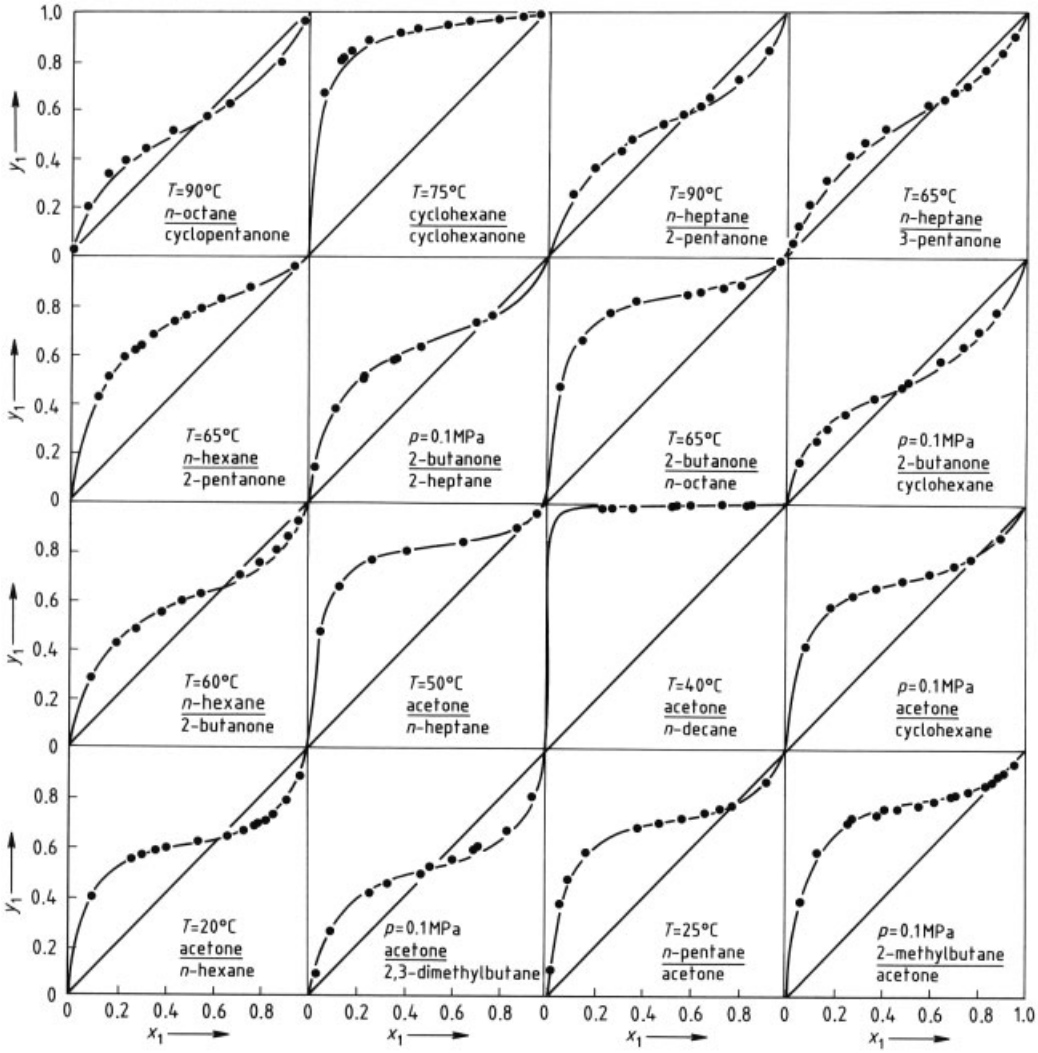

Figure 10. Results of UNIFAC predictions (curves) compared with experimental data (dots) for 16 ketone – alkane mixtures x_1 = mole fraction of underlined component in the liquid phase; y = mole fraction of underlined component in the vapor phase Only two interaction parameters ($a_{CH_2,\,CH_2CO}$ = 476.4 K, $a_{CH_2CO,\,CH_2}$ = 26.76 K) have been used for all diagrams [177].

For practical purposes the classical UNIFAC method with its large interaction parameter matrix provides a useful tool for the prediction of vapor – liquid equilibria of unknown systems and also for the selection of solvents in extractive and azeotropic distillation [183]. The matrix of UNIFAC interaction parameters is given in [178].

Extensions of the UNIFAC Method for Polymer Solutions and for Gas Solubilities. In order to apply the UNIFAC method to polymer solutions Oishi and Prausnitz proposed the inclusion of a free volume term ln γ_i^{FV} in Equation (4.39) [184]:

$$\ln\gamma_i = \ln\gamma_i^{C} + \ln\gamma_i^{R} + \ln\gamma_i^{FV} \tag{4.47}$$

For gas – liquid equilibria the activity coefficient approach (Eq. 4.30) employing the pure liquid as reference state cannot apparently be applied when the system temperature exceeds the critical temperature of one of the components, as is the case with solubilities of so-called permanent gases (e.g., nitrogen, oxygen, methane,

ethylene) in liquids at ambient temperature. A hypothetical liquid state of the gaseous component i may then be defined by extrapolating the standard fugacity f_i^0 beyond the critical temperature. Using this hypothetical liquid state NOCON et al. [185], [186] obtained satisfactory predictions of solubilities of nitrogen, oxygen, carbon dioxide, methane, and ethylene in hydrocarbons and alcohols with a UNIFAC method based on Equation (4.47) (i.e., by including a free volume term).

4.2.2. Vapor – Liquid and Gas – Liquid Equilibria with Equations of State

4.2.2.1. Basic Considerations

Taking the ideal gas as standard state for the fugacities (Eq. 4.29), the equilibrium condition of equal fugacities in the vapor and liquid phases yields for component i:

$$y_i \varphi_i^{\mathrm{v}} = x_i \varphi_i^{\mathrm{l}} \tag{4.48}$$

With this relation the same thermodynamic model can be used for both phases. The fugacity coefficient can be calculated from the rigorous relation

$$\ln \varphi_i = \frac{1}{RT} \int_0^p \left[\left(\frac{\partial V}{\partial n_i} \right)_{T,p,n_{j \neq i}} - \frac{RT}{p} \right] \mathrm{d}p \tag{4.49}$$

employing an equation of state for V. The advantage of this approach is that no standard-state fugacity is necessary. This is particularly useful when the mixture consists of both sub- and supercritical components. A further advantage is the possibility of calculating other mixture properties such as densities, enthalpies, and entropies of the two phases. Numerous equations of state have been published but accurate prediction of phase equilibria of mixtures is not yet possible from pure-component data only. Hence, binary interaction parameters (see Section 3.1.3.1) must be obtained from experimental data for mixtures.

Modern statistical mechanics have achieved remarkable progress in the last two decades [187], [188]. Nevertheless, computation time is still considerable. Adequate force models are available only for rather simple molecules. As a consequence, the molecular methods have not yet become established for engineering calculations.

4.2.2.2. Data Correlation

An extensive compilation of high-pressure vapor – liquid equilibria and gas – liquid equilibria data has been presented by KNAPP et al. [189]. They tested the capability of four common equations of state (Lee – Kesler – Plöcker, LKP; Benedict – Webb – Rubin – Starling, BWRS; Redlich – Kwong – Soave, RKS; Peng – Robinson, PR; see Section 3.1.1) in representing experimental data for a large number of binary systems. The calculations of KNAPP et al. show that there is no principal advantage of any of these equations. They yield satisfactory results when the components involved are not highly polar or do not differ too much in size.

More generally applicable are the so called ΔG^{E} mixing rules. HURON and VIDAL [190] proposed using the following definition of the excess Gibbs energy:

$$\Delta G^E = RT \left(\ln \varphi - \sum x_i \ln \varphi_{i0} \right) \tag{4.50}$$

where the fugacity coefficient φ accounts for the real behavior of the mixture:

$$\ln \varphi_i = \frac{1}{RT} \int_0^p \left[V - \frac{RT}{p} \right] \mathrm{d}p \tag{4.51}$$

The fugacity coefficients φ_{i0} of the pure components are given by Equation (4.49). Substitution of ΔG^{E} in Equation (4.50) by Equation (4.37) leads to an expression from which mixing rules for the parameter a of cubic equations of state (Section 3.1.1) can be derived. When infinite pressure ($p \to \infty$) is chosen as reference state, it can be assumed that the parameter b is equal to the molar volume V. With this assumption and the linear mixing rule for b_{m} (Eq. 3.39) the Huron – Vidal (HV) mixing rule for the RKS equation of state (Eq. 3.11) can be derived:

$$a_{\mathrm{m}} = b_{\mathrm{m}} RT \left[\sum x_i \cdot \frac{a_i}{b_i RT} - \frac{\sum x_i \ln \gamma_i}{0.6931} \right] \tag{4.52}$$

Other reference states may be chosen. For example with $p = 0$ the modified Huron – Vidal (MHV) mixing rules have been derived [191].

With a constant value for the ratio V/b the MHV1 mixing rule [192] for the RKS equation of state is obtained:

$$a_{\mathrm{m}} = b_{\mathrm{m}} RT \left[\sum x_i \cdot \frac{a_i}{b_i RT} + \frac{\sum x_i \ln \gamma_i + \sum x_i \ln \frac{b_{\mathrm{m}}}{b_i}}{A} \right] \tag{4.53}$$

Depending on the value of V/b different values for A will result. In the PSRK method (predictive Soave – Redlich – Kwong [193], Section 4.2.2.3) with V/b = 1.1 the value of A is -0.64663. Equation (4.53) allows any G^{E} model for the liquid phase to be incorporated (see Section 4.2.1.2) into the RKS equation of state.

Based on the ΔG^{E} mixing rules several other models have been proposed; they differ mainly in the reference pressure and in the V/b ratio. They are all well suited for the representation of vapor – liquid and gas – liquid equilibria, also at high pressures and for strongly polar systems. For the latter systems they should be preferred to equations of state combined with density-dependent mixing rules (e.g., [194]) or with association models (e.g., [195], [196]).

4.2.2.3. Prediction of Vapor-Liquid and Gas-Liquid Equilibria

Since any type of G^{E} model can be introduced into Equation (4.53), group contribution methods can be applied, too. The PSRK group contribution equation of state [193] combines the UNIFAC method (Section 4.2.1.3) with the RKS equation of state. Thus all group interaction parameters tabulated for UNIFAC can be used. Moreover new groups for gaseous components have been included: NH_3, CO_2, CH_4, O_2, Ar, N_2, H_2S, H_2, CO, SO_2, NO, N_2O, SF_6, He, Ne, Kr, Xe, HCl, HBr [193], [197].

Compared to other group contribution equations of state proposed for estimating vapor – liquid equilibria up to the critical region [198], [199], [200] the PSRK method can be applied over a wider range of temperature and pressure. In addition, many more groups and their parameters have been defined for the PSRK method. As an example of its predictive power, a comparison of predicted gas solubilities is given in Figure 11. It shows the Henry coefficients H_{12} (Eq. 4.54) of several hydrocarbon gases (1) in the liquid solvent (2) eicosane.

$$H_{12} = \lim_{x_i \to 0} \frac{f_1}{x_1} \tag{4.54}$$

where f_1 is the fugacity of gaseous component (1) and x_1 the mole fraction of gaseous component (1) in the liquid.

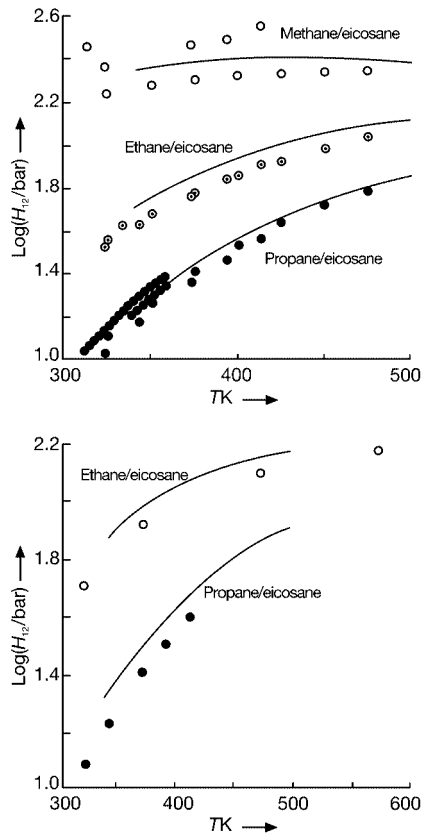

Figure 11. Results of PSRK predictions (lines) compared with experimental data (•) from the Dortmund Data Bank) for Henry coefficients of hydrocarbon gases in eicosane (n-$C_{20}H_{42}$)

4.2.3. Liquid – Liquid Equilibria

Phase splitting of a liquid mixture into two liquid phases occurs when a single liquid phase is thermodynamically unstable. For the computation of liquid – liquid equilibria, equality of fugacities for each component in the two phases is usually applied. Since this equilibrium condition is only necessary, but not sufficient, it should

be combined with an analysis of thermodynamic stability in order to find the true equilibrium. Using Equation (4.30) for the fugacities f_i' and f_i'' in phases I and II with the same reference state for f_i^0 yields the following relation:

$$x_i'\gamma_i' = x_i''\gamma_i'' \tag{4.55}$$

The products $x_i'\gamma_i'$ and $x'\prime_i\gamma'\prime_i$ are the so-called activities. Since Equation (4.55) holds for each component of a liquid – liquid system, it should be possible to predict liquid – liquid equilibria when the ΔG^{E} function (Eq. 4.37) and consequently the activity coefficients of the individual components in the multicomponent system are known (e.g., from vapor – liquid equilibria or from prediction methods developed for these phase equilibria, such as UNIFAC). This procedure entails several problems, however:

1) Whereas in vapor – liquid equilibrium calculations activity coefficients serve as a correction of Raoult's law (cf. Eqs. 4.34 and 4.35), they are the sole numerical input into calculations of liquid – liquid equilibria. Therefore, the accuracy of activity coefficients γ_i for modeling and predicting liquid – liquid equilibria has to be considerably higher than for vapor – liquid equilibria.
2) Systems with phase splitting often show rather high temperature dependence of activity coefficients, which is in general predicted less reliably.
3) The majority of vapor – liquid equilibrium data and the associated activity coefficient data have been determined at other, usually higher temperatures than are of interest for liquid – liquid equilibria.

Modifications of the UNIFAC method, either by creating a separate UNIFAC parameter matrix for liquid – liquid equilibrium prediction [201] or by including experimental liquid – liquid equilibria data into the determination of interaction parameters [181], [202], have achieved limited success.

The use of equations of state has been tested with promising results for the calculation of liquid – liquid equilibria, for example, extrapolation of liquid – liquid equilibria to higher pressures and representation of vapor – liquid – liquid equilibria [203].

Octanol – Water Distribution Coefficients. The distribution coefficient of toxic substances between octanol and water is an important property in toxicology and environmental science since it correlates with the distribution of substances between a lipid and an aqueous phase. For the estimation of octanol – water distribution coefficients, Hansch and Leo [204] developed a much used group contribution method. Another such method has been proposed by Wienke and Gmehling [205]. It is based on the UNIFAC method and has the advantage that other equilibrium constants relevant to the distribution of chemicals in the environment, i.e., Henry constants and water solubilities, can be predicted with the same parameters.

4.2.4. Solid – Liquid Equilibria

Solid – liquid equilibria may be roughly classified into systems without miscibility of the components in their solid states (eutectic systems) and systems that form mixed crystals. With nonelectrolytes the first type of behavior is much more frequent. For these systems the following relation between the solubility of pure solid *i* (expressed as liquid mole fraction x_i^{l}) and temperature *T* holds:

$$\ln\left(x_i^{\mathrm{l}}\gamma_i^{\mathrm{l}}\right) = -\frac{\Delta H_i^{\mathrm{sl}}}{RT}\left(1-\frac{T}{T_{\mathrm{m},0i}}\right) \tag{4.56}$$

where ΔH_i^{sl} is the enthalpy of melting and $T_{\mathrm{m},0i}$ the melting point of pure component *i* (terms of higher order have been neglected in Eq. 4.56). The liquid-phase activity coefficient γ_i^{l} can be estimated with group contribution methods such as UNIFAC [206], [207]. At high concentrations of *i* the equilibrium temperature *T* (i.e., the freezing point of a liquid solution or melt) can be predicted quite reliably without knowing the value of the liquid-phase activity coefficient γ_i^{l}, because γ_i^{l} generally approaches unity at these concentrations. With this assumption, rearrangement of Equation (4.56) (with some simplifications) yields the well-known Raoult equation for the lowering of the freezing point of solutions at low solute concentrations:

$$T_{\mathrm{m},i} - T = \frac{RT_{\mathrm{m},0i}}{\Delta H_i^{\mathrm{sl}}}\left(1-x_i\right) \tag{4.57}$$

5. Thermodynamic Data of Chemical Reactions

5.1. Definitions

Thermodynamic data of chemical reactions are highly important in process design for two reasons:

1) Molar reaction enthalpies (heats of reaction) ΔH_{r} are required in the design of chemical reactors, since the heat effect connected with the reaction must be known to control the reaction temperature.
2) The molar Gibbs energy ΔG_{r} (also called the free enthalpy) of a chemical reaction is directly related to the reaction equilibrium; knowledge of ΔG_{r} allows prediction of equilibrium concentrations.

5.1.1. Reaction Enthalpy

Reaction enthalpies can be determined from the enthalpies of formation ΔH_f of the compounds taking part in the reaction. Thus, for the reaction

$$\alpha\mathrm{A} + \beta\mathrm{B} \rightleftharpoons \gamma\mathrm{c} + \delta\mathrm{D} \tag{5.1}$$

the reaction enthalpy ΔH_{r} is given by

$$\Delta H_{\mathrm{r}} = \gamma\Delta H_{\mathrm{fC}} + \delta\Delta H_{\mathrm{fD}} - \alpha\Delta H_{\mathrm{fA}} - \beta\Delta H_{\mathrm{fB}} \tag{5.2}$$

or by the general form

$$\Delta H_{\mathrm{r}} = \sum \nu_i \Delta H_{\mathrm{f}i} \tag{5.3}$$

where ν_i is the stoichiometric coefficient of compound *i* (according to the rules of chemical thermodynamics, ν_i is positive for products and negative for educts) and $\Delta H_{\mathrm{f}i}$ is the enthalpy of formation of compound *i*.

For the calculation of reaction enthalpies the formation enthalpies of many compounds have been tabulated for a standard state of 1.01325 bar = 1 atm at 298.15 K [208–215]. They are designated as $\Delta H^0_{\mathrm{f298}}$ (with superscript 0 for standard pressure and subscript 298 for standard temperature, 298 K). Since 1982 0.1 MPa = 1 bar has been recommended as standard pressure, but in chemical thermodynamics and relevant data compilations 1.01325 bar is still in use. Generally the standard formation enthalpy $\Delta H^0_{\mathrm{f298}}$ of a compound is given for its most stable state at standard pressure (1.01325 bar at 298.15 K) . For the elements $\Delta H^0_{\mathrm{f298}}$ is zero by definition in their most stable state under standard conditions (for gases: the ideal gas state).

The standard reaction enthalpy $\Delta H^0_{\mathrm{r298}}$ for Equation (5.1) is thus given by

$$\Delta H^0_{\mathrm{r298}} = \gamma\Delta H^0_{\mathrm{f298,C}} + \delta\Delta H^0_{\mathrm{f298,D}} - \alpha\Delta H^0_{\mathrm{f298,A}} - \beta\Delta H^0_{\mathrm{f298,B}} \tag{5.4}$$

or in general

$$\Delta H^0_{\mathrm{r298}} = \sum \nu_i \cdot \Delta H^0_{\mathrm{f298},i} \tag{5.5}$$

At temperature T the heat of reaction ΔH^0_{rT} is calculated from $\Delta H^0_{\mathrm{r298}}$ with the aid of the molar heat capacities (ideal gas state) c^0_{pi} of the reactants:

$$\Delta H^0_{\mathrm{r}T} = \Delta H^0_{\mathrm{r298}} + \int_{298.15}^{T} \Delta c_{p\mathrm{r}} \mathrm{d}T \tag{5.6}$$

with $\Delta c_{p\mathrm{r}} = \sum \nu_i c_{pi}$. This is in fact a heat balance. Therefore in the case of phase changes between standard-state and reaction conditions (T, p) for one or several reactants the respective phase change enthalpies ΔH^{trs} also have to be included yielding the following general expression:

$$\Delta H^0_{\mathrm{r}T} = \Delta H^0_{\mathrm{r298}} + \int_{298.15}^{T} \Delta c_{p\mathrm{r}} \mathrm{d}T + \sum \nu_i \Delta H_i^{\mathrm{trs}} \tag{5.7}$$

5.1.2. Gibbs Energy of Reaction and Chemical Equilibrium

Gibbs energies of reaction are calculated analogously to reaction enthalpies. For the reaction of Equation (5.1) the Gibbs energy ΔG_{r} is equal to the sum of the Gibbs energies of formation ΔG_{f} of the reactants multiplied by their stoichiometric coefficients:

$$\Delta G_{\mathrm{r}} = \gamma\Delta G_{\mathrm{fC}} + \delta\Delta G_{\mathrm{fD}} - \alpha\Delta G_{\mathrm{fA}} - \beta\Delta G_{\mathrm{fB}} \tag{5.8}$$

The general form of Equation (5.8) is

$$\Delta G_{\mathrm{r}} = \sum \nu_i \Delta G_{\mathrm{f}i} \tag{5.9}$$

In order to determine the standard Gibbs energy of reaction $\Delta G^0_{\mathrm{r}298}$ with Equation (5.8) or (5.9) the relevant standard quantities of formation $\Delta G^0_{\mathrm{f}298}$ have to be used.

Another way of calculating $\Delta G^0_{\mathrm{r}298}$ uses the definition

$$\Delta G_{\mathrm{r}} = \Delta H_{\mathrm{r}} - T \Delta S_{\mathrm{r}} \tag{5.10}$$

Application of Equation (5.10) requires $\Delta H^{\,0}_{\mathrm{r}298}$ from Equation (5.5) and $\Delta S^{\,0}_{\mathrm{r}298}$ obtained from standard entropies $S^{\,0}_{298}$ according to Equation (5.11):

$$\Delta S^0_{\mathrm{r}298} = \sum \nu_i S^0_{298} \tag{5.11}$$

Data for $\Delta G^0_{\mathrm{f}298}$ and $S^{\,0}_{298}$ can be found for many compounds in data collections, e.g., in [208], [211], [213], [214].

Conditions for equilibrium at constant pressure and temperature yield the following relation between the Gibbs energy of reaction $\Delta G^0_{\mathrm{r}T}$ (ideal gas at standard pressure = 1.01325 bar = 101.325 kPa) and equilibrium constant K:

$$\Delta G^0_{\mathrm{r}T} = -RT \ln K\,(T) \tag{5.12a}$$

For the chemical reaction of Equation (5.1) K is defined as

$$K = \frac{f^{\gamma}_{\mathrm{C}} f^{\delta}_{\mathrm{D}}}{f^{\alpha}_{\mathrm{A}} f^{\beta}_{\mathrm{B}}} \cdot \frac{f^{0\alpha}_{\mathrm{A}} f^{0\beta}_{\mathrm{B}}}{f^{0\gamma}_{\mathrm{C}} f^{0\delta}_{\mathrm{D}}} \tag{5.13}$$

with the standard fugacity f^0_i of the reactants A, B, C, D (equal to the standard pressure of 1 atm). K is a function of temperature only, but not of pressure. At low pressure where real gases do not show much difference from ideal gas phase behavior, the fugacities f_i can be considered as ideal gas partial pressures. The "thermodynamic" (true) equilibrium constant K is then practically equal to the numerical value of the equilibrium constant K_p expressed in partial pressures $p_i = y_i\, p$ (y_i is the mole fraction of component i in the gas phase) of the reactants as:

$$K_p = \frac{p^{\gamma}_{\mathrm{C}} p^{\delta}_{\mathrm{D}}}{p^{\alpha}_{\mathrm{A}} p^{\beta}_{\mathrm{B}}} \approx K \cdot (1\,\mathrm{atm})^{\sum \nu_i} \tag{5.14}$$

For all those reactions for which values of either $\Delta G^0_{\mathrm{f}298}$ or $\Delta H^0_{\mathrm{f}298}$ and S^0_{298} are available for each reactant the value of K at 25 °C (298.15 K) can easily be determined:

$$\ln K_{298.15} = -\frac{\Delta G^0_{\mathrm{r}298}}{298.15\,R} \tag{5.12b}$$

The dependence of K on temperature is given by the Van't Hoff Equation:

$$\frac{\mathrm{d} \ln K}{\mathrm{d}T} = \frac{\Delta H^0_{\mathrm{r}T}}{RT^2} \tag{5.15}$$

Integration of Equation (5.15) with $\Delta H^0_{\mathrm{r}T}$ as a function of temperature according to Equation (5.6) yields a general relation for calculating K at temperature T when it is known at temperature T_0 (e.g., 298.15 K):

$$\begin{aligned} \ln K\,(T) = {} & \ln K_{T_0} - \frac{\Delta H^0_{\mathrm{r}T_0}}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right) \\ & - \frac{1}{RT}\int_{T_0}^{T} \Delta c_{\mathrm{pr}}\,\mathrm{d}T \\ & + \frac{1}{R}\int_{T_0}^{T} \frac{\Delta c_{\mathrm{pr}}}{T}\,\mathrm{d}T \end{aligned} \tag{5.16}$$

For a first approximation the integrals in Equation (5.16) can be neglected which means that the reaction enthalpy ΔH_{r} is considered to be independent of temperature. With K known at T_0 = 298.15 K this first approximation is

$$\begin{aligned} & \ln K\,(T) \\ & = \ln K_{298.15} + \frac{\Delta H^0_{\mathrm{r}298}}{R}\left(\frac{1}{298.15} - \frac{1}{T}\right) \end{aligned} \tag{5.17}$$

In a second approximation the sum of the molar heat capacities Δc_{pr} is considered to be constant [$\Delta c_{\mathrm{pr}}(T) = \overline{\Delta c_{\mathrm{pr}}}$]:

$$\begin{aligned} \ln K\,(T) = {} & \ln K_{298.15} + \frac{\Delta H^0_{\mathrm{r}298}}{R}\left(\frac{1}{298.15} - \frac{1}{T}\right) \\ & + \frac{\overline{\Delta c_{\mathrm{pr}}}}{R}\left(\ln\frac{T}{298.15} + \frac{298.15}{T} - 1\right) \end{aligned} \tag{5.18}$$

For approximate estimations Equations (5.17) and (5.18) are sufficient; accurate predictions, especially at high temperatures, require evaluation of the integrals in Equation (5.16).

Concerning the effect of pressure on equilibrium concentrations the influence of real

gas behavior on K_p has to be considered. At low to moderate pressures (up to several hundred kilopascal) K_p is approximately equal to $K(1\,\text{atm})^{\sum \nu_i}$ (Eq. 5.14). But with higher pressure the difference between the numerical values of K and K_p increases and cannot be neglected at pressures > ca. 1 MPa. This effect can be predicted when the pVT behavior of the reaction mixture is known. Using the relation

$$f_i = \varphi_i y_i p = \varphi_i p_i \tag{5.19}$$

where φ_i is the fugacity coefficient for component i, Equation (5.13) is transformed as follows:

$$K = \frac{p_C^\gamma p_D^\delta}{p_A^\alpha p_B^\beta} \cdot \frac{\varphi_C^\gamma \varphi_D^\delta}{\varphi_A^\alpha \varphi_B^\beta} \cdot \left(\frac{1}{\text{atm}}\right)^{\sum \nu_i} = K_p K_\varphi \cdot \left(\frac{1}{\text{atm}}\right)^{\sum \nu_i} \tag{5.20}$$

The parameter K_φ represents the effect of real gas behavior on chemical equilibrium. The magnitude of this effect on high-pressure equilibria can be seen from Table 12, where data for the methanol synthesis equilibrium are given.

Table 12. Effect of pressure on equilibrium of methanol synthesis at 300 °C ($CO + 2\,H_2 \rightleftharpoons CH_3OH$) $K = 2.316 \times 10^{-4}$

p, bar (MPa)	K_φ	K_p, atm^{-2}
10 (1)	0.96	2.41×10^{-4}
25 (2.5)	0.90	2.57×10^{-4}
50 (5)	0.80	2.90×10^{-4}
100 (10)	0.61	3.80×10^{-4}
200 (20)	0.38	6.09×10^{-4}
300 (30)	0.27	8.58×10^{-4}

5.2. Group Contribution Methods for the Estimation of Enthalpies and Gibbs Energies of Formation

Especially in processes for new products, thermodynamic data (ΔH_f^0, ΔG_f^0) are often not available for all compounds and their estimation is extremely helpful.

Chemical reactions can be considered as processes in which molecular bonds are broken and others are formed. The thermodynamic functions for a chemical reaction can therefore be considered as the sum of the contributions of those chemical bonds which are altered (i.e., broken or formed) by the reaction. In a similar approach the structural groups of the reacting molecules can be used as the basis for estimating thermodynamic reaction quantities, in particular of organic substances.

Various such group contribution methods have been proposed for the estimation of standard enthalpies, Gibbs energies of formation, and standard entropies (Table 13). The method of Joback [216], [217] is explained in more detail.

In the Joback method ΔH_{f298}^0, ΔG_{f298}^0, and the molar heat capacities c_p^0 are calculated via Equation (5.21), (5.21a), and (3.46):

$$\Delta H_{f298}^0 = 68.29 + \sum s_i I_{Hi} \quad [\text{kJ/mol}] \tag{5.21}$$

$$\Delta G_{f298}^0 = 53.88 + \sum s_i I_{Gi} \quad [\text{kJ/mol}] \tag{5.21a}$$

where s_i is the number of structural groups of type i and I_{Hi} and I_{Gi} are the increments of structural groups i for enthalpy and Gibbs energy of formation, respectively.

The structural groups are the same as in the Joback method for the estimation of critical data (Chap. 2). The values of the group contributions I_{Hi} for ΔH_{f298}^0; I_{Gi} for ΔG_{f298}^0 ; and I_a, I_b, I_c, and I_d for the coefficients of the polynomial for c_p^0 are given in Table 14.

As is evident from Equations (5.21) and (5.21a), standard formation enthalpies ΔH_f^0 and standard Gibbs energy of formation ΔG_f^0 can be obtained directly for 298.15 K only. In order to calculate these quantities and the equilibrium constant K for other temperatures, the molar heat capacities c_p^0 have to be evaluated as a function of temperature with Equation (3.46); ΔH_f^0 and K can be then determined with Equations (5.6) and (5.16), respectively.

Deviations of ΔH_{f298}^0 and ΔG_{f298}^0 calculated with the Joback method from literature data are mostly below 10 kJ/mol; in a few cases, however, differences of more than 20 kJ/mol were found [218]. With the methods of Benson [219], [220] and of Yoneda [221] errors are somewhat smaller because they use larger numbers of groups which allow higher differentiation in structure. Application of these methods is, however, rather tedious. In both methods ΔG_{f298}^0 is obtained via Equation (5.10), formulated for the reaction of formation for which ΔH_{f298}^0 and the standard entropy S_{298}^0 are estimated with group contributions. In the same way, molar heat capacities c_p^0

Table 13. Group contribution methods for the estimation of ΔH_f^0, ΔG_f^0, and S^0

Authors	Estimated parameters	Form in which contributions are given	Average error, kJ/mol	Secondary sources for increments; remarks
Benson [219], [220]	ΔH_f^0, S^0	I_H (298), I_S (298); I_c (T)	10	[218]
Constantinou and Gani [225]	ΔH_f^0, ΔG_f^0	I_H (298); I_G (298)	10	ΔH_f^0 and ΔG_f^0 at 298 K only
Franklin [222]	ΔH_f^0	I_H (T)	15	[227]; in kcal/mol
Joback [216], [217]	ΔH_f^0, ΔG_f^0	I_H (298); I_G (298); $I_c = f(T)$	15	this article (Table 14); [218]
van Krevelen and Chermin [224]	ΔG_f^0	$I_G = A + BT$	15	[226], [227]; in kcal/mol
Verma and Doraiswamy [223]	ΔH_f^0	$I_H = A + BT$	15	[227]; in kcal/mol
Yoneda [221]	ΔH_f^0, S^0	I_H (298), I_S (298); $I_c = f(T)$	10	[218]

Table 14. Increments for the calculation of enthalpies and Gibbs energies of formation by Joback's method [216], [217]

Type of increment	I_H, kJ/mol	I_G, kJ/mol	Molar heat capacities ∗, J mol^{-1} K^{-1}			
			Δ_a	Δ_b	Δ_c	Δ_d
Nonring increments						
$-CH_3$	−76.45	−43.96	1.95 E+1	−8.08 E−3	1.53 E−4	−9.67 E−8
$>CH_2$	−20.64	8.42	−9.09 E−1	9.50 E−2	−5.44 E−5	1.19 E−8
>CH−	29.89	58.36	−2.30 E+1	2.04 E−1	−2.65 E−4	1.20 E−7
>C<	82.23	116.02	−6.62 E+1	4.27 E−1	−6.41 E−4	3.01 E−7
$=CH_2$	−9.63	3.77	2.36 E+1	−3.81 E−2	1.72 E−4	−1.03 E−7
=CH−	37.97	48.53	−8.00	1.05 E−1	−9.63 E−5	3.56 E−8
=C<	83.99	92.36	−2.81 E+1	2.08 E−1	−3.06 E−4	1.46 E−7
=C=	142.14	136.70	2.74 E+1	−5.57 E−2	1.01 E−4	−5.02 E−8
≡CH	79.30	77.71	2.45 E+1	−2.71 E−2	1.11 E−4	−6.78 E−8
≡C−	115.51	109.82	7.87	2.01 E−2	−8.33 E−6	1.39 E−9
Ring increments						
$-CH_2-$	−26.80	−3.68	−6.03	8.54 E−2	−8.00 E−6	−1.80 E−8
>CH−	8.67	40.99	−2.05 E+1	1.62 E−1	−1.60 E−4	6.24 E−8
>C<	79.72	87.88	−9.09 E+1	5.57 E−1	−9.00 E−4	4.69 E−7
=CH−	2.09	11.30	−2.14	5.74 E−2	−1.64 E−6	−1.59 E−8
=C<	46.43	54.05	−8.25	1.01 E−1	−1.42 E−4	6.78 E−8
Halogen increments						
−F	−251.92	−247.19	2.65 E+1	−9.13 E−2	1.91 E−4	−1.03 E−7
−Cl	−71.55	−64.31	3.33 E+1	−9.63 E−2	1.87 E−4	−9.96 E−8
−Br	−29.48	−38.06	2.86 E+1	−6.49 E−2	1.36 E−4	−7.45 E−8
−I	21.06	5.74	3.21 E+1	−6.41 E−2	1.26 E−4	−6.87 E−8
Oxygen increments						
−OH (alcohol)	−208.04	−189.20	2.57 E+1	−6.91 E−2	1.77 E−4	−9.88 E−9
−OH (phenol)	−221.65	−197.37	−2.81	1.11 E−1	−1.16 E−4	4.94 E−8
−O− (nonring)	−132.22	−105.00	2.55 E+1	−6.32 E−2	1.11 E−4	−5.48 E−8
−O− (ring)	−138.16	−98.22	1.22 E+1	−1.26 E−2	6.03 E−5	−3.86 E−8
>C=O (nonring)	−133.22	−120.50	6.45	6.70 E−2	−3.57 E−5	2.86 E−9
>C=O (ring)	−164.50	−126.27	3.04 E+1	−8.29 E−2	2.36 E−4	−1.31 E−7
O=CH− (aldehyde)	−162.03	−143.48	3.09 E+1	−3.36 E−2	1.60 E−4	−9.88 E−8
−COOH (acid)	−426.72	−387.87	2.41 E+1	4.27 E−2	8.04 E−5	−6.87 E−8
−COO− (ester)	−337.92	−301.95	2.45 E+1	4.02 E−2	4.02 E−5	−4.52 E−8
=O (except as above)	−247.61	−250.83	6.82	1.96 E−2	1.27 E−5	−1.78 E−8
Nitrogen increments						
$-NH_2$	−22.02	14.07	2.69 E+1	−4.12 E−2	1.64 E−4	−9.76 E−8
>NH (nonring)	53.47	89.39	−1.21	7.62 E−2	−4.86 E−5	1.05 E−8
>NH (ring)	31.65	75.61	1.18 E+1	−2.30 E−2	1.07 E−4	−6.28 E−8
>N− (nonring)	123.34	163.16	−3.11 E+1	2.27 E−1	−3.20 E−4	1.46 E−7
−N= (nonring)	23.61					
−N= (ring)	55.52	79.93	8.83	−3.84 E−3	4.35 E−5	−2.60 E−8
=NH	93.70	119.66	5.69	−4.12 E−3	1.28 E−4	−8.88 E−8
−CN	88.43	89.22	3.65 E+1	−7.33 E−2	1.84 E−4	−1.03 E−7
$-NO_2$	−66.57	−16.83	2.59 E+1	−3.74 E−3	1.29 E−4	−8.88 E−8
Sulfur increments						
−SH	−17.33	−22.99	3.53 E+1	−7.58 E−2	1.85 E−4	−1.03 E−7
−S− (nonring)	41.87	33.12	1.96 E+1	−5.61 E−3	4.02 E−5	−2.76 E−8
−S− (ring	39.10	27.76	1.67 E+1	4.81 E−3	2.77 E−5	−2.11 E−8

∗ E = exponent of 10: E + 1 = 10^1, E − 2 = 10^{-2} etc.

are determined when Equations (5.6) and (5.16) are applied for ΔH_r^0 and K at other temperatures.

Straightforward estimation of ΔH_f^0 and ΔG_f^0 at different temperatures is also possible [222–224]. Application of these older methods is convenient because they do not require estimation of heat capacity data. Results from these methods are, however, in kilocalories per mole.

5.3. Applications

5.3.1. Reaction Enthalpy

In the determination of reaction enthalpies with Equation (5.3) the enthalpy of formation is often not known for all reactants. The following procedure can then be adopted:

Example. Calculation of reaction enthalpy for ethoxylation of *n*-butanol at 390 K.

Monoethers of ethylene glycol (e.g., *n*-butoxyethanol) are produced by ethoxylation of alcohols, i.e., by addition of ethylene oxide:

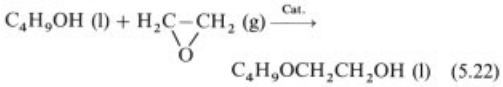

(5.22)

The design of the reactor for the ethoxylation of *n*-butanol at 390 K is dependent on the enthalpy of this reaction (ΔH_{r390}). Determination of ΔH_{r390} consists of three steps:

1) Calculation of ΔH_{r298}^0 from the standard formation enthalpies ΔH_{f298}^0 with Equation (5.5)
2) Calculation of ΔH_{r390}^0 with Equation (5.6)
3) Calculation of the ΔH_{r390} with *n*-butanol and *n*-butoxyethanol as liquids and ethylene oxide in the gaseous state, using Equation (5.7)

Step 1. ΔH_{r298}^0 with Equation (5.5).

ΔH_{f298}^0	*n*-butanol	− 274.9 kJ/mol
ΔH_{f298}^0	ethylene oxide	− 52.67 kJ/mol
ΔH_{f298}^0	*n*-butoxyethanol:	no data available

Estimation with Joback's method [216], [217]

Group i	s_i	I_{Hi}	$s_i I_{Hi}$
$-CH_3$	1	− 76.45	− 76.45
$-CH_2-$	5	− 20.64	−103.20
−OH	1	−208.04	−208.04
−O−	1	−132.22	−132.22

$\sum s_i I_{Hi} = -519.91$ kJ/mol

Equation (5.21): $\Delta H_{f298}^0 = 68.29 - 519.91 = -451.62$ kJ/mol

Equation (5.5): $\Delta H_{r298}^0 = -451.62 + 274.9 + 52.67 = -124.05$ kJ/mol

Step 2. ΔH_{r390}^0 with Equation (5.6)
Sum of molar heat capacities:

	ν_i	a	b	c	d
n-Butanol	−1	4.543	4.1782×10^{-1}	-2.242×10^{-4}	4.62×10^{-8}
Ethylene oxide	−1	−37.79	3.682×10^{-1}	-3.467×10^{-4}	20.86×10^{-8}
n-Butoxyethanol	1	28.255	5.4462×10^{-1}	-2.220×10^{-4}	1.52×10^{-8}
		61.502 $= \Delta a$	-2.414×10^{-1} $= \Delta b$	3.489×10^{-4} $= \Delta c$	-23.96×10^{-8} $= \Delta d$

$$\int_{T_0}^{T} \Delta c_{pr} \mathrm{d}T = \int_{T_0}^{T} \left(a + bT + cT^2 + dT^3\right) \mathrm{d}T = \Delta a\,(T - T_0) + \frac{\Delta b}{2}\left(T^2 - T_0^2\right) + \frac{\Delta c}{2}\left(T^3 - T_0^3\right) + \frac{\Delta d}{4}\left(T^4 - T_0^4\right)$$

$$\int_{298.15}^{390} \Delta c_{pr} \mathrm{d}T = 924.0\,\mathrm{J\,mol^{-1}}$$

$$\Delta H_{r390}^0 = -124.05 + 0.924 = -123.13\,\mathrm{kJ/mol}$$

Step 3. ΔH_{r390} with Equation (5.7)

$$\sum \nu_i \Delta H_i^{trs} = \Delta H_{n-\mathrm{butoxyethanol}}^{lv} - \Delta H_{n-\mathrm{butanol}}^{lv} = 48.517 - 43.781 = 4.736\,\mathrm{kJ/mol}$$

$$\Delta H_{r390} = -123.13 + 4.74 = -118.4\,\mathrm{kJ/mol}$$

5.3.2. Chemical Equilibrium

With Equation (5.12 a) the equilibrium constant K of a chemical reaction can be determined. The necessary thermodynamic data have been tabulated for many inorganic and organic compounds. However, in using results of such calculations, inaccuracies in the original data must be considered. Thus enthalpies and Gibbs energies of formation are derived from calorimetric measurements: formation enthalpies of some organic compounds are usually determined from measured heats of combustion. With combustion enthalpies of ca. ≥ 1000 kJ/mol, inaccuracies of about ± 1 kJ/mol in the final values of ΔG_r^0 and ΔH_r^0 have to be expected independent of the absolute value of the resulting quantities. If the Gibbs energy of formation has to be estimated for one or more of the reactants (cf. Section 5.2), the error in ΔG_r^0 is ≥ 10 kJ/mol.

In Table 15 an example of the effect of errors in ΔG_r^0 on K and on equilibrium concentrations is given. Obviously, K values obtained in this way are too inaccurate to be used for exact process calculations without experimental verification.

Table 15. Deviations of equilibrium constant K caused by errors in ΔG_r^0 ($\Delta G_r^0 = 0.0$ kJ/mol at 500 K)

Error in ΔG_r^0, kJ/mol	K
+ 1.0 (max.)	0.786
− 1.0 (min.)	1.272
+ 10.0 (max.)	0.090
− 10.0 (min.)	11.08

For preliminary investigations and for feasibility studies, however, predicted Gibbs energies of reaction with limited accuracy are helpful. For $-\Delta G_r^0-$ values > 40 kJ/mol, the equilibrium is either highly favorable for the reaction (for negative ΔG_r^0) or the reaction is practically impossible (for positive ΔG_r^0).

With knowledge of ΔG_r^0 it is possible to predict whether a reaction is impossible or thermodynamically feasible, i.e., whether the reaction equilibrium may yield the desired products if equilibrium can be achieved. Equilibrium may not, however, be reached in practice. For example, the velocity of the equilibrium reaction may be so low that no detectable conversion occurs. In other cases consecutive reactions of the equilibrium reaction may be so fast that product recovery becomes practically impossible.

The application of ΔG_r^0 for the selection of processes and process conditions will now be demonstrated for hydrocarbon pyrolysis. Figure 12 shows the temperature dependence of the Gibbs energy of formation for several hydrocarbons. All of the hydrocarbons apart from acetylene show increasing instability with higher temperatures, i.e., increasing decomposition into the elements carbon (as graphite) and hydrogen. The much lower temperature dependence of ethylene and benzene means that these compounds are less stable than paraffinic hydrocarbons (e.g., C_6H_{14} and $C_{20}H_{42}$) at low temperature and more stable at high temperature. Therefore, in pyrolysis of naphtha (C_5–C_{10} paraffins) for the production of ethylene (steam cracking), temperatures of 500 – 900 °C are employed. Apparently it is pointless to try to develop a catalyst for the synthesis of ethylene from naphtha at lower temperatures (e.g., 300 – 400 °C). On the other hand naphtha pyrolysis yielding acetylene requires higher temperatures, preferably > 1150 °C; with methane as raw material even higher temperatures are necessary.

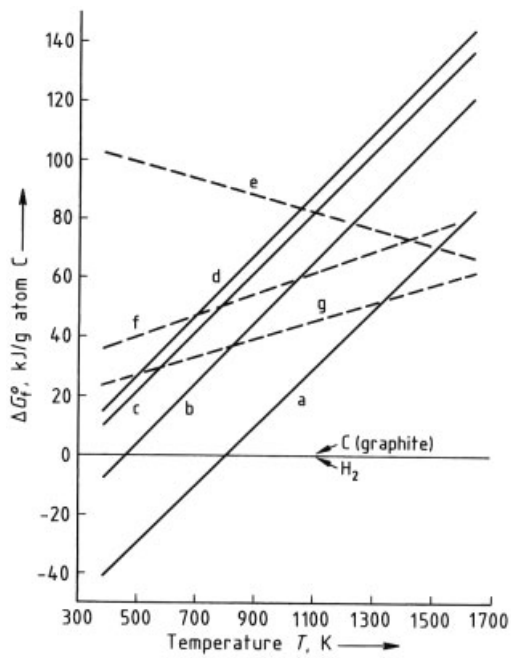

Figure 12. Temperature dependence of the Gibbs energy of formation for selected hydrocarbons
a) Methane; b) Ethane; c) Hexane; d) Eicosane (n-$C_{20}H_{42}$); e) Acetylene; f) Ethylene; g) Benzene

Due to increasing instability at higher temperature, reaction time has to be kept short to prevent decomposition; fast cooling of the reaction mixture after leaving the reactor is also necessary.

Example. Calculation of equilibrium constant for the reaction of 2-propanol to diisopropyl ether

2-Propanol is produced by reacting propene with water. In a subsequent reaction diisopropyl ether is formed at ca. 150 – 250 °C:

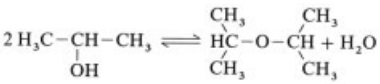

The loss in 2-propanol due to this reaction might be reduced by altering the reaction temperature. The temperature dependency of the equilibrium constant must therefore be determined. This can be done in three steps:

1) Determination of ΔG^0_{r298} (with Eq. 5.8) and of the equilibrium constant K_{298} (with Eq. 5.12 b)
2) First approximation for K_T at temperature T (Eq. 5.17) with ΔH^0_{r298} by using Equation (5.5).
3) Second approximation for K_T with Equation (5.18)

Step 1. ΔG^0_{r298} with Equation (5.8) and K_{298} with Equation (5.12 b)
Sum of the standard Gibbs energies and standard enthalpies of formation; data from [214].

	ν_i	ΔG^0_{f298}, kJ/mol	ΔH^0_{f298}, kJ/mol
2-Propanol	– 2	– 173.71	– 272.77
Diisopropyl ether	1	– 121.96	– 319.03
Water	1	– 228.77	– 242.00
		ΔG^0_{r298} = – 3.31	ΔH^0_{r298} = – 15.96

$$\ln K_{298} = -\frac{-3310}{8.314 \times 298.15} = 1.3353$$

$K_{298} = \mathbf{3.80}$

Step 2. First approximation of $\ln K = f(T)$ with Equation (5.17).

$$\ln K(T) = 1.3353 + \frac{-15960}{8.314}\left(\frac{1}{298.15} - \frac{1}{T}\right)$$

T, °C	$\ln K(T)$	$K(T)$
0	1.9246	6.852
25	1.3353	3.801
50	0.8372	2.310
100	0.0412	1.042
150	–0.5667	0.567
200	–1.0461	**0.351**
250	–1.4338	0.238

Step 3. Higher accuracy for $\ln K = f(T)$ with Equation (5.18)
Example: Reaction temperature T = 473.15 K (= 200 °C), calculate $\overline{\Delta c_{pr}} = \sum \nu_i c_{pi}$ (385.65K) with Joback's method [216], [217].

	ν_i	$c_{pi}{}^0_{385}$, J mol^{-1} K^{-1}
2-Propanol	–2	106.79
Diisopropyl ether	1	191.71
Water	1	34.35
		$\overline{\Delta c_{pr}}$ = 12.48

Evaluation of the last term of the sum in Equation (5.18):

$$\frac{\overline{\Delta c_{pr}}}{R}\left(\ln\frac{T}{298.15} + \frac{298.15}{T} - 1\right) = \frac{12.48}{8.314}\left(\ln\frac{473.15}{298.15} + \frac{298.15}{473.15} - 1\right) = 0.138$$

exp (0.138) = 1.148; K (473.15 K) = 1.148 0.351 = **0.403**

Comparison with the result from Step 2 (K = 0.351) shows that the contribution of the last term of Equation (5.18) is only minor.

Although formation of diisopropyl ether is less favored at higher temperature (250 °C), the equilibrium concentration of this product is still considerable. Moreover, the velocity of the sequential reaction increases with temperature. Therefore, a decrease in diisopropyl ether formation is not to be expected when the reaction temperature is increased.

In the above calculation tabulated data of ΔG^0_{f298} are used, in the following K_{298} is determined from ΔG^0_{f298} estimated with Joback's method.
Calculation of K_{298} in Step 1 using ΔG^0_{r298} determined with Joback's method

	ν_i	ΔG^0_{f298}, kJ/mol
2-Propanol	− 2	− 164.88
Diisopropyl ether	1	− 110.24
Water	1	− 228.56
		$\Delta G^0_{r298} = -9.04$

$$\ln K = -\frac{-9040}{8.314 \times 298.15} = 3.6469$$
$$K_{298} = 38.4$$

The value of K obtained with Joback's method is ten times higher than that determined with tabulated ΔG^0_{f298} data.

6. Transport Properties of Pure Compounds and Mixtures

Transport properties largely determine the type and dimensions of equipment in the process industries. They are therefore of great importance in the engineering design of many processes – not just those involved in the chemical industry. Although transport properties are macroscopic nonequilibrium quantities, they may, to a certain extent, be related to static equilibrium quantities. In particular, kinetic theory has been very useful in explaining the mechanisms of heat, mass, and momentum transport in dilute gases. For a detailed description of transport phenomena, see → Transport Phenomena.

From the ratios of the transport coefficients of the liquid to the gas phase given below it is evident that transport processes in the two phases differ considerably:

Density ratio	ϱ^l/ϱ^g	ca. 10^3
Viscosity ratio	η^l/η^g	10 – 100
Thermal conductivity ratio	λ^l/λ^g	10 – 100
Diffusion coefficient ratio	D^l/D^g	ca. 10^{-4}

6.1. Viscosity

The *dynamic viscosity* η of a fluid is defined as the ratio of the shear stress to velocity gradient. The SI unit of dynamic viscosity is $Pa \cdot s = N \cdot s/m^2$ but poise is also still in use where $1\,P = 1\,g\,s^{-1}\,cm^{-1} = 10^{-1}\,N \cdot s/m^2$ and $1\,cP = 1\,mPa \cdot s$

The *kinematic viscosity* ν [m^2/s] is defined by

$$\nu = \frac{\eta}{\varrho} \tag{6.1}$$

and is still often expressed in Stokes [cm^2/s].

A review on methods for predicting and correlating viscosity data has been given by Monnery et al. [228]. The following discussion is restricted to Newtonian fluids and thus includes most organic liquids and solvents but excludes many polymer solutions, melts, and slurries. Typical viscosity behavior of a pure substance (ethanol) is shown in Figure 13.

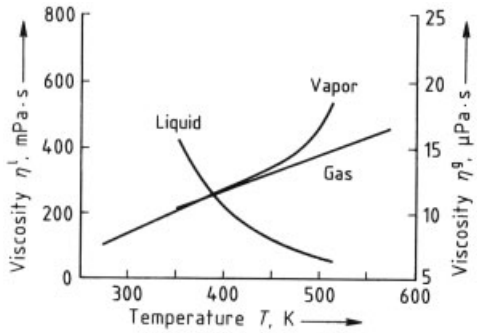

Figure 13. Viscosity behavior of ethanol [232]

6.1.1. Viscosity of Pure Gases

Gases at Low Pressure (< 1 MPa). Kinetic theory yields the following relation for gases consisting of rigid, noninteracting, elastic spherical molecules with a Maxwellian velocity distribution:

$$\eta = C' \frac{M^{1/2}}{\sigma^2} T^{1/2} \tag{6.2}$$

where η is expressed in $\mu Pa \cdot s$ and σ is the collision diameter in 10^{-10} m [Å]; M is the molecular mass in g/mol and C' a constant ($C' = 2.6693$). Chapman and Enskog described the effect of intermolecular forces by considering the potential energy of atoms and molecules [229]. They found that viscosity can be expressed in the general form

$$\eta = C' \frac{M^{1/2}}{\sigma^2 \Omega_v} T^{1/2} \tag{6.3}$$

where Ω_v, the collision integral, can be calculated from an appropriate potential energy function, such as that of Lennard – Jones (see Section

2.1). Despite the simplifying assumptions used when deriving Equation (6.3), the accuracy of the viscosity of nonpolar gases is quite acceptable (ca. 1 %).

Information on the potential energy function and the molecular parameters for more complex molecules is, however, scarce. Values for the molecular parameters and solutions for Ω_v for various compounds are tabulated in [230], [231].

Figure 14 shows viscosity data of some organic gases and vapors. In polar gases dipole–dipole interactions must be taken into account. Consequently the collision integral is calculated employing suitable potential functions, e.g., the Stockmayer potential [233]. The accuracy in the estimation of viscosity of polar gases by Equation (6.3) is usually better than 2 %.

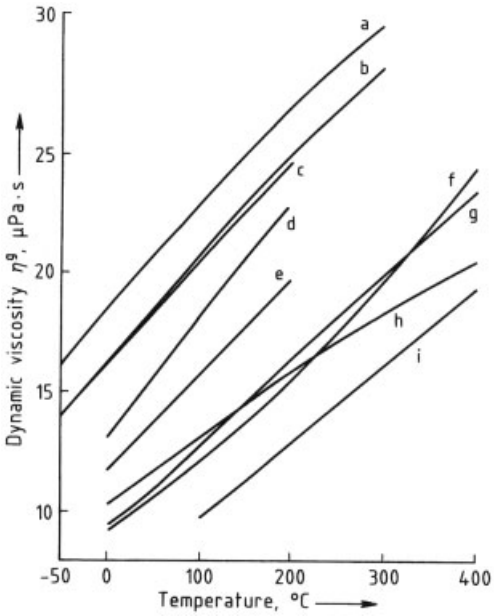

Figure 14. Viscosity data of some gases and vapors [232]
a) Helium; b) Carbon dioxide; c) Nitrogen; d) Hydrogen chloride; e) Hydrogen sulfide; f) Water; g) Ammonia; h) Methane; i) Hydrogen cyanide

Corresponding States Methods. Several methods employ the principle of corresponding states by defining a critical viscosity η_c. With the relation $\sigma = \text{const.} \times V_c^{1/3}$, Equation (6.3) transforms into:

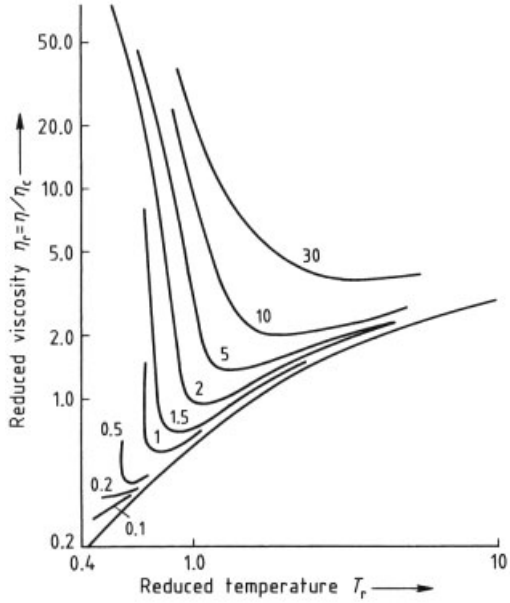

Figure 15. Generalized reduced viscosities as a function of reduced temperature [235]
Values on the curve denote the reduced pressure $p_r = p/p_c$.

$$\eta = C'' \left(\frac{M^3 p_c^4}{T_c} \right)^{1/6} \tag{6.4}$$

Figure 15 shows the generalized reduced viscosities as a function of reduced temperature. In the accurate method proposed by Lucas [236], [238] the basic relation for the reduced viscosity η_r is

$$\begin{aligned} \eta_r &= \frac{\eta}{\eta_c} \\ &\quad F_P^0 F_Q^0 \cdot \\ &= \left[0.807 T_r^{0.618} - 0.357 \exp(-0.449 T_r) \right. \\ &\quad \left. + 0.34 \exp(-4.058 T_r) + 0.018\right] \cdot \end{aligned} \tag{6.5}$$

where

$$\frac{1}{\eta_c} = 1.76 \left[10^{-4} T_c M^{-3} p_c^{-4}\right]^{1/6} \tag{6.6}$$

η is in μPa·s and p_c is in MPa. The correction factors F_P^0 and F_Q^0 account for polarity and quantum effects, respectively, and can be obtained from critical data. In terms of the reduced dipole moment μ_r

$$\mu_r = 524.6 \frac{p_c}{T_c^2} \mu^2 \tag{6.7}$$

three different equations have been formulated for F_P^0:

$$F_P^0 = 1\ 0\ \leq \mu_r < 0.022 \tag{6.7a}$$

$$F_P^0 = 1 + 30.55\,(0.292 - z_c)^{1.72}$$
$$0.022 \leq \mu_r < 0.075 \tag{6.7b}$$

$$F_P^0 = 1 + 30.55\,(-z_c)^{1.72}\,[0.96 + 0.1\,(T_r - 0.79)]$$
$$0.075 \leq \mu_r \tag{6.7c}$$

The factor F_Q^0 can be set equal to unity except in the case of quantum gases (He, H_2, D_2). The errors in predicting gas viscosities are usually $< 3\,\%$ for nonpolar compounds and somewhat higher for polar substances. Table 16 gives an overview on important estimation methods.

Effect of Pressure on Gas Viscosity. Above ca. 1 MPa, pressure has an increasing influence on gas viscosity particularly in the vicinity of the critical point.

LUCAS has suggested an efficient method for the prediction of the viscosity of dense gases with errors usually $< 5\,\%$ [236], [237]. The method correlates a reduced viscosity η/η_0 (where η_0 is the low-pressure viscosity) with pressure. The Lucas technique offers the advantage that the viscosity is expressed in terms of temperature and pressure, whereas other methods employ density as a variable and require an additional equation of state. Finally, if no experimental data are available, the low-pressure viscosity η_0 can be estimated with the same model (according to Eq. 6.5).

Example. Calculation of viscosity of gaseous carbon monoxide with the Lucas method
$M = 28.01$ g/mol, $T_c = 132.9$ K, $\mu = 0.1$ Debye, $p_c = 3.5$ MPa [233].
From Equation (6.6) $1/\eta_c = 7.02 \times 10^{-2}\,(\mu\text{Pa} \cdot \text{s})^{-1}$
Since $\mu_r = 1.04 \times 10^{-3}$ (Eq. 6.7) it follows that $F_P^0 = 1$.
Hence using Equation (6.5) $\eta^g = 15.1\ \mu\text{Pa} \cdot \text{s}$
Experimental value $\eta^g = 15.4\ \mu\text{Pa} \cdot \text{s}$ at 250 K [238]
Relative deviation: $-2.0\,\%$

6.1.2. Viscosity of Gas Mixtures

Extending the kinetic theory of Chapman and Enskog, viscosities of multicomponent gas mixtures can be approximated by

$$\eta_m = \frac{\sum y_i \eta_i}{\sum y_i \Phi_{ij}} \tag{6.8}$$

WILKE [239] and more recently BROKAW [240], [241] and REICHENBERG [242] have published methods for calculating the binary interaction parameter Φ_{ij}. Deviations with respect to experimental data are usually $< 5\,\%$ [230]. LUCAS has developed a calculation scheme that may be used if no pure component viscosity data are available [236], [237]. These methods are described in more detail, e.g., in [234].

6.1.3. Viscosity of Pure Liquids

Many methods have been published for estimating liquid viscosities. These methods can be applied in the temperature range slightly above the triple point up to reduced temperatures of about 0.75 – 0.8 at moderate pressures. A method for the prediction of viscosities at the melting point is given in [247].

No single method is clearly superior to the others. Since the method of Van Velzen et al. [248] is well established it is presented here. This group contribution method is based upon the following empirical relation (η in mPa · s):

$$\eta = 10^{\left[B\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]} \tag{6.9}$$

where T_0 is a fictive temperature and B is a parameter characteristic for the molecular structure of the substance concerned. It is determined using an equivalent chain length N^+

$$N^+ = N + \sum n_i I_N \tag{6.10}$$

where N is the total number of carbon atoms, I_N the increment, and n_i the number of increment I_N. The parameter B is given by

$$B = B_a + \sum I_B \tag{6.11}$$

For molecules with $N^+ \leq 20$ and $N^+ > 20$:

$$N^+ \leq 20:$$
$$T_0 = 28.86 + 37.439 N^+ - 1.3547\,(N^+)^2 + 0.02076\,(N^+)^3 \tag{6.12a}$$

$$N^+ > 20:$$
$$T_o = 8.164 N^+ + 238.59 \tag{6.12b}$$

Table 16. Selected methods for estimating low-pressure viscosity of gases

Method	Range of application		Information required [a]	Error, %
	Substances	Temperature		
Kinetic theory	polar and nonpolar		σ, μ, Ω_v, ε_0/k	ca. 2
Chung 1984 [243], 1988 [244]	not quantum gases		T_c, V_c, μ, σ	2
Bromley, Wilke 1951 [245]	polar compounds		T_c, V_c	ca. 5
Reichenberg 1975 [246]		$T_r < 3$	T_c, p_c, μ	1 [b], 3 [c]
Lucas 1980 [237], [238]	not highly associated gases		T_c, p_c, z_c, μ	1 [b], 3 [c]

[a] In addition to the corresponding parameters and increments.
[b] Nonpolar.
[c] Polar.

and

$$N^+ \leq 20:$$
$$B_a = 24.79 + 66.885 N^+ - 1.3173 \left(N^+\right)^2 - 0.00377 \left(N^+\right)^3 \quad (6.13a)$$

$$N^+ > 20:$$
$$B_a = 530.59 + 13.740 N^+ \quad (6.13b)$$

The structure parameters are given in Table 17. Since the method is only fairly accurate (errors 10 – 15 %) it cannot easily be applied to complex molecules [234]. Moreover, the predictions of viscosity for the first members of homologous series often yield higher errors.

Example. Calculation of viscosity of liquid diethylamine $(CH_3CH_2)_2NH$, using the method of Van Velzen et al. [248].

With $N = 4$:	$N^+ = 4 + 1.390 + 0.461 \times 4 = 7.234$
For secondary amines:	$I_B = 25.39 + 8.744\,N^+ = 88.64$
From Equation (6.13 a):	$B_a = 438.27$
From Equation (6.12 a):	$T_0 = 236.66\,K$

At 283.2 K $\eta^l_{cal} = 0.4306$ mPa · s, $\eta^l_{exp} = 0.3878$ mPa · s [249]; rel. deviation: 11.0 %
At 310.8 K $\eta^l_{cal} = 0.2944$ mPa · s, $\eta^l_{exp} = 0.2732$ mPa · s [249]; rel. deviation: 7.8 %

At higher temperatures ($T_r > 0.75$) the method of Letsou and Stiel [254] should be used. Their method is based upon the principle of corresponding states and valid in the range $0.76 < T_r < 0.98$. Only liquids at saturation pressure can be considered. The critical properties as well as the acentric factor must be known.

An overview of selected estimation methods for the viscosity of liquids is given in Table 18. In addition, viscosities of some liquids are presented as a function of temperature in Figure 16.

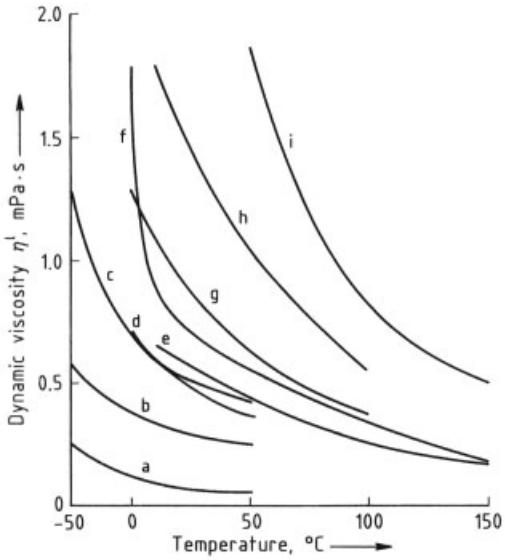

Figure 16. Dynamic viscosity η^l of some liquids [232]
a) Hydrogen chloride; b) Acetone; c) Trichloromethane; d) Methanol; e) Benzene; f) Water; g) Decane; h) Formic acid; i) Aniline

6.1.4. Viscosity of Liquid Mixtures

The viscosity of a liquid mixture is not usually a linear function of composition and is very difficult to predict. A first estimate of the viscosity of liquid mixtures can be obtained by

$$\ln \eta = \sum x_i \ln \eta_i \quad (6.14)$$

Methods to date do not allow confident predictions, particularly when dealing with highly po-

Table 17. Parameters of structural groups in the method of Van Velzen et al. [248] for estimating viscosities of liquids

Group		I_N	I_B
n-Alkanes		0	0
Isoalkanes		$1.389 - 0.238\,N$	15.51
Saturated hydrocarbons with two methyl groups in iso position		$2.319 - 0.238\,N$	15.51
n-Alkenes		$-0.152 - 0.042\,N$	$-44.94 + 5.410\,N^{+}$
n-Alkadienes		$-0.304 - 0.084\,N$	$-44.94 + 5.410\,N^{+}$
Isoalkenes		$1.237 - 0.280\,N$	$-36.01 + 5.410\,N^{+}$
Isoalkadienes		$1.085 - 0.322\,N$	$-36.01 + 5.410\,N^{+}$
Hydrocarbon with one double bond and two methyl groups in iso position		$2.626 - 0.518\,N$	$-36.01 + 5.410\,N^{+}$
Hydrocarbon with two double bonds and two methyl groups in iso position		$2.474 - 0.560\,N$	$-36.01 + 5.410\,N^{+}$
Cyclopentanes	$N < 16$	$0.205 + 0.069\,N$	$-45.96 + 2.224\,N^{+}$
	$N \geq 16$	$3.971 - 0.172\,N$	$-339.67 + 23.135\,N^{+}$
Cyclohexanes	$N < 17$	1.48	$-272.85 + 25.041\,N^{+}$
	$N \geq 17$	$6.517 - 0.311\,N$	$-272.85 + 25.041\,N^{+}$
Alkyl benzenes	$N < 16$	0.60	$-140.04 + 13.869\,N^{+}$
	$N \geq 16$	$3.055 - 0.161\,N$	$-140.04 + 13.869\,N^{+}$
Polyphenyls		$-5.340 + 0.815\,N$	$-188.40 + 9.558\,N^{+}$
Alcohols:			
Primary		$10.606 - 0.276\,N$	$-589.44 + 70.519\,N^{+}$
Secondary		$11.200 - 0.605\,N$	497.58
Tertiary		$11.200 - 0.605\,N$	928.83
Diols (correction)			557.77
Phenols (correction)		$16.17 - N$	213.68
−OH on side chain to aromatic ring (correction)		−0.16	213.68
Acids	$3 \leq N < 11$	$6.795 + 0.365\,N$	$-249.12 + 22.449\,N^{+}$
	$N \geq 11$	10.71	$-249.12 + 22.449\,N^{+}$
Iso acids	$3 \leq N < 11$	$6.795 + 0.365\,N - 0.24\,n_{\mathrm{ic}}$	$-249.12 + 22.449\,N^{+}$
	$N \geq 11$	$10.71 - 0.24\,n_{\mathrm{ic}}$	$-249.12 + 22.449\,N^{+}$
		(n_{ic} = number of methyl groups in iso position)	
Acids with aromatic nucleus in structure (correction)		4.81	$-188.40 + 9.558\,N^{+}$
Esters		$4.337 - 0.230\,N$	$-149.13 + 18.695\,N^{+}$
Esters with aromatic nucleus in structure (correction)		$-1.174 + 0.376\,N$	$-140.04 + 13.869\,N^{+}$
Ketones		$3.265 - 0.122\,N$	$-117.21 + 15.781\,N^{+}$
Ketones with aromatic nucleus in structure (correction)		2.70	$-760.65 + 50.478\,N^{+}$
Ethers		$0.298 + 0.209\,N$	$-9.39 + 2.848\,N^{+}$
Aromatic ethers		$11.5 - N$	$-140.04 + 13.869\,N^{+}$
Halogenated compounds:			
Fluoride		1.43	5.75
Chloride		3.21	−17.03
Bromide		4.39	$-101.97 + 5.954\,N^{+}$
Iodide		5.76	−85.32
Special configurations (corrections):			
$C(Cl)_x$		$1.91 - 1.459x$	−26.38
−CCl−CCl−		0.96	0
$-C(Br)_x-$		0.50	$81.34 - 86.850x$
−CBr−CBr−		1.60	−57.73
CF_3, in alcohols		−3.93	341.68
CF_3, in other compounds		−3.93	25.55
Aldehydes		3.38	$146.45 - 25.11\,N^{+}$
Aldehydes with an aromatic nucleus in structure (correction)		2.70	$-760.65 + 50.478\,N^{+}$
Anhydrides		$7.97 - 0.50\,N$	−33.50
Anhydrides with an aromatic nucleus in structure (correction)		2.70	$-760.65 + 50.478\,N^{+}$
Amides		$13.12 + 1.49\,N$	$524.63 - 20.72\,N^{+}$
Amides with an aromatic nucleus in structure (correction)		2.70	$-760.65 + 50.478\,N^{+}$

Table 17. (Continued)

Group	I_N	I_B
Amines:		
Primary	$3.581 + 0.325\,N$	$25.39 + 8.744\,N^{+}$
Primary amine in side chain of aromatic compound (correction)	-0.16	0
Secondary	$1.390 + 0.461\,N$	$25.39 + 8.744\,N^{+}$
Tertiary	3.27	$25.39 + 8.744\,N^{+}$
Primary amines with NH_2 group on aromatic nucleus	$15.04 - N$	0
Nitro compounds:		
1-nitro	$7.812 - 0.236\,N$	$-213.14 + 18.330\,N^{+}$
2-nitro	5.84	$-213.14 + 18.330\,N^{+}$
3-nitro	5.56	$-338.01 + 25.086\,N^{+}$
4-nitro; 5-nitro	5.36	$-338.01 + 25.086\,N^{+}$
Aromatic nitro-compounds	$7.812 - 0.236\,N$	$-213.14 + 18.330\,N^{+}$

Table 18. Selected methods for estimating the viscosity of pure liquids

Method	Range of application		Information required ∗	Error, % ∗∗
	Substances	Temperature		
Thomas 1946 [250]	no strongly polar and naphthenic compounds but sulfur-containing substances	$T > T_m + 5$ K $T < T_b$ $T_r < 0.75$	T_c, ϱ^1	20
Morris 1964 [251]	no sulfur-containing compounds, no highly branched substances	$T > T_m + 5$ K $T_r < 0.75$	T_c	14
Van Velzen et al. 1972 [248]	no sulfur-containing compounds, not for the first members of homologous series	$T_r < 0.75$		10 – 15
Orrick – Erbar 1974 [252]	no sulfur- or nitrogen-containing compounds	$T_r < 0.75$	T_c, M, ϱ^1	15
Przezdziecki – Sridhar 1985 [253]	no sulfur- or nitrogen-containing compounds, no alcohols	$T_r < 0.75$	T_c, p_c, V_c, M, ω, ϱ^1	15
Letsou – Stiel 1973 [254]	nonpolar, weakly polar compounds at saturation	$0.76 < T_r < 0.98$	T_c, p_c, M, ω	10
Joback – Reid 1987 [255]	hydrocarbons	$T > T_m + 5$ K $T_r < 0.75$		10
Tatevskii et al. 1987 [256]	no sulfur-, nitrogen-, or fluorine-containing compounds	$T > T_m + 5$ K $T_r < 0.75$		18

∗ In addition to the corresponding increments.
∗∗ In some cases errors may be considerably higher.

lar compounds or components differing greatly in size.

On the basis of Eyring's theory, McAllister published a relation for the kinematic viscosity of binary and ternary mixtures which allows estimations for chemically similar compounds with average errors of 15 % [257]. For details and recommendations see [258].

Viscosity data can be found, e.g., in [259]. The pressure – temperature dependence of viscosity for 50 pure fluids together with a comprehensive discussion of theories for dense fluids are given in [260].

6.2. Thermal Conductivity

The thermal conductivity λ is defined as the proportionality constant between heat flux and temperature gradient. It is usually expressed in terms of $W\,m^{-1}\,K^{-1}$.

6.2.1. Thermal Conductivity of Gases at Low Pressure

This chapter discusses methods for the prediction of thermal conductivities at pressures up to 1 MPa. The lower boundary is the region of extremely low pressures ($< 10^{-2}$ kPa) where the

mean free path of molecules becomes large compared to the macroscopic dimensions of the apparatus (Knudsen domain). Thermal conductivities of some organic gases and vapors are shown as a function of temperature in Figure 17. In the limiting case of noninteracting monatomic hard spheres, kinetic theory leads to the basic relation

$$\lambda^{\mathrm{g}} = 2.63 \times 10^{-23} \frac{T^{1/2}}{M^{1/2}\sigma^2 \Omega_{\mathrm{v}}} \tag{6.15}$$

where T is the absolute temperature, M the molecular mass, σ the characteristic dimension of the molecule, and Ω_{v} the collison integral. Equation (6.15) gives the temperature dependence of λ^{g} at low pressure.

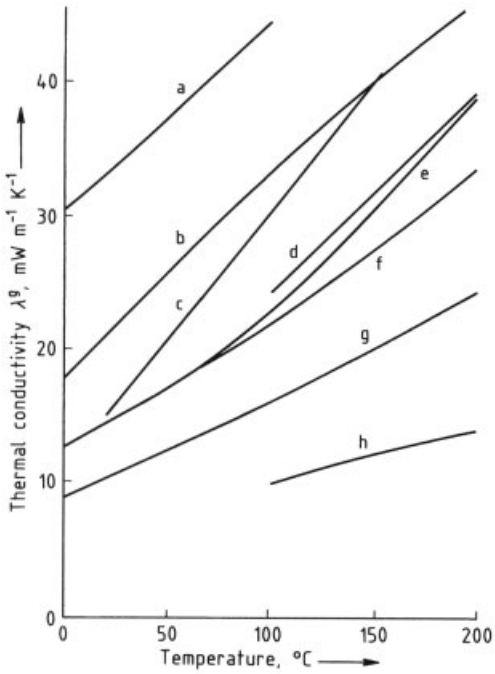

Figure 17. Thermal conductivities λ^{g} of some organic gases and vapors (data from [261])
a) Methane; b) Ethane; c) Dimethylamine; d) Water; e) Ethanol; f) Dimethyl ether; g) Chloromethane; h) Trichloromethane

The thermal conductivity λ and the viscosity η of a gas are closely related. Successful prediction methods (e.g., the Eucken model and the Mason and Moschick approach) have been proposed based on the dimensionless Eucken factor E

$$E = \frac{\lambda M}{\eta c_v} \tag{6.16}$$

where c_{v} is the molar heat capacity at constant volume. For further details see, e.g., [234].

Table 19. Coefficients A, B, and C for the Roy and Thodos method for predicting thermal conductivities of liquids (Eq. 6.20) [234]

Substance	$A\,T_{\mathrm{r}} + B\,T_{\mathrm{r}}^2 + C\,T_{\mathrm{r}}^3$
Saturated hydrocarbons ∗	$-0.152\,T_{\mathrm{r}} + 1.191\,T_{\mathrm{r}}^2 - 0.039\,T_{\mathrm{r}}^3$
Olefins	$-0.255\,T_{\mathrm{r}} + 1.065\,T_{\mathrm{r}}^2 + 0.190\,T_{\mathrm{r}}^3$
Acetylenes	$-0.068\,T_{\mathrm{r}} + 1.251\,T_{\mathrm{r}}^2 - 0.183\,T_{\mathrm{r}}^3$
Naphthalenes and aromatics	$-0.354\,T_{\mathrm{r}} + 1.501\,T_{\mathrm{r}}^2 - 0.147\,T_{\mathrm{r}}^3$
Alcohols	$1.000\,T_{\mathrm{r}}^2$
Aldehydes, ketones, ethers, esters	$-0.082\,T_{\mathrm{r}} + 1.045\,T_{\mathrm{r}}^2 + 0.037\,T_{\mathrm{r}}^3$
Amines and nitriles	$0.633\,T_{\mathrm{r}}^2 + 0.367\,T_{\mathrm{r}}^3$
Halides	$-0.107\,T_{\mathrm{r}} + 1.330\,T_{\mathrm{r}}^2 - 0.223\,T_{\mathrm{r}}^3$
Cyclic compounds (e.g., pyridine, thiophene, ethylene oxide, dioxane, piperidine)	$-0.354\,T_{\mathrm{r}} + 1.501\,T_{\mathrm{r}}^2 - 0.147\,T_{\mathrm{r}}^3$

∗ Not recommended for methane.

For the estimation of the thermal conductivity of gases, Roy and Thodos [262] used an approach based upon the reduced thermal conductivity λ_{r}

$$\lambda = \frac{\lambda_{\mathrm{r}}}{\left[210\,(M^3 T_{\mathrm{c}}/p_{\mathrm{c}}^4)^{1/6}\right]} \tag{6.17}$$

with the pressure p_{c} expressed in bar. As proposed by Eucken, the reduced thermal conductivity is separated into two terms

$$\lambda_{\mathrm{r}} = \lambda_{\mathrm{r}}^{\mathrm{trans}} + \lambda_{\mathrm{r}}^{\mathrm{int}} \tag{6.18}$$

The term $\lambda_{\mathrm{r}}^{\mathrm{trans}}$ considers the translational energy of the molecules and is determined by correlating dilute gas properties with temperature. The second term $\lambda_{\mathrm{r}}^{\mathrm{int}}$ takes internal (i.e., rotational and vibrational) energy contributions into account and employs specific polynomial coefficients for different classes of compounds.

$$\lambda_{\mathrm{r}}^{\mathrm{trans}} = 8.757\,[\exp(0.0464 T_{\mathrm{r}}) - \exp(-0.2412 T_{\mathrm{r}})] \tag{6.19}$$

$$\lambda_{\mathrm{r}}^{\mathrm{int}} = \beta\left[A T_{\mathrm{r}} + B T_{\mathrm{r}}^2 + C T_{\mathrm{r}}^3\right] \tag{6.20}$$

The coefficients A, B, and C are tabulated for nine classes of organic compounds (e.g., paraffins, olefins, alcohols). The additional parameter β is determined by employing group increments I_{i}

$$\beta = \sum I_i \tag{6.21}$$

The method of Roy and Thodos yields estimates for both polar and nonpolar organic substances at ambient pressure. The constants for this method are given in Tables 19 and 20. The only specific data required are the molecular mass and the critical temperature. The average error is ca. 5 %.

Example. Estimation of the thermal conductivity of 2-butanol, $CH_3CH_2CH(OH)CH_3$, at 480 K by the method of Roy and Thodos.
$M = 74.122$ g/mol, $T_b = 372.66$ K, $T_c = 535.95$ K, $p_c = 41.95$ bar [249].

$$\beta = (-CH_3) + (-CH_2-) + (-CH_2-) + (-CH_2) + (-OH)$$
$$= 0.73 + 2.0 + 3.18 + 3.68 + 4.62 = 14.21$$

with $A = 0$, $B = 1$, $C = 0$ from Table 19

From Equation (6.19):	$\lambda^{trans} = 2.073$
From Equation (6.20):	$\lambda^{int} = 11.398$
From Equation (6.21):	$\lambda = 0.0316$ W m^{-1} K^{-1}
Experimental:	$\lambda = 0.0324$ W m^{-1} K^{-1} [263]
Relative deviation: -0.8 %	

Other more sophisticated methods have been suggested based upon the principle of corresponding states, e.g., [264]. These methods employ the Eucken factor and require much more information on the properties of the substances considered, i.e., the critical properties (T_c, p_c, and V_c), the heat capacity at constant volume c_v, and the viscosity.

6.2.2. Thermal Conductivity of Gases at High Pressure

Typical $p\,\varrho\,\lambda$ behavior of fluids is displayed in Figure 18. Some methods used for gases at low pressures, particularly the corresponding states methods using the Eucken factor, have been extended to correlate high-pressure thermal conductivity data. For the region close to the critical point, however, other methods must be chosen.

The method of Chung et al. [266] is typical and will be briefly described here. These authors extended their equation for low-pressure thermal conductivity data:

$$\lambda^g = \frac{31.2\theta\eta^0}{M}\left(\frac{1}{G_2} + B_6 y\right) + qB_7 y^2 T_r^{1/2} G_2 \tag{6.22}$$

with λ expressed in terms of the reduced density function y (V, V_c in cm^3/mol)

$$y = \frac{V_c}{6V} \tag{6.23}$$

and the parameters

$$q = 3.586 \times 10^{-3}\,\frac{(T_c/M)^{1/2}}{V_c^{2/3}}$$

η^0	viscosity at low pressure, Pa · s
G_1, G_2	generalized functions of reduced density
B_1–B_7	parameters which are functions of the acentric factor ω, the reduced dipole moment μ (Eq. 6.31), and the association factor k (Eq. 6.24)
θ	function of T_r, c_v/R, and ω (Eq. 6.27)

Additional equations:

$$B_i = a_i + b_i\omega + c_i\mu_r + d_i k \tag{6.24}$$

The parameters a_i, b_i, c_i and d_i are listed in Table 21.

$$G_1 = \frac{1 - 0.5y}{(1-y)^3} \tag{6.25}$$

$$G_2 = \frac{B_1\left[1 - \exp(-B_4 y)\right] + B_2 G_1 \exp(B_5 y) + B_3 G_1}{y\,(B_1 B_4 + B_2 + B_3)} \tag{6.26}$$

$$\theta = 1 + \alpha\left(\frac{0.215 + 0.28288\alpha - 1.061\beta + 0.26665Z}{0.6366 + \beta Z + 1.061\alpha\beta}\right) \tag{6.27}$$

$$\alpha = c_v/R - 1.5 \tag{6.28}$$

$$\beta = 0.7862 - 0.7109\omega + 1.3168\omega^2 \tag{6.29}$$

$$Z = 2.0 + 10.5T_r \tag{6.30}$$

$$\mu_r = \frac{131.3\mu}{(T_c V_c)^{1/2}} \tag{6.31}$$

Data for the associaton factor k are given, e.g., in [234, p. 396].

Since no appropriate correlation for the polar factor is available, the application of the Chung method is limited to nonpolar and a few polar substances. Errors are usually < 8 % [234].

Table 20. Parameters for the Roy and Thodos method for predicting thermal conductivities of gases [234]

Type of group	ΔI
Paraffinic hydrocarbons	
Base group, methane	0.73
First methyl substitution (1)	2.00
Second methyl substitution (2)	3.18
Third methyl substitution (3)	3.68
Fourth and successive methyl substitutions (4)	4.56
Type of substitutions[a]	
$1 \leftarrow 2 \rightarrow 1$	3.64
$1 \leftarrow 2 \rightarrow 2$	4.71
$1 \leftarrow 2 \rightarrow 3$	5.79
$2 \leftarrow 2 \rightarrow 2$	5.79
$1 \leftarrow 3 \rightarrow 1$	3.39
$\downarrow$ 1	
$1 \leftarrow 3 \rightarrow 1$	4.50
$\downarrow$ 2	
$1 \leftarrow 3 \rightarrow 1$	5.61
$\downarrow$ 3	
Olefinic, acetylenic hydrocarbons	
First double bond $1 \leftrightarrow 1$	−1.19
$1 \leftrightarrow 2$	−0.65
$2 \leftrightarrow 2$	−0.29
Second double bond $2 \leftrightarrow 1$	−0.17
Any acetylenic bond	−0.83
Halides	
First halogen substitution on methane	
Fluorine	0.26
Chlorine	1.38
Bromine	1.56
Iodine	2.70
Second and successive halogen substitutions on methane	
Fluorine	0.38
Chlorine	2.05
Bromine	2.81
Substitutions on ethane and higher hydrocarbons	
Fluorine	0.58
Chlorine	2.93
Aldehydes, ketones	
$-CH_2-CH_3 \rightarrow -CH_2-CHO$	1.93
$-CH_2-CH_2-CH_2- \rightarrow -CH_2-CO-CH_2-$	2.80
Alcohols, type of −OH substitution	
On methane	3.79
$1 \leftarrow 1$	4.62
$2 \leftarrow 1$	4.11
$3 \leftarrow 1$	3.55
$4 \leftarrow 1$	3.03
$1 \leftarrow 2 \rightarrow 1$	4.12
Naphthenes	−1.0

Type of group	ΔI
Aromatics	
Benzene	13.2
Methyl-substituted benzenes	$13.2 - n5.28$[b]
$C \approx 5.21 \times 10^{-2} M + 1.82 \times 10^{-3} M^2$	$M < 120$[c]
Ethers	
$-CH_2OH \rightarrow -CH_2-O-CH_3$	2.46
Nitriles Type of −CN addition	
On methane	5.43
$CH_3-CH_3 \rightarrow CH_3-CH_2-CN$	7.12
$-CH=CH_2 \rightarrow -CH=CH-CN$	6.29
Acids and esters	
$-CH_2-O-CH_3 \rightarrow -CH_2-O-CH=O$	0.75
$-CH_2-O-CH_2- \rightarrow -CH_2-O-\overset{\overset{O}{\|}}{C}-$	0.31
Primary amines, type of substitution	
On methane	2.60
$1 \leftarrow 1$	3.91
$1 \leftarrow 2 \rightarrow 1$	5.08
$2 \leftarrow 2 \rightarrow 1$	7.85
$1 \leftarrow 3 \rightarrow 1$	6.50
$\downarrow$ 1	
Secondary amines	
$CH_3-NH_2 \rightarrow CH_3-NH-CH_3$	3.31
$-CH_2-NH_2 \rightarrow -CH_3-NH-CH_3$	4.40
Tertiary amines	
$CH_3-NH-CH_3 \rightarrow (CH_3)_3 \equiv N$	2.59
$-CH_2-\overset{\overset{H}{\|}}{N}-CH_2- \rightarrow -CH_2-\overset{\overset{CH_3}{\|}}{N}-CH_2-$	3.27
$-CH_2-\overset{\overset{H}{\|}}{N}-CH_3-CH_2-N-(CH_3)_2$	2.94
Cyclic, ring member	
$-CH_2-$	4.25
$-CH=$	3.50
$-NH-$	4.82
$-N=$	3.50
$-O-$	3.61
$=S=$	7.01
$C = \sum \Delta C - 7.83$	

Type of carbon atom:
[a] 1 = primary; 2 = secondary; 3 = tertiary; 4 = quaternary.
[b] n = number of methyl groups.
[c] M = molecular mass, g/mol.

Table 21. Constants (Eq. 6.24) for Chung's method, [234, p. 522] ∗

i	a_i	b_i	c_i	d_i
1	2.4166 E + 0	7.4824 E − 1	−9.1858 E − 1	1.2172 E + 2
2	−5.0924 E − 1	−1.5094 E + 0	−4.9991 E + 1	6.9983 E + 1
3	6.6107 E + 0	5.6207 E + 0	6.4760 E + 1	2.7039 E + 1
4	1.4543 E + 1	−8.9139 E + 0	−5.6379 E + 0	7.4344 E + 1
5	7.9274 E − 1	8.2019 E − 1	−6.9369 E − 1	6.3173 E + 0
6	−5.8634 E + 0	1.2801 E + 1	9.5893 E + 0	6.5529 E + 1
7	9.1089 E + 1	1.2811 E + 2	−5.4217 E + 1	5.2381 E + 2

∗ E = exponent of 10: E + 1 = 10^1, E + 2 = 10^2 etc.

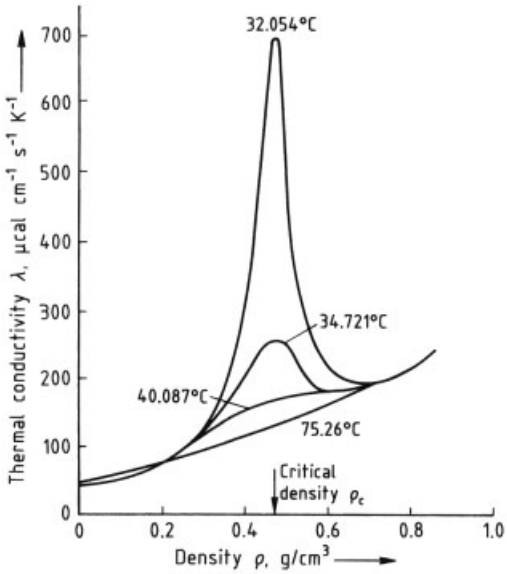

Figure 18. Isotherms of the thermal conductivity of carbon tetrachloride [265]

Example. Calculation of thermal conductivity of carbon dioxide at 473.15 K and 30 MPa $M = 44.01$ g/mol, $\omega = 0.239$, $T_c = 304.1$ K, $V_c = 93.9\ \mathrm{cm^3/mol}$ [234], $\mu = 0$ Debye; hence from Equation (6.31): $\mu_r = 0$, $V = 112.95\ \mathrm{cm^3/mol}$, $\varrho = 8.8537 \times 10^{-3}\ \mathrm{mol/cm^3}$, $c_v = 39.5\ \mathrm{J\,mol^{-1}\,K^{-1}}$ [267], $\eta^0 = 23.04 \times 10^{-6}$ (at 473.15 K, 0.1 MPa).

Calculation:	$G_1 = 1.4560$, $G_2 = 0.6489$, $\alpha = 3.251$, $\beta = 0.692$, $\theta = 1.8687$, $z = 18.34$
	$\lambda = 0.04968\ \mathrm{W\,m^{-1}\,K^{-1}}$
Experimental:	$\lambda = 0.05181\ \mathrm{W\,m^{-1}\,K^{-1}}$ [238]
Relative deviation	= − 4.1 %

The reliable estimation of thermal conductivities in dense gases is only possible for nonpolar substances; information about the critical properties, the acentric factor, the constant-volume heat capacity c_v at low pressure, and low-pressure viscosity data are required for the prediction. The accuracy of estimation methods is usually < 10 % [234]. An overview of estimation methods is given in Table 22.

6.2.3. Thermal Conductivity of Gas Mixtures

As a first approximation the thermal conductivity of a gas mixture can be described as a linear function of composition. Deviations increase with increasing polarity of the components and increasing difference in size and structure.

Methods for the prediction of mixture data are based upon the theoretically derived Wassiljewa equation:

$$\lambda_m = \frac{\sum y_i \lambda_i}{\sum y_j A_{ij}} \tag{6.32}$$

The binary interaction parameters A_{ij} can be determined employing the methods of Mason – Saxena or Lindsay [234]. Errors are usually < 8 % according to [234], but greater deviations are possible for components differing greatly in size and polarity. Nevertheless, reasonable estimates of thermal conductivities of mixtures can be obtained from pure component data using a linear combination of λ_i in mole fraction.

6.2.4. Thermal Conductivity of Liquids

Thermal conductivities of liquids are significantly higher than those of gases due to their much higher density. Liquid thermal conductivities λ^l are usually in the range

Table 22. Selected methods for estimating the thermal conductivity of gases at low pressure

Method	Range of application		Information required *	Error, %
	Substances	Temperature and pressure		
Eucken and modifications	polar and nonpolar compounds, no associating gases		M, c_v, η	10 – 15
Stiel – Thodos 1964 [270]	no associating gases	$T_r < 10$ $10^{-4} < p_r < 0.2$	M, T_c, p_c, c_v, η	11
Misic – Thodos 1961 [268]	hydrocarbons only		M, T_c, p_c, c_p	8
Roy – Thodos 1968, 1970 [262], [269]	preferably for polar compounds		M, T_c, p_c	8
Ely – Hanley 1983 [264]		low- and high-pressure version	M, T_c, V_c, z_c, ω, c_v	5 – 7
Chung et al. 1980 [266]	weakly polar but preferably nonpolar compounds	$T_r < 5$	M, T_c, ω, c_v (η)	5 – 7

* In addition to the corresponding increments.

0.1 – 0.17 W m^{-1} K^{-1} (Fig. 19) and are, like liquid densities, only a weak function of pressure. Comprehensive compilations of thermal conductivity data are given in [271], [272]. Since the numerical variation of λ^l values is small, estimation is comparatively easy, except for strongly polar or associating substances. Empirical methods usually provide estimates of considerable accuracy, often comparable to experimental uncertainty.

Sato's Method. SATO found that a simple empirical relationship exits between λ^l at the normal boiling point T_b and the molecular mass M [273]:

$$\lambda^l(T_b) = \frac{1.11}{M^{1/2}} \tag{6.33}$$

In order to extend this relation the equation [274] can be employed to give

$$\lambda^l = B\left[3 + 20\,(1 - T_r)^{2/3}\right] \tag{6.34}$$

where B is a substance-specific parameter [279]. Combination of Equations (6.33) and (6.34) yields

$$\lambda^l = \frac{1.11}{\sqrt{M}} \cdot \frac{3 + 20\,(1 - T_r)^{2/3}}{3 + 20\,(1 - T_b/T_c)^{2/3}} \tag{6.35}$$

The only information required is the critical temperature and the normal boiling point. Errors are usually < 15 %, but are generally higher for strongly polar substances.

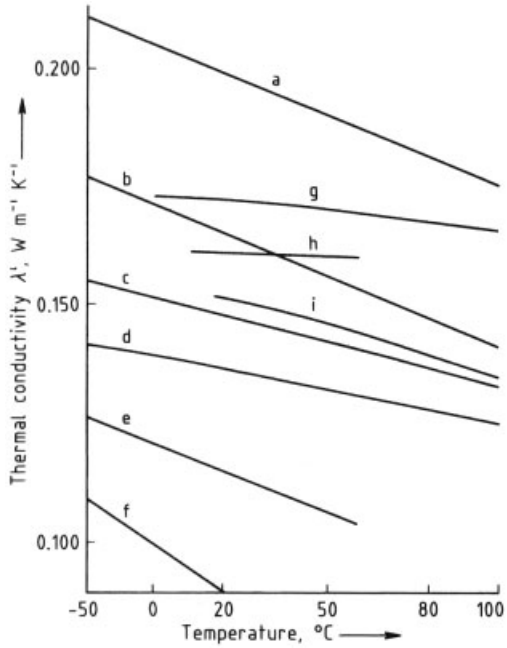

Figure 19. Thermal conductivities λ^l of some organic liquids (data from [261])
a) Methanol; b) Ethanol; c) 1-Butanol; d) 1-Octanol; e) Trichloromethane; f) Chlorodifluoromethane; g) Aniline; h) Pyridine; i) Nitrobenzene

Method of Latini and Baroncini [275], [276]. This estimation method is based upon the relation

$$\lambda^l = \frac{A^*}{M^\beta}\,\frac{T_b^\alpha}{T_c^\gamma}\,\frac{(1 - T_r)^{0.38}}{T_r^{1/6}} \tag{6.36}$$

Specific data required for the application of the method are the normal boiling point T_b, the criti-

cal temperature T_c, and the molecular mass. For chlorofluorohydrocarbons the method is applicable up to $T_r < 0.9$. The exponents α, β, γ and the parameter $A*$ are given in Table 23 for various organic compounds. According to [234] the errors in estimating liquid thermal conductivity are usually $< 10\,\%$.

Table 23. Constants for the Latini – Baroncini method (from [234])

Compounds	$A*$	α	β	γ
Saturated hydrocarbons	0.0035	1.2	0.5	0.167
Olefins	0.0361	1.2	1.0	0.167
Cycloparaffins	0.0310	1.2	1.0	0.167
Aromatics	0.0346	1.2	1.0	0.167
Alcohols	0.00339	1.2	0.5	0.167
Organic acids	0.00319	1.2	0.5	0.167
Ketones	0.00383	1.2	0.5	0.167
Esters	0.0415	1.2	1.0	0.167
Ethers	0.0385	1.2	1.0	0.167
Chlorofluorohydrocarbons:				
R 20, R 21, R 23	0.562	0	0.5	−0.167
Others	0.494	0	0.5	−0.167

Both estimation methods can be employed reliably up to $T_r < 0.65$. Numerous other methods have been published, but are either of limited applicability or require much more specific information without offering significant improvement in reliability and accuracy.

An overview of methods for estimating the thermal conductivities of liquids is given in Table 24.

Example. Estimation of thermal conductivity of liquid isopropylbenzene, C_9H_{12}, at 323 K.
$M = 120.194$ kg/kmol, $T_b = 425.56$ K, $T_c = 631.13$ K [249]

Experimental: $\lambda^l = 0.1187\ \mathrm{W\,m^{-1}K^{-1}}$ [276]
Calculation:
Sato's method (Eq. 6.35): $\lambda^l = 0.125\ \mathrm{W\,m^{-1}K^{-1}}$
Relative deviation=5.1 %
Latini's method (Eq. 6.36): $\lambda^l = 0.119\ \mathrm{W\,m^{-1}K^{-1}}$
Relative deviation=0.2 %

Effect of Temperature on Thermal Conductivity. At low to moderate pressures (< 1 MPa) and over a limited range of temperature, the thermal conductivity decreases linearly with temperature

$$\lambda^l = A - BT \tag{6.37}$$

For more accurate description over wider temperature ranges, higher order polynomial series are recommended, preferably with $(1 - T_r)$ as an independent variable.

Effect of Pressure on Thermal Conductivity. The thermal conductivity increases significantly at pressures above $p_r > 0.5$. MISSENARD suggested an empirical approach for the dimensionless thermal conductivity normalized by the thermal conductivity at low pressure (p^0) [281]:

$$\frac{\lambda^l(p_r)}{\lambda^l(p^0)} = 1 + C p_r^{0.7} \tag{6.38}$$

Values of the generalized parameter $C = f(T_r, p_r)$ are reported by the author for several compounds.

6.2.5. Thermal Conductivity of Liquid Mixtures

The thermal conductivity of organic liquid mixtures λ_m can usually be approximated with sufficient accuracy by linear combination of the thermal conductivities of the pure components.

For binary systems JAMIESON et al. recommended the following relation in terms of the weight fraction w_i [272]:

$$\lambda_m = w_1\lambda_1 + w_2\lambda_2 - \alpha(\lambda_2 - \lambda_1)\left(1 - w_2^{1/2}\right) w_2 \tag{6.39}$$

which employs an adjustable binary parameter α. If no experimental mixture data are available, α is set equal to unity.

For multicomponent mixtures LI suggested an equation with volume fractions [282]:

$$\lambda_m = 2\sum_i^r \sum_j^r \Phi_i \Phi_j \frac{1}{(1/\lambda_i + 1/\lambda_j)} \tag{6.40}$$

where the volume fraction Φ_i of component i is defined by

$$\Phi_i = \frac{x_i V_i}{(\sum x_j V_j)} \tag{6.41}$$

6.3. Diffusion Coefficients

The diffusion coefficient D is defined as the proportionality factor between mass flux and the

Table 24. Selected methods for estimating the thermal conductivities of liquids

Method	Range of application		Information required	Error, %
	Substances	Temperature		
Robbins – Kingrea 1962 [277]	no nitrogen- or sulfur-containing compounds	$0.4 < T_r < 0.9$	$T_b, T_c, c_p, \varrho^l, \Delta H_b^{lv}$	5
Missenard 1965 [278]			$M, T_b, c_p^l, \varrho_b^l$	8
Sato – Riedel 1973 [279]	no low molecular mass hydrocarbons, no highly polar compounds	$0.5 < T_r < 0.75$	M, T_b, T_c	15
Baroncini – Latini 1978, 1984 [275], [276]	no nitrogen- or sulfur-containing compounds, no aldehydes $50 \leq M \leq 250$		M, T_b, T_c	8
Nagvekar – Daubert 1987 [280]	no sulfur-containing compounds	$0.3 < T_r < 0.8$	T_c	6

concentration gradient; it is usually given in [cm^2/s]. In this discussion only molecular diffusion is considered with concentration gradients (i.e., primarily gradients in chemical potential) as driving force in systems which are free of external force fields and are homogeneous with respect to temperature and pressure. Diffusion is treated comprehensively in [233], [283].

6.3.1. Diffusion Coefficients of Gases at Low and Moderate Pressures

Molecular Theory. Statistical mechanics lead, via integration of the Boltzmann equation, to the Chapman and Enskog equation for the mutual diffusion of the components *i* and *j*:

$$D_{ij} = 3/16\,(2\pi kT)^{1/2}\,\frac{(1/M_i + 1/M_j)^{1/2}}{\pi N \sigma_{ij}^2 \Omega_D}\,f_D \quad (6.42)$$

where k is the Boltzmann constant; M_i, M_j the molecular masses of components *i, j* respectively; N the number of molecules per unit volume; σ_{ij} a characteristic length; and f_D a correction term. Introducing the ideal gas law into this equation gives a basic relation that is valid for spherical molecules in dilute gases:

$$D_{ij} = 0.01333\,\frac{T^{1.5}\,[1/M_i + 1/M_j]^{0.5}}{p}\,\frac{1}{\sigma_{ij}^2 \Omega_D} \quad (6.43)$$

This expression (p in MPa) contains the characteristic length σ_{ij} (in 10^{-10} m) and the dimensionless collison integral Ω_D, which can be determined from an appropriate intermolecular potential. For monatomic gases the potential function of Lennard and Jones may be chosen (values of σ_{ij} and Ω_D for the Lennard – Jones potential are given for some components, e.g., in [234]).

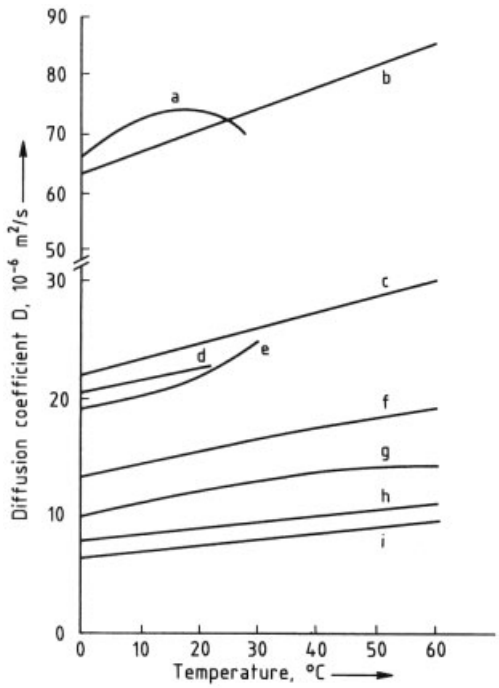

Figure 20. Diffusion coefficients of selected gases in air at ambient pressure
a) Hydrogen; b) Helium; c) Water; d) Ammonia; e) Methane; f) Methanol; g) Ethanol; h) Benzene; i) Hexane

Employing simple combination rules, i.e.,

$$\sigma_{ij} = \frac{\sigma_i + \sigma_j}{2} \text{ and } \varepsilon_{ij} = (\varepsilon_i \varepsilon_j)^{1/2} \quad (6.44)$$

diffusion coefficients can be determined with reasonable accuracy without considering concentration effects, even for complicated molecules. For mixtures containing highly polar compounds more sophisticated intermolecular potential functions must be applied. Diffusion coefficients of some components in air are

shown as a function of temperature in Figure 20.

Method of Fuller et al. FULLER and coworkers suggested an incremental method for D_{ij} [cm^2/s] based on Equation (6.43) [283], [284]:

$$D_{ij} = \frac{1.013\,[1/M_i + 1/M_j]^{1/2}}{10^4 p} \cdot \frac{T^{1.75}}{\left[(I_\mathrm{D})_i^{1/3} + (\Sigma I_\mathrm{D})_j^{1/3}\right]^2} \tag{6.45}$$

with p in MPa where $(\Sigma I_\mathrm{D})_i$ and $(\Sigma I_\mathrm{D})_j$ are the diffusion volumes of components i and j, respectively which are given by FULLER et al. for simple molecules such as the noble gases, nitrogen, and oxygen [284]. Some values for diffusion volumes (ΣI_D) follow:

He	2.67	CO	18.0
Ne	5.98	CO_2	26.9
Ar	16.2	N_2O	35.9
Kr	24.5	NH_3	20.7
Xe	32.7	H_2O	13.1
H_2	6.12	SF_6	71.3
D_2	6.84	Cl_2	38.4
N_2	18.5	Br_2	69.0
O_2	16.3	SO_2	41.8
Air	19.7		

For other molecules (ΣI_D) is calculated by summing up the increments I_D for the elements and structural groups in the molecule. Values for these increments follow [284]:

C	15.9	F	14.7
H	2.31	Cl	21.0
O	6.11	Br	21.9
N	4.54	I	29.8
Ring, aromatic	−18.3	S	22.9
Ring, heterocyclic	−18.3		

The authors claim an average accuracy in the prediction of diffusion coefficients of ca. 4 %.

Example. Calculation of D_{12} of ethylene – water using Fuller's method.

1) C_2H_4: M = 28.05 g/mol; $(\Sigma I_\mathrm{D})_{C_2H_4}$ = 2 C + 4 H = 2×15.9 + 4×2.31 = 41.04
2) H_2O: M = 18.02 g/mol; $(\Sigma I_\mathrm{D})_{H_2O}$ = 2 H + 1 O = 2×2.31 + 1×6.11 = 10.73

From Equation (6.45): D_{12} = 0.242 cm^2/s (at 0.1 MPa, 328.4 K)

Experimental: D_{12} = 0.236 cm^2/s (at 328.4 K) [234]
Relative deviation: 2.7 %

Temperature Dependence of Binary Diffusion Coefficients. Molecular theory (Eq. 6.42) predicts the following form for the temperature dependence of binary diffusion coefficients:

$$D \sim T^{1.5} \tag{6.46}$$

[neglecting $\Omega_\mathrm{D}(T)$]. For real systems the exponent is often about 1.8.

Effect of Pressure on Binary Diffusion Coefficients of Gases. In general, diffusion coefficients decrease with increasing pressure according to Equation (6.43). At higher pressures ($p >> p_\mathrm{c}$) the influence of pressure becomes more complex and corresponding states methods must be employed [285], [286].

Effect of Concentration on Diffusion Coefficients of Gases. Binary diffusion coefficients can usually be assumed to be independent of composition. A discussion of the theory of diffusion in multicomponent gas mixtures is given, e.g., in [233], [283], [287]. A procedure for predicting multicomponent diffusion phenomena is given in [288].

6.3.2. Diffusion Coefficients of Liquids

The Stokes – Einstein equation provides the basis for many empirical correlations of diffusion coefficients in liquids. This relation describes the movement of a large hypothetical spherical particle through a solvent consisting of infinitely small molecules. With these assumptions binary diffusion coefficients are given as a function of solvent viscosity η_j and the "molecular" radius of the solute r_i:

$$D_{ij} = \frac{kT}{6\pi\eta_j r_i} \tag{6.47}$$

The effect of solvent viscosity on binary diffusion coefficients is shown in Figure 21.

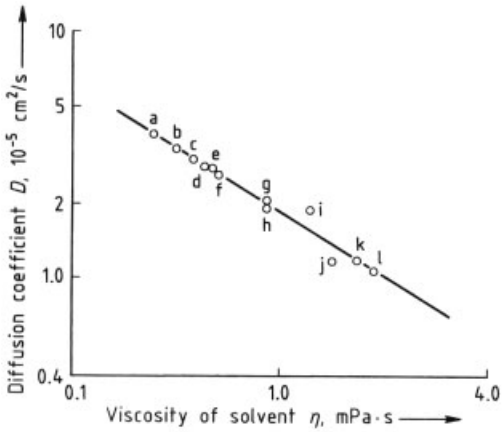

Figure 21. Diffusivity of carbon tetrachloride in various solvents [289]
a) Hexane; b) Heptane; c) Isooctane; d) Methanol; e) Toluene; f) Benzene; g) Carbon tetrachloride; h) Cyclohexane; i) Ethanol; j) Dioxane; k) Kerosine; l) Decalin

Since Equation (6.47) only holds for simple systems, a variety of empirical correlations have been published. They are usually applicable to solvent – solute systems with low to moderate viscosity ($< 20 - 40$ mPa · s).

Diffusion Coefficients at Infinite Dilution. The diffusion coefficient D_{ij}^{∞} for solute i at infinite dilution in solvent j is a useful reference value for the determination of binary diffusion coefficients of liquid mixtures.

The estimation technique of Hayduk and Minhas comprises several equations for different types of solutions [290]:

1) Nonelectrolyte systems:

$$D_{ij}^{\infty} = \frac{P_j^{0.5}}{P_i^{0.42}} \frac{1.55 \times 10^{-8} T^{1.29}}{V_{0j}^{0.23} \eta_j^{0.92}} \tag{6.48}$$

where V_{0j} [cm^3/mol] is the liquid molar volume of solvent j at its normal boiling point, η_j its viscosity [mPa · s], and P the parachor (see Section 7.1) [cm^3 g$^{0.25}$ s$^{-0.5}$ mol^{-1}]. From the definition of the parachor as $P \equiv \sigma_i^{0.25} V_{0i}$ (Eq. 7.2 where σ_i is the surface tension of component i), Equation (6.48) can be rewritten as

$$D_{ij}^{\infty} = 1.55 \times 10^{-8} \frac{V_{0j}^{0.420.27}}{V_{0i}} \frac{T^{1.29}}{\eta_j^{0.92}} \frac{\sigma_j^{0.125}}{\sigma_i^{0.105}} \tag{6.49}$$

In many cases the surface tensions σ_i and σ_j are similar and their ratio can be set equal to unity, making this expression very convenient.

2) n-Paraffin solutions:

$$D_{ij}^{\infty} = 1.33 \times 10^{-7} \frac{T^{1.47}}{V_{0i}^{0.71}} \eta_j^{c} \tag{6.50}$$

where the exponent $c = (10.2/V_{0i}) - 0.791$. According to the authors errors are < 4 %.

3) Organic solutes in water:

$$D_{ij}^{\infty} = 1.25 \times 10^{-8} \left(V_{0i}^{-0.19} - 0.292\right) T^{1.52} \eta_j^{\delta} \tag{6.51}$$

where the parameter $\delta = (9.58/V_{0i}^{l}) - 1.12$. According to the authors the errors are usually < 10 %.

Example. Calculation of infinite dilution diffusion coefficients using the Hayduk – Minhas method.

1) Acetone in ethyl acetate [249]:

	Acetone	Ethyl acetate
M, g/mol	58.08	88.106
$V^{l}{}_{0i}$, cm^3/mol	73.52	97.83
η_i, mPa · s		0.4508
σ_i, N/m	0.02332	0.02375

From Equation (6.49): $D_{12}^{\infty} = 2.59 \times 10^{-5}$ cm^2/s
Experimental: $D_{12}^{\infty} = 3.18 \times 10^{-5}$ cm^2/s at 293 K [235]
Relative deviation: -18.6 %

2) Toluene in n-hexane [249]:

	Toluene	n-Hexane
M, g/mol	92.14	86.177
$V^{l}{}_{0i}$, cm^3/mol	106.9	131.60
η_i, mPa · s		0.2942

Substituting $c = -0.6956$ in Equation (6.50):
$D_{12}^{\infty} = 4.90 \times 10^{-5}$ cm^2/s
Experimental: $D_{12}^{\infty} = 4.12 \times 10^{-5}$ cm^2/s at 298 K [235]
Relative deviation: 18.9 %

3) Aniline in water [249]:

	Aniline	Water
M, g/mol	93.128	18.02
V^1_{0i}, cm^3/mol	77.42	
η_i, mPa · s		1.002

Substituting $\delta = -0.996$ in Equation (6.51):
$D^{\infty}_{12} = 1.021 \times 10^{-5}$ cm^2/s
Experimental: $D^{\infty}_{12} = 0.92 \times 10^{-5}$ cm^2/s at 293 K [235]
Relative deviation: 11.0 %

The application of the method of Hayduk et al. is somewhat more complicated for systems with associating compounds (e.g., organic acids). Table 25 gives an overview of estimation methods.

Effect of Composition on Binary Diffusion Coefficients in Liquids. The dependence of binary diffusion coefficients on concentration is shown in Figure 22. In mixtures with compounds differing greatly in size and polarity the effect of concentration on diffusion coefficients cannot be neglected. Even a linear combination of diffusion coefficients at infinite dilution does not yield accurate results. The thermodynamic correction term ($\partial \ln a_i \,/\, \partial \ln x_i$) is common to many estimation methods as in the Darken equation [291] for a binary mixture of components i and j

$$D_{ij} = \left(x_i D^{\infty}_{ij} + x_j D^{\infty}_{ji}\right) \frac{\partial \ln a_i}{\partial \ln x_i} \tag{6.52}$$

where x is the mole fraction and a the activity; $a_i\,(x_i)$ can be obtained from vapor – liquid equilibrium data. From the Gibbs – Duhem equation it follows that

$$\frac{\partial \ln a_i}{\partial \ln x_i} = \frac{\partial \ln a_j}{\partial \ln x_j} \tag{6.53}$$

Application of the NRTL equation (Section 4.2.1.2) yields for a binary mixture:

$$\frac{\partial \ln a_1}{\partial \ln x_1} = 1 - 2x_1x_2 \times \left[\frac{\tau_{21}G^2_{21}}{(x_1 + x_2 G_{21})^3} + \frac{\tau_{12}G^2_{12}}{(x_2 + x_1 G_{12})^3}\right] \tag{6.54}$$

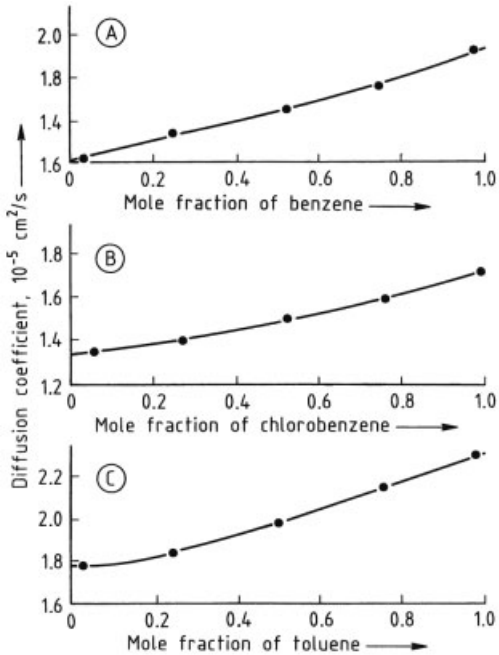

Figure 22. Effect of composition on binary diffusion coefficients [294]
A) Benzene – carbon tetrachloride at 298.41 K; B) Chlorobenzene – bromobenzene at 299.93 K; C) Toluene – chlorobenzene at 300.11 K

For systems showing ideal mixing behavior the differential quotient $\partial \ln a_i \,/\, \partial \ln x_i$ is equal to zero. In view of the fact that Equation (6.53) in some cases overcorrects for nonideal behavior, RATHBURN and BABB recommended the introduction of an exponential constant f for the thermodynamic correction [292]:

$$D_{ij} = \left(x_i D^{\infty}_{ij} + x_j D^{\infty}_{ji}\right) \left(\frac{\partial \ln a_i}{\partial \ln x_i}\right)^f \tag{6.55}$$

The values for f originally proposed by RATHBURN and BABB (in parentheses) have been reevaluated by SIDDIQI and LUCAS [293]: $f = 0.6$ (0.6) for positive deviations from Raoult's law and $f = 0.4$ (0.3) for negative deviations.

Another correlation has been proposed by VIGNES [295]:

$$D_{ij} = \left[\left(D^{\infty}_{ij}\right)^{xj} \left(D^{\infty}_{ji}\right)^{xi}\right] \frac{\partial \ln a_i}{\partial \ln x_i} \tag{6.56}$$

SIDDIQI and LUCAS found that the average error was less than 5 % for the prediction of 79 data sets. For systems with polar components they report deviations of 11 %.

Table 25. Selected methods for estimating diffusion coefficients in liquid solutions at infinite dilution

Method	Range of applicability	Information required *	Error, % **
Wilke – Chang 1955 [296]	not reliable for water as solute	$M_2, \eta_2, V^1_1, (T_b)$, association parameters	11 (w), 20 (o)
Scheibel 1954 [297]	usually appropriate for aqueous solutions	$M_2, \eta_2, V^1_1, (T_b)$	11 (w), 20 (o)
Othmer – Thakar 1953 [298]	preferably for aqueous solutions	V^1_{01}, η_{water}	5 (w)
Reddy – Draiswamy 1967 [299]		$M_2, V^1_{01}, V^1_{02}\ \eta_2$	22 (o), 25 (w)
Lusis – Ratcliff 1968 [300]	preferably for alkane – alkane systems, not for aqueous solutions	$V^1_{01}, V^1_{02}, \eta_2$	29 (o)
Tyn – Calus 1975 [301]	different correlations for different types of systems	$M_2, V^1_{01}, V^1_{02}, P_1, P_2, \eta_2$	9
Nakanishi 1978 [302]	not for highly viscous solvents ($\eta < 20 - 30$ mPa · s)	$V^1_{01}, V^1_{02}, \eta_2$	14
Hayduk – Minhas 1982 [290]	different correlations for different types of systems	$M_2, V^1_{02}, V^1_{01}, \eta_2, P_1, P_2, (\sigma_i)$	10
Siddiqi – Lucas 1986 [293]	including dissolved gases	$V^1_{01}, V^1_{02}, \eta_2$	13 (o), 20 (w)

* Subscripts: 1 = solute; 2 = solvent.
** w = aqueous solutions; o = organic solutions.

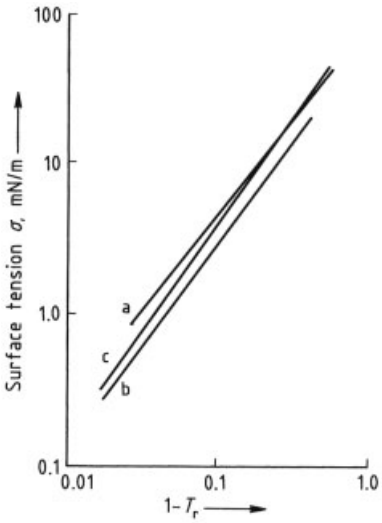

Figure 23. Variation of surface tension with temperature [304]
a) Acetic acid; b) Diethyl ether; c) Ethyl acetate

7. Surface Tension of Liquids

Surface and interfacial tension influence mass and heat transfer processes considerably. In this chapter only the liquid – vapor interface is discussed.

Surface tension σ is expressed in units of [N/m] or preferably in [mN/m] ($\equiv$[dyn/cm]). Surface tensions of organic liquids vary between 20 and 50 mN/m at moderate pressure ($p >> p_c$); experimental data can be found in the compilation of JASPER [303]. Water exhibits an extraordinarily high surface tension of 72 mN/m at room temperature. The variation of surface tension with temperature of some organic compounds is shown in Figure 23. Salt melts typically have surface tensions of ca. 100 – 200 mN/m and metal melts of 300 – 1000 mN/m. At low to moderate pressures ($p < p_c$) the influence of inert gases can be neglected.

7.1. Surface Tension of Pure Liquids

Two basically empirical categories of methods are used for the estimation of surface tension. One type uses the parachor P and the other employs the principle of corresponding states.

Parachor Methods. The parachor is a characteristic parameter which, according to its definition in Equation (7.1), is a function of surface tension. It has been defined by SUDGEN [305] as

$$P_i = \sigma_i^{1/4} \frac{M_i}{\varrho_i^l - \varrho_i^v} \tag{7.1}$$

where ϱ_i^l and ϱ_i^v are the liquid and saturated vapor densities of pure compound *i,* respectively. At temperatures considerably lower than the critical temperature the density of the vapor ϱ_i^v can be neglected compared to the density of the liquid ϱ_i^l, thus simplifying Equation (7.1) to

$$P_i = \sigma_i^{1/4} \frac{M_i}{\varrho_i^{\mathrm{l}}} = \sigma_i^{1/4} V_i^{\mathrm{l}} \tag{7.2}$$

The parachor is virtually independent of temperature; it is approximately an additive quantity which can be determined by summing up increments. Several such methods have been proposed. The method of QUALE [306] is recommended (for increments see Table 26).

Surface tensions predicted with this method usually show errors of less than 5 % for nonpolar and weakly polar compounds and 5 – 10 % for compounds with hydrogen bounds [307].

Example. Calculation of surface tension of acetamide, CH_3CONH_2, at 378.2 K from parachor estimated with the method of Quale [306]
Structural groups: 1 (CH_3-), 1 ($-CONH_2$) (see Table 26)
$P = 55.5 + 91.7 = 147.2\ \mathrm{cm^3\ g^{0.25}\ s^{-0.5}\ mol^{-1}}$
With $M = 59.068$ g/mol, $\varrho^{\mathrm{l}} = 977.05\ \mathrm{kg/m^3}$, $p_{0i}^{\mathrm{lv}} = 12.4$ hPa [308] and $\varrho^{\mathrm{v}} = p_{0i}^{\mathrm{lv}} M/(RT) = 0.023\ \mathrm{kg/m^3}$ from Equation (7.1): $\sigma = 0.0352$ N/m
Experimental: $\sigma = 0.03696$ N/m at 378.2 K [308]
Relative deviation: -4.9 %
(Deviation between Eqs. 7.1 and 7.2 = 0.01 %)

Corresponding States Methods. These methods are based upon a relation proposed by VAN DER WAALS [309]:

$$\sigma = k T_{\mathrm{c}}^{1/3} p^{1/32/3} (1 - T_{\mathrm{r}})^m \tag{7.3}$$

where k and m are constants, T_{c} is the critical temperature, p_{c} the critical pressure, and T_{r} the reduced temperature. Employing the Stiel polar factor χ (see Section 2.4), HAKIM et al. [310] transformed Equation (7.3) into a relation which is particularly useful for alcohols (p_{c} in bar, σ in N/m):

$$\sigma = p^{1/32/3} T^{1/31/3} \left(\frac{1 - T_{\mathrm{r}}}{0.4} \right)^m \frac{Q_{\mathrm{P}}(\chi, \omega)}{1000} \tag{7.4}$$

with the functions Q_{p} and m in terms of the Stiel polar factor χ and the acentric factor ω

$$Q_{\mathrm{P}} = 0.156 + 0.365\omega - 1.754\chi - 13.57\chi^2 - 0.506\omega^2 + 1.287\omega\chi \tag{7.5}$$

$$m = 1.21 + 0.5385\omega - 14.61\chi - 32.07\chi^2 - 1.656\omega^2 + 22.03\omega\chi \tag{7.6}$$

Errors are usually between 5 and 10 %.

Table 26. Quale's increments for the estimation of the parachor P_{i} [$\mathrm{cm^3\ g^{0.25}\ s^{-0.5}\ mol^{-1}}$] [306]

Group	Increment
Carbon – hydrogen	
C	9.0
H	15.5
CH_3	55.5
CH_2 in $-(CH_2)_n-$	
$n < 12$	40.0
$n > 12$	40.3
Alkyl groups	
1-Methylethyl	133.3
1-Methylpropyl	171.9
1-Methylbutyl	211.7
2-Methylpropyl	173.3
1-Ethylpropyl	209.5
1,1-Dimethylethyl	170.4
1,1-Dimethylpropyl	207.5
1,2-Dimethylpropyl	207.9
1,1,2-Trimethylpropyl	243.5
C_6H_5	189.6
Special groups	
$-COO-$	63.8
$-COOH$	73.8
$-OH$	29.8
$-NH_2$	42.5
$-O-$	20.0
$-NO_2$	74
$-NO_3$ (nitrate)	93
$-CO(NH_2)$	91.7
$R-[-CO-]-R'$ (ketone)	
Number of C atoms in $R + R'$:	
$R + R' = 2$	51.3
$R + R' = 3$	49.0
$R + R' = 4$	47.5
$R + R' = 5$	46.3
$R + R' = 6$	45.3
$R + R' = 7$	44.1
$-CHO$	66
O (not noted above)	20
N (not noted above)	17.5
S	49.1
P	40.5
F	26.1
Cl	55.2
Br	68.0
I	90.3
Ethylenic bonds	
Terminal	19.1
2,3-position	17.7
3,4-position	16.3
Triple bond	40.6
Ring closure	
Three-membered	12
Four-membered	6.0
Five-membered	3.0
Six-membered	0.8

A simpler relation based on Equation (7.3) was proposed by BROCK and BIRD [311]; it should only be used for nonpolar or weakly polar liquids.

Example. Calculation of surface tension of 1-butanol, $CH_3CH_2CH_2CH_2OH$, at 303.2 K according to Hakim et al. [310].
$M = 74.122$ g/mol, $p_c = 44.13$ bar, $T_c = 563.0$ K, $\chi = -0.07$, $\omega = 0.593$ [308].
From Equation (7.6): $m = 0.8981$.
From Equation (7.5): $Q_p = 0.1974$.
From Equation (7.4): $\sigma = 0.02314$ N/m.
Experimental: $\sigma = 0.02378$ N/m [308].
Relative deviation: $-2.7\,\%$.

7.2. Surface Tension of Liquid Mixtures

Organic Systems. Surface tensions of mixtures of organic compounds can be estimated by the simple mixing rule:

$$\sigma_m = \sum x_i \sigma_i \tag{7.7}$$

where x_i is the mole fraction of component i or by the more general relation

$$\sigma_m^\beta = \sum x_i \sigma_i^\beta \tag{7.8}$$

with the exponent β.

An exponent $\beta = -1$ is sometimes used:

$$\sigma_m = \left(\sum \frac{x_i}{\sigma_i}\right)^{-1} \tag{7.9}$$

Equation (7.2) can be extended for nonaqueous mixtures:

$$\sigma_m = \left[\sum P_i \left(\frac{\varrho_m^l x_i}{M}\right)\right]^4 \tag{7.10}$$

where x_i and y_i are the mole fractions of compound i in the liquid and vapor phases, respectively, and

$$\frac{1}{\varrho_m} = \left(\sum x_i/\varrho_i\right) \tag{7.11}$$

Again, the parachor of the pure compounds is employed (see Section 7.1). When experimental surface tension data for the mixture are available, P_i can also be used as a correlation parameter. The accuracy of Equation (7.10) is usually $< 10\,\%$. If the parachor is applied as an adjustable parameter using experimental information, precision is of course much higher. The concentration dependence of the surface tension of some binary organic solutions is shown in Figure 24.

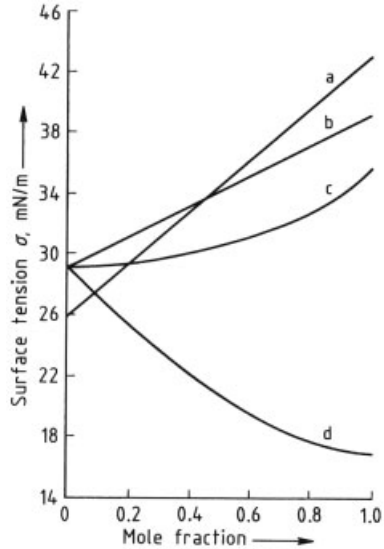

Figure 24. Surface tension of binary organic mixtures at 298 K [311]
a) $\underline{Nitrobenzene}$ – carbon tetrachloride;
b) $\underline{Acetophenone}$ – benzene;
c) $\underline{Nitrobenzene}$ – benzene;
d) $\underline{Diethylether}$ – benzene
The mole fraction refers to the underlined component.

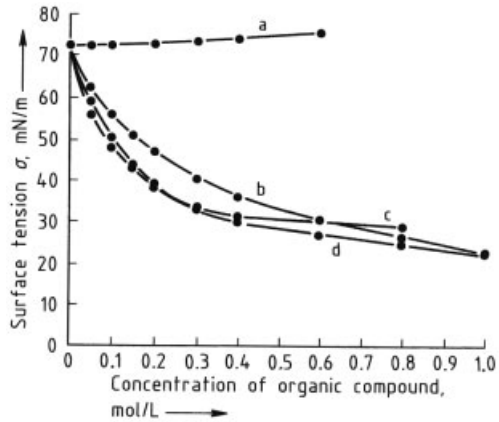

Figure 25. Surface tension of aqueous solutions of organic compounds (data from [312])
a) Sucrose; b) Methanol; c) Acetonitrile; d) Ethanol

A large number of more sophisticated methods have been published based on the principle of corresponding states or on phenomenological or statistical thermodynamics. These methods are discussed in detail in [307].

Aqueous Systems. The surface tension of water is strongly affected by small amounts of dissolved organic compounds because hydrophobic molecules are repelled by the water molecules and are therefore enriched at the surface. The surface tension of aqueous solutions of organics therefore depends on the hydrophobicity of the organic solute with a highly nonlinear influence of solute concentration. In general, the surface tension of water is greatly decreased by hydrophobic solutes (Fig. 25).

For estimating the effect of organic solutes on the surface tension of water MEISSNER et al. [313] have proposed a method which is based on the following equation:

$$\sigma_{\mathrm{m}} = \sigma_{0\mathrm{w}} \left[1 - 0.411 \log \left(1 + \frac{x_i}{A}\right)\right] \tag{7.12}$$

where $\sigma_{0\mathrm{w}}$ is the surface tension of pure water, x_i the molar concentration of the organic component, and A is a substance-specific constant. Values of A ($\times 10^4$) for a number of organic substances follow:

Propionic acid	26	Ethyl propionate	3.1
1-Propanol	26	Propyl acetate	3.1
2-Propanol	26		
Methyl acetate	26	Valeric acid	1.7
		Isovaleric acid	1.7
Propylamine	19	Pentanol	1.7
Methyl ethyl ketone	19	Isopentanol	1.7
Butyric acid	7	Propyl propionate	1.0
Isobutyric acid	7	Caproic acid	0.75
Butanol	7	Heptanoic acid	0.17
Isobutanol	7	Octanoic acid	0.034
Propyl formate	8.5	Decanoic acid	0.0025
Ethyl acetate	8.5		
Methyl propionate	8.5		
Diethyl ketone	8.5		

The accuracy of this method is reported to be better than 15 %. Another more precise but more complicated estimation method was suggested by TAMURA et al. [314]. This method is applicable to binary mixtures only and is discussed in detail in [307].

8. References

General References

Methods

1. S. Bretsznajder: *Prediction of Transport and Other Physical Properties of Fluids,* Pergamon Press, Oxford 1971.
2. T. E. Danner, R. P. Daubert (Eds.): *Manual for Predicting Chemical Process Design Data,* Pennsylvania State University 1987.
3. G. Hecht et al.: *Berechnung thermodynamischer Stoffwerte von Gasen und Flüssigkeiten,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1966.
4. C. Jochum, M. G. Hicks, J. Sunkel (eds.): *Physical Property Prediction in Organic Chemistry,* Springer Verlag, Berlin 1988.
5. R. C. Reid, T. K. Sherwood: *The Properties of Gases and Liquids,* 2nd ed., McGraw-Hill, New York 1966.
6. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
7. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
8. Z. Sterbacek, B. Biskup, P. Tausk: *Calculation of Properties Using Corresponding-State Methods,* Elsevier, Amsterdam 1979.
9. S. Weiß et al. (eds.): *Verfahrenstechnische Berechnungsmethoden,* part 7: Stoffwerte, VCH Verlagsgesellschaft, Weinheim 1986.

General References

General Data Collections

10. D. Behrens, R. Eckermann (eds.): *Chemistry Data Series,* Dechema, Frankfurt/Main 1977.
11. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* 4th ed., **vols. 1 – 3**, Springer Verlag, Berlin 1997.
12. R. P. Daubert, T. E. Danner: Design Institute for Physical Property Data (DIPPR)/American Institute of Chemical Engineers (AIChE): *Data Compilation Tables of Properties of Pure Compounds,* Pennsylvania State University 1985.
13. H. Landolt, R. Börnstein: *Zahlenwerte und Funktionen aus Naturwissenschaft und Technik,* 6th ed., Springer Verlag, Berlin 1950.
14. *Physical Sciences Data,* vols. 1 – 44, Elsevier, Amsterdam 1987.
15. Thermodynamics Research Centre (TRC): *Selected Values of Properties of Chemical Compounds,* Texas A & M University, College Station, Texas, 1994.
16. Y. S. Touloukian, C. Y. Ho (eds.): *CINDAS Data Series on Material Properties,* vols. 1 – Hemisphere, Washington 1981.
17. Y. S. Touloukian, C. Y. Ho: *Thermophysical Properties of Matter,* vols. 1 – 13, IFI/Plenum Press, New York 1970.

18. VDI-Wärmeatlas, 7th ed., VDI-Verlag, Düsseldorf 1994.
19. D. R. Lide (ed.): *Handbook of Chemistry and Physics,* 76th ed., CRC Press, Boca Raton, Florida, 1995.

General References
Universal Handbooks
20. *Beilsteins Handbuch der Organischen Chemie,* 4th ed., Springer Verlag, Berlin 1958.
21. *Gmelin Handbook of Inorganic and Organometallic Chemistry,* 8th ed.,VCH Verlagsgesellschaft, Weinheim 1990.

General References
Thermodynamic Data of Pure Compounds
22. O. Knacke, O. Kubaschewski, K. Hesselmann: *Thermochemical Properties of Inorganic Substances,* vol. 1, Springer Verlag, Berlin 1973, vol. 2, Verlag Stahleisen, Düsseldorf 1977.
23. T. Barin: *Thermochemical Data of Pure Substances,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1995.
24. M. W. Chase et al.: *Joint Army Navy Air Force (JANAF) Thermochemical Tables,* 3rd ed., American Chemical Society, American Institute of Physics, Washington, D.C., 1986.
25. J. D. Cox, G. Pilcher: *Thermochemistry of Organic and Organometallic Compounds,* Academic Press, London 1970.
26. J. B. Pedley, R. D. Naylor, S. P. Kirby: *Thermochemical Data of Organic Compounds,* 2nd ed., Chapman and Hall, New York 1986.
27. F. D. Rossini: *Selected Values of Chemical Thermodynamic Properties,* National Bureau of Standards, Circular 500, Washington 1952.
28. D. R. Stull, G. C. Sinke: "Thermodynamic Properties of the Elements," in: *Advances in Chemistry Series,* vol. 18, American Chemical Society, Washington, D.C., 1956.
29. D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds,* Krieger, Malabar, Fla. 1987.
30. A. L. Suris: *Handbook of Thermodynamic High Temperature Process Data,* Hemisphere, Washington 1987.
31. A. Tamir, E. Tamir, K. Stephan: *Heats of Phase Change of Pure Components and Mixtures,* Elsevier, Amsterdam 1983.
32. Y. S. Touloukian et al.: *Thermodynamic and Transport Data Properties of Gases, Liquids and Solids,* McGraw-Hill, New York 1959.
33. Y. S. Touloukian, T. Makita: *Specific Heat. Nonmetallic Liquids and Gases,* IFI/Plenum Press, New York 1970; Suppl. 1976.
34. Y. S. Touloukian, E. H. Buyco: *Specific Heat. Nonmetallic Solids,* IFI/Plenum Press, New York 1970.
35. Further reference: [55]

General References
Thermodynamic Data of Mixtures
36. W. Arlt, M. E. A. Macedo, P. Rasmussen, J. M. Sorensen: *Liquid – Liquid Equilibrium Data Collection,Chemistry Data Series,* **vol. V, parts 1 – 4,** Dechema, Frankfurt/Main 1979 – 1987.
37. C. Christensen, J. Gmehling, P. Rasmussen, U. Weidlich, T. Holderbaum: *Heats of Mixing Data Collection, Chemistry Data Series,* **vol. III, parts 1 – 4,** Dechema, Frankfurt/Main 1984 – 1991.
38. H. Engels: *Phase Equilibria and Enthalpies of Electrolyte Solutions, Chemistry Data Series,* **vol. XI,** Dechema, Frankfurt/Main 1990.
39. J. Gmehling, U. Onken et al.: *Vapor – Liquid Equilibrium Data Collection,* **vol. 1: part 1 (1977), part 1 a (1981), part 1 b (1988), part 2 a (1985, reprint), part 2 b (1987), part 2 c (1982), part 2 d (1982), part 2 e (1988), part 2 f (1990), part 3/4 (1979), part 3 a (1993), part 3 b (1993), part 4 a (1994), part 5 (1982), part 5 a (in prep.), part 6 a (1980), part 6 b (1980), part 6 c (1984), part 7 (1980), part 8 (1984)**, Dechema, Frankfurt/Main 1977 – 1994.
40. J. Gmehling et al.: *Activity Coefficients at Infinite Dilution, Chemistry Data Series,* **vol. IX, parts 1 – 4,** Dechema, Frankfurt/Main 1986 – 1994.
41. J. Gmehling, J. Menke, J. Krafczyk, K. Fischer: *Azeotropic Data,* parts I – II, VCH Verlagsgesellschaft, Weinheim 1994.
42. IUPAC: *Solubility Data Series,* Pergamon Press, Oxford 1979.
43. H. Knapp, R. Döring, L. Öllrich, U. Plöcker, J. M. Prausnitz: *Vapor – Liquid Equilibria for Mixtures of Low-Boiling Substances, Chemistry Data Series,* **vol. VI, parts 1 – 4,** Dechema, Frankfurt/Main 1982 – 1989.
44. H. Knapp, M. Teller, R. Langhorst: *Solid – Liquid Equilibrium Data Collection, Chemistry Data Series,* **vol. VIII,** Dechema, Frankfurt/Main 1987.
45. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Organic Compounds,* 4th ed.,

vols. 1 – 2, Princeton [vol. 1(1958), vol. 2 (1965)].
46. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* Pergamon Press, Oxford 1963, reprinted 1979.
47. N. K. Voskresenskaya (ed.): *Handbook of Solid – Liquid Equilibria in Systems of Anhydrous Inorganic Salts,* Keter Press, Jerusalem 1970.
48. J. F. Zemaitis, D. M. Clark, M. Rafal, N. C. Scrivner: *Handbook of Aqueous Electrolyte Thermodynamics,* AIChE, New York 1986.
49. Further reference: [31]

General References
Critical Data
50. D. Ambrose: "Correlation and Estimation of Vapour Liquid Critical Properties, I. Critical Temperatures of Organic Compounds," National Physical Laboratory, Teddington, *NPL Rep. Chem.* **92** (1978) corr. March 1981, Teddington, *NPL Rep. Chem.* **98** (1979).
51. K. H. Simmrock, R. Janowsky, A. Ohnsorge: *Critical Data of Pure Substances, Chemistry Data Series,* **vol. II, parts 1 – 2,** Dechema, Frankfurt/Main 1986.

General References
Vapor Pressure
52. T. Boublik, V. Fried, E. Hala: *The Vapor Pressures of Pure Substances,* 2nd ed., Elsevier, Amsterdam 1984.
53. S. Ohe: *Computer Aided Data Book of Vapor Pressure,* Data Book Publishing, Tokyo 1976.
54. I. Wichterle, J. Linek: *Antoine Vapour Pressure Constants of Pure Compounds,* Academia, Praha 1971.
55. B. J. Zwolinski, R. C. Wilhoit: *Handbook of Vapour Pressures and Heats of Vaporisation of Hydrocarbon and Related Compounds,* Thermodynamic Research Center, Department of Chemistry, Texas 1971.

General References
pVT Behavior
56. J. Cholinski, A. Szafranski et al.: *Second Virial Coefficients for Organic Compounds,* Institute of Physical Chemistry, Polish Academy of Sciences, Warszawa 1985.
57. J. H. Dymond, E. B. Smith: *The Virial Coefficients of Pure Gases and Mixtures* : A Critical Compilation, Clarendon Press, Oxford 1980.

General References
Transport Data
58. K. Stephan, T. Heckenberger: *Thermal Conductivity and Viscosity Data of Fluid Mixtures, Chemistry Data Series,* **vol. X,** Dechema, Frankfurt/Main 1989.
59. K. Stephan, K. Lucas: *Viscosity of Pure Dense Fluids,* Plenum Press, New York 1979.
60. D. S. Viswanath, G. Natarjan: *Data Book on the Viscosity of Liquids,* Hemisphere, New York 1989.

General References
Selected Compounds
61. J. Barthel, R. Neuder: *Electrolyte Data Collection: Conductivities, Transference Numbers, Limiting Ionic Conductivities, Chemistry Data Series,* **vol. XII, parts 1 – 3,** Dechema, Frankfurt/Main 1992 – 1997.
62. R. R. Dreisbach: "Physical Properties of Chemical Compounds," vols. I – III, in: *Advances in Chemistry Series,* vols. 15, 22, 29, American Chemical Society, Washington, DC [vol. I (1955), vol. II (1959), vol. III (1961)].
63. E. W. Flick: *Industrial Solvents Handbook,* 3rd ed., Noyes Data Corporation, New Jersey 1985.
64. A. L. Horvath: *Physical Properties of Inorganic Compounds,* Fletcher & Son, Norwich, UK, 1975.
65. J. A. Riddick, W. B. Bunger, T. K. Sakano: *Organic Solvents. Physical Properties and Methods of Purification,* 4th ed. **, vol. II,** Wiley, New York 1986.
66. R. M. Stephenson, S. Malanowski: *Handbook of the Thermodynamics of Organic Compounds,* Elsevier, Amsterdam 1987.
67. Thermodynamics Research Centre (TRC): *Selected Values of Properties of Hydrocarbons and Related Compounds,* vol. 1, Texas A & M University, College Station, Texas, 1977.
68. Thermodynamics Research Centre (TRC): *TRC Thermodynamic Tables–Hydrocarbons,* vols. 1 – 11, Texas A & M University, College Station, Texas, 1942.
69. Thermodynamics Research Centre (TRC): *TRC Thermodynamic Tables–Non-Hydrocarbons,* vols. 1 – 8, Texas A & M University, College Station, Texas, 1955.
70. J. Timmermanns: *The Physico-Chemical Constants of Pure Organic Compounds,* Elsevier, New York [vol. 1 (1950), vol. 2 (1965)].

71. Y. S. Touloukian (ed.): *Thermophysical Properties of High Temperature Solid Materials,* vols. 1 – 6, MacMillan, New York 1967.
72. W. Hao, H. S. Elbro, P. Alessi: *Polymer Solution Data Collection, Chemistry Data Series,* **vol. XIV, parts 1 – 3,** Dechema, Frankfurt/Main 1982 – 1992.

General References

References for Chapter 1

73. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* 4th ed., **vols. 1 – 3,** Springer Verlag, Berlin 1997.
74. D. R. Lide (ed.): *Handbook of Chemistry and Physics,* 76th ed., CRC Press, Boca Raton, Florida, 1995.
75. H. Landolt, R. Börnstein: a. " Eigenschaften der Materie in Ihren Aggregatzuständen," *Zahlenwerte und Funktionen aus Naturwissenschaft und Technik,* 6th ed., **vol. II** Springer Verlag, Berlin 1956; b. " Makroskopische und technische Eigenschaften der Materie," *Zahlenwerte und Funktionen aus Naturwissenschaft und Technik,* 6th ed., new series, **Group IV** Springer Verlag, Berlin 1974.
76. D. Behrens, R. Eckermann (eds.): *Chemistry Data Series,* Dechema, Frankfurt/Main 1977.
77. U. Onken, J. Rarey-Nies, J. Gmehling: "The Dortmund Data Bank: A Computerized System for Retrieval, Correlation, and Prediction of Thermodynamic Properties of Mixtures," *Int. J. Thermophys.* **10** (1989) 739 – 747.
78. C. Jochum, M. G. Hicks, J. Sunkel (eds.): *Physical Property Prediction in Organic Chemistry,* Springer Verlag, Berlin 1988.
79. Fachinformationszentrum (FIZ) Chemie, Postfach 12 60 50, D-10583 Berlin, Germany; http://www.fiz-chemie.de DECHEMA e.V., Informationssysteme und Datenbanken, Postfach 15 01 04, D-60061 Frankfurt/Main, Germany; http://www.dechema.de Design Institute for Physical Property Data (DIPPR), STN International, 2540 Olentangy River Road, P.O.Box 2228, Columbus, Ohio, 43202, USA. STN International, Postfach 2465, D-76012 Karlsruhe, Germany; http://www.fiz-karlsruhe.de STN International, Tokyo Headquarters, 5-3 Yonbancho, Chiyoda-ku, Tokyo 102, Japan. Physical Properties Data Service (PPDS), Institution of Chemical Engineers, 165 – 171 Railway Terrace, Rubgy, CV21 3HQ, England.
80. W. Kistiakowsky, *Z. Phys. Chem.* **107** (1923) 65.
81. J. Timmermanns: *The Physico-Chemical Constants of Pure Organic Compounds,* Elsevier, New York [vol. 1 (1950), vol. 2 (1965)].
82. Thermodynamics Research Centre (TRC): *Selected Values of Properties of Chemical Compounds,* Texas A & M University, College Station, Texas, 1982.
83. J. A. Riddick, W. B. Bunger, T. K. Sakano: *Organic Solvents. Physical Properties and Methods of Purification,* 4th ed., vol. II, Wiley, New York 1986.

General References

References for Chapter 2

84. L. I. Stiel, G. Thodos, *Chem. Eng. Data Ser.* **7** (1962) 234.
85. J. S. Rowlinson: *Liquids and Liquid Mixtures,* 2nd ed., Butterworth, London 1969.
86. J. O. Hirschfelder, C. F. Curtiss, R. B. Bird: *Molecular Theory of Gases and Liquids,* 4th print, Wiley, New York 1967.
87. L. C. Nelson, E. F. Obert, *Trans. ASME* **76** (1954) 1057.
88. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
89. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
90. A. L. Lydersen: "Estimation of Critical Properties of Organic Compounds," *Univ. Wisconsin Coll. Eng., Eng. Exp. Stn. Rept. 3,* Madison, Wisc., April 1955.
91. K. G. Joback, MS Thesis, Massachusetts Institute of Techn., Cambridge, Mass., 1984.
92. K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **57** (1987) 233.
93. J. Timmermanns: *The Physico-Chemical Constants of Pure Organic Compounds,* Elsevier, New York [vol. 1 (1950), vol. 2 (1965)].
94. D. Ambrose: "Correlation and Estimation of Vapour Liquid Critical Properties, I. Critical Temperatures of Organic Compounds," *NPL Rep. Chem. UK Natl. Phys. Lab. Div. Chem. Stand.* **92** (1978) (corr. March 1980).
95. D. Ambrose, *NPL Rep. Chem. UK Natl. Phys. Lab. Div. Chem. Stand.* **98** (1979).
96. R. F. Fedors, *Chem. Eng. Commun.* **16** (1982) 149.
97. L. Constantinou, R. Gani, *AIChE J.* **40** (1994) 1697.

98. A. P. Kudchadkar, G. H. Alani, B. J. Zwolinski, *Chem. Rev.* **68** (1968) 659.
99. K. H. Simmrock, R. Janowsky, A. Ohnsorge: *Critical Data of Pure Substances, Chemistry Data Series,* **vol. II, part I,** Dechema, Frankfurt/Main 1986.
100. J. F. Mathews, *Chem. Rev.* **72** (1972) 71.
101. W. C. Edmister, *Pet. Refin.* **37** (1958) no. 4, 173.
102. S. Weiß et al. (eds.): *Verfahrenstechnische Berechnungsmethoden,* part 7: Stoffwerte, VCH Verlagsgesellschaft, Weinheim 1986.
103. L. Constantinou, R. Gani, J. P. O'Connell, *Fluid Phase Equilib.* **103** (1995) 11.

General References

References for Chapter 3

104. J. S. Rowlinson: *Liquids and Liquid Mixtures,* 2nd ed., Butterworth, London 1969.
105. J. M. Prausnitz, R. N. Lichtenthaler, E. Gomes de Azevedo: *Molecular Thermodynamics of Fluid-Phase Equilibria,* Prentice-Hall, Englewood Cliffs, N.Y., 1986.
106. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
107. L. C. Nelson, E. F. Obert, *Trans. ASME* **76** (1954) 1057.
108. G. Soave, *Chem. Eng. Sci.* **27** (1972) 1197.
109. D. Y. Peng, D. B. Robinson, *Ind. Eng. Chem. Fundam.* **15** (1976) 59.
110. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
111. J. H. Dymond, E. B. Smith: *The Virial Coefficients of Pure Gases and Mixtures: A Critical Compilation,* Clarendon Press, Oxford 1980.
112. K. S. Pitzer, R. F. Curl, *J. Am. Chem. Soc.* **79** (1957) 2369.
113. C. Tsonopoulos, *AIChE J.* **20** (1974) 263 – 272; **21** (1975) 827.
114. J. G. Hayden, J. P. O'Connell, *Ind. Eng. Chem. Process Des. Dev.* **14** (1975) no. 3, 209.
115. B. J. Lee, M. G. Kesler, *AIChE J.* **21** (1975) 510.
116. U. Plöcker, H. Knapp, J. M. Prausnitz, *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) 324.
117. E. Bender: *Die Berechnung von Phasengleichgewichten mit der thermischen Zustandsgleichung,* Habilitationsschrift, Ruhr Universität, Bochum 1971.
118. A. Polt, Ph. D. thesis, Universität Kaiserslautern 1987.
119. J. Partington: *An Advanced Treatise on Physical Chemistry,* vol. I: Fundamental Principles: The Properties of Gases; Longmans, Green & Co., New York 1949.
120. F. Schroeder in [119]
121. J. A. Riddick, W. B. Bunger: *Organic Solvents, Physical Properties and Methods of Purification,* 3rd ed., **vol. II,** 1970; 4th ed., 1986, J. Wiley, New York.
122. L. Constantinou, R. Gani, J. P. O'Connell, *Fluid Phase Equilib.* **103** (1995) 11.
123. R. D. Gunn, T. Yamada, *AIChE J.* **17** (1971) 1341.
124. R. W. Hankinson, G. H. Thomson, *AIChE J.* **25** (1979) 653.
125. L. C. Yen, S. S. Woods, *AIChE J.* **16** (1966) 95.
126. V. L. Bhirud, *AIChE J.* **24** (1978) 880, 1127.
127. S. Weiß et al. (eds.): *Verfahrenstechnische Berechnungsmethoden,* part 7: Stoffwerte, VCH Verlagsgesellschaft, Weinheim 1986.
128. J. M. Prausnitz, R. N. Lichtenthaler, E. Gomes de Azevedo: *Molecular Thermodynamics of Fluid-Phase Equilibria,* 2nd ed., Prentice-Hall, Englewood Cliffs, N.Y., 1986
129. W. B. Kay, *Ind. Eng. Chem. Ind. Ed.* **28** (1936) 1014.
130. J. M. Prausnitz, R. D. Gunn, *AIChE J.* **4** (1958) 430, 494.
131. J. Joffe, *Ind. Eng. Chem. Fundam.* **10** (1971) 532.
132. S. Angus et al. (eds.): *IUPAC International Thermodynamic Tables of the Fluid State: Carbon Dioxide,* Pergamon Press, Oxford 1973.
133. K. G. Joback, MS thesis, Massachusetts Institute of Techn., Cambridge, Mass., 1984.
134. K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **57** (1987) 233.
135. S. W. Benson et al.: "Additive Rules for the Estimation of Thermochemical Properties," *Chem. Rev.* **69** (1969) 279; *J. Phys. Colloid Chem.* **52** (1948) 1060. S. W. Benson: *Thermochemical Kinetics,* Wiley, New York 1968.
136. C. J. Dobratz, *Ind. Eng. Chem. Ind. Ed.* **33** (1941) 759.
137. D. N. Rihani, L. K. Doraiswamy, *Ind. Eng. Chem. Fundam.* **4** (1965) 17.
138. T.-P. Thinh, T. K. Trong, *Can. J. Chem. Eng.* **54** (1976) 344.
139. Y. Yoneda, *Bull. Chem. Soc. Jpn.* **52** (1979) 1297.
140. B. K. Harrison, W. H. Seaton, *Ind. Eng. Res.* **27** (1988) 1536 – 1540.

141. C. F. Chueh, A. C. Swanson, *Chem. Eng. Prog.* **69** (1973) no. 7, 83; *Can. J. Chem. Eng.* **51** (1973) 596.
142. A. Missenard, *Compte Rendue* **260** (1965) 5521.
143. V. Ruzicka, Jr., E. S. Domalski, *J. Phys. Chem. Ref. Data* **22** (1993) 579, 619.
144. A. Bondi, *Ind. Eng. Chem. Fundam.* **5** (1966) 443.
145. T.-F. Yuan, L. J. Stiel, *Ind. Eng. Chem. Fundam.* **9** (1970) 393.
146. A. S. Teja, *J. Chem. Eng. Data* **28** (1983) 83.

General References

References for Chapter 4

147. S. Weiß et al. (eds.): *Verfahrenstechnische Berechnungsmethoden,* part 7: Stoffwerte, VCH Verlagsgesellschaft, Weinheim 1986.
148. D. Ambrose, J. F. Counsell, A. J. Davenport, *J. Chem. Thermodyn.* **6** (1974) 693.
149. T. Boublik, V. Fried, E. Hala: *The Vapor Pressures of Pure Substances,* 2nd ed., Elsevier, Amsterdam 1984.
150. B. J. Lee, M. G. Kesler, *AIChE J.* **21** (1975) 510.
151. M. Gomez-Nieto, G. Thodos, *Ind. Eng. Chem. Fundam.* **16** (1977) 254.
152. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
153. J. A. Riddick, W. B. Bunger: *Organic Solvents, Physical Properties and Methods of Purification,* 3rd ed., vol. II, 1970; 4th ed., 1986, J. Wiley, New York.
154. Y. Ogata, M. Tsuchida, *Ind. Eng. Chem.* **49** (1957) no. 3, 415.
155. K. G. Joback, MS thesis, Massachusetts Institute of Techn., Cambridge, Mass., 1984.
156. K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **57** (1987) 233.
157. L. Constantinou, R. Gani, *AIChE J.* **40** (1994) 1697.
158. J. E. Haggenmacher, *J. Am. Chem. Soc.* **68** (1946) 1633.
159. A. Vetere in [88], p. 209].
160. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
161. A. Tamir, E. Tamir, K. Stephan: *Heats of Phase Change of Pure Components and Mixtures,* Elsevier, Amsterdam 1983.
162. L. Riedel, *Chem. Ing. Tech.* **26** (1954) 679.
163. N. H. Chen, *J. Chem. Eng. Data* **10** (1965) 207.
164. W. Wagner, *Cryogenics* **13** (1973) 470.
165. R. Kobayashi, J. Magee, A. Sivaraman, *Fluid Phase Equilib.* **16** (1984) 1.
166. K. M. Watson, *Ind. Eng. Chem. Ind. Ed.* **35** (1943) 398.
167. G. M. Wilson, *J. Amer. Chem. Soc.* **86** (1964) 127.
168. H. Renon, J. M. Prausnitz, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 413; *AIChE J.* **15** (1969) 785.
169. D. S. Abrams, J. M. Prausnitz, *AIChE J.* **21** (1975) 116.
170. O. Redlich, E. L. Derr, G. Pierotti, *J. Am. Chem. Soc.* **81** (1959) 2283.
171. G. M. Wilson, C. H. Deal, *Ind. Eng. Chem. Fundam.* **1** (1962) 20.
172. E. L. Derr, C. H. Deal, *Ind. Chem. Eng. Symp. Ser.* **3** (1969) 40.
173. K. Tochigi, D. Tiegs, J. Gmehling, K. Kojima, *J. Chem. Eng. Jpn.* **23** (1990) 453.
174. A. Fredenslund, R. L. Jones, J. M. Prausnitz, *AIChE J.* **21** (1975) 1086.
175. A. Fredenslund, J. Gmehling, P. Rasmussen: *Vapour Liquid Equilibria Using UNIFAC-A Group-Contribution Method,* Elsevier, Amsterdam – New York 1977.
176. A. Bondi: *Physical Properties of Molecular Liquids, Crystals and Glasses,* Wiley, New York 1968.
177. J. Gmehling, P. Rasmussen, A. Fredenslund, *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 118.
178. H. Hansen et al., *Ind. Eng. Chem. Res.* **30** (1991) 2352.
179. B. Larsen, P. Rasmussen, A. Fredenslund, *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 2274.
180. U. Weidlich, J. Gmehling, *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 1372.
181. J. Gmehling, J. Li, M. Schiller, *Ind. Eng. Chem. Res.* **32** (1993) 178.
182. H. V. Kehaian, S. I. Sandler, *Fluid Phase Equilib.* **17** (1984) 139.
183. B. Kolbe, J. Gmehling, U. Onken, *Inst. Chem. Eng. Symp. Ser.* **56** (1979); *Ber. Bunsen Ges. Phys. Chem.* **83** (1979) 1133.
184. T. Oishi, J. M. Prausnitz, *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) no. 3, 333.
185. G. Nocon, U. Weidlich, J. Gmehling, U. Onken, *Ber. Bunsen Ges. Phys. Chem.* **87** (1983).
186. G. Nocon, U. Weidlich, J. Gmehling, J. Menke, U. Onken, *Fluid Phase Equilib.* **13** (1983) 381.
187. K. E. Gubbins, *Chem. Eng. Prog.* **85** (1989) 38.

188. M. Luckas, K. Lucas, *Fluid Phase Equilib.* **45** (1989) 7 – 13.
189. H. Knapp et al.: *Vapor – Liquid Equilibria for Mixtures of Low-Boiling Substances, Chemistry Data Series,* **vol. VI,** Dechema, Frankfurt/Main 1982.
190. M.-J. Huron, J. Vidal, *Fluid Phase Equilib.* **3** (1979) 255.
191. S. Dahl, M. L. Michelsen, *AIChE J.* **36** (1990) 1829.
192. M. L. Michelsen, *Fluid Phase Equilib.* **60** (1990) 213.
193. T. Holderbaum, J. Gmehling, *Fluid Phase Equilib.* **70** (1991) 251.
194. W. B. Whiting, J. M. Prausnitz, *Fluid Phase Equilib.* **9** (1982) 119. D. Luedecke, J. M. Prausnitz, *Fluid Phase Equilib.* **22** (1985) 1.
195. A. Anderko, *Fluid Phase Equilib.* **45** (1991) 39.
196. W. G. Chapman, K. E. Gubbins, G. Jackson, M. Radosz, *Ind. Eng. Chem. Res.* **29** (1990) 1709.
197. J. Gmehling, J. Li, K. Fischer, *Fluid Phase Equilib.* **141** (1997) 113.
198. S. Skjøld-Jørgensen, *Fluid Phase Equilib.* **16** (1984) 317.
199. S. Skjøld-Jørgensen, *Ind. Eng. Chem. Res.* **27** (1988) 110.
200. K. Tochigi, K. Kurihara, K. Kojima, *Ind. Eng. Chem. Res.* **29** (1990) 2142.
201. F. Magnussen, P. Rasmussen, A. Fredenslund, *Ind. Eng. Chem. Process Des. Dev.* **20** (1981) 331.
202. J. Gmehling, *Fluid Phase Equilib.* **107** (1995) 1.
203. F. Zielke, D. A. Lempe, *Fluid Phase Equilib.* **141** (1997) 63.
204. C. Hansch, D. A. Leo, *Substituent Constants for Correlation Analysis in Chemistry and Biology,* Wiley, New York 1979.
205. G. Wienke, J. Gmehling, *Toxicol. Environ. Chem.* **65** (1998) 57.
206. A. Jakob, R. Joh, C. Rose, J. Gmehling, *Fluid Phase Equilib.* **113** (1995) 117.
207. J. Gmehling, T. F. Anderson, J. M. Prausnitz, *Ind. Eng. Chem. Fundam.* **17** (1978).

General References

References for Chapter 5

208. H. Landolt, R. Börnstein: *Zahlenwerte und Funktionen aus Naturwissenschaft und Technik,* 6th ed., Springer Verlag, Berlin 1950.
209. G. J. Janz: *Thermodynamic Properties of Organic Compounds,* Academic Press, New York 1967.
210. M. W. Chase et al.: *Joint Army Navy Air Force (JANAF) Thermochemical Tables,* 3rd ed., Journal of Physical and Chemical Reference Data, Suppl. 14,1, American Chemical Society, American Institute of Physics, Washington, D.C., 1986.
211. J. D. Cox, G. Pilcher: *Thermochemistry of Organic and Organometallic Compounds,* Academic Press, London 1970.
212. F. D. Rossini: *Selected Values of Chemical Thermodynamic Properties,* National Bureau of Standards, Circular 500, Washington 1952.
213. D. R. Stull, G. C. Sinke: "Thermodynamic Properties of the Elements," in: *Advances in Chemistry Series,* vol. 18, American Chemical Society, Washington, D.C., 1956.
214. D. R. Stull, E. F. Westrum, G. C. Sinke: *The Chemical Thermodynamics of Organic Compounds,* Wiley, New York 1969.
215. Thermodynamics Research Centre (TRC): *Selected Values of Properties of Chemical Compounds,* Texas A & M University, College Station, Texas, 1982 *Selected Values of Properties of Hydrocarbons and Related Compounds,* **vols. 1,** Texas A & M University, College Station, Texas, 1977; *TRC Thermodynamic Tables – Non-Hydrocarbons,* **vols. 1 – 8,** Texas A & M University, College Station, Texas, 1955; *TRC Thermodynamic Tables – Hydrocarbons,* **vols. 1 – 11,** Texas A & M University, College Station, Texas, 1942.
216. K. G. Joback: MS thesis, Massachusetts Institute of Techn., Cambridge, Mass., 1984.
217. K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **57** (1987) 233.
218. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
219. S. W. Benson: *Thermochemical Kinetics,* Wiley, New York 1968.
220. S. W. Benson et al., *Chem. Rev.* **69** (1969) 279.
221. Y. Yoneda, *Bull. Chem. Soc. Jpn.* **52** (1979) 1297.
222. J. L. Franklin, *Ind. Eng. Chem.* **41** (1949) 1070; *J. Chem. Phys.* **21** (1953) 2029.
223. K. K. Verma, L. K. Doraiswamy, *Ind. Eng. Chem. Fundam.* **4** (1965) 389.
224. D. W. Van Krevelen, H. A. G. Chermin, *Chem. Eng. Sci.* **1** (1951) 66; *Chem. Eng. Sci.* **1** (1952) 238.
225. L. Constantinou, R. Gani, *AIChE J.* **40** (1994) 1697.
226. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.

227. R. C. Reid, T. K. Sherwood: *The Properties of Gases and Liquids,* 2nd ed., McGraw-Hill, New York 1966.
228. W. D. Monnery, W. Y. Svrcek, A. K. Mehrota, *Can. J. Chem. Eng.* **73** (1975) 3.

General References

References for Chapter 6

229. S. Chapman, T. G. Cowling: *The Mathematical Theory of Nonuniform Gases,* Cambridge, NY, 1939.
230. S. Weiß et al. (eds.): *Verfahrenstechnische Berechnungsmethoden,* part 7: Stoffwerte, VCH Verlagsgesellschaft, Weinheim 1986.
231. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* 3d ed., McGraw-Hill, New York 1977.
232. *VDI Wärmeatlas,* 7th ed., VDI Verlag, Düsseldorf 1994.
233. J. O. Hirschfelder, C. F. Curtiss, R. B. Bird: *Molecular Theory of Gases and Liquids,* 4th print, Wiley, New York 1967.
234. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
235. O. A. Uyehara, K. M. Watson, *Natl. Pet. News* **36** (1944) R 714.
236. K. Lucas: *Phase Equilibria and Fluid Properties in the Chemical Industry,* Dechema, Frankfurt 1980, p. 573.
237. K. Lucas, *Chem. Ing. Tech.* **53** (1981) 959.
238. *VDI-Wärmeatlas,* 7th ed., VDI-Verlag, Düsseldorf 1994.
239. C. R. Wilke, *J. Chem. Phys.* **18** (1950) 517.
240. R. S. Brokaw, NASA Tech. Note D-4496, Apr. 1968.
241. R. S. Brokaw, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 240.
242. D. Reichenberg, *Symp. Transp. Prop. Fluids and Fluid Mixtures,* Natl. Eng. Lab., Glasgow, UK, 1979; see also [234], p. 404.
243. T.-H. Chung, L. L. Lee, K. E. Starling, *Ind. Eng. Chem. Fundam.* **23** (1984) 8.
244. T. H. Chung, L. L. Lee, K. E. Starling, *Ind. Eng. Chem. Res.* **27** (1988) 671.
245. L. A. Bromley, C. R. Wilke, *Ind. Eng. Chem.* **43** (1951) 1641.
246. D. Reichenberg, *AIChE J.* **16** (1970) 854; *AIChE J.* **21** (1975) 181.
247. E. N. da C. Andrade, *Endeavour* **13** (1954) 117.
248. D. Van Velzen, R. L. Cardozo, H. Langenkamp, *Ind. Eng. Chem. Fundam.* **11** (1972) 20.
249. J. A. Riddick, W. B. Bunger: *Organic Solvents, Physical Properties and Methods of Purification,* 3rd ed., vol. II, 1970, 4th ed., 1986, J. Wiley, New York.
250. L. H. Thomas, *J. Chem. Soc.* 1946, 573.
251. P. S. Morris, MS thesis, Polytechnic Institute of Brooklyn, Brooklyn, N.Y., 1964.
252. C. Orrick, J. H. Erbar in [234], p. 456.
253. J. W. Przezdziecki, T. Sridhar, *AIChE J.* **31** (1985) 333.
254. A. Letsou, L. I. Stiel, *AIChE J.* **19** (1973) 409.
255. K. G. Joback, R. C. Reid, *Chem. Eng. Commun.* **57** (1987) 233.
256. V. M. Tatevski, A. V. Abramenkov, O. E. Grikina, *Russ. J. Phys. Chem.* **61** (1987) 1454.
257. R. A. McAllister, *AIChE J.* **6** (1960) 427.
258. J. B. Irving: "Viscosities of Binary Liquid Mixtures: A Survey of Mixture Equations," *NEL Rep. GB* **630** (1977); "Viscosities of Binary Liquid Mixtures: The Effectiveness of Mixture Equations," *NEL Rep. GB* **631** (1977).
259. Y. S. Touloukian, C. Y. Ho: "Viscosity. Thermophysical Properties of Matter," *The TPRC Data Ser.,* **vol. 2,** Plenum Press, New York 1975.
260. K. Stephan, K. Lucas: *Viscosity of Pure Dense Fluids,* Plenum Press, New York 1979.
261. W. Blanke (ed.): *Thermophysikalische Stoffgrößen,* Springer Verlag, Berlin 1989.
262. D. Roy, G. Thodos, *Ind. Eng. Chem. Fundam.* **7** (1968) 529.
263. N. B. Vargaftig: *Tables on the Thermophysical Properties of Gases and Liquids,* 2nd ed., Hemisphere, Washington 1975.
264. J. F. Ely, J. M. Hanley, *Ind. Eng. Chem. Fundam.* **22** (1983) 90.
265. L. A. Guildner, *Proc. Natl. Acad. Sci. USA* **44** (1958) 1149.
266. T. H. Chung, L. L. Lee, K. E. Starling, *Ind. Eng. Chem. Fundam.* **19** (1980) 186.
267. S. Angus et al. (eds.): *IUPAC: International Thermodynamic Tables of the Fluid State: Carbon Dioxide,* Pergamon Press, Oxford 1973.
268. D. Misic, G. Thodos, *AIChE J.* **7** (1961) 264; *J. Chem. Eng. Data* **9** (1963) 540.
269. D. Roy, G. Thodos, *Ind. Eng. Chem. Fundam.* **9** (1970) 71.
270. L. I. Stiel, G. Thodos, *AIChE J.* **7** (1961) 611; *AIChE J.* **10** (1964) 26.
271. Y. S. Touloukian, C. Y. Ho: "Thermal Conductivity. Thermophysical Properties of Matter," *The TPRC Data Ser.,* **vol. 3,** Plenum Press, New York 1972.

272. D. T. Jamieson, J. B. Irving, J. S. Tudhope: *Liquid Thermal Conductivity: A Data Survey to 1973,* H. M. Stationery Office, Edinburgh 1975.
273. T. Maejima in [234], p. 550.
274. L. Riedel, *Chem. Ing. Tech.* **21** (1949) 349; **23** (1951) 59, 321, 465.
275. G. Latini, M. Pacetti, *Therm. Conduct.* **15** (1978) 245.
276. C. Baroncini, G. Latini, P. Pierpaoli, *Int. J. Thermophys.* **5** (1984) no. 4, 387.
277. L. A. Robbins, C. L. Kingrea, *Hydrocarbon Process. Pet. Refiner* **41** (1962) 133.
278. A. Missenard, *Comptes Rendus* **260** (1965) 5521.
279. K. Sato in [234], p. 550
280. M. Nagvekar, T. E. Daubert, *Ind. Eng. Chem. Res.* **26** (1987) 1362.
281. A. Missenard, *Rev. Gen. Therm.* **5** (1970) no. 101, 649.
282. C. C. Li, *AIChE J.* **22** (1976) 927.
283. W. Jost: *Diffusion in Solids, Liquids, Gases,* Academic Press, New York 1960.
284. E. N. Fuller, J. C. Giddings, *J. Gas Chromatogr.* **3** (1965) 222. E. N. Fuller, K. Ensley, J. C. Giddings, *Ind. Eng. Chem.* **58** (1966) no. 5, 18.
285. L. S. Tee, G. R. Kuether, R. C. Robinson, W. E. Stewart, *Am. Pet. Inst. Div. Refin.* May 1966.
286. S. Takanishi, *J. Chem. Eng. Jpn.* **7** (1974) 417.
287. E. L. Cussler: *Diffusion: Mass Transfer in Fluid Systems,* Cambridge University Press, Cambridge 1984, Chaps. 3, 7.
288. D. Reinhardt, K. Dialer, *Chem. Eng. Sci.* **36** (1981) 1557.
289. W. Hayduk, S. C. Cheng, *Chem. Eng. Sci.* **26** (1971) 635.
290. W. Hayduk, B. S. Minhas, *Can. J. Chem. Eng.* **60** (1982) 295.
291. L. S. Darken, *Trans. Am. Inst. Min. Metall. Eng.* **175** (1948) 184.
292. R. E. Rathburn, A. L. Babb, *Ind. Eng. Chem. Process Des. Dev.* **5** (1966) 273.
293. M. A. Siddiqi, K. Lucas, *Can. J. Chem. Eng.* **64** (1986) 839.
294. H. R. Kamal, L. N. Canjar, *AIChE J.* **8** (1962) no. 3, 329.
295. A. Vignes, *Ind. Eng. Chem. Fundam.* **5** (1966) 189.
296. C. R. Wilke, P. Chang, *AIChE J.* **1** (1955) 264.
297. E. G. Scheibel, *Ind. Eng. Chem.* **46** (1954) 2007.
298. D. F. Othmer, M. S. Thakar, *Ind. Eng. Chem.* **45** (1953) 589.
299. K. A. Reddy, L. K. Draiswamy, *Ind. Eng. Chem. Fundam.* **6** (1967) 77.
300. M. A. Lusis, G. A. Ratcliff, *Can. J. Chem. Eng.* **46** (1968) 385.
301. M. T. Tyn, W. F. Calus, *J. Chem. Eng. Data* **20** (1975) 106.
302. K. Nakanishi, *Ind. Eng. Chem. Fundam.* **17** (1978) 253.

General References

References for Chapter 7

303. J. J. Jasper, *J. Phys. Chem. Ref. Data* **1** (1972) 841.
304. D. B. Macleod, *Trans. Faraday Soc.* **19** (1923) 38.
305. S. Sugden, *J. Chem. Soc.* **125** (1924) 1177.
306. Q. R. Quale, *Chem. Ref.* **53** (1953) 439.
307. R. C. Reid, J. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
308. J. A. Riddick, W. B. Bunger, T. K. Sakano: *Organic Solvents. Physical Properties and Methods of Purification,* 4th ed., vol. II, Wiley, New York 1986.
309. J. D. Van der Waals, *Z. Phys. Chem.* **13** (1894) 716.
310. D. I. Hakim, D. Steinberg, L. I. Stiel, *Ind. Eng. Chem. Fundam.* **10** (1971) 174.
311. J. R. Brock, R. B. Bird: *AIChE J.* **1** (1955) 174.
312. W. Blanke (ed.): *Thermophysikalische Stoffgrößen,* Springer Verlag, Berlin 1989.
313. H. P. Meissner, A. S. Michaelis, *Ind. Eng. Chem. Ind. Ed.* **41** (1949) 2782.
314. M. Tamura, M. Kurata, H. Odani, *Bull. Chem. Soc. Jpn.* **28** (1955) 83.

Estrogens → Hormones
Ethane → Hydrocarbons

Construction Materials in Chemical Industry

HUBERT GRÄFEN, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 2
2. Material Requirements 2
2.1. Processability and Joining 3
2.2. Mechanical Stability and its Dependence on Temperature 3
2.3. Corrosion Resistance 5
2.4. Resistance to Wear 6
3. Choice of Materials 6
4. Quality Assurance through Material Tests and Checking of Fabrication and Functioning 8
5. Properties and Applications of Materials 9
5.1. Steels 9
5.1.1. Unalloyed and Low-Alloy Steels for Vessels and Pipelines 9
5.1.2. Steels with High-Temperature Strength 11
5.1.3. Heat-Resistant Steels 12
5.1.4. Steels for Low Temperatures 13
5.1.5. Steels Resistant to Pressurized Hydrogen 13
5.1.6. Stainless Steels 14
5.1.6.1. Technical Properties 18
5.1.6.2. Chemical Properties 20
5.1.6.3. Development State of Stainless Cr–Ni Steels 24
5.2. Cast Iron 25
5.3. Nickel and Nickel Alloys 26
5.3.1. Nickel–Copper Alloys 26
5.3.2. Nickel–Chromium Alloys 27
5.3.3. Nickel–Molybdenum and Nickel–Molybdenum–Chromium Alloys . 27
5.4. Aluminum and Aluminum Alloys 28
5.5. Copper and Copper Alloys 31
5.6. Lead and Lead Alloys 33
5.7. Zinc and Zinc Alloys 34
5.8. Tin and Tin Alloys 35
5.9. Titanium, Zirconium, Niobium, and Tantalum 35
5.10. Organic Materials 38
5.10.1. Selection Criteria 39
5.10.2. Properties and Application Criteria 39
5.10.3. Thermosetting Plastics 41
5.11. Inorganic Nonmetallic Materials 42
5.11.1. Glass 42
5.11.2. Graphite 43
5.11.3. Refractory and Acid-Resistant Bricks 43
5.11.4. Engineering Ceramics 44
6. References 45

1. Introduction

Components of chemical plant are generally subjected to thermal, chemical, and mechanical stresses. The combination of these stresses places very heavy demands on plant materials, especially with regard to corrosion. Thus even unalloyed and low-alloy steels have to meet very strict quality requirements. They must have high purity, for example, and their nonmetallic inclusions must be finely dispersed, since they affect the ability of plant to withstand corrosion cracking, such as stress corrosion cracking and damage by hydrogen. By far the most important materials are the highly alloyed stainless steels and nickel-based alloys, though aluminum, copper and their alloys, and refractory metals, and organic and inorganic materials are also important. In this article the applications of materials in process plant manufacture are described and development trends are discussed.

2. Material Requirements

In the choice of materials for production plant in the chemical industry, there are three basic considerations:

1) The processibility of the material in its commercially available form (sheet, piping, profiles, etc.)
2) The ability of materials to withstand production processes. This is a complex property that includes mechanical stability and its dependence on temperature, resistance to corrosion, and possibly resistance to wear also.
3) The costs of materials, of their processing,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_07

and of the inspections of the chemical apparatus during its useful lifetime.

As so many factors have to be taken into consideration, choosing materials is not easy, especially if the plant is to be used for a recently developed chemical process. The prospect of a suitable choice is best where the material scientist, chemical engineer, plant designer, and plant engineer have worked closely together [1], [2].

2.1. Processability and Joining

Shaping (e.g., bending, rounding, and flanging), separating (e.g., cutting and machining) and joining (e.g., welding, bonding, and pipe rolling) are particularly important processes in the fabrication of chemical plant. Coating and processes that modify the properties of materials (e.g., quenching, tempering, nitriding, age hardening) are also widely used. The choice of fabrication processes depends on the properties of the material concerned, e.g., on its suitability for cold shaping or welding. Fabrication must not affect materials so drastically that their resistance to the conditions of subsequent use is significantly impaired.

Joint welding is by far the most important joining process in chemical plant fabrication, while build-up welding is used widely both for coating in original plant fabrication, and for repairs [3–5]. The weldability of a component depends on the weldability of the material and on the design and fabrication of the part [6]. Although each factor may be decisive on its own, the interplay of factors must never be overlooked. It is thus pointless to choose a steel with the highest possible yield strength unless one has checked that dangerous peak stress will not occur (through deficiencies of design) and that welding defects will not be caused by the use of welding techniques unsuited to the material.

Bonding is used for materials that cannot be welded or whose properties are changed excessively by welding [7], [8]. As adhesive bonds cannot be exposed to elevated temperatures, this method of joining parts has acquired little importance for chemical apparatus.

2.2. Mechanical Stability and its Dependence on Temperature [9]

Chemical apparatus in use is subjected to widely differing mechanical stresses. The possible variations of mechanical stress with time are shown schematically in Figure 1. These stresses may be monoaxial, biaxial, or triaxial.

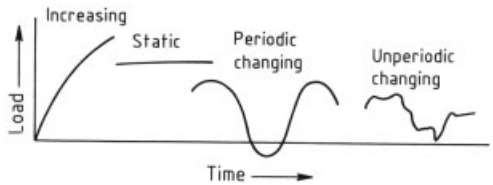

Figure 1. Forms of the time function of mechanical stress

The behavior of materials under polyaxial stress – the states of stress in plant components are almost always of this kind – is seldom known as such, because strength data are usually available only as yield strength and tensile strength for loads exerted monoaxially in a tensile test. Therefore, where a particular exposure is concerned, the material's mechanical stability must be assessed by comparing its known mechanical strength values with a stress calculated according to the appropriate theory of strength [10].

In the fabrication of chemical apparatus preference is given to materials that are easily shaped, since they react to the application of excessive force by undergoing energy-consuming changes of shape instead of simply breaking. Temperature has an important influence on the mechanical stability of materials. As the temperature rises, strength decreases, ease of shaping increases, and creep behavior becomes the main factor determining mechanical stability.

Thus, above a limit temperature, which depends on the material, the yield strength or tensile strength at the envisaged operating temperature can no longer serve as a characteristic value for calculating the stress of a chemical apparatus (Fig. 2, see next page) [9].

The strength of metals increases with decreasing temperature. Therefore, components intended for use at room temperature might be expected to be more resistant to loads exerted at lower temperatures. That this is only partly true is explained by the deformation behavior of materials. The characteristic strain values of metals tend to drop suddenly as the tempera-

ture decreases, and this is accompanied by an increase in the tendency to undergo brittle fracture. The inherent risk of brittle fracture is not accounted for directly by strength calculations.

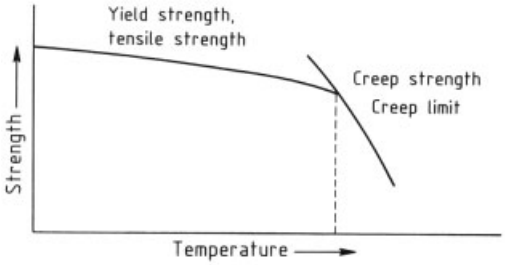

Figure 2. Dependence of characteristic strength data on temperature (schematic)

Susceptibility to brittle fracture is represented by values for elongation at rupture and reduction of area in the tensile test, and also by the transition temperature obtained from a plot of the notched-bar impact energy versus temperature (Fig. 3) (see also → Mechanical Properties and Testing of Metallic Materials). In choosing a material to suit a particular set of requirements, adequate safety must be ensured by taking this transition temperature into consideration, as well as the influences exerted on the material during its production and processing. In the case of high-strength materials, it is also advisable to perform fracture toughness testing [11].

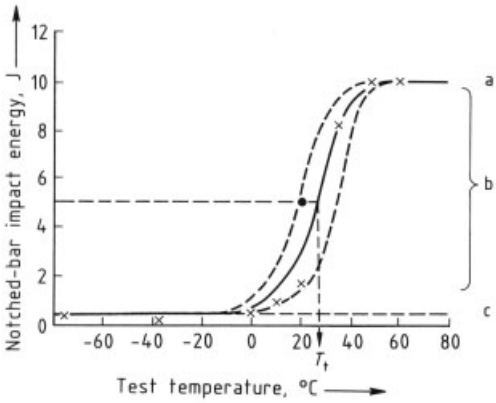

Figure 3. Transition temperature T_t of the notched-bar impact energy
a) High level, fracture on working; b) Transition range, mixed fracture; c) Low level, brittle fracture

Where alternating stresses occur, determining suitable characteristic material values that will enable a part to be designed so that it satisfies safety requirements is often difficult, because the combinations of loads exerted in practical use cannot be accounted for fully by measuring the alternating stress endurance limit (e.g., by the Wöhler method). The shape and surface condition of the part must also be considered (Fig. 4). More realistic criteria can be obtained by determining the resistance to service conditions of components themselves. The AD (Arbeitsgemeinschaft Druckbehälter) leaflets [12] on calculations for pressure vessels or their components are based on the assumption that these vessels are normally subjected to static loads. If the pressure fluctuates, the resultant additional stresses can be taken into account according to AD leaflet S 1, in which allowance for them is made by reducing the permissible stresses (Fig. 5). Prediction of useful lifetime is particularly difficult when alternating stresses and creep stresses are superimposed.

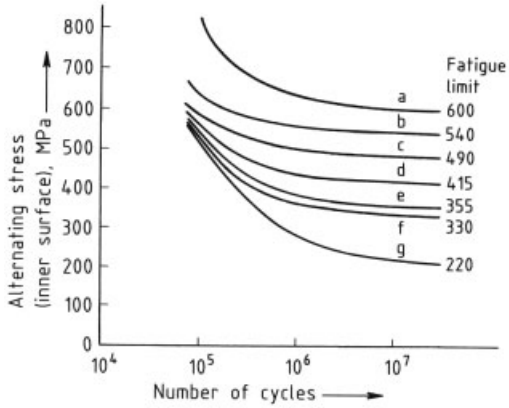

Figure 4. Influence of the surface quality on alternating stress tests on high-presure tubes (nominal bore 6 mm; nominal pressure 250 MPa) made of alloy 30CrMoV9 (material no. 1.7707; HB30 ≈ 3000 MPa; tensile strength = 900 – 1100 MPa)
a) Cold-worked; b) Nitrided; c) Electropolished; d) Nitrided; e) Polished; f) and g) As-delivered

In such cases there is still no accepted method for prediction of useful life. It would also appear unrealistic to seek "universally valid" theories and better to simulate typical stress processes in test specimens similar to plant components, thus providing the design engineer with useful life data for particular kinds of materials and problems. According to the available experience it is best to design plant and components so that they

can withstand the stress that gives the shortest useful life.

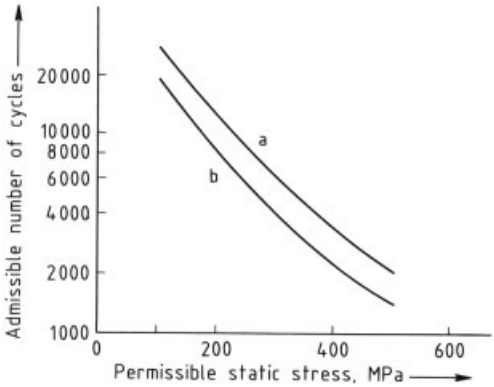

Figure 5. Admissible number of cycles for pressure vessels as a function of stress
a) Range of alternating compressive stress = 90 %;
b) Range of alternating compressive stress = 100 %

2.3. Corrosion Resistance (→ Corrosion)

The longevity of most apparatus depends not only on mechanical loads but also on the nature and spectrum of the corrosive ambient medium. As a rule the characteristic symptom of corrosion damage is not the loss of material but an impairment of function and load-bearing capacity. Uniform corrosion, for example, may cause a considerable loss of mass before the serviceability of a part begins to suffer. Localized or selective corrosion proceeds rapidly towards the inside of a part, such that a notching effect is exerted on parts that bear mechanical loads. Corrosion of this kind may rapidly cause vessels to leak, or parts to fail through low-ductility fracture.

If the surface becomes creviced or pitted, the stress becomes nonuniform, and the operational stability of the component under alternating stresses is impaired. This also reduces the static fatigue resistance under constant load, as shown by the example in Figure 6 [13]. The static fatigue resistance of copper samples exposed simultaneously to corrosion and mechanical loading is impaired to a much greater extent that of precorroded samples that have already undergone a considerably greater loss of mass and are then stressed mechanically. This exemplifies the interaction between corrosion and creep. For iron-based and nickel-based alloys, this interrelation affects the static load fatigue resistance of parts exposed to corrosion processes at high temperatures.

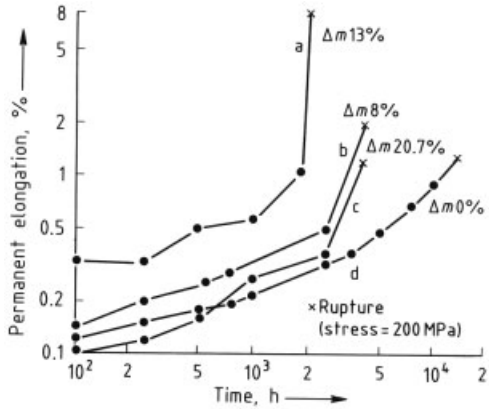

Figure 6. Creep rupture behavior of copper at 95 °C
a) 0.5 M H_2SO_4 (air saturated); b) 0.5 M H_2SO_4 (air); c) Air (precorroded in 0.5 M H_2SO_4); d) Air
Δm = Mass loss, wt %

Basically it may be assumed that the behavior of a part exposed to corrosive influences depends largely on the stability of the material, which, in turn, depends mainly on the film which forms at the surface. Efforts are therefore being made to find alloys that can improve the conditions for surface film formation. It is also necessary to ensure that the microstructures of alloys remain as stable as possible, so that demixing (which, for example, could impair the passivation capability of alloys) cannot occur in welding, hot forming, etc. This is particularly important because precipitation may strengthen the tendency towards local activation (local removal of the passive layer), whereby local corrosion may be initiated.

Highly localized corrosion that results from simultaneous chemical attack and mechanical stress is particularly serious. The criterion of failure here is the occurrence of low-ductility fracture. Stress corrosion cracking, fatigue cracking, and water-induced cracking are phenomena of this kind. They are particularly important with regard to serviceability, useful life, and operational reliability.

The special danger of anodic stress corrosion cracking and hydrogen-induced cracking is that they can rarely be detected while the cracks are spreading and before the apparatus starts to leak

or a part breaks. Although anodic stress corrosion cracks outwardly resemble brittle fractures, the metal itself retains its ductility.

The initiation of cracking by the various stress corrosion cracking mechanisms that result from the action of a given corrodent on a given material depends very much on the mechanical stresses involved.

In the case of fatigue cracking the loss of fatigue strength for finite life and loss of fatigue strength have caused considerable difficulties in the dimensioning of parts.

2.4. Resistance to Wear (see also → Abrasion and Erosion)

Since all production processes, except chemical reactions, involve physical operations such as comminution, conveyence, and separation of phases, the wear resistance of materials is important for many types of chemical apparatus. Models of the various kinds of wear have been developed [14]; in practice, however, the conditions are complex, for not only may severals kinds of wear occur simultaneously, but overlap and interaction of wear and corrosion processes can also arise.

Thus, as a rule, only practical trials reveal the actual stress conditions involved. Compared to corrosion, however, wear is less problematic because it does not cause cracklike damage, and its effects are more easily repaired, e.g., by build-up welding.

3. Choice of Materials

Designers of chemical apparatus must pay careful attention to materials as well as the purpose for which the apparatus is intended. This is important at all stages, from initial planning to the detailed drawing of the finalized design. For a selection process, however, not just the function and properties needed by the components, but also the operations by which the components are fabricated from the material, must be taken into consideration. Not only do these operations call for definite properties on the part of materials, they may also affect the behavior of components under service conditions.

The demands made on chemical plant in use are becoming increasingly strict, and the variety of fabrication processes and range of materials available are growing. These factors, together with the need to use economical fabrication techniques, have given the choice of materials a complexity that calls for systematic consideration. A basic procedure for the selection of materials for chemical plant is shown in Figure 7 (see next page). Clearly the demands of the application and those of the fabrication process must be taken into account as fully as possible so that the overall requirements are correctly formulated and properly interpreted in the fixing of verifiable property data. At the same time the behavior of materials under operating conditions and those of fabrication must be known so that one can judge reliably whether or not a given material is suitable for a given purpose.

Often, these preconditions cannot be met entirely. In many cases, therefore, model experiments and serial testing are necessary as empirical aids to selection. Establishing and then interpreting the combination of mechanical and corrosion-chemical exposures, all of which are often supplemented by wear, is particularly difficult for chemical plant. Frequently, corrosion behavior is treated in the selection procedure as the decisive criterion. In fact, experience in the operation of chemical plant has shown that much of the damage that occurs to chemical apparatus arises from the use of materials that are insufficiently resistant to corrosion under practical conditions, and that damage of purely mechanical origin is generally much less frequent [15].

Therefore, technical rules exist for designing and dimensioning apparatus to meet combined mechanical and thermal stresses. The standardized calculation procedures based on strength and ductility data are intended to provide the designer with a proven basis on which to work. However, where wear processes (corrosion and abrasion) are concerned, that is not possible; here, instead, special design criteria must be worked out in each individual case.

A shortlist of materials can be drawn up from corrosion tables, which give the corrosion behavior of specific materials [16], [17]. Nevertheless, the most extensive tables can never take into consideration all the conditions of a given practical case. Almost always there is more than one corrosive agent, because chemi-

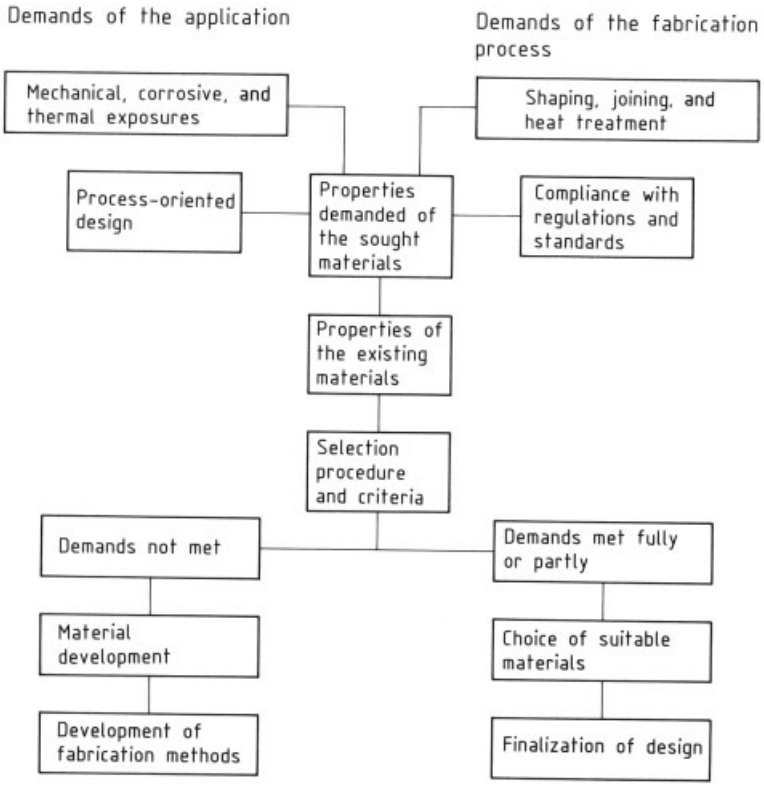

Figure 7. Basic procedure for the choice of materials for chemical apparatus

cal processes normally involve starting materials and end products, intermediates and byproducts, often with unknown properties, and solids that promote wear. Furthermore, reactions are influenced by pressure and temperature. Thus the corrosion behavior of materials cannot be given in tables with any degree of accuracy, and terms such as resistant or nonresistant may be inappropriate in an actual application. The results of laboratory tests are consideraly more useful.

The most reliable predictions are those based on plant tests performed either in a plant engaged in practical production or in a pilot plant. Often, however, such tests are very laborious. Even in cases where tests have been carried out, corrosion may still occur, possibly because the stability of a material has been impaired by excessive cold forming or by excessive heat input in welding, or because the states of the material in the test plant and in the subsequently constructed chemical apparatus were not identical.

In view of this complexity, close cooperation between material scientists, chemical engineers, chemists, designers, apparatus manufacturers, and – where chemical apparatus subject to compulsory testing, e.g., pressure vessels, is concerned – the representatives of the officially recognized supervisory authorities, is particularly desirable.

Because corrosion reactions are so complex and depend on so many factors, an assessment of the corrosion behavior of a given material – corrodent system will differ greatly according to the method of investigation. False interpretations can be avoided and comparison improved if one has a good knowledge of the kinds of information provided by the chosen test method as well as knowing its limitations. It is therefore understandable that standardized corrosion tests are few in number or contain more general instructions. Frequently, they cover only the states of materials that are most favorable from the aspect of corrosion chemistry (that is to say the ho-

mogeneous structural states existing at the time of delivery), describing, for example, the testing of stainless steels for resistance to intergranular corrosion (DIN 50 914) or the testing of stainless steels in boiling nitric acid (DIN 50 921 and ASTM A 262-70). A comparison of the sensitivities of unalloyed and low-alloyed steels to intergranular stress corrosion cracking is given in DIN 50 915.

A common feature of all investigation methods is that they refer to the material, and not directly to the stresses that occur in practice. Supplementary chemical or electrochemical tests simulating the conditions that occur in practice are therefore necessary in the selection of materials. Chemical tests are performed on samples exposed to gaseous and liquid corrodents at the operating pressure and temperature, whereas in electrochemical corrosion tests, the dependence of corrosion on potential is investigated and provides information on the effects of variables which alter the potential.

The results of chemical corrosion tests and electrochemical corrosion tests may differ fundamentally. The differences arise because in chemical tests the corrosion potential may vary with time. In electrochemical tests the potential is fixed, so that, though they are more informative in some respects, their practical value may be limited in others.

The characteristic stability data needed in the choice of materials can be obtained in two ways:

1) By inserting suitable material specimens in pilot plants or in existing, but not yet entirely satisfactory, production plants,
2) By performing laboratory corrosion tests.

With regard to the corrosive medium in corrosion tests, careful attention must be paid to a number of important factors, such concentration, temperature, dissolved gases, impurities, solid matter, and rate of flow. As there are so many parameters, tests should use suitable specimens in a plant and under conditions as close as possible to those of practical use.

The specimens should be placed at several representative locations. In a distillation column, for example, they should be placed in the pit, in the vicinity of the feed point, and in the head.

The shape of a specimen used depends on the types of corrosion expected. Where general corrosion predominates, it is sufficient to use welded sheets that have the surface quality of the parts to be used in practice. If stress corrosion cracking is involved, plastically and elastically stressed specimens or tuning fork specimens are used.

Material specimens are generally attached to interior parts, such as an agitator or a thermometer protection tube. To prevent polarization of specimens through contact with plant components, and consequent falsification of measurements, the fixing screws are placed in insulating sheaths.

In addition to the exposure of test specimens, the parts of a pilot plant itself should be examined as a source of further information on corrosion behavior. For critical plant units such as heat exchangers, it has been found advisable to remove a tube from time to time, to cut it open, and to examine the inside and outside surfaces for corrosion symptoms.

If stability tests cannot be performed in pilot installations, laboratory tests under conditions as close as possible to those of practice must be carried out. They should also be carried out as a supplementary measure in cases where plant tests have not clearly revealed the time dependency of the corrosion processes, and severer conditions of attack may therefore help to complete the picture.

General guidelines on the conduct of corrosion tests are given in DIN 50 905. These guidelines should be followed as a route to reciprocally comparable results that can be transferred to the plant component in question.

For further details on chemical and electrochemical corrosion tests, see → Corrosion.

4. Quality Assurance through Material Tests and Checking of Fabrication and Functioning

To ensure that a chemical apparatus will be properly fabricated and that its functioning free from disturbances, many tests must be carried out before and during fabrication, before the apparatus is put into operation, and while it is in actual use. At the stage of testing of materials when they leave the factory and arrive at the plant manufacturer's establishment it is normally fairly easy to ascertain whether or not they are of the required

quality, as the required property data will have been agreed on the basis of quality standards or similar specifications. The same cannot be said of fabrication checks, which include, in particular, checking the construction for compliance with the blueprints and examining the welding work. Here the test engineer has more latitude, especially in interpreting the results of nondestructive tests. It must be emphasized that the first step in fabrication testing should be a thorough visual inspection; this gives the tester an overall impression of the care taken by the manufacturer. Checking compliance with legislation and official regulations is mainly the responsibility of the Technical Control Associations and the plant safety inspection departments of the large chemical firms. Functioning tests check the behavior of the apparatus under conditions close to those of practice. Of exceptional importance are the regular inspections of plant in use that are intended to reveal incipient damage early enough to ensure that plant shut-down can be avoided and any necessary repairs need not be undertaken in haste. Once again, visual inspection takes precedence. Much emphasis is also placed on nondestructive methods – thickness measurement, for example, and tests that reveal initial cracking [18]. Testing of plant in service also provides knowledge that may be useful in selecting materials for new chemical apparatus by revealing weaknesses of design, materials, or fabrication.

5. Properties and Applications of Materials

5.1. Steels (→ Steel)

Steels are still the materials most commonly used for chemical plant. Their variety of alloy compositions and the range of variation of their properties permit an exceptional degree of adaptation to practical requirements.

5.1.1. Unalloyed and Low-Alloy Steels for Vessels and Pipelines

Sheet and piping made of unalloyed and low-alloyed steels with carbon contents up to ca. 0.25 wt % are used extensively in chemical plant for vessels (pressure vessels, storage vessels, etc.) and pipelines that are not exposed to particularly severe corrosion. The main standards and designations of several typical steels are compiled in Table 1. Comparable steels with comparable properties are described in the standard specifications of other countries – see, for example, SAE, AISI, ASTM (United States), BS (Great Britain), NF (France), SS (Sweden), UNI (Italy) and NBN (Belgium). The steels are normally processed in the normalized state.

Owing to the introduction of weldable fine-grained structuralsteels with yield strengths of

Table 1. Quality standards for unalloyed and low-alloy steels widely used in chemical plant

Standard	Short title	Typical steels	
		Short name	Material number
DIN 17 100	Steels for general structural purposes	RSt 37 – 2	1.0038
		St 44 – 2	1.0044
DIN 1629	Seamless unalloyed circular steel tubes for special requirements	St 37.0	1.0254
		St 44.0	1.0256
		St 52.0	1.0421
DIN 17 155	Steel plates and strips for pressure purposes	H I	1.0345
		H II	1.0425
		15 Mo 3	1.5415
DIN 17 102	Weldable, finegrained construction steels (normalized)	StE 285	1.0486
		StE 355	1.0562
		WStE 460	1.8935
DIN 17 175	Seamless tubes of heat-resistant steels	St 35.8	1.0305
		St 45.8	1.0405
		15 Mo 3	1.5415
Stahl-Eisen-Werkstoffblatt 087 – 81	Structural steels resistant to weathering	WTSt 37 – 2	1.8960
		WTSt 37 – 3	1.8961

≥ 360 MPa, boiler plate consisting of unalloyed steels, whose yield strengths are considerably lower, is used on a smaller scale than it was formerly. The use of high-strength heat-treated structural steels with yield strengths of ≥ 700 MPa [19–23] will presumably increase. Pipes consisting of welded steel strip will be used increasingly instead of seamless pipes [24].

Steels are suitable for fusion welding by all processes, but the welding filler metals must be suited to the base material. Specific welding guidelines must be adhered to increasingly as the alloy content and yield strength rise; these guidelines can be found by consulting either the standards or manufacturers of steels and welding filler materials. Processing guidelines for weldable fine-grained structural steels are compiled in DIN standards.

Normalized fine-grained structural steels are now very popular. They combine the excellent strength and toughness imparted by fine-grain hardening [25] with the advantages offered by modern ladle metallurgy in respect of purity and microalloying.

The blast or shot injection of calcium compounds to reduce the sulfide content of steel (TN process) and the addition of cerium and zirconium to produce globular, finely distributed sulfide inclusions should also be mentioned. These processes improve purity, block segregation, weldability, and toughness.

In special cases, these desulfurization methods enable the sulfur content to be reduced to below 0.005 %. Consequently, and through the influence exerted on the nature of the precipitation, the anisotropy of the steel's properties is reduced considerably, as may be seen from the percentage reductions of area at fracture of fine-grained structural steels in the thickness direction of the sheet (Fig. 8) [26]. Figure 9 [26] shows that the fatigue resistance is also improved.

Where damage occurs, the leak-before-failure behavior imparted by high toughness is exceptionally favorable because the risk of sudden failure as a result of unstable crack propagation is practically eliminated. The high purity of fine-grained steels, together with the fine dispersion of their nonmetallic inclusions, notably the sulfides, is very favorable with regard to corrosion resistance. Coarse sulfides promote the formation and growth of cracks both in hydrogen-

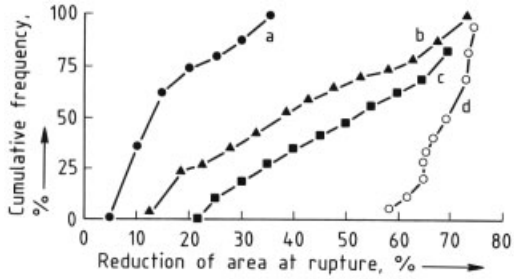

Figure 8. Influence of sulfur content on the reduction of area at rupture in the thickness direction of StE 355 steel plates
a) 40 – 55 mm plate, 0.015 – 0.054 % S; b) 40 mm plate, ≤ 0.010 % S; c) ≥ 50 mm plate, ≤ 0.006 % S; d) 20 – 50 mm plate, TN treated, 0.002 % S

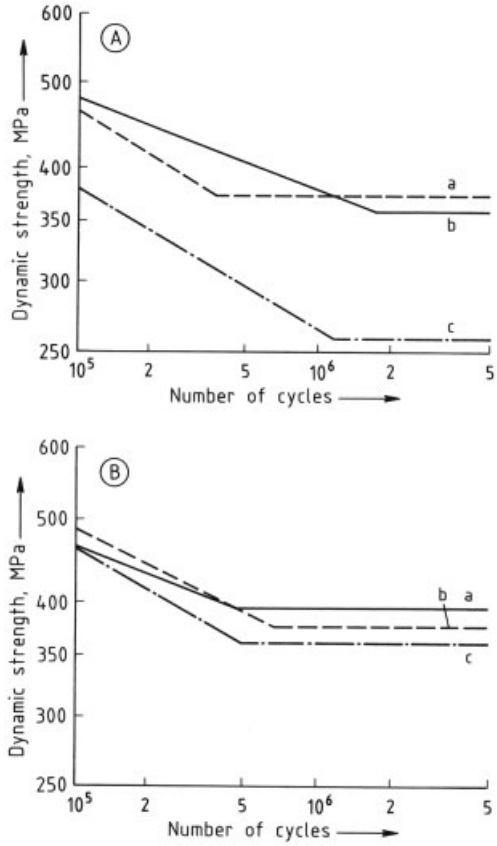

Figure 9. Influence of the TN process on the dynamic strength of steel StE 355 (stress ratio $R = 0$, thickness of plate: 50 mm)
A) Untreated; B) Calcium treated (TN process); a) Transverse; b) Longitudinal; c) Normal

induced crack formation and in stress corrosion cracking [27]. Upon chemical exposure to aqueous solutions, the sulfides, especially under the influence of the hydrolytic acidification that occurs in cracks, may be converted to H_2S, which is a vigorous promoter of hydrogen cracking.

It is now believed that crack growth in anodic stress corrosion cracking is also promoted by hydrogen. That would explain why steel StE 355, in particular, behaves favorably towards such crack-initiating media as alkalis, nitrates, and liquid ammonia. This steel is now a standard material for these media. For greater safety, the finished apparatus should be subjected to stress relief annealing.

Considerable progress has been made through efforts to improve the economy of structural steels by raising their strength. Starting from steel St 52-3, it has been possible to raise the yield strength from 355 to ca. 900 MPa. The first step in this direction consisted in raising the alloy contents of normalized structural steels. But the scope for improvement in this way was limited by the fact that increasing the alloy content impairs the cold cracking resistance in welding. Agreement has therefore been reached that the yield strength of normalized steels should not exceed 500 MPa.

Considerably higher yield strengths, even at low alloy contents, are obtainable by quenching and tempering. In the case of heat-resistant structural steels, air tempering is preferred. Water quenching followed by tempering, on the other hand, is more favorable for steel needing high toughness. By far the most important representative of this group of steels in terms of quantity is StE 690, whose minimum yield strength is 690 MPa (Euronorm 137). Steels with comparable yield strengths are also used on a large scale for pressure vessels and other chemical apparatus, such as the pressure-bearing parts of multistory vessels. StE 890 has been used to an increasing extent since the mid 1980s [28].

5.1.2. Steels with High-Temperature Strength

High-temperature structural steels are used in the chemical industry mainly for heavily stressed parts of steam boilers (e.g., collectors, drums, piping) and for pressure vessels and tubes operating at up to ca 800 °C [29]. These materials are also very important in reactor technology. According to their maximum service temperatures that can be maintained for long periods, the following groups are distinguished:

1) For temperatures up to ca. 400 °C: unalloyed steels, in some cases melted as fine-grained structural steels, whose main strength-related property is the high-temperature yield strength. These steels are standardized in DIN 17 155 (boiler plate types), DIN 17 175 (pipes) and DIN 17 102 (fine-grained structural steels) and are suitable for all fusion welding methods. Typical steels belonging to this group are the grades HI (1.0345), St 35.8 (1.0305), and WStE 355 (1.0565).
2) For temperatures of 400 – 550 °C: unalloyed and low-alloy steels having good long-term, high-temperature strength-related property values. The main alloying elements are manganese (up to 1.3 wt %), chromium (up to 2.5 wt %), and molybdenum (up to 1.2 wt %), vanadium also being used occasionally (up to 0.5 wt %). These materials are standardized in DIN 17 155, DIN 17 175, DIN 17 240 (nuts and bolts), and DIN 17 245 (steel castings). These steels are suitable for fusion welding, though preheating and post-welding annealing may be necessary, details of which can be found in the standards.
3) For temperatures of 550 – 600 °C: martensitic steels containing up to ca. 12 wt % chromium and additions of molybdenum, vanadium, nickel, and tungsten [30]. These are standardized in DIN 17 459 (sheet, pipes, forgings) and in DIN 17 245 (steel castings). These steels are supplied in the quenched and tempered state; a typical representative is X 20 CrMoV 12 1 (1.4922). Being air-hardenable, these steels must be kept at 250 to 450 °C during welding, after which they must be cooled to ca. 120 °C and then immediately annealed.
4) Temperatures of 600 – 800 °C: austenitic steels containing 16 – 21 wt % chromium, 11 – 32 wt % nickel, and additions of molybdenum, tungsten, niobium, tantalum, aluminum, and other elements, e.g., X 8 CrNiNb 16 13 (material no. 1.4961), X 8 CrNiMoNb 16 16 (material no. 1.4981),

Table 2. Approximate temperature limits for the use of steels in a weakly oxidizing flue-gas atmosphere

Steel	Material number	Standard	Approximate temperature, °C
St 35.8	1.0305	DIN 17 175	500
St 45.8	1.0405	DIN 17 175	500
15 Mo 3	1.5415	DIN 17 155	530
13 CrMo 44	1.7335	DIN 17 155	560
10 CrMo 9 10	1.7380	DIN 17 175	590
X 20 CrMoV 12 1	1.4922	DIN 17 175	600
X 8 CrNiNb 16 13	1.4961	DIN 17 459	750
X 8 CrNiMoNb 16 16	1.4981	DIN 17 459	750

and X 8 NiCrAlTi 32 21 (material no. 1.4959), which are standardized in DIN 17 459.

These steels are welded with filler metals of the same composition, and at the lowest possible heat input to avoid hot cracking of the weld metal.

As austenitic steels are very expensive, they are used only in the zones which reach the highest temperatures, outside of which low-alloy steels are used. Special techniques have to be used to form welds between ferritic and austenitic steels because of the diffusion processes that occur in welding and because of the differences in thermal expansion [31].

For temperatures above 800 °C only iron – chromium – nickel alloys and nonferrous alloys based on cobalt or nickel are suitable [32], [33]; see also Section 5.3.

Table 2 gives approximate figures for the highest temperatures at which a number of steels with high strength at elevated temperature can be used in a weakly oxidizing flue-gas atmosphere. Long-term heat stability values are compiled in [34]; the importance of these values in calculating the useful lives of components subjected to this exposure is discussed in [35].

5.1.3. Heat-Resistant Steels

Heat-resistant steels are those that have good strength-related property data and are distinguished by exceptional resistance to exposure for short or long periods to hot gases or combustion products at temperatures exceeding 550 °C, and which are thus resistant to scaling [36]. Rotary furnaces, cracking units, and muffle furnaces are examples of chemical plant in which these conditions arise.

Resistance to scaling is obtained mainly by alloying with chromium, but further improvements are possible if silicon and aluminum are added. The most commonly used heat-resistant steels are standardized in Stahl-Eisen-Werkstoffblatt 470-76. Distinctions are made between the ferritic steels, such as X 10 CrAl 7 (1.4713), X 10 CrAl 13 (1.4724) and X 10 CrAl 24 (1.4762); the ferritic-austenitic steel X 20 CrNiSi25 4 (1.4828); and austenitic steels such as X 15 CrNiSi 20 12 (1.4828), X 15 CrNiSi 25 20 (1.4841) and X 10 CrNiAlTi 32 20 (1.4876).

The above-mentioned publication also gives long-term heat resistance values for periods of up to 10^5 h. The highest service temperatures listed, which extend to 1200 °C, apply to air. They may be reduced greatly by admixtures to the air, e.g., of water vapor or sulfur-containing or carburizing matter, because the reaction products do not form sufficiently thick surface layers.

Within certain temperature ranges, heatresistant steels tend to become brittle. Ferritic steels with ≥ 12 wt % chromium do so between 400 and 530 °C. This so-called 475 °C embrittlement can be eliminated by annealing briefly above 600 °C. In the case of ferritic steels with ≥ 17 wt % chromium and austenitic steels, an intermetallic iron – chromium σ-phase, which causes severe embrittlement, is formed between 600 and 900 °C. It endangers particularly the weld interfaces.

Within the same temperature range, austenitic steels precipitate chromium carbides, which further reduce the toughness. The steel most severely affected is X 15 CrNiSi 25 20 (1.4841), which should be used only above 900 °C. The σ-phase and the carbides can be redissolved by annealing above 1000 °C, followed by quenching.

The heat-resistant steels can be welded by the usual methods, provided the guidelines for alloyed steels are followed. Suitable filler metals are listed in Stahl-Eisen-Werkstoffblatt 470-76. It should be noted that in ferritic steels with ≥ 12 wt % chromium, coarse grains, which can no longer be removed by heat treatment, are formed at temperatures above 900 °C.

Heat-resistant cast steel, as used in the manufacture of pipes by centrifugal casting, for example, is standardized in Stahl-Eisen-Werkstoffblatt 471-76.

5.1.4. Steels for Low Temperatures

Refrigeration is very important in the chemical industry, e.g., in the fractional distillation of hydrocarbons or storage and transportation of liquid gases. Steels for the pressure vessels must still have sufficient toughness at the lowest operating temperature. The degree of low-temperature toughness depends particularly on the steel's composition and heat treatment. Notched-bar impact energies, measured on DVM (Deutscher Verband für Materialforschung und -prüfung e. V., Berlin) longitudinal specimens, of ≥ 40 and ≥ 60 J/cm^2, for cast steel and other steels, respectively, are regarded as evidence of sufficient low-temperature toughness. Hence the lowest service temperature of a steel with good low-temperature tenacity is that at which the notch impact energy is still above this limit. Steels for low temperatures can be divided into four groups:

1) Unalloyed aluminum-killed steels for service temperatures down to -50 °C in the normalized state and -80 °C in the heat-treated state; e.g., TTSt 35 (1.1101).
2) Unalloyed and low-alloy weldable fine-grained structural steels in the normalized state for service temperatures down to -60 °C, e.g., TStE 380 (1.8910).
3) Nickel-alloyed heat-treatable steels with 1.5 – 9 wt % Ni for service temperatures of -100 to -190 °C, e.g., 12 Ni 19 (1.5680).
4) Austenitic chromium – nickel steels for service temperatures extending close to absolute zero; e.g., the steels specified in DIN 17 440 and DIN 17 441.

The relevant quality standards are DIN 17 173 and DIN 17 174 for groups 1 and 3 and DIN 17 102 for group 2, and Stahl-Eisen-Werkstoffblatt 685 – 82 for cast steel with good low-temperature toughness.

In the welding of steels for low temperatures it is necessary to use filler materials that give a weld metal whose strength and toughness are equal to those of the base material [37–39].

5.1.5. Steels Resistant to Pressurized Hydrogen [40], [41]

Steels for high-pressure plant in the chemical industry can divided into the following two groups:

1) Steels for components subjected to purely mechanical loads or to pressurized hydrogen, in both cases at temperatures of ≤ 200 °C,
2) Steels for components exposed to pressurized hydrogen at temperatures > 200 °C,

For the conditions of group 1, as present in the production of low-density polyethylene (pressures of up to about 400 MPa), mainly unalloyed and low-alloy heat-treatable steels are used; these are standardized in DIN 17 200 and in the case of large forgings, in Stahl-Eisen-Werkstoffblatt 550-76. If relatively good corrosion resistance is required, stainless heat-treatable chromium and chromium – nickel steels according to DIN 17 440, and occasionally hardenable stainless steels, are used instead [42]. The periodic alternation of the pressure subjects the components to a severe fatigue-inducing stress. Their shape stability therefore depends decisively not just on the choice of materials, but also on their being designed to suit the conditions of exposure as well as on satisfactory fabrication and installation [43].

Hot pressurized hydrogen, which is needed in many high-pressure syntheses (e.g., ammonia synthesis and pressure hydrogenation) dissociates on the surface of many steels, diffuses into the metal, and reacts with carbon to form methane, thus causing decarburization, embrittlement, and cracking:

$$Fe_3C + 2\,H_2 \rightarrow 3\,Fe + CH_4$$

This reaction between hydrogen and cementite occurs mainly at the grain boundaries, the

Table 3. Composition of pressurized-hydrogen-resistant steels

Steel type	Material number	Composition, wt %				
		C	Cr	Mo	V	Others
25 CrMo 4	1.7218	0.22/0.29	0.90/1.20	0.15/0.25		
16 CrMo 9 3	1.7281	0.12/0.20	2.0/2.5	0.30/0.40		
26 CrMo 7	1.7259	0.22/0.30	1.5/1.8	0.20/0.25		
24 CrMo 10	1.7273	0.20/0.28	2.3/2.6	0.20/0.30		
10 CrMo 11	1.7276	0.08/0.12	2.7/3.0	0.20/0.30		
10 CrMoV 10	1.7766	0.15/0.20	2.7/3.0	0.20/0.30	0.10/0.20	
20 CrMoV 13 5	1.7779	0.17/0.23	3.0/3.3	0.50/0.60	0.45/0.55	
X 20 CrMoV 12 1	1.4922	0.17/0.23	11.0/12.5	0.80/1.20	0.25/0.35	
X 8 CrNiMoVNb 16 13	1.4988	0.04/0.10	15.5/17.5	1.10/1.50	0.60/0.85	Ni: 12.5/14.5 N: 0.07/0.13

grains thus losing their cohesion. Hydrogen and methane accumulate, giving rise to high pressure whose cleaving action leads to internal microcracks. Together with the stresses exerted on the component and the loss of mechanical strength, this results finally in pronounced brittle fracture.

The resistance of carbon steels to attack by pressurized hydrogen depends partly on the ambient conditions, but also on the microstructure, cold working, effects of welding, impurities, and heat treatment. Elements that form stable carbides, such as chromium, molybdenum, tungsten, vanadium, titanium, and niobium, can be added to the steel to prevent the reaction between hydrogen and cementite. The pressurized-hydrogen-resistant steels now in use contain up to 16 wt % Cr, and also in many cases 0.2 – 1.5 wt % Mo. Some types are additionally alloyed with up to 0.85 wt % V (Table 3). The low-alloy ferritic steels are suitable for fusion welding, provided that they are preheated. The chromium steels have to be kept at 250 – 450 °C throughout welding, and partially annealed immediately afterwards. The austenitic steel (1.4988) tends to suffer hot cracking on welding. Short-term tests give very little indication of the limits to the use of steels of this kind, because the time factor is much too important and the attack of pressurized hydrogen on steels has a significant incubation period. As an illustration, Figure 10 gives stability limits for various steels, as compiled by Nelson [44]. It can be seen that even at low hydrogen partial pressures (e.g., 2 MPa) an unalloyed steel should be used only below 300 °C. Austenitic steels with 18 % Cr and ca. 9 % Ni have good resistance to pressurized hydrogen; they can be exposed to hydrogen throughout the range of temperatures used in normal high-pressure processes.

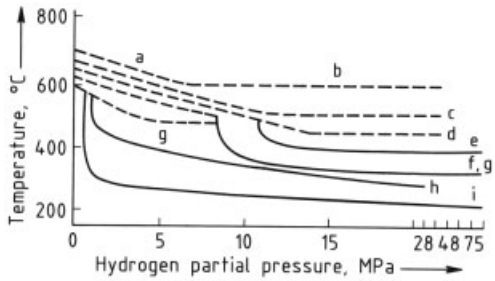

Figure 10. Resistance diagram for the attack of pressurized hydrogen on steels (Nelson diagram)
a) 1.25 Cr 0.5 Mo steel; b) 5.0 Cr 0.5 Mo steel; c) 8.0 Cr 0.5 Mo steel; d) 2.25 Cr 1.0 Mo steel; e) 2.0 Cr 0.5 Mo steel; f) 1.25 Cr 0.5 Mo steel; g) 1.0 Cr 0.5 Mo steel; h) 0.5 Mo steel; i) Mild steel

5.1.6. Stainless Steels [45]

Among the highly alloyed steels, the most important group comprises those that are chemically resistant. They have chromium contents ≥ 14 wt %. As the steels of this group resist heat as well as chemicals, this group also includes a considerable number of the heat-resistant steels discussed in Section 5.1.3. Often they contain nickel in addition to chromium. The more highly alloyed corrosion-resistant (acid-resistant) steels must additionally withstand general corrosion and localized corrosion in relatively aggressive corrodents (salt solutions; acids, even at fairly high concentrations). Characteristic alloying components, apart from chromium and nickel, are molybdenum, copper, and in some cases silicon.

There are many kinds of steels and alloys, each being represented by a material number (DIN 17 007) or by chemical composition (DIN 17 006).

The mechanical properties of steels are determined by the microstructure, which depends on composition and heat treatment. Stainless steels are divided according to microstructure into four groups:

1) Martensitic (hardenable) steels with > 0.12 wt % C and ≤ 15wt % Cr. They are hardened at temperatures $> 1000\,°C$ (e.g. cutting steels). After being hardened they can be improved by tempering at 500 to 600 °C – their strength thus being reduced to a desired lower level – and henceforth combine high strength with good ductility.
   In the heat treatment, a ferritic structure, with precipitated carbides of high chromium content of type $M_{23}C_6$ or M_7C_3, is formed. Fixation of chromium and the formation of chromium-depleted zones reduces the resistance to corrosion. To equalize the chromium content of the ferritic matrix again it is necessary to anneal the steel for a fairly long time at elevated temperature. Although this restores the corrosion resistance to some extent, it impairs the strength. As some of the chromium is still bound as carbide, the corrosion resistance after annealling procedure is still considerably poorer than that of the martensite.
2) Ferritic steels with body-centered cubic lattice (α-phase). Those of greatest importance are the ferritic chromium steels with ca. 17 wt % chromium. Their mechanical and technical properties, however, are unsatisfactory; when welded, they tend to become coarse-grained and brittle. Nevertheless they have the advantage of resisting stress corrosion cracking in chloride-containing media.
3) Austenitic steels with face-centered cubic lattice (γ-phase). Particularly important are the austenitic chromium – nickel steels with ca. 18 wt % Cr and 10 wt % Ni, but without addition of molybdenum. Their strength is relatively low and their ductility very good. In chloride-containing corrodents, however, they undergo transgranular stress corrosion cracking.
4) Ferritic – austenitic steels combine good mechanical properties with improved resistance to stress corrosion cracking. They are therefore used mainly under conditions which may cause fatigue corrosion and stress corrosion cracking in austenitic steels.

Pure iron has a ferritic structure. The alloying elements that may be added to iron are either ferrite-forming (chromium, molybdenum, silicon, titanium) or austenite-forming (carbon, nitrogen, nickel, manganese).

The austenitic chromium – nickel steels, with or without molybdenum, are used in the solution-heat-treated state (solution heat treatment temperature 1000 – 1100 °C). At solution heat treatment temperatures these steels are close to the boundary of the $\alpha + \gamma$ field in the iron – chromium – nickel phase diagram (Fig. 11). As the content of chromium increases, more nickel or nitrogen must be added to maintain the austenitic structure. If the content of molybdenum is raised, then it is also necessary to increase the content of austenitizers and/or to reduce the chromium content.

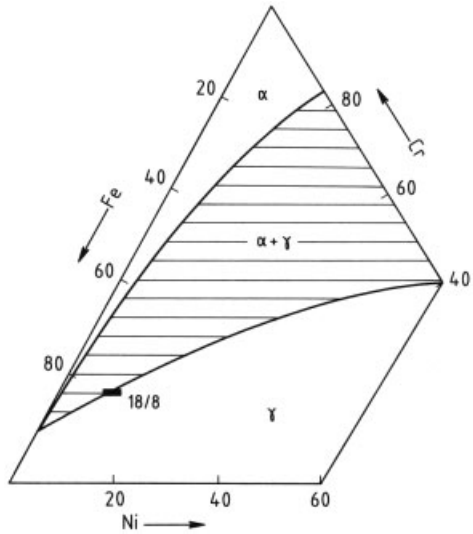

Figure 11. Ternary phase diagram of Fe – Cr – Ni (section at 1100 °C)

With increasing chromium content, and especially increasing molybdenum content, the tendency to segregate intermetallic compounds (σ-phase, χ-phase, Laves phase, Fe_2Mo), increases greatly (see Fig. 12). If the precipitation tendency is strong enough, intermetallic compounds may already be precipitated during the

relatively short-lived heat exposure of welding. Hence the addition of molybdenum, which is desirable for reasons of corrosion chemistry, has an upper limit. In the case of austenitic chromium – nickel steels it is about 6 wt % Mo.

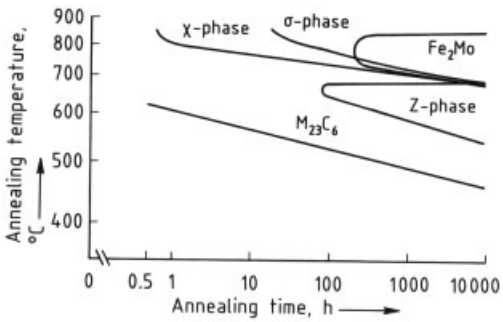

Figure 12. Precipitation fields of chromium-rich carbides $M_{23}C_6$, σ- and χ-phase, molybdenum-rich Laves phase (Fe_2Mo), and of a complex nitride (Z-phase) for the steel X 2 CrNiMo 18 14 3

Addition of nitrogen retards the precipitation of the intermetallic phases and compounds considerably, as well as improving the mechanical properties by introducing nitrogen atoms into interstices in the metal lattice. Nitrogen-alloyed austenitic and ferritic – austenitic steels are therefore of industrial importance.

In ferritic – austenitic steels, the ferritizers chromium and molybdenum are enriched in the ferrite phase, whose nickel content is correspondingly depleted. The austenitic phase contains, conversely, more nickel and less chromium and molybdenum. These differences of concentration are relatively slight, however, and are only important for the use of ferritic – austenitic steels in borderline cases. The alloy composition of these steels is also chosen such that the ferritic and austenitic phases are present in the structure in approximately equal proportions and the chromium content of the austenite is not less than 17 wt %.

Annealing within certain temperature ranges may cause carbides, nitrides, and intermetallic compounds to be precipitated in the microstructure of stainless steels. Owing to the special importance of chromium and molybdenum in imparting corrosion resistance to stainless steels, the precipitation of phases and compounds that contain these elements exert strong effects. The segregation of chromium- and molybdenum-rich phases depletes the matrix in these elements. The matrix therefore loses its resistance to corrosion.

In ferritic chromium steels and austenitic chromium – nickel steels, the chromium-rich σ-phase may precipitate on annealing at 600 – 900 °C. For ferritic chromium steels with up to ca. 18 wt % chromium this precipitation has no importance, while in the case of molybdenum-free austenitic chromium – nickel steels it is, at the most, important only in connection with the welded material. For reasons of welding technique (prevention of hot cracking) the welded material of normal austenitic steels almost always contains some δ-ferrite, which decomposes when the steel is annealed. In a corrosion exposure this decomposed δ-ferrite may be selectively dissolved. Damage must be expected where a coherent δ-ferrite network is present (proportion of δ-ferrite in the structure > 10 %).

The designations of the most important stainless steels for chemical plant are compiled in DIN 17 440 (Table 4, see next page).

Chemical engineering also uses a number of other special materials notable for their high resistance to pitting corrosion and stress corrosion cracking, as well as to mineral acids (Table 5, see next page).

Particularly high molybdenum contents are present in the steels listed in Table 6 (see next page). The stability of the austenite is due to nickel and nitrogen.

Unlike ferritic and austenitic steels, ferritic – austenitic steels have a two-phase structure, which, in contrast to the composition of austenitic steels, is obtained by raising the contents of ferrite-stabilizing elements, such as chromium and silicon, and reducing the austenite-stabilizing nickel content (Table 7, see page after next). By virtue of their high yield strength of at least 450 MPa, steels of this kind are used for components which, while exposed to corrosive media, are additionally subjected to wear (cavitation, erosion) and vibration. Being more resistant than commercial austenitic steels to stress corrosion cracking in neutral chloride solutions, they are also being used increasingly to handle aggressive cooling water. A commonly used material of this group is X 2 CrNiMoN 22 5 (1.4462), which, in addition to about 22 wt % Cr, has a nickel content of about 5.5 wt %, an Mo content of about 3 wt %, and an austenitic structure proportion of about 60 % [46], [47].

Table 4. Composition of stainless steels

Short name (DIN)	US Standard (AISI)	Material number	Composition, wt %				
			C	Cr	Mo	Ni	Others
Ferritic and martensitic steels							
X 7 Cr 13	403	1.4000	<0.08	12.0 – 14.0			
X 7 CrAl 13	405	1.4002	<0.08	12.0 – 14.0			Al 0.10 – 0.30
X 10 Cr 13	410	1.4006	0.08 – 0.12	12.0 – 14.0			
X 15 Cr 13		1.4024	0.12 – 0.17	12.0 – 14.0			
X 20 Cr 13	420	1.4021	0.17 – 0.22	12.0 – 14.0			
X 40 Cr 13		1.4034	0.40 – 0.50	12.0 – 14.0			
X 45 CrMoV 15		1.4116	0.42 – 0.48	13.8 – 15.0	0.45 – 0.60		V 0.10 – 0.15
X 8 Cr 17	430	1.4016	<0.10	15.5 – 17.5			
X 8 CrTi 17	439	1.4510	<0.10	16.0 – 18.0			Ti ≥ 7× % C
X 8 CrNb 17		1.4511	<0.10	16.0 – 18.0			Nb ≥ 12× % C
X 6 CrMo 17	434	1.4113	<0.07	16.0 – 18.0	0.90 – 1.20		
X 12 CrMoS 17	430 F	1.4104	0.10 – 0.17	15.5 – 17.5	0.20 – 0.30		S 0.15 – 0.35
X 22 CrNi 17	431	1.4057	0.15 – 0.23	16.0 – 18.0		1.5 – 2.5	
Austenitic steels							
X 12 CrNiS 18 8	303	1.4305	<0.15	17.0 – 19.0		8.0 – 10.0	S 0.15 – 0.35
X 5 CrNi 18 9	304	1.4301	<0.07	17.0 – 20.0		8.5 – 10.0	
X 5 CrNi 19 11	305/308	1.4303	<0.07	17.0 – 20.0		10.5 – 12.0	
X 2 CrNi 18 9	304 L	1.4306	<0.03	17.0 – 20.0		10.0 – 12.5	
X 10 CrNiTi 18 9	321	1.4541	<0.10	17.0 – 19.0		9.0 – 11.5	Ti > 5× % C
X 10 CrNiNb 18 9	347	1.4550	<0.10	17.0 – 19.0		9.0 – 11.5	Nb > 8× % C
X 5 CrNiMo 18 10	316	1.4401	<0.07	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	
X 2 CrNiMo 18 10	316 L	1.4404	<0.03	16.5 – 18.5	2.0 – 2.5	11.0 – 14.0	
X 10 CrNiMoTi 18 10	316 Ti	1.4571	<0.10	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	Ti > 5× % C
X 10 CrNiMoNb 18 10	316 Cb	1.4580	<0.10	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	Nb > 8× % C
X 5 CrNiMo 18 12	316	1.4436	<0.07	16.5 – 18.5	2.5 – 3.0	11.5 – 14.0	
X 2 CrNiMo 18 12	316 L	1.4435	<0.03	16.5 – 18.5	2.5 – 3.0	12.5 – 15.0	
X 2 CrNiMo 18 16	317 L	1.4438	<0.03	17.0 – 19.0	3.0 – 4.0	15.0 – 17.0	
X 2 CrNiN 18 10	304 LN	1.4311	<0.03	17.0 – 19.0		9.0 – 11.5	N 0.12 – 0.20
X 2 CrNiMoN 18 12	316 LN	1.4406	<0.03	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	N 0.12 – 0.20
X 2 CrNiMoN 18 13	316 LN	1.4429	<0.03	16.5 – 18.5	2.5 – 3.0	12.0 – 14.5	N 0.14 – 0.22

Table 5. Composition of special alloyed stainless steels

Steel type, short name (DIN)	Material number	Composition, wt %						
		C	Si	Mn	Cr	Ni	Mo	Other
X 3 CrNiMoN 17 13 5	1.4439	≤0.04	≤1	≤2	17.5	13.5	4.5	N
X 2 NiCrAlTi 32 20	1.4558	≤0.04	≤0.7	≤1	21.5	33.5		Ti, Al
X 1 NiCrMoCu 25 20	1.4539	≤0.02	≤1	≤2	20	25	4.5	Cu
X 1 CrNiMoN 25 25 2	1.4465	≤0.02	≤0.7	≤2	25	23.5	2.3	Al, Cu
X 1 NiCrMoCuN 31 27	1.4563	≤0.02	≤0.7	≤2	27	31	3.5	Cu
NiCr21Mo	2.4858	≤0.03	≤0.5	≤1	21.5	42	3	Ti, Al

Table 6. High-alloy stainless steels containing ≥ 6 wt % molybdenum

DIN designation	ASTM designation	Composition, wt %						Trade mark
		C	Cr	Ni	Mo	Cu	N	
	S 31254	0.02	20	18	6.1	0.7	0.2	Awesta 254 SMO
1.4529		0.02	20	25	6.5	1	0.15	Nirosta 4529
		0.02	17	16	6.3	1.6	0.16	VEW A 963
		0.03	20	24	6.0			AC 6 X

Table 7. Composition of ferritic – austenitic (duplex) steels

Name/material number	Short designation (DIN)	Typical composition, wt %						
		C	Si	Mn	Cr	Mo	Ni	N
1.4462	X 2 CrNiMoN 22 5 3	0.012	0.34	0.65	22.1	3.0	5.7	0.14
Ferralium 255	X 2 CrNiMoCu 25 5	0.019	1.4	1.4	26	3.3	5.5	0.2
Noridur 9.4460	G–X 3 CrNiMoCuN 24 6	0.03	1.5	1.5	25	3.1	6	+
A 905	X 3 CrMnNiMoN 25 4 4	0.034		5.8	26	2.3	3.7	0.36

A specially developed austenitic steel that contains about 4 wt % Si (X 2 CrNiSi 18 15, 1.4361) and has already undergone trials under practical conditions has remarkably good resistance to highly concentrated nitric acid and chromic acid. As Figure 13 shows, its advantages are particularly pronounced at HNO_3 concentrations above 75 %, where it is superior to stainless steels of all other types [48]. In the meantime a variant has been developed for extremely severe HNO_3 exposure (5.3 wt % Si, 17.5 wt % Ni, 0.015 wt % C) [49].

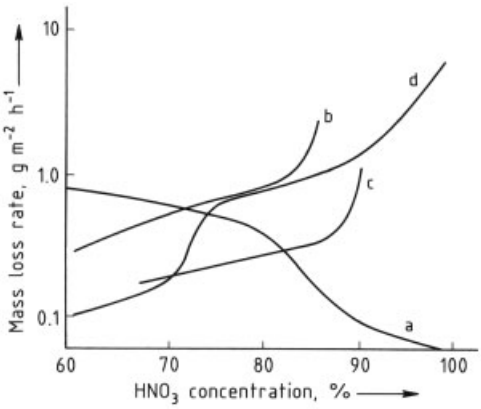

Figure 13. Area-related mass loss rate of stainless steels in boiling nitric acid (exposure time: 50 – 90 d)
a) X 2 CrNiSi 18 15; b) X 2 CrNi 18 9; c) X 1 CrMo 26 1; d) X 1 CrNi 25 21

Ferritic steels have also undergone special development. Because the mechanical properties of the welded joints were unsatisfactory, the commercial types used until now were rarely chosen for chemical plant. A substantial improvement may be expected from the titanium-stabilized chromium and chromium – molybdenum steels of low carbon and nitrogen content and from the nickel-alloyed chromium – molybdenum steels. Progress in the field of ferritic steels has been marked by the use of special methods for obtaining low C + N contents in their manufacture (ELI = extra low interstitial steels) [50]. Three groups can now be distinguished: 1) steels with 25 – 28 wt % Cr, without nickel, in some cases with an addition of Mo, and in all cases with a C + N content limited to 0.015 wt % (superferrites); 2) molybdenum-containing steels with 18 wt % Cr, which, although the proportions of the aforementioned interstitial elements have been lowered, are additionally stabilized with titanium or niobium; and 3) stabilized molybdenum- and nickel-containing ferritic steels with high Cr contents, such as X 1 CrNiMoNb 28 4 2 (Remanit 4575) and X 2 CrNiMoTi 25 4 4 (Monit).

The superior toughness of the last-mentioned steel types, as compared with that of a conventional Cr steel, is illustrated by the transition temperature of the notch impact strength.

Stabilized ferritic steels should be of particular interest in connection with the handling of corrosive media, especially chloride-containing cooling waters (e.g., river water), owing to their good corrosion properties, suitability for welding, and availability as semifinished products, as well as for economic reasons [51].

For reasons of cost, and because of certain design problems, stainless cast steel is often preferred to stainless rolled and wrought steels as a material for chemical plant. The chemical composition and mechanical properties of stainless cast steels are fairly similar to those of the rolled and wrought steels. Table 8 lists several important types, especially those used for chemical pumps.

5.1.6.1. Technical Properties

The mechanical properties of the four groups of stainless steels are shown in Table 9.

Table 8. Cast stainless and special steels

Steel type	Designation	Material number	Composition, wt %							
			C	Si	Mn	Cr	Ni	Mo	Cu	Others
Cast ferritic stainless steel	G-X 8 CrNi 13	1.4008	≤0.09	≤1.0	≤1.0	13.0	1.5	≤0.5		
Cast martensitic stainless steel	G-X 5 CrNi 13 4	1.4313	≤0.07	≤1.0	≤1.5	13.0	4.0	≤0.7		
Cast austenitic stainless steel	G-X 6 CrNi 18 9	1.4308	≤0.07	≤2.0	≤1.5	19.0	10.0			
	G-X 6 CrNiMo 18 10	1.4408	≤0.07	≤1.5	≤1.5	19.0	11.0	2.5		
	G-X 6 CrNiMo 1713	1.4448	≤0.07	≤1.0	≤2.0	17.0	13.5	4.5		
	G-X 5 CrNiNb 18 9	1.4552	≤0.06	≤1.5	≤1.5	19.0	10.0			Nb ≥ 8 × % C
	G-X 5 CrNiMoNb 18 10	1.4581	≤0.06	≤1.5	≤1.5	19.0	11.5	2.3		Nb ≥ 8 × % C
	G-X 7 CrNiMoCuNb 18 18	1.4585	≤0.08	≤1.5	≤2.0	17.5	20.0	2.3	2.1	Nb ≥ 8 × % C
Cast special austenitic steel	G-X 7 NiCrMoCuNb 25 20	1.4500	≤0.08	≤1.5	≤2.0	20.0	25.0	3.0	2.0	Nb ≥ 8 × % C
	G-X 3 CrNiSiN 2013	9.4306	≤0.04	4.5	4.5	20.0	13.0	≤0.05		N
Cast ferritic – austenitic stainless steel	G-X 3 CrNiMoCu 24 6	9.4460	≤0.04	≤1.5	≤1.5	25.0	6.0	2.4	3.1	N

Table 9. Mechanical properties of the four main steel types

Steel type	Ferritic [a]	Martensitic [b]	Austenitic [c]	Austenitic – ferritic [d]
Yield strength, MPa	ca. 300	450 – 600	ca. 230 [e]	ca. 500 [f]
Tensile strength, MPa	450 – 650	600 – 950	ca. 600	ca. 750
Fracture strain, %	ca. 25	14 – 18	45	ca. 30
Notch impact energy, J [g]	30 – 55		ca. 150	

[a] Normal steels with 16 – 18 wt % Cr, annealed.
[b] Quenched and annealed.
[c] Solution annealed.
[d] Annealed at 1100 °C and quenched.
[e] 0.2 % yield strength.
[f] 1 % yield strength;
[g] DVM specimen, longitudinal.

Between their melting points and room temperature, ferritic steels are pure or mainly ferritic. The production of a fine-grained structure requires forming at temperatures below 800 °C, heating beyond 800 °C, and cooling. Heating above 1000 °C (welding) enlarges the grain, thus embrittling the steel and causing precipitation of carbides. Temperatures above 1100 °C affect the structure to such an extent that the steel breaks when subsequently loaded. These steels cannot be hardened by heat treatment. Being less expensive than austenitic steels, and more resistant to chloride-containing solutions, they are used for pipelines intended for chloride-containing waters, as well as for heat exchangers and condensers. Martensitic steels exist only in the pearlitic and martensitic forms, without intermediate stages. They are hardened by heating to 980 – 1100 °C, depending on carbon content, cooling in oil or air, and tempering at a temperature above 600 °C. The steels containing less than 0.4 wt % carbon serve as heat-treatable materials with good mechanical properties, while those with higher carbon contents are used as hardened steels. Certain types are resistant to very high temperatures and to pressurized water. Having good resistance to erosion and cavitation as well as high fatigue strength, they are used also for propellers and impellers. Austenitic steels with low chromium and nickel contents are so unstable that martensite may be formed at low temperatures or through cold

forming. High chromium contents favor the formation of ferritic structures. Fine granulation results from recrystallization during hot forming. The grain-growth rate is much lower than that of ferritic steel. The final heat treatment consists of heating to 1000 – 1100 °C and quenching in water or air. Hardening by heat treatment is not possible; the strength is improved by cold or hot – cold forming. The most conspicuous properties of these steels are a low yield strength, high resistance to fracture, high strength at elevated temperatures, resistance to pressurized water, heat resistance, permanent strain at low loads, and a tendency to creep. An important application is the fabrication of pressure vessels.

In austenitic – ferritic steels the ratio of ferrite to austenite is determined by the composition and heat treatment. It increases with the annealing temperature and rate of cooling. Attention is drawn to the risk of 475 °C embrittlement (hardness increase, loss of toughness, and loss of chemical resistance through separation into phases of high iron and chromium content [52] and σ-phase formation), as well as to the fact that, in comparison with purely austenitic steels, these steels are less sensitive to stress corrosion cracking. Their applications include pump manufacture. Special grades with high wear resistance are used for bushings.

Ferritic, austenitic, and ferritic – austenitic stainless steels can be welded by practically all the well-known methods. The argon arc (TIG, MIG) process, resistance welding (spot or seam welding), and the plasma arc welding and electron beam welding are particularly suitable for ferritic chromium steels. Austenitic chromium – nickel steel is recommended as the electrode material in addition to material of the same composition as the base. If, for reasons of corrosion (particularly the risk of stress corrosion cracking), either of these materials cannot be used, an austenitic filler can serve for the lower layers and a ferritic filler for the cover pass. In view of the marked grain growth tendency, as little heat as possible should be introduced. It is therefore advisable to use thin electrodes and low currents.

As scale formation may be accompanied by the formation of chromium-depleted surface zones, protective gas should be used to prevent scale formation on welding. Cleanliness of the seam, which must be without splashes and undercutting, is also important. Where the demands are particularly severe, as in the case of components subjected to vibration, the surface quality can be improved by subsequent grinding.

Unstabilized 17 wt % chromium steels that have been welded must always be annealed since they are otherwise sensitive to intergranular corrosion, even in mild corrodents. Where this treatment is impracticable, stabilized grades should be used. Although the latter benefit from heat treatment, it is not absolutely necessary. Stabilized steels and molybdenum-containing steels also exhibit lower growth tendency. Apart from imparting resistance to intergranular corrosion, additions of stabilizing elements also improve the resistance to pitting corrosion.

The principal welding methods for austenitic steels are TIG, MIG, and submerged-arc welding. Apart from using steels that have no intergranular susceptibility after welding (stabilized steels and ELC types) it is particularly important to avoid hot cracking and loss of corrosion resistance through precipitation of intermetallic phases. Appropriate recommendations can be found in the processing guidelines.

5.1.6.2. Chemical Properties

The resistance of stainless steels to corrosive media is determined by their passive behavior, which depends in turn on the alloying elements and their concentrations.

In this connection, chromium is particularly important, but the effects of nickel, molybdenum, and copper are also significant. In particular, chromium reduces the passivation current density and the residual current density of the steel in the passive state. The main effect of molybdenum is a reduction of the passivation current density and facilitation of repassivation, which promotes the transition from the active to the passive state.

In general it is considered that a corrosion rate of 0.1 mm/a (0.1 g m^{-2} h^{-1}) represents the highest degree of corrosion resistance attainable under practical conditions. Corresponding data can be found in stability tables and diagrams and in manufacturers' literature. Often, however, stability diagrams, of which Figure 14 is an example, apply only to chemically pure corrodents. In the case of sulfuric acid, for example, even relatively small amounts of oxidizing agents(NO_2^-,

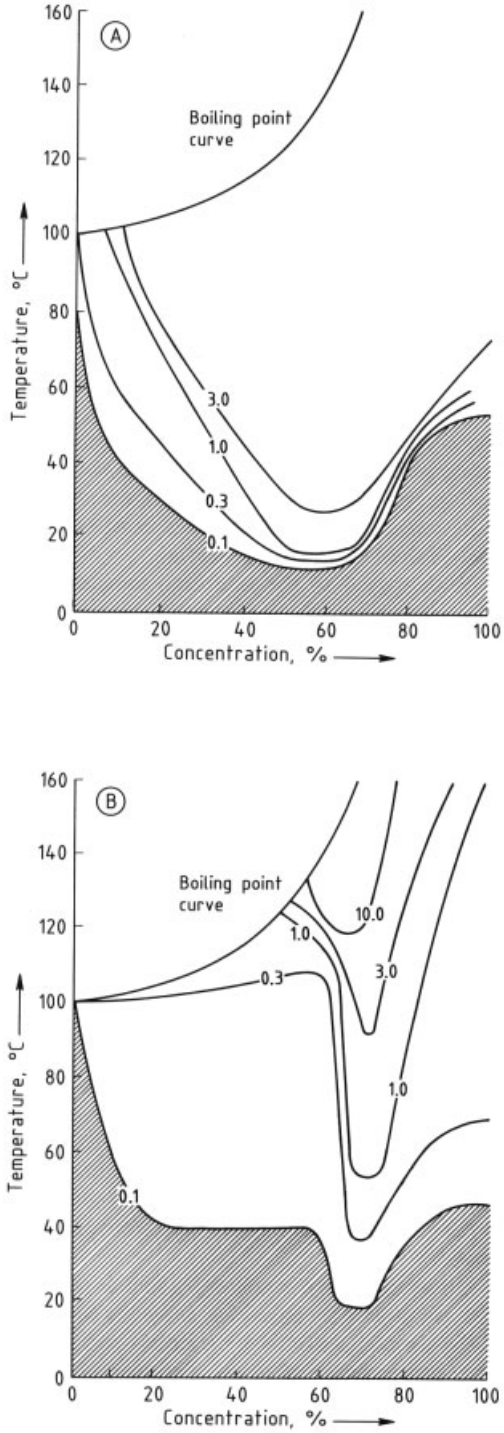

Figure 14. Isocorrosion curves of chromium – nickel – molybdenum stainless steels in sulfuric acid
A) 0.05 % C, 17 % Cr, 13 % Ni, 5 % Mo; B) 0.05 % C, 20 % Cr, 25 % Ni, 3 % Mo, 2 % Cu

Fe^{3+}) permit the use of stainless steels in sulfuric acid at concentrations up to ca. 50 % and temperatures considerably higher than those shown in Figure 14. Thus figures for resistance to pure corrodents should in many cases be regarded as initial approximations. The influences of ventilation, i.e., the corrodent's oxygen content, and of the state of motion (e.g., laminar or turbulent flow) must be taken into account also.

Uniform surface corrosion is of relatively minor importance for stainless steels. Localized corrosion, however, may lead quickly to component failure even where the corrosion rate is low, and must therefore be avoided as far as possible. Normally, the susceptibility of metallic materials to localized corrosion increases with the resistance to general corrosion. In localized corrosion of stainless steels, the locally active areas are very small compared to the area of the passive surface. Therefore, the local anodic dissolution current density is very high. Damage caused to passive materials by stress corrosion cracking and fatigue corrosion, which occur with virtually no loss of mass, is particularly hazardous. The corrosion may penetrate even relatively thick components within a short time. The types of localized corrosion that are most important in stainless steels are intergranular corrosion (grain-boundary attack), pitting corrosion (pit formation), stress corrosion (corrosion cracking), and fatigue corrosion (corrosion fatigue cracking).

Intergranular corrosion is a selective process in which the areas close to the grain boundaries are corroded. In all cases it has one of the following causes:

1) Depletion of alloying elements important to chemical resistance in areas close to grain boundaries, caused by formation of precipitates, in which these elements have accumulated, at the grain boundaries. In joint welding, these precipitates may be formed in the heat-affected zones bordering on the weld seams.
2) Chemical attack on precipitates at the grain boundaries, e.g., by concentrated nitric acid on precipitates of titanium carbonitrides in titanium-stabilized steels.
3) Preferential attack by accumulations of accompanying and trace elements in the steel, e.g., phosphorus, silicon, and boron (which

stimulate the anodic metal dissolution at the grain boundaries).

4) Cell formation between the matrix and precipitates at the grain boundaries.

In stainless steels, the depletion of the alloying elements chromium and molybdenum resulting from precipitation of carbides with high chromium and molybdenum contents is the most important cause of intergranular corrosion in industry.

At low temperatures the solubility of carbon in austenitic chromium – nickel steels and nickel – chromium alloys is very low, but it increases markedly as the temperatures rises. The carbon dissolved in the austenitic solid solution at high annealing temperatures may precipitate at the grain boundaries at lower annealing temperatures as a carbide of type $M_{23}C_6$ with high chromium and molybdenum content. In this way zones of depleted chromium and molybdenum content, which grow together to form coherent chromium-depleted areas as the carbide precipitation progresses, are formed round the precipitated carbides. Through the precipitation of chromium-rich carbides (75 – 90 wt % Cr) the chromium content in the vicinity of the grain boundaries falls below the so-called resistance limit of ca. 12 wt % chromium.

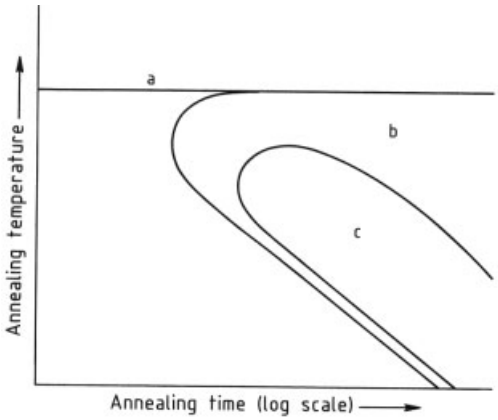

Figure 15. Diagram of grain-boundary disintegration
a) Solubility limit for carbon; b) Precipitation of chromium-rich carbides; c) Intergranular corrosion

The susceptibility to intergranular corrosion is indicated by the grain-boundary corrosion fields in the plot of annealing time versus annealing temperature (Fig. 15, see above). The shape and sizes of these fields are best described by the highest temperature that still causes susceptibility to intergranular corrosion and by the shortest annealing time leading to sensitization of grain boundaries. This shortest annealing time (represented by the "nose" of the grain-boundary corrosion field) is situated in the temperature range around 650 °C for molybdenum-free austenitic steels and in the range around 750 °C for molybdenum-containing steels.

When the temperature falls after welding, it passes through the range in which austenitic chromium – nickel steels become susceptible to intergranular corrosion. The following steps are taken to make the steel resistant to intergranular corrosion after welding:

1) The carbon content is reduced to levels at which neither temperature nor the exposure time causes sensitization (ELC steels).
2) Stabilization with additions of the carbide-forming element titanium or niobium/tantalum. The affinity of these alloying elements for carbon is greater than that of chromium. They form special carbides or carbonitrides of the type M (C, N). Due to the affinity of both elements for nitrogen, the steel's nitrogen content, as well as its carbon content, must be taken into account in calculating the stabilization ratio.

In ferritic chromium steels, the rate at which chromium-rich carbides are precipitated is very high. After annealing at temperatures above 850 °C, even very rapid cooling cannot prevent their precipitation, and the formation of a structure susceptible to intergranular corrosion. In the temperature range around 800 °C, however, the very high diffusion rates permitted by the body-centered cubic lattice of the α mixed crystal causes rapid replenishment of chromium by diffusion from the chromium-rich matrix into the chromium-depleted grain boundary areas. As a reduction of the carbon content would not eliminate intergranular susceptibility until an extremely low level had been reached (< 0.005 wt %), stabilization, even of ferritic chromium steels, by addition of niobium/tantalum, and especially titanium, has great industrial importance. Pitting corrosion of stainless steels occurs only when they are in the passive state, and is then caused almost exclusively by chloride ions in aqueous solution. It is therefore also known as chloride corrosion.

In the initiating step of pitting corrosion chloride ions are adsorbed locally by the passive layer, whereupon the passivity breaks down. The mechanism of this corrosion has not yet been fully explained.

In the pitting corrosion of stainless steels, the influence of potential is important. Pitting corrosion only occurs when the potential exceeds the pitting potential. In the presence of a corrosive medium, pitting corrosion therefore occurs only if the corrosion potential of the steel is more positive than the pitting potential.

With regard to the attacking medium, the main factors by which the pitting potential is lowered (and therefore the risk of pitting corrosion heightened) are increase of chloride concentration and temperature, and reduction of pH [53].

With regard to the material, the pitting potential can be shifted towards more positive potentials by raising the chromium content and especially the molybdenum content; the alloying elements nickel and manganese, on the other hand, have practically no influence.

The effects of chromium and molybdenum are described, and the resistances of the various alloys estimated, with the help of the pitting resistance equivalent PRE [54]:

$$\mathrm{PRE} = \mathrm{wt}\,\%\,\mathrm{Cr} + 3.3 \times \mathrm{wt}\,\%\,\mathrm{Mo}$$

Over a wide range of percentages of these elements, the pitting potential depends linearly on the pitting resistance equivalent, in as far as the structure is that given by solution annealing. Deviations are attributable to homogenization of the structure by relatively high nitrogen contents. Stress corrosion cracking, which is likewise initiated mainly by chlorides, accounts for most of the damage suffered by stainless steels. The basic prerequisites of stress corrosion cracking are as follows:

1) The action of a specific corrodent on a material susceptible to stress corrosion cracking, i.e., the presence of a critical material/corrodent system.
2) Static tensile stresses, possibly superimposed by infrequent dynamic loads. Only tensile loads are effective. Pressure does not cause stress corrosion cracking; in fact, in many cases it is an effective means of preventing corrosion of this type.

Stress corrosion cracking of stainless steels with mainly transgranular progression of the corrosion cracks may occur in the following corrodents:

1) Chloride solutions (threshold temperature 50 °C);
2) Alkaline solutions (threshold temperature at 40 – 50 wt % NaOH, 100 °C). In addition, intergranular cracking occurs at high temperatures in dilute alkaline solutions.

If the steel is in the sensitized state in which it is susceptible to intergranular cracking, and the tensile stress and potential exceed critical values, intergranular stress corrosion cracking occurs also in, e.g., water at high temperatures (Fig. 16) [55].

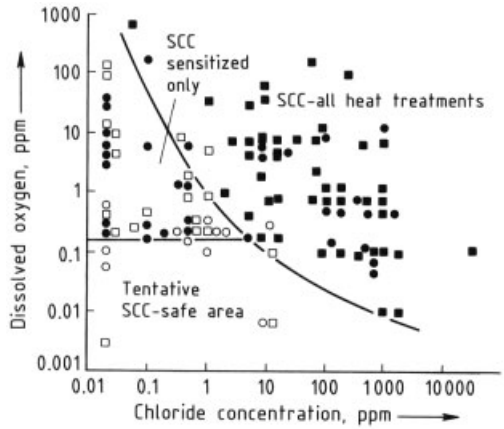

Figure 16. Concentration ranges of dissolved oxygen and chloride that may lead to stress corrosion cracking (SSC) of type 304 stainless steels in water at temperatures in the range 260 – 300 °C. Applied stresses in excess of yield strength and test times in excess of 1000 h, or strain rates greater than $10^{-5}\,\mathrm{s}^{-1}$
● = Sensitized, SCC; ○ = Sensitized, no SCC; □ = Annealed, SCC; □ = Annealed, no SCC

Under conditions that cause stress corrosion cracking the following materials can be used instead of normal stainless austenitic 18 Cr 10 Ni steels: stainless ferritic chromium steels with 17 – 28 wt % chromium, with or without molybdenum; austenitic – ferritic steels (duplex steels), whose use is restricted in some cases to temperatures below 300 °C; austenitic steels with increased nickel and/or molybdenum content; and nickel-based alloys. Apart from molybdenum, an alloying element that decisively improves the stress corrosion cracking resistance

of stainless austenitic steels is nickel (at nickel contents exceeding about 40 wt %, stress corrosion cracking no longer occurs). Corrosion fatigue is the term used for crack damage which occurs in a component exposed simultaneously to alternating mechanical stress and a corrodent. It may be caused in all metallic materials by unspecific corrodents, differing in this respect very much from stress corrosion cracking, for whose occurrence specific corrodents are responsible. The main symptom of corrosion fatigue is the loss of endurance strength.

Fatigue cracking of stainless steels differs according to whether they are in the active or passive states. In the active state numerous cracks are formed, whereas in the passive state few cracks, or perhaps just a single precrack, are formed. Damage to stainless steel plant through fatigue cracking is, however, relatively rare [56].

5.1.6.3. Development State of Stainless Cr – Ni Steels

The corrosion resistance of stainless steels is benefitting from the attainment of particularly high degrees of purity, as given by electroslag remelting, for example. At the same time the strength-related properties are being improved without loss of ductility.

To make possible the increases in Cr and Mo content on which these improvements depend it is necessary to increase the nitrogen content. This is achieved by renitrogenizing the steel under pressure and simultaneously adding alloying elements that increase the nitrogen's solubility [57], [58], e.g., manganese.

Of two steels now undergoing trials, namely X 3 CrNiMoNoN 23 27 6 4 and X 3 CrNiMnMoNbN 23 17 8 4, the latter is superior in respect of strength-related properties in consequence of precipitation hardening by a finely dispersed intermetallic Z-phase (0.2 % yield strength at room temperature ca. 500 MPa; at 500 °C, > 300 MPa). This steel, which is without Nb and therefore free from precipitation, is superior in ductility (notch impact energy at room temperature > 300 J), though at 500 °C it still has a 0.2 % yield strength of ca. 250 MPa [59].

The stress corrosion cracking behavior of steels of this type is shown in Figure 17. The steel with the highest N content is the one most resistant to boiling 35 % $MgCl_2$ solution. In boiling seawater all the steels achieved times to failure of more than 1000 h at loads above the 0.2 % yield strength [60].

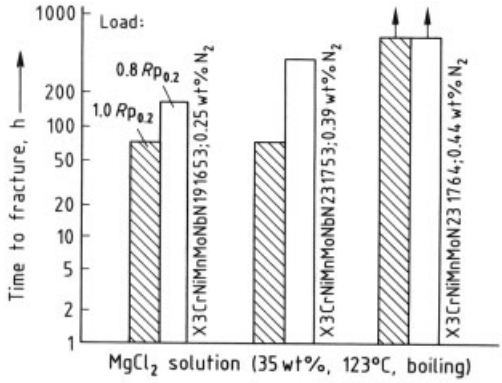

Figure 17. Stress corrosion behavior of nitrogen-containing stainless steels in boiling $MgCl_2$ solution (35 wt %, 123 °C)

The steel X 3 CrNiMnMoNbN 23 17 5 3 (1.4565) is commercially available. With a pitting resistance equivalent of about 48 (in which the high nitrogen content is taken into account) it is more resistant to pitting corrosion than any other stainless steel (cf. Fig. 18). Only NiCr 22 Mo 9 Nb, better known as Incoloy 625, performs comparably well.

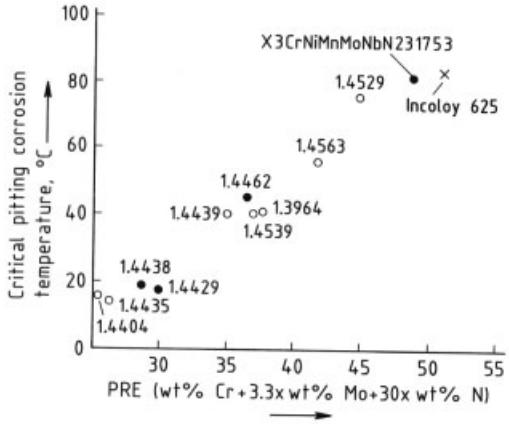

Figure 18. Pitting resistance equivalent (PRE) and critical pitting corrosion temperature for various steels and Incoloy 625 in a 10 wt % solution of $FeCl_3 \cdot 6\,H_2O$

Some mechanical properties of alloy 1.4565 are as follows:

0.2 % Yield strength	$\geq 500\,N/mm^2$
0.2 % Yield strength at 300 °C	$\geq 300\,N/mm^2$
Notch impact energy at − 269 °C	$\geq 200\,J$
Elongation	> 50 % at room temperature
Reduction of area	> 70 % at room temperature

As the nitrogen contents of the aforementioned steels in the molten state are below the solubility limit, these materials have good welding behavior. The filler metals used for them at present are relatively highly alloyed nickel-based materials of the Ni – Mo – Cr type.

5.2. Cast Iron [60], [61] (→ Steel)

The term cast iron is used for unalloyed, low-alloy, and highly alloyed iron – carbon cast materials containing ca. 2 – 5 wt % carbon and 0.8 – 3 wt % silicon, the carbon being mainly in the form of graphite. The fundamental structure of unalloyed and low-alloy cast iron is ferritic, ferritic – pearlitic, or pearlitic; that of highly-alloyed cast iron may be any of these, but also austenitic. The main alloying elements are nickel, chromium, and copper. The main quality standards are DIN 1691, for unalloyed and low-alloy lamellar graphite cast iron; DIN 1693, for unalloyed and low-alloy spheroidal graphite cast iron; and DIN 1694, for austenitic cast iron.

Mechanical properties of the various types of cast iron are listed in Table 10. The strength of gray cast iron depends on the thickness of the casting's cross section (wall thickness). Empirical values for the strength of castings are given in Figure 19.

The exceptionally low plastic deformability and impact strength of lamellar graphite cast iron in comparison with nodular graphite cast iron are explained by the graphite lamellae, which act as internal notches. Unalloyed and low-alloy lamellar graphite cast iron and nodular graphite cast iron are attacked only slightly by pure sulfuric acid and pure phosphoric acid. Their resistance to soil corrosion and seawater attack is also good. Depending on their composition and the form of the graphite, austenitic cast materials [62], [63] withstand corrosion by alkalis and sulfuric acid. The austenitic grades that withstand moderately high and high temperatures are used for pumps, valves, cylinder linings, waste-gas pipes, and furnace parts, for example. The corrosion-resistant grades are used in the food, artificial silk, and plastics industries as well as for pipes and vessels. Mention should also be made of the thermal shock stability of GGL(laminar)NiCr 30 3 (material no. 0.6676) and the low-temperature ductility of GGG(nodular)NiMn 23 4 (material no. 0.7673).

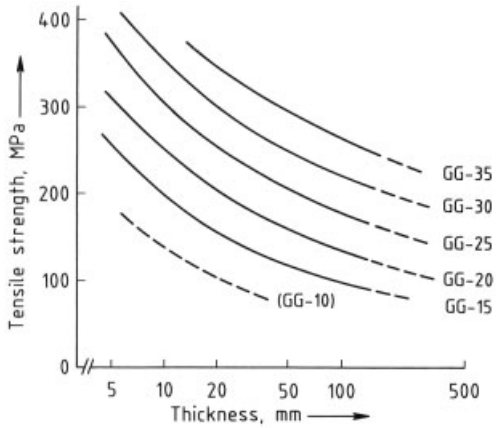

Figure 19. Mean tensile strength of gray cast iron versus the section thickness of the casting (according to DIN 1691)

A special grade that withstands hot nitric acid and hot sulfuric acid is the iron – silicon material containing 14 – 18 wt % silicon. Its hardness and brittleness constitute a great disadvantage because they restrict the choice of shaping techniques to casting and grinding.

Lamellar and nodular graphite cast iron can be welded hot with filler materials of the same

Table 10. Mechanical properties of the various types of cast iron

Cast iron type	Tensile strength σ_u, MPa	Fracture strain, %	Notch impact energy, J
Unalloyed and low-alloyed laminar	100 – 450	< 1	
Unalloyed and low-alloyed nodular	400 – 800	2 – 22	9 – 14 (− 20 °C)
Austenitic laminar	140 – 280	1 – 3	
Austenitic nodular	390 – 500	1 – 40	7 – 34

composition, and cold with filler materials of different composition [64], [65].

5.3. Nickel and Nickel Alloys [66–68]

(→ Nickel, → Nickel Alloys)

Owing to their excellent corrosion resistance to various aggressive media, nickel and its alloys are of growing importance in chemical engineering. Furthermore, Ni – Cr alloys, especially those with additions of silicon and aluminum, have good high-temperature strength by virtue of their high creep strengths, and good heat-resistance to hot gases and combustion products.

The most important unalloyed nickel grades are Ni 99.6 (2.4060), nickel content ≥ 99.6 wt %; LC-Ni 99.6 (2.4061), nickel content ≥ 99.6 wt %; Ni 99.2 (2.4066), nickel content ≥ 99.2 wt %; LC-Ni 99 (2.4060), nickel content ≥ 99.0 wt %.

The nickel content (according to DIN 17 740) may include up to 1 wt % cobalt. Ni 99.6 may contain up to 0.08 wt % carbon, and Ni 99.2 up to 0.1 wt % carbon; after prolonged exposure to temperatures above about 300 °C, the carbon may be precipitated as grain boundary graphite, thus making the material brittle [69]. Frequent use is therefore made of LC (low-carbon) Ni, which has a maximum carbon content of 0.02 wt % [70], [71].

Plant consisting of nickel or nickel-plated steels is used mainly to produce and process alkalis [71], [72]. The good resistance of pure nickel to sodium hydroxide solution can be seen in Figure 20. Nickel is also used to handle acid halides [73], seawater, brackish water, dilute air-free and nonoxidizing mineral acids and salt solutions, and fatty acids. It can also be exposed to dry chlorine or hydrogen chloride at temperatures up to 535 °C.

Nickel has good high-temperature strength. Up to about 300 °C its yield strength and tensile strength undergo practically no change [69]. Nickel is very sensitive to sulfur and sulfur compounds [69], [70], [74]. Above 400 °C sulfur penetrates nickel along the grain boundaries and forms a eutectic mixture of nickel and nickel sulfide, embrittling the material. This process is rapid at 550 – 650 °C. The surface must therefore be cleaned thoroughly before welding and each annealing treatment. Sulfur-free lubricants must be used in hot and cold forming. The furnace atmosphere for heat treatment should be sulfur-free and reducing or neutral; if freedom from sulfur cannot be ensured, the furnace atmosphere should be slightly oxidizing.

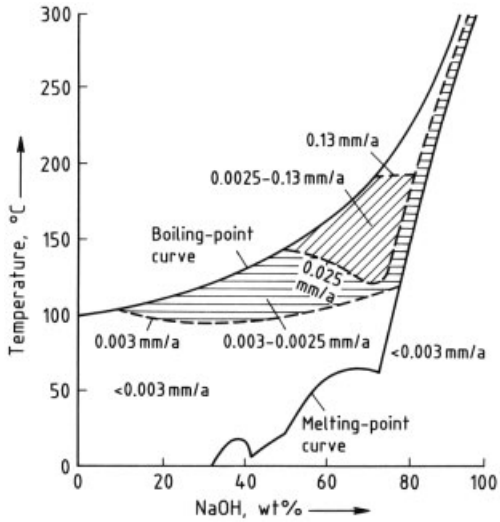

Figure 20. Corrosion resistance of pure nickel in sodium hydroxide, related to concentration and temperature [72]

All the welding methods normally used for steel, except submerged-arc welding [74], are suitable for nickel. Suitable weld filler materials are available for combinations with unalloyed or low-alloy steels, austenitic stainless steels, and nickel – copper, copper – nickel, and nickel – molybdenum alloys. Nickel parts can also be joined by brazing and soldering.

Cast nickel [70] has the characteristic properties of forged nickel, but is always slightly alloyed with carbon, silicon, and manganese to improve the flow of the molten metal and density of the castings.

5.3.1. Nickel – Copper Alloys

In chemical engineering only two alloys with more than 50 wt % nickel have gained importance, namely NiCu30Fe (2.4360, Inconel alloy 400) and NiCu30Al (2.4375), which are standardized in DIN 17 743. Both are sold as sheets, strip, pipe, rods, wires, and forgings.

The alloy NiCu30Fe can be exposed to seawater and brackish water, solutions of nonoxi-

dizing salts, alkalis [75], cold or warm nonoxidizing acids [73], and dry bromine. It is also resistant to hydrofluoric acid [76], though stress corrosion cracking may occur in hydrofluoric acid vapors containing air, as well as to a lesser extent in the liquid phase. Stress corrosion cracking has been caused in these alloys by corrosive media containing mercury or its compounds. Where there is a risk of stress corrosion cracking, stress relief annealing should be carried out at 550 °C, and in exceptional cases at up to 650 °C, and followed by slow cooling in the furnace. As NiCu30Fe is expensive, solid sheet is often replaced in the manufacture of process plant by steel sheet plated with this alloy.

NiCu30Al is an age-hardenable alloy with similar corrosion behavior to NiCu30Fe. Because of the risk of stress corrosion cracking it should be used in the age-hardened state; it cannot be exposed to hydrofluoric acid, however.

Both materials are weldable by the usual methods; they can also be brazed. Soft soldering is less frequently used. Like pure nickel, both alloys are sensitive to sulfur and sulfur compounds above ca. 400 °C.

5.3.2. Nickel – Chromium Alloys [68]

Nickel – chromium alloys, with various iron contents and up to ca. 9 wt % molybdenum and 2 wt % copper, are normally used where the corrosion resistance of highly alloyed stainless steels is inadequate. They exhibit good resistance to acids and to pitting corrosion, crevice corrosion, and stress corrosion cracking. The types most important in chemical engineering, together with their main applications, are listed in Table 11.

The molybdenum-free alloy NiCr15Fe, like nickel, has good resistance to caustic alkalis, even surpassing it in this respect under oxidizing conditions in consequence of its high chromium content [77].

NiCr15Fe, NiCr23Fe and NiCr29Fe have greater high-temperature strength than nickel and can be exposed to oxidizing sulfur-free atmospheres at temperatures up to 1100 °C, and to reducing sulfur-free atmospheres at up to 1150 °C. These alloys are also used in nitriding gases (480 – 600 °C) and under carburizing conditions (800 – 950 °C). Like pure nickel and the other nickel alloys, they are sensitive to sulfur and its compounds at temperatures above ca. 400 °C; this must be taken into account where processing, e.g., by welding, is envisaged. Nickel – chromium – iron – molybdenum alloys combine the good resistance of nickel – molybdenum alloys under reducing conditions with the good resistance of nickel – chromium alloys under oxidizing conditions. This explains their good resistance to acids and salt solutions. With the low-carbon grades and those stabilized with titanium or niobium there is virtually no risk of intergranular corrosion.

Chromium – nickel casting alloys (60/40, 50/50 and 35/65) are used in chemical engineering and have good resistance to corrosion by fuel ash [78].

5.3.3. Nickel – Molybdenum and Nickel – Molybdenum – Chromium Alloys

Nickel – molybdenum and nickel – molybdenum – chromium alloys are among the chemically most resistant metallic materials. The main

Table 11. Nickel – chromium alloys

DIN designation	Material no.	U.S. trade name	Exposure tolerated
NiCr15Fe	2.4816	Inconel alloy 600	seawater, cooling water, alkalis, hot gases, combustion products
NiCr23Fe	2.5851	Inconel alloy 601	as 2.4816
NiCr29Fe	2.4642	Inconel alloy 690	alkalis
LC-NiCr15Fe	2.4817		
NiCr20CuMo	2.4660	Carpenter 20 Cb-3	sulfuric acid, phosphoric acid
NiCr21Mo	2.4858	Inconel alloy 825	seawater, cooling water, alkalis
NiCr21Mo6Cu	2.4641		sulfuric acid, phosphoric acid, seawater, cooling water
NiCr22Mo6Cu	2.4618		as 2.4641
NiCr22Mo7Cu	2.4619	Hastelloy alloy G-3	acids, mixed acids
NiCr22Mo9Nb	2.4856	Inconel alloy 625	acids, seawater, cooling water
NiCr21Fe18Mo	2.4603	Hastelloy alloy G-30	acids, HNO_3 – HF pickling acid

representatives of this group are listed in Table 12.

Table 12. Nickel – molybdenum and nickel – molybdenum – chromium alloys

Material no.	Designation	US trade name
2.4617	NiMo28	Hastelloy alloy B-2
2.4610	NiMo16Cr16Ti	Hastelloy alloy C-4
2.4819	NiMo16Cr15W	Hastelloy alloy C-276
2.4608	NiCr21Mo14W	Hastelloy alloy C-22

Nickel – molybdenum alloys are not passifiable, since they contain no chromium. Their good resistance to reducing acids is explained by the low rate of corrosion in the active state, for which molybdenum is responsible.

The alloy NiMo30 (2.4810), consisting of nickel, 26 – 30 wt % molybdenum and 5 – 7 wt % iron withstands hydrochloric acid at all concentrations. Nickel – molybdenum alloys of this kind cannot be used under oxidizing conditions, however, because even atmospheric oxygen increases the rate of corrosion considerably. Alloying with chromium makes them passive in oxidizing corrodents, and therefore resistant to corrosion under these conditions also. In addition they retain their low rates of corrosion in the active state; therefore, like NiMo16Cr16Ti (2.4610) and NiCr21Mo14W (2.4608), for example, they can be used under reducing and oxidizing conditions. NiCr21Mo14W has proved itself under severely oxidizing conditions.

Since the corrosion resistance of nickel – molybdenum and nickel – molybdenum – chromium alloys depends on molybdenum and/or chromium, precipitation of intermetallic compounds containing these elements may make the materials susceptible to intergranular corrosion. Precipitation of molybdenum-rich compounds leads to intergranular susceptibility in the active state (in hydrochloric acid, for example), while precipitation of compounds rich in chromium causes susceptibility to intergranular corrosion and intergranular crack formation in the passive state [79].

Exact proportioning of the alloying elements, but, above all, minimization of the silicon and carbon contents (< 0.03 wt %) and additional stabilization [80], [81], give alloys that tolerate hot forming, annealing, and welding without acquiring intergranular sensitivity; this is demonstrated by the time – temperature precipitation diagram of Hastelloy alloy C-4 (Fig. 21).

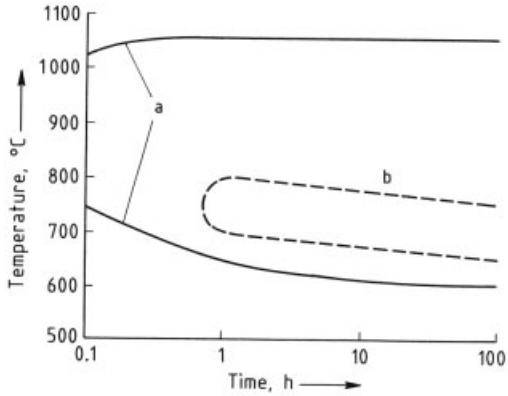

Figure 21. Time – temperature sensitization diagrams for Hastelloy alloy C-276 (a) and Hastelloy alloy C-4 (b); test for intercrystalline corrosion in accordance with ASTM G-28

Hot and cold forming of these high-nickel materials is difficult owing to their high water resistance and work-hardening tendency. They are weldable by all the commonly used methods. In the production of weld joints it is advisable, in the interest of good corrosion resistance, to start by making a thin root pass and then to weld the filler and cover passes while cooling the root side of the weld. If the design and size of a component permit heat treatment, solution annealing at 1150 °C, followed by water quenching, is recommended [79]. Steel can be explosion-cladded with any of these alloys, and they are often used for this purpose. Above 400 °C they are sensitive to sulfur and sulfur compounds, like all materials with high nickel contents; this must be taken into account in connection with annealing and welding.

5.4. Aluminum and Aluminum Alloys

(→ Aluminum, → Aluminum Alloys)

The high strength and low density of aluminum make it highly suitable for lightweight structures. Its economic value in the chemical industry, however, depends more on its corrosion behavior [82], [83]. Pure and very pure aluminum according to DIN 1712 comprise the grades Al 99, Al 99.5, Al 99.8, and Al 99.99.

The good corrosion resistance of aluminum, in spite of its very negative standard electrode

potential of − 1.66 V, is due to the formation of nonconducting Al_2O_3 passive films. The corrosion resistance depends on the purity of the material, which for chemical plant should not be below 99.5 wt % Al. As the protective Al_2O_3 surface layer is largely insoluble in the pH range of 4.5 – 8.8 (Fig. 22), aluminum materials have very good resistance to corrosion in approximately neutral aqueous media. The corrosion behavior and the main applications are listed in Table 13 [82].

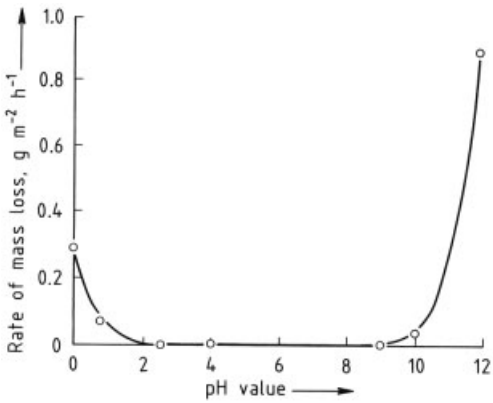

Figure 22. Influence of pH on the solubility of the Al_2O_3 film on aluminum

Because the Al_2O_3 passive film is nonconducting, relatively thick corrosion-inhibiting Al_2O_3 layers can be produced by anodic oxidation. They consist of a very thin and almost nonporous dielectric underlayer (barrier film) and a fine-pored top layer (Fig. 23), which can be dyed and compacted [84]. Hard anodizing is an anodic oxidation process that gives particularly hard and wear-resistant oxide films for technical purposes [85]. Apart from aluminum itself, numerous wrought and cast alloys based on Al – Mg, Al – Mn, Al – Mg – Si, and Al – Zn – Mg are suitable for hard anodizing.

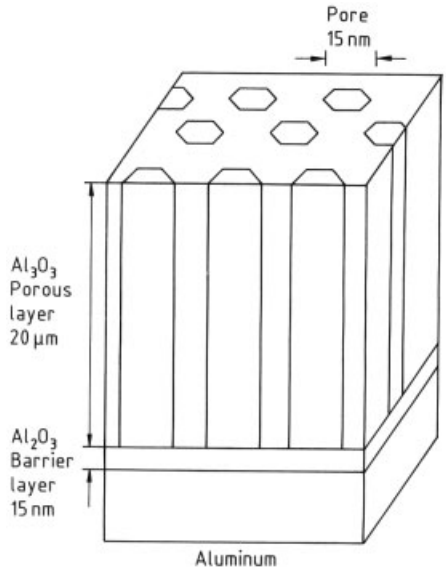

Figure 23. Structure of an oxide film on aluminum, produced by anodizing

High-purity aluminum is used in chemical engineering mainly as a cladding material.

Table 13. Examples for the corrosion behavior of aluminum alloys in various media

Corrosive environment	Corrosion resistance*			Application
	Al 99.5	AlMg	AlCuMg	
Ethanol				
water-containing	1 – 2			
dry	5			
Acetylene, dry	1		2 – 3	Pressure bottles
Liquid ammonia, dry	1 – 2	1 – 2		Cooling elements
Ethane	1	1		Pressure bottles
Atmosphere				
industrial	2 – 3	3 – 2	3	Construction, cars
marine	1 – 2	1	3 – 5	Shipbuilding
Benzene	1	1	1	Tanks, apparatus
Gasoline	1	1	1 – 3	
leaded, wet	3 – 6			
Distilled water	1 – 2	1 – 2		
Freon	1	1		Cooling machines
Ice	1	1	2	Cooling plant
Seawater	2 – 3	1 – 2	3 – 5	

* 1 = good resistance; 2 = resistant; 3 = low resistance; 4 = still usable; 5 = barely resistant; 6 = not resistant

Super-pure aluminum and AlSi 12 are outstandingly resistant to highly concentrated nitric acid (Fig. 24), even at high temperatures. They also withstand concentrated acetic acid well [86].

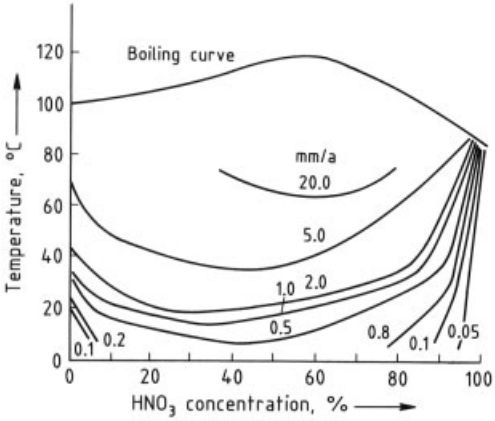

Figure 24. Corrosion behavior of Al 99.5 in nitric acid

The main alloying elements that are used to increase the strength of aluminum are Mg, Si, Zn, Cu, and Mn. Wrought alloys are standardized in DIN 1725, Part 1, and cast alloys in DIN 1725, Part 2. Both groups include self-hardening (alloy hardening) and hardenable (precipitation hardening) materials.

Aluminum – manganese alloys are of the self-hardening type. Alloys containing 0.8 – 1.5 wt % Mn are easy to process and exhibit corrosion resistance similar to that of pure aluminum.

The main representatives of the self-hardening materials are Al – Mg alloys, whose strength is further improved by cold forming. Their corrosion resistance is approximately equal to that of pure aluminum (Al99.5), which they even surpass where seawater is concerned (cf. Table 13). AlMg3, AlMg5, AlMg2Mn0.8, and AlMg4.5Mn therefore belong to the group of aluminum materials known as seawater resistant.

However, with increasing magnesium content (especially for Mg >5 wt %), the susceptibility of Al – Mg alloys to intergranular corrosion increases through the formation of Al_8Mg_5 grain-boundary precipitates (Fig. 25). There is also a risk of susceptibility to stress corrosion cracking [87].

The stability can be raised by reducing the magnesium content and adding manganese. A typical representative of the materials obtained in this way is AlMg4.5Mn, which has good strength-related properties and corrosion resistance and is used both for purposes involving seawater contact and in chemical engineering. Additions of Mn and Cr improve the resistance to chloride-induced pitting corrosion [88].

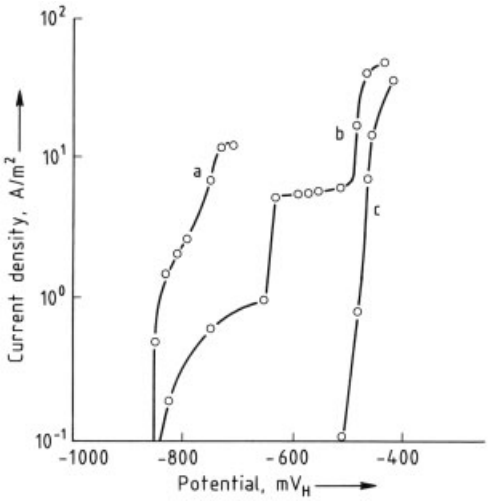

Figure 25. Breakthrough potentials for intergranular corrosion
a) Al_8Mg_5; b) AlMg7 phase with grain boundary precipitates; c) AlMg7 phase, homogenized

Age-hardenable Al – Cu – Mg alloys (2.8 – 4.8 wt % Cu, 0.4 – 1.8 wt % Mg) have high strength (e.g., AlCu4Mg2: σ_u= 440 MPa) but low corrosion resistance. Apart from having little resistance to atmospheric corrosion, they are particularly susceptible to stress corrosion cracking [89].

Age-hardenable Al – Mg – Si alloys with good heat forming behavior (0.4 – 3.5 wt % Mg, 0.3 – 1.5 wt % Si) are more resistant to corrosion than Al – Cu – Mg alloys. Although relatively high silicon contents raise the strength, they reduce the resistance to intergranular corrosion by causing the formation of heterogeneous grain boundary precipitates. Therefore, where corrosion is expected, it is preferable to choose materials with low silicon contents (e.g., AlMgSi0.5). These materials, however, are not fully resistant to pitting and intergranular corrosion in media with high chloride contents.

After artificial ageing, the alloys based on Al – Zn – Mg, and especially Al – Zn – Mg – Cu,

have higher strength (up to 530 MPa) than all other aluminum materials, but they are susceptible to pitting and intergranular corrosion in chloride-containing media.

Important cast aluminum alloys for mechanical and chemical plant are those based on Al – Si (5 to 20 % Si), Al – Mg (3 to 10 % Mg), or Al – Mg – Si or Al – Si – Mg; apart from having high strength and adequate ductility, they are distinguished by good corrosion resistance, including resistance to seawater.

Various treatments and methods are used to protect aluminum from corrosion. High-strength aluminum alloys, for example, are frequently plated with pure aluminum (Al99.5). Nonmetallic inorganic coatings such as chromate and phosphate increase the corrosion resistance and also serve as primers for paints [83].

The thickness of the natural oxide film on aluminum can be increased by boiling the metal in deionized water or by steaming it, as a result of which an AlOOH (boehmite) layer with a thickness of 1 – 2 µm is formed above the Al_2O_3 barrier layer [84]. Another possibility is the use of impressed current or sacrificial anodes (zinc) to reduce the potential below the values at which the risk of pitting or intergranular corrosion is critical [84].

5.5. Copper and Copper Alloys [90], [91]

(→ Copper and → Copper Alloys)

The good corrosion behavior of copper and copper-based alloys has led to their extensive use under moist atmospheric conditions and in the handling of drinking water, industrial water, and water at high temperatures. Their many outdoor applications as fresh water pipes, fittings, condensers, heat exchangers in seawater desalination plants, in chemical apparatus, and for many other purposes are explained not only by good resistance to corrosion but also by their good workability, strength-related properties, and high thermal and electrical conductivity.

In accordance with its position in the electrochemical series (standard potential $Cu/Cu^+ = 0.34$ V) copper (DIN 1787) has good resistance to corrosion, as do high-copper alloys. The good resistance to approximately neutral to alkaline aqueous media (not including water containing NH_3) results from the formation of oxide films which consist of Cu_2O or CuO (depending on the nature of the medium and the corrosion potential) and which afford good protection. In outdoor applications, including those in marine climates, copper is substantially resistant; this explains its use in the construction industry. Its main field of application, however, is plant for drinking, cooling, and industrial water [82]. Here, too, it has good resistance, though under unfavorable conditions (as when deposits are formed or the nature of the water is unsuitable) pitting corrosion cannot be entirely ruled out [92].

In inorganic and organic acids the corrosion rate depends largely on the presence of oxidizing agents. In nonoxidizing acids in the absence of oxygen and at room temperature this rate remains low; in oxidizing acids such as sulfuric acid, it increases with the oxygen content.

In aqueous ammonia solutions under conditions in which surface films are not formed (high alkalinity), copper is attacked severely. In approximately neutral solutions capable of forming surface films, especially aqueous media containing NO_2^-, some susceptibility to stress corrosion cracking, even on the part of pure copper, cannot be entirely excluded [93].

Alloying, particularly with Zn, Al, Ni, or Sn, raises the strength of copper. In the case of Cu – Zn alloys (DIN 17 660) single phase α-phase alloys (Zn ≤ 37 wt %) and $(\alpha+\beta)$-alloys (37 – 46 wt % Zn) are distinguished, the former being more resistant to corrosion. The $(\alpha+\beta)$-alloys tend to suffer preferential attack on the β-phase, which has the higher Zn content. The alloys of both types, but above all the $(\alpha+\beta)$-alloys, suffer dezincification corrosion in chloride-containing waters (Fig. 26, see next page). Dezincification of the α-phase alloys is counteracted by adding As, Sn, Sb, and P as additional alloying elements [94]. Among the ternary alloys, CuZn28Sn and CuZn20Al (2 wt % Al) have been used successfully for many years as materials for cooler, condenser, and heat exchanger tubes in cooling waters with relatively low contamination levels. CuZn20Al is remarkable also for its good resistance to seawater and also to erosion – corrosion.

In cooling water containing NH_3 or H_2S there is a risk of stress corrosion cracking. The critical stress for ammonia-inducedstress corrosion

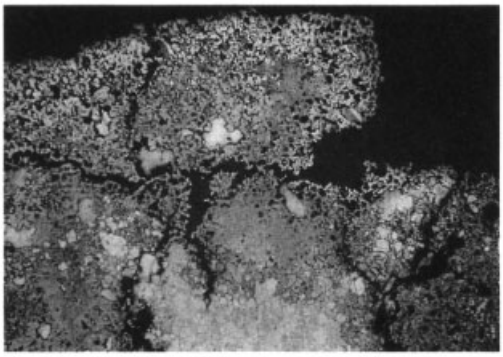

Figure 26. Dezincification of α-brass in tap water

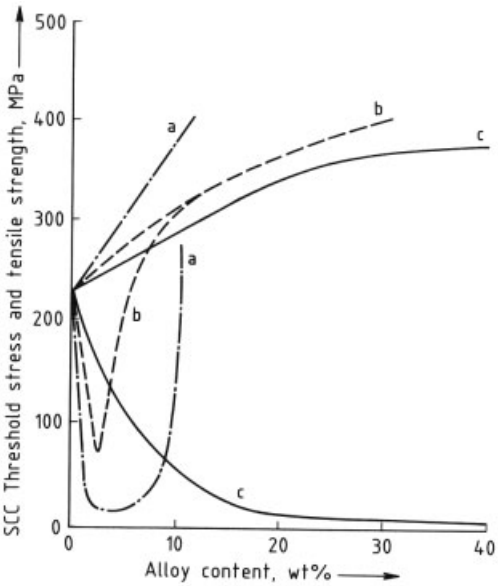

Figure 27. Influence of the content of alloying elements on ammonia-induced SCC of copper alloys
a) CuAl; b) CuNi; c) CuZn
Upper curves: tensile strength; Lower curves: SCC treshold stress

cracking in Cu – Zn materials decreases with increasing Zn content, and is only about 10 MPa for Cu70Zn30 (Fig. 27) [95]. Residual stresses in components that have not been sufficiently annealed may therefore suffice to initiate crack formation. To prevent stress corrosion cracking, stress relief annealing, followed by testing for residual stresses according to DIN 50 916, is necessary.

Of the Cu – Al alloys standardized in DIN 17 665, the homogeneous α-alloys (Al $\leq$ 7.8 wt %) are among those that withstand corrosion, examples being CuAl5, CuAl5As and CuAl8. In their resistance to seawater they even surpass pure copper [82]. In addition to the increased Al content, their improved resistance results from the formation of adherent, highly protective oxide films that contain Cu_2O and Al_2O_3 [96].

The two-phase ($\alpha + \beta$)-alloys (Al >7.8 wt %) are less resistant. The heterogeneous multiphase alloys (Al >10 wt %), which, apart from the α-phase, also contain the γ_2-phase and (depending on the heat treatment) martensitic phases, suffer not only corrosion at relatively high rates, but also preferential dealuminization. Additions of Ni, Fe, and Mn improve the corrosion behavior of the heterogeneous alloys (e.g., CuAl11Ni, CuAl10Fe, CuAl9Mn). Aluminum bronzes are used in process plant manufacture, where their applications include pump parts (gears) and fittings, including those of seawater pumps, and condenser tubing.

Although the homogeneous α-alloys are those most resistant to uniform corrosion and dealuminization, they show heightened susceptibility to ammonia-induced stress corrosion cracking; this applies particularly to Cu – Al materials with about 4 wt % Al (Fig. 27). The heterogeneous materials with Al >8 wt %, on the other hand, have good resistance to stress corrosion cracking, but are sensitive to selective corrosion [95].

Technically interesting Cu – Sn materials (tin bronzes) standardized in DIN 17 662 contain up to 9 wt % and up to 14 wt % Sn, respectively, depending on whether they are wrought alloys or cast alloys [82]. Their strength depends on the tin content and degree of cold forming. With favorable working properties and resistance to alternating stresses, together with resistance to corrosion in neutral salt solutions and alkaline solutions (except those containing NH_3), they are used for screws, springs (CuSn2), pipes (CuSn6) and chemical plant components (CuSn4, CuSn6).

Cast materials consisting of binary Cu – Sn bronzes and multicomponent alloys with additions of Pb (Sn – Pb bronzes) or Zn (gun metal) have proved themselves in marine technology, machinery, and process plant. Owing to their lubrication properties, which are outstanding in some cases, they are used for bearings, screws, worm gears, toothed wheels, pumps, and turbine blades, etc.

Copper – nickel alloys (DIN 17 664 and DIN 17 658) form a continous mixed-crystal series, which improves their corrosion behavior. The Cu – Ni materials of technical interest (CuNi10, CuNi20, CuNi30), which generally contain iron and manganese additions, are among the copper-based materials most resistant to corrosion. At levels of up to about 1.5 wt %, iron further improves the corrosion behavior. The good corrosion resistance of the frequently used materials CuNi10Fe and CuNi30Fe results from the formation of oxidic surface films consisting of Cu_2O on the metal side and complex corrosion products with high percentages of Fe and Ni on the solution side [97].

The high corrosion resistance of Cu – Ni – Fe materials, especially in seawater, but also in ammonia-containing media (brackish water), explains their growing importance in marine technology (seawater desalination plants) and as tube materials for condensers, coolers, and heat exchangers [82]. These alloys also withstand erosion – corrosion and cavitation well. Chloride-induced pitting corrosion occurs occasionally under unfavorable conditions. In comparison with other copper-based materials, Cu – Ni – Fe alloys with Ni contents of ≥ 10 % exhibit substantially better resistance to ammonia-induced intergranular stress corrosion cracking (Fig. 27). They are immune to stress corrosion cracking in chloride-containing media [97].

As the Cu – Ni – Fe alloys of technical interest tend to form Fe – Ni-rich grain boundary precipitates at 350 – 650 °C [98], it is assumed that the homogenized, precipitation-free materials are those most resistant to intergranular stress corrosion cracking. Hence the homogenized state has always been preferred for component fabrication. Investigations into the influence of the state of precipitation on the stress corrosion cracking behavior of the alloy CuNi10Fe1.5 in ammonia-containing solutions have shown, however, that specific heat treatments are capable of producing material states considerably superior to the homogeneous states [97] (Fig. 28). It has also been demonstrated that the stress corrosion cracking behavior is determined exclusively by the hardness of the material [97].

The variety of copper-based materials is such that reliable usedepends on very careful choice: the material must be exactly suited to the conditions of use.

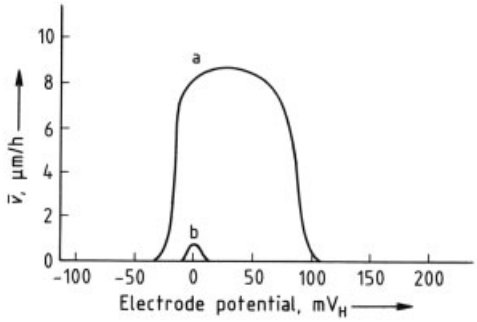

Figure 28. Mean crack velocity $\bar{v}$ of CuNi10Fe1.5 in 1 M $[(NH_4)_2SO_4 + NH_3]$ (pH 9, 50 °C) as a function of the electrode potential
a) Homogenized; b) Aged (50 h, 500 °C)

5.6. Lead and Lead Alloys [82], [99](→ Lead and → Lead Alloys)

Lead owes its good corrosion resistance to its ability to form dense, firmly adherent surface films consisting of lead sulfates, carbonates, or oxides, depending on the corrosive medium [99].

In outdoor applications and approximately neutral waters the protection from corrosion arises from the formation of basic lead carbonates of low solubility ($Pb(OH)_2 \cdot 2\,PbCO_3$), which may also contain lead sulfates. In acids and alkalis, lead normally suffers severe surface corrosion (Fig. 29). However, it is distinguished by outstanding resistance to sulfuric acid and good resistance to phosphoric acid and chromic acid; this results from the low solubility of the sulfate, phosphate, or chromate surface films [99]. Lead is therefore a useful material for fittings, pipes, pump parts, and other mechanical components and also for leadings (reaction vessels, electrolysis tanks) [82], [100].

For greater strength and finer grain, lead is alloyed with antimony (0.5 – 13 wt % Sb). Lead – antimony alloys can be age hardened (precipitation hardening), which gives them further strength (hard lead). These alloys are used particularly where mechanical stresses must be withstood in addition to corrosion, as in the case of accumulator plates, pumps, valves, and impellers [82]. The corrosion resistance can be further improved by small additions of As and Se.

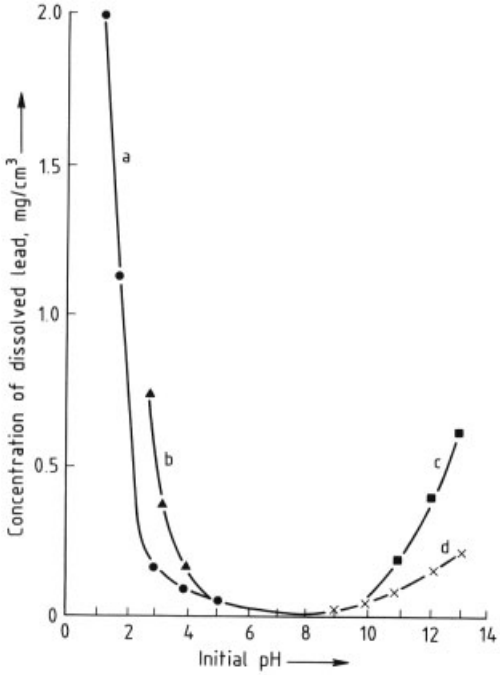

Figure 29. Dependence of the corrosion behavior of lead in aqueous solutions on pH
a) HNO_3; b) Acetic acid; c) $Ba(OH)_2$; d) NaOH

Compared to pure lead, alloys based on Pb – Cu or Pb – Cu – Sn, and also alloys of the Pb – Cu – Sn – Pd type with extremely low contents of Cu (0.01 – 0.1 wt %), Sn (0.05 – 1.12 wt %), and Pd (0.10 wt %), have improved resistance to corrosion by hot and boiling sulfuric acid [101]; see Table 14.

Table 14. Linear corrosion rate of multicomponent lead alloys in boiling sulfuric acid

Alloy	Corrosion rate, mm/a		
	50 % H_2SO_4	70 % H_2SO_4	80 % H_2SO_4
Pb 99.9 Cu	0.48	8.23	[a]
Pb Cu Pd (0.06, 0.1)	0.12	0.21	0.22
Pb Cu Au (0.06, 0.1)	1.86	0.10	0.30
Pb Sb Pd (1.1, 0.1)	0.17	0.19	[a]
Pb Cu Sn Pd (0.05, 0.12, 0.10)	0.01	0.10	0.26
Pb Cu Sn Pd (0.10, 0.13, 0.2)	0.01	0.05	0.19
Pb Cu Sn Au (0.04, 0.05, 0.10)	0.15	0.23	1.95
Pb Ni Sn Pd (0.10, 0.10, 0.10)	0.09	0.28	2.80
Pb Te Sn Pd (0.10, 0.10, 0.10)	0.09	0.29	3.50

[a] Dissolves.

As the solubilities of Cu and Pd in lead are extremely low, these alloying elements are present in the precipitated state, and fine dispersion of the particles is beneficial. The precipitates are capable of reducing the hydrogen overvoltage and accelerating oxygen reduction, thus shifting the potential of the metal into the passive range and consequently improving its corrosion resistance (Fig. 30) [101].

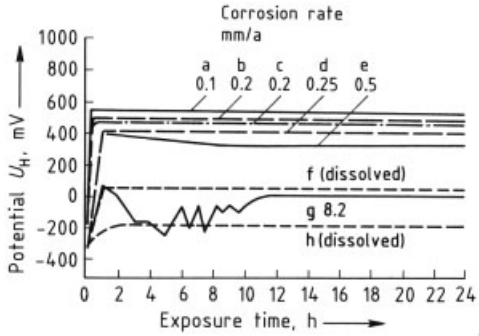

Figure 30. Potential – time curves of lead alloys in boiling 70 % H_2SO_4
a) Pb Cu Sn Pd (0.05 0.12 0.1); b) Pb 99.9 Cu – 0.1 Pd; c) Pb 99.985 – 0.1 Pd; d) Pb Cu Sn Au (0.04 0.05 0.1); e) Pb Cu Sn (0.12 0.12); f) Pb 99.985 – 0.1 Au; g) Pb 99.9 Cu; h) Pb 99.985

Accordingly, alloys consisting of highpurity lead and copper, or where particularly good corrosion resistance is needed, of lead and palladium, are used for chemical apparatus. The Pb – Cu – Sn – Pd alloys also have relatively high creep strength [101].

5.7. Zinc and Zinc Alloys (→ Zinc Alloys)

In DIN 1706 a distinction is made between highpurity zinc (Zn 99.995, Zn 99.99, Zn 99.95) and technical zinc (Zn 99.5, Zn 98.5 and Zn 97.5).

Through outdoor exposure and even mild chemical exposure (contact with drinking water, for example) zinc forms protective surface films that consist mainly of basic zinc carbonates and which adhere to the metallic zinc fairly firmly. Zinc is severely attacked by hot water and steam. A fairly high corrosion rate can also be expected in industrial air containing SO_2 [82]. Zinc is not resistant to acids and strong alkalis.

The main use of zinc is the protection of steel from corrosion, where it is used mainly in the form of metallic coatings, but also as sacrificial anodes.

The most important zinc alloys are the die casting alloys GD-ZnAl 4 and GD-ZnAl 4Cu 1;

in plant manufacture they are of only minor importance.

5.8. Tin and Tin Alloys (→ Tin Alloys)

The purity requirements for tin as specified in DIN 1704 range from 98 to 99.90 wt % Sn.

Due to the formation of oxidic surface films of low solubility, tin has good corrosion behavior in outdoor applications and in water [82].Tin is attacked severely by halogens, halogen compounds, and alkalis. It is substantially resistant to numerous foods and drinks. It also has good resistance to corrosion by organic acids (acetic, citric, maleic, tartaric, and lactic acid), especially in the absence of oxygen. Tin is used mainly in the food industry as a protective metallic coating on appliances and containers.

Tin – lead alloys with 30 – 60 wt % Sn (e.g., SnPb40, PbSn50, PbSn40), possibly with additions of Sb and Cu, are used as soft solders (DIN 1707). Tin alloys with Sn contents of 80 – 91 wt % are important bearing materials (sliding bearings); examples are Sn80Sb12 (5 – 7 wt % Cu, 11 – 13 wt % Sb, 1 – 3 wt % Pb) and Sn80Sb18 (16 – 20 wt % Sb, 1 – 3 wt % Pb).

5.9. Titanium, Zirconium, Niobium, and Tantalum [102], [103]

The refractory metals of groups 4 and 5 of the periodic table have now gained considerable importance in the fabrication of chemical plant. Having excellent passivity, which results from the formation of oxidic surface films, such materials as titanium, tantalum, niobium, and zirconium are outstandingly resistant to many media.

In VdTÜV-Werkstoffblatt 230/1 unalloyed titanium is allocated to four groups having different degrees of purity and strength (Table 15). As the forming behavior of types belonging to group IV is limited, only those belonging to groups I to III are normally used in chemical engineering. The corrosion resistance of these types can be improved by alloying with 0.15 – 0.25 wt % palladium [104].

In the United States, unalloyed titanium (C.P. Ti = commercially pure titanium) is classified in four groups (grades 1 to 4; see Table 16) but the percentages of permissible additives are greater than those laid down in VdTÜV-Werkstoffblatt 230/1. Two palladium-alloyed titanium types are termed grade 7 and grade 11.

The high price of Pd explains the introduction of the molybdenum- and nickel-containing

Table 15. Grades of unalloyed titanium

Material	Material number	Max. impurity content, wt %					Yield strength
		Fe	O	N	C	H	(1 %), MPa
Ti I	3.7025	0.15	0.12	0.05	0.06	0.013	200
Ti II	3.7035	0.20	0.18	0.05	0.06	0.013	270
Ti III	3.7055	0.25	0.25	0.05	0.06	0.013	350
Ti IV	3.7065	0.30	0.35	0.05	0.06	0.013	410

Table 16. U.S. Classification of titanium

Type	ASTM B-265-79	Composition, wt % (max.)						
	grade	Fe	O	N	C	H	Pd	Others
C.P.	1	0.20	0.18	0.03	0.10	0.015		
C.P.	2	0.30	0.25	0.03	0.10	0.015		
C.P.	3	0.30	0.35	0.05	0.10	0.015		
C.P.	4	0.50	0.40	0.05	0.10	0.015		
C.P.-Pd	7	0.30	0.25	0.03	0.10	0.015	0.12 – 0.25	
C.P.-Pd	11	0.20	0.18	0.03	0.10	0.015	0.12 – 0.25	
C.P.-Mo + Ni	12	0.30	0.25	0.03	0.08	0.015		0.2 – 0.4 Mo 0.6 – 0.9 Ni
Ticorex-A∗		0.03	0.05	0.008	0.004	0.0015		0.05 Ru 0.5 Ni

∗ Nippon Mining Co.

alloy grade 12, which, like the ruthenium- and nickel-containing alloy Ticorex A, is superior to unalloyed titanium in crevice corrosion resistance.

Finally, mention should be made of a titanium alloy containing 5 % tantalum which has been developed to the production stage in Japan. It has improved resistance to nitric acid.

A Werkstoffblatt for the alloy tantalum – 2.5 tungsten, in which growing interest is being shown on account of its strength, is in preparation.

The mechanical properties of titanium, zirconium and tantalum, and also a selection of their physical data, are compiled in Tables 17 and 18.

The fundamental corrosion-chemical properties of these materials for chemical plant are

1) The stability of titanium under oxidizing conditions
2) The stability of zirconium under reducing and alkaline conditions
3) The stability of tantalum under oxidizing and reducing conditions.

In the case of titanium, which is the most favorable material in this group with respect to density, particular attention must be drawn to its resistance to chlorides and oxidizing agents. For example, titanium has been used successfully for heat exchangers exposed to seawater, brackish water, nitric acid, acetic acid, chromic acid, moist chlorine, chlorine dioxide, bleaching solutions, sodium chlorate, and sodium chlorite.

The use of palladium as an alloying element extends the range of applications of titanium in acids with reducing effects, such as sulfuric acid and hydrochloric acid, and improves the resistance to crevice corrosion [104].

Corrosion resistance data for titanium, zirconium, and tantalum are compiled in Table 19 (see next page). A common feature of the special metals is a pronounced sensitivity to hydrofluoric acid and fluorides, with a risk of hydrogen embrittlement in the case of extreme exposures. Titanium cannot be used in hydrochloric acid or sulfuric acid. However, its resistance to these media can be improved considerably by adding oxidizing agents. Titanium and zirconium cannot be used in fuming nitric acid, because of the possibility of pyrophoric reactions and stress corrosion cracking.

Of the refractory metals, tantalum has by far the best corrosion behavior in hot, concentrated mineral acids, apart from hydrofluoric acid. In hot, concentrated sulfuric acid its corrosion resistance is comparable with that of glass or cast ferrosilicon and is surpassed only by that of noble metals such as gold and platinum. It is also highly resistant to hydrochloric acid and phosphoric acid and has no pyrophoric tendency in

Table 17. Mechanical properties of titanium, zirconium, and tantalum

Property	Titanium		Zirconium	Tantalum
	3.7035 group II	3.7055 group III	grade 702	
Tensile strength, MPa	400 – 550	470 – 600	295 – 442	274
0.1 % Yield strength, MPa	280	360	197 – 295	196
Creep strength				
150 °C (10^5 h), MPa	150	170	176	255
250 °C (10^5 h), MPa	110	130	117	225
Elongation, %	22	18	25 – 35	15
Hardness HB 30, MPa	1400	1600		
Hardness HV 30, MPa			1180 – 1570	825 – 1470

Table 18. Some physical properties of tianium, zirconium, and tantalum

Property	Titanium 3.7025	Zirconium	Tantalum (Ta-ES)
Density, g/cm^3	4.5	6.53	16.6
mp, K	1 975	2 125	3 271
Thermal conductivity, J m^{-1} s^{-1} K^{-1}	17	21	55.25
Modulus of elasticity (300 K), GPa	108	94.176	172.5

Table 19. Corrosion resistance of titanium, zirconium, and tantalum

Medium	Concentration, wt %	T, °C	Corrosion rate, mm/a		
			Titanium	Zirconium	Tantalum
Hydrochloric	5	20	<0.05	<0.05	<0.001
acid	15	35	2.4	<0.08	<0.001
(aerated)	37	35	15.0	<0.08	<0.001
Sulfuric acid	10	35	1.2	<0.05	<0.001
(aerated)	40	35	8.5	<0.05	<0.001
Nitric acid (fuming)			inflamm. in air	inflamm. in air	resistant
	0.001	all	not resistant	not resistant	not resistant
Sodium	10	100	<0.05	<0.05	1.0
hydroxide	40	80	<0.1	<0.05	not resistant

fuming nitric acid. Tantalum is, however, attacked by oleum, even at room temperature, and also by hot alkaline solutions. Although it has high ductility, its strength can be increased considerably by alloying with 2 – 10 wt % W, with little effect on the corrosion resistance. Tantalum – tungsten alloys exhibit good resistance to sulfuric acid, which may be seen in Figure 31 [105]. The alloy most favorable in this respect is Ta2.5W.

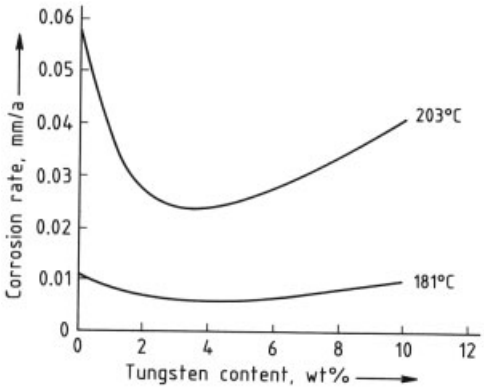

Figure 31. Corrosion rates of tantalum – tungsten alloys in sulfuric acid (96 %)

At normal pressure tantalum and its alloy with 2.5 wt % tungsten are completely resistant to hydrochloric acid at all concentrations and temperatures [106]. If, however, hydrochloric acid is handled under pressure in tantalum apparatus, damage through hydrogen absorption, and finally through hydrogen embrittlement, is possible. An autoclave insert failed in consequence of hydrogen embrittlement after being in use for only 8 h. The range of conditions under which tantalum is at risk through hydrogen embrittlement is shown in Figure 32.

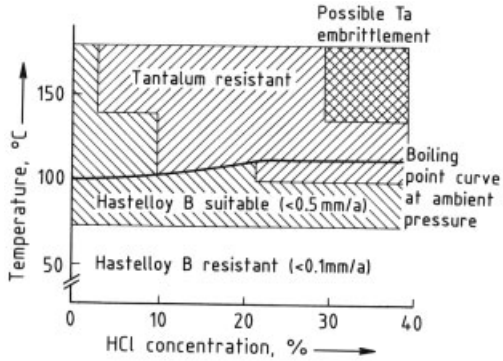

Figure 32. Comparison of the corrosion resistance of tantalum and Hastelloy alloy B in hydrochloric acid

Alloying tantalum with niobium gives favorably priced materials with similar resistance to corrosion that have a good chance of becoming established in the chemical industry and of playing a part similar in importance to that of tantalum itself. This applies particularly to use in nitric acid. Figure 33 shows the corrosion rates of tantalum – niobium alloys in sulfuric acid (70 %) at 165 °C. As, however, the rate of corrosion increases with the niobium content, only alloys containing up to 40 wt % Nb are of technical interest. The mechanical properties of tantalum, tantalum alloys, and niobium are shown in Table 20.

Table 20. Mechanical properties of tantalum, tantalum alloys, and niobium

Property	Ta	Ta 2.5 W	TaNb 60/40	Nb type 2
Tensile strength, MPa	225	290	276	170
0.2 % Yield strength, MPa	140	205	193	105
Elongation, %	15	25	25	25

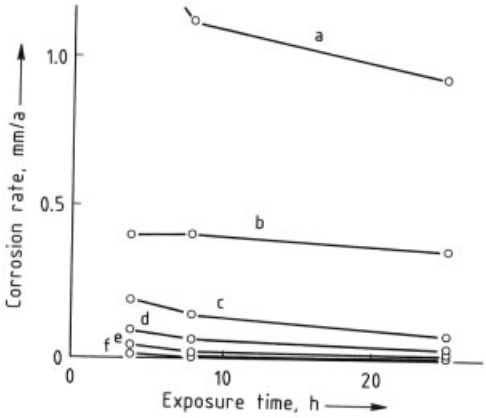

Figure 33. Corrosion rates of tantalum – niobium in boiling 70 vol% sulfuric acid at 165 °C
a) Ta–75 Nb; b) Ta–60 Nb; c) Ta–50 Nb; d) Ta–40 Nb; e) Ta–25 Nb; f) Ta

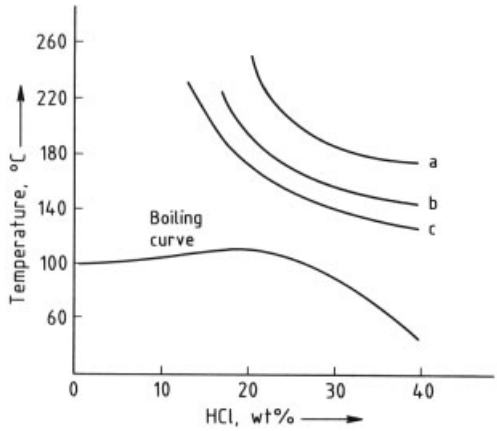

Figure 34. Corrosion of zirconium in hydrochloric acid
a) 5 mm/a; b) 0.5 mm/a; c) 0.13 mm/a

Under oxidizing conditions zirconium is often less resistant than titanium. Its resistance to nonoxidizing media nevertheless greatly exceeds that of titanium. Zirconium is thus practically passive in hot, fluoride-free sulfuric acid. Zirconium, like titanium, is resistant to many organic acids, but is severely attacked by hydrofluoric acid. Unlike the other refractory metals, zirconium is resistant to hot alkaline solutions and is therefore the most suitable material for these media in many cases.

Zirconium also has good resistance to corrosion by hydrochloric acid (Fig. 34), though it is vital to ensure that oxidizing agents, such as Fe(III) or Cu(II) compounds, are not present. At 30 °C and an acid concentration of 32 %, as little as 5 ppm of such a compound increases the corrosion rate ten-fold (from 7.6×10^{-4} mm/a to 7.6×10^{-3} mm/a) [107], [108]. For critical corrosion conditions, heat treatment for 3 h at 750–790 °C, especially after welding, is advisable in order to maximize the resistance [109].

Refractory metals have a wide range of applications in the chemical industry. They are used for reactors, columns, agitators, pipelines, fittings, bellows, and pumps, for example. Sensing heads, such as thermo-feelers, magnetic floats and membranes, that must also function reliably under adverse conditions are protected by cladding with special metals. Heat exchangers made from special metals and consisting of coils, rods, candle-shaped elements, or tube bundles are used to cool or heat corrosive media.

The components may consist entirely of the special metal or of suitable composites. In many cases nonadherent lining or sheathing of the substrate material is sufficient; cladding is appropriate where heavy loads are exerted, good heat transmission is needed, or negative pressure exists. As the special metals tend to form intermetallic phases with steel at high temperatures (1000 °C), they are normally applied by explosion cladding, possibly with intermediate layers (e.g., copper) [110], [111].

Although special metals have good welding behavior, their reactions with atmospheric gases must be taken into account. Hence welding is only possible under an inert gas or in high vacuum. It is essential to avoid alloying with iron-based materials in welding. Since the melting point of tantalum, for example, is roughly twice as high as that of steel, specially designed butt-welds are necessary in the processing of explosion-cladded metal sheet (cf. Fig. 35, next page) [112].

5.10. Organic Materials [113–115]

To an increasing extent the growing requirements on the availability and reliability of chemically exposed plant components are necessitating the use of particularly corrosion-resistant materials. In this respect the chemical nature of organic materials gives them certain advantages over metallic and inorganic nonmetallic materials, provided the operating temperatures do not reach critical levels. In many cases they

are well suited for plant that handles the media most commonly encountered at chemical works, namely hydrochloric acid, caustic soda solution, hypochlorite solution, sulfuric acid, waters, and salt solutions of all kinds. The main types of organic materials are thermoplastics, thermosetting plastics, elastomers, glass-fiber-reinforced plastics, and foamed plastics. Composite and surface coating materials are being used increasingly too (see → Corrosion).

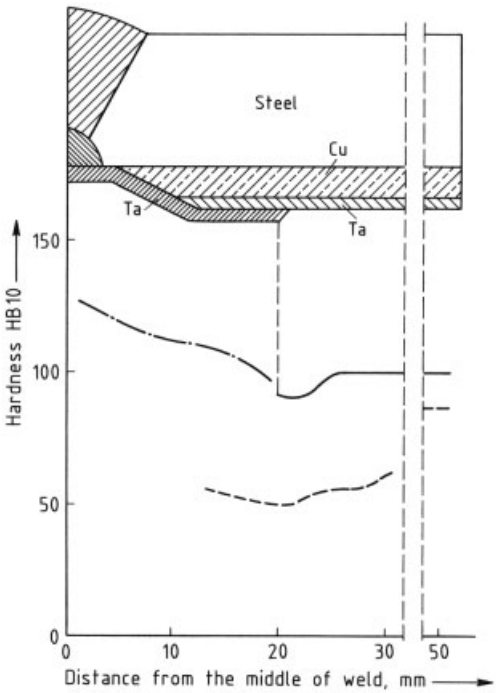

Figure 35. Design of a butt-weld for explosion-cladded sheet (mild steel/tantalum with copper intermediate layer) –·– Hardness curve in tantalum cover strip; —- Hardness curve in tantalum cladding; – – – Hardness curve in copper intermediate layer

One of the most important processing methods for plastics is welding. A survey of the welding methods is given in [116]. A comprehensive account of the design of plastic components, with detailed consideration of strength calculations and dimensioning, is contained in VDI-Taschenbuch T 21 [117].

5.10.1. Selection Criteria

The stability required of plastics varies according to whether a plastic component must simply withstand corrosion, or fulfil structural functions also. In general no risk arises if the physical properties of a coating or facing consisting of a synthetic material are impaired through the action of a medium; but the same process in a statically or dynamically stressed part may lead to its premature failure. Similarly, swelling and stress corrosion cracking may be nonhazardous, and the quality of joints relatively unimportant, in nonstructural plastics, but hazardous, or critical, in those with structural functions. The permeability of polymers is less serious in components than in coatings, especially those on metallic substrates. Thus the envisaged conditions of use must be carefully considered before materials are chosen. As important parameters, the solutions to which the components will be exposed, including all their consituents (such as solvents, amines, phenols, compounds with high vapor pressures, oxidizing agents), the maximum and long-term service temperatures, the pressures and vacuum conditions, and the static and dynamic loads, must receive special attention. Only when the various factors have been established can a list of foreseeably suitable materials be drawn up for more detailed consideration.

5.10.2. Properties and Application Criteria

Unlike the surface removal corrosion of metallic materials, the action of liquid media on organic materials leads mainly to swelling, caused by absorption of the medium. In many cases the degree of swelling of an organic material in individual media and the dependence of swelling on time and temperature as well as its possible reversibility constitute adequate criteria for the determination of suitability. The stability data published in manufacturers' and other literature are usually based on swelling tests. These data are adequate where coatings or linings, e.g., of rubber or plastics, are concerned, and in a number of other cases.

Where static or dynamically loaded parts are concerned, swelling tests alone must be considered tentative, enabling one merely to draw up

a shortlist of materials for further consideration. Here it is far more important to ascertain how the strength of materials is affected by prolonged exposure to chemicals, the degree to which creep occurs, and whether or not the conditions are favorable to stress corrosion cracking. Losses of long period creep resistance caused by liquid media are preferably expressed as resistance factors. A resistance factor is the ratio of the time to failure in the test medium to the time to failure in water [118]. In some media this ratio depends on the mechanical stresses exerted on the sample; in other cases it is largely independent of them. Resistance factors for glass-reinforced plastics can be derived from the differences in creep behavior in tensile tests under various ambient conditions (as ratios between creep moduli, for example) [119].

The stress corrosion cracking of organic polymers is generally an entirely physical process, with diffusion and swelling processes and internal and/or external stresses playing a major part [120]. The term is used only where cracking caused by exposure to chemicals, like that of metals, occurs exclusively in the presence of tensile stresses. Tests are performed in various media and at various temperatures on panels or pipes subjected to constant deformation or constant stress in the long period tensile creep test [118]. Susceptibility to stress corrosion cracking is increased by exposure to wetting agents, organic acids, solvents, and oxidizing media, by shaping and jointing operations, by internal and external stresses, by prolonged exposure to high temperatures in processing, by temperature alteration, and by notching. In many cases the sensitivity of finished parts to stress corrosion cracking can be reduced by thermal after-treatment.

The thermoplastics most widely used in chemical engineering are PVC [poly(vinyl chloride)], PE (polyethylene), PP (polypropylene) and PB (polybutene).

Poly(vinyl chloride) has been used for decades due to its outstanding resistance to chemicals. It is harder and tougher than other thermoplastics, but at temperatures below 20 °C it has less impact strength than polyethylene (PE) and polybutene (PB). Above its softening temperature of 80 °C, poly(vinyl chloride) can be shaped by plastic deformation. The most favorable deformation range is 110 – 130 °C. Original shaping (e.g., by injection molding, extrusion, or welding) requires higher temperatures, however (170 – 200 °C). The upper service temperature limit is 60 °C, or possibly a lower temperature, depending on the chemical exposure.

Polyethylene is a partly crystalline material with high impact strength at temperatures down to ca. – 100 °C. High-density polyethylene is preferred for chemical plant because it has greater strength and stiffness than the low-density types. It can be shaped thermoplastically at temperatures above 130 °C and welded at ca. 200 °C. The upper service temperature limit is 60 °C, or 80 °C if glass fibers are incorporated. The resistance to nonoxidizing media is excellent. The swelling caused by hydrocarbons leads to losses of strength that can be taken into account with the aid of resistance factors. The susceptibility to stress corrosion cracking increases with the degree of crystallization, but falls with increasing molecular mass. High molecular mass material is less easily processed, however, and its properties may suffer if it is exposed to elevated processing temperatures for long periods.

Polypropylene (PP) has properties similar to those of polyethylene, but greater impact, notched impact, and shock resistance, especially at temperatures below 0 °C. The upper service temperature limit is 80 °C, or 100 °C if glass fibers are present. Polypropylene is almost as resistant to chemicals as hard polyethylene, but more sensitive to oxidizing media and more strongly swelling in organic solvents. In common with hard polyethylene it has a certain permeability to organic media.

The applications of polypropylene include exhaust ducts, internal chimney tubing, separators, rotary filters, filter fabrics and pressure plates, hydrocyclones, and gas scrubbers.

Polybutene (PB) is one of the more recent polymeric materials that have gained acceptance in chemical engineering. Its strength properties lie between those of soft polyethylene (LD-PE) and polypropylene. Its long period creep resistance exceeds even that of polypropylene. Although polybutene is soluble in several organic

solvents at elevated temperatures, its susceptibility to stress corrosion cracking is low. The upper service temperature limit is 90 °C. The main applications of polybutene are high-temperature pressure pipes, filter pressure plates, strip-wound vessels, and pumps.

Polytetrafluorethylene (PTFE) is the thermally and chemically most resistant plastic used in chemical engineering. Its service temperature range extends from − 200 to 250 °C. It is chemically resistant to virtuallly all media except molten alkalis and elemental fluorine. However, fillers added to the plastic to improve its extrudability or mechanical strength may reduce its resistance to chemicals. The outstanding abhesive behavior and exceptionally favorable sliding properties deserve special mention. Applications include pipe linings, bellows, hoses, pump parts, tube bundle heat exchangers, steam release nozzles, conveyor belts, seals, stuffing box packings, and filter fabrics.

Poly(vinylidene fluoride) (PVDF) is a partly crystalline plastic of high thermal stability. It can be used at temperatures up to 100 °C, or, if glass-filled, up to about 120 °C. PVDF is resistant to most organic and inorganic acids. In alkalis, especially caustic soda solution, it tends to suffer stress corrosion cracking. Its main application is in pipelines used to convey acids.

Elastomers and ebonite based on natural or synthetic rubber [121], [82] are used mainly for surface protection (see → Corrosion). Elastomers also have many applications in chemical engineering as seals, membranes, bellows, vibration dampers, hoses, and Moineau pump stators, as well as being used for space-saving folding containers for the transportation of corrosive materials by land and sea.

5.10.3. Thermosetting Plastics

The main materials in this group are phenolic, furane, epoxy, vinyl ester, and unsaturated polyester (UP) resins. They are filled with glass fibers, graphite, or quartz and processed to tamping compositions, hot molding compounds, sheet, and putties. Being brittle, they are rarely used directly for mechanically stressed parts unless reinforcing materials and fillers have been incorporated. Having high cross-link density, thermosetting plastics are insoluble and cannot be melted, and their strength and long period creep resistance are without the marked dependence on temperature exhibited by thermoplastics and elastomers. Individual parts can be joined to one another with catalytically hardening cements.

Solid parts made from glass-reinforced reaction resins are distinguished not only by chemical stability but also by relatively low weight, together with high strength, and by good emergency running behavior. They are being used increasingly for composite structures exploiting the high stability and heat resistance of glass-fiber-reinforced reactive resins with the superior chemical resistance of thermoplastic coatings and linings.

For exposure to aggressive media, parts consisting entirely of glass-reinforced plastics usually receive an at least 2.5 mm thick coating consisting mainly of resin. This protective coating, in which the proportion of glass is low, acts as a barrier to corrosion and protects the load-bearing, glass-reinforced structure from the action of destructive media. The resins used are unsaturated polyester, vinyl ester, epoxy, or furane resins.

Table 21 gives a general idea of the corrosion resistance of reaction resins used in the fabrication of apparatus and pipes. Glass-fiber-reinforced resins are used for corrosion-resistant articles of all kinds, especially pipes for waste water and waste air disposal, stacks, siphons, vats, transportation and storage containers, down-pipes, filter pressure plates, and frames. The upper service temperature limits extend to about 120 °C, depending on the resin in question.

Table 21. Chemical resistance of reaction resins ∗

	Resistance towards			
Type of resin	Acids	Alkalis	Solvents	Oxidizing agents
Polyester	+/0	+/−	0/−	+
Vinyl ester	+/0	+	0/−	+
Epoxy	0/−	+	+/0	+/−
Furane	+	+	+	0/−

∗ + = resistant; 0 = conditionally resistant; − = not resistant

More frequently, however, parts consisting of glass-reinforced plastics are provided with linings. The lining materials most widely used at present are the thermoplastics polypropylene (PP), unplasticized polyvinyl chloride (PVC-U) and polyvinylidene fluoride (PVDF).

In the fabrication of composite structures, thermoplastic liners are normally produced as a first step and than reinforced with a glass-filled reaction resin. Thermoplastic linings consisting of sheet material are placed end to end and joined by welding. The normal welding techniques are hot gas, heated tool, and hot gas extrusion welding.

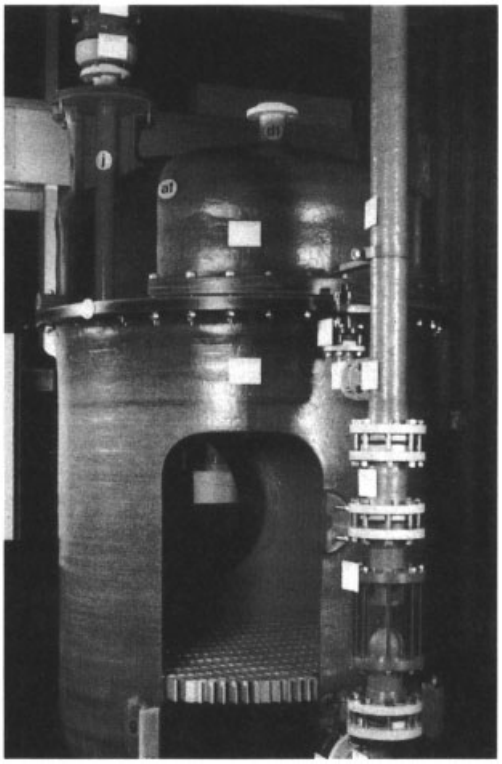

Figure 36. PVC-lined UP-GF apparatus

Figure 36 depicts a PVC-lined UP-GF apparatus, seen here as a demonstration exhibit with parts removed to show structural details. Another application is shown in Figure 37. This is a storage facility for solutions containing HCl and NaOCl, consisting of UP-GF tanks and pipes with PVC liners. The design of parts consisting of glass-reinforced plastics is described in [122]. Properties, processing, and applications are described in [123]. The use of organic materials in the form of corrosion-resistant coatings and linings represents an important application in chemical engineering (see → Corrosion).

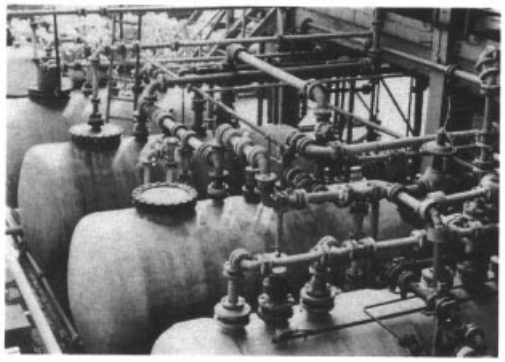

Figure 37. UP-GF tanks and pipes with PVC liners, containing HCl and NaOCl solutions

5.11. Inorganic Nonmetallic Materials

5.11.1. Glass [124], [125]

Pipelines, apparatus, and even entire plants are now being produced increasingly from borosilicate glass in order to solve exceptionally critical corrosion problems. A complete range of glass parts for assembly by the user, ranging from simple pipe sections and moldings to fittings, vessels and column components, and even to complex heat exchangers and their accessories, is now commercially available.

Borosilicate glass has a very wide range of uses and can be exposed permanently to temperatures up to 200 °C. It is permanently stable to virtually all media except hydrofluoric acid and strong alkalis. Above 200 °C it is attacked fairly severely by all acids, especially phosphoric acid [126]. Other advantages include the smoothness of its surfaces, catalytic inertness, and nontoxicity.

Glass can be combined with other materials – plastics and metals, for example – thus opening up a wide range of technical possibilities. Glass apparatuses for heating purposes can be provided with elements consisting of stainless steel, highly corrosion-resistant nickel alloys, titanium, or tantalum. Pumps and valves of all kinds can be sealed with polytetrafluoroethylene components. PTFE is also used in the form

of bellows to compensate for thermal expansion and to damp vibration. Important applications of borosilicate glass include distillation and rectification units, absorbers and gas scrubbers, nitric acid concentrators, acid recovery plant, cleaning units, and piping.

Although quartz glass is very resistant to thermal shock due to its extremely low coefficient of thermal expansion, its high cost limits technical use to special articles (e.g., observation windows).

The use of enamel coatings, made from vitreous enamels and partly crystalline enamels, to protect tanks, apparatus, columns, pipelines, pumps and other items from corrosion is described in → Corrosion).

5.11.2. Graphite [82], [127], [128]

Graphite has been used as a special material in chemical engineering since the mid-1930s. Heat exchangers, simply cemented together from panels and strip, were the first application. Now, all structural elements, such as pipes, solid and hollow cylinders, slabs, panels, and profiles, can be produced from impregnated graphite (hard-fired carbon or electrographite).

As the various graphite production methods give porous materials, impregnation with phenol – formaldehyde or furane condensation resins by the vacuum-pressure process is necessary if impermeability to gases and liquids is required, the effect of this treatment being that all the pores are closed. The quality of the finished material depends on the properties of the chosen resin. This applies particularly to the permeability, corrosion resistance, and upper service temperature limits (which are in the region of 165 °C). For greater stability and higher service temperature limits, alternative impregnating agents, such as polytetrafluorethylene, are used. Carbon filling of the pores in graphite can be carried out by impregnating with cokable materials, which are then decomposed thermally, this treatment being repeated several times. The resulting material has good corrosion resistance and a high service temperature limit, but is less strong and more brittle than the resin-impregnated grades.

Resin-impregnated electrographite is the synthetic carbon material most widely used in chemical engineering. It is highly resistant to corrosion by organic and inorganic acids, alkalis, alcohols and many solvents, and other organic compounds. It is easily worked, has high thermal conductivity, and withstands even severe temperature shocks. As graphite is a brittle material with practically no plastic deformation, the parts must be so designed that tensile and shear stresses are avoided as far as possible.

The main applications of graphite are heat exchangers of all kinds, columns, sprinkling coolers, centrifugal pumps, valves, pipelines, and small parts of many kinds. The use of graphite in plant for production of dry hydrogen chloride and in the treatment of waste water and waste air deserves special mention.

5.11.3. Refractory and Acid-Resistant Bricks [129–132]

This term covers ceramic materials that are still free from deformation at temperatures exceeding 1600 °C and still do not soften at those exceeding 1800 °C. Bricks are normally shaped and fired, but are sometimes used in the form of castings. Apart from refractoriness, such properties as softening under pressure, thermal conductivity, chemical resistance, resistance to temperature alternation, abrasion resistance, shape stability, and slagging tendency are of great importance.

Refractory bricks serve to enclose spaces in which reactions take place and must last as long as possible under the given conditions. In view of the variety of these conditions they must be chosen for then intended application in accordance with their chemical and technological properties. The following classification of refractory materials is based on their main consituents: products of high silica content, alumina-based materials, basic refractory bricks, neutral refractory products, carbon bricks, and silicon carbide bricks.

For corrosion protection in difficult cases, chemical apparatus can be lined with corrosion-resistant bricks, tiles, or moldings. These are attached to the wall or to one another with mortar or putty. Sealing and insulating layers may be used additionally between the wall of the apparatus and the lining.

5.11.4. Engineering Ceramics [133], [134]

Oxide Ceramics. The most important oxide ceramic material is aluminum oxide, Al_2O_3. Beryllium oxide has applications as a moderator in nuclear reactors, and zirconium dioxide as a rocket material. In Table 22 the main mechanical and thermal properties of aluminum oxide are compared with those of hard metal. A conspicuous property of Al_2O_3 is its great hardness. The strength values are practically constant over a wide temperature range. This is because the increase in plasticity (and hence in the ability to dissipate local peak stresses) which occurs as the temperature rises compensates for the thermally induced loss of strength. On the other hand the impact strength (unnotched) of 0.2 J/cm^2 at 20 °C, is low compared to the notched impact strength of 100 J/cm^2 for structural steel. Thus Al_2O_3 is a brittle material.

Table 22. Properties of Al_2O_3 ceramic and a hard metal

Property	Al_2O_3 ceramic	WC Hard metal with 6 % Co
Density, g/cm^3	3.9	14.9
Bending strength, MPa	300	1900
Compression strength, MPa	2500	5000
Breaking elongation in compression test, %	<0.1	ca. 0.5
Modulus of elasticity, GPa	390	620
Vickers hardness	1750	1600
Thermal conductivity, W m^{-1} K^{-1}	30	80
Coefficient of thermal expansion, 10^{-6}/K	8	5

Alumina has very good corrosion resistance, withstanding acids, molten metals, many glasses, and slag. It has advantages over metals particularly where corrosive and thermal stresses are superimposed on mechanical stresses.

One of the commonest applications of alumina ceramics is therefore the sealing of passages leading into vessels and spaces in which corrosive media are present. In the form of slide rings and mechanical seals alumina is superior to all other materials in resistance to corrosion and wear. Its use is therefore leading to the replacement of conventional stuffing boxes on rotating shafts by mechanical seals, especially where high rotation speeds occur. Aluminum oxide can be used in oxidizing and reducing atmospheres and under vacuum at working temperatures of up to 1950 °C.

Silicon Carbide. Silicon carbide has not only high strength (bending strength up to 650 MPa), but also a low coefficient of thermal expansion of only 4.8×10^{-6} K^{-1} and a high thermal conductivity of 42 W K^{-1} m^{-1}. These properties explain the very good resistance of silicon carbide products to changes of temperature. Since silicon carbide also has excellent resistance to corrosion, it is used as a structural material for components subjected to severe thermal and chemical loads.

Silicon carbide is used mainly as a grinding agent, e.g., for grey cast iron, nonferrous metals, glass, and ceramics. With great resistance to temperature alternation and corrosion, SiC thermocouple protection tubes are used in contact with liquid aluminum in aluminum holding furnaces. Silicon carbide products are used for the muffles of indirectly heated smelting furnaces for zinc, aluminum, copper and their alloys, and for tank linings. With good electrical conductivity at high temperatures, coupled with resistance to oxidation, SiC heating elements are used in oxidizing atmospheres at temperatures up to 1500 °C.

Silicon Nitride. The most remarkable properties of silicon nitride include good resistance to chemical attack by acids and molten nonferrous metals, resistance to temperature fluctuation (coefficient of thermal expansion 3×10^{-6} K^{-1}), and heat resistance. The bending strength of hot-pressed silicon nitride is about 700 MPa at 20 °C and still as high as 290 MPa at 1400 °C. At 1000 °C dense silicon nitride still has a 100-h long-term creep resistance of 200 MPa. In the case of the aluminum industry, materials are needed that are not wetted by the molten metal, have good resistance to temperature fluctuation, and high strength at temperatures up to ca. 800 °C. As silicon nitride meets these requirements, it is used for, among other things, thermocouple protection tubes and ascending pipes in low-pressure casting. The hardness of silicon nitride gives it high wear resistance, which, together with the low friction coefficient ($\mu = 0.1 - 0.2$), favors its use as a material for bearings and sliding contacts. Slide rings, roller bearings, and

tube drawing plugs are typical applications. Silicon nitride is also used as a cutting material.

Ceramic materials are often used in the chemical industry in the form of thermally sprayed coatings that afford protection against wear and corrosion. This field of application is described in → Corrosion).

6. References

1. E. Gaube et al.: "Werkstoffe für den Apparatebau in der chemischen Technik," in: *Winnacker-Küchler,* 4th ed., **vol. 1,** pp. 452 – 503
2. H. Titze: *Elemente des Apparatebaues,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York 1967.
3. K. Wellinger, F. Eichhorn, P. Gimmel: *Schweißen,* Kröner, Stuttgart 1964.
4. O. Becken: "Grundlagen und Anwendung," in: *Handbuch des Schutzgasschweißens,* part I, Deutscher Verlag für Schweißtechnik, Krefeld 1969.
5. J. Ruge: "Werkstoffe," in: 2nd ed., **vol. 1,** Springer Verlag, Berlin – Heidelberg – New York 1980.
6. W. Schönherr: "Über die Eignung der Werkstoffe zum Schweissen und thermischen Trennen," *Schweissen Schneiden* **23** (1971) 441 – 442.
7. G. Habenicht: *Kleben, Grundlagen, Technologie, Anwendung,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York – London – Paris – Tokyo – Hongkong – Barcelona 1990.
8. R. J. Schliekelmann, F. Mittrop: *Metallkleben – Konstruktion und Fertigung in der Praxis,* Deutscher Verlag für Schweißtechnik, Krefeld 1971.
9. M. Pfender: "Über das Festigkeitsverhalten metallischer Werkstoffe und Konstruktionsteile," in: *Das Gesicht des Bruches metallischer Werkstoffe,* vols. 1 and 2, Allianz-Versicherungs-AG, München 1956.
10. K. Wellinger, H. Dietmann: *Festigkeitsberechnung,* Kröner, Stuttgart 1969.
11. D. Broek: *Elementary Engineering Fracture Mechanics,* Noordhoff International Publ., Leyden 1974. H. Blumenauer, G. Pusch: *Technische Bruchmechanik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1982 (distributed by Springer, Vienna – New York).
12. Merkblätter der Arbeitsgemeinschaft für Druckbehälter.
13. M. Litzkendorf: *Über den Zusammenhang zwischen Korrosion und Festigkeitsverhalten hochlegierter Chrom- und Chrom-Nickel-Stähle,* Dissertation, TH Darmstadt 1974.
14. K. Wellinger, H. Uetz: "Gleitverschleiß, Spülverschleiß, Strahlverschleiß unter der Mitwirkung von körnigen Stoffen," *VDI Forschungsh.* **449** (1955) .
15. T. Günther *Chem. Ing. Tech.* **42** (1970) 774 – 780.
16. DECHEMA: *DECHEMA-Werkstoff-Tabelle,* Frankfurt, published since 1953. D. Behrens (ed.): *DECHEMA Corrosion Handbook,* VCH Verlagsgesellschaft, Weinheim , vol. 1 (1981) – vol. 6 (1990), to be continued.
17. F. Ritter: *Korrosionstabellen metallischer Werkstoffe,* 4th ed., Springer Verlag, Wien 1958.
18. H. Gräfen, K. Steiger: "Zerstörungsfreie Prüfungen bei Betrieb und Instandhaltung von Chemieanlagen," *DECHEMA Monogr.* **99** (1985) 137 – 157.
19. H. Adrian, F. Brühl: "Die Entwicklung der hochfesten schweißbaren Stähle für den Stahlbau und ihre Anwendung und Verarbeitung," *Stahl Eisen* **86** (1966) 645 – 662, 670 – 672.
20. K. Schaar: "Hochfeste Stähle im Druckbehälterbau," *Tech. Überwach. (Essen)* **7** (1966) 145 – 150.
21. K. H. Piehl: "Herstellung, Eigenschaften, Verarbeitung und Anwendung von hochfesten Feinkornbaustählen," *Rheinstahl-Tech.* **7** (1969) 71 – 83.
22. F. Heisterkamp, L. Meyer: "Mechanische Eigenschaften perlitarmer Baustähle," *Thyssenforschung* **3** (1971) 44 – 65.
23. F. Heisterkamp, D. Lauterborn, H. Hübner: "Technologische Eigenschaften, Verarbeitbarkeit und Anwendungsmöglichkeiten perlitarmer Baustähle," *Thyssenforschung* **3** (1971) 66 – 76.
24. E. Weber, W. Rudat: "Herstellung der Stahlrohre," in: *Rohrleitungen, Theorie und Praxis,* Springer Verlag, Berlin – Heidelberg – New York 1967.
25. W. Dahl, H. Hengstenberg, H. Behrens: "Sprödbruchverhalten unlegierter und niedriglegierter Baustähle im Kerbschlagbiege- und Zugversuch in Abhängigkeit von Korngrößen, Temperatur,

Vorverfestigung und Alterung," *Stahl Eisen* **88** (1968) 578.
26. C. Straßburger: "Stahl – auch weiterhin ein moderner Werkstoff," *Thyssen Tech. Ber.* **13** (1987) no. 1, 89 – 102.
27. E. Lenz, N. Wieling: "Einflußgrößen auf die Riß-initiierung und auf das Rißwachstum in niedriglegierten Feinkornbaustählen in Hochtemperaturwasser," *Vorträge der 20. Sitzung des Arbeitskreises Bruchvorgänge, Frankfurt 1988,* Deutscher Verband für Materialforschung und -prüfung, Berlin 1988, pp. 127 – 167.
28. D. Uwer, H. Dißelmeyer: "Erfahrungen mit der Herstellung, Verarbeitung und Anwendung des hochfesten wasservergüteten Baustahles XAB090," *Schweissen Schneiden* **38** (1986) no. 9, 430 – 436.
29. C. Florin et al.: "Über die Grundlagen der Warmfestigkeit, dargestellt an legierten Stählen," *Arch. Eisenhüttenwes.* **41** (1970) 777 – 787.
30. H. Wisniowski: "Hochwarmfeste 12 %-Chromstähle," *DEW Tech. Ber.* **9** (1969) 117 – 134.
31. H. Wirtz: "Anwendung," in: *Verhalten der Stähle beim Schweißen,* 2nd ed., vol. 2: Deutscher Verlag für Schweißtechnik GmbH, Düsseldorf 1977.
32. K. Bungart: "Hochtemperaturwerkstoffe auf Nickel- und Kobaltbasis," *DEW Tech. Ber.* **9** (1969) 146 – 162.
33. H. Kiessler: "Normung hochwarmfester Stähle und Legierungen," *DEW Tech. Ber.* **9** (1969) 110 – 117.
34. *Ergebnisse deutscher Zeitstandversuche,* Verlag Stahleisen mbH, Düsseldorf 1969.
35. W. Ruttmann, H. Kaes: "Bedeutung der Änderung der Langzeitwerte," *Mitt. Ver. Großkesselbetr.* **48** (1968) 1 – 9.
36. K. Bungardt, H. Krainer, H. Schrader: "Nichtrostende, hitzebeständige sowie hochwarmfeste Stähle und Legierungen," *Stahl Eisen* **84** (1964) 1796 – 1811.
37. D. Schneider: "Schweißverbindungen an kaltzähen Stählen im Druckbehälterbau," *Tech. Überwach. (Essen)* **7** (1966) 317 – 321.
38. K.-A. Ebert: "Betrachtungen zum Lichtbogenhandschweißen der kaltzähen Stähle," *Schweissen Schneiden* **18** (1966) 125 – 137.
39. H. Jesper, K. Achtelik: *Kaltzähe Stähle,* **publication no. 33,** Nickel-Informationsbüros GmbH, Düsseldorf 1964.
40. H. Spähn: "Druckwasserstoffangriff auf unlegierter und niedriglegierter Stähle im Temperaturbereich oberhalb 200 °C," in D. Kuron (ed.): *Wasserstoff und Korrosion,* Bonner Studien-Reihe, Verlag Irene Kuron, Bonn 1986, pp. 155 – 201.
41. H. H. Buchter: *Apparate und Armaturen der Chemischen Hochdrucktechnik,* Springer Verlag, Berlin – Heidelberg – New York 1967.
42. W. Wessling, F. Schmöning, H. E. Bock: "Aushärtbare nichtrostende Stähle," *Sie und wir,* Company brochure, Stahlwerke Südwestfalen, 1970, no. 6, pp. 41 – 59.
43. H. Sigwart: "Werkstofftechnische Grundlagen der Konstruktion," in Dubbel: *Taschenbuch für den Maschinenbau,* 13th ed., vol. 1, Springer, Berlin – Heidelberg – New York 1970.
44. American Petroleum Institute, Refining Department: *Steels for Hydrogen Service at Elevated Temperatures and Pressures in Petroleum Refineries and Petrochemical Plants,* API-Publication 941, 3rd ed., May 1983.
45. *Nichtrostende Stähle,* 2nd ed., Verlag Stahleisen, Düsseldorf 1989.
46. H. Wehner, H. Speckhardt: "Zum Ausscheidungs- und Korrosionsverhalten eines ferritisch-austenitischen Chrom-Nickel-Molybdän-Stahles nach kurzzeitiger Glühbehandlung unter besonderer Berücksichtigung des Schweißens," *Z. Werkstofftech.* **10** (1979) 317 – 332.
47. W. Wessling, H. E. Bock, W. Fuchs: "Neuere Entwicklungen von korrosionsbeständigen Sonderstählen für das Chemiewesen," *Z. Werkstofftech.* **4** (1973) 186 – 195.
48. E.-M. Horn, A. Kügler: "Entwicklung, Eigenschaften, Verarbeitung und Einsatz des hochsiliciumhaltigen Stahls X2CrNiSi1815," *Z. Werkstofftech.* **8** (1977) 362 – 370.
49. G. Hochörtler, E.-M. Horn: "Austenitic Stainless Steel With Approximately 5.3 % Silicon," *Met. Corros. Proc. Int. Congr. Met. Corros. 8th* **2** (1981) 1471.
50. R. Oppenheim, H. Kiesheyer, *Thyssen Edelstahl Tech. Ber.* **8** (1982) 97 – 100, 111 – 114.
51. K. Fäßler, H. Spähn: "Werkstoffauswahl bei wasserseitiger Beanspruchung," in: "Wärmeaustauscher," *DECHEMA Monogr.* **87** (1980) 51 – 72.
52. E.-M. Horn, D. Kuron, H. Gräfen: "Lochkorrosion an passiven Legierungssystemen der Elemente Eisen,

Chrom und Nickel," *Z. Werkstofftech.* **8** (1977) no. 2, 37 – 68.
53. E.-M. Horn, *VDI Ber.* **235** (1975) 39 – 45.
54. K. Lorenz, G. Medawar, *Thyssenforschung* **1** (1969) no. 3, 97 – 108.
55. R. W. Staehle, M. O. Speidel (eds.): *Handbook on Stress Corrosion Cracking and Corrosion Fatigue,* ARPA (Advanced Research Projects Agency), Washington 1978.
56. H. Spähn, *Z. Phys. Chem. (Leipzig)* **234** (1967) 1.
57. M. A. Harzenmoser, P. J. Uggowitzer: "Neue aufgestickte austenitisch-rostfreie Stähle und Duplexstähle," in: *Moderne Stähle,* vol. 1, Verlag der Schweizerischen Akademie der Werkstoffwissenschaften, Zürich 1987, pp. 205 – 218.
58. F. P. Pickering: "Physical Metallurgical Developments of Stainless Steels," in: *Stainless Steels 84,* The Institute of Metals, London 1985.
59. G. Gümpel, T. Ladwein, E. Michel, F. H. Strom: "Entwicklungen bei austenitischen Stählen mit erhöhten Festigkeitseigenschaften für den Einsatz im chemischen Apparatebau," *Thyssen Edelstahl Tech. Ber.* **14** (1988) no. 1, 12 – 25.
60. *Gußeisen Handbuch,* Gießerei-Verlag GmbH, Düsseldorf 1963.
61. K. Röhrig, D. Wolters: "Gußeisen mit Lamellengraphit und karbidischen Gußeisen," in: *Legiertes Gußeisen,* vol. 1, Gießerei-Verlag GmbH, Düsseldorf 1970.
62. *Die Ni-Resist-Gußeisenwerkstoffe,* publication no.16, 3rd ed., International Nickel Deutschland GmbH, Düsseldorf 1968.
63. O. Nickel: "Eigenschaften und Anwendungen austenitischer Gußeisenwerkstoffe in Erdölraffinerien und petrochemischen Werken," *Erdöl Kohle Erdgas Petrochem.* **18** (1965) 121 – 127, 197 – 205.
64. H. Grundmann: *Schweißen von Gußeisenwerkstoffen und Stahlguß,* Deutscher Verlag für Schweißtechnik GmbH, Düsseldorf 1971.
65. U. Draugelates: "Konstruktionsschweißen von Eisengußwerkstoffen mit Kugelgraphit, Leistungsstand und Verbindungsgüte," *VDI Ber.* **469** (1982) 95 – 103.
66. K. E. Volk (ed.): *Nickel und Nickellegierungen,* Springer Verlag, Berlin – Heidelberg – New York 1970.
67. W. Betteridge: *Nickel and Its Alloys,* MacDonald and Evans Ltd., Estover – Plymouth 1977.
68. U. Heubner et al.: *Nickel Alloys and High-Alloy Special Stainless Steels,* VDM, Expert Verlag, Sindelfingen 1987.
69. G. R. Pease: *Metallkundliche Vorgänge beim Schwei-ßen von Nickel und Nickellegierungen,* publication no. 6, International Nickel Deutschland GmbH, Düsseldorf 1957.
70. *Nickel-Eigenschaften, Verarbeitung und Verwendung,* publication no. 32, 1st ed., International Nickel Deutschland GmbH, Düsseldorf 1964.
71. *Beständigkeit von Nickel und Nickellegierungen gegenüber Ätzalkalien,* publication no. 8, 3rd ed., International Nickel Deutschland GmbH, Düsseldorf 1963.
72. W. Z. Friend: *Corrosion of Nickel and Nickel-Base Alloys,* J. Wiley and Sons, New York – Chichester – Brisbane – Toronto 1980.
73. Wiggin, Huntington: *Korrosionsbeständige Legierungen,* publication no. 3385, Nickel Alloys International S.A., Brüssel 1969.
74. *Schweißen von Nickel und seinen Legierungen,* publication no. 9, 1st ed., International Nickel Deutschland GmbH, Düsseldorf 1967.
75. *Nickel-Kupfer-Legierungen,* publication no. 20, 2nd ed., International Nickel Deutschland GmbH, Düsseldorf 1966.
76. *Korrosionsbeständigkeit nickelhaltiger Werkstoffe gegenüber Fluor, Fluorwasserstoff, Flußsäure und anderen Fluorverbindungen,* publication no. 59, 1st ed. , International Nickel Deutschland GmbH, Düsseldorf 1970.
77. A. J. Sedricks: *Corrosion of Stainless Steels,* J. Wiley and Sons, New York – Chichester – Brisbane – Toronto 1979.
78. *Chrom-Nickel-Legierungen gegen Brennstoffaschen-Korrosion,* publication no. 2942 D , International Nickel Deutschland GmbH, Düsseldorf 1965.
79. H. Gräfen, G. Böhm: "Die interkristalline Korrosion von Ni-Mo- und Ni-Mo-Cr-Legierungen," *Z. Metallkd.* **51** (1960) 245 – 252; *Nickel-Ber.* **19** (1961) 3 – 9.
80. H. Gräfen: "Eigenschaften von Ni – Cr – Mo-Legierungen und ihre Verbesserungsmöglichkeiten," *Chem. Ing. Tech.* **35** (1963) 229 – 235.
81. R. Köcher: "Neue Werkstoffe für den Chemie-Apparatebau und ihre schweißtechnische Verarbeitung," *VDI Z.* **112** (1970) 749 – 808.
82. W. Schatt (ed.): *Werkstoffe des Maschinen-, Anlagen- und Apparatebaues,* VEB-Verlag, Leipzig 1983.

83. Aluminium-Zentrale (ed.): *Aluminium-Taschenbuch,* Aluminium-Verlag, Düsseldorf 1984.
84. W. Huppatz, in: H. Gräfen et al. (eds.): *Die Praxis des Korrosionsschutzes,* Expert Verlag, Grafenau 1981, pp. 64 – 81.
85. A. Oechslin: "Hartanodisieren von Aluminiumlegierungen," *Schweiz. Alum. Rundsch.* **19** (1969) 8.
86. F. F. Berg: *Korrosionsschaubilder,* VDI-Verlag, Düsseldorf 1965.
87. W. Hornig, Diplomarbeit, Universität Erlangen-Nürnberg 1977.
88. H. Holten, H. Sigurdsson, *Werkst. Korros.* **28** (1977) 475 – 477.
89. E. Wendler-Kalsch, *VDI-Ber.* **365** (1980) 17 – 35.
90. *Kupfer – Eigenschaften, Verarbeitung, Verwendung,* Deutsches Kupfer-Institut e.V., Berlin 1961.
91. *Kupfer und Kupferlegierungen in Industrie und Handwerk,* Deutsches Kupfer-Institut, e.V., Berlin 1952.
92. O. v. Franqué, *Sanit., Heiz., Klimatech.* 1984, nos. 1 + 2, 34 – 40, 94 – 97.
93. E. Wendler-Kalsch in: *Korrosion in Kalt- und Warmwassersystemen der Hausinstallation,* DGM, Oberursel 1984, pp. 51 – 70.
94. E. E. Langenegger, F. P. A. Robinson: "The Role of Arsenic in Preventing the Dezincification of α-Brass," *Corrosion* (Houston) 25 (1969) 137 – 143.
95. D. H. Thompson, A. W. Tracy, *Metals Trans.* 1949, 100. A. W. Blackwood, N. S. Stoloff, *ASM Trans. Q.* **62** (1969) 677.
96. E. Wendler-Kalsch, *Z. Werkstofftech.* **13** (1982) 129 –137.
97. H. Vogel, Dissertation, Universität Erlangen-Nürnberg 1985.
98. H. Vogel, E. Wendler-Kalsch, *Z. Metallkd.,* **75** (1984) 217 – 221.
99. W. Hofmann: *Lead and Lead Alloys,* Springer Verlag, Berlin – Heidelberg – New York 1970.
100. W. W. Krysko: "Homogen verbleiter Stahl als Konstruktionswerkstoff," *Werkst., Korros.* **15** (1964) 631 – 633, 726 – 728, 991 – 995; **17** (1966) 235 – 238.
101. H. Gräfen, D. Kuron: "Neuentwicklung von Bleilegierungen aufgrund elektrochemischer Untersuchungen über das Korrosionsverhalten in Schwefelsäure," *Werkst., Korros.* **21** (1970) 3 – 16.
102. M. Hörmann, D. Lupton, H. Heintze, E.-M. Horn: "Sondermetalle und ihre Anwendung im Chemieapparatebau," *Z. Werkstofftech.* **18** (1987) 139 – 147, 186 – 194.
103. K. Rüdinger: "Moderne Werkstoffe, Auswahl – Prüfung – Anwendung, Übersichten über Sondergebiete der Werkstofftechnik für Studium und Praxis," *Z. Werkstofftech.* **9** (1978) 181 – 188.
104. K. Rüdinger: "Technologische Eigenschaften und Korrosionsverhalten einer Titanlegierung mit 0,2 % Palladium," *Werkst., Korros.* **16** (1965) 109 – 115.
105. H. Schussler: *Corrosion Data Survey on Tantalum,* Fansteel Inc., North Chicago, Ill., 1972, p. 104.
106. H. Dickmann, E.-M. Horn, U. Gramberg: Erfahrungen mit hochschmelzenden Metallen in der Chemietechnik, VDI-W Meeting, Köln, Dec. 6 – 7, 1984, pp. 87 – 95.
107. D. R. Knittel, R. T. Webster: "Corrosion Resistance of Zirconium and Zirconium Alloys in Inorganc Acids and Alkalis," in: E. W. Kleefisch (ed.): "Industrial Applications of Titanium and Zirconium," *ASTM Spec. Tech. Publ.* **728** (1981) 191 – 203.
108. R. D. Knittel: "Zirconium," in: P. A. Schweitzer (ed.): *Corrosion and Corrosion Protection Handbook,* Marcel Dekker, New York – Basel 1983, pp. 209 – 210.
109. T. Günther: "Zusammenhang zwischen Gefüge und Korrosionsverhalten bei Zirconium," *Werkst. Korros.* **30** (1979) publication no. 5, 308 – 321.
110. R. Köcher: "Nichteisenmetalle im Apparate- und Anlagenbau," *Chem. Ing. Tech.* **55** (1983) no. 10, 752 – 762.
111. U. Gramberg, E.-M. Horn, K.-O. Cavalar: "Explosionsplattierte Bleche in Anlagen, Erfahrungen aus der Chemietechnik," VDI-W Meeting: Explosionsplattieren – ein modernes Verfahren zur Herstellung von Hochleistungsverbundsystemen, Düsseldorf, Dec. 1, 1983, pp. 29 – 35.
112. R. Köcher, *Werkst. Korros.* **28** (1977) 166 – 173.
113. *Kunststoff-Handbuch* **(12 vols.),** Hanser, München 1963 – 1975.
114. H. Domininghaus: *Die Kunststoffe und ihre Eigenschaften,* 3rd ed., VDI-Verlag, Düsseldorf 1988.
115. W. Nestler: "Kunststoffe im Apparatebau," *Kunststoffe* **60** (1970) 719 – 732.
116. Th. Hadick: "Die schweißtechnische Praxis," in: *Schweißen von Kunststoffen für Praktiker und Konstrukteur,* vol. 6, Deutscher Verlag für Schweiß-technik GmbH, Düsseldorf 1970.

117. R. Taprogge: *Konstruieren mit Kunststoffen,* VDI-Taschenbuch 21. VDI-Verlag, Düsseldorf 1971.
118. G. Ehrbar: "Kunststoffe im chemischen Apparatebau," *Kunststoffe* **53** (1963) 845 – 854.
119. E. Dolfen: Bemessungsgrundlagen für tragende Bauelemente aus glasfaserverstärkten Kunststoffen, insbesondere aus durch Glasseidenmatten bewährten Polyesterharzen, Dissertation, TH Aachen, 1968.
120. F. Fischer: "Spannungsrißbildung und Spannungsrißkorrosion bei Kunststoffen," *Z. Werkstofftech.* **1** (1970) no. 2, 74 – 83.
121. S. Boström: *Kautschukhandbuch,* **vol. 5,** Verlag Berliner Union, Stuttgart 1962, pp. 340 – 369.
122. R. Grünewald: "Konstruieren mit glasfaserverstärkten Kunststoffen," *Kunststoffberat. Rundsch. Tech.* **5** (1967) 356 – 361.
123. H. Haferkamp: *Glasfaserverstärkte Kunststoffe,* VDI-Taschenbuch 17. VDI-Verlag, Düsseldorf 1970.
124. S. Lohmeyer et al.: "Kontakt und Studium," in: *Werkstoff Glas,* vol. 22, Lexika-Verlag, Grafenau/Württemberg 1979.
125. H. Scholze: *Glas,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York 1977.
126. A. Peters: "Chemisches Verhalten und physikalische Daten von Borosilikatglas," *Haus Tech. Vortragsveröff.* **211** (1969) 6 – 10.
127. F. Brandmair: "Zweckmäßige Apparatekonstruktionen aus Kunstkohle und Elektrographit und deren Einsatzgebiete," *Werkst. Korros.* **17** (1966) 10 – 17.
128. H. Wurmseher: "Graphit im chemischen Apparatebau," *Tech. Mitt.* **62** (1969) 226 – 230.
129. F. Harders, S. Kienow: *Feuerfestkunde,* Springer Verlag, Berlin – Göttingen – Heidelberg 1960.
130. K. Konopicky: *Feuerfeste Baustoffe,* Verlag Stahleisen, Düsseldorf 1957.
131. J. Agst: *Die feuerfesten Baustoffe,* 6th ed., Scimarowsky and Agst, Kapellen/Moers 1971.
132. F. K. Falcke: *Kleines Handbuch des Säureschutzbaues,* Verlag Chemie, Weinheim 1966.
133. H. Lahnang, H. Scholze: "Keramische Werkstoffe," in: *Keramik,* part 2: 6th ed., Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1983.
134. H.-J. Bargel et al. in H.-J. Bargel, G. Schulze (eds.): *Werkstoffkunde,* 5th ed., VDI-Verlag, Düsseldorf 1988.

Contact Lenses → Ophthalmological Preparations

Corrosion

Hubert Gräfen, Bayer AG, Leverkusen, Federal Republic of Germany

Elmar-Manfred Horn, Bayer AG, Leverkusen, Federal Republic of Germany

Hartmut Schlecker, Bayer AG, Leverkusen, Federal Republic of Germany

Helmut Schindler, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 2
2. Basic Electrochemistry 3
2.1. Electrochemical Equilibrium . . . 3
2.2. Electrochemical Kinetics 5
2.2.1. Corrosion Elements 6
2.2.2. Mixed Electrode and Mixed Potential 7
2.2.3. Overpotential and Polarization . . . 7
2.3. Current Density – Potential Curves 8
2.4. Hydrogen Overpotential and Acid Corrosion 11
2.5. Oxygen Overpotential and Oxygen Corrosion 12
2.6. Protective Film Formation and Passivity 14
3. Types of Electrolytic Corrosion . 16
3.1. Uniform and Nonuniform Surface Corrosion 16
3.2. Forms of Localized Corrosion without Applied Mechanical Stress 17
3.2.1. Pitting Corrosion 17
3.2.2. Crevice Corrosion 21
3.2.3. Selective Corrosion 22
3.2.3.1. Intergranular Corrosion 23
3.2.3.2. Spongiosis 26
3.2.3.3. Dezincification 27
3.2.3.4. Line- and Layer-Type Corrosion . . 28
3.3. Forms of Local Corrosion under Mechanical Stress 30
3.3.1. Stress Corrosion Cracking 30
3.3.1.1. Unalloyed and Low-Alloy Steels . . 31
3.3.1.2. Stainless Steels 33
3.3.1.3. Nonferrous Metals 35
3.3.1.4. Special Types of Stress Corrosion Cracking 35
3.3.1.5. Hydrogen-Induced Cracking at Low Temperatures 37
3.3.1.6. Hydrogen Corrosion at Elevated Temperatures 39
3.3.2. Corrosion Fatigue 41
3.3.3. Differentiation Between Types of Corrosion Cracking 43
4. High-Temperature Corrosion . . . 44
4.1. Principles of High-Temperature Corrosion 44
4.1.1. Thermodynamic Principles 44
4.1.2. Kinetics of Scaling 45
4.2. Oxidation of Alloys 47
4.3. Sulfurization of Metals and Alloys 47
4.4. Corrosion by Combustion Gases . 48
4.5. Effect of Ash, Deposits, and Salts 49
4.6. Corrosion by Synthesis Gases . . 50
4.7. Oxidation Behavior of Technical Alloys 51
4.7.1. Steels 51
4.7.2. Nickel Alloys 54
4.7.3. Cobalt Alloys 54
5. General Corrosion Protection Procedures 55
5.1. Corrosion Protection by Coating 55
5.1.1. Metal Coatings 55
5.1.2. Inorganic Nonmetallic Protective Coatings 59
5.1.2.1. Enamel 60
5.1.2.2. Brick Lining 62
5.1.2.3. Thermally Sprayed Coatings 62
5.1.3. Organic Lining and Coating Materials 63
5.1.3.1. Rubber Linings 64
5.1.3.2. Rubber – Plastic Composite Linings 65
5.1.3.3. Linings Consisting of Thermally Crosslinkable Thermosetting Plastics 66
5.1.3.4. Reinforced Reaction Resin Coatings, Crosslinkable with Catalysts . 66
5.1.3.5. Thermoplastic Linings 67
5.1.3.6. Paints and Powder Coatings 68
5.2. Inhibitors 69
6. Electrochemical Corrosion Protection 70
7. Corrosion Testing 71
7.1. Introduction 71
7.2. Test Methods 73
8. References 75

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_08

1. Introduction

When in use, all materials are exposed to environmental effects that can impair or even destroy their usefulness or that of components made out of them. Most of the damage in the case of metals is caused by corrosion – reactions with the environment which can lead to a measurable change in the material (cf. DIN 50 900, Part 1 [2]). The reactions are mainly electrochemical, although they can also involve chemical or metallo-physical processes.

Material changes due exclusively to mechanical effects are attributable to wear rather than to corrosion. However, a corrosive reaction with the environment can be brought about by erosion of protective coatings (erosive corrosion).

The tendency of metals to change with loss of energy into the disordered, more stable thermodynamic state represented by compounds is the reason for corrosion. However, kinetic barriers, mainly provided by the formation of protective layers, often make metals resistant to corrosion and allow them to be used in technological applications.

The types of material changes brought about by corrosion are manifold. A distinction is made between uniform and nonuniform surface corrosion, pitting, selective corrosion, and cracking. Of these, the local corrosion processes are particularly critical.

Since corrosion reactions generally occur on the metal surface, they are termed interfacial processes and can be represented by a phase diagram. A distinction is made between the metal phase, the liquid, gaseous, or solid medium, and the phase boundary (Fig. 1). The corrosion process takes place at the metal – medium phase boundary and is therefore a heterogeneous reaction in which the structure and condition of the reaction surface play a significant role. Essential points, for example, are whether the surface is uncoated, whether it is covered with an adhesive, compact or loose, porous coating, or whether its properties have been changed by machining and processing. Furthermore, the corrosive medium must be transported to the surface, and the corrosion products removed. Hence, material transport phenomena, including free convection and diffusion into surface layers, must be taken into account.

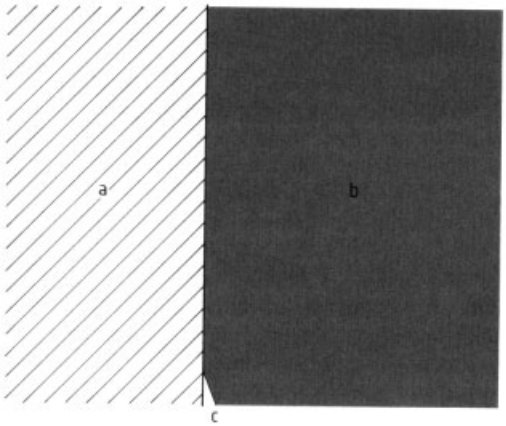

Figure 1. Corrosion as a phase interface reaction
a) Metal; b) Solid, liquid, or gaseous medium; c) Phase boundary

2. Basic Electrochemistry (See also → Electrochemistry)

If the corrosive medium is an electrolyte solution, the resulting corrosion reactions are electrochemical. This means that material transport in the form of metal ions and charge exchange in the form of electrons take place at the metal – solution phase boundary because of the conductivity caused in the liquid phase by mobile anions and cations and the electron conductivity of the metals.

Aqueous corrosive media are electrolyte solutions, and hence by far the greatest proportion of damage arising through corrosion is due to electrochemical processes. Since in metals these are largely identical with electrolysis processes, metal corrosion in aqueous media is best described as electrolytic corrosion. The most important definitions are set out in DIN 50 900, Part 2 [3].

2.1. Electrochemical Equilibrium

When a piece of metal is dipped into water a number of processes take place. Metal atoms leave the solid phase and enter the liquid as cations. This disturbs the equilibrium in the metal phase; the metal now contains a surplus of electrons, thus becoming electrically negative, while a positively charged ion cloud forms

in the liquid. This cloud is held close to the metal surface by the attractive force of the electrons, resulting in an electrical double layer in which an electrical field, directed from the ion cloud to the metal surface, exists. The dissolution of the metal is self-inhibitory because all departing metal ions must overcome this field. With increasing concentration of the ion cloud, the exit energy released is less and less capable of supplying the energy needed to overcome the electrical field, which results in a dynamic equilibrium: the rate of metal dissolution equals the rate of metal deposition, with the exchange current density i_0. The double layer results in a potential difference between metal and solution (Fig. 2A).

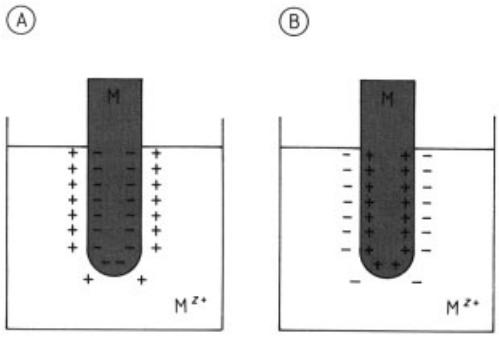

Figure 2. Formation of a potential difference between electrode and electrolyte
A) $\mu_{M_z^+}$ (metal) $> \mu_{M^{z+}}$ (solution);
B) $\mu_{M^{z+}}$ (metal) $< \mu_{M^{z+}}$ (solution)
μ = Chemical potential

If ions of the metal are already present in the liquid when the metal electrode is immersed, far fewer metal ions will go into solution. If the concentration is higher than that of the equilibrium double layer, metal ions will deposit on the metal surface and return to the solid state, and the direction of the potential difference at the phase boundary is reversed (Fig. 2B).

A potential difference also forms with an electrode that is not attacked by the electrolyte solution (e.g., platinum): through electron loss or gain, the dissolved substances are changed into an oxidized or reduced form (redox couple). An example is the following redox system:

$$Fe^{3+} + e^- \rightarrow Fe^{2+}$$

The immersed inert electrode acts as an electron donor or acceptor. Electrons are exchanged at the metal surface until an electrochemical equilibrium is reached between the electrons in the metal and in the solution.

Gas electrodes are a further important type of electrode in corrosion. These involve the formation, on the metal surface, of gases from dissolved ions, or of ions from dissolved gases. In this instance the metal has the function of consuming or providing electrons. In pure form such processes occur on metals (e.g., platinum) which themselves exhibit no reaction with electrolyte solutions. Of greatest importance in corrosion processes are hydrogen and oxygen electrodes. The reduction of hydrogen ions in acid solutions ($2\,H^+ + 2\,e^- \rightarrow H_2$) and the reduction of dissolved oxygen in approximately neutral solutions ($O_2 + 4\,e^- + 2\,H_2O \rightarrow 4\,OH^-$) are the most important cathodic reactions.

Standard Potentials and Electrochemical Series. The potential of the electrodes (half cells) described above is measured by connecting them by means of a solution bridge to a second electrode, whose potential is used as a reference. The potential difference between the two half cells, with zero current flow, is then measured. An electrode potential U measured in this way differs from the absolute potential $\Delta\varphi$ by an unknown constant which is dependent on the reference electrode. The potential difference arising at the point of contact of the electrolyte solutions (diffusion potential) can largely be eliminated by interposing concentrated salt solutions with ions of identical mobility. The additional potential changes caused by the instrument leads between electrode and instrument in each of the two cells can be disregarded since they are in opposing directions.

The hydrogen electrode in a given concentration (platinized platinum electrode immersed in a solution with a H^+ activity of 1 and purged with H_2 at 100 kPa) is generally used as the standard electrode; its potential is arbitrarily set at zero. The potentials of electrodes measured against the standard hydrogen electrode are designated by the symbol U_H. Figure 3 (see next page) shows the determination of the electrode potential of silver. The potentials of other metal electrodes, relative to the standard hydrogen electrode, in a solution of their own salts with metal-ion activity 1 are called standard potentials.

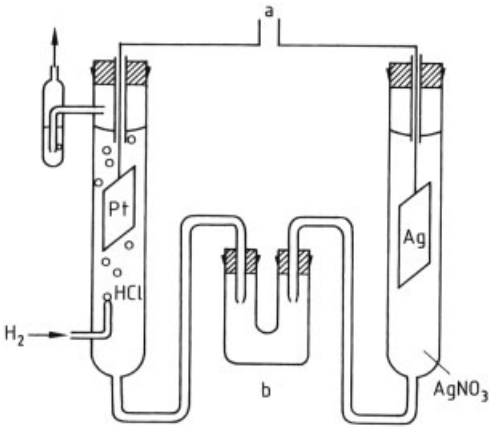

Figure 3. Determination of the electrode potential of silver a) High-impedance voltmeter; b) Solution bridge

The electrochemical series of metals is a listing of the standard potentials U_H^0 of metals determined at 25 °C. Some of the most important standard potentials are listed in Table 1.

Table 1. Some important standard potentials

Electrode	U_H^0, V
Mg/Mg^{2+}	− 2.350
Al/Al^{3+}	− 1.660
Zn/Zn^{2+}	− 0.762
Fe/Fe^{2+}	− 0.440
Ni/Ni^{2+}	− 0.236
Sn/Sn^{2+}	− 0.141
Pb/Pb^{2+}	− 0.126
H_2/H^+	0.000
Cu/Cu^{2+}	+ 0.345
O_2/OH^-	+ 0.402
Ag/Ag^+	+ 0.800
Hg/Hg^{2+}	+ 0.861
Cl_2/Cl^-	+ 1.359
Au/Au^{3+}	+ 1.500

Equilibrium electrode potentials at other concentrations are calculated from the Nernst equation:

$$U_H = U_H^0 + \frac{RT}{zF} \cdot \ln a_{M^{z+}} \quad (1)$$

where R is the gas constant, z is the valency of the metal ions, F is the Faraday constant, and $a_{M^{z+}}$ is the activity of potential-determining metal ions in the solution. For a redox electrode the equation is

$$U_H = U_H^0 + \frac{RT}{zF} \cdot \ln \frac{a_{ox}}{a_{red}} \quad (2)$$

The change in equilibrium potential U_H^0 in this instance depends on the activity quotient of the oxidized and reduced species in solution. The equation for a gas electrode, e.g., the hydrogen electrode, is as follows:

$$U_H = U_H^0 + \frac{RT}{zF} \ln a_{H^+} - \frac{RT}{zF} \ln p_{H_2} \quad (3)$$

The deviations from U_H^0 are due to the activity of the hydrogen ions and the hydrogen partial pressure. If both are unity, then $U_H^0 = U_H$ and the conditions of the standard hydrogen electrode prevail.

2.2. Electrochemical Kinetics

If two half cells with differing electrode potentials are joined together, an electrochemical element (voltaic cell) arises with a current flow from the more negative to the more positive electrode, caused by the potential difference between the two electrodes (Fig. 4). Since the electrons released at the anode through metal dissolution migrate to the cathode and are consumed there (in this instance by metal deposition) the flow of current continues as long as the reaction partners remain available. The constant electrical field arising at one electrode at equilibrium, described in Section 2.1 (double layer), is not evident in this case and, therefore, cannot slow down the process. Accordingly, an electrochemical reaction can only take place when the electrode potential deviates from the equilibrium potential, whereby the rate of dissolution of the metal or reaction of the electrons is described by a current intensity I, which is equivalent to the respective mass conversion in accordance with Faraday's Law. For the dissolved metal mass ΔM and the transported electric charge Q in A · s the following equation applies:

$$-\Delta M = A_r Q / zF \quad (4)$$

where A_r is the relative atomic mass of the metal, z is the ion valency, and F is the Faraday constant (96 520 A · s or 26.8 A · h). This constant gives the amount of charge required for the dissolution of 1 mol of a metal. For example, 1 A · h dissolves 1.04 g of pure iron. As a rule of thumb, the following applies for unalloyed steels:

$$0.1\,\mathrm{mA\,cm^{-2}} \simeq 1\,\mathrm{g\,m^{-2}\,h^{-1}}$$

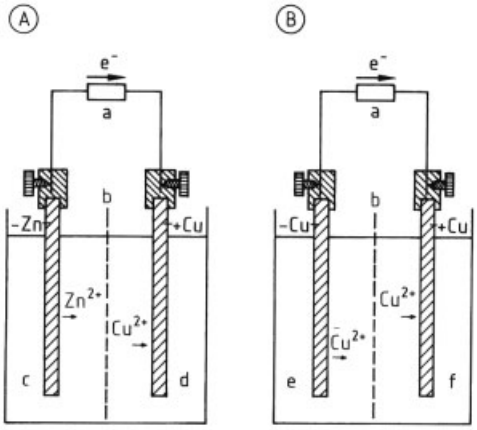

Figure 4. Electrode cells
A) Daniell cell; B) Concentration cell
a) External resistance; b) Diaphragms; c) $ZnSO_4$ solution; d) $CuSO_4$ solution; e) Dilute $CuSO_4$ solution; f) Concentrated $CuSO_4$ solution

In contrast to electrolysis with current supply, in spontaneous corrosion with current yield, the dissolving electrode is known as the anode and its counterpart, at which, for example, cations are deposited, as the cathode. That is, the electrode at which negative charges enter the electrolyte solution is the cathode.

Voltaic cells can also be formed by combining metal – salt electrodes with metal – gas electrodes or metal – redox electrodes. The metal dissolution that occurs at the anode corresponds to the partial process that leads to material destruction in electrochemical corrosion. This also occurs when the electrolytic solution initially contains no metal ions, which is usually the case in corrosion processes, because only one element is needed, which is formed by the coupling of an electron-delivering (anode) and an electron-consuming (cathode) electrode. For this to occur, the cathode potential must be the higher of the equilibrium potentials of the individual electrode reactions. Therefore, noble metals such as gold do not corrode in aqueous solutions because, with their high anodic potentials, more positive cathode reactions are no longer available.

2.2.1. Corrosion Elements

A corroding metal – solution system forms anode and cathode areas, whose sizes range from a few lattice constants to macroscopic areas. Figure 5 shows a comparison between a galvanic and a corrosion cell. The corrosion cell, like the galvanic cell, can arise only when at least two electrode processes take place, one providing the electrons and the other consuming them. This corresponds to the composition of a galvanic cell consisting of at least two half cells.

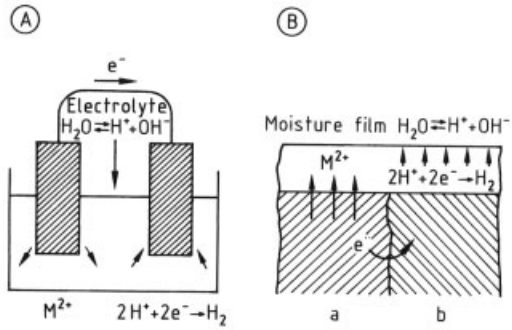

Figure 5. Comparison of a galvanic cell and a corrosion cell
A) Galvanic cell; B) Corrosion cell
a) Anode; b) Cathode
Anode process: $M \rightarrow M^{2+} + 2\,e^-$ (dissolution of metal)
Cathode process: $2\,H^+ + 2\,e^- \rightarrow H_2$ (evolution of hydrogen)

Metal destruction (dissolution) takes place in the material area with the more electronegative potential (anode) since this is where the metal ions from the surface enter the solution. The electrons released in the process migrate to the electropositive cathode and are consumed there at the phase boundary by reducible substances present in the solution (e.g., cations, dissolved oxygen).

The anodic process (anode reaction) thus consists of metal dissolution (oxidation)

$$M \longrightarrow M^{2+} + 2\,e^-$$

while the cathodic process (cathode reaction) is a reduction:

a) $2\,H^+ + 2\,e^- \longrightarrow H_2$ hydrogen ion reduction
b) $O_2 + 2\,H_2O + 4\,e^- \longrightarrow 4\,OH^-$ oxygen reduction
c) $Cu^{2+} + 2\,e^- \longrightarrow Cu$ metal reduction
d) $Fe^{3+} + e^- \longrightarrow Fe^{2+}$

In practice, the cathode reactions (a) and (b) are particularly important because of their frequency. In the corrosion of metals in acid solutions, the reduction of hydrogen ions is the main reaction, whereas oxygen reduction is the main

reaction in the corrosion of metals in neutral, aerated electrolyte solutions (e.g., salt solutions, seawater, atmosphere).

2.2.2. Mixed Electrode and Mixed Potential

A corroding metal surface represents a collection of corrosion elements, with the constancy, the spread, and the attendant macroscopic interface of the anode and cathode surfaces depending on the system. The distribution of local anodic and cathodic current densities i_a and i_c at the surface of an electrode can differ appreciably. Possibilities are $i_a = i_c$, $i_a - i_c = + i_E$ and $i_a - i_c = - i_E$. The current i_E per unit area of the corrosion element can thus be zero, positive, or negative. Only in certain borderline cases are the anode and cathode surfaces physically separated from one another (local elements). The principle of additivity (a superimposition of all electrode reactions taking place on a metal surface, see Section 2.3), first formulated by WAGNER and TRAUD [4], makes these statements comprehensible. A multiple or mixed electrode of this kind is homogeneous when i_E is zero, and is heterogeneous when, as a result of locally differing current densities, it is nonzero. The mixed potential of a homogeneous mixed electrode is identical over the entire surface, whereas a heterogeneous mixed electrode exhibits different mixed potentials.

Considering the corrosion system as a whole, the sum of the cathode currents must equal the sum of the anode currents for reasons of electroneutrality. Thus the cathode and anode reactions affect one another reciprocally, and the slowest (most inhibited) determines the overall reaction rate. In many cases the cathode reaction is the slowest (cathode-controlled corrosion). Lack of external current can only arise at a given mixed potential, the free corrosion potential.

2.2.3. Overpotential and Polarization

The corrosion behavior of metals cannot be predicted from the position of their standard potentials in the electrochemical series because the potential of an electrode changes with the current density. If an electrode in which only one electrode process takes place is termed a working electrode and the resultant potential a working potential, then the difference between working potential and the Nernst equilibrium potential is called overpotential; it is caused by reaction restraints. In general, polarization is defined as the shift in potential of working electrodes within a corrosion element. In such an element, at least two electrode reactions occur, whose overpotentials are superimposed, resulting in the polarization effect.

Transfer overpotential is the inhibition of transfer of charge carriers (ions and electrons) as a result of the electrical field on the metal surface. The potential barrier between metal and solution can only be overcome by ions of sufficiently high energy, which requires a corresponding deviation from the equilibrium potential of the electrode reaction. In reversible electrochemical reactions, anodic current dominates for $\Delta U > 0$, and cathodic current for $\Delta U < 0$. Near the equilibrium potential, the external current is the difference of the two partial current densities:

$$i = i_0 \left[\exp\left(\frac{\Delta U \cdot zF\alpha^+}{RT} \right) - \exp\left(\frac{\Delta U \cdot zF\alpha^-}{RT} \right) \right] \quad (5)$$

where i_0 is the exchange current density at the equilibrium potential of the reaction, and α is the charge transfer factor. For high overpotentials the second exponential expression is negligible and the equation simplifies to

$$i = i_0 \exp\left(\pm \frac{\Delta U \cdot zF\alpha}{RT} \right) \quad (6)$$

Taking logarithms and transforming gives the Tafel equation,

$$\log i = \log i_0 \pm \frac{\Delta U \cdot zF\alpha}{RT} \cdot \frac{1}{2.3} \quad (7)$$

$$\Delta U = \pm\, 2.3 \frac{RT}{zF\alpha} \log i_0 \pm 2.3 \frac{RT}{zF\alpha} \log i \quad (8)$$

which in simplified form is written:

$$\Delta U = a + b \log i \quad (9)$$

For the anodic dissolution of metals the following applies:

$$a = -2.3 \frac{RT}{\alpha zF} \log i_0 \text{ and } b = 2.3 \frac{RT}{\alpha zF}$$

Since a and b are constants, plotting the anode and cathode overpotential curves of a simple

electrode reaction in semi-logarithmic representation results in straight lines (Tafel lines) when the overpotential ΔU is plotted on the linear and the current density i on the logarithmic axis (Fig. 6). Individually, however, they cannot be directly measured; their superimposition gives a sum curve, the actual overpotential curve, as shown by the dashed line in Figure 6.

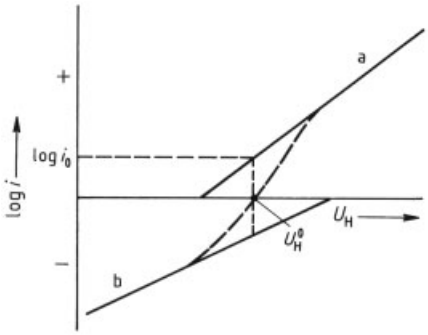

Figure 6. Overpotential curve for a simple electrode reaction
a) Overpotential curve for the anode reaction; b) Overpotential curve for the cathode reaction
U_H^0 = Equilibrium potential; i_0 = Exchange current density

The value of α generally lies between 0.45 and 0.55; as an approximation a value of 0.5 can be used. The constant b indicates the gradient of the straight lines. For $z = 2$, b is almost always 0.05 – 0.06 V (for $z = 1$, b is 0.1 – 0.12 V). Since in a number of electrode reactions the situation is complicated by the occurrence of multistep reactions, the Tafel equation has only limited applicability.

In the dissolution and deposition of most metals, the activation energy is small and hence the transfer overpotential is low. In transition metals such as iron and nickel it is much higher. Conspicuously high overpotentials indicate electrode processes in which hydrogen and oxygen are involved.

The effect of concentration c, resulting from the concentration changes at the electrode, leads to a concentration overpotential:

$$\Delta U = 2.3 \frac{RT}{nF} \log c \tag{10}$$

At 18 °C, $2.3\,RT/F = 0.058$ V. Consequently, a concentration shift (more precisely a change in activity) by a factor of 10 results in a potential shift of 58, 29, or 19 mV, depending on the number of electrons n involved (change of valency). The concentration overpotential plays an important role where a reaction partner (e.g., oxygen) is consumed by corrosion. In this case the corrosion rate is determined by the rate of transport (by convection and diffusion) of the oxygen required for the cathodic reaction. Since for each electrode reaction the supply of species to, and removal from, the interface is essential, any barrier to such diffusion processes leads to a diffusion overpotential.

When one of the reaction steps at the metal surface is a purely chemical reaction, any inhibition of that reaction leads to a reaction overpotential. It is conceivable, for example, that an oxidizing agent needed to sustain the cathodic reaction is formed by a chemical reaction. If this is the slowest individual process the reaction overpotential will be determined by this reaction and the transport of material to the electrode.

A further barrier to corrosion reactions is provided by electrical resistance. When the anodic and cathodic reactions at the metal surface take place with locally different current densities (heterogeneous mixed electrode), resistance in the current circuits can cause a measurable drop in potential (resistance polarization). This resistance polarization is a linear function of the current. Resistance polarization frequently arises through the formation of passive films. The resulting relationship between the change in potential and the current usually no longer follows Ohm's law, but instead is subject to a logarithmic relationship.

2.3. Current Density – Potential Curves

Theoretical predictions of the overpotential (polarization) at a given current density are generally not possible, and it is impossible to determine the current densities to be expected during corrosion (a measure of the corrosion rate) from the equilibrium potentials. The relationship between potential and current density can normally only be determined experimentally.

The results of such measurements are known as current density – potential curves. They represent cumulative curves given by the superimposition of the current density – potential curves of the individual reactions. For simple electrodes with defined electrode processes, these are the

overpotential curves (see Fig. 6). For metals exposed to electrolytic attack, superimposition of several overpotential curves gives the actual current density – potential curves that are of significance in corrosion testing and research. Figure 7 illustrates how two overpotential curves superimpose to form a current density – potential curve when both reactions take place at a metal surface. The potential is plotted on the abscissa and the current density on the ordinate. This presentation corresponds to the normal practice of measuring current density at a given potential. Figure 7 shows the superposition of the overpotential curves of a hydrogen and an iron electrode. The metal surface on which the two reactions occur appears to be without external current, although current flows on the metal surface. However, these are internal processes that have no outward effect since the anodic current must be the same as the cathodic current. In the example given this means that the extent of metal dissolution must be equal to the extent of hydrogen evolution. When there is no resistance in the circuit, which can be assumed for electrolytes with good conductivity, this condition is met at the common free corrosion potential (mixed potential) U_{corr}. The U_{corr} is the intersection of the measured cumulative current – potential curve with the abscissa and is also called the rest potential when the electrode reactions are homogeneous.

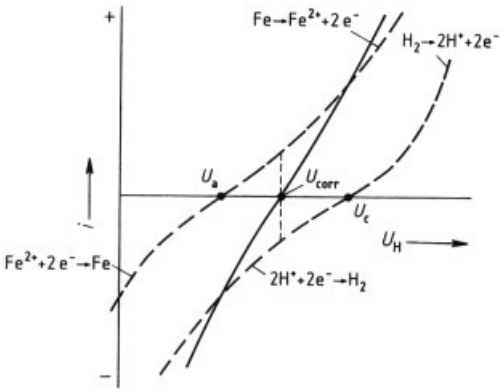

Figure 7. Superposition of two overpotential curves to form a current density – potential curve
U_{a} = Equilibrium potential of the overpotential curve for the anode reaction; U_{c} = Equilibrium potential of the overpotential curve for the cathode reaction; U_{corr} = Corrosion potential (rest potential)

Since only the cathode branch (hydrogen reduction) is of interest on the overpotential curve of the hydrogen electrode (right) and only the anode branch (metal dissolution) on the iron electrode (left), the curve can be simplified, as shown in Figure 8. Electrical resistance in the circuit of a dissolving metal electrode always leads to reduced metal dissolution.

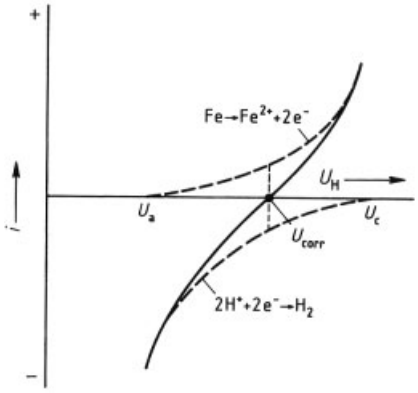

Figure 8. Superposition of the anodic and cathodic partial reactions to form a current density – potential curve (simplified version compared to Fig. 7)
– – – – = Partial current curves; —— = Cumulative current curve; U_{corr} = Corrosion potential

Even though normal corrosion without external current is only shown by the point of intersection of the cumulative current density – potential curve and the abscissa (U_{corr}), the anode and cathode branches, i.e., the full course of the curve measurable with external current supply, are still of interest since they reflect all the peculiarities of the polarization curves involved. From such diagrams, therefore, conclusions can often be drawn as to the corrosion performance of metals under given environmental conditions.

The procedure for determining a current density – potential curve is to immerse the metal electrode in the corrosive medium. After an incubation period, the mixed potential corresponding to the outwardly currentless state becomes established on the metal. If a circuit is now placed on the electrode and external current is applied, the result is a shift in potential (polarization). Current density – potential curves are therefore also known as polarization curves. Depending on the direction of the external current imposed, this is termed anodic or cathodic polarization (anodic: potential becomes positive; cathodic: potential becomes negative).

Determination of Corrosion Current Density. Simple, purely transfer-related electrode reactions give cumulative current density – potential curves of the type shown by the unbroken line in Figure 9. At the points p_1 and p_2 it swings into the overpotential curves of the respective part reactions, because beyond these points, only the anodic or the cathodic reaction exists. At these points the equilibrium potentials of the reverse reaction are exceeded and superimposition no longer occurs. These pure overpotential curves thus form linear Tafel lines, which, after reflection of the cathode curve in the x-axis, can be made to intersect by extrapolation in the direction of the abscissa. The ordinate section at the point of intersection is then $\log i_{\mathrm{corr}}$, i.e., the log of the corrosion current density at the free corrosion potential (rest potential), from which the corrosion rate can be calculated by Faraday's law (see Section 2.2).

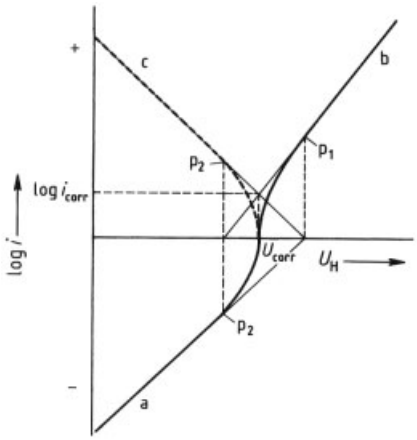

Figure 9. Determination of the corrosion current density i_{corr}
a) Overpotential curve for the cathode reaction; b) Overpotential curve for the anode reaction; c) Reflection of curve a

The simple correlation between polarization resistance and corrosion rate at the free corrosion potential developed by WAGNER and TRAUD [4], was later extended by STERN and GEARY for practical application [5]. By this method the corrosion rate is estimated from the gradient of the current density – potential curve (polarization resistance):

$$i_{\mathrm{corr}} = \frac{1}{2.3\left(\frac{1}{b_{\mathrm{a}}}+\frac{1}{b_{\mathrm{c}}}\right)}\left(\frac{\mathrm{d}i}{\mathrm{d}U}\right)_{i=0} = B\left(\frac{\mathrm{d}i}{\mathrm{d}U}\right)_{i=0} \quad (11)$$

where i_{corr} = corrosion current density in $\mathrm{A/cm^2}$; b_{a}, b_{c} = constants (gradients of the Tafel lines); and $(\mathrm{d}i/\mathrm{d}U)_{i=0}$ = gradient of the current density – potential curve at the rest potential in $\mathrm{A\,cm^{-2}\,V^{-1}}$.

This procedure is useful when quantitative determination based on Tafel lines is no longer possible; for example, when the corrosion products form deposits. The polarization resistance is measured at the rest potential (free corrosion potential) within a range of ≤ 20 mV. The constant B is often ca. 20 mV and can be determined for exact measurements.

An example of determining the corrosion rate by this method is shown in Figure 10 [6]. However, for very small corrosion currents that are independent of potential in some ranges (e.g., in passive metals), this method cannot be used to determine the corrosion rate. Nor is the method applicable when non-transfer-related part reactions occur, e.g., when the corrosion current is dependent on diffusion processes at the specimen surface. Also, the constant B can be time-dependent when dense protective films form, which further restricts the applicability of the method [7].

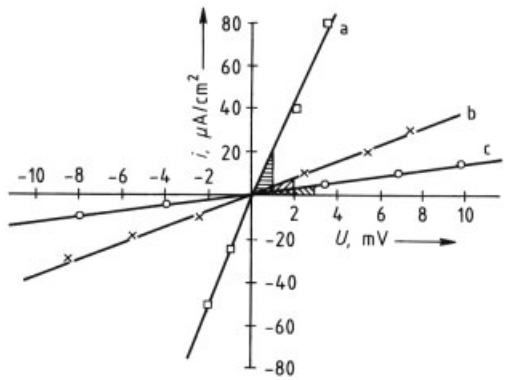

Figure 10. Determination of the corrosion rate i_{corr} from polarization resistances

$$i_{\mathrm{corr}} = 0.02\,\mathrm{V}\cdot\frac{\Delta i[\mathrm{A/cm^2}]}{\Delta U[\mathrm{V}]}$$

a) $t = 80\,^{\circ}\mathrm{C}$, $i_{\mathrm{corr}} = 0.48\ \mathrm{mA/cm^2}$; b) $t = 60\,^{\circ}\mathrm{C}$, $i_{\mathrm{corr}} = 0.09\ \mathrm{mA/cm^2}$; c) $t = 40\,^{\circ}\mathrm{C}$, $i_{\mathrm{corr}} = 0.023\ \mathrm{mA/cm^2}$

For local corrosion phenomena, determination of the corrosion rate is of secondary importance. In fact, in processes limited to specific potential ranges, the threshold potential (i.e., the potential above which local corrosion begins), which depends on both the corrosive medium and the alloy composition, is more important.

2.4. Hydrogen Overpotential and Acid Corrosion

The cathodic reduction of hydrogen ions

$$2\,H^+ + 2\,e^- \rightleftharpoons H_2$$

takes place in a number of complex steps. Under certain conditions, acid corrosion can induce hydrogen embrittlement when special effects influence the kinetics of hydrogen reduction, which takes place up to the stage of molecular hydrogen formation (recombination). In many metals, considerable overpotential is needed to initiate hydrogen evolution. Furthermore, the activity of metal surfaces can be impaired by contamination with impurities. This principle is applied in deliberate inhibition (e.g., use of pickling inhibitors, see Section 5.2). A slower recombination rate generally leads to a higher concentration of adsorbed hydrogen at the surface, which in turn can result in increased diffusion of atomic hydrogen into the metal. This may eventually lead to embrittlement in various metals and alloys.

In aqueous solution, the hydrogen ion from the dissociation of water ($H_2O \rightleftarrows H^+ + OH^-$) or acids (e.g., $HCl \rightleftarrows H^+ + Cl^-$) is hydrated, forming a hydronium ion:

$$H^+ + H_2O \longrightarrow H_3O^+$$

The hydronium ion (H^+_{aq}) must first diffuse through the liquid to the metal surface

$$H^+_{aq} \longrightarrow (H^+_{aq})_{metalsurface}$$

For discharge of the hydronium ion to occur, an electron must pass through the metal–electrolyte phase boundary; the discharge reaction (Volmer reaction) is therefore known as a transfer reaction:

$$(H^+_{aq})_{metalsurface} + e^- \longrightarrow H_{ad}$$

Initially, the hydrogen remains adsorbed in atomic form on the metal surface. For the formation of molecular hydrogen, there are two possibilities: either two adsorbed hydrogen atoms combine chemically to form an adsorbed hydrogen molecule (Tafel reaction):

$$H_{ad} + H_{ad} = (H_2)_{ad}$$

or an electrochemical reaction takes place between an adsorbed hydrogen atom, a hydronium ion, and an electron (Heyrovský reaction):

$$H_{ad} + (H^+_{aq})_{metalsurface} + e^- \longrightarrow (H_2)_{ad}$$

Thus there are two possibilities for the formation of adsorbed molecular hydrogen: two Volmer reactions and one Tafel reaction, or one Volmer reaction and one Heyrovský reaction (Fig. 11).

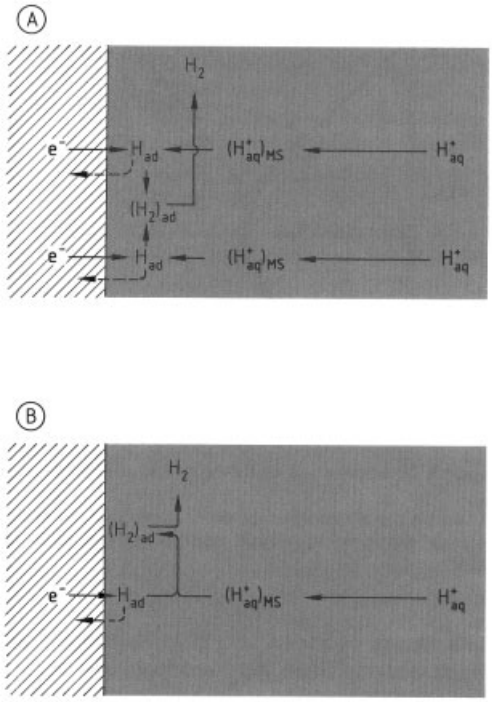

Figure 11. Reaction schematic of cathodic hydrogen formation
A) By the Volmer–Tafel mechanism; B) By the Volmer–Heyrovský mechanism (after KAESCHE, see general references)
Hatched area = Metal; shaded area = Boundary layer solution
$(H^+_{aq})_{MS}$ = Hydronium ion H_3O^+ on metal surface

The formation of adsorbed molecular hydrogen is followed by the transition of H_2 molecules into the gaseous phase:

$$(H_2)_{ad} \longrightarrow H_2$$

This reaction sometimes consists of several distinguishable individual steps. It is occasionally observed with metals which dissolve in acid,

for example, that the hydrogen bubbles always occur at a few spots where bubble formation is presumably energetically favorable. In this case, the reaction consists of diffusion to those spots with subsequent desorption. For very slow hydrogen evolution, bubble formation is unnecessary because the hydrogen molecules can dissolve rapidly enough in the aqueous phase, diffuse through it to the liquid surface, and be desorbed into the ambient gaseous phase. The transition of atomic hydrogen into the metallic phase is discussed in Section 3.3.1.5.

The considerable overpotential for hydrogen evolution observed in many metals (in Hg and Pb, > 0.5 V) is due to inhibition in one of the part reactions. The slowest partial process determines the rate of the overall reaction. There is practically no inhibition in platinum, which possesses the lowest hydrogen overpotential. Other "active" metals are palladium, strontium, and rhodium.

2.5. Oxygen Overpotential and Oxygen Corrosion

Acid corrosion arises only rarely in pure form because it requires the corrosive solution to be completely free of air. However, this is not usually the case, and in acid solutions an additional electrode process often occurs:

$$O_2 + 4\,e^- + 2\,H_2O \rightleftharpoons 4\,OH^-$$

An important feature of a corrosion process is the reduction of the dissolved oxygen by formation of hydroxyl ions (OH^-). Since the equilibrium potential is linked through the concentration of the OH^- ions with the concentration of the hydrogen ions (or H_3O^+ ions) by the ionic product of water

$$(a_{H^+}) \cdot (a_{OH^-}) = K_w = 10^{-14} \quad (\text{at } 25\,°C)\,,$$

the equilibrium potential is pH dependent. The standard potential is ca. 0.4 V at 25 °C, a partial pressure of 100 kPa, and an OH^- ion activity of 1 (pH 14). Equilibrium potentials under other conditions can be calculated by the Nernst equation:

$$U_H = U_H^0 - \frac{RT}{F} \ln a_{OH^-} + \frac{RT}{4F} \ln p_{O_2} \qquad (12)$$

Interpretation of the equation for $p_{O_2} = 100$ kPa is shown in Figure 12, in which, for comparison, the equilibrium potential of the hydrogen electrode for $p_{H_2} = 100$ kPa is also given. It can be seen that the potential lines run parallel, and that at all pH values the equilibrium potential of the oxygen electrode is more than 1.2 V higher than that of the hydrogen electrode. This explains why the presence of oxygen often promotes electrochemical corrosion. Quite apart from the fact that a higher cathode potential normally increases the potential difference and hence the current density in local elements, anodic destruction processes are now possible at potentials that are above that of the hydrogen cathode and which cannot occur in the absence of oxygen. An example is the behavior of copper, which is only attacked by electrolytes when oxygen is present. On a bare copper surface, the loss of metal is directly proportional to the quantity of cathode-converted oxygen. In practice, however, there is usually fairly rapid formation of oxidic protective films (consisting at room temperature mostly of Cu_2O and at higher temperatures mainly of CuO), which restrict corrosion to a harmless extent. Only when the formation of these films is inhibited can the oxygen corrosion of copper progress linearly as a function of time.

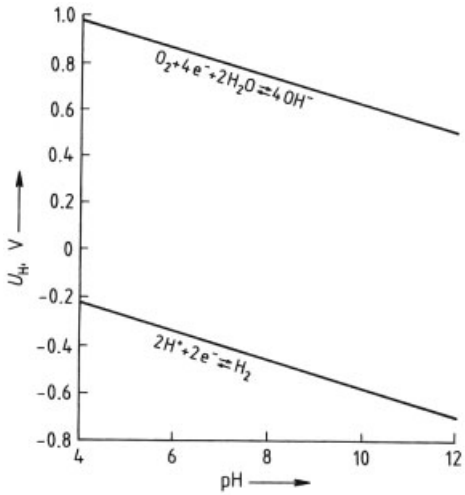

Figure 12. Equilibrium potentials for the oxygen and hydrogen electrodes

The working potential of the oxygen electrode generally differs greatly from its equilibrium potential. This leads to a considerable over-

potential, which in most cases is a concentration overpotential. Oxygen is consumed at the cathode, leading to depletion of oxygen in the cathode boundary layer. If the cathodic process is to continue, additional oxygen must be supplied from the electrolyte. The amount of oxygen converted electrochemically – and, therefore, the corrosion rate – depends on the rate of subsequent oxygen supply, which is diffusion controlled and is determined by:

1) The diffusion coefficient, which increases with increasing temperature:
2) The difference between the oxygen concentration at the metal surface and in the corrosive medium, which increases with increasing oxygen content of the solution:
3) The thickness of the boundary layer, which depends on whether the flow is laminar or turbulent, with smaller thicknesses being associated with higher flow rates.

The quantity of oxygen diffusing to the cathode per unit time and area is directly proportional to the diffusion coefficient and the difference in oxygen concentration, and inversely proportional to the thickness of the boundary layer. Therefore, an increase in temperature and a higher concentration of oxygen in the electrolyte increase the corrosion current density, whereas an increase in the boundary layer thickness reduces the current density.

The occurrence of a concentration overpotential in oxygen corrosion is clearly shown by the cathodic overpotential curve. As demonstrated in Figure 13, a diffusion-limited current density appears on the curve, which is indicated by the parallelism of the cathodic current curve and the potential axis (diffusion current density). In this region the current density can only be further increased if the concentration of available oxygen is increased. The figure also shows the behavior of copper towards corrosive solutions containing oxygen.

If the corrosion potentials in metals (mixed potentials) are within the range of the diffusion-limited current, the corrosion current density is independent of the potential. This leads to the conclusion that normally the rate of oxygen corrosion is determined solely by oxygen diffusion to the cathode and that other factors are only of minor influence. In particular, the pH of the solution in this instance does not have the importance often attached to it in practice.

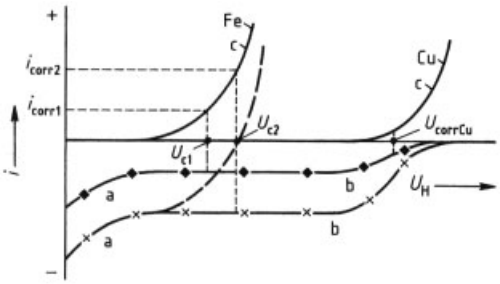

Figure 13. Current – potential curve for oxygen corrosion
—— Partial current curve; – – – Current – potential curve at oxygen concentration 2; –♦–♦– Oxygen concentration 1; –×–×– Oxygen concentration 2
a) $2\,H^+ + 2\,e^- \rightarrow H_2$;
b) $O_2 + 2\,H_2O + 4\,e^- \rightarrow 4\,OH^-$;
c) $M \rightarrow M^{2+} + 2\,e^-$

The following equation shows that areas in which a mainly cathodic electrode process prevails tend to alkalize through the accumulation of hydroxyl ions:

$$O_2 + 4\,e^- + 2\,H_2O \longrightarrow 4\,OH^-$$

If phenolphthalein is used as an indicator for hydroxyl ions and potassium hexacyanoferrate(III) as an indicator for iron(II) ions, the anode and cathode areas can be made visible in the case of corroding iron. If, for example, an iron nail is placed in a gelatine solution mixed with these two indicators, after a while the head and point of the nail start to turn blue because of the formation of Prussian blue, whereas the middle of the nail turns red. This experiment shows that it is mainly the anodic process that takes place at the mechanically worked points, whereas the cathodic process with the formation of hydroxyl ions occurs in the middle of the nail.

This separation of the two distinct processes in oxygen corrosion can also be observed when a ring of rust forms under a saltwater droplet on an iron surface some distance from the outer edge of the droplet and not at the point of minimum oxygen supply. The dissolved iron ions then combine as they diffuse with hydroxyl ions to form sparingly soluble iron hydroxide, which turns to rust on absorption of oxygen, and deposits (Fig. 14).

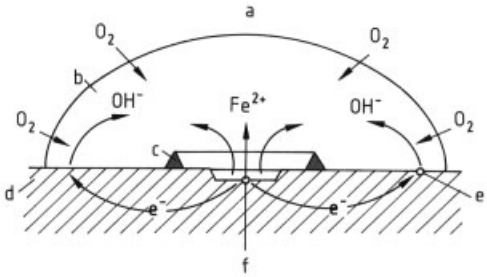

Figure 14. Mechanism of corrosion under a saltwater droplet
a) Air; b) Drop of salt solution; c) Rust ring; d) Iron; e) Cathodic oxygen reduction $1/2\,O_2 + H_2O + 2\,e^- \rightarrow 2\,OH^-$; f) Anodic metal dissolution $Fe \rightarrow Fe^{2+} + 2\,e^-$

The process can be written as follows:

$$Fe \rightarrow Fe^{2+} + 2\,e^-$$
$$2\,Fe^{2+} + O \longrightarrow 2\,Fe^{3+} + O^{2-}$$
$$O + 2\,e^- \longrightarrow O^{2-}$$
$$O^{2-} + H_2O \longrightarrow 2\,OH^-$$
$$Fe^{2+} + 2\,OH^- \longrightarrow Fe(OH)_2$$
$$2\,Fe(OH)_2 + O + H_2O \longrightarrow 2\,Fe(OH)_3$$
$$Fe(OH)_3 \longrightarrow FeO(OH) + H_2O$$

Using a simple test arrangement, EVANS showed that different metal surface aeration in oxygen corrosion results in the formation of local elements (aeration elements) [8]. An improved model experiment is illustrated in Figure 15.

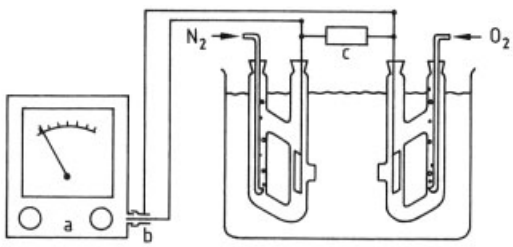

Figure 15. Basic circuit diagram of an aeration cell
a) Millivoltmeter; b) Input with inner conductor and screen; c) Short circuit resistance

Although the nitrogen-purged iron electrode (anode) is corroded, no corrosion is found on the oxygen-purged iron electrode (cathode). The actual effect consists of inhibition of the anodic process through the formation of OH^- ions at the O_2 electrode. Consequently, the term aeration element is somewhat misleading, in that the element formation in this corrosion process is the result of a change in pH. This can be readily proven since element formation does not occur in sufficiently buffered solutions.

2.6. Protective Film Formation and Passivity

The corrosion reactions described above require an active, i.e., mainly bare, metal surface. In practice, however, many metals tend to form protective films which can have a major influence on the electrolytic process. Dense films provide protection and are thus referred to as protective coatings or films. A protective iron oxide film is known to form, for example, in boiler pipes. In contrast, porous rust layers provide no protection. The good resistance of aluminum, which is the result of an Al_2O_3 film forming in air, can also be produced artificially by anodic oxidation.

In addition to these marcroscopic protective films, a number of metals possess optically invisible, pore-free oxide films with particularly good protective action; this gives rise to the passivity of metals. The occurrence of passivity can best be observed electrochemically. Figure 16 (see next page) shows an anodic metal dissolution curve with the typical drop in current at the onset of passivity. The metal initially dissolves actively in the solution. With increasing potential the dissolution current rises until a potential (passivation potential, often called Flade potential U_{Fl}) is reached at which there is a sudden sharp drop in metal dissolution. The very low metal dissolution remains practically constant over a relatively large potential range, the passive range, before rising again at the beginning of the so-called transpassive range. To achieve the passive state, the redox system (as Fig. 16 shows) must provide a cathodic current that is at least as large as the anodic current needed to form the passive film at the passivation potential. A pore-free oxide film of ca. 10 nm thickness then forms at the electrode surface, suppressing the previous intense metal dissolution.

The minimal residual anodic current at the electrode in the passive range is caused by the slow dissolution and constant reformation of the passive film. The passivation potential and the dissolution in the passive range differ for each metal or alloy. Of economic significance are metals with a moderate passivation potential and

minimum dissolution in the passive state. Figure 17 shows current – potential curves of iron and chromium in sulfuric acid. The passivation of chromium begins more than 500 mV before that of iron and requires a considerably lower current density. The good passivation capability of chromium can be imparted to iron by alloying with > 12 wt % chromium. This forms the basis for the development of all stainless and acid-resistant steels. Further improvements can be obtained by adding nickel, molybdenum, and copper to these alloys.

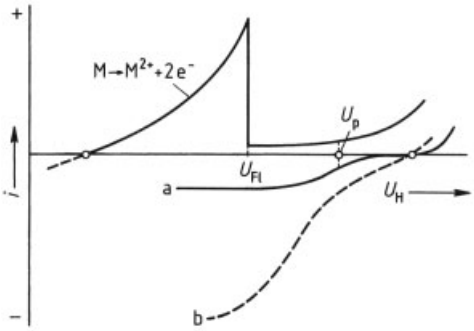

Figure 16. Current – potential curve for metal dissolution with drop in current due to passivation
a) Redox system 1: conservation of passivity; b) Redox system 2: generation of passivity
U_P = Rest potential in the passive state; U_{Fl} = Flade potential

Under certain circumstances the passivation capacity of various alloys results in special corrosion phenomena. These arise mainly because the passive film exhibits local defects as a result of external influences (special corrosive media) or structure-related effects (deposits, grain boundary disturbances, segregation, etc.). This leads to localized corrosion processes, including selective corrosion, intergranular corrosion, and pitting corrosion. Under the influence of static mechanical tensile stress, various types of stress corrosion cracking can arise, while alternating mechanical stress leads to corrosion fatigue, which occurs in both fully active and passive alloy systems.

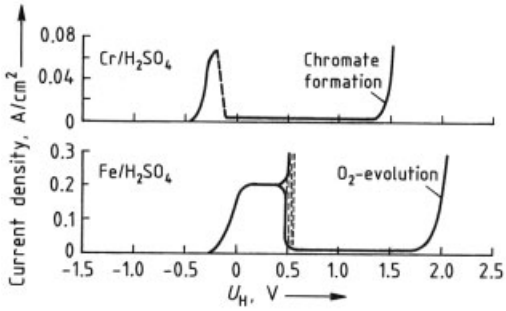

Figure 17. Passive behavior of chromium and iron in sulfuric acid

Figure 18 shows schematically the different types of corrosion together with the possible combinations which can arise in practice. Corrosion fatigue is not considered here because mechanistically it is not confined to passive alloys. It is therefore treated separately (Section 3.3.2).

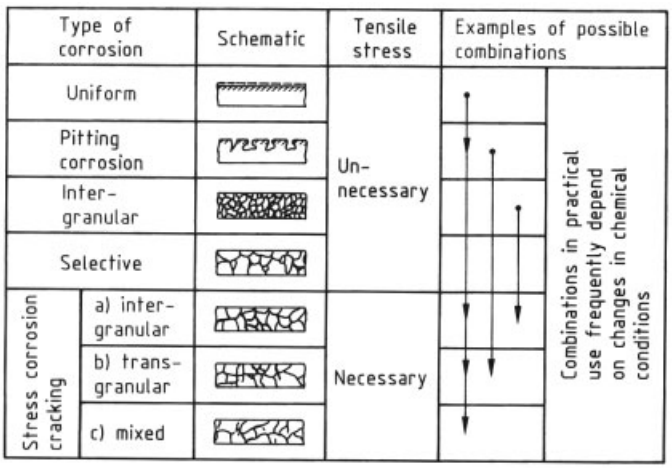

Figure 18. Types of corrosion

3. Types of Electrolytic Corrosion

The surface of a material or component is subject to various environmental influences when in use. The mechanical, thermal, and chemical (corrosive) parameters may act individually, or two or even three factors can combine, the significance of each differing according to the type of stress. The corrosive factor, as shown in Figure 19, can be divided into chemical and electrochemical components. Electrochemical corrosion can take place with or without mechanical stress. The types of electrochemical corrosion that occur are also shown in Figure 19 and are discussed in more detail in the following sections.

3.1. Uniform and Nonuniform Surface Corrosion

DIN 50 900, Part 1 [2] defines uniform surface corrosion as corrosion with practically equal mass loss over the entire surface (homogeneous mixed electrode), and nonuniform corrosion (shallow pit formation) as corrosion with locally different mass losses. The cause of nonuniform corrosion is the presence of corrosion elements (heterogeneous mixed electrode) [3]. The common feature of both types of corrosion is the geometry of the damaged area. The surface spread of these areas is generally greater than the depth. Accordingly, compared with localized types of corrosion such as pitting, these types of corrosion are limited in extent.

Uniform corrosion is without doubt the easiest form of corrosion to recognize and to control. In the ideal case, material loss runs parallel to the surface of the corroded component, the reduction in wall thickness being given by:

$$\Delta M_{\mathrm{tot}} / \varrho \cdot A$$

where ΔM_{tot} = total mass of the dissolved metal, A = area of corroded surface, and ϱ = density of the metal. In terms of the corrosion current densities, this implies the equality

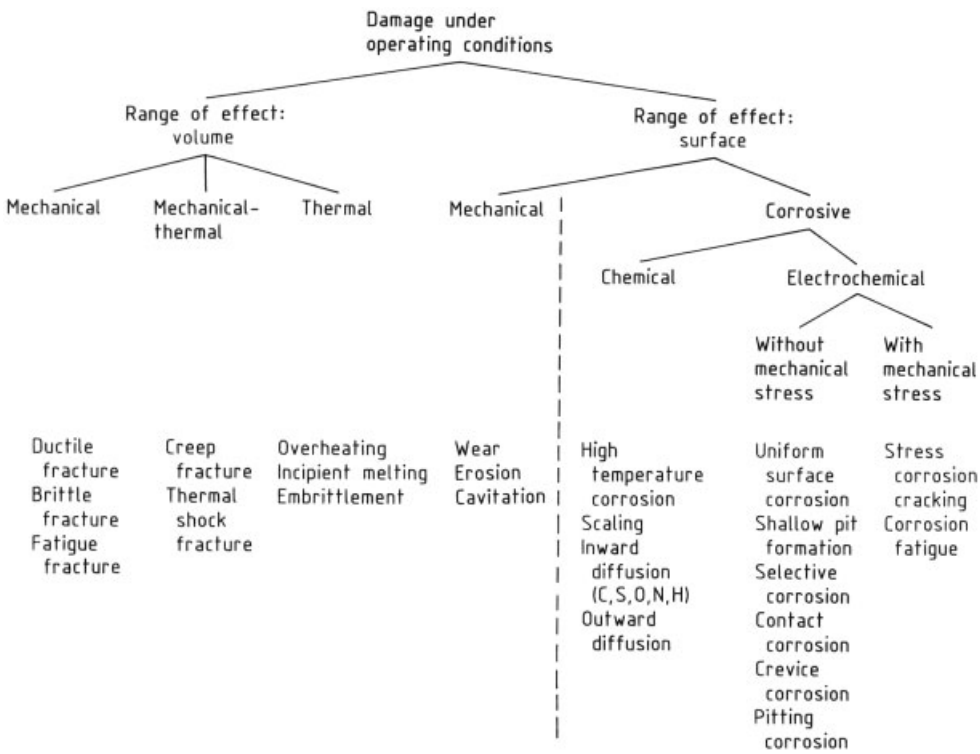

Figure 19. Overview of damage to metallic materials under operating conditions

of the real local corrosion current density I_R and the mean corrosion current density i_R, ($I_R = i_R$), where the following relationship exists [9]:

$$i_R = \frac{1}{A} \int I_R \cdot dA$$

A practical example of almost uniform surface corrosion is as follows [10]: a pipe made of unalloyed steel St 35 (material no. 1.0308) used at ca. 90 °C for the transport of service water showed material erosion of the inner surface after three years in operation; a layer of corrosion product (mainly iron oxide) had formed on the inside with a practically constant thickness over the entire area. Given the operating conditions, the material corrosion had to be due to oxygen corrosion.

A further example is illustrated in Figure 20. A thermometer protection tube made of the highly corrosion resistant nickel alloy NiMo 30 (material no. 2.4810) was immersed at ca. 20 °C in an aqueous, acidic chloride solution to which sodium nitrite was added, releasing nitrogen oxides. The protection tube, consisting of a larger-diameter tube and a smaller-diameter tube extension welded together, suffered severe corrosion with the exception of the weld joint. A chromium-containing filler metal of the type nickel alloy NiMo 16 Cr (material no. 2.4812) had been used. The uniform surface corrosion of the chromium-free alloy without passivation capability NiMo 30 is due to the presence of oxidants. Under these conditions, the weld containing chromium is considerably more resistant.

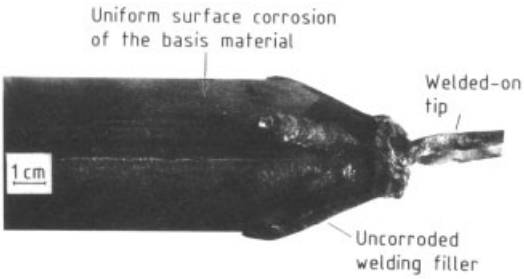

Figure 20. Uniform corrosion of a thermometer protection tube made of NiMo 30 (material no. 2.4810)
Welding filler: NiMo 16 Cr (material no. 2.4812)

In practice, uniform surface corrosion is easy to control in the majority of cases, especially because the reduction in wall thickness during operation can be monitored by nondestructive testing or by suspending a specimen of the material in the corrosive agent and monitoring the corrosion [11]. Suitable measures to combat uniform corrosion include greater wall thicknesses, a more resistant material, and cladding or protective coatings.

In the case of nonuniform corrosion (shallow pit formation), the real corrosion current density I_R and mean corrosion current density i_R are not equal. Accordingly, nonuniform corrosion is characterized by a locally differing corrosion rate. The conditions for this may occur, for example, during the course of corrosion through a nonhomogeneous covering of the reaction product on the material.

3.2. Forms of Localized Corrosion without Applied Mechanical Stress

This general term covers pitting corrosion, crevice corrosion, and selective corrosion, including the subgroups of intergranular corrosion, spongiosis, dezincification, and line- and layer-type corrosion.

3.2.1. Pitting Corrosion

In DIN 50 900, Part 1 [2], pitting in the broadest sense is defined as the form of corrosion in which craterlike, surface-hollowing, or pinpricklike cavities occur (see Fig. 21). Outside the pitted areas practically no surface corrosion is observed. The depth of the pitted area is generally equal to or greater than its diameter. Pitting always arises when the material surface is covered with a corrosion-resistant coating that has pores or defects, the latter being there initially or arising through a local reaction with the corrosive medium. The following are regarded as being protective layers:

1) A natural oxide scale or rolling scale in unalloyed or low-alloy steels;
2) Protective films (e.g., paint, bitumen coatings, or oil films);
3) Metallic coatings (e.g., zinc);
4) Salt layers formed under operating conditions, such as the lead sulfate layers that form on lead in sulfuric acid, or the copper(I) oxide or malachite layer formed on copper in drinking water;

5) A film produced by the adsorption of inhibitors (e.g., chromates);
6) The passive film, usually consisting of only few molecular layers, such as found in iron, nickel, chromium, and iron- and nickel-based alloys with passivation capacity.

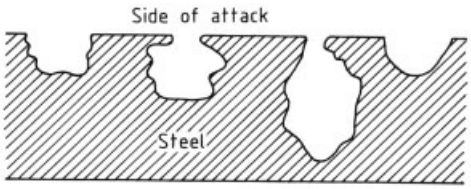

Figure 21. Shapes of pits in pitting corrosion

For pitting to occur in metals coated with protective layers as in (1) to (3), there must be defects in the protective coating and an electrolyte solution acting as a corrosive medium. The resulting corrosion element consists of the cathode (intact coating) and the anode (pores in the coating). The attack on the base material in the region of the pores acts downwards. Depth progression depends on the size and activity of the cathode surface. Pitting in cases (4) and (5) in corrosive media is due to an incomplete protective layer on the material surface. The intensity of corrosion depends on the area ratio between anode (uncoated surface) and cathode (coated surface).

Pitting on materials with a passive coating, such as unalloyed or low-alloy steels, stainless and acid-resistant ferritic chromium steels, and stainless and acid-resistant austenitic chromium – nickel steels, is caused by electrolytes that contain chloride, bromide, and, in some cases, iodide ions. Since chloride ions in particular are responsible for this kind of corrosion, it is often called *chloride corrosion*. Halide ions (excluding fluoride, which does not cause pitting) are adsorbed at discrete locations of the passive coating (formation of cavity nuclei). After an induction period the passive coating is penetrated and pitting corrosion initiated.

An example of localized corrosion in the form of pits in natural protective coatings is the corrosion that arises through cracks in the rolling scale of unalloyed or low-alloy steels. If, for example, the rolling scale of a steel plate has spalled at a tiny point, and the sheet is immersed in a salt solution containing oxygen, an electric current flows between the oxide film (the cathode) and the bare metal (the anode). At the anode, iron dissolves as iron(II) ions, whereas at the cathode oxygen is reduced:

$$1/2\,O_2 + H_2O + 2\,e^- \longrightarrow 2\,OH^-$$

This oxygen reduction, which determines the flow of current and hence the corrosion of the iron at the uncoated spot, continues as long as oxygen can diffuse unhindered to the cathode (see Section 2.5). This local dissolution process can even perforate sheet and piping. This kind of corrosion can be avoided by completely removing the rolling scale.

Similarly, corrosion processes take place at pores and defects in electrically conducting coatings that are more noble than the base material. In paints, the propagation rate of cavities depends on the possibility of substance transport through the coating.

Pitting Corrosion in Stainless Steels. The conditions that lead to chloride-induced pitting corrosion in stainless steels in chemical plants are well-known and have been extensively examined for this group of materials [12].

Corrosion Parameters. Figure 22 (see next page) shows the current density – potential curve of a material with passivation capacity in an initially chloride-free corrosive medium. Starting from the free corrosion potential U_{corr}, the steel passes, with increasing anode polarization after exceeding the passivation potential, through the passive range and reaches the breakthrough potential at U_{b}. While in the active range, the alloying elements go into solution in their lowest valencies, during corrosion above U_{b}, iron(III) and chromium(VI) compounds are formed. If chlorides are now added to the corrosive medium, an increase in current is observed at U_{p}, this being due to pitting corrosion. Chlorides (and also bromides and iodides) thus decrease the passive range of the material.

The pitting potential shifts to more negative values with increasing chloride ion concentration, rising temperature, and decreasing pH; i.e., susceptibility to pitting increases. On the other hand, the susceptibility to pitting in a given corrosive medium can be decreased by higher flow rates of the medium.

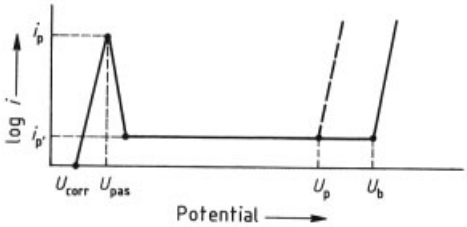

Figure 22. Current density – potential curve of a passifiable alloy, with and without pitting corrosion
U_{corr} = Corrosion potential; U_{pas} = Passivation potential; U_p = Pitting potential; U_b = breakthrough potential; i_p = Passivation current density; $i_{p'}$ = Passive current density

The dependence of the pitting potential U_p on various factors has been explained for a number of austenitic chromium – nickel – (molybdenum) steels by using diffusion-annealed specimens [13]. It was demonstrated, for example, that the pitting potential U_p is directly proportional to the logarithm of chloride ion activity. This is shown in Figure 23 for a nitrogen-alloyed steel with high molybdenum content (X 3 CrNiMoN 17 13 5, material no. 1.4439). Note that pitting occurs, albeit at high potentials, even at chloride ion concentrations of 2 ppm. A temperature change results in a parallel shift of the characteristic curve. The temperature dependence of the pitting potential is shown in Figure 24, in which U_p is plotted against the reciprocal of absolute temperature. Figure 25 shows that an increase in the pH of the corrosive medium shifts the pitting potential U_p to more positive values, to the extent of 55 – 60 mV per decade of the hydrogen ion concentration. This value corresponds to the factor $2.3\,RT/nF$ for $n = 1$ at 20 °C in Equation (10), which describes the dependence of the overpotential on the concentration of the potential-determining ion. A change in hydrogen ion activity by a factor of 10 makes $\Delta U = 58$ mV (see Section 2.2.3):

$$\Delta U = (2.3RT/nF)\log a$$

Pitting in austenitic chromium – nickel steels can be avoided in many cases by lowering the chloride ion concentration, reducing the temperature, and raising the pH of the corrosive medium as high as possible. If these measures cannot be taken, or cannot be taken to an adequate degree, other materials must be chosen.

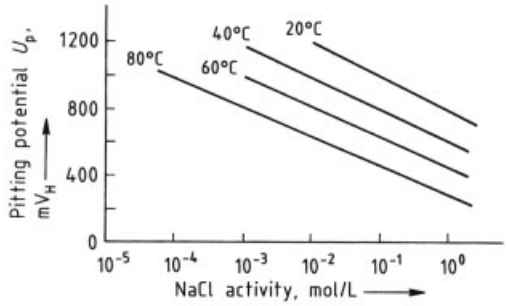

Figure 23. Pitting potential of diffusion-annealed X 3 CrNiMoN 17 13 5 (material no. 1.4439) in NaCl solution as a function of chloride ion activity and temperature

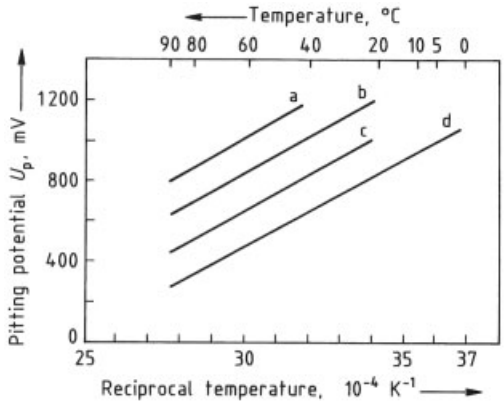

Figure 24. Pitting potentials U_p of diffusion-annealed X 3 CrNiMoN 17 13 5 (material no. 1.4439) in NaCl solutions as a function of temperature
a) 0.001 N NaCl solution; b) 0.01 N; c) 0.1 N; d) 1 N

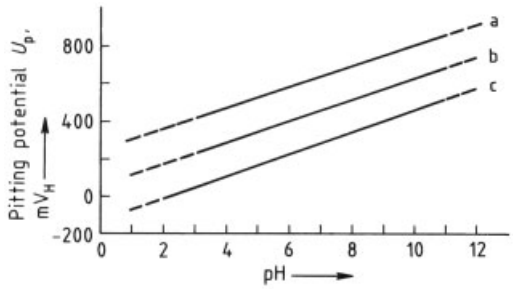

Figure 25. Pitting potentials U_p of diffusion-annealed X 3 CrNiMoN 17 13 5 (material no. 1.4439) in NaCl solutions as a function of temperature
a) 1 N, $t = 40$ °C; b) 0.1 N, $t = 80$ °C; c) 1 N, $t = 80$ °C

Material Parameters. For a given steel, the susceptibility to pitting decreases with increasing surface finishing. A ground surface is more

susceptible to pitting than a pickled one, which in turn is more susceptible than a polished surface. A further decisive point is structural uniformity. Rolled surfaces (i.e., surfaces parallel to the rolling direction) are less prone to pitting than short-transverse-cut sections exhibiting open microsegregations. Oxide and sulfide inclusions, carbide precipitates, and intermetallic phases also increase susceptibility to pitting. Cold forming can also cause a marked increase in the incidence of pitting, although the influence of cold forming has not yet been finally established and contrary results have also been found.

Effect of Alloying Elements. For a standard chromium – nickel steel (e.g., X 6 CrNiTi 18 10, material no. 1.4541) with 17.0 – 19.0 wt % chromium and 9.0 – 11.5 wt % nickel, the resistance to pitting can be improved by increasing the contents of chromium and molybdenum, but is not affected by higher nickel concentrations. In stainless steels, 3 % chromium is equivalent to ca. 1 % molybdenum. This finding led to the formulation of a *pitting resistance equivalent* (PRE), which is defined as [14]:

$$\mathrm{PRE} = \mathrm{wt\%chromium} + 3 \times \mathrm{wt\%molybdenum}$$

Some investigations have suggested a factor of 3.3 for molybdenum [15]. Figure 26 shows that U_p is directly proportional to the PRE. Therefore, austenitic steels with the highest possible chromium and molybdenum contents exhibit extremely high pitting resistance.

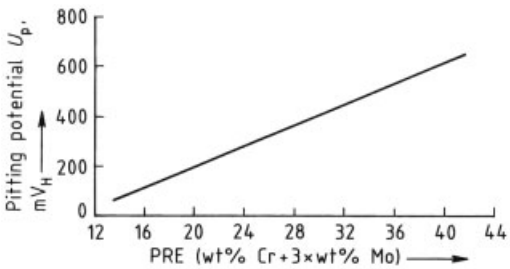

Figure 26. Pitting potential U_p of austenitic chromium – nickel – molybdenum steels in a NaCl solution as a function of the pitting resistance equivalent (PRE)

The positive effect of nitrogen on the resistance to pitting corrosion of stainless austenitic steels has recently been studied in greater detail. Examinations of austenitic materials with different Cr, Mo, and N contents indicated a PRE factor of ca. 13 for nitrogen [16]; other authors give values of 20 or even 30 [17]. However, the effectiveness of nitrogen is clearly linked to specific Cr and Mo percentages. To what extent nitrogen exercises its own pitting-inhibiting effect or whether its structure-stabilizing influence is the cause has yet to be clarified.

The high-alloy, austenitic stainless steels X 2 CrNiMoN 17 13 5 (material no. 1.4439) [18], X 1 CrNiMoN 25 25 2 (material no. 1.4465) [19], [20] and X 1 CrNiMoN 25 22 2 (material no. 1.4466) [21], which have PREs of ca. 31, are widely used in chemical engineering [22–26]. Even higher PREs are exhibited by the materials X 1 NiCrMoCu 25 20 5 (material no. 1.4539) [20], [27], X 1 NiCrMoCu 31 27 4 (material no. 1.4563) [20], X 2 CrNiMoCuN 20 18 6 [20], [28], and X 1 NiCrMoCuN 25 20 6 (material no. 1.4529) [20], [29]. These high-alloy materials have also proved successful in a wide variety of applications, such as chemical apparatus, flue gas desulfurization, and marine technology [25], [30–35].

Factors affecting pitting corrosion in austenitic chromium – nickel – (molybdenum) steels are listed in Table 2. An example of pitting corrosion is given in Figure 27, which shows a section of a forged T-piece made of X 5 CrNiMoTi 25 25 (material no. 1.4577) into which a thermocouple protection tube was inserted. In the gap between the steel wall and the protection tube, a sublimable, chloride-containing salt accumulated causing considerable pitting corrosion. The problem was solved by using the same material under changed plant conditions.

Table 2. Parameters affecting the pitting corrosion of austenitic chromium – nickel steels

Favoring pitting	Hindering pitting
Material parameters	
Rough surface	polished surface
Inhomogeneous structure	homogeneous structure
Cold forming	
Corrodent parameters	
High chloride ion concentration	low chloride ion concentration
High temperature	low temperature
Low pH value	higher pH value
Stationary attacking agent	higher flow rate
Alloying parameters	
Low chromium contents	high chromium contents
Low molybdenum contents	high molybdenum contents

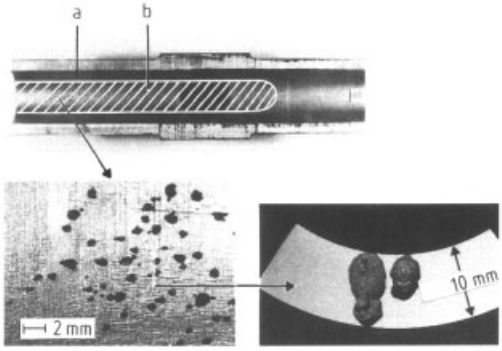

Figure 27. Pitting corrosion on a high-pressure T-piece of material no. 1.4577 at 180 °C and 13 MPa as the result of accumulation of chloride in the gap between the pipe (a) and the thermometer protection tube (b)

Pitting corrosion is not restricted to stainless steels; other passive metals tend to undergo pitting in solutions containing halides. These include, for example, aluminum and aluminum alloys, nickel alloys with passivation capacity such as NiMo16Cr16Ti, and titanium. Nickel-based alloys are far less prone to corrosion attack by chlorides than corrosion-resistant Cr–Ni austenitic steels and are thus seldom at risk. Aluminum alloys show some susceptibility and have pitting potentials that are determined by the chloride ion activity and temperature but not by the pH. Because of the low electron conductivity of the oxide scale of aluminum, however, the corrosion rate remains low if it is not enhanced by contact corrosion with metals such as copper, iron, nickel, and their alloys, or with electrically conducting deposits with rest potentials more positive than the pitting potential of aluminum (intensification of the cathode part reaction) [36]. Under extreme conditions (e.g., in a solution with 65 % $ZnCl_2$, 5 % HCl, and 1 % $FeCl_3$ at 100 °C) [37], titanium is prone to pitting corrosion. So far pitting has not been reported for tantalum and niobium.

3.2.2. Crevice Corrosion

Crevice corrosion is also a form of localized corrosion. It amounts to locally intensified corrosion in crevices and thus results exclusively from the design of a specific component [38], [39]. In unalloyed steels it usually occurs as nonuniform or uniform corrosion, and in stainless steels it can also occur as pitting corrosion. It can be regarded as a variation of pitting corrosion.

Although the mechanism of crevice corrosion has not yet been fully explained, the electrolyte in the crevice contains less oxygen than the electrolyte outside the crevice because diffusion is impeded. The vast majority of crevice corrosion phenomena can thus be explained by the formation of an aeration element (see Section 2.5). Through the lack of oxygen in the crevice, anodic regions form in which metal ions go into solution. As a result, the conditions in the crevice electrolyte usually change, this being towards greater corrosiveness (e.g., lower pH through hydrolysis). The time dependence of oxygen depletion as a function of the crack width and depth is shown in Figure 28 [40].

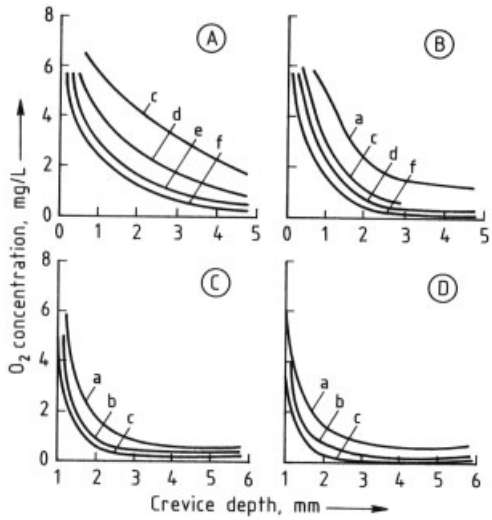

Figure 28. Dependence of oxygen depletion in crevice corrosion on crack width and depth
Crack width: A) 1.65 mm; B) 0.74 mm; C) 0.40 mm; D) 0.15 mm
Time: a) 6 min; b) 15 min; c) 30 min; d) 1 h; e) 1.5 h; f) 2 h

Whether or not crevice corrosion occurs largely depends on the crevice geometry. Crevice widths are particularly critical when the distance between the crevice-forming surfaces is less than 1 mm. Since the reactions taking place in the crevice cannot be influenced from the outside, damage due to crevice corrosion can only be avoided through the prevention of crevices.

In ferritic and austenitic stainless steels, crevice corrosion is almost always initiated by

local activation. This can be induced in a crevice by oxidant depletion, if necessary supplemented by halides. The passivity then breaks down. The access of oxidants to the material surface, and hence the passivity, may also be hindered by local deposits.

Numerous incidences of damage in crevices in components are due to crevice corrosion; for example, in the critical areas between the tubes and tube sheets in heat exchangers. Other commonly found defects are crevices caused by incomplete welding, especially of root passes, which can lead to crevice corrosion under critical corrosion conditions (Fig. 29). Typical damage due to crevice corrosion on a plate heat exchanger made of X 5 CrNiMo 17 13 (material no. 1.4449) is illustrated in Figure 30. The medium was an almost saturated sodium chloride solution containing small amounts of free chlorine. The product temperature was 50–70 °C. The first plates were already porous after five weeks at the points where the grooves touch those of the neighboring sheet. The damage is due to heavy local corrosion in the region of the points of contact (crevices). Remarkably, pipes operating with the same product and made of the same material, but which do not present the conditions for crevice corrosion, have been in use for more than ten years with negligible corrosion.

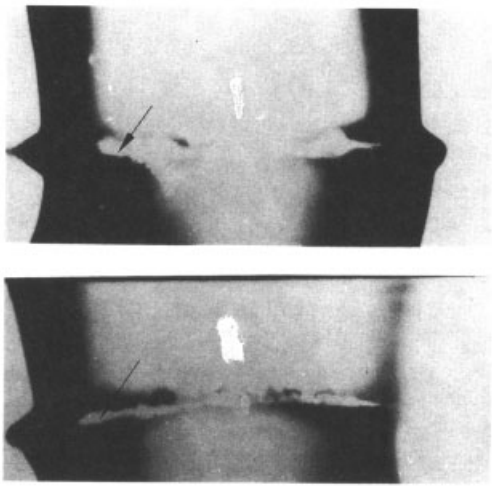

Figure 29. Crevice corrosion of a pipe (material no. 1.4571) (radiograph)

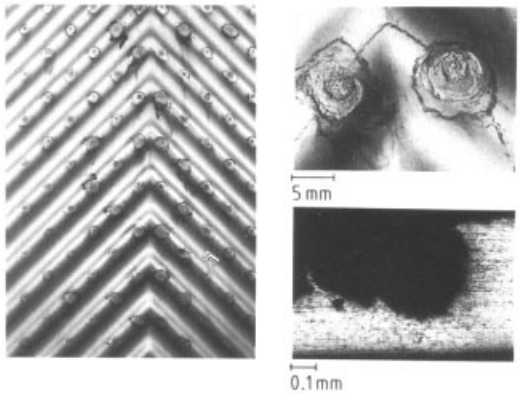

Figure 30. Crevice corrosion of a plate heat exchanger (material no. 1.4449)

The conditions for crevice corrosion are not only present when the crevice is formed by metallic materials, but also when the surface of a passive material is coated with nonconductors such as seals, packings, and product deposits, under which the material can then be prone to pitting corrosion. Corrosive media containing chloride, which can neutralize passivity, are particularly dangerous in this respect.

An example is a horizontal copper pipe for drinking water, in which typical pitting corrosion damage has formed under deposits. The first perforations appeared after eight-months' use. The deposits consisted primarily of residual solder paste and iron oxides and were found solely in the bottom half of the pipe. After removing the deposited layers by pickling, numerous pitted spots could be seen, mainly under pimples in the protective coating. Since the cause of damage can only be improper handling of solder paste during soldering, this kind of damage, which was frequently found some years ago in the Federal Republic of Germany, ceased to exist after the changeover to brazing. The corrosion-inducing effect of iron oxides is obviously due to their impairing the formation of a dense copper(I) oxide layer. Iron oxides do not cause corrosion damage when they come into contact with previously formed, dense protective layers.

3.2.3. Selective Corrosion

Selective corrosion is defined as corrosion in which specific microstructural constituents, re-

gions close to grain boundaries, or alloying components are preferentially attacked [2].

Intergranular corrosion, spongiosis, dezincification, and line- and layer-type corrosion are classified in this subgroup. Of particular importance is intergranular corrosion.

3.2.3.1. Intergranular Corrosion

Intergranular corrosion in electrolyte solutions arises when the resistance of zones close to the grain boundary, compared with the resistance of areas further from the grain boundary, is reduced due to microstructural changes in the grain boundary regions. This type of corrosion is particularly characteristic of passive materials, in which large differences between the corrosion rates of the matrix and the grain boundary areas are possible if the chemical composition of the grain boundary zones is changed by discontinuous precipitation processes in the supersaturated solid solution. Intergranular corrosion is characterized by the fact that grain boundary zones are corroded preferentially in a groovelike manner until individual grains are loosened and fall out of the microstructure. Given heavy intergranular corrosion, the entire structure can collapse [41–43].

The reduced resistance of the grain boundary regions is due to the following:

1) Depletion of alloying elements that are important for the maintenance of the corrosion resistance in the grain boundary regions through precipitation of phases containing these elements along the grain boundaries;
2) Formation of microstructural components at the grain boundaries that are preferentially attacked in a corrosive medium;
3) Accumulation at the grain boundaries of impurities which stimulate anodic dissolution.

Of these causes, depletion of alloying elements has the greatest practical significance. This occurs when grain boundary precipitates are formed that contain one or more alloying elements in a markedly higher concentration than the matrix.

Precipitation processes of this kind are always caused by heat treatments, such as sensitizing annealing, that are inappropriate for the alloy in question. For the austenitic chromium – nickel – (molybdenum) steels used for the fabrication of chemical plant equipment, the critical temperature range is 400 – 800 °C. Chromium depletion through formation of chromium-rich carbides, mostly of the type $M_{23}C_6$, is the main cause of intergranular corrosion in these steels. The precipitation of chromium nitrides – of importance is only the chromium-rich nitride Cr_2N [44] – can initiate intergranular corrosion, especially in ferritic steels. Since the intermetallic phases in stainless steels contain appreciably less chromium than carbides and nitrides and their deposition is far slower, the chromium depletion related to these phases is minimal.

Molybdenum depletion at grain boundaries through carbides rich in molybdenum and through intermetallic phases can also lead to intergranular corrosion. For standard molybdenum steels, molybdenum depletion is always linked to a far greater chromium depletion and therefore plays only a minor role. In special high-alloy molybdenum-containing steels, and especially in nickel-based alloys with very high molybdenum contents, molybdenum depletion is a major cause of susceptibility to intergranular corrosion. This is particularly significant in the chromium-free alloy NiMo 30 (material no. 2.4810).

In NiMo 30, molybdenum-rich carbides and intermetallic phases of the type Ni_4Mo and Ni_3Mo precipitate readily [45]. Figure 31 shows intergranular corrosion in a welded pipe. In the meantime, NiMo 30 has been superseded

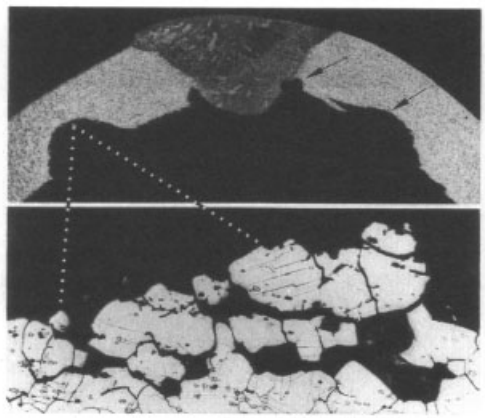

Figure 31. Intergranular corrosion in the weld heat-affected zone of a pipe (made of NiMo 30) caused by hydrochloric acid

by NiMo 28 (material no. 2.4617), which resists the formation of grain boundary precipitates in the weld heat-affected zone.

In the high-silicon austenitic chromium–nickel special steel X 1 CrNiSi 18 15 4, which has high resistance to concentrated nitric acid [46–48], susceptibility to intergranular corrosion depends on the separation of a carbide $M_{23}C_6$ and on a π-phase of type $M_{11}(C,N)_2$, containing chromium, nickel, silicon, carbon, and nitrogen [49].

The susceptibility to intergranular corrosion of a chemically resistant steel depends on the carbon content, the duration of sensitizing, the sensitizing temperature, and the temperature of prior solution annealing; it is represented in a time – temperature diagram. Curves of this type are shown in Figure 32 for an austenitic steel with ca. 18 % chromium, 8 % nickel, and various carbon contents [50]. The curves encompass the respective area of susceptibility.

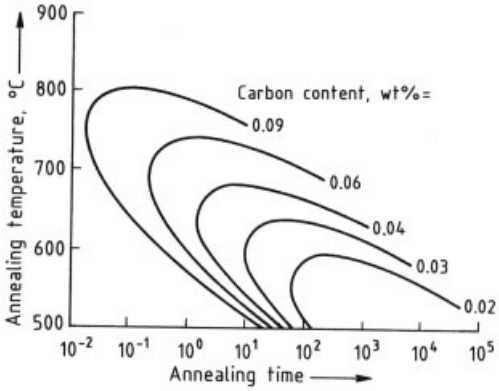

Figure 32. Time – temperature sensitization curves of an unstabilized austentic steel (18 % Cr, 8 % Ni) as a function of carbon content

The effect of the solution annealing temperature is especially pronounced in stabilized steels (i.e., those that contain titanium and niobium for bonding of the carbon). With increasing solution annealing temperature, the curves encompassing the area of susceptibility shift to lower sensitizing temperatures and longer annealing times, as shown in Figure 33 [51]. The reason for this shift is that with increasing solution annealing temperature, the amount of carbon which is in solution and, therefore, available for precipitation in the critical temperature range, also increases. At the same time, grain coarsening occurs, reducing the grain boundary surfaces available as precipitation sites. This also leads to greater chromium depletion because the increased amount of available carbon on the grain surfaces can then bind more chromium. However, the considerably lower susceptibility of stabilized steels to grain disintegration, compared with unstabilized steels, is due to the formation of carbides, which leads to a lower concentration of dissolved carbon. Since the solubility of the carbides (TiC and NbC) decreases with decreasing temperature, proneness to intergranular corrosion may be eliminated by annealing (stabilization) these steels at a lower temperature (900 – 950 °C) to remove dissolved carbon by precipitation of the carbides. Another possibility of lowering the susceptibility to grain disintegration is to reduce the carbon content to below 0.03 % (ELC grades).

The carbon solubility is further reduced by a higher nickel content, thus increasing the tendency to precipitate. This shifts the grain disintegration areas to shorter tempering times and higher temperatures, as demonstrated in Fig-

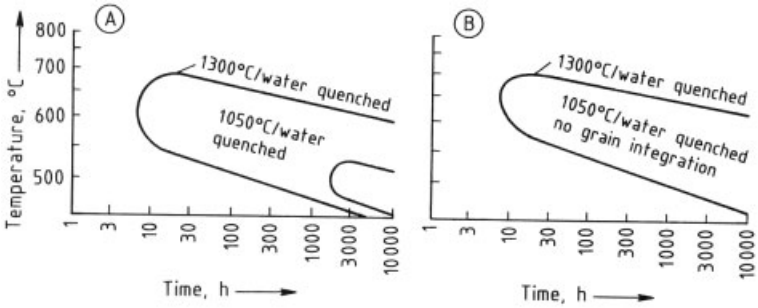

Figure 33. Time – temperature sensitization curves of stabilized steels as a function of solution annealing and sensitizing temperatures (test criterion: Strauss test)
A) Material no. 1.4541; B) Material no. 1.4550

ure 34, which compares the grain disintegration curves of X 2 CrNi 19 11 (material no. 1.4306, 11 % Ni), the titanium-stabilized grade X 10 NiCrAlTi 32 20 (material no. 1.4876, 32 % Ni), and the alloy NiCr 15 Fe (> 72 % Ni). The carbon content of the materials is ca. 0.02 % in each case. Higher nitrogen contents also shift the beginning of precipitation to shorter times (Fig. 35). In practice, however, grain disintegration is only expected above ca. 0.2 % N [44].

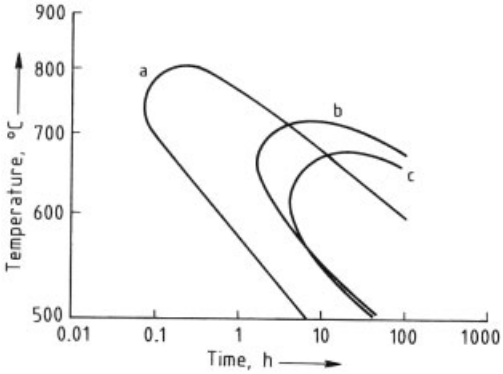

Figure 34. Time – temperature sensitization curves of austentic steels
a) NiCr 15 Fe, 0.017 % C; b) X 10 NiCrAlTi 32 20, 0.021 % C, 0.35 % Ti; c) X 2 CrNi 19 11, 0.020 % C

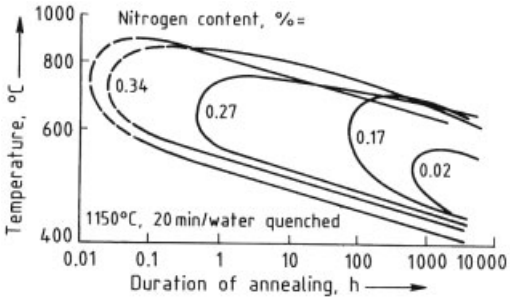

Figure 35. Time – temperature sensitization curves of unstabilized, nitrogen-containing, austenitic steels with ca. 18 % Cr, 10 % Ni, and 0.01 % C

Of the ferritic steels, those containing ca. 17 wt % chromium are of industrial significance. Final heat treatment is performed at 750 – 850 °C with subsequent air or water cooling. In unstabilized grades the carbon is practically all precipitated as $M_{23}C_6$, without the formation of chromium-depleted zones. As the annealing temperature rises, the carbon solubility increases, the carbides dissolve completely or partially, and reprecipitate on cooling as $M_{23}C_6$ at the grain boundaries, giving rise to chromium-depleted grain boundary zones if the cooling rate is sufficiently high. In contrast to austenitic steels, rapid cooling from a high temperature is critical for ferritic steels; the higher mobility of the atoms in the body-centered cubic lattice causes a high precipitation rate of the carbides, and rapid post-diffusion of chromium into the depleted zones occurs. Ferritic steels with 17 % Cr can therefore suffer intergranular corrosion in the high-temperature zone immediately adjacent to the weld if this is cooled so rapidly after welding that chromium depletion is not eliminated; an example is shown in Figure 36. Ferritic chromium steels can be stabilized by addition of titanium or niobium, but compared with the austenitic standard steels, a comparatively higher concentration of these elements is necessary [52]. So-called superferrites also exist, such as X 1 CrNiMoNb 28 4 2 (material no. 1.4575), characterized by C + N contents ≤ 0.040 %, which greatly improves the weldability of these materials.

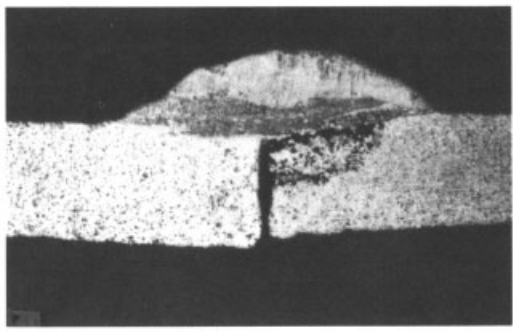

Figure 36. Intergranular corrosion in the weld heat-affected zone of a 17 % Cr steel

A special form of intergranular corrosion is observed in titanium-stabilized stainless steels containing ca. 17 wt % chromium: it is caused by precipitation of titanium carbonitrides at the grain boundaries. The carbonitrides are dissolved by strongly oxidizing solutions such as nitric acid, leaving grooves (Fig. 37). In such cases preference should be given to the niobium-stabilized grade since NbC is hardly attacked by oxidizing acids [53].

Marked susceptibility to grain disintegration can also be observed in the highly corrosionresistant nickel-based alloys, such as

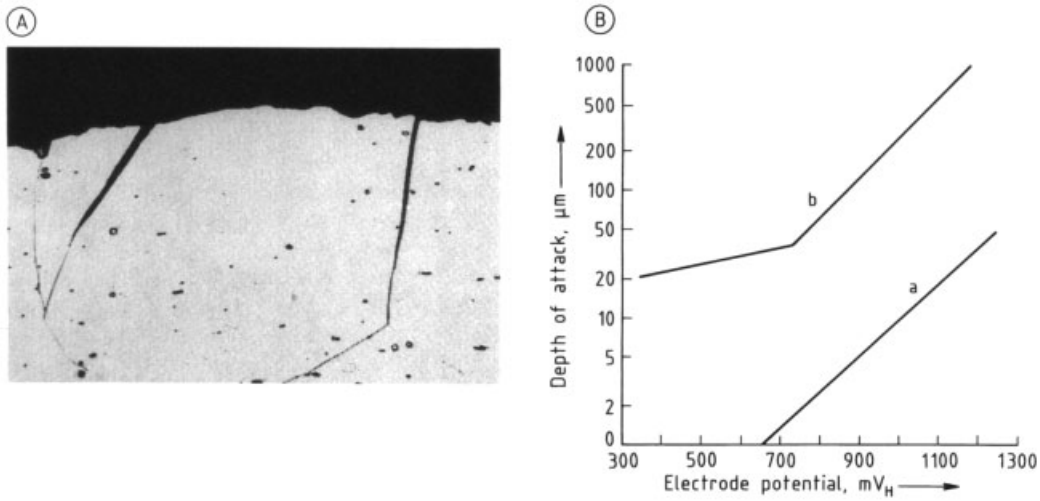

Figure 37. Intergranular corrosion of a Ti-stabilized 17 % Cr steel (X 7 CrTi 17) in boiling sulfuric acid
A) Mode of attack; heat treatment 30 min, 1300 °C, water quenched; test time 24 h; $U_H = 650$ mV
B) Depth of attack as a function of electrode potential; heat treatment: a) 30 min, 850 °C, water quenched; b) As a) + 30 min, 1300 °C, water quenched

the chromium- and molybdenum-containing type NiMo 16 Cr (material no. 2.4812) (Fig. 38), which is, however, no longer in use. In addition to carbides, grain boundary precipitates in this material also consist of an intermetallic chromium- and molybdenum-rich intermetallic phase, which leads to depletion of chromium and molybdenum in the grain boundary zones. The critical temperature range for precipitation is very wide (500 – 1150 °C) [45]. More recent types of this family of alloys, such as NiMo16Cr16Ti (material no. 2.4610) and NiCr21Mo14W (material no. 2.4602), exhibit a considerably reduced susceptibility due to their lower carbon and silicon contents (both < 0.1 %); thus, there is no further risk of intergranular corrosion in practical applications [54–58].

Intergranular corrosion can also occur in other metals and alloys in connection with grain boundary precipitation, for example, in AlMg5 alloy, NiCr30Fe, or titanium (in methanol with traces of water and HCl). Solution-annealed austenitic chromium – nickel steels can suffer intergranular corrosion in strong oxidants such as chromium-containing, azeotropic, or superazeotropic nitric acid, even though no precipitation can be detected. It is assumed that phosphorus contents above 0.01 wt % and silicon contents above 0.1 wt % are responsible for this form of corrosion [41].

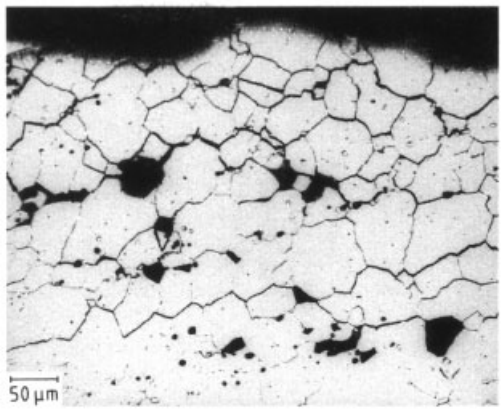

Figure 38. Intergranular corrosion of NiMo 16 Cr as a result of improper heat treatment

3.2.3.2. Spongiosis

Spongiosis, also known as graphitization, is a form of corrosion found only in gray cast iron. Because of inadequate formation of a protective layer, ferrite and pearlite are converted primarily to iron(II) oxide hydrate, whereas the graphite structure and the distribution and percentage of the phosphide eutectic remain unchanged [2], [59]. In this instance the metallic bond is lost, leading to complete loss of strength in the damaged areas of the component (Fig. 39). The re-

markable feature of this process is that the outer form of the component is so well maintained that tool marks can often be seen.

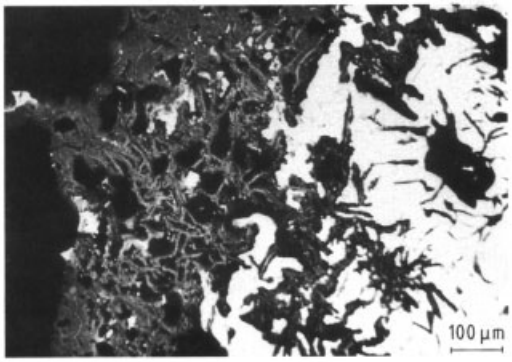

Figure 39. Spongiosis of a cast iron pump used for well water
The light areas in the picture are ferrite, the dark areas are graphite.

Spongiosis occurs primarily in cast iron pipes laid in acid earth or in calcareous loamy soil. It can be avoided by applying adhesive, crack-free protective coatings, such as those based on tar. Damage caused by spongiosis is also observed in cast-iron feed-water preheaters and pumps for drinking water and seawater [59].

When ferritic or pearlitic cast iron comes into contact with pure water or aqueous salt solutions, spongiosis occurs. It begins without exception in the immediate vicinity of the graphite, giving rise to a porous coating consisting of graphite, carbides, sulfides, phosphides, and nitrides. A sufficiently high oxygen content in the corrosive medium then leads to a greater deposition of corrosion products on this primary coating, which thereby becomes sealed, finally forming a protective coating [60]. The advance of spongiosis is therefore to be expected when the corrosive medium has a low oxygen concentration.

3.2.3.3. Dezincification

Dezincification is the selective corrosion of copper – zinc alloys (brass) with more than 15 % zinc, in which porous copper deposits form on the corroded parts [2], [61]. Although less common, similar phenomena have also been observed in other copper alloys such as copper – aluminum (aluminum stripping), copper – tin (detinning), and copper – nickel (nickel stripping) [62–65].

Dezincification can occur, for example, in tap water, river water, seawater, and liquids containing chloride. The rate of this corrosion process is increased by plastic deformation of the material [66]. Dezincification can occur by two mechanisms:

1) going into solution of both alloy constituents (copper and zinc) through dissolution of the solid solution and cementation of the more noble component (copper);
2) selective dissolution of the less noble constituent from the solid solution and accumulation of the exposed atoms through surface diffusion.

It is assumed that the two mechanisms work either simultaneously or one after the other [67], [68].

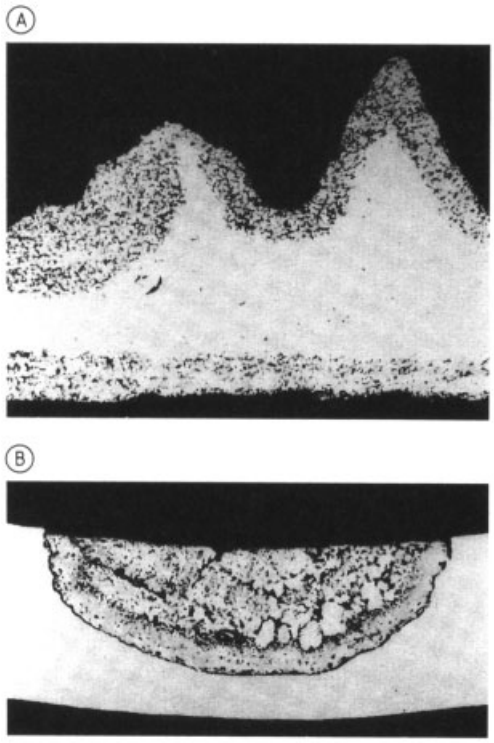

Figure 40. Types of dezincification
A) Layer dezincification of a brass water faucet (GK-CuZn 37Pb); B) Plug-type dezincification in a brass pipe (CuZn 28 Sn 1)

According to their appearance, a distinction is made between layer dezincification and plug-type dezincification (Fig. 40, see previous page). In layer dezincification the entire surface is corroded practically uniformly. This form of corrosion is observed when a component made of a copper – zinc alloy comes into contact with nobler metals in an electrolyte (e.g., seawater). Crevices and cracks such as threads and surface damage promote corrosion [61].

Far more critical is plug-type dezincification, which mainly arises in heat exchanger tubes upon deposition of sand grains, oil residues, rust, mussels, and algae with formation of aeration elements. Even thick tubes can be perforated in this way [61], [69]. Dezincification can be impeded by adding small amounts of arsenic or phosphorus.

3.2.3.4. Line- and Layer-Type Corrosion

In these forms of selective corrosion the attack occurs practically parallel to the rolling direction. Layer corrosion is observed in the aluminum alloys AlMg, AlZnMg, AlCuMg, and AlZnMgCu. The material undergoes corrosion in chloride-containing media by exfoliation parallel to the surface. This occurs if a rolling process gives rise to a laminar structure of intermetallic compounds or when lamellar zones of differing concentration were already present in the ingot [70], [71]. Figure 41 shows layer corrosion in an aluminum alloy [71].

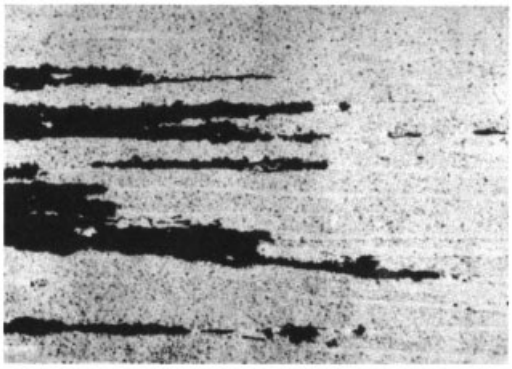

Figure 41. Layer corrosion of an Al – Zn – Mg alloy

In contrast to layer corrosion in aluminum alloys, line-type corrosion can occur in many materials, although it is also linked to forged and rolled products. The cause of line-type corrosion may be due to both the banded structure of precipitations and microsegregation, i.e., insufficient concentration compensation.

Because of the high degree of deformation, sheet products exhibit a particularly marked sandwich structure with regard to the concentration distribution of the alloying elements. If the corrosion resistance of a material depends on the concentration of the segregating alloying element, local differences in corrosion depth can occur if fabricated parts are subjected to stress in the fiber direction. In practice this is the case at the face of rolled-in tubes of heat exchangers, the cutting edges of sieve trays, or valves and valve seats in forged armatures [43], [72–76].

A good example is provided by support rigs of tube bundles made of material no. 1.4571, which after several months' exposure in a hot,

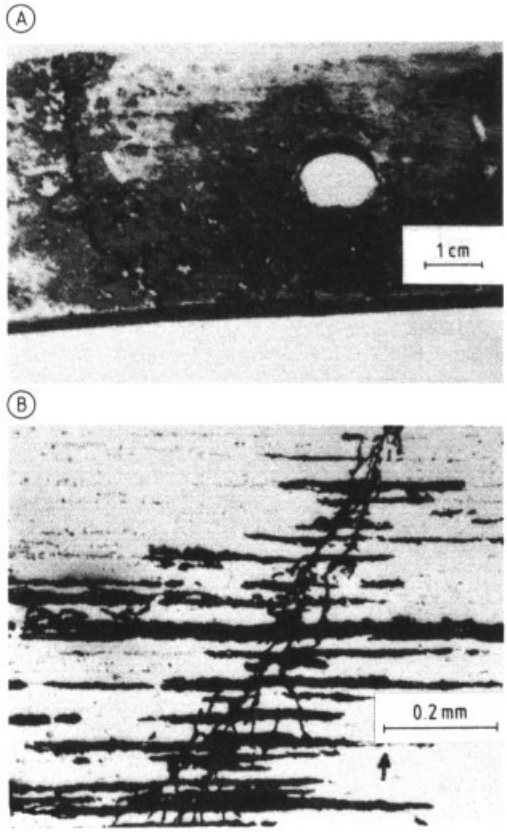

Figure 42. Transgranular stress corrosion cracking of a steam coil support
A) Macroscopic view; B) Microsection showing selective corrosion along the ferrite lines (→)

weakly acidic, chloride-containing suspension had fractured because of transgranular stress corrosion cracking (see Section 3.3.1.2). One detail of the fractured fixtures, with many cracks, on which clear signs of pitting corrosion can be seen, is shown in Figure 42 (see previous page). Metallographic examination (see Fig. 42B) showed that the transgranular cracks produced selective corrosion, which spread along the ferrite lines. At the point marked by an arrow in Figure 42B it can be seen that the extended ferrite islands were not dissolved; instead, the corrosion spread along the segregated austenite rims [43].

Figure 43 shows a nozzle head made of NiMo 30 (material no. 2.4810), which was manufactured from a nonuniformly annealed bar. After 200 h in operation, signs of corrosion that look like pitting appear. Figure 43B shows line- or band-like precipitation in the fiber direction which is responsible for the pitting-like attack.

Figure 44 shows selective corrosion caused by segregation. The SEM photographs of the

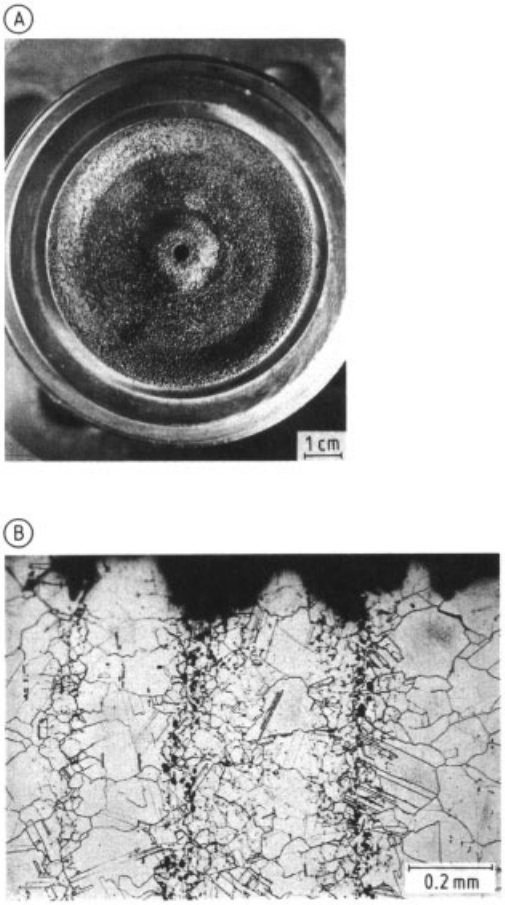

Figure 43. Selective corrosion of a NiMo30 (material no. 2.4810) nozzle head in a region exhibiting lines of precipitates
A) Nozzle head; B) Microstructure of nozzle head

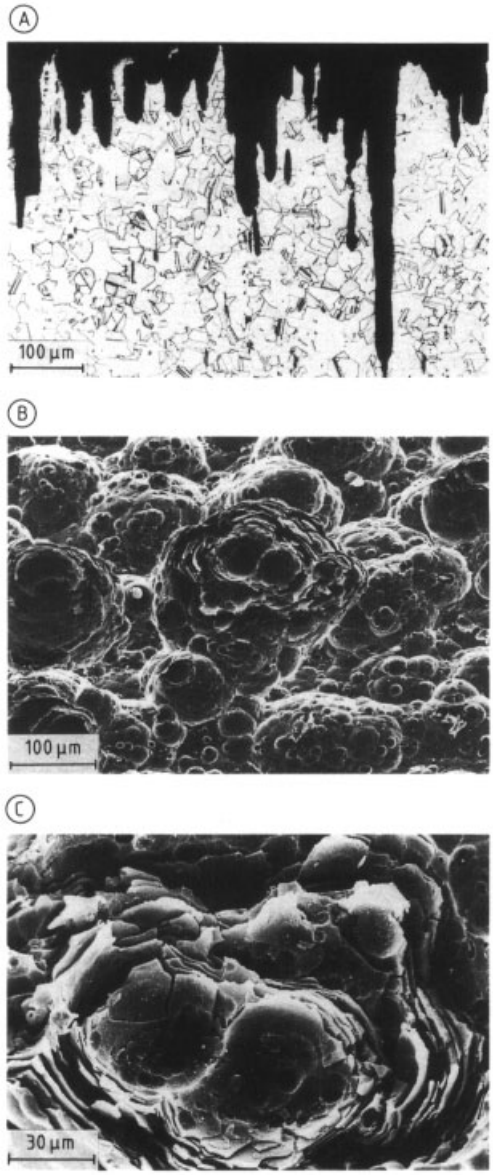

Figure 44. Selective corrosion caused by segregation on the face of a heat exchanger tube in boiling, highly concentrated HNO_3. Material: X 1 CrNiSi 18 15 4
A) Longitudinal section; B) and C) Scanning electron micrographs of the attacked face

face of a heat exchanger tube made of X 1 CrNiSi 18 15 4 which was subjected to boiling, highly concentrated nitric acid for 60 d showed evidence of line-type corrosion. The sides of the corrosion crevices are coated with several layers of the corrosion product SiO_2 [10], [49].

Selective corrosion caused by differences in concentration can be reduced by homogenizing heat treatment or by laying one weld pass over the edges of sheets and tubes.

3.3. Forms of Local Corrosion under Mechanical Stress

3.3.1. Stress Corrosion Cracking

Stress corrosion cracking is one of the most unpleasant forms of corrosion because it usually occurs unexpectedly and can very quickly lead to cracking in equipment such as reactors, vessels, and piping, and to fractures in components of all kinds. Depending on the alloy system and corrosive medium, the characteristics are inter- or transgranular brittle cracking, branching out to varying degrees into the material and eventually causing fracture of the remaining cross section. General surface corrosion arising in this way is either nonexistent or immeasurably small. Deposited corrosion products are seldom found.

Normal anodic stress corrosion cracking is caused by a combination of mechanical tensile stress and local electrolyte dissolution processes when certain conditions are met. Firstly, the corrosive medium must have a specific effect on the respective alloy, and in addition the alloy in contact with the electrolyte in this material/corrosive medium system must be prone to stress corrosion cracking. The tensile stress must also be sufficiently high. Susceptible systems, for example, are stainless austenitic steels in chloride-containing solutions or unalloyed and low-alloy steels in nitrate solutions. In contrast, unalloyed and low-alloy steels are not susceptible to stress corrosion cracking in chloride solutions.

Protective layers are a prerequisite for stress corrosion cracking to occur on the surface of a component under the influence of an electrolyte. Local crack-inducing corrosion processes are caused by selective degradation of these protective or passive layers. In this instance, the corrosive medium can cause local perforation of the layer (e.g., media containing chlorides with passive, stainless austenitic CrNi steels), or its passivity can be impaired by the influence of the prevailing tensile stress to the extent that only a locally incomplete protective layer is formed (e.g., solutions of alkali hydroxides or nitrates with unalloyed steels). However, this simplified view of the formation of local anodes in stress corrosion cracking merely explains the initial stage during the induction period; it sheds no light on the actual cracking period, for which there are now several interpretations.

The most common interpretation of the mechanism of cracking is based on a "periodic electrochemical – mechanical process". This suggests that cracking is an alternating sequence of relatively slow anodic dissolution in the crack base and sudden mechanical crack propagation. In some alloys intermittent cracking has actually been found, but in many other cases no evidence of stepwise cracking has been produced.

A further theory on the cracking mechanism is that it is a purely electrochemical process in the base of the crack (active path corrosion), the rapid crack propagation rate observed in most cases being due to greatly increased dissolution of the material as a result of plastic deformation at the crack tip with simultaneous passivation of the crack walls. Greatly increased dissolution during the plastic deformation of a metal has been demonstrated [78].

A distinction is made between anodic and hydrogen-induced cathodic stress corrosion cracking, in which atomic hydrogen generated from a cathodic process in a corrosion reaction penetrates the metal lattice and, in combination with internal and external stresses, causes transgranular cracking and in heat-treated steels preferentially intergranular cracking. This kind of stress corrosion cracking thus belongs to the category of hydrogen damage or embrittlement (see Section 3.3.1.5).

A more recent theory on the mechanism of anodic stress corrosion cracking, based partly on tests on steels, combines the electrochemical process of local metal dissolution with hydrogen embrittlement at the base of the crack caused by the atomic hydrogen forming during corrosion, which may be of major significance to crack propagation.

In addition to classic stress corrosion cracking, systems are known in which the static tensile stress is not sufficient to create cracking (e.g., carbonate – bicarbonate solutions). In such cases, critical slow strain rates are necessary to initiate stress corrosion cracking. This type of corrosion is called strain-induced or nonclassical stress corrosion [79–82].

3.3.1.1. Unalloyed and Low-Alloy Steels

In unalloyed and low-alloy steels, the corrosive media (alkali, nitrates) produce passive protective layers which have a certain instability at high-angle grain boundaries and are thus decomposed under sufficient stress, initiating local corrosion. The resulting stress corrosion cracking is therefore an intergranular process in these steels (Fig. 45). It is assumed that the special activity of the grain boundaries is determined by accumulation of alloy phases. This theory is supported by a number of test results. It has been found, for example, that cementite and nitrides or carbonitrides at the grain boundaries in unalloyed steels produce a high susceptibility to stress corrosion cracking [83].

The susceptibility to stress corrosion cracking of a steel with 0.011 wt % carbon subject to different annealing conditions, measured under constant tensile stress in boiling $Ca(NO_3)_2$ solution, is shown in Table 3 [83]. The marked improvement when the carbon is kept in solution is clearly evident. Elements promoting passivity such as chromium and molybdenum clearly reduce the tendency to cracking, although in alloyed steels of this kind, the type of heat treatment also plays a major role [84].

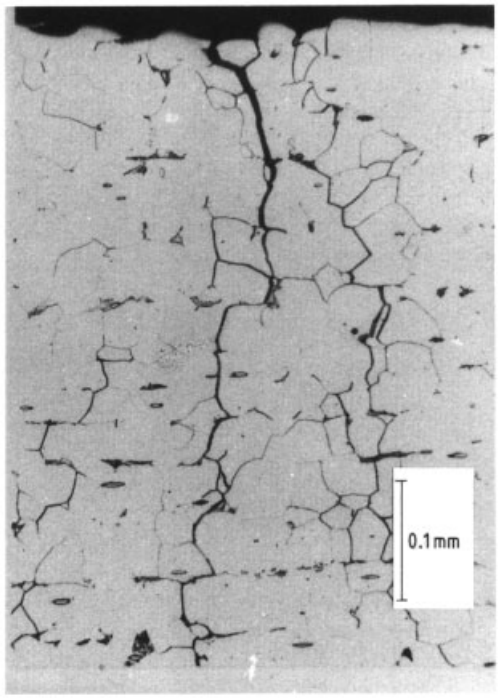

Figure 45. Intergranular stress corrosion cracking of Thomas steel (normalized) in a nitrogen-purged 20 wt % NaOH solution at 110 °C

On the basis of these observations the induction phase in intergranular cracking can be interpreted as the preferential dissolution of inhomogeneous grain boundaries with formation of deep cavities with corresponding stress peaks, which initiate the actual cracking. Basically, a lower stress limit for the induction of intergranular stress corrosion cracking in a given corrosive medium exists for each steel and for each structural state resulting from annealing, below which stress corrosion cracking cannot be initiated.

Table 3. Influence of grain interface cementite on the stress corrosion behavior of a mild, unalloyed steel [a]

Heat treatment	Time to fracture in minutes at a tensile stress of			
	$0.3 \cdot \sigma_u$ [b]	$0.4 \cdot \sigma_u$	$0.5 \cdot \sigma_u$	$0.6 \cdot \sigma_u$
Normalized (920 °C, 20 min, air)	288	30	104	48
Normalized + 720 °C, 24 h, liq. N_2 (carbon held in solution)	> 15 000	> 15 000	132	54
Normalized + 730 °C, 2 h, water + 500 °C, 4 h, water (tertiary cementite precipitation)	96	66	18	12

[a] Analysis of the steel: C 0.011 %, Si 0.02 %, Mn < 0.005 %, P 0.007 %, Si < 0.005 %, Ni 0.02 %; N, Al, Cr, and Mo lie beneath the detection limit.
[b] Ultimate tensile strength σ_u (normalized) = 213 MPa.

Current density – potential curves for stressed and unstressed tensile test specimens in crack-inducing media show that the unstressed specimen exhibits no rise in current until ca. 900 mV U_H, which indicates the start of grain boundary corrosion (Fig. 46) [85]. At 20 % of the tensile strength σ_u, corrosion begins at 400 mV, i.e., a shift of 500 mV in the direction of the actual corrosion potential in the steel. At a stress of $0.4\,\sigma_u$ the grain boundary breakthrough potential reaches the corrosion potential, and the steel can now rupture without external polarization through stress corrosion cracking. As it does, the critical tensile stress for dissolution is reached.

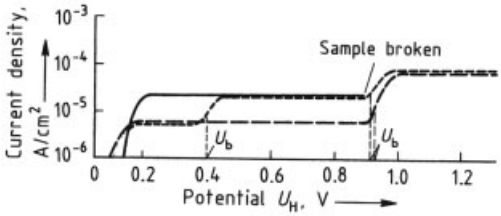

Figure 46. The influence of load on the grain boundary breakthrough potential U_b
Load: ––– unloaded; ----- $0.2 \cdot \sigma_u$; —— $0.4 \cdot \sigma_u$ ($\sigma_u = 438$ MPa)

In practice, the considerable influence of the stress level can also be seen by the fact that weld seams are preferentially attacked by stress corrosion cracking on account of shrinkage stresses and the stresses caused by structural changes due to the effect of heat input during welding.

The susceptibility to stress corrosion cracking of unalloyed and low-alloy steels is thus indicated by the extent to which the grain boundary breakthrough potential shifts under tensile stress. If, for a low tendency to shift, the stress limit needed to induce cracking is greater than the yield strength, industrial application of a steel of this kind is possible given careful monitoring of the nominal stress. The structural composition formed through annealing is also of importance, as is shown in Figure 47 for the fine-grained structural steel StE 355 [86].

Stress corrosion cracking in alkali-metal hydroxide solutions shows a number of variations. Whereas in nitrate solutions the grain boundary breakthrough potential only means a restriction of the potential range of stress corrosion cracking towards the cathodic direction (there is no limit potential for cracking at the anodic side), in alkali there is both a cathodic and an anodic limit potential, as shown in Figure 48 [87]. In this instance, stress corrosion cracking, as illustrated by the current density – potential curve, is restricted to the potential range corresponding to formation of a passive layer that is initially disturbed at the grain boundaries. Susceptibility disappears once complete passivation has been attained. In alkali, initiation of crack-

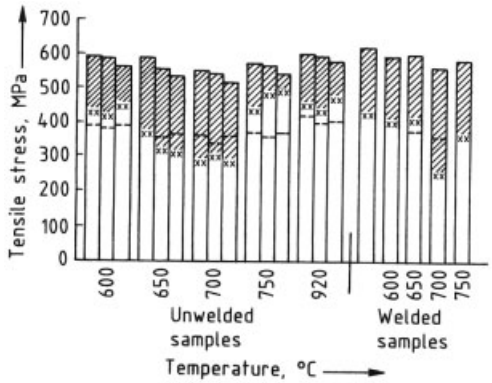

Figure 47. Threshold tensile stresses for stress corrosion cracking of the fine-grained structural steel StE 355 as a function of the annealing temperature as determined with unwelded and welded samples
Plain area = Tensile strength; –– Yield strength; xx Threshold stress for stress corrosion cracking; Hatched area = Range of stress corrosion cracking

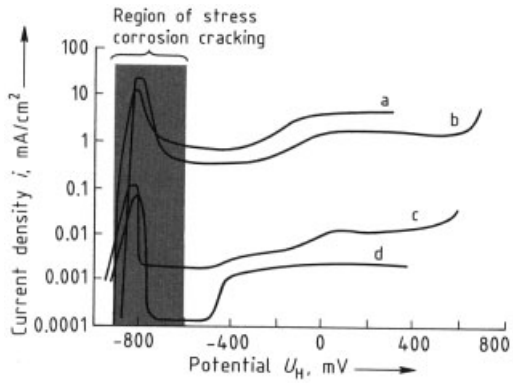

Figure 48. Current density – voltage curves of unalloyed steels in alkaline solution and position of stress corrosion cracking
a) 33 % NaOH, boiling, holding time 60 s per measuring point [88]; b) 35 % NaOH, boiling, holding time 30 s per measuring point [214]; c) 35 % KOH, 80 °C, holding time 5 h per measuring point [87]; d) Anodic partial current – voltage curve for dissolution of iron in 4 N KOH (ca. 20 %), 90 °C [77]

ing requires tensile stresses comparable to the yield strength. In contrast to nitrate solutions, therefore, the limit stress to initiate stress corrosion cracking in alkali is characterized in most cases by the requirement for small amounts of permanent elongation [88].

In alloyed air-hardened steels (e.g., 13 CrMo 4 4), thorough tempering at 750 °C is needed after welding to eliminate the hardness structure. Only when the hardness is below 220 HV are these steels resistant to stress corrosion cracking.

In addition to the known corrosive effects of alkali-metal hydroxide and nitrate solutions, intergranular stress corrosion cracking in unalloyed and low-alloy steels in contact with ammonium carbonate [82], [89], [90] and crude methanol (methanol with a low concentration of impurities) has now also been observed [91], [92]. When this group of materials comes into contact with various other aggressive substances, stress corrosion cracking occurs, primarily with transgranular characteristics.

Figure 49. Stress corrosion cracking in an unalloyed steel storage tank for liquid ammonia

In the past, cases of this sort of crack damage in storage tanks and transportation containers for liquid ammonia (both at room temperature and at 0 °C) were found in increasing numbers (Fig. 49) and simulated in laboratory experiments [93–96]. In a few cases the formation of cracks was even observed in areas where the vapor phase had condensed, and in low-temperature storage (– 33 °C) [97].

Transgranular stress corrosion cracking also occurs in solutions of $CO-CO_2-H_2O$ containing oxygen [98], [99]. Unalloyed and low-alloy steels may also be sensitive to anodic stress corrosion cracking in aqueous CO_2 solutions. Note, however, that considerable hydrogen permeation has been observed under critical conditions ($t > 40\,°C, p_{CO_2} > 1$ MPa, $\sigma > 0.5 \cdot \sigma_{ys}$); a certain amount of hydrogen-induced stress corrosion cracking in the overall cracking activity must therefore be assumed [100, 101].

3.3.1.2. Stainless Steels

Chloride-containing corrosive media can give rise to transgranular stress corrosion cracking in austenitic stainless steels (e.g., steels of composition 18 % Cr, 10 % Ni, or 18 % Cr, 12 % Ni, and 2 % Mo). It occurs particularly often through the high chloride content of cooling water (usually river water) in heat exchanger tubes: in many cases in the region of salt deposits that arise through local evaporation with insufficient water throughput or incomplete filling of the tube [102].

Alkali-metal hydroxide solutions can also cause stress corrosion cracking in austenitic stainless steels (Fig. 50; see next page), albeit only at elevated temperature, often above the boiling point of the corrosive medium (Fig. 51; see next page). Therefore, damage occurs only occasionally and primarily in pressure vessels. However, in Figure 52 (see next page) lower temperature limits are given [103].

It must be assumed in transgranular stress corrosion cracking that local accumulations of foreign atoms in areas with lattice defects outside the grain boundaries, i.e., dislocations and slip lines, are responsible for local activation, which can often be recognized by the formation of etch pits in the presence of corrosive media or through a particular proneness to pitting.

Less prone to stress corrosion cracking are high-alloy austenitic steels that contain more chromium and nickel or have additional copper and molybdenum contents. Particularly resistant are nickel-based alloys with chromium and/or molybdenum, in which stress corrosion cracking

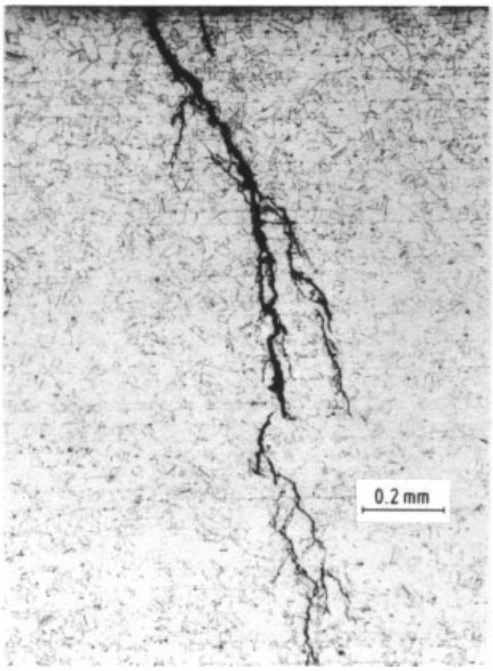

Figure 50. Stress corrosion cracking of a loop sample of X 10 CrNiMoNb 18 10 (material no. 1.4580) caused by an alkaline solution at 280 – 300 °C and 3 MPa

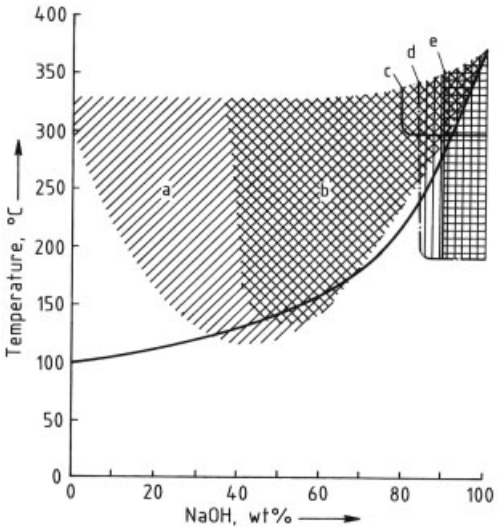

Figure 51. Stress corrosion cracking regions of various metallic materials in sodium hydroxide
a) 18/8 CrNi; b) 18/8/2 CrNiMo; c) Monel; d) Nickel; e) Inconel
—— Boiling curve at 100 kPa

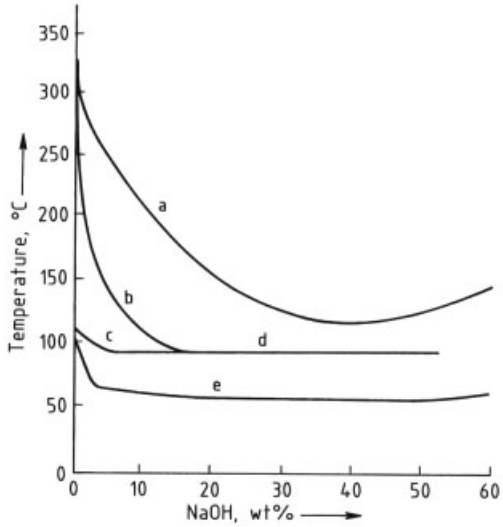

Figure 52. Stress corrosion cracking (SCC) of austenitic stainless steels in sodium hydroxide solution: temperature and concentration limits [103]
a) Rapid stress corrosion failure (one day), unsensitized; b) Delayed SCC failure, unsensitized type 304; c) Delayed SCC failure, sensitized type 304; d) Delayed SCC failure type 304, both sensitized and unsensitized (100 – 300 d); e) Tentative safe SCC limit

occurs only under extreme stress. Commercial ferritic stainless chromium steels also show very high resistance, but for other reasons (e.g., welding sensitivity) are seldom used in equipment fabrication. Ferritic – austenitic steels are readily applicable in approximately neutral halide solutions. Apart from chlorides, which are responsible for most of the damage, bromides and iodides can also cause stress corrosion cracking, although such cases are rare.

A tensile stress limit must also be exceeded for transgranular stress corrosion cracking to occur in stainless steels. The limit stresses are shown in Table 4 for three standard steels in boiling 42 % $MgCl_2$ solution – the standard laboratory solution for investigating stress corrosion cracking [104]. The limit stresses are always lower than the yield stress, as the comparison with the 0.2 % yield strength shows. Components are usually designed with a 150 % safety margin based on the yield strength. The nominal stresses calculated for the individual steels are also listed in Table 4. The limit stress is higher than the nominal stress only for the steel with the material no. 1.4550. This does not, however, ensure safety against stress corrosion cracking,

Table 4. Comparison of the yield strength and the nominal stress σ_N present with normal laying out with the limit stress necessary to initiate stress corrosion cracking for three austenitic steels

Steel		0.2 % yield strength, MPa	Limit stress, MPa	$\sigma_N = \frac{\sigma_{0.2}}{1.5}$, MPa
Mat. no.	Designation (DIN 17 006)			
1.4550	X 6 CrNiNb 18 10	180	160	120
1.4435	X 2 CrNiMo 18 14 3	200	130	132
1.4429	X 2 CrNiMoN 18 133	280	100	155

since stress peaks and local yield processes cannot be avoided in these highly ductile steels, which, for example, are not usually heat-treated after welding. It must also be expected that, as pitting begins, the required degree of stress is reached because of the excess stress at the edges of the pits. Transgranular stress corrosion cracking can occur in the presence of halides under two conditions:

1) Below the limit stress through shifting of the corrosion potential towards the pitting potential by oxidants, whereby, depending on the corrosive medium, the increase in potential need sometimes only be small, and subsequent exceeding of the limit stress through excess stress at the location of pitting;
2) Above the limit stress through the occurrence of local yield processes with the formation of local corroded spots on the steel surface with high current density and due to perforation of the surface layer by slip lines as the start of cracking.

If stress-relief annealing is to be carried out on stainless austenitic steels to avoid stress corrosion cracking, temperatures above 900 °C are required because of the high strength at elevated temperatures. The use of more resistant materials can therefore be regarded as a more suitable alternative.

3.3.1.3. Nonferrous Metals

In the presence of specific corrosive media, nonferrous alloys are also prone to stress corrosion cracking. Some aluminum alloys that contain Zn and Mg, for example, are sensitive to chloride solutions and seawater, and copper alloys, especially brass, are attacked by moist ammonia, causing cracking [105]. In aluminum alloys cracking is intergranular, and the formation of the cracks is related to the degree of chemical activity of the grain boundaries, as shown in Figure 53 for alloy AlMg 9 as a function of heat treatment [106].

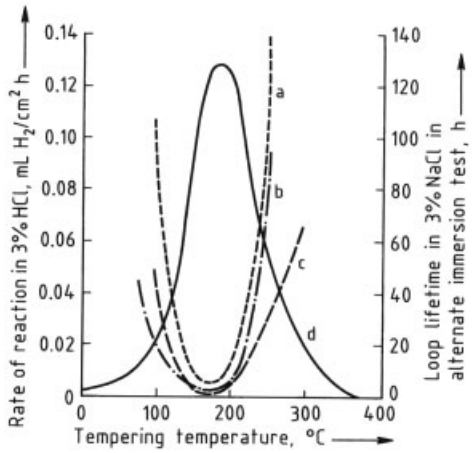

Figure 53. Stress corrosion cracking behavior (a – c) of loop samples in 3 % NaCl solution (lifetime in alternate immersion test) and chemical activity (hydrogen evolution) (d) of the grain boundaries in hydrochloric acid for 1 mm aluminum alloy sheet with 9.1 % Mg

At a tempering temperature of 180 °C, the activity of the grain boundaries is particularly high, and the life of loop-shaped specimens in the alternating immersion test in sodium chloride solution is particularly short. In Cu – Zn alloys, cracking unter stress corrosion cracking conditions depends on the structure, intergranular cracking being prevalent in α alloys and transgranular in β alloys.

3.3.1.4. Special Types of Stress Corrosion Cracking

An uncommon type of stress corrosion, which is of importance in practice, is transgranular cracking of copper-containing stainless austenitic Cr – Ni – Mo steels in sulfuric acid. Cracking de-

pends on the concentration and temperature of the acid and the level of mechanical stress [107]. In boiling sulfuric acid, the critical concentration range in which the transgranular stress corrosion cracking, depicted in Figure 54, occurs is 20–60 %.

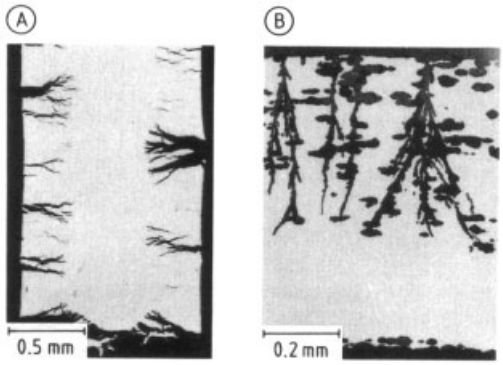

Figure 54. Transgranular stress corrosion cracking of stainless steel X 5 CrNiMoCuNb 18 18 in sulfuric acid
A) Tensile specimen, stress $0.3 \cdot \sigma_{ys}$, 50 % H_2SO_4, boiling;
B) Bent-beam specimen, 30 % H_2SO_4, boiling

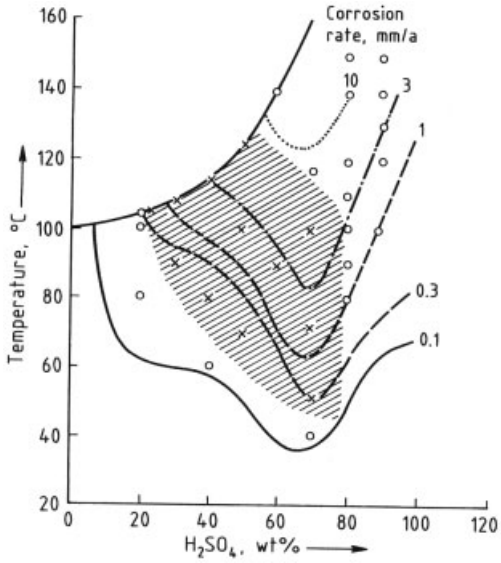

Figure 55. Iso-corrosion curves of X 5 CrNiMoNuNb 18 18 and field of SCC
Hatched area = Stress corrosion cracking region; ○ = Sample without stress corrosion cracking; × = Sample with stress corrosion cracking; —— = Boiling curve

In addition to the iso- corrosion curves of the steel X 5 CrNiMoCuNb 18 18, Figure 55 shows the stress corrosion cracking zone. The data refer to aerated sulfuric acid as the corrosive medium. A notable feature is that cracking does not occur at all temperatures and concentrations, but only above a given corrosion rate, which under the conditions in this case is around 0.2 mm/a. In many cases copper deposits on the steel, which is due to the prior corrosion of the copper-containing solid solution.

The formation of copper-containing protective layers, which apparently exhibit local defects under the influence of tensile stress, is a prerequisite for this type of corrosion. There must therefore first be a minimum corrosion rate, as otherwise the required concentration in the solution for copper deposition cannot arise. However, if the corrosion rate is too high, a copper layer will not form to a sufficient extent and no cracking occurs. In this case, too, the susceptibility of copper-containing, austenitic stainless steels to stress corrosion cracking can be eliminated by increasing the nickel content to over 32 %.

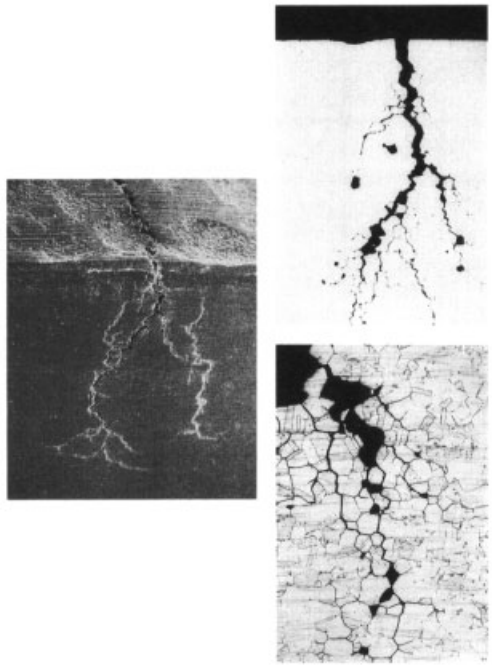

Figure 56. Intergranular stress corosion cracking of a loop sample of X 3 CrNiMoN 17 13 5 (material no. 1.4439) caused by saturated ammonium nitrate solution at 180 °C

Transgranular stress corrosion cracking is induced not only in austenitic stainless steels but also in nickel alloys by aqueous alkaline solu-

tions. Since the initiation temperature for cracking is high, often exceeding the boiling point of the alkaline solution, it occurs primarily in pressure vessels. The precise conditions are set out in Figure 51 [108]. In ammonium nitrate solutions a type of intergranular stress corrosion cracking (see Fig. 56, see previous page) of high-alloy austenitic stainless steels is observed at temperatures close to the decomposition point of ammonium nitrate [109].

3.3.1.5. Hydrogen-Induced Cracking at Low Temperatures

The absorption of atomic hydrogen at temperatures below 100 °C changes the toughness and ductility of steel; the effect increases with increasing steel strength. This can be recognized by the disappearance of the distinctive yield point and the loss of reduction of area and elongation in tensile testing.

The quantity of hydrogen H_{ab} being absorbed by the metallic phase from an electrolyte solution in which hydrogen ions are discharged, corresponds only to the hydrogen gas pressure in the solution (generally 100 kPa, given undisturbed equilibrium between H_{ad} and H_2 gas):

$$H^+_{aq} \rightarrow (H^+_{aq})_{surface} \xrightarrow[e-]{\text{reduction}} H_{ad} \begin{cases} \nearrow H_{ad} + H^+_{aq} + e^- \rightarrow H_2 \uparrow \\ \rightarrow 2\,H_{ad} \rightarrow H_2 \uparrow \\ \searrow H_{ab} \end{cases}$$

Only when the steps involving recombination of atomic hydrogen are inhibited can an increased H_{ad} concentration and hence a higher value for H_{ab} occur. Some substances result in a particularly high retention of atomic hydrogen on a steel surface in corrosion reactions by inhibiting recombination. They are known as promoters or poisons, and include compounds of S, P, As, Se, Te, and Sb (e.g., thiocyanates and thiosulfates) as well as CO and cyanides. They cause a considerable increase in the concentration of hydrogen absorbed in the metal, which formally corresponds to an increased hydrogen pressure in the gas phase. This can give rise to considerable embrittlement of steel even under normal pressure. Apart from AsH_3, COS, HCN, and CO, above all H_2S is responsible for many cases of damage, in particular in the chemical, petroleum, petrochemical, and electroplating industries.

In mild steels, blisters, cracks, and terraced fractures (caused by slag and sulfide lines) can occur under sufficiently high hydrogen uptake and static tensile stress below the yield point at low temperature (e.g., room temperature): this results from the formation of high pressure hydrogen in structural defects as shown in Figure 57 [110]. Residual stress is often sufficient to generate local plastic deformation and initiate cracking.

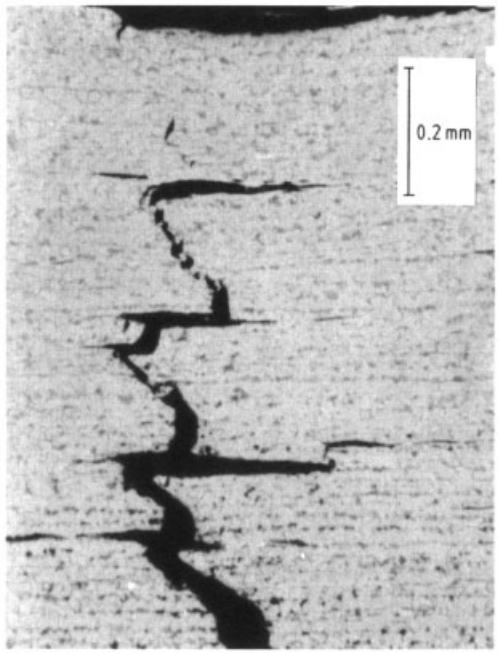

Figure 57. Terrace-shaped fractures in a natural gas pipeline of unalloyed steel (X 42 API Std 5 LX) caused by condensate containing H_2S

Cracking occurs far more readily and rapidly in intermediate-strength steels. In cold parts of high-pressure petrochemical plants, cracks are often found in shaped components made of heat-treatable steel through the attack of condensates containing H_2S. The same applies to equipment for drilling and extracting natural gas [111].

Hydrogen-induced embrittlement in high-strength steels is less dependent on the occurrence of local internal hydrogen pressure zones, because they cannot induce local plastic deformation in such steels. It depends on the interaction between dissolved and mobile hydrogen and stress fields, precipitation regions, and crack

tips during a local plastic deformation process. The proneness to embrittlement by plastic deformation under the influence of low-pressure molecular hydrogen was reported by HOFMANN et al. [112].

With pure molecular hydrogen, atomic hydrogen can originate on places of plastic deformation (notches, scratches) by interaction between molecules and the resulting active surface of steels. The absorbed hydrogen causes crack initiation and propagation. This occurs particularly under low-cycle fatigue conditions. Damage occurs, e.g., at inner die marks in high-pressure gas cylinders for hydrogen transport that are often filled and emptied [113], [114].

A purely optical differentiation between hydrogen-induced cracks and cracks produced by anodic stress corrosion cracking is often extremely difficult. Figure 58 shows a case that occurred in a spiral-welded natural gas pipeline made of St 52.3 as a result of condensate containing H_2S and CO_2 in the region of the weld and in a neighboring stress zone. The observation that weld seam areas and cold-formed zones are particularly prone to cracking is a general one. The proneness to cracking of welded equipment in the presence of nascent hydrogen can usually be eliminated by stress-relieving annealing.

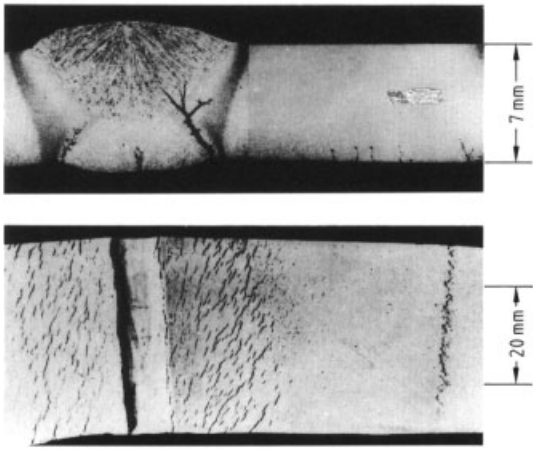

Figure 58. Hydrogen-induced crack formation in a spiral-welded natural gas pipeline (made of St 52.3) caused by condensate containing H_2S and CO_2

A striking feature is provided by the "fish eyes" sometimes seen in fractures, indicating that hydrogen was involved in the bursting process (see Fig. 59A). They arise when the yield limits are exceeded at the edges of hydrogen-filled cavities through diffusion of atomic hydrogen into the surrounding area, leaving a confined, speck-shaped brittle fracture with a bright fracture surface. In this process the hydrogen dissociates at the plastically deforming cavity wall.

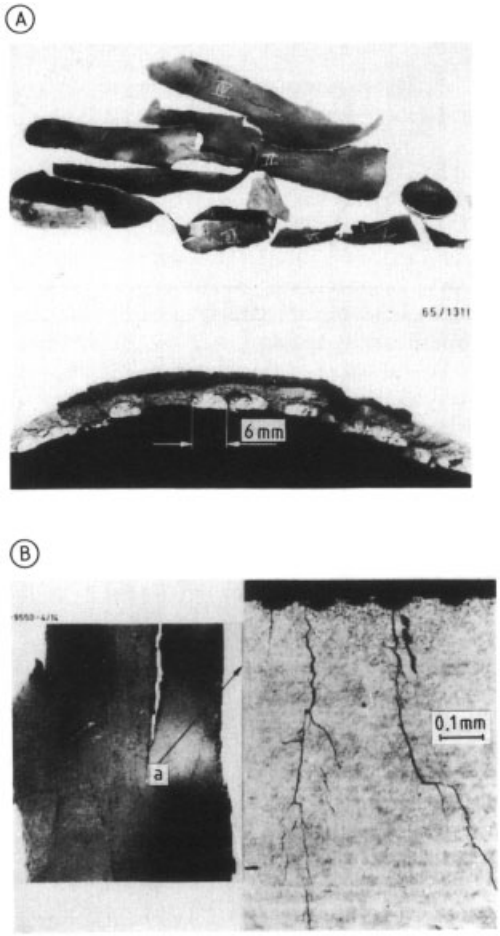

Figure 59. An exploded carbon monoxide lightweight steel bottle made of a heat-treatable manganese steel ($\sigma_{ys} = 840$ MPa)
A) General view and fracture surface; B) Fragment showing course of crack: a) Micrograph

An example of the tremendous effect hydrogen-induced cracking can have on high-strength steel is the explosion of a light steel bottle (40 L volume) upon filling (Fig. 59). It was made of a low-alloy heat-treatable manganese

steel with a yield strength of 840 MPa and a tensile strength of 950 MPa. The bottle had been in use for 2 1/2 years, filled with technically pure CO up to 20 MPa [108], [115]. In addition to the presence of CO as promoter, a major part in this damage was played by the existence of a corrosive water sump in the bottle since the reaction leading to the formation of atomic hydrogen is electrolytic. Numerous cracks were found in the area covered with water.

Just how important the absolute absence of moisture films is, especially in bottles made of high-strength steel, is shown by the damage in the form of hydrogen-induced stress corrosion cracking caused after several years of being filled with pure CO_2 (8 – 10 MPa) to a number of fire extinguisher bottles, which burst. The initiator and source of hydrogen was the weakly acidic carbonic acid formed through the presence of water, which acted on the mostly horizontally-positioned bottles along the inner surface of the cylinder and caused cracking [115]. Ultrahigh-strength steels are so prone to hydrogen-induced stress corrosion cracking that they can suffer brittle fracture under the effect of atmospheric humidity.

If there is a danger of electrolyte-induced atomic hydrogen formation and it is not possible to eliminate the electrolyte (e.g., by drying), the use of steels as soft as possible and careful stress-relieving annealing of the components should be considered. Since austenitic Cr – Ni steels in standard supply form and after normal processing are practically unaffected by hydrogen they can be used in difficult cases.

Because of their tendency to hydride formation, the embrittlement of titanium, zirconium, and tantalum by hydrogen is particularly pronounced. The proneness to gas absorption restricts the use of these metals. They are not only attacked by nascent hydrogen, they also become brittle at elevated temperatures in gases containing nitrogen and oxygen. In addition to uniform corrosion, these metals also show hydrogen embrittlement in corrosion reactions in acidic solutions (Fig. 60). In the case of tantalum, hydrogen embrittlement can occur after long periods in use, even when the corrosion rate is practically negligible (e.g., several thousandths of a millimeter per year) [116].

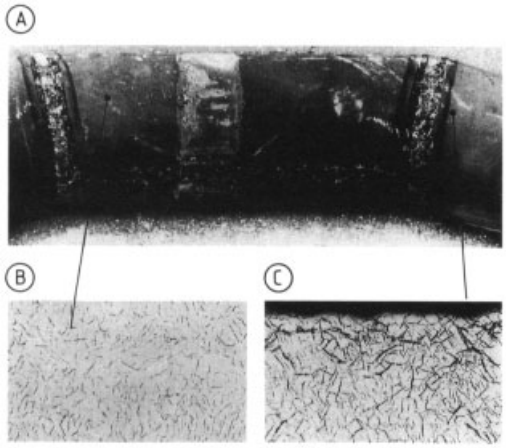

Figure 60. Corrosion of a weld of a titanium pipe
A) General view; B) and C) Embrittlement as the result of hydrogen absorption (structure exhibiting hydride needles); bending angle up to fracture 12° (A) and 4° (B)

3.3.1.6. Hydrogen Corrosion at Elevated Temperatures

The effect of pressurized molecular hydrogen on steel must be borne in mind in a number of high-pressure syntheses, e.g., in high-pressure hydrogenation and ammonia production. At sufficiently elevated temperatures (> 230 °C) pressurized hydrogen decarburizes carbon steels. The results of this process are embrittlement, poorer mechanical properties, grain-boundary loosening and a splitting effect at defect sites caused by reaction products. As a consequence, brittle fractures occur (Fig. 61).

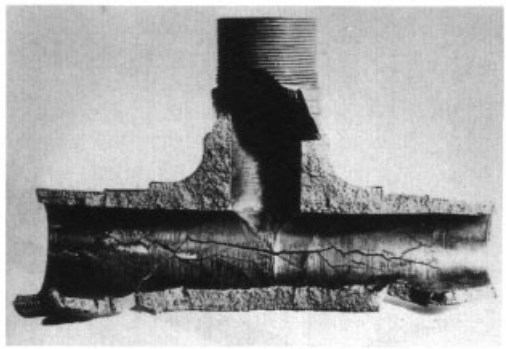

Figure 61. Brittle fracture of a fitting (T-piece) as a result of pressurized hydrogen attack

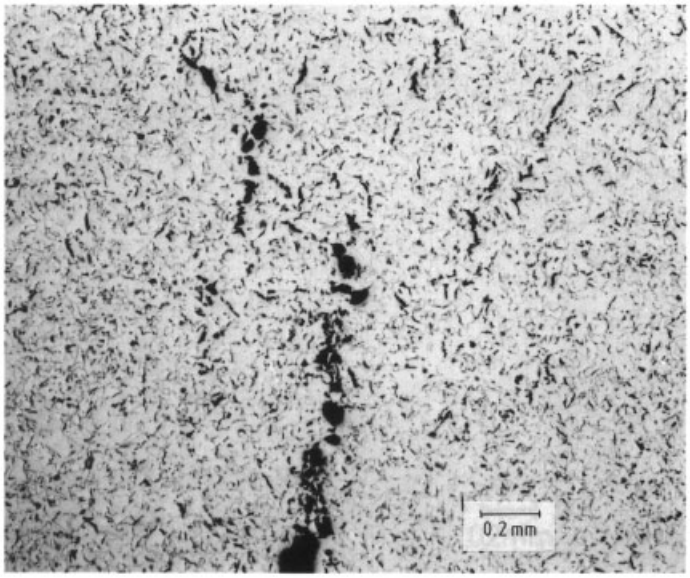

Figure 62. Damage to a St 35.8 pipe (NH_3 plant) by pressurized hydrogen

Figure 62 shows decarburization and grain loosening in the microstructure of an unalloyed steel as a result of pressurized hydrogen attack. The visible changes within the corroded areas cause significant changes in mechanical properties. In the decarburized zone, the tensile strength, elongation, reduction in area, and hardness drop to very low levels as the carbon content decreases.

Even the first observations of damage of this kind in the development of ammonia synthesis showed that the effect of pressurized hydrogen on steel is both chemical and physical [117]. The physical effect presumably takes place in different stages. Since molecular hydrogen does not diffuse in the steel, the first step must be surface adsorption with dissociation at elevated temperature. In this process the metallic wall acts as a catalyst. With pure, dry pressurized hydrogen, for example, dissociation produced by yielding phenomena on the steel surface can supply critical quantities of diffusible hydrogen even at room temperature.

The next step after formation of adsorbed hydrogen is absorption of the hydrogen. The ensuing chemical effect of hydrogen on steel is marked by decarburization of pearlite and formation of hydrocarbons, usually methane:

$$Fe_3C + 2\,H_2 \longrightarrow 3\,Fe + CH_4$$

The reaction generally occurs in areas adjacent to the grain boundaries, with the grains losing their cohesion. In addition to methane, recombined hydrogen molecules accumulate at the loosened points producing a high pressure, which leads initially to internal microcracking. Component stress and lowered mechanical properties finally combine to produce marked brittle fracture.

Apart from the ambient conditions, the resistance of carbon steels to pressurized hydrogen attack depends on the grain size, cold forming, welding effects, impurities, and heat treatment. To prevent the reaction of hydrogen with pearlite, carbon-bonding elements such as chromium, molybdenum, tungsten, vanadium, titanium, and niobium, whose carbides are considerably more stable than cementite, can be added to the steel. These alloying elements considerably increase the resistance of steels to hydrogen attack [118].

Steels resistant to pressurized hydrogen that are now in use contain 2 – 6 % chromium and often a further 0.2 – 0.6 % molybdenum. Some types are additionally alloyed with vanadium up to 0.85 % and tungsten up to 0.45 %. Short-term testing gives little information on where the limits of application of such steels lie, especially since corrosion by pressurized hydrogen

normally starts after a considerable incubation period. However, experience gained with a wide variety of installations does provide reliable indicators. Figure 63, for example, shows the resistance limits for various steels [119]. Even at low hydrogen partial pressures (e.g., 2 MPa) a carbon steel should not be used above a temperature of 300 °C; small amounts of molybdenum are sufficient to increase resistance appreciably. Austenitic steels with 18 % chromium and ca. 10 % nickel possess good resistance to pressurized hydrogen attack and can be used with hydrogen over the the entire temperature range of common high-pressure processes.

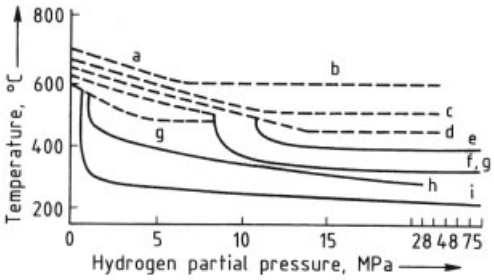

Figure 63. Resistance diagram for the attack of pressurized hydrogen on steels (Nelson diagram) [119]
a) 1.25 Cr 0.5 Mo steel; b) 5.0 Cr 0.5 Mo steel; c) 8.0 Cr 0.5 Mo steel; d) 2.25 Cr 1.0 Mo steel; e) 2.0 Cr 0.5 Mo steel; f) 1.25 Cr 0.5 Mo steel; g) 1.0 Cr 0.5 Mo steel; h) 0.5 Mo steel; i) Mild steel

3.3.2. Corrosion Fatigue

If in addition to alternating mechanical stresses a component is exposed to a constant corrosive effect through the environment, the resulting damage is called corrosion fatigue. The Wöhler curve describes the fatigue behavior of a metallic material without corrosion. The corrosive influence shifts the area of fatigue strength for finite life ($< 10^7$ load cycles) to lower numbers of cycles and shows in addition, as depicted in Figure 64, no further fatigue strength [120]. Accordingly, only a fatigue strength for finite life depending on different parameters can be achieved under a load combination of this kind.

The corrosion process itself is localized by slip processes at the surface generated by the mechanical stress. The corrosive effect can also be recognized from the corrosion products that often deposit in cracks. These are found extensively at the crack beginning, the area of slower crack propagation, but are not much in evidence at the crack tip, the area of mostly rapid propagation. Extensively corroded cracks indicate a high number of cycles to failure, low load amplitude, and severe corrosion. The component finally breaks when the remaining cross section is stressed to such an extent by the continuing corrosion fatigue that it cracks through spontaneous rupture. As an example Figure 65 shows damage to a pipe elbow made of 13 CrMo 4 4 steel caused by alternating thermal stress with simultaneous exposure to steam (low frequency, long service period, marked corrosion).

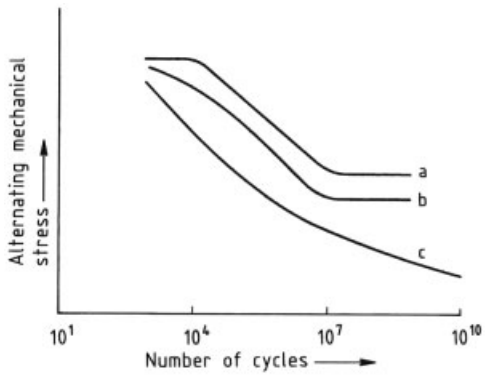

Figure 64. Wöhler curves, in air, after precorrosion, and under corrosion fatigue conditions
a) Fatigue strength in air; b) Fatigue strength after precorrosion; c) Corrosion fatigue

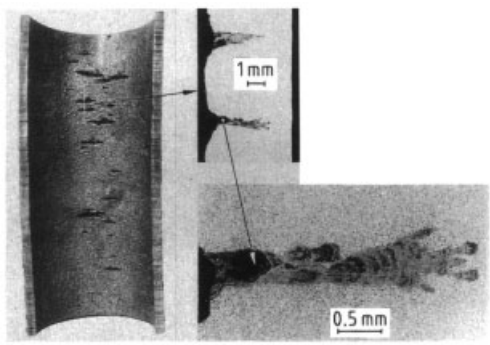

Figure 65. Cracks in a pipe elbow made of 13 CrMo 4 4 caused by alternating thermal stress and by corrosion
Conditions: 340 °C, steam, 10^5 h (11.4 a)

In addition to the metallographic examination, microfractography can also be used to identify the cause of damage since in areas where corrosion of the rupture surface and deposition

are not too extensive, striations are often found in corrosion fatigue (Fig. 66), indicating the action of alternating stress [121]. The decision as to whether pure fatigue failure or corrosion fatigue is involved can be facilitated by the appearance of corrosion phenomena or products on the rupture surface.

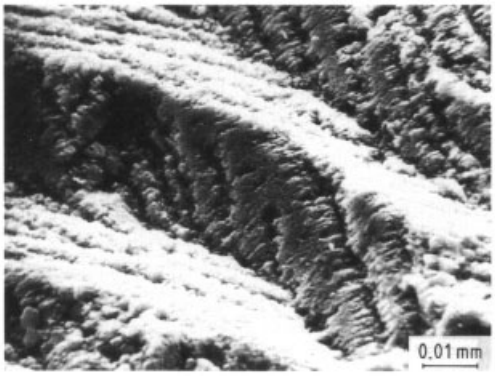

Figure 66. Fatigue striations on the fracture surface of an oil cooler pipe made of brass SoMs 76, cracked by corrosion fatigue due to river water

Electromechanical methods have also been employed to study the mechanism of corrosion fatigue [122]. If the potential changes occurring at the surface of the specimen during corrosion fatigue are recorded, three stages of corrosion are found for steels with an active surface.

In the *preliminary stage*, potential drops in the negative direction are observed, the cause of which can be seen as the appearance of slip lines on the specimen surface with increased activity, i.e., increased rate of reaction with the electrolyte.

In the *second stage*, these local anodes increase in number and area, forming notches with high mechanical stress and hence high dissolution, which is marked by a further denobling of the potential.

In the *final stage*, interaction of localized corrosion and increased notching finally leads to cracking and thus to the spontaneous rupture of the remaining cross section.

If corrosion fatigue of unalloyed and low-alloy steels is allowed to occur potentiostatically and the changes in current are measured as corrosion progresses, it is seen that marked current peaks only occur at the stage of rapid crack propagation [123].

In passive steels the current rises only after local activation, at which point the cracking stage begins immediately. This is because only one crack forms in the otherwise passive surface, causing extreme localization of the corrosion process and hence a correspondingly high current density at the crack base, which explains the extraordinarily rapid crack propagation. Stainless steels can of course become fully active in appropriate corrosive media, whereupon their corrosion fatigue behavior is similar to that of active steels. Whereas a single smooth break generally occurs in the passive state, without any further signs of corrosion attack [124], the indicators in corrosion fatigue in the active state are numerous cracks, a fissured break due to the converging of various crack levels, and general corrosion attack. A comparison of active and passive corrosion fatigue is shown in Figure 67 [125].

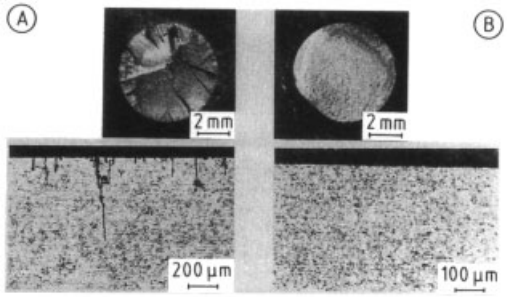

Figure 67. Active and passive corrosion fatigue
A) Active state: severe corrosive attack, fissuring of the fracture surface, several cracks; B) Passive state: no evidence of corrosive attack, smooth fracture surface, no secondary cracks

Since a frequently repeated minimum stress amplitude is needed to initiate corrosion fatigue in stainless steels in passivating solutions, it must be assumed that this load limit is connected with the mechanical stress capacity of passive layers. Only when particularly marked slip starts at one point on the surface will the layer be cracked. At that point a constantly repetitive process begins, in which a new activation process at the same spot always follows repassivation. Each new passivation process consumes metal, deepens the corrosion, and increases the stress peak until, because of the constantly rising stress, repassivation is no longer possible. The resistance of passive metallic materials under fatigue

conditions in electrolytes is therefore largely dependent on three factors:

1) the extent of the load amplitude in relation to the mechanical strength of the material and the passive layer;
2) the passivation properties of the electrolyte;
3) the passivation tendency of the material, as determined by its alloying elements.

The mechanical properties of the alloy determine the occurrence of slip. The higher the fatigue strength in air the greater the stress amplitude needed to initiate sufficiently extensive slip, i.e., the higher the resistance to corrosion fatigue in the passive state.

In the case of "protective layers" the mechanical strength and the adhesion of these layers and their healing capacity in corrosive media are of special significance in corrosion fatigue. For thicker layers, which themselves are subject to fatigue, a special Wöhler curve applies, whose path is determined by the fatigue strength of the layer if it is lower than that of the base material (Fig. 68). If the operating stress is below the fatigue strength of the material but above that of the protective layer, the layer cracks at the points of highest stress, and local corrosion begins. The corrosion that forms provides the starting point for subsequent cracks. Corrosion fatigue of metallic materials with protective layers is therefore preceded by a period of strain-induced corrosion, formerly known as stress-induced corrosion [126].

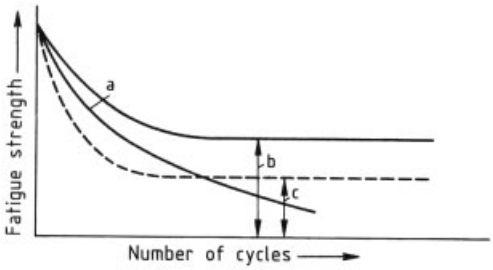

Figure 68. Cracking of a protective layer as a precondition for corrosion fatigue (strain-induced corrosion)
a) Corrosion fatigue; b) Fatigue strength in air; c) Fatigue strength determined by the protective layer

To increase the resistance to corrosion fatigue, the same steps can be taken that give high structural strength (good surface, no notches or grooves, avoidance of stress peaks through good geometrical design and processing). However, to these are added the increased corrosion resistance of the materials used, since greater chemical resistance plays a decisive role in achieving longer life.

An especially noteworthy feature is a process known as low-cycle corrosion fatigue, in which a stress producing constantly repetitive deformation amplitudes and corrosion cause damage after only a low alternating load cycle. The interaction of plastic alternating strain and a corrosive electrolyte solution leads to rupture, especially in components with notches or fissures, after only a small number of load cycles if the corrosion is long-acting due to the low cycle frequency, especially in the tensile phase (slow change in load). Embrittlement and crack propagation also be caused under these conditions by gases (e.g., H_2) and media that show hardly any corrosive effect under static load (see Section 3.3.1.5).

3.3.3. Differentiation Between Types of Corrosion Cracking

The term stress corrosion cracking often refers to very different types of cracking under corrosive attack. A clear distinction must be made between their corrosion mechanisms. The different types of corrosion cracking are set out in Table 5 (see next page). The distinction between the types of corrosion cracking is made difficult by overlapping. This overlapping is found in particular between nonclassical stress corrosion cracking and corrosion fatigue and between corrosion fatigue in the passive state and strain-induced corrosion cracking.

In the more recent literature the term *strain-induced corrosion* is commonly found, which is a generic term for nonclassical stress corrosion cracking, corrosion fatigue with low cycle frequency, and strain-induced corrosion cracking. For the purposes of a clear distinction between the different types of cracking this term should be replaced by the following:

1) strain-induced stress corrosion cracking for nonclassical stress corrosion;
2) low cycle corrosion fatigue for strain-induced fatigue corrosion with low-frequency load cycles on materials and protective layers;

Table 5. Types of corrosion cracking [127]

Type of cracking	Stress corrosion cracking		Corrosion fatigue		Strain-induced corrosion cracking [a]
	Classical (stress-induced)	Nonclassical (strain-induced)	Active state	Passive state	
Type of stress	constant elongation (strain rate $\dot{\varepsilon}=0$)	continuous strain ($\dot{\varepsilon}>0$)	pulsating strain		alternating elongation (with nonstatic load)
Specimen and test method	smooth clamped specimen (constant elongation)	notched tensile test specimen (constant load)	smooth tensile specimen, pulsating or oscillating test		any form of specimen, alternating load (mechanical and chemical)
		smooth tensile test specimen (slow strain rate)			
Corrosion system	critical system parameter values		no critical system parameter values		
					only in systems with film formation
Characteristics of cracking	inter- or transgranular		mostly transgranular		transgranular
Number of cracks	often many cracks		many cracks	few single cracks	few cracks

[a] Formerly known as stress-induced corrosion

3) strain-induced corrosion cracking for corrosion with cracking of materials with protective layers as a result of layer damage on elongation change and corrosion during shutdown periods (non-stationary operation).

4. High-Temperature Corrosion

4.1. Principles of High-Temperature Corrosion

4.1.1. Thermodynamic Principles

The reaction of a metal M in normal state (100 kPa) with an oxidizing gas X_2 produces a solid reaction product MX, which usually forms a layer on the surface ($M/MX/X_2$). The free enthalpy of such a reaction is negative for practically all metal oxides, i.e., oxides are stable in atmospheres containing oxygen. Metals therefore are not stable, and oxidation or corrosion occurs always, although it is not visible in many cases because of the low reaction rate.

With rising temperature the free enthalpy approaches zero. This means that at a given temperature, equilibrium exists between metal, metal oxide, and oxygen (100 kPa pressure). All phases are stable next to one another, i.e., under the prevailing conditions the decomposition pressure of the metal oxide is equal to the oxygen pressure of the surrounding atmosphere. At lower oxygen pressures the compound decomposes into its elements.

At room temperature, therefore, practically all metals and metal alloys – with the exception of gold, whose oxide decomposition pressure is greater than the oxygen partial pressure of air – are unstable in air and form surface oxide films. At higher temperatures, all metal compounds decompose in principle. This is also the case at moderate temperatures for the oxides of silver, mercury, platinum, and palladium, whose oxygen partial pressures (decomposition pressures) between 140 °C and ca. 800 °C exceed the oxygen pressure of air. At elevated temperatures, however, this also applies to the oxides of copper, iron, and other less noble metals. For most oxides, however, the oxygen partial pressures required for dissociation are too low to be achieved experimentally, or else the decomposition temperatures are above the boiling points of the metals or metal compounds.

The decomposition pressure of a metal oxide p_{O_2}, i.e., the oxygen partial pressure of the surrounding atmosphere at which the oxide is just stable at a given temperature, can be calculated from Equation (13):

$$G_0 = RT \ln\left(1/p_{O_2}\right) = 4.574 T \log p_{O_2} \quad (13)$$

Equilibrium values (e.g., decomposition pressures) therefore determine the existence of compounds in metal oxidation, but the rate at which they form is a kinetic problem. Because of the formation of layers of low porosity that separate metal and gaseous phase, the reaction rates of metals vary appreciably. Corrosion resistance in metals therefore means that the oxidation rates under the conditions in which they are used in practice are low.

4.1.2. Kinetics of Scaling

When a clean metal surface is subjected to attack from an oxidizing gas X_2, e.g., oxygen, nitrogen, or chlorine, an oxidized film forms between gas and metal, provided that the reaction product is not fragile or volatile and the oxygen is not completely soluble in the structure. This film is usually pore-free and gas-tight, and separates the reaction partners M and X_2. However, the reaction does not come to a standstill at elevated temperatures. Thus, at least one of the reaction partners must diffuse through the MX film. The mechanism and the rate of diffusion of the reaction partners through the reaction products are decisive factors in the reaction of metals with gases. They are closely linked with imperfections in the MX layer.

All crystalline solids exhibit defects in and departure from the ideal lattice structure, particularly at elevated temperature. Either lattice points remain unoccupied (vacancies) or lattice elements deposit between the regular lattice points (interstitial lattice points). These point defects determine material transport in a solid. In addition, there are a number of linear and face defects (dislocations, grain boundaries, etc.), which although of importance to the mechanical properties of a solid are less significant for material transport.

When metals react with gases, the main corrosion products are ionic compounds which can be stoichiometric or nonstoichiometric. Generally, only defect ions (ion conductors) arise in stoichiometric compounds (e.g., AgCl, NaCl). Four border cases of imperfections are possible: When cation vacancies in the lattice and cations at interstitial lattice sites are found in an undisturbed anion lattice, the cations are mobile. In the opposite case the anions are mobile. In compounds with anion and cation vacancies both can migrate, as they can when an equal number of cations and anions are present at interstitial lattice sites.

Nonstoichiometric compounds show electron defects in addition to ion defects because of electroneutrality. In chemical terms this can be equated to the occurrence of ions of different valency (e.g., Fe^{3+} in FeO). Since the electrons are more mobile than ions by a factor of $10^3 - 10^6$, these compounds are practically pure electron conductors (semiconductors). They can be divided into electron-surplus conductors, and electron-deficient conductors, according to whether the electric current is transported by free electrons or by defect electrons.

The mechanism of metal oxidation can be divided into an initial phase and a phase of diffusion-controlled scaling. During the initial phase, reaction steps occur that are largely controlled by processes taking place at the phase interfaces, such as:

1) cleavage of the molecules of the oxidizing medium into atoms and chemisorption;
2) formation of a nonepitactic, polycrystalline, primary oxide film;
3) occurrence of monocrystalline, epitactic nuclei on the metal surface and the spread of these nuclei through lateral propagation.

These steps of the first reaction phase take place rapidly at elevated temperatures and high partial pressures of the oxidizing medium with formation of protective films. After the formation of thin layers, a protective film propagates, as is observed in the scaling of Fe, Ni, and Co at moderate and elevated temperatures, in the form of a dense layer. The reaction can only then continue if at least one of the reactants can diffuse through the scale. At elevated temperatures, positive partial pressures of the scaling atmosphere, and a certain layer thickness, the phase boundary reactions are generally rapid. Therefore, transport processes such as the diffusion of oxygen to the metal or the transport of metal through the initial layer or scale, become the slowest and hence the rate determining process (Fig. 69). After layer formation the oxidation reaction thus continues by way of diffusion of the metal or nonmetal in the form of ions, electrons, or defect electrons through the film.

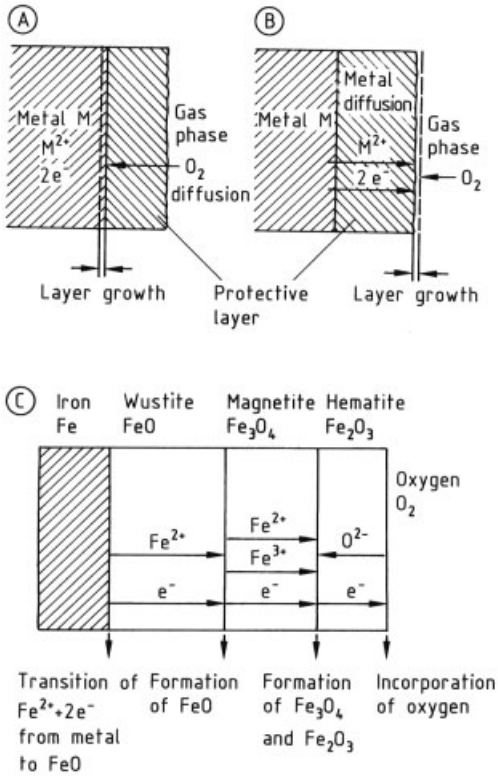

Figure 69. Transport processes and layer growth
A) Layer growth with preferential oxygen diffusion; B) Layer growth with preferential metal diffusion; C) Diffusion processes and phase interface reactions for the oxidation of iron at $> 570\,^{\circ}\mathrm{C}$

Time Laws. For the formation of very thin films (about 10 nm) at low temperature, logarithmic, reciprocal-logarithmic, or cubic time laws can be deduced. If diffusion determines the rate of formation of thick and compact scale, a parabolic time law applies

$$\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{k'}{x} \tag{14}$$

which in integrated form reads:

$$x^2=2k't \tag{15}$$

where x is the thickness of the scale, t is the reaction time, and k' is the scale constant. If the rate is determined by phase boundary reactions a linear law applies

$$dx/\mathrm{d}t=k \text{ or } x=kt \tag{16}$$

where k is a constant. Knowledge of the rate-determining step is needed if the oxidation rate is to be influenced. This can be determined from detailed kinetic measurements. As the scale increases in thickness, however, diffusion in the scale will at some point become the rate-determing partial step. The mechanism of scale formation also determines whether a crack in the film surface layer can mend or not. If metal ions and electrons migrate outwards, the oxide film propagates at the oxide – atmosphere interface and the crack quickly closes. But when oxygen ions migrate inwards the oxide film propagates at the metal – oxide interface and the crack cannot close.

The mechanical properties of the oxide film, in particular the formability and state of stress, determine the occurrence of cracks. If the formability of the film is poor the contact between metal and oxide is lost as long as ion migration of the metal is from the inside to the outside. Oxygen migration from the outside to the metal – oxide interface leads to formation of a new oxide scale, which propagates into the metal.

Wagner's Theory of Metal Oxidation. Assuming that only ions and electrons migrate in the scale and not neutral atoms, Carl Wagner established a formula in the 1930s whereby the parabolic scale constant of a pure metal can be calculated from the free formation enthalpy of the corrosion product, the electrical conductivity of the protective layer, and the transport numbers of cations, anions, and electrons in the film [128]:

$$k=\frac{300}{96500N\cdot\varepsilon}\int_{p_{x_2}^{(\mathrm{i})}}^{p_{x_2}^{(\mathrm{a})}}\frac{1}{Z_2}\,(U_1+U_2)\,U_3x\,\mathrm{d}\ln p_{x_2} \tag{17}$$

where:

k = rational scale constant in mol $\mathrm{cm}^{-1}\,\mathrm{s}^{-1}$
N = Loschmidt's number
ε = elementary charge
Z_2 = charge of the nonmetal in the protective film
U = transport number with index 1 for cation, 2 for anion and 3 for electron ($U_1+U_2+U_3=1$)
x = electrical conductivity in $\Omega^{-1}\cdot\mathrm{cm}^{-1}$

p = partial pressure of the nonmetal X_2 with index i at the interface M/MX and index a at the interface MX/X_2.

The rate of oxidation with rate-determining diffusion in the scale is calculated from this equation. This is the maximum oxidation rate for a gas-tight protective layer. The equation applies for the oxidation, sulfurization, nitriding, and halogenation of a metal ($X_2 = O_2$, S_2, N_2, halogen). Two border cases can be differentiated with the equation. For $(U_1 + U_2) \simeq 1$ (ion conductors), k is determined by the parameter U_3, the electron transport number. For $U_3 \simeq 1$ (electron conductors), k is determined by $(U_1 + U_2)$.

4.2. Oxidation of Alloys

Whereas the oxidation rate of pure metals for the case of rate-determining diffusion of ions and electrons in the protective film can be predicted, this is not possible for the oxidation of alloys. It will probably never be possible to predict the oxidation rate of alloys because too many effects can overlap. The following effects must be considered in the oxidation of alloys:

1) the influence of the base metal,
2) composition of the alloy,
3) formation enthalpy of the oxides, sulfides, and halides in question,
4) effect of dissolved oxides of the alloy components on the disorder of protective oxide or sulfide films,
5) formation of mixed oxides,
6) formation of ternary oxides or sulfides,
7) diffusion within the alloy,
8) solubility and diffusion of oxygen in the metal phase (internal oxidation),
9) relationship between the formation rate of different oxides.

Adding more noble metals has little influence on the oxidation of the base metal. The more noble metal will accumulate in metallic form, e.g., Cu in steel (danger of red shortness), at the alloy – scale interface. Higher contents can lead to a marked retardation of oxidation because diffusion within the alloy could become the slowest partial step.

If the oxidation rate of a metal is to be changed appreciably this can only be achieved by adding less noble metals. In this case the oxide of the alloying element is more stable than that of the base metal. The oxidation rate will be lower if the concentration is high enough for the formation of a sealed oxide film, and the diffusion rate of the ions and electrons in this oxide is lower than in the oxide of the base metal. Examples of such oxides are Cr_2O_3, Al_2O_3, and SiO_2. The metals of these oxides provide the scale resistance of heat-resistant alloys. The metals Ag, Cu, Ni, and Fe can dissolve oxygen. If these metals contain small amounts of less noble metals (Al, Cr, Si, Ti) these can be oxidized within the metal phase (internal oxidation). Internal oxidation does not occur as long as the concentration of the alloying elements is so high that a sealed oxide film of the alloying element forms on the surface.

4.3. Sulfurization of Metals and Alloys

The reaction of metals and alloys with sulfur is basically identical to oxidation with oxygen. Thus, Wagner's formulae are also applicable here. However, the sulfurization rate of most metals is far higher than the oxidation rate in air or oxygen because of the following:

1) Most sulfides show far greater disorder than the corresponding oxides. Therefore, substance transport in them is greater.
2) Some sulfides form relatively low-melting eutectics with metals. The protective effect of the film is thereby lost. Examples of this are:

Fe – FeS	*mp* 965 °C
FeO – FeS	*mp* 940 °C
Fe – FeO – FeS	*mp* 925 °C
Co – Co_4S_3	*mp* 880 °C
Ni – Ni_2S_3	*mp* 645 °C

The low melting point of the Ni – Ni_2S_3 eutectic is the reason for the high sensitivitiy of high-nickel materials to sulfur. The resistance to attack by sulfur can be improved by adding chromium. However, the protective effect of chromium is not as high for attack by S_2 as it is for attack by O_2 (see Figs. 70 and 71, next page) [129, 130].

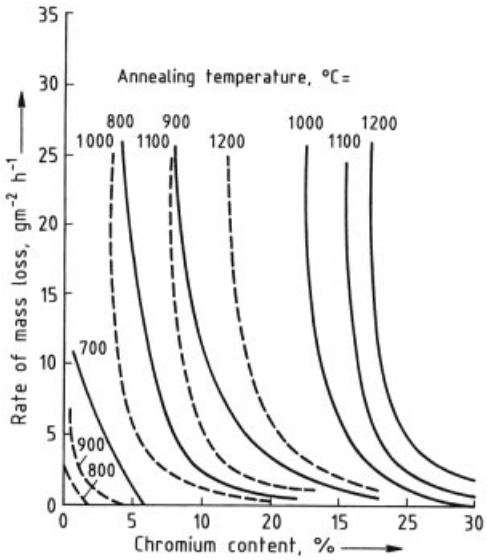

Figure 70. Influence of the silicon and chromium contents on the scaling resistance of steel during annealing in air for 120 h
—- 0.5 – 1.0 % Si; – – – 2.0 – 3.0 % Si

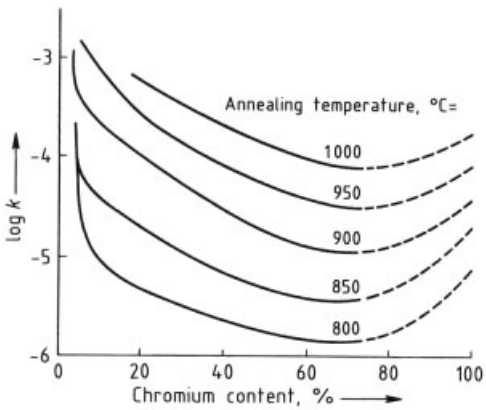

Figure 71. Dependence of the sulfurization constant k of steels on the chromium content in the temperature range 800 – 1000 °C
$k = \left(\frac{\Delta m}{A}\right)^2 \cdot \frac{1}{t}$
Δm = Mass increase, g; A = Surface area, m^2 ; t = time, min

4.4. Corrosion by Combustion Gases [131]

The effect of combustion gases from fossil fuels on metals depends on whether

1) the fuel has been burnt completely (air surplus) or not (air deficiency),
2) the fuel contains sulfur,
3) the combustion gases contain ash or dust particles.

The effect on the scaling behavior of steel of air surplus and deficiency from sulfur-containing fuels and sulfur-free fuels is shown schematically in Figure 72. In sulfur-free gases, oxidation with surplus air is somewhat greater than in incompletely burnt gases. In gases containing sulfur the situation is reversed. In combustion gases containing oxygen, the sulfur content of the gas has practically no influence, whereas the scaling rate increases in incompletely burnt gases [132]. The behavior depicted in Figure 71 is observed in many cases. To understand this behavior, thermodynamic and kinetic considerations are necessary.

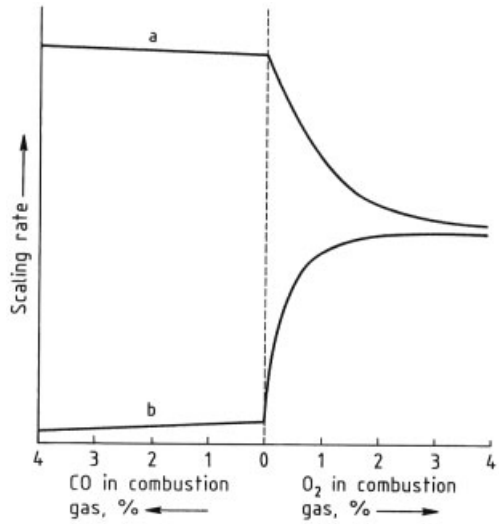

Figure 72. Scaling of soft steel at 850 °C in combustion gases from sulfur-containing and sulfur-free fuel
a) 0.15 % SO_2; b) SO_2-free

The isothermal section through the Fe – O – S system (Fig. 73) shows how the stability of the individual phases changes as a function of the oxygen and sulfur pressure. For a given sulfur content of a gas, the oxides are stable at high oxygen pressure; at low oxygen pressure, however, the sulfides are stable. That is why oxides are stable in fully burnt gases containing sulfur, and sulfides are stable in incompletely burnt gases because of their very low oxygen

pressure. These thermodynamic considerations alone, however, are not sufficient to avoid sulfide formation in gases containing sulfur and oxygen. Kinetic measurements show that where oxides exist, both oxide and sulfide form simultaneously if scaling does not follow a parabolic time law, i.e., if the diffusion of ions and electrons in the oxide film is not the slowest partial step [131].

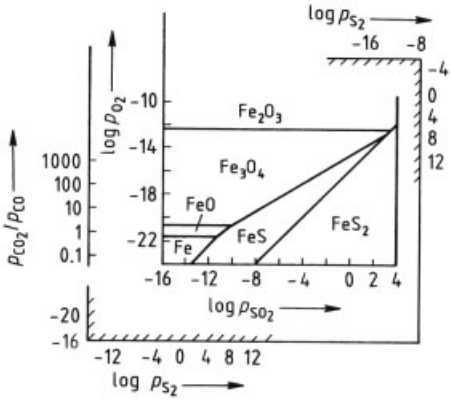

Figure 73. Isothermal section through the Fe – O – S system at 700 °C

4.5. Effect of Ash, Deposits, and Salts

Many corrosion problems arise through deposition of ash and dust on the metal surface. These deposits can react at elevated temperatures, even in the solid state, with the protective films of metals and alloys to form new compounds with different transport properties. However, corrosion becomes particularly intense when fusible phases form because either the deposits contain components with relatively low melting points, or low-melting eutectics arise between the deposits and protective films. A well-known phenomenon is oil ash containing vanadium (Fig. 74) [133], which in the presence of surplus oxygen creates a critical oxidation rate. Other oxides such as PbO, MoO_3, and B_2O_3 can also form low-melting eutectics. Alkali-metal sulfates are also often responsible for increased corrosion (Fig. 75) [134]. The reactions are extremely complex. Since most salts and oxides in molten state are dissociated into ions, electrochemical processes, similar to those in aqueous solutions, take place. These reactions can therefore be investigated with the same electrochemical methods as are employed to study corrosion in aqueous media (Fig. 76) [135].

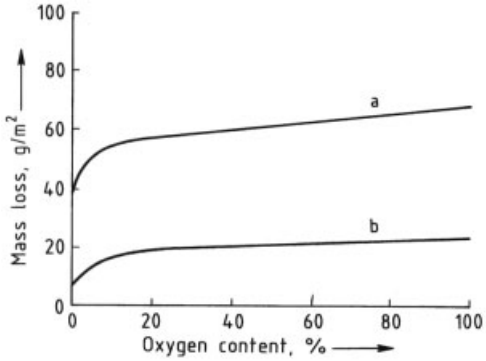

Figure 74. Dependence of the mass increase of steel X 20 CrNiSi 25 4 (material no. 1.4821) after annealing for 6 h in V_2O_5-containing oil ashes on the oxygen content of the gas
a) Synthetic oil ash (V_2O_5 : Na_2O = 3), annealing temperature 800 °C; b) Natural oil ash (V_2O_5 : Na_2O = 1), annealing temperature 700 °C

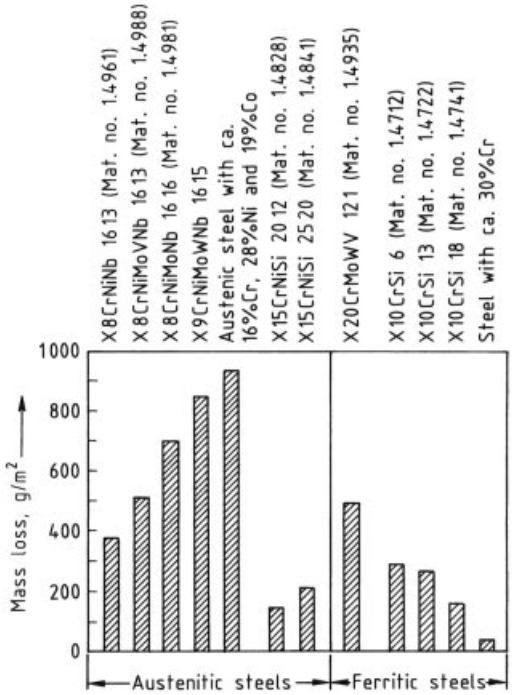

Figure 75. Corrosion rates of various high temperature and heat-resistant steels embedded in pure potassium sulfate after annealing for 700 h at 650 °C in an air – steam environment containing SO_2 (dew point 165 °C)

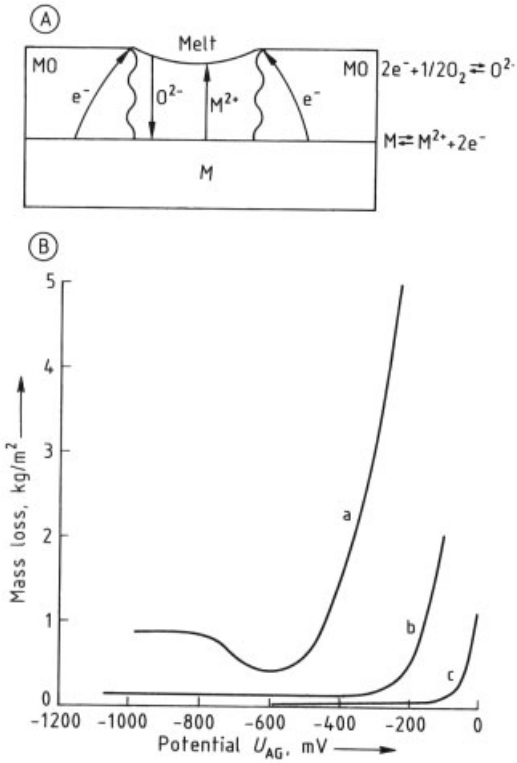

Figure 76. High-temperature corrosion in salt melts
A) Schematic of corrosion; B) Dependence of the corrosion in salt melts on the potential (at 750 °C for 23 h)
a) 10 CrMo 9 10; b) X 10 CrAl 18; c) X 10 CrAl 24

4.6. Corrosion by Synthesis Gases

In high-pressure synthesis the corrosive effect of hydrogen is not the only problem: there is also the problem of corrosive gas mixtures involved in the reaction. Their reaction products must also be taken into account. In ammonia synthesis, for example, there is the additional aspect of the nitriding effect of ammonia, which results in the formation of a nitride layer and embrittlement of the material beneath this layer. The behavior of various steels in an ammonia – hydrogen gas mixture is shown in Figure 77 [136].

After Armco iron, austenitic steels are least prone to corrosion. Ammonia has a nitriding and embrittling effect on unalloyed and lowalloy steels at 250 °C, and in austenitic steels corrosion begins at ca. 350 °C (Fig. 78) [137]. However, a standard 18-8 CrNi steel shows only a very shallow nitriding depth, and experience has proven that this steel can be used for years in the ammonia cycle-gas at 450 °C. After ten years the penetration depth of nitration is 1 mm at most.

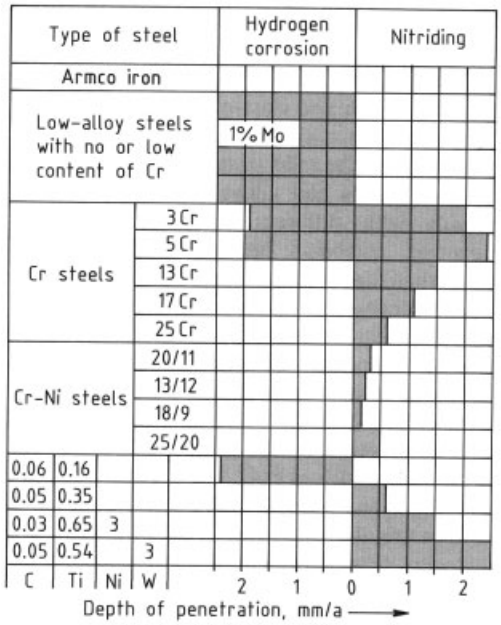

Figure 77. The effect of a $H_2 - NH_3$ gas mixture on various steels ($t = 450 - 520$ °C; $p = 32.5 - 90$ MPa)

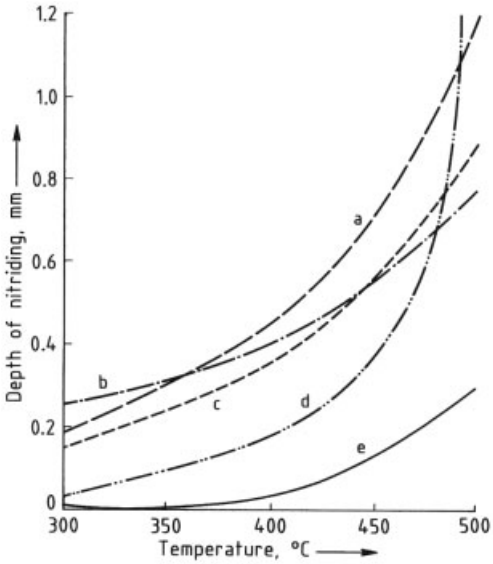

Figure 78. Depth of nitriding of various steels following nitridation in ammonia for an annealing time of 300 h
a) 13 CrMo 4 4; b) 17 CrMoV 10; c) 10 CrMo 9 10; d) X 8 Cr 17; e) X 8 CrNiNb 16 13

Carbon monoxide is also important in that it is frequently used as synthesis gas, e.g., in methanol synthesis. Under pressure, carbon monoxide attacks unalloyed and low-alloy steels above 130 – 140 °C, forming iron pentacarbonyl. Above 350 °C, corrosion practically ceases again because the carbonyl becomes unstable. In high-alloy chromium and chromium – nickel steels the damage is appreciably less. Chromium steels with 30 % Cr and austenitic steels with 25 % Cr and 20 % Ni are completely stable. Figure 79 shows the corrosive effect of CO on various steels in a short-term test with pure $CO - H_2$ mixtures at two CO partial pressures [138]. A striking feature is the shift in the corrosion maximum of the 18-9 CrNi steel at the higher CO partial pressure. This possibly corresponds to the beginning of nickel carbonyl formation, which after a period of time leads to a chromium-enriched and, therefore, more stable surface layer.

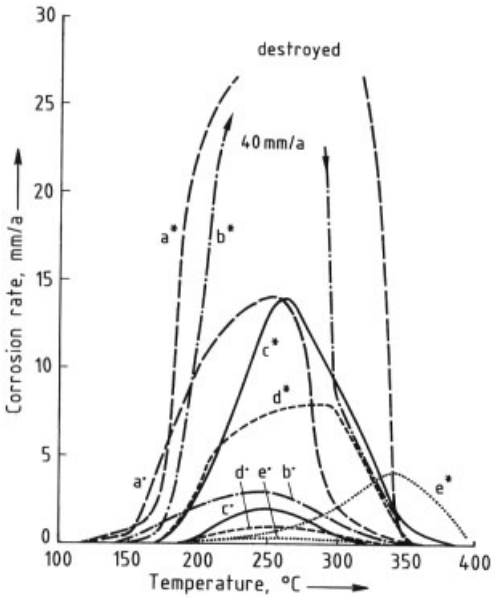

Figure 79. Corrosion behavior of various steels in $CO - H_2$ gas mixtures at two CO partial pressures
a) Boiler plate; b) 3.3 Cr/0.3 Mo; c) 13.1 Cr/0.34 Mo; d) 16.2 Cr; e) 18-9 CrNiCO partial pressure: · = 12.5 MPa; ∗ = 35 MPa

If the iron carbonyl produced on CO attack is carried into the synthesis cycle, carbonyl decomposition can occur at elevated temperatures with the formation of active, pyrophoric iron, which in turn catalyses thereaction

$$2\,CO \longrightarrow CO_2 + C$$

Since this reaction is highly exothermic, individual apparatus parts can overheat [139]. In this way expansion has occurred in high-pressure vessels; the carbonyl produced must therefore be removed.

4.7. Oxidation Behavior of Technical Alloys

4.7.1. Steels

The scaling of steels in air has so far been investigated more thoroughly than in other corrosive media. Behavior in air is often the basis for a comparison of steels with respect to their applicability. However, since the corrosion conditions in technical gases can differ appreciably from those in air, extrapolation of parameters determined in air to the service behavior of steels is only possible to a limited extent.

The time function of oxidation is described by the equation:

$$\Delta m^n = kt$$

where Δm is the mass loss or increase, t the time, k the scale constant, and n a dimensionless exponent. The exponent n indicates the type of rate-determining reaction step and gives information on resistance. Values of $n \geq 2$ generally indicate resistance capacity, and values < 2 often imply lack of resistance. If n and k are known, metal losses can be calculated for any corrosion time. However, n is frequently variable because of transition from resistance to lack of resistance (breakdown), e.g., through a change in the oxide composition of alloy steels. Unalloyed steels can be used in air up to 550 °C and low-alloy steels up to ca. 600 °C. The applicability of high-alloy steels is determined by the alloy contents, with special importance attached to Cr, Si, and Al, as demonstrated in Figures 70, 80, and 81 [140]. Water vapor and carbon dioxide in air generally worsen the scaling behavior of steels. The resistance of steels in water vapor is of particular importance in steam boilers and heat exchangers. To date it has been investigated at temperatures up to 800 °C.

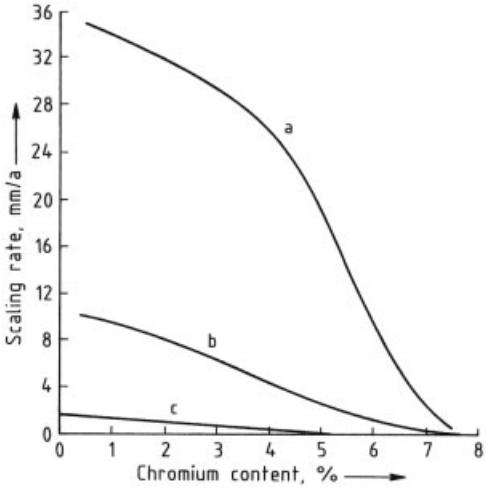

Figure 80. Influence of the chromium content of steels containing 0.15 % C and 0.7 – 0.9 % Si on the scaling resistance in air
Annealing temperature, °C: a) 800; b) 700; c) 600

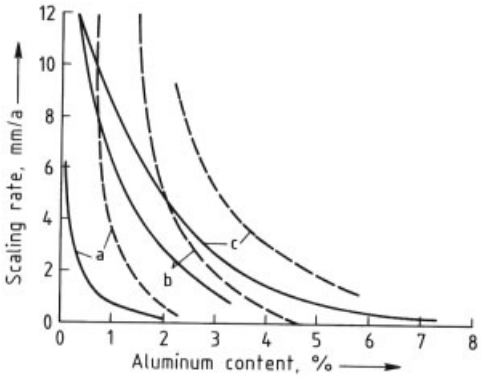

Figure 81. Influence of the Al and Cr contents on the scaling resistance of steel when annealed in air for 120 h
a) 6 % Cr; b) 2 % Cr; c) 0 % Cr
—— 600 °C; – – – 700 °C

Scaling in water vapor on low-alloy steels is similar to that measured in air at a temperature some 50 – 100 °C higher. This applies to high-alloy steels even at elevated temperatures. Of significance here is the shift of breakdown towards greater scaling: the breakdown in low-alloy steels can be about 50 °C lower and in high-alloy steels up to 150 °C lower. The improved scaling behavior provided by Cr, Si, and Al corresponds to that obtained in air. A positive effect has also been found for molybdenum. In steels with higher chromium content, nickel makes a pronounced improvement.

Resistance to carbon dioxide is required of certain components in gas-cooled nuclear reactors, e.g., in heat exchanger tubing and fuel element cladding, where temperatures up to 800 °C exist. Data on high-temperature corrosion in carbon dioxide provide further indication of the performance of heat-resistant steels in exhaust gases arising from combustion of pure gases, especially when corrosion tests are run in carbon dioxide containing water vapor. Basically the same time laws apply to scaling in carbon dioxide as to scaling in air, but with carbon dioxide, graphite deposits in the scale can occur in unalloyed and low-alloy steels, and carburization in austenitic steels. The penetration of carbon

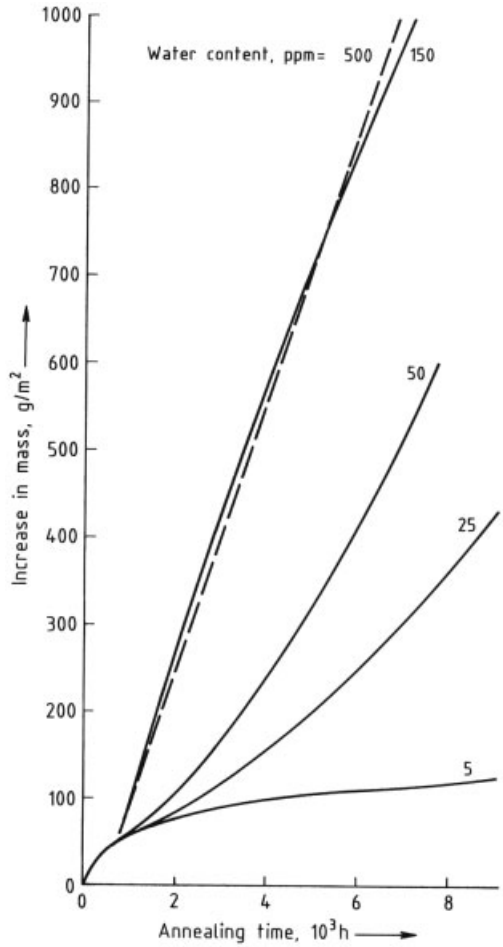

Figure 82. Scaling behavior of a steel similar to 13 CrMo 4 4 and dependence on the water content of carbon dioxide at 500 °C

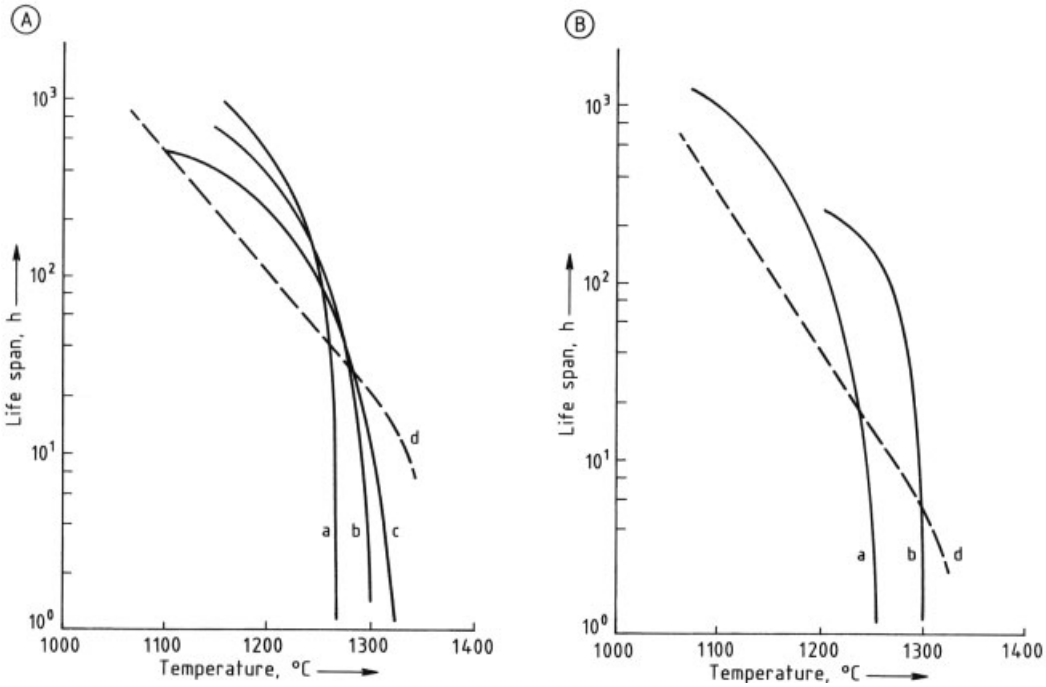

Figure 83. Dependence of the lifespan of Ni – Cr alloys on temperature in various gases
A) NiCr 80 20; B) NiCr 30 20 (Fe-containing)

Curve	Gas composition, vol %					
	H_2	CO	CO_2	CH_4	N_2	O_2
a	40	20			40	
b	20	13	7	2	58	
c	25	7	4	3	59	2
d			(air)			

into an alloy can lead to the following processes, which alter properties and reduce scaling resistance:

1) lower melting point (by ca. 350 °C in Ni – Cr alloys);
2) carbide formation with chromium depletion: cubic $Cr_{23}C_6$, triclinic Cr_7C_3, and orthorhombic Cr_3C_2;
3) formation of a brittle, intermetallic σ-phase.

Graphite contents in the scale lead to a loosening of the oxide film and accelerate the oxidation rate.

The addition of carbon monoxide to carbon dioxide reduces the scaling rate. The effect is particularly marked in the low-temperature region in unalloyed and low-alloy steels. Water content, on the other hand, has a negative effect. In addition to high carbon dioxide pressure, it is the cause of premature breakdown, i.e., the formation of noncovering oxide films, whereby in addition to the time of breakdown, the scaling rate after breakdown is also influenced. For dry gas ($< 5 - 10$ ppm H_2O) scaling follows a parabolic time law; higher water contents (see Fig. 82, previous page [141]) give practically straight lines, with gradients that increase with increasing water content. Goodison et al. found in low-alloy steel at 420 °C and 2.9 MPa gas pressure the following relation between water content and scaling rate after breakthrough [142]: mass increase in g/m^2 after 1000 h = 5 × water content in ppm. In contrast, methane and hydrogen contents delay break-

through in unalloyed and low-alloy steels. Niobium and small amounts of rare earths (ca. 0.1 % cerium or yttrium) also have a positive effect on the scaling resistance of steels.

4.7.2. Nickel Alloys

Nickel alloys with more than 15 % Cr show good oxidation properties, which are due largely to the presence of an outer Cr_2O_3 layer. The adhesive strength of oxide films can also be greatly improved by small additions of cerium, lanthanum, and calcium (ca. 0.1 %) (Ni – Cr heat conducting alloys). The influence of temperature and atmospheric composition on the performance of Ni – Cr and Ni – Cr – Fe heat conducting alloys is shown in Figure 83 (see previous page) [143]. The sharp drop in life span is caused by the formation of liquid carbide eutectics through carbon absorption. Generally, higher application temperatures than in air are possible, as is shown by the curve drawn for performance in air.

Nickel has a high affinity for sulfur and combines with it to form the eutectic Ni – Ni_3S_2, which melts at 645 °C. However, even at ca. 100 °C below this temperature, annealing in sulfurous gases leads to penetration of sulfur into the nickel, especially along the grain boundaries (Fig. 84).

In reducing gases, intergranular damage due to sulfur (H_2S) is more pronounced than in oxidizing gases (SO_2). Generally, the resistance of nickel – chromium alloys to sulfur attack increases with increasing chromium content. Suitably good values are achieved with alloys NiCr 50 50 and NiCr 60 40.

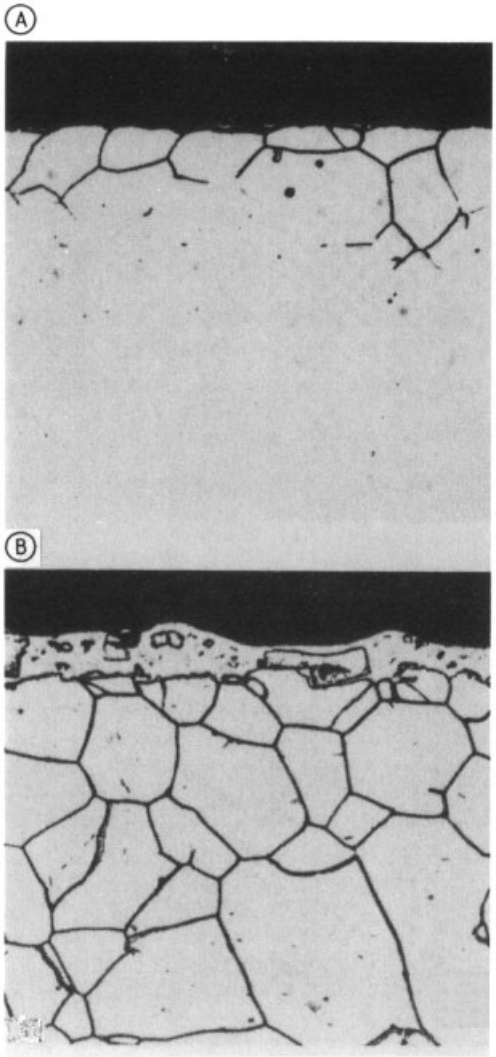

Figure 84. Cross section of a nickel sheet after annealing in nitrogen containing 0.5 % sulfur dioxide
A) Annealing at 600 °C; B) Annealing at 640 °C

4.7.3. Cobalt Alloys

Oxidation of Co – Cr alloys is in many ways similar to that of Ni – Cr alloys. The rate of cobalt oxidation is faster than for nickel. Under equivalent conditions CoO is even richer in cation vacancies than NiO.

Small Cr contents increase the rate of reaction, but at 20 % Cr the reaction rate starts to decrease and exhibits a minimum at 25 – 30 % Cr (see Fig. 85). The minimum value depends on the pressure. More chromium is needed to stabilize a protective film since the diffusion coefficient of chromium in cobalt is lower than for chromium in nickel. However, since the adhesion strength of the film on Co – Cr alloys is poorer than on Ni – Cr alloys – despite the identical oxidation rate of the Cr-containing Co alloys with Cr_2O_3 protective film – the practical oxidation resistance is lower. Other alloying elements, as Figure 85 shows, have little influence on scale resistance [144].

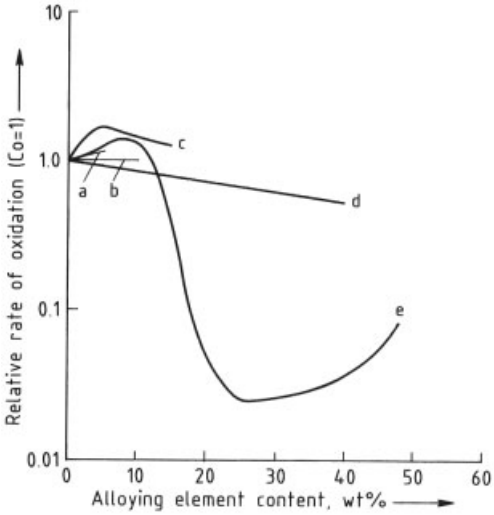

Figure 85. Influence of alloying elements on cobalt oxidation at 1000 °C
a) Al; b) Mo; c) W; d) Ni; e) Cr

5. General Corrosion Protection Procedures

5.1. Corrosion Protection by Coating

The best possible corrosion protection can be achieved through the choice of suitable materials (see → Construction Materials in Chemical Industry). Given extensive corrosion attack, this can mean compulsory use of expensive materials that may not always prove satisfactory in mechanical terms. A way out of this dilemma is provided by a composite system: the underlying function of such a system is provided by a cheap base material with good mechanical properties while the corrosion resistance is imparted by a coating material with relatively low thickness.

5.1.1. Metal Coatings

Of the various procedures for forming metal–metal composite systems, cladding is the most significant. Materials of greater thickness are combined by various processes with the base material. Compared with cladding, lining or the other coating methods are of minor importance. A distinction can be made between the following:

Cladding: roll cladding, explosive cladding, weld overlay cladding;
Lining;
Coating: immersion, electrolytic metal deposition, diffusion deposition, hot dip metal coating, evaporation coating, and metal spray coating.

Roll Cladding. Strip and sheet metal are mainly produced by this method. Prerolled base and coating materials with clean surfaces are heated to the required temperature and rolled in contact with one another. Strength-reducing oxidation in the bonding zone can be avoided either by rolling the base metal/cladding sandwich within a hermetically sealed envelope of low-cost steel or by interleaving layers of nickel or mild steel between the base and cladding. Steel as a base metal can be clad with stainless steels, silver, nickel, copper, aluminum, and their alloys. Roll-clad semifinished products can be further processed by well-established fabrication methods, although care must be taken in hot forming and heat treatment.

The process for cladding tubes is carried out in two stages: backward impact extrusion of a multilayer solid followed by forward extrusion of the tube blank [145]. Forming with high-pressure hammers facilitates the binding process [146]. For unalloyed steel as the base material, stainless steel and titanium have so far been examined as cladding materials; interleaving sheaths such as Cu–Ni–Fe alloys are required when cladding with titanium.

Explosive Cladding. By using high-energy shock waves, composite systems can be produced from materials which for metallurgical or technological reasons can be neither roll-clad nor weld overlay plated [147]. In this process, the schematic for which is given in Figure 86, the cladding material impacts the substrate at high speed. If the collision angle reaches a critical value, a liquid metal jet, which is essential for a solid bond to form, is ejected from the contact region. Practically all metals can be combined with one another by explosive cladding. Because of the small depth of the binding zone, the mechanical properties are generally not impaired. While cladding of flat or tubular products is the state of the art, problems still arise when

using this method to produce other curved composite systems.

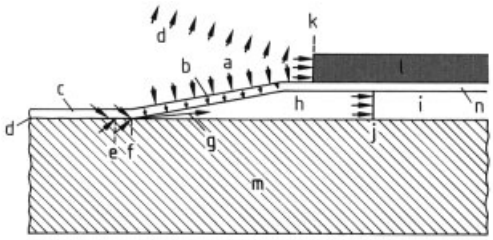

Figure 86. Schematic of explosive cladding
a) Expanding detonation products (swathes); b) Accelerated coating sheet; c) Cladded coating sheet; d) Undulating binding zone; e) Metal interface layers flowing in a laminar or turbulent manner under high pressure; f) Collision point, very high pressure (1000 – 10 000 MPa); g) Jet of material at high speed; h) Air in a collision state; i) Air at atmospheric pressure; j) Advancing collision front in the air; k) Detonation front in the explosive; l) Unused explosive; m) Base plate or substrate; n) Sheet to be clad

Like roll-clad products, explosively clad metals can be further processed by all fabrication procedures, although special welding conditions may be required. The problems of heat treating systems of this kind are exemplified by the properties listed in Table 6 for a combination of fine-grained structural steel and titanium.

Table 6. Heat treatment of fine-grained structural steel and titanium

Treatment	Temperature, °C
WStE 460	
Hot forming	> 800
Normalizing	900 – 940
Recrystallization annealing	> 600
Stress-relieving annealing	530 – 580
Titanium	
Carbon diffusion from steel into titanium	> 550
Iron diffusion from steel into titanium	> 800
Hot forming	> 650
Recrystallization annealing	> 500
Stress-relieving annealing	400 – 500

Joining of semifinished clad products is an important step in the production of components and equipment. This presents no particular problems in roll cladding because base material and coating can be welded seperately. However, to obtain the desired properties of the system the respective recommendations must be observed [148].

Joining explosive-clad semifinished products, on the other hand, presents a number of problems because metal combinations are often involved which cannot be combined thermally. Thus, in such cases design measures must be taken to prevent the substrate and cladding materials from mixing [149].

Weld Overlay Cladding is primarily economical for fairly thick plates and large forgings. On account of their low cost and only slight penetration, two processes are particularly significant: submerged arc welding [150] with strip electrodes and plasma hot-wire welding [151].

In submerged arc welding the strip electrode is fed from a coil into the flux bed and fused in the arc. The width of the weld deposit corresponds approximately to the width of the electrode, and mixing with the substrate depends on the process parameters. The procedure is preferably used to clad stainless steels. Because of burn off and mixing, the electrode must be overalloyed to compensate for the loss of important elements.

In plasma hot-wire deposition welding, an oscillating plasma jet under inert gas melts the surface of the base material. Wires of the cladding material are fed into the molten layer, melt through the passage of current, and form the cladding. The heat input in this process is small. Consequently mixing is very limited, as are thermal influences on the base material. There is practically no burn off, and the method is far more widely applicable than submerged arc welding [152], [153].

There are two nonwelding methods in which the coating material is also in molten form. In homogeneous lead coating, lead is melted with burners and applied to the steel surface. Intermediate coatings, e.g., tin solder, are needed to improve adhesion.

Centrifugally cast composite pipes are used for special applications (e.g., cracking tubes in the petrochemical industry) [154]. Initially the casting of the outer shell is produced in a rotating cast-iron mold. After a suitable cooling period, during which the inner surface of the casting must be protected against oxidation, the melt to form the inner shell of the composite pipe is poured in. Practical experience has been gained with a composite of unalloyed or low-alloy steels with high-alloy, corrosion-resistant grades.

Lining. A further possibility of protecting equipment and components against corrosion

consists of lining with corrosion-resistant materials. This lining is generally attached to the base material (as long as a thick-walled self-supporting liner is not preferred) by local welding. In this case strips of the lining material are welded to the base material and the edges of the lining material welded to these strips. If the two materials cannot be welded to each other, a form-fitting bond is used, whereby the lining material is mechanically locked into conical slots in the base material. This corrosion protection measure has declined: one reason is the design problems presented by a lining subjected to alternating thermal stress.

A special lining method is the Resista Clad process, which allows thin liners (0.15 – 2.5 mm) to be applied during manufacture. The base and lining materials (e.g., steel, nickel alloys, titanium, tantalum) are attached to one another by resistance welding with an intermediate layer. Depending on the service conditions, the distance between the weld seams can vary from 20 to 150 mm. Apparatus such as vessels, columns, and heat exchangers can be operated in vacuum (seam distance ca. 20 mm). Depending on the coating material, service temperatures of up to 350 °C are possible [155–157]. Apparatus and equipment manufactured by the Resista Clad method have also been successfully used in flue gas desulfurization [158].

Coating. With regard to the protection they offer against corrosion, the methods listed above for producing coatings [159] are generally only of minor importance in chemical engineering because they cannot guarantee adequate resistance under the extreme conditions encountered in some areas of the chemical industry.

Through immersion, metals such as copper or silver are deposited from an aqueous solution or molten salt onto the material to be protected. The coating thickness is ca. 1 µm and the adhesion strength is not always satisfactory.

In electrolytic methods, deposition of metals is improved by using a constant or pulsating direct current along the principle of electrolysis [160]. Voltages of 1 – 10 V and current densities of 5 – 100 mA/cm^2 are used. Depending on the base metal, coating metals such as nickel, copper, zinc, tin, chromium, silver, and gold can be applied to the base material either directly or with intermediate layers. Nickel, for example, is deposited on steels both directly and on an intermediate copper layer [161]. A chromium coating on nickel or copper intermediate layers also offers good application possibilities.

Coating with tantalum by molten salt electrolysis [162] has gained importance in chemical engineering. This process has proved successful primarily for fairly small parts since it provides compact, nonporous coatings with good adhesion and thicknesses above 200 µm. Base materials include steels, copper, nickel, and their alloys [163], [164].

In the diffusion coating process the coating material is deposited at elevated temperature from the pulverized or gaseous state. The metal diffuses into the lattice of the base metal with formation of a high-alloy surface zone. To produce zinc or aluminum coatings (sherardizing and alitizing), the part to be coated is embedded in powders of the corresponding metal chlorides. Chromium coatings are deposited by inchromizing from gaseous or molten chromium(II) chloride.

Hot-Dip Metal Coating is being used increasingly in automated, continuous plant, in particular for the production of semifinished products, but also for mass-produced parts. The part is immersed in a molten bath of a coating metal with a comparatively low melting point. This is a common method for coating with zinc, tin, aluminum, and lead.

Evaporation Coating [165] involves the production of very pure, nonporous surface coatings by condensation of vacuum-evaporated metallic materials (physical vapor deposition, PVD). However, the extremely low coating thickness permits application of evaporation-coated parts in chemical engineering only in exceptional cases.

Metal Spray Coating involves wire melt spraying and powder spraying from gas, arc spraying, or plasma spraying torches. Powder spraying is important for metals (and several nonmetals) that cannot be made into wires. Common to all these processes are the melting of the coating material in the heat source as it leaves the torch, the atomization of the coating material, and the impinging and adhesion of the metal droplets on the bare metal surface of the object

to be coated. As the adhesion of sprayed metal coats is entirely mechanical, the adherence depends very much on the physical nature of the substrate. A roughened and clean (shot-blasted) surface is therefore a prerequisite for good adherence.

The most important materials used for chemical plant include metals with high melting points and good resistance to corrosion, such as tantalum and molybdenum, but also titanium and nickel alloys, including hard metal alloys.

Alloys based on Ni – Cr and Co – Cr, containing borides, silicides, and carbides, and applied by powder flame spraying, are used to protect components from corrosion and wear. The inherent pore volume of a coating can be reduced to about 1 % by a subsequent densifying heat treatment [166]. Figure 87 shows the structure of a flame-sprayed and subsequently densified hard metal coating consisting of Ni – Cr – Si – B. The Si and B form fluxes by alloying and lower the melting point. Exceptional hardness (60 – 70 R_C) and good adherence can be obtained when W_2C is introduced into the surface of the substrate. A conveyor screw made of mild steel (St 37) protected by Ni – Cr – Si – B provides an example of the improvements in durability achievable by these techniques. The useful life in this instance was increased 12-fold. The life of an iron sludge pump casing spray-coated with a W_2C/Ni – Cr – Si – B alloy, which was subsequently densified at 1050 °C was extended 10 times. Note that subsequent thermal densification may distort the component and alter the structure of the material.

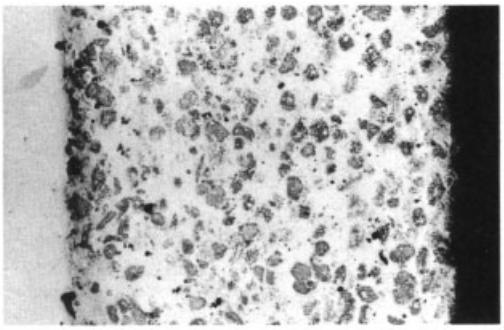

Figure 87. Microstructure perpendicular to the surface of a flame-sprayed and densified coating of a Ni – Cr – Si – B alloy

A prerequisite for good resistance to corrosion is that the protective coating must be dense and free from pores. Flame-sprayed coatings, with their low density and numerous oxide inclusions, are therefore unsuitable in many cases for exposure to aggressive media. Plasma-sprayed coatings are denser and more homogeneous.

Plotting current density against potential curves is a quick way of investigating the electrochemical behavior of thermally sprayed coatings and gaining an initial impression of their resistance to corrosion. Figure 88 shows current density – potential curves for plasma-sprayed Ni – Ta – Cr – B alloys (containing τ-borides) in 0.1 N sulfuric acid. All three alloys show good resistance to corrosion by virtue of their chemical composition and the fact that the coatings are very dense. Figure 89 shows the microstructure of a coating of this type on unalloyed steel. Although alloys forming τ-borides are still under development, they will become increasingly important in the protection of chemical plant components [167].

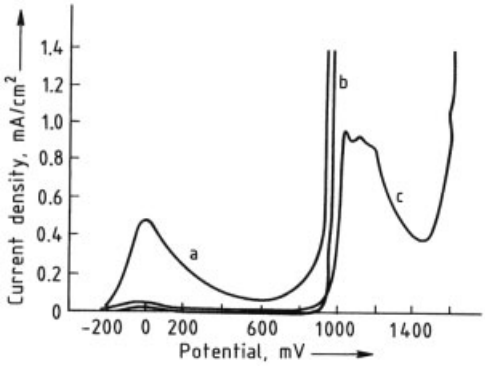

Figure 88. Current density – potential curves for various alloys containing τ-borides in 0.1 N H_2SO_4
a) Ni25Ta8CrB; b) Ni20Cr3TaB; c) Ni38Ta7CrB

The high-velocity flame spraying (Jet-Kote) process, which became commercially available only a few years ago, is similar to powder flame spraying, but differs in that the spray particles are accelerated to high velocities in the gun, with the result that the adherence and density of the coating are superior to those achieved by conventional powder flame spraying. However, some kinds of spray powder cannot be applied by this technique. Figure 90 shows the good adherence and low porosity of a Hastelloy C-4 coating applied by the Jet-Kote process.

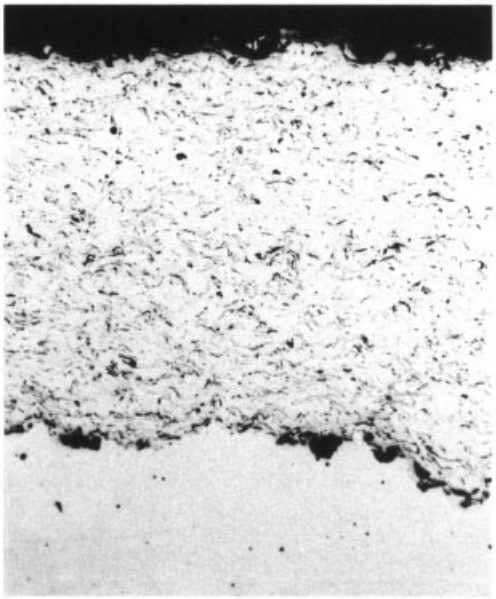

Figure 89. Microstructure perpendicular to the surface of a plasma-sprayed Ni – Ta – Cr – B alloy containing τ-borides

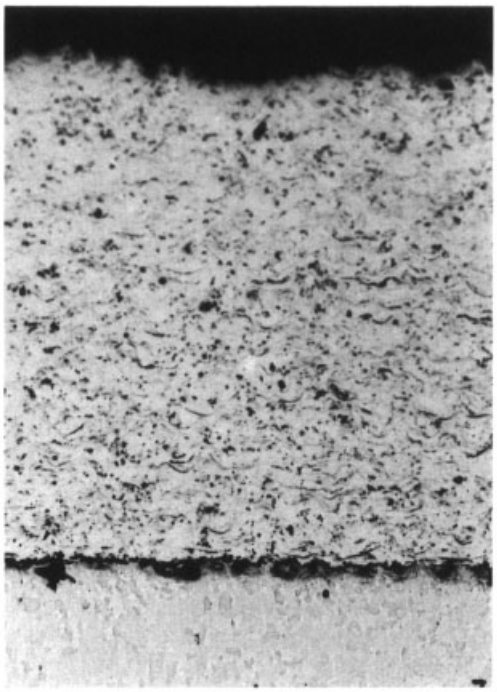

Figure 90. Hastelloy C-4 coating applied by powder flame spraying (Jet-Kote process)

Spray coatings consisting of metals with high melting points are applied mainly by plasma spraying, as used particularly for oxide ceramic coatings. Coat thicknesses of 0.1 to several millimeters are feasible and good adherence is obtained. The coating is relatively dense, but contains individual closed pores. Vacuum or low-pressure plasma spraying is a further development of atmospheric plasma spraying. The process is carried out in a pressure chamber (Fig. 91), which is evacuated to 20 Pa. While the spraying is in progress, argon is introduced until the pressure reaches about 5 kPa. The resulting inertization prevents the reactions between molten particles and the surrounding atmosphere that occur in atmospheric plasma spraying [168]. This modification gives the process substantial advantages over conventional plasma spraying. In particular, the method enables reactive metals such as Ti, Ta, and Nb to be applied as very dense, nonporous coatings. At present the process suffers from the limitations imposed by the dimensions of the chamber and the high cost (€0.75 million).

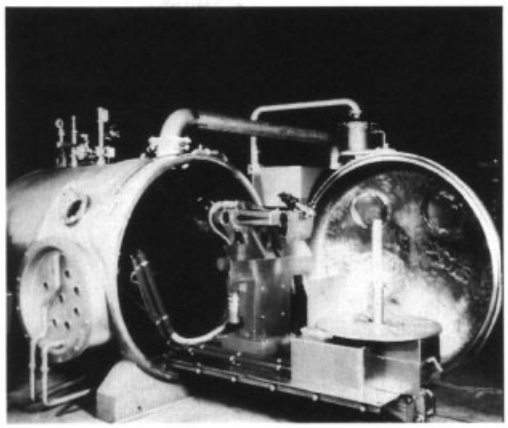

Figure 91. Vacuum chamber for plasma spraying of Ti, Ta, and Nb

5.1.2. Inorganic Nonmetallic Protective Coatings

Important inorganic nonmetallic coatings used for the protection of chemical plant include, above all, enamels with high silicon dioxide and low boric acid contents, brick linings based on silicates and carbon, as well as thermally

sprayed coatings based on metal oxides and metal carbides (see Table 7). Metal nitrides and metal borides are also used, but less often because chemically they are less resistant. Having low thicknesses of $\leq 20\,\mu m$, coatings formed on metal and plastic substrates by chemical and physical deposition from the gas phase (CVD and PVD) are of only minor importance in corrosion protection. The same applies to protective coatings formed by reaction of the metal surface with acidic phosphate, chromate, or oxalate solutions; these are used particularly as passivating primers for paints. Lining with tubes and other shaped parts consisting of fusion-cast basalt [169] has gained some importance as a means of protecting plant from corrosion and abrasion, and the same is true of the use of tungsten carbide platelets for screw cladding [170].

Table 7. Inorganic nonmetallic protective coatings used in chemical engineering

Material	Applications [a]
Enameling (0.8 – 2.2 mm)	
Vitreous or partly crystalline	Acidic and organic media up to 230 °C (340 °C [b]) (except HF and H_3PO_4) Alkaline media up to 60 °C
Brick lining	
Acid-resistant bricks	Neutral, acidic, and organic media up to 200 °C
Carbon bricks and electrographite, impregnated	Nonoxidizing acidic, alkaline, or organic media up to 180 °C
Carbon bricks and electrographite, unimpregnated	Nonoxidizing media up to 400 °C
Refractory bricks	Lining of reaction zones in which temperatures of up to ca. 1600 °C occur
Thermally sprayed coatings	
Metal oxides	Wear protection, neutral and acidic media up to 100 °C
Metal carbides	Wear protection, neutral and alkaline media up to 80 °C
Ceramic – metal mixtures (cermets)	Special applications and intermediate coatings
CVD and PVD coatings	Corrosion-inhibiting wear protection

[a] The indicated temperature limits depend on the agressiveness and concentration of the medium and may in certain cases be too high. If no experience is available, it is advisable to carry out corrosion tests.
[b] With increased compressive prestressing, which extends the temperature range to 340 °C and enhances the resistance to thermal shock [175].

The resistance of protective materials depends mainly on whether they are based on oxides or on silicates, consist of carbon, or have been sealed by impregnation with synthetic resins (see Table 8). In general it can be assumed that the oxide and silicate coating and lining materials used for chemical plant are outstandingly resistant to acids (except hydrofluoric acid and phosphoric acid), salt solutions, and organic compounds of all kinds. As a rule they resist aqueous alkaline solutions up to 50 °C. Unimpregnated carbon materials withstand temperatures up to 400 °C, while those that are impregnated with synthetic resins are stable up to 180 °C. With silicate materials the chemical attack, which occurs almost entirely on the surface, depends greatly on the pH and water content of the attacking medium. Water-free acids, except phosphoric acid, do not attack silicate materials. The resistance of carbon materials depends on the resistance of the resin impregnating agents and putty joints.

Table 8. Main properties of the constituents of the enamel/mild steel system

Property	Enamel	Boiler plate
Adhesive strength, steel/enamel, MPa	100	
Compressive strength, MPa	800 – 1000	2000 (60 % compression)
Tensile strength, MPa	70 – 90	400
Coefficient of thermal expansion (20 – 400 °C), K^{-1}	80 – 95×10^{-7}	135×10^{-7}
Compressive prestress, MPa	130	
Elastic modulus, GPa	70	210
Vickers hardness, MPa	6000	1100
Fracture strain, %	0.15 – 0.3	25
Yield strength, MPa		200
Impact force (DIN 51 155), N	40 – 80	
Coefficient of thermal conductivity, $kJ\,m^{-1}h^{-1}K^{-1}$	3.35	188
Specific heat (10 – 100 °C), $kJ\,m^{-1}h^{-1}K^{-1}$	0.84	0.46
Specific resistivity, $\Omega \cdot cm$	$10^{12} - 10^{14}$	0.002
Dielectric strength (room temperature), kV/mm	20 – 30	
Density, g/cm^3	2.5	7.8
Softening point, °C	790	
Melting point, °C	960	1510
Enamel thickness, mm	1 – 2	

5.1.2.1. Enamel

An enameling is a highly stable combination consisting of enamel (an inorganic oxide-based glass flux) and a metallic substrate. Known in the process industries as chemical enamel, the combination of enamel with steel has for decades been firmly established in chemical processing plants. Its stability is such that many chemical

processes are dependent on it, and many products can only be manufactured economically in enameled apparatus [171]. By virtue of their structure [172] – highly stable silicate glass in a firm bond, created by fusion, with supporting steel – the traditional domain of enamel coatings is processes that involve reactions with acids at high temperatures. However, the development of enamel formulations and enameling technology have considerably extended this traditional field of application: enamel is now being used increasingly as an all-round material for exposure to neutral and alkaline media as well [173]. This applies equally to vitreous enamels and to the partly crystalline ceramic enamels. Table 8 lists the main properties of the partners in the enamel/unalloyed steel system. The stability of this system, represented as iso-corrosion curves for a corrosion rate of 0.1 mm/a, is shown in Figures 92 and 93 for characteristic acids and alkali. Stability tests are performed at atmospheric pressure according to DIN/ISO 2743, 2744, and 2745, and under pressure according to DIN 51 174. The temperature dependence of the corrosion rate obeys the Arrhenius equation. A corrosion rate of 0.1 mm/a is normally the highest at which enameled apparatus can be used economically, and it should not be exceeded, especially in cases where intermediate exposure to alkali is involved, e.g., in neutralization reactions and cleaning processes. Vitreous and ceramic enamels differ little in their resistance. If a glassy cover coating is not present, the chemical resistance of partly crystalline enamels is in some cases considerably lower than that of glassy coatings of comparable composition. Ceramic enamels are considerably superior to vitreous enamels in mechanical behavior, especially impact resistance, and to some extent also in resistance to abrasion. The acid resistance of enamels can be improved considerably by adding small amounts of SiO_2 to the acid, which has a marked inhibiting effect [174].

Special enamels are now available which, for example, are 300 % more resistant to alkali (though less resistant to acids than normal enamels), which show considerably less tendency to attract deposits because they have substantially improved surface quality (owing to an increase in surface tension), or which, because they contain refractory ingredients, can be fired on heat-resistant substrates at temperatures up to 1200 °C [175]. These high-temperature enamels give additional protection from scale formation and corrosion to alloys and metals that are intrinsically heat-resistant, thus simultaneously prolonging the useful lifetime of the plant and enabling it to withstand high temperatures. A notable improvement in the already high standard

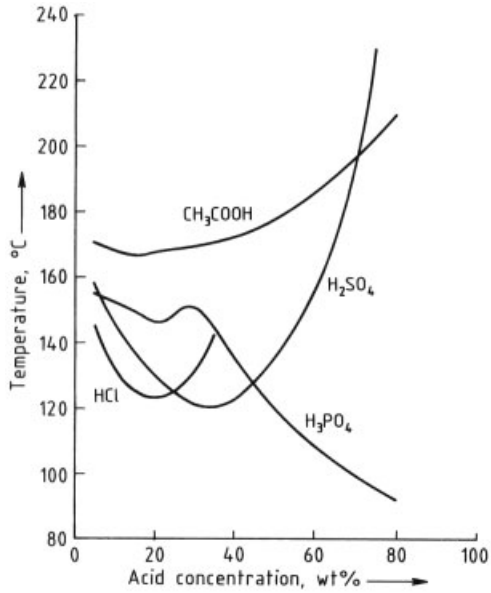

Figure 92. Iso-corrosion curves (0.1 mm/a) for chemical enamel in acids

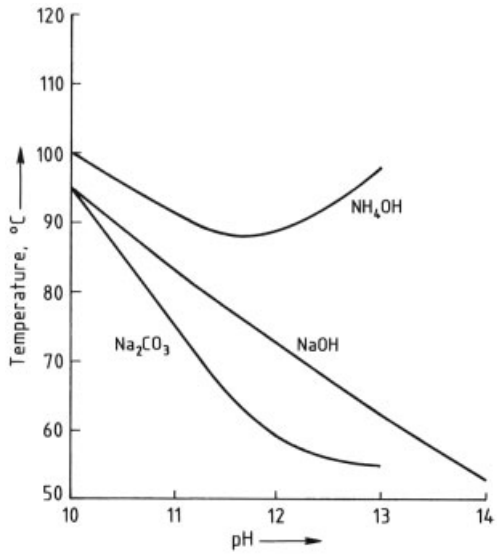

Figure 93. Iso-corrosion curves (0.1 mm/a) for chemical enamel in alkali

of chemical enameling is expected to result from the use of an enamel that combines the property patterns of vitreous and ceramic enamels. This product, introduced at Achema 1988 [176], has a reinforcing mechanism which is independent of the matrix material and absorbs the energy that tends to cause fractures. This "in situ" micro-reinforcement enables the mechanical stability of chemical apparatus coated with this enamel to surpass even that of plant coated with good ceramic enamels, without concessions having to be made in regard to chemical stability.

5.1.2.2. Brick Lining

Acid-resistant bricks, carbon bricks, and electrographite afford additional protection to parts of plants which, owing to the severity of the chemical or thermal exposure, are not adequately protected by coatings and linings.

Refractory bricks and other materials are used, both with and without a lining between the steel and bricks, to enclose reaction chambers. Depending on the conditions to be withstood, the products used include basic and neutral materials with high silicon dioxide contents, carbon bricks, and silicon carbide bricks. The structure of a chemically resistant brick lining is shown schematically in Figure 94.

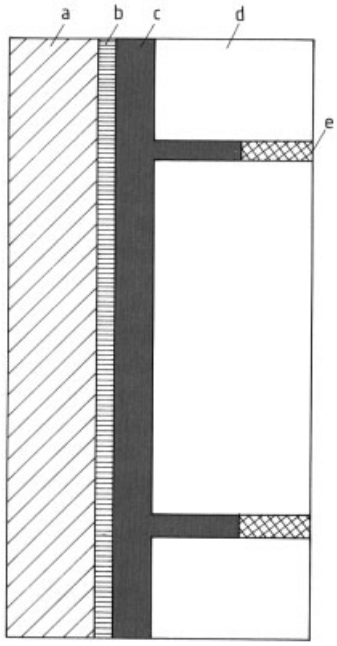

Figure 94. Schematic of a chemically resistant brick lining a) Substrate (concrete or steel); b) Resistant sealing compound; c) Bonding cement; d) Bricks or tiles; e) Pointing cement

5.1.2.3. Thermally Sprayed Coatings

Improvements in sprayable materials and spraying techniques have already resulted in the widespread use of thermally sprayed coatings consisting of inorganic nonmetallic materials [177]. These can be applied reliably by flame spraying, detonation spraying, or plasma spraying (see also → Metals, Surface Treatment, Chap. 5.). Metal oxides, metal carbides, or their mixtures are melted in a plasma flame at ca. 20 000 °C and sprayed onto the surfaces of the workpieces; coatings with a thickness of 0.1 mm to several millimeters can be produced in this way. Subsequent heat treatment of the coating can make it almost nonporous through the formation of a glassy phase. Most of the pores,

Table 9. Classification of spraying methods

Method	Energy source	Operating temperature, K	Partial velocity, m/s	Spraying material		Capacity, kg/h
				Product form	Material	
Flame spraying	fuel gas, acetylene – oxygen	3200	50 – 200	wire	mostly metals	5 – 8 (metals)
				powder	all types	1 – 2 (ceramics)
Jet-Kote high-speed flame spraying	fuel gas, slow-burning	2900	350 – 600	powder	metals and ceramics	2 – 6
Detonation spraying	fuel gas	3500	650	powder	all types	4 – 6
Plasma spraying	electricity	3000 – 20 000	350 – 450	powder	all types	4 – 8 (metals) 2 – 4 (ceramics)
Arc spraying	electricity	4000	150	wire	metals	15 – 20

which account for 1 – 20 % of the volume, are closed. Corrosion-resistant coatings can be built up without difficulty by spraying several coats on top of one another. In Table 9 the principle spraying processes are classified according to rate of application, supplementary spraying materials, and source of energy. Figure 95 shows a polished section through a wear-resistant coating on a channel entrance in a graphite exchanger [178].

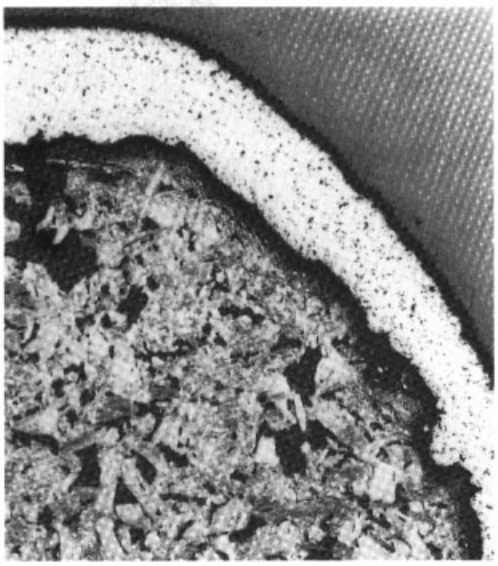

Figure 95. Polished section through a plasma-sprayed coating on a channel entrance in a graphite heat exchanger

Coatings of these types can be produced on carbon steels and stainless steels. Another application is shown in Figure 96. The fiber deflection roll shown here, which is coated with Al_2O_3, lasts three times longer than an otherwise identical roll with hard chromium plating.

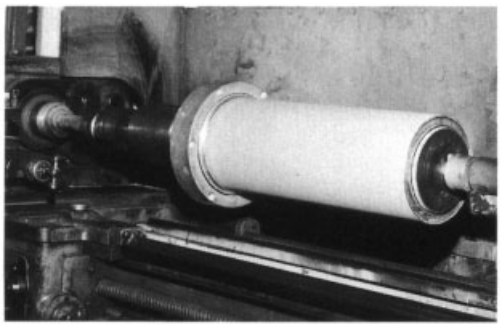

Figure 96. Fiber deflection roll, spray coated with Al_2O_3

5.1.3. Organic Lining and Coating Materials

The range of high-molecular lining and coating materials has increased substantially, both with respect to types and grades of materials and with respect to their properties and applications. In general it can be said that modification of well-known and proven materials by copolymerization has improved important mechanical properties of linings and coatings, such as their impact resistance, while compounding with randomly distributed conductive carbon has enhanced the electrical conductivity of plastics so much – without appreciably impairing their mechanical and chemical stability – that the linings or coatings can now even be exposed to combustible liquids and vapors (Figs. 97, 98, 99) [179].

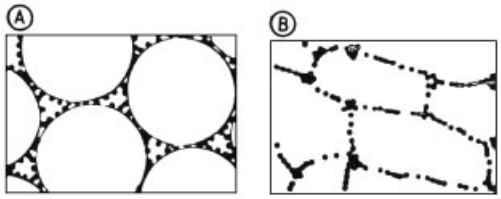

Figure 97. Schematic of materials with covered particle structure [180]. The diameter of the electrically conducting particles is considerably smaller than that of the plastic particles
A) Original structure; B) Deformed structure (e.g., after processing)

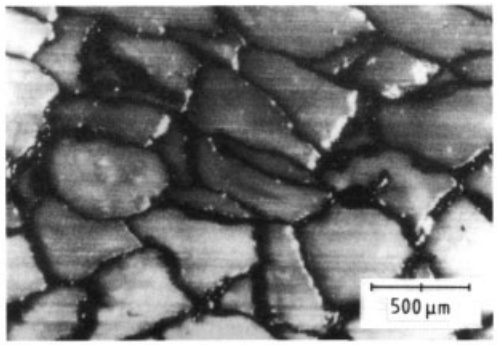

Figure 98. Section through an electrically conducting pressed sheet, consisting of high-density polyethylene filled with carbon black and having a covered particle structure

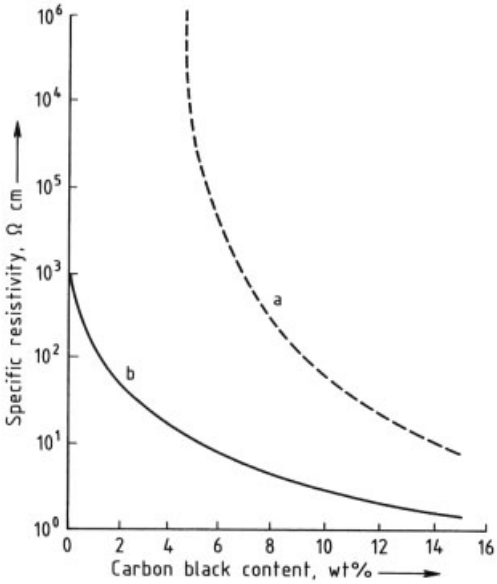

Figure 99. Specific resistivity as a function of carbon black content
a) Uniform carbon black distribution; b) Covered particle structure

Table 10 gives a general idea of the thermal stability of coatings as a function of chemical exposure. Standards and guidelines concerning organic lining and coating materials can be found in DIN 28 053, 55 929; VDI 2531 – 2539; TRbF 401 – 403; and DVS 2202, 2301.

Effects of Media. Polymers, unlike metals, are subject to swelling rather than corrosion. Chloride-induced pitting corrosion, to which austenitic Cr – Ni steels are particularly susceptible, is unknown in polymeric materials. Stress corrosion cracking, however, is possible under conditions that cause the medium to attack the polymer chemically [181]. Stress corrosion cracking can only occur, however, if, at the same time, stresses resulting from manufacture, coating or lining, or the conditions of use are present.

Where polymeric materials are concerned, attention should be paid to diffusion and permeation, irrespective of stability behavior. Even if the protective material is stable, gaseous or vaporous substances may penetrate organic coatings and linings and, if corrosive, attack the substrate, thus resulting in loss of adhesion, blister formation, and failure of the coating or lining to protect the substrate. The factors which influence this process include temperature, temperature gradient, the vapor pressure of the medium, the thickness of the lining, and the chemical affinity of the lining material and of the medium to which it is exposed.

Protective Coatings and Linings. Protective coatings or linings consisting of organic materials are applied to metals, especially steel, to glass-filled UP, VE, and EP resins, and to concrete and masonry. They protect the substrate from chemical attack. Recommendations for the design and production of the items to be protected can be found in VDI-Richtlinie 2532 for metals and in VDI-Richtlinie 2533 for concrete and masonry; recommendations on the choice of coating materials and methods are given in VDI-Richtlinie 2531.

5.1.3.1. Rubber Linings

Rubber linings (see also VDI-Richtlinie 2537) consist of ebonite or soft rubber based on natural or synthetic polymers. They account for the largest group of materials used for surface protection in chemical plant. Synthetic elastomers are becoming more and more important since they have greater chemical and mechanical stability than natural materials, and less susceptibility to mechanical damage.

Halogen butyl and self-vulcanizing systems based on the chloroprene rubber Baypren have become established and proved their worth, especially where they have been used for in situ treatment of large items such as storage tanks, waste water purification plant and flue gas desulfurization equipment [182]. The sulfochlorinated polyethylene Hypalon has good resistance to acidic and alkaline oxidizing agents and to chromic acid and chlorine bleaching solution.

Being more highly cross-linked, ebonite linings are even more resistant to swelling and permeation than soft rubber linings. As, however, they have to be vulcanized in workshop autoclaves, their use is generally restricted to relatively small items whose volume does not exceed about 100 m^3.

Plant components of up to about 1000 m^3 in volume can also be ebonite-lined on site if hot water or piped steam is available for vulcanization. For adequate cross-linking, the steel must

Table 10. Organic lining and coating materials used in chemical engineering

Material [a]	Application [b]			
	Acids, salt solutions	Alkaline solutions	Oxidizing agents	Organic media
Thermoplastic linings (1 – 4 mm)				
PVC with plasticizer	50 °C	50 °C	50 °C	
PVC without plasticizer	60 °C	60 °C	60 °C	
PE-HD	60 °C	60 °C		30 °C
PP with contact adhesive	50 °C	50 °C		30 °C
PP with fabric lamination	80 °C	80 °C		30 °C
PVDF (partly fluorinated)	100 °C		80 °C	30 °C
E/CTFE (partly fluorinated)	100 °C	100 °C	80 °C	30 °C
FEP, PFA (fully fluorinated)	120 °C	120 °C	120 °C	120 °C
PTFE (fully fluorinated)	200 °C	200 °C	200 °C	200 °C
Rubber linings (3 – 4 mm)				
Soft rubber based on natural and synthetic rubbers (NR, CR, IIR, NBR, CSM, FPM)	Nonoxidizing acidic to alkaline media up to 90 °C			
Ebonite based on natural and synthetic rubbers (NR, NR/IR, SBR, NBR)	Nonoxidizing acidic to alkaline media up to 110 °C			
Graphite-filled thermosetting plastic linings (4 – 5 mm)				
PF	120 °C			120 °C
FU	120 °C	120 °C		120 °C
EP	100 °C	120 °C		80 °C
Reinforced reaction resin coatings				
Glass fiber laminate coatings (3 – 0.5 mm)				
UP-GF and VE-GF	70 °C	50 °C	30 °C	30 °C
FU-GF	70 °C	70 °C		70 °C
EP-GF	70 °C	70 °C	30 °C	30 °C
Grouts and sprayed coats, reinforced with glass flakes (1 – 2 mm)				
UP and VE, based on bisphenol A	100 °C			
VE, based on epoxy – novolak acrylate	150 °C			
Paint and powder coatings (0.2 – 2 mm)				
Liquid coating materials				
Catalyst-curing resins (PUR, EP, UP, VE)	Nonoxidizing acidic to alkaline media up to 80 °C			
Thermosetting resins (PF, EP, PF/EP)	Nonoxidizing acidic to alkaline or organic media up to 120 °C			
Thermoplastic powders				
PE-HD	Nonoxidizing acidic to alkaline media up to 50 °C			
E/VAL	Nonoxidizing acidic to alkaline media up to 50 °C			
PA	Nonoxidizing neutral and alkaline or organic media up to 60 °C			
PVDF (partly fluorinated)	Acidic to neutral media up to 80 °C			
E/TFE, E/CTFE (partly fluorinated)	Acidic to neutral media up to 80 °C			
PFA, FEP (fully fluorinated)	Acidic to alkaline or organic media up to 150 °C			

[a] E/TFE Ethylene/tetrafluoroethylene, E/CTFE Ethylene/chlorotrifluorethylene, EP Epoxide, E/VAL Ethylene/vinyl alcohol, FEP Tetrafluorethylene/hexafluoropropylene, FU Furane, PA Polyamide, PCTFE Polychlorotrifluroethylene, PE-HD High density polyethylene, PF Phenol formaldehyde, PFA Perfluoro-alkoxyalkane, PP Polypropylene, PTFE Polytetrafluoroethylene, PUR Polyurethane, PVC Poly(vinyl chloride), PVDF Poly(vinylidene fluoride), UP Unsaturated polyester, VE Vinyl or phenacrylic ester, UP-GF Fiberglass-reinforced unsaturated polyester, VE-GF Fiberglass-reinforced vinylester, FU-GF Fiberglass-reinforced furane, EP-GF Fiberglass-reinforced epoxide, CR Chloroprene rubber, CSM Chlorosulfonyl polyethylene, FPM Vinylidene fluoride/hexafluoropropylene copolymer, IIR Isobutane – isoprene rubber, NBR Nitrile – butadiene rubber, NR Natural rubber, SBR Styrene – butadiene rubber.
[b] The indicated temperature limits depend on the aggressiveness, concentration, and penetration capability of the medium and may in individual cases be too high. If no experience is available, corrosion tests are recommended.

be heated to 90 – 95 °C, which is only possible if effective exterior insulation is provided. The largest tank known to have been ebonite-lined on site has a capacity of 8000 m^3. It has a triple-ply lining, which was applied as a 5-mm thick single sheet and consists of a soft rubber adhesive layer, a diffusion-proof ebonite layer, and a corrosion-resistant soft rubber layer. Since 1982 this tank has been continuously exposed to concentrated hydrochloric acid containing chlorine and solvents without suffering damage. It was installed at the New Martinsville works of Mobay in the United States.

5.1.3.2. Rubber – Plastic Composite Linings

The stability of rubber linings, especially to organic compounds that diffuse readily in rubber

and damage it severely, can be improved by the simultaneous use of plastics. In such cases a supplementary lining consisting of a thermoplastic chosen for its resistance to organic or oxidizing compounds is applied on top of a soft rubber or ebonite lining. Where the item to be protected will be exposed to temperatures of up to 50 °C in practical use, the polypropylene or polyvinylidene chloride sheets, which are the preferred materials, can be bonded to the rubber reliably with an isocyanate cross-linking chloroprene rubber adhesive. For higher temperatures, as reached in reaction vessels, for example, a bond must be formed by vulcanizing the rubber in contact with the plastic [183] or by placing a fabric on the rubber, applying a heat-resistant epoxy resin to it, adding the thermoplastic sheet, and applying pressure by means of a vacuum [184].

5.1.3.3. Linings Consisting of Thermally Crosslinkable Thermosetting Plastics

Phenol, furan, and epoxy resins are the main representatives of this group. After addition of plasticizers, graphite, and fibers they are applied like rubber linings as 4 – 6 mm thick sheets and cross-linked three-dimensionally in pressure vessels. In common with rubber linings they can also be applied to items made of steel or fiberglass-filled plastics. The cross-link densities of the finished linings are so high that the material is neither soluble nor capable of melting, and its strength and creep resistance are not dependent on temperature to the same extent as those of thermoplastics. Thermosetting plastics are used mainly to line columns, flues, and other plant items that have to withstand severe exposure to chemicals and heat.

5.1.3.4. Reinforced Reaction Resin Coatings, Crosslinkable with Catalysts

Fiber-Reinforced Laminate Coatings. Fiber-reinforced reaction resins (see also VDI-Richtlinie 2536) are used mainly to coat large storage tanks on site. Generally the coating is a fiberglass-mat-reinforced laminate, but sometimes it is applied by the fiber spraying technique. The coatings are based on unsaturated polyester, vinyl ester, epoxy, and furan resins, and, like those described in Section 5.1.3.3, they are thermosetting. The structure of a laminate coating is shown in Figure 100. Plant protected in this way generally consists of steel, but in some cases of concrete. Coatings highly resistant to acids, alkali, or even solvents can be produced, depending on the type of resin chosen. Storage tanks for concentrated chlorine bleaching solution have been protected successfully with fiberglass-reinforced unsaturated polyester resins based on bisphenol A fumarates or vinylesters based on bisphenol A acrylates. For more than 10 years fiberglass-filled vinyl esters based on epoxy – novolak acrylates have been protecting tanks of up to 5000 m^3 for the treatment of acidic or alkaline wastewater containing organic compounds. Fiberglass – furan (FU) laminates withstand solvents best, even at elevated temperatures. At the moment, however, the lack of ductility exhibited by FU resins restricts their application to containers of up to 100 m^3 in volume. Because of the differences between the expansion coefficients of the substrate and coating, fiberglass laminates can only be exposed to temperatures of up to 70 °C.

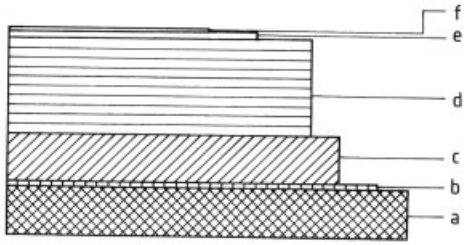

Figure 100. Structure of a laminate coating (schematic) a) Substrate (steel or concrete); b) Primer (ca. 0.1 mm); c) Grout (1 mm); d) Glass mats (2 mm); e) Glass veil; f) Sealing coat (0.2 mm)

Grouts and Spray Coats Reinforced with Glass Flakes. Thermosetting unsaturated polyester (PE) and vinyl ester (VE) resins reinforced by glass flakes, which have a thickness of only a few microns and lie parallel to the surface of the substrate, are applied as grouts and sprayed coats, especially to protect gas ducts in flue gas desulfurization units, even in cases where condensates form [185]. Other applications of these easily applied coatings, which, when silanized flakes are used, are also

relatively diffusion-proof, include the covers of large rubber-lined, brick-lined, or laminate-coated tanks. Details of their structure are given in Figure 101.

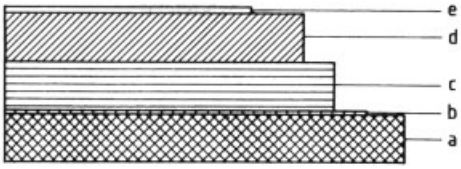

Figure 101. Structure of a grout coating reinforced with glass flakes (schematic)
a) Steel; b) Primer (ca. 0.1 mm); c) Grout (ca. 0.9 mm, 3×2-mm glass flakes); d) Grout (ca. 0.9 mm, 3×2-mm glass flakes); e) Topcoat (ca. 0.1 mm)

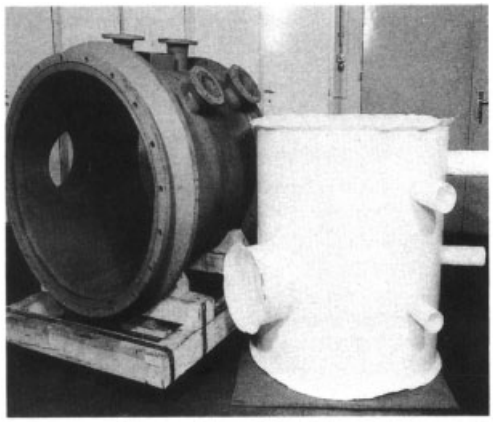

Figure 102. PTFE lining with welded-in sockets, before assembly

5.1.3.5. Thermoplastic Linings

These are available as semi-finished goods in the form of foils, sheets, panels, and tubes. The preferred materials are poly(vinyl chloride) with or without plasticizers (PVC-U, PVC-P), high-density polyethylene (PE-HD), polypropylene (PP), poly(vinylidene fluoride) (PVDF), ethylene chlorotrifluoroethylene (E/CTFE), tetrafluoroethylene/hexafluoropropylene (FEP), and polytetrafluoroethylene (PTFE). They are used to protect plant made of steel, concrete, or fiberglass-reinforced unsaturated polyester, vinyl ester or epoxy resins, and their application in the manufacture of vessels and other plant is increasing rapidly. Except for PTFE, which can be processed only by the sintering technique and calls for special methods where forming and welding are required, all of these thermoplastic lining materials can be formed, welded, and bonded without difficulty. Figure 102 shows a loose-insert lining of PTFE peeled foil with welded-in sockets which merely have to be attached to the flanges of the vessel and its sockets after the insert has been introduced. If the lined apparatus is to have a long useful life, especially at high service temperatures, the apparatus must be well insulated against loss of heat and should have ventilation openings in the wall beneath the lining. Figure 103 shows how the resistance of such linings to permeation by gases and steam depends on the density of the PTFE and therefore on the production process [186].

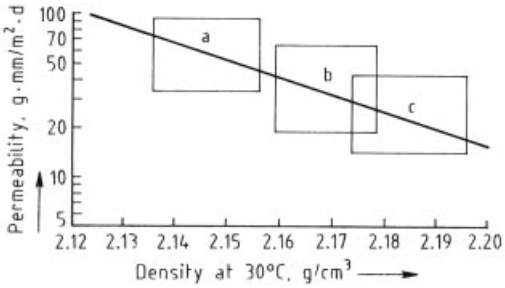

Figure 103. Permeability of PTFE in water at 137.5 °C and 0.34 MPa as a function of density
a) Ram-extruded products; b) Paste-extruded products; c) Peeled foil

Thermoplastic lining materials, with the exception of PTFE, are attached to the substrate with contact adhesives or with heat-resistant systems based on unsaturated polyester (UP), vinyl or phenacrylic ester (VE), or epoxide (EP), the adhesive or resin composition being applied to a glass or synthetic fiber fabric, the underside of which is fused into the pipe or panel. For service temperatures above 80 °C it is particularly desirable to use glass fabric because this improves resistance to hydrolysis. The use of knitted glass fiber is always recommended if the laminated sheet is likely to be deformed considerably while being processed, as in the fabrication of round-bottomed vessels, columns, or large pipes. The structure of a thermoplastic/fiberglass laminate is shown schematically in Figure 104. The weld joint is additionally reinforced by a

fabric-reinforced cover tape. This reinforcement is particularly necessary in the case of linings that must be antistatic or electrically conductive, especially because these cannot be tested for pores and cracks with high-voltage instruments. Liners of this kind having good ductility are now available as copolymers with carbon contents of less than 5 wt %. They have superseded the brittle homopolymers in which up to 25 % of carbon had to be incorporated [188]. Figure 105 shows the structure of Bekaplast, a composite consisting of a thermoplastic liner and concrete.

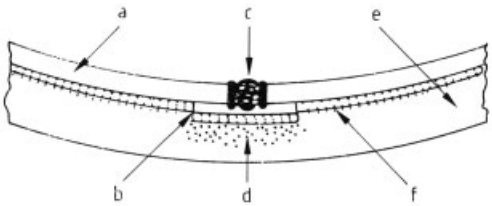

Figure 104. Structure of a fiberglass/thermoplastic laminate [187]
a) Thermoplastic; b) Thermoplastic patch strip; c) Weld; d) Carbon-filled resin target; e) FRP; f) Fiberglass

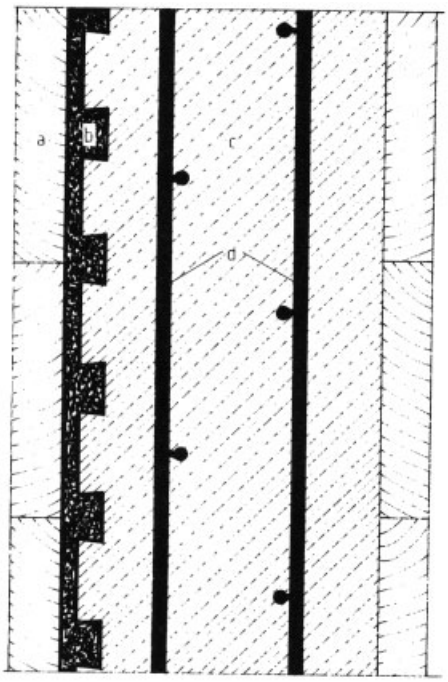

Figure 105. Structure of a Bekaplast lining [189]
a) Casing; b) Bekaplast panel; c) Concrete; d) Reinforcement

5.1.3.6. Paints and Powder Coatings

Liquid Coating Materials. Liquid coating materials are classified as physically drying (here the coating material gels after the solvent or dispersing agent has evaporated), thermally cross-linking, or catalytically cross-linking. Their applications depend on the polymer base, the thickness at which the coat can be applied, and–where heated plant items are concerned–on the magnitude and direction of the heat flow. In the chemical industry tube bundle heat exchangers coated on the cooling water side with stoving finishes are an important application of liquid coating materials in which they have shown that they are able to withstand the respective operating conditions continuously for up to ten years. Large storage tanks are being coated increasingly with epoxy-based high-solids formulations or with solvent-free polyurethane (PU), EP, UP, or VE resins by the low- or high-pressure process, and with or without heat, depending on the technical requirements of the system. The PU and EP resins are also modified with mineral coaltar pitch, particularly to reduce their cost and improve their stability to aqueous media [190].

Powder Coatings. Temperature sensitive plant items can be coated reliably with thermoplastics or thermosetting plastics by the techniques of fluidized-bed sintering, powder sintering, flame spraying, and electrostatic powder coating. The range of applications is similar to that of catalytically and thermally cross-linking resins. Details of surface protection by powder plastics that are melted to form coatings can be found in VDI-Richtlinie 2538.

In powder sintering, the powder is sprinkled, tipped, or flung onto the metal parts after they have been heated to 200 – 400 °C. Excess powder is removed and the plastic coat formed by melting is reheated until it becomes a continuous film.

In fluidized-bed sinter coating the heated substrate is dipped into a fluidized bed of the plastic powder. A 0.1 – 1.5 mm-thick film is formed, depending on the heat content of the part.

In electrostatic powder spraying, the plastic powder is sprayed by compressed air from a spray nozzle to which a high DC voltage is

applied. The current gives the plastic particles an electric charge which causes most of them to land on the earthed substrate as soon as this enters the vicinity of the polymer dust cloud. The difference in polarity between the substrate and the particles and the throwing power ensures that the metal object is uniformly coated, even on surfaces that do not face the cloud. The residual powder is separated in a cyclone and reused. Films with a thickness of 50 – 1000 µm can be obtained, depending on the coating equipment and properties of the powder. The powder, which adheres to the metal through Coulomb forces, is melted in a stove to form a continuous film. Not only components, but also complete apparatuses, are treated in this way.

5.2. Inhibitors [191], [192]

Inhibitors are substances that reduce or eliminate the aggressiveness of a corrosive medium and are either already contained in the corrosive medium or are specifically added to it. A distinction is made between electrochemical, chemical, and physical inhibitors [193].

Electrochemical inhibitors retard or prevent the anodic and/or cathodic partial reactions, i.e., they influence the reaction at the metal/corrosive medium interface. Chemical inhibitors can react both with the material and form protective coatings and with the medium itself or its constituents and thus diminish its aggressiveness. Physical inhibitors form adsorption layers on the metal surface, which block the corrosion reaction. Inhibitors that influence the electrochemical electrode reactions are subdivided according to their mode of action and site of action in the area of the metal/medium phase boundary, the subdivision being between interface inhibitors, electrolyte film inhibitors, membrane inhibitors, and passivators.

Industrial application of inhibitors involves acidic, neutral, and alkaline solutions as well as the gas phase [194]. Physical and electrochemical inhibitors can be used for acids, whereas all three types are employed in neutral and alkaline solutions. Gas and vapor phase inhibitors (i.e., substances with high vapor pressure) have gained importance, especially in atmospheric corrosion protection in storage and transportation [195].

The following compounds are used as inhibitors in acid solutions: amines, aminoimidazolines, amino- and nitrophenols, aminotriazole, aldehydes, benzothiazol, dibenzyl sulfoxide, dithiophosphonic acids, guanidine derivatives, ureas, phosphonium salts, sulfonium salts, sulfonic acids, thioethers, thioureas, and thiocarbanoyl disulfides. Amino alcohols, aminobenzimidazole, benzoates, quinoline derivatives, cinnamates, fatty amines, polyether amines, silicates, and triazoles are used as inhibitors in neutral or weakly alkaline solutions, while for strongly alkaline solutions, aldehydes and fatty amines are used.

Inhibitors exist for the protection of metals and their alloys [196–200] and for the most varied of uses, including pickling acids, service water, brine, coolants, oils, heat-transfer agents, and electroplating baths [201–203]. Often, using a combination of several inhibitors gives a synergistic effect, whereby the protective action achieved generally far exceeds the sum of the effectivenesses of the individual components.

For the chemical industry inhibitors in open and closed water circulation systems and in oil refining are highly significant [203]. In oil refineries, for example, the cracking that may occur in heat-treatable steels through the action of weakly acidic solutions containing H_2S and H_2O can be eliminated by the addition of ammonia saturated with oxygen. The H_2S is oxidized to polysulfide, which forms a good protective coating on the steel [111].

Corrosion inhibitors for mineral oils include alkyl imidazoles, amines, aminopyridines, quaternary ammonium salts, alkyl sulfonates, borate esters, hydroxy amines, naphthenic acids, phosphate esters, and stearates.

Oxygen and carbon dioxide corrosion can cause serious economic loss in water circulation systems in which water is heated, evaporated, and condensed. Oxygen-binding additives offer good protection (hydrazine, Levoxin, sodium sulfite) and are used extensively in boiler feed water preparation. Carbon dioxide corrosion can be controlled by neutralizing and film-forming amines and by vapor phase inhibitors. In practice derivatives of phosphoric acid, organophosphates, and chelates have proved successful as inhibitors in open cooling circulation systems.

Note, however, that there are conditions under which inhibitors can give rise to detrimental

local corrosion (pitting corrosion). This is the case when the amount of inhibitor is insufficient. Under these conditions only part of the surface can be covered, thus giving rise to a local element. Corrosive attack is particularly extensive at the uncovered anode areas because of increased corrosion current density, and deep cavities penetrate into the material. Similarly, if the inhibitor is too readily reduced at the cathodic areas of the metal surface, increased corrosion can result because compact protective films are not formed. Since there are no universally applicable inhibitors they must be carefully selected and examined for each specific case. In doing so, inhibition of metal dissolution is not the only point to be considered; there is also hydrogen absorption [203].

In many cases corrosimeters are successfully used to monitor inhibition.

6. Electrochemical Corrosion Protection

Potential plays an essential role in corrosion processes. Not only the corrosion rate but also the occurrence of local corrosion phenomena, such as pitting and stress corrosion cracking, depend on the potential. If the critical potential range for corrosion is known, a condition can be established through polarization in which no corrosion occurs or it is negligible. Depending on the direction of the polarization this electrochemical corrosion protection is anodic or cathodic [204].

In addition to the well-known application of cathodic corrosion protection to underground pipelines, there has been an increased use for the internal protection of containers and pipes. Initially, galvanic anodes were used to this effect, like the ones now used, for example, to protect the interiors of tankers and boilers. However, since these anodes are often subject to heavy inherent corrosion, especially with the highly aggressive media often found in the chemical industry, external current systems with insoluble anode material have now largely replaced these.

The possibilities of cathodic protection are limited in two respects. The cathodic polarization required for protection can lead to cathode corrosion in some systems. Hydrogen evolution can also cause damage such as embrittlement of the steel and debonding of a protective coating (e.g., paint). This kind of protective coating is often combined with cathodic protection in order to keep the current demand low. The attack of acidic solutions on equipment parts cannot therefore be prevented in general by a cathodic protection system. Cathodic polarization without evolution of large amounts of hydrogen is at best a possibility with copper alloys in acidic solutions [205].

For these reasons the cathodic corrosion protection of chemical plant parts is restricted to neutral solutions, service water, and alkali. Various types of anode are available for the protection of container interiors. Platinized disk electrodes, bar anodes, and titanium basket anodes are used [206], [207]. They are now supplied ready for installation and are designed in such a way that in continuous operation they can emit 8 – 10 A at a current density of 6 – 8 A/dm^2.

Aluminum and ferrosilicon anodes are suitable for the protection of hot-water tanks [208]. The use of aluminum anodes in a hot-water apparatus by the Guldager electrolysis method [209] has additional effects: an indirect protective action on subsequent piping, since a protective film (1 – 1.5 mm thick) is built up by anode-formed aluminum oxide hydrates. In addition to alkalization at the cathode, a shift in the $CaCO_3$–CO_2 equilibrium leads to partial softening of water, $CaCO_3$ being deposited at the cathode.

Corrosion protection through imposition of a passive state in metallic materials by anodic polarization has found some applications in the chemical industry.

The practical design of equipment for anodic protection requires intensive laboratory studies to determine the passivation current densities and the potential range of passivity as a function of various parameters such as temperature, concentration, and rate of flow for the respective application. The properties of the cathode material must also be studied.

The protective current must be potentiostatically controlled if the material to be protected shows rapid activation on disconnection. Intermittent operation may also be chosen, whereby the protection current connects or disconnects on dropping below or rising above a limit potential.

All common systems can be used as reference electrodes, e.g., calomel, Ag/AgCl, and Mg/MgO electrodes, which help to monitor the

apparatus and control the on-off processes. Platinum, platinized materials, tantalum, and lead are used as cathode materials in acidic solutions. Nickel is suitable in alkaline solutions.

Anodic protection against acidic solutions has been used in a number of chemical processes and in the transportation and storage of liquids. Unalloyed steels can be protected in this way in salt solutions with nitrates and sulfates and in nitric and sulfuric acid [210], although there are limits imposed in sulfuric acid by temperature and concentration [211]. Stainless chromium and chromium – nickel steels are particularly suited to anodic protection. It has so far been practiced with H_2SO_4, oleum, and H_3PO_4.

In the production of sulfuric acid, including heat recovery and the reconditioning and recycling of spent acids, it is necessary to handle acids at elevated temperatures and various concentrations. Corrosion damage that considerably impairs the availability of plant has occurred in sulfuric acid coolers, for example. Damage of this kind can be prevented by anodic protection.

The commonly used austenitic stainless steels exhibit satisfactory resistance to corrosion by sulfuric acid at low concentrations (< 20 %), and high concentrations (> 70 %) below a critical temperature. If at high sulfuric acid concentrations (> 90 %) the temperature exceeds ca. 70 °C, corrosion, differing in severity according to the composition of the steel concerned, occurs, and the steel may alternate between the active and passive states [212].

Anodic protection enables materials to be used under unfavorable conditions, provided they are passifiable in sulfuric acid. In the handling of sulfuric acid at concentrations of 93 – 99 %, Cr – Ni steels (material nos. 1.4541 and 1.4571) can be used economically at temperatures up to 160 °C. This allows operation within a range of temperatures (120 – 160 °C) suitable for heat recovery.

The anodic protection technique now enables air coolers and tube bundles in sulfuric acid plants to be protected from corrosion reliably and economically. In 1966 anodic protection was provided in the Federal Republic of Germany for the air coolers of a sulfuric acid production plant for the first time. Since then, a combined cooler surface area exceeding 10 000 m^2 in air-cooled and water-cooled sulfuric acid plants has been protected in this way worldwide. The installed initial electrical direct current output of the potentiostats is > 25 kW, corresponding to an energy requirement of 2.5 W/m^2 for the surface needing protection [213].

Owing to the narrowness of the potential range within which stress corrosion cracking occurs in unalloyed and low-alloy steels in alkali, anodic protection against this particular kind of corrosion can be provided by impressed current. As the current densities needed to maintain passivity are quite low, this form of protection is also economically favorable, particularly for large plant equipment that cannot be annealed. Cathodic protection, though theoretically possible, cannot be recommended, since excessive surface corrosion caused by removal of top layers from the steel surface would have to be tolerated as the alternative to stress corrosion cracking. Further reasons are that higher current densities are needed and that a power failure would immediately end the protection. A passivated surface, on the other hand, does not become active immediately, so that anodic protection can also be operated intermittently.

Anodic inhibition of stress corrosion cracking was first provided on an industrial scale in a large plant for the production of hydrogen by electrolysis of potassium hydroxide solution [87]. In November 1968, after preliminary trials on a laboratory scale, the chemical industry's first anodically protected large-scale plant, a sodium hydroxide solution evaporator with a capacity of 142 t, was put into operation. In the meantime further plants have been equipped in the same way [215]. The structure of the evaporator and the arrangement of the counter electrodes and reference electrodes are shown schematically in Figure 106.

Because of its good passivity, titanium is also suited for anodic protection, e.g., in mineral and organic acids; tantalum cathodes are generally used [216].

7. Corrosion Testing [217–219]

7.1. Introduction

In the design of chemical apparatus, apart from process technology, material selection is especially important. With careful attention paid to economy, material selection must take account

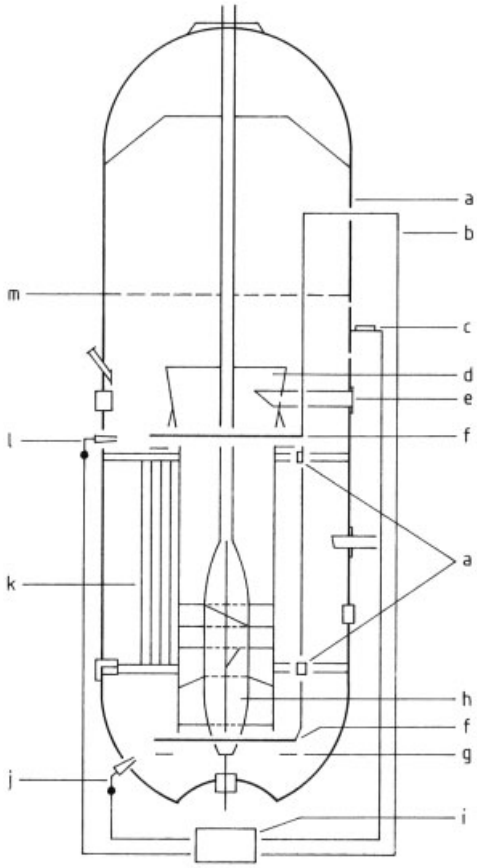

Figure 106. Anodic protection of a caustic vaporizer (content 115 m^3, surface area 2400 m^2) against stress corrosion cracking
a) PTFE; b) Cathode; c) Anode; d) Central tube e) Liquor entry; f) Ring electrode; g) Insulation; h) Stirrer; i) Potentiostat; j) Electrode E_2; k) Vaporizing tubes; l) Electrode E_1; m) Filling level

of all the requirements arising out of the operating function of the chemical apparatus and the manufacturing process of the product. It is often difficult to comprehend and take account of the interaction of mechanical and chemicocorrosion stresses, which are often compounded by wear.

The resistance parameters needed to select a material can be obtained in two ways:

1) by incorporating suitable test specimens in test plants or in existing, not yet optimally designed production plants;
2) by corrosion testing in the laboratory.

Both methods of corrosion testing are similiar in that on the corrosive medium side a number of important parameters such as concentration, temperature, dissolved gases, impurities, solids, and the flow rate must be carefully considered. Because of the many parameters involved, when material specimens are incorporated in a test or production plant, efforts should always be made to ensure that the specimens are subjected as close as possible to operating stress. The material specimen is geared to the types of corrosion expected.

Material specimens are mostly attached to internals, e.g., to an agitator or a thermometer protection tube. To avoid unwanted polarization of the specimens through contact with plant components and a subsequent incorrect measurement, the specimens are bolted with insulating sleeves. In many cases it is sufficient to wrap the bolts with PTFE tape. At low flow rates and low temperatures the specimens can be incorporated with the help of plastic binding or tape.

If corrosion tests in pilot plants are not possible, laboratory tests sufficiently close to practice should be carried out. These should also be carried out when operating tests do not clearly indicate the time function of corrosion processes and thus more intense corrosion conditions can complete the picture.

Three groups of variables determine the corrosion of metallic materials [220]:

1) material variables (chemical composition, heat treatment, and surface condition);
2) corrosive medium variables (pH, temperature, flow rate);
3) potential.

In a laboratory test used as the basis for material selection the biggest problem is the correct choice of these variables, which is the determining factor for simulating operating conditions.

Chemical corrosion tests focus primarily on resistance to surface and selective corrosion. In general the effect of material and corrosive medium variables can be understood with these methods. The variable potential, on the other hand, is more or less undefined and can experience time changes depending on the properties of the various partial reactions involved in corrosion. The fluctuation range of the potentials found in practice cannot be taken into account in the immersion test. Often, therefore, the results obtained in chemical corrosion testing using electrochemical methods must be fur-

ther differentiated to take account of the variable potential. If the corrosion rate determined in chemical corrosion testing depends heavily on the potential, chemical testing has little to offer, but if there is only little potential dependence it is more reliable.

By means of electrochemical corrosion tests, the dependence of corrosion on the potential can be investigated and indicators obtained as to the parameters which influence the potential. There can be fundamental differences between the results of chemical and electrochemical corrosion tests. This is due to the fact that in chemical tests the potential can change with time. This is deliberately avoided in electrochemical tests, with partly increased predictability and partial restriction of the potential practical uses, e.g., in the event of strong time dependence. Dependence on the following variables is a major consideration in electrochemical corrosion tests: potential U_H (in relation to the standard hydrogen electrode), current density i (in relation to the geometric surface of the specimen), and time t.

7.2. Test Methods

Process-Related Corrosion Tests. As mentioned above, tests to determine the corrosion resistance of materials in industrial corrosive media can be performed in the laboratory, in pilot plants, and in existing production plants. With regard to the choice of locations for specimens it should be borne in mind that reactors and other apparatus may be attacked differently by the liquid and vapor phases of the corrosive agent and at the three-phase interface of the liquid, vapor, and material. As a rule, therefore, material specimens must be exposed at each of these phases. Corrosion testing principles are described in [221–223].

The characteristics of individual forms of corrosion are taken into consideration by providing appropriate corrosion specimens. Welded coupons having the surface quality of the material used later in practice are sufficient for determining uniform corrosion rates and acquiring general information on the type of local corrosion. Resistance to crevice corrosion can be determined by using specimens as described in ASTM G 78 [224]. Conditions of heat transfer can be simulated by using hot-wall/cool-wall specimens under temperature-controlled conditions [222].

Sensitivity to stress corrosion cracking can be determined with elastically or plastically prestressed specimens, such as those of the Jones, U-bent, and C-ring types (Fig. 107, see next page) [225–227]. Specimens sensitive to stress corrosion cracking can also be produced by introducing heavy tensile stresses into the surface by means of stress-inducing grinding [228].

Laboratory corrosion tests are discussed in [221], [223]. To adjust the parameters of temperature and/or pressure it may be necessary to perform corrosion tests in an autoclave.

In pilot and production plants, the flow conditions are determined by the nature of the apparatus and process, but in laboratory tests they have to be individually chosen. To determine the effects of static or very gently moving media it is sufficient to stir the medium with an agitator. If, however, exposure of the material to flowing media is expected, special corrosion tests are essential for simulation: for example, circulation tests with pipe or channel flow and use of rotating discs or cylinders as specimens [229].

Special laboratory investigations of the resistance to stress corrosion cracking may include the use of machinery. In such cases tensile test specimens are tested under defined mechanical conditions, such as constant stress or constant strain rate [127].

Standardized Methods. Standardized corrosion tests are performed in the laboratory with standardized solutions; they are used mainly to determine general corrosion characteristics, as well as in fundamental investigations and quality control. Various methods [230–235], the choice of which depends on the type of material concerned, are used to determine the resistance of highly alloyed steels and nickel-based alloys to intergranular corrosion. There is also a standardized method for determining resistance to pitting and crevice corrosion [236]. The general resistance of various materials to stress corrosion cracking (SCC) can likewise be determined in test solutions. Depending on the material and corrodent, the following standards are used:

1) Stainless steels, transgranular SCC: $MgCl_2$ solution [217], [237]; NaCl solution [242];

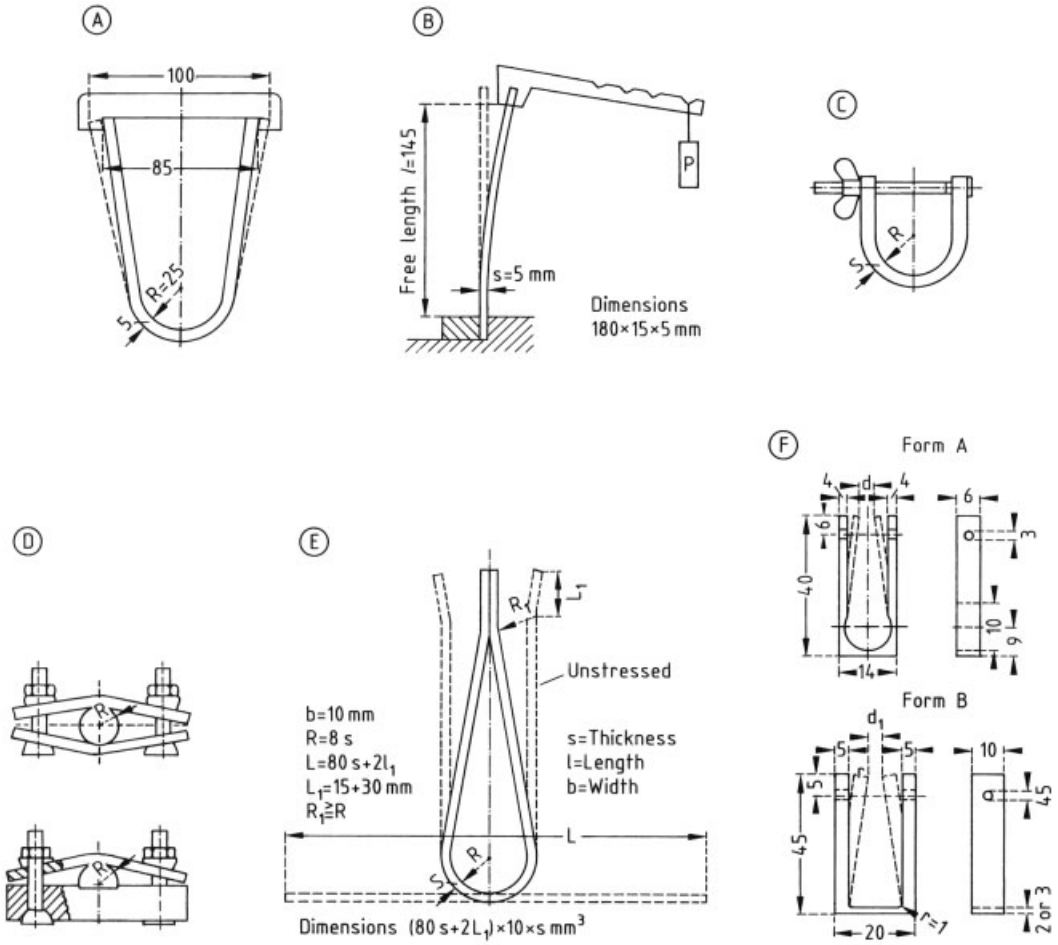

Figure 107. Stress corrosion cracking specimens
A), C), and E) Typical U-bent specimens; B) Cantilever beam specimen; D) Bent beam specimen, three point loaded; F) Tuning fork specimen

2) Unalloyed and low-alloy steels: intergranular SCC [225]; hydrogen-induced cracking H_2S [238];
3) Cu materials in NH_3 [239]; in NH_4^+ solution [240];
4) Al materials [241], [242].

The standard methods also include those used particularly, but not exclusively, to test anticorrosion coatings in corrosive atmospheres, e.g., those containing SO_2 and Cl^-.

Field Tests. Field tests include those in which specimens are surrounded by aggressive soils, atmospheres, or waters (e.g., seawater). Atmospheric and water tests, which are performed both on unprotected and on protected (coated) materials, require special methods [243, Parts 1 and 2]. Here it is important that control specimens be tested simultaneously in order to predetermine the corrosive conditions at the testing site.

Atmospheric Corrosion Testing. It is customary to perform tests in special climate-chambers in addition to field tests. The tests are used for comparison, but are also valuable for determining the behavior of anticorrosive films and coatings. The conditions used to obtain the appropriate atmospheres, constant or alternating condensed water climates, with and without the

presence of such additional substances as sulfur dioxide and salt spray, at various pH values, are specified in [244–250].

Electrochemical Measurement. Electrochemical methods are used mainly in the laboratory [251–253]. In special cases (potential measurements; corrosion meters) they can also be used in industrial plants.

Electrochemical methods, in contrast to chemical methods, provide mainly qualitative data on corrosion systems. They are very suitable for studying corrosion mechanisms and the influences of parameters, but also for investigating the value and effectiveness of active corrosion protection measures. Current density versus potential curves (see Chap. 2, Figs. 13, 17) provide information on the effects of alloy constituents and inhibitors. Quantitative data on loss of mass rates can be gained by extrapolating the Tafel lines or by measuring the polarization resistance (see Section 2.3; Figs. 9, 10). The latter method also forms the basis of a number of commercially available corrosion meters, which are used to determine corrosion rates and to investigate the effectiveness of inhibitors.

Depending on the purpose of an investigation, such tests are performed either at a given potential (potentiostatic tests) or at a potential which changes in a defined manner (potentiodynamic tests). Galvanostatic or galvanodynamic tests are also performed frequently. Besides investigation of the relationship between current density and potential, simple measurement of the (free) corrosion potential U_{corr} (open loop potential), or its dependence on time, is a useful way of studying corrosion systems, for the level at which the potential of the material settles is very important with regard to the damage suffered. Potential measurements are also fundamental, though not alone sufficient, in the investigation of contact corrosion [254], [255]. Electrochemical methods are also used successfully in the basic investigation of local corrosion, e.g., crevice and pitting corrosion.

A special method for determining the sensitivity of stainless steels to pitting corrosion in industrial corrodents is the measurement of the pitting passivation potential U_{pp}. Here the potential, which is initially adjusted to the free corrosion potential U_{corr}, is slowly increased until, at the pitting potential U_{p} (see Section 3.2.1) pitting – recognizable from the steep increase of the current (Fig. 108) – is forced to occur. After the rate of pitting corosion has been stabilized by a further increase of potential the scanning direction is reversed until the current acquires values similar to those of the passive current density. The potential at which this occurs is known as the pitting passivation potential U_{pp}. Comparison of U_{pp} with U_{corr} indicates whether pitting corrosion will occur ($U_{\mathrm{pp}} < U_{\mathrm{corr}}$) or not ($U_{\mathrm{pp}} > U_{\mathrm{corr}}$) in the corrosion system concerned [256]. The main advantage of this process is that the repassivation potential U_{pp} is more easily reproduced than the pitting potential U_{p}, and that within wide limits it is independent of the rate of change of the potential.

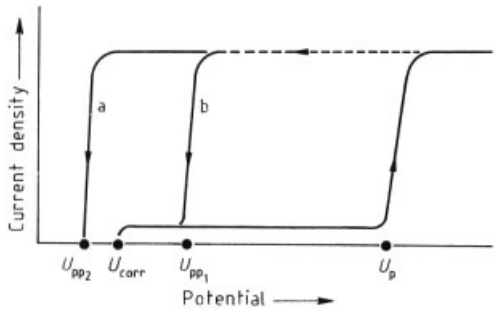

Figure 108. Determination of susceptibility to pitting corrosion (schematic potentiodynamic current density versus potential curves)
a) $U_{\mathrm{pp}} < U_{\mathrm{corr}}$: pitting corrosion susceptibility; b) $U_{\mathrm{pp}} > U_{\mathrm{corr}}$: no pitting corrosion susceptibility

8. References

General References

1. H. H. Uhlig: *The Corrosion Handbook,* J. Wiley & Sons, New York/Chapmann & Hall Ltd., London 1948. F. N. Speller: *Corrosion, Causes and Prevention,* McGraw-Hill, New York – London 1951. F. Ritter: *Korrosionstabellen metallischer Werkstoffe,* 4th ed., Springer Verlag, Wien 1958. F. Tödt: *Korrosion und Korrosionsschutz,* 2nd ed., de Gruyter, Berlin 1961. K. Vetter: *Elektrochemische Kinetik,* Springer Verlag, Berlin – Göttingen – Heidelberg 1961. F. L. Laque, H. R. Copson: *Corrosion Resistance of Metals and Alloys,* 2nd ed., Reinhold Publ., New York 1963. U. R. Evans: *Einführung in die Korrosion der Metalle,* Verlag Chemie,

Weinheim 1965. U. R. Evans: *The Corrosion and Oxidation of Metals; Scientific Principles and Practical Applications,* **1st. and 2nd suppl. vols.,** Edward Arnold Ltd., London 1960, 1968, 1976. H. Kaesche: *Die Korrosion der Metalle,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York 1979. Mannesmann-Röhren-Werke, *Lexikon der Korrosion,* **vols. 1 and 2,** Düsseldorf 1970. H. E. Hömig: *Metall und Wasser. Eine Einführung in die Korrosionskunde,* Vulkan-Verlag, Essen 1971. H. Gräfen, F. Kahl, A. Rahmel (eds.): *1. Korrosionum: Die Bedeutung der Korrosion für Planung, Bau und Betrieb von Anlagen der chemischen und petrochemischen Technik sowie in der Mineralölindustrie,* Verlag Chemie, Weinheim 1974. B. J. Monitz, W. I. Pollock (eds.): *Process Industries Corrosion,* NACE, Houston, Texas, 1986. R. N. Parkins (ed.): *Corrosion Processes,* Applied Science Publishers Ltd., London – New York 1982. D. Kuron (ed.): *Wasserstoff und Korrosion,* Bonner Studien Reihe, Verlag I. Kuron, Bonn 1986. R. Gibala, R. F. Hehemann (eds.): *Hydrogen Embrittlement and Stress Corrosion Cracking.* Papers presented at a symposium held at Case Western Reverse University on June 1 – 3, 1980, ASM, Ohio 1984. G. Kortüm: *Lehrbuch der Elektrochemie,* 5th ed., Verlag Chemie, Weinheim 1972. C. H. Hamann, W. Vielstich: *Elektrochemie I,* Verlag Chemie, Weinheim 1975. Verein Deutscher Eisenhüttenleute (ed.): *Prüfung und Untersuchung der Korrosionsbeständigkeit von Stählen,* Verlag Stahleisen, Düsseldorf 1973. A. Rahmel, W. Schwenk: *Korrosion und Korrosionsschutz von Stählen,* Verlag Chemie, Weinheim 1977. H. Gräfen et al. in W. J. Bartz, Techn. Akademie Esslingen (ed.): *Die Praxis des Korrosionsschutzes, Kontakt und Studium,* **vol. 64,** Expert Verlag, Grafenau 1981. DIN-Taschenbuch 219:Korrosion und Korrosionsschutz, Beurteilung – Prüfung – Schutzmaßnahmen, Beuth-Verlag, Berlin – Köln 1987. W. Fischer (ed.): *Korrosionsschutz durch Information und Normung,* Kommentar zum DIN-Taschenbuch 219, Schriftenreihe der Arbeitsgemeinschaft Korrosion e. V., Verlag I. Kuron, Bonn 1988. Annual Book of ASTM Standards 1988, Section 3; *Metals Test Methods and Analytical Procedures,* vol. 03.02, Wear and Erosion, Metal Corrosion, American Society for Testing and Materials, Philadelphia 1988. W. v. Baeckmann, W. Schwenk, W. Prinz: *Handbuch des kathodischen Korrosionsschutzes,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1989. P. J. Gellings: *Introduction to Corrosion Prevention and Control for Engineers,* Delft University Press, Delft 1976. L. L. Shreir: *Corrosion,* 2nd ed., **vols. 1 and 2,** Newnes-Butterworth, London – Boston 1976. *Corrosion Data Survey Metals Section,* 6th ed., NACE, Houston, Texas, 1985. D. Behrens (ed.): *DECHEMA Corrosion Handbook, Corrosive Agents and Their Interaction with Materials,* **vols. 1 – 5,** VCH Verlagsgesellschaft, Weinheim – New York 1987. M. Fischer, K. Hauffe, W. Wiederholt: *Passivierende Filme und Deckschichten,* Springer Verlag, Berlin 1956. K. Hauffe: *Oxidation von Metallen und Metallegierungen,* Springer Verlag, Berlin 1956. O. Kubaschewski, B. E. Hopkins: *Oxidation of Metals,* Butterworth, London 1962. J. Bernard: *L'Oxydation des Metaux,* **vols. 1 and 2,** Gauthier-Villars, Paris 1962 and 1964. M. Pfeiffer, M. Thomas: *Zunderfeste Legierungen,* Springer Verlag, Berlin 1963. P. Kofstad: *High-Temperature Oxidation of Metals,* J. Wiley, New York 1966. J. C. Scully: *Fundamentals of Corrosion,* Pergamon Press, Oxford 1966. Z. A. Foroulis: *High Temperature Metallic Corrosion of Sulfur and its Compounds,* The Electrochemical Society, New York 1970. H.-J. Engell, A. Rahmel: "Verzunderung von Metallen durch Gase. 50 Jahre Tamman'sche Zunderformel," *Korrosion* **23** (1971). A. Rahmel, H. Manenc: "Mechanische Eigenschaften und Haftung von Zunderschichten. Einfluß auf die Oxidation von Metallen," *Korrosion* **24** (1973).

Specific References

2. DIN 50 900,Korrosion der Metalle. Begriffe, Part 1, Allgemeine Begriffe, Beuth-Verlag, Berlin – Köln 1982.
3. DIN 50 900,Korrosion der Metalle. Begriffe, Part 2, Elektrochemische Begriffe, Beuth-Verlag, Berlin – Köln 1984.
4. C. Wagner, W. Traud, *Z. Elektrochem. Angew. Phys. Chem.* **44** (1938) 391 – 402.
5. M. Stern, A. L. Geary, *J. Electrochem. Soc.* **104** (1957) 56 – 63.
6. E. Heitz, *Werkst. Korros.* **19** (1968) 773 – 781.
7. E. Heitz, W. Schwenk, *Werkst. Korros.* **27** (1976) 241 – 245.
8. U. R. Evans, *J. Inst. Met.* **30** (1923) 239.

9. H. Kaesche, *Z. Metallkd.* **61** (1970) 94 – 101.
10. U. Gramberg, E.-M. Horn, *VDI-Ber.* **243** (1975) 75 –92.
11. H. Gräfen, K. Gerischer, E.-M. Horn, *Z. Werkstofftech.* **4** (1973) no. 4, 169 – 186.
12. E.-M. Horn, D. Kuron, H. Gräfen, *Z. Werkstofftech.* **8** (1977) 37 – 68.
13. E.-M. Horn, *VDI-Ber.* **235** (1975) 39 – 49.
14. K. Lorenz, G. Medawar, *Thyssenforschung* **1** (1969) no. 3, 97 – 108.
15. E. Wallis: *Einfluß des Molybdängehaltes auf die Lochkorrosionsbeständigkeit hochlegierter nichtrostender Stähle und Nickel – Chrom – Eisen-Legierungen mit unterschiedlichen Gefügezuständen unter Berücksichtigung der Wirksummenrelation Massen-% Cr + 3,3 Massen-% Mo,* Dissertation, TU Clausthal 1988.
16. M. Renner, U. Heubner, M. B. Rockel, E. Wallis, *Werkst. Korros.* **37** (1986) 183 – 190.
17. G. Herbsleb, *Werkst. Korros.* **33** (1982) 334 – 340.
18. DIN 17 440,issued 07. 85; VdTÜV-Werkstoffblatt 405, issued 12. 87.
19. VdTÜV-Werkstoffblatt 486, issued 09. 87.
20. Stahl-Eisen-Werkstoffblatt des Vereins Deutscher Eisenhüttenleute, SEW 400, issued 04. 88.
21. VdTÜV-Werkstoffblatt 415, issued 09. 84.
22. A. Bäumel, E.-M. Horn, G. Siebers, *Werkst. Korros.* **23** (1972) no. 11, 973 – 983.
23. H. Brandis, G. Lennartz, R. Oppenheim, *DECHEMA-Monogr.* **78** (1975) 235 – 258.
24. A. Diebold, H. Weingerl, *Werkst. Korros.* **25** (1974) no. 3, 175 – 179.
25. J. C. Carlen, B. Kwarnbäck, *Werkst. Korros.* **25** (1974) no. 9, 653 – 658.
26. W. Stewen, H. Jörgens, M. Remke, M. Litzkendorf, *Glückauf* **121** (1985) no. 13, 1030 – 1033.
27. VdTÜV-Werkstoffblatt 421, issued 02. 82/03. 84/09. 86.
28. VdTÜV-Werkstoffblatt 473, issued 06. 85.
29. M. B. Rockel, M. Renner, *Werkst. Korros.* **35** (1984) 537 – 542.
30. *Nickel Topics* **34** (1981) no. 4, 15.
31. J. C. Crum, E. L. Hibner, R. W. Ross, Jr., *Mater. Perform.* **20** (1981) no. 2, 9 – 13.
32. Sandvik Steel AB: Sandvik Sanicro 28 – References and booked orders for the process industry, Oct. 1985.
33. Avesta Technical Information: Avesta 254 SMO – Referenzen und Anwendungsgebiete, issued 02. 1982.
34. W. Heimann, *VDI-Ber.* **674** (1988) 41 – 63.
35. R. Kirchheiner, W. Römer: VDM-Report no. 9, Oct. 1987.
36. H. Kaesche, *Werkst. Korros.* **14** (1963) 557 – 574.
37. W. R. Fischer: *Zur Lochfraßkorrosion,* Dissertation, TH Braunschweig 1964.
38. K. Fässler, *VDI-Ber.* **235** (1975) 51 – 68.
39. K. Fässler: "Gesichtspunkte der konstruktiven Gestaltung chemisch beanspruchter Apparate," in H. Gräfen, F. Kahl, A. Rahmel (eds.): *1. Korrosionum, die Bedeutung der Korrosion für Planung, Bau und Betrieb von Anlagen der chemischen und petrochemischen Technik sowie der Mineralölindustrie,* Verlag Chemie, Weinheim 1974, pp. 136 – 165.
40. R. Münster: "Erscheinungsformen der Korrosion," in [39] pp. 3 – 27.
41. G. Lennartz, *VDI-Ber.* **235** (1975) 169 – 182.
42. G. Herbsleb, *VDI-Ber.* **243** (1975) 103 – 118.
43. E.-M. Horn, T. Günther, *Chem. Ing. Tech.* **41** (1969) no. 18, 991 – 999.
44. G. Grützner: *Über die interkristalline Korrosion stickstofflegierter 18/10 Chrom-Nickel-Stähle,* Dissertation, RWTH Aachen 1971.
45. H. Gräfen, G. Böhm, *Z. Metallkd.* **51** (1960) 245 –252.
46. A. Bäumel, E.-M. Horn, H. Gräfen, *Proc. Int. Cong. Met. Corros. 5th* 1972, 934 – 941.
47. A. Kratzer, B. Pieger, H. Tischner, E.-M. Horn: "A Low Carbon Stainless Steel for Use in Nitric Acid at Higher Concentrations and Temperatures," *ICMC, 9th International Congress on Metallic Corrosion,* Toronto, Canada, June 3 – 7, 1984.
48. E.-M. Horn, H. Kohl, *Werkst. Korros.* **37** (1986) 57 –69.
49. E.-M. Horn, A. Kügler, *Z. Werkstofftech.* **8** (1977) 362 – 370, 410 – 417.
50. H. J. Rocha, *DEW-Tech. Ber.* **2** (1962) no. 1, 16 – 24.
51. Mannesmann-Röhren-Werke: *Lexikon der Korrosion,* **vol. 1,** Düsseldorf 1970, p. 55.
52. I. Class, H. Gräfen, *Werkst. Korros.* **11** (1960) no. 9, 529 – 547.
53. G. Herbsleb, *Werkst. Korros.* **19** (1968) 406 – 412.
54. Publication H-2007 and H-2019 of Cabot Wrought Products Division, now Haynes International, 1983/1984.
55. E. Scheil, I. Class, H. Gräfen, DE 1 210 566, 1961, US 3 203 792, 1964.
56. R. B. Leonard, *Corrosion (Houston)* **25** (1969) no. 5, 222 – 228.

57. R. W. Kirchner, W. L. Silence, *Mater. Prot. Perform.* **10** (1971) no. 1, 11 – 15.
58. R. W. Kirchner, F. G. Hodge, *Werkst. Korros.* **24** (1973) no. 12, 1042 – 1049.
59. Kh. G. Schmitt-Thomas, G. Fenzel, *Maschinenschaden* **38** (1965) 94 – 97.
60. E. Fot, *Schweiz. Arch.* **30** (1964) 329 – 344, 384 – 398.
61. E. Kauczor, *Metall (Berlin)* **20** (1966) 1165 – 1168.
62. R. Heidersbach, *Corrosion (Houston)* **24** (1968) 38 –44.
63. "Quaterly Report: Clarification of De-alloying Phenomena," *Corrosion* (Houston) 26 (1970) 445 –447.
64. W. B. Brooks, *Corrosion (Houston)* **24** (1968) 171.
65. L. Leontaritis, E.-M. Horn, *Werkst. Korros.* **31** (1980) 179 – 185.
66. P. Rothenbacher, *Corros. Sci.* **10** (1970) 391 – 400.
67. H. G. Feller, *Z. Metallkd.* **58** (1967) 875 – 885.
68. E. E. Langenegger, F. P. A. Robinson, *Corrosion (Houston)* **25** (1969) 59 – 66.
69. G. Schikorr: "Häufige Korrosionsschäden an Metallen und ihre Vermeidung," in Landesgewerbeamt Baden-Württemberg (ed.): *Schriften zur Gewerbeförderung,* vol. 3, Verlag K. Wittwer, Stuttgart.
70. F. Tödt: *Korrosion und Korrosionsschutz,* 2nd ed., De Gruyter, Berlin 1961, p. 418.
71. D. G. Evans, P. W. Jeffrey: "Exfoliation Corrosion of AlZnMg Alloys," in R. W. Staehle, B. F. Brown, J. Kruger, A. Agrawal (eds.): *Localized Corrosion,* International Corrosion Conference Series, Nat. Assoc. of Corrosion Eng., Houston, Texas 1974, pp. 614 – 622.
72. H. Weingerl, H. Straube, R. Blöch, *Werkst. Korros.* **27** (1976) 69 – 77.
73. E. Schürmann, H.-J. Voss, *Arch. Eisenhüttenwes.* **48** (1977) 129 – 132.
74. A. Diebold, H. Weingerl, *Werkst. Korros.* **28** (1977) 240 – 243.
75. A. Bäumel, *Schweißen Schneiden* **27** (1975) 227 – 230.
76. A. Bäumel, *Werkst. Korros.* **27** (1976) 687 – 693.
77. W. Schwarz, W. Simon, *Ber. Bunsenges. Phys. Chem.* **67** (1963) 108 – 117.
78. K. Gerischer, *Z. Metallkd.* **46** (1955) 661.
79. R. N. Parkins: *5th Symp. Linepipe Research,* Amer. Gas Ass. Cat. No. L 30 174, Report U, Houston 1974.
80. E. Wendler-Kalsch, *Werkst. Korros.* **31** (1980) 534 –542.
81. W. Schwenk: "Einflußgrößen bei der Korrosion von Stählen unter Rißbildung mit besonderer Berücksichtigung der mechanischen Belastungsparameter," *Mannesmann – Forschungsbericht,* no. 944, 1983.
82. H. Diekmann et al., *Stahl Eisen* **103** (1983) no. 18, 895 – 901.
83. R. Münster, H. Gräfen, *Arch. Eisenhüttenwes.* **36** (1965) 227 – 284.
84. W. Rädeker, B. N. Miskra, *Werkst. Korros.* **21** (1970) 691 – 698.
85. H. Gräfen, *Mitt. Ver. Großkesselbesitzer* **73** (1961) 280 – 289.
86. P. Drodten, *Rheinstahl-Tech.* **10** (1972) no. 3, 97 –106.
87. H. Gräfen, D. Kuron, *Arch. Eisenhüttenwes.* **36** (1965) 285 – 291.
88. K. Bohnenkamp, *Arch. Eisenhüttenwes.* **39** (1968) 361 – 368.
89. E. Wendler-Kalsch, *Corros. Sci.* **23** (1983) no. 6, 601 – 612.
90. K. J. Kessler, E. Wendler-Kalsch, *Werkst. Korros.* **28** (1977) 78 – 85.
91. K. Matsukura, K. Sato: "Effects of Metallurgical Factors of Low Carbon Steel Sheets on Stress Corrosion Cracking in Methanol Solutions," *Tetsu to Hagane* **63** (1977) no. 6, 1016 – 1025.
92. E. Wendler-Kalsch: "Korrosion und Spannungsrißkorrosion niedriglegierter Stähle in Methanol, (Project-no. G 6.1/7)," *FEKKs Symposium,* Lahnstein 1987.
93. B. E. Wilde: Stress Corrosion Cracking of ASTM A 517, Steel in Liquid Ammonia: Environmental Factors, National Association of Corrosion Engineers, vol. 37, no. 3 (1981).
94. H. Gräfen et al., *Werkst. Korros.* **36** (1985) 203 – 215.
95. L. Lunde: "Ammonia Plant Safety," *AIChE* **24** (1984) 154; *Symp. on Safety in Ammonia Plants and Related Facilities AIChE,* Denver, CO, August 29 – 31, 1983, report no. iFE/KR/E-83/007.
96. K. Fäßler, H. Spähn: "Grundlagen der Spannungsrißkorrosion unlegierter Stähle in flüssigem Ammoniak – Einflußgrößen und Gegenmaßnahmen," Essen, VGB-Konf. Werkstoffe, 205 – 228 (1989).
97. A. Heuser, H. Spähn, G. H. Wagner, *Materialprüfung* **31** (1989) no. 3, 73 – 79.
98. M. Kowaka, S. Nagata, *Boshoku Gijutsu* **21** (1972) no. 4, 165 – 171.

99. H. Gräfen, H. Schlecker: "CO–CO_2–H_2O," *GWF Gas Wasserfach Gas Erdgas* **126** (1985) no. 4, 195–204.
100. G. Schmitt:, *GWF Gas Wasserfach Gas Erdgas* **122** (1981) no. 2, 49–54.
101. G. Schmitt, H. Schlerkmann: *Met. Corros. Proc. Int. Congr. Met. Corros. 8th,* Mainz 1981, vol. 1, p. 426.
102. H. Gräfen, *Werkst. Korros.* **23** (1972) 247–254.
103. M. O. Speidel in A. J. Sedriks: *Corrosion of Stainless Steels,* Wiley & Sons, New York 1979, p. 173.
104. H. Spähn, G. H. Wagner, U. Steinhoff, *TÜ* **14** (1973) no. 10, 292–299.
105. L. Logan, *J. Res. Nat. Bur. Stand. (U.S.)* **56** (1956) R.P. 2662, 159.
106. A. Beerwald, H. Gröber, *Z. Aluminium* 1940, 502–510. K. Matthaes, *Korrosion* **9** (1958) 5–21.
107. H. Gräfen, *Werkst. Korros.* **16** (1965) 876–879.
108. H. Gräfen, H. Spähn, *Chem. Ing. Tech.* **39** (1967) 138–146.
109. E.-M. Horn, H. Schlecker: unpublished results.
110. W. Dahl, H. Stoffels, H. Hengstenberg, C. Düren, *Stahl Eisen* **87** (1967) 125–136.
111. I. Class, *Werkst. Korros.* **6** (1955) 237–245.
112. W. Hofmann, W. Rauls, *Arch. Eisenhüttenwes.* **32** (1961) 169–171. W. Hofmann, W. Rauls, *Arch. Eisenhüttenwes.* **34** (1963) 925–934.
113. G. Enterlein, M. Kesten, D. Schlegel, K. F. Windgassen, *Z. Werkstofftech.* **13** (1982) 290–297.
114. H. Gräfen, T. Günther, *Z. Werkstofftech.* **10** (1979) no. 11, 373–390.
115. H. Spähn, G. H. Wagner, U. Steinhoff, *TÜ* **14** (1973) no. 9, 260–264.
116. H. Böhm, *Schweiz. Arch.* **33** (1967) 339–363.
117. C. Bosch: Nobel lecture, 21. 5. 1932 in Stockholm, *VDI Z.* **77** (1933) 305–317; *Chem. Fabr.* **6** (1933) 127–142.
118. I. Class, *Stahl Eisen* **80** (1960) 1117–1135.
119. API Refining Department: *Steels for Hydrogen Service at Elevated Temperatures and Pressures in Petroleum Refineries and Petrochemical Plants,* API Publication 941, May 1983 (Copyright 1967 by G. A. Nelson).
120. H. Spähn, *VDI-Ber.* **235** (1975) 103–115.
121. U. Gramberg, T. Günther, H. Palla, *Werkst. Korros.* **26** (1975) 461–464.
122. H. Spähn, *Metalloberfläche* **16** (1962) 267–272.
123. H. Spähn, *Metalloberfläche* **16** (1962) 335–340.
124. K. Risch, *Z. Werkstofftech.* **17** (1986) 6–17.
125. H. Spähn, *Z. Phys. Chem. (Leipzig)* **234** (1967) 1–25.
126. K. Wellinger, K. Lehr, *Mitt. VGB* **49** (1969) 190–201; *Tech. Wiss. Berichte MPA-Stuttgart* (1969) no. 69-02.
127. DIN 50 922 (Oct. 1985), Korrosion und Metalle. Untersuchung der Beständigkeit von metallischen Werkstoffen gegen Spannungsrißkorrosion.
128. C. Wagner, *Z. Phys. Chem. Abt. B* **21** (1933) 25–41; *Z. Phys. Chem. Abt. B* **32** (1936) 447–462.
129. E. Houdremont, G. Bandel, *Arch. Eisenhüttenwes.* **11** (1937/38) 131–138.
130. S. Mrowec, T. Welec, T. Werber, *Oxid. Met.* **1** (1969) 93–120.
131. A. Rahmel, *VDI-Ber.* **235** (1975) 145–154.
132. A. Preece in: *High-Temperature Steels and Alloys for Gas Turbines,* London 1952, (Spec. Rep. Iron Steel Inst. no. 43), pp. 149–152.
133. K. Wickert, *Nickel Ber.* **24** (1966) 177–186.
134. A. Rahmel in H. R. Johnson, D. J. Littler (eds.): *The Mechanism of Corrosion by Fuel Impurities,* Plenum Press, New York 1963, pp. 556–570.
135. U. Jäkel, W. Schwenk, *Werkst. Korros.* **26** (1975) 521–529. A. Rahmel, E. Tarar-Moisescu, *Werkst. Korros.* **26** (1975) 513–520.
136. V. Cihal: "Corrosion Mechanism in Ammonia Synthesis Equipment," *First International Congress on Metallic Corrosion,* London 1962, pp. 591–596.
137. U. Jäkel, W. Schwenk, *Werkst. Korros.* **22** (1971) 1–7.
138. V. van Rossum, *Chem. Ing. Tech.* **25** (1953) 481–487.
139. L. Raichle, *Chem. Ing. Tech.* **28** (1956) 203–213.
140. Mannesmann-Röhrenwerke, Düsseldorf: *Lexikon der Korrosion,* **vol. 1,** pp. 133–134.
141. C. G. Stevens, J. Board, *Br. Corros. J.* **4** (1969) 80–85.
142. D. Goodison, R. J. Harris, P. Goldenbaum, *Brit. Corros. J.* **4** (1969) 293–300.
143. H. Pfeiffer, G. Sommer, *Werkst. Korros.* **13** (1962) 667–677.
144. O. Kubaschewski, B. E. Hopkins: *Oxidation of Metals,* Butterworths, London 1962.
145. J. Möller, *Metall* **26** (1972) 820–825.
146. J. Möller, *Ind. Anz.* **96** (1974) no. 38, 856–857.
147. U. Richter, *Schweißtechnik (Vienna)* **7** (1975) 116–120.

148. DIN 8553 (April 1970): Verbindungsschweißen plattierter Stähle.
149. U. Richter, *Schweißen & Schneiden* **25** (1973) 218–220.
150. J. Chene, P. Luginbühl, *Tech. Rundsch. Sulzer* **2** (1973) 73–78.
151. W. Ruckdeschel, *Schweißtechnik (Zürich)* **63** (1973) 229–241.
152. B. Bouaifi, J. Krohn, U. Draugelates, *DECHEMA-Monogr.* **103** (1986) 165–175.
153. B. Bouaifi, I. Graf, U. Draugelates, *Prakt. Metallogr.* **25** (1988) no. 11, 543–554.
154. D. Fuchs, H. Preisendanz, P. Schüler, *DEW-Tech. Ber.* **13** (1973) 137–142.
155. Pfaudler-Werke AG, Data Sheet 249-1: "Resista-Clad, ein Verfahren, das Metalle dauerhaft verbindet."
156. The Pfaudler Company, Data-Sheet DS49-300-1: "Resista-Clad, Weld-Bonded Cladding Process," 1984.
157. The Pfaudler Company, Data-Sheet DS49-301-1: "Resista-Clad, Physical & Chemical Performance Data," 1985.
158. M. X. Cerny, *Power Eng.* (1987) no. 8, 36–37.
159. W. Burckhardt, *Technik* **26** (1971) 697–704; **27** (1972) 563–570, 528–531; **30** (1975) 239–245.
160. L. L. Shreir, *Ind. Finish (London)* **8** (1955) 261–267, 389–400.
161. L. Winkler, *Metall (Berlin)* **31** (1977) 506–509.
162. J. Wurm, *DECHEMA-Monogr.* **76** (1974) 1486–1504, 81–89.
163. A. W. Berger, *CAV* (1980) 82–84.
164. K. Breitwieser, R. Engelmann, *Maschinenmarkt* **85** (1979) 393–400.
165. W. Paatsch, *Metall (Berlin)* **30** (1976) 332–336.
166. H. Gräfen, *VDI-Ber.* **624** (1986) 273.
167. H. Gräfen: "Werkstoffe des Chemieanlagenbaus, Stand der Entwicklung," in *50 Jahre Werkstofftechnik an der Fakultät für Maschinenwesen,* Colloquy at the Technische Universität München (1989) pp. 38–63.
168. K. Kreisel, E. Protogerakis, *DECHEMA-Monogr.* **103** (1986) 191–205.
169. H. Schindler: "Nichtmetallische Werkstoffe für die Chemietechnik," *Fortschritte der Verfahrenstechnik,* **vol. 15,** VDI-Verlag GmbH, Düsseldorf 1977, pp. 445–459.
170. H. Schindler: "Nichtmetallische Werkstoffe für die Chemietechnik," *Fortschritte der Verfahrenstechnik,* **vol. 18,** VDI-Verlag GmbH, Düsseldorf 1980, pp. 439–456.
171. H. Gräfen, U. Gramberg, H. Schindler, *Werkst. Korros.* **30** (1979) 297–307.
172. A. H. Dietzel: *Emaillierung,* Springer Verlag, Berlin–Heidelberg–New York 1981.
173. R. Lorentz: "Korrosion von Chemieemail durch wäßrigneutrale Medien," *Mitteilungen des Vereins Deutscher Emailfachleute e.V.,* **vol. 34/5,** (1986) 65–76.
174. H. Scharbach: "Glas- und Glaskeramikemail für den technischen Einsatz," *Swiss Chem.* **5** (1983) 55–60.
175. Pfaudler-Werke AG, Schwetzingen, Printed publication FCP5M 12/87.
176. Literature supplied by Pfaudler-Werke AG, Schwetzingen, at ACHEMA 1988 in Frankfurt.
177. E. Kreisel, E. Protogerakis: "Thermisches Spritzen in Chemieapparate-Maschinenbau-Anlagen, Werkstoffe, Anwendungen," *Chem. Ing. Tech.* **59** (1987) no. 2, 118–122.
178. Sigri GmbH, Meitingen, Printed publication 853 01 10, 1983.
179. K. H. Möbius: "Füllstoffhaltige elektrisch leitfähige Kunststoffe," *Kunststoffe* **78** (1988) no. 1, 53–58.
180. A. Malliaris, D. T. Turner: "Influence of Particle Size on the Electrical Resistivity of Compacted Mixtures of Polymeric and Metallic Powders," *J. Appl. Phys.* **42** (1971) no. 2, 614–618.
181. H. Schindler, H. Gräfen, *Chem. Ing. Tech.* **57** (1985) no. 7, 597–602.
182. H. Gräfen: "Fortschritte beim Einsatz organischer Werkstoffe im Korrosionsschutz," *VDI-Ber.* **670** (1988) 451–468.
183. E. Oelke: "Kunststoffe in Chemieanlagen," *VDI-Gesellschaft Kunststofftechnik,* VDI-Verlag GmbH, Düsseldorf 1986, pp. 167–183.
184. T. Hasky in: "Oberflächenschutz mit organischen Werkstoffen im Behälter-, Apparate- und Rohrleitungsbau," *VDI-Gesellschaft Werkstofftechnik,* VDI-Verlag GmbH, Düsseldorf 1980, pp. 101–129.
185. E. Schacht: "Erfahrungen mit Gummierungen und Beschichtungen in Anlagen zur Rauchgasentschwefelung," *VDI-Ber.* **674** (1988) 263–281.
186. A. Swozil, G. Ullmann, *Chem. Ing. Tech.* **52** (1980) no. 4, 292–298.
187. J. E. Niesse: "Innovations in Organic Linings," *Chem. Eng. Prog.* **82** (1986) no. 6, 55–62.
188. G. Dexheimer, H. Schindler, internal report, Bayer AG, 1989.

189. "Bekaplast-Säureschutzverkleidung," company brochure, Steuler Industriewerke, Höhr-Grenzhausen.
190. H. R. Rottkämer in: [7] , 169 – 184.
191. J. I. Bregmann: *Corrosion Inhibitors,* The Macmillan Comp., New York 1963.
192. H. J. Rother: "Die Praxis des Korrosionsschutzes," *Corrosion Inhibitors,* The Maxmillan Comp., New York 1963, pp. 250 – 266.
193. H. Fischer, *Werkst. Korros.* **23** (1972) 445 – 465.
194. K. Risch, *Werkst. Korros.* **25** (1974) 727 – 734.
195. E. Rabald, *Werkst. Korros.* **5** (1954) 368 – 392.
196. Corrosion Inhibition, Proceedings of the Int. Conference on Corrosion Inhibition, NACE Houston 1988
197. D. Kuron, H.-J. Rother, R. Holm, S. Storp, *Werkst. Korros.* **37** (1986) 83 – 93.
198. M. N. Desai, *Werkst. Korros.* **23** (1972) 483 – 487.
199. M. N. Desai, *Werkst. Korros.* **24** (1973) 707 – 716.
200. J. C. Kora, S. C. Makwana, K. C. Koshel, N. K. Patel, *Werkst. Korros.* **25** (1974) 753 – 756.
201. M. Brooke, *Chem. Eng.* **12** (1954) 230 – 234.
202. D. Kuron, H. J. Rother, H. Gräfen, *Werkst. Korros.* **32** (1981) 409 – 421.
203. G. Schmitt, B. Olbertz, *Werkst. Korros.* **29** (1978) 451 – 456. G. Schmitt, B. Olbertz, *Werkst. Korros.* **35** (1984) 99 – 106. G. Schmitt, B. Olbertz, K. H. Kurtz, *Werkst. Korros.* **35** (1984) 107 – 110.
204. H. Gräfen, G. Herbsleb, F. Paulekat, W. Schwenk, *Werkst. Korros.* **22** (1971) 16 – 31.
205. E.-M. Horn, R. Kilian, H. Stiepel, H. Gräfen, *Werkst. Korros.* **23** (1972) 967 – 973.
206. J.-W. Kühn, V. Burgsdorff, H. Richter in W. v. Baeckmann, W. Schwenk (eds.): *Handbuch des kathodischen Korrosionsschutzes,* Verlag Chemie, Weinheim 1971, pp. 342 – 354.
207. H. Gräfen, F. Paulekat in: W. v. Baeckmann, W. Schwenk, W. Prinz, see general references.
208. W. von Baeckmann, *Chem. Ztg. Chem. Appar.* **87** (1963) 395 – 404.
209. U. Heinzelmann in: [206] , pp. 330 – 341.
210. W. P. Banks, J. P. Sudbury, *Corrosion (Houston)* **19** (1963) 300 – 307.
211. J. M. Stammen, *Mater. Proc.* **7** (1968) no. 12, 33 – 35.
212. F. Paulekat, H. Gräfen, D. Kuron, *Werkst. Korrosion* **33** (1982) 254 – 262.
213. D. Kuron, H. Gräfen, *Chem. Ing. Tech.* **60** (1988) 604 – 612.
214. M. J. Humphries, R. N. Parkins, *Corros. Sci.* **7** (1967) 745 – 761.
215. H. Gräfen et al., *Werkst. Korros.* **22** (1971) 16 – 31.
216. J. B. Cotton, *Werkst. Korros.* **11** (1960) 152 – 155.
217. Verein Deutscher Eisenhüttenleute (ed.): *Prüfung und Untersuchung der Korrosionsbeständigkeit von Stählen,* Verlag Stahleisen, Düsseldorf 1973.
218. DIN-Taschenbuch 219,see general references.
219. Annual Book of ASTM Standards 1988, see general references.
220. H. Gräfen, G. Herbsleb, in: [39] , pp. 55 – 69.
221. DIN 50 905:Korrosion der Metalle. Korrosionsuntersuchungen Part 1 Grundsätze; Part 2 Korrosionsgrößen bei gleichmäßiger Flächenkorrosion; Part 3 Korrosionsgrößen bei ungleichmäßiger und örtlicher Korrosion ohne mechanische Belastung; Part 4 Durchführung von chemischen Korrosionsversuchen ohne mechanische Belastung in Flüssigkeiten im Laboratorium (1987).
222. ASTM G 4-84:Standard Method for Conducting Corrosion Coupon Tests in Plant Equipment (1984).
223. ASTM G 31-72:Standard Practice for Laboratory Immersion Corrosion Testing of Metals (1985).
224. ASTM G 78-83:Standard Guide for Crevice Corrosion Testing of Iron-Base and Nickel-Base Stainless Alloys in Seawater and other Chloride Containing Aqueous Environments (1983).
225. DIN 50 915:Prüfung von unlegierten und niedriglegierten Stählen auf Beständigkeit gegen interkristalline Spannungsrißkorrosion (1985).
226. ASTM G 30-79:Standard Practice for Making and Using U-Bend Stress-Corrosion Test-Specimens (1984).
227. ASTM G 38-73:Standard Practice for Making and Using C-Ring Stress-Corrosion Test Specimens (1984).
228. K. Risch, *Werkst. Korros.* **36** (1985) 55 – 63.
229. DIN 50 920:Korrosion der Metalle. Korrosionsuntersuchungen in strömenden Flüssigkeiten. Part 1 Allgemeines (1985).
230. DIN 50 914:Prüfung nichtrostender Stähle auf Beständigkeit gegen interkristalline Korrosion. Kupfersulfat-Schwefelsäure-Verfahren. Strauß-Test (1984).
231. Stahl-Eisen-Prüfblatt 1877: Prüfung der Beständigkeit hochlegierter, korrosionsbeständiger Werkstoffe gegen interkristalline Korrosion, VDEh (1979).

232. DIN 50 921:Korrosion der Metalle, Prüfung nichtrostender austenitischer Stähle auf Beständigkeit gegen örtliche Korrosion in stark oxidierenden Säuren. Korrosionsversuch in Salpetersäure durch Messung des Massenverlustes (Prüfung nach Huey) (1984).
233. ASTM A 262-86:Standard Practices for Detecting Susceptibility to Intergranular Attack in Austenitic Stainless Steels (1986).
234. Euronorm 114-72.
235. ASTM G 28-85:Standard Test Methods of Detecting Susceptibility to Intergranular Attack in Wrought, Nickel-Rich, Chromium-Bearing Alloys (1985).
236. ASTM G 48-76:Standard Test Methods for Pitting and Crevice Corrosion Resistance of Stainless Steels and Related Alloys by the Use of Ferric Chloride Solution (1980).
237. ASTM G 36-87:Standard Practice for Evaluating Stress-Corrosion-Cracking Resistance of Metals and Alloys in a Boiling Magnesium Chloride Solution (1987).
238. NACE Standard TM-01-077.
239. DIN 50 916:Prüfung von Kupferlegierungen. Spannungsrißkorrosionsversuche mit Ammoniak. Part 1 Prüfung von Rohren, Stangen und Profilen (1976); Part 2 Prüfung von Bauteilen (1985).
240. ASTM G 37-85:Standard Test Method for Use of Mattsson's Solution of pH 7.2 to Evaluate the Stress-Corrosion Cracking Susceptibility of Copper-Zinc Alloys (1985).
241. ASTM G 47-79:Standard Test Method for Determining Susceptibility to Stress-Corrosion Cracking of High-Strength Aluminum Alloy Products (1984).
242. ASTM G 44-88:Standard Practice for Evaluating Stress Corrosion Cracking Resistance of Metals and Alloys by Alternate Immersion in 3.5 % Sodium Chloride Solution (1988).
243. DIN 50 917:Korrosion der Metalle. Naturversuche. Freibewitterung (1979).
244. DIN 50 017:Klimate und ihre technische Anwendung. Kondenswasser-Prüfklimate (1982).
245. DIN 50 018:Korrosionsprüfungen. Beanspruchung im Kondenswasser-Wechselklima mit schwefeldioxidhaltiger Atmosphäre (1978).
246. DIN 50 021:Korrosionsprüfungen. Sprühnebelprü-fungen mit verschiedenen Natriumchloridlösungen (1975).
247. ASTM B 117-85:Standard Method of Salt Spray (Fog) Testing (1985).
248. ASTM G 87-84:Standard Practice for Conducting Moist SO_2 Tests (1984).
249. ASTM G 91-86:Standard Practice for Monitoring Atmospheric SO_2 Using the Sulfation Plate Technique (1986).
250. ASTM G 85-85:Standard Practice for Modified Salt Spray (Fog) Testing (1985).
251. DIN 50 918:Korrosion der Metalle. Elektrochemische Korrosionsuntersuchungen (1978).
252. ASTM G 3-74:Standard Practice for Conventions Applicable to Electrochemical Measurements in Corrosion Testing (1981).
253. ASTM G 5-87:Standard Reference Test Method for Making Potentiostatic and Potentiodynamic Anodic Polarization Measurements (1987).
254. DIN 50 919:Korrosion der Metalle. Korrosionsuntersuchungen der Kontaktkorrosion in Elektrolytlösungen (1984).
255. ASTM G 71-81:Standard Guide for Conducting and Evaluating Galvanic Corrosion Tests in Electrolytes (1986).
256. D. Kuron, H. Gräfen, *Z. Werkstofftechn.* **8** (1977) 182 – 191.

Cortisone → Hormones
Corundum → Abrasives
Corundum → Aluminum Oxide
Cosmetics → Skin Cosmetics
Cotton → Cellulose
Cottonseed Oil → Fats and Fatty Oils

Abrasion and Erosion

Klaus Schneemann, Hüls AG, Marl, Federal Republic of Germany

1. Introduction 1
2. Types of Wear and Wear Mechanisms 1
3. Behavior of Materials 8
3.1. Metals 9
3.2. Plastics and Elastomers 13
3.3. Ceramics 15
4. Surface Treatment and Coatings . . . 15
5. Practical Examples of Abrasion and Erosion Damage 17
5.1. Sliding Wear, Elastic Rolling Wear, and Oscillation Wear 18
5.2. Abrasion Wear 19
5.3. Damage by Particle Erosion 20
5.4. Solid Particles-Free Erosion 22
5.5. Damage Caused by Erosion – Corrosion 23
6. References 25

1. Introduction

Practical experience with industrial equipment, machinery, and plant has shown that components have only limited service lives. Damage and ultimate failure of the component can occur as a result of changes in the material that originate at the surface, even if the components are designed such that long-term action of the forces alone causes neither fracture nor undue deformation.

If the reactions responsible for the damage are of electrochemical or predominantly chemical nature, the term corrosion is normally used, whereas mechanical damage to the surface of the component is defined as wear. Attempts to avoid a loss of material due to wear, or at least to reduce the loss, concentrate on making the affected surface more resistant to wear. This can be achieved by mechanical, thermal, or thermochemical treatment of the surface or by applying or depositing metallic coatings. Under some circumstances the wear conditions can be changed by design measures so that the danger for the affected component surface is eliminated or reduced to a tolerable level.

With few exceptions (e.g., running-in of bearings), wear in engineering means an undesired change that causes very high costs every year; in a highly developed, industrialized country this can amount to ca. 1 – 2 % of the gross national product [1].

Plant construction typical of the chemical industry plays an insignificant role, and wear is correctly known as "the problem child of mechanical engineering" [2].

Wear, friction, and lubrication are described under the term tribology as the science of the study, industrial application, and modification of the phenomena and processes occurring between surfaces which are acting against each other and moving relative to one another; this includes boundary surface interactions between solids, and between solids and their gaseous or liquid surroundings. Since at least two components of a system are involved in wear, it is not a pure material characteristic, but only a system characteristic. Wear itself is generally understood as progressive loss of material from the surface of a solid body caused by mechanical action, i.e., contact and relative motion with a solid, liquid, or gaseous phase.

2. Types of Wear and Wear Mechanisms

The treatment of wear must take the diversity of tribological processes into account, and this requires precise analysis of the loads and of the appearance of the damage. It is usual to subdivide the large number of wear processes into types of wear and wear mechanisms, in which different mechanisms have to be allocated to one and the same type of wear.

The kinematic conditions and the types of materials involved in the wear determine the types of wear, such as sliding wear, elastic rolling wear, impact wear, and shock wear.

Oscillating wear stress, or oscillation wear, is caused by oscillating sliding and by oscillat-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_09

Table 1. Classifaction of types of wear (based on DIN 50 320)

System	Tribiological loading		Type of wear		Active mechanisms[a] (individual or combined)			
					Adhesion	Grooving	Surface destruction	Tribochem. reactions
Solid-intermediate material-solid[b]	Sliding, rolling, impact	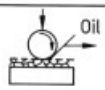					●	○
Solid-solid	Sliding		Sliding wear		●	○	○	●
	Rolling		Rolling wear		○	○	●	○
	Oscillation		Oscillating wear (fretting)		●	●	●	●
	Impact		Impact wear		○	○	●	○
Solid-particle			Abrasive impact wear[c]	Abrasion		●	●	○
	Sliding[d]		Abrasive sliding wear[c]			●		○
Solid-particles-solid	Sliding		Three-body abrasive wear		○	●	●	○
	Rolling	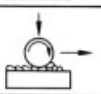			○	●	●	○
	Impact				○	○	●	○
Solid-particle-liquid	Flow	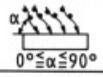	Hydroabrasive wear	Erosion		●	●	○
Solid-particle-gas	Flow		Sliding jet wear		○	●	●	○
	Flow, impact	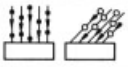	Impact and inclined jet wear		○	●	●	○
Solid-liquid	Flow, vibration		Cavitation erosion				●	○
	Impact	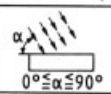	Drop impingement erosion				●	○
	Flow	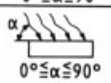	Liquid erosion				○	●
Solid-gas	Flow		Gas erosion					●

[a] ● Principle active mechanisms; ○ additional active mechanisms. [b] With complete separation of the two solid surfaces. [c] Two-body abrasive wear. [d] Also rolling.

ing sliding parts when rolling or even on impact. Wear mechanisms characterize the energetic and material interactions between the individual elements of a wear system. DIN 50 320 differentiates between adhesion, abrasion, surface destruction, and tribochemical reaction as the principle mechanisms (Table 1) [3].

Adhesion. If the micro-roughnesses on the "clean" surfaces of two solid bodies are in contact with each other, high surface pressures are generated locally; relative movements cause cold welding which can lead to separation of material. Material transfer, scales, and shear dimples are typical wear patterns.

Abrasion. When two bodies of different hardnesses slide against each other, abrasion causes the softer surface of the base body to be scratched by the harder counter body. This harder body can be a micro-roughness of the counter body or a wear particle that has already been removed. The penetration of the harder counter body and the relative movement produce wear grooves or furrows in the softer body, for which reason it is also called furrow wear (Fig. 1). If the softer material is plastically deformed and is displaced to the furrow edges, the process is called microplowing. Material is not removed in a single event but by the simultaneous action of many abrasive particles or by repeated furrowing by one particle [4].

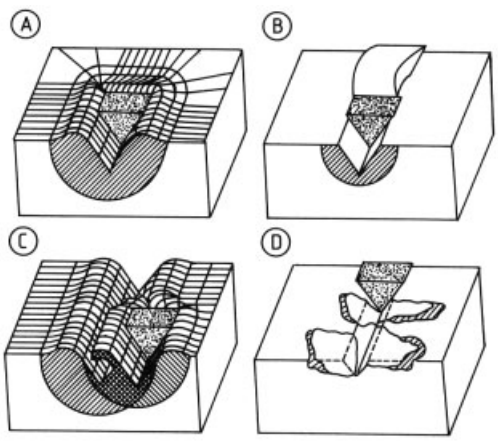

Figure 1. Mechanisms of material damage in furrow wear A) Microplowing; B) Microcutting; C) Microfatigue; D) Microgouging

More often the material is removed by cutting; so-called microchips are obtained. In ductile materials these two processes are predominant, but in relatively brittle materials microparticles break away and microgouging occurs.

Depending on the number of components involved, a distinction is made between two-body and three-body abrasion wear. Sliding, shock, and rolling are possible types of wear, see Table 2 [5].

Oscillating mechanical loads in the surface of a solid can lead to surface destruction by material fatigue. This appears as formation and propagation of cracks, and may lead to separation of material particles.

Tribological stress in the system can lead to tribochemical reactions in addition to purely physical processes. As a result layers can appear on or between the touching surfaces. In general these wear mechanisms do not act in pure form in practice; they usually occur in combinations.

According to [4] the relative proportions of microplowing and microcutting can be determined from the profile of a wear groove (Fig. 2). The so-called f_{ab} value is defined by

$$f_{\mathrm{ab}}=\frac{A_{\mathrm{v}}-(A_1+A_2)}{A_{\mathrm{v}}}$$

so that for the two extreme cases, pure microplowing and pure microcutting, f_{ab} becomes 0 and 1, respectively.

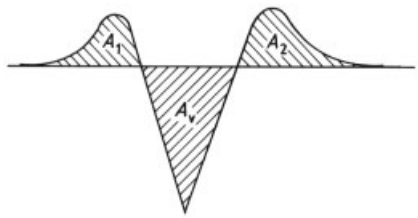

Figure 2. Profile of a wear groove

The linear wear intensity $W_{1/\mathrm{s}}$ is defined as

$$W_{1/\mathrm{s}}=\frac{\Delta h}{s}$$

where Δh represents the reduction of thickness of the worn component and s is the wear path. Using the f_{ab} value allows the wear intensity to be expressed as

$$W_{1/\mathrm{s}}=\varphi f_{\mathrm{ab}}\cdot\frac{p}{H_{\mathrm{def}}}$$

Table 2. Comparison of the characteristics of abrasive sliding, abrasive impact, and three-body abrasive wear

Classification	Abrasive sliding wear			Abrasive impact wear	Three-body abrasive wear		
	Two-body abrasive wear						
Symbol	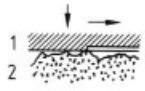	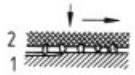	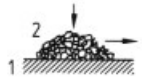	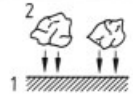	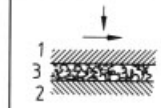	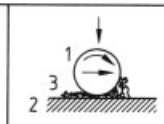	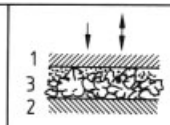
Body 1	Wearing material surface						
Counterbody 2	Surface peaks of solid abrasive	Bound abrasive	Free abrasive	Coarse-grained abrasive	Wearing material surface		
Intermediate material 3					Particulate solid		
Surrounding medium	Liquid or gas						
Relative motion of body and counterbody	Sliding	Sliding	Sliding, rolling, flow	Impact, sliding	Sliding	Rolling	Impact, sliding
Motion of abrasive	Sliding	Sliding	Sliding, rolling	Impact, sliding	Sliding, rolling		Sliding, rolling, impact
Relative wear intensity	High	High	Low	Very high	Low, high	Low, high	Very high
Practical examples	Rock drilling	Pressing of refractory bricks (grinding, honing)	Solids transport in channels, chutes, pipes	Discharge points in solid transport, deflector plates	Undesired processes in bearings, joints		Comminution and grinding processes
Test systems	Tool/stone	Abrasive paper test	Wear pot test	Buffing gear for mass goods	Block/ring block/wheel	Roller/roller roller/surface ball mill	Model jaw crusher

where p is the surface pressure, H_{def} is the hardness of the wearing material in the strain-hardened state, and φ is the form factor. The f_{ab} value is a function of the effective surface deformation caused by the abrading particles, the ductility of the worn material, and the material's strain-hardening behavior [6]. The value decreases with increasing ductility, increasing strain-hardening capability, and decreasing surface deformation.

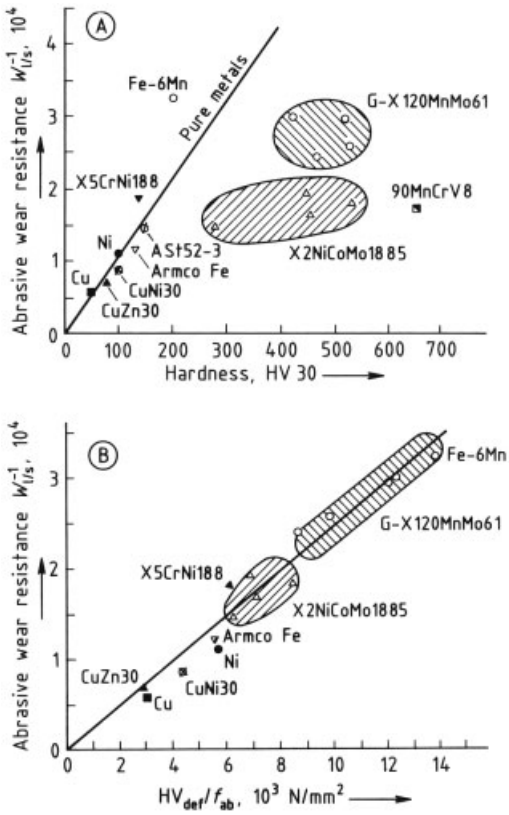

Figure 3. Dependence of the abrasive wear resistance of metallic materials on the hardness of the material (A) and on the ratio of the particle hardness to the factor f_{ab} (B), measured by the abrasive paper method, (SiC, 80 grain, $p = 3.54$ MPa)

According to the above equation, a high wear resistance means a small f_{ab} value and a high material hardness in the strain-hardened state. According to this model, hardness alone often fails to provide a reliable assessment of the expected wear resistance.

Figure 3 A shows the wear resistance $1/W_{1/\mathrm{s}}$ as a function of hardness for various metals, and Figure 3 B shows the resistance as a function of the ratio $\mathrm{HV}_{\mathrm{def}}/f_{\mathrm{ab}}$ in accordance with the above equation. For different materials with the same hardness, the wear resistance increases with increasing strain-hardening capability; see Figure 4 [4].

This assumes that in the test method the wear takes place at the so-called upper shelf of the wear characteristic. In the treatment of wear problems, Wahl was the first to draw attention to the connection between abrasive particle hardness and wear [7]. He found that at a certain value of the particle hardness the wear increases sharply, and at higher values remains virtually constant; see Figure 5.

Note, however, that the pure furrowing described here occurs in tests with bonded particles, a condition which occurs relatively rarely in practical situations. More typical is the case where unbonded, loose particles are carried over the surface of the component, and additional rolling motion is also possible.

With increasing particle mobility, the wear caused by loose particles represents the transition from abrasive sliding wear to erosion (abrasive – erosive attack).

Erosion. Erosion phenomena are caused by flow, in which the flowing fluid itself is erosive or the fluid carries particles that are unable to follow the filaments of flow at surface irregularities. In this case a shock component is often superimposed on the plowing mechanism; this causes local deformation and destruction.

If the fluid is a gas, tribochemical reactions with the solid are only possible at fairly high temperatures where the ablation mechanism can act through the processes of sublimation, vaporization, or fusion.

Pure liquid erosion occurs relatively rarely because even at high flow velocities the forces acting are hardly sufficient to cause removal of material; exceptions to this are additional corrosive elements and the involvement of cavitation.

Cavitation. Since the term cavitation is used to describe the processes occurring in the liquid and also the resulting damage to the material surface, the latter is often called cavitational erosion to avoid confusion. These processes are caused by the formation and subsequent collapse of vapor or gas bubbles in liquids. When the pressure falls, dissolved gases evolve from small gaseous

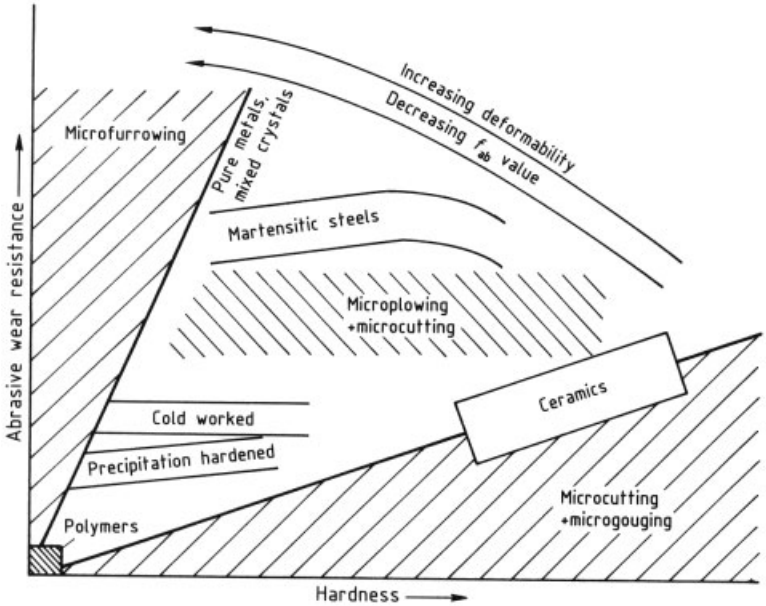

Figure 4. Schematic representation of the abrasive wear resistance in the upper shelf and the wear mechanisms plotted versus the material hardness

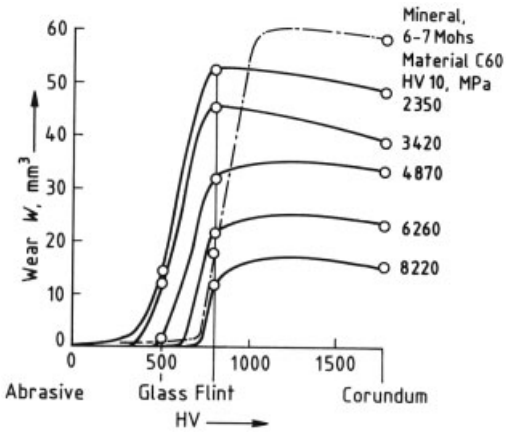

Figure 5. Lower and upper shelf wear characteristic (abrasive paper method, 80 grain, $p = 0.37$ MPa)

nuclei, or evaporation takes place if the pressure falls locally below the vapor pressure. In regions of higher pressure in the liquid the bubbles implode. If this implosion takes place directly at the surface of the component, the cyclic, locally confined compressive stresses finally lead to microscopic fatigue phenomena. This resembles the behavior of materials under vibratory stress. If the vapor bubbles formed do not implode directly at the surface the shock intensity is attenuated by the interposed liquid layer [8]. Damage is therefore only caused by bubbles which collapse directly at or close to the surface.

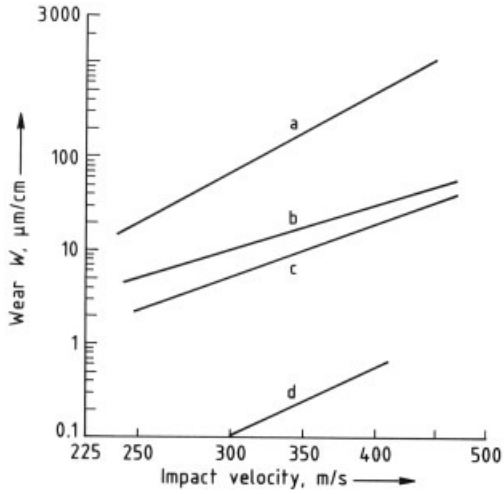

Figure 6. Dependence of drop impingement wear (wear depth per centimeter impinging liquid column) on the impact velocity for various materials
a) Glass; b) Aluminum; c) Polyurethane; d) Sintered alumina

If the pressure differences are caused solely by flow, the resulting process is called flow cav-

itation. However, vibrations of sufficiently high frequency can initiate vibrational cavitation.

The material damage described can take place by purely mechanical means, but in practice it is often considerably intensified by superimposed corrosion stress.

Comparable damage can occur through the repeated impact of drops or liquid jets at high velocity. After the deformability has become exhausted in tough materials, initial cracks appear; these develop into breaks and progress to deep fissuring. This wear, known as drop impingement, depends on the impact velocity (Fig. 6) and the angle of impact [9].

In chemical technology and process engineering, the most important instances of erosion are those where the fluid contains solid particles. If the fluid is a liquid this type of wear is called hydroabrasive wear or scouring wear; with gaseous fluids it is called jet wear. In both types of wear, plowing is active as the essential mechanism, and with steeper angles of impact there is an additional component causing destruction of the material. Tribochemical reactions also occur.

During hydraulic transport of solids in pipelines, the flow path lines near the surface travel essentially parallel to the surface. The erosive effect of the particles is generally slight provided that the flow remains undisturbed. Flow separation at irregularities causes turbulence, which can force the particles against the wall and produce erosion. Typical irregularities in pipelines are the protruding roots of welds. Furthermore, in turbulent flow more collisions between particles occur, leading to impacts with the enclosing walls.

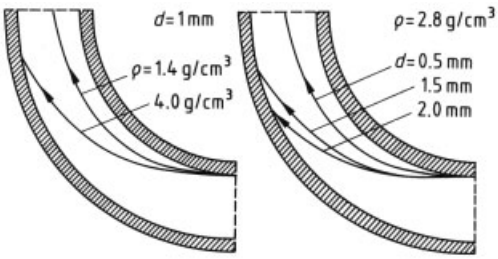

Figure 7. Path of motion of hydraulically transported particles of various densities and diameters (calculated for water at 18 °C and a flow rate of 2 m/s)

During hydraulic transport the particles follow the streamlines of the fluid better than in pneumatic transport due to the smaller difference in densities. Sliding jet wear is always present in horizontal pipeline transport due to the influence of gravity.

Where changes in direction of flow occur, (e.g., at pipe bends) the particles, due to their higher density, are pressed to an increasing extent against the surface, where they have an erosive action (Fig. 7) [10].

The wear increases distinctly with increasing velocity and can be described by the empirically determined expression

$$W(\alpha) = C v^{n(\alpha)}$$

where α is the angle of impact, C is a constant that depends on the material and the abrasive, v is the velocity, and n is a velocity exponent which depends on the material, the abrasive, and the angle. This equation is also valid for jet wear, in which material is removed by impacting and/or plowing particles in a carrier gas. Because of the large number of forms which occur in practice it is necessary to distinguish between the sliding jet, in which the particle slides over the surface of the workpiece, and the impact jet, in which the particle strikes perpendicularly; all intermediate forms of jet ($0° \leq \alpha \leq 90°$) are called inclined jets.

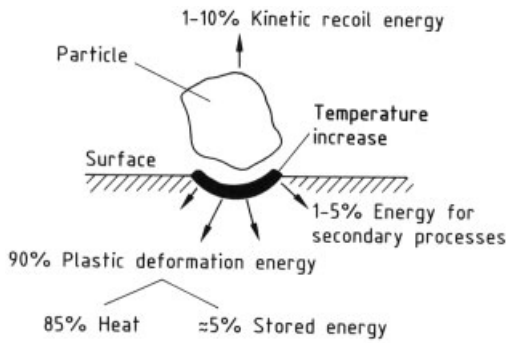

Figure 8. Energy balance for the impact, without breakage, of a hard particle on a ductile metal surface

The basic process in jet wear is the impact of a particle against the surface at some arbitrary angle. About 90 % of the kinetic energy of the particle is converted into plastic deformation and heat (Fig. 8) [11]. The main energy conversion takes place in the softer body [12].

Estimates give values of a few thousand megapascals [13] for the pressures occurring on impact and of microseconds [14] for the impact times (Fig. 9). The fatigue mechanism is dominant where the impact is predominantly perpendicular, while at shallow incident jet angles the particle has a plowing action on the surface under the influence of a normal force component [15]. These result in different requirements for the material involved:

1) Very hard materials are best for withstanding plowing processes
2) Impact processes cause little damage if the energy can be absorbed elastically or plastically.

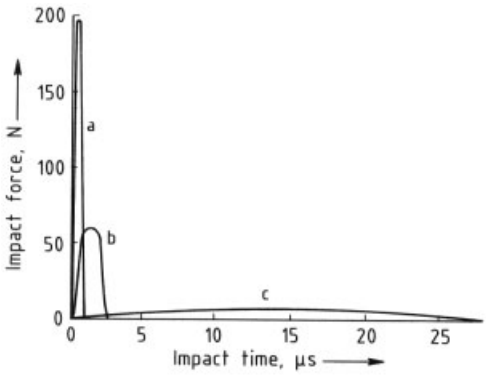

Figure 9. Variation of the calculated impact force of a sphere (diameter 1 mm, velocity 100 m/s) with time for:
a) C60H; b) St 37; c) Rubber

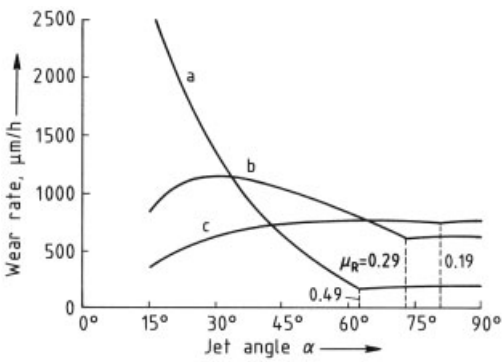

Figure 10. Influence of the jet angle on the wear rate
a) Rubber; b) St 37; c) C60H

Since hard materials normally have low ductility, wear is low at small angles, but fragments are removed easily with perpendicular impact. Soft, ductile materials exhibit higher wear under furrowing conditions. At $\alpha = 0°$, the vertical component of the force theoretically approaches zero, so after a maximum the wear should also approach zero, but this situation does not actually occur in practice (Fig. 10) [15].

Erosion – Corrosion. Under certain corrosive conditions many metals form covering layers. If these are sufficiently dense they act as protective films against corrosive removal of material. An example of this is the protective layer of iron oxide formed in unalloyed or low-alloy boiler tubes. Erosion – corrosion is understood as the combined action of mechanical surface removal and corrosion. With some soft and loose layers the shear forces obtained with pure flowing liquids at medium flow velocities are sufficient to damage the protective layer without the involvement of abrasive solid particles. Where drop impingement or cavitation are involved the mechanical removal of material is understandable. On the other hand, diffusion-controlled corrosion inhibition may be nullified in flowing fluids as a result of saturation of the boundary layer [16]. The resulting purely chemical removal of material therefore no longer corresponds to the usual term of erosion – corrosion and is now called flow-induced corrosion [17].

3. Behavior of Materials

The many outward forms assumed by wear mechanisms make it necessary to design the component to be resistant to wear, at least on the surface under attack. This can be accomplished by selecting a suitable base material or by modifying the surface by mechanical, thermal, or thermochemical treatment or by applying protective coverings of other materials [18]. Adequate wear resistance of a component is normally only one of several criteria in the requirement profile. Adequate strength, toughness, corrosion resistance, and other properties, such as ease of repair, are requirements which must also be met. Economic criteria also have to be considered. In general, theoretical considerations and laboratory tests do not lead to a quick solution, and in most cases the final suitability is proved under operational conditions. Wear problems can often be overcome by changes in de-

sign or in the process technology. Metals, plastics, ceramics, and protective coatings or layers are compared below purely from the aspect of material technology.

3.1. Metals

There is still no comprehensive correlation available between wear and hardness of materials. Depending on the wear mechanism, other properties, such as the state of the material's microstructure, also play a significant role. Because of the simplicity of the hardness test it is, however, useful to know which wear processes depend essentially on hardness alone. In any case, the loading of the material is confined to the surface region, and in abrasive wear a penetration process occurs that is similar to the hardness tests. Unlike hardness tests, an additional tangential force component also has to be taken into account in the case of wear, which initiates effects such as increase in temperature (possibly associated with phase changes), reaction layer formation, and strain hardening by plastic deformation, and can therefore lead to permanent changes in the tribological system.

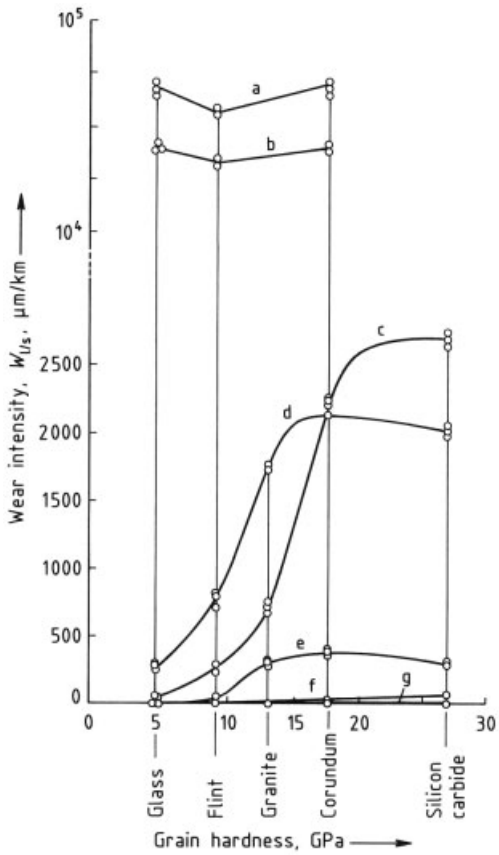

Figure 11. Dependence of wear intensity on the hardness of the abrasive particles for counterbody furrowing for various materials (abrasive paper method, $p = 0.1$ MPa, $v = 0.25$ m/s) a) Rubber; b) Polystyrene; c) Cast basalt; d) Steel St 37; e) C60H; f) Hard metal G4; g) Hard metal H2

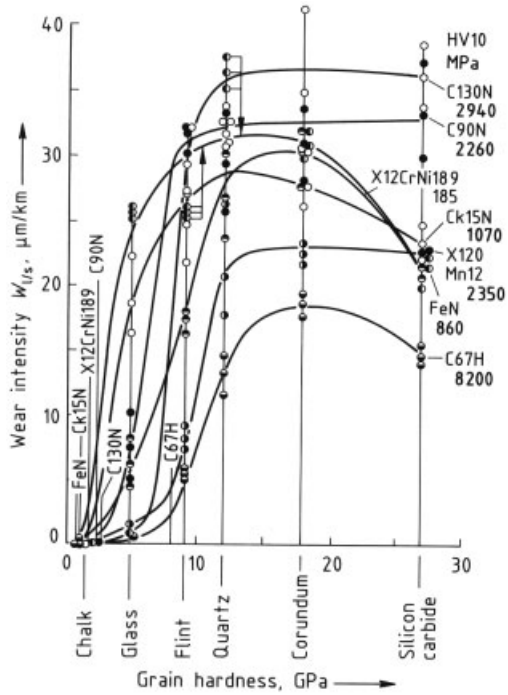

Figure 12. Dependence of wear intensity on the hardness of the abrasive particles for particle furrowing for various materials

It is usual to subdivide furrow wear into:

1) counterbody furrowing, in which the mineralogical grains are fixed on the surface of the counterbody (e.g., a grinding wheel), and
2) particle furrowing, in which the wear is caused by freely moving particles.

In trials both types show a similar dependence of the wear rate on the hardness of the attacking particle (lower shelf/upper shelf characteristic); see Figs. 11 [19] and 12 [20].

For steels an increase in wear resistance is observed with increasing hardness, whereby in-

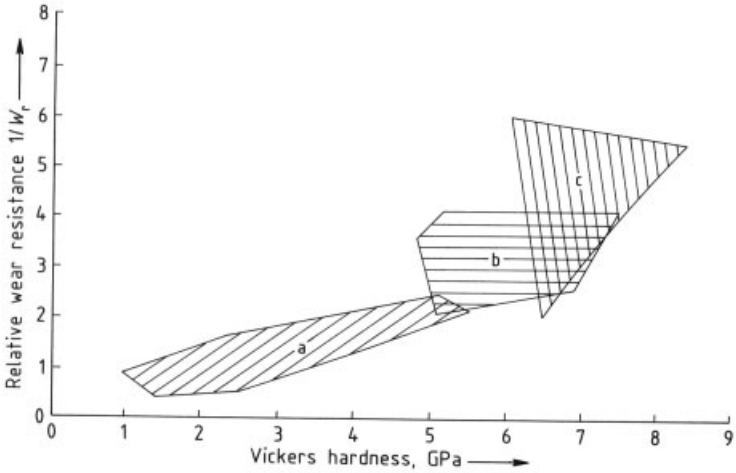

Figure 13. Wear resistance of cast iron towards particle furrowing (wear pot method, quartz sand)
a) Martensitic cast iron alloyed with Cr and Ni; b) Unalloyed and low-alloy pearlitic cast iron; c) Steels with 0.15 – 0.5 % C and gray cast iron with lamellar or spheroidal graphite

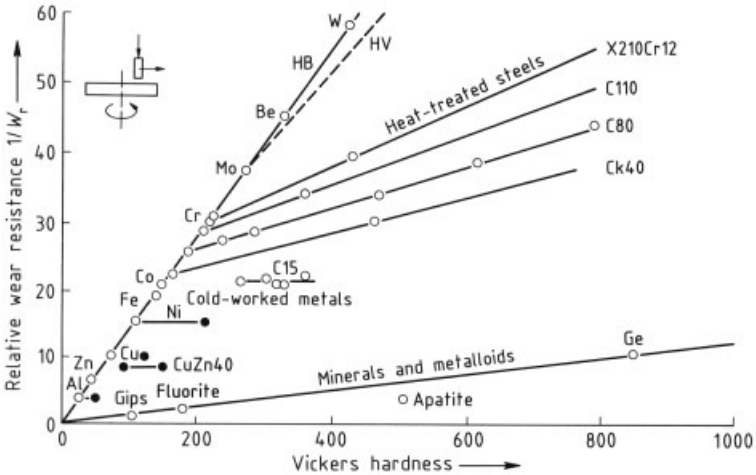

Figure 14. Linear relationship between wear resistance and hardness in the upper shelf region (abrasive paper method, corundum, 180 grain, $p = 0.94$ MPa)

creasing the hardness by alloying elements has a greater effect than increasing the hardness by heat treatment.

With cast iron the wear resistance also tends to increase with increasing hardness. The maximum resistance values increase in the following order: gray cast iron with lamellar or spheroidal graphite, pearlitic cast iron, and martensitic cast iron (Fig. 13) [21].

In the upper shelf region and in the presence of normal abrasive particles, all pure metals exhibit a linear relationship between resistance and material hardness (Fig. 14) [22].

Jet Wear. Materials in chemical plants are usually transported through pipelines and converted in enclosing containers. Compressors and fans are used as the conveying units when gaseous materials are transported, while pumps are normally used for liquids. The resulting wear phenomena are therefore associated with the type of flow involved, and the various types of wear are covered by the term erosion.

Even at very high velocities pure gases are only capable of damaging the enclosing solid body when the temperature is high enough to lead to thermally induced reactions. Problems with gas conveying are therefore only to be expected when entrained solid particles are present. In jet wear the material is elastically or elastically/plastically deformed by impact of the particles against the surface.

Apart from particle velocity the most important factor in jet wear is the jet angle: brittle metals exhibit increasing wear with increasing angle of impact, with a maximum at 90°; whereas with tough metals, wear increases up to an angle of 15° – 40°, after which the loss decreases at steeper angles. The ductility and strain hardening at the surface are thought to be responsible for this behavior (Fig. 15) [23].

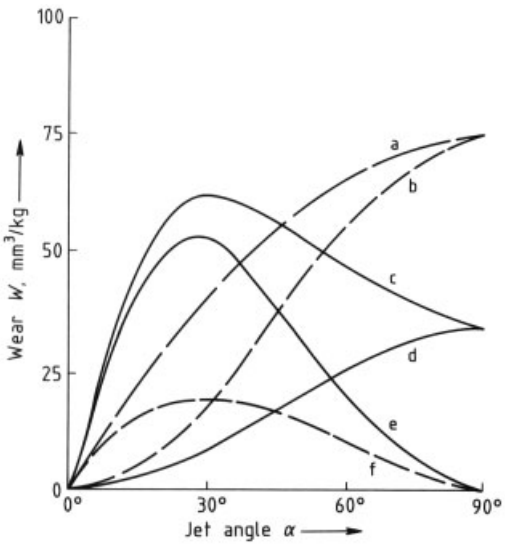

Figure 15. Impact jet and sliding jet components in jet wear a) Experimental curve for cast iron (2.8 % C, 21 % Cr); b) Experimental curve for St 37; c) Impact jet component for cast iron; d) Impact jet component for St 37; e) Sliding jet component for cast iron; f) Sliding jet component for St 37

The relationship between wear resistance and material hardness known from abrasion is valid only at small jet angles because of the comparable wear mechanisms. According to Bitter [24], abrasion is dominant with the sliding jet and surface destruction with the impacting jet.

For the first-named conditions Wahl [25] gives an appraisal of working characteristics (Table 3) which relates predominantly to steel and cast iron.

Wear with Liquids. The forces that particle-free liquids exert on metal surface are so low at the usual flow velocities of a few meters per second that they do not cause any removal of material. According to Heil, no erosion effects could be detected with plain carbon steel, even at 60 m/s, in a test method similar to that with the rotating disc [26]. Damage caused by liquid flow alone is therefore relatively rare in practice. The situation is different if, for example, a liquid under high pressure emerges with high velocity into a space of lower pressure; it may have an erosive action on the metal at the exit point. This applies to an even greater extent if the fluid forms dense protective layers with the metal. If these layers are removed mechanically, they must be reformed, otherwise erosion will act in combination with corrosion, which is known as erosion – corrosion. This situation occurs very frequently in practice as many metals can only be used, for example, with water because a thin, usually oxidic layer protects the metal against corrosive attack. Table 4 gives guide values for maximum permissible flow velocities for pure water and, for comparison, for the more aggressive seawater [27].

Since technical pure liquids usually contain very fine solid particles, it is often difficult in practice to differentiate between pure fluid erosion and so-called hydro-abrasive wear. As with jet wear, the number, hardness, and velocity of entrained particles determine the extent to which the protective covering layer is removed and the extent of the pure metal wear. If mechanical stressing due to entrained particles is predominant then hardness and strength are of decisive importance for metals, as shown in Figure 16 for copper alloys [28].

Cavitation and Drop Impingement. According to Grein all materials succumb to the exceptionally high mechanical stress occurring during cavitation if the intensity is sufficiently high [29]. In soft materials such as aluminum, cavitation erosion appears in the form of plastic indentations caused by single implosions. With higher strength metals the damage starts by roughening of the surface after a certain incubation time; as the destruction progresses the

Table 3. Performance properties of wear resistant materials

Material	Tensile strength, MPa	Fracture strain, %	Wear resistance
Low-alloy austenitic manganese steel	600	15	very low
Austenitic manganese hard steel	550	50	medium
Unalloyed steels	420	20	extremely low
High-strength low-alloy steels	600	18	extremely low
Pearlitic steels	500 – 600	12	extremely low
Martensitic steels	500 – 600	15 – 20	medium
Ledeburitic steels	600	2	high
Heat-treated 66 (gray cast iron)	250	< 0.5	very low
Heat-treated 666 (nodular iron)	500	3	low
Pearlitic white cast iron	250	< 0.5	high
Martensitic white cast iron	350	< 0.5	very high
High-chromium white cast iron	500	< 0.5	very high
Co – Cr – W – C alloy	500	< 0.5	very high
Hard metal	500	< 0.5	extremely high

interconnected holes eventually form a spongy structure.

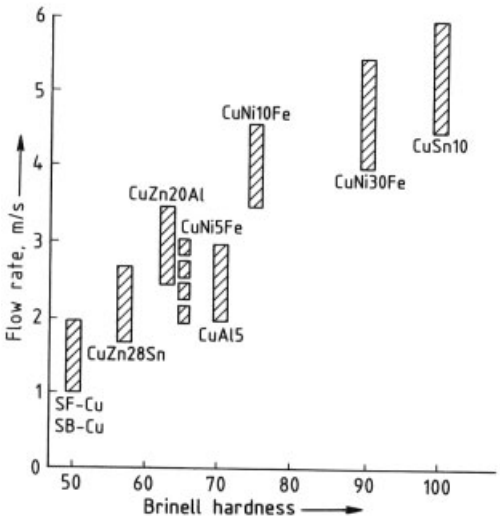

Figure 16. Maximum tolerable flow rates (water) as a function of material hardness for various copper alloys

In general, the strength and the fatigue strength are the most important material parameters for resistance to cavitation; hardness is therefore often used as a measure. With similar microstructures, the metal with the higher hardness has the more favorable behavior, and where different metals have the same hardness, the resistance improves with increasing toughness. As cavitation attack takes place locally, the behavior is determined by the structural constitution to a greater extent with this type of erosion than with others. A homogeneous, fine-grained microstructure is required; soft inclusions and soft, or very brittle, grain boundaries reduce the resistance.

Table 4. Maximum permissible flow rates for pure water and seawater

Material	Flow rate, m/s	
	Pure water	Seawater
Aluminum	1.2 – 1.5	1.0
Copper	1.8	1.0
Copper + As	2.1	1.0
Copper + Fe	4.0	1.5
CuZn28Sn	2.0 – 2.4	1.5 – 2.0
Al bronze	ca. 3.0	ca. 2.0
CuNi10Fe	5.0	2.4
CuNi30Fe	6.0	4.5
Steel	3 – 6	2 – 5
Nickel alloys	30	15 – 25
Plastics	6 – 8	6 – 8

The following factors increase the resistance to cavition erosion:

1) High corrosion resistance
2) Homogeneity
3) Ductility
4) Compressive residual stress
5) High strain-hardening capability
6) Smooth surface
7) Fine-grained structure
8) Fine-grained hard inclusions
9) High content of hard components
10) Layer structure
11) Wrought structure

while the following reduce the resistance:

1) Low corrosion resistance
2) Heterogeneity
3) Brittleness
4) Tensile residual stress

5) Low strain-hardening capability
6) Rough surface
7) Coarse-grained structure
8) Coarse-grained hard inclusions
9) High content of soft components
10) Dendritic structure
11) Cast structure

The data provided by PILTZ [30] and REINGANS [31] can be used as guides for the selection of materials.

As with other types of erosion, the superposition of a corrosion process also has to be taken into account where damage by cavitation occurs. Removal of material by corrosion after destruction of protective covering layers often represents the more intensive attack, and the corrosion resistance of the material is then the dominant property.

A comparable stress occurs with drop impingement, characterized by repeated, short-time liquid impacts. The comments made above about cavitation also apply here to the material behavior.

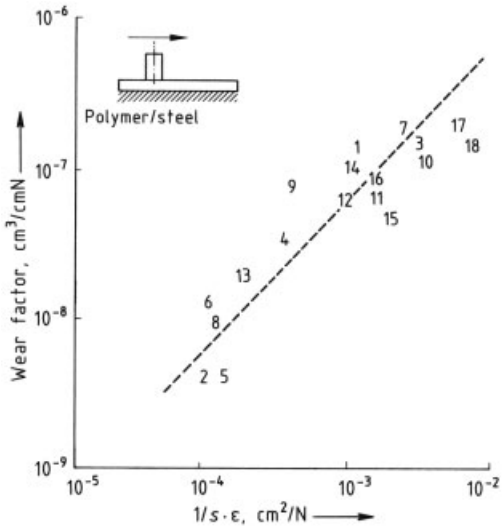

Figure 17. Relationship between wear and the reciprocal product of tensile strength S and fracture strain ε for polymers
1) Poly(methyl methacrylate); 2) Low-density polyethylene; 3) Polystyrene; 4) Polyoxymethylene; 5) Polyamide 66; 6) Polypropylene; 7) Epoxy resin; 8) Polytetrafluoroethylene; 9) PMMA – acrylonitrile copolymer; 10) Polyester; 11) Polychlorotrifluoroethylene; 12) Polycarbonate; 13) Polyamide 11; 14) ABS; 15) Poly(phenylene oxide); 16) Polysulfone; 17) Poly(vinyl chloride); 18) Poly(vinylidene chloride)

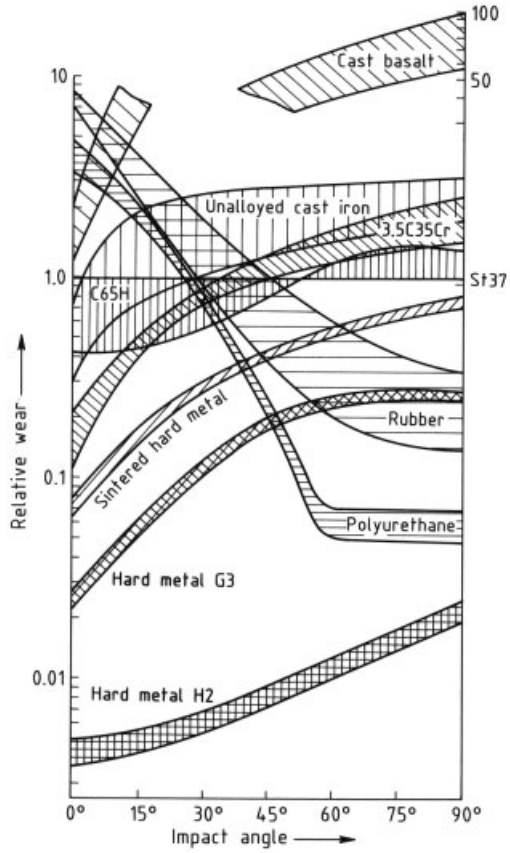

Figure 18. Wear relative to St 37 as a function of impact angle for various materials (abrasive: quartz sand, particle size 0.2 – 1.5 mm, HV = 1250)

Since residual stresses counteract the external loading of the material they can increase resistance to destruction perceptibly. In this sense, strain hardening has a similar effect to carburization and nitriding [32].

3.2. Plastics and Elastomers

The relationship between wear and surface hardness obtained for metals would predict a comparatively poor behavior for polymers. However, their special structural features give rise to properties that can play a special role in wear.

The viscoelastic deformation behavior is characterized by time-, temperature- and velocity-dependent deformation processes. Relatively low levels of hardness and strength, high plasticity, low thermal conductivity, and

Table 5. Jet wear of polymers and metals

Material	Hardness	Abrasive *	Relative wear W/W_{St37}
Steel T 80 H	590 HV	II	0.109
Polyurethane	18 Shore D	II	0.143
Poly(vinyl chloride)	5 Shore D	II	0.143
Polyurethane	34 Shore D	II	0.403
Poly(vinyl chloride)	10 Shore D	II	0.42
Rubber	17 Shore D	II	0.57
Poly(vinyl chloride)	14 Shore D	II	0.96
Steel St 37	122 HV	I	1.0
Low-pressure polyethylene	60 Shore D	I	1.06
Steel St 34	124 HV	I	1.07
Poly(vinyl chloride)	17 Shore D	II	1.12
Polyamide 6, Grilon R 50	62 Shore D	I	1.33
Polyamide 6, Grilon R 70	64 Shore D	I	1.33
Copper	99 HV	I	1.36
High-pressure polyethylene	42 Shore D	II	1.4
Low-pressure polyethylene	58 Shore D	II	1.4
Polyamide 11, Rilsan Besvo	71 Shore D	I	1.81
Low-pressure polyethylene	58 Shore D	I	2.0
Low-pressure polyethylene	60 Shore D	II	2.0
Polyamide 6, Ultramid	70 Shore D	I	2.21
Aluminum	39 HV	II	2.68
Brass	150 HV	I	2.76
Aluminum	29 HV	II	3.23
Polyamide 11	69 Shore D	I	3.31
Poly(vinyl chloride)	52 Shore D	II	4.2
Poly(vinyl chloride)	78 Shore D	II	6.3
Resitex	89 Shore D	II	8.2
Poly(vinyl chloride)	76 Shore D	II	8.5
Glass	6 – 7 Mohs	II	9.7
Lead	4 HV	II	10.5
Plexiglas	85 Shore D	II	10.75
Pertinax	92 Shore D	II	18.5
Epoxy resin with glass fiber	86 Shore D	II	19.5
Epoxy resin with hardener and quartz powder	84 Shore D	II	31

* Abrasive I: sand, HV = 5000 MPa, particle size ≤ 0.9 mm; Abrasive II: sand, HV = 7200 – 8100 MPa, particle size 0.3 – 0.5 mm

high thermal expansion are effects of the weak secondary bonding forces between the macromolecules and their coiled structures.

In particular, the low tendency to adhesion gives polymers their good slip characteristics with steels as the sliding partners – in the absence of additional abrasive particles – because of the low frictional forces involved, and the slip system is characterized by additional emergency running properties. Polyamide and PTFE occupy the prime positions here as they possess good cohesive linkage properties compared with other unreinforced polymers [33], [34].

If abrasive sliding stress is present, the dependence on hardness known for metals cannot really be depicted in the same way. It has been demonstrated that polymers exhibit a good relationship between wear resistance and crack propagation energy, or even between wear and the product of tensile strength and fracture strain; see Figure 17 [35].

Due to their material properties, polymers have proved successful where streams of small particles cause impact stress in addition to sliding wear, i.e., with abrasive impact wear and with erosive attack. Although polymers generally have poor resistance to abrasive sliding attack, their ductility, especially of elastomers, leads to a behavior superior to that of metals when the impacting component is dominant (Fig. 18) [36]. Their behavior therefore differs significantly depending on the angle of impact. The material becomes heated due to internal friction, which can lead to complete failure at high jet intensities.

The preferred elastomers include the polyurethanes and synthetic rubbers because of their outstanding resistance to wear. In polyurethanes, greater resistance is found in the hardness range 70 – 95 Shore A, whereas normal grades of rubber reach their optimum between 50 and 70 Shore A [37]. It is not possible to

separate the influencing factors systematically with respect to tribological behavior because of the large number of additives, types of rubber, and applications (see Table 5) [38], [39].

If, for rubber and C 60 H steel, the amount of wear relative to St 37 steel is plotted versus the impact angle and the hardness of the jet material then, according to [40], it is possible to show the very different wear behavior of these two materials (Fig. 19).

3.3. Ceramics

Ceramic materials have gained increasing importance over the last few decades; in addition to corrosion resistance and suitability for high temperatures the resistance to wear is the important property of these materials. The abovementioned properties of the sintered body can be varied over a wide range by adjusting the powdered raw material.

The nonmetallic, mechanically resistant materials which are constituents of the ceramic materials can be subdivided into oxidic and nonoxidic materials [41]. High melting points and hardnesses are the outstanding properties, as shown in Table 6. :

Table 6. Hard materials

Material	*mp*, °C	ϱ, g/cm^3	Vickers hardness $HV_{0.2}$
Oxides			
Al_2O_3	2050	3.9	2300
ZrO_2	2700	5.7	1100
TiO_2	1860	4.2	1000
Carbides			
TiC	3150	4.9	3200
TaC	3780	14.5	1790
WC	2600	15.7	2080
SiC	2180	3.2	3000
B_4C	2450	2.5	3500
Nitrides			
TiN	2950	5.2	2450
TaN	3090	13.8	3230
Si_3N_4	1900	3.2	1400
BN (cubic)	3000 *	3.48	8000
Borides			
TiB_2	2900	4.4	3480
ZrB_2	2990	6.0	2200
Carbon			
Diamond	3750	3.5	8000 – 10 000

* Sublimes.

Alumina is the most important oxidic abrasion-resistant material. Metal carbides are in some ways superior to oxides with respect to hardness and melting point, but they are much more brittle than the oxides and are only used in isolated instances as wearing bodies. Silicon carbide is characterized by its low thermal expansion and high thermal conductivity and has proved to be more resistant to thermal shock than oxides. Zirconia is tougher than alumina; its modulus of elasticity is only about half as large and is comparable with that of steel. Zirconia is therefore very suitable for compound structures with steel. At present the applications of ceramic sintered materials in chemical plant construction are slide rings, pump parts, and slide bearings.

4. Surface Treatment and Coatings

The fact that wear starts at the surface of the workpiece suggests that only the tribologically stressed surface should be strengthened instead of making the entire component out of wear-resistant material. It should be borne in mind that the base material has to fulfill other, for example, load-bearing functions as well as surface stressing, i.e., the surface which has been made wear-resistant must not interfere unduly with the component.

If the material involved can be hardened then various methods can be used to harden the surface to a certain depth by structural transformation, making it resistant to wear. These include induction, flame hardening, HF impulse, electron, and laser jet methods. In case hardening, a steel which originally cannot be hardened is made hardenable by inward diffusion of carbon, sometimes also with nitrogen; hardening is then carried out by rapid quenching. The hardness achieved is governed essentially by the contents of dissolved C and N; the usual depths of application are 0.5 to 2 mm.

Nitride layers are produced after treatment in a salt bath or in a gas atmosphere by inward diffusion, usually below 600 °C. Unlike transformation hardening, which is effected by lattice distortion as a result of embedded carbon atoms, nitriding and boriding provide the surface with an exceptional increase in hardness by the formation of an intermetallic bonding layer. These bonding layers also increase the general corrosion resistance, but they are so thin that they can break by mechanical point loading.

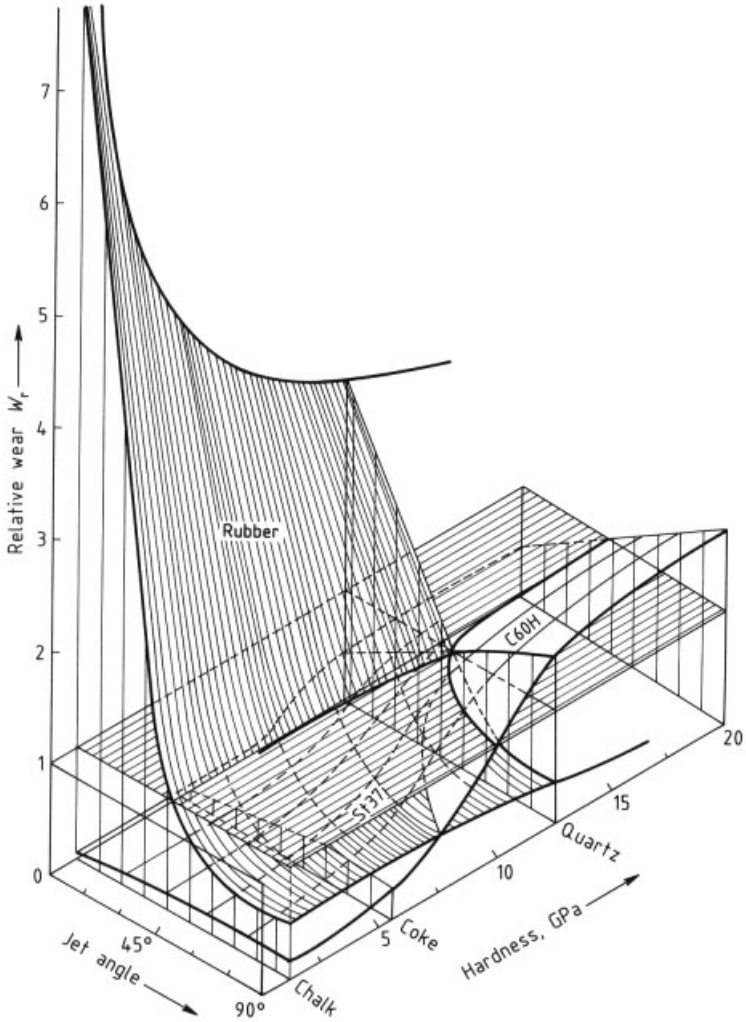

Figure 19. Dependence of jet wear on the jet angle and the hardness of the impinging mineral

Hardfacing, on the other hand, is a thick layer process. By means of various welding techniques, wear resistant alloys in rod, wire, or powder form build strong metallurgical bonds in the fused state with the surface, which is also fused. Mixing should be kept as low as possible. There is an exceptionally large variety of additives available on the market and, among other things, the requirement for mechanical workability must be taken into account. Most filler metals are alloyed on an Fe-base, but Ni or Co can also be the main constituents where high-temperature and/or additional corrosion resistance is required. Where flat or rotationally symmetrical surfaces have to be coated over large areas, automatic or fully-mechanized welding is much cheaper than manual welding [42].

Sheet metal plating by roll-bonded or explosion plating is not always as effective as weld plating with respect to adhesion. As the plating material must be very plastically deformable to achieve an intimate bond with the supporting material, highly wear resistant and high-strength materials are unsuitable for the purpose. Nevertheless, this method is widely used in chemical technology because to a large extent corrosion

stresses are superimposed on the wear processes. The selection of highly corrosion resistant alloys applied to mechanically heavily stressed base bodies provides satisfactory solutions to the majority of problems in the construction of vessels and pipes [43].

Thermal spraying can cause even greater adhesion problems than those of the above-mentioned coating methods. Unlike hardfacing, after the coating additive has been fused in a flame, arc, or plasma it is projected against the cold or only slightly heated base. Adhesion to the surface and the density of the layer itself can differ greatly depending on the intensity of the spraying and the possibility of reaction with the surrounding medium on the way to the base material. From a micrograph it can be seen that, depending on the process used, the layers have a particular structure caused by agglomeration of many fine globules of molten material, which become deformed on impact and leave a visible porosity in the composite material.

The most important groups of materials for wire flame spraying and arc spraying are

1) Low- and high-alloy steels
2) Bronzes, nonferrous metals such as nickel, copper, zinc, Monel metal, brass, aluminum

Powder flame spraying is a versatile method of thermal coating and also handles

1) Special steel alloys with embedded mechanically resistant materials
2) Ceramics with fairly high melting points
3) So-called self-flowing Cr – Ni – B – Si alloys which, after the actual spraying process – without any adhesion layer – are submitted in a second operation to subsequent treatment at about 1100 °C (liquid-phase sintering); this produces virtually pore-free layers with good adhesion characteristics.

Substantially higher temperatures are reached in the plasma process, so protective layers can be produced from high-melting oxides, carbides, and borides. The higher thermal energy and higher velocity produce improved homogeneity and adhesive strength.

The quality of this method is improved by the use of an inert atmosphere under reduced pressure. Individual alloying elements are not burnt off, oxide and nitride formation is suppressed, and adhesion and layer structure is improved.

Typical materials and applications for vacuum plasma spraying are:

1) M – Cr – Al – Y alloys for protection against corrosion by hot gases
2) Carbides (WC, Co – Cr_3C_2, NiCr – TiC – NbC) for protection against abrasive wear and erosion
3) Oxides and refractory metals for the formation of thick protective coatings [44]

The use of chemical vapor deposition (CVD) and physical vapor deposition (PVD) for forming coatings of carbides, borides, nitrides, and oxides has increased. The CVD method has been furthest developed for the deposition of TiC, TiN, CrC, WC, and Al_2O_3 [45–48].

Surface treatment and coating of metals are discussed in more detail in → Corrosion, Chap. 5. and in → Metals, Surface Treatment.

5. Practical Examples of Abrasion and Erosion Damage

Chemical industry with its manifold process steps involves conveying products from vessels or tanks through pipelines to other vessels where new products are produced by chemical reactions, and then to still further apparatuses where the solutions or mixture of products are separated from one another. The process steps also include mechanical processing such as comminution, kneading, and granulating. Production of large quantities requires a continuous material flow. This can be achieved most easily with fluid phases which are transported in conveying units. Erosion and abrasion are therefore the predominant types of wear in the main chemical plant components, while sliding wear, elastic rolling wear, and oscillation wear are found more in drives and machinery.

Transport velocities remain within defined limits and material flows are relatively constant, so characteristic wear data which have been determined under realistic conditions are quite readily transferable. Wear, as a form of damage, does not play a dominant part in chemical plants. Nevertheless, it does occur and the following examples give an impression of the diversity.

5.1. Sliding Wear, Elastic Rolling Wear, and Oscillation Wear

The pulsation in gases which have been compressed in reciprocating compressors must be adequately damped before the gases are transferred to heat exchangers for cooling. This damping was not sufficient in the case of oxygen compression described here, so that the baffles were continuously striking against the supporting spacer tubes; they deformed them increasingly and caused sliding wear to the tubes seated tightly against them (Fig. 20).

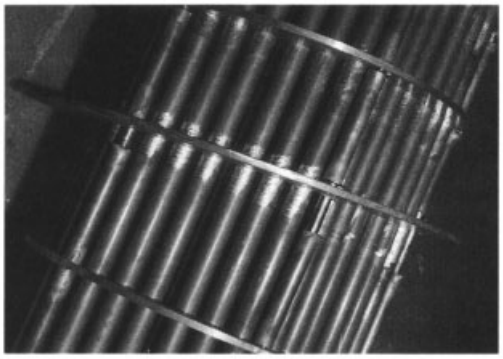

Figure 20. Sliding wear on copper tubes of a heat exchanger for oxygen

Instances of damage caused by sliding occur quite frequently in piston machines. In the rotary compressor shown (Roots type) for ammonia compression the damage occurred after approximately 60 000 h operation (Fig. 21).

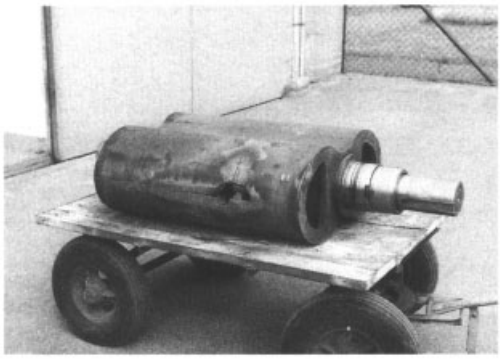

Figure 21. Sliding wear damage on the rotary piston of a Roots compressor for ammonia

The centrifugal pump is the most important hydraulic machine for setting a fluid in motion and for increasing the pressure of the transported medium. Its roller bearings are highly stressed parts with a limited life. This can be shortened considerably if axial shocks are transmitted from the flowing liquid as a result of large numbers of stops and starts, i.e., predominantly discontinuous operation. This is often indicated by running tracks displaced towards the edges (Fig. 22).

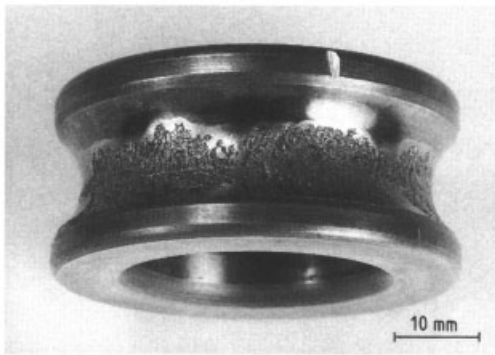

Figure 22. Surface destruction of a ball-bearing inner race as a result of intermittent axial overloading

Predominantly transient stressing (radial impacts) occurred at the piston of an oxygen compressor where the surface of a stuffing box chamber ring showed tribochemical reaction zones as a result of heating in contact with oxygen (Fig. 23). Because of the danger of spontaneous ignition of metals in the presence of compressed oxygen any processes involving friction in such machines must be treated with great caution.

Figure 23. Tribochemical reaction and surface destruction on the brass piston of an oxygen compressor

5.2. Abrasion Wear

Where abrasive impact wear occurs in hammer mills, such as those used for comminution of raw coal in power stations, continuous replacement of plain plates of unalloyed steel is preferred to the application of sustantially more expensive materials with only moderately improved service lives.

Three-body abrasive wear occurs relatively frequently in machines used for process engineering as the materials processed often contain additives with abrasive properties. This is particularly true for the large number of barrel extruders used for kneading and milling processes. It is found, for example, in the draining of unvulcanized rubber, in which the moisture content is reduced from 60 to 10 % under a pressure of 6 – 8 MPa. Screw elements made of 13 % Cr steel withstood these conditions for several years without any signs of wear. However, when the unvulcanized rubber contains carbon black the service life of screws made of X 40 Cr 13 is only about one year, even after hard-facing with Hastelloy C (16 % Cr, 16 % Mo, remainder Ni), Figure 24. This carbon black can be processed, for example, in a wet comminution mill: 1.5 t/h carbon black is ground with 30 t/h water at 3000 rpm between a stator and a rotor adjusted to a gap of approximately 0.4 mm. Both parts have impact ribs with special profiling and various flow diverters. In this case the most economical solution was again to design the elements as wearing parts made of 18 % Cr steel and replace them after 7000 – 10 000 h operation (Fig. 25).

Figure 24. Three-body abrasive wear, caused by carbon black, on a draining screw, hard-faced with Hastelloy C, used for unvulcanized rubber

Figure 25. Impact ribs (18 % Cr steel) of a wet comminution mill eroded by water containing carbon black

The screw housings which enclose the screw shafts are exposed to high stresses, similar to those in the screw combs of the extruders. The product, which is processed under high pressures, stresses the surface to an exceptional extent, especially in the so-called kneading zones of worm extruders. If glass fibers or glass balls are added to the product (in this case, polypropylene), even wear-protecting layers (60 % Ni, 18 % Co, 13 % Cr, 2 % Mo, 1 % C) applied by centrifugal casting have only limited service life (Fig. 26). A suitable method of coating with appropriate protective material should be selected to suit the type of stressing.

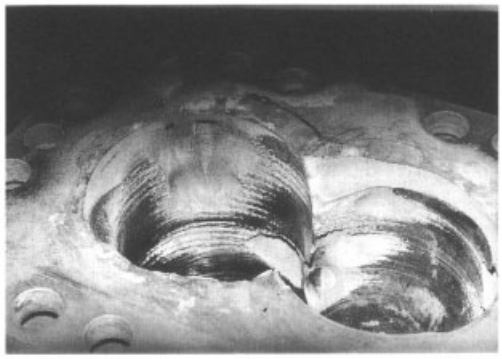

Figure 26. Abrasive wear caused by glass-fiber-containing polypropylene on the high-alloy protective coating (60 % Ni, 18 % Co, 2 % Mo, 1 % C) of a screw housing

Permanent magnet pumps with no seals are being used increasingly for the sake of imission protection. The medium to be transported, in this case liquefied petroleum gas, flows through a gap of less than 1 mm around the non-contact internal rotor, which is fitted with permanent

magnets. Magnetite particles (Fe_3O_4) originating from the pipeline adhere to the outer surface of the internal rotor and cause permanent sliding abrasion at the isolation shell and the rotor, (Fig. 27); finally, this causes the entire pump system to leak.

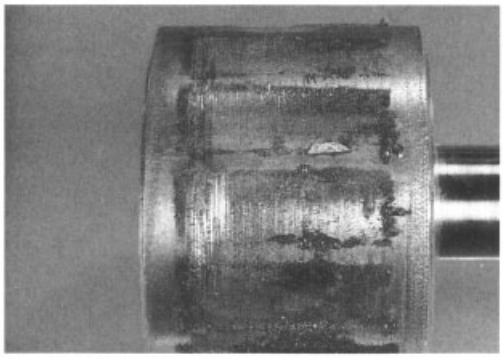

Figure 27. Abrasive wear on the inner rotor of a canned pump caused by adhering magnetite particles

5.3. Damage by Particle Erosion

Abrasive wear is found predominantly in conveying equipment, but in pipelines and apparatus erosive wear is more typical, and is caused by tribological stressing during the flow of fluids, usually with the involvement of solid particles. Unlike the welded joints in apparatus, the butt welds in pipelines can normally only be made from the outside. Depending on geometrical factors (e.g., edge misalignment, difference in wall thickness) and the welding method, weld seam roots are often produced which have a very considerable sag. This can cause a great reduction in the flow cross section and also, depending on the flow conditions, marked turbulences behind the roots. Figure 28 shows a section of a pipe made of St 35.8 (unalloyed steel) in which the carbon black content in a liquid mixture of organic compounds caused a breach in the wall after 15 months' operating time as a result of hydroabrasion. The wall retained its original thickness only a few centimeters behind the point of turbulence, which illustrates the effect of the angle of incidence of the particle on the surface.

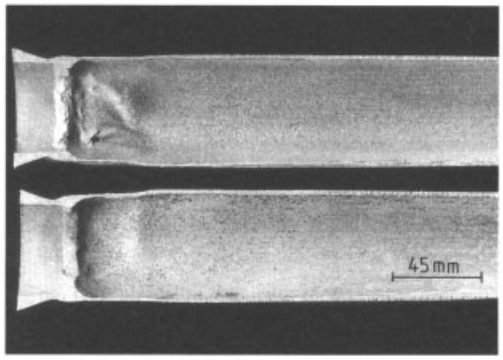

Figure 28. Hydroabrasive wear caused by a liquid containing carbon black in an unalloyed steel pipe as a result of turbulences behind a weld

High mechanical stressing occurs when dispersions of plastics are atomized at high pressure. Under some circumstances the acceleration in a cone leading to the jet hole (a few tenths of a millimeter in diameter) and passage through the hole can change the geometry of a jet in a very short time by erosion. This can affect hard metals and oxide ceramics as well as steel. Due to the forces of acceleration produced in centrifugal pumps, they are particularly liable to hydroabrasive wear. Corrosion-resistant pumps made of cast CrNi (1.4408) had to be replaced after only 6 months' operating time; this was due to the erosion damage shown in Figure 29 because the sharp-edged CuBi catalyst suspended in the aqueous solution had almost eroded through the pump cover.

Figure 29. Pump cover (18-8 CrNi cast) eroded by sharp-edged catalyst particles in the liquid

Even a ducted-wheel pump made of G-X3 CrNiMoCu 24 6 (Noridur 9.4460) hardly lasted any longer, but the tempered and quenched cast

alloy G-X170 CrMo 25 2 (Niroloy NL 262) had a service life of several years.

When solid particles are carried in a gas stream and not in a guided liquid stream, any changes in flow direction are very important as, due to the large difference in density, the solid particles are pressed with high force on to the tube surface which guides the flow. The velocities of the transported gas may be very high and then, as a result of intensive sliding jet wear at localized points, lead to short term, but usually accelerated, loss of wall thickness and to rupture of the tube bends.

An example of this is provided by bends from tube furnaces for the production, for example, of ethylene. Because of the high operating temperatures of over 900 °C, the required creep strength and oxidation resistance are usually obtained by using high-alloy materials (e.g., 30 % Cr, 30 % Ni) and considerably increasing the wall thickness at the bends. Figure 30 shows localized penetration of this type. The wear stressing occurs during repeated decoking with superheated steam.

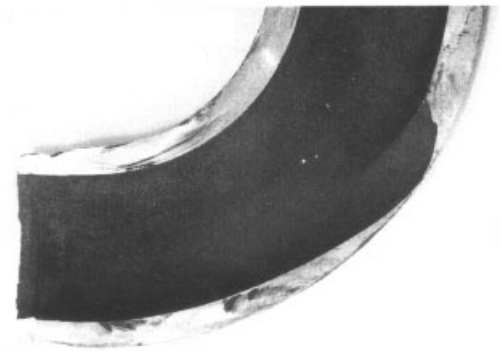

Figure 30. Bend in a tube furnace for ethylene production, worn by sliding jet wear (material: CrNi 30/30 alloy)

A comparable situation occurs in waste heat boilers that utilize the high temperature of synthesis gas (CO, H_2) produced in an oil gasification plant for generating steam. Condensing and solidifying particles of slag often form a dense and tightly adhering coating in the gas-carrying tubes, with the result that over a long period the system provides its own protection against jet wear on the heat-resistant, low-alloy steels used. When there are changes in the mode of operation, this protective action often fails locally – the gas velocity apparently plays an important role here – and the protective layer and the tube elbow are worn away in a very restricted area, as if with a milling cutter (Fig. 31).

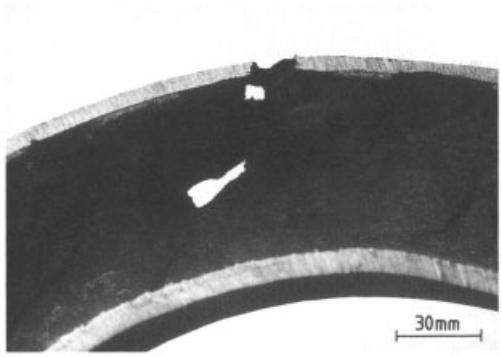

Figure 31. Tube elbow of a waste heat boiler whose protective coating has been worn away in a restricted area by inclined jet wear, which ultimately penetrated the tube

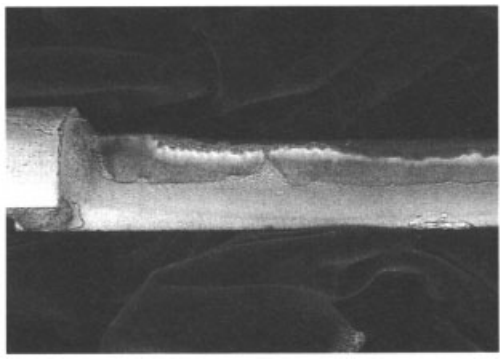

Figure 32. Plasma-coated protective shell of a steam superheater tube, damaged by impact jet wear at ca. 1000 °C

Superimposition of flow and transient stressing by particles in the gas flow, which is characteristic of impact wear and inclined jet wear, is not found to any great extent in chemical plants but does occur in power stations. It affects the first contact heating surfaces where the ash particles, at temperatures up to 1000 °C rising at ca. 10 m/s with the flue gas, lead to impact jet wear. These tubes, which are under high steam pressure, are therefore usually protected with heat-resistant wear-protecting half shells. In the case described, comparative tests have shown that high-alloy austenitic alloys (with 25 % Cr, 20 % Ni, and Si) last longer than high-chromium ferritic steels that have been plasma coated with an oxide layer (Fig. 32). It is assumed that metal-

lurgical changes and chemical reactions caused by the high temperature combine with the purely mechanical stressing to form the complex outward appearance.

However, if the wear-induced break occurs in steam tubes, the expanding steam escaping at high velocity causes erosion on adjacent tubes. Although in this case there are again solid particles entrained in the flue gas, microscopic observation of the affected surface shows that, unlike the situation with particle jet wear, there is an ripple structure, typical of steam jet action (Fig. 33).

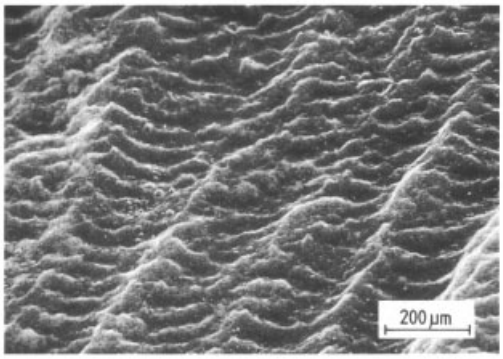

Figure 33. Scanning electron micrograph of the surface of a steam tube (material: 12 % Cr steel) damaged by a steam jet

5.4. Solid Particles-Free Erosion

Heat transfer systems are one of the basic requirements of thermal process technology. Depending on the purpose of the heat exchange, the terms evaporator, liquefier or condenser, superheater, boiler, and many others, are also used.

If the flows of material taking part in the heat transfer are separated from one another by walls, the process is called indirect heat exchange. In general, the flow conditions and prevailing thermal conditions are sufficiently defined and maintained, so only in critical cases does mechanical damage occur by erosion alone.

Liquid erosion damage is to be expected if heating steam contains a proportion of condensate – again dependent on the angle of incidence. When compared with austenitic steels, steam-heated, high-pressure pipes made of low-alloy ferritic steel prove to be insufficiently resistant if the heating is carried out with incompletely dried steam at 2 MPa (Fig. 34).

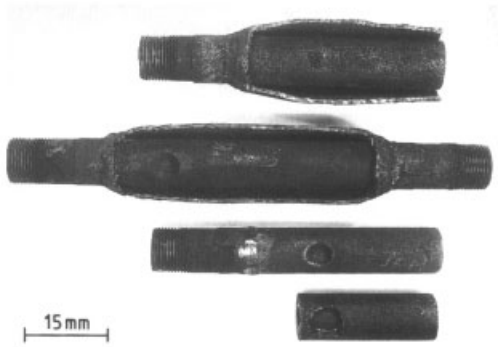

Figure 34. Steam-heated high-pressure pipe showing liquid erosion damage due to condensate carried with the steam

Rotary compressors include liquid ring pumps, in which a rotor eccentric to the housing opens and closes chambers of different sizes with the aid of a liquid ring; they are used as compressors and vacuum pumps. Under some circumstances cavitation of spontaneously vaporizing liquid can lead to damage to the rotor at the narrowest gap between the rotor and the housing, as in the example shown in Figure 35, where the rotor is made of nodular cast iron GGG 42. The different resistance to cavitational stressing can be seen clearly on the surface, where the harder iron phosphide eutectic is largely retained, while the iron matrix has been worn away (Fig. 36). There is also the possibility that a certain corrosive component has contributed to the damage, so this may be a case of cavitation corrosion.

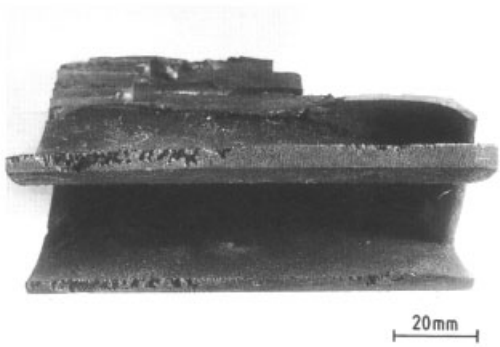

Figure 35. Cavitation damage on a cast iron pump rotor

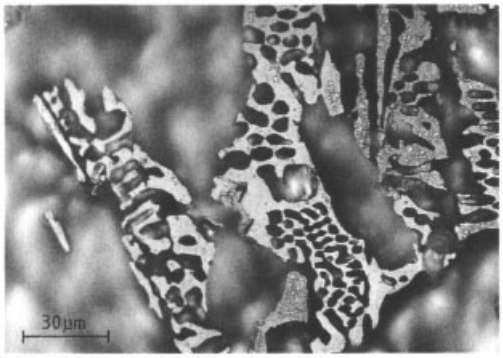

Figure 36. Micrograph of the rotor shown in Figure 35; although the iron matrix has been removed by cavitation, the more resistant iron phosphide eutectic remained unaffected

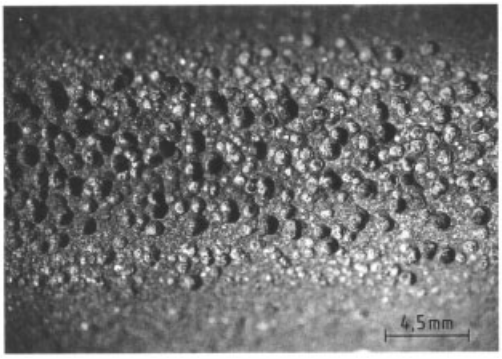

Figure 37. Cavitation erosion in a steel (St. 35.8) heat exchanger tube due to excessive heating

Comparable conditions also caused the damage shown in the following examples. Damage occurred only in the region of the gas inlet of a shell-and-tube heat exchanger in which the tubes carry water in closed circuit for cooling. This was because local vaporization of the water followed immediately by collapse of the vapor bubbles led to deep holes, while immediately after the gas inlet point the existing protective layer remained completely undamaged on the unalloyed steel. The weakening visible in the micrograph resembles that of pitting corrosion, and clearly the already broken protective layer remains electrochemically active so that a corrosive removal mechanism is superimposed. The hammering effect of the micro jet becomes obvious from Figure 37. In steam turbines even a chromium steel X 20 Cr 13 with otherwise good resistance properties is permanently damaged by impingement of droplets at a water content of approximately 8 % at room temperature (Fig. 38).

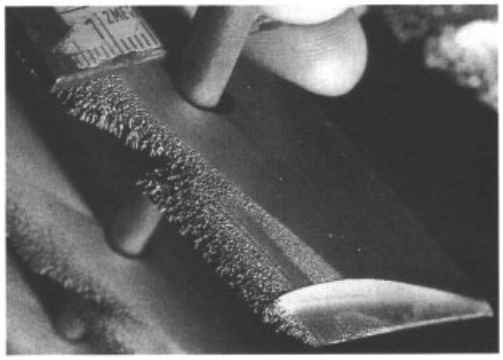

Figure 38. Impingement attack on a steam turbine blade (X 20 Cr 13)

A type of drop impingement can also lead to purely mechanical damage and deformation on such parts where a liquid-filled gap, in this case between parts of a canned motor pump made of steel 1.4408, is submitted to transient stressing in the axial direction with simultaneous oscillating radial movement without contact between the metal surfaces (Fig. 39).

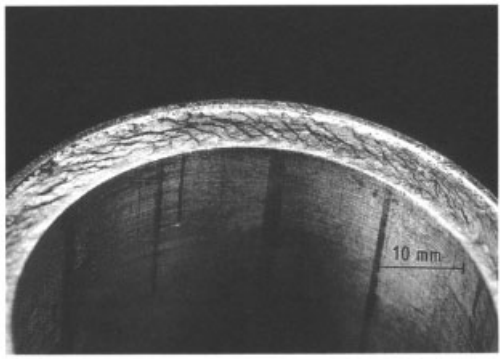

Figure 39. Surface destruction of the case (18-8 CrNi steel) of a canned motor pump

The reason for this unusual type of stressing was incorrect assembly in which the jointed parts had not been correctly secured.

5.5. Damage Caused by Erosion – Corrosion

The chemical resistance of many metals to electrolytes is achieved because the reaction products adhere to the surface, hindering ionic trans-

port that controls the electrochemical dissolution process. Guide values can therefore be provided, for example, for the maximum velocity of drinking water up to which certain copper alloys, steels, and titanium retain their protective coatings, and only when the measured values are exceeded is removal of the protective layer to be expected. In many cases unalloyed steel and cooling water operating under defined flow conditions represent the most economical solution for heat exchangers in the chemical industry. Even then, the alternating actions of erosion and corrosion can cause significant damage due to the design and the associated changes in flow direction, although with smooth flow it is possible to maintain a protective covering layer for many years.

Cases of this type of damage can be clearly demonstrated in heat exchangers from a CO conversion process in which the gas preheats the recirculating water (nominal pressure 6.4 MPa) to a maximum of 220 °C: excessive water velocities through perforated impact plates, at changes in direction, and at tube openings through baffles prevented the formation of a complete protective layer; this led to progressive wear through liquid erosion accelerated by electrolytic dissolution of the metal (Fig. 40). It was possible to eliminate this type of erosion on the internal surfaces of vessels completely and permanently when they were spray-coated with stainless Cr – Ni steel.

Figure 40. Due to the high flow velocity of the cooling water, the protective layer was removed from heat exchanger tubes (St. 35.8) and sheets (St. 37), leading to intensive erosion – corrosion

Loss of material by preferential erosion – corrosion caused by the flow conditions can be seen at a product outlet connection of a wash column made of 15 Mo 3 (DN 300). The high velocity together with the slightly corrosive medium is entirely responsible for the loss in wall thickness (Fig. 41).

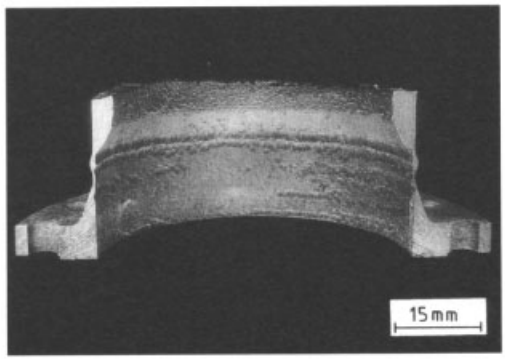

Figure 41. Erosion – corrosion due to turbulence of the liquid caused by the reduction in diameter of a pipe (nominal diameter 300 mm, 15 Mo 5)

Occasionally the free fall of a corrosive liquid is also sufficient to remove protective layers and promote the corrosion process, as can be seen in a melting apparatus for dodecanedioic acid (Fig. 42).

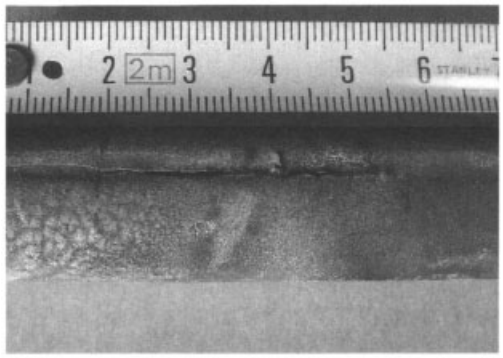

Figure 42. Erosion – corrosion caused by dripping molten acid (material: 18-8 CrNi steel)

Crystalline acid with a moisture content of 10 – 20 % is fed through a filling hopper and melted by heating tubes made of 18-8 CrNi steel, and the vapor produced is driven off. This damage, a combination of erosion and corrosion, occurred at the region of highest temperature. The complex stress profile of the extruder screw may also be demonstrated with the example of a kneader for unvulcanized rubber, in which, as a result of a localized escape of fumaric acid

and very small quantities of water, serious damage occurred within one day due to superimposed erosion and corrosion of a kneading element made of quenched and tempered ionitrided chromium steel X 35 CrMo 17 (Fig. 43).

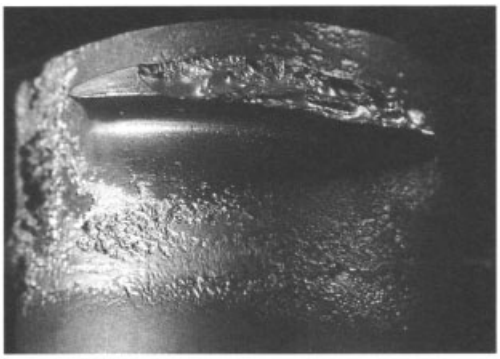

Figure 43. Superposition of erosion and corrosion in a kneading element of a rubber extruder (material: X 35 CrMo 17, ionitrided)

Conclusion. In chemical plants the problems of wear play a significantly smaller role than those of errosion. Damage in the component equipment and pipelines is met relatively rarely, although abrasion and erosion can occasionally be very detrimental to the potential availability of the plant components at risk. However, it is a different question with machines and conveying units, in which exceptionally high wear stresses can sometimes result from the influence of high flow velocities and high accelerations at housing walls, and especially at sliding seal elements. The manufacturing industries make use of the most modern materials technology to control these problems. Nevertheless, design measures also have to be chosen so that components which are exposed to particular stresses and have limited service lives are designed as easily replaceable wearing parts.

6. References

1. BMFT-Report: *Damit Rost und Verschleiß nicht Milliarden fressen,* BMFT, Bonn 1984.
2. B. Genath: "Der Verschleiß ist das Sorgenkind im Maschinenbau," *VDI-Nachr.* **52** (1971) no. 47, 1, 6, 7.
3. DIN 50 320:Dec. 1979.
4. K. H. Zum Gahr: "Grundlagen des Verschleißes," in: *Metallische und nichtmetallische Werkstoffe und ihre Verarbeitungsverfahren im Vergleich,* part III, *VDI-Ber.* **600.3** (1987) 29 – 56.
5. H. Uetz, K. Sommer: "Abrasiv-Gleitverschleiß," in: H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 108 – 157.
6. K. H. Zum Gahr: "Einfluß des Makroaufbaus von Stahl/Polymer-Faserverbundwerkstoffen auf den Abrasivverschleiß," *Z. Werkstofftech.* **16** (1985) 297 – 305.
7. H. Wahl: "Verschleißprobleme im Braunkohlenbergbau," *Braunkohle Wärme Energie* **3** (1951) 75 – 87.
8. R. Schulmeister: "Zur Untersuchung der Werkstoffzerstörung durch Kavitation und Korrosion mit Ultraschall-Koppelschwingern," *Metalloberfläche* **21** (1967) no. 1, 17 – 25.
9. H. Rieger: *Kavitation und Tropfenschlag,* Werkstofftechnische Verlagsgesellschaft, Karlsruhe 1977.
10. H. Brauer, E. Kriegel: "Probleme des Verschleißes von Rohrleitungen beim pneumatischen und hydraulischen Feststofftransport," *Maschinenmarkt* **71** (1965) no. 68, 140 – 151.
11. J. M. Hutchings: "Some Comments on the Theoretical Treatment of Erosive Particle Impacts," *Proc. 5th Int. Conf. on Erosion by Solid and Liquid Impact,* Cambridge 1979.
12. G. Gommel: *Stoßuntersuchungen Stahlkugel/Stahlplatte im Zusammenhang mit Strahlmittelzertrümmerung und Strahlverschleiß,* Dissertation, TH Stuttgart 1966.
13. J. S. Rinehard, J. Pearson: *Behavior of Metals Under Impulsive Load,* American Society of Metals, Cleveland 1954.
14. K. Wellinger, H. Uetz: "Strahlverschleiß," *Tech. Rundsch.,* 8 (1958) 1 – 8.
15. J. Föhl: "Strahl- und Spülverschleiß," in: K. H. Zum Gahr (ed.): *Reibung und Verschleiß,* Deutsche Gesellschaft für Metallkunde, Oberursel 1983, pp. 157 – 176.
16. H.-G. Heitmann, W. Kastner: "Erosionskorrosion in Wasser-Dampf-Kreisläufen – Ursachen und Gegenmaßnahmen, *VGB Kraftwerkstech.* **62** (1982) 211 –219.
17. H. Tischner: "Korrosionserscheinungen in strömenden Medien am Beispiel von Chemiepumpen," *Chem. Ing. Tech.* **62** (1989) no. 3, 220 – 228.

18. C. Razim, C. Düll, W. Räuchle: "Über die Beeinflussung der Bauteil-Grundeigenschaften durch Verschleißschutzschichten," *VDI-Ber.* **333** (1979) 11 – 22.
19. K. Wellinger, H. Uetz: "Gleitverschleiß, Spülverschleiß, Strahlverschleiß unter der Wirkung von körnigen Stoffen," *VDI-Forschungsh.* **449, ed. B**(1955) no. 21.
20. M. Y. Gürleyik: *Gleitverschleißuntersuchungen an Metallen und nichtmetallischen Hartstoffen unter Wirkung körniger Gegenstoffe,* Dissertation, TH Stuttgart 1967.
21. F. Henke: "Niedrig- und hochlegierter verschleiß-fester Vergütungsstahlguß," *Gießerei-Prax.* 1975, no. 23/24, 377 – 407.
22. M. M. Krushchov, M. A. Babichev: "Experimental Fundaments of Abrasive Wear Theory," *Russ. Eng. J.* (Engl. Transl.) (1964) no. 6, pp. 43 – 48.
23. H. Uetz, K. J. Groß: "Strahlverschleiß," in: H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 236 – 278.
24. J. G. A. Bitter: "A Study of Erosion Phenomena," Part I: *Wear* **6** (1963) 5 – 21; Part II: *Wear* **6** (1963) 69 – 190.
25. W. Wahl: "Unterschiedliche Werkstoffbewährung bei abrasiv beanspruchten Bauteilen," *VDI-Ber.* **600.3** (1987) 245 – 286.
26. K. Heil: *Erosionskorrosion an unlegierten Eisenwerkstoffen in schnellströmenden Wässern,* Dissertation, TH Darmstadt 1979.
27. D. Kuron: "Korrosion durch Kühlwasser und Schutzmaßnahmen," in W. J. Bartz (ed.): *Die Praxis des Korrosionsschutzes,* Expert Verlag, Grafenau 1981.
28. H. Sick: "Die Erosionsbeständigkeit von Kupferwerkstoffen gegenüber strömendem Wasser," *Werkst. Korros.* **23** (1972) no. 1, 12 – 18.
29. H. Grein: "Kavitation – eine Übersicht," *Sulzer Forschungsh.* 1974, 87 – 112.
30. H. H. Piltz: *Werkstoffzerstörung durch Kavitation,* VDI-Verlag, Düsseldorf 1966.
31. W. J. Rheingans: "Cavitation in Hydraulic Turbines," Symp. on Erosion and Cavitation, *ASTM Spec. Tech. Publ.* **307** (1962) 17 – 31.
32. K. H. Habig: *Verschleiß und Härte von Werkstoffen,* Carl Hanser Verlag, München – Wien 1980, p. 215.
33. H. Uetz, J. Wiedemeyer: *Tribologie der Polymere,* Carl Hanser Verlag, München – Wien 1985.
34. J. Wiedemeyer: "Deutung des tribologischen Verhaltens ungeschmierter Thermoplaste auf der Basis von Modellrechnungen sowie experimentellen Ergebnissen," *Fortschr. Ber. VDI Z. Reihe 5* **96** (1985).
35. J. K. Lancaster: "Abrasive Wear of Polymers," *Wear* **14** (1969) 223 – 239.
36. K. Wellinger, H. Uetz, G. Gommel: "Verschleiß durch Wirkung von körnigen mineralischen Stoffen," *Materialprüfung* **9** (1967) no. 5, 153 – 160.
37. H.-D. Ruprecht: "Elastomere-Polyurethane," in: H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 438 – 450.
38. H. P. Lachmann: "Elastomere-Gummi," in H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 451 – 465.
39. H. Brauer, E. Kriegel: "Untersuchungen über den Verschleiß von Kunststoffen und Metallen," *Chem. Ing. Techn.* **35** (1963) 697 – 707.
40. K. Wellinger, H. Uetz: "Verschleiß durch körnige mineralische Stoffe," *Aufbereit. Tech.* **4** (1963) 193 – 204, 319 – 335.
41. E. Dörre: "Nichtmetallische Hartstoffe," in H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 451 – 465.
42. W. Wahl, I. Kretschmer, J. Wabnegger: "Auftragschweißen," in H. Uetz (ed.): *Abrasion und Erosion,* Carl Hanser Verlag, München – Wien 1986, pp. 374 –394.
43. H. Gräfen: "Beschichtungen in der Chemietechnik," in: Beschichtungen für Hochleistungs-Bauteile, *VDI-Ber.* **624** (1986) 273 – 296.
44. H.-M. Höhle: "Thermische Spritzverfahren," in: Beschichtungen für Hochleistungs-Bauteile, *VDI-Ber.* **624** (1986) 71 – 83.
45. F. Wendl: "Aktuelle Trends bei der Oberflächenbehandlung von Werkzeugen in der Kunststoffverarbeitung," *Thyssen Edelstahl Tech. Ber.* **15** (1989) no. 2, 110 – 125.
46. H.-A. Mathesius: "Herstellen von verschleißfesten Schichten mit Hilfe von CVD-Verfahren," in: Beschichtungen für Hochleistungs-Bauteile, *VDI-Ber.* **624** (1986) 37 – 48.
47. H. Weiß: "Elektrochemische Beschichtung und Sonderverfahren der Oberflächentechnik," *Thyssen Edelstahl Tech. Ber.* **15** (1989) no. 2, 85 – 114.
48. H. Simon, M. Thoma: *Angewandte Oberflächentechnik für metallische Werkstoffe,* Carl Hanser Verlag, München – Wien 1985.

Mechanical Properties and Testing of Metallic Materials

ALI FATEMI, The University of Toledo, Toledo, Ohio 43606, United States

1. Introduction 2
1.1. Concepts of Stress and Strain 2
1.1.1. Stress 2
1.1.2. Strain 2
1.2. Elastic Deformation and Properties 3
1.2.1. Young's Modulus 3
1.2.2. Shear Modulus 3
1.2.3. Poisson's Ratio 3
1.2.4. Relation Between Elastic Constants . 3
1.3. Plasticity and Flow 4
2. Mechanical Testing 4
2.1. Tensile Test 4
2.1.1. Significance and Use 4
2.1.2. Apparatus and Specimen 4
2.1.3. Stress – Strain Diagram 5
2.1.4. Strength Properties 5
2.1.5. Ductility 6
2.1.6. Toughness and Resilience 6
2.1.7. True Stress – True Strain Diagram . . 7
2.1.8. Strain-Rate Sensitivity 8
2.1.9. Temperature Effects 8
2.2. Compression Test 8
2.3. Bend and Flexure 9
2.3.1. Elastic Bending Stress and Deflection 9
2.3.2. Bend and Flexure Test 9
2.4. Shear and Torsion 10
2.4.1. Direct Shear Stress and Test 10
2.4.2. Torsional Shear Stress and Test 10
2.5. Hardness Tests 11
2.5.1. Rockwell Hardness Test 11
2.5.2. Brinell Hardness Test 12
2.5.3. Vickers Hardness Test 13
2.5.4. Knoop Hardness Test 13
2.5.5. Correlation of Hardness with Strength 13
2.6. Impact Testing 13
2.6.1. Notched Bar Impact Test 13
2.6.2. Dynamic Tear Test 14
2.7. Creep Test 15
2.8. Stress Relaxation Tests 16
3. Fatigue Loading 17
3.1. Principles and Terminology 17
3.1.1. Fatigue Phenomena 17
3.1.2. Fatigue Fracture Surfaces 17
3.1.3. Fatigue Behavior 17
3.2. Cyclic Stress-Controlled Fatigue . . 18
3.2.1. Stress – Life Curves 18
3.2.2. Fatigue Tests 18
3.2.3. Mean Stress Effects 19
4. Fracture Mechanics 20
4.1. Principles and Terminology 20
4.2. Fracture Toughness 20
5. References 23

Symbols:

a	crack length
A	cross-sectional area
A_0	initial cross-sectional area
A_f	final cross-sectional area
C	distance to neutral surface in beam bending
d	diameter
D	indenter diameter
E	elastic (Young's) modulus
f	function
F	force or load
F_{max}	maximum load
G	shear modulus or modulus of rigidity
h_0	initial height
h_f	final height
I	moment of inertia
K	stress intensity factor, strength coefficient
K'	strain-rate constant
K_c	fracture toughness
K_{Ic}	plane-strain fracture toughness
l	length
l_0	initial gauge length
l_f	final gauge length
Δl	change in length
m	strain-rate sensitivity
M	bending moment
n	strain-hardening exponent
N	number of cycles
r	radius
R	stress ratio
S	shear strength
t	thickness, indentation depth, time
Δt	time increment
T	torque
v	displacement

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_10

V	shear force
w	width
Y	dimensionless parameter
γ	shear strain
γ_{max}	maximum shear strain
δ_{max}	maximum deflection
ε	normal (axial) strain
ε_e	elastic normal strain
ε_{eng}	engineering strain
ε_p	plastic normal strain
ε_{true}	true strain
$\dot{\varepsilon}$	creep rate
$\dot{\varepsilon}_{true}$	true strain rate
$\Delta\varepsilon$	strain increment
μ	Poisson's ratio
σ	normal (axial) stress
σ_a	alternating stress or stress amplitude
σ_{eng}	engineering stress
σ_f	fracture stress
σ_{fat}	fatigue strength for $\sigma_m = 0$
σ_m	mean stress
σ_{max}	maximum stress
σ_{min}	minimum stress
σ_{true}	true stress
σ_u	ultimate tensile strength
σ_{ys}	yield strength
$\Delta\sigma$	stress range
τ	shear stress
τ_{ave}	average shear stress
τ_{max}	maximum shear stress
Φ	displacement angle or angle of twist

1. Introduction

1.1. Concepts of Stress and Strain

1.1.1. Stress

Stress σ is defined as the force divided by the cross-sectional area on which the force acts

$$\sigma = \frac{F}{A} \tag{1}$$

where F is the force and A is the area. It represents the intensity of the reaction force at any point in the body as imposed by service loads, assembly conditions, fabrication, and thermal changes.

If the imposed load is perpendicular to the area over which it acts, the resulting stress is a normal (or axial) stress σ. There are two general types of normal stress: tensile and compressive. A pulling force produces tensile normal stress, whereas a pushing force produces compressive normal stress. When the load acts parallel to the area, the resulting stress is referred to as shear stress τ.

In SI units, stress is measured in newtons per square meter or pascals.

1.1.2. Strain

When stresses are applied to a body, they cause deformation of the body and result in changes in dimensions or distortion; the body is then said to be strained. Strains resulting from changes in dimension are normal strains ε, whereas strains resulting from distortion of the body are shear strains γ. A normal strain is, by definition, the axial deformation (elongation or contraction) per unit length

$$\varepsilon = \frac{\Delta l}{l} \tag{2}$$

where Δl is the change in length and l is the length. A normal strain resulting from elongation is called a tensile strain, and a normal strain resulting from contraction is called a compressive strain.

A change in the shape of a body is caused by the application of shearing stresses. Shearing stresses cause relative displacement of the upper and lower surfaces of the body, as shown in Figure 1. The angular change between the two perpendicular lines is defined as the shear strain γ. Strain is a nondimensional quantity, but it is frequently expressed in meters per meter or in percent.

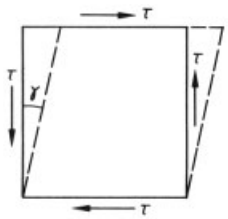

Figure 1. Distortion of a body as a result of the application of shear stress.
—- undeformed; – – – deformed

1.2. Elastic Deformation and Properties

Elastic deformation is a temporary deformation that is fully recoverable when the load is removed. The amount of elastic deformation that a metal can undergo is small because metal atoms are displaced from their original positions during elastic deformation, but not to the extent that they take up new positions.

1.2.1. Young's Modulus

If a long steel wire is pulled by an increasing load, the length of the wire increases proportionately with the load. The relationship between stress resulting from the load ($\sigma = F/A$) and strain resulting from elongation of the wire ($\varepsilon = \Delta l/l$) is given by Hooke's law:

$$\sigma = E\varepsilon \tag{3}$$

where the proportionality constant E is a material property, called the elastic modulus or Young's modulus. This value represents the stiffness of the material (its resistance to elastic strain), and for most typical metals its magnitude is in the range of 45 – 410 GPa. A stress – strain diagram representative of elastic deformation for loading and unloading is shown in Figure 2.

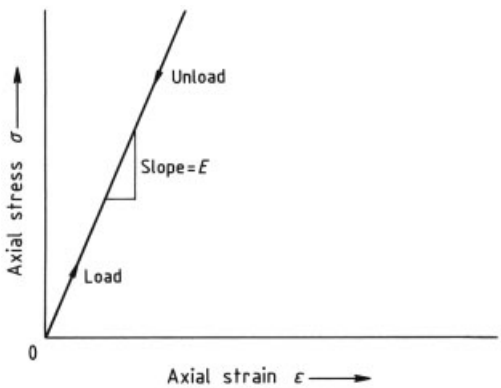

Figure 2. Elastic axial stress – strain diagram

As expected, imposition of a compressive stress also evokes elastic behavior. However, stress – strain characteristics and Young's modulus are virtually the same for both tensile and compressive stresses in the elastic range.

1.2.2. Shear Modulus

Shear stress and strain are related through:

$$\tau = G\gamma \tag{4}$$

where G is a material property called the shear modulus, sometimes referred to as the modulus of rigidity. The shear modulus represents the slope of the linear elastic region of the shear stress – strain diagram, as shown in Figure 3. For many metals, G is ca. $0.4\,E$.

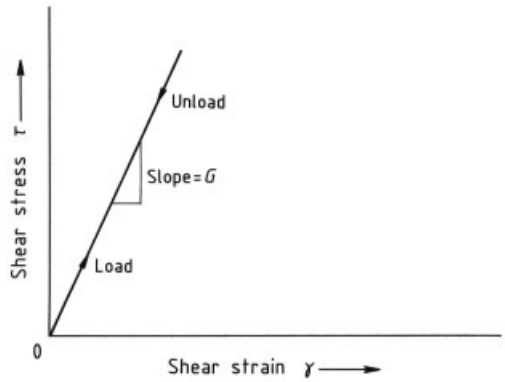

Figure 3. Elastic shear stress – strain diagram

1.2.3. Poisson's Ratio

When a uniaxial tensile force is applied to a rod or bar, it results in extension (or contraction in the case of a compressive force). Existing simultaneously with this elongation, however, is a lateral or transverse contraction (or extension for a compressive force) in the directions normal to the applied stress. The ratio of the magnitude of the transverse change per unit length (transverse strain) to the magnitude of the longitudinal change in length per unit length (longitudinal strain) is called Poisson's ratio μ:

$$\mu = \frac{|\text{Lateral strain}|}{|\text{Longitudinal strain}|} \tag{5}$$

Poisson's ratio is a material elastic property and for typical metals is on the order of 0.3.

1.2.4. Relation Between Elastic Constants

The three elastic constants (E, G, and μ) in linear elasticity are related to each other by the equation:

$$G = \frac{E}{2\,(1+\mu)} \tag{6}$$

Therefore, if two of the elastic constants are given or measured, the third can be calculated. Typical values of elastic constants for metallic materials are given in Table 1.

Table 1. Elastic constants for some common metallic materials

Material	Young's modulus E, 10^3 MPa	Shear modulus G, 10^3 MPa	Poisson's ratio μ
Aluminum	71	26	0.34
Brass	110	41	0.34
Copper	128	48	0.34
Iron	207	80	0.29
Magnesium	45	17	0.29
Nickel	200	77	0.30
Silver	83	30	0.37
Titanium	113	43	0.32

1.3. Plasticity and Flow

Many materials undergo permanent or nonrecoverable deformation when stressed beyond a certain level. This is called plastic deformation; it remains in the material after the load or stress has been removed. Plastic deformation results from permanent displacement of atoms or molecules from their original positions in the lattice. This displacement is due to the breaking of bonds between atoms or molecules and the net movement of large numbers of them relative to one another.

A constant load of sufficient magnitude applied to a material results in continuously increasing deformation; this phenomenon is called flow.

Figure 4 illustrates the idea of plastic deformation. If the material is stressed to σ_1 and then released, the strain returns to zero (elastic behavior). However, when the material is stressed to level σ_2 and then released, the material recovers to the extent $(\varepsilon - \varepsilon_\mathrm{p})$, where ε_p is the plastic strain remaining after the load or stress has been removed. Therefore, for stress levels beyond the elastic limit, the total strain is composed of elastic, ε_e, and plastic, ε_p, portions:

$$\varepsilon = \varepsilon_\mathrm{e} + \varepsilon_\mathrm{p} \tag{7}$$

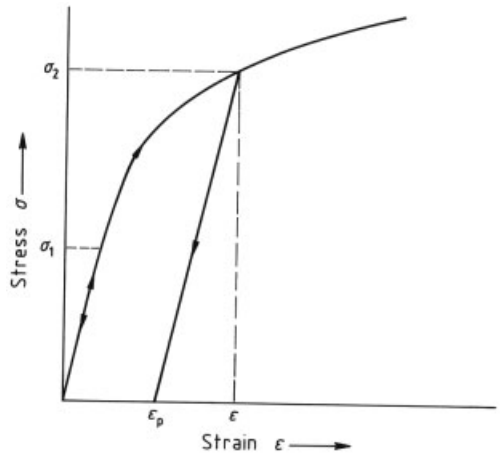

Figure 4. Plastic deformation preceded by elastic deformation

2. Mechanical Testing

The mechanical properties of materials determine their characteristics and behavior under applied service forces or loads. Specifically, the mechanical behavior of a material characterizes its response (or deformation) when subjected to an applied load. Important mechanical properties include stiffness, strength, ductility, and hardness. In this chapter, several tests that are used to determine these properties are discussed.

2.1. Tensile Test

2.1.1. Significance and Use

The tensile test (or pull test) provides information on the strength and ductility of materials under a uniaxial tensile stress condition. The results of this test are often used to evaluate the mechanical properties of materials and to compare materials. Details of tensile testing are given in ASTM E8M-89 [1].

2.1.2. Apparatus and Specimen

A typical setup for a tensile test is shown in Figure 5. A typical specimen has a diameter of 12.5 mm and a gauge length of 62.5 mm as

shown in Figure 6 [1]. The ends of the specimen are screwed into a pair of grips, which are part of the tensile testing machine (Fig. 5). The grips are then pulled apart, often by mechanical means. The load on the specimen is recorded continuously by means of a load cell, which is often a calibrated stiff spring. Extension of the gauge section is measured by an extensometer, which is anchored to the specimen gauge section at the start of the test. Raw data are recorded in terms of load and extension.

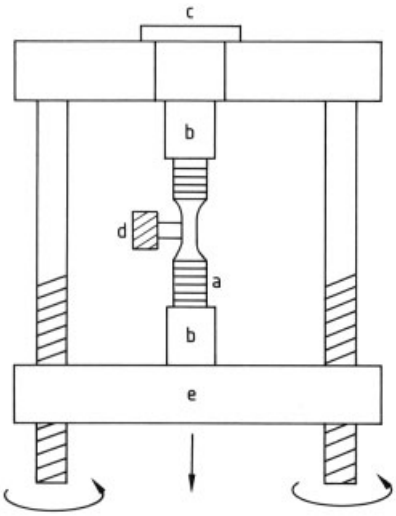

Figure 5. Schematic of a tension test setup
a) Specimen; b) Grips; c) Load cell; d) Extensometer; e) Moving crosshead

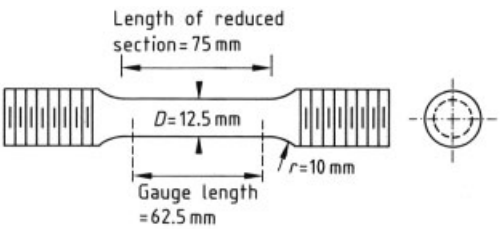

Figure 6. Typical round tension test specimen [1]

2.1.3. Stress – Strain Diagram

Load versus gauge extension data are converted to engineering stress σ_{eng} and engineering strain ε_{eng} as follows:

$$\sigma_{\text{eng}} = \frac{\text{Load}}{\text{Initial area}} = \frac{F}{A_0} \tag{8}$$

$$\varepsilon_{\text{eng}} = \frac{\text{Change in length}}{\text{Initial length}} = \frac{\Delta l}{l_0} \tag{9}$$

A typical engineering stress – strain diagram for a mild steel is shown in Figure 7, which is divided into two distinct regions of elastic deformation and plastic deformation.

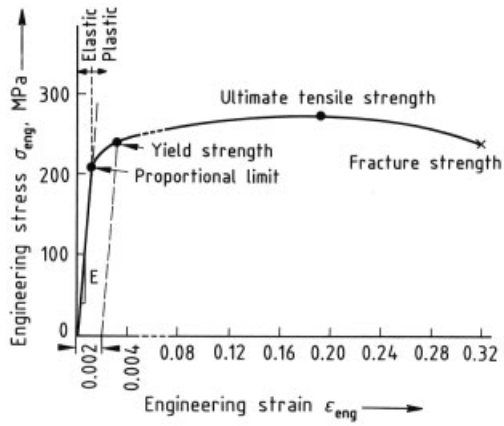

Figure 7. Engineering stress – strain diagram for a mild steel

Stretching of the metal in the elastic portion of the stress – strain curve (the initial linear portion) is fully recoverable when the load is removed. At higher forces the material behaves in a plastic manner. Plastic deformation is permanent; upon removal of the load, only the elastic deformation (which is often very small compared to the plastic deformation) is recovered. The plastic region is the nonlinear portion of the stress – strain diagram.

2.1.4. Strength Properties

The *proportional limit* is the limit within which Hooke's law (Eq. 3) applies; it is shown in Figure 7. Young's modulus E is determined from the slope of this linear portion of the curve and measures the stiffness of a material (see Section 1.2.1).

The stress level at which plastic deformation begins is referred to as the *yield point*. At this point the material undergoes rather rapid and extensive plastic deformation with little accompanying increase in load. Since the transition from elastic to plastic behavior is usually gradual, accurate assessment of the yield point is often difficult. Therefore, the *yield strength* is generally

found by an offset method. A line parallel to the initial linear elastic line is drawn from an offset point (typically at a strain of 0.002). The intersection of this line with the stress – strain curve determines the yield strength for the material, as shown in Figure 7. The yield strength of a metal is a measure of its resistance to plastic deformation and is probably the most important mechanical property used in design.

The *ultimate tensile strength* is the maximum strength reached in the engineering stress – strain curve, as shown in Figure 7. Even though the ultimate tensile strength is relatively unimportant for selection or fabrication of materials, its value is often reported in handbooks because it is easy to measure and some of the other material mechanical properties that are more difficult to measure, can be estimated from it. All deformation up to this point in the stress – strain curve is uniform throughout the specimen gauge section. However, at this maximum stress, deformation does not remain uniform, and one region starts deforming more than others (a large local decrease in cross section). This phenomenon is called *necking* and is prominent in ductile materials. Figure 8 shows a schematic representation of this local deformation.

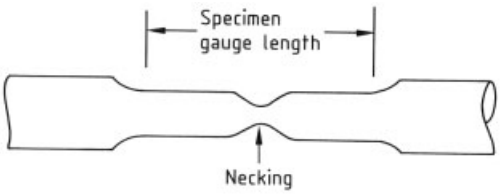

Figure 8. Necking during tension testing

As soon as a localized neck begins to form, a drop in the force occurs due to the decrease in the cross-sectional area at this point. With the onset of necking, the stress distribution changes from uniaxial to triaxial.

After the neck has developed, further plastic deformation is constrained to its vicinity and fracture ultimately occurs at this point. The *fracture* or *rupture strength* corresponds to the stress at fracture, as shown in Figure 7.

2.1.5. Ductility

Ductility is another important mechanical property; it measures the amount of deformation that a material has sustained at fracture. A material that undergoes very little deformation before rupture is said to be *brittle*, whereas one that undergoes a large amount of plastic deformation before rupture is said to be *ductile*. Engineering stress – strain curves for a brittle and a ductile metal are shown in Figure 9.

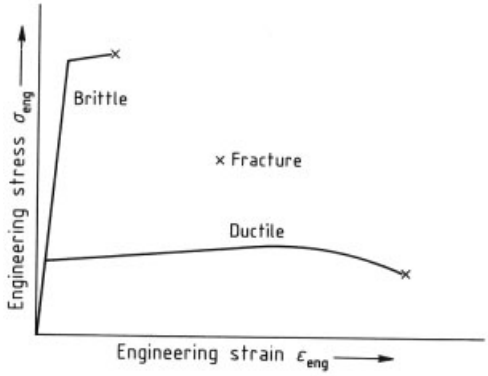

Figure 9. Engineering stress – strain curves for a ductile and a brittle material

Ductility is expressed by two measures: percent elongation and percent reduction in area. *Percent elongation* describes the amount that the specimen stretches before fracture and is calculated from

$$\%\text{elongation} = \frac{l_f - l_0}{l_0} \times 100 \tag{10}$$

where l_0 is the initial gauge length and l_f is the final gauge length. The *percent reduction in area* describes the amount of thinning that the specimen undergoes during the test and can be determined from the following equation:

$$\%\text{reduction in area} = \frac{A_0 - A_f}{A_0} \times 100 \tag{11}$$

where A_0 is the initial gauge area and A_f is the final gauge area. Both percent elongation and percent reduction in area also give an index of metal quality. If defects such as porosity and inclusions are present, decreases in percent elongation and percent reduction in area are often observed.

Tensile properties for some metallic engineering materials are listed in Table 2 [2].

2.1.6. Toughness and Resilience

Toughness is a measure of the ability of a material to absorb energy before fracturing and

Table 2. Tensile properties for some metallic materials [2]

Material	Treatment	Yield strength σ_{ys}, MPa	Ultimate tensile strength σ_u, MPa	Elongation, %	Reduction in area, %
Steel alloys					
1015	as-rolled	315	420	39	61
1050	as-rolled	415	725	20	40
1080	as-rolled	585	965	12	17
4340	Q + T ∗ (205 °C)	1675	1875	10	38
4340	Q + T ∗ (650 °C)	855	965	19	60
301	annealed plate	275	725	55	
403	annealed bar	275	515	35	
Aluminum alloys					
2024	T3	345	485	18	
2024	T6, T651	395	475	10	
7075	T6	505	570	11	
7075	T73	415	505	11	
Titanium alloys					
Ti – 5Al – 2.5Sn	annealed	805	860	16	40
Ti – 6Al – 4V	annealed	925	995	14	30

∗ Quenched and tempered

is given by the area under the engineering stress – strain curve. Toughness is particularly important for materials subjected to dynamic or impact loads, such as automobile bumpers, shock absorbers, and airplane landing gear. Under dynamic loads, toughness is determined by means of an impact test in which a certain load is applied suddenly to a specimen (see Section 2.6).

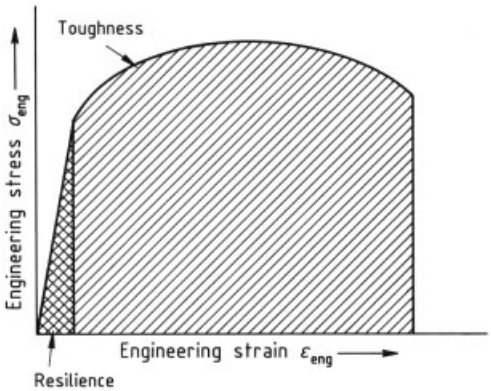

Figure 10. Schematic representation of toughness and resilience

Resilience is a measure of a material's ability to absorb energy in the elastic range. The *modulus of resilience* is given by the area under the linear portion of the uniaxial stress – strain curve. A high-resilience material has high strength and low elastic modulus. A schematic representation of toughness and resilience is shown in Figure 10.

2.1.7. True Stress – True Strain Diagram

In the engineering stress – strain diagram (Fig. 7), engineering stress decreases beyond the ultimate tensile strength, which seems to indicate that the material is becoming weaker. However, in reality the material's strength increases. The reason for the decrease of stress in the engineering stress – strain curve is that the calculated stress is based on the original cross-sectional area, whereas the area actually changes continually with plastic deformation. Calculation of the true values of stress σ_{true} and strain ε_{true} are based on the instantaneous area and length of the specimen as follows:

$$\sigma_{true} = \frac{\text{Load}}{\text{True area at the time}} = \frac{F}{A} \tag{12}$$

$$\varepsilon_{true} = \int_{l_0}^{l} \frac{dl}{l_0} = \ln\left(\frac{l}{l_0}\right) = \ln\left(\frac{A_0}{A}\right) \tag{13}$$

True and engineering stress and strain are related by the following equations:

$$\sigma_{true} = \sigma_{eng}\,(1+\varepsilon_{eng}) \tag{14}$$

$$\varepsilon_{true} = \ln\,(1+\varepsilon_{eng}) \tag{15}$$

The true stress – strain curve is compared to the engineering stress – strain curve in Figure 11. As

can be seen from the figure, the true stress necessary to sustain increasing strain continues to increase after necking.

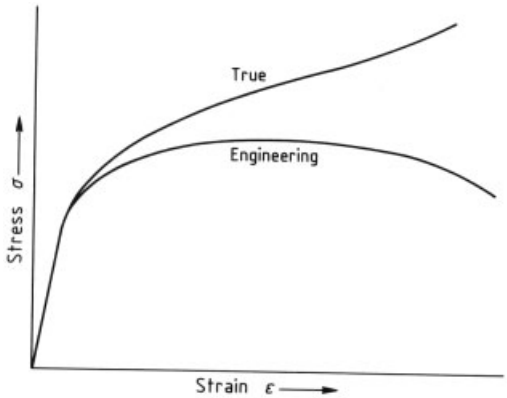

Figure 11. Comparison of true and engineering stress – strain diagrams

The true stress – strain curve is important in certain instances involving high plastic deformation of ductile materials, for example, in metal forming. If log σ_{true} is plotted against log $\varepsilon_{\text{true}}$, a straight line is obtained, as shown in Figure 12. The equation of this line is given by

$$\sigma_{\text{true}} = K\varepsilon_{\text{true}}^{n} \tag{16}$$

where K is a material property called the *strength coefficient* and n is the *strain-hardening exponent*, which gives a measure of the material's work-hardening behavior. The higher the strain-hardening exponent, the more difficult the material is to work (deform), i.e., higher stress is necessary for a given strain.

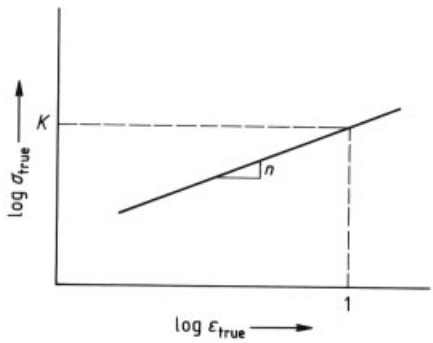

Figure 12. Plot of log σ_{true} versus log $\varepsilon_{\text{true}}$

2.1.8. Strain-Rate Sensitivity

An increase in loading rate (strain rate) usually results in an increase in flow stress of the material and is often adequately represented by the following empirical equation:

$$\sigma_{\text{true}} = K'\,(\dot{\varepsilon}_{\text{true}})^{m} \tag{17}$$

where $\dot{\varepsilon}_{\text{true}}$ is the true strain rate ($d\varepsilon_{\text{true}}/dt$), m is the *strain-rate sensitivity*, and K' is a constant. The value of m varies between zero (indicating a strain-rate-insensitive material) and unity (indicating a highly strain-rate-sensitive material). Strain-rate sensitivity generally increases with increasing temperature. The effect of strain rate on the tensile properties of a mild steel at room temperature is shown in Figure 13 [3].

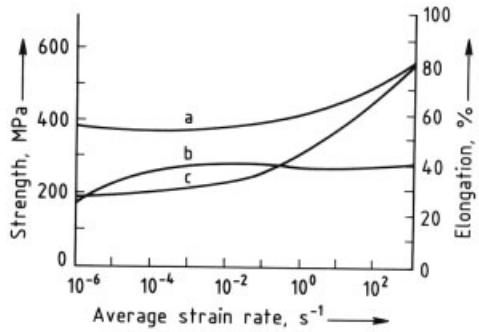

Figure 13. Effect of strain rate on tensile properties of a mild steel [3]
a) Ultimate tensile strength; b) Total elongation; c) Yield strength

2.1.9. Temperature Effects

In general, tensile properties are significantly affected by temperature. The modulus of elasticity, yield strength, and ultimate tensile strength all decrease with increasing temperature. However, the ductility of the material, as measured by percent elongation or percent reduction in area, generally increases with increasing temperature.

2.2. Compression Test

Compressive properties are used in the analysis of structures subjected to compressive loads or bending loads that result in compressive normal

stresses, as well as in processes involving large compressive deformations such as metal working (e.g., rolling of steel). Mechanical properties obtained from a compression test generally include Young's modulus, yield strength, and compressive strength. These properties are obtained by a test conducted in a manner similar to the tensile test, except that the test specimen used is often in the form of a solid circular cylinder and the force is compressive. The resulting stress – strain behavior in the plastic region is similar to its tensile counterpart, but no necking occurs and the mode of fracture is different from that for tension.

Equations (8) and (9) are also used for calculating the compressive stress and strain. In a compression test, the specimen contracts along the load direction; therefore, the final length l_f is smaller than the initial length l_0 in Equation (9), resulting in a compressive strain that is negative. In compression testing, the specimen geometry and loading should be such that the specimen does not buckle; otherwise, failure may occur by instability rather than by crushing. Details of compression testing are given in ASTM E9-89 a [4]

2.3. Bend and Flexure

2.3.1. Elastic Bending Stress and Deflection

The stress in an elastically deformed beam or bar can be calculated by using the flexure formula

$$\sigma = \frac{MC}{I} \tag{18}$$

where M is the applied bending moment, C is the distance from the neutral plane (the plane with zero stress) to the point at which the stress is being calculated, and I is the moment of inertia for the beam cross section. The value of the bending moment M depends on the magnitude of the applied load, the length of the beam, and the location of the supports. The moment of inertia I is a geometrical property of the cross section. The maximum stress $\sigma_{\max}$ and deflection $\delta_{\max}$ for two common beam geometries and support configurations are shown in Figure 14.

For the cantilever beam (Fig. 14 A), maximum stress occurs at the supported end of the beam, with tension on top and compression on the bottom, and maximum deflection at the free end of the beam. For the simply supported configuration (Fig. 14 B), maximum stress and deflection occur at the middle of the beam, with tension on the bottom and compression on top. Note that the equations given in Figure 14 apply when the load magnitude is such that only elastic deformation results.

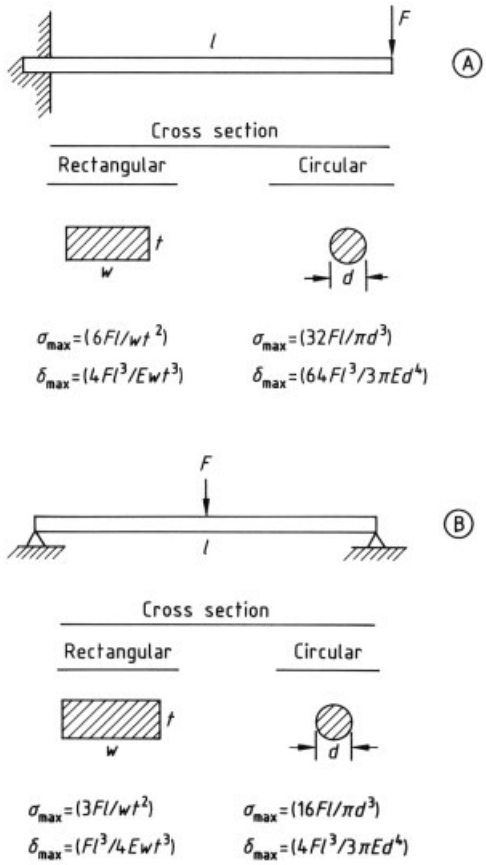

Figure 14. Maximum stress $\sigma_{\max}$ and maximum deflection $\delta_{\max}$ in selected beams
a) Cantilever beam with end load; b) Simple supports with center load

2.3.2. Bend and Flexure Test

The bend test is a method of evaluating the ductility of a material, as evidenced by its ability to resist cracking during bending. Because the loading in such a test consists of both elastic and plastic deformation (particularly in ductile materials), the flexure and deflection equations given in Section 2.3.1 cannot be used.

In a bend test, the test specimen is a bar with either rectangular or circular cross section. Bending forces are usually applied through one of the three arrangements shown in Figure 15 [5]. Bending is continued to a specified angle of bend and inside radius of curvature. If complete fracture does not occur, the convex surface of the bar is examined for cracks, and the number and size of the cracks can be used as the criterion for failure. Therefore, the bend test does not provide a quantitative method of evaluating the ductility of a material. Details of the bend test can be found in ASTM E290-87 [5].

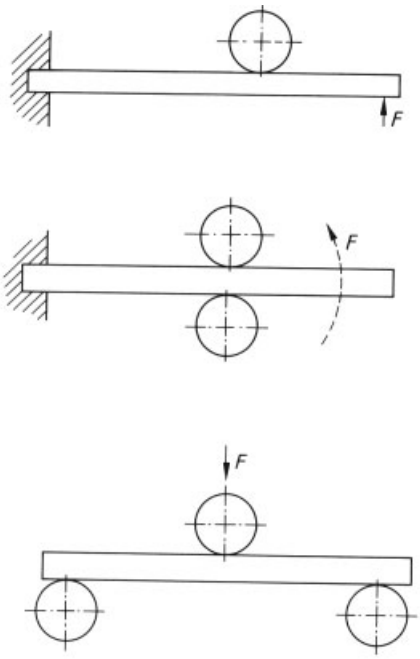

Figure 15. Bend test arrangements [5]

2.4. Shear and Torsion

2.4.1. Direct Shear Stress and Test

Mechanical joints, such as rivets, bolts, and pins, often experience direct shearing stress, which is defined as the shearing force divided by the cross-sectional area of the rivet, bolt, or pin:

$$\tau_{\text{ave}} = \frac{V}{A} \tag{19}$$

where V is the shear force and A is the cross-sectional area. Equation (19) gives the average shear stress that the rivet, bolt, or pin experiences. Rivets, bolts, or pins that resist shearing forces at a single cross section are called single-shear joints, and those resisting shearing forces at two cross sections are called double-shear joints, as illustrated in Figure 16. The average shearing stress for each condition is also given in Figure 16.

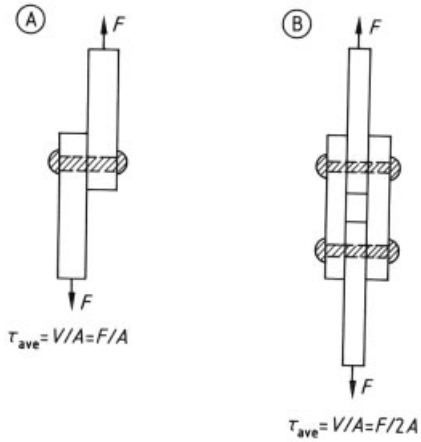

Figure 16. Single- and double-shear joints
a) Single; b) Double

The shear strength of a joint is determined by a shear test in which the joint is usually subjected to double-shear loading (Fig. 16 B). The load required to fracture the joint is determined, and the shear strength is calculated from

$$S = \frac{F_{\max}}{2A} = \frac{F_{\max}}{2\,(\pi d^2/4)} = \frac{2F_{\max}}{\pi d^2} \tag{20}$$

where S is the maximum shear strength, $F_{\max}$ is the maximum load in the test, and d is the diameter of the joint.

2.4.2. Torsional Shear Stress and Test

A torque applied to a circular bar results in twisting of the bar, as shown in Figure 17. Point A on the circular bar moves to point A′ when a torque T is applied. Details of shear testing are given in ASTM B 565-87 [6].

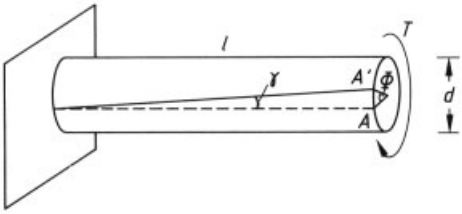

Figure 17. Torsion of a solid bar

A shear strain is produced that varies linearly with the radius of the bar. The maximum shear strain $\gamma_{\max}$ occurs on the lateral surface and is given by

$$\gamma_{\max} = \frac{\Phi d}{2l} \tag{21}$$

where d is the diameter, Φ is the displacement angle, and l is the length.

For elastic twisting, the displacement angle Φ, which is referred to as the angle of twist, can be calculated from

$$\Phi = \frac{32Tl}{\pi d^4 G} \tag{22}$$

In elastic twisting, the shear stress also varies linearly with the radius. The maximum shear stress $\tau_{\max}$ occurs on the lateral surface and is given by

$$\tau_{\max} = \frac{16T}{\pi d^3} \tag{23}$$

The shear stress and strain are related by the shear modulus G discussed in Section 1.2.2.

A torsional test is analogous to a tensile test in that shear stress τ is plotted against shear strain γ. The slope of the initial linear portion is the shear modulus G (Fig. 3), which is related to Young's modulus and Poisson's ratio by Equation (6), discussed in Section 1.2.4. Because of this relationship, if Young's modulus and Poisson's ratio are determined from a tensile test, a torsional test is not required to determine G.

2.5. Hardness Tests

Hardness tests are probably the most frequently performed mechanical tests because they are nondestructive and usually no special specimen need be prepared.

Hardness tests measure the resistance to penetration of the surface of a material by an indenter with a force applied to it. The hardness of materials depends on the type of binding forces between atoms, ions, or molecules and, like strength, increases with the magnitude of these forces. Because indentation occurs by plastic deformation in metals and alloys, hardness is inherently related to the material's plastic resistance and, therefore, to its yield strength and wear resistance.

A hard indenter (such as hardened steel, tungsten carbide, or diamond) of a standard shape (either round or pointed) is pressed into the surface of the material under a known force. The resulting area or depth of indentation is measured, which in turn is related to the hardness number. This process is shown schematically in Figure 18.

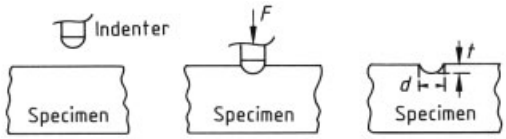

Figure 18. Hardness test

Depending on the material tested and the type of test, various combinations of load and indenters are used. The most common methods of hardness testing for metals are the Rockwell, Brinell, Vickers, and Knoop methods. Table 3 summarizes the type of indenters and types of impressions associated with these four common hardness tests.

2.5.1. Rockwell Hardness Test

The Rockwell hardness test is widely used; it has scales (Rockwell A, Rockwell B, etc.) for different hardness ranges. A small-diameter steel ball is used as the indenter for soft materials, and a diamond cone for hard materials. The depth of penetration is measured automatically by the testing machine and converted to a Rockwell hardness number. Therefore, each measurement requires only a few seconds. The Rockwell scale is designated by an R with appropriate scale letter as a subscript (i. e., R_B indicates Rockwell

Table 3. Hardness tests [7]

Test	Indenter	Shape of indentation: Side view	Shape of indentation: Top view	Load	Formula for hardness number
Brinell	10-mm sphere of steel or tungsten carbide	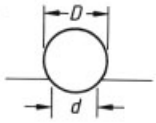	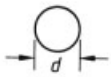	F	$\mathrm{BHN} = \dfrac{2F}{\pi D(D - \sqrt{D^2 - d^2})}$
Vickers	diamond pyramid	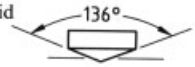		F	$\mathrm{VHN} = \dfrac{1.72F}{d_1^2}$
Knoop microhardness	diamond pyramid	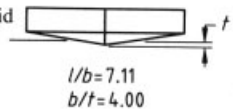	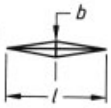	F	$\mathrm{KHN} = \dfrac{14.2F}{l^2}$
Rockwell A, C, D	diamond cone	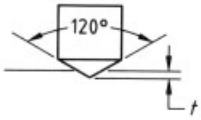		60 kg $R_A =$ 150 kg $R_C =$ 100 kg $R_D =$	100 – 500 t
B, F, G	$\frac{1}{16}$-inch-diameter steel sphere			100 kg $R_B =$ 60 kg $R_F =$ 150 kg $R_G =$ 100 kg $R_E =$	130 – 500 t
E	$\frac{1}{8}$-inch-diameter steel sphere				

B). For each scale, hardness values range up to 100. Details of the Rockwell hardness test are given in ASTM E18-89 a [8].

2.5.2. Brinell Hardness Test

In the Brinell hardness test, a single scale covers a wide range of material hardness. A hard spherical indenter (10 mm in diameter) is forced into the surface of the metal (a 500 kg load is used for softer metals such as copper and aluminum; a 3000 kg load for hard metals such as steel, iron, and alloys). The Brinell hardness number (BHN) is a function of both the magnitude of the load and the diameter of the resulting indentation. The diameter of indentation is measured with a special microscope. A schematic of the Brinell hardness test is shown in Figure 19.

The Brinell hardness number is defined as the load divided by the surface area of the indentation and is calculated from the following equation:

$$\mathrm{BHN} = \frac{F}{\left(\frac{\pi}{2}\right) D \left(D - \sqrt{D^2 - d^2}\right)} = \frac{F}{\pi D t} \tag{24}$$

where F is the applied load in kilogram-force; and D is the diameter of the indenter, d the diameter of the impression, and t the indentation depth – all in millimeters. This hardness testing method is relatively insensitive to irregularities in the material surface because the impression left by the indenter is large. Details of the Brinell hardness test are given in ASTM E10-84 [9].

2.5.3. Vickers Hardness Test

The Vickers hardness test is termed a microhardness test because it forms a small indentation of microstructural dimensions. A small diamond pyramid is pressed into the surface of the specimen, and the diagonal of the square impression is measured. Based on this measurement, the Vickers hardness number (VHN) is read from a chart. Because a small indentation is made, Vickers hardness testing is often used for measurements of small regions, for example, in assessing the relative hardness of various phases in multiphase alloys. A microscope is required to obtain the measurement. Details of the Vickers hardness test can be found in ASTM E92-82 [10].

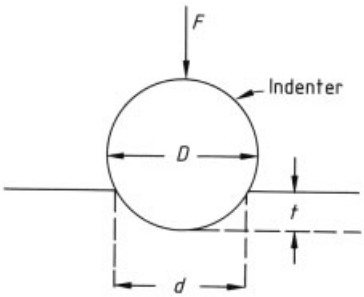

Figure 19. Schematic of the Brinell hardness test

2.5.4. Knoop Hardness Test

The Knoop test is also a microhardness test, and the resulting impression requires measurement under a microscope. A skewed diamond indenter is used in this test, and the Knoop hardness number (KHN) is calculated as the force divided by the projected indentation area. The Knoop hardness scale is approximately equivalent to the Vickers hardness scale. The Knoop hardness test is often used in measuring the microhardness of small areas of metallic material microstructures and for brittle materials such as ceramics. Details of the Knoop hardness test are given in ASTM E384-84 [11].

2.5.5. Correlation of Hardness with Strength

Since hardness and tensile strength are both indicators of a metal's resistance to plastic deformation, a strong correlation exists between them. For example, a strong linear correlation exists between the Brinell hardness number and the ultimate tensile strength σ_u of steel, as follows:

$$\sigma_u = 3.45\,\mathrm{BHN} \tag{25}$$

where tensile strength is given in megapascals. Figure 20 shows the conversion among Brinell, Vickers, and Rockwell tests, as well as the correlation between the hardness numbers and ultimate tensile strength; conversion tables are given in ASTM E140-88 [13].

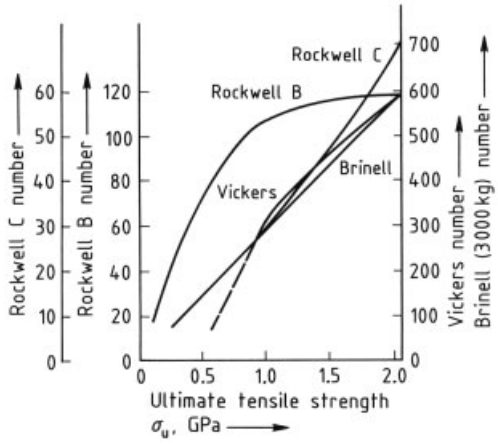

Figure 20. Correlation of hardness with tensile strength and conversion among Brinell, Vickers, and Rockwell tests [12]

2.6. Impact Testing

Impact tests measure a material's resistance to failure under an impact load. The ability of the material to resist an impact load without failure is also referred to as *toughness*. This resistance to failure by fracture is particularly important for notched components. Two common methods of impact testing are the notched bar impact test and the dynamic tear test.

2.6.1. Notched Bar Impact Test

The notched bar impact test is the most common laboratory test to measure impact energy. A Charpy V-notch specimen, shown in Figure 21, is often used. The Charpy V-notch specimen is placed in an impact testing machine, as shown

in Figure 22 [14]. A heavy pendulum is released from a known height h_0. After striking and breaking the specimen, the pendulum stops at a lower, final height h_f. If the mass of the pendulum and its initial and final elevations are known, the energy absorbed by the fracture can be determined. The energy is usually expressed in joules. In general, materials that have both high strength and high ductility have large impact energies (or high toughness).

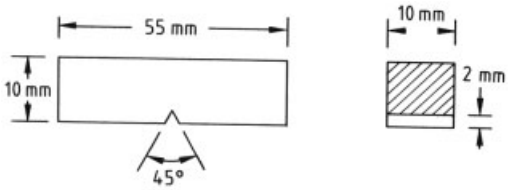

Figure 21. Charpy V-notch specimen

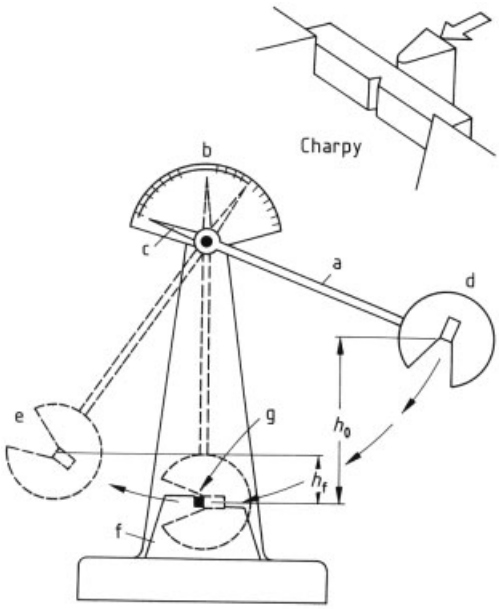

Figure 22. Standard Charpy impact test machine [14]
a) Hammer; b) Scale; c) Pointer; d) Starting position; e) End of swing; f) Anvil; g) Specimen

The impact energy of a material is usually very sensitive to temperature. At high temperature, the material behaves in a ductile manner, and a large absorbed energy is required for fracture. At low temperature, the material behaves in a brittle manner, and little deformation is observed at fracture. Therefore, the absorbed energy is low. The temperature at which the material changes behavior from ductile to brittle is called the *transition temperature*. The transition temperature can fall between ca. −100 and +100 °C, depending on the alloy and the test conditions. Figure 23 shows the effect of temperature on the impact energy of a typical material. Details of the notched bar impact test can be found in ASTM E23-88 [15].

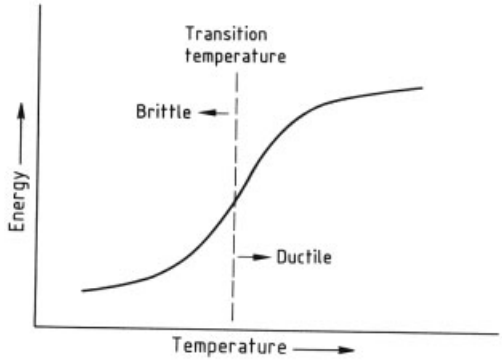

Figure 23. Effect of temperature on impact energy

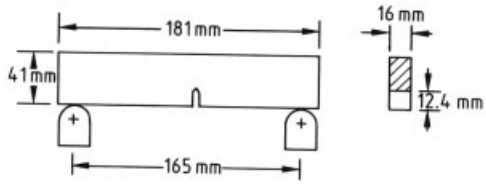

Figure 24. Dynamic tear test specimen and supports

2.6.2. Dynamic Tear Test

The dynamic tear test is conducted to determine the resistance of a metallic material to rapid progressive fracture. This resistance is measured as dynamic tear energy, which is the total energy required to fracture a dynamic tear specimen.

A single-edge notched beam specimen is used as the dynamic tear specimen and is simply supported on its ends, as shown in Figure 24.

The specimen is then impact loaded in the middle by a pendulum or drop weight. The total energy during separation of the specimen as it fractures is measured as the dynamic tear energy for the material tested. Details of dynamic tear testing can be found in ASTM E604-83 [16].

2.7. Creep Test

Permanent deformation of materials at elevated temperature when subjected to a constant externally applied load or stress below the yield strength is called *creep*. In metallic materials, this phenomenon usually occurs at temperatures of about 40 % of the melting point. The creep test measures the resistance of a material to deformation under a static load at high temperature. A cylindrical specimen is placed in a furnace, and a constant stress or load is applied while a constant temperature is maintained. As soon as the constant load or stress is applied the specimen undergoes a small elastic strain. After this initial instantaneous elastic deformation, the strain increases with time, as shown in Figure 25. The creep curve has three distinct stages. Stage I is the primary stage, during which the strain rate ($d\varepsilon/dt$) decreases, which indicates that the material is experiencing an increase in hardness. The second stage is the linear region, which indicates a constant strain rate and is referred to as the steady-state region. This stage of creep usually has the longest duration. The slope of the steady-state region of the creep curve is the creep rate $\dot{\varepsilon}$:

$$\dot{\varepsilon} = \frac{\Delta\varepsilon}{\Delta t} \tag{26}$$

In stage III, the strain rate increases due to the increase in true stress as a result of a reduction in cross-sectional area (necking), until failure occurs. This failure is called rupture, and the time required for failure to occur is the rupture time. The creep rate of a material is a function of both the applied stress level and the test temperature. Figure 26 shows the variation of the creep curve with stress and temperature.

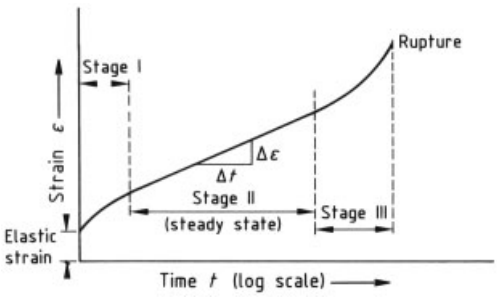

Figure 25. Typical creep curve of strain versus time at constant load or stress and constant elevated temperature

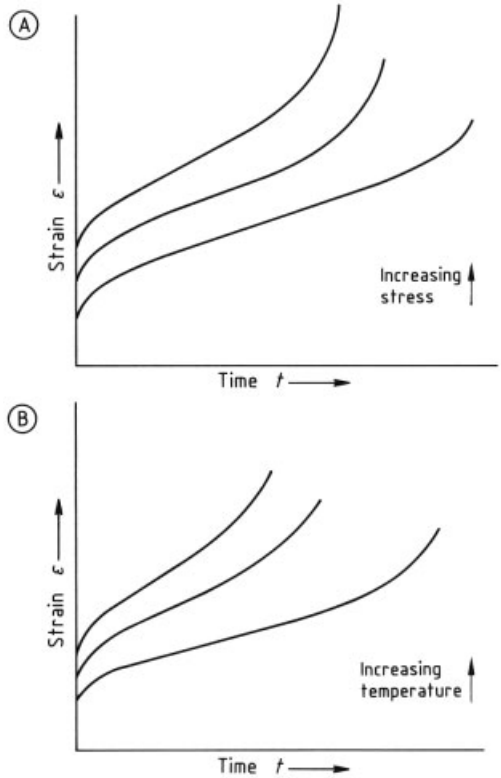

Figure 26. A) Variation of creep curve with stress; B) Variation of creep curve with temperature

To predict the expected lifetime of a component for a given combination of stress and temperature, stress rupture curves are often plotted. These curves are obtained from a series of creep tests by plotting the logarithm of the stress versus the logarithm of rupture life (Fig. 27). Details of the creep test can be found in ASTM E139-83 [17].

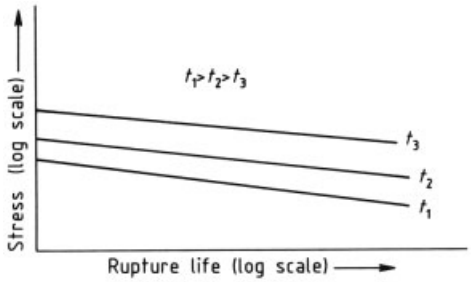

Figure 27. Stress rupture curves

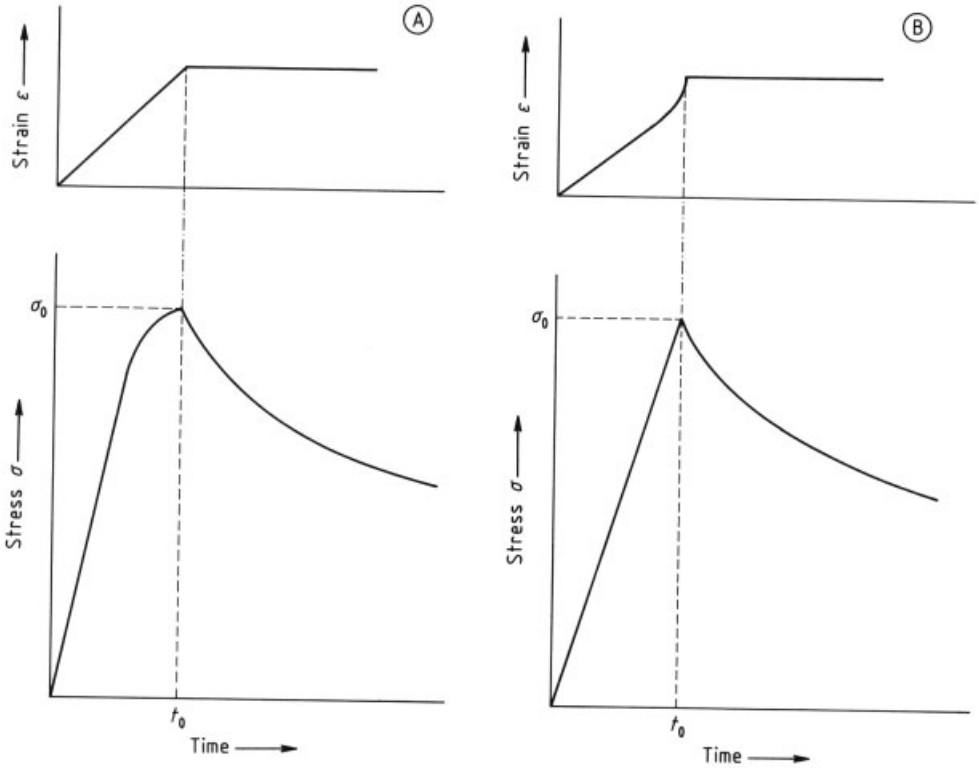

Figure 28. Stress relaxation test
A) At constant strain rate; B) At constant load rate

2.8. Stress Relaxation Tests

Stress relaxation refers to the time-dependent decrease in stress under a constant constraint and environment. Stress relaxation test data are often necessary for the design of mechanically fastened joints such as bolted and riveted assemblies, and shrink-fit components, for which permanent tightness is of major concern.

Stress relaxation tests for a material can be conducted under tension, compression, bending, or torsion loads. In general, a specimen is subjected to an increasing load until a predetermined strain is reached. The time at which this initial strain is attained is taken as the reference (zero) time. Then, while the predetermined strain is held constant, the stress during the test is monitored continuously. This procedure under constant strain rate as well as constant load rate is shown in Figure 28 [18].

Relaxed stress, which is the initial stress minus the remaining stress at a given time, is then plotted against time (Fig. 29). The absolute value of the slope of the stress relaxation curve at a given time gives the relaxation rate. Stress relaxation testing is described in ASTM E328-86 [18].

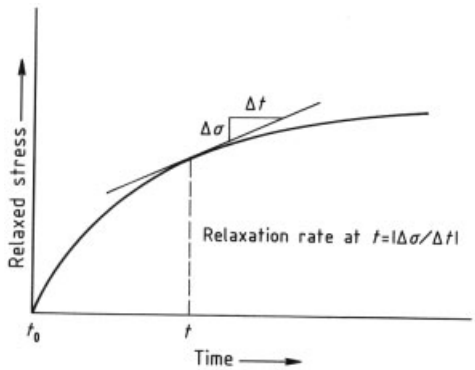

Figure 29. Typical relaxation curve

3. Fatigue Loading

3.1. Principles and Terminology

3.1.1. Fatigue Phenomena

In many applications and service conditions, metallic components and parts are subjected to repetitive or cyclic stresses. Repeated stress usually arises in moving parts that rotate, bend, or vibrate. Under cyclic stress, the material usually fails at a stress level far below its maximum static strength (yield strength or ultimate tensile strength). Failures that occur under repeated or cyclic stress are called fatigue fractures.

3.1.2. Fatigue Fracture Surfaces

A typical fatigue failure usually originates at a point of stress concentration. Stress concentrations generally result from either geometrical discontinuities in a part (e.g., keyways, holes, and fillets) or metallurgical defects (e.g., inclusions and voids). Once a fatigue crack has nucleated, the effect of stress concentration increases and the crack grows during, and as a result of, cyclic stress. The crack growth rate is primarily a function of the magnitude, and frequency of the cyclic stress. As the crack grows, the stressed area decreases in size, which results in increased magnitude of the stress, leading ultimately to rapid fracture of the remaining area. Figure 30 is a schematic representation of this process for a typical fatigue failure in rotating bending.

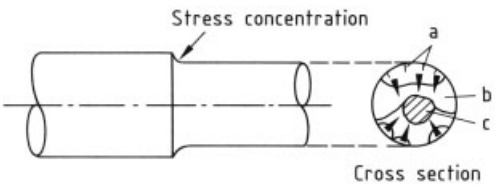

Figure 30. Schematic of a typical fatigue failure process in rotating bending
a) Crack nucleation; b) Crack growth region; c) Final fracture region

As can be seen in Figure 30, the fracture surface has three distinct regions: (1) the crack nucleation site or sites, (2) the crack growth region, and (3) the final fracture region. Since the crack surfaces are pressed together repeatedly during crack growth, the crack growth region is smooth, whereas the final fracture region usually exhibits a rough surface. In addition, the crack growth region generally contains line markings referred to as "beach marks", which result from different periods of crack extension and various positions of the crack front during its growth. The size of the crack growth region, compared to the final fracture region, depends mainly on the magnitude of the cyclic stress and the material toughness. As the magnitude of stress decreases or toughness increases, the size of the crack growth region on the fracture surface increases. In contrast to the gross deformation observed in a tensile failure for a ductile material (i.e., necking; see Section 2.1.4), typical fatigue fractures do not exhibit gross plastic deformation. Therefore, fatigue failures are often characterized as brittle failures, even for ductile materials.

3.1.3. Fatigue Behavior

The applied stress history can be completely random or simple and repetitive. Examples of a random history and a simple sinusoidal history are given in Figure 31.

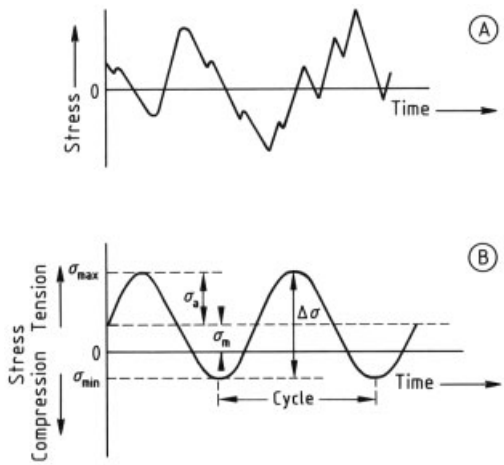

Figure 31. Cyclic variation of stress with time
A) Random history; B) Sinusoidal history

The applied cyclic stress can be axial, flexural (bending), or torsional. The standard stress variables in fatigue loading are shown in Figure 31 B and are defined as follows:

$$\text{Stress range} = \Delta\sigma = \sigma_{\max} - \sigma_{\min} \tag{27}$$

$$\text{Stress amplitude} = \sigma_a = \frac{\Delta\sigma}{2} = \frac{\sigma_{max} - \sigma_{min}}{2} \quad (28)$$

$$\text{Mean stress} = \sigma_m = \frac{\sigma_{max} + \sigma_{min}}{2} \quad (29)$$

$$\text{Stress ratio} = R = \frac{\sigma_{min}}{\sigma_{max}} \quad (30)$$

For completely reversed stressing the mean stress σ_m is zero, and $R = -1$.

Depending on the number of cycles that a part must withstand during its intended lifetime, the fatigue behavior can be classified into low-cycle or high-cycle regimes. The strain – life approach is generally used to characterize the low-cycle (typically less than 10^5 cycles) fatigue of materials, whereas the stress – life approach (see Section 3.2.1) is often used for high-cycle fatigue applications. Cyclic plasticity is an important consideration in the treatment of low-cycle fatigue, whereas strain is primarily elastic in the high-cycle fatigue regime.

The strain – life approach or strain-controlled fatigue is mainly used to evaluate the fatigue life of notched parts because significant plastic strains usually exist at the root of a notch. In this approach, the fatigue life of the notched part is related to the life of a small smooth specimen cycled to the same strain level as the material at the root of the notch. Further details of strain-controlled low-cycle fatigue testing are given in ASTM E606-80 [19].

The treatment of material fatigue behavior in this section covers only stress-controlled, constant-amplitude cyclic loading, shown in Figure 31 B. For treatment of variable amplitude or random loading and strain-controlled fatigue, see [20].

3.2. Cyclic Stress-Controlled Fatigue

3.2.1. Stress – Life Curves

Traditionally, the stress – life method has been widely used in design against fatigue. A series of tests are performed under stress-controlled conditions by using smooth polished specimens. The data from these tests are then plotted in the form of a stress – life (S – N) curve in which the stress causing failure is plotted against the number of cycles to failure. Figure 32 shows typical S – N curves for a steel and an aluminum alloy [21]. As Figure 32 indicates, the higher the stress amplitude, the smaller is the number of cycles to failure. For steel, leveling off occurs in the S – N curve with no decrease in fatigue strength as the number of cycles increases. The stress level at which the S – N curve becomes horizontal is called the *fatigue limit* or *endurance limit*. The endurance limit is the stress below which fatigue failure does not occur. Ferrous materials usually have S – N curves exhibiting an endurance limit. For steel, the fatigue limit is estimated to be about half the ultimate tensile strength. For materials such as the aluminum alloy in Figure 32, the S – N curve continues to decline and such materials do not have an endurance limit.

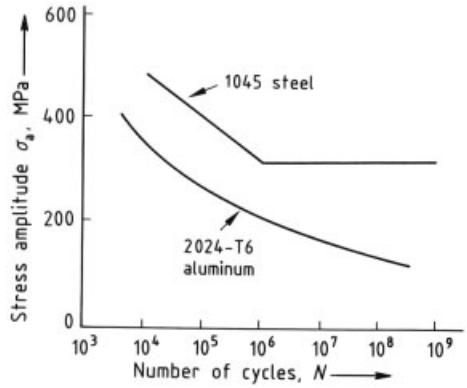

Figure 32. Typical S – N curves for a steel and an aluminum alloy [21]

3.2.2. Fatigue Tests

The rotating bending fatigue test is the most commonly used small-scale test, in which the smooth polished specimen is subjected to alternating compression and tension stresses of equal magnitude while it rotates. The apparatus is depicted schematically in Figure 33 [20]. A dead weight W produces a uniform bending moment over the test section of the specimen as it rotates. The stress is then calculated from the flexure formula (see Section 2.3.1).

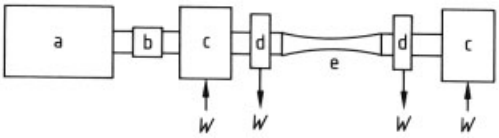

Figure 33. Schematic of rotating bending fatigue test machine [20]
a) Motor; b) Flexible coupling; c) Main bearings; d) Load bearings; e) Test piece

This type of test, however, cannot be used to simulate loading conditions with a mean stress (i.e., $R \neq -1$) because in rotating bending, the loading is completely reversed ($R = -1$). Therefore, for fatigue testing with a mean stress, an axially loaded specimen, shown in Figure 34, is often used [2]. Pulsating tension or tension – compression axial stress is provided by means of a hydraulic or mechanical actuator. A standard procedure for constant amplitude axial fatigue testing is given in ASTM E466-82 [22].

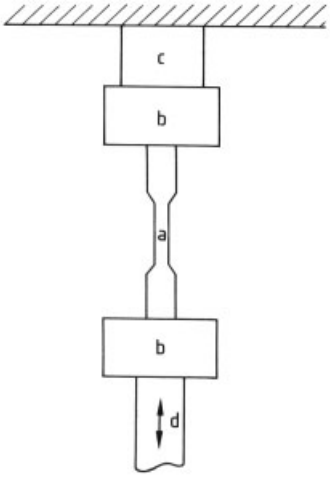

Figure 34. Axial fatigue loading [2]
a) Specimen; b) Grips; c) Load cell; d) Moving actuator

3.2.3. Mean Stress Effects

Mean stress can substantially influence a material's fatigue behavior. In general, tensile mean stresses ($\sigma_m > 0$) are detrimental to fatigue behavior, whereas compressive mean stresses ($\sigma_m < 0$) are beneficial. Typical S – N diagrams with differing mean stress levels are shown in Figure 35.

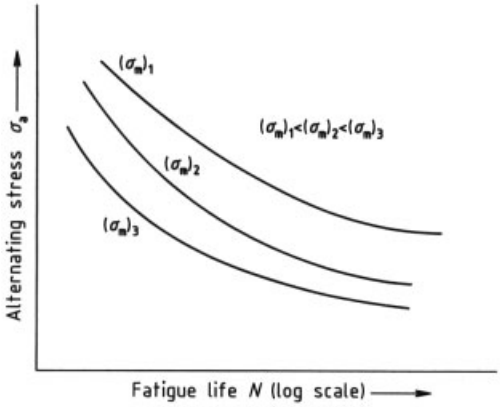

Figure 35. Effect of mean stress on fatigue life

As can be seen from Figure 35, cyclic life decreases with increasing mean stress level at a given alternating stress σ_a. To account for the effect of mean stress on long-life fatigue strength, one of the following empirical relations is generally used:

$$\text{Goodman}: \sigma_a = \sigma_{fat}\left(1 - \frac{\sigma_m}{\sigma_u}\right) \tag{31}$$

$$\text{Gerber}: \sigma_a = \sigma_{fat}\left[1 - \left(\frac{\sigma_m}{\sigma_u}\right)^2\right] \tag{32}$$

where σ_{fat} and σ_a are fatigue strengths for the conditions $\sigma_m = 0$ and $\sigma_m \neq 0$, respectively. These equations are shown graphically in Figure 36. Most experimental data fall between the two lines plotted in Figure 36. Therefore, the Goodman relation is a more conservative criterion of mean stress.

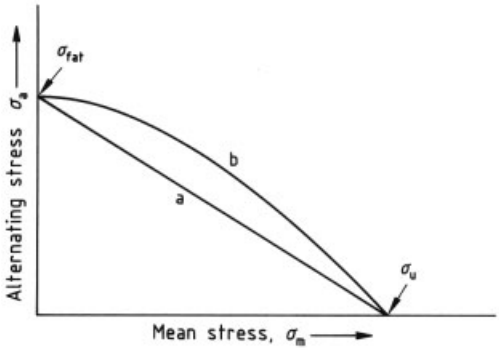

Figure 36. Relations for mean stress effects
a) Goodman; b) Gerber

4. Fracture Mechanics

4.1. Principles and Terminology

Fracture mechanics is used to study the behavior of components and structures containing cracks or other flaws. If the geometry and size of the crack or flaw are known, the fracture strength of the part or structure can be evaluated by applying the principles of fracture mechanics.

Figure 37 shows the three modes of loading under which a crack can extend. Mode I, the opening or tensile mode, is the most common. Mode II is the sliding or shear mode, and mode III, the tearing or antiplane mode.

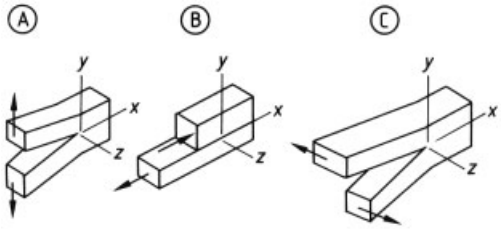

Figure 37. Crack extension modes
A) Mode I; B) Mode II; C) Mode III

Stresses in the vicinity of the crack tip, shown in Figure 38, are given by

$$\sigma_y = \frac{K}{\sqrt{2\pi r}} f_1(\theta) \tag{33}$$

$$\sigma_x = \frac{K}{\sqrt{2\pi r}} f_2(\theta) \tag{34}$$

$$\tau_{xy} = \frac{K}{\sqrt{2\pi r}} f_3(\theta) \tag{35}$$

where θ and r are geometric parameters, defined in Figure 38, and $f_1(\theta)$, $f_2(\theta)$, and $f_3(\theta)$ are functions of θ.

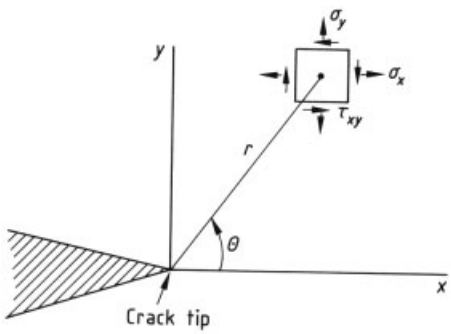

Figure 38. Stresses in the vicinity of a crack tip

These equations indicate that the stresses at a given point near the crack tip are dependent only on K, which is a stress field parameter, defined as the *stress intensity factor*, that depends on crack geometry, configuration, and loading. The general form of K is given as

$$K = \sigma Y \sqrt{\pi a} \tag{36}$$

where σ is the nominal applied stress, a is crack length, and Y is a dimensionless parameter that depends on specimen and crack geometry. Typical mode I stress intensity expressions for several common specimen configurations, shown in Figure 39, are as follows [20]: for a center-cracked plate in tension (Fig. 39 A),

$$K_I = \sigma\sqrt{\pi a} \cdot \sqrt{\sec(\pi a/w)}, \text{ where } \sigma = \frac{F}{wt} \tag{37}$$

for a single-edge cracked plate in tension (Fig. 39 B),

$$K_I = \sigma\sqrt{\pi a}\left[1.12 - 0.231\left(\frac{a}{w}\right) + 10.55\left(\frac{a}{w}\right)^2 - 21.72\left(\frac{a}{w}\right)^3 + 30.39\left(\frac{a}{w}\right)^4\right], \quad \text{where } \sigma = \frac{F}{wt} \tag{38}$$

and for a single-edge cracked beam in pure bending (Fig. 39 C),

$$K_I = \sigma\sqrt{\pi a}\left[1.12 - 1.40\left(\frac{a}{w}\right) + 7.33\left(\frac{a}{w}\right)^2 - 13.08\left(\frac{a}{w}\right)^3 + 14.0\left(\frac{a}{w}\right)^4\right], \quad \text{where } \sigma = 6M/tw^2 \tag{39}$$

The subscript I in these equations denotes mode I (opening or tensile mode) loading. Stress intensity factor expressions for other common crack geometry and loadings can be found in [2], [20].

4.2. Fracture Toughness

Fracture toughness is the critical value of the stress intensity factor and refers to the condition in which a crack extends in a rapid (unstable) manner. Fracture toughness can therefore be considered the limiting value of the stress intensity factor and is designated K_c. If the fracture

toughness and the size and geometry of the crack in a material are known, the fracture stress σ_f can be calculated from

$$\sigma_f = \frac{K_c}{Y\sqrt{\pi a}} \tag{40}$$

In general, the interaction of material properties (i.e., K_c), design stress, and crack size controls the fracture condition in a cracked component. If any two of these three parameters are known, the third can be calculated from Equation (40).

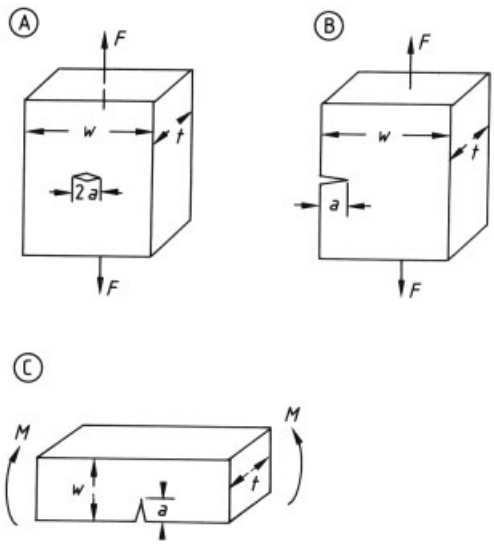

Figure 39. Several common crack configurations
A) Center-cracked plate in tension; B) Single-edge cracked plate in tension; C) Single-edge cracked beam in pure bending

Although fracture toughness can be considered a material property, its value depends on thickness. In general, a material can exhibit ductile behavior in a thin sheet, whereas the same material may fracture in a brittle manner in a thick plate. The thicker the material, the lower is the fracture toughness, as shown in Figure 40. As thickness increases, K_c approaches an asymptotic minimum value called the *plane-strain fracture toughness*, denoted by K_{Ic}. Values of K_{Ic} for several metallic materials are given in Table 4 [23].

In general, the fracture toughness of metals and alloys decreases as yield strength increases. Therefore, even though the fracture resistance of a component may be increased by choosing a higher strength material, if a crack (or flaw) exists, the fracture resistance can actually decrease. Also, the fracture toughness of metals depends on strain rate and environmental conditions such as temperature and corrosion. In general, as temperature decreases, fracture toughness usually decreases, whereas yield strength generally increases.

Table 4. Values of K_{Ic} for some metallic materials [23]

Material	K_{Ic}, MPa · m$^{1/2}$	Yield strength σ_{ys}, MPa
Steel		
4340	99	860
4340	60	1515
52100	14	2070
Aluminum		
2024	26	455
7075	24	495
7178	33	490
Titanium		
Ti–6Al–4V	115	910
Ti–6Al–4V	55	1035

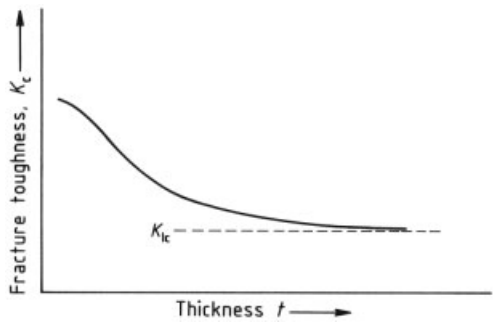

Figure 40. Variation of fracture toughness with thickness

Fracture Toughness Testing Because K_{Ic} represents a lower limiting value of fracture toughness it can be considered an important material property. By using this property, a relationship between failure stress and crack or defect size can be obtained (Eq. 40).

Different specimen configurations can be used for K_{Ic} testing. Acceptable specimen configurations for K_{Ic} testing according to the ASTM standard are the bend, arc-shaped, compact, and disk-shaped compact specimens (Fig. 41) [24].

The crack length a is nominally equal to the thickness, which is usually half of the width w. For a valid plane-strain fracture toughness test,

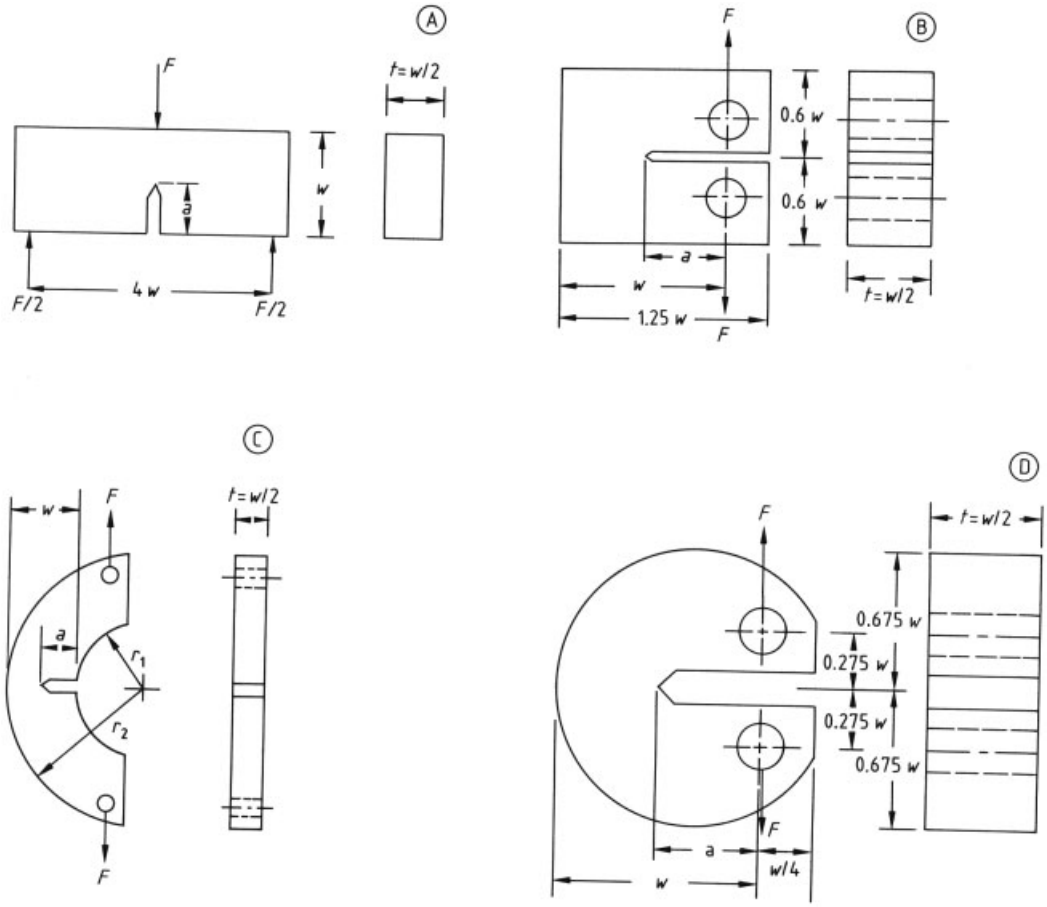

Figure 41. Configurations of K_{Ic} test specimens [24]
A) Bend specimen; B) Compact specimen; C) Arc-shaped specimen; D) Disk-shaped compact specimen

the specimen thickness and crack length should satisfy the following criteria:

$$\text{thickness}\, t\, \text{and crack length}\, a \geq 2.5 \left(\frac{K_{\mathrm{Ic}}}{\sigma_{\mathrm{ys}}}\right)^2 \tag{41}$$

The test sample is usually fatigue loaded to extend the machined notch for a sharp initial crack. A clip gauge is then placed at the crack mouth in order to monitor crack displacement as the specimen is loaded. Typical load – displacement plots are shown in Figure 42 [24].

To determine a valid K_{Ic} value, the following procedure is used [24]: the secant $0F_5$ is drawn through the origin of the test plot with a slope that is 5 % lower than the tangent line 0A (see Fig. 42). If the load at every point on the load – displacement that precedes F_5 is lower than F_5, then $F_5 = F_{\mathrm{Q}}$. This situation is shown in Figure 42, type I. However, if a maximum load precedes F_5 and exceeds it, this maximum load is F_{Q} (types II and III in Fig. 42). Once F_{Q} has been determined according to this procedure, K_{Q} is calculated from it. For the compact tension specimen (Fig. 41 B), K_{Q} is given by [24]

$$K_{\mathrm{Q}} = \frac{F_{\mathrm{Q}}}{t\sqrt{w}}\, f\left(\frac{a}{w}\right) \tag{42}$$

where

$$f\left(\frac{a}{w}\right) = \frac{\left(2+\frac{a}{w}\right)\left(0.886+4.64\frac{a}{w}-13.32\frac{a^2}{w^2}+14.76\frac{a^3}{w^3}-5.6\frac{a^4}{w^4}\right)}{\left(1-\frac{a}{w}\right)^{3/2}} \tag{43}$$

Similar K_Q expressions for other specimen geometries are given in [24]. If K_Q satisfies Equation (41), then $K_Q = K_{Ic}$. Otherwise, a thicker or more deeply cracked specimen should be used. For further details, see [24].

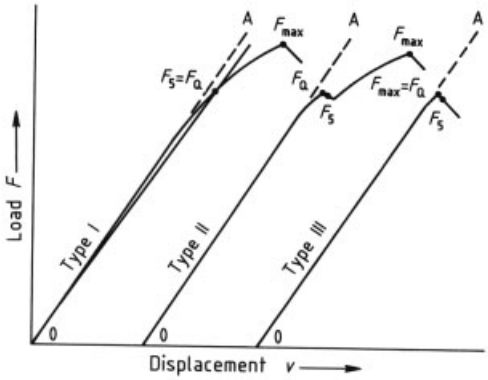

Figure 42. Load – displacement plots in a K_{Ic} test [24]

5. References

1. Annual Book of ASTM Standards, vol. 3.01, 1989, Designation E8M-89: Standard Test Methods for Tension Testing of Metallic Materials.
2. R. W. Hertzberg: *Deformation and Fracture Mechanics of Engineering Materials,* 3 rd ed., Wiley Interscience, New York 1989, p. 21.
3. M. J. Manjoine: "Influence of Rate of Strain and Temperature on Yield Stresses of Mild Steel," *J. Appl. Mech.* **66** (1944) A211-A218.
4. In [1], Designation E9-89a: Standard Test Methods for Compression Testing of Metallic Materials at Room Temperature.
5. In [1], Designation E290 – 87: Standard Test Method for Semi-Guided Bend Test for Ductility of Metallic Materials.
6. In [1], Designation B565-87: Standard Method for Shear Testing of Aluminum and Aluminum-Alloy Rivets and Cold-Heading Wire and Rods.
7. H. W. Hayden, W. G. Moffatt, J. Wulff: *The Structure and Properties of Materials,* **vol. III,** Wiley Interscience, New York 1965, p. 12.
8. In [1], Designation E18-89a: Standard Test Methods for Rockwell Hardness and Rockwell Superficial Hardness of Metallic Materials.
9. In [1], Designation E10-84: Standard Test Method for Brinell Hardness of Metallic Materials.
10. In [1], Designation E92-82: Standard Test Method for Vickers Hardness of Materials.
11. In [1], Designation E384-84: Standard Test Method for Microhardness of Materials.
12. R. A. Flinn, P. K. Trojan: *Engineering Materials and Their Applications,* 3 rd ed., Houghton Mifflin Co., 1986, p. 84.
13. In [1], Designation E140-88: Standard Hardness Conversion Tables for Metals.
14. H. W. Hayden, W. G. Moffatt, J. Wulff: *The Structure and Properties of Materials,* **vol. III,** Wiley Interscience, New York 1965, p. 13.
15. In [1], Designation E23-88: Standard Test Methods for Notched Bar Impact Testing of Metallic Materials.
16. In [1], Designation E604-83: Standard Test Method for Dynamic Tear Testing of Metallic Materials.
17. In [1], Designation E139-83: Standard Practice for Conducting Creep, Creep – Rupture and Stress – Rupture Tests of Metallic Materials.
18. In [1], Designation E328-86: Standard Methods for Stress Relaxation Tests for Materials and Structures.
19. In [1], Designation E606-80: Standard Recommended Practice for Constant Amplitude Low-Cycle Fatigue Testing.
20. H. O. Fuch, R. I. Stephens: *Metal Fatigue in Engineering,* Wiley Interscience, New York 1980.
21. H. W. Hayden, W. G. Moffatt, J. Wulff: *The Structure and Properties of Materials,* **vol. III,** Wiley Interscience, New York 1965, p. 15.
22. In [1], Designation E466-82: Standard Practice for Conducting Constant Amplitude Fatigue Tests of Metallic Materials.
23. J. E. Shigley, C. R. Mischke: *Mechanical Engineering Design,* 5th ed., McGraw-Hill, New York 1989, p. 224.
24. In [1], Designation E399-83: Standard Test Method for Plane-Strain Fracture Toughness of Metallic Materials.

Mechanical Wear → Abrasion and Erosion
Medical Diagnostic Reagents → Diagnostic Reagents

Nondestructive Testing

KANJI ONO, Department of Materials Science and Engineering, University of California, Los Angeles, California 90024, United States

1. Introduction 1
2. Methods of Nondestructive Testing 2
2.1. Material Identification 2
2.2. Radiography 4
2.2.1. Radiation 4
2.2.2. Attenuation 4
2.2.3. Imaging 6
2.2.4. Recording Images 6
2.2.5. Screens 8
2.2.6. Radiographic Quality 8
2.2.7. Radiographic Practices 9
2.2.8. Radiographic Interpretation 10
2.2.9. Other Techniques 10
2.2.10. Radiation Safety 11
2.3. Ultrasonics 11
2.3.1. Waves 12
2.3.2. Generation and Detection 13
2.3.3. Inspection Methods 16
2.3.4. Display Methods 16
2.3.5. Interpretation 17
2.3.6. Inspection Standards 18
2.4. Acoustic Emission 19
2.4.1. Detection 19
2.4.2. Material Behavior 19
2.4.3. Source Location 20
2.4.4. Acoustic Emission Monitoring . . . 20
2.5. Penetrant Methods 21
2.6. Magnetics and Electromagnetics . 22
2.6.1. Magnetic Particle Inspection 22
2.6.2. Magnetic Leakage Field Inspection . 24
2.6.3. Other Magnetic Inspection Methods 24
2.6.4. Eddy-Current Inspection 24
2.6.5. Microwave Inspection 27
2.7. Thermal and Optical Methods . . 28
2.7.1. Thermal Inspection 28
2.7.2. Visual and Optical Inspection 28
2.7.3. Optical Holography 29
2.8. Leak Testing 29
2.8.1. Pressurized Gas Leaks 30
2.8.2. Pressurized Liquid Leaks 30
2.8.3. Vacuum System Leaks 31
3. Uses 31
3.1. Metallic Pressure Vessels and Piping 31
3.2. Composite Pressure Vessels and Piping 32
3.3. Weldments 33
3.4. Other Uses 33
4. References 34

1. Introduction

Nondestructive testing of materials has become an indispensable part of modern industry. Beginning with the oil and whiting method, a forerunner of the penetrant testing method used for crack detection, a variety of tests – based on numerous physical principles – have been developed. At the start, the primary task was to find flaws and discontinuities. With the advent of fracture mechanics, the measurement of flaw dimensions became more urgent. Increasing use of fiber-reinforced plastics in the past decade has posed new challenges in material characterization.

Techniques of nondestructive evaluation have made substantial advances. However, many problems still remain because the current design of components and structures demands higher efficiency and the material capability is exploited to its fullest extent. For example, a bonded joint of composite materials is a desirable design approach, yet no nondestructive testing method is available to assure the strength of the joint.

This article provides brief summaries of widely used methods of nondestructive testing, covering basic concepts, techniques in common use, and examples of industrial applications. Less well developed but promising methods are also introduced. Because the field of nondestructive testing is broad, all methods of importance to the chemical industry are included, but those applicable primarily to the aerospace, nuclear, and steel industries are covered only incidentally.

Two major problems, corrosion and cracking, confront nondestructive test engineers in the chemical and petroleum industries. Fer-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b01_11

rous and corrosion-resistant nonferrous alloys used in vessels and piping are often subjected to chloride- and hydrogen-containing environments. Hydrogen-induced cracking and stress corrosion cracking are found frequently. Cracking of welds from residual stress or poor weld quality is another persistent problem during service. Many vessels and pipes are constructed from carbon steel, and general corrosion from long-term service must be monitored. In addition, fiber-reinforced plastic components require special attention because traditional nondestructive testing methods cannot adequately evaluate their reliability.

Nondestructive testing attempts to discover indications of discontinuities and abnormal conditions in materials and structures. If an indication is found to be beyond an acceptable limit, it is a flaw and the material (or structure) must be repaired or rejected. A very small gas hole in a weld is innocuous and can be allowed to remain. Beyond a certain size, however, it is no longer tolerated, while cracks of any size are typically rejected. Flaws and discontinuities are treated as interchangeable because the size limits are set by design considerations.

Flaws are introduced at three stages. The first is during primary manufacturing. Shrinkage cavities, inclusions, cracks, blowholes, and other faults occur in castings. Flaws in forgings may be traced to ingots that contain segregations, pipes (shrinkage cavities), inclusions and, hydrogen-induced cracks or hydrogen flakes. During forging, forging bursts (internal tears), surface flaws (such as cold shuts, laps, folds, and seams), burning, and decarburization are often the result of improper setup and control. Rolling and drawing of ingots and billets generate flaws similar to those of the forging process. In addition, lamination and chevron defects may occur. The second stage is fabrication into structures, in which welding is the most important process. Because the fusion of metals is induced, many flaws are also introduced. These include gas porosity, slag entrapment, incomplete fusion, incomplete penetration, and cracks. In addition, surface flaws such as undercut, mismatch, and underfill occur. The third stage of flaw generation is during service. Overloading and fatigue loading produce cracks, and corrosion leads to pitting, general reduction of wall thickness, and stress corrosion cracking. Exposure to hydrogen results in hydrogen-induced cracking, and fluid flow produces erosion damage. High-temperature environments in furnaces and reactors cause creep and oxidation, leading to wall thinning and cracking. The nature and origin of the flaws listed here are discussed in [1], [2].

Detailed information on nondestructive testing can be found in introductory texts [3], [4], training guides [6], handbooks [7, vols. 1–6], codes and standards [8–10], and a series of monographs and conference proceedings [11], [12–15]. Several journals and a newsletter are published in English [16], and publications in other languages are also available. More than a dozen national societies for nondestructive testing, such as the American Society for Nondestructive Testing (ASNT), the British Institute for Nondestructive Testing (BINDT), Deutsche Gesellschaft für Zerstörungsfreie Prüfung e. V. (DGZfP), and the Japanese Society for Nondestructive Inspection (JSNDI), exist worldwide, and the International Institute of Welding (IIW) and the International Standards Organization (ISO) are active in establishing standards. The Institut für Zerstörungsfreie Prüfverfahren (IZfP; Saarbrücken, Federal Republic of Germany) and the Electric Power Research Institute (EPRI) NDT Center are research institutes dedicated to studies of nondestructive testing.

2. Methods of Nondestructive Testing

2.1. Material Identification

To verify that an alloy is within specified composition limits, quantitative chemical analysis is the only answer. However, a system of rapid identification of metals and alloys is useful in confirming the alloy type or in sorting mixed lots of alloys. Several techniques are used singly or in combination. The basic technique is to use mass density (specific gravity) for preliminary grouping into light (Al, Mg, Ti), medium (steel, Cu, Ni, Zn), heavy (Pb, Mo, Ag), and very heavy (W, Ta, Pt) alloys. Some alloys have distinct colors, but most have a white or grayish white color.

X-Ray Fluorescent Analysis. Each element generates characteristic X-rays at several specific energy levels when it is irradiated with electromagnetic radiation having a higher energy

spectrum. Characteristic X-ray energies range from 1 keV for sodium to 118 keV for uranium K radiation, and from 1 keV for zinc to 17 keV for uranium L radiation. Fluorescent X-rays are detected with an energy-discriminating proportional counter or solid-state detector and are electronically processed to identify the chemical composition of an alloy sample. Because of the high absorption of low-energy X-rays, only the surface of the sample can be analyzed.

Radioactive isotopes are often used as the source of radiation for exciting fluorescent X-rays. An isotope source of 1 – 100 millicuries (mCi) can be small and portable. Some sources must be replaced periodically. Regular X-ray sources or small, air-cooled X-ray tubes can also be used. For X-ray energy analysis, a solid-state detector commonly used in a scanning electron microscope (SEM) for elemental detection can be employed, but such a detector requires constant cooling with liquid nitrogen. Another limitation arises from the resolution of the X-ray energy analyzer; elements lighter than aluminum cannot be evaluated. In the electron microprobe analyzer, energetic electrons are used to generate fluorescent X-rays; a high-resolution energy spectrum analyzer is employed, which is capable of detecting an element as light as boron. Small samples can be analyzed in an SEM or microprobe.

Chemical Spot Testing. Chemical spot testing utilizes chemical color reactions, which are carried out on a sample, on filter paper, or on a spot plate. A test kit consists of about 100 solutions. Carbon steels, for example, can be identified by placing a drop of 50 % nitric acid on a fresh surface. After 5 min, the spot is rinsed with water; a brown spot indicates carbon steel. See [17] for a listing of solutions and identification tests.

Spark Testing. The spark testing technique classifies ferrous alloys according to their chemical composition by visual examination of the sparks thrown off when a sample is held against a high-speed grinding wheel. The test is fast and economical, yet an experienced operator can identify a variety of alloys with reasonable accuracy [1], [2], [18].

Hardness Testing. Hardness testing is not strictly nondestructive because an indenter is forced into the surface of a test piece. The resulting indentation is, however, often sufficiently small to be left on a finished product. The indenter is a ball, cone, or pyramid made of hardened steel or diamond. Constant force is applied on the indenter, and the diameter, length, or depth of the indentation is used as a measure of material hardness. The Brinell, Rockwell, Vickers, and Knoop hardness tests are commonly employed [19] (see also → Mechanical Properties and Testing of Metallic Materials). Because each alloy has a certain hardness range, knowing the hardness aids in material identification. Hardness measurement is also a nondestructive test method for the yield and tensile strength of ductile alloys based on empirical correlations. In brittle ceramic materials, cracks emanate from the corners of a Vickers hardness indentation. The length of these cracks has been found to be related to the fracture toughness of ceramics. Thus, nearly nondestructive determination of ceramic fracture toughness can be conducted.

Electrical and Magnetic Testing. Electrical resistivity can be used to classify metals and alloys and the state of heat treatment of many alloys. One technique involves use of a four-point probe with four equally spaced tips. The probe passes a direct current (I_{dc}) between the two outermost tips and measures the potential (V_{t}) between two inner tips. For samples whose thickness is more than three times the distance between two tips (S_{t}), the electrical resistivity (ϱ_{e}) is

$$\varrho_{\mathrm{e}} = 2\pi S_{\mathrm{t}} V_{\mathrm{t}} / I_{\mathrm{dc}}$$

Another technique uses the eddy-current principle (see Section 2.6.4). A probe containing a coil(s) is placed on the surface of a sample. Changes in coil impedance are analyzed and indicated as electrical conductivity (the inverse of resistivity) values on the analyzer output device, for example, a digital readout.

The magnetic properties of alloys can be checked with a hand magnet to classify them as strongly magnetic, weakly magnetic, or nonmagnetic. Austenitic stainless steels and some nickel alloys are weakly magnetic or nonmagnetic, depending on cold working and heat treatment. The eddy-current instrument for electrical

conductivity testing can also be used for quantitative evaluation of magnetic permeability.

Heated and cold electrodes are brought in contact with the surface to evaluate the thermoelectric properties of a sample [9, E 977]. This is useful in sorting alloys by comparative measurements.

Composite Phase Testing. In fiber-reinforced plastic composites, the amount of fiber reinforcement is an important material variable. Destructive testing for fiber content uses acid digestion and ignition techniques to eliminate the resin matrix. Ultrasonic velocity has been shown to depend on fiber content, although few dedicated instruments are in use. An instrument based on elastostatic principles has also been developed [38]; it can determine the fiber content and the porosity fraction from differential compressibility measurements (i.e., measuring the thickness of a sample under different static loading conditions) together with destructive calibration procedures. When the density and composition of matrix and fiber phases differ, as in the case of a resin matrix with glass fibers, X- or gamma-ray absorption can be used to determine the fiber content.

2.2. Radiography

Radiography utilizes the differential attenuation of penetrating radiation to find internal flaws such as porosity and compositional changes, as well as variations in thickness and cracks. X-rays, gamma-rays, and neutrons are used and are imaged by photographic films, fluoroscopic screens, and more recently, electronic imaging devices. Thickness changes of 1 – 2 % and flaw sizes of a few tenths of a millimeter can typically be detected. This method examines a large volume of material simultaneously, and most welds are radiographed routinely. However, radiation safety measures must be rigorously enforced. For details, see [7, vol. 3], [20], and more than 20 ASTM standards [9].

2.2.1. Radiation

X-rays are generated when high-energy electrons strike a metallic target. When the electrons are decelerated at the target, a continuous spectrum of X-ray photons is produced, having energies lower than that of the electrons. The peak X-ray intensity occurs at about one-third of the maximum photon energy. At photon energies of less than 118 keV, characteristic X-rays are also produced and their energies are specified by the target element. A common choice of target material is tungsten, but copper, iron, and cobalt are also used. Tubes are used to generate X-rays, with tube voltages up to 1000 kV. These consist basically of a diode with a filament for electron emission (cathode) and a target (anode), between which an accelerating potential is applied. The cathode has a focusing cup to control the size of the electron beam. The anode is fluid cooled, and the target is usually tilted to direct the X-rays generated. The vacuum envelope is made partly of glass or ceramic for electrical insulation and has a thin beryllium X-ray window.

Higher energy X-rays are generated by using an electron linear accelerator (a linac). Electron energies of 1 – 30 MeV can be obtained with commercial equipment, which employs microwave power to accelerate electrons in a series of cylindrical cavities. Betatrons, as well as Van de Graaff and resonant transformers, have also been utilized, but they are larger and less convenient than linacs.

Gamma rays are produced by radioisotopes such as cobalt-60 and cesium-137. Although radioisotope sources require radiation shielding and safe handling equipment, no power supply is needed. This results in portability, but the radiation intensity is low and long exposure times are common. Radioisotopes (e.g., ^{60}Co) are also a convenient source of the high-energy radiation needed to penetrate thicker test pieces.

Neutrons are produced by nuclear reactors, deuterium accelerators, or isotopes such as californium-252. Because of the capability of neutrons to detect explosives, neutron radiography is used by the military. Industrial applications are uncommon.

2.2.2. Attenuation

Both X-rays and gamma rays are attenuated in a substance by absorption and scattering. At lower energy levels, an incoming photon is absorbed by ionization of an atom (photoelectric effect).

This effect decreases with increasing photon energy E_p as $E_p^{-3.5}$, except when the energy corresponds to the electronic binding energy, at which point absorption increases greatly. Such increases are called absorption edges. Another absorption process, in which the photon energy is used to create an electron – positron pair, occurs above 1 MeV. Scattering of a photon by an electron occurs either coherently without changing the kinetic energy (Rayleigh scattering) or incoherently with loss of the photon energy and changes in photon direction (Compton scattering). Figure 1 shows changes in attenuation for several materials as a function of energy.

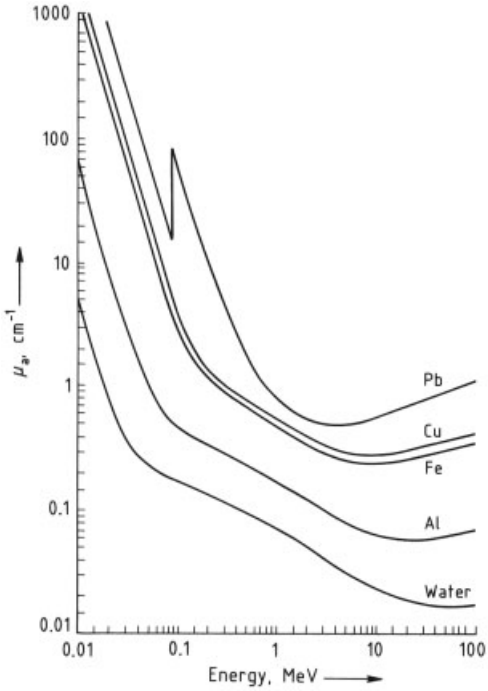

Figure 1. Absorption coefficient μ_a versus energy for various materials

Radiation intensity I decreases exponentially with the thickness x of a uniform material:

$$I = I_0 \exp(-\mu_a x) = I_0 \exp(-N_a \sigma_a x)$$

where I_0 is the incident intensity, μ_a the linear attenuation coefficient, N_a the number of atoms per unit volume, and σ_a the atomic attenuation coefficient. When μ_a is divided by mass density ϱ, the mass attenuation coefficient (μ_a/ϱ) is obtained to express attenuation per unit areal weight. In steels and nickel or copper alloys, photoelectric absorption is dominant at a photon energy of less than 100 keV. In light metals (aluminum alloys), this range shifts to less than 40 – 50 keV. The atomic attenuation coefficient increases with atomic number. The exponential attenuation equation is valid when the photon energy spectrum remains unaffected by passage through the material. This condition exists when a collimator is used in front of the radiation detector to exclude scattered radiation (i.e., the so-called narrow-beam geometry). Here, all scattered photons are assumed to be absorbed. In conventional radiographs, scattered photons reach the recording medium; this condition is known as broad-beam geometry. The broad-beam attenuation coefficient is smaller than the narrow-beam attenuation coefficient and depends on the photon energy spectrum and the penetration depth. Published data on attenuation coefficients are based on the narrow-beam geometry. When estimating radiographic exposures, narrow-beam data may result in inaccuracy, especially with low-energy X-rays and thick test pieces.

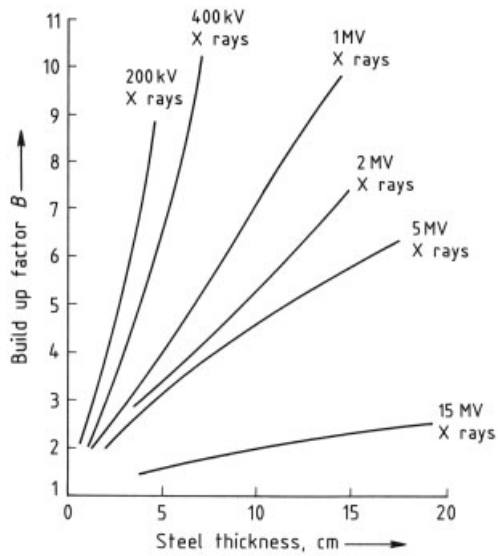

Figure 2. Buildup factor as a function of thickness

The radiation intensity I_B in the broad-beam geometry is expressed as

$$I_B = I_0 B \exp(-\mu_a x)$$

where B is the buildup factor, which is equal to the ratio of the total intensity (direct beam plus

scattered beam) to the direct-beam intensity. The buildup factor increases with increasing thickness of the test piece, and the rate of increase of B is higher for lower energy radiation (see Fig. 2 for B in steel).

When X-rays are generated at the target of an X-ray tube, the energy spectrum contains the low-energy portions. As the beam passes through the tube window, higher proportions of low-energy X-rays are absorbed, hardening the spectrum. This spectrum change is quantified by equivalent filtration with a thickness of aluminum and is called inherent filtration. When harder (higher average X-ray energy) radiation is required, a filter is inserted into the beam path.

After electromagnetic radiation is emitted from the source, the beam spreads as it travels. As the distance from the source increases, the same radiation intensity is distributed over a larger area and the intensity varies inversely with the square of the distance. If the intensity is I_1 at a distance x_1, the intensity I_2 at x_2 is

$$I_2 = I_1 \, (x_1/x_2)^2$$

which is called the inverse-square law.

2.2.3. Imaging

Radiographic images are formed in analogy to geometric optics because of straight-line propagation of X and gamma rays. A point source of radiation is generally used. However, the X- and gamma-ray sources are finite in size, typically ranging from 0.5 to 5 mm. When an image (or shadow) of an object at the source-to-object distance x_o is formed at the source-to-image distance x_i, the image is enlarged by a factor x_i/x_o. This enlargement factor is nearly unity in conventional radiography because the film is placed immediately behind the object or test piece. By using a microfocus X-ray tube with a source size of 1 – 10 µm, a ten- to hundredfold magnification can be realized in microradiography (X-ray microscopy) applications. Here, a test piece is placed near the source and the film is further away.

Image distortion occurs when a two-dimensional object is imaged on a nonparallel image plane. A three-dimensional object always gives a distorted image because the enlargement factor is not constant over the object.

The finite spot size of X- and gamma-ray sources produces regions of partial shadow at the edges of an object. These are known as penumbra and result in a geometric unsharpness U_g, given by

$$U_g = F_s \, (x_i - x_o) \, / x_o$$

where F_s is the source spot size (see Fig. 3). Geometric unsharpness and image distortion are reduced by using a larger source-to-object distance x_o. Increasing x_o, however, reduces the radiation intensity, and the exposure time must therefore be increased.

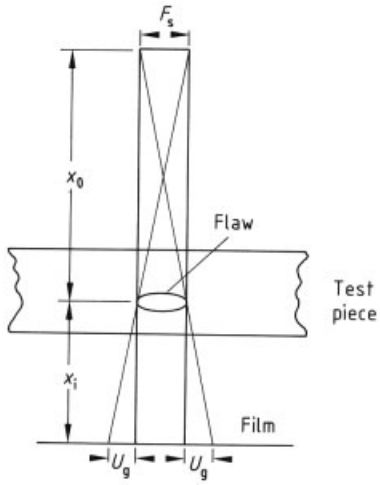

Figure 3. Geometrical unsharpness in radiography

2.2.4. Recording Images

The most common method of recording radiographic images is the use of X-ray films. Films make a permanent record of images of large areas. These have a thin emulsion coating on both sides of a film base or a thin clear plastic support. The emulsion is typically 25 µm thick and contains a microscopic suspension of silver halide grains. Both X-ray and gamma-ray energy cause a latent image to form. A visible image is revealed by a developer that converts the exposed silver halide to metallic silver. The image is then fixed by washing off unexposed halide grains and drying. Most X-ray films are for direct exposure to X-rays. Some are used with fluorescent screens and are more sensitive to visible light emitted by the screens.

The exposure of a film causes an inherent unsharpness, which results from the scattering of secondary electrons in the film emulsion. Typical values of inherent unsharpness are 0.01 – 0.05 mm for 100 – 250-kV X-rays.

Darkening of a film is measured as density D by transmitting light through it. The density is defined as

$$D = \log (I_0/I_t)$$

where I_0 is the intensity of the incident light and I_t is that of transmitted light. Only 1 % of light is transmitted at a density of 2, and 0.01 % at $D = 4$. Since low values of D give low sensitivity and high values make viewing difficult, standard procedures restrict D values. Valid density ranges for single and composite viewings are 1.5 – 4 and 2 – 4, respectively. For $D > 2$, a high-intensity illuminator is required. The density of an exposed film is dictated by exposure, film characteristics, and film development conditions. Exposure is the product of the radiation intensity and the time of exposure. In radiography, relative exposure is generally used, with a certain exposure condition including X-ray source, energy, filters, source-to-film distance, film type, and developing condition, being set as reference.

The characteristic curve, also known as the H and D curve after Hurter and Driffield, expresses the relation between the relative exposure applied to a film and the resulting density. The relative exposure is plotted on a logarithmic scale. Figure 4 shows examples of such curves, which are useful for determining suitable exposure conditions and comparing the speed and contrast of various film types. Speed describes how fast a particular density can be produced. In Figure 4, type 1 has the lowest speed and type 4 the highest. High-speed film tends to have coarse grains, whereas low-speed film has fine grains. Such graininess results from clustering of individual silver halide grains. Film contrast, also called film gradient, is the difference in density due to a change in exposure and is expressed by the slope G of the characteristic curve. A higher contrast gives a larger difference in density for the same variation in exposure, allowing better resolving of details, but the range of usable exposure conditions is narrow. For type 1 film in Figure 4, the log of relative exposure must be in the range 4.75 – 5.15 to obtain high contrast and densities of 2 – 4. This film has a narrow latitude.

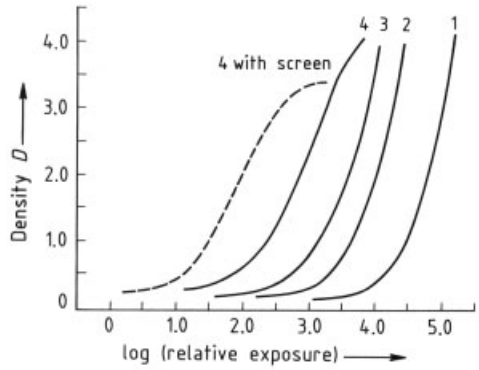

Figure 4. Characteristic curves for films

For a given range of densities on the characteristic curve, the average slope G_{av} can be defined. A change in density ΔD for a given exposure time is related as

$$\Delta D = -0.43 G_{av} \mu_a \Delta x / B$$

where μ_a is the attenuation coefficient for the radiation used, Δx is the change in thickness, and B is the buildup factor. A change in D of more than 0.006 is required for a perceptible density change, so the minimum thickness change that can be detected radiographically is expressed as

$$-(2.3 \times 0.006)\, B/\mu_a G_{av}$$

That is, higher attenuation coefficients, obtained with low-energy radiation, and steeper slopes of the characteristic curve increase radiographic sensitivity. The buildup factor, on the other hand, decreases the sensitivity due to non-image-forming scattered radiation. Both B and μ_a increase with decreasing radiation energy, but μ_a is affected much more than B.

The shape of characteristic curves is affected by the development time: film contrast increases with development time, which also gives some increase in speed. The energy spectrum of the radiation also affects the speed. The characteristic curve is shifted to the higher exposure side as the radiation energy increases from less than 100 kV to several hundred kV or more.

For a given combination of radiation source, source-to-film distance, material to be radiographed, type of film, and film development technique, a radiographic exposure chart for a

particular density can be prepared. When the thickness of a test piece varies within the area of coverage, a latitude chart may be used to help determine the correct exposure. The latitude chart is similar to the exposure chart except that a band corresponding to a range of densities is given (Fig. 5).

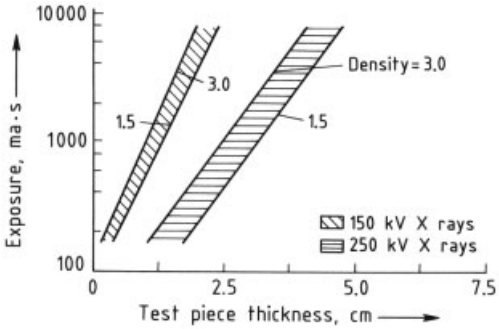

Figure 5. Latitude chart for type 2 film

2.2.5. Screens

Lead foil screens are commonly used in direct contact with films to produce an intensifying action at as low as 90 kV, depending on the thickness of the material and the foil. The intensification arises through emission of secondary electrons by the lead upon X-ray photon bombardment. Higher energy radiation gives greater intensification. A screen in front of the film also filters out scattered radiation arising from the test piece and improves radiographic quality. Lead screens, 0.13 mm thick, are generally used in front and back for radiography of 6-mm-thick (or more) steel with sources of more than 135 kV. For higher energy radiation (e.g., ^{60}Co) 0.25-mm-thick lead screens are used. The rear screen also protects the film from backscattered radiation. For this purpose, a lead sheet is used in the back of the film holder as well. Other screen materials include copper, gold, tantalum, and lead oxide.

Fluorescent screens are used only when the exposure would be extremely long without them. They are employed in combination with film optimized for high sensitivity to emitted light. The screens are coated with phosphors such as $CaWO_4$ and ZnCdS and can produce 2 – 25-fold intensification. However, light scattering within the phosphor gives rise to screen unsharpness, reducing image definition on the film. Typical values of screen unsharpness for commercially available screens are 0.1 – 0.4 mm. Another drawback of fluorescent screens is the appearance of screen mottle on the finished radiograph. This arises from statistical variations in fluorescence and appears as a larger scale graininess.

2.2.6. Radiographic Quality

Image quality is influenced by many variables that contribute to radiographic contrast and image clarity. Superior contrast is obtained by a large change in thickness Δx, lower radiation energy, less scattered radiation, exposure in the high-film-gradient region, adequate film development, and use of films with high gradients. Good image clarity results from low total unsharpness, which is defined as the square-root of the sum of the squares of geometrical, inherent, and screen unsharpnesses. Motion unsharpness also arises due to movement of the test piece. Typical values of total unsharpness with good film techniques for 100-kV X-rays are 0.05 mm; 200-kV X-rays, 0.10 mm; 400-kV X-rays, 0.20 mm; iridium-192 gamma rays, 0.25 mm; and cobalt-60 gamma rays, 0.50 mm.

Practical indicators of image quality are known as penetrameters or image quality indicators. One ASTM design [9, E 1025] is based on rectangular or round sheets with thickness T of 0.13 – 1.3 mm, having holes of size 1 T, 2 T, and 4 T (1 T and 2 T for round penetrameters). A penetrameter whose thickness corresponds to 2 % of that of the test piece is generally used. Another type of penetrameter [9, E 477] employs a set of six wires of different sizes. The penetrameter material is the same as that of the test piece. The penetrameter is placed directly on the test piece, facing the source and remote from the area of inspection. Image quality is usually expressed as the size of the smallest penetrameter feature, such as hole size or wire diameter, that is clearly visible in the processed radiograph. The minimum acceptable quality level corresponds to the visibility of a 2 T hole with a 2 % thickness penetrameter. This quality level is expressed as 2-2 T. Equivalent penetrameter sensitivity (in percent) is defined as $100\sqrt{Th}/x\sqrt{2}$, where x is the test

piece thickness, T is the penetrameter thickness, and h is the hole size. Other measures of equivalent sensitivity are also used.

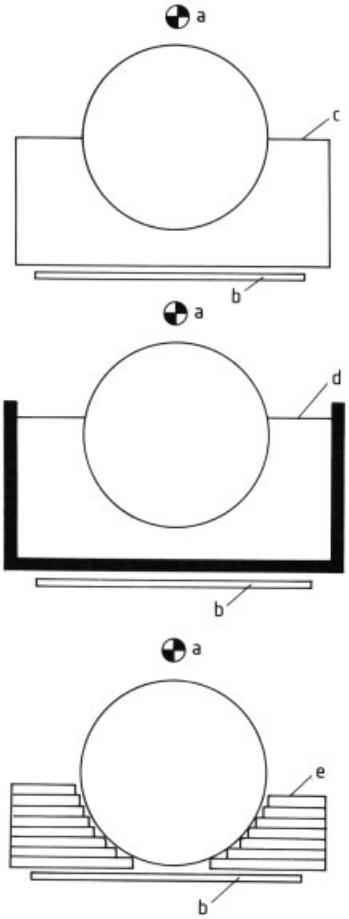

Figure 6. Three methods of equalizing radiographic density
a) Radiation source; b) Film; c) Solid cradle; d) Liquid absorber; e) Shims

2.2.7. Radiographic Practices

The positions of the radiation source, the test piece, and the film are chosen to best reveal expected flaws with minimum geometrical distortion. In the case of cracklike flaws, the X- or gamma-ray beam must be oriented parallel to the flaw. The orientation matters little for voidlike flaws, but the size or shape of the image will change unless the beam passes normal to the plane of the flaw.

Scattered radiation decreases radiographic sensitivity. Scattering from wall reflections and internal holes or openings can be reduced by using lead masks that cover the holes or openings and by surrounding the test piece with metallic shot or a liquid absorber. Backscatter from objects behind a film (away from the source) is eliminated by a lead screen or sheet. A collimator is placed near the source and reduces the beam size to cover only the test piece, thus eliminating spurious scattering.

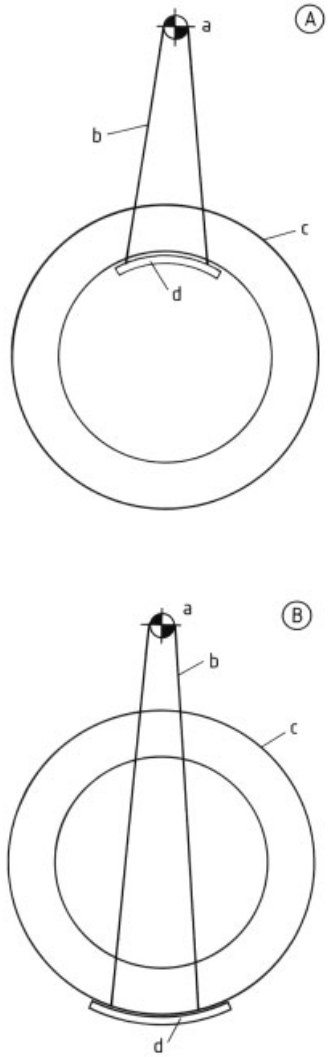

Figure 7. Radiographic inspection
A) Single-wall and B) Double-wall radiographic technique
a) Radiation source; b) Beam; c) Pipe; d) Film

Flat test pieces can best be radiographed by a normally incident beam with films below the

test pieces. Curved plates can be inspected similarly, but minimum distortion is achieved with the source at the center of the radius of curvature. Inspection of a cylindrical test piece requires section-equalizing techniques with a solid cradle, liquid absorber, or shim stocks (Fig. 6). Double-film techniques with films of different speed, and multiple exposures with different radiation energies can be employed instead of these section-equalizing techniques. Overexposure of the outer portion of the radiographic image can be corrected either by exposing a film of lower sensitivity or by using lower energy radiation. Tubular sections are inspected by single- or double-wall techniques (Fig. 7). Single-wall techniques use an external radiation source and the film inside the cavity. Alternatively, the source is placed inside the cavity and the film on the outside. For the latter, a panoramic X-ray source that emits radiation 360° around its axis may be used. Single-wall techniques usually take a single image through a wall or weld and provide good radiographic sensitivity. Double-wall, double-image techniques are used when single-wall techniques are difficult to employ, e.g., for small cross sections (< 9 cm). These produce radiographs on which the images of both walls are recorded and are also applied to the inspection of diagonally opposed corner welds in rectangular box sections. Circumferential butt welds can be inspected with the offset or corona techniques (Fig. 8). For smaller diameters, two images can be exposed, but the single-image technique gives a better radiograph for large-diameter pipes.

Inspection of complex shapes requires multiple exposures with different viewing directions. Combinations of the above-mentioned techniques are chosen to reduce the image to a simple shape for easy interpretation. Welds are usually inspected to reveal flaws in weld metals and in heat-affected zones. A beam direction normal to the surface of the weld is used. For T-joints or lap joints, the beam direction is inclined to both sides of the weld joint. When two adjacent welds are made, two beam directions are usually required, but the images may be recorded on a single film.

2.2.8. Radiographic Interpretation

The radiographic appearance of the types of flaw usually found in castings and weldments has been compiled in volumes of reference radiographs, which are graded according to flaw size and severity. The ASTM has produced 12 volumes of reference radiographs [9]. Radiographic acceptance criteria are defined from these reference radiographs, permitting acceptable limits of flaw types, sizes, and densities. Linear-type flaws, such as cracks and incomplete fusion, are prohibited regardless of size in most criteria [8].

Of the service-induced flaws, corrosion in a flat or tubular section can be detected because the loss of material results in dark areas in the radiograph. In contrast, tight cracks from stress corrosion and hydrogen effects are difficult to detect. Similarly, short fatigue cracks are hard to locate.

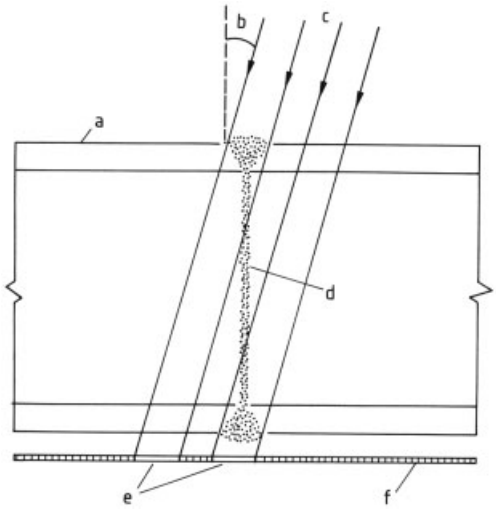

Figure 8. Corona (offset) radiographic technique
a) Pipe; b) Offset angle; c) Radiation beam; d) Weld; e) Images of weld; f) Film

2.2.9. Other Techniques

Computerized tomography, developed for medical applications, can also be used in inspection of lightweight composites. Conventional tomography equipment is expensive and as yet impractical for routine commercial use. Film-based tomography may, however, become a useful method in the future [21].

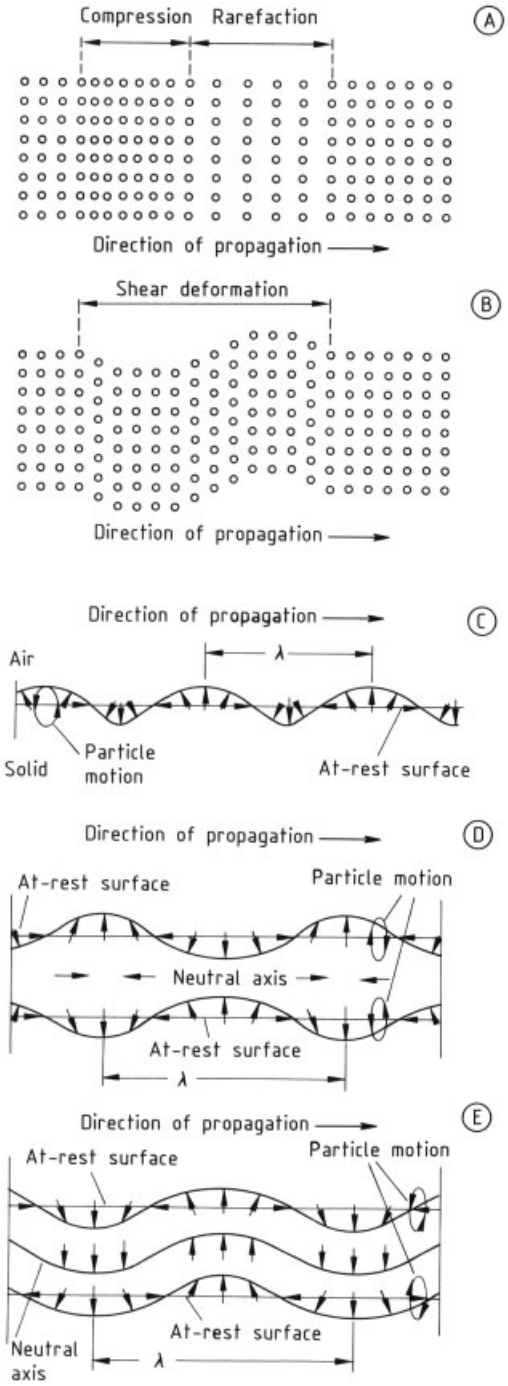

Figure 9. Ultrasonic waves
A) Compressive wave; B) Shear wave; C) Surface wave; D) Symmetrical plate wave; E) Asymmetrical plate wave

Real-time radiographic inspection techniques have advanced substantially within the past decade because of improvements in electronic image processing. However, they are five to ten times less sensitive to cracks and are used mainly in electronic component inspection and airport baggage screening [9, E 1000].

Microfocus X-ray tubes can readily provide a magnification of two to ten times because of their small spot size (ca. 10 µm). Although their low X-ray output requires long exposure times or use of an electronic image intensifier (combining the real-time fluoroscopic and image-processing equipment), a microfocus unit allows the inspection of small flaws in ceramics and composites.

Inspection of thick sections requires high-energy radiation sources. Steel sections up to 5 cm can be radiographed with medium energy X-rays (250 – 400 kV) or iridium-192, and up to 10 cm, with 1-MV X-rays or cobalt-60. Higher energy radiation improves radiographic sensitivity for thick sections, but requires specialized equipment.

2.2.10. Radiation Safety

Exposure to radiation is harmful to health and must be strictly controlled. Radiation safety is practiced by adhering to established procedures of handling radiation equipment and radioactive materials, by incorporating adequate shielding design and safety interlock systems, by thoroughly educating personnel about safe practices, by monitoring radiation levels of the working environment, and by enforcing the limit on allowable exposure. Proper licensing requirements must be maintained.

2.3. Ultrasonics

Ultrasonic testing uses high-frequency mechanical (elastic) waves in the range of 20 kHz to 50 MHz. Elastic waves are reflected and refracted at interfaces where elastic moduli and density are discontinuous. Cracks, voids, nonmetallic inclusions, and back surfaces reflect incident waves. By sending a pulse and receiving its reflection, the presence and location of flaws can be defined in analogy to radar. This is the pulse-echo technique. When laminations, disbonding, and other flaws are present, the intensity of transmitted waves is attenuated.

By measuring the intensity on the back surface, the through-transmission technique detects flaws via attenuation. When the material under inspection is uniform, the time of transit of elastic waves through the material can be converted to the thickness by determining the wave velocity on a calibration sample. This time-of-flight technique is used in thickness meters. Time-of-flight and attenuation measurements on samples of known thickness are used to characterize material properties, such as grain size of metals or fiber and void content in composites.

The ability to locate and measure internal flaws, the ability to internally inspect three-dimensional objects from a single surface, and nonhazardous operation are advantages of ultrasonic testing. Disadvantages include the complexity in interpreting some results and the difficulty of inspecting parts with rough surfaces, irregular or small shapes, or near surface flaws. Operator training requirements are also higher than for most other nondestructive testing methods; see [9], [22] for further details.

2.3.1. Waves

Elastic waves travel with characteristic velocity in a gas, liquid, or solid. In gases and liquids, only compressive waves exist; i.e., the medium is alternately compressed or rarefied. Waves travel parallel to the direction of compression and rarefaction; hence, they are also called longitudinal waves. In solids, compressive waves and shear waves are present. In shear or transverse waves, elastic displacement occurs perpendicular to the direction of wave propagation. These waves are shown schematically in Figure 9 A and B. Here, the distance between two points of the same phase is the wavelength λ_u.

Most ultrasonic testing utilizes frequencies of 1 – 25 MHz. Higher frequencies allow better spatial resolution due to the shorter wavelength. However, more attenuation occurs above 10 MHz in steels as the wavelength approaches the grain size. In highly attenuating fiber-reinforced composites, frequencies of 1 MHz or lower are commonly used. In most structural steels and wrought alloys, attenuation is low, and ultrasonic testing has good penetrating power to depths of 1 – 10 m.

The velocity V_c of plane compressional waves in an isotropic medium is given by

$$V_c = \sqrt{(\lambda + 2\mu)/\varrho}$$

where λ and μ are Lamé constants and ϱ is mass density. Similarly, the shear wave velocity V_s is

$$V_s = \sqrt{\mu/\varrho}$$

Wavelength is equal to wave velocity V divided by frequency f. For steels, $V_c \approx 6$ mm/μs, and for $f = 1$ MHz, $\lambda_u \approx 6$ mm; for water, $V_c \approx 1.5$ mm/μs and $\lambda_u \approx 1.5$ mm. Table 1 lists typical values of V for various materials.

Table 1. Typical ultrasonic velocities

Material	V_c, mm/μs	V_s, mm/μs
Carbon steels	5.94	3.24
304 stainless steel	5.64	3.07
Cast iron	3.5 – 5.6	2.2 – 3.2
Inconel	5.82	3.02
Brass (70/30)	3.83	2.05
Titanium (pure)	6.10	3.12
Glass	5.77	3.43
Poly(methyl methacrylate) (Plexiglas)	2.67	1.12

By measuring wave velocities, the elastic constants, Young's modulus E, and Poisson's ratio ν, can be found:

$$E = \frac{\varrho V_s^2 \left[3\,(V_c/V_s)^2 - 4\right]}{(V_c/V_s)^2 - 1}$$

$$\nu = \frac{(V_c/V_s)^2 - 2}{2\,(V_c/V_s)^2 - 2}$$

Surface (or Rayleigh) waves are also used in ultrasonic testing. These travel along the surface with approximately 90 % of the shear wave velocity. The wave energy is concentrated within about one wavelength of the surface, as shown in Figure 7 C. Rayleigh waves follow curved surfaces and corners, but are reflected when the ratio of the radius of curvature r to the wavelength λ_u is less than 1.6. About half the wave energy is reflected for $r/\lambda_u < 0.6$. In comparison to bulk waves (compressive and shear), Rayleigh waves are attenuated less and are well suited to inspecting parts that have complex contours.

Lamb waves (plate waves) propagate in sheets and plates in which the wavelength and thickness are of the same order of magnitude.

Symmetric and asymmetric modes of propagation exist (Fig. 7 D and E). These waves have different phase and group velocities, each dependent on frequency and thickness. Until recently, the application of Lamb waves to ultrasonic testing has been limited due to their complex propagation characteristics.

When ultrasonic waves are normally incident on a plane interface of two materials, 1 and 2, part of the wave energy is transmitted through the interface, and the remainder is reflected. The ratio of transmitted energy to incident wave energy, known as the transmission coefficient T_i, is expressed in terms of the acoustic impedance Z, the product of mass density and wave velocity ($Z = \varrho V_c$):

$$T_i = 4Z_1Z_2/(Z_1+Z_2)^2$$
$$= 4\varrho_1\varrho_2V_{c1}V_{c2}/(\varrho_1V_{c1}+\varrho_2V_{c2})^2$$

The reflection coefficient $R_i = 1 - T_i$. At an interface of water and steel, $T_i = 0.12$ and $R_i = 0.88$; that is, nearly 90 % of the wave energy is reflected.

When a plane wave is obliquely incident on an interface, both compressional and shear waves are reflected or refracted regardless of the nature of the incident wave. This is known as mode conversion, and the angles of reflection and refraction are defined by Snell's law (see Fig. 10):

$$\frac{\sin\theta_{c1}}{V_{c1}} = \frac{\sin\theta_{s1}}{V_{s1}} = \frac{\sin\theta_{c2}}{V_{c2}} = \frac{\sin\theta_{s2}}{V_{s2}}$$

For water (1) and steel (2) with an incident angle θ_{c1} of less than 45°, compressional waves in water are reflected at θ_{c1}; refracted compressional and shear waves exit from the interface at θ_{c2} and θ_{s2}. Note that no shear wave exists in liquids. The angle

$$\theta_{c1} = \sin^{-1}(V_{c1}/V_{c2}) = 14.5°$$

is called the first critical angle because θ_{c2} becomes 90° and the compressional waves in steel no longer exist for higher values of θ_{c1}. The second critical angle is at $\sin^{-1}(V_{c1}/V_{s2}) = 27.4°$, above which no shear wave is observed in steel and total reflection occurs. Between the first and second critical angles, only shear waves are refracted in steel. This is advantageous because interpretation is easier with only one wave mode, and higher intensity waves are produced for shear waves refracted at 35 – 60°. Partly due to geometrical considerations, 45° and 60° shear wave transmission paths are most commonly used. For thin sections, 70° shear waves are also employed.

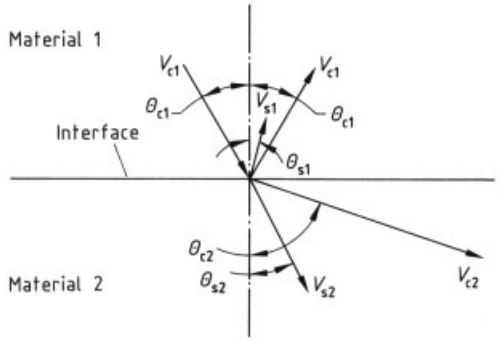

Figure 10. Mode conversion of ultrasonic waves

Useful depth of inspection is determined by the attenuation of ultrasonic waves. The acoustic pressure decreases exponentially with depth, as in the case of X-rays. However, the attenuation of ultrasonic waves is usually expressed in decibels per meter (dB/m). General ranges of inspection are up to 1 – 10 m in steels, in nickel, titanium and magnesium alloys with fine microstructures, and in all aluminum alloys; these have attenuation rates of 0.1 – 1 dB/m. Most cast metals have a lower inspection range of 0.1 – 1 m. Stainless steel weldments and fiber-reinforced composites have high attenuation, and the inspection range extends to less than 0.1 m.

In highly attenuating materials such as cast nonferrous materials and fiber-reinforced composites, the wave velocities are frequency dependent. This dispersion affects waveforms as they propagate because higher frequency contents are attenuated.

2.3.2. Generation and Detection

Ultrasonic waves are generated and detected by using a transducer element that converts electrical signals to mechanical ones and vice versa [23]. A thin disk of piezoelectric crystal or polarized ceramic is used. Quartz, lithium sulfate, and lithium niobate are used as single-crystal elements, whereas lead zirconate – titanate is the most common ceramic element. The transducer

element is enclosed in a protective casing with a thin ceramic or plastic faceplate and an attenuative backing such as a tungsten – epoxy mixture (see Fig. 11 A). The backing is needed to dampen the resonance of the transducer element and to broaden the frequency response. The casing also acts as an electrical shield. A thicker faceplate (a delay tip) may be used to separate the front-surface signal from the large excitation pulse. The frequency characteristics of the packaged transducer or search unit are dictated primarily by the thickness resonance of the transducer element, which occurs when the thickness equals one-half of the wavelength, causing constructive interference of waves reflected at the surface.

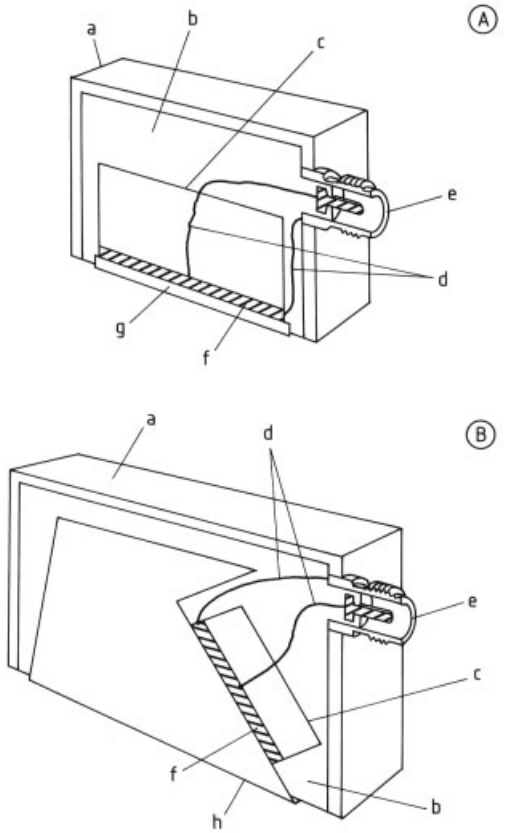

Figure 11. Typical ultrasonic search units
A) Straight-beam unit; B) Angle-beam unit
a) Case; b) Epoxy potting; c) Tungsten-loaded backing; d) Electrical connections; e) Cable connector; f) Transducer element; g) Faceplate; h) Plastic wedge

To generate ultrasonic pulses with a search unit, an electronic pulse generator is used to produce a train of short pulses or tone bursts (up to ten oscillations). These are repeated 50 to 2000 times per second. A short pulse contains a wide spectrum of frequencies, while the frequency spectrum of a tone burst is centered at its base frequency. When the search unit is excited with a pulse (or tone burst), the frequency spectrum of the ultrasonic wave output is the product of the frequency spectra of the input pulse (or tone burst) and the search unit.

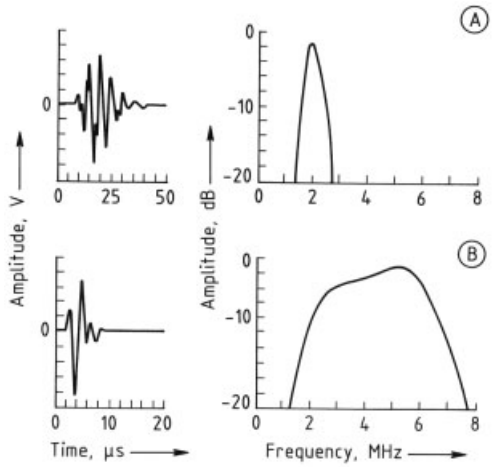

Figure 12. Ultrasonic waveforms and their frequency spectra
A) Narrow band; B) Broad band

The waveform of the output can be obtained from the frequency spectrum by inverse Fourier transform. This waveform W_0 can also be expressed as the convolution of the input waveform W_i and the input response W_s:

$$W_0\,(t) = \int W_i\,(t-\tau)\,W_\mathrm{s}\,(\tau)\,\mathrm{d}\tau$$

where W_s is the characteristic waveform for an infinitely short pulse (delta-function shape). Typical waveforms and the frequency spectra of narrow-band and broad-band search units are shown in Figure 12. High peak voltages (300 – 500 V) are needed for a short pulse because only part of the pulse energy is used. This often leads to the failure of thin, high-frequency transducer elements. With a tone burst, the base frequency must be matched to the nominal resonance frequency of the search unit. Since the tone burst is limited in bandwidth, the peak voltage can be much lower to obtain the same ultrasonic energy output. Alternatively, a much

stronger wave can be generated with a tone burst than with a short pulse for a given applied voltage.

The sound field immediately in front of the faceplate is basically a plane wave, but the intensity varies depending on the position and distance from the search unit because of diffraction effects. This region is known as the near field, which extends to a distance N_x given by

$$N_x = D_t^2 / 4\lambda_u$$

where D_t is the diameter of the transducer element. For flaw detection, the near field is difficult to use because of the intensity variation (near-field effect). Near-field effects in the material under inspection can be eliminated by placing a delay tip between it and the search unit, whereby the near-field effect is confined to the delay tip. The region beyond N_x is known as the far field and a spherical wave propagates as shown in Figure 13. No diffraction fringe occurs, and the wave intensity decreases with the square of the distance. The spread φ of the far-field sound beam is dictated by λ_u/D_t and is given by

$$\varphi = 2\sin^{-1}(1.22\lambda_u/D_t)$$

For example, $\varphi = 44°$ in steel at 1 MHz with $D_t = 19$ mm. For the same D_t at 5 MHz, φ decreases to 8.7°. A smaller value of D_t allows flaws to be more accurately located at short distances, but this advantage is lost as the beam spreads with increased penetration. A higher frequency beam spreads less for a given value of D_t or allows the use of a smaller D_t.

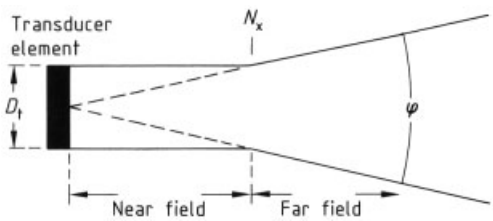

Figure 13. Near- and far-field regions

For special laboratory studies, ultrasonic pulses can be generated by means of the thermoelastic effect with a strong laser beam. Electromagnetic transducers allow noncontact generation and detection. They are suited for automated inspection of pipes and plates at elevated temperature or with rough surfaces, for which noncontact generation is desirable for ease of use. Commercially available units can be used up to 3 MHz. A strong magnetic field is required, and a special powering unit must be used. Nonmagnetic test pieces lower the efficiency. Limits in the frequency range of operation and lower signal-to-noise ratios are some of the disadvantages.

By using the principle of refraction, ultrasonic search units can be designed to generate angle beams. A plastic wedge is housed with a transducer element as shown in Figure 11 B. The direction of the beam is valid for a specified material. Surface wave search units are also made with the same design concept.

In the pulse-echo mode, the same search unit is used for generation and detection of ultrasonic waves. Dual-element search units are also used which house separate transducer elements for detection and generation. Two separate search units are employed for the through-transmission mode, in which one is dedicated for receiving the ultrasonic waves. Received ultrasonic waves produce low-level electrical signals. Electronic amplifiers with low noise levels (even for a wide bandwidth of 0.1 – 50 MHz) are used to provide good signal-to-noise ratios. When a single search unit is used for transmission and receiving, an input protection circuit is required and the input amplifier must be capable of rapid recovery from an overload due to the excitation pulse.

The sound beam can be focused by means of an acoustic lens attached to the front surface of the transducer. This requires a water path between the lens and the test piece. Transducers can be spherically focused to a spot or cylindrically to a line. For the latter, a rectangular transducer element is used. Spherical focusing achieves the same for pipe and tubing inspection. Both can increase near-surface resolution without increasing the transducer frequency. Focusing also reduces the effects of surface roughness and contour.

Focal distance is given in water path length. Path length in a solid is converted to an equivalent path length in water by multiplying by the ratio of sound velocities. Focal distance should then be equal to (water path) + (solid path) × (sound velocity in solid) ÷ (sound velocity in

water). Search units of various focal distances (typically, 1 – 25-cm water path) are available.

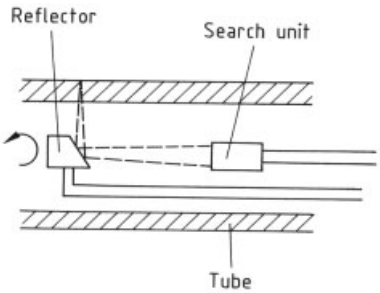

Figure 14. Ultrasonic inspection of pipes

2.3.3. Inspection Methods

When a search unit is placed directly over the surface of a part under test in contact inspection, a thin layer of liquid couplant is used to provide a low loss path for wave transmission. Water, oils, glycerol, greases, and resins are used as couplants. Certain soft rubbers have been specially formulated to provide dry coupling of a search unit. In immersion inspection, the part under inspection and the search unit are placed in a tank of water. Here, a water path exists in place of a couplant layer. A search unit is usually fitted with a waterproof connector and attached to a manipulator. It is usually set manually but is increasingly used under computer control.

In conventional immersion inspection, the search unit and the test piece are placed in a water-filled tank. The search unit is connected via a manipulator and an extension tube to an electromechanical scanning device (or a probe manipulator bridge). The basic scanning device provides X – Y movements of prearranged scanning patterns. Advanced units add computer control that adapts to the shape of a test piece. Turntables and roller drives allow efficient scanning of round disks and cylindrical test pieces. A rotating reflector inside a tube can be used to reflect the ultrasonic beam of an immersion search unit toward the tube wall, allowing the inspection of thickness changes or flaws by means of normal or shear beams (Fig. 14) [24]. Scanning is synchronized with ultrasonic data acquisition, and flaw indications are displayed at corresponding locations. An X – Y recorder has been used for such a display, but recent equipment can store the scan position and ultrasonic data from 10^7 – 10^8 locations in digital form. Real-time or post-test analysis of the stored data provides detailed color displays.

Wheel-type search units consist of a stationary transducer element, a liquid path inside a rubber tire, and a rolling wheel to provide continuous contact with a test piece. The transducer element can be placed for straight-beam or angle-beam inspection. The wheel unit can be stationary and a test piece moved past it, or it may be moved over a fixed test piece.

A search unit can be fitted with a squirter, a nozzle through which water streams out under pressure, which provides a column of water for a sound beam to reach the test piece. The squirter technique can be combined with robotics. This combination eliminates the size limitation of immersion testing, can be adapted to complex contours, and is suited for automatic operation. Another method, known as the bubbler method, uses a small tank with overflowing water. A search unit is placed in this tank, and a test piece is moved over it.

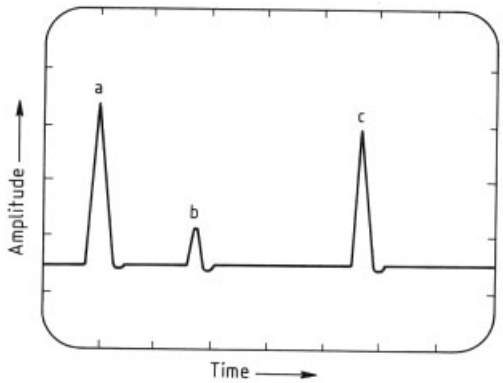

Figure 15. A-scan display for pulse-echo technique
a) Front reflection; b) Flaw echo; c) Back reflection

2.3.4. Display Methods

Results of ultrasonic testing are commonly presented in A-scan, B-scan, and C-scan and, more recently, with reconstructed images. The A-scan presentation is a plot of amplitude versus time, as shown in Figure 15. Front reflection, back reflection, and a flaw echo are indicated on the time axis, which corresponds to depth. The height of the flaw echo is related to the size of the flaw. In

the absence of a flaw, an A-scan over a longer period also shows the ultrasonic attenuation of the test piece. Multiple back reflections, whose heights decrease exponentially, are detected.

The B-scan presentation is used to display the depth and length of a flaw. The transducer is moved along the front surface, as shown in Figure 16. When a flaw echo exceeds a predetermined threshold, an indication is recorded on the display along with indications of the front- and back-surface signals. Convenient visual observation is possible with manual scanning and a storage oscilloscope. Mechanized scanning allows more accurate determination of the position of a flaw.

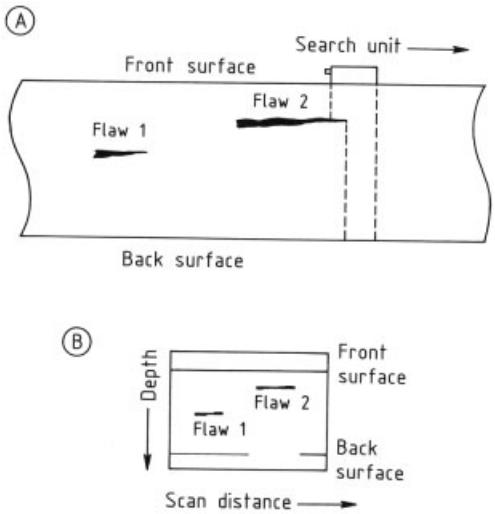

Figure 16. A) Pulse-echo technique; B) B-scan display

The C-scan presentation is most useful in identifying defective areas in a plane view of a test piece. The transducer is moved back and forth over the front surface (Fig. 17). When a flaw echo above a threshold is present, it is recorded on an X – Y recorder, which is also scanned synchronously. The use of several electronic timing gates, each sensing the flaw echo over a specified time interval, enables scanning for flaws at predetermined depth ranges. Newer inspection systems are usually equipped with a computer that controls scanning, ultrasonic data acquisition (in digitized form), and display. These allow data storage, display of C-scan records at different depths, and a variety of image processing. For example, flaws at different depths can be displayed in different colors.

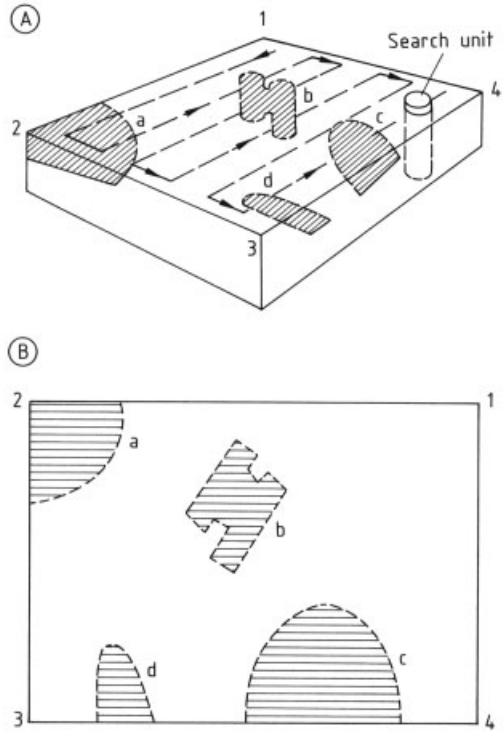

Figure 17. A) Pulse-echo technique; B) C-scan display

In specialized inspection systems, A-scan records are accumulated along with location data during a more complex mechanical scanning of various shapes. For tubular products, helical scanning is used, either internally or externally. For pressure vessels, remote-controlled crawlers are used. Industrial robots are used on contoured surfaces. Three-dimensional images of flawed regions can be reconstructed from these data, and different perspectives of flaw images can be obtained. A reconstructed view of flaws in a weld is shown in Figure 18. In both reconstructed imaging and C-scan displays, the use of color enhances visual perception.

2.3.5. Interpretation

Echo intensity in an A-scan display depends on the size, shape, and distance of the flaw. For disks of various sizes in water, the intensity of the echo varies as shown in Figure 19 [22]. The reduction in echo intensity results from attenuation and spreading of the beam. Compensation for this reduction is provided by a distance – amplitude

correction (DAC) circuit in most ultrasonic instruments. By comparing the echo intensities of flaws with those of flat-bottom holes in reference blocks, flaw sizes can be estimated. Another approach is to use the ratio of echo intensity to the back-surface reflection as an indicator of an unacceptable flaw. An echo intensity equal to 50 % of the back reflection is one such rejection criterion.

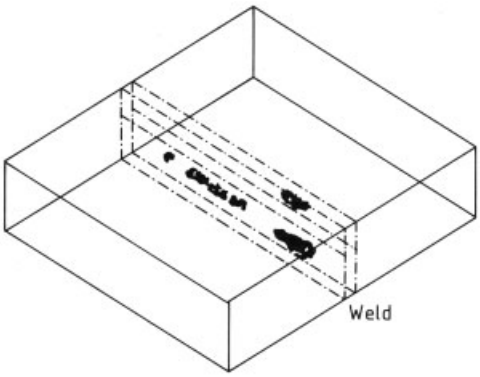

Figure 18. Reconstructed image of weld flaws

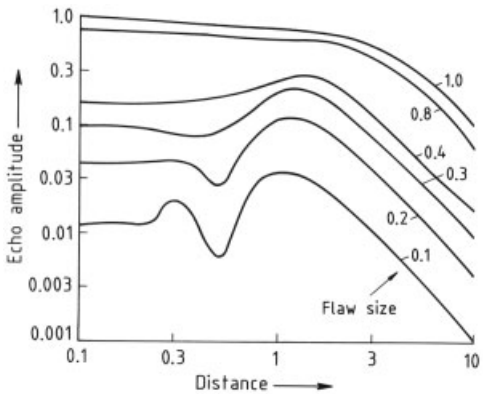

Figure 19. Distance – echo amplitude – flaw size diagram

The shape of the echo depends on the shape, orientation, and sound-reflecting characteristics of an interface. Smooth metal – air interfaces that are normal to the beam produce sharp echoes. Curved or rough interfaces (pores, cracks, laminations) produce broadened echoes. Changing the beam orientation and the position of the search unit often provides clues to flaw shape and orientation. Loss of back reflection is another factor in flaw size estimation. This is due to the reflection of a sound beam by a flaw, which may not produce a flaw echo if the reflected beam is directed away from the search unit. In this case, the loss of back reflection indicates the presence of a flaw. Scattering from slags, inclusions, and large grain structures also results in the loss of back reflection.

The large grain structures of anisotropic materials (e.g., stainless steel or brass) produce beam scattering, which is indicated as random echoes between the front and back reflections. These are called grass and make the discrimination of small flaw echoes difficult. Spurious reflections are also produced by fillets. A choice of suitable inspection position is important for complex geometries.

2.3.6. Inspection Standards

Ultrasonic inspection is performed under specified procedures established by national bodies, such as the ASTM, DIN, and JIS. They require the use of standard reference blocks, which may be of different sizes prepared from different alloys, containing holes, slots, or notches. Two types of standard blocks are widely used. One type is the area – amplitude or distance – amplitude block, including ASTM E 127 blocks. Each of the ASTM blocks is 51 mm in diameter and has a 19-mm-deep flat-bottom hole. The size of the holes varies from 0.40 to 6.4 mm in diameter, and the distance from the surface to the bottom of the hole varies from 3.2 to 152 mm. By changing the hole size at a given distance, these blocks can be used to relate the amplitude of the flaw signal to the area of a flaw, and to check the linearity and sensitivity of a pulse-echo inspection system. For a given hole size at various depths, distance – amplitude blocks can evaluate variations of echo amplitude with distance for straight-beam inspection in a given material. These blocks must be made from the same material as the test piece in order to estimate flaw sizes at various depths. However, a set of blocks can be used as reference blocks for performance calibration of a test system.

Another reference block is known as the IIW type. It is a nearly rectangular steel plate (25 × 100 × 360 mm) with a curved edge of radius 100 mm at one end. The block has a notch, a slot, and large and small holes (50 and 1.5 mm diameter), which are used together with the curved

edge to determine the sensitivity, propagation angle and beam spread of angle-beam search units. The IIW block can also be used for simple sensitivity evaluation of straight-beam units. A miniature version based on the same concept is also used.

2.4. Acoustic Emission

Acoustic emission (AE) refers to transient elastic waves generated from localized sources within a material or structure. Such transient waves originate at, for example, cracks, defective welds, and yielding regions in metals and at delaminations, fibers, and matrix fractures in composite materials. Sensitive ultrasonic sensors detect acoustic emission signals, from which the location and activity level of the source can be evaluated. Acoustic emission requires the application of a stimulus to a test piece. In pressure vessels and piping, hydrostatic pressurization is usually employed to activate acoustic emission sources. Acoustic emission has become the inspection method of choice for fiberglass-reinforced composite vessels and piping and, more recently, for metallic vessels, including industrial gas transport trailers and railroad tank cars; see [7, vol. 5], [25], [26].

2.4.1. Detection

In acoustic emission testing, signals in the ultrasonic frequency range of 30 kHz to 2 MHz are usually detected. Airborne noise interferes with AE measurements at lower frequencies, and signal attenuation makes the higher frequency range difficult to use. In composite testing, the 30 – 150-kHz range is common, and the 100-kHz to 1-MHz range is used for metals. Both burst-type (pulselike) and continuous-type (random noise) signals are generated. Burst-type emissions arise from distinct events involving the release of elastic energy, such as crack advances, fiber fracture, and delamination. Continuous-type emissions are produced by many overlapping events and are observed from plastic deformation of metals.

Sensors used for acoustic emission detection are 40 – 60 dB more sensitive than those used for ultrasonic testing. One type enhances the sensitivity by using the resonance of a transducer element, which is usually a piezoelectric disk. These are resonant or narrow-band sensors and are most commonly employed in acoustic emission testing. The second type has a broadened frequency response and uses a backing material behind a transducer element; its construction is essentially identical to that of ultrasonic sensors. For specialized high-fidelity capture of signal waveforms, a conical piezoelectric element with matched backing and a capacitive sensor can be used.

Acoustic emission activity is measured by the rate of burst-type emissions (AE event count rates) and by the averaged signal intensity of continuous-type emissions (e.g., root-mean-square voltages of amplified signals). The cumulative number or rate of oscillations that cross a predetermined threshold value (called AE counts or AE count rates), some measures of signal energy, the duration of burst-type signals, and the distribution of peak amplitudes are also employed.

2.4.2. Material Behavior

Plastic deformation of most structural alloys emits acoustic emission that reaches a maximum near the yield stress and diminishes with work hardening. The composition, microstructure, and heat treatment of alloys affect acoustic emission behavior. Cold-worked metals produce hardly any acoustic emission because AE phenomena are irreversible; therefore, acoustic emission is absent until previously applied stress levels are exceeded. Some alloys (e.g., austenitic stainless steels and 2219 aluminum alloy) have inherently low acoustic emission activity during plastic deformation. Stressing of various types of composite materials produces very high levels of acoustic emission. Matrix cracking is dominant at low stress. Glass-fiber fracture and delamination generate intense signals at higher stress.

Subcritical crack growth emits acoustic emission at increasing rates as the crack nears criticality. Some ductile cracks and fatigue crack growth have lower acoustic emission activity. Such cracks can still be located by detecting secondary AE sources, such as crack face interference and crushing of corrosion and oxida-

tion products. Decohesion and fracture of nonmetallic inclusions are also prominent sources of acoustic emission.

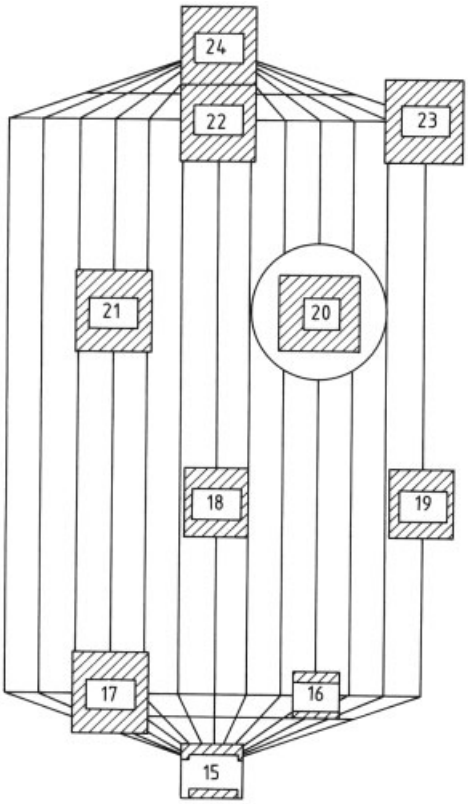

Figure 20. Zone location results from pressurization test of a hydrocracker reactor
Sensor 20 indicates a significant flaw requiring immediate inspection; sensors 17 – 19 and 21 – 24 show flaws that require follow-up evaluation

Other important acoustic emission sources include corrosion (gas evolution), stress corrosion (cracking), martensitic transformation, solidification (hot cracking), oxidation (oxide cracking and spalling), gas and fluid leakage, and welding processes (metal and slag cracking, transformation, and contamination). Magnetization of steels produces magnetoacoustic emission, which can be used to evaluate heat treatment and residual stress.

2.4.3. Source Location

Acoustic emission monitoring utilizes the structure under test to discover flaws. Acoustic emission signals emitted from the flaws travel through the structure to sensors mounted on its surface. When the structure is hot, a waveguide is used to mount a sensor. The flaws are located by analyzing the signals received. Two general approaches are employed in flaw location. The zone location method uses multiple sensors. The flaw is presumed to exist within the zone that belongs to the sensor receiving the strongest AE signals. This method is suitable for locating leaks that produce continuous emissions and flaws in highly attenuating media, such as fiber-reinforced plastics, but is also used in testing metal vessels. Typical results are shown in Figure 20 [27]. The discrete source location method also employs multiple sensors, but uses triangulation techniques based on the exact differences in time between the arrival of a given event at different sensors. It provides more accurate positions of acoustic emission sources. This approach requires an AE signal to reach at least two (linear location) or three (planar location) sensors. Attenuation of the signals must be low to avoid the use of an excessive number of sensors and signal processing channels. Most metal vessels and piping can be inspected by this approach. An example is shown in Figure 21 [28].

2.4.4. Acoustic Emission Monitoring

Acoustic emission testing of metallic pressure vessels is conducted by using one of the source location methods. The vessels are pressurized according to applicable code specifications. Sources of emissions are usually crack growth, yielding, delamination, corrosion effects (corrosion product cracking or spalling, local yielding of thinned sections), stress corrosion cracking, and embrittlement. Most weld defects also exhibit acoustic emission when stressed. When metals are ductile, AE activity is low and AE test results should be evaluated carefully. Once the metals have been embrittled by the environment or by low temperature, even early stages of stressing activate AE sources. In fact, acoustic emission is the best method for detecting hydrogen-induced cracking and stress corrosion cracking. The flaws located are usually

confirmed by other nondestructive test methods. In small, thick-walled metallic vessels, the use of acoustic emission examination alone has been approved in lieu of ultrasonics or radiography [8, Sect. VIII]. Highway gas-trailer tube testing can be performed with acoustic emission instead of the formerly used hydrostatic recertification process. Many chemical and petrochemical vessels, including tank cars and tank trailers, are monitored by acoustic emission during and between service without being emptied, thus minimizing the length of plant shutdown [39].

Composite vessels and piping are monitored with acoustic emission. This is one of the most successful AE applications (see Section 3.2).

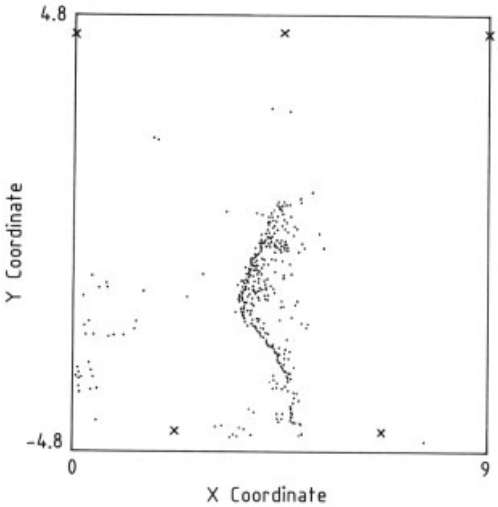

Figure 21. Discrete source location from pipe subjected to sour gas solution
X marks the sensor position; dots indicate located acoustic emissions sources on an X – Y plot

2.5. Penetrant Methods

Inspection with liquid penetrants relies on the capillary action of a surface-wetting fluid and provides a reliable method for locating minute surface flaws down to micron sizes. Surface cracks, laps, porosity, shrinkage areas, laminations, and other discontinuities can be detected as long as they are open to the surface. The shape, size, and composition of the test piece have no effect, but materials with rough or porous surfaces cannot be inspected; see [1], [2], [7, vol. 2].

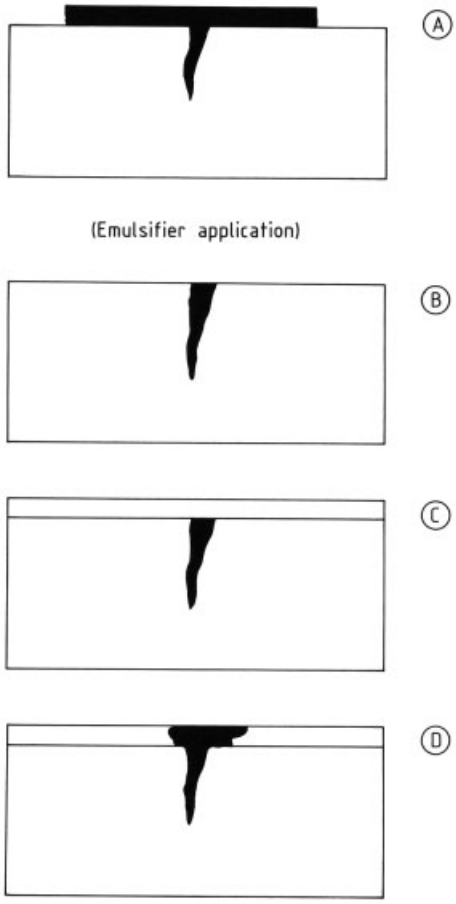

Figure 22. Steps of penetrant testing
A) Penetrant application; B) Surface Cleaning; C) Application of Developer; D) Inspection
Emulsifiers are used only for postemulsifiable penetrants.

Penetrant testing consists of (1) surface preparation, (2) penetration, (3) penetrant removal, (4) development, and (5) inspection; see Figure 22. Initial cleaning of the surface removes oil, grease, dirt, surface coatings, rust, scale, and other foreign objects that give rise to false indications or interfere with wetting by a penetrant. Chemical etching may be required to eliminate a cold-worked layer. A film of liquid penetrant is applied to cover the surface to be inspected. The film remains on the surface for a penetration (dwell) period specified for the penetrant used. Excess penetrant remaining on the surface is then removed, leaving penetrant that has seeped into discontinuities. For postemulsifiable penetrants, application of emulsifier is required before removal. A developing agent is

applied to cover the surface in a thin coating. The developer draws penetrant out of the discontinuities and broadens the apparent width of the visible indications. The test piece is then viewed under suitable lighting, either in white light for visible-dye penetrants or ultraviolet (black) light with darkened background for fluorescent penetrants.

Fluorescent and visible-dye penetrants are made with water-washable, postemulsifiable, or solvent-removable fluids. Different levels of flaw detection sensitivity are available with any combination, but visible-dye penetrant systems are less sensitive than fluorescent ones. Water-washable penetrants contain emulsifiers and can be removed by water. This requires careful washing so that penetrants in flaws are not removed. Postemulsifiable penetrants require application of emulsifier following a penetrant dwell period. Excess penetrants become washable with water and are rinsed off. Penetrant in the flaws remains water insoluble, providing excellent retention capability. Although postemulsifiable penetrants are more expensive, they provide greater sensitivity and can reveal hairline cracks and shallow discontinuities. Solvent-removable penetrants are used for limited, special applications because the use of solvent is labor intensive and in some locations requires an extensive pollution abatement system. Minimum facilities are needed especially for solvent-removable, visible-dye penetrants, which are convenient for use in remote locations.

Reference photographs of typical flaw indications are available to establish inspection criteria [9, E 433]. The smallest flaws that can be detected by penetrant inspection with a laser and a photodetector are cracks of ca. 1-µm width. Reliable visual inspection of 25-µm cracks can be made. Most inspection has been done by the human eye, but automated scanning equipment has increased inspection speed substantially for quantity inspection of identical or similarly shaped parts [29].

The quality of penetrant inspection systems can be monitored by the use of cracked-block comparators or suitable materials with known flaws. A number of comparators have been developed, including cracked chrome-plate panels, fractured glass panels, unglazed ceramic disks, and nickel – copper laminates. Quenched aluminum plates are the best known example.

Inspection of high-strength steels and titanium- or nickel-based alloys with penetrants requires special attention to the sulfur and chlorine content of all materials used. These elements should be kept below 1 %. Some penetrants are shock sensitive in the presence of liquid oxygen. Special penetrants must be used if the surface will come in contact with liquid oxygen. Another limitation of penetrant systems is that most penetrants cannot be used much above ambient temperature because they may be baked onto the surface and lose efficiency. A crayon penetrant system has been developed for high-temperature use up to 270 °C.

2.6. Magnetics and Electromagnetics

2.6.1. Magnetic Particle Inspection

Magnetic particles become attached to the surface and to some subsurface discontinuities when a ferromagnetic test piece is magnetized. The particles (e.g., iron or soft ferrite), which have high permeability and low coersive force, are coated with colored or fluorescent pigments for increased visibility. They are applied as dry particles or as suspensions in water or oil. Magnetic leakage fields of a discontinuity collect the particles, indicating its location, size, and shape by color contrast or fluorescence. Nonmagnetic materials cannot be tested with this method. Such materials include austenitic stainless steels and many heat-resistant nickel alloys; see [1], [2], [7, vol. 6], [9, E 709].

In the continuous method, magnetization is applied continuously while the magnetic particles are administered. Because the high current used causes heating, it is used for only a short time during which the magnetic particles must be applied. In the residual method, a pulse magnetization is applied, after which magnetic particles are brought in contact with the test piece. The residual method can be used when the test piece has adequate remanence to retain a leakage field for attracting magnetic particles. Most structural steels fall into this category, but low-carbon steels and soft magnetic materials require the use of the continuous method.

Magnetic leakage fields are formed when a discontinuity lies normal to the applied magnetic

field. The direction of magnetization must be selected to reveal expected flaws. Several methods of magnetization are used: (1) the flow of direct or alternating current through a test piece, and (2) the placement of a test piece within the magnetic field of a magnetizing coil or yoke. Because of the eddy-current effect (see Section 2.6.4), an a.c.-induced magnetic field concentrates near the outer surface and is used for surface flaw detection.

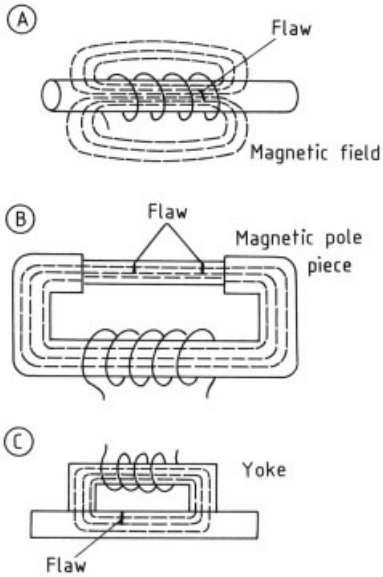

Figure 23. Longitudinal magnetization methods
A) Encircling coil; B) Between magnetic pole pieces; C) Magnetic yoke

To detect a transverse or circumferential flaw (e.g., a weld crack in a butt-welded pipe joint) a longitudinal magnetic field is necessary. To produce the required field, a magnetizing coil can be wound over the pipe (Fig. 23 A), the pipe can be placed between two pole pieces of a magnet (Fig. 23 B), or the two poles of a magnetizing yoke (a C-shaped electromagnet with flexible poles) can be placed across the flaw (Fig. 23 C). The test piece can also be placed inside a stationary magnetizing coil. To reveal a longitudinal flaw (e.g., an axial crack in a welded pipe) the magnetic field must exist circumferentially. This can be induced by diametrically applying a magnetizing yoke. Alternatively, a high current can be passed axially parallel to the flaw to produce a circular magnetic field (Fig. 24 A). By using two plod contacts placed along the weld, a current can be passed parallel to the crack (Fig. 24 B). For short hollow cylinders, electrical current may be passed through a central conductor.

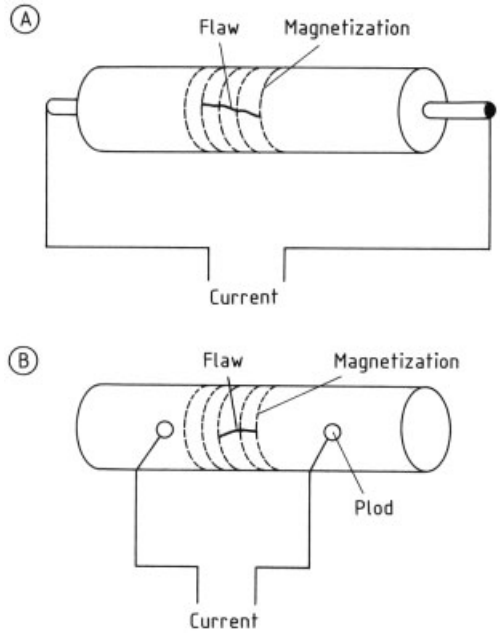

Figure 24. Circular magnetization methods
A) Current passing through a sample (head-shot); B) Current passing locally through two plods

The levels of magnetization of ferrous materials must be of the order of 1 T (10 000 gauss). Typical ferromagnetic materials have a relative permeability of more than 240. For direct, circular magnetization, the required magnetization can be achieved by 200 – 350 A per linear centimeter in diameter. For longitudinal magnetization of a short test piece having a length-to-diameter ratio (L/D) of 2 – 15, the number of required ampere-turns is 45 000/(L/D).

Interpretation of magnetic particle indications requires the ability to reject irrelevant indications. Reference photographs of indications on steel castings have been compiled and can be used to establish acceptance and rejection criteria.

Variations of magnetic particle testing have been developed. The magnetic rubber method polymerizes in situ a liquid rubber containing magnetic particles. During curing, the particles concentrate in the region of a leakage field. This allows the inspection of recessed areas (e.g., bolt

holes) and makes a permanent record. Magnetic printing and painting are also useful. These employ magnetic tape and paint for recording and indication.

2.6.2. Magnetic Leakage Field Inspection

Magnetic leakage fields from a surface of subsurface flaw can be detected and measured by Hall-effect probes, magnetic diodes, moving coils, and magnetometers. Leakage fields can in principle provide quantitative information on the flaw. The leakage fields have tangential and perpendicular components, and their distribution for simple flaw shapes has been calculated. Added to the leakage fields are large, but slowly varying bias magnetic fields. Most materials, however, have nonlinear magnetic properties that vary with processing conditions. Thus, estimation of flaw size and shape is difficult except by methods that use reference flaws for comparison; see [7, vol. 4], [30].

Hall-effect probes are made with submillimeter dimensions and have high sensitivity. Some have a built-in amplifier and a sensing element as small as $100 \times 25\ \mu m$. The probe size must not be much larger than the dimensions of the leakage field. Two probes are often used to cancel the effect of bias fields. Magnetic diodes work similarly to Hall-effect probes. A moving coil probe senses the rate of change of magnetic flux lines through the coil. The coil is moved uniformly over the surface of the sample. It must be small to achieve as high a sensitivity as Hall-effect probes.

A flux-gate magnetometer or a Förster probe utilizes the nonlinear magnetic hysteresis of a ferrite core, on which two coils are wound [3, pp. 328 – 332]. The primary coil of the probe is excited at a carrier frequency, and the second harmonics in the secondary coil are detected. The output is sensitive to the magnetic field acting on the ferrite core and has a sensitivity hundreds of times greater than that of Hall-effect probes. Nuclear magnetic resonance magnetometers are also very sensitive. Because they use microwave frequencies, these are confined to specialized studies.

Applications of magnetic leakage fields use either continuous field excitation or residual leakage fields (fields existing after the applied magnetic field is removed). The latter require high-sensitivity probes such as flux-gate magnetometers. Magnetic leakage field methods are used extensively in the inspection of tubular shapes. Excitation fields and sensing probes can be placed on the outside or inside to search for axial and transverse flaws. For internal inspection, a drive section, a magnetizing and sensing section, and a recording or data output section are combined into a form of inspection "pig," which is universally jointed to permit movement through bent tubes. For smaller diameter tubes ($<$ 5-cm inside diameter), permanent magnets can be employed for magnetization. Flaws such as pitting, holes, splits, hydrogen damage, and cracks in heat exchangers, pipelines, and chemical plant tubing can be inspected by using this method.

2.6.3. Other Magnetic Inspection Methods

The magnetic properties of a ferromagnetic material, such as initial permeability, saturation induction, coersive force, and remanence, are affected by its metallurgical history. Of these, permeability and coersive force are most often used for material characterization (i.e., hardness, heat treatment, alloy content, and grain structure). Applications include sorting of steel, ferrite content determination in stainless steel welds and castings, and detection of ferromagnetic impurities [31].

The magnetic Barkhausen effect and magnetoacoustic emission originate from the motion of magnetic domain boundaries [32]. These are also suited for material characterization and detection of residual stresses.

2.6.4. Eddy-Current Inspection

A changing electromagnetic field induces eddy currents in an electrical conductor. The eddy currents create a secondary electromagnetic field, which is detected through its effects on the primary exciting coil or by means of a separate sensor. In nonmagnetic materials or ferromagnetic materials that are magnetized to more than 98 % saturation, the electrical conductivity of the materials is the primary property controlling the secondary field. In unsaturated magnetic materials, magnetic permeability has a larger effect

on the secondary field. Moreover, the permeability of a material is sensitive to a variety of material properties, including heat treatment, cold working, composition, and magnetization level. Eddy-current testing of ferromagnetic materials has been hampered by this variation in permeability and by the difficulty of applying fields high enough to achieve saturation. This method can measure a combination of electrical conductivity and magnetic permeability, as well as any parameter or discontinuity that affects these two properties. Consequently, the detection of heat treatment and hardness changes and the sorting of dissimilar alloys can be performed. The detection of corrosion and flaws in tubular products has been an important application of this method; see [7, vol. 4], [28].

Eddy-current testing can be carried out without direct contact, and signal processing is entirely electronic. Thus, an automated inspection system can achieve high inspection rates.

Because of the eddy current, externally applied electromagnetic fields are attenuated in a conductor. At the skin depth D_s, the field strength decreases to $1/e$, where

$$D_s = 1/\sqrt{\pi f \mu_m \sigma_e}$$

where f is the frequency, μ_m the magnetic permeability, and σ_e the electrical conductivity. In terms of the electrical resistivity ($\varrho_e = 1/\sigma_e$) in $\mu\Omega \cdot$ cm, and the relative permeability μ_r (which is unity for nonmagnetic materials, and 200 – 800 in structural steels), D_s (in mm) can be expressed as

$$D_s = 50\sqrt{\varrho_e / f \mu_r}$$

Typical values of D_s are given in Figure 25 as a function of frequency.

The basic element of an eddy-current instrument is an inductor or a coil. When its inductance is L_0 and its resistance is R, the coil impedance Z_c is given by

$$Z_c = \sqrt{(2\pi f L_0)^2 + R^2}$$

Z_c is affected by a conductor in proximity, and such changes are usually plotted in a normalized impedance plane representation (Fig. 26). The abscissa is the resistive component ($R/2\pi f L_0$), and the ordinate is the inductive component ($2\pi f L/2\pi f L_0$). When a solid rod is inserted into an encircling coil, the coil impedance Z_c varies from A (no rod) to B (a fill factor or ratio of rod area to coil area of 0.49), C (0.64), and D (0.81) through E (1.0). Both the magnitude Z_c and the phase angle ($\tan^{-1} 2\pi f L/R$) vary with changing rod diameter D_r along the dashed lines. The value given for the curve of the unit fill factor is the ratio of the frequency f to the characteristic frequency f_c and is equal to

$$f/f_c = f \mu_r \sigma_e D_r^2 / 5066$$

where σ_e is given in $\mu\Omega^{-1}\,\mathrm{cm}^{-1}$ and D_r in mm. Thus, when σ_e varies, Z_c changes along the solid curve. These variations in Z_c can be measured by detecting the voltage and current in the coil. This example gives the basis for an absolute (single) coil inspection method for variations in rod diameter. The same method can be used to inspect variations in alloy composition or heat treatment for constant-diameter rods and longitudinal flaws in such rods. These flaws lower the effective conductivity of the rod in the direction of the eddy currents. A basically identical method can inspect the diameter and thickness of tubing by means of an encircling coil.

For practical instruments, two coils are used in a bridge circuit to detect small differences. A reference rod is placed in one coil, and the other coil is used to inspect a rod or tubing. Another approach is to use one coil for excitation and the other for sensing. These coils are usually inductively coupled, but they are separated for the remote field eddy-current techniques. Differential sensing methods can also be used for the coupled-coil approach [33].

Inspection frequencies of 200 Hz to 6 MHz are used in eddy-current testing. Lower frequencies have greater penetration depth, but the sensitivity to flaws and speed of inspection decrease. For surface flaw inspection of nonferrous alloys, frequencies in the megahertz range can be used. Ferromagnetic materials require lower frequencies because of the poor penetration (high μ_r).

A single coil or a set of coils can be used as a probe to be placed over the flat surface of a conductor. This arrangement is used to measure conductivity by first calibrating the instrument with electrical conductivity standards of known values. The frequency is chosen so that the thickness of test pieces and standards is at least 2.6 times the skin depth. For conductivity measure-

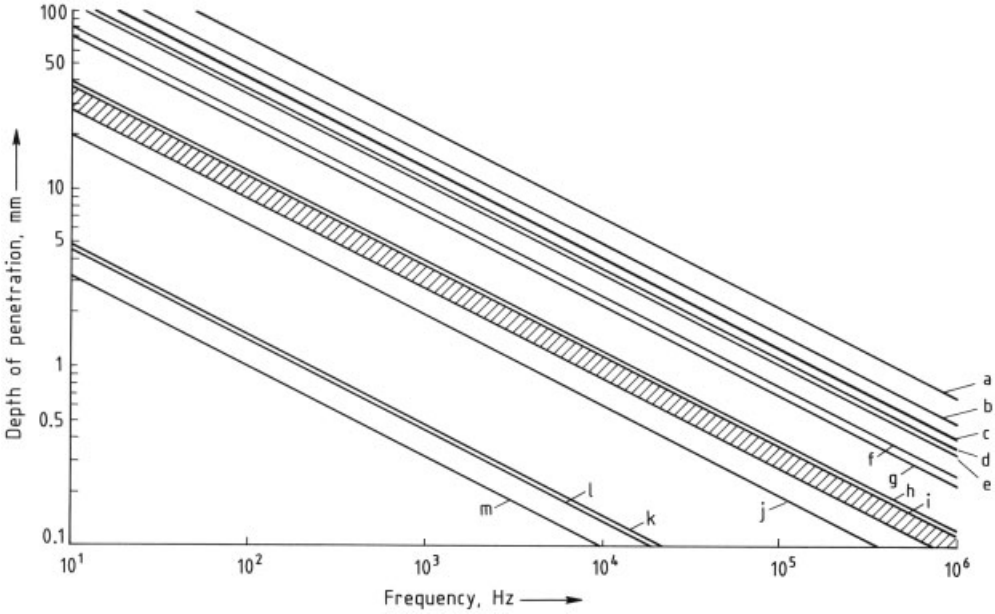

Figure 25. Skin depth versus frequency for various metallic materials
a) Hastelloy; b) Inconel; c) 300 stainless steels; d) 400 stainless steels, titanium; e) Zirconium, Monel; f) Nickel – silver (20 Ni); g) Lead; h) Brass (30 Zn); i) Aluminum alloys; j) Copper (pure); k) Carbon steel; l) Alloy steel; m) Iron (pure)

ments, the coil-to-metal distance (lift-off) is kept constant or compensated.

Changes in lift-off can be measured while the conductivity effect is kept negligible by controlling the frequency. This allows measurement of thickness for nonconductive coatings on conductive materials. The thickness of sheet materials can be measured by adjusting the frequency so that it is of the same order as the skin depth D_s.

The encircling coil technique passes a tube lengthwise through a coil energized at one or more frequencies (Fig. 27 A). The electrical impedance of the coil is modified by the proximity of the tube, the tube dimensions, electrical conductivity, saturating magnetic field, magnetic permeability, and metallurgical or mechanical discontinuities in the tube. Encircling coils are used for small-diameter tubing (< 8 cm) because the sensitivity to flaws deteriorates for larger diameter tubes.

A probe with one or more sensors in close proximity to a test piece is used in the probe technique (Fig. 27 B). The probe is generally small and examines only a limited area. For welded tubular products, only the weld is typically examined.

A coil can be placed inside a tube (Fig. 27 C). The internal coil can be concentric, the same type as an encircling coil, for exciting and sensing circumferential eddy currents. Two concentric coils can be used as an exciter and a sensor, each of which may be placed inside or outside. When one is outside and the other inside, this is a transmission technique and the sensor is affected only by those electromagnetic fields that have passed through the wall. The two coils can also be located inside or outside. This reflection technique is suited for detecting primarily inner or outer surface flaws.

A smaller probe can be used to scan helically the inner or outer surface of a tube. These rotating probe techniques are suitable for detecting corrosion and cracks in tubing (Fig. 27 D).

Reference standards for flaws are required to set eddy-current test conditions. Samples with actual flaws and those with drilled holes and machined notches of known size are employed. Notches are milled, filed, or electrodischarge machined on the outer or inner surface.

For magnetic materials, a saturation magnetic field is applied to reduce the permeability and its variation. For tubular products, an encircling saturation coil is employed. Permanent magnets and U-shaped electromagnets are used in other instances.

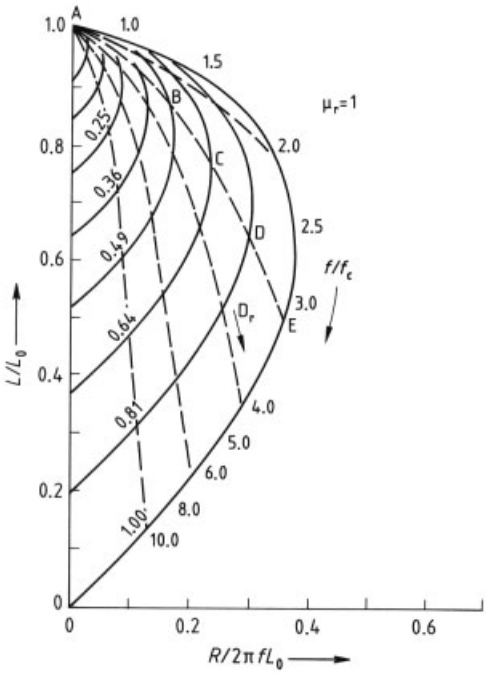

Figure 26. Normalized coil impedance plot of inductive versus resistive components for variations in frequency or conductivity (solid lines for fill factors of 1, 0.81, 0.64, . . .) and diameter (dashed lines for $f/f_c = 2$, 3, 4, 6, and 10) of a nonmagnetic solid cylinder

For eddy-current inspection of small surface flaws, various probe designs have been developed. These typically have a ferrite core and concentrate electromagnetic fields to a gap in the core. The gap is placed on the surface of a test piece, which increases the flaw detection sensitivity.

2.6.5. Microwave Inspection

Microwaves are electromagnetic radiation with frequencies of 300 MHz to 300 GHz. For nondestructive testing, frequencies of 1 – 20 GHz are generally employed. By measuring attenuation and reflection, this method can be used to analyze the moisture content of dielectric materials; to detect chemical changes, voids, and discontinuities in ceramics, polymers, and fiberglass-reinforced composites; and to gauge thickness. For conductive materials, skin depth is very low and most of the incident wave is reflected. Only reflection techniques can be used for metals. For nonconductive materials, transmission, reflection, and standing-wave techniques are used. The techniques can employ different modes of microwave generation. Frequency may be fixed, swept, or pulse modulated. Continuous waves or pulsed waves are generated [7, vol. 4].

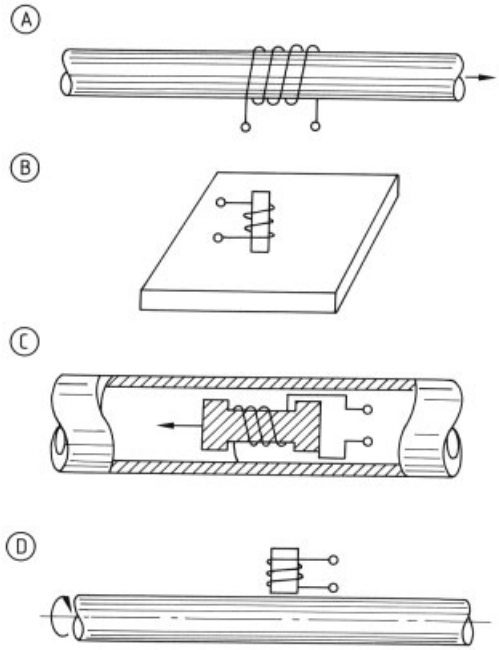

Figure 27. Examples of inspection coil arrangement
A) Encircling coil; B) Probe coil (with or without core); C) Inside coil (with or without core); D) Probe coil (rotating probe or tube)

Basic transmission techniques use a microwave generator, two antennas for transmitting and receiving, and a phase-sensitive detector or peak-amplitude detector. With the phase-sensitive detector, the in-phase and out-of-phase (quadrature) components can be determined for a fixed-frequency system. Attenuation and phase change may be used to determine the moisture content, dimensional changes, and property variation of dielectric materials. The presence of local anomalies can also be detected. By using a pulse-modulated transmission and a peak-amplitude detector, the velocity of propagation

can be measured. This is useful in evaluating polymerization, esterification, distillation, vulcanization, and evaporation in polymers.

Reflection techniques use one antenna. When a metal backing exists or a reflector is placed behind a dielectric sample, property measurements similar to those made with transmission techniques can be achieved. Instruments for dielectric property determination are referred to as microwave dielectrometers. Both fixed-frequency and swept-frequency (also known as frequency-modulated, FM) reflection techniques are also used to detect delaminations (if a detectable air gap exists), inclusions, voids, and other three-dimensional anomalies in dielectric materials.

Since microwaves are reflected at an interface, standing waves are produced when parallel interfaces or reflectors exist; for example, the front and back faces of a plastic plate. The patterns of standing waves are used for distance measurements and thickness gauging.

Moisture measurements are perhaps the most widely used application of microwaves. Water has a high dielectric loss compared to most other dielectric materials, and its content can be determined by sensing dielectric constants with transmission or reflection techniques. Transmission techniques may be employed at high water levels (50 – 85 %). Reflection measurements with a resonant cavity can achieve high sensitivity over a narrower range of moisture content.

Another use of microwaves occurs in eddy-current inspection. In contrast to the usual frequencies up to 10 MHz, a ceramic crystal resonator is used in a sensor, extending the frequency to several gigahertz. Resolution of surface flaws is greatly improved.

2.7. Thermal and Optical Methods

2.7.1. Thermal Inspection

Heat flow and temperature changes are detected to locate flaws and abnormal heat distribution. Temperature can be measured by many types of substances and devices such as heat-sensitive paints, papers, and phosphors; liquid crystals; thermocouples; thermistors; and bolometers. These have been used in limited applications because direct contact is required and only point-by-point measurements are possible with some methods. Massive changes in thermal conductivity from thermal fatigue cracking, delamination, and stress corrosion cracking may be detected by using coatings of liquid crystals and heat-sensitive substances and applying certain heat flow patterns [1], [2].

Noncontact thermography utilizes IR radiation emitted by the surfaces of a test piece. The principle used is similar to radiation pyrometry, but modern thermographic cameras can obtain false-color images of the surface, with each color representing a narrow range of temperature having intervals as small as 0.1 °C. Infrared optics are also available to image microscopic areas. Thermographic cameras can be employed to evaluate the temperature distribution of a component in service. When an abnormal condition exists, such an examination can reveal it. Transient heat flow can also disclose an anomaly in thermal conductivity that may result from delamination and unbonding. Temperature changes on the surface can be induced by an external heat source. Higher temperatures are observed where the section thickness has been reduced by corrosion. This method is applicable to inspection of piping.

Another method of inducing a temperature differential is to apply mechanical vibration. This is known as vibrothermography [40] and can be applied to flaw detection in composite materials subjected to fatigue loading.

A scanning high-power laser can be combined with an IR detector. The unbonding of a coated layer can be revealed because the temperature of the coating at the irradiated unbonded spot rises more than that at well-bonded regions. Photoacoustic spectrometry employs modulated laser beams, irradiating mainly particulate matter. Thermal excitation leads to detectable emission of acoustic waves. This method has been used in evaluating ceramic powders and chemicals.

2.7.2. Visual and Optical Inspection

Evaluating the surface of a structure with the human eye is generally the first step in any nondestructive inspection [42]. Visual inspection is facilitated by using a magnifier, a borescope, a fiberscope, or a videoscope. Video technology has also become an integral part

of this method for examination and recording of images. Borescopes and fiberscopes are used for remote viewing. Borescopes are low-magnification devices with a rigid tube support for optical lenses. Their length may be extensible. Fiberscopes are made with a bundle of flexible optical fibers and provide magnification up to 200×. These are available in diameters as small as 0.35 mm and are often combined with a video camera. Videoscopes employ a small image sensor, which is typically a charge-coupled device consisting of many light-sensitive elements. The image is sent electronically to a video monitor. Videoscopes provide higher resolution and greater depth of field than fiberscopes. Images from fiberscopes and videoscopes are displayed on a monitor screen and recorded on a video recorder.

Long-distance microscopes have a large working distance, up to a few hundred times that of conventional microscopes. This allows the examination of small objects 1 – 2 m away with 5-µm resolution.

By using a laser as the light source and scanning rapidly over a surface, transmitted, reflected or scattered light can be used to characterize the surface by imaging the flaws and discontinuities and by measuring the surface roughness. Diffraction of light scattered from certain kinds of surfaces (e.g., weave patterns) can be used to detect unacceptable surface patterns.

Optical interferometry is used to obtain surface profiles via interference fringes [44]. Noncontact displacement measurements can be made with a resolution of 10 nm. Two-dimensional surface profiles are also obtained by laser speckle reflectometry [43], which is employed mainly for measurement of roughness.

2.7.3. Optical Holography [41]

Optical holography utilizes coherent monochromatic light from a laser. First, both the amplitude and the phase of coherent light from a test piece are recorded on a film or plate along with the reference beam (Fig. 28 A). Three-dimensional images are formed from the holographic film or plate when it is irradiated with the same type of light (Fig. 28 B). The regenerated image can then be used as a three-dimensional template against which any deviations in the shape or dimensions of the test piece can be observed and measured. With no part of the optical system or test piece position changed, the test piece is irradiated with the same reference beam and an inspector views it through the hologram. Interference patterns are seen with optical fringes surrounding anomalies where the test piece has moved short distances. In a double-exposure technique, the surface of the test piece is slightly distorted between exposures, which are recorded on a single holographic film, thereby producing a holographic fringe pattern. Another method utilizes the interfering holographic image effects of vibrating test surfaces.

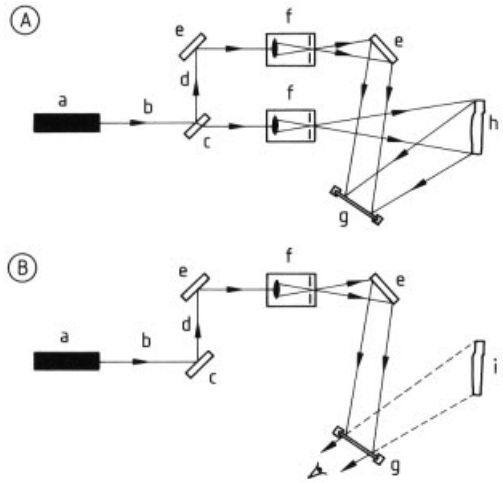

Figure 28. Basic holography setup
A) Recording hologram; B) Reconstructing image
a) Laser; b) Laser beam; c) Beam splitter; d) Reference beam; e) Mirrors; f) Spatial filters; g) Photographic plate; h) Test piece; i) Virtual image of test piece

These methods are used to inspect brazed or adhesively bonded honeycomb-sandwich structures, coatings, and tires for lack of bonding, delamination, and other flaws. They are also useful in vibration analysis of rotating and vibrating components.

2.8. Leak Testing

Leak testing detects the leakage of a liquid or gas from an internally pressurized vessel or the leakage of a gas into a vacuum vessel. The location of a leak, the amount of leakage, or both can be determined. A number of methods have

been developed with different sensitivities and detection principles [1], [2], [7, vol. 1].

2.8.1. Pressurized Gas Leaks

Two approaches are used to detect leaks in pressurized gas systems. The amount of gas lost is determined by the measurement of weight or pressure. Both methods are insensitive, and leak location is not possible. If two vessels are available, a differential pressure gauge can be used to increase the sensitivity of the pressure-gauging method.

The other approach is to detect the escaping gas; many techniques are available for this. Acoustic methods rely on sound or ultrasound generated by turbulent flow through a leak. If the leak is large, audible sounds are emitted. Stethoscopes can be used for smaller leaks. Still smaller leaks can be detected with ultrasonic probes, which operate for airborne ultrasound in the range of 30 – 45 kHz. Airborne ultrasound can be detected from as far away as 30 m and works well for overhead gas pipelines. Parabolic sound collectors are useful. With low background noise, turbulent leaks on the order of $1\,\mathrm{mPa}\cdot\mathrm{m}^3\,\mathrm{s}^{-1}$ can be detected. An acoustic emission sensor (30 – 100 kHz) can be coupled directly to the leaking vessel. By using multiple AE sensors, a leak can be located. This method is effective in leak detection especially in buried gas pipelines.

Bubble testing is applied to small vessels that can be immersed in a liquid. Oils and glycerol are more sensitive to leaks than water, and reduced pressure above the liquid is also helpful. Bubble-forming solutions are applied to the surface of larger or fixed-pressure vessels and piping. Soap solutions can be used for simple tasks, but specially formulated solutions have been developed for more consistent bubbling. Sensitivity on the order of $10^{-2}\,\mathrm{mPa}\cdot\mathrm{m}^3\,\mathrm{s}^{-1}$ and qualitative estimates of leak size are feasible.

Flow detection is applicable to a vessel if it can be placed in a larger enclosure with an orifice. The flow of leaking gas from the orifice is measured by using a flow meter. Leaks of $1\,\mu\mathrm{Pa}\cdot\mathrm{m}^3\,\mathrm{s}^{-1}$ can be detected.

Specific-gas detectors are used to respond to halogens, combustible and odorous gases, gases that cause chemical reactions, and certain other gases. The human nose detects very low levels of odorous gases but quickly becomes insensitive. A halogen-sensitive detector uses a diode that ionizes halogen atoms and detects the ionic current. Air is drawn into the diode through a sniffing probe. If a vessel does not contain halogens, a tracer gas containing halogen compounds can be mixed with the gas in the vessel for leak detection. Combustible gas and sulfur hexafluoride gas detectors are similar to the halogen detector. Common uses of the halogen and combustible gas detectors are for air-conditioner maintenance and in unventilated, enclosed areas where leaked gases accumulate. Leaked gas can react chemically with a coating or tape placed on the surface of a vessel to change its color. Ammonia is generally used for this method. Another type of gas detector works like a thermocouple vacuum gauge and a Pirani gauge [45]. Gases with a different thermal conductivity (e.g., hydrogen) can be sensed relative to air by this method.

Mass spectrometers detect the distribution of atomic or molecular weights of different components of a gas. Helium mass spectrometers are most commonly used in leak detection. Either a sniffing probe is used to draw leaked gas from a helium-filled vessel (Fig. 29 A) or an evacuated enclosure may be used to collect the leakage from a helium-filled vessel (Fig. 29 B). When a vessel can be evacuated (even partially), it can be placed in an enclosure filled with helium (Fig. 29 C). This is the most sensitive method, and a sensitivity of $\mathrm{nPa}\cdot\mathrm{m}^3\,\mathrm{s}^{-1}$ is often achieved.

2.8.2. Pressurized Liquid Leaks

Liquid leaks are visually detectable if leak rates are high. Dye penetrants may be introduced into a vessel and procedures for penetrant testing applied on the outside of the vessel. As in the detection of chemically reactive gases, the vessel surface can be coated with pH-sensitive paint, paste or tape, and color changes can be monitored. Liquid level loss monitoring is simple and is used commonly for underground tanks. The liquid-containing tank can be pressurized and sealed, and the pressure drop monitored.

Acoustic methods are also useful when turbulent flow occurs. Acoustic emission tech-

niques are used to detect leaks through valves, steam condensate lines, and flat-bottomed tanks. Metal and plastic pipelines are monitored for leaks by means of multiple acoustic emission sensors during hydrostatic testing.

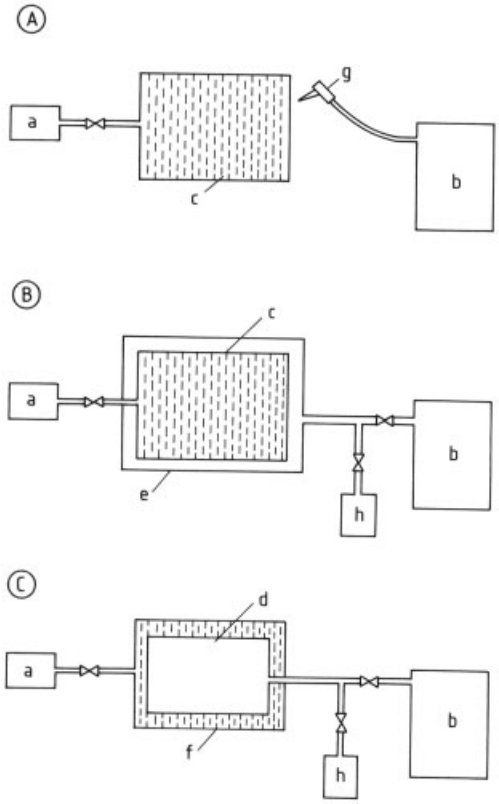

Figure 29. Helium leak testing
a) Helium source; b) Mass spectrometer; c) System under test (helium filled); d) System under test (evacuated); e) Evacuated enclosure; f) Helium-filled enclosure; g) Probe; h) Vacuum pump

2.8.3. Vacuum System Leaks

Vacuum systems are normally tested in the vacuum condition. Most vacuum-sealing systems are not suited for internal pressurization, and container walls may not withstand pressure reversal. Helium mass spectrometers are generally used for leak detection. Vacuum pressure gauges and halogen detectors may also be used. The vacuum system is connected to the helium detector, and a stream of helium gas is directed at a suspect area. To determine the total leak rate, the system is enclosed in a container filled with helium.

In testing high-vacuum systems, virtual leaks may be a substantial source of leakage. These involve gradual desorption of gases from surfaces or components within a vacuum system.

3. Uses

3.1. Metallic Pressure Vessels and Piping

Pressure vessels are constructed in various sizes and shapes, but invariably these are manufactured and inspected according to the American Society of Mechanical Engineers (ASME) Boiler and Pressure Vessel Code [8]. Section V of the ASME code establishes methods of nondestructive testing and covers visual, radiographic, ultrasonic, acoustic emission, magnetic particle, penetrant, and eddy-current methods.

The code specifies inspection methods and acceptance criteria for plates, forgings, and tubes, before and during fabrication, and for vessels and piping after fabrication. The most likely locations of flaws are at various weldments (see Section 3.3).

Acoustic emission testing is used to localize defective areas during initial hydrotesting. It is also employed for in-service inspection and continuous monitoring, in which process fluids are used for pressurization. Results indicate the locations of active AE sources, which are graded according to their severity. The severity of the source is classified according to the average energy of the most energetic emissions, energy release rates, number of AE events during a load hold period, number of events during a load increase, number of high-amplitude events, etc. These provide real-time indications of defective areas of the pressure vessel being tested and allow prevention of catastrophic failure of the vessel. The defective areas identified are then inspected by using other test methods. Acoustic emission testing is applied on various types of vessels, including ammonia spheres and tanks, hydroformer reactors, and catalytic cracking units [8, Section V], [34].

Surface examination for cracks, corrosion, and contamination relies mainly on visual methods, which are used for all welds not internally tested after a hydrotest. Magnetic particle and

penetrant methods are also applied for crack detection. For surface inspection of vessel interiors, plant shutdown is required.

Generally, radiographic and ultrasonic methods are used for internal flaw detection. Both X- and gamma-ray sources are used in radiography. Access to the vessel interior is often difficult, which limits the use of radiography. For heavy steel sections (> 10 cm), field inspection becomes difficult because high-energy sources are needed. Ultrasonic methods, especially with shear waves, are widely used for weld inspection. Thickness measurement using the normal beam is commonly practiced for corrosion monitoring. Those areas having a high corrosion rate or history of attack are inspected thoroughly.

3.2. Composite Pressure Vessels and Piping

The performance of fiber-reinforced plastic vessels and piping has been poor, and many failures have been recorded. Apart from acoustic emission, no satisfactory test exists for determining the structural adequacy of fiber-reinforced plastic equipment. Radiography is difficult because resins have low absorption coefficients. Fiber-reinforced plastics are anisotropic, and ultrasonic attenuation is very high, making pulse-echo techniques hard to use. In addition, the various types of life-limiting flaws are quite different and some of them are not amenable to conventional ultrasonic inspection. Fiber fracture and fiber – matrix debonding are difficult to detect with ultrasonics.

The use of acoustic emission for testing fiber-reinforced plastic vessels and piping has been highly successful, and standard test procedures have been established [9, E 1067], [28], [35], [36]. Test vessels and piping are pressurized up to 150 % of the maximum allowable working pressure. The procedures are designed to locate substantial flaws, which are then evaluated by other techniques such as ultrasonic testing (delamination) or visual (resin loss) and penetrant (matrix cracking) inspection. A vessel (or piping) to be tested can be new, in-service, or repaired, and different steps are specified for atmospheric, vacuum, and pressurized vessels. A vessel that has been in service must be preconditioned by reducing the operating pressure; for this, the maximum operating pressure within the previous year must be known. Acoustic emission instrumentation should have a sufficient number of channels to localize sources by using the zone location method; that is, many AE sensors are installed to completely cover the vessel with the corresponding zone marked around each sensor. Acoustic emission activities of flaws within each zone are detected, and the zone represents the approximate position of these flaws. High-frequency (100 – 200 kHz) sensors are used for zone location. Two or more low-frequency (25 – 75 kHz) sensors are used to evaluate the adequacy of coverage of the high-frequency sensors. If a low-frequency sensor detects acoustic emission whereas none of the high-frequency sensors do, the latter must be relocated. Sensors are positioned to detect structural flaws at critical sections of the test vessel, such as high-stress areas, geometrical discontinuities, nozzles, manways, repaired regions, support rings, and visible flaws.

Pressurization of a vessel during AE testing proceeds in steps, with pressure hold periods. For atmospheric vessels, the pressure is held at 50, 75, 87.5, and 100 % of the test pressure. Pressure vessels are stressed with 10 % increments, with depressure increments (also 10 %) above 30 % of the test pressure. A test is terminated whenever a rapid increase in AE activity indicates an impending failure. Acoustic emission data from the high-frequency sensors are used for evaluation, whereas the low-frequency sensors generally detect acoustic emission from significant flaws. Detected flaws are graded according to using several criteria, including emissions during pressure hold periods, felicity ratio (the ratio of the load at the onset of significant emissions to the maximum prior pressure), total AE counts, high-amplitude events, and long-duration events. Emissions during hold indicate continuing permanent damage and lack of structural integrity. For in-service vessels, the felicity ratio criterion (when it is less than 0.95) is an important measure of previous damage. High-amplitude events indicate structural (fiber) damage, especially in new vessels. Long-duration events are characterized by measured area of the rectified signal envelope (MARSE), which is an indicator of combined signal and amplitude duration. Large MARSE values result from

delamination, adhesive bond failure, and crack growth.

3.3. Weldments

Nondestructive testing methods used for completed fusion weldments include (1) visual inspection, (2) radiography, (3) ultrasonic pulse echo, (4) magnetic particle and leakage field testing, (5) liquid penetrant testing, (6) leak testing, and (7) acoustic emission testing.

For many noncritical welds, integrity is assured mainly by visual inspection to look for cracks, bead thickness, bead contour, undercut, overlap, and spatter. For critical welds, both faces and root surfaces are examined, especially for cracks, undercut, root penetration, and unfilled craters. These are tested further by radiography and ultrasonics for internal flaws.

Radiography is commonly used for detecting porosity, slag entrapment, and inclusions. A round or oval dark spot represents the image of a pore. Slag inclusions appear along the weld edge as irregular or continuous dark lines, whereas tungsten inclusions give rise to single or clustered light spots. Cracks are sometimes visible in radiographs as dark narrow irregular lines, but the lack of any radiographic image of cracks does not assure their absence. Radiography may be unable to detect incomplete fusion and incomplete penetration because of their small effects on X-ray absorption. They appear as very narrow dark lines.

Ultrasonic pulse-echo techniques are used effectively for the detection and location of planar defects such as laminations, unbonded areas, cracks, hidden surfaces, and other flaws. They are also used to reveal a lack of root penetration, porosity, and unbonded sidewalls of fusion zones. Ultrasonic testing and radiography are complementary techniques, and both are used for inspection of critical welds. When no access to the opposite side of the weldment is available, ultrasonic testing is the only option for internal flaw inspection.

Shear wave beams are generally used in weld inspection. To detect longitudinal flaws (along the weld), the search unit is moved along a zigzag scanning path either with sharp changes in direction or with right-angle changes. To detect transverse flaws in welds, the search unit is placed on the base metal surface at the edge of the weld. The sound beam is directed into the weld by angling the search unit at ca. 15°. The scanning of the search unit is parallel to the weld.

Magnetic particle testing and penetrant testing are used mainly for surface-breaking flaws (also for subsurface flaws in magnetic particle testing). These methods are relatively inexpensive but can reveal even fine cracks with clarity. In magnetic particle testing, the magnetic field is applied in two mutually perpendicular directions.

Leak testing of welded vessels uses a tracer gas under pressure or vacuum. Welds are tested for leak location or for leak rate. Acoustic emission tests for large-scale structures can narrow the areas of inspection by locating sources of acoustic emission. Typically, the structure is stressed by pressurization while AE sensors are mounted in arrays on the surface of the structure. When the applied stress exceeds the previously applied stress level, acoustic emission activities increase rapidly. The positions of such activities are identified by triangulation or the zone-location method. Often, the source of acoustic emission is localized to specific weld regions which are inspected further by ultrasonic and radiographic methods.

3.4. Other Uses

Polymeric materials are adversely affected by improper curing, inclusions, porosity, environmental degradation, machining and impact damage, and fretting. These are detected by specialized application of various nondestructive testing methods. Ultrasonic propagation characteristics are used to reveal anisotropy of elastic moduli, film and fiber orientation, structural relaxation, glass transition, degree of cross-linking, and state of cure in polymers.

Dielectric measurements are used to monitor the curing process [37]. In situ radio-frequency impedance is determined for this purpose. Microwave transmission and reflection methods are used for flaw detection (cracks, voids, inclusions) and for monitoring moisture, thickness, and cure. Optical techniques are employed to measure the state of mechanical strain in transparent polymers (photoelasticity). This is used

widely in studying stress concentration effects in model components.

Thermographic techniques are useful in testing polymers because they have low thermal conductivities. Infrared spectroscopy provides information on molecular structure and has also been used to measure stress concentration effects near crack tips.

Radiography of polymers requires low-energy (ca. 20 kV) radiation because polymers have low absorbance. To enhance contrast, high atomic number penetrants are used. Test pieces are immersed in fluids such as tetrabromoethane and diiodobutane. This is only beneficial for surface-breaking flaws. Although difficult to use, neutron radiography is effective for polymers.

Paper and paperboard products require special testing methods during production. Radiation thickness gauging with beta emitters (^{85}Kr and ^{90}Sr) is used for the measurement of basis weight (areal weight). For determining ash in paper due to fillers and coating minerals, the attenuation and fluorescence of low-energy photons (4 – 5 keV) are used.

Various on-line optical methods are employed in determining opacity, surface roughness, color, and reflectance. Infrared absorption in the 1.95-µm region is used for moisture measurement. For heavyweight paper, moisture is measured by microwave methods.

The mechanical properties of a paper can be probed by means of ultrasonic wave propagation. An online caliper system obtains the thickness and ultrasonic propagation data and the results are correlated to the mechanical strength of the paper.

4. References

1. *Metals Handbook,* "Inspection and Quality Control," 8th ed., **vol. 11,** Amer. Soc. Metals, Metals Park, Ohio 1976.
2. *Metals Handbook,* "Nondestructive Evaluation and Quality Control," 9th ed., **vol. 17,** ASM International, Metals Park, Ohio 1989.
3. D. E. Bray, R. K. Stanley: *Nondestructive Evaluation,* McGraw-Hill, New York 1989.
4. R. Halmshaw: *Nondestructive Testing,* Arnold, London 1987.
5. *Nondestructive Testing,*"A Survey," NASA SP-5113, NASA, Washington D.C. 1973.
6. *Nondestructive Testing Study Guides,* "Basic Magnetic Particle Testing, Penetrant Testing, Eddy-Current Testing and Radiography (1983)," General Dynamics, Convair Div., 1977 – 1983.
7. *Nondestructive Testing Handbook,* 2nd ed., **vol. 1, "Leak Testing," 1982; vol. 2, "Liquid Penetrant Tests," 1982; vol. 3, "Radiography," 1985; vol. 4, "Electromagnetic Methods," 1987; vol. 5, "Acoustic Emission Testing," 1987; vol. 6, "Magnetic Particle Testing," 1989,** ASNT, Columbus, Ohio.
8. ASME Boiler and Pressure Vessel Code, ASME, New York 1989.
9. *1989 Annual Book of ASTM Standards,* **vol. 3.03, "Nondestructive Testing,"** ASTM, Philadelphia 1989.
10. *Redi-Reference Guide,* "An Issue of Materials Evaluation," published yearly by ASNT, Columbus, Ohio.
11. R. S. Sharpe (ed.): *Research Techniques in Nondestructive Testing,* **vol. 1 1970, vol. 2 1973, vol. 3 1977, vol. 4 1980, vol. 5 1981, vol. 6 1982, vol. 7 1984, vol. 8 1985,** Academic Press, London.
12. D. O. Thomson, D. E. Chimenti (eds.): *Quantitative NDE,* **vol. 1 – 7,** Plenum Publishing, New York(1982 – 1988).
13. *Proc. World Conference on Nondestructive Evaluation,* **11th,** Las Vegas, Nov. 1985. J. Boogard, G. M. van Dijk (eds.): *Non-Destructive Testing (Proc. World Conference on Nondestructive Evaluation),* Elsevier, Amsterdam 1989.
14. *17th Symposium on Nondestructive Evaluation,* San Antonio, April 1989.
15. J. M. Farley, R. W. Nichols (eds.): *Nondestructive Testing,* vol. 1 – 4, Pergamon Press, Oxford 1988.
16. Materials Evaluation; NDT International; Research in Nondestructive Evaluation; Brit. J. Nondestructive Testing; Soviet J. Nondestructive Testing; NTIAC Newsletter.
17. F. Feigl, U. Anger: *Spot Tests in Inorganic Analysis,* 6th ed., Elsevier, Amsterdam 1972.
18. G. Tschorn: *Spark Atlas of Steels,* Macmillan Publ. Co., New York 1963.
19. A. R. Fee, R. Sagabache, E. L. Tabolski, *Metals Handbook,* "Hardness Testing," 9th ed., **vol. 8,** ASM International, Metals Park 1985, pp. 69 – 113.

20. R. Halmshaw: *Industrial Radiography-Theory and Practice,* Applied Science Pub., Englewood, NJ 1982.
21. R. A. Armistead, R. N. Yancey, *Mater. Eval.* **47** (1989) 487 – 491. B. D. Hansche, *Mater. Eval.* **47** (1989) 741 – 745.
22. J. Krautkrämer, H. Krautkrämer: *Ultrasonic Testing of Materials,* 3rd ed., Springer Verlag, Berlin 1983.
23. M. G. Silk: *Ultrasonic Transducers for Nondestructive Testing,* A. Hilger, Bristol 1984.
24. C. Broere et al., in [15], vol. 4, pp. 2424 – 2432.
25. Progress in Acoustic Emission, (a) I (1982), (b) II (1984), (c) III (1986), (d) IV (1988), Japan Soc. Non-Destructive Inspection, Tokyo.
26. Many articles in *J. Acoustic Emission,* vols. 1 – 9 (1982 – 1990).
27. R. K. Miller et al., *J. Acoustic Emission* **8** (1989) 25 – 29.
28. R. Davies, in [25, (c)], pp. 9 – 25.
29. K. Göbbels, G. Ferrano, in [15] vol. 4, p. 2763.
30. W. Lord (ed.): *Electromagnetic Methods of Nondestructive Testing,* Gordon and Breach, New York 1985.
31. D. C. Jiles, *NDT Int.* **21** (1988) 311 – 319.
32. K. Ono, in [25, (c)]pp. 200 – 212.
33. V. S. Cecco, F. L. Sharp, *NDT Int.* **22** (1989) 217 – 221.
34. T. J. Fowler, in [25, (c)]pp. 150 – 162.
35. *Recommended Practice for Acoustic Emission Testing of Fiberglass Reinforced Plastic Tanks/Vessels,* Soc. Plastics Industry, New York 1982.
36. T. J. Fowler, *Chem. Proc. (Chicago)* March 1984, 24 – 27.
37. F. I. Mopsik et al., *Mater. Eval.* **47** (1989) 448 – 453.
38. C. A. Salvado, *Proc. Design and Manufacturing of Advanced Composites,* ASM International, Metals Park 1989, pp. 111 – 119.
39. T. J. Fowler, J. A. Blessing, P. J. Conlisk, *AECM-3* (Third Int. Symp. Acoustic Emission from Composite Materials, Paris 1989), ASNT, Columbus 1989, pp. 16 – 27.
40. S. S. Russell, E. G. Henneke, *NDT Int.* **17** (1984) 19 – 25.
41. J. W. Wagner, in [22], pp. 405 – 431.
42. R. C. Anderson, *Inspection of Methods: Visual Examination,* vol. 1, ASM, Metals Park, Ohio 1983.
43. E. P. Chiang in C. P. Grover (ed.): *Optical Testing and Metrology,* **vol. 661,** SPIE International Soc. Optical Eng., Bellingham 1986, pp. 249 – 261.
44. R. Jones, C. Wykes, *Holographic and Speckle Interferometry,* Cambridge University Press, Cambridge 1983.
45. In [7, vol. 1], pp. 346 – 368.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Patents

Standard Article

Christian Lang[1]

[1]LangRaible IP Law Firm, München, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b01_13.pub2
Article Online Posting Date: July 15, 2005

Abstract | Full Text: **HTML**

Abstract

This article provides an overview on patents and protective rights in general. First of all information is given which protective rights exist and what can be protected by these rights. Afterwards, general principles of patent law are presented which are common to many national patent laws and international treaties regarding industrial property throughout the world. A main focus of this presentation is directed to the description of the proceedings according to the Patent Cooperation Treaty (PCT) and the European Patent Convention. Furthermore, the particularities of US patent law are considered.

The article contains sections titled:

1. Introduction
2. Industrial Property and Types of Protective Rights
2.1. Technical Protective Rights (Patents, Utility Models)
2.2. Trademarks
2.3. Design Patents/Industrial Designs
2.4. Copyright
2.5. Others
3. Basic Principles of Patents
3.1. Invention
3.2. Exclusive Privilege
3.3. Principle of Territoriality
3.4. Period of Protection
3.5. Ways of Establishing Patents
3.6. Priority
3.7. Novelty
3.8. Inventive Step/Nonobviousness
3.9. Industrial Application
4. System of International Treaties
4.1. Paris Convention
4.2. Patent Cooperation Treaty
4.3. Regional Treaties
5. Patent Cooperation Treaty (PCT)
5.1. Application
5.2. Search
5.3. Preliminary Examination
5.4. Entry into National/Regional Phase
6. European Patent Convention
6.1. Member States
6.2. Application Proceedings
6.2.1. Application
6.2.2. Examination Procedure
6.3. Opposition Proceedings

6.4. Appeal
7. U.S. Patents
8. Other National Law

[Top of Page]

1. Introduction

As a result of new technical developments in the field of information technology and data processing, a large amount of information is available in a very short time all over the world, e.g., via internet. Furthermore, mobility of persons and goods via modern means of transportation means that services and goods are offered worldwide. While these developments, known under the term “globalization”, may have many positive results with respect to technical development and economic growth, the participants in this system must be aware of the consequences of these changes with respect to the protection of intellectual property.

This may be the reason for the increasing number of patents and other protective rights for intellectual property worldwide. As statistics of the World Intellectual Property Organization (WIPO) show, the number of patent, trademark, and design patent applications has been continuously growing for the few last decades.

Accordingly, people working in the field of research and development and responsible managers are forced to deal with protective rights in the field of industrial property. This is not only necessary because companies and private inventors need to apply for protection for their own intellectual property but also with respect to defence against claims of third parties that limit economic freedom.

This article provides an overview of the patent system in the world, especially in Europe, and the role of patents in the overall system of protective rights in the field of industrial property. While this overview may help to find some guidelines for dealing with patents, it cannot replace professional advice with respect to specific questions or problems in order to avoid loss of rights. Since patents and the claims thereof are based on a number of national and international acts and treaties, this system is too complex to be completely described in this short article. In particular, requirements during application proceedings with respect to observing time limits, payment of fees, avoiding publishing of novelty-destroying documents, etc. are so complex that they cannot be represented in full detail at a glance. However, this article should provide enough information to enable decision-making on how to deal with innovations and inventions and to facilitate communication with specialists such as patent attorneys.

[Top of Page]

2. Industrial Property and Types of Protective Rights

“Patent” is a term which is commonly used for protective rights in the field of industrial property, although different protective rights are meant. However, patents usually only cover technical inventions, while other intellectual property is protected by different protective rights. In the following, an overview of the various protective rights and the subject matter of protection thereof is given.

2.1. Technical Protective Rights (Patents, Utility Models)

While in Europe and most other countries, the protection of inventions by patents is restricted to technical inventions, this does not apply to U.S. patent law. According to 35 U.S. Code § 101, inventions which can be protected by patents or utility patents are defined as follows: “Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefore, subject to the conditions and requirements of this title.” The Court of Appeal for the Federal Circuit decided in the decision State Street Bank & Trust Company versus Signature Financial Group that all kinds of inventions, also including business methods, are covered by this provision of the U.S. Code. Contrary to the practice in the United States, in Europe and most other countries the subject matter for which protection can be sought is restricted to technical inventions.

Although it is not explicitly mentioned in patent acts or treaties, the requirement for the technical character of an invention according to European and almost worldwide practice is derived from the fact that one of the preconditions for obtaining a patent refers to novelty with respect to the prior art in the specific field of technology. Therefore, according to the European practice, one requirement for obtaining a patent is the technical character of the invention. Technical inventions comprise technical methods including methods for producing a product or working methods of machines and the like, as well as products, chemical substances, devices, apparatuses, arrangements of devices, components thereof, etc. However, business methods are also excluded from patent protection as well as other specific subject matters. This is discussed in more detail below.

In some countries, in addition to patents, other possibilities for protecting technical inventions exist. For example, in Germany it is possible to obtain, in addition to a patent, a so-called petty patent or utility model (*Gebrauchsmuster*). This type of protective right is different in the duration of the protection as well as in specific aspects with regard to the requirements for obtaining protection. For example, the German utility model merely covers technical devices or products but not methods.

2.2. Trademarks

Besides patents, trademarks are the most important protective rights. Trademarks protect names, signs, figures or the like and, in general, means of marking by which the origin of products or services is proved. Due to this main function of a

trademark, the means of marking must be distinctive in order to fulfil the object of indicating the origin of the product or service.

Accordingly, a trademark only protects the designation of a product or process in order to allow the consumer to identify a product or service as a specific product or service from a certain supplier, but does not provide any protection with respect to the product or service itself. Therefore, a competitor would be able to offer the same product or service if a different trademark or means of marking is used.

2.3. Design Patents/Industrial Designs

Design patents or industrial designs are intellectual property rights which protect the aesthetic effect of a creation, i.e., the outer appearance of a product or the package thereof. Compared to the subject matter of patents, the protection of industrial designs or design patents is directed at the shape, form, colors, etc. of a product. While the scope of protection of a patent normally covers different technical embodiments of a technical invention, the design patent is directed at the aesthetic impression of the product. Accordingly, a product which is only protected by a design patent can be copied by a competitor if the overall appearance is different.

2.4. Copyright

Copyrights cover artistic and scientific merits such as works of literature, music, painting, visual arts, performing arts, etc. The copyright is directed to the intellectual content and the form of expression of the artistic or scientific creation. As a consequence, copyrights merely comprise creative works which have the necessary character of an artistic or a scientific creation due to the mental effort of creating the creative works and subsequently cover only the individual form of expression.

Compared to the subject matter of patents and design patents, it must be considered that not every creation comprises the intellectual or aesthetic content which is necessary to fulfil the requirements for being a creation under the scope of copyright and that in addition the scope of protection is restricted to the actual form of expression. Thus, the design of most industrially applicable products can be copied if there is no protection by an industrial design or a design patent, since these products often cannot be regarded as creative works in the sense of copyright. Furthermore, the copyright of works of literature does normally not cover the whole content of the ideas described but only the form of expression.

2.5. Others

In addition to the above-mentioned protective rights, further legal provisions exist which may play a role for avoiding copying and counterfeiting. For example, provisions with respect to unfair competition may be used for acting against imitators. However, these legal provisions generally require specific unfair practices that can only be determined in the individual case. Thus, this kind of protection suffers from the problem of proof as well as from the legal uncertainty.

[Top of Page]

3. Basic Principles of Patents

As mentioned above, the subject matter of patents is in general, with the exception of a wider range of patentable subject matters in the USA, directed to technical solutions in form of devices, apparatuses, products, substances, and technical methods. In the following, some basic principles of patents are discussed.

3.1. Invention

Since patents protect technical solutions, one precondition for a patent is that an invention has been made. Inventions can only be made by persons, and not by legal entities. Accordingly, the invention is always related to a person who made it. Consequently, the right to a patent always belongs to the inventor.

However, the inventor is able to transfer the right to the patent to anybody else, for example, his employer. The transfer of the rights to the patent can be executed in different ways, for example, by a contract or according to specific legal provisions, as for example in Germany under the law relating to inventions of employees. Accordingly, it is important to determine in advance of a patent application, who is entitled to file the patent application, since applications made by persons not having the right to the patent may be or become invalid.

3.2. Exclusive Privilege

After granting or registration of the patent, the patent confers on its proprietor the exclusive right to produce, offer, or use the product or a method protected by the patent. Other individuals are not allowed to produce, offer, sell or use the subject matter of the patent. However, this exclusive privilege is mainly the right to prevent others from doing something, but not the permission for the patent proprietor for any operation. If the use of a patent infringes rights of third parties or violates the law, the patent does not create a certificate of exemption.

A patent and the exclusive privilege given thereby gives a patentee the opportunity to act against infringers by claiming omission of use, payment of damages, and providing information as to the circumstances of the infringement.

3.3. Principle of Territoriality

The exclusive privilege given by a patent is always restricted to the geographical limits of the patent. Patents can only be granted with respect to a specific territory, e.g., the territory of a state. However, as is described in detail below, some states constitute a union for filing and granting of patents based on one application for several member states, e.g., the European Patent Organization.

3.4. Period of Protection

In addition to the geographical limits, patents are also limited with respect to their duration. The exclusive right which results from the patent is the gratification or remuneration for the patentee or the inventor for disclosing the invention and making it available for the public. However, after a predetermined period of time during which the patentee can solely preserve the yield of his invention, the public should benefit from the invention. Consequently, patents generally have a maximum period of protection of normally 20 years from the day of filing. After expiration of the maximum period of protection, the patent loses its effects and everybody is allowed to use or produce the subject matter of the patent.

3.5. Ways of Establishing Patents

As a matter of principle, different ways of establishing industrial rights or protective rights exist. First, one can distinguish between rights by fact and formal rights. Rights by fact are automatically generated by means of facts, e.g., by creating artificial works in the case of copyright. Formal rights are generated by formally applying for the right at authorities and the subsequent granting or registration of the right. This group of rights comprises patents or utility patents, utility models, trademarks, and design patents.

Accordingly, in order to obtain a patent it is necessary to file a patent application with an office or authority. Patent protection cannot be obtained by simply creating something. It is also necessary to file an application. However, the further proceeding of the filing of the application differs from country to country. In some countries, it is not necessary according to national law to have the patent examined with respect to the substantive law. In these countries, patents are registered after passing formal examination with respect to the formal requirements. No examination of the preconditions for a valid patent with regard to substantive law must be carried out. Such patents are also called registered rights, since they are substantively unexamined. Contrary to that practice, in most countries, patents must pass through substantive examination before being granted. During the substantive examination, it is verified that the subject matter for which protection is sought fulfils the requirements with respect to the substantive law, i.e., novelty and inventive step, as discussed below. The advantage of examined patents is the higher degree of legal certainty with respect to the validity of the protective right. In case of substantively unexamined and merely registered protective rights or patents, the validity must be examined in case of an infringement.

For this reason, it is recommended when facing a patent to first verify that the patent was substantively examined before registration or granting. For example, petty patents or utility models (*Gebrauchsmuster*) in Germany are unexamined protective rights that are only registered after filing of the application. However, the proprietor of the utility model is allowed to claim omission of using the subject matter of the utility model due to the fact that he is in possession of the registered utility model. However, this does not provide any information on the legal validity of the utility model, which must be examined regarding validity in the case of an infringement.

3.6. Priority

Since it is conceivable that the same invention is made by different persons without their having knowledge of each other, there must be the possibility to decide who has the right to the patent in this case. For this problem, two different solutions have been developed, namely, the principles “first to file” and “first to invent”.

According to the first-to-invent principle, the inventor who made the invention first has the right to the patent. This principle is used in U.S. patent law. Due to the difficulty of proving the point of time of the invention, this leads, in the case of interference, to very time-consuming proceedings in which the point of time of the first invention is determined. The first-to-file principle avoids such problems by giving the right to the patent to the person who filed the application first. This principle applies to the patent law in most countries, except for the USA.

Accordingly, it is very important to secure an early filing date for an invention in order to ensure that the right to the patent is allocated. Although the time of filing a patent application is not as important with respect to the question who is allocated the right to the patent in the USA, the filing date is nevertheless important with respect to the defence function of a U.S. patent, since the U.S. patent application forms prior art with respect to subsequent applications of third parties.

According to the first-to-file doctrine, the date of filing determines who has the better rights with regard to an invention. For this reason, it would be important for an inventor or applicant to file the patent application all over the world (principle of territoriality) as soon as possible. Since this would cause enormous difficulties for the applicant, as early as the late 1800s, some countries made an agreement with respect to industrial property rights and established the Paris Convention. This convention provides a set of rules that are described in more detail below. One of these rules refers to the priority of inventions and that an applicant can claim the priority of a former application in a member state of the Paris Convention if the subsequent application in a foreign country is filed within a time period of twelve months. Thus, it is possible to save the right to the patent by a first application in a member state of the Paris Convention for all other member states of the Paris Convention for the following twelve months. This dramatically facilitates the procedure for the applicant, since the applicant is no longer urged to immediately file the application in all countries which are in his sphere of interest concerning the protection of his invention. Thus, the time of filing a first application in a member state of the Paris Convention is of great importance, not only for patent protection in this country, but also for further applications, almost worldwide.

3.7. Novelty

As a common principle for patents, it is a precondition for obtaining a patent that the subject matter for which protection is sought is novel with respect to prior art. Novelty means that the same invention or the subject matter of the invention is not known to the public. To determine whether the invention is already known or not, the filing date or the priority date of the first application that is claimed is used as time reference. Accordingly, the novelty test, which must be carried out during substantive examination of the patent, is carried out by comparing the claimed invention to the subject matter disclosed before the priority date of the patent application. All subject matters known or disclosed before the priority date of the patent application are called “prior art”. If the claimed invention with all its features is known from one document or disclosure of prior art, it is not novel with respect to that prior art.

3.8. Inventive Step/Nonobviousness

Besides the requirement of novelty, the claimed invention must also be based on an inventive step or, in other words, should not be obvious for a person skilled in the art in order to be granted as patent or be legally valid after registration. The requirement of an inventive step or nonobviousness postulates that a certain inventive effort should be necessary to merit remuneration of a patent or exclusive right.

In order to assess the inventive step, it must be determined whether a person skilled in the art would have been able to find the technical solution of the claimed invention without being inventive. For this purpose, the objective object for the claimed invention must be investigated in the light of prior art. If a person skilled in the art were able to easily find the claimed solution when considering the known prior art or suggested combinations thereof, the patentee or inventor should not be awarded a patent for this invention.

The person skilled in the art, who is taken as a reference for assessing the inventive step, is not a real person, but a fictive figure having average knowledge in the specific area of technology of the invention. This hypothetical skilled person is assumed to have theoretical knowledge of every prior art document or disclosure, however not to have very high other abilities. Instead of this, the skills are not more than average, rather below average.

3.9. Industrial Application

A further requirement for patents regarding substantive patent law is industrial applicability of the invention. Only those inventions that are industrially applicable should be regarded as patentable matter. However, this can be construed as given in most cases. Only specific exemptions of patentability, such as methods for treatment of the human or animal body by surgery or therapy and diagnostic methods, are deemed to be not acceptable for industrial application according to the European Patent Convention, for example. Generally, the requirement of industrial applicability does not constitute a high barrier to obtaining patent protection.

[Top of Page]

4. System of International Treaties

4.1. Paris Convention

The protection of technical inventions, as well as trademarks and other aspects of industrial property protection, was internationalized by the Paris Convention, which was established by a conference held in Paris in 1883.

The Paris Convention created the priority right that allows the priority date of a first application in a member state of the Paris Convention to be claimed for subsequent applications in other member states of the Paris Convention during a period of twelve months after the first filing. During this period actions such as filing of other applications for the same invention or publishing of the invention cannot lead to a rejection of the application claiming priority of the first application, since the later application is deemed to be “filed” at the priority date or, better, is allocated this priority date. As this priority date lies chronologically before the subsequent actions of filing another application by a third person or publishing of the invention, neither the publication nor the filing of another application can impede patentability of the application claiming priority due to missing novelty or anticipating the invention, respectively.

Besides the rules for claiming the priority of a first application in one of the member states of the Paris Convention for subsequent filings in other member states, this treaty guarantees that every member state adjudicates the same rights to applicants or inventors from abroad as to its own citizens or residents. Thus, the member states of the Paris Convention guarantee that there is no difference with respect to the protection granted for an invention regardless of whether the applicant or inventor is a citizen or resident of the member state or not.

Further, the Paris Convention guarantees minimum standards regarding formal and substantive law so that applicants or inventors can, to a certain degree, be sure that an invention that is prosecutable in their home country can also be protected in foreign member states of the Paris Convention.

4.2. Patent Cooperation Treaty

A further simplification to obtaining international patent protection was achieved by the Patent Cooperation Treaty. This treaty, which has so far (2005) been signed by about 125 member states allows the applicant to obtain protection in all member states of the PCT by filing only one application. Thus, the applicant is not forced to file many parallel applications in different countries. By the single PCT application almost the same rights can be reserved as by corresponding national applications.

However, after the international phase including search for prior art, the application must be transferred into national or regional applications for further examination and granting, as discussed in more detail below.

4.3. Regional Treaties

In addition to the PCT system, different regional alliances were formed in order to facilitate and to improve patent protection. In Europe, the European Patent Organization was founded by the European Patent Convention, and in the Eurasian region, the Eurasian Patent Office was established. Furthermore, in Africa it is possible to obtain a so-called ARIPO patent or an OAPI patent for different African member states. In the Gulf region, the Gulf Cooperation Council established rules for granting patents in force in the Gulf neighboring states. By these regional treaties it is not only possible to file an application but the examination procedure is also centrally managed. Thus, the effort for the applicant is further reduced.

[Top of Page]

5. Patent Cooperation Treaty (PCT)

The PCT created the possibility to file only one application for a great number of member states. Thus, the early stage of application is facilitated if protection is sought in numerous countries.

5.1. Application

Since the PCT application is deemed a national application in every member state of the Patent Cooperation Treaty, the application must fulfil some minimum requirements regarding formal and substantive aspects of the application.

Applicant/Inventor. The application must indicate the applicant and preferably the inventors. One of the applicants must be a citizen or resident of a member state of the PCT. The application must be filed at the national patent office of the citizen or resident of the PCT member state. In addition, it is also possible to file the application at the international office of the World Intellectual Property Organization (WIPO).

Description. The application must designate the invention and describe the subject matter of the invention. The description of the invention must include information about the technical field of the invention, the prior art the invention is based on, and an explanation of the invention by which the object of the invention and the technical solution can be understood. Furthermore, if applicable, drawings must be added to support the description.

Claims. In addition to the formal request for obtaining patent protection and the description of the invention, the most important part of the application is formed by the claims defining the subject matter of the invention. The claims contain a definition of the novel and inventive solution by denomination of essential features of the invention that are novel, at least in their combination. Two different types of claims can be distinguished: product claims directed to products as devices, apparatuses, substances and the like, and method claims covering methods of producing something or methods of carrying out a specific method for operating a machine or the like. Claims dealing with using something known for a new application are also allocated to the method claims. Usually, the claims are drafted in the so-called two-part form having a preamble and a characterizing part. The preamble contains the features that are known from the closest prior art, while the characterizing part is formed by the features that are novel with respect to the closest prior art. Preamble and characterizing part are separated by the wording “characterized by” or “characterized in that”. While it is sufficient to draft only one claim, usually several claims are filed. If there is more than one claim, the claims can be divided into main or independent claims and subclaims or dependent claims. The dependent claims refer to one or more preceding claims, in particular independent claims, so that the subject matter of the dependent claim is defined by the features of the subclaim plus the features of the claims the dependent claim is referring to. Thus, the scope of protection is basically defined by the independent claim(s) having the broadest scope of protection, while the subclaims define further options and restrictions of the claimed subject matter.

Priority. If the priority of a former national, regional, or PCT application is claimed, the filing date of this earlier application as well as the file number and country of application must be indicated.

Further Requirements. Inventions regarding biotechnology may include additional information with respect to the deposition of biological material or the disclosure of nucleotide and amino acid sequences by sequence listings.

The application can be filed in any language, which is allowed by the receiving office for this purpose. However, if the application language is not a language of publication within the PCT system, the application must be translated within a predetermined period of time.

Finally, official fees must be paid for the handling of the application as well as conducting the search for prior art.

5.2. Search

After the formal examination of the application by the receiving office, the application or a search copy of the application is transferred to the international search authority for carrying out the search for prior art.

The international search authorities are national or regional patent offices such as the European Patent Office that are designated as an international search authority by the PCT assembly. For each receiving office the PCT application is filed with, one or more international search authorities are determined.

The international search authority carries out the international search with respect to the PCT application. The search authority looks for any publication or disclosure regarding the subject matter of the application that was disclosed or published prior to the filing date of the PCT application. Further, patent applications published after the filing date of the searched PCT application but having an earlier application or priority date are also considered.

The examiner of the international search authority prepares an international search report including citations of the documents found by the search. Usually, the citations are classified by the examiner according to their relevance for the claimed invention. Documents of category X are assumed to be novelty destroying for the claimed invention, i.e., all features of one or more claims of the PCT application are disclosed in this single citation. Documents classified as category Y should be relevant with respect to the inventive step or the obviousness issue. Accordingly, the claimed invention is assumed to be obvious to a person skilled in the art because of the disclosure of this document alone or in combination with other documents. Further categories indicate additional specific types of documents.

Normally, the international search report should be established prior to the publication of the PCT application, which takes place 18 months after the filing date or priority date of the application. In this case, the search report is published with the specification of the application. This kind of document has a publication number of the type WO followed by the date of year and a serial number and closing with the expression A1. If the international search report was not prepared in time, the PCT application is nevertheless published 18 months after the filing or priority date but the publication number of this

document is different. Instead of A1, the publication number ends in A2.

If the international search report has not been published with the specification of the international application, the report is published after completion as a so-called A3 document.

In addition to the search report listing the documents of prior art, a written opinion is prepared indicating the opinion of the examiner with respect to novelty, inventive step/nonobviousness, and industrial applicability. This written opinion is mainly based on the categories of documents mentioned above. The written opinion is, however, not published with the search report, but is confidential for a period of 30 months after the filing or priority date of the PCT application.

5.3. Preliminary Examination

In addition to the search for prior art an international preliminary examination is carried out on request of the applicant. The request or demand must be made within three months after the transmittal of the international search report and of the written opinion delivered with the search report to the applicant or 22 months after the priority date, whichever period expires later.

Since during the search, a first opinion regarding novelty, inventive step/obviousness, and industrial applicability was established in the written opinion, the preliminary international examination is based on this written opinion. However, the applicant has the opportunity to file a reply including arguments supporting novelty, inventive step, and industrial applicability or to file amendments to the claims in order to obtain a positive preliminary international examination report.

The preliminary international examination report prepared after finishing of the examination is, however, not binding for the selected states for which the preliminary international examination was carried out. The selected designated states are allowed to make their own decision with respect to novelty, inventive step, and industrial applicability in the national or regional phase following the international phase of the PCT application. Accordingly, the importance of the international preliminary examination has strongly declined.

5.4. Entry into National/Regional Phase

As indicated above, the international PCT application merely includes filing of the application but does not lead to a granted patent at the end of the international phase. Therefore, the PCT application must be transferred to national or regional applications at the end of the international phase. The international phase is completed, for most member states of the PCT, 30 or 31 months after the priority or filing date of the PCT application. For some countries, the former 20/21-month time limit still applies. Within the time limit for entry into the national or regional phase (20/21- or 30/31-month time limit), the applicant must furnish a copy of the international application and a translation thereof, if necessary, and must pay the national fee to each designated office or member state. Each national or regional office is allowed to carry out an additional search for prior art and to examine the application according to national requirements. Therefore, the main advantage of a PCT application can be seen in the fact that the applicant gains time and information for deciding whether it is appropriate to proceed with the application in specific countries or states. With the search result, the applicant receives information about prior art for assessing the probability of being successful in the prosecution of the patent application regarding novelty and inventive step.

[Top of Page]

6. European Patent Convention

As outlined above, the European Patent Organisation is a regional federation for facilitating obtaining patent protection by filing and granting of only one application and patent, respectively, for various European countries.

6.1. Member States

On January 1, 2005 the European Patent Organization had 30 member states (http://www.european-patent-office.org/epo/members.htm):

AT	Austria	IE	Ireland
BE	Belgium	IS	Iceland[a]
BG	Bulgaria[b]	IT	Italy
CH	Switzerland	LI	Liechtenstein
CY	Cyprus	LT	Lithuania[c]
CZ	Czech Republic[b]	LU	Luxembourg
DE	Germany	MC	Monaco
DK	Denmark	NL	Netherlands
EE	Estonia[b]	PL	Poland[d]
ES	Spain	PT	Portugal
FI	Finland	RO	Romania[e]
FR	France	SE	Sweden
GB	United Kingdom	SI	Slovenia[f]
GR	Hellenic Republic	SK	Slovakia[b]

HU Hungary[f] TR Turkey

For information on the EPC accessions see:
[a]http://www.european-patent-office.org/news/info/2004_10_10_e.htm
[b]http://www.european-patent-office.org/news/info/2002_05_08_e.htm
[c]http://www.european-patent-office.org/news/info/2004_10_11_e.htm
[d]http://www.european-patent-office.org/news/pressrel/2004_03_01_e.htm
[e]http://www.european-patent-office.org/news/info/2002_12_19_e.htm
[f]http://www.european-patent-office.org/news/pressrel/2002_12_02_e.htm

Further, the European Patent can be extended to Albania, Bosnia and Herzegovina, Croatia, Latvia, and Former Yugoslavian Republic of Macedonia and Serbia Montenegro, although these countries are not yet members of the European Patent Organization.

6.2. Application Proceedings

Similar to the PCT proceedings, the European patent application makes it possible to obtain the same effect as different national applications by a single application. However, different to the PCT system, the European Patent Application leads to the granting of a patent for the designated member states of the European Patent Organization. Thus, contrary to the PCT application, no further examination proceedings must be carried out in the single designated countries. Instead, an examined patent can be obtained for all designated member states by one application and one examination proceeding.

Accordingly, it is very convenient for an applicant to obtain patent protection in different countries in Europe, since only one examination proceedings for all these states must be conducted instead of passing the examination in various national proceedings.

6.2.1. *Application*

The European patent application may be filed directly with the European Patent Office in Munich or at its branch in The Hague, if this is not precluded by national legislative regulatory provisions ensuring that secret inventions are not communicated abroad. Further, the European Patent Application may be filed with the national patent offices or industrial property offices, if the national law so permits.

The European patent application may be filed in any language of the member states if the applicant is a citizen of such a member state or has his residence in the member state, but a translation into one of the official languages of the European Patent Office has to be filed within three months. The official languages of the European Patent Office are German, English, and French.

The application must include a request for the grant of a European patent indicating the identity of the applicant who is applying for the patent. Further, a designation and description of the invention must be added. In addition, patent claims defining the subject matter of the claimed invention as well as a summary of the invention for easier information of the public and facilitating the search for prior art should be submitted. Regarding drafting of the claims, see Section 5.1 with respect to the Patent Cooperation Treaty, since the requirements are almost identical. If appropriate, drawings can be added to visualize the technical features of the invention, which very often can be described much easier by drawings than by words. However, the function of the description of the invention and of the drawings is merely to support the claims and to enable a person skilled in the art to carry out the invention and to interpret the claims if there are uncertainties as to the meaning of the wording. In the end, the claims define the scope of protection given by the subsequently granted patent.

Although the European Patent Organization is not a member of the Paris Union but a special agreement within the meaning of Article 19 of the Paris Convention, it is possible to claim the priority of applications filed in member states of the Paris Convention. In a similar way, a European patent application can form the basis for a priority claim in any other member state of the Paris Union.

For inventions in the field of biotechnology, information about the deposition of biological material or sequence listings may be added.

6.2.2. *Examination Procedure*

Formal Examination and Search. After filing of the application with the European Patent Office, the application is formally examined in order to check whether the formal requirements are fulfilled, e.g., payment of fees, etc. When the application has passed formal examination, a search for prior art is conducted. The search is directed to documents or disclosures with respect to the claimed invention all over the world in the period before the filing date of the European patent application.

Publication. Eighteen months after the date of filing or the priority date if priority is claimed, the European patent application is published. Together with the publication of the specification, the search report is published as well. Only if the search report has not been finished by the publication date of the application, a subsequent publication is effected. The relevant documents are similarly designated by the endings A1, A2, and A3, as outlined above for PCT applications.

Designation and Substantive Examination. Within six months after the publication of the search results, the applicant must file a request for substantive examination and pay the designation fees for the countries which should be designated for the application.

After receiving the request for substantive examination, the European Patent Office or the examiner in charge starts the substantive examination, i.e., the claimed invention, as defined in the claims of the application, is compared to the documents of prior art found during the search. The comparison of the application and the prior art is carried out with

respect to novelty and inventive step.

Novelty. The novelty test comprises a comparison of the disclosure of a single document with the features of the respective claims. If every feature of a claim is disclosed in a single document of prior art, this prior-art document is novelty-destroying for the invention claimed in this claim.

Inventive Step. Regarding the inventive step, the examiner must assess whether the claimed invention is rendered obvious for a person skilled in the art by one or more documents or disclosures of prior art. For this purpose, the European Patent Office established the so-called problem/solution approach, which defines a method of assessing the inventive step.

According to the problem/solution approach, the closest prior art must be defined first. This can be prior art which differs only in a few features or is directed at the same problem in the same field of technology. By the differences between the closest prior art and the claimed invention, an objective object of the claimed invention is defined, which results from the advantages given by the claimed invention compared to the closest prior art. Thus, the problem of the problem/solution approach is found.

Starting from the objective object of the claimed invention, it is asked whether a person skilled in the art would have to be inventive in order to arrive at the claimed invention or whether the claimed solution was proposed by prior art, namely, the document of the closest prior art and/or additional documents. Usually, two or three documents of prior art can be combined in order to find the claimed solution of the invention, since the combination of more than two or three documents can be regarded as inventive by itself. Regarding obviousness, it is important that a person skilled in the art should not only have the possibility to arrive at the claimed invention by knowledge of prior art documents, but rather would have been able to find the solution without any additional considerations. Therefore, it is usually necessary to have a hint or proposal in the prior art pointing to the claimed solution. It is not sufficient proof of obviousness if the person skilled in the art only could have arrived at the claimed solution. It is rather necessary for obviousness that a skilled person would actually have arrived at the alleged invention because of suggestions or proposals in the prior art.

Further Requirements for Patentability. During the substantive examination, the other requirements for patentability are also investigated. Thus, the examination also covers the requirement of industrial applicability as well as the verification whether the claimed invention is a patentable invention according to Article 52 EPC and does not fall under the exceptions of patentability according to Article 53 EPC.

Patentable Subject Matter. According to Article 53 EPC, patents cannot be obtained for inventions which are contrary to “ordre public” or morality or refer to plant or animal varieties or essentially biological processes for the production of plants or animals with the exception of microbiological processes or the products thereof.

Furthermore, according to Article 52 EPC, the following shall not be regarded as inventions within the meaning of the European Patent Convention:

- Discoveries, scientific theories, and mathematical methods
- Aesthetic creations
- Schemes, rules, and methods for performing mental acts, playing games, or doing business, and programs for computers
- Presentations of information

as well as claimed inventions relating to such subject matters or activities as such. Furthermore, methods for treatment of the human or animal body by surgery or therapy and diagnostic methods practiced on the human or animal body shall not be regarded as inventions within the European Patent Convention.

The exceptions named above only apply to the extent to which the claimed invention relates to these subject matters or activities as such. This leads to the problem of defining the meaning of “as such”.

Software and Computer Programs. Especially in the field of computer technology and computer-implemented inventions which use for reduction to practice computer or computerlike processors, problems arise with respect to the question what are patentable inventions and what are inventions which are excluded from patent protection because they refer to programs of computers as such.

According to the practice of the European Patent Office, it is not sufficient to only use a known computer system for carrying out an invention which otherwise is nontechnical, e.g., a business method, to obtain a patent. Moreover, it is necessary that the invention contains a specific technical aspect which contributes to novelty and inventive step. Thus, business methods using only computer programs or computer systems which are known per se for carrying out the invention are not patentable within the European Patent System if no additional technical effect is present.

Procedural Aspects. During the examination proceedings, the applicant is allowed to amend the claims, in particular to restrict the claims over the prior art. The applicant can also amend the specification, but he must make sure that the application is not extended beyond the content originally filed.

During the examination procedure, it is also possible to carry out oral proceedings if requested by the applicant. This gives the applicant the opportunity to make his arguments in an oral speech instead of writing.

At the end of the examination procedure, the European patent is either granted or rejected. If the patent is granted, the applicant must validate the European patent in the designated states by filing a translation of the patent specification into the corresponding national languages and by paying the national fees.

6.3. Opposition Proceedings

According to Article 99 EPC, anyone can file an opposition against a European patent within a time period of nine months after the publication of grant. This gives the public or third parties the opportunity to present observations as to the patentability or validity of the patent. In particular, this is the opportunity to present additional prior art which was not considered during the examination procedure, because it does not exist in form of a written document, for example. In particular, the opponent has the possibility to assert a public use of the invention which was not known to the examiner for practical reasons.

The opposition must clearly point out the reasons why the opponent is of the opinion that the granted patent should be revoked. Furthermore, an opposition fee must be paid.

After filing of the opposition, the patentee has the possibility to argue against the opposition and/or to file amended claims in order to restrict the patent over the prior art that became known due to the opposition.

At the end of the opposition proceedings, the patent is maintained, revoked as a whole, or partially revoked.

6.4. Appeal

An appeal can be lodged against the decisions of the European Patent Office during the examination procedure, i.e., rejection of the patent or partial rejection of the patent application, and against decisions made in the opposition proceeding, in order to have a second instance to decide about the case. The appeal must be submitted in writing at the European Patent Office within two months after notification of the decision and a written statement setting out the grounds of appeal must be filed within an additional term of two months. Furthermore, the appeal fee must be paid.

The decisions of the Boards of Appeal of the European Patent Office are final decisions and cannot be challenged.

[Top of Page]

7. U.S. Patents

As outlined above, there are some differences between U.S. and European patent practice.

First, the system for determining who has the right to the patent is different. While according to European or Japanese practice, the first application makes the running, in the USA, the first inventor has the right to the patent. Therefore, it is necessary to conduct interference proceedings by which it is determined who is the first inventor.

Another important difference is that different subject matters can be protected by a patent compared to European practice. While in Europe or Japan, business methods are not patentable, this is possible in the USA. After the Supreme Court's decision Diamond versus Chakrabarty, the variety of patentable matter was widely extended.

Another particularity of U.S. patent law is the grace period of 12 months before the U.S. application. During this grace period, publication of the invention is not detrimental to novelty.

Opposition proceedings are not provided for in the U.S. patent law. Instead, reissue and re-examination proceedings exist. While the reissue proceedings are directed to a correction of faults in a granted patent, the re-examination allows a new examination with respect to the preconditions of patentability. However, only patent documents and written documents can be used for supporting missing patentability. Furthermore, a third person requesting re-examination is not a party of the proceedings. Therefore, a new proceeding, namely, inter-partes re-examination, was established which is comparable to the opposition proceedings according to the European Patent Convention.

[Top of Page]

8. Other National Law

Since every national patent act comprises specific national rules in addition to or sometimes in place of the principles of patent law outlined above, it is impossible to consider all these particularities of national law in the present article. For this reason, it is necessary to contact specialists in case of concrete and precise questions with respect to patent issues in order to consider all national aspects and to avoid loss of rights. However, the principles of patent law shown above are valid in most countries or the national provisions are similar to those described for the PCT and the EPC, so that it should be possible to understand the different national procedures bearing the information given above in mind.

[Top of Page]

References

General References

1. H. Götting, *Gewerblicher Rechtsschutz*, 6. Auflage, Auflage, CH Beck, München 1998.
2. K.-D. Rippe, *Europäische und internationale Patentanmeldungen, Praxisleitfaden*, 3. Auflage, Carl Heymanns Verlag, Köln, Berlin, Bonn, München 2003.
3. R. L. Mayer, *Das US-Patent*, 3. Auflage, Carl Heymanns Verlag, Köln, Berlin, Bonn, München 2003.
4. U. Singer, D. Stauder. *The European Patent Convention, A Commentary*, 3rd ed., Carl Heymanns Verlag, Köln, Berlin, Bonn, München 2003.
5. *Manual for the Handling of Applications for Patents, Designs, and Trademarks Throughout the World*, Octroibureau Los

en Stigter, Amsterdam, Update Service.
6. *PCT Applicants Guide*, World Intellectual Property Organization, Geneva, Switzerland, Update Service.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Solids Technology, Introduction

For unit operations in solids technology see → Particle Size Analysis and Characterization of a Classification Process; → Crystallization and Precipitation; → Drying of Solid Materials; → Size Reduction; → Spraying and Atomization of Liquids; → Size Enlargement; → Solids Handling.

Klaus Borho, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Friedrich Richard Faulhaber, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Reinhard Polke, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Peter Thoma, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Objectives of Solids Technology . . . 1
2. Stages of Implementation 3
2.1. First Stage: Formulation of Working Hypotheses 3
2.2. Second Stage: Testing of Working Hypotheses 6
2.3. Third Stage: Development of Processing Strategy 6
2.4. Fourth Stage: Verification of Processing Strategy 9
3. Measurement Methods 9
3.1. Measurement of Product Properties 10
3.1.1. Application-Related Product Properties 10
3.1.2. Production-Related Product Properties 10
3.1.3. Measurement of the Dispersity Variables 10
3.2. Measurement of Product-Specific Characteristics 11
3.3. Measurement of Process Parameters 11
4. Research and Development 11
5. References 11

1. Objectives of Solids Technology

Solids technology is concerned with the production of disperse solid products. Characteristically, the properties of solid products depend not only on chemical composition but also on the state of dispersion. The significance attached to the objective of systematically adjusting product characteristics, often referred to as product engineering or formulation, is indicated by the fact that almost all studies in recent years describe this as an important task for the future [1], [2]. Production processes are composed of individual process steps (unit operations) which include all necessary activities from forming the solid up to packaging the product for sale. Other articles of this encyclopedia deal with the scientific principles, the various options in terms of equipment, and the scale-up methods for individual unit operations. The significance of particle technology becomes clear from a survey by Davies [3]: about 60 % of all of DuPont's chemical products are disperse substances, that is, powders, granular materials, dispersions, slurries and pastes. About half of the remaining 40 % pass through a disperse state in the course of their production. The situation for the European chemical industry is estimated to be similar.

This introduction to particle technology is intended to demonstrate an approach, based on the underlying principles of unit operations, to the design of a solids process. This involves optimization of the interaction of the various unit operations with a view to the goal of solids technology, namely, the economical preparation of products having specific characteristics in terms of production and application.

The principles of particle technology were only developed in the 1900s by the pioneers of process engineering and their students in various countries. Rumpf [4] introduced the concept of the "property function" which identified the relationship between the state of dispersion of a product (particle size, crystal modification, etc.) and its properties (quality: color strength, activity, etc.).

The production of disperse solid products is characterized by the generation and repeated modification of disperse states. In the following, the term "disperse state" or "dispersity variable" refers not only to particle size, but also

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_01

includes all the variables which characterize the product state, such as the morphology of the particles, interfacial characteristics, and agglomeration strength.

A typical example of a solids processing chain consists of the unit operations crystallization, filtration, drying, and size conversion. Each unit operation modifies the physical state of the product within the processing chain (Fig. 1 A). Crystallization changes the state from solution to suspension (→ Crystallization and Precipitation); filtration changes it from suspension into filter cake (→ Filtration); drying converts the filter cake into a powder or lumpy product (→ Drying of Solid Materials); and size conversion converts powders into granular materials (→ Size Enlargement) or lumpy product into powder (→ Size Reduction).

At the same time the state of dispersion of the solid changes (Fig. 1 B). The original suspension contains individual particles or loose floccules. On filtration, these are compacted significantly and form moist, loose agglomerates. The drying process produces considerably stronger agglomerates, which can be formed into larger secondary agglomerates or reduced in size in the size conversion stage.

The state of dispersion affects the production properties of a product (Fig. 1 C). It constitutes a crucial factor that determines whether specific processing operations are possible or defines the cost of carrying out these operations. Production properties can be characterized by the following terms:

Crystallizability
Filterability
Washability
Dryability
Pumpability
Storability in bins

The user of solid chemicals requires the end products to have well-defined properties which, as has already been shown, depend on the state of dispersion of the material. Application-related properties such as color strength of dyes or activity of pharmaceuticals are deployed only if further characteristics are present:

Wettability
Dispersibility
Rate of dissolution

These are complemented by additional application-related properties required by customers who further process the products:

Resistance to attrition
Dusting
Flowability

The ultimate objective of modifying the state of dispersion throughout the production process is to develop specific application-related product properties. Therefore, the production process affects the application properties both in intended, and quite often unintended, ways. Many application properties become apparent only in the end product, although they originate at various stages throughout the production process.

Just to illustrate the concepts of these properties, which initially may seem rather abstract, consider the following examples:

- A tablet which fails to dissolve or cannot be digested and therefore does not release its active ingredient
- A paint whose pigment settles to the bottom of the can as a solid layer
- Powdered instant coffee which clumps together on the spoon
- Crystals which cannot be filtered
- A powder which cakes in bins

Products with these properties clearly do not meet the ultimate objective of the process. A process engineer must realize that choice of equipment and process conditions will affect the physical states of the products, which in turn are responsible for the application properties.

2. Stages of Implementation

The various stages in the development of a process are shown in Figure 2. In the first stage the property profile is defined. A concept covering the interrelationships of product properties and dispersity variables is formulated and this concept is then used to derive working hypotheses. In the second stage the working hypotheses are subjected to experimental tests on a laboratory scale. In the third stage the process steps (unit operations) must be selected and worked out in detail so as to achieve an overall optimal processing chain. In a final stage, the process is validated, if necessary, on a pilot scale.

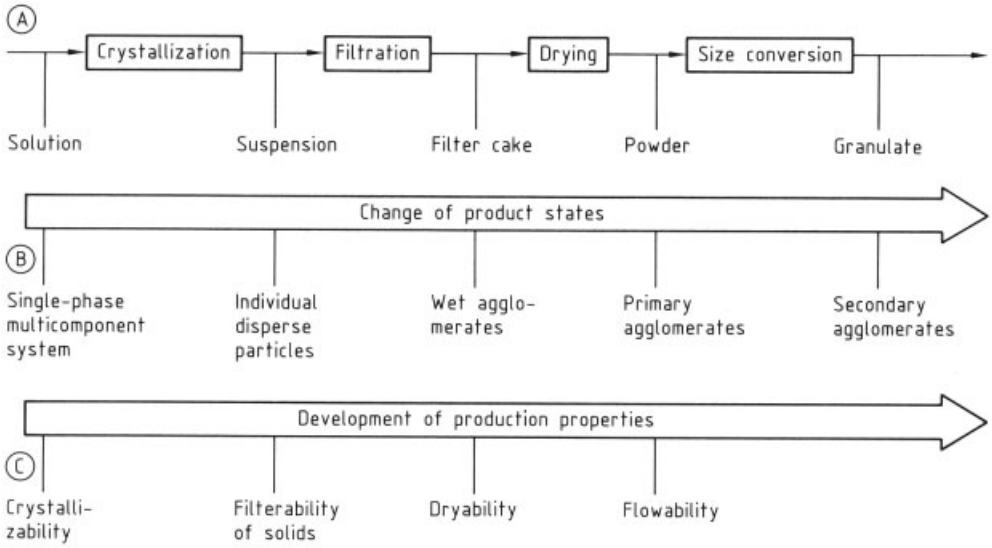

Figure 1. A) A typical solids technology processing chain; B) Change of product states during the processing chain; C) Development of production properties during the processing chain

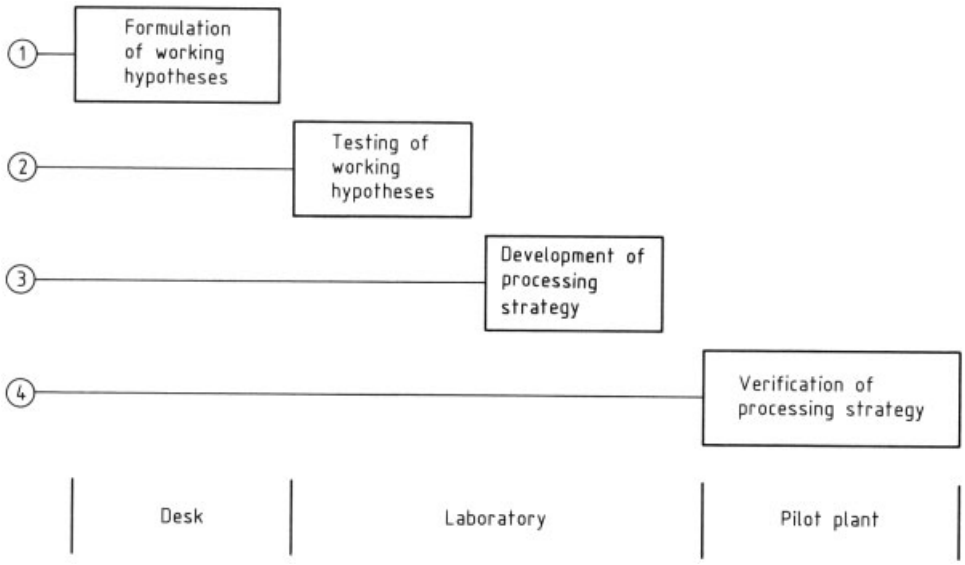

Figure 2. Four stages in process development and scale of experimental work

2.1. First Stage: Formulation of Working Hypotheses

The formulation of working hypotheses involves a number of steps (Figure 3). In developing a process, the first step is to specify the objective, that is, the characteristic profile of the end product. This observation may seem trivial at first, but in practice it often proves difficult to define a final specification of the characteristic profile at the start of process development. This is due to the fact that the desired product properties, in almost every case, represent a compromise between what is possible in terms of process technology and what is acceptable in terms of manufacturing costs. Final specifications are therefore arrived at only during the development process.

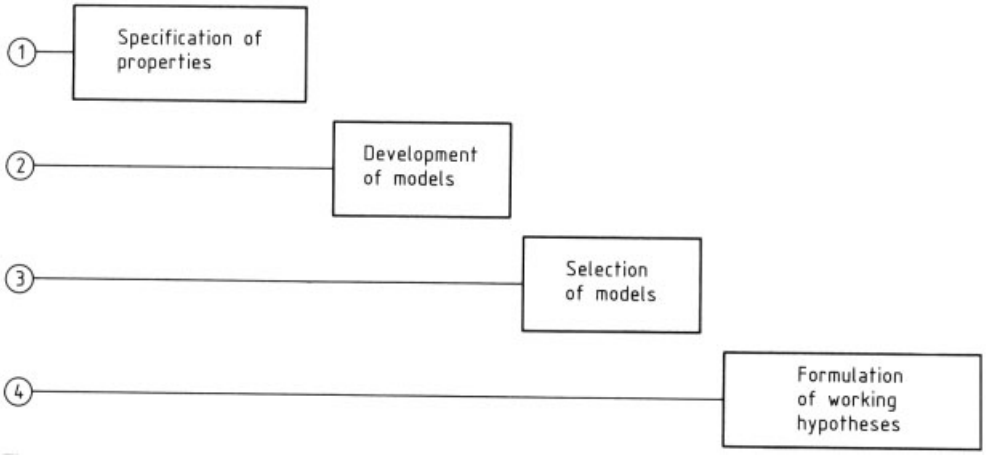

Figure 3. Four steps in the formulation of working hypotheses

Once the characteristic profile is available, Step 2 requires a conceptual model to be developed to define in what state the end product must be if it is to give the desired final properties. In many cases the desired properties can be reduced to simple physical principles. Thus a description of the required product condition which at least points in the right direction is possible [4], [5]. This state must be achieved by means of process engineering. This frequently results in a number of conceptual models and numerous parameters that act by means of physical mechanisms.

Given the high costs involved, empirical testing by measuring all the parameters of all the concepts produced is not feasible. Therefore, this only leaves the option of selecting those physical variables, on the basis of a chosen concept, which are likely to have the most significant effect on the desired product properties. Thus, various working hypotheses are formulated and then tested experimentally. The iterative cycle – model concept – working hypothesis – verification – may have to be repeated several times before a working hypothesis can be accepted. The point of the procedure described is not to confirm models, but to use models to enable physical parameters to be found rapidly and thus to obtain a solution to the overall problem. To select the right model is therefore of key importance. The process engineer must have been trained in drawing up models and must be fully conversant with the available concepts.

This procedure may seem over complicated and laborious. Fortunately, the route described can be truncated in practice. In the most favorable case the relationships between product properties and physical variables are known. Another frequently effective approach is to make use of extensive knowledge acquired in the academic sector by resorting to existing conceptual models.

The following example is presented to clarify these ideas. For the sake of simplicity, it involves a single product property as the target variable. Initial laboratory samples of a new product show that the product has the required activity but that its flowability is very poor. The aim is therefore to improve the flow properties. Step 1 in Figure 3 is thus defined.

The qualitative relationships between flow properties and physical variables have been addressed in fundamental research which shows how flow properties depend on the structure of the bulk material and the adhesion mechanisms acting between particles [6]. Therefore, a conceptual model for the first processing stage (Step 2 in Figure 3) already exists in the present example. Only the adhesion mechanisms can be affected by the process to a significant extent. These can be classified into various groups:

1) Adhesion mechanisms without material bridges
   a) Van der Waals forces
   b) Electrostatic forces
   c) Mechanical interlocking

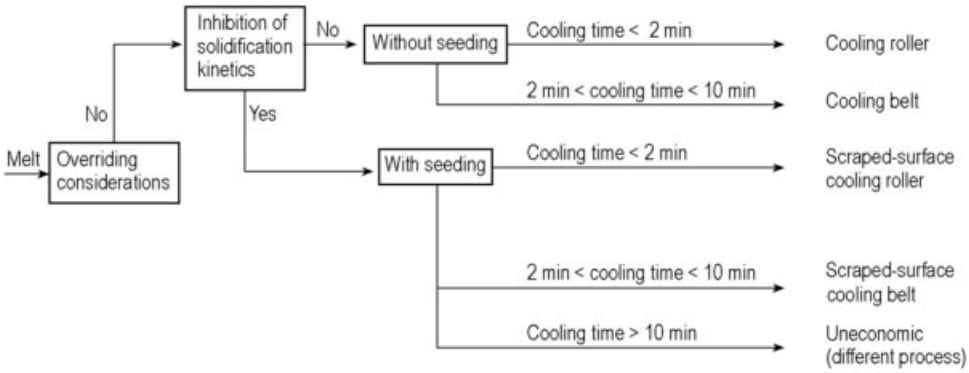

Figure 4. Melt conditioning: cold plate test

2) Adhesion mechanisms with solid bridges
   a) Crystallization of dissolved substances
   b) Recrystallisation of dissolved substances
   c) Sintering processes
   d) Chemical reactions
3) Adhesion mechanisms with liquid bridges

In the following discussion, two specific adhesion mechanisms which influence the flowability have been selected (Step 3, Fig. 3): Van der Waals forces (1a) and crystallization of dissolved substances (2a). Physically relevant variables for these two mechanisms are listed in the following:

1 a) Van der Waals forces
- Particle-size distribution
- Particle shape
- Interparticle distance
- Interaction constant
- Hardness

2 a) Crystallization of dissolved substances
- Particle-size distribution
- Solubility of product and impurities
- Moisture transport

This leaves the further task, in Step 4, of selecting those variables which have a suitably strong effect on the property in question. In the present case the following working hypotheses are formulated:

1) The poor flow behavior is caused by Van der Waals forces. Possible remedies:
   a) Larger particles
   b) Different crystal modifications (hardness or shape)
2) The poor flow behavior is due to solid bridges formed by crystallization of dissolved substances. The flow behavior can be improved by reducing the impurity content.

The severe restriction to a few adhesion mechanisms and hence in turn to a few physical variables highlights the hypothetical nature of this procedure. What is determined in this stage is therefore not the complete property function as defined by Rumpf, but merely the essential parameters.

The next stage requires experimental tests to decide which hypotheses should be rejected and which can be retained for further process development.

2.2. Second Stage: Testing of Working Hypotheses

Samples have to be prepared which will ensure that the postulated main parameters are varied over a wide enough range. It is necessary for the required product property to be quantified as far as possible as a function of the physical variables. On the one hand this assumes experimental techniques which permit straightforward preparation of such samples, and on the other hand it assumes effective measurement methods which are capable of quantifying the relationships.

The large number of samples required means that this procedure can only be carried out on a laboratory scale. Only at this level is it possible for sufficiently varied samples to be produced

rapidly at reasonable cost in amounts which, as experience has shown, should not exceed 1 kg. At the same time, the laboratory apparatus must be able to set or to modify the physically relevant variables of the product. The laboratory apparatus should not, in this context, be seen as a scaled-down version of production equipment.

This stage can be clarified further, again taking the example of flow properties. To test the working hypothesis 1 (Van der Waals forces, particle size, crystal modification), samples containing different particle sizes and different crystal modifications must be prepared. This can be achieved by classifying or by controlling particle size during crystallization. Different crystal modifications are obtained by maintaining special crystallization or drying conditions. In this way, the relationship between flowability, particle size and crystal modification is established. Working hypothesis 2 (solid bridges, impurities) can be tested by more thorough washing of the filter cake or by reducing the concentration of impurities in the mother liquor.

This stage may be concluded with the result that working hypothesis 1 is accepted and working hypothesis 2 should be rejected. The studies of hypothesis 1 provide information on the minimum particle size and crystal modification required to achieve the desired flowability.

2.3. Third Stage: Development of Processing Strategy

The result reached in the previous stage is that the relationship between the properties and the state of dispersion of the end product has been determined. At this stage it is then necessary to establish which process steps and process parameters will lead to the desired state of dispersion. The experimental testing of the working hypotheses will have provided some relevant partial data, since some initial experience and experimental values for production properties will have been gathered in the process of preparing the laboratory samples. Additional measurements of material-specific characteristics, within the relevant boundary conditions, allow the range of feasible technologies to be considerably reduced. Consequently, only a few variants remain which require more thorough empirical testing. This test procedure must provide all the necessary or supplementary data required to ensure that the process goal of achieving specific product properties can be attained economically. Initially the main effort will be directed at those unit operations which have a dominant effect on the desired properties.

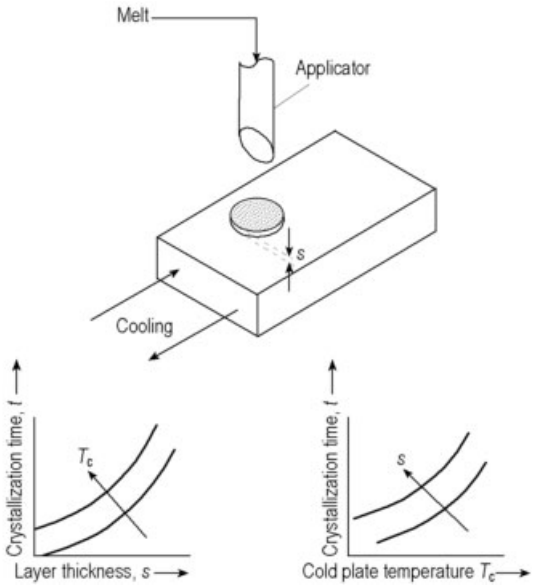

Figure 5. Laboratory cold plate (melt conditioning)

In some cases a product can be defined so effectively by means of material-specific characteristics, that possible solutions in terms of equipment can be represented in the form of decision trees. Figure 4 illustrates this, in a simplified manner, for the example of melt conditioning. The implementation options, in terms of equipment, depend to a great extent on the solidification behavior of the melt. This behavior can be described physically by the crystallization of the melt, involving the individual processes of nucleation and growth, which depend on a number of parameters. The effort involved in describing these individual processes would be excessive for the purpose of equipment selection. A simple laboratory apparatus for obtaining the necessary data is shown in Figure 5. Varying the parameters melt temperature, cold plate temperature, layer thickness and seed conditions provides information on the cooling times with the corresponding layer thicknesses, whether nucleation is inhibited (i.e., seeding is required), and

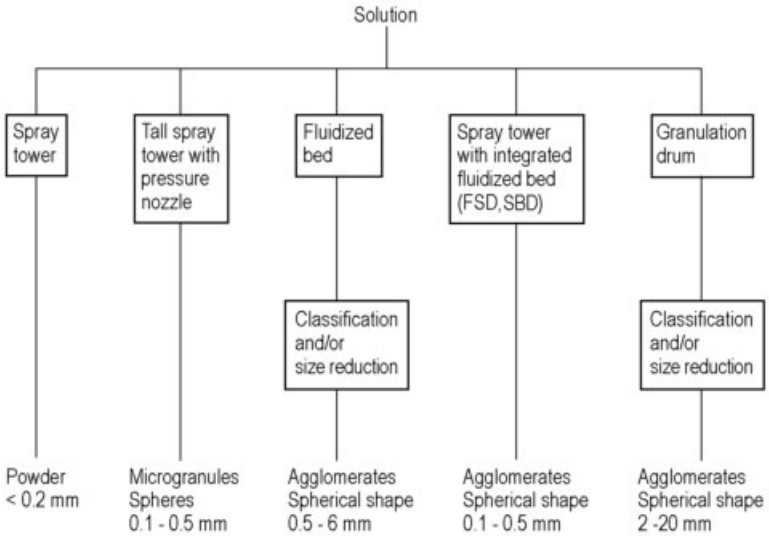

Figure 6. Product from solution

on the adhesion behavior of the layer. The decisions required on the tree in Figure 4 can therefore be made.

The selection of dryers is based on the required end product properties as shown, for example, in Figure 6 for a solution as the starting material. In this case the required end product properties have a very significant impact on the choice of process and equipment. If the studies of the dependence of end product properties on the various parameters show that only a porous morphology and a particle size of less than 500 µm allow the required dissolution rate to be achieved, then an FSD (see Fig. 6) process might be the appropriate solution.

To specify process parameters and to aid in deciding which equipment to use if a number of alternatives are available, one requires a large number of material-specific characteristics:

Moist product
- Consistency
- Moisture
- Meterability
- Formability
- Sprayability

Product during drying
- Sticking behavior
- Encrustation behavior
- Reaction with drying gas
- Crystal modification conversion
- Temperature/time stability
- Sorption isotherm

Safety
- Flash point
- Dust explosiveness
- Minimum ignition energy
- Electrostatic chargeability

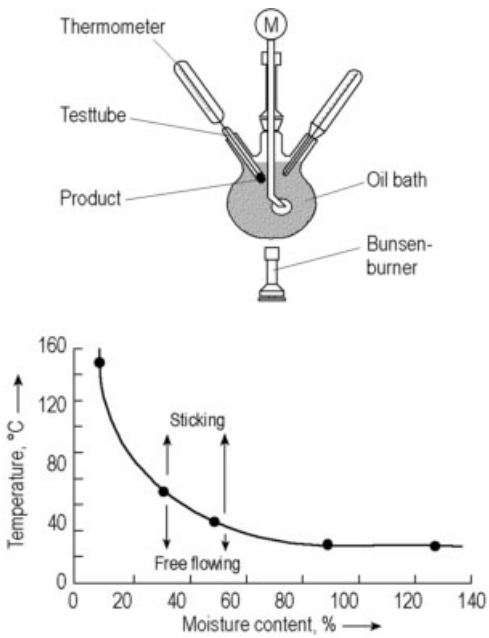

Figure 7. Measurement of sticking behavior

One example worth discussing is that of sticking behavior. The flow properties of products change as a result of changes in the adhesive forces acting between them, as discussed in Section 2.1. Particularly important for drying is the behavior of the products as a function of moisture content and temperature. This behavior can be determined very simply with the laboratory equipment shown in Figure 7. A sample whose moisture content has been adjusted beforehand is slowly heated in a test tube. Its temperature is monitored by means of a thermometer, and the product is agitated. Visual inspection reveals at what point the flow behavior and the adhesion behavior on the wall change. This information is important for specifying temperatures [7].

Note that the overall optimum is not necessarily represented by the sum of individually optimized process steps. This means that the entire process chain has to be economically optimized. This point may again seem trivial, but it can require possible procedural variations to be studied in parallel or the same process chain to be iterated repeatedly. In extreme cases, new working hypotheses will have to be introduced even at this stage or objectives may have to be modified, i.e., the stages described above must be repeated.

The data obtained should contain information on how sensitive the process is to any deviations in the individual unit operations, with a view to possible scale-up. The development stage is finally complete with a preliminary process flow diagram including a measuring and control plan.

At this stage, too, rapid and cost-effective experimental techniques are required, and these must be sufficiently sophisticated that the products have the same properties as industrially produced products and that the parameters measured allow sufficiently accurate scale-up. This is too costly on the pilot scale, given the long timescale required for setting up the equipment and the handling of relatively large quantities of product. These objectives must therefore likewise be addressed on a laboratory scale. The fundamental research carried out over the last decades will, on the whole, allow these objectives to be met.

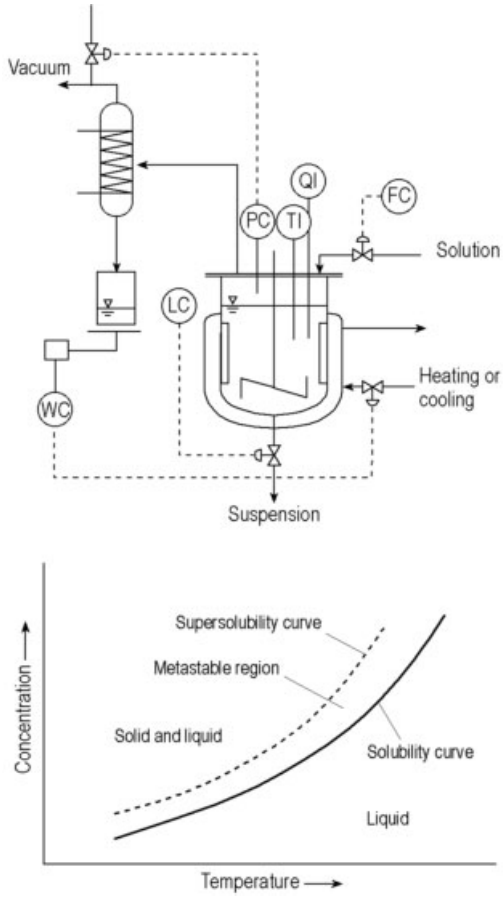

Figure 8. Measurement of solubility (crystallization)

The result of applying this stage to our example of improving the flow properties could be that the most promising process chain is crystallization – filtration – drying. The process objective of flowability should be attained if the desired crystal modification and/or a previously required particle size can be achieved during crystallization and retained during subsequent processing. The detailed design of the crystallization operation involves choices between batch or continuous cooling, and vacuum or evaporation crystallization. Figure 8 shows the corresponding laboratory equipment. The centerpiece of the equipment is a 1-L stirred vessel made of glass, which can be heated and cooled via a double jacket. At the bottom of the vessel there is a valve through which representative product samples can be drawn off during and after a crystallization experiment. More recently, fast in-line sensors have become available for

this purpose (see QI in Fig. 8). The particle size distribution is determined by the processes occurring during nucleation and growth, which in turn are effected by supersaturation and mechanical stresses to which the crystals are subjected. The sensitivity of these parameters in terms of the result can be studied in the laboratory equipment shown.

The other unit operations involved are studied in a similar manner. In the present example it is very important to consider the movement of product in and between the process steps, as attrition or size reduction is possible at every point, thereby potentially compromising the process objective.

Optimization involves selecting the technology and the process conditions of each unit so as to ensure that the process objective of flowability is likewise achieved reliably and most cost-effectively. To do this, a sensitivity analysis is carried out to establish what tolerance limits must be observed for the operational states to ensure that the set objective is indeed achieved. A commercial assessment of various options can be based on the costs for energy, ancillary materials, repairs, and personnel, which can be derived from the analysis of the process, and on methods for estimating the capital expenditure.

Some of the costs of the process design stage in the laboratory may be saved if the overall process or individual steps of the process can be simulated, as is the case at present for processes involving fluids. Simulation would then allow relatively complex process structures to be represented by means of simple laboratory experiments, and the product condition to be predicted. This would permit experiments and optimizations to be carried out on a computer and consequently provide for more effective process development. This is one of the reasons why many research groups are active in this field.

2.4. Fourth Stage: Verification of Processing Strategy

The process development steps so far will not, as a rule, have addressed every aspect, e.g. the effects of impurity enrichment and the effect of recycle streams over extended production runs, the effects of different mixing kinetics and residence times, heat and mass transfer, temperature, and shear gradients. It is precisely in disperse systems that these aspects can be particularly significant. If the process includes such features, they are studied on a pilot plant scale and the processing strategy is verified. This then means that all the information required for definitive process development will have been provided.

3. Measurement Methods

The procedure as described for developing solids technology requires informative measuring techniques. If a process is to be characterized it is necessary to measure both product properties and process conditions. This is shown in more detail in → Process Control Engineering with reference to the phase model described there.

- The product is characterized by the *product properties* and/or those *dispersity variables* which effect them.
- Selection and design require *product-specific parameters* which characterize the product behavior within the unit. These include safety characteristics and toxicological properties.
- To ensure that the eventual plant will operate reproducibly and reliably, the process conditions must be specified, i.e., the essential *process parameters*, their nominal values and tolerance limits have to be determined (sensitivity analysis).

The development of novel processes and the causal analysis in the event of product properties deviating from the target values is more likely to require and justify the comparatively costly measuring methods for the physical variables than would be the case for production monitoring, where measurements of auxiliary variables or production-related properties are often sufficient.

3.1. Measurement of Product Properties

3.1.1. Application-Related Product Properties

The desired characteristics of intermediates or final products of a process are ultimately ob-

tained by measurements of the product properties. Properties summarize the effect of a number of variables; this applies both to application-related and production-related properties. Examples include the Jenike shear test to assess flowability, or measurement activity to assess the efficacy of a catalyst.

Most of the application-related properties are evaluated by means of product-specific test methods, such as the mechanical strength of concrete, the color strength of pigments, and the gloss resulting from matting agents. Other properties cannot be represented on metric scales, e.g. taste or smell. In these situations, statistical methods of subjective assessment are resorted to.

Frequently, the same quality characteristics are assessed by different test procedures, involving application-specific measuring equipment. The dispersibility of tablets is not evaluated in the same way as the dispersion ability of pesticides or of pigments. Such measurements are often not particularly informative, owing to the great variety of test equipment used or because they are single-point measurements, and take virtually no account of actual functional relationships. In this situation test methods would be desirable - especially for product development - which allow for more generally valid physical insights into the mechanisms producing the product properties.

3.1.2. Production-Related Product Properties

Properties which are less important to the end users but do matter to the manufacturer could lend themselves to being determined by generally valid measuring methods. The flowability of products, for example, can be assessed in the Jenike shear cell, irrespective of whether it is the feeding of pharmaceutical products to a tabletting press or the storage and discharge of fertilizers from a silo which is being assessed.

Cross-exchanges of measuring methods among individual sectors are most helpful and are among the aims set by professional bodies. One particular advantage of measuring methods which can be used in wider contexts is the physically meaningful information they provide. This is an aspect worth striving for.

3.1.3. Measurement of the Dispersity Variables

The most important variables, albeit often not sufficient, for describing the disperse state are the particle size and the particle size distribution. The particles in question must be assessed as they occur in the industrial process or the application, i.e., the particle size distribution must not be altered either by sampling, preparation, or measurement. For example, particles whose deposition from gases is to be considered should be measured on an aerosol; once they are deposited, agglomerates may form. Additionally preparation or handling might well destroy agglomerates and change the condition for deposition.

The development of particle measurement techniques is characterized by progress in the fields of optics, acoustics, and electronics. For example, the principle of light scattering, known since 1855, is utilized for measuring particle size distributions, by making full use of laser technology, microelectronics, and fast calculations. The familiar subjective assessment of particles and collections of particles under a microscope is now complemented by quantitative image analysis. Depending on the problem, other physical variables may be required for characterizing the condition of the product, for example the shape of the particles, the specific surface area, the porosity or, in the case of nanoparticles, the interfacial properties.

Quantitative characterization of product properties and dispersity variables is an indispensable tool in testing working hypotheses, in establishing the property functions or in sensitivity analyses [8]. This requires not only a critical assessment of the measurement techniques, but also of the measurement-related procedures, i.e., sampling and preparation prior to a measurement, which at present are still essential for most measurement methods.

3.2. Measurement of Product-Specific Characteristics

For process steps and equipment to be selected it is necessary to characterize, for each unit, a material in terms of the most important material

properties or process-related parameters. For example, in crystallization studies, the solubility and the width of the metastable region are measured; for drying applications, a large number of characteristics are discussed in Section 2.3.

3.3. Measurement of Process Parameters

The design and verification of a process requires the process parameters to be measured and maintained (sensitivity analysis). With a view to retaining them into production, the measurements at the design stage should establish what information must be available or should be available, and how accurately, rapidly and frequently it is measured to safeguard the process. The sensitivity analyses form the basis of the measuring and control plan. While even simple pressure or temperature measurements may present difficulties because of encrustation, the difficulty with solids-loaded systems is that, as yet, few sensors exist which can be used to measure viscosity, throughput, or concentrations of different components, let alone catalytic activity.

More recently, optical sensors in particular have provided new options for in-line process monitoring by characterizing the state of dispersion and the concentrations or composition (spectroscopy) [9]. Increasing insight into a process additionally provides the opportunity to employ model-based measurement techniques involving easily measured auxiliary variables so as to control the process.

Measurement technology developments in the field of ultrasonic absorption and ultrasonic spectrometry are likely, in the near future, to yield even deeper insights into what happens in the course of a process. The advantage of both optical sensors and ultrasonic sensors is that even disperse systems can be characterized in the original state, several variables being recorded simultaneously, some of which may interact with one another.

4. Research and Development

Solids technology is a relatively new industrial science. This is demonstrated by the fact that most of the solids process engineers active today are disciples of the "founding fathers" in this field. On the one hand, being a young science, it is in a state of rapid development; on the other hand, it can already pride itself on a rich fund of basic and applied research. Moreover, research and development in this field are characterized by a lively exchange of ideas between academic and industrial research institutions. Academic research is primarily devoted to elucidating fundamental relationships, whereas industrial research is concerned mainly with gathering more information on the applicability of the models produced in the academic sector. Fruitful collaboration in this manner clearly is the only possible way to progress further in solids technology. Numerous examples could be presented to demonstrate the advances such collaboration has already achieved. However, this was considered inappropriate at this point for two reasons: firstly, given the many researchers involved, there is the insoluble problem of making a representative selection; secondly, the following specialized articles present the opportunity of applying the ideas outlined in this introduction to the individual areas and to the field of solids technology as a whole.

5. References

Specific References

1. Frontiers in Chemical Engineering, Research Needs and Opportunities, National Research Council, Washington, 1988.
2. Technologien des 21. Jahrhunderts, press release 25/92, Bundesministerium für Forschung und Technologie, 1992.
3. B. J. Enmie, J. Green, R. Davies, *Chem. Eng. Prog.* **90** (1994) 31 – 43.
4. H. Rumpf, *Staub-Reinheit-Luft* **27** (1967) 3 – 13.
5. K. Borho, R. Polke, K. Wintermantel, H. Schubert, K. Sommer *Chem.-Ing.-Tech.* **62** (1991) 907 – 915; 4th World Congr. of Chem. Eng. Karlsruhe, Germany (1991) 93 – 98.
6. H. Rumpf: *Mechanische Verfahrenstechnik,* Carl Hanser Verlag, München 1975.
7. W. Bühler, W. Liedy, *Chem. Eng. Process* **26** (1989) 27 – 34.
8. J. Krekel, R. Polke, *Chem.-Ing.-Tech.* **64** (1992) no. 6, 27 – 34.
9. R. Polke, 5th World Congr. of Chem. Eng. San Diego USA (1996) VI, 551 – 556.

Particle Size Analysis and Characterization of a Classification Process

See also Solids Technology, Introduction

KURT LESCHONSKI, Technische Universität Clausthal, Clausthal, Federal Republic of Germany

1. Introduction 3
2. Graphic Representation 3
2.1. Particle Characteristics and Types and Measures of Quantity 4
2.2. Cumulative and Density Distributions 6
2.3. Distribution Functions 7
2.4. Calculation of Specific Surface Area 8
3. Characterization of a Classification or Separation Process 8
3.1. Classification Based on Error-Free Distribution Curves and Mass Balances 9
3.1.1. Density and Cumulative Distribution Curves 9
3.1.2. Mass Balances 10
3.1.3. Indirect Evaluation of v_1 and v_2 . . . 11
3.1.4. Cut Size 11
3.1.5. Grade Efficiency Curve (Tromp's Curve) 13
3.1.6. Sharpness of Cut and Other Parameters 13
3.2. Influence of Stochastic Errors 15
3.2.1. Indirect Evaluation of v_1 and v_2 . . . 16
3.2.2. Evaluation of the Grade Efficiency Curve, $T(x)$ 18
3.3. Influence of Systematic Errors . . . 19
3.3.1. Splitting in the Classifier 20
3.3.2. Incomplete Dispersion of the Feed Material 20
3.3.3. Comminution in the Classifier 20
4. Sample Splitting 21
5. Sedimentation Analysis 21
5.1. Calculation of the Equivalent Settling Rate Diameter, x_w 22
5.2. Methods of Sedimentation Analysis 23
5.2.1. Suspension Techniques 24
5.2.2. Superimposed Layer Techniques . . . 24
6. Classification Methods 25
6.1. Sieve Analysis 25
6.2. Air Classification 26
6.2.1. Cross-flow Classifiers 26
6.2.2. Counterflow Classifiers 28
7. Counting Methods 28
7.1. Image Analysis 29
7.2. Electrical Sensing Zone Method . . 30
7.3. Measurement by Light Scattering or Absorption 30
8. Fraunhofer Diffraction 31
9. References 33

Symbols

a	acceleration
A	Cunningham factor
A	area
A_1	fine particles in the feed material (Eq. 49)
A_2	coarse particles in the feed material (Eq. 50)
A_3	fine particles in the coarse fraction (Eq. 51)
A_4	coarse particles in the coarse fraction (Eq. 52)
A_5	fine particles in the fine fraction (Eq. 53)
A_6	coarse particles in the fine fraction (Eq. 54)
C_V	volume concentration
C_N	number concentration
D	characteristic fractal dimension
E	expected value
E_M	Ecart of Mayer [22], [23]
E_T	Ecart Terra
f	focal length
g	gravity constant
H	height
I	imperfection
I	intensity of light
I_0	intensity of incident light
k	constant
k_S	surface area shape factor
k_V	volume shape factor
K_0	corrected cumulative distribution of the feed material
K_1	corrected cumulative distribution of the fine fraction

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_02

K_2 corrected cumulative distribution of the coarse fraction
L length
m mass
m_∞ mass of particles at time $t = \infty$
m spread parameter of the power function
m_0 mass of the feed material
m_1 mass of the fine fraction
m_2 mass of the coarse fraction
m_s mass of solids
m_t mass of particles at time t
$\dot{m}$ mass flow rate
$\dot{m}_0$ mass flow rate of the feed material
$\dot{m}_1$ mass flow rate of the fine fraction
$\dot{m}_2$ mass flow rate of the coarse fraction
n spread parameter of the RRSB function
n number of intervals
q density distribution
q_0 number density distribution
q_0 density distribution of the feed material
q_1 density distribution of the fine fraction
q_2 density distribution of the coarse fraction
Q cumulative distribution
Q_0 cumulative distribution of the feed material
Q_1 cumulative distribution of the fine fraction
Q_2 cumulative distribution of the coarse fraction
Q_3 mass or volume cumulative distribution
r radius
r_i radius of the liquid surface
R residue
Re Reynolds number
s standard deviation
s_g logarithmic standard deviation
S surface area
S_m mass-related surface area
S_V volume-related surface area
t first derivation of T
t student factor
t time
T selectivity curve
T_tot overall classification efficiency
u_p peripheral velocity
U perimeter
v_a velocity of air
v_r radial velocity
v_1 relative amount of the fine fraction, fine yield
v_2 relative amount of the coarse fraction, coarse yield
V volume
w settling rate
w_at centrifugal settling rate
w_g stationary settling rate
x' position parameter of the RRSB function
x_a analytical cut size
x_min minimum diameter
x_min2 minimum diameter of the coarse fraction
x_max maximum diameter
x_max1 maximum diameter of the fine fraction
x_O cut size with $1 - Q_1(\mathrm{x_O}) = Q_2(\mathrm{x_O})$
x_A equivalent projected area diameter
x_S equivalent surface area diameter
x_t cut size
x_V equivalent volume diameter
$\mathrm{x_w}$ equivalent settling rate diameter
$x_{A\mathrm{m}}$ equivalent medial position projected area diameter
$x_{A\mathrm{s}}$ equivalent stable position projected area diameter
x_{50} median diameter
$x_{50,\,r}$ median of the $Q_r(x)$ distribution
z dimensionless variable of the error function
α angle
α dimensionless particle size $= \pi x/\lambda$
α weighted sum of variances
Δx interval of diameter
ε size-dependent error
η sharpness of cut parameter
η dynamic viscosity
η_i ordinate for evaluation of $v_1 = \eta_i/\xi_i$
Θ scattering angle
κ sharpness of cut
λ stride length
λ wavelength of light
λ mean free path
ξ integration variable for size
ξ_i abscissa for evaluation of $v_1 = Q_{1\mathrm{i}} - Q_{2\mathrm{i}}$
ϱ_d suspension density
ϱ_fl fluids density
ϱ_s solids density
σ^2 variance
τ dead flux
φ ratio of weighted density distributions
ψ sphericity
ω angular velocity

1. Introduction

Particle size analysis or particle characterization comprises the measurement and the quantitative description of physical properties of particulate matter. Therefore, it characterizes on the one hand the properties of single stationary or moving particles and on the other hand those of a bulk material.

This article describes the main areas of particle characterization, the representation of particle size analysis data, the characterization of a classification or separation process, and the methods of particle size analysis.

Only a limited number of different measuring techniques dominate instrumentation; therefore, this article does not describe all possible techniques, but concentrates on the most important ones.

Any unit operation of particle technology changes certain physical properties of solid or liquid particles coarser than ca. 10^{-7} m. In heat and mass transfer operations, the state of the process is described adequately by average values of the molecular state. In all particle technology processes, however, the distribution of certain characteristic particle properties must be analyzed. The main task of particle size analysis is to measure these distributions.

Disperse or particulate systems consist of a finely divided particulate phase in a uniform carrier or dispersion medium [1]. The particulate phase and the carrier medium may be solid, liquid, or gaseous. Therefore, particle size analysis predominantly investigates such systems as suspensions, emulsions, aerosols, or aerodispersions. However, the particulate phase may also consist of bubbles or pores.

The range of particle sizes to be measured covers more than five to six orders of ten, with smallest particle sizes 10^{-9} to 10^{-7} m. Many physical properties of particulate matter, e.g., tendency to agglomerate, settling characteristics, solubility, and permeability, depend strongly on particle size.

2. Graphic Representation [2]

The individual elements of particulate matter always form a distribution, the elements of which can be classified according to their size. The distribution is, therefore, characterized by a quantity, e.g., the number or the mass of particles present in individual size classes. The elements can be solid particles, droplets, bubbles, or pores.

Many methods can be used to analyze particulate matter. Each group of measurement techniques based on the same physical principle provides, in general, different distribution curves. The reason for this must partly be attributed to the fact that in particle size analysis every possible physical principle has been used to determine these distributions. For several reasons, the results are generally not the same [3]. One of the main problems still encountered with most methods in use is their unknown absolute accuracy. However, apart from this the distributions differ because (1) different physical properties are used to characterize the "size" of individual particles, and (2) different types and different measures of quantity are chosen.

2.1. Particle Characteristics and Types and Measures of Quantity

In a graphic representation of particle size analysis data, the independent variable plotted on the abscissa describes the physical property chosen to characterize the size of the particles (Table 1). The dependent variable characterizes the type and measure of quantity and is plotted on the ordinate.

Table 1. Types and measures of quantity

Type of quantity		Measure of quantity	
		density distribution	cumulative distribution
Number:	$r = 0$	$q_0(x)$	$Q_0(x)$
Length:	$r = 1$	$q_1(x)$	$Q_1(x)$
Area:	$r = 2$	$q_2(x)$	$Q_2(x)$
Volume:	$r = 3$	$q_3(x)$	$Q_3(x)$
General description:	$r = r$	$q_r(x)$	$Q_r(x)$

Different types of quantity are
$r = 0$: number
$r = 1$: length
$r = 2$: area, surface area
$r = 3$: volume, mass
Measures of quantity are
Q_r: cumulative distribution
q_r: density distribution

The density distribution is obtained from the cumulative distribution by differentiation:

$$q_r(x) = \frac{\mathrm{d}Q_r(x)}{\mathrm{d}x}$$

If particles having the same physical property are counted, the *type of quantity* used is a number; if the particles are weighed, the type of quantity is a mass. Particles laid end to end form a line of a certain length, but they also represent a certain projected area. Thus, length and area are other ways of quantifying particle size.

Therefore, the following types of quantity are distinguished: number, length, area, volume, or – at constant particle density – mass. This is the decisive difference from ordinary statistics, where the type of quantity that is used is always a number. Since number and volume distributions of the same powder may differ considerably, the type of quantity measured or chosen when representing the data must be taken into account.

However, a distinction must also be made between two *measures of quantity.* This is necessary because in a counting method either all particles smaller than a certain size may be counted, or only those found in certain size classes. The different relative amounts of particles measured in certain size intervals form the so-called *density distribution,* $q_r(x)$, which represents the first derivative of the *cumulative distribution*, $Q_r(x)$. In this notation x characterizes a physical property uniquely related to particle size. The subscript, r, indicates the type of quantity chosen.

Table 1 shows the types of distribution that can be distinguished:

The physical properties that characterize particles are as follows:

1) linear dimension
2) surface area, and projected area
3) volume
4) mass
5) settling rate
6) response of electrical, optical, or acoustical field

These are the one-, two-, or three-dimensional particle properties, such as linear dimensions, chords of different definition, the projected area of particles, their surface area, or their volume. Other characteristics are the mass of particles, their settling rate or terminal velocity in a viscous fluid, or the response of an electrical or an optical field to the presence of particles.

Each of these properties can be used to characterize the size of a particle. In principle, however, the properties can be applied usefully to the characterization of particles only if a unique relationship exists between the physical property used and a one-dimensional property unequivocally defining "size." With irregular particles, this can be achieved only approximately.

So-called *equivalent diameters* are defined, which are diameters of spheres that yield the same value of a certain physical property when analyzed under the same conditions as the irregularly-shaped particle. The relationship between the physical property chosen and the equivalent particle diameter must be either evaluated theoretically or measured experimentally.

Therefore, an equivalent diameter represents the diameter of a sphere that has the same physical property as the irregularly-shaped particle. One distinguishes, for instance, between

x_A:	equivalent projected area diameter
x_S:	equivalent surface area diameter
$x_{A\mathrm{m}}$:	equivalent projected area diameter for mean particle position
$x_{A\mathrm{s}}$:	equivalent projected area diameter for stable particle position
x_V:	equivalent volume diameter
x_w:	equivalent settling rate diameter

For a given sphere, all equivalent diameters are the same; they differ from each other if the particles are irregularly-shaped. Under certain assumptions and after a shape factor has been defined that describes the deviation in shape from a sphere, the equivalent diameters given above can be intercorrelated using the shape factor *sphericity,* as defined by WADELL [4]:

$$\Psi = x_V^2 / x_S^2 \tag{1}$$

If Cauchy's theorem [5] is used and a force balance is applied to a particle settling in a viscous fluid under laminar conditions, the following is obtained:

$$\begin{array}{llll}
x_S & & = x_{A\mathrm{m}} & = x_w\,\psi^{-3/4} & = x_V\,\psi^{-1/2} \\
x_{A\mathrm{m}} & = x_S & & = x_w\,\psi^{-3/4} & = x_V\,\psi^{-1/2} \\
x_w & = x_S\,\psi^{3/4} & = x_{A\mathrm{m}}\,\psi^{3/4} & & = x_V\,\psi^{1/4} \\
x_V & = x_S\,\psi^{1/2} & = x_{A\mathrm{m}}\,\psi^{1/2} & = x_w\,\psi^{-1/4} &
\end{array}$$

In principle, the following inequality holds:

$$x_{A\mathrm{s}} > x_{A\mathrm{m}} > x_S > x_V > x_w$$

2.2. Cumulative and Density Distributions

The graphic representation of particle size analysis data is shown in Figures 123. The term x represents an arbitrary equivalent diameter, the dimension of which is a length such as µm.

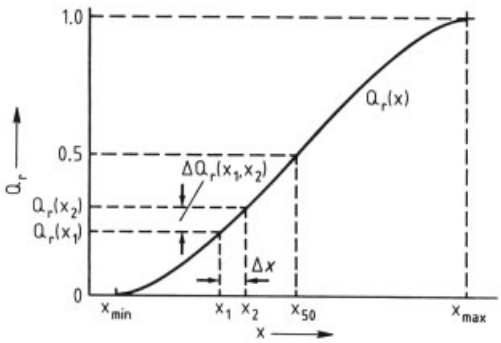

Figure 1. Cumulative distribution $Q_r(x)$

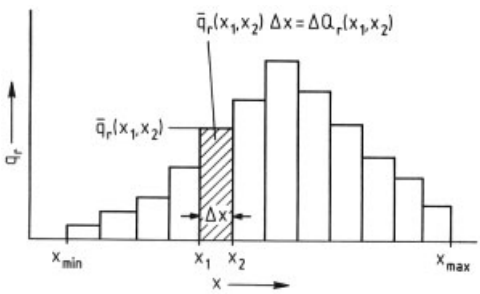

Figure 2. Histogram of density distribution $q_r(x)$

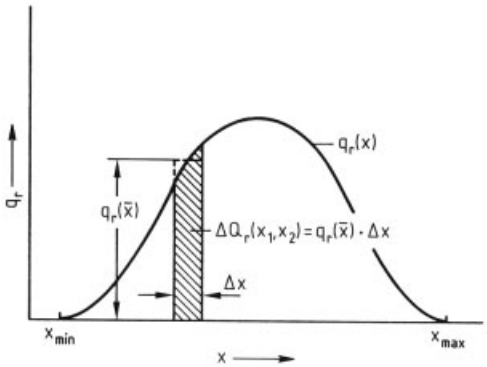

Figure 3. Density distribution $q_r(x)$

Figure 1 shows the normalized cumulative distribution $Q_r(x)$. Each individual point of $Q_r(x)$ specifies the relative amount of particles smaller than or equal to a certain equivalent diameter. The kind of quantity chosen is indicated by the subscript *r*. The normalized curve $Q_r(x)$ has values between 0 and 1:

$Q_r(x = x_{\min}) = 0$
$Q_r(x = x_{\max}) = 1$
(condition of normalization)

Figure 2 shows the normalized *histogram* of a density distribution, $\bar{q}_r(x_1, x_2)$. The difference ΔQ_r, of two values of the cumulative distribution, $Q_r(x)$, between two equivalent diameters x_1 and $x_2 = x_1 + \Delta x$ is represented by

$$\Delta Q_r(x_1, x_2) = Q_r(x_2) - Q_r(x_1) \tag{2}$$

The histogram can then be obtained from

$$\bar{q}_r(x_1, x_2) = \frac{\Delta Q_r(x_1, x_2)}{\Delta x} = \frac{Q_r(x_2) - Q_r(x_1)}{x_2 - x_1} \tag{3}$$

The value $\bar{q}_r(x_1, x_2)$ then represents the constant height of the rectangular shaded area of Figures 2 and 3 of width Δx. Therefore, the shaded area is given by

$$\bar{q}_r(x_1, x_2) \cdot \Delta x = \Delta Q_r(x_1, x_2) \tag{4}$$

and represents the relative amount of particles within the size interval, Δx.

If the curve $Q_r(x)$, as plotted in Figure 1, can be differentiated, the density distribution, $q_r(x)$, is obtained from

$$q_r(x) = \frac{\mathrm{d}Q_r(x)}{\mathrm{d}x} \tag{5}$$

This function, $q_r(x)$, is plotted in Figure 3. It represents the first derivative or the slope of $Q_r(x)$. It should be noted that $q_r(x)$ has the dimension of an inverse length. Irrespective of the type of quantity plotted, the overall area underneath $q_r(x)$ equals unity or 100 % (condition of normalization).

$$\int_{x_{\min}}^{x_{\max}} q_r(x)\,\mathrm{d}x = Q_r(x_{\max}) - Q_r(x_{\min}) = 1 \tag{6}$$

Each of the three curves shown in Figures 1, 2, 3 represents the information needed for further evaluation of the data. If the cumulative distribution, $Q_r(x)$, has been obtained by multiple analyses, statistical methods allow the most probable distribution curve to be found. If certain assumptions with regard to an analytical function

that fits $Q_r(x)$ are possible, regression analysis using least squares methods yield the most probable parameters of the analytical function. The main problems in evaluating more data from these curves originate mainly from two sources:

1) The distribution curve, as plotted, has been analyzed with several methods of size analysis using different types of quantity and different physical properties defining "size." Therefore, the values plotted on both the ordinate and the abscissa must be recalculated and reduced to the same form. It may be necessary, for instance, to transform a number distribution of volumes into a volume distribution of equivalent sizes.
2) The distribution curves obtained must be correlated to other material properties, such as strength, solubility, or pore volume. Usually one figure, or at the most two figures, must then represent the size distribution adequately. The problem of defining an average particle size (a position parameter) and another parameter describing the spread of the distribution (a spread parameter) then arises. Some of the existing possibilities are discussed in [2].

2.3. Distribution Functions

Several analytical particle size distributions have been found to approximately describe empirically-determined particle size distributions. Special graph papers have been developed in which cumulative distributions following these functions form a straight line. Three of the best known analytical distribution functions are the Gaudin–Schumann distribution, the RRSB function, and the Gaussian distribution.

1) The power function

$$Q_3(x) = (x/x_{\max})^m \tag{7}$$

is also known as the *Gaudin–Schumann distribution* [6], [7]. It contains a position parameter, $x_{\max}$, and a spread parameter, m. The latter parameter indicates the slope of the straight line, which is obtained when the above function is plotted on a $\log Q_3 - \log x$ paper.

2) The *Rosin, Rammler, Sperling, Bennett (RRSB) function* has the following equation [8–12]:

$$1 - Q_3(x) = \exp\left(-\left(x/x'\right)^n\right) \tag{8}$$

x' represents the position parameter where the function $1 - Q_3(x') = e^{-1} = 0.368$. The spread parameter, n, represents the slope of the straight line obtained in a special graph paper when log log $(1/R)$ is plotted on the ordinate and log x is plotted on the abscissa. The straight line is represented by

$$\log\log(1/R) = n\log x - n\log x' + \log\log e \tag{9}$$

3) The third analytical function is the *normal distribution* or *Gaussian distribution*, given here in its general form [11]:

$$q_r^*(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-z^2/2\right) \tag{10}$$

with z being a dimensionless variable describing particle size. The cumulative distribution is obtained from

$$Q_r^*(z) = \int_{-\infty}^{z} q_r^*(\xi)\,\mathrm{d}\xi \tag{11}$$

Normal distributions differ by the definition of z and the distinction between (1) the Gaussian distribution with linear abscissa and (2) the logarithmic normal distribution.

In the *Gaussian distribution with linear abscissa*, z is defined as

$$z = (x - x_{50,r})/s \tag{12}$$

where $x_{50,r}$ is the median of the $Q_r(x)$ distribution and s represents the standard deviation.

The quantity $Q_r*(z)$ is tabulated for certain z values; therefore, the standard deviation, s, can be calculated from

$$z = 1: \quad Q_r^*(1) = 0.841;$$

$$s = x_{84,r} - x_{50,r} \tag{13}$$

$$z = -1: \quad Q_r^*(-1) = 0.159;$$

$$s = -(x_{16,r} - x_{50,r}) \tag{14}$$

In the logarithmic normal distribution, the following substitution is made:

$$z = \ln\left(x/x_{50,r}\right)/s \tag{15}$$

or according to HERDAN [11]:

$$z = \log\left(x/x_{50,r}\right)/\log s_g = \ln\left(x/x_{50,r}\right)/\ln s_g \tag{16}$$

By comparison of Equations (15) and (16), Equation (17) is obtained:

$$s = \ln s_g \quad \text{or} \quad s_g = e^s \tag{17}$$

The standard deviation can be calculated from

$$s = \ln\left(x_{84,r}/x_{50,r}\right) = \ln\left(x_{50,r}/x_{16,r}\right) = \ln s_g \tag{18}$$

2.4. Calculation of Specific Surface Area

Specific surface area may be defined in two ways:

1) The ratio of surface area, S, to volume, V, is defined as the volume-related surface area, S_V:

$$S_V = S/V \tag{19}$$

2) The ratio of surface area, S, to mass, m, is defined as the mass-related surface area, S_m:

$$S_m = S/m \tag{20}$$

S_m and S_V differ by the solids density, ϱ_s:
S_m and S_V differ by the solids density, ϱ_s:

$$S_V = S_m\,\varrho_s \tag{21}$$

The volume-related surface area S_V of particles of a size interval $\Delta x = x_2 - x_1$ can be calculated from a number distribution $q_0(x)$:

$$S_V\left(x_1, x_2\right) = \int_{x_1}^{x_2} k_S\left(x\right) x^2\, q_0\left(x\right) \mathrm{d}x \Big/ \int_{x_1}^{x_2} k_V\left(x\right) x^3\, q_0\left(x\right) \mathrm{d}x \tag{22}$$

where $k_S(x)$ = surface area shape factor and $k_V(x)$ = volume shape factor

In general, k_S and k_V are assumed to be independent of size, and a new shape factor, f, is introduced according to HEYWOOD which takes into account deviations in shape from spherical particles [13]. If Equation (23) is introduced

$$6f = k_S\left(x\right)/k_V\left(x\right), \tag{23}$$

Equation (24) is obtained:

$$S_V\left(x_1, x_2\right) = 6f\,\frac{\int_{x_1}^{x_2} x^2\, q_0\left(x\right) \mathrm{d}x}{\int_{x_1}^{x_2} x^3\, q_0\left(x\right) \mathrm{d}x} \tag{24}$$

3. Characterization of a Classification or Separation Process

[14], [15]

Classification is a basic unit operation of particle technology. As shown in Figure 4, the feed material consists of particular matter of size, settling rate, density, or shape distribution $q_0(\xi)$; this is separated in a classification process into a fine or light $q_1(\xi)$ and coarse or heavy fraction $q_2(\xi)$.

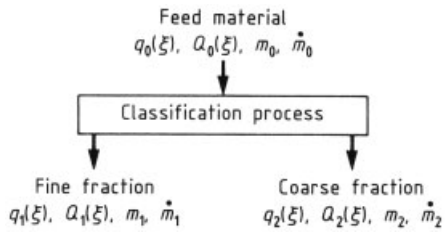

Figure 4. Fraction and distributions produced in a one-step classification process

The subscripts 0, 1, and 2 should not be mistaken for the subcripts used in Chapter 2. Actually $q_{r,0}(\xi)$ or $q_{r,1}(\xi)$ should have been employed; however, because the quantity used in the calculation of the mass balances of Section 3.1.2 is always assumed to be the same, the first subscript, r, has been omitted.

The parameter ξ represents any kind of physical property that is uniquely correlated to the "size" of a particle. The "size" of a particle is usually represented by the diameter of a sphere equivalent in physical behavior to the irregularly-shaped particle (equivalent diameter; see Section 2.2). The equivalent spherical

diameter is called x in this and the following chapters. The parameter ξ may, however, also represent a settling rate in a gas or a liquid, a density, or even a shape-related piece of information.

According to Figure 4, the feed material is separated into at least two parts, with one part mainly containing particles below a specific value, and the other part containing particles above that value. Irrespective of the classification methods used, the classification itself can be described completely by using the density and/or the cumulative distribution curves of the feed material and the product fractions [15]. Therefore, it is important to determine these curves as accurately as possible. In Section 3.1 the characterization of a classification is described under the assumption that the distribution curves describing the feed material and the products, as well as the overall mass balance, are free of errors. This is the traditional approach; it has been applied in the way described throughout the world for many years. The errors involved in particle size analysis have been included in the characterization of a classification process by LESCHONSKI and HERMANN [16]; Section 3.2 describes this method and its improvements.

Section 3.3 deals with systematic changes of the feed material in the classifier, for instance, due to comminution of the feed material in the classification zone. Systematic changes lead to abnormal grade efficiency or classifier selectivity curves, $T(x)$, which no longer extend between 0 and 1.

The general description of a classification process, based on error-free distribution curves and an error-free overall mass balance, was first attempted in the processing of minerals or coal; it has been the subject of many papers [17]. In the following chapters, the description of a separation process is explained by using classification as an example. However, the same curves and equations are valid for separations based on density, shape, or any other, e.g., size-dependent physical property.

In classification processes, the characteristic physical property of a single particle is described by its equivalent settling rate diameter, x_w, plotted on the abscissa. It has the dimension of a length and is usually given in 10^{-6} m.

3.1. Classification Based on Error-Free Distribution Curves and Mass Balances

3.1.1. Density and Cumulative Distribution Curves

In all classification processes a given feed material (subscript 0) is classified into at least two parts, which are called the fine (subscript 1) and the coarse (subscript 2) fractions. If an ideal classification were possible, the fine fraction would contain particles below or equal to a certain size, x_t, the so-called *cut size*, and the coarse fraction would contain all particles above that size. Such a classification process is shown in Figure 5.

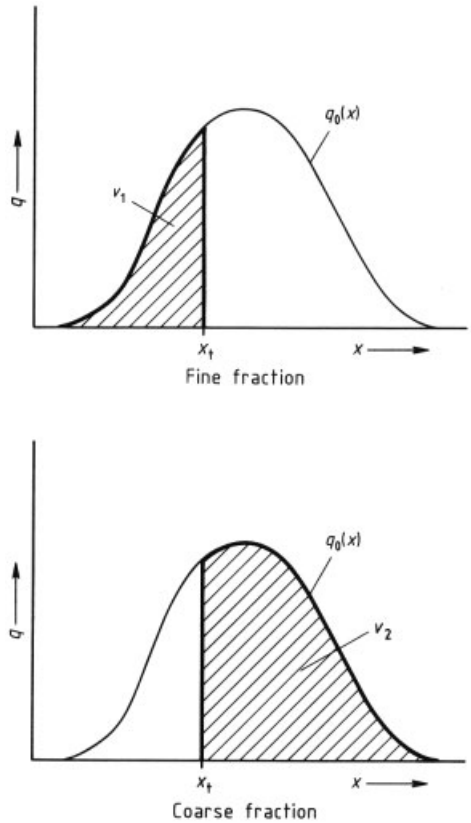

Figure 5. Weighted density distributions of feed, fine, and coarse fractions of an ideal classification

The shaded areas beneath the weighted density distributions of the fine and the coarse product represent the overall relative amount of the fine fraction, v_1, also known as *fine yield*, and of the coarse fraction, v_2, also known as *coarse yield*, the sum of which equals 100 % or 1.

In reality, however, in a certain range of sizes ($x_{\min 2} \leq x \leq x_{\max 1}$), particles of the same size, x, are present in both the fine and the coarse fraction, as shown in Figure 6.

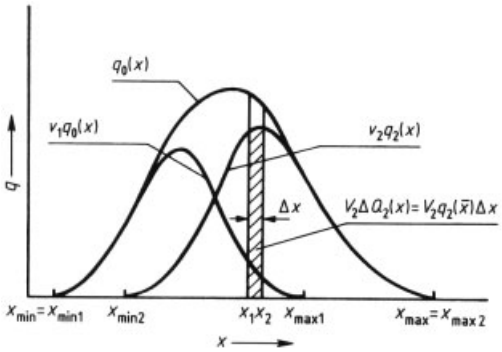

Figure 6. Weighted density distributions of the feed material $q_0(x)$, the fine fraction $[v_1q_1(x)]$, and the coarse fraction $(v_2q_2(x))$

If the density distribution curves of the fine, $q_1(x)$, and the coarse fraction, $q_2(x)$, are multiplied by their relative amounts, v_1 and v_2 (see Eqs. 27 and 28), the weighted density distribution curves $v_1q_1(x)$ and $v_2q_2(x)$ are obtained. These curves overlap and intersect each other in the size range

$$x_{\min 2} \leq x \leq x_{\max 1}$$

The point of intersection is attributed to the so-called cut size, x_t. The particles below and above x_t in the coarse or fine fraction, respectively, have been incorrectly classified. These particles are commonly called *misplaced particles* or *misplaced material*. The less precise a classification process has been, the greater the areas representing the misplaced materials are. The amount of misplaced material in the coarse fraction is represented by the integral:

$$v_2 \int_{x_{\min 2}}^{x_t} q_2(x)\,\mathrm{d}x = v_2 Q_2(x_t) \tag{25}$$

The amount of misplaced material in the fine fraction can be calculated from Equation (26):

$$v_1 \int_{x_t}^{x_{\max 1}} q_1(x)\,\mathrm{d}x = v_1 Q(x_{\max 1}) - v_1 Q_1(x_t)$$

$$= v_1 [1 - Q_1(x_t)] \tag{26}$$

The classification, as described, can also be represented by using the cumulative distribution curves $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$, as shown in Figure 7.

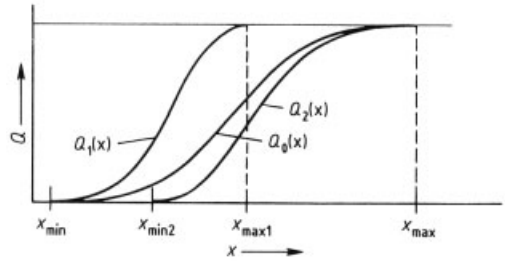

Figure 7. Cumulative distribution curves of the feed material $Q_0(x)$ and the fine fraction $Q_1(x)$ and coarse fraction $Q_2(x)$

3.1.2. Mass Balances

In the representation of density distribution curves, the relative mass is often used as the quantity plotted on the ordinate. In principle, other quantities such as number, length, area, or volume may be used, without change in the grade efficiency or classifier selectivity curves. Therefore, the distribution curves and all other quantities are based on particle mass. It should be repeated that in the equations to follow the density and cumulative distribution curves are assumed to be free of errors.

Mass Balance in the Size Range from $x_{\min}$ to $x_{\max}$. In the classification process, the mass, m_0, or the mass flow rate, $\dot{m}_0$, of the feed material is split into the mass, m_1, or the mass flow rate, $\dot{m}_1$, of the fine material and the mass, m_2, or the mass flow rate, $\dot{m}_2$, of the coarse material.

Therefore, the relative amount of fine or coarse material, v_1 and v_2, is obtained from

$$v_1 = m_1/m_0 = \dot{m}_1/\dot{m}_0 \tag{27}$$

$$v_2 = m_2/m_0 = \dot{m}_2/\dot{m}_0 \tag{28}$$

Because of

$$m_0 = m_1 + m_2 \quad \text{or} \quad \dot{m}_0 = \dot{m}_1 + \dot{m}_2 \tag{29}$$

the following is obtained:

$$v_1 + v_2 = 1 \tag{30}$$

In Figures 5 and 6, v_1 and v_2 are represented by the areas beneath the weighted density distribution curves of the fine $[v_1 q_1(x)]$ and the coarse $[v_2 q_2(x)]$ material. The area beneath the density distribution curve of the feed material $q_0(x)$ equals unity. For example:

$$\int_{x_{\min}}^{x_{\max 1}} v_1\, q_1\,(x)\,\mathrm{d}x = v_1\,[Q_1\,(x_{\max 1}) - Q_1\,(x_{\min})] = v_1 \tag{31}$$

Mass Balance in a Small Size Range. In the size range $x_{\min 2} \leq x \leq x_{\max 1}$ the quantity

$$\mathrm{d}Q_0\,(x) = q_0\,(x)\,\mathrm{d}x \tag{32}$$

is split into two fractions:

$$\mathrm{d}Q_0\,(x) = v_1\,\mathrm{d}Q_1\,(x) + v_2\,\mathrm{d}Q_2\,(x) \tag{33}$$

Replacing $\mathrm{d}Q\,(x)$ by $q\,(x)\,\mathrm{d}\,x$ yields

$$q_0\,(x) = v_1 q_1\,(x) + v_2 q_2\,(x) \tag{34}$$

If only two density distributions and the relative amount of the fine or the coarse material are known, Equation (34) may be used to determine the missing third density distribution or to check and control values of this density distribution, if also measured. Equation (34) is also used to construct the set of density distribution curves in Figure 6. For instance, if the density distributions of the feed and the coarse material have been chosen or measured, the density distribution of the fine material can be calculated or constructed from

$$v_1 q_1\,(x) = q_0\,(x) - v_2 q_2\,(x)$$

This must be taken into account when diagrams are plotted, as shown in Figure 6.

Mass Balance below a Certain Size. Integrating Equation (34) between $x_{\min}$ and x yields

$$Q_0\,(x) = v_1 Q_1\,(x) + v_2 Q_2\,(x) \tag{35}$$

3.1.3. Indirect Evaluation of v_1 and v_2

In many cases of practical application, v_1 and v_2 cannot be calculated from the relevant masses or mass flow rates because these values are not available, are difficult to measure, etc. However, if representative samples of the three product streams have been measured, Equation (30), (34), and (35) may be used to calculate v_1 or v_2. Introducing, for instance, Equation (30) into Equation (35) and solving with respect to v_1 yields

$$v_1 = 1 - v_2 = [Q_0\,(x) - Q_2\,(x)]\,/\,[Q_1\,(x) - Q_2\,(x)] \tag{36}$$

If the cumulative distributions $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$ are free of errors (i.e., the mass balance according to Equation (36) leaves no remainder), then v_1 or v_2 will be constant and independent of size x. In reality, however, the calculated values of v_1 will vary due to errors of size analysis and sampling when measuring $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$. In this case, average values of v_1 and v_2 must be calculated by using the procedures suggested in Section 3.2.1.

3.1.4. Cut Size

In principle, any value of x between $x_{\min 1}$ and $x_{\max 2}$ (Fig. 6) can be used as cut size, x_{t}. However, its definition should be correlated as closely with the field of application of the product as possible.

Reference Cut Size. In many cases of practical application, the grade efficiency curve (see Section 3.1.5) is not determined and the fine product is available for analysis only. Then a so-called reference cut size is defined which represents a particle size, x_y, at y % weight undersize $[Q_3(x_y) = y\,\%]$. The value y is, for instance, taken to be 97 % or even 99 %. It should be noted that x_y does not in the true sense represent a cut size of a classifier because the other products and the relative amounts of the fine or the coarse material have not been taken into account.

Three definitions of cut size are commonly used: (1) the median of the grade efficiency curve x_{50}, (2) the analytical cut size x_{a}, and (3) the overlap cut size x_0.

Median of the Grade Efficiency Curve. In Figure 6 the weighted density distribution curves intersect at a certain size x_t. This is the cut size.

$$x_t = x_{50} \tag{37}$$

where

$$v_1 q_1(x_{50}) = v_2 q_2(x_{50}) \tag{38}$$

Particles of size x_{50} are equally present in the coarse and the fine fractions.

The grade efficiency curve, or classifier selectivity curve, $T(x)$, as defined in Section 3.1.5, has a value of 0.5 or 50 % at this particular size. Therefore, the cut size, x_{50}, also represents the median value of the grade efficiency curve.

Analytical Cut Size. The analytical cut size x_a represents the value of x at which a vertical line divides the shaded areas I and II in Figure 8 beneath the $v_1 q_1(x)$ and $v_2 q_2(x)$ curves into two equal parts. At this cut size, the coarse and fine material contains equal quantities of misplaced material.

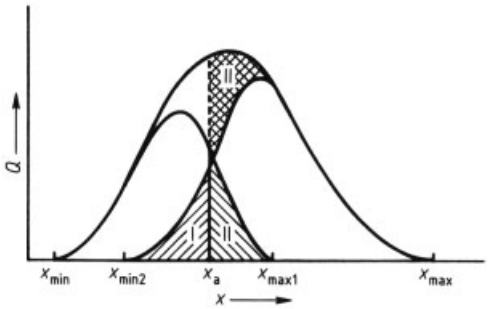

Figure 8. Definition and representation of the analytical cut size x_a

The integration yields

$$v_2 \int_{x_{min2}}^{x_a} q_2(x)\,dx = v_1 \int_{x_a}^{x_{max1}} q_1(x)\,dx$$

$$= \int_{x_a}^{x_{max1}} (q_0(x) - v_2 q_2(x))\,dx \tag{39}$$

or

$$v_1 = 1 - v_2 = Q_0(x_a) \tag{40}$$

An analytical air classifier can be seen as a black box, similar to the situation shown in Figure 4, operated under defined conditions. A certain mass, m_0, of the feed material is fed to the classifier, and the mass, m_2, of the coarse product only is quantitatively obtained. Therefore, it is possible to calculate $m_2/m_0 = v_2$.

The relative amount, v_2, of the coarse material as determined by the experiment is usually understood to equal the so-called residue R, that is, the amount of particles coarser than x_t. Equation (40) shows that the cut size x_t must equal the analytical cut size, x_a, if $R = v_2$. Then v_2 represents one point of the size distribution of the feed material to be measured.

Overlap Cut Size. The value x_O is taken as the cut size of technical air classifiers, where v_1 or v_2 are unknown and where $Q_1(x)$ and $Q_2(x)$ can be obtained relatively easily. The value x_O represents a cut size, x_t, where $Q_2(x_O)$ equals the oversize fraction of the fine product: $1 - Q_1(x_O)$

$$1 - Q_1(x_O) = Q_2(x_O) \tag{41}$$

The determination of x_O is shown in Figure 9.

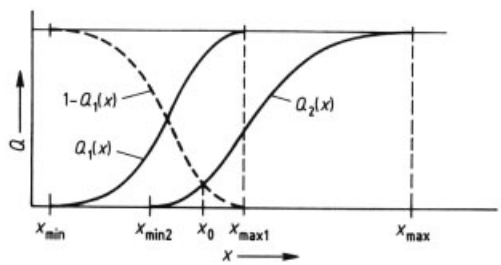

Figure 9. Determination of x_O

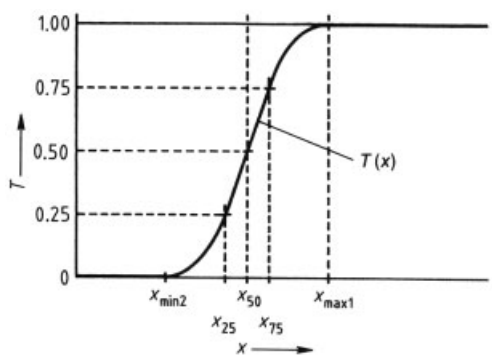

Figure 10. Grade efficiency or classsifier selectivity curve, $T(x)$

3.1.5. Grade Efficiency Curve (Tromp's Curve)

To describe the efficiency of a classification process, the so-called grade efficiency curve, $T(x)$, is normally derived from the density distribution curves of Figure 6 [17].

The *grade efficiency* or *classifier selectivity*, T, represents the ratio of the relative amount of material of a certain size present in the coarse material ($v_2 q_2(x)\,\mathrm{d}x$), to the relative amount of the same size initially present in the feed material: $q_0(x)\,\mathrm{d}x$. Therefore, the grade efficiency is equal to

$$T(x) = v_2 q_2(x) / q_0(x) \tag{42}$$

The grade efficiency $T(x)$ can be obtained from the ratio of the ordinates of the coarse and the feed materials, as shown in Figure 6. It is equal to 0 at particle sizes below $x_{\min 2}$; it is equal to 1 at particle sizes above $x_{\max 1}$, and is 0.5 at $x = x_{50}$.

If T is plotted against particle size, x, with x being the equivalent settling rate diameter, the resulting curve is the *grade efficiency curve*, or *classifier selectivity curve* (Fig. 10). It should start at 0 and remain there between $x_{\min 1}$ and $x_{\min 2}$, and it should reach the value of 1 at particle sizes equal to or coarser than $x_{\max 2}$.

Attempts have often been made to fit analytical functions to grade efficiency curves. In most cases these attempts have not been very successful. A universally applicable function does not seem to be available as yet. Tromp [17], Mayer [18] and Eder [19] succeeded in special cases; Trawinski [20] also published an analysis of useful functions to fit grade efficiency curves.

3.1.6. Sharpness of Cut and Other Parameters

In principle, the sharpness or the quality of a classification process is better when the overlapping size range $x_{\min 2} \le x \le x_{\max 1}$ is smaller, or when the amounts of misplaced material are smaller. A considerable number of parameters have been defined to indicate quantitatively the sharpness or the lack of sharpness of a cut in a classification process.

These parameters can be used meaningfully only when they are selected with regard to the technical application for which the classification process has been used. Therefore, one parameter alone is in many cases not as adequate for a complete description of the classification process as a series or even a combination of different parameters. Most parameters that are given quantify only a part or parts of the information obtainable from the grade efficiency curve. In 1972, Leschonski formed three groups of parameters which suffice to include the existing parameters: (1) parameters formed with characteristic particle sizes, (2) parameters derived from cumulative distribution curves, and (3) parameters derived from the overall course of the grade efficiency curve.

Parameters from Characteristic Particle Sizes. Parameters of this type indicate a difference or a ratio of characteristic particle sizes taken from the grade efficiency curve. Some examples follow:

Ecart Terra [21]:

$$E_{\mathrm{T}} = (x_{75} - x_{25})/2 \tag{43}$$

F. W. Mayer [22], [23]:

$$E_{\mathrm{M}} = x_{90} - x_{10} \tag{44}$$

Imperfection [24]:

$$I = (x_{75} - x_{25})/2\,x_{50} \tag{45}$$

K. Grumbrecht [25]:

$$\kappa_{75/25} = x_{75}/x_{25} \tag{46}$$

Th. Eder [24]:

$$\kappa_{25/75} = x_{25}/x_{75} \tag{47}$$

$$\kappa_{35/65} = x_{35}/x_{65} \tag{48}$$

where x subscript y indicates the value of size x where the grade efficiency curve has a value of $T = y\,\%$.

Parameters from Cumulative Distribution Curves [7]. These parameters can be determined from the cumulative distribution curves $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$ and the relative amounts of the fine material, v_1 or v_2. The parameters have the advantage that the grade efficiency curve is not required for their determination. In principle, a distinction is made between six different areas beneath the three density distribution curves, as shown in Figure 11.

From these areas a number of characteristic parameters can be derived:

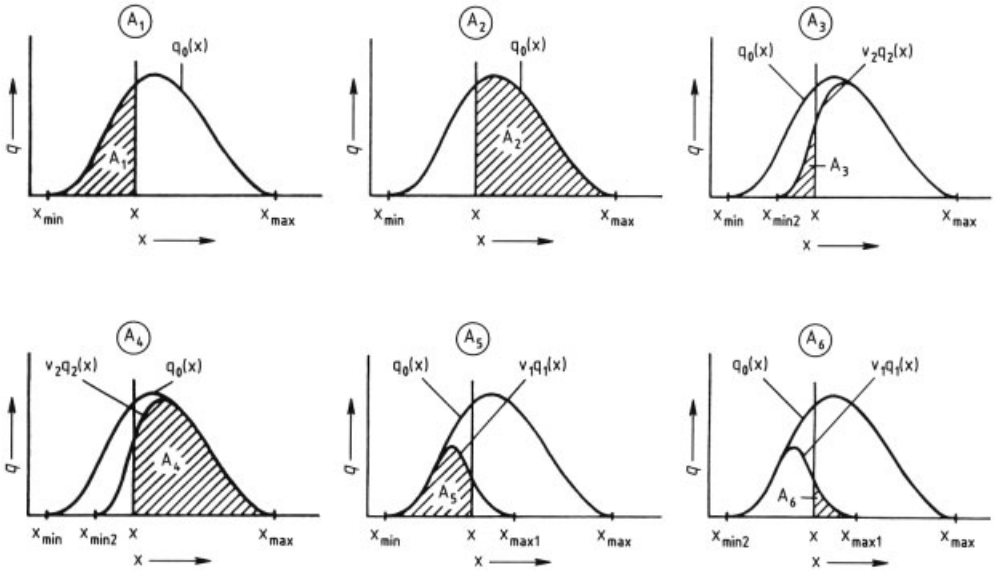

Figure 11. Definition and representation of six areas beneath the three density distribution curves

1) Parameters composed of the areas A_1–A_6, as shown in Figure 11. These parameters indicate either the amount of fine or coarse particles below or above a certain size in the feed, the fine or the coarse fractions. The following distinctions are made:
   – Fine particles in the feed material:

$$A_1 = Q_0(x) = \int_{x_{\min}}^{x} q_0(\xi)\,\mathrm{d}\xi \tag{49}$$

   – Coarse particles in the feed material:

$$A_2 = 1 - Q_0(x) = \int_{x}^{x_{\max}} q_0(\xi)\,\mathrm{d}\xi \tag{50}$$

   – Fine particles in the coarse fraction:

$$A_3 = v_2 Q_2(x) = \int_{x_{\min}}^{x} v_2 q_2(\xi)\,\mathrm{d}\xi \tag{51}$$

   – Coarse particles in the coarse fraction:

$$A_4 = v_2\left[1 - Q_2(x)\right] = \int_{x}^{x_{\max}} v_2 q_2(\xi)\,\mathrm{d}\xi \tag{52}$$

   – Fine particles in the fine fraction:

$$A_5 = v_1 Q_1(x) = \int_{x_{\min}}^{x} v_1 q_1(\xi)\,\mathrm{d}\xi \tag{53}$$

   – Coarse particles in the fine fraction:

$$A_6 = v_1\left[1 - Q_1(x)\right] = \int_{x}^{x_{\max}} v_1 q_1(\xi)\,\mathrm{d}\xi \tag{54}$$

   The parameters A_1–A_6 vary with size; A_3 and A_6 are plotted vs. x in Figure 12. The misplaced coarse material in the fine fraction, A_6 drops from v_1 at small sizes to 0 at $x_{\max 1}$; the misplaced fine material in the coarse fraction, A_3, on the other hand, increases from 0 at $x_{\min 2}$ to v_2. The sum of A_3 and A_6 represents the total misplaced material; it forms a minimum at $x = x_{50}$ [25].
2) Relative parameters may also be formed from ratios of the areas A_1 to A_6, as for instance:
   – the *retrieval* or *recovery* of fine particles related to those initially present in the feed material (or related to the feed undersize)

$$A_5/A_1 = v_1 Q_1(x)/Q_0(x) \tag{55}$$

   – the *retrieval* or *recovery* of coarse particles related to those initially present in the feed material (or related to the feed oversize)

$$A_4/A_2 = v_2\left(1 - Q_2(x)\right)/\left(1 - Q_0(x)\right) \tag{56}$$

3) More complicated relative parameters have been suggested by Madel [26], Rosin and

RAMMLER [27], and WHITE [28]; they are derived from differences of relative parameters according to paragraph 2:

$$\eta = \frac{A_4}{A_2} - \frac{A_3}{A_1} = \frac{A_5}{A_1} - \frac{A_6}{A_2} = \frac{v_2\,[Q_0\,(x) - Q_2\,(x)]}{Q_0\,(x)\,(1 - Q_0\,(x))}$$

$$= \frac{v_1\,[Q_1\,(x) - Q_0\,(x)]}{Q_0\,(x)\,(1 - Q_0\,(x))} \tag{57}$$

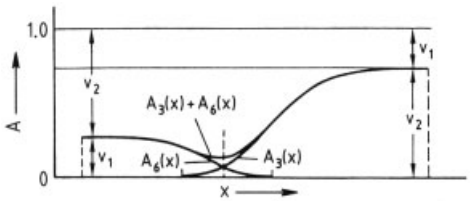

Figure 12. Course of $A_6\,(x)$, $A_3\,(x)$, and $(A_3\,(x) + A_6\,(x))$

One or more of the given parameters or other parameters must be used to quantify the quality of the process, depending on the separation or classification process to be characterized. The parameters should be chosen from those that best characterize the aim of the separation or classification process.

Parameters from the Overall Course of the Grade Efficiency Curve. The most important parameter of this group is the so-called *overall classification or separation (collection) efficiency*, T_{tot}. This parameter is generally used to describe the quality of dust removal systems (overall collection efficiency, → Dust Separation, Chap. 2.). However, it can also be used for classification processes. It corresponds to the already defined relative amount of coarse material, v_2, and may be calculated from the grade efficiency curve $T\,(x)$ and the density distribution curve of the feed material, $q_0(x)$, as follows:

$$T_{\mathrm{tot}} = v_2$$

$$= \int_{x_{\min}}^{x_{\max}} T\,(x)\,q_0\,(x)\,\mathrm{d}x = \int_0^1 T\,(x)\,\mathrm{d}Q_0\,(x) \tag{58}$$

If the density distribution $q_0(x)$ is available as a histogram, the integration may be replaced by a summation:

$$T_{\mathrm{tot}} = v_2 = \sum_{i=1}^{n} T_i\,(\bar{x}_i)\,\Delta Q_{0i} \tag{59}$$

The use of variance and the expected value of the grade efficiency curves for its characterization have been suggested [29], [30]:

If $t\,(x)$ is the first derivative of $T\,(x)$ with regard to x

$$t\,(x) = \mathrm{d}T\,(x)\,/\mathrm{d}x \tag{60}$$

then the expected value $E\,(x) = x$ and the variance, s^2, of the grade efficiency curve may be determined in the usual way as follows:

$$E\,(x) = \bar{x} = \int_{x_{\min}}^{x_{\max}} x\,t\,(x)\,\mathrm{d}x = \int_0^1 x\mathrm{d}T\,(x) \tag{61}$$

$$s^2 = \int_{x_{\min}}^{x_{\max}} (x - \bar{x})^2\,t\,(x)\,\mathrm{d}x$$

$$= \int_0^1 (x - \bar{x})^2\,\mathrm{d}T\,(x) \tag{62}$$

Further parameters have been suggested by MAYER [31–33].

3.2. Influence of Stochastic Errors

The mass balances given in Section 3.1.2 and the evaluation of the grade efficiency curve as described in Section 3.1.5 are valid only under the assumption that the relevant Equations (30) and (33) – (35) leave no remainder. However, in reality Equation (34) does not hold with, e.g., the density distributions as obtained by conventional particle size analysis for the feed, fine, and coarse material. A size-dependent error $E\,(x)$ is observed, and Equation (34) must be rewritten as follows:

$$q_0\,(x) - v_1 q_1\,(x) - v_2 q_2\,(x) = E\,(x) \tag{63}$$

The same applies to Equation (35), which yields

$$Q_0\,(x) - v_1 Q_1\,(x) - v_2 Q_2\,(x) = \varepsilon\,(x) \tag{64}$$

The size-dependent errors $E\,(x)$ and $\varepsilon\,(x)$ are due to errors of particle size analysis when either $q_0(x)$, $q_1(x)$, and $q_2(x)$ or $Q_0(x)$, $Q_1(x)$, and

$Q_2(x)$ are measured. Further errors may arise when v_1 or v_2 are measured directly.

Errors of particle size analysis may be caused not only by erroneous sampling and sample splitting, but also (and sometimes mainly) by errors produced by handling the instruments of particle size analysis.

If the average grade efficiency, T, of a small size class, Δx_i, is calculated, as usual, with $x_{i+1} = x_i + \Delta x_i$ and the average size $\bar{x}_i$:

$$\bar{x}_i = (x_i + x_{i+1})/2 \tag{65}$$

from

$$T(x_i) = \frac{v_2 [Q_2(x_{i+1}) - Q_2(x_i)]}{Q_0(x_{i+1}) - Q_0(x_i)} \tag{66}$$

results are obtained as shown in Figure 13 [34], which reveals a remarkable difference to the grade efficiency curve shown in Figure 10. The results shown in Figure 13 have been obtained with the Acucut B18-classifier from Donaldson Co. Inc., Minneapolis, Minnesota, USA, in a modified form suggested and investigated by Leschonski and Boeck [34]. A fine limestone has been used as feed material. The cumulative size distributions of the feed, fine and coarse material have been obtained with the X-ray sedimentometer, Sedigraph 5000, of Micromeritics Instr. Corp., Norcross, Georgia, USA. The points calculated for different average sizes, $\bar{x}_i$, scatter considerably, and a reliable average grade efficiency curve cannot be obtained.

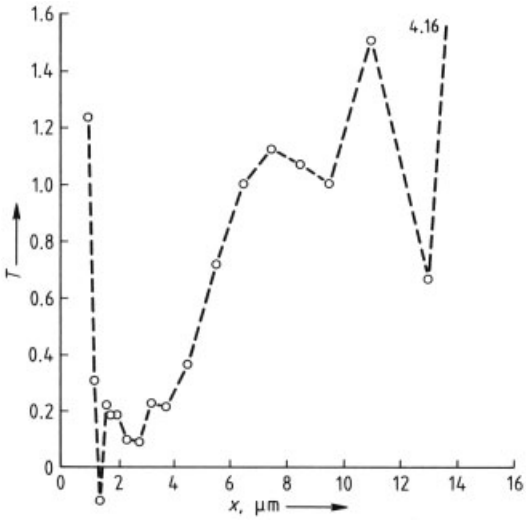

Figure 13. Uncorrected grade efficiency curves, $T(x_i)$

Grade efficiency curves, as shown in Figure 13, are not unusual. Published grade efficiency curves generally follow a rather smooth course because either a graphic or numerical regression technique has been used. Herrmann and Leschonski [16] have suggested a statistically sound correcting method that yields the best possible answer, i.e., the best estimate of the grade efficiency curve from the experimental evidence as obtained by size analysis.

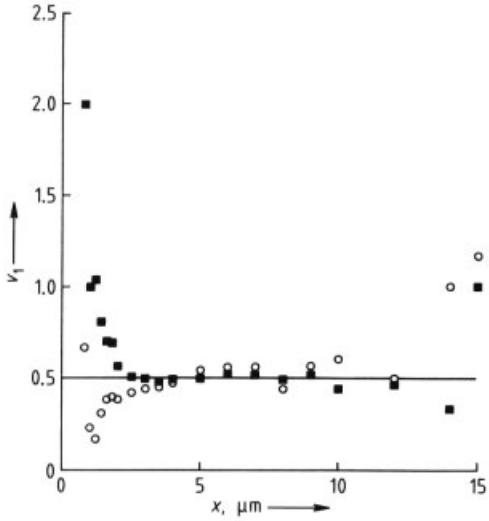

Figure 14. $v_1(x)$ as calculated from erroneous cumulative size distributions $Q_0(x)$, $Q_1(x)$, $Q_2(x)$
■ $\bar{v}_1 = 0.502$; ○ $\bar{v}_1 = 0.506$

3.2.1. Indirect Evaluation of v_1 and v_2

In many practical cases, the relative amounts of the fine and the coarse materials, v_1 and v_2, cannot be obtained from the solids mass flow rates directly, as suggested in Equation (9) and (10). However, if sampling is possible from the three streams of the feed, the fine, and the coarse material, the cumulative size distributions [$Q_0(x)$, $Q_1(x)$, and $Q_2(x)$] of the feed material and the two products may be obtained. As shown in Section 3.1.3, v_1 or v_2 may be calculated from Equation (36). According to this equation, v_1 and/or v_2 must be constant and independent of x if the cumulative distributions $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$ are free of errors. In reality this is not true, which consequently leads to values of v_1

and v_2 that vary with size x (Figure 14). Therefore, suitable equations to calculate an average value v_1 from $v_1(x)$ have been suggested by a number of authors [25], [35–38].

Equation (36) may be interpreted as a linear equation, if it is rewritten as follows:

$$Q_0(x) - Q_2(x) = v_1 [Q_1(x) - Q_2(x)] \tag{67}$$

If $\eta_i = Q_{0\,i} - Q_{2\,i}$ is plotted on the ordinate against $\xi_i = Q_{1\,i} - Q_{2\,i}$ on the abscissa, a series of points is obtained that approximates a straight line through the origin (Fig. 15).

$$\eta_i = v_1 \cdot \xi_i \tag{68}$$

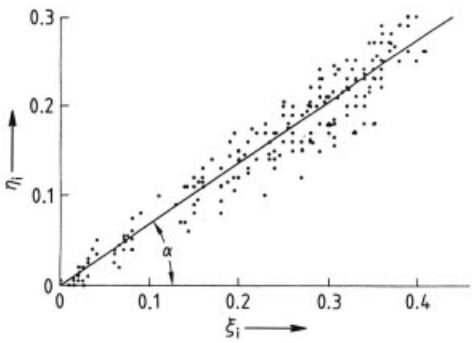

Figure 15. $\eta_i = v_1 \cdot \xi_i$ as obtained from multiple analysis of a hydrocyclone cut [38]
$\tan\alpha = 0.67991$

Regression analysis yields the average slope, for instance: $\bar{v}_1$, for n values of ξ_i and η_i, from

$$\bar{v}_1 = \sum_{i=1}^{n} \xi_i \eta_i \Big/ \sum_{i=1}^{n} \xi_i^2 \tag{69}$$

Equation (69) coincides with the solution suggested by GRUMBRECHT [35], obtained when $\bar{v}_1$ is calculated under the assumption that

$$\sum_{i=1}^{n} \varepsilon_i^2 = \min \tag{70}$$

and the constraint that $v_1 + v_2 = 1$ holds. Equation (68) is found when Equation (64) is substituted into Equation (69) and this equation is solved by using Lagrange's method.

However, the method has the general disadvantage that Equation (68) is valid only under the assumption that the errors of particle size analysis are attributed only to the η_i values. The ξ_i values are taken to be free of errors. However, strictly speaking, neither η_i nor ξ_i are free of errors, and the minimum of the squared errors should be calculated perpendicular to the straight line drawn through the measured points under the angle α (Fig. 16).

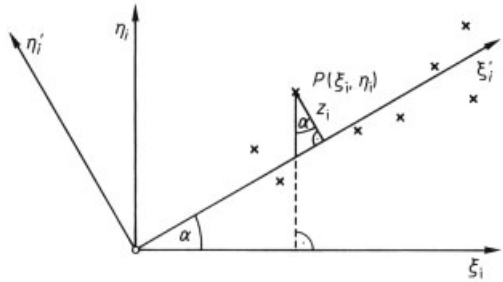

Figure 16. Definition of the deviation z_i

The most probable $\bar{\alpha}$, is obtained with $\tan\bar{\alpha} = \bar{v}_1$, from

$$\sum z_i^2 = \sum (\xi_i \sin\alpha - \eta_i \cos\alpha)^2 \overset{!}{=} \min \tag{70a}$$

The solution yields

$$\tan 2\bar{\alpha} = \frac{2 \sum_{i=1}^{n} \xi_i \eta_i}{\sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \eta_i^2} = a \tag{71}$$

with

$$\bar{\alpha} = \frac{\arctan a}{2} \tag{72}$$

Numerical values of $\bar{v}_1$ calculated by using Equation (68) or (70) for the same set of results differ very little. Therefore, in most cases of application it is advisable and legitimate to use the much simpler Equation (68).

Regression analysis not only yields average values of $\bar{v}_1$, but also its standard deviation or variance, as well as its confidence interval. In the case of Equation (68), the variance, s^2, may be calculated from Equation (73) [39]:

$$s^2 = \frac{1}{(n-1)\sum \xi_i^2} \left(\sum \eta_i^2 - \frac{(\sum \xi_i \eta_i)^2}{\sum \xi_i^2} \right) \tag{73}$$

with the confidence interval being defined in the usual way by

$$\pm t \cdot s / \sqrt{n} \tag{74}$$

where t equals the student factor, i.e., approximately 1.96 with a probability of 95 % and $n > 25$. Equation (73) is, in fact, a well-known equation in regression analysis, but unfortunately has not been used in the characterization of a classification process. It should be remembered that

$$1/v_1 = \dot{m}_0/\dot{m}_1 \tag{75}$$

with $1/v_1$ being the so-called recirculating load. Therefore, in a continuous classification process, $1/v_1$ represents the average ratio of the mass throughput of the feed, $\dot{m}_0$, to the mass throughput of the fine material, $\dot{m}_1$.

3.2.2. Evaluation of the Grade Efficiency Curve, *T* (*x*)

As shown in Figure 13, erroneous cumulative size distributions $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$ lead to nonideal grade efficiency curves. Therefore, Herrmann and Leschonski suggested a statistical correction of the measured values that not only reduces the error as obtained, but also allows the calculation of the best statistical estimate of the grade efficiency curve, $T(x)$, based on the given data of particle size analysis [16]. According to this, corrected cumulative distribution values $K_0(x)$, $K_1(x)$, and $K_2(x)$ are derived from the measured cumulative distribution values $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$, for which the mass balance holds:

$$K_0(x) - v_1 K_1(x) - v_2 K_2(x) = 0 \tag{76}$$

A least squares method was applied with $1/\sigma^2$ as weight for the square of the deviation $(Q-K)^2$ to derive Equations (77) – (80) for the correction, where σ^2 represents the variance or its estimated value s^2, as calculated from q_1 values, obtained by multiple analysis of the same product. The statistical correction yields the following equations:

$$K_0(x) = Q_0(x) - \frac{\sigma_0^2(x)}{\alpha(x)}\varepsilon(x) \tag{77}$$

$$K_1(x) = Q_1(x) - \frac{v_1\sigma_1^2(x)}{\alpha(x)}\varepsilon(x) \tag{78}$$

$$K_2(x) = Q_2(x) - \frac{v_2\sigma_2^2(x)}{\alpha(x)}\varepsilon(x) \tag{79}$$

with

$$\alpha(x) = \sigma_0^2(x) + v_1^2\sigma_1^2(x) + v_2^2\sigma_2^2(x) \tag{80}$$

The correction terms on the right-hand side of Equations (77) – (79) depend not only on the error variance, σ^2, of each individual measured value relative to the weighted sum of the variances, $\alpha(x)$ (Eq. 80), but also on the error $\varepsilon(x)$ (Eq. 64). When the suggested correction method is applied, the error distributions σ^2 or s^2 must be known for the particular case of application and the particle size analysis instruments used for the determination of $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$. These curves can be obtained by repeated analyses of the feed and the fine and coarse products. As shown in Figure 17, the error distributions, $s(x)$, may differ for the three streams.

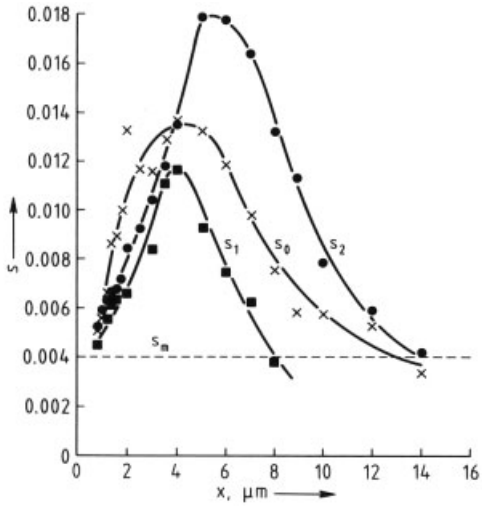

Figure 17. Error distributions $s_0(x)$, $s_1(x)$, and $s_2(x)$

In Figure 17 the average error, $\bar{s}$, as obtained from 21 measurements of $Q_0(x)$, $Q_1(x)$, and $Q_2(x)$, is plotted against the equivalent settling rate diameter, $x = x_w$. On a preliminary basis, the three differing curves may be described by Equation (81):

$$\sigma^2(x) = s_m^2 + C(x_{50})\,Q(x)\,[1 - Q(x)] \tag{81}$$

If the suggested correction is applied to grade efficiency curves, as shown in Figure 13, the solid lines of Figure 18 are obtained.

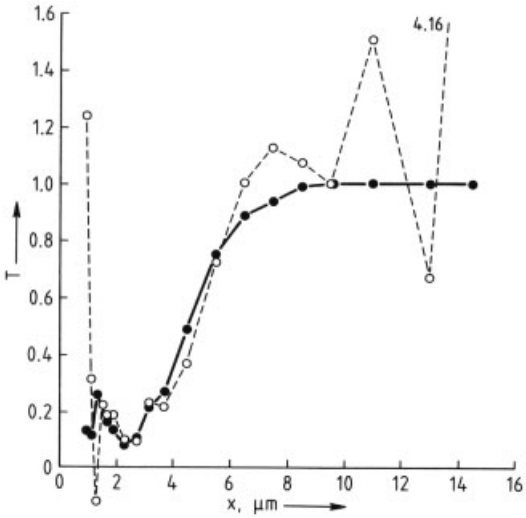

Figure 18. Corrected grade efficiency curves using $\sigma^2(x)$ but only one set of results

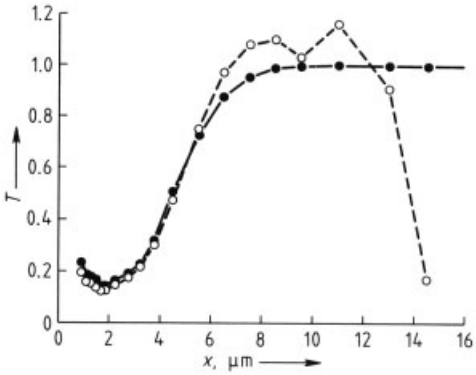

Figure 19. Average and average corrected grade efficiency curves

Apart from the error distributions $s^2(x)$, only one set of results of particle size analysis has been used in this presentation. The suggested correction smooths the grade efficiency curves especially well in the coarse size range.

From the results obtained, the obvious question can be answered, whether the use of Equations (77) – (80) can be dropped altogether, if an average grade efficiency curve is calculated from average cumulative size distributions of the products. The answer is no, even though each product had been analyzed 21 times in this case.

The dotted line in Figure 19 shows the average grade efficiency curve. While the fine end seems to be right, the coarse end neither reaches unity nor does it seem to follow the right course. The best answer is obtained with multiple analyses and an application of the suggested correction. The full line in Figure 19 indicates the difference.

This example clearly shows the necessity for multiple analyses. These can be performed on an economical basis only if particle size distributions are analyzed quickly and, if possible, on-line. Therefore, the development and use of rapid-response methods or on-line methods of particle size analysis are more important than generally recognized and acknowledged. A complete theory of the correction method and further statistical measures have been published [40], [41].

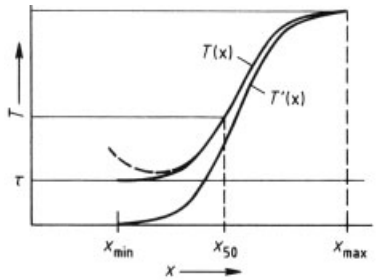

Figure 20. Influence of a splitting process in the classifier on $T(x)$

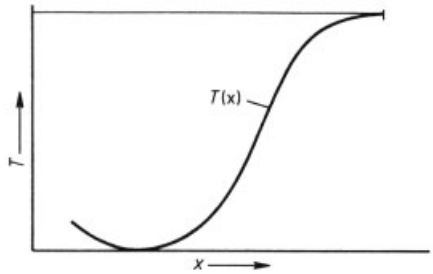

Figure 21. Deviation from the ideal grade efficiency curve caused by incomplete dispersion of the feed material in the classifier

3.3. Influence of Systematic Errors

Various effects can cause systematic deviations from the ideal course of a grade efficiency curve, as shown in Figures 20, 21, 22. Such effects are

1) systematic analytical errors of sampling and sample splitting

2) superposition of classification and splitting within the classifier
3) undispersed agglomerated fine particles which are transferred to the coarse product
4) comminution of the feed material in the classifier.

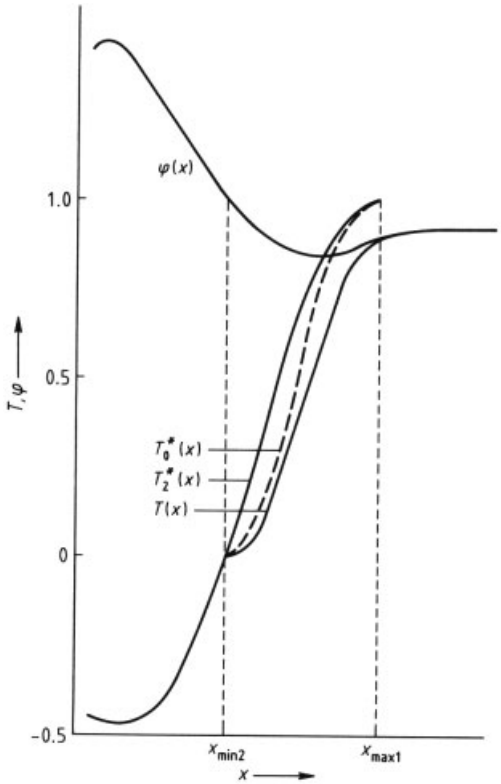

Figure 22. $\varphi(x)$ and $T(x)$ if comminution occurs in a classification process

Under the assumption that sampling from the product streams and the splitting of the samples of the three products have been performed with greatest care, the error listed under 1) may be disregarded. The others, however, must be taken into account.

3.3.1. Splitting in the Classifier

If the grade efficiency curve, as shown in Figure 20, does not drop to 0 at small particle sizes, but ends parallel to the abscissa at a certain grade efficiency $T = \tau$, a splitting process is superimposed on the classification. A modified grade efficiency curve $T'(x)$ may then be calculated [18]:

$$T'(x) = (T(x) - \tau) / (1 - \tau) \tag{82}$$

In a hydrocyclone classification, the value of τ describes the so-called *dead flux*. $T'(x)$ resembles the ideal grade efficiency curve of Figure 10; it extends between 0 and 1. When this technique is used, it should be remembered, the effective grade efficiency curve is represented by $T(x)$ only; however, the use of $T'(x)$ and τ may simplify the description of the grade efficiency curve and may give additional information. In many technical applications, it is usually advisable to minimize τ.

3.3.2. Incomplete Dispersion of the Feed Material

If the feed material is not properly dispersed before it enters the classification zone, or even within the classification zone, coarse agglomerates of fine particles may be taken by the classifier as coarse particles and may consequently appear in the coarse product. On the other hand if the samples are well dispersed, e.g., in a sedimentation liquid, during the analysis of the three products, the agglomerates disappear. The analysis then detects more fine particles than the classifier did, and the calculated grade efficiency curve deviates from the ideal one, as shown in Figure 21.

The classification process itself is not to blame for this result. Better dispersion of the feed material, if possible before entering the classification zone, reduces this effect. Therefore, the feed material should pass a separate dispersion zone before it enters the classifier.

3.3.3. Comminution in the Classifier

If the feed material is comminuted in the classification zone, i.e., if the classifier partly acts as a grinding machine, the fine feed material differs from the one obtained from sampling the original feed. New fine material is produced and an equivalent amount of coarse particles disappears.

This error has already been described and is defined in Equation (63):

$$q_0(x) - v_1 q_1(x) - v_2 q_2(x) = E(x) \quad (63)$$

It can be used to formulate a new variable, $\varphi(x)$, which indicates whether the deviations observed are either stochastic or systematic and which course they follow [42]:

$$\varphi(x) = \frac{v_1 q_1(x) + v_2 q_2(x) +}{q_0(x)} = 1 - \frac{E(x)}{q_0(x)} \quad (83)$$

If the sum of the weighted density distributions of the fine and the coarse material equals the density distribution of the feed, then $E(x) = 0$ and $\varphi(x) = 1$. Stochastic deviations of $\varphi(x)$ from a value of 1 indicate stochastic errors, $E(x)$, in the mass balance of Equation (63). The suggested error correction and multiple analysis should then be used in the evaluation of the grade efficiency curve, as described in Section 3.2. However, if $\varphi(x)$ deviates systematically from 1, then comminution of the feed material in the classifier might be responsible. A typical example is given in Figure 22.

Table 2 shows the relevant equations and an explanation of the subscripts chosen. If $q_2^*(x)$ and $q_0^*(x)$ are introduced into the equation for $T_2^*(x)$ and $T_0^*(x)$ (Table 2), the following is obtained:

$$T_2{}^*(x) = T(x) + [1 - \varphi(x)]$$

$$= 1 - v_1 q_1(x) / q_0(x) \quad (90)$$

$$T_0{}^*(x) = T(x) / \varphi(x) = 1 / \left(\frac{v_1 q_1(x)}{v_2 q_2(x)} + 1 \right) \quad (91)$$

While $T_0^*(x)$ can be calculated from the fine and the coarse materials, $T_2^*(x)$ depends on the fine and the feed materials. The function $T_0^*(x)$ covers the size range between $x_{\min 2}$ and $x_{\max 1}$ and is not influenced by the course of $\varphi(x)$ as soon as $T(x) = 0$. Therefore, neither $T(x)$ nor $T_0^*(x)$ gives evidence of the increased amount of fine material at small particle sizes due to comminution. This information can be obtained only from $T_2^*(x)$. Therefore, $T_2^*(x)$ should be calculated as soon as $T(x)$ remains smaller than 1 at coarse particle sizes.

4. Sample Splitting

Particle size analysis may be performed only if a suitable sample of particulate matter to be analyzed has been taken from the bulk material in the factory or delivered to the laboratory. The sample mass may vary from less than 1 mg to more than several kilograms depending on the method of analysis or the particular instrument chosen, and the average size of the particles and their spread. The analysis sample must be taken from the parent population of particulate matter. Its composition should equal the composition of the parent population in all aspects; it should be a so-called *representative sample.* Since the overall mass of the parent population may be of arbitrary size, the mass of the analysis sample must be obtained by repetitive splitting of the original sample. The answer obtained after the analysis sample has been analyzed may represent the particle size distribution, the surface area, etc., of the parent population more or less accurately, depending on the quality of the splitting process.

Differences may occur due both to statistical variations of the composition of the analysis samples and to systematic deviations in the composition caused by the splitting process.

Statistical variations of the samples can generally be decreased by using an appropriate amount of particulate material for the analysis, if the instrument that is applied permits this. Systematic and statistical errors can also be reduced if the analysis sample consists of a large number of much smaller samples.

The methods applied in sample splitting may be subdivided into two groups, those for powders and those for suspensions and emulsions. Sample splitting of powder is achieved by *coning* and *quartering*, or by repeatedly splitting the laboratory sample (10 – 100 kg) in stationary or rotating rifflers. Sample splitting from suspensions or emulsions can be achieved by withdrawing an appropriate small amount of suspension or emulsion from the bulk material after the laboratory sample has been mixed thoroughly.

5. Sedimentation Analysis

Sedimentation methods use the effect that single particles settle, for instance, under the influence of gravity in a static liquid medium with

Table 2. Definition of $\varphi(x)$ and various grade efficiencies

Measured values		Calculated values	Equation
Feed material:	$q_0(x)$	$q_0^*(x) = v_1\, q_1(x) + v_2\, q_2(x) = q_0(x) - E(x)$	(84)
Fine material:	$q_1(x)$, v_1		
Coarse material:	$q_2(x)$, v_2	$q_2^*(x) = \frac{q_0(x) - v_1\, q_1(x)}{v_2} = q_2(x) - \frac{E(x)}{v_2}$	(85)
		$\varphi(x) = \frac{q_0^*(x)}{q_0(x)} = 1 - \frac{E(x)}{q_0(x)}$	(86)
Grade efficiencies:		$T(x) = \frac{v_2\, q_2(x)}{q_0(x)}$	(87)
		$T_2{}^*(x) = \frac{v_2\, q_2^*(x)}{q_0(x)} = T(x) + (1 - \varphi(x))$	(88)
		$T_0{}^*(x) = \frac{v_2\, q_2(x)}{q_0^*(x)} = \frac{T(x)}{\varphi(x)}$	(89)

constant velocity after a short and negligible period of acceleration. The settling rates that are measured may be used to calculate the size of a sphere, e.g., the equivalent settling rate diameter x_w, travelling with the same speed in the same environment.

In most cases the analysis starts with a *homogeneous suspension* in which the particles are evenly, or better, randomly distributed throughout the liquid volume. The individual solid particles then start to settle, with settling rates depending on the particles' size, shape, and density. The suspension density, which was constant at the beginning of the experiment, starts to change, forming an image of the cumulative distribution of settling rates with respect to the height of the sedimentation column, which can be used to analyze the sample. The physical methods used to measure the time- and height-dependent change of suspension densities define and distinguish the different instruments of sedimentation analysis.

5.1. Calculation of the Equivalent Settling Rate Diameter, x_w

The liquid used in a sedimentation analysis should neither physically nor chemically affect the particles to be analyzed. Furthermore, it should not cause reagglomeration of the solid particles, which in most cases is guaranteed by adding small amounts of suitable surfactants. The liquid density must be smaller than the solids density, and its viscosity should match the size distribution under investigation, avoiding, for instance, extremely long measuring times for the smallest particles [44].

The stationary motion of a single spherical particle in a static, infinite liquid under the influence of gravity can be calculated from the force balance: weight minus buoyancy force equals the drag force. With spherical particles and laminar flow conditions, i.e., under the validity of Stokes law, the drag coefficient c_d equals

$$c_\mathrm{d} = 24/Re \tag{92}$$

The dependency between equivalent settling rate diameter, x_w, and stationary settling rate, w_g, in the gravity field can then be calculated from

$$x_w^2 = \frac{18\eta w_g}{(\varrho_\mathrm{s} - \varrho_\mathrm{fl})\, g} \tag{93}$$

or

$$v_\mathrm{g} = \frac{(\varrho_\mathrm{s} - \varrho_\mathrm{fl})\, g x_w^2}{18\eta} \tag{94}$$

The Reynolds number, *Re*, as used in Equation (92), equals

$$Re = w_g x_w \varrho_\mathrm{fl}/\eta \tag{95}$$

The equivalent settling rate diameter, x_w, of particles settling in a centrifugal field with the acceleration

$$a = r\omega^2 \tag{96}$$

where the centrifuge rotates with constant angular velocity, ω, can be obtained from

$$x_w^2 = \frac{18\eta}{(\varrho_s - \varrho_{fl})} \cdot \frac{\ln r/r_i}{\omega^2 t} \tag{97}$$

where r_i represents the radius of the liquid surface. The equivalent settling rate diameter, x_w, as calculated using Equations (93) and (97) describes the settling behavior of a sphere in an infinite static liquid. In reality, however, settling takes place in the neighborhood of 10^6 other particles. Furthermore, the particles are in most cases not spherical and the vessels used have limited dimensions. Walls and the bottom of the vessel decrease settling rate, but for practical applications the influence of both may be neglected.

The settling rate increases under certain conditions with increasing solids volume concentrations C_V; it reaches a maximum at approximately $C_V = 1\,\%$ and then decreases to 0 [45–47]. The settling rate maximum disappears if fine particles form the suspension and if the distribution of settling rates is widely spread. The settling rate is decreased from the stationary settling rate, w_g, at increasing solids concentration by the increased liquid flow that is caused by the settling of the particles into a confined liquid volume. The settling rate only increases, if hydrodynamicallylinked particles form clusters that settle faster than neighboring single particles [45]. Therefore, there is an upper limit to solids concentration in sedimentation analysis.

The solids volume concentration should be kept smaller than $C_V = 2 \times 10^{-3} = 0.2\,\%$. An upper limit, where deviations in settling rate from w_g may still be small, is $C_V = 1\,\%$.

With particles smaller than 10^{-6} m and settling in liquids, the Brownian motion is no longer negligible. This diffusion process produces wider size distributions than are present in reality. It can be avoided only if the analysis is repeated in a centrifuge.

Settling of small particles in air is influenced by the *slip velocity*. If the mean free path, $\bar{\lambda}$, of the air molecules is of the order of the average size of the particles, the settling rate is increased and the *Cunningham correction* must be taken into account [48]:

$$w/w_g = 1 + A\bar{\lambda}/x \tag{98}$$

A = 1.73 for oil droplets in air,
A = 0.7 for rough sphere in air

In addition, the settling rate may be strongly influenced by *density convection currents.* They may be caused either by temperature gradients in the liquid or density differences in the suspension. Therefore, convection currents can be avoided only if the density of the suspension increases toward the bottom of the vessel. Cooling or evaporation of liquid at the top of the liquid column should be avoided, as well as temperature differences greater than 0.01 °C/min caused by the temperature control of thermostats. Differences in suspension density are caused mainly by the inclined walls of the sedimentation vessel and solid obstacles, e.g., a weighing pan or a withdrawal tube immersed in the suspension.

5.2. Methods of Sedimentation Analysis

The methods and instrumentation used in sedimentation analysis can be subdivided into two principal groups:

1) *Suspension techniques:* at the beginning of the analysis the suspension is evenly distributed throughout the liquid volume
2) *Superimposed layer techniques:* at the beginning of the analysis the suspension is superimposed in a thin layer on top of a solids-free liquid column.

- Suspension
  - Gravity field
    - Incremental
    - Cumulative
  - Centrifugal field
    - Incremental
    - Cumulative
- Superimposed layer
  - Gravity field
    - Incremental
    - Cumulative
  - Centrifugal field
    - Incremental
    - Cumulative

Both methods may be applied in a gravity or in a centrifugal field. Furthermore, the measurement of the amount of particles present in a certain size class, or smaller than that certain size, is either performed in a measuring volume, which is small in comparison to the overall settling height, or includes all particles above or

below the plane of measurement. The first group comprises the *incremental methods*, the second one the *cumulative methods*. Therefore, eight principal possibilities are distinguished:

The measurements may be performed at a fixed height with respect to time, or after a fixed time with respect to settling height; in this way, the number of possible methods may be at least doubled. The number of instruments is much larger than the number of principal methods because different physical methods are used to measure the concentration, mass, or volume of particles in a thin suspension layer or below or above the height of measurement. The measurements are performed gravimetrically; they use the scattering and absorption of electromagnetic radiation, as well as buoyancy, hydrodynamic pressure, and other physical variables that depend on solids concentration, surface, or number.

5.2.1. Suspension Techniques

Pipette Method. One of the best known incremental gravitational suspension methods is the pipette method [49–52]. With this method, the cumulative weight or mass undersize, Q_3, is determined from the ratio of two volume concentrations measured at a fixed height at the beginning of the analysis and after verying times t, (Eq. 99).

$$Q_{3\,t} = C_{V\,t}/C_{V0} \tag{99}$$

Fixed suspension volumes are withdrawn, the solids contents of which are determined, e.g., by weighing after the liquid is evaporated.

Diver Method. Berg developed the diver method [53]. A small spherical diver, the density of which lies between the density of the solids-free liquid and the density of the homogeneous suspension, is submersed in a suspension that has already formed its settling-rate-dependent concentration gradient. The diver then settles to the suspension layer of the same density and its position is measured. The suspension density and Q_3 are correlated according to Equation (100):

$$Q_3 = (\varrho_\mathrm{d} - \varrho_\mathrm{fl}) / [(\varrho_\mathrm{s} - \varrho_\mathrm{fl})\, C_{V0}] \tag{100}$$

Therefore, the missing settling rate is determined from the diver position and the length of time the experiment has taken.

Optical Measurement. Other incremental techniques use the attenuation of a beam either of white light or of γ-rays by the particles. Lambert–Beer's law of attenuation yields the theoretical background for photosedimentometers [54].

Sedimentation Balance. The best known example for cumulative suspension methods is the sedimentation balance first suggested by Odén in 1916 [52], [55]. This method determines $Q_3(w_g)$ from the mass of particles, m_t, collected on a submerged balance pan at the lower end of a suspension column. The quantity $Q_3(w_g)$ can be calculated from the curve m_t vs. t by using Equation (101):

$$m_\infty Q_{3t} = m_t - t\,\mathrm{d}m_t/\mathrm{d}t \tag{101}$$

The methods mentioned have been applied in both gravitational and centrifugal applications. Centrifugal methods, which are generally used for particles smaller than ca. 5×10^{-6} m, apply the same techniques as gravitational ones; however, the mathematical evaluation of the data is more sophisticated.

5.2.2. Superimposed Layer Techniques

The superimposed layer technique was first applied in 1930 by Marshall [56]. The superposition of a thin layer of suspension on top of a solids-free liquid free from disturbances is the basic problem of the superimposed or stratified layer technique.

In most cases the analyst experiences instabilities at the interface between the suspension and the solids-free liquid during or after the superposition. These instabilities depend on a number of variables [57]. To achieve a stable superposition free of instabilities, the *density gradient technique* must be applied [58].

Incremental concentration measurements in gravitational or centrifugal instruments and the direct measurement of a density distribution can be performed best with conventional photometers. Examples for cumulative gravity methods

have been the *Micromerograph* [59], where the superposition is performed in air, the *Woods–Hole–Rapid–Sediment Analyzer* [60], and other methods [61], [62].

The methods are applicable mainly to size distributions with particles coarser than ca. 60×10^{-6} m.

Modern centrifuges no longer use beakers, but disk-shaped rotors. A typical example is the *Joyce–Loebl centrifuge*. A perspex rotor is filled with solids-free liquid at the beginning of the experiment, and the suspension sample to be analyzed is introduced near the axis of rotation with a hypodermic syringe. While adhering to the rotating wall, the suspension is radially accelerated and finally forms a thin suspension ring on top of the solids-free liquid, from which the particles start to settle. Instabilities can be avoided only if the density of the solids-free liquid surpasses at any radial position the suspension density formed by the particles and the surrounding liquid at slightly smaller radii.

6. Classification Methods

With all classification methods of particle size analysis, a given sample of particulate matter is separated into at least two size or settling rate classes, one comprising particles smaller and the other coarser than a certain size. The characterization and a general description have already been given in Chapter 3. Sieve analysis and air classification are the most important classification techniques used in particle characterization.

6.1. Sieve Analysis

The principle of sieve analysis is generally well known. Several sieves, one on top of each other, form a set, with the size of the sieve openings decreasing from top to bottom. A representative sample of the material to be analyzed is fed to the coarsest sieve, and the sieve column is vibrated either mechanically or electro-mechanically. The particles pass through the sieve openings, and the coarser ones are retained on the sieves. As soon as the particles cease to pass through the openings, the vibration is stopped. The amount of particles on each individual sieve is weighed, and the histogram of the density distribution is obtained by weight.

The mass of particles retained between two sieves can be measured relatively simply and with little error. The problem of sieve analysis lies in the determination of the cut sizes of the individual sieves. In most cases of practical application, the nominal width of a sieve opening is taken to represent its cut size. However, since the size of the openings of a sieve follows a distribution, the cut size does not necessarily coincide with the nominal width of the sieve openings, and a calibration procedure becomes necessary for high-quality sieve analyses.

Sieve cloths are standardized in most countries and internationally:

Federal Republic of Germany:	
DIN 4188	woven wire test sieves
DIN 4187	perforated plate test sieves
United Kingdom:	
BS 410	woven wire test sieves
United States:	
ASTME 11	woven wire test sieves
ASTME 161	micromesh sieves (electroformed sieves)

The range of the widths of sieve openings varies from 5.5 µm (ASTM 161) or 25 µm (DIN 4188) to 125 mm, thus covering more than four powers of ten.

The particles can pass through the openings of a sieve cloth only if the sieve cloth and the particles move relative to each other. The relative motion may be produced either by a motion of the sieve cloth itself or when the particles are transported to the sieve opening with an air or liquid flow. Therefore, the quality of sieving depends on how often a particle is presented to a free sieve opening within a given period of time.

The sieving process also depends on further variables, which are correlated with the material to be sieved and the sieving principle (→ Screening). The latter includes intensity and frequency of relative motion, sieving aids, and the aperture width distributions at the sieves. Variables that depend on the material to be sieved are the solids loading on each individual sieve, the size distribution that is mainly the amount of particles, the size of which is similar to the width of the sieve openings, and particle shape and their tendency to clog the openings of the sieve. Furthermore, sieving depends on the

tendency of fine particles to agglomerate, i.e., on the absolute size and the moisture content of the material to be sieved.

Many authors have investigated theoretically and experimentally how these variables influence sieving results [63–68]. However, until now a universal formula has not been found, describing the course of sieving with time, which takes into account the above variables.

Cut size, sharpness of cut, and the amount of material that has passed through a sieve depend on and increase with sieving time. As soon as a reasonable sharpness of cut has been obtained, cut size and undersize weight should be determined independently. A simple and reasonably accurate method to measure an equivalent to a cut size has been suggested [70], [71]. A counting and weighing technique is applied to the fraction of particles passing through the sieve just after stopping the shaking or sieving action. The average volume diameter, x_V, of this fraction is taken to be the cut size of the sieve [71]. It can be calculated from the mass, m_s, of n particles collected:

$$x_V = \sqrt[3]{\frac{6\,m_s}{\pi\,\varrho_s n}}$$

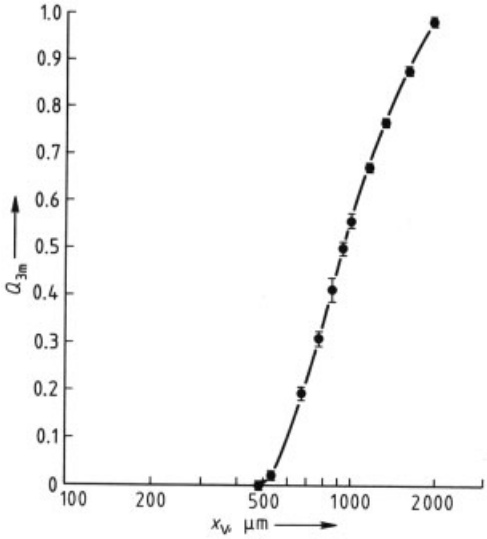

Figure 23. BCR certificate with extremely low scatter. BCR sample 131 was used and regression analysis performed.

A subcommittee of the Bureau Commutative de Référence (BCR) has prepared a number of quartz samples for calibration purposes, using the above technique [72].

Results obtained for a coarse quartz are given in Figure 23. The cumulative distribution by mass shows an extremely low scatter. When this sample is used as calibration material, it is fed to a stack of sieves and sieving is performed under the same conditions as those chosen for the calibration. The over- or undersize weight values of each size are determined and the analytical cut sizes, x_a, can easily be obtained from the cumulative distribution curve of the calibration material (see Figure 23).

6.2. Air Classification

Particles introduced into the flow field of the classification zone of an air classifier travel on trajectories to different points or areas of the classification zone, and are separated into at least two fractions, a fine and a coarse one. These trajectories depend on size, or more accurately, on settling rate. A distinction is made between the *counterflow* and the *cross-flow principle,* which may be applied independently or in combination. The terms counterflow and cross-flow have been used to describe the movement of the coarse particles either against or partly perpendicular to the main air flow (→ Air Classifying).

Air classification as applied in particle size analysis should at least fulfill the following conditions:

1) It should be possible, either to precalculate the cut size from the main variables or to use a calibration curve for its determination.
2) Sharpness of cut should be as high as possible, and/or the amount of misplaced material should be the same in the fine and in the coarse fraction.
3) The coarse fraction should be recovered without any losses.

6.2.1. Cross-flow Classifiers

Cross-flow classifiers have been used mainly for the sampling of fine particles from aerosols and their size-dependent deposition on a carrier medium suitable for further size analysis. The principle is explained in Figure 24.

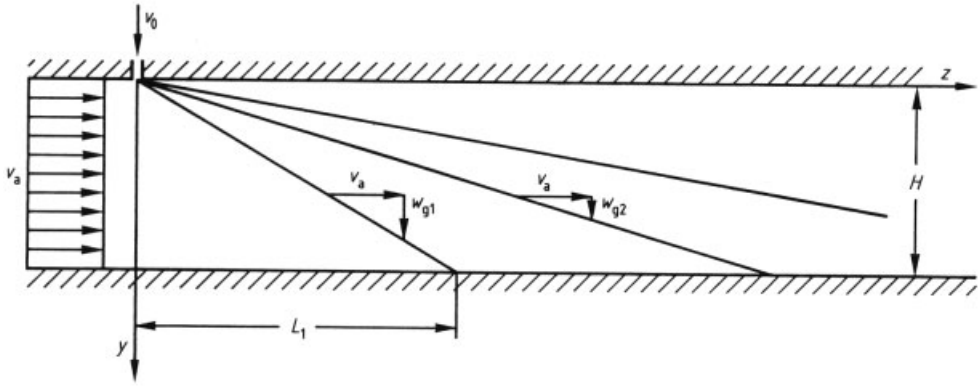

Figure 24. Principle of a cross-flow classifier

A cross-flow classifier basically consists of a horizontal tube of rectangular cross section. Solids-free air enters it from one side, and the particles to be classified are fed into the airstream through a small slit in the top wall. If plug flow and a negligible particle entry velocity, v_0, are assumed, the particle trajectories are straight lines, the slope of which is equal to the ratio of settling rate, w_g, to the average air velocity, v_a, if the particles settle perpendicular to the flow under the influence of gravity. If the assumed plug flow is replaced with a laminar flow profile of the same average velocity, the particle trajectory changes, but the position L_1 at which a particle of settling rate w_{g1} meets the bottom remains the same. The straight lines, i.e., the particle trajectories in Figure 24, are represented by Equation (103):

$$y = z \cdot w_g / v_a \tag{103}$$

and the point of deposition is equal to:

$$L = v_a H / w_g \tag{104}$$

Thus, the particles are spread at the bottom in a pattern that depends on settling rate. The particles may be removed, for instance, by adherence to a transparent, sticky tape and analyzed microscopically with respect to their geometrical size distribution. Typical examples for gravity applications have been suggested [73–75].

The above cross-flow principle may also be used with forces other than gravity that cause the particles to move perpendicular to the flow. These are centrifugal forces, electrical forces as in electrostatic precipitators, and thermal forces as used in thermal precipitators.

Aerosol Centrifuge. The first aerosol centrifuge using the cross-flow principle was the *Conifuge* [76–79]. Its design has been changed to a spiral channel in a plane plate [80], which has later been improved [81]. A comprehensive description of the development of aerosol centrifuges is given in [82].

Impactor. For several years the use of impactors has become increasingly important. These instruments can be used for the analysis of solid particles or droplets. Its principle is shown in Figure 25. One stage of the impactor consists of a jet of diameter d arranged at a distance s perpendicular to a plate of diameter D. The air flow passing the jet impinges on the plate and bypasses it, forming a so-called stagnation point flow. Some of the coarse particles entrained in the air flow cannot follow the flow lines because of their inertia; they collide with the plate surface and are retained there. The impact depends not only on the velocity of the air and the particle when leaving the jet, but also on the particle's starting position in the jet orifice and the other variables already mentioned when the principle was described. A size-dependent grade efficiency curve is obtained, a great number of which have been calculated [83], [84]. The first impactor was described in 1945 [85]. If several impactor stages are connected in series with the cut size of each stage continuously decreasing, a *cascade impactor* is obtained.

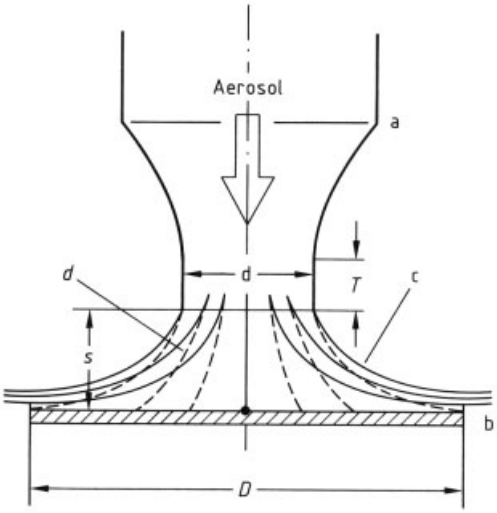

Figure 25. Principle of an impactor
a) Nozzle; b) Plate; c) Stream line; d) Particle trajectory

6.2.2. Counterflow Classifiers

A distinction is made between counterflow classifiers in the gravity or centrifugal field.

Conventional *gravity counterflow classifiers* consist of a long, cylindrical tube of circular or rectangular cross section arranged vertically. The tube is tapered at its bottom, leading into a receptacle for the coarse material. Air is introduced at the bottom end, dispersing the particles which are then introduced into the rising flow. Fine particles follow the flow to the top end, while the coarse ones settle [86–89].

Centrifugal counterflow classifiers are commonly called *spiral classifiers.* The classification takes place in a flat, rotationally symmetrical classification zone. Air is introduced tangentially at its periphery, passing through it on spiral flow lines. Fine particles leave the classification zone with the air through a central orifice, while the coarse particles circle at the outer periphery and are removed from there continuously. The cut size of the classification process can be calculated from a radial force balance, which yields the following:

$$v_r = w_{at} \tag{105}$$

The radial velocity component of the air, v_r, equals the centrifugal settling rate, w_{at}, of the particle, obtained with the acceleration a:

$$a = u_{\mathrm{p}}^2/r \tag{106}$$

Spiral classifiers were first investigated by Rumpf [90]; a laboratory spiral classifier has also been developed [91].

7. Counting Methods

In all counting methods, the measure of quantity is a number. To characterize the "size" of an irregularly shaped particle, many different physical particle properties are used. Number distributions determined by counting methods can be obtained only if the size of each particle is measured from its image. The physical properties used to measure size are as follows:

1) characteristic linear dimensions, chord lengths, the circumference, or the area of the projected areas of the particles
2) particle volume
3) the disturbance of an electromagnetic field caused by the particle passing through it; the most important techniques use the disturbance of an electrical field (Coulter principle) or of a light beam (optical methods).

The measurement may be performed with the particle ifself or its image; hence, the distinction is made between *direct* and *indirect counting methods*, respectively. Proper dispersion of a representative sample to be analyzed is one of the basic assumptions of all counting methods. Each single particle must be presented to the measuring zone of the instrument used. Since this can only be achieved at a negligible solids concentration, an error caused by the simultaneous presence of two or more particles in the measuring zone must be accepted. If a Poisson process is assumed to govern the particle approach to the measuring zone, and the probability for two particles being present in the zone is 5 %, the number concentration, C_N, per volume, V, of fluid may be calculated from Equation (107):

$$C_N = 0.05/V \tag{107}$$

In most cases the sample size is fairly small and statistical errors may arise because only a limited number of particles are measured. In addition, all

counting methods have a lower limit of recognition, because, e.g., optical microscopes cannot produce an image if the particle size drops below the wavelength of light. In direct counting methods, among other reasons, a lower size limit exists when the signal created by the particle is of the same order as the electronic noise of the amplifier. The total number of all particles passing through the measuring zone is, therefore, unknown in most cases, and the resulting number distributions are shifted toward coarser sizes.

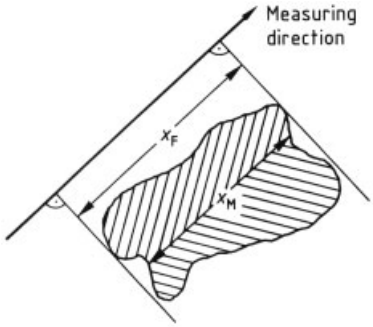

Figure 26. Feret and Martin diameter of a particle projection

7.1. Image Analysis

Image analysis is an indirect counting method. The scale may vary from 1 : 1000 and more (image reduction) to 10^5 : 1 (image magnification), depending on the optical method used, (camera or optical or electron microscope). The problems and methods used in sample preparation are described in [92].

Of all particle size analysis methods, image analysis presumably yields the most definite size and shape description. However, if high accuracy is required, it is one of the most time-consuming and labor-intensive methods in use. Particle size is determined from the projected, hence two-dimensional, image of the particle. Figure 26 shows the sharp outline of the projected area of a particle. The so-called *Feret diameter,* x_F, is represented by the distance of two tangents to the particle contour perpendicular to the direction of measurement. The so-called *Martin diameter,* x_M, drawn parallel to the direction of measurement bisects the projected area. With an irregularly-shaped particle outline, each of these definitions yields different particle sizes if the direction of measurement changes. A distribution of *Feret* diameters is obtained, and an average value, $\bar{x}_F$, if the direction of measurement is rotated over 360°. The average *Feret* diameter, $\bar{x}_F$, is equal to the equivalent circumferential diameter, x_U, that is, the diameter of a circle of the same perimeter, U [93]:

$$\bar{x}_F = U/\pi = x_U \tag{108}$$

Fractals. Mandelbrot introduced the use of fractals into the description of particle images [94]. With a pair of compasses set at a certain stride length, λ, estimation of the perimeter of the profile is made by walking the compasses around it to construct a polygon, the perimeter of which becomes the estimate of the profile at resolution, λ [95]. If this procedure is repeated at different stride lengths, λ, Figure 27 is obtained.

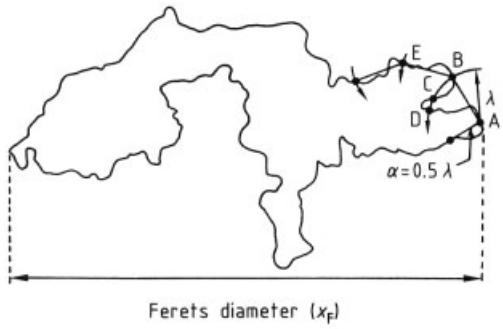

Figure 27. Demonstration of fractal construction

Equation (109) describes the relationship between perimeter, U, and stride length, λ:

$$U = k\lambda^{1-D} \tag{109}$$

where D is the characteristic fractal dimension.

Fourier Analysis. The particle outline may also be described with Fourier analysis [96].

The deviation of irregularly-shaped particles from spheres can quantitatively be described by shape factors, ψ, which can be defined as the ratio of two different equivalent spherical diameters, x_α and x_β, with α and β representing two different physical properties measured on the same particle os the same fraction of particles.

$$\psi_{\alpha,\beta} = x_\alpha / x_\beta \tag{110}$$

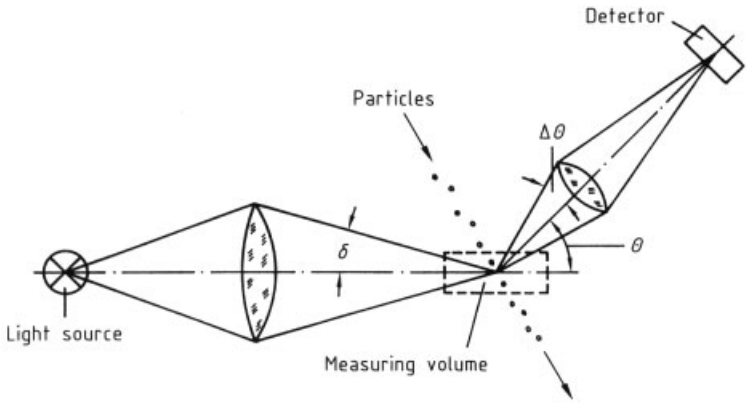

Figure 28. Optical setup for measuring scattered light

One of the best known definitions of a shape factor is the *sphericity,* ψ [4]. The sphericity ψ represents the ratio of the surface of the sphere of the same volume to the surface of the particle.

$$\psi = \pi x_V^2 / \pi x_S^2 = (x_V / x_S)^2 = \psi_{V,S}^2 \qquad (111)$$

Shape factors may be used for substituting the abscissa of a size distribution, or more generally, for replacing one equivalent diameter by another.

Automatic Image Analysis. Image analysis can also be performed automatically [97]. A video camera is then used to evaluate the particle image. The cathode ray of the camera scans the image and produces brightness-dependent voltage information. The electrical analog or digitized signal is then, for instance, transformed into a better image, rich in contrast. Further mathematical algorithms yield other image modifications, such as shading, erosion, or dilatation. The corrected data are evaluated with respect to different chord lengths, equivalent diameters, or projected area.

7.2. Electrical Sensing Zone Method

In electrical sensing, a homogeneous electrical field is set up in a cylindrical orifice. A particle passing through the orifice changes its electrical resistance and a voltage impulse is created, the height of which is directly proportional to the volume of the particle. With one particle after the other passing through the orifice, the number density distribution of particle volumes can be measured [98].

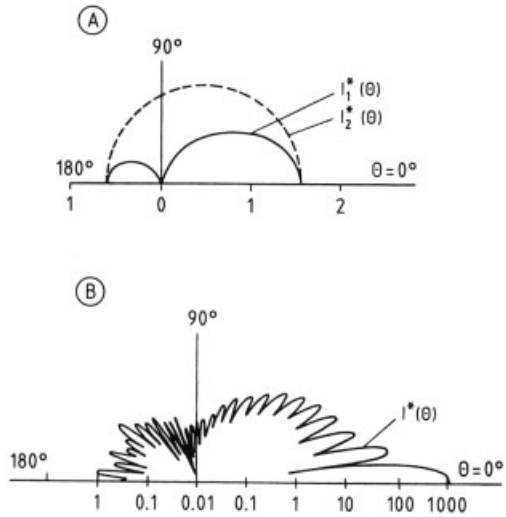

Figure 29. Polar diagrams of scattered light for two values of α
A) $\alpha = 1$, $n = 1.46$; B) $\alpha = 30$, $n = 1.33$

7.3. Measurement by Light Scattering or Absorption

Small particles scatter and absorb electromagnetic radiation. If a planar electromagnetic wave interferes with a homogeneous, isotropic sphere, it is scattered three-dimensionally. Scattering represents the deflection of a light beam by refrac-

tion, reflection, and diffraction. Absorption attenuates the light by converting it into different kinds of energy.

The schematic setup used for scattered light methods is shown in Figure 28. A small measuring volume receives the incident light beam from a light source and a suitable optical arrangement. An individual particle, passed through the measuring volume by a gas or a liquid stream, creates scattered light which is transmitted by a second optical system to a photoelectrical detector. The detector signal depends on the aperture, $\Delta\Theta$, of the second optical system and its position (angle Θ) with respect to the forward direction of light, where $\Theta = 0$.

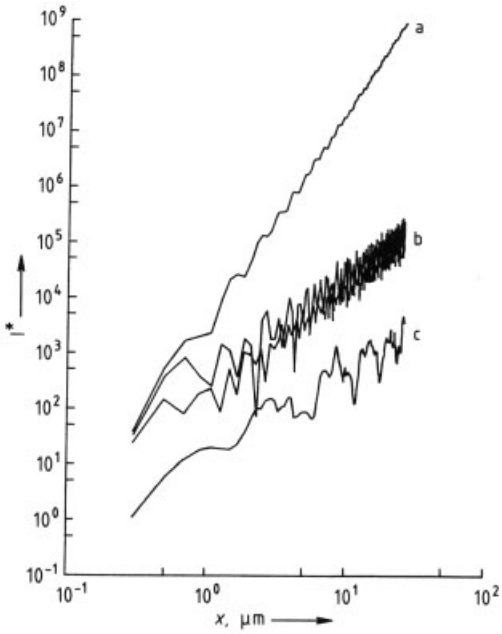

Figure 30. Scattered intensities of monochromatic light at different angles ($\lambda = 0.4\,\mu$m, $n = 1.46$)
a) $\theta = 0°$; b) $\theta = 15°$; c) $\theta = 30°$; d) $\theta = 90°$

The theory of light scattering was first described in 1908 [99]. This theory yields the distribution of scattered light in space for a sphere from the Maxwell field equations. Results are given in Figure 29. This figure shows the polar diagram of the scattered light. The two graphs indicate the differences obtained in the scattered light, with respect to the dimensionless particle size, $\alpha = \pi x/\lambda$. The value x represents the diameter of the sphere, and λ the wavelength of light. Figure 29 A shows dipole scattering in the so-called Rayleigh range. At higher α, the number of the maxima and minima of the polar diagram increases, as well as the intensity of light in the forward direction.

To use light scattering for the characterization of particles, a monotonic dependency is needed between the intensity of the scattered light, $I*$, and particle size, x. Figure 30 [100] shows the theoretical dependency for monochromatic light with $\lambda = 0.4\,\mu$m for different values of Θ. It is realized that an approximate monotonic dependency is obtained only at $\Theta = 0°$. At $x < 0.7\,\mu$m; i.e., in the Rayleigh range, the scattered light drops proportionally to x^6.

Figure 31 shows similar curves for white light [101].

Dependencies close to monotonic ones are obtained only for $\Theta = 90°$ (Figure 31 C). Contrary to monochromatic light, however, the curves strongly depend on the refractive index, i.e., the material to be measured.

8. Fraunhofer Diffraction

In 1817 Fraunhofer proposed an optical setup for the generation of diffraction patterns, which was similar to that shown in Figure 32 [102]. It consists of a monochromatic light source, optical means to expand the diameter of the beam, a measuring zone, a lens, and a suitable detector. Spherical monosized particles passing through the measuring zone, dispersed in a gas or a liquid, create a radially symmetrical diffraction pattern, which consists of an extremely bright central spot, followed by a succession of dark and bright rings. This intensity distribution was first described in 1835 with an equation similar to Equation (111) which is generally called the Fraunhofer approximation [103]. With $\Theta < 8°$, $\tan\Theta = r/f$ and $k = \pi/\lambda f$, the following is obtained:

$$I(r,x)/I_0 = \left(kx^2/2\right)^2 \left(J_1(krx)/krx\right)^2 \qquad (112)$$

where f represents the focal length, and $J_1(krx)$ is the first kind, first-order Bessel function.

If the particles passing the measuring zone form a number density distribution, $q_0(x)$, the diffraction patterns are superimposed and a radially symmetrical, diffuse diffraction pattern with the following intensity distribution is obtained:

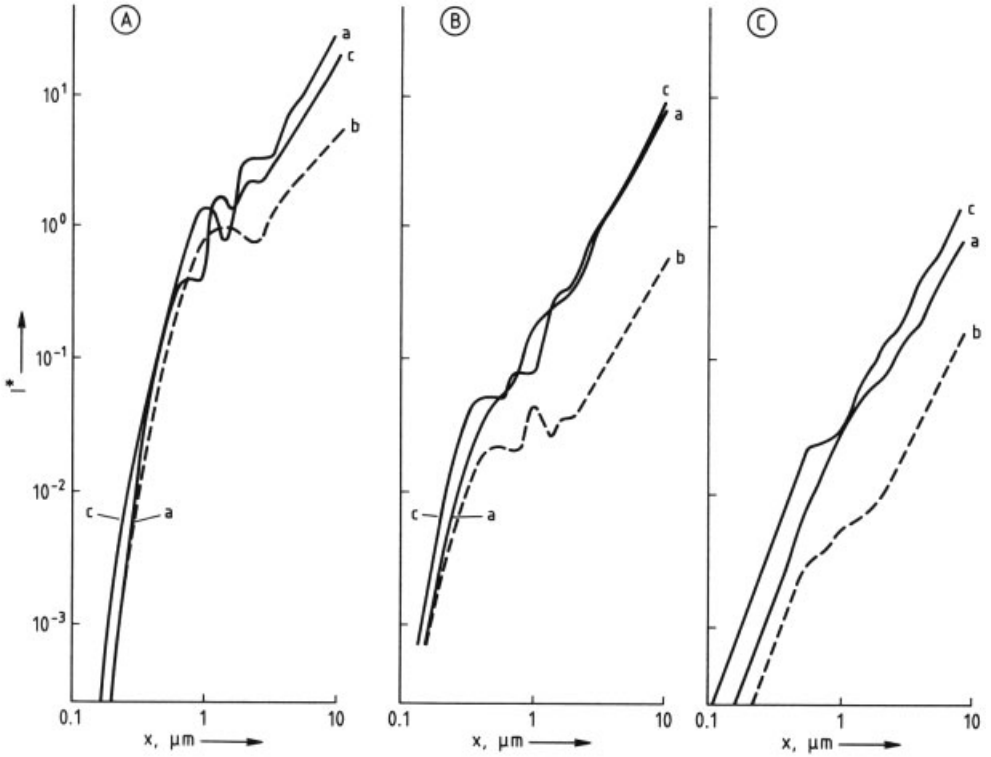

Figure 31. Scattered intensities of white light at $\theta = 17°$ (A), $\theta = 47°$ (B), and $\theta = 90°$ (C) and different refractive indices of spheres
a) $n = 1.46$; b) $n = 1.46 - 0.15$; c) $n = 1.8$

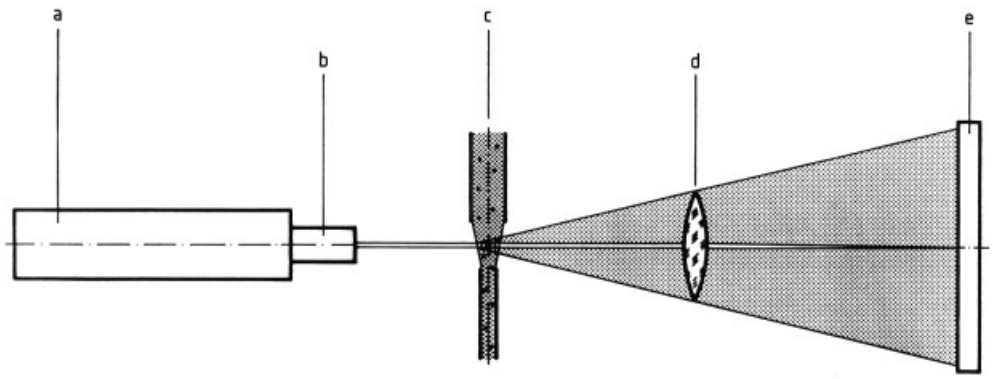

Figure 32. Optical setup for Fraunhofer diffraction pattern analysis
a) Laser; b) Beam expander; c) Aerosol beam; d) Lens; e) Detector

$$I(r) = \int_{x_{\min}}^{x_{\max}} N_0 \, q_0(x) \, I(r,x) \, \mathrm{d}x \tag{113}$$

Equation (112) represents a first kind, first-order *Fredholm integral equation.* Its solution with respect to $q_0(x)$ was first given in 1955 [104], and then in 1956 [105], using the Mellin transformation [106]. The above solution has been used to develop a new diffraction pattern analyser (HELOS, Sympatec, Federal Republic of Germany). At present, three other instruments are available

on the world market, manufactured by CILAS, France; Leeds and Northrup, United States; and Malvern, Great Britain.

9. References

1. W. H. Westphal: *Physikalisches Wörterbuch,* Springer-Verlag, Berlin-Göttingen-Heidelberg 1952, pp. 238 – 239.
2. K. Leschonski: "Representation and Evaluation of Particle Size Analysis Data," *Part. Charact.* **1** (1984) 89 – 95.
3. K. Leschonski: "Particle Characterization, Present State and Possible Future Trends," *Prepr. 1. World Congress Particle Technology,* Part I, Particle Characterization, Nürnberg 1986, pp. 1 – 16.
4. H. Wadell: "Volume, Shape and Roundness of Rock Particles," *J. Geol.* **40** (1932) 443 – 451.
5. A. Cauchy: "Nôte sur divers theorèmes relatifs à la rectification des courbes," *C. R. Hebd. Seances Acad. Sci.* **13** (1841) 1060 – 1065.
6. A. M. Gaudin: "An Investigation of Crushing Phenomena," *Trans. Am. Inst. Min. Metall. Pet. Eng.* **73** (1926) 253 – 316.
7. R. Schumann: "Principles of Communition I – Size Distribution and Surface Calculations," Amer. Inst. of Mining and Metallurgical Engineers, Techn. Publ. No. 1189 (1940), pp. 1 – 11.
8. P. Rosin, E. Rammler, U. Sperling: "Korngrößenprobleme des Kohlenstaubs," *Bericht C 52 des Reichskohlenrates,* Berlin 1933, pp. 1 – 33.
9. E. Rammler: "Gesetzmäßigkeiten in der Kornverteilung zerkleinerter Stoffe," *Z. Ver. Dtsch. Ing., Beih. Verfahrenstech.* **5** (1937) 101 – 168.
10. I. G. Bennett: "Broken Coal," *J. Inst. Fuel.* **10** (1936) 23 – 39.
11. G. Herdan: *Small Particle Statistics,* Butterworths, London 1960.
12. M. Langemann: "Zur Berechnung der spezifischen Oberflächen von Dispersionen," *Chem. Ing. Tech.* **27** (1955) 27 – 32.
13. H. Heywood: "Numerical Definitions of Particle Size and Shape," *Chem. Ind.* **15** (1937) 149 – 154.
14. K. Leschonski: "Classification of Particles in Gases," *IFPRI-Report,* 1 981 K Leschonski, Clausthal.
15. *Ullmann,* 4th ed., **2,** 35 – 42.
16. K. Leschonski, H. Herrmann: "Einfluß und Berücksichtigung von Fehlern der Partikelgrößenanalyse bei der Ermittlung von Trennkurven," *Proc. 2. Europ. Symp. "Partikelmeßtechnik,"* Nürnberg 1979, pp. 41 – 58.
17. F. K. Tromp: "Neue Wege in der Beurteilung von Steinkohlen," *Glückauf* **73** (1937) 125 – 131; 151 – 156.
18. F. W. Mayer: "Die Teilungskurve und ihre wichtigsten Anomalien," *Bergakademie* **9** (1957) 1 – 13.
19. Th. Eder: "Probleme der Trennschärfe," *Aufbereit. Techn.* **2** (1961) 104 – 109;136 – 148; 443 – 446; 484 – 495.
20. H. Trawinski: "Die mathematische Formulierung der Tromp-Kurve," *Aufbereit. Tech.* **17** (1976) 248 – 254; 449 – 459.
21. H. Heidenreich: *Die Erfolgsrechnung im Aufbereitungsbetrieb,* Verlag Glückauf, Essen 1954.
22. F. W. Mayer: "Die Trennschärfe von Sichtern," *Zem. Kalk-Gips* **19** (1966) 259 – 268.
23. F. W. Mayer: "Die Trennschärfe von im Mahlkreislauf arbeitenden Sichtern anhand von Tromp-Kurven," *Tech. Mitt.* **59** (1966) 246 – 249.
24. Th. Eder: "Zur einheitlichen Kennzeichnung der Trennschärfe," *Montanzeitung, Zeitschrift für Bergbau und Hüttenwesen* **67** (1951) 163 – 165.
25. K. Grumbrecht, F. W. Mayer: "Die Lage des Trennschnittes bei Trennvorgängen in der Aufbereitung," *Glückauf* **96** (1960) 186 – 188.
26. H. Madel: "Über den Trennungsgrad von Windsichtern," *Zement* **19** (1930) 958 – 961.
27. P. Rosin, E. Rammler: "Windsichter und ihre Untersuchung," *Glückauf* **68** (1932) 529 – 537.
28. J. White: *J. Chem. Metall. Min. Soc. S. Africa* (1925).
29. H. Rumpf: "Über die Feinheitsbestimmung von technischen Stäuben," *Staub* **25** (1965) 15 – 22.
30. S. Pethö, E. Tompos: "Über die statistische Auswertung der Tromp-Kurven," *Bergakademie* **21** (1969) 430 – 433.
31. F. W. Mayer: "Die Grundlagen der Trennschärfenermittlung von Aufbereitungsvorgängen mit Hilfe der Tromp'schen Verteilungszahlen," *Glückauf* **88** (1952) 4 – 13.
32. F. W. Mayer: "Allgemeine Grundlagen der T-Kurven," *Aufbereit. Tech.* **8** (1967) 429 – 440; 673 – 678; **9** (1968) 14 – 23.

33. F. W. Mayer: "Berechnung des neuen Trennschärfe-Kennwertes," *Aufbereit. Tech.* **12** (1971) 82 – 90; 203 – 211.
34. K. Leschonski, Th. Boeck: "Die Bedeutung des Dispergierens und Dosierens bei Feinsttrennung in Windsichtern," *Prepr. Europ. Symp. Particle Technology 1980,* **vol. B,** Amsterdam 1980, pp. 746 – 761.
35. K. Grumbrecht: "Über die Genauigkeit der Ermittlung von Mengenausbringen nach graphischen Methoden in der Steinkohleaufbereitung," *Glückauf* **89** (1953) 175 – 178.
36. P. Moiset: "Über die Bestimmung der funktionellen Eigenschaften einer Klassiermaschine," *Aufbereit. Tech.* **2** (1961) 395 – 403.
37. K. Koulen, H. Schneider: "Zur Berechnung des Gewichtsausbringes bei der Sichtung," *Aufbereit. Tech.* **6** (1965) 586 – 589.
38. K. Leschonski: "Bewertung des Trenneffektes bei der Trennung von Suspensionen," *Proc. 17. Diskussionstagung Mechanische Flüssigkeitsabtrennung,* 21. 10. 80, TH Dresden, DDR.
39. H. Heidenreich: "Zur Berechnung des Gewichtsausbringens in der Kohlenaufbereitung," Bergbau-Arch. **24** (1963) 27 – 34.
40. H. Herrmann: "Ein allgemeines Verfahren zur Verbesserung des Meßfehlers in der Partikelgrößenanalyse," *Chem. Ing. Tech.* **51** (1979) 1140 – 1141.
41. H. Herrmann: "Zur Ermittlung der Trennkurve– Fehleranalyse und Fehlerverbesserung," *Chem. Ing. Tech.* **52** (1980) 466 – 467.
42. H. J. Smigerski: "Die Bilanzierung von Haufwerkstrennungen mit Hilfe mikroskopischer Korngrößenmessungen," *VI Verfahrenstech.* **11** (1977) 546 – 548.
43. *Ullmann,* 4th ed., **5,** 725 – 753.
44. DIN 66 111,"Grundlagen der Sedimentationsanalyse".
45. B. H. Kaye, R. P. Boardman: "Cluster Formation in Dilute Suspensions," *Proc. 3rd. Congr. Europ. Tech. Chem. Eng. London* 1962, pp. A 17 – 21.
46. R. Johne: "Einfluß der Konzentration einer monodispersen Suspension auf die Sinkgeschwindigkeit ihrer Teilchen," *Fortschr. Ber. VDI Z., Reihe 9,* **11** (1966).
47. B. Koglin: "Untersuchung zur Sedimentationsgeschwindigkeit von Einzelteilchen," *Freiberg. Forschungsh. A,* **A 484** (1971) 35 – 44.
48. E. Cunningham: "On the Velocity of Steady Fall of Spherical Particles through Fluid Medium," *Proc. R. Soc. London Ser. A,* **83** (1909/10) 357 – 365.
49. D. S. Jennings, M. D. Thomas, W. Gardner: "A New Method of Mechanical Analysis of Soils," *Soil Sci.* **14** (1922) 485 – 499.
50. G. W. Robinsion: "A New Method for the Mechanical Analysis of Soils and other Dispersion," *J. Agric. Sci.* **12** (1922) 306 – 321.
51. D. H. M. Andreasen: "Zur Kenntnis des Mahlgutes. Kap. D: Untersuchung über zulässige Meß-unsicherheit nebst Berichten über die rechnerische Behandlung der Meßresultate," *Kolloidchem. Beih.* **27** (1928) 384 – 409.
52. K. Leschonski: "Vergleichende Untersuchungen der Sedimentationsanalyse," *Staub* **22** (1962) 475 – 486.
53. S. Berg: "Studies on Particle-Size Distribution," *Ingenioervidensk. Skr.* **2** (1940) 5 – 243. Akad. Tekn. Videnskaber u. Dansk Ing. Sorening, Kopenhagen.
54. H. E. Rose, H. B. Lloyd: "On the Measurement of the Size Characteristics of Powders Photoextinction Methods, Part 1," *J. Soc. Chem. Ind. London* **65** (1946) 52 – 58.
55. S. Odin: "Eine neue Methode zur Bestimmung der Körnerverteilung in Suspensionen," *Kolloid Z.* **28** (1916) 33 – 47.
56. C. E. Marshall: "A New Method for Determining the Distribution Curve of Polydisperse Colloidal Systems," *Proc. Roy. Soc. (London) Ser A* **126** (1930) 427 – 439.
57. H. Rumpf, W. Alex, R. Johne, K. Leschonski: "Korngrößenanalyse feiner Teilchen, eine kritische Betrachtung der Methoden," *Ber. Bunsenges. Phys. Chemie* **71** (1967) 253 – 270.
58. *An Introduction to Density Gradient Centrifugation,* Beckmann Instruments Inc., Spinco Div., Palo Alto, Cal. USA 1960.
59. F. S. Eadin, R. E. Payne: "Particle Size Distribution Analysed Quickly, Accurately," *Iron Age* **174** (1954) 99 – 102.
60. J. M. Zeigler, G. G. I. R. Whitney, C. R. Hayes: "Woods Hole Rapid Sediment Analyser," *J. Sediment. Petrol.* **30** (1960) 490 – 495.
61. B. Bienek, H. Huffmann, H. Meder: "Korngrößenanalyse mit Hilfe von Sedimentationswaagen," *Erdöl Kohle Erdgas Petrochem.* **18** (1965) 509 – 513.

62. J. Brezina: "Sedimentation of Stratified Suspensions above the Stokes Range and its use for Grain Size Analysis," *2nd Int. Symp. Part. Size Anal.,* Bradford 9./11. 09. 1970.
63. K. T. Whitby: *The mechanics of fine sieving,* Ph. D. Thesis. Univ. of Minnesota 1945.
64. F. A. Shergold: "Effect of Sieve Loading on the Results of Sieve Analysis of Natural Sands," *Trans. Soc. Chem. Ind.* **65** (1946) 245 – 249.
65. F. G. Carpenter, V. R. Deitz: "Methods of Sieve Analysis with Particular Reference to Bone Char," *J. Res. Natl. Bur. Stand. U.S.* **45** (1950) 328 – 346.
66. W. Batel: "Kritische Betrachtungen zur Teilchengrößenbestimmung durch Siebanalyse, Windsichten, Sedimentieren und den Blaine-Test," Chem. Ing. Tech. **29** (1957) 581 – 589.
67. B. H. Kaye: *Physical Problems of Particle Size Analysis,* Ph. D. Thesis, London Univ. 1962.
68. H. E. Rose: "The Derivation of an Equation for the Performance of a Screen, " *DECHEMA Monogr.* **79** (1976) 351 – 369.
69. A. H. M. Andreasen: "Eine selbstregistrierende Sedimentationswaage, die die Sedimentationskurve direkt aufzeichnet," *Sprechsaal* **60** (1927) 869 – 871.
70. J. Andersen: "Über Maschenweiten und Korngrößen," *Zement* **20** (1931) 224 – 226, 242 – 245.
71. K. Leschonski: "Sieve Analysis, the Cinderella of Particle Size Analysis Methods," *Powder Technol.* **24** (1979) 115 – 124.
72. Subcommittee 4.13.1 of Bureau Commutative de Référence, Rue de la Loi 200, B-1049 Bruxelles, Belgium.
73. H. Polley, K. D. Friedberg: " Zur gravimetrischen Konzentrationsbestimmung von Fraktionen des Schwebestaubes mit einem Spaltkanalgerät," *Silikosebericht NRW, Arbeitsgemeinschaft Staub- und Silikosebekämpfung,* **vol. 7,** Verlag H. Bösmann GmbH, Detmold 1969, pp. 65 – 71.
74. H. Boose: "Gerät zur fraktionierten Abscheidung von Staub für Korngrößenanalysen," *Staub Reinhalt. Luft* **22** (1962) 109 – 112.
75. V. Timbrell: "The Terminal Velocity and Size of Airborne Dust Particles," *Br. J. Appl. Phys.* **5** (1954) Suppl. 3, 86 – 90.
76. K. F. Sawyer, W. H. Walton: "The Conifuge – A Size Separating Sampling Device for Airborne Particles," *J. Sci. Instrum.* **27** (1950) 272 – 276.
77. C. H. Keith, J. C. Derrick: "Measurement of the Particle Size Distribution and Concentration of Cigarette Smoke by the "Conifuge" ", *J. Colloid Sci.* **15** (1960) 340 – 356.
78. A. Goetz, J. R. Stevenson, O. Preining: "The Design and Performance of the Aerosol Spectrometer," *J. Air Pollut. Control. Assoc.* **10** (1960) 378 – 383, 414, 416.
79. W. Stöber, U. Zessack: "Zur Theorie einer konischen Aerosolzentrifuge," *Staub Reinhalt. Luft* **24** (1964) 295 – 305.
80. W. Kast: "Staubmeßgerät zur Schnellbestimmung der Staubkonzentration," *Staub Reinhalt. Luft* **21** (1961) 215 – 223.
81. W. Stöber, H. Flachsbart: "High Resolution Aerodynamic Size Spectrometry of Quasi-Monodisperse Latex Spheres with a Spiral Centrifuge,"*J. Aerosol Sci.* **2** (1971) 103 – 116.
82. W. Stöber: "Spiralenzentrifugen als Aerosolmeßgeräte," *Proc. 2. Europ. Symp. "Partikelmeßtechnik,"* Nürnberg, 24./26. 09. 1979, 91 – 148. Engl. Fassung: *Proc. Symp. "Fine Particles,"* US Env. Prot. Agency, Minneapolis, 28./30. May, 1975, Academic Press, New York 1976, pp. 351 – 397.
83. V. A. Marple: *A Fundamental Study of Inertial Impactors,* Ph. D. Thesis, Univ. of Minnesota, Particle Technology Laboratory, Publ. No. 144 (1970).
84. V. A. Marple, B. Y. H. Liu: "Characteristics of Laminar Jet Impactors," *Environ. Sci. Techn.* **8** (1974) 648 – 654.
85. K. R. May: "The Cascade Impactor; An Instrument for Sampling Coarse Aerosols," *J. Sci. Instrum.* **22** (1945) 187 – 195.
86. H. W. Gonell: "Ein Windsichtverfahren zur Bestimmung der Kornzusammensetzung staubförmiger Stoffe," *VDI Z.* **72** (1928) 945 – 950.
87. P. S. Roller: "Measurement of Particle Size with an Accurate Air Analyzer. The Fineness and Particle Size Distribution of Portland Zement," *Proc. Am. Soc. Test. Mater.* 1932, no. 59, 607 – 628.
88. M. Weilbacher: *Untersuchungen zur Schwerkraft- und Fliehkraftwindsichtung für Teilchengrößenanalysen,* Diss. Universität Karlsruhe 1968.
89. K. Leschonski, H. Rumpf: "Principle and Construction of Two New Air Classifiers for Size Analysis," *Powder Technol.* **2** (1968/69) 175 – 185.

90. H. Rumpf: *Über die bei der Bewegung von Pulvern in spiraligen Luftströmungen auftretende Sichtwirkung,* Diss. TH Karlsruhe 1939.
91. K. A. Gustavson: "Stoftcentrifug för Kornstorleksanalyse," *Tek. Tidskr.* **78** (1948) 667 – 670.
92. W. Alex: "Prinzipien and Systematik der Zählverfahren in der Teilchengrößenanalyse," *Aufbereit. Tech.* **13** (1972) 105 – 111, 168 – 182, 639 – 652, 723 – 732.
93. C. N. Davies: "Measurement of Particles," *Nature* (London) 195 (1962) 768 – 770.
94. B. B. Mandelbrot: *"Fractals, Form, Chance, and Dimension,"* W. H. Freeman and Comp., San Francisco 1977.
95. B. H. Kaye: "Multifractal Description of a Rugged Fineparticle Profile," *Part. Charact.* **1** (1984) 14 –21.
96. R. Weichert, D. Huller: "Volumenbestimmung und Formerkennung unregelmäßig geformter Partikeln mittels dreidimensionaler Bildanalyse," *Proc.2. Europ. Symp. "Partikelmeßtechnik,* NMA, Nürnberg 24./26. 09. 79, pp. 600 – 615.
97. K. Hermes, U. Kesten: "Die Bildanalyse als modernes Verfahren der Partikelmeßtechnik," *Chem. Ing. Tech.* **53** (1981) 780 – 786.
98. Coulter-Counter: Equipment Description.
99. G. Mie: "Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen," *Ann. Phys.* **4. Folge 25** (1908) 377 – 445.
100. Th. Boeck: *Entwicklung eines photometrischen online Oberflächenmeßverfahrens für trockene, disperse Feststoffe,* Diss. TU Clausthal 1983.
101. R. Broßmann: *Die Lichtstreuung an kleinen Teilchen als Grundlage einer Teilchengrößenbestimmung,* Diss. TH Karlsruhe 1966.
102. J. Fraunhofer: "Bestimmung des Brechungs- und Farbzerstreuungsvermögens verschiedener Glasarten," *Gilberts Ann. Phys.* **56** (1817) 193 – 226.
103. G. B. Airy: "On the Diffraction of an Object Glass with Circular Aperture," *Trans. Cambridge Philos. Soc.* **5** (1835) 283 – 290.
104. J. H. Chin, C. M. Sliepcevich, M. Tribus: "Determination of Particle Size Distributions in Polydispersed Systems," *J. Phys. Chem.* **59** (1955) 845 – 848.
105. K. S. Shifrin: *Tr. VZLTI (Leningrad)* **2** (1956) 153 – 162.
106. E. C. Titchmarsh: *Introduction to the Theory of Fourier Integrals,* Claredon Press. Oxford 1924.
107. M. Heuer, K. Leschonski: "Erfahrungen mit einem neuen Gerät zur Messung von Partikelgrößenverteilungen aus Beugungsspektren," *3. Europ. Symp. "Partikelmeßtechnik,"* NMA Nürnberg, 9./11. 05. 84, pp. 515 – 537.

Parting Agents → Release Agents

Crystallization and Precipitation

See also Solids Technology, Introduction

JOHN W. MULLIN, University College London, Torrington Place, London WC1E 7JE, United Kingdom

1. Introduction . . . 3
2. System Properties . . . 4
2.1. Solutions and Solubility . . . 4
2.2. Saturation and Supersaturation . . . 4
2.3. Crystal Size and Solubility . . . 6
2.4. Effect of Impurities . . . 6
3. Phase Equilibria . . . 6
3.1. One-Component Systems . . . 6
3.2. Two-Component Systems . . . 7
3.2.1. Eutectics . . . 7
3.2.2. Solid Solutions . . . 8
3.2.3. Compound Formation . . . 9
3.3. Three-Component Systems . . . 9
3.4. Multicomponent Systems . . . 11
3.5. Phase Transformations . . . 11
4. Kinetics and Mechanisms of Crystallization . . . 13
4.1. Crystal Nucleation . . . 13
4.1.1. Primary Nucleation . . . 13
4.1.2. Secondary Nucleation . . . 15
4.1.3. Nucleation Measurements . . . 15
4.1.4. Induction Periods . . . 17
4.2. Crystal Growth . . . 17
4.2.1. Measurement of Growth Rate . . . 18
4.2.2. Expression of Growth Rate . . . 19
4.2.3. Dependence of Growth Rate on Crystal Size . . . 20
4.3. Growth – Nucleation Interactions . . . 20
4.4. Crystal Habit Modification . . . 21
4.5. Inclusions in Crystals . . . 22
4.6. Caking of Crystals . . . 22
5. Crystallization from Solutions . . . 23
5.1. Cooling Crystallizers . . . 23
5.1.1. Nonagitated Vessels . . . 23
5.1.2. Agitated Vessels . . . 23
5.1.3. Direct-Contact Cooling . . . 24
5.2. Evaporating Crystallizers . . . 24
5.3. Vacuum (Adiabatic Cooling) Crystallizers . . . 25
5.4. Continuous Crystallizers . . . 25
5.5. Crystal Yield . . . 27
5.6. Controlled Crystallization . . . 27
5.7. Comparison of Batch and Continuous Crystallization . . . 29
5.8. Crystallizer Modeling and Design . . . 29
5.8.1. Population Balance . . . 29
5.8.2. Design and Scaleup Problems . . . 31
6. Crystallization from Melts . . . 31
6.1. Single Stage Processes . . . 32
6.2. Multistage Processes . . . 32
6.3. Column Crystallizers . . . 34
6.4. High-Pressure Crystallization . . . 35
6.5. Prilling and Granulation . . . 36
7. Crystallization from Vapors . . . 36
8. Precipitation . . . 37
8.1. Solubility Products . . . 37
8.2. Ostwald's Rule of Stages . . . 38
8.3. Development of Precipitates . . . 38
8.3.1. Ripening . . . 38
8.3.2. Agglomeration . . . 39
8.3.3. Precipitate Morphology . . . 39
8.3.4. Coprecipitation . . . 40
8.4. Precipitation Techniques . . . 40
8.4.1. Reaction Precipitation . . . 40
8.4.2. Salting Out . . . 42
8.5. Precipitation Methods and Equipment . . . 43
9. Fractional Crystallization . . . 43
9.1. Recrystallization from Solutions . . . 44
9.2. Recrystallization from Melts . . . 44
9.3. Recrystallization Schemes . . . 44
9.4. Recrystallization from Supercritical Fluids . . . 46
9.5. Separation of Enantiomers and Racemates . . . 46
10. Miscellaneous Crystallization Techniques . . . 47
10.1. Salting-Out Crystallization . . . 47
10.2. Reaction Crystallization . . . 47
10.3. Adductive and Extractive Crystallization . . . 47
10.4. Spray Crystallization . . . 47
10.5. Spherical Crystallization . . . 48
10.6. Freeze Crystallization . . . 48
11. References . . . 49

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_03

Symbols

a	activity
$a*$	activity of a saturated solution
$a_{\pm}$	mean ionic activity
A	particle surface area, preexponential factor, m^2
b	kinetic order of secondary nucleation
B	secondary nucleation rate, $m^{-3}\,s^{-1}$
c	solution concentration
$c*$	equilibrium saturation concentration
C	specific heat capacity
D	diffusivity, m^2/s
D	homogeneous distribution coefficient
g	kinetic order of growth
G	crystal growth rate, m/s
i	relative kinetic order ($=b/g$)
J	nucleation rate, $m^{-3}\,s^{-1}$
k	Boltzmann constant
K	rate constant, solubility product
L	crystal size, m
M	relative molecular mass
M_T	total magma density, kg/m^3
n	kinetic order of primary nucleation
n^0	population density of nuclei ($=B/G$), m^{-4}
N	impeller rotational speed, Hz
q	heat of crystallization, J/kg
r_c	size of critical nucleus, m
R	ratio of molecular masses, mass deposition rate
S	supersaturation ratio ($=c/c*$)
S'	activity supersaturation
v	molar volume, m^3/mol
$\bar{v}$	mean linear growth velocity, m/s
V	water lost by evaporation, kg/kg
W	initial mass of water, kg
x	mole fraction
Y	crystal yield, kg
z	valence
α	surface roughness factor, volume shape factor
β	surface shape factor
γ	interfacial tension, J/m^2
$\gamma_{\pm}$	mean ionic activity coefficient
λ	latent heat, heterogeneous distribution coefficient
ν	number of moles of ions
σ	relative supersaturation, $=(S-1)$
τ	induction period, residence time, batch time, s

1. Introduction

Few branches of the chemical and process industries do not, at some stage, employ crystallization or precipitation for production or separation purposes. Vast quantities of crystalline substances are manufactured commercially: sodium chloride and sucrose, for example, have worldwide production rates exceeding 10^8 t/a, while annual rates for fertilizer chemicals such as ammonium nitrate, potassium chloride, ammonium phosphates, and urea exceed 10^6 t. Although crystalline products of the pharmaceutical, organic fine chemical, and dye industries are produced in relatively low tonnages, they still represent a valuable and important industrial sector. Crystallization is also a key operation in, for example, the desalination of seawater, the concentration of fruit juices, and the removal of unwanted materials or recovery of valuable constituents in many industrial processes (e.g., sodium sulfate removal from viscose spin-bath liquors, recovery of metal salts from electroplating baths). Crystallization is increasingly employed in the production of materials for the electronics industry; applications of precipitation have been extended to biotechnology, e.g., for processing proteins.

The unit operation of crystallization is governed by very complex, interacting variables. It is a simultaneous heat- and mass-transfer process with a strong dependence on fluid and particle mechanics. Crystallization occurs in a multiphase, multicomponent system and involves particulate solids whose size and size distribution are incapable of definition and vary with time. The solids are suspended in a solution which can fluctuate between a so-called metastable equilibrium and a labile state that is unstable and prone to change; the solution composition can also vary with time. Nucleation and growth kinetics, the key processes in this operation, are often influenced profoundly by traces of impurities: a few parts per million may alter the product beyond recognition.

Five main types of information are generally required to design a crystallization process:

1) Solubility and phase relationships
2) Metastability limits
3) Nucleation characteristics
4) Crystal growth characteristics

5) Hydrodynamics of crystal suspensions

Solubility and phase relationships influence the choice of crystallizer and method of operation. These data must be obtained by using the materials to be encountered in the plant because traces of impurity often have a considerable effect on phase relationships. Metastable limits define acceptable operating conditions for the minimization of uncontrolled nucleation and encrustation of heat-exchange surfaces. The processes of nucleation and growth are both exceedingly complex; they are influenced greatly by temperature, supersaturation, and impurities. A knowledge of these system characteristics is essential in design. Crystal suspension velocities must also be known so that liquor circulation rates in fluidized-bed crystallizers and agitation rates in stirred vessels can be specified. Because the crystals are present in large quantities, settling is hindered; further complications can arise if their shapes are irregular.

Many different methods are available for crystallization. Crystals can be grown from the liquid (solution or melt) or vapor phase (desublimation, see → Sublimation), but in all cases the state of supersaturation must first be achieved. The method used to obtain supersaturation depends on the characteristics of the crystallizing system: some solutes are readily deposited from solution when cooled, whereas others may crystallize only after some solvent has been removed. The addition of another substance to the system to alter the equilibrium conditions is used frequently in precipitation processes. Supersaturation is sometimes achieved as a result of a chemical reaction between two or more substances and one of the reaction products is precipitated.

Comprehensive accounts of industrial crystallization theory and practice have been presented in several books [1–8].

2. System Properties

2.1. Solutions and Solubility

Water is almost exclusively used as the solvent for the industrial crystallization of inorganic substances from solution. It is also used whenever possible for the crystallization of organic compounds, but in this particular field for a variety of reasons some other solvent must often be employed. A guide to the selection of the "best" solvent for a given crystallization operation is given in [7] and an algorithm for the prediction of an optimal solvent or solvent mixture for cooling crystallizers is proposed in [9]. Compilations of solid – liquid solubility data are available in [5], [7] , [10–12].

2.2. Saturation and Supersaturation

A saturated solution is a solution that is in thermodynamic equilibrium with the solid phase of its solute at a specified temperature. However, solutions frequently contain more dissolved solute than that given by the equilibrium saturation value and are then said to be supersaturated. The degree of supersaturation can be expressed by the concentration difference Δc:

$$\Delta c = c - c^* \tag{1}$$

where c is the actual solution concentration and $c*$ the equilibrium saturation value. Other common expressions are the supersaturation ratio S and the relative supersaturation σ which are both dimensionless:

$$S = c/c^* \tag{2}$$

$$\sigma = \Delta c/c^* = S - 1 \tag{3}$$

Solution concentration may be expressed in a variety of units, but for general mass balance calculations, units such as kilograms of anhydrate per kilogram of solvent or kilograms of hydrate per kilogram of free solvent are most convenient. The former avoids complications if different phases (e.g., anhydrates and hydrates) can crystallize over the temperature range considered. The latter simplifies yield calculations when a single hydrate phase crystallizes.

Although the terms S and σ are dimensionless, their magnitudes depend on the units used to express solution concentration. For example, a supersaturated solution of sucrose at 20 °C contains 2.45 kg of sucrose per kilogram of water, and the corresponding value for $c*$ is 2.04; therefore, the value of S ($= c/c*$) is 1.20. However, if the composition is expressed as kilograms of sucrose per kilogram of solution ($c = 0.710$, $c* = 0.671$), the value of S becomes 1.06 [7].

None of the above expressions represents exactly the true thermodynamic supersaturation.

The fundamental driving force for crystallization is the difference between the chemical potential of a given substance in the transferring and the transferred state, i.e., in solution (state 1) and in the crystal (state 2). For an unsolvated solute crystallizing from a binary solution, this may be written as

$$\Delta\mu = \mu_1 - \mu_2 \tag{4}$$

The chemical potential μ is defined in terms of the standard potential μ_0 and the activity a by

$$\mu = \mu_0 + RT \ln a \tag{5}$$

The fundamental dimensionless driving force for crystallization may, therefore, be expressed as

$$\Delta\mu/RT = \ln\,(a/a^*) = \ln S' \tag{6}$$

where $a*$ is the activity of a saturated solution, S' the activity supersaturation, R the gas constant, and T the absolute temperature; i.e.,

$$S' = \exp\,(\Delta\mu/RT) \tag{7}$$

For electrolyte solutions, use of the mean ionic activity $a_\pm$ is more appropriate; this is defined by

$$a = a_\pm^\nu \tag{8}$$

where ν is the number of moles of positive and negative ions in one mole of solute ($\nu = \nu_+ + \nu_-$). Therefore

$$\Delta\mu/RT = \nu \ln S_a \tag{9}$$

where

$$S_a = a_\pm / a_\pm^* \tag{10}$$

Alternatively, supersaturation can be expressed as

$$\sigma_a = S_a - 1 \tag{11}$$

and Equation (9) becomes

$$\Delta\mu/RT = \nu \ln\,(1 + \sigma_a) \tag{12}$$

For low supersaturations ($\sigma_a < 0.1$), the following approximation is valid:

$$\Delta\mu/RT \simeq \nu\sigma_a \tag{13}$$

In practice, however, supersaturation is generally expressed directly in terms of solution concentration (Eqs. 1 – 3). The relationship between concentration-based supersaturation and fundamental (activity-based) supersaturation can be expressed by means of the relevant concentration-dependent activity coefficient ratios; for details see [13].

Metastability. The state of supersaturation is an essential feature of all crystallization operations. Ostwald [14] introduced the terms *labile* and *metastable* supersaturation to describe supersaturated solutions in which spontaneous (primary) nucleation (see Section 4.1) would or would not occur, respectively. Miers and Isaac [15] proposed a representation of the metastable zone by means of a solubility – supersolubility diagram (Fig. 1). The lower, continuous, equilibrium solubility curve can be determined accurately, but the position of the upper, broken, supersolubility curve is less certain because it is influenced considerably by factors such as the rate at which supersaturation is generated, the intensity of agitation, and the presence of crystals or impurities. The width of the metastable zone is usually expressed as a temperature difference $\Delta\theta$, which is related to the corresponding concentration difference Δc by the local slope of the solubility curve, $dc*/d\theta$:

$$\Delta c = (dc^*/d\theta)\,\Delta\theta \tag{14}$$

The measurement of metastable zone width is discussed in Section 4.1.3.

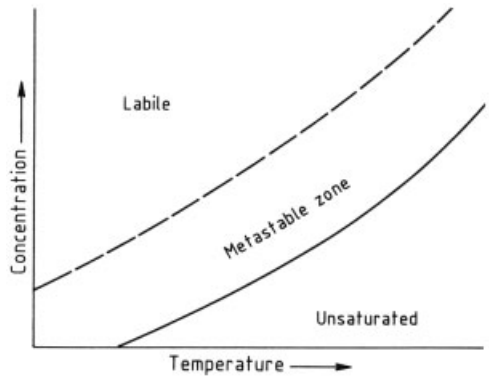

Figure 1. Solubility – supersolubility diagram

Supersaturation Measurement. The supersaturation of a solution can be calculated simply from Equations (1) – (3) if the actual solution concentration and the corresponding equilibrium saturation concentration at a given temperature are known.

Many ways exist for measuring supersaturation (i.e., concentration), but all of these are not readily applicable to industrial crystallization. If chemical analysis is difficult, measurement of a concentration-dependent property of the system (e.g., density or refractive index) may be possible. Both of these properties can usually be measured with high precision on a sample transferred under controlled laboratory conditions.

On the other hand, an in situ, preferably continuous method of concentration measurement is usually required for a crystallizer operating under laboratory or pilot-plant conditions. Although the above properties are temperature-dependent, they can often be measured with sufficient accuracy for supersaturation determination. In industrial crystallization, temperature and feedstock concentration can fluctuate and thus make the assessment of supersaturation difficult. Under these conditions, a crude method based on a mass balance coupled with feedstock and exit-liquor concentrations and crystal production rates averaged over several hours may be adequate.

2.3. Crystal Size and Solubility

If the solute particles dispersed in a solution are small enough, the solute concentration may greatly exceed the normal equilibrium saturation value. The relationship between particle size and solubility, first derived for liquid – vapor systems by Thomson (1870), utilized by Gibbs (1890), and applied to solid – liquid systems by Ostwald (1900) and Freundlich (1909), may be expressed in the form

$$\ln\left[\frac{c(r)}{c^*}\right] = \frac{2M\gamma}{\nu RT\varrho r} \tag{15}$$

where $c(r)$ is the solubility of particles with radius r, $c*$ the normal equilibrium solubility of the substance, R the gas constant, T the absolute temperature, ϱ the density of the solid, M the relative molecular mass of the solute in solution, and γ the interfacial tension of the crystallization surface in contact with its solution. The quantity ν represents the number of moles of ions formed from one mole of electrolyte. For a nonelectrolyte, $\nu = 1$. For most inorganic salts in water, the solubility increase becomes significant only for particle sizes smaller than ca. 1 µm. For example, for barium sulfate at 25 °C: $T = 298$ K, $M = 0.233$ kg/mol, $\nu = 2$, $\varrho = 4500$ kg/m^3, $\gamma = 0.13$ J/m^2, $R = 8.3$ J mol^{-1} K^{-1}. Thus for a 1-µm crystal ($r = 5 \times 10^{-7}$ m), $c/c* = 1.005$ (i.e., 0.5 % increase); for 0.1 µm, $c/c* = 1.06$ (i.e., 6 % increase); and for 0.01 µm, $c/c* = 1.72$ (i.e., 72 % increase). All such calculated values should be treated with caution, however, not only because of the potential unreliability of γ values, but also because the Gibbs – Thomson effect may cease to be influential at extremely small crystal sizes [7].

2.4. Effect of Impurities

Industrial solutions invariably contain dissolved impurities which can increase or decrease the solubility of the prime solute considerably [7]. Therefore, the solubility data used to design crystallization processes must relate to the actual system used. Impurities can also have profound effects on other crystallization characteristics such as nucleation and growth (cf. Chap. 4).

3. Phase Equilibria

3.1. One-Component Systems

Temperature and pressure are the two variables that can affect the phase equilibria in a one-component system. The phase diagram in Figure 2 depicts equilibria between the solid, liquid, and vapor states of water: ice, water, and steam, respectively. At the *triple point* B (0.6 kPa, 0.01 °C), all three phases are in equilibrium.

The *sublimation curve* AB records the vapor pressure of ice, the *vaporization curve* BC records the vapor pressure of liquid water, and the *fusion curve* BD records the effect of pressure on the melting point of ice. The fusion curve for ice is very unusual because an increase of pressure increases the melting point of most other one-component systems.

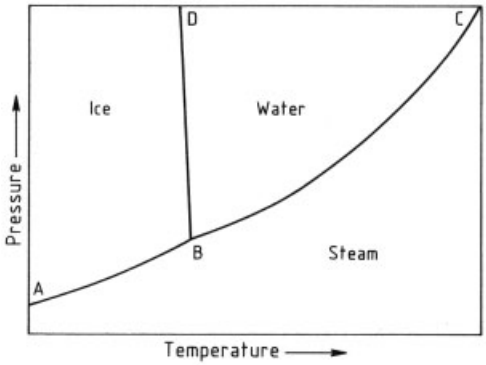

Figure 2. Phase diagram for water

Polymorphs. A single substance can crystallize in more than one of seven crystal systems, 32 classes, or 230 point groups. Because these crystalline forms differ in their lattice arrangement, they can exhibit not only different basic shapes but also different physical properties. A substance capable of crystallizing in more than one different crystalline form is said to exhibit *polymorphism*. Carbon, for example, has three polymorphs: graphite (hexagonal) and diamond (regular), and fullerenes. Calcium carbonate has three: calcite (rhombohedral), aragonite (orthorhombic), and vaterite (trigonal).

Each polymorph is composed of a common component but constitutes a separate phase. Since only one polymorph is thermodynamically stable at a specified temperature and pressure, all the other polymorphs are potentially capable of being transformed into the stable polymorph. Some polymorphic transformations are rapid and reversible, others are not.

Polymorphs may be *enantiotropic* (interconvertible) or *monotropic* (incapable of transformation). Graphite and carbon, for example, are monotropic at ambient temperature and pressure, whereas ammonium nitrate has five enantiotropic polymorphs over the temperature range − 18 to 125 °C:

$$\text{Regular} \underset{125\,^\circ\text{C}}{\rightleftharpoons} \text{Trigonal} \underset{84\,^\circ\text{C}}{\rightleftharpoons} \text{Orthorhombic(I)} \underset{32\,^\circ\text{C}}{\rightleftharpoons} \text{Orthorhombic(II)} \underset{-18\,^\circ\text{C}}{\rightleftharpoons} \text{Tetragonal}$$

At a given temperature, the more stable phase will always have the lower solubility in any given solvent. Similarly, at a given pressure, the more stable phase will always have the higher melting point. Polymorphic transformations are further discussed in Section 3.5.

3.2. Two-Component Systems

Three variables (temperature, pressure, and concentration) can affect the phase equilibria in a two-component (binary) system. The effect of pressure, which is usually negligible, is generally ignored so that the relevant data can be shown on a two-dimensional temperature – concentration plot.

Three important basic types of binary system – eutectics, solid solutions, and systems with compound formation – are discussed. They are also referred to in the discussion of melt crystallization (Chap. 6). Although the terminology used is specific to melt systems, the types of behavior described may also be exhibited by aqueous solutions of salts, for example. In fact, no fundamental difference exists between a melt and a solution [7].

3.2.1. Eutectics

An example of a binary eutectic system AB is shown schematically in Figure 3. A eutectic (Greek: *eu tektos* = easily melted) is the mixture of components that has the lowest crystallization temperature in the system. The concentration of component B in system AB is plotted as the abscissa and temperature as the ordinate. Point A is the crystallization temperature (freezing point) of pure component A, and point B that of component B. Curves AE and EB represent the crystallization temperatures of all mixtures of the two components A and B. Above the two curves, all mixtures of A and B exist only in the liquid state, i.e., as a melt. If a melt represented by point X is cooled along the vertical line XZ, crystals will start to be deposited at point Y (with proper initiation); theoretically, these crystals should be pure component B. On further cooling, more crystals of pure component B will be deposited until, at the eutectic point E, the system solidifies completely. The horizontal line passing through point E is called the *solidus*. The liquid-phase composition during cooling changes continuously along curve

BE, which is called the *liquidus.* For example, if mixture X is cooled to point Z, the crystals C will be pure component B and the liquid L will be a mixture of A and B. The mass proportion of solid phase (crystal) to liquid phase (residual melt) at temperature Z is quantified by the ratio of the lengths LZ to CZ. This relationship is known by various names such as the *mixture rule* or the *lever rule.* Similar reasoning can be applied to mixtures located in the region of curve AE, in which case the crystals would be pure component A.

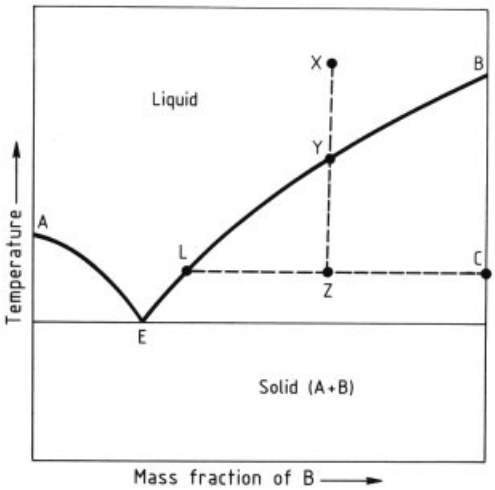

Figure 3. Phase diagram for a binary eutectic system AB
For explanation of symbols, see text.

The eutectic point E is common to both curves. A liquid of this composition cooled to the eutectic temperature crystallizes with unchanged composition, and continues to deposit such crystals until the whole system solidifies. Although a eutectic in a given system has a fixed composition, it is not a chemical compound but simply a physical mixture of the individual components. The component crystals are often clearly visible under a low-power microscope.

3.2.2. Solid Solutions

The second common type of binary system is the continuous series of solid solutions. The term *solid solution* or *mixed crystal* refers to an intimate mixture, on the molecular scale, of two or more components. The components of a solid-solution system cannot be separated as easily as those of a eutectic system, as shown in Figure 4. Points A and B represent the crystallization temperatures (freezing points) of pure components A and B, respectively. The upper curve (the liquidus) represents the temperature at which mixtures of A and B begin to crystallize on cooling. The lower curve (the solidus) represents temperatures at which mixtures begin to melt on heating. A melt of composition X begins to crystallize on reaching temperature Y. At temperature Z, the system consists of a mixture of crystals with composition C and a liquid with composition L. The ratio of crystals to liquid is again given by the mixture rule (see Section 3.2.1). However, the crystals do not consist of a single pure component (as in a simple eutectic system) but are an intimate mixture (a solid solution) of components A and B. To purify the crystals further, they must first be melted; the resulting melt must then be cooled and recrystallized. This sequence may have to be repeated many times to achieve the desired purity.

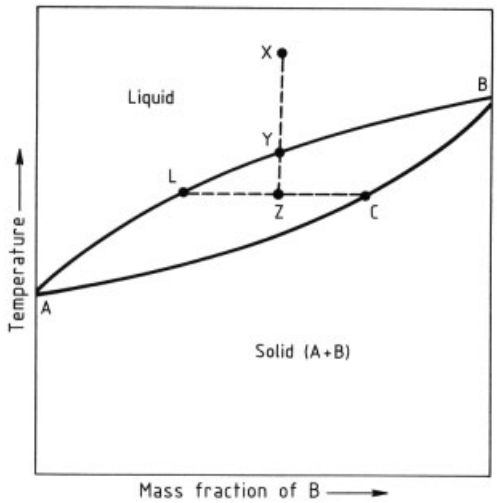

Figure 4. Phase diagram for a binary system AB composed of a continuous series of solid solutions
For explanation of symbols, see text.

In brief, a simple eutectic system may be purified in a single-stage crystallization operation, whereas a solid-solution system always needs a multistage operation.

3.2.3. Compound Formation

The solute and solvent of a binary system can combine to form one or more different compounds, for example, hydrates in aqueous solutions. If the compound can coexist in stable equilibrium with a liquid phase of the same composition, it is said to have a *congruent* melting point; i.e., melting occurs without change in composition (Fig. 5). If it cannot, the melting point is said to be *incongruent* (Fig. 6).

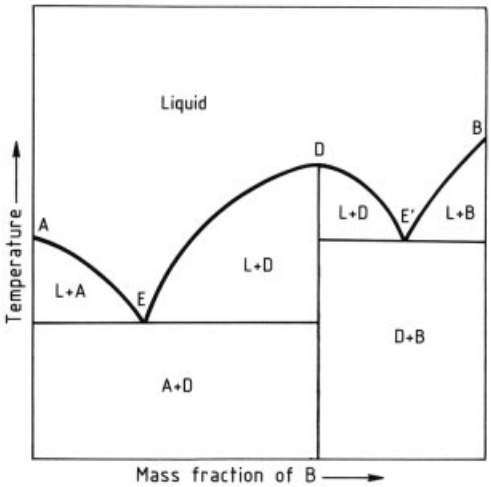

Figure 5. Phase diagram of a binary system AB that forms a compound D with a congruent melting point
Symbols indicate the phases coexisting in equilibrium: L = liquid; E, E′ = eutectics.

In Figure 5, the heating – cooling cycle follows the vertical line through point D; i.e., melting and crystallization occur without any change of composition. In Figure 6, however, compound D decomposes at a temperature T_1 below its theoretical melting point T_2. Thus, if compound D is heated, melting begins at temperature T_1, but it is not complete. At temperature T_1, the system of overall composition D contains crystals of pure component B in a melt of composition C. If this mixture is then cooled, a solid mixture of B and C is obtained. Subsequent heating and cooling cycles result in further decomposition of compound D. Inorganic salt hydrates are of interest as heat-storage materials, particularly for storage of solar heat in domestic and industrial space heating. (→ Heat Storage Media) Ideally, the hydrate should have a congruent melting point so that sequences of crystallization – melting – crystallization can be repeated indefinitely. Incongruently melting hydrate systems tend to stratify on repeated temperature cycling, with a consequent loss of efficiency. This occurs because melting yields a liquid phase that contains crystals of a lower hydrate or the anhydrous salt, which settle to the bottom of the container and fail to redissolve on subsequent heating. Calcium chloride hexahydrate, although not a true congruently melting hydrate, appears to be one of the most promising materials [16], [17]. Sodium sulfate decahydrate, sodium acetate trihydrate, and sodium thiosulfate pentahydrate, which have incongruent melting points, are also worthy of mention [18], [19].

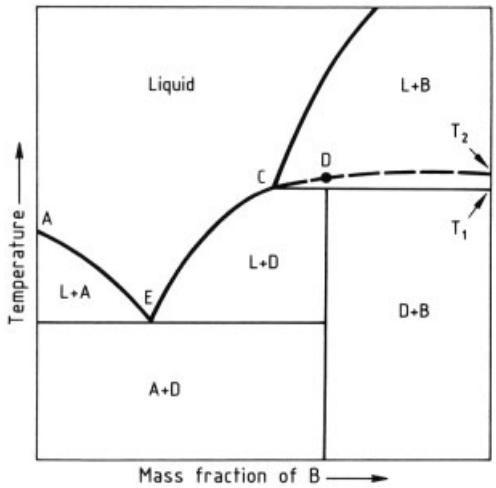

Figure 6. Phase diagram of a binary system AB that forms a compound D with an incongruent melting point
Symbols indicate the phases coexisting in equilibrium: L = liquid; E = eutectic; T_1 = decomposition temperature of D; T_2 = theoretical melting point of D.

3.3. Three-Component Systems

The phase equilibria in three-component (ternary) systems can be affected by four variables: temperature, pressure, and the compositions of any two of the three components. The effect of pressure (usually negligible in normal ranges) is generally ignored, and the phase equilibria are plotted on an isothermal triangular diagram.

Equilibrium relationships in three-component systems can be represented on a temperature – concentration space model as shown

in Figure 7A. The ternary system 2-, 3-, and 4-nitrophenol (*o* -, *m*-, and *p*-nitrophenol), in which compound formation does not occur, has been chosen for illustrative purposes. The three components are referred to as O, M, and P, respectively. Points O′, M′, and P′ on the vertical edges of the model represent the melting points of the pure components *o*- (45 °C), *m*- (97 °C), and *p*-nitrophenol (114 °C). The vertical faces of the prism represent temperature – concentration diagrams for the three binary eutectic systems O – M, O – P, and M – P, which are all similar to that shown in Figure 3.

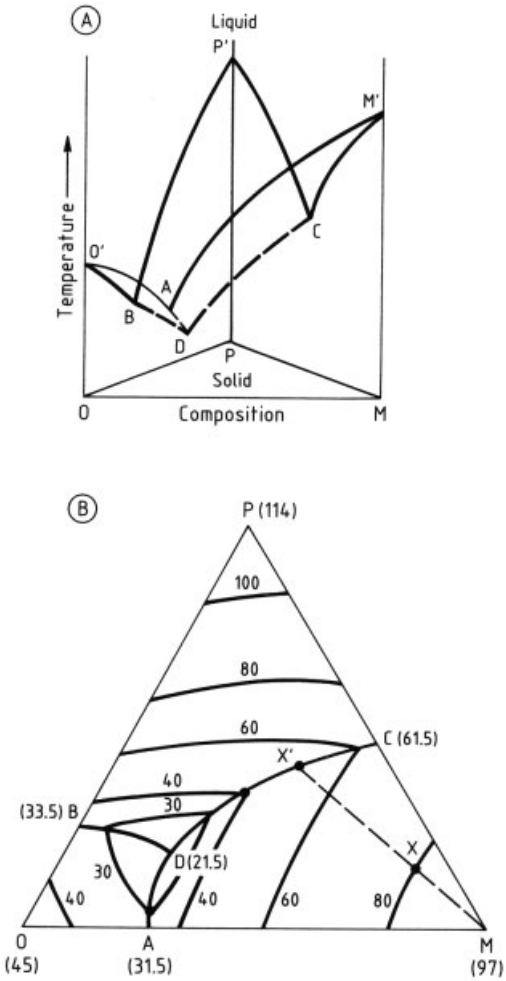

Figure 7. Eutectic formation in the ternary system *o*-, *m*-, and *p*-nitrophenol
A) Temperature – concentration space model; B) Projection on a triangular diagram
Numerical values represent temperatures in °C.

The binary eutectics are represented by points A (31.5 °C; 72.5 % O, 27.5 % M), B (33.5 °C; 75.5 % O, 24.5 % P), and C (61.5 °C; 54.8 % M, 45.2 % P). Curve AD within the prism represents the effect of addition of component P to the O – M binary eutectic A. Similarly, curves BD and CD denote the lowering of freezing points of the binary eutectics B and C, respectively, upon addition of the third component. Point D, indicating the lowest temperature at which solid and liquid phases can coexist in equilibrium in this system, is a ternary eutectic point (21.5 °C; 57.7 % O, 23.2 % M, 19.1 % P). At this temperature and concentration, the liquid freezes to form a solid mixture of the three components. The section of the space model above the freezing point surfaces formed by the liquidus curves represents the homogeneous liquid phase. The section below these surfaces down to a temperature represented by point D denotes solid and liquid phases in equilibrium. The section of the model below this temperature represents a completely solidified system.

Figure 7B is the projection of the curves AD, BD, and CD in Figure 7A onto the triangular base of the prism. The apexes of the triangle represent pure components O, M, and P, and their melting points are indicated in parentheses. Points A, B, and C on the sides of the triangle indicate the three binary eutectic points; point D is the ternary eutectic point. The projection diagram is divided by curves AD, BD, and CD into three regions which denote the three liquidus surfaces in the space model. The temperature falls from the apexes and sides of the triangle toward the eutectic point D; several isotherms showing points on the liquidus surfaces are also drawn. The phase reactions occurring when a given ternary mixture is cooled can be traced. A molten mixture with a composition X starts to solidify when the temperature is reduced to 80 °C. Point X lies in the region ADCM, so pure *m*-nitrophenol is deposited when the temperature is decreased. The composition of the remaining melt changes along line MX′ in the direction away from point M which represents the deposited solid phase (the mixture rule). At X′, where line MX′ intersects curve CD, the temperature is about 50 °C and *p*-nitrophenol also starts to crystallize. On further cooling, both *m*- and *p*-nitrophenol are deposited, and the composition of the liquid phase changes in the di-

rection X′D. When the melt composition and temperature reach point D, the third component (*o*-nitrophenol) also crystallizes, and the system solidifies without further change in composition. Similar reasoning can be applied to the cooling or melting of systems represented by points in other regions of the diagram.

Many different types of phase behavior are encountered in ternary systems that consist of water and two solid solutes. One simple case is considered here: the system $KNO_3 - NaNO_3 - H_2O$ at 50 °C (Fig. 8). The salts do not form hydrates nor do they combine chemically. Point A represents the solubility of KNO_3 in water at 50 °C (46.2 g per 100 g of solution) and point C the solubility of $NaNO_3$ (53.2 g per 100 g of solution). Curve AB indicates the composition of saturated ternary solutions in equilibrium with solid KNO_3 and curve BC those in equilibrium with solid $NaNO_3$. The upper area enclosed by ABC represents the region of unsaturated homogeneous solutions. Three other triangular areas are constructed by drawing straight lines from point B to the two remaining apexes of the triangle; the compositions of the phases within these regions are marked on the diagram. At point B, the solution is saturated with both KNO_3 and $NaNO_3$.

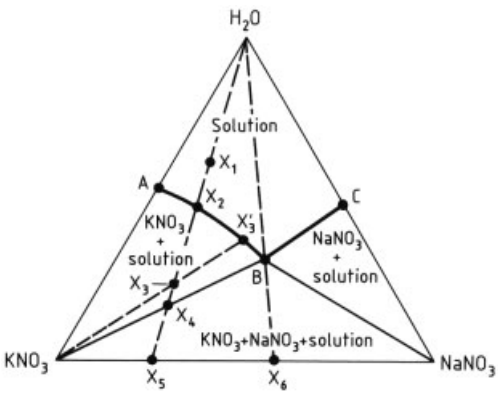

Figure 8. Phase diagram for the ternary system $KNO_3 - NaNO_3 - H_2O$ at 50 °C

If water is evaporated isothermally from an unsaturated solution represented by point X_1, the solution concentration increases along the line X_1X_2. Pure KNO_3 is deposited when the concentration reaches point X_2. If more water is evaporated to give a system of composition X_3, the solution composition is represented by point X_3' on the saturation curve AB, and by point B when composition X_4 is reached; further removal of water causes deposition of $NaNO_3$. All solutions in contact with solid will thereafter have a constant composition B, which is referred to as the *eutonic point* or *drying-up point* of the system. After complete evaporation of water, the composition of the solid residue is indicated by point X_5 on the base line. Similarly, if an unsaturated solution, represented by a point located to the right of B in the diagram, is evaporated isothermally, only $NaNO_3$ is deposited until the solution composition reaches point B. The salt KNO_3 is then also deposited, and the solution composition remains constant until evaporation is complete. If water is removed isothermally from a solution of composition B, the composition of deposited solid is given by point X_6 on the base line, and it remains unchanged throughout the evaporation process.

3.4. Multicomponent Systems

The more components a system contains, the more complex are the phase equilibria and the more difficult it is to represent phase reactions graphically. Specific descriptions of multicomponent solid – liquid diagrams and their uses are dealt with comprehensively in several books and monographs [7], [20–23]. Techniques for predicting multicomponent solid – liquid phase equilibria have also been described [22–26].

3.5. Phase Transformations

Metastable crystalline phases frequently crystallize prior to an expected, more stable phase, in accordance with Ostwald's rule of stages (cf. Section 8.2). The more common types of phase transformation that occur in crystallizing and precipitating systems include those between polymorphs (Section 3.1) and solvates. Transformations can occur in the solid state, particularly at temperatures near the melting point of the crystalline solid, or due to the intervention of a solvent.

A stable phase has a lower solubility than a metastable phase, as indicated by the solubility curves in Figures 9A and 9 B for enantiotropic

(reversible) and monotropic (irreversible) systems, respectively. Transformation cannot occur between the metastable (I) and stable (II) phases in the monotropic system in the temperature range shown, but it is possible above the transition temperature in an enantiotropic system. Below the transition temperature metastable (I) can transform to stable (II). Above the transition temperature (II) can transform to (I).

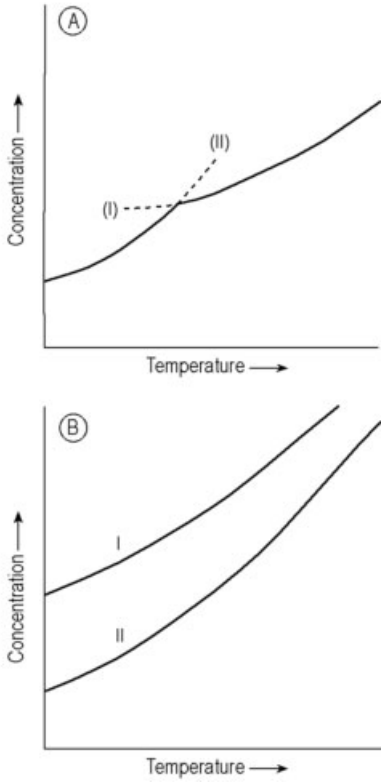

Figure 9. Solubility curves for substances with two polymorphs I and II
A) Enantiotropic polymorphs; B) Monotropic polymorphs
Broken lines in A indicate the metastable solubility curves of the two enantiomorphs.

Transformation is not inevitable even though a system may have entered a condition that would theoretically allow it. Transformation can only be ensured if crystals of a more stable solid phase are present, e.g., by deliberate seeding or by spontaneous nucleation. Theoretical analyses of both solid-state and solvent-mediated transformations are presented in [27], [28].

Polymorphs and Solvates. The ability of a single compound to crystallize in more than one crystallographic form (polymorphism) is encountered in a wide range of industries including pharmaceuticals, dyestuffs, agrochemicals, photochemicals, and other specialty compounds, both organic and inorganic. Many compounds may also crystallize in the form of different solvates, e.g., hydrates. These are not strictly speaking polymorphs, since they are not chemically identical, but like polymorphs they are capable of transforming from one form to another and hence can conveniently be included in the following general considerations.

As described in Section 3.1 (One-Component Systems), all polymorphs of a given substance, although chemically identical, exhibit different physical properties that can considerably affect the end-use of the material. For example, one polymorph of a pharmaceutical compound may have quite significant differences in solubility, bioavailability and pharmacological action from another. It is essential, therefore, when investigating a new chemical substance with a view to eventual industrial production, to investigate at an early stage the possibility of the existence of different polymorphs or solvates and how they may be isolated. Without such an investigation, new phases may make a sudden appearance years after the development of the first form and this can often be very inconvenient. However, the existence of different polymorphs or solvates need not always be considered as a problem. Indeed, if discovered early enough, they may have potential commercial advantages in allowing increased patent coverage and greater flexibility in, for example, the formulation of pharmaceuticals.

For monotropic polymorphs (Figure 9B) only one form is thermodynamically stable at all temperatures in the range considered. All other forms are metastable and potentially capable of transforming into the stable form. For enantiotropic polymorphs (Figure 9A) two or more forms are thermodynamically stable over different temperature ranges. In these cases it is essential to identify the transition temperatures as accurately as possible since they mark boundaries that should be avoided when planning processing operations. Furthermore, compounds with transition temperatures within the ambient range are likely to cause problems in subsequent storage and use.

Laboratory procedures that can be adopted in the preliminary search for possible polymorphs or solvates include: crystallizing from a wide range of solvents (polar, non-polar, hydrophilic, and hydrophobic) at different temperatures; chilling saturated solutions rapidly; precipitation by rapid quenching with a liquid nonsolvent; heating excess solid with a high boiling solvent; crystallization from the melt or by sublimation, and so on. The identity and purity of all product crystals should then be checked by appropriate analytical techniques.

Once different polymorphs or solvates have been isolated and identified, samples can be used as seed crystals in subsequent crystallization operations to promote the production of a specific form. It is important to identify any process conditions that could result in polymorphic transformation, e.g., the initiation of a solvent-mediated transformation in monotropic systems during a drying operation. Solid phase transformations can be temperature dependent, but they can also occur during energetic processes such as grinding or tabletting.

The process implications of polymorphism in organic compounds and some general recommendations for the batch cooling crystallization of a desired polymorph may be summarized as follows [28]. First, it is necessary to isolate and identify each polymorph and to generate solubility data in more than one solvent in order to determine if the system is monotropic or enantiotropic. This information will help in the selection of a solvent for the industrial process, although the ultimate choice will have to include considerations of process yield, solvent recovery costs, hazards, etc.

For enantiotropic systems the temperature range of the crystallization process will be determined by the particular polymorph required. If it is metastable below the transition temperature, crystallization should begin just above transition temperature, where the kinetics are relatively slow. Seeding with the desired polymorph at this point recommended. If the required polymorph is stable below the transition temperature, seeding should be commenced just below the transition temperature.

Seeding is also recommended for monotropic systems. For the stable polymorph, the temperature after seeding should be held constant for a predetermined time to allow the solution to desupersaturate, after which an appropriate cooling profile (see Section 5.6 Controlled Crystallization) should be adopted to maintain a constant controlled supersaturation. If the metastable polymorph is required, and the solution is supersaturated with respect to the stable form, the solution should be seeded with the metastable form and cooled rapidly to avoid any solvent-mediated transition. As noted above, the drying conditions for metastable polymorphs must be chosen carefully to avoid any solvent-mediated transformation occurring.

4. Kinetics and Mechanisms of Crystallization

4.1. Crystal Nucleation

Nucleation, i.e., the creation of crystalline bodies within a supersaturated fluid, is a complex, often ill-defined event, and nuclei may be generated by many different mechanisms. Numerous nucleation classification schemes have been proposed; most distinguish between two basic modes:

1) *Primary nucleation* – in the absence of crystals
2) *Secondary nucleation* – in the presence of crystals

Several comprehensive reviews of general nucleation phenomena are given in [5], [7], [29], [30].

4.1.1. Primary Nucleation

Classical theories of primary nucleation are based on sequences of bimolecular collisions and interactions in a supersaturated fluid, which result in the buildup of lattice-structured bodies that may or may not achieve thermodynamic stability. This type of primary nucleation is referred to as *homogeneous*, although the terms *spontaneous* and *classical* have also been used.

Ample experimental evidence demonstrates that ordered solute clustering can occur in supersaturated solutions prior to the onset of homogeneous nucleation [6], [31–34]. Concentration gradients develop readily in supersaturated solutions of citric acid [33] under the influence of

gravity; theoretical analysis of this phenomenon estimates the size of the clusters at 4 – 10 nm [34].

Primary nucleation may also be initiated by suspended particles of foreign substances (e.g., dust or other debris). This mechanism is generally referred to as *heterogeneous* nucleation.

Most primary nucleation in industrial crystallization is almost certainly heterogeneous rather than homogeneous; i.e., it is induced by the foreign solid particles that are invariably present in working solutions. The mechanism of heterogeneous nucleation is not well understood, but it probably begins with adsorption of the crystallizing species on the surface of solid particles, thus creating apparently crystalline bodies larger than the critical nucleus size. These stable particles then grow into macrocrystals.

Homogeneous Nucleation. Consideration of the energy involved in solid-phase formation and in creation of the surface of an arbitrary spherical crystal of radius r in a supersaturated fluid leads to the relationship

$$\Delta G = 4\pi r^2\gamma + (4\pi/3)\, r^3 \Delta G_v \tag{16}$$

where ΔG is the overall excess free energy associated with the formation of the crystalline body, γ is the interfacial tension between the crystal and its surrounding supersaturated fluid, and ΔG_v is the free energy change associated with the phase change. The first term on the right-hand side of Equation (16), representing the surface contribution, is positive and proportional to r^2. The second term, representing the volume contribution, is negative and proportional to r^3. The overall dependence of ΔG on r is shown in Figure 10. Any crystal smaller than the critical nucleus size r_c is unstable and will tend to dissolve. Any crystal larger than r_c is stable and will tend to grow.

Combination of Equations (15) and (16) and expression of the rate of nucleation J in the form of an Arrhenius reaction rate equation give

$$J = A\exp\left[\frac{-16\pi\gamma^3 v^2}{3k^3T^3\,(\ln S)^2}\right] \tag{17}$$

where A is a preexponential factor, γ is interfacial tension, v is molar volume, k is the Boltzmann constant, T is absolute temperature, and S is the supersaturation ratio. Equation (17) not only demonstrates the powerful effect of supersaturation on homogeneous nucleation, predicting an explosive increase in the nucleation rate beyond some so-called critical value of S, but also indicates the possibility of nucleation at any level of supersaturation.

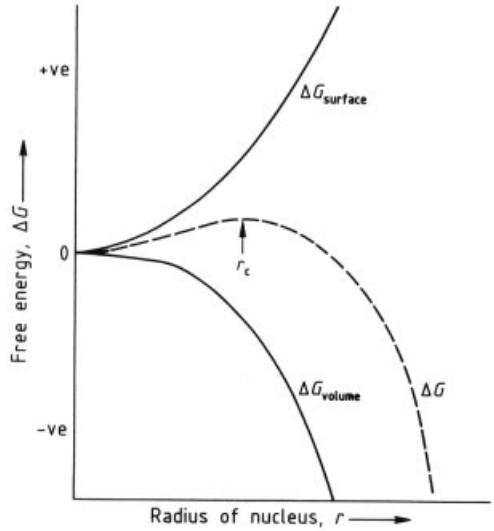

Figure 10. Free energy diagram for homogeneous nucleation demonstrating the critical nucleus size

Heterogeneous Nucleation. The presence of foreign particles (heteronuclei) enhances the nucleation rate of a given solution. Equations similar to that for homogeneous nucleation (Eq. 17) have been proposed to express this enhancement. However, the result is simply a displacement of the nucleation rate vs. supersaturation curve as shown in Figure 11, indicating that heterogeneous nucleation occurs more readily, i.e., at lower supersaturation.

For primary nucleation in industrial crystallization, classical relationships like those based on Equation (18) have little practical use. All that can be justified is a simple empirical relationship such as

$$J = K_n \Delta c^n \tag{18}$$

which relates the primary nucleation rate J to the supersaturation Δc (Eq. 1). The primary nucleation rate constant K_n and the order of the nucleation process n depend on the physical properties and hydrodynamics of the system. Values of n are usually greater than 2.

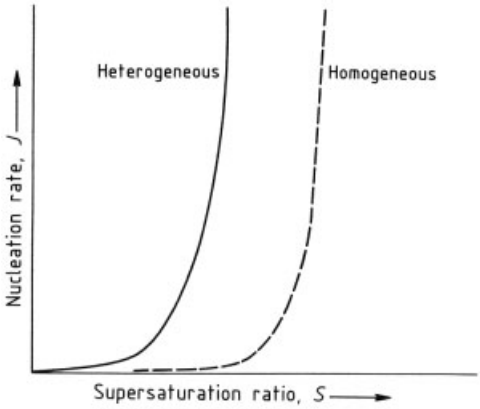

Figure 11. Effect of supersaturation on the rates of homogeneous and heterogeneous nucleation

4.1.2. Secondary Nucleation

Secondary nucleation, by definition, can take place only if crystals of the species under consideration are already present. Since this is mainly the case in working crystallizers, secondary nucleation has a profound influence on virtually all industrial crystallization processes.

Apart from deliberate or accidental introduction of tiny seed crystals to the system, and productive interactions between existing crystals and quasicrystalline embryos or clusters in solution, the most influential mode of new crystal generation in an industrial crystallizer is *contact* secondary nucleation. The contact in this case can be between the existing crystals themselves, between crystals and the walls or other internal parts of the crystallizer, or between crystals and the mechanical agitator.

Secondary nucleation rates in industrial crystallizers are most commonly correlated by empirical relationships such as

$$B = K_{\mathrm{b}} M_{\mathrm{T}}^{j} N^{l} \Delta c^{b} \tag{19}$$

where B is the rate of secondary nucleation (birthrate), K_{b} is the birthrate constant, M_{T} is the slurry concentration (magma density), N is a term that gives some measure of the intensity of agitation in the system (e.g., the rotational speed of an impeller), and Δc is the supersaturation. The exponents j, l, and b vary according to the operating conditions.

4.1.3. Nucleation Measurements

One of the earliest attempts to derive meaningful nucleation kinetics for solution crystallization was proposed by NÝVLT [35]. The method is based on the measurement of metastable zone widths (Section 2.2) by using a simple apparatus (Fig. 12) that consists of a 50 mL flask fitted with a thermometer and a magnetic stirrer, located in an external cooling bath. A solution of known composition is heated to about 5 °C above its saturation temperature and then cooled at a slow steady rate. The temperature at which nuclei appear is recorded. The difference between the saturation and nucleation temperatures is the maximum allowable undercooling (the metastable zone width) corresponding to the particular cooling rate used. Both primary and secondary nucleation can be studied this way.

The typical results [7] shown in Figure 13 demonstrate that seeding has a considerable influence on the nucleation process. The difference between the slopes of the two lines indicates that primary and secondary nucleation occur by different mechanisms.

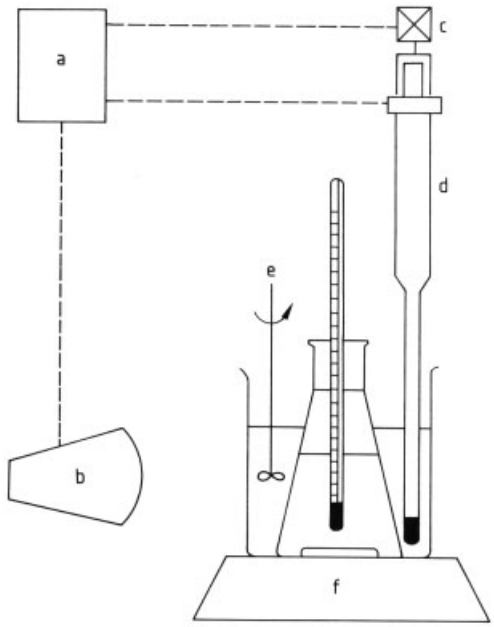

Figure 12. Simple apparatus for measuring metastable zone widths [35]
a) Control unit; b) Hot – cold air blower; c) Rotating magnet; d) Contact thermometer; e) Water bath stirrer; f) Magnetic stirrer

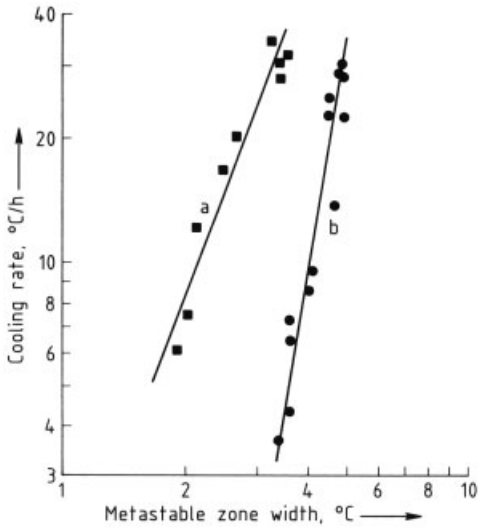

Figure 13. Metastable zone width of aqueous ammonium sulfate solutions as a function of cooling rate [7]
a) Seeded solution; b) Unseeded solution

Solution turbulence also affects nucleation. In general, agitation reduces the metastable zone width. For example, the metastable zone width for gently agitated potassium sulfate solutions is about 12 °C; vigorous agitation reduces this to ca. 8 °C. The presence of crystals also induces secondary nucleation at supercooling around 4 °C. The relation between supercooling $\Delta\theta$ and supersaturation Δc is given by Equation (14).

Useful information on secondary nucleation kinetics for crystallizer operation and design can best be determined from model experiments that employ techniques such as those developed for MSMPR (mixed-suspension mixed-product removal) crystallizers (see Section 5.8). In a real crystallizer, however, both nucleation and growth proceed together and interact with other system parameters in a complex manner [7], [35], [36].

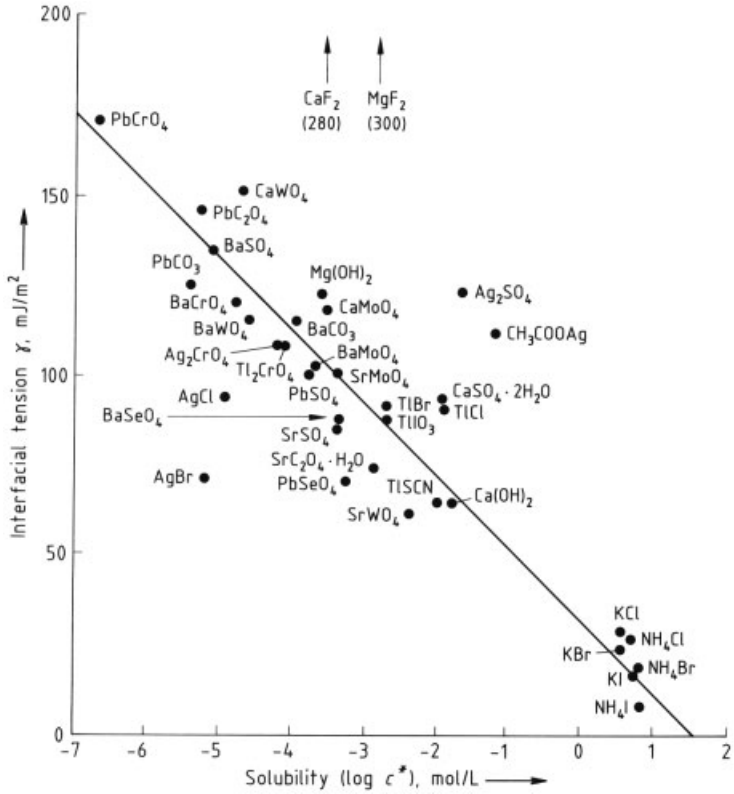

Figure 14. Interfacial tension as a function of solubility [37]

4.1.4. Induction Periods

A delay occurs between achievement of supersaturation and detection of the first newly created crystals in a solution. This so-called induction period τ is a complex quantity that involves both nucleation and growth components. However, if the assumption is made that τ is essentially concerned with nucleation (i.e., τ is inversely proportional to J), then from the classical nucleation equation (Eq. 6) it follows that [7]

$$\log\tau \propto \gamma^3/T^3\ (\log S)^2 \tag{20}$$

Thus, for a given temperature T, a plot of log τ vs. $(\log S)^{-2}$ should yield a straight line which, if the data truly represent homogeneous nucleation, will allow calculation of the interfacial tension γ. The effect of temperature on γ can also be evaluated. An attempt has been made to derive a general correlation between interfacial tension and the solubility of inorganic salts (Fig. 14, see previous page) [37]. The link between interfacial tension and solubility can be substantiated on the basis of regular solution theory [9]. However, the success of this method for evaluating interfacial tension depends on precise measurement of the induction period τ, which presents problems if τ is less than a few seconds.

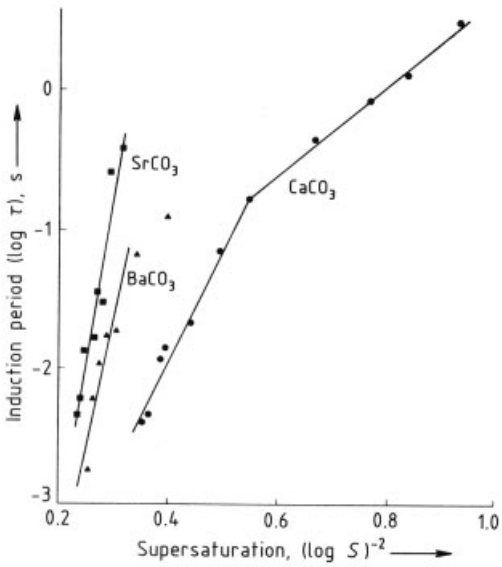

Figure 15. Induction period as a function of initial supersaturation [39]

Short induction periods can be determined by a stopped-flow technique that detects rapid changes in the conductivity of a supersaturated solution [39] (e.g., during a precipitation reaction). Typical results for $CaCO_3$, $SrCO_3$, and $BaCO_3$, produced by mixing an aqueous solution of Na_2CO_3 with a solution of the appropriate chloride, are shown in Figure 15 (see below). The slopes of the linear (high-supersaturation) regions are used to calculate the interfacial tensions (80 – 120 mJ/m^2), which compare reasonably well with the values predicted from the interfacial tension – solubility relationship in Figure 13.

4.2. Crystal Growth

(For a detailed discussion of crystal growth, see → Crystal Growth.)

Once a stable nucleus has been created in a solution, it is capable of growing into a crystal. In its simplest form, crystal growth may be considered a two-step process involving (1) mass transport, either by diffusion or convection from the bulk solution to the crystal face, followed by (2) a surface reaction in which the growth units are integrated into the crystal lattice. Either step may control the overall growth process, although convective mass transport is unimportant for crystals smaller than ca. 10 µm because these crystals are scarcely affected by turbulent eddies and diffusional mass transport predominates.

For diffusion-controlled growth, the theoretical linear crystal growth rate G for a nonionic species may be calculated from

$$G = Dv\Delta c/L \tag{21}$$

where D is the diffusion coefficient, v the molar volume, Δc the supersaturation, and L the crystal size; i.e., the diffusion-controlled growth rate is directly proportional to the degree of supersaturation [7]. However, most crystal growth processes are not simply diffusion-controlled because the surface reaction (integration) step generally plays a contributing role. The two most common surface processes are referred to as spiral and polynuclear growth. *Spiral growth* (also known as BCF growth because it was first postulated by Burton, Cabrera, and Frank [40]) stems from the emergence of screw dislocations on a crystal face, as illustrated in Figure 16.

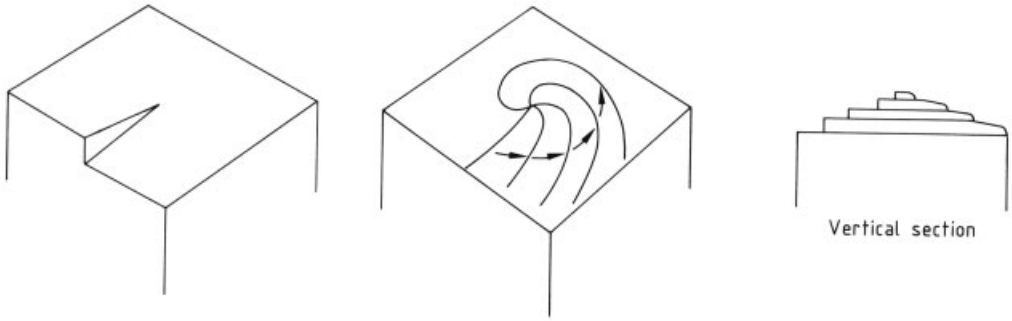

Figure 16. Spiral (BCF) crystal growth beginning at a screw dislocation

Growth proceeds at relatively low supersaturation at a rate proportional to the square of the supersaturation.

Polynuclear growth develops from surface (monolayer) nucleation on the edges, corners, and faces of a crystal. These surface nuclei spread across the crystal face, and further nuclei develop on them. This mechanism, illustrated in Figure 17, is sometimes referred to as the B + S (birth and spread) model. Polynuclear growth is related to supersaturation by a complex exponential relationship [41].

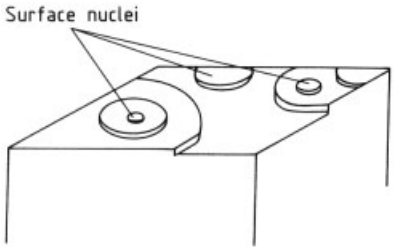

Figure 17. Polynuclear (B + S) crystal growth initiated by surface nucleation

At first sight, therefore, identifying the mechanism of growth by examining the relationship between the experimentally determined growth rate and supersaturation might appear to be possible. In practice, however, this is not easy because of the many errors inherent in growth rate measurement and, consequently, the relatively poor reproducibility of the data.

Several comprehensive reviews of modern theories of crystal growth are available [5], [7], [41–43] (see also → Crystal Growth).

4.2.1. Measurement of Growth Rate

A variety of methods have been used to measure crystal growth rate; they are divided into two main categories: (1) direct measurement of the linear growth rate of a chosen crystal face and (2) indirect estimation of an overall linear growth rate from mass deposition rates measured either on individual crystals or on groups of freely suspended crystals [4], [7], [35], [44], [45].

Face Growth Rates. Different crystal faces grow at different rates. In general, high-index faces grow faster than low-index faces. Changes in growth environment (temperature, supersaturation, pH, impurities, etc.) can also have a profound effect on growth. Differences in individual face growth rates give rise to habit (shape) changes in crystals (see Section 4.4).

Equipment for precise measurement of individual crystal face growth rates is depicted in Figure 18. A fixed crystal in a glass cell is observed with a traveling microscope. Solution temperature, supersaturation, and velocity are precisely controlled [7].

The solution velocity past the fixed crystal is frequently an important growth-determining parameter. It is sometimes responsible for the so-called size-dependent growth effect often observed in agitated vessel and fluidized-bed crystallizers. Large crystals have higher settling velocities than small crystals and, if their growth is diffusion-controlled, they tend to grow faster. Other reasons for size-dependent growth rates are discussed in Section 4.2.3. Examples of salts that exhibit solution-velocity-dependent growth rates include the alums, nickel ammonium sul-

fate, and potassium sulfate. However, salts such as ammonium sulfate and ammonium or potassium dihydrogen phosphate are not affected by solution velocity.

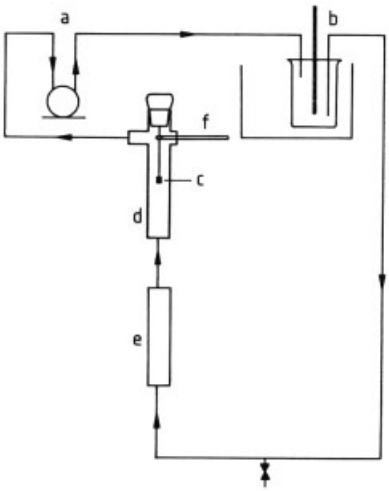

Figure 18. Equipment for measuring the growth of a single crystal
a) Pump; b) Saturator; c) Single crystal fixed on a wire; d) Glass cell; e) Rotameter; f) Thermometer

Overall Growth Rates. In the laboratory, growth rate data for crystallizer design can be measured in fluidized beds or agitated vessels. A typical laboratory fluidized-bed crystallizer is shown in Figure 19. Crystal growth rates are measured by growing large numbers of carefully sized seeds in fluidized suspension under strictly controlled conditions. A warm undersaturated solution of known concentration is circulated in the crystallizer and supersaturated by cooling to the working temperature. About 5 g of closely sized seed crystals with a narrow size distribution and a mean size of ca. 500 µm is introduced into the crystallizer, and the upward solution velocity is adjusted so that the crystals are maintained in a reasonably uniform fluidized state in the growth zone. The crystals are allowed to grow at a constant temperature until their total mass is ca. 10 g. They are then removed from the crystallizer, washed, dried, and weighed. The final solution concentration is measured, and the mean of the initial and final supersaturations is taken as the average for the run. This assumption does not involve any significant error because the solution concentration is usually not allowed to change by more than about 1 % during a run. The overall crystal growth rate is then calculated in terms of mass deposited per unit area per unit time at a specified supersaturation.

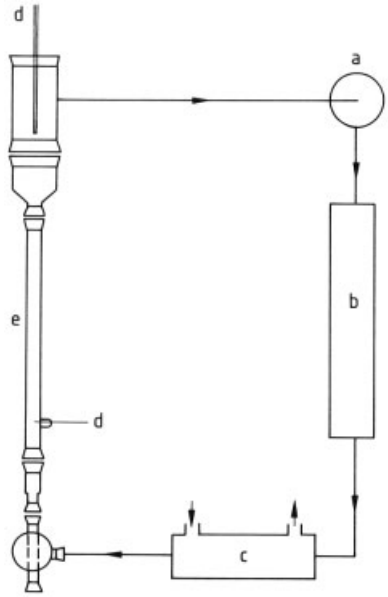

Figure 19. Laboratory-scale fluidized-bed crystallizer
a) Pump; b) Heater; c) Cooler; d) Thermometer; e) Crystal bed

4.2.2. Expression of Growth Rate

Crystal growth rates are commonly expressed in three different ways:

1) As mass deposition rate R, e.g., $\mathrm{kg\,m^{-2}\,s^{-1}}$
2) As mean linear growth velocity $\bar{v}$ (= dr/dt), m/s
3) As overall linear growth rate G (= dL/dt), m/s

The relationships between these quantities are

$$R = K_g \Delta c^g = \frac{1}{A} \cdot \frac{\mathrm{d}m}{\mathrm{d}t} = \frac{3\alpha\varrho G}{\beta} = \frac{6\alpha\varrho}{\beta} \cdot \frac{\mathrm{d}r}{\mathrm{d}t} = \frac{6\alpha\varrho\bar{v}}{\beta} \tag{22}$$

where L is some characteristic dimension of the crystal (e.g., the equivalent sieve aperture size), r is the radius of the equivalent sphere, and ϱ is the density of the crystal. The order of the growth process, g, is generally between 1 and 2. The volume and surface shape factors (α and β,

respectively) are related to particle mass m and surface area A, respectively, by

$$m = \alpha \varrho L^3 \tag{23}$$

and

$$A = \beta L^2 \tag{24}$$

For spheres and cubes, $6\,\alpha/\beta = 1$.

4.2.3. Dependence of Growth Rate on Crystal Size

A considerable amount of experimental evidence indicates that crystal growth kinetics often depend on crystal size. Apart from cases of solution velocity dependence (Section 4.2.1), this condition may result from the crystal size dependence of the surface integration kinetics. Different crystals of the same size can also have different growth rates, for example, because of differences in surface structure or perfection. Furthermore, small crystals ($< 50\,\mu$m) of many substances grow much more slowly than larger crystals, and some do not grow at all [46].

The behavior of very small crystals has considerable influence on the performance of continuously operated industrial crystallizers because new crystals with a size of $1-10\,\mu$m are constantly generated by secondary nucleation. These subsequently grow to populate the full crystal size distribution. Therefore, the ability to predict the growth rates of small crystals is useful in assessing the performance of crystallizers.

4.3. Growth – Nucleation Interactions

Crystal nucleation and growth in an industrial crystallizer cannot be considered in isolation because they interact with one another and with other system parameters in a complex manner. These factors are summarized in Figure 20.

For a complete description of the crystal size distribution of the product in a continuously operated crystallizer, both the nucleation and the growth processes must be quantified, and the laws of conservation of mass, energy, and crystal population must be applied. The importance

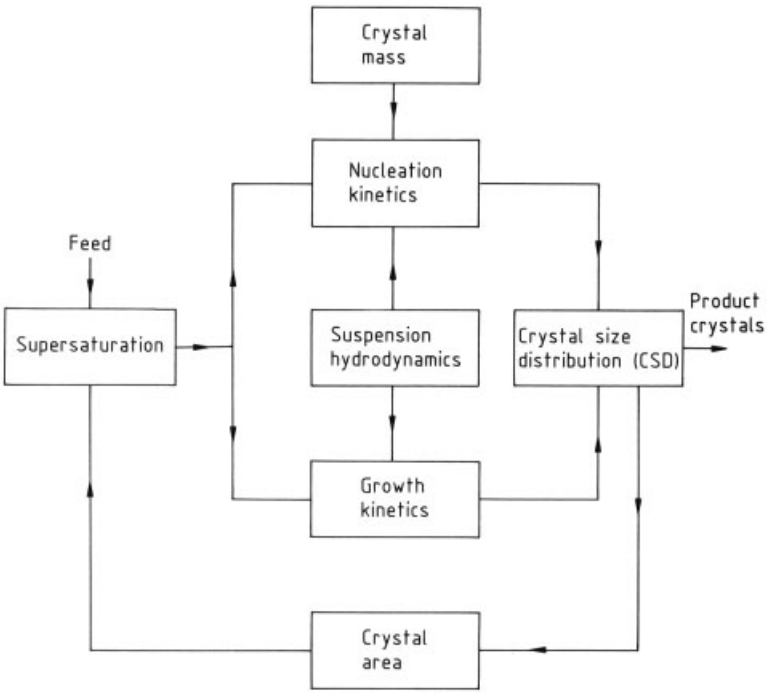

Figure 20. Complex interactions in a continuous crystallization process

of population balance, in which all particles are accounted for, was first stressed in the pioneering work of RANDOLPH and LARSON [36] (cf. Section 5.8.1).

4.4. Crystal Habit Modification

Changes in the face growth rates of crystals give rise to changes in their habit (shape). The growth kinetics of individual crystal faces usually depend to various extents on supersaturation, so that crystal habit can sometimes be controlled by adjusting operating conditions. The most common cause of habit modification, however, is the presence of impurities. The crystallizing solution may already contain impurities (e.g., raffinose which induces crystallization of characteristic flat sucrose crystals from beet sugar feedstocks), or they can be added deliberately.

In some cases, mere traces (ca. 1 mg/kg) of an impurity can cause significant changes in the crystal habit, as is the case for $Fe(CN)_6^{4-}$ which causes sodium chloride to crystallize in dendritic form. On the other hand, habit modification sometimes occurs only in the presence of large quantities of impurity, e.g., > 5 wt % of biuret is needed to change the shape of urea crystals from needles to bricks. Transition-metal ion species, such as complexes of Cr(III) and Fe(III), are particularly active impurities at < 10 mg/kg for a variety of inorganic salts in aqueous solution. Surface-active agents can also change crystal habit: anionic surfactants (e.g., alkyl sulfates; alkyl and arylalkyl sulfonates) and cationic surfactants (e.g., quaternary ammonium salts) are commonly used. Low molecular mass organic phosphonates and high molecular mass polyelectrolytes, polyacrylamides, poly(vinyl alcohols), etc., also find specific application. A change of solvent frequently results in a change of crystal habit because the solvent represents a massive impurity in a crystallizing system.

In the simple Kossel model of crystal growth, three basic sites may be considered for the adsorption of an impurity species: at a kink (Fig. 21A), at a step (layer front, Fig. 21B), or on a ledge (face) between steps (Fig. 21C). Adsorption at kinks, the key sites at which growth units are incorporated into the lattice, effectively reduces both their number and the velocity at which the step advances. Consequent increases in interkink distance diminish the importance of surface diffusion in the growth process and cause the step velocity to depend on kink density; this results in polygonization of the growth layers [47]. Adsorption at a step also reduces both the number of sites available for growth and the step velocity. However, unless the adsorbed units are closer together than the diameter of a critical two-dimensional nucleus, the advancing step can still squeeze between impurity species and progress relatively unimpeded. Step adsorption, therefore, implies a critical concentration of dissolved impurity, below which the step velocity is unaffected and above which it decreases rapidly. Adsorption of an impurity on a face is likely to be effective in cases where surface processes, particularly diffusion, play an important role in crystal growth. The result is a decrease in the surface flux of growth units to the step and, thus, in the step velocity.

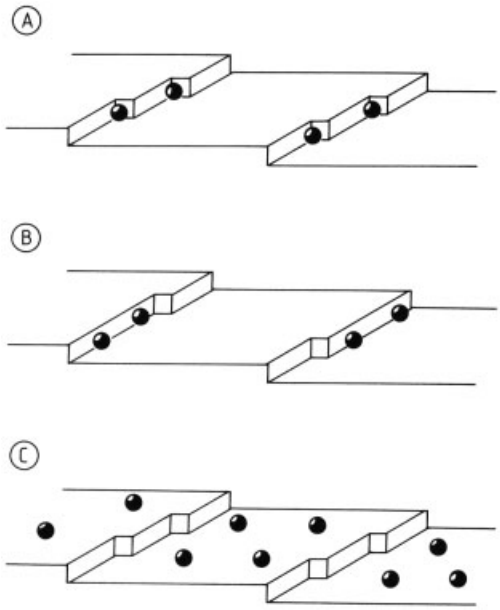

Figure 21. Kossel model of crystal growth showing sites for impurity adsorption on a crystal surface
A) At a kink; B) At a step; C) On a ledge (face)

In addition to physical obstruction, the effects of impurity adsorption and change of solvent on crystal growth may be explained in terms of changes in crystal surface properties. These changes can be quantified by the α factor, also called the surface entropy factor or surface roughness factor, which is related to the interfa-

cial tension γ between the growing crystal and its solution [48].

In brief, at low values of α, i.e., at low interfacial tension γ, the crystal surface is inherently rough, on the molecular scale; under these conditions, growth is controlled only by the rate of diffusion of growth units to the crystal face. At high values of α and γ, the surface is inherently smooth and growth cannot proceed at low supersaturation unless growth centers (e.g., emergent screw dislocations) exist on the face. Growth under these conditions may then proceed by the spiral (BCF) mechanism. At intermediate values of α and γ, growth may proceed by the polynuclear (B + S) mechanism (see Section 4.2).

The interfacial tension γ can be changed by additives. Adsorption on a crystal face, for example, reduces γ and could, therefore, change the growth mechanism. The dependence of growth rate on supersaturation thus changes, and if this occurs differently on different faces of the crystal, the habit can consequently change. Similar reasoning can be applied to the removal of a particular impurity from the crystallizing solution. Furthermore, since γ is a solvent-dependent property (it increases as solubility decreases), a change of solvent can have the same effect on crystal habit as an additive.

Some form of habit modification is employed in a large proportion of all industrial crystallization and precipitation operations to control the type of crystal produced and to improve the rheological properties of the slurry, downstream processes such as filtration or washing and the handling properties of the dried product and its stability on storage. This may be done by controlling the rate of crystallization (e.g., by adjusting the rate of cooling or evaporation, the degree of supersaturation, or the temperature at which crystallization occurs). Alternative methods involve choosing an appropriate solvent, adjusting the pH of the solution, or deliberately adding (or perhaps removing) some habit-modifying impurity to (or from) the system. A combination of several of these methods may have to be used.

The search for a habit modifier is complex, and the results of small-scale laboratory investigations may not always be useful for large-scale industrial application; in some cases, they can even be misleading. Realistic pilot-plant trials on batches greater than about 100 L, however, generally yield useful information.

Reviews on the industrial applications of habit modification and the selection of habit modifiers are given in [7], [49], [50].

4.5. Inclusions in Crystals

Inclusions are small pockets of solid, liquid, or gaseous impurities entrapped in crystals. They usually occur randomly throughout the crystal, but sometimes a regular pattern may be observed.

A simple technique for observing inclusions is to immerse the crystal in an inert liquid of similar refractive index or, alternatively, in its own saturated solution. In the latter case, the inclusion can be identified under the microscope by raising the temperature slightly to dissolve the crystal. If the inclusion is a liquid, concentration streamlines will be seen as the two fluids meet; if it is a vapor, a bubble will be released [7].

Crystals produced industrially may contain significant amounts of included mother liquor, in extreme cases up to 1 wt %, which can significantly affect product purity. Stored crystals may cake because of liquid seepage from inclusions in broken crystals, which leads to subsequent recrystallization (see Section 4.6). To minimize inclusions, the crystallizing system should be free of dirt and other solid debris to prevent its incorporation into the crystals. Vigorous agitation or boiling should be avoided because it can lead to the formation of air or vapor inclusions. Ultrasonic irradiation may suppress adherence of bubbles or particles to a growing crystal face and, hence, reduce inclusion formation. Fast crystal growth is probably the most common cause of inclusion formation; this means that, in general, high supersaturation should be avoided.

Reviews and accounts of crystal inclusions are available in [7], [51–53].

4.6. Caking of Crystals

Crystalline materials frequently cake (i.e., cement together) on storage. The size of the crystals, as well as their shape, moisture content, and storage conditions (time, temperature and humidity fluctuations, pressure, etc.), can all contribute to the caking tendency.

In general, caking is caused initially by dampening of the crystal surfaces in a storage container, e.g., because of inefficient drying or an increase in atmospheric humidity above some critical value that depends on both substance and temperature. For example, at 15 °C, crystals of $Na_2SO_4 \cdot 10\,H_2O$ become damp when atmospheric humidity exceeds ca. 93 % saturation; the same applies to NaCl at 78 % saturation and $CaCl_2 \cdot 6\,H_2O$ at 32 % saturation. The presence of a hygroscopic trace impurity in the crystals, therefore, can greatly influence their tendency to absorb atmospheric moisture. Moisture may also be released from inclusions if crystals fracture under storage conditions (Section 4.5). If crystal surface moisture later evaporates, e.g., because of atmospheric temperature or humidity changes, adjacent crystals become firmly joined together with a cement of recrystallized solute.

Precautions to reduce caking include efficient drying, packaging in airtight containers, and avoiding compaction on storage. In addition, crystals may be coated with an inert dust that acts as a moisture barrier; e.g., crystals of table salt are sometimes coated with finely powdered magnesium carbonate. However, the crystals themselves play a dominant role in caking. Small crystals are more prone to cake than large crystals because of the greater number of contact points per unit mass, but actual size is less important than size distribution and shape. The narrower the size distribution and the more granular the shape, the lower is the tendency of crystals to cake. Crystal size distribution can be controlled by adjusting operating conditions of the crystallizer (Section 5.4), and crystal shape may be influenced by the use of habit modifiers (Section 4.4).

Methods employed for the conditioning of crystals and testing procedures for caking tendency are discussed in [54–56]. A comprehensive account of caking inhibition by trace additives is given in [57].

5. Crystallization from Solutions

Solution crystallizers are generally classified according to the method by which supersaturation is achieved, e.g., cooling, evaporation, vacuum (adiabatic cooling), reaction, salting out. The designation *controlled* denotes supersaturation control; *classifying* refers to classification of product size. The term mixed-suspension mixed-product removal is abbreviated as MSMPR (see Section 5.8).

5.1. Cooling Crystallizers

5.1.1. Nonagitated Vessels

The simplest type of cooling crystallizer is the unstirred tank: a hot feedstock solution is charged to the open vessel where it is allowed to cool, often for several days, by natural convection. Metallic rods may be suspended in the solution so that large crystals can grow on them and reduce the amount of product that sinks to the bottom of the crystallizer. The product is removed by hand.

Because cooling is slow, large interlocked crystals are usually obtained and retention of mother liquor is unavoidable. As a result, the dried crystals are generally impure. Because of the uncontrolled nature of the process, product crystals range from a fine dust to large agglomerates.

Labor costs are generally high, but the method is economical for small batches because capital, operating, and maintenance costs are low. However, the productivity of this type of equipment is low and space requirements are high.

5.1.2. Agitated Vessels

Installation of an agitator in an open-tank crystallizer generally results in smaller, more uniform crystals and reduced batch time. The final product tends to have a higher purity because less mother liquor is retained by the crystals after filtration and more efficient washing is possible. Water jackets are usually preferred to coils for cooling because the latter often become encrusted with crystals and cease to operate efficiently. Where possible, the internal surfaces of the crystallizer should be smooth and flat to suppress encrustation.

An agitated cooler is more expensive to operate than a simple tank crystallizer, but it has a much higher productivity. Labor costs for product handling may still be rather high. The design

of tank crystallizers varies from shallow pans to large cylindrical tanks.

The large agitated cooling crystallizer shown in Figure 22A has an upper conical section which slows down the upward velocity of liquor and prevents the crystalline product from being swept out with the spent liquor. An agitator located in the lower region of a draft tube circulates the crystal slurry (magma) through the growth zone of the crystallizer. If required, cooling surfaces may be provided inside the crystallizer.

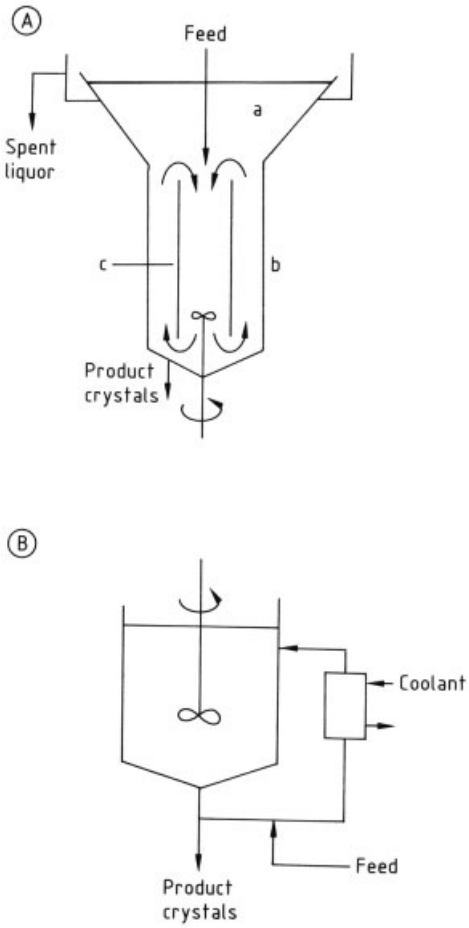

Figure 22. Cooling crystallizers
A) Internal circulation through a draft tube; B) External circulation through a heat exchanger
a) Calming section; b) Growth zone; c) Draft tube

Use of external circulation allows good mixing inside the crystallizer and high rates of heat transfer between the liquor and coolant (Fig. 22B). An internal agitator may be installed in the crystallization tank if needed. The liquor velocity in the tubes is high; therefore, small temperature differences are usually adequate for cooling purposes and encrustation on heat-transfer surfaces can be reduced considerably. The unit shown may be used for batch or continuous operation.

5.1.3. Direct-Contact Cooling

The use of a conventional heat exchanger and the problems caused by crystal encrustation can be avoided by employing direct-contact cooling (DCC) in which supersaturation is achieved by allowing the process liquor to come into contact with a cold heat-transfer medium. Other potential advantages of DCC over conventional indirect-contact methods include better heat transfer and smaller cooling load. However, problems are also associated with DCC crystallization; these include product contamination from the coolant and the cost of extra processing required to recover the coolant for further use.

A solid, liquid, or gaseous coolant can be used; heat exchange may occur via the transfer of sensible or latent heat. The coolant may or may not boil during the operation, and it can be miscible or immiscible with the process liquor. Thus, four basic types of DCC crystallization are possible [7]:

1) Immiscible, boiling, solid or liquid coolant: heat is removed mainly by transfer of latent heat of sublimation or vaporization.
2) Immiscible, nonboiling, solid, liquid, or gaseous coolant: mainly sensible heat transfer.
3) Miscible, boiling, liquid coolant: mainly latent heat transfer.
4) Miscible, nonboiling, liquid coolant: mainly sensible heat transfer.

Crystallization processes employing DCC have been used successfully in the dewaxing of lubricating oils, the desalination of water, and the production of inorganic salts from aqueous solution [7].

5.2. Evaporating Crystallizers

If the solubility of a solute in a solvent is not appreciably decreased by lowering the temper-

ature, the appropriate degree of solution supersaturation can be achieved by evaporating some of the solvent. Evaporation techniques have been used for centuries to crystallize salts; the simplest method – utilization of solar energy – is still employed commercially throughout the world [58]. Common salt is produced widely from brine in steam-heated evaporators, and similar evaporating crystallizers, often in multiple-effect series, are used in sugar refining. Many types of forced-circulation evaporating crystallizers are now in large-scale use [1–8].

Evaporating crystallizers are normally operated under reduced pressure to aid in solvent removal, minimize heat consumption, or decrease the operating temperature of the solution; they are best described as "reduced-pressure evaporating crystallizers".

5.3. Vacuum (Adiabatic Cooling) Crystallizers

A vacuum crystallizer operates on a slightly different principle from the reduced-pressure evaporating crystallizer described in the previous section.

Supersaturation is achieved in a vacuum crystallizer by simultaneous evaporation and adiabatic cooling of the feedstock. A hot, saturated solution is fed into an insulated vessel maintained under reduced pressure. If the feed liquor temperature is higher than the boiling point of the solution under the low pressure existing in the vessel, the liquor cools adiabatically to this temperature. The sensible heat and any heat of crystallization liberated by the solution evaporate some of the solvent and concentrate the solution.

5.4. Continuous Crystallizers

Many different continuously operated crystallizers are available, but the majority can be divided into three basic types: forced-circulation, fluidized-bed (Oslo), and draft-tube agitated units. A small selection of the large number of commercial types available is described.

Forced-Circulation Crystallizers. A *Swenson forced-circulation crystallizer* that operates under reduced pressure is shown in Figure 23. A high recirculation rate through the external heat exchanger is used to provide good heat transfer and minimize encrustation. The crystal magma is circulated from the lower conical section of the evaporator body through the vertical tubular heat exchanger and reintroduced tangentially into the evaporator below the liquor level to create a swirling action and prevent flashing (sudden evaporation). Feedstock enters on the pump inlet side of the circulation system. Product crystal magma is removed below the conical section.

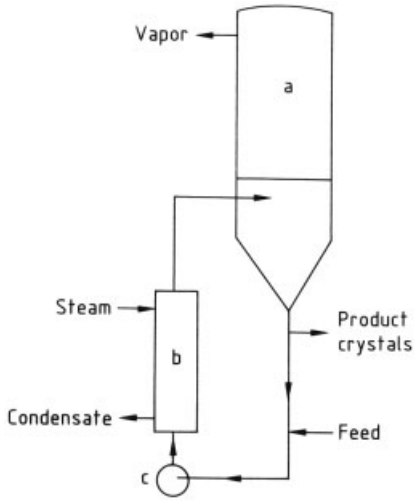

Figure 23. Forced-circulation Swenson crystallizer
a) Evaporator; b) Heat exchanger; c) Pump

Fluidized-Bed Crystallizers. In an *Oslo fluidized-bed crystallizer,* a bed of crystals is suspended in the vessel by the upward flow of supersaturated liquor in the annular region surrounding a central downcomer (Fig. 24). Although originally designed as classifying crystallizers, fluidized-bed Oslo units are now frequently operated in a mixed-suspension mode to improve productivity, although this reduces product crystal size. A cooling-type Oslo crystallizer operates in the classifying mode as follows. The hot, concentrated feed solution is fed into the vessel at a point directly above the inlet to the circulation pipe. Saturated solution from the upper regions of the crystallizer, together with the small amount of feedstock, is circulated through the tubes of the heat exchanger, which is cooled by forced circulation of water or brine. On cool-

ing, the solution becomes supersaturated, but not enough for spontaneous nucleation to occur; great care, in fact, is taken to prevent this. Product crystal magma is removed from the lower regions of the vessel.

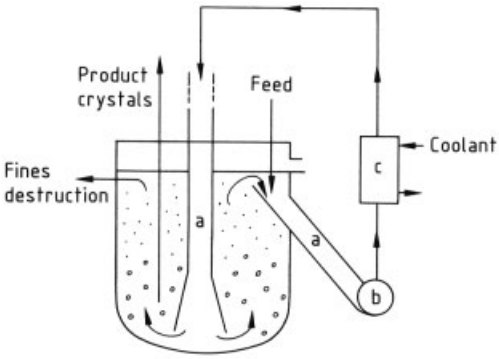

Figure 24. Oslo cooling crystallizer
a) Downcomer; b) Pump; c) Heat exchanger

Draft-Tube Agitated Vacuum Crystallizers. A *Swenson draft-tube-baffled* (*DTB*) *vacuum unit* is shown in Figure 25. A relatively slow-speed propeller agitator is located in a draft tube which extends to a few inches below the liquor level in the crystallizer. Hot, concentrated feedstock enters at the base of the draft tube. The steady movement of magma and feedstock up to the surface of the liquor produces a gentle, uniform boiling action over the whole cross-sectional area of the crystallizer. The degree of supercooling thus produced is very low ($< 1\,^{\circ}C$), and in the absence of violent vapor flashing, both excessive nucleation and salt buildup on the inner walls are minimized. The internal baffle in the crystallizer forms an annular space in which agitation effects are absent. This provides a settling zone that permits regulation of the magma density and control of the removal of excess nuclei. An integral elutriating leg may be installed underneath the crystallization zone (as depicted in Fig. 25) to effect some degree of product classification.

The *Standard–Messo turbulence crystallizer* (Fig. 26) is another successful draft-tube vacuum unit. Two liquor flow circuits are created by concentric pipes: an outer ejector tube with a circumferential slot and an inner guide tube. Circulation is effected by a variable-speed agitator in the guide tube. The principle of the Oslo crystallizer is utilized in the growth zone; partial classification occurs in the lower regions, and fine crystals segregate in the upper regions. The primary circuit is created by a fast upward flow of liquor in the guide tube and a downward flow in the annulus; liquor is thus drawn through the slot between the ejector tube and the baffle, and a secondary flow circuit is formed in the lower region of the vessel. Feedstock is

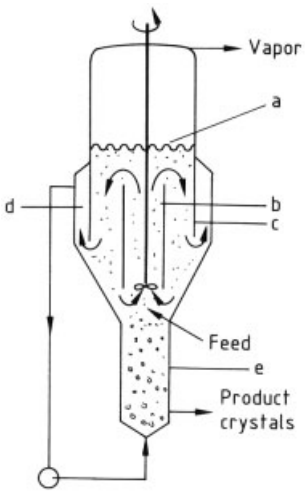

Figure 25. Swenson draft-tube-baffled (DTB) crystallizer
a) Boiling surface; b) Draft tube; c) Baffle; d) Settling zone; e) Elutriating leg

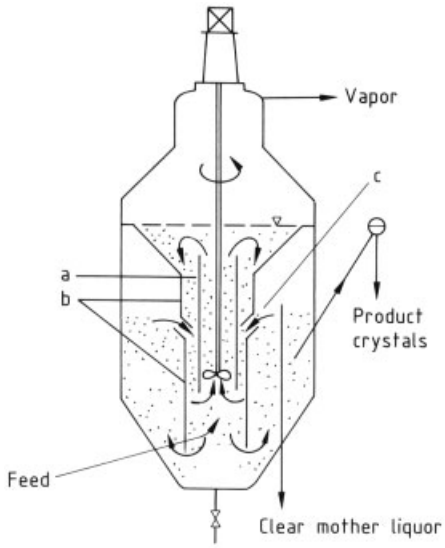

Figure 26. Standard–Messo turbulence crystallizer
a) Draft tube; b) Downcomer; c) Circumferential slot

introduced into the guide tube and passes into the vaporizer section where flash evaporation takes place. Nucleation, therefore, occurs in this region, and the nuclei are swept into the primary circuit. Mother liquor can be drawn off via a control valve, thus providing a means of controlling crystal slurry density.

The *Escher–Wyss Tsukishima double-propeller (DP) crystallizer* (Fig. 27) is essentially a draft-tube agitated crystallizer with some novel features. The DP unit contains an annular baffled zone and a double-propeller agitator which maintains a steady upward flow inside the draft tube and a downward flow in the annular region. Very stable suspension characteristics are claimed.

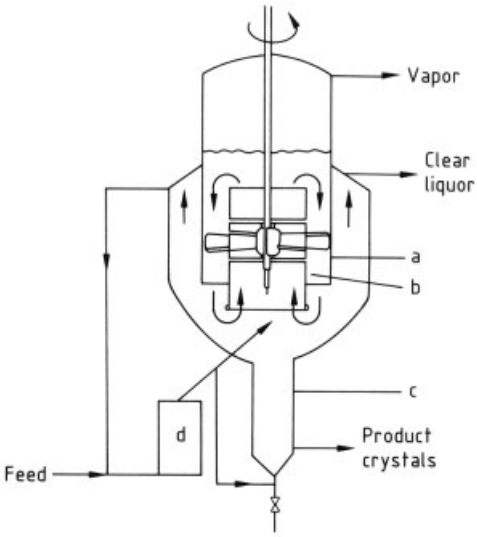

Figure 27. Escher–Wyss Tsukishima double-propeller (DP) crystallizer
a) Baffle; b) Draft tube; c) Elutriating leg; d) Heat exchanger

5.5. Crystal Yield

The crystal yield for simple cooling or evaporating crystallization can be estimated from the solubility characteristics of the solution. For aqueous solutions, the following general equation applies:

$$Y = \frac{WR\,[c_1 - c_2\,(1-V)]}{1-c_2\,(R-1)} \tag{25}$$

where c_1 is the initial solution concentration, kg anhydrous salt per kg water; c_2 is the final solution concentration, kg anhydrous salt per kg water; W is the initial mass of water, kg; V is the water lost by evaporation, kg per kg of original water present; R is the ratio of molecular masses of hydrated to anhydrous salts; and Y is the crystal yield, kg. The actual yield may differ slightly from that calculated by Equation (25). For example, if the crystals are washed with fresh solvent on the filter, losses may occur through dissolution. On the other hand, if mother liquor is retained by the crystals, an extra quantity of crystalline material will be deposited on drying. Furthermore, published solubility data usually refer to pure solvents and solutes. Because pure systems are rarely encountered industrially, solubilities should always be checked on the actual working liquors.

Before Equation (25) can be applied to vacuum (adiabatic cooling) crystallization, the quantity V must be estimated:

$$V = \frac{qR\,(c_1 - c_2) + C\,(t_1 - t_2)\,(1+c_1)\,[1-c_2\,(R-1)]}{\lambda\,[1-c_2\,(R-1)] - qRc_2} \tag{26}$$

where λ is the latent heat of evaporation of the solvent, J/kg; q is the heat of crystallization of the product, J/kg; t_1 is the initial temperature of the solution, °C; t_2 is the final temperature of the solution, °C; C is the specific heat capacity of the solution, J kg^{-1} K^{-1}; and c_1 and c_2 have the same meaning as in Equation (25).

5.6. Controlled Crystallization

Carefully selected seed crystals are sometimes added to a crystallizer to control the final product crystal size. The effect of rapid cooling on an unseeded solution is shown in Figure 28A. The solution cools at constant concentration until the limit of the metastable zone is reached, where nucleation occurs. The temperature increases slightly due to the release of latent heat of crystallization, but cooling reduces it and more nucleation occurs. The temperature and concentration subsequently fall as indicated. In such a process, nucleation and growth cannot be controlled.

Figure 28B demonstrates the slow cooling of a seeded solution in which temperature and solution composition are controlled within the metastable zone throughout the cooling cycle. Crystal growth occurs at a controlled rate only

on the added seeds; spontaneous nucleation is avoided because the system is never allowed to become labile. This batch operating method is known as *controlled crystallization*; many modern large-scale crystallizers operate on this principle.

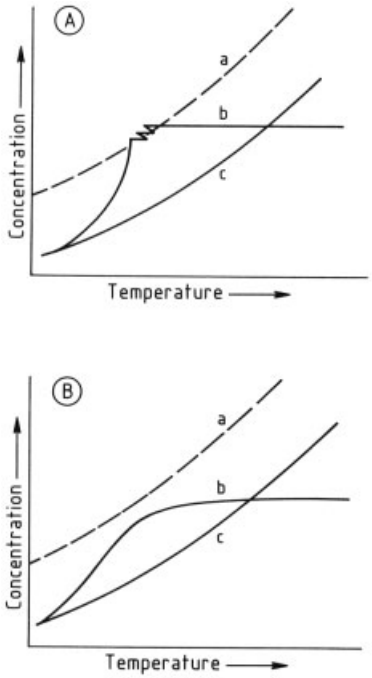

Figure 28. Effect of seeding on cooling crystallization
A) Rapid cooling of an unseeded solution; B) Slow cooling of a seeded solution
a) Supersolubility curve of the solute; b) Cooling curve of the solution; c) Solubility curve of the solute

If crystallization occurs only on the added seeds, the mass M_s of seeds of size L_s that can be added to a crystallizer depends on the required crystal yield Y (Eq. 26) and the product crystal size L_p:

$$M_s = Y L_s^3 / \left(L_p^3 - L_s^3\right) \tag{27}$$

The product crystal size from a batch crystallizer can also be controlled by adjusting cooling or evaporation rates. Natural cooling (Fig. 29A), for example, produces a supersaturation peak in the early stages of the process when rapid, uncontrolled heavy nucleation inevitably occurs. However, nucleation can be controlled within acceptable limits by following a cooling path that maintains a constant low level of supersaturation (Fig. 29B).

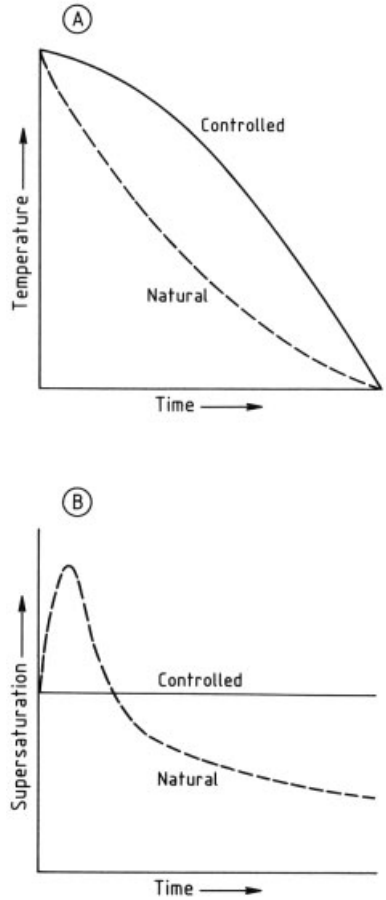

Figure 29. Natural and controlled cooling in batch crystallization
A) Temperature profile; B) Supersaturation profile

The calculation of optimum cooling curves for different operating conditions is complex [59], but the following simplified relationship is often adequate for general application:

$$\theta_t = \theta_0 - (\theta_0 - \theta_f)\,(t/\tau)^3 \tag{28}$$

where θ_0, θ_f, and θ_t are the temperatures at the beginning, end, and any time t during the process, respectively, and τ is the overall batch time. The potential benefits of controlled cooling are sometimes diminished by the occurrence of secondary nucleation, although a fines destruction loop may be installed on the crystallizer to combat this problem [7], [62].

5.7. Comparison of Batch and Continuous Crystallization

Although continuous, steady-state operation is often regarded as the ideal procedure for much processing plant equipment, this is not always true for crystallization processes. Batch operation often offers considerable advantages, such as simplicity of equipment and minimization of encrustation on heat-exchanger surfaces. In many cases, only a batch crystallizer can produce the required crystal form, size distribution, or purity. On the other hand, the operating costs of a batch system can be significantly higher than those of a comparable continuous unit, and problems of product variation from batch to batch may be encountered.

The particular attraction of a continuous crystallizer is its built-in flexibility for control of temperature, supersaturation, nucleation, crystal growth, and all the other parameters that influence crystal size distribution. A continuous crystallizer, however, does not discharge its product under equilibrium conditions (a batch unit can, if the batch time is adjusted appropriately) so the product slurry may have to be passed to a holdup tank to allow equilibrium to be reached. Omission of this step may cause problems due to further crystallization occurring in other parts of the plant (e.g., unwanted deposition in pipelines and effluent tanks). A holdup (aging) tank may also be necessary in a continuous system if the product exists in a phase (polymorph, hydrate, etc.) that differs from the one which appears initially.

A distinct advantage of batch crystallization, widely acknowledged in the pharmaceutical industry, is that the crystallizer can be cleaned thoroughly at the end of each batch to prevent contamination (seeding) of the next charge with any undesirable phase that might have arisen from transformation, rehydration, dehydration, air oxidation, etc., during the batch cycle. Continuous crystallization systems often self-seed undesirably after a certain operating time, which necessitates frequent shutdown and washout.

Semicontinuous crystallization processes often combine the best features of both batch and continuous operation. For example, a rapid mixer with a short residence time (possibly an on-line device) can discharge its product slurry into an agitated residence tank.

Continuous crystallizers may also be operated in a linked series of well-mixed vessels, with the crystal magma flowing from one stage to another. The most frequent objective of series or "cascade" operation is to economize on heat utilization, e.g., by dividing the overall temperature gradient over several stages and operating each successive stage at a lower temperature down the cascade. In cooling crystallization encrustation problems can be significantly reduced, because smaller temperature drops are needed across the individual heat exchangers. Accounts of cascade operation are given in [7], [36], [63], [64].

5.8. Crystallizer Modeling and Design

5.8.1. Population Balance

As described in Section 4.3, the processes of growth and nucleation interact in a crystallizer, and both contribute to the final crystal size distribution (CSD) of the product. In such assessments, the utility of the population balance [36] is widely acknowledged.

Application of the population balance is described most easily with reference to the simple, idealized case of an MSMPR (mixed-suspension mixed-product removal) crystallizer operated continuously in the steady state. The assumptions are made that no crystals are present in the feed stream, that all crystals are of the same shape, that crystals do not break down by attrition, and that crystal growth rate is independent of crystal size.

The relationship between crystal size L and population density n (number of crystals per unit size per unit volume of the system), derived directly from the population balance over the system [36], is

$$n = n^0 \exp\left(-L/G\tau\right) \tag{29}$$

where n^0 is the population density of nuclei (zero-sized crystals) and τ is the residence time. Equation (29) describes the crystal size distribution for steady-state operation. Rates of nucleation B and growth G ($= \mathrm{d}L/\mathrm{d}t$) are conventionally written in terms of supersaturation as

$$B = k_1 \Delta c^b \tag{30}$$

and

$$G = k_2 \Delta c^g \tag{31}$$

These empirical expressions can be combined to give

$$B = k_3 G^i \tag{32}$$

where

$$i = b/g \tag{33}$$

in which b and g are the kinetic orders of nucleation and growth, respectively, and i is the relative kinetic order. The relationship between nucleation and growth may be expressed as

$$B = n^0 G \tag{34}$$

or

$$n^0 = k_4 G^{i-1} \tag{35}$$

Experimental measurement of crystal size distribution (recorded on a number basis) in a steady-state MSMPR crystallizer can thus be used to quantify nucleation and growth rates. A plot of log n vs. L should give a straight line of slope $-(G\tau)^{-1}$ with an intercept at $L = 0$ equal to n^0 (Eq. 29 and Fig. 30A); if the residence time τ is known, the crystal growth rate G can be calculated. Similarly, a plot of log n^0 vs. log G should give a straight line of slope $i - 1$ (Eq. 35 and Fig. 30B); if the order g of the growth process is known (Eq. 33), the order of nucleation b can be calculated.

The total mass M_T of crystals in the system, the so-called magma density (mass of crystals per unit volume of the system), is given by

$$M_T = 6\alpha \varrho n^0 (G\tau)^4 \tag{36}$$

where α is the volume shape factor (Eq. 23) and ϱ is the crystal density. The peak of the mass distribution, the dominant size L_D of the CSD, is given by

$$L_D = 3G\tau \tag{37}$$

and can be related to the crystallization kinetics by

$$L_D \sim \tau^{(i-1)/(i-3)} \tag{38}$$

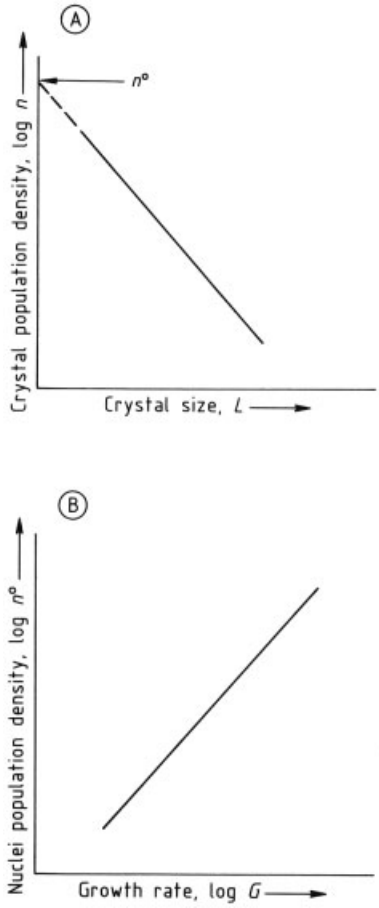

Figure 30. Population plots for a continuous mixed-suspension mixed-product removal (MSMPR) crystallizer
A) Crystal size distribution, slope $= -(G\tau)^{-1}$; B) Nucleation and growth kinetics, slope $= i - 1$
For explanation of symbols, see text.

This interesting relationship [36] enables the effect of changes in residence time to be evaluated. For example, if $i = 2$ (a typical value for many inorganic salt systems), a doubling of the residence time would increase the dominant product crystal size by only ca. 15 %. To double the residence time, however, either the crystallizer volume would have to be doubled or the volumetric feed rate, and hence the production rate, would have to be halved. Therefore, residence time adjustment is usually not a very effective means of controlling product crystal size. Population-balance-based CSD modeling can be applied to crystallizer configurations other than MSMPR [36] and has become a distinct, self-contained branch of reaction engineering. The

ability to simulate process configurations, however, presently exceeds the capabilities for measuring CSD on-line, predicting growth and nucleation rates, maintaining crystal magmas in a well-mixed state, and measuring supersaturation.

5.8.2. Design and Scaleup Problems

Industrial crystallizers are commonly designed by using data measured on laboratory-scale (1 – 10 L) or pilot-scale (50 – 250 L) units. In difficult cases, information may have to be obtained for both scales of operation. One of the main problems in crystallizer scaleup is characterization of particle-fluid hydrodynamics and assessment of their effects on the kinetics of nucleation and crystal growth.

In fluidized-bed crystallizers, for example, evaluation of the crystal suspension velocity is necessary. This parameter is related to crystal size, size distribution, and shape, as well as bed voidage and other system properties such as density differences and viscosity. Possible ways of estimating crystal suspension velocity are discussed in [7]. In practice, however, determination of suspension velocities on actual crystal samples by simple experimental techniques is often advisable.

In agitated vessels, the "just-suspended" agitator speed N_{JS} must be established, i.e., the minimum rotational speed necessary to keep all crystals in suspension. Not only do the crystals have to be kept in suspension, but the development of "dead spaces" in the vessel must also be avoided because they are unproductive zones and regions of high supersaturation in which vessel surfaces can become encrusted. Fluid and crystal properties, together with vessel and agitator geometries, are important in establishing N_{JS} values [7].

Agitated vessel crystallizers are often scaled up successfully on the crude basis of either constant power input per unit volume or constant agitator tip speed. A refinement of the latter criterion, for draft-tube agitated vessels, is to maintain the quantity $(\mathrm{TS})^2/(\mathrm{TO})$ constant, where TS is the agitator tip speed and TO is the turnover time, i.e., TO is the vessel volume divided by the volumetric circulation rate [65].

Crystallizer design procedures based on the population balance and other methods, together with practical examples, are described in [4], [7], [8], [36], [64].

6. Crystallization from Melts

Melt is the common name given to a liquid or a liquid mixture at a temperature near its freezing point. Melt crystallization is the process of separating the components of a liquid mixture by cooling until a quantity of crystallized solid is deposited from the liquid phase.

Melt crystallization is often considered to be commercially attractive, compared with distillation, for the separation of close-boiling organic substances. It offers the potential for low-energy separation because enthalpies of fusion are generally much lower than enthalpies of vaporization. A further advantage is that it operates at much lower temperatures than distillation, and this can be beneficial for processing thermally unstable substances. Technical limitations to the theoretical possibilities for melt crystallization are discussed in [66].

The basic requirement of melt crystallization is that the composition of the crystallized solid differ from that of the liquid mixture from which it is deposited. The ease or difficulty of separating one component from a multicomponent mixture by crystallization is best represented by a phase diagram as in Figures 3 and 4, both of which depict binary systems: the former shows a eutectic, and the latter shows a continuous series of solid solutions. These two systems behave quite differently on freezing; as described previously, a eutectic system can deposit a pure component (Section 3.2.1), whereas a solid solution can only deposit a mixture of components (Section 3.2.2).

Other more complex phase diagrams may be encountered in organic melts, but a survey [67] of several hundred binary organic mixtures reported that more than 50 % of these mixtures exhibited simple eutectic behavior, about 25 % formed intermolecular compounds (Fig. 5) and about 12 % formed solid solutions of one kind or another.

6.1. Single Stage Processes

Two basic techniques of melt crystallization are

1) Gradual deposition of a crystalline layer on a chilled surface in a static or laminar flowing melt
2) Fast crystallization of discrete crystals in the body of an agitated vessel.

An example of category 1 is found in the Proabd refiner [68] which is essentially a batch cooling process. A static liquid feedstock is progressively crystallized onto extensive cooling surfaces (e.g., fin-tube heat exchangers supplied with a cold heat-transfer fluid) located inside a crystallization tank. As crystallization proceeds, the remaining liquid becomes increasingly impure and, in some cases, crystallization may be continued until virtually the entire charge has solidified. The crystallized mass is then slowly melted by circulating a hot fluid through the heat exchanger. The impure fraction melts first and drains out of the tank. As melting proceeds, the melt runoff becomes progressively richer in the desired component, and fractions may be taken off during the melting stage if required. A typical flow diagram, based on a scheme for the purification of naphthalene, is shown in Figure 31. In this case, the circulating fluid is usually cold water which is heated during the melting stage by steam injection.

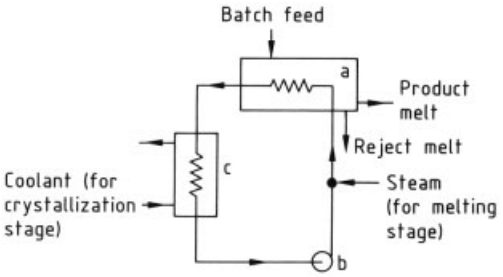

Figure 31. Batch cooling crystallization of melts: flow diagram for the Proabd refiner
a) Crystallizer; b) Pump; c) Heat exchanger

Another example in the first category is the rotary drum crystallizer which usually consists of a horizontally mounted cylinder that is partially immersed in the melt or supplied with feedstock insome other way. The coolant enters and leaves the inside of the hollow drum through trunnions. As the drum rotates, a crystalline layer forms on the cold surface and is subsequently removed with a scraper knife. Two feed and discharge arrangements are shown in Figure 32 (see next page). Rotary drum behavior and design have been discussed in [69], [70].

Analysis of general layer freezing processes shows that the structure and impurity levels of growing crystal layers are determined primarily by mass-transfer effects at the layer front [71]. Effective distribution coefficients described by a single parameter that combines growth velocity, mass-transfer coefficient, and concentration of mother liquor can be determined by simple laboratory tests and employed to facilitate process scaleup.

An example of melt crystallization in category 2 is the scraped-surface heat exchanger, which is basically a cylindrical tube surrounded by a cylindrical heat-exchange jacket. The tube is fitted with close-clearance scraper blades and rotates at relatively low speed. Two basic types are available: the large (>200 mm in diameter, >3 m long) slow-speed (<10 rpm) unit and the small (<150 mm in diameter, <1.5 m long) high-speed (>500 rpm) machine. Both types, but especially the latter, can handle viscous magmas, and operate at temperatures as low as $-80\,°C$. They are widely used, for example, in the manufacture of margarine (crystallization of triglycerides), dewaxing of lubricating oils (crystallization of higher *n*-alkanes), and large-scale processing of many organic substances (naphthalene, *p*-xylene, chlorobenzenes, etc.). The magma emerging from a scraped-surface crystallizer generally contains very small crystals (often $<10\,\mu m$) which can cause separation and subsequent reprocessing problems unless the crystals are first grown to a larger size, e.g., in a separate holdup tank.

6.2. Multistage Processes

The single-stage crystallization described in Section 6.1 might not always be sufficient to achieve the required purity of the final product, in which case further separation, melting, washing, or refining may be required. Two basic models can be considered:

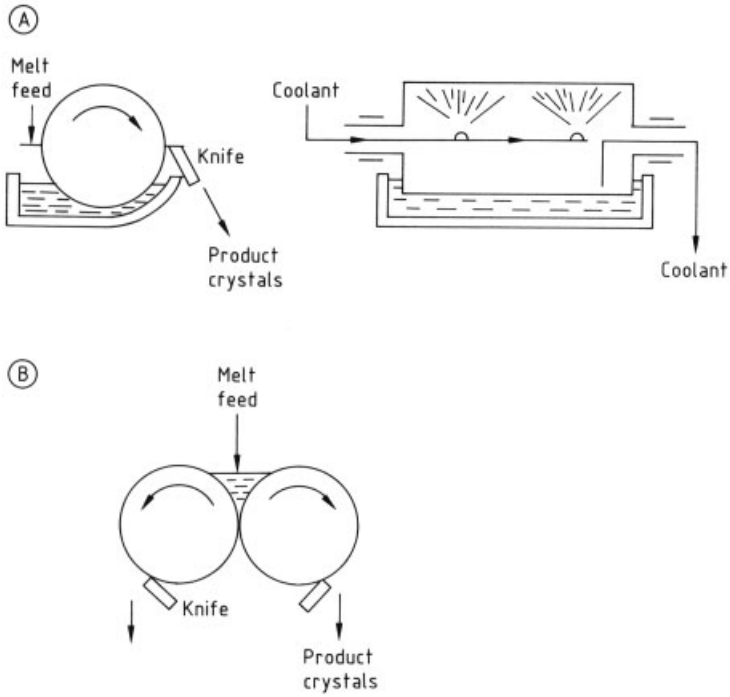

Figure 32. Two feeding modes (A) and (B) for drum crystallizers

1) a repeating sequence of crystallization, melting, and recrystallization steps (cf. Section 9.2)
2) a single crystallization step followed by countercurrent contacting of the crystals with a relatively pure liquid stream

Scheme 1 is preferred if the concentration of impurities in the feedstock is high; it is essential if the system forms a continuous series of solid solutions. Scheme 2 is applicable if the concentration of impurities is low. Some industrial operations, however, require a combination of both systems. For a comparative analysis of different types of multistage crystallization schemes, see [72].

Sulzer MWB Process. A highly successful industrial example of multistage operation is the Sulzer MWB process [73] that operates by crystallization on a cold surface; it also has features that permit it to operate effectively as a multistage separation device. Consequently, it can be used to purify solid solutions. An effective multistage countercurrent scheme is illustrated for four-stage operation in Figure 33A. Stage 1 is fed with melt L_2 and recycle liquor $L_1 - L$ (L denotes the impure reject liquor stream). Crystals C_1 are deposited in stage 1, and after melting, they are mixed with melt L_3 and fresh feedstock. Crystallization of this mixture yields crystals C_2 and melt L_2 in stage 2. Similar patterns are followed in stages 3 and 4, and the final high-purity stream C_4 is remelted and split into product C and recycle melt $C_4 - C$. Only one crystallizer, a vertical multitube heat exchanger, is required in this scheme. The crystals do not have to be transported; they remain deposited on the internal heat-exchange surfaces in the vessel, until they are melted for further processing. The intermediate storage tanks and crystallizer are linked by a control system consisting of a program timer, actuating valves, pumps, and the cooling loop (Fig. 33B). The process has been used on a large scale in the purification of a range of organic substances (e.g., chloro- and chloronitrobenzenes, nitrotoluenes, cresols, and xylenols) and in the separation of fatty acids.

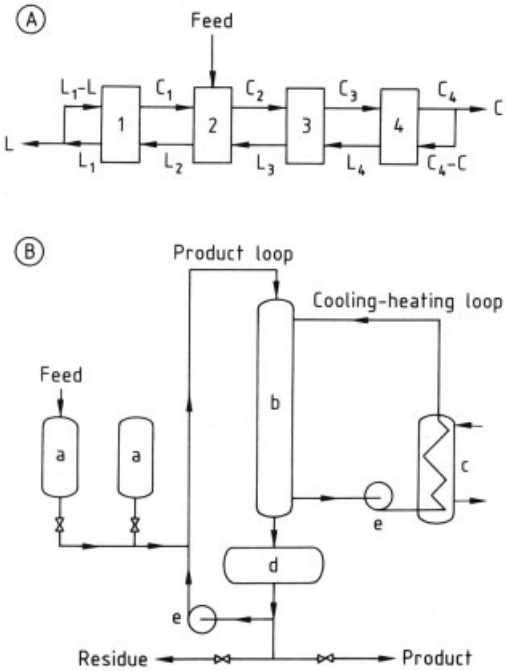

Figure 33. Sulzer MWB process
A) Multistage flow diagram
Symbols: C = crystals, L = liquor
B) Schematic of practical layout
a) Residue melt storage tanks; b) Crystallizer; c) Heat exchanger; d) Collecting tank; e) Pump

6.3. Column Crystallizers

Because melt feedstock components can form both eutectic and solid-solution systems with one another, sequences of washing, partial or complete melting, and recrystallization are often necessary to produce one of the components in near-pure form. However, the operation of a sequence of melt crystallization steps can be time-consuming and costly, especially if the liquid feedstock has to be cooled until it crystallizes and the crystals have to be separated from the residual melt, washed, and then remelted before the cycle can be repeated. Many attempts have been made to effect some of these events in a single unit, such as the column crystallizer developed by SCHILDKNECHT in the late 1950s.

The basic features of a *Schildknecht column* are shown in Figure 34. Liquid feedstock enters the column continuously at an arbitrary midpoint. Freezing at the bottom of the column is accomplished with a suitable refrigerant fluid, melting at the top, with a hot fluid or an electrical heating element. The crystals and liquid pass through the column countercurrently, the solid phase being transported downward by a helical conveyor fixed on a central shaft. The purification zone is usually operated at a virtually constant temperature between those of the freezing and melting sections. Crystals are formed mainly in the freezing section, but they can also deposit on the inner surface of the column from which the helical conveyor removes them by scraping. During conveyance, crystals come into contact with the counterflowing liquid melt and are thus subjected to surface washing. The reverse mode of operation has also been used, i.e., upwardly flowing liquid in contact with crystals being conveyed downward. In this case, the locations of the freezer and melter are the reverse of those depicted in Figure 34. Comprehensive studies of the modeling of column crystallizers appear in [74], [75].

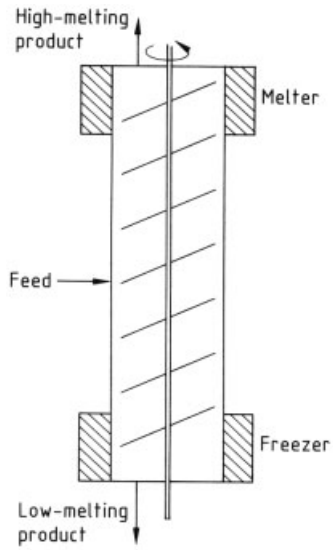

Figure 34. Schildknecht column

Although successful, the Schildknecht column is basically a laboratory apparatus; no large-scale industrial applications have been reported. A melt crystallizer of the wash column type, however, was developed by Phillips Petroleum Company in the 1960s for large-scale production of *p*-xylene [76]. The key features of the *Phillips pulsed-column crystallizer* are shown

in Figure 35. A cold slurry feed, produced in a scraped-surface chiller, enters at the top of the column. Crystals in the vertical bed are pulsed downward by a piston, and impure mother liquor leaves through a wall filter. The upwardly flowing wash liquor is generated by the bottom heater which melts pure crystals before they are removed from the column.

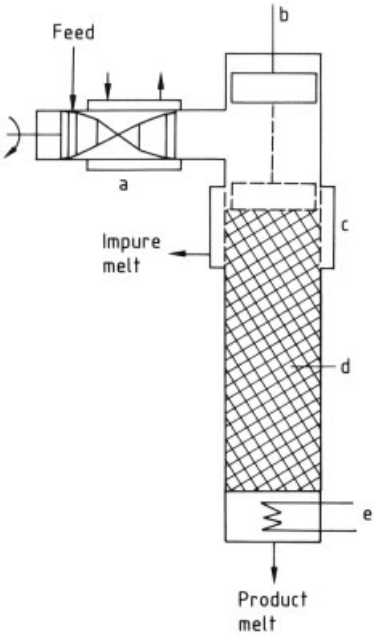

Figure 35. Phillips pulsed-column crystallizer
a) Scraped-surface chiller; b) Piston; c) Wall filter; d) Crystal bed; e) Heater

The *Brodie purifier* [77] developed in the late 1960s has several features of the column crystallizers described above, but it also has the potential to deal effectively with solid-solution systems. It is essentially a center-fed column which can convey crystals from one end to the other (Fig. 36). As the crystals are conveyed through the unit, their temperature is gradually increased by a deliberately imposed temperature gradient along the flow path; they are thus subjected to partial melting which encourages the release of low-melting impurities. The interconnected scraped-surface heat exchangers are of progressively smaller diameter to maintain reasonably constant axial flow velocities and prevent backmixing. The vertical purifying column acts as a countercurrent washer in which falling, near-pure crystals meet an upflow of pure melt. The Brodie purifier has been used in the large-scale production of high-purity 1,4-dichlorobenzene and naphthalene.

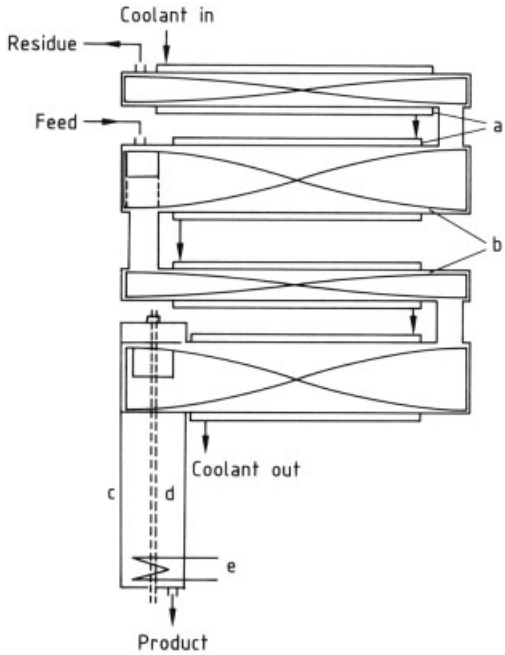

Figure 36. Brodie purifier
a) Cooling jackets; b) Scraper conveyors; c) Purifying column; d) Slow agitator; e) Heater

The *Tsukishima Kikai* (*TSK*) *countercurrent cooling crystallization process* [78] is, in effect, a development of Brodie technology. The flow sheet in Figure 37 (see next page) shows three conventional cooling crystallizers connected in series. Feed enters the first-stage vessel and partially crystallizes. The slurry is then concentrated in a hydrocyclone before passing into a Brodie purifying column (Fig. 36). After passage through a settling zone in the crystallizer, clear liquid overflows to the next stage. Slurry pumping and overflow of clear liquid in each stage result in a countercurrent flow of liquid and solid. The process has been applied in the large-scale production of *p*-xylene.

6.4. High-Pressure Crystallization

Pressure, as well as temperature, can alter the equilibrium relationships in a crystal – melt system. The potential of high-pressure crystallization as a purification technique has been demon-

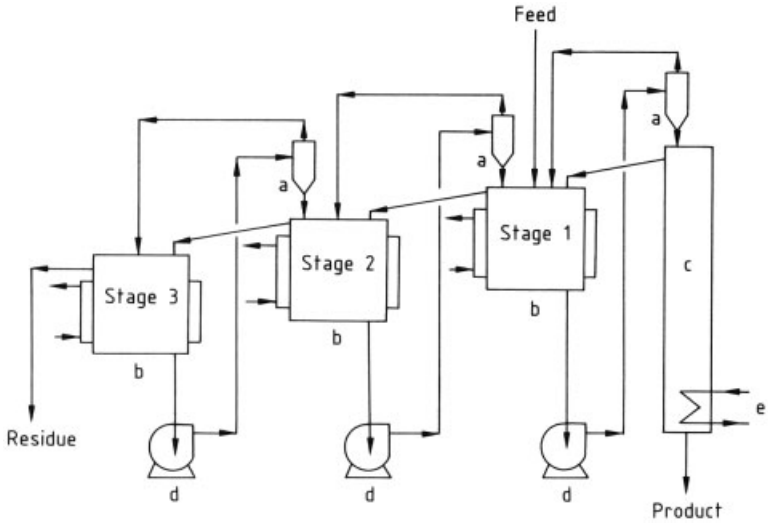

Figure 37. Tsukishima Kikai (TSK) countercurrent cooling crystallization process
a) Hydrocyclone; b) Conventional crystallizer; c) Brodie purifying column; d) Pump; e) Heater

strated with impure organic melt feedstocks subjected to pressures up to 300 MPa under adiabatic conditions. As the pressure and temperature of the charge increase, fractional crystallization ensues and the impurities are concentrated in the liquid phase which is then discharged from the pressure chamber. At the end of the cycle, further purification is possible because residual impurities in the compressed crystalline plug may be "sweated out" when the pressure is released. So far only relatively small (≈ 2 L) pressure chambers have been employed, but a single-cycle operation lasting less than 5 min is claimed to be sufficient to effect substantial purification in a wide range of organic binary melt systems [79].

6.5. Prilling and Granulation

Prilling is a melt spray crystallization process that results in the formation of solid spherical granules. It is employed widely in the manufacture of fertilizer chemicals such as ammonium nitrate and urea. In the ammonium nitrate prilling process [80], a very concentrated solution, containing ca. 5 % water, is sprayed at 140 °C into the top of a 30-m-high, 6-m-diameter tower in which the droplets fall countercurrently to an upwardly flowing air stream that enters the base of the tower at 20 °C. The solidified droplets (prills), which leave the tower at 80 °C, contain ca. 4 % water and must be dried to an acceptable moisture content at < 80 °C to prevent the occurrence of polymorphic transitions (see Section 3.1).

In a development of the melt granulation technique for urea [81], molten urea is sprayed at 148 °C onto cascading granules in a rotary drum and seed granules (< 0.5 mm) are thereby built up to product size (2 – 3 mm). Heat released by the solidifying melt is removed primarily by evaporation of a fine mist of water sprayed into air that is passed through the granulation drum. Problems associated with the design of prilling towers are discussed in [82], [83].

7. Crystallization from Vapors

In the rapidly expanding field of single-crystal growth, a tendency exists nowadays to classify the various processes of growth from the vapor according to the manner in which vapor is generated [84], e.g., sublimation, chemical vapor transport (CVT), chemical vapor deposition (CVD) (→ Crystal Growth, Chap. 8.).

Sublimation processes are characterized by vapor production resulting from heating the solid phase and subsequent crystallization of the sublimate by condensation under conditions below the triple point. Substances such as

zinc selenide and gallium arsenide are grown by this technique. Sublimation – desublimation processes are also used on a large scale by the chemical industry to produce a wide range of organic and inorganic substances (→ Sublimation).

8. Precipitation

Precipitation is widely used in the laboratory for chemical analysis and in industry for the manufacture of paints, pigments, pharmaceutical and photographic chemicals, etc. In the production of ultrafine crystalline powders, precipitation is often considered an attractive alternative to comminution, particularly for heat-labile substances (i.e., substances that are unstable when heated).

No generally accepted, unambiguous definition of the term precipitation exists; it may refer simply to very fast crystallization, although precipitation is often an irreversible process, i.e., many precipitates are virtually insoluble substances produced by a chemical reaction. The products of conventional crystallization, on the other hand, can often be redissolved when the original conditions of temperature and concentration are restored. Nevertheless, precipitation and crystallization have much in common and are governed by the same laws.

8.1. Solubility Products

The solubility of a sparingly soluble electrolyte in water may be expressed in terms of the concentration solubility product K_c. For example, if such an electrolyte dissociates in solution into x cations and y anions according to

$$\mathrm{M}_x\mathrm{A}_Y \rightleftharpoons x\,\mathrm{M}^{z+} + x\,\mathrm{A}^{z-} \tag{39}$$

where z^+ and z^- are the valences of the metal cation M and the anion A, respectively, then for a saturated solution

$$(c_+)^x\,(c_-)^y = \text{constant} = K_c \tag{40}$$

where c_+ and c_- are the ionic concentrations. For 1 – 1, 2 – 2, etc., electrolytes (i.e., $x = y = 1$, $c_+ = c_- = c*$, the equilibrium solubility), Equation (40) becomes

$$c^* = (K_c)^{1/2} \tag{41}$$

In general,

$$c^* = (K_c/x^x y^y)^{1/(x+y)} \tag{42}$$

For a 2 – 1 electrolyte, therefore,

$$c^* = (K_c/4)^{1/3} \tag{43}$$

A more fundamental approach to the solubility product involves the use of activities rather than concentrations; the activity solubility product K_a is then defined by

$$(a_+)^x\,(a_-)^y = \text{constant} = K_a \tag{44}$$

where a_+ and a_- are the ionic activities, or by

$$(c_+\gamma_+)^x\,(c_-\gamma_-)^y = K_a \tag{45}$$

where γ is the ionic activity coefficient. Therefore,

$$K_a = K_c\,(\gamma_\pm)^\nu \tag{46}$$

where $\gamma_\pm$ is the mean ionic activity coefficient and ν $(= x + y)$ is the number of moles of ions produced by one mole of electrolyte. In practice, K_a and K_c may be assumed equal for concentrations up to about 10^{-3} mol/L, but above this, significant differences can occur. The activity of an ion depends on the concentration of all the other ions in solution, so the presence of a foreign electrolyte can greatly influence the value for $\gamma_\pm$ of a sparingly soluble salt.

A number of cases that appear anomalous when the simple solubility product is used can be explained when activity coefficients are considered [85]. For instance, the addition of a common ion generally decreases the solubility of a given salt, but cases are known in which addition of a common ion increases the solubility of the salt. This is because a large increase in ionic concentration can bring about a reduction in the activity coefficients. Thus, from Equation (45), an increase in c_- will result in a decrease in c_+ (i.e., precipitation of the sparingly soluble salt) if γ_+ and γ_- remain fairly constant; however, an increase in c_- to a value that reduces both γ_+ and γ_- must result in an increase in c_+ if K_a is to remain constant. The addition of a salt without a common ion often increases the solubility; this again occurs because the increased ionic concentration reduces the activity coefficients.

8.2. Ostwald's Rule of Stages

Ostwald's rule of stages, or rule of successive transformations, was originally stated as follows [86]: "An unstable chemical system does not spontaneously transform directly into that state which, under the given conditions, is the most stable of all the possible states, but into that which most closely resembles its own, i.e., into the state whose formation from the original is accompanied by the smallest loss in free energy."

Although many exceptions to Ostwald's rule have been recorded and little theoretical support has been found, it provides a useful guide to the possible behavior of precipitating systems. The rule appears, in fact, to be a manifestation of the frequently noted behavior of chemical systems that, if more than one reaction is thermodynamically possible, the resulting reaction is not the one which is thermodynamically most likely, but the one with the fastest rate. In other words, kinetics are often more important than thermodynamics [7].

Therefore, a metastable solid phase may be precipitated first and then, at some later stage, transformed into a more stable phase. Transformations such as those of one polymorph to another, of one hydrate to another hydrate or to an anhydrous form, and of an amorphous precipitate to a crystalline phase are quite common (Section 3.5).

8.3. Development of Precipitates

Precipitation, like all crystallization processes, consists of three basic steps: (1) the creation of supersaturation, followed by (2) the generation of nuclei and (3) the subsequent growth of these nuclei to detectable size. The kinetics of nucleation and growth are described in Sections 4.1 and 4.2, respectively. The agglomeration of small crystals into clusters and changes in size distribution caused by ripening can also play important roles in the development of a precipitate.

8.3.1. Ripening

When solid particles are dispersed in their own saturated solution, the smaller particles tend to dissolve and the resulting solute is then deposited on the larger particles. Thus, the small particles disappear, the large ones grow larger, and theoretically, the particle-size distribution should ultimately become monodisperse. The reason for this behavior is that the solid phase in the system adjusts itself so as to achieve a minimum total surface free energy. This process of particle coarsening is called ripening, or more frequently "Ostwald Ripening" after the first proposer of the mechanism [86].

The driving force for ripening is the difference in solubility between small and large particles, as given by the size–solubility (Gibbs–Thomson–Ostwald–Freundlich) relationship (Eq. 15), although the effect only becomes significant for particle sizes $< 1\,\mu\text{m}$ (see Section 2.2). If mass transport occurs between the particles in a polydisperse precipitate and if the growth kinetics are diffusion controlled, all particles of size

$$r\prime = 2v\gamma c^*/\nu RT\,(c - c^*) \tag{47}$$

are in equilibrium with the bulk solution ($\mathrm{d}r/\mathrm{d}t = 0$) [7]; v is molar volume, γ interfacial tension and ν the number of ions. All particles smaller than r' will dissolve ($\mathrm{d}r/\mathrm{d}t < 0$) and all particles larger than r' will grow ($\mathrm{d}r/\mathrm{d}t > 0$). The speed at which ripening occurs depends to a large extent on particle size r and solubility. For diffusion-controlled growth kinetics, the linear growth velocity may be expressed approximately as

$$dr/\mathrm{d}t \simeq \gamma \mathrm{v}^2 Dc^*/3\nu RTr^2 \tag{48}$$

where D is the diffusion coefficient. However, because ripening generally occurs at very low supersaturation, it is more likely to be controlled by surface reaction than by diffusion; under these circumstances, ripening could be considerably slower than indicated by Equation (48). Analyses of ripening mechanisms have been made in [5], [87], [88].

Ripening changes the particle-size distribution of a precipitate over a period of time, even in an isothermal system, but the change can be accelerated by controlled temperature fluctuation. This process, known as temperature cycling, has been utilized to alter the physical characteristics of precipitates [5], [7].

8.3.2. Agglomeration

Small particles in liquid suspension tend to agglomerate into clusters. A theoretical basis for the rate of agglomeration of colloidal particles in suspension was first proposed by SMOLUCHOWSKI [89]. Two types of behavior may be distinguished: *perikinetic* (static fluid with particles in Brownian motion) and *orthokinetic* (agitated dispersions), which can both be important in precipitation processes. In agitated precipitators, orthokinetic agglomeration becomes more important as the particle size and the shear rate increases.

Double-layer repulsion forces and van der Waals attraction forces operate independently in disperse systems. Repulsion forces decrease exponentially over a distance corresponding to the thickness of the ionic double layer, whereas attraction forces decrease, over a larger distance from the particle surface, as an inverse power of the distance. Consequently, attraction normally predominates at very small and very large distances, and repulsion over intermediate distances [90].

The assessment and modeling of agglomeration kinetics in precipitation processes are discussed in [5], [91], [92].

8.3.3. Precipitate Morphology

The morphological development of a precipitate is a complex combination of a variety of processes including nucleation, habit modification, phase transformation, ripening, and agglomeration. The most influential system parameters are supersaturation and the concentration of active impurities, although pH can also exert a profound effect in some aqueous systems.

The dominant influence of supersaturation on the particle-size characteristics of a precipitate has been summed up in the so-called Weimarn laws of precipitation [93] which, while open to theoretical criticism, provide very useful guidelines for batch precipitation behavior. They are illustrated in Figure 38:

1) As the concentration of reactants increases, the median particle size of the precipitate (determined at a given time after mixing the reactants) increases to a maximum and then decreases. As the time interval increases, the maximum is displaced to lower initial supersaturation and higher median particle size.
2) For a completed precipitation, the median size of the precipitate crystals decreases as the initially created supersaturation S is increased.

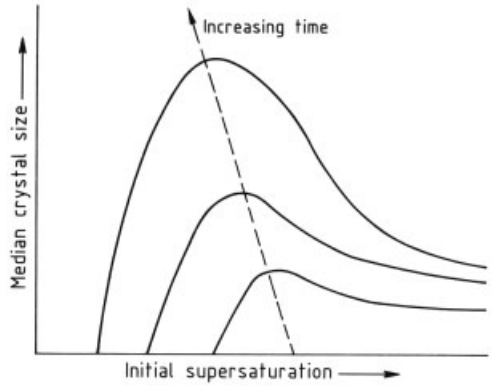

Figure 38. Illustration of Weimarn's laws of precipitation

For a given system, maximum agglomeration often appears to occur at a certain level of supersaturation; this behavior has been linked to the character of the adsorption layer surrounding a growing crystal which consists of loosely bonded, partially integrated groups of the crystallizing species [7].

An oversimplified but graphic example of the interactive effects of supersaturation, nucleation, and growth in the development of precipitated particles is given by WALTON [94], who considered the homogeneous nucleation of three different systems at an arbitrary value of supersaturation $S = 100$ (Eq. 2). If the number of particles nucleated is $10^6/\text{cm}^3$, the maximum precipitated particle sizes expected for three different solubilities $c*$ of 10^{-7}, 10^{-4}, and 10^{-1} mol/L, were concluded to be 1, 10, and 100 µm, respectively.

Alternatively, the approximate particle size can be estimated if the above three solutions are all assumed to have the same concentration, e.g., 1 mol/L. The first ($S = 10^7$) would nucleate homogeneously, forming approximately 1-nm particles and producing a colloidal system (gel) which could remain stable for long periods before the primary particles agglomerated. The second ($S = 10^4$) would also nucleate homogeneously, forming primary particles

around 0.1 µm which would agglomerate easily and develop rapidly into a conventional precipitate. The third solution ($S = 10$) would probably nucleate heterogeneously, yielding approximately 10-µm crystals which could, under favorable conditions, remain discretely dispersed.

8.3.4. Coprecipitation

All precipitates are contaminated to some extent with materials originally present in the mother liquor. The general term coprecipitation may be used to describe the many different types of impurity incorporation that can occur, including surface adsorption and lattice entrapment of foreign ions and solvent molecules, as well as physical inclusion of pockets of mother liquor.

The adsorption of salts having an ion in common with the precipitate roughly obeys the Paneth–Fajans–Hahn adsorption rule which postulates that the less soluble the salt, the more easily is it incorporated into a precipitate. For example, barium chloride is more readily adsorbed by barium sulfate than is barium iodide, which is more soluble than the chloride. The dissociability of the adsorbed salt is also important; adsorption decreases as the degree of dissociation of the adsorbed salt increases.

The distribution of an impurity between solid (i.e., solid solution) and liquid phases may be represented by the Chlopin [95] equation:

$$\frac{x}{y} = D\left(\frac{a-x}{b-y}\right) \tag{49}$$

where a and b are the amounts of components A and B in the original solid, x and y are the amounts of A and B in the crystallized solid, and $a - x$ and $b - y$ are the amounts of A and B retained in the solution. D is a distribution coefficient. Alternatively, the logarithmic Doerner–Hoskins [96] equation may be used:

$$\ln(a/x) = \lambda \ln(b/y) \tag{50}$$

The constant λ has been called a heterogeneous distribution coefficient to distinguish it from the homogeneous distribution coefficient D in Equation (49). Under ideal conditions $D = \lambda$.

If component A is the impurity, $\lambda > 1$ indicates that the impurity will be enriched in the precipitate; conversely, if $\lambda < 1$ it will be depleted. The effect of precipitation rate on λ is shown in Figure 39 [94]. In enrichment systems, $\lambda \to \lambda_e = D_e$ as the precipitation rate tends to zero. For fast rates of precipitation $\lambda \to 1$. In depletion systems, an analogous situation exists, with $\lambda \to \lambda_d = D_d$ for very low precipitation rates and $\lambda \to 1$ for rapid precipitation.

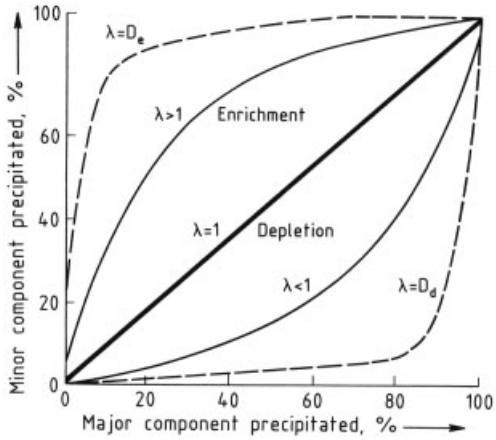

Figure 39. Effect of precipitation rate on the heterogeneous distribution coefficient λ [94]
Broken lines denote very slow precipitation; thickened line, very rapid precipitation.

Both the Chlopin and the Doerner–Hoskins relationships have been widely used to correlate the results of fractional precipitation schemes (see Chap. 9) although neither is entirely satisfactory from a theoretical point of view.

8.4. Precipitation Techniques

8.4.1. Reaction Precipitation

A common method for producing a precipitate is to mix two reacting solutions together quickly to create a highly supersaturated system. One practical difficulty, however, is to maintain reasonably uniform conditions throughout the reaction vessel. The choice of method used to mix the reactants is, therefore, very important because zones of excessive supersaturation should not be allowed to develop. The sequence of reactant mixing can also be of critical importance: addition of A to B to produce a precipitate C often yields a very different product from the addition of B to A. Factors such as the development of local pockets of reactants in nonstoichiometric

ratios and undesirable pH levels can have highly detrimental effects on precipitation.

Primary nucleation does not necessarily commence as soon as the reactants are mixed, even when the level of supersaturation is very high. The mixing stage is followed by a time lag (induction period) before nuclei appear, which depends on the temperature, supersaturation, efficiency of mixing, state of agitation, and presence of impurities (see Section 4.1.4). After the induction period, rapid desupersaturation ensues (Fig. 40), during which primary (homogeneous and heterogeneous) and secondary nucleation may occur together. The predominant process at this stage, however, is growth of the nuclei. Later, if sufficient time is allowed to elapse, particle coarsening may occur as a result of ripening or agglomeration.

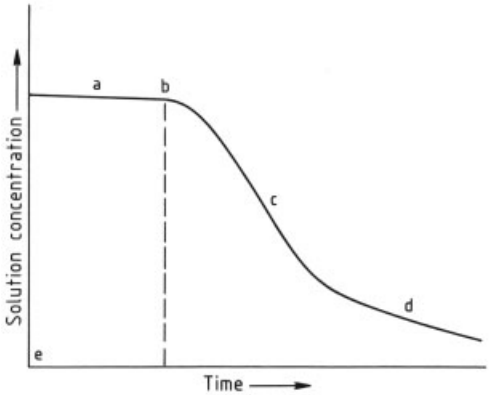

Figure 40. A typical desupersaturation curve
a) Development of nuclei; b) Induction period; c) Growth of crystals; d) Ripening of precipitate; e) Equilibrium saturation concentration

Precipitation of solid particles resulting from a chemical reaction between gases or liquids is a standard method for the preparation of many industrial chemicals. Precipitation occurs because the gaseous or liquid phase becomes highly supersaturated with respect to the solid phase. A crude precipitation operation, therefore, can be transformed into a crystallization process by careful control of the degree of supersaturation.

Reaction precipitation – crystallization is used widely in industries where waste gases containing valuable or noxious substances are produced. For instance, ammonia can be recovered from coke-oven gases by converting it into ammonium sulfate via reaction with sulfuric acid. In this case, agitation within the vessel is effected by a combination of the vigorous nature of the exothermic reaction and air sparging. The heat of reaction is removed by the evaporation of water added to the reaction zone [1].

Problems of precipitation process plant designs are discussed in [5].

Precipitation from Homogeneous Solution. For the purpose of gravimetric analysis, when a solid must be efficiently separated from a liquid, precipitation is generally carried out slowly from dilute solution. However, substances such as the hydroxides and basic salts of aluminum, iron, and tin demand extremely high dilution and excessively long time for coarse filterable particles to be produced. Precipitation from homogeneous solution (PFHS) offers a useful way of overcoming these difficulties [97].

Briefly, the technique consists of slowly generating the precipitating agent homogeneously within a well-mixed solution by means of a chemical reaction. Undesirable concentration effects are thus eliminated, a dense granular precipitate is formed, and coprecipitation is minimized. For example, silver chloride crystals can be produced from an aqueous solution of silver nitrate by reaction with allyl chloride (3-chloropropene):

$$CH_2{=}CHCH_2Cl + H_2O \longrightarrow Cl^- + CH_2{=}CHCH_2OH + H^+$$
$$Cl^- + Ag^+ \longrightarrow AgCl$$

An example of a PFHS reaction in nonaqueous solution is the precipitation of silver iodide in ethanol:

$$2\,C_2H_5I + 2\,AgNO_3 + C_2H_5OH \longrightarrow$$
$$2\,AgI + (C_2H_5)_2O + HNO_3 + C_2H_5NO_3$$

Other PFHS methods used to produce crystalline precipitates by controlled generation of the required anions in an appropriate aqueous solution include the hydrolysis of dimethyl oxalate ($C_2O_4^{2-}$), triethyl phosphate (PO_4^{3-}), dimethyl sulfate (SO_4^{2-}), and thioacetamide (S^{2-}).

Precipitation from homogeneous solution plays an important role in modern analytical chemistry. It is also used to investigate coprecipitation and nucleation because the slow, controlled precipitation allows a close approach to equilibrium between the solid and the solution.

Industrial applications of PFHS have so far been limited, but it appears to be a promising technique. It generally improves fractional precipitation methods and has been applied to the difficult separation of radium and barium used in the production of carriers for radioactive materials. A further use is in the preparation of monodisperse suspensions of pigments and polishing agents [97], [98].

8.4.2. Salting Out

A solution can be made supersaturated with respect to a given solute by addition of a substance generally referred to as the precipitant, which reduces the solubility of the solute in the solvent. The precipitant may be a liquid, solid, or gas. This operation is known by a variety of terms, salting out being the most common. The term watering out is used in the pharmaceutical industry to describe the precipitation – crystallization of organic substances from water-miscible organic solvents by the controlled addition of water. The term solventing out has been applied to the use of water-miscible organic solvents to precipitate electrolytes from aqueous solution [99].

The properties required of a precipitant are that it be miscible with the solvent of the original solution, that the solute be relatively insoluble in it, and that the final solvent–precipitant mixture be easily separable if it contains valuable components. The beneficial effects of using a precipitant prediluted with the system solvent to avoid excessive nucleation and encourage the development of larger precipitated particles are described in [100].

Salting out has many advantages. For example, highly concentrated initial solutions can be made by dissolving an impure crystalline material in a suitable solvent. If the solute is very soluble in the chosen solvent, dissolution may be effected at low temperature, which is advantageous when heat-labile substances are processed. Solute recovery is usually high. Purification is often better than in straightforward crystallization because the mixed mother liquor often retains more undesirable impurities than the original solvent does. On the other hand, salting out has the disadvantage that a recovery unit may be needed to handle fairly large quantities of mother liquor in order to separate valuable solvents and precipitants.

Several potential large-scale applications of salting out have been reported [7]. Examples are the production of pure inorganic salts with liquid organic precipitants (particularly anhydrous salts from aqueous solution at ambient temperature when a hydrated species is the thermodynamically stable phase [101]), the treatment of seawater with alcohols to recover fertilizer-grade double salts (e.g., hydrated potassium magnesium sulfate) [102], and the separation of inorganic salt mixtures [99], [103].

Gases or solids may be used as precipitants provided they are soluble in the original solvent and do not react with the solute to be precipitated. Ammonia can assist in the production of potassium sulfate by the reaction of calcium sulfate and potassium chloride: in a pure aqueous medium the yield is low, but in aqueous ammonia the yield is greatly improved. Hydrazine acts in a similar manner [104].

An example of the use of a solid precipitant is the addition of sodium chloride crystals to salt out organic dyes from aqueous solution. The sodium chloride acts in the solution phase, i.e., it must dissolve in the water present before it can act as a precipitant; its precise mode of action is probably quite complex.

Crystalline salts can be added to solutions to precipitate other salts, for example, as a result of the formation of a stable salt pair. This behavior is encountered when two solutes AX and BY, usually without a common ion, react in solution and undergo double decomposition (*metathesis*).

$$AX + BY \rightleftharpoons AY + BX$$

The four salts AX, BY, AY, and BX constitute a reciprocal salt pair. One of these pairs, AX – BY or AY – BX, is stable and is composed of compatible salts which can coexist in solution; the other salt pair is unstable and composed of incompatible salts which cannot coexist. This principle is used in the large-scale production of potassium sulfate by addition of solid glaserite ($3\,K_2SO_4 \cdot Na_2SO_4$) to a concentrated aqueous solution of potassium chloride [105]. Conversion to a stable salt pair occurs; the Na_2SO_4 and KCl remain in solution and K_2SO_4 precipitates.

8.5. Precipitation Methods and Equipment

Although continuous, steady-state operation is often regarded as ideal for process plant equipment, this is not always true for crystallization [7] and perhaps even less so for precipitation. The advantages and disadvantages discussed in Section 5.7 also apply to precipitation.

Most industrial precipitation units are constructed simply. The main aim is usually to mix reacting fluids rapidly and allow the development of a precipitate with certain desirable physical and chemical characteristics. Good mixing is essential to smooth-out supersaturation peaks in local regions. Both micromixing and macromixing are involved. *Micromixing* is concerned with mixing at or near the molecular level, and is influenced by fluid physical properties and local conditions. *Macromixing*, which is concerned with bulk fluid movement and blending, is influenced by agitator speed, vessel geometry, etc. The two types of mixing should strictly be considered independently, but the overall effects of mixing on the precipitation process are extremely complex [5], [101], [102].

The position of the feedstock entry point(s) can have a great influence on the precipitate quality [7]. Several choices are available. For example, if two reactant feedstocks A and B are involved, A could be fed on or near the surface of reactant B already in the vessel. This is the so-called "single-jet" mode (Fig. 41A). Alternatively, A could be introduced into the intensely agitated zone near the impeller blades (Fig. 41B). The latter procedure often results in the production of larger primary crystals since good mixing keeps local levels of supersaturation low at the first point of contact between the reactants and minimizes the nucleation rate. The sequence of reactant additions can often have a significant effect on the characteristics of the precipitate produced, i.e., reactant B could be introduced as a single jet into reactant A charged first into the vessel, if required.

Both reactant streams A and B may be introduced to the vessel simultaneously, in the so-called "double-jet" mode, and again several choices emerge. For example, A and B could be introduced together near the surface (Fig. 41C), or near the impeller blades (not illustrated). On the other hand, the two streams could be premixed before entering the vessel as a single jet (Fig. 41D) at some appropriate point. Premixing of feedstocks is often used because it can provide a means of exerting some control over the initial supersaturation levels. Impinging jets or in-line mixers may be used for rapid premixing.

9. Fractional Crystallization

A single crystallization operation performed on a solution or melt can fail to produce a pure crystalline product for a variety of reasons. For example, the impurity may have solubility characteristics similar to those of the desired pure component, and both substances consequently cocrystallize. Alternatively, the impurity may be present in such large amounts that the crystals inevitably become contaminated. Furthermore, a pure substance cannot be produced in a single crystallization stage if the impurity and the required substance form a solid solution (see Section 3.2.2). Recrystallization (i.e., repeated crystallization steps) from a solution or melt is, therefore, widely employed to increase crystal purity.

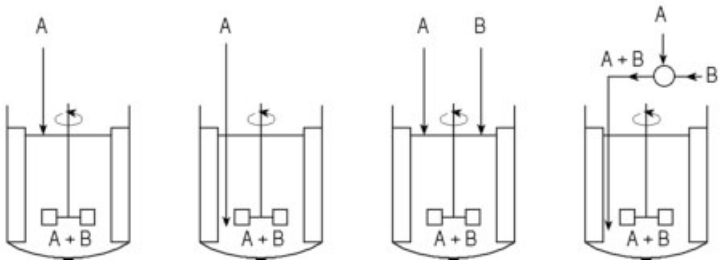

Figure 41. Some possibilities for introducing reactant feedstock streams to a precipitation vessel

9.1. Recrystallization from Solutions

Most of the impurities from a crystalline mass can often be removed by dissolving the crystals in a small amount of fresh hot solvent and cooling the solution to produce a fresh crop of purer crystals. However, the impurities must then be more soluble in the solvent than the main product is. Recrystallization may have to be repeated many times before crystals of the desired purity are obtained. A simple recrystallization scheme is

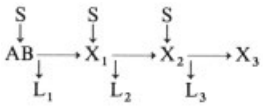

An impure crystalline mass AB (A is the less soluble, desired component) is dissolved in the minimum amount of hot solvent S and then cooled. The first crop of crystals X_1 will contain less impurity B than the original mixture, and B is concentrated in the liquor L_1. To achieve a higher degree of crystal purity, the procedure can be repeated.

In such a sequence, losses of the desired component A can be considerable, and the final amount of "pure" crystals may easily be a small fraction of the starting mixture AB. Many schemes have been designed to increase both the yield and the separation efficiency of fractional recrystallization. The choice of solvent depends on the characteristics of the required substance A and the impurity B. Ideally, B should be very soluble in the solvent at the lowest temperature employed and A should have a high temperature coefficient of solubility, so that high yields of A can be obtained from operation within a small temperature range.

9.2. Recrystallization from Melts

Melts can be fractionally recrystallized by schemes similar to those described in Section 9.1 for solutions, although a solvent is not normally added. Usually, simple sequences of heating (melting) and cooling (partial crystallization) are followed by separation of the purified crystals from the residual melt. Selected melt fractions may be mixed at intervals according to the type of scheme employed, and fresh feedstock may be added at different stages if necessary. Several such schemes have been proposed for purification of fats and waxes (→ Waxes, Chap. 4.2.3.2.) [109].

As described in Section 3.2, eutectic systems can theoretically be purified by single-stage crystallization, whereas solid solutions always require multistage operations. Countercurrent fractional crystallization processes in column crystallizers are described in Section 6.3.

9.3. Recrystallization Schemes

A number of fractional crystallization schemes have been devised [7], [97], the triangular scheme shown in Figure 42 (see next page) [110] demonstrates their essential features. Individual fractions, represented by the circles, are designated by a horizontal row number and a diagonal column letter. Consider the separation of two components A and B: if a constant fraction of A is precipitated at each operation and fractions are combined as in the triangular scheme, then the fraction x of the original A which appears at any given point in the triangular scheme is given by the binominal expansion

$$[x+(1-x)]^n = 1 \tag{51}$$

Similarly, for component B, where a constant fraction y is precipitated at each step, the distribution of B in various fractions is given by the expansion

$$[y+(1-y)]^n = 1 \tag{52}$$

Fractionation schemes are simplified when fractions of repeating composition recur at regular intervals. The first fraction in the above scheme that can have the same composition (i.e., A : B ratio) as the original mixture is 2 b. In this case, every fraction has the same composition as the fraction vertically above it in the scheme. Fraction 2 b will also be identical in composition to the original if the fraction of A which reaches that point is the same as the fraction of B reaching that point, i.e.,

$$2x\,(1-x) = 2y\,(1-y) \tag{53}$$

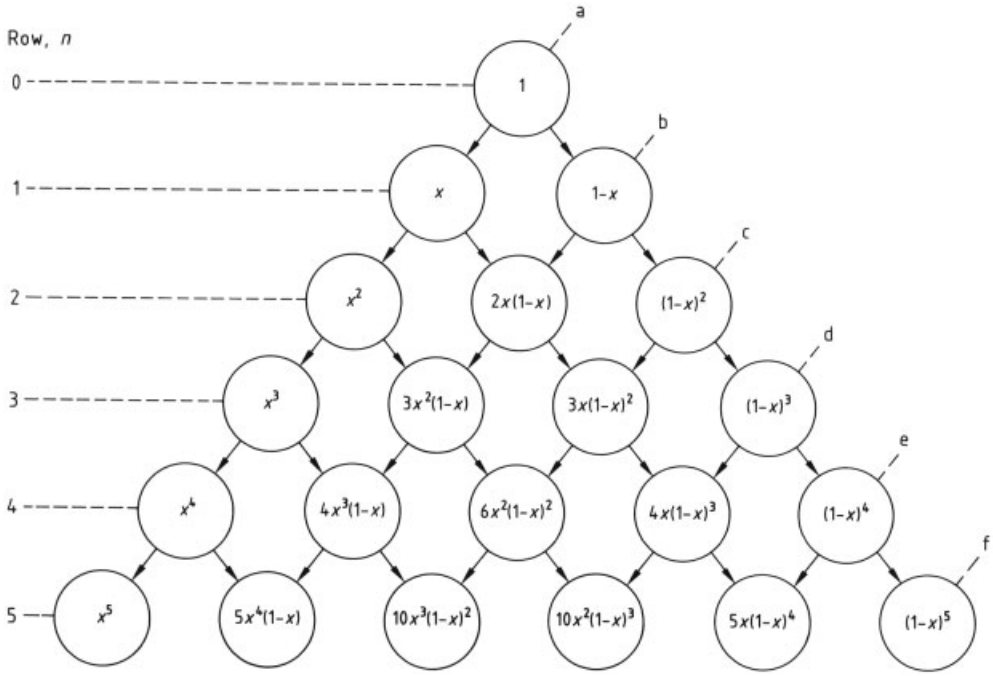

Figure 42. Triangular fractional crystallization scheme [110]

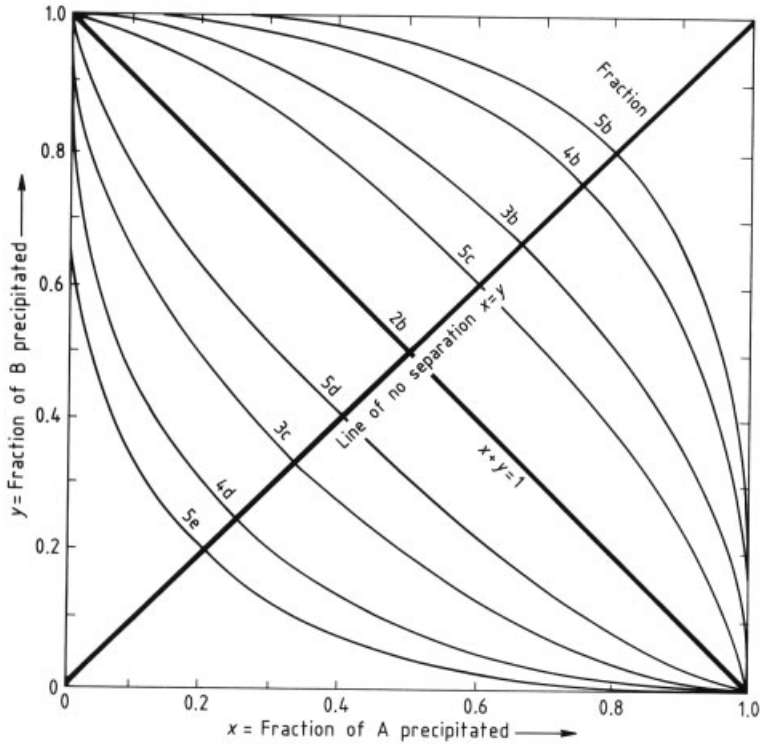

Figure 43. Operating curves for systems with repeating composition [110]
Curves were calculated by assuming that $\lambda = 7.21$; λ is the heterogeneous distribution, coefficient defined in Equation (50), Section 8.3.4.

This equation has two solutions ($x = y$ and $x + y = 1$). The first represents a line of no separation which, when drawn as in Figure 43, divides the diagram into two zones: enrichment of component B in the precipitate occurs in the upper, left-hand zone (enrichment being expressed by the ratio y/x) and enrichment of A in the lower, right-hand zone. The second solution to the equation ($x + y = 1$), represented by the other diagonal, gives repeating compositions for fractions in the same vertical columns and also gives enrichment of $(y/x)^n$ in the end fractions (column a in Fig. 42). This second diagonal may be called the operating line for a repeating composition in fraction 2 b. Operating lines can be calculated in a similar manner for repeating compositions in the other fractions, as depicted in Figure 43 [110]. Examples of the use of such a diagram are described in [110], [111].

9.4. Recrystallization from Supercritical Fluids

Solute solubilities in supercritical fluids (SCF) can undergo considerable changes with relatively small changes in pressure, and at constant pressure they generally pass through a minimum with temperature. These characteristics give the possibility of separating different solutes by manipulating pressure or temperature in the so-called "cross-over" region (Fig. 44). Rapid depressurization, e.g., through a nozzle, creates high supersaturation, fast nucleation and consequently large numbers of small crystals.

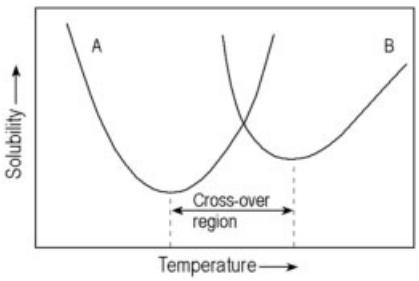

Figure 44. Solubility curve minima exhibited by two different solutes A and B in a supercritical fluid

Of all the solvents considered in recent years as possible SCFs for crystallization processes, the only two that now command any notable attention are water and CO_2 primarily because they are non-flammable, non-toxic, low-cost, and readily available. Of these, CO_2 is attracting greater support on account of its more accessible critical point (31 °C and 74 bar) compared with that of water (374 °C and 220 bar).

One problem with SCFs as crystallization solvents is that the solubilities of most organic compounds are generally low. Whilst the supersaturations developed can be high when expressed as the ratio $c/c*$ (see Section 2.2 Saturation and Supersaturation) they are usually low when expressed as a mass concentration driving force ($c - c*$) and the consequent low yields and productivity can make the process appear commercially unattractive. Nevertheless, SCF crystallization could still be profitable in, for example, the separation of isomers and polymorphs and the purification of high-value products such as pharmaceuticals.

Potential uses for processing with SCFs are discussed in [112–114] and examples of exploratory crystallizations using supercritical CO_2 are described in [115], [116].

9.5. Separation of Enantiomers and Racemates

The resolution of racemic (optically inactive) mixtures is becoming an important operation in the manufacture of chiral (optically active) pharmaceutical and agrochemical products because most of the specific activity usually resides predominantly in only one of the enantiomers in the equimolal mixture.

A crystalline racemate may be either a *conglomerate* (a physical mixture of two enantiomorphs) or a *racemic compound* (two enantiomers homogeneously distributed in the crystal lattice). Conglomerates, the much rarer of the two types of racemate, can be separated by recrystallization from solutions or melts.

Crystallization from solution is the most common procedure. A mixture of D and L enantiomers dissolved in a solvent S constitutes a ternary system in which the equilibria are best represented on a triangular diagram (see Fig. 8). The recrystallization procedure, which is possible only for conglomerates, involves alternate sequences of seedings using crystals of pure D and L enantiomers.

Crystallization from the melt, usually at low temperature, may also be a possibility. A temperature – composition phase diagram for a conglomerate is the same as that for a two-component eutectic system (Section 3.2.1) with perfect symmetry arising from the thermodynamic identity of the D and L enantiomers (melting point, enthalpy of fusion, etc.) giving a eutectic of equimolal racemic composition. Separation is effected by alternate D and L seeding. Examples of the crystallization procedures are described in [117]. General reviews of the industrial manufacture of optically active compounds are given in [118], [119].

10. Miscellaneous Crystallization Techniques

10.1. Salting-Out Crystallization

A solute can be deposited from solution by the addition of another substance (a soluble solid, liquid, or gas) which effectively reduces the original solute solubility. The process is commonly referred to as "salting out," although it is often applied to electrolytes and nonelectrolytes alike. For convenience this topic is dealt in Section 8.4.2.

10.2. Reaction Crystallization

The production of a solid crystalline product as the result of chemical reaction between gases and/or liquids is a standard method for the preparation of many industrial chemicals. Crystallization occurs because the gaseous or liquid phase becomes supersaturated with respect to the reactant.

Reaction crystallization is practiced widely, especially in industries where valuable waste gases are produced. For instance, ammonia can be recovered from coke-oven gases by converting it into ammonium sulfate by reaction with sulfuric acid [1]; sodium bicarbonate is formed by the interaction between brine and flue gases containing carbon dioxide [120]. A study of the crystallization reaction kinetics of a range of calcium phosphates is reported in [121].

10.3. Adductive and Extractive Crystallization

The simple crystallization of a binary eutectic system can only produce one of the components in pure form, while the residual mother liquor composition progresses towards that of the eutectic (Section 3.2.1). There is often a need, however, to produce both components in pure form, and one way in which this may be achieved is to add a third component to the system which forms a compound with one of the binary components. A typical *adductive crystallization* sequence would be as follows. A certain substance X is added to a given binary mixture of components A and B so that a solid complex, say $A \cdot X$, is precipitated. Component B is left in solution. The solid and liquid phases are then separated, and the solid complex is split into pure A and X, e.g., by the application of heat or by dissolution in some suitable solvent. Urea and thiourea, for example, have the property of forming complexes (adducts) with a wide range of hydrocarbons. Separation processes for other organic mixtures based on the formation of adducts have been described in [122].

An alternative to adductive crystallization for separating a binary eutectic mixture into its component parts is to alter the solid – liquid phase relationships by introducing a third component, usually a liquid called the solvent. This process is generally known as *extractive crystallization*.

Both adductive and extractive crystallization procedures are feasible for the separation of a wide range of close-boiling organic mixtures [122], [123] and for the recovery of inorganic salts from concentrated aqueous solution [124].

10.4. Spray Crystallization

Spray crystallization is similar to spray drying. The shape and size of the solid particles depend to a large extent on those of the spray droplets. The spray crystallization of a solution also has some features in common with prilling (see Section 6.5). Pilot plant investigations with a wide range of inorganic salts are described in [125].

10.5. Spherical Crystallization

An interesting technique for transforming small crystals into dense spherical agglomerates during the crystallization process, hence the name spherical crystallization, appears to have potential application in the pharmaceutical industry.

The method is based on the addition to a crystallizing system of a small amount of an immiscible liquid that preferentially wets the developing fine crystals and encourages them to compact into spherical agglomerates $\approx 250-1000$ mm. Chloroform appears to be the preferred organic liquid for use in association with crystallization from aqueous solution [126]. Spherical agglomerates of salicylic acid, other pharmaceutical substances and precipitated calcium carbonate have been successfully made by this technique [127].

10.6. Freeze Crystallization

Crystallization by freezing, more commonly known as freeze crystallization, is a process in which heat is removed from a solution to form crystals of the solvent rather than the solute. Subsequent steps include separation of crystals from the concentrated solution, washing the crystals with near-pure solvent, and finally melting them to generate visually pure solvent. The product of freeze crystallization can be either the melted crystals (near-pure solvent), as in water desalination, or the concentrated solution, as in the concentration of fruit juice or coffee extract. Freeze crystallization is applicable in principle to a variety of solvents and solutions; however, because it is most commonly applied to aqueous systems, the following account refers exclusively to the freezing of water.

One of the more obvious advantages of freezing over evaporation for removal of water from solutions is the potential for saving heat energy: the enthalpy of crystallization of ice (334 kJ/kg) is only one-seventh of the enthalpy of vaporization of water (2260 kJ/kg). Process energy consumption, however, may be reduced below that predicted by the phase-change enthalpy by utilizing energy recycle methods, such as multiple-effect or vapor compression, commonly employed in evaporation. In freeze crystallization plants operating by direct heat exchange, vapor compression has been employed to recover refrigeration energy by using the crystals to condense the refrigerant evaporated in the crystallizer.

Another advantage of freeze crystallization, important in many food applications, is that the volatile flavor components normally lost during conventional evaporation can be retained in the freeze-concentrated product. In fact, freeze crystallization is used mainly in the food industry. Despite earlier enthusiasm, large-scale applications in desalination, effluent treatment, dilute liquor concentration, solvent recovery, etc., have not yet emerged.

All freeze separation processes depend on the formation of pure solvent crystals from solution, as described for eutectic systems in Section 3.2.1; this allows single-stage operation. Solid-solution systems, which require multistage operation, are not usually handled. Several types of freeze crystallization can be designated according to the kind of refrigeration system used.

Indirect-Contact Freezing. In indirect-contact freezing, for instance, the liquid feedstock is crystallized in a scraped-surface heat exchanger (Section 6.1) fitted with internal rotating scraper blades and an external heat-transfer jacket through which a liquid refrigerant flows. The resulting ice–brine slurry passes to a wash column where the ice crystals are separated and washed before melting. One of a number of commercial systems based on this type of freezing process is described in [107].

Direct-contact freezing processes utilizing inert, immiscible refrigerants have been investigated widely for desalination purposes. A typical scheme is shown in Figure 45 (see next page) [129]. Seawater, precooled close to its freezing point, is fed continuously into the crystallization vessel where it comes in direct contact with a liquid refrigerant (e.g., *n*-butane) which vaporizes and causes ice crystals to form due to the exchange of latent heat. The ice–brine slurry is fed to a wash column where it is washed countercurrently with fresh water. The emerging brine-free ice is melted by the enthalpy of vapor condensation released from the compressed refrigerant. Energy input is required for the compressors.

Vacuum freezing processes do not require a conventional heat exchanger, and the problems of scale formation on heat-transfer sur-

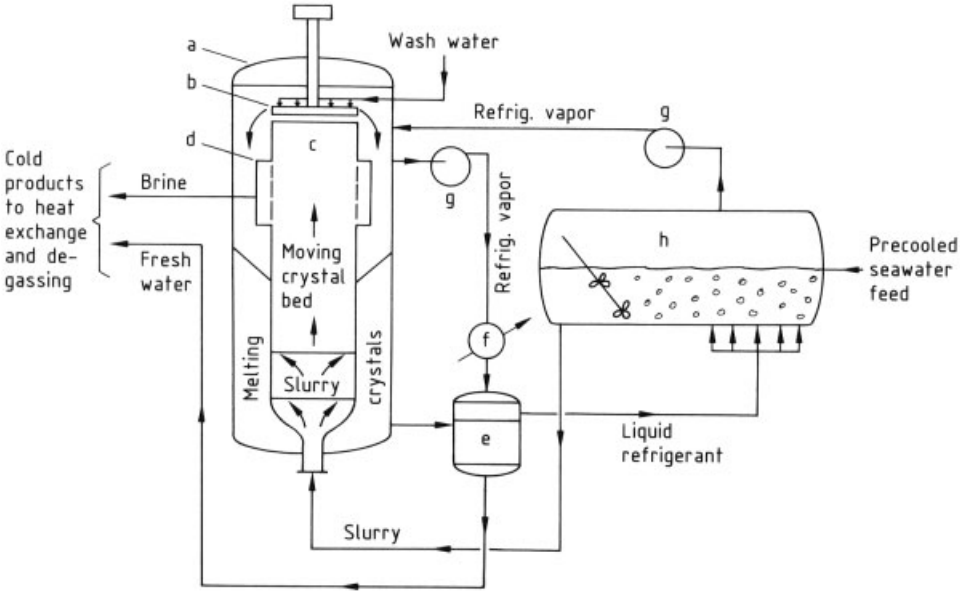

Figure 45. Desalination of seawater by freezing [129]
a) Washer–melter; b) Scraper; c) Wash column; d) Screens; e) Decanter; f) Heat exchanger; g) Compressor; h) Crystallizer

faces are avoided. Cooling is effected by flash-evaporating some of the solvent when the liquid feedstock enters a crystallization vessel maintained at reduced pressure. Vacuum freezing is potentially attractive for aqueous systems (7 kg of ice can be produced for every kilogram of water evaporated) but has not yet achieved commercial success.

A general review of freeze concentration as an industrial separation process is given in [130]. The potential of freeze crystallization in the recycling and reuse of wastewater is reviewed in [131]. The kinetics of ice crystallization in aqueous sugar solutions and fruit juice are described in [132], [133].

The subject of gas hydrates has become highly topical in recent years [134], particularly since the discovery of vast amounts of natural gas hydrates under ocean floor sediments at depths > 500 m and in polar permafrost regions. Gas hydrates are clathrate compounds in which variable (non-stoichiometric) amounts of gas, e.g., methane, ethane and propane, are trapped within ice crystal lattice 'cages'. The amount of entrapped gas increases with lowering temperature and increasing pressure. It has been estimated that worldwide the amount of methane trapped in gas hydrates is around 2×10^{16} m^3 at STP, which is roughly equivalent to twice the mass of carbon in all conventional gas, oil and coal deposits combined. Although the commercial recovery of methane from natural hydrate sources has not yet been achieved, very active international research programmes are at present being pursued [134].

11. References

1. A. W. Bamforth: *Industrial Crystallization,* Leonard Hill, London 1965.
2. G. Matz: *Kristallisation,* 2nd ed., Springer Verlag, Berlin 1969.
3. J. Nývlt: *Industrial Crystallization –State of the Art.* VCH Verlagsges., Weinheim 1982
4. S. J. Jančić, P. A. M. Grootscholten: *Industrial Crystallization,* Reidel, Dordecht 1984.
5. O. Söhnel, J. Garside: *Precipitation,* Butterworth-Heinemann, Oxford, 1992.
6. A. S. Myerson (ed): *Handbook of Industrial Crystallization,* Butterworth-Heineman, Boston 1993.
7. J. W. Mullin: *Crystallization,* 4th ed., Butterworth-Heinemann, Oxford 2001.
8. A. Mersmann (ed.): *Crystallization Technology Handbook,* Marcel Dekker, New York 1995.
9. K. K. Nass, *Ind. Eng. Chem. Research* **33** (1994) 1580 – 1584.

10. A. Seidell in W. F. Linke (ed.): *Solubilities of Inorganic and Metal Organic Compounds,* 4th ed., **vol. 1,** Van Nostrand, New York 1958, **vol. 2,** Am. Chem. Soc., Washington, D.C., 1965.
11. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* Pergamon, London 1963.
12. M. Broul, J. Nývlt, O. Söhnel: *Solubilities in Binary Aqueous Solutions,* Academia, Prague 1981.
13. J. W. Mullin, O. Söhnel, *Chem. Eng. Sci.* **32** (1977) 683 – 686.
14. Wilhelm Ostwald, *Z. Phys. Chem. (Leipzig)* **34** (1900) 493 – 503.
15. H. A. Miers, F. Isaac, *J. Chem. Soc.* **89** (1906) 413 – 454.
16. H. Feilchenfeld, S. Sarig, *Ind. Eng. Chem. Process Des. Dev.* **24** (1985) 130 – 133.
17. H. Kimura, *J. Cryst. Growth* **73** (1985) 53 – 62.
18. F. Grønvold, K. K. Meisingset, *J. Chem. Thermodyn.* **14** (1982) 1083 – 1098.
19. H. Kimura, K. Junjiro, *Sol. Energy* **35** (1985) 527 – 534.
20. A. Findlay, A. N. Campbell: *The Phase Rule and its Applications,* 9th ed., Longmans, London 1951.
21. J. E. Ricci: *The Phase Rule and Heterogeneous Equilibrium,* Van Nostrand, New York 1951.
22. H. R. Null: *Phase Equilibrium in Process Design,* Wiley-Interscience, New York 1970.
23. J. Nývlt: *Solid-Liquid Phase Equilibrium,* Academia, Prague, Elsevier, Amsterdam 1977.
24. H. Hörmeyer, *Ger. Chem. Eng. (Engl. Transl.)* **6** (1983) 277 – 281.
25. C. L. Kusik, H. P. Meissner, E. L. Field, *AIChE J.* **25** (1979) 759 – 762.
26. B. Sander, P. Rasmussen, A. Fredenslund, *Chem. Eng. Sci.* **41** (1986) 1197 – 1202.
27. P. T. Cardew, R. J. Davey, A. J. Ruddik, *J. Chem. Soc. Faraday Trans. 2,* **80** (1984) 659 – 668 and *Proc. R. Soc. London Ser. A* **398** (1985) 415 – 428.
28. K. K. Nass, *AIChE Symposium Series No. 284* **87** (1991) 72 – 81.
29. R. F. Strickland-Constable: *Kinetics and Mechanism of Crystallization,* Academic Press, London 1968.
30. A. C. Zettlemoyer (ed.): *Nucleation,* Marcel Dekker, New York 1969.
31. A. R. Ubbelohde: *Melting and Crystal Structure,* OUP, Oxford 1965.
32. V. A. Garten, R. B. Head, *Philos. Mag.* **8** (1963) 1793 – 1803 and **14** (1966) 1243 – 1253.
33. J. W. Mullin, C. L. Leci, *Philos. Mag.* **19** (1969) 1075 – 1077.
34. M. A. Larson, J. Garside, *Chem. Eng. Sci.* **41** (1986) 1285 – 1289.
35. J. Nývlt, O. Söhnel, M. Matuchová, M. Broul: *The Kinetics of Industrial Crystallization,* Academia, Prague 1985.
36. A. D. Randolph, M. A. Larson: *Theory of Particulate Processes,* 2nd. ed. Academic Press, New York 1988.
37. A. E. Nielsen, O. Söhnel, *J. Cryst. Growth* **11** (1971) 233 – 242.
38. P. Bennema, O. Söhnel, *J. Crystal Growth* **102** (1990) 547 – 556.
39. O. Söhnel, J. W. Mullin, *J. Cryst. Growth* **44** (1978) 377 – 382.
40. W. K. Burton, N. Cabrera, F. C. Frank, *Philos. Trans. R. Soc. London Ser. A.* **243** (1951) 299 – 358.
41. A. E. Nielsen, *J. Cryst. Growth* **67** (1984) 289 – 310.
42. P. Bennema, in J. W. Mullin (ed.): *Industrial Crystallization,* Plenum Press, London 1976, pp. 91 – 112.
43. A. A. Chernov, *Progr. Crystal Growth and Charact.* **26** (1993) 121 – 151.
44. W. Wöhlk: "Meßanordnungen zur Bestimmung von Kristallwachstumsgeschwindigkeiten", *Fortschr. Ber. VDI Z., Reihe 3* (1982) no. 71.
45. J. Garside, A. Mersmann, J. Nývlt: *Measurement of Crystal Growth Rates,* Euro. Fed. Chem. Eng. Working Party on Crystallization, 1990.
46. P. D. B. Bujac in J. W. Mullin (ed.): *Industrial Crystallization,* Plenum Press, London 1976, pp. 23 – 31. A. H. Janse, E. J. de Jong in J. W. Mullin (ed.): *Industrial Crystallization* Plenum Press, London 1976, pp.145 – 154.
47. G. H. Gilmer, P. Bennema, *J. Cryst. Growth* **13/14** (1972) 142 – 153.
48. J. R. Bourne, R. J. Davey, *J. Cryst. Growth* **36** (1976) 278 – 286.
49. G. M. van Rosmalen, G. J. Witkamp, C. H. de Vreugd in J. Nývlt, S. Zacek (eds.): *Industrial Crystallization 87,* Elsevier, Amsterdam 1987, pp. 15 – 20.
50. R. J. Davey et al., *Nature,* **353** (1991) 549 – 551 and *J. Chem. Soc. Faraday Trans.* **87** (1991) 3409 – 3414.
51. G. Deicha: *Lacunes des Cristeaux et leurs Inclusions Fluides,* Masson, Paris 1955.
52. H. E. C. Powers, *Sugar Technol. Rev.* **1** (1970) 85 – 190.

53. G. Mantovani et al., *L'Industria Saccarifera Italiana* **78** (1985) 7 – 14 and 79 – 86.
54. N. Burak, *Chemistry and Industry,* (1966) 844 – 850.
55. T. Tanaka, *Ind. Eng. Chem. Proc. Des. Dev.* **17** (1978) 241 – 246.
56. Y. L. Chen, J. Y. Chou, *Powder Technology* **77** (1993) 1 – 6.
57. L. Phoenix, *Brit. Chem. Eng.* **11** (1966) 33 – 38.
58. R. K. Multer, *Chem Eng. (N.Y.)* **89** (1982) March, 87 – 89.
59. J. W. Mullin, J. Nývlt, *Chem. Eng. Sci.* **26** (1971) 369 – 377.
60. A. G. Jones, J. W. Mullin, *Chem. Eng. Sci.* **29** (1974) 105 – 118.
61. A. G. Jones, *Chem. Eng. Sci.* **29** (1974) 1075 – 1087.
62. A. G. Jones, A. Chianese, J. W. Mullin in S. J. Jancic, E. J. de Jong (eds.): *Industrial Crystallization 84,* Elsevier, Amsterdam 1984, pp. 191 – 195.
63. J. Robinson, J. E. Roberts, *Can. J. Chem. Eng.* **35** (1957) 105 – 112.
64. J. Nývlt: *Design of Crystallizers,* CRC Press, Boca Raton, 1992.
65. R. C. Bennett, H. Fiedelman, A. D. Randolph, *Chem. Eng. Prog.* **69** (1973) no. 7, 86 – 93.
66. K. Wintermantel, G. Wellingdorf in A. Mersmann (ed.): *Industrial Crystallization 90,* GVC-VDI, Düsseldorf 1990, pp. 703 – 708.
67. M. Matsuoka in J. Garside, R. J. Davey, A. G. Jones (eds.): *Advances in Industrial Crystallization,* Butterworth-Heinemann, Oxford 1991, pp. 229 – 244.
68. J. G. D. Molinari in M. Zeif, W. R. Wilcox (eds.): *Fractional Solidification,* Dekker, New York 1967, pp. 393 – 400.
69. J. C. Chaty, H. A. O'Hern, *AIChEJ* **10** (1964) 74 – 78.
70. V. G. Ponomorenko, G. F. Potebnya, V. I. Bei, *Theor. Found. Chem. Eng. (Engl. Transl.)* **13** (1980) 724 – 729.
71. G. Wellinghoff, K. Wintermantel, *Chem. Ing. Tech.* **63** (1991) 881 – 891.
72. G. R. Atwood, *AIChE Sympos. Ser.* **65** (1969) no. 95, 112 – 121.
73. S. J. Jancic in J. Nývlt, S. Zacek (eds.): *Industrial Crystallization 87,* Elsevier, Amsterdam 1989, pp. 57 – 70.
74. W. C. Gates, J. E. Powers, *AIChE J.* **16** (1970) 648 – 655.
75. J. D. Henry, J. E. Powers, *AIChE J.* **16** (1970) 1055 – 1063.
76. D. L. McKay, in Reference [68], pp. 427 – 439.
77. J. A. Brodie, *Mech. Chem. Trans. Inst. Engrs. Australia* **7** (1971) no. 1, 37 – 44.
78. K. Tagegami, N. Nakamaru, M. Morita, in Reference [62], pp. 143 – 146.
79. M. Moritoki, *Int. Chem. Eng.* **20** (1980) 394 – 401.
80. W. H. Shearon, W. B. Dunwoody, *Ind. Eng. Chem.* **45** (1953) 496 – 504.
81. L. M. Nunelly, F. T. Cartney, *Ind. Eng. Chem. Prod. Res. Dev.* **21** (1982) 617 – 620.
82. P. von Schweizer et al. *Separatabdruck aus Chimia* **29** (1975) 78 – 84.
83. P. Scheffler, R. Henning, *Chem. Tech.* **31** (1979) 610 – 614.
84. M. M. Faktor, I. Garrett: *Growth of Crystals from the Vapor,* Chapman and Hall, London 1974.
85. S. Lewin: *The Solubility Product Principle,* Pitman, London 1960.
86. Wilhelm Ostwald: *Lehrbuch Allgemeine Chemie,* **vol. 2,** Engleman, Leipzig 1896, p. 401.
87. E. Hanitzch, M. Kahlweit in *Symposium on Industrial Crystallization,* Inst. Chem. Engrs., London 1969, pp. 130 – 141.
88. G. Matz, *Ger. Chem. Eng. (Engl. Transl.)* **8** (1985) 255 – 265.
89. M. von Smoluchowski, *Z. Phys. Chem. (Leipzig)* **92** (1918) 129 – 168.
90. T. F. Tadros, *Adv. Colloid Interface. Sci.* **12** (1980) 141 – 150.
91. M. J. Hounslow, *AIChE J.* **36** (1990) 1748 – 1752.
92. A. G. Jones, J. Hostomský, S. Wachi, *Chem. Eng. Comm.* **146** (1996) 1053 – 130.
93. P. P. von Weimarn, *Chem. Rev.* **2** (1926) 217 – 235.
94. A. G. Walton: *The Formation and Properties of Precipitates,* Interscience, New York 1967.
95. V. Chlopin, *Zeit. Anorg. Allg. Chem.* **143** (1925) 97 – 117.
96. H. A. Doerner, W. M. Hoskins, *J. Am. Chem. Soc.* **47** (1925) 662 – 675.
97. L. Gordon, M. L. Salutsky, H. H. Willard: *Precipitation from Homogeneous Solution,* Wiley-Interscience, New York 1959.
98. P. F. S. Cartwright, E. J. Newman, D. D. Wilson, *Analyst (London)* **92** (1967) 663 – 679.
99. Z. B. Alfassi, *Sep. Sci. Technol.* **14** (1979) 155 – 161.
100. J. W. Mullin, N. Teodossiev, O. Söhnel, *Chem. Eng. Process.* **26** (1989) 93 – 97.
101. H. Hoppe, *Chem. Process Eng. (London)* **49** (1968) no. 12, 61 – 63.

102. J. R. Fernandez-Lozano, *Ind. Eng. Chem. Process Des. Dev.* **15** (1976) 445 – 449.
103. Z. B. Alfassi, S. Mosseri, *AIChE J.* **30** (1984) 874 – 876.
104. J. R. Fernandez-Lozano, A. Wint, *Chem. Eng. London* **349** (1979) 688 – 690.
105. D. E. Garrett, *Chem. Eng. Prog.* **54** (1958) no. 12, 65 – 68.
106. R. W. Rice, R. E. Baud, *AIChE J.* **36** (1990) 293 – 298.
107. B. Marcant, R. David, *AIChE J.* **37** (1991) 1698 – 1710.
108. J. Franke, A. Mersmann, *Chem. Eng. Sci.* **50** (1995) 1737 – 1753.
109. Y. H. Hui (ed.): *Bailey's Industrial Oil and Fat Products,* 5th ed., (5 vols), Wiley, New York 1996.
110. E. F. Joy, J. H. Payne, *Ind. Eng. Chem.* **47** (1955) 2157 – 2161.
111. M. L. Salutsky, J. G. Sites, *Ind. Eng. Chem.* **47** (1955) 2162 – 2166.
112. G. T. Liu, K. Nagahama, *Ind. Eng. Chem. Res.* **35** (1996) 4626 – 4634.
113. A. A. Clifford, K. D. Bartle, *Chem. Ind.* (1996) 449 – 452.
114. B. Yu. Shekunov, M. Hamma, P. York, *J. Crystal Growth.* **198/199** (1999) 1345 – 1351.
115. E. M. Berends, O. S. L. Bruinsma, G. M. van Rosmalen, *J. Crystal Growth.* **128** (1993) 50 – 56.
116. C. Y. Tai, C. S. Cheng, *Chem. Eng. Res. Des.* **75** (1997) 228 – 232.
117. J. Jacques, A. Collet, S. H. Willen: *Enantiomers, Racemates and Resolutions,* Wiley, New York 1981.
118. R. A. Sheldon: *Chirotechnology,* Marcel Dekker, New York 1993.
119. A. N. Collins, G. N. Sheldrake, J. Crosby (eds.): *Chirality in Industry,* Wiley, New York 1985.
120. T. P. Hou: *Manufacture of Soda.* 2nd ed. Reinhold, New York 1942.
121. H. Tsuge, S. Yoshizawa, M. Tsuzuki, *Chem. Eng. Res. Dev.* **74** (1996) 797 – 802.
122. V. K. Jadhav, M. R. Chivate, N. S. Tavare, *Chem. Eng. Data* **36** (1991) 249 – 251 and **37** (1992) 232 – 235.
123. S. R. Dye, K. M. Ng, *AIChE J.* **41** (1995) 1456 – 1470.
124. D. A. Weingartner, S. Lynn, D. N. Hanson, *Ind. Eng. Chem. Res.* **30** (1991) 490 – 501.
125. D. E. Metheny, S. W. Vance, *Chem. Eng. Progr.* **58** (1962) no. 6, 45 – 48.
126. C. E. Capes, J. P. Sutherland, *Ind. Eng. Chem. Proc. Des. Dev.* **6** (1967) 146 – 154.
127. Y. Kawashima et al. *Powder Technology,* **39** (1984) 41 – 47 and **46** (1986) 61 – 66.
128. W. H. van Pelt, *Food Eng.* (1975) Nov., 77 – 79.
129. A. J. Barduhn, *Chem. Eng. Progr.* **71** (1975) no. 11, 80 – 87.
130. H. A. C. Thijssen, in A. Spicer (ed.): *Advances in Preconcentration and Dehydration of Foods,* Wiley-Interscience, New York 1974.
131. J. A. Heist, *AIChE Symp. Ser.* **77** (1981) no. 209, 259 – 272.
132. A. M. Omran, C. J. King, *AIChE J.* **20** (1974) 799 – 801.
133. A. G. Patil, *Int. Sugar J.* **95** (1993) 349 – 355.
134. E. D. Sloan: *Clathrate Hydrates of Natural Gases,* Marcel Dekker, New York 1993.

Crystallization, Industrial → Crystallization and Precipitation
Crystallography, X-ray → Structure Analysis by Diffraction
Cumene → Hydrocarbons
Cumene → Phenol
Curium → Transuranium Elements

Drying of Solid Materials

See also Solids Technology, Introduction

EVANGELOS TSOTSAS, Institut für Thermische Verfahrenstechnik, University of Karlsruhe, Karlsruhe, Federal Republic of Germany

VOLKER GNIELINSKI, Institut für Thermische Verfahrenstechnik, University of Karlsruhe, Karlsruhe, Federal Republic of Germany

ERNST-ULRICH SCHLÜNDER, Institut für Thermische Verfahrenstechnik, University of Karlsruhe, Karlsruhe, Federal Republic of Germany

1. Theoretical Fundamentals of the Drying Process . . . 2
1.1. Concepts, Definitions . . . 2
1.2. Characteristics of Moist Solids . . . 3
1.3. Drying Rate Curves for Convection Drying . . . 6
1.4. Drying Rate Curves for Contact Drying . . . 11
1.5. Drying Rate and Moisture-Composition Curves for Solids Wetted by Liquid Mixtures . . . 13
2. Drying Methods and Dryer Types . . . 15
2.1. Convection Drying . . . 15
2.1.1. Flowing Gas . . . 15
2.1.2. The Solid is Aerated . . . 18
2.1.3. Large-Scale Agitation of the Solid . . . 21
2.1.4. The Solid Moves in the Drying Agent . . . 23
2.1.5. The Material is Sprayed . . . 24
2.2. Contact Drying . . . 25
2.2.1. Flat and Strip Materials . . . 25
2.2.2. Low-Viscosity Materials . . . 26
2.2.3. Pasty Materials . . . 27
2.2.4. Granular Materials . . . 28
2.3. Radiant Heat or Infrared Drying . . . 29
2.4. Dielectric Heating . . . 30
2.5. Vacuum and Freeze Drying . . . 30
3. Selecting, Sizing, and Energy Requirements of Dryers . . . 31
3.1. Choosing the Type of Dryer . . . 31
3.1.1. The Role of the Material Properties of the Solid . . . 31
3.1.2. Production Rate . . . 31
3.1.3. Dryer Ventilation . . . 32
3.2. Sizing the Dryer . . . 33
3.2.1. Batch Dryers . . . 33
3.2.2. Continuous Dryers . . . 34
3.3. Heat and Driving Power Requirements . . . 35
4. References . . . 36

Symbols

A	surface area of the drying solid, m^2
C	excess air factor in Eq. (38)
C	radiation coefficient, $W\,m^{-2}\,K^{-4}$
c	specific heat, $J\,kg^{-1}\,K^{-1}$
d	particle diameter, m
h	enthalpy, J/kg
L	dryer length, m
l	modified mean free path, m
M	mass, kg
$\dot{M}$	mass flow rate, kg/s
$\tilde{M}$	molecular mass, kg/kmol
$\dot{m}$	drying rate, $kg\,m^{-2}\,s^{-1}$
N	number of moles, kmol
P	pressure, Pa
$\dot{Q}$	heat transfer rate, W
$\dot{q}$	heat flux, W/m^2
R	gas constant, $J\,kg^{-1}\,K^{-1}$
s	width of sample, m
T	temperature, K
t	time, s
u	velocity of drying agent, m/s
v	velocity of liquid in capillaries, m/s
$\dot{W}$	power, W
X	moisture content in solid, kg moisture/kg dry solid
$\tilde{x}$	mole fraction in the liquid phase
Y	humidity of the drying agent, kg moisture/kg drying agent
$\tilde{y}$	mole fraction in the gas phase
z	axial coordinate in a dryer, m
z	distance from the free surface of sample, m
α	heat-transfer coefficient, $W\,m^{-2}\,K^{-1}$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_04

β	mass-transfer coefficient, m/s
γ	accommodation coefficient
δ	diffusion coefficient, m^2/s
δ	surface roughness, m
ζ	dimensionless axial coordinate in a dryer
Λ	mean free path, m
λ	thermal conductivity, $W\,m^{-1}\,K^{-1}$
μ	ratio of molecular masses
$\dot{\nu}$	normalized drying rate, Eq. (23)
ξ	normalized moisture content, Eq. (24)
ϱ	density, kg/m^3
τ	dimensionless time
Φ	dimensionless number, Eq. (29)
φ	relative humidity
φ	surface coverage
ψ	porosity

Subscripts

a	ambient
bed	bed
crit	at the critical point
eq	in equilibrium
f	final, at the outlet
g	gas (drying agent)
h	maximum hygroscopic moisture content
heat	from the heater
i	initial, at the inlet
l	liquid (moisture)
lost	heat losses to the surroundings
max	maximum
p	at constant pressure
r	radiation
s	dry solid
v	vapor
vent	ventilator
W	wall
w	wetting
1	component index (isopropyl alcohol)
I	in the first period of drying

Superscripts

0	effective coefficient, Eqs. (15), (17)
$*$	saturated
$\sim$	molar quantities

1. Theoretical Fundamentals of the Drying Process

1.1. Concepts, Definitions

Drying denotes the separation of volatile liquids from solid materials by vaporizing the liquid and removing the vapor. The liquid that is to be removed is usually water, but it could also be a solvent such as alcohol or acetone, or a mixture of such solvents. The solid material that is to be dried can be a natural product such as wood, a semifinished or a finished good (such as paper). The removal of water from other fluids such as refrigerants or from gases, such as natural gas, is also considered a drying process, but this topic is not treated in this article.

The vaporization of liquids requires the supply of heat. Accordingly, drying can be considered a *thermal separation process.* The removal of liquids from solids without the application of heat, which is the case in a centrifuge, does not come under the strict definition of drying.

The product that is to be dried is denoted as the moist solid, or simply as the solid. The substance that carries the necessary heat is called the drying agent. This substance could be air, an inert gas, or superheated steam. Heat could also be supplied by radiation, by hot surfaces, or by microwaves. The moisture content of the solid is denoted by X and measured in kg liquid per kg of dry solid. For the humidity of a gaseous drying agent the symbol Y is used (in kg vapor per kg of dry gas). The saturation humidity is denoted by $Y*$. The mass flux of the vapor leaving the surface of the solid per unit time is called the drying rate, and is denoted by the symbol $\dot{m}$ (in $kg\,m^{-2}\,s^{-1}$). The drying rate is usually determined by measuring the change of moisture content with time dX/dt. It follows that

$$\dot{m} = -\frac{M_s}{A}\frac{dX}{dt}$$

M_s is the mass of the dry solid, and A is the portion of its surface area that is in contact with the drying agent. The drying rate $\dot{m}$ depends on the conditions of drying and on the moisture content X. The drying conditions are specified by factors such as the air pressure, temperature, and humidity, the radiator temperature, the temperature of a heating surface, or the strength of the microwaves. The relationship of the drying rate

$\dot{m}$ and the moisture content X under constant drying conditions – this is $\dot{m}(X)$ – is called the *drying rate curve.*

1.2. Characteristics of Moist Solids

Almost all industrial products have to be dried one or more times in the course of their manufacture. Consequently, the variety and characteristics of moist solids are manifold. However, only two of these characteristics have a direct influence on the drying rate curve $\dot{m}(X)$. They relate to the questions:

1) Is the liquid in the solid freely mobile or is it bonded to the solid by sorption?
2) Does vaporization take place at the surface or in the interior of the solid?

The first characteristic is described by the thermodynamic equilibrium (sorption isotherms), the second by the kinetics of liquid migration in the interior of the solid (capillarity, diffusion). Both topics will be treated thoroughly subsequently. In any case it should be borne in mind that drying is not only a thermal separation process; it is also a means to manufacture specified products and to influence their quality. Paper is an example of such a specified product whose quality is controlled by the choice of drying conditions. Further, the risk of possible product damage during drying must be reduced. This is particularly important for sensitive products such as foodstuffs. In this context, a large number of product characteristics differing from material to material should be accounted for. Some remarks on the handling of temperature-sensitive materials are given in Section 3.1

Free and Bonded Moisture in Solids. If free (unbound) liquid is in contact with its own vapor, then the vapor pressure is equal to the saturation pressure for the respective temperature $P_v*(T)$, e.g., $P_v* = 0.1$ MPa for water at 100 °C. In the presence of an inert gas (e.g., air), the total pressure at the surface of the liquid P is equal to the sum of the saturation pressure P_v* and of the partial pressure of the inert gas P_g:

$$P = P_v^*(T) + P_g \tag{1}$$

In most practical cases the total pressure P is prescribed. The partial pressure of the inert gas can be obtained from Eq (1):

$$P_g = P - P_v^*(T) \tag{2}$$

The ratio of the partial pressure to the total pressure is equal to the mole fraction of the respective component. It follows:

$$\tilde{y}_v^*(T) = \frac{N_v}{N_v + N_g} = \frac{P_v^*}{P} \tag{3}$$

and correspondingly

$$\tilde{y}_g = \frac{N_g}{N_v + N_g} = \frac{P_g}{P} \tag{4}$$

The number of moles N (expressed in kmol), is related to the mass of the respective component M (expressed in kg) by the molecular mass $\tilde{M}$ (which is given in kg/kmol) by

$$M_v = N_v \tilde{M}_v \quad M_g = N_g \tilde{M}_g \tag{5a,b}$$

($\tilde{M}_v = 18.01$ kg/kmol for water and $\tilde{M}_g = 28.96$ kg/kmol for air) From Eqs. (1) – (5) the saturation humidity of the inert gas [$Y*(T)$ in kg vapor per kg inert gas] can be obtained:

$$Y^*(T) = \frac{\tilde{M}_v}{\tilde{M}_g} \frac{P_v^*(T)}{P - P_v^*(T)} \tag{6}$$

The humidity of the unsaturated inert gas is given by the relationship

$$Y = \frac{\tilde{M}_v}{\tilde{M}_g} \frac{\varphi P_v^*(T)}{P - \varphi P_v^*(T)} \tag{7}$$

where φ is the relative humidity, defined as

$$\varphi = P_v / P_v^*(T) \tag{8}$$

The limits of φ are 0 and 1, which correspond to completely dry air and vapor-saturated air, respectively. In summary, the thermodynamic equilibrium between the unbound moisture within a solid that is in contact with a gas – vapor atmosphere is characterized by $\varphi = 1$ and $Y = Y*(T)$. This is the case regardless of the magnitude of the moisture content X.

If the moisture is bonded (bound) to the solid material, the air humidity under conditions of thermodynamic equilibrium (denoted by Y_{eq}) depends not only on the temperature T, but also on the moisture content X_{eq}, i.e., $Y_{eq} = Y_{eq}(T, X_{eq})$. The parameter Y_{eq} is always

smaller than $Y*$ and consequently, the partial vapor pressure in the gas – vapor mixture $P_{v,eq}$ is always smaller than the saturation vapor pressure $P_v*(T)$. The diminishing of the vapor pressure at the surface of materials with a high moisture content is associated with the phenomenon called *capillary condensation*. With a low moisture content (down to the limit $X \to 0$), the intermolecular attraction between the fluid and the solid plays a dominant role. One speaks of *Langmuir sorption*. For practical purposes the equilibrium moisture content of a solid X_{eq} is depicted as a function of the relative humidity of the air φ at various temperatures T, i.e., $X_{eq} = X_{eq}(\varphi, T)$. Such curves are called *sorption isotherms* (Fig. 1).

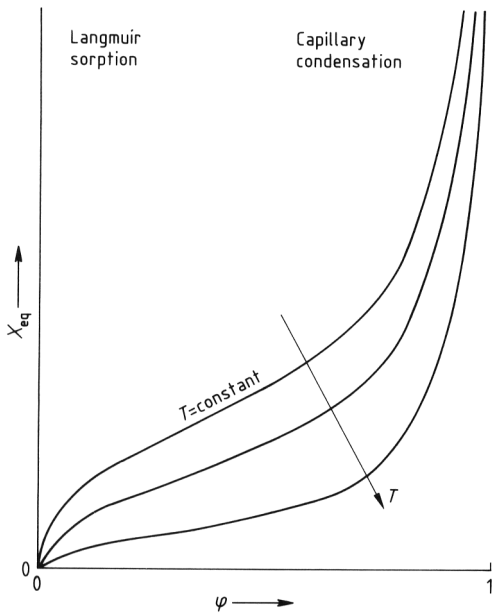

Figure 1. Sorption isotherms $X_{eq}(\varphi, T)$ (qualitative graph)

Figure 2 shows two sorption isotherms for potatoes. To dry potatoes to a residual moisture content of $X_{eq} = 0.10$, for example, cold air with a temperature of $T = 0\,°C$ and a low relative humidity of $\varphi = 0.24$, or hot air with a temperature of $100\,°C$ and a high relative humidity of $\varphi = 0.77$ could be used.

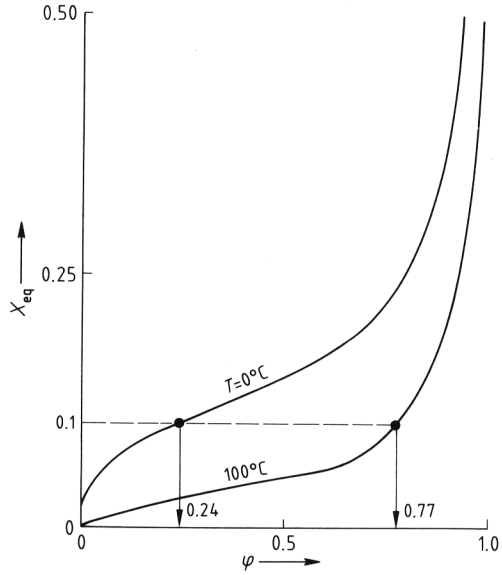

Figure 2. Sorption isotherms for potatoes [1]

Sorption isotherms also aid in choosing the drying conditions when using steam as the drying agent. Taking the potatoes as an example, if the steam temperature is $100\,°C$, then its pressure should not exceed 77 kPa in order to attain a residual moisture of $X_{eq} = 0.10$.

Solids that contain a considerable amount of bonded moisture at normal conditions are usually called hygroscopic. Sorption isotherms for many such solids are to be found in [1]; some are presented in Figure 3.

Figure 3. Sorption isotherms for various materials [1]
A) Cereal foods and tobacco
a) Tobacco leaves; b) Tobacco; c) Macaroni; d) Flour; e) Bread; f) Crackers
B) Textile fibers
a) Beryllium alginate fiber (25 °C); b) Calcium alginate fiber (25 °C); c) Nitrate fiber (25 °C), copper fiber (25 °C), viscose fiber (25 °C), wool (worsted, 25 °C); d) Casein fiber (20 °C), wool (35.6 °C); e) Jute; f) Cotton (mercerized, 20 °C), silk; g) Flax (30 °C), hemp; h) Cotton, fluffed (20 °C); i) Acetate fiber (25 °C); j) Linen; k) Perlon fiber (25 °C), nylon fiber (25 °C); l) Cellulose acetate fiber; m) Pe-Ce fiber (20 °C)
C) Leather, rubber, catgut, feathers
a) Sheepskin; b) Leather (sole, oak tanned); c) Catgut; d) Gold beater skin; e) Feathers; f) Latex dipped cord; g) Reclaimed rubber; h) Rubber

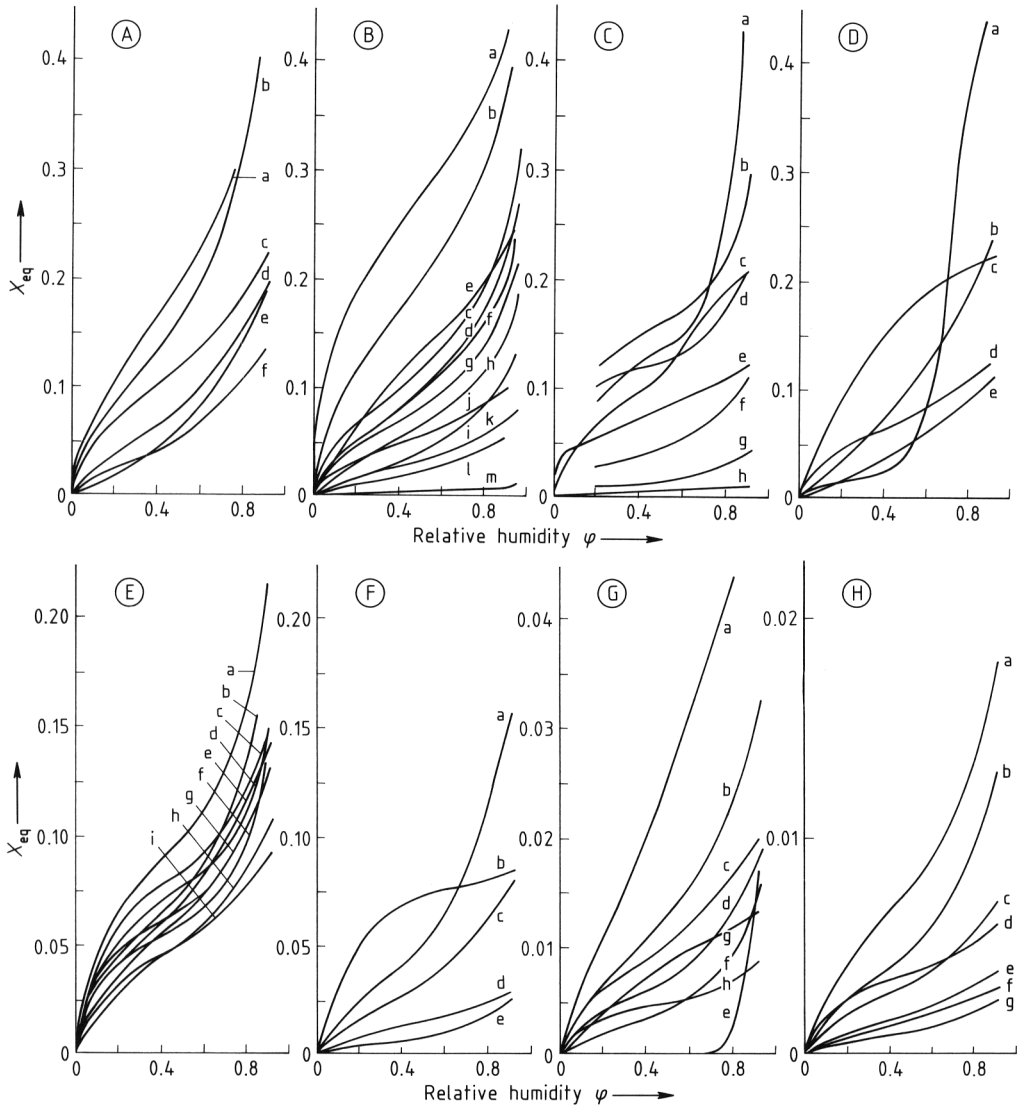

Figure 3. Continued.
D) Binding agents, adsorption agents, soaps
a) Activated charcoal (Novitkol, 400 kg/m^3); b) Soap; c) Silicagel; d) Glue, starch; e) Gelatin
E) Paper and wood shavings
a) Wood shavings; b) Manila paper; c) Kraft paper; d) Cellulose paper and printing paper from sulfite pulp; e) Writing paper; f) Book paper; g) Writing paper (high quality, white); h) Newsprint, filter paper; i) Brazil wood paper, offset press paper
F) Plastics and carbon black
a) Poly(vinyl alcohol) powder (25 °C); b) Carbon black (25 °C); c) 6-Polyamide (20 °C); d) Polyacrylonitrile powder (50 °C); e) Mixed polymer powder (85 % poly(vinyl chloride), 14 % poly(vinyl acetate), 1 % maleic acid (50 °C)
G) Building materials, soils, asbestos
a) Cement mortar 2040 kg/m^3 ; b) Diatomaceous earth; c) Concrete 2300 kg/m^3 ; d) Lime mortar 1800 kg/m^3 ; e) Gypsum 1340 kg/m^3 ; f) Lime stucco 1600 kg/m^3 ; g) Kaolin h) Asbestos
H) Plastics
a) Polystyrene, polymerized in blocks (40 °C); b) Poly(vinyl chloride) powder (50 °C); c) Mixed polymer powders (85 % poly(vinyl chloride), 15 % poly(vinyl acetate) (50 °C); d) Polyethylene powder (25 °C); e) Polytrifluoroethylene powder (50 °C); f) Polyethylene granules with carbon black (25 °C); g) Polyethylene granules (25 °C)

Liquids that are bound to solids by sorption are in a state of lower free energy than those that are unbound. Hence, in the drying of hygroscopic substances, the enthalpy of wetting Δh_w must be included in addition to the vaporization enthalpy Δh_v. The so-called differential heat of wetting can be calculated from the relationship

$$\Delta h_w = -R_v \left[\frac{d \ln \varphi}{d\,(1/T)} \right]_{X_{eq}=\text{constant}} \tag{9}$$

where R_v is the gas constant of the vapor. When the logarithm of the relative humidity is plotted versus the reciprocal value of the absolute temperature at constant equilibrium moisture content, straight lines are obtained (*sorption isosteres*). The sorption isosteres and the differential heat of wetting for potatoes are plotted in Figures 4 and 5, respectively. To calculate the amount of heat needed to accomplish drying to a final moisture content $X_{eq,\,f}$ from an initial moisture content $X_{eq,\,i}$, the integral heat of wetting is needed.

$$\overline{\Delta h_w} = \frac{1}{X_{eq,i} - X_{eq,f}} \int_{X_{eq,f}}^{X_{eq,i}} \Delta h_w \, dX_{eq} \tag{10}$$

The total enthalpy change Δh_{total} is the sum of the vaporization enthalpy and the integral enthalpy of wetting:

$$\Delta h_{total} = \Delta h_v + \overline{\Delta h_w} \tag{11}$$

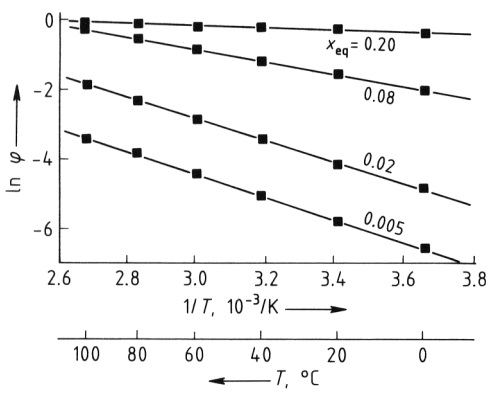

Figure 4. Sorption isosteres for potatoes
The graph can be used for the determination of the differential heat of wetting.

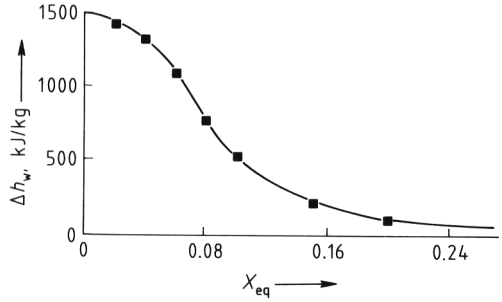

Figure 5. Differential heat of wetting Δh_w for potatoes as a function of their equilibrium moisture content X_{eq} [1]
The points for $X_{eq} = 0.20$, 0.08, and 0.02 can be calculated from the slope of the corresponding straight lines in Figure 4, according to Eq. (9).

1.3. Drying Rate Curves for Convection Drying

In most cases, drying rate curves are experimentally determined using air as the drying agent. During each experiment of this kind the air conditions, that is the total pressure P, the temperature T, the relative humidity φ, and the air velocity u are kept constant (Fig. 6). The sample is weighed at specified intervals of time Δt, the reduction of mass ΔM is calculated ($\Delta M = M_s\,\Delta X$), and the drying rate is obtained:

$$\dot{m} = -\frac{1}{A}\frac{\Delta M}{\Delta t} = -\frac{M_s}{A}\frac{\Delta X}{\Delta t} \tag{12}$$

For most moist solids, especially those having capillary porosity, the drying rate $\dot{m}$ depends upon the moisture content X in a manner similar to that shown in Figure 7 [1], [2].

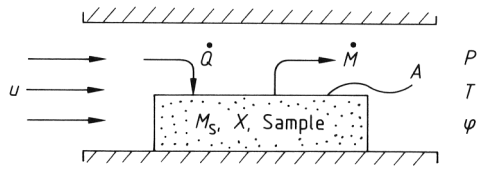

Figure 6. Schematic of a drying tunnel to measure the drying rate curve

In the first drying period, the drying rate remains practically constant. During this period, unbound liquid is vaporized from the surface of the solid and carried away by the drying agent. When the moisture content is reduced below a critical value X_{crit}, the surface of the solid dries out, and further evaporation takes place in the

interior of the porous solid. The drying rate decreases with decreasing moisture content. This is called the second drying period. The residual moisture in the solid is bound to it by sorption. The drying rate decreases rapidly with decreasing moisture content and tends to zero as the hygroscopic equilibrium moisture content $X_{eq}(T, \varphi)$ is approached. The regime between the maximum hygroscopic moisture content X_h and the equilibrium value of X_{eq} is designated as the third drying period.

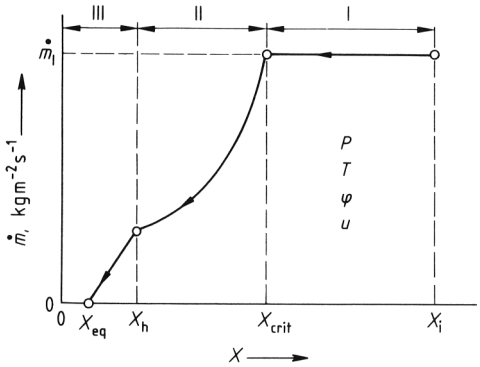

Figure 7. Typical drying rate curve $\dot{m}(X)$, showing the first, second, and third period of drying

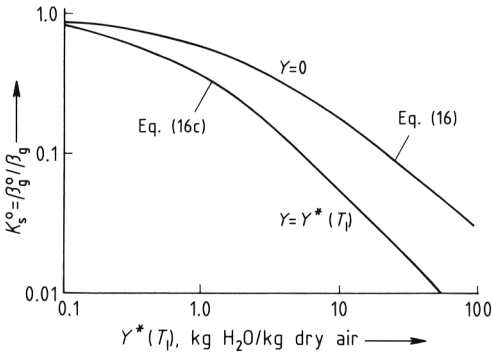

Figure 8. Stefan correction K_S^0 as a function of the saturation humidity $Y*(T_I)$ for the limiting cases $Y=0$ and $Y=Y*(T_I)$
Curves are calculated for water vapor and air, $\mu = 1.608$.

First Drying Period. The drying rate during the first period (denoted by $\dot{m}_I$) is dependent only upon the conditions of drying and not upon the characteristics of the solid: there is only evaporation from the surface of the solid. The temperature of the solid, denoted by T_I, is established so that the heat flux from the drying agent to the solid $\dot{q}$ is equal to the product of the drying rate and the vaporization enthalpy,

$$\dot{q} = \dot{m}_I \Delta h_v \tag{13}$$

The drying agent, usually air, is completely saturated at the surface of the solid,

$$Y\,(\text{at the surface of the solid}) = Y^*\,(T_I) \tag{14}$$

Assuming that the humidity in the bulk of the drying agent is equal to Y, where $Y < Y*(T_I)$, the drying rate $\dot{m}_I$ is proportional to the humidity difference $Y*(T_I) - Y$. The proportionality constant is the effective mass-transfer coefficient β_g^0. Hence,

$$\dot{m}_I = \varrho_g \beta_g^0 \left[Y^*\,(T_I) - Y\right] \tag{15}$$

In this expression ϱ_g is the density of air (or other drying agent). The coefficient β_g^0 is related to the real mass-transfer coefficient β_g by

$$\beta_g^0 = \beta_g \frac{\ln \frac{1+\mu Y^*(T_I)}{1+\mu Y}}{\mu\left[Y^*\,(T_I) - Y\right]} \tag{16}$$

in which

$$\mu = \tilde{M}_g / \tilde{M}_v \tag{16a}$$

is the ratio of the molecular masses of the inert gas and the vapor ($\mu = 28.96/18.01 = 1.606$ for air and water vapor). For low levels of humidity ($Y < Y*(T_I) < 0.05$), Equation 16 reduces to

$$\beta_g^0 = \beta_g \tag{16b}$$

For the case $Y \to Y*(T_I)$,

$$\beta_g^0 = \beta_g \frac{1}{1+\mu Y^*\,(T_I)} \tag{16c}$$

is obtained. The effective mass-transfer coefficient β_g^0 is always smaller than the real one β_g. Only at sufficiently low humidities Y are the two coefficients equal. This fact is illustrated in Figure 8, where the dependence of the so-called *Stefan correction* $K_s^0 = \beta_g{}^0/\beta_g$ on the saturation humidity is depicted.

The saturation humidity $Y*(T_I)$ can be found from the energy balance given by Equation 13, where the heat flux $\dot{q}$ is given by

$$\dot{q} = \alpha^0\,(T - T_I) \tag{17}$$

In this equation T_I is the solid temperature during the first drying period, T is the temperature

in the bulk of the drying agent, and α^0 is the effective heat-transfer coefficient. The latter is related to the real heat-transfer coefficient α by the following relationship:

$$\alpha^0 = \alpha \cdot \frac{\Delta h_v}{c_{pv}(T - T_I)} \ln\left[1 + \frac{c_{pv}(T - T_I)}{\Delta h_v}\right] \tag{18}$$

In this equation Δh_v is the vaporization enthalpy of the liquid and c_{pv} is the heat capacity of the vapor. The value of α^0 is smaller than the value of α. If the temperature difference $(T - T_I)$ is sufficiently small, these two heat-transfer coefficients are practically the same. Figure 9 shows the relationship between $\alpha^0/\alpha = K_A$ and $c_{pv}(T - T_I)/\Delta h_v$. The former is sometimes called the *Ackermann correction*, while the latter is termed the *phase change number.*

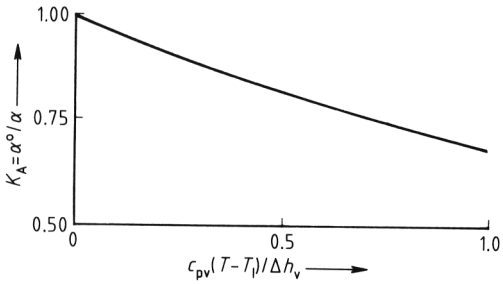

Figure 9. Ackermann correction K_A as a function of the phase change number $c_{pv}(T - T_I)/\Delta h_v$

The temperature that the solid reaches in the first drying period T_I can be found by inserting the equation for the mass transfer (Eq. 15) and the equation for the heat transfer (Eq. 17) into the energy balance (Eq. 13). The result is

$$\alpha^0 (T - T_I) = \varrho_g \beta_g^0 [Y^*(T_I) - Y] \Delta h_v \tag{19}$$

For solids wetted by water that are to be dried by air, the approximation

$$\frac{\alpha^0}{\beta_g^0} \frac{1}{\varrho_g} \approx c_{pg} + Y c_{pv} \tag{20}$$

is useful provided that the saturation humidity $Y*(T_I)$ is not too high. The quantities c_{pg} and c_{pv} are the heat capacities of air and water vapor, respectively. Whenever Equation (20) is valid, Equation (19) can be transformed into one that is written in the coordinates of the Mollier diagram:

$$\frac{h^*(T_I) - h}{Y^*(T_I) - Y} = c_{pl} T_I \tag{21}$$

In this relationship c_{pl} is the heat capacity of the liquid and h is the enthalpy of the moist air

$$h = c_{pg} T + Y\left(\Delta h_v^0 + c_{pv} T\right) \tag{22}$$

where Δh_v^0 is the vaporization enthalpy at 0 °C. Equation (21) is the locus of points of the *adiabatic saturation line*, crossing the curve with $\varphi = 1$ at the point $(h*, Y*)$ (Fig. 10).

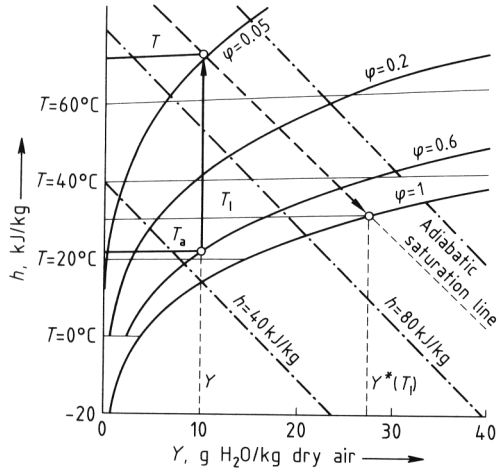

Figure 10. Mollier diagram for moist air
Evaluation of the solid's temperature T_I in the first period of drying

If the temperature of the solid T_I and either the heat-transfer coefficient α^0 or the mass-transfer coefficient β_g^0 are known, then the drying rate can be determined numerically either with Equations (17) and (13) or with Equation (15).

Numerical Example. Consider the drying agent to be ambient air ($T_a = 22$ °C, $\varphi = 0.60$) that has been heated to $T = 72$ °C. The heat-transfer coefficient at the air – solid interface is 20 W m^{-2} K^{-1} ($= \alpha^0$), the moisture is water. What is the drying rate during the first drying period?

The temperature of the solid during the first drying period T_I can be evaluated with the help of the Mollier diagram (Fig. 10). Condition of ambient air: $T_a = 22$ °C, $\varphi = 0.60 \Rightarrow Y = 0.010$ kg H_2O/kg dry air. Condition of preheated air: $T = 72$ °C, $Y = 0.010$ kg H_2O/kg dry air $\Rightarrow \varphi = 0.05$. Following the adiabatic saturation line, a value of $T_I = 30$ °C is obtained. The

corresponding vaporization enthalpy of water is $\Delta h_v = 2\,430\,300$ J/kg (tabulated value). Using Equations (13) and (17), the drying rate during the first drying period $\dot{m}_I = 1.24\,\mathrm{kg\,m^{-2}\,h^{-1}}$ is obtained.

Second and Third Drying Periods. The first drying period is completed when the solid's moisture content X has reached the critical value X_{crit} (cf. Fig. 7). After this point the capillary forces are no longer sufficient to transport the liquid to the surface of the solid. The liquid–vapor interface (drying front) moves inside the solid. The dried portion of the solid near the surface thermally insulates the moist inner portions of the solid. At the same time these dried regions impede the transport of the vapor to the bulk of the drying agent. These are the reasons why the drying rate decreases while the solid's temperature rises after the completion of the first drying period. The critical moisture content X_{crit} must usually be determined experimentally; Table 1 gives representative values for various materials.

Table 1. Critical moisture content of solids when dried in air [3]

Solid	Layer thickness, cm	Critical moisture content X_{crit}
Sulfite pulp	0.6 – 1.9	0.6 – 0.8
Paper, white eggshell	0.02	0.41
Paper, fine book	0.0125	0.33
Paper, coated	0.01	0.34
Paper, newsprint		0.6 – 0.7
Beaverboard	0.43	>1.2
Poplar wood	0.42	1.2
Wool fabric, worsted		0.31
Wool, undyed serge		0.08
Sole leather	0.63	> 0.9
Chrome leather	0.1	1.26
Sand (50 – 150 mesh)	5	0.05
Sand (200 – 325 mesh)	5	0.1
Sand (through 325 mesh)	5	0.21
Sea sand (on trays)	0.63	0.03
	1.27	0.047
	2.5	0.059
	5.0	0.06
Brick clay	1.6	0.14
Kaolin		0.14
Barium nitrate (crystals)	2.5	0.07
Carbon pigment	1	0.4
Copper carbonate	2.5 – 3.8	0.6
Iron blue pigment	0.63 – 1.9	1.1
Lithopone press cake	0.63	0.064
	1.27	0.08
	1.9	0.12
	2.5	0.16
Prussian blue pigment		0.4
Gelatin ($X_i = 4.0$)	0.25 – 0.5 (moist)	3.0
White lead		0.11
Rock salt	2.5	0.07

In Figure 11, measurements of the moisture content within a paper sample during drying are depicted. The width of the sample was $s = 30$ mm, its depth – measured from the surface which is in contact with the drying agent – is denoted by z. After 7 or 8 h, drying is still taking place in the first drying period. The surface of the sample is moist, the moisture content is almost evenly distributed over the entire depth of the sample. About 9 h after the beginning of the experiment, the critical point is reached. The surface of the sample dries out and the second drying period begins. After 65 or 95 h, drying takes place in the third drying period. All moisture is bonded, the drying rate is very small. The drying rate curve, associated with the experiment in Figure 11 is depicted in Figure 12 (curve d). Additionally, three other curves for thinner samples are shown. The thinner samples dry more rapidly in the second and third drying periods than the thicker ones.

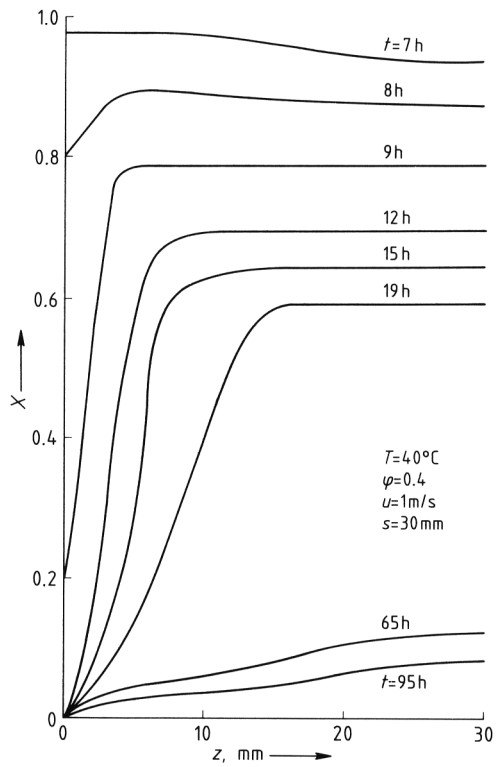

Figure 11. Moisture profile in paper stock as a function of the duration of drying [1]

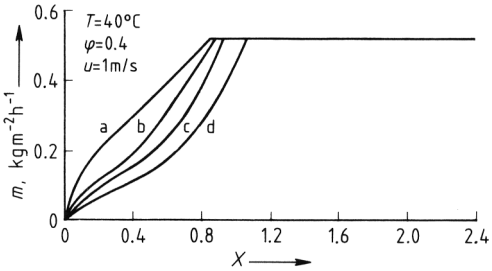

Figure 12. Drying rate curves for paper stock of various thicknesses [1]
Sample thickness s in mm a) 10; b) 15; c) 20; d) 30

The history of the temperature at various positions within a layer of powder is presented in Figure 13. In the first drying period the entire sample takes on the temperature T_{I} that can be read from a Mollier diagram. This temperature is usually referred to as the adiabatic saturation temperature. In the second and third drying periods, the temperature rises continuously in the various internal layers of the solid with the result that they approach the temperature of the drying agent. Large differences in moisture content and temperature can occur in the solid after the first drying period has been completed. Such differences can produce fissures in the solid, as well as scaling and discoloration.

Drying rate curves $\dot{m}(X)$ are usually determined experimentally. If the process is sufficiently slow, the sample can be periodically weighed. In rapid processes, such as the drying of paper by a transverse flow of air, which is completed in a matter of a few seconds, the humidity of the exhaust air can be determined continuously with an infrared spectrometer. Figure 14

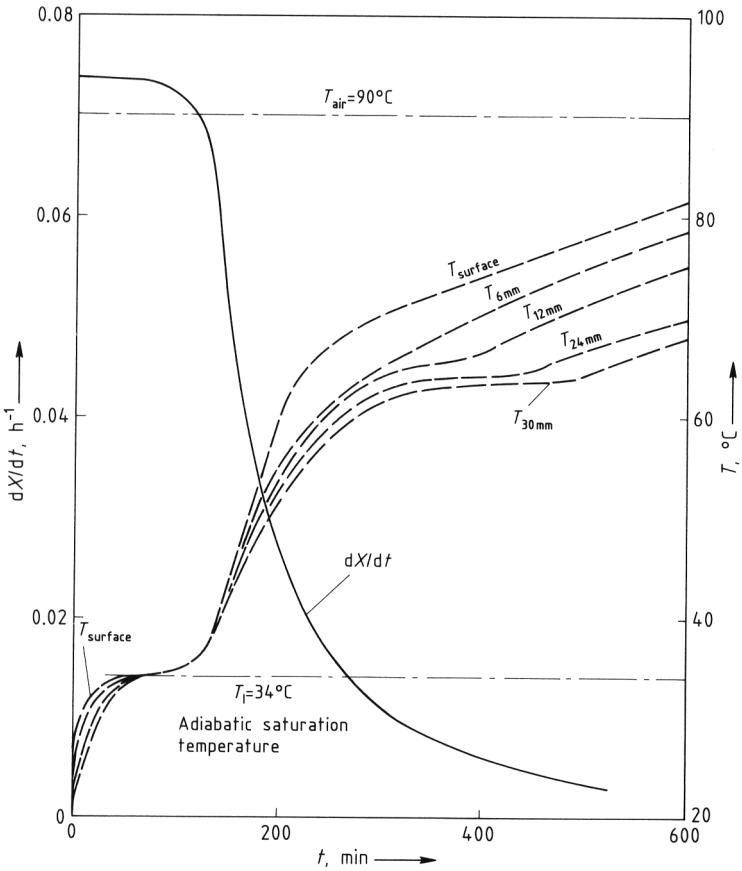

Figure 13. Drying rate curve and temperatures at various depths of the sample during the drying of powdered $CaCO_3$ [1]

shows drying rate curves that were obtained in this way for a textile fabric through which air was blown [4].

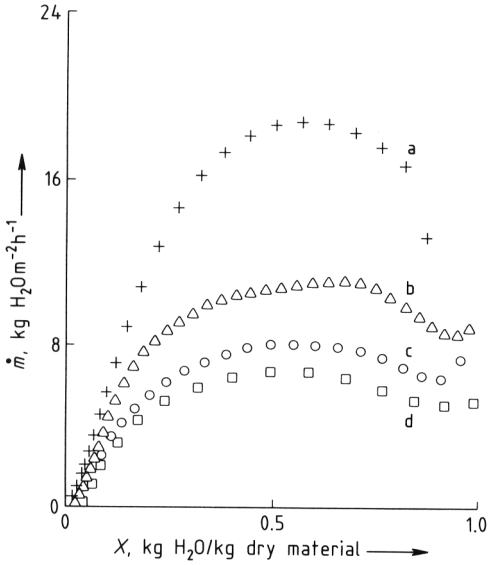

Figure 14. Drying rate curves for a textile fabric dried by passing air through it [4]
a) $T = 71.7\,°C$; b) $T = 39.4\,°C$; c) $T = 25.3\,°C$; d) $T = 22.8\,°C$

As already discussed, drying rate curves are obtained from laboratory tests that are conducted under constant drying agent conditions. In reality, the condition of the drying agent changes with time and location in any commercial drying apparatus. Because it is seldom possible to simulate all of the conditions that can occur in a production dryer, methods of interpolating the laboratory data are needed. In this context, the use of a drying rate curve is very helpful. The normalized drying rate is defined by

$$\dot{\nu} = \dot{m}/\dot{m}_{\mathrm{I}} \tag{23}$$

and the normalized moisture content by

$$\xi = \frac{X - X_{\mathrm{eq}}}{X_{crit} - X_{\mathrm{eq}}} \tag{24}$$

When $\dot{\nu}$ is plotted versus ξ, most drying rate curves measured in air under different conditions coincide (Figs. 15 and 16). In such cases, only one drying rate curve has to be determined by experiment.

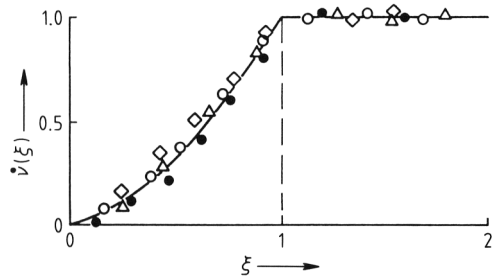

Figure 15. Normalized drying rate curve for pottery clay
The air temperature T and the relative humidity φ were varied [1]:
○ $T = 45\,°C$, $\varphi = 0.537$; △ $T = 15\,°C$, $\varphi = 0.537$
● $T = 25\,°C$, $\varphi = 0.187$; ◇ $T = 25\,°C$, $\varphi = 0.758$

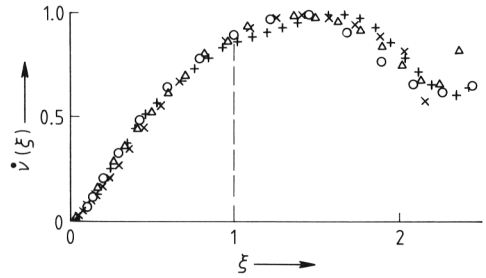

Figure 16. Normalized drying rate curve calculated from the data of Figure 14
× $T = 71.7\,°C$; + $T = 39.4\,°C$; △ $T = 25.3\,°C$; ○ $T = 22.8\,°C$

1.4. Drying Rate Curves for Contact Drying

In contact drying the heat necessary to vaporize the moisture in the solid is transferred by direct contact with a heated surface. The process of drying can take place in an atmosphere containing only the liquid's vapor, or air can also be present. Vacuum dryers are an example of the former, tube dryers for peat an example of the latter. Figure 17 shows a laboratory tray dryer equipped with an agitator. The entire experimental setup is placed on a balance and can be used in order to measure drying rate curves of porous, granular materials in vacuum. During each experiment the temperature of the heating surface T_{w} and the chamber pressure P are held constant. The same apparatus can be used in order to measure drying rate curves in the presence of inert gas.

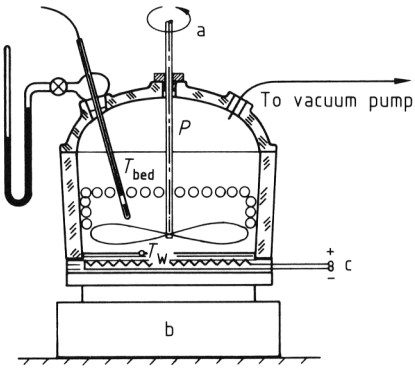

Figure 17. Laboratory tray dryer to determine the drying rate during contact drying
a) Mixer; b) Balance; c) Heater

The drying rate curves shown in Figure 18 were obtained with the apparatus of Figure 17 during vacuum contact drying of moist granular aluminum silicate; the particle diameter d was varied. Figure 19 gives similar results for peat in the presence of air (at normal pressure). In both cases, the drying rate depends upon the efficiency of the heat transfer between the heating surface and the granular bed, as well as upon the intensity of mechanical mixing. The reduction of the drying rate with decreasing moisture content is a consequence of the increased blocking of the heating surface by the particles that have already been dried. Consequently, the drying rate curve appears to be a property of the agitated bed as a whole, and not of the individual porous particles.

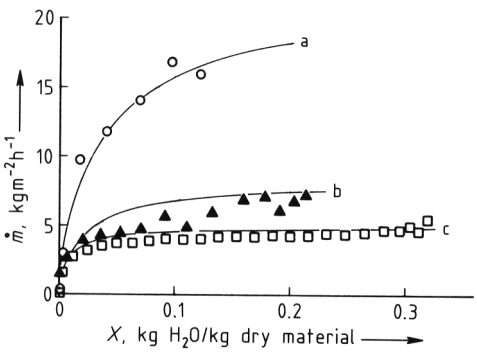

Figure 18. Drying rate curves for nonhygroscopic, vacuum-dried aluminum silicate particles
Conditions: $P = 2.63$ kPa, $T_w = 80\,°C$, and mixer speed = 45 rpm
a) $d = 0.83$ mm; b) $d = 3.25$ mm; c) $d = 6.60$ mm

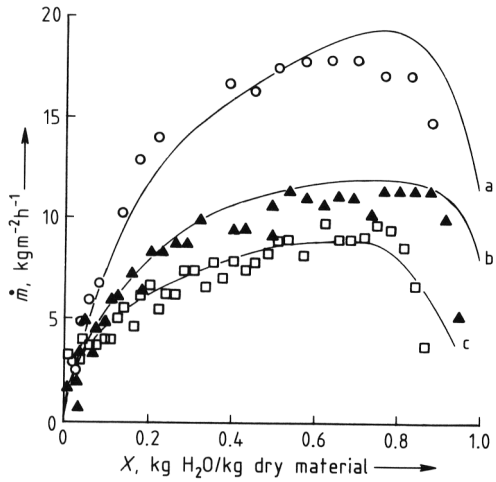

Figure 19. Drying rate curves for hygroscopic peat in the tray dryer of Figure 17
Conditions: $P = 0.1$ MPa (air), $T_w = 130\,°C$, and mixer speed = 40 rpm
a) $d = 0.75$ mm; b) $d = 3.0$ mm; c) $d = 6.0$ mm

Contact drying becomes faster as the temperature difference between the heated wall and the bed ($T_w - T_{bed}$) increases or the mixing speed is raised. For coarse-grained materials ($d \approx 10$ mm), the drying rate is directly proportional to the temperature difference, for fine materials ($d \approx 100\,\mu$m) it depends approximately upon the square root of this difference $(T_w - T_{bed})^{0.5}$. In contrast, the mixing intensity has a stronger influence on the drying rate of fine-grained materials than on the drying rate of coarse-grained ones. The physical explanation for this behavior can be found in [5]. The maximum drying rate that can be expected is given by

$$\dot{m}_{max} = \alpha_{max}\,(T_w - T_{bed})\,/\,\Delta h_v \tag{25}$$

The maximum possible heat-transfer coefficient, α_{max}, can be found from

$$\alpha_{max} = \alpha_r + \varphi_w \frac{4\lambda}{d}\left[\left(1+\frac{2\,(l+\delta)}{d}\right)\ln\left(1+\frac{d}{2\,(l+\delta)}\right) - 1\right] \tag{26}$$

with

$$l = 2\Lambda\left(\frac{2}{\gamma} - 1\right) \tag{27}$$

and

$$\alpha_r = 4C_{w,bed}\left(\frac{T_w + T_{bed}}{2}\right)^3 \tag{28}$$

λ	molecular thermal conductivity of vapor or inert gas
d	mean particle diameter
δ	surface roughness of the particles (in most cases $0 < \delta < 10\,\mu\text{m}$ [5], [6])
φ_w	fraction of the heated surface that is covered ($\varphi_\text{w} \approx 0.8$ for random beds of spherical particles)
Λ	mean free path of the vapor or inert gas molecules
γ	accommodation coefficient ($\gamma \approx 0.9$ for air or water vapor at moderate temperatures)
$C_{\text{w, bed}}$	coefficient for thermal radiation between the heated surface and the bed

The radiant heat-transfer coefficient α_r has a value of approximately $5\,\text{W}\,\text{m}^{-2}\,\text{K}^{-1}$ at room temperature. For high vacuum, the molecular conduction, the second term in Equation (26), is insignificant so that $\alpha_\text{max} = \alpha_\text{r}$ is obtained. If this limit is reached, the drying rate can only be increased if the temperature difference $(T_\text{w} - T_\text{bed})$ is made larger; in this instance, the degree of mixing has no effect on the drying process.

1.5. Drying Rate and Moisture-Composition Curves for Solids Wetted by Liquid Mixtures

A large variety of solids are moistened with liquid mixtures containing water and other substances such as alcohol, acids, or esters. In the course of drying such solids, the composition of the moisture generally changes because different liquids evaporate at different rates. This is termed *selective drying*, to denote the preferential removal of individual components from the mixture. However, under certain circumstances the composition of the moisture remains constant during drying (*unselective drying*). The properties of the dried item are often very strongly dependent upon the composition of the residual moisture. Moreover, there are accepted industry and health standards that the final moisture content must meet. Pharmaceutical items must not contain any toxic residues; foodstuffs should be dried to remove their water but not their (highly volatile) aromas. Although lacquers should dry with a constant composition, no fixed rules can be given because the requirements in individual cases differ.

The drying rate as well as the selectivity of the drying process are controlled by five interrelated physical mechanisms:

Phase equilibrium between the gaseous, liquid, and solid phases
Diffusion in the gaseous phase
Diffusion in the liquid phase
Capillary transport of moisture
Heat transfer

The combined influence of these five mechanisms on selectivity of drying has been theoretically and experimentally examined using simple liquid mixtures [7], [8]. According to the results of such investigations, the influence of vapor–liquid equilibria and gas-phase diffusion on the one hand, and the influence of capillary transport and liquid-phase diffusion on the other hand act together.

Furthermore, it is necessary to differentiate between liquid mixtures containing only volatile components and those containing at least one nonvolatile component. A mixture of alcohol, water, and glycerol is an example of the second kind. In all drying processes with a sufficiently high drying rate and a large thickness of the solid (large samples, coarse-grained products), the selectivity is determined by the capillary motion and the diffusion of the liquid constituents. This is the case when the dimensionless number

$$\Phi_1 = \frac{\delta_1}{v\,(d/2)\,1} \tag{29}$$

is considerably greater than unity. This parameter is equal to the ratio of two lengths: the penetration depth of the concentration profile, $\varepsilon = \delta_1/v_1$, and the depth of the sample, $d/2$. The velocity of the capillary flow v_1 is directly proportional to the drying rate $\dot{m}$, i.e., $v_1 = \dot{m}/(\varrho_1 \cdot \psi)$. (Note: ϱ_1 is the density of the liquid phase, δ_1 is the diffusion coefficient in this phase, and ψ is the porosity of the sample.) If the solution within the pores of the solid contains only volatile components, such as water, alcohol, or acetone, then for $\Phi_1 \ll 1$ "sharp" drying proceeds unselectively (in sharp drying the drying agent has a high temperature, and the drying rate is high). But should the liquid in the pores contain one or more nonvolatile components, then the volatile component that diffuses most quickly through the nonvolatile constituent of the mixture disappears preferentially. An example is the water in a solution with glycerol and alcohol. The relative volatility of the individual components, as well

as the speed of diffusion into the gaseous drying agent, is inconsequential in this case. This situation is important because of the possibility of supressing the loss of aroma during drying of foodstuffs.

If the dimensionless number Φ_1 is considerably larger than unity (low temperature of drying agent, low drying rate), then the selectivity is solely determined by the relative volatility and by the velocity of diffusion into the gaseous phase. The more volatile components disappear first, as long as "dynamic azeotropy" does not occur. The appearence of such azeotropic points is due to the interaction of relative volatility and gaseous diffusion. For the solution of isopropyl alcohol and water, the (static) azeotropic mole fraction of isopropyl alcohol is about 0.65. However, using dry air as drying agent a dynamic azeotropic mole fraction of 0.41 can be attained; the reason for this behavior is that water diffuses into the air much faster than alcohol. At this composition the drying process would be unselective for all values of the parameter Φ_1.

Figures 20 and 21 clarify the drying behavior of solids that are wetted by solutions with only volatile components. The experimental drying rate and composition curves for a clay cylinder are shown in Figure 20. The clay has been soaked in a solution of isopropyl alcohol and water. The drying agent is dry air ($T = 60\,°C$, $u = 0.2$ m/s). Because of the relatively large diameter of the sample, the parameter Φ_1 remains much lower than unity during the first drying period. Consequently, the process is controlled by the diffusion in the liquid phase. In the absence of an inert liquid component the drying process occurs unselectively, i.e., the mole fraction of the alcohol, $\tilde{x}_1$, remains constant, as illustrated in Figure 20B. This situation changes in the course of the second drying period: with decreasing moisture content, the liquid – vapor interface retreats inside the solid, so that the drying rate and the width of the moist region diminish. The result is that Φ_1 increases. Towards the end of the drying process a pronounced change in the composition of the remaining moisture is recorded. The extent and trend of this selectivity are dictated by the diffusion occurring in the gaseous phase, and by the thermodynamic equilibrium. For an initial mole fraction of alcohol $\tilde{x}_{1,i} > 0.41$, the criterion for dynamic azeotropy, the solution becomes depleted of water, and for $\tilde{x}_{1,i} < 0.41$, alcohol is lost.

The region of selective drying can be considerably extended if the thickness of the sample is decreased. In this manner, the drying of a thin-walled hollow cylinder made of sintered bronze with dry air ($T = 60\,°C$ and $v = 0.2$ m/s), occurs selectively from the beginning (see Fig. 21).

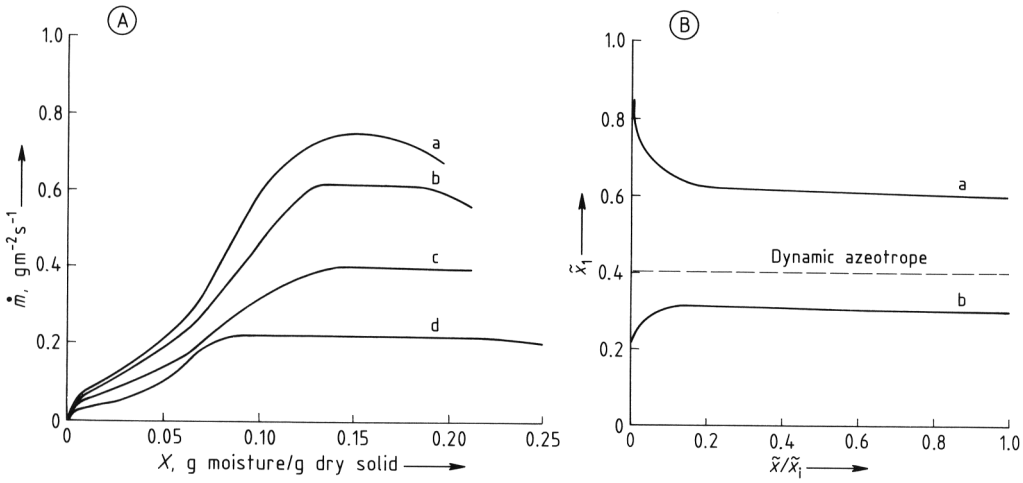

Figure 20. Drying rate (A) and composition curves (B) for a clay cylinder wetted by isopropyl alcohol ('1') and water
Conditions: diameter of sample = 39.2 mm, length = 95.7 mm; drying agent: dry air, $T = 60\,°C$, $u = 0.2$ m/s [7]
A) a) $\tilde{x}_{1,i} = 1.0$; b) $\tilde{x}_{1,i} = 0.6$; c) $\tilde{x}_{1,i} = 0.3$; d) $\tilde{x}_{1,i} = 0.0$
B) a) $\tilde{x}_{1,i} = 0.6$; b) $\tilde{x}_{1,i} = 0.3$

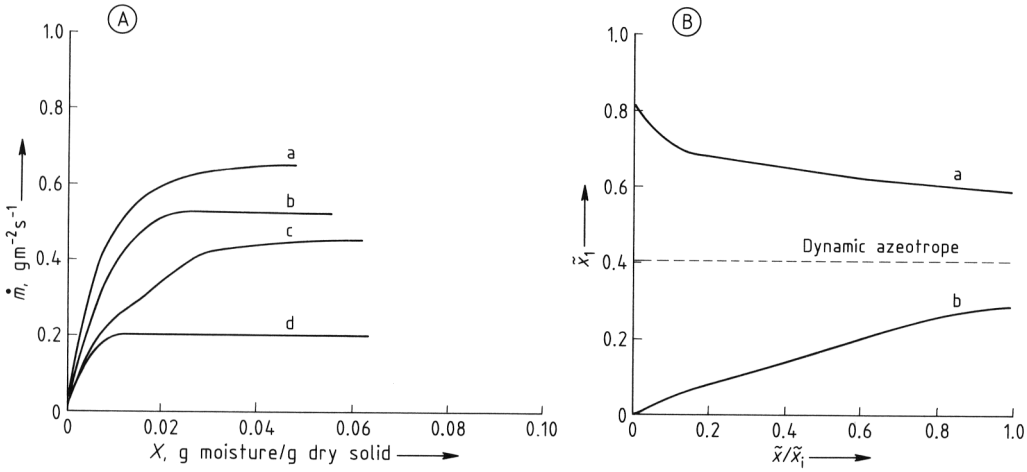

Figure 21. Drying rate (A) and composition curves (B) for a hollow cylinder of sintered bronze wetted by isopropyl alcohol ('1') and water
Conditions: outer diameter of sample = 31.0 mm; inner diameter of sample = 30.0 mm; length = 95.7 mm; drying agent: dry air, $T = 60\,°C$ and $u = 0.2$ m/s [7]
A) a) $\tilde{x}_{1,i} = 1.0$; b) $\tilde{x}_{1,i} = 0.6$; c) $\tilde{x}_{1,i} = 0.3$; d) $\tilde{x}_{1,i} = 0.0$
B) a) $\tilde{x}_{1,i} = 0.6$; b) $\tilde{x}_{1,i} = 0.3$

2. Drying Methods and Dryer Types

In this chapter the most important methods of drying are arranged according to the way the heat is transferred. In a convection dryer the liquid is vaporized by the heat that is transferred from the drying agent. In the case of a contact dryer the heat is conducted from a heated surface to the solid. The solid may be transported over the heat-transfer surface, or it may rest upon it. Radiant heating, in which the heat is supplied from a radiation source that is remote from the surface of the solid but with an unobstructed view of it, is also used. Special methods of drying include dielectric drying and freeze drying. The myriad of dryer types is a consequence of the different behavior that the solid exhibits during drying, the particular product needs, and many economic considerations.

2.1. Convection Drying

The methods of convection drying differ from one another in the manner by which the moist solid contacts the drying agent, which is usually hot air.

2.1.1. Flowing Gas

Drying with a flowing gas is particularly suitable for materials which should not be mechanically stressed during drying.

Drying Oven (Kiln). With small quantities of moist solids, the simplest, cheapest dryer is a drying oven similar to that shown in Figure 22.

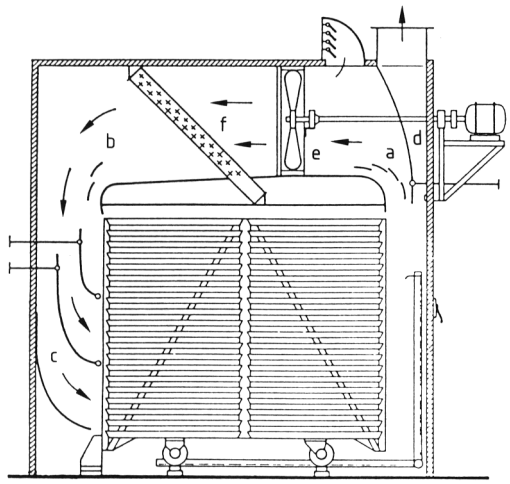

Figure 22. Compartment dryer (Courtesy of Babcock BSH)
a) – c) Turning vanes; d) Air-exhaust duct with damper; e) Fan; f) Heaters

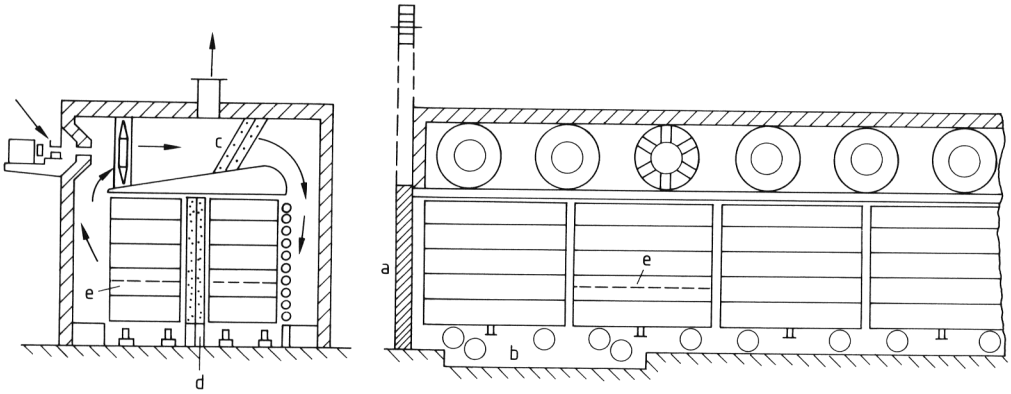

Figure 23. Double-truck dryer for two rows of carts
a) Loading door; b) Cart pool; c) Heater; d) Intermediate heater; e) Racks for loading

The solid is placed upon racks or in trays that are mounted on a cart. A fan circulates warm air through a heater and then through the drying racks or trays. Flow dividers ensure that the air is uniformly distributed. Drying proceeds uniformly and without overdrying portions of the material. The drying process for forced convection dryers is controlled by the amount of exhaust and intake air, as well as its velocity and temperature. Such dryers are employed for processing sensitive materials that require long drying times, i.e., gentle drying. The kilns for some grades of lumber can be 200 m^3 or larger. The atmosphere in such chambers can best be regulated by a feedback system with suitable sensors.

Tunnel Dryers. Large quantities of materials are dried in a tunnel in a continuous process. The cart is placed within the tunnel at its entrance and conveyed, usually continuously, but sometimes step by step, along the tunnel to its exit. The moving air can be cocurrent or countercurrent with the direction of the solids.

Sometimes the flow of the drying agent is changed, and it is blown perpendicular to the solid's direction of motion. This can lessen the excessive drying of edges for materials such as plasterboard, and permits different drying conditions in different portions of the tunnel (Fig. 23). Applications of this drying method are freshly lacquered chassis parts and glass plates that have a layer of leather glued to them; they are carried on moving hooks through the drying tunnels. Goods that are in the form of large sheets, such as plasterboard or wood particle board, are moved by a conveyor system consisting of many parallel rollers that are driven by an interconnecting chain. Many layers of such a roller system can form decks to use the tunnel's space effectively.

Spiral Belt Dryer. A dryer of this type (Fig. 24) is often ideal for materials that require a long, undisturbed drying time. The moist solid is placed upon a circulating belt at a position outside the dryer. The solid remains undisturbed while it dries. The belt enters at the top of the dryer and moves in a spiral fashion toward the bottom. Several blowers rotate about the vertical axis of the dryer and service a particular elevation of it. The air is blown past the solid, onto the heating pipes, and then sucked back across the solid. It is possible to regulate both the temperature and air speed of each vertical drying zone separately.

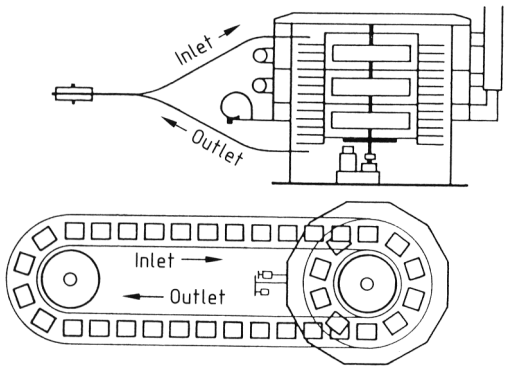

Figure 24. Spiral belt dryer (Courtesy of Babcock BSH)

Segmented Rotating Tray (Wiped-Tray) Dryer. The continuous drying of large quantities of crystalline, granular, or pasty materials that must be dried gently can be effectively done with a segmented rotating tray dryer of the type shown in Figure 25. The solid can have a broad grain-size spectrum. A lazy Susan that slowly rotates around the central axis of the dryer is its main component. Each tray is subdivided into pie-shaped sections by radial slots – 60 – 120 mm wide – through which the material can fall. The moist solid is placed upon the topmost tray in an even layer. A stationary wiping arm comes into contact with the dried surface material and pushes it circumferentially along the disk toward the radial slots. The material that falls through a slot lands upon the disk below, where it initially forms a heap or ball. An arm is used to spread the material into a uniform layer, and the procedure continues until the dried material reaches the bottom of the dryer, where it is removed. Pasty and sticky materials can form clumps in the early stages of drying. In order to achieve uniform drying, such clumps are subdivided between each disk by an appropriate device.

In the dryer shown in Figure 25, the blowers are centrally located and force the air over the material on one disk and back again over another disk after the air passes through a heating unit. The dryer can have several zones for the circulating air, each having a different temperature. The dimensions for these types of dryers range from 1.2 m in diameter with 8 m^2 of useful drying surface to 10 m in diameter with 1500 m^2. The rotational speed of the dryer is specified by the drying time and the number of vertical stages.

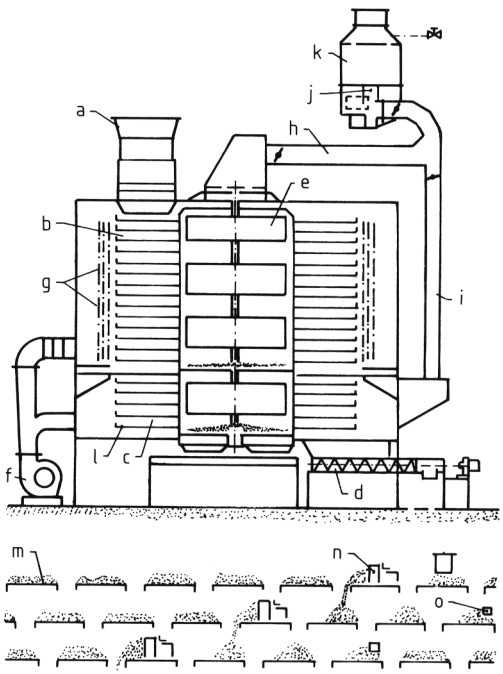

Figure 25. Segmented rotating tray (wiped-tray) dryer
Top: vertical, axial section
Bottom: Detail of the wiping and distributing method
a) Loading device; b) Drying zone; c) Cooling zone; d) Unloading conveyor; e) Turbo fans; f) Fresh air fan; g) Heaters; h) Exhaust duct; i) Cooling air duct; j) Wet washer; k) Demister; l) Segmented tray; m) Segment; n) Wiper; o) Distributor

Disking Dryers. A circulating disking dryer is advantageous for materials that must be constantly turned over. Like the wiped-tray dryer, it contains a set of vertically stacked, circular trays. In addition, fixed disks of different diameters, in effect, disk-harrow the material on the trays. In this way, the material that is to be dried is both agitated and transported. Attached to the drying chamber is a separate chamber for the fans and heaters to supply and heat the air. An important application of this kind of dryer is for materials moistened by volatile solvents.

Jet Dryers. Dryers of the type shown in Figure 26 blow hot air out of slots or circular holes at a high speed onto the surface of the moist solid. As a consequence of the high air speeds resulting from large rates of circulated air, very large mass-transport coefficients and high drying rates of the liquid near the surface, i.e., the first stage of drying, are achieved. This minimizes the redistribution of the moisture within the solid because of the increased exchange of

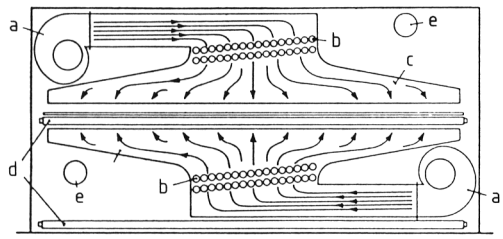

Figure 26. Impinging jet dryer (Courtesy of Deutsche Babcock Anlage)
a) Fresh air fan; b) Heaters; c) Nozzle boxes; d) Roller system loaded with strip material; e) Exhaust

mass. Jet dryers are mostly used to dry flat materials such as wood veneer, cardboard, foils, textiles, or photographic papers.

Continuous sheets of materials, such as coated papers, foils after printing, or photographic film, are transported through the dryer on a cushion of air that can be over 100 m long. Pieces of veneer are placed onto a conveyor belt; extruded pellets, for example, either dry or wet feed, are transported on a continuous stainless steel belt. The speed of the drying air must be adjusted so that the pieces are not blown away. A tentering frame dryer has hooks or clamps for holding the cloth that are attached to a chain that moves through the dryer. The material is laterally constrained in this way, with the result that the fabric undergoes a favorable tensile treatment while drying.

2.1.2. The Solid is Aerated

If air can be blown through a layer of solids, or if the solid can be processed to make this possible, aeration is the most efficient form of convection drying. Because nearly all of the surface of the moist particles is exposed to the drying agent, maximum heat and mass transfer are achieved for the available conditions. High drying rates are possible even under relatively mild thermal and mechanical conditions.

Through-Circulation Batch Dryer. Agricultural products are often dried by this device. Grain or hay are placed into a container that has a perforated bottom, or onto a rack, and warm air is blown through it. Uniformity of the drying within a pile of cereal can be facilitated with stirring forks that simultaneously transport the material.

More than 100 t/h of grain can be continuously dried in inclined tubular dryers that also aerate the material. There are dryers that resemble silos, and these are filled from the top. The grain slides past roof-like inserts (Fig. 27), as it moves toward the bottom where it is removed from a funnel-shaped collector.

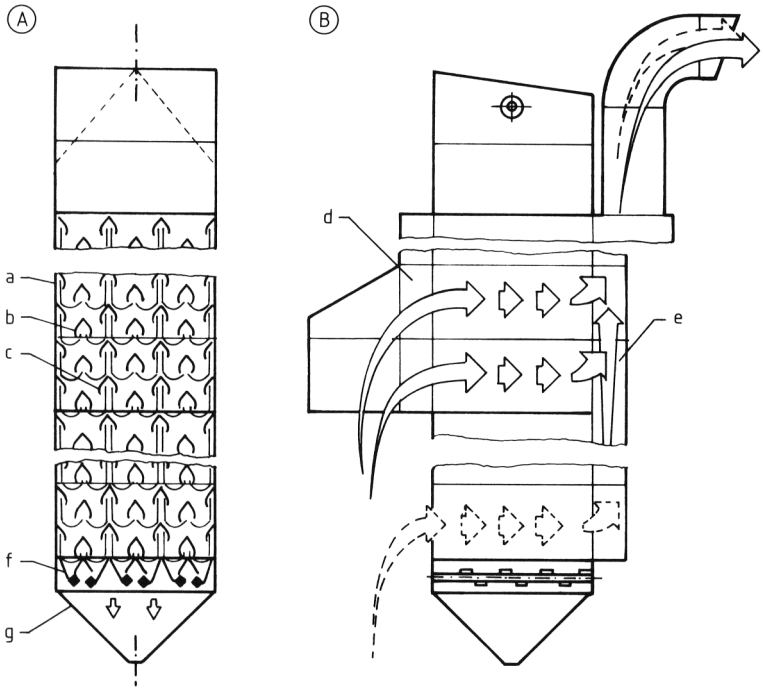

Figure 27. Vertical gravity bed continuous flow dryer
A) Construction of the dryer; B) Air flow (view perpendicular to that in A)
a) Chute; b) Fresh air; c) Exhaust; d) Inlet air duct; e) Exhaust duct; f) Unloading device; g) Funnel

The gas that is used as the drying agent flows through a duct and into the layer of grain via the roof-shaped forms. The gas is subsequently sucked back through similar V-shaped openings that are placed somewhat higher. The uppermost portion of the dryer is not ventilated and is called the sweating zone. Here the material slides past heaters that are usually filled with hot water. The sweating of the grain considerably reduces the necessary drying time. In the lower portion of the duct, a cooling zone is often provided. Fodder is dried with air diluted with (warm) exhaust fumes, but legumes and grains used for food or producing oil are dried with air that is indirectly heated to prevent contamination.

Aerated, Stationary Rack Kiln. Small quantities of fragile materials can be placed on racks that can be put into a kiln. Air is blown through the material. Such ovens are loaded and unloaded by hand.

Aerated, Moving Rack Kiln (Simplizior Dryer). The Simplizior dryer (Fig. 28) is a semicontinuously operating rack dryer that is mainly used by the food industry to treat vegetable products such as parsley, carrots, spinach, mushrooms, as well as spices. Because the moisture content of the drying air increases on its way through the dryer, the racks are moved through the dryer countercurrently to the air. They are loaded with the moist goods and installed in the upper part of the dryer. Then they move toward the bottom, where they are unloaded. The racks are finally loaded with new moist goods, and sent back to the top of the dryer to begin another drying round. The motive power can be supplied by a hand crank in small units, while large installations have power systems that are automatically controlled. In this way the material in all of the racks reaches the same final condition, and the process approaches the efficiency of a fully continuous process.

The dryer that is illustrated in Figure 28 requires removal of the racks above the intermediate heating unit and their reinsertion below it. Intermediate heating of the air is necessary because the air cools after passing through the lower racks, and because very high drying temperatures cannot be used due to the sensitivity of the goods. (The Simplizior dryer is also produced without an intermediate heater.) Useable surface areas up to 60 m^2 are common.

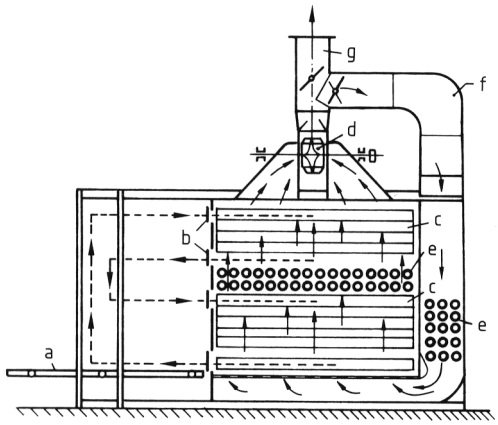

Figure 28. Through-circulation floating rack compartment dryer
Simplizior (Courtesy of Deutsche Babcock Anlage)
a) Elevator; b) Flaps; c) Rack piles; d) Fan; e) Heaters; f) Recirculation duct; g) Exhaust duct
– – → Path of solid; ⟶ Path of air

Belt Dryers. In belt dryers a loading device that is especially designed for the product is used to place the moist solid on the surface of a belt which passes through a drying chamber that resembles a tunnel. At the end of this chamber the material falls from the belt into a chute for further processing. In some installations the material falls onto another belt that moves in the opposite direction to the first one. Depending upon the characteristics of the material to be dried, multiple passages through the dryer are possible. In this way the material is mixed while drying, and new surfaces are exposed to the drying agent. A shorter, more gentle drying process is the result.

Centrifugal or axial flow blowers are used to aerate the moist materials. The air stream can enter the solid from below or above. The drying agent in belt dryers is commonly supplied laterally so that they are operated as convection dryers with partly recirculating air. The drying agent is heated indirectly with steam or hot water. Fuel oil or natural gas can also be used to indirectly or directly heat the drying agent.

Stepwise heating of the solid is possible with a unidirectional belt dryer, and this allows further control of the drying process. Belt dryers are ideal for friable, molded, granular, or crys-

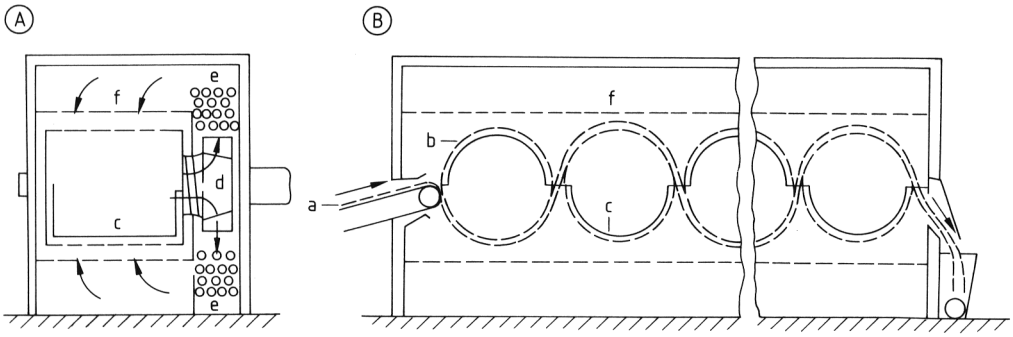

Figure 29. Through-circulation perforated drum dryer for loose, fibrous materials (Courtesy of Fleißner Egelsbach)
A) End view; B) Side view
a) Loading conveyor; b) Perforated drum; c) Cover plates; d) Fan; e) Heaters; f) Air distributor

talline products. They are utilized in all branches of industry.

Aerated Perforated Drum Dryers. This type of machine is employed to dry solids that can form porous layers on curved surfaces. Examples of such materials are cellulose fibers, wool, and cotton. The dryer consists of a series of closely spaced drums with diameters ranging up to 2 m and lengths of 6 m, which turn about a horizontal axis (Fig. 29). The cylindrical surfaces of these drums are made of perforated sheet metal. The material to be dried is transported to the top of one drum and is transferred to the bottom of the next; it moves to the top of the following drum, etc. It is held against the drum by a flow of high velocity air that is sucked into the drums through guides and hemicylindrical baffles at their periphery.

Air-permeable paper is produced on rotating cylinder dryers that consist of only one drum with a surface made of metal screening. The efficacy of these dryers is dependent upon the porosity of the layer of moist solid, the maximum permissible air temperature, and the vacuum that can be achieved within the drum which, in turn, determines the quantity of air that passes through the paper. It takes only seconds to dry the paper in this way.

Rotating Drum Dryers. The drum of a rotating drum dryer (Fig. 30) is longer in relation to its diameter than that of the perforated drum dryers and rotates about a horizontal axis that is slightly inclined. Consequently, the material that is loaded at the upper end slowly wanders toward the lower end, where it usually falls into a loading hopper. Longitudinal protrusions (flights) inside the kiln distribute the material more or less uniformly over the peripheral surface for drying and promote transport in some instances. These protrusions are often shaped like crosses or quadrants and reduce the drying time in two ways: (1) they constantly stir the solid and (2) intermittently cause it to fall away from the wall as a shower of fine particles through the drying agent, which streams axially along the kiln. Some of these lifting flights do not perform well when the solid is quite dry because they create too much dust, but are very effective for pasty or sludge-like materials. If the drum must be cleaned often, or blockage of the drum can occur, flights that have a relatively simple shape are preferable.

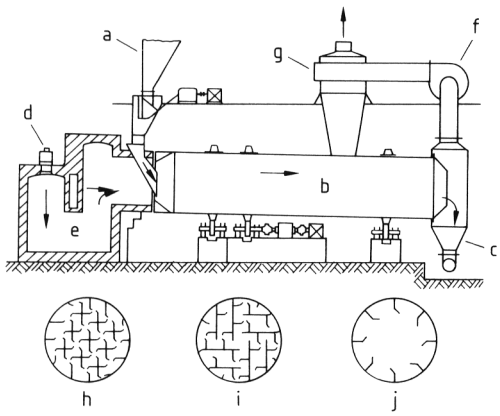

Figure 30. Cocurrent direct-heat rotary dryer and various flight arrangements
a) Feed chute; b) Drum; c) Discharge; d) Burner; e) Combustion chamber; f) Exhaust fan; g) Dust separator; h) Cross-shaped flights; i) Quadrant flights; j) 45° lip flights

The moist material and the drying agent flow either cocurrently or countercurrently. This type of kiln is often direct-fired. They have diameters between 0.3 and 6 m and are used principally for granular and crumbly materials. Pasty substances and slurries are often transformed into crumbs in a short time after they come in contact with the hot drying agent. Some liquid or semi-liquid materials form small clumps when mixed with an already dried portion of the same materials. Then they can be dried with this kind of dryer. To increase the residence time in rotary drum dryers, circumferential baffles are used.

Roto-Louvre Dryer. This particular type of drum dryer has internal guide vanes (Fig. 31) that convey the drying medium in crosscurrent to the moist solid in very close proximity. The solid is simultaneously well stirred. The channels for the drying gas are tapered toward the rear of the dryer to achieve a good flow distribution. This results in a conical dryer, which does not need to be inclined.

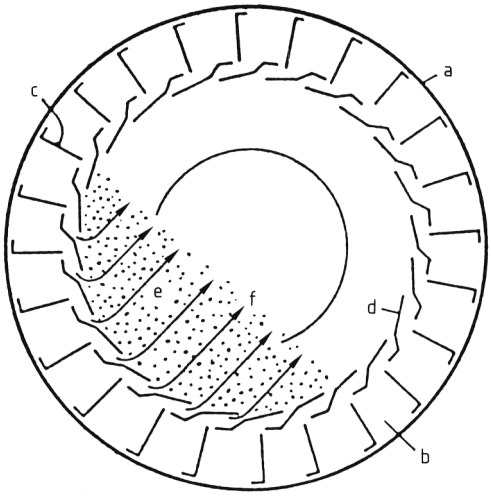

Figure 31. Cross section through a Roto-Louvre dryer (Courtesy of Dunford & Ellioth, London)
a) Rotating drum; b) Inlet ducts for hot air; c) Radial guide vanes; d) Tangential guide vanes; e) Moist solid; f) Air flow through the solid

Aerated, Double-Screw Dryer. The moisture in clumps of materials and coarsely fibered or flaky granules can be extracted in an aerated double-screw dryer (Fig. 32). Two parallel screws that rotate closely together convey the solid over a perforated floor through which the drying agent flows upward. The dryer can be completely sealed so that the drying agent flows in a closed circuit to facilitate the recovery of volatile solvents.

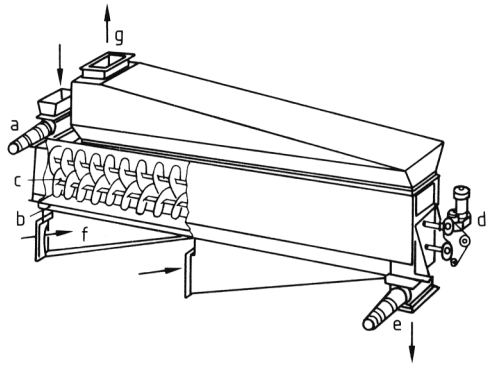

Figure 32. Twin-screw dryer (Courtesy of Werner Pfleiderer Maschinenfabrik, Stuttgart)
a) Product feed valve; b) Fine screen plate; c) Conveyor screws; d) Screw drive; e) Product discharge valve; f) Fresh air distributor duct; g) Exhaust air connector

2.1.3. Large-Scale Agitation of the Solid

Spouting Bed Dryers. These units are noted for their high drying efficiency. One model is illustrated in Figure 33. The solid moves horizontally in a chute and the drying agent flows vertically through a perforated floor to fluidize the solid. These machines can operate continuously, because the solid that enters the dryer via an adjustable opening is transported through it while suspended in the drying agent. The throughput can be increased by belts fitted with scoops to collect and move the material.

To achieve fluidization of the solid with minimum air velocity, vibrating fluidized beds are used. The dryer is mounted on a chute that vibrates under the action of a shaking device. The lengthwise oscillations of the chute serve to transport the material. The residence time of the solid within the dryer can be modified by adjusting the amount of vibration. If the airflow is divided into two sequential streams, the first can be used to dry the solid, while the second can cool it for further processing or packaging. Versions of these kinds of dryers that have a closed circuit for the drying agent can be used to recover solvents.

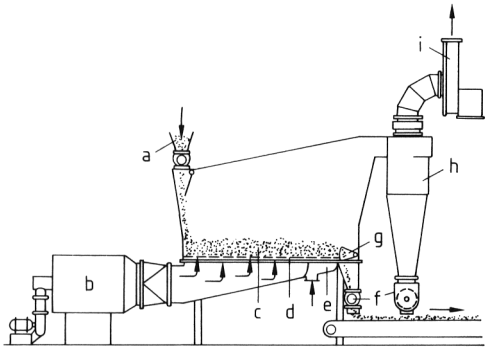

Figure 33. Fluidized bed dryer (Courtesy of Büttner-Schilde-Haas)
a) Fluidized bed; b) Heater; c) Solids feeder; d) Gas distributor; e) Cooling zone; f) Discharge; g) Control baffle; h) Dust separator; i) Fan

Materials that can be suspended in the drying agent are usually powders, crystals, and granular or short-fibered products that remain finely divided or have only a slight tendency to stick or cake. Pastes and slurries are mixed with previously dried material and are then easily fluidized. Solutions and suspensions can be dried by spraying them into the swirling layer with nozzles. The minute solid particles that are present are then covered by a thin film of liquid, which drys quickly, primarily due to the surface evaporation that characterizes the first period of drying. The small particles grow in size by conglomerating with others. Particles are added to the dryer when it is operating continuously to act as seeds. Small dried particles are recirculated; large particles are pulverized, classified, and recirculated. Particles between 0.5 and 5 mm in diameter can be produced from various solutions in this way.

Spin-Flash Dryer. Pastes and high viscosity liquids are often dried best in the type of dryer that is pictured in Figure 34. The dryer is loaded by a screw conveyor or a pump. The material to be dried is first stirred and broken up by a multi-armed paddle that turns at 50 – 500 rpm. Air is fed in tangentially at the bottom. A kind of fluidized bed is produced in the region of the stirrer. The air flow entrains the smaller, drier particles, and the larger, wetter ones fall back into the stirred region, where they are reduced in size.

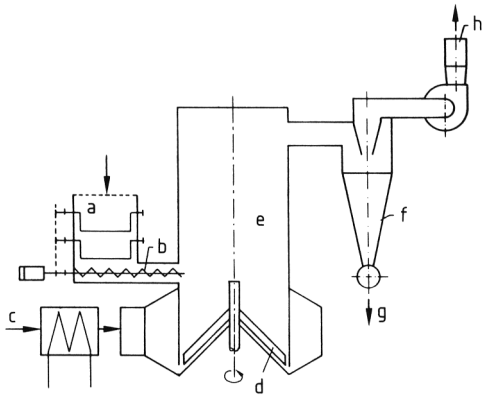

Figure 34. Spin-flash dryer (Courtesy of Anhydro AS, Soborg-Kopenhagen)
a) Moist solid bin; b) Screw feeder; c) Air inlet; d) Stirrer; e) Main drying chamber; f) Material separator; g) Product discharge; h) Exhaust exit

Centrifugal Dryer. Figure 35 shows a centrifugal dryer whose motion serves to subdivide the moist solid and bring it into effective contact with the drying agent. The dryer consists of a stationary closed vessel that has at its base a centrifuge, or possibly two if the unit is large. They whirl the moist solid around in the dryer where the drying agent, usually air mixed with the combustion products of coal, oil, or natural gas, can perform its drying function. The fumes are diluted with enough fresh air to attain the needed inlet temperature and to increase the gas volume sufficiently so that the humidity does not become too high.

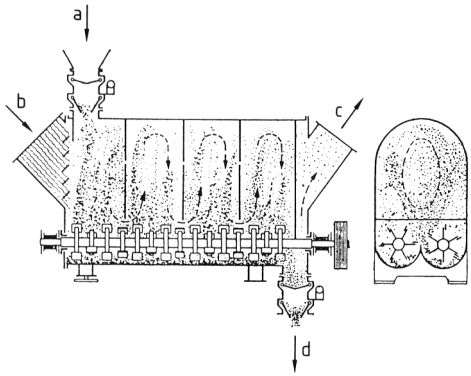

Figure 35. Centrifugal dryer (Courtesy of Hazemag, Münster/Westfalen)
a) Loading zone; b) Hot gas inlet; c) Exhaust outlet; d) Material discharge

There are also dryers that combine the mechanical and aerodynamical methods to entrain the solid in the vortical motion of the drying agent. Machines that swirl the solid are used to dry granules of plastics, salts, coal, and other chemicals.

Centrifugal-Impact Dryer. The unit pictured in Figure 36 functions as a centrifugal dryer and an impact pulverizer. Materials that have an edge length of less than 500 mm and a moisture content of less than 30 % are reduced to a grain size between 0 – 10 mm and dried in gases that reach 900 °C. Within the dryer a rotating wheel with impact bars catapults the particles onto hardened plates, which shatter the clumps of moist solid. The pulverized pieces of the moist solid are dried in a stream of hot gases. The smallest particles are entrained in the stream and pass out of the dryer with the drying agent and must be recovered. The coarser particles fall to the bottom of the dryer, where they are removed. Dryers of this type are often used in the sand and gravel industry.

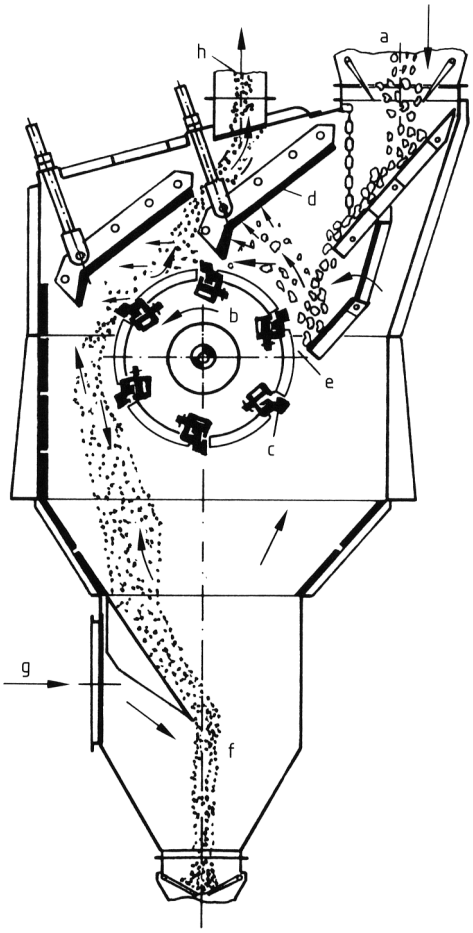

Figure 36. Centrifugal-impact dryer (Courtesy of Hazemag mbH, Münster/Westfalen)
a) Moist solid loading; b) Impact rotor; c) Impact bar; d) Striker plate; e) Adjustable slit; f) Discharge collector; g) Hot-air inlet; h) Exhaust gas exit

2.1.4. The Solid Moves in the Drying Agent

Pneumatic Conveyor Dryers. Materials that can be pneumatically transported can be dried simultaneously. The simplest form for a dryer encompassing this dual role for the drying agent consists of a vertical tube in which granular or pulverized materials are dried while suspended in a gas or air stream (Fig. 37). The available drying time is only a few seconds: only fine materials, with their high rates of heat and mass transfer, or coarse products, with only surface moisture to be removed, are used in such dryers. Solids that contain internal moisture can only be dried to a limited extent by this method. Sometimes such materials can be dried in a multistaged gas-lift dryer.

The drying agent and the solid move cocurrently in air-lift dryers. The moist solid comes into contact with the hottest drying gases when it first enters the dryer. Because the solid is completely submerged in the hot drying gas, the conditions for high heat- and mass-transfer rates are good. The solid has the wet bulb temperature as long as surface evaporation is dominant. Only when this phase of drying is completed does the temperature of the solid begin to rise. This warming process is limited, however, because the drying agent has been cooled by evaporating the moisture. Therefore, sensitive materials can be dried without damage in such dryers, even though the gas inlet temperatures are relatively high. The limited residence time is also a factor. Both organic and inorganic salts that have some moisture after having been centrifuged or filtered can be dried effectively with pneumatic conveyor dryers. Plastic powders, granules, foodstuffs, fodder, wood chips, sand, and quartz can also be dried in this way.

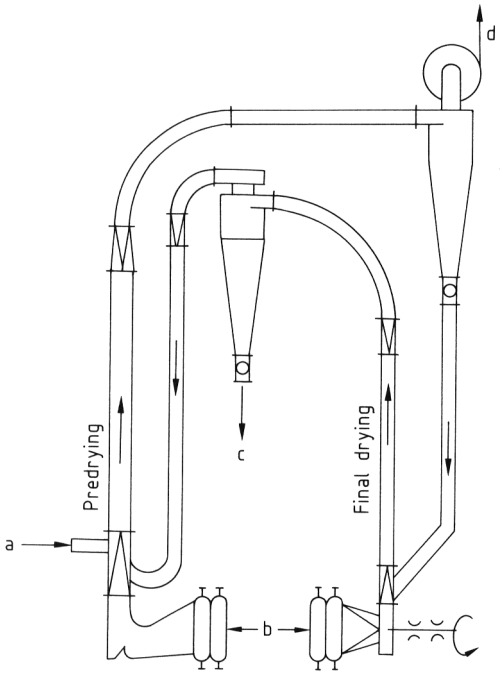

Figure 37. Two-stage pneumatic conveyor dryer (Courtesy of H. Orth, Böhl/Pfalz)
a) Wet feed; b) Fresh air; c) Material discharge; d) Exhaust

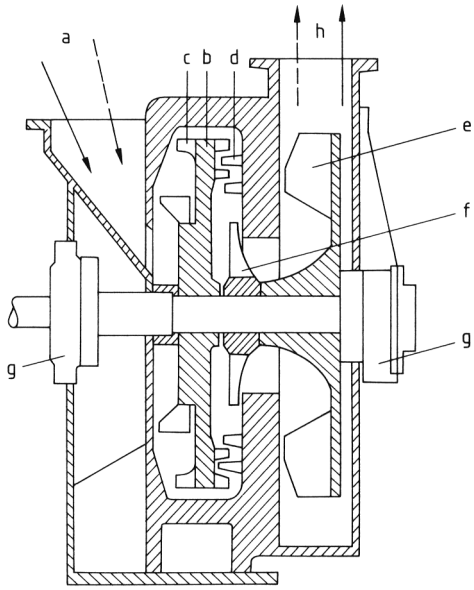

Figure 38. Atritor mill dryer (Courtesy of Alfred Herbertg, Coventry, UK)
a) Moist solid and heating gas; b) Rotor; c) Hammer segments; d) Stator; e) Fan; f) Finger separator; g) Bearing; h) To the separator
– – → Heating gas; ⟶ Moist solid

Mill Dryers. As the name of the dryer implies, in mill dryers (Fig. 38) the moist solid is simultaneously ground and dried. Although the energy consumption of these devices is relatively large, even that used for grinding the solid ultimately appears in the form of the heat that is necessary for drying. Hot gases also flow through the dryer to achieve the desired drying times. The mill shown in Figure 38 has a loading chute in which the moist material slides to the front surface of a rotating grinding wheel. The particles that this disk produces are carried away by the hot gases into a second grinding chamber of the dryer. Here the solid and entraining gas pass through a series of stationary and rotating pegs that further reduce the particle size. The gases and solids are ultimately forced out of the dryer by the blower.

This kind of dryer can produce powder from granular, caked, or partially liquid materials that can have a moisture content up to 80 % and a representative size of 50 mm. Mill dryers are used to dry and comminute peat clumps at power plants.

2.1.5. The Material is Sprayed

Spray drying is used for the drying of pastes, suspensions, or solutions. The moist material is sprayed into the drying agent, and is converted into a powder that is entrained by the gas stream. The volatile liquids vaporize quickly. The gas and the dry powder can be separated at the exit of the dryer.

In this kind of drying process, a uniform fog of moist material should be produced within the dryer. Each type of solid affects the design used to produce a uniform product. The most important types are the disk atomizer (Fig. 39A) and pressure nozzle (Fig. 39B, C). (→ Spraying and Atomizing of Liquids).

Centrifugal disks atomize liquids by extending them into thin sheets, which are discharged at high speeds from the periphery of the rapidly rotating (4000 – 15 000 rpm), specially designed disks (diameter 50 – 350 mm). The speed depends upon the desired size of the particles. This type of atomizer is used for pastes and suspensions because these materials would damage

nozzles by abrasion, or even clog them. Very thick pastes can be treated in disk atomizers but they require a high-pressure pumping system to introduce them into the dryer.

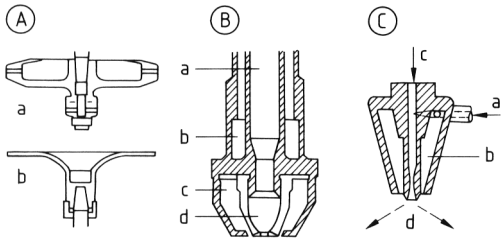

Figure 39. Devices for spraying
A) Disk atomizer
a) Suspended installation; b) Standing installation
B) Pressure nozzle
a) Fluid delivery; b) Heating or cooling jacket; c) Nozzle support; d) Nozzle orifice
C) Two-fluid nozzle
a) Tangential high-pressure air supply; b) Conical expansion and swirl chamber; c) Liquid supply line; d) Spray zone

Powders are produced from a spray nozzle, either a simple nozzle operating under sufficient liquid pressure or a two-fluid nozzle that concurrently sprays streams of liquid and gas – usually air. The resulting mist contains individual particles with diameters between 20 and 300 µm.

Figure 40 shows a two-fluid nozzle that exploits the fact that the speed of sound (critical speed) in a gas – liquid mixture is much smaller than in the gas or liquid phase. The two fluids leave the mixing chamber, which is at a low pressure, and enter an atmosphere where the mixture suddenly expands. The shock wave that occurs at the end of the mixing chamber causes the liquid – with suspended solids in some cases – to be distributed into a fine mist with a narrow size spectrum.

Because the liquid is transformed into minute drops, a very large surface area is placed in contact with the drying agent. Very short drying times – from fractions of a second to a maximum of a few seconds – are the result.

One or more powderizing systems can be installed at the top of a cylindrical structure, which is often as high as 20 m and several meters in diameter. Hot gases are introduced into the drying tower at the bottom or the top. The final product falls into a conical hopper at the bottom of the dryer, where it is unloaded.

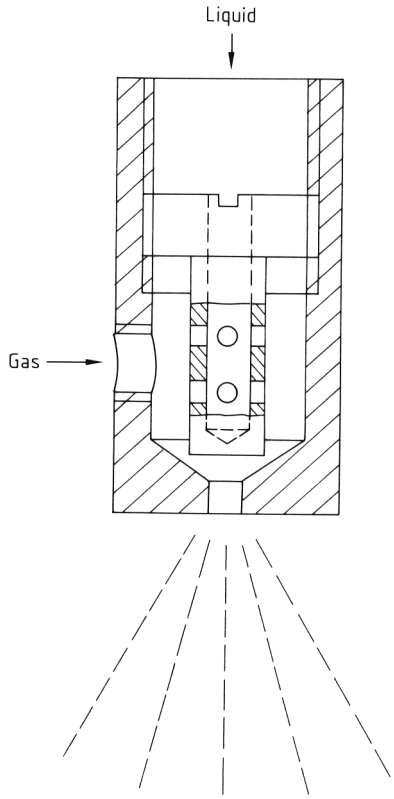

Figure 40. Two-fluid sonic nozzle
(Courtesy of Caldyn, Ettingen, FRG)

Spray dryers are particularly suitable for drying solids that are temperature sensitive. Examples are milk products, baby foods, eggs, blood and blood plasma, pharmaceutical products, chemicals, dyes, plastics, glues, tannin, and soaps. An advantage of these dryers is the high solubility of the powders that are produced.

2.2. Contact Drying

In contact drying, the heat is directly transferred to the solid from a heated surface upon which it rests. The solid may be either stationary or be continually transferred from one hot surface to another.

2.2.1. Flat and Strip Materials

Materials that are flat or in strips, such as textiles, paper, or cardboard, are dried with a drum dryer. The moist material is wrapped around a

series of rotating, horizontal cylinders that are usually heated internally with steam. The steam and its condensate enter and leave the cylinder through hollow-journals, often the same journal. Materials that have a low strength when moist, e.g., felt and cotton, move between the drums on endless belts. The belts press the solid against the drying drums and simultaneously absorb some of the moisture. This moisture must be subsequently removed from the belts (Fig. 41). The possibility of introducing air between the drums and the moist material in rapidly rotating drum dryers is reduced by the pressure that the belts exert against the material. Consequently, the contact resistance to heating is reduced, and the heat transfer between the drum and product is improved.

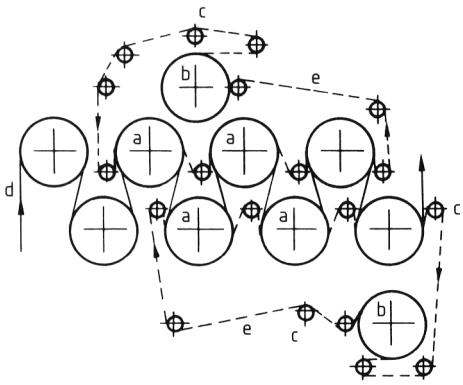

Figure 41. Schematic of a multicylinder dryer with guiding belt
a) Drying cylinders for the moist solid; b) Drying cylinders for the guiding belt; c) Guide rollers; d) Moist material; e) Guide belt

2.2.2. Low-Viscosity Materials

Roller Dryers. In order to contact dry solutions of organic or inorganic material, a continuously operating roller dryer is almost always used (Fig. 42). They are usually selected because of their effective utilization of heat.

The material flows onto the drying roll as a thin layer. The liquid is vaporized by the heat coming from the heated rollers. Shortly before the dried solid reaches the end of the dryer, it is peeled away from the hot surface by a scraper to produce a film, coarse or fine flakes, or a powder. Some of the methods of applying the moist material to the rollers are shown in Figure 43. Each one has its particular uses, depending upon the adhesion and consistency of the moist solid.

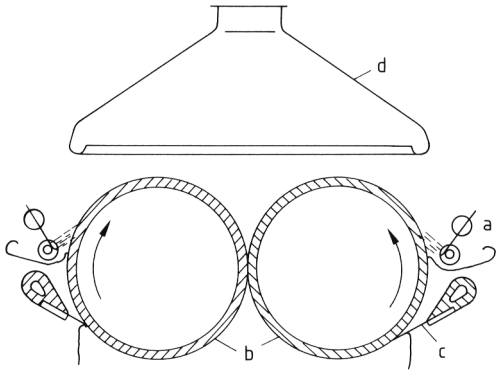

Figure 42. Double-roller spray dryer (Courtesy of Escher Wyss)
a) Spraying apparatus; b) Drying rollers; c) Scraping knife; d) Moisture exhaust

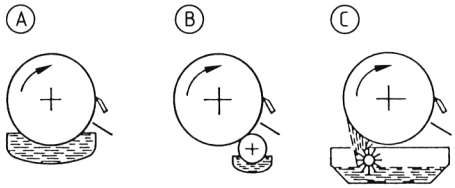

Figure 43. Methods for feeding the moist solid for a roller dryer
A) Submerged roller; B) Distributing roller; C) Barbed roller

The rollers are usually heated internally with steam, although hot water or heat carrier oils are sometimes used. Induction-heated drying cylinders are also available; they can produce controllable surface temperatures between 40 and 400 °C.

Hoods can be placed over the dryer to reduce the ambient pressure and facilitate drying. The dryer can also be completely encased to make it airtight and dustproof. In some applications where increased drying efficiency is needed, the roller is additionally heated with a stream of hot air or exposed to a radiant heater (Fig. 44).

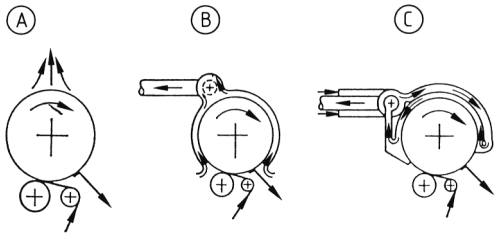

Figure 44. Aerating a single-roller dryer
A) Dryer without a hood; B) Dryer with hood and exhaust; C) Dryer with hood, warm air supply, and exhaust

2.2.3. Pasty Materials

Roller dryers are also effective for drying pasty or creamy materials if they can be satisfactorily applied to the rollers. A proven single-roller device is shown in Figure 45.

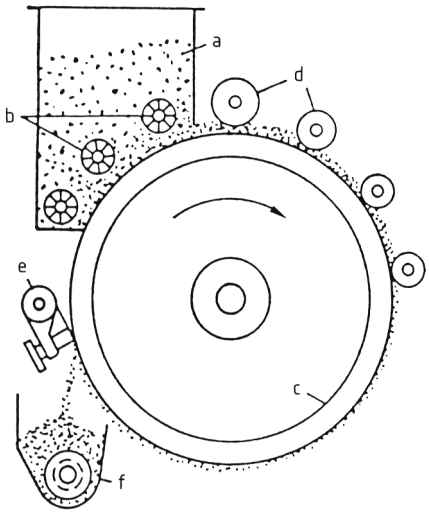

Figure 45. Single-roller dryer for mealy materials
a) Feed container; b) Stirrers; c) Drying roller; d) Pressurizing rollers; e) Scraping knife; f) Discharge

Grooved Roller Dryer. A special type of roller dryer reduces the moisture of pastes by 8 – 10 % and produces short segments for further drying steps. The paste is not rolled out into a thin layer, instead it is forced by a special roller into the grooves of the drying roller. As the roller turns, the material is pressed and thickened. Finally, a serrated scraper removes the dried material. To dry the solid completely a belt dryer is added. The drying of materials such as stearates, carbonates, clay, kaolin, white lead, or titanium dioxide are some special applications of this kind of dryer.

Hollow-Screw Heat Exchanger. Hollow-screw heat exchangers (Fig. 46) have proved useful for the continuous drying of paste. This dryer consists of a trough that contains one or two hollow screws. The screws turn in the same or the opposite directions, and partially overlap. Screws turning in the same direction are preferred for pastes because of the greater overlap and the self-cleaning action. Hot water, saturated steam, or hot organic liquids serve to provide the heat.

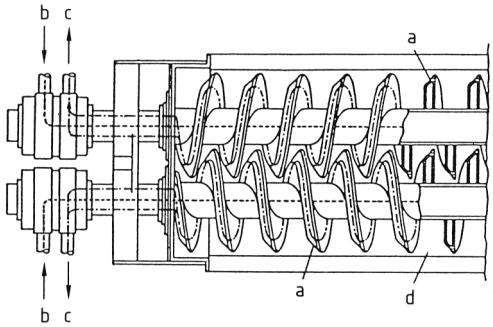

Figure 46. Hollow-screw heat exchanger (Courtesy of Lurgi)
a) Hollow-screw conveyors; b) Inlets for the heating agent; c) Outlets for the heating agent; d) Trough

List AP Dryer (All Phase Dryer). The dryer shown in Figure 47 can be used for materials that can range from fluid or pasty to putty-like or granular during the drying process. It is also suitable for substances that are sticky or form crusts. A suitably designed horizontal housing surrounds the main roller, which has disks attached to it. This roller rotates in the opposite direction to a cleaner roller that has rectangular hoops that stir the material. The cleaner roller's speed is four times that of the main roller, whose surface is cleaned by it. Kneadable solids adhere to the disk protrusions on the main roller and scrape against the housing to scour and clean it. The throughput of the device can be controlled by the orientation of the rollers. The housing, its lid, the hollow main roller, and the disk elements are heated with steam, hot water, or hot oil.

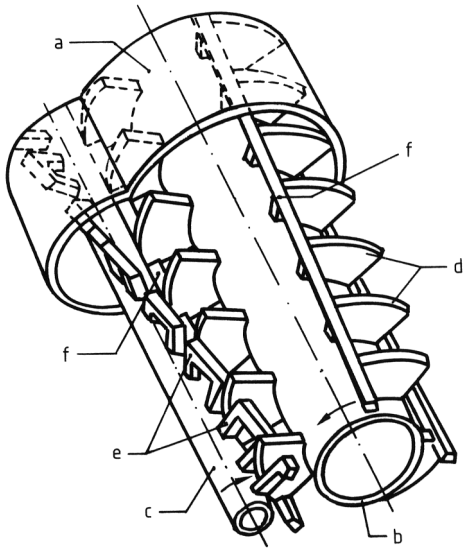

Figure 47. Contact dryer with stirrer and kneader (Courtesy of List, Prattein)
a) Double-walled container; b) Main shaft; c) Cleaning shaft; d) Cleaning disk; e) Square-hoop stirrer; f) Kneading bars

is provided by the flow of vapor that ensues in the tubes.

Screw Conveyor Dryers. Inside the drum, screw conveyor dryers (Fig. 48) have a rotor that is fitted with many small paddles that constantly force the moist solid into contact with the hot wall, as well as move it along the axis. The residence time of the solid in the dryer is regulated by changing the angle of the paddles.

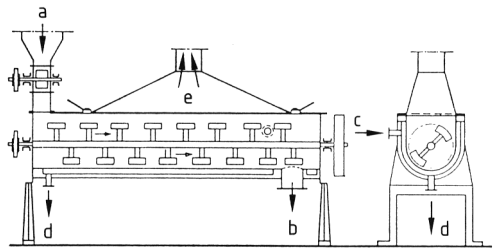

Figure 48. Screw conveyor dryer (Courtesy of Büttner-Schilde-Haas)
a) Moist solid loading; b) Dried solid discharge; c) Inlet for heating agent; d) Outlet for heating agent; e) Solvent vapors

2.2.4. Granular Materials

Rotary Dryer. The heat necessary for drying is transferred through the peripheral walls of a rotary dryer in contact drying. Only a small amount of air is necessary to carry off the moisture that is taken from the solid. Accordingly, the air velocity in these units is quite low. This is advantageous when drying materials that dust easily or form dust during drying. If necessary, ducting can be installed to introduce gases into the drum where the solid is drying.

Rotary dryers are preferred for drying washed ores, coal, and cement. Such materials are constantly turned over and mixed by lifting flights attached inside the drum.

Rotary contact dryers are produced in two main types. The first is constructed with integral heating pipes within the drum. The solid falls in a shower past these hot pipes once during each revolution of the drum. In the second type, the moist solid is in tubes that are attached to the walls of a drum into which steam is fed. The steam condenses on the surface of the tubes, heating the moist solid within. The drum is inclined to the horizontal so that gravity transports the material down the tubes, but some assistance

Cone-Worm Dryer. A cone-worm dryer operates along the inclined wall of a vertical funnel-shaped container. The chamber is sealed and equipped with a heating jacket. The screw serves to transport the material to the top and to produce a gentle stirring of the moist solid.

Tray Dryers. When contact drying is employed, the trays (Fig. 49) are heated. Different heating intensities on the different trays provide the optimum conditions for drying. In this way, for example, the material can be cooled in the lowest levels of such dryers. The vapor is removed by warm air flowing countercurrently.

Spiral Tube Dryers. The drying of solids that range from granular to powdery is facilitated by spiral tube dryers (Fig. 50). The material is only momentarily in contact with the heated surface of the tube and is transported pneumatically. Residence time in the dryer is but a few seconds. These dryers require only a small amount of gas. Therefore, the drying of solids wetted by solvents can be done very economically with inert gases flowing in a complete circuit that includes a solvent recovery system.

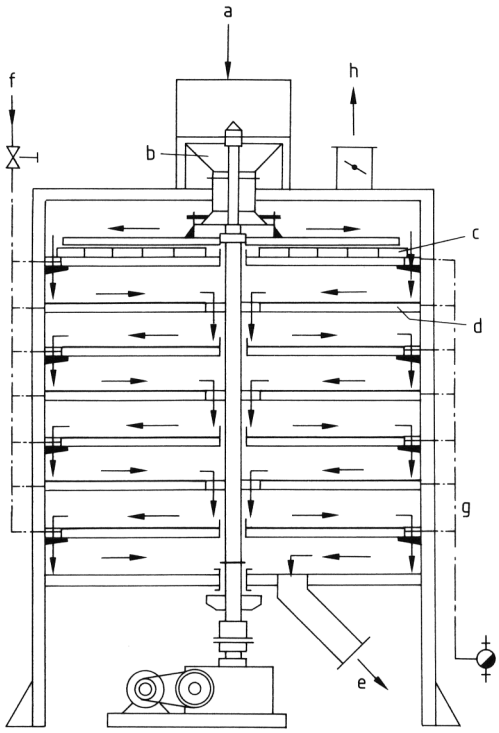

Figure 49. Tray dryer (Courtesy of Büttner-Schilde-Haas)
a) Moist solid loading; b) Rotating tray feeder; c) Rakes (only drawn for topmost level); d) Heated tray; e) Product discharge; f) Steam; g) Condensate; h) Exhaust

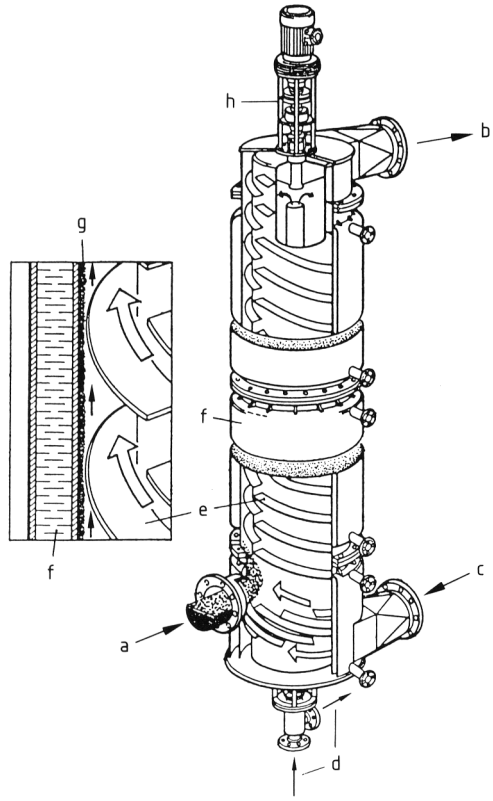

Figure 50. Spiral tube dryer (Ruhrchemie system, greatly shortened) (Courtesy of Werner & Pfleiderer)
a) Moist solid; b) Dry solid; c) Transporting air; d) Heating liquid; e) Air baffles; f) Heating jacket; g) Product path; h) Compactor drive

Spiral tube dryers find application in processing plastic granules, starches, bread crumbs, and copper powder.

2.3. Radiant Heat or Infrared Drying

Very short drying times are possible with infrared radiation sources that transmit large amounts of energy per unit of surface area to the solid in the form of heat. This method of drying is costly because of the electrical energy. It is only economical for drying long runs of similar materials that have a sufficiently thin surface film.

The intensity and emisson spectrum of the radiation must be optimized with respect to the moist solid to keep the energy costs low and to prevent excessive heating of the solid. Ovens with conveyor systems are often divided into zones to achieve this optimization; the individual radiation sources can be adjusted to match the conditions of the solid (→ Radiation Heating).

The main use of this method is the drying of painted surfaces because warming occurs rapidly, forcing the solvents within the paint to move to the surface, which inhibits the formation of an undesirable surface skin. This method of drying is also important for drying thin-walled ceramic products, which can be dried quickly but without cracks by this method.

A combination of radiant and contact drying is effective if the drying time or the size of the dryer must be reduced. An example of such a combination dryer is shown in Figure 51. The dryer is provided with a means for jet drying a woven belt and a radiant heater to predry the material. The rapid drying of the fabric surface

hinders the undesirable migration of dyes and other chemicals to the surface.

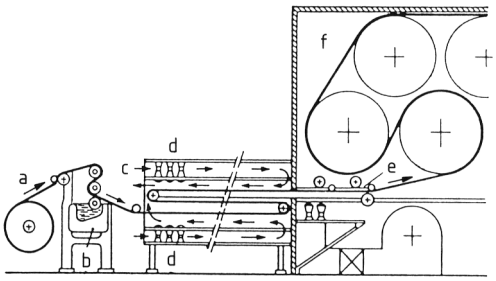

Figure 51. Predryer for textiles (Courtesy of Philips)
a) Fabric; b) Impregnating trough; c) Air inlet; d) Infrared radiators; e) Sheet-metal reflector; f) Main dryer

2.4. Dielectric Heating

Drying with high-frequency heating has an advantage over other methods of drying: the temperature within the moist solid rapidly rises. The internal temperature can be maintained at a prescribed value regardless of the surface temperature. When wood is dried by this means, the moisture distribution can be adjusted so that the surface is always more moist than the interior. As a result, compressive stresses are induced in the surface fibers, strengthening the product, whereas the other means of drying tend to weaken the material and produce cracks. Only a few hours are needed to dry wood and ceramic goods with high frequency (→ Electrically Generated Heat).

Relatively high energy and installation costs are associated with this method of drying so that it is only cost effective to dry particularly valuable materials such as special hardwoods, large ceramic pieces, or temperature-sensitive foodstuffs and gourmet items.

2.5. Vacuum and Freeze Drying

Almost all dryers that operate at atmospheric pressure and supply the necessary heat by conduction or radiation can be converted to a vacuum dryer. The principal difference in the vacuum dryers are their seals and the means to produce the vacuum. Continuously operating vacuum dryers require, in addition, special devices for loading and unloading. Drying under reduced pressure is advantageous for materials that are temperature sensitive or easily decomposed because the vaporization temperature is reduced. Also, drying times in vacuum dryers at the maximum temperature are shorter. When drying materials containing organic solvents in vacuum dryers, the solvents are recovered more economically than with convection dryers because the humid drying agent need only be cooled to the condensation temperature of the solvent at the reduced pressure. Vacuum dryers are most often used to dry pharmaceutical products and foodstuffs. If the substance to be dried is poisonous or potentially explosive, vacuum drying lessens the emission of poisonous gas or vapor into the environment.

The simplest form of a vacuum dryer for batch drying is the vacuum shelf dryer. The moist solid lies on a heated plate. Improved heat transfer with higher efficiency is obtained in the vacuum tumble dryer (Fig. 52), in which the moist solid is constantly agitated and mixed. An example of a continuously operating vacuum dryer is a roller dryer with one or two rollers. Air-lift dryers that operate at reduced pressures are effective when large amounts of materials must be dried at low temperature and only a relatively small amount of moisture must be removed.

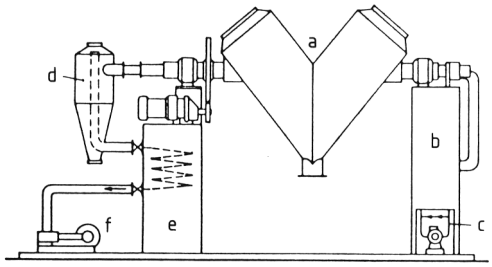

Figure 52. Vacuum tumble dryer (Courtesy of Patterson-Kelley)
a) Rotating container; b) Heating boiler; c) Pump; d) Dust separator; e) Condensor; f) Vacuum pump

Freeze Drying. Freeze drying is characterized by the removal of water (sublimation of ice) from the solid at temperatures below 0 °C and reduced pressure. Freezing halts nearly all the chemical and biological processes in the

material so that it is biochemical, physiological, and therapeutic characteristics remain essentially unchanged. Accordingly, this method of drying is used to process high-value foods, fruit juices, spices, tea, coffee, pharmaceuticals, virus and bacteria cultures, vaccines, and preparations containing protein.

A schematic diagram of a freeze dryer is shown in Figure 53. Usually the freezer is a separate first stage and then the solid is vacuum-dried, but in some systems one apparatus performs both functions. Shallow pans that contain the solid are heated by conduction in the dryer. The vapors condense as ice upon reaching the cooling surfaces. These surfaces must be thawed at the end of the process when the condenser is isolated from the drying chamber by closing a valve. The dryer is then evacuated with the aid of a diffusion pump.

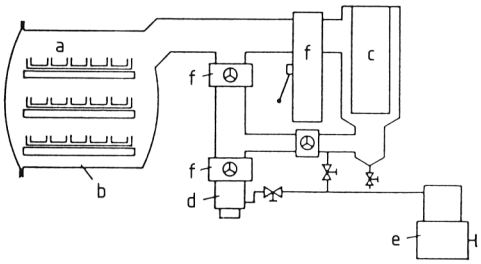

Figure 53. Oven vacuum drying (Courtesy of E. Leybolds) a) Heated trays with moist solid; b) Drying oven; c) Condenser; d) Diffusion pump; e) Booster pump; f) Shutoff valves

The continuous operation of a freeze dryer that processes foodstuffs is accomplished with a vacuum disk dryer. Such a unit is characterized by considerably shorter drying times than a vacuum chamber dryer (→ Refrigeration Technology).

3. Selecting, Sizing, and Energy Requirements of Dryers

3.1. Choosing the Type of Dryer

The choice of a particular dryer type is mainly determined by the characteristics and production rate of the material that requires drying. Often several types would be suitable so that operating and economic aspects become the decisive criteria.

3.1.1. The Role of the Material Properties of the Solid

The mechanical properties of the moist solid when it arrives at the dryer determine the ways the material can be loaded into and transported through the dryer. Lumpy, granular, and crumbly materials can remain stationary (as in simple drying chambers, Simplizior, tunnel, or screw dryers), or they may be moving in respect to their supporting elements (as in fully-automated rack dryers, belt dryers, air-lift dryers, drum dryers, cyclones, and spiral tube dryers). Liquefied and pasty materials can be converted into fine drops (spray dryer), or thin layers (roller dryer), or they can be transformed into a crumbly material by partial drying (grooved roller dryer).

The thermal sensitivity of the material determines the temperature at which the heat can be transferred safely and the residence time in the dryer. If low temperatures are necessary, moisture removal can be achieved with a vacuum dryer. Small residence times are possible in forced convection dryers, spiral tube dryers, and spray dryers.

Materials that are chemically insensitive can be dried in direct contact with the hot drying agent, either air or combustion products. If the moist material is chemically sensitive, an inert gas must be used as the drying agent or some other heating method must be applied.

In the event that an organic solvent is present, it may be economical to recover it. In some cases the solid or the moistening liquid may be flammable or toxic. In such cases, completely enclosed dryers, such as vacuum dryers or special contact dryers, are to be recommended.

3.1.2. Production Rate

The throughput of the solid determines whether the dryer is to be operated in a continuous or batch mode. Batch operation requires a high energy consumption and worker effort per unit mass of dry product, but installation costs are relatively low. Nevertheless, this method of drying (a simple drying oven or chamber) is often

economically viable for products with a small throughput or often changing properties.

When it is necessary to dry a well-specified product at a high production rate, a continuous dryer is advantageous. Even though the initial costs of such installations are high, the standardized physical and geometric characteristics of the product permit optimization of the energy requirements for this mode of operation. Moreover, the manpower necessary to operate the system is quite small. Most of the common dryer types are able to operate in the continuous mode. The efficiency of a dryer is largely determined by the amount of moisture that can be removed per unit area or volume of the dryer, and the amount of heat that is necessary to vaporize the liquid on a unit mass basis. Table 2 gives information about these parameters.

Table 2. Vaporizing effectiveness and heat consumption of dryers

Dryer	Vaporizing effectiveness, kg $H_2O\,m^{-2}\,h^{-1}$ *	Heat consumption, kJ/m^{3}h vaporized H_2O
Air-lift dryer	$d = 0.5$ mm: 100 kg m^3 h^{-1}	
(d = particle diameter)	$d = 1$ mm: 20 kg m^3 h^{-1}	3700 – 8800
	$d = 5$ mm: 4 kg m^3 h^{-1}	
Belt dryer		4000 – 5000
Centrifugal dryer	90 kg m^3 h^{-1}	3300 – 3500
Compartment dryer	0.1 – 12 (ventilated)	
	0.1 – 15 (flown over)	5600 – 13 000
Cylinder dryer	7 – 25	2900 – 5700
Drum dryer	25 – 50 kg m^3 h^{-1}	4000 – 8000
Kiln	20	5000 – 5800
Roller dryer	5 – 60	2900 – 5700
Rotating steam tube contact dryer		3800 – 6300
Roto-Louvre dryer	35 – 75	3500 – 10 000
Screw conveyor dryer	5 – 15 (high humidity)	
	0.5 – 2.5 (low humidity)	3400 – 5600
Spiral tube dryer	500 kg H_2O m^3 h^{-1}	3400
Spray dryer	1.5 – 48 kg m^3 h^{-1}	4600 – 11 000
Tray dryer	4 – 8	
Vacuum chamber dryer	0.15 – 1	2900 – 4600
Vacuum paddle dryer	10 – 15	2900 – 4200
	0.4 – 0.6 **	10 000
Wiped tray dryer	3.5 – 4	3500 – 9200

* If not noted otherwise.
** With intensive terminal drying.

3.1.3. Dryer Ventilation

The ventilation of a dryer is important, and the various possibilities are given according to a scheme by GÖRLING in Figure 54.

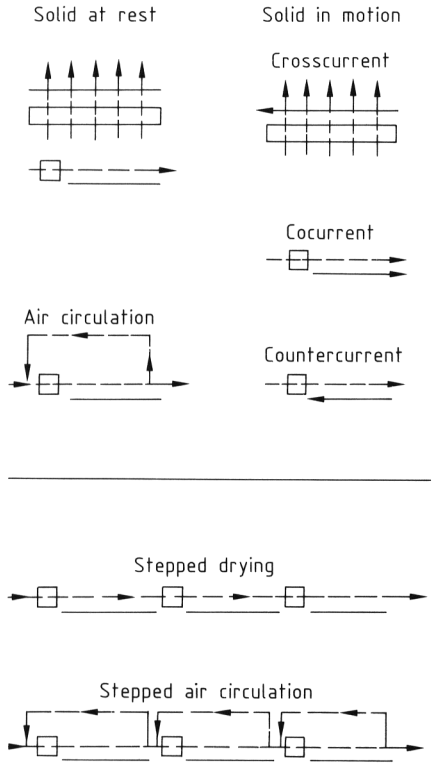

Figure 54. Various arrangements between solid, heaters, and drying agent
--, ⟶ Solid; – – → Drying agent; □ Heaters

Stationary Solids. Solids that remain stationary in a convection dryer usually encounter a flow that is parallel to the surface on which they rest. However, it is possible to blow the air through a layer of granular material. This requires more energy to realize a given amount of air flow, but shorter drying times are possible so that the total energy consumption can remain satisfactory.

The usual manner of using air as a drying agent is to constantly circulate it past the moist solid while adding only a fraction of fresh air to the system. This is economical and minimizes the possibility of uneven drying or overheating the solid, which otherwise usually occurs near the air inlet duct. Through such a circulating system, the drying process can be satisfactorily matched to the characteristics of the solid by adjusting the temperature and speed of the air. Moreover, the influence of the daily atmospheric conditions are easy to compensate with this system.

Moving Solids. If the solid is moving in the dryer, the stream of air can be flowing in the same, opposite, or perpendicular direction to produce cocurrent, countercurrent, or crosscurrent drying, respectively.

In *cocurrent drying* the hot, dry air encounters the solid in its most moist condition at the dryer's entrance. As a consequence of the large temperature and humidity differences between the solid and the air at the beginning of the drying process, the drying rate is high. However, as a result of increasing humidity and decreasing temperature of the air along its path, the drying rate decreases continuously and becomes quite small near the end of the dryer. Thus with cocurrent drying, a small residual moisture content in the solid is not realizable with economical amounts of drying agent. Cocurrent drying is economically attractive if the dry solid is sensitive to high temperatures, the moisture removal does not have to include an extensive part of the hyroscopic moisture, and the initial high temperature in the dryer is not deleterious to the properties of the solid.

In *countercurrent dryers* the hot, fresh air comes in contact with the solid when it is at the end of its drying process; the cooled, moist air flows past the fresh, moist solid at the beginning. At first the solid is dried slowly, which is beneficial for products such as clay, but at the end of the drying process the solid is exposed to very high temperatures. Countercurrent dryers are not suitable for temperature-sensitive products.

A matching of the drying air temperature with the requirements of the material to be dried is not possible in one-stage *crosscurrent dryers*. For this reason the solid must not be sensitive to high temperatures. The advantage of the method is that short drying times can be achieved. Using crosscurrent dryers which are subdivided in zones with different air temperatures, the conditions in cocurrent as well as in countercurrent dryers can be approximated.

As with stationary solids, moving solids can be dried more economically and uniformly when a circulating system for the drying agent is used. A further refinement consists of heating the air at several places along its path through the dryer. In this manner, the dryer is subdivided into several zones, and the temperature of the solid in each zone can be easily regulated. However, the change of humidity and the velocity of the drying agent are prescribed. In order to freely adjust the air velocity as well, the zone construction can be combined with the circulation of the drying agent. In each zone of this type of dryer, the air is circulated in large amounts and can be kept at the desired temperature by individual heaters. The humidity is adjusted by adding small amounts of air from the next zone, which is then mixed with the primary air. This method can be used for stationary or moving solid.

3.2. Sizing the Dryer

After the type of dryer that is needed has been determined, then its size (major dimensions) must be specified, which cannot always be done theoretically. The engineer often needs to turn to values obtained from existing units or experimental tests.

3.2.1. Batch Dryers

Drying with Constant Air Properties. The conditions of the air are practically constant during the entire drying process when large amounts of air are used or a large surplus of fresh air is circulated. If the normalized drying rate curve $\dot{\nu}(\xi)$ is known, then the drying time t can be calculated:

$$\frac{A\dot{m}_{\mathrm{I}}}{M_{\mathrm{s}}(X_{\mathrm{crit}}-X_{\mathrm{eq}})}t=\int_0^{\xi_{\mathrm{i}}}\frac{\mathrm{d}\xi}{\dot{\nu}(\xi)} \tag{30}$$

A is the surface area and M_{s} is the mass of the dry solid. The drying rate in the first period of drying is $\dot{m}_{\mathrm{I}}$ and is determined from Equation (15):

$$\dot{m}_{\mathrm{I}}=\varrho_{\mathrm{g}}\beta_{\mathrm{g}}^{0}\left[Y^{*}(T_{\mathrm{I}})-Y\right]$$

Drying with Variable Air Properties. If the fresh air surplus is small or the amount of circulating air is limited, the conditions of the drying agent change considerably as it flows in the dryer. The normalized moisture content of the solid ξ is not only a function of time but also depends upon the path length of the drying agent z. In the analysis of the drying process, it is useful to introduce the dimensionless drying time τ

$$\tau = \frac{\varrho_g \beta_g^0 A \left[Y^* (T_I) - Y (z = 0)\right]}{M_s (X_{crit} - X_{eq})} \tag{31a}$$

and the dimensionless path length ζ

$$\zeta = \frac{\varrho_g \beta_g^0 A}{\dot{M}_g} \cdot \frac{z}{L} \tag{31b}$$

in which $\dot{M}_g$ is the mass flow rate of the drying agent and L is the length of the dryer. The fundamental equation for the temporal and spatial distribution of the solid's moisture $\xi(\zeta, \tau)$ is

$$\frac{\partial^2 \xi}{\partial \tau \partial \zeta} + \dot{\nu}(\xi) \frac{\partial \xi}{\partial \tau} - \frac{1}{\dot{\nu}(\xi)} \cdot \frac{d\dot{\nu}}{d\xi} \cdot \frac{\partial \xi}{\partial \zeta} \cdot \frac{\partial \xi}{\partial \tau} = 0 \tag{32}$$

This equation assumes that the partial pressure of the vapor is small with respect to the total pressure, and that the normalized drying rate curve $\dot{\nu}(\xi)$ is independent of the condition of the drying agent. In addition, the critical moisture content X_{crit} must be independent of the drying rate. Solutions to this equation are given in the literature [9]. A qualitative picture of the temporal and spatial moisture distribution in the solid is given in Figure 55. The initial moisture content is ξ_i, and $\xi = 1$ corresponds to the critical moisture content. At $\tau = \tau_{crit}$, the solid's critical moisture content is achieved at the entrance of the dryer, $\zeta = 0$. Subsequently, the location of the critical moisture content of the solid $\zeta_{crit}(\tau)$ is shifted through the dryer. Whenever $\zeta_{crit} > \zeta_L$, the entire solid would be in the second or possibly third stage of drying.

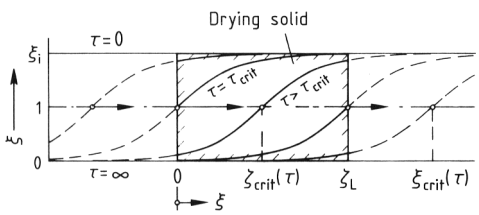

Figure 55. Moisture distribution in a solid as a function of dimensionless time τ

In order to estimate the extent of drying, it is often sufficient to approximate the normalized drying rate curve $\dot{\nu}(\xi)$ with the linear approximation

$$\dot{\nu} = \xi \text{ for } 0 \le \xi \le 1$$
$$\dot{\nu} = 1 \text{ for } 1 \le \xi \le \xi_i$$

In this case Equation 32 has an analytical solution. For

$$\tau_{crit} = \xi_i - 1 \tag{33}$$

in the region $\zeta \ge \zeta_{crit}$ and $0 \le \tau/\tau_{crit} < \infty$

$$\xi(\zeta, \tau) = \xi_i - (\xi_i - 1) \exp\left[-(\zeta - \zeta_{crit})\right] \tag{34}$$

is obtained. In the region $\zeta \le \zeta_{crit}$ and $0 \le \tau/\tau_{crit} < \infty$ the result is

$$\xi(\zeta, \tau) = \frac{\xi_i}{1 + (\xi_i - 1) \exp\left[-\xi_i (\zeta - \zeta_{crit})\right]} \tag{35}$$

The dimensionless location of the critical moisture content ζ_{crit} can be calculated by

$$\zeta_{crit} = \ln(\tau/\tau_{crit}) \tag{36}$$

for $\tau/\tau_{crit} < 1$, and by

$$\zeta_{crit} = \frac{1}{\xi_i} \ln \frac{\xi_i \exp\left[(\xi_i - 1)(\tau/\tau_{crit} - 1)\right] - 1}{\xi_i - 1} \tag{37}$$

for $\tau/\tau_{crit} \ge 1$. Even though Equations (36) and (37) encompass all values of ζ_{crit} between $\pm\infty$, only the values between 0 and ζ_L have a physical meaning.

3.2.2. Continuous Dryers

Useful suggestions and short-cut methods for the sizing of drum, convection, and spray dryers can be found in [10]. The following theoretical analysis utilizes the same simplifying assumptions that were used in the analysis of batch drying (see Section 3.2.1).

Drying with Constant Air Properties. The drying time t is determined by Equation (30). This corresponds to the residence time of the solid in the dryer. If u is the speed of the conveyor, then the necessary dryer length is $ut = L$.

Drying with Variable Air Properties. In contrast to batch drying, the location of the solid's critical moisture content within the dryer is independent of time. The basic equation for the spatial variation of the moisture in a cocurrent dryer is

$$\frac{d\xi}{d\zeta} + (\xi + C)\, \dot{\nu}(\xi) = 0 \tag{38}$$

In Equation (38) C is the so-called *excess air factor*, which is defined as

$$C=\frac{[Y^*(T_I)-Y(\zeta=0)]\,\dot{M}_g-(X_i-X_{eq})\,\dot{M}_s}{(X_{crit}-X_{eq})\,\dot{M}_s} \tag{39}$$

If $C=0$, the air exactly reaches saturation when the solid has reached its equilibrium moisture content. In practice, $C>0$ is used. To obtain an order of magnitude, Equation (38) can be integrated with a simple function for the normalized drying rate curve. For this purpose the linear approximation $\dot{\nu}(\xi)=1$ for $\xi\geq 1$ and $\dot{\nu}(\xi)=\xi$ for $\xi\leq 1$ is used. For the region $\xi\geq 1, \zeta\leq\zeta_{crit}$ the result is

$$\xi=(1+C)\exp[-(\zeta-\zeta_{crit})]-C \tag{40}$$

For the region $\xi\leq 1, \zeta\geq\zeta_{crit}$

$$\xi=\frac{C\exp[-(\zeta-\zeta_{crit})\,C]}{1+C-\exp[-(\zeta-\zeta_{crit})\,C]} \tag{41}$$

is obtained. In both cases

$$\zeta_{crit}=\ln\frac{\xi_i+C}{1+C} \tag{42}$$

holds. Countercurrent drying can be analyzed with the following basic equation

$$\frac{d\xi}{d\zeta}+(C^*-\xi)\,\dot{\nu}(\xi)=0 \tag{43}$$

in which $C*=C+\xi_i+\xi_f$; the subscripts denote the initial and final moisture content of the solid. Integration of Equation (43) for the region $\xi\geq 1, \zeta\leq\zeta_{crit}$ yields

$$\xi=(1-C^*)\exp(\zeta-\zeta_{crit})+C^* \tag{44}$$

and for the region $\xi\leq 1, \zeta\geq\zeta_{crit}$

$$\xi=\frac{C^*\exp[-(\zeta-\zeta_{crit})\,C^*]}{C^*-1+\exp[-(\zeta-\zeta_{crit})\,C^*]} \tag{45}$$

The dimensionless location of the critical moisture content is given by

$$\zeta_{crit}=\ln\frac{C^*-1}{C^*-\xi_i} \tag{46}$$

Figure 56 is an example of the variation of the moisture content ξ along the length of a dryer for cocurrent and countercurrent operation. For the calculation the values $\xi_i=2$, $\xi_f=0.2$, and $C=0$ have been used. The cocurrent dryer must be 1.7 times longer than the countercurrent unit.

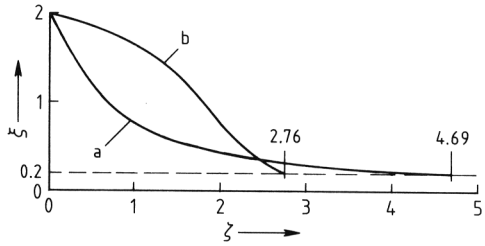

Figure 56. Dimensionless moisture content versus dimensionless location ζ in a dryer with
a) Cocurrent; b) Countercurrent flow ($\xi_i=2$, $\xi_f=0.2$, and $C=0$)

3.3. Heat and Driving Power Requirements

The heat consumption of dryers is determined by

1) the heat needed to vaporize or desorb the moisture from the solid
2) the heat needed to warm the solid up to the drying temperature and to heat the vapor up to its exit temperature
3) the heat lost to the surroundings

The heat consumption is determined from an energy and mass balance of the system. Figure 57 shows the situation for a batch dryer. Mass balance is

$$\dot{M}_g\,(Y_i-Y_f)=M_s\frac{dX}{dt} \tag{47}$$

Energy balance is

$$\dot{Q}_{Heat}+\dot{W}_{Vent}=\dot{M}_g\,(h_f-h_i)+\dot{M}_s\frac{d\,(h_s+Xh_l)}{dt}+\dot{Q}_{lost} \tag{48}$$

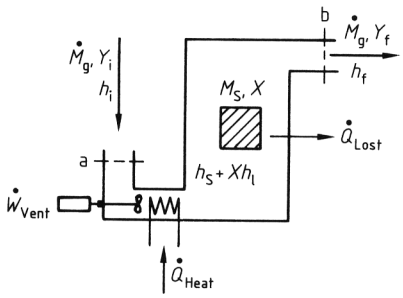

Figure 57. Schematic drawing of a batch dryer for the derivation of overall mass and energy balances
a) Inlet; b) Outlet

In these equations, h_i and h_f are the entering (initial) and exiting (final) enthalpies of the moist drying agent, respectively; h_s is the enthalpy of the dry solid and h_l the enthalpy of the liquid. $\dot{Q}_{heat}$ is the heat requirement from the heaters; $\dot{W}_{vent}$ is the power required by the ventilation unit; and $\dot{Q}_{lost}$ is the heat lost to the surroundings.

The corresponding equations for a continuous dryer, shown schematically in Figure 58, are

$$\dot{M}_g\,(Y_f - Y_i) = \dot{M}_s\,(X_i - X_f) \tag{49}$$

for the mass balance and

$$\dot{Q}_{Heat} + \dot{W}_{Vent} = \dot{M}_g\,(h_f - h_i) + \dot{M}_s\,[(h_{sf} - h_{si}) + (X_f h_{lf} - X_i h_{li})] + \dot{Q}_{lost} \tag{50}$$

for the energy balance.

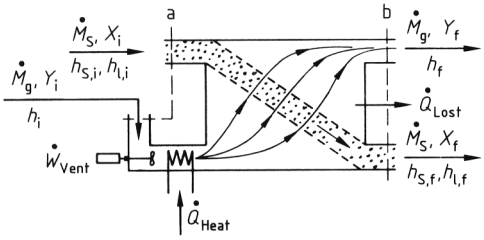

Figure 58. Schematic drawing of a continuous dryer for the derivation of overall mass and energy balances
a) Inlet; b) Outlet

Approximate values for the heat consumption of the most important types of dryers are given in Table 2.

Driving Power Consumption. Not only is energy needed as heat to the dryer, also both the solid and the drying agent must be transported. Most of the power needed to move these materials in a convection dryer is consumed by the fans or blowers. This power has already been accounted for in the overall balances of Equation (50). The vacuum pump absorbs most of the power in a vacuum dryer. A guide to the driving power requirements of the various kinds of dryers is given in [10].

4. References

1. O. Krischer, W. Kast: *Die wissenschaftlichen Grundlagen der Trocknungstechnik,* 3rd ed., Springer, Berlin-Heidelberg-New York 1978.
2. R. B. Keey: *Drying: Principles and Practice,* Pergamon Press, Oxford 1972.
3. R. H. Perry, D. Green (eds.): *Perry's Chemical Engineer's Handbook,* 6th ed. McGraw-Hill, New York-Toronto-London, 1984, pp. 20-1 – 20-58.
4. P. Gummel: *Durchströmungstrocknung: Experimentelle Bestimmung und Analyse der Trocknungsgeschwindigkeit und des Druckverlustes luftdurchströmter Textilien und Papiere,* Dissertation, Karlsruhe 1977.
5. E.-U. Schlünder, N. Mollekopf: "Vacuum Contact Drying of Free Flowing Mechanically Agitated Particulate Material," *Chem. Eng. Process* **18** (1984) 93 – 111.
6. E.-U. Schlünder, "Heat Transfer to Packed and Stirred Beds from the Surface of Immersed Bodies," *Chem. Eng. Process.* **18** (1984) 31 – 53.
7. F. Thurner, E.- U. Schlünder: "Progress Towards Understanding the Drying of Porous Materials Wetted with Binary Mixtures," *Chem. Eng. Process.* **20** (1986) 9 – 25.
8. J. Schwarzbach, E.-U. Schlünder: "Microconvection due to Marangoni-type Instabilities in Porous Media – Effect on Pervaporation and Selective Drying," *Proc. of 3rd World Congress of Chemical Engineering,* Tokyo 1986, 560 – 563.
9. D. A. van Meel: "Adiabatic Convection Batch Drying with Recirculation of Air," *Chem. Eng. Sci.* **9** (1958) 36 – 44.
10. K. Kröll: *Trockner und Trocknungsverfahren,* 2nd ed., Springer, Berlin-Heidelberg-New York 1978.

Size Reduction

See also Solids Technology, Introduction

SIEGFRIED BERNOTAT, Institut für Mechanische Verfahrenstechnik der Technischen Universität Braunschweig, Braunschweig, Federal Republic of Germany (Chap. 3)

KLAUS SCHÖNERT, Institut für Aufbereitung und Veredelung der Technischen Universität Clausthal, Clausthal, Federal Republic of Germany (Chaps. 1, 2)

1. Introduction 2
2. Fundamentals 3
2.1. Particle Breakdown 3
2.2. Material Properties 8
2.3. Process Description 14
3. Size-Reduction Equipment 15
3.1. Crushers 16
3.1.1. Jaw Crushers 16
3.1.2. Gyratory (Cone) Crushers 18
3.1.3. Roll Crushers 19
3.1.4. Impact and Hammer Crushers 20
3.2. Grinding Media Mills 21
3.2.1. Tumbling mills 22
3.2.2. Vibration Mills 25
3.2.3. Planetary Mills 26
3.3. Roller and Ring-Roller Mills 26
3.3.1. Operation Characteristics 26
3.3.2. Roller Mills 28
3.3.3. Ring-Roller Mills 29
3.4. Impact and Jet Mills 32
3.4.1. Impact Mills 33
3.4.2. Jet Mills 35
3.5. Cutting Mills 37
3.6. Wet Grinding 37
4. References 40

Symbols

a	half the crack length, half the diameter of the contact surface
A	cross sectional area of specimen
A_B	fracture surface area
b_{ij}	breakage distribution coefficient of fragments
B	breakage function, sample thickness
B_f	breakage fraction
c	specific heat
c_v	volume concentration of solid
d	particle size, diameter of ball or disk
d_L	diameter of perforations
D	mill diameter, roller diameter
E	energy, Young's modulus
E_B	energy of fracturing
E_M	energy applied relative to mass,
E_u	energy utilization
F	force
F_{sp}	specific contact pressure
g	acceleration due to gravity
G	energy release rate
h	stroke
H	height of the crusher chamber
k_{opt}	optimal recirculation factor
l	crack length
L	length of cylindrical specimen, mill length, roller length
$\dot{m}$	mass throughput
m_i	mass fraction of interval i
M	mass
M_0	total particle mass
M_p	mass of fragments below lower limit of size range
n_c	critical rotational speed
p_m	maximum compressive stress
P	power
q_A	relative area of the ascending zone
q_L	relative area of the perforations
Q	heat produced during fracturing, particle-size distribution
r	radius of curvature
r_s	amplitude of oscillation
R	crack resistance
s	gap width
S	newly created surface area
S_M	specific surface area
t	time
u	peripheral speed of roller
u_p	velocity of particle in gap
U_{el}	elastic energy of stress field
T	driving torque

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_05

v	velocity
V	volume
$\dot{V}$	volumetric flow rate
w_{i}	specific rate of breakage of fraction *i*
W	work
W_{ad}	specific work (or energy) required in additional process steps
W_{B}	energy of fracturing
W_i	Bond index
W_{M}	specific work (or energy) of comminution
W_{t}	total work
x	particle size
x_{F}	particle size of feed
x_{P}	particle size of product
x_{V}	equivalent diameter of a sphere with same volume as particle
z	displacement
α	crushing angle
β	angle of force action
δ	length of fracture zone
ν	Poisson number
ΔS	newly created surface area
ΔT_m	maximum temperature increase
Δx	particle-size range
ϱ	density
ϱ_{s}	density of solid
σ	tensile stress in unperturbed state
σ_{B}	breaking stress
σ_{m}	maximum tensile stress
σ_{y}	yield stress
τ_{m}	maximum shear stress
ω	angular velocity

1. Introduction

Size reduction, also known as comminution, is defined as the mechanical breakdown of solids into smaller particles without changing their state of aggregation. Processes such as the atomization of liquids or the separation of gas into bubbles are not considered in this article.

Products that have undergone size reduction surround us daily. Some examples are foods such as flour, sugar, and spices; construction materials such as cement, lime, plaster and sand; and pigment particles in lacquer and paint. Throughputs in industrial size reduction range from a few kilograms to hundreds of tons per hour. Particles ranging in size from blocks with edge lengths of 1 – 2 m down to very fine particles with a diameter of 1 µm are subjected to size reduction.

The two main objectives of size reduction are:

1) *Production of a specified particle-size distribution or a specific surface area.* The maximum permitted amount of residual coarse material at the upper particle-size distribution limit or the maximum permitted amount of fines at the lower limit is also stipulated. These requirements are always applied if the dispersity properties of the material have a pronounced effect on product quality (e.g., pigment color intensity, taste of chocolate, cement strength) and properties (e.g., agglomeration, miscibility, and solubility [1], [2]).
2) *Treatment of multicomponent materials.* In the case of inhomogeneous materials such as ores, the content of a valuable substance in a particle frequently depends on the particle size. As a result, the size must be reduced sufficiently so that the components are obtained as particles of maximum purity which can then be separated.

Size-reduction devices have been in use ever since people began to eat cereal crops. The exploitation of water, wind, and steam power led to the development of size-reduction equipment. However, a thorough treatment of size reduction from a scientific and engineering viewpoint began only in the late 1950 s. The work of Rumpf and coworkers in the Federal Republic of Germany deserves special mention [3]. Additional research groups now exist in Germany, the United States, and Japan [4]. The European Symposia on Size Reduction provide an international insight into this field [4–6]. They cover research and progress in the basic understanding and mathematical description of size reduction as well as developments in machinery, processes, and protection against wear.

A comprehensive survey of size-reduction equipment is given in [7]. Reviews of size reduction are given in [8–18].

2. Fundamentals

Areas covered in this chapter include the basic mechanism of particle breakdown, the com-

minution results obtained on a laboratory scale using particles under defined loading conditions, and a description of size-reduction processes.

2.1. Particle Breakdown

Particle breakdown describes the sequence of events that leads to particle fracture:

1) Loading of the particle
2) Particle deformation
3) Buildup of a nonuniform stress field
4) Occurrence or nonoccurrence of fracturing

Division of the process into the above steps facilitates understanding of the point at which the many variable parameters exert their effects. Examples of such parameters are particle shape, particle size, material properties (e.g., elasticity), temperature, and type of stress. Stress may be applied between two surfaces (*compression*) or at a single surface (*impacting*). Figure 1 summarizes the various parameters and their effects, which are discussed below.

The *state of stress* plays a key role in particle breakdown and determines whether inelastic deformation or fracturing occurs. The latter two phenomena are termed the particle reaction. The particle reaction, in turn, affects the state of stress; this is especially important in inelastic deformation. Furthermore, the state of stress depends on the loading force, type of comminution tool used (flat plates or cutters), particle shape, and mechanical properties of the material.

In size-reduction equipment, the particles are stressed by *contact forces.* Generally, the cause of the stress (i.e., contact between neighboring particles or between particles and a moving part of the machine) is unimportant. The conditions in various types of equipment (e.g., crushers, ball mills, or impact mills) differ with respect to the number and direction of contact forces and the loading velocity.

Deformation and stress are related by the *mechanical properties of the material*, which are described by three limiting cases: elastic, plastic, and viscous behavior. Plastic and viscous behavior are both inelastic. In practice, however, the behavior of materials subjected to loads sufficient to cause breakdown seldom corresponds to only one of these cases.

Elastic behavior (described in the simplest case by Hook's law) is characterized by reversibility and time independence. Because of the latter, the stress in the contact region depends solely on the contact force and not on the loading velocity. Therefore, impacting and slower compression have the same effects. Elastic waves of significant intensity occur only at impact velocities above 300 m/s. Such values are seldom achieved in mills.

The limiting case of *plastic behavior* is characterized by irreversibility, as well as time independence. However, for plastic materials encountered in practice, a slight time dependence exists but can be neglected. Plastic deformation produces internal stresses and embrittlement. As a result, particles that undergo this type of deformation (e.g., metal powders) must usually be stressed repeatedly to break them. Because the loading velocity has little or no effect on the stress, only a slight advantage is obtained by using impact instead of compression.

Viscous behavior is distinguished from the two other limiting cases by its strong dependence on strain velocity and the related temperature. Plastics such as polyethylene and polystyrene are typical examples of this class of materials. The instantaneous stress is a function of the previous course of deformation. Therefore, under certain conditions, the maximum stress in a sphere can occur during removal of the load. The loading velocity plays an important role in materials exhibiting viscous behavior. In high-pressure polyethylene, a very rapid deformation lasting a few microseconds produces a four- to fivefold higher stress than a slow deformation lasting a few seconds. Consequently, in this case, impacting is superior to other methods of stress application. Cooling reduces the viscous component of deformation. Regions undergoing plastic deformation experience an increase in temperature, which in turn increases viscous behavior. Because the particles are generally stressed many times before breaking, intermittent cooling is advantageous. In cases of highly viscous behavior, application of stress with a cutting blade is more successful than impact stress. However, particle fineness is limited by equipment design.

The deformation of *irregularly shaped particles* and the resultant stresses cannot be calculated. However, to a close approximation, the

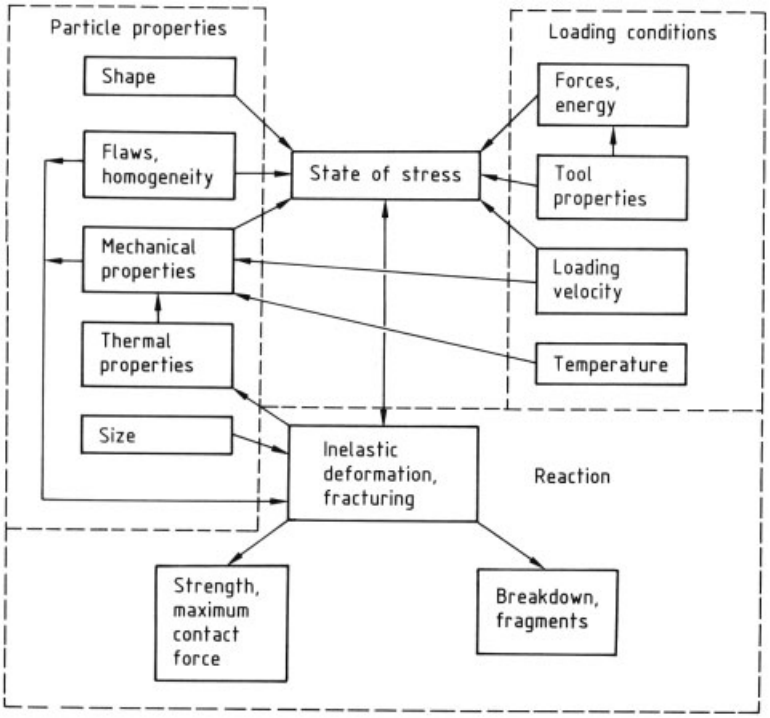

Figure 1. Factors affecting particle breakdown in size reduction

contact region of the particle can be considered spherical. Solutions obtained by using the Hertz – Huber theory are valid only for linear elastic behavior and small deformations. Nevertheless, important information results from them and is discussed in the following.

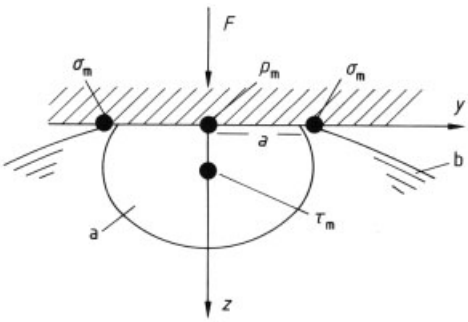

Figure 2. Explanation of the results of the Hertz – Huber theory for the contact region of a spherical particle
a) Region of shear and compressive stress; b) Surface of stressed particle
a = radius of the contact surface; F = applied force; p_m = maximum compressive stress; σ_m = maximum tensile stress; τ_m = maximum shear stress

Beneath the contact surface, a region exists in which only compressive and shear forces are found. The maximum compressive stress p_m occurs at the center of the contact surface, whereas the maximum tensile stress σ_m occurs at its perimeter, and the maximum shear stress τ_m along the radius at a distance roughly one-quarter the diameter of the contact surface (Fig. 2). These stresses have the following ratio:

$$p_m : \tau_m : \sigma_m = 1 : 0.3 : 0.17$$

This ratio is a function of the Poisson number ν, and the above values are for $\nu = 0.25$.

Brittle materials exhibit primarily elastic behavior up to the breaking point. Here, cracking starts in the area of tensile stress at the perimeter of the contact area. During compression, cracks extend between the contact areas (Fig. 3 B). During impacting, cracks diverge within the particle (Fig. 3 A) [19].

Despite the existence of triaxial compressive stresses in the contact region, most of the fines are formed here. Since the compressive stresses are not equal, the cracks may trigger fragmentation. Theoretical studies have shown that a com-

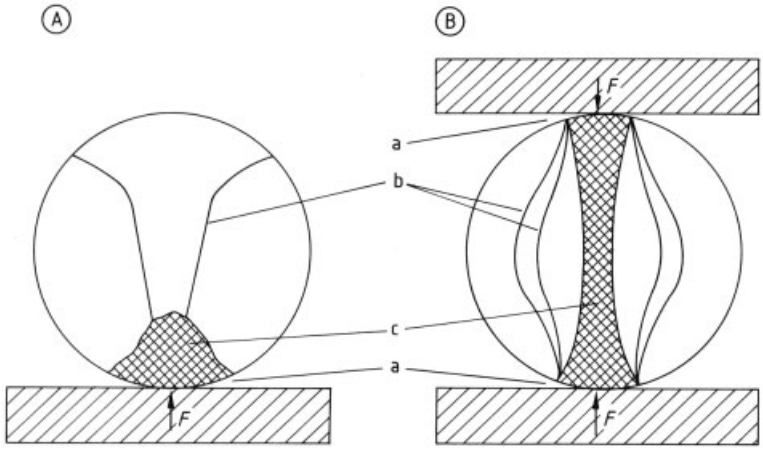

Figure 3. Fracturing in brittle glass spheres subjected to impact (A) and compression (B) loading
a) Contact surface at which stress is applied; b) Cracks; c) Fines

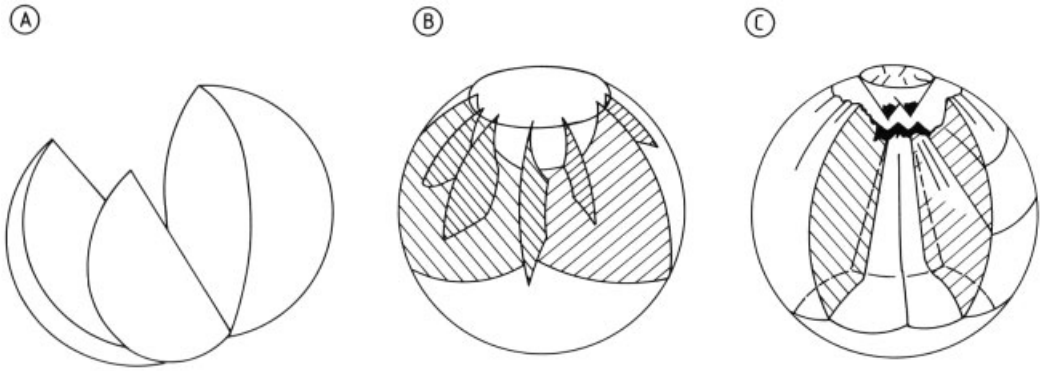

Figure 4. Meridian cracks in inelastically deforming poly(methyl methacrylate) spheres
A) Slow compression at 20 °C; B) Impacting at 90 m/s and 100 °C; C) Impacting at 90 m/s and −196 °C

pact fracture field, and thus fines, are formed at high compressive stress [19].

By contrast, in plastic behavior, a conical region is pushed into the sphere. Sideways displacement of material produces peripheral tensile stresses and cracks in meridian planes (Fig. 4). Because stresses are determined by the depth of cone penetration, additional shear stress does not significantly affect the fracturing process. Thus, additional frictional stress has no effect, it leads to increased energy losses. The idea that friction is an advantage must be viewed with skepticism. However, frictional stresses are useful for the disintegration of agglomerates.

Smaller particles contain only small flaws. Thus, larger stresses are required to reach the yield stress beneath the contact area and initiate fracturing. Deformation behavior, therefore, depends on particle size. For a brittle material, a transition from brittle to plastic behavior is observed in a defined particle-size range. These transitions are evident in loading diagrams. Multiple peaks indicate fracturing and brittle behavior (Fig. 5 A), whereas smooth curves indicate plastic deformation (Fig. 5 B). Figure 6 shows the transition zones between brittle and plastic behavior for various materials. Corresponding particle strengths are also given and are defined as the breaking force divided by the nominal particle cross section. If the contact surface is assumed to be one-tenth of the nominal cross section, the resulting stresses are the same order of magnitude as the molecular bond strength. Reduction of these materials to sizes below this transition region is very difficult.

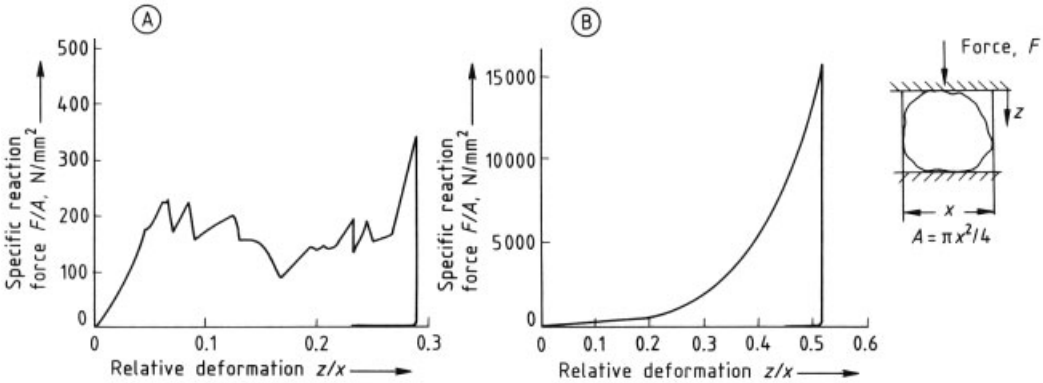

Figure 5. Loading diagram for cement clinker particles with a diameter of 16.1 µm (A) and 2.5 µm (B)

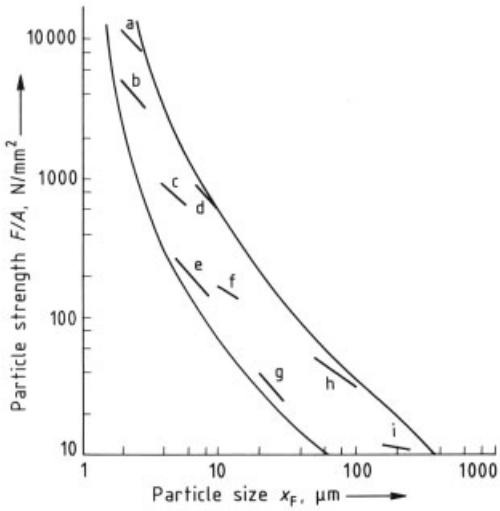

Figure 6. Particle strength of various materials in the transition zone between brittle and plastic deformation
a) Boron carbide; b) Crystalline boron; c) Quartz; d) Cement clinker; e) Limestone; f) Marble; g) Coal; h) Cane sugar; i) Potash salts

All solids display *material flaws* such as crystal dislocations, grain boundaries, voids, and cracks which lead to an increase in stress when a load is applied. Consequently, they are the starting point for inelastic deformation and fracturing. The material flaws within a particle largely determine its reaction and, thus, its strength, breaking energy, and fragment-size distribution. The increased stress also explains why the observed particle strength is far below the theoretical or molecular strength (which is usually one- to two-tenths of Young's modulus). The increased stress at the tip of the cracks can be estimated. According to the notch tension theory, for elastic behavior with a uniaxial tensile stress perpendicular to the crack the following applies (Fig. 7):

$$\sigma_m/\sigma = 1 + 2\sqrt{a/r}$$

where σ is the stress in the unperturbed state, a is half the crack length l, and r is the radius of curvature of the tip of the crack. Substitution of typical values (e.g., $a = 1$ µm, $r = 1$ nm, $\sigma = 100$ N/mm^2) gives very high stresses that exceed the yield stress. This leads to the conclusion that a zone of inelastic deformation, called the fracture zone, forms around the tip of the flaw and later around the tip of the crack.

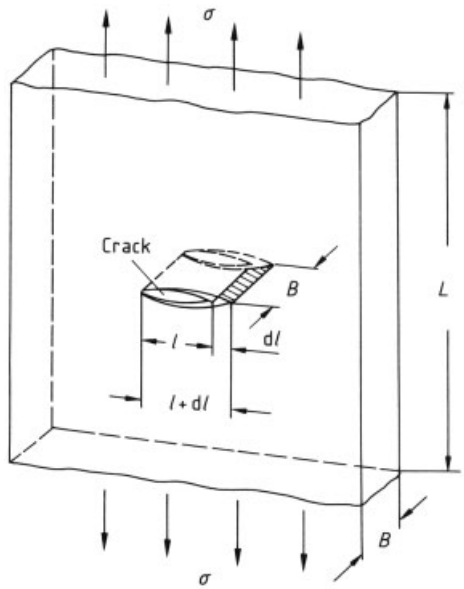

Figure 7. Schematic of a specimen with a crack produced by tensile stress

Fracturing involves *energetic considerations*. Energy is required for the inelastic deformation occurring in the fracture zone, for the electronic processes involved in material separation, and for the creation of new surfaces. The required energy, relative to the fracture surface area A_B ($A_B = B \cdot l$, where B is the thickness of the material, see Fig. 7), is termed the *crack resistance* R and, at maximum crack propagation velocity, is a material constant.

The crack generally propagates so quickly that energy cannot be supplied externally; it can only come from the stress field. To determine the energy supply from the stress field, the decrease in elastic energy of the stress field dU_{el} caused by a differential increase in crack length dl is considered

$$G = -dU_{el}/dA_B$$

where G is the *energy release rate.* The fracture surface A_B corresponds to half the total surface area produced by the fracture. The parameter G is a function of the specimen shape, the position of the crack, the type of loading, and the mechanical properties of the material. Theoretical solutions for calculating G are available for simple cases. For a crack or fracture of length $l = 2a$, which propagates perpendicular to a tensile stress in an infinitely large plate (Fig. 7) at a velocity less than one-tenth the velocity of the longitudinal wave, the following applies:

$$G = \pi a \sigma^2 / E$$

where E is Young's modulus. The energy release rate increases with crack length, provided the stress remains constant; this is always the case for relatively large specimens.

The expression $G \geq R$ represents the differential energy balance. The parameter G can be calculated from measured values of a and σ. In addition, the heat value of the fracturing process can be measured and used to determine the inelastic deformation energy. These values agree within a few percent, which implies that the fracturing energy is primarily determined by inelastic deformation. For glass and similar materials, the crack resistance ranges from 1 to 10 J/m^2; for plastics, it is 10 – 10^3 J/m^2; and for metals, 10^2 – 10^5 J/m^2. The specific surface energy of the newly created surface is, however, only 0.01 – 0.5 J/m^2. These figures clearly show that changing the surface energy via the medium surrounding the solid, for example with *grinding aids*, cannot have any effect on the fracturing process. Effective aids for very fine grinding in ball mills improve processing conditions by preventing agglomerate formation and adhesion of the ground material.

A *minimum specimen size* is necessary to ensure that a fracture crosses the entire particle. This can be estimated by using a cylindrical model of cross section A and length L. Fracturing starts at the breaking stress σ_B. Thus, the integral energy balance is

$$(\sigma_B^2/2E)\ A \cdot L > R \cdot A$$

$$L \geq 2R \cdot E/\sigma_B^2$$

Typical values for the minimum specimen length of glass are 10 – 100 µm and of steel 2 – 20 mm. Therefore, in the case of fine particles (<100 µm), the integral energy balance is not met. Furthermore, a single contact stress produces several breaks rather than a single fracture. Therefore, to ensure the disintegration of fine particles, additional energy must be supplied after the initial fracture, or stress must be applied repeatedly.

The *length of the fracture zone* δ can be estimated if it is much smaller than the length of the crack. The following applies [20]:

$$\delta \approx 0.4R \cdot E/\sigma_y^2$$

where σ_y is the yield stress. Values of δ for glass and brittle minerals are 1 – 10 nm and for brittle plastics 1 – 10 µm. The length of the fracture zone determines the minimum separation between fractures and, consequently, the size of the smallest fragments (ca. 5 – 10 δ).

Inelastic deformation causes an *increase in temperature.* At the maximum crack propagation velocity (1000 – 2000 m/s for glass, 500 – 800 m/s for plastics), a sudden temperature change occurs. The maximum temperature increase can be calculated from adiabatic considerations:

$$\Delta T_m = Q/\delta^3 \varrho c$$

where Q is the heat produced during fracturing, ϱ the density, and c the specific heat. Measurements give the following values; 4000 K

for quartz, 3000 K for glass, 1300 K for limestone, 200 K for sodium chloride, and 150 K for poly(methyl methacrylate) [21].

Extreme conditions exist in the fracture zone. Structural changes (mechanical activation) can be expected, which may be either favorable or harmful to the product [22]. Different findings show a higher reactivity of the fracture surface (e.g., the amorphous structure of the fracture surfaces of crystalline quartz or sugar) and a decrease in the molecular mass of polymers. The original state of the fracture surface changes immediately after size reduction due to the action of the atmosphere. Microcracks and microstructures disappear. Gas adsorption measurements made immediately after size reduction yield substantially larger surface area values than later measurements. Values increase by the following factors: quartz 1.6, calcite 1.6, calcium fluoride 2.0, lithium fluoride 2.7, glass 2.7, sugar cane 3.1, and sodium chloride 6.4 [23].

2.2. Material Properties

To plan, predict, optimize, and control a process, knowledge of the relevant material properties is required. In the case of size reduction, this poses several difficulties. The material is not a continuous phase but a dispersion composed of a group of particles with various shapes and sizes. The energy applied to the group of particles reaches only a portion of them. The size of this fraction is unknown, and the type of energy transfer can be described only approximately. The reaction of a particle to an applied stress depends on its size and shape, as well as the conditions under which stress is applied. As described in Section 2.1 (Fig. 1), the effect of applied stress is not solely a function of the energy supplied.

The material properties relevant to size reduction can be divided into two groups:

1) Parameters describing the resistance to breakage: particle strength as well as the energy and probability of fracturing, and the breakage fraction
2) Parameters pertaining to the effect of the applied stress: breakage function (size distribution of the resulting fragments), increase in surface area, and energy utilization

The irregular shape of the particles and the effects of particle size cause special difficulties. The above parameters are defined in such a way that they can be measured and are of practical use. They cannot be derived from known material properties such as Young's modulus, tensile strength, yield strength, or hardness.

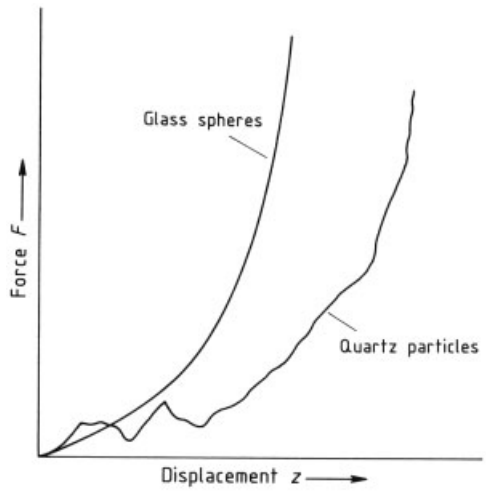

Figure 8. Loading diagram for glass spheres and quartz particles

Material properties that are important during size reduction depend on the nature of the feed in the mill and the loading conditions.

Feed parameters are

1) type, origin, and prior treatment of material;
2) particle size;
3) particle shape; and
4) particle homogeneity.

Loading conditions include

1) type of applied stress as described by Rumpf [24]: stress can be applied between two surfaces (compression and shear stress), to a single surface (impacting), via the surrounding medium, or via nonmechanical energy transfer;
2) particle configuration: single particles or groups of particles;
3) stress intensity quantified by the energy of fracturing relative to volume or mass, the reaction force relative to surface area, and the reduction ratio (see page 10);
4) loading velocity: in impacting, the energy supplied relative to the mass is a function of the impact velocity ($W_M = v^2/2$);

5) frequency of stress application: single or multiple applications;
6) shape and hardness of the surface of size-reduction tools: smooth or molded surfaces; and
7) temperature.

In the usual strength tests, material breakdown occurs when the applied load reaches a critical value. This value can be used to derive a measurement of strength or energy of fracturing. However, the loading diagram of small, irregularly shaped particles displays numerous peaks indicating fractures (Fig. 8). Therefore, the first large peak must be defined as the breaking point.

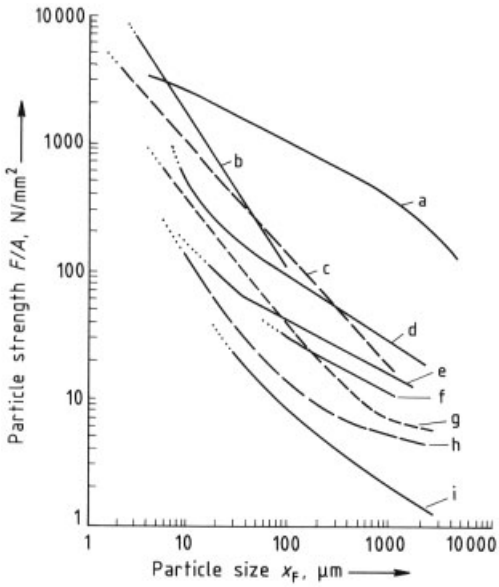

Figure 9. Particle strength of various materials as a function of particle size
a) Glass spheres; b) Boron carbide; c) Crystalline boron; d) Cement clinker; e) Marble; f) Cane sugar; g) Quartz; h) Limestone; i) Coal

Parameters at the Breaking Point. The particle strength and the energy of fracturing are derived from the breaking point. *Particle strength* is the force at the breaking point relative to the nominal cross section $\pi x_V^2/4$ (x_V is the equivalent diameter of a sphere which has the same volume as the particle in question). The *energy of fracturing* E_B is the energy supplied up to the breaking point, relative to the mass of the particle.

Although particle strength has the same dimensions as stress, it does not represent the stress on the particle because the contact area is unknown. Figure 9 shows particle strengths for various materials [25]. At particle sizes below several millimeters, the strength increases sharply because as the particle size decreases, the material flaws become smaller and the particles more homogeneous. Because the crack resistance remains constant, the differential energy balance necessitates that increasing stress be applied to initiate fracturing (see Section 2.1). However this leads to considerable deformation because the compressive and shear stresses in the contact region are much higher than the tensile stresses that initiate fracturing. As a result, the tensile stresses increase only slightly or may even be prevented from increasing.

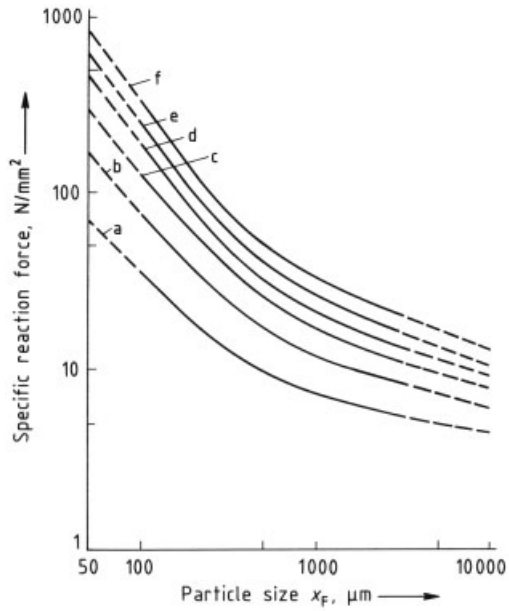

Figure 10. Specific reaction force of quartz particles as a function of particle size and reduction ratio
Reduction ratio, x_F/s (particle size divided by gap width): a) 1.5; b) 2; c) 2.5; d) 3; e) 3.5; f) 4

Parameters at Loads above the Breaking Point. In contrast to the customary strength tests, energy can be supplied to a particle under compression or impacting at levels that exceed the breaking point. This yields a larger number of finer particles. In the case of compression, both the particles and their fragments exert a reaction

force against energy transfer, which is related to the nominal cross section.

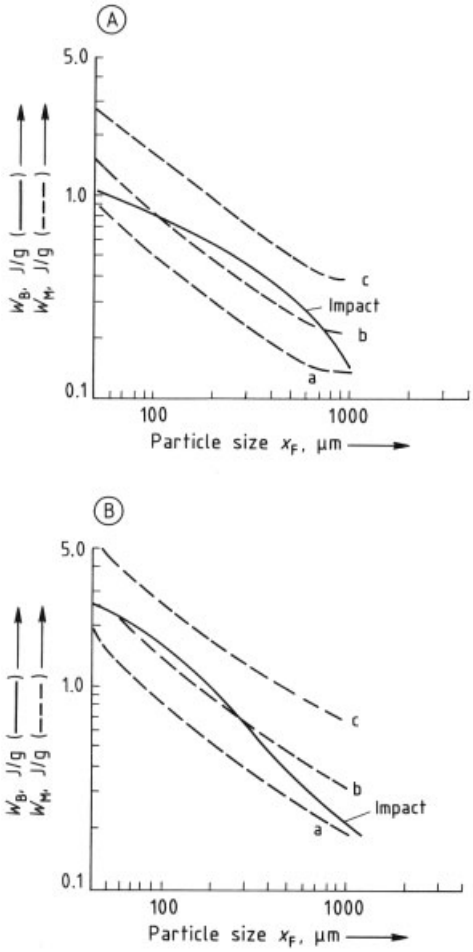

Figure 11. Energy of fracturing W_B and specific work of comminution W_M for quartz (A) and limestone (B) Reduction ratio, x_F/s: a) 1.2; b) 1.3; c) 1.5

To distinguish parameters above the breaking point from those related to the breaking point, the terms *specific reaction force* (measured force relative to the nominal cross section) and *specific work of comminution* W_M (work done relative to the mass) are used. The parameters at loads above the breaking point are not resistance parameters but are useful, measurable quantities. For example, the *reduction ratio* x_F/s (feed particle size divided by gap width) provides a useful standard when compression is employed. Reaction force and work of comminution can be measured as functions of the reduction ratio. An example is shown in Figure 10, which again displays the typical increase of reaction force with decreasing particle size. Figure 11 compares the energy of fracturing W_B, measured by means of impacting, with the specific work of comminution at three reduction ratios for limestone (Fig. 11 A) and quartz (Fig. 11 B). To a first approximation, the work of comminution for reduction ratios of 1.2 – 1.3 is equal to the energy of fracturing. This is probably also valid for other brittle materials.

Probability of Fracturing. The probability of fracturing indicates the fraction of particles broken under a specific load. Determination of breakage by visual inspection is simple in the case of spheres, but more difficult in the case of irregularly shaped particles. Figure 12 shows the probability of fracturing of glass spheres. The increase in particle strength at lower particle sizes displaces the curves to the right, in the direction of higher impact velocities.

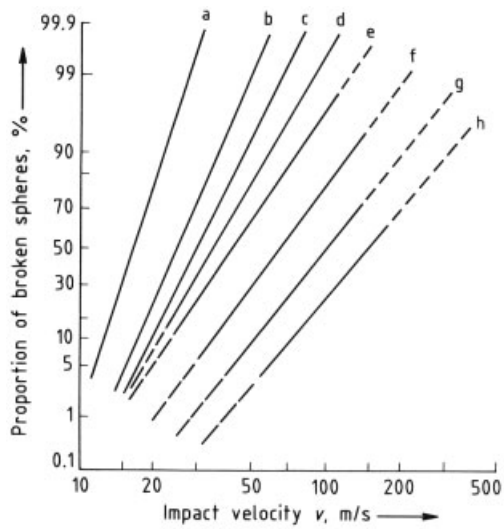

Figure 12. Probability of fracturing for glass spheres Particle size x_F, mm: a) 8; b) 4; c) 2.3; d) 1.6; e) 1.2; f) 0.5; g) 0.2; h) 0.1

Breakage Fraction. The breakage fraction B_f can be considered a modified fracture probability for irregularly shaped particles. It is also a very informative resistance parameter for practical applications. A narrow particle-size fraction between $x_F - (\Delta x/2)$ and $x_F + (\Delta x/2)$ is subjected to stress by a specified method, and the

mass of fragments M_p below the lower limit of the size range is determined; x_F is the feed particle size, defined as the mean value of the particle-size range Δx. The value of M_p depends on the particle-size range Δx. Extrapolation of the function M_p/M_0 (M_0= total particle mass) to $\Delta x = 0$ gives the breakage fraction:

$$B_f = \lim_{\Delta x \to 0} (M_P/M_0) = \lim_{\Delta x \to 0} f\,(x_F, \Delta x, W_M)$$

The breakage fraction is obtained with sufficient accuracy with the aid of fractions obtained by using standard laboratory screens whose mesh sizes decrease by a factor of $\sqrt{2}$. Theoretical consideration of crack initiation based on (1) the Hertz theory describing stress in spheres and (2) Weibull statistics for the distribution of flaws shows that the dependence of the breakage probability on the specific work requirement obeys a Rosin – Rammler – Sperling – Bennett distribution (→ Particle Size Analysis and Characterization of a Classification Process, Chap. 2.3.) [26]. Comparison with experimental results confirms that this relationship also holds for a wide range of irregular particles. Empirical studies show that the dependence of the breakage probability on the work requirement approximates to a normal, multiparameter, logarithmic distribution [27].

The breakage fraction in a stressed bed of particles depends on the particle-size distribution and other factors [28]. Figure 13 shows results obtained with a binary mixture of quartz particles. The addition of coarse particles increases the breakage fraction of fine particles and vice versa. The coarse particles concentrate the force flux on the smaller ones, whereas the fine particles distribute the force flux over the larger surface of the coarse particles.

Resistance to Grinding (Grindability). The specific work of comminution required to achieve a defined degree of fineness in a given process is termed the resistance to grinding. The Bond index is an example of such a parameter (see Section 2.3). The applicability of these parameters to other grinding processes is limited.

Breakage Function. The most important result of size reduction is the size distribution of the fragments. The fragment size distribution is referred to as the breakage function when the feed consists of particles of equal size or a narrow size fraction. The breakage function is determined by applying stress to single particles or particle beds; it is also determined by grinding in size-reduction equipment.

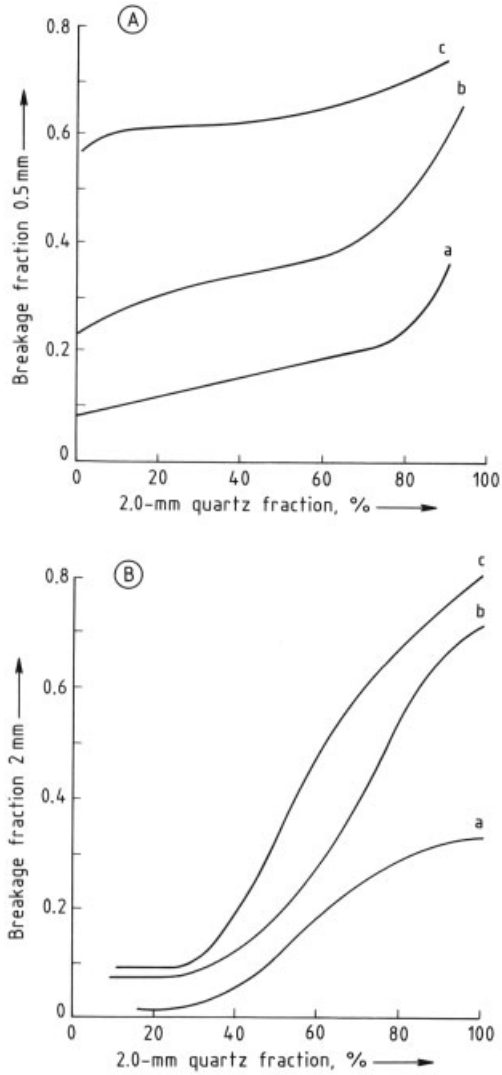

Figure 13. Breakage fractions of 0.5-mm particles (A) and 2.0-mm particles (B) in binary mixtures of 0.5-mm and 2.0-mm quartz fractions
Compression, MPa: a) 14; b) 28; c) 56

To determine the breakage function, stress is applied under predetermined conditions to narrow size fractions, and the resulting particle-size distribution Q is determined. The term $Q\,(x_F - \Delta x/2)$ gives the mass fraction below the

lower limit of the size class Δx. The breakage function B results from the following limiting expression:

$$B\,(x, x_{\mathrm{F}}) = \lim_{\Delta x \to 0} \left[Q\,(x, x_{\mathrm{F}})\,/\right.$$

$$\left. Q\,(x_{\mathrm{F}} - \Delta x/2, x_{\mathrm{F}})\right]$$

The width of a fraction corresponding to a $\sqrt{2}$ relation is sufficiently small. The breakage function depends on the specific work of comminution and other loading conditions:

$$B = B\,(x, x_{\mathrm{F}}, W_{\mathrm{M}}, \ldots)$$

The breakage functions for impacting and compression are different, as explained in Section 2.1. However, experimental results indicate that they may be the same in the fine particle range, i.e., when $x < 0.1\ x_{\mathrm{F}}$. These fine fragments are formed primarily in the contact region where conditions are the same for both types of applied stress.

When individual particles are compressed in the size range $0.05\,\mathrm{mm} < x_{\mathrm{F}} < 5\,\mathrm{mm}$ and the fragment size x is expressed relative to the initial particle size x_{F}, the breakage function depends on the reduction ratio (see Fig. 14):

$$B\,(x, x_{\mathrm{F}}, W_{\mathrm{M}}) = B\,(x/x_{\mathrm{F}}, x_{\mathrm{F}}/s)$$

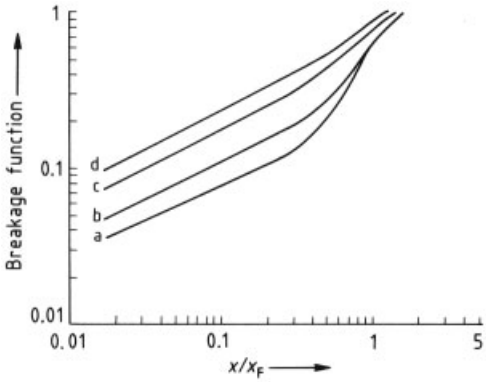

Figure 14. Breakage function of single particles of limestone subjected to compressive stress
Reduction ratio, x_{F}/s: a) 1.5; b) 2; c) 3; d) 4

The breakage function depends principally on the supplied energy and, thus, on the specific work of comminution. Generally, in a bed of particles, only energy up to the energy of fracturing can be transmitted to a particle when an increased load is applied. This is because the fragments can avoid additional stress in the voids as long as the solid fraction remains below 60 – 65 %. In this case, the breakage function is not dependent, or only slightly dependent, on the energy supplied to the particle bed; on the other hand, the breakage fraction increases. Therefore, in particle beds a material-dependent breakage function is always determined, which is referred to as the *natural breakage function.* This does not apply to impact and jet mills, in which stress is applied to individual particles.

Repeated attempts have been made to describe the breakage function mathematically. However, generally valid functions still have to be found. For example, SCHUBERT and coworkers stress that the breakage function can always be represented as the sum of a three-parameteric, logarithmic probability distribution [27].

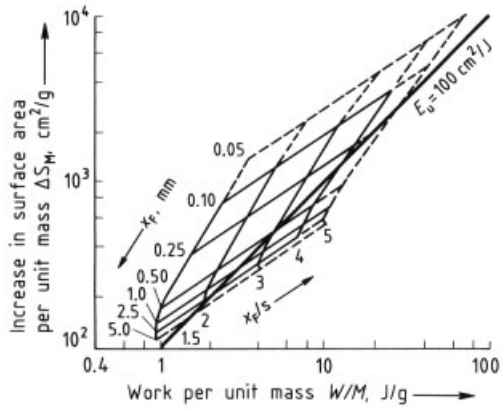

Figure 15. Dependence of surface area on specific work of comminution W_{M} for size reduction of individual cement clinker particles

Energy Utilization. The quotient of the newly created surface area ΔS and the work supplied W is called the energy utilization E_{u}:

$$E_{\mathrm{u}} = \Delta S/W = S_{\mathrm{M}}/W_{\mathrm{M}}$$

Because the value of the specific surface area S_{M} depends on the method of measurement, this parameter must be specified. The gas adsorption method (BET method, → Adsorption, Chap. 4.2.) is not suitable because it also measures internal surface area and surface microstructures. More suitable methods include measuring the permeability of the bulk particles

(Blaine method), the photoextinction method, and calculation from the particle-size distribution (→ Particle Size Analysis and Characterization of a Classification Process, Chap. 2.4.).

Figure 15 illustrates results obtained for the size reduction of individual cement clinker particles. The increase in specific surface area is plotted against the specific work of comminution. Particle size and reduction ratio are variable parameters. The straight line with a gradient of 1 corresponds to constant energy utilization. The measured curves have a smaller gradient, which means that energy utilization decreases as more energy is supplied. This has also been found to be true for size reduction in a particle bed [29].

Because of this phenomenon, size-reduction equipment combined with a downstream classifier must operate at minimal stress levels with multiple applications of stress. However, this means that the amount of final product decreases and the amount that must be recirculated to the mill increases. If the total energy requirement (milling, classifying, transporting) is taken into account, an optimal recirculation factor k_{opt} can be obtained (where k_{opt} is the ratio of the mill and product throughputs) [30]. The optimal recirculating level depends on the ratio of the specific energy required for the additional process steps (W_{ad}) to the consumption energy of the mill (W_{M}) (Fig. 16). Calculations also show that deviations from the optimal recirculating level of $\pm 20\,\%$ lead to an additional total energy consumption of ca. 2 %. At larger deviations, the energy requirements increase significantly.

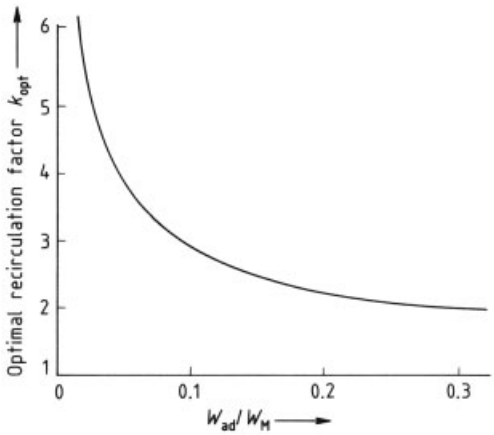

Figure 16. The optimum recirculation factor k_{opt} as a function of the energy required by additional process steps W_{ad}

Figure 15 also shows that energy utilization is higher with smaller particles than with larger ones. This result is at first surprising, because the smaller particles are stronger and require more energy when comminuted in size-reduction equipment. However, this is understandable on a physical basis because the stored stress energy can be utilized more efficiently by smaller particles. Cement clinker particles have an energy utilization of $100-200\,\mathrm{cm^2/J}$ when stress is applied to individual particles. In a ball mill, values are considerably lower, $20-40\,\mathrm{cm^2/J}$. The energy utilization of compression and impacting is compared in Figure 17. Compression shows better energy utilization. However, this is true only for brittle materials, not for viscous substances. Viscous materials cannot be broken down by compression; they must be impacted.

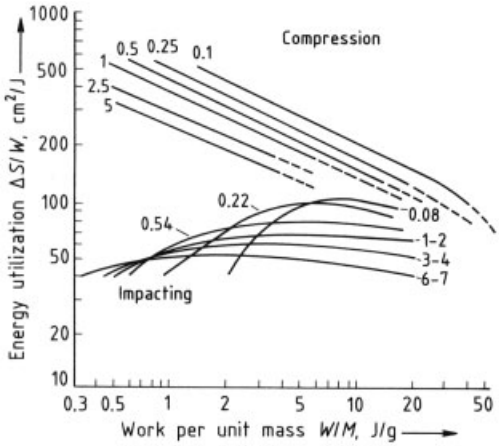

Figure 17. Energy utilization of limestone particles
Values on the curves indicate particle size x_{F} in millimeters.

If application of stress to individual particles is assumed to be the most energy-efficient method of size reduction, the efficiency of other size-reduction processes can be judged by comparison with results from single particle tests [31]. The efficiency is the energy requirement of the ideal process divided by that of the size-reduction machine under the assumption that the products in both cases have the same particle-size distribution.

The ideal process is defined as a multistage sequence in which stress is applied to single particles, it can be described by a mathematical model [30]. The model requires experimen-

tal determination of the breakage function, the breakage fraction, and the energy required to comminute the various particle sizes under different loads. Such data exist only for a few materials. Standard values for the efficiency of size-reduction equipment follow:

Jaw and roll crushers	0.7 – 0.9
Impact crushers	0.3 – 0.4
Roller-ring mills	0.007 – 0.15
Ball mills	0.05 – 0.1
Impact mills	0.01 – 0.1

The poor values for roller-ring and ball mills are due to the fact that stress is applied to a bed of particles. Furthermore, in ball mills the stress is applied stochastically and with a wide variation in intensity. Roller-ring mills require energy for extensive internal circulation of the feed. Impact mills are not very efficient because of the energy needed for ventilation and the limited loading attainable in fine grinding.

2.3. Process Description

An attempt is made to relate the particle size to the energy consumption with three simple relationships, which are referred to as the laws of comminution.

In 1867, Rittinger postulated that the increase in surface area is proportional to the energy consumption and that the energy utilization (the ratio of these two quantities) remains constant (Rittinger's hypothesis). Based on more recent results with the fracturing process, Rumpf [32] showed that in the case of a geometrically similar distribution of cracks, the product ($W_M \cdot x$) is independent of particle size, and geometrically similar fracture surfaces are formed. The energy utilization is then a constant as postulated by Rittinger. In 1885, Kick described a model system which assumed that strength is constant and that the propagation of fracture surfaces is geometrically similar. The energy utilization was concluded to be inversely proportional to the particle size x (Kick's law of similarity). Results obtained from single-particle tests confirm this but also show that strength depends on particle size. Bond [33] developed the following relationship:

$$W_M = W_i \left(10/\sqrt{x_{80,P}} - 10/\sqrt{x_{80,F}}\right)$$

where $x_{80,F}$ and $x_{80,P}$ are the particle sizes of the feed and product in micrometers for 80 % cummulative undersize. The parameter W_i is a measure of grindability (Bond index). The Bond index W_i corresponds to the specific work required to grind from a theoretical infinite size to 80 % passing 100 µm, i.e., x_{80}= 100 µm. Typical values for W_i (in kW · h/t) are: corundum 64; basalt 22; iron ore 16 – 20; cement clinker 15 – 18, feldspar, limestone, and quartz 13 – 15; lead and chrome ores 11 – 13. Additional values are given in [7], [10]. Grindability is determined in a laboratory mill of specified size, according to a set procedure. The Bond relationship permits a first approximation for the design of ball mills; a number of correction factors give better agreement with actual values [33].

Mathematical modeling allows a better description of the size-reduction process. A sucessful model provides answers to the following questions:

1) What state of dispersity can be attained for known material properties and types of stress?
2) How can experimental results from test apparatus or laboratory-scale mills be scaled up to production plants [34]?
3) From what viewpoint should new mills and plans be designed [35] ?
4) What effect do various parameters have on the results of size reduction [36], [37]?
5) How can a plant be controlled?

Extensive development of methods for the mathematical modeling of size-reduction processes began in the late 1960 s. A detailed description is not given here; only the general concept is presented.

The most common method of mathematical modeling, the *population balance model*, describes the change in particle-size distribution with time. To do this, the complete particle-size distribution is divided into fractions (intervals). A mass balance is set up for the individual fractions, and the decrease in mass for each fraction is assumed to be proportional to the mass already present in that fraction. The parameter m_i is the mass fraction of particle-size interval i and w_i is the specific rate of breakage. The term b_{ij} is the breakage distribution parameter of the fragments, i.e., the mass fraction of fragments

from fraction j that move into size fraction i. The stepwise size-reduction process is described by a system of coupled first-order differential equations:

$$\mathrm{d}m_i/\mathrm{d}t = -w_i\, m_i + \sum_{j=1}^{i-1} b_{ij}\, w_j\, m_j;$$
$$i = 1, 2, \ldots, n$$

where the class with the coarsest particles is denoted by the index 1. The first term $(-w_i\, m_i)$ describes the decrease in mass produced by size reduction, and the second describes the increase in mass due to fragments from coarser fractions. If w_i and b_{ij} are constant and $w_i \neq w_j$, this system of equations can be solved explicitly [38]. The problem in mathematical modeling is the determination of the size-reduction coefficients w_j and b_{ij}. For example, dividing the particle-size distribution into ten fractions requires 45 coefficients. Experimental determination is very time-consuming. Usually, the desired parameters are estimated from a small number of measurements (back-calculation method) [39–41]. The size-reduction coefficients, especially size-reduction rates, are not constant, e.g., in wet milling in ball mills [42]. On the other hand, the distribution (transfer) coefficients are less strongly dependent on milling conditions.

The performance of continuously operated plants and milling circuits can also be simulated by mathematical modeling. More complex methods are required because transport of particles through the mill must also be considered. The residence-time distribution or the convection – dispersion model is used [39], [43].

3. Size-Reduction Equipment

The first major step in the development of size-reduction equipment was the utilization of steam power to drive crushers and mills. Development proceeded independently for different kinds of materials (e.g., crushers for road and railroad construction, comminution equipment for ore preparation, and mills for processing grain). New types of machines were invented. Subsequent developments were an increase in equipment size and the utilization of new materials of construction. New methods of construction (e.g., welding instead of casting), and new fabrication techniques (e.g., flame cutting) were also introduced. A major advance in improving energy utilization was the introduction of classification and the development of efficient air classifiers capable of performing sharp separations (→ Air Classifying). Developments have been largely determined by emphasis on efficient energy utilization. However, this has not led to the development of new types of mills. Instead, existing types of mills have been used with new feed materials and new processing techniques [44–49].

Size-reduction equipment cannot be operated without auxiliary devices, such as bunkers, feeders, feed-control and conveying equipment, ventilators, pipes, classifiers, and dust separators. This auxiliary equipment is often responsible for weaknesses and disturbances in size-reduction plants. Additional areas to which attention must be paid include protection against explosions (dust explosion) and, above all, against wear [50].

Size reduction is always associated with wear. In some cases, wear can be minimized, for example, by selecting suitable materials or by modifying the grinding method and the intensity of stress application. However, wear is often intentionally concentrated in one area of the mill, which is then equipped with easily changeable parts. Typical values for wear are 0.5 – 5.0 g/kW · h for dry grinding and 10 – 80 g/kW · h for wet grinding of hard materials. (Wear is expressed as the loss in gram per tonne of material passing through the mill relative to the specific work of comminution in kilowatt hours per tonne.)

A comprehensive, detailed description of size-reduction equipment is to be found in [7]. Aspects such as sturdiness, bearings, lubrication, power supply, overload protection, and other design fundamentals are covered. Information on size of construction and operating parameters such as speed of rotation, throughput and power consumption is also given.

A broad selection of different types of equipment is available as a result of historical developments and the wide range of applications based on particle size, throughput, and material properties. No uniform criteria are available for classifying size-reduction equipment. Furthermore, the system used for naming machines is not standardized, so that a particular type of machine may have several names. The classification

scheme used in the following sections is based on particle size as well as similarities between process types and machine construction.

A clearly defined boundary between crushers and mills does not exist. The term *crusher* is applied to devices which perform coarse size reduction. The term *mill* describes machines which perform fine size reduction where the particle size of the feed is a few centimeters or less.

3.1. Crushers

Crushers can process particles up to 1.5 m in edge length. They are used in mines and quarries; for ore preparation; and for the production of road construction materials (ballast and fine gravel) and raw materials for cement. Their function is to produce particles ready for transport or to perform a primary reduction step prior to subsequent comminution. A more recent application is the recycling of rubble, road surfacing, and metal turnings or scrap.

The use of conveyor belts in mines means that the crusher must be located at the site where mining occurs. This has led to the development of portable or movable crushing installations. The crushers weigh up to 700 t and are equipped with tires, rails, caterpillar treads, or hydraulic advancing mechanisms.

Crushers can be classified into four types: jaw crushers, gyratory (cone) crushers, roll crushers, and impact crushers.

3.1.1. Jaw Crushers

Jaw crushers compress the feed between a stationary and a movable surface.

Double-Toggle Crusher. The double-toggle crusher (Fig. 18) was invented by E. W. Blake in 1858 and is also known as the *Blake jaw crusher.* The movable jaw (a) is pivoted at point (b) and stress is applied to the particles in the crusher opening when the jaw moves. In order to smooth out the discontinuous energy requirements, large fly wheels are mounted on the eccentric shaft. Movement of the jaw is brought about by means of a system of bent levers equipped with a lifting lever (c) and pressure (toggle) plates (d) as well as an eccentric drive (g). A return spring (f) is provided to open the crusher discharge opening or to brace the toggle plate. The narrowest gap width can be changed slightly (by a factor of ca. 1.5) by moving the toggle-plate pivot (e).

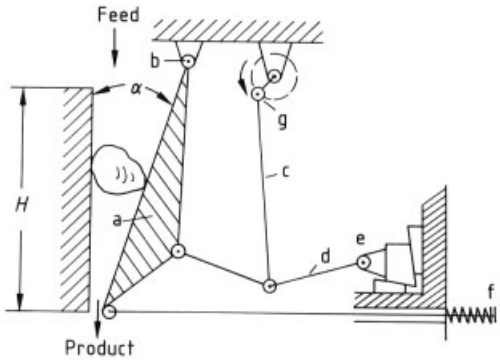

Figure 18. Schematic of a double-toggle jaw crusher
a) Movable jaw; b) Jaw bearing; c) Lifting lever; d) Pressure (toggle) plate; e) Toggle-plate bearing; f) Spring; g) Eccentric

The most important operating parameter is the crushing or nip angle α. It must be so small (14 – 22°) that the particles are held in the crusher cavity by friction when the load is applied. The *reduction ratio* (the ratio of feed particle size to product particle size, e.g., $x_{80,\,\mathrm{F}}$: $x_{80,\,\mathrm{P}}$ where x_{80} is the particle size for 80 % cumulative undersize) is determined by setting the angle α and the height of the crusher cavity H. Reduction ratios of up to 8 are achieved by jaw crushers. The smaller the angle α at a constant stroke, the greater is the distance that the particle may fall as the crusher cavity opens. This means that α affects the throughput of the crusher. Consequently, α has a small value in the region of the discharge opening of fine crushers because here the danger of plugging is particularly high.

Since material in the crusher chamber is transported by gravity, an optimal speed of rotation exists in relation to the throughput. At high rotation speeds, the movable jaw returns before the material has moved downward along the entire theoretical path.

Overhead Eccentric Jaw Crusher. In the overhead eccentric jaw crusher (Fig. 19), also called the *single-toggle crusher*, the movable jaw (a) is directly attached to the eccentric pivot (b). At its lower end, the jaw is supported by a pivot pin (c) above a toggle plate (d) which

is attached to the pivot (e). Therefore, the motion of the jaw shows a significant tangential component compared to the Blake crusher. Although this is advantageous for moving the feed, it increases wear. Variable positioning of the toggle-plate pivot (e.g., $\beta <$ or $> 90°$, where β is the angle between the jaw and the toggle plate) allows the lift and the motion of the jaw in the exit aperture to be varied and adjusted to the type of crushing required (coarse or fine crushing with specially desired particle shapes). The single-toggle crusher is the simplest and lightest of the jaw crushers, but is suitable only for producing low crushing forces.

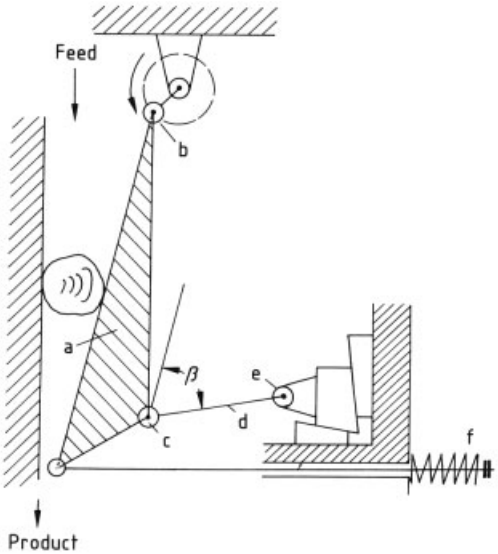

Figure 19. Schematic of an overhead eccentric jaw crusher
a) Movable jaw; b) Eccentric jaw bearing; c) Pivot pin; d) Toggle plate; e) Toggle-plate bearing; f) Spring

In both single- and double-toggle crushers, unbreakable objects may lodge in the crusher chamber. Therefore, they are equipped with overload protection devices which usually consist of a torque overload device built into the fly wheel. Hydropneumatic support of the toggle-plate bearing is sometimes used and has the advantages that the gap width can be easily adjusted or controlled and that the gap can be opened wide.

Impact Jaw Crushers. In an impact jaw crusher, the crusher cavity is inclined (Fig. 20). The jaw (a) below it is attached to the pivot pin (b) and is moved by two laterally mounted, split rods (d), which engage with the crossbar (e). The spring (f) between (c) and (d) serves as an overload protector. The rods are driven by the eccentric bearing (c).

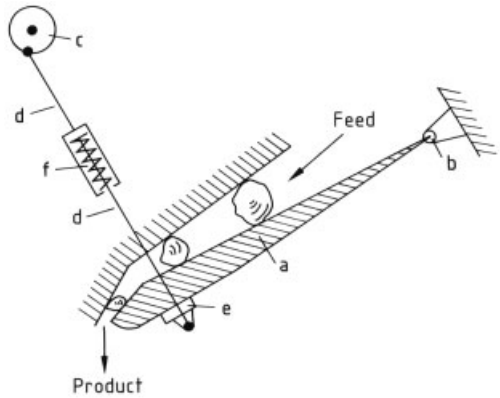

Figure 20. Schematic of an impact jaw crusher
a) Movable jaw; b) Jaw bearing: c) Eccentric; d) Connecting rod; e) Crossbar; f) Spring

As a result of the larger stroke and higher rotation speed (400 rpm compared to 200 – 250 rpm for other jaw crushers), stronger impact is achieved. As a result, hard, tough materials, such as ferrochrome or ferromolybdenum, can be processed.

3.1.2. Gyratory (Cone) Crushers

Stress is applied in a gyratory (cone) crusher in the same way as in the jaw crusher, i.e., between a stationary and a movable surface.

Primary Gyratory Crusher. In the primary gyratory crusher, stress is applied to the feed by pressure as the conical head periodically approaches the bowl (a) (Fig. 21). This periodic movement occurs because the shaft (b) of the crushing head (c) follows a conical surface, i.e., gyrates. The shaft is suspended from a bearing (d) and is guided along a circular path by the lower bearing (e). The angle between the cone axis and the vertical is ca. 1 – 3°. The bearing (e) is set in an eccentric hole in the lower bearing bushing (f). The narrowest gap between the head and the bowl rotates at the same rate as the head shaft; however, the head does not roll along the circumference of the bowl. The gap width can be varied slightly by axial displacement of

the head, particularly if the crusher shaft is supported vertically by a hydropneumatic device, as is now common practice. This feature simultaneously serves as overload protection when, for example, a piece of iron enters the crusher. The ring-shaped form of the crushing chamber allows a constant power input and substantially higher throughputs than in jaw crushers.

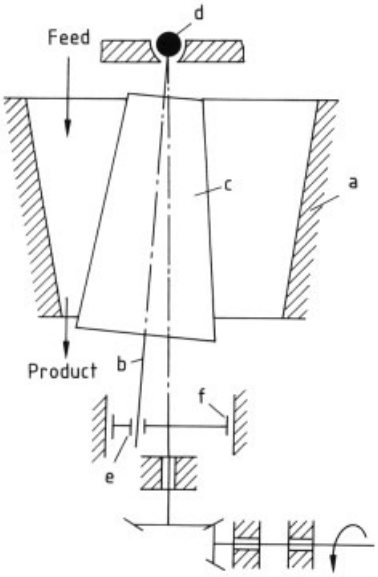

Figure 21. Schematic of a primary gyratory crusher
a) Bowl; b) Shaft; c) Head; d) Upper bearing; e) Lower bearing; f) Eccentric

Table 1. Applications of gyratory crushers [7]

Parameter	Primary gyratory crusher	Shallow cone crusher
Feed particle size x_F, mm	150 – 1800	25 – 300
Product particle size x_F, mm	25 – 250	5 – 40
Throughput $\dot{m}$, t/h	35 – 3500	10 – 600
Specific work of comminution W_M, kW · h/t	0.15 – 0.5	0.4 – 2.2
Type of applied stress	pressure	pressure/impact

The primary gyratory crusher is a large, heavy, and thus expensive machine. Consequently, it is only used for special materials and when a certain throughput is required (e.g., as a primary crusher at $>600\,m^3/h$). It is employed for the coarse and intermediate comminution of hard and moderately hard materials. Operating parameters are listed in Table 1.

Symons Cone Crusher. The shallow cone crusher, also called the *Symons cone crusher*, has a shallower crusher cavity than the primary gyratory crusher (Fig. 22). The crushing head (a) is supported from below (b), requiring less machine height. The design of the bowl (d) produces high reduction ratios up to 18. The relatively long, parallel gap preceding the discharge aperture ensures a uniform product particle size and good particle shape. The shallow cone crusher cannot handle such large feed particles as the primary gyratory crusher because the diameter of the discharge aperture is larger than that of the feed aperture. As in the impact jaw crusher, the stroke is larger and the speed of rotation higher than in the gyratory crusher. The stroke h is ca. 2 – 5 times the gap width s; the speed of rotation is 200 – 300 rpm compared to 100 – 200 rpm for primary gyratory crushers. These values are favorable for material transport and give a better shaped particle, i.e., more cubic. Gap adjustment and overload protection can be provided by supporting the bowl (d) with springs or by supporting the bearing (b) with springs or a hydraulic device.

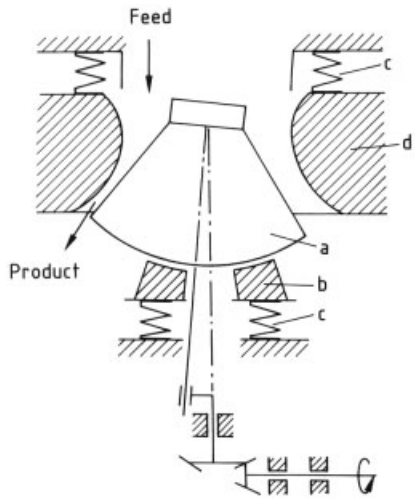

Figure 22. Schematic of a shallow cone crusher
a) Head; b) Bearing; c) Spring; d) Bowl

Shallow cone crushers are used mainly for the fine crushing of hard and moderately hard materials. Operating parameters are listed in Table 1.

Disk Crusher. The disk crusher (Fig. 23) was developed to overcome the disadvantages of the shallow cone crusher (i.e., plugging and low throughput) and is basically a centrifugal gyratory crusher. The crushing head (a) is extended to form a disk, resulting in a radial gap. The support (b) is driven at an angular velocity ω_1. A gear (c) ensures that the narrowest gap rotates with velocity ω_2. The feed falls onto the center of the rotating disk and is transported through the gap by centrifugal force. The stator disk (d) rotates and is driven either by friction in the gap or by a separate motor. To provide overload protection, the cover and the stator disk are supported with springs; they can both be moved axially to adjust the gap.

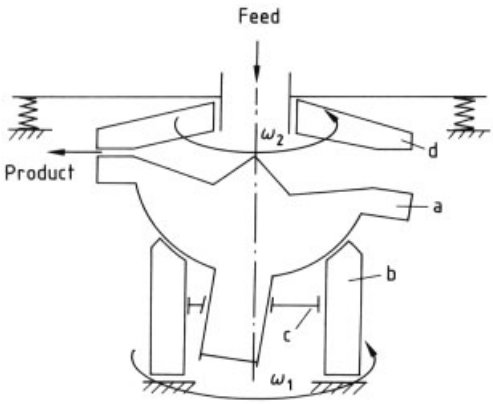

Figure 23. Schematic of a disk crusher
a) Crushing disk; b) Support for crushing disk; c) Gear; d) Stator disk

3.1.3. Roll Crushers

Double-Roll Crushers. Double-roll crushers consist of two rollers, which rotate toward each other and are separated by an adjustable gap (Fig. 24). One roller (a) is held by stationary bearings, the bearings of the other (b) are held by springs or a hydraulic system and can move along a linear or circular path. The material is fed in from above. The rollers, which may have a shaped surface, draw in the feed and apply stress to it. The crushers are designated as cam, toothed, gear, knob grooved, or helical according to the roll profile. The profile improves the drawing in of the feed and therefore helps to increase the reduction ratio. Rollers with diameters of up to 2 m can handle feed sizes up to 1 m. With soft to moderately hard materials, throughputs up to 3500 t/h may be obtained.

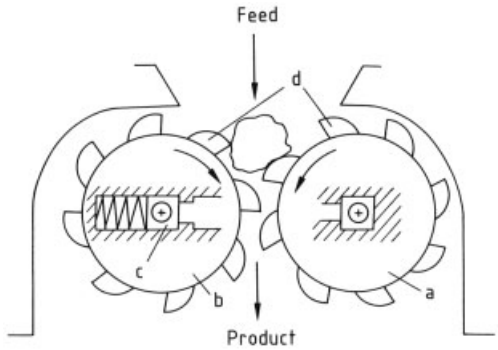

Figure 24. Schematic of a double-roll crusher
a) Roller held by stationary bearing; b) Roller held by movable bearing; c) Movable bearing; d) Tooth

Both rollers usually rotate at the same rate, but they may be set at different speeds in the case of soft, sticky, wet materials such as coal, lignite, peat, salt, clay, or pumice. This has the advantage of producing an additional shearing or cutting stress. The peripheral velocity of the roller controls the amount of material passing through the gap. For coarse crushing, values range up to 10 m/s.

Single-Roll Crusher. In the single-roll crusher, a shaped crushing roller acts against a crushing plate, which is held either by springs or a hydraulic system. This simple design is suitable for coarse crushing of moderately hard to soft materials.

In the *continuous crusher* (Fig. 25), the toothed roller (a) acts against a conveyor belt (b) rather than a crushing plate. This feature makes the crusher especially suitable as a mobile comminution device for use in mines and quarries. This crusher has long been used underground for coarse crushing of coal and salt, and later for the comminution of limestone and abrasive materials such as rubble.

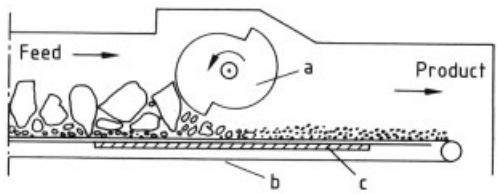

Figure 25. Schematic of a continuous single-roll crusher
a) Crushing roller; b) Mechanical conveyor; c) Support for conveyor

3.1.4. Impact and Hammer Crushers

Impact Crushers. A typical impact crusher is depicted in Figure 26. The feed slides through the feed chute (a) and is hit by the impact bars (c), which are firmly attached to the rotor (b) and occupy its entire width. The particles and fragments are thrown against the impact faces or plates (d), and are then redirected to the impact bar zone. The impact plates are mounted flexibly, so that they can retract when a large piece of material enters the crusher. A grinding path (e) can also be included; it primarily applies additional stress to coarse particles and promotes a better particle shape. The crusher housing is open at the bottom. The reduction ratio and the particle size are determined by the peripheral speed of the rotor and not by the gap between the rotor and the impact path. Reduction ratios of 20 or higher are possible.

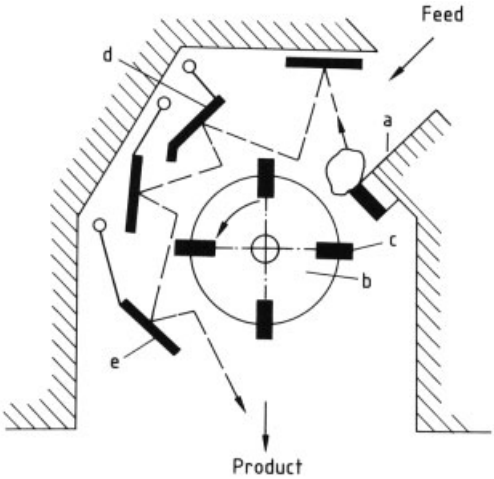

Figure 26. Schematic of an impact crusher
a) Feed chute; b) Rotor; c) Impact bar; d) Impact plate; e) Grinding path

Hammer Crushers. In the case of hammer crushers (Fig. 27), impact devices (hammers) (a) are attached to the rotor via pivots so that they are deflected when they hit strong, particularly large particles. In most cases the crushing zone is surrounded by grate bars (b), so that fragments which are larger than the openings of the grating are retained in the crushing zone. Hammer crushers with rotor diameters up to 3 m are available and can crush ca. 1500 t/h of limestone.

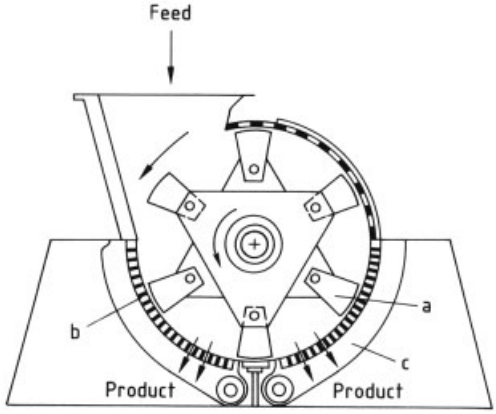

Figure 27. Schematic of a hammer crusher with a grating
a) Pivoted hammer; b) Grate bar; c) Support for grating

Although hammer crushers wear more quickly than impact crushers, they can process moist materials more efficiently. Impact and hammer crushers operate with peripheral velocities between 20 and 50 m/s (maximum velocity 70 m/s). In principle, only soft to moderately hard materials can be processed because of wear considerations at these high peripheral velocities. However, these crushers are simpler than jaw and cone crushers and units with equivalent throughputs are much smaller in size. Consequently, impact crushers have been developed for the processing of hard rock such as basalt, diabase, and graywacke. The peripheral velocity is then decreased to 25 – 35 m/s.

The *crusher with a vertical shaft* (Fig. 28) has been specially developed for low-wear operation. The feed is introduced into the center (a) of the rotating disk (b). The particles are accelerated by radial bars (c) and fly off the rim of the disk against a surrounding stator (d) where stress is applied by impact. To reduce wear, the acceleration bars and the stator have recesses in which the particles collect and interparticle contact is thus maximized. However, interparticle contact produces relatively little size reduction. The crusher is often used to improve particle shape. The reduction ratio can be improved by using a metal stator.

Special hammer crushers known as *shredders* are used in the disintegration of scrap metal for the recovery of raw materials. A shredder is depicted in Figure 29 A. Two rollers (a) control feed entry; their speed is variable and the clearance between them can be adjusted according

to the feed particle size. Stress is applied to the pieces of scrap on the anvil (b) at the full peripheral velocity. Two types of anvil are shown in Figure 29 B. In the toothed design, correspondingly longer and shorter hammers are used, so that they can mesh with the anvil. The fragments or deformed pieces are thrown against a grating (d). Pieces which are not sufficiently reduced in size can be removed from the shredding zone via a door (e).

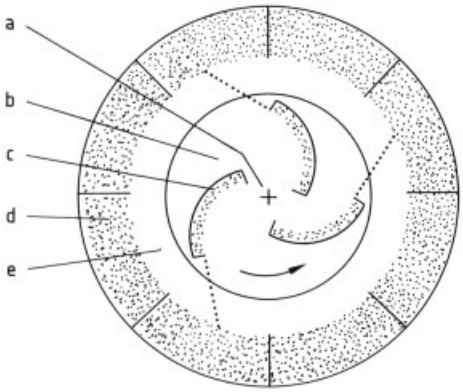

Figure 28. Schematic of a crusher with a vertical shaft viewed from above
a) Central feed inlet; b) Rotating disk; c) Acceleration bars with particle cushions; d) Fixed ring of particles; e) Product discharge

3.2. Grinding Media Mills

The size reduction devices discussed in this section consist of a mill vessel containing a moving grinding medium. Balls, rods, short bars (cylpebs), or coarse particles of the feed material itself (autogenous grinding) are used as grinding media. The feed trapped between the media particles is ground by their movement.

In the case of *tumbling mills* (Fig. 30 A), the tubular mill vessel rotates around its own axis. The grinding medium is elevated by the vessel wall, and then rolls, slides, or falls toward the lowest point of the vessel.

The mill vessel in a *vibration mill* moves in a circular path, without rotating around its own axis (Fig. 30 B). This sets the grinding medium in motion and produces a grinding action. The pathway of the moving grinding media depends on the diameter of the circular oscillation and the charge of the grinding media.

In the *planetary mill* (Fig. 30 C), the mill vessels rotate with an angular velocity ω_2 on a carrier system which itself rotates with an angular velocity ω_1. Depending on the combination of rotational speeds selected, the type of grinding media motion can be adjusted to correspond to that of either a tumbling mill in a centrifugal force field or to a vibraton mill with a very large diameter of oscillation.

In theory, the *agitated ball mill* (Fig. 30 D) can also be classified in this group, because the agitator moves a grinding medium. However, since this type of mill is primarily operated as a wet mill (the material is fed and ground as a suspension), it is described in Section 3.6.

3.2.1. Tumbling mills

Tumbling mills have a horizontally supported cylindrical (or conical – cylindrical) mill vessel. The mills can be named, for example, according to (1) the shape of the grinding medium or the mill vessel (ball, rod, drum, tube, or cone mills); (2) to its connection with a classifier (open or closed circuit); or (3) according to the method used for product removal (air stream mill). The length to diameter ratio (L/D) varies from ca. 0.25 (autogenous mill) to ca. 6 (continuous mill). Mill sizes range from 300×200 mm ($L \times D$) for laboratory use, up to 18×6 m. In most cases, the mill is operated continuously in combination with a classifier (closed circuit), which determines the particle-size distribution of the resulting product.

Tumbling mills are suitable for dry or wet grinding. General uses are the fine grinding of hard (to soft) and abrasive materials such as ores, cement clinker, coal, quartz sand, and limestone at high throughputs (up to 500 t/h for closed circuit mills, the product stream then being ca. 200 t/h).

Figure 31 shows a vertical section through a tube mill. The mill vessel is supported at both ends with trunnion bearings (a). The inside of the vessel is lined (b). The lining primarily protects against wear, but it also assists the motion of the grinding medium (better elevation by means of lifter bars) and classifies the grinding balls. The large balls are mainly found near the mill inlet (c) and the smaller ones near the outlet (d) (classification lining). This classification effect is nec-

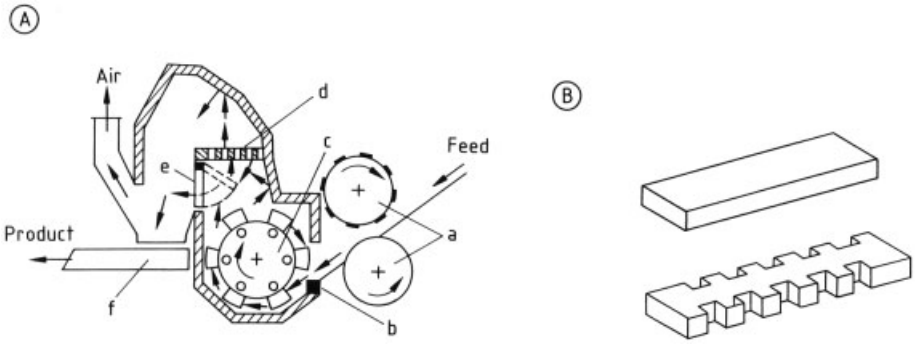

Figure 29. Schematic of a shredder (A) and two anvil designs (B)
a) Feed rollers; b) Anvil; c) Rotor with hammers; d) Grating; e) Hinged door for removing uncrushed material; f) Vibration feeder

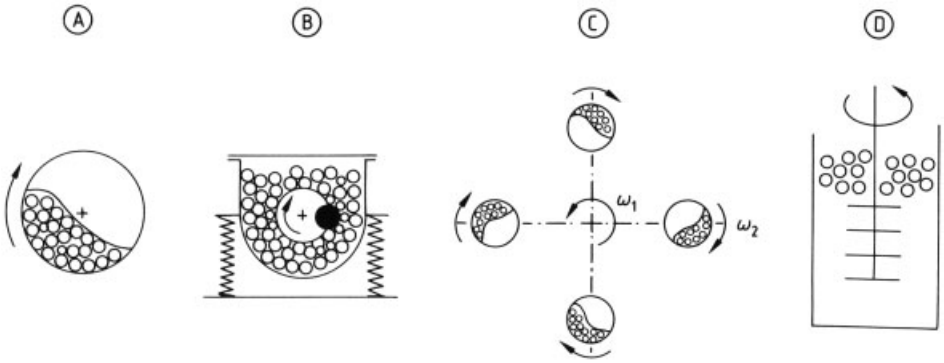

Figure 30. Schematic of mills using grinding media
A) Tumbling mill; B) Vibration mill; C) Planetary mill; D) Agitated ball mill

essary because an increasing degree of fineness along the mill axis requires smaller grinding media; it is often achieved by dividing the mills into chambers (multicompartment mills). The partition walls (e) have perforations which are large enough to allow the feed to pass through but are small enough to retain the grinding medium.

In the absence of an air stream, the particles can only pass through the walls if an axial force is exerted. This is the case in the ascending zone of the solids. The *throughput* $\dot{m}$ through a perforated barrier can be estimated as described in [51]:

$$\dot{m} \approx 0.35\,(1 - 0.6 d_L/d)\, q_L\, q_A\, \varrho_S D^2 \sqrt{g d_L}$$

where d is the diameter of the ball, d_L the diameter of the perforations, q_L the relative area of the perforations, q_A the relative area of the ascending zone, ϱ_S the density of the solid, D the mill diameter, and g the acceleration due to gravity. Thus, critical factors for the throughput are the relative area of the perforations (as a constructional parameter) and the size of the ascending zone (as an operating parameter).

The *motion of the grinding medium* can be described as cascading, falling (or cataracting), or centrifugal (Fig. 32).

Centrifugal motion (Fig. 32 C) begins at the critical rotational speed at which the centrifugal acceleration is equal to the acceleration due to gravity. The critical speed in revolutions per minute (n_c) is commonly defined as

$$n_c = 42.3/\sqrt{D}$$

where D is the mill diameter in meters. Tube mills are operated at speeds below this value. Because a significant amount of slippage may occur between the mill vessel and its contents, the type of motion cannot be deduced directly from the speed of rotation.

The *power consumption* of the mill can be explained by means of the motion of the mill

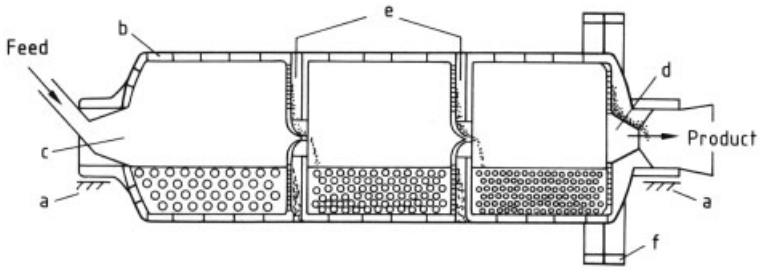

Figure 31. Vertical section through a tube mill
a) Trunnion bearing; b) Liners; c) Tube inlet; d) Tube outlet; e) Intermediate diaphragm with lifter; f) Drive gear

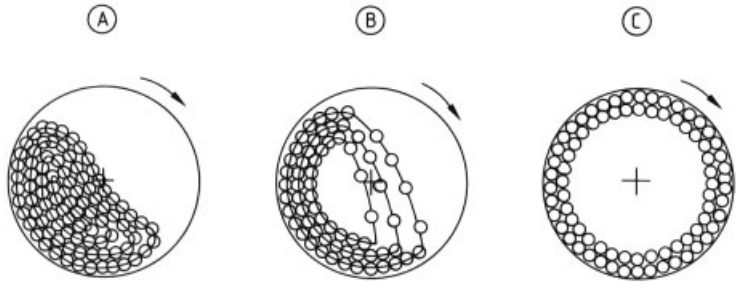

Figure 32. Types of motion in a ball mill
A) Cascading; B) Falling or cataracting; C) Centrifugal

charge. When the mill is stationary, the center of gravity of the charge is adjacent to the midpoint of the mill (Figs. 32A and B). The weight of the charge times the unknown lever arm gives the driving torque from which the following performance formula can be derived:

$$P \sim n/n_c M \sqrt{D}$$

where P is the operating power and M the mass of the charge. The mass fraction of the feed is usually less than a fifth of the total mass of the charge. This relationship was described by Blanc and Eckhardt in 1928. The corresponding proportionality constant has values ranging from 6 to 9, if $n/n_c = 0.75$. If the mass is given in tonnes and the diameter in meters, the calculated power is in kilowatts. The mass of the charge can be expressed with the aid of the mill volume. The effect of the length L and diameter D of a mill on power consumption is given by

$$P \sim n/n_c L D^{2.5}$$

The parameter P is only proportional to the speed of rotation up to a certain value, i.e., $n/n_c = 0.68 - 0.75$. Above this range the power consumption decreases sharply due to the centrifugal motion of individual balls as shown in Figure 33 [52].

The relationship between the charge of grinding medium (expressed as the ratio of the bulk volume of the grinding medium to the mill volume) and power consumption was investigated by Rose and Sullivan [53]. A maximum was found at a loading ratio of 0.4 (Fig. 34). Optimal operating values are in the range 0.27 – 0.33.

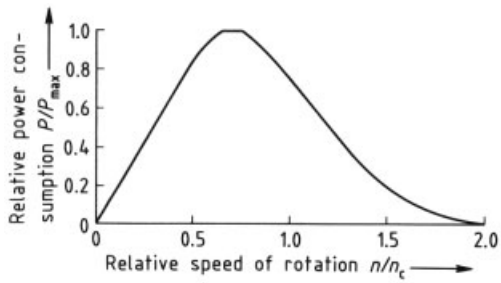

Figure 33. Dependence of power consumption on the relative speed of rotation for dry grinding with lifters

The amount and the motion of the grinding medium determine the power consumption of tube mills. However, the energy must be efficiently transmitted to the particles by selecting

a correct size for the grinding medium. For example, the impact energy of a ball must be sufficient to break the largest particle, but must not compact the resulting fines into agglomerates. Examples of empirical formulas for the selection of grinding media size are

$$d = 28\ \sqrt[3]{x} \quad \text{or} \quad d = 6 \log\left(x_{\mathrm{P}}\ \sqrt[3]{x_{\mathrm{F}}}\right)$$

where x_{F} and x_{P} are the feed and product particle sizes in millimeters. To avoid sliding and oscillation of the ball charge, the diameter of the balls should be at least $\leq D/20$ and is usually 20 – 80 mm.

The literature often states that wet grinding is more energy efficient than dry grinding by as much as 20 – 30 %. However, systematic comparison has shown that under comparable operating conditions a maximum energy saving of 10 % can be expected [52]. In wet grinding, wear is up to 10 times higher. For wet grinding the volume fraction of solid c_{v} should be in the range 0.35 – 0.4.

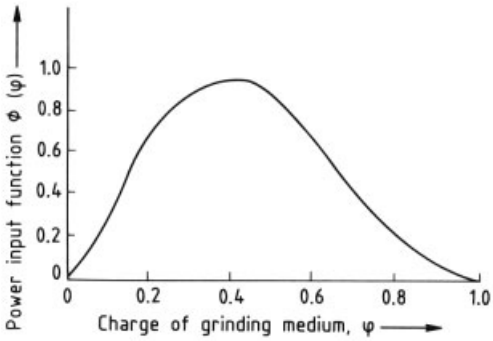

Figure 34. Dependence of power consumption on the charge of grinding medium

Autogenous Mills. Tumbling mills with a small length: diameter ratio of 0.25 – 0.4 which use the coarse feed particle as grinding bodies (possibly with up to 10 % of a steel grinding medium) are termed autogenous mills (Fig. 35). They are used to process ores, refuse, and slag. They can either be operated wet (*Hydrofall mills*) or dry (*Aerofall mills*) in a continuous mode. The mill diameters are large compared to tube mills (up to 14 m with a power consumption of ca. 15 MW) and their relative speed of rotation is somewhat higher (n/n_{c} = 0.75 – 0.8). Lifting and breaking bars increase the elevation of the feed. Addition of steel grinding balls (diameter 75 – 150 mm) makes operation independent of fluctuations in the size of the feed particles. It also increases impact stress which leads either to a decrease in mill diameter or to an increase in throughput.

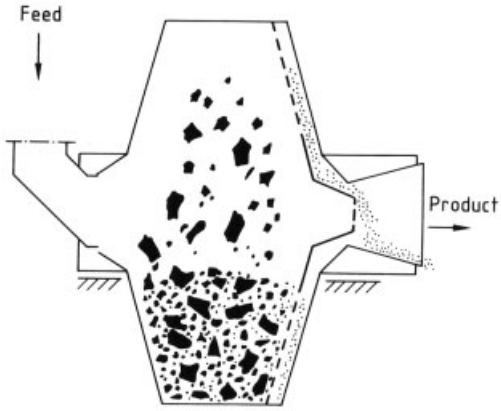

Figure 35. Schematic of an autogenous mill

3.2.2. Vibration Mills

In a typical vibration mill, the mill vessel is a tube (a) which is filled 65 – 80 % with spheres, rods, or cylpebs (Fig. 36). The tube is held in a frame (b), which is supported by springs (c). It is moved in an almost circular fashion by an unbalance (d).

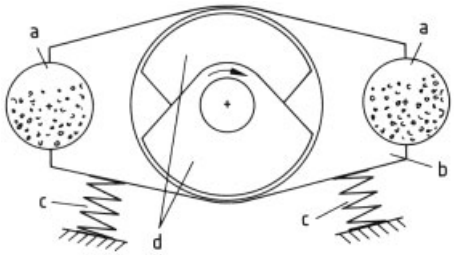

Figure 36. Schematic of a vibration mill
a) Mill tube; b) Frame; c) Spring; d) Unbalance

Vibration mills are much smaller than tube mills (maximum diameter = 0.65 m, length = 4 m) and their throughput is lower (maximum is ca. 15 t/h). Like tube mills, they are suitable for processing hard and abrasive materials. The vibration mill can easily be adjusted to handle different feeds (for example, by changing the oscillation diameter or the residence time). The feed particle size is smaller than that of the tube

mill (5 – 10 mm compared to 25 – 40 mm). The throughput largely controls the fineness of the product. As a result, the mill is usually operated in a continuous mode without a classifier.

Vibration mills are used to grind fired clay, bauxite, silicon carbide, quartz sand, barium ferrite and even limestone and talc. Wet grinding is possible but is rarely carried out in practice. Special applications include the grinding of dried lignite (because of the low risk of dust explosion); use as a leaching reactor in hydrometallurgy; and activation grinding, (i.e., very frequent, prolonged application of stress to produce crystal modification, see Section 2.1).

Under typical operating conditions (oscillation diameter 6 – 14 mm, speed of rotation 1000 – 1500 rpm) and with a grinding medium charge > 50 % of the mill volume, each particle of the grinding medium follows an almost circular path. The circle diameter is roughly equal to the oscillation diameter and the angular velocity is equal to that of the mill. In addition, the entire charge rotates slowly at a speed about one hundreth that of the mill and in the opposite direction [54].

Above oscillation diameters of ca. 20 mm (which are not yet used on an industrial scale), a different type of motion takes place. It is characterized by rotation of the center of gravity of the grinding-medium charge at the frequency of the mill and in the same direction. Raasch has derived an expression for the associated power consumption of the mill, based on a force balance [55]:

$$P = c/r_s M r_s^2 \omega^3$$

where r_s is the amplitude of oscillation, M is the mass of the grinding medium, and c is the distance between the center of gravity of the grinding-medium charge and the center of the mill tube. The parameter c is not constant and its value accounts for the variation of the power consumption as a function of the grinding medium charge which is similar to that shown in Figure 34. Even though the type of motion is different at lower oscillation diameters, the equation may also be used in this range. The equation can also be derived from the impact energy of the charge [55]. The author's own studies have shown that the product fineness is primarily determined by the energy per impact.

At high grinding medium charges (above ca. 70 % of the mill volume), the power consumption decreases as the amount of solid increases. This is because the amplitude of oscillation decreases and the added substrate moves the center of gravity of the solid charge toward the center of the tube. With lower grinding medium charges, the diameter of oscillation is scarcely affected and the power consumption increases with the amount of solid.

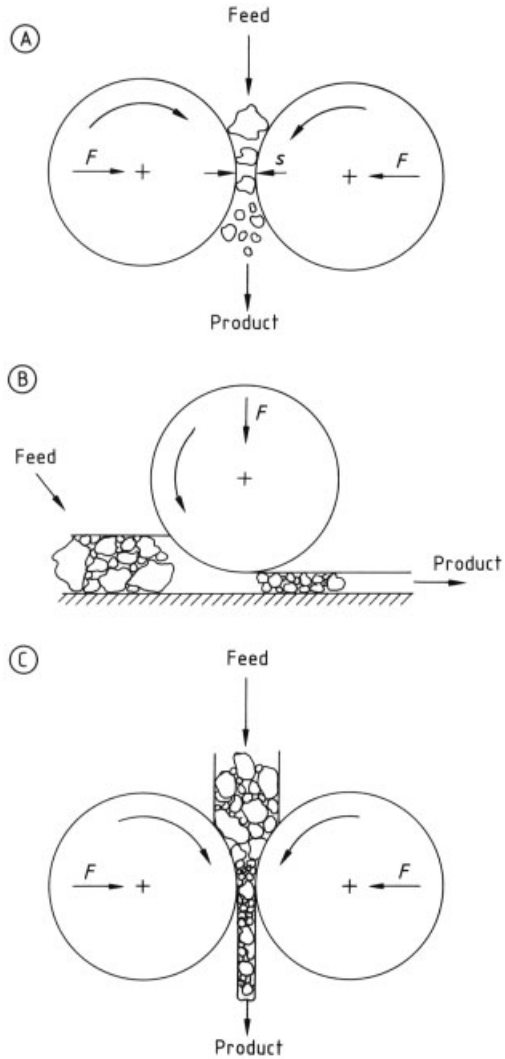

Figure 37. Schematic of a roller mill (A), a ring-roller mill (B), and a high-compression roller mill (C)
F = force; s = gap width

3.2.3. Planetary Mills

In planetary mills (Fig. 30 D), high stress intensities can be obtained as a result of the greater acceleration. Considerable design problems are associated with this device on a production scale, especially if it is to be operated in a continuous mode. This has hindered the broad application of this system. Although new developments have been made and their associated advantages have been reported, e.g., [56], the only units available commercially are laboratory models. They are used for the batchwise preparation of solids for chemical analysis. The use of grinding media and mill vessels made of semiprecious stone prevents iron contamination.

3.3. Roller and Ring-Roller Mills

In *roller mills* stress is applied to the feed in the gap between two counterrotating rollers (Fig. 37 A). In a *ring-roller mill*, different types of rollers roll over a compressed layer of the feed (particle bed), which is supported on a grinding ring (Fig. 37 B).

Roller mills have two modes of operation: at the start of compression, the feed between the rollers consists of a layer of single particles (Fig. 37 A) or, in the case of small particles, a bed of particles (Fig. 37 C). In the latter case, a higher pressure is applied in the gap and, as a result, agglomerates are usually formed. The term *high-compression roller mill* has been applied to this mode of operation [48], [49].

3.3.1. Operation Characteristics

Even though ring-roller, roller, and high-compression roller mills are very different in construction, they can be discussed together because of the similarity between the zones in which stress is applied. The conditions in the grinding gap furnish information concerning throughput, power consumption, and the product fineness. Operating characteristics of the three mill types are listed in Table 2. The feed particle size x_F has a maximum value relative to the diameter of the roller D (see Table 2). The *throughput* between the rollers is given by

$$\dot{V} = c_v\, s L u_p$$

or

$$\dot{V} = c_v\, s/D \cdot D^2\, (L/D)\, u \cdot u_p/u$$

where $\dot{V}$ is the volumetric flow rate of the solid, c_v its volume concentration, s the gap width, D the roller diameter, L the length of the roller, u the peripheral speed of the roller, and u_p the velocity of a particle in the gap.

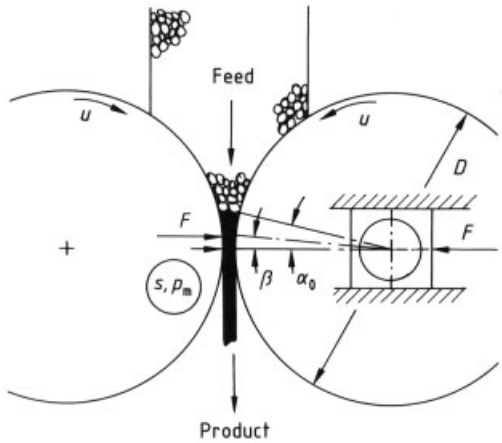

Figure 38. Comminution of material between two rollers D = roller diameter; F = grinding force; p_m = maximum pressure in gap; s = gap width; u = peripheral velocity; α_o = compression angle; β = angle of force action

The term $c_v\, s/D$ is dimensionless and is called the *specific throughput*. It is specific to the mill and the feed material, and is a characteristic of the drawing-in process which is a complex interaction between the feed and the roller. The term $D^2\,(L/D)\,u$ contains design parameters. The term u_p/u is approximately equal to 1 because, in most cases, slippage between the feed and the roller can be prevented. The equation shows that, without slippage and with constant drawing-in conditions, the throughput increases linearly with the peripheral speed of the roller. However, an upper limit exists, above which the expression is no longer valid. Operating throughputs for roller mills range from 2 to 6 m/s and for high-compression roller mills from 0.5 to 1.0 m/s. In the case of geometrically similar mills (i.e., equal values of L/D and u, and equal drawing-in conditions), the throughput is proportional to the square of the roller diameter. A similarly simple

Table 2. Comparison of roller, ring-roller mills, and high-compression roller mills

Parameter	Roller mill	High-compression roller mill	Ring-roller mill
Stress application	single particles	particle bed	particle bed
Adjustable parameter	gap width	force	force
Specific grinding force *F/DL*, MPa∗		1 – 5	0.5 – 1
Volume fraction of solid in gap	<0.10	0.7 – 0.9	0.6 – 0.7
Gap width	>0.5 mm	*D*/100 – *D*/50	*D*/200 – *D*/70
Maximum feed particle size x_F	*D*/20	*D*/40	*D*/40
Feed material	moderately hard to soft	hard to soft	moderately hard to soft

∗ *D* = roller diameter; *F* = applied force; *L* = roller length

method cannot be used for estimating throughput in ring-roller mills, because the feed is not forced through. An estimate for the throughput of ring-roller mills with disk-shaped grinding rings is cited in the literature as

$$\dot{V} \sim d^{2.65}$$

where *d* is the diameter of the disk.

The *driving torque T* of a single roller in a roller or ring-roller mill is given by

$$T = F\,D/2\sin\beta$$

where β is the angle of force action (Fig. 38). The angle β can be determined by measurement of the grinding force *F* and the torque, values are between 2.5 and 8 °.

Multiplication of the driving torque by the angular velocity gives the *power consumption P*:

$$P = 2\sin\beta \cdot D^2 L/Du \cdot F/LD$$

The term *F/LD* is called the *specific grinding force* F_{sp}. It has the dimensions of pressure and is a measure of the maximum pressure in the gap ($p_m \sim F_{sp}$). Measurements give $p_m \approx 45\,F_{sp}$.

The equation for power consumption is also suitable for scale-up purposes. For geometrically similar mills and equal specific grinding force, they give the result $P \sim D^2$. Ring-roller mills with disk-shaped grinding rings are operated so that the radial acceleration at the rim of the disk remains constant, i.e., $u \sim \sqrt[3]{d}$. Thus, for ring-roller mills $P \sim d^{2.5}$.

Grinding in roller mills approximates very closely the comminution of single particles. As a result the reduction ratio *x/s* determines the specific energy input (see also page 13) and thus the product fineness. In the case of particle-bed comminution, the pressure in the gap is the major parameter affecting product fineness. Use of the above equations and the equation defining the specific work of comminution shows that the reduction ratio is primarily determined by the specific work:

$$W_M = P/V\,\varrho_s \sim F_{sp} \sim p_m$$

3.3.2. Roller Mills

The design of roller (roll) mills is equivalent to that of roll crushers (Section 3.1.3). The movable bearing is usually supported hydropneumatically; this provides overload protection and allows easy adjustment of the grinding force. The rollers are 0.5 to 1.5 m in diameter and, in contrast to crushers, are smooth or grooved. The reduction ratio lies between 3 and 5. The feed must have a narrow particle-size distribution to ensure that stress is not only applied to the coarsest particles in the mill gap. The size of the coarsest product particles is of the same order of magnitude as the gap width. Due to constructional factors, the gap width has a minimum value (0.5 mm) and ultrafine grinding is therefore not possible.

Roller mills are especially suited for the abrasives industry and the milling of cereals. They are also used for processing clays, especially for the comminution of hard mineral inclusions. Operating parameters are listed in Table 2.

Roller Frame. The roller frame is a roller mill whose rollers operate at different speeds of rotation (friction). The peripheral speed of the faster roller can range up to 8 m/s. The ratio of

rotation speeds lies between 1 : 1.3 and 1 : 3. The rollers are often shaped (grooved) and their diameters are ca. 0.25 m.

Roller frames are used primarily for milling cereals. A double-roller frame is depicted in Figure 39. The feed passes over grooved rollers (a) to ensure a uniform feed coating. Scrapers (c) or brushes (d) remove adhering feed particles from the crushing rolls (b). Machines with rollers turning at the same speed are called *crushing roller frames.* They are used to produce flakes from vegetable products (e.g., oat flakes).

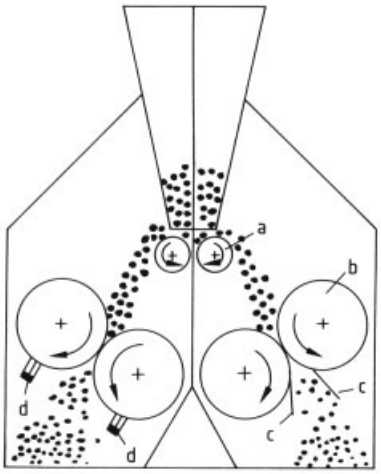

Figure 39. Schematic of a double-roller frame
a) Grooved feed roller; b) Crushing roller; c) Scraper; d) Brush

Roller frames are also used for the comminution of hard waxes or brittle materials in the chemical industry, because the amount of fines in the product is low. This is especially true at low reduction ratios x_F/s (see page 13).

High-Compression Roller Mills. The construction of a high-compression roller mill differs from that of a roller mill in that it is stronger because it must exert a higher grinding force. Because agglomerates form as the feed passes through the mill, additional comminution equipment is required. The desired degree of fineness cannot be obtained in a single pass even at very high pressure. Consequently, the mill must work in conjunction with a classifier, the classifier determines the product fineness. Particle-size distributions with upper limits of ca. 0.01 – 1 mm can be achieved.

Mills with diameters D up to 2.8 m have been built, but a diameter of 1 – 1.5 m is most common. The L/D ratio lies between 0.3 and 1.0. Throughputs range from 500 to 550 t/h.

The mills can operate with dry feeds or moist feeds containing up to 15 – 20 vol % water. Materials such as cement clinker, limestone, marble, slag, foundry sand, quartz, silicon carbide, ores, and kimberlite have been processed with high-compression roller mills. The specific grinding force lies between 2 and 5 N/mm^2 depending on the hardness of the material. The energy required for deagglomeration ranges from 0.2 to 0.8 kW · h/t, depending on the material, and corresponds to ca. 2 – 5 % of the energy expended in the comminution process. The specific wear is roughly 10 – 20 times smaller than that of a ball mill. This is because a major portion of the applied energy is distributed within the particle bed and only a small amount of material contacts the rollers.

3.3.3. Ring-Roller Mills

Many different mill designs are available for applying stress to bulk solids between rollers and a grinding surface. They differ in the shape of the roller and the grinding ring as well as in the way in which the grinding force is exerted. The ring, pendular roller, bowl, and ring ball mills depicted in Figure 40 operate in closed circuit with an integrated classifier. An air current carries the ground material to the classifier which returns the coarse material to the grinding ring. The classifier determines the final degree of fineness; particle sizes smaller than 30 – 40 µm can be obtained.

Depending on the required fineness and the feed material, the recirculation rates of the milling system may reach very high values. In a plant for milling cement raw material, recirculation rates up to 18 were observed (i.e., the amount of recirculated material was 18 times the product output) [46], [57]. However, usual values range from 3 to 6. A large fraction of the total energy consumption (ca. 40 %) can be attributed to recirculation of material. However, the total specific energy requirement of such mills is ca. 25 % less than that of tube mills [45].

A milling process with external recirculation is described in [7], [46]. Here, the ground mate-

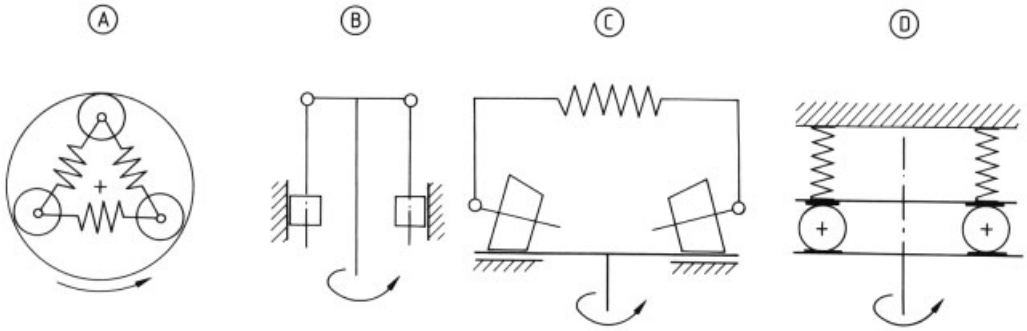

Figure 40. Schematics of the most important industrial ring-roller mills
A) Ring mill; B) Pendular mill; C) Bowl mill; D) Ring ball mill

rial is transported to a classifier above the mill by means of a bucket conveyor. This significantly lowers the transport energy requirements. The large volumes of air required for material transport are often available as hot gas supplied at a temperature up to 600 °C, drying by as much as 20 wt % is then possible.

Ring Mill. The ring mill depicted in Figure 41 comprises three rollers (a) which support a freely rotating grinding ring (b). One of the three rollers is driven. The grinding ring rotates as a result of friction and in turn drives the two other rollers. Usually, the rollers are cambered and the grinding ring is correspondingly concave. The springs (c) produce the applied pressure. The feed is fed in from lateral chutes but remains on the grinding ring due to centrifugal force, until it is displaced by additional material. This American design has not been able to penetrate the European market.

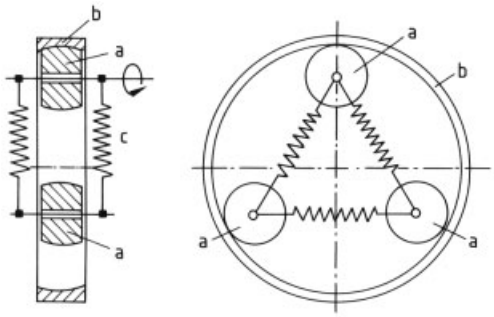

Figure 41. Schematic of a ring mill with three rollers
a) Roller; b) Grinding ring; c) Spring

Pendular Mill. The pendular mill (Fig. 42) has 2 – 5 suspended, swinging rollers (a) that are pressed in the direction of the stationary grinding ring (b) by centrifugal force (peripheral speed ca. 7 m/s, maximum diameter of the grinding face 2.5 m). The rollers are driven by rotation of the shaft (c). A stop prevents direct contact between the rollers and the ring. The feed flows in front of the rolls under the action of gravity. In front of each roller is a blade resembling a plowshare, which brings up the feed from below and directs it toward the roller. The product is carried upward by a stream of air to the classifier. These mills are used to grind soft to moderately hard materials, especially sticky, temperature-sensitive, or low-density substances. Examples are lignite, sulfur, barite, ceramic materials, and clay.

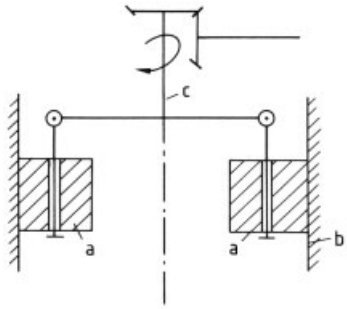

Figure 42. Schematic of a pendular mill
a) Pendular roller; b) Stationary grinding ring; c) Shaft

Ring-Roller Mills with Powered Grinding Rings. The most widely used ring-roller mill designs have powered disk- or pan-shaped grinding rings (Fig. 43). The feed falls onto the rotat-

ing grinding ring (a) and is driven radially under the rollers (b) by centrifugal force. After being rolled over repeatedly, the material is taken from the edge of the disk and carried upward to the classifier by a stream of air (c). These designs have found two main applications – the grinding of coal fed into furnaces and the grinding of limestone in the cement industry. In the grinding of coal, large amounts of exhaust gas provide an inert atmosphere and can be used for drying and for transporting the fines to the burner. Maximum throughput is 80 t/h. Mills with disk diameters of up to 6 m and rollers with diameters of ca. 3 m are used for grinding limestone. Because of their lower energy requirements, these mills have largely displaced tube mills. Maximum througput is 400 t/h. Additional uses are the grinding of cement, phosphate, magnesite, burnt lime, and talc.

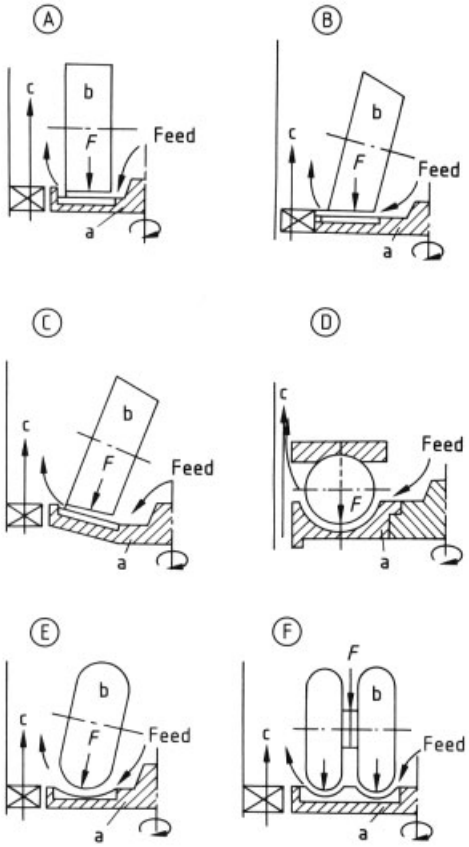

Figure 43. Schematic of ring-roller mills with powered grinding rings
a) Grinding ring; b) Roller; c) Airstream
F = grinding force

Edge Runner Mill. The edge runner mill (Fig. 44) is the oldest of the ring-roller mills. It was developed for gold ore preparation in Chile and is therefore also called a *Chilean mill.* It is now used in grinding and mixing plants in the ceramic and refractory ceramic industries as well as in the production of olive oil. Two heavy cylindrical rollers called mullers (a) roll over a pan (b) and may be positioned at different distances from the axis of rotation. Either the muller or the pan can be powered. The design with a powered pan has the advantage that the mullers are individually suspended and therefore the beds can be loaded uniformly with different particle bed depths. In addition, material transport is accomplished by centrifugal force. The weight of the muller and, in the case of the driven mullers, the momentum of rotation supply the grinding force.

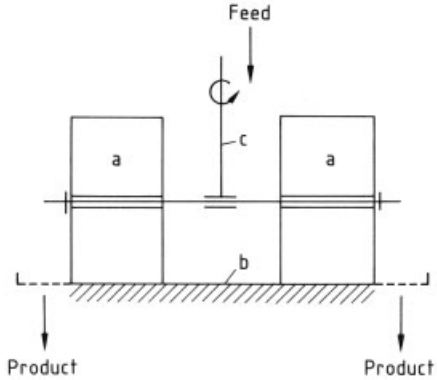

Figure 44. Schematic of an edge runner mill
a) Muller; b) Pan; c) Shaft

The peripheral speed of the pan is a function of the radius. In contrast, all points on the surface of the muller have the same peripheral speed. As a result the muller and the pan run at the same speed only at one particular value of the radius. At a larger radius the pan is faster, at a smaller radius the surface of the muller is faster. Thus, the applied pressure is reinforced by friction that promotes the homogenization and mixing of the materials.

Ring-roller mills with conical rollers (Fig. 45) have flat or pan-shaped grinding rings (a). A velocity difference exists between the ring and roller. The grinding force is controlled by a hydraulic device (b) which also permits lifting

of the rollers during startup. A stop (c) prevents contact of the rollers with the grinding ring. The rollers can often be swung out (d) in order to change them.

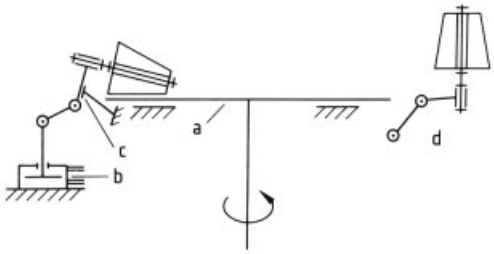

Figure 45. Schematic of a ring-roller mill with conical rollers
a) Grinding ring; b) Hydraulic device for pressure control; c) Stop; d) Roller swung out for replacement

Ring-roller Mills with Cambered Rollers and Pan-Shaped Grinding Rings. In the mill shown in Figure 46, the rollers (a) are cambered and move in the correspondingly pan-shaped grinding rings (b). The contact pressure is applied by a system of springs or hydropneumatically (c). The mill housing does not have openings for shafts or axles. Consequently, no complicated seals are required because a slight excess pressure exists within the housing; pressure-resistant operation (coal grinding) is possible.

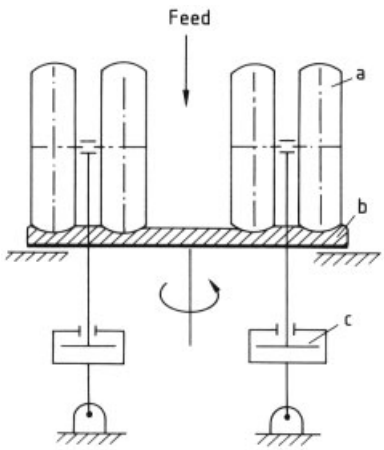

Figure 46. Schematic of a ring-roller mill with cambered rollers
a) Roller; b) Grinding ring; c) Hydraulic device for applying pressure

Ring Ball Mill. The ring ball mill is basically a large axial ball bearing (Fig. 47). The lower race (b) is powered. The upper race (c) is stationary and is held against the housing by an array of springs or by a hydraulic system. The balls (a) lay on the lower race next to each other without spacers and can touch each other. Heavy-walled, wear-resistant, cast, hollow spheres are used with diameters up to 1.25 m. Because there are no bearings inside the mill, the gas temperature can be somewhat higher than in other roller mills (up to 800 °C for the entering gas). Openings for shafts or axles in the housing are also unnecessary.

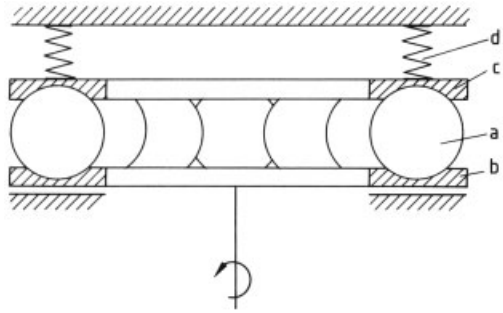

Figure 47. Schematic of a ring ball mill
a) Ball; b) Lower race; c) Upper race; d) Spring system

3.4. Impact and Jet Mills

In impact and jet mills, stress is applied at a single surface; the particles fly against an impact plate or a rapidly moving grinding tool impacts slowly moving particles. Stress can also result from interparticle collisions (see Section 2.2). In jet mills, gas jets are used to accelerate the particles.

Stress application differs from that in previously described mills where stress is primarily applied by means of pressure. The main differences between impacting and compression are

1) stress is applied to individual particles and not to a particle bed,
2) the energy supplied to the individual particle is proportional to the square of its velocity,
3) values of the fracturing energy are limited to ca. 5 J/g (corresponding to 100 m/s), and
4) the velocity at which stress is applied is higher, however this is mainly relevant to ma-

terials which exhibit viscous behavior (Section 2.1).

Very fine particles can be generated under dry conditions with impact and jet mills. However, the final particle fineness (in the range 1 – 10 µm) is limited by two factors. The first factor is that the finest particles flow around the moving impact tools of the mill in the air stream and can no longer be struck by them. The second factor is the slowing down of the particles; for example, a 10-µm particle in stationary air slows down from 100 to 0 m/s in a distance of 10 mm. This means that any particle can lose its kinetic energy over a short distance, so that little size reduction occurs when a collision takes place [3]. Jet mills are somewhat better suited to deal with this problem, because they use interparticle collisions that take place at relatively high particle concentrations.

3.4.1. Impact Mills

The range of products processed in impact mills is extraordinarily large. Impact mills are found in all branches of industry which deal with bulk materials or powders. A few examples of the many substrates are cocoa, herb tea, plant seeds, spices, fertilizers, plant protectants, resins, waxes, gypsum, limestone, wood, and lignite.

Several limits are set for the economic industrial use of impact mills. The product must not be abrasive (usually < 4 on the Mohs hardness scale). Hard minor components must be considered, for example the quartz content must not exceed 0.15 %. Brittle materials (i.e., substances which exhibit elastic behavior up to the breaking point, see Section 2.1) are especially suitable. The material should not be too heat sensitive. However, some heat-sensitive materials can be ground in an impact mill because of the large volume of air used and the relatively low number of collisions. Moisture, or oil or fat content can cause coating or plugging of the mill.

Impact mill designs can be divided into beater mills, hammer mills, and mills with or without peripheral grinding faces. In Figure 48 the mills are classified according to the final product fineness. The fineness is given by x_{97}, the size at which 97 % of the material passes.

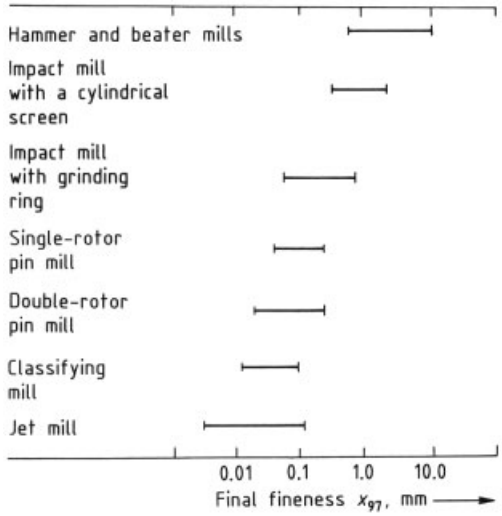

Figure 48. Classification of impact and jet mills according to product fineness

In impact mills, the ventilator action of the rotor ensures that a stream of air flows through the mill (provided, of course, that entry of air into the feed inlet is not obstructed or blocked). Volumetric flow rates up to 360 000 m^3/h can be handled by beater mills, which are specially designed for high gas throughputs. However, in conventional impact mills with a maximum diameter of ca. 1.5 m, the flow rates range up to 20 000 m^3/h. Consequently, a large portion of the total energy supplied (ca. 30 – 80 %) must be used for air or gas transport.

Beater Mills. Beater mills (Fig. 49 A) serve two purposes, they operate as blowers and simultaneously as impact mills. They are used for grinding lignite fed into steam boilers. Hot flue gas is used to dry the lignite and provide an inert atmosphere in the mill. Lignite and gas enter axially, pass radially through the blower cascade, and collect in the housing. From there, they flow to a classifier which returns the coarse material to the mill. The mill gas carries the fines to the burner. The largest mills have rotor diameters of ca. 4 m and process ca. 100 t/h.

Blower – Beater Mills. Blower – beater mills (Figs. 49 B and C) are also used for grinding furnace feeds. Separate devices are employed for comminution and gas transport. As in hammer mills, comminution is performed by

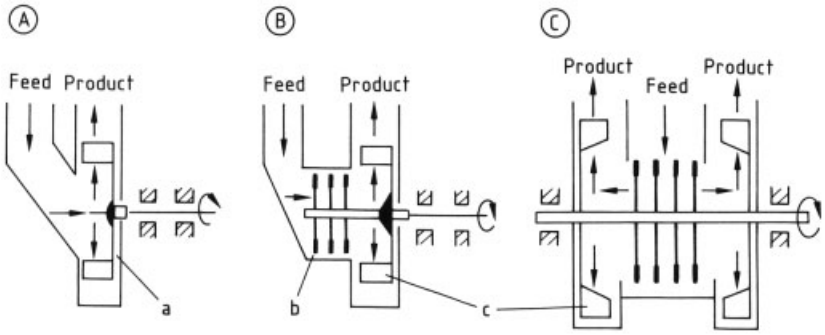

Figure 49. Schematic of beater and blower – beater mills
A) Beater mill with horizontal wheel shaft; B) Blower – beater mill; C) Double flow blower – beater mill
a) Beater wheel; b) Swinging hammer; c) Fan

a rotating array of swinging beaters attached to a shaft.

Hammer Mills. Hammer mills are basically similar to hammer crushers (Fig. 27) but their construction is less sturdy. The boundary between hammer mills and hammer crushers is not clearly defined but can be taken to be a diameter of ca. 1.5 m. The hammers are usually simple iron plates. The screen openings surrounding the rotor range between 2 and 10 mm. Applications include rough milling of grain, chipping of wood, and processing of waste such as wood turnings, fruit crates, paper bags and corrugated cardboard.

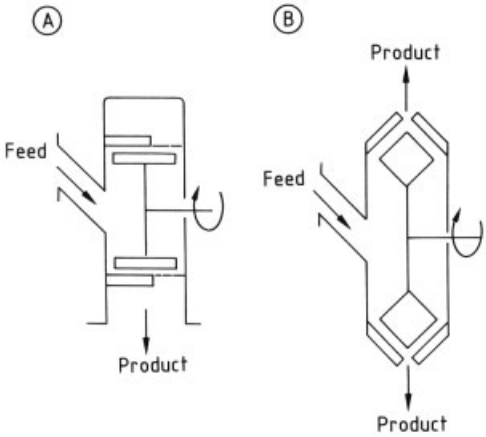

Figure 50. Schematic of impact mills with horizontal shafts
A) Mill with grinding bars and screen; B) Mill with a V-shaped grinding chamber and peripheral product outlets

Universal Mills. The characteristic feature of the mills depicted in Figure 50 is a horizontal axis that is surrounded by a grinding face. The feed is fed in laterally into the center of the rotor and then moves radially outward. The material is crushed by the comminution tools of the rotor, accelerated, and crushed once again at the grinding face. The grinding face holds back the particles until they are sufficiently fine.

The grinding face consists of a perforated or slotted screen, which is usually interrupted at intervals by impact bars. The bars direct the material sliding along the screen back toward the circular impact zone. Findings on the operation of screen-impact mills are given in [58].

The side wall which holds the feeder is constructed as a door. This permits easy access to the grinding chamber for cleaning purposes or for changing the grinding tools. These mills are referred to as universal mills because the grinding devices and grinding surfaces can easily be changed to adapt the mill for different feeds. These mills are extraordinarily versatile in their application. Their names correspond to their construction, e.g., cross-flow, turbo, blower, beater, pendular hammer, or disintegrator mill.

Peripheral speeds range from 30 to 120 m/s. With diameters up to 1.5 m they can achieve throughputs up to 10 t/h depending on the product fineness required.

Impact Mills with Vertical Shafts. Impact mills with vertical shafts are depicted in Figure 51. Figure 51 A shows a design with a conical rotor and stator. The gap width can be adjusted by axial movement of the stator. The impact bars are arrayed in multiple tiers along

the circumference of the rotor. Air and feed flow down through the mill cocurrently. In a version with a cylindrical rotor, flow is upward.

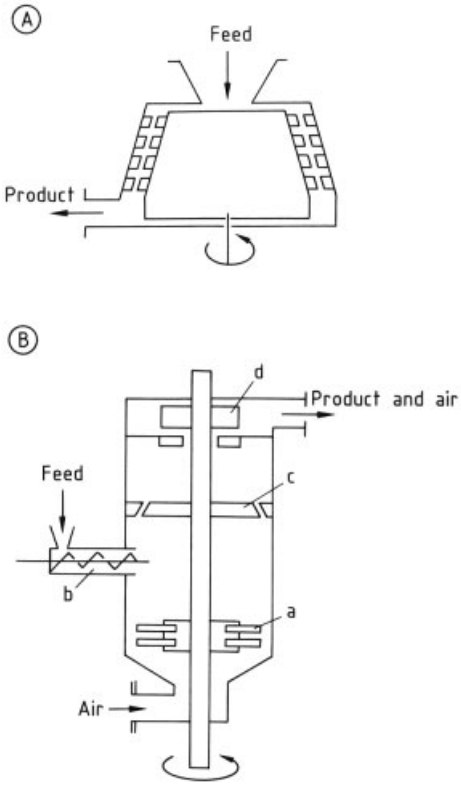

Figure 51. Schematic of impact mills with vertical shafts
A) Impact-bar mill with a conical rotor; B) Impact mill with rejector classifier
a) Impact hammer; b) Screw feeder; c) Classifying whizzer blades; d) Fan

Disk-shaped rotors with peripheral impact tools (hammers) (Fig. 51 B) can, as in roller mills, have an airstream which carries the feed to a classifier. The degree of fineness is determined by the adjustable classifier.

Pin or Pin-Beater Mills. The impact devices used in pin mills (Fig. 52) are a large number of round pins or elongated squares (beaters). The devices are arranged in concentric rings of varying radius. A stator of corresponding design meshes with the rotating rows of pins.

The pin mill differs from the universal mill in that the rotor is not surrounded by a grinding face. Thus, the degree of fineness is determined primarily by the peripheral speed of the rotor.

The number of collisions is relatively low because there is no grinding face (ca. 5 – 8 collisions for three rows of pins). Consequently, materials which are temperature-sensitive or tend to cake are often processed in these mills.

Designs are available in which two arrays of pins rotate in opposite directions, so that the impact velocity is theoretically doubled (maximum 230 m/s).

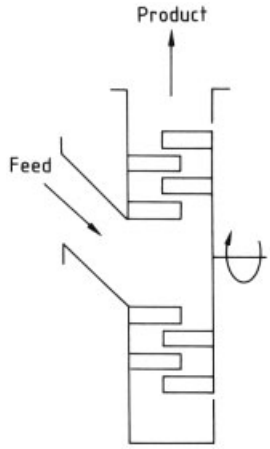

Figure 52. Schematic of a pin mill

3.4.2. Jet Mills

In jet mills the feed particles are accelerated by means of propellant jets (air, gas, or steam) with high velocities (sonic or supersonic speeds, i.e., 500 – 1200 m/s). Comminution occurs by interparticle collision or by impact with an impact surface (target). These mills are always operated in closed circuit with an air classifier. The particles do not attain the velocity of the gas because they cannot reach the center of the jet. In addition, a single particle collision is insufficient for obtaining the desired final particle size for particles smaller than ca. 20 µm. The specific energy consumption ranges from 300 to 3000 kW · h/t and is mainly determined by the generation of propellant [e.g., 0.6 MPa (6 bar), 3000 – 12 000 m^3/h depending on design] and inefficient particle acceleration.

The actual milling or comminution volume is small compared to the sections which provide the propellant gas and separate the fine particles.

Spiral Jet Mill. In the spiral jet mill (Fig. 53 A), the feed is sucked into the mill by means of an injector (a). The propellant is fed via a circular distributor (b) to individual nozzles (c) distributed on the circumference of the mill. The angles of the nozzles and the gas velocity are adjusted such that the feed rotates around the chamber in the form of a ring without touching the wall. The particles are slowed and accelerated at the margins of the jets and interparticle collisions occur. The gas motion allows size classification as in spiral air classifying. The equilibrium between the centrifugal force acting on a particle and the radial component of the drag of the spiral flow determines the cut point (→ Air Classifying, Chap. 2.). The fine particles leave the mill through the central opening.

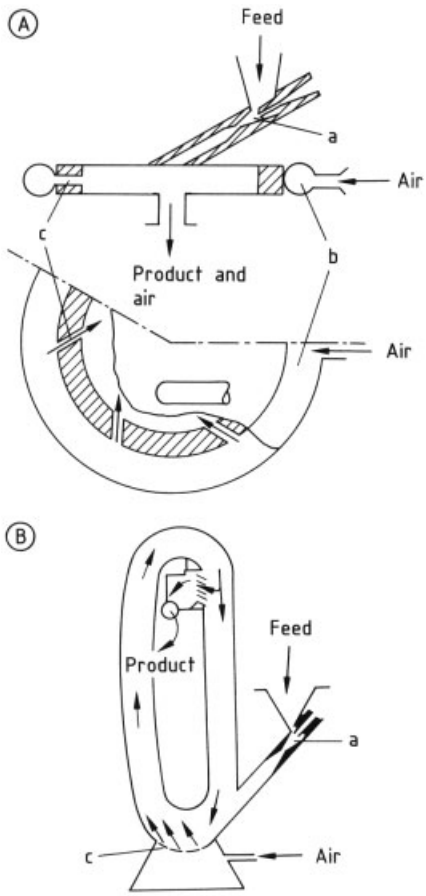

Figure 53. Schematics of the spiral jet mill (A) and the oval-tube jet mill (B)
a) Injector; b) Air-supply pipe; c) Nozzle

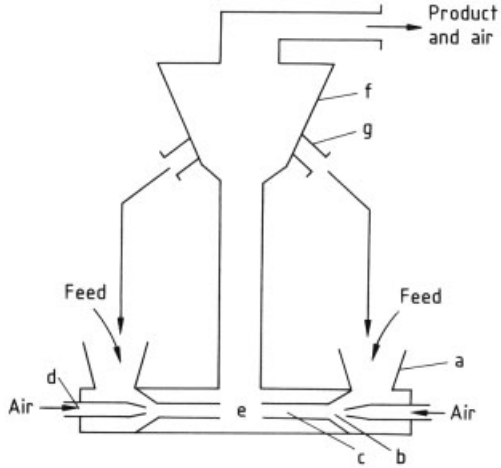

Figure 54. Schematic of an opposed jet mill
a) Feed funnel; b) Injector; c) Acceleration pipe; d) Propellant inlet; e) Grinding chamber; f) Classifier; g) Outlet for recycling coarse fraction

Spiral jet mills can handle explosive powders as well as temperature- and contamination-sensitive materials because they have no moving parts and there is no contact between the feed and the wall. The mills have a maximum diameter of 800 mm, and therefore do not require much space; they also offer the advantage of easy cleaning and sterilization. Throughputs range from 1 – 500 kg/h. Their main application is the comminution of chemicals and pharmaceuticals.

Oval-Tube Jet Mill. The oval-tube jet mill (Fig. 53 B) is based on the same comminution principle as the spiral jet mill. The particles are classified by a sharp change of direction of the flow stream. Particles that can follow this bend can leave the mill. This mill is used in the preparation of minerals (e.g., talc and mica), because of its sturdy construction.

Opposed Jet Mill. In the opposed jet mill (Fig. 54), the particles are accelerated through two pipes (c) and directed against each other in the central grinding chamber (e). The feed streams may be the coarse material from the classifier and the feed, as shown. Alternatively, the feed may first pass through a classifier; the coarse material is then distributed between the two pipes. Acceleration and comminution occur in a restricted region which can easily be protected against wear. Consequently, abrasive

material such as quartz or zircon sand can be ground in this type of mill.

The opposed jet mill differs from the spiral jet and oval-tube mills in that it has somewhat lower energy requirement and achieves a higher degree of fineness. In the case of an independently powered air classifier, better control of coarse particles is possible.

Fluidized-Bed Jet Mill. The fluidized-bed jet mill (Fig. 55) does not require an armored acceleration tube or a grinding chamber, as do the previously described opposed jet mills. The grinding zone (a) consists of a fluidized bed. Jets of air are blown into the bed of particles toward the center of the mill through 3 – 6 nozzles (e) arranged on the perimeter of the cylindrical mill vessel (b). The jets fluidize the particles. The particles collide with each other at the margins of the jets and in the center. The air carries the ground particles to a classifier (c) in the upper part of the cylinder. The classifier determines the particle size of the product. The mill cannot be used with nonfluidizable particles.

The particles come into contact almost exclusively with other particles, thus this mill is especially suited for abrasive substances (e.g., abrasives, ceramic raw materials, as well as metal carbides, oxides, and nitrides) and materials that must be free of contaminants (e.g., powdered phosphors, dental materials, silicic acid, and silicon).

Target Mills. In particle jets that are directed against each other, the probability that the particles will not collide is relatively high. Consequently, mills have also been developed in which the particles accelerated in the gas jet are directed against a solid surface or target. Wear can be controlled by using suitable construction materials, because the impact area is not very large and the target can be made sufficiently thick.

3.5. Cutting Mills

Paper and wood waste, tough plastic films, rubber, leather, and feeds cannot be comminuted with any of the mills described so far. Equipment with a cutting action is required for these applications as well as for the disintegration of bundles of plant fibers, the production of wood chips, and cellulose pulp. Shredders, chippers, knife-choppers, and, especially for fibers, pulp machines are used.

Shredders. The construction of shredder mills resembles that of impact crushers (see page 20). The rotor and stator are equipped with blades that pass by each other, and are separated by a narrow clearance. The peripheral speed ranges from 5 to 20 m/s. The basket screen determines the final product fineness.

Other designs use disks mounted on two parallel shafts and mesh with each other to give a shearing action (Fig. 56 A). Each disk varies in radius along its circumference. The part with the largest radius forms a cam, which helps to draw in the feed. The cams are set at different positions along the shaft to ensure uniform energy consumption during rotation. In case of plugging or overload, the direction of motion reverses automatically.

In another design, which can be used in the outlets of hoppers, two propeller-like blades work against a stationary knife beam (Fig. 56 B).

Pulp Mills. Mechanical separation of fibers is of great importance in the paper and cellulose industries. If possible, only shear stress should be exerted in the direction of the fiber. No tensile stress should be applied since this could lead to tearing of the fiber. The paper pulp is prepared in a pulp machine, in which stress is applied between drums equipped with blades and blades attached to the bottom of a trough. Many different designs of this type of machine are described in [59] (see also → Paper and Pulp).

3.6. Wet Grinding

Wet grinding is defined as the comminution of solids suspended in a fluid. The principal reasons for using this technique are

1) the feed is already in the form of a suspension,
2) the product is desired as a suspension,
3) the feed material has to be finely ground and has a tendency to agglomerate, or
4) the material is explosive or toxic.

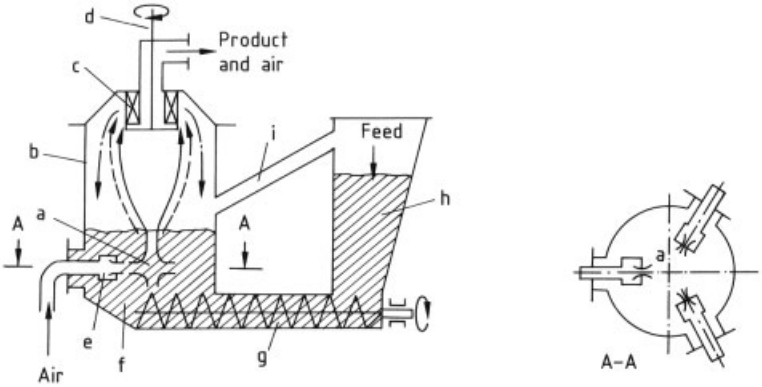

Figure 55. Schematic of a fluidized-bed jet mill
a) Grinding zone; b) Grinding vessel; c) Classifier; d) Drive for classifier; e) Nozzle; f) Fluidized bed; g) Screw feeder; h) Hopper; i) Pressure equalization tube

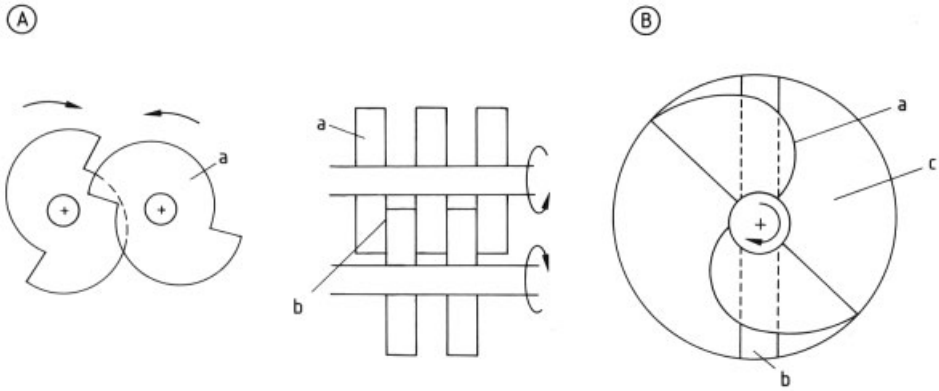

Figure 56. Principle of shredders
A) Disk knife system: a) Cutting disk; b) Cutting edge
B) Top view of a rotating shredder: a) Cutting edge; b) Cutting beam; c) Funnel outlet

Wet grinding combines the actual comminution of the solid particles with the breaking up of agglomerates and the dispersion of the individual particles. In some machines comminution is the primary objective, whereas in others deagglomeration is the primary function.

All the previously described comminution devices with the exception of crushers, impact mills, and jet mills are used for wet grinding. They operate primarily with low-to-moderate viscosity suspensions. However, roller and high-compression roller mills can only be employed at moisture levels of 20 % or less; although some roller mills are made for wet grinding [60], [61].

The Szego mill, the roller frame, the agitated ball mill, and the disk mill have been specially designed for wet grinding.

Szego Mill. The Szego mill is basically a centrifugal roller mill (Fig. 57). The roller shafts (c) are held between two disks (a), which rotate inside a cylindrical housing (b). Because the roller shafts can be moved radially, centrifugal force presses the rollers against the stator and they roll over it. A helical channel is cut into the surface of the rollers and, together with gravity, provides for material transport.

A production-scale mill was investigated with coal – oil and coal – water suspensions (60 – 70 wt % coal) [62]. The mill can also be used for the dry grinding of materials such as mica, talc, wheat flour, cocoa, and soybeans. A similar design is reported in [7] and has been used in the Soviet Union for the wet grinding of ores.

Another horizontal design employs rods that are driven by a thick central shaft. This system is comparable to a needle bearing.

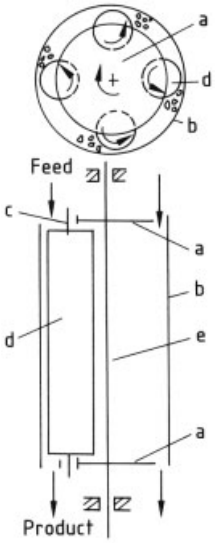

Figure 57. Schematic of the Szego mill
a) Rotating disk; b) Housing; c) Shaft of grinding roller; d) Helical grooved roller; e) Central shaft

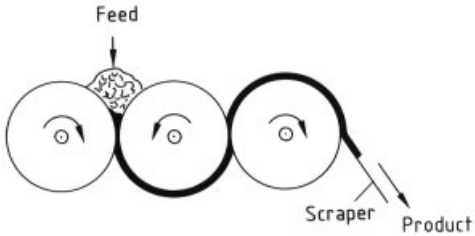

Figure 58. Schematic of a triple-roller frame

Roller Frames. Roller frames with two or more smoothly polished rollers (Fig. 58) are finding application in the processing of moderately viscous suspensions, for example in the printing ink industry or the manufacture of chocolate. The rollers turn at different speeds, each being faster than the one following it. In this way the suspension coating is transferred between the rollers without any difficulty. In some machines, one of the rollers is replaced by a plate, usually called the bar. The gap width is adjusted by changing the grinding force, which is usually applied hydraulically. The smallest gap width is 20 µm. Coarse particles are stressed between the rollers, according to the gap width. Fine particles can also be stressed if particle bridges form at sufficiently high concentrations, otherwise only the shear stress transmitted by the medium is effective. Multiple passes through the gap are usually required to achieve the desired degree of fineness.

Agitated (Stirred) Ball Mills. The agitated ball mill (Fig. 59) was developed from the "sandmill" and has recently gained importance in wet grinding. It consists of a horizontal or vertical cylindrical housing (a), which is filled to ca. 80 % with glass beads, steel balls, or a ceramic grinding medium, with sizes ranging from a few tenths of a millimeter to several millimeters. An agitator shaft runs along the axis of the mill and is equipped with perforated disks, rings, or pins. The agitator (b) rotates with a peripheral speed of 4 – 20 m/s. The suspension is forced through the mill by a pump. A screen (c) or slot holds back the grinding media at the product discharge. Product fineness of less than 2 µm can be obtained. Examples of applications are the paint and lacquer industry, coatings for the paper industry, ceramic materials, chocolate production, and the disruption of cells in biotechnology. Mill volumes may be up to 1 m^3.

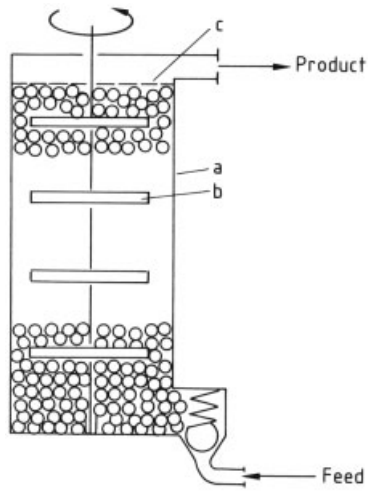

Figure 59. Schematic of an agitated ball mill
a) Cylindrical mill housing; b) Agitator; c) Screen

Other designs employ a narrow conical or cylindrical grinding gap with smooth walls (gap width is ca. four times the diameter of the grinding balls). The grinding balls move through the gap together with the suspension and the recycled within the mill by appropriate devices. The

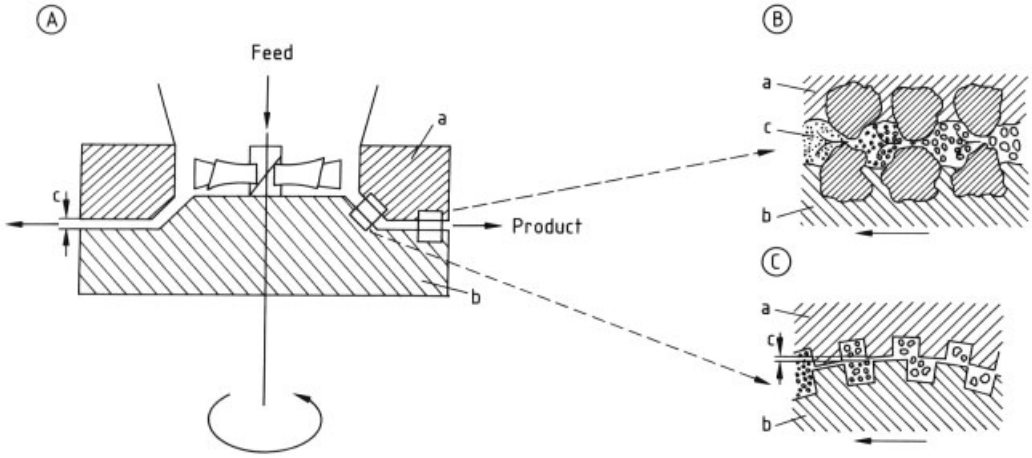

Figure 60. Schematic of a disk mill (A) showing close-ups of the gaps of a corundum disk mill (B) and a toothed colloid mill (C)
a) Stator; b) Rotor; c) Gap width

narrow gap permits high energy densities to be achieved.

Under certain operating conditions (especially with aqueous suspensions), the specific work measured in a laboratory-scale mill can be used for scale-up purposes. It determines the attainable fineness [63].

Agitator mills are sometimes used for dry grinding. The separation of the grinding medium then takes place outside the mill. The grinding medium is fed in with the feed. A considerable amount of additional equipment is required: conveying devices and intermediate bunkers for the grinding media as well as classifiers and dust separators for the fine material.

Disk Attrition Mills. Disk attrition mills (Fig. 60) are rotor – stator configurations with maximum diameters of ca. 500 mm. Large mills are equipped with a horizontal drive. The profiled or perforated disks can be made of steel, sintered material, or corundum (Fig. 60 B). They rotate with peripheral velocities up to 35 m/s. The gap width can be very finely adjusted (in the micrometer range) and is in the order of 30 – 500 µm.

Examples of applications are the comminution of pharmaceuticals and chemicals to finenesses in the range of 10 – 20 µm. The mills are also used in the comminution of leather, hides, bone, cardboard, and wood, as well as for the pulping of foods such as fruit, meat, bacon, and almonds. Throughputs are usually a few hundred kilograms per hour but depend on the application; several tonnes per hour may be reached. A special rotor – stator configuration with a conical clearance, the toothed colloid mill (Fig. 60 C), is used primarily for the homogenization, dispersion, and mixing of creams, salves, toothpaste, baby food, and mayonnaise.

Kneaders are required to disperse pigments and fillers in high-viscosity suspensions. These devices are discussed elswehere (→ Mixing of Highly Viscous Media).

4. References

1. H. Rumpf, *Staub-Reinhalt. Luft* **27** (1967) 3 – 13.
2. R. Polke, *Part. Charact.* **4** (1987) 54 – 62.
3. H. Rumpf, K. Schönert in *Ullmann,* 4th ed., **2,** 1 – 23.
4. *6th European Symposium "Comminution" Preprints,* **Part II,** Nürnberger Messe- und Ausstellungsgesellschaft, Nürnberg 1986.
5. *European Symposium "Particle Technology," Preprints,* **Vol. A,** Dechema, Frankfurt 1980.
6. *4th Europäisches Symposium "Zerkleinern," Dechema Monographie,* **vol. 79, Part A,** Verlag Chemie, Weinheim 1976.
7. K. Höffl: *Zerkleinerungs- und Klassiermaschinen,* Springer Verlag, Berlin-Heidelberg-New York-Tokyo 1986.

8. K. Schönert in *Winnacker-Küchler,* **1,** 80 – 93.
9. K. Dialer, V. Onken, K. Leschonski: *Grundzüge der Verfahrenstechnik und Reaktionstechnik,* Hanser Verlag, München-Wien 1986.
10. M. Zoog: *Einführung in die Mechanische Verfahrenstechnik,* Verlag Teubner, Stuttgart 1987.
11. H. Schubert: *Aufbereitung fester mineralischer Rohstoffe,* **vol. 1,** VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1975.
12. C. L. Prasher: *Crushing and Grinding Process Handbook,* J. Wiley and Sons, New York 1987.
13. B. Beke: *The Process of Fine Grinding,* M. Nijhoff, W. Junk Publishers, Den Haag – Boston – London 1981.
14. J. Beddow: *The Production of Metal Powders by Atomisation,* Heyden and Sons, London 1978.
15. A. J. Lynch: "Mineral Crushing and Grinding Circuits" in *Developments in Mineral Processing,* Elsevier, Amsterdam 1977.
16. V. C. Marshall: *Comminution, Report of a Working Party* Chap. 2, Inst. of Chemical Engineers, London 1975.
17. G. C. Lowrison: *Crushing and Grinding,* Butterworth, London 1974.
18. R. H. Snow, B. H. Kaye, C. E. Capes, G. C. Sresty in R. H. Perry, D. Green (eds.): *Perry's Chemical Engineers Handbook,* McGraw Hill, New York 1984, pp. 8-7 – 8-60.
19. H. Horii, S. Nemat-Nasser, *JGR J. Geo-phys. Res.* **90** (1985) 3105 – 3125.
20. D. S. Dugdale, *J. Mech. Phys. Solids* **8** (1960) 100.
21. R. Weichert, K. Schönert, *J. Mech. Phys. Solids* **26** (1978) 151 – 161.
22. V. Boldyrev, K. Meyer: *Festkörperchemie, Beiträge aus Forschung und Praxis,* VEB Verlag für Grundstoffindustrie, Leizig 1973.
23. W. Schwenk: *3rd European Symposium "Zerkleinern," Cannes (1971), Dechema-Monogr.,* **vol. 69,** Verlag Chemie, Weinheim 1972, pp. 121 – 138.
24. H. Rumpf, *Chem. Ing. Tech.* **37** (1965) 187 – 202.
25. W. Hess: *Einfluß der Schubbeanspruchung und des Verformungsverhaltens bei der Druckzerkleinerung von Kugeln,* Dissertation Universität Karlsruhe 1980.
26. R. Weichert, *Int. J. Miner. Process.* **22** (1988) 1 – 8.
27. H. Schubert, *Aufbereit. Tech.* **28** (1987) 237 – 246.
28. N. Hoffmann et al., *Chem. Ing. Tech.* **48** (1976) 329.
29. G. Göll, J. Hanisch, *Aufbereit. Tech.* **28** (1987) 582 – 590.
30. K. Schönert in [4] .
31. C. J. Stairmand in [6] .
32. H. Rumpf, *Aufbereit. Tech.* **16** (1975) 59 – 71.
33. F. C. Bond, *Aufbereit. Tech.* **5** (1964) 211 – 218.
34. P. S. B. Stewart, C. J. Restarick, *Proc. Australas. Inst. Min. Metall.* **239** (1971) 81 – 92.
35. L. G. Austin, *Ind. Eng. Chem. Process Des. Dev.* **12** (1973) 121 – 129.
36. V. K. Gupta, P. C. Kapur, *Powder Technol.* **10** (1974) 217 – 223.
37. R. K. Jaspan et al. in *4th Europäisches Symposium "Zerkleinern," Dechema Monographie,* **vol. 79, Part A,** Verlag Chemie, Weinheim 1976. S. G. Malghan, D. W. Fuerstenau in *4th Europäisches Symposium "Zerkleinern," Dechema Monographie,* **vol. 79, Part A,** Verlag Chemie, Weinheim 1976.
38. K. J. Reid, *Chem. Eng. Sci.* **20** (1965) 953 – 963.
39. L. G. Austin, P. T. Luckie, *Powder Technol.* **5** (1971/1972) 267 – 271.
40. A. L. Mular, *Bulletin Canad. Mining and Metall.* (1970) 821 – 826.
41. R. R. Klimpel, L. G. Austin, *Ind. Eng. Chem. Fundam.* **9** (1970) 230 – 237.
42. J. A. Herbst, T. S. Mika, *Proc. 11th Intern. Symp. Comp. Applications in the Minerals Industry,* Tucson Arizona (1973) E 78 – E 123. K. Shoji, L. G. Austin, *Powder Technol.* **10** (1974) 29 – 35.
43. K. Schönert, *Chem. Ing. Tech.* **6** (1971) 361 – 367. R. P. Gardner, K. Verghese, *Powder Technol.* **11** (1975) 87 – 88.
44. J. Decasper, *ZKG Zement Kalk Gips* **52** (1980) 219 – 222.
45. D. Hochdahl, *ZKG Zement Kalk Gips* **54** (1982) 1 – 10.
46. F. Feige et al., *ZKG Zement Kalk Gips* **55** (1983) 628 – 632.
47. K. Schönert, *ZKG Zement Kalk Gips* **51** (1979) 1 – 9.
48. H. Kellerwessel, *Aufbereit. Tech.* **27** (1986) 555 – 559.
49. M. V. Seebach, N. Parzelt, *ZKG Zement Kalk Gips* **40** (1987) 337 – 344.
50. DIN 50 320. H. Metz (ed.): *Abrasion und Erosion,* Hanser Verlag, München–Wien 1986.

51. G. Roth: *Transport in Drehrohren und Kugelmühlen,* Dissertation Techn. Universität Karlsruhe 1982.
52. J. Leluschko: *Untersuchungen zum Einfluß der Flüssigkeit bei der Naßmahlung in Kugelmühlen,* Dissertation Techn. Universität München 1985.
53. H. E. Rose, R. M. E. Sullivan: *Ball, Tube and Roll Mills,* Constable, London 1958.
54. K. E. Kurrer, E. Gock, *Chem. Ing. Tech.* **57** (1985) 64 – 65.
55. J. Raasch, *Chem. Ing. Tech.* **36** (1964) 125 – 130.
56. H. Meiler et al. in [4] .
57. L. T. Schneider et al., *ZKG Zement Kalk Gips* **38** (1985) 705 – 708.
58. D. Landwehr, M. H. Pahl, *Aufbereit. Tech.* **28** (1987) 57 – 65, 188 – 192.
59. K. Rosenfeld: *1st European Symposium "Zerkleinern,"* Verlag Chemie – VDI Verlag, Weinheim, Düsseldorf 1963.
60. T. Loesche, D. Anstey in [4] .
61. D. Schwechten, K. Schönert in [4] .
62. O. Trass et al., *Powder Technol.* **40** (1984) 269 – 282.
63. H. Weit, *Chem. Eng. Technol.* **10** (1987) 398 – 404; *Chem. Ing. Tech.* **6** (1988) 318 – 319.

Spraying and Atomizing of Liquids

See also Solids Technology, Introdution.

PETER WALZEL, Universität Essen, Essen, Federal Republic of Germany

1. Introduction 2
2. Fundamentals of Drop Formation . . 2
2.1. Dripping and Jet Disintegration 3
2.2. Sheet or Lamella Disintegration 7
2.3. Dispersion of Liquids by Gas 8
3. Atomizing Apparatus 9
3.1. Single-Substance Pressure Nozzles . . 9
3.2. Pneumatic Atomizers 12
3.3. Atomization with Propellants 13
3.4. Rotary Atomizers 13
3.5. Ultrasonic Atomization 14
3.6. Electrostatic Atomization 14
4. Drop Size Measurements and Drop Size Distributions 15
5. References 16

Symbols

A	m^2	cross section area of the nozzle
A_e	m^2	cross section area of emerging jet
a	m/s^2	centrifugal acceleration
C_1, C_2, C_3	-	constants
d	m	drop diameter
d_{32}	m	sauter mean diameter
$d_{v,0.5}$	m	volume median diameter
d_{th}	m	thread or jet diameter
d'	m	nozzle diameter of liquid orifice
d'_C	m	diameter of swirl chamber
$d'_h \approx 2S_p$	m	hydraulic diameter slit
d'_p	m	prefilming diameter of gas orifice
E	V/m	strength of electrical field
f	1/s	frequency
g	m/s^2	acceleration of gravity
j, k	–	exponents
$K = \delta_x$	m^2	sheet thickness parameter
l_{th}	m	length of threads
m	–	exponent
$\dot{m}$	kg/s	mass flow rate of liquid
$\dot{m}_g$	kg/s	mass flow rate of gas
n	–	exponent
Δp	$kg\,s^{-2}\,m^{-1}$	pressure drop of the nozzle
Q	C	electrical charge
r	m	radius of atomizer
S_p	m	width of gas slit
s	m	spacing between threads
U	V	voltage
t	s	time
v	m/s	gas velocity
v_{rel}	m/s	relative gas velocity
$\dot{V}$	m^3/s	volumetric flow rate
w	m/s	liquid velocity
x	m	distance from orifice in jet direction
δ	m	sheet thickness
ε	$C\,V^{-1}\,m^{-1}$	dielectrical constant of the fluid
η	$kg\,m^{-1}\,s^{-1}$	viscosity of the liquid
η_g	$kg\,m^{-1}\,s^{-1}$	viscosity of the gas
θ	–	spray angle
λ	m	wavelength
ϱ	kg/m^3	density of the liquid

(Continued)

ϱ_g	kg/m^3	density of the gas
σ	kg/s^2	surface tension
ω	1/s	angular frequency
$a*$	$= a\,\eta^4/\sigma^3\,\varrho$	Kapitza number
Bo	$= d'^2\,\varrho\,g/\sigma$	Bond number
d^*_{th}	$= d_{th}\sqrt{r\,\omega^2\,\varrho/\sigma}$	dimensionless thread diameter
$d*$	$= d\sqrt{r\,\omega^2\,\varrho/\sigma}$	dimensionless drop diameter
$\Delta p*$	$= \Delta p\,d'/\sigma$	Laplace number
$\dot{V}*$	$= \dot{V}\sqrt{\varrho/(r\,\sigma)}/2\pi\,r$	throughput number
$\dot{V}^*_{th}$	$= \dot{V}^*\left(a^3\,\varrho^5/\sigma^5\right)^{1/4}$	throughput number of threads
Re_p	$= d'\sqrt{\Delta p\,\varrho}/\eta$	pressure Reynolds number
$s*$	$= s\sqrt{\varrho\,a/\sigma}$	dimensionless thread spacing
We	$= w^2\,d'\,\varrho/\sigma$	Weber number
We_g	$= v^2\,d'_p\,\varrho_g/\sigma$	gas Weber number
We_r	$= \varrho\,r^3\,\omega^2/\sigma$	Weber number for rotary atomizers
We_δ	$= w^2\,\delta\,\varrho/\sigma$	sheet Weber number
Z	$= \eta/\sqrt{d'\,\varrho\,\sigma}$	Ohnesorge number of the nozzle
Z_d	$= \eta/\sqrt{d\,\varrho\,\sigma}$	Ohnesorge number of drops
Z_r	$= \eta/\sqrt{r\,\varrho\,\sigma}$	Ohnesorge number for rotary atomizers
α	$= A_e/A$	contraction coefficient
κ	$= \delta x/A$	sheet number
μ	$= \dot{V}/\left(A\sqrt{2\,\Delta p/\varrho}\right)$	discharge coefficient
μ_m	$= \dot{m}/\dot{m}_g$	mass flow ratio of liquid to gas
φ	$= w/\sqrt{2\,\Delta p/\varrho}$	velocity coefficient

1. Introduction

Atomization is the formation of drops from liquids in a gas phase or vacuum. The gas density ϱ_g and liquid density ϱ differ by one or more orders of magnitude. In most cases, the disintegration processes differ fundamentally from the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_06

dispersion processes in systems with small density differences.

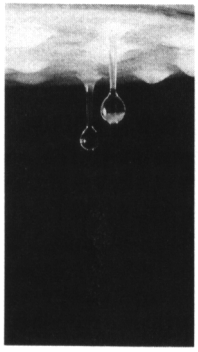

Figure 1. Dripping from the underside of a porous penetrated plate by the field of gravity (water)

In chemical engineering, liquid atomizers are employed in four major fields:

1) Atomization of liquid fuels
2) Production of granular products
3) Execution of mass-transfer operations
4) Coating of surfaces

Depending on the given problem, the desired characteristics of the spray differ. A narrow drop-size distribution is often desired in the production of granular products. A small average drop size is usually necessary in mass-transfer processes. Intensive, turbulent mixing between gas and liquid, as well as a small drop size, is required in combustion and mass-transfer operations. In coating, all of the drops should precipitate on a body to form a uniform film.

Frequently, the atomized products are not pure liquids in the usual sense but rather solutions, melts, or suspensions, thus demanding different atomization techniques than do pure liquids. Trouble is often caused by wear and plugging, which however can be reduced to a minimum by suitable apparatus. With highly abrasive suspensions, easily exchangeable hardmetal or ceramic inserts are used at the points of high flow velocity. To avoid plugging, nozzles that produce small drops in spite of relatively large orifices are used: hollow cone nozzles, air blast nozzles, and rotary cup atomizers are especially advantageous.

Although there is a large variety of atomization apparatus and nozzles on the market, specialized designs are only considered here to a limited extent. Instead, the emphasis is on the fundamentals of drop formation and their significance in practice, which apply to all nozzles. In general, an estimation of the drop size is possible at least for low viscosity liquids and liquids obeying Newtonic flow behavior.

2. Fundamentals of Drop Formation

Drops can be formed generally by dripping, jet disintegration, sheet disintegration, and dispersions of liquids by gases.

2.1. Dripping and Jet Disintegration

In these drop formation ranges, the influence of the ambient gas is subordinate except for very high discharge velocities.

Dripping. The simplest case of drop formation is *dripping* directly from a solid surface. This occurs for low liquid throughputs, and a static equilibrium between the gravitational (or mass forces) and the surface force develops. The drop constricts from the base as soon as the gravitational force exceeds the surface force. Thus, drops of very uniform size are produced. This is often important in dosing, e.g., in pharmaceutical applications. Dripping is also applied in pastil forming machines on a large technical scale up to mass flow rates of 1 t/h. The process of rapid drop formation is supported by flow oscillations at the exit holes. The oscillations are caused by a vibrating piston in a shaft or by a rotating perforated cylinder which periodically closes and opens the channels to the exit holes during rotation [119].

When dripping occurs from the underside of horizontal plates [2], [3], drops with diameter d arise upon complete wetting (Fig. 1)

$$2.9 < d\sqrt{\varrho g/\sigma} < 3.3 \tag{1}$$

where the surface tension σ and the density ϱ of the liquid in the gravitational field with acceleration g control the process. Drops can be obtained by dripping from horizontal edges, at limited flow rates such as at the end of a vertical draining surface [4], with diameter

$$d = 2.7\sqrt{\sigma/\varrho g} \tag{2}$$

In these dripping processes the influence of the viscosity η on drop size is small due to low deformation velocities.

In practice, dripping from capillary tubes or tubular nozzles is a frequent case (Fig. 2). Here the drop diameter d depends mainly on the capillary tube diameter d' and the liquid – nozzle wetting [5], [6]. Figure 3, Curve a, shows the relationship between d/d' and the Bond number,

$$\sqrt{B_0} = d' \sqrt{\varrho g/\sigma} \tag{3}$$

In the case of low dripping frequency the drop size can be estimated from [120]

$$d/d' = (6/B_0)^{1/3} \tag{4}$$

For $\sqrt{B_0} > 6$ the droplet size is identical to that of drops formed on the underside of a saturated horizontal surface. The drop formation and equilibrium contours are described in [121].

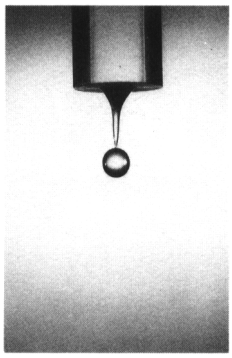

Figure 2. Dripping from tubular nozzle at very low flow velocities (water)

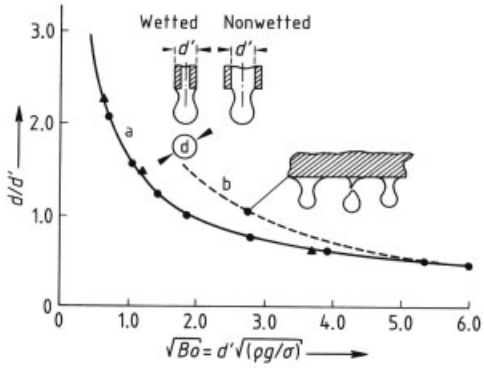

Figure 3. The ratio of drop diameter to capillary diameter (d/d') as a function of the Bond number B_0
a) Small capillary tubes [5], [6]; b) Wetted horizontal surfaces

Jet Disintegration. With higher liquid throughput, i.e., increased discharge velocity w, liquid jets or threads form and first disintegrate into drops at some distance from the outlet point; this is then called *laminar jet disintegration* or Rayleigh-type desintegration (Fig. 4). First considering small orifice diameters, $B_0 < 0.5$, jet formation occurs only above a certain discharge velocity with the pertinent minimum dynamic pressure $p_\varrho \sim \varrho w^2$. The ratio p_ϱ / p_σ is called the Weber number; when $p_\sigma \sim \sigma / d'$:

$$We = w^2 d' \varrho / \sigma \tag{5}$$

For well defined jet formation at the exit nozzles with horizontal axis [7]

$$We > 8$$

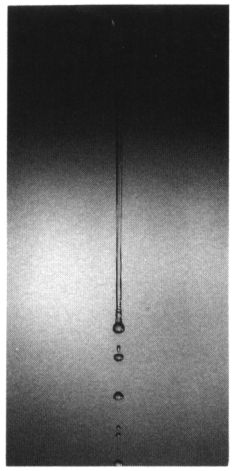

Figure 4. Laminar jet desintegration from tubular nozzle (water)

A bundle of many capillary tubes or bores in a metal sheet enable to increase the liquid throughput; in this way, a shower is obtained. If the bores are too close together, however, jet coalescence takes place [8].

Drops form from liquid threads because the liquid threads are unstable and on their own form centrally symmetrical waves; these grow quickly when the wavelength is $\lambda > \pi d_{\mathrm{th}}$ [9]. The jet diameter d_{th} can be set equal to the nozzle diameter d' with little error. The fastest growing waves are those with the optimal wavelength [10],

$$\lambda_{\mathrm{opt}} = \pi\, d_{\mathrm{th}} \sqrt{2 + 6\,Z} \tag{6}$$

with $Z = \eta / \sqrt{\varrho\, \sigma\, d_{\mathrm{th}}}$, where Z is the Ohnesorge number. For low viscosity liquids $\lambda_{\mathrm{opt}} = 4.4\, d_{\mathrm{th}}$. This produces drops with diameter

$$d = d_{\mathrm{th}} \sqrt[6]{44 + 6Z} \tag{7}$$

or in the case of low viscosity $d \approx 1.9\, d_{\mathrm{th}}$. The disintegration time of threads for complete drop formation is given by the following equation [10]:

$$t = 12 \sqrt{\frac{\varrho\, d_{\mathrm{th}}^3}{\sigma}} + 36\, \frac{\eta\, d_{\mathrm{th}}}{\sigma} \tag{8}$$

Within the laminar jet disintegration range, the jet length before complete disintegration increases with the discharge velocity w: $l_{\mathrm{th}} = wt$. In natural jet disintegration all the drops do not have the same size, but rather, although relatively narrow, a drop size distribution arises. This is due to occasionally formed satellite drops between the main drops, to drop coalescence, and to the natural disturbance frequency. Monodisperse drops, i.e., drops of a uniform size, can be produced by oscillatory stimulation, either by jet pulsation or nozzle vibration [11–15] (see Fig. 5). The required stimulation frequency is $f = w/\lambda_{\mathrm{opt}}$.

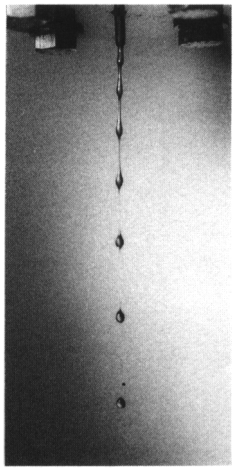

Figure 5. Laminar pulsed jet desintegration at a tubular nozzle forming droplets of uniform size (viscous oil)

When the exit diameters of vertical tubular ducts are larger, i. e., $B_0 > 0.5$, smaller exit velocities or *We*-numbers are sufficient in order to form a liquid thread which then underlies considerable extension and diameter reduction until breakup occurs.

On edges covered with liquid or on the underside of porous horizontal plates with sufficient liquid throughput, dripping points form on their own with regular spacing. The drop size is

$$d = 2.85\, d' \sqrt{\frac{1}{B_0}} = 2.85 \sqrt{\frac{\sigma}{\varrho g}}$$

(Fig. 3, Curve b). The same process occurs on rotating atomizers such as atomizer disk-cups or rotating porous rings [16], [17]. Centrifugal acceleration, $a = r\omega^2$, then replaces the gravitational acceleration g. The liquid threads formed at the dripping points also disintegrate according to the laws of jet disintegration. On edges covered with liquid or on rotating atomizer cups, the average spacing s between the liquid threads is $5.6 < s\sqrt{\varrho a/\sigma} < 7.0$ [4]. On the horizontal underside of saturated porous plates, in the field of gravity a regular triangular dripping pattern forms, with a jet spacing of $s^* = s\sqrt{\varrho g/\sigma} = 7.6$ [3] (see Fig. 6). The minimum throughput $\dot{V}_{\mathrm{th}}$ per jet in case of low viscous liquids was found to be $\dot{V}^* = 1$, with $\dot{V}^*_{\mathrm{th}} = \dot{V}_{\mathrm{th}}\, (a^3 \varrho^5/\sigma^5)^{0.25}$ and $\dot{V}_{\mathrm{th}}$ being the throughput per thread. Under these conditions the droplet size can be given as $1.1 < d* < 1.3$, with $d^* = 1.1\, d\sqrt{\varrho g/\sigma}$. This limit also holds for tubular nozzles with larger Bond numbers, $B_0 > 4.0$.

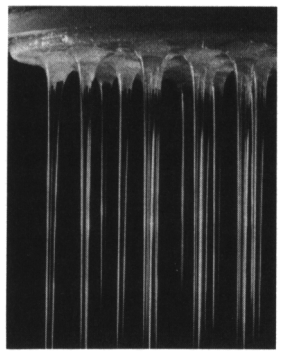

Figure 6. Formation of regular spaced laminar jets on the underside of a porous penetrated plate by the field of gravity (water)

As the liquid throughput is increased, the thread thickness at the detachment points increases with a small increase in drop size. This is due to the greater jet length, which permits pronounced jet contraction. If the throughput is increased further, the liquid threads progressively

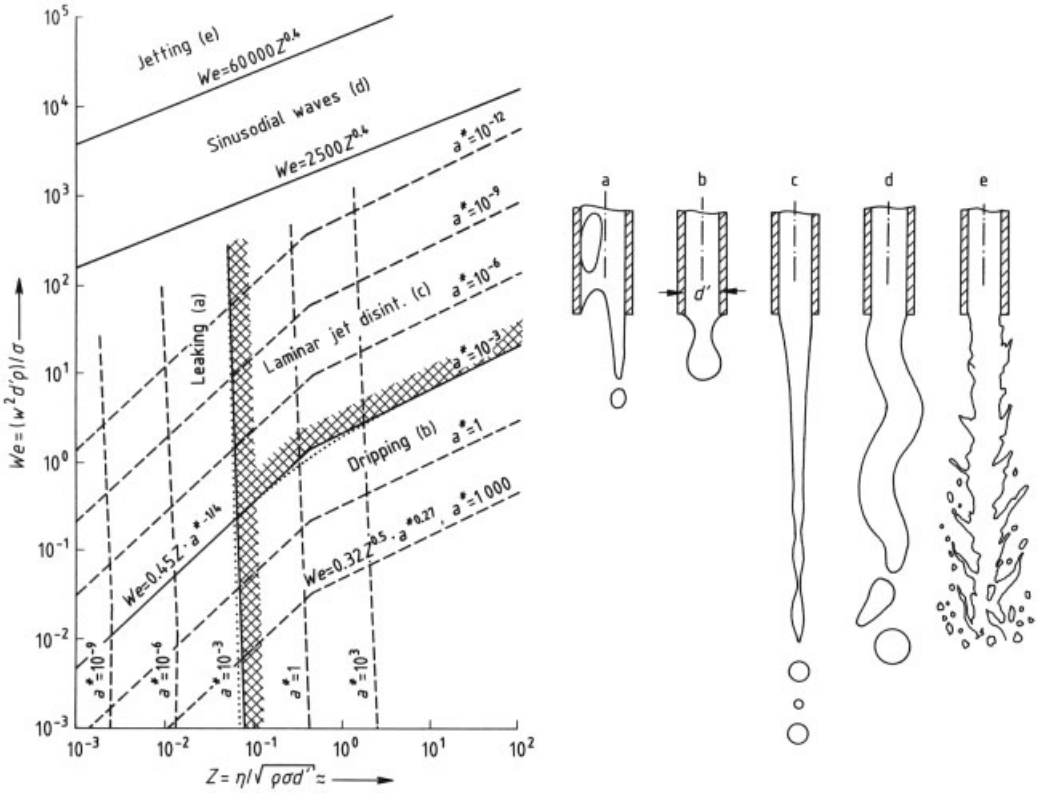

Figure 7. Diagram showing regions of various flow conditions for capillary tubes or nozzles either vertical in the gravitational field or in a centrifugal field with radial tube axis
At low discharge velocities the flow condition also depends on the Kapitza number $a*$. As an example, the boundaries for $a* = 10^{-3}$ are shaded.
$a* = g\,\eta^4/\sigma^3\,\varrho$ or $a* = a\,\eta^4/\sigma^3\varrho$ and $a = r\,\omega^2$

coalesce until finally a liquid sheet is formed on the dripping edges. In the case of plates, with increasing throughput jet coalescence also takes place and finally leads to an irregular pattern of thick jets [3].

In addition to the Weber number and the Ohnesorge number, a further characteristic number takes account of the influence of the acceleration, namely the Kapitza number $a* = a\eta^4/\sigma^3\varrho$, which describes the flow conditions for *capillary tubes* vertical to the gravitational field or in a centrifugal field. In the case of drop or jet formation due to gravity the Kapitza number is defined as $a* = g\,\eta^4/\sigma^3\varrho$. Figure 7 shows a diagram for the various flow conditions. When Z (i.e., liquid properties and nozzle diameter) is given, as the Weber number (i.e., flow velocity w) increases, the sequence of flow conditions is dripping, laminar jet disintegration, sinusodial jet disintegration, and turbulent jet disintegration. If the capillary tubes are large, the liquid column in the capillary becomes unstable and can break: the liquid partially leaks out. This occurs when $d'\sqrt{\varrho g/\sigma} > 5.2$ for $Z < 0.43 \cdot a*^{1/4}$ [18], [19].

Sinusodial jet disintegration is a transitory condition during which the jet is helically deformed by aerodynamic forces [20]. A distribution of drop sizes is produced, but the sizes lie in the range of the jet diameter. With further increasing throughput there is *turbulent flow* in the jet (Figs. 8, 9, 10), which leads to smaller drops. This also occurs at very low ambient pressures, hence the conclusion that the gas pressure has only a subordinate influence. The turbulence in the jet can be significantly pronounced by special flow duct geometries, as will be shown in Section 3.1. At extremely high discharge veloc-

ities, on the order of magnitude of the velocity of sound of the gas, the gas influence becomes increasingly noticeable.

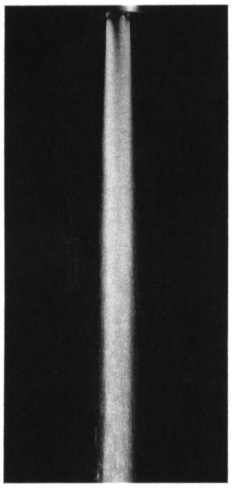

Figure 8. Visual appearance of turbulent jet from a tubular nozzle at high exit velocities (water)

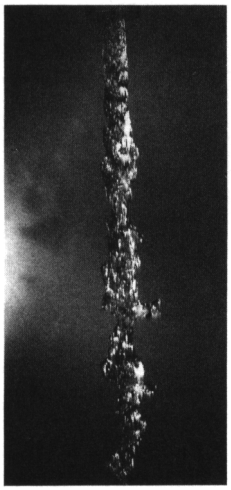

Figure 9. Short time photograph of a turbulent jet emerging from a tubular nozzle at high exit velocities (water)

2.2. Sheet or Lamella Disintegration

The production of fine drops using capillary tubes has the disadvantage that fine nozzles and high pressures are needed. A better approach is to first produce thin sheets or lamella, which subsequently break up into drops by several breaking mechanisms:

1) Edge contraction
2) Aerodynamic wave formation
3) Turbulent disturbances

Frequently these disintegration mechanisms are superimposed.

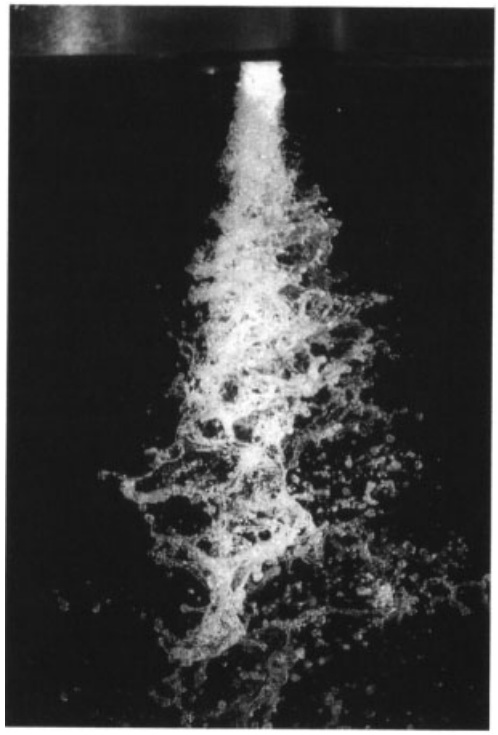

Figure 10. Jet with high degree of turbulence from a knee shaped nozzle duct (water)

In contrast to liquid threads, in which wave formation with $\lambda > \pi\, d_{\mathrm{th}}$ leads to a smaller surface and thus to increasing instability, sheets are stable because each deformation enlarges the surface and must work against a restoring force [21–23].

Most industrial atomizers form sheets that expand transversely to the main flow direction and thus decrease in thickness δ as the distance x increases. The product δx is about constant and is designated as the sheet thickness parameter $K = \delta\, x$ [24], [25]. If K is related to the cross-sectional area of the nozzle outlet (orifice) A, the sheet number $\kappa = \delta\, x/A$ is obtained [26]. With circular orifices, $\mathrm{A} = d'^{\,2}\,\pi/4$ and $\kappa = 4\,\delta x/\pi\, d'^{\,2}$. For small Weber numbers, $\sqrt{\kappa}\, We = \sqrt{\kappa}\, d'\, \varrho w^2/\sigma < 370$, *edge contraction*

takes place. The sheet edge contracts in the form of a liquid bulge, which breaks up into relatively coarse drops according to the principles of jet disintegration. At the site of the bulge the propagation velocity w of the sheet is equal to the contraction velocity, or $w^2\,\varrho\,\delta/\sigma = 2$ (see also Fig. 11).

Figure 11. Tulip formation and edge contraction of a lamella formed at an hollow cone nozzle at low exit velocities (water)

For $\sqrt{\kappa}\,We > 370$, the sheet begins to oscillate due to the influence of the gas. With increasing distance, the quickly growing amplitudes cause the sheet to break up. According to one model, liquid threads are formed by contraction from the broken-up sheet parts and these threads then break up into drops [27], [28].

For low viscosity liquids the drop size is

$$d_{32} = C_1\,(\delta\,x)^{1/3}\left(\frac{\sigma}{v_{\mathrm{rel}}^2\,\varrho}\right)^{1/3}\left(\frac{\varrho}{\varrho_{\mathrm{g}}}\right)^{1/6} \tag{9}$$

Here, $1.2 < C_1 < 1.7$, v_{rel} is the relative velocity between sheet and gas, ϱ_{g} is the gas density, and d_{32} is the Sauter mean diameter of the drop size distribution. The Sauter mean diameter has the same ratio of surface to drop volume as the whole drop distribution. This *aerodynamic wave formation* takes place for most sheet nozzles and most rotating atomizers when they are operated in the range of sheet formation under conditions of low turbulence. At very high velocities aerodynamic wave formation recedes and *turbulent disintegration* comes up which shows a stronger decrease of drop size with increasing discharge velocity, depending on the degree of turbulence in the flow.

2.3. Dispersion of Liquids by Gas

If a gas jet strikes a liquid surface with sufficient relative velocity v_{rel}, then the liquid is dispersed. Disintegration of drops in gas flows has been extensively examined [29–31], [122], [123]. Dispersion occurs when the dynamic pressure of the gas exceeds the internal pressure in the drops. The ratio of dynamic pressure to surface pressure is described by the gas Weber number, $We_{\mathrm{dg}} = v_{\mathrm{rel}}^2\,d\varrho_{\mathrm{g}}/\sigma$. Blowing first excites the drop surface into oscillation, thus promoting disintegration. Increasing the viscosity further damps the oscillation. For $Z_{\mathrm{d}} = \eta/\sqrt{\varrho\,d\,\sigma} > 4$, dispersion of drops in gas flows is no longer possible. In terms of the Weber number, the splitting limit can be approximated by

$$We_{\mathrm{dg}} = 12 + 14\left(\frac{\eta}{\sqrt{d\,\varrho\,\sigma}}\right)^{0.8} \tag{10}$$

At high relative velocities, the drop is blown up to a hemispherical sheet, which subsequently bursts into small drops. The splitting modes arising at various Weber and Ohnesorge numbers are described in more detail in [32], [122], [123].

A liquid film attached to a solid surface that is blown over by a gas forms waves from whose peaks drops detach. Although this process has been examined repeatedly [33–36], a uniform method of calculation has yet to be found. According to Kataoka and co-workers [37], for drops produced in a two-phase film flow in tubes with a diameter d′,

$$\frac{d_{32}}{d'} = 8\cdot 10^{-3}\,(We_{\mathrm{g}}\cdot Z)^{-2/3} \tag{11}$$

where $We_{\mathrm{g}} = d'v^2\varrho_{\mathrm{g}}/\sigma$, $Z = \eta/\sqrt{\varrho\,\sigma\,d'}$, and v is the gas velocity. Drop formation on pneumatic atomizers is described most often by the relationship

$$\frac{d_{32}}{d'_{\mathrm{p}}} \approx We_{\mathrm{g}}^{-k}\,(1+\mu_{\mathrm{m}})^{2\,k} \tag{12}$$

where d'_{p} is a characteristic nozzle dimension. The mixing process between a gas flow with high velocity and a liquid flow with low velocity only enables limited mixing or dispersion velocities $v_{\mathrm{e}} = v_{\mathrm{g}}\,(1+\mu)$. This conservation of linear momentum explains the effect of the loading in Eq. (12). At low loadings $\mu_{\mathrm{m}} = \dot{m}/\dot{m}_{\mathrm{g}}$ a minimum drop size is obtained.

3. Atomizing Apparatus

The basic principles of drop formation can be applied to a great number of technically employed atomizers.

3.1. Single-Substance Pressure Nozzles

Single-substance pressure nozzles are the simplest atomization devices. They convert the liquid pressure Δp into velocity. The geometry of the nozzle determines the conversion efficiency. The discharge velocity is $w = \varphi\sqrt{2\,\Delta p/\varrho}$; φ is the velocity coefficient. The throughput of single-substance pressure nozzles is $\dot{V} = A\alpha\varphi\sqrt{2\,\Delta p/\varrho}$ where $\alpha = A_{\mathrm{e}}/A$ is the contraction number. A_{e} is the cross-sectional area of the emerging jet and A is the nozzle cross-sectional area. The discharge coefficient is $\mu = \alpha\varphi$. The quantities α, φ and μ depend on the nozzle geometry and on the pressure Reynolds number $Re_{\mathrm{p}} = d'\sqrt{\Delta p\,\varrho}/\eta$, respectively. At high pressures or low viscosities , φ and α and therefore μ reach a constant limiting value, which is frequently used for calculation in practice. In practically at all single-substance pressure nozzles, variation in throughput requires a pressure change according to $\Delta p \approx \dot{V}^2$ and droplet sizes simultaneously change such that $d_{32} \approx \dot{V}^{-2/3}$; hence, throughput variation is often only possible by switching single nozzles on or off.

Sheet Nozzles. When fine droplets are required, large orifice diameters of the nozzle are still desirable to minimize the plugging behavior of the spraying system. It was found that sheet-forming nozzles are able to produce fine droplets relative to the orifice diameter even with a relatively low pressure drop. Particularly thin sheets and correspondingly fine droplets are formed at hollow cone or swirl nozzles as long as low viscous liquids are sprayed, e.g., $Z = \eta/\sqrt{\sigma\,\varrho\,d'} < 0.05$. The larger the spray angle the thinner the lamella at a given distance from the nozzle and the smaller the droplets. For more viscous liquids, fan jet nozzles are usually employed despite of the thicker lamellas at a given distance from the nozzle because they still ensure a relatively high discharge velocity due to lower friction in the flow channel.

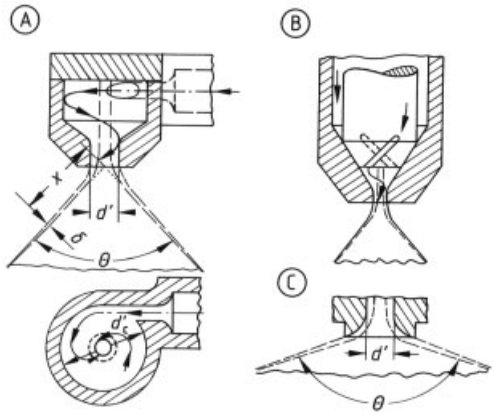

Figure 12. Hollow cone nozzles, also called swirl atomizers
A) Nozzle with tangential inlet (side and top views); B) Nozzle with slitted cone insertion; C) Deviation aperture

Hollow cone nozzles have a swirl chamber into which the liquid flows tangentially through one or more inlets (see Fig. 12 A and B). A gas core forms in the middle of the swirl chamber. At the orifice, with diameter d', a hyperbolic sheet forms; with increasing distance it becomes thinner and breaks up into drops by aerodynamic wave formation (Fig. 13). A minimum pressure is required [26], [38]. For liquids with low viscosity this minimum pressure is given by

$$\sqrt{\varkappa}\,\Delta p^{*}\varphi^{2} > 185 \tag{13}$$

with $\Delta p* = \Delta p\,d'/\sigma$. Common values for the velocity coefficient for hollow cone nozzles operated at high Re_{p} numbers lie in the range $0.7 < \varphi < 0.9$. The sheet number is obtained from the spray angle θ,

$$\varkappa = \frac{\mu}{2\,\pi\,\varphi \sin\,(\theta/2)} \tag{14}$$

with $\mu = 4\,\dot{V}/d'^2\pi\,(2\,\Delta p/\varrho)^{1/2}$. In practice, the dependence of $\dot{V}$ on Δp noted in the nozzle manufacturer's pamphlet is best taken. For low viscosity liquids, sheet number and discharge velocity coefficient can also be calculated from potential theory [39], [40]. Considerable deviations, however, arise even at relatively low viscosities [41]. In this case, the velocity coefficient can then drop even to $\varphi = 0.35$ [42].

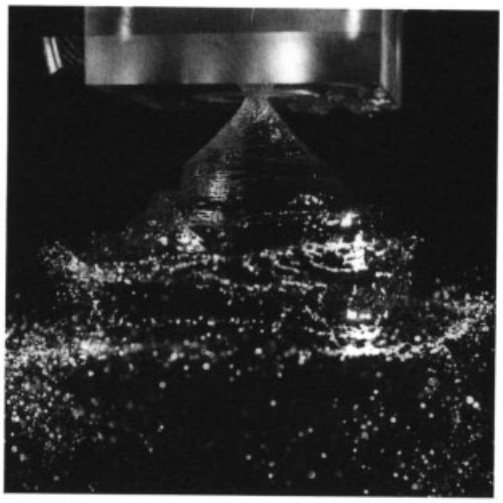

Figure 13. Aerodynamic wave formation on a lamella formed at an hollow cone, swirle type nozzle resp. at high exit velocities (water)

The spray angle decreases as the viscosity increases. The throughput, however, increases because in the nozzle the annular area occupied by the liquid increases while the gas core area shrinks. The drop size is approximately [26]

$$\frac{d_{32}}{d'} = 1.2\left(\frac{\varkappa}{2\varphi^2\Delta p^*}\right)^{1/3}\left(\frac{\varrho_g}{\varrho}\right)^{-1/6} \tag{15}$$

Since the velocity coefficient φ increases with increasing Reynolds number or pressure, respectively, and κ decreases with increasing Reynolds number, the droplet size can also be calculated by equations presented in [124]. Further examinations on swirl nozzles are described in [125–130]. In conical spray jets formed at hollow cone nozzles or swirl nozzles as shown in Figure 12 A and 12 B a graduation arises, which leads to retro-transport of small drops in a back-flow zone. The large drops remain on the jet edge (Fig. 13) [43], [44], [45].

The spray angle can be enlarged by a special aperture as in Figure 12 C, where the Coanda Effect is utilized [26].

To adjust the throughput, part of the liquid can be removed from the swirl chamber [46] without major effect on the drop size [126].

Fan Jet Nozzles. Sector-shaped sheets can be produced by suitable nozzle geometries. Figure 14 shows two designs for fan jet nozzles. A frequent design consists of a nozzle with a spherical bore slitted by a cross groove (Fig. 14 A). Another construction called tongue nozzle is shown in Figure 14 B. The circular jet is deflected by an inclined plate producing a sector shaped lamella.

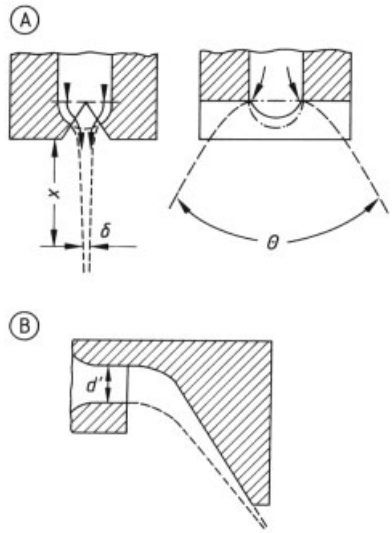

Figure 14. Fan jet nozzles
A) Nozzle with spherical bore and transverse slit; B) Tongue nozzle

The flow through such a nozzle causes streamlines that diverge in the slit direction but converge transversally, forming a sector-shaped sheet fan (Fig. 15). Discharge coefficient, velocity coefficient, and sheet number (μ, φ, κ) must be determined by experiment [24], [26]. Whereas μ and κ can be calculated from data available from nozzle manufacturers, φ must be estimated or measured by determination of the momentum flux. This can be done by measuring either the nozzle thrust or the force caused by the impingement of the jet on a plate. With low viscosity liquids or for $Re_p > 1000$, fan jet nozzles have much higher κ values, in the range

$0.4 < \kappa < 0.7$, than do hollow cone nozzles. For lower Re_p numbers or highly viscous liquids, however, the velocity coefficient decreases much less than in the case of hollow cone nozzles. The discharge coefficient μ also decreases, which counteracts an increase in sheet thickness. For these reasons, fan jet nozzles are also suitable for highly viscous liquids. The sheet number can be estimated by

$$\kappa = \frac{360}{2\pi\theta}\frac{\mu}{\varphi} \tag{16}$$

The projection surface A of the contour in the flow direction should be used for the nozzle cross-sectional area. The equivalent nozzle diameter is then $d' = \sqrt{4A/\pi}$. Calculation of the drop size is equal to Equation (15). For low viscosity liquids the velocity coefficient lies in the range $0.8 < \varphi < 0.95$.

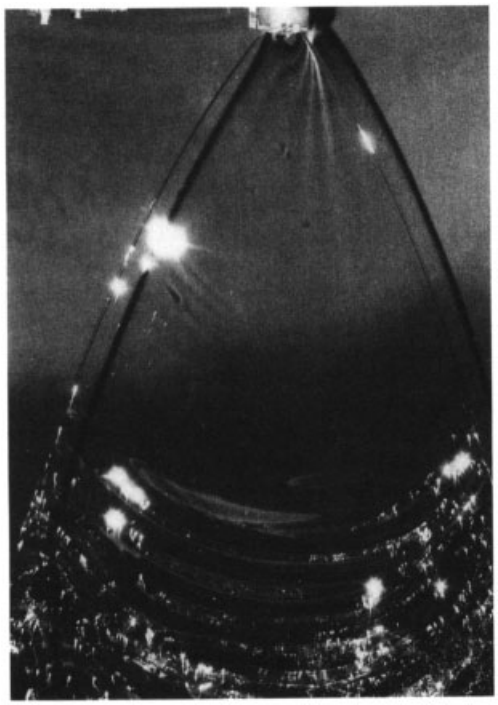

Figure 15. Aerodynamic wave formation of a fluid sheet emerging from a fan jet nozzle (viscous oil)

Fan jet nozzles are employed for atomizing lacquers and the like because of the still relatively high φ values even at higher viscosities (airless atomization). They are also used for coating of moving sheets and in washing plants, which require high jet impulse. Deceleration of the spray jet by the ambient gas is less than in the case of hollow cone nozzles [47].

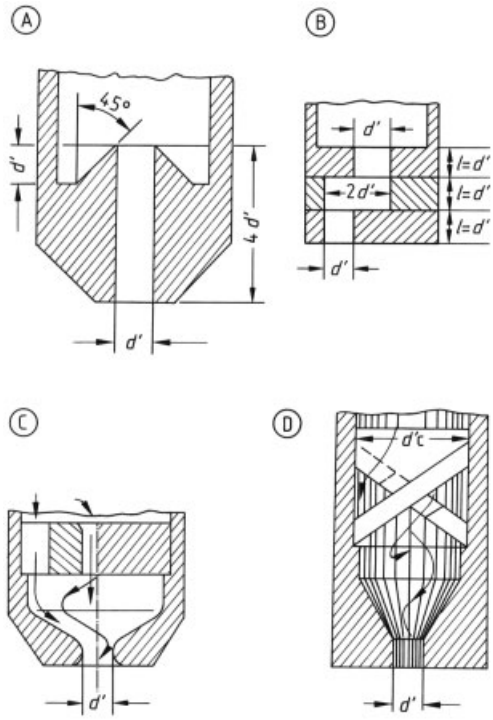

Figure 16. Full cone nozzles (turbulence nozzles)
A) Nozzle with Borda aperture; B) Nozzle with knee shaped, sharp edged duct; C) Nozzle with turbulence-amplifying and swirl producing insertion to avoid the formation of an air core, the insert has a central drill hole; D) Nozzle with sharp edged, swirl and turbulence inducing insert

Turbulence Nozzles. For uniform wetting of large surfaces, full cone nozzles are frequently employed. Compared with swirl nozzles they produce large drops [48]. By insertion of a swirl body in the hollow cone nozzle a swirl and strong turbulence are simultaneously produced (see Fig. 16 C and D). The swirl also makes large spray angles, up to 120°, possible. Formation of an air core must be prevented because otherwise a hollow jet forms. Other possibilities of turbulence amplifications are the Borda aperture, as in Figure 16 A, or sharp edged deflections in the nozzle duct (see Fig. 16 B [108]). Jet disintegration is caused by the strong internal turbulence. Here, both temporally and spatially irregular sheet structures develop and form drops (see Fig. 10). This process is also effective in vacuum.

Turbulent jets emerging from smooth tube nozzles disintegrate to fine drops only at high discharge velocities. The spray angle is small. According to Glaser [49] the relative drop size from a nozzle with a turbulence-amplification Borda aperture (Fig. 16 A) is

$$\frac{d_{32}}{d'} = 400\,\Delta p^{*-3/4}\left(\frac{\varrho}{\varrho_g}\right)^{-1/5} Z^{1/7} \tag{17}$$

in the range $20\,000 < \Delta p* < 100\,000$ and $2\times10^{-3} < Z < 7\times10^{-2}$. With extremely high discharge velocities, of the order of the velocity of sound in the gas, fine drops can also be produced by smooth nozzles, e.g., diesel injection [50–53], [131].

For a nozzle as in Figure 16 D spraying low viscosity liquids into air in the range $2\times10^4 < \Delta p* < 1.5\times10^5$, measurements with a laser diffraction spectrometer showed that

$$\frac{d_{32}}{d'} = C_3\,(\Delta p^*)^n \tag{18}$$

For the ratio $d'_c/d' = 2.7$, a spray angle $\theta = 110°$ and the parameters $C_3 = 23$, $n = -0.6$, and $\mu = 0.72$ were found.

3.2. Pneumatic Atomizers

By using gases, small drops can be produced even at low pressures because of high relative velocities. The cross section of the liquid channel plays a subordinate role, so it can be selected large enough to avoid clogging. Most atomizer nozzles with external mixing depend on the principle of *prefilming* [54]. First, the liquid is spread by the gas flow as a thin film on the atomizer surface. A wave flow arises on the film surface and is blown over an edge, where the actual drop formation occurs [55]. The thinner the liquid film on the detachment edge the smaller the drops.

The drop sizes can by calculated from equations proposed in [124] for prefilming nozzles

$$\frac{d_{32}}{d'_h} = \left[0.33We_g^{-0.6}\left(\frac{\varrho}{\varrho_g}\right)^{0.1} + 0.068Z_p^{0.5}\right]\left[1+\mu_m\right] \tag{19}$$

The gas Weber number and Z_p are derived from the pneumatic diameter as $We_g = \varrho_g v^2 d'_p/\sigma$, where d'_p is the diameter of the detachment lip or prefilming diameter, respectively. d'_h is the hydraulic mean diameter of the gas exit duct, i.e., $d'_h \approx 2\,S_p$, S_p being the slit width of the gas duct. The data were obtained from the ratio of $d'_h/d'_p = 0.22$. It must be taken into account that calculation formulas are only valid for certain geometries. See also [56], [57] and Figure 17 A.

Normal loading of twin fluid atomizers with external mixing lies in the range $0.1 < \mu_m < 2$. The spray angle lies in the range $20° < \theta < 40°$. To enlarge the spray angle, the air is frequently fed with swirl (Fig. 17 B). For atomization of lacquers using twin fluid atomizers, see [58]. The energy demand of pneumatic nozzles in their usual operation range is higher than that of single substance or simplex pressure nozzles. Droplet size control is easier because of additional adjustable parameters such as gas pressure and cross-sectional area of the gas slit.

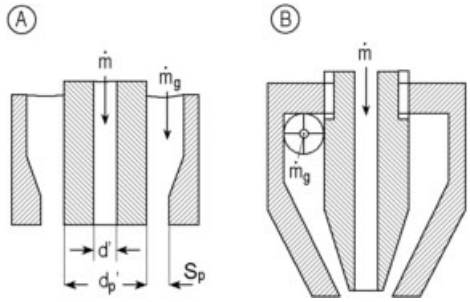

Figure 17. Pneumatic or twin-fluid nozzles with external mixing
A) Nozzle with parallel flow; B) Nozzle with adjustable gas annulus and tangential gas inlet for spray angle $\theta > 30°$

While nozzles with external mixing produce suction in the liquid channel, in nozzles with internal mixing (see Fig. 18 A and B) the liquid must be fed or dosed under pressure. The two-phase mixture emerges from the nozzle aperture with the velocity of sound, which is in the range $20 < w < 50$ m/s. Then the gas suddenly expands to ambient pressure, suddenly accelerating and breaking up the large drops that have been carried along (see Section 2.3 and also [59–61]). Nozzles with internal mixing chambers are able to produce fine drops even at high ratios of liquid-to-air mass flow rates, $\mu_m > 2$ and even with high-viscous liquids, i.e., $\mu < 0.3$ Pa s. They show less tendency to fluctuations and oscillations in the spray pattern. Their disadvantage is the need for a more complex flow control sys-

tem. The spray pattern of pneumatic atomizers was examined extensively in [62–64].

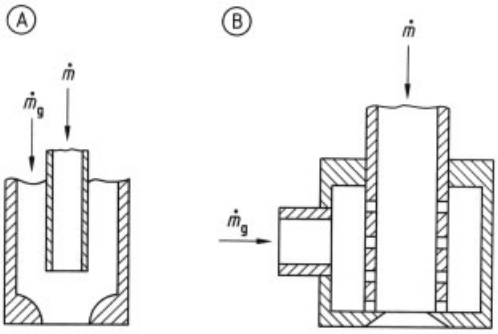

Figure 18. Pneumatic nozzles with internal mixing
A) Nozzle with concentric ducts for gas and liquid; B) Nozzle with predispersion chamber [60]

3.3. Atomization with Propellants

Spray cans contain propellants, formerly halogenated hydrocarbons, now simple hydrocarbons, e.g., butane, together with the substance to be sprayed. When the valve is opened, the liquid components flow through a nozzle, usually a hollow cone nozzle. The propellant vaporizes in the nozzle and promotes disintegration of the emerging sheet by the expanding bubbles of gas [65]. Liquid nitrogen and carbon dioxide can be used as propellants as well [66].

3.4. Rotary Atomizers

Rotary atomizers are used most frequently to make fine droplets $d < 100$ µm from suspensions and liquids with a tendency to clogging. For that purpose they are driven by special motors with speeds up to 20 000 min^{-1}. Rotary atomizers produce a bell-shaped spray, which, e.g., is utilized in spray drying. A tower with a large diameter is required so that the drops flying furthest do not touch the side of the tower before drying. The drying gas therefore often is supplied through an annular clearance near to the atomizer in order to achieve a more conical spray pattern together with high mass-transfer rates.

The liquid is fed centrally on rotating cups, disks, or plates. Figure 19 shows two types of rotating atomizers. The liquid flows to the edge of the atomizer due to the centrifugal force. The drop formation mechanisms depend on the liquid feed rate. The processes correspond to the flowing from edges in the gravitational field (Section 2.1). For calculation, however, instead of the acceleration of gravity g, the centrifugal acceleration $a = r\omega^2$ is used, where r is the atomizer radius and ω the angular frequency of rotation. Dripping is attained at low throughputs. The drop size is calculated with Equation (2). Increasing the throughput causes regular liquid threads to form at equal intervals along the circumference of the atomizer. This occurs when the flow rate per detachment point is large enough to form a liquid thread. Compare also [4], [67–70]:

$$1.0 < \dot{V}^*_{\text{th}} < 1.5 \tag{20}$$

$\dot{V}^*_{\text{th}} = \dot{V}^*_{\text{th}} \left(a^3 \varrho^5 / \sigma^5\right)^{1/4}$ is the nondimensional throughput per liquid thread. The number of threads formed per unit of length at the detachment edge is given by $N = (L/6)\,(\varrho a/\sigma)^{1/2}$. L is equal to the circumference of rotating discs. In the case of vaned discs, L is the number of vanes multiplied by the vane height. The dimensionless droplet size in the thread region lies in the range of $1.3 < d^* < 2.0$. The drops are smaller than in the dripping regime. If the desired droplet size is known, the necessary centrifugal acceleration $a = r\omega^2$ can be calculated. From the given throughput the total required length of all detachment edges can be obtained. The width of droplet size distribution is small as expected from Rayleigh-type desintegration. However, coalescence sometimes causes double or triple droplets leading to multimodal distributions [4]. In addition, from low viscous liquids satellite drops with $d_s \approx d_{32}/4$ are also formed.

In chemical production plants rotary atomizers are mostly operated in the range of sheet formation because of higher throughput. The drop size distribution is much broader than in the thread regime, $2 < d_{\max}/d_{32} < 3$. Furthermore, a considerable amount of fine material is obtained. The transition from thread to sheet formation occurs when the liquid feed per length of the detachment edge $\dot{V}/L$ at a given speed of the atomizer exceeds the contraction velocity of the liquid film, i.e., when $We_\delta > 2$. The transition derived is:

$$\dot{V}/L > (2\sigma/\varrho)^{3/5}\ (3\eta/\varrho a)^{1/5} \tag{21}$$

In that case lamella are formed which desintegrate by combined mechanisms, i.e., aerodynamic wave formation, turbulence, and destruction in the ambient gas. Allthough most atomizers are operated in the lamella regime, it is difficult to recommend a calculation method which also considers the geometrical shape of the atomizer. An estimate on the droplet size in the lamella regime may be obtained from [71].

$$d_{32}^{*} = 1.44\left(We_{\mathrm{r}} \cdot Z_{\mathrm{r}} \dot{V}^{*}\right)^{0.20} \tag{22}$$

(For definition of symbols see the abbreviations list at the beginning of this article.)

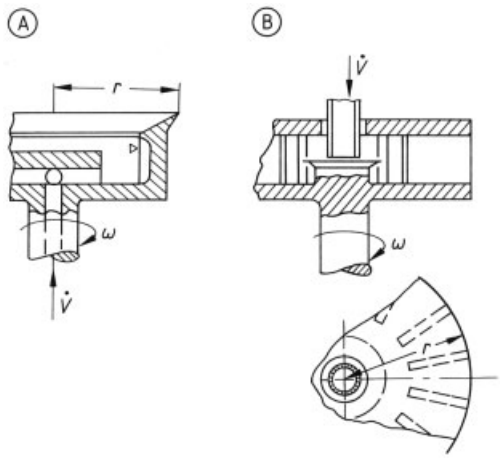

Figure 19. Rotary atomizer
A) Cup atomizer with distribution nuts and liquid feed through a hollow shaft; B) Vaned atomizer disk

The relationships in Equations (19) – (23) can be used for vaned disks when in the definition for $\dot{V}*$ – the entire quantity [(circumference + vane height) × vane number] – is inserted instead of the circumference $2\ \pi\ r$. Especially on vaned disks the liquid distribution is often imperfect so that local thick liquid strands cause thick drops. Special constructions of rotary atomizers are the rotating porous body [16], [17], the twin roller atomizer [72], [73], and the spin top atomizer with pneumatic drive [74].

Lamella production and broad droplet size distributions can be avoided when grooved cups or multiple-hole cylinders are employed. Grooves [132] and bores are able to maintain the laminar thread regime over a significantly larger throughput range. The number of threads are then given by the number of grooves or bores at the circumference of the atomizer. The aforementioned geometry is used, e.g., to spray paints in carcas finishing [75] often in combination with electrostatical charging of the droplets. The latter has also been designed as a multiple-hole cylinder for spraydrying operations [133] in order to achieve narrow particle size distributions despite of high mass flow rates.

3.5. Ultrasonic Atomization

Various principles can be differentiated in ultrasonic atomization [76], namely

1) Capillary waves [77–80]
2) Cavitation [79]
3) Stationary ultrasonic waves in a gas [81], [134]
4) Excited jet disintegration (Section 2.1)

The *capillary wave* is the most important; it enables production of drops with extremely low residual impulse. The drop size can be approximated by

$$d_{32} \approx 0.8\ \sqrt[3]{\frac{\pi\sigma}{\varrho f^{2}}} \tag{23}$$

where f is the excitation frequency. With water, when $f = 100$ kHz, a medium drop size of $d_{32} = 25$ µm results. Here the atomization rate of $10\ \mathrm{cm^3/cm^2\ s}$ ($= \mathrm{cm\ s^{-1}}$) is obtained at an oscillator amplitude $A = 5$ µm. Capillary wave atomization is only suitable for low viscosity liquids, $\eta < 20$ mPa s, and produces a drop distribution.

3.6. Electrostatic Atomization

In addition to the dispersion of liquids by high potentials or field forces, the charging of drops produced mechanically is also sometimes advantageous [82], [83]. In both cases, charging causes the spray particles to drift apart. In this way, coalescence is avoided. Subsequently, the droplets can be deposited with a high yield on an oppositely charged, or simply a grounded body. If a drop is charged, it breaks apart when the

repulsive force of the charges on the surface exceeds the surface force [84–89],

$$d\pi\sigma \leq \frac{Q^2}{8\pi\varepsilon d^2} \tag{24}$$

where Q is the charge on the drop and ε the dielectric constant of the liquid. The same mechanism is valid for the detachment of drops from highly curved liquid surfaces, e.g., the tip of a capillary tube. Even at relatively low potentials a high field strength arises and produces high charges.

The throughput of electrostatic atomizers is relatively low, $\dot{m}_{\max} \approx 1$ kg/h. However at high field strengths, $E \geq 10^{10}$ V/m, drops can be produced in the range $0.1 < d < 5$ µm, the diameter depending on the liquid characteristics. The usual potentials lie in the ranges $5 < U < 100$ kV. An overview can be found in [135–137].

4. Drop Size Measurements and Drop Size Distributions

Drop sizes can be measured by short time photography. A holographic figure, which enables evaluation of the spatial spray and presents no problems with the focus depth, is even better. In this way, in combination with double exposures at defined time intervals, the magnitude and direction of the velocity can also be determined [90]. The laser diffraction method can measure drop sizes in the range $0.5 < d < 500$ µm quickly and online. The method presents the drop size distribution found in the measuring volume (laser beam).

In addition to the measurement of single drops by the scattered light counting method [91], [92], the laser phase anemometry is applied for simultaneous velocity and drop size determination [93–96], [138–140]. The local droplet flux can also be determined, informing about the actual production rates in each size class and on given trajectories. Mechanical methods [97] are sometimes also employed.

All known distribution functions are used as suitable model functions for the measured drop size distributions: RRSB [98], [99] and log normal distribution [100], [101] for example. The ULLN distribution has the advantage that it permits the definition of a maximum drop diameter [102], [103]. For the significance of drop size distributions in jet scrubbers, see [104].

An approximation of the distribution width in different atomization processes in given in [102], [105–107] and Table 1.

Table 1. Distribution widths in atomization processes

Atomization process	Measure of distribution	
	$d_{v,0.5}/d_{32}$	$d_{\max}/d_{v,0.5}$
Dripping	1	1.1
Natural laminar jet disintegration	1.05 – 1.1	1.4
Sheet nozzles	1.1 – 1.3	1.5 – 2.0
Turbulence nozzles	1.3 – 1.5	2.5 – 3.0
Twin-fluid atomizers	1.1 – 1.5	2.5 – 3.5
Rotary atomizer		
threads	1.05 – 1.1	1.4 – 2.0
sheets	1.1 – 1.5	2.5 – 3.5
Capillary wave atomizer	1.05 – 1.3	1.5 – 2.5

The efficiency of energy transformation from pressure to surface energy depends on the process and on the ratio of the Sauter mean droplet size to the orifice diameter [141]. This ratio is a main indicator to the plugging behavior of a spraying system. In the case of simplex pessure nozzles, hollow cone nozzles have highest efficiency rates of $0.01 < e < 0.05$ over an acceptable operation range. In the case of pneumatic nozzles the efficiency lies within $0.01 < e < 0.002$. Rotary atomizers may reach efficiencies up to $e = 0.05$. Highest efficiency rates of up to $e = 0.3$ are achieved by applying the Rayleigh-type jet disintegration at capillaries. However, large droplets relative to the cappilary diameter, i.e., $d/D = 1.9$, are formed and often give rise to plugging of the capillaries when small drops are required.

Special cases are the atomization of metals [109–111], [142] for the production of fine powders or spray-compacted metal parts [143–145] and drop impingement [112], [113], [146], [147]. Destruction of droplets impinging on walls is very important in disintegrators, washing towers and fuel suppying systems in motors [148], [149]. For the application of similarity theory to atomizers see [114]. The drag coefficients of drops in gas flows can be found in [115–118].

The application of spray processes is also treated in Combustion, Adsorption, Dust Sepa-

ration and Scrubbers, Particle Size Analysis and Drying of Solid Materials.

5. References

General References

1. R. Clift, J. R. Grace, M. E. Weber: *Bubbles, Drops, and Particles,* Academic Press, New York 1978. The *Journal of Atomization and Spray Technology* (Elsevier Applied Sience), which began in 1985 with vol. 1, is a good source of information on spraying and atomization of liquids. Other good sources are the *Proceedings of the International Congresses of Liquid Atomization and Spraying Systems* (ICLASS): 1) Tokyo (1978), 2) Madison, Wisc., (1982), and 3) London (1985). Sendai, Japan (1988), 5) Gaithesburgh, USA (1991), 6) Rouen, France (1994), 7) Seul, Corea (1997).

Specific References

2. Y. Tanasawa, S. Toyoda, *Technol. Rep. Tohoku Univ.* **21** (1955) 135.
3. P. Walzel, U. Klaumünzner, *Ger. Chem. Eng. (Engl. Transl.)* **4**(1981) 154 – 160.
4. M. R. Mehrhardt, Dissertation, TU Berlin 1978.
5. W. D. Harkins, F. E. Brown, *J. Am. Chem. Soc.* **41** (1919) 499 – 524.
6. M. Hozawa, T. Tsukada, N. Imaishi, K. Fujinawa, *J. Chem. Eng. Jpn.* **14** (1981) no. 5, 358 – 364.
7. M. R. Lindblad, J. M. Schneider, *J. Sci. Instrum.* **42** (1965) 635 – 658.
8. P. Walzel, *Chem. Ing. Tech.* **52** (1980) no. 8, 652 – 654.
9. Lord Rayleigh, *Proc. London Math. Soc.* **10** (1978) 4/13.
10. C. Weber, *Z. Angew. Math. Mech.* **11** (1931) 136 – 154.
11. R. N. Berglund, Y. H. Lin, *Environ. Sci. Technol.* **7** (1973) no. 2, 147 –153.
12. H. Hausner, P. W. Taubenblat: *Proc. Int. Powder Met. Conf. Princeton, N. J.,* (1976).
13. P. Walzel, *Chem. Ing. Techn.* **51** (1979) no. 5, MS 692 – 679.
14. P. Schümmer, K. Tebel, *Chem. Ing. Tech.* **12** (1981) MS 961/81.
15. Ch. Müller, F. Widmer, *Swiss. Chem.* **8** (1986) no. 2 a, 31 – 43.
16. P. Schmidt, *Chem. Ing. Tech.* **39** (1967) nos. 5/6, 375 – 379.
17. W. Gösele, *Chem. Ing. Tech.* **40** (1968) nos. 1/2, 37 – 43.
18. M. Dosoudil, *Chem. Ing. Tech.* **43** (1971) no. 21, 1172 – 1176.
19. P. Walzel, H. Michalski, *VI Verfahrenstechnik* **14** (1980) no. 3, 157 – 159.
20. W. v. Ohnesorge, *Z. Angew. Math. Mech.* **16** (1936) 355 – 358.
21. W. N. Bond, *Proc. Phys. Soc. London* **47** (1935) no. 4, 549 – 558.
22. G. I. Taylor, *Proc. R. Soc. London. Ser. A* **253** (1959) 313 – 321.
23. H. B. Squire, *Br. J. Appl. Phys.* **4** (1953) 167 – 169.
24. N. Dombrowski, D. Hasson, D. E. Ward, *Chem. Eng. Sci.* **12** (1960) 35 – 50.
25. D. Hasson, J. Mizrahi, *Trans. Inst. Chem. Eng.* **39** (1961) 415 – 422.
26. P. Walzel, *Ger. Chem. Eng. (Engl. Transl.)* **7** (1984) 1 – 12.
27. R. P. Fraser, P. Eisenklam, N. Dombrowski, D. Hasson, *AIChE J.* **8** (1962) 672 – 680.
28. N. Dombrowski, W. R. Johns, *Chem. Eng. Sci.* **18** (1963) 203 – 214.
29. J. O. Hinze, *AIChE J.* **1** (1955) no. 3, 289 – 295.
30. F. C. Haas, *AIChE J.* **10** (1964) no. 6, 920 – 924.
31. R. S. Brodkey: *The Phenomena of Fluid Motion,* Addison-Wesley, Reading, Mass., 1967.
32. S. A. Krzeczkowski, *Int. J. Multiphase Flow* **6** (1980) 227 – 239.
33. J. W. Miles, *J. Fluid Mech.* **3** (1957) 185 – 204.
34. S. Ostrach, A. Koestel, *AIChE J.* **11** (1965) no. 2, 294 – 303.
35. G. F. Hewitt, N. S. Hall-Taylor: *Annular Two Phase Flow,* Pergamon Press, Oxford 1970.
36. P. Andreussi, J. C. Asali, T. J. Hanratty, *AIChE J.* **31** (1985) 119 – 126.
37. J. Kataoka, M. Ishii, M. Kaichiro, *Trans. ASME* **105** (1983) 230 – 238.
38. I. C. P. Huang, *J. Fluid Mech.* **43** (1970) 305 – 319.
39. G. J. Taylor, *Proc. 7th Int. Congr. Appl. Mech.* **2** (1948) 392 – 400.
40. W. C. Nieuwkamp, *Proceedings ICLASS 85,* **III C/1** (1985) pp. 1 – 9.
41. M. Horvay, Dissertation, TU Karlsruhe 1981.
42. N. K. Rizk, A. H. Lefebvre, *Atom. Spray Technol.* **2** (1983) no. 3, 219 – 233.
43. P. H. Rothe, J. A. Block, *Int. J. Multiphase Flow,* **3** (1977) 263 – 272.
44. S. Y. Lee, R. S. Tankin, *Int. J. Heat Mass Transfer* **27** (1984) no. 3, 351 – 361.

45. K. Lampe, K. Baukhage, *Chem. Ing. Tech.* **67** (1995) 481 – 483.
46. K. Lohoff, *Wärme-Gas, Int.* **31** (1982) nos. 2 – 3, 87 – 93.
47. F. E. J. Briffa, N. Dombrowski, *AIChE. J.* **12** (1966) no. 4, 708 – 717.
48. E. Sada, K. Takahashi, K. Morikawa, S. Ito, *Can. J. Chem. Eng.* **56** (1978) 455 – 459.
49. H. W. Glaser, *Brennst. Wärme Kraft* **38** (1986) no. 5, 193 – 200.
50. R. P. Grant, *AIChE. J.* **12** (1966) no. 4, 669 – 678.
51. R. E. Phinney, *Phys. Fluids* **16** (1973) no. 2, 193 – 196.
52. G. T. Sato, H. Tanabe, H. Fujimoto, *Proceedings ICLASS 82* (1982) 229 – 235.
53. F. Ruiz, N. Chigier, *Proceedings ICLASS 85,* **VI B/ 3/1**.
54. N. Dombrowski, R. P. Fraser, *Philos. Trans. R. Soc. London, Ser A* **247** (1954) 101 – 130.
55. R. P. Fraser, P. Eisenklam, N. Dombrowski, *Br. Chem. Eng.* **2** (1957) 610 – 613.
56. J. Y. Deysson, J. Karian, *Proceedings ICLASS 78,* (1978) 243 – 249.
57. S. Nukiyama, Y. Tanasawa, *Nippon Kikai Gakkai Ronbunshu* **5** (1939) 1 – 4.
58. K. W. Thomer, Dissertation, Stuttgart 1981.
59. T. Sakai, M. Kuroda, M. Saito, *J. Chem. Eng. Jpn.* **12** (1979) no. 2, 86 – 90.
60. J. M. Chawla, *Proceedings ICLASS 85,* LP/1 A/5/ 1 – 7.
61. R. W. Tate, *Proceedings ICLASS 85,* II C/1/1 – 13
62. K. Bauckhage, *Chem. Ing. Tech.* **62** (1990) no. 8, 613 – 625.
63. V. Uhlenwinkel: "Zum Ausbreitungsverhalten der Partikel bei Sprühpaktierung von Metallen," Dissertation Universität Bremen, 1992.
64. H. M. Liu, "Berechnungsmodelle für die Geschwindigkeiten und die Abkühlung von Tropfen im Srühkegel eine Stahl-Zerstäubungsanlage," Dissertation Universität Bremen, 1990.
65. N. Namiyama, N. Tokuka, P. Eisenklam, G. T. Sato, *Proceedings ICLASS 85,* III B/2/1 – 9.
66. J. Nagel, O. Molerus, K. E. Wirth, *Chem. Ing. Tech.* **69** (1997) no. 6, 1257 – 1258.
67. J. O. Hinze, H. Milborn, *J. Appl. Mech.* **17** (1950) 145 – 153.
68. S. Matsumoto, K. Saito, Y. Takashima, *J. Chem. Eng. Jpn.* **7** (1974) no. 1, 13 – 19.
69. S. Matsumoto, Y. Takashima, *Kagaku Kogaku* **33** (1969) 357 – 362.
70. S. Matsumoto, D. W. Belcher, E. J. Crosby, *Proceedings ICLASS 85,* /I A/1/1 – 21.
71. S. J. Friedman, F. A. Gluckert, W. R. Marshall, *Chem. Eng. Prog.* **48** (1952) no. 4, 181 – 191.
72. A. R. Singer, A. D. Roche, L. Day, *Powder Metall.* **23** (1980) 673 – 682.
73. P. Walzel, *Ger. Chem. Eng. (Engl. Transl.)* **5** (1982) 121 – 129.
74. T. P. Mitchell, R. L. Stone, *J. Phys. E.* **15** (1982) 565 – 569.
75. G. Schäfer, *Ind.-Lackierbetr.* **49** (1981) no. 6, 193 – 199.
76. P. Schmidt, *Maschinenmarkt* **91** (1985) no. 72, 1419 – 1421.
77. W. Eisenmenger, *Acustica* **9** (1959) 327 – 340.
78. U. Reimann, R. Pohlmann, *Forsch. Ingenieurwes.* **42** (1976) no. 1, 1 – 7.
79. M. K. Li, H. S. Fogler, *J. Fluid Mech.* **88** (1978) no. 3, 499 – 528.
80. L. H. Berger, *Proceedings ICLASS 85,* I A/2/1 – 13.
81. H. Haberfelner, *Z. Umwelttechn.* **22** (1978) 10.
82. K. Heberlein, *Met. Oberfläche* 39 (1985) 355 – 358.
83. E. Moser, H. Ganzelmeier, K. Schmidt, *Nachr. Deut. Pflanzenschutz* **33** (1981) no. 10, 145 – 157.
84. B. Vonnegut, R. L. Neubauer, *J. Coll. Sci.* **7** (1952) 616 – 622.
85. V. G. Drozin, *J. Coll, Sci.* **10** (1955) 158 – 164.
86. W. Simm, *Chem. Ing. Tech.* **41** (1969) no. 8, 503 – 507.
87. A. Kelly, *J. Appl. Phys.* **47** (1976) no. 12, 5264 – 5270.
88. A. Kelly, *J. Appl. Phys.* **49** (1978) no. 5, 2621 – 2628.
89. A. G. Bailey, *Atom. Spray. Technol.* **2** (1986) no. 2, 95 – 134.
90. M. Schäfer, *Fortschr. Ber. VDI Z* **3** (1986) 126.
91. K. Leschonski, *Chem. Ing. Tech.* **50** (1978) no. 3, 194 – 203.
92. G. Schuch, H. Umhauer, *VI Verfahrenstechnik* **14** (1980) no. 4, 237 – 241.
93. R. Semiat, A. E. Dukler, *AIChE J.* **27** (1981) no. 1, 148 – 153.
94. F. Durst, *Trans. ASME* **104** (1982) 284 – 296.
95. G. Dibelius, H. Voß, *VDI Ber.* **487** (1983) 131 – 136.
96. K. Bauckhage, V. Wassmansdorff, *Chem. Ing. Tech.* **57** (1985) no. 8, MS 1388 – 1385.
97. A. Buerkholz, *Chem. Ing. Tech.* **45** (1973) no. 1, 1 – 7.
98. E. Rammler, A. Bahr, *Freiberg. Forschungsh. A* **A 590** (1978) 95 – 109.
99. J. S. Chin, A. H. Lefebvre, *Proceedings ICLASS 85,* IV A/1/1 – 12.

100. S. Matsumoto, Y. Takashima, *J. Chem. Eng. Jpn.* **4** (1971) no. 3, 257 – 263.
101. L. P. Bayvel, *Atom. Spray. Technol.* **1** (1985) 3 – 20.
102. R. A. Mugele, *AIChE. J.* **6** (1960) no. 1, 3 –8.
103. A. Lekic, R. Bajramovic, J. D. Ford, *Can. J. Chem. Eng.* **54** (1976) no. 5, 399 – 402.
104. D. D. Schegk, F. Löffler, *Chem. Ing. Tech.* **59** (1987) no. 4, 319 – 322.
105. R. P. Fraser, P. Eisenklam, *Trans. Inst. Chem. Eng.* **34** (1956) 294 – 319.
106. R. A. Mugele, H. D. Evans, *Ind. Eng. Chem.* **43** (1951) no. 6, 1317 – 1324.
107. D. E. Dietrich, *Proceedings ICLASS 78* (1978) 69/77.
108. P. Walzel, *Chem. Ing. Tech.* **52** (1980) no. 12, 985, MS 857 – 880.
109. H. A. Kuhn, A. Lawley: *Powder Metallurgy Processing,* Academic Press, New York – San Francisco – London 1978.
110. J. K. Beddow: *The Production of Metal Powders by Atomisation,* Heyden & Son, London – Philadelphia – Rheine 1978.
111. G. Rai, E. Lavernia, N. J. Grant, *J. Met.* **37** (1985) no. 8, 22 – 26.
112. P. Walzel, *Chem. Ing. Tech.* **52** (1980) no. 4, 338 – 339.
113. R. Reske, *Fortschr. Ber. VDI Z. Reihe 7* (1987) Nr. 115.
114. P. Walzel, *Proceedings ICLASS 82* (1982) 187 – 194.
115. A. Reinhart, *Chem. Ing. Tech.* **36** (1964) no. 7, 740 – 746.
116. R. R. Hughes, E. R. Gilliland, *Chem. Eng. Prog.* **48** (1952) no. 10, 497 – 504.
117. K. A. Cliffe, D. A. Lever, *AIChE Symp. Ser.* **80** (1984) 236, 61 – 66.
118. J. A. Wesselingh, *Chem. Eng. Process* **21** (1987), 9 – 14.
119. P. Walzel: *Spray '96 (preprints)* Universität Bremen 1996, p. 1/5.
120. S. Middleman: *Modeling of Axisymetric Flow* Academic Press, San Diego1995.
121. J. F. Padday, A. R. Pitt, *Philos. Trans. Royal Soc. London, Ser. A* **275** (1972) 489 – 529.
122. M. Pilch, C. A. Erdmann, *Int. J. Multiphase Flow* **13** (1987) 741 – 757.
123. L. P. Hsiang, G. M. Faeth, *Int. J. Multiphase Flow* **18** (1992) 635 – 652.
124. A. Lefebre: *Atomization and Sprays* Hemisphere Publ., New York1989.
125. H. D. Dahl, VDI Fortschr. Ber. Reihe 3, no. 302; and Ph. Thesis University Stuttgart 1992.
126. M. Löffler-Mang, Düseninnenströmung, Tropfenentstehung und Tropfenausbreitung bei rücklaufgeregelten Drall-Druckzerstäuberdüsen, Ph. Thesis University Stuttgart 1992.
127. M. Horvay, Theoretische und experimentelle Untersuchung über den Einfluß des inneren Strömungsfeldes auf die Zerstäubbungseigenschaften von Drall-Druckzerstäuberdüsen, Ph. Thesis University Karlsruhe 1985.
128. T. Richter, P. Walzel, *Chem. Ing. Tech.* **61** (1989) no. 4, 319 – 321.
129. A. J. Yule, J. J. Chinn, ICLASS '94, Rouen 1994, preprints III 1, p. 334 – 228.
130. A. J. Yule, J. J. Chinn, ILASS '97, Florence 1997, preprints III 4, p. 223 – 341.
131. A. J. Yule, D. G. Salters, ICLASS '94, Rouen 1994, preprints II 2, p. 236 – 243.
132. S. Wilhelm, Tropfenbildung an Lamellen und Filmen, Ph. Thesis University Essen 1992; VDI Fortschrtsber. Reihe 3, no. 312.
133. T. Schröder, Tropfenbildung an Gerinneströmungen im Schwere- und Zentrifugalfeld, Ph. Thesis University Essen 1997.
134. K. Bauckhage, ILASS '97, Florence 1997, preprints III 5, p. 229 – 2351.
135. H. Wiggers, P. Walzel, *Chem. Ing. Tech.* **69** (1997) 1066 – 1073.
136. A. G. Bailey: *Electrostatic Spraying of Liquids,* Res. Stud. Press Ltd Taunton, Somerset 1988.
137. D. Michelson: *Electrostatic Atomization,* Adam Hilger, Bristol–New York 1990.
138. G. Schulte, ILASS '97, Florence 1997, preprints VI 5, p. 375 – 381.
139. C. Tropea et al., ILASS '97, Florence 1997, preprints VI 6, p. 382 – 388.
140. F. Schöne et al., *Chem. Ing. Tech.* **62** (1990) no. 12, 1058 – 1059.
141. P. Walzel, ILASS '97, Florence 1997, preprints VII 1, p. 407 – 413.
142. K. Baukhage, *Chem. Ing. Tech.* **64** (1992) no. 4, 322 – 332.
143. V. Uhlenwinkel, U. Fritschig, K. Baukhage, Proceedings 5th Internat. Conf. on Liquid Atomization and Spray Systems, ICLASS-Americas, Gaithersburgh, July 15 – 18, 1991.
144. K. Baukhage et al., Proceedings Powder Metallurgy Conf., London, July 2 – 6, 1990.
145. V. Uhlenwinkel, K. Baukhage, B. Seuren, *Chem. Ing. Tech.* **64** (1992) no. 1, 86 – 87.
146. Chr. Mundo, M. Sommerfeld, C. Tropea, *Int. J. Multiphase Flow* **21** (1995) no. 2, 151 – 173.

147. M. Rein, *fluid Dyn. Res.* **12** (1993) 61 – 93.
148. M. Marengo et al., ILASS '97, Florence 1997, preprints VIII 5, p. 465 – 472.
149. W. Samenfink et al., ILASS '97, Florence 1997, preprints VIII 6, p. 480 – 486.

Spunbonded Fabrics → Nonwoven Fabrics
Sputtering → Thin Films
Stains → Wood, Surface Treatment

Size Enlargement

See also Solids Technology, Introduction

Karl Sommer, Technische Universität München, Lehrstuhl für Maschinen- und Apparatekunde Weihenstephan, Freising-Weihenstephan, Federal Republic of Germany

1. Introduction 2
2. Binding Mechanisms of Agglomerates 4
2.1. Tensile Strength of Agglomerates Derived from Attractive Forces . 4
2.2. Attractive Forces in Size Enlargement in Air 5
2.2.1. Electrostatic Forces 6
2.2.2. Van der Waals Forces 6
2.2.3. Capillary Forces 7
2.2.4. Solid Bridges 7
2.2.5. Influence of Particle Surface Roughness on Binding 9
2.2.6. Strengthening of Binding Forces . . 10
2.3. Attractive Forces of Particles in a Liquid Medium 10
2.3.1. DLVO theory 10
2.3.2. Flocculation by Polymeric Flocculants 11
3. Characterization of Agglomerates 12
3.1. Particle Size, Particle Size Distribution 12
3.2. Agglomerate Density and Porosity 12
3.2.1. Definition 12
3.2.2. Densities 13
3.2.3. Porosities 13
3.3. Agglomerate Strength 15
3.4. Redispersion 15
4. Size Enlargement Processes 17
4.1. Growth Agglomeration (Pelletization) 17
4.1.1. Agglomeration Kinetics 17
4.1.2. Pan and Drum Agglomeration . . . 18
4.1.3. Mixer and Fluidized Bed Agglomeration 19
4.1.3.1. Pan and Drum Mixers 19
4.1.3.2. Horizontal Mixers 19
4.1.3.3. High-Speed Mixers 19
4.1.3.4. Fluidized-Bed Agglomeration . . . 19
4.2. Agglomerate Formation from Moist Material 20
4.2.1. Introduction 20
4.2.2. Agglomeration Equipment for Plastic Materials 21
4.3. Drying Processes 22
4.3.1. Atomization Drying 23
4.3.1.1. Fundamentals of Atomization Drying 23
4.3.1.2. Fluidized-Bed Spray Drying 24
4.3.1.3. Combination Processes 24
4.3.2. Contact Drying 25
4.3.3. Vacuum Drying 26
4.4. Spherical Agglomeration 27
4.5. Pressure Agglomeration 29
4.5.1. Compression Equations 30
4.5.2. Stamp Presses 31
4.5.3. Roller Presses 32
4.5.4. Tableting 34
4.5.4.1. Tableting Equipment 34
4.5.4.2. Tablet Pressing 35
4.5.4.3. Tableting Problems 36
4.5.4.4. Effect of Compression Rate and Humidity 36
5. References 38

1. Introduction

In the broadest sense of the term, size enlargement includes all processes in which fine particles, dispersed in either gas or liquid, aggregate to form a coarser product. The collection of particles that results is called an agglomerate or granule. Depending on the process, the size of the agglomerate is between 0.02 and 50 mm. In most cases, the preferred particle shape is spherical. In many processes, the product is a cylindrical section such as a tablet or some other regular geometrical shape. Size enlargement is used in many industries, such as fertilizer production, iron ore, nuclear fuel, pulverized fuel ash, ceramics, light-weight aggregate production, carbon blacks, catalysts, pesticides, and pharmaceutical products.

There are other reasons for the production and use of agglomerates:

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_07

1) Plant contamination and nuisance dusts in the workplace are reduced
2) Hygiene is easier to maintain
3) Air quality is easier to achieve, and in some controls can be avoided
4) The danger of dust explosions is reduced
5) With the increase in air and water pollution control regulations there is a need to install dust and sludge removal equipment. As a result, there is a need to process and dispose of the resulting fines. With agglomeration processes, perhaps with the addition of a binder, dust and sludge can be recycled or disposed of without pollution.

There are many problems in the handling of fine particles. The flow properties and consequently dosage control are poor. Size enlargement can eliminate these disadvantages while retaining the desired particulate properties. The chemical industry, with countless intermediate and final products, has a great need for size enlargement technology. Some examples are the agglomeration of simple or complex fertilizers, to avoid segregation; agglomeration of fillers used in plastics, to add exact amounts of iron oxide, zinc oxide, stearate, or silicate to the molten plastic; production of dust-free products, especially those with toxic or corrosive materials. Size enlargement increases the commercial worth because of the improved physical properties. A typical example is detergents. Some materials, such as fertilizers, pigments, pesticides, and instant foods, are sold in agglomerated form, but instantly decompose when added to liquids [1], [2]. Many raw materials are only suitable for processing or end use as coarse particles. These coarse solids have the desired strengths and porosities, for example, catalysts.

Ore beneficiation produces a large number of fine-grained products. Examples are flotation concentrates from selective beneficiation processes, dusty products from roasting, blast furnace by-products, filter cakes, and particulates from dust removal equipment. Ores or ore mixtures must be agglomerated for smelting to obtain optimal flow conditions in the roasting process. The principal area of application in the ceramic industry is in the preparation of pressed articles made of barium titanate and manganese – zinc, nickel – zinc and barium ferrites [3]. These materials are initially fine powders, which must be agglomerated to obtain good dimensional replication in the pressed articles.

In addition, size enlargement can cause a delayed action, which is important for pharmaceutical and agricultural applications. This allows good dosage or improves the appearance of the product.

Table 1. Agglomeration processes

Growth Agglomeration (Agitation Methods). Fine particles are brought into contact with each other in a flowing system or in air when the concentration is higher. This is usually done in the presence of liquid and binders. The particle size enlargement occurs by coalescence or accretion (snowballing) based on capillary forces. In a few exceptional cases, the major cohesive force is the van der Waals force. Usually the agglomerates are spherical with diameters between 0.5 and 20 mm. Typical equipment types are inclined drums, cones, pans, paddle mixers, and plowshare mixers. The maximum throughput is about 100 – 200 t/h for iron ore pellets and 50 t/h for fertilizers.
Spray Agglomeration (Spray Methods). This is one of the most commonly used methods in the chemical, pharmaceutical, and food industries. Pumpable suspensions are atomized, and the liquid is evaporated from the droplets by means of hot air, as a preliminary drying step. The first cohesive forces are the capillary forces, which are followed by crystal bridges at the contact points. The agglomerates are 20 – 500 µm. For chemical applications, throughputs of up to 50 t/h are possible.
Selective Agglomeration (Spherical Agglomeration). A second immiscible phase is added to the suspension. This wets the solid phase and binds the particles together by means of capillary forces. As a result, rounded flocs or agglomerates form with diameters up to 5 mm. Selective agglomeration can be achieved for mixtures of solids. In the case of coal suspensions, throughputs up to and exceeding several tonnes can be achieved.
Pressure Agglomeration (Pressure Methods, Compaction). Particles with only slight amounts of moisture are formed in tablets and briquettes in stamp presses, tablet presses, and roller presses. The principal binding force is van der Waals attraction. The agglomerates have uniform shapes and range in size from a few millimeters (pharmaceutical tablets) to decimeter size (fuels). In the case of smooth rollers, the resulting flakes are broken up into the desired size. The throughput for ores is roughly 100 t/h; for chemicals, up to 30 t/h.
Sintering (Thermal Methods). Fine particles are made into a paste by adding moisture and then processed in a horizontal sintering oven into sinter (burnt agglomerates). This is especially common in the mining and preparative industries. The final sintered product usually has an irregular shape and is usually coarser than other agglomerates. The mechanism of binding is the formation of solid bridges at the contact points. Sintering plants have throughputs up to 1000 t/h.

The product properties desired determine which enlargement process is used. Size enlargement processes are classified by the principal mechanism by which the particles are made to come together [4], [5] (Table 1). The selection of a specific process is only possible if the user clearly defines the properties required of the product. Special attention must be paid to the following points: state of the material, desired agglomerate size, size distribution, and agglom-

erate shape, required strength and porosity, and required throughput.

State of the Material. The starting material may be fluid, pumpable, sprayable, pasty, or dry. The product may be dryable. Although temperature sensitive, it may have to be dried. Frequently the cost of drying determines the economic feasibility of the process [6].

Desired Agglomerate Size, Size Distribution, and Agglomerate Shape. Many processes permit only a narrow range of agglomerate size. Spray agglomeration gives the finest products; briquetting and sintering, the coarsest. In the case of pressure agglomeration with smooth rollers, the resulting flakes must be broken up. The cost of this size reduction and the recycling of the resultant fines must be considered. The shape of the agglomerate may be of critical importance to the process. Some examples are flow through a bed of solid catalyst, in the case of bunkering and for storage.

Required Strength and Porosity of the Agglomerate. Many end uses require a large specific surface area and ready dispersibility (e.g., instant food products). Examples are pigments, agricultural sprays, food, and pharmaceuticals. Other products must be dispersible but with a delay time. Catalyst and iron ore pellets, however, must withstand high stresses without being crushed. The type of binding material (and thus binding force) has a major effect on strength and porosity.

Required Equipment Throughput. Some size enlargement processes, for example, tableting, are not suitable for producing high tonnages.

Other Factors. Amont other factors are compressibility drying behavior (including possible cracking), and high- and low-temperature behavior. For ore pellets, there are properties important in smelting, reducibility, swelling, and shrinking [7].

Once an appropriate process has been selected, laboratory tests must be run to determine the fluid dosage that gives a suitable moisture level for agglomeration. Laboratory tests must be run to determine applicability and dosage of agglomeration aids such as binders and wetting agents (wetting rate). After the laboratory tests, a pilot plant must be set up, and the process checked on a larger scale. At this point the production process, involving recycling, can be tested, allowing prediction of the technical and economic feasibility of the process on an industrial scale [8], [9].

2. Binding Mechanisms of Agglomerates

The attractive forces between particles cause the size enlargement process and are essential to the strength and dispersibility of the agglomerate. Rumpf [10] has given a complete summary of binding mechanisms. Schubert [11] considers the presence or absence of material bridges as the overriding criterion for binding mechanisms (Fig. 1). Material bridges can be further categorized as solid or liquid. Solid bridges include those formed by sintering, chemical reaction at contact points, hardening of binders, and the crystallization of dissolved materials. Fluid bridges are formed by capillary forces between the particle and a fluid or high viscosity binder. In the case where there are no material bridges van der Waals forces are the principal binding force. In general, for similar materials with the same surface ions, electrostatic forces are repulsive.

Mechanical interlocking is rarely a mechanism for agglomeration, for example, the balling up of "burrs."

Attractive forces can be calculated for model systems, such as smooth, fixed, ideal spheres. Even though the values calculated for these models apply only crudely for real systems, they do indicate the effects of important parameters on the agglomeration process. Models are essential for the understanding of the size enlargement process. To calculate the actual attractive forces between real particles is not possible because of their irregular shape and usually rough surfaces [11].

2.1. Tensile Strength of Agglomerates Derived from Attractive Forces

Rumpf [12] presents the following equation, based on a statistical, geometric approach:

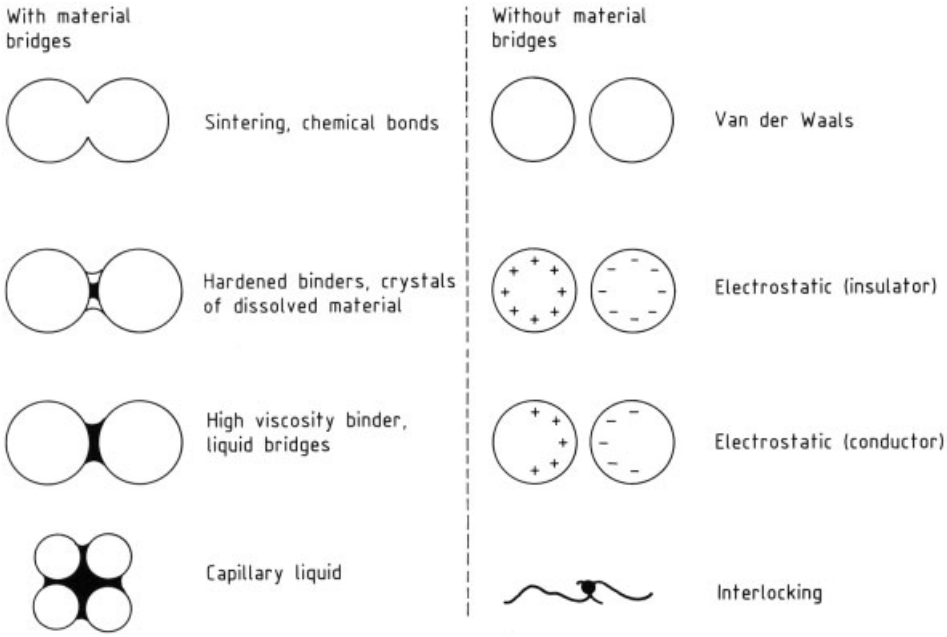

Figure 1. Schematic representation of major binding mechanisms in agglomerates [11]

$$\sigma_z = \frac{1-\varepsilon}{\pi x^2} \cdot k \cdot F \tag{1}$$

where σ_z is the tensile strength of randomly packed, equally sized spheres with diameter x; ε is the porosity. F is the attractive force at the point of contact, and k is the coordination number.

RUMPF [13] has derived a more general equation for the nonspherical particle-size distribution using the approximation of SMITH et al. [14] for the coordination number ($\varepsilon \cdot k \approx \pi$)

$$\sigma_z = \frac{1-\varepsilon}{\varepsilon} \cdot \frac{\bar{F}}{x_s^2} \tag{2}$$

where $\bar{F}$ is the average attractive force at the contact points and x_s is the equivalent diameter. Equation (2) can be experimentally confirmed [15] for moist agglomerates with force transmission via fluid bridges. In the case of short range forces, such as van der Waals forces, surface roughness at the contacts causes problems. In most cases of this type, the strength is measurably less than that calculated using Equation (2) [15]. Even though the van der Waals forces can only be calculated for model systems (Section 2.2.2.). Equation (2) does give an indication of the effect of various parameters.

The strength of fluid-saturated agglomerates can be calculated by using the capillary pressure curve [16]. The primary force in fluid-filled capillaries is the capillary pressure p_k. A negative pressure, it is equal to the tensile stress transmitted to the pores. The tensile strength is given by the following equation only for the case when the degree of saturation $S \approx 0.8$ (Section 2.2.3.), and the stress is transmitted to the pores.

$$\sigma_z = S \cdot p_k \tag{3}$$

2.2. Attractive Forces in Size Enlargement in Air

Agglomeration with mixers, drums, and pans; fluidized bed agglomeration; compaction; and tableting. The dominant force acting on particles larger than 1 cm is gravity. As the size decreases, the effect of gravity rapidly decreases – in proportion to the third power of the particle diameter – and attractive forces become more important. For example, for a 1 µm particle the van der Waals forces are larger than

the weight by about six orders of magnitude. Figure 2 shows glass spheres about 10 µm in diameter, which adhere to the surface of a glass fiber, solely by van der Waals attraction.

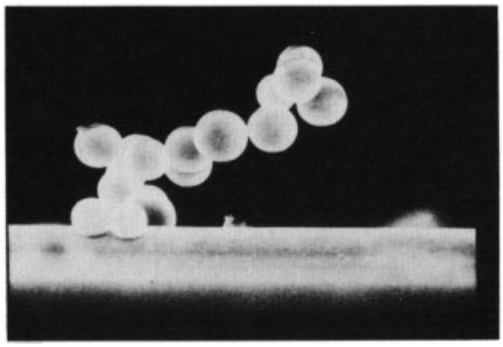

Figure 2. Glass spheres on a glass fiber from Polke [17]

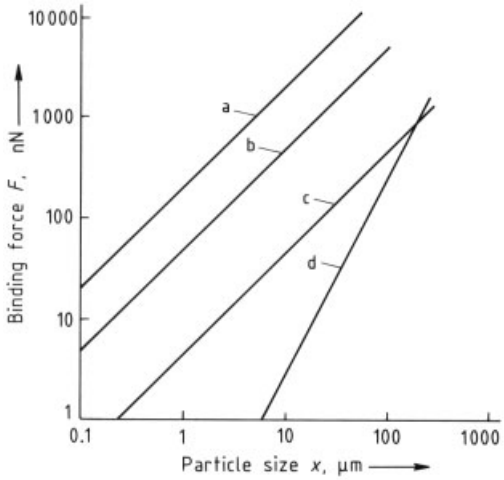

Figure 3. Theoretical binding forces derived from different mechanisms calculated for two identical ideal spheres [18]
Separation $a = 0$, $z_{el} = z_0 = 4 \cdot 10^{-8}$ cm
a) Liquid bridges $V_L/2\,V_S = 10^{-4}$; b) van der Waals $\hbar\omega = 5$ eV; c) Conductor $U = 0.5$ V, $\varepsilon_0 = 8.86 \times 10^{-12}$ As/Vm; d) Insulator $\sigma_1 = \sigma_2 = 100\,e$ µm^{-2}

2.2.1. Electrostatic Forces [18]

The attractive force F_{el} between two charged surfaces is derived from the energy of the two electrostatic fields. The force has been calculated by Krupp [19] as well as Moser [20] for the case of two electrically conducting spheres (Figs. 3 and 4).

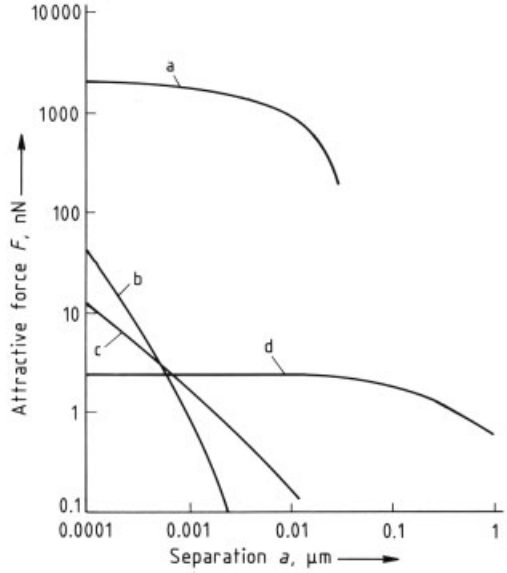

Figure 4. Theoretical attractive forces derived from various mechanisms calculated for two equal-sized spheres as a function of distance between surfaces
Particle size 10 µm
a) Liquid bridges $V_L/2\,V_s = 10^{-4}$, $\gamma = 0.072$ Nm; b) van der Waals $\hbar\omega = 5$ eV; c) Conductor $U = 0.5$ V; $\varepsilon_0 = 8.86 \times 10^{-12}$ As/Vm; d) Insulator $\sigma_1 = \sigma_2 = 100\,e$ µm^{-2}

$$F_{el} = \frac{1}{4}\varepsilon \cdot \varepsilon_0 \cdot \pi \frac{U^2 \cdot x}{a + z_{oel}} \tag{4}$$

ε is the relative dielectric constant of the medium. In the case of air ε is 1; x is the diameter of the sphere, and a is the distance between the surfaces, U is the potential difference, z_{oel} is a fitting parameter of the order of magnitude of an atomic dimension, for example, 4×10^{-8} cm.

For nonconductors the attractive forces are noticeably smaller than for conductors. In nonconductors the charge is not concentrated on the surface, but can extend below the surface to a depth exceeding 1 µm. Two spheres with surface charges σ_1 and σ_2 exert an attractive force

$$F_{el} = \frac{\pi}{4 \cdot \varepsilon \cdot \varepsilon_0} \cdot \frac{\sigma_1 \cdot \sigma_2 \cdot x^2}{\left(1 + \frac{a + z_{oel}}{x}\right)^2} \tag{5}$$

In the case of ideal nonconductors in contact ($a = 0$), the quantity z_{oel}/x is small compared to 1 and can be neglected.

2.2.2. Van der Waals Forces [21]

There are two principal methods of calculating the van der Waals attraction. In the micro-

scopic theory the attractive energy is calculated as the sum of the interaction energy of two atoms [22]. The assumption is made that the interactive forces are additive and do not influence each other. Another, more physically satisfactory theory starts with the optical and electrical properties of the two interacting particles and calculates the van der Waals force from them [23].

KRUPP [19] gives the attraction for the case of two spheres in which the distance a is less than 500 Å.

$$F = \frac{\hbar\bar{\omega}}{32\cdot\pi}\cdot\frac{x}{(a+z_0)^2} \tag{6}$$

where $\hbar\bar{\omega}$ is the Lifschitz – van der Waals constant. Depending on the material, the value is 0.2 – 9 eV and must be determined experimentally. It can be calculated in a few cases [19] (see Figs. 3 and 4).

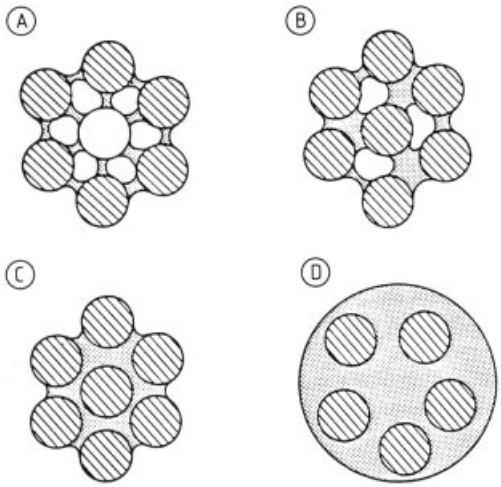

Figure 5. Agglomerate binding mechanisms due to mobile liquids [24]
A) Pendular; B) Funicular; C) Capillary; D) Droplet

2.2.3. Capillary Forces

Because of surface tension, there is an attractive force between wetted particles, which can hold agglomerates together. NEWITT and CONWAY-JONES [24] have classified binding mechanisms based on mobile liquids into four types: pendular, funicular, capillary, and droplet (Fig. 5). Liquid bridging is the term used to describe the case in which the fluid droplets are concentrated at the contact points and are separated from each other [25]. Figure 6 shows a liquid bridge between two spherical particles. The simplifying assumption is made that the boundary of the bridge is a circular arc. For the case of wetted particles in contact, PIETSCH [26] gives the following equations:

$$F = \gamma\cdot x\cdot\pi\cdot\sin^2\beta\cdot\left[1+\frac{x}{4}\left(\frac{1}{R_1}-\frac{1}{R_2}\right)\right] \tag{7}$$

$$R_1 = \frac{x\,(1-\cos\beta)+a}{2\cos\beta}$$

$$R_2 = \frac{x}{2}\sin\beta + R_1\,(\sin\beta - 1) \tag{8}$$

where γ is the surface tension, β is one-half the central angle (see Fig. 6), and x is the diameter (see Figs. 3 and 4). SCHUBERT [27], [28] has calculated the appropriate differential equations exactly corresponding to the attractive forces transmitted by rotationally symmetric fluid bridges. The numerical results with the parameters separation distance, bridge volume, and contact angle are presented in chart form [29]. The maximum divergence of these values is less than approximately 20 %.

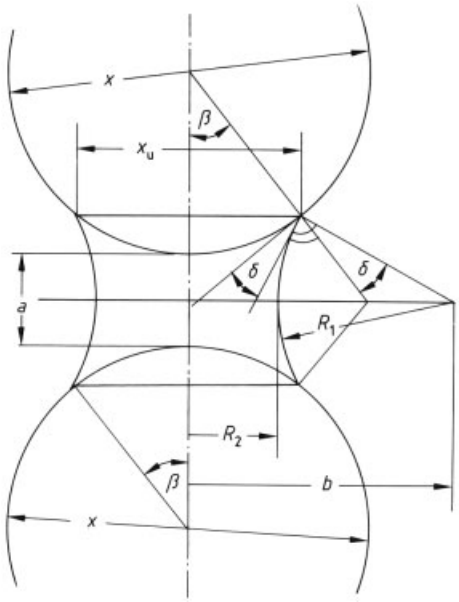

Figure 6. Liquid bridges between two equal-sized spherical particles

2.2.4. Solid Bridges

In many cases, the attractive forces due to mobile liquids are only an aid in forming an agglomerate. They are crucial to the process of formation,

but are not often important for the physical properties of the final product. These properties are produced by binding processes that follow agglomerate formation. In this step, solid bridges are formed by pressure or thermal treatment. The major mechanisms are as follows:

1) Crystallization of dissolved material
2) Hardening of fluid binders
3) Local melting and coalescence processes at the contact points
4) Particle deformation in combination with sintering and chemical reactions

Salt Bridges. If wet agglomerates were only held together by surface forces and the capillary pressure, the cohesive force would disappear when the liquid was evaporated. To give the dry agglomerates a definite amount of strength, dissolved salts are often used. These crystallize out during drying and form salt bridges at the contact points between the particles. The strength of such an agglomerate can be estimated to an order of magnitude if the total volume of the material in the salt bridges V_{salt} is known [31].

$$\sigma_{\mathrm{z}} = \frac{V_{\mathrm{salt}}}{V_{\mathrm{solid}}} \cdot (1-\varepsilon) \cdot \sigma_{\mathrm{s}} \tag{9}$$

$V_{\mathrm{salt}}/V_{\mathrm{solid}}$ is the volume ratio of the crystallized salt to the total volume of the particles (V_{solid}), and σ_{s} is the tensile strength of the dissolved salt. In this idealized case, there is only a small amount of water, primarily at the contact points (pendular state). Alternatively, the assumption is made that the solubility of the salt is so high that it remains completely dissolved in the water at the contact points during the drying process [32], [33].

Binders. The mechanical properties that can be obtained by using binders, which harden and increase in strength with time, are similar to those obtainable with salt bridging. However, the strength can be better controlled. Capes [34] has shown that the addition of a binder, for example, corn starch, decreases crust formation during drying, leading to a more even distribution of the solid forming bridges, noticeably increasing the strength (Fig. 7). Spherical agglomerates that contain both, a binder and salt crystals, are more resistant to breakage under pressure than agglomerates that have only one of the two.

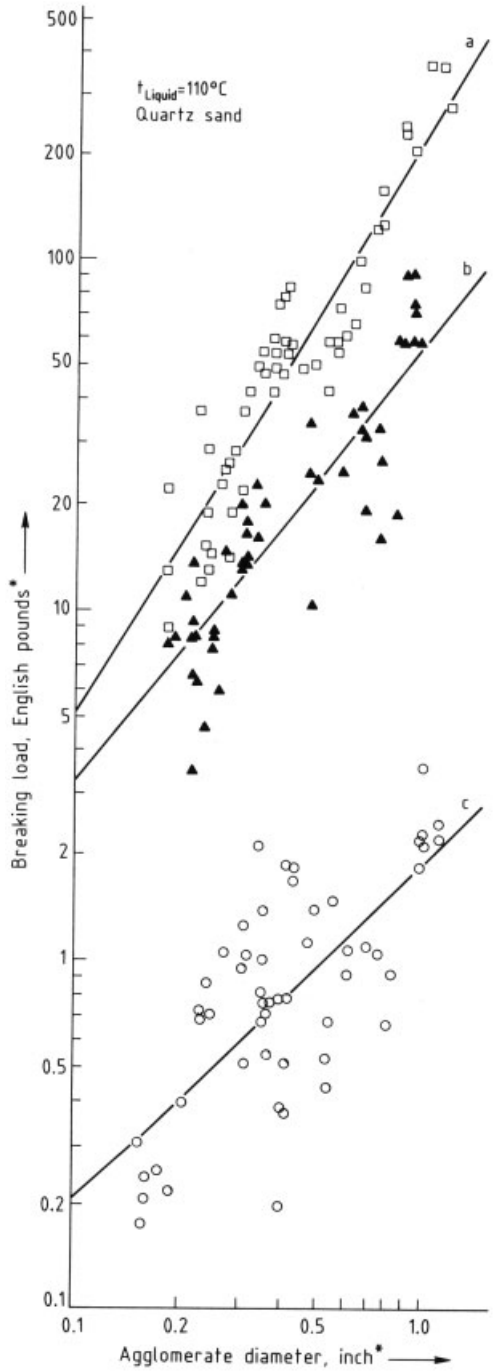

Figure 7. Breaking load of spherical agglomerates with NaCl salt bridges and cornstarch as binder from Capes [34]
∗ To convert English pound into kilograms multiply by 0.4536; to convert inch into millimeter multiply by 25.4
a) NaCl solution and cornstarch; b) Distilled water and cornstarch; c) NaCl solution

Sinter Bridges. Sinter bridges can form in the contact area between particles when the temperature exceeds roughly 60 % of the melting temperature of the material that makes up the particles. The strength of the sinter bridges is of a similar order of magnitude as the strength of the material which makes up the particles. For this reason, the maximum attractive force present depends primarily on the diameter $2b$ of the neck of the sinter bridge. During the sintering process, the sintered area grows with time t.

Only the initial growth stage of a sinter bridge is of interest for agglomeration. Complete sintering, down to pore-free material, which is important in metallurgy, is not considered here.

Research on sintering began in 1949 with the work of Frenkel [35] on the sintering of viscous materials. In this case the kinetics are determined solely by the reduction of surface free energy and viscous dissipation energy. At almost the same time, Pines [36] published an article on the sintering of solids, which showed that the actual molecular diffusion mechanism could be derived from a mathematical treatment of the diffusion of vacancies in the crystal lattice. The growth of the sinter neck was explained by the diffusive migration of vacancies from concave points of contact to convex parts of the particle's surface. In a similar way, atoms migrate from the particle surface to the sinter neck. This results in diffusion both in the bulk material (volume diffusion) and on the surface (surface diffusion). There is also gas-phase material transport because of the difference in equilibrium vapor pressures near concave surfaces and convex surfaces [36–41], [72]. An exact description of all the physical processes that occur in the various stages of sintering is extremely difficult because of the complex geometry and many factors that affect the surface, such as adsorbed and oxide layers.

2.2.5. Influence of Particle Surface Roughness on Binding

The idealized-spheres model cannot completely describe what actually happens. In actuality, surfaces are rough, and this can have a major effect on the attractive forces. The effect of surface roughness can be visualized by means of the following model. A semicircular peak with radius r sits on the surface of a sphere with diameter x and is in contact with another sphere. As a first approximation, the electrostatic and van der Waals forces of the individual sections can be assumed to be superimposable.

Electrostatic Forces (Fig. 8) (with roughness $r \ll x$).

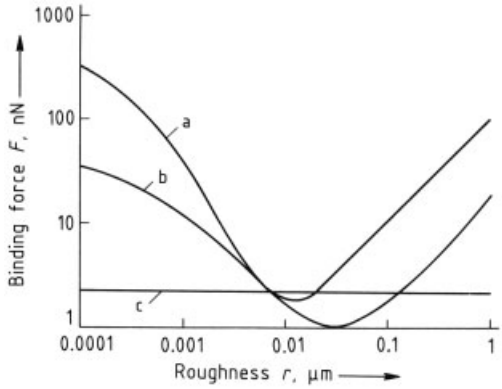

Figure 8. Effect of a hemispherical roughness on the surface of a sphere in the contact region on the binding force
Particle size 10 µm
a) van der Waals $\hbar\omega = 5$ eV; b) Conductor $U = 0.5$ V; c) Insulator $\sigma = 100\,e$ µm^{-2}

Electrical Conductor

$$F_{\mathrm{el}} \approx \frac{1}{4}\pi \cdot \varepsilon \cdot \varepsilon_0 U^2 \left[\frac{x}{r + z_{\mathrm{oel}}} + \frac{4 \cdot r}{z_{\mathrm{oel}}}\right] \tag{10}$$

Nonconductor

$$F_{\mathrm{el}} \approx \frac{1}{4}\pi \cdot \frac{\sigma^2 \cdot x^2}{\varepsilon \cdot \varepsilon_0} \tag{11}$$

Van der Waals Forces (see Fig. 8)

$$F = \frac{[plank]\,\bar{\omega}}{32\pi}\left[\frac{x}{(r+z_0)^2} + \frac{2r}{z_0^2}\right] \tag{12}$$

Capillary Forces. For capillary forces, the effect of roughness is closely related to the size of the bridges. If the fluid bridge is so small that only the projection is wetted, and not the surface of the rest of the particle, then the appropriate model is the attraction of two contacting spheres with different diameters (x and $2r$) [16, p. 150]. If the fluid bridge is so large that the projection is completely covered, so that it acts merely as a spacer ($a = r$), then the appropriate model is for two large spheres with diameter x [16, p. 139].

For a given particle size, the van der Waals force passes through a minimum, which is roughly 300 times smaller than for an idealized, smooth sphere. A similar result applies for the electrostatic attractive force between conductors. However, the attractive force between insulators is largely independent of roughness. Even though van der Waals forces are an order of magnitude greater than electrostatic forces in the smooth sphere model, electrostatic forces can be dominant when roughness is present. Like for van der Waals forces, the attractive force in the case of small fluid bridges is almost exclusively determined by the size and shape of the roughness projections. The fact that these are mostly unknown explains the difficulty in trying to measure the attractive forces between real particles. In comparison to ideal spheres, results in real systems commonly deviate by orders of magnitude.

2.2.6. Strengthening of Binding Forces

Deformation in the contact area, caused by external or internal forces, can affect the binding of real particles, just as roughness can. Strengthening of the binding forces means an increase relative to the value calculated for a model system with no particle deformation [38]. Strengthening of binding forces can also occur directly without any external agent. This may be brought about by the binding forces of the undeformed system alone, or for example, the dead weight of a pile of agglomerates. In addition, external force can cause a rearrangement or further deformation in the contact areas. This is especially the case for tableting and compacting. The mechanisms for strengthening of binding forces are classified as follows [44]:

1) Rearrangement
2) Elastic deformation at the contact point
3) Plastic deformation at the contact point
4) Viscoelastic flattening
5) Sintering

The driving forces (or energies) are:

1) The binding force itself, deformation or rearrangement increasing it relative to the binding force of the undeformed system F_0
2) An external force F_{C0}, the sum of the binding force F_0 and the external force F_{C0} being the total force

$$F_t = F_0 + F_{C0} \tag{13}$$

3) The interfacial energy

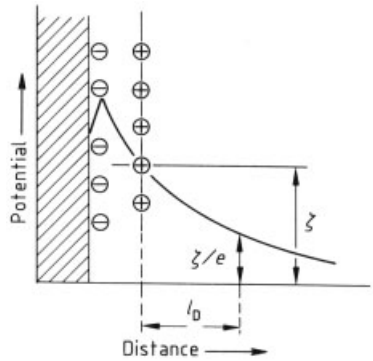

Figure 9. Potential function of the electrical double layer around a particle

2.3. Attractive Forces of Particles in a Liquid Medium

In principle, the attractive forces between particles in liquids and in gases are based on the same mechanisms. This similarity is especially true in regard to the bonding mechanisms of the material bridges if the changes with time because of solution processes and corrosion in liquids are considered and if the surface tension for particles in gases is replaced for particles in liquids by the interfacial tension when considering capillary forces. However, the interaction of the van der Waals forces and the repulsive electrostatic forces, which are caused by absorbed ions, give rise to anomalies in liquid media.

2.3.1. DLVO theory

For particles suspended in a fluid, the medium acts as a dielectric that has an effect on the dipole – dipole interaction between the particles. As a result the van der Waals attraction is modified. Equivalent to Equation (6), Equation 14 expresses the van der Waals potential V_{vdw} as the product of a constant related to the type of material and a geometric term [22].

$$V_{vdw} = -\frac{A_{PlP}}{12} \cdot \frac{x}{a + z_0} \tag{14}$$

An approximate relationship for the Hamaker – van der Waals constant A_{PlP} between the particle P and the liquid l has been given by **Derjagin** and **Landau** [45] and **Vervey** and **Overbeck** [46].

$$A_{\mathrm{PlP}} \approx \left(\sqrt{A_{\mathrm{PP}}} - \sqrt{A_{\mathrm{ll}}} \right)^2 \tag{15}$$

The resultant Hamaker – van der Waals constant increases as the difference between the Hamaker – van der Waals constants for the particle A_{PP} and the liquid A_{ll} increases. The Lifschitz constants for the usual macroscopic theory for the corresponding dielectric spectra have a similar relationship.

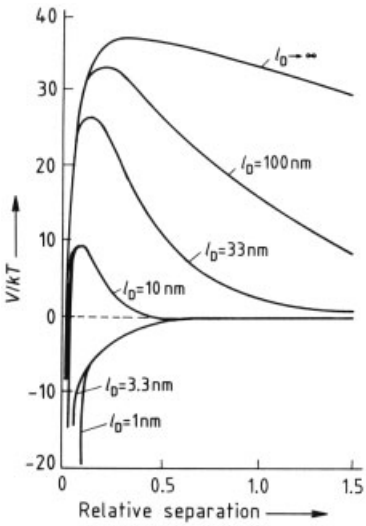

Figure 10. Interactive potential function according to DLVO theory [45], [46]

The source of the electrostatic charge in a liquid medium is a layer of tightly adsorbed ions on the particle surface (Stern layer). This layer causes an electrostatic repulsion. Over the tightly adsorbed ionic layer there is a diffuse layer of counter ions (Gouy – Chapman layer) as shown in Figure 9. A characteristic value for the thickness of the double layer is the Debye length l_{D} over which the potential decreases by a factor of $1/e$. The potential V_{el} of the electrostatic repulsion increases with the number of adsorbed ions and the resulting increase of the Stern layer potential V_0. The combination of the van der Waals attraction and the electrostatic repulsion determines the interaction potential as a function of the interparticle distance (Fig. 10). A potential barrier prevents particle agglomeration when the electrostatic charge is high enough. While the van der Waals forces cannot be varied much in a given system (although one possibility is perhaps a suitable adsorbed layer) the electrostatic repulsion can be greatly decreased, for example, by the addition of electrolytes. In this case the van der Waals forces dominate, and the particles agglomerate.

2.3.2. Flocculation by Polymeric Flocculants

The use of synthetic polymers for the agglomeration in suspensions is steadily gaining in industrial importance as replacements for inorganic electrolytes. Synthetic polyelectrolytes are used in a large number of diverse applications. Polymeric flocculants are classified by their charge. Polymers with a negative charge are termed anionic, those with a positive charge are termed cationic. There is also a group of nonionic water-soluble polymers such as poly(ethylene oxide), poly(vinyl alcohol), and polyacrylamide.

Two mechanisms have been proposed [47] for the action of polymers in promoting interparticle attraction [47]. The macromolecule also affects the electrostatic repulsion, which in turn allows aggregation of the particles due to van der Waals forces (Section 2.3.1). The polymer must have a charge opposite to that of the surface to be adsorbed in a flat configuration and neutralize the surface charge. If there is an excess of polymer, long range electrostatic attraction of other particles can occur.

The second mechanisms is based on bridging. In this case, only a few segments of the polymer are attached to the surface, while the rest project out into the liquid. The distance they extend must exceed the range of the repulsive forces. This type of binding can only take place at high solids levels or when the particles are sufficiently close. **Pusch** has used the scanning electron microscope to show that polymer bridges of considerable length, as well as flat layers covering the surface and cross-links can occur on the particle surface, depending on the polymer concentration [47]. The polymer molecules do not appear to act as individuals, but are themselves agglomerated. These aggregates are somewhat flattened and are very elastic. They fill the interparticle regions, forming thin threads, bridges or configurations resembling pendular water bridges.

Because of their porosity, the strength of flocs produced with the aid of flocculants is quite low. The shear forces encountered in flow through a pipe, nozzle, or pump can break the aggregates up into the original particles (Fig. 11). However, polymer particles show a considerable tendency to reagglomerate.

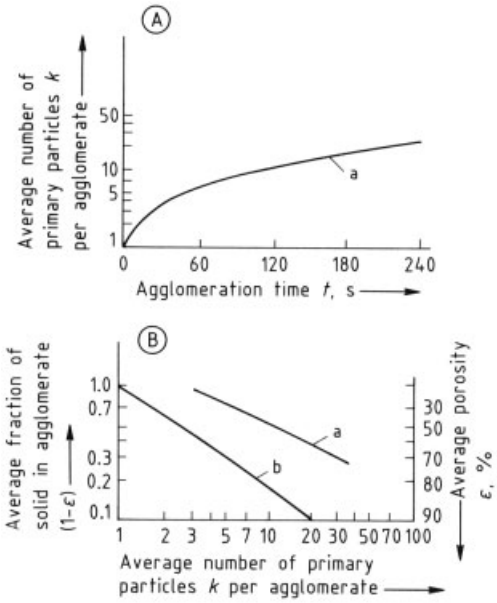

Figure 11. Effect of gravity on polymer induced agglomerates [47]
A) a) Effect of shear force on agglomerates of fine solids; B) a) Quartz; b) 100 % cationic polymer

3. Characterization of Agglomerates

An understanding of the properties of agglomerates is of importance for the various steps encountered during further processing as well as the application by the end-user:

1) Process optimization
2) Quality control
3) Production of specific properties

3.1. Particle Size, Particle Size Distribution [49]

Agglomerates can be characterized by their size, shape, and size distribution, as is usually done with finely dispersed materials. For most applications the optimal shape is spherical because this is the shape least subject to breakage and abrasion. Cylindrical shapes (pellets) are also common. In the optimal case, the size distribution is narrow and the fines fraction small. Size and size distribution can be determined by a large number of techniques [48]. For practical applications, a quantitative size criterion that is relevant to the end use should be found. Therefore, great care must be taken in selecting a suitable measurement process. Even when a suitable process has been established, serious problems in sampling and sample preparation can be encountered.

Up to the present, mechanical screening is the most common analytical method because it covers a broad size range. In addition to an analytical method, screening may be part of the process. In this case, the product stream is divided into three fractions: fines, product, and oversize. Care must be taken that screening does not result in excessive abrasion or even agglomerate breakage.

Even though agglomerate size is recognized as one of the most important parameters, particle size analysis is seldom used in industry for process control. The principal reason for this is that product application during additional processing as well as by the customer depends on other parameters. Therefore, an application related test is often a better standard of quality [50]. The pigment industry offers one example. The color in the standard formulation is required as a definitive test of the quality of pigment agglomerates.

3.2. Agglomerate Density and Porosity

3.2.1. Definition

Porosity is defined as the ratio of the void volume V_v to the total volume V_{tot}.

$$\varepsilon = V_v / V_{tot} \tag{16}$$

Density is defined as the ratio of the mass m to the volume V of a particular material.

$$\varrho = m / V \tag{17}$$

For porous particles, agglomerates and piles of agglomerates, the difficulty in defining porosity and density lies in the fact that there are various methods for determining volume.

Figure 12 illustrates the different kinds of pores. Pores may be either accessible or inaccessible. Some pores have a constant diameter, while others tend to narrow down and are accessible through narrow capillaries. Other pores are open enough to permit flow through the material. Surface roughness also has to be considered.

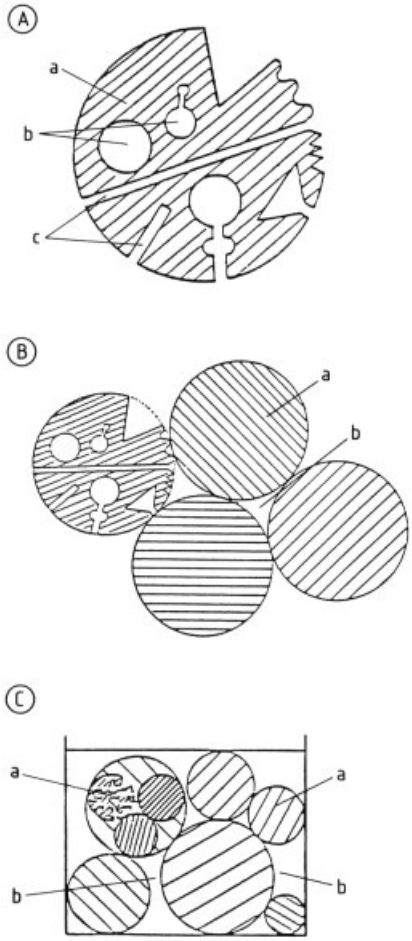

Figure 12. Structure of packed agglomerates composed of individual particles
A) Particle: a) Solid; b) Inaccessible pores; c) Accessible pores
B) Agglomerate: a) Particle; b) Interstitial volume
C) Bulk agglomerates: a) Agglomerate; b) Interagglomerate volume

3.2.2. Densities

Solid Density ϱ_S. The solid density is defined as the density of the pure solid. This value determines the sedimentation rate of solids without any included void space, i.e., in wetting fluids. The density of such solids is measured with a pycnometer. If the solid may contain inaccessible void space this must be made accessible by grinding.

$$\varrho_S = \frac{m_S}{V_S} \tag{18}$$

Apparent Solid Density ϱ'_S. The apparent solid density is defined for solids that contain inaccessible void space V_{cP}. This value determines the sedimentation rate of such solids in a wetting fluid, under the assumption that the void volume does not become accessible by some dissolution process. The apparent density is determined with a pycnometer directly, i.e., without grinding.

$$\varrho'_S = m_S / (V_S + V_{cP}) \tag{19}$$

Particle Density ϱ_P. Both accessible and inaccessible pores are taken into account in the particle density, which determines the sedimentation rate of solids in nonwetting fluids.

$$\varrho_P = m_S / V_P \tag{20}$$

Agglomerate Density ϱ_A. Agglomerate density controls the sedimentation rate of agglomerates in a nonwetting fluid.

$$\varrho_A = m_S / V_A \tag{21}$$

Bulk Density ϱ_B. A large group of particles or agglomerates is characterized by the bulk density, which is determined by measuring the total mass of bulk material contained in a given volume.

$$\varrho_B = m_S / V_B \tag{22}$$

3.2.3. Porosities

Particle Porosity ε_P. The ratio of the pore volume (accessible V_{oP} and inaccessible V_{cP}) to the total particle volume V_P is the particle porosity.

$$\varepsilon_P = \frac{V_{oP} + V_{cP}}{V_P} \tag{23}$$

If the inaccessible pore volume cannot be measured, the apparent particle porosity is used, instead of the particle porosity.

$$\varepsilon'_P = V_{oP}/V_{cP} \tag{24}$$

Agglomerate Porosity ε_A. The ratio of the volume of the void space between the particles V_{iP} and the volume of the agglomerate V_A is the agglomerate porosity. The agglomerate porosity allows the difference in sedimentation rate of the same particles in wetting and nonwetting fluids to be determined. The rate of sedimentation is always greater in wetting fluids. The agglomerate porosity is the ratio of the particle density to the agglomerate density.

$$\varepsilon_A = V_{iP}/V_A \tag{25}$$

$$\varepsilon'_A = 1 - \varrho_A/\varrho'_S \tag{26}$$

where ε'_A is the apparent agglomerate porosity.

Bulk Porosity ε_B. In a group of agglomerates, there is an interstitial volume V_{iA} between the agglomerates themselves. From this value the bulk porosity ε_B is defined as follows:

$$\varepsilon_B = V_{iA}/V_B \tag{27}$$

The apparent bulk porosity is defined as:

$$\varepsilon'_B = 1 - \varrho_B/\varrho'_s \tag{28}$$

Total Porosity ε_{tot}. The total porosity of the system must be distinguished from the bulk porosity.

$$\varepsilon_{tot} = \frac{V_{oP} + V_{cP} + V_{iP} + V_{iA}}{V_B} \tag{29}$$

If the inaccessible pore volume is not included in the measurement, then the quantity calculated is the apparent total porosity ε'_{tot}.

$$\varepsilon'_{tot} = \frac{V_{oP} + V_{iP} + V_{iA}}{V_B} \tag{30}$$

The different porosity values are related as follows:

$$(1 - \varepsilon_{tot}) = (1 - \varepsilon_P) \cdot (1 - \varepsilon_A) \cdot (1 - \varepsilon_B) \tag{31}$$

$$\varepsilon'_{tot} = 1 - \varrho_B/\varrho'_S \tag{32}$$

In addition, that pore sizes are different must be considered. Figure 13 illustrates the pore radius size distribution function for a bulk agglomerate system. In general, pore sizes are distinguished clearly for three systems: individual particles, agglomerates, and bulk systems. This is the result, in ideal cases, of three different maxima in the distribution function [51–55].

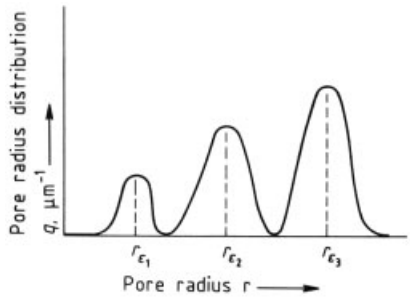

Figure 13. Idealized pore size distribution [51]

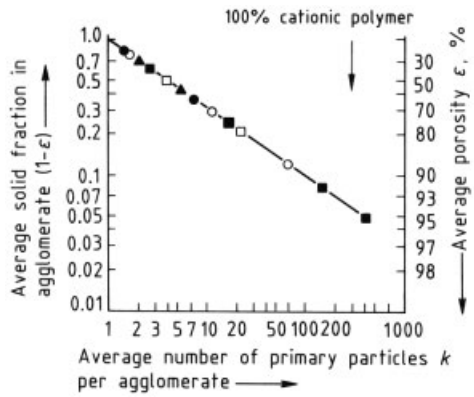

Figure 14. Structure of agglomerates in suspension [58]

Type of solid	x, µm
● Glass spheres	16
▲ Quartz	14.5
■ Coal	8.5
○ Cornstarch	13
□ Cellulose	14

Because of the limited strength of flocs, the determination of the size and porosity of flocs formed by agglomeration in liquids is especially difficult. None of the usual methods can be used in this case. In general, another measurable parameter is used, one which depends on both the size and porosity. At Technische Hochschule Karlsruhe, a method has been devised to determine the average size and porosity of agglomerates in undisturbed suspensions. Simultaneous measurements are made with a photometer and a sedimentation balance [56], [57]. Figure 14 shows the average floc density during flocculation as a function of particle size. The process is characterized by the increase in the number of particles per agglomerate [58].

3.3. Agglomerate Strength

In general, the term strength refers to all of the properties of the agglomerate that act against whatever stresses the agglomerate may encounter: abrasion, fracturing, compression, or impact. The breakability that is desired for the end use of the product determines the upper limit of the strength that the agglomerate may have. At the same time, the agglomerate should have as high a strength as possible to prevent breakage during handling. Because the handling stresses are of the same order of magnitude as the breakability, the permissible stress range is narrow.

The measurement of agglomerate strength is divided into two categories: agglomerates smaller than 5 mm, and agglomerates whose diameter is greater than 5 mm. A number of scientific and industrial tests are used for the larger agglomerates [59–61]. Figure 15 illustrates some of the usual methods that are used in the pharmaceutical industry: the diametral compression test, the bending test, the drop test, and the abrasion test. There are numerous types of commercial apparatus that can be used to determine the strength of a pressed product with these tests. The strength values obtained on different machines are not usually comparable with each other. Some reasons for this are differences in geometry and the rate at which the stress is applied, as well as differences in external factors, handling, and calibration.

The strength of smaller agglomerates can be determined by the rate of deagglomeration in a specific shear field [60] or with a dispersion apparatus specific to the industrial application [61–65]. This could mean the ratio of optically determined surface areas after various types of dispersion processes, or it could mean the relative values of the desired property, such as the color intensity or average agglomerate size after two dispersants have been used.

Pneumatic transfer over a specific distance under simulated industrial conditions can be used as a measure of agglomerate abrasion.

3.4. Redispersion

When fine particulate properties are required for processing or application, as with pigments, the agglomerates must be easily broken down mechanically. If the agglomerate is dissolved during use, the required redispersibility also depends on the wettability. Good wetting and rapid dissolution are desirable for instant products and pharmaceuticals, while slow dissolution is desired for fertilizers.

Today, a large number of foods, such as coffee, cocoa, and cereal are sold as instant products. Often this property can only be obtained by means of agglomeration. Different manufacturers may define this property in different ways, therefore a method to measure it quantitatively is required. A quantitative method of measuring redispersion rate is necessary for industrial powders that are added to and dissolved in liquids. The reconstitution process takes places in the following steps:

1) Fluid soaks into the bulk material and the individual agglomerates
2) Agglomerates sink in the liquid
3) Agglomerates break up
4) Particles dissolve in the liquid

None of these steps can be clearly isolated from the others. They do not happen sequentially, but rather occur simultaneously, affecting each other. Therefore, determination of the individual properties separately is difficult.

Penetration of the Liquid [66–70]. The first step in the reconstitution process is the penetration of the bulk material by the liquid. This results from capillary fluid transport through the

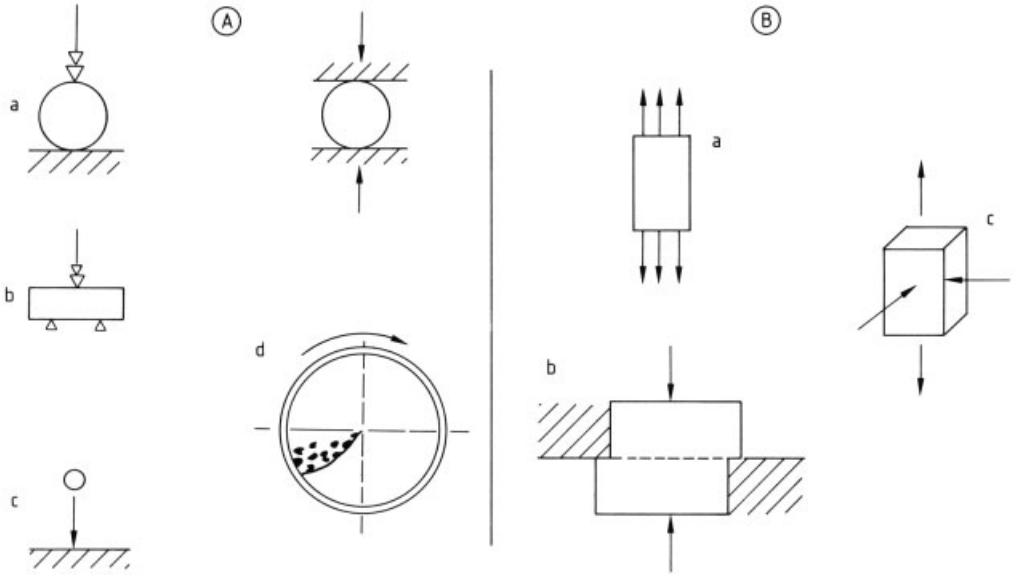

Figure 15. Operation of strength measurement techniques that simulate actual industrial conditions
A) Industrial tests: a) Compressive strength; b) Bending test; c) Drop test; d) Abrasion test
B) Deformation stress: a) Tensile strength; b) Shear strength; c) Shearing strength

void spaces in the bulk material. If the contact angle $\delta = 0$, wetting is rapid and complete. For most powdered foods, δ is greater than zero so that liquid penetration of the bulk is slower. Agglomeration of these powdery foods reduces the time required for the liquid to penetrate the bulk [50].

Sinking. If the wetting angle of a powder is 0°, then the powder can sink only if its apparent density is greater than the density of the liquid or the material dissolves or disperses in the liquid. If the powder dissolves, inaccessible void spaces are opened up and the apparent density increases with time. A powder with a wetting angle of greater than 0° requires a higher apparent density to sink than one with an angle of 0°.

Dispersion [71–73]. Dispersion means the breakup of the agglomerates into their primary particles. This process should be as rapid as possible. Dispersion greatly increases the solid – liquid contact area, and consequently the rate of dissolution. To obtain rapid agglomerate breakup, the bridges between the particles should be weak. However, weak bridges may allow more abrasion, which would tend to slow down liquid penetration. PFALZER [68] therefore recommends having a large number of weak interparticle bridges. The weak bridges would be easily dissolved, while the large number of bridges would provide the necessary mechanical stability. SCHUBERT [71] seeks an optimum in promoting the desired, antithetical properties of rapid breakup and low abrasion. As strength increases, the liquid penetration time decreases because of the reduced amount of abrasion. However, the time required for the agglomerate to break up increases because of the stronger interparticle bridges. There is a relative minimum in the overall time and a corresponding optimal agglomerate strength. The optimal strength must be determined for various porosity levels. The optimal porosity is the one that gives the shortest overall reconstitution time.

So-called burst agents are added to highly compacted pharmaceutical preparations. These agents are materials that swell to considerable degree when they absorb water. Some typical release agents are starch and modified starches. The action of a burst agent is only apparent when the swollen dimensions of the burst agent exceed the pore size of the material in which it is incorporated. There are no literature references to the application of burst agents in granulated products. Any improvement in dispersibil-

ity from such an application cannot be predicted with certainty. The effect of such a burst agent is influenced by its concentration, its swelling properties, the particle size distribution, and the agent's distribution in the instant product.

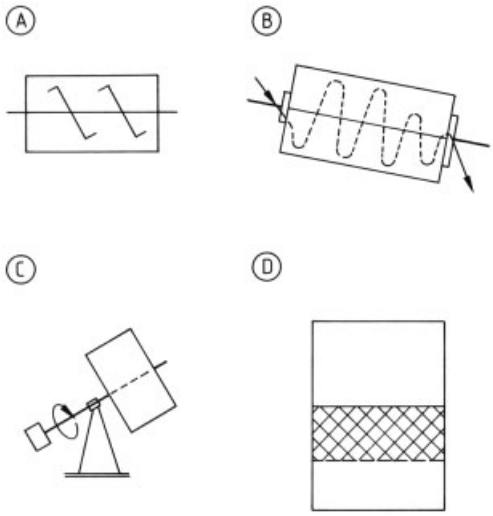

Figure 16. Schematic representation of equipment used for growth agglomeration
A) Mixer; B) Drum granulator; C) Pan granulator; D) Fluidized bed granulator

4. Size Enlargement Processes

4.1. Growth Agglomeration (Pelletization)

Growth agglomeration is defined as the growth of more or less solid agglomerates in either of two environments. The first is a rotating apparatus that produces both a mixing and a rolling motion. The second is a turbulently agitated suspension of particles that generates interparticle collisions. There is a stable accumulation if the attractive forces in the industrial equipment always are greater than the destructive forces present in the system. The binding forces in this type of agglomerate are primarily capillary forces. Binders are often added to the liquid used to form the agglomerates. The strength of the final product is usually a function of the processing that occurs after agglomeration, such as drying and sintering. The most common types of equipment for growth agglomeration are pans, drums, mixers, and fluidized beds (Fig. 16). Growth agglomeration can result from several mechanistic processes: coalescence of fine particles (starting materials and recycled fines), i.e., the nucleation phase, followed by growth through accumulation of further starting materials and the debris from other particles. The course of these processes with time is described by agglomeration kinetics.

4.1.1. Agglomeration Kinetics

Capes and Danckwarts [74] and Kapur and Fürstenau [75], [76] have developed kinetic models for particle growth. If certain assumptions are made for the probability of agglomeration, changes in particle size distribution for a batchwise operated drum can be predicted theoretically. Capes and Danckwarts use the following assumptions in their model of particle growth:

1) The smallest size fraction in the feed material will be abraded
2) The abraded material from this process will be distributed among the remaining agglomerates according to their size
3) The rate of growth of the diameter of each agglomerate is proportional to the difference between its diameter and the diameter of the smallest agglomerate present in the system.

Then they obtained a dimensionless form of the agglomerate size distribution, which is not explicitly time dependent, and which they called "self-preserving."

Kapur and Fürstenau determine the change in an agglomerate size distribution as a function of time by means of a coalescence model. This approach is similar to that used by Smoluchowski [77] in his droplet coalescence theory. Kapur and Fürstenau make the following assumptions:

1) Mixing of all the fractions is completely homogeneous
2) Particle size distribution is relatively narrow
3) Coalescence rate is independent of granule size
4) Probability of collision between the granules is independent of their size

This model leads to the same results as Capes and Danckwarts, the so-called "self-preserving" distribution. These kinetic models have the disadvantage that it is difficult or impossible to apply them to continuous systems. Over 90 % of growth agglomeration is performed in continuously operated mixers, drums, and pans. The agglomeration model can be modified by dividing the process into two stages. The first is the aggregation stage, in which the particles coalesce. The second stage is called the snowballing stage (Fig. 17), in which agglomeration proceeds by layerwise deposition. By suitable selection of particle size classes, a system of differential equations can be written which describe the change in particle size distribution as a function of time [78].

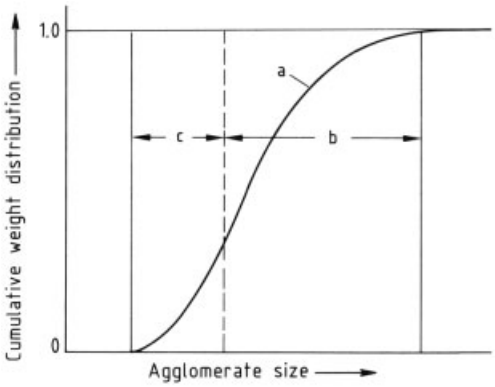

Figure 17. Division of the agglomerate size distribution in the aggregation and snowballing zones
a) Agglomerate size distribution in agglomerator; b) Snowballing zone; c) Aggregation zone

4.1.2. Pan and Drum Agglomeration [79–85]

Growth granulation using a pan or drum granulator converts finely dispersed material to a coarser product. The fine feed (diameter < 1000 µm) enter the disk rotating on an inclined axis or the rotating trommel, along with some added liquid. They are exposed to a rolling action (Fig. 18). Because the rolling action is primarily random, spherical agglomerates form, with diameters up to 50 mm.

In the pan agglomerator, the agglomerator nuclei and small agglomerates move toward the base of the disk. Because of increased friction, due to their largely irregular shape, the particles are transported higher. The larger, more rounded agglomerates roll easily over the smaller ones. Ultimately, at a certain size they are removed. In this way the pan classifies the agglomerates. The overflow product is so uniform in size that subsequent classification is usually unnecessary. Classifying pans producing granules having a narrow size range may have diameters up to 7.5 m and have iron-ore-pellet capacities greater than 100 t/h.

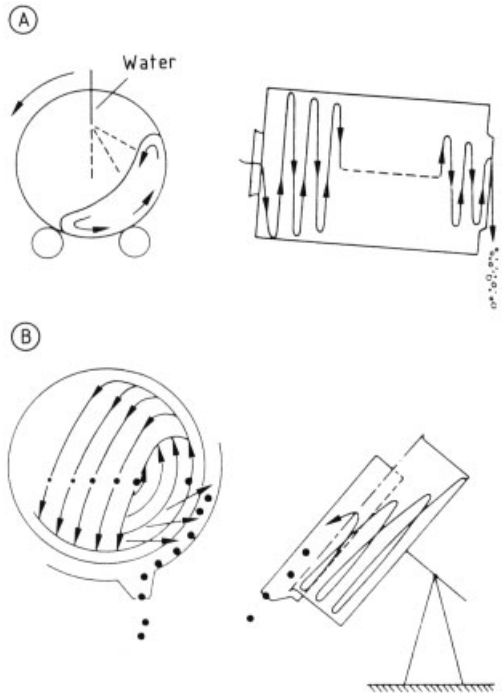

Figure 18. Schematic diagram of A) pan and B) drum agglomeration

In the agglomeration drum, the inclination of the axis is too slight to cause classification. It serves primarily to transport the feed through the cylindrical drum. The agglomerate size distribution is substantially broader than that obtained with a pan. To obtain a specific agglomerate diameter, the desired size agglomerates must be screened out of the exit stream. The oversize material must be reduced in size and fed back into the drum along with the fines to form a closed loop. The principal users of agglomeration drums are the iron and fertilizer industries. Sizes range up to 10 m in length and 3.5 m in diameter.

4.1.3. Mixer and Fluidized Bed Agglomeration

In theory, all solid and solid – liquid mixers are suitable for agglomerate production. In contrast to pan and drum agglomerators, the powder is rolled about in mixers by means of a mechanical impeller or, in the case of a fluidized bed, pneumatically. When a suitable amount of agglomeration liquid is sprayed in, most of the particles form either solid agglomerates or a low-density instant product due to the stresses generated by stirring.

4.1.3.1. Pan and Drum Mixers

Horizontal pan and vertical drum mixers have been constructed with capacities up to $5\,m^3$. Their first application was in the agglomeration of fertilizers [86]. A typical example is the Eirich countercurrent mixer-granulator. This device has a rotating mixing drum supported by a vertical or slightly tilted axle. A mixing impeller rotates in the opposite direction. When used for fertilizers, throughputs of up to 30 t/h, with residence times of 2 – 3 min, can be obtained. If the conventional impeller is replaced with a specially designed impeller, intensive mixing action can be obtained, even permitting the agglomeration of molten materials while they are cooling. Drum granulation of a similar type is widely used for pulpy, plastic, or pasty starting materials in the pharmaceutical industry. In some cases, dry material is added to produce small agglomerates as feed for tableting. The agglomerates produced by pan or drum mixers are usually stable, with high strength, after drying.

4.1.3.2. Horizontal Mixers [87]

The mixing action necessary for agglomeration is provided in the case of the horizontal mixer by a horizontal mixing shaft. This shaft may be equipped with paddles, screw strips, or special mixing blades that resemble plowshares. A typical example is the Lödige plowshare mixer. The rolling action is therefore more gentle than that in a pan or vertical drum mixer. As a result, the agglomerates are usually more porous and have a lower strength. Horizontal mixers can be operated either continuously or batchwise.

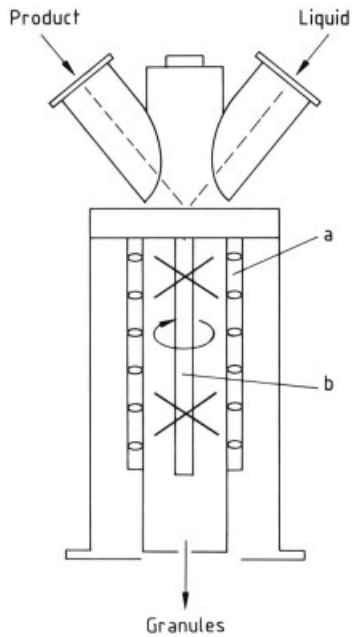

Figure 19. Schugi mixer
a) Roller cage; b) Rotor with blades

4.1.3.3. High-Speed Mixers

High-speed mixers are continuous mixers with impellers operating at relatively high speeds. The finished product has a brief residence time in the unit, only a few seconds. Porous, easily dispersible agglomerates are formed with the aid of sprayed in liquid. The product has a size range of 0.5 – 2 mm and finds application in the food industry, especially as instant products (Fig. 19).

4.1.3.4. Fluidized-Bed Agglomeration [90–92]

In this process, pumpable solutions, suspensions, pastes, or melts are converted into agglomerates mostly in combination with drying (see Section 4.3.1.3.). Fluidized bed drying technology has been known for several decades. Even though agglomerate formation was observed early in this process, the process was not used for agglomeration until 1970. The first use was the production of feed material for tablet production in the pharmaceutical industry [88].

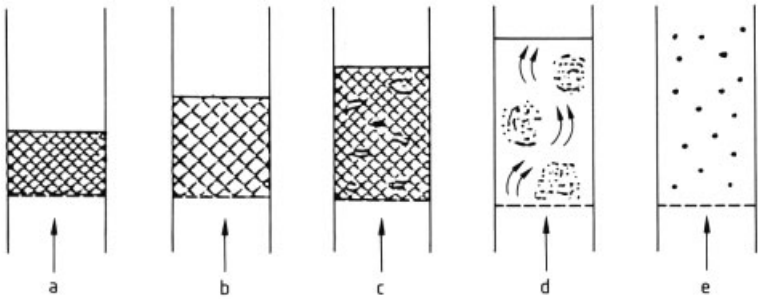

Figure 20. Schematic diagram of fluidized bed formation
a) Solid bed; b) Homogeneous fluidized bed; c) Inhomogeneous fluidized bed; d) Transition zone; e) Pneumatic transport

The range of application has grown from pharmaceuticals to the chemical, ceramics, and food industries and is steadily growing in importance. In fluidized bed agglomeration, the feed particles are moved about by flowing gas (Fig. 20). Externally heated gas (air) is forced or drawn upward through a porous distributor plate and then through the particle bed. A heterogeneous solid – gas fluidized bed is formed. The solid particles are intensively mixed by the rising gas bubbles[89]. This promotes good heat and material exchange. The agglomeration fluid is sprayed with either single- or two-fluid nozzles, which are located above the bed surface or project just below it. The motion of the bed serves to distribute the fluid uniformly among the particles. A portion of the agglomerates in the bed are drawn through an opening beneath the bed surface or through the overflow port. The product, in most cases, is separated into oversize, product, and fines. The fines and crushed oversize are recycled to the fluidized bed. To produce round agglomerates in the fluidized bed, the fixed perforated plates are replaced by rotating disks. The gas flow is either through perforations in the outside of the disk, or through a slot between the rotating disks or a slot between the rotating disk and the stationary housing. This process is called rotor agglomeration (Fig. 21). Many manufacturers offer fluidized bed equipment with a rectalinear cross section. These offer nearly the same possibilities of variation with regard to construction and operation as does equipment with a round cross section.

4.2. Agglomerate Formation from Moist Material

4.2.1. Introduction

The production of fine, dry powders entails high drying costs, which are greatly increased by the addition of fluid during agglomeration. Economic considerations (energy savings up to 50 %) would favor combining the drying and granulation process segments. This would prevent having to granulate a dry powder or enable the granulation to be done before drying. The goals of this process are the following:

1) Processing of pasty products without repulping to reduce drying energy
2) Production of a dust-free agglomerate
3) Production of the desired size of agglomerate having the required product characteristics

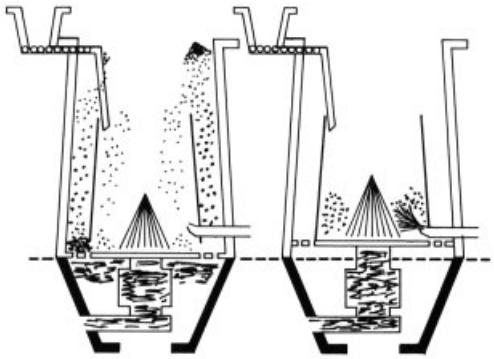

Figure 21. Schematic representation of rotor agglomeration

One possible way of achieving the first goal lies in combining the granulation process with fluidized bed drying. In conventional fluidized bed drying (see Section 4.3.1.3.), dust from the filter is remixed with the bed material, producing a dusty product. A better method is to mix the dust from the filters into the moist feed so that it is incorporated into the agglomerate (Fig. 22).

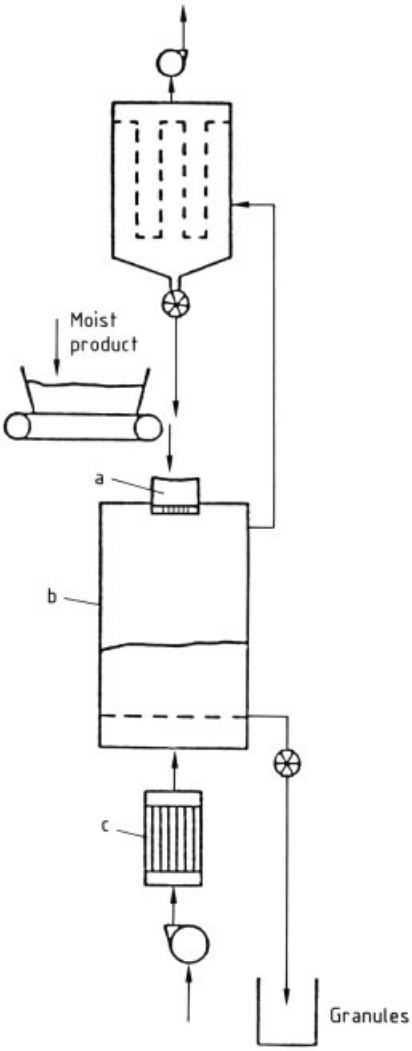

Figure 22. Agglomeration prior to drying
a) Screen granulator; b) Fluidized bed dryer; c) Air heater

4.2.2. Agglomeration Equipment for Plastic Materials

There are a large number of different machines for this purpose. They differ in their construction, the configuration of the kneading surfaces, and their kinematics. There are more than 50 types commercially available (Fig. 23).

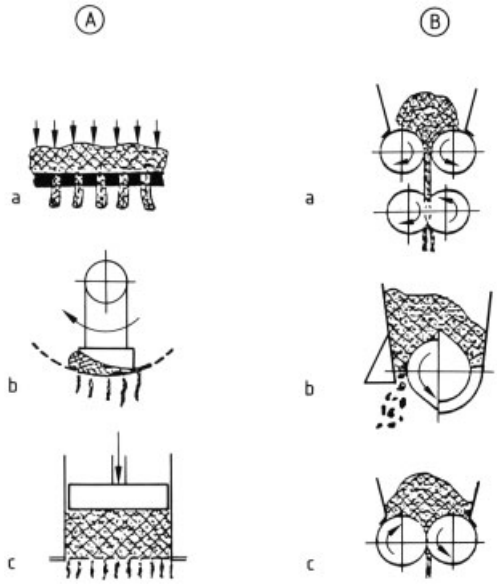

Figure 23. Operation of machines for agglomeration of moist materials
A) Extrusion in pellet mills: a) Application of pressure; b) Frictional press; c) Pressure by volume reduction
B) Breaking of clumps: a) Cutting of sheets; b) Crushing or milling; c) Squeezing into molded products

Extrusion through Screens and Perforated Plates. Pasty products, when pressure is applied, are forced to flow through the provided openings. Pressure can be applied by the following physical processes:

Internal Friction (*Friction Formation*). The amount of pressure is greatly dependent on the properties of the product. The maximum pressure is ca. 50 kPa, which is relatively low. The product is strongly sheared by the frictional forces.

Volume Reduction (*Shape Formation*). The amount of pressure that can be applied is only dependent on geometric factors. Pressures up to 2.0 MPa can be generated. In this case there is little shearing.

Movement of Excluded Volume (*Shape Formation*). High pressure can be generated in the space allowed. The product must be capable of flowing.

Breaking up of Product Clumps. The following methods are used for pasty products:

Cutting. Thin cakes, produced by rollers, are cut by the action of a knife roller. The method is also applicable to harder clumps.

Scalping, breaking. A rapidly rotating knife roller either peels off or breaks off by inertia, pieces of the product stream. The shape of the pieces cannot be well controlled. The method is not suitable for soft pastes.

Squeezing. The product is squeezed into the grooves of a grooved roller made of an elastic material. The shaped material is released when the grooves are retracted. This method is suitable for soft pastes.

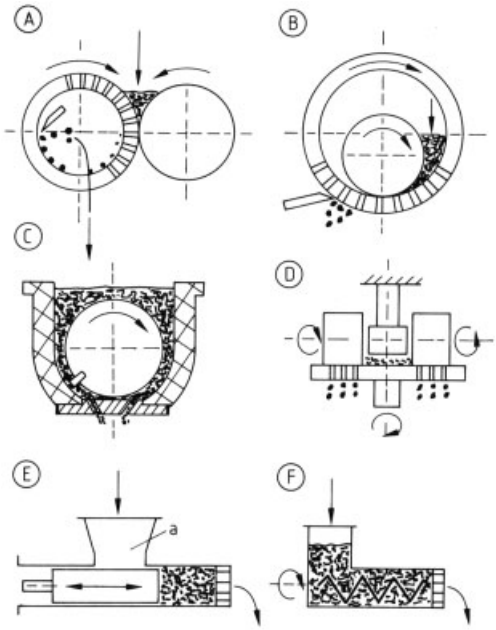

Figure 24. Operation of pellet mills
A) Roll pellet mill; B) Ring collar mill; C) Haas-granulator; D) Disk collar mill; E) Extrusion mill; F) Screw mill

Pellet mills are used to shape pasty materials such as filter cakes and others. The most important types of machines are illustrated in Figure 24. In all cases, the material is forced through openings and in so doing is compressed more or less into granules, also called pellets. In the roll pellet mill, the material is introduced into the space between a perforated roller and a solid one. It is compressed and forced through cylindrical openings. Inside the perforated roller, the granules are cut off by a knife and flow outward in an axial direction. A variation of this type of machine has two perforated rollers.

In the *ring collar mill,* the material is introduced inside and forced out through the openings in the die ring by the collar. The granules are then cut off on the outer side by a knife. This type of machine can have various sized collars as well as different numbers of collars. The force exerted in a ring collar mill can be up to 15 t [93]. Through puts up to 15 t/h for feed materials can be achieved with 22 mm openings.

The so-called *Haas granulator* is commonly used before belt drying. In this machine, the in and out motion of a roller equipped with a ridge forces the material through a perforated plate into a collection trough. In the *disk collar mill* the collar as well as the perforated plate move. In the extrusion and *screw mills* the perforated disk is stationary.

The swivel press has a rectilinear body and is open on the top. The bottom is closed off with a removable perforated plate. It is also equipped with a swivel arm. The feed is distributed across the perforated plate by a conveyor. Due to the swivel motion of the displacement arm, a triangular space bounded by the wall, the perforated plate, and the displacement member is enclosed. A squeezing or load stroke forces the material through a perforated plate. Because the plates are removable, various geometries can be selected. If the product is highly resistant to extrusion through the plate, it may be squeezed out sideways. It is drawn in again on the return stroke. After the pressure stroke, the swivel arm quickly moves back. The return angle for the no load and the pressure strokes is adjustable. If there is a foreign body in the paste, the machine stops after the overload alarm sounds. Because of the low height, it can easily be removed by hand. The following parameters are involved in the construction and operation of the machine; geometry of the perforated plate (number of openings and their configuration), number of strokes, extrusion rate, return angle, extrusion pressure, and throughput [93–95].

4.3. Drying Processes

In the case of fluid raw materials, agglomeration can take place during drying, i.e., drying and ag-

glomeration are combined into one process step. The available methods are classified as follows:

1) Atomization drying
2) Contact drying
3) Vacuum drying

Atomization drying includes tower spray drying [98], fluidized bed spray drying, and multistage spray drying. Contact drying includes roller and belt drying. Under vacuum drying, only freeze drying and contact vacuum drying are described.

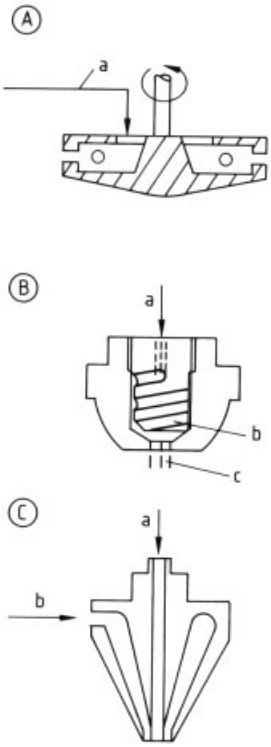

Figure 25. Types of atomizers
A) Centrifugal atomization disk: a) Moist product (not under pressure)
B) One-fluid nozzle: a) Moist product (under pressure up to 10 MPa); b) Mouthpiece; c) Nozzle opening
C) Two-fluid nozzle: a) Moist product (not under pressure); b) Air for atomization

4.3.1. Atomization Drying

Atomization drying processes are important for industry. A number of arrangements are used. The principal methods are classified as follows:

1) Simple spray drying
2) Fluidized bed spray drying
3) Spray drying with an integrated or downstream fluidized bed

The feed to atomization processes are pumpable products, such as solutions, suspensions, dispersions, and pumpable pastes. These are converted into dry powder by spraying them into a hot gas stream.

4.3.1.1. Fundamentals of Atomization Drying

Four different types of equipment are used to achieve atomization (Fig. 25):

1) Ultrasonic atomizer
2) Centrifugal atomizer
3) Two-fluid nozzle
4) One-fluid nozzle (pressure nozzle)

The use of the ultrasonic atomizer in drying is limited to special cases.

In centrifugal atomization, a wet product is accelerated to velocities between 100 and 200 m/s. A thin, fluid film enters the hot gas stream of the dryer through an opening in the atomization disk. It then breaks up into droplets. The resulting particle size is independent of throughput over a wide range. The average particle can be varied by changing the rotational speed.

A narrow particle size distribution can be obtained with a one-fluid nozzle. The average drop size is controlled by the diameter of the nozzle opening, the atomization pressure, the viscosity, the solids content, and the throughput requirements. One-fluid nozzles tend to clog easily.

In a two-fluid nozzle, the energy required for atomization is provided by the air stream. In most cases the air stream is compressed air, although in special applications steam or inert gas may be used. The wet product is introduced into the two-fluid nozzle under little or no pressure. Energy requirements are higher for the two-fluid nozzle than for the one-fluid type. With two-fluid nozzles, fine grained dry products can be obtained, although the particle size distribution is relatively broad. Compared to one-fluid nozzles, two-fluid nozzles present fewer handling and control problems. Special designs, which lessen the risk of clogging [96], are used for fluidized beds.

The average particle size has a major effect on the instant properties of a product. The drying process is slower for larger particles compared to smaller ones. This has an effect on the design of a drying tower. The longer residence time worsens the solubility properties. Temperature control is important for the properties of a spray-dried product. In the case of sensitive products, such as powdered milk, too high a temperature can damage the product. BRUMMELHUIS [97] has shown the effect of the temperature of the air leaving the dryer on various product properties using as the example the atomization drying of whole milk concentrate. Lower dryer exhaust temperatures improve the instant properties. The apparent density increases. The free fat content and the solubility index decreases. However, there is an increase in the moisture content of the powder, so that a second drying step is required.

4.3.1.2. Fluidized-Bed Spray Drying

If gas flows through a solid, either wet or dry, resting on a porous platform, the piled material begins to loosen. At a certain velocity it begins to whirl about. Vibrators can be used to aid fluidization. The pumpable product is sprayed into the receiver on top of the material already there. Either one- or two-fluid nozzles are installed over the pile or brought into it. In addition, stirrers and disintegrators are used to loosen up the product and to break up the larger clumps.

The separation of solid particles can be done externally by cyclones, filters, or wet scrubbers or internally by an integrated filter. The use of an integrated filter has the advantage that the product falls directly into the receiver. The risk of having moisture condense in the exhaust is minimized because of the short pathway. However, moisture condensing in the filter can cause stickiness and fluctuations in flow rate. Integrated filters are available with automatic shakers or continuously acting blow-off devices.

The units can be operated either in suction or pressure mode. Suction operation is clearly predominant among the major manufacturers because no dust escapes under low-pressure conditions. Also, suction spray towers can be operated without any material escaping. The dry product can be removed by means of belts, rotary bin values, overflow, or exit pipes. The product can subsequently be classified by screening or sintered, and the undersize material can be recycled. If necessary, the oversize material can be broken up and recycled. Product removal can be combined with classification. This is done with an ascending tube, which discharges at the screen. The tube is used as a sorter. By regulating the flow velocity in the tube, the size of the particles that are removed can be controlled.

Plants using only fluidized beds are mainly used to dry solutions or melts of organic acids, organic or inorganic salts, tanning agents, and dyes. Binders can be added to the solution or melt to improve cohesion. Exhaust temperatures are lower than in spray tower drying. The dry product comes in contact with the high-temperature, incoming gas. This is similar to the situation in countercurrent spray tower drying. As a result, however, the same disadvantages are present.

4.3.1.3. Combination Processes

During the last ten years, spray towers with downstream or integrated fluidized bed dryers have become more common. This is due to their lower energy usage, the primarily dust-free product, and the reduced thermal stress. The disadvantage of these types of processes is their high capital costs. A spray drying process with a downstream fluidized-bed drying stage (Fig. 26) is called a two-stage drying process. Lower residual product moisture with less harm to product quality as well as more efficient energy utilization can be achieved. The solids leave the drying tower with 5 – 8 % residual moisture. The final drying step takes place under mild conditions with low energy usage. Two-stage drying gives unsatisfactory results when the product with high residual moisture has a tendency to stick or form hard clumps. In these cases it may be possible to dry the product to a lower moisture in the drying tower. An agglomeration fluid is then sprayed in, either directly on the wet product or into the downstream fluidized bed (Fig. 27). As an alternative to having a fluidized bed dryer follow a tower drying stage, one can be integrated into a drying tower. In this case, rather than removing the powder from the spray tower, it is allowed to fall, with its high moisture

content, directly into a fluidized bed. In the fluidized bed zone, drying takes place in a cloud of fine particles. This powdering permits the drying of sticky products with a high moisture content without their sticking to the walls (Fig. 28).

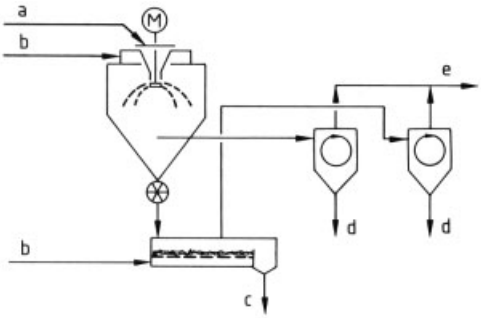

Figure 26. Spray dryers with downstream fluidized bed dryers
a) Moist product; b) Drying gas; c) Dry product, coarse fraction; d) Dry product, fine fraction; e) Exhaust gas

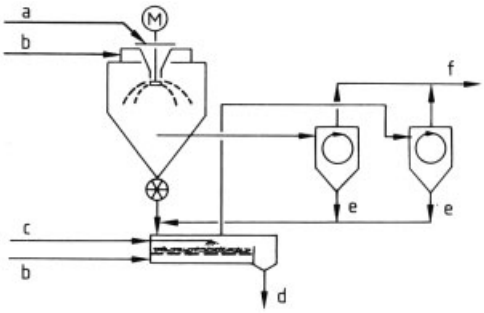

Figure 27. Spray dryer with downstream fluidized bed dryer with injection of agglomeration liquid
a) Moist product; b) Drying gas; c) Agglomeration liquid; d) Dry product, agglomerates; e) Recycling of fines; f) Exhaust gas

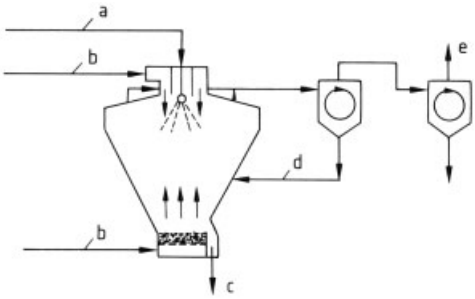

Figure 28. Spray dryer with integrated fluidized bed
a) Moist product; b) Drying gas; c) Dry product, agglomerates; d) Recycling of fines; e) Exhaust gas

4.3.2. Contact Drying

Contact drying is roller drying (Fig. 29). Pumpable products, including low- and high-viscosity materials as well as pastes, can be dried on a roller [99]. Metal rollers are heated internally with steam. The heat used for drying is obtained by condensing the steam. The resulting condensate must, therefore, be removed from the lowest point within the roller. The product to be dried is applied as a thin layer onto the hot roller. This coating process can be done in different ways. The principal methods are classified either as roller application methods or as spray methods.

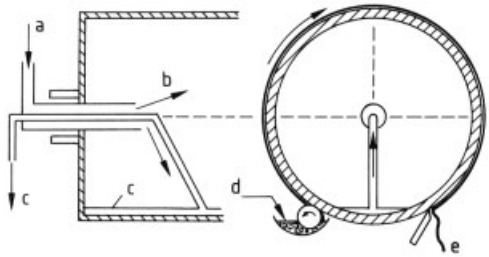

Figure 29. Roller dryer [99]
a) Hot steam; b) Steam; c) Condensate; d) Liquid product; e) Dry product

The roller can dip directly into the material to be dried. This type of equipment is referred to as a sump roller dryer. Temperature and concentration gradients are caused by heat transfer to the fluid. These can be removed by means of an agitator. The disadvantages of this design are the nonuniform coating on the roller and exposing the product to heat, the latter having quite negative effects in the case of heat-sensitive products.

An improvement is achieved by using a so-called coating roller (Fig. 30). In this case, only the coating roller dips into the liquid. It coats a liquid film of uniform thickness on the drying roller. Another variation of these design makes use of an additional roller, called a transfer roller. This roller may be coated with hard rubber and cooled internally with water. In addition to single roller dryers, there are also heated twin rollers (dual roller dryers) in use (Fig. 31). Examples of this type are the sump dual roller dryer and the dual roller dryer with spray coating. The dry product is scraped off by a knife. It may be in the form of a film, or flocs, flakes, or powder [99]. The products are usually removed with a

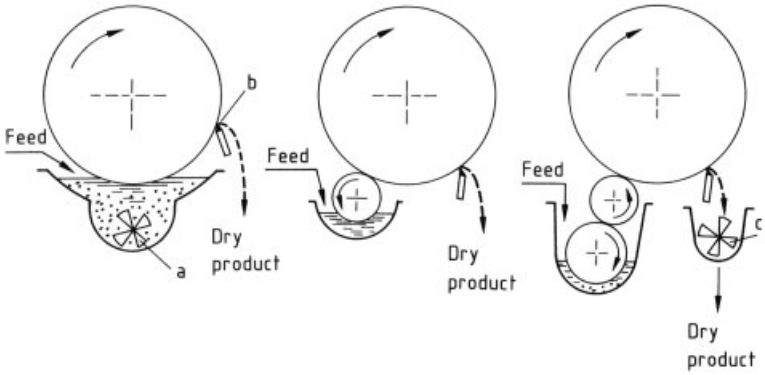

Figure 30. Roller dryers with various feed systems [99]
a) Stirrer; b) Blade; c) Screw conveyer

screw conveyor. If necessary the particle size can be reduced to the desired value by means of a hammer mill or crusher. In designing this type of dryer, the product feed point and the product removal point must be separated. For this reason, only 70 – 80 % of the circumferance is effectively used for drying.

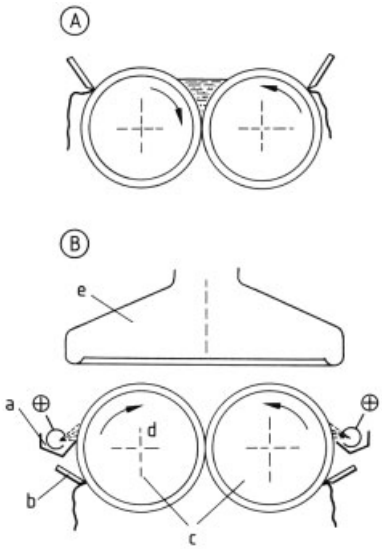

Figure 31. A) Dual roller steam dryer and B) Dual roller dryer with spray coating
a) Distributor trough; b) Scraper; c) Drying rollers; d) Steam space; e) Exhaust hood

4.3.3. Vacuum Drying

In the case of temperature-sensitive products, the external pressure must be lowered to avoid drying at a high temperature. This is called vacuum drying. One simple type of vacuum drying is vacuum roller drying. In this method, either one or two rollers are installed in a vacuum housing. The resulting vapor precipitates in a condensor located between the vacuum chamber and the pump. The product is removed by a screw conveyor (Fig. 32).

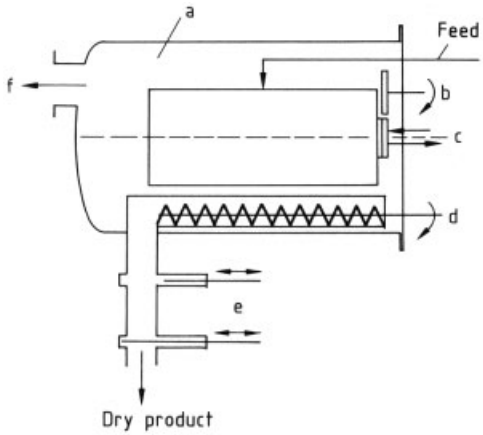

Figure 32. Vacuum roller dryer [99]
a) Vacuum chamber; b) Roller drive; c) Heat; d) Screw conveyer drive; e) Air lock; f) To condenser and vacuum pump

Freeze drying is also widely used for sensitive products. Freeze drying, also called sublimation drying, is defined as the drying of frozen material. At pressures below 0.6 kPa, ice goes directly into water vapor, without passing through the liquid state. Dissolved material (for example, salts) lower the freezing point. Actual sublimation is only possible at the eutectic temperature (the point of maximum freezing point depression) [99]. Kluge and Heiss [100] used the

nonenzymatic browning reaction to demonstrate qualitatively that freeze drying was the best drying process. The biggest disadvantage of this process is its high capital and operatings costs. Products with good instant properties can be produced by freeze drying because sublimation produces a highly porous solid structure. The product must first be frozen before drying. The rate of freezing has a strong effect on both product quality and the rate of drying in the freeze dryer. Rapid freezing produces small ice crystals, a strong product structure, and fine pores. On the other hand, these factors increase the resistance to the flow of moisture during drying. Freeze drying can be performed batchwise or continuously. Figure 33 shows a cross sectional diagram of a batch freeze dryer. Continuously operating freeze dryers are illustrated in Figure 34. A vibrational freeze dryer and a disk freeze dryer are shown.

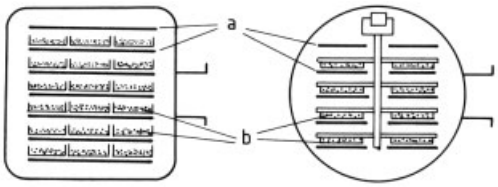

Figure 33. Cross-sectional view of batch freeze dryers
a) Heating surfaces; b) Product trays

4.4. Spherical Agglomeration [101]

Another possible method of forming agglomerates in a fluid medium is the addition of a second immiscible liquid [102]. This binding liquid must be able to wet the particles better than the liquid in which they are suspended. This process, called spherical agglomeration takes place in a three-phase system. The three phases are the suspending liquid, the immiscible binding liquid, and the solid dispersed phase. This phenomenon, which occurs during the progressive addition of the bridging liquid, is illustrated in Figure 35. The finely dispersed droplets of binding liquid encounter the particles in a dynamic system. The particles cling to the droplets and become concentrated in a manner not unlike flotation. This process may take place in a few seconds or last for hours, depending on the wetting properties and composition of the system. Bemer [103], [107] calls this process the flocculation phase. He used 19 µm glass spheres in carbon tetrachloride and a water – glycerin mixture as the binding liquid. Flocs roughly 100 – 300 µm in size were formed. This was followed by an extended period of apparent stability. This was followed suddenly by a period of rapid agglomerate growth. The agglomerates grew by coalescence until they arrived at a stable state, in which the cohesive and dispersive forces were in equilibrium.

Farnand [104] has stated three factors that control the spherical agglomeration process, without describing their effect in a quantitative manner. These are, in order of importance:

1) Relative interfacial free energies
2) Ratio of binding liquid to solid
3) Type and intensity of mixing

The effect of interfacial free energy on spherical agglomeration has not been described in the literature. During the wetting stage, flocs, agglomerate nuclei, spherical agglomerates, even pastes can form depending on the composition of the three-phase system and the ratio of binding liquid to solid [104–110].

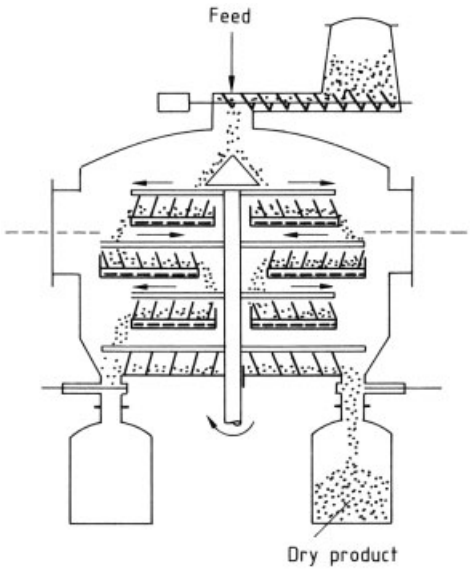

Figure 34. Continuous freeze dryer

The effect of the wetting properties of the solid and liquid components makes the agglomerates sensitive to composition. For certain combinations of materials, the agglomeration process can be made selective. Therfore in a mixture

of solids, particles with a certain composition can be made to form agglomerates, while the remaining material can be separated by screening. By changing the liquids, additional components can be obtained in succession in a way resembling preparation by flotation. For this reason spherical agglomeration is also called selective agglomeration.

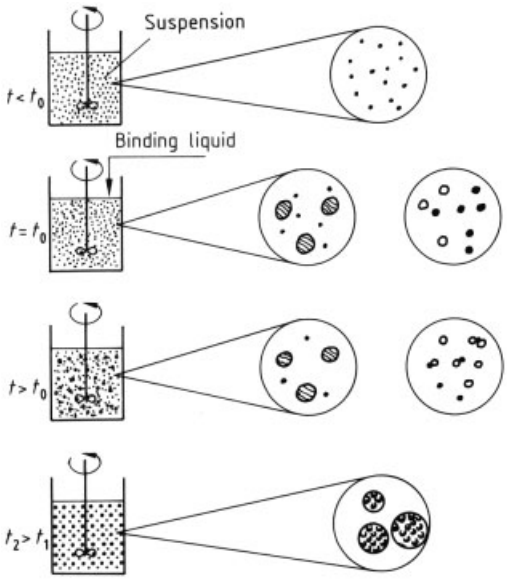

Figure 35. Stages of spherical agglomeration according to Bröckel [103]

Application Cases. A major area of application of spherical agglomeration is the removal of fine coal particles dispersed in water. The fraction of fines produced during coal preparation is increasing. This is due to the increased use of mechanized mining methods and finer milling to produce a cleaner final product. Coal particles in aqueous suspension can be agglomerated with special oils and removed. Inorganic or ash components are left behind.

Olifloc Process. The flowsheet of the olifloc process (Bogenschneider, Ruhrkohle AG, 1976) is shown in Figure 36. The pilot plant was designed for a throughput of ca. 3 t/h. This process is a further development of the Convertol process developed by the Essen Mining Association in the 1950s.

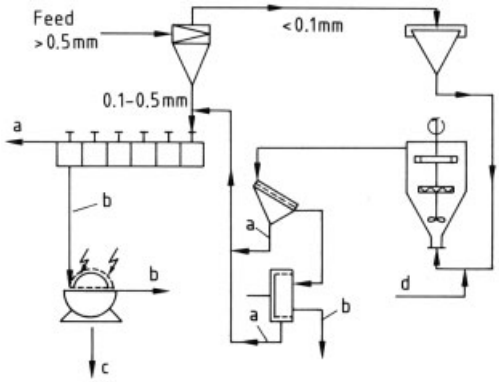

Figure 36. Process flowsheet of the olifloc pilot plant
a) Gangue; b) Coal; c) Water; d) Oil

Operation. The local slurry with a particle size of $< 500\,\mu m$ is classified in a hydrocyclone. The fraction $< 100\,\mu m$ is sent to a clarifier to increase its solids concentration. The underflow with a solids content of 200 – 400 g/L is removed and mixed with oil, which serves as a binder. The components are intensively mixed and agglomerate in the upper section of the mixer. An inclined dewatering screen is used to separate the agglomerates from the gangue, which remains behind in the water. The residual moisture in the oil – coal agglomerates is lowered to 15 % with a scrubbing centrifuge. The centrifugates from cyclone and screen go to flotation to further recover the coal content. Bogenschneider states that the residual moisture in a coal preparation plant can be reduced to about 3 % by using this process.

Spherical Agglomeration Process. Spherical agglomeration, as illustrated in Figure 37, was developed in the 1960s by the National Research Council of Canada [111].

Operation. Feed slurries are mixed in 2 high capacity mixers (a) with, for example, diesel fuel in order to completely disperse the binder. The agglomerates form in the agglomeration vessel (b). They are partially separated from the gangue by sedimentation in vessel (c). An inclined dewatering screen also serves to separate the agglomerates from the gangue. In this variation of the process, the agglomerates are pelletized with hot heavy fuel oil on a disk pelletizer. The pellets reach a diameter of ca. 6 mm with high strength.

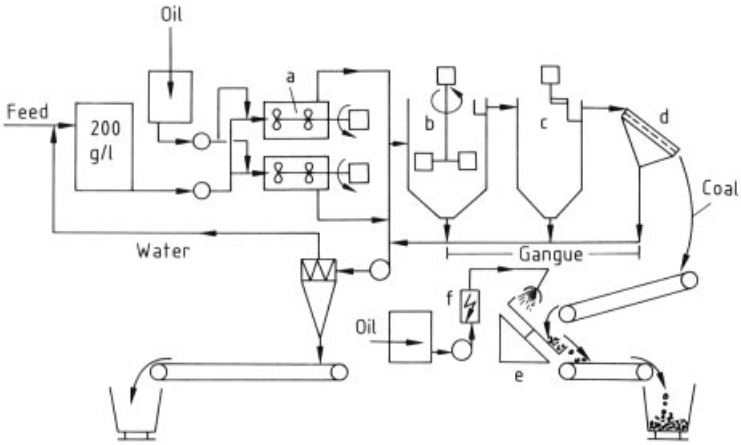

Figure 37. Process flowsheet for the spherical agglomeration process
a) High capacity mixer; b) Agglomeration and settling vessel; c) Stripping vessel; d) Dewatering screen; e) Disk pelletizer; f) Continuous heater

Pelletizer – Separator. The pelletizer – separator (Fig. 38) [111] has a simple design. It generates solid pellets 1 – 6 mm in diameter by churning with rotating paddles. Nearly all fine coal slurries can be selectively agglomerated by adding 9 – 15 wt % of heavy oil. The binder is almost completely recovered in the pellets. An industrial plant with a capacity of 10 t/h has been in operation in Japan since 1972.

Operation. A coal – water suspension and heavy fuel oil are fed simultaneously from the left. After mixing in the first third of the vessel, the coal is selectively agglomerated in the second third. Next it is separated from the gangue on an inclined dewatering screen.

4.5. Pressure Agglomeration

The use of liquids in size enlargement is dependent on the compatibility of the product with the liquid. In some cases, the necessity of adding a drying step increases the cost. Pressure agglomeration is used when it is desirable to proceed without the use of liquids. During compression the number of interparticle contact points is increased in the substrate. Then by plastic deformation the contact points are converted to areas of interfacial attraction. Van der Waals forces or sintering are the principal factors affecting strength in the case of compression. Both increase with increasing surface contact area. Pressed articles with high strength can be obtained with materials that can be deformed plastically. It is difficult to compress elastic particles with unfavorable properties. Compressive properties can be improved by addition of a binder. If a binder is used, the compressive properties depend on what type of binder it is and its binding capacity. The particle-size distribution of the starting material is also a factor. Above all, coal, low-temperature coke and ores are briquetted with binders.

The pressure agglomeration process has a long history in the preparative and coal industries as the concept of briquetting. Coal tar pitch, bitumen, or 50 % solids waste sulfite liquors are added to plasticize coal powder. Fine grained ores and other fine grained metal containing materials are mixed with fine limestone, magnesium hydroxide, or clay before being compressed. In addition to coal and ores, large amounts of ceramic precursors, fertilizers, plant protectants, and pharmaceuticals are produced with pressure agglomeration processes.

In *roller compaction,* the powder is introduced in between two smooth or mold rollers and is compressed. The product is either thin sheets (flakes) or molded shapes. In addition, the process is widely used as an intermediate step in agglomeration processes. One example is the pregranulation of feed materials for tableting.

The oldest process is *compression* with *stamp* and *die*. This process is widely used in powder

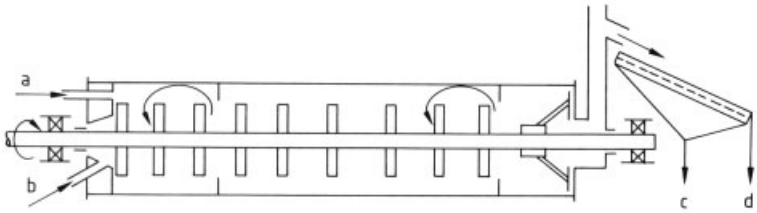

Figure 38. The Shell pelletizer-separator
a) Heavy fuel oil; b) Coal water suspension; c) Gangue; d) Pellets

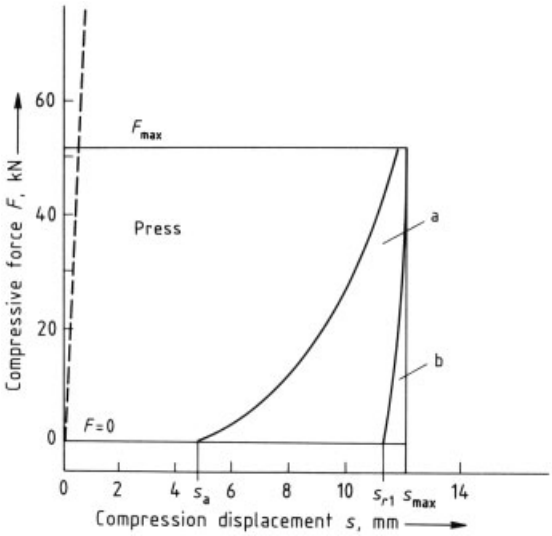

Figure 39. Force – displacement diagram for compression, cornstarch compression rate $v_V = 0.24$ cm/s
a) Compression energy E_a plastic; b) Compression energy E_2 elastic

metallurgy and the coal industry, as is roller compaction.

In *isostatic compression,* the powdered starting material is loaded into an elastic membrane and is compressed from all sides by fluid pressure. The resulting pressed object is homogeneous and has little internal stress. Compared to tableting, the pressure is relatively low. Because of its cost, this process is not suitable for high volume applications. It is used primarily in powder metallurgy.

Vibration compaction is especially suited for the production of long pressed articles with uniform properties. It is used in the ceramic industry and, for example, for the production of reactor fuel rods. The substrate is homogeneously compressed by the wall impact of the vibrating container as well as the uniform pressure due to interparticle collisions.

4.5.1. Compression Equations

Figure 39 illustrates the force – displacement diagram for compression of cornstarch, with its elastic and plastic regions. There are over 15 different mathematical models in the literature [113], [114] for compression. These described the volume reduction of a powdery material when a compressive force is applied. Depending on the magnitude of the compressive force or the type of material, most of these models have only a limited validity. The best known compression equations are listed in Table 2.

Table 2. Mathematical models of compression [115]

Athy	$\frac{V-V_\infty}{V_0} = \frac{V_0-V_\infty}{V_0}\,\mathrm{e}^{-k_1\sigma^c}$	(a)
Ballhausen	$ln\,\frac{V_\infty}{V-V_\infty} = K_2\sigma_c + K_3$	(b)
Balshin	$ln\sigma_c = -K_4\frac{V}{V_\infty} + K_5$	(c)
Cooper	$\frac{V_0-V}{V-V_\infty} = k_6\,\mathrm{e}^{-k_7\sigma_c} + k_8\,\mathrm{e}^{-k_9\sigma_c}$	(d)
Führer	$(s-k_{10})(\sigma_c + k_{11}) = k_{12}$	(e)
Gurnham	$\sigma_c = k_{13}\,\mathrm{e}^{k_{14}}/V$	(f)
Heckel	$ln\,\frac{V}{V-V_\infty} = k_{15}\sigma_c + \frac{V_0}{V_0-V_\infty}$	(g)
Jones	$ln\sigma_c = -k_{16}\left(\frac{V}{V_\infty}\right)^2 + k_{17}$	(h)
Kawakita	$\frac{V_0-V}{V_0} = \frac{k_{18}k_{19}\sigma_c}{1+k_{20}\sigma_c}\quad k_{18} = \frac{V_0-V_\infty}{V_0}$	(i)
Konopicky	$ln\,\frac{V}{V_0-V} = k_{21}\sigma_c + \ln\frac{V_0}{V_0-V_\infty}$	(j)
Murray	$ln\,\frac{V}{V-V_\infty} = k_{22}\left(\frac{V_\infty}{V-V_\infty}\right)^{1/3} + k_{23}\sigma_c$	(k)
Nishishara	$ln\left(\frac{V_0}{V}\right) = -\left(\frac{\sigma_c}{k_{24}}\right)^{1/k_{25}}$	(l)
Nutting	$ln\left(\frac{V_0}{V}\right) = k_{26}\sigma_c^{k_{27}}$	(m)
Smith	$\frac{1}{V} - \frac{1}{V_0} = k_{28}\sigma_c^{1/3}$	(n)
Tanimoto	$\frac{V_0-V}{V} = \frac{k_{29}\sigma_c}{V} + \frac{k_{30}\sigma_c}{\sigma_c+k_{31}}$	(o)
Terzaghi	$\frac{V-V_\infty}{V_\infty} = k_{32}\ln(\sigma_c+k_{33}) - k_{34}(\sigma_c+k_{35}) - k_{36}\sigma_c + k_{37}$	(p)
Tsuwa	$\frac{V_0-V}{V_0} = \frac{V_0-V_\infty}{V_0}\,\frac{1/k_{38}\sigma_c}{1+1/k_{39}\sigma_c}$	(q)
Walker	$\frac{V}{V_\infty} = k_{40} - k_{41}\ln\sigma_c$	(r)
Leuenberger	$P = P_{\max}(1-\mathrm{e}^{\gamma\cdot\sigma_c}\varrho^r)$	(s)
Ryshkewitch	$F = F_0\cdot 3^{-b\cdot\varepsilon}$	(t)
σ_c	Compressive pressure	
V	Volume $V(\sigma_c)$	
V_∞	Minimum possible volume ($\sigma_c \to \infty$)	
V_0	Initial volume	
k_i	Constants specific to the model and material	
s	Compressive displacement	
P	Deformational strength of the compressed body	
$P_{\max}$	Deformational strength of the compressed body when $\varrho_r = 1$	
ϱ_r	Relative density $(1-\varepsilon)$	
F	Strength of compressed body	
F_0	Strength of compressed body at porosity = 0	
b	Constant related to composition	

Aggregates of powdery material can be described by their compressibility, that is, the ease of volume reduction under pressure, or by their pressability, that is to say, the ease with which they can be pressed into a molded object with sufficient strength. All of the compression equations presented in Table 2 have the disadvantage that they describe the compressibility of a powder rather than its pressability. For practical applications, the pressability is the more important parameter. For this reason, a mathematical model was developed which linked the two parameters of a powder, the compressibility and the pressability (Table 2, Eq. (s)). Equation (s) is in good agreement [115] with the experimental data. This has been demonstrated for a number of pharmaceutical products and additives. These equations are derived from mathematical models based on binding and nonbinding contact points. In addition, there are empirically derived equations in the literature. Various workers [116] cite the formula found by Ryshkewitch [117] for ceramic materials. This is the relationship (Table 2, Eq. (t)) between the strength and the porosity of a molded object.

4.5.2. Stamp Presses

Closed-form stamp presses, as shown in Figure 40, work in the following way: the material to be compacted is loaded into a mold, which is closed on all sides except one. A stamp, which closely fits the mold, is applied and presses the material into the desired form of a briquette. The stamp is operated automatically with an electric motor. The pressure can be transmitted hydraulically with oil or water, or by means of a screw spindle. The finished briquette can be

forced out of the mold in the original direction or in the opposite direction by a counter stamp. In the preparative industry a typical size for cylindrical briquettes is 12×7 cm. Typical values for compressive force are 318 t supplied by an hydraulic oil pump of 56 kW capacity. Throughput is 4 – 6 t/h. An alternative high-capacity press is an automatic table press (couffinhal press). It is used for producing the briquettes with binders in the construction industry. The press consists of a rotating, horizontal, circular platform in which are set molds. After each pressing step, the platform is rotated to the next mold by means of guide rollers. The pressing action is commonly assisted by vibration. The production capacity for cubic or round briquettes can reach 20 t/h or even higher. The tablet presses of the pharmaceutical and catalyst industries are types of stamp presses, as are some extruders.

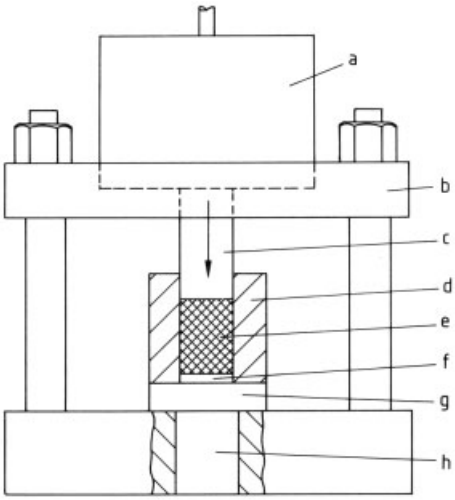

Figure 40. Hydraulic stamp press [118]
a) Hydraulic cylinder; b) Press frame; c) Upper stamp; d) Mold; e) Substrate; f) Closing plate; g) Bottom plate; h) Exit port

4.5.3. Roller Presses

In a roller press, the material is introduced between two oppositely turning rollers and is compressed (Fig. 41). Smooth rollers produce flakes. This is called roller compaction. Specially shaped rollers produce pressed objects. This is termed roller briquetting. Gravity feeding from bins is the preferred method in the beginning stages of roller press compaction. With increasing fineness of the feed, screw feeders are used. The specific operational configuration, such as the number of screw feeders, the size of the feed opening and the screw feeder pressure, is determined experimentally for the most part.

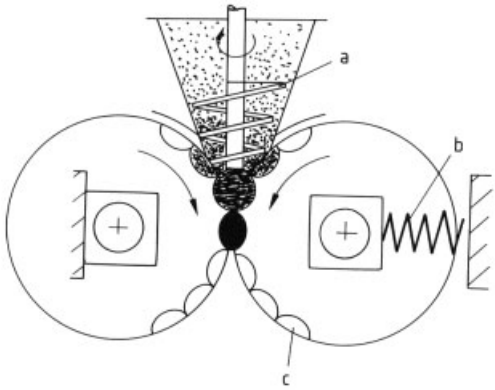

Figure 41. Operation of a roller press with screw feeder
a) Screw feeder; b) Applied pressure; c) Smooth or shaped roller

An essential factor in pressure agglomeration by roller presses is the external compressive forces. These generate a pressure distribution in the space between the rollers, which has a maximum. An average pressure can be calculated from the pressure distribution. Usually the specific compressive force is specified rather than the pressure. This is defined as the total compressive force relative to the width of the roller (linear force). The specific compressive force usually has a range of 10 to 140 kN/cm for rollers 1000 mm in diameter. The specific compressive force required for various materials is shown in Table 3.

The rotational speed of roller presses is limited by behavior of the feed as trapped air is expelled. This can only be determined experimentally by simulation of industrial conditions. A fine grained, loose material has a large void volume. This is diminished drastically during pressure agglomeration. As a result, a correspondingly large volume of air is released. In order to maintain a steady compression process, the air must be able to escape upward from the space between the rollers. This is opposed by the downward flow of the feed. The air removal problem also directly affects the maximum throughput of the roller press, which is

primarily a function of the circumferential speed of the roller. The problem of air removal is especially acute with smooth rollers. Many variations in shape have been tried in order to find a remedy. The circumferential speeds conforming to the air removal properties of different materials are listed in Table 4.

Table 3. Specific compressive forces [119]

Material	Specific compressive force (for 10 cm diameter roller), kN/cm
Coal	10 – 30
Ceramic raw materials	40
Rock salt	60 – 80
Mixed fertilizer (with urea)	40 – 60
Mixed fertilizer (without phosphate rock and ground slag)	50 – 80
Mixed fertilizer (with phosphate rock and ground slag)	≥80
Calcium cyanamide (nitrolime)	60
Potassium chloride (120 °C)	50
Potassium chloride (20 °C)	70
Potassium sulfate (70 – 100 °C)	70
Scrubber sludge	95
Magnesite (cold and hot)	110 – 130
Quicklime	130
Spongy iron (cold and hot)	130 – 140
Ores (cold with binder)	60 – 80
Ores (hot without binder)	120 – 140

The specific energy requirement is dependent on the power consumption and the throughput. While the throughput is generally of a presupposed size, the power consumption must be determined. It depends on the torque and the rate of rotation. Both of these quantities must be determined experimentally for the material being processed (Table 5).

Table 5. Specific energy requirements [119]

Material	Specific energy, kW h/t
Scrubber sludge	11 – 12
Potassium chloride	7 – 8
Caustic magnesite	18 – 20
Spongy iron	8 – 12
Ore	7 – 9
Quicklime	16 – 19
LD-process dust (hot)	7 – 9
Coal (binder)	2 – 3
Coal (hot)	4 – 6
Mixed compost (100 – 120 kN/cm)	12 – 13
Mixed compost (< 80 kN/cm)	9 – 11

The type of material feeding system is governed by the properties of the material such as its bulk density and flow properties. It is always desirable to have a continuous and uniform material feed in order to have optimal loading of the press. Either gravity feed equipment (hoppers) or power feeders (screw feeders) can be used. In cases where the material tends to cake or form bridges, the flow of material may be assisted by vibrators installed in the hoppers. If the material has good flow properties, a hopper is sufficient. It is provided with a slide plate in order to regulate the amount being fed. This is done by using the slide plate to change the effective size of the loading area cross section. Screw feeders are used in cases where adequate feeding cannot be obtained by gravity. This may occur if the bulk density is too low or the loading capacity is insufficient. Screw feeders of this type transport the material directly into the loading area. In this manner, they perform a type of precompression. The machines must be designed as precisely as possible in order to obtain optimal product properties. For this reason the interrelation between the compressive characteristics of the material (rheology) and the machine parameters, especially the compressive force and the roller diameter, must be known [120–127].

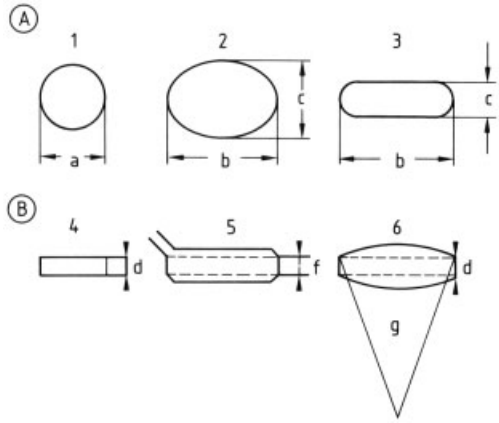

Figure 42. The most important tablet shapes [128]
A) Top view: 1) Round; 2) Oval; 3) Oblong; a) Diameter; b) Length; c) Width
B) Cross section: 4) Biplanar; 5) Biplanar with beveled edges; 6) Domes (biconvex); d) Tablet height; e) Facette; f) Median height; g) Dome radius

Table 4. Roller rotational speeds [119]

Material	Maximum peripheral speed, m/s	Briquette size, cm^3	Type of roller surface	Roller diameter, mm	Roller width, mm
Scrubber sludge	0.35	5 – 10		750	600
Potassium chloride	0.70		molded	1000	1250
LD-process dust (fine)	0.25	20		650	220
Magnesium oxide (natural)	0.40	5 – 7		650 – 750	320
Lead – zinc oxide	0.27	100		750	265
Zirconium tetrachloride	0.17		smooth	500	200
Dolomite dust	0.17	6.5		650	250
Quicklime (< 5 mm)	0.50	10		650 – 1000	250 – 540

4.5.4. Tableting

Tablets originated in the field of pharmacy. From a technological viewpoint, they were defined as an individual dose of a medicinal substance. These were produced in various shapes by compression of powders or granules (Fig. 42).

K. Munzel, I. Buchi, and O.-E. Schultz [128] give the following definition. "Tablets are rigid aggregates or pressed articles (pressed bodies) produced by compression by mechanical means. They are made from dry dispersions such as powders, macroscopic crystals and/or granules and the medicinal material that they contain. As a rule, these pressed articles exhibit radial symmetry or at least have a plane of symmetry."

The origin of the term tablet is the Latin word "tabyletta" (the diminutive form of tabula), which means "little board." This designation is derived from the usually flat shape of early tablets. It is only appropriate today in very rare cases, since most tablets are cylindrical. Even though tableting today is in the domain of pharmaceutical practice, other areas of application are growing. For example, it is used in chemistry for the production of catalysts and in the ceramic industry for the production of reactor packing material. More recently it has been used for making carriers for immobilized enzymes in bioreactors.

The production capacity of tableting machines has been developed further. Centrally rotating machines produce about 700 000 tablets per hour today. For a tablet weight of 500 mg, this gives an impressive hourly production of 350 kg/h.

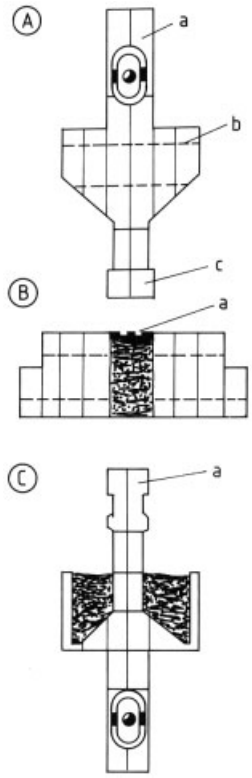

Figure 43. Working components of tableting machines [130]
A) Upper punch: a) Set pin; b) Support plate; c) Punch head
B) Die: a) Die bore
C) Lower punch: a) Punch head

4.5.4.1. Tableting Equipment

In order to produce a tablet, the material to be used is loaded into a mold (die) and then compressed by two punches (upper and lower punches) as shown in Figure 43. The size and geometric shape of the stamp affects the weight

and shape of the tablet as well as its outer surface. As a result, it also affects the properties of the tablet, such as its compressive strength and disintegration time. The care and maintenance of the pressing equipment is of great importance to the success of the tableting process.

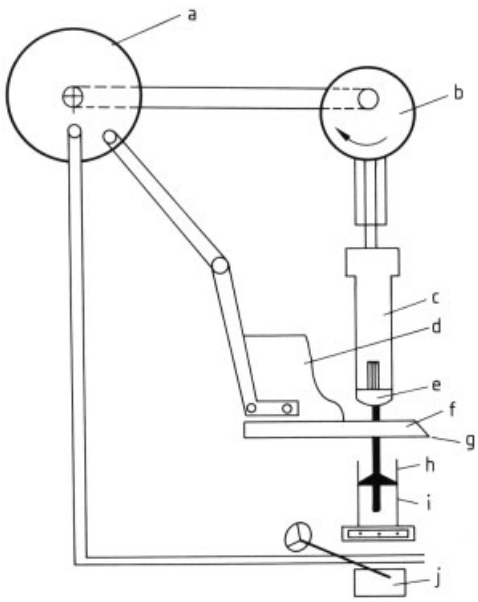

Figure 44. Diagram of an eccentric press [130]
a) Driving wheel; b) Eccentric disk; c) Upper punch support; d) Filling device with hopper; e) Upper punch; f) Die; g) Die plate; h) Lower punch; i) Lower punch support; j) Loading depth adjustment

4.5.4.2. Tablet Pressing

Two principal types of tablet presses are in use: the eccentric press and the rotating press.

Eccentric Press (Fig. 44). The compressive force is transmitted from a cam to the substrate by the upper punch. The lower punch is usually stationary during the compression process. It can, however, be adjusted to change the depth of the loaded material. There can also be a precompression step by means of an upward motion of the lower punch, instead of the unidirectional pressure of the upper punch. After the completion of the pressing process, the tablet is removed by an expulsion stroke of the lower punch (Fig. 45). With their simple design, eccentric presses only have a production rate of 3000 tablets/h, which is rather low. In order to increase capacity, multiple arrays of eccentric presses are used. They are clearly less expensive than rotating presses and are preferred for the production of smaller batches as well as the production of larger tablets, such as catalyst supports.

Rotating Press. It operates in the same way as a revolver. The compressive force is generated between an upper and a lower pressure roll. In contrast to the usual type of eccentric press, the tablet is compressed from both above and below (Fig. 46). The depth of loading is controlled by a system which uses either a wire strain gauge or a piezoelectric sensor. The pairs of punches along with their dies are mounted in a rotating die platform. Each pair of punches produces one

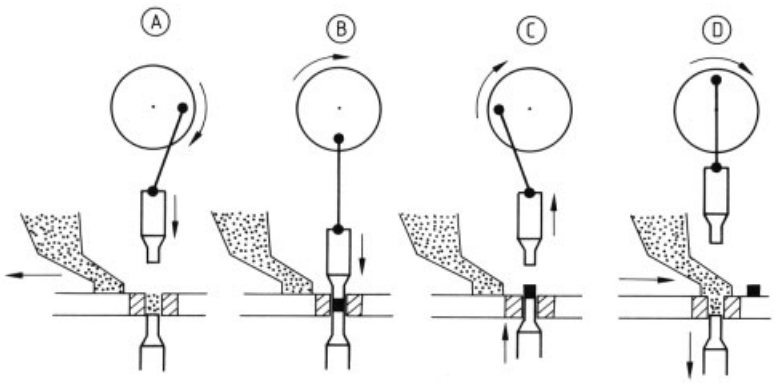

Figure 45. Schematic illustration of the compression process in an eccentric tablet press [128]

tablet per revolution. The capacities of the rotating presses range from $10-700\times10^3$ tablets/h, depending on the number of punch pairs and the rate of rotation. For very high speed presses the persistence of the flow velocity of the material being tableted can present problems. For this reason, pre-granulation to produce a free-flowing feed material is of considerable importance.

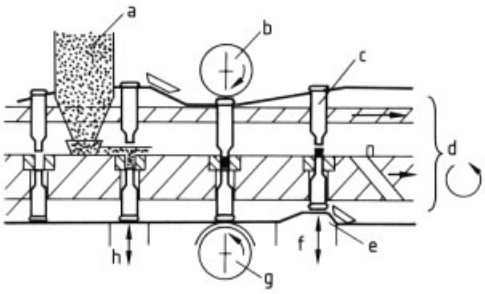

Figure 46. Schematic illustration of the compression process in a rotating tablet press [128]
a) Hopper (stationary); b) Upper pressure roller; c) Upper punch; d) Rotating die platform; e) Lower punch; f) Tablet removal mechanism; g) Lower pressure roller; h) Loading depth adjustment

4.5.4.3. Tableting Problems

If the tablet press (rotating or eccentric) runs properly without material and the adjustable parameters such as tablet thickness are set correctly, then any perturbations of the process are limited to those related to the properties of the substrate. The weight of the product can be faulty. This is exhibited by an intolerable fluctuation in the weight of the tablets. The cause usually is poor flow properties of the substrate. It can form bridges in the hopper which interrupt the flow, and consequently the loading process. In addition, undesirable vibrations can cause segregation that impairs the uniformity of the tablets. Too high a moisture content of the substrate or too little lubrication leads to sticking, crackling, and coating. The resulting inadequate release of the tablet from the surface of the stamp causes surface defects on the tablet (nonuniform color, roughness). Coating occurs when the tablet breaks up into upper or lower fragments which form layers on the surfaces.

The interaction of the particle size distribution, the crystal shape and other physical properties of the substrate is the principal factor affecting the compressive strength, the disintegration time, and the release of the active ingredient.

The properties of the tablet are thus substantially affected by the properties of the fillers. In case of difficulties therefore, there is always the option of changing the formulation (binder, lubricant). If this is not possible, the other process variables such as the pressure and speed and other environmental conditions must be investigated.

4.5.4.4. Effect of Compression Rate and Humidity [133]

For practical application, the most important major parameters affecting tableting and tablet properties are humidity and compression rate. Even under extreme conditions, such as high humidity, high press speed, and difficult substrates, it should be possible to avoid tablet defects. A range of techniques and types of measurements are used to investigate the tableting process and the resulting tablet properties. The aim is to comprehend the microscopic processes that are occurring during tableting and to be able to draw conclusions about tablet properties from them [137]. The example of tableting powdered limestone in moist air (45 % relative humidity) shows the effect of the speed of compression on the compressive behavior. The force displacement curves are shown in Figure 47. Tablets with constant weight and the same diameter were produced. The average values for the three different compression rates v_v which were used were 0.3, 2.5, and 9.0 cm/s. The compressive pressure was held constant at ca. 110 MPa. In tableting, the compression rate is defined as the ratio of the displacement of the punch from the first application of a detectable force to the point of maximum force to the time required for this displacement to occur. The compression pressure is plotted as a function of the compressive displacement of the lower stamp or the volume fraction of solid $(1-\varepsilon)$ in the pressed object. Compression occurs earlier with increasing compression rate and the rate of pressure increase is lower. However, the rates of increase become equal at higher levels of compression. The principal reason for the lower rate of pressure increase at high compression rates is the frictional effect of air rapidly

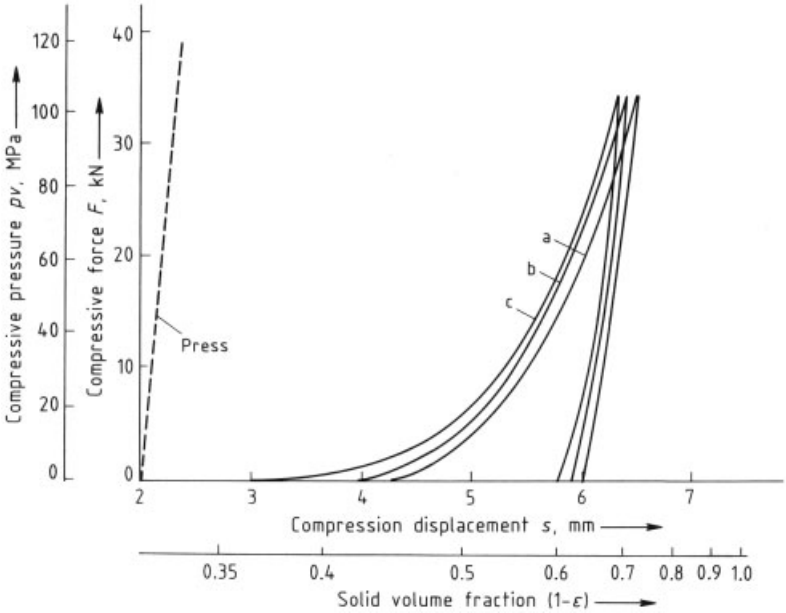

Figure 47. Force – displacement diagrams of powdered limestone (x_{50} = 13.5 µm) for tableting for three compression rates v_v [133]
a) v_v = 0.3; b) v_v = 2.5; c) v_v = 9.0 cm/s; Parameters of tablets: m = 2.25 g; $h_{loading}$ = 10 mm; $\varnothing_{Tablet}$ = 20 mm

escaping from the powder pressed between the punch and the die. In addition, sliding and rearrangement processes of the particles are hindered by turbulence in the powder due to rapid compression.

The density of tablets made with the same pressure is lower for rapidly made tablets. The elastic recovery when they are released and removed from the die is larger. The ratio of the energy incorporated into the tablet to the energy delivered by the machine is consistent with this. The relative energy incorporation is higher for slow compression compared to rapid compression. The elastic recovery of the tablets after release and removal from the die is limited by the recovery of the elasticly stressed particles. On one hand, consequently, the contact area between the particles is reduced, resulting in a loss of strength. On the other hand, there is a reduction of stress inside the tablet, which has a favorable effect on strength. Strength studies must be done to determine which mechanism predominates.

The example of the effect of compression rate on the shape of the compression diagram illustrates how such a diagram can be used to characterize changes in compression parameters, for the same material under otherwise constant conditions. The first conclusions can be drawn concerning the trends in properties of the resulting tablets by determining the compression energy and the elastic recovery behavior. It should be noted that the compression diagram is only an integrated representation of all the complex processes which occur during tableting. The evidence obtained from the compression diagram must therefore first be checked and proven by the appropriate types of measurements. Figure 48 shows the diametrical compressive strength as a function of compression rate v_v for limestone tablets produced under vacuum (6×10^{-4} MPa) and in moist air (45 % and 80 % relative humidity). The compressive pressure p_v was 110 MPa. The production and testing conditions for the tablets being studied are abbreviated as illustrated in the following example. The designation air/vacuum tablet means that the tablet was made in air and tested under vacuum.

The following conclusions, which are applicable to actual tableting, can be drawn from this investigation of strength:

1) Substances that exhibit problems when tableted at high humidity and high compression rates should be tableted at low humidities.
2) Certain compression rates may not be exceeded.
3) Strength testing must be done by a test that is designed to match the actual requirements of the tablet as closely as possible. Furthermore, testing must be done in a controlled atmosphere, otherwise the test results cannot be compared with each other.
4) Low humidity is favorable for resistance to stress when the tablets are stored or transported [134–139].

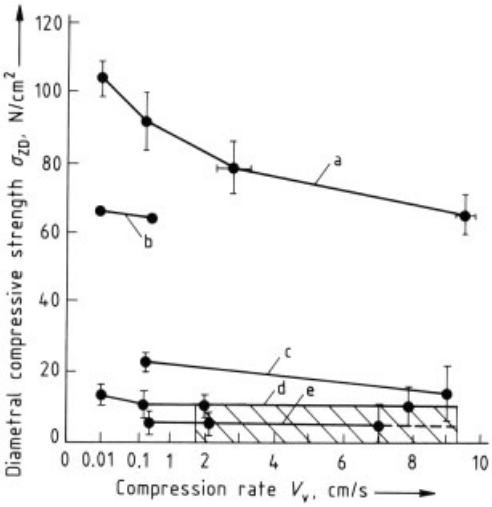

Figure 48. Diametrical compressive strength as a function of compression rate [133]
Limestone $x_D = 13.5\,\mu m$; $m_{Tablet} = 2.25$ g; $\varnothing_{Tablet} = 20$ mm; $h_{Tablet} = 3.6 - 4.4$ mm; $p_v = 110$ MPa
The rectangular region marked by diagonal lines is the region where 10 – 50 % of the tablets produced are defective.
a) Vacuum/vacuum; b) Air/vacuum; c) Vacuum/air (45 % rel. humidity); d) Air/air (45 % rel. humidity); e) Air/air (100 % rel. humidity)

5. References

General References

1. R. H. Perry, C. H. Chilton: *Chemical Engineer's Handbook,* McGraw Hill, New York 1973.
2. A. W. Adamson: *Physical Chemistry of Solids,* Wiley-Interscience, New York 1967.
3. B. Ries, *Ind. Anz.* **21** (1961) 1 – 4.
4. P. J. Sherrington, R. Oliver: *Granulation,* Heyden & Son, London–Philadelphia–Rheine 1981.
5. C. E. Capes: *Particle Size Enlargement,* Elsevier, Amsterdam–Oxford–New York 1980.
6. W. Herrmann, K. Sommer: *Preprints 3rd Int. Symposium Agglomeration,* Nürnberg 1977.
7. G. Astarita, *Chem. React. Eng. Proc. Eur. Symp. 5th* 1972, Amsterdam.
8. W. Herrmann: *Das Verdichten von Pulvern zwischen zwei Walzen,* Verlag Chemie, Weinheim 1973.
9. K. Sommer, W. Herrmann, *Chem. Ing. Tech.* **50** (1978) 518 – 524.

Specific References

10. H. Rumpf, *Chem. Ing. Tech.* **30** (1958) 144 – 158, 329 – 336.
11. H. Schubert, *Chem. Ing. Tech.* **51** (1979) 266 – 277.
12. H. Rumpf in W. A. Knepper (ed.): *Agglomeration,* Wiley-Interscience, New York 1962, pp. 379 – 414.
13. H. Rumpf, *Chem. Ing. Tech.* **42** (1970) 538 – 540.
14. W. O. Smith et al., *Phys. Rev.* **34** (1929) 1271 – 1274.
15. H. Schubert, *Powder Technol.* **11** (1975) 107 – 119.
16. H. Schubert: *Kapillarität in porösen Feststoffsystemen,* Springer Verlag, Berlin–Heidelberg–New York 1982.
17. R. Polke: *Zur Haftung zwischen Festkörpern bei erhöhten Temperaturen,* Dissertation, Technische Universität Karlsruhe, 1971.
18. H. Krupp: *Static Electrification,* The Institute of Physics, London–Bristol 1971.
19. H. Krupp, *Adv. Colloid Interface Sci.* **1** (1967) 111 – 234.
20. St. Moser: *Van der Waals- und elektrostatische Haftkräfte,* Dissertation, Technische Universität Karlsruhe, 1976.
21. D. Langbein: "Van der Waals Attraction in and between Solids," *Festkörperprobleme* **13** (1973) 85 – 108.
22. H. C. Hamaker, *Physika* **4** (1937) 1058 ff.
23. E. M. Lifschitz, *Sov. Phys. JETP (Engl. Trans.)* **2** (1956) 73 ff.
24. D. M. Newitt, J. W. Conway-Jones, *Trans. Inst. Chem. Eng.* **36** (1985) 422 ff.
25. N. G. Stanley-Wood: *Enlargement and Compaction of Particulate Solids,* Butterworths, London 1983.
26. W. Pietsch, H. Rumpf, *Chem. Ing. Tech.* **39** (1967) 885 – 893.

27. H. Schubert, *Chem. Ing. Tech.* **45** (1973) 396 – 401.
28. H. Schubert, *Chem. Ing. Tech.* **46** (1974) 333 – 334.
29. H. Schubert: *Untersuchungen zur Ermittlung von Kapillardruck und Zugfestigkeit von feuchten Haufwerken aus körnigen Stoffen,* Dissertation, Technische Universität Karlsruhe, 1972.
30. W. Schütz, H. Schubert, *Chem. Ing. Tech.* **52** (1980) 451 – 453.
31. H. Rumpf, *Chem. Ing. Tech.* **30** (1958) 144 – 158.
32. W. Pietsch: *Festigkeit und Trocknungsverhalten von Granulaten,* Dissertation, Technische Universität Karlsruhe, 1965.
33. J. Charè: *Trocknung von Agglomeraten bei Anwesenheit auskristallisierender Stoffe,* Dissertation, Technische Universität Karlsruhe, 1976.
34. C. E. Capes, *Powder Technol.* **4** (1970 – 1971) 77 – 82.
35. J. Frenkel, *J. Phys.* **9** (1945) 385 – 391.
36. B. Pines, *Z. Techn. Fiz.* **16** (1946) no. 6, 137.
37. G. Kuczynski, *J. Appl. Phys.* **20** (1949) 1160 – 1163.
38. H. Rumpf, K. Sommer, K. Steier, *Chem. Ing. Tech.* **48** (1976) 300 – 307.
39. F. Steiner et al., *Polym. Prepr. Am. Chem. Soc. Div. Polym. Chem.* **11** (1970) no. 2, 1168 – 1175.
40. Wei Hsuin Yang, *J. Appl. Mech.* **33** (1966) 395 – 401.
41. H. Hertz: *Gesammelte Werke,* Barth, Leipzig 1895.
42. B. Dahneke, *J. Colloid Interface Sci.* **40** (1972) no. 1, 1 – 13.
43. F. Faulhaber: *Einfluß der Mikroplastizität auf das Bruchverhalten von Quarzkugeln,* Dissertation, Technische Universität Karlsruhe, 1966.
44. W. Schütz, H. Schubert, *Chem. Ing. Tech.* **48** (1976) MS 375 – 376.
45. B. V. Derjagin, L. D. Landau, *Acta Physicochim. USSR* **14** (1941) 633 ff.
46. E. J. W. Verwey, J. Th. G. Overbeck: *Theory of the Stability of Cyophobic Colloids,* Elsevier, Amsterdam 1948.
47. W. Pusch: *Agglomeration mit polymeren Flokkungsmitteln,* Dissertation, Technische Universität Karlsruhe, 1982.
48. K. Leschonski, Kurs "Partikelgrößenanalyse" 1984.
49. K. Sommer, R. Fath: "Automatische Bildanalyse für die Qualitätskontrolle in der Lebensmittelindustrie," *ZFL* **38** (1987) no. 7, 290 – 292.
50. L. Pfalzer, W. Bartusch, R. Heiss, *Gordian* **74** (1974) 80 – 83, 126 – 129.
51. R. Polke, W. Herrmann, K. Sommer, *Chem. Ing. Tech.* **51** (1979) 283 – 288.
52. E. A. Flood: *The Solid/Gas Interface,* Marcel Dekker, New York 1967.
53. A. Baiker, W. Richarz, *Chem. Ing. Tech.* **49** (1977) 399 – 403.
54. W. W. Washburn, *Proc. Natl. Acad. Sci. USA* **68** (1971) 115 – 116.
55. *Powder Technol.* **9** (1974) no. 4 (special number).
56. B. Koglin, *Proc. 1st Int. Conf. on Particle Technology,* Chicago, Aug. 1973, pp. 272 – 278.
57. B. Koglin, *Powder Technol.* **17** (1977) 219 – 227.
58. W. Pusch, B. Koglin, *Chem. Ing. Tech.* **55** (1983) 137 – 139.
59. H. Rumpf, W. Herrmann, *Aufbereit. Tech.* **11** (1970) no. 3, 117 – 127.
60. H. Reichert: "Vortrag zum Jahrestreffen der Verfahrensingenieure," Stuttgart 1977.
61. H. Schubert, *Chem. Ing. Tech.* **8** (1970) 541 – 545.
62. H. Schubert, *Powder Technol.* **11** (1975) 107 – 119.
63. U. Kaluza, *DEFAZET Dtsch. Farben Z.* **29** (1975) 102 – 116.
64. R. Polke, R. Rieger, *Chem. Ing. Tech.* **50** (1978) 149 – 154.
65. H. Reichert: *Desagglomeration organischer Farbpigmente in Couetteströmungen,* Dissertation, Technische Universität Karlsruhe, 1973.
66. W. Mohr, *Milchwissenschaft* **15** (1960) 215 – 222.
67. J. J. Mol, *Int. Dairy Fed. E-Doc 90,* Paris 1973.
68. L. Pfalzer, W. Bartusch, R. Heiss, *Chem. Ing. Tech.* **45** (1973) 510 – 516.
69. H. Schubert, *VT Verfahrenstechnik* **12** (1978) 296 – 301.
70. H. Schubert, B. Bauer, A. Löflath, paper read at the ZDS Instantisieren, Solingen 1986.
71. H. Schubert, *Chem. Ing. Tech.* **47** (1975) 86 – 94.
72. H. Schubert, *ZFL* **36** (1985) no. 3, 149 – 152.
73. A. Krefeld, J. M. Van der Vertroog, *Neth. Milk Dairy J.* **17** (1963) 209 – 232.
74. C. E. Capes, P. V. Danckwarts, *Trans. Inst. Chem. Eng.* **43** (1965) T 116 – T 130.

75. P. C. Kapur, D. W. Fürstenau, *Ind. Eng. Chem. Process Des. Dev.* **5** (1966) 5 – 10.
76. P. C. Kapur, D. W. Fürstenau, *Ind. Eng. Chem. Process Des. Dev.* **8** (1969) 56 – 62.
77. M. Smoluchowski, *Z. Phys. Chem. (Leipzig)* **92** (1918) 129 – 168.
78. W. Herrmann, K. Sommer, *Preprints 3rd Int. Symposium Agglomeration,* Nürnberg 1977.
79. W. Dötsch: *Agglomerationskinetik zur Stimulation von Agglomerationsprozessen im Agglomeriesteller,* Dissertation, Technische Universität München, 1987.
80. H. Klatt, *Zem. Kalk Gips* **11** (1958) 144 – 154.
81. H. B. Ries, *Aufbereit. Tech.* **16** (1975) 639 – 646.
82. U. N. Bhrany, R. Manz, *Aufbereit. Tech.* **18** (1977) 641 – 647, *Zem. Kalk Gips* **23** (1970) 407 – 412.
83. M. Papadakis, J. P. Bombled, *Rev. Mat. Construct.* **549** (1961) 289 – 299.
84. W. Pietsch, *Chem. Tech. (Leipzig)* **19** (1967) 259 – 266.
85. H. E. Rose, R. M. E. Sullivan, *Ball, Rube and Rod Mills,* Constable, London 1957.
86. J. Nordengren: *Granulation of phosphate fertilizers–theory and practice,* Proc. Fertil. Soc., London 1947.
87. H. Leuenberger, G. Imanidis, *Chem. Ing. Tech.* **56** (1984) 322 – 323.
88. J. R. Farzand, I. E. Puddington, US 3 528 809, 1970.
89. J. Werther, *Powder Technol.* **15** (1976) 155 ff.
90. N. A. Shakhova, B. G. Yevdokomov, N. M. Ragozina, *Process Technol. Int.* **17** (1972) 946 – 947.
91. P. E. Wurster, *J. Am. Pharm. Assoc.* **48** (1959) 451 ff.
92. E. P. Nichols, GB 1 039 177, 1966.
93. H. Schwanghardt, *Aufbereit. Tech.* **12** (1969) 713 – 722.
94. H. Schwanghardt, *Aufbereit. Tech.* **12** (1973) 818 – 830.
95. W. Herrmann, *Chem. Ing. Tech.* **48** (1976) 333.
96. R. Herbener, *Chem. Ing. Tech.* **59** (1987) 112 – 117.
97. I. A. J. Brümmelkuis, *Milchwissenschaft* **30** (1975) 75 – 80.
98. J. Pisecky, V. Westergaard, *Milchwissenschaft* **26** (1971) 557 – 560.
99. H. G. Kessler: *Lebensmittelverfahrenstechnik,* Verlag A. Kessler, Freising 1976.
100. G. Kluge, R. Heiss, *Verfahrenstechnik (Mainz)* **1** (1967) 251 – 260.
101. U. Bröckel, Kurs "Agglomeration" Weihenstephan 1987.
102. N. K. Stuart, World Coal, August 1979.
103. G. G. Bemer, F. J. Zuiderweg, Int. Symposium *Fineparticles Processing,* Las Vegas 1980.
104. J. R. Farnand, *J. Sep. Process Technol.* **3** (1982) 1 – 15.
105. S. Bosa: Dissertation, University of Technology of the Netherlands, Delft 1983.
106. K. V. S. Sastry, Agglom. 77, Am. Inst. of Minery Met. and Petroleum Eng. 1977.
107. G. G. Bemer: Dissertation, University of Technology of the Netherlands, Delft 1983.
108. Y. Kawashima, C. E. Capes, *Powder Technol.* **10** (1974) 85 – 92.
109. C. E. Capes, J. P. Sutherland, *Ind. Eng. Chem. Process Des. Dev.* **6** (1967) 1.
110. J. R. Farnand, H. M. Smith, I. E. Puddington, *Can. J. Chem. Eng.* **39** (1961) 4.
111. B. Bogenscheider, H. J. Behrenbeck, K. H. Kubika, *Glückauf* **112** (1976) 12.
112. K. Sartor, Seminar "Verdichten von Stäuben," Graz 1977.
113. K. H. Lüdde, K. Kawakita, *Pharmazie* **21** (1968) 393 ff.
114. K. Kawakita, K. H. Lüdde, *Powder Technol.* **4** (1970/1971) 61 ff.
115. H. Leuenberger, E. N. Hiestand, H. Sucker, *Chem. Ing. Tech.* **59** (1987) 77 – 78.
116. R. Spang: Dissertation no. 5081, ETH Zürich 1973.
117. E. Ryshkewitch, *J. Am. Ceram. Soc.* **6** (1953) 65 ff.
118. W. John, *Ullmann,* 4th ed., **14,** 317.
119. R. Zisselmar, Seminar: "Kompaktieren," TA Wuppertal 1986.
120. W. Herrmann, *Das Verdichten von Pulvern zwischen zwei Walzen,* Verlag Chemie, Weinheim 1973.
121. J. R. Johanson, *Trans. ASME* (1965) 12.
122. J. R. Johanson, *Proc. of 9th Biennal Briquetting Conference,* 1965 (IBA).
123. J. R. Johanson, *Conference of International Briquetting Association,* 27. – 29. Aug. 1969, Sun Valley, Idaho.
124. A. W. Jenike, *Exp. Station Bulletin,* **No. 108, vol. 52,** Oct. 1961.
125. O. Molerus: *Schüttgutmechanik,* Springer Verlag, Heidelberg 1985.
126. G. Hauser, K. Sommer, CHISA, Prag, Sept. 1984.
127. K. H. Sartor, H. Schubert, *Chem. Ing. Tech.* **50** (1978) 708 – 709.
128. K. H. Bauer, K. H. Frömming, C. Führer: *Pharmazeutische Technologie,* Thieme Verlag, Stuttgart 1986.

129. K. Münzel, J. Büchi, O. E. Schultz: *Chemisches Praktikum,* Thieme Verlag, Stuttgart 1959, p. 727.
130. H. Sucker, P. Fuchs, P. Speiser: *Pharmazeutische Technologie,* Thieme Verlag, Stuttgart 1978.
131. N. G. Stanley-Wood: *Enlargement and Compaction of Particulate Solids,* Butterworths, London 1983.
132. D. Train, *J. Pharm. Pharmacol.* **8** (1956) 745.
133. K. Sator, *Chem. Ing. Tech.* **50** (1978) 708 – 709.
134. K. H. Lüdde, K. Kawakita, *Pharmazie* **21** (1968) 393 ff.
135. K. Kawakita, K. H. Lüdde, *Powder Technol.* **4** (1970/1971) 61 ff.
136. R. Spang: Dissertation no. 5081, ETH Zürich 1973.
137. E. Ryshkewitch, *J. Am. Ceram. Soc.* **6** (1953) 65 ff.
138. K.-H. Sartor, H. Schubert: Vortrag auf dem Jahrestreffen der Verfahrensingenieure, Stuttgart 1977.
139. W. Herrmann, *Chem. Ing. Tech.* **43** (1971) 418 – 425.

Solids Handling

See also Solids Technology, Introduction

Roy D. Marcus, Cargo Carriers Ltd., Isando, Republic of South Africa

F. Rizk, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

Stephen J. Meijers, University of Pretoria, Pretoria, Republic of South Africa

1. Introduction 2
2. System Selection 3
2.1. Selection of a Transportation System 3
2.2. Flow Properties of Bulk Solids . . . 4
2.3. Properties of Bulk Solids 5
3. Flow of Solids from Bins and Silos . 7
3.1. Common Flow Problems 7
3.2. Flow Characteristics 7
3.3. Stress Distribution in a Silo 8
3.4. Measurement of Flow Characteristics 8
3.5. Shear Cell Analysis 9
4. Mechanical Conveyors 9
4.1. Belt Conveyors 9
4.2. Oscillating Conveyors 11
4.3. Screw Conveyors 12
4.4. Bucket Elevators 12
4.5. En Masse Conveyors 13
4.6. Shuttle Belt Conveyors 15
4.7. Apron Feeders 15
4.8. Rotary Table Feeders 15
5. Hydraulic Conveying 16
5.1. Pump Selection 16
5.1.1. Centrifugal Pumps 16
5.1.2. Reciprocating Pumps 16
5.2. Classification of Slurries 17
5.2.1. Settling Slurries 17
5.2.2. Nonsettling Slurries 19
5.3. Principles of Design Calculations . . 20
6. Pneumatic Conveying 21
6.1. Classification of Pneumatic Conveying Systems 22
6.1.1. State Diagram 22
6.1.2. Vertical and Horizontal Flow 24
6.2. Pressure Drop Components 24
6.3. Modes of Pneumatic Conveying and Product Characteristics 25
6.4. Pneumatic Conveying Systems . . . 25
6.4.1. Positive Pressure System 26
6.4.2. Negative Pressure System 26
6.4.3. Combined Positive and Negative Pressure Systems 26
7. References 26

Symbols

$A = \pi D^2/4$ cross-sectional area of pipe or product, m^2
$A*$ cross-sectional area of a particle normal to flow, m^2
ΔA effective cross-sectional area
c solids velocity, m/s
C_d delivered concentration
C_{vd} concentration delivered by volume
$C_D = \frac{2F_D}{\varrho w^2 A^*}$ drag coefficient
D pipe inner diameter, mm
D_{min} minimum cone diameter, m
d particle diameter, mm
D_s shaft diameter, mm
d_v volume equivalent particle diameter, mm
f friction factor
f_c unconfined yield stress, N/m^2
$f_{c,crit}$ critical unconfined yield stress, N/m^2
F_D drag resistance, N
$\dot{G}$ solid mass flow rate, kg/h, t/h
K constant
k effective pipe wall roughness, mm
L, ΔL length, length of element, m
L_v vertical length, m
$\dot{m}$ mass flow rate, kg/h
$M_I = (\sigma_1 + \sigma_2)/2$
$M_{II} = f_c/2$
n flow index for power law fluid or number of revolutions per minute
p pressure, kPa
Δp_L, Δp_z, Δp_{tr} pressure change due to air, solids, transportation, respectively, kPa
Δp pressure difference, kPa
$Re = vD/\nu$ Reynolds number related to pipe diameter

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_08

$Re_p = vd/\nu$ Reynolds number related to particle diameter
$Re*$ generalized Reynolds number
Re_{crit} critical Reynolds number
s pitch of screw, m, or slip or velocity ratio
S_p surface area of particle, m^2
U shape factor
$V_p = \pi d^3/6$ volume of particle, m^3
$\dot{V}_p = \dot{G}/\varrho*$ solids volume flow rate, m^3/s
v flight or superficial velocity, m/s
$\bar{v}$ average fluid velocity, m/s
v_m velocity of mixture, m/s
v_D deposition velocity, m/s
v_{min} minimum conveying velocity, m/s
v_s gas conveying velocity at point of particle settling velocity, m/s
$w = (v - c)$ relative velocity, slip velocity, m/s
β trough angle, degrees
γ rate of shear
δ angle of surcharge, degrees
$\varepsilon = 1 - \frac{\varrho^*}{\varrho_p}$ voidage
ε_H head loss gradient
η dynamic viscosity, Ns/m^2
η_a apparent viscosity, Ns/m^2
η_e effective viscosity, Ns/m^2
η_p plastic viscosity, Ns/m^2
θ cone angle, degrees
ν kinematic viscosity, m^2/s
ϱ air (gas) density, kg/m^3
$\varrho* = \varrho_p(1 - \varepsilon)$ conveying bulk density, kg/m^3
ϱ_f fluid density, kg/m^3
$\varrho = \varrho_p - \varrho_f$ density difference between particle and fluid, kg/m^3
ϱ_B bulk density, kg/m^3
ϱ_p particle density, kg/m^3
σ normal stress, N/m^2
$\bar{\sigma}_1$ bridging stress, N/m^2
σ_1 principal or major consolidating stress, N/m^2
τ shear stress, N/m^2
τ_o shear stress at pipe wall, N/m^2
τ_y yield shear stress for Bingham plastic, N/m^2
φ_e effective angle of internal friction, degrees
φ_w angle of wall friction, degrees
φ_i kinematic angle of friction, angle of internal friction, degrees
φ filling efficiency
$\psi = \pi d_v^2/S_p$ sphericity

1. Introduction

Handling bulk materials plays an important role in the chemical industry. The effective operation of most modern chemical manufacturing plants depends on the ability to move raw materials, semiprocessed and fully-processed products, both within the manufacturing environment and ultimately to the end user.

Bulk materials handling comprises a wide range of techniques incorporating a number of handling modes. An attempt is made here to highlight some of these modes by providing some basis on which both the method of operation and the basic conveying requirements and parameters are given. In the limited space available, it is not possible to do justice to any one system or to provide full details on all aspects of the various techniques. The reader is advised to obtain more extensive information if a full system design is required; adequate references are listed to facilitate further reading.

2. System Selection

Selection of a specific handling system for bulk materials requires a knowledge of the physical and chemical properties of the material to be handled. Frequently, both the material characteristics and the prevailing plant conditions play an important role in determining the flow behavior of the product, thereby influencing the type of equipment to be selected.

It is thus essential to have prior knowledge of the product flow behavior before any reliable system selection can be attempted. As particle characterization techniques are improved, so the selection of suitable equipment to handle the material is simplified. Nevertheless, there are still various conveying techniques which require pilot plant testing in order to obtain the necessary flow parameters for reliable system selection and design. It is frequently essential that a representative sample (of the product to be conveyed) be tested under realistically simulated conditions.

Two of the most salient features for selecting a suitable handling system are a knowledge of the *distance* and the *throughput* required of

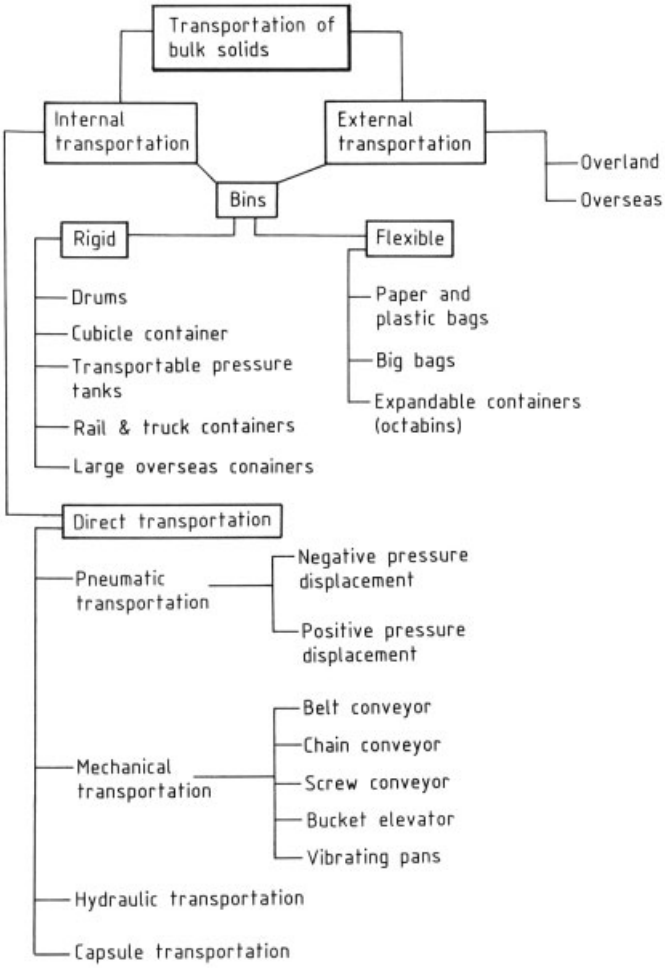

Figure 1. Categorization of the transportation system

the system. When specifying distance, details of the *route profile* are important: elevation, horizontal and vertical conditions to be traversed, head room, and route flexibility are critical issues for effective system selection. The handling system will often incorporate storage facilities, which also have to be carefully designed to meet a number of requirements.

In terms of solid mass flow rate, it must be clearly identified whether the system can operate in a batch mode or if continuous flow is mandatory, together with an estimation of peak loading. Future expansion must also be taken into account; this could include, for example, increasing the length or the flow rate.

2.1. Selection of a Transportation System

Essentially two categories of bulk materials handling systems exist: indoor (or in-plant) transportation and long distance external transportation. The type of equipment selected can sometimes be suitable for both types of operation, but internal transportation systems are usually associated with specialized equipment which is normally designed to meet fairly stringent process requirements. The categorization of the transportation system is illustrated by Figure 1. Within the overall classification of bins there is a number of techniques used to transport, weigh, load, and unload these systems (Fig. 2).

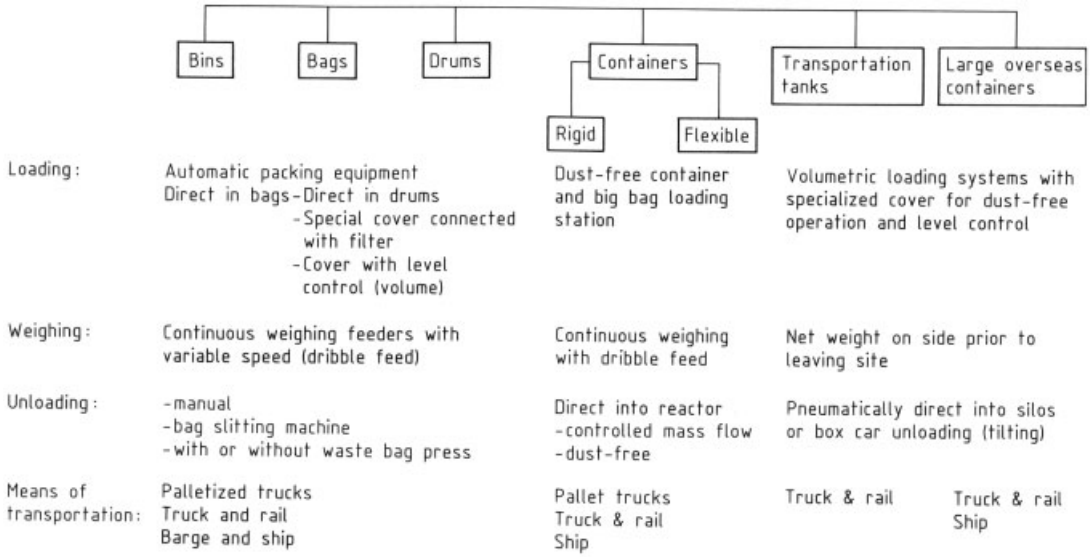

Figure 2. Loading, unloading, and transportation techniques

The ultimate selection of suitable hardware depends on the conveying rate. A basis for selection has been made for so-called continuous conveying systems (Fig. 3). As an example, the appropriate conveying system for a given flow rate as a function of the conveying length can be selected, using Figure 3, as follows:

1) For a solid mass flow rate $\dot{G} = 300$ kg/h and conveying lengths $L = 2.5$, 25, and 250 m, the following results are found:
   $L = 2.5$ m: systems (b), (d), and (f) may be used;
   $L = 25$ m: system (a), (b), and (d) are applicable;
   $L = 250$ m: only system (a) can be considered.
2) For a solid mass flow rate $\dot{G} = 10\,000$ kg/h and a length $L = 200$ m, Figure 3 shows that systems (a), (b), and (c), i.e., a pneumatic conveyor, air-activated gravity conveyor, or a belt or chain conveyor would be acceptable. The final selection, however, depends on the route which the conveyor is to traverse. If a requirement to raise the material vertically for, say, $L_v = 35$ m, is then included in the route, systems (b) and (d) must be combined with system (e).

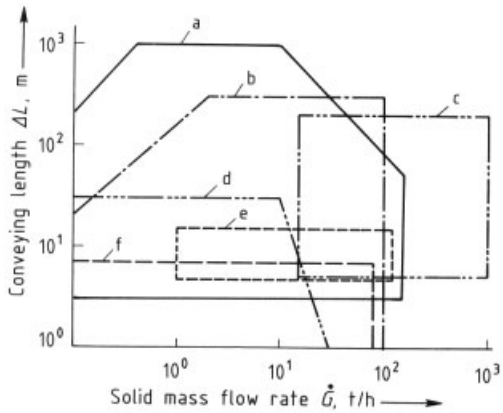

Figure 3. Transportation system selection chart:
a) Pneumatic conveyors; b) Air-activated gravity conveyors; c) Belt and chain conveyors; d) Vibrating pans; e) Elevators; f) Screw conveyors and vibrating conduits

The final selection will also depend on laboratory tests which could indicate that the product might not be ideal for, say, a pneumatic conveying system.

2.2. Flow Properties of Bulk Solids

The flow characteristics of bulk solid materials depend on their physical and chemical properties, on the pressure to which the solids are sub-

jected, and the duration of the pressure. When fine particles are pressurized, the particles are drawn closer together, resulting in deformation (elastic/plastic) with a possible simultaneous reduction in porosity and voidage; these effects can drastically change the flow properties of the solid.

The main characterizing factors are the angle of friction and the so-called flow function; both can de determined by means of a shear cell analysis, which is discussed later (Section 3.5).

Product Grouping. In the selection of conveying systems, the flowability of the product yields two main categories of bulk solids:

Group I. This group includes free-flowing materials, i.e., noncohesive products: those that do not undergo any plastic deformation when subjected to high pressures. When the load is removed, the particles return to their original condition in terms of both shape and flow characterisitics.

Group II. All products which undergo plastic deformation when subjected to external pressure: cohesive products. The degree of deformation is strongly influenced by both temperature and moisture. When the load is removed the particles do not regain their original shape, thereby yielding poor flow conditions.

Fluidization Characteristics. The ability of the material to fluidize and whether the product has an affinity to trap air or gas is of major importance when designing and selecting handling and storage systems. Note that in various cases noncohesive products flow extremely well and tend to flood when withdrawn from bins.

Angle of Friction. There are two angles of friction, the internal angle φ_{i} and the hopper wall-product angle φ_w. The internal angle of friction yields the friction between the particles being sheared. The angle of friction between the hopper wall and the product is critical in determining whether the product will flow or not.

Flow Function. For both short and long residence times in a silo, the flow function must be taken into account. Additional complications occur when the product is stored at elevated temperature or when humidity could influence the moisture content of the product.

2.3. Properties of Bulk Solids

An attempt will be made here to highlight some of the more important physical properties of bulk solids which influence flowability.

Voidage. All bulk solids, irrespective of their shape, are packed together in a particular volume, leaving areas of free space. The percentage of the total volume not occupied by the particles is referred to as voidage ε. The actual volume of solids is then $(1-\varepsilon)$. The magnitude of the voidage varies from bulk solid to bulk solid and depends considerably on particle shape. For spheres of uniform size, the voidage $\varepsilon = 0.48$.

Bulk Density. An important parameter required for the design of both storage and conveying systems is the bulk density ϱ_{B}, defined as the ratio of the mass of the product to the total volume occupied by the solid.

There are two types of bulk density: loose bulk density and packed bulk density. Each is a function of the measuring technique and will yield results which depend on the ultimate process to which the material is being subjected. In the former the measurement is made when the product is poured into a container of known volume. In packed bulk density the product is tamped in the container, thereby reducing the voidage.

Particle Density. The actual particle density ϱ_{p} is defined as the ratio of the mass of the individual particle to its volume. Numerous techniques are available for measuring particle density. In the case of fine particles these techniques are relatively complex.

In terms of the definitions of voidage and bulk density ϱ_{B} above, particle density can also be defined as:

$$\varrho_{\mathrm{B}} = \varrho_{\mathrm{p}}\,(1-\varepsilon)$$

Particle Size and Shape. The design of a number of conveying systems relies on knowledge of the particle size and in many cases the so-called particle size distribution. For a number of

products the determination of particle size does not present a problem: e.g., for pellets, spheres, and cubes there is obviously no doubt about the particle size.

Difficulties arise when trying to determine the particle size of irregularly-shaped solids, and it is usual to determine the size distribution. Several techniques are used to determine particle size distribution (→ Particle Size Analysis and Characterization of a Classification Process).

The data obtained from sizing are then plotted on a cumulative graph in which the cumulative percentage mass passing is plotted as a function of the particle size (Fig. 4). From the cumulative graph it is possible to find the 50 % passing value, which is then taken as the average diameter of the particles.

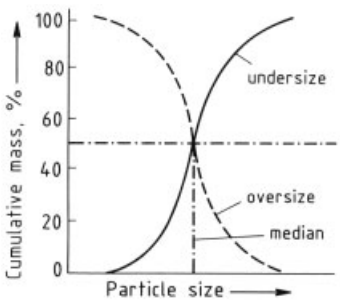

Figure 4. Typical particle size analysis

Particle shape determination is more complex and in many cases knowledge of the shape is an important consideration. Two materials with identical size distribution, density, and chemical composition may behave completely differently as a result of a variation in particle shape [1].

Although much research is being conducted on particle shape determination [2], [3] the most common shape factor is the sphericity factor:

$$\psi = \frac{\text{Surface area of a sphere with same volume as particle}}{\text{surface area of particle}} = \frac{\pi\,(6V_\mathrm{p}/\pi)^{2/3}}{S_\mathrm{p}}$$

where V_p and S_p are the particle volume and surface area, respectively. From the above definition, $\psi = 1$ for a spherical particle.

Cohesion and Agglomeration between Particles. Cohesion between particles is of fundamental importance when determining the flow properties of the bulk solid, which can be described as either free flowing or cohesive. Cohesion can be defined as the resistance of a bulk solid to shear at zero compressive normal stress [4].

A number of mechanisms promote cohesive forces in bulk solids. In particular, van der Waal's forces, flow turbulence [5], static electrical effects, moisture content, and particle interactions are all important contributors.

Cohesive forces are the most important, but the tendency of particles to agglomerate can also present the designer with major problems: e.g., if a system is designed to meet a particular particle size and due to agglomeration the resulting particle size is significantly larger.

The cohesion of a material can be obtained from shear cell analysis (Section 3.5.).

3. Flow of Solids from Bins and Silos

For the successful operation of any materials handling system the flow of solids from bins and silos must be controlled. This section highlights some of the important features relating to the suitable design of such storage facilities.

3.1. Common Flow Problems

Two common flow problems are associated with the incorrect design of a silo or hopper: a) bridging; b) rat holing.

Bridging results in formation of an arch or a bridge across the opening of a hopper. The strength of the bridge is such that it supports the total load of solids above it. Bridging occurs when the outlet of the silo is too small on the basis of a calculated minimum diameter D_min required to effect uninterrupted solids flow, although the cone angle θ might be of the same magnitude as that obtained from silo design theory. Bridging can also occur when θ is too large.

Rat-holing is characterized by the formation of a channel through the solid mass and is the result of incorrect selection of the cone angle. Rat-holing will occur even though the hopper opening conforms to the correct specifications.

The various bridging and rat-holing phenomena are shown in Figure 5.

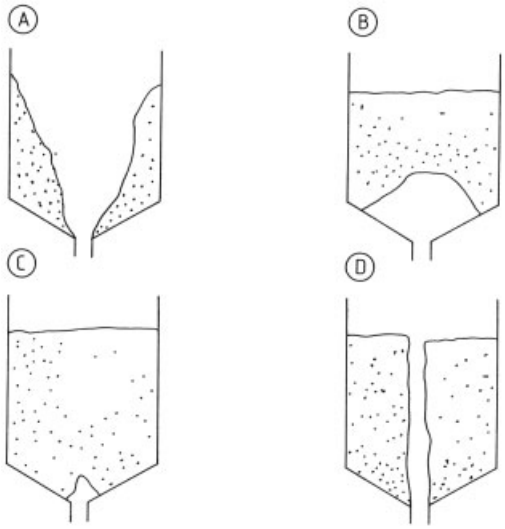

Figure 5. Typical flow problems from silos
A) Clinging; B) Arching; C) Bridging; D) Rat-holing

3.2. Flow Characteristics

The flow of a product from a storage bin conforms essentially to two types of flow: (A) mass flow or (B) funnel flow, the characteristics of which are depicted in Figure 6.

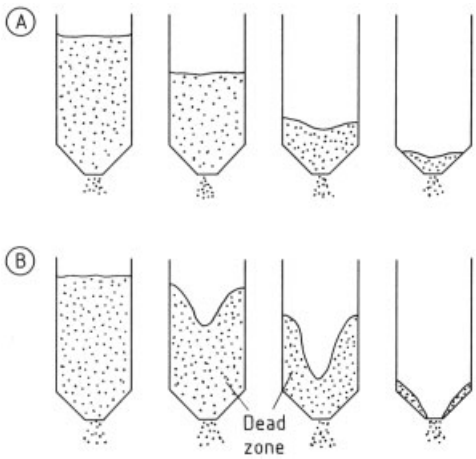

Figure 6. Flow of a solid from a silo A) Mass flow; B) Funnel flow

Mass flow from a bin implies that the total volume of the stored solids is in motion; uniform and steady state flow can be attained. A mass flow bin is devoid of channeling, surging, flooding, hang-ups, or bridging. The flow is also independent of the head of the stored solids and there is minimum consolidation of product, no dead space, and minimal degradation or spoilage. The flow from a mass flow bin is aptly described as first in-first out.

Funnel flow is best understood by considering the flow from a bin with a flat bottom or gentle slope: flow occurs in a channel in the center of the bin and is always from the top of the stored solids to the center. During the flow of the solids through the channel, a dead zone occurs in all other parts of the bin. Funnel flow leads to first in-last out and has disadvantages such as flooding, material degradation, and consolidation.

The ultimate selection between mass flow and funnel flow must, in addition to taking into consideration the product characteristics and the process, recognize that a mass flow bin is generally more costly and will occupy considerably more head room.

Factors Influencing Flow. Three essential factors must be considered when designing a storage hopper or bin:

1) *Geometric form of the hopper*: the elements which must be considered include a) cone angle; b) size of outlet; c) shape (circular or rectangular); d) hopper construction material.
2) *Product characteristics*: a) particle size and shape; b) particle size distribution; c) particle density and bulk density; d) cohesiveness of the product; d) fluidizability; f) floodability; g) deaeration characteristics.
3) *Additional factors*: a) influence of humidity; b) temperature of product and process; c) storage time; d) ambient conditions.

3.3. Stress Distribution in a Silo

In order to appreciate the various forces imposed by the stored solids on the wall of the silo, the stress distribution for a typical conical bottom silo will be discussed using Figure 7. The stress distribution for a silo carrying a liquid is shown as the hydrostatic pressure. Clearly, the stress distribution is uniform and a function of the liquid level in the silo.

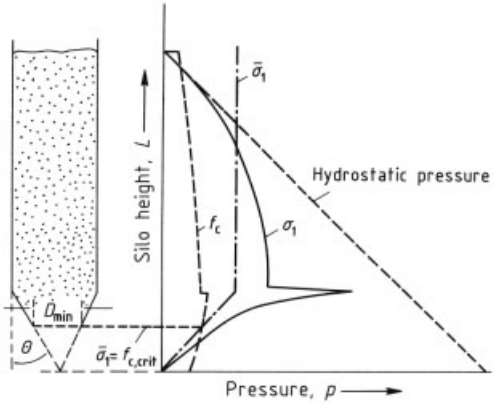

Figure 7. Stress distribution in a silo
a) Principal stress, σ_1; b) Unconfined yield stress, f_c; c) Bridging stress, $\bar{\sigma}_1$; d) Critical unconfined yield stress $f_{c,\,crit}$; θ = cone angle; D_{min} = min. cone diameter

Substituting the liquid with a granular product results in a stress distribution known as the principal stress or major consolidating stress curve. At the highest level in the silo the stress distribution for the solid is similar to that of the liquid, but the curve for the former deviates with increasing depth because of the friction between the particles and the wall of the silo.

Above a certain head of solid the stress distribution remains constant until the transition zone, in which the silo changes shape from a cylinder to a cone. Beyond the transition the stress in the cone decreases linearly towards the outlet. Theoretically, the stress reaches zero at the peak of the cone.

An important feature of the storage of products in a silo is the advent of consolidation. The stress arising from consolidation is referred to as the unconfined yield stress f_c (see Fig. 7).

A critical component in the design of a silo is knowledge of the minimum opening to ensure free flow of the solid. The determination of this minimum diameter depends on the so-called critical unconfined yield stress $f_{c,\,crit}$, which occurs at the intersection of the so-called bridging stress $\bar{\sigma}_1$ and the unconfined yield stress.

3.4. Measurement of Flow Characteristics

The final design procedures for silos and hoppers depend on the acquisition of certain fundamental flow properties of the material. Although there are several techniques available to provide design information, perhaps the most common and successful technique is the Jenike shear cell which enables evaluation of the following [6]:

1) The effective angle of internal friction, φ_e
2) The kinematic angle of friction, φ_i
3) The major consolidating stress, σ_1
4) The unconfined yield stress, f_c
5) The bulk density of the product as a function of bulk solids head.

Other techniques include the ring shear cell developed by Walker, the powder test of Hosakawa, and the torsional shear cell of Peschl [7].

3.5. Shear Cell Analysis

Once an accepted testing procedure has been performed on the bulk solids using a shear cell, the observations obtained are plotted on a shear stress versus a normal stress diagram (Fig. 8).

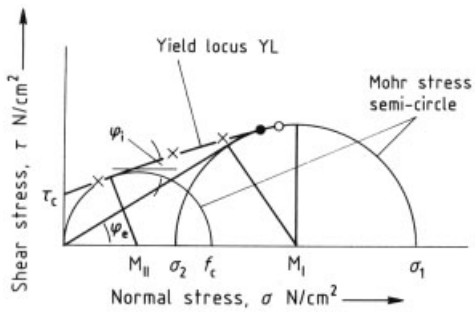

Figure 8. Typical shear-stress diagram

From the shear cell analysis it is possible to draw three yield loci, obtained by a Mohr semicircle technique as shown in Figure 8. When the large Mohr semicircle is drawn tangential to the obtained yield locus and passed through the initial consolidating load, the major consolidating stress σ_1 is obtained. The small Mohr semicircle passes through the origin and the tangent to the yield locus curve yields the unconfined stress f_c. In addition, from tests carried out using the plate of the material to be used for the construction of the silo, the shear cell analysis permits the construction of another diagram from which the friction factor between the bulk solid and the

silo construction material is obtained. Once the various flow charcteristics have been obtained the exact design of a silo, in which all the major dimensions are specified, can be carried out.

4. Mechanical Conveyors

Mechanical conveying techniques are the most widely used form of materials handling in the chemical industry. Although a large number of devices are available, most systems conform to the basic elements in which either belts, chains, or moving flights are used to move material.

Mechanical conveyors have distinct advantages in terms of the ability to effect accurate control in the monitoring of material from one process to another. Further, by careful selection of construction materials, it is possible to design systems to meet a wide variety of operating conditions including corrosive environments and extremes of temperature.

4.1. Belt Conveyors

Belt conveyors are the most widely used and versatile mode of mechanical conveying systems. They are able to handle larger tonnages over longer distances and at a lower cost than most forms of conveying system. A belt conveyor is an endless length of flexible material stretched between two drums and supported at intervals. The first belt conveyor recorded was that used for the transport of wheat almost 200 years ago [8, p. 127].

The belt conveyor system consists of several main components:

1) *Belt.* The belt consists of a carcass and a cover; numerous forms of construction are used.
2) *Idlers.* Troughing idlers enable the belt load to be distributed, and the belt volume and mass to be optimized. Idlers are strategically located along the length of the conveying system to ensure maximum belt support whilst minimizing the number of idlers.
3) *Drive Unit.* The belt is driven by one drum at the turnabout side of the conveyor. The drive may be positioned at either the feed end or the head end.
4) *Loading and discharging.* The belt is usually loaded by a gravity feeder such as a hopper, a preceding conveyor, or a vibratory feeder. The art of loading is to provide a steady flow of material onto the belt. Discharge of the bulk material is usually by gravity dumping, but the trajectory of the material must be taken into account.
5) *Belt cleaners.* This accessory to the belt conveyor is probably the most important. There are several types of belt cleaners designed to minimize the amount of carryover, increasing the life of the conveying system.

Belt Speed. The capacity of a belt conveying system is clearly a function of belt speed. The average belt speed in practice ranges from 1 m/s to 5 m/s, although some belts operate at speeds of up to 10 m/s [8, p. 141]. The belt speed is not only a function of the belt width, but also of the type of material to be conveyed (see Fig. 9).

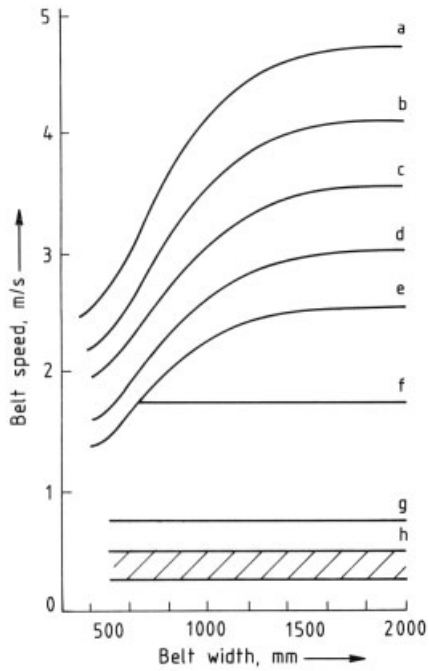

Figure 9. Conveyor belt speed as a function of belt width
Conveyed material: a) Fine: free-flowing, nonabrasive; b) Fine: mildly abrasive or lumpy; c) Granular: abrasive or lumpy, mildly abrasive; d) Granular: very abrasive or lumpy, moderately abrasive; e) Lumpy and very abrasive; f) Belts used with belt-propelled trippers; g) Belts used with ploughs; h) For picking and feeding belts

Angle and Length of Incline. The angle of the incline and the length of the inclined portion of the conveyor may be obtained when the horizontal length and the rise of the inclined portion is known (see Fig. 10).

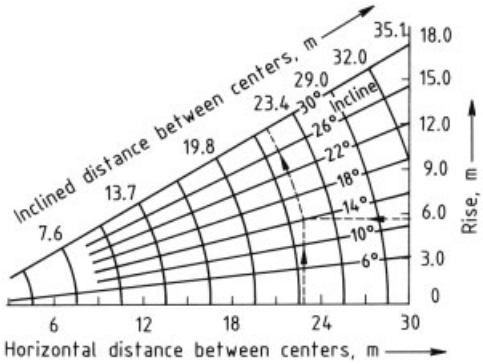

Figure 10. Inclined length of a belt conveyor

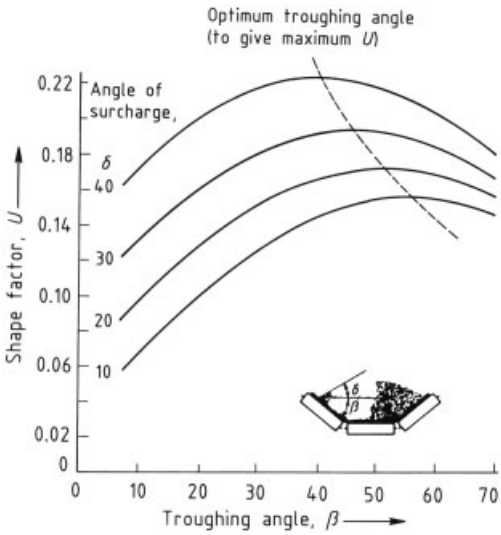

Figure 11. Shape factor for a three idler system

The use of Figure 10 is best illustrated with an example. Consider a belt conveyor with a horizontal distance between centers of 46 m and a vertical rise of 11.6 m. To use the chart for this example, divide the figures by two giving 23 m and 5.8 m, respectively. The intersection of a vertical line from the 23 m mark and a horizontal line from the 5.8 m mark occurs at a point corresponding to approximately the 14° line to a radius of 23 m. Multiplying this by 2 results in an incline length of 46 m for the conveyor. The angle is 14° for this set of conditions.

The belt width and speed are dependent on a number of factors including the capacity, angle of inclination, belt tension, average particle size, and configuration. A number of combinations are usually considered before a final decision is made on the design.

Idler Selection. A typical chart for a three idler system is given for rollers of identical length in Figure 11; the optimum trough angle β for a maximum shape factor U is indicated. Only open belt conveyors, i.e., without any capsulation, have been considered in this text. Conditions do not, however, vary for closed or pipe conveyors.

4.2. Oscillating Conveyors

The vibratory conveyor accelerates the particle in both the horizontal and vertical directions. This mode of transportation must not be confused with that of vibratory and reciprocating feeders. The major difference is the range of frequency and amplitude under which they operate, as can be seen below.

Device	Frequency, Hz	Amplitude, mm
Vibratory feeder	13 – 60	12 – 1
Vibratory conveyor	3 – 17	50 – 5
Reciprocating Conveyor	1 – 3	300 – 50

These conveyors are ideally suited to the handling of granular, free-flowing materials and hot, fine, dusty, lumpy, stringy, or other materials which are awkward to handle. The basic operation of the vibratory conveyor is shown in Figure 12 [9]. The relationship between the frequency and the amplitude of the oscillation is vitally important when tuning and selecting a vibratory conveyor.

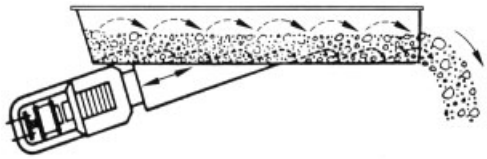

Figure 12. Arrangement of a vibratory conveyor

Some typical arrangements for several products using a vibratory conveyor are given in Table 1. This table should be used as a guide only if applied to the products mentioned. Four main types of vibratory conveyor are available.

Table 1. Typical data for vibratory conveyors [10]

Material	Approximate size, mm	Average bed depth, mm	Average transport velocity, m/s
Alumina	0.15	75	0.15
Bagasse	0.25 – 5	150	0.4
Carbon black	1.5 ∗	75	0.18
Cement clinker	6 – 10	125	0.36
Cereal	6 – 10	150	0.36
Coal	18 – 26	125	0.3
Crumb rubber	6	100	0.3
Detergent powder	0.15	75	0.25
Glass cullet	3 – 12	100	0.3
Gravel	6 – 10	125	0.33
Limestone	10.3	100	0.36
Milk powder	0.075	35	0.13
Plastic pellets	3 – 6	100	0.36
Sand-damp	0.8	100	0.4 – 0.45
Sand-dry	0.8	75	0.25 – 0.3
Salt (table)	0.4 – 0.8	50	0.3
Steel shot	1.5 – 3	50	0.36
Steel turnings	6 – 12	100	0.28
Sugar (granulated)	0.5 – 0.8	60	0.25
Tobacco	Cut	250	0.36
Wood chips	10	250	0.4

∗ Pelletized

Positive Mechanical Drive. The low frequency-high amplitude feature is predominant in this mode of mechanical drive. These features enable high loads to be transported over long distances. The drive mechanism is coupled to the trough to prevent any restrictions to movement. Because there is no restrictive movement, large vibratory forces can be transmitted to surrounding structures, requiring that the mechanical drive unit be substantially secured. Conveying speeds vary between 0.2 and 0.8 m/s.

Eccentric Mass Mechanical Drive. Comparatively high frequencies and variable amplitudes are predominant in eccentric mass conveyors. The amplitudes can vary from 1 to 10 mm depending on the load. This layout consists of a rotating single eccentric load or, more often, two masses of equal configuration placed 180° apart on a rotating shaft. The twin load can be easily adjusted to produce the desired motion.

Electromagnetic Drive. This technique relies on a number of electromagnets to provide the required motion. By applying impulses at the correct frequency to the electromagnets it is possible to establish the conveying motion. The amplitudes are generally small compared with other forms of drives; the conveying velocity is up to 0.3 m/s.

Hydraulic Drive. The advantage of using hydraulic drives (in some cases pneumatic) is that they are not susceptible to explosions, enabling vibratory conveyors to be used in areas of high fire or explosion risk. These conveyors are capable of heavier duty cycles than electromagnetic conveyors.

One of the biggest advantages of vibratory conveyors is that they can convey products such as sugar and milk powder under conditions that almost eliminate degradation. They are also well suited to abrasive products and are amenable to fine tuning and control. With the incorporation of sensing devices and a feedback loop it is possible to use these conveyors for very accurate dosing in many chemical operations.

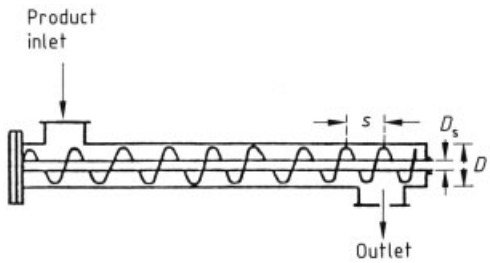

Figure 13. Arrangement of a screw conveyor

4.3. Screw Conveyors

The screw conveyor moves mass products at a controllable rate. The screw conveyor, commonly known as the Archimedes screw, was first used 2000 years ago as a water pump.

The screw conveyor consists primarily of a rotating helicord placed in a stationary trough. Material is propagated by the action of the rotating screw. Inlet and outlet ports control the

loading and delivery of the product. The helicord flight conveyor screw is made from a helix, formed from a flat bar. This helix is then mounted on a pipe or round bar. A typical screw conveyor arrangement, showing the more important parameters of such a conveyor, is illustrated in Figure 13.

There is a diversified number of uses for screw conveyors, such as controlled heating or cooling, mixing, and blending.

The mass flow rate can be calculated using the following expression:

$$\begin{aligned}\dot{G} &= \dot{V}_{\mathrm{p}}\varrho_{\mathrm{B}} \\ &= \Delta A\, s\, n\, \varrho_{\mathrm{B}}\, 60\varphi \\ &= \Delta A\, c\, \varrho_{\mathrm{B}}\, \varphi\end{aligned}$$

where $\dot{G}$ = solids mass flow rate, $\dot{V}_{\mathrm{p}}$ = bulk solids volumetric flow rate, ϱ_{B} = bulk density, ΔA = cross-sectional area available for solid flow, s = pitch, n = revolutions per minute, φ = filling efficiency, c = solids velocity = $sn/60$. The slip factor is assumed to be unity.

The filling efficiency is defined as the ratio between the volume of product in the screw conveyor at any one time to the volume which is available to convey product. The filling efficiencies are illustrated schematically in Figure 14.

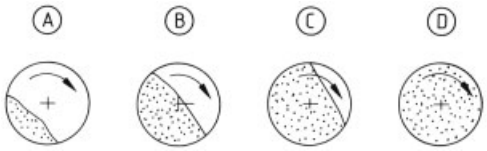

Figure 14. Screw conveyor filling efficiency
A) 25 %; B) 50 %; C) 75 %; D) 100 %

4.4. Bucket Elevators

Bucket elevators are the most commonly used device in the continuous unloading of ships, tankers, and other large storage depots. They have a distinct advantage over other methods of high payload conveyors in that they are completely maneuverable, simple, efficient, and reliable. The continuous elevator system consists of a digging element, a horizontal transfer element, and a junction element.

The design and selection of bucket elevators depends predominantly on the capacity requirements and characteristics of the material. There are two distinct categories of bulk elevators: the continuous discharge and the centrifugal discharge (spaced buckets) type [11].

The centrifugal type bucket elevator is used for free-flowing or granular products. The name arises from the discharge of the product which occurs at the head of the elevator. As the bucket passes over the head the product is discharged by centrifugal forces. The buckets usually have a 60° vertical/lip angle. The speed of the carrying chain must be high enough to create sufficient centrifugal force; speeds of 1.2 to 2.0 m/s are normal.

Materials that range from light to heavy and from fine to large lumps are usually handled by a continuous bucket elevator, where the buckets are spaced close together. These elevators can handle high loading factors and continuous loading. Normal speeds for continuous elevators range between 1.0 and 1.3 m/s.

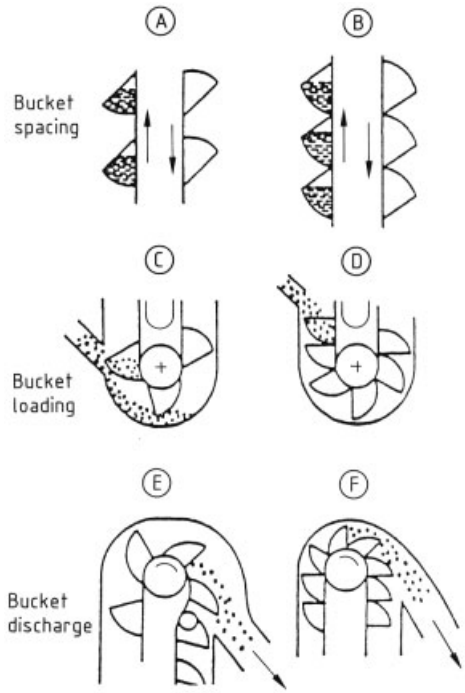

Figure 15. Arrangement of bucket elevators
A) Spaced bucket; B) Continuous bucket; C) Sloping and direct loading; D) Direct loading; E) Gravity discharge; F) Centrifugal discharge

Tables 2 and 3 give data for the sizing and power requirements of centrifugal and con-

Table 2. Size and power requirements for a centrifugal bucket elevator

Pitch, mm	Width, mm	Velocity, m/s	Bucket volume, dm^3	Solid mass, flow t/h	Relative power consumption, kW/m	Additional power consumption, kW
250	200	0.3	6.3	5.5	0.25	0.37
250	400	0.3	12.5	11.0	0.30	0.37
320	300	0.3	18.0	12.0	0.32	0.37
320	500	0.3	30.0	20.0	0.38	0.37
400	400	0.3	38.0	20.5	0.18	0.74
400	600	0.3	57.0	31.0	0.21	0.74
400	800	0.3	76.0	41.0	0.27	0.74
400	1000	0.3	95.0	51.0	0.32	0.74
500	600	0.3	93.0	40.0	0.29	0.74
500	800	0.3	121.0	52.0	0.32	0.74
500	1000	0.3	151.0	65.0	0.41	0.74

Table 3. Size and power requirements for a continuous bucket elevator

Bucket volume, dm^3	Solids mass flow, t/h	Velocity c, m/s	Related power consumption, kW/h	Additional power, kW
400	45.0	0.3	0.32	0.74
600	70.0	0.3	0.38	0.74
400	61.0	0.3	0.35	1.10
600	94.0	0.3	0.47	1.10
800	124.0	0.3	0.54	1.10
1000	155.0	0.3	0.69	1.10

Bulk density taken as 0.5 kg/cm^3, filling efficiency 0.8 and 1.0 for spaced and continuous bucket elevators, respectively.

tinuous bucket elevators, as obtained from DIN 22 201 and DIN 22 203, respectively.

A typical arrangement of bucket elevators can be seen in Figure 15.

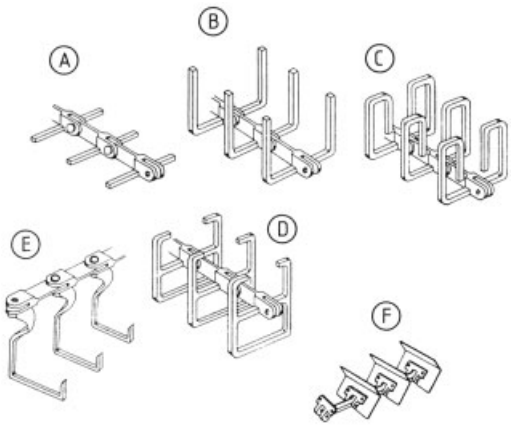

Figure 16. Flight arrangements for en masse conveyors [10] A) Flat flights; B), C), D) Skeleton flights for vertical conveying; E) Suspended flights; F) Solid peaked flights

4.5. En Masse Conveyors

The en masse, or continuous flow, conveyor was first used in England in the 1920s. Transport of the bulk product relies on the friction between the product and the chain or flight and the friction between the particles themselves. If the friction is said to be 100 % then there would be zero slip between the product and the conveyor and hence ideal efficiency would result [9].

The product is introduced to the system to move with the flights at a steady speed. Since the movement is relative to the flight or chain there is a minimum amount of frictional movement between the particles and hence degradation is minimized. En masse conveyors can operate in the vertical, horizontal, and inclined modes. The most common applications are, however, horizontal conveyance of products. Distances in excess of 100 m are not uncommon, and the conveyors may reach capacities of 1000 t /h.

The flight profile depends on the characteristics of the material to be conveyed. Flat flights are most commonly used in the horizontal direction, whilst skeletal flights are used in the verti-

cal direction (Fig. 16) [10, p. 322]. The material from which the flights are made plays an important role in the design of the system.

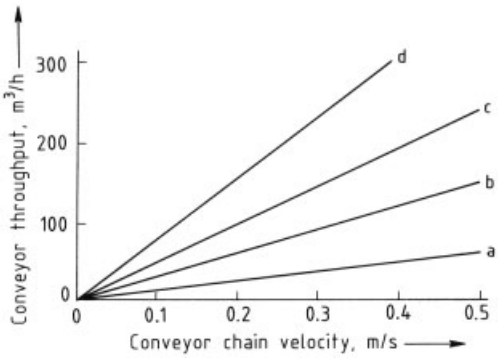

Figure 17. En masse conveyor throughput
Conveyor width, mm: a) 190; b) 290; c) 390; d) 490

The flights are housed in a trough or casing which consists of a steel box with or without a removable lid. The product introduction points are at the top surface of the box or trough and the discharge point is normally at the bottom. The volumetric conveying rate of an en masse conveyor is dependent upon the velocity of the chain, the cross-sectional area of the mass being conveyed, and the slip between the flight and the material. The slip, or the ratio of product velocity to chain velocity, is the efficiency at which the product is conveyed.

The volumetric flow rate is then:

$$\dot{V}_{\mathrm{p}} = A\,v\,s$$

and the mass flow rate is simply:

$$\dot{m} = \dot{V}_{\mathrm{p}}\,\varrho_{\mathrm{B}}$$

where

A = cross sectional area of product, m^2
v = flight velocity, m/s
s = slip or velocity ratio
ϱ_{B} = bulk density, kg/m^3

For horizontal conveyors with a product of large particle size, the slip s can usually be taken as unity, whereas when dealing with vertical and incline systems the slip is between 0.6 and 0.85 [10, p. 324].

The velocity of the conveyor depends on the nature of the material. Products which are easily fluidized have flight velocities in the region of 0.25 m/s while larger products can have flight velocities of up to 0.5 m/s.

Typical capacities for an en masse conveyor are shown in Figure 17 [12].

4.6. Shuttle Belt Conveyors

A shuttle belt conveyor is a simple horizontal track-mounted belt conveyor which receives material from some type of fixed conveying configuration and distributes the material along the length of a bin or a stock pile. Shuttle belt conveyors are often used for lumpy, sticky or highly abrasive materials that would be undesirable in, for example, tripper chutes. They require little head room and can be designed in a variety of belt widths and lengths [8, p. 87].

4.7. Apron Feeders

Apron feeders are essentially an endless apron of overlapping pans supported by suitable intermediate supports. The pans are attached to two or more strands of chain and the apron acts as the conveying medium. These feeders can be used to handle practically any loose bulk material such as ores, sand, gravel, coal, stone, or industrial refuse. They are suitable for sharp, heavy, or hostile materials which are incompatible with other types of material handling systems. There are four basic types of apron feeders:

1) Feeders which are suitable for handling both light and heavy materials where maximum lump size is limited and impact and service are not severe.
2) Feeders which are designed for heavier service than type 1 apron feeders; they can handle larger lump sizes and withstand higher impacts. These feeders are compact and economical, requiring minimum headroom.
3) Feeders which are designed for medium duty but are more flexible as to length and application than the self-contained types 1 and 2. They are mostly used for light materials, such as coal.
4) Heavier duty feeders which are extremely ruggedly built to withstand severe impact and abrasion. They can handle materials containing a high percentage of large lumps, such as ore and rock.

The general layout of an apron feeder is shown in Figure 18 below [8, p. 89].

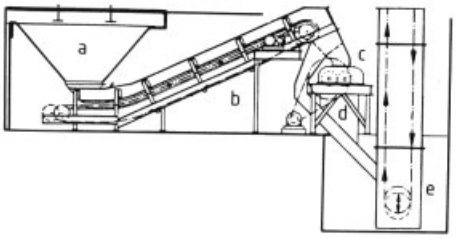

Figure 18. General layout of an apron feeder
a) Hopper; b) Apron feeder; c) Crusher; d) Bypass chute. e) Elevator

4.8. Rotary Table Feeders

The rotary table feeder consists of a power-driven circular plate rotating directly below a bin opening. The volume of the material delivered to the feeder is regulated by a type of feed collar located directly above the table. These feeders are normally used with vertical bins for the handling of materials which tend to arch, such as damp sand or wood chips. The general layout of the rotary table feeder is shown in Figure 19 [8, p. 118].

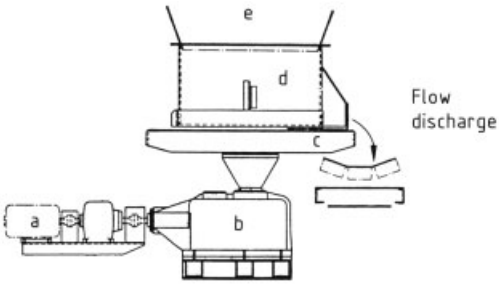

Figure 19. Layout of rotary table feeder
a) Motor; b) Spiral bevel gear drive; c) Revolving table; d) Collar; e) Hopper

Numerous other types of feeder are available and a brief description of the more common types is given below.

Belt feeders consist of a short length of belt operating over idlers. They can handle nearly all types of materials and deliver the product at a uniform rate evenly distributed over the belt.

*Reciprocating feeders*consist of reciprocating driven plates which operate under a head of material. The motion delivers a forward pulsating action to the material, thereby promoting flow. Non-sticky materials and large particle size products can be handled.

Bar flight feeders are simply bars or flights attached to two strands of chain which slide along the flat surface of troughs. They are used for free-flowing non-abrasive materials.

Rotary vane feeders are one of the most common types of feeding devices for the controlled feeding of a product from a storage vessel. They consist of radial vanes rotating in a close-fitting housing and are generally used for the fine free-flowing materials which are not very abrasive.

5. Hydraulic Conveying

Hydraulic conveying technology has made significant advances since its appearance in the 1600s. Today it provides the end user with a technique to move product reliably in the form of a slurry. Although there has been significant progress in the development of theoretical predictions over the past twenty years, there are still uncertainties in the prediction of pipeline characteristics. The information presented in this text can by no means be used as accurate design guidelines but will provide the user with an opportunity for sizing a system [13].

5.1. Pump Selection

There are two main categories of pumps: reciprocating and centrifugal. Reciprocating pumps can be divided into two sub-categories: plunger configuration and piston configuration. Pump selection depends on three factors: the pressure required, the flow rate required, and the nature of the slurry, taking particle characterization into account. The discharge pressure determines whether a centrifugal or a positive displacement pump should be used.

Table 4 outlines the characteristics of various pumps [14].

5.1.1. Centrifugal Pumps

There are two basic configurations for the centrifugal pump pipeline: low pressure pumps spaced equally along the pipeline, and high pressure pumps stationed at only a few positions

Table 4. Performance capabilities of various pump types

Type	Max. [a] pressure, MPa	Max. [b] flowrate, m^3/h	Mechanical efficiency, %	Max. particle size, mm	Comments
Plunger	24 – 28	12 540	85 – 90	2.4	For abrasive slurries [c]
Piston	17 – 21	36 800	85 – 90	2.4	
Centrifugal		681 400	40 – 75	150	Series installation, low capital cost

[a] Maximum pressure refers to commercially viable slurry pipeline volumes
[b] Positive displacement pumps are only capable of these maximum flow rates at pressures considerably lower than the maximum pressure.
[c] Defined as having a Miller No. > 50.

along the pipeline. Some of the advantages of centrifugal pumps over positive displacement pumps include:

1) Large flow rates and large particle sizes (up to 150 mm)
2) Ability to take high discharge pressures
3) Relatively smooth pressure characteristics
4) Long life, especially with abrasive particles
5) The capital costs of multiple staged of centrifugal pumps can be significantly lower than those of positive displacement pumps for high flow rates

High pressure centrifugal pumps were used predominantly for the transportation of tailings. These pumps are, however, being used increasingly to transport coarse mineral concentrates and coal [15].

For a homogeneous slurry the transport velocity is usually chosen to promote turbulent flow. The system can be operated, however, with a completely nonsettling homogeneous slurry where laminar flow exists, but the transition area must be avoided due to inherent instability.

Heterogeneous slurries are operated above the particle settling velocities. The approximate settling velocity can be obtained from numerous equations, the better being that of Cave's modification of Durand's equation [15].

Some high pressure centrifugal pipeline installations are given in Table 5.

5.1.2. Reciprocating Pumps

Plunger pumps are used for high pressure conveying due to the relatively small plunger diameters. Their use for abrasive materials can only be affected if a suitable plunger flushing system is installed. The flushing system keeps the conveyed material from contact with the surface of the plunger and hence minimizes wear. Plunger pumps operate between 80 and 120 strokes per minute, depending on the particle characteristics. Plunger pumps are relatively inexpensive but require more maintenance than piston pumps.

Piston diaphragm pumps are used for the conveyance of very abrasive materials. The product to be conveyed does not come into direct contact with the moving parts of the pump, from which it is separated by a diaphragm that reduces rapid wear propagation. The normal life expectancy of a diaphragm can be up to 6000 h [16].

5.2. Classification of Slurries

There are two categories of slurries, namely *settling* and *nonsettling*. The characteristics of solid-liquid flow are not the same as those of Newtonian fluids for the following reasons:
The effects of the particles on the mixture are combined with the liquid properties, and the flow characteristics will be determined by the particular conditions under which the slurry finds itself [13].

5.2.1. Settling Slurries

This category of slurry can be further divided into four flow regimes:

1) Homogeneous flow
2) Heterogeneous flow
3) Saltation
4) Moving bed

Table 5. Some high pressure centrifugal pipeline installations [15]

Owner (Location)	Slurry type	Solids SG *	Particle size distribution, % ± (mm)	Water, wt %	Capacity, t/h	Pipe diameter, mm	Pipe length, km	Number of pumps in series	Displacement pressure, kPa
BHP Minerals (Australia)	Coal Tailing	1.8	50 – 0.075	52	81	150	5.6	4	4135
BHP Minerals (Australia)	Sand Tailing	3.9	100 – 1	50	200	–	–	5	2700
Alcoa Australia (Australia)	Alumina (sand) Tailings	2.7	50 – 0.28	55	127	150	3.3	6	3550
N.Z. Steel Mining (New Zealand)	Iron Sands	4.9	100 – 0.3	50	1000	325	3.6	6	3450
Consolidated Coal (U.S.A.)	ROM Coal	–	100 – 100	–	425	300	–	7	3500
Consolidated Coal (U.S.A.)	ROM Coal	–	100 – 100	–	425	300	–	7	3500
Hanna-National Steel (U.S.A.)	Taconite Tailings	3.0	100 – 3.5	30 – 40	1500	590	9.7	7	2400
Hanna Butter (U.S.A.)	Taconite Tailings	3.0	100 – 3.5	35 – 40	850	340	7.3	6	3880
Nchanga Consolidated Copper Mine (Zambia)	Copper Tailings	2.7	100 – 0.85	40	1300	600	8.2	4	2545
Folldal-Werk (Norway)	Pyrites Tailings	4.9		33	20	90	1.0	3	2661
Funtana Raminosa (Italy)	Copper Tailings	2.7		50	37	100	1.55	3	2954

* Solids specific gravity

An *homogeneous slurry* has a uniform size distribution or concentration of solid particles across the section of pipe through which it flows [13]. If the ratio of concentration of particles at any plane y to the concentration at some arbitrary reference plane at a height a above the bottom of the pipe is greater than 0.8, then the slurry is homogeneous.

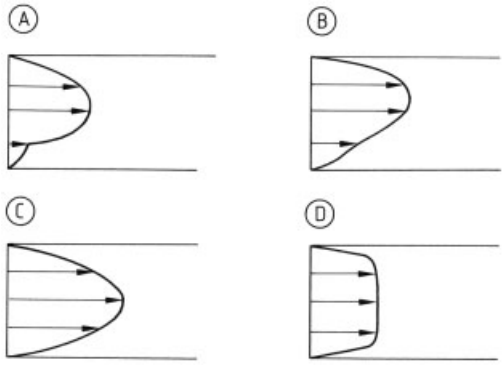

Figure 20. Flow regimes [17]
A) Sliding bed; B) Saltation; C) Suspended heterogeneous; D) Homogeneous

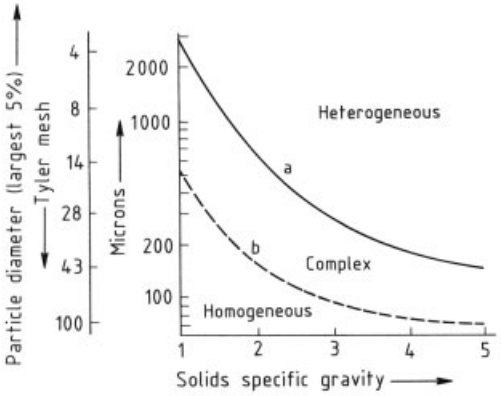

Figure 21. Slurry flow regimes (velocity = 1.2 – 2.1 m/s) as a function of solids size and specific gravity [12]
a) Based on thick slurries with fine (0.325 mesh) vehicle; b) Based on thin slurries with graded particle size

In *heterogeneous slurries* the concentration of particles increases towards the bottom of the pipe. A critical deposition velocity exists, at which point the particles are just held in suspension. At lower velocities the particles fall out of suspension and are deposited on the bottom of the pipe.

Saltation flow occurs when the particles have an energy lower than the critical minimum and have fallen out of suspension. The particles move along the bottom of the pipeline in consecutive bounces.

The *moving bed regime* occurs when the particles are found on the bottom of the pipe and move by sliding along the bottom of the pipe.

The various flow regimes are shown schematically in Figure 20. The flow regime may be classified by analyzing the particle diameter and density (Fig. 21).

The influence of the velocity on the flow regime is shown in Figure 22. Note that most practical slurries have both homogeneous and heterogeneous flow patterns, due to the range of particle size in any given slurry.

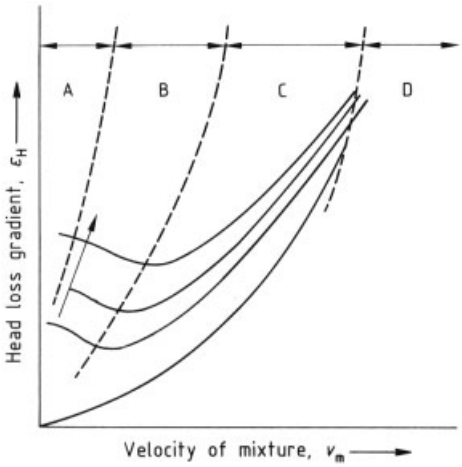

Figure 22. Head loss gradient versus mixture velocity [18], for given values of d, D, and specific gravity
A) Stationary bed; B) Fully moving bed, C_d increasing; C) Heterogeneous flow, critical deposition velocity; D) Homogeneous flow

5.2.2. Nonsettling Slurries

Nonsettling slurries comprise *Newtonian* and *non-Newtonian* fluids.

Newtonian Fluids. The viscosity of a Newtonian fluid is independent of the flow conditions. Newtonian fluids are not common, however, in the hydraulic transport of solids.

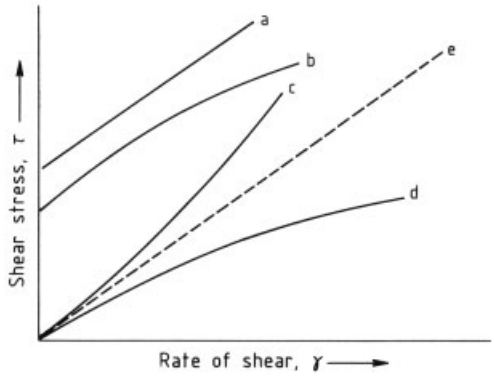

Figure 23. Shear stress – shear strain curves for typical time-independent non-Newtonian fluids
a) Bingham plastic; b) Yield pseudo-plastic; c) Dilatant; d) Pseudo-plastic; e) Newtonian fluid, for comparison

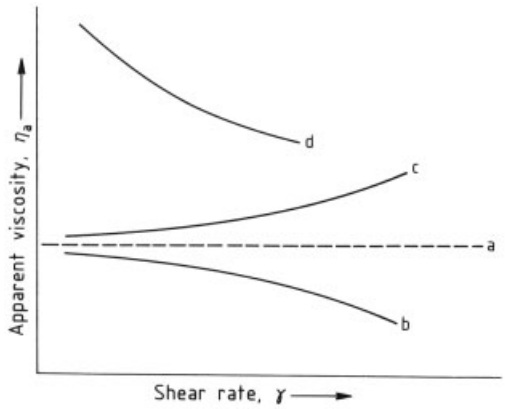

Figure 24. Apparent viscosity versus shear rate for time-independent fluids
a) Newtonian; b) Pseudo-plastic; c) Dilatant; d) Bingham

Non-Newtonian Fluids. This flow pattern is characterized by a continually changing viscosity and can be divided into two categories:

Time-independent fluids require two parameters to define them, unlike Newtonian fluids which require only one. Those parameters needed for classification of flow are known as the rheology of the fluid [13].

The shear stress τ is a function of the applied rate of the shearing strain for those fluids in laminar flow, and is defined as:

$$\tau = K \cdot \frac{\mathrm{d}u}{\mathrm{d}y}$$

where K is a constant and du/dy is the rate of change of velocity u relative to the distance y from the pipe wall.

The time-independent fluids can be divided into several categories (Fig. 23) [13].

The viscosity used to describe the stress-strain relationship of the various time-independent fluids depends on the applied rate of shearing strain and is illustrated in Figure 24. There is still, however, a great deal of controversy in acquiring the viscosity from analytical curves.

Time-dependent fluids. There are two categories of time-dependent fluids, thixotropic and rheopetic fluids. These fluids are indicated in Figure 25 in the stress-strain curves.

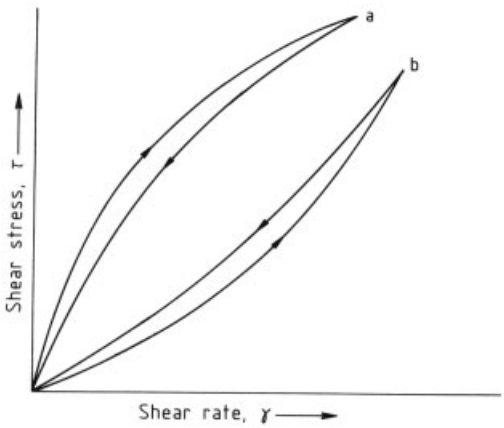

Figure 25. Shear diagram for time-dependent fluids
a) Thioxotropic; b) Rheopetic

5.3. Principles of Design Calculations

The complete area of design principles cannot be treated in this text; some of the more important parameters only will be discussed.

The drag resistance F_{D} of a solid particle in a fluid medium is proportional to the fluid density, the cross-sectional area of the particle and the square of the relative flow velocity. The drag coefficient C_{D} is a factor of proportionality and is a function of the Reynolds number [19].

$$F_{\mathrm{D}} = C_{\mathrm{D}}\,(Re_{\mathrm{p}})\,\frac{\varrho_{\mathrm{f}}}{2}\,(v-c)^2\,\frac{\pi d^2}{4}$$

where $(v-c)$ is the velocity difference between the liquid and the product, ϱ_{f} is the fluid density, and d is the particle diameter.

For a circular pipe, with a laminar fluid flowing through it, of diamter D flow index n, and effective wall roughness k, the average volocity $\bar{v}$ is:

$$\bar{v} = \frac{n}{(3n+1)}\,\frac{D}{2}\left(\frac{1}{k}\,\frac{D\Delta p}{4L}\right)^{\frac{1}{n}}$$

from which the pressure drop Δp per length L is expressed as:

$$\frac{\Delta p}{L} = \frac{4k}{D}\left(\frac{3n+1}{n}\cdot\frac{2\bar{v}}{D}\right)^n \tag{1}$$

where k is the effective pipe wall roughness.

Substituting for the shear stress at the pipe wall

$$\tau_{\mathrm{o}} = \frac{D\Delta p}{4L}$$

gives:

$$\tau_{\mathrm{o}} = k\left(\frac{3n+1}{4n}\right)^n\left(\frac{8\bar{v}}{D}\right)^n$$

Manipulating this equation for a non-Newtonian fluid by defining the effective viscosity as the wall shear stress divided by the average shear rate at the boundary, it can be seen that for a power law fluid the effective viscosity η_{e} is [20]:

$$\eta_{\mathrm{e}} = k\left(\frac{3n+1}{4n}\right)^n\left(\frac{8\bar{v}}{D}\right)^{n-1}$$

The Reynolds number (as a function of the effective viscosity) at which the transition from laminar to turbulent flow occurs is a function of the flow index n [20]. Figure 26 shows the approximate relationship between the critical Reynolds number and the flow index n [21]. It is then possible to obtain the velocity at which the transition from laminar to turbulent flow occurs for a non-Newtonian fluid.

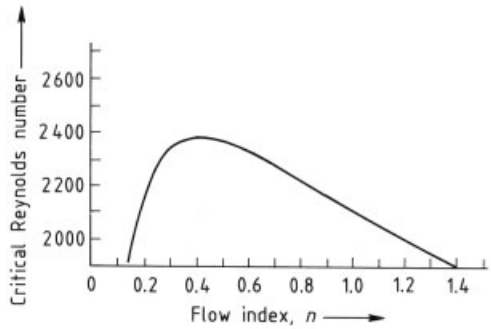

Figure 26. Relationship between critical Reynolds number and flow index n for power law fluids [21]

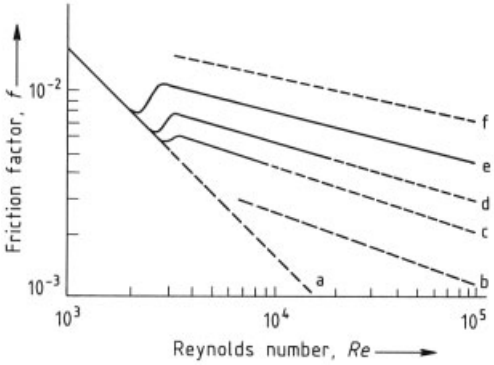

Figure 27. Friction factor design chart for power law fluids [21]
—- Experimental region; - - - - Extrapolated region
a) $n = 0$; b) $n = 0.2$; c) $n = 0.4$; d) $n = 0.6$; e) $n = 1.0$; f) $n = 2.0$

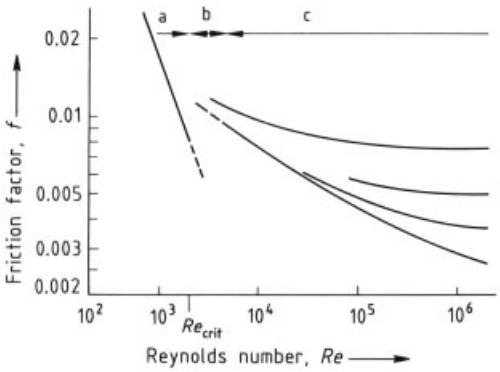

Figure 28. Moody chart for the determination of friction factors for the flow of fluids in pipes of circular cross-section
a) Laminar flow, $f = 16/Re$; b) Critical flow; c) Turbulent flow

If the flow occurs, however, in the turbulent region, the calculations are not so simple. Dodge and Metzner [22] suggested that the generalized Reynolds number, defined as follows, be used for pressure drop calculations:

$$Re^* = \frac{\varrho_f D^n \bar{v}^{2-n}}{\eta}$$

Using the above Reynolds number and the design chart (Fig. 27) [23], the pipe friction factor f can be calculated and the pressure drop obtained by the normal method.

For laminar flow the pressure drop is simply calculated using Equation (1) above.

The pressure drop calculation for a Bingham plastic is given by the equation for laminar flow:

$$\frac{\Delta p}{L} = \frac{32\,\eta_p\,\bar{v}}{D^2} + \frac{16\,\tau_y}{3D}$$

For turbulent flow a convenient method is to obtain the friction factor from the common Moody diagram (Figure 28) and to use the plastic viscosity for the Reynolds number.

The most predictable method of designing an installation is clearly to use theoretical correlations backed up by extensive practical loop tests.

6. Pneumatic Conveying

The pneumatic transport of bulk solids represents one of the most widely used material handling techniques in the chemical industry. In addition to providing a relatively compact (high density) handling technique, pneumatic conveying offers the following *advantages*:

1) *Flexibility*. The system can follow any route; a change in flow direction is achieved by simple addition of a pipe bend; it is relatively easy to extend pipeline length.
2) *Pollution-free transport.* Because solids are contained in a closed conduit, toxic materials can be transported safely and the product is also protected from environmental contamination.
3) *Reduction in manpower and minimal maintenance*. Pneumatic conveying systems have virtually no moving parts, require minimal maintenance, and have relatively modest manpower requirements.
4) *Control.* With the latest feeding techniques, pneumatic conveying systems can be easily integrated into modern chemical plants employing sophisticated process control.
5) *Range of products transported*. Pneumatic conveying techniques have been succesfully applied to a wide range of products from large lumps of ore to ultrafine submicron powders. Modern handling techniques have facilitated the transportation of flakes, powders, and granules.

Although the technique provides the user with a wide range of attractive features, there are also certain *disadvantages*:

1) *High power* consumption, when compared with other mechanical handling systems.

2) *Limited distance*. Pneumatic handling techniques are generally limited to the indoor transportation of materials, although there are systems operating which move products over a maximum distance of 4000 m.
3) *Limited throughput*. Pneumatic conveying techniques are best suited to handling products at rates of hundreds, rather than thousands, of tonnes per hour.

6.1. Classification of Pneumatic Conveying Systems

There is much confusion about the classification of pneumatic conveying systems. Scientific classification techniques have been confused by commercial vendors highlighting certain flow features of their own systems; it is not always clear in exactly which flow regime such systems operate.

6.1.1. State Diagram

In order to eliminate any confusion pneumatic conveying systems are now classified according to the so-called state diagram (Fig. 29), constructed by plotting the pressure gradient $\Delta p/L$ for a gas-solid suspension against the superficial air velocity v.

A state diagram can only be obtained by conducting tests in a test loop and monitoring the pressure loss over a certain length of pipe in the so-called established flow regime (after the solids have been accelerated). The superficial air velocities are obtained by using standard air flow measuring techniques.

The flow conditions for a horizontal pipe are depicted in Figure 29 A. The following considerations are important:

Dilute Phase and Dense Phase Flow: The state diagram distinguishes clearly between these two flow regimes. Dilute phase flow is characterized by fully suspended flow in which all particles are moved along the pipe supported by the gas flow (Fig. 29 A, a).

The transition between dilute and dense phase flow is known as the saltation point; at velocities below the saltation velocity the solids will settle out of suspension. Dense phase flow is thus characterized by the solids moving as a bed along the bottom of the conveying pipe. A number of flow profiles can occur in the dense phase mode, including plug flow, dune flow, and slug flow. These flow profiles are generally a function of the solids being conveyed.

Between dilute phase flow and the so-called steady state dense phase flow there is a region known as the unsteady state. This flow region is normally characterized by a surging type flow pattern in which there is a continual buildup and breakdown of dunes. The flow in this region is normally associated with pressure surging. A state diagram with five flow regimes is shown in Figure 30 A. For each regime, a typical flow pattern appears in Figure 30 B.

The state diagram (Fig. 29) also illustrates that the pressure gradient at the saltation point is at a minimum. In dilute phase flow it is desirable to operate as close to the saltation velocity as possible; in general, dense phase flow occurs at a much higher pressure gradient.

Two important observations must be noted: an increasing solids mass flow rate results in an increase in the pressure gradient; the locus drawn through the saltation points for each solids flow rate increases with increasing air velocity (see Fig. 30 B). This observation is critical in terms of uprating a system, when it is important to ensure that the air compressor is capable of providing the additional pressure requirements and that there is also a corresponding increase in air flow rate. By not increasing the air flow rate it is quite conceivable that the system could move from dilute phase flow at the lowest solids mass flow rate to dense phase flow at the higher solids mass flow rate. Many problems could then arise, particularly if the system has not been designed to withstand a higher operating pressure.

In general, dense phase flow (which employs far less conveying air) is more energy efficient than dilute phase systems. The added pressure requirements, however, dictate the need for more expensive equipment. Dense phase conveying systems employ high pressure feeders which require high pressure compressors. However, the most popular pneumatic conveying systems still operate in the dilute phase mode.

Because dilute phase systems employ more conveying air and as the requirements are for all the solids to move in a suspended flow mode,

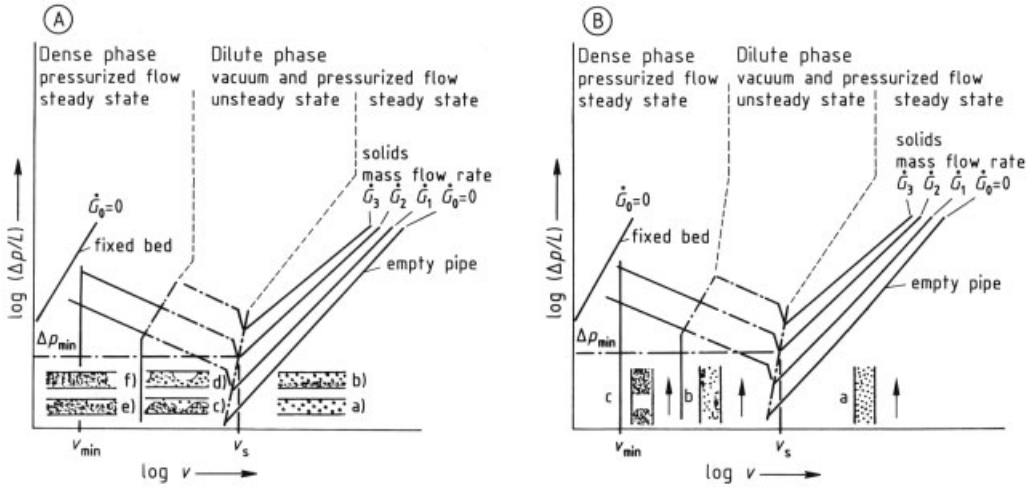

Figure 29. State diagram for A) Horizontal conveying;
a) Fully dispersed; b) Layer dispersed; c) Dune flow; d), e) Plug flow; f) Full bore plug flow
B) Vertical conveying; a) Uniformly dispersed; b) Non uniformly dispersed; c) Full bore plug flow
Solids Mass flow rates $\dot{G}_3 > \dot{G}_2 > \dot{G}_1$

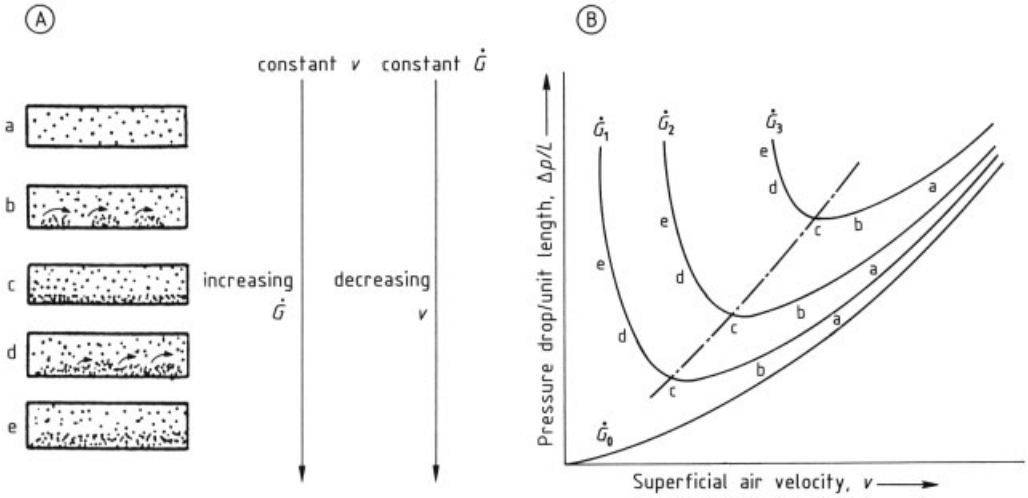

Figure 30. State diagram with corresponding flow regimes, $\dot{G}_3 > \dot{G}_2 > \dot{G}_1$

dilute phase systems operate at conveying velocities which are substantially greater than those employed in the dense phase systems. Typically, for, say, ordinary portland cement moving as a dilute phase, flow air velocities of the order of 25 m/s are employed. In dense phase flow air velocities of the order of 10 m/s are normal practice. Dilute phase systems are, due to the higher velocities, limited to conveying products in which product degradation is unimportant. The high velocities cause very high solids degradation because of particle collisions and, more importantly, because of to particle-pipe wall interaction.

With dense phase conveying there is a number of techniques used to promote the so-called plug phase conveying. External mechanical devices are placed along the pipe to promote the movement of solids as discrete plugs. Plug phase or low velocity conveying has been developed to ensure minimal product degradation. The ability of a product to move as a discrete plug depends

on the physical and chemical characteristics of the product itself. Vendors will often market a dense phase system for a product which either has a tendency to self plug generation or for products whose characteristics are not amenable to plug formation at all.

6.1.2. Vertical and Horizontal Flow

The state diagram for vertical flow (Fig. 29 B) has the same overall characteristics as that for horizontal flow (Fig. 29 A). There are, however, some important differences between horizontal and vertical flow. On average the saltation point in vertical flow occurs at a much lower velocity, and the overall pressure loss in a vertical pipe is substantially lower than that in a horizontal pipe.

These differences can be ascribed to particle-wall interaction. In the case of horizontal flow the conveying air must have sufficient energy to overcome the tendency of the particles (which have a density some 1000 times greater than that of the conveying air) to drag along the bottom of the pipe. Designers are thus cautioned not to base their design criteria on a vertical pipeline if it is intended to move the product through both vertical and horizontal pipes.

6.2. Pressure Drop Components

The components of the pressure drop required for the transportation of a gas-solid suspension are depicted in Figure 31. For simplicity the pressure drop may be broken into components:

$$\Delta p_{\mathrm{tr}} = \Delta p_{\mathrm{L}} + \Delta p_{\mathrm{Z}}$$

where Δp_{tr} is the total pressure drop required to effect the transportation of solids, Δp_{L} is the pressure drop due to air alone, and Δp_{Z} is the additional pressure drop due to the presence of the solid.

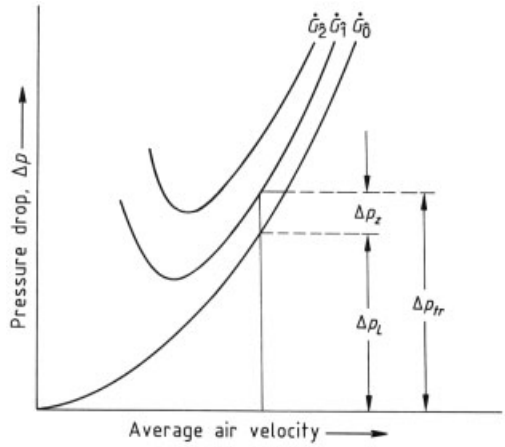

Figure 31. Components of pressure drop required to transport product, constant solids mass flow rates $\dot{G}_1$, $\dot{G}_2$

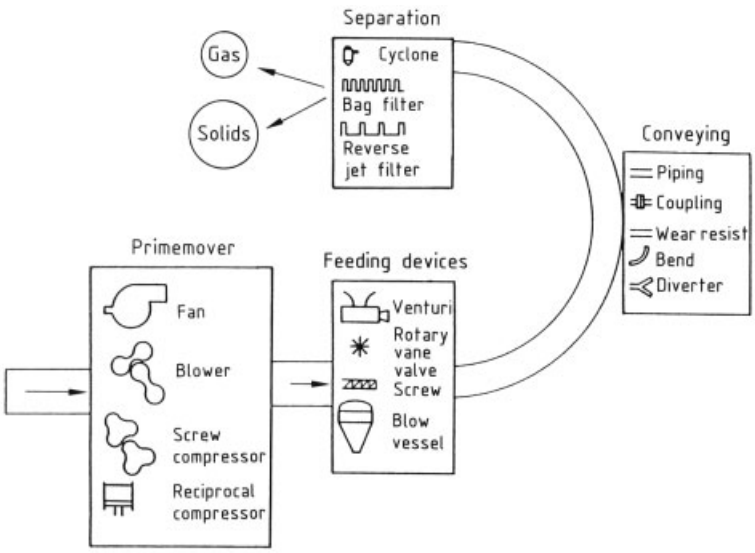

Figure 32. Principal components of a pneumatic conveying system

The greater contribution to the transportation pressure drop is because of the air alone and because the solids contribute a relatively minor portion of this pressure drop (see Fig. 31).

6.3. Modes of Pneumatic Conveying and Product Characteristics

Although the state diagram provides a useful description of dilute and dense phase conveying modes, there is still the question of product characteristics and whether a product is amenable to either dilute or dense phase conveying or both.

Products can be categorized into four regimes: The so-called Group A materials (e.g., cement, lime, fly ash) are considered to be good candidates for dense phase conveying. Group B materials (e.g., sugar, sodium sulphate) are more granular in nature and can be conveyed in the dense phase mode; special attention must be given to the method of feeding the product into the pipe line. (Without controlled feeding, these products tend to flood the pipeline and the system invariably blocks almost instantaneously.) Group C materials are those ultrafine submicron powders (e.g., talc powder) that, because of their particle size, tend to build up on the conveying pipe wall leading eventually to complete pipe blockage. These products, although amenable to dense phase conveying, must be transported in a flexible rubber pipe; the continuous movements of the pipe wall minimize any buildup. The Group D materials are those products (e.g., coal, mine ore) with large lump sizes. Because of their size (lumps up to 80 mm have been conveyed) it is suggested that the only reliable conveying mode is to transport the products as a fully suspended flow in the dilute phase mode.

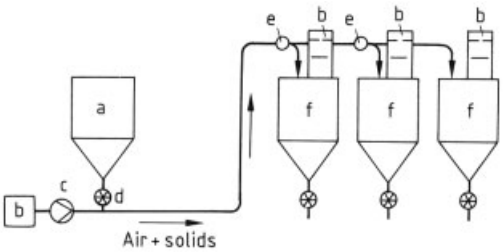

Figure 33. Typical positive pressure pneumatic system
a) Storage hopper; b) Filter; c) Air supply; d) Feeder; e) Diverter valves; f) Discharge hoppers

6.4. Pneumatic Conveying Systems

Certain basic components are required for a pneumatic conveying system. The solid product, at atmospheric pressure, must be fed into a flowing gas stream, conveyed over a prescribed route, and on reaching its destination, is separated from the conveying gas. A basic pneumatic conveying system consists of the following components (Fig. 32):

1) *Prime mover*. Prime movers vary from low pressure fans to blowers and positive displacement compressors or vacuum pumps, and are selected according to system requirements.
2) *Feeding system*. Perhaps the most important item in any pneumatic conveying system. The feeder introduces the solid into a positive or negative pressure environment, simultaneously providing an airtight seal. Feeders vary in construction and operation and are selected according to the nature of the product, the process, and the system operating pressure.
3) *The conveying pipe*. With more sophisticated feeding techniques it is now possible to convey highly abrasive materials and the selection of the conveying pipe is rather critical. A large number of different piping materials is available and piping materials should be selected on the basis of both pressure and the particular requirements of the product. In certain cases, pipe walls are specially prepared to prevent the formation of so-called angle hairs in the conveying of, say, plastic pellets.
4) *The gas-solid separation system*. In order to separate the solid from the conveying gas stream, it is necessary to select a suitable separation system which is also dependent upon system requirements. Gas-solid separation systems vary from cyclone separators to highly sophisticated reverse jet bag filters. The separation efficiency is determined by the process and the possible pollution problems arising from inadequate separation. Separating devices are selected on the basis of separation efficiency and are designed to ensure complete separation of the

solid from the conveying stream. Although the gas-solid separation system is located at the end of the process, it is a vital component to the effectiveness of a pneumatic conveying system.

There are four basic configurations in pneumatic conveying systems, each being arranged to meet a particular industrial application.

6.4.1. Positive Pressure System

A positive pressure system facilitates the feeding of the product into a positive pressure air stream. These systems are normally used where a product is conveyed from a single pick-up point to multiple delivery points. Such a system is illustrated in Figure 33. Alternatively, there are also positive pressure systems that pick up product from several feeding points.

6.4.2. Negative Pressure System

Negative pressure or vacuum systems are well suited to applications where product has to be picked up from several feed points and delivered to a single receiving point. These systems are limited in capacity and distance because of the vacuum which can be generated. A schematic of a negative pressure system is shown in Figure 34.

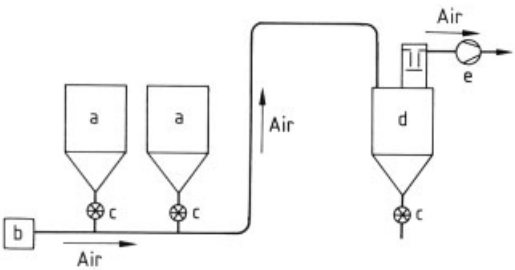

Figure 34. Typical negative pressure pneumatic system
a) Storage hoppers; b) Filter; c) Feeder; d) Discharge hopper; e) Exhauster

Vacuum systems are extensively used for the movement of toxic products, where any hole in the pipeline would cause pollution because of product spillage. Vacuum systems are also extensively used to handle product from open stockpiles, rail and road vehicles, and are often used as primary feeder for positive pressure systems.

6.4.3. Combined Positive and Negative Pressure Systems

The advantages of the negative and positive pressure systems can be consolidated into one system (Fig. 35). The hardware is arranged to facilitate the use of a single blower/compressor in which both the intake and outlet are used to provide the negative and positive pressure air flow, respectively.

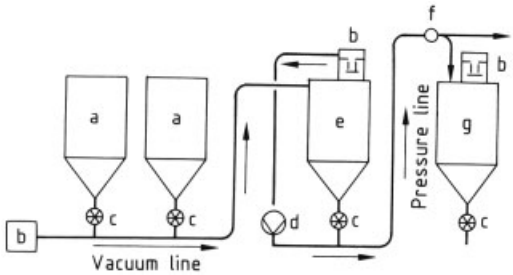

Figure 35. Combined positive and negative pressure pneumatic system
a) Storage hoppers; b) Filter; c) Feeder; d) Exhauster/blower; e) Intermediate hopper; f) Diverter to alternative discharges; g) Discharge hopper

7. References

1. N. Clark, T. E. Durney, T. P. Meloy, "Recent Advances in Particle Shape Characterisation," *J. Powder Bulk Solids Technol.* **8** (1984) 21 – 24.
2. H. H. Hausner, "Powder Characteristics and their Effect on Powder Processing," *Powder Technol.* **30** (1981) 3 – 8.
3. T. P. Meloy, "A Hypothesis for Morphological Characterisation of Particle Shape and Physiochemical Properties," *Powder Technol.* **16** (1977) 233 – 253.
4. C. R. Woodcock, J. S. Mason: *Bulk Solids Handling,* Chapman and Hall, London 1987, p. 30.
5. R. G. Boothroyd: *Flowing Gas-Solid Suspension,* Chapman and Hall, London 1971, p. 34.
6. P. C. Arnold, A. G. McLean, A. W. Roberts: *Bulk Solids: Storage, Flow and Handling,* Tunra Bulk Solids Handling Research Associates, 1981.

7. R. L. Carr, "Evaluating Flow Properteis of Solids," *Chem. Eng. NY,* **72** (1965) 163 – 168.
8. LINK BELT – Materials Handling and Processing Equipment, 1958.
9. H. A. Bolz, G. E. Hagemann: *Materials Handling Handbook,* Ronald Press, New York 1958, p. 24.24.
10. C. R. Woodcock, J. S. Mason: *Bulk Solids Handling,* Chapman and Hall, London 1987.
11. T. Kobayashi, "The Mitsubishi Continuous Ship Unloader," *Bulk Solids Handling,* **5** (1985) no. 3, 63, Trans. Tech. Publications.
12. G. C. King, "The Application and Design of En-masse Conveyors," Proc. Solidex 80 conf, Harrogate, UK, March/April 1980.
13. C. G. Verkerk, *An Investigation into the Design of Slurry Pipelines with Reference to Low and High Concentration Slurries,* MSc Dissertation, University of the Witwatersrand, Johannesburg, S. Africa, July 1982, pp. 22 – 60.
14. E. J. Wasp, J. P. Kenny, R. L. Gandhi, "Solid-liquid flow Slurry Pipeline Transportation," Trans. Tech. Publications, 1977.
15. C. I. Walker, "High Pressure Centrifugal Pumps and their Pipeline Applications," *Bulk Solids Handling,* **3** (1983) no. 1, 149, Trans. Tech. Publications.
16. G. Wallrafen, "Piston Pumps for the Hydraulic Transport of Solids," *Bulk Solids Handling,* **3** (1983) no. 1, 149, Trans. Tech. Publications.
17. E. Anderson, S. Bouchardy, G. Capitaine, J. Davies, M. Lavitte, J. C. Leyssieux, R. Meslah: *Evolution of Pipeline Transport Techniques* **4,** Pipelines Geneva Research Centre, March 1977.
18. P. J. Baker, B. E. A. Jacobs: *A Guide to Pipeline Slurry Systems,* BHRA Fluid Engineering, Cranfield, Bedford, England 1979.
19. M. Weber, "Principles of Hydraulic and Pneumatic Conveying in Pipes," *Bulk Solids Handling,* **1** (1981) no. 1, 2, Trans. Tech. Publications.
20. C. R. Woodcock, J. S. Mason: *Bulk Solids Handling,* Chapman Hall, London 1987, p. 142 – 145.
21. E. J. Wasp, J. P. Kenny, R. L. Gandhi: *Solid-Liquid Flow Slurry Pipeline Transportation,* Gulf Publishing Co, Houston, TX, 1979.
22. D. W. Dodge, A. B. Metzner, "Turbulent Flow of non-Newtonian System," *AIChE J.,* **5** (1959) 1 – 9.
23. C. G. Verkerk, "Some Practical Aspects of Correlating Empirical Equations to Experimental Data in Slurry Pipeline Transport," *Bulk Solids Handling,* **5** (1985) no. 4, 21, Trans. Tech. Publications.

Solid – Liquid Separation, Introduction

For unit operations in solid – liquid separation see → Filtration; → Centrifuges, Filtering; → Centrifuges, Sedimenting; → Hydrocyclones; → Sedimentation

CHRISTIAN ALT, München, Federal Republic of Germany

1. General Considerations 1
2. Separation Processes 4
2.1. Sedimentation 4
2.2. Filtration 5
3. References 7

1. General Considerations

A wide variety of equipment and machinery is available for the separation of solids and liquids. Counting the many possible optional variations, a potential user can choose from among more than 100 separation devices. Since some of the equipment was developed to solve specific problems, its use is confined mainly to those or to other closely related tasks. However, most equipment is universally applicable and lends itself to comparisons.

The equipment considered in this article is used principally in the following separation processes:

1) screening (wet screening or screen filtration)
2) sedimentation
3) filtration
4) expression, and
5) electrokinetic or magnetic separation.

In this classification, equipment for elutriation occupies a special place. Although it includes separation of the two phases of a solid – liquid mixture, its major task is selective separation within the solid phase (→ Elutriation).

Each of the listed separation processes can be further categorized by additional characteristics. Figure 1 shows a classification according to the forces that are necessarily active in every separation process. Accordingly, separation processes are basically differentiated by the motion of the phases to be separated, and the equipment can be classified into *sedimentation apparatus* and *filters*. Nevertheless, this classification is not always unambiguous, since a large number of devices comprise both mechanisms.

Separators are used for coarsely dispersed solid – liquid mixture systems with particle sizes of several millimeters down to finely dispersed systems of molecular dimensions. In addition, highly concentrated slurries are separated as well as dilute suspensions. For the evaluation of separation equipment such inherent parameters as different geometries of the particles, interfacial effects, and electrostatic interactions are considered first; they can be influenced by pretreatment. Therefore, a number of important factors must be observed.

A systematic collection of factors to be considered may be helpful for selecting separation equipment and avoiding basic mistakes. Some criteria for the use of liquid separation processes are listed below; see also [1]:

1) Material Factors

1.1) Feed

Solids:
- concentration
- density
- particle size
- particle size distribution
- particle shape
- particle structure (aggregates, flocs and crystals)
- stability and hardness
- solubility
- corrosiveness
- chemical composition

Liquids:
- chemical composition
- density
- viscosity
- pH value
- volatility, vapor pressure

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_09

Transport mechanism		Free motion of dispersed particles	Free motion of liquid
Designation		Sedimentation Flotation	Filtration
Driving forces	Gravity	Thickener/Clarifier	Screen drainage Dewatering screens Hydrostatic filter Microstrainer
	Pressure gradient		Cake filter Pressure filter Vacuum filter Capillary filter Expression device Deep bed filter Cross-flow filter Strainer
	Centrifugal field	Static wall Hydrocyclone Rotating wall Solid bowl centrifuge Rim centrifuge Scroll discharge centrifuge Disk-type centrifuge	Centrifugal filter Peeler centrifuge Continuous solids discharge
	Electric field	Electrophoresis	Electro-osmosis
	Magnetic field	Low-field magnetic separator High-gradient magnetic separator	

Figure 1. Classification of equipment for solid – liquid separation

Impurities
Foaming tendency
Temperature
Toxicity

1.2) Wash Liquor

Material properties; see "liquids"

1.3) Equipment

Corrosion protection
Abrasion protection
Load capacity

2) Performance Requirements

Amount of feed
Residual moisture
Effluent clarity
Wash liquor concentration
Residual mother liquor
Solids transportability

3) Operating Conditions

Continuous versus discontinuous
Throughput
Feed variation (concentration and amount)
Residual moisture
Solids breakage
Washing (separate collection of effluents)
Foaming tendency
Temperature range
Auxiliary processes:
- flocculation
- precoating
- filtration aids
- foam suppression

Preceding and subsequent process steps

4) Equipment Conditions

Energy requirement
Auxiliaries' requirement
Space requirement
Pollution (e.g., gases or noise)
Corrosion tendency
Abrasion tendency
Personnel requirements

Automatic control
Durability
Spare parts
Accessories
Special adaptation
Apparatus combination
Availability

This list does not claim to be complete for extreme cases. However, very often only a fraction of the factors mentioned above must be observed.

The properties of materials to be separated sometimes vary widely in similar processes and fluctuate frequently in the course of production, which further complicates comparisons and evaluations made for the purpose of selecting the optimal separation equipment.

Nevertheless, some generally valid relations narrow the comparison of different separation processes. A separation process utilizing gravity usually requires less energy than most other processes. Thus, for example, a sedimentation tank can be operated with only a few joules per cubic meter of liquid to be separated, while filters and centrifuges rarely require less than 10 000 J per cubic meter of liquid to be separated. Both machines require less than half the energy needed to operate a drier, although the latter is capable of achieving a lower residual moisture content. Thus, energy consumption and residual moisture content determine to some extent the type of separation process to be used. Another factor to be considered is the intensity of operation of the separation equipment; thus, the energy requirement for the same apparatus can be reduced considerably if, for example, a higher residual moisture content of the solids can be tolerated.

The particle size of the dispersed solids also shows a relationship between material parameters and the equipment that can be used. Figure 2 surveys proven application ranges of typical separators. The plot along the ordinate is roughly proportional to the optimum feed concentration for the various separator devices; in Figure 2 the ordinate extends to a solids concentration of about 40 wt %. However, in individual cases a lower or even a higher feed concentration may be preferable.

A comparison of separation processes should also include *investment costs.* In addition to the purchase price of each piece of equipment and of the necessary auxiliaries, one must consider the costs of installation, erection, and connections as well as safety, both of the operating personnel and of the environment. The investment cost divided by the throughput always decreases as the capacity of the equipment increases. As an example, the net purchase price of filters is

$$I_A = k_1 A^{-k_2} \tag{1}$$

where I_A is the purchase price per square meter of filter area, k_1 and k_2 are constants that depend on the type of equipment and the material of construction, and A is the filter area.

Only an expert is able to compare easily and effectively separation processes and associated equipment, considering all nuances of the factors and criteria mentioned above. If, in addition, economic factors are to be taken into account, it is almost impossible to select the optimum separation equipment without expert knowledge and years of experience. Therefore, it is, frequently impossible to avoid the type of laboratory studies that until recently were practiced routinely. Certain types of testing equipment are useful, in such cases, e.g., long-tube apparatus, hand filter plates or compression filter cells, and bottle centrifuges. Frequently, pilot-plant or small-scale production testing cannot be avoided. The corresponding production technique can then be simulated, if necessary.

Even the relatively favorable energy consumption in the solid-liquid separation, mentioned at the outset, should be compared in relation to the subsequent process steps. Initially, the energy consumption of the equipment under consideration certainly forms the basis of planning studies, especially for new plants and within the context of cost saving measures. However, it should be kept in mind that the situation can be altered by overriding factors. For example, dilutions should be avoided in a production process wherever possible, since the solids-related separation costs are increased by reductions in feed concentration. Nevertheless, dilution of a suspension prior to separation may be useful; it can, for example, advance flocculation. In certain filtration processes a more compact filter cake can result from a dilute suspension than from a concentrated one.

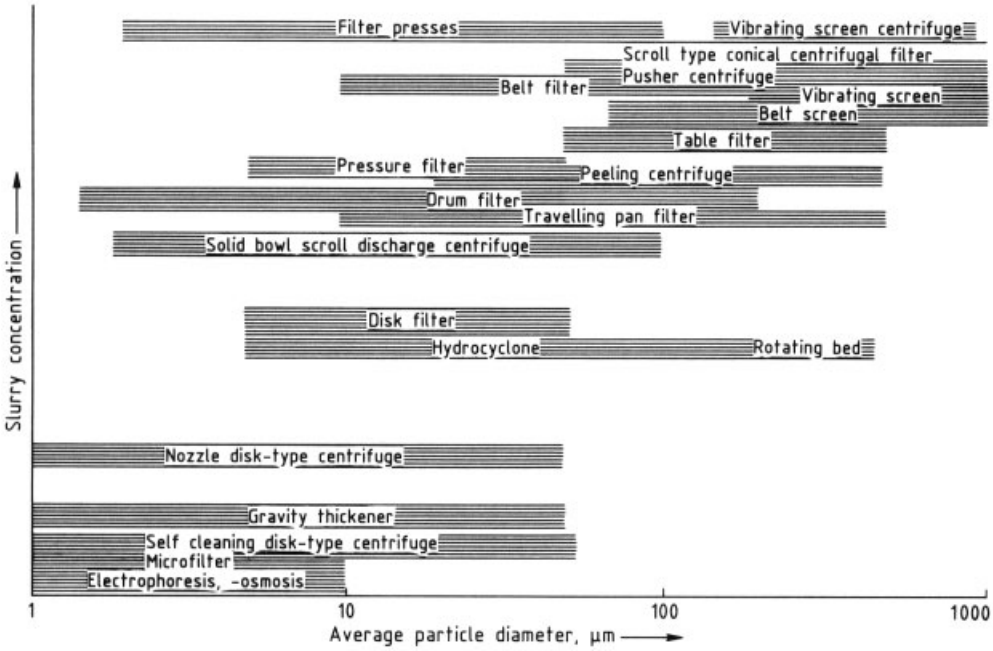

Figure 2. Ranges of application of solid – liquid separation equipment as a function of particle size and slurry concentration

2. Separation Processes

Each separation process exhibits its own specific behavior based on physical and structural situations which characterize the ranges of application from the outset [2].

2.1. Sedimentation

Gravity Sedimentation. Sedimentation in the gravitational field should be considered in principle only for highly dilute dispersions with relatively finely divided solids particles. It is, indeed, the simplest separation process and entails the lowest cost. Even broadly dispersed and extremely structured, flaky solids particles can be separated at relatively high throughputs with very few problems, if suitably long sedimentation times are provided.

Thickeners and clarifiers are, as a rule, not operated in a free-standing, single-step mode, but often as preliminary stages if both of the phases to be separated are to be obtained with high efficiency. The liquid contents (residual moistures) of the sludges are comparatively high. Nevertheless, the degree of solids separation is good and may be as high as 90 to 100 %. The degree of liquids separation, on the other hand, is lower.

At throughputs of up to $1\,m^3\,m^{-2}\,d^{-1}$ the specific performances are relatively modest and require suitably large units at higher throughputs. Concrete basins are cheap, but require a lot of space. Because of their low operating costs, simplicity of automation devices, and the small number of operators required, thickeners are used widely in the separation and wastewater treatment plants.

Centrifugal Sedimentation. Sedimentation in a centrifugal force field clearly requires more complicated equipment and more careful assessment than gravity sedimentation. For example, with solid bowl centrifuges and disk-type centrifuges operating at an acceleration of 1000 g (g: acceleration due to gravity), high sedimentation rates can be achieved, even with fine-grained particles, which permit high throughputs per machine unit. With scroll discharge of solids, in continuous decanter centrifuges, about 200 m^3/h of wastewater slurry can be dewatered in large modern units. With the aid of flocculation, the sedimentation rate can be increased further by about a factor of 10. The flocculating agent,

which in this case remains in the separated solids, may be undesirable at times; for that reason, separation processes free of flocculating agents, for example, in disk-type centrifuges, are then preferred. Also suitable, although at reduced throughput, are special modifications of solid bowl scroll discharge centrifuges. For further information, see → Centrifuges, Sedimenting. Ternary systems – two liquid phases and one solid phase – can be successfully and completely separated both in differently modified solid bowl scroll discharge centrifuges and in disk-type separators with peripheral solids discharge through jets or annular slots.

A comparison of sedimentation equipment should usually also include the wear that is expected. Centrifuges are considerably more subject to wear than gravity basins; therefore, a number of devices are available to protect against wear, especially for the highly stressed conveyors in continuous decanter centrifuges.

Hydrocyclones. Hydrocyclones are simple devices and require very little expenditure. To be sure, they do not provide sharp phase separations; for this reason, they are used more as pre-stages for concentrating dilute suspensions ahead of a filter and for similar purposes (→ Hydrocyclones.

A key disadvantage of most sedimentation equipment is that sludges cannot be washed. If the washing of solids is strictly required, solid bowl scroll discharge centrifuges and disk-type centrifuges are applicable, though only to a limited extent.

2.2. Filtration

A very large number of different filters are available on the market. They can be used for coarse-particle suspensions of up to a several millimeters particle size and even with a broad particle size distribution on the one hand, and for colloidal suspensions on the other hand. The very fine-particle suspensions have become amenable to mechanical separation because membranes of various configurations have become available. In addition, filters can be used for almost any solids concentration.

In contrast to sedimentation equipment, filters are independent of phase density. Even slurries with only minute differences in density can be filtered without problems.

Commensurate with their importance, filters are also available for extremely high performance. Filtration areas can be up to 1000 m^2 in size and permit throughputs of several hundred cubic meters per hour.

Particle Size and Solids Concentration. *Coarse-grained suspensions* are reserved for vacuum filters or occasionally for hydrostatic filters. Both devices are energetically more favorable and simpler than pressure filters. For *slurries containing fine particles* deep-bed filters and dynamic (cross flow) filters are available in addition to prevalent pressure filters. The *feed concentration* largely determines the preferred type of filter. Pressure filters generally require higher solids concentrations than the others which are selected more on the basis of what is expected subsequently of the separated phases. For example, it is difficult to further concentrate solids retained in deep-bed filters; in dynamic filters such concentration is routine.

When highly concentrated slurries contain solids that are finely divided or otherwise difficult to separate, such as slimy particles, they can also be separated by precoat filtration, preferably on drum filters, centrifugal discharge plate or candle filters; this involves the use of filter aids. However, relatively dry solids obtained in this way contain the filter aids, except for those shaved off the drum filter.

Coarse-grained and simultaneously fast settling solids are less suitable for drum and disk filters since the solids form a sediment in the tanks. Likewise, slurries with relatively large fractions of fine particles are more difficult to handle in these filters since the finest particles, which build up the filter cake first, may frequently blind the filter medium.

Residual Moisture. Filters in comparison to thickeners clearly yield lower residual moisture. However, minor differences in individual filters should be noted. Thus, lower residual moisture is obtained in pressure filters, filter presses, and especially in press filters. In this connection, dynamic filters give above-average residual moisture. Deep-bed filters generally are not intended for the production of dry solids with low residual moisture, but are preferred to clarify liquids.

Particularly low residual moisture is obtained in centrifugal filters, since centrifugal forces act on the retained liquid more efficiently than do the other liquid displacement mechanisms usually employed in filters.

Bleeding. If everything else is equal, bleeding of solids to the filtrate is higher in centrifuges than in filters. Therefore, slurries containing a relatively wide particle size distribution are preferably conveyed to filters.

Solids Discharge. Typical cake filters occasionally have problems with removing the filter cake from the filter medium. For example, if the filter cake clings strongly to the filter medium, whether because the solids have adhesive surface properties or because some of the solids are entrapped at higher rate within the interstices of the filter medium, the filter cake no longer drops off the filter medium. Only drum and disk filters with scrapers and the like permit safe removal; with pan, table, leaf, and candle filters, as well as with filter presses, cake removal can present difficulties.

Washing of the Filter Cake and Filter Medium. Some separations also require complete removal of the liquid that remains in the filter cake after dewatering, for example, when a corrosive mother liquor or solvent is involved. For this purpose, filter cakes can be washed. Good possibilities for washing exist in belt, plate, and table filters, because the washing liquid does not remove the filter cake from the filter medium. Newer spraying devices make it possible to wash the filter medium in pressure leaf filters and filter presses as well; for the latter, high pressure spraying devices with up to 10 MPa water pressure are on the market.

Accessibility of Filter Area. It is often desirable that the filter area be accessible to operating personnel for instant intervention in case of a breakdown. Vacuum filters offer the best conditions for access. In line with the general trend toward process automation, however, this need may recede more and more into the background.

Cake Compression. Compressible filter cakes are less suited to pressure filtration than to vacuum filtration, since the relatively high pressure drops which exist in pressure filters favor consolidation of the filter cake. The same is true for centrifuges. However, by combining clarifiers and filters, as in solid bowl centrifuges in which the lip ring is perforated for filtration, compression can become almost insignificant.

Durability. Some filters are distinguished by long durability. This includes particularly filter presses. However, drum and disk filters also have a relatively long durability. The reason for this lies in the design of these filters which minimizes component parts exposed to abrasion and wear. Centrifuges, in contrast, are more subject to wear, especially in those component parts which come into contact with solids at high velocities.

Economy. Some filters have relatively favorable economics such as nutsche filters and candle filters. Centrifugal-discharge filters and filter centrifuges are considerably more expensive. Considering their specific throughputs, however, it is largely found that they also increase with construction cost. Many cost advantages are balanced in this way. However, minimizing costs is always an important development objective; it results in improved and novel filters. An example is seen in the electrokinetic filter, much discussed recently.

Throughput. The throughputs attainable per filter surface area for equivalent products are generally greater in centrifuges than in the conventional filters. For the same reason, slurries which are difficult to filter, can be more efficiently treated in centrifuges (Fig. 3) [3]. Accordingly, filter cakes with small cake formation times also require shorter cycles in centrifuges.

Energy Consumption. The energy consumption per mass rate of liquid or solid separated is only slightly higher in centrifuges than in filters, given the same conditions. Only extremely simple designs such as nutsche filters behave differently.

Occasionally, however, filters consume more energy, for example, if greater demands are made with respect to the degree of washing rate or residual moisture, which frequently requires additional equipment. As a rule, filters and filter centrifuges, including the necessary

auxiliary equipment, consume between 1 and 10 kW h per ton of dry solids. The lower values apply to coarse-grain and crystalline products. The higher values must be expected for fine-grain and especially for pasty materials. The observed behavior of filters often varies widely with a corresponding broad scatter in the separation efficiencies.

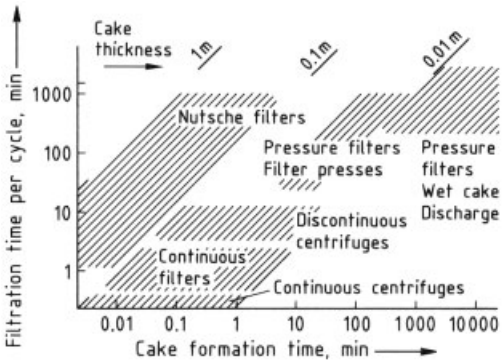

Figure 3. Typical ranges of application of cake filtration equipment
The cake formation time is the time required for the buildup of the first centimeter of filter cake.

Quantitative Comparison. Because the effectiveness of a separation is strongly related to the product, it is difficult to find quantitative comparisons of different separators. A causative factor is the impossibility of making precise studies on identical products in industrial-scale operations.

Some examples, however, have recently been published [4]. In the separation of a pigment that was difficult to filter, one comparison was made between a filter press and a centrifugal discharge plate filter. Higher residual moistures resulted in the latter, especially in opposite membrane filter presses. Of course, to achieve equal elutriation the centrifugal discharge plate filter requires two to three times the amount of wash water. The availability was equal in both cases. Cleanup of the centrifugal discharge plate filter following possible product changes is disproportionately more expensive.

Another comparison between a pusher centrifuge and a belt filter, using a crystalline product, indicated a better durability for the filter. Bleeding of solids is also lower in the filter than in the centrifuge which is provided with relatively large wide slots for economic reasons. Total energy consumption, however, was greater for the filter. Total operating costs of the filter are lower because of the high wear of the centrifuge.

3. References

1. D. A. Dahlstrom: "Selection of Solid-Liquid Separation Equipment", in H. S. Muralidhara (ed.): *Advances in Solid-Liquid Separation,* Battelle Press, Columbus, OH, 1986.
2. H. F. Trawinski: "Current Liquid-Solid Separation Technology", *Filtr. Sep.* **17** (1980) no. 4, 326 – 335.
3. W. Gösele: "Filterapparate–eine Übersicht", *Aufbereit. Techn.* **18** (1977) no. 5, 210 – 214.
4. W. Gösele: "Vergleich von ausgeführten Anlagen zur Fest-Flüssig-Trennung", *Preprints GVC-Dezember-Tagung 1987,* Verein Deutscher Ingenieure, Düsseldorf 1987.

Filtration

See also → Solid – Liquid Separation, Introduction

WALTER GÖSELE, Heidelberg, Federal Republic of Germany (Chaps. 1 – 7)

CHRISTIAN ALT, München, Federal Republic of Germany (Chaps. 8 – 11)

1. Terminology 3
2. Filtration Models 4
2.1. Cake Filtration: Calculation of the Pressure Drop 4
2.1.1. Definition of Filter Resistance and Cake Permeability: The Darcy Equation 4
2.1.2. The "Cake Filter Equation" 5
2.1.3. Evaluation of Experiments with Linear Diagrams 6
2.1.4. Compressible Cake Filtration 8
2.2. Blocking Filtration and other Modes of Filtration 10
2.3. Deep Bed Filtration 11
2.4. Cross-Flow Filtration 14
3. Washing of Filter Cakes 15
3.1. Basic Effects, Mass Balances 15
3.2. Example of Experimental Results . 15
3.3. Test Procedures and Pitfalls 17
3.4. "Intermediate" Deliquoring before Cake Washing 18
4. Deliquoring of Filter Cakes 18
4.1. Deliquoring by Gas Pressure 18
4.1.1. Equilibrium Saturation of Filter Cakes 19
4.1.2. Kinetics of Deliquoring by Gas Pressure 20
4.1.3. Approximate Solution for Coarse, Incompressible Cakes 21
4.1.4. Practical Scale-Up of Deliquoring by Gas Pressure 22
4.1.5. Shrinking and Cracks in Filter Cakes 23
4.2. Deliquoring by Expression 23
5. Optimal Cycle Time 24
6. Understanding the Filter Resistance 26
6.1. The Equation of Carman and Kozeny 26
6.2. Interparticle Forces, DLVO Theory 27
6.3. Mathematical Simulation of Cake Formation 28
7. Solving Filtration Problems in Small-Scale Tests 29
7.1. Laboratory Tests 29
7.2. Handling of "Unfilterable" Suspension 30
8. Filtration Equipment 32
8.1. Bag Filters 33
8.2. Belt Filters 33
8.3. Candle Filters 38
8.4. Deep-Bed Filters 40
8.5. Disk Filters 42
8.6. Drum Filters 43
8.7. Leaf and Plate Filters 50
8.8. Nutsche (Pan) Filters 53
8.9. Pressure Plate Filters 54
8.10. Tubular Filters 59
8.11. Special Filter Types 60
9. Filter Selection 62
10. Filter Media 63
11. Filter Aids 65
12. References 66

Symbols

a intercept in Figure 5, s/mL
A cross-sectional area, filter area, m^2
b slope in Figure 5, s/mL
B creep constant
C creep constant
d_{pore} capillary diameter, m
d_S Sauter mean diameter of particles, m
E electrostatic potential at the solids surface, V
H cake thickness, m
J Boucher's filterability index, m^{-3}
I ionic strength, mol/L
K_H proportionality volume of cake/volume of filtrate, m^3/m^3
K_m proportionality mass of cake/volume of filtrate, kg/m^3
K_N proportionality number of particles/volume of filtrate, m^{-3}
k permeability = $1/\alpha_h$, m^2
L distance from the inlet face of the filter bed, m
m mass of dry filter cake, kg
m_i molar concentration of ion "i", mol/L

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_10

N	number of pores in the filter medium
n	compressibility index
ΔP	dimensionless pressure difference
p_c	capillary pressure, Pa
p_{ce}	capillary entry pressure, Pa
p_{ci}	capillary suction pressure, Pa
p_L	pressure in the liquid, Pa
p_S	compressive stress on the solids, Pa
q	exponent describing blocking behavior
r	distance from the solids surface, m
r_D	Debye length, m
S	saturation (volume of liquid/volume of pores)
S_R	reduced saturation
S_∞	irreducible saturation
t_{reg}	regeneration time for cleaning and preparing a run, s
U_c	consolidation rate
V	volume of filtrate, m^3
$\dot{V}$	volumetric flow rate, m^3/h
z_i	valency of ion "i"
α_H	cake resistance relative to cake thickness, m^{-2}
α_m	cake resistance relative to dry mass, m/kg
β	resistance of filter medium, m^{-1}
ε	porosity = pore volume/total volume
η	viscosity, Ns/m^2
Θ	dimensionless time
λ	filter coefficient of deep bed filters, m^{-1}
ζ	surface potential, V

Indices

av	average
e	end of filtration
H	related to cake thickness (height)
L	liquid
loc	local
m	related to cake mass
N	number
S	solid
w	wash liquor
pore	pore (volume)

1. Terminology

Filtration is the separation of solid particles or liquid ones (droplets) from liquids and gases with the help of a *filter medium* also called a *septum*, which is essentially permeable to only the fluid phase of the mixture being separated. In earlier times, this process was carried out with felts, and the word "filter" has a common derivation with "felt".

Often however, purification of a liquid or gas is called filtration even when no semipermeable medium is involved (as in electrokinetic filtration). This chapter deals only with the filtration of solid – liquid mixtures (suspension, slurries, sludges). For the treatment of gases by filtration, → Dust Separation.

The liquid more or less thoroughly separated from the solids is called the *filtrate, effluent, permeate* or, more rarely, *clean water*. As in other separation processes, the separation of phases is never complete: Liquid adheres to the separated solids (*cake* with *residual moisture*) and the filtrate often contains some solids (*solids content in the filtrate* or *turbidity*).

The purpose of filtration may be clarification of the liquid or solids recovery. In *clarification* the liquid is typically the valuable product and the solids are of minor quantity and are often discarded without further treatment. If however the solids are to be recovered, they very often have to be washed, deliquored and dried (see Fig. 1). In this article, *washing* means the cleaning of a product (filter cake) and it is distinguished from cleaning parts of the filter itself, which will be called *rinsing* (e.g., rinsing a filter screen or a filter cloth by jets of water). A further distinction is to be made between *washing* and *extraction* or *leaching*. *Washing* eliminates liquid contaminants from the pores between the particles of a filter cake. *Extraction* recovers soluble matter from the solid particles themselves (→ Liquid – Solid Extraction). The term *drying* means thermal drying, while the elimination of liquid from the filter cake by mechanical forces is called *deliquoring* or *dewatering*, e.g., deliquoring by gas pressure or by expression.

Filtration is effected by application of *vacuum, pressure*, or of *centrifugal force* (see Fig. 2). *Vacuum filtration* requires a vacuum pump. The pump evacuates the gas from a filtrate receiver, where the filtrate is separated from the gas. The filtrate is drained either by a barometric leg of at least 8 to 10 m or by a pump that is able to "run on snore" (i.e. with a deficiency of feed liquid so that it tends to draw in air). *Pressure filtration* typically only requires a pump for delivering of the suspension and the filter is placed

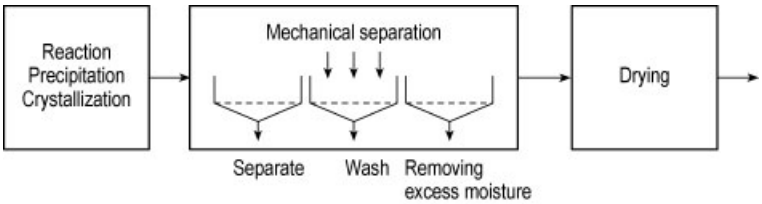

Figure 1. Solids processing chain

within a pressure vessel, hence less easily accessible. *Centrifugal filtration* is done in perforated centrifuge rotors (→ Centrifuges, Filtering).

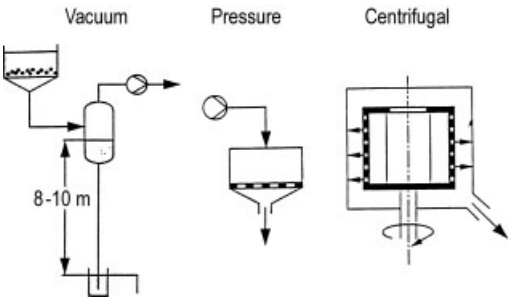

Figure 2. Driving forces in filtration

Vacuum filters have the great advantage that the cake is freely accessible. This facilitates automatic cake handling. However, vacuum filters cannot handle hot liquids, or solvents with high vapor pressure. The pressure difference across vacuum filters is very limited, and the residual moisture of the filter cake is higher than with pressure filters. *Pressure filters* in turn are preferred when the product must be kept in a closed system for safety reasons, or if the residual moisture content is important. The handling of the filter cake is obviously more difficult in a pressure filter. Filtration by *centrifugal force* requires more technical equipment, but as a general rule it yields solids with lower residual moisture (→ Centrifuges, Filtering).

Various models can describe the physical process of filtration. This chapter concentrates on four idealized filtration models depicted in Figure 3.

– *Cake filtration* is the most frequently used model. Here it is assumed that the solids are deposited on a filter medium as a homogeneous porous layer with a constant permeability. Thus, if the flow rate dV/dt is constant, the pressure drop will increase linearly, proportional to the quantity of solid deposited. This model can be applied particularly well for all hard, particulate solids.

– *Blocking filtration*: The pressure drop is caused by solid particles blocking pores. Soft, gelatinous particles retained by a sieve exhibit such a behavior. If the flow rate dV/dt is constant, the pressure drop increases exponentially with the quantity filtered, the number of open pores asymptotically approaching zero. The pores may belong to a filter medium (screen or filter layer) or it may be pores within a filter cake of coarse particles, which are blocked by migrating fine particles.

– *Deep bed* or *depth filtration*: Solid particles are retained in a deep filter layer. This takes place for example in sand filters for clarification of drinking water, which retain even colloidal particles. The typical effect of deep bed filtration is adhesion of solids to the grains of the filter layer, comparable to charcoal adsorption. Only rather big particles are retained by the screening effect. When the filter bed has been saturated with solids, the solids concentration in the filtrate leaving the bed progressively approaches that of the incoming suspension.

– *Cross-flow filtration*: In cross-flow filtration the suspension flows with high speed tangentially to the filter surface, preventing the formation of a cake. Only a small flow of liquid passes through the filter medium. A certain layer of solids accumulates in the boundary layer on the filter surface, and reduces the flow of filtrate. After an initial period, a dynamic equilibrium is established between convective transport of solids to the filter surface and removal of solids by turbulence and by diffusion.

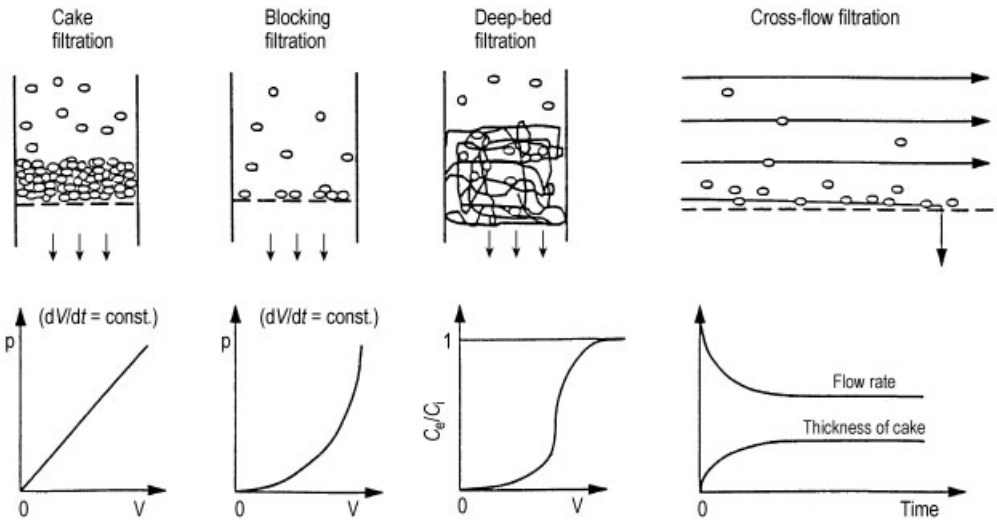

Figure 3. Filtration models

Surface filtration is the antonym to *depth filtration*. The solids are retained on the surface of a filter medium. Generally the models of cake filtration or of blocking filtration can be applied.

Screening designates a classification process, which retains the particles below a certain size and lets pass the smaller ones (→ Screening). Often the term *screening filtration* is also used to designate a surface filtration with a screen as a filter medium. Its mode of action resembles screening (or straining) as long as the filter medium is clean, but it is clearly a cake filtration as soon as a layer of solids has formed.

2. Filtration Models

2.1. Cake Filtration: Calculation of the Pressure Drop

2.1.1. Definition of Filter Resistance and Cake Permeability: The Darcy Equation

The resistance to flow of a filter cake can be described by Darcy's law [1] (see Fig. 4). Consider a liquid flowing through a filter cake (or a stream of water percolating through soil as considered by Darcy). The pressure drop Δp of this flow is proportional to:

1) The flow rate per unit area V/A
2) The cake thickness H
3) the viscosity η of the liquid
4) A constant α_{H} describing the "specific filter resistance" of the cake:

$$\Delta p_1 = \left(\frac{\dot{V}}{A}\right) \cdot H \cdot \eta \cdot \alpha_{\mathrm{H}} \tag{1}$$

The SI-units are $\left[\mathrm{Pa} = \left(\frac{\mathrm{m}^3/\mathrm{s}}{\mathrm{m}^2}\right) \cdot \mathrm{m} \cdot \mathrm{Pa} \cdot \mathrm{s} \cdot \mathrm{m}^{-2}\right]$

The unit of α_{H} must therefore be m^{-2} in order to satisfy the Darcy equation (1). The reciprocal of the filter resistance α_{H} is also called permeability k of the filter cake:

$$k = \frac{1}{\alpha_H} \left[\mathrm{m}^2\right] \tag{2}$$

Sometimes it is more convenient to define cake thickness in terms of solid mass per unit area (unit kg/m^2). This leads to a slightly different definition of filter resistance, the factors being:

1) The flow rate per unit area V/A
2) The cake thickness m/A
3) The viscosity η of the liquid
4) A constant α_{m} with the unit m/kg describing the resistance of the cake.

Then the following expression is obtained instead of Equation (1)

$$\Delta p_1 = \left(\frac{\dot{V}}{A}\right) \cdot \left(\frac{m}{A}\right) \cdot \eta \cdot \alpha_{\mathrm{m}} \tag{3}$$

For practical reasons the viscosity η is very often not measured separately. Then it is legitimate to include it in a term $\alpha_{\mathrm{H}}\eta$ (unit mPa · s/m^2) or $\alpha_{\mathrm{m}}\eta$ (unit mPa · s · m/kg), respectively.

Using this latter term $\alpha_{\mathrm{H}}\eta$ or $\alpha_{\mathrm{m}}\eta$ filter resistances lie between 10^{11} mPa · s/m^2 (filtering very rapidly) and 10^{16} mPa · s/m^2 (nearly unfilterable), or between 10^8 and 10^{13} mPa · s · m/kg, respectively.

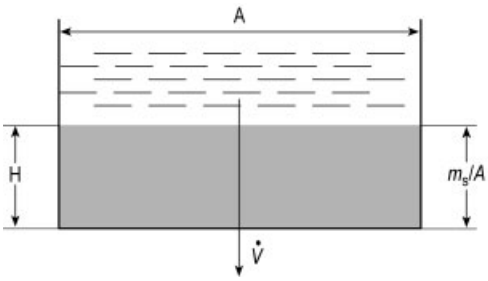

Figure 4. Definition of filter resistance
$\eta\alpha_{\mathrm{H}}$ = Cake thickness H related filter resistance; $\eta\alpha_{\mathrm{m}}$ = Solid mass related filter resistance

2.1.2. The "Cake Filter Equation"

The pressure drop in a filter is composed of a pressure drop Δp_1 across the cake according to Equation (1) and a pressure drop Δp_2 across the filter medium, which can be written as:

$$\Delta p_2 = \beta \cdot \eta \cdot \left(\frac{\dot{V}}{A}\right) \tag{4}$$

where b (unit m^{-1}) is the resistance of the filter medium.

The total pressure drop is therefore:

$$\Delta p = \Delta p_1 + \Delta p_2 = \alpha_{\mathrm{H}}\eta \cdot H \cdot \frac{\dot{V}}{A} + \beta\eta\frac{\dot{V}}{A} \tag{5}$$

or

$$\Delta p = \Delta p_1 + \Delta p_2 = \alpha_{\mathrm{m}}\eta \cdot \frac{m \cdot \dot{V}}{A^2} + \beta\eta\frac{\dot{V}}{A} \tag{6}$$

If the suspension is homogeneously mixed, the cake height H (or m/A) will be proportional to the quantity of filtrate. The concentration is described by a factor K:

$$K_{\mathrm{H}} = \frac{H \cdot A}{V} \tag{7}$$

or

$$K_{\mathrm{m}} = \frac{m}{V} \tag{8}$$

this gives

$$\Delta p = \frac{\alpha_{\mathrm{H}}\eta K_{\mathrm{H}}}{A^2} \cdot V \cdot \frac{\mathrm{d}V}{\mathrm{d}t} + \frac{\beta\eta}{A} + \frac{\mathrm{d}V}{\mathrm{d}t} \tag{9}$$

or

$$\Delta p = \frac{\alpha_{\mathrm{m}}\eta K_{\mathrm{m}}}{A^2} \cdot V \cdot \frac{\mathrm{d}V}{\mathrm{d}t} + \frac{\beta\eta}{A} + \frac{\mathrm{d}V}{\mathrm{d}t} \tag{10}$$

Equations (9) and (10) are identical under the generalization $\alpha_{\mathrm{H}} \cdot K_{\mathrm{H}} = \alpha_{\mathrm{m}} \cdot K_{\mathrm{m}}$. Nevertheless the distinction between both equations is useful for clarity.

The differential Equations (9) and (10) can be integrated either for constant flow rate or for constant pressure. Integration for constant flow rate $\mathrm{d}V/\mathrm{d}t$ = const gives the trivial solution:

$$\Delta p = \frac{\eta \cdot \dot{V}}{A} \cdot \left(\frac{\alpha_{\mathrm{H}} K_{\mathrm{H}}}{A} \cdot V + \beta\right) \tag{11}$$

or

$$\Delta p = \frac{\eta \cdot \dot{V}}{A} \cdot \left(\frac{\alpha_{\mathrm{m}} K_{\mathrm{m}}}{A} \cdot V + \beta\right) \tag{12}$$

For Δp = const. the integration yields:

$$\mathrm{d}t = \frac{\alpha\eta K}{A^2 \cdot \Delta p} \cdot V \cdot \mathrm{d}V + \frac{\beta\eta}{A \cdot \Delta p} \cdot \mathrm{d}V \tag{13}$$

$$t = \frac{\alpha_{\mathrm{H}}\eta K_{\mathrm{H}}}{2A^2 \cdot \Delta p} \cdot V^2 + \frac{\beta\eta}{A \cdot \Delta p} \cdot V \tag{14}$$

or

$$t = \frac{\alpha_{\mathrm{m}}\eta K_{\mathrm{m}}}{2A^2 \cdot \Delta p} \cdot V^2 + \frac{\beta\eta}{A \cdot \Delta p} \cdot V \tag{15}$$

2.1.3. Evaluation of Experiments with Linear Diagrams

Linear Diagram in the Differential Form. It is often very helpful to plot the instantaneous resistance to flow as a function of the quantity filtered. The resistance is characterized by the pressure drop Δp related to the instantaneous flow rate $\mathrm{d}V/\mathrm{d}t$. Using Equation (5) or (6) this can be written as:

$$\frac{\mathrm{d}t}{\mathrm{d}V} = \frac{\alpha_{\mathrm{H}}\eta K_{\mathrm{H}}}{\Delta p A^2} \cdot V + \frac{\beta\eta}{\Delta p A} = \frac{\alpha_{\mathrm{m}}\eta K_{\mathrm{m}}}{\Delta p A^2} \cdot V + \frac{\beta\eta}{\Delta p A} = b \cdot V + a \tag{16}$$

As shown in Figure 5 left diagram [2], interpolation gives a straight line with slope b and intercept a. The intercept

$$a = \frac{\beta\eta}{\Delta p}$$

represents the resistance at the very first moment of filtration, before a cake is formed, hence the resistance of the filter medium including the boundary layer to the cake.

The slope b contains the filter resistance α according to

$$b = \text{slope} = \frac{\Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)}{\Delta V} = \frac{\alpha\eta K}{A^2} \tag{17}$$

Inserting K_H or $\mathrm{K_m}$ respectively, gives

$$\alpha_\mathrm{H}\eta = \frac{\Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)}{\Delta V} \cdot \frac{A^2}{K_\mathrm{H}} = \frac{\Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)}{\Delta V} \cdot \frac{A \cdot V}{H}$$
$$= \Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)_\mathrm{e} \cdot \frac{A}{H_\mathrm{e}} \tag{18}$$

and

$$\alpha_\mathrm{m}\eta = \frac{\Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)}{\Delta V} \cdot \frac{A^2}{K_\mathrm{m}} = \frac{\Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)}{\Delta V} \cdot \frac{A^2 \cdot V}{m}$$
$$= \Delta\left(\Delta p \cdot \frac{\mathrm{d}t}{\mathrm{d}V}\right)_\mathrm{e} \cdot \frac{A^2}{m_\mathrm{e}} \tag{19}$$

Linear Diagram in Integrated Form. Another approach starts from the integrated filter Equations (14) or (15). The experimental results are plotted in the form of

$$\frac{t}{V} = f\,(V)$$

If the pressure is constant during the experiment, a straight line should be obtained

$$\frac{t}{V} = \frac{\alpha_\mathrm{m}\eta K_\mathrm{m}}{2A^2 \cdot \Delta p} \cdot V + \frac{\beta\eta}{A \cdot \Delta p}$$
$$= \frac{\alpha_\mathrm{H}\eta K_\mathrm{H}}{2A^2 \cdot \Delta p} \cdot V + \frac{\beta\eta}{A \cdot \Delta p} = \frac{b}{2} \cdot V + a \tag{20}$$

This kind of plot is still very popular in practice because the evaluation is often easier than with Equation (16). Experimental values measured with a bucket and a stopwatch can be inserted directly into Equation (20), while for a differential diagram they have to be converted into momentary flow rates $\mathrm{d}V/\mathrm{d}t$.

On the other hand Equation (20) is correct only for $\Delta p = \text{const}$. Furthermore, the resulting diagram shows less clearly the deviations from linearity than a diagram based on instantaneous flow rates. If for example a cake stops growing and keeps a constant resistance this gives a well discernible plateau in a differential diagram, but only a gradual hyperbolic bend in the integrated diagram.

Example. The interpolation of the test results shown in Figure 5 leads to the intercept a and the slope b

$$a = 3.6\mathrm{s/mL} = 3.6 \times 10^6\,\mathrm{s/m^3} \tag{21}$$

and

$$b = \frac{28\mathrm{s/mL}}{150\mathrm{mL}} = 1.87 \cdot 10^{-11}\,\mathrm{s/m^6} \tag{22}$$

The following parameters are known:

Filtration pressure $p = 1\,\mathrm{bar} = 10^5\,\mathrm{N/m^2}$
Filter area $A = 20\,\mathrm{cm^2} = 20 \times 10^{-4}\,\mathrm{m^2}$
Viscosity $\eta = 0.001\,\mathrm{Pa \cdot s} = 10^{-3}\,\mathrm{Ns/m^2}$
Cake height $H_\mathrm{e} = 37\,\mathrm{mm}$
Cake mass $m_\mathrm{e} = 24.3\,\mathrm{g}$
Volume of filtrate $V_\mathrm{e} = 154\,\mathrm{cm^3}$

Solids in suspension

$$K_\mathrm{H} = \frac{A \cdot H_\mathrm{e}}{V_\mathrm{e}} = \frac{20 \cdot 10^{-4} \cdot 0.037}{154 \cdot 10^{-6}} = 0.481 \tag{23}$$

or

$$K_\mathrm{m} = \frac{m_\mathrm{e}}{V_\mathrm{e}} = \frac{24.3 \cdot 10^{-3}}{154 \cdot 10^{-6}} \cdot \frac{\mathrm{kg}}{\mathrm{m^3}} = 158\,\frac{\mathrm{kg}}{\mathrm{m^3}} \tag{24}$$

This gives for the filter medium resistance

$$\beta = \alpha \cdot \frac{\Delta p \cdot A}{\eta} = 3.6 \cdot 10^6 \cdot \frac{10^5 \cdot 20 \cdot 10^{-4}}{10^3}\,\mathrm{m^{-1}}$$
$$= 7.2 \cdot 10^{11}\,\mathrm{m^{-1}} \tag{25}$$

From Equations (18) or (19) follows the cake filter resistance

$$\alpha_\mathrm{H} = b \cdot \frac{A^2 \cdot \Delta p}{K_\mathrm{H}\eta} = 0.187 \cdot 10^{11} \cdot \frac{400 \cdot 10^{-8} \cdot 10^5}{0.481 \cdot 10^{-3}}\,\mathrm{m^{-2}}$$
$$= 1.56 \cdot 10^{14}\,\mathrm{m^{-2}} \tag{26}$$

$$\alpha_\mathrm{m} = b \cdot \frac{A^2 \cdot \Delta p}{K_\mathrm{m}\eta} = 0.187 \cdot 10^{11} \cdot \frac{400 \cdot 10^{-8} \cdot 10^5}{158 \cdot 10^{-3}}\,\mathrm{m^{-2}}$$
$$= 4.73 \cdot 10^{11}\,\frac{\mathrm{m}}{\mathrm{kg}} \tag{27}$$

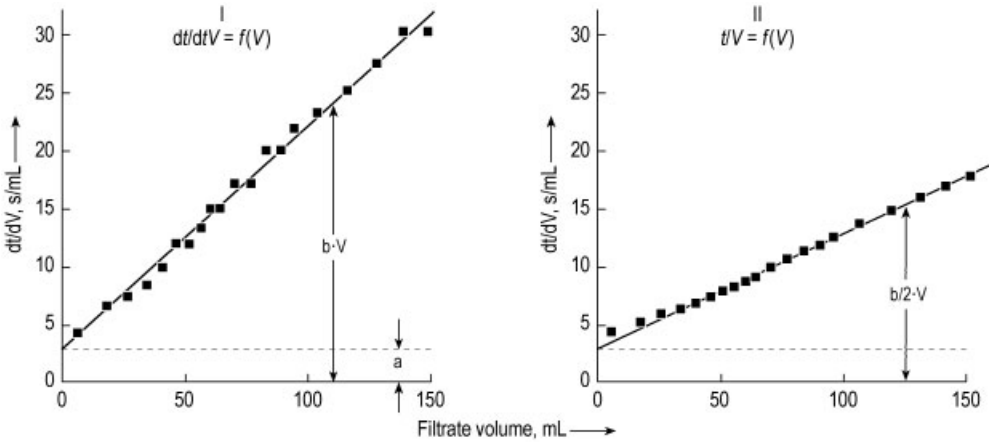

Figure 5. Linear plots according to Eqs. (8) and (11) [41]

Deviations from Linearity. Experimental data very often differ from the linear characteristic shown in diagrams "I" or "II" of Figure 5. The kind of deviation indicates the nature of secondary effects (see Fig. 6):

- A: Theoretical linear curve without secondary effect.
- B: Some solids have settled before the filtration started, apparently increasing the resistance of the filter medium.
- C: At the beginning of filtration the filtrate has been turbid. The solids loss diminishes the amount of cake, apparently indicating a negative medium resistance. (When no turbidity is found in the filtrate, this curve may indicate that solids accumulate inside the filter medium, blocking it in the long run!). Similar curves may be observed when the solids float and clear liquid is filtered in the beginning.
- D: The solids settle out completely, increasing the speed of cake growth. At the end of filtration clear liquid flows through the cake of constant resistance.
- E: Only coarse particles settle. After some time only the remaining fines are filtered and the resistance increases progressively.
- F: Fine particles trickle through the cake and block the pores of the cake or of the filter medium. The process can perhaps be described as a blocking filtration (see Section 2.2).
- Further causes of deviations might be:

Determination of the beginning or the end point of filtration may not be correct. The end point is particularly difficult to determine if there is no sight glass in the filter. It is therefore common practice to neglect some points at the end of the measurement.

Unusual properties of the suspension like foam bubbles or oil droplets, non-Newtonian rheology of the liquid or non-isometric, plate-like particles with special orientation can produce a great variety of effects.

The sediment may be so voluminous that it fills the whole suspension. The process then resembles rather an expression than a filtration (Section 4.2).

Freshly precipitated solids sometimes continue to agglomerate or to "ripen" during filtration, changing their effective particle size.

2.1.4. Compressible Cake Filtration

Most filter cakes are compressible, which means that their resistance increases with growing pressure. An x-fold increase in filtration pressure therefore normally gives rise to a less than x-fold increase in flow rate. The compression of the cake is caused by the compressive stress p_s on the solids which is caused by the drag of the flowing liquid (see Fig. 7). The loss in liquid

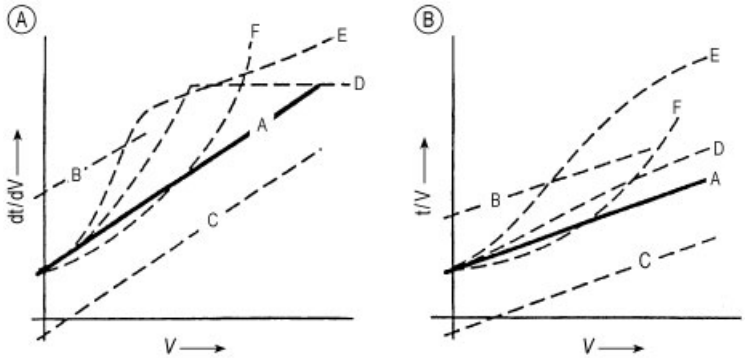

Figure 6. Typical deviations from linearity [41]
A) Differential plot according to Eq. (16); B) Integral plot according to Eq. (20)

pressure p_L translates into solids pressure p_S, the sum of both being constant:

$$\partial p_S = -\partial p_L \tag{28}$$

$$p_S + p_L = p_0 \tag{29}$$

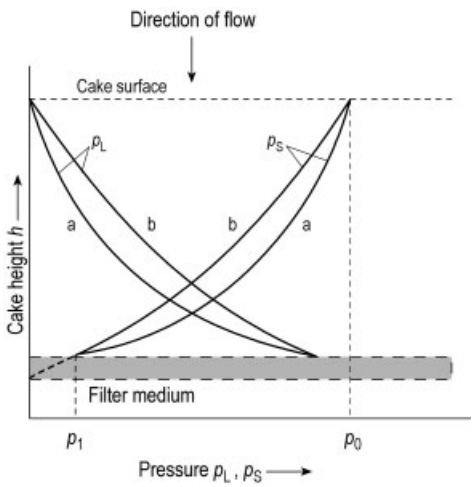

Figure 7. Hydraulic pressure p_L and solids pressure p_S as function of cake height. Curve (a) represents a more compressible cake than curve (b)

The *local filter resistance* α_{loc} in the cake is a function of the local compressive stress p_S and it is therefore low at the surface of the cake and high near the interface with the filter medium. This $\alpha_{loc} = f(p_S)$ can be measured in the Compression Permeability cell (CP cell) described in Chapter 7, where a sample of the cake is subject to a known compressive stress.

However, only the *average resistance* α_{av} of a filter cake, all layers combined, is important for practical scale-up purposes. It can be calculated from α_{loc} measured in the CP-cell [3]

$$\alpha_{av} = \frac{\Delta p}{\int_0^p \frac{dp}{\alpha_{loc}}} \tag{30}$$

Normally however it is much easier to measure α_{av} directly with filtration experiments in a pressure filter as described in Section 7.1. The uneven distribution of local porosity and filter resistance is then ignored.

The dependence of filter resistance α_{av} or α_{loc} on pressure can be approximated over a limited pressure range by

$$\alpha = \alpha_0 \cdot \left(\frac{p_S}{\text{unit pressure}} \right)^n \tag{31}$$

where α_0 is the resistance at unit pressure drop and n is a compressibility index (equal to zero for incompressible cakes). Both parameters can be determined easily from a logarithmic plot $\alpha = f(p_S)$, where the slope indicates the compressibility factor n. Introducing α_{loc} according to this approximation into Equation (19), it can be shown that for $0 < n < 1$

$$n_{av} = n_{loc} \tag{32}$$

$$\alpha_{0,av} = (1 - n) \cdot \alpha_{0,loc} \tag{33}$$

Thus the approximation Equation (31) has the advantage that it applies to both the local and the average resistance. However, such an approximation is valid only for a limited range of pressures and Equations (32) and (33) are restricted to $0 < n < 1$.

TILLER and coworkers [4] give an integration of the filter equation also for $n > 1$. According to their equations the flow rate quickly reaches a constant value with increasing pressure. For practical purposes it can be approximated that such "supercompactible" filter cakes with $n > 1$ (e.g., a great number of waste water sludges) yield the same flow rate independently of the pressure applied: increasing pressure only increases the compressed layer adjacent to the filter cloth. (The hypothetical cake which yields the highest flow at an "optimal pressure" and less flow at higher pressures apparently does not exist. In practical application, however, the output of a filtration may be best with moderate pressure, since this avoids excessive blocking of the filter medium).

For practical purposes Equations (31) – (33) are often ignored and compressible cakes are treated with the same equations as incompressible cakes, provided α is defined as average specific cake resistance under the conditions of operation.

Concerning deliquoring of compressible cakes, it is important to know that the porosity is unevenly distributed after filtration. The filter resistance often is concentrated in a thin, compressed layer facing the filter medium, the rest of the cake being very porous and wet. Deliquoring by expression (see Section 4.2) will then be very effective, even at moderate pressures, since it acts also on the upper layer of the cake.

2.2. Blocking Filtration and other Modes of Filtration

According to the idealized idea of blocking filtration, every particle in the suspension is retained on the filter medium and blocks one pore. The remaining number of open pores is then given by the equation

$$N(t) = N_0 \left(1 - K_\mathrm{N} \cdot V_\mathrm{L}\right) \tag{34}$$

where N_0 is the number of pores in the clean filter medium, $N(t)$ is the remaining number of open pores and K_N (m^{-3}) is the concentration of particles by number in the suspension. At constant pressure drop the volume of filtrate $V(t)_{\Delta p}$ = const will follow the equation [5]

$$V(t) = \frac{1}{K_\mathrm{N}} \cdot \left[1 - \exp\left(-K_\mathrm{N} \cdot \dot{V}_0 \cdot t\right)\right] \tag{35}$$

with $\dot{V}_0$ (m^3/s) being the flow rate at the start of filtration with a clean filter medium.

In real filtration processes not every particle in the suspension will block a pore of the filter medium. Some particles will be retained by adhesion to the walls of a pore without blocking it totally; others will pass through the filter medium without being retained at all. Different formulae (or "filtration laws") exist to describe such moderate blocking. A generalized model is based on the equations

$$\frac{\mathrm{d}^2 t}{\mathrm{d}V^2} = \mathrm{const} \cdot \left(\frac{\mathrm{d}t}{\mathrm{d}V}\right)^q \tag{36}$$

for constant pressure filtration and

$$\frac{\mathrm{d}\left(\Delta \mathrm{p}\right)}{\mathrm{d}V} = \mathrm{const} \cdot \left(\Delta \mathrm{p}\right)^q \tag{37}$$

for constant rate filtration.

The exponent q in this equation varies between 0 and 2 and describes the "blocking speed":

$q = 0$ cake filtration, slow blocking
$q = 1$ so-called intermediate filtration
$q = 3/2$ so-called standard blocking filtration
$q = 2$ complete blocking.

A summary of the corresponding equations for flow rates and filtrate quantities is given in [6], [7]. Figure 8 visualizes some examples: Surface straining represents the classical example of complete blocking ($q = 2$), but also depth straining follows the same characteristic. Normal depth filtration in contrast (with particles much smaller than the filter pores) leads to pressure drop by diminishing pore size and can be described by the standard blocking equation ($q = 3/2$).

In most filtrations it is highly probable that more than one filtration mode occurs. Cake filtration for example often is preceded by an initial period before cake build-up which can be described as standard blocking. This period is characterized by depth filtration (particles penetrating into the filter medium) and/or by fines passing through the filter medium (creating initial turbidity of the filtrate). For scientific investigations it may then be helpful to identify the filtration mechanism more precisely by parameter-fitting the experimental data and quantifying the above-mentioned "blocking speed".

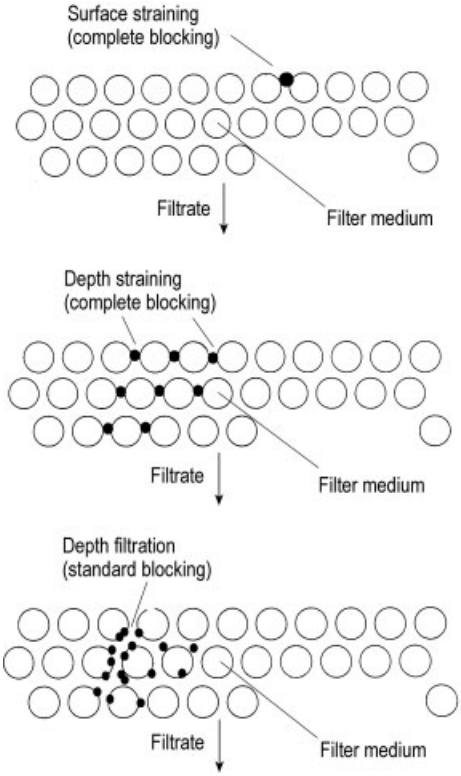

Figure 8. Three possible modes of filtration: Surface straining and depth straining are described as "complete blocking" ($q = 2$). Depth filtration of particles much smaller than the pore size is best described as "standard blocking" ($q = 3/2$) [6]

Simplified Evaluation of Experimental Data. A very pragmatic approach is often used to describe the mode of filtration. The pressure drop at constant flow rate is depicted on a semi-logarithmic scale as a function of the volume filtered. Normally the values can be approximated by a straight line according to [8]

$$\log\left(\frac{p}{p_0}\right) = -J \cdot V_{\mathrm{L}} \tag{38}$$

For the flow rate at constant pressure drop the same approximation holds

$$\log\left(\frac{\dot{V}}{\dot{V}_0}\right) = -J \cdot V_{\mathrm{L}} \tag{39}$$

Here p_0 is the initial pressure drop of the clean filter; and the gradient J, which represents something like the above-mentioned blocking speed, is "Boucher's filterability index". If $J = 1$ Equation (38) and (39) are identical to an "intermediate filtration" with $q = J = 1$. For other values of J however the physical meaning of J is unclear. For scientific purposes this approach is therefore not so well accepted. For practical purposes, however, it is very useful.

2.3. Deep Bed Filtration [9], [10]

In deep bed filtration (as opposed to cake filtration and cross-flow filtration), the solid particles are separated mainly by deposition within the pores of the filter medium. The filter medium may consist of

- a 0.5 – 3 m layer of coarse grains (e.g., sand 0.3 – 5 mm) or
- a layer of a few centimeters of fibers (e.g., wound or resin bonded cartridge filters) or
- sheets of a few mm thickness (e.g., made of cellulose).
- a layer of granular filter aid (e.g., precoat layer)

All these filter elements have pores that are larger than the particles to be retained. The particles stick in the pores by adhesion and their accumulation weakens the filtering action and increases the pressure drop, so that the filter must be cleaned or replaced periodically. This is why deep bed filtration should be used not to recover solids from a suspension such as cake filtration, but instead to produce a very clean effluent from suspensions with low solids loading (typically < 0.1 g/L) and with very fine particles. Depending on the particle size, the prevailing effects of retention are summarized in Figure 9:

1) Particles larger than the pores are trapped mechanically. This is typically true for particles $\gg 10\,\mu$m.
2) Particles 1 – 10 µm in diameter hit the solid surface mainly due to inertia effects and stick there due to surface forces.
3) Particles $< 0.1\,\mu$m reach the solid surface mainly because of diffusion and again they stick there due to surface forces.

In the range between 0.1 and 1 µm the effects of both inertia and diffusion are small and a minimum in transport and hence in the effectiveness of deep bed filters is observed.

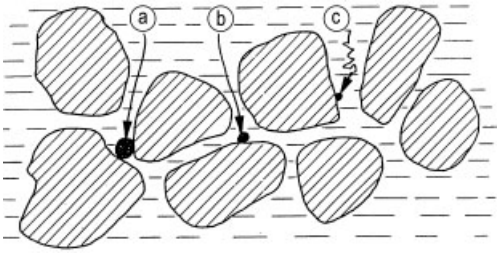

Figure 9. Retention mechanisms in a deep bed filter
a) Particles $> 10\,\mu m$ are retained by mechanical interception; b) Particles of ca. $1\,\mu m$ size are subject to inertial impact + adhesion; c) Particles $< 1\,\mu m$ follow mainly Brownian diffusion + adhesion

Thus the transport of particles to the solid surface is rather well explained and can be described mathematically. The mechanism of adhesion to the filter grains however and the resulting "sticking probability" is less well understood. (Some aspects will be discussed in Section 6.2). Consequently the clarification effect must be summarized in an empirical "filter coefficient" λ describing the local decrease in concentration of the suspension flowing through the bed:

$$-\frac{\partial C}{\partial L} = \lambda C \tag{40}$$

where c is the concentration of the suspension and L is the distance from the inlet face of the filter. For uniform conditions, this differential equation can be integrated:

$$c = c_0 \cdot \exp\left(-\lambda_0 \cdot L\right) \tag{41}$$

here λ_0 is the initial filter coefficient of a clean, homogeneous filter bed. As soon as the bed is loaded with solids, its efficiency will diminish and that is why the solution of the differential equation becomes rather difficult. Different models exist to describe the process, but they are rarely used for practical purposes. To find a suitable filter medium in a deep bed filter which shows good retention efficiency over a long cycle time, laboratory tests over a realistic cycle time have to be carried out with filter layers of realistic depth.

The *pressure drop* in a deep bed filter can be interpreted as an effect of two different phenomena of blocking filtration:

1) The filter media exhibit a resistance to flow, which is increased by the solids deposit in the pores. The quantity of deposit is generally small compared to the pore volume, and the additional pressure loss per unit depth can be described as proportional to the local specific deposit. For constant solids concentration at the inlet and constant retention the increase in this pressure drop with time is therefore approximately constant.
2) In addition there is often a pressure drop due to deposits on the surface of the filter bed. This is a typical blocking filtration and can be described by Boucher's law (Eqs. 38 and 39).

The total pressure drop of a deep bed filter at constant flow rate can then be approximated by an equation of the type

$$p\,(t) = \mathrm{const}_1 \cdot t + \mathrm{const}_2 \cdot e^{J \cdot t} \tag{42}$$

The pressure drop should be measured in a test filter with a vertical height close to that of the full-scale unit (Fig. 10 A) [11]. For the case that the driving force is gravity, the pressure profile is shown in Figure 10 B. In the static equilibrium, at flow zero, 1 m of pressure head is gained for every meter of depth. The downstream consumers may then draw a constant flow rate. When flow has started, the pressure drop within the media increases linearly with depth. As solids are deposited in the pores, the local pressure loss will increase in the upper layers and the pressure line becomes distorted. When the pressure line touches the atmospheric pressure value, the required flow cannot be maintained and the filter should be cleaned.

Cleaning of Deep Bed Filters. Sand filters can be cleaned by backflushing. The flow is reversed to wash off deposited solids and flush them, e.g., to a wastewater station. Often air scour is used to increase the turbulence and reduce the amount of water necessary for fluidization. A typical set of data of a sand filter for cleaning of water could be [12]:

Medium depth	1 m
Medium size	0.5 – 1 mm (as uniform as possible)
Flow rate during filtration	$5 - 10\,m^3\,m^{-2}\,h^{-1}$
Flow rate air scour	$60\,m^3\,m^{-2}\,h^{-1}$ for 10 min
Flow rate water backflow	$15 - 30\,m^3\,m^{-2}\,h^{-1}$ for 3 – 8 min after air scour

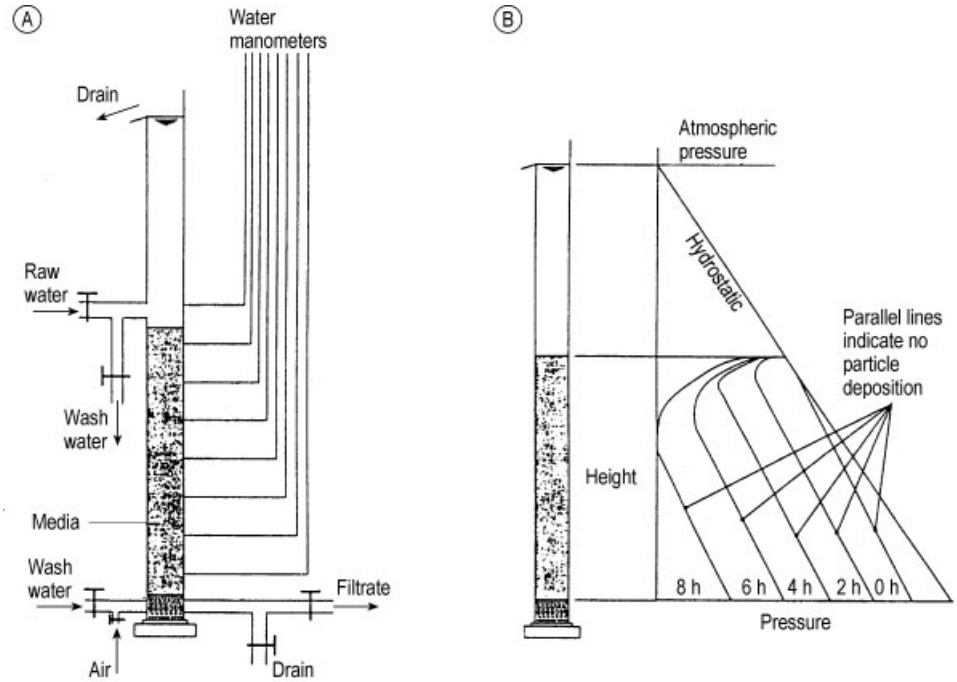

Figure 10. A) Laboratory test equipment for sand filters
B) Pressure profiles in test filter [11]

During backflow the sand grains are classified, the smallest ones are entrained to the top, the coarse ones sink to the bottom of the filter bed. This effect is undesired, since it leads to premature blocking of the top layer. Therefore it is important to use sand with grain sizes as uniform as possible.

Sizing of a Deep Bed Filter. The sizing of a deep bed filter consists of the following steps:

- Preselection of the filter material by comparative laboratory tests with small layers of different materials under standard conditions and with the filtrate quality as a criterion.
- Selection of the right grain size in tests with a filter column of 1 m height and a realistic flow rate and pressure drop. If necessary the grain size must be adapted so that the bed is saturated (the filtrate gets turbid) at the same time as the maximum pressure drop is reached. If the maximum pressure drop is reached first, the grains should be coarser and vice versa.

Special Variations of Deep Bed Filters. As mentioned above the sand grains are classified during backflush and the uppermost layer is most efficient in collecting and becomes clogged rapidly by captured solids. This disadvantage is diminished when a second layer of coarse material with lower specific weight is added on top of the sand layer. During backflow these low density grains collect at the bed surface and form a coarse layer able to retain a great quantity of dirt (so-called *multimedia filters* or *dual media filters*, e.g., with a layer of anthracite particles 1 – 2 mm in size on top of a main layer of sand grains 0.5 – 1 mm in size.)

Other designs of sand filters (especially in the United States) convert the disadvantage of classification into an advantage by operating in the upflow mode with filtering and backflushing from below. This offers the advantage that large quantities of incoming dirt are retained in the lower layer of sand, which contains the coarser grains and larger pores. These upflow filters however need special safeguards against breakthrough of dirt when the pressure drop becomes high.

In order to increase the retention capacity in a given volume, it has also been proposed to use porous grains instead of sand grains. Small pieces of polymeric foam have indeed this effect [13]. Cleaning by simple backflush is, however, not possible with this material, it must be cleaned by backflush combined with compression, and this method has therefore not yet found practical application.

2.4. Cross-Flow Filtration

In cross-flow filtration the deposition of particles on the filter membrane is prevented by a strong flow parallel to the filter surface. This so called cross-flow is achieved in most cases by pumping the suspension through bundles of membrane tubes (see Section 8.10, Figs. 77, 78). Alternatively the cross-flow is achieved by rotating inserts (Section 8.9, Fig. 76). The elimination of a cake enables to filter very fine particles, which otherwise would form a cake with prohibitively high resistance. Even submicron, nonparticulate matter can be retained by semipermeable membranes according to this principle. In this chapter a short overview will only be given, covering both the filtration of particles and the separation of non-particulate matter. For more details see → Membranes and Membrane Separation. Depending on the size of the "particles" retained, a distinction is made between microfiltration, ultrafiltration, nanofiltration, and reverse osmosis. Typical parameters of cross-flow filtration are summarized in Table 1.

Microfiltration retains solid particles (see Section 8.11). According to the usual definition of "solid particles" these are visible in the microscope, hence > 0.1 µm in size. In steady-state cross-flow filtration these particles are conveyed onto the membrane by convection due to the filtrate flow and transported away from it by hydrodynamic lift forces due to the parallel shear flow (forces F_Y and F_L in Fig. 11). For particles below a certain size the lift force becomes smaller than the convection, $F_L < F_Y$, and they are deposited on the membrane [14]. After deposition they are retained by van der Waals' adhesion forces F_A and not easily swept away even if there is no filtrate flow (irreversible cake formation). Steady state microfiltration therefore requires a parallel flow strong enough to flush off the finest fraction of particles in the suspension. (In most practical applications colloidal particles are also present and they are transported according to the mechanisms of ultrafiltration.)

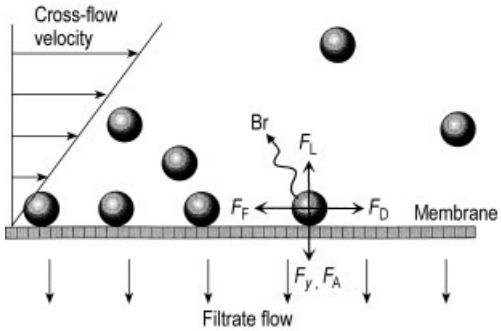

Figure 11. Forces and transport effects onto a particle in cross-flow filtration
Br = Brownian motion; F_A = adhesion to the membrane; F_D = drag from the cross-flow; F_F = friction force; F_L = hydrodynamic lift force; F_Y = drag from then filtrate flow

Ultrafiltration retains colloidal particles or macromolecules. The cut-off sizes are ca. < 0.1 µm. Such particles are subject to Brownian diffusion (Br in Fig. 11) and only weakly influenced by the hydrodynamic lift force [15]. The smaller these particles are, the more they are transported away from the filter medium by diffusion. This is why in ultrafiltration the smaller particles are preferably swept away and the bigger ones are collected on the membrane (contrary to microfiltration). The cut-off size of an ultrafiltration is usually defined as the molar mass (in g/mol) of a test suspension (generally protein or dextran molecules of defined molar mass in the range of $10^3 - 10^5$ g/mol). In practical application the cut-off size depends not only on the pore size of the membrane, but to a large degree on a "gel" layer on the membrane consisting of retained colloids.

Nanofiltration is a relatively new technique combining features of ultrafiltration and reverse osmosis with a high selectivity. Its name is derived from its approximate cut-off size of some nanometers or more exactly molar masses of 200 – 1000 g/mol. This is achieved with special nanofiltration membranes which still have pores of a defined size, but their retention depends

on the electrostatic charge of the molecules to be separated (bivalent anions are typically retained).

Reverse osmosis retains molecules or ions using selective membranes without pores. Certain molecules permeate through the membrane because they are soluble in the membrane material. Other molecules are not (or less) soluble and are retained (or concentrated) on the upstream side of the membrane. A particular feature of reverse osmosis is the high pressure required to overcome the osmotic pressure of the retained molecules. An important application is desalination of seawater. In the food industry it is applied to concentrate juices at low temperatures.

3. Washing of Filter Cakes

3.1. Basic Effects, Mass Balances

The purpose of washing is to remove the mother liquor left behind in the cake after filtration when this liquid is regarded as contaminant or must be recovered as a valuable component. Wash liquid flows through the filter cake and displaces the mother liquor, a mixture of both liquids leaving the cake as a wash filtrate. The concentration in this filtrate may help to explain some basic facts. In Figure 12 this concentration is reported as a function of the filtrate volume V. In the beginning it is $c_0 = 1$ (concentration of contaminant in the mother liquor). Four characteristic curves are reported:

Curve 1 represents the idealized piston flow. The pore volume V_{pore} is ideally displaced and leaves the cake with the concentration $c_0 = 1$. In this idealized case the washing would be finished with a quantity of wash liquid V_{w} equal the pore volume or a

$$\text{“wash ratio''} = \frac{V_{\mathrm{w}}}{V_{\mathrm{pore}}} = 1 \tag{43}$$

Curve 2 takes into account that even in an idealized cake the flow passes through streamlines with different lengths and speeds. Because of this “axial dispersion” the required wash volume is several times the pore volume or

$$\text{“wash ratio''} = \frac{V_{\mathrm{w}}}{V_{\mathrm{pore}}} > 1 \tag{44}$$

Curve 3 is still more realistic; taking into account that a filter cake contains stagnant zones or dead-end pores, that are not reached by the flow. They deliver their contaminant content by diffusion to the effluent. Diffusion is an asymptotic process and depends on the time elapsed, not on the quantity of liquid. This curve is typical for most washing processes. The first part depends on the “wash ratio”, the end however is an asymptotic process depending on time.

When the curves 1, 2 and 3 were drawn, it was supposed that the mass m of contaminant eliminated was contained in the pore volume V_{pore}. The integral below the curves is therefore the same for all three curves:

$$m = \int_0^\infty c \cdot dV = c_0 \cdot V_{\mathrm{pore}} \tag{45}$$

Curve 4 finally depicts an example where the washing resembles an extraction (→ Liquid – Solid Extraction) because the contaminant is not contained in the pores only, but also in the solid (liquid inclusions or soluble solid matter). The integral under curve 4 is therefore not related to the pore volume V_{pore} but to a higher total quantity of contaminant.

Theoretically such curves could be used to determine the residual mother liquor content in the cake from analyses of the filtrate. In practice this is impossible since the required residual contents in the cake are much smaller than the precision of such a mass balance. Nevertheless analyses of the filtrate are often used to judge the progress of washing. This is possible if an empirical correlation between the two concentrations in the filtrate and in the cake exists for the particular process. The validity of such empirical correlation is, however, restricted to the particular equipment and operating parameters.

3.2. Example of Experimental Results

An illustrative example of a washing process is given in Figure 13 A and B [16]. A kaolinite suspension contaminated with NaCl was filtered to varying cake thickness and washed with varying quantities of water. After every test the residual salt content in the cake was measured and the results are reported as a function of washing time

Table 1. Typical parameters of cross-flow filtrations

	Microfiltration	Ultrafiltration	Nanofiltration	Reverse osmosis
Cut-off size	> 100 nm	10 – 100 nm $10^3 - 10^5$ g/mol	> 1 nm $200 - 10^3$ g/mol	$< 10^3$ g/mol
Transmembrane pressure	0.02 – 0.5 MPa	0.2 – 1 MPa	0.5 – 3 MPa	2 – 20 MPa
Permeate flow	$50 - 1000\,\mathrm{L\,m^{-2}\,h^{-1}}$	$< 100\,\mathrm{L\,m^{-2}\,h^{-1}}$	$< 100\,\mathrm{L\,m^{-2}\,h^{-1}}$	$10 - 35\,\mathrm{L\,m^{-2}\,h^{-1}}$
Cross flow speed	2 – 6 m/s	1 – 6 m/s	1 – 2 m/s	< 2 m/s
Important mechanism of retention	screening by the pores of the membrane	screening by the membrane and the gel layer	electrostatic repulsion and screening	solubility and diffusion in the membrane
Important mechanism of transport	hydrodynamic lift force	back-diffusion	back-diffusion	back-diffusion

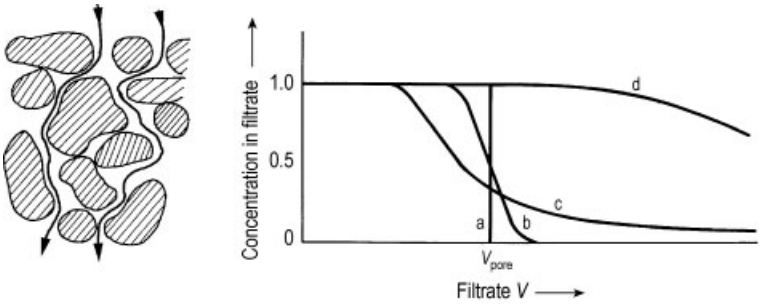

Figure 12. Concentration of contaminant in the effluent of a filter cake during washing
a) Idealized piston flow; b) Effect of axial dispersion; c) Effect of diffusion from dead-end pores; d) Effect of contaminant extracted from the solid

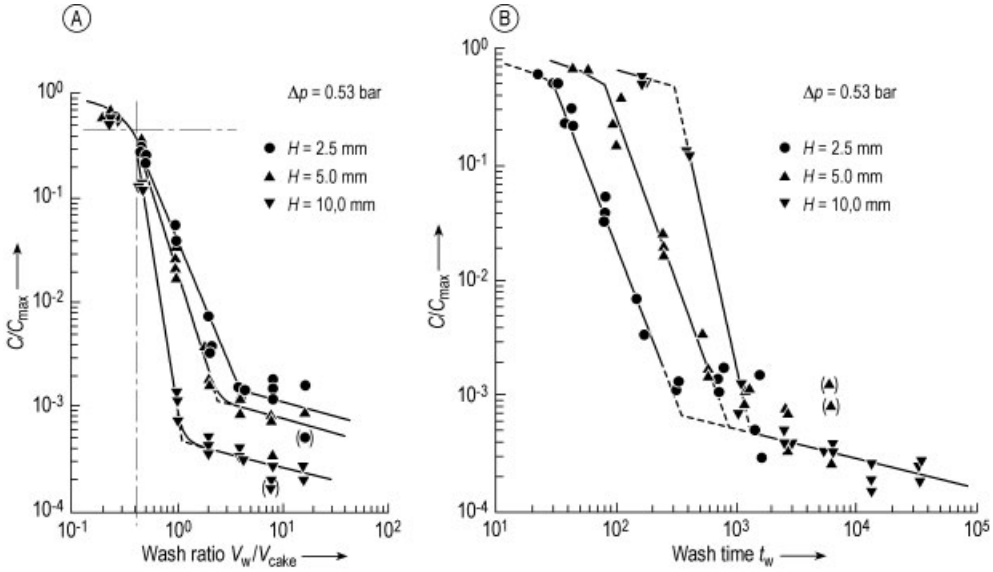

Figure 13. Residual concentration of salt in filter cakes of kaolin. The same test results are reported either as a function of the quantity of wash water (A) or as a function of washing time (B) [16]

and of wash ratio. (In contrast to Fig. 12 the wash ratio is related to the solids mass, not the pore volume. Also the residual content of the cake is reported, not the concentration in the filtrate.)

Figure 13 A shows how at the very beginning of the washing process the residual concentration depends on the wash ratio, disregarding the cake thickness. In this example this is true up to the dash-dotted line which corresponds to less than a one-fold displacement of the pore volume. As the wash ratio increases to several times the pore volume, the mechanism of axial dispersion becomes more important, and the effect of a given quantity of wash liquid becomes better for thick cakes than for thin ones.

Figure 13 B shows how at the end of an experiment the washing time is the most relevant parameter for the result. In this example this seems to be true below a residual concentration of

$$\frac{c}{c_0} \approx 10^{-3}$$

3.3. Test Procedures and Pitfalls

As can be seen in Figure 13, washing of filter cakes gives scattering results even under laboratory conditions. This explains why only few experimental results are found in the literature: many experimentators are discouraged by inconclusive test results and never publish them. The reason of scattering results are stochastic influences:

Fingering. In most washing processes the wash liquid has a lower viscosity than the mother liquor. It tends to flow through the cake in finger-like streams past isles of viscous fluid, as shown in Figure 14 [16]. This is a stochastic process, and the local concentration in the cake and the momentary concentration in the effluent will vary stochastically. Scale-up from a laboratory test filter with only small-scale fingering can be misleading as well as the results of few small samples taken from a big cake.

Instabilities, Shrinking. If small electrodes are placed in a filter cake during the washing process, they indicate wildly varying local conductivity. This reveals the existence of small cracks in changing size and positions even in cakes with no visible fissures or shrinkage. This is not surprising because the contact forces between the particles change dramatically when the ion content of the surrounding liquid is washed off (see Section 6.2). The phenomenon has not yet been investigated thoroughly, but its stochastic effects on the washing results should be similar to the above-mentioned "fingering".

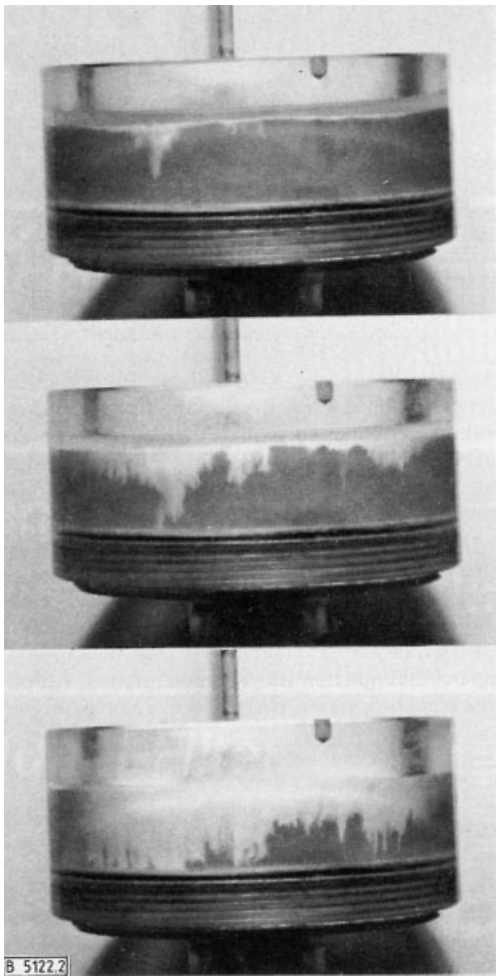

Figure 14. Visualization of "fingering" flow in a transparent laboratory nutsche filter. A wash liquid of lower viscosity displaces dark mother liquor [16]

If shrinking during washing presents a serious problem, it can be helpful to reslurry the cake in wash liquid and filter it again. Many nutsche filters (see Section 8.8) are used this way. The second cake (with less ions in the liquid during cake build-up) is often less porous and less prone

to shrinking (but also less permeable) than the original one.

In order to get meaningful results from washing experiments in spite of all these pitfalls, the following procedure is useful (see Fig. 15). Wash liquid flows at constant pressure through the cake. The quantity of filtrate and its composition are recorded automatically. Often the pH and the conductivity Q of the filtrate are used as indicators for concentration; sometimes its color may be more informative. The washing is continued until the filtrate reaches a pre-set criterion for purity (chosen according to some preliminary experience so that cake purity is near to the required specification). At this point the flow of wash water is automatically stopped. This should be done automatically, since it will probably happen during the night. The next morning, the cake is deliquored by air blowing and its residual content is analyzed. In small-scale experiments the whole cake should be analyzed, in large-scale experiments several samples must be taken from different parts of the cake.

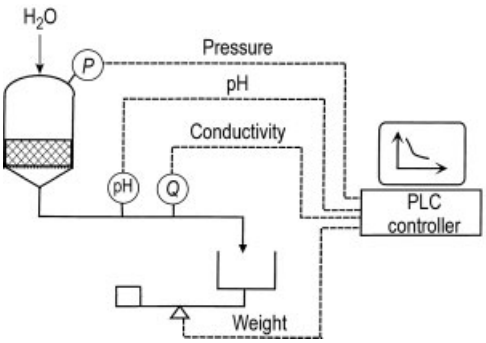

Figure 15. Test arrangement for washing test

The following results must then be evaluated:

- The required washing time,
- The required quantity of wash liquid, and
- The resulting purity of the cake.

Each of these criteria taken alone will scatter stochastically, but all three together will often allow a meaningful interpretation.

3.4. "Intermediate" Deliquoring before Cake Washing

Normally it is advantageous to deliquor the cake partly before washing, thus reducing the amount of mother liquor to be washed out. Sometimes however this "intermediate" deliquoring has detrimental effects:

If deliquoring is done by gas pressure, cracks may appear which let pass the washing liquid (see Section 4.1.5). In this case any deliquoring before the washing has to be avoided carefully. On belt filters (see Section 8.2) an overlap of filtering and washing zones is therefore often accepted, even if mother liquid and wash liquid get mixed. Some nutsche filters (Section 8.8) have even been equipped with optical sensors to avoid premature deliquoring [17]. They measure the light reflected by the liquid surface and detect when the cake surface runs dry, so that the gas pressure can be released before cracks appear.

If deliquoring is done by compression (see Section 4.2), the cake may get rather impermeable. Thus, according to a rule-of-thumb the applied pressure should be raised monotonously during the sequence of filtering – intermediate deliquoring – washing – final deliquoring, i.e., the squeezing pressure for intermediate deliquoring should be lower than the pressure for washing.

4. Deliquoring of Filter Cakes

4.1. Deliquoring by Gas Pressure

The moisture in the pores of a filter cake can be displaced by gas flowing through the cake under pressure. The residual saturation with moisture as a function of time then asymptotically approaches a final equilibrium value as represented in Figure 16. Thus, an equilibrium is established with regard to mechanical displacement of the liquid. Thermal drying by airflow will of course further reduce the moisture down to total dryness if the air flows long enough. But generally thermal effects are small as compared with mechanical effects because of the comparably small mass of air flowing through a filter cake [18].

The final "equilibrium" moisture for mechanical dewatering depends on the nature of the cake and the applied pressure difference. The initial deliquoring speed however depends in addition on the cake thickness. For both values the flow rate of gas is apparently without importance, if

it is not to produce the pressure difference. In the next sections first the equilibrium conditions will be examined.

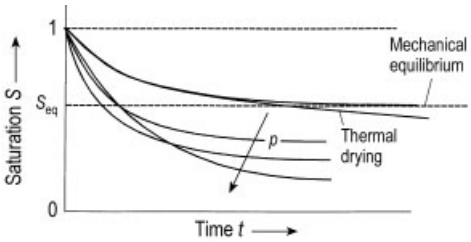

Figure 16. Deliquoring by gas pressure – residual saturation as a function of time

4.1.1. Equilibrium Saturation of Filter Cakes

A plot of saturation of the cake versus pressure difference in an equilibrium state is called a *capillary pressure curve*. It can be measured in a device according to Figure 17 [19]. The saturated filter cake (b) is placed in the filter (a) on a semipermeable membrane (c) (the membrane is permeable for liquid, but it is impermeable for gas at the applied pressure). Gas pressure is applied via port (e). The quantity of liquid displaced from the cake is drained by the valve (d) and collected. (The standpipe is used to obtain an exact liquid level when the pressure gauge is adjusted to zero).

A typical capillary pressure curve is shown in Figure 18 (incompressible cake of glass beads with a mean diameter of 79 µm and water). Starting from the completely saturated cake ($S = 1$) the gas pressure is slowly increased, at each set point the equilibrium is established and the quantity of displaced liquid is registered, yielding a point of the deliquoring curve. Liquid is displaced from the pores when the pressure exceeds a certain threshold pressure. This is the capillary entry pressure p_{ce} (in this example 0.062 bar). Each time when a pore is emptied, the gas pressure overcomes the capillary pressure at the neck of the pores with a diameter $d_{\mathrm{pore,neck}}$:

$$p_{\mathrm{ce}} = \frac{4 \cdot \sigma \cdot \cos\delta}{d_{\mathrm{pore,neck}}} \tag{46}$$

Further increase of the pressure displaces liquid from finer pores and reduces the moisture down to an irreducible saturation S_∞ after infinite time beyond which no further reduction is reached. This corresponds to the amount of liquid, which is trapped in isolated domains within the cake.

Figure 18 also shows the *imbibition curve*. When the pressure is reduced again, and provided the bottom of the cake is still in contact with the expelled liquid, then the cake will be reimbibed by capillary suction. At equal moisture content the suction pressure will be smaller than the capillary pressure difference for deliquoring. This is easily explained because the capillary pressure for imbibition depends on the larger "waist"-diameter of the pores (see Fig. 18):

$$p_{\mathrm{ci}} = \frac{4 \cdot \sigma \cdot \cos\delta}{d_{\mathrm{pore,waist}}} \tag{47}$$

Also imbibition does not go to full saturation, because some air remains trapped in the pores.

The following conclusions can be drawn from the capillary pressure curve:

1) Deliquoring is caused by pressure difference between gas and liquid. Gas flow is normally required to maintain this pressure difference, but not for the deliquoring itself. No gas flow is necessary if the cake is placed on a semipermeable membrane, which is impermeable for gas. (This principle can also be used technically for dewatering without gas flow [20]).
2) A certain threshold pressure must be exceeded before liquid is displaced from the pores; this is the capillary entry pressure p_{ce}. This implies technical consequences:
   Vacuum filters are restricted to pressure differences < 1 bar. This means that cakes cannot be deliquored by suction if their p_{ce} is near or above this limit. A pressure filter with $\Delta p > 1$ bar has to be considered for such cakes, even if its installation is much more expensive than that of a vacuum filter. Theoretically suction can dewater filter cakes with pores $>$ ca. 3 µm: with pure water and a small contact angle ($\delta \approx 0$) Equation (46) yields $p = 1$ bar for $d_{\mathrm{pore}} = 2{,}9$ µm.

Filter cakes of submicronic particles like wastewater sludge have such a high capillary entry pressure that their pores cannot be dewatered mechanically: with pure water and a small

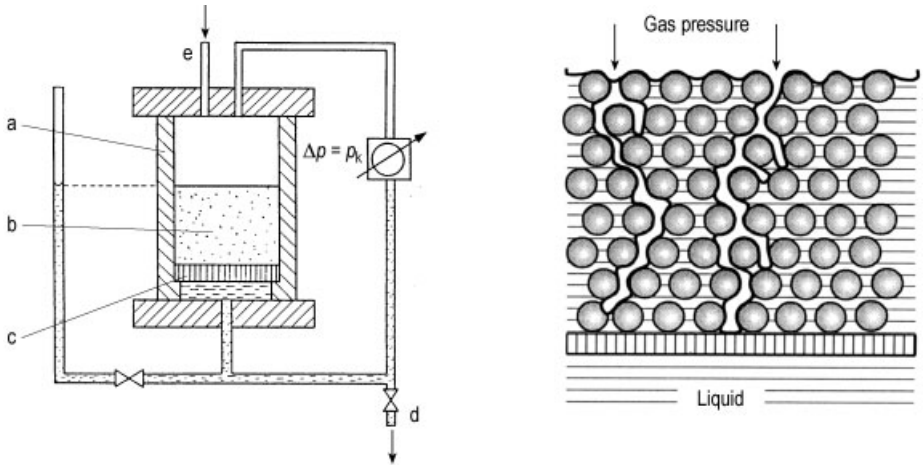

Figure 17. Test installation for measuring the capillary pressure p_k [19]
a) Filter vessel; b) Filter cake; c) Semipermeable membrane; d) Draine; e) Entry for gas

contact angle Equation (46) yields $p > 10$ bar for $d_{\text{pore}} < 0.29\,\mu\text{m}$.

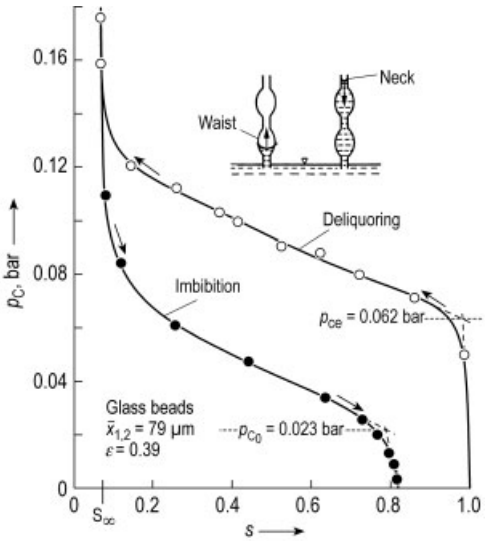

Figure 18. Example of a capillary pressure curve for imbibition and for deliquoring [19]

4.1.2. Kinetics of Deliquoring by Gas Pressure

In most cases there is no semipermeable membrane below the cake but a "normal" filter medium that is permeable for gas, like a filter cloth. Gas and liquid flow simultaneously through the cake. The gas exerts capillary pressure onto the adjacent liquid and both gas and liquid flow in the same direction, however with different speeds and different pressures (the local pressure difference between gas and liquid is equal to the local capillary pressure). The difference in saturation between bottom and top of an incompressible filter cake is normally rather small, see Figure 19 left. This is in contrast to the deliquoring by gravity (or by a centrifugal field), shown in Figure 19, right. In this latter case always a saturated bottom layer exists, corresponding to the capillary suction height of the cake. Even in equilibrium this layer remains saturated. Sometimes this height is negligibly small. In scraper-type centrifuges (→ Centrifuges, Filtering) for example, the saturated bottom layer must in any case be smaller than the residual layer which the scraper does not remove.

The permeability of a completely saturated cake is the same if either liquid or gas flows through the pores. When, however, gas and liquid flow simultaneously through the cake, the pores filled with liquid and the pores filled with gas form two separate capillary systems, each of them with reduced permeability. The local permeabilities (for liquid or air) in relation to the permeability of the saturated cake are called "relative permeabilities" k_{rel}. They depend on the local saturation of the cake and can be represented as:
for the wetting fluid (the liquid)

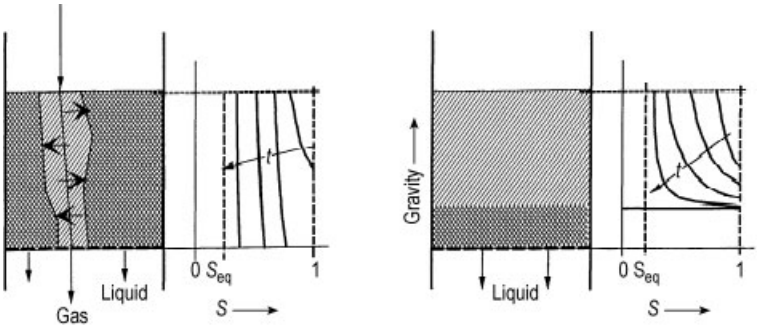

Figure 19. Kinetics of deliquoring by gas blowing (left) and by gravity or a centrifugal field (right)

$$k_{\mathrm{rel,w}} = \frac{k_{\mathrm{w}}}{k} = \frac{\alpha}{\alpha_{\mathrm{w}}} \tag{48}$$

for the non-wetting fluid (the gas)

$$k_{\mathrm{rel,n}} = \frac{k_{\mathrm{n}}}{k} = \frac{\alpha}{\alpha_{\mathrm{n}}} \tag{49}$$

Such relative permeabilities are shown schematically in Figure 20 as a function of cake saturation [21]. The permeability for the liquid (the wetting phase) declines to zero at a finite saturation. This value corresponds to the irreducible saturation S_∞. At the other extreme the flow of air ceases when the saturation is above 90 %. Some of the voids are filled with air without allowing an appreciable airflow.

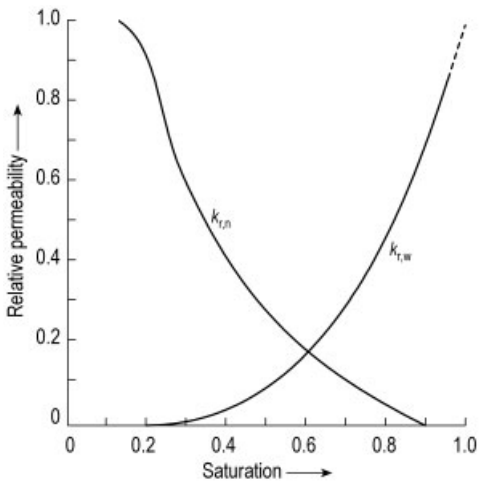

Figure 20. Relative permeability as a function of saturation [21]

If the curves for capillary pressure and for the relative permeabilities as a function of saturation are known, the local values of capillary pressure, gas permeability, and liquid permeability are defined as functions of the local saturation. The deliquoring kinetics by gas flow can then be described mathematically. The resulting system of interdependent differential equations is however extremely complex and not suited for practical application. For scientific purposes approximate solutions have been verified by comparing them to experiments [19].

4.1.3. Approximate Solution for Coarse, Incompressible Cakes

For the particular case of incompressible cakes with threshold pressures much smaller than the applied gas pressure $p_{\mathrm{ce}} \ll p$ (hence for rather coarse solid particles), WAKEMAN has described the process of deliquoring by gas pressure by three dimensionless parameters [22]:
reduced saturation

$$S_{\mathrm{R}} = \frac{S}{1 - S_\infty} \tag{50}$$

dimensionless time

$$\Theta = \frac{p_{\mathrm{ce}} \cdot t}{\alpha_{\mathrm{H}} \cdot \eta \cdot H^2 \cdot (1 - S_\infty) \cdot \varepsilon} \tag{51}$$

dimensionless pressure difference

$$\Delta P = \frac{p_{\mathrm{entry}}}{p_{\mathrm{ce}}} - \frac{p_{\mathrm{exit}}}{p_{\mathrm{ce}}} \tag{52}$$

According to this approach, the cake is characterized by its porosity e, the filter resistance α_{H} (of the saturated cake) and the capillary pressure curve which itself is approximately described by

the capillary entry pressure p_{ce} and the "irreducible" saturation S_∞ (see Fig. 18). The relative permeabilities are described by generalized interpolation formula. With these approximations the residual saturation and the theoretical gas flow rate can be read from the dimensionless charts, shown in Figures 21 and 22. An example how to apply these charts is given in [23]. However, the validity of these charts is limited to $p_{ce} \ll p$ and to incompressible cakes. Also the gas flow rate indicated on the chart does not take into account the possibility of cracks in the cake which can increase the gas flow tremendously (see Section 4.1.5).

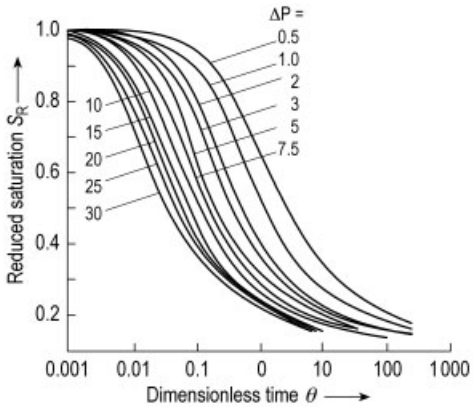

Figure 21. Reduced cake saturation vs. dimensionless time [23]

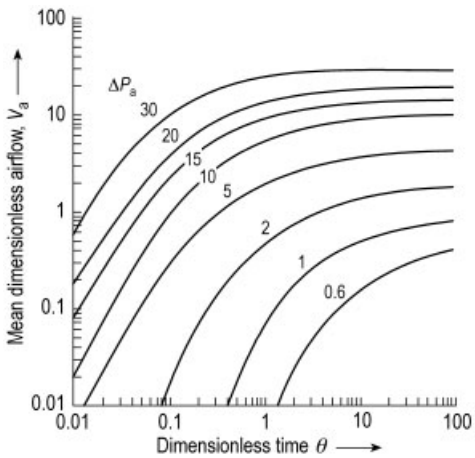

Figure 22. Reduced air (gas) flow rate vs. dimensionless time (WAKEMAN and PURCHAS 1986; reproduced from [23])

4.1.4. Practical Scale-Up of Deliquoring by Gas Pressure

The deliquoring kinetics are normally investigated by measuring the residual moisture as a function of cake thickness, blowing time, and pressure difference. The results are interpolated to give the parameters necessary for scale-up. Extrapolation beyond the investigated range of cake thickness, blowing time, and pressure should be avoided. In addition it is important to measure and to scale-up the gas flow from realistic experiments, since in most cases the capacity of the compressor is the limiting factor, not the pressure. Particular attention has to be paid to the possible formation of cracks. In case of doubt, a filter test with sufficiently large filter area is recommended, because the cracks may not be pronounced and obvious on a small laboratory filter. Typically the cracks appear at a characteristic liquid saturation which then represents the limit for deliquoring by gas blowing with technically reasonable gas flow rates (see Section 4.1.5).

Interpolation of the test results with a minimum of tests is made easier if the residual moisture is reported as a function of a dimensionless time Θ. The results from a given kind of filter cake are then supposed to fit approximately a straight line in a logarithmic chart. As a first approach, when nothing is known about the deliquoring behavior of the cake, the experimental values are reported as a function of

$$\Theta_0 = \frac{\Delta p \cdot t}{\eta \cdot H \cdot (\alpha \cdot H + \beta)} \approx \text{const} \cdot \frac{\Delta p \cdot t}{H^2} \tag{53}$$

Here $\Delta p = p_{\text{entry}} - p_{\text{exit}}$ is the pressure difference applied. This approach gives generally good results for centrifugal dewatering, as shown in Figure 23 (the pressure difference being calculated from the centrifugal force) [24]. For dewatering by gas blowing, the fit is however often unsatisfactory (see Fig. 24) [24]. A dimensionless time taking into account the capillary entry pressure p_{ce} will always permit a good fit. Two equations are proposed in the literature [24]:

$$\Theta_1 \approx \frac{\Delta p \cdot t}{\eta \cdot H \cdot (\alpha \cdot H + \beta)} \cdot \frac{p_{\text{entry}}}{p_{\text{exit}}} \cdot \left(1 - \frac{p_{ce}}{\Delta p}\right)$$
$$\approx \text{const} \cdot \frac{\Delta p \cdot t}{H^2} \cdot \frac{p_{\text{entry}}}{p_{\text{exit}}} \cdot \left(1 - \frac{p_{ce}}{\Delta p}\right) \tag{54}$$

and [25]

$$\Theta_1 \approx \frac{\Delta p \cdot t}{\alpha \cdot \eta \cdot H^2} \cdot \frac{2}{\varepsilon} \cdot \left(1 - \frac{p_{\mathrm{ce}}}{\Delta p}\right)$$
$$\approx \mathrm{const} \cdot \frac{\Delta p \cdot t}{H^2} \left(1 - \frac{p_{\mathrm{ce}}}{\Delta p}\right) \tag{55}$$

Here p_{entry} and p_{exit} are the pressures applied above and below the cake and p_{ce} is the capillary entry pressure. This p_{ce} is not measured, but derived as a "best fit" to the experiments (this is why the fit is generally rather good, see Fig. 25 [24]).

4.1.5. Shrinking and Cracks in Filter Cakes

Deliquoring of compressible filter cakes by gas pressure very often produces cracks in the cake, a problem of great practical impact. The mechanism of crack formation is represented in Figure 26. During filtration the viscous forces from the flowing liquid compress the cake. As described in Section 2.1.4, only the layer near the filter medium is subjected to the full amount of pressure, while the "upper" layer of the cake remains uncompressed and porous. When filtration is finished, the liquid surface reaches the cake surface. Before gas enters into the capillaries of the cake, the capillary entry pressure p_{ce} must be overcome. The surface tension of the liquid thus exerts a pressure $p \leq p_{\mathrm{ce}}$ onto the cake, comparable to an elastic membrane spread over the cake. As long as p_{ce} is not exceeded, the cake is compressed monoaxially like by action of a piston, the cake thickness is reduced, but no cracks appear. However, when p_{ce} is exceeded, gas penetrates into the largest capillaries, and the solid structure is compressed by lateral forces. This may lead to lateral compression of the cake and hence to cracks. The criterion for cracks to occur thus probably depends on the lateral compressive strength of a cake after monoaxial consolidation. But until now the precise mechanism of crack formation is not understood and the great variety of crack patterns shown in Figure 27 [26] cannot be explained.

Cracks in a filter cake are very detrimental to further deliquoring and also to subsequent washing (see Section 3.4) because gas and wash liquid flow through the cracks without great resistance and without effect. The only sure remedy against cracks is compression of the cake before gas blowing. When the remaining "shrinking potential" (this is the remaining shrink when the cake is reduced to total dryness) is small enough, no cracks will appear anymore [27]. Some filters are equipped for this purpose with membranes or with pressure belts (see Sections 8.6 and 8.10). According to a proposal of Shirato and coworkers this compression can also be achieved by adding a layer of very fine material (hence a cake layer with high p_{ce}) onto the surface of the cake [28].

Some remedy against cracks is also possible with nutsches equipped with agitators (see Section 8.8). With their paddles they smear over the surface of the cake during deliquoring so that the cracks are closed as soon as they appear. This does, however, not eliminate the cracks totally but only down to a certain depth below the surface of the cake, so that some cracks can exist invisibly below the surface. Without such equipment the formation of cracks can be reduced (but only to a very limited degree) if the cake is formed with high final filtration pressure, and hence with somewhat less porosity.

4.2. Deliquoring by Expression

When soft filter cakes are compressed, the void volume of the cake is reduced by expelling liquid, while the voids remain saturated with liquid. Such compression often is applied in combination with a subsequent deliquoring by gas pressure; it reduces the subsequent gas flow and prevents the formation of cracks (see Section 4.1.5). But there are many applications where deliquoring by gas pressure is not possible, and expression is the only way to reduce the liquid content by mechanical means. Classical examples for applications of mechanical forces are juice and oil pressing from fruits and also the dewatering of wastewater sludge in membrane filter presses (see Section 8.10) or in pressure belt filters (see Section 8.2). Such sludges form filter cakes with very high capillary entry pressure p_{ce} so that deliquoring by gas blowing is technically not feasible.

Mathematical description of expression generally starts from modifications of the Terzaghi model for soil mechanics. The ratio of liquid collected to the amount of liquid, which can be expressed, is [29]

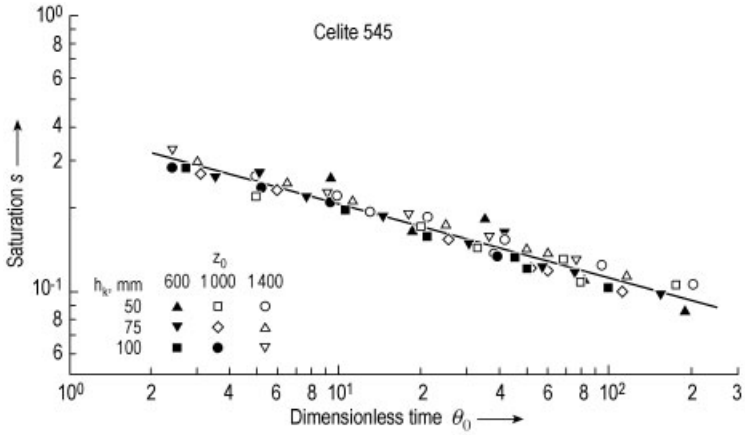

Figure 23. Residual saturation of filter cakes dewatered by centrifugal force as a function of dimensionless time Θ_0 [24]

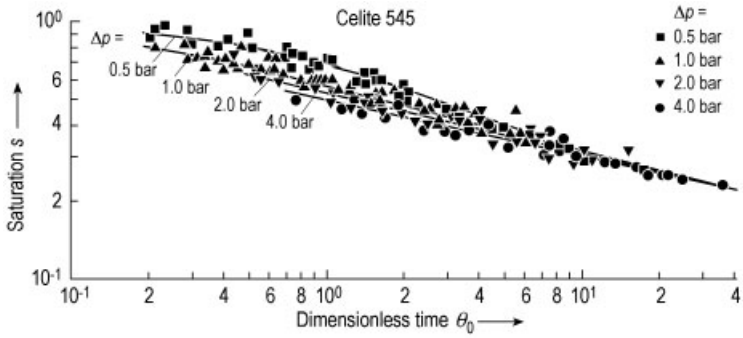

Figure 24. Residual saturation of filter cakes dewatered by gas blowing as a function of dimensionless time Θ_0 [24]

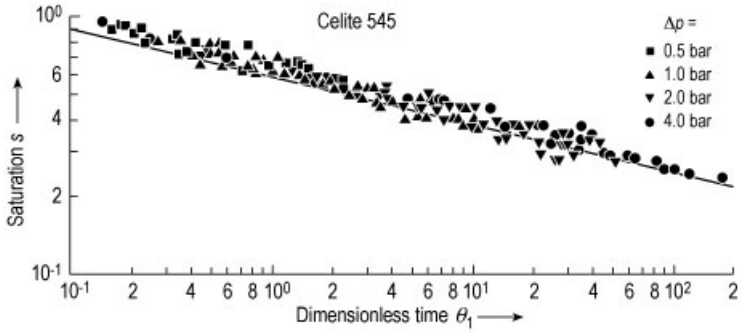

Figure 25. Residual saturation of filter cakes dewatered by centrifugal force as a function of the modified dimensionless time Θ_1 [24]

$$U_c = \frac{H_1 - H(t)}{H_1 - H_\infty} = 1 - B \cdot \exp(-C \cdot t) \qquad (56)$$

where U_c is the consolidation rate, $H(t)$ the cake thickness, H_1 the original cake thickness and H_∞ the thickness after infinite time. B and C are "creep" constants and t_c is the consolidation (=compression) time. A comparison of different compression models can be found in [30].

5. Optimal Cycle Time

When sizing a filter installation, the right cycle time has to be chosen (example: the same task

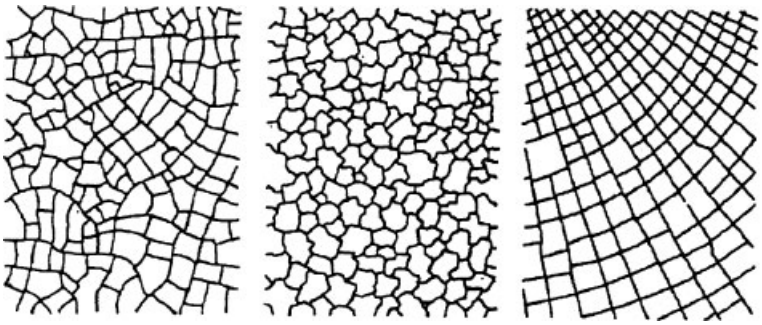

Figure 26. Formation of cracks by gas blowing through a compressible cake

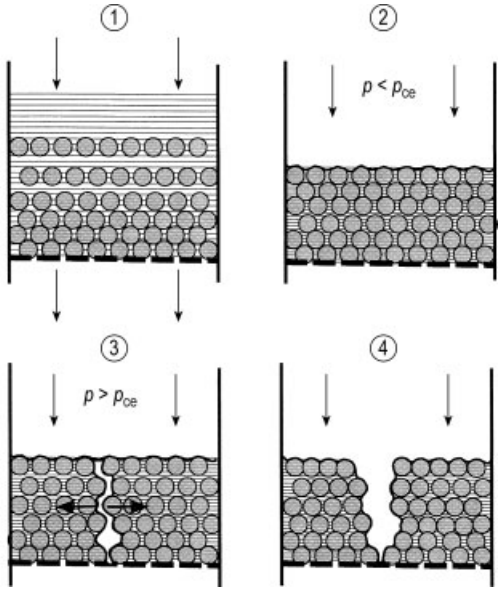

Figure 27. Patterns of shrinking cracks [26]

can be performed in a small filter which must be cleaned frequently or a bigger one with longer cycles). With a long filtration time the filter cake becomes thick and the flow rate per filter area declines. On the other hand for very short filtration the downtime for frequent cleaning will reduce the capacity. This leads to two questions: (1) What is the optimal cycle time with the highest overall throughput? (2) What is the optimal cycle time with the lowest overall cost?

1) Figure 28 shows the quantity of filtrate produced as a function of time. Before the start of filtration a time t_{reg} is needed for regenerating the filter (extracting the cake from the previous cycle, cleaning the filter, rearranging the filter elements, etc.). The highest overall throughput is obtained when the straight line from the start of this regeneration to the end of filtration is tangential to the yield curve. In other words: the filtration should be stopped when the instantaneous flow (as indicated by a flow meter) falls below the mean throughput according to the definition:

$$\dot{V}_{\mathrm{mean}} = \frac{V}{t + t_{\mathrm{reg}}} \tag{57}$$

This criterion can for example be implemented in a computerized process control. For design purposes it can be useful to calculate the optimal cycle time in advance. For this purpose a rule-of-thumb will be derived

assuming cake filtration and negligible resistance of the filter medium ($\beta = 0$). The quantity of filtrate according to Equations (14) or (15) is then

$$V = c_1 \cdot t^{1/2} \tag{58}$$

and Equation (57) becomes

$$\dot{V}_{\mathrm{mean}} = \frac{c_1 \cdot t^{1/2}}{t + t_{\mathrm{reg}}} \tag{59}$$

The highest $\dot{V}_{\mathrm{mean}}$ is obtained if the filtration is stopped at t_{opt}, which is defined by the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{\dot{V}_{\mathrm{mean}}}\right) = 0 = \frac{1}{2c_1} \cdot t_{\mathrm{opt}}^{-1/2} - \frac{t_{\mathrm{reg}}}{2c_1} \cdot t_{\mathrm{opt}}^{-3/2} \tag{60}$$

This leads to the rule-of-thumb for maximum throughput:

$$t_{\mathrm{opt},1} = t_{\mathrm{reg}} \tag{61}$$

2) Quite analogous considerations are valid for the optimal cycle time in terms of cost per quantity filtered. During filtration, the running cost is proportional to time (capital cost and pumping energy), cost = $c_2 t$. In Figure 28 all costs are converted with this proportionality factor c_2 into an equivalent running time: $t = \mathrm{cost}/c_2$.
Regeneration represents a considerable cost, including downtime, labor, new filter elements, and disposal of waste. If this cost is called c_{reg}, the mean specific cost is

$$\frac{\mathrm{cost}}{\mathrm{filtrate}} = \frac{c_2 \cdot t + c_{\mathrm{reg}}}{V} \tag{62}$$

With the approximation of Equation (58) this is

$$\frac{\mathrm{cost}}{\mathrm{filtrate}} = \frac{c_2}{c_1} \cdot t_{\mathrm{opt},2}^{1/2} + \frac{c_{\mathrm{reg}}}{c_1} \cdot t_{\mathrm{opt},2}^{-1/2} \tag{63}$$

The minimum is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{cost}}{\mathrm{filtrate}}\right) = 0 = \frac{c_2}{2c_1} \cdot t_{\mathrm{opt},2}^{-1/2} - \frac{c_{\mathrm{reg}}}{2c_1} \cdot t_{\mathrm{opt},2}^{-3/2} \tag{64}$$

As result the rule-of-thumb for maximal yield per cost is obtained:

$$c_2 \cdot t_{\mathrm{opt},2} = c_{\mathrm{reg}} \tag{65}$$

or "cost for the filter run = cost for regenerating the filter".

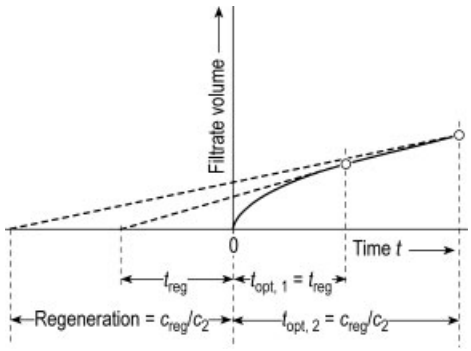

Figure 28. Optimal cycle time derived from the yield curve. $t_{\mathrm{opt},1}$ for maximal flow per time and $t_{\mathrm{opt},2}$ for maximal yield per cost

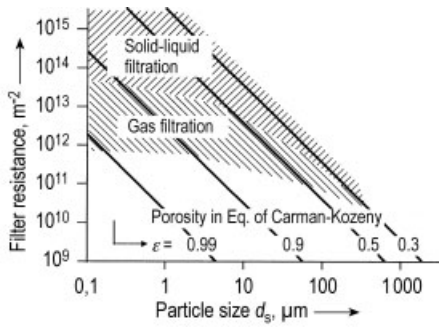

Figure 29. Filter resistance vs. particle size
The range of usual values is compared to the calculated lines according to Carman – Kozeny, Equation (66)

These rules have been derived here for constant pressure filtration. They can also be derived for filter runs with different but constant flow rates until a pre-set final pressure. Practitioners therefore apply the rule-of-thumb rather generally (e.g., also for filtration with a pump with a characteristic curve anywhere between constant pressure and constant flow).

6. Understanding the Filter Resistance

6.1. The Equation of Carman and Kozeny

The resistance α of a filter cake depends on the size and number of pores in the cake. In a first

approximation it can be related to the size of particles and the porosity of their arrangement according to the classical Kozeny equation [31]. (The theoretical basis of the Kozeny equation is subject to criticism, see for example [32], nevertheless the Kozeny equation is a very useful approximation. More recent correlations are given in [33].) For spherical particles this equation is:

$$\alpha_{\mathrm{H}} \approx 5 \cdot \frac{(1-\varepsilon)^2 \cdot 36}{\varepsilon^3 \cdot d_{\mathrm{S}}^2} \tag{66}$$

where ε is the porosity of the cake and d_{S} is the Sauter mean diameter of the particles, i.e., the diameter giving the same specific surface. The filter resistance resulting from this equation is depicted in Figure 29 as a function of particle size d_{S} and porosity ε. In practical applications the filter resistances cover a very wide range of porosity. Fine particles, especially dry dust, often form cakes with surprisingly high porosities (see Fig. 30). The same quantity of powder (filter aid, mean particle size $\approx 5\,\mu\mathrm{m}$) has settled in water with different pH and in air. The settling volume (and the clarity of the supernatant) is quite different. Filter resistance depends therefore to a large degree on cake porosity and hence on the surface forces producing this porosity!

6.2. Interparticle Forces, DLVO Theory

According to the theory of Derjaguin, Landau, Verwey, and Overbeek (= DLVO theory after the researchers' initials), the forces between suspended particles of equal material are described by repulsion due to electrostatic charges and attraction due to van der Waals' forces:

- Particles in suspension (especially in polar liquids like water) have an electrically charged surface, due to adsorption and dissolution of ions. The order of magnitude of this charge ζ is $-120\,\mathrm{mV} < \zeta < 120\,\mathrm{mV}$, and it depends on the pH of the liquid. The charge reaches relatively far, depending on the ion content of the surrounding liquid. The reach is characterized by the Debye length r_{D}. In pure distilled water this length is exceptionally long with $0.9\,\mu\mathrm{m}$. In most technical aqueous solutions it is much smaller, in sea water for example only 0.4 nm [34].
- Van der Waals forces are part of the cohesive forces between the molecules within the solid particle. They reach also a certain distance beyond the surface, but their range is less than that of electrostatic forces. The decline is proportional to the distance according to

$$E_{\mathrm{attraction}} = -\frac{\mathrm{const}}{r} \tag{67}$$

  When two particles approach to distance zero, these attractive forces become very strong and should in principle become equal to the cohesive force inside the particle. The reality is, however, more complicated because there are always absorbed molecules on the surface.

An example of the superposition of both forces according to DLVO theory is represented in Figure 31. The graph shows the resulting energy potential. A positive gradient of the curves describes attraction, negative gradient repulsion. At very small distances, attraction prevails. At larger distance electrostatic repulsion prevails (if it is not zero).

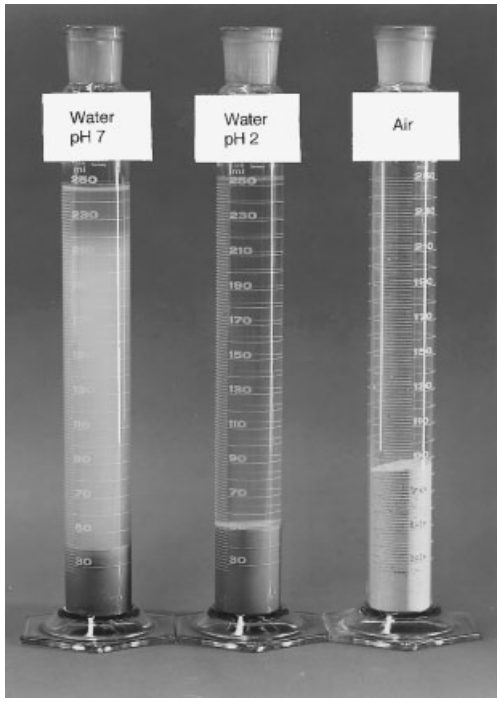

Figure 30. Settling volumina of the same quantity of filter aid (mean particle size ca. 5 µm); left: in water at pH 7; center: in water at pH 2; right: dry powder compacted by its own gravity [44]

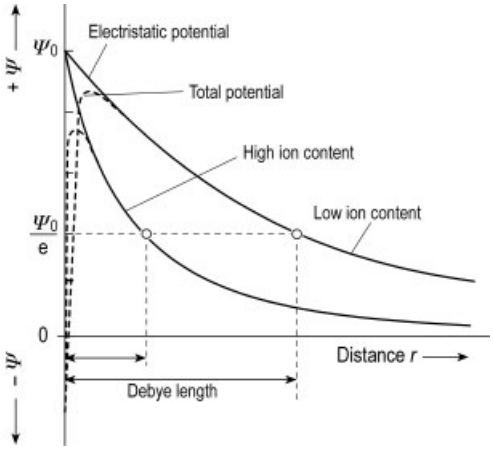

Figure 31. Energy potential vs. particle distance (DLVO-theory)
Attraction and repulsion between two particles in suspension. The superposition of electrostatic repulsion and attraction by van der Waals forces is represented by the resulting energy potential. High ion content in the surrounding liquid reduces the reach of electrostatic repulsion (Debye length).

Concerning interparticle contact this means:

- Electrostatically charged particles repulse one another. Seen from their equally charged neighbors they show some kind of smooth and slippery repulsive skin. (The smooth repulsive skin can be made visible with modern atomic force microscopy, see, e.g., [35].) The Debye length characterizes the thickness of this skin (see Fig. 32).
- Without electric charge (as for example in air), there is no such skin. The particles touch their neighbors with their rough surfaces and tend to adhere due to van der Waals' forces. This explains why the dry powder in Figure 30 has the lowest packing density.
- Particles in aqueous suspension with weak electric charge most probably have oppositely charged patches on their surface what may enhance mutual adhesion and the formation of loose flocs.

The surface charge of suspended particles can be measured as a "zeta potential" (→ Colloids, Chap. 4.; → Emulsions, Chap. 11.6.). This potential varies with the suspending liquid, especially with its pH: High pH means a high concentration of OH^- ions, which can be absorbed, creating a negative surface charge. Low pH means high concentration of protons and a positive surface charge. At a certain pH the surface is neutral (point of zero charge, isoelectric point). At this particular pH, the particles can more easily agglomerate and are easier to filter. This optimal pH is often determined empirically without knowledge of the zeta potential.

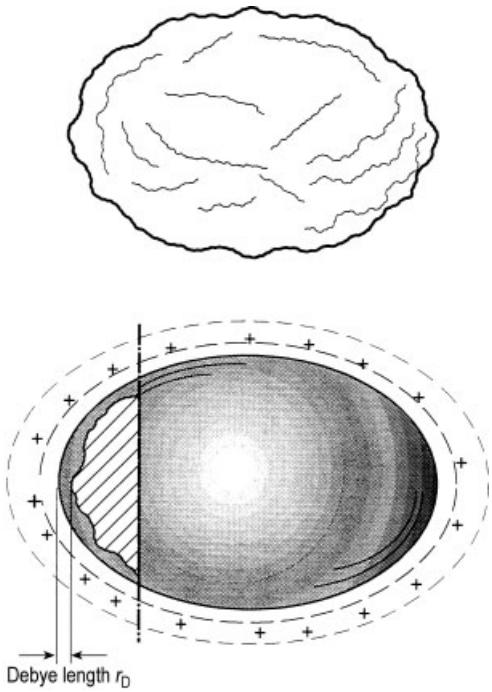

Figure 32. A particle in suspension
Seen from a particle with the same surface charge, it is covered with a repulsive skin of a thickness near the Debye length. This skin has no roughness and no friction; hence it is rather slippery.

6.3. Mathematical Simulation of Cake Formation

Different attempts have been made to simulate numerically the structure and porosity of filter cakes. Until now such calculations are restricted to spherical particles with homogeneous surface. Starting from randomly chosen locations the trajectories of particles in a filter flow are calculated (generally in two dimensions). When they touch the cake surface, the deposition is simulated using either a sticking angle [36], [37], or a friction angle and an adhesive force [38], [39], or by estimating the interparticle forces from van der Waals' and electrostatic forces according to the DLVO theory [40]. After calculating the deposition of a large number of particles, the packing

density of the resulting cake is found. These calculations help to understand and explain some empirical facts:

1) High adhesion forces and large friction angles make the particles adhere at the first contact and produce a randomly stacked "house of cards" with high porosity and permeability. Low adhesion forces or even repulsion between the particles make them slip to more stable positions and produce dense packing. Figure 33 shows as an example of the cake structure calculated for different friction angles and adhesive forces [39].
2) An opposite effect can be observed when the particles are very small (colloidal). Here the Debye length is often not negligible compared to the particle size and the repulsion can be strong enough to keep the particles at a distance. In this case the porosity and permeability will increase by repulsion and decrease by adhesion, contrary to what was said above. This is for example observed with clay soil in marshland near the sea. The permeability of such soil is reduced after flooding with salt water and increases again when the salt is washed out, i.e., the Debeye length increases with falling ion content. For industrially filterable suspensions however the effect is not yet reported in the literature. It seems that for particles > 0.1 µm either the Debye length is to small (at high ion content) or the zeta potential is not strong enough (as in distilled water) for this effect.

A comprehensive simulation of cake formation has not yet been proposed. The above mentioned approaches are restricted to spherical particles with uniform surface charge. Often, however, the electric charge is unevenly distributed, and there may even exist oppositely charged spots on the same particle. This would explain why many suspensions form loose flocs when the stirrer is stopped. This is quite often observed even if the pH is not exactly at the isoelectric point (the term "isoelectric range" would therefore be more appropriate than "isoelectric point"). The flocs disappear when the stirrer starts again, but nevertheless they probably have an impact on filterability because the particles are prearranged in a way that they touch with their rough, adhesive spots and have less tendency to slip into densely packed positions when deposited on the filter cake.

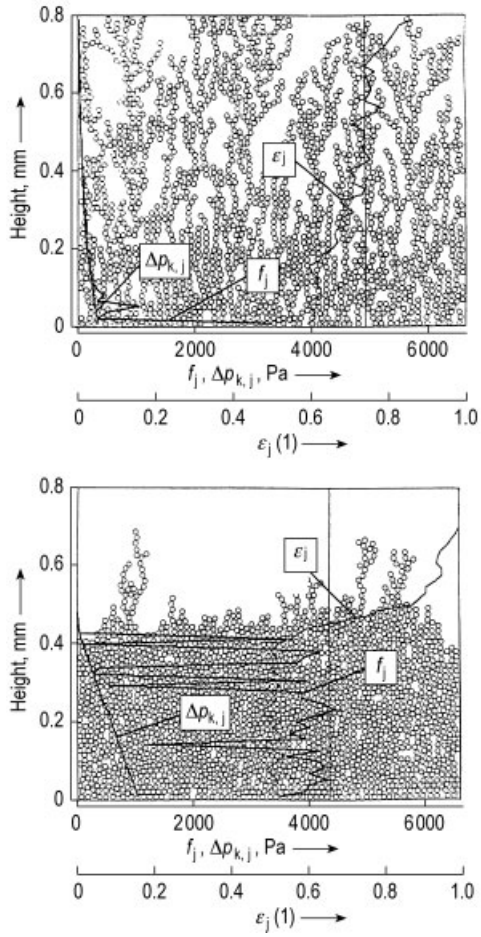

Figure 33. Simulation of the cake build-up for two different adhesion forces and friction angles (upper diagram for 4×10^{-8} N adhesion force and 20° friction angle, lower diagram for 1×10^{-8} N and 5°) [39]

7. Solving Filtration Problems in Small-Scale Tests

7.1. Laboratory Tests

The most common technique for laboratory filter tests uses a laboratory pressure filter according to Figure 34 [41]. It consists of a pressure

vessel about 16 cm long with a filtration area of 20 cm^2 (corresponding to a diameter of 50 mm). Suspension (250 – 300 cm^3) is charged into the filter, and the lid is closed. Filtration starts when pressurized gas (air or nitrogen) is applied from above. This should be done quickly (within a few seconds) to prevent sedimentation. The filtrate is collected in a vessel on a balance or in a graduated cylinder and its quantity is registered as a function of time. When the filtration is finished, the compressed gas is applied for about another minute to deliquor the cake. The quantity of gas applied during deliquoring is measured with a flow meter or it is estimated from the capacity of the feed line and its pressure drop on gas withdrawal from the feed. After filtration and deliquoring the filter is opened, the cake thickness is measured and the cake is removed to determine its wet and dry mass and calculate the moisture content. The quantity of filtrate obtained is interpreted in a linear diagram (see Section 2.1.3) and the corresponding filter resistance is calculated. Several measurements at different pressures (e.g., 0.5 + 1 + 6 bar) provide the compressibility coefficient "*n*" from Equation (31). For more details see [42], [43].

Another very simple test is the bench leaf test. The filter leaf includes a filter cloth fixed on a plate (100 cm^2) with grooves and sealing around the edges. It is connected to an evacuated filtrate receiver via a rubber hose with a valve. The filter leaf is plunged into the suspension and filtrate is sucked through the leaf by applying of vacuum. It requires, however, considerable skill to get well reproducible results with this arrangement, since the solids are sucked to the leaf against the effect of settling, and their composition depends on the way the leaf is plunged into the suspension. This is why the bench leaf test is used less frequently nowadays, only for sizing rotary vacuum filters, which operate in a quite similar way.

For research purposes it may be interesting to measure the filter resistance as a function of compressive pressure. To this behalf the Compression-Permeability Cell (CP cell) has been developed (see Fig. 35). This is a cylindrical cell with a porous bottom and a porous piston from above. Mechanical load is placed on the piston and the filter cake is compressed. Then liquid is percolated through the cake with moderate pressure and the permeability is calculated from pressure drop and flow rate according to Equation (1) or (3).

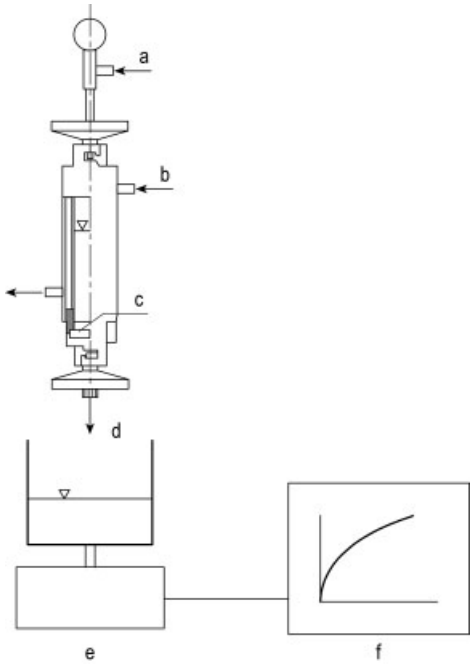

Figure 34. Bench pressure filter with balance and recorder for yield curve
a) Compressed air; b) Heating fluid; c) Filter medium; d) Filtrate; e) Balance; f) Recorder

7.2. Handling of "Unfilterable" Suspension

In industrial filtration often suspensions are encountered with a high filter resistance. Several possibilities for handling of these "unfilterable" suspensions are given below.

Optimization of Upstream Steps (Crystallization, Precipitation). Considerable improvements in filterability are achieved by producing coarser particles. Good knowledge is available about crystallization (→ Crystallization). If, however, the solids are formed by precipitation, i.e., by mixing two reactants, this quick process often depends strongly on the mixing conditions and the reaction speeds. In this case the filterability must be improved empirically. Improvement has been achieved with the following parameters:

1) Continuous mixing instead of batch mixing,

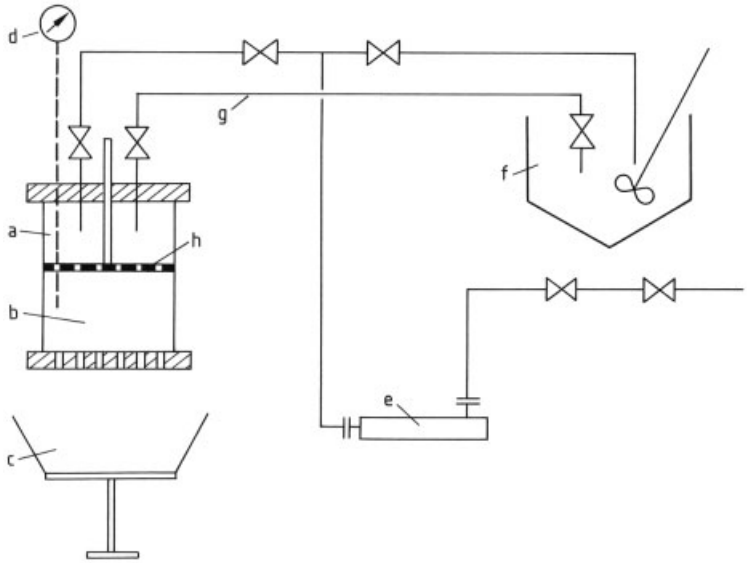

Figure 35. Compression – permeability cell
a) Filter cell; b) Filter cake; c) Filtrate receiver; d) Pressure recorder; e) Pump; f) Slurry tank; g) Recirculation; h) Movable plate

2) Recycling of suspended solids to the zone of solids formation,
3) Variation of temperature, mixing ratio, stirring speed etc.

Application of Flocculants (Polyelectrolytes). Synthetic, water-soluble polymers are highly effective as flocculating agents. They agglomerate fine particles in aqueous dispersion to voluminous flocs, which settle and filter easily. As a general rule, however, the resulting filter cake or sediment has a higher porosity and residual moisture than without flocculation.

Synthetic flocculants are commercially available as nonionic, anionic, or cationic grades with more or less ionic character. A charge opposite to the zeta potential of the solids to be separated gives theoretically the best effect. In practice it is, however, easier to observe the flocculation visually without knowledge of the surface charge. Samples of the suspension are prepared in graduated cylinders and gently mixed with different flocculants in 0.1 % solution. The high dilution is necessary to reduce the viscosity and facilitate mixing. The required quantity of polyelectrolyte is generally in the range of 0.5 – 10 kg per ton dry solids, equivalent to 10 – 100 mL of solution for 1 L suspension.

Criteria for the flocculation effect are settling speed, clarity of the supernatant, and aspects of the flocs (small, dense flocs are sometimes preferred, because they are more stable).

Adaptation of pH. As explained in Section 6.2, the interparticle forces depend on the pH of the surrounding liquid. With a pH in the isoelectric range the particles can aggregate and their filter resistance and settling speed is increased.

Checking of Alternatives to Cake Filtration. Possible alternatives to filtration are processing in:

– Settling centrifuges (→ Centrifuges, Sedimenting) or static settlers. Even poorly filterable suspensions sometimes settle readily, either the solids have high specific weight or big, soft flocs make an impermeable cake (e.g., wastewater sludge).
– Cross-flow filtration (see Section 2.4). This alternative however yields the solids as a thickened suspension only.
– Evaporation of the liquid (→ Evaporation). This alternative may be attractive if the required energy is not to high (highly concentrated suspension or a liquid with moderate

heat of evaporation, like many organic solvents).

Selection of Filter Aids. Filter aids are inert powders added to the liquid to be filtered and increasing the porosity and permeability of the cake (see Chap. 11). They are very helpful, provided the presence of filter aid in the solid can be accepted. Filter aids are used in two ways:

- As a precoat layer to protect the filter medium and improve filtrate clarity and
- As a body feed to increase flow rates.

Preliminary laboratory tests help to identify the proper kind of filter aid. The selection procedure is done in three steps:

1) Selection of the right material according to the chemical resistance and purity (perlite, diatomaceous earth, cellulose, carbon).
2) Selection of the grade (particle size) of the filter aid. The particle size should be as coarse as possible to give low filter resistance, but fine enough to prevent the dirt particles from trickling through the pores of the cake. For selection of the grade a precoat layer is prepared by filtering a diluted suspension of filter aid. Then the solution to be cleaned is filtered through this layer. After filtration the precoat layer is broken in half. The dirt should form a separate layer on the top and should not have penetrated into the precoat.
3) Selection of the quantity of filter aid. As a first guess the quantity of filter aid is calculated which gives a cake volume equal to the volume of dirt to be separated. It is then admixed to the suspension and the filterability of this mixture determined. If the flow rate is to low, the next trial should be made with a higher quantity of filter aid.

8. Filtration Equipment

Filters can be classified in accordance with different criteria:

Solids retention:	surface, deep-bed
Filtration technique:	cake, cross-flow, screening
Driving force:	pressure, vacuum, expression, magnetic or electric field, capillary
Operation:	discontinuous, continuous, quasicontinuous, automatic self-controlling, programmed, precoating
Filter element:	1) bag, belt, candle/cartridge, disk, drum, leaf/plate, nutsche, press plate, tube 2) horizontal, vertical, single-element, multi-element, stationary, rotating 3) open, closed
Solids discharge:	scraper removal, vibration, centrifugation, back blow, tipping, manual
Application:	rapid settling systems, moderate and slow settling systems, clarification of liquids, pastes, pulps, sludges, and nonfluid systems.

Because of numerous overlaps with respect to process engineering and design, any systematic arrangement entails a loss of clarity. In the descriptions that follow, the design element is given top priority.

The principles of operation and the design of filter equipment have changed little for many years. Only with recent efforts to improve the economics of processes have some crucial design changes and additions taken place. The most important of these involve

1) automation of operating modes and cycles
2) increases in specific throughput capacities and separation efficiencies, especially aimed at reducing residual moisture
3) optimization in the context of the overall process
4) improvement of operational reliability
5) increased concern for environmental protection and worker safety
6) adaptation to new, more difficult separation tasks
7) development of new separation methods for existing filtration problems

For example, classical filter presses have been replaced by completely automatic, unattended equipment. Mechanical dewatering devices have been installed on belt and drum filters. Classical open nutsches have been modified into "process filters," which combine several downstream operations (e.g., washing, dewatering, and drying) in a single apparatus. For difficultly filterable suspensions, dynamic filters with membranes have been designed; together with elec-

trokinetic filters, these open up new possibilities for mechanical liquid–solid separations.

Important advances in the performance of existing filters have come about through better understanding of the theoretical relationships in all areas, such as filtration, washing, and dewatering and expression. In particular, the automation and optimization of processes, based on knowledge about the interactions of the many process parameters, are currently being pursued in extensive programs.

8.1. Bag Filters

The filter element in this type of filter is a bag made of the filter medium. A typical bag has an opening at one end. The bags are mounted in various kinds of containers; they allow an uncomplicated design, are easy to service, and are relatively versatile in view of the large selection of filter media available. For example, solids in the range of 80 – 800 µm are successfully collected with bags made of fabrics woven from monofilament yarns; solids between 1 and 200 µm are separated with needled felt bags or by precoat filtration (in which deep-bed filtration is active at the same time).

As batch filters, are opened by hand, they are employed only where relatively small charges with extremely low solids contents are to be clarified. They therefore find use chiefly in clarifying and polishing filtration (sterilization), where the solids to be removed are impurities or undesirable byproducts. Applications include chemical and pharmaceutical products, paints and varnishes, juices, edible oils, waxes and resins.

Bags open on one side are charged from the inside, so that the solids are collected as a filter cake in the bag. Service times range from hours to days. In pressure bag filters, where the bags are inserted in supporting cages, pressures between 0.5 and 1 MPa are common; in special designs, the pressures are up to 2.5 MPa. Capacities are in the range of $1\,m^3\,m^{-2}\,min^{-1}$ at 1.6 MPa. Filter units with several bags in housings are available at total capacities of up to $2.4\,m^3$/min ($4.8\,m^3$/min maximum) and down to quite small values.

Pocket-shaped filter bags are charged from outside; they are slipped over collapsible filter inserts attached to the filtrate pipe (Fig. 36). The open corner of the filter bag is sealed against the filtrate pipe. In order to increase the filtration area, the bags are made up to three times wider than the inserts, so that folds are produced when they are slipped on; in this way, the space requirement per square meter of filtration area is greatly reduced. Filter units up to $250\,m^2$ are available. After filtration, the bags can be cleaned by spraying and blowing with compressed air.

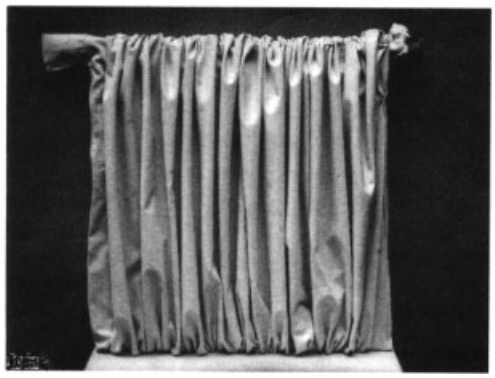

Figure 36. Pocket-type bag filter

8.2. Belt Filters

A belt filter resembles an ordinary belt conveyor driven by one of the guide rollers. Slurry is fed onto the filter belt from above and the filtrate drains or is pulled by suction from the bottom of the belt. Belt filters are operated as simple gravity, vacuum and pressure filters. Whether operation is continuous or intermittent (semicontinuous), slurry is fed onto one end of the belt. The filtrate drains into vacuum tanks, which can be stationary or can move intermittently with the belt. The cake can be washed and dewatered in downstream zones. If necessary, the filter belt is washed during recycling. In another design, the belt is made up of a number of individual cells, which bear the filter cake. In this way, a more careful separation of the filtration stages (see also Fig. 1) is achieved.

In general, belt filters are applied for readily filterable suspensions that contain coarser, and therefore easily sedimented, solids (groups F and M in Table 2), where the filtrate and the wash liquor are to be collected separately and the solids must have gentle handling. The maximum

Table 2. Filterability of cakes

Filtration characteristics	Group of suspension, symbol				
	Fast (F)	Medium (M)	Slow (S)	Dilute (D)	Very dilute (VD)
Initial cake formation rate, min/cm	0.005 – 0.1	0.1 – 1	1 – 10	10 – 100	no cake
Slurry concentration, %	> 20	10 to 20	1 to 10	< 5	< 0.1
Settling rate	rapid	medium	slow	slow	
Leaf test rate, kg/m^2 h	> 2500	250 to 2500	25 to 250	< 25	no test
Filtrate rate, m^3/m^2 h	> 10	0.5 to 10	0.0025 to 0.05	0.025 to 0.05	0.025 to 0.05
Typical slurry	crystalline solids	salts	pigments	wastewater	water

filtration area is 120 m^2. Fluctuations in product call for control of the belt speed or slurry rate.

For especially sensitive filtrations, belt filters are used with filter media that remove the liquid from the cake by capillary action alone. The filter cloth runs over felt or similar filter media.

Vacuum Belt Filters. A continuous vacuum belt filter consists of the moving filter cloth supported on a profiled elastomer transport belt (Fig. 37). To provide better sealing for the vacuum space, sliding belts can be placed under the transport belt, or a specially shaped belt can travel along with the filter cloth. The difficulties met with in ensuring a good vacuum with moving seal systems are overcome when the filter belt or filtrate box move intermittently. Figure 38 shows an example. When the belt is stationary, slurry is fed to the filter and at the same time the filter cake is washed, dewatered and, if necessary, pressed. In the intermittent movement of the belt, the discharge roller shifts the belt, while the retaining roller holds the belt in place. The movement is compensated by movement of the compensating roller in the opposite direction. The vacuum is let down when the belt is moving.

Another continuous vacuum belt filter has vacuum tanks of pan-shaped sections each 1.4 m long. Each section is divided into two parts with flexible connections to the vacuum system (vacuum pumps). In this way the individual filtration stages are divided into arbitrary lengths as needed. The vacuum tanks are designed with grids that allow the filtrate to drain freely. During filtration, the tanks move along with the belt at the same speed, as the vacuum builds up. At the end of the co-moving path, the vacuum is broken and the tank vented, while the filter cloth moves on. Slurry and wash liquors used are fed continuously.

Vacuum belt filters are usually installed in open frames, or in closed housings if no vapor is to be released into the environment. Good filtration capacities and washing effects can be achieved only with uniform loading of the filter medium or filter cake. For this reason, the feed and wash liquor are distributed over the width of the filter belt by flat nozzles or washing grooves. Downstream dewatering of the filter cake is done by double-belt expression (see twin-belt press, page 36) or a plate press.

Because of the versatility of vacuum belt filters, they have found use in nearly every field of liquid–solid separations where slurries are relatively easy to separate.

Even difficultly filterable slurries with comparatively finely dispersed solids can be separated on belt filters if suitable belt materials are employed and the solids do not immediately block the medium. In such cases, pressure is most commonly used instead of vacuum (see page 36).

Belt filters for gravity and suction filtration offer low equipment costs. With a suitable filter medium (nonwovens are generally used), they can successfully clean up slurries that are not too highly concentrated (solids down to 50 mg/m^3) and of the filterability group F in Table 2. Feeds include machine-tool cooling and cutting lubricants (emulsions), electrolysis cell slimes, beer (for removal of turbidity and yeast), wastewaters, and chemical solutions.

A *gravity* or *hydrostatic filter* with a continuous belt (Fig. 39) consists of a tank, in which a moving support belt made of coarse woven fabric or a flexible belt forms a depression. The filter belt proper rests on this support belt. Operation is intermittent: slurry is fed in until the

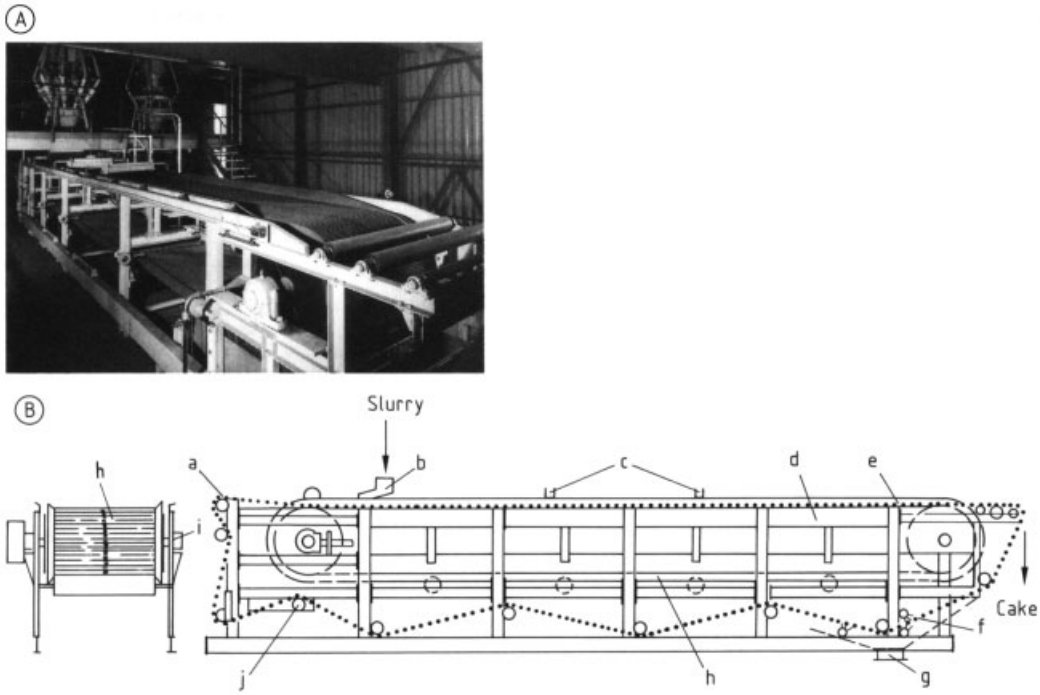

Figure 37. A), B) Vacuum belt filter (reproduced with permission of Dorr – Oliver)
a) Filter cloth take-up assembly; b) Feed box; c) Cake wash; d) Vacuum pan; e) Filter cloth; f) Filter cloth wash; g) Drip pan drain; h) Drainage belt; i) Tensioning device; j) Filter cloth aligning; k) Drainage belt

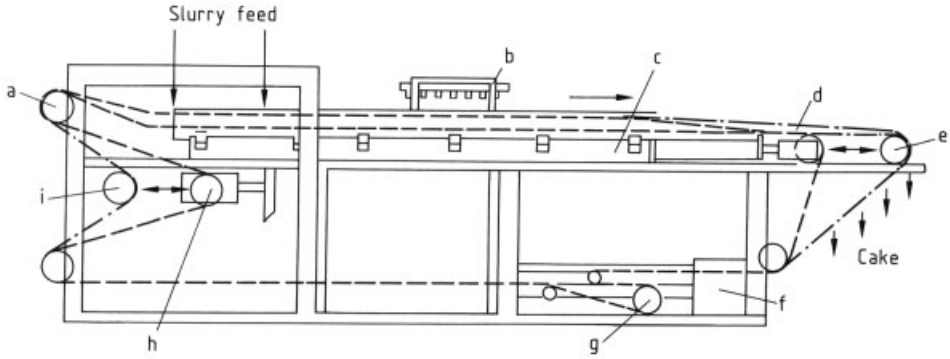

Figure 38. Jerking-type vacuum belt filter (BAS Sonthofen)
a) Filter cloth take-up assembly; b) Cake wash; c) Vacuum trays; d) Filter cloth; e) Movable discharge roller; f) Cloth wash assembly; g) Blocking roller; h) Tensioning assembly; i) Floating roller

increasing resistance to flow causes the liquid above the filter belt to reach a predetermined depth. The hydrostatic head promotes the filtration rate. The belt is then advanced, the filter cake carried out of the tank, and new filter medium transported into it. The filtrate flow rate is $0.2\,\mathrm{m^3\,m^{-2}\,min^{-1}}$.

There are two designs of *suction belt filters*, also called *flat-bed filters*. One, used chiefly for wastewater treatment, consists of a receiving tank for slurry, with a drag chain conveyor. The perforated or slotted floor of the tank holds the filtrate receiver, which is connected to the pump. The filter medium is guided between the scrapers and the tank floor. After a filter cake has built

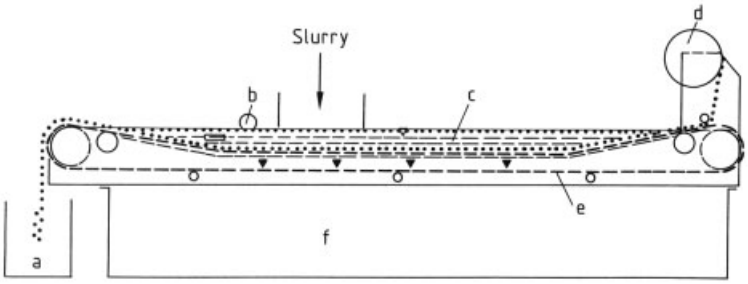

Figure 39. Hydrostatic belt filter
a) Sludge bin; b) Float control; c) Filter trough; d) Filter medium; e) Carrier belt; f) Filtrate tank

up to a predetermined thickness, the increasing pressure drop causes the conveyor to begin moving. It advances some 20 – 50 cm along with the filter cloth. These filters are made with filtration areas up to 21 m^2.

The other design consists of a vacuum tank with lid; the paper filter tape is fed between the tank and the lid. This type of filter also operates intermittently. At a filter area of 1 m^2 a throughput of 0.1 – 0.5 m^3/h can be achieved, depending on the amount of solids in the suspension and the presence of filter aids.

Pressure Belt Filters. A pressure belt filter of the flat bed type consists of a filtrate receiver and a pressure chamber. The filter medium is guided between the two chambers, which are situated in a housing. Operation is necessarily intermittent, since the pressure chamber, which is provided with slotted gates, is pressurized during filtration. In smaller units, the use of slotted gates is replaced by pivoting up the pressurized section. Pressure belt filters are also made with expression devices (up to some 2 MPa).

A *twin-belt pressure filter* is a belt filter with a second, co-moving belt, which exerts an additional mechanical compressive load on the filter cake with pressure rollers. Such a filter can be employed provided the cake is deformable. This condition holds especially for bulky flocculated slimes such as occur with filter cakes made up of relatively fine solid particles (mainly of organic origin) having a broad particle size distribution. Twin-belt pressure filters have proved themselves for the *dewatering of sewage sludges* that have been treated with polyelectrolytes (based on poly(acrylic acid) and polyacrylamide) to produce relatively large, stable flocs.

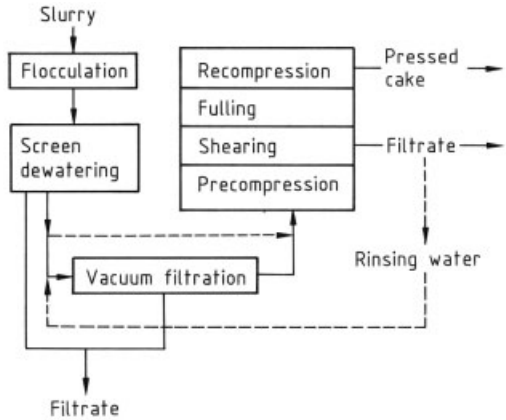

Figure 40. Schematic processing in twin-belt pressure filters

This type of filter has made it possible to reduce the relatively high 70 vol % residual moisture in municipal sewage sludges by as much as 10 vol %, corresponding to a decrease of 35.7 % in the moisture content per cubic meter of sludge. Further applications have also been tested, and this class of filter has been adopted for some dewatering jobs, such as slimes from coal washing, large-scale animal husbandry, cellulose and pulp production, electrolysis, and off-gas treatment. An average dewatering capacity for designs now current is 6 m^3 of municipal sewage sludge per meter of belt width and hour. With sludges that are readily dewaterable, the influent rate can be increased to 15 m^3/h.

Machines with multistage dewatering are used in nearly all these sample applications. Figure 40 shows a simplified diagram. The flocculated sludge is put through preliminary dewatering in a simple screen drainage (rotating or flat screens), since part of the liquid bound up in the

Figure 41. Twin-belt pressure filter (reproduced with permission of Flottweg)

sludge is released in the flocculation step. Vacuum filtration usually follows before the filter cake reaches the expression zone.

A special feature of twin-belt pressure filters is that the belts change direction several times, on rollers that shear and press the cake, thus promoting its compaction. Usually the filtrate produced in the several stages, which is not entirely clear, is recycled to the first stage, where the solids still dispersed in the filtrate are reflocculated. The filtrate resulting here is largely free of solids and can be passed on to, for example, biological treatment.

Quite a variety of arrangements of the individual process stages and design features has come into being. The performance of these machines thus varies widely, being influenced as well by the properties of the influent sludges.

Twin-belt pressure filters will find additional use in food treatment, for instance in juice extraction from fruits and vegetables. Figure 41 shows a typical arrangement of that twin-belt press filters including horizontal press section and several turns of the belts. Here the recovery of juice may reach 65 – 70 % in a 2500 mm machine and 7 – 10 t/h throughput.

In the modification of Figure 42 A, the sludge enters the expression zones from below, through a straining zone (gravity dewatering). Passing over rolls that vary in diameter, the cake is sheared and compressed (up to somewhat over 100 kPa). The modification of Figure 42 B applies a further increased compressive load with a third moving flat belt.

Figure 42 C shows a filter in which an overflow (for gentle handling of the flocculated solids) directly connects the flocculation vessel to the gravity filtration and dewatering stages; the filter cake passes downward through the successive expressing zones. The last stage is supplemented by a device that exerts a controlled linear pressure.

Cake compression in twin-belt pressure filters is rather modest in comparison with the compressions achieved in membrane filter presses (see page 58). The twin-belt devices have the advantage of continuous operation and relatively low cost. A recent development demonstrates the possibility of reaching pressures in twin-belt expressing filters that are similar to those in membrane filters (up to 2 MPa). A filtration zone with filter belts converging in wedge fashion is followed by compaction as the cake turns around a drum pressurized by hydraulically actuated elements.

Experience and theoretical analyses (see Section 4.2) lead to the conclusion that optimal performance is attained in twin-belt pressure filters

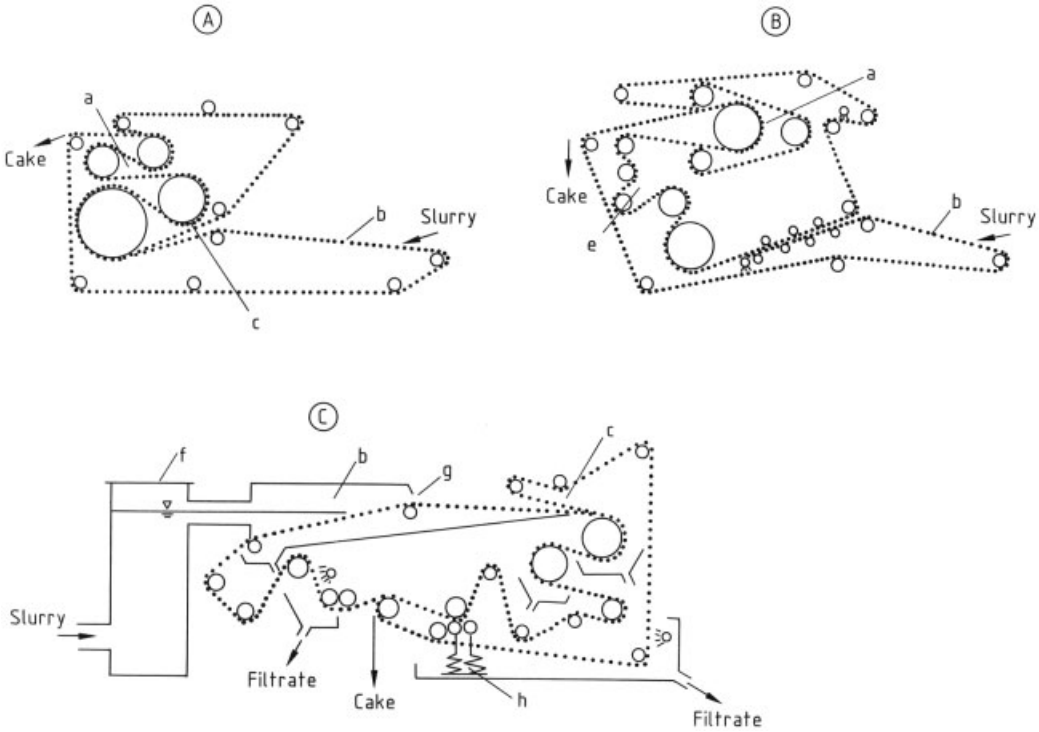

Figure 42. Modifications of twin-belt pressure filters (A, B, and C)
a) Press zone; b) Gravity strainer zone; c) Pre-compression zone; d) Press belt high compression; e) Low compression zone; f) Flocculation; g) Cake break-up; h) Line compression

when the cake is relatively thin, the filtrate drains on both sides, the expressing times are long (especially for readily compressible cakes), and the expressing force is as great as possible.

8.3. Candle Filters

Candle filters (cartridge filters) consist of tubular elements, mostly covered by sleeves of a filter medium. They are widely used in the filtration of weakly to extremely weakly concentrated slurries that contain solids more or less as impurities at concentrations down to the ppm range. This class includes all filters with cylindrical elements mounted, singly or multiply, more or less vertically in tanks (Fig. 43 A).

All candle filters have the fundamental advantages that the filter elements are relatively simple in construction and cheap to fabricate, and that the devices are adaptable to requirements imposed on separation efficiency and throughput capacity. The elements are mounted either suspended from an overhead tube sheet or upright on a plate (except where each element has its own housing). There are advantages to the suspended design, but great care is required with elements that break easily.

Most candle filters are operated as pressure filters. The elements can be cleaned by backwashing, or they may be replaceable. Some multiple-element filters offer filtration areas of far more than 100 m^2. As typical clarifying filters, these devices have a broad field of application (groups D and VD in Table 2). The filtration surface can take a wide variety of shapes, depending on the job being done.

Sheet Candles. A simple candle of this type consists of wire mesh or perforated screen made of AISI 316 stainless steel, phosphor bronze or man-made fibers supported on a tubular metal core (Fig. 43 B). Slotted screens and wire-wound filters can also be employed. Such elements are used mainly to collect weakly-concentrated solids from liquids such as oils and soap and

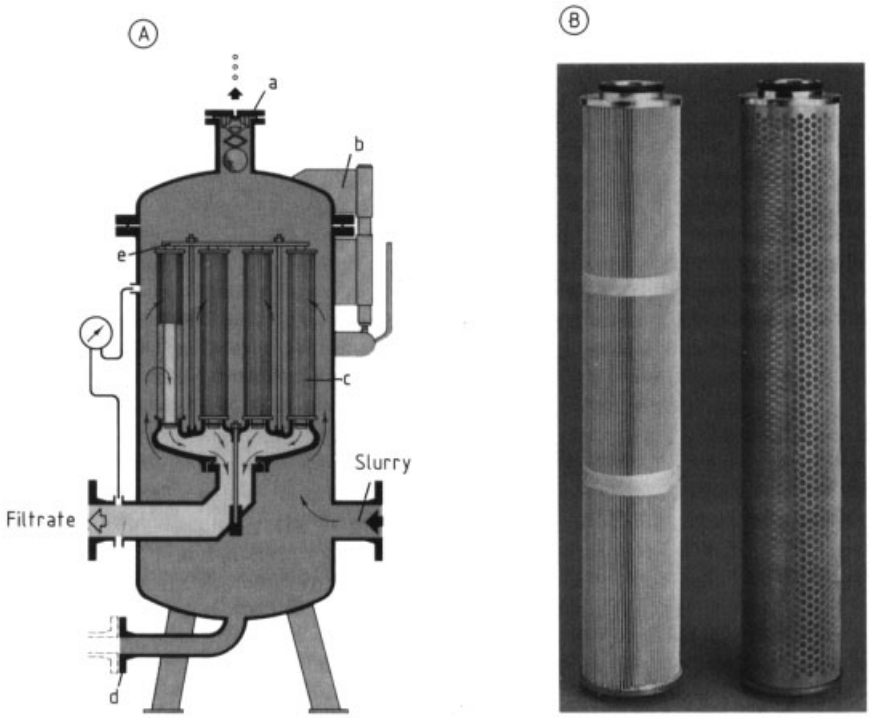

Figure 43. A) Sectional view of candle filters and B) tubular filter elements (reproduced with permission of Faudi Filter) a) Deaeration; b) Opening device; c) Filter candle d) Discharge; e) Tensioning plate

dyestuff solutions; mesh openings range from 0.01 mm to 0.5 mm (down to 2 µm in special meshes). Backwashing is desirable for wire-mesh elements.

Disposable sleeves of paper or nonwoven synthetics are used to retain finer solids. In part, these filter media simultaneously act as deep bed filters, the solids being deposited in the medium itself. These simple filter elements offer comparatively high flow rates. Available filter units cover the range from a few liters per minute to several hundred cubic meters per hour. Enclosed in pressure-tight vessels, they can be operated at absolute pressures of up to 10 MPa. The pressure drop across such a filter is around 10 kPa.

Gap-Type Candle Filters. Gap-type candle filters can be made of disks, profiled wire, or tubes (Fig. 44). Most are held together by straps. In one design, the gap-forming elements are slipped or welded onto a tube; filters of this type are used as well strainers. A special drinking water filter element is the gravel-filled prepack cartridge.

A continuously operating gap-type cartridge filter has several filter elements, charged on the inside, arranged on a rotating mount that brings them one after another to a stationary backwash device.

Cartridge Filters. Cartridges are designed for deep-bed filtration and are thus used mainly for fine filtration of dilute slurries, concentrations of less than 0.01 % and particle sizes less than 20 µm. They usually comprise replaceable cartridges made of a variety of materials such as sintered metals (AISI 316 stainless steel), ceramics (also usable at high temperatures), plastics, cellulose or glass fiber (resistant to most liquids). Wound cartridges of cotton, cellulose or manmade polymer fibers, often combined with nonwovens, are distinguished by relatively large spaces (honeycomb) for retaining solids. The simple construction and low material costs open

up technically and economically interesting possibilities.

Large-Area Filters. Large-area filter elements are made by folding the filter medium along the direction of the cartridge (Fig. 45) or by using pouch and disk constructions. The filter medium is again mounted on a center tube. Elements can also be built up in several layers. Folded-paper cartridges of this type offer 170 m^2 of filtration area per unit. Pouch- and disk-type filters have far less area per unit (up to 0.5 m^2) but can be cleaned and are therefore reusable. Up to 3.45 m^2 of filtration area per unit is available in pouch filters. A disk filter consists of disks of pulp, metal screen, or membrane filter medium slipped on a central support, alternating with support disks, which have slots to admit the slurry into the elements.

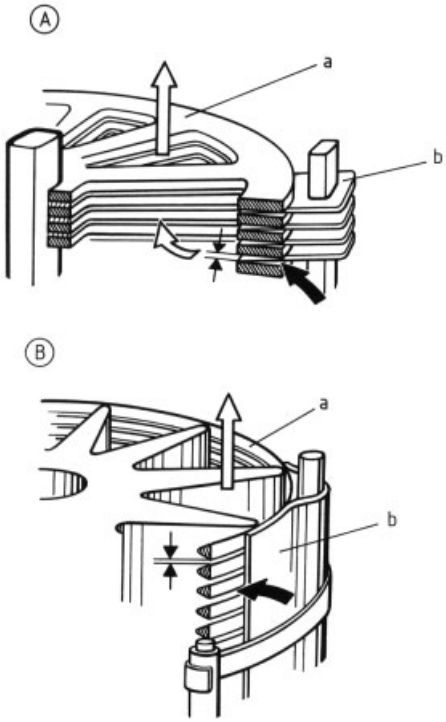

Figure 44. Gap-type candle filter
A: a) Steel disk package (turned for solids discharge); b) Fixed gap cleaner
B: a) Steel wire spiral (turned for solids discharge); b) Fixed scraper

Other Designs. A number of other cartridge designs have been devised for special tasks. For clarifying difficultly filterable liquids, such as some pharmaceutical and biological liquids (blood, among others), elements made of fibrous bases or materials coated with activated carbon give satisfactory performance in the micrometer range (microfiltration).

A recently developed element is assembled from seven individual tubes having slots for the passage of filtrate (Fig. 46). A tube bundle is covered with a suitable high-pressure fabric, which fits tightly around the curved surfaces during filtration. During the backwash cycle, the fabric is lifted off the tubes; this movement causes the cake residue to be thrown off. The filter medium has a tubular weave, that is, it forms a seamless tube that can withstand several bars of backwash pressure. Therefore, its cleaning does not cause any concern. This large-area filter element is stable and can be used especially well as a backwash filter with a cake thickness between 3 and 50 mm. With up to 200 m^2 of filtration area, this candle-bag filter is suitable for thickening and clarification of suspension as well as for solids treatment, such as washing, extraction, or steam drying. Finally, the cake can be reslurried or discharged as dry solids. Typical applications are catalyst or activated carbon removal, filtration of pigments or impurities and treatment of polymer intermediates.

Design and Selection. Candle filters are generally designed to the flow capacity, filtration pressure and separating power stated by the manufacturer. The selection of the filter medium calls for some experience.

It can often be advantageous to use several units with small filtration areas in place of one larger-area unit. With several units in parallel, for example, one element at a time can be cut out of the line for washing while the remaining ones stay in the filtration stage.

8.4. Deep-Bed Filters

This section deals only with depth filters having beds of sand, gravel or similar loose materials.

A depth filter is usually a simple container with a porous bottom, on which a bed of filter medium between 0.5 and 1.5 m deep is placed. If the bed material is gravel or sand, such a filter is also called a sand filter. Figure 47 shows

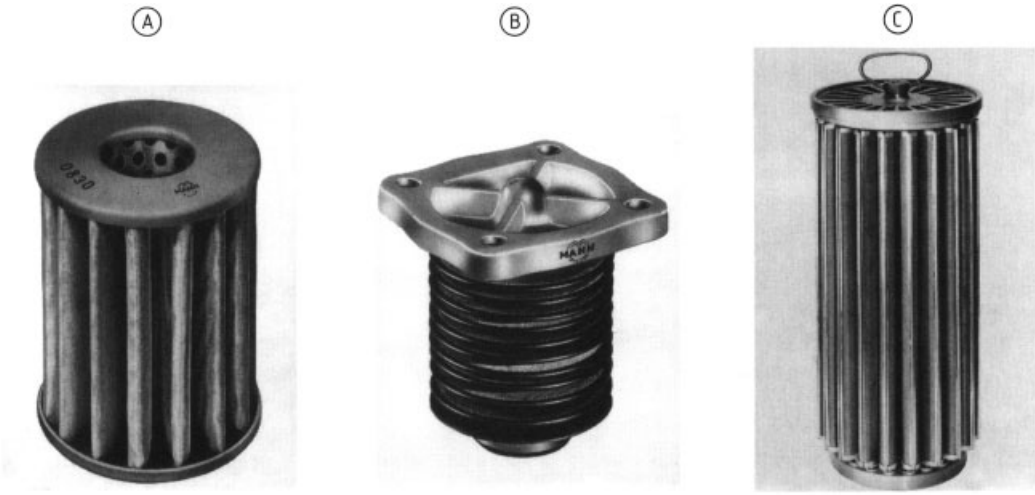

Figure 45. Candle-type strainer filters (reproduced with permission of Mann & Hummel)
A) Radial fold filter; B) Disk filter; C) Multitube filter

a simple form of a gravity sand filter. The filter bed (c) is supported by the screen (a), and the filtrate is discharged through pipe (e). Water, steam or air can be admitted through pipes (g) for cleaning. Wash water overflows at (b), if at all. When agitation is used in cleaning, the wash liquor is discharged at (f). The cooling or heating coils (e) make it possible to temper the filter. Depth filters are also operated as pressure filters, either in order to increase the throughput or when they are installed in pressurized lines. To increase the filtration area of pressure sand filters, several beds can be arranged one above another.

Most filters of this type are used in the treatment of drinking water and process water. Basin-type sand filters have been introduced where quantities are very large, since these fixed structures offer relatively large areas and depths.

Flow in a depth filter is usually from top to bottom. The flow is reversed in order to clean the bed. The basic operating modes differ little even though a variety of techniques are possible, especially for cleaning.

With finer-grained bed materials used to collect finely dispersed solids, the topmost layer of the bed tends to become plugged, so that cleaning may become necessary before the entire bed is loaded with solids. In such cases, layers differing in particle size can be used; a new cleaning method is needed, however, to avoid mixing the fractions. In such multilayer filters, the liquid can be introduced at different points. This approach gives a 70 % increase in flow velocity, substantially longer service time, and enhanced sludge capacity.

When a depth filter is employed to clarify juices, solutions and similar valuable liquids, enclosed tanks are used instead of open vessels, and higher pressures can also be applied. Pressure sand filters usually permit two to three times as great a pressure drop as gravity filters, and so finer-grained bed materials can be used as well. With a few exceptions, bed depths in such filters are much less than the depths usually found in water-treatment plants. In industry, depth filters with beds of special materials such as activated carbon or ion-exchange resins are often installed so that dissolved substances can be removed simultaneously.

In a *continuous process* (Fig. 48), the influent is initially in an unpacked internal space, whence it enters the bed through conical internals. The clarified liquid (filtrate) drains from the bed into an external space. The filter medium is hydraulically removed at the bottom and cleaned in a regeneration tank at the top. The clean sand is returned to the filter vessel.

For smaller units (up to roughly 600 m^3/h), a backwash-cleaning filter has been developed for largely unattended operation. When solids deposited in the filter bed cause the differential

pressure to increase, the liquid level in the tank will also rise until its hydrostatic pressure initiates backwashing. Because the cross-sectional area of the backwash pipe is larger than that of the influent pipe, the flow velocity can be roughly four times as great (44 m/h).

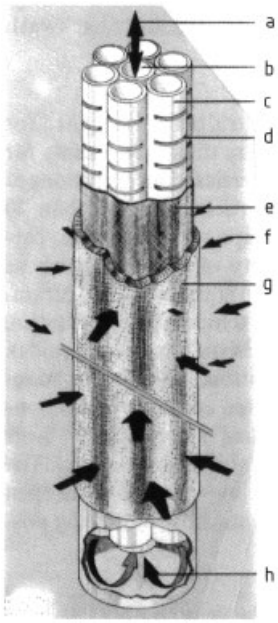

Figure 46. Candle-bag FUNDA-BAC filter (reproduced with permission of Dr. Müller, Männedorf)
a) Filtrate/blowback-air; b) Central tube; c) Concentric tubes; d) Horizontal slots for filtrate entry–blowback; e) Filter medium; f) Slurry; g) Filter cake; h) Filtrate coming down concentric tubes and rising in central tube

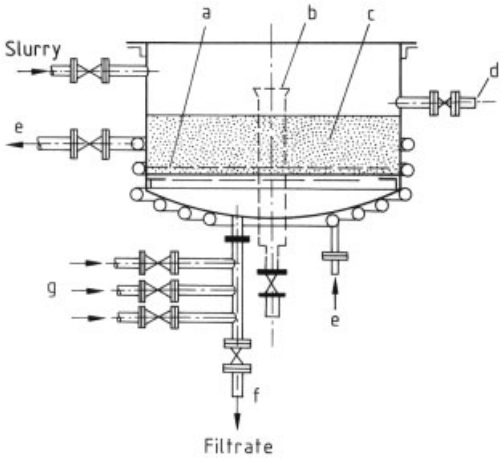

Figure 47. Open "low" sand filter
a) Screen; b) Effluent; c) Sand bed; d) Wash water; e) Heating or cooling system; f) Filtrate; g) Cleaning pipes

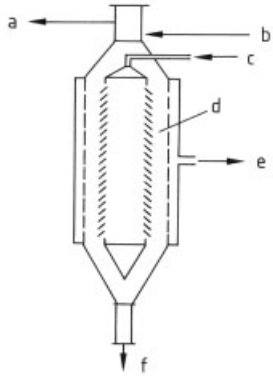

Figure 48. Continuous sand filter
a) Sludge; b) Sand and sludge; c) Slurry; d) Sand bed; e) Filtrate; f) Sand and sludge

8.5. Disk Filters

The element in a disk filter is a rotating vertical disk (Fig. 49); one or more such disks are arranged on a shaft, with roughly half of each disk submerged in a slurry tank (Fig. 50). The device allows continuous filtration. The disks are made up of 8 – 10 individual sectors, each of which is an independent filter. The interior space of a sector is connected to the filtrate discharge by a rotary valve, so that a cake can form on the outside when the sector is submerged in the tank. After the submersion part of the cycle, air is drawn or blown through the sector in order to effect dewatering; finally, the cake is removed by simultaneous application of a scraper and an air blast, or by spraying.

The disks are 1.2 – 3.6 m in diameter; in a special design, disks up to 5 m in diameter were used. Multi-disk filters have filtration areas of up to 280 m^2. In devices with relatively large diameters, hydrostatic effects may cause nonuniform cake thickness.

When several disks are mounted in a single tank, the tank is fitted with an agitator (usually a paddle mixer) or an air turbulizer in order to prevent sedimentation of the solids, which would lead to uneven cakes and poor dewatering. If not with an agitator, the multi-disk filter is built with a divided tank. The slurry level in the tank dictates the disk rotation speed.

The disk filter is commonly operated as a vacuum filter. The tank is open and readily accessible. If vapors generated at elevated temperatures

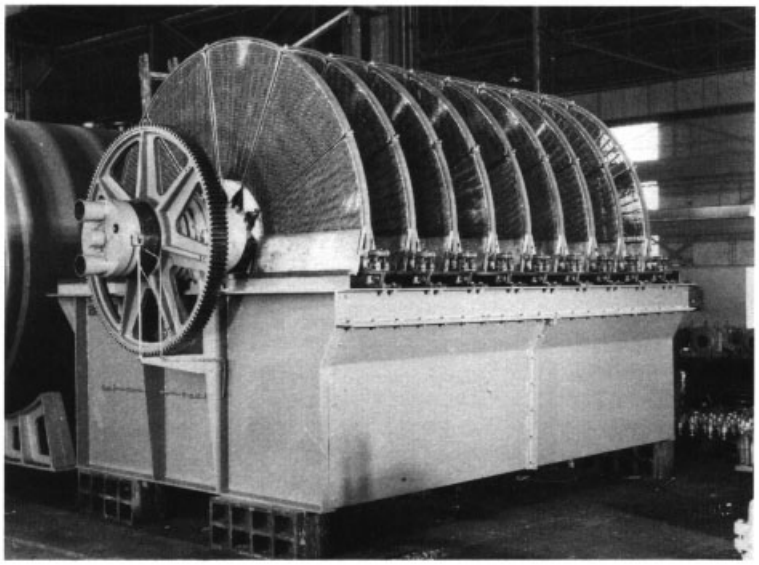

Figure 49. Vacuum disk filter

must be kept from escaping into the environment, the filter can be mounted in a pressure-tight housing and operated under pressure; for hyperbaric operation, see also Section 4.1.4.

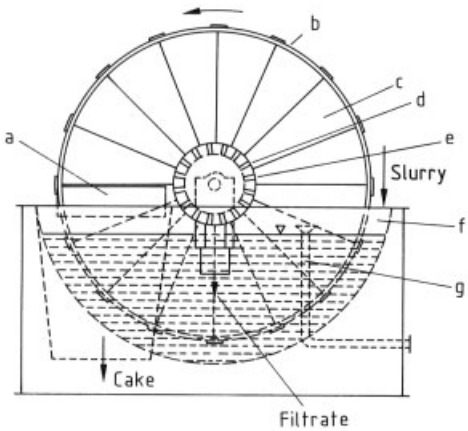

Figure 50. Vacuum disk filter
a) Scraper; b) Filter disk; c) Trapezoidal sectors; d) Outlet nipples ; e) Automatic valve; f) Filter tank; g) Overflow

In principle, the filter cake in a vacuum disk filter can be washed, but it is certainly less effective than in, e.g., belt filters. The basic advantage of the continuous vacuum filter is that it requires little space per unit of filtration area. Another advantage is that the filter elements are readily accessible. In some cases, this type of filter has the drawback that cracks develop in the cake and cannot be smoothed.

Applications center on high-tonnage products where large filtration areas are needed, e.g., iron ore slimes, flotation concentrates, white water from papermaking machines, blast-furnace dust from scrubbers, and aluminum hydrate (groups M and S, Table 2).

8.6. Drum Filters

Wrapping the filtration surface around a cylinder yields significant advantages that have made drum filters one of the most widely used types. Their advantages are continuous operation, easy charging, a choice of cake-removal methods, low sensitivity to suspension variations, good possibilities for washing in some cases, satisfactory residual moisture, relatively simple designs, reliability, and economy.

In one revolution, the drum passes through the processing stages, thus allowing continuous movement of filter cake and filtrate. Filters can be operated as gravity, vacuum or pressure filters.

Gravity Drum Filters (Revolving Screens). Applications of gravity drum filters are limited to simple screening, where cake formation is not

crucial to the filtration operation, that is, where the cake does not greatly increase the filtration resistance. Thus the nature of the solids governs the application: gravity drum filters are suitable for filtering coarse-grained or fibrous mixtures, for example, treating surface waters, papermaking wastewaters, and the like. They can be charged on the inside or the outside. When the filter is internally charged, the untreated liquid flows through a distributing channel to the inside of the drum. As the drum turns, the solids collected by gravity move upward, separate and are removed, for example with a screw conveyor.

The drum is covered with a screen fabric. With microscreens, it is possible to collect solids that are somewhat more difficult to filter, such as occur in sewage treatment.

A filter charged from outside is used in the treatment of recycle waters and wastewaters in paper and pulp manufacturing. The drum is almost completely submerged in the slurry. While the solids are held up on the screen surface (an easily removable cake forms in the case of fibrous solids), the virtually fiber-free filtrate runs into the lowermost part of the drum. For a drum roughly 2 m in diameter turning at a peripheral speed of 3 – 15 m/min, throughputs of 0.5 – 5.0 m^3/min per meter of drum width are possible. Activated sludge flocs are collected with a drum 3 m in diameter and 3 m long, lined with a microscreen having a mesh opening of 10 µm; the throughput is 630 m^3/h.

Vacuum Drum Filters. In order to increase throughputs and extend the range of applications, the inside or outside of the drum is enclosed in a housing connected to a vacuum pump. In the simplest case, the drum has heads on either end, so that the entire inside space is under a uniform vacuum. Roughly 30 % of the drum is submerged in a tank, which fits fairly closely around it. Slurry in the tank is held at a constant level by means of an adjustable overflow. The feed is distributed by a stirrer, whose oscillating movement along the drum is intended to prevent partial settling of the suspended solids. In the submerged part of the drum, the vacuum pulls liquid through the filter medium while a cake builds up on the outside. The cake exposed on the top part of the drum is dewatered by air pulled through it. Shortly before the drum re-enters the tank, the cake is removed by an air blast from inside the drum, falling through a chute into a collector.

Single-cell vacuum drum filters are built with filtration areas of up to 40 m^2. They are suitable for all slurries that can be held suspended in the tank, that is, slurries that do not contain readily sedimenting solids (group S in Table 2), and where the cake need not be washed. They feature a high working vacuum, since there are no high losses; good dewatering, since relatively long blowing times are possible; and simple construction.

Drum filters generally have continuously variable drives so that they can accommodate a range of product properties. The filter medium (a finely-woven or fine-pored fabric) is supported on a widely spaced wire mesh.

Because of its simple design, the single-cell drum filter is well-suited to precoat filtration (see also page 65 and Chap. 11). In a special version where the common path of filtrate and air is short, it is preferred in the processing of cloudy wine and beverages. Throughputs varies from 0.2 – 1.2 $m^3\,m^{-2}\,h^{-1}$, depending on the kind of beverage. In order to produce a uniform precoat, which may be 10 – 12 cm thick (this layer is continuously scraped off during filtration), the concentration of precoat material (diatomaceous earth) is low, the depth of submersion is slight, and the drum is rotated rapidly. If the drum shell is divided into sections 50 – 60 mm wide and a tank is added to receive the wash liquor, cake washing is also possible in single-cell drum filters.

A special variety of this filter is the *suction roll* used in papermaking. It has a suction box, which restricts the vacuum to a relatively narrow zone at the top. Suction rolls make it possible to increase the speed of the papermaking machine from 250 m/min with couch rolls to over 1000 m/min.

Multi-compartment Drum Filters. Washing, as well as the separate recovery of filtrates from the several process steps, are made easier in multi-compartment drum filters. The filtration surface in such a device is subdivided into flat, closed cells, which have their filtrate sides connected through separate pipes to vacuum pump(s) (Fig. 51). The compartments usually have perforated inserts bearing, for example, drainage channels on the cake side and filtrate collection channels on the vacuum side. The

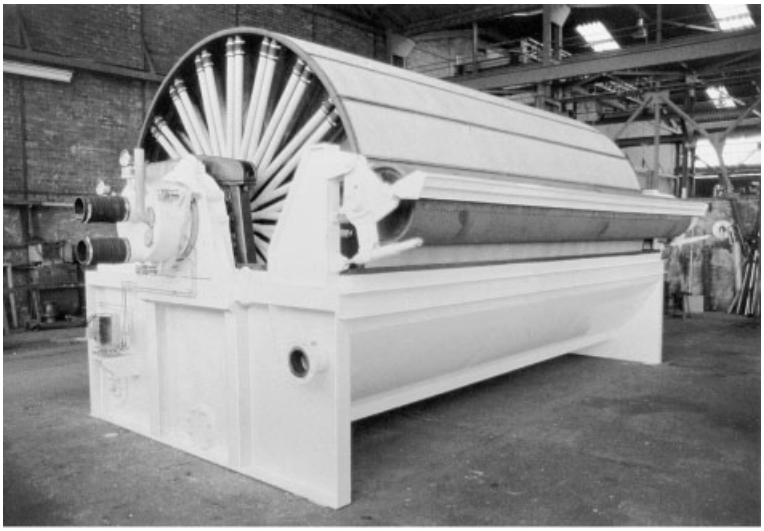

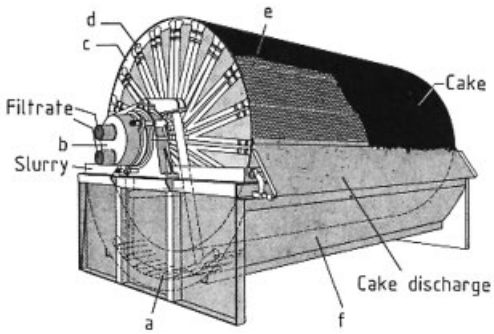

Figure 51. A) Vacuum drum filter (reproduced with permission of Dorr – Oliver)
B) Vacuum drum filter
a) Agitator; b) Valve; c) Drum; d) Filter pipes; e) Division strips; f) Tank

compartment insert bears a fabric, which supports the actual filter medium. While this fabric is installed separately for each compartment, the filter cloth itself covers as much of the drum periphery as possible.

The filtrate pipes rotate with the drum and have a sliding seal connection (rotary valve) to the fixed lines leading to the supply tanks and pumps. The valve also serves as control means, allotting the desired process stages to the several sectors of the drum. The control arrangement makes it possible to set up any desired filtration cycle; this is an undisputed advantage of the compartment filter.

The slurry feed is on the ascending side of the drum. An overflow pipe provides level control and thus maintains steady filtration conditions; the pipe can be adjustable. A stirrer in the bottom of the tank has the function of preventing the solids from settling in order to keep from upsetting the filtration. It is not always desirable to start filtration immediately after the drum is submerged, since with sedimenting feeds this would mean picking up the fine fraction first and thus impairing the permeability of the filter medium. The cake emerging from the tank should first be dewatered and then washed.

If cracks appear in the cake on dewatering, they will interfere with uniform washing. The cake surface can then be smoothed with pressure rolls. Washing should be spread over a large portion of the drum circumference, and so with large quantities of wash liquor there is a danger that the cake will slip off. To prevent this, washing belts are placed around the drum. Like the expressing belts described further on (see

page 36), these surround the drum and have the dual function of closing up cracks and effecting good distribution of the wash liquor. The wash liquor is usually delivered to the cake through drip tubes, channels or nozzles.

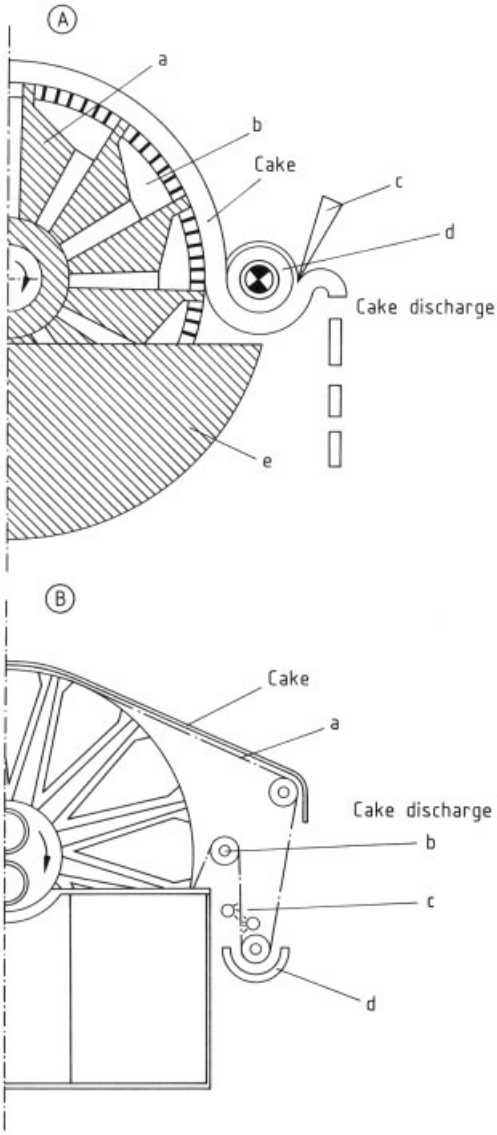

Figure 52. Cake discharge assemblies in vacuum filters
A) Roller discharge; a) Filter drum; b) Filtrate pipe; c) Scraper; d) Roller; e) Slurry tank;
B) Belt discharge; a) Filter cloth; b) Tensioning device; c) Cloth wash; d) Wash liquid receiver

A variety of devices and techniques are used to remove various cakes from the filter medium. *Scraper discharge* is the earliest and simplest method (Fig. 51 B); it is best suited to rigid, non-deformable cakes at least 4 mm thick. Lifting of the cake can be reinforced by an air blast supplied below the filter medium through the filtrate pipes.

In *string and chain discharge*, parallel strings or chains are passed over the drum in the circumferential direction. After the last dewatering step, these pass over a deflecting roll with a relatively small diameter, and the cake then drops off. Chain discharge is suitable for relatively thick cakes such as are produced in the filtration of coarsely crystalline materials. "Felted" cakes, on the other hand, call for string discharge.

Roller Discharge. Thin filter cakes, down to 1 mm and thinner, can be taken up by rollers (Fig. 52 A). Contact between roller and drum should not be too firm. The roller turns at the same speed or slightly (3 – 5 %) faster. The cake is removed from the roller by a scraper, comb or other means; part is left on the roller to promote separation from the drum.

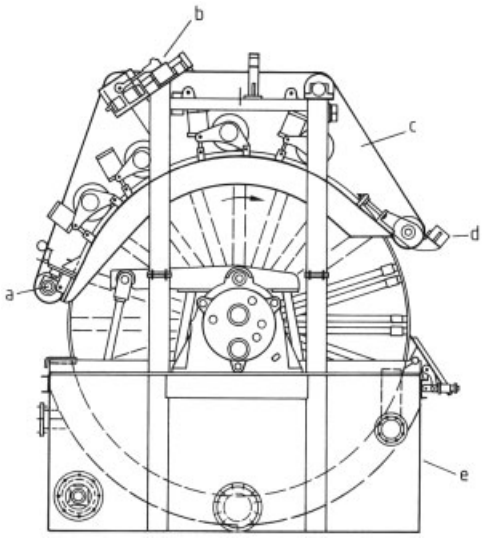

Figure 53. Combined vacuum drum pressure belt filter (Dorr – Oliver, Germany)
a) Press belt drive; b) Filter cloth aligning assembly; c) Press rolls; d) Press belt tensioning device; e) Vacuum drum filter

Belt discharge involves leading the filter cloth with the cake over small rollers (Fig. 52 B). This method can be used for continuous filtration of slurries containing fine solids that would generally block the filter cloth too quickly. Spraying the cloth from both sides makes it possible

to handle finely-dispersed suspensions without precoating.

When filtrates are volatile or tend to produce vapors, or when an improvement in filter capacity is desired (see page 48 and Fig. 58), it may be useful to enclose the filter and pressurize the housing slightly.

Expression devices have been added to drum filters (Fig. 53). Besides lowering the cake moisture by 20 – 200 %, compression also promotes cake removal. The rubber expressing belt is held against the drum with several pneumatically adjustable rollers; forces per unit length of up to some 250 N/cm are used. The belt path is such that the cake gradually enters the region of maximum pressure.

For the filtration of finely dispersed feeds, the drum filter can also be operated as a precoat filter (Chap. 11). Before filtration proper, a 5 – 6 cm layer of filter aid is applied from a slurry. It may be advantageous to smooth the precoat layer. During filtration, the filter aid must remain adhering to the filter cloth over the entire periphery of the drum, so as to maintain the suction in all the compartments of the drum. No air blast is used to help in cake removal. During filtration, the scraper is advanced 0.05 – 3 mm per drum revolution until a permanent precoat thickness of a few mm is reached.

If the suspension is one that settles rapidly (group F of Table 2), it can be supplied to the tank from a box located at the top of the drum (*top-feed filter*). This arrangement creates sealing problems with respect to the drum and the cake. In general, the lower tank is either eliminated or filled with wash liquor. The *inner-face vacuum drum filter* (Fig. 54) is also suitable for this type of feed; sedimentation and filtration in such a device are in the same direction, so that no stirring is needed. The filter medium is on the inside of the drum shell, while the compartments are on the outside. The drum is open at one end to allow slurry feeding and cake removal; the closed end bears the filtrate pipes and rotary valve. By virtue of gravity, the coarse fraction of solids is deposited first on the filter cloth, forming a more permeable ground layer, which separates more easily from the cloth.

Vacuum multi-compartment filters are made in standard series (about 30 sizes) with filtration areas up to 100 m^2 and drum diameters up to 4.2 m. The smaller sizes, roughly 0.25 m^2, do good service as laboratory filters. The filters are irreplaceable for many continuous filtration jobs, being distinguished by reliability, low operating costs, versatility and adaptability.

Rotation speeds depend on filter size and desired cake moisture content and range from 0.1 – 3 revolutions per minute.

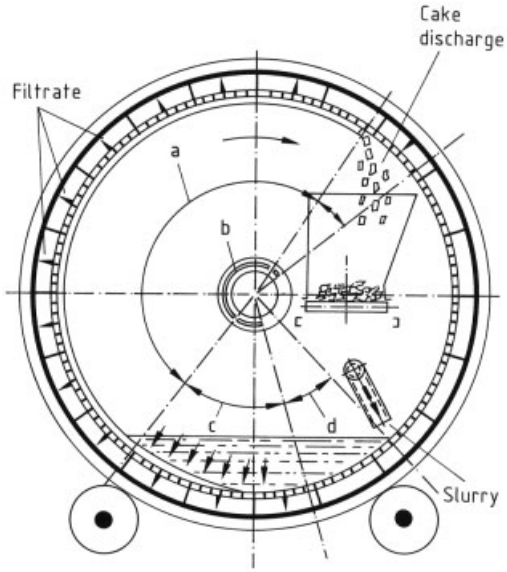

Figure 54. Inner face vacuum filter
a) Dewatering zone; b) Rotary valve; c) Settling zone; d) Suction zone

Sizing of Vacuum Drum Filters. The design of such filters is normally directed to the throughput of solids as the significant component of the feed not least of the fact that deliquoring and washing of the produced cake herein are more or less uncomplicated procedures.

Based on the filtration equation (Eq. 15) (Section 2.1.2) the material balance using mass ratios of the slurry, the cake and the filtrate and the substitution of the effective filtration time (s) by

$$t = 60\frac{a}{n} \tag{68}$$

where a is the submerged to the total drum area and n (min^{-1}) the number of drum revolutions, lead to the available throughput of dry solids (kg) per unit filter area (m^2) and unit time (h)

$$(M_S)_A = C\sqrt{\frac{aKn\Delta p}{\alpha_m \eta}} \tag{69}$$

here $(M_S)_A$ is the specific mass of solids ($kg\,m^{-2}\,h^{-1}$), $C = 3600\sqrt{\frac{60}{\pi}}$ a conversion fac-

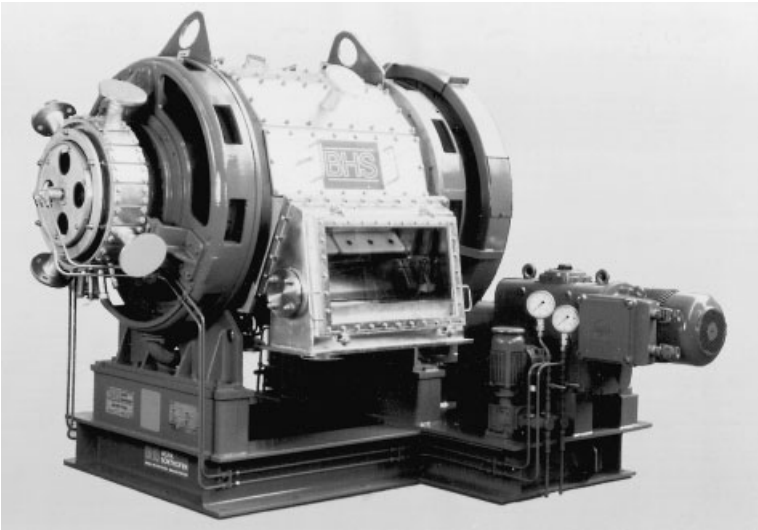

Figure 55. Pressure drum filter (Fest-Filter for antibiotics production, 1 m^2 filtration area (reproduced with permission of BHS Sonthofen)

tor ($s^{1/2}\,min^{-1/2}\,h^{-1}$), $K = \varrho_1 c_M \left(1 - \frac{c_M}{c_{Mc}}\right)$ a concentration constant (kg/m^3), ϱ_1 the density of the liquid (kg/m^3), c_M and c_{Mc} the mass ratios of solids in the slurry or in the cake after the filtration cycle, α_m the average specific resistance of the cake (m/kg), η the viscosity of the liquid (Pa s) and Δp the pressure drop through the cake (Pa). The specific flow resistance of the filter medium is disregarded because of the comparatively low influence.

The equation gives a parabolic increase in cake production with increasing rotation speed; in other words, the maximum solids output corresponds to the highest possible rotation speed. On the other hand, the cake thickness decreases exponentially thereby. Since this thickness is limited by the cake discharge device, the maximum solids production is also limited. The limit further depends on the material values contained in Equation (69).

Table 3 presents some examples illustrating the performance range of the vacuum drum filter.

Pressure Drum Filters. To realize the advantages of continuous drum operation in pressure filtration, with its higher throughputs, and to make the filtration of volatile liquids more economical, FEST developed a drum filter with a pressure housing around the drum (Fig. 55). The drum turns at a continuously variable speed. The annular space between drum and housing is sealed with stuffing boxes and partitioned into four or five chambers. The drum shell consists of compartments (Figs. 56 and 57), which are connected to a rotary valve by filtrate pipes. Slurry under pressure is continuously pumped into the filtration space. The cake builds up in the filter compartments; as the drum rotates, the cake is mechanically compacted by passing through a wedge, then moves on to the washing chamber. If no provision is made for multistage washing, the next step is dewatering by blowing in a separate chamber, which is followed by the cake removal section. In the unpressurized region, the cake is removed by back-blowing and pivoting scrapers. Washing of the filter cloth can also take place in this zone.

The advantages are: continuous operation, relatively high throughputs, filtrations up to near the boiling point of the liquid, no vapor losses, very good washing, countercurrent washing when multistage arrangements are used, low cake moisture, no cracking before washing, and mechanical compaction. The filter is naturally more expensive than a conventional rotary drum filter, so that additional filtration costs are calculated as some 50 %. It finds use mainly in the chemical industry (filtration of dyes, vitamins, inorganic and organic chemicals, pesticides, insecticides), in the production of pharmaceuticals

Table 3. Performance of vacuum drum filters (Dorr–Oliver).

Material	Feed concentration, % solids	Filtrate rate, kg $m^{-2}h^{-1}$	Cake moisture, %	Discharge
Calcium carbonate	20	34 – 50	40 – 50	scraper
Coal (flotation concentrate)	25	360 – 500	25	scraper
Coal (refuse)	30	120 – 145	25 – 30	scraper
Blast furnace flue dust	45	45	25 – 30	scraper
Iron oxide (pigment red)	13	17	30	roll
Corn starch	12 – 28	130 – 165	30 – 40	belt
Wine lees	low	82 – 820 (L/m^2h)	–	precoat

(antibiotics), and in vegetable processing (extraction). Table 4 gives some throughput figures, which show substantial gains in some cases. Because cleaning is more complicate, it is recommended that the product not be changed.

Table 4. Some throughputs of pressure drum filters, kg m^{-2} h^{-1}, based on dry cake (BHS, Sonthofen, Germany)

	Throughput	Cake moisture, %
Antibiotics	385	14
Copper sulfate	525	24
Sodium hydrosulfite	700	20
Zinc stearate	45 – 160	60 – 70
PVC (suspension)	420	51
Titanium dioxide	150	45

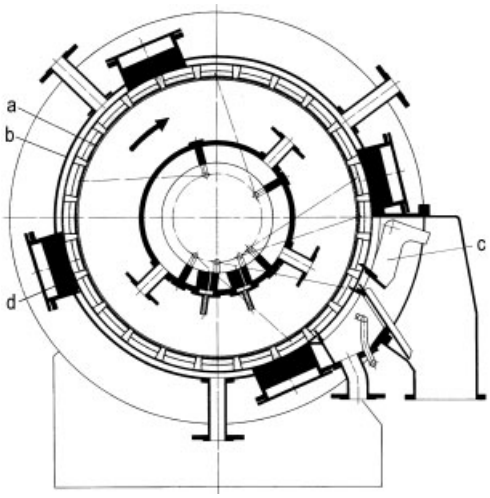

Figure 56. Sectional view of Fest-Filter (reproduced with permission of BHS Sonthofen)
1) Filter drum; 2) Pressure tight housing; 3) Scraper; 4) Partitioning element pressed by compressed air

Increased capacity sometimes results from the use of pressure drum filters created by housing conventional vacuum drum filters in pressure tanks (Fig. 58). These devices have long been known and have recently come to be called *hyperbaric filters*. Many such designs provide better dewatering than the straight vacuum filter.

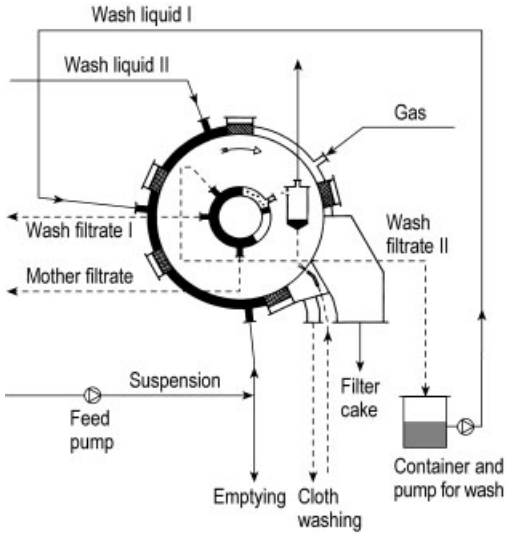

Figure 57. Schematic operation of Fest-Filter

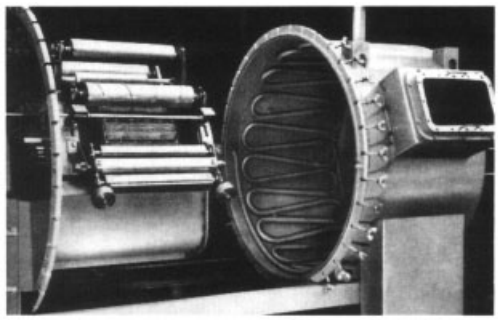

Figure 58. Pressurized drum filter – Hyperbar filter

8.7. Leaf and Plate Filters

The element in a leaf or plate filter is essentially a leaf-shaped structure covered with filter medium on one or both sides. Slurry is supplied to the leaf from the outside, and the filtrate can drain through the void inside, while the cake collects on the outside. Filter elements are usually assembled into units of varying size. They can be operated as open-tank devices, immersed in the feed, with the filtrate being withdrawn by suction (open-tank leaf filter, Moore filter). Because the cake often drops spontaneously into the feed tank, these devices are also used as filter thickeners. Alternatively, the filter elements are enclosed in pressure vessels (which may be heatable) and operated as pressure filters, especially for influents of relatively low concentration (clarifying filtration). They are also well-suited to precoat filtration.

The shape and arrangement of the filter leaves or plates are adapted to the properties of the cake and the filter operating requirements. If the cake is easily detached from the filter medium, a vertical leaf configuration can be chosen, so that the cake falls off the medium by gravity (removal can be promoted by an air blast or vibrating). Horizontal leaves are suitable for poorly-adhering filter cakes.

Possible element shapes are innumerable, from simple perforated and mesh constructions to multi-layer designs with internal stiffening. The design of Figure 59 is typical for bed elements which are preferred because of their pressure resistance. Each element consists of five fabric plies with graded mesh openings. The core ply, which allows the filtrate to drain, bears a supporting fabric on either side, with the filter cloths proper on the outside.

In another type, the effective filtration area in each element is made up of a base with perforated plates, which is inserted in a massive frame to withstand greater mechanical loads. The filter cloth or a wire mesh is placed over the base.

For clarifying filtration with filter beds, woven fabrics, filter paper or membranes, the filter media are placed right on the horizontal plate structure. The stack of filter elements to retain solids down to some micrometers is placed about a cylindrical pipe enclosed within a pressure-tight vessel (see Figs. 60 and 61).

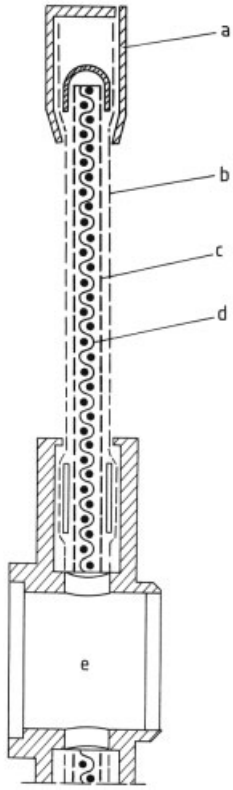

Figure 59. Wire filter leaf
a) Binding; b) Filter medium; c) Metal mesh; d) Wire mesh support (Chamber screen); e) Filtrate manifold

Leaf filters are made in a smooth design with up to 60 m^2 of filtration area. In large plant units, the leaf assemblies can be moved mechanically or hydraulically (Fig. 62), or the housings are designed for easy cake loosening and removal. Because the pressure on the cake is relieved when the tank is opened, a vacuum is often pulled on the filtrate pipes at this time so that the cake will not drop off the leaves until the desired moment. In the Kelly filter, vertical filter leaves were arranged axially in the cylindrical pressure housing, so that the leaf size decreases to either side. The leaves are vertical in smaller units or horizontal in larger ones (up to 140 m^2). The cylindrical housing of the Sweetland filter is divided; the lower half can be pivoted to the side to allow cake removal by hosing off.

Where process conditions do not allow the slurry heel (slurry remaining in the pressure housing after filtration is complete) to be recycled to the filter feed or otherwise utilized, some

designs have one or more filter leaves in the bottom of the pressure space to filter this residual feed. These leaves do not function during the main part of the cycle. The heel is usually driven through those filter leaves by pressurized gases (compressed air).

Figure 60. Horizontal plate filter (deep-bed filter) (reproduced with permission of Seitz-Filter-Werke)

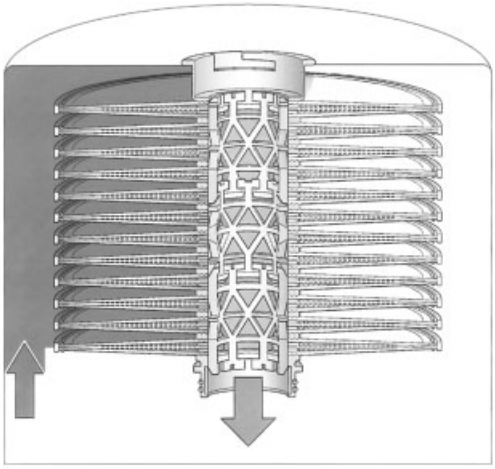

Figure 61. Sectional view of horizontal plate filter

Many designs address the problem of cake removal. Modern filtration systems are generally provided only with completely automatic operation; with vertical leaves as shown in Figure 63, automatic removal is often a simple step.

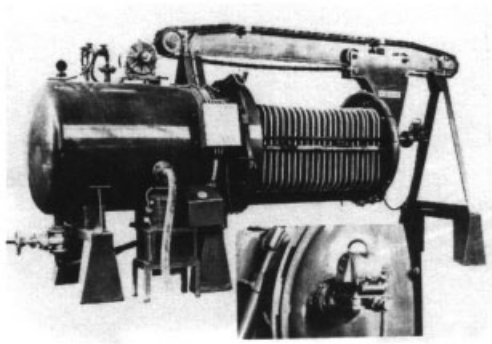

Figure 62. Horizontal-tank pressure leaf filter (Niagara)

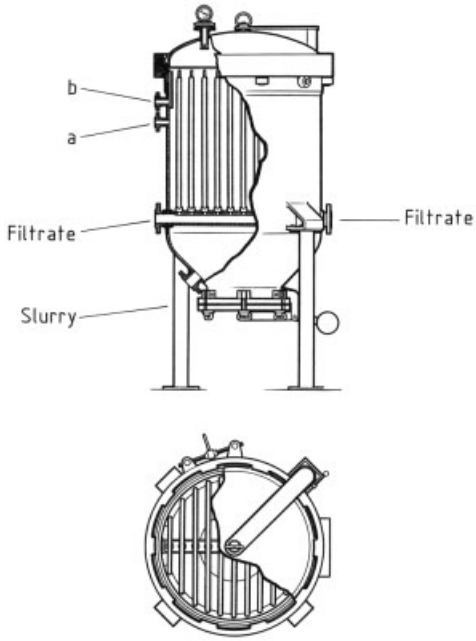

Figure 63. Vertical pressure leaf filter
a) Valve security; b) Overflow

In *horizontal* systems, flowable and nonflowable cakes must be distinguished. Flowable cakes are the rule in precoat filtration for clarification; here the cake is commonly discarded as waste. The round filter leaves or plates, with slurry feed from the top, are arranged on a hollow shaft in a pressure vessel and positioned a few centimeters apart with spacing disks (Fig. 64). Cake thickness is controlled through the filtration pressure or with special sensors. Cakes in these devices are nearly always so loose that they

can be discharged by briefly spinning the stack of plates; separation can be helped by backwashing if necessary. Cake removal, required when leaves cease to function, is effected by vertical vibration in addition to rotation. Filters of this type are also called *centrifugal-discharge filters* up to 170 m^2. They are often classified as plate filters.

A further variation in operation of pressure plate filters is shown in Figure 64. If the cake does not flow off the rotating plates, it is discharged by fixed scrapers during cake formation on the rotating plates.

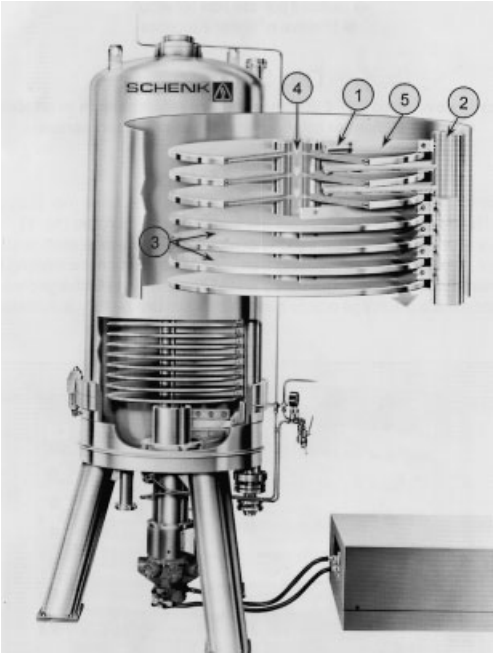

Figure 64. Continuous pressure plate-filter (reproduced with permission of Schenk Filterbau)
1) Feed; 2) Scraper; 3) Plates; 4) Filtrate discharge; 5) Cake discharge

To empty the filter completely before removing the cake, most filters of this group have leaves at the bottom of the stack (scavenger filtration elements) to allow separate filtration of the slurry heel.

For relatively dry filter cakes, a conical discharge is preferred (Fig. 65, left). The stack of plates in such a device must be driven from the top of the tank, which, in addition facilitates the sealing of the driving shaft.

Nonflowable cakes are generally obtained in cake filtration where filter aids must be avoided. Help in these instances comes only from mechanical cleaning devices. Washing in these pressure filters usually involves some difficulty, because the slurry heel must first be removed and the housing filled with wash liquor. Special devices have been developed to reduce the consumption of wash liquor.

In the pressure leaf filters described above, the feed is pumped into the housing, and its distribution over the leaves is naturally somewhat random. The slurry can also be supplied to the leaves through a hollow center shaft, with the filtrate draining into the housing. In this way, uniform loading of the filter elements is ensured.

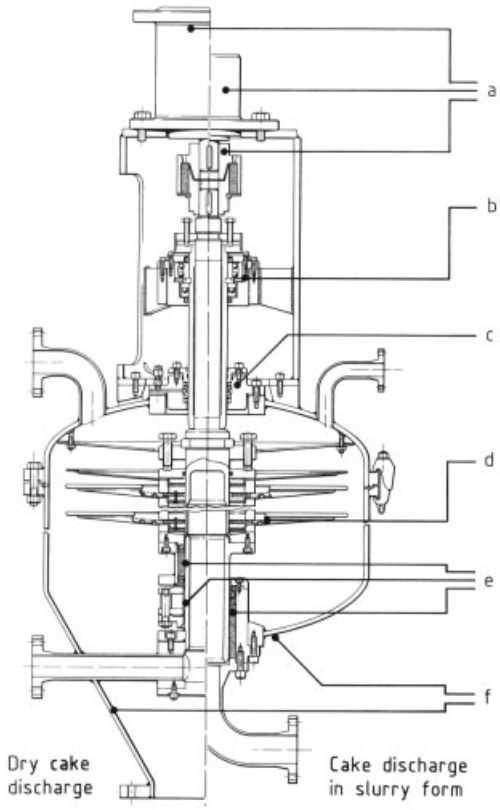

Figure 65. Centrifugal-discharge pressure plate filter (STAWAG BIOTECH/Switzerland)
a) Filter drive; b) Upper bearing; c) Upper seal; d) Filter plates; e) Lower bearing and seal assembly; f) Filter vessel

Range of Application. The applications of pressure leaf filters are extraordinarily broad,

and extension to special filtration problems is relatively easy. As cake filters, these devices are employed for filtering all kinds of chemical products, such as glycerol, caustic soda solution, phosphoric acid, or pharmaceuticals. Run times between cleaning periods are several hours, and throughputs of $1-4\,m^3\,m^{-2}\,h^{-1}$ (up to $6\,m^3\,m^{-2}\,h^{-1}$ in extreme cases) have been achieved. Pressure leaf filters are applied in group S and D of Table 2.

When used as precoat filters, these units are especially well-suited for the collection of relatively fine-grained solids from low-concentration feeds (beverages, cane syrup, or water). Average throughputs are $0.3-1\,m^3\,m^{-2}\,h^{-1}$. Run times are strongly dependent on the level of impurities to be removed; in mixed and thick juices in sugar refining, runs of 7 – 16 h (depending on, e.g., the quantity of diatomaceous earth used as filter aid) are achieved.

Centrifugal-discharge filters are available with filtration areas of $1-60\,m^2$ (up to $170\,m^2$ from one manufacturer); the leaves are 420 – 1500 mm in diameter.

8.8. Nutsche (Pan) Filters

The simplest type of nutsche filter is an open tank or box with a porous bottom, into which the slurry is charged or flows continuously. Filters of this kind are common everywhere, especially in laboratories. They do not impose stringent requirements on suspension properties, and they can function even without difficulty with highly-concentrated or rapidly-settling slurries. As a rule, they are operated manually. For poorly filterable feeds, suction can be applied to the filtrate side.

Nutsches are widely applied in the chemical industry because all the steps in filtration, including washing, dewatering, and drying if practiced (with hot air), can be performed in a clean, distinct manner. There has accordingly been no lack of efforts to automate the operation of nutsches or make it continuous.

In order to increase filtrate flow rates, an enclosed nutsche can also be pressurized with air or inert gas. Pressure nutsches are especially good with volatile solvents. They can be drained mechanically through a retractable bottom. Several process steps can be carried out with no change of apparatus when nutsches in process service are fitted with a variety of accessories (e.g., agitator, mixer, heating and cooling devices); in this way operations such as mixing, stirring, reacting, crystallization filtration, extraction and drying can be performed in sequence. Low equipment costs, non-polluting operation, and optimal techniques make these devices suitable for small product batches. Nutsches can be placed on pivots and repositioned for the various process steps (Fig. 66).

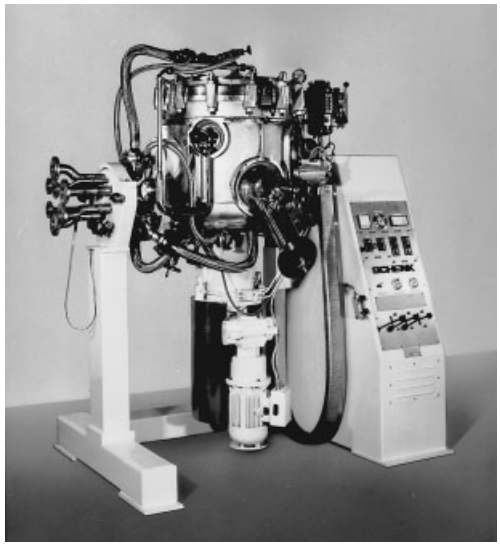

Figure 66. Pressure nutsche filter – multi purpose filter – in the crystallization step (reproduced with permission of Schenk Filterbau)

Nutsches are good candidates for filtration with relatively thick cakes. To keep cracks from appearing, some filters have smoothing attachments, which can also provide advantageous mixing of wash liquor and cake and can assist in cake discharge.

Tanks with internals such as agitators and solids-discharge devices (plough) are more versatile nutsches. These accessories can homogenize the filter cake and render it especially washable (Fig. 67).

In some advanced developments, intended for comparatively high throughputs, a number of individual nutsches circulate on either a belt or a carousel. (A system of the first type is usually classed as a belt filter.) This kind of nutsche has proved useful especially in phosphoric acid

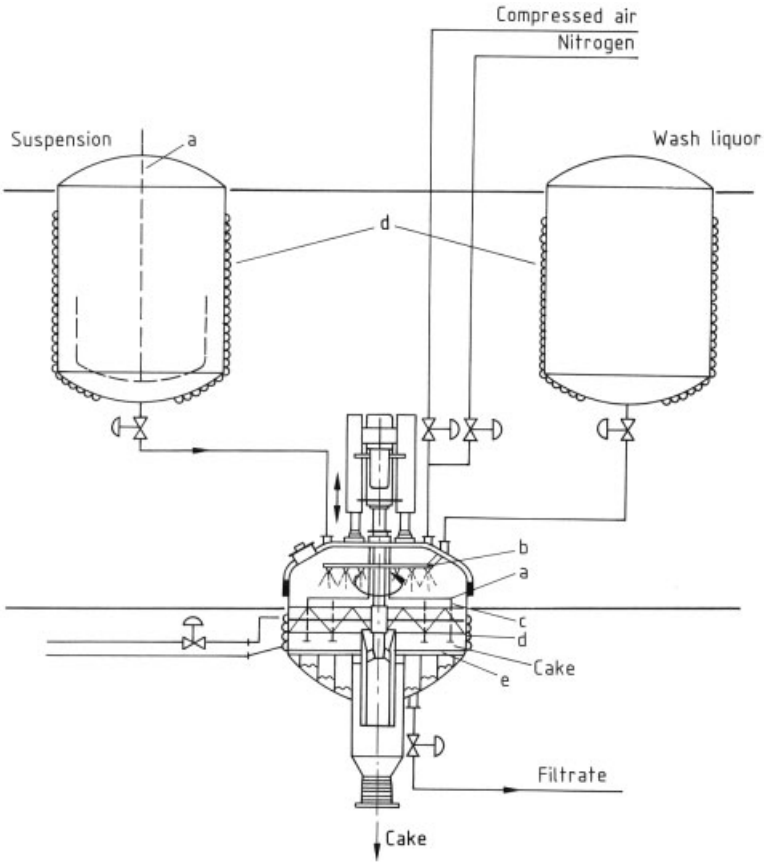

Figure 67. Pressure Nutsche filter
a) Agitator; b) Wash device; c) Screw conveyer; d) Heating pipes; e) Filter plate

manufacturing, where it is employed to remove gypsum in the sulfuric acid digestion of phosphate rocks. Careful countercurrent washing is practiced in order to attain a high degree of separation. Equipment sizes range up to a filtration area of some 200 m^2.

Automated nutsches share the advantages stated earlier: They are insensitive to slurry properties and thus suitable for groups F, M, and S in Table 2, and they provide a clear separation of the process steps and thus insure a uniform, reliable filtrate composition.

For the filtration of slimy feeds, it is advantageous to keep the filter cake in motion, for example by mechanical vibration or an appropriate slurry flow, so that the cake will remain permeable for a longer time.

8.9. Pressure Plate Filters

The filter press is one the simplest of all pressure filters; it allows relatively easy cake removal after the individual filtration steps. It consists of a number of flush plates, up to 100 or more, and frames whose width (between 25 and 150 mm) can be matched to the service conditions. The plates and frames form rather flat pressure filtration spaces with a relatively large filtration area (*plate-and-frame filter press*, Fig. 68 A). During filtration, the plates and frames must be squeezed by mechanical or hydraulic means in order to seal the filtration spaces (Fig. 69). The faces of the plates are studded or grooved to permit filtrate drainage.

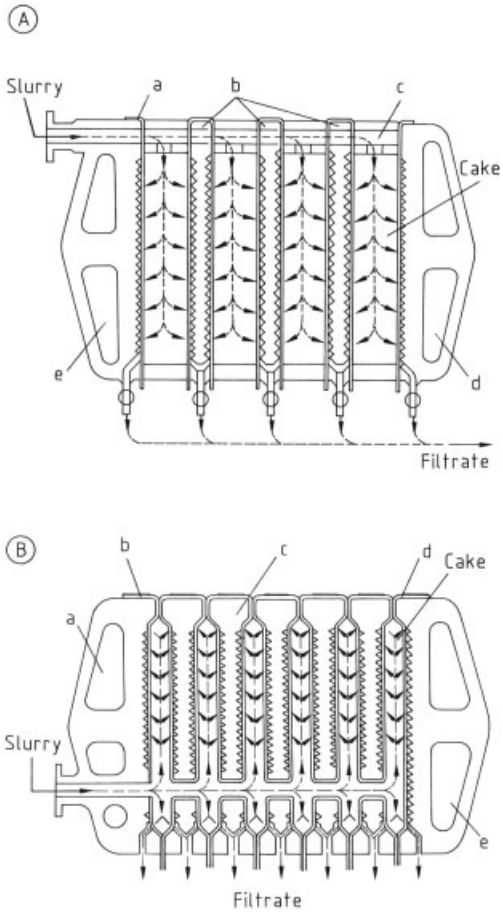

Figure 68. Filter press – sectional view
A) Plate-and-frame filter press
a) Filter cloth; b) Filter plates; c) Hollow frames; d) Movable end half plate; e) Fixed end half plate
B) Recessed-plate filter press
a) Fixed end half plate; b) Filter cloth fastening; c) Recessed filter plates; d) Filter cloth; e) Movable end half plate

The plates can also be made with a raised edge (*recessed-plate filter press*, Fig. 68 B). There are several essential differences between the two types: In the plate-and-frame press, slurry enters through channels in the corners of the plates and frames. The filter cloths can therefore be shaped as flat strips and pulled over the plates; this arrangement is advantageous when the cloths are changed. In recessed-plate presses, slurry is introduced into a center orifice of the plates. The filter cloths must be made in two parts and inserted in the slurry feed channels; this arrangement involves much more effort. Further, it is harder to vary the cake thickness in recessed-plate devices than in plate-and-frame presses, where all that is necessary is to replace the frames. The recessed-plate press has, however, the advantage that only about half of the filter elements need be moved to open the press; the equipment also costs much less for a given filtration area. Because of the deflection of the filter cloths at the edge of the recessed plate, thinner cakes (up to about 35 mm) are generally produced than in plate-and-frame devices. With either type of press, it must be noted that capillary flow takes place despite the relatively high density of the filter cloths, and this flow – at least in filtration with solvents – causes the undesirable escape of filtrate through the seals between plates.

The filtrate is either discharged separately from each plate through cocks (open discharge) or collected in a channel (closed discharge), which is preferred when toxic or volatile materials are handled.

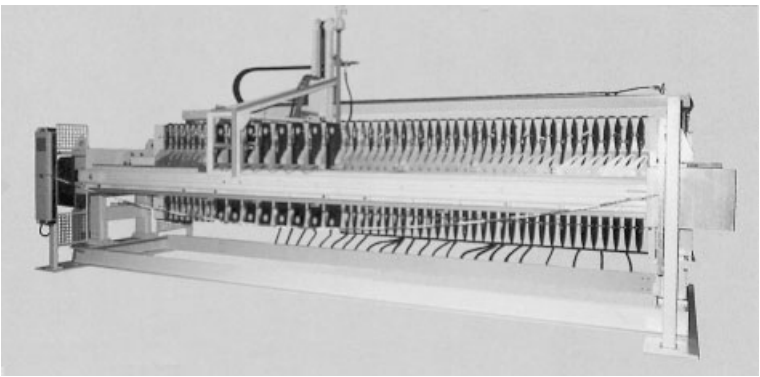

Figure 69. Filter press (reproduced with permission of Eberhard Hoesch & Söhne)

At cake washing time, valves are changed over and special lines carry wash water only to every other plate, so that the liquor can flow transversely through the cake to the opposite plate. Occasionally the wash water is also delivered through the slurry channel; this arrangement is, of course, not as effective. After washing, the cake can be dewatered by passing steam or (possibly heated) air through it.

For cake discharge, the press is opened and the plates are moved, one by one, through a predetermined distance. A variety of mechanical devices are now available for this purpose. If the cake is to drop out by itself, it should not adhere too tightly to the filter cloth and should not be too densely filtered. Cake separation can be assisted by rapping, pulling out the filter cloth, or other means. An automatic filter press for high throughputs has more than 100 chambers with a filtration area of 1000 m^2 and more.

Filter presses are generally insensitive to slurry filtration properties, with perhaps one exception: In large presses with long slurry channels, varying sedimentation may cause a slurry of group S (see Table 2) to fill the filtration spaces unevenly. Under some conditions, an undesired classification occurs, with a detrimental effect on washing, and may even lead to pressure differences between chambers, causing the plates to deform or break. In such cases, it is desirable to provide a hole for pressure equalization or a supporting button.

In addition to filter presses with vertical plates, some designs also feature horizontal plates.

Filtration Cycle (see Chap. 5). Cake filtration in filter presses involves two indistinctly separated stages. At the start of filtration there is no cake, and the hydraulic resistance is therefore very low, so that the filtration pressure cannot rise very high at this point. The process is similar to constant-volume filtration (see Section 2.1.2). The pressure increases as the cake thickness grows, until the rated pressure (of the pump) is reached. Constant-pressure filtration then begins, with a declining quantity of filtrate. The exact sequence of filtration in filter presses is still unclear and cannot be calculated, because not enough research has been done, but a rough estimate suggests that the highest throughput is achieved if

$$t_f = t_k Z \tag{70}$$

where t_f is the filtration time, t_k is the time for filling and emptying per chamber, and Z is the number of chambers.

On the basis of a cost calculation, the maximum number of chambers is

$$Z = \left[\frac{t_f\,(C_M - C_p)}{t_k C_p}\right]^{\frac{1}{2}} \tag{71}$$

where C_M is the cost of mechanical equipment (fixed and movable end plates, adjusting device) and C_p is the cost per plate (chamber). For conventional designs, Z is found to be around 60.

To determine t_f, the total cycle time including washing and other steps, experimental data must be used. Unfortunately, experience has shown that measurements in laboratory-scale filter presses cannot generally be extended to arbitrary sizes, since there may be discrepancies in the filling of the filtration chambers and thus in the pressure relationships.

Filter presses are commonly charged at pressures of up to 2 MPa, more in special cases. Centrifugal or diaphragm pumps are preferred for the cake filtration of highly-concentrated slurries. The possibility of employing relatively high filtration pressures, along with the mechanical advantages in charging and discharging, lend filter presses some advantages over other types of pressure filter, even though these devices are always operated batchwise:

1) Insensitivity to feed fluctuations
2) Relatively high throughputs, even with difficultly filterable slurries, so that far smaller amounts of filter aids are needed
3) Low cake moisture
4) Denser cakes that are easier to handle
5) Clear filtrate
6) Good separation of different filtrates (wash liquor, etc.)
7) Arbitrary filtration steps (washing, steam or air blowing, etc.)
8) Low maintenance costs, few spare parts, low depreciation
9) Simple operation
10) Comparable equipment costs per square meter of filtration area

The drawbacks are obvious: Batch operation; personnel requirements in cases where cake does

not drop off by itself; relatively long cleaning times.

Applications. Even though they are batch devices, the good washing and low cake moisture available in filter presses have kept them in wide use. Contributing factors, along with comparatively long service times and good reliability, include mechanical plate movement, used almost exclusively for the opening and discharge of large units; automatic filter-cloth cleaning with high-pressure water sprays; and the availability of special plate designs. Plates up to 2.6 m on a side give far more than 1000 m^2 of filtration area in one press. The design of the plates and the use of polypropylene as the main plate material have remedied many of the shortcomings of filter presses, such as plate fractures, corrosion, and plate mass.

Filter presses can be employed almost anywhere, even if the mode of operation and the equipment must be adapted to the feeds in question. Filter presses are used successfully in either cake or clarifying filtration, depending on the concentration and properties of the solids; in the latter case, filter aids are usually added, or filter beds are employed instead of cloth media.

In the production of chemicals (e.g., dyes), ceramics and raw materials, and in wastewater treatment, nuclear technology and other fields, cake filtration is applied chiefly to groups M and S of Table 2. Slurry feed rates in these fields depend on the widely varying slurry composition and range from 0.2 – 1 $m^3\ m^{-2}\ h^{-1}$ on the average.

Figure 70. Filter press for sheet filtration – deep bed pressure filter (reproduced with permission of Seitz-Filter-Werke)

Clarification is carried out in *sheet filter presses*, which resemble plate-and-frame filter presses in design (Figs 70, 71, and 72). Because of the very much lower solids concentration, service times are much longer (several hours). For this reason, automation is commonly dispensed with. Filtersheets include ordinary precoats and special papers (see page 64 and Chap. 11).

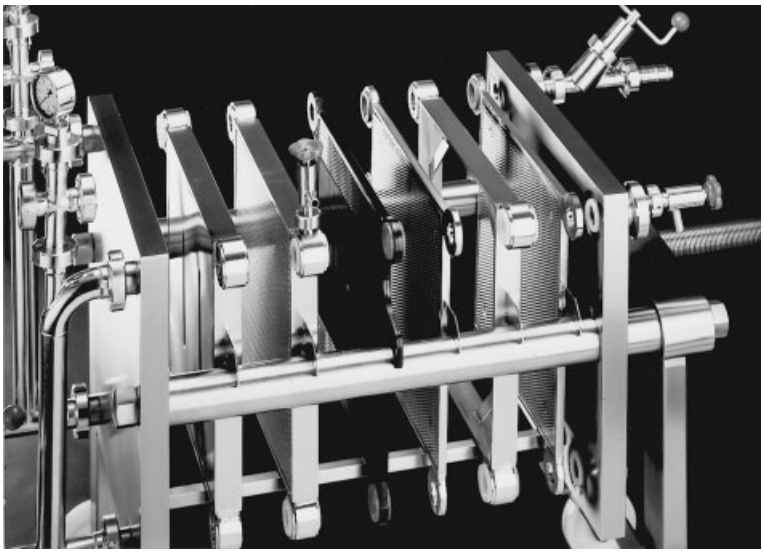

Figure 71. Sheet filter press opened (reproduced with permission of Seitz-Filter-Werke)

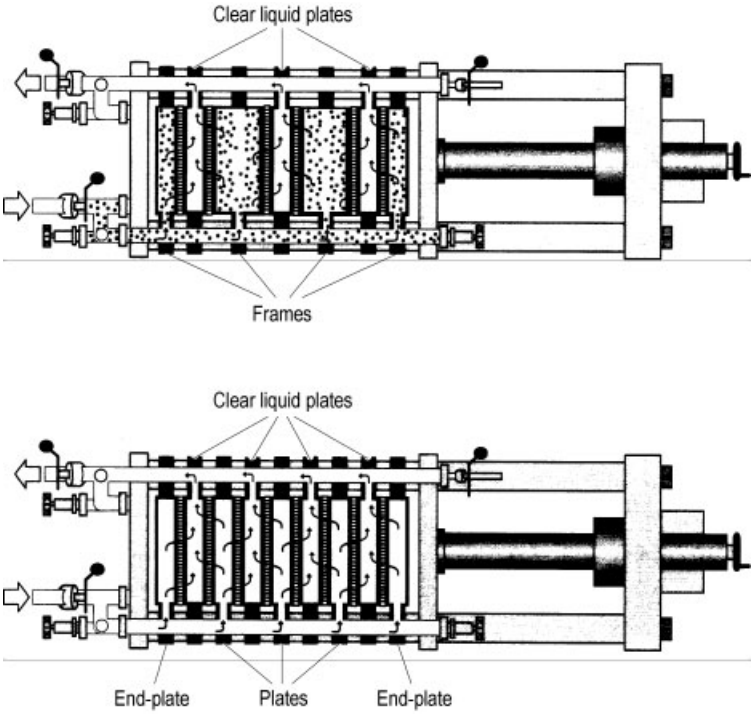

Figure 72. Clarifying filter press – schematic operation
Top: Filtration with filter aids; Bottom: Sheet filtration

The most important applications of clarifying filtration include beer, wort, wine, fruit juices, pharmaceutical liquids, or water sterilization (group D of Table 2). Here too, filter loading varies widely, depending both on the properties of the solids being removed and their concentration in the feed and on the nature of the liquid. They range from $0.05\,m^3\,m^{-2}\,h^{-1}$ for viscous varnishes up to $10\,m^3\,m^{-2}\,h^{-1}$ for dilute suspensions, such as beverages.

Membrane Filter Presses. In a membrane press, the filter plate is coated with an elastic material (rubber). After filtration, the membranes are pressurized so that they mechanically compress the filter cake. In this way the cake moisture is reduced by ca. 1 % – 20 %, depending on the cake compressibility. Figure 73 shows an example of a recessed-plate membrane press. A similar expedient is also possible in a plate-and-frame press. The pressing either shortens filtration time or the filtration pressure can be made smaller, which is an eminent practical advantage.

Membrane filter presses can be operated at up to 2 MPa; they produce a uniformly dewatered cake.

Plate Press with Belt Discharge (Automatic Press). This type of filter press, originally developed in Russia, combines positive automatic cake discharge with cake compression. The horizontal filter plates are stacked and, with the filter frames, combined into a three-chamber system: filtrate chamber, cake chamber, and high-pressure water space. The water space is separated from the cake by a membrane. For cake discharge, a cloth belt on the frame is advanced until the cake from each plate is completely stripped off as the belt is deflected into the next lower plate (Figs. 74 and 75).

Filter Press with Stationary and Rotating Plates. This filter contains alternating stationary and rotating plates on a shaft (Fig. 76). A velocity difference is produced in the cake between the plates, leading to breakdown of the cake. A thin layer of filter cake, whose thick-

ness depends on the filter medium employed, is left on the filtration surface. In the extreme case where cake formation is totally prevented, this is pure cross-flow filtration (see Section 2.4). The intermediate states are referred to as dynamic filtration.

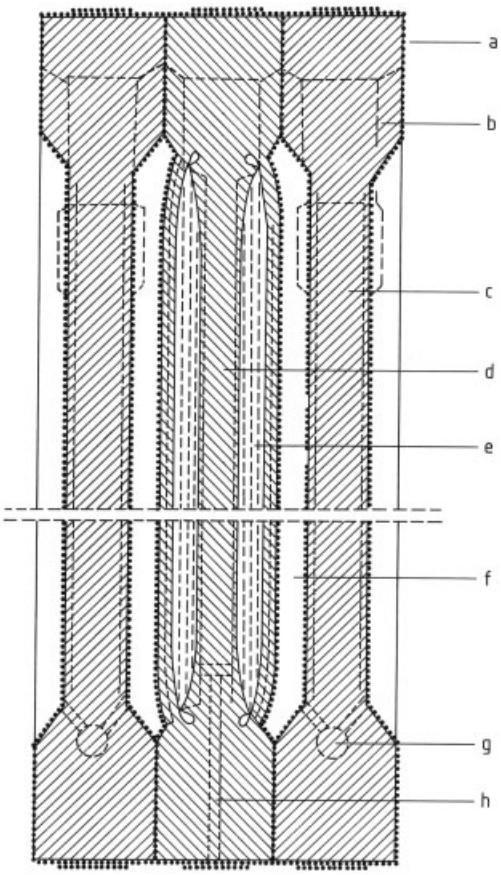

Figure 73. Membrane filter press plate assembly
a) Filter cloth; b) Slurry corner feed; c) Recessed plate; d) Membrane recessed plate; e) Press medium; f) Cake; g) Filtrate; h) Air or water inlet (Pressuring membranes)

If the filter cake is thixotropic, the shearing makes it flowable, so that it can be conveyed by pressure through the system of plates. According to studies of some dye suspensions, this filter can collect the same quantity of solids as a conventional filter press with roughly 20 times as large a filtration area.

8.10. Tubular Filters

Tubes as filter elements are similar in structure to cartridges. This section deals only with tubular filters whose dimensions do not usually correspond to the concept of a filter cartridge (diameter either too large or too small). This group includes, on the one hand, tubular pressure filters and, on the other, tubular membranes used in cross-flow filtration.

Tube Filters. The tubular pressure filter – the *tube press* – developed in England, consists of a pressure tube with a filter tube inserted in it. After cake filtration in the annular space, a membrane placed inside the pressure tube compresses the cake. A pressure of up to 10 MPa can be applied. Next (or immediately after filtration), the cake can be washed. For cake discharge, the filter tube is swung downward and the cake is loosened from the filtration surface with compressed air. This filter has the advantage of the cylindrical filtration chamber, which is economical for relatively high pressures.

In another type of tubular pressure filter, filtration and compaction are similar but the tube is positioned horizontally, so that the cake can be scraped out of the filter tube.

A third type of tubular pressure filter has a cylindrical rotor inside the filter tube. Slurry is continuously compacted in the annular space. The solids are retained in the annular space on the filtration surfaces of the tube and the rotor, where they are subjected to shear. In this way, the cake is caused to move downward to the discharge. As in the filter press with rotating plates (see preceding section), the filter cake must remain flowable after substantial dewatering; that is, it must be thixotropic. In one or more washing zones, wash liquor can be fed radially through the filter tube. Several filter sizes with filtration areas from 0.25 – 1.6 m^2 are tested.

Cross-Flow Tubular Filters. The filter elements in the tubular cross-flow filter are small tubes or hoses, a few millimeters in diameter, which are bundled into systems of tubes carrying parallel flows (Fig. 77). The filtrate passes the system through walls of the tubes whereas the suspension flows tangentially to them dragging the retained particles, see Chapter 2.4. The

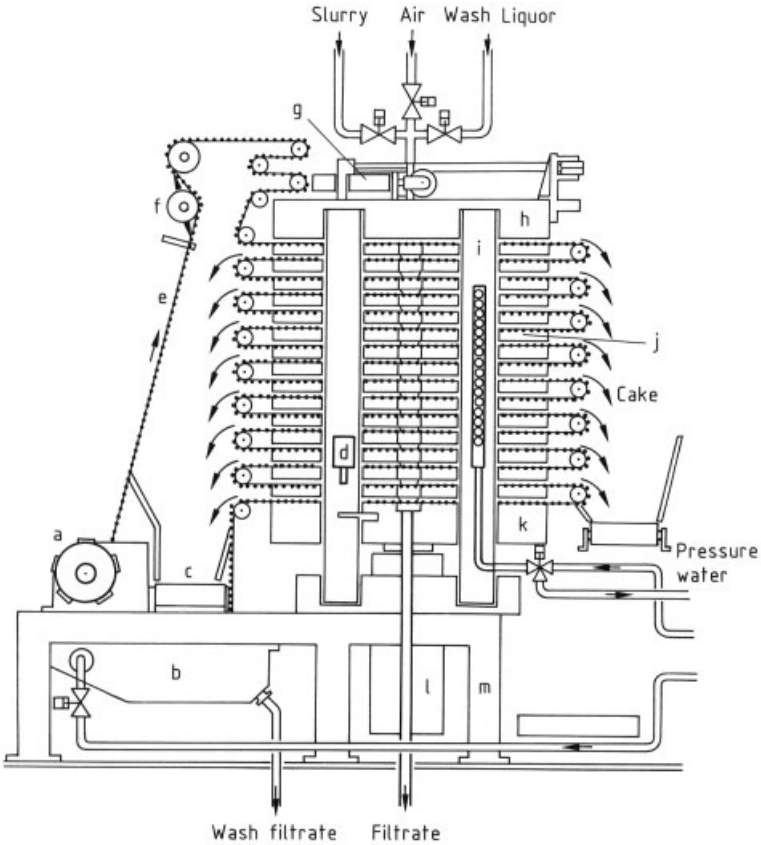

Figure 74. Automatic filter press – systemic processing
a) Cloth drive; b) Cloth wash; c) Cake removal; d) Pressure control; e) Filter cloth; f) Filter cloth aligning assembly; g) Tensioning device; h) Yoke plate; i) Yoke anchor; j) Filter plate with inflatable membrane; k) Lifting table; l) Hydraulic closure device; m) Filter preframe

tubes can be made of various porous materials, preferably polymers or ceramics, and have quite narrow pore-size ranges, from a few micrometers down to hundredths of a micrometer. They are, above all, microfilters for the concentration of dilute suspensions and for clarifying and sterilizing filtration. Such filter systems with the finest membrane pore sizes are also suitable for ultrafiltration and reverse osmosis; multilayer or asymmetric membranes are used in them (→ Membranes and Membrane Separation Processes).

Throughput rates range up to $3000\,\mathrm{L\,m^{-2}\,h^{-1}}$, depending on pore size and filtration pressure. The possible applications vary with capacity. The principal users are chemical, pharmaceutical and foodstuffs producers and, increasingly, bioengineering concerns.

Figure 78 shows an example of a complete microfiltration unit including cross-flow elements and the programmed control unit for batch and continous operation.

8.11. Special Filter Types

Coarsely-dispersed systems with relatively high solids contents are dewatered with conventional screens like the vibrating screens used for classification. One such filter is the vertical sieve containing a *bent screen* plate, which is employed to separate coarse and fibrous solids at lower concentrations from process waters and wastewaters with screen filtration. The sieve, which is equipped with an adjustable screen slope, is charged from the top, so that the slurry flows

Figure 75. Automatic filter press (reproduced with permission of Eberhard Hoesch & Söhne)

downward by gravity. The screen, with slots varying from 0.125 – 2.5 mm wide, retains the solids, which drop downward. The simple, inexpensive design has made this device a widely-used one in food preparation, the textile and manmade fiber industry, paper and pulp manufacturing, and the chemical industry.

Several new filtration mechanisms are being explored with an eye to filtration problems that have not yet been solved. These include, in particular, the dewatering of suspensions having relatively high concentrations of extremely fine solids (below 10 µm down to colloidal sizes). The conventional separation techniques – flocculation and microfiltration – generally yield filter cakes or concentrates that require considerable energy for complete phase separation (e.g., by drying). Research on ionized phases has shown that the electrokinetic effects of *electro-osmosis* for liquid transport and *electrophoresis* for solids transport along an electric field allow separation, even of difficultly filterable suspensions. Both these effects have been known for a long time.

A filter based on the above processes consists of separating chambers with anodes, where the ionized solids collect, and cathodes made as filter media, where the electrolytic filtrate is discharged [45]. Several separating chambers form filter systems similar to filter presses.

Difficultly filterable suspensions are also filtered by *high-gradient magnetic separation* (HGMS) (→ Magnetic Separation). This type of filter employs a steel wool medium. In a very strong magnetic field, high field gradients are generated at the individual filaments. These gradients attract fine magnetic particles. This HGMS filter thus resembles a depth filter. Successes have been achieved in the purification of kaolin and wastewaters.

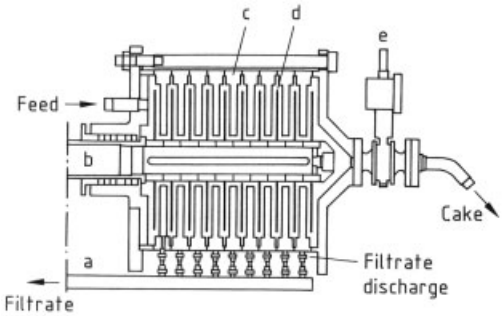

Figure 76. The dynamic filter press
a) Filtrate discharge pipe; b) Main shaft; c) Filter plate; d) Agitating disk; e) Cake discharge valve

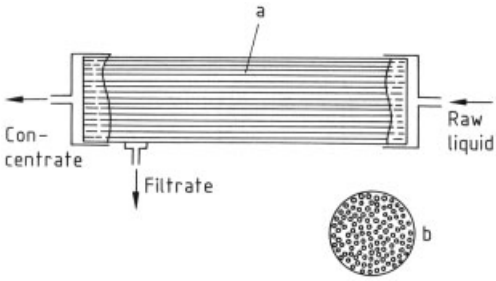

Figure 77. Schematic of tubular membrane filter
a) Tube bundle; b) Cross section

Figure 78. Microfiltration equipment (Seitz-Microflow, reproduced with permission of Seitz-Filter-Werke)

9. Filter Selection

Because certain qualities, largely related to design, characterize the ranges of application of filters, some guidelines can be set forth. These together with a few other points can provide a basis for filter selection. In general, a good selection takes account of factors such as slurry properties, throughput, or cake formation rate (see Table 2). Other points to be considered are the density and viscosity of the liquid, the size, size distribution and shape of the solids particles, the sensitivity of the product to impurities, corrosion problems, the sensitivity of the equipment to sedimentation and to product fluctuations, the cost of operation, and so on. Table 5 can be used in the selection of special vacuum filter types under a number of criteria. The tabulated data should, however, be regarded only as rough guidelines. Filter types are shown in Figure 79 together with the particle sizes for which they can be used.

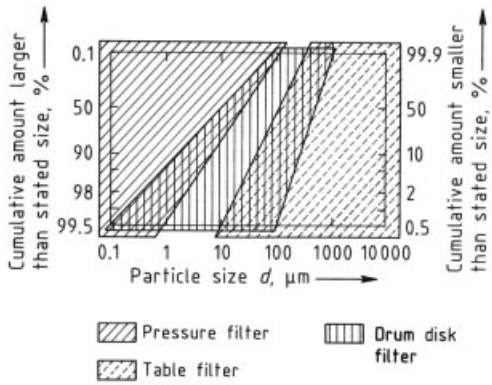

Figure 79. Filter selection as a function of solids particle sizes

The higher cost of pressure filtration in comparison with vacuum filtration is usually not justified except when the cake has a resistance of $3.5 \cdot 10^{11}$ m/kg. Flocculated solids, which should not lose their structure since it is advantageous for filtration, require careful handling, especially in feeding to the filter. Because most such solids are somewhat compressible, not too much pressure should be applied, and so vacuum filtration is commonly preferred for these materials.

A safe selection without trials or appropriate experience is possible only in the rarest instances.

Little general information and few computational formulas can be given in the field of filtration, which is one of the oldest operations in process engineering. Regrettably, filtration (like other essential processes involving the handling of bulk solids, dumped and packed beds, etc.) must deal with the effects of material properties that can be determined only with difficulty or cannot be measured in a predictable way, and these effects are so large and pervasive that there is no comprehensive, unified way of allowing for

Table 5. Applicability of filters

Filter type	Suspension of fine particles	Operation			
		Pressure	Washing	Compression	Continuous (c) automatic (a)
Belt			×		c
Belt with compression	×		×	×	c + a
Nutsche	×	×	×	×	a
Tilting pan			×		c
Leaf/plate	×	×			a
Drum	×				c
Pressure drum	×	×	×		c
Membrane filter press	×	×	×	×	a
Twin-belt pressure filters	×	×	×	×	c
Tube press	×	×		×	a

them. There is no doubt, however, that selection criteria exist and that they imply at least certain filter categories, so that in many cases the number of open options is not very large.

10. Filter Media

The outcome of filtration, that is, the degree to which the filter separates the two phases, is governed above all by the filter medium. Where the available filter media are not suited to the properties of the phase (solid) being separated, filter aids can close the gap.

The filter medium should be selected to provide the desired separating effect and to fit the special properties of the suspension (such as the particle size of the solids and the viscosity of the liquid). The selection of a medium should also take account of service conditions such as cleaning, washing, reaction to back-blowing, and chemical and mechanical stability, which are no less important than the slurry properties. In cake filtration, cake removal must also be considered.

In line with the multiple purposes of filtration, media must perform the following tasks:

1) Optimal retention of all solid particles outside the filter medium
2) Retention of the solid particles, mainly inside the filter medium, by virtue of effects generally lumped together as "entrapment," which essentially have to do with adhesion forces and electrical forces.

Often these functions cannot be unambiguously separated; this is particularly the case with slurries having a very broad particle-size spectrum. In such cases, some concession nearly always has to be made in terms of separation efficiency, or further operating practices, such as careful cleaning or frequent replacement of the filter medium, have to be introduced.

Filter media can be characterized in terms of the following overall criteria:

1) Cut size, that is, the particle size that just passes through the filter medium.
2) Permeability; high permeability means low pressure drop.
3) Chemical stability with respect to the filtrate.
4) Blocking tendency, particularly for fabrics in cake filtration.
5) Mechanical strength in relation to loads imposed in back-blowing or the movement of filter cloths.
6) Smooth surface to promote cake removal.

As is appropriate after many years' experience in filtration, a great many types of filter media have been introduced in practice. New materials and processing methods continue to offer a wide range of improvements. The diversity of filter-media properties has led to a number of classifications in the literature. The following seems to be an advantageous one:

Perforated plates, screens, slotted screens
Fabrics
Beds, felts and other nonwoven materials
Packed beds and precoat layers
Porous materials
Membranes

Within these groups, there are differences in properties, especially with regard to chemical

stability (including aging). However, often physical and biological requirements are also imposed, such as high-temperature stability, water uptake, swelling capacity, weathering behavior (stability to light), and resistance to insects (moths) and bacteria (rotting).

Aside from material properties, the overall filtration qualities can be represented only in an approximate way, because they depend heavily on the properties of the suspension being filtered. Even the often-used permeability and/or mean pore size figures should be considered – at best – rough reference values for the behavior during filtration. A reliable assessment is usually possible only if practical tests or, within limits, laboratory trials (e.g., with the test leaf filter, Chap. 7) are performed.

Perforated Plates, Screens, Slotted Screens. Aside from screens used in screen filtration, filter media in this group serve mainly to support filter cloths or papers, which as a rule have finer pores.

With perforated plates, the largest possible free surface area is usually desired so that the pressure differences to be overcome in filtration will be small. Free areas range from some 20 – 30 % for ordinary plates with circular holes to over 90 % for high-quality screens with square openings in the millimeter range. The free area drops to less than 10 % for very fine sieves with openings a few micrometers across.

Slotted screens with very small openings, used chiefly as strainers and candle filters (see Section 8.3), are made by stacking plates, rings or other elements around a permeable cylinder. The fine gaps between the elements allow only the liquid to pass. Strainers can be fabricated more simply by winding profiled (wedge) wires or strings on a core. These units have little free area, roughly 10 %. As a result, they are economic only when solids concentrations are very low.

Fabrics. By far the most widely used filter media are woven fabrics. Manmade fibers dominate, but fibers of glass, minerals, and metals are also used to an increasing extent, especially in view of their good chemical stabilities.

The advantages of multifilament-yarn fabrics include the depth action of the filter medium, which permits slight penetration of solids into the medium, thus effecting some reduction in pore size and also promoting bridge formation on the medium. The ease of working with fabrics is an incomparable advantage, of course.

The fabric weave is important for filtration proper as well as service properties. Another essential factor is the structure of the fiber, which varies widely in natural and synthetic materials.

Where particularly high strengths are needed, as in twin-belt pressure filters, monofilament-yarn fabrics are preferred. High precision in fabrication allows mesh openings in the micrometer range, especially with metal fibers.

Within some limits it is possible to start with typical fabric data, such as weight per unit area, numbers of warp and filling yarns, fiber diameter, fabric thickness and fiber density, and calculate the pore size and hydraulic resistance. The resistance calculation is based on the pore radii of the fabric and the yarn, and Poiseuille's law is used as a first approximation.

The permeability of a clean, unused fabric can be calculated (with about the same reliability) from Kozeny's equation (66); the starting data are the densities and porosities of the yarn and the fabric. But empirical relations should give better agreement with experimental values.

However, it is vital to know how the fabric influences cake formation and the narrowing and possible blocking of the pores. The resulting change in hydraulic properties of the filter medium is described in Section 2.2.

Nonwovens, Papers, and Felts. These filter media consist of relatively short fibers (staple fibers a few centimeters long) formed into a random layer and then compacted in some way. *Nonwoven media* are usually fabricated in fairly thick layers and therefore function as depth filters (→ Nonwoven Fabrics).

Papers made of cellulose fibers are available in a bewildering range of grades for use in laboratories, small plants, and daily life (→ Paper and Pulp). Heavier papers and paperboards (2 – 6 mm thick) can also be fabricated from mixed materials, such as cellulose plus diatomaceous earth or activated carbon. These filter materials meet a wide range of requirements and are employed in large quantities for the clarification of beer, wine, or juices, and in the dyes and coatings, food, and cosmetics industries. By virtue of the raw materials and fabrication processes, one-

time use of the filter medium is often economic; this feature is especially desirable in sterile filtration.

Conventional *felts* of relatively short natural and manmade fibers are usually not adequate to the filtration and service loads in plant equipment. Felts that incorporate woven-fabric reinforcement are often preferred for this reason. Some loss in terms of the basic advantage of felt must be accepted when these composite materials are used. As a rule, needled felts have lower resistances than woven fabrics. They have now become the most commonly employed nonwovens (→ Felts).

Packed Beds and Precoat Layers. Loose beds of granular materials can also be used in clarifying and some cake filtrations. For economic reasons, cheap natural products such as sand, gravel, or coal are used, but it is desirable to remove the excessively fine-grained fractions from them by washing. Common particle sizes are 3 – 50 mm for gravel, 0.35 – 0.45 mm for fine sand, 0.45 – 0.55 mm for medium sands, and over 0.55 mm for coarse sands. Other candidate materials include minerals (quartz, limestone), coke, anthracite, slag, wood pulp, glass beads, wadding, or fibers.

The hydraulic properties are calculated with the relations developed in the theory section, in particular Darcy's law, Section 2.1.1, or – if the appropriate material data are known – the Kozeny equation (66).

Rigid Porous Materials. Porous materials, which are fundamentally similar in structure, are suitable for any kind of filtration. Their mechanical stability allows their use as filter media without supporting structures (up to pressures of 2 MPa and more). They are preferably employed as filter cartridges or in the form of sheets and more recently, honeycomb and other three-dimensional shapes. Because they are generally too valuable to use just once, and because they can withstand vigorous backwashing without special preparation, solids should be collected mainly by surface filtration. Deep-bed filtration is recommended only when the particles are very small in comparison with the pores and are collected by an adsorptive mechanism. In the intermediate region, solid particles block the pores too quickly, and backwashing may not adequately remove them. Materials for porous filters include minerals, diatomaceous earth, porcelain, clay, coal, graphite, corundum, rubber, plastics, and metals, including highly resistant nickel – chromium – molybdenum alloys. The most important criterion for selection is chemical and thermal stability.

Porous materials made from equal-sized particles are distinguished by uniform porosity. In order to reduce the blocking tendency, layers with differing particle sizes can be superimposed with the fine-pored layer toward the filter cake; thus only the topmost layer need be cleaned, and the filter offers better service qualities.

Porcelain filters are available with pore radii from 0.08 to 6 μm and porosities of 30 – 70 %. Sintered metal filters range from 1 to over 100 μm in pore size and from 30 to 60 % in porosity. Plastic filter media with about the same pore spectra can easily be fitted to the equipment. Their relatively poor temperature stability, however, restricts their use. Besides poly(vinyl chloride) and polyethylene, porous masses can also be fabricated from polytetrafluoroethylene, with its excellent stability.

Membranes. At present a great number of distinct membranes are on the market, not just for ultrafiltration and reverse osmosis, but also for microfiltration (→ Membranes and Membrane Separation Processes). They are made of plastics and are suitable for clarifying as well as pure surface filtration. If very small pores are needed, membranes are assembled from two layers, a coarse-pored supporting layer (with, e.g., 15 – 20 μm pore openings) topped with a filter layer (0.2 μm pores). All membranes have a relatively narrow range of pore sizes.

A special form of membrane filtration employs hollow fibers made of, e.g., polyamides, with diameters of 50 μm and wall thicknesses of 13 μm. The solution being separated is introduced into the hollow fibers under an appropriate pressure. Pure liquid escapes through the wall of the fiber, while the dissolved material remains behind and can easily be carried out of the fiber.

11. Filter Aids

Some suspensions contain solids that quickly fill or block the openings in the filter medium. This

group includes suspensions substantially made up of relatively fine, soft (especially gelatinous), compressible solids. Such feeds can still be filtered well if the blocking of the filter medium is prevented by the inclusion of inert, readily filterable granular materials. These filter aids form a layer permeable to the filtrate, and at the same time they carry out the functions of a loose filter cake. The particles being retained are deposited on the filter aid. This process must still leave sufficiently free openings through which the filtrate can pass. The main candidates for filter aids are thus highly porous materials. Diatomaceous earths are very economical examples; their fineness is characterized by $10\,\% > 40\,\mu m$ at the high end and between 1 and 7 μm at the low end. Also used are perlites and, for special jobs, Fuller's earths, powdered glass, coal preparations, cellulose fibers, wood pulp, paper stock, bagasse (a sugar cane residue), talc, or plastics, although these products are less effective than diatomaceous earths and perlites. The requirements for a good filter aid are uniform quality, correct particle size and shape, and the highest possible wet volume (specific volume in the wet condition, usually greater than the wet volume per mass of filter aid).

In diatomaceous earths, particles consisting of needle-shaped diatoms have proved advantageous; disk-shaped particles filter poorly and produce too dense a cake. Unsuitable constituents such as clay, or sand must be removed from the raw materials, because they hinder filtration by blocking the pores.

A synthetic filter aid is manufactured from volcanic rocks. The raw material, bound with some water, is ground, then suddenly heated to a high temperature; the grains inflate to form spheres. Regrinding yields half-shell-shaped fragments, whose large surface area makes them suitable as a filter aid once the fines are screened out. Their action in filtration is somewhat worse than that of diatomaceous earth, with a porosity 20 – 40 % lower.

In *precoating*, a layer of filter aid several millimeters thick is deposited on the filter medium before filtration proper. After the filtration is completed, the filter aid is often discarded along with the collected solids.

Several steps take advantage of the effect of the loosened filter cake. First, a precoat some 1 – 2 mm thick, containing 200 – 800 g/m^2 of filter aid, is filtered onto the medium (base layer) so that a clear filtrate can be obtained right from the start of filtration. Filter aid is also added to the slurry continuously during filtration. In batch operation, this is done by stirring in the slurry tank; in continuous filtration, dry or wet filter aid can be metered into the slurry delivery line. No generally valid rules exist for the proper ratio of filter aid to solids concentration. A variety of empirical values are available for some processes and solids types. If there is too little filter aid, the filtration goes badly; if too much, the filtrate flow rate drops off too sharply.

The main applications for filtration with filter aids are the precoat technique for poorly filterable slimes of chemical and mineral products and the clarifying filtration of beverages (beer, wine, juices) containing soft, gelatinous impurities, as well as the clarification of gelatins, cane juice, or edible oils.

12. References

1. H. Darcy: "Les fontaines publiques de la ville de Dijon", Paris 1856.
2. VDI-Richtlinie "Filtrierbarkeit von Suspensionen" VDI 2762, 1997.
3. L. Svarovsky: *Solid – Liquid Separation,* Butterworths, London 1977, p. 175.
4. M. Tiller, T. Cleveland, R. Lu, *Ind. Eng. Chem. Res.* **38** (1999) 590 – 595.
5. P. H. Hermans, H. L. Bredée, *Rec. Trav. Chim.* **54** (1935) 680 – 700.
6. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology,* VCH Verlagsgesellschaft, Weinheim 1996, pp. 64 ff.
7. K. Luckert, *Wiss. Z. Techn. Univ. Magdeburg* **36** (1992) no. 5/6, 74 – 80.
8. P. L. Boucher, *J. Proc. Inst. Civ. Eng.* (1946 – 1947) 415 – 445.
9. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology,* VCH Verlagsgesellschaft, Weinheim 1996, pp. 177 ff.
10. K. J. Ives, in A. Rushton (ed.): *Mathematical Models and Design Methods in Solid-Liquid Separation, NATO ASI series E No. 88,* Martinus Nijhoff, Dordrecht 1985, pp. 90 ff.
11. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology,* VCH Verlagsgesellschaft, Weinheim 1996, p. 198.

12. G. Dierckx in: *La Filtration Industrielle des Liquides, Tome III,* Société Belge de Filtration, Liège 1978, pp. 295 – 314.
13. A. Palinski, W. Uhl, R. Gimbel, *Vom Wasser* **89** (1997) 175 – 189.
14. J. Altmann, S. Ripperger, *J. Membrane Sci.* **124** (1997) 119 – 128.
15. S. Ripperger, J. Altmann, *Filtrieren Separieren* **11** (1997) no. 3, 109 – 113.
16. W. Bender, *Chem. Ing. Tech.* **55** (1983) no. 11, 823 – 829.
17. K.-H. Steiner, *Chem. Ing. Tech.* **61** (1989) no. 1, 1 – 8.
18. H. Anlauf, *Fortschrittsber. VDI, Reihe 3,* **114** (1986).
19. H. Schubert: *Kapillarität in porösen Feststoffsystemen,* Springer Verlag, Heidelberg 1982, p. 184.
20. H. Anlauf, *Maschinenmarkt* **95** (1989) no. 8, 26 – 30.
21. R. J. Wakeman: "Cake Dewatering," in L. Svarovsky (ed.): *Solid – Liquid Separation,* Butterworths, London 1977, p. 300.
22. R. J. Wakeman, *Filtr. Sep.* **16** (1979) no. 6, 655 – 669.
23. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology,* VCH Verlagsgesellschaft, Weinheim 1996, pp. 348 ff.
24. D. Redeker, K.-H. Steiner, U. Esser, *Chem. Ing. Tech.* **55** (1983) no. 11, 829 – 839.
25. I. Nicolaou, W. Stahl, *Aufbereit. Tech.* **33** (1992) no. 6, 328 – 338.
26. J. Walker, *Spektrum Wissensch.* **12** (1986) 188 – 197.
27. W. Tiedemann, *Fortschrittsber. VDI, Reihe 3,* **453** (1996).
28. M. Shirato, T. Murase, E. Iritani, S. Nakatsuka, *Filtr. Sep.* **24** (1987) no. 2, 115 – 119.
29. A. Rushton, A. S. Ward, R. G. Holdich: *Solid – Liquid Filtration and Separation Technology,* VCH Verlagsgesellschaft, Weinheim 1996, pp. 455 ff.
30. D. Leclerc, S. Rebouillat, in A. Rushton (ed.): Mathematical Models and Design Methods in Solid – Liquid Separation, NATO ASI series E No. 88, Martinus Nijhoff, Dordrecht 1985, pp. 356 ff.
31. J. Kozeny, *Ber. Math.-naturwiss. Abt. Akad. Wien.* (1927) 271 – 306.
32. A. E. Scheidegger: *The Physics of Flow Through Porous Media,* 3rd ed., University of Toronto Press, Toronto 1974, pp. 137 ff.
33. H. Rumpf, A. R. Gupte, *Chem. Ing. Tech.* **43** (1971) no. 6, 367 – 375.
34. J. Gregory in K. J. Ives (ed.): *The Scientific Basis of Flocculation,* Sijdhoff & Noordhoff Alphen aan de Rijn 1978, p. 91.
35. H.-J. Jacobasch, P. Weidenhammer, *Chem. Ing. Tech.* **68** (1996) 1590 – 1594.
36. D. Houi, R. Lenormand, in F. Family, D. P. Landau (eds.): *Kinetics of Aggregation and Gelation,* Elsevier, Amsterdam 1984.
37. P. Schmitz, B. Wandelt, D. Houi, M. Hildenbrand, *J. Membrane. Sci.* **84** (1993) 171 – 183.
38. W. Höflinger, C. Stöckelmayer, A. Hackl, *Filtr. Sep.* (1994) Dec., 807 – 811.
39. W. Höflinger, C. Stöckelmayer, *Staub Reinhalt. Luft* **55** (1995) 423 – 428.
40. W.-M. Lu, C. C. Lai, K.-J. Hwang, *Sep. Technol.* **5** (1995) 45 – 53.
41. "Filtrierbarkeit von Suspensionen," *VDI-Richtlinie* **2762** (1997).
42. H. Anlauf, *Filtrieren Separieren* **8** (1994) no. 2, 63 – 66, 69 – 70.
43. H. Anlauf, *Filtrieren Separieren* **8** (1994) no. 3, 116 – 118, 121 – 126.
44. W. Gösele, *Filtrieren Separieren* **9** (1995) no. 1, 14 – 22.
45. F. M. Tiller, C. S. Yeh in H. S. Muralidhara (ed.): *Advances in Solid – Liquid Separation,* Batelle Press, Columbus OH 1986, pp. 1 – 37.

Sedimentation

See also → Solid–Liquid Separation, Introduction

YORAM ZIMMELS, Department of Civil Engineering, Technion IIT, Haifa, Israel

1. Scope 5
2. Theory 5
2.1. Steady Sedimentation of Nondiffusing Particulates 5
2.1.1. Monodisperse Particulates 5
2.1.2. Polydisperse Particulates 10
2.1.3. Aggregated Dispersions 12
2.2. Time Dependence of Steady Sedimentation Systems 15
2.3. Nonsteady Sedimentation 15
2.3.1. System Characterization 15
2.3.2. Computer Program for Simulation of Nonsteady One-Dimensional Sedimentation of Polydisperse Mixtures . 17
2.3.3. Graphical Output 18
2.3.4. Transient Sedimentation 23
2.3.5. Sedimentation in Turbulent Suspension 26
2.4. Sedimentation of Diffusing Particulates 26
2.4.1. Monodisperse System 26
2.4.2. Polydisperse System 28
2.5. Sedimentation of Fluid Particles . . 29
2.6. Generalized Correlations for Hindered Sedimentation Fluidization and Flow through Fixed Beds 30
2.7. Enhancement of Sedimentation . . . 32
2.8. Sedimentation in Inclined Channels 33
2.9. Principle of Hydrodynamic Particle Competition 35
3. Processes and Equipment Involving Sedimentation 35
3.1. Settling Pools 35
3.2. Suspension of Particles in Fluids . . 36
3.2.1. Suspension of Particles in Mixed Containers 36
3.2.2. Suspension of Particles in Conduits . 38
3.3. Fluidization and Sedimentation . . . 40
3.4. Slurry Transportation in Horizontal Pipes 41
3.5. Sedimentation in Centrifugal Fields 48
3.5.1. Ultracentrifugation 48
3.5.2. Magnetic and Magnetocentrifugal Sedimentation 58
3.5.3. Centrifugal Sedimentation in Flowing Dispersions 60
3.5.4. Solid – Liquid Separation by Sedimentation 63
4. References 64

Symbols

a_f	acceleration of fluid
a_i	constant ($i = 1, 2, \ldots$)
a_1, a_2, a_3	constants
$a(d)$	size-dependent constant
A, A_1	constants
A_d	constant dependent on ratio of viscosities
A_s	area under a schlieren peak; also used to denote area across which the dispersion flows in a settling pool
b	length of shorter wall of inclined channel
b_j	constants ($j = 1, 2, \ldots$)
$b*$	dimensionless length
$b(d)$	size-dependent constant
B_d, B_v	functions of volume fraction and viscosity
B_p	pool width
B_1	constant
C	concentration of particles
C_a	concentration at a reference level y_a or at the center of a conduit
$C_D, C_{D\varphi}, C_{D0}$	drag coefficient: general, generalized concentration dependent, and at infinite dilution, respectively
C_h, C_{ht}	concentration in the vehicle and in the heterogeneous part of the suspension, respectively
C_m	concentration at $r = r_m$
$C_0, C(t)$	initial and time-dependent concentration, respectively

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_12

C_1, C_2	concentrations at regions 1 and 2 in differential sedimentation system
d, d_i	particle size: general and characterizing the ith size fraction, respectively
d_L	limiting size for capture in a settling pool
d_{max}	limiting size of particles for which $G(d > d_{max}) = 1$
d_{50}	equiprobable size of particles, defined by $d_{50} = d\,[G(d) = 0.5]$
D	Stokes – Einstein diffusion coefficient; also used to denote pipe diameter
D'	diffusion coefficient
D_a, D_T	diameter of impeller and tank, respectively
D_c	diameter or characteristic size of container
D_e	eddy diffusion coefficient
D_m	fluid momentum transfer coefficient
D_s	mass-transfer coefficient
D'_{sed}	diffusion coefficient obtained from measurement in an ultracentrifuge
$D_\varphi(r, \eta), D_d(r, \eta)$	position- and η-dependent diffusion coefficient in the fluid and the dispersion frame of reference, respectively
$\overline{D_\varphi(r)}$	expectation of $D_\varphi(r, \eta)$ with respect to η
E	electric field strength
f	friction factor
f_a	activity coefficient of diffusing particles
f_D	drag coefficient of a particle
f_l, f_s	fluid and slurry friction factors, respectively
f_p	driving force density
$f(\xi)$	distribution function of ξ
$f(\varphi), f_D(C)$	hydrodynamic hindrance factors
$f_1(\varphi), f_2(\varepsilon), f_3(\varepsilon)$	functions of φ and ε
F_D	hydrodynamic drag force
$F_{D\varphi}$	generalized concentration-dependent hydrodynamic drag force
$F_G, F_\omega, F_q, F_M, F_E$	gravitational, centrifugal, electrical, magnetic, and electrical polarization forces, respectively
F_i	ith force acting on a particle
F_R	reaction force due to accelerated fluid
F_γ	force associated with γ
F_μ	diffusion driving force
Fr	Froude number
Fr_L	modified Froude number
g	gravitational acceleration
g'	effective acceleration
$G(d)$	grade efficiency
h	height of sedimentation column in nonsteady sedimentation models
h_c	channel height in a disk centrifuge
h_s	height of a schlieren peak
h_1	thickness of one layer of a sedimentation column
H	magnetic field strength
H_a	height of impeller above tank floor
H_j	height
H_j^*	dimensionless height
H_p	pool height
i	integer
j, j'	integers denoting current and start-off layer numbers
J, J_{max}	particle flux (and flux densitiy) and maximum particle flux in the dispersion frame of reference, respectively
J_D, J_F	particle flux density due to diffusion and sedimentation, respectively
k	number of time increments elapsed
$k*$	surface effect retardation constant
k_B	Boltzmann constant
K, K_1, K_2, K_3, K_w	constants
K_a	design parameter of agitation tanks
K_{ac}	agitation concentration factor
K_f	Karman constant
l_m	eddy mixing length
L	length
L_p, L_e	length of straight pipe and equivalent pipe length of fittings, respectively; is also used as length of inclined channel
L_t	total effective pipe length
m, m_f	mass and mass of fluid, respectively
M_i	ith component of magnetization
M_p, M_f, M_φ	magnetization of particles, fluid, and dispersion, respectively

M_r molecular mass
n exponent; also an integer
n_i number of particles in the ith fraction per unit volume of dispersion
$\tilde{N}$ Avogadro number
N number of particles per unit volume of dispersion; also used to denote number of sedimentation zones in an inclined channel
P pressure; also used to denote electrical polarization of particles (i.e., dipole moment per unit volume) and power
P_{EP}, $P_{E\varphi}$ electrical polarization (i.e., dipole moment per unit volume) of particles and dispersion, respectively
P_0 power number
ΔP_{exp}, ΔP_{cal} measured and calculated pressure drop, respectively
ΔP_l, ΔP_t, ΔP_i pressure drop across a bed of particles in the laminar, turbulent, and intermediate flow regimes, respectively
q_v, q_s, $q_{v\varphi}$, $q_{s\varphi}$ charge per unit volume and charge per unit area of particles and dispersion (subscript φ)
$q(t)$ volume flow rate of clarified fluid per channel per unit length
Q flow rate in a pool; also used to denote volumetric flow rate in a centrifuge
Q_p particle flux in a centrifuge
$Q(t)$ volumetric rate of fluid clarification in an inclined channel
r radial distance from axis of rotation; also used to denote position vector and radial coordinate
$\bar{r}$ radial distance to a point in the boundary which moves at the same velocity as particles in the dispersion ahead
r_m, r_p radial distance of meniscus and of a point in the region of concentration plateau, respectively
R gas constant
Re, Re_φ Reynolds number and generalized concentration-dependent particle Reynolds number
Re_a Reynolds number of an agitator
S cross-sectional area of particle perpendicular to direction of flow
S_g, $S_{g(20,W)}$ sedimentation coefficient and its value at standard state, respectively
S_{gD} differential sedimentation coefficient
S_p shape factor
$S(i, j', k)$ array of sedimentation distances (at time t)
$S(t)$ time-dependent distance
$S_j^*(t)$ dimensionless instantaneous volumetric rate of formation of suspension devoid of particles faster than (or equal to) particles of the jth fraction
t time
t_H, t_R time for sedimentation to the pool floor and for crossing the pool, respectively
T absolute temperature
u_f flow velocity of a carrier fluid
U flow velocity of fluid in a conduit; also used as velocity of particles
U_c superficial velocity; also used to denote critical deposition velocity of particles on a pipe wall
U_d, U_f velocity of particles and fluid, respectively, in the dispersion frame of reference
U_{di}, $U_{\varphi i}$ velocity of ith particles in the dispersion and the fluid frame of reference, respectively
U_{exp} measured velocity of fluidizing gas
U_g free sedimentation velocity of d_{50} particles
U_h transition velocity from homogeneous to heterogeneous flow
U_j velocity of a particle
U_R crossflow velocity of fluid in a settling pool
U_x, U_y x and y components of particle velocity in the dispersion frame of reference
U_φ velocity of particles in the fluid frame of reference
$U_{\varphi a}$ velocity of aggregates in the fluid frame of reference
U_0 $= U_\varphi$ at $\varphi = 0$
U'_i energy per diffusing particle
U'_{M1}, U'_{M2}, U'_{H1}, U'_{H2} magnetic energies
$U*$ friction velocity
U^*_{dj} dimensionless velocity

$U_{\mathrm{d}}(t)$, $U_{\mathrm{d}}(\infty)$ time-dependent and terminal settling velocity in the dispersion frame of reference
v velocity of slurry; also used to denote particle specific volume
V, V_{p} general volume and volume of a particle, respectively
V_{b} volume of dispersion in centrifuge bowl
V_{v} volume per unit length of inclined channel sedimentation vessel
V_{T} volume of tank
V_{φ} volume of dispersion
w weight fraction of solids
w_{s} weight of dispersed phase
x_{s} function of $U_{\mathrm{d}}/U_{\mathrm{c}}$
y height
y_{a} reference level above conduit bottom
y_{m} distance from bottom wall of conduit to a point in the dispersion

Greek Symbols

α size parameter that determines the four size fractions in the computer model
α_{s} safety factor
α_{φ} concentration-dependent factor
$\alpha_1, \alpha_2, \alpha_3, \alpha_4$ variables in the equation of motion
β_{s} constant
β_{φ} concentration-dependent factor
$\beta(t)$ time-dependent force transmission parameter
γ interfacial tension; also used to denote $U_{\mathrm{d}}/U_{\mathrm{c}}$
$\gamma_{\mathrm{p}}, \gamma_{\mathrm{f}}, \gamma_{\varphi}$ physical parameters (size, density, etc.) of particles, fluid, and dispersion, respectively
Γ adsorption density
δ $= \varrho_{\mathrm{p}} - \varrho_{\varphi}$; also used to denote thickness of viscous sublayer
ε void fraction
ε_{m} effective magnetic porosity
ζ parameter
η parameter denoting size, density, etc.
η_{p} pool efficiency factor
θ angle of inclination
λ resistance to flow of a carrier fluid in a pipe
μ chemical potential
μ_{d} viscosity of fluid of the dispersed phase
μ_{f} viscosity of pure fluid
$\mu_{\varphi}, \mu'_{\varphi}$ concentration-dependent extended viscosities
$\mu_{\varphi\mathrm{d}}$ apparent "Taylor" viscosity of a continuous fluid containing dispersed fluid particles
μ_0 magnetic permeability of vacuum
μ'_{f} effective viscosity of a continuous fluid surrounding a single spherical fluid particle
μ^*_{d} extended viscosity inclusive of interfacial effects
ν_{f} kinematic viscosity
ξ, $\bar{\xi}$ distribution variable characterizing particles (i.e., size, density, etc.) and its expectation, respectively
ϱ density
$\varrho_{\mathrm{p}}, \varrho_{\mathrm{p}i}, \varrho_{\mathrm{f}}$ density of particle, ith particle, and pure fluid, respectively
ϱ_{φ} density of dispersion
σ_{k} surface parameter of viscosity
Σ, Σ_{φ} general and extended concentration-dependent index of centrifuge size
τ_{w} shear stress at wall
τ_{φ} factor accounting for degree of particle fixation in space
φ total volume fraction occupied by particles in dispersion
φ_{a} volume fraction occupied by aggregates in dispersion
φ_i, $\varphi_i(d_i)$ volume fraction occupied by ith particle fraction and by ith size fraction (d_i) in dispersion, respectively
$\varphi_{\max}, \varphi_{\mathrm{m}}$ volume fraction at maximum particle and aggregate flux, respectively
φ_{s} volume fraction occupied by particles in a single aggregate
φ' probability density function
$\varphi(i, j)$ time-dependent volume fraction occupied by ith particles in the jth layer
$\varphi(i, j')$ invariable volume fraction of ith particles that start sedimentation at the jth layer
$\varphi(j)$ $= \sum_i \varphi(i, j)$
$\varphi_{\mathrm{s}}(i, j)$ volume fraction occupied by ith particles in the sediment
Φ cumulative size distribution
χ_{p} susceptibility of particles
ω angular velocity
$\omega_1, \omega_{\varphi}$ effective, concentration-dependent, angular velocities

1. Scope

Sedimentation is the motion of particulates in a fluid which is generated by a force field. This motion often results in increased concentration of the particulates, either in the form of a concentrated dispersion or as a sediment.

Sedimentation involves the motion of particulates under the action of driving forces that are generated by force fields, such as gravitational, centrifugal (→ Centrifugation, Sedimenting), magnetic (→ Magnetic Separation), and electrical (→ Electrostatic Separation) fields. The use of a particular field depends on the process required and its technical and economical feasibility. To date, sedimentation techniques can handle particulates with a variety of properties (e.g., size and density). This is because the force fields that are available today can produce force densities equivalent to 300 000 g and higher. Such high force densities are readily produced by commercial ultracentrifuges and, in some cases, also by electromagnetic fields. Thus large particulates, as well as those having dimensions comparable to molecular sizes, can be handled by conventional sedimentation techniques. This capacity is applied in chemical, biochemical (→ Biochemical Separations), mineral, and environmental processing technologies. It is also used as a standard means of biomedical testing and sample preparation. A world without sedimentation phenomena, whether natural or manmade, would be incomprehensible. In some way, sedimentation exists around us every day in the form of natural (i.e., rain, snow, dust, sedimentation of silt in rivers) or artificial (i.e., processing and pollution products) precipitation. Living in a world where gravity prevails means that sedimentation exists, to different degrees of significance, in every system in which one or more phases are dispersed in a fluid.

2. Theory

Sedimentation can be part of a process in which particulates dispersed in a fluid are involved, or else it can be used to separate them from the dispersion in the form of a concentrated sludge or sediment. Analysis of sedimentation requires a distinction between steady-state, continuous and nonsteady, time-dependent systems, in conjunction with either mono- or polydisperse dispersions. Depending on the frame of reference, time-dependent sedimentation may be part of a flow system running continuously at steady state. In this article, the theory of steady sedimentation of mono- and polydisperse mixtures is reviewed first; it is then extended to include nonsteady conditions (for a theoretical treatment of the motion of particles in fluids, see → Elutriation).

2.1. Steady Sedimentation of Nondiffusing Particulates

2.1.1. Monodisperse Particulates

The term *monodisperse particulates* implies that all dispersed particulates are identical. The following material considers the steady-state "terminal velocity" of monodisperse spherical particulates uniformly dispersed in a Newtonian fluid [1]. Let U_φ denote the velocity of the particulates in the fluid frame of reference. Thus, U_φ is the velocity of particulates relative to the fluid being displaced by their motion; it can be evaluated in the creeping (i.e., laminar) flow regime where $Re < 1$ ($Re = 0.2$ defines the upper limit of the creeping flow regime) by Equation (1), and in intermediate and turbulent flow where $Re > 1$, by Equation (2). Equation (2) also provides a good approximation for creeping flow, but in this regime Equation (1) is preferable. The flow regimes that apply to Equations (1) and (2) pertain to flow around the particles and not to the whole dispersion. Thus, turbulent flow means the formation of eddies due to unstable flow around the particles and not due to other sources, such as agitation or turbulent flow in a conduit.

$$U_\varphi = \frac{(\varrho_\mathrm{p} - \varrho_\mathrm{f})\, g\prime\, d^2}{18\mu_\mathrm{f}} f(\varphi)\,,$$
$$f(\varphi) = \frac{(1-\varphi)}{\left(1+\varphi^{1/3}\right)\, \exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right)} \tag{1}$$

$$U_\varphi = \left(\frac{-a_2 + \sqrt{a_2^2 + 4a_1 a_3^{1/2}}}{2a_1}\right)^2 \tag{2}$$

$$a_1 = 0.63\,(d\varrho_f)^{1/2}$$

$$a_2 = 4.80\mu_f^{1/2}\exp\left(\frac{5}{6}\cdot\frac{\varphi}{1-\varphi}\right)$$

$$a_3 = 4.0\,(\varrho_p - \varrho_f)\,(1-\varphi)\,g'\,d^2/\left[3\left(1+\varphi^{1/3}\right)\right]$$

where φ is the volume fraction occupied by the particulates in the dispersion; d is particle diameter (i.e., a measure of particle size); ϱ_p and ϱ_f are densities of particulates and fluid, respectively; μ_f is the viscosity of the pure fluid (i.e., at $\varphi = 0$); and g' is the effective acceleration defined by Equations (3) – (5):

$$g' = f_p/\delta \tag{3}$$

$$\delta = \varrho_p - \varrho_\varphi \tag{4}$$

$$\varrho_\varphi = \varphi\,\varrho_p + (1-\varphi)\,\varrho_f \tag{5}$$

where f_p is the driving force density (i.e., force per unit volume) due to which steady motion occurs. In *gravitational* and *centrifugal fields,* $g' = g$ and $g' = \omega^2 r$, respectively, where g is acceleration due to gravity, ω is angular velocity, and r is radius of rotation. In *magnetic fields,* g' can be evaluated by Equation (6):

$$g' = \mu_0 \nabla\,[(M_p - M_\varphi)\cdot H]\,/\,(\varrho_p - \varrho_\varphi) \tag{6}$$

where M_p and M_φ represent the magnetization (i.e., magnetic moment per unit volume) of particulates and of the dispersion, respectively, and H is magnetic field strength. The electrical counterpart of Equation (6) is obtained by replacing $\mu_0(M_p - M_\varphi)$ by $P_{Ep} - P_{E\varphi}$ and H by E, where P_{Ep} and $P_{E\varphi}$ denote polarization of particulates and dispersion, respectively, and E is electric field strength. Let U_d be the velocity of particulates relative to the dispersion frame of reference. Then, if the dispersion is at rest relative to its container, U_d is also the velocity in the container frame of reference. Figure 1 shows the definition of U_φ and U_d. The velocity U_f of the displaced fluid in the dispersion frame of reference is given by

$$U_f = U_\varphi \varphi \tag{7}$$

Hence,

$$U_d = U_\varphi - U_f = U_\varphi\,(1-\varphi) \tag{8}$$

Note that Equation (8) applies when by continuity the displaced fluid flows counter to the particulates so as to maintain volumetric balance (→ Elutriation). Other forms of $f(\varphi)$ are known [2–4], particularly for sedimentation involving low Reynolds numbers [2]. The correlation given by Equation (9) is attributed to Richardson and Zaki [5]:

$$f(\varphi) = (1-\varphi)^{n-1} \tag{9}$$

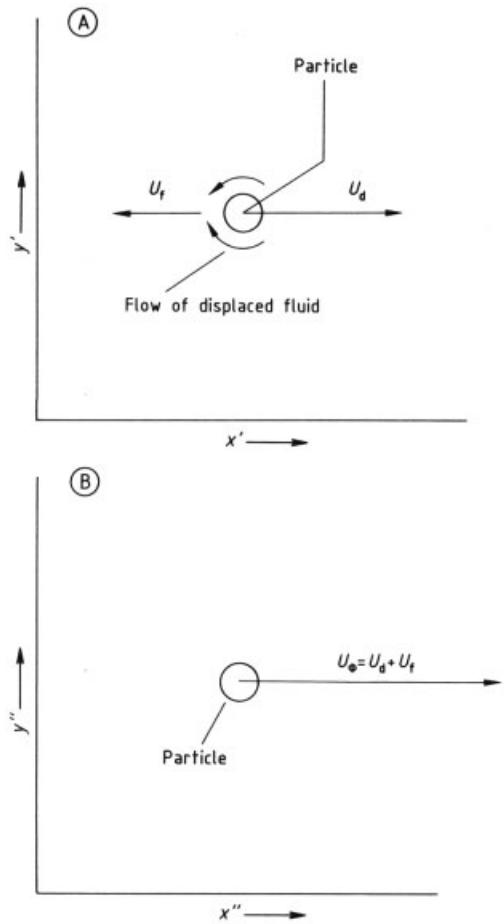

Figure 1. Definition of U_φ and U_d in the fluid and dispersion frame of references
A) Dispersion frame of reference x', y'; B) Fluid frame of reference x'', y''

The results of Garside and Al-Dibouni [6] suggest a value of $n = 5.1$. Dimensional analysis indicates that n should depend solely on wall effects for both creeping and turbulent flow, and on wall effects and Reynolds number Re evaluated at $\varphi = 0$ in the intermediate flow regime [5].

The results [7] corresponding to flow ranges of *Re* excluding wall effects, are given in the following material:

Range of *Re*	*n*
$Re < 0.2$	4.65
$0.2 < Re < 1.0$	$4.4\,Re^{-0.03}$
$1.0 < Re < 500$	$4.4\,Re^{-0.1}$
$Re > 500$	2.4

Note that n is highest in laminar flow and drops as turbulence increases. For more information on generalized $f(\varphi)$ functions, see Section 2.6. For dilute dispersions [2] ($\varphi \to 0$), $f(\varphi)$ given by Equation (1) reduces to

$$f(\varphi) = \left(1-\varphi^{1/3}\right) / (1-\varphi) \tag{10}$$

whereas Equation (9) for $n = 5.1$ gives

$$f(\varphi) = (1 - 5.1\varphi) / (1-\varphi) \tag{11}$$

Figure 2 shows a plot of $(1-\varphi)f(\varphi)$ vs. φ for four correlations, two of which are given by Equations (1) and (9) [2]. Equation (1) [3] (curve b) predicts a steeper decrease in $f(\varphi)$ for $\varphi < 0.2$ than Equation (9) does; the reverse is true for $\varphi > 0.2$.

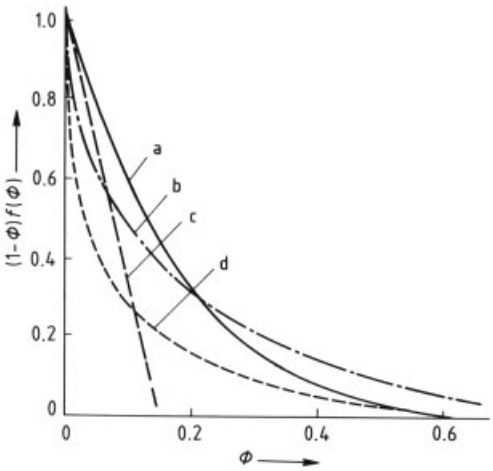

Figure 2. A plot of $(1-\varphi)\cdot f(\varphi)$ vs. φ
a) Equation (9); b) Equation (1); c) Batchelor's theory of dilute suspension [8]; d) Model of simple cubic arrays (Sangani and Acrivos [9])

The discrepancy between the four correlations is clear. Motion of particulates at very low Reynolds numbers may involve factors other than driving and hydrodynamic drag forces. In such cases, the discrepancy between theory and experimental results may not be due entirely to the inadequacy of the theory. In this article, $f(\varphi)$ is defined according to Equation (1) because of the vast amounts of experimental data that support it [3].

Sedimentation volume flux density J is defined as the volume of particulates crossing a unit area per unit time. The flux density is given in the dispersion frame of reference by

$$J = U_{\mathrm{d}}\varphi = U_{\varphi}\,(1-\varphi)\,\varphi \tag{12}$$

In the laminar flow regime,

$$J = \frac{(\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}})\, g\prime d^2}{18\mu_{\mathrm{f}}}\,(1-\varphi)\,\varphi \cdot f(\varphi)\,, Re < 1 \tag{13}$$

and in the intermediate and turbulent regimes,

$$J = \left(\frac{-a_2 + \sqrt{a_2^2 + 4a_1 a_3^{1/2}}}{2a_1}\right)^2 \cdot (1-\varphi)\,\varphi, \quad Re \geq 1 \tag{14}$$

Figure 3 shows a plot of J vs. φ calculated for sedimentation of particles ($\varrho_{\mathrm{p}} = 3000\,\mathrm{kg/m^3}$, different sizes) in water [1], [10]. The dark circles are results obtained by using the generalized correlation of Barnea and Mizrahi [3] (→ Elutriation, Chap. 3.1.), and the open circles were obtained by using Equation (14). The two sets of calculated results are in good agreement; hence, Equation (14) can be used safely. The volume fraction $\varphi_{\max}$ that maximizes the flux density varies with particle size. Calculated values of $\varphi_{\max}$ vs. d are plotted in Figure 4. In the laminar flow regime, $\varphi_{\max} = 0.1792$ and is independent of particle size. At this volume fraction, $J_{\max} = 0.2994\,\varphi_{\max}\,U_{\mathrm{d0}}$, where U_{d0} is the free sedimentation velocity (i.e., at $\varphi \to 0$). Therefore, $J_{\max}$ is 29.94 % of the flux limit (in the given fluid and at $\varphi_{\max}$), which is expected if the hindrance due to φ could have been eliminated so that free sedimentation prevails. Increase of particle size or transfer of motion into the turbulent regime results in an increase of $\varphi_{\max}$ (see Fig. 4):

$$\log\varphi_{\max} = A_1 \log d + B_1 \tag{15}$$

where for settling in dry air (18 °C), $A_1 = 0.1320$, $B_1 = -0.9505$ and for settling in water (25 °C), $A_1 = 0.1309$, $B_1 = -0.9946$.

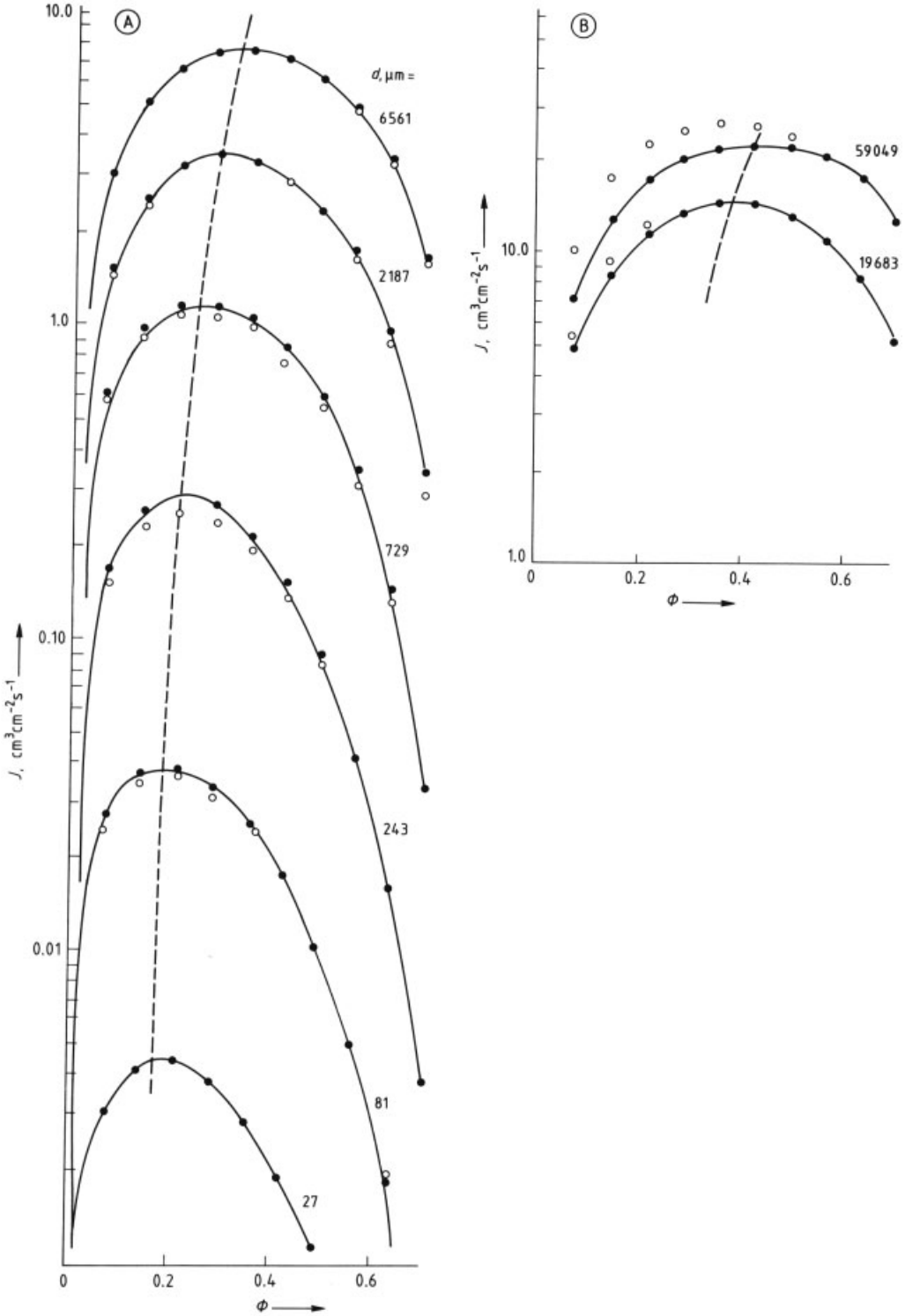

Figure 3. Calculated flux densities J vs. φ for different particle sizes, $\varrho_p = 3000\,\mathrm{kg/m^3}$, sedimentation in water [1], [10] -•- General correlation; -○- Dallavalle

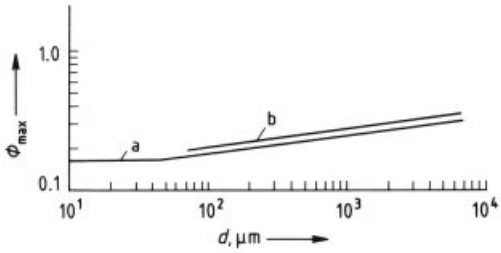

Figure 4. Calculated values of $\varphi_{\max}$ vs. d (data from Fig. 3 [1], [10])
a) Water; b) Air

These data may be used for process evaluation regarding the capacity of phases involving sedimentation. The following points are suggested for consideration in evaluation of sedimentation processes:

1) For most practical purposes, $0.1792 < \varphi_{\max} < 0.40$.
2) The flux density of finer and less dense particles is sensitive to changes in φ and $\varphi_{\max}$ is independent of particle size in the laminar flow regime.
3) The flux density of larger and more dense particles is less sensitive to variation of φ in the vicinity of $\varphi_{\max}$. Dispersions containing larger particles may be processed at higher levels of φ.

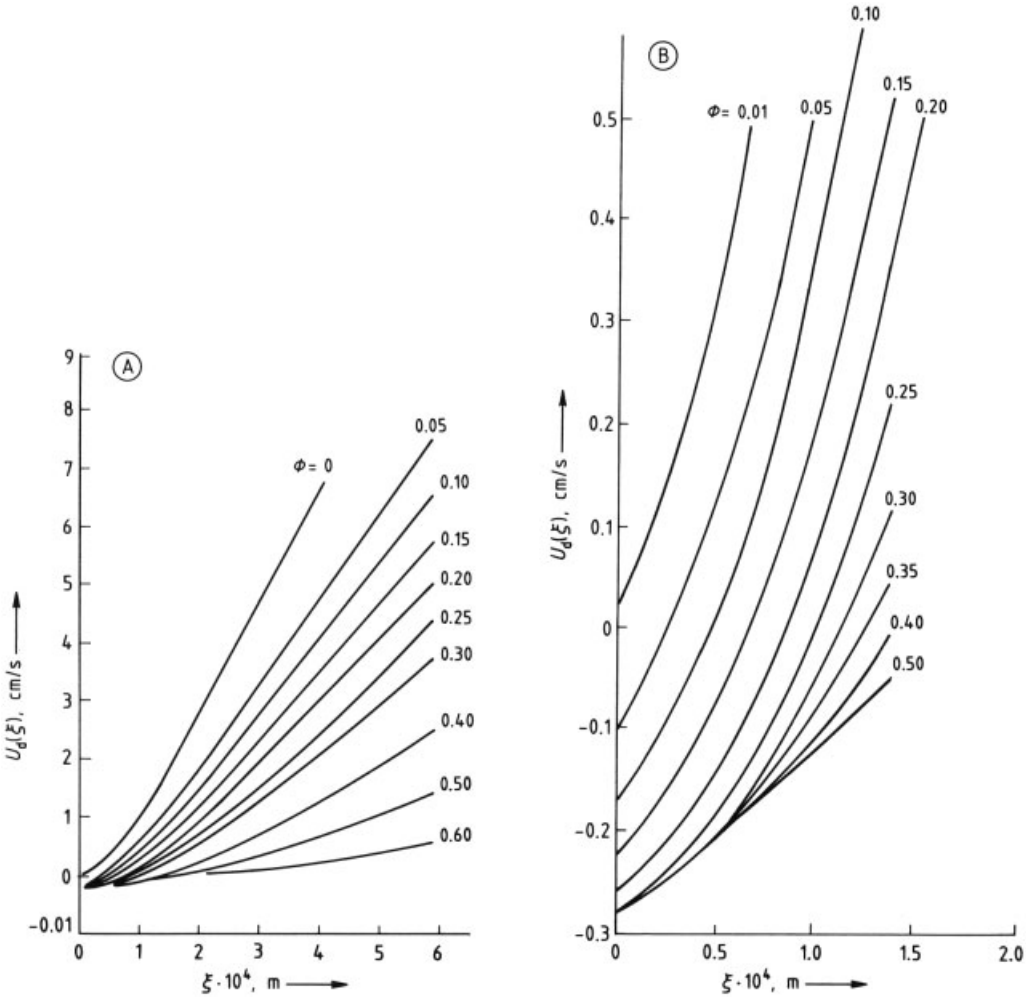

Figure 5. Calculated values of $U_{\mathrm{d}}(\xi)$ vs. ξ
A) Truncated normal distribution $\bar{\xi} = 10^{-4}$ m, $\sigma = 3\,\bar{\xi}$ [11]; B) Detail of A at the finer size end

4) The gain in particle flux density, obtained due to increase in particle size, is largest in the vicinity of $Re = 1$.

For further details, see [1].

The evaluation of flux densities is important in mass balance calculations. If the incoming flux exceeds the limiting outgoing flux in a sedimentation zone, then undesirable accumulation and surging are expected, which can produce instability and process shutdown.

2.1.2. Polydisperse Particulates

The term *polydisperse particulates* means that the dispersed particulates are not identical and may differ in properties such as size, density, shape, polarizability, and charge. Alternatively, it implies that properties of the dispersed particulates can be described by distribution functions.

Distribution Functions. (For a review of useful distribution functions, see → Particle Size Analysis and Characterisation of a Classification Process, Chap. 2.3. and → Elutriation, Chap. 3.7.). Because particle size is a dominant factor in many sedimentation systems, and in powder technology generally, a wealth of size distribution functions is available from the literature. If a known distribution function cannot be made to fit particle-size data, then the data must be divided into appropriate size fractions small enough to be assumed uniform. In general, information related to the distribution of properties that are significant in the sedimentation of polydisperse particle mixtures, must be obtained and processed to a usable form prior to implementation of sedimentation theories or to evaluation of velocity and flux.

Sedimentation Velocities and Flux Densities.

Discrete Distributions. If the total volume fraction occupied by the particulates in a dispersion is expressed in terms of uniform fractions as

$$\varphi = \sum_{i=1}^{n} \sum_{j=1}^{m} \ldots \varphi_{ij} \ldots (d, \varrho, \ldots) \tag{16}$$

where the dots denote the inclusion of summation and the indexing of variables other than d and ϱ. Thus, if three variables characterize the polydisperse mixture, then three summation symbols, three indexes, and three independent variables will be used in Equation (16). For the sake of clarity, only one independent variable (i.e., size) is considered here, for which

$$\varphi = \sum_{i=1}^{n} \varphi_i \, (d_i) \tag{17}$$

Equations (1) and (2) can also be used to determine the sedimentation velocities of fractions that are part of a polydisperse mixture, with φ given by Equation (17). This involves the assumption that hydrodynamic hindrance to motion of the particulates is (to a good approximation) solely dependent on the total volume fraction occupied by the whole distribution, i.e., that it is independent of the distributions characterizing the dispersion. Thus although $U_{\varphi i}$, pertaining to the ith fraction, is assumed to be independent of effects of the distributions, U_{f} and hence also $U_{\mathrm{d}i}$ are directly affected by them. The effect of interparticle collision is neglected; otherwise, $U_{\varphi i}$ must be corrected by a collision factor $\alpha\,(\varphi)$ which is assumed here at a value of unity.

$$U_{\mathrm{f}} = \sum_{i=1}^{n} U_{\varphi i} \, \varphi_{\mathrm{i}} \tag{18}$$

$$U_{\mathrm{d}i} = U_{\varphi i} - U_{\mathrm{f}} = U_{\varphi i} - \sum_{i=1}^{n} U_{\varphi \mathrm{i}} \varphi_{\mathrm{i}} \tag{19}$$

Equations (18) and (19) reflect the effect of counterflowing fluid set in motion by the whole polydisperse mixture. These currents have the strongest effect on slower particles. Observe that in Equations (18) and (19), $U_{\varphi i}$ represents vector quantities; hence, the terms $U_{\varphi i} \, \varphi_i$ should be added accordingly. Equation (19) facilitates an extended definition of "equal sedimentation" (or equal settling) particles:

$$U_{\mathrm{d}i} = U_{\mathrm{d}j}, \quad i \neq j \tag{20}$$

Thus different particles satisfy this definition under different conditions, i.e., at different values of φ and different distribution functions.

Continuous Distributions. If the distribution variable is denoted by ξ and its probability density function by $f(\xi)$, the differential volume fraction $\mathrm{d}\varphi$, U_f, and $U_\mathrm{d}(\xi)$ are given by

$$\mathrm{d}\varphi = \varphi f(\xi)\,\mathrm{d}\xi, \quad \int_{-\infty}^{\infty} f(\xi)\,\mathrm{d}\xi = 1 \tag{21}$$

$$U_\mathrm{f} = \varphi \int_{-\infty}^{\infty} U_\varphi(\xi) \cdot f(\xi)\,\mathrm{d}\xi \tag{22}$$

$$U_\mathrm{d}(\xi) = U_\varphi(\xi) - \varphi \int_{-\infty}^{\infty} U_\varphi(\xi) \cdot f(\xi)\,\mathrm{d}\xi \tag{23}$$

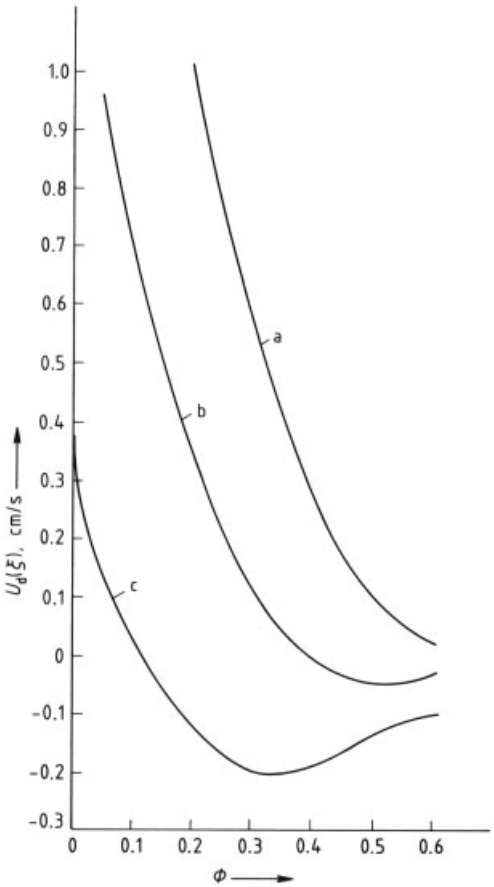

Figure 6. Calculated values of $U_\mathrm{d}(\xi)$ vs. φ for three particle sizes [11] (data derived from Fig. 5 [11])
a) $d = 2.17 \times 10^{-4}$ m; b) $d = 1.39 \times 10^{-4}$ m; c) $d = 0.61 \times 10^{-4}$ m

The total flux density in the dispersion frame of reference is then

$$J = \int_{-\infty}^{\infty} U_\mathrm{d}(\xi)\,\mathrm{d}\varphi = \varphi \int_{-\infty}^{\infty} \left[U_\varphi(\xi) - \varphi \int_{-\infty}^{\infty} U_\varphi(\xi) f(\xi)\,\mathrm{d}\xi \right] f(\xi)\,\mathrm{d}\xi \tag{24}$$

and the flux density of the $\xi_1 < \xi < \xi_2$ fraction is given by

$$J(\xi_2) - J(\xi_1) = \varphi \int_{\xi_1}^{\xi_2} \left[U_\varphi(\xi) - \varphi \int_{-\infty}^{\infty} U_\varphi(\xi) f(\xi)\,\mathrm{d}\xi \right] f(\xi)\,\mathrm{d}\xi \tag{25}$$

where $J(\xi_i)$ denotes cumulative flux density, i.e., for $-\infty < \xi \leq \xi_i$, $i = 1, 2$. Letting $\xi_1 \to -\infty$ gives $j(\xi_2)$, i.e., (without the subscript) the cumulative flux density distribution $J(\xi)$. Equations (19) and (25) show that the velocity and flux density of a given fraction depend on φ as well as on $f(\xi)$. This can significantly affect the velocity and flux of slower fractions, even to the extent of changing their direction of motion. Figure 5 [11] shows a plot of $U_\mathrm{d}(\xi)$ vs. ξ at different levels of φ, which were calculated by assuming a normal size distribution. The distribution was truncated on the left at $\xi = 0$; average particle size was $\bar{\xi} = 10^{-4}$ m, and the standard deviation was $\sigma = 3\,\bar{\xi}$. The large standard deviation was selected with a view to intensifying the effects of polydispersity on particle motion [11]. The figure shows that at the finer size end, $U_\mathrm{d}(\xi) < 0$ and that the range in which $U_\mathrm{d}(\xi) < 0$ persists expands as φ is increased. Clearly an increase in φ causes a corresponding decrease in the absolute value of $U_\mathrm{d}(\xi)$. Figure 6 shows a plot of $U_\mathrm{d}(\xi)$ vs. φ which was derived from Figure 5 for three particle sizes [11]. Reversal of the direction of velocity of 61- and 139-µm particles occurs at $\varphi = 0.12$ and $\varphi = 0.40$, respectively. The minima in the corresponding curves reflect the decrease in velocity (absolute value) occurring at higher values of φ. Figure 7 shows a plot of critical size $\xi\,[U_\mathrm{d}(\xi) = 0]$ vs. φ at two levels of standard deviation of the normal distribution. Note that the critical size is defined as one

at which the particle is stationary in the dispersion frame of reference. The value $\xi\,[U_{\mathrm{d}}(\xi)=0]$ increases with φ and also with an increase in σ. For further details, see [1] and [11].

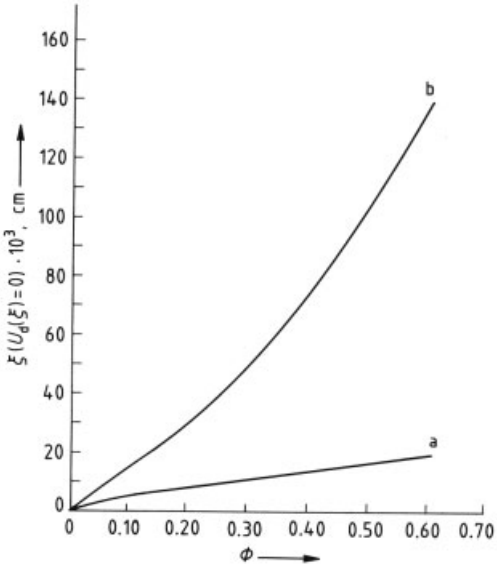

Figure 7. Critical values of particle size $\xi\,(U_{\mathrm{d}}(\xi)=0)$ vs. φ Normal size distribution, $\bar{\xi}=10^{-4}$ m, at two levels of σ [11]: a) $\sigma=3\,\bar{\xi}$; b) $\sigma=9\,\bar{\xi}$

2.1.3. Aggregated Dispersions

For details of the theory of sedimentation of aggregated dispersions [12], see → Elutriation, Chap. 3.3. The hydrodynamic hindrance function $f\,(\varphi)$ (Eq. 1) is given by Equation (26) for sedimentation of monodisperse aggregates:

$$f\,(\varphi_{\mathrm{a}})=\frac{\varphi_{\mathrm{s}}-\varphi}{\left[1+(\varphi/\varphi_{\mathrm{s}})^{1/3}\right]\exp\left(\frac{5}{3}\cdot\frac{\varphi}{\varphi_{\mathrm{s}}-\varphi}\right)},\quad \varphi_{\mathrm{s}}>\varphi \tag{26}$$

where φ_{s} is the volume fraction occupied by particulates within an aggregate and φ_{a} is the volume fraction occupied by the aggregates in the dispersion. The meaning of φ is unchanged, i.e., the volume fraction occupied by particles in the dispersion irrespective of their aggregation.

In the turbulent flow regime where Equation (2) is applied, a_1 remains unchanged, and a_2 and a_3 are given by

$$a_2=4.80\mu_{\mathrm{f}}^{1/2}\exp\left(\frac{5}{6}\cdot\frac{\varphi}{\varphi_{\mathrm{s}}-\varphi}\right)$$
$$a_3=4.0\,(\varrho_{\mathrm{p}}-\varrho_{\mathrm{f}})\,(\varphi_{\mathrm{s}}-\varphi)\cdot g\prime d^2/\left\{3\left[1+(\varphi/\varphi_{\mathrm{s}})^{1/3}\right]\right\} \tag{27}$$

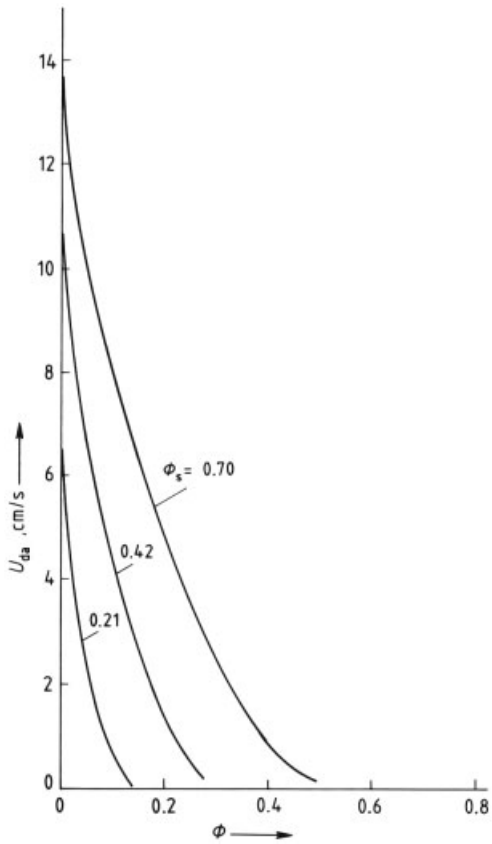

Figure 8. Calculated velocities of monodisperse aggregates vs. φ for three levels of φ_{s}, $d=183\,\mu\mathrm{m}$ [12]

For polydisperse aggregates, the same calculation procedure suggested for unaggregated particles (see Eqs. 19 – 25) can be applied, with Equations (26) and (27) replacing their counterparts in Equations (1) and (2). The following calculated results show the effects of aggregation on sedimentation velocities and flux densities in the dispersion frame of reference. For monodisperse aggregates, $d=183\,\mu\mathrm{m}$ (where d is the diameter of the aggregates), $\varrho_{\mathrm{p}}=2500\,\mathrm{kg/m^3}$, $\varrho_{\mathrm{f}}=1000\,\mathrm{kg/m^3}$, and $\mu=0.001\,\mathrm{kg\,m^{-1}\,s^{-1}}$ were used in the calculations.

Figure 8 shows a plot of U_{da} vs. φ at three levels of φ_{s}; U_{da} is the velocity of aggregates

in the dispersion frame of reference. The strong effect of φ_s is clear, and the combined effect of φ and φ_s intensifies at lower values of φ_s. The flux density of aggregated particles is given by

$$J = U_{\varphi a}\,(\varphi_s - \varphi)\,\varphi/\varphi_s \tag{28}$$

where $U_{\varphi a}$ (similar to U_φ) is the sedimentation velocity in the fluid frame of reference. Figure 9 shows a plot of J vs. φ calculated for different values of φ_s by using the data given above. Low φ_s values lower the flux densities considerably. Figure 10 shows plots of maximum flux density $J_{\max}$ vs. φ_s calculated for two aggregate sizes. The strong effect of φ_s is clear to the extent of possibly overriding the effect of aggregate size. Equation (29) relates $J_{\max}$ to φ_s and d (183 µm $< d <$ 0.01 m):

$$\ln J_{\max} = a\,(d)\,\ln\varphi_s + \ln b\,(d)\,,\ 0 < \varphi_s < 0.7 \tag{29}$$

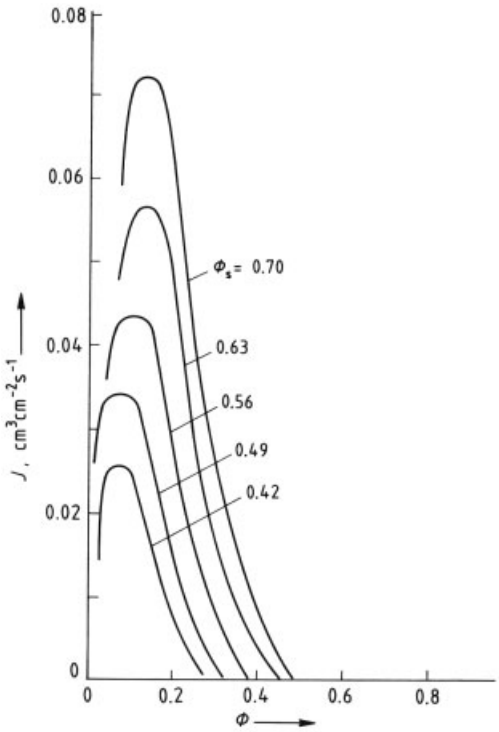

Figure 9. Calculated values of J vs. φ for different levels of φ_s, d = 183 µm [12]

Values of $a\,(d)$ and $b\,(d)$ vs. d are given in Table 1. Figure 11 shows a plot of volume fraction φ_m that yields $J_{\max}$ vs. φ_s for $0 < d < 0.01$ m and $0 < \varphi_s < 0.7$. Aggregate density is plotted on the horizontal axis. Equation (30) correlates the results well [$\varphi_m = (0.0391 \ln d + 0.340)\,\varphi_s$, 183 µm $< d <$ 0.01 m]:

$$\begin{aligned}\varphi_m &= [0.0391 \ln\,(0.0183) + 0.340]\,\varphi_s \\ &= 0.1836\,\varphi_s,\ d < 183\,\mu m\end{aligned} \tag{30}$$

Equation (30) shows that for a high packing density (φ_s = 0.6) of $d <$ 183 µm aggregates in the creeping flow regime, φ_m = 0.11 is recommended. This value is substantially lower than the value of 0.1792 that maximizes the flux of nonaggregated particles in the same flow regime (see Section 2.1.1).

Table 1. Values of a and b obtained from linear regression of Equation (29) for different aggregate sizes

d, cm	$a\,(d)$	$b\,(d)$	Correlation factor
0.01832	2.0018	0.1471	> 0.9999
0.1353	1.7205	1.8013	> 0.9999
0.3679	1.6217	4.2131	> 0.9999
1.0000	1.5638	8.3352	> 0.9999

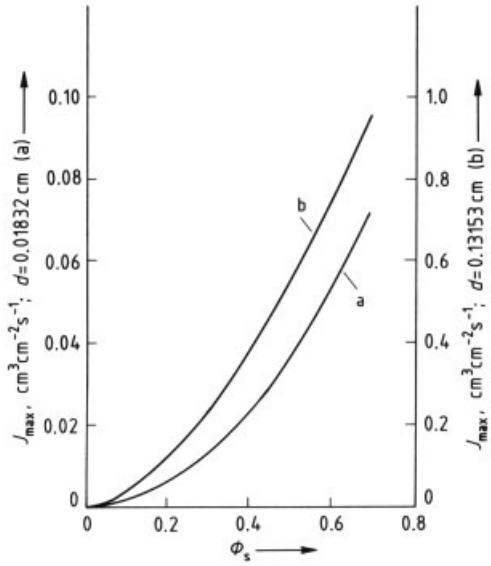

Figure 10. Calculated values of $J_{\max}$ vs. φ_s for two aggregate sizes [12]

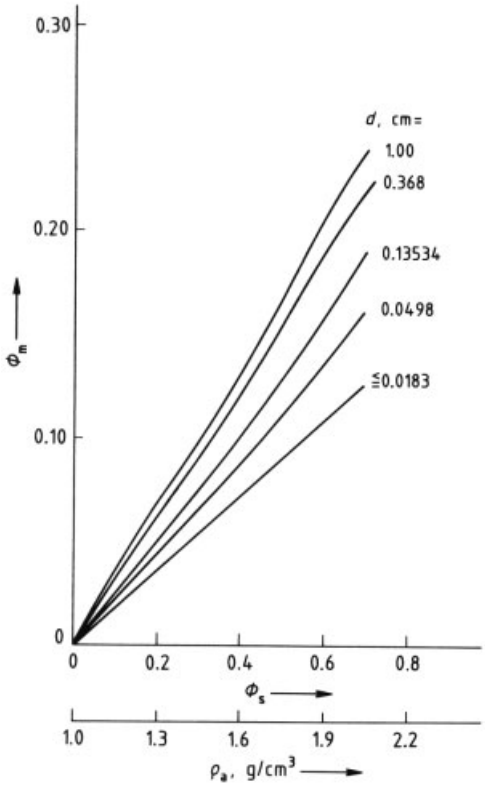

Figure 11. A plot of φ_m vs. φ_s for different aggregate sizes with $d \leq 0.01$ m [12]

Aggregation results in size enlargement, but at the same time the density of the aggregate may be lower than that of the unaggregated particles. This is particularly significant for particles that are "near-density" in the sense that their density is close to that of the fluid. Near-density particles can be found in biological, agricultural, and chemical systems, e.g., when the water content of particles (which are dispersed in water) is high. In near-density polydisperse aggregate systems, the effects of φ_s, φ, and the distribution on sedimentation are expected to be pronounced.

Figure 12 shows a plot of U_{da} vs. φ_s for different values of ϱ_p, where the diameter of the aggregates is 0.01 m, $\varphi = 0.36$, and the distribution of aggregate packing density is given below. The dominant effect of φ_s (at this high value of φ) in determining the magnitude and direction of U_{da} is demonstrated clearly for aggregates as large as 0.01 m. Figure 13 shows a plot of "partition values" $\varphi_s(U_{da} = 0)$ vs. d, calculated for $\varrho_p - \varrho_f = 400\,\mathrm{kg/m^3}$ at $\varphi = 0.36$. The discrete distribution of ϱ_s characterized by higher packing density of aggregates is given in the following material:

φ_s	0.400	0.475	0.550	0.625	0.700
$\psi(\varphi_s)*$	0.05	0.10	0.15	0.30	0.40

$*$ The parameter $\psi(\varphi_s)$ denotes the distribution of φ_s.

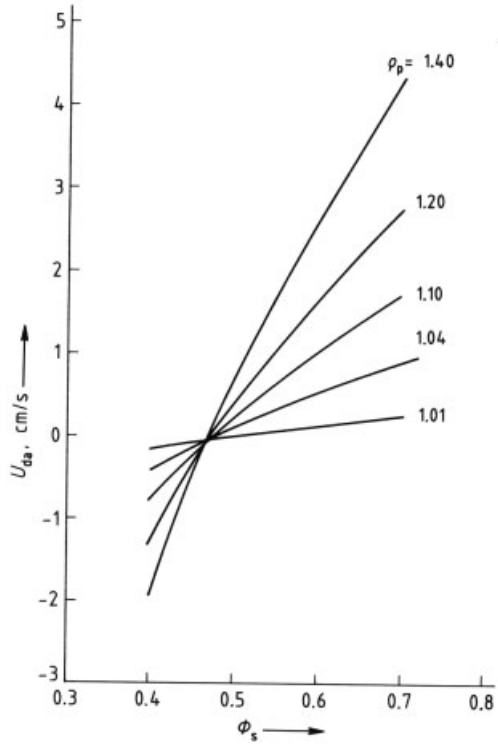

Figure 12. A plot of calculated U_{da} vs. φ_s for different values of ϱ_p, 0.01 m aggregates, $\varphi = 0.36$, distribution of packing density φ_s given below [12]

The smallest aggregates cannot undergo sedimentation at $\varphi_s = 0.54$, and decreasing φ_s to 0.46 is sufficient to prevent sedimentation of aggregates as large as 0.010 m. In summary: if higher sedimentation flux is required, then higher packing density of aggregates, with the smallest scatter about the packing density mean, should be used. The formation of large aggregates may be insufficient to produce a desirable sedimentation flux if the combination of φ_s, size, density, and the value of φ is unfavorable. For additional information, see [1].

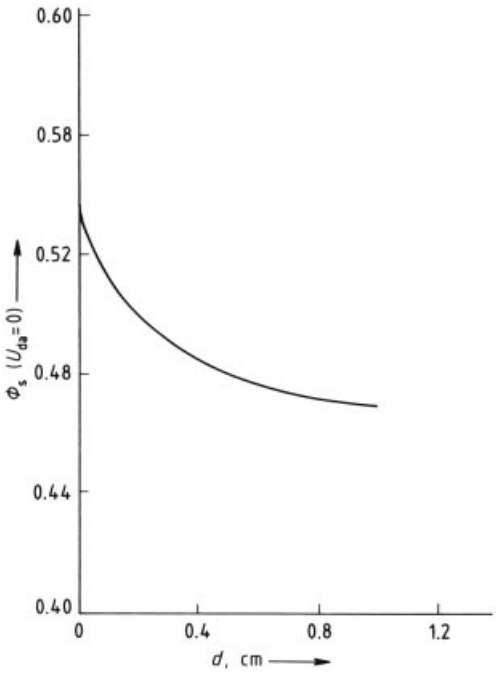

Figure 13. A plot of calculated partition values $\varphi_s(U_{da}=0)$ vs. d; $\varphi=0.36$, $\varrho_p-\varrho_f=400\,kg/m^3$ [12]

2.2. Time Dependence of Steady Sedimentation Systems

Steady sedimentation systems are supposed to be time independent. This must be true at least in one frame of reference. However, in other frames of reference the system may no longer be time independent. Figure 14 shows schematically a dispersion flowing steadily in a pipe. A force field applied across the flow produces crossflow sedimentation. For an observer in the pipe frame of reference, the flow system, excluding the rising sediment, is time independent. This frame of reference is an Eulerian representation of flow. For an observer moving with the stream (i.e., in the fluid frame of reference), particle sedimentation appears time dependent, sedimentation being a batchwise rather than a continuous process (see Fig. 14 B). A frame of reference moving with the fluid is a Lagrangian representation of flow. If the dispersion is monodisperse, then in the Lagrangian representation, the time-dependent sedimentation involves steady sedimentation. This does not hold if polydisperse mixtures are involved, where in this case nonsteady sedimentation is observed. Thus, analysis of sedimentation systems that are steady in an Eulerian representation is likely to require the use of a Lagrangian representation to uncover their nonsteady features. This applies to analysis of the multiphase flow of slurries and dispersions, reactors, contacting devices, separation processes involving particulates, and sedimentation phenomena from flowing streams in nature. Thus the ability to analyze batch nonsteady sedimentation provides the means for analysis of sedimentation in steady flow systems. In the following, nonsteady sedimentation is considered.

2.3. Nonsteady Sedimentation

2.3.1. System Characterization

Continuity must be satisfied by all parts of a distribution of a polydisperse mixture undergoing sedimentation. Thus in the conduit frame of reference and with respect to the ith fraction,

$$\nabla\,[(U_{di}+v)\,\varphi_i]+\frac{\partial\varphi_i}{\partial t}=0 \tag{31}$$

where v is the velocity of the dispersion relative to the conduit. If the distribution is continuous, then from Equations (21) and (23),

$$\nabla\,[(U_d+v)\,\varphi f\,(\xi)]+\frac{\partial\,[\varphi f\,(\xi)]}{\partial t}=0 \tag{32}$$

$$\nabla\left[\left(U_\varphi\,(\xi)-\varphi\int_{-\infty}^{\infty}U_\varphi\,(\xi)\cdot f\,(\xi)\,\mathrm{d}\xi+v\right)\varphi f\,(\xi)\right]+\frac{\partial\,[\varphi f\,(\xi)]}{\partial t}=0 \tag{33}$$

where as before ξ is a distribution parameter such as size, density, etc.

The justification for using U_φ in nonsteady sedimentation depends on the relaxation time involved in attaining this velocity. If the relaxation time for attaining, for example, 99 % of $U_\varphi(\xi)$ is significantly shorter than that associated with changes in $U_\varphi(\xi)$, then the assumption that terminal velocity conditions apply for the time- or position-dependent $U_\varphi(\xi)$ is justified. In other words, if the hydrodynamic response

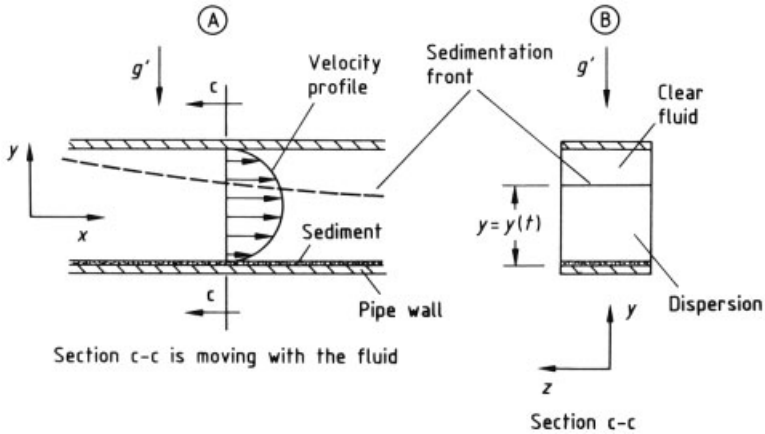

Figure 14. Steady flow of a dispersion in a pipe
A) Flow profile; B) Section c – c in the fluid frame of reference

to changes in U_φ is significantly faster than the time rate of these changes, then using terminal sedimentation velocities values for the time- and position-dependent particle velocities provides a good approximation and is justified. Hence, Equations (1) and (2) can be used for evaluation of $U_\varphi(\xi)$ in nonsteady sedimentation. If the relaxation time associated with hydrodynamic response to changes in $U_\varphi(\xi)$ is not short enough, then the use of Equations (1) and (2) may introduce an unacceptable error. Sedimentation of particles in liquids generally involves short relaxation times, whereas this is not always true for sedimentation in gases.

Simulation of nonsteady sedimentation of polydisperse mixtures requires solution of Equation (31) or (33) for each fraction ξ at every point in the dispersion at a given time. The number of simultaneous equations is $n_1 \times n_2 \times n_3$, where n_1, n_2, and n_3 denote the number of fractions, the number of segments (i.e., volume elements) into which the dispersion is divided, and the number of time increments, respectively. In the following material, an outline and sample outputs of a computer program using an algorithm of direct tracing of each sedimenting fraction are described briefly [13].

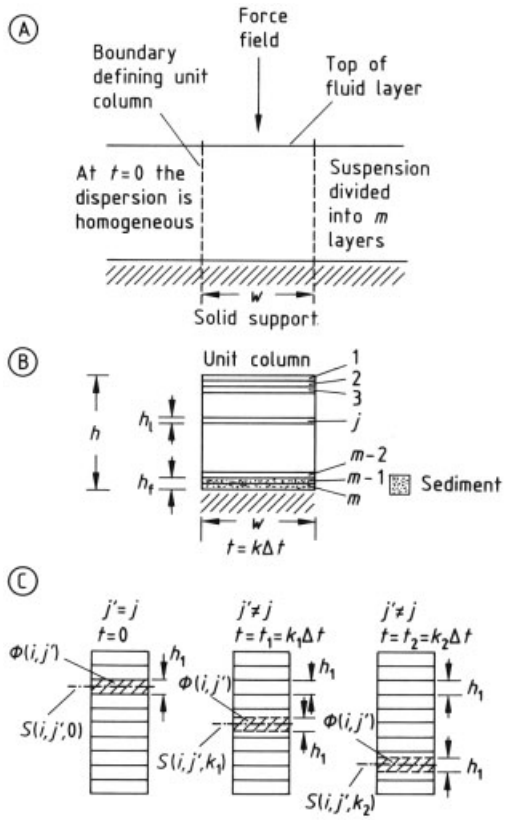

Figure 15. Characterization of the sedimentation system
A), B) Definition of a unit sedimentation column; C) Definition of sedimentation variables

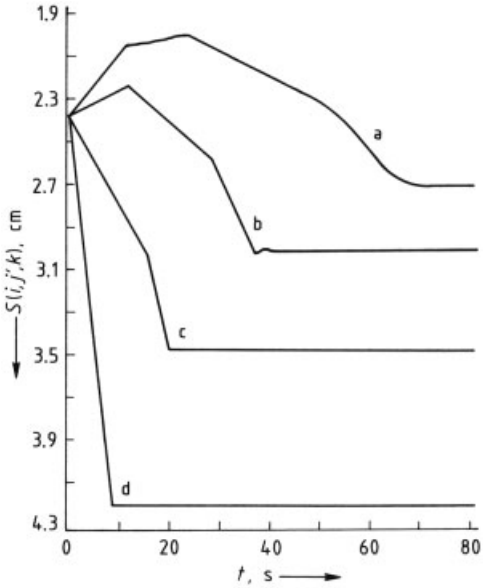

Figure 16. Trajectories of the four $\varphi(i, j')$ fractions, a) $i = 1$; b) $i = 2$; c) $i = 3$; d) $i = 4$, starting sedimentation at the tenth layer ($j' = 10$), $\varphi = 0.30$, $h = 0.05$ m

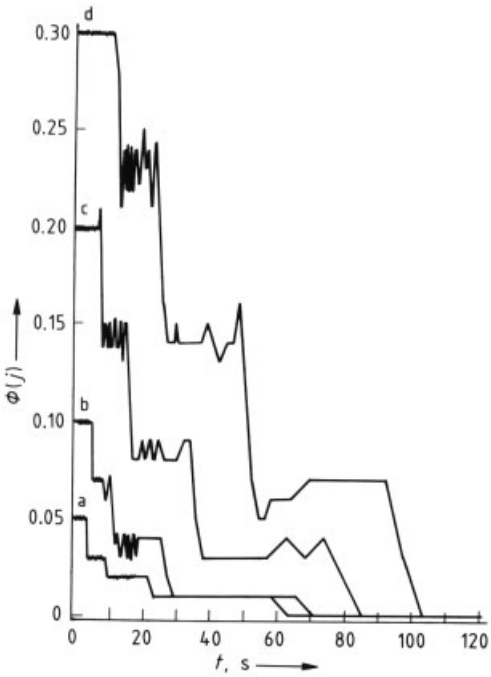

Figure 17. A plot of $\varphi(j)$ vs. t for $j = 10$ and four levels of φ:
a) $\varphi = 0.05$; b) $\varphi = 0.1$; c) $\varphi = 0.2$; d) $\varphi = 0.3$; $h = 0.05$ m

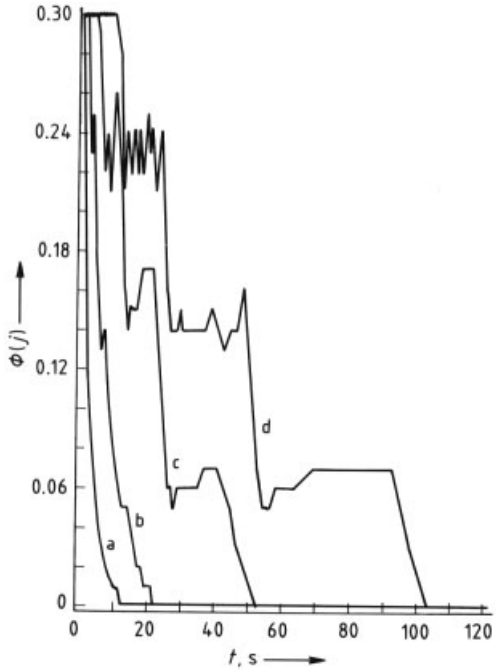

Figure 18. A plot of $\varphi(j)$ vs. t for different layers j:
a) $j = 1$; b) $j = 2$; c) $j = 3$; d) $j = 4$, $\varphi = 0.30$; $h = 0.05$ m

2.3.2. Computer Program for Simulation of Nonsteady One-Dimensional Sedimentation of Polydisperse Mixtures

The computer model incorporates a principle of tracing positions and concentrations, along the sedimentation trajectory, of all size fractions, given initial conditions and elapsed time from start of sedimentation [13]. The fluid is divided into m thin layers (of thickness h_1) which are set perpendicular to the direction of sedimentation (see Fig. 15 B). Each layer is assumed uniform with respect to composition and concentration, both being time dependent. The sedimentation system is defined with respect to two arrays: the (i, j') "primed array" and the (i, j) unprimed array (see Fig. 15 C). The (i, j') primed array is invariable and characterizes the volume fraction $\varphi(i, j')$ of the ith fraction originating at the jth layer at the onset of sedimentation ($t = 0$). The (i, j) unprimed array characterizes the time-variable volume fraction $\varphi(i, j)$ of the ith size fraction in the jth layer at time t. The fraction $\varphi(i, j')$ occupies a space of one layer and moves through the

dispersion as one invariable unit (see Fig. 15 C). The distance traveled by $\varphi(i, j')$ is denoted by $S(i, j', k)$, where k denotes the number of time increments elapsed. The fraction $\varphi(i, j)$ consists of all contributions from $\varphi(i, j')$ fractions present in the jth layer at time t. Note that a contribution of $\varphi(i, j')$ means that a nonvanishing part of it entered and remained in the jth layer at time t. The total layer concentration at time t is given by

$$\varphi(\mathrm{j}) = \sum_{i=1}^{n} \varphi(i,j)$$

where summation is over all size fractions.

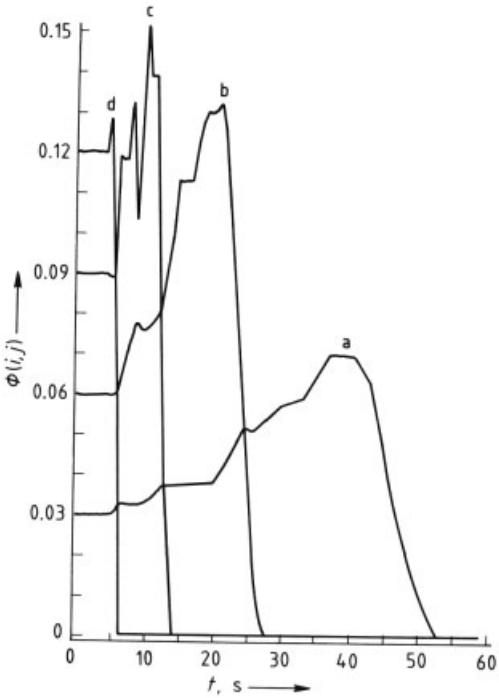

Figure 19. A plot of $\varphi(i, j)$ vs. t in the $j = 5$ layer for: a) $i = 1$; b) $i = 2$; c) $i = 3$; d) $i = 4$, $\varphi = 0.30$, $h = 0.05$ m, $\alpha = 9.0$

The tracing algorithm is rather involved; therefore, only a sample of outputs is presented here. The following data apply. One-dimensional sedimentation takes place in water at room temperature, with the number of layers $m = 20$; at the onset of sedimentation, the dispersion is uniform consisting of four $i = 1, 2, 3$, and 4 size fractions $d_i = 20 \times 9^{(i-1)/4}$ µm (giving a size range from 20 to 103.9 µm), $\varphi(i, j') = 0.1\, i$, at $t = 0$, $\varphi = 0.3$. The field is gravitational, acting perpendicular to the m layers. Packing density of the sediment is $\varphi = 0.6$. Height of the sedimentation column h is either 0.05 or 0.00625 m. High φ and skewed size distribution were selected so that the effects of concentration and distribution on sedimentation patterns would be pronounced. The computer graphical outputs provided the following information:

1) particle trajectories $S(i, j', k)$ and residence time in dispersion,
2) profiles of total concentration $\varphi(j)$ as a function of time and position,
3) profiles of specific concentrations $\varphi(i, j)$ as a function of time and position, and
4) structure and composition of sediment as a function of time and position.

2.3.3. Graphical Output

Trajectory and Composition of Dispersion and Sediment. Figure 16 shows plots of sedimentation trajectories $S(i, j', k)$ vs. time for the four size fractions positioned at the tenth layer ($j' = 10$) at $t = 0$, the column height being 0.05 m. The largest particles $d_4 = 103.9$ µm follow practically a steady sedimentation, reaching the sediment in less than 2 s. The smallest particles $d_1 = 20$ µm follow an inverted trajectory, first counterflowing the large particles and then changing direction toward the sediment. This fraction required 25 s for velocity inversion to occur, 57 s to return to the tenth layer, and 65 s to reach the sediment. The $d_2 = 34.6$ µm fraction also follows a smaller inverted trajectory. The curves disclose abrupt changes of slope. In the ith ($i = 1, 2, 3, 4$) curve, $5 - i$ such changes occur, including the final slope change due to transfer into the sediment. Thus the smallest fraction $i = 1$ shows four abrupt slope changes, the first, second, and third being due to disappearance of d_4 (at 10 s), d_3 (at 25 s), and d_2 (at 55 s) fractions from the layer in which the d_1 fraction moves. The increase in slope after its abrupt change indicates a shock, or discontinuity, occurring due to an abrupt fall in φ.

Thus the full sedimentation record of the four size fractions is available from Figure 16, i.e., starting position, trajectory, and position in the sediment, which shows a clear segregation effect.

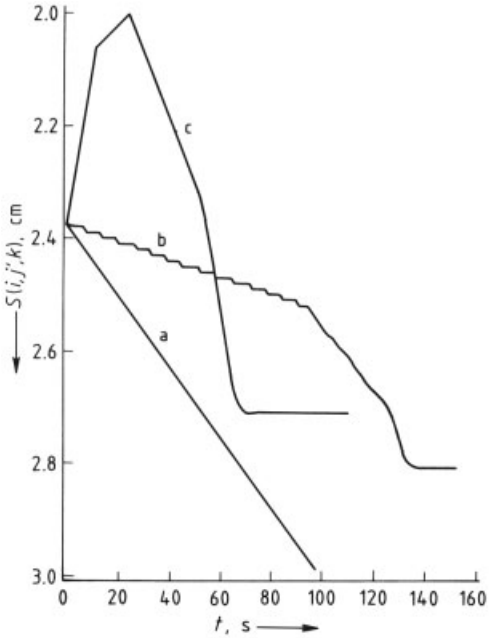

Figure 20. A plot of $S(i, j', k)$ vs. t of the 20 μm particles ($i = 1$), originating at the $j = 10$ layer, $\varphi = 0.30$, $h = 5.0$ cm, for three levels of the distribution parameters:
a) $\alpha = 1$; b) $\alpha = 2.5$; c) $\alpha = 9$

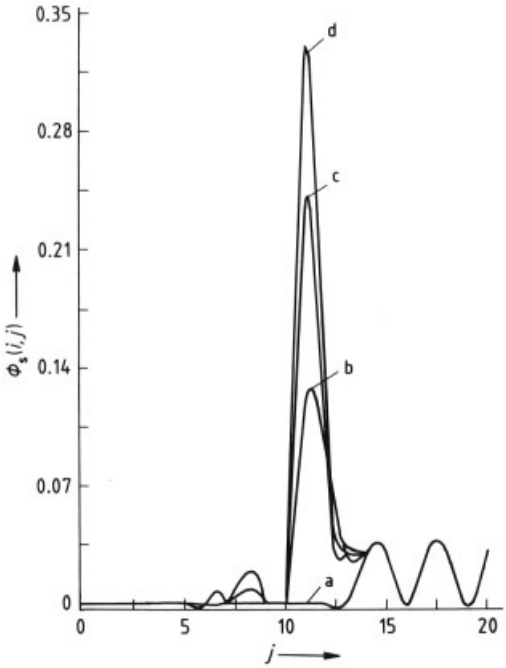

Figure 21. A plot of $\varphi_s(i, j)$ vs. j of the 20 μm particles, at different elapsed times from onset of sedimentation:
a) $t = 3.8$ s; b) $t = 10.1$ s; c) $t = 12.2$ s; d) $t = 15$ s, $\varphi = 0.30$, $h = 0.625$ cm, $\alpha = 9.0$

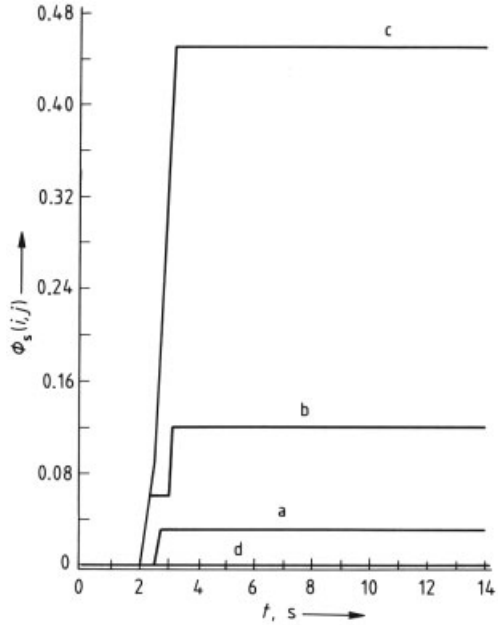

Figure 22. A plot of $\varphi_s(i, j)$ vs. t for:
a) $i = 1$; b) $i = 2$; c) $i = 3$; d) $i = 4$, $\varphi = 0.30$, $h = 0.625$ cm, $\alpha = 9.0$

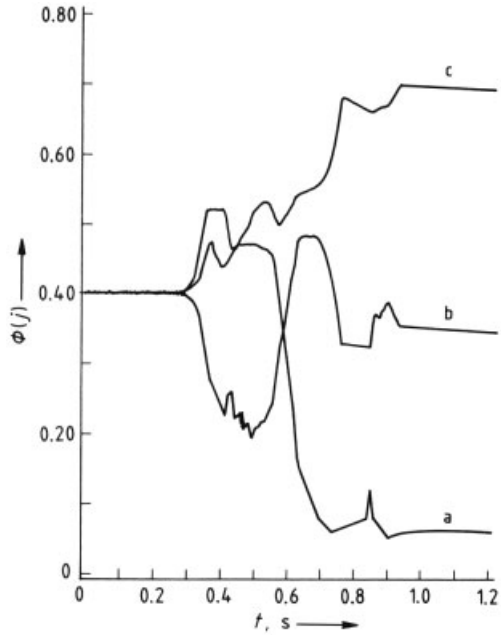

Figure 23. A plot of $\varphi(j)$ vs. t in the
a) $j = 9$; b) $j = 10$; c) $j = 11$ layers, $d_i = 60 \times \alpha^{(i-1)/4}$, $\alpha = 1.05$, $\varphi = 0.40$, $h = 5.0$ cm, sedimentation in air

Figure 17 is a plot of $\varphi(j)$ vs. t in the tenth layer $j = 10$, for four levels of initial (i.e., at $t = 0$) concentration: 0.05, 0.10, 0.20, and 0.30. A stepwise decrease of $\varphi(j)$ can be observed for the four levels of φ. Furthermore, in a given step, $\varphi(j)$ exhibits a train of concentration waves. Each step corresponds to partial or full exit of one size fraction from the tenth layer, the largest first and the smallest last. The concentration waves are an integrated effect of the contribution from each size fraction, with their amplitudes diminishing after a decrease in φ. The train of concentration waves intensifies with increasing layer depth, i.e., larger j. An example is shown in Figure 18 where $\varphi(j)$ vs. t is plotted for different layers j and $\varphi = 0.3$. In the first layer $j = 1$, the amplitude of the concentration waves is small, whereas in the tenth layer $j = 10$, large waves exist. Furthermore, in the upper layers ($j = 1$ and 2), sedimentation steps [wherein $\varphi(j)$ fluctuates] are small, whereas in lower ones ($j = 5$ and 10), they expand considerably. A plot of $\varphi(i, j)$ vs. t for the four size fractions $i = 1$, 2, 3, and 4, in the fifth $j = 5$ layer, given $\varphi = 0.3$, is shown in Figure 19. Observe that at $t = 0$, $\varphi(i, j) = 0.03 \times i$; for example, the initial volume fraction of the $d_1 = 20\,\mu$m, $i = 1$ fraction in the fifth layer was 0.03. Time-variable composition occurs due to accumulation and size segregation effects, prior to depletion of the layer from particles. For example, after 20 s the layer consists of $\varphi(1, 5) = 0.04$, $\varphi(2, 5) = 0.13$, $\varphi(3, 5) = \varphi(4,5) = 0$ fractions, indicating depletion of the third and fourth fractions, with simultaneous upgrading of the second and first fractions. Concentration waves of individual size fractions pass the fifth layer, increasing in relative intensity for smaller particle sizes. The maximum amplitude of these waves is of the order of twice the initial concentration. Superposition of the individual concentration waves yields total concentration waves of the type shown in Figure 18. The individual as well as total concentration waves expand as they move down the layers. This has been verified by outputs not shown here. The effect of the range of size distribution in the polydisperse mixture was tested by using a size parameter α, with the size range being defined by $d_i = 20\,\alpha^{(i-1)/4}\,\mu$m for $\alpha = 1.0$, 2.5, and 9.0. Figure 20 shows the variation of particle trajectories $S(i, j', k)$ of the $i = 1$ fraction vs. time, for the three values of α tested. At $t = 0$, particles are in the tenth layer and $\varphi = 0.3$. The $\alpha = 1.0$ curve corresponding to monodisperse 20 μm particles is a straight line as expected. Slope transitions, but no inversions, can be observed in the $\alpha = 2.5$ curve. Increasing α to 9.0 (size range from 20 to 103.9 μm) produces an inverted trajectory with four slope transitions (as also shown by the highest curve in Fig. 16). Increasing α from $\alpha = 1$ decreases the overall sedimentation velocity initially, and then the reverse occurs. Thus the sedimentation of the 20-μm particles is slower at $\alpha = 2.5$ than at $\alpha = 9.0$. The computer program also provided information regarding composition of the sediment and effects of perturbation of concentration in a given position on the stability of the sedimentation system.

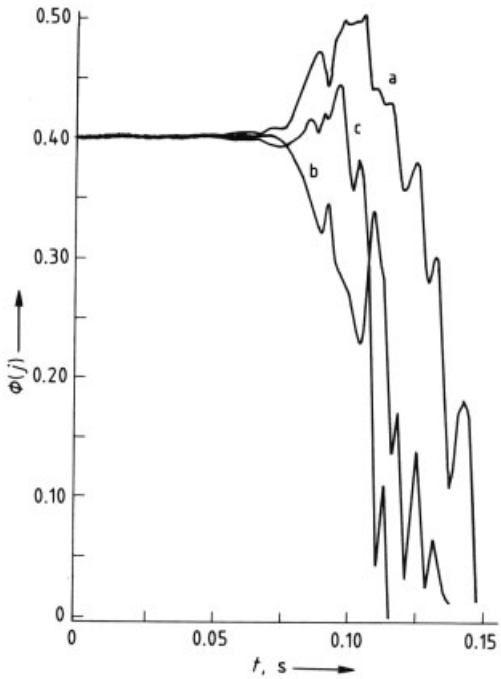

Figure 24. A plot of $\varphi(j)$ vs. t, data as in Figure 23 except for $\alpha = 3.0$

Figure 21 shows the volume fraction of the 20-μm fraction $i = 1$ in the sediment $\varphi_s(i, j)$ vs. the layer number j for different elapsed times of sedimentation, $\alpha = 9$, $h = 0.625$ cm, $\varphi = 0.30$. Accumulation of this fraction around the $j = 11$ layer (and other fractions around other layers, see Fig. 22) increases with time, indicating a size segregation effect. Figure 22 shows a plot of $\varphi_s(i, j)$ vs. t in the $j = 14$ layer; the rest

of the data remain unchanged. The figure shows that after 3 s the sediment in the 14th layer consists of 75 % $i = 3$ particles, 20 % $i = 2$ particles, 5 % $i = 1$ particles, and no $i = 4$ particles. Compositions of other sediment layers can be obtained in a similar manner.

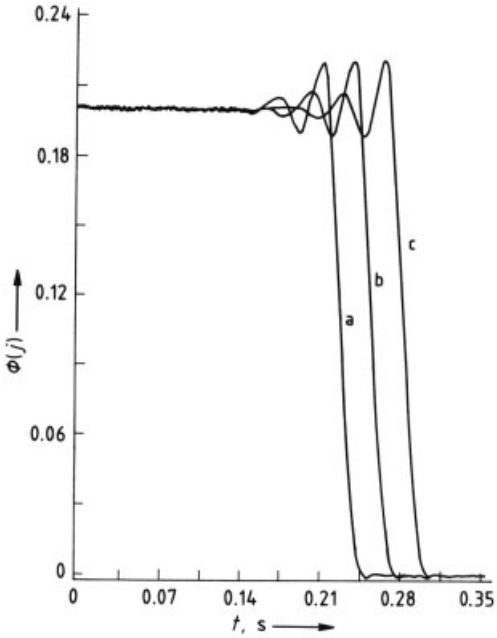

Figure 25. A plot of $\varphi(j)$ vs. t, data as in Figure 23 except for $\varphi = 0.20$

Effect of Distribution and Total Volume Fraction on Sedimentation Behavior. The effect of size range on sedimentation trajectories of particles in water is illustrated in Figure 20. High φ dispersions of polydisperse mixtures in air are characteristic of fluidization and sedimentation systems in which volume fractions occupied by particles range typically from 0.4 to 0.6. High φ dispersions are affected mainly by deviation of size distribution from uniformity, being prone to instabilities caused by perturbation of local concentrations. In the following examples, size distributions that satisfy $d = 60 \times \alpha^{(i-1)/4}\,\mu\text{m}$ and $i = 1$, 2, 3, and 4 are used, the fluid being air $\mu_f = 182 \times 10^{-7}\,\text{kg m}^{-1}\,\text{s}^{-1}$, $\varrho_f = 1000\,\text{kg/m}^3$, $\varrho_p = 3000\,\text{kg/m}^3$. Figures 23 and 24 contain plots of $\varphi(j)$ vs. t for the $j = 9$, 10, and 11 layers, and $\alpha = 1.05$ and 3.0, respectively. At $t = 0$, $\varphi = 0.40$.

Figure 23 shows that a narrow size distribution (i.e., low α) and high value of φ generate strong fluctuations of $\varphi(j)$ in the $j = 9$ and 10 layers, and particle accumulation in the $j = 11$ layer. At $t = 0.4$ s, $\varphi(9)$ increases above 0.50, and at $t = 0.8$ s, the eleventh layer becomes practically blocked with a dense phase of particles $\varphi(11) \geq 0.6$. After 0.9 s, the ninth layer is almost depleted of particles, the tenth layer remains close to $\varphi(10) \simeq 0.40$, and the eleventh layer is highly packed with particles, i.e., $\varphi(11) \geq 0.6$. Fluidization of such polydisperse particle system would be inherently unstable, since no single sedimentation velocity can be balanced by the drag of the fluidizing gas. Higher packing density (i.e., $\varphi(j)$) in a given layer decreases sedimentation velocity there. Hence, if the velocity of the fluidizing gas is unchanged, material in this layer will be lifted and probably separated from lower, less dense layers. This can be related to the phenomenon of slugging which is well documented in fluidization engineering [14], [15].

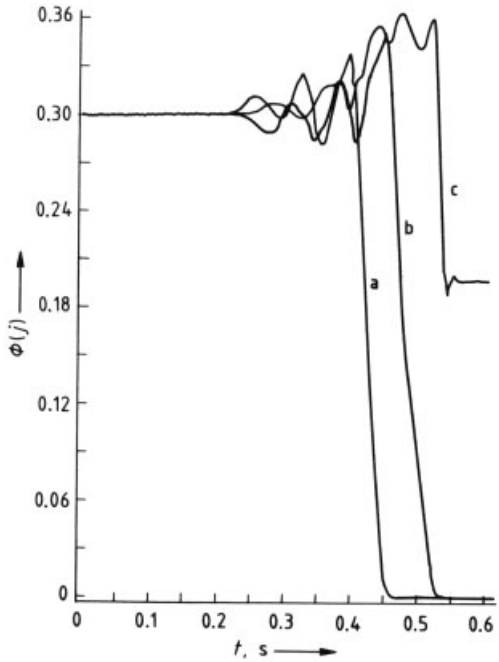

Figure 26. A plot of $\varphi(j)$ vs. t, data as in Figure 23 except for $\varphi = 0.30$

Figure 24 shows similar results of $\varphi(j)$ vs. t but for a wider size distribution (i.e., $\alpha = 3.0$). Here the largest increase [to $\varphi(j) \simeq 0.50$] in con-

centration occurs in the $j = 9$ layer after 0.10 s of sedimentation. This rise may be sufficient to effect instability, but the overall effect has diminished compared to $\alpha = 1.05$ (Fig. 23). Further increase of α to 9.0 (not shown here) eliminates the surging of concentration. This indicates that strong sedimentation instability is expected in dispersions that are characterized by high initial particle concentration and narrow size distribution. Thus the computer model for nonsteady sedimentation of polydisperse mixtures predicts that for high concentration and narrow distribution, instability is an inherent property of the sedimentation system.

Figures 25, 26, and 23 show for $\alpha = 1.05$ the effect of increasing φ on the formation of surging concentrations in the $j = 9$, 10, and 11 layers. The rest of the data remain unchanged. If at $t = 0$, $\varphi = 0.20$, the maximum surge (occurring in all three layers) in $\varphi(j)$ is 10 % above its initial value (see Fig. 25). The surge intensifies at $\varphi = 0.30$, reaching in the $j = 11$ layer a 20 % increase in concentration over the initial value of 0.30 (see Fig. 26). Further increase of φ (for the same $\alpha = 1.05$ size distribution) to 0.4 produces the effect already shown in Figure 23, where the $j = 11$ layer accumulates particles to the extent of becoming a sort of plug capable of obstructing further sedimentation. The

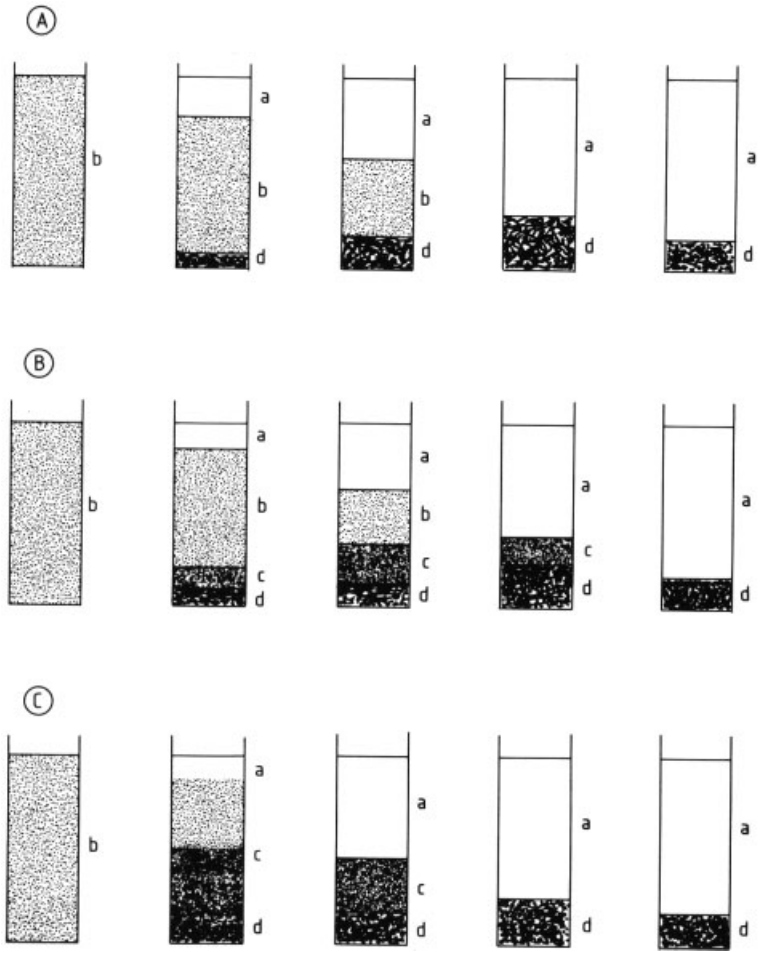

Figure 27. Effect of particle size distribution on profile of batch sedimentation [17]
A) Uniform size and density; B) Narrow size distribution; C) Wide size distribution
a) Clear liquid; b) Dispersion; c) Polydisperse mixture; d) Sediment

above results agree with the strong dependence of velocity and particle flux on φ; see Equations (1), (2), (13), (14), and (24), and Figures 3 and 6, in particular, for sedimentation in the laminar flow regime. Self- or *internal perturbation* occurs in polydisperse mixtures, which as shown here can produce substantial instability due to surges in φ. Such surges can develop into a self-accelerating process in which further increase in φ results from the incoming flux being increasingly higher than the one capable of leaving. Clearly, perturbations induced by external sources (i.e., *external perturbations*) are expected to produce similar concentration-dependent instabilities. Further computer outputs (not shown here) verify that the sedimentation system becomes more sensitive to external perturbation as the particle concentration (prior to the perturbation) increases. The complexity of nonsteady batch sedimentation has been known for a long time [16]. Figure 27 [17] shows three categories of batch sedimentation defined for uniform size and density (A), narrow size distribution (B), and wide size distribution (C). Figure 28 describes [2] the formation of discontinuities of composition, or "shock," in batch sedimentation of a polydisperse particle system consisting of three size fractions. The existence of such discontinuities is predicted by the computer model which, as shown above, also provides in-depth information about the internal structure and dynamics of nonsteady sedimentation as a function of position and time.

2.3.4. Transient Sedimentation

Definition. Transient sedimentation means nonsteady sedimentation of particles under the action of constant driving forces. If the driving forces vary along the sedimentation trajectory, then transient sedimentation is associated with velocity changes at every point along the trajectory. When the relaxation time associated with transient motion is sufficiently small relative to that associated with velocity changes, the effect of transient sedimentation may be neglected. This is assumed in earlier sections (Section 2.3.3) of this article.

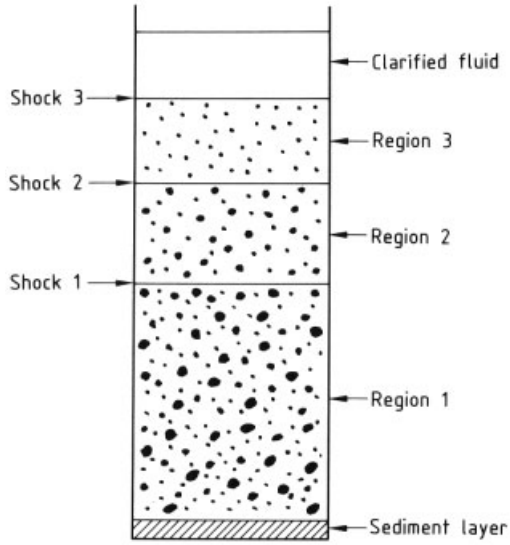

Figure 28. Formation of distinct regions that are separated by composition discontinuities ("shocks") in batch sedimentation of a mixture of three different particle fractions. Region 1 contains all three fractions, region 2 is devoid of the fastest settling particles, and region 3 contains only the slowest settling particles [2]

Equation of Motion. Newton's law of motion reads

$$\sum_j \mathrm{d}\,(\mathrm{m}U)_j\,/\mathrm{d}t = \sum_i F_i \tag{34}$$

where the summation is over all forces F_i acting on accelerated masses m_j whose velocity is U_j at time t. All masses directly or indirectly accelerated by the driving forces are included in the momentum balance equation. The forces acting on a particle can be divided into body and surface forces, as well as driving and retarding forces. If the motion of particles is coupled with that of the fluid, then the accelerated mass of the fluid must be also accounted for in Equation (34). This results in a so-called added or virtual mass [18–22], which accounts for the fact that part of the driving force must be expected to accelerate the fluid displaced by motion of the particles. The significance of this added mass becomes apparent when the density of the fluid is comparable to, or higher than, that of the particles, an extreme example being dispersion of bubbles in a liquid. The average velocity of displaced fluid counterflowing particles in the dispersion frame of reference is given (see Eqs. 7 and 8) by

$$U_f = U_d \varphi / (1-\varphi) \tag{35}$$

In a given volume V of dispersion, the mass of flowing fluid is $m_f = \varrho_f V(1-\varphi)$; hence the average fluid momentum in V is $m_f U_f = \varrho_f V \varphi U_d$. Thus the fluid momentum per particle is $\varrho_f V_p U_d$, where V_p is particle volume and the number of particles in V is $V\varphi/V_p$. Hence, the equation of motion of monodisperse particles in a fluid is given by

$$d\,[V_p(\varrho_p + \varrho_f) U_d]/dt = \sum_i F_i \tag{36}$$

If V_p, ϱ_f, and ϱ_p are constants, then Equation (36) reduces to

$$V_p \varrho_p (1 + \varrho_f/\varrho_p)\, dU_d/dt = \sum_i F_i \tag{37}$$

Equation (37) applies to a sedimentation system at constant total volume of incompressible particles and fluid. If sedimentation involves variable density of either particles or fluid, or variable volume of particles, then Equation (36) must be applied. Changes in particle density and size can occur due to aggregation, flocculation, and crystallization (or dissolution) processes. Properties of the fluid can be changed by thermodynamic means such as temperature gradients. For nonsteady sedimentation of polydisperse mixtures, Equation (36) must be applied to each fraction of the distribution. Thus for a distribution consisting of n size fractions, n equations are required to describe the system.

Driving Forces. At steady state, the local properties of a dispersion consisting of two phases can be expressed as a weighted average of properties of the individual phases:

$$\gamma_\varphi = \varphi \gamma_p + (1-\varphi)\gamma_f \tag{38}$$

where γ_φ, γ_p, and γ_f are the values of this property in a given small volume of the dispersion for the dispersion, for particles, and for fluid, respectively. Equation (5) is a special case of Equation (38), i.e., for $\gamma = \varrho$. The postulated steady state implies that the forces exerted on the particles are transmitted fully to the fluid, thus generating pressure gradients having the same effect as Archimedes levitational force. If the degree of force transmission $\beta(t)$ is time dependent, then γ_φ is given by

$$\gamma_\varphi = \beta(t)\varphi\gamma_p + (1-\beta(t)\varphi)\gamma_f,\ 0 \le \beta(t) \le 1 \tag{39}$$

$$\beta(t) = F_D/F_\gamma = F_D/(\varrho_p V_p g'),\ 0 \le \beta(t) \le 1 \tag{40}$$

where F_D and F_γ are the drag force and the force associated with the property γ, respectively, and g' (see Eqs. 1–3) is acceleration due to F_γ. Thus at the onset of sedimentation when no drag exists $\beta(t) = 0$, whereas at terminal velocity conditions $\beta(t) = 1$. The following categories of driving forces may be used in sedimentation systems.

Gravitational Force:

$$F_G = (\varrho_p - \varrho_\varphi) V_p g \tag{41}$$

At the onset of sedimentation of monodisperse particles $F_G = (\varrho_p - \varrho_f) V_p g$, whereas at steady-state terminal velocity conditions $F_G = (\varrho_p - \varrho_f)(1-\varphi) V_p g$. Thus full transmission of the particle weight to the fluid results in a decrease of F_G by a factor of $1-\varphi$.

Centrifugal Force:

$$F_\omega = (\varrho_p - \varrho_\varphi) V_p \omega^2 r \tag{42}$$

where r is the radius of rotation and ω is the angular velocity.

Electrical Charge Force:

$$F_q = \int_{V_p} E(q_v - q_{v\varphi})\, dV_p + \int_{S_p} E(q_s - q_{s\varphi})\, dS_p \tag{43}$$

where q_v and q_s are charge densities per unit volume and unit area of the particles. For monodisperse particle mixtures at $\beta(t) = 1$,

$$q_{v\varphi} = \varphi q_{vp} + (1-\varphi) q_{vf},$$
$$q_{s\varphi} = \varphi q_{sp} + (1-\varphi) q_{sf}$$

where subindex f denotes fluid.

Forces due to Magnetic Polarization:

$$F_M = \mu_0 V_p \nabla [(M_p - M_\varphi) \cdot H] \tag{44}$$

For monodisperse particle mixtures and $\beta(t) = 1$, $M_\varphi = \varphi M_p + (1-\varphi) M_f$; hence, $F_M = \mu_0 V_p \nabla [(1-\varphi)(M_p - M_f) \cdot H]$.

Forces due to Electrical Polarization:

$$F_{\mathrm{E}} = V_{\mathrm{p}} \nabla \left[\left(P_{\mathrm{Ep}} - P_{\mathrm{E}\varphi} \right) \cdot E \right] \tag{45}$$

For monodisperse particle mixtures and $\beta(t) = 1$, $P_{\mathrm{E}\varphi} = \varphi P_{\mathrm{Ep}} + (1-\varphi) P_{\mathrm{Ef}}$; hence, $F_{\mathrm{E}} = V_{\mathrm{p}} \nabla \left[(1-\varphi) (P_{\mathrm{Ep}} - P_{\mathrm{Ef}}) \cdot E \right]$.

Other forces such as those arising from interaction of moving charges (i.e., currents) and magnetic fields may also be relevant.

Note that for polydisperse mixtures, Equation (38) can be applied, with γ_{p} being replaced by its average value $\bar{\gamma}_{\mathrm{p}} = \frac{1}{\varphi} \int_0^{\varphi} \gamma_{\mathrm{p}} \, \mathrm{d}\varphi$. For example, in Equation (41), $\varrho_{\mathrm{pi}} - \varrho_{\varphi} = \varrho_{\mathrm{pi}} - \left[\varphi \bar{\varrho}_{\mathrm{p}} + (1-\varphi) \varrho_{\mathrm{f}} \right]$ should be used for the ith fraction.

Retarding Forces. Equations (46) and (47) give the *hydrodynamic drag force* exerted on a spherical particle in the laminar $Re < 1$ and general flow regimes, respectively.

$$F_{\mathrm{D}} = -3\pi d \mu'_{\varphi} U_{\varphi}, \; \mu'_{\varphi} = \mu_{\varphi} \left(1 + \varphi^{1/3} \right),$$
$$\mu_{\varphi} = \mu_{\mathrm{f}} \exp \left(\frac{5}{3} \cdot \frac{\varphi}{1-\varphi} \right) \tag{46}$$

$$F_{\mathrm{D}} = -\frac{\pi}{8} \left[0.63 \left(U_{\varphi} \, \mathrm{d}\varrho_{\mathrm{f}} \right)^{1/2} + 4.80 \mu_{\varphi}^{1/2} \right]^2 \cdot d \left(1 + \varphi^{1/3} \right) U_{\varphi} \tag{47}$$

In Equations (46) and (47), F_{D} is the drag force existing in a fully developed flow around the particle, which corresponds to terminal velocity (here U_{φ}) flow conditions. Under transient flow conditions (such as those existing in turbulent flow) where flow around the particles is not steady, the history of flow may be significant. This is accounted for by an additional term from Basset [23], [24]. The effects of Basset force and virtual mass on the amplitude of particle oscillations in an oscillating flow were reported in [25].

Force due to Wall Hindrance:

$$F_{\mathrm{W}} = -f \left(d / D_{\mathrm{c}} \right) \tag{48}$$

where f denotes a function and D_{c} is the diameter or characteristic distance between the walls. The particle wall hindrance effect is accounted for in F_{D} by the factor $1 + \varphi^{1/3}$ [3].

Reaction due to Coupled Acceleration of Displaced Fluid:

$$F_{\mathrm{R}} = \int_{m_{\mathrm{f}}} a_{\mathrm{f}} \, \mathrm{d}m_{\mathrm{f}} \tag{49}$$

where a_{f} is the acceleration of the differential fluid mass $\mathrm{d}m_{\mathrm{f}}$. In polydisperse mixtures, the motion of faster particles is expected to displace not only fluid but also slower particles. Hence, the coupled acceleration of displaced particles must be taken into account. For nonspherical particles, appropriate shape factors should be applied to d, where d is the diameter of an equivalent sphere [26], [27].

Solution of Equation of Motion of Monodisperse Particles in a Constant Force Field. If the driving forces are expressed as in Equation (3) by an equivalent acceleration g' and retarding forces are exclusively hydrodynamic, then Equation (37) takes the following form:

$$V_{\mathrm{p}} \varrho_{\mathrm{p}} \left(1 + \varrho_{\mathrm{f}} / \varrho_{\mathrm{p}} \right) (1-\varphi) \, \mathrm{d}U_{\varphi} / \mathrm{d}t = \left(\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}} \right) (1-\varphi) V_{\mathrm{p}} g' + F_{\mathrm{D}} \tag{50}$$

where $U_{\mathrm{d}} = (1-\varphi) U_{\varphi}$ (see Equation 7) and $\varrho_{\mathrm{p}} - \varrho_{\varphi} = (\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}}) (1-\varphi)$ if $\beta(t) = 1$. Solution of Equation (50) in the laminar flow regime, where F_{D} is given by Equation (46) gives

$$U_{\mathrm{d}}(t) = (1-\varphi) U_{\varphi} \left[1 - \exp(-At) \right] \tag{51}$$

$$A = \frac{18 \mu'_{\varphi}}{(\varrho_{\mathrm{p}} + \varrho_{\mathrm{f}}) d^2 (1-\varphi)}, \quad Re < 1$$

where U_{φ} is given by Equation (1) and μ'_{φ} is defined in Equation (46).

The distance $S(t)$ traveled at t is then [10]

$$S(t) = U_{\varphi} (1-\varphi) \left\{ t - A^{-1} \left[1 - \exp(-At) \right] \right\}, \quad Re < 1 \tag{52}$$

In the general flow regime (i.e., for $Re > 0$), the equation of motion is given by

$$\mathrm{d}U_{\varphi} / \mathrm{d}t = \alpha_1 + \alpha_2 \left(\alpha_3 U_{\varphi}^{1/2} + \alpha_4 \right)^2 U_{\varphi} \tag{53}$$

$$\alpha_1 = \left(\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}} \right) \left(\varrho_{\mathrm{p}} + \varrho_{\mathrm{f}} \right)^{-1} g'$$
$$\alpha_2 = (3/4) \left(1 + \varphi^{1/3} \right) \left[\left(\varrho_{\mathrm{p}} + \varrho_{\mathrm{f}} \right) d^2 (1-\varphi) \right]^{-1}$$
$$\alpha_3 = 0.63 \left(d \varrho_{\mathrm{f}} \right)^{1/2}$$
$$\alpha_4 = 4.80 \mu_{\varphi}^{1/2}$$

The solution of Equation (53) is given in [10].

Figures 29 and 30 show a plot of $U_d(t)/U_d(\infty)$ vs. t for $d = 243\,\mu m$, $\varrho_p = 3000\,kg/m^3$ particles settling in a gravitational field in air (18 °C) and water (25 °C), respectively. Clearly, transient sedimentation can be significant in air for it takes (at $\varphi = 0$) the 243-µm particles nearly 1 s to almost reach their terminal velocity $U_d(\infty)$. During this time, the particles covered approximately 4 m. Increased φ decreases relaxation time of the transient sedimentation. Increasing φ from 0.0 to 0.56, a value in the range of φ used in fluidized beds, decreases relaxation time by an order of magnitude. Changing the fluid from air to water also results in a decrease of relaxation time by more than an order of magnitude. In this example, 0.03 s is required for the 243-µm particles to attain nearly their terminal velocity, i.e., $U_d(t)/U_d(\infty) > 0.99$. The inverse of the slope of the $\varphi = 0$ curve (Fig. 30) at $U_d(t)/U_d(\infty) = 0.5$ is 0.0168 s. If the time Δt required for a driving force to double the velocity of a particle [e.g., from $0.4\,U_d(\infty)$ to $0.8\,U_d(\infty)$] satisfies $\Delta t \gg 0.0168$ s, then neglecting the effect of transient sedimentation is justified, and terminal velocity conditions can be assumed along the sedimentation path. Alternatively, if the time t required for a particle to attain a given velocity U_d in a variable force field satisfies $t \gg t_0$, where t_0 is the time required to attain U_d during transient sedimentation in a constant force field, then the effect of transient sedimentation can be neglected, and the assumption that $\beta(t) = 1$ is justified.

2.3.5. Sedimentation in Turbulent Suspension

Hitherto, the flow regime has been postulated to be solely dependent on the motion of particles relative to the fluid. Thus the nature of flow around the particles determines whether it is laminar or turbulent. However, the suspension may be set in motion by external means such as agitators and pumps, in which case the flow is different from that induced by particle motion. In the extreme, for a completely turbulent suspension, the flow in the vicinity of the particles may be totally independent of their motion. Examples are turbulent agitation so as to maintain a uniform particle suspension, turbulent slurry transportation in pipes, and turbulent pneumatic conveying. In such cases it can be assumed that the fully turbulent drag coefficient applies; hence, the drag force is given by

$$F_D = -\frac{\pi}{8} f_D \varrho_f U_\varphi^2 d^2 \left(1 + \varphi^{1/3}\right) \tag{54}$$

where f_D is the drag coefficient, having a value of 0.40 at high values of Re for the particle. If a time-averaged steady state is assumed and $V_p = \pi d^3/6$ Equation (50) gives

$$U_\varphi = \left[\frac{4}{3} \cdot \frac{gd}{f_D} \cdot \frac{(\varrho_p - \varrho_f)}{\varrho_f} \cdot \frac{(1-\varphi)}{\left(1+\varphi^{1/3}\right)}\right]^{1/2} \tag{55}$$

where turbulence is postulated to eliminate only the effect of viscosity, not the hydrostatic (see Eq. 41) and the interparticle wall hindrance (see Eq. 48) effects.

Because f_D is taken at its lowest value, the highest values of U_φ are obtained for a given driving force field. In this sense, Equation (55) provides a conservative evaluation of U_φ.

2.4. Sedimentation of Diffusing Particulates

2.4.1. Monodisperse System

The diffusion equation in the dispersion frame of reference can be derived from continuity as follows:

$$\partial C/\partial t + \nabla \cdot J = 0 \tag{56}$$

$$J = J_0 + J_F \tag{57}$$

$$J_0 = -D' \nabla C \tag{58}$$

$$J_F = U_d C \tag{59}$$

where C is concentration, D' is the diffusion coefficient, and J_D and J_F are flux densities due to diffusion and sedimentation drift in the force field, respectively; C is related to φ by

$$\varphi = C v, \quad v = \tilde{N} V_p \tag{60}$$

where N and v are the Avogadro number and the partial molar volume, respectively. Combining Equation (56) – (59) with Equations (7) and (60) gives

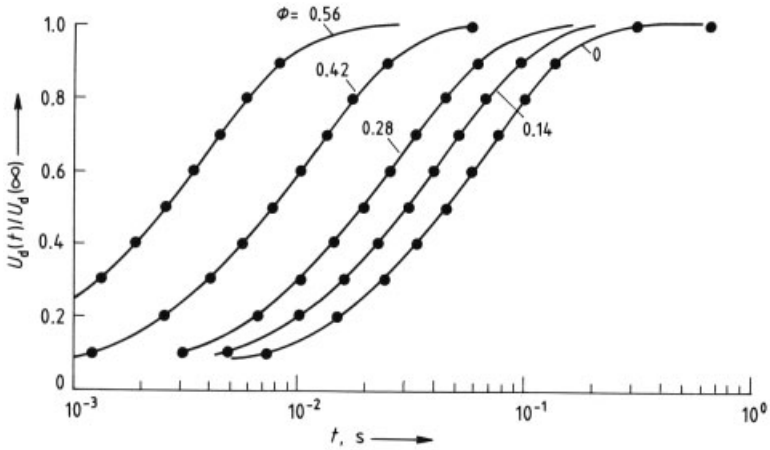

Figure 29. A calculated plot of $U_{\mathrm{d}}(t)/(U_{\mathrm{d}}(\infty)$ vs. t for different values of φ, gravitational settling in air (18 °C), $d = 243\,\mu\mathrm{m}$, $\varrho_{\mathrm{p}} = 3000\,\mathrm{kg/m^3}$

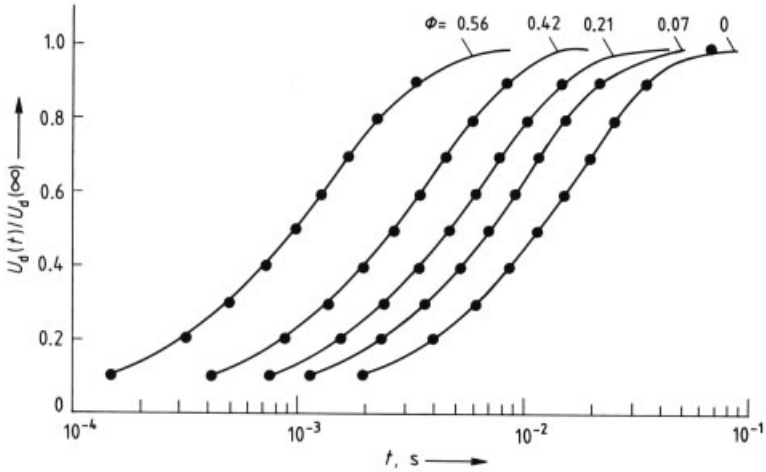

Figure 30. A calculated plot of $U_{\mathrm{d}}(t)/U_{\mathrm{d}}(\infty)$ vs. t for different values of φ, gravitational settling in water (25 °C) $d = 243\,\mu\mathrm{m}$, $\varrho_{\mathrm{p}} = 3000\,\mathrm{kg/m^3}$

$$\partial C/\partial t - \nabla\cdot\left(D'\nabla C\right) + \nabla \cdot \left[U_\varphi\left(1 - C\tilde{N}V_{\mathrm{p}}\right)C\right] = 0 \tag{61}$$

where U_φ is given by Equation (1). Equation (61) can be solved provided that information regarding D' is available. In the following, some properties of D' are outlined which can help to theoretically predict its value in diffusion sedimentation systems. The driving force F_μ per diffusing particle is often derived by using the gradient of the chemical potential μ [28]:

$$F_\mu = -\frac{1}{N}\nabla\mu \tag{62}$$

where μ is the chemical potential of 1 mol of particles. The hindrance force due to diffusion, as to other driving forces, in the laminar flow regime is given by Equation (46). Particles are assumed to be large, compared to fluid molecules, so that Stokes drag law can be applied. The chemical potential may be expressed as [29]

$$\mu = RT\left(1 + \sum_i U_i'/k_{\mathrm{B}}T\right)\ln C + RT\ln f_{\mathrm{a}} \tag{63}$$

where R and T are the gas constant and the absolute temperature, respectively; $k_{\mathrm{B}} = R/\tilde{N}$ is the

Boltzmann constant; f_{a} is the activity coefficient; and U_i' are energies (per particle), such as interfacial and electromagnetic energies, which may be associated with the particles. Combining Equations (62) and (63) gives

$$F_\mu = -k_{\mathrm{B}}T\left(1+\sum_i U_i'/k_{\mathrm{B}}T\right)\nabla \ln C - k_{\mathrm{B}}T\nabla \ln f_{\mathrm{a}} \tag{64}$$

Hence for steady diffusion where $F_\mu + F_{\mathrm{D}} = 0$ is satisfied U_φ is given by

$$U_\varphi = \frac{-k_{\mathrm{B}}T\left(1+\sum_i U_i'/k_{\mathrm{B}}T\right)\nabla \ln C - k_{\mathrm{B}}T\nabla \ln f_{\mathrm{a}}}{3\pi\, d\, \mu_\varphi'} \tag{65}$$

Note that the effect of hindrance due to acceleration of fluid displaced by the diffusing particles has been neglected. For steady diffusion and vanishing total flux, by using Equations (61) and (65),

$$D' = D\left(1+\sum_i U_i'/k_{\mathrm{B}}T + \nabla \ln f_{\mathrm{a}}/\nabla \ln C\right) f_{\mathrm{D}}(C) \tag{66}$$

where D is the well-known [28], [30] Stokes–Einstein diffusion coefficient

$$D = k_{\mathrm{B}}T/(3\pi\mu_{\mathrm{f}}d) \tag{67}$$

and $f_{\mathrm{D}}(C)$ is the hydrodynamic hindrance factor

$$f_{\mathrm{D}}(C) = \frac{1-C\tilde{N}V_{\mathrm{p}}}{\left[1+\left(C\tilde{N}V_{\mathrm{p}}\right)^{1/3}\right]\exp\left(\frac{5}{3}\cdot\frac{C\tilde{N}V_{\mathrm{p}}}{1-C\tilde{N}V_{\mathrm{p}}}\right)} \tag{68}$$

Note that $1/(3\pi\, d\, \mu_\varphi')$, which appears in Equation (65), is a generalization of the well-known particle mobility, i.e., drift velocity per unit driving force in the fluid frame of reference. In the dispersion frame of reference, this mobility is $(1 - C\tilde{N}V_{\mathrm{p}})/(3\pi\, d\, \mu_\varphi')$. In centrifugation technology, particle mobilities are also referred to as sedimentation coefficients, which are given in terms of velocity per unit acceleration g' [31]. In this case, the mobility S_{g} in the dispersion frame of reference is given by

$$S_{\mathrm{g}} = V_{\mathrm{p}}(\varrho_{\mathrm{p}} - \varrho_\varphi)\left(1-C\tilde{N}V_{\mathrm{p}}\right)/\left(3\pi\, d\, \mu_\varphi'\right)$$

In centrifugal fields where the driving force is radial, the diffusion equation takes the following form

$$\partial C/\partial t = D'\left(\frac{\partial^2 C}{\partial r^2} + \frac{1}{r}\cdot\frac{\partial C}{\partial r}\right) - S_{\mathrm{g}}\omega^2\left(r\frac{\partial C}{\partial r} + 2C\right) + \frac{\partial C}{\partial r}\left(\frac{\partial D'}{\partial C}\cdot\frac{\partial C}{\partial r} - C\omega^2 r\frac{\partial S_{\mathrm{g}}}{\partial C}\right) \tag{69}$$

and at steady state,

$$-D'\nabla C + U_\varphi\left(1-C\tilde{N}V_{\mathrm{p}}\right)C = \mathrm{const.} \tag{70}$$

Equation (70) can be solved, given the boundary conditions, for the steady diffusion and sedimentation profile of particle concentration.

2.4.2. Polydisperse System

Let φ' denote the probability density function of a random variable η (e.g., size, density, etc.) that characterizes the particle population:

$$\varphi' = \frac{1}{\varphi}\cdot\frac{\partial\varphi}{\partial\eta} \tag{71}$$

$$\int_{-\infty}^{\infty}\varphi'\mathrm{d}\eta = 1 \tag{72}$$

$$\mathrm{d}\varphi = \varphi\varphi'\mathrm{d}\eta \tag{73}$$

Let φ' be a function of position r and $D_\varphi(r, \eta)$ be the diffusion coefficient of sedimenting particles in the fluid frame of reference. The expectation $\overline{D_\varphi(r)}$ (i.e., weighted average of the population) of $D_\varphi(r, \eta)$ at position r is given by [32]

$$\overline{D_\varphi(r)} = \int_{-\infty}^{\infty} D_\varphi(r,\eta)\,\varphi'\mathrm{d}\eta \tag{74}$$

The diffusion coefficient $D_{\mathrm{d}}(r, \eta)$ of the fraction η at position r in the dispersion frame of reference is given by [29]

$$D_{\mathrm{d}}(r,\eta) = D_\varphi(r,\eta) - \varphi\overline{D_\varphi(r)} \tag{75}$$

Equation (75) provides the transformation from the frame of reference of the fluid (which is displaced by the diffusing particles) to that of the dispersion. Equation (75) shows that at a given position in the diffusion system, the diffusion coefficient of particles pertaining to the η fraction

depends on the expected diffusion coefficient of the whole distribution (in the fluid frame of reference) and on φ. Only in the fluid frame of reference can the diffusion of a particular fraction be assumed to be independent of the diffusion of the rest of the distribution. Furthermore, Equation (75) shows that reverse diffusion can occur provided $D_\varphi(r, \eta) - \varphi\overline{D_\varphi(r)} < 0$ is satisfied. Thus, faster diffusing particles can affect slower ones to the extent of reversing their local diffusion coefficient and also their diffusion flux. If reverse diffusion occurs, then the sedimentation and diffusion system will be divided into two counterdiffusing groups. Analysis of such systems involves a model of nonsteady diffusion that can be solved numerically. This, however, is outside the scope of this article.

2.5. Sedimentation of Fluid Particles

Fluid particulates differ from solid ones in view of their deformability and their capacity to maintain internal flow. Properties that may affect sedimentation of fluid particulates are listed below:

1) The stability of the fluid – fluid interface depends on flow conditions; slip velocity can occur.
2) Fluid particulates may be deformed; larger drops are more susceptible to deforming forces.
3) Fluid particulates can undergo coalescence or disintegration.
4) The viscosity of the dispersed fluid phase is finite; no limitation is imposed on its value relative to that of the continuous fluid phase.
5) Interfacial effects may be significant, for example, due to adsorption of surface-active agents or contaminants.

Barnea and Mizrahi extended their generalized approach toward the fluid dynamics of solid particulate systems [3] to dispersed liquids in fluid systems [33], [34]. If the assumptions made regarding the general properties of solid dispersions remain valid also for liquid in fluid dispersions, then the theory available for solids can be used for liquid dispersions once an appropriate modified viscosity is known. This modified viscosity must account for the finite viscosity of the dispersed phase and the hindrance effect of neighboring drops [35]. Equation (76) gives the effective viscosity μ'_f of a continuous fluid surrounding a single spherical drop, i.e., at $\varphi \to 0$ [35], [36]:

$$\mu'_\mathrm{f} = \mu_\mathrm{f} \frac{2/3 + \mu^*_\mathrm{d}/\mu_\mathrm{f}}{1 + \mu^*_\mathrm{d}/\mu_\mathrm{f}} \tag{76}$$

where

$$\mu^*_\mathrm{d} = \mu_\mathrm{d} + \sigma_\mathrm{k}, \sigma_\mathrm{k} = -\frac{1}{3} k^* \, \partial\gamma/\partial\Gamma,$$

μ_d is the viscosity of the dispersed phase, $k*$ is the surface effect retardation constant, and γ and Γ are the interfacial tension and the adsorption density of the dissolved material, respectively. Equation (76) shows that $\mu'_\mathrm{f} \to \mu_\mathrm{f}$ if $\mu_\mathrm{d} \to \infty$, whereas $\mu'_\mathrm{f} \to 2\,\mu_\mathrm{f}/3$ for $\mu_\mathrm{d} \to 0$. For dilute dispersions, the variation of the apparent viscosity of the continuous fluid $\mu_{\varphi\mathrm{d}}$ with φ is given in [37].

$$\mu_{\varphi\mathrm{d}} = \mu_\mathrm{f} \left(1 + 2.5\varphi A_\mathrm{d}\right),$$
$$A_\mathrm{d} = \left(\mu_\mathrm{f} + 2.5\mu_\mathrm{d}\right) / \left(2.5\mu_\mathrm{f} + 2.5\mu_\mathrm{d}\right) \tag{77}$$

Equation (78) is an extension of Equation (77) due to Leviton and Leighton [38]:

$$\mu_{\varphi\mathrm{d}} = \mu_\mathrm{f} \exp\left[2.5 A_\mathrm{d}\left(\varphi + \varphi^{5/3} + \varphi^{11/3}\right)\right] \tag{78}$$

Further extension of $\mu_{\varphi\mathrm{d}}$ can be obtained by use of the generalized drag coefficient $C_{\mathrm{D}\varphi}$ and Reynolds number Re_φ for monodisperse dispersions [3], [33], [34] (see Section 2.6). The drag coefficient $C_{\mathrm{D}\varphi}$, which depends on hydrostatic effects and on the distribution of particles in the dispersion, is unaffected by changing the solid into fluid particulates. The finite internal viscosity of each particle and the effect of this finite viscosity on hydrodynamic interactions between neighboring particles (Taylor correction) affect Re_φ. The parameter $\mu_{\varphi\mathrm{d}}$, given by Equation (79), accounts for these effects:

$$\mu_{\varphi\mathrm{d}} = \mu_\mathrm{f} B_\mathrm{d} B_\mathrm{v}, B_\mathrm{d} = \exp\left(\frac{5\varphi A_\mathrm{d}}{3\,(1-\varphi)}\right),$$
$$B_\mathrm{v} = \frac{\frac{3}{2} B_\mathrm{d} + \mu^*_\mathrm{d}/\mu_\mathrm{f}}{B_\mathrm{d} + \mu^*_\mathrm{d}/\mu_\mathrm{f}} \tag{79}$$

where the Taylor correction is accounted for by B_d and the effect of finite viscosity of the dispersed phase on that of the continuous phase

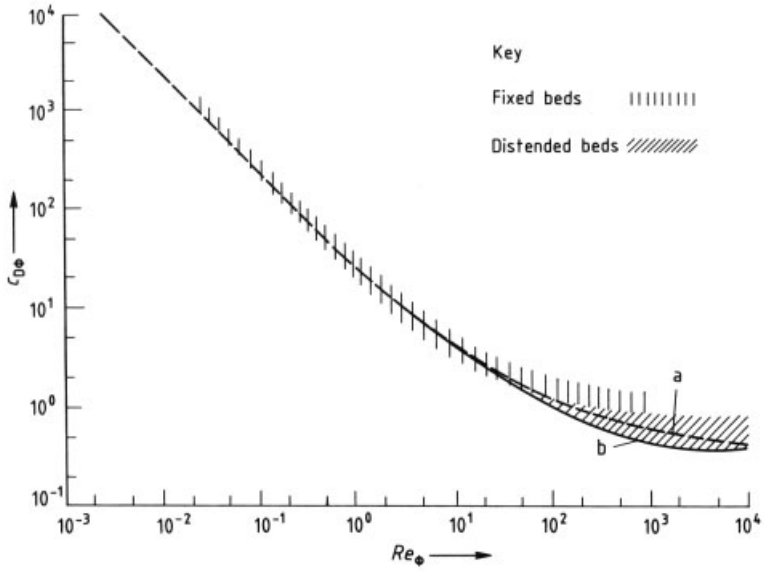

Figure 31. A plot of $C_{D\varphi}$ vs. Re_φ for calculated and experimental data (compiled from [41])
a) Equation 87; b) Lapple – Shepherd curve for single sphere

by B_v, in accordance with Equation (76). Substitution of $\mu_{\varphi d}$ for μ_φ in Equations (1) and (2) facilitates evaluation of steady sedimentation velocities of fluid particulates that are dispersed in a continuous fluid.

2.6. Generalized Correlations for Hindered Sedimentation Fluidization and Flow through Fixed Beds

Hindered sedimentation fluidization and flow through fixed beds pertain to the class of physical phenomena in which fluid is forced to flow between particles. This flow can be modeled as being inside a bundle of pipes that are formed between particles (when they are part of a fixed bed) or around individual particles with no reference to specific path of flow. Foscolo et al. [39] proposed a model for a fluidized bed that "enables its steady state particulate expansion to be predicted as a function of superficial velocity (U_c) from initial (packed bed) conditions to the final fully expanded (single suspended particle) state." Their results, which can also be applied to hindered sedimentation, are given by Equation (80) for laminar and turbulent flow, and Equation (81) for intermediate flow:

$$U_c/U_o = \left[\frac{\varepsilon^4}{\alpha\,(1-\varepsilon)+\varepsilon^3}\right]^m, \; \varepsilon = 1-\varphi \tag{80}$$

For laminar flow, $Re < 0.2$, $\alpha = 4.00$, and $m = 1.0$. For turbulent flow, $Re > 500$, $\alpha = 3.55$, and $m = 0.5$.

$$U_c/U_o = \frac{\left[0.0777 Re\,(1-0.0194\,Re)\,\varepsilon^{4.8}+1\right]^{0.5}}{0.0388\,Re}, \quad 0.2 \le Re \le 500 \tag{81}$$

where U_c is superficial velocity, related to U_φ by

$$U_c = U_\varphi\,\varepsilon = U_d, U_o = U_\varphi\,(\varphi = 0)\,, \text{ and } Re = \frac{\varrho_f U_o d}{\mu_f}$$

Equations (80) and (81) were derived by using an extended tortuosity function and the model of flow through a bundle of pipes in accordance with the Blake – Kozeny flow model. Because $\varepsilon = 1 - \varphi$, these equations provide another set of generalized $f(\varphi)$ functions such as that given by Equation (1). Pressure drop per unit length is given by

$$\Delta P_l/L = 17.3\left(\mu_f U_c/d^2\right)(1-\varepsilon)\,\varepsilon^{-4.8},\quad Re<0.2 \tag{82}$$

$$\Delta P_t/L = 0.336\left(\varrho_f U_c^2/d\right)(1-\varepsilon)\,\varepsilon^{-4.8},\quad Re>500 \tag{83}$$

$$\Delta P_i/L = (\Delta P_l + \Delta P_t)/L,\quad 0.2 \le Re \le 500 \tag{84}$$

The model of flow around individual particles was used to derive a generalized correlation for hindered sedimentation fluidization and flow through fixed beds [40]. Figure 31 shows the drag coefficient $C_{D\varphi}$ vs. the Reynolds number Re_φ [41]. Equations (85) and (86) provide definitions of $C_{D\varphi}$ and Re_φ:

$$Re_\varphi = U_\varphi \varrho_f d/\mu_\varphi,\quad \mu_\varphi = \mu_f \exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right) \tag{85}$$

$$C_{D\varphi} = 2F_{D\varphi}/\left[\varrho_f S U_\varphi^2\left(1+\beta_\varphi \varphi^{1/3}\right)\right],\quad S=\pi d^2/4 \tag{86}$$

The subscript φ denotes that the variable is given for a dispersion at φ. For fixed beds $\beta_\varphi = 2.6$, whereas for hindered sedimentation $\beta_\varphi = 1$. To evaluate U_φ and $F_{D\varphi}$, a correlation must be made between $C_{D\varphi}$ and Re_φ, and β_φ must be specified. To this end, the Dallavalle correlation [40–42] can be used:

$$C_{D\varphi} = \left(0.63 + 4.8 Re_\varphi^{-1/2}\right)^2 \tag{87}$$

A plot of Equation (87) is shown in Figure 31. Equation (87) agrees with experimental data for fixed beds in the laminar flow regime, but it involves a substantial discrepancy in the turbulent $Re > 500$ regime. For hindered sedimentation, Equation (87) also fits experimental data in the turbulent regime and for the full range of practical values of φ. Equation (88) is an extension of Equation (87) [40]:

$$C_{D\varphi} = \left(0.63 + 4.8 Re^{-1/2} + \alpha_\varphi \tau_\varphi Re_\varphi^{1/8}\right)^2 \tag{88}$$

$$\alpha_\varphi = (\varphi/0.6)\left[1/8 + \sum_{i=2}^{n} a_i\,(\varphi/0.6)^i\right],\quad i = 2, 3, \ldots, n \tag{89}$$

where α_i represents constants by which $\alpha_\varphi = \alpha(\varphi)$ can be determined and τ_φ relates to the degree of constraint of particle motion. For particles that are fully constrained, i.e., fixed in space as in fixed beds, $\tau_\varphi = 1$; for hindered sedimentation where no constraints other than hydrodynamic drag exist, $\tau_\varphi = 0$. If partial fixation of particles prevails, then $0 \le \tau_\varphi \le 1$. Using only the first term of α_φ provides (Eq. 90) a good first-order fit between Equation (88) and experimental data (Fig. 31) for both fixed and hindered sedimentation curves:

$$C_{D\varphi} = \left(0.63 + 4.8 Re_\varphi^{-1/2} + \tau_\varphi\,(\varphi/4.8)\,Re_\varphi^{1/8}\right)^2 \tag{90}$$

At $\varphi = 0$ Equation (90) reduces to Equation (87), and at $\varphi = 0.6$ corresponds well with literature data for fixed beds $\tau_\varphi = 1$ and for hindered sedimentation $\tau_\varphi = 0$. Equation (91) gives $\beta_\varphi = \beta(\varphi)$ in the form of an infinite series:

$$\beta_\varphi = 1 + \tau_\varphi\,(\varphi/0.6)\left[1.6 + \sum_{j=2}^{m} b_j\,(\varphi/0.6)^j\right],\quad j = 2, 3, \ldots, m \tag{91}$$

The first constant of β_φ satisfies $\beta_\varphi = 2.6$ at $\varphi = 0.6$, $\tau_\varphi = 1$, $b_j = 0$ for $j > 2$, as required. Taking only the first term as a first-order approximation yields

$$\beta_\varphi = 1 + 8\tau_\varphi \varphi/3 \tag{92}$$

and the corresponding extended drag coefficient is then

$$C_{D\varphi} = C_{D0}/\left[1 + \varphi^{1/3} + (8\tau_\varphi/3)\,\varphi^{4/3}\right] \tag{93}$$

where $C_{D0} = C_{D\varphi}\,(\varphi = 0)$ (see Equation 86). Note that in the laminar flow regime Equation (94) is preferable:

$$C_{D\varphi} = 24/Re_\varphi \tag{94}$$

The pressure drop per unit length across the particle assembly (e.g., as a fixed bed or as a suspension) is the sum of individual (i.e., per particle) drag forces:

$$\Delta P/L = N F_{D\varphi},\quad N = \varphi/\left(\frac{\pi d^3}{6}\right) \tag{95}$$

Equations (96) and (97) give the results for the laminar and intermediate or turbulent flow regimes, respectively:

$$\Delta P/L = \frac{18\mu_f U_c \varphi \left[\left(1+\beta_\varphi \varphi^{1/3}\right) \exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right)\right]}{d^2\,(1-\varphi)} \tag{96}$$

$$\Delta P/L = \frac{\left(\frac{3}{4}\right)\left(0.63 + 4.8 Re_\varphi^{-1.2} + \tau_\varphi \varphi\, Re_\varphi^{1/8}/4.8\right)^2 \varrho_f U_c^2\, \varphi\, \left(1+\beta_\varphi \varphi^{1/3}\right)}{d\,(1-\varphi)^2} \tag{97}$$

Note that Equation (97) also provides a good approximation in the laminar flow regime; hence, if Re_φ is not known, it can be used as a first trial. In cases involving polydisperse mixtures and consisting of i_n size fractions, Equation (95) should be replaced by Equation (98), with due changes in Equations (96) and (97).

$$\Delta P/L = \sum_{i=1}^{i_n} n_i F_{D\varphi i}, \quad n_i = \varphi_i / \left(\frac{\pi d_i^3}{6}\right), \quad N = \sum_{i=1}^{i_n} n_i \tag{98}$$

Extending the above model, in which the contribution to pressure drop due to the drag exerted on individual particles has been taken into account, to polydisperse mixtures is straightforward. This applies to fixed, distended, and stable fluidized beds, and to hindered sedimentation in the full range of flow regimes and volume fractions occupied by the particles. Moreover this generalized model does not involve any assumptions as to the appropriate averaging of particle size, which is necessary when the model of flow through a bundle of pipes is used.

2.7. Enhancement of Sedimentation

Sedimentation of small and near-density particles can be very slow; hence, a way is needed to enhance the throughput of devices involving sedimentation of such particles. *Near density* means that the density of particles is close to that of the fluid. The following options for enhancement of sedimentation rate should be considered:

1) size enlargement by processes such as flocculation (→ Flocculants), coagulation, precipitation, and crystallization (→ Crystallization and Precipitation);
2) increase of driving forces, i.e., of the effective acceleration g'; and
3) decrease of sedimentation retention time by decreasing the sedimentation path and increasing the area available for sedimentation and for deposition of sediment.

Size enlargement is a well-known method in solid – liquid separation and water treatment operations [17], [42], [43]. Size enlargement (or the reverse, i.e., dispersion) can occur spontaneously through a change in the properties of the suspension (i.e., pH and ionic strength) and through addition of flocculants (or dispersants) that bind particles (or disintegrate aggregates) into floccules.

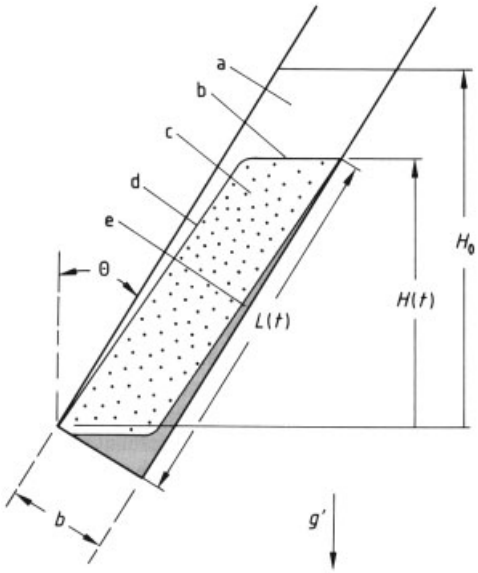

Figure 32. Formation of flow regions during sedimentation of monodisperse particles in an inclined channel [2]
a) Clear fluid; b) Horizontal interface between fluid and suspension; c) Suspension; d) Clear fluid above inclined sedimentation front; e) Sediment

Increase of driving force can be achieved by applying force fields (of higher intensity) to

which the particles are or can be made susceptible. *Centrifugal fields* are often employed when separation of slowly settling particles by density or size differences is desirable. *Sedimentation of polarizable particles* toward polarized elements occurs in magnetic (→ Magnetic Separation) and electrical separation (→ Electrostatic Separation) and in filtration (→ Filtration). *Attachment of polarizable particles to nonpolarizable ones* renders the aggregate susceptible to sedimentation in the relevant polarizing field.

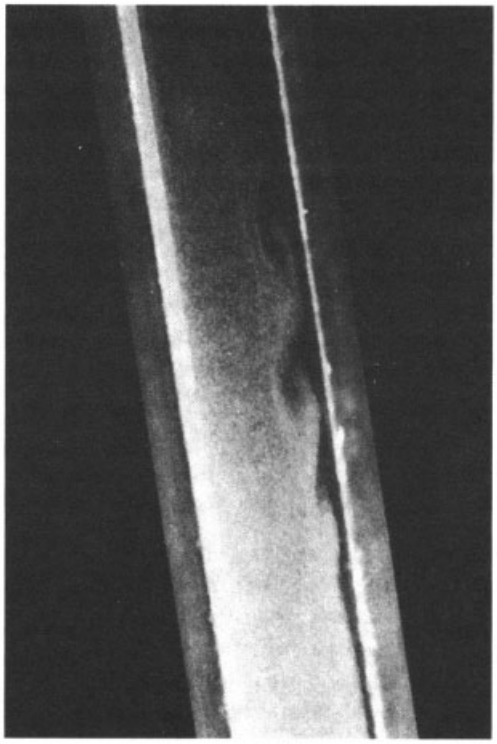

Figure 33. Formation of unstable waves during sedimentation in an inclined channel [2]

A *decrease of sedimentation retention time* can be achieved by increasing the volumetric density of collecting elements toward which the particles undergo sedimentation. This involves a corresponding decrease in the working volume available for holding and flow of the dispersion. Enhanced sedimentation in inclined channels is a good example of this approach (see Section 2.8). Enhanced sedimentation occurs also in filtration processes where particles are partially or fully carried by the filtering fluid toward the filtration cake on which they are deposited. In this case, drag forces that drive particles can reinforce or counteract other sedimentation forces, depending on the geometry of the system. In electromagnetic separation and filtration (→ Magnetic Separation, Chap. 3.2.1.) of polarizable particles, a matrix of highly polarizable elements is distributed in the separation chamber, which is placed in a high-intensity field, thus decreasing the sedimentation path of particles and increasing the area available for their collection [44].

2.8. Sedimentation in Inclined Channels

Enhanced sedimentation of blood corpuscles in inclined tubes was observed by Boycott in 1920; since then, this phenomenon of enhanced sedimentation in inclined channels has been a subject of long-standing research [45]. Reviews are given in [3] and [46]. Enhanced sedimentation in inclined channels is used commercially in *lamella thickeners* [17], [43], *sedimentation centrifuges* [47], [48], and *sedimentation cones* [17] (→ Biochemical Engineering).

Figure 32 shows five regions characterizing the flow field in an inclined channel sedimentation system [2]. Particles reach the final sediment by first forming a deposit on the inclined channel floor, which slides at a faster rate toward the channel bottom. In this way, the higher retention time associated with sedimentation in suspension is cut down in favor of the lower retention time associated with the sliding deposit. The inclined channel system incorporates the principle of increased volume density of collection elements for the sedimenting particles or else increased surface area for their deposition. The displaced fluid flows toward the top of the channel. A thin clear layer is then formed above the inclined sedimentation front, and a thicker clear layer develops above the shorter horizontal sedimentation front at the top (see Fig. 32). Ponder [49] and Nakamura and Kuroda [50] proposed the so-called PNK theory for the clarification rate in inclined channels. This theory states that the rate of production of clarified fluid is given by the product of the settling velocity of particles and the projection of the channel area available for particle deposition on a plane set perpendicular to the settling direction. For the

parallel plate geometry shown in Figure 32, this theory predicts

$$g(t) = U_{\mathrm{d}}\,(b\cos\theta + L\sin\Theta) \tag{99}$$

where $g(t)$ is the volume rate of clarified fluid per channel per unit length L_{p} perpendicular to the plane of the drawing, U_{d} is given by Equation (7) for monodisperse particles, θ is the angle of inclination from the vertical axis, and b and L are the dimensions of the shorter and the larger lower walls of the channels, respectively (see Fig. 32). Let the volume per unit length L_{p} of the sedimentation vessel be V_{v}, and assume that it is large enough to contain a number of channels for which Equation (99) applies. From this, the steady-state rate of clarification for such volume (fully packed with inclined channels) is

$$Q(t) = V_{\mathrm{v}} U_{\mathrm{d}}\,(\cos\theta/L + \sin\theta/b) \tag{100}$$

Equation (100) shows that $Q(t)$ is inversely related to L as well as to b. Hence, given L, the use of narrower channels, with due regard to design constraints, facilitates a corresponding increase in $Q(t)$. For analysis of sedimentation in inclined channels, see [51–56]. Equations (99) and (100) are valid for laminar flow. If instabilities followed by deformation of sedimentation fronts occur, then the efficiency of the process can be curtailed significantly. Figure 33 shows an example of the formation of unstable waves in the sedimentation fronts during sedimentation of 130-µm glass beads, $\varrho_{\mathrm{p}} = 2440\ \mathrm{kg/m^3}$, which were suspended in a synthetic Newtonian lubricant, $\varrho_{\mathrm{f}} = 1050\ \mathrm{kg/m^3}$, $\mu_{\mathrm{f}} = 0.01\ \mathrm{kg\,m^{-1}\,s^{-1}}$. For further information regarding linear stability analysis of sedimentation in inclined channels, see [2] and [57–59]. The sedimentation of polydisperse suspensions in vessels having inclined walls is considered in [60]. Figure 34 shows the different sedimentation regions that develop during sedimentation of a polydisperse particle mixture containing N discrete fractions. Equation (101) gives the dimensionless instantaneous volumetric rate $S_j^*(t)$ at which a suspension devoid of particles faster than or equal to the jth fraction is formed [60]:

$$S_j^*(t) = U_{\mathrm{d}j}^* \left[1 + \frac{H_j^*}{b^*}\sin\theta\right] \frac{b^*}{\cos\theta} \tag{101}$$

where all starred quantities have been rendered dimensionless by using the initial height of the suspension H_0 as the characteristic length and the velocity (at $\varphi = 0$) of the fastest settling particle as the characteristic velocity. For steady-state continuous settling, letting $j = N$ in Equation (101) gives the dimensionless rate of formation of clear fluid. This can be used for design purposes or for evaluation of the maximum capacities that can be expected from existing equipment. For batch sedimentation [60],

$$\mathrm{d}H_j/\mathrm{d}t = -U_{\mathrm{d}j}\left[1 + \frac{H_j}{b}\right]\sin\theta \tag{102}$$

where at $t = 0$, $H_j = 1$; $j = 1, 2, \ldots, N$. The rate of descent of the region containing all particles ($j = 1$) is fastest, with the one corresponding to $j = N$ being the slowest. Equations (101) and (102) can also be applied to continuous distributions by letting $N \to \infty$.

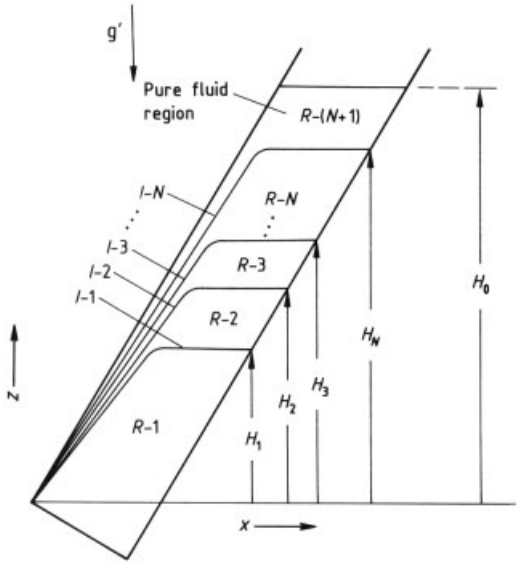

Figure 34. Formation of $N + 1$ flow regions during sedimentation of polydisperse particles (comprising N fractions) in an inclined channel [60]. Region j, $1 \le j \le N$ contains particles pertaining to fraction i, $i + 1, \ldots, N$, $i = j$. Fraction $j = N + 1$ denotes clear fluid. Larger i denotes slower settling particles

Sedimentation phenomena of polydisperse mixtures in inclined channels are generally considered by assuming low volume fractions of solids. Furthermore, the analysis is concerned mainly with discontinuities between regions and not with processes occurring inside these sedimentation regions. As with nonsteady sedimentation, the formation of concentration waves can

also be expected in sedimentation in inclined channels. Such waves can provide further hindrance to sedimentation and cause instabilities (see Section 2.3.3).

2.9. Principle of Hydrodynamic Particle Competition

The principle of hydrodynamic particle competition (HPC) states that hydrodynamic hindrance to motion of particles in the direction of a driving force is enhanced due to the competition of neighboring particles for available flow space. The probability of particles, entering a flow space is determined not only by their own response to the driving force but also by their relative response compared to other particles. Thus according to this principle, less responsive particles may be driven out of a flow zone by more responsive ones, notwithstanding their being driven by a positive force toward this zone. The term "response" is used in the most general way; for example, larger particles are more responsive to body forces. In the analysis of particle motion (e.g., in separation processes), assessing whether a certain part of a distribution can reach a given position in space in a given time may be of interest. In this context, the expected effect of changing operating parameters on the probability of separation must be determined. This probability is affected by "side variables," which do not constitute the basis for separation but inherently enter the force terms. The relative HPC effect of such variables is expected to be more pronounced due to the following:

1) increase of φ;
2) shift of size distribution toward the finer size end;
3) decrease of difference between particles and fluid with respect to the variable that constitutes the basis for particle motion;
4) increase in standard deviation of variables that constitute the basis for particle motion, of side variables, and of particle size; and
5) increased skewness of distribution variables, specifically those directly affecting particle motion.

3. Processes and Equipment Involving Sedimentation

3.1. Settling Pools

The settling pool can be used as an idealized model for crossflow sedimentation (see Section 2.2). Figure 35 shows an ideal settling pool in which plug flow is assumed to prevail. The time t_{H} required for a particle introduced at the feed port to reach the pool floor is related to the pool height H_{p} by

$$H_{\mathrm{p}} = \int_0^{t_{\mathrm{H}}} U_{\mathrm{d}}\,\mathrm{d}t = \int_0^{t_{\mathrm{H}}} (U_{\varphi} - U_{\mathrm{f}})\,\mathrm{d}t \tag{103}$$

If φ is low throughout the sedimentation path, then $U_{\mathrm{d}} = U_{\varphi} =$ const. can be assumed; hence

$$t_{\mathrm{H}} = H_{\mathrm{p}}/U_{\mathrm{d}} \tag{104}$$

For particles that travel simultaneously the full length L and depth H_{p} of the pool, $t_{\mathrm{H}} = t_{\mathrm{R}}$, with the residence time t_{R} related to the crossflow velocity U_{R} by

$$t_{\mathrm{R}} = L/U_{\mathrm{R}} \tag{105}$$

$$U_{\mathrm{R}} = Q/(B_p H_{\mathrm{p}}) \tag{106}$$

where Q is the volumetric flow rate across the pool and B_{p} is the pool width. Thus the limiting particle size reaching the pool floor within the distance L can be determined from

$$U_{\varphi} = \frac{Q}{A_{\mathrm{s}}}, \quad A_{\mathrm{s}} = BL \tag{107}$$

where A_{s} is the pool area across which sedimentation occurs. Equation (107) shows that in ideal pools, the ability to capture a particle is independent of pool depth. This property is also known as the area principle. Real pools and systems that are simulated by sedimentation pools deviate from ideality due to the effects of increased particle concentration, particle distribution, flow profiles other than plug flow, and inhomogeneous force fields. The actual surface area required for a given limiting particle size is then

$$A_{\mathrm{s}} = \frac{Q}{\eta_{\mathrm{p}} U_{\varphi}} \tag{108}$$

where η_p denotes the efficiency, i.e., the ratio of theoretical to required surface area. The customary assumption is that $0.5 \leq \eta_p \leq 1.0$, but lower and in theory even vanishing values of η_p can be expected in pools and crossflow systems that involve sedimentation of polydisperse mixtures at elevated concentration. For fixed values of A_s and η_p, Q/U_φ is constant. Hence, increasing the concentration of the suspension, when the limiting size d_L is fixed, decreases U_φ and also Q. All particles having size larger than d_L are expected to settle in the pool. Thus the captured fraction is given by

$$\Phi(d \geq d_L) = 1 - \Phi(d \leq d_L) = \int_0^{d_L} \varphi' \mathrm{d}d_c,$$

and the total mass of solids capture is then

$$\begin{aligned} &\varrho_p \varphi [1 - \Phi(d_L)] Q/A_s \\ &= \eta_p \varrho_p \varphi [1 - \Phi(d_L)] U_d(d_L) \\ &= \eta_p \varrho_p [1 - \Phi(d_L)] \varphi (1 - \varphi) U_\varphi(d_L) \end{aligned}$$

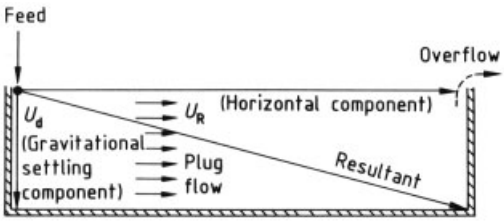

Figure 35. Plug flow and sedimentation of particles in an ideal settling pool [17]

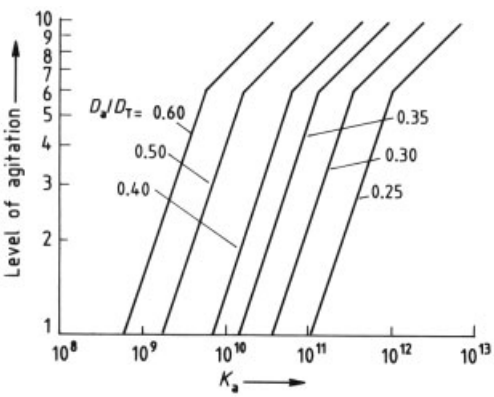

Figure 36. A plot of level agitation vs. K_a for different values of D_a/D_T [17] after [62]

For a given d_L, the maximum mass that can be captured is obtained by maximizing $\varphi(1-\varphi) U_\varphi(d_L)$, which gives $\varphi_{max} = 0.1792$ in the laminar regime. However, if d_L is not fixed, then $\Phi(d_L)\varphi(1-\varphi) U_\varphi(d_L)$ must be maximized instead.

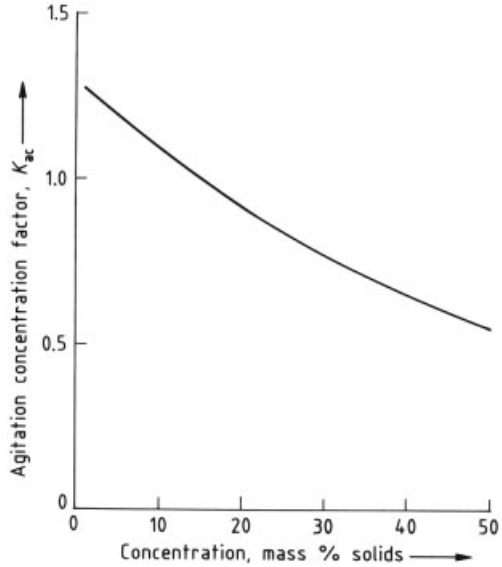

Figure 37. A plot of agitation concentration factor K_{ac} vs. concentration [17] after [62]

3.2. Suspension of Particles in Fluids

3.2.1. Suspension of Particles in Mixed Containers

Particles must often be suspended in fluids in which sedimentation must be prevented at various degrees. Uniform suspensions of particles in fluid streams can be achieved by *fluidization, intense agitation,* and *turbulent flow.* If particles are suspended by turbulent agitation, then the process can be expressed in terms of the *eddy diffusion coefficient* D_e:

$$D_e \cdot \frac{\mathrm{d}C}{\mathrm{d}y} = -CU_y \tag{109}$$

where C is particle concentration at height y and $U_y = U_d$. Equation (109) is incomplete in the sense that it relates to a mass-transfer process that is restricted exclusively to the direction of sedimentation. The following equation also accounts for mass transfer across the direction of sedimentation:

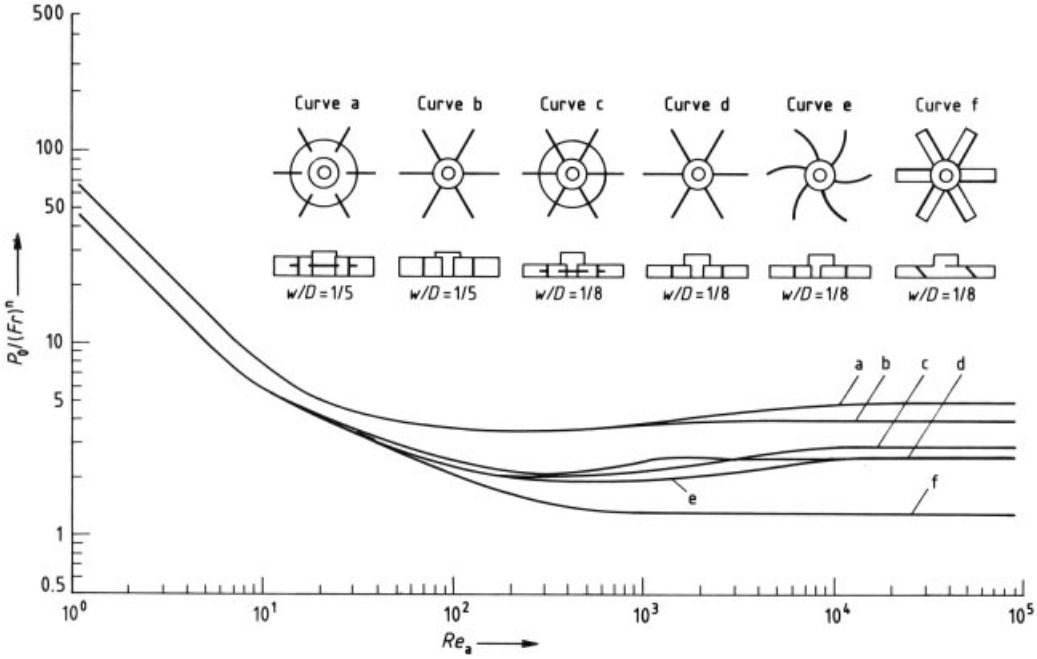

Figure 38. Turbine power correlation [63], [64]: $P_0/(Fr)^{\mathrm{n}}$ vs. Re_{a}

$$\frac{\partial}{\partial Y}\left(D_{\mathrm{e}}\cdot\frac{\partial C}{\partial Y}\right)+D_{\mathrm{e}}\cdot\frac{\partial^2 C}{\partial x}=U_{\mathrm{x}}\cdot\frac{\partial C}{\partial X}-U_{\mathrm{y}}\cdot\frac{\partial C}{\partial Y} \tag{110}$$

where U_{x} is the time-averaged component of particles velocity in a direction perpendicular to U_y (i.e., horizontal if U_y is vertical). Increased intensity of agitation decreases concentration gradients. The power required for complete dispersion in a tank is given as follows [17], [61]:

$$P/V_{\mathrm{T}}=0.092gU_{\varphi}\frac{D_{\mathrm{T}}}{D_{\mathrm{a}}}\left(\frac{\varphi}{1-\varphi}\right)^{1/2}$$

$$\cdot\exp\left(5.3\cdot\frac{H_{\mathrm{a}}}{D_{\mathrm{T}}}\right)(\varrho_{\mathrm{p}}-\varrho_{\mathrm{f}}) \tag{111}$$

where P, V_{T}, D_{a}, D_{T}, and H_{a} are the power, volume of slurry in the tank, agitator diameter, tank diameter, and distance of agitator impeller from the bottom of the tank, respectively. For turbulent agitation, U_{φ} is given by Equation (55). Equation (111) shows that the power density is directly related to the sedimentation velocity at $\varphi\to 0$ and to $[\varphi/(1-\varphi)]^{1/2}$. This equation provides conservative estimates of power consumption because in many cases, a completely uniform dispersion is not required. A less conservative estimation of power requirements is given by the *scale of agitation* approach [62], which classifies the degree of suspension into ten levels. In level 1, particles are partially suspended, being in contact with the bottom of the agitator; in level 3, all particles are suspended; and in level 10, uniform suspension prevails. In level 5, the lower part of the agitated volume (i.e., closer to the impeller) is practically uniform, whereas the upper part contains concentration gradients. Figure 36 shows the observed levels of agitation vs. K_{a}, where K_{a} (Eq. 112) relates to design parameters of the tank and properties of the slurry:

$$K_{\mathrm{a}}=154K_{\mathrm{ac}}\omega^{3.75}D_{\mathrm{a}}^{2.81}/U_0 \tag{112}$$

where ω is expressed in revolutions per minute and K_{ac} is the agitation concentration factor that is plotted vs. concentration in Figure 37. Figure 38 shows a plot of $P_0/(Fr)^{\mathrm{n}}$ vs. Re_{a} for six-blade turbine generator data [63], [64], where Re_{a} is the Reynolds number of the agitator,

$$Re_{\mathrm{a}}=D_{\mathrm{a}}^2\omega\varrho_{\mathrm{f}}/\mu_{\mathrm{f}} \tag{113}$$

P_0 is the power number,

$$P_0 = \frac{P}{\varrho_f \omega^3 D_a^5} \tag{114}$$

and Fr is Froude number, i.e., the ratio of inertial to gravitational force:

$$Fr = \omega^2 D_a / g \tag{115}$$

In baffled tanks, $Fr = 1.0$; otherwise, an appropriate value of n must be selected [17].

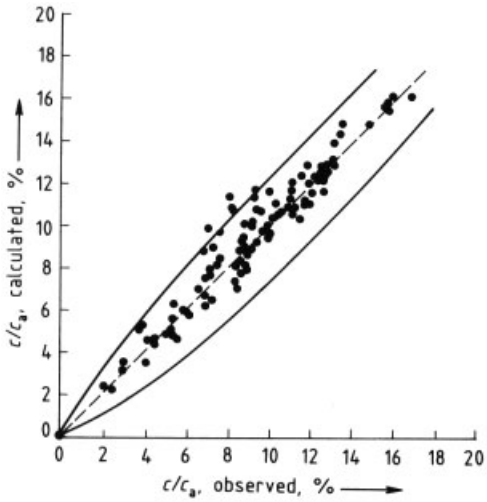

Figure 39. A plot of calculated vs. observed values of C/C_a, for 1.19 – 0.59 mm and 0.595 – 0.297 mm size fractions of coal [65], [69]
—— = 95 % confidence limit for probe size analyses

3.2.2. Suspension of Particles in Conduits

Horizontal Open Channels. Suspension and sedimentation of particles in conduits occur in flows of slurries and dispersions in channels or pipes [65]. As with agitators and mixers, suspension of particles against their sedimentation is achieved by turbulent flow. Equation (109) was used by O'Brien [66] in the early 1930s to describe the mechanism of particle suspension in turbulent transportation of particles in conduits. The eddy diffusion coefficient D_e given in Equation (109) is also a mass-transfer coefficient. This mass-transfer coefficient D_s can be related to the eddy viscosity or fluid momentum transfer coefficient D_m as follows:

$$D_s = \beta_s D_m = \beta_s k_f U^* y \, (1 - y/y_m) \tag{116}$$

$$U^* = \sqrt{\tau_w / \varrho_f} = U / \sqrt{f/2} \tag{117}$$

$$k_f = l_m / y \tag{118}$$

where β_s is a constant believed by some researchers to be unity; l_m is the eddy mixing length, which has the same role as the mean free path in the kinetic theory of gases; y and y_m are distances from the conduit bottom wall to a given point and to the point of maximum flow velocity, respectively; $U*$ is the frictional velocity defined by Prandtl; τ_w is shear stress at the wall; U is mean flow velocity in the conduit; f is the friction factor of the flow system; and k_f is the Karman constant rated at 0.40 for pure fluids but lower for slurries due to suppression of the intensity of turbulence [67]. Substituting Equation (116) in Equation (109) and integrating give

$$\ln (C/C_a) = \frac{U_d}{\beta_s k_f U^*} \ln (y'/y'_a) \tag{119}$$

$$y' = y_m/y - 1, \quad y'_a = y_m/y_a - 1$$

where C and C_a are particle concentrations at a level y and at a reference level y_a above the conduit bottom, respectively.

Equation (119) shows that C/C_a depends on y and on the ratio of the sedimentation velocity at y to the group $\beta_s k_f U*$, which is a measure of the intensity of turbulence. Good agreement between experimental data for open channels and Equation (119) has been achieved [67].

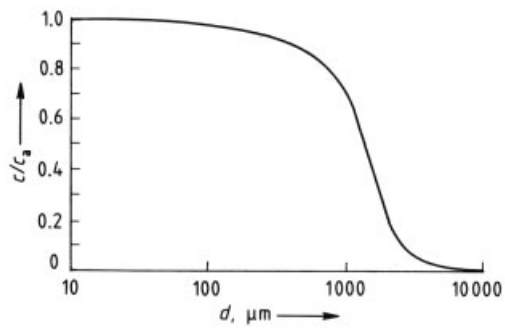

Figure 40. A plot of C/C_a vs. particle size d, 30.5 cm pipe inner diameter, flow velocity 1.8 m/s, 50 wt % coal

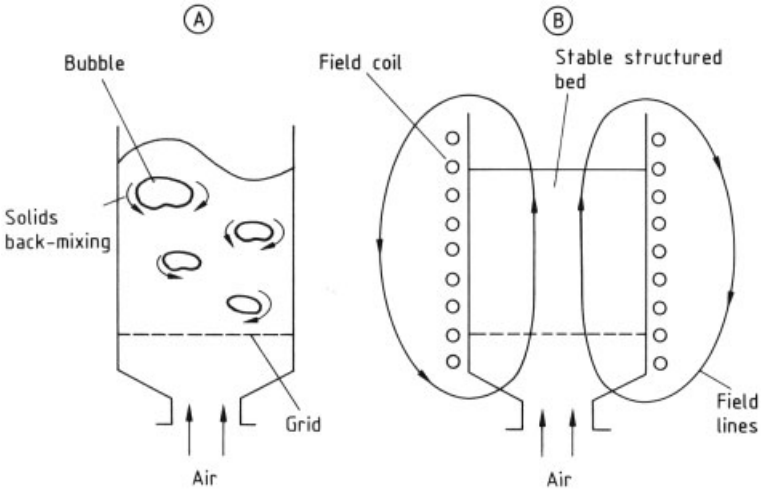

Figure 41. A) Ordinary and B) magnetically stabilized fluidized beds [76]

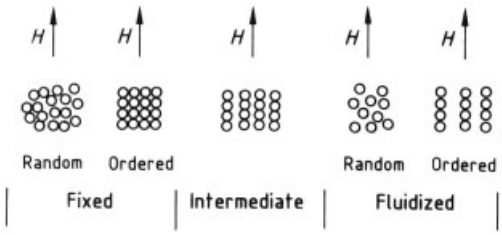

Figure 42. Particle structures in fixed, intermediate and fully fluidized magnetizable beds [75]

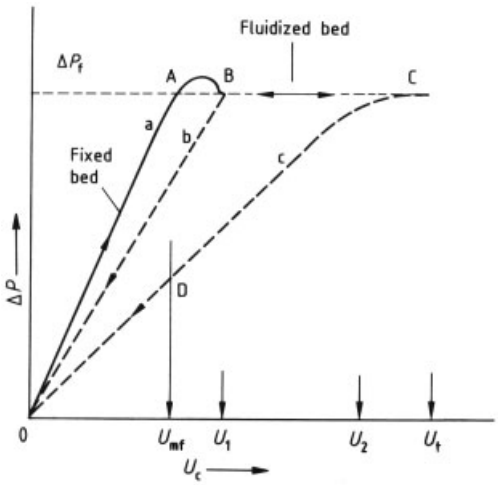

Figure 43. A schematic plot of ΔP vs. U_c showing fluidization and defluidization curves at different field intensities [75]
a) Fluidization curve; b) Defluidization curve, $H = 0$; c) Defluidization curve, $H > 0$

Closed Channels and Pipes. A correlation for slurry flow in pipes was developed by an engineering team of the Consolidated Coal Company in the United States [68]. For the design of an Ohio coal pipeline, they collected data from a 30.5-cm pipeline and arrived at the following correlation:

$$\ln\,(C/C_a) = -4.145\,U_d/\,(\beta_s k_f U^*) \tag{120}$$

where C_a is the concentration at the pipe axis. A plot of calculated vs. observed values of C/C_a for two (1.19 – 0.59 mm, i.e., 14/28 mesh; 0.595 – 0.297 mm, i.e., 28/48 mesh) size fractions at two positions $L_p = 0$ and $L_p = 160$ km (L_p is distance along the pipe) is shown in Figure 39 [65], [69]. The reliability of Equation (120), within a 95 % confidence level, in predicting the distribution of relative concentration across the flow profile is demonstrated clearly. Often the definition of the state of flow depends on setting accepted values for C/C_a where, for example, C is taken at the topmost part of the pipe. Thus letting $C/C_a \geq 0.8$ across the full flow profile can be considered a definition of homogeneous flow in which vertical concentration gradients can be neglected. Because C/C_a is a power function of $U_d/U*$, the effect of sedimentation velocity and hence also of particle size can be critical. This is seen in Figure 40 [65], [70], where C/C_a vs. particle size d is plotted for a 30.5-cm pipe carrying a 50 wt %

coal slurry at a flow velocity $U = 1.82$ m/s (C is concentration at the top of the pipe). By using the $C/C_a \geq 0.8$ criterion, the slurry can be considered homogeneously flowing if $d < 700\,\mu$m. Increasing d by an order of magnitude under the same operating conditions results in practically complete sedimentation. Flow velocities and concentration of solids have a strong effect on C/C_a, whereas variation of pipe diameter has little effect [70]. Increasing flow velocities and using higher slurry concentration provide enhanced conditions for operating at larger C/C_a. This is due to a decrease in U_d (following the increase in C) and an increase in $\beta_s k_f U*$ (following the increase in flow velocities).

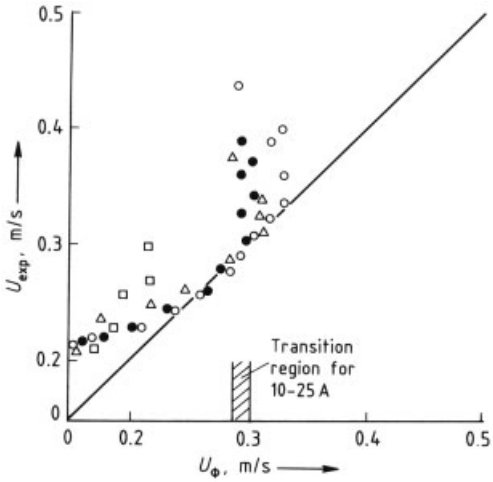

Figure 44. A plot of measured sedimentation velocity U_{exp} vs. calculated velocity U_φ, at different current levels [75]: □ 5 A; △ 10 A; ● 15 – 20 A; ○ 25 A

3.3. Fluidization and Sedimentation

In *fluidization,* particulates are suspended in a fluid stream so that their time-averaged velocity in the dispersion frame of reference vanishes. However, in the fluid frame of reference, particles undergo a time-averaged steady sedimentation. In stable fluidization, no displaced fluid exists because the fluid provides the driving force for suspension of the particles, whereas in sedimentation (which takes place in a container at fixed volume), fluid is displaced due to particle motion. Thus in fluidization of monodisperse particles, $U_d = 0$, $U_\varphi = U_f$, and the right-hand side of Equation (8) does not hold. Therefore, in cases of particulate fluidization Equations (1) and (2) can be used to evaluate gas velocity directly or, alternatively, to evaluate sedimentation velocity in the fluid frame of reference. Note that particulate fluidization means stable fluidization of randomly and uniformly dispersed particles. However, if particles attain ordered structures, then their sedimentation velocities can be expected to deviate significantly from those predicted for random structures. Thus, different structures formed by particles in sedimentation and fluidization systems, at the same levels of φ, are expected to yield different sedimentation velocities. In lateral segregation in sedimentation of bidisperse suspensions, formation of particle streams in a suspension containing a mixture of dense and neutrally buoyant particles has been observed [71]. A remarkably enhanced sedimentation rate was also noted for the denser particles. Similar phenomena have been observed in a bidisperse mixture containing particles more and less dense than the fluid [72]; the fingering effect also occurs in the sedimentation of bidisperse mixtures containing particles that are all more dense than the fluid [73]. The settling rate of glass spheres in carbon tetrachloride can be increased up to sixfold by adding poly(vinyl chloride) (PVC) particles to the dispersion [74].

Notwithstanding the fact that these phenomena are of great interest from the point of view of sedimentation and fluid mechanics, they seem to have relatively little importance in enhancing overall sedimentation rate [2]. The effect of structuring the particle assembly on sedimentation and fluidization velocities, has been demonstrated in a magnetically stabilized fluidized-bed system [75]. Figure 41 depicts schematically an ordinary (Fig. 41 A) and a magnetically stabilized fluidized bed (Fig. 41 B) [76]. The bed contains magnetizable particles, and the field is generated by a coil wound around the fluidization vessel. Low-intensity uniform magnetic fields are sufficient to stabilize a bed of ferromagnetic particles such as magnetite (Fe_3O_4) [76]. In the absence of a magnetic field or for sufficiently high fluid velocities, the bed becomes unstable and functions as an ordinary unstable bed. Figure 42 shows different structures that can be formed in the fixed-, fluidized-, and intermediate-bed states [75]. The structures are characterized by the degree of alignment of par-

ticles with the field H. Figure 43 shows a typical ΔP vs. U_c plot of fluidization and defluidization curves at different field strengths, where ΔP and U_c are the pressure drop across the bed and the superficial velocity of the gas, respectively. The structural effect of particle alignment with the field is clearly seen in defluidization path III, $H > 0$. Along this path, ΔP is significantly lower compared to that required to fluidize the bed, despite the gas velocity being higher between points C and D than the minimum fluidization velocity U_{mf}. The field produced by the coils was $H = 2.395 \times 10^5$ A/m. Figure 44 is a plot of U_{exp} vs. U_φ at different currents in the coils, where U_{exp} is the measured sedimentation velocity in the fluid frame of reference and U_φ is the calculated (i.e., theoretical) value. The bed consists of 197 – 297-µm polyester-encapsulated magnetite particles, defined as magnetic sand, $\varrho_p = 2670\,\mathrm{kg/m^3}$, having a magnetic permeability of 1.64 µm and a shape factor $S_p = 0.85$. The volume fraction occupied by particles φ varied in the range of 0.45 – 0.50. The transition zone for the 10 – 25-A current levels (see Fig. 44) denotes transition from a stable to a bubbling bed. Initial bubbling observed in the vicinity of $U_\varphi = 0.3$ m/s indicates that ordered structures are turning into random ones. Only at this transition zone is $U_{exp} = U_\varphi$ expected to be satisfied. In the region prior to the transition zone, $U_{exp} > U_\varphi$ reflects the reduced resistance to the flow of the aligned particle structure. In the posttransition zone, $U_{exp} > U_\varphi$ persists due to bypassing of gas in the form of bubbles. The results in Figure 44 conform with this. At $U_\varphi \simeq 0.3$, $U_{exp} = U_\varphi$ for the 15 – 25-A current levels, and at $U_\varphi = 0.15$ m/s, U_{exp} is 46 % higher than U_φ. Figure 45 shows the *structural effect* in terms of calculated effective porosity $\varepsilon_{eff} = \varepsilon + \varepsilon_m$, where ε is the actual bed porosity and ε_m is the magnetic structural porosity, $\varepsilon_m \leq 0.1\,\varepsilon$, at the transition velocity $\varepsilon_m = 0$. The structural effects persist along defluidization paths (see Fig. 43). Figure 46 shows plots of ΔP_{exp} vs. ΔP_{cal} for fluidization (A) and defluidization (B) paths; ΔP_{exp} and ΔP_{cal} are pressure drops across the bed obtained experimentally and calculated by the Blake – Kozeny correlation, respectively. A good fit between ΔP_{exp} and ΔP_{cal} was obtained in the fluidization path; deviations increasing with current level were observed in the defluidization path. The lower pressure drops observed indicate that the increased alignment of particles persists throughout the transition from fluidized- to fixed-bed states. In summary, sedimentation of polarizable particles in a polarizing field (set parallel to the sedimentation velocity) can be enhanced by increasing the alignment of particles with the field lines.

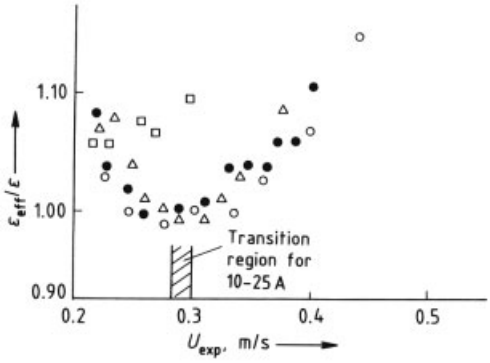

Figure 45. A plot of $\varepsilon_{eff}/\varepsilon$ vs. U_{exp}. [75], symbols as in Figure 44

3.4. Slurry Transportation in Horizontal Pipes

Slurry transportation in pipes must provide the necessary conditions to maintain a predesigned proportion of the solids, which are being carried by the fluid, in suspension. Prevention of undesirable sedimentation across the flow profile can be achieved at the expense of increasing pressure gradients along the pipe. Thus a definite relation must exist between pressure gradients associated with the axial flow and concentration gradients (formed by inertial effects, i.e., sedimentation) across it. Slurry flows in pipes are commonly characterized by the so-called *critical velocity.* This velocity marks the transition from a state in which particles are deposited on the pipe wall to a state in which they are in full inhomogeneous suspension in the stream.

Figure 47 is an idealized plot of $\ln(\Delta P/L)$ vs. $\ln U$ for the flow of a heterogeneous slurry, where $\Delta P/L$ is the pressure drop per unit pipe length due to frictional losses and U is the mean flow velocity. At high velocities, the slurry tends to behave as a uniform fluid, such as that represented by line AC. The critical or deposition

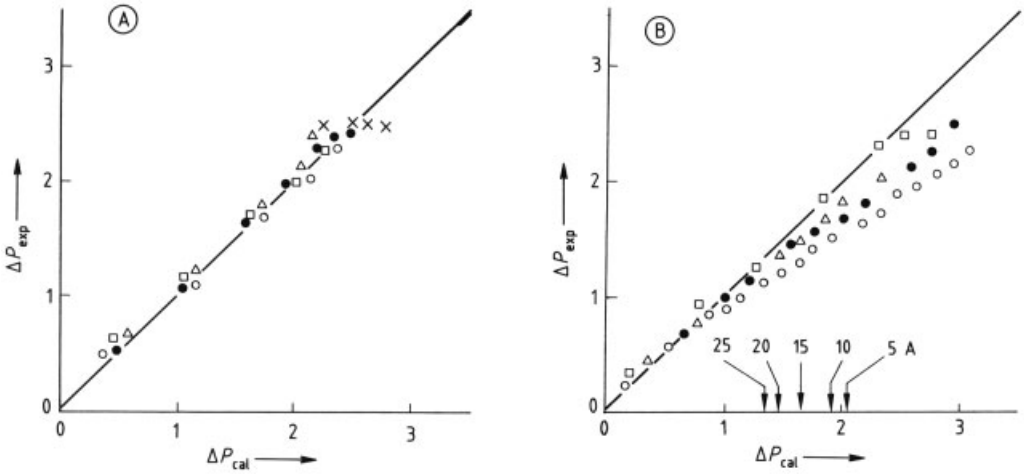

Figure 46. A plot of ΔP_{exp} vs. ΔP_{cal} for different current levels
A) Fluidization path; B) Defluidization path
× Fluidization, other symbols as in Figure 44

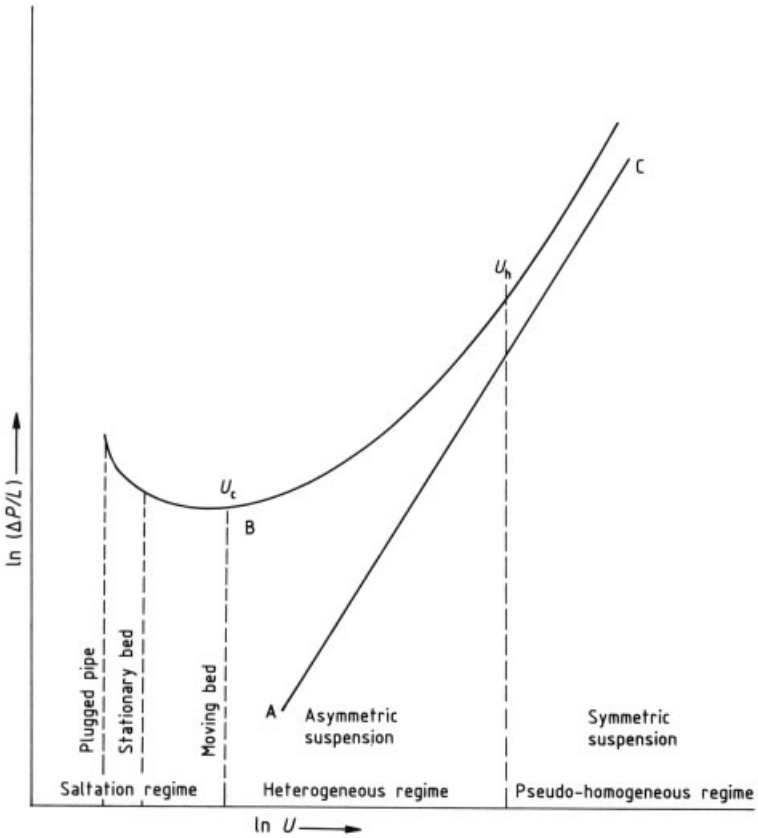

Figure 47. A schematic plot of $\ln(\Delta P/L)$ vs. $\ln U$ for slurry flow in horizontal pipes

velocity U_c corresponds to the minimum (point B) in the heterogeneous slurry curve. Above U_c the slope is positive in accordance with the effect of increasing flow velocity on increasing friction losses, whereas below U_c the slope is negative due to the existence of particle deposits that narrow the space available for flow. The critical velocity is a strictly turbulent phenomenon; the transport of slurries in pipelines must be carried out above the critical velocity, otherwise excessive sedimentation is likely to cause instabilities, erosion, and finally plugging of the pipe. In the transportation of polydisperse slurries, excessive segregation of large from small fractions and the ensuing action of the pipe as a classifier are undesirable and should be avoided. In Section 3.2, $C/C_a \geq 0.8$ was selected as a criterion for homogeneity of the flowing suspension. Similarly, heterogeneous flow can be defined by $C/C_a \leq 0.1$ (see Fig. 40). The condition $C/C_a \leq 0.1$ applies also to the moving bed and saltation regime in the $U < U_c$ range (see Fig. 47). Note that *saltation* implies a process whereby particles advance along the pipe by being alternately lifted from, and deposited on, a bed that may be either moving or stationary. Other criteria that depend on sedimentation velocity have been suggested [77], [78] in the form of $U_d/U* > 0.2$ and $U_d/U* > 0.13$, respectively. Using Equation (120) with $\beta_s = 1.0$ and $k_f = 0.4$, gives $C/C_a < 0.13$ and $C/C_a < 0.27$, respectively [65]. Figure 48 shows a different approach for determination of slurry flow regimes [79]. Data are given for 1.2 – 2.1-m/s flow velocities as a function of particle size and density. This figure summarizes data collected from commercial pipelines. It shows that homogeneous flow of dilute slurries is expected provided that particle size is smaller than 70 µm, whereas a twofold increase in this limiting size is expected for dense slurries. In a more conservative estimate, a limiting size of 200 µm, above which heterogeneous flow prevails, has been suggested [65]. Figure 48 shows that for dilute slurries (given the particle density), the heterogeneous flow regime is extended in terms of a two- to fourfold increase in particle size. This is expected in view of the increased sedimentation rates prevailing in low φ slurries (see Eqs. 1 and 2). The effect of particle size and flow rate, drawn for a specific set of ϱ_p, φ, and D data, on the formation of different types of slurry flows is shown in Figure 49 [17], [80]. The curve of critical velocity is marked by U_c. Homogeneous and heterogeneous suspensions characterize the regions marked A and B, respectively. Saltation and moving bed, and saltation and stationary bed prevail in the regions marked C and D, respectively. Schematic particle concentration profiles are also shown in Figure 49. Many correlations have been produced over the years for estimation of critical velocity. These correlations show either explicitly or implicitly the dependence of U_c on the sedimentation velocities of the transported particles. In formulating these correlations, use is often made of free sedimentation velocities. This does not conform to the substantial dependence of sedimentation velocity on φ. Hence although experimental data can be made to fit empirical correlation, imposing $\varphi = 0$ on sedimentation velocity does not seem warranted. The necessary correction factors that make the correlations work may not disclose the true mechanisms that control the flow of slurries and the critical velocity; more work is needed to resolve this discrepancy. In 1942, Wilson contended that U_c is proportional to the one-third power of U_d (evaluated at $\varphi = 0$) and suggested the following correlation

$$U_c = \frac{[K_w\,(\varrho_p/\varrho_f)\,\varphi\, g\, D\, U_d\,(\varphi = 0)]^{1/3}}{1 + \varphi\,(\varrho_p/\varrho_f - 1)} \tag{121}$$

where D is pipe diameter and K_w is a constant [81]. The pioneering work of Durand and coworkers [82], [83] provided a widely recognized empirical correlation. This correlation depends on a modified slurry Froude number Fr_L:

$$U_c = Fr_L\,[2\,g\,D\,(\varrho_p/\varrho_f - 1)]^{1/2} \tag{122}$$

Figure 50 shows a plot of Fr_L vs. particle size d for different levels of φ: Fr_L is most sensitive to d for particles smaller than 0.5 mm; for particles larger than 2 mm, Fr_L is practically independent of particle size. It has a value of 1.34, regardless of the system properties, with U_c being a function of $D^{1/2}$ and of ϱ_p/ϱ_f. These data [82], [83] were obtained using coal, sand, and gravel slurries containing 0.2 – 25-mm particles and flowing in pipes of 3.8 – 5.8-cm diameter. A plot of Fr_L vs. φ (Fig. 51) discloses a weak dependence of Fr_L and hence of U_c on φ [65], [70]. The plot shows the excellent fit provided by the Durand correlation to the experimental results of several researchers [82–86]. Results similar

to the ones plotted in Figure 50 were reported for transportation of tailings [17], [87]. The significance of sedimentation velocities in affecting U_c is also reflected in the pressure drops necessary to maintain a stable flow. The pressure drop ΔP in a pipe due to frictional losses is given by

$$\Delta P = f_s\,(L_t/D)\,2\varrho_f U^2 \tag{123}$$

where the slurry friction factor in heterogeneous flow is given by

$$f_s = f_l \left\{ 1 + K\varphi \left[\frac{g\,D}{U^2}\,(\varrho_p/\varrho_f - 1)\,/\,\sqrt{C_D} \right]^{3/2} \right\} \tag{124}$$

where f_s and f_l are friction factors for the slurry and the pure fluid, respectively; K is a constant ranging from 80 to 150 (for $U > U_c$, $K = 82$ [17], [65]); C_D is the particle drag coefficient, evaluated at $\varphi = 0$; L_t is the total effective length of the pipe:

$$L_t = L_p + L_e \tag{125}$$

where L_p and L_e are the length of straight sections of the pipe and the equivalent length of other components along the flow path (e.g., fittings, constrictions, valves, elbows, lees), respectively. If U is proportional to $U_d^n, n \leq 1$, then ΔP (via U^2) is proportional to U_d^{2n}. Equation (124) can be arranged as follows:

$$\frac{f_s - f_l}{f_l} \cdot \frac{1}{\varphi} = 82 \left[\left(\frac{U^2}{g\,D} \right) \cdot \frac{\varrho_f}{\varrho_p - \varrho_f} \sqrt{C_D} \right]^{-3/2}, \quad U > U_c \tag{126}$$

A plot of

$$\frac{f_s - f_l}{f_l} \cdot \frac{1}{\varphi} \text{ vs. } \frac{U^2}{g\,D} \cdot \frac{\varrho_f}{\varrho_p - \varrho_f} \sqrt{C_D}$$

is shown in Figure 52 [65]. Lower flow velocities, when D, ϱ_p, ϱ_f, C_D, and φ are fixed, give a larger relative increase of f_s (up to two orders of magnitude), whereas at high velocities this increase may be as small as 10 %. This increased friction factor (i.e., over pressure) results from concentration gradients across the pipe which arise due to sedimentation effects. NEWITT et al. studied the flow of different slurries containing polydisperse particle-size mixtures [80]. Table 2 gives details of materials used in these studies, which produced the flow regimes depicted in Figure 49. The range of densities of solids tested was 1180 – 4600 kg/m^3, the size range being 20 – 5970 µm. They proposed a direct relation between U_c and sedimentation velocity:

$$U_c = 17 U_\varphi\,(\varphi = 0) \tag{127}$$

Combining Equations (123) and (127) and letting $U = (1 + \alpha_s)\,U_c$, where $\alpha_s > 0$ denotes a safety factor, give

$$\Delta P = 578\,f_s \left(\frac{L_t}{D} \right) \varrho_f\,[(1 + \alpha_s) \cdot U_\varphi\,(\varphi = 0)]^2 \tag{128}$$

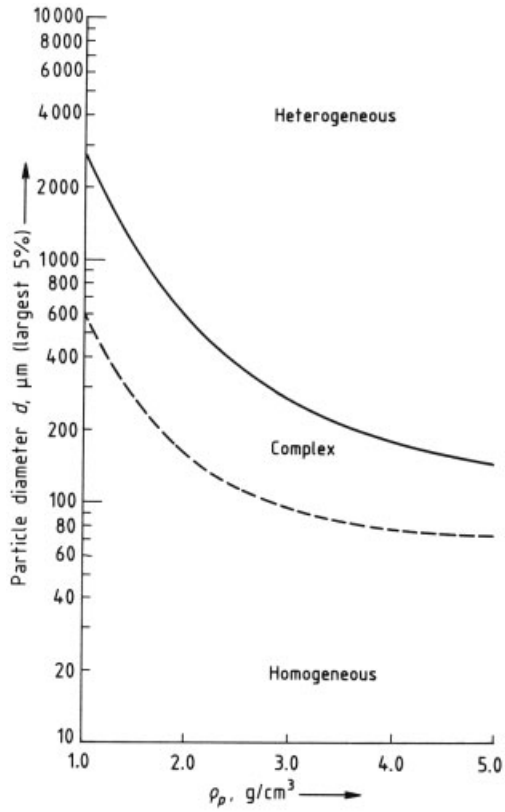

Figure 48. Slurry flow regimes as a function of particle diameter d and particle density ϱ_p, velocity = 1.2 – 2.1 m/s [78], [79]
—— Based on thick slurries with fine (< 44 µm) vehicle; – – – – Based on thin slurries or slurries with graded particle size

Because for turbulent flow $U_\varphi \sim d^{1/2}$, according to the Newitt model $\Delta P \sim d$. A linear dependence of pressure drop on particle size has direct implications on pipeline economics. The cost of size reduction prior to transportation

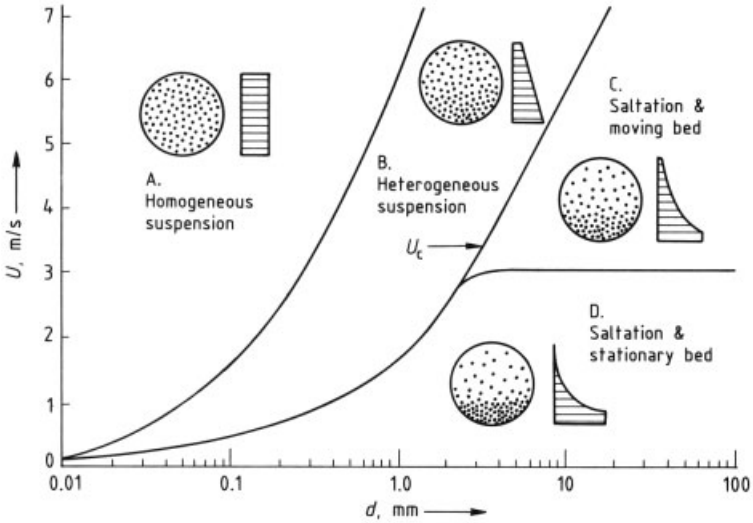

Figure 49. Flow regimes in a slurry pipeline defined in a plot of flow velocity U vs. particle size d [17], [80]

should be evaluated against expected savings in energy consumed by the pumps. The longer the pipe, the more favorable size reduction may be. Constraints on size reduction due to specification of end users and increased cost of dewatering should also be considered. Optimization involves determination of the pipe diameter that yields a minimum annual cost, where the annual cost consists of operating and investment costs. Increasing pipe diameter requires larger investment at lower operating cost; hence an optimum can be found. An illustration of single-phase pipe flow optimization is given in [88]. A nearly linear dependence of U_c on U_φ has also been reported [89].

$$U_c = (gD)^{0.5} \left[\frac{0.094\varphi\,(\varrho_p/\varrho_f - 1)}{\lambda} \cdot \frac{\varrho_f U_\varphi^3\,(\varphi=0)}{g\mu_f} \right]^{0.375} \tag{129}$$

where λ is resistance to the flow of carrier fluid through the pipe:

$$\lambda = 0.3164\,Re^{-0.25} \tag{130}$$

Equation (129) shows that $U_c \sim U_\varphi{}^{1.125}$, where U_φ is evaluated at $\varphi = 0$.

Published correlations diverge widely in predictions of critical deposition velocity U_c [90]. The value of U_c has been postulated as

$$U_c = f\,(d, D, \varphi, (\varrho_p - \varrho_f)\,g, \varrho_f, \mu_f)$$

or, in dimensionless form,

$$U_c = f\left[\left(\frac{d}{D}\right), \frac{D\,\varrho_f\,\sqrt{gd\,(\varrho_p/\varrho_f - 1)}}{\mu_f}, \varphi\right] \tag{131}$$

Table 2. Properties of materials used by Newitt et al [80]

Material	Density ϱ_p, kg/m^3	Terminal velocity U_d, m/s	Mean particle size d, cm
Perspex	1180	0.024	0.0584
	1180	0.036	0.1092
	1180	0.067	0.1981
Coal (anthracite)	1400	0.068	0.1321
	1400	0.102	0.2286
	1400	0.148	0.5969
Sand	2640	–	0.0020
	2640	0.001	0.0097
	2640	0.028	0.0208
	2640	0.113	0.0762
	2560	0.079	0.0483
Gravel	2600	0.192	0.1981
	2600	0.314	0.3124
	2550	0.326	0.3810
Manganese dioxide	4100	0.174	0.0965
	4100	0.253	0.1575
Zircon sand	4600	0.025	0.0109

The result of this analysis is

$$U_c = \left[\frac{15}{4} \cdot \frac{\varphi}{(1-\varphi)^{1-2n}} \cdot C_D U_0^2 \left(\frac{D}{d}\right) \left(\frac{D\varrho_f}{\mu_f}\right)^{1/8} \cdot \frac{1}{x_s}\right]^{8/15} \quad (132)$$

where $U_0 = U_\varphi(\varphi = 0)$, n is the exponent in $U_d = U_0(1-\varphi)^n$ which depends on the particle Reynolds number, $2.3 \leq n \leq 4.6$ [91], and x_s is a function of U_d/U_c.

$$x_s = \frac{2}{\sqrt{\pi}} \left[\frac{2}{\sqrt{\pi}} \gamma \exp\left(-\frac{4\gamma^2}{\pi}\right) + \int_\gamma^\infty \exp\left(-\frac{4\gamma^2}{\pi}\right) d\gamma\right] \quad (133)$$

where $\gamma = U_d/U_c$.

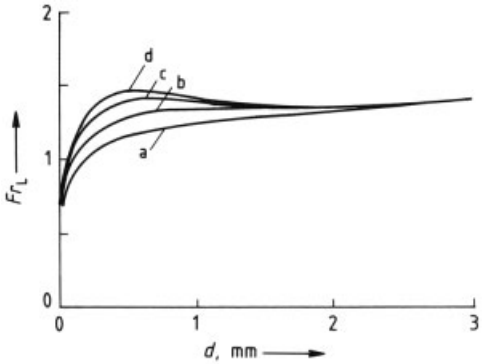

Figure 50. Variation of the modified slurry Froude number Fr_L as a function of particle diameter d [65], [82]
a) $\varphi = 2\,\%$; b) $\varphi = 5\,\%$; c) $\varphi = 10\,\%$; d) $\varphi = 15\,\%$

Figure 53 shows a plot of x_s vs. U_d/U_c calculated from Equation (133). The decrease in x_s is nearly linear in the 0.5 – 1.5 range of U_d/U_c. A simplified version of Equation (132) is given by

$$U_c = \left[5\varphi(1-\varphi)^{2n-1} \left(\frac{D}{d}\right) \frac{D\varrho_f \sqrt{gd(\varrho_p/\varrho_f - 1)}}{\mu_f} \cdot \frac{1}{x_s}\right]^{8/15} \sqrt{gd(\varrho_p/\varrho_f - 1)} \quad (134)$$

The following is a correlation obtained from regression analysis of published data [91], [90]:

$$U_c / \sqrt{gd(\varrho_p/\varrho_f - 1)} = 1.85\,\varphi^{0.1536} \cdot (1-\varphi)^{0.3564} \left(\frac{D}{d}\right)^{0.378} Re^{0.09}\, x_s^{0.3} \quad (135)$$

Equations (134) and (135) provide the fewest deviations from experimental data compared to other known correlations [91]. Transition velocity from heterogeneous to homogeneous flow (U_h in Fig. 47) depends on sedimentation velocities (see Section 3.2 and Eq. 120). Other correlations that do not make use of an arbitrary ratio of concentrations are listed below. These correlations show a weaker dependence on sedimentation velocity compared to those concerned with critical deposition velocity. Equation (136) was derived by Govier and Charles [92], [93]:

$$U_h = 11.9\,(U_0 D)^{1/2}\, d^{-1/4}, \quad U_0 = U_\varphi(\varphi = 0) \quad (136)$$

whereas according to Newitt [92], [80], U_h is given by

$$U_h = 38.7\,(g D U_0)^{1/3} \quad (137)$$

The plot of $U_0/U*$ vs. $dU*/\nu_f$ in a generalized phase diagram for evaluation of transport of suspensions in pipes is shown in Figure 54; ν_f is the kinematic viscosity $\nu_f = \mu_f/\varrho_f$ [65], [94]. The solid line indicates initial particle movement from a state of rest on the pipe wall. Regions of formation of transverse and longitudinal waves, and of heterogeneous and homogeneous flow, are also indicated (δ denotes the thickness of the viscous sublayer and is given by $\delta = 11.6\,\nu_f/U*$ [65].

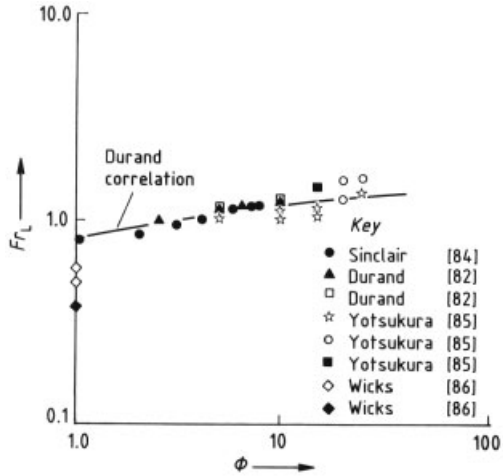

Figure 51. A plot of Fr_L vs. φ, literature data, and the Durand Correlation [65], [70]

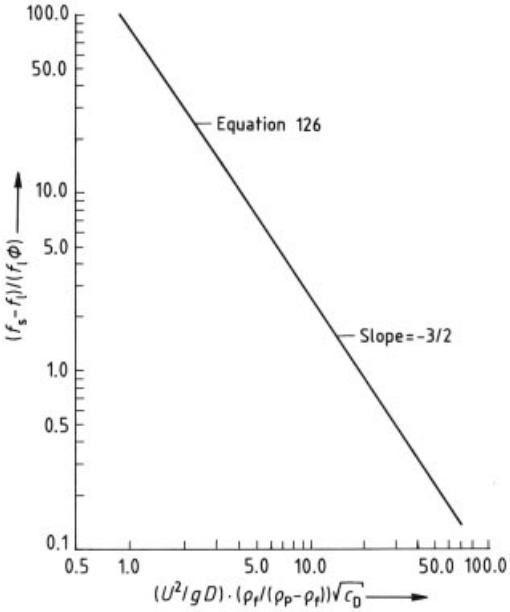

Figure 52. A plot of $(f_s - f_1)/(f_1\varphi)$ vs. $\frac{U^2}{gD}\frac{\varrho_f}{\varrho_p - \varrho_f}C_D\sqrt{\pi}$ [65]

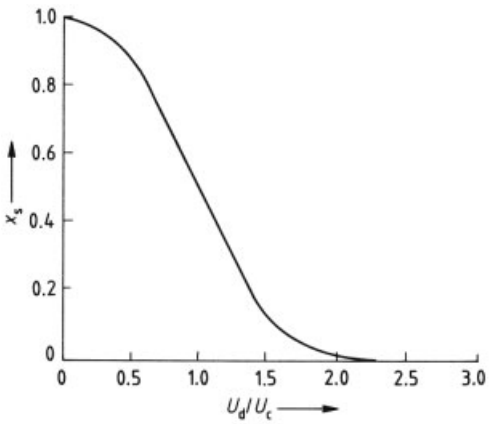

Figure 53. A plot of x_s vs. U_d/U_c [90], [91]

This plot shows that the effect of $U_0/U*$ depends on d/δ, being largest for $d/\delta < 1$ and smallest in the $d/\delta > 6$ range. Particle Reynolds numbers are marked on the solid line indicating the Stokes, intermediate, and turbulent flow regimes associated with particle motion.

Centrifugal forces affect sedimentation of particles suspended in slurries that flow in bent pipes. Examples are flow in elbows and in helical pipes. If the resulting value of g' exceeds g, then prevention of deposition or transition from homogeneous to heterogeneous flow must occur at the expense of increasing the frictional velocity $U*$.

In the transportation of *polydisperse slurries,* the effect of attendant distributions must be accounted for. WASP et al. used the concept of a two-phase vehicle [65], [69], [95]. According to this, the slurry is divided into homogeneous and heterogeneous parts. The concentration and the size distribution of particles occurring at the top of the pipe are assumed to exist throughout. This concentration is considered part of the homogeneous or "vehicle" portion of the flow. The remainder of the polydisperse mixture is considered as a heterogeneous suspension transported by this vehicle. Analysis of flow requires the determination of particle split between the vehicle and the heterogeneous portion of the slurry. This split depends on $U*$, which cannot be determined until the friction factor is found; hence an iterative solution is called for. Applying Equation (120), with $\beta_s = 1$, $k_f = 0.4$, to the ith fraction gives

$$C_i = C_{ai}\exp\left(-10.362\,U_{0i}/U^*\right) \tag{138}$$

in which C_i is added to the vehicle and the balance to the heterogeneous part of the suspension. Thus $C_h = \sum_{i=1}^{n} C_i$ and $C_{ht} = \sum_{i=1}^{n} C_{ai} - C_i$, where C_h and C_{ht} are total concentrations in the vehicle and in the heterogeneous part of the suspension, respectively. Thus in each iteration, U_{0i} for all particle fractions and $U*$ must be determined, and the contribution of each heterogeneous fraction must be accounted for, the total frictional loss being due to both the vehicle and the heterogeneous part of the slurry [65], [92]. Again, evaluation of free sedimentation velocity (i.e., U_φ at $\varphi = 0$) does not conform to the physics of the flow and sedimentation systems. The successfulness of using U_0 instead of U_φ at $\varphi > 0$, particularly for dense slurries, seems to be due primarily to the adjustment of empirical constants.

Furthermore, slurry flow and particle sedimentation in dense slurries are likely to involve non-Newtonian rheology. Details of slurry pipeline hydraulics, treatment of different slurry rheologies and description of computational models are given in [96].

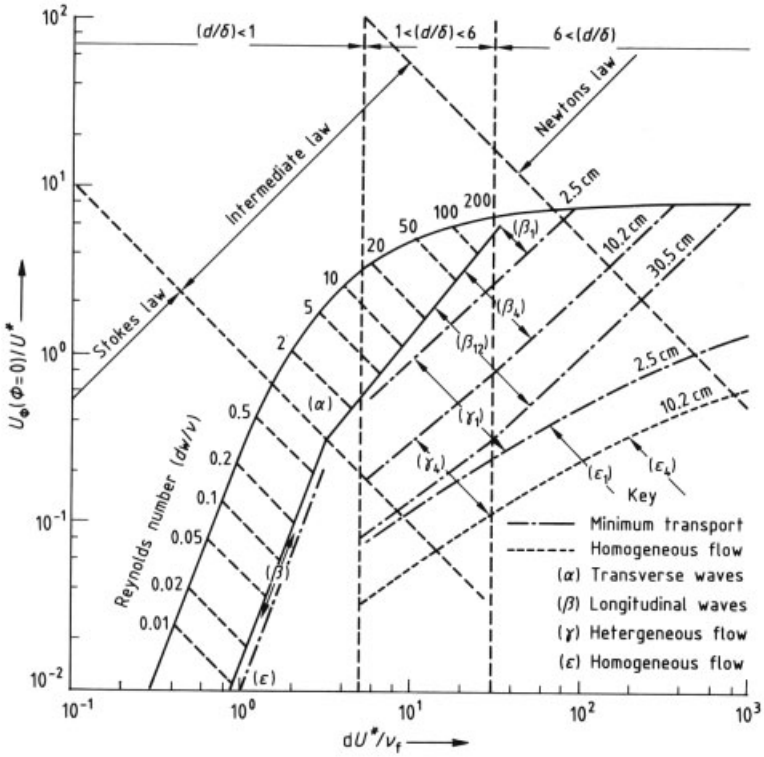

Figure 54. A generalized phase diagram for transport of suspensions [65], [94]
Range of β, γ, and ε phases apply to density ratio $(\varrho_P - \varrho_f)/\varrho_f = 1.65$

3.5. Sedimentation in Centrifugal Fields

The use of centrifugal fields to enhance sedimentation of fine and slow-settling particles is well known. Sedimentation of diffusing particles in centrifugal fields is used for separation, concentration, pellet formation, sample preparation, and as an analytical means for particle characterization. *Ultracentrifugation* generally implies sedimentation of diffusing particles in centrifugal fields inclusive of those that provide the highest (i.e., ultra) g forces known to date in commercial centrifuges. Because this technique incorporates many aspects of sedimentation, it is introduced first.

3.5.1. Ultracentrifugation

Ultracentrifugation is used extensively in biochemistry (→ Biochemical Separations, Chap. 4.2.), where analysis of molecules, macromolecules, cells, and other microscopic matter of organic origin is often required [31], [97]. Particulates in this category are generally dispersed in a carrier fluid, the dispersions thus formed being uniform under ordinary conditions. Analysis of the contents of such dispersions often requires redistribution of the dispersed entities so that their characteristic signals can be picked up. This can be achieved during or after sedimentation in ultracentrifugal fields. The ultracentrifuge generally consists of a driving mechanism, control unit, centrifugation chamber, and rotor that holds the sample containers. Different types of rotors are available, and the purpose of the runs dictates rotor selection.

Figure 55 shows the LKB Ultrospin preparative ultracentrifuge, and typical material bands produced in sample tubes that were run in different rotors. This high-speed programmable unit is available in three models that can be oper-

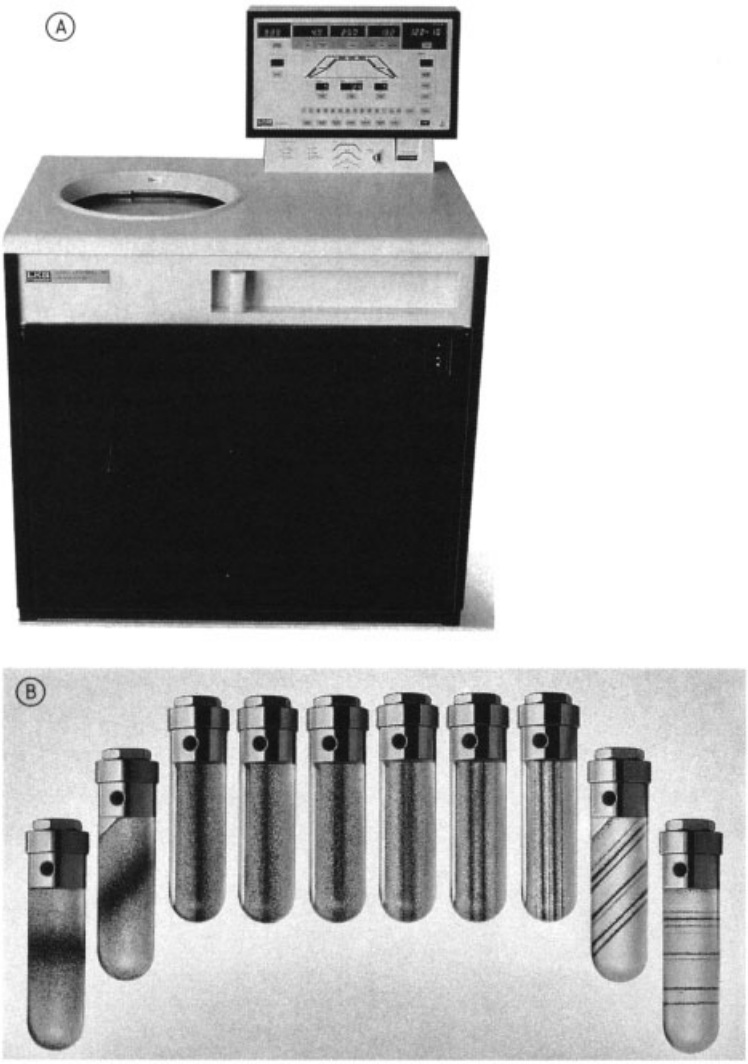

Figure 55. A) The LKB Ultrospin preparative ultracentrifuge; B) Material patterns in sample tubes spun with different types of rotors

ated at maximum speeds of 55 000, 70 000, and 83 000 rpm, which are equivalent to ratings of 393 600 g, 504 000 g, and 615 700 g as maximum values of g', respectively.

Figures 56, 57, 58 show typical *fixed-angle, swing-out,* and *vertical rotors,* respectively. One of the many available models of each type of rotor is shown at the side of each figure. The tilted tubes in the *fixed-angle rotor* provide conditions for sedimentation of diffusing particles that can be considered as inclined channels (see Section 2.8); observe the inclined bands in the second and ninth tubes of Figure 55. *Swing-out rotors* provide ordinary conditions for batch sedimentation; note the bands of particles in the first and tenth tubes in Figure 55. *Vertical rotors* produce sedimentation across the tube, see bands in the third to eighth tubes in Figure 55. Standard ultracentrifugation equipment, such as the Beckman series, is not designed to provide g' levels exceeding 300 000 g. The Beckman *elutriator rotor* incorporates counterflow of the carrier fluid to separate slower settling from fast-settling particles. Figure 59 shows the elutriator rotor and

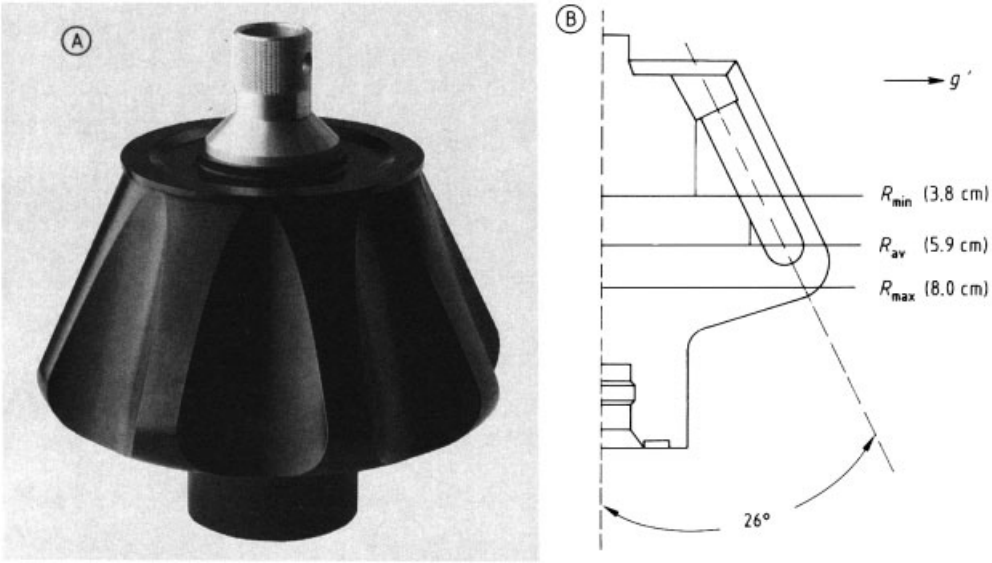

Figure 56. Fixed-angle rotors
A) Apparatus; B) Cross-section

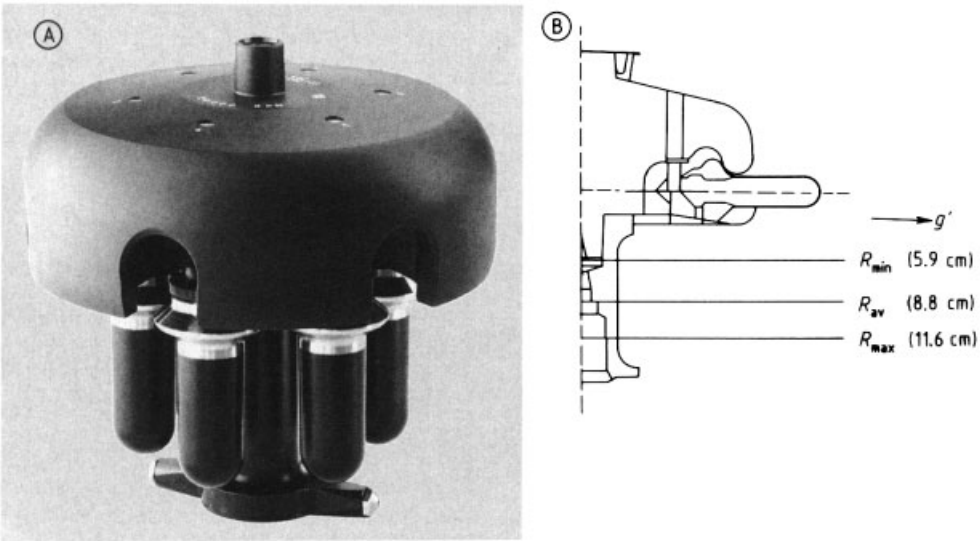

Figure 57. Swing-out rotors
A) Apparatus; B) Cross-section

the mechanism of elutriation. The suspension is introduced via a narrow inlet port into a tapered cell (stage 1), where it is subjected to centrifugal force and to a counteracting hydrodynamic drag of the carrier fluid (stage 2). Larger and heavier particles undergo faster sedimentation against the fluid currents, whereas slower ones are washed away and thus separated (stage 3). This high-recovery method is used for separation of whole cells and large subcellular particles. Ultracentrifugal analysis can provide information on diffusion coefficient D and sedimentation coefficient S_g which were defined in conjunction with Equation (69). These parameters can then be used to evaluate molecular mass and to predict sedimentation under different op-

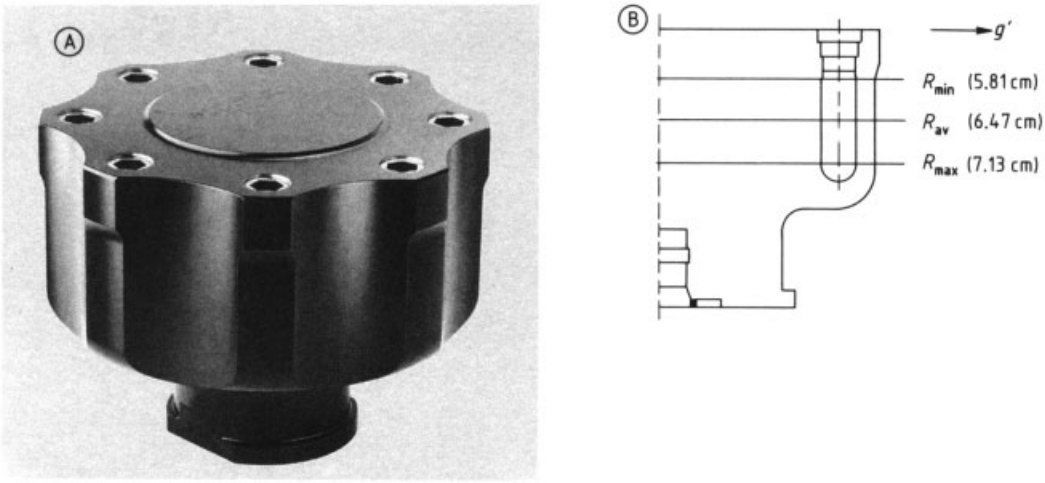

Figure 58. Vertical rotors
A) Apparatus; B) Cross-section

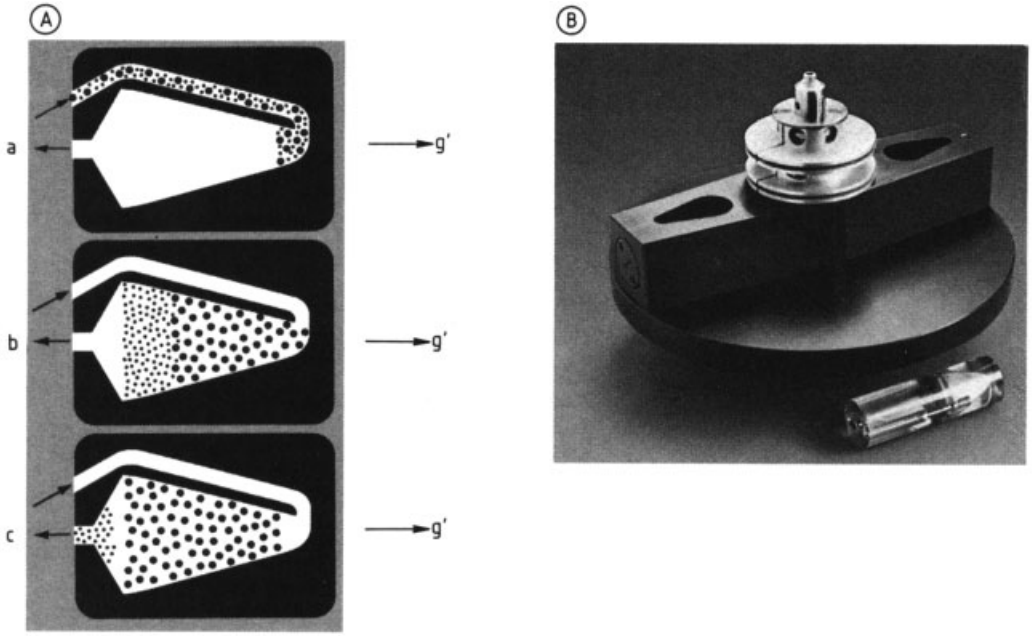

Figure 59. The Beckman elutriator rotor
A) Mechanism of elutriation (schematic):
a) Sample suspended in medium enters chamber; b) Sedimentation tendency of particles balanced by counterflow; c) Flow increased. Slow sedimenting particles elutriated from chamber;
B) Rotor

erating conditions. Measurement of D and S_g can be done by either the sedimentation velocity or the sedimentation equilibrium method. In the sedimentation velocity method which involves relatively high rotor speeds, a moving boundary that separates particle-devoid fluid from the dispersion is driven by centrifugal force toward the cell boundaries. If $\varrho_P - \varrho_\varphi > 0$, motion is away from the axis of rotation and toward the cell bottom. Figure 60 shows a plot of concentration and concentration gradient profiles vs. distance from the axis of rotation [31]. As time proceeds, the boundary of fluid devoid of particles moves to the right and the variance of concentration gradients (i.e., their spread) increases due to diffusion.

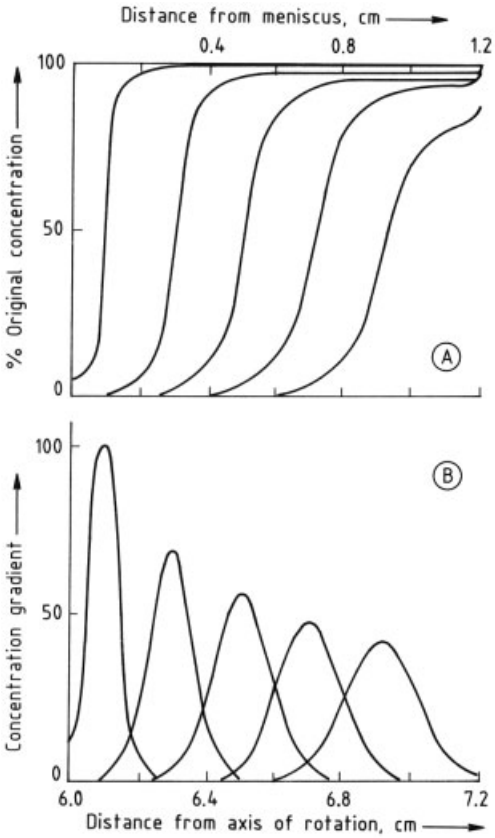

Figure 60. Schematic plots of: A) concentration, and B) concentration gradient vs. distance from axis of rotation at different elapsed times from the start of a sedimentation velocity experiment [31]

Such boundary motion is generated in sedimentation velocity experiments. In sedimentation equilibrium, sufficient time is allowed for the system to reach steady, zero-flux conditions in the sedimentation cell. Relatively low rotor speeds are used to bring the dispersion to a state where sedimentation drift is balanced by counterdiffusion throughout the cell. Figure 61 shows schematically the variation of concentration and concentration gradient profiles vs. distance from the axis of rotation occurring during a typical sedimentation equilibrium experiment. The curves are numbered in order of increasing time. A typical equilibrium profile is characterized by equilibrium concentration having approximate values of $C_0/2$, C_0, and $2\,C_0$ at the cell meniscus, midpoint, and bottom (6.0, 6.6, and 7.2 cm, respectively, in Fig. 61).

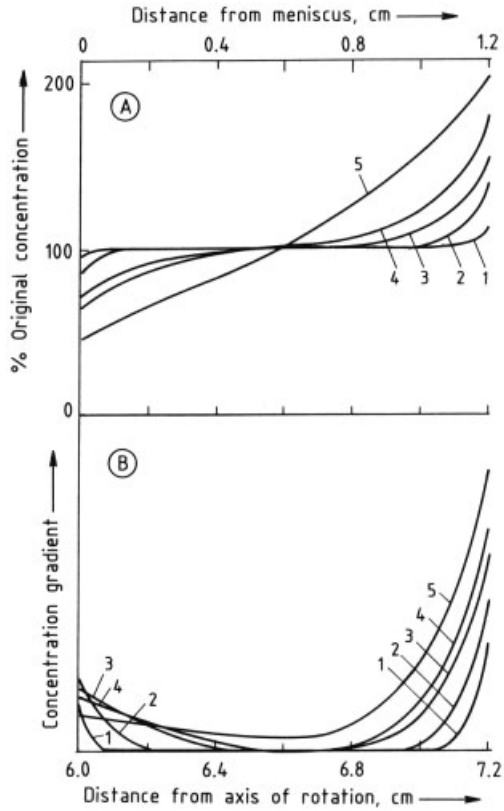

Figure 61. Schematic plots of: A) concentration, and B) concentration gradient vs. distance from axis of rotation, at different elapsed times from the start of a sedimentation equilibrium experiment

Sedimentation Coefficient. The sedimentation coefficient is defined as sedimentation velocity per unit acceleration g'. In the fluid frame

of reference, the sedimentation coefficient $S_{g\varphi}$ of monodisperse particles is then given by

$$S_{g\varphi} = \frac{U_\varphi}{g\prime} = \frac{U_\varphi}{\omega^2 r} \tag{139}$$

whereas in the dispersion frame of reference,

$$S_g = \frac{U_d}{\omega^2 r} = \frac{U_\varphi\,(1-\varphi)}{\omega^2 r} = \frac{U_\varphi\left(1 - C\tilde{N}V_p\right)}{\omega^2 r} \tag{140}$$

The sedimentation coefficient (in the dispersion frame of reference) of the *i*th particle fraction pertaining to a polydisperse mixture depends on coefficients of the rest of the distribution as follows:

$$S_{gi} = S_{g\varphi i} - \sum_i S_{g\varphi i}\,\varphi_i \tag{141}$$

Because $U_d = dr/dt$,

$$S_g = \frac{dr/dt}{\omega^2 r} = \frac{1}{\omega^2} \cdot \frac{d\ln r}{dt} \tag{142}$$

The dimension of S_g is seconds.

In ultracentrifugation, one unit of sedimentation coefficient is called one Svedberg. One Svedberg equals 10^{-13} s. The rate of change of S_g with C is given by

$$\partial S_g/\partial C = \frac{1}{\omega^2 r}\left[\frac{\partial U_\varphi}{\partial C}\left(1 - C\tilde{N}V_p\right) - U_\varphi \tilde{N}V_p\right] \tag{143}$$

in which Equation (140) has been used. Combining Equations (69) and (143) gives

$$\partial C/\partial t = D'\left(\frac{\partial^2 C}{\partial r^2} + \frac{1}{r}\cdot\frac{\partial C}{\partial r}\right)$$

$$- S_g\,\omega^2\left(r\cdot\frac{\partial C}{\partial r} + 2C\right)$$

$$+ \frac{\partial C}{\partial r}\left\{\frac{\partial D'}{\partial C}\cdot\frac{\partial C}{\partial r} - \left[\frac{\partial U_\varphi}{\partial C}\right.\right.$$

$$\left.\left.\cdot C\left(1 - C\tilde{N}V_p\right) - U_\varphi C\tilde{N}V_p\right]\right\} \tag{144}$$

The relationship $\frac{\partial U_\varphi}{\partial C} = U_0 \cdot \frac{\partial f(\varphi)}{\partial C}$ can be obtained from Equation (1), and if the dependence of D' on C is known (or use is made of Equation 66), then Equation (144) can be solved without the need to impose the constraint that D' and S_g be constants. In dilute dispersions, provided that concentration and concentration gradients are low, the assumption that D' and S_g are constants seems to be justified. In the range $r < r_1$ and $r > r_2$, $\partial C/\partial r$ and thus $\partial^2 C/\partial r^2$ practically vanish (see Fig. 62). Hence, Equation (143) reduces to

$$\partial C/\partial t = -2\omega^2 S_g C$$

which upon integration gives

$$C\,(t) = C_0 \exp\left(-2\omega^2 \int_o^t S_g\,dt\right) \tag{145}$$

$C(t)$ and C_0 are concentrations at time t and $t = 0$, respectively. If S_g is constant, then

$$C\,(t) = C_0 \exp\left(-2\omega^2 S_g t\right) \tag{146}$$

In the sedimentation of nondiffusing monodisperse particles, the transition zone would be infinitely sharp, i.e., $r_2 - r_1 \to 0$ in Figure 62. Under these conditions, the velocity of the resultant moving boundary would be the same as the sedimentation velocity of particles ahead of the boundary within the dispersion. In sedimentation of diffusing particles, the boundary tends to spread out as time elapses (see Fig. 60). In this case, the peak of the concentreation gradient profile does not move at the same velocity as particles within the dispersion. The point $\bar{r}$ that does satisfy this condition (i.e., having the same velocity as particles that move within the suspension) is given by the square root of the normalized second moment of the concentration gradient curve [31], [98]:

$$\bar{r}^2 = \int_{r_m}^{r_p} r^2\,(\partial C/\partial r)\,dr/C_p\,,\quad C_p = \int_{r_m}^{r_p} (\partial C/\partial r)\,dr \tag{147}$$

where r_m and r_p are the distances of the meniscus and of a point in the dispersion at which $\partial C/\partial r = 0$ and $C = C_p$, respectively. In deriving Equation (147), a meniscus devoid of particles is assumed. The actual difference between $\bar{r}$ and the location of the peak may be small. In such cases, use of the latter can be justified. As the transition zone, i.e., the peak in Figure 62 (see also Fig. 60), moves away from the meniscus, interparticle spacings become larger due to the increased intensity of the centrifugal field. This phenomenon is known as radial dilution, the end result being time-decreasing concentration. Using Equation (142) for $r = \bar{r}$ in the plateau region (see Fig. 62) gives $\partial C/\partial t = -\frac{2C}{r}\cdot\frac{\partial \bar{r}}{\partial t}$, which

upon integration yields the radial dilution equation:

$$C(t)/C_0 = (r_m/\bar{r})^2 \tag{148}$$

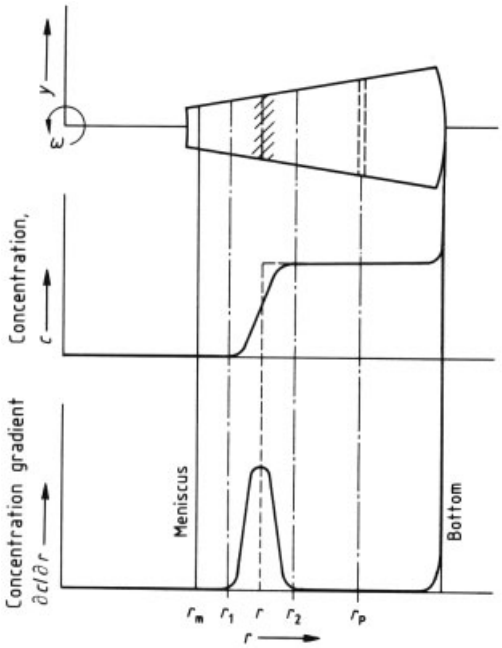

Figure 62. Concentration and concentration gradient profiles during sedimentation in a sector cell [31], [99]

Thus the advance of the concentration transition zone (i.e., the peak) from the meniscus at r_m to a new position $\bar{r}$ is associated with a decrease in concentration of the dispersion from C_0 to $C(t)$. In a polydisperse mixture that forms a series of moving boundaries [99],

$$C_0 = \sum_{i=1}^{n} (\bar{r}/r_m)_i^2 \, C_i(t) \tag{149}$$

where i denotes the ith component of a mixture comprising n components. If the dependence of S_g on concentration is significant, then radial dilution must also be accounted for. The suggestion has been made that the standard sedimentation coefficient $S_{g(20,w)}$ be defined with respect to sedimentation in pure water at 20 °C; $S_{g(20,w)}$ is related to S_g by [100]

$$S_{g(20,w)} = S_g \cdot \frac{\mu_f}{\mu_{f(20,w)}} \cdot \frac{(1 - v\,\varrho_\varphi)_{20,w}}{1 - v\,\varrho_\varphi} \tag{150}$$

where v is the partial specific volume, which can be considered as the volume occupied by 1 g of the dispersed phase (e.g., solute) in the dispersion [97].

$$v = (dV_\varphi/dW_s)_{V_\phi \to \infty} \tag{151}$$

where V_φ is the volume of the dispersion and W_s is the weight of the dispersed phase. Note that $v = 1/\varrho_p$, where ϱ_p is the density of the dispersed phase in the dispersion. Different values of v and ϱ_p may apply to materials in pure and in dispersed (or dissolved) states.

Equation (150) does not account for the full effect of concentration and is valid if $\varphi \to 0$, i.e., $C \to 0$. In view of Equations (1) and (140), $S_{g(20,w)}$ can be related to S_g as follows:

$$S_{g(20,w)} = S_g \frac{[U_\varphi(1-\varphi)]_{20,w}}{U_\varphi(1-\varphi)}$$

$$= S_g \frac{(U_0)_{20,w}}{U_0} \cdot \frac{[f(\varphi)(1-\varphi)]_{20,w}}{f(\varphi)(1-\varphi)} \tag{152}$$

Equation (152) shows that in transforming S_g to $S_{g(20,w)}$, the combined effect of U_0 and of $f(\varphi)(1-\varphi)$ must be accounted for. In the moving-boundary method (Eq. 142), $\ln \bar{r}$ is plotted vs. t and the slope yields $\omega^2 S_g$. If S_g is constant, a straight line is obtained. Figure 63 shows a plot of log r vs. t for sucrose, which indicates a constant S_g over the range of distance measured [101]. In the *transport method,* S_g is determined by measuring the total amount of solute transported across an arbitrary surface in the sedimentation cell. This method is useful for measuring S_g of small molecules. Material balance in the sector $r_m \le r \le r_p$ (see Fig. 62) gives [31]

$$\frac{1}{2}\beta_c C_0 (r_p^2 - r_m^2) - \int_{r_m}^{r_p} \beta_c C r dr$$

$$- \int_0^t \beta_c r_p C \omega^2 r_p S_g dt = 0 \tag{153}$$

where β_c is a constant depending on cell geometry. The first, second, and third terms represent the amount of dispersed material at $t = 0$ and $t > 0$, and the amount accumulated from $t = 0$ to t, respectively. Combining Equations (146) and (153), if S_g is constant, and integrating give

$$C_0 (r_p^2 - r_m^2) - 2 \int_{r_m}^{r_p} C r dr = r_p^2 C_0$$

$$\cdot \left[1 - \exp\left(-2\omega^2 S_g t\right)\right] \tag{154a}$$

Rearranging Equation (154 a) yields

$$S_g = -\frac{1}{2\omega^2 t} \cdot \ln \left[2 \int_{r_m}^{r_p} C r \mathrm{d}r / \left(C r_p^2 C_0 \right) + r_m^2 / r_p^2 \right] \tag{154b}$$

If t is given, and $C = C(r)$ is known in the $r_m \leq r \leq r_p$ range, then S_g can be determined directly from Equation (154). Equation (154) is valid, irrespective of the formation or absence of a clear boundary between solution and supernatant, if a plateau region exists at $r = r_p$. Thus it is useful in the ultracentrifugation of small molecules that do not form well-developed moving boundaries. Application of Equation (154) requires direct measurements of C. This can be done by using optical absorption or analytical chemistry techniques. BALDWIN [102] transformed Equation (154 b) so that methods in which $\partial C/\partial r$ is measured (i.e., *schlieren optics*) can be applied [31], [97]. The result is as follows:

$$S_g = -\frac{1}{2\omega^2 t} \ln \left\{ 1 + \left[r_m^2 \int_{r_m}^{r_p} (\partial C/\partial r)\, \mathrm{d}r - \int_{r_m}^{r_p} r^2 (\partial C/\partial r)\, \mathrm{d}r \right] / \left(r_m^2 C_0 \right) \right\} \tag{155}$$

A plot of the negative of the logarithmic expression vs. t for data corresponding to Equation (154) or (155) is expected to give a straight line, whose slope is $2\omega^2 S_g$ [103], [104].

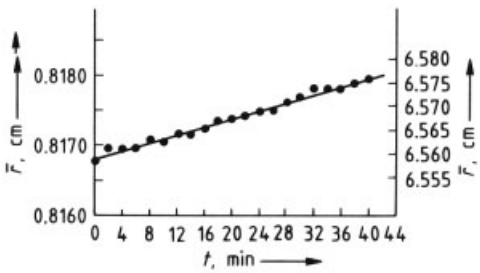

Figure 63. A plot of log $\bar{r}$ and $\bar{r}$ vs. t for sedimentation of sucrose [31], [101]

Differential Sedimentation Rates of Monodisperse Particles. Differential sedimentation rates can be used to determine the dependence of S_g on concentration. Differential sedimentation rates are defined with respect to movement of a boundary formed between solutions (or, generally, dispersions) that are identical except for having different concentrations. Figure 64 shows schematically a system for measuring differential sedimentation rates; a_1 and a_2 mark imaginary moving planes of "zero flux." These planes move at the same velocity as local particles; hence no particles cross them, and particle flux (in the plane frame of reference) vanishes. A boundary is shown between the upper (concentration C_1) and lower (concentration C_2) sections. Because no flux enters or leaves the control zone between a_1 and a_2, the associated amount of solute (e.g., particles) is independent of time. The volume of the control zone is nevertheless time dependent. Differential sedimentation rates in the C_1 and C_2 sections give rise to motion of the boundary in the control zone. Let the variation of $\omega^2 r$ across the control zone be negligible and assume that S_g is solely a function of C such that $\partial S_g/\partial C < 0$. If $C_1 = C_2$, then (see Fig. 64 A) the relative position of the boundary b (which here is also imaginary) with respect to the control zone is time independent. If $C_1 < C_2$ (Fig. 64 B), then $S_{g1} > S_{g2}$ and the solute enters at a faster rate than it leaves the boundary b. Accumulation of solute in b occurs until the concentration rises to C_2. The net result is displacement of the boundary in the frame of reference of the control zone toward a_1. If $C_1 > C_2$, then $S_{g1} < S_{g2}$ and particles in the C_2 zone will move faster than the boundary. An unstable low-density zone will be formed below the boundary, which is then expected to be deformed and probably eliminated by convection currents. Stable motion of the boundary toward a_2 (see Fig. 64 C) can occur if $\partial S_g/\partial C > 0$ and $C_2 > C_1$ are satisfied. From continuity (see Eq. 31), because the system is one dimensional and U_d is a sole function of C,

$$\mathrm{d}C/\mathrm{d}t + C\mathrm{d}U_d/\mathrm{d}r + U_d\, \mathrm{d}C/\mathrm{d}r = \mathrm{d}C/\mathrm{d}t + C\, (\mathrm{d}U_d/\mathrm{d}C)\, (\mathrm{d}C/\mathrm{d}r) + U_d\, (\mathrm{d}C/\mathrm{d}r) \tag{156}$$

Multiplying Equation (156) by $(\mathrm{d}r/\mathrm{d}C)/(\omega^2 r)$ gives

$$S_{gD} + S_g + C\mathrm{d}S_g/\mathrm{d}C = 0 \tag{157}$$

where $S_{gD} = (\mathrm{d}r/\mathrm{d}t)/(\omega^2 r)$ is the differential sedimentation coefficient of the boundary that moves at velocity $\mathrm{d}r/\mathrm{d}t$. For $\mathrm{d}S_g/\mathrm{d}C < 0$ and $C_1 < C_2$, the boundary will move toward a_1 at a slower rate than the sedimentation rate of particles in zone C_1.

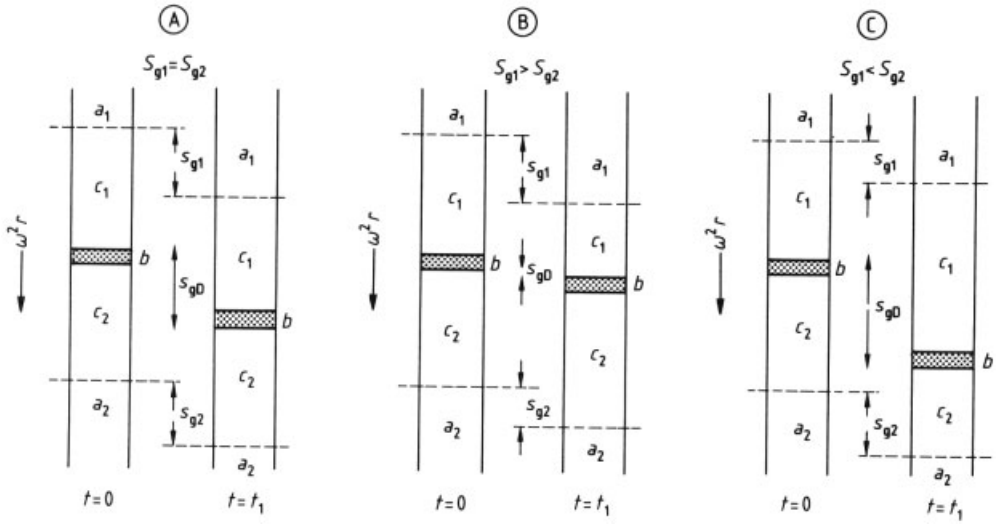

Figure 64. Schematic setup for measurement of differential sedimentation rates [31]

Diffusion Coefficient. The diffusion equation, for the static (i.e., $U_d = 0$) case and constant diffusion coefficient, is obtained readily from Equation (61):

$$\partial C/\partial t = D' \nabla^2 C \tag{158}$$

Equations (159) and (161) present two integrated forms of Equation (158) for a one-dimensional (x-oriented) diffusion,

$$\mathrm{d}C/\mathrm{d}x = \left(C_0/\sqrt{4\pi D' t}\right) \exp\left(-\zeta^2\right) \tag{159}$$

$$\zeta = x/\sqrt{4D't} \tag{160}$$

$$C = \frac{C_0}{2}\left[1 - \frac{2}{\pi}\cdot\int_0^y \exp\left(-\zeta^2\right)\mathrm{d}\zeta\right] \tag{161}$$

Figure 65 shows a plot of time-dependent concentration (Eq. 161) and concentration gradient (Eq. 159) profiles around a static and initially sharp boundary between solution and solvent. Various methods for measuring D' have been described in detail [105]. Equation (159) is suitable for use when the experimental system provides direct values of dC/dx, whereas Equation (161) requires direct determination of C. Because Schlieren optics is the standard optical system in ultracentrifuges [97] and this method provides direct data on concentration gradients, Equation (159) seems to be a convenient choice.

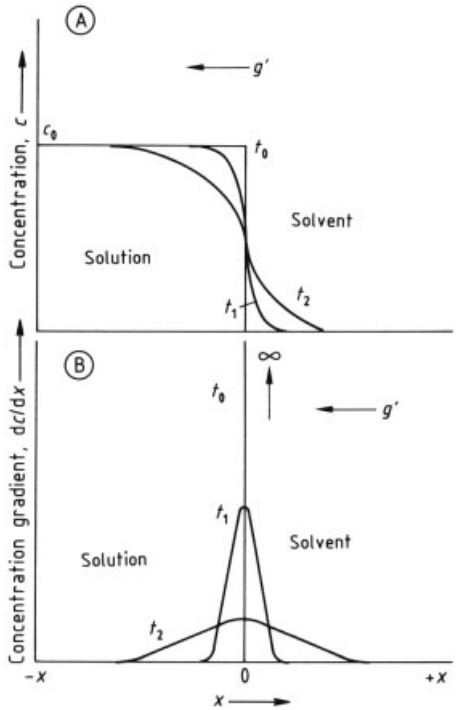

Figure 65. A plot of: A) Equation (161), and B) Equation (159). Area under the time dependent curve in B) is constant [97]

At the center of the boundary, $x = 0$ and the value of dC/dx is at a maximum.

$$(\mathrm{d}C/\mathrm{d}x)_0 = C_0/\sqrt{4\pi D' t}, \quad x=0 \tag{162}$$

Schlieren optics provides peaks that are directly related to the dC/dx peaks shown in Figure 65. The area under the schlieren peak is directly proportional to C_0, which is given by the area under the peak in Figure 65. If h_s and A_s denote the height and the area of the schlieren peak, then

$$D' = \frac{(A_s/h_s)^2}{4\pi t} \tag{163}$$

A plot of $(A_s/h_s)^2$ vs. $4\pi t$ gives D' as the slope of a straight line. Note that Equation (159) can be obtained as a solution for diffusion of a concentration pulse [28]. In diffusion that follows a pulse, the variable dC/dx in Equation (159) (and Fig. 65) is equivalent to the concentration. Alternatively, the pulse effect is to transform ordinary concentration gradients that evolve from a boundary into concentration profiles and vice versa. Determination of D' by ultracentrifugation can be done by starting with Equation (144) and applying simplifying assumptions such as constant D'.

The fact that the boundary is moving and its concentration gradient profile may deviate from a Gaussian curve introduces factors that must be corrected. This is the reason for the superiority of the static over the dynamic method. At low speed the effect of S_g on the shape of the diffusing boundary may be neglected, but at higher speed [31], [97], [106]

$$D' = D'_{\mathrm{sed}} \left(1 - \omega^2 S_g t\right) \tag{164}$$

where D'_{sed} denotes the sedimentation coefficient derived from experimental ultracentrifugal data. Allowance for a variation of S_g with concentration must be made, and an advanced version for determination of D' has been suggested [107], [108]. The main reason for presenting the methods of evaluating D' in an ultracentrifuge is the desirability of obtaining S_g and D' from the same equipment. Otherwise the standard static methods for determination of D' are preferable.

Determination of Molecular Mass. Ultracentrifugation techniques provide the means for evaluating the molecular masses of materials which range from as low as 300 up to 50×10^6 or higher. The following three methods and their attendant equations can be used to determine molecular mass [31], [97]:

1) Sedimentation – diffusion:

$$M_r = \Omega S_{g0}/D, \quad S_{g0} = S_g\,(C \to 0) \tag{165}$$

2) Equilibrium:

$$M_r = 2\left(\Omega/\omega^2\right) \mathrm{d}\ln C/\mathrm{d}r^2 \tag{166}$$

3) Approach to equilibrium:

$$M_r = \Omega\,(\mathrm{d}C/\mathrm{d}r)_m \,/\, \left(\omega^2 C_m r_m\right) \tag{167}$$

where

$$\Omega = RT\,/\,(1 - v\,\varrho_\varphi) \tag{168}$$

Equations (165) – (167) are valid at infinite dilution. Note that in Equation (165) – the Svedberg equation – D (not D') is given by Equation (67), hence the requirement of infinite dilution. Furthermore, at infinite dilution $\varrho_\varphi \to \varrho_f$, where ϱ_f is the density of the carrier fluid which may be a solution containing dissolved species other than the one being tested. In the following, an extended Svedberg equation with regard to the hydrodynamic effect of concentration is derived by using the microscopic particle drift model [31]. The drift of molecules in centrifugal fields is assumed to involve full transmission of stresses from the sedimenting molecules to the carrier fluid. Under these conditions $F + F_D = 0$ prevails along the sedimentation path (see Section 2.3.4). Hence, using Equations (42) and (46) for monodisperse molecules gives

$$\varrho_p V_p\,(1-\varphi)\,(1-\varrho_f/\varrho_p)\,\omega^2 r = 3\pi\, d\mu'_\varphi U_\varphi \tag{169}$$

where μ'_φ is given by Equation (46). Multiplying both sides of Equation (169) by $\tilde{N}$; substituting the molecular mass M_r for $\tilde{N}\varrho_p V_p$, $S_g/(1-\varphi)$ (Equation 140) for $U_\varphi/(\omega^2 r)$, RT/D for $3\pi\tilde{N}d\mu_f$, v for $1/\varrho_p$, and $C\tilde{N}V_p$ for φ; and rearranging give

$$M_r = \frac{RT\,S_g}{D\,(1 - v\,\varrho_f)} \cdot \frac{1}{\left(1 - C\tilde{N}V_p\right) f_D\,(C)} = \frac{\Omega\,(S_g/D)}{\left(1 - C\tilde{N}V_p\right) f_D\,(C)} \tag{170}$$

where $f_{\mathrm{D}}(C)$ is given by Equation (68) and S_{g} is the sedimentation coefficient at concentration C.

Combining Equations (66) and (170) gives

$$M_{\mathrm{r}} = \frac{RT\,S_{\mathrm{g}}}{D'\,(1 - v\varrho_{\mathrm{f}})} \cdot \frac{\left(1 + \sum_i U_i'/k_{\mathrm{B}}T + \nabla \ln f_{\mathrm{a}}/\nabla \ln C\right)}{\left(1 - C\tilde{N}V_{\mathrm{p}}\right)} \tag{171}$$

Note that D' is the diffusion coefficient at concentration C. For dilute solutions of the sedimenting molecules, $(1 - C\tilde{N}V_{\mathrm{p}})f_{\mathrm{D}}(C)$ can be approximated by Equation (10) or (11). Equation (170) reduces to the Svedberg equation for $c \to 0$. Use of the diffusion coefficient as in Equation (171) calls for further considerations. If $C \to 0$, then D' is equal to D only if $\sum_i U_i'/k_{\mathrm{B}}T = 0$. For example, electromagnetic and interfacial energies may be associated with large sedimenting molecules or colloids, in which case the above condition does not hold. Consider now the Svedberg equation for polydisperse molecules or colloidal mixtures. In this case, combining Equations (42) and (46) for molecules pertaining to the ith fraction and using

$$U_{\varphi i} = U_{\mathrm{d}i} + U_{\mathrm{f}} = U_{\mathrm{d}i} + \sum_{j=1}^{n} U_{\mathrm{d}j}\,\varphi_j/(1 - \varphi)$$

give

$$V_{\mathrm{p}}\,(\varrho_{\mathrm{p}i} - \varrho_{\varphi})\,\omega^2 r = 3\pi d_i\,\mu_{\varphi}' \left[U_{\mathrm{d}i} + \sum_{j=1}^{n} U_{\mathrm{d}j}\,\varphi_j/(1 - \varphi)\right] \tag{172}$$

$$\varrho_{\varphi} = \sum_{j=1}^{n} \varrho_{\mathrm{p}j}\,\varphi_j + \varrho_{\mathrm{f}}\,(1 - \varphi) \tag{173}$$

By using the same procedure as above and recalling that

$$S_{\mathrm{g}i} = U_{\mathrm{d}i}/\omega^2 r,$$

$$M_{\mathrm{r}i} = \frac{RT\left[S_{\mathrm{g}i} + \sum_{j=1}^{n} S_{\mathrm{g}j}\,\varphi_j/(1 - \varphi)\right]}{D\,(1 - \mathrm{v}_i\,\varrho_{\varphi})} \cdot \frac{(1 - \varphi)}{f_{\mathrm{D}}(C)} \tag{174}$$

where

$$\varphi_j = \tilde{N}\,C_j V_{\mathrm{p}j} \quad \text{and} \quad \varphi = \sum_{j=1}^{n} \varphi_j$$

Note that Equations (170) and (174) are extensions of the Svedberg equation accounting only for the hydrodynamic effect of concentration.

If intermolecular interactions other than hydrodynamic exist, then further corrections are needed. For determination of molecular mass, see [31] and [97].

3.5.2. Magnetic and Magnetocentrifugal Sedimentation

Magnetic forces can be made comparable to diffusion and centrifugal forces that act on particles down to the colloidal size range [29], [109]. Magnetic forces that act on polarizable particles can arise due to magnetic field gradients, as well as to the formation of concentration gradients in the dispersion of such particles (which are placed in an externally uniform field). Nonuniform fields generate forces that directly affect the sedimentation coefficient in the same way the centrifugal field does. This force per particle can be expressed by

$$F_{\mathrm{H}} = \mu_0 V_{\mathrm{p}} \cdot \sum_{i=1}^{3} M_i\,\nabla H_{\mathrm{e}i} \tag{175}$$

where i denotes the coordinate system used (for example, $i = 1, 2, 3$ can represent the cartesian x, y, z system) and subscript e means that the field H is external. High g' fields (see Eq. 6) can be generated by using high-intensity, high-gradient magnetic fields (→ Magnetic Separation, Chap. 3.2.1.). The formation of nonuniform concentration profiles in an externally uniform (or nonuniform) magnetic field generates a magnetic diffusion effect that can be expressed in terms of the energy U_i' (see Eq. 66). In externally uniform magnetic fields [29],

$$U_{\mathrm{M1}}' = \mu_0 V_{\mathrm{p}}\,H \cdot M \tag{176}$$

$$U_{\mathrm{M2}}' = \mu_0 V_{\mathrm{p}} C \cdot \sum_{i=1}^{3} H_i\,\partial M_i/\partial C \tag{177}$$

where subscript M denotes that magnetization via concentration is the thermodynamic unconstrained variable, and U_{M1}' and U_{M2}' are the

magnetic energies of particles and those due to particle-particle interaction, respectively.

In an externally nonuniform field,

$$U'_{\mathrm{H1}} = \mu_0 v V_{\mathrm{p}} C M^2 \tag{178}$$

$$U'_{\mathrm{H2}} = \mu_0 v V_{\mathrm{p}} C^2 \cdot \sum_{i=1}^{3} M_i \, \partial M_i / \partial C \tag{179}$$

where subscript H denotes that the field strength is the thermodynamic unconstrained variable. In many cases, $U'_{\mathrm{M2}} \ll U'_{\mathrm{M1}}$ and $U'_{\mathrm{H2}} \ll U'_{\mathrm{H1}}$ can be assumed. The significance of U'_{H1} (being directly proportional to CM^2) increases with concentration and particle magnetization, i.e., for highly permeable particles in concentrated dispersions; U'_{M1} becomes dominant at high field strengths. The value of U'_{M1} can be made comparable to $k_{\mathrm{B}}T$ for a wide range of particle magnetization and size. For example, at 298 K, 10-nm paramagnetic particles having susceptibility as low as $\chi_{\mathrm{p}} = 10^{-6}$ require fields of 30 T to this end, whereas only 1 T is sufficient if particle size is 0.1 µm. If particles are made of ferri- or ferromagnetic material, then the field can be substantially reduced. Thus for a 1-nm magnetite particle, the above condition is fulfilled at a field intensity of 2 T, the magnetite magnetization being 0.56 T. Note that in ferrofluids, particle size is generally an order of magnitude larger. Hence for ferromagnetic particles in the colloidal range and larger, U'_{M} and U'_{H} can be made to exceed $k_{\mathrm{B}}T$, thus providing the means for magnetically controlled diffusion. For example, imposing a sufficiently high magnetic field on a mixture of colloids having differential magnetizations and subjecting it to a centrifugal field can be used for subsequent separation. This subject is yet to be developed.

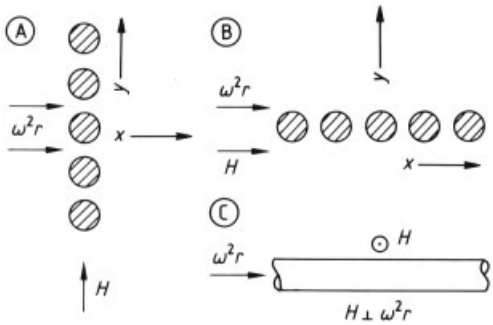

Figure 66. Three alternative magnetocentrifugal separation systems using arrays of ferromagnetic wires set at different orientations towards the centrifugal field [109]

Magnetocentrifugal separation (*MCS*) incorporates the combined action of centrifugal and nonuniform magnetic fields [109]. Particles undergo sedimentation in a centrifugal field and then are subjected to a high-gradient magnetic field along the sedimentation path. High gradients can be produced by placing a ferromagnetic matrix in an external field. Such a matrix can be an array of ferromagnetic wires, spheres, etc. Three alternative MCS systems using a onedimensional array of wires with different orientations toward the centrifugal field are shown in Figure 66. Note that the high-gradient zones are generated in the vicinity of the wires and that typical wire diameters range between 10 and 1000 µm depending on particle size. Figure 67 shows schematic trajectories of two particles of different sizes and the same susceptibility. In Figure 67 A, $\omega = 0$ and $u_{\mathrm{f}} > 0$; in Figure 67 B, $\omega > 0$ and $u_{\mathrm{f}} = 0$. In the first case, which is typical of high-gradient magnetic separation (HGMS), particle trajectory depends on particle size, whereas in the second case (MCS) it does not. Note that the wire is perpendicular to the plane of the drawing. Thus in MCS, if particle concentration is low, trajectories depend solely on magnetization. In magnetocentrifugal separation, particles of higher magnetization can be captured and separated from those that flow past the magnetic barrier. Figure 68 is a schematic of a MCS cell. The "magnetic" fraction is filtered out from the "nonmagnetic" fraction that flows on toward the cell bottom. The MCS technique can also be used for density separation in magnetizable fluids [109].

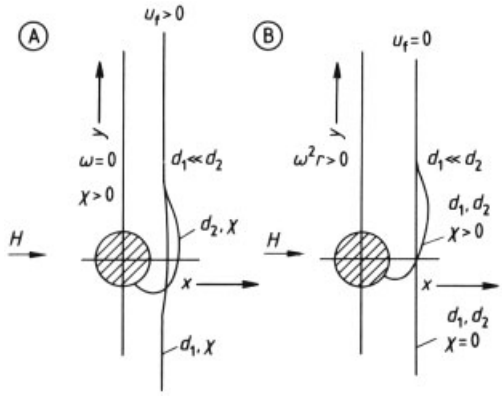

Figure 67. Trajectories of two particles of different sizes and the same susceptibility [109]
A) HGMS system: $U_{\mathrm{f}} > 0$, $\omega^2 r = 0$; B) MCS system: $u_{\mathrm{f}} = 0$, $\omega^2 r > 0$

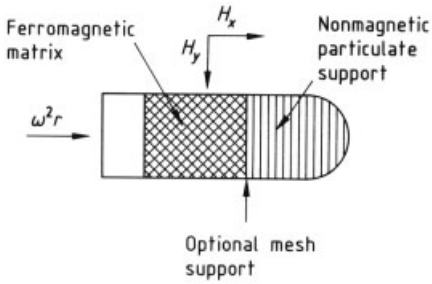

Figure 68. MCS separation cell [109]
Ferromagnetic matrix is on top of a nonmagnetic support

3.5.3. Centrifugal Sedimentation in Flowing Dispersions

Trajectory, Grade Efficiency, and Cut Size. In Section 3.5.1 and 3.5.2, ultracentrifugation and magnetocentrifugal separation involving sedimentation of particles in a constrained fluid, are treated. An exception is the elutriator rotor for separation of whole cells. This section deals with cases involving the combined effect of fluid flow and centrifugal field on particle sedimentation. Figure 69 shows schematically a simple tubular centrifuge. Feed is introduced axially, and clarified fluid overflows the exit port. Although the flow of the dispersion is nearly steady in the bowl frame of reference, for an observer in the dispersion frame of reference particles undergo nonsteady sedimentation. Description of the sedimentation of a polydisperse mixture consisting of n fractions requires the simultaneous solution of a set of n pairs of the following equations:

$$\mathrm{d}r_i/\mathrm{d}t = U_{\mathrm{d}i}(r), \quad \mathrm{d}z/\mathrm{d}t = v(r_i, z), \quad i = 1, 2, \ldots, n \tag{180}$$

where v denotes the axial fluid velocity. The use of the computer model presented in Section 2.4 is another alternative for a numerical solution. Such detailed solutions are, however, outside the scope of this chapter. Instead a set of simplifying assumptions that facilitate acceptable approximation of centrifuge performance is applied. Monodisperse particles, fine enough so that creeping flow sedimentation prevails, are assumed, and the theoretical background reviewed in [110], [111] is followed. With $g' = \omega^2 r$ and U_φ given by Equation (1), the radial sedimentation velocity is given by

$$U_{\mathrm{d}}(r) = \mathrm{d}r/\mathrm{d}t = U_\varphi(1 - \varphi) \tag{181}$$

If L/r_3 is large enough, dispersion entry and fluid discharge effects on axial flow can be neglected. In this case, the axial flow velocity is solely a function of r, which is given by [112]

$$\mathrm{d}z/\mathrm{d}t = \frac{Q}{\pi\left(r_3^2 - r_1^2\right)} K_1 \left(r_1^2 \ln\frac{r}{2} + \frac{r_3^2 - r_1^2}{2}\right) \tag{182}$$

where K_1 is a bowl constant

$$K_1 = \frac{r_3^2 - r_1^2}{\frac{3}{4} r_1^2 + \frac{1}{4} r_3^4 - r_1^2 r_3^3 - r_1^4 \ln(r_1/r_3)}$$

and Q is the volumetric flow rate of the liquid. Particle trajectories can be determined by dividing Equation (181) by Equation (182) followed by integration. Considerable simplification is achieved by using an axial plug flow model, i.e.,

$$\mathrm{d}z/\mathrm{d}t = \frac{Q}{\pi\left(r_3^2 - r_1^2\right)} \tag{183}$$

Substituting Equation (1), with $g' = \omega^2 r$, in Equation (181) and integrating the ratio of Equations (181) and (183) between the limits $r = r$ at $z = 0$ and $r = r_3$ at $z = L$ give

$$\ln(r_3/r) = \frac{K_2 d^2 L \pi \left(r_3^2 - r_1^2\right)}{Q} \tag{184}$$

$$K_2 = \frac{(\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}})\,\omega^2}{18\,\mu_{\mathrm{f}}} (1 - \varphi)\, f(\varphi) \tag{185}$$

The limit of separation is

$$d_{\max} = \left[\frac{Q \ln(r_3/r_1)}{K_2 \pi L \left(r_3^2 - r_1^2\right)}\right]^{1/2} \tag{186}$$

where all particles $d < d_{\max}$ are expected to escape sedimentation if they start at $r = r_1$, $z = 0$. For a given size d, all particles in the annulus between r and r_3 will reach the bowl wall, where r is determined from Equation (184).

$$r = r_3 \exp\left[-\pi K_2 L \left(r_3^2 - r_1^2\right) d^2/Q\right] \tag{187}$$

For a given size fraction, the ratio of particles deposited on the bowl wall, (i.e., between $z = 0$ and $z = L$) to the feed is defined as the grade efficiency $G(d)$ of this size (For a definition of grade efficiency, → Particle Size Analysis

and Characterization of a Classification Process, Chap. 3.1.4.). If the dispersion feed is uniform, $G(d)$ is given by

$$G(d) = (r_3^2 - r^2) / (r_3^2 - r_1^2) = \frac{r_3^2}{r_3^2 - r_1^2} \cdot \left[1 - \exp\left(-2K_2K_3d^2\right)\right] \quad (188)$$

where r is given by Equation (187) and K_3 is a constant equal to the residence time of the slurry in the bowl because

$$K_3 = \frac{\pi\left(r_3^2 - r_1^2\right)L}{Q} = \frac{V_b}{Q} \quad (189)$$

in which V_b denotes the volume of dispersion in the bowl. The equiprobable cut size d_{50} is defined as the size that has an equal probability of escape or deposition on the bowl wall, i.e., $d_{50} = d\,[G(d) = 0.5]$. Hence, when $G(d) = 0.5$, Equation (188) gives

$$d_{50}^2 = \frac{1}{2K_2K_3} \ln \frac{2r_3^2}{r_1^2 + r_3^2} \quad (190)$$

The value of d_{50} is often used as a criterion related to the efficiency of separation (→ Particle Size Analysis and Characterization of a Classificatin Process, Chap. 3.1.4.). The effect of particle concentration on d_{max}, r, $G(d)$, and d_{50} occurs via the constant K_2 (Eq. 185). Increase of φ increases d_{max}, r, and d_{50}, and decreases $G(d)$.

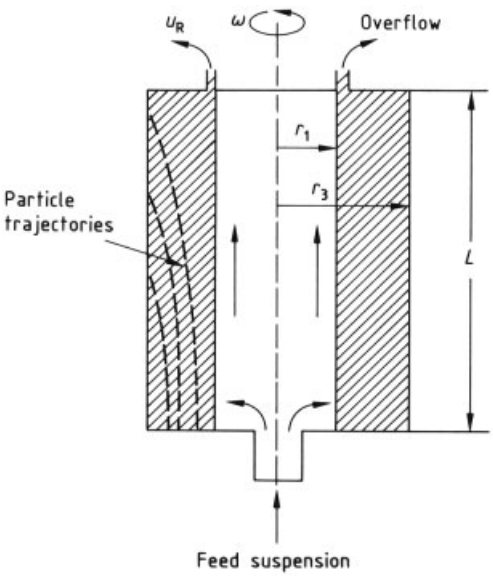

Figure 69. Schematic diagram of a simple tubular centrifuge [110]

The Sigma Concept [111]. A simplified relation of efficiency of the centrifuge in terms of d_{50}, total volumetric flow rate Q, and index of machine size Σ has been developed [113]. This relation involves the sigma concept for which a free sedimentation reference velocity (at $g' = g$) of d_{50} particles is defined by

$$U_g = \frac{(\varrho_p - \varrho_f)\, g\, d_{50}^2}{18\mu_f} = d_{50}^2 K_2 \left(g/\omega^2\right) / [(1 - \varphi)\, f(\varphi)] \quad (191)$$

Combining Equations (189) – (191) gives

$$Q = 2U_g\Sigma \quad (192)$$

$$\Sigma = \left(\frac{\omega_1^2}{g}\right) \pi L \left(r_3^2 - r_1^2\right) / \ln \frac{2r_3^2}{r_1^2 + r_3^2} \quad (193)$$

where ω_1 is defined as an effective angular speed

$$\omega_1^2 = \omega^2 (1 - \varphi)\, f(\varphi) \quad (194)$$

In dilute dispersions, $\omega_1 \to \omega$ and Σ is a constant of the centrifuge. However as Equations (193) and (194) show, it is also generally a function of slurry concentration. Thus Equation (192) facilitates scaleup (in the form of Eq. 195) of a given type of centrifuge as well as "scaleup" in the sense of increasing concentration.

$$Q_1/\Sigma_1 = Q_2/\Sigma_2, \text{ for a constant } d_{50} \quad (195)$$

The parameter Σ has the dimension of area and can be viewed as representing half the area of an ideal settling pool that is required to handle gravitational sedimentation of the same particles. This parameter relates to the grade but not to the throughput of particles, and to this end the following extension is proposed. The particle flux Q_p is related to the fluid flux by $Q_p = \varphi Q$; hence, by using Equation (192),

$$Q_p = 2U_g\Sigma_\varphi$$

$$\Sigma_\varphi = \left(\frac{\omega_\varphi^2}{g}\right) \pi L \left(r_3^2 - r_1^2\right) / \ln \frac{2r_3^2}{r_1^2 + r_3^2} \quad (196)$$

$$\omega_\varphi^2 = \omega^2 \varphi (1 - \varphi)\, f(\varphi) = \varphi\omega_1^2 \quad (197)$$

Equations (196) and (197) show that from the point of view of flux, $\varphi \to 0$ gives a vanishing effective angular speed; hence given ω, an infinitely large sedimentation area [i.e., $(r_3^2 - r_1^2) \to \infty$] must be provided if Σ_φ is to remain finite. On the other extreme, high values of φ call for increased ω so that the d_{50} specifications can be met, but this incurs increased operating and possibly investment costs. Hence the optimum is expected at intermediate values of φ. The estimation of centrifuge capacity, by using classical analysis of gravitational thickening and sedimentation flux curves, has been described [114].

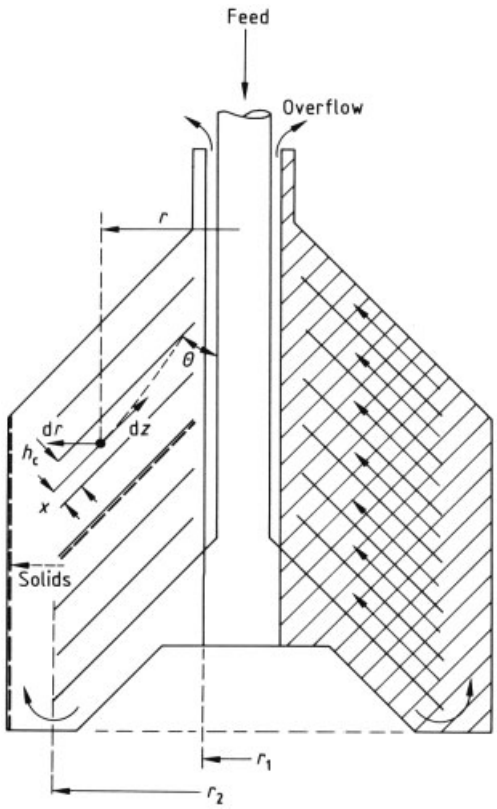

Figure 70. Structure of sedimentation and flow zones in a disk centrifuge, schematic [110]

Centrifugal Sedimentation in Rotating Inclined Channels [111]. The enhanced sedimentation of particles in inclined channels in a constant force field is described in Section 2.8. In this section, the position-dependent centrifugal field is considered. The operation of disk centrifuges involves sedimentation in inclined channels, or chambers, that are formed between disks. The inclined disks are mounted on a rotating shaft, as shown in Figure 70. An exact analysis of flow and sedimentation in each channel requires data about the flow profile of the fluid and the solution of nonsteady sedimentation equations. Instead, a simplified plug flow model of flow in the channels is assumed. The flow velocity along the channel, in the z direction (see Fig. 70), is related to the total flow rate Q by

$$dz/\mathrm{d}t = (Q/n) \,/\, (2\pi r h_c) \tag{198}$$

where h_c is the spacing between adjacent disks. The sedimentation velocity toward the upper disk downward facing wall is related to the radial velocity by

$$\mathrm{d}x/\mathrm{d}t = (\mathrm{d}r/\mathrm{d}t)\cos\theta \tag{199}$$

where θ is the angle between the disk and the shaft. Combining Equations (181), (198), and (199) gives

$$\mathrm{d}x/\mathrm{d}z = K_2 d^2 r^2 \frac{2\pi n h_c}{Q}\cos\theta \tag{200}$$

$$\mathrm{d}x/\mathrm{d}r = -K_2 d^2 r^2 \frac{2\pi n h_c}{Q}\cot\theta \tag{201}$$

In deriving Equation (201), use was made of $\mathrm{d}z = -\mathrm{d}r/\sin\theta$. Integration of Equation (201) between $x = x_1$ at $r = r_2$ and $x = h_c$ at $r = r_1$ gives

$$h_c - x_1 = K_2 d^2 \frac{2\pi n h_c}{Q}\left(r_2^3 - r_1^3\right)\cot\theta \tag{202}$$

Equation (202) determines the limiting size given the entry point $x = x_1$ and $r = r_2$ – or, alternatively, the limiting entry point given the size d – for which the sedimentation trajectory still crosses the upper wall of the channel (i.e., at $x = h_c$ and $r = r_1$). The size limit $d = d_{\max}$ of separation is readily obtained by letting $x = x_1 = 0$, $r = r_2$ at the entry point. Under specific operating conditions, $d > d_{\max}$ particles would not report to the overflow. The grade efficiency of particles of size d is given by

$$G(d) = \frac{h_c - x_1}{h_c} \tag{203}$$

where x_1 satisfies Equation (202). Because $d_{\max}$ is obtained by substitution of $x_1 = 0$ in Equation (202),

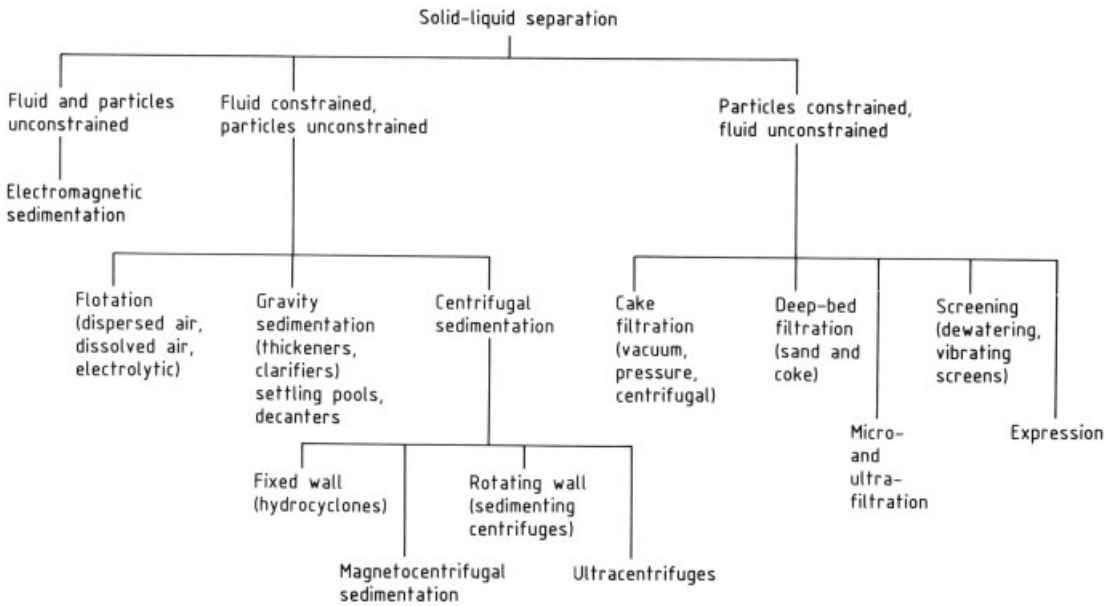

Figure 71. Classification of solid – liquid separation techniques

$$(h_c - x_1)/h_c = d^2/d_{max}^2$$

hence,

$$G(d) = d^2/d_{max}^2 \text{ for } d \leq d_{max} \tag{204}$$

and

$$G(d) = 1 \quad \text{for } d > d_{max}$$

Letting $G(d) = 0.5$ and using Equation (204) give

$$d_{50} = d_{max}/\sqrt{2} \tag{205}$$

The Σ factor for a disk centrifuge is obtained by combining Equations (191), (192), (202) for $x_1 = 0$ and $d = d_{max}$, and (205).

$$\Sigma = \frac{2}{3} \cdot \frac{\omega_1^2}{g} \pi n \left(r_2^3 - r_1^3\right) \cot\theta \tag{206}$$

Replacing ω_1 (Eq. 194) by ω_φ (Eq. 197) in Equation (205) gives Σ_φ instead of Σ.

Centrifugal Sedimentation in Rotating Fluid Streams. Centrifugal sedimentation in rotating fluid streams is well known and is used for solid – liquid (→ Centrifugation, Sedimenting), solid – solid (→ Air Classifying, → Elutriation), liquid – gas, and solid – gas (→ Dust Separation) separation. The fluid stream is forced to rotate in the separation zone as it flows through, thus producing centrifugal forces due to its own flow. Such flow ordinarily occurs through devices having fixed curved walls. This technique is used in various dry and wet separators, classifiers, and elutriators (→ Air Classifying; → Elutriation; → Centrifugation, Sedimenting). Analysis of sedimentation in rotating streams requires data from the flow field and the solution of n simultaneous nonsteady sedimentation equations, where n denotes the number of sedimentating particle fractions. Computer models of the type described in Section 2.3 can also be used to this end. The present state of the art provides mainly empirical correlations for prediction of the performance of such devices with respect to the products of feed containing relatively concentrated polydisperse mixtures.

3.5.4. Solid – Liquid Separation by Sedimentation

Sedimentation methods are versatile in solid – liquid separation processes. Solid – liquid separation often involves sedimentation, filtration (→ Filtration), or both. Large-scale sed-

imentation operations have had long-standing success in water treatment (→ Biochemical Engineering) and reclamation. Different literature sources provide outlines for the selection of equipment for solid – liquid separation. Sedimentation and filtration are described under the heading of mechanical separation, along with practical detailed information about related equipment, in [115]. A systematic approach to the solution of solid – liquid separation problems is contained in [116]. A comparison of sedimentation and filtration for solid – liquid separation and time scales for sedimentation in different groups of equipment are provided in [117]. These time scales, given for different sludge concentrations, provide initial elimination and selection options for sedimentation equipment, which must then be processed further by experts in the field. Figure 71 shows an extended version of the diagram provided by SVAROVSKY [111] in which the solid – liquid separation is classified according to whether or not the fluid and particles are constrained. Centrifugal sedimentation is introduced in Sections 3.5.1, 3.5.2, 3.5.3, and thickeners and clarifiers are considered under → Wastewater. Flotation (→ Flotation) can be viewed as reverse sedimentation or "buoyant sedimentation," in which particle-laden bubbles float or undergo negative sedimentation. In electromagnetic sedimentation, fluid and particles are generally unconstrained, with magnetocentrifugal sedimentation being an exception. Sedimentation techniques in many ways complement filtration and vice versa; in solid – liquid separation processes, sedimentation usually precedes filtration as the first step. This is due to the fact that often none of these techniques provides a completely viable economic solution if used as the sole separation step.

4. References

1. Y. Zimmels: "Accelerated and Steady Particle Flows in Newtonian Fluids," *Encyclopedia of Fluid Mechanics,* **vol. 5, Chap. 3,** Gulf Publishing Co., Houston 1986, pp. 93 – 153.
2. R. H. Davis, A. Acrivos, *Ann. Rev. Fluid Mech.* **17** (1985) 91 – 118.
3. E. Barnea, J. Mizrahi, *Chem. Eng. J. (Lausanne)* **5** (1973) 171 – 189.
4. P. U. Foscolo, G. L. Gibilaro, S. P. Waldram, *Chem. Eng. Sci.* **38** (1983) no. 8, 1251 – 1260.
5. J. F. Richardson, W. N. Zaki, *Trans. Inst. Chem. Eng.* **32** (1954) 35 – 53.
6. J. Garside, M. R. Al-Dibouni, *Ind. Eng. Chem. Process Des. Dev.* **16** (1977) 206 – 214.
7. J. M. Coulson, J. F. Richardson: *Chemical Engineering,* 3rd ed., vol. 2, Pergamon Press, Oxford 1978.
8. G. K. Batchelor, *J. Fluid Mech.* **52** (1972) 245 – 268.
9. A. S. Sangani, A. Acrivos, *Int. J. Multiphase Flow* **8** (1982) 343 – 360.
10. Y. Zimmels, *Powder Technol.* **34** (1983) 191 – 202.
11. Y. Zimmels, *AIChE J.* **29** (1983) no. 4, 669 – 676.
12. Y. Zimmels, *J. Appl. Phys.* **56** (1984) no. 7, 2133 – 2141.
13. Y. Zimmels, *Powder Tech.* **56** (1988) no. 4, 227 – 250.
14. J. F. Davidson, D. Harrison: *Fluidization,* Academic Press, New York 1971.
15. D. Kuni, O. Levenspiel: *Fluidization Engineering,* J. Wiley & Sons, New York 1969.
16. S. Mirza, J. R. Richardson, *Chem. Eng. Sci.* **34** (1979) 447 – 454.
17. E. G. Kelly, D. J. Spotiswood: *Introduction to Mineral processing,* J. Wiley & Sons, New York 1982.
18. H. Lamb: *Hydrodynamics,* Dover, New York 1945.
19. L. M. Milne Thompson: *Theoretical Hydrodynamics,* 5th ed. Macmillan Publ. Comp, London 1974.
20. V. L. Streeter: *Fluid Dynamics,* McGraw-Hill, New York 1948.
21. Chia-Shun Yih: *Fluid Mechanics,* McGraw-Hill, New York 1969.
22. G. B. Wallis: *One Dimensional two phase flow,* McGraw-Hill, New York 1969.
23. A. B. Basset: *Hydrodynamics,* vol. 2, Dover, New York 1961.
24. G. Rudinger: "Fundamentals of Gas Particle Flow," *Handbook of Powder Technology,* Elsevier, Amsterdam 1980.
25. A. T. Hjelmfelt, L. F. Mockros, *Appl. Sci. Res.* **16** (1966) 149 – 161.
26. T. Allen: *Particle Size Measurements,* 3rd ed., Chapman and Hall, London 1981.
27. B. H. Kaye: *Direct Characterization of Fine Particles,* J. Wiley & Sons, New York 1981.
28. E. L. Cussler: *Diffusion Mass Transfer in Fluid Systems,* Cambridge University Press, Cambridge 1984.

29. Y. Zimmels: "Diffusion of Polarizable Colloids in Electromagnetic Fields," *J. Colloid Interface Sci.,* in press.
30. R. B. Bird: *Theory of Diffusion, Advances in Chemical Engineering,* **vol. 1,** Academic Press, New York 1956, pp. 155 – 239.
31. H. K. Schachman: *Ultracentrifugation in Biochemistry,* Academic Press, New York 1959.
32. P. L. Meyer: *Introductory probability and Statistical Applications,* 2nd ed., Addison-Wesley, Reading, Mass. 1970.
33. E. Barnea, J. Mizrahi, *Can. J. Chem. Eng.* **53** (1975) 461 – 467.
34. E. Barnea, J. Mizrahi, *Ind. Eng. Chem. Fundam.* **15** (1976) 120 – 125.
35. V. G. Levich: *Physicochemical Hydrodynamics,* Prentice Hall, Englewood Cliffs, N.J. 1962.
36. J. Newman, *J. Chem. Eng. Sci.* **22** (1967) 83.
37. G. I. Taylor, *Proc. R. Soc. London Ser. A* **138** (1932) 41.
38. A. Leviton, A. Leighton, *J. Phys. Chem.* **40** (1936) 71.
39. P. U. Foscolo, L. G. Gibilaro, S. P. Waldram, *Chem. Eng. Sci.* **38** (1983) no. 8, 1251 – 1260.
40. Y. Zimmels: "A Generalized Approach to Flow through Fixed Beds Fluidization and Hindered Sedimentation," *Chem. Eng. Commun.,* **67** (1988) 19 – 42.
41. E. Barnea, R. L. Mednick, *Chem. Eng. J. (Lausanne)* **15** (1978) 215 – 227.
42. J. M. Dallavalle: *Micromeritics,* 2nd ed., Pitman, (London) 1948.
43. N. L. Weiss (ed.): *SME Mineral Processing Handbook,* Society of Mining Engineers, American Institute of Mining, Metallurgical and Petroleum Engineers Inc., New York 1985.
44. R. Gerber, R. Briss: *High Gradient Magnetic Separation,* Research Studies Press, Chichester 1983.
45. A. E. Boycott, *Nature (London)* **104** (1920) 532.
46. W. D. Hill: *Boundary-Enhanced Sedimentation due to Setting Convection,* Ph. D. Thesis, Carengie-Mellon Univ. Pittsburgh, PA 1974.
47. R. H. Perry, C. H. Chilton (eds.): *Chemical Engineers Handbook,* 6th ed., McGraw-Hill, New York, 1984.
48. L. Svarvosky (ed.): *Solid-Liquid Separation,* Chap. 7, Butterworths, London 1977.
49. E. Ponder, *Q. J. Exp. Physiol.* **15** (1925) 235 – 252.
50. N. Nakamura, K. Kuroda, *Keijo J. Med.* **8** (1937) 256 – 296.
51. W. D. Hill, R. R. Rothfus, K. Li, *Int. J. Multiphase Flow* **3** (1977) 561 – 583.
52. I. Robinstein, *Int. J. Multiphase Flow* **6** (1980) 473 – 490.
53. R. F. Probstein, R. Yung, R. Hicks in M. P. Freeman, J. A. Fitzpatrick (eds.): "A model of lamella settlers," *Physical Separation,* Engineering Foundation, New York, 1981, pp. 53 – 92.
54. W. F. Leung, R. F. Probstein, *Ind. Eng. Chem. Process Des. Dev.* **22** (1983) 58 – 67.
55. A. Acrivos, E. Herbolzheimer, *J. Fluid Mech.* **92** (1979) 435 – 457.
56. E. Herbolzheimer, A. Acrivos, *J. Fluid Mech.* **108** (1981) 485 – 499.
57. E. Herbolzheimer, *Phys. Fluids* **26** (1983) 2043 – 2054.
58. R. H. Davis, E. Herbolzheimer, A. Acrivos, *Phys. Fluids* **26** (1983) 2055 – 2064.
59. W. F. Leung, *Ind. Eng. Chem. Process Des. Dev.* **22** (1983) 68 – 73.
60. R. H. Davis, E. Herbolzheimer, A. Acrivos, *Int. J. Multiphase Flow* **8** (1982) no. 6, 571 – 585.
61. J. Weisman, L. Efferding, *AIChE J.* **6** (1960) 419 – 426.
62. L. E. Gates, J. R. Morton, P. L. Pondy, *Chem. Eng. (N.Y.)* **83** (1976) 144 – 150.
63. F. Aerstin, G. Street: *Applied Chemical Process Design,* Plenum Press, New York 1978.
64. R. L. Bates, P. L. Pondy, R. R. Corpstein, *Ind. Eng. Chem. Process Des. Dev.* **2** (1963) no. 4, 310 – 314.
65. E. J. Wasp, J. P. Kenny, R. L. Gandhi: *Solid – Liquid Flow Slurry Pipeline Transportation,* 1st ed.,1978, Trans. Tech. publications, Clausthal, Federal Republic of Germany, Gulf Publishing Co., Houston 1979.
66. M. P. O'Brien, *Trans. Am. Geophys. Union* **14** (1933) 487 – 491.
67. V. A. Vanoni, *Trans. ASCE* **111** (1946) 67 – 133.
68. H. M. Ismail, *Trans. ASCE* **117** (1952) 409 – 447.
69. E. J. Wasp et al.: "Cross Country Coal Pipeline Hydraulics," Pipe Line News (July 1963) pp. 20 –30.
70. E. J. Wasp et al.: "Deposition Velocities, Transition Velocities and Spatial Distribution of Solids in Slurry Pipelines," *1st. Int. Conf. on Hydr. Transport of Solids in Pipes,* BHRA Fluid Engn., Cranfield U.K., paper H4, Sept. 1970.
71. R. L. Whitmore, *Br. J. Appl. Phys.* **6** (1955) 239 – 245.

72. R. H. Weiland, R. R. McPherson, *Ind. Eng. Chem. Fundam.* **18** (1979) 45 – 49.
73. R. H. Weiland, Y. P. Fessas, B. V. Ramarao, *J. Fluid Mech.* **142** (1984) 383 – 389.
74. Y. P. Fessas, R. H. Weiland, *Int. J. Multiphase Flow.* **110** (1984).
75. W. Resnick, Y. Zimmels, D. Boadi: *Magnetic Structural Effects on Flow through Beds of Magnetizable Particles,* In MT10, Boston, MA, Oct. 1987.
76. R. E. Rosensweig: *Ferrohydrodynamics,* Cambridge University Press, Cambridge 1985.
77. D. G. Thomas, *AIChE J.* **8** (1962) 373 – 378.
78. M. E. Charles, G. S. Stevens: "The Pipeline Flow of Slurries-Transition Velocities," *2nd Int. Conf. on Hyd. Transport of Solids in Pipes,* BHRA Fluid Eng. Cranfield, U.K., paper E3, Sept. 1972.
79. T. C. Aude et al., *Chem. Eng.* (June 28, 1971) 74.
80. D. M. Newitt et al., *Trans. IChemE* **33** (1955) 93 – 102.
81. W. E. Wilson, *Trans. ASCE* **107** (1942) 1576 – 1594.
82. R. Durand: "Basic Relationships of the Transportation of Solids in Pipes," *Experimental Research, Proceedings,* International Hydraulics Convention, Minnesota 1953, pp. 89 – 103.
83. R. Durand, E. Condolios: "Hydraulic Transport of Coal and Solid Materials in Pipes," *Proceedings of Colloquium on Hydraulic Transport of Coal* (Nov. 5 – 6, 1952). National Coal Board London England 1953, pp. 39 – 52.
84. C. G. Sinclair: "The Limit-Deposit Velocity of Heterogeneous Suspensions," *Interactions between Fluid and Particles,* Inc. Chem. Eng. London 1962.
85. N. Yotsukura: *Some Effects of Bentonite Suspensions on Sand Transport in Smooth 4-inch Pipe,* Ph. D. Thesis, Colorado State University, 1961.
86. M. Wicks: "Transportation of Solids at low Concentration in Horizontal Pipes," *ASCE Int. Symposium on Solid – Liquid Flow in Pipes,* Univ. of Pennsylvania, March 1968.
87. R. E. McElvain, I. Cave: "Transportation of Tailings," *World Mining Tailings Symposium,* 1972.
88. M. S. Peters, K. D. Timmerhaus: *Plant Design and Economics for Chemical Engineers,* 3rd ed., McGraw-Hill, New York 1980, pp. 377 – 383.
89. H. Braur, E. Kriegel: "Hydraulic Conveying of Solids in Horizontal Pipes," *Bänder Bleche Rohre* **6** (1965) no. 6, 315 – 324; H. Braur, E. Kriegel: *Proc. Hydrotransp.* **2** (1972) E1 – 1.
90. A. R. Oroskar, R. M. Turian, *AIChE J.* **26** (1980) no. 4, 550 – 558.
91. G. A. Wani in N. P. Cheremisinoff (ed.): "Critical Velocity in Multisize Particle Transport Through Pipes," *Encyclopedia of Fluid Mechanics,* **Chap. 4, vol. 5,** Gulf Publ. Co., Houston 1986, pp. 155 – 187.
92. R. Darby in N. P. Cheremisinoff (ed.): "Hydrodynamics of Slurries and Suspensions," *Encyclopedia of Fluid Mechanics,* **Chap. 2, vol. 5,** Gulf Publ. Co., Houston 1985, pp. 49 – 91.
93. G. W. Govier, M. W. Charles, *Eng. J.* **44** (1951) 50.
94. D. G. Thomas, *AIChE J.* **10** (1964) 303 – 308.
95. E. J. Wasp et al.: "Hetero-Homogeneous Solid-Liquid Flow in the Turbulent Regime," *ASCE Int. Symposium on Solid Liquid Flow,* Univ. of Pennsylvania, March 1968.
96. R. W. Hanks in N. P. Cheremisinoff (ed.): "Principles of Slurry Pipeline Hydraulics," *Encyclopedia of Fluid Mechanics,* **Chap. 6, vol. 5,** Gulf. Publ. Co., Houston 1986, 213 – 276.
97. T. J. Bowen: *An Introduction to Ultracentrifugation,* Wiley-Interscience, London 1970.
98. R. J. Goldberg, *J. Phys. Chem.* **57** (1953) 194.
99. R. Trautman, V. N. Schumaker, *J. Chem. Phys.* **22** (1954) 551.
100. T. Svedberg, K. O. Pedersen: *The Ultracentrifuge,* Oxford Univ. Press, London, Johnson Reprint Corporation, 1940.
101. E. G. Pickels, W. F. Harrington, H. K. Schachman, *Proc. Natl. Acad. Sci. U.S.* **38** (1952) 943.
102. R. L. Baldwin, *Biochem. J.* **55** (1953) 644.
103. R. V. W ebber, *J. Am. Chem. Soc.* **78** (1956) 536.
104. R. Trautman, *J. Phys. Chem.* **60** (1956) 1211.
105. L. J. Gosting, *Adv. Protein Chem.* **11** (1956) 429.
106. O. Lamm, *Z. Phys. Chem. Ser. A* **143** (1929) 177.
107. H. Fujita, *J. Chem. Phys.* **24** (1956) 1084.
108. K. Kawahara, *Biochemistry* **8** (1969) 2551.
109. Y. Zimmels, *J. Appl. Phys.* **60** (1986) no. 6, 2162 – 2174.
110. L. Svarovsky in L. Svarovsky (ed.): "Separation by Centrifugal Separation," *Solid Liquid Separation,* Chemical Engineering

series, Butterworths, London 1977, pp. 125 – 147.
111. L. Svarovsky, in L. Ricci (ed.): "Advances in Solid – Liquid Separation II: Sedimentation Centrifugation and Flotation," *Separation Techniques 2: Gas/Liquid/Solid Systems,* Chemical Engineering, McGraw-Hill, New York 1980, pp. 19 – 31.
112. H. K. Schachman, *J. Colloid Phys. Chem.* **52** (1948) 1034 – 1045.
113. C. M. Ambler, *Chem. Eng. (N.Y.)* **48** (1952) 150 – 158.
114. P. A. Vesilind in L. Ricci (ed.): "Estimating Centrifuge Capacities," *Separation Techniques 2: Gas/Liquid/Solid Systems,* Chemical Engineering, McGraw-Hill, New York 1980, pp. 169 – 172.
115. *Solid/Liquid Separation Equipment Scale-Up,* edited by D. B. Purchas, Uplands press, 1977.
116. F. M. Tiller, "How to select Solid – Liquid Separation Equipment," in *Separation Techniques 2: Gas/Liquid/Solid Systems,* Chemical Engineering, McGraw-Hill, New York 1980, pp. 42 – 45.
117. H. G. W. Pierson in L. Svarovsky (ed.): "The Selection of Solid – Liquid Separation Equipment," *Solid – Liquid Separation,* Chemical Engineering Series, Butterworths, London 1977, pp. 313 – 326.

Dust Separation

EBERHARD SCHMIDT, Bergische Universität – GH Wuppertal, Wuppertal, Germany

FRIEDRICH LÖFFLER, Universität Karlsruhe (TH), Karlsruhe, Germany

1. Introduction 2
2. Evaluation of Dust Collection Equipment 3
3. Centrifugal Collectors (Cyclones) 4
3.1. General 4
3.2. Mode of Operation and Basic Designs 4
3.3. Design Calculations 5
3.3.1. Collection Efficiency 5
3.3.2. Pressure Drop 7
3.4. Operating Characteristics 7
4. Scrubbers 9
4.1. General 9
4.2. Mode of Operation and Basic Designs 9
4.3. Design Calculations 11
4.3.1. Collection Efficiency 11
4.3.2. Pressure Drop 14
4.4. Droplet Separators 15
5. Filters 16
5.1. General 16
5.2. Fibrous Filters (Deep-Bed Filters) 16
5.2.1. Mode of Operation and Basic Designs 16
5.2.2. Design Calculations 18
5.2.2.1. Collection Efficiency 18
5.2.2.2. Pressure Drop 20
5.3. Fabric Filters (Surface Filters) . . 21
5.3.1. Mode of Operation 21
5.3.2. Basic Designs 23
5.3.3. Operating Characteristics 25
5.3.4. Design Calculations 27
5.4. Granular-Bed Filters 27
5.4.1. General 27
5.4.2. Mode of Operation 28
5.4.3. Basic Designs 28
5.4.4. Design Calculations 29
5.4.4.1. Collection Efficiency 30
5.4.4.2. Pressure Drop 31
6. Electrical Precipitators 31
6.1. General 31
6.2. Mode of Operation 31
6.2.1. Principles 31
6.2.2. Charging of Particles 32
6.2.3. Effect of Dust Resistivity 33
6.3. Basic Designs 34
6.4. Design Calculations 35
7. References 36

Symbols

a plate spacing, tube radius
A surface area
b specific scrubbing liquor ratio
c particle concentration
c_{m} limiting dust load
C_{D} drag coefficient
d_{G} diameter of granules
d_{l} droplet diameter
D_{F} fiber diameter
e diameter of flow tube
E overall collection efficiency
f circular cross-sectional area
f factor
f' fitting parameter
F filtration area
h adhering fraction
H height
k_1, k_2 fitting parameters
$k_{\mathrm{a}}, k_{\mathrm{b}}$ fitting parameters
K_1 residual resistance of filter medium after cleaning
K_2 specific resistance of filter cake
L collection length
m specific cleaned volume
m particle mass
M filter mass per unit filter area
$\dot{M}$ mass flow rate
n empirical fitting factor
N number of cyclones
Δp pressure drop
P penetration
$q(x)$ particle-size distribution
r radius
R resultant of external forces
Re Reynolds number
t time
T absolute temperature

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_13

$T(x)$	fractional collection efficiency
U_0	gas velocity, superficial velocity
v	velocity
v	air to cloth ratio
v_{rel}	relative velocity between the droplet and the gas
$\dot{V}$	volume flow rate
w	particle velocity
$w(x)$	migration velocity
W	mass of dust deposited per unit filter area
x	particle size (diameter)
x_{cr}	critical particle diameter
x_{t}	cut size
Z	thickness
ε	porosity
η	collision efficiency
η	transport parameter
μ	dynamic viscosity
μ_{m}	maximum dust load
ξ	pressure-drop coefficient
ϱ_{g}	gas density
ϱ_{F}	fiber density
φ	collection efficiency of single fiber or granule
ψ	inertial (Stokes) parameter

Subscripts

c	collected
D	diffusion
e	entry
f	fine
F	fiber
g	gas
i	internal
l	scrubbing liquor
p	particle
r	radial
φ	tangential

1. Introduction

Whether desired or not, many industrial processes inevitably generate particles entrained by a carrier gas. These particles may be solid (dust) or liquid (droplets or mist). As a rule, the particles must be separated from the transporting gas, either because they are the desired product or because they are an undesirable or even harmful emission.

Product recovery by means of dust (gas – solid) separation has been practiced long and intensively in process engineering. The importance of emission control has grown, and standards have become more stringent. Rapid increases in industrial production and a better knowledge of potential hazards and harmful effects have led to more stringent regulations for the control of air pollution. One example is the emission limits established in "TA Luft 1986" [9], an administrative regulation issued by the Federal Republic of Germany, which tolerates a maximum emission of 50 mg/m^3 of inert dust only. In many cases, which are defined in these regulations, the specified emission value is much lower than this. For dust that is harmful to health, the limits are geared to the potential hazard and run from 5 mg/m^3 (e.g., antimony, lead), through 1 mg/m^3 (e.g., arsenic, nickel) and 0.2 mg/m^3 (e.g., cadmium, mercury), to 0.1 mg/m^3 (carcinogens). In future these will be supplemented by regulations concerning particle size classes.

Because dust concentration upstream of collection equipment is frequently in the range of 1 – 10 g/m^3, collection efficiencies greater than 99.99 % are often necessary to meet the above limits. Such values can be achieved only through the most painstaking selection and design and, in many cases, only by combining several types of equipment. One difficulty is that the size distribution of the particles to be collected generally extends over a wide range, from < 0.1 µm to > 100 µm; the harmful components (e.g., heavy metals), which demand especially efficient collection, often fall in the fine range below 1 µm.

Dust collection methods used in practice can be classified as

1) Inertial and centrifugal separation (gas cyclones)
2) Wet collection (scrubbers)
3) Filtration (fibrous filters, fabric filters, granular-bed separators, etc.)
4) Electrostatic collection (precipitators).

All of these collectors separate particles from gas by a common principle: forces acting on the particles cause them to enter regions from which the gas cannot transport them away. These regions may be the inner wall of a cyclone, the droplet surface in a scrubber, the fiber or grain surface in a filter, or the collecting electrode in a precipitator.

An elementary theory of gas–solid separation thus consists of calculating particle trajectories by solving the equation of motion:

$$m \cdot \mathrm{d}w/\mathrm{d}t = R \tag{1}$$

where m is the particle mass, w is the particle velocity, t is time, and R is the resultant of the external forces acting on the particle (such as gravitational, hydrodynamic and electrical forces).

As a rule, Equation (1) is solved numerically. A problem arises because the flow field of the gas must be entered into the equation and this field is often not known precisely. Furthermore, a random component of motion may also be involved, allowance for which can be difficult. In practice, therefore, approximations are used or a purely empirical approach is taken, in which design is based on experimental findings for a given type of equipment. The drawback of the empirical method is that the extent to which experience can be extrapolated to new situations is unclear.

2. Evaluation of Dust Collection Equipment

To evaluate dust collection equipment, both the energy consumption (or pressure drop Δp) and the collection efficiency must be determined. The latter can be found by integral techniques (for the totality of particles) or differential methods (for each size fraction).

A general plan for characterizing a separation can be found in DIN 66 142 [10]. Here, only the case is considered where the solid material entering along with the crude gas is assumed to be separated into two components. The nomenclature has been adapted to common usage in separation technology; important parameters are depicted and defined in Figure 1.

The *overall collection efficiency* (integral mass balance) E, also known as the total or cumulative collection efficiency, is given by

$$E = \frac{\dot{M}_\mathrm{c}}{\dot{M}_\mathrm{e}} \tag{2a}$$

or

$$E = 1 - \frac{\dot{M}_\mathrm{f}}{\dot{M}_\mathrm{e}} \tag{2b}$$

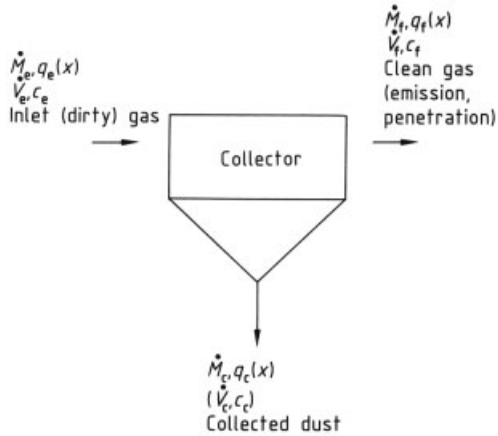

Figure 1. Notation used in dust collectors
$\dot{M}$ = mass flow rate of particles; $\dot{V}$ = volume flow rate of gas; c = particle concentration; $q(x)$ = particle-size distribution (density distribution)
Subscripts: c = collected (coarse); e = entry (inlet); f = fine (penetrating)

In view of the importance of emissions, especially for collection efficiencies close to 100 %, the *penetration* P is more informative:

$$P = 1 - E \tag{3}$$

or

$$P = \frac{\dot{M}_\mathrm{f}}{\dot{M}_\mathrm{e}} \tag{4}$$

As a rule, dust loads upstream and downstream of the collection equipment are measured experimentally and E is determined as follows:

$$E = \frac{\dot{V}_\mathrm{e} c_\mathrm{e} - \dot{V}_\mathrm{f} c_\mathrm{f}}{\dot{V}_\mathrm{e} c_\mathrm{e}} \tag{5}$$

with $\dot{V}_\mathrm{e} = \dot{V}_\mathrm{f}$, then

$$E = 1 - \frac{c_\mathrm{f}}{c_\mathrm{e}} \tag{6}$$

or

$$P = \frac{c_\mathrm{f}}{c_\mathrm{e}} \tag{7}$$

Note that, if a collection device is rated in terms of overall collection efficiency, this quantity depends not only on the equipment design and operating conditions, but also on the type of dust and, especially, the particle-size distribution. The stated overall collection efficiency or overall penetration thus always refers to a partic-

ular case; this severely restricts its applicability to other situations.

A more useful and informative measure is the *fractional collection efficiency,* also called the "grade efficiency," "separation efficiency," or "separation function." The fractional collection efficiency $T(x)$ measures collection as a function of particle size, x; it is defined as

$$T(\mathrm{x}) = \frac{\mathrm{d}\dot{M}_\mathrm{c}(x)}{\mathrm{d}\dot{M}_\mathrm{e}(x)} = 1 - \frac{\mathrm{d}c_\mathrm{f}(x)}{\mathrm{d}c_\mathrm{e}(x)} \tag{8}$$

Hence,

$$T(x) = 1 - \frac{Pq_\mathrm{f}(x)}{q_\mathrm{e}(x)} \tag{9a}$$

or

$$T(x) = \frac{Eq_\mathrm{c}(x)}{q_\mathrm{e}(x)} \tag{9b}$$

Procedures for measuring $T(x)$ are based on Equations (9 a) and (9 b); Equation (9 a) is employed most commonly in separation technology.

Typical fractional collection curves are presented in subsequent chapters on individual types of collection equipment. The *cut size* x_t is the particle size at which a given fractional collection efficiency is achieved. A value of 50 % is often used ($x_{50,\mathrm{t}}$) but other cut sizes (e.g., $x_{90,\mathrm{t}}$, $x_{99,\mathrm{t}}$) may be chosen depending on technical requirements and specific applications.

One advantage of the fractional collection efficiency is that, being a ratio, it is independent of the quantity (e.g., number or mass) used to express the amount of particulate material in the particle-size ranges upstream and downstream from the collection equipment. The fractional collection efficiency is a property of the apparatus and can thus be used to estimate the expected overall collection efficiency for any given particle size distribution; this is important in the selection and sizing of dust collectors:

$$E = 1 - P = \int_{x_\mathrm{min}}^{x_\mathrm{max}} T(x) \cdot q_\mathrm{e}(x)\,\mathrm{d}x \tag{10}$$

The overall collection efficiency calculated by means of Equation (10) is, of course, expressed in the same quantity as the particle-size distribution of the inlet gas $q_\mathrm{e}(x)$; it is usually a mass fraction but can also be a number fraction.

3. Centrifugal Collectors (Cyclones)

3.1. General

The simplest way to separate particles from gases is to allow them to settle out in so-called gravity settlers into zones of low gas velocity [11]. Because this method is technically feasible only for coarse particles that are significantly larger than 100 µm, gravity settlers now have negligible importance and are not discussed here.

Even before the end of the 19th century, collection by means of centrifugal forces was known to be much more efficient than simple gravity separation. This led to the development of centrifugal collectors; the most commonly used form employs flow reversal and is called a gas cyclone.

Because of their simple design, reliable operation, small space requirement, and low cost, cyclones are used widely in many branches of industry. They are frequently employed in material recovery from gas recycle systems, pneumatic conveying, and other areas.

Cyclones can be operated at pressures of 0.001 – 10 MPa and temperatures exceeding 1200 °C. They currently provide the most common method of particle collection at high temperature (> 1000 °C) for use on an industrial scale. Growing interest in dust removal from hot gases has given a new impetus to research and development work on cyclones.

The advantages of cyclones over other collectors are offset by the fact that they are less efficient collectors for particles below ca. 1 – 5 µm in size. Therefore, in view of the increasingly stringent requirements on collecting capacity, the use of cyclones is restricted; they are often used as first-stage collectors (precollectors).

3.2. Mode of Operation and Basic Designs

Figure 2 shows the most common type of cyclone with flow reversal. The gas path is designed to impart a twist to the particle-laden gas entering the rotationally symmetrical apparatus. A turbulent, three-dimensional, rotational flow is produced. In the separating chamber spiral, downward flow takes place at larger radii. Flow

is then reversed, and the gas spirals upward at smaller radii, and reaches the exit duct. The particles entrained in the revolving gas stream experience centrifugal forces hundreds to thousands of times greater than the force of gravity. Thus, the larger particles migrate outward to the cyclone wall, where they collect as strands; they are carried downward along the tapered body and out of the separating chamber by boundary-layer flow. The conical shield (d), often built into the bottom of the tapered body, prevents already collected material from being reentrained. The particles that are not collected are carried inward by the gas flow and removed through the exit duct (vortex finder).

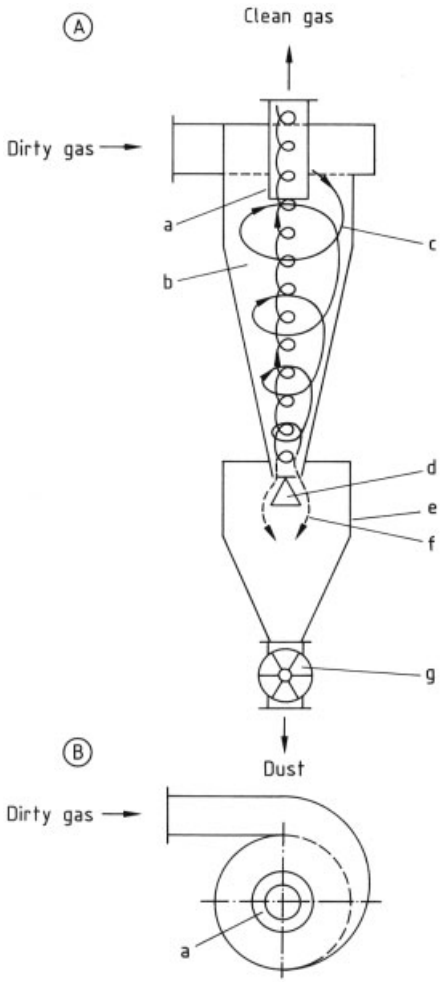

Figure 2. Schematic of a reverse-flow cyclone
A) Vertical section; B) Top view
a) Exit duct; b) Separating chamber; c) Gas flow pattern; d) Conical shield; e) Dust hopper; f) Dust; g) Lock

Cyclones vary in the form of the swirl-producing inlet (tangential, axial with swirl vanes); the shape of the separating chamber (cylindrical, conical); the design of the exit duct; and their dimensions. Diameters range from 5 m to 2 cm (the latter for measurement purposes); flow rates vary from 1 to 100 000 m^3/h. Reviews of cyclone construction are given in [12], [13].

3.3. Design Calculations

Important parameters in cyclone design and operation are the fractional collection efficiency $T(x)$ (defined in Eq. 8) and the energy consumption, which is measured in terms of the pressure drop Δp.

3.3.1. Collection Efficiency

In theory, particle trajectories inside the cyclone can be calculated to determine the fractional collection efficiency; such calculations have been attempted [14], [15]. This method provides insight into processes occurring in the cyclone which improve the understanding of particle separation. However, because the flow field is complicated and turbulence must be considered, this procedure is quite complex and not yet adequately proven. Thus far, it has been primarily of scientific interest and has not been used practically.

Models for Low Dust Loads. Several models have been devised for the approximate calculation of collection efficiency at low dust loads [16–22]. They describe the flow field in more or less simplified terms. The flow field is shown schematically in Figure 3. An especially important parameter is the tangential velocity component v_φ, which initially increases with decreasing radius and can be approximated in this region by

$$v_\varphi \cdot r^n = \text{constant} \tag{11}$$

Here r is the radius and n is an empirical exponent that takes into account the loss of angular momentum due to friction. In a loss-free vortex, n would be unity; in real gas flow, it lies between 0.5 and 0.8. The value of n depends on

other factors, such as wall roughness, particulate load, and cyclone geometry [23].

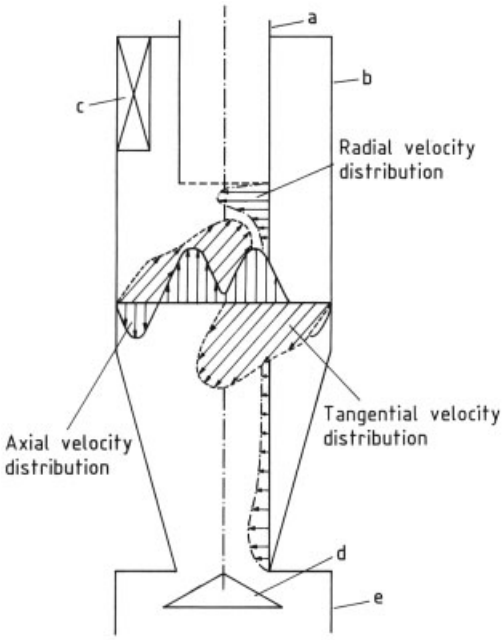

Figure 3. Flow field in a gas cyclone
a) Exit duct; b) Separating chamber; c) Gas inlet; d) Conical shield; e) Dust hopper

Because of a strong "vortex" in the cyclone, the tangential velocity depends mainly on the radius and not on the height. In the core of the cyclone, the tangential component decreases rapidly and approaches zero at the axis.

Design models are based either on a residence-time concept [17] or a separation-zone concept [18], [19]. Recent theoretical work combines both concepts [20–22].

The separation-zone model proposed by BARTH [18] and improved by MUSCHELKNAUTZ [19] is now employed widely because of its simplicity and ease of use. However, this approach does not give the complete fractional efficiency curve; only a *critical particle diameter* x_{cr} can be calculated [16]:

$$x_{\mathrm{cr}} = \left[\frac{18\mu_{\mathrm{g}} \cdot v_{\mathrm{r,i}} \cdot r_{\mathrm{i}}}{(\varrho_{\mathrm{p}} - \varrho_{\mathrm{g}}) \cdot v_{\varphi,\mathrm{i}}^2} \right]^{\frac{1}{2}} \tag{12}$$

where μ_{g} is the dynamic viscosity of the gas, ϱ_{p} is the particle density, ϱ_{g} is the gas density, r_{i} is the internal radius of the exit duct, $v_{\mathrm{r,i}}$ is the radial velocity at radius r_{i}, and $v_{\varphi,\mathrm{i}}$ is the tangential velocity at radius r_{i}.

According to this concept, all particles whose diameter x is less than x_{cr} pass through the device; all particles with $x > x_{\mathrm{cr}}$ are collected.

In theory, the fractional efficiency curve jumps from 0 to 1 at x_{cr} (curve a, Fig. 4). BARTH assumed that x_{cr} corresponds to the median diameter $x_{50,\mathrm{t}}$ of the separation function $T(x)$, i.e., the 50 % value on the effective S-shaped fractional efficiency curve (curve b, Fig. 4). On the basis of experimental results (curve c, Fig. 4), the suggestion has been made that the actual cut diameter be estimated as $x_{50,\mathrm{t}} \approx 1.3\, x_{\mathrm{cr}}$ [19].

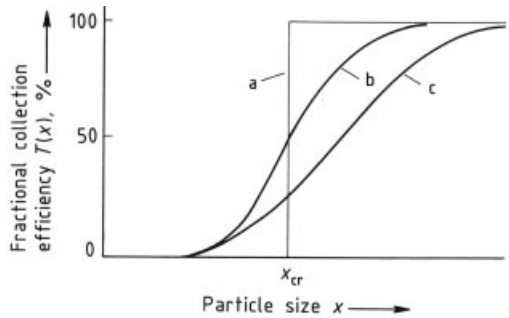

Figure 4. Fractional efficiency curves for cyclones
a) Theoretical cut diameter [18]; b) Assumed effective curve [18]; c) Experimental curve [19]

Although using Equation (12) to predict the critical diameter involves some uncertainty and the complete fractional efficiency curve cannot be derived, experience shows that this formula has an important advantage: experimental values of x_{50} can be converted easily and reliably to different conditions, i.e., the change in the cut diameter produced by a change in cyclone geometry, gas velocity, or gas temperature (viscosity) can be estimated.

In a more detailed approach [21], [22], particle motion is treated as a process in which diffusional particle transport is superimposed on a predetermined motion. Particle flux calculations are based on a modified flow field [23]. The advantage of this model is that the complete fractional efficiency curve can be obtained and the entire cyclone geometry allowed for. The computing time needed for numerical calculation is short, and processing can be done with an ordinary personal computer. Comparison has shown good agreement with measured values.

Models for High Dust Loads. The computational techniques described up to this point have all presumed that the particle concentration entering the cyclone is low and, therefore, has a negligible effect on flow conditions in the separating chamber. This is the case up to a particle load of ca. 10 g/m^3. In practice, however, cyclones often operate at higher concentration. Experiment has shown that the collection efficiency of a cyclone improves with increasing particulate load in the inlet gas, even though the tangential velocity (which is responsible for collection in the vortex) and the centrifugal acceleration decrease simultaneously.

At higher particle concentration, additional collection mechanisms contribute to particle separation, and these must also be taken into account in design.

By analogy with pneumatic conveying, MUSCHELKNAUTZ [19] reasons that a fluid stream can transport only a limited particle mass flow. If a limiting dust load c_{m} is exceeded, the excess particle mass is accelerated to the wall immediately after entering the cyclone and moves down into the dust trap in the form of a strand. Collection due to a dust load in excess of c_{m} is assumed to be selective only at low total loads and increasingly nonselective (i.e., independent of particle size) at higher loads. If c_{m} is exceeded and the particles thrown to the wall are collected in the inlet, the remaining particles can be collected selectively in the separating chamber. Thus, particle collection at high concentration can be considered as two consecutive processes. Equations for calculating the overall collection efficiency of a cyclone are given in [16], [24].

Along with particle collection due to excess loading, agglomeration of small particles with large ones can be assumed to occur at high particle concentration; this also improves overall collection efficiency. A model that takes account of agglomeration is described in [21], [25]. Larger particles migrate to the wall of the cyclone at a higher radial velocity than the small particles. This difference in settling velocity leads to collisions between particles of different sizes and hence to agglomeration. Separation of the small particles adhering to the larger ones is thus improved.

3.3.2. Pressure Drop

The total pressure drop Δp of a cyclone consists of pressure losses in the inlet, separating chamber, and exit duct; it is usually calculated as follows:

$$\Delta p = \xi \frac{\varrho_{\mathrm{g}}}{2} v_{\mathrm{i}}^2 \tag{13a}$$

where

$$\xi = \xi_{\mathrm{inlet}} + \xi_{\mathrm{separator}} + \xi_{\mathrm{exit}} \tag{13b}$$

Here ξ is the pressure-drop coefficient, ϱ_{g} is the density of the gas, and v_{i} is the axial velocity in the exit duct.

The calculation of ξ is described in [16]. Experience has shown that the pressure drop in the exit duct accounts for as much as 90 % of the total cyclone pressure drop. Usually, ξ varies from 5 to 40, depending on cyclone geometry; values of 10 – 20 are common. Typical values for Δp are in the range of 500 – 2000 Pa.

The collection efficiency and the pressure drop depend on cyclone geometry and volume flow rate. These quantities can be combined and used to find the "optimum" cyclone with either a minimum pressure drop [26] or minimum equipment volume [19], [27] for a given cut diameter. Calculations of this kind are not always implemented, however, because they can yield unrealistically long cyclone designs, especially for fine particles. Empirical shapes are, therefore, generally used.

3.4. Operating Characteristics

The collection performance of a cyclone depends chiefly on its geometry, gas throughput, inlet gas concentration, and properties of the material to be separated. The change in cut size under different operating conditions can be estimated with Equation (12). For example, an increase in gas viscosity (e.g., when the temperature is raised) decreases performance. If the cyclone geometry is altered (e.g., exit-duct diameter or inlet cross section is diminished), the tangential velocity component $v_{\varphi,\mathrm{i}}$ at the exit-duct radius increases so that a greater centrifugal force acts on the particles. As a result, the cut size $x_{50,\mathrm{t}}$ decreases and particle collection improves.

Collection efficiency is also improved if the volume flow rate through the cyclone increases. Figure 5 shows some fractional efficiency curves for an experimental cyclone in which the inlet velocity v_e, and thus the volume flow rate $\dot{V}$ are varied. The cut size decreases with increasing $\dot{V}$. The following relation between cut size and gas throughput can be derived from Equation (12):

$$x_{50,t} \sim \frac{1}{\sqrt{\dot{V}}} \tag{14}$$

When the tangential velocity in the cyclone increases, however, the pressure drop Δp also increases. To achieve good collection performance and acceptable energy consumption when gases with high flow rates have to be cleaned, several small cyclones connected in parallel are preferable to a large cyclone because the reduction in cyclone diameter shifts the cut size toward smaller particle sizes. Figure 6 presents typical fractional efficiency curves for three cyclones with different diameters but the same pressure drop (the gas flow rate $\dot{V}$ decreases at the smaller diameters). Cleaning a gas with a given flow rate, therefore, requires more cyclones as the cyclone diameter decreases. If geometrically similar cyclones are assumed and Δp is constant, the result of Equations (12) and (13) is that the cut size for N cyclones in parallel decreases in proportion to $(1/N)^{1/4}$.

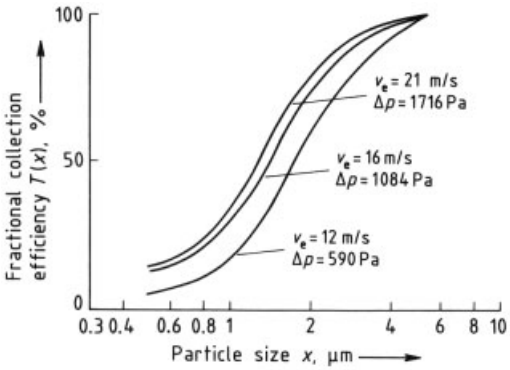

Figure 5. Fractional efficiency curves for a cyclone at various inlet velocities v_e and pressure drops Δp
Particle concentration = 0.5 g/m^3; external radius of cyclone = 95 mm; radius of exit duct = 37 mm; cyclone height = 612 mm; cross-sectional area of inlet duct = 0.003 m^2

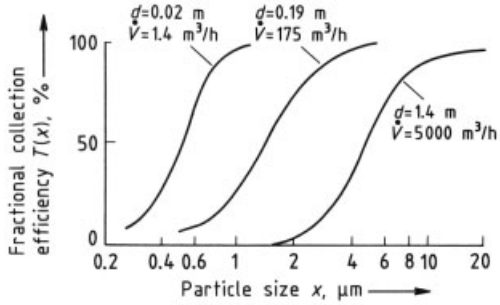

Figure 6. Fractional efficiency curves for cyclones of varying diameter d at a constant pressure drop
$\dot{V}$ = volume gas flow rate

Systems composed of many small cyclones are often called *multicell collectors* or *multicyclones.* An example is shown in Figure 7.

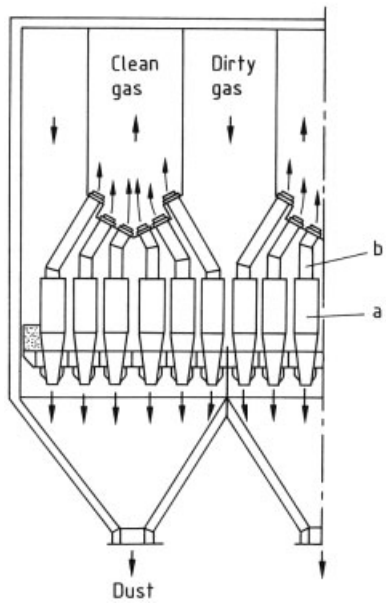

Figure 7. Multicyclone (Lurgi, Multiklon) with bent clean-gas pipes
a) Separating cell; b) Clean-gas pipe

Cyclones with diameters of a few centimeters and particle cut sizes less than 1 µm are used mainly for testing and measuring purposes rather than industrial dust collection.

Substantial differences are often found between calculated and measured collection efficiencies. This discrepancy occurs because the cyclone design is unsuitable for the internal flow pattern and, thus, unfavorable for collection.

An especially important point is the correct design of the dust removal system. Long vertical tubes have proved valuable, as have conical shields which must be mounted below the cyclone cone. The cyclone should not be too long, and the ratio of the dust discharge opening to the exit-duct radius should be greater than one; this prevents the inner vortex core from being deflected toward the wall and reentraining already collected particles.

In multicell collectors, the cells should be loaded uniformly; short-circuit flow at the dust removal ends of the cells leads to poorer separation and must be prevented.

Finally, no welded joints or other flow-obstructing ridges should protrude into the cyclone. Such irregularities can yield secondary flow patterns that reduce collection performance. Detailed information on cyclone design can be found in [28].

4. Scrubbers

4.1. General

In wet collectors (scrubbers), both particulate and gaseous components can be separated from the carrier gas. Only particle collection is discussed below; for gas separation see → Absorption.

Even very fine particles ($\leq 0.1\,\mu m$) can be collected in scrubbers. The particles are brought into contact with a scrubbing liquor, generally water; they bind to the liquor and are removed from the gas stream in bound form as a slurry. Thus two operations must take place in a scrubber: (1) incorporation (binding) of particles into the scrubbing liquor and (2) separation of the dust-laden liquor. In most scrubbers, especially those with the best collection efficiency for fine particles, the scrubbing liquor is in droplet form. The first operation can, therefore, be regarded as the agglomeration of (small) solid particles and (larger) liquid particles. These larger droplets are readily collected from the gas stream. However, care must be taken to ensure the efficient separation of dust-laden liquid droplets, especially in high-energy scrubbers which use relatively fine droplets.

Scrubbers are relatively easy to adapt to various operating conditions. They are employed mainly when products are to be recovered from a wet phase or when the danger of dust explosion or other hazards forbids dry collection. They have proved suitable chiefly for moderately large-volume gas flow rates (often $<30\,000\,m^3/h$).

The solid materials collected in scrubbers are transferred from a gas into a liquid, usually water; therefore, their use entails either wastewater treatment (which may also be dictated by the presence of absorbed gases such as hydrogenchlorate or fluorine) or recycling of the scrubbing liquor in the plant. In either case, water consumption has to be minimized by setting up a closed liquor cycle and optimizing collection conditions. Finally, the danger of corrosion or icing should be mentioned.

4.2. Mode of Operation and Basic Designs

Many scrubber designs are available; examples can be found in [29–32]. However, most designs fall into five basic types, shown in Figure 8 (see next page).

The *spray tower* (Figure 8A) is the oldest type of scrubber. The dirty gas flows upward through a large tube at a relatively low speed of ca. 1 m/s. The washing fluid is dispersed countercurrent to the gas by spray nozzles that are mounted at different levels. The dust-laden liquid leaves the bottom of the scrubber. Packings, e.g., glass beads, are sometimes used to improve collection efficiency (packed-bed scrubbers). In any case, a droplet separator must be integrated in the clean-gas duct.

The *ejector venturi scrubber* (Fig. 8B) is basically a water-jet pump. Water is sprayed into the scrubber at a speed of ca. 25 – 35 m/s and provides a draft for moving the gas which often attains a flow rate of 10 – 20 m/s. Two successive stages are often used because the collection efficiency of ejector venturi scrubbers is not much higher than that of spray towers.

Self-induced spray scrubbers (Fig. 8C) are available in many designs. In contrast to other scrubbers, the dust-laden gas is fed through the wash fluid. The wash fluid is thus atomized, entrained, and mixed with the gas. This effect is reinforced by a redirecting vane. Entrained, dust-laden fluid collects inside the scrubber in the

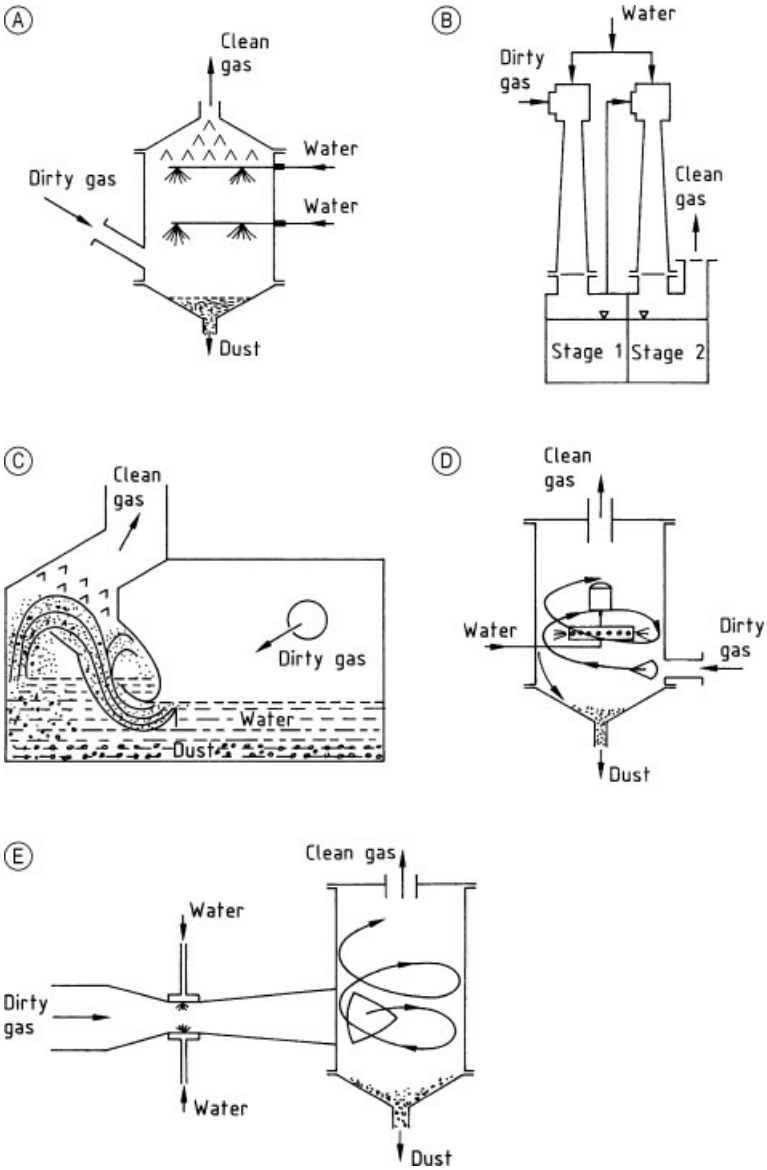

Figure 8. Basic types of scrubbers
A) Spray tower; B) Ejector venturi scrubber; C) Self-induced spray scrubber; D) Rotational scrubber; E) Venturi scrubber

collection tank. The resulting slurry must be removed at intervals. Water consumption is limited to that required to replace losses associated with the gas or slurry removal.

In *rotational scrubbers* (Fig. 8D), the washing fluid is dispersed radially in the gas by rotating spray nozzles. The gas spirals upward and exits through a tube, as in a cyclone. The dirty water is removed from the bottom of the scrubber.

The *venturi scrubbers* (Fig. 8E) have the highest collection efficiency. The gas is accelerated to up to 150 m/s in the venturi throat (orifice). The washing fluid enters the middle of the throat just in front of its narrowest point or is fed radially into the throat. The liquid is dispersed by

Table 1. Operating data for scrubbers

Parameter	Packed-bed scrubber	Ejector venturi scrubber	Self-induced spray scrubber	Mechanical scrubber	Venturi scrubber
Cut size, µm∗	0.7 – 1.5	0.8 – 0.9	0.6 – 0.9	0.1 – 0.5	0.05 – 0.2
Relative velocity, m/s	1	10 – 25	8 – 20	25 – 70	40 – 150
Pressure drop, kPa	0.2 – 2.5		1.5 – 2.8	0.4 – 1.0	3 – 20
Water consumption, L/m^3	0.05 – 5	5 – 20∗∗		1 – 3∗∗	0.5 – 5
Energy consumption, kW · h/1000 m^3	0.2 – 1.5	1.2 – 3	1 – 2	2 – 6	1.5 – 6

∗ Particle density $\varrho_P = 2.42$ g/cm^3.
∗∗ For each stage.

the inflowing gas. The dust-laden water droplets are separated from the gas in a cyclone.

Typical operating data are listed in Table 1. Increased collection (i.e., reduction in critical particle diameter) is associated with greater relative velocity and higher energy consumption. With venturi scrubbers, for example, a cut size (50 % value on the fractional efficiency curve) of about 0.1 µm can be attained, but the pressure loss increases to 20 kPa.

Scrubbers operate by transporting dust particles to the scrubbing liquor and binding them to it. The wettability of the particles does not play a crucial role in binding; wettable particles penetrate the interior of the droplet, while nonwettable particles adhere to the surface. The transport step is decisive in separation. Diffusion and condensation may certainly take place, but in general, inertial forces appear to dominate [33]. Furthermore, because the scrubbing liquor is assumed to be mostly in droplet form, the transport process can be described in clear terms.

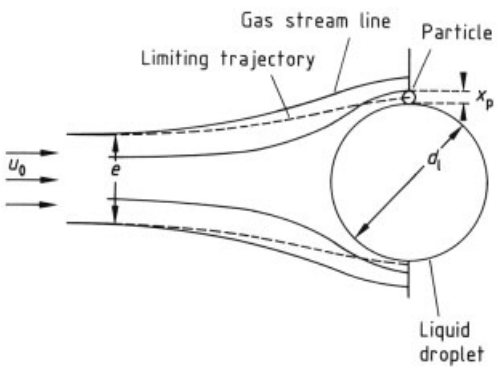

Figure 9. Trajectories and flow lines for flow past a spherical droplet
U_0 = gas velocity; e = diameter of flow tube; d_l = droplet diameter; x_P = particle diameter

The gas flow lines diverge as they approach the liquor droplet (Fig. 9, see below). By virtue of their inertia, the particles follow the flow lines only partially, if at all; instead, they fly toward the droplet surface. Particles strike the droplet if their initial position in the undisturbed flow upstream of the droplet is inside a flow tube with diameter e. Particles located outside the tube bypass the droplet.

A *limiting trajectory* can thus be defined as the line separating particles that collide with the droplet from those that pass by it. The *collision efficiency* η is defined as

$$\eta = \left(\frac{e}{d_l}\right)^2 \tag{15}$$

where d_l is the droplet diameter. Observations indicate that practically all particles striking the droplet adhere to it [34]; therefore, the collision efficiency η corresponds to the single-droplet collection efficiency. Collection in the scrubber is then a result of the overall action of individual droplets. Figure 10 (see next page) gives fractional efficiency curves and overall collection efficiencies for the scrubbers shown in Figure 8. Measurements were performed with sillitin, a very fine dust composed of silicon dioxide [30].

4.3. Design Calculations

4.3.1. Collection Efficiency

Development and sizing of scrubbers are often purely empirical because the processes that occur in these complicated pieces of equipment are complex and difficult to simulate with theoretical models. However, Barth [35] and Calvert [36] have proposed models that are quite helpful because (1) they allow clear identification of important parameters and trends; (2) they have

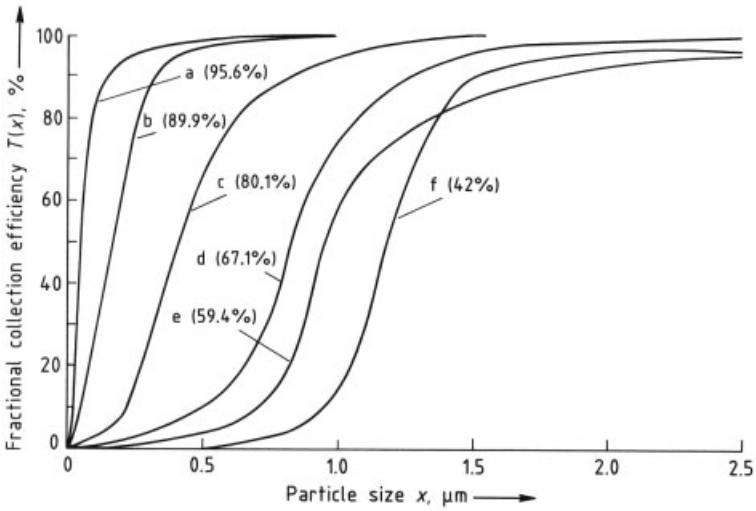

Figure 10. Fractional collection efficiencies for various scrubbers [30]
Measurements were made using sillitin, mean particle size = 1.5 µm, particle density = 2.42 g/cm^3; percentages indicate the overall collection efficiency for sillitin.
a) Venturi scrubber, Δp = 10 kPa; b) Two-stage rotational scrubber; c) Single-stage rotational scrubber; d) Self-induced spray scrubber; e) Single-stage ejector venturi scrubber; f) Packed-bed spray scrubber

been at least partly tested by experiment [5], [37]; and (3) they contain geometrical dimensions that are important for size calculations.

In these models, the scrubbing liquor is assumed to be in droplet form, and inertial forces are considered crucial in the transport of dust particles to the droplet (cf. Section 4.2). The calculation consists of three steps:

1) Calculation of single-droplet collection
2) Calculation of the gas volume cleaned by the single droplet
3) Calculation of the change in dust concentration due to the action of all droplets

The model depicted in Figure 9 is the basis for calculating the single-droplet collection efficiency; the limiting trajectory is found by solving the equations of motion for dust particles numerically. Important parameters are the Reynolds number and the Stokes parameter.

The Reynolds number *Re* for flow around the droplet is given by

$$Re = \frac{v_{\mathrm{rel}} \cdot d_{\mathrm{l}} \cdot \varrho_{\mathrm{g}}}{\mu_{\mathrm{g}}} \tag{16}$$

where ϱ_{g} is the gas density, μ_{g} is the viscosity of the gas, and v_{rel} is the relative velocity between the droplet and the gas.

The inertial parameter (Stokes parameter) ψ describes the dynamics of the particles:

$$\psi = \frac{\varrho_{\mathrm{p}} \cdot U_0 \cdot x_{\mathrm{p}}^2}{18 \mu_{\mathrm{g}} \cdot d_{\mathrm{l}}} \tag{17}$$

where ϱ_{p} is particle density and x_{p} is particle size.

The numerical results presented in Figure 11 have been confirmed experimentally. To a good approximation [37], they can be calculated as

$$\eta = \left(\frac{\psi}{\psi + k_1} \right)^{k_2} \tag{18}$$

where η is the collision efficiency (Eq. 15) and k_1 and k_2 are constants that depend on the Reynolds number. For most operating conditions, the collection process takes place primarily in the range $Re < 40$. Here the results can be approximated well with $k_1 = 0.65$ and $k_2 = 3.7$ [37]. A more detailed investigation into this topic can be found in [38].

As the droplets move through the apparatus, they clean a certain volume of gas. The ratio of the cleaned volume to the droplet volume, the *specific cleaned volume m,* is the most important parameter for sizing a scrubber:

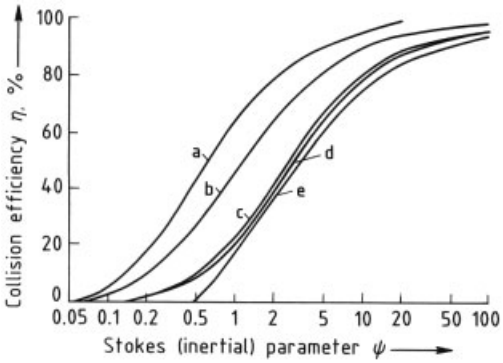

Figure 11. Calculated collision frequencies η vs. Stokes (inertial) parameter ψ for various Reynolds numbers *Re*
Values of *Re*: a) $\gg 1$, potential flow; b) 60, 80; c) 40; d) 10, 20; e) < 1, viscous flow

$$m = \frac{3}{2d_1} \int_{t_1}^{t_2} \eta\,(t)\, v_{\mathrm{rel}}\,(t)\, \mathrm{d}t \tag{19}$$

where t is time.

The value of m is found by (numerically) solving the equation of motion (Eq. 1) for the droplet. In this way, the relative path length between the droplet and the gas in the time difference $t_2 - t_1$ (residence time) is determined.

The results depend on scrubber geometry, operating conditions (gas velocity and direction), and particle properties (size, density, initial velocity). Equation (19) thus provides an important and useful basis for sizing scrubbers. The drawback of this procedure is that the calculated results cannot be generalized but are valid only for the selected boundary conditions. Furthermore, numerical calculation requires a certain amount of computational effort, but this should not be a problem in view of the present state of computer technology.

An optimal droplet size usually exists for the specific cleaned volume, as can be seen in Figure 12. Under other operating conditions, this optimum can also occur at droplet sizes much smaller than 100 µm [5].

In the third step, the total effect of all the droplets in changing the dust concentration is calculated. The result is the fractional or overall collection efficiency.

The particle concentration c in the scrubber is expressed by

$$c = c_{\mathrm{e}} \cdot \exp\,(-b \cdot m) \tag{20}$$

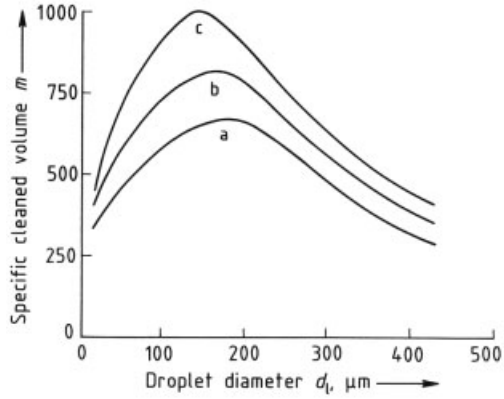

Figure 12. Specific cleaned volume vs. droplet diameter d_{l} for different particle sizes
Particle size x_{p}, µm: a) 2.0; b) 2.5; c) 3.2
Velocity of the entering droplets $w_{\mathrm{e}} = 7$ m/s; gas velocity $U = 7$ m/s

where c_{e} is the particle concentration in the gas inlet to the scrubber; m is the specific cleaned volume; and b is the specific scrubbing-liquor ratio ($b = \dot{V}_{\mathrm{l}}/\dot{V}_{\mathrm{g}}$, where $\dot{V}_{\mathrm{l}}$ is the volume flow rate of scrubbing liquor and $\dot{V}_{\mathrm{g}}$ is the volume flow rate of gas). Typical b values are $(0.5-5)\times 10^{-3}$, i.e., 0.5 – 5 L of liquor per cubic meter of gas (see Table 1).

To find the fractional separation efficiency $T\,(x_{\mathrm{p}})$, m is calculated as a function of particle size x_{p}:

$$T\,(x_{\mathrm{p}}) = 1 - \exp\,[-b \cdot m\,(x_{\mathrm{p}})] \tag{21}$$

The overall collection efficiency is finally obtained by integrating the product of the fractional efficiency and the particle density (cf. Eq. 10). The complete droplet-size distribution must be used for these calculations. If a mean, often arbitrarily selected value is used, considerable errors can occur [39].

The model proposed by Calvert [36] utilizes the same basic approach as that of Barth [35]. The effort required for the numerical computations is shortened substantially by fitting the equations to experimental values, so that the equation of motion for the droplets need not be solved. However, this procedure entails some loss of general applicability and information content.

For the venturi scrubber, the following equation is given:

$$T\,(x_{\mathrm{p}}) = 1 - \exp\left[-\frac{2 \cdot \dot{V}_{\mathrm{l}} \cdot d_{\mathrm{l}} \cdot \varrho_{\mathrm{l}} \cdot v}{55 \cdot \dot{V}_{\mathrm{g}} \cdot \mu_{\mathrm{g}}}\, F\,(\psi, f')\right] \tag{22}$$

where v is the maximum gas velocity, ψ is the inertial parameter (Eq. 17), and f' is an empirical fitting parameter, often in the range of 0.25 – 0.5.

The following approximation is given for the function $F(\psi, f')$ in the range $1 \leq \psi \leq 4$:

$$F(\psi, f') \approx 0.312 \cdot \psi \cdot f'^2 \tag{23}$$

Equation (22) can also be used, along with a pressure-drop equation, for optimization [5]. In general, model calculations indicate that higher relative velocities and increased addition of scrubbing liquor lead to better collection efficiency; this is also observed experimentally (Table 1).

4.3.2. Pressure Drop

Calculation of the pressure drop is of interest chiefly for venturi scrubbers. Empirical formulas and model calculations are available for this purpose.

The *empirical formulas* are based on the procedure for evaluating flow in pipes:

$$\Delta p = \xi \frac{\varrho_g}{2} v_K^2 \tag{24}$$

where v_K is the gas velocity in the venturi throat and ξ is the pressure-drop coefficient:

$$\xi = k_a + k_b \frac{\dot{V}_l}{\dot{V}_g} 10^3 \tag{25}$$

where k_a and k_b are empirical constants that depend on the equipment and thus cannot be generalized; typical values are $0 \leq k_a \leq 0.5$ and $0 < k_b \leq 1.7$.

In *model calculations,* the pressure-drop contributions of the gas and droplets are added. The droplet contribution is taken into account by solving the equation of droplet motion and

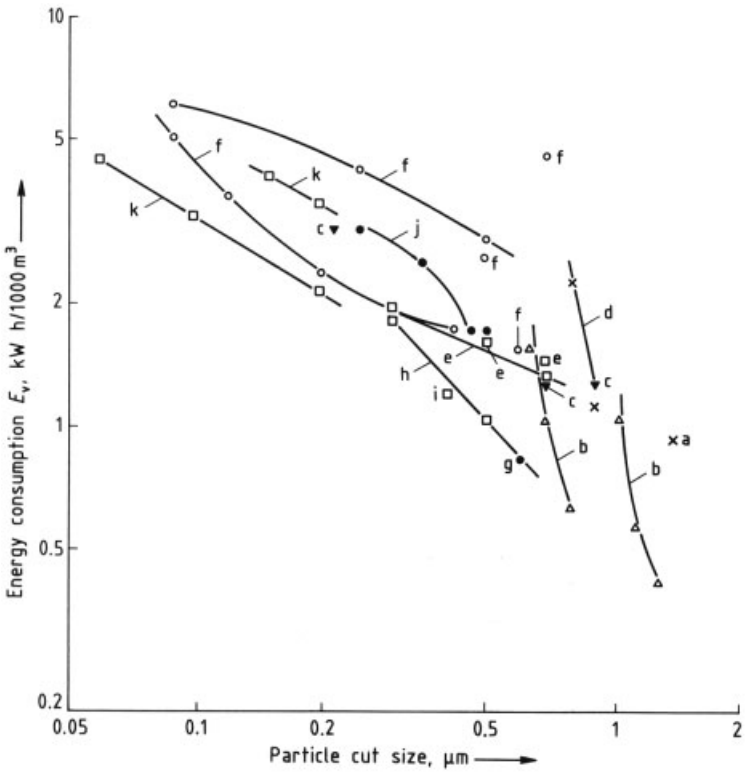

Figure 13. Energy consumption of scrubbers as a function of cut size
a) Wet cyclone; b) Spray tower; c) Self-induced spray scrubber; d) Ejector venturi scrubber; e) Low-pressure venturi scrubber; f) Rotational scrubber; g) Impinger; h) Low-pressure venturi–impinger scrubber; i) Wet cyclonette; j) Scrubber based on different principles; k) High-pressure venturi scrubber

by introducing the laws of two-phase flow [40]. The resulting formulas can be solved only by numerical techniques; they are in good agreement with experiment. No further details are discussed here; however, the extent of agreement between calculation and experiment depends on whether the scrubbing liquor is uniformly distributed over the total flow cross section or whether a significant fraction flows down the walls of the scrubber as a film.

Experiments have repeatedly shown that energy consumption grows as the particle cut size decreases, i.e., as collection efficiency increases. Empirical equations have also been derived on the basis of this observation [41]. Such equations are not, however, well suited to the sizing or optimization of scrubbers, because the several independent variables (equipment dimensions, dust properties, operating conditions) are not explicitly stated. The correlation between energy consumption and cut size holds only for a given size and type of equipment. Considerable discrepancies can occur between different types of scrubbers (Fig. 13) [30].

4.4. Droplet Separators

Efficient dust collection in a scrubber requires separation of the dust-laden droplets in a separator that is either integrated into the scrubber or connected downstream. Because droplet sizes generally range from 1 µm to several hundred micrometers, commercial separators utilize inertial forces to separate droplets from the gas stream. Layers of wire mesh or fabric, stacks of plates, or centrifugal devices are employed [5], [42].

Lamellar (plate) separators are often used because of their simple construction and low pressure drop. These devices consist of parallel channels with multiple deflections and liquidretaining grooves (Fig. 14). The droplet-laden gas stream undergoes many changes of direction in the channels. By virtue of their inertia, droplets collide with the wall to create a film which is "peeled" off in the retaining grooves. To prevent product caking, the plates can also be irrigated; a second downstream stage is then installed to trap the droplets entrained at the leading edges of the plates in the first stage. The plates can have various profiles. Typical channel widths are 20 – 30 mm; gas velocities are 5 – 10 m/s for horizontal flow and 2 – 3 m/s for vertical flow; pressure drops are a few hundred pascals. The usual cut diameters are ca. 5 – 10 µm.

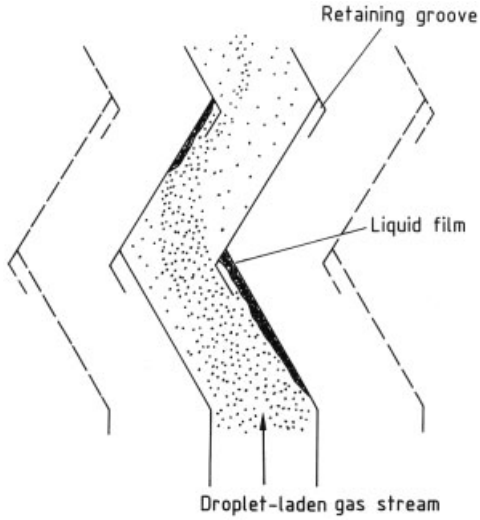

Figure 14. Droplet collection by stacks of plates

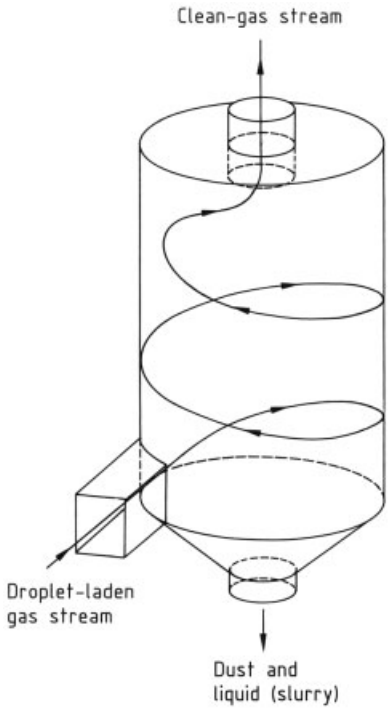

Figure 15. Centrifugal droplet separator

Centrifugal separators (cyclones) are often employed if high separation efficiencies are required, especially downstream of venturi scrubbers. Figure 15 (see previous page) shows a simple form.

Like all inertial collectors, lamellar and centrifugal separators become more efficient as flow velocity increases. However, an upper velocity limit must not be exceeded; otherwise, collected liquid is reentrained (cyclones), or the formation of secondary droplets due to impact on the liquid film becomes significant (lamellar separators). These limits must be determined empirically for each type of equipment and application.

5. Filters

5.1. General

Filters constitute an important aid in modern particle collection technology, especially when highly efficient collection in the finest particle-size range is required. Many versions are available. Of all collectors, these have the widest spectrum of applications and thus a large share of the market.

With regard to application, design, and mode of operation, filters for dust collection can be classified as deep-bed filters [43] or surface filters [44], [45].

Deep-bed filters are used where dust loads are low, on the order of several milligrams per cubic meter of gas. Typical applications include air conditioning and ventilation, laboratory exhaust systems, clean rooms, and respiratory protection. Such filters usually consist of a relatively loose fiber mat with a pore fraction over 90 % (often > 99 %). As the gas flows through the fibrous layer, particle collection takes place in the interior of the layer where the dust accumulates (deep-bed filtration). After becoming saturated with dust, these filters are usually discarded; some can be cleaned by washing or blowing. Typical gas velocities are 0.1 – 3 m/s.

Surface filters are used for high dust loads, on the order of several grams to hundreds of grams per cubic meter; this is the typical range for industrial dust removal and air pollution control. The filter media used to be mainly woven fabrics; today, nonwovens such as mats or needled felts, often with surface coatings, are preferred. The pore-volume fraction in these media is 70 – 90 %. Collection occurs mainly at the filter surface, in the resulting dust layer (filter cake). The dust bed is the true filter medium and is highly efficient: emission values of a few milligrams per cubic meter or less can be attained. Because the pressure drop increases as the cake grows, these filters must be cleaned periodically. Collected dust, which may be the product, is recovered. Depending on operating conditions, cleaning can take place at intervals of a few minutes to several hours. Typical gas velocities are 0.5 – 5 cm/s.

Combinations of surface and deep-bed filters have been discussed with a view to satisfying extreme emission limits (e.g., 0.1 mg/m^3).

Because filter design and functioning as well as system design differ greatly for deep-bed and surface filters on the one hand, and those using fibrous layers and granules on the other, three groups are discussed separately.

5.2. Fibrous Filters (Deep-Bed Filters)

5.2.1. Mode of Operation and Basic Designs

The fibrous layers in filter-medium systems consist mainly of synthetic fibers (glass or polymers). Table 2 summarizes typical geometrical data. The spacing between the fibers is large compared with the size of the particles to be collected. This implies that the filters do not function as a sieve; the purely geometrical blocking action is not crucial for particle collection. Various mechanisms (diffusion, inertia, electrostatic attraction) cause the particles to be transported to the fibers, where they must then adhere.

Table 2. Geometrical data for deep-bed

Parameter	Coarse filter	Submicron filter
Fiber diameter D_F, µm	50 – 100	1 – 5
Mat thickness Z, cm	1 – 3	0.1 – 0.3
Fiber volume fraction $(1-\varepsilon)$, %	$\lesssim 1$	$< 5 - 10$
Mean fiber spacing	$9\,D_F$	$3\,D_F$

Filter mats are installed in a variety of ways in accordance with plant requirements and available space. Figure 16 shows the most important types. In the simplest case (A), the flat filter

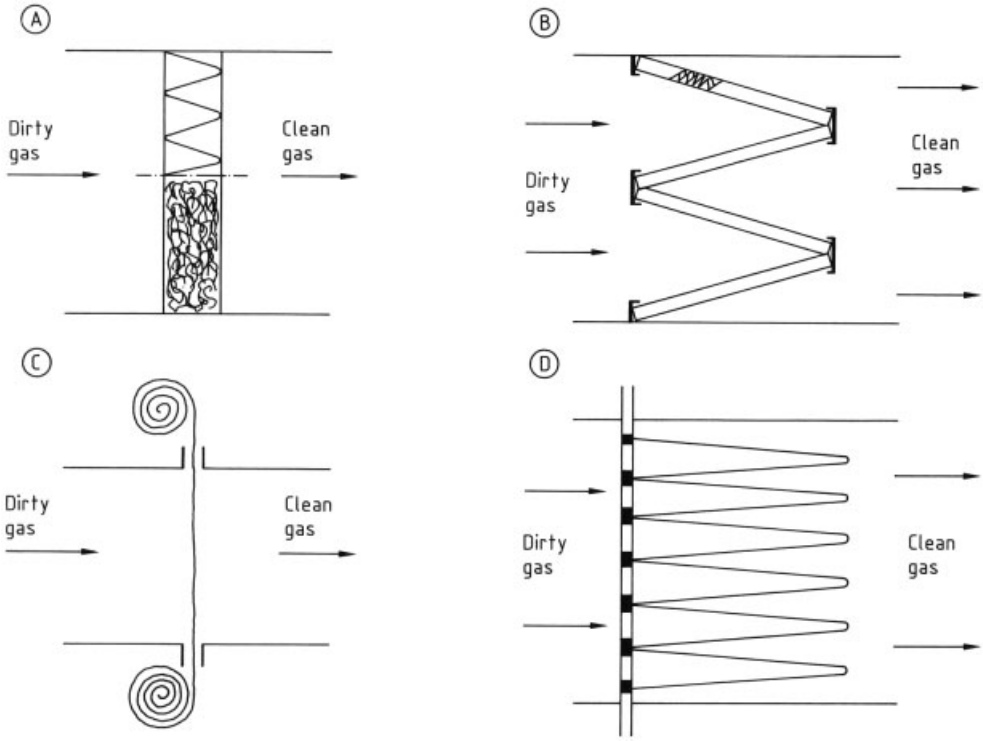

Figure 16. Designs of fibrous deep-bed filters
A) Flat filter; B) Stationary zigzag filter; C) Roll filter; D) Pocket filter

mat is inserted perpendicular to the direction of flow. To increase filtration area (lower flow velocity → lower pressure drop → longer lifetime for a given channel cross section), the filter is often folded and mounted in a frame. A further gain in filtration area is obtained by a zigzag placement of frames with folded filter mats (B).

Because the filters become plugged with collected dust over time, they must be replaced at certain intervals (usually several months to years). The method used for filter mounting is therefore very important. The mount must be tight but easy to handle. Filter replacement can be effected, for example, with a roll filter arrangement (C). This rather costly design has been increasingly superseded by stationary zigzag filters (B) or by pocket filters (D). This avoids the danger of releasing collected dust when the filter strip is advanced.

The main application for fibrous deep-bed filters is inlet-air cleaning, i.e., for low dust loads. However, since the early 1980s high-efficiency filters have been used in exhaust-gas systems where particle loads are higher but the clean gas must have especially low particle concentration, e.g., for toxic or harmful dusts composed of heavy metals or radioactive substances. Two-stage submicron filters are then used (Fig. 17). The first stage (a) is periodically cleaned by reverse blowing with a reciprocating nozzle carriage (b). The second stage (c) is replaced after it is saturated with dust; because of the low dust concentration, replacement intervals are sufficiently long. The usual cleaning procedure for such a system involves shutting off the gas stream, so for continuous operation at least two units must be connected in parallel. Development work aimed at finding application limits, optimal designs, materials, and operating conditions is still under way. Problems must be expected for very fine, highly adhesive dust, or if the folds in the first filter stage are too narrow. The collection efficiency for such a system may be very good; emission values of 0.1 mg/m^3 or less can be achieved even with a dirty-gas dust load of 10 g/m^3 or more.

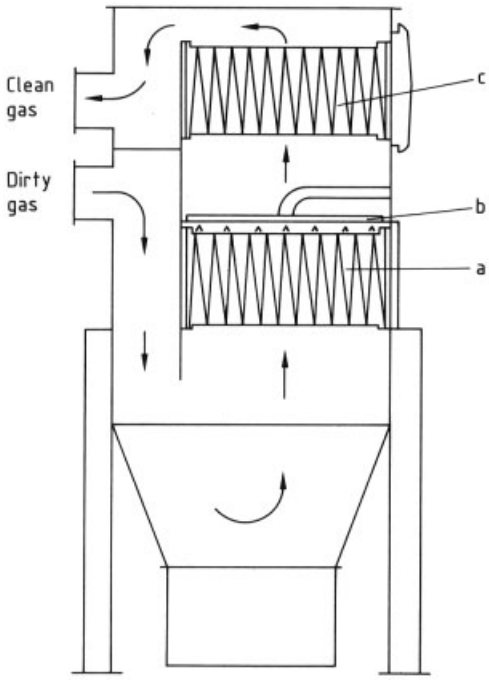

Figure 17. Exhaust-gas cleaning system employing submicron filters
a) First filter stage; b) Nozzle carriage for cleaning filter; c) Second filter stage

The range of collection efficiencies for deep-bed filters varies widely and depends on the design and operating conditions. It extends from a single prefilter for relatively coarse dust to submicron particulate filters for clean rooms, where the DIN standard requires a collection efficiency of at least 99.97 % for 0.3-µm particles [46].

With regard to temperature, deep-bed filters are used mainly at $< 100\,°C$. In paint shops, temperature resistance up to 120 °C is required. Some designs can also be used up to 300 °C. These filters generally employ glass fibers in metal cartridges with appropriately temperature-resistant sealing materials.

Another application of deep-bed filters is the collection of liquid droplets, especially oil or acid mist. Here, the collected liquid flows continuously downward out of the filter, so that the pressure drop is stable. A variant of this use is the collection of soluble dusts with simultaneous solvent feed. Again, the collected material flows continuously out of the filter layer.

5.2.2. Design Calculations

The models for calculation presented in this section refer to unloaded filters, i.e., to the initial conditions. With regard to collection efficiency, this is usually the critical state because collection generally improves with increasing dust load (electret filters are an exception). The methods proposed in the literature for calculating filtration behavior over time are not discussed because they are very complex and still of limited practical use.

5.2.2.1. Collection Efficiency

The fractional collection efficiency $T(x)$ of a fibrous layer is given by

$$T(x) = 1 - \exp\left[-f\prime \cdot \varphi(x)\right] \tag{26}$$

For fibers of circular cross section,

$$f\prime = \frac{4}{\pi} \cdot \frac{1-\varepsilon}{\varepsilon} \cdot \frac{Z}{D_F} \tag{27a}$$

$$f\prime = \frac{4}{\pi} \cdot \frac{1}{\varepsilon} \cdot \frac{M}{D_F \cdot \varrho_F} \tag{27b}$$

$$f = f\prime \cdot \varepsilon \tag{27c}$$

where ε is the porosity, Z is mat thickness, D_F is fiber diameter, M is filter mass per unit filter area, and ϱ_F is the fiber density.

The geometric factor f represents the ratio of the projected fiber area to the filter face area. For prefilters $f \approx 3-10$; for high-efficiency submicron filters, $f \approx 100-300$.

The factor $\varphi(x)$ is the collection efficiency of a single fiber inside the mat. Because collection in a medium (depth) filter includes transport and adhesion of the particles to the fiber,

$$\varphi = \eta \cdot h \tag{28}$$

where η is the collision efficiency and h is the adhering fraction. This factor is included because at the usual filtration velocities and especially for particle sizes $> 1\,\mu m$, not all the particles hitting the fiber surface for the first time adhere [47], [48].

Since the 1930s, research has been concerned with determining η. The importance of the adhering fraction h was not recognized until the

1960s, but research on this factor has been increasing.

Particles can be transported to the fiber surface in three ways: Brownian diffusion, inertial forces, and electrostatic interaction (Fig. 18).

1) *Brownian Diffusion.* The random motion of particles about their mean trajectory [Fig. 18 (a)] is due to the thermal motion of gas molecules (Brownian motion); it can be calculated with Fick's diffusion equations. For engineering purposes, this effect is important chiefly for particles smaller than ca. 0.5 µm and velocities lower than 10 cm/s.

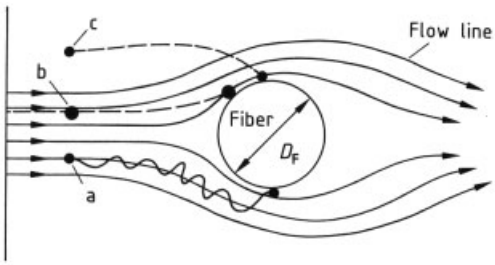

Figure 18. Transport mechanisms for particle collection on fibers
a) Brownian diffusion; b) Inertial forces; c) Electrostatic interaction

This diffusion effect η_D is governed by the following relation:

$$\eta_D \sim \left(U_0 \cdot D_F \cdot x \cdot \frac{\mu_g}{T}\right)^{-2/3} \tag{29}$$

where U_0 is the velocity of the incident gas stream, μ_g is the viscosity of the gas, and T is the absolute temperature.

2) *Inertial forces* [Fig. 18, (b)] cause a predetermined movement of particles toward the fiber. These processes can be described by the equation of motion (Eq. 1).
The decisive parameter for the inertial effect is the inertial (Stokes) parameter

$$\psi = \frac{\varrho_P \cdot x^2 \cdot U_0}{18 \cdot \mu_g \cdot D_F} \tag{30}$$

Inertial effects are important mainly for particles larger than 0.5 – 1 µm. As ψ increases, the collision efficiency number rises rapidly. The Reynolds number for flow past the fiber is also important because the collision efficiency number increases with the Reynolds number.

3) *Electrostatic interactions* [Fig. 18 (c)] can also be described by the equation of motion.

Figure 19 presents measurements of the fractional collection efficiency of a fibrous layer [49]. Characteristic influences of the diffusion and inertial effects can be seen at particle diameters < 0.5 and > 0.5 µm, respectively. In the transition region (ca. 0.1 – 0.5 µm), both effects are very slight. This minimum in the fractional efficiency curve is typical; it represents the critical particle-size range for the filter. For this reason, filter tests are often performed with particle sizes ca. 0.3 µm.

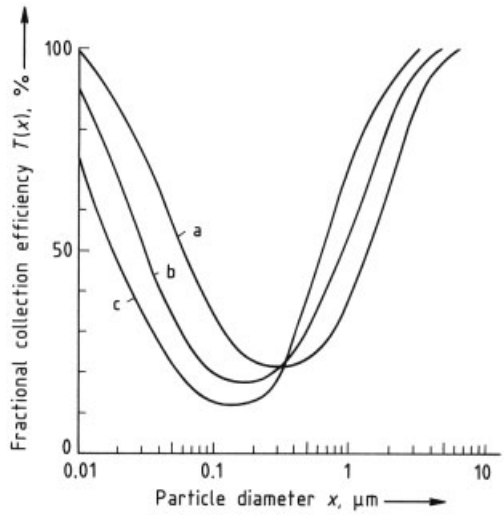

Figure 19. Fractional efficiency curves for a glass-fiber filter at various superficial velocities
Superficial velocity U_0, m/s: a) 0.055; b) 0.2; c) 0.5
Values for $x > 0.5$ µm were determined with a quartz aerosol and for $x < 0.2$ µm with a sodium chloride aerosol

One way of overcoming the efficiency minimum is to utilize electrostatic interactions [Fig. 18, (c)]. Electrostatic forces act over long ranges and can thus attract particles from the flow even if they are far from the fiber.

Calculation of electrostatic effects is very complicated because many configurations of charges and fields are possible. "Electret" filters have attracted much interest. These devices consist of fibers on which a charge is imposed during manufacture [50]. The charge is localized not at the fiber surface but in its interior. Such fibers greatly improve collection, even of uncharged particles (Fig. 20). If the particles carry only a small charge (e.g., are in equilibrium with the

ion concentration in air), even the size range below 0.1 µm is collected quite efficiently [Fig. 20, curve (a)]. The advantage of the electret filter is that, with comparatively slight pressure drops, the collection efficiency is much higher than that of conventional filters, especially in the critical transition region. Its drawback is that its behavior deteriorates with time because, as more particles are deposited on the fibers, the electrostatic effects weaken (depending on particle properties). Further experimental and theoretical results can be found in [49].

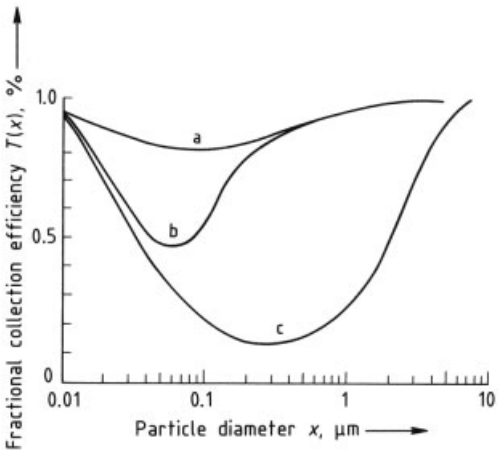

Figure 20. Fractional efficiency curves for an electret filter spun from a polymer solution
a) Fibers and particles charged; b) Fibers charged, particles uncharged; c) Fibers and particles uncharged
Superficial velocity $U_0 = 0.1$ m/s

To obtain the single-fiber collection efficiency φ from the transport parameter η, information is needed about the adhering fraction h (Eq. 28); this is important not only for single fibers but also for fibrous beds. Extensive studies have shown that the adhering fraction generally depends on flow velocity, fiber geometry, particle geometry (size and shape), and surface properties of the fiber and particle [48]. Bouncing of particles from the fibers can begin at velocities lower than 10 cm/s. As the velocity and particle size increase, the adhering fraction usually decreases. Therefore, adhesion in the particle-size range > 1 µm should always be given special attention.

Besides bouncing, resuspension of already collected particles can cause severe problems. The concentration of bio-aerosols, for example, can reach higher values in the "clean gas" than in the original raw gas [51].

The fiber configuration in the filter layer is another aspect that was formerly neglected. Theoretical calculations usually assume a homogeneous distribution of fibers; this is certainly not the case, and uneven distribution strongly affects the collection efficiency. As the inhomogeneity increases, collection usually becomes poorer. More detailed information is given in [52], [53].

5.2.2.2. Pressure Drop

The pressure drop generated in flow through a filter medium can be described by the "drag" model. The pressure drop is attributed to the drag involved in flow past the fibers; in view of the low volume fraction of fibers (< 5 %), this is certainly realistic in physical terms.

Accordingly,

$$\frac{\Delta p}{Z} = \frac{2}{\pi} \cdot \frac{\varrho_g \cdot U_0^2}{D_F} \cdot \frac{(1-\varepsilon)}{\varepsilon^2} \cdot C_D\,(Re) \tag{31}$$

where Z is the thickness of the fiber layer,

ε is the porosity of the fiber layer, and U_0 is the incident flow rate. The drag coefficient C_D depends on the Reynolds number for flow past the fibers:

$$Re < 1: \quad C_D = \frac{8\pi}{Re\,(2 - \ln Re)} \tag{32a}$$

$$1 \le Re \le 50: \quad C_D \approx \frac{10}{Re} + 1.5 \tag{32b}$$

where

$$Re = \frac{U_0 \cdot D_F \cdot \varrho_g}{\mu_g \cdot \varepsilon} \tag{32c}$$

Essential assumptions for derivation of these equations are that the fibers are circular–cylindrical and homogeneously distributed and that the incident flow is perpendicular to the fiber axes. Furthermore, the equations hold only for the initial filtration state, i.e., with no dust on the filter. Within these constraints, agreement with experiment is good. For practical purposes, however, time-dependent behavior is often more important, i.e., the dependence of pressure drop on the quantity of dust trapped. However, because no reliable way of calculating this relationship is

known, the time dependence must be determined by experiment. Optimization can be based on the pressure-drop equation (Eq. 31) and the efficiency equation (Eq. 26) [5] and can give useful information on the design of medium filters.

5.3. Fabric Filters (Surface Filters)

Surface filters are used widely because of their excellent collection performance, even for the finest particles; their adaptability to the most diverse conditions; and significant advances in the development of new filter media (e.g., improved chemical resistance, heat resistance, and cleaning properties). Cake filters are available with filtration areas from a few to several hundred thousand square meters. For a properly designed, well-maintained filter, emission values can be held to a few milligrams per cubic meter (in the best cases less than 1 mg/m^3). The allowable temperature range extends to ca. 250 °C for PTFE and glass-fiber filters. Special metal fibers permit operation up to ca. 600 °C; sintered silicon carbide and other ceramic filters can function to 1000 °C, but they cost substantially more. Filtration rates are typically 40 – 180 m^3 h^{-1} m^{-2} and reach 300 m^3 h^{-1} m^{-2} in exceptional cases; typical pressure drops are 1 – 3 kPa.

5.3.1. Mode of Operation

The principal surface-filter media are nonwovens and (needled) felts that are composed of randomly arranged glass, plastic, or ceramic fibers. Woven materials, chiefly constructed from glass-fiber yarns or metal wires, are also employed.

Only at the beginning of filtration do these media trap particles in the interior of the fibrous layer. As more dust is deposited, it bridges the fibers and forms a dust layer (filter cake) on the surface.

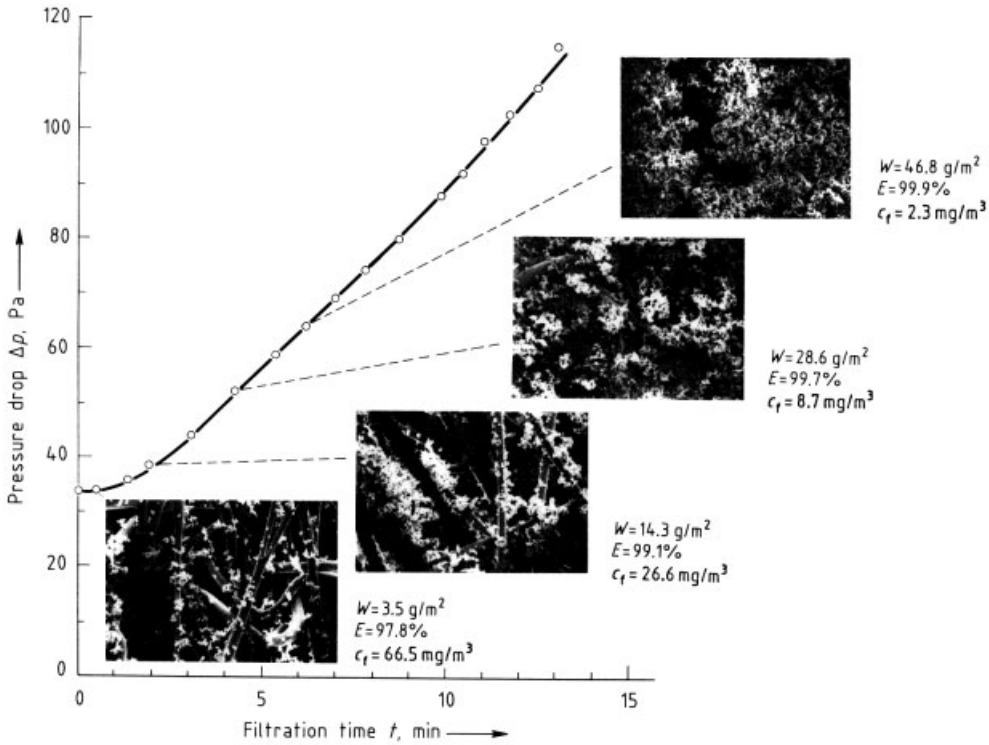

Figure 21. Time dependence of filter-cake formation, pressure drop, and dust collection on a needled felt
W = amount of dust deposited; E = total collection efficiency; c_f = dust content of clean gas; dust content of dirty gas = 3 g/m^3; gas velocity = 150 m/h

Figure 21 shows a typical example of filtration behavior for a needled felt, in which the initial emission value was 66.5 mg/m^3 and decreased to 2.3 mg/m^3 after 46.8 g/m^2 of dust had been deposited.

The fractional efficiency curves in Figure 22 clearly show the improvement in collection performance as more dust is collected. The new, clean filter gives a typical depth-filter curve (a); after 5 min, the efficiency is over 99 % throughout the particle-size range covered (f). Since the cake-formation phase lasts only a short time for the dust loads usually encountered in practice, the interception effect inside the cake is dominant. This process is called surface filtration.

As the amount of dust deposited in and on the filter grows, the pressure drop also increases. After a certain time, or when a predetermined pressure-drop value is reached, the dust must be removed (cleaned off) to regenerate the filter. Filtration periods may be 10 min or less when the amount of incident dust is large (i.e., high concentration and high velocity); if the dust load is small, filtration can last several hours. As a rule, filters should be regenerated only when absolutely necessary (cf. Section 5.3.2); this saves operating costs and reduces emissions.

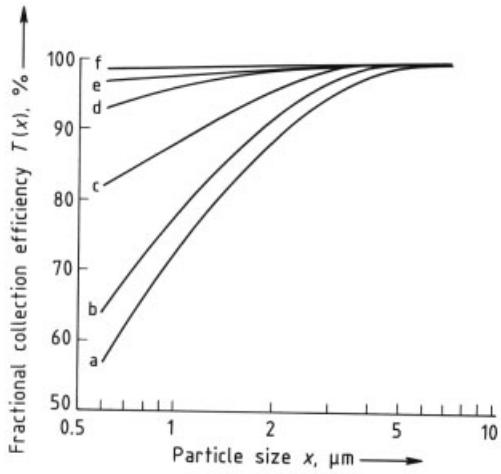

Figure 22. Variation in fractional collection efficiency due to dust deposition on a polyester needled felt
Amount of dust deposited, g/m^2: a) 4, $\Delta p = 33$ Pa; b) 14; c) 29; d) 47; e) 75; f) 95, $\Delta p = 110$ Pa
Dust content of dirty gas = 3 g/m^3; gas velocity = 150 m/h

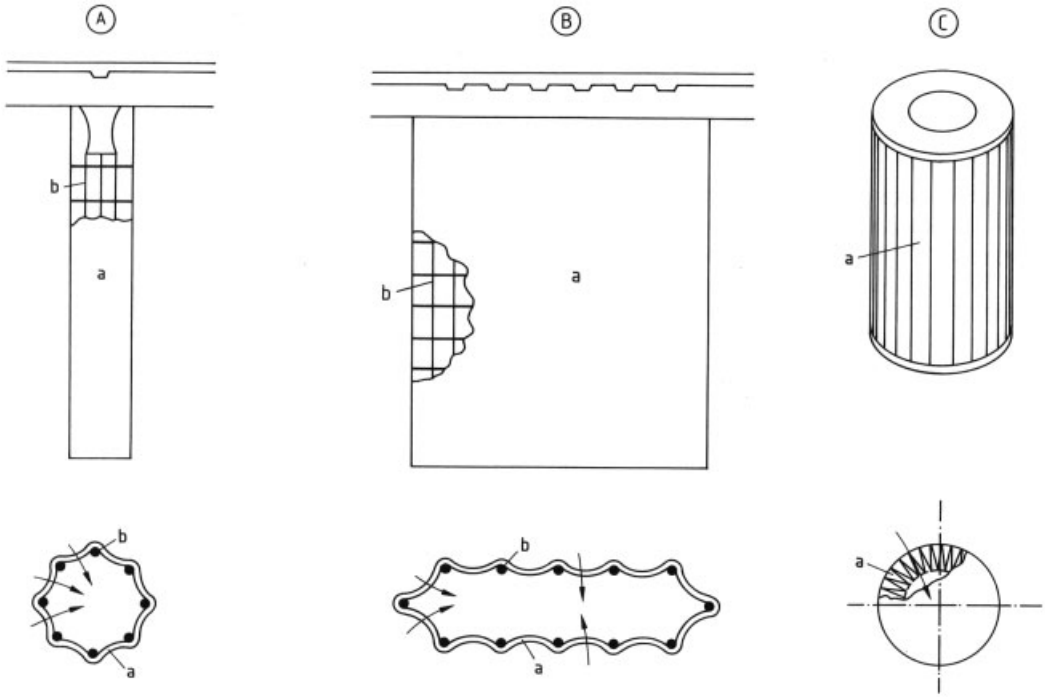

Figure 23. Basic filter designs
A) Round bag filter; B) Envelope filter; C) Cartridge filter a) Flexible filter medium; b) Supporting cage

In modern commercial filter media, surface coating or treatment is used to attempt collection on the surface right from the start and to prevent particles from penetrating into the medium. This approach yields high initial collection efficiencies, avoids the danger of plugging, and facilitates removal of the filter cake.

5.3.2. Basic Designs

The three most important basic surface-filter designs are round bag, envelope, and cartridge (Fig. 23, see previous page) if flexible media are used. A filter unit contains the requisite number of filter elements connected in parallel. Designs differ primarily in the method of cleaning and sometimes in the way the dirty gas is introduced [45], [54], [55].

Round (tubular) bag filters are used most widely. If gas flow during filtration is inward, as shown in Figure 23A, a retainer (supporting cage) is required. Often, however, flow is outward and, in this case, only a few stiffening rings need be incorporated in the bag. Typical bags are 100 – 300 mm in diameter and 1.5 – 10 m long. The filtration area per bag can thus be as much as 10 m^2.

The classical design, introduced in 1887 by the Beth company, is a *multichamber filter with shaker cleaning* (Fig. 24). Dirty gas is fed into the bags from below; it then flows through the filter medium (f) and is led to a clean-gas duct (d) at the top of the device. Particle collection takes place on the inside of the bags. For cleaning, the dirty-gas supply is cut off from one chamber at a time, and the dust is loosened by shaking or rapping the bag support. The dust removed from the bags falls into a dust hopper; purge air drawn into the bags from outside aids this process. Because the bags being cleaned are cut off from the gas flow, several filter chambers must always be used in parallel.

Large multichamber filtration plants (e.g., for cleaning flue gas from furnaces or use in the cement and metallurgical industries [44]) are often designed as *baghouses* (Fig. 25). The bags are usually large (300 mm in diameter, 10 m long). The dirty gas flows from the inside to the outside of the bags, which are cleaned by low-pressure reverse flushing (Fig. 25B). The reversal of flow for cleaning causes the bags to collapse and the dust to drop out. In connection with the surface-

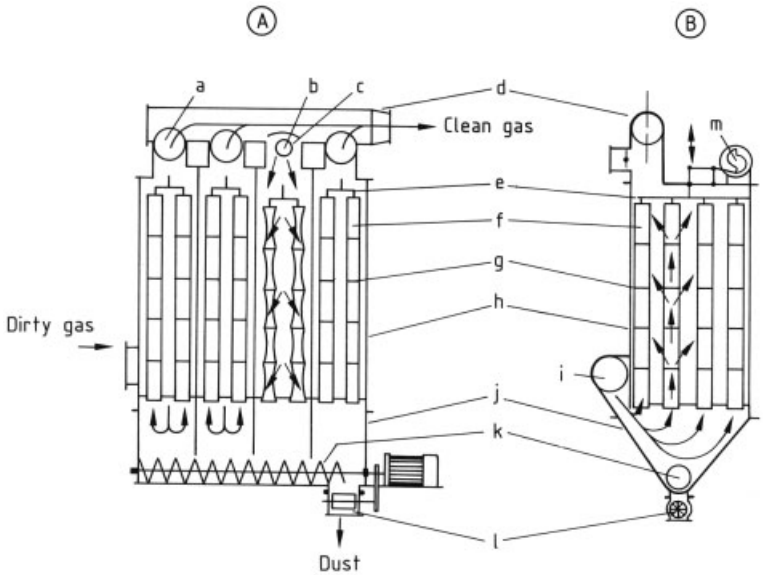

Figure 24. Multichamber bag filtration plant with shaker cleaning
The two vertical sections (A) and (B) are at right angles to each other
a) Valve open for filtration; b) Valve closed for cleaning; c) Purge air inlet; d) Clean-gas duct; e) Bag support; f) Filter bag; g) Stiffening ring; h) Filter housing; i) Dirty-gas duct; j) Dust hopper; k) Screw conveyor; l) Air lock; m) Shaker mechanism

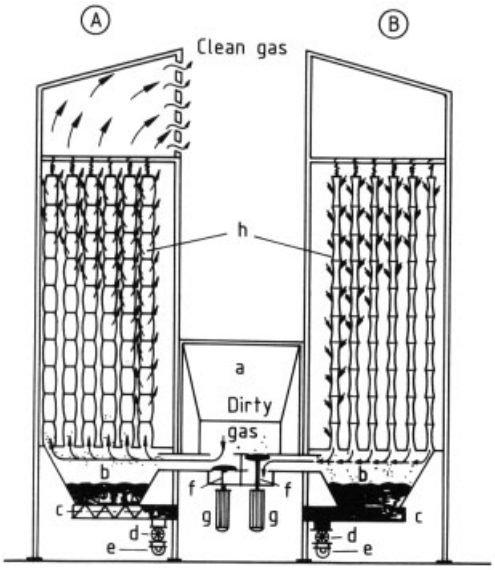

Figure 25. Baghouse with low-pressure reverse flushing
A) Operating filter chamber; B) Filter chamber being cleaned
a) Dirty-gas pressure duct; b) Dust hopper; c) Hopper screw conveyor; d) Air lock; e) Screw conveyor; f) Flushing gas duct; g) Cleaning valve; h) Filter bag

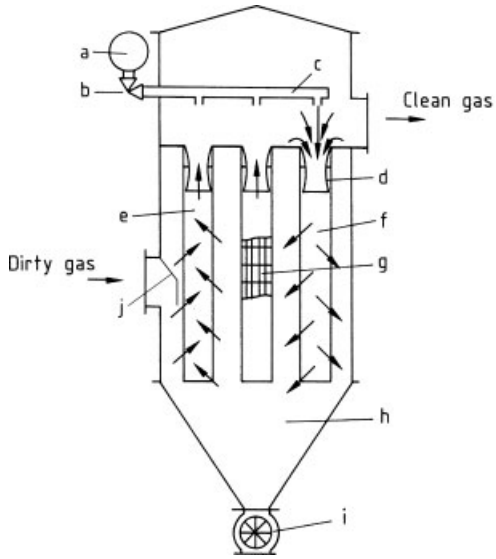

Figure 26. Round bag filter with reverse-pulse cleaning
a) Compressed-air tank; b) Right-angle diaphragm valve; c) Compressed-air duct with nozzles; d) Venturi nozzle; e) Filter bag in operation; f) Filter bag being cleaned; g) Retainer; h) Dust hopper; i) Air lock; j) Deflecting shield

coated filters mentioned in Section 5.3.1, this design offers a particularly economical approach.

Pulse-jet filters (bag filters with reverse-pulse cleaning) became more widely used as progress was made in needle felts. This design dominates in small and intermediate-size units. The gas always flows from outside to inside (Fig. 26), so the bags must be pulled over supporting cages. For cleaning, a jet of compressed air is admitted to the bag through its open top end. The pressure needed in the air tank depends on the design (0.3 – 0.7 MPa). The compressed-air jet suddenly raises the pressure in the bag, which is rapidly inflated and gas flow is reversed. The dust is thus dislodged from the outside of the bag; the process lasts ca. 0.1 – 0.3 s. Because single bags or single rows of bags must be removed from the filtration process for only a short time, the filter does not have to be divided into chambers. Cleaning is done on-line; however, off-line cleaning, in which the dirty-gas flow is cut off in a multichamber filter, may be preferable in extreme cases, e.g., for very fine dusts. Further details on the design and properties of pulse-jet filters can be found in [44], [45], [54].

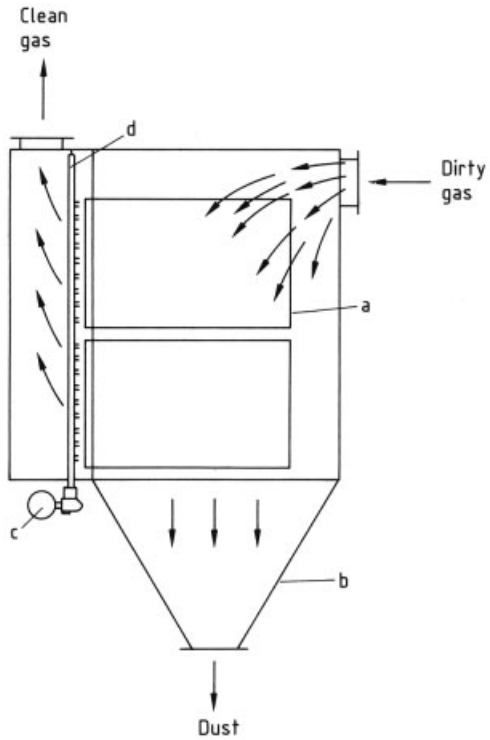

Figure 27. Schematic of an envelope filter unit with reverse-pulse cleaning
a) Envelope filter; b) Dust hopper; c) Compressed-air tank; d) Compressed-air duct with nozzles

Envelope filters (Fig. 23B) are not as widely employed as round bag filters. They are used mainly to remove dust from small amounts of gas, e.g., in silo venting, and as bunker top filters [54].

Flow is always from outside to inside. The filter medium is stretched over a flat, rectangular, supporting cage; dust is collected on the outer surface of the medium. Envelope filters are usually smaller than round bags; filtration areas are often 0.1 – 1.5 m^2. An envelope filtration unit is shown in Figure 27 (see previous page). The usual cleaning method is intermediate-pressure reverse flushing or reverse pulsing. As a rule, multichamber design is not necessary.

Cartridge filters are a further development in surface filters. The filter medium is folded in star fashion (Fig. 23C) to pack more filtration area into a unit volume. With a standard cartridge height [56] of 606 mm, the filtration area is between 5 and 20 m^2, depending on the depth and angle of the folds. Gas flow is always from outside to inside. A multichamber cartridge filter unit is shown in Figure 28.

The filter cake on the outside of the filter is removed either by cutting off the dirty-gas stream and flushing at low pressure with circulating nozzles (multichamber design) or by applying a reverse pulse (on-line). Cartridge filters are very compact devices that offer a relatively large filtration area in the enclosed volume. As a result, low filtration velocities are possible; this is an advantage with respect to emission levels and costs. Until now, however, cartridge filters have been useful only for easily removable dust; their development continues.

5.3.3. Operating Characteristics

Particle collection in a surface filter occurs mainly in the dust layer formed on the surface (Section 5.3.1); efficiencies greater than 99.9 % are possible independent of particle size. Therefore, this highly efficient cake must be retained on the surface as long as possible.

Figure 29A shows that the pressure drop Δp increases as more dust is deposited, and periodic cleaning is necessary. After cleaning, the pres-

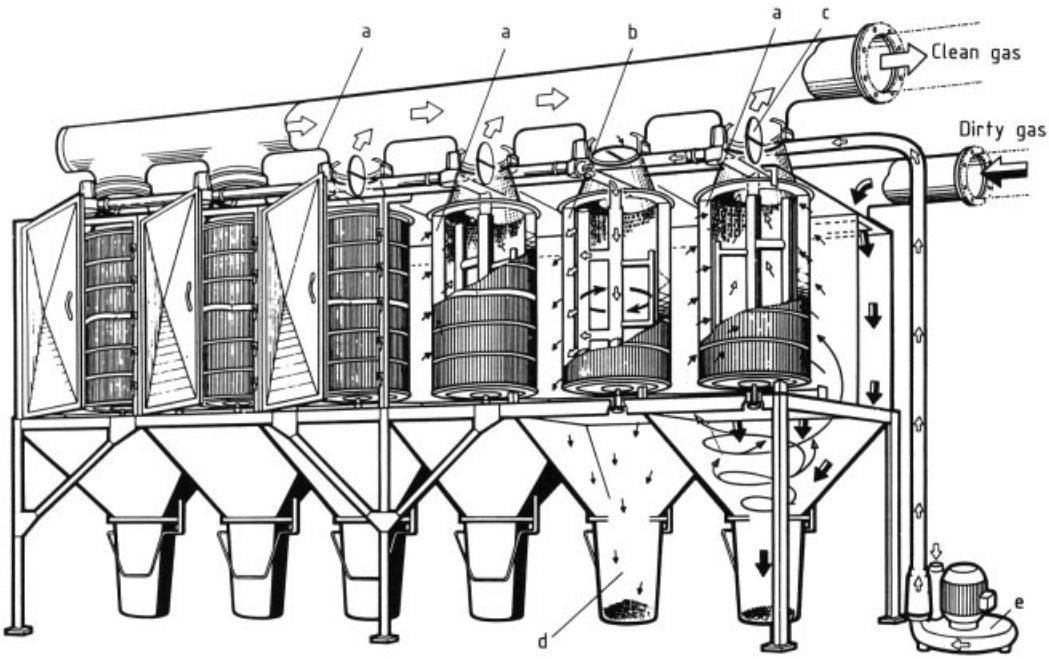

Figure 28. Multichamber filtration unit (Mann & Hummel)
a) Filter in operation (valve open, dust collects on the outside of the filter); b) Filter cleaning (valve closed, cleaning air is pumped outward through the filter and dust falls into the hopper); c) Valve; d) Dust hopper; e) Pump for cleaning air

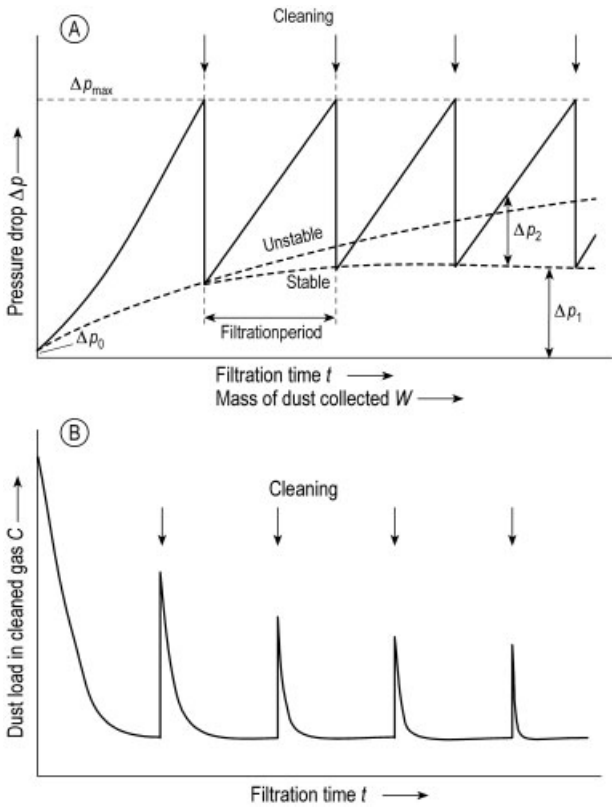

Figure 29. Pressure drop (A) and dust content (B) in clean gas vs. filtration time or mass of dust collected at constant filtration velocity. The values Δp_1 and Δp_2 denote the pressure-drop contributions of the filtration medium and filter cake, respectively

sure drop falls to a value Δp_1 that depends on the flow resistance of the filter medium (including residual dust not removed); Δp_1 is higher than the initial pressure drop Δp_0 of the new, clean filter. For stable filter operation, Δp_1 must reach an approximately constant value after an induction period that can last several days to weeks. The filter medium, filtration velocity, and the method and intensity of cleaning determine whether stable filtration is attained for given gas and dust properties.

During and immediately after cleaning, the particle content of the clean gas increases (Fig. 29B). This occurs partly because the filter cake, which acts as a layer of filter aid, has been removed and must be rebuilt. Especially with reverse-pulse cleaning, some of the dust passes through the filter medium because after the pressure pulse inside the bag has decayed, the filter medium falls back against the retainer and knocks part of the dust through the filter (seepage effect).

Many laboratory and plant studies have shown that the peak concentrations after cleaning may be crucial for the emission value [5], [57]. The maintenance of low emission levels thus demands that cleaning be done as seldom and as carefully as possible. Accordingly, filtration velocities should not be too high, and pressure drop – rather than time – should be used as a criterion for cleaning.

Influencing the operational behavior of surface filters by raw-gas conditioning (dosing of additives and precoating) and other techniques (e.g., electrostatic and acoustic methods) is described in [58]. In some cases considerable potential for improvement exists with regard to pressure drop, filter-medium regenerability, and particle emissions. At present, however, these methods can not be regarded as the industrial state of the art.

5.3.4. Design Calculations

The main purpose of surface-filter design is to determine the required filtration area *F,* which is given by

$$F=\frac{\dot{V}}{v} \tag{33}$$

where $\dot{V}$ is the volume flow rate of dirty gas and v is the filtration velocity; v, the ratio of flow rate to filtration area, is often called the *air to cloth ratio* and is expressed in $m^3\,h^{-1}\,m^{-2}$ of m/min. A proper selection of filter load determines the pressure drop and, ultimately, the emission value.

The filtration velocity is selected on the basis of empirical values because of a lack of generally valid theoretical models. The VDI 3677 standard gives guideline values [55]. A base load value can also be employed, which is modified for actual conditions by means of various correction factors [54].

$$v=v_0 \cdot c_1 \cdot c_2 \ldots c_n \tag{34}$$

where v is the effective filtration velocity, v_0 is the base value, and $c_1 - c_n$ are correction factors. This method is used chiefly for filters with reverse-pulse cleaning. The correction factors always refer to a defined table of base values. They differ from one author to another, being purely empirical and specific in nature.

In describing the time dependence of the pressure drop (Fig. 29), Δp is assumed to be the sum of the contributions of the medium (Δp_1) and the filter cake (Δp_2):

$$\Delta p=\Delta p_1+\Delta p_2 \tag{35}$$

Because flow through the filter takes place at low Reynolds numbers (< 1), Darcy's law is also assumed to hold:

$$\Delta p=K_1 \cdot \mu_g \cdot v + K_2 \cdot \mu_g \cdot W \cdot v \tag{36}$$

where K_1 is the residual resistance of the filter medium after cleaning, K_2 is the specific resistance of the filter cake, W is the mass of dust collected per unit filter area, and μ_g is the viscosity of the gas. Since W is proportional to filtration time, the second term in Equation (36) increases with time and describes the time dependence of Δp.

Important parameters in Equation (36) are the resistances K_1 and K_2; as a rule, they are not constants but functions of several variables. Prediction of these parameters is still problematic because the functions are not known well enough [5], [54]. Even when K_1 and K_2 can be determined by experiment, great care must be taken to ensure that conditions in the experimental device and the full-scale plant are comparable.

5.4. Granular-Bed Filters

5.4.1. General

In granular-bed filters, the dust-laden gas stream flows through the bed and is thereby cleaned. The granular bed functions as a filter medium and in most cases can be considered a depth or deep-bed. Under some conditions, however, dust bridges and then a filter cake are formed, so that surface filtration can occur.

Besides pure particle collection, the removal of gaseous components in granular-bed filters ("dry sorption" of gases such as sulfur dioxide, hydrogen chloride, and hydrogen fluoride) has attracted increasing interest (→ Adsorption). Granules for these applications must be made of appropriate sorbents (often calcium compounds). As a result of more stringent demands for air pollution control, the combined collection of dust and noxious gases may become still more important for small and medium-sized installations. However, only particle collection is discussed here.

The applications of granular-bed filters are determined by the material properties of the granules making up the bed. Examples of such materials are gravel, sand, ceramics, and activated carbon.

Granular-bed filters are used chiefly for the removal of hot, chemically aggressive, abrasive, or adhesive dust or when danger of fire exists. Main applications are in the lime and cement, rock and soil, metallurgical, chemical, and nuclear industries.

The service temperature range is limited by the materials used for plant construction rather than by the filter granules; in some cases, the behavior of the collected particles is also important. Operation up to 450 °C presents no prob-

lem; with special designs, 800 °C or more is attainable.

5.4.2. Mode of Operation

Granular beds can be either fixed or moving. Table 3 presents typical operating data.

Table 3. Typical data for granular-bed filters

Parameter	Value
Granule diameter, mm	0.5 – 5
Bed heights, cm	5 – 20
Porosity, %	40 – 50
Incident flow rate, m/s	0.5 – 2.5
Pressure drop (no dust), Pa	500 – 1500

Fixed-bed granular filters are operated batchwise; i.e., when dust deposition has raised the pressure drop to a predetermined limit, the dirty-gas stream is cut off and the bed is regenerated in one of two ways. Either the dust is removed internally by reverse flushing and agitation or stirring, or else it is removed externally by screening or other means. For continuous operation, these techniques always require several chambers in parallel, which are regenerated in rotation.

In *moving-bed granular filters*, the granules move (slowly) through the filter chamber by gravity, either steadily or at intervals. Fresh granules are added continuously from above. Regeneration takes place in an external loop. This filtration method operates without interruption of the dirty-gas stream. Multichamber design for regeneration is thus unnecessary.

Dust collection occurs mainly in the interior of the bed. Particles are transported to the filter granules by the same mechanisms as in fibrous depth filters: diffusion, inertial forces, and (less often) electrostatic attraction (cf. Section 5.2.2). For successful collection, the particles must adhere to the granules. Sieve effects become significant only after the deposition of sufficient dust and the formation of bridges or a cake.

Once a critical load or a critical pressure drop has been reached, dust can penetrate to the clean-gas side of the bed. With fixed beds, dust bridges can collapse; with moving granular beds, a danger exists that friction between granules will loosen collected dust particles, which will be carried out of the filter (reentrained).

Fluidized-bed granular filters have not gained a significant market share until now, but research continues.

5.4.3. Basic Designs

Fixed-Bed Filters. An example of a fixed-bed filter is the so-called *Drallschicht* (*DS*) *filter* shown in Figure 30 [59]. A cyclone-like precollector (b) removes coarse particles. Collection takes place as the gas flows downward through the bed (d). For regeneration, the dirty-gas stream is interrupted by adjusting the appropriate shutoff valve (e), causing clean gas to flow back through the bed while the bed is raked (g). Removed dust is collected partly in the hopper of the precollector and partly in another filter connected in parallel (multichamber design).

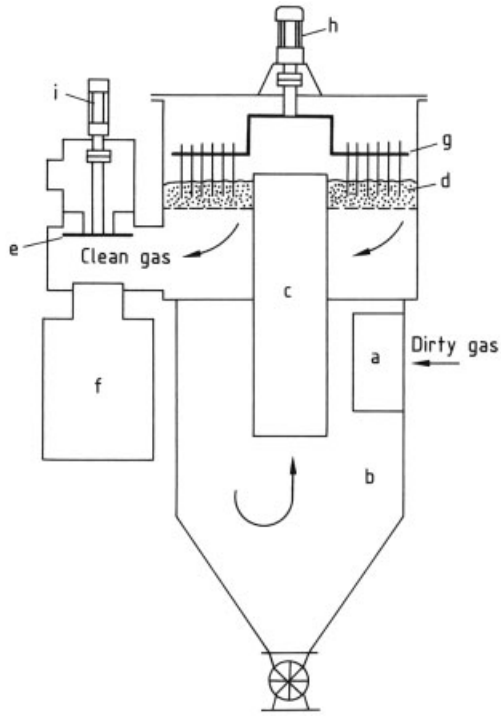

Figure 30. "Drallschichtfilter" in operation (Lurgi)
a) Tangential dirty-gas inlet; b) Precollector; c) Exit duct; d) Granular bed; e) Changeover valve; f) Clean-gas duct; g) Rake; h) Drive; i) Valve actuator motor

In the *granulate-tube filter* (Fig. 31), the granular filter material is contained in perforated double-wall pipes (c); the gas flows horizontally

through the pipes. For regeneration (right-hand side of Fig. 31), the dirty-gas stream is again cut off. Purge gas sets the granules in motion and conveys them in a closed loop; at the same time, clean gas flows backward (horizontally) through the pipes and carries the loosened dust to the hopper (g).

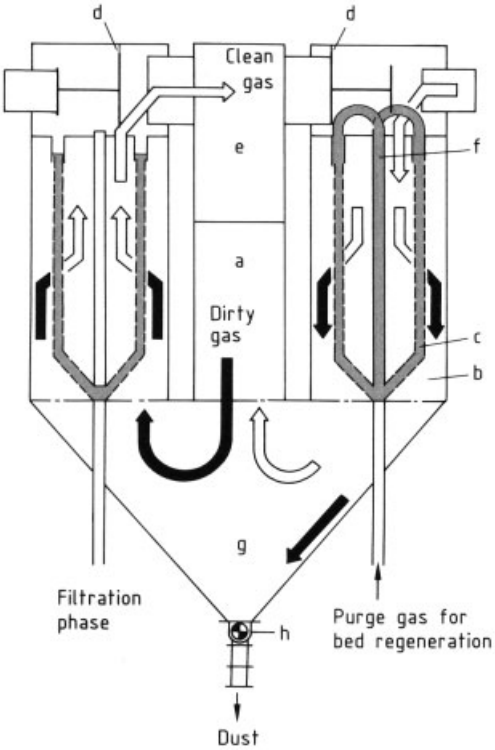

Figure 31. Granulate-tube filter (Lurgi)
a) Dirty-gas duct; b) Filter space; c) Perforated granulate tubes; d) Shutoff valve; e) Clean-gas duct; f) Riser pipe; g) Hopper; h) Screw conveyor

Moving-Bed Filters. An example of a moving-bed bilter is the *countercurrent granular-bed filter* (Fig. 32). Gas and filter material move countercurrently. The dust content of the dirty gas is first lowered in a cyclone-like precollector. The gas then flows radially into the filter bed where it is redirected and slowed. The gas then flows upward through the fresh filter material (b) into the clean-gas plenum (a). To increase filtration area, the filter material is distributed in several conical containers stacked one above another; gas flows through these beds in parallel.

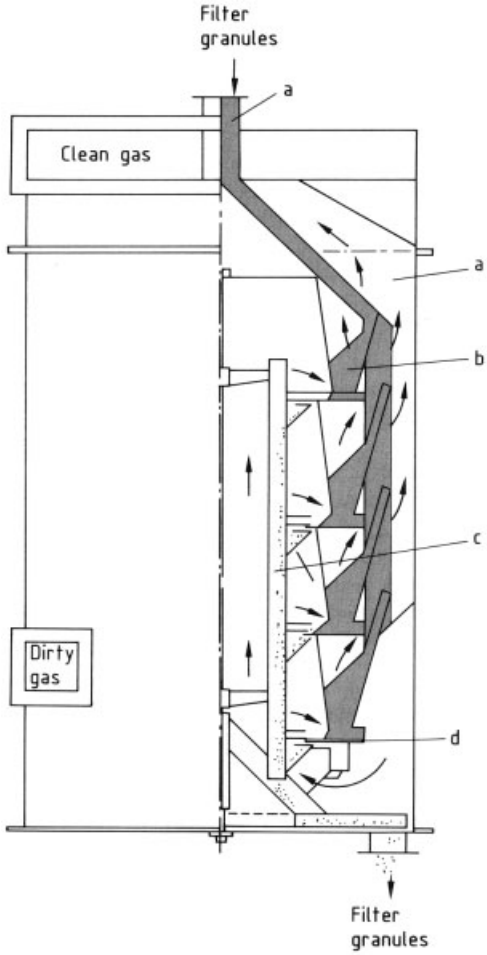

Figure 32. Vertical section of a countercurrent filter (Rüskamp)
a) Clean-gas plenum; b) Granular bed; c) Downpipe; d) Scraper
Arrows indicate direction of gas flow

Dust-laden filter material is removed at intervals by a scraper (d), collected in the tube (c), regenerated externally in vibrating sieves, and returned to the top of the collector. This design and mode of operation make a multichamber device unnecessary.

5.4.4. Design Calculations

The design of granular-bed dust filters is mainly empirical (as in collection generally), and development of theories is not yet complete. However, existing formulas still yield valuable information about important variables and trends and,

therefore, are discussed briefly. A detailed discussion can be found in [5].

5.4.4.1. Collection Efficiency

The initial fractional collection efficiency (no dust load in the bed) is described by

$$T(x) = 1 - \exp\left[-1.5\frac{(1-\varepsilon)}{\varepsilon}\cdot\frac{H}{d_G}\cdot\varphi(x)\right] \qquad (37)$$

where ε is the porosity of the bed, H is its height, d_G is the diameter of the filter granules, and $\varphi(x)$ is the single-granule collection efficiency.

The dependences of $T(x)$ on filter geometry (ε, H, d_G) have been confirmed experimentally [60], [61]. As the dust load in the bed increases, fractional efficiency increases, at least up to the critical load. The fractional efficiency curves in Figure 33 were measured with relatively small filter granules (d_G= 0.28 mm) and at low filtration velocities [62]. Under these conditions, the shift from collection in the bed to collection in the dust layer (i.e., to surface filtration) is relatively quick. As a result, fractional efficiencies are very high and only slightly dependent on particle size.

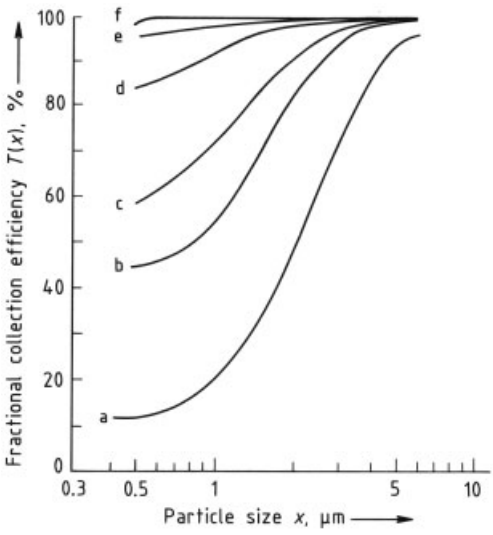

Figure 33. Fractional efficiency curves for a limestone bed (height 5 mm, granule diameter 0.28 mm) with various dust loads of quartz F 500 at a gas velocity of 0.055 m/s
Mass of dust W collected per unit filter area, g/m^2 (pressure drop Δp, Pa): a) 0 (108, initial trial); b) 6.98 (117); c) 11.97 (141); d) 17.95 (244); e) 25.93 (452); f) 50.85 (1110)

The difficulty in calculating fractional efficiency with Equation (37) is determination of the single-grain collection efficiency $\varphi(x)$. Transport mechanisms are the same as in fibrous-bed filters (diffusion, electrostatic attraction, inertia; see Section 5.2.2), but defining the flow field near a filter granule (which is assumed to be spherical) is more difficult.

Because the filter granules are close together, the flow fields derived for isolated spheres are too inexact, and correction factors are introduced to allow for the packing density [60]. At higher velocities, even flow fields modified in this way do not closely approximate reality.

The experimental fractional efficiency curves in Figure 34 essentially confirm theoretically predicted effects and trends. Below about 0.5 µm and 0.5 m/s (sodium chloride particles), diffusion effects are dominant. In this region, collection efficiency increases with decreasing particle size x, decreasing velocity U_0, and decreasing filter-granule size d_G. Inertial forces control collection to the right of the minimum (Quartz F 500), but here the curves pass through a maximum. The maximum shifts toward smaller particle sizes as the velocity increases, due to inadequate adhesion of particles when they collide with the filter granules. This phenomenon still occurs in much thicker beds and becomes even more evident if Equation (37) is used to calculate the single-grain efficiency from measured fractional efficiencies [61]. These tests, performed on beds up to 9.7 cm thick and composed of 4-mm spheres, show that the dust particles bounce off at velocities as low as 2 m/s.

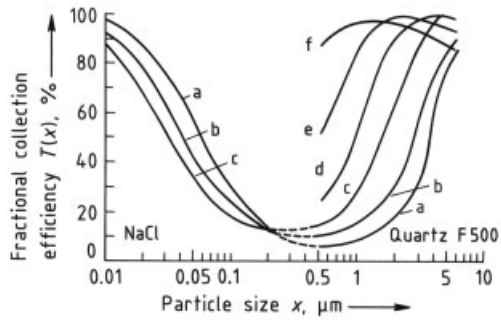

Figure 34. Fractional efficiency curves for a bed of glass beads at various filtration velocities
Bed height H = 20 mm; bead diameter d_G= 0.58 mm Filtration velocity U_0, m/s: a) 0.05; b) 0.10; c) 0.25; d) 0.5; e) 1.0; f) 1.5

5.4.4.2. Pressure Drop

The pressure drop across the granular bed in its initial unloaded state can be calculated fairly accurately. Ergun's equation is used primarily because it holds for all Reynolds numbers and has been tested for granular beds.

$$\Delta p = 150 \frac{(1-\varepsilon)^2}{\varepsilon^3} \cdot \frac{\mu_g \cdot U_0}{d_G^2} H + 1.75 \frac{(1-\varepsilon)}{\varepsilon^3} \cdot \frac{\varrho_g \cdot U_0^2}{d_G} H \quad (38)$$

where U_0 is the superficial velocity and μ_g is the gas viscosity. The marked dependence on porosity ε should be noted. Inaccuracy in the determination of ε strongly affects Δp.

The pressure drop increases with time due to incorporation of dust into the bed. This process has been described by a few equations which apply only to special cases and are not generally valid. Accordingly, the time-dependent behavior of the pressure drop must be determined experimentally.

6. Electrical Precipitators

6.1. General

An especially efficient way of separating particles from gases is based on the forces exerted on charged particles in an electric field. This principle is important when inertial forces are no longer effective, i.e., for fine particles with a diameter of ca. $< 1\,\mu m$. The benefits of such processes have been known and patented for over a hundred years. Large-scale implementation, however, did not start until early in the 20th century, when sufficiently powerful high-voltage equipment was developed. The developmental history, principles, and current status of this field have been described in detail [2], [5], [63–65].

Electrical collectors are often simply called *electrofilters;* the term *electrostatic precipitator* has been adopted even though the processes are by no means electrostatic and sizable electric currents are involved.

Because the required capital investment is rather large, the main application of these precipitators is in dust removal from large gas streams (up to several million cubic meters per hour). These includes flue gases from power plants and refuse incinerators (furnaces), cement plants (rotary kilns and mill drying), the iron and steel industry (blast furnaces and converters), foundries, nonferrous metal refineries (furnaces), expanded-clay aggregate plants, and chemical plants. Collection efficiencies over 99.9 % can be obtained. Emission levels significantly lower than 20 mg/m^3 can be achieved if the dust properties are favorable. Pressure drops are generally very slight (< 500 Pa).

6.2. Mode of Operation

6.2.1. Principles

The collection principle used in electrical precipitators is depicted in Figure 35 and involves three successive steps:

1) charging of particles,
2) transport of particles to the collecting electrode, and
3) removal of particles gathered on the collecting electrode.

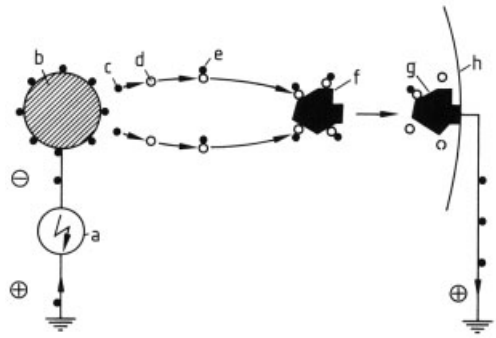

Figure 35. Schematic diagram of charging and collection process
a) High-voltage power supply; b) Emission electrode; c) Electron; d) Neutral molecule; e) Ionized molecule; f) Charged dust particle; g) Collected dust particle; h) Collecting electrode

Most particles carry an electrostatic charge, but it is usually too small for the collection process. Additional charge must, therefore, be supplied. This is accomplished by gas ions that are generated in a corona discharge. Figure 36 shows

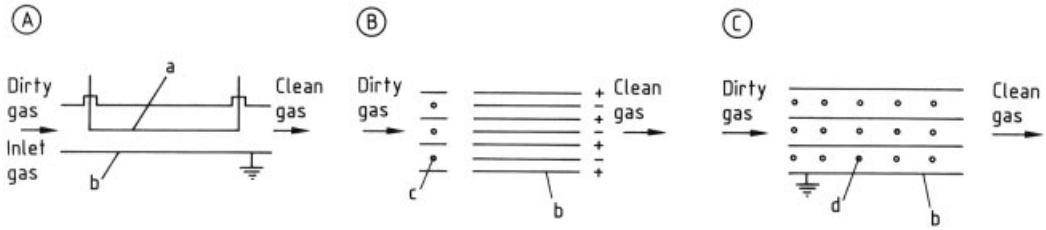

Figure 36. Schematic arrangements of emission and collecting electrodes in electrical precipitators
A) Wire-in-tube precipitator; B) Two-stage parallel-plate precipitator; C) Single-stage parallel-plate precipitator
a) Emission electrode; b) Collecting electrode; c) Positive emission electrode; d) Negative emission electrode

the configurations used for charging and collection: wire-in-tube (A) and parallel-plate (B and C) precipitators.

The parallel-plate types are by far more widely used and may be either two-stage or single-stage devices. In the two-stage unit (Fig. 36B), the charging and collection regions are separate. These systems operate advantageously with a positive emission electrode and are preferably used in air-conditioning applications because a positive corona does not form ozone. However, the currents which can be obtained are not as high as with a negative corona. Higher currents are necessary for high collection efficiency and for operation with high dust concentration. Therefore, for dust removal from industrial exhaust gases, the single-stage negative-corona design is preferred (Fig. 36C).

Typical voltages in industrial filters are 20 – 100 kV, with channel widths of 200 – 500 mm.

Electric charges released at the emission electrode attach themselves to the particles that are to be collected. The charged particles in the electric field between the emission and collecting electrodes are transported to the collecting electrode and retained. Liquid particles flow downward as a film, while solid particles form a dust layer that must be removed from time to time by rapping or flushing with a liquid. The loosened dust falls into a hopper beneath the collecting electrode. Dust removal is done on-line, i.e., the gas stream is not cut off. Consequently, previously collected dust can be reentrained and transported further. The collector length is therefore subdivided into several (three to six) successive zones, which are cleaned at different times in order to minimize dust penetration.

6.2.2. Charging of Particles

Production of Charge Carriers. The ions needed for particle charging are produced by corona discharge at the emission electrodes, which are made of thin wires or tapes with barbs. Triggering the corona discharge requires a minimum voltage, the *corona inception voltage,* which depends on the electrode geometry as well as the gas species and its properties; a typical value is in the range 10 – 30 kV.

If the applied voltage is raised, the current increases until electrical breakdown (sparking) occurs at the *breakdown voltage* (Fig. 37). The breakdown voltage depends on electrode geometry and gas properties. To obtain optimal collection efficiencies, high field strengths and high

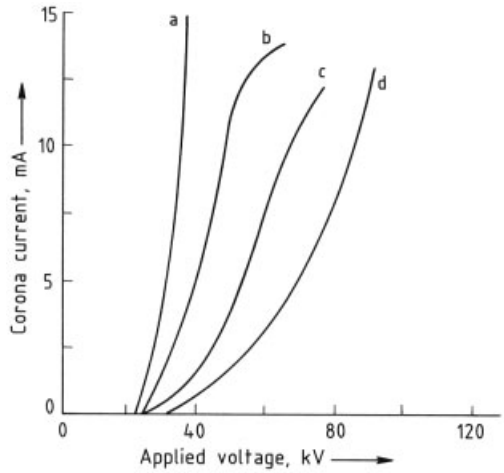

Figure 37. Current–voltage curve for a negative corona in nitrogen, nitrogen–oxygen mixtures, and air (electrical tube precipitator)
a) Nitrogen 100 %; b) Nitrogen 98 %, oxygen 2 %; c) Nitrogen 95 %, oxygen 5 %; d) Air

corona currents are required. The d.c. voltage is, therefore, adjusted to establish a condition as close as possible to electrical breakdown. A trend in the use of pulsed voltage (pulse energization), in which the voltage peaks exceed the breakdown voltage [66].

To allow precipitator operation, the difference between the corona inception voltage and the breakdown voltage must not be too small. As the gas temperature increases, however, this difference becomes smaller and smaller, so that stable operation at atmospheric pressure ($\pm$ 10 kPa) is difficult above ca. 400 °C. However, the difference becomes larger with increasing pressure; this point is especially important in cleaning gases from high-pressure fluidized-bed combustors: at pressures > 1 MPa, operation is possible even at 800 °C.

Attachment of Charge. Depending on particle size, the attachment of gas ions to particles occurs by one of two mechanisms: field charging or diffusion charging.

Field charging dominates in the particle-size range above ca. 1 µm. The ions move toward the collecting electrodes along the electric field lines. As they move, the ions collide with particles and become attached to them. The charging process is time-dependent; 80 % of the maximum charge is acquired after about 0.1 s. The maximum charge is proportional to the surface area of the particle, i.e., to x^2. For a 2-µm particle, this amounts to some 250 elementary charges.

Diffusion charging results from random thermal motion of the ions (Brownian motion) and is important chiefly for particles smaller than 1 µm. This process is slower than field charging: acquisition of 80 % of the maximum charge takes approximately 1 s. In this range, the maximum charge is proportional to the particle size x.

6.2.3. Effect of Dust Resistivity

Charged particles are transported to the collecting electrode and must adhere to it. The electrical properties of the particles, particularly resistance, are important for the behavior of both the particle on impact and the resulting dust layer.

The favorable range of resistivity for collection is $10^4 - 10^{11}\,\Omega\,\mathrm{cm}$. Particles with a lower resistance ($< 10^4\,\Omega\,\mathrm{cm}$) give up their charge quickly when they collide with the collecting electrode; they change their polarity and are repelled from the electrode back into the gas stream. If the resistance is too high ($> 10^{11}\,\Omega\,\mathrm{cm}$), the particles form a layer that continues to accumulate more charge, thus leading to a weakening of the electric field and "reemission." Reemission results from discharge in the porous dust layer (back corona).

Dust resistivity depends on the geometry of the dust layer, particle characteristics, and gas properties (especially the dew point, Fig. 38); thus it should be measured under conditions as close as possible to those used in practice. The resistivity is made up of two contributions, surface resistivity and volume resistivity, which both depend on temperature but in opposite ways. The temperature–resistance curves therefore have a maximum (Fig. 38).

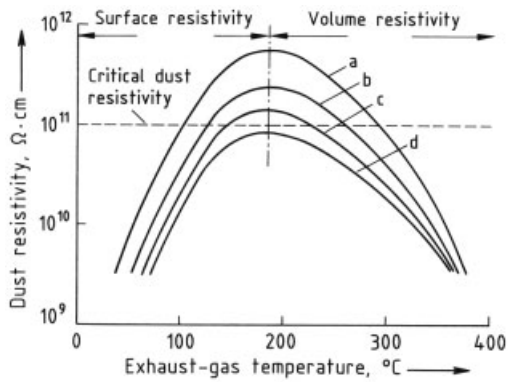

Figure 38. Electrical resistivity of exhaust gas as a function of temperature
Dew point of gas, °C: a) 20; b) 40; c) 50; d) 60

The resistivity of fly-ash particles is affected by their sulfur content. Some coals are so low in sulfur that the resulting fly-ash dust has a high resistance, which makes successful precipitator operation difficult. Consequently, electrical precipitation has been abandoned in power plants fired with these coals and bag filters have been installed. Problems caused by high dust resistivity can be overcome through gas conditioning, e.g., by spraying water into the gas or adding sulfur trioxide in low concentrations. A sulfur

trioxide level of 10 – 15 ppm has proved most suitable; it is completely adsorbed by the fly ash so that the sulfur oxide emission level does not increase [67]. Another way of promoting the collection of high-resistance dusts is the pulsed-voltage technique discussed in Section 6.2.2.

6.3. Basic Designs

Wire-in-tube precipitators (Fig. 39) are employed for small quantities of gas and especially for the collection of liquids (e.g., tar mists and acid mists). The emission electrodes (wires) are stretched along the central axis of parallel tubes. The inside walls of the tubes serve as collecting electrodes. If the tubes are set up vertically, the collected liquid runs down the walls as a film. If solid particles are collected, they are frequently rinsed out by flushing the tubes with water. The tubes are usually 0.2 – 0.3 m in diameter and 2 – 5 m long. They may have a circular (Fig. 39B) or a hexagonal cross section (Fig. 39C). Since there is a conductive liquid film om the tubes, plastics such as PVC can be used instead of steel. This minimizes corrosion problems.

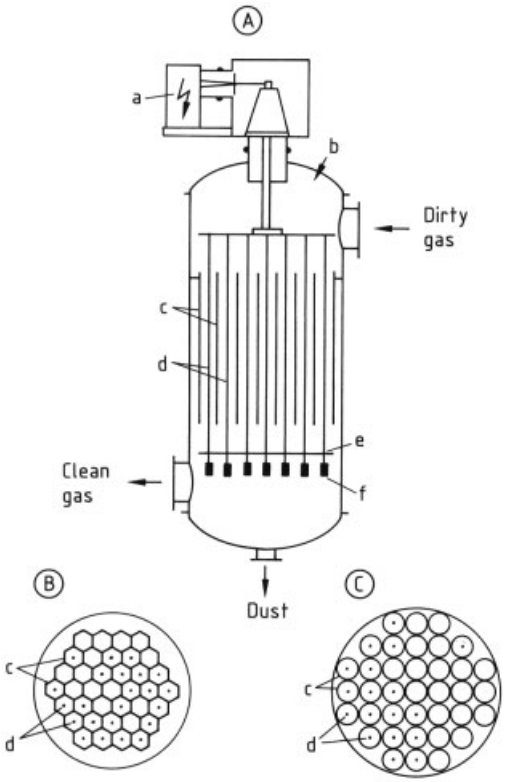

Figure 39. Tube precipitator consisting of individual collectors in parallel
A) Vertical section; B) Tubes with a hexagonal cross section (honeycomb precipitator); C) Tubes with a circular cross section
a) High-voltage power supply; b) Flushing; c) Tube walls (collectors); d) Emission electrodes; e) Electrode spacer; f) Weight

Parallel-plate precipitators contain many plane or profiled plates that are suspended at uniform intervals (200 – 500 mm) as collecting electrodes (Fig. 40, see next page). Gas flows horizontally in the channels between plates. Emission electrodes hang in the midplane between neighboring plates. Particle removal may be either dry or wet.

In *dry electrostatic precipitators,* the dust layer collected on the plates is removed periodically by rapping the plates with hammers; the dust falls into the hopper beneath the plates. The total plate length is divided into several (three to six) zones for two purposes: (1) dust reentrainment is reduced because the zones can be cleaned at different times; and (2) separate zone-by-zone power controls allow the voltage to be adjusted in accordance with dust concentration, which declines exponentially along the precipitator.

In *wet electrostatic precipitators,* the collecting electrodes are cleaned by flushing with water. The dust flows down off the plates with the water; reentrainment of the dust is thus effectively prevented.

If the dirty gas has a very high dust content, the corona or the electric field strength is weakened, i.e., collection performance deteriorates. Precollectors are then used to lower dust concentration. However, these should not significantly alter the particle-size distribution because dust removal from the plates is better with a wider distribution than with fine particles only. Simple deflecting vanes have proved suitable for this purpose [68].

If dust concentration in the clean gas must be particularly low, combinations of several types of dust collectors are employed: flue gases from blast furnaces can be cleaned up to 5 – 10 mg/m^3 by a series consisting of a cyclone, a scrubber, and a wet electrostatic precipitator.

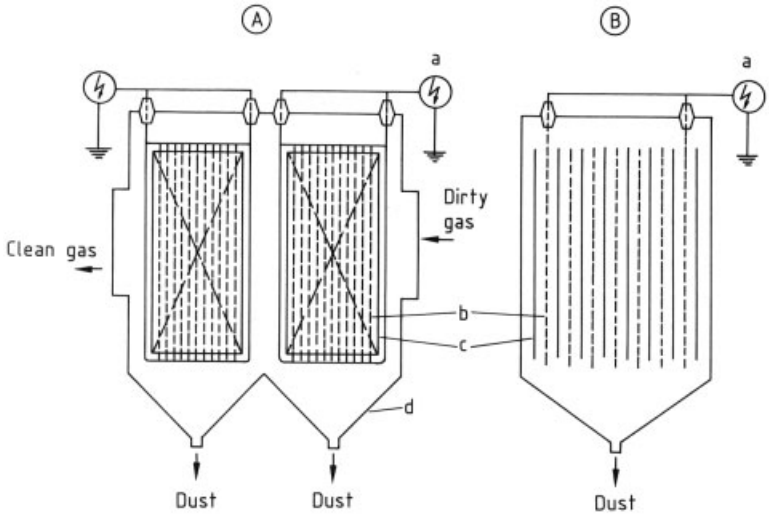

Figure 40. Schematic of a two-zone parallel-plate electrical precipitator
Vertical sections A and B are shown at right-angles to each other
a) High-voltage power supply; b) Emission electrode; c) Collecting electrode; d) Hopper

6.4. Design Calculations

Calculations for precipitator design are concerned solely with collection efficiency; the very small pressure drop is not important.

In 1922, DEUTSCH derived the precipitator equation that bears his name. The equation has been challenged, but it is still used as the basis for sizing precipitators and for evaluating and interpreting measurements. If turbulent flow in the precipitator is assumed (i.e., the distributions of gas velocity and particle concentration transverse to the direction of flow are uniform) and if reentrainment and reemission are neglected, the Deutsch equation for fractional efficiency is

$$T(x) = 1 - \exp\left[-\frac{2 \cdot L}{a \cdot v} w(x)\right] \tag{39a}$$

where L is the collection length, a is the plate spacing (parallel plate precipitators) or the radius of the tube (tube precipitators), v is the gas velocity, and $w(x)$ is the so-called migration velocity. Equation (39 a) can also be written as

$$T(x) = 1 - \exp\left[-\frac{A}{\dot{V}} w(x)\right] \tag{39b}$$

where A is the surface area of the collecting electrodes and $\dot{V}$ is the volume gas flow rate. The ratio $A/\dot{V}$ is also called the specific collecting area.

The migration velocity $w(x)$ depends on the particle size and charge, the electric field, and the gas viscosity. It can be derived theoretically by solving the equation of particle motion [5]. In practice, migration velocity is calculated from the measured fractional efficiencies, by means of Equation (39); it often lies in the range of 10 – 30 cm/s. The parameter $w(x)$ is basically a transport coefficient that characterizes particle transport to the collecting electrode; it is thus vital for sizing electrical precipitators. Deutsch's assumption that $w(x)$ is constant along the precipitator has not been confirmed experimentally [69]. The mean value of w (the effective migration velocity), recalculated with Equation (39) from measured fractional efficiencies, increases with gas velocity and channel width [70]. The physical significance of these phenomena and their incorporation in the design equation are not yet completely understood because particle behavior inside the precipitator is still not clear. For this reason, the search for an optimal channel width has led to differing results in the range of 40 – 60 cm.

Nevertheless, Deutsch's equation continues to be of central importance in design. A variety of correction factors in the exponent of Equation (39) have been suggested [71]. One problem in interpreting experimental studies is that measurements usually yield overall collection

efficiencies, not fractional efficiency curves. The resulting w values thus represent integral mean values, which depend on particle-size distribution. As the distribution in the dirty gas becomes broader, the overall efficiency for a given precipitator declines. As one of the few fractional measurements shows, the shape of the fractional efficiency curve is rather complicated but also informative (Fig. 41) [72]. The ascending part of the curve, to the left of the minimum at ca. 0.4 µm, can be accounted for by the dependence on flow resistance of particle motion relative to the gas (Cunningham correction). The decline of the curve past ca. 3 µm has been explained by the typical behavior of incompletely combusted coal particles, which are repelled from the collecting electrode because of their low resistivity (cf. Section 6.2.3).

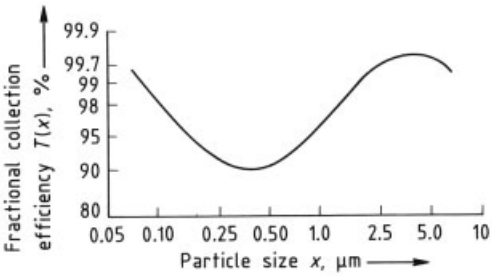

Figure 41. Experimentally determined fractional efficiency curve for an electrical precipitator (specific collecting area = 70 $m^2\,s\,m^{-3}$) [72]

Improved particle measuring instruments have been developed recently; their use in determining particle motion and fractional efficiency curves for electrostatic precipitators should contribute to the understanding of this and other processes [73].

7. References

General References

1. F. Baum: *Luftreinhaltung in der Praxis,* Oldenbourg Verlag, München 1988.
2. H. Brauer (ed.): *Handbuch des Umweltschutzes und der Umweltschutztechnik,* **vol. 3,** Springer Verlag, Heidelberg 1996.
3. F. A. L. Dullien: *Introduction to Industrial Gas Cleaning,* Academic Press, San Diego 1989.
4. W. Licht: *Air Pollution Control Engineering,* 2nd ed., Marcel Dekker, New York 1988.
5. F. Löffler: *Staubabscheiden,* Thieme Verlag, Stuttgart 1988.
6. A. Ogawa: *Separation of Particles from Air and Gases,* **vol. 2,** CRC Press, Boca Raton, FL 1984.
7. E. Schmidt et al. (ed.): *High Temperature Gas Cleaning,* Institut für Mechanische Verfahrenstechnik und Mechanik, Karlsruhe 1996.
8. J. Seville: *Gas Cleaning in Demanding Applications,* Blackie, London 1997.

Specific References

9. Technische Anleitung zur Reinhaltung der Luft (TA Luft) 1986, Allgemeine Verwaltungsschrift zum Bundesimmissionsschutzgesetz.
10. DIN 66142, Darstellung und Kennzeichnung von Trennungen disperser Güter.
11. R. H. Perry, O. Green (eds.): *Perry's Chemical Engineers' Handbook,* 6th ed., McGraw-Hill, New York 1984.
12. VDI-Richtlinie 3676, Massenkraftabscheider, Beuth Verlag, Berlin 1999.
13. M. Bohnet, *Chem. Ing.-Tech.* **54** (1982) 621 – 630.
14. F. Boysan, W. H. Ayers, J. Swithenbank, *Trans. Inst. Chem. Eng.* **60** (1982) 222 – 232.
15. R. Kirch: *Der Einfluß der Turbulenz auf die Partikelbewegung im Gaszyklon,* VDI Verlag, Düsseldorf 1988.
16. E. Muschelknautz, V. Greif, M. Trefz: "Zyklone zur Abscheidung von Feststoffen aus Gasen," *VDI-Wärmeatlas,* 7th ed., VDI Verlag, Düsseldorf 1994.
17. D. Leith, W. Licht, *AIChE Symp. Ser.* **68** (1972) 196 – 206.
18. W. Barth, *Brennst. Wärme Kraft* **8** (1956) 1 – 9.
19. E. Muschelknautz, *Chem. Ing. Tech.* **44** (1972) 63 – 71.
20. P. W. Dietz, *Powder Technol.* **31** (1982) 221 – 226.
21. H. Mothes, F. Löffler, *Chem. Eng. Process* **18** (1984) 323 – 331.
22. T. Lorenz: *Heißgasentstaubung mit Zyklonen,* VDI-Verlag, Düsseldorf 1994.
23. P. Meißner, F. Löffler, VDI Ber. **294** (1978) 69 – 75.
24. W. Rentschler: *Abscheidung und Druckverlust des Gaszyklons in Abhängigkeit von der Staubbeladung,* VDI-Verlag, Düsseldorf 1991.
25. E. Schmidt, C. Wadenpohl, F. Löffler, *Chem. Ing. Tech.* **64** (1992) 76 – 78.
26. H. Rumpf, K. Borho, H. Reichert, *Chem. Ing. Tech.* **40** (1968) 1072 – 1082.

27. M. Bohnet, *Chem. Ing. Tech.* **56** (1984) 416 – 417.
28. W. Krambrock, *Chem. Ing. Tech.* **51** (1979) 493 – 496.
29. VDI-Richtlinie 3679, Naßabscheider für partikelförmige Stoffe, Beuth Verlag, Berlin 1998.
30. K. Holzer, *Staub Reinh. Luft* **34** (1974) 361 – 365.
31. K. Holzer, *Chem. Ing. Tech.* **51** (1979) 200 – 207.
32. M. Wicke: *Aufbau, Leistung und Betriebsverhalten von Naßentstaubern,* VDI-Verlag, Düsseldorf 1970.
33. F. Ebert, H. Büttner, Proc. 1. World Congr. PARTEC, Nürnberg (1986).
34. F. Löffler, G. Schuch, *Filtr. Sep.* **18** (1981) 70 – 74.
35. W. Barth, *Staub Reinh. Luft* **19** (1959) 175 – 180.
36. S. Calvert, *J. A. P. C. A.* **24** (1974) 929 – 934.
37. G. Schuch, F. Löffler, *Verfahrenstechnik Mainz* **12** (1978) 302 – 306.
38. M. Schmidt, Dissertation Universität Karlsruhe, 1993.
39. C.-D. Schegk, F. Löffler, *Chem. Ing. Tech.* **59** (1987) 319 – 322.
40. S. Ripperger, G. Dau, *VT Verfahrenstechnik* **14** (1980) 164 – 168.
41. K. T. Semrau, *Chem. Eng.* **26** (1977) 87 – 91.
42. A. Bürkholz: *Droplet Separation,* VCH Verlagsgesellschaft, Weinheim 1989.
43. R. C. Brown: *Air Filtration,* Pergamon Press, Oxford 1993.
44. R. P. Donovan: *Fabric Filtration for Combustion Sources,* Dekker, New York 1985.
45. E. Schmidt: *Abscheidung von Partikeln aus Gasen mit Oberflächenfiltern,* VDI-Verlag, Düsseldorf 1998.
46. DIN 24 184, Typprüfung von Schwebstoffiltern, 1990.
47. C. N. Davies: *Air Filtration,* Academic Press, London 1973.
48. R. Hiller: *Der Einfluß von Partikelstoß und Partikelhaftung auf die Abscheidung in Faserfiltern,* VDI-Verlag, Düsseldorf 1981.
49. H. Baumgartner Elektrofaserschichten für die Aerosolfiltration, VDI-Verlag, Düsseldorf 1987.
50. H. Baumgartner, F. Löffler, H. Umhauer, *JEEE Trans. Electr. Ins.* **21** (1986) no. 3, 477 – 486.
51. R. Maus: *Verhalten von Bioaerosolen bei der Abscheidung in Tiefenfiltern,* Shaker Verlag, Aachen 1997.
52. H. Jodeit: *Untersuchungen zur Partikelabscheidung in technischen Tiefenfiltern,* VDI-Verlag, Düsseldorf 1985.
53. E. Schweers: *Einfluß der Filterstruktur auf das Filtrationsverhalten von Tiefenfiltern,* VDI-Verlag, Düsseldorf 1993.
54. F. Löffler, H. Dietrich, W. Flatt: *Dust Collection with Bag Filters and Envelope Filters,* Vieweg, Braunschweig 1988.
55. VDI-Richtlinie 3677, Filternde Abscheider, Beuth-Verlag, Berlin 1997.
56. DIN 71 459, Filtereinsätze für Trockenluftfilter, 1973.
57. E. Schmidt, B. Weiß, *Gefahrstoffe-Reinhalt. Luft* **58** (1998) 35 – 40.
58. E. Schmidt, T. Pilz, *Filtr. Sep.* **33** (1996) 409 – 415.
59. G. L. Kinchett, *Filtr. Sep.* **22** (1985) 378 – 380.
60. G. J. Tardos, N. Abuaf, C. Gutfinger, *J. Air Pollut. Control Assoc.* **28** (1978) 354 – 363.
61. S. L. Goren, T. D'Ottavio, *Aerosol Sci. Technol.* **2** (1983) 91 – 108.
62. W. Peukert, F. Löffler, Proceedings of International Conference on Advanced Coal Power Plant Technology, Düsseldorf, 2 – 4 Dec. 1987.
63. J. Böhm: *Electrostatic Precipitation,* Elsevier, Amsterdam 1982.
64. VDI-Richtlinie 3678, Elektrische Abscheider, Beuth-Verlag, Berlin 1998.
65. K. R. Parker: *Applied Electrostatic Precipitation,* Blackie, London 1997.
66. K. Darby, *2nd Int. Conf. Electr. Precipitation,* Kyoto 1984, pp. 575 – 584.
67. G. Mayer-Schwinning, R. Rennhack, *Chem. Ing. Tech.* **52** (1980) 375 – 383.
68. K. Arras, *Zem. Kalk Gips,* **29** (1976) 241 – 247.
69. K. Kinkelin, *Fortschr. Ber. VDI Z. Reihe 15,* **24** (1982).
70. H. Wiggers, *Staub Reinhalt. Luft* **42** (1982) 292 – 294.
71. G. Cooperman, *Atmos. Environ.* **14** (1984) 277 – 285.
72. S. Maartmann, *Staub Reinhalt. Luft* **34** (1974) 353 – 355.
73. H.-J. Schmid: *Zum Partikeltransport in Elektrischen Abscheidern,* Shaker Verlag, Aachen 1999.

Dye Carriers → Textile Auxiliaries

Solid – Solid Separation, Introduction

For unit operations in solid – solid separation see → Screening; → Elutriation; → Air Classifying; → Mineral Sorting; → Magnetic Separation; → Electrostatic Separation; → Gravity Concentration; → Dense-Medium Separation; → Flotation.

Henry E. Cohen, Imperial College of Science and Technology, Royal School of Mines, London, United Kingdom

1. Introduction 1
2. Preparatory Processes 2
2.1. Comminution 2
2.2. Size Separation 3
2.2.1. Screening 3
2.2.2. Classification 4
3. Physical Separation Processes 4
3.1. Radiation Sorting 4
3.2. Gravity Separation 5
3.3. Magnetic Separation 5
3.4. Electrostatic Separation 6
3.5. Separation Based on Surface Phenomena 6
4. Process Design Economics 7
5. References 7

1. Introduction

The need to separate solids occurs in many different industrial activities and for a variety of materials. Examples include the separation of food grains such as nuts, beans, peas, or rice from contaminants or waste such as shells, husks, stones, or vermin. Other examples are the removal of print from recycled paper, the separation of granulated plastics and copper in recycling electric cables, or the recovery of precious-metal contacts from scrapped electronic equipment.

By far the largest application of solid – solid separation, in both tonnage of throughput and value of product is in the utilization of mineral resources. Physical methods of separation are essential for the successful recovery of metals, industrial minerals, and fuels (the values) from their ores. "Ores" are defined as any kind of rock or mineral deposit from which marketable products can be extracted profitably. To meet subsequent processing and user requirements, the desired values need to be separated from unwanted or harmful ore constituents (waste or tailings) by methods collectively known as mineral processing.

Solid – solid separation relies on differences in physical and chemical properties of the constituents of an ore to devise engineering solutions for separating wanted values from the rest of the ore. In industrial practice, no separation process can attain 100 % recovery of the values in a feed within acceptable processing costs. Hence it is prudent to restrict the number of unit processes in a treatment scheme so as to achieve an optimized balance between costs of processing, recovery of values, grade of the product, transport charges and attainable product price. In most cases, a trade-off is necessary between product grade and percentage recovery of values, because an increase in one generally entails a decline in the other. Total literation may be possible, but that would result in losses of values due to excessive comminution. Similarly, 100 % recovery of values may be possible with less comminution, but the grade of the product would be low due to the presence of unliberated particles. Thus, the choice of processes of preparation and separation is necessarily a compromise conditioned by financial constraints.

For attaining best separation, the differences in physical and chemical properties of values and tailings need to be displayed and enhanced as much as possible by appropriate preparatory treatment. Physical liberation from each other is essential for effective separation of the constituents of an ore. The product quality (grade) and the proportion of the values extracted from an ore (recovery) depend on the degree of liber-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_14

ation and the enhancement of differential properties achieved during preparation.

Broadly, all ores can be divided into two categories:

1) Ores that contain high-priced values (e.g., gold or other precious metals, nickel, tin, diamonds, emeralds, etc.) in low concentrations. A typical gold ore contains a few grams of gold per tonne of rock.
2) Ores that contain low-priced values (e.g., iron, manganese, chromium, gypsum, china clay, etc.) in relatively high concentrations. A typical iron ore contains 60 % iron, or nearly two-thirds of a tonne of iron per tonne of rock.

These two categories influence selection, application, and performance of methods of solid–solid separation. The highest possible percentage recovery (better than 90 %) of a small content of high-priced values is essential (e.g., with gold ores) to cover the cost of treatment. By contrast, modest recoveries (60 – 70 %) can be tolerated with low-priced values, but the product grade must be high enough to qualify for good market prices. For china clay, practically 100 % purity (i.e., 100 % grade) is necessary for its acceptance as a paper-coating material and relatively low recovery has to be accepted.

2. Preparatory Processes

Essential preparatory processes include *comminution* for solid rocks (or other heterogeneous solids), *dry* or *wet scrubbing* for unconsolidated materials such as sand or clay deposits, and *size separation processes* for mixed populations of particles.

Although preparation processes necessarily precede solid–solid separation, complex alternating sequences of preparation and separation are commonly employed. Such sequences, termed stage processing, are advantageous if the values show some inhomogeneous liberation behavior or if different values in the ore require separate recovery after dissimilar preparation.

For example, the tin mineral cassiterite (SnO_2) may occur in granite in natural crystal sizes from 10 mm down to 0.05 mm. During comminution to liberate the cassiterite from waste minerals such as silicates, iron oxides, etc., breakage to < 5 mm may produce substantial proportions of coarse, liberated cassiterite particles. Having to break such particles further during comminution to liberate smaller crystals of cassiterite would be wasteful in terms of processing cost and possible loss of values. Hence it is good industrial practice to interrupt comminution and first separating a coarse concentrate of liberated cassiterite. Further comminution of the remaining ore, followed by separation of a product (or of a waste reject) can be repeated several times at progressively smaller size ranges.

This is practiced with ores where locked middlings persist over wide size ranges. Repetition ends when the remaining values are insufficient to justify the costs of further treatment. Complex sequences of preparation and separation may also be necessary for ores that contain several different value minerals. For example, ores containing sulfides of copper, lead, and zinc require repeated and different surface conditioning for selective recovery of these minerals into separate concentrates.

2.1. Comminution

Comminution serves two related, but separate purposes: to liberate minerals from each other and to render the resulting particles within defined size ranges suitable for subsequent separation, or to meet user requirements. For both purposes, comminution usually needs to be closely linked with size separation. In industrial practice, comminution–also called size reduction–is usually divided into crushing and grinding (→ Size Reduction). *Crushing* employs processes of compression or impact shattering and is applied to coarse particles, from meters down to centimeters in size. Power consumption is relatively low, up to ca. 3 kW · h per tonne of throughput. *Grinding* employs processes of abrasion, with or without impact, and is applied to fine particles, from centimeters down to submicrometers in size. Power consumption is relatively high, commonly in the range of 10 – 20 kW · h and up to 200 kW · h per tonne for very fine products. Grinding is usually the most energy-intensive process in any sequence of treatment and can account for more than 70 %

of the total energy consumed (hence of the total costs incurred).

This presents one of the reasons for stage processing. If substantial proportions of waste material can be liberated from values at relatively coarse particle sizes, separating and discarding such waste ahead of the expensive grinding stage may be economically advantageous.

For example, in the treatment of iron ores in Minnesota, some 40 % of the feed tonnage was removed as coarse waste before grinding. Magnetic separators were used to produce a clean nonmagnetic waste and a magnetic "middling" which passed to the grinding stage. About 10 % of the total iron in the feed was lost with the coarse waste, but this loss was far outweighed by the savings in grinding costs.

The use of repeated comminution combined with size separation is common, because many separation processes work best within specific ranges of particle sizes. For example, recovery in froth flotation usually declines sharply for particles coarser than 0.2 mm or finer than 0.02 mm. Hence preparation needs to be arranged so as to achieve liberation of the greatest possible proportion of the values in a flotation feed within these size limits.

2.2. Size Separation

Size separation is necessary for various aspects of preparation. For efficient handling, for good separation, and subsequent use or disposal of products, ores are generally divided into three size groups: oversize, product size (one or several), and undersize. *Oversize material* is too coarse for an intended process of separation. It may contain values that need further comminution because they are insufficiently liberated. There may be coarse free values that should be extracted by some other suitable method, or coarse gangue low enough in value content not to merit costs of further treatment. User requirements may also limit oversize contents. For example, china clay for coating paper needs to be at least 80 % finer than 2 µm. *Product sizes* are those most suitable for a separation process; they may also be defined according to size specifications imposed by users. For example, commercial specifications for iron ores usually include upper and lower size limits. The term *undersize* refers to particles that are too small for a given process of separation or for commercial acceptance of a product. Particle-size analysis is treated elsewhere in this volume.

All three terms–oversize, product size, and undersize–are relative and can cover widely different size ranges for different materials and purposes. For example, in treating gold ores, an upper product size limit of about 0.1 mm is usually necessary to ensure adequate liberation. Particles coarser than 0.1 mm are oversize and require further grinding. In contrast, when jigs are used for gravity concentration of tin ores, particles smaller than around 0.2 mm are undersize. The jig would be inefficient in treating -0.2 mm particles (i.e., particles < 0.2 mm), therefore these are treated separately by other means.

The two dominant industrial processes for performing size separation are *screening* and *classification.* They can be used for wet particles carried in water (or other fluids) or for dry particle populations that are free-flowing.

2.2.1. Screening

In screening (or sieving), each individual in a particulate material is presented to apertures which may be circular, square, or rectangular and which act as two-dimensional gauges (→ Screening). Ability to pass through depends on size but is also affected by particle shape which may be platy, nearly equidimensional, needle-like, or irregular. Thus, with increasing aspect ratios, the orientation of a particle relative to the aperture becomes an important factor which blurs the definition of particle size. For example, consider a square screen aperture of 10×10 mm. With favorable presentation, this aperture would permit passage of a 9-mm square disk, a 9-mm cube, or a rectangular block 9 mm square and 90 mm (or more) in length. Screening assigns particles roughly into size groups, each group having passed through a larger aperture and having failed to pass through the next smaller aperture used in a given process.

The difficulty of passage increases as particle dimensions approach aperture size, and this can lead to blockage. Therefore, good industrial practice is to avoid aperture sizes that are near the dominant particle size in a given product.

Screening is used generally for coarse particles, from meters down to about 1 mm and even smaller sizes occasionally (e.g., 0.02 mm for china clay). Under laboratory conditions, woven wire sievescan be used down to 0.015 mm for some minerals. Fragile electroformed perforated sheets and molecular sieves can be employed down to submicrometer sizes, but then throughput is very slow.

2.2.2. Classification

Classification is commonly employed in industry for size separation of particles below about 1 mm. The process relies on different terminal velocities (different settling rates) for particles of different sizes moving through fluids (usually water; → Elutriation) or through gases (usually air; → Air Classification). Particles accelerating under the influence of gravity (or some other force) encounter increasing resistance from the enveloping medium. Terminal velocities are reached when the forces of acceleration and resistance are equal. Larger particles then move away from smaller ones, and separate size groups can be collected.

However, differences in specific gravity must yield different terminal velocities. Hence, each classified size group in heterogeneous populations of particles includes larger particles of lower specific gravity and smaller particles of higher specific gravity. This distinguishes a classified size fraction from a screened product, which contains the same size range for all particles, irrespective of specific gravities. Classified feed is preferred for some separation processes, e.g., gravity separation on shaking tables.

Classification used for the recirculation of oversize material in grinding circuits can cause undesirable overgrinding of value minerals. If the latter have higher specific gravities than the associated gangue minerals, they may undergo selective recirculation, with further grinding, even though they are fine enough for processing.

3. Physical Separation Processes

Physical processes of solid – solid separation utilize differences in the physical properties of mixed solids and may be grouped broadly as

1) Radiation sorting,
2) Gravity separation,
3) Magnetic separation,
4) Electrostatic separation,
5) Separation based on surface phenomena.

Physical separation methods normally produce concentrates of unaltered minerals that have retained their natural characteristics and crystal structures.

Another important group of processes is based on chemical reactivity, e.g., acid leaching, solvent extraction, or other methods of selective dissolution that entail the partial or total destruction of ore constituents. These processes are termed chemical separation or hydrometallurgy (→ Liquid – Liquid Extraction; → Liquid–Solid Extraction).

3.1. Radiation Sorting

Radiation sorting depends on differential responses to externally applied radiations. Some processes utilize differential radiation emissions, with or without externally applied stimuli. Many types of radiation can be used for solid–solid separation, including visible, ultraviolet, and infrared light; X-rays, gamma-rays, and electron beams; ultrasound; and a range of radio frequencies. The methods of separation rely on differential effects of transmission, absorption, reflection,modulation, or emission For example, reflection of visible light can be used to separate particles by color. Similarly, transmission of light can be used to separate opaque from transparent species. Some particles in mixed populations emit radiation selectively under ultraviolet illumination, e.g., the tungsten mineral scheelite fluoresces bright blue and the mineral calcite fluoresces bright pink under ultraviolet illumination. The selective emissions from radioactive substances are another example.

The necessary process mechanisms include a sequence of presentation, inspection, and sorting (→ Mineral Sorting). Each particle, in free flight or traveling on a conveyor, must be presented individually to one or several radiation sources. The particle is inspected, and the resultant information is analyzed to trigger an accept – reject

signal which operates a sorting mechanism (e.g., a moving splitter, an air or water jet, or a trap). The size range of particles suitable for such sorting extends from about 150 mm down to about 2 mm, but individual feeds must be sized relatively closely. Throughput rates are proportional to particle size, and separation is most efficient if the ejected material is a minority constituent.

3.2. Gravity Separation

Processes of gravity separation utilize differences in specific gravity and are divided into two groups::

1) Sink – float separation and
2) Differential acceleration.

Sink – float separation, also called *dense medium separation* or *heavy medium separation,* employs heavy liquids or suspensions, which have effective densities intermediate between the specific gravities of species in a mixed population (→ Dense Medium Separation). For example, quartz (*d* 2.65) can be separated from hematite (*d* 5.1) by using a heavy liquid such as tetrabromoethane which has a specific gravity of 2.95. If a mixture of quartz and hematite is immersed in tetrabromoethane, the quartz floats on the liquid and the hematite sinks to the bottom. Similarly, the heavy liquid may be replaced by a suspension of finely pulverized heavy minerals in water. Suspensions with effective relative densities up to about 3.5 are achievable with suspensions of barites, galena, magnetite, or ferrosilicon. For the separation of coal from shale, quartz sand suspensions with effective densities in the range of 1.4 – 1.6 are used. Heavy suspensions are much more economical for industrial use than organic heavy liquids, such as tetrabromoethane, which are employed mainly in laboratory work. In industrial use, the heavy suspensions are usually recovered, cleaned, and recycled to the separation process.

The second major group of gravity separation processes relies on *gravitational acceleration* to exploit differences in specific gravity; a difference of not less than 1 is normally required for efficient separation. Gravitational acceleration causes particles of higher specific gravity to advance faster from rest than particles of lower specific gravity, irrespective of their sizes. Path lengths must be short so that the particles cannot reach terminal velocities (i.e., zero acceleration). Thus, gravity separation, which employs differential acceleration over short path lengths, differs from size classification, which depends on differential terminal velocities (→ Gravity Concentration).

Various separator designs employ films of flowing water or oscillatory motion to produce repeated acceleration over short path lengths. The flow velocities and amplitudes of oscillation must match the dynamic characteristics of the particles; hence, size-classified feeds are usually necessary. Throughput capacities range from hundreds of tonnes per hour (e.g., jigs for coarse feed) to less than 0.5 t/h (e.g., shaking tables for very fine feeds).

3.3. Magnetic Separation

Most natural minerals are more or less feebly paramagnetic and a few are very weakly diamagnetic. Magnetic separation processes can separate species of relatively higher magnetic susceptibility from others of lower susceptibility. The strength of the available magnetic force limits the applicability of a separator, and the difference in susceptibility determines the possible efficiency of separating a given mixed feed. Magnetic separators employ various methods of presenting feed to convergent magnetic fields which provide directional tractive forces. Counterforces, such as fluid drag, gravitational acceleration, and others, are arranged singly or jointly to act in a divergent direction from the magnetic forces. The balance of forces is arranged so that the more susceptible particles (the "magnetics") follow the magnetic force and the less susceptible particles (the "non-magnetics") follow the counterforces (→ Magnetic Separation).

Magnetic separation can be carried out dry or in fluid suspension. The magnetic force experienced by a particle is mass related; hence, no theoretical upper or lower size limit exists. However, problems of physical handling usually impose an upper limit of about 150 mm. Electrostatic charges in dry feed and fluid drag in wet feed generally impose lower size limits around 0.01 – 0.05 mm. Throughput rates depend on particle size, but for different types

of separators they can range from hundreds of tonnes to a few kilograms per hour.

3.4. Electrostatic Separation

Electrostatic separation is based on differences in surface electrical properties, particularly surface conductivity, and the associated ability to acquire, retain, or lose charges. Dry particles can acquire charges by induction, convection, or conduction. If different species in a mixture are charged selectively, they can be separated by application of external electrical forces (→ Electrostatic Separation).

For example, conduction charging can arise through spontaneous migration of electrons in a heterogeneous population of pulverized particles. Electrons will pass from a poorer to a better conductor; thus, the latter acquires a negative charge and the former is left with a net positive charge. If this population passes between a pair of electrode plates, one positively and one negatively charged, each particle will experience repulsion from the plate of like charge and attraction from the plate of opposite charge. The geometry, level of charge, and speed of passage can be arranged so that the particles separate into two streams.

In another example, a mixed particle population of conductors and nonconductors can be charged by means of a corona discharge from a wire electrode. The nonconductors retain the charge, but the conductors lose their charge during passage over a rotating drum. A second charged electrode can then "pin" the charged particles to the drum and "lift" the discharged particles. A typical application is the separation of zircon (a nonconductor) from rutile (a conductor) in the treatment of beach sand.

Various separator designs utilize different combinations of charging and discharging to achieve discrete products. Full liberation, total absence of surface moisture and a restricted size range of feed are usually important for attaining good separation. Several repeated treatments are normally needed to attain acceptable clean products. Throughput rates range up to several tonnes per hour.

3.5. Separation Based on Surface Phenomena

Surface chemical properties of solids can be manipulated selectively in various ways to generate differential behavior of the different constituents in a mixed population of particles. For example, traces of reagents can be selectively adsorbed on species, rendering them either hydrophobic or hydrophilic. The hydrophobic particles can be made to coagulate and to settle from a suspension. This is known as *selective flocculation* (→ Flocculants, Chap. 3.; see also → Sedimentation).

More important, hydrophobic particles can be attached to air bubbles which are made to stream through a pulp suspension. The air bubbles transport the hydrophobic particles to the surface of the pulp, forming a scum or froth which can be collected as a concentrate. This process, known as *froth flotation*, is one of the most important and widespread methods of solid–solid separation in the mineral industry (→ Flotation). A variety of reagents have been developed for specifically enhancing the selectivity of flotation. The use of reagents usually necessitates stringent control of pH and of dissolved ions in the water.

With different reagent regimes, froth flotation has such diverse applications as the recovery of sulfides of copper, lead, zinc, nickel, molybdenum, antimony, and other nonferrous metals; oxides of iron, titanium, tungsten, tin, etc.; calcium fluoride, phosphates, various carbonates, silicates, and many other mineral compounds. Flotation is an extremely versatile process, limited only by the development of selective reagents for specific applications. It operates most efficiently on particles in the range of 0.2 – 0.02 mm, but coarser and finer sizes can be treated successfully under favorable conditions.

The throughput capacity of individual flotation machines varies, according to unit size, from kilograms to tonnes per hour. Multiple units in very large flotation plants, achieve throughputs in the range of 30 000 – 50 000 t/d. The process route is usually arranged so that a first flotation, the rougher stage, is followed by one or more cleaner stages for the float product. The cleaners reject misplaced sink material. Similarly, one or more scavenger stages for the sink residue can rescue misplaced float material. These stages

may be interspersed with regrinding of some products, so as to achieve higher product grades through improved liberation. Reagent additions can be made at various stages, so as to improve selectivity.

Another application of surface chemistry, *oil agglomeration,* employs various reagents for specific conditioning of mineral species so that they are selectively taken into an organic fluid that is immiscible with water.

4. Process Design Economics

Under normal economic conditions, the cost of processing must be less than the resultant added value of the product. Processing costs are minimized when the simplest possible process route is used to attain the best possible recovery, at the lowest product grade acceptable to the customer. Variations in the three basic cost components–capital, energy, and labor–can influence process choices in relation to specific local conditions. The resultant costs for any separation route consist of only two elements:

1) processing costs, which are definable as capital and operating charges, and
2) loss of values due to inevitable process inefficiencies. The recoverable values are always lower than 100 % of the values contained in the feed.

Both cost elements are related to the balance between grade and recovery, as well as to the complexity of the processing route.

For high-priced values which are minority constituents of the feed, high recovery is more important than grade; hence, the process design should aim at the earliest possible removal of concentrates. This points to treatment in stages so that values can be extracted as soon as they are liberated in any significant quantity. Repeated further preparation of the remaining bulk of the material and separation of values should continue until the residual values cannot justify the cost of further processing. Thus, values are removed in stages, until the remaining waste is of sufficiently low grade for disposal.

Low-priced values which are major constituents of a feed must be worked up to high-grade specifications, but high recovery can be sacrificed. To keep processing costs as low as possible, the process design should aim at early gangue removal, especially to avoid fine grinding. This requires stage separation of liberated gangue and retention of values for further processing to attain the desired grade. Thus, waste is removed in stages until the product reaches market specification.

Detailed factors influencing the selection of appropriate processes of solid – solid separation are discussed in the other articles to which reference is made here.

5. References

1. N. L. Weiss (ed.): *SME Mineral Processing Handbook,* Society of Mining Engineers, AIMMPF, New York 1985, pp. 7 – 5 to 7 – 29.
2. B. A. Wills: *Mineral Processing Handbook,* 3rd ed., Pergamon Press, Oxford 1985, pp. 522 – 530.
3. G. Tarjan: *Mineral Processing,* Akademiai Klado, Budapest 1981, pp. 419 – 423.
4. P. Somasundranan (ed.): *Advances in Mineral Processing,* Society of Mining Engineers, Boulder, Col., 1986.
5. L. White (ed.): *E & MJ Second Operating Handbook of Mineral Processing,* McGraw-Hill, New York 1980.
6. E. G. Kelly, D. J. Spottiswood: *Introduction to Mineral Processing,* Australian Mineral Foundation, Auckland 1995.
7. D. Bradley: *The Hydrocyclone,* Pergamon Press, London 1965.
8. H. E. Cohen: "Energy Usage in Mineral Processing," *Trans. Inst. Min. Metall. Sect. C* **92**(1983) C160-164.
9. J. C. Guillaneau, J. Villeneuve, P. Blot: *"Advances in the Design and Optimisation of Mineral Processing Plants*," Proc., APCOM 92, 23rd International Symposium on the Application of Computers and Operations Research in the Mineral Industry, Tucson, 1992.
10. T. Yalcin (ed.): *Innovations in Mineral Processing,* ACME, Sudbury 1994.
11. A. F. Taggart: *Handbook of Mineral Dressing,* Wiley, New York 1945.

Screening

See also → Solid – Solid Separation, Introduction

PAUL SCHMIDT, Universität Essen, Essen, Federal Republic of Germany

1. General 2
1.1. Applications of Screening 2
1.2. Fundamentals of Screening 3
2. Theory of Screening 4
2.1. Free-Flow Screening, Scalping . . . 6
2.1.1. Adhesive Forces between Particles. . 6
2.1.2. Ultrasonic screening 7
2.2. Probability Approach to Screening 7
2.3. Screening as a First-Order Reaction 8
2.4. Empirical Screening Function 8
2.5. Applying Dimensional Analysis to Screening 9
2.5.1. Free-Flow Screening with Vibration . 9
2.5.2. General Application of Dimensional Analysis 10
2.6. Particle – Mesh Interaction 11
2.7. Oblique Flow to the Mesh 11
2.8. Fluid-Flow Screening 12
3. Screening Machines 12
3.1. Vibrating Screens 12
3.2. Sifters 14
3.3. Trommel Screens 17
3.4. Special Screens 17
3.5. Sizing of Screen Surfaces 19
3.6. Flow Screens 20
3.6.1. Pneumatic Screens 20
3.6.2. Wet Screens 22
4. Screen Arrays 22
5. Screen Surfaces 22
5.1. Cleaning of Screens 24
5.2. Electrostatics in Screening 25
6. Safety Requirements 25
7. References 25

Symbols

a	acceleration, fraction of fines in the feed
A	area
c	concentration, constant, correction factor
d	wire gauge
e	amplitude
E	modulus of elasticity
f	frequency, fines in fine fraction, function
F	force
Fr	Froude number
g	acceleration due to gravity ($g = 9.81$ m/s^2)
h	height
H	depth of bed
k	sieving rate constant
K	acceleration number, throw number
l	lenght
L	screen length
m	mass
$\dot{m}$	mass flow, specific capacity
n	number, rpm
p	pressure
P	probability, power
$q(x)$	particle-size distribution
$Q(x)$	cumulative particle size distribution
R	radius
t	time
T	grade efficiency
u	velocity of fluid flow
v	velocity
V	volume
$\dot{V}$	volume flow
w	mesh size
x	particle size
α	casting angle
β	slope
δ	angle of screening
Δ	difference
η	efficiency, viscosity
κ	sharpness of cut
μ	coefficient of friction
Π	dimensionless number
ϱ	density, angle of friction, angle of radial inclination
σ	normal stress
τ	shear stress
φ	open screen area
Ψ	sphericity
ω	angular velocity

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_15

Indices

a	side of rectangle
b	side of rectangle
B	width
c	critical
C	near-mesh
F	fines, fluid
G	coarse
h	horizontal, transport
H	bond, adhesive
H	depth of bed
K	plugging, jamming
L	length
m	mean
m	mass
$\dot{m}$	mass flow
M	medium
n	normal
R	radius
S	screen, segregation
T	separation, cut size
v	vertical
V	volume
w	mesh
0	feed

1. General

Sieving or screening is a method of separating particulate materials. The separation is according to size and is referred to as classification. Separation results from the repeated, random comparison of each individual particle of size x with the characteristic size w of the screen opening, as shown in Figure 1. Particles may be larger or smaller than the mesh size. So-called near-mesh particles, however, present problems. Both the feed to the screen and the resulting fractions are described by the particle-size distribution $q(x)$, the cumulative particle-size distribution $Q(x)$ or by the parameters x_{m} and n of a distribution, where by the RRSB distribution is usually chosen. The sharpness of cut is shown by the grade efficiency curve $T(x)$ or the $\kappa_{25/75}$ value (for further information, see → Particle Size Analysis and Characterization of a Classification Process, Chap. 3.1.5.). In some cases, screening can be used to separate according to composition or particle shape. This is referred to as sorting. Most articles are published in manuals, encyclopedias, and journals of chemical engineering, food technology, agriculture, particle and powder technology, mining, and building materials. For an overview, see [23].

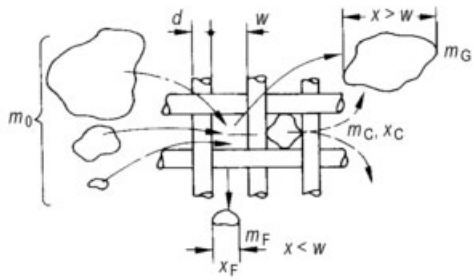

Figure 1. Screening as a comparison between particles of size x in the feed material m_0 [particle size distribution $q(x)$ or $Q(x)$] and the mesh with aperture w. Partition into oversize, coarse, or nonpassing (m_{G}) and fines or passing (m_{F}); near-mesh particles (m_{C}) with $x \approx w$ cause blocking of the screen

1.1. Applications of Screening

Most processes in which particulate materials are handled employ some screening method to obtain a product of desired quality. Coarse screening, with mesh size $w \gtrsim 100$ mm, is used to process earth and rock as well as ore and coal. Intermediate screening, with $100 > w > 10$ mm, is crucial in the preparation of building materials such as gravel, sand, and crushed stone. Fine screening, with $10 > w > 1$ mm, is employed in all industries: agriculture, food processing, particle-board manufacture, etc. Here and in ultrafine screening ($1 > w > 0.025$ mm) the methods and machines have been developed extensively. Some examples are the screening of plastic powders, toners, fillers, pigments, pharmaceuticals, and flour. Especially narrow size fractions are required in screening abrasives.

Screening assists both size reduction and agglomeration processes. Screening machines are responsible for economically removing particles of the proper size from the process stream. Screening machines are also used for solid – liquid separation as dewatering screens and screening centrifuges. Even the generation of fluidized beds is sometimes performed by vibratory screens, for example, in drying particulate

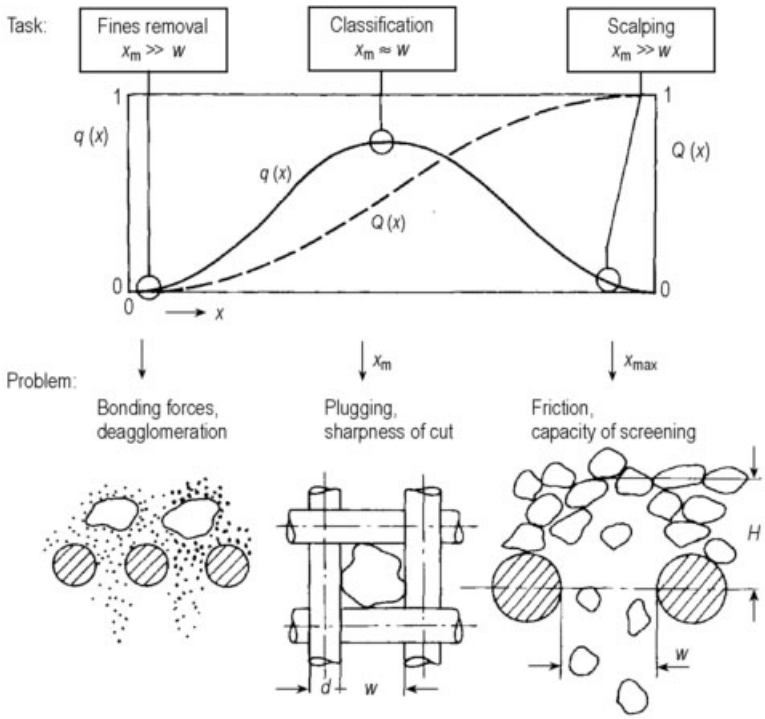

Figure 2. Principal practical applications of screening, which depend on the relationship between mean particle size x_m and mesh size w

solids. Worldwide, about 5×10^9 t of product are screened per year. About 100 companies produce screening machines, and about 50 manufacture wire cloth.

From a practical viewpoint, dry screening finds application in three categories, as shown in Figure 2.

1) *Scalping or trash removal* is the process in which relatively small amounts of foreign matter, lumps, and coarse impurities are removed from the product. The particles are predominantly smaller than the mesh size, $x_m \ll w$. The main problem in this case is to obtain a high specific screen capacity $\dot{m}$. The process often involves a trickle flow of the product through the meshes with flow resistance on side of the sieve.
2) *Classification* is the separation of material into fine and coarse fractions. The problem here is to achieve a high sharpness of cut and avoid blinding of the screen by jammed particles. The latter occurs because many particles are close to the mesh size, so-called near-mesh material $x \approx w$. In thick-layer screening the capacity is limited by layer separation; therefore, segregation, mixing, or blending of the layer is the dominant operation.
3) *Fines removal or dedusting* is the removal of a small amount of dust that clings to the particles, $x_m \gg w$. The problem here is deagglomerating and overcoming adhesive forces. Fines removal normally requires a long screening time. Theoretically, this is segregation.

1.2. Fundamentals of Screening

Batchwise screening is generally used for analytical purposes, where it is referred to as sieving. In most cases, screening is a continuous process. As shown in Figure 3, a feed stream $\dot{m}_0$ is transported across a screen and separated into the passing fraction or fines m_F and the nonpassing fraction or oversize m_G. The coarse fraction consists of oversize grains as well as misplaced

particles, i.e., fines m_{FG}. If the screen is intact, the passing fraction should not contain any misplaced particles. However, because mesh size can vary by the manufacturing tolerance $\pm \Delta w$ from the nominal mesh size w, some misplaced particles can be found in the fine fractions also; their mass is m_{GF}. A hole due to a damage of the screen may be the reason for this, especially in the case of very coarse misplaced particles.

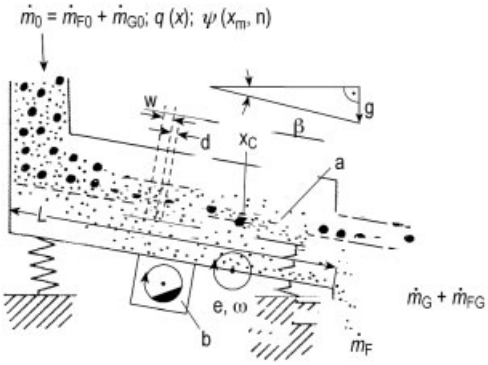

Figure 3. Operation of screen classification
The vibrator (b) causes a mechanical fluidization of the bed, which results in a stratification, and moves the feed material $\dot{m}_0$ across the screen (a). Separation into nonpassing material m_G and passing material m_F occurs by diffusion and under the combined influence of vibratory acceleration $a = e\,\omega^2$ and gravity g. All other parameters relevant to the process of vibrational screening are also shown.

To give a judgment about a screen, one needs its efficiency and its capacity or screen surface rating. The two characteristics must be considered together because they behave reciprocally.

Many formulae can be found in the literature for calculating the efficiency η of screening. For a concentration c_{F0} of fines in the feed, c_{FG} in the overflow, and c_{FF} in the underflow it is given by

$$\eta = [(c_{F0} - c_{FG})(c_{FF} - c_{F0})/ c_{F0}(1 - c_{F0})(c_{FF} - c_{FG})]\,100\,\% \quad (1)$$

Often, the fine fraction contains no misplaced grains; then Equation (1) simplifies to

$$\eta = [(c_{F0} - c_{FG})/c_{F0}(100 - c_{FG})]\,100\,\% \quad (2)$$

Example: Screening of a feed material with $m_0 = 1000$ g (or $\dot{m} = 1000$ g/s) consisting of fines ($m_{F0} = 500$ g) and coarse material ($m_{G0} = 500$ g). The overflow contains fines ($m_{FG} = 90$ g), and, due to screen damage, the underflow includes some coarse material ($m_{GF} = 10$ g). This gives the concentrations

$$c_{F0} = 0.5,\ c_{FG} = 90/(490 + 90) = 0.15517, \text{ and}$$
$$c_{FF} = 410/(410 + 10) = 0.97617.$$

According to Equations (1) and (2) the efficiencies are

$$\eta_{(1)} = [(0.5 - 0.15517)(0.97619 - 0.5)]/ [(1 - 0.5)(0.97617 - 0.15517)] = 0.800 \text{ or } 80.00\,\%$$

and

$$\eta_{(2)} = (0.5 - 0.15517)/0.5(1 - 0.15517)] = 0.8163 \text{ or } 81.63\,\%.$$

Note: The screening efficiency calculated depends on the formula used.

The fraction of fines in the feed, in the overflow and in the underflow is determined by analysis. Unfortunately, these analyses are not absolute. They are performed mainly on laboratory sieves, whose efficiency is often of the same order of magnitude as that of production screens.

The other important parameter for a screen is its specific capacity, also called the screen surface rating $\dot{m}$. It is the ratio of mass flow to screen surface area for a particular separation, measured in kilograms per square meter per second or tonnes per square meter per hour. The mass flow usually refers to the feed stream $\dot{m}_0$. It is a function of the cut or mesh size, size distribution, the type of feed particle, the type of screen, its operating parameters, and desired efficiency. The specific capacity together with the efficiency determines the screen size required for a screening machine.

2. Theory of Screening

Screening is very easy to do. Its use is shown in illustrations from ancient times and from the Middle Ages. However, a precise analytical treatment is extremely difficult, because a large number of parameters influence the screening process (see Figure 3 and Section 2.5.2). In spite of many efforts to find a universal theory of screening, there are about a dozen theories in existence. So even today, when screening machines

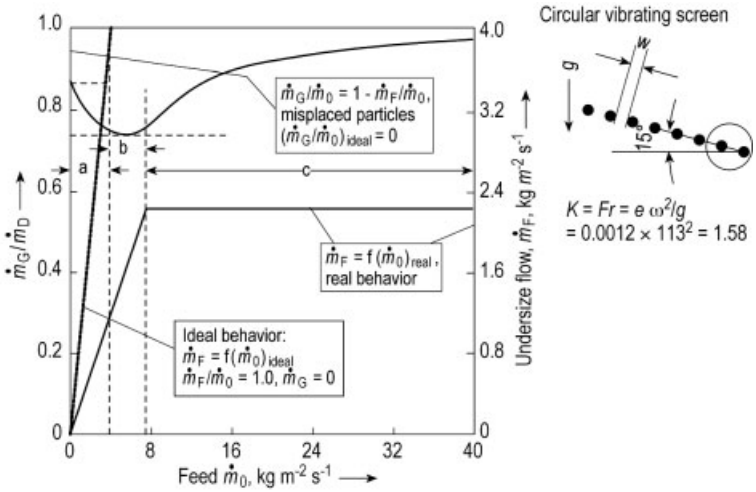

Figure 4. Screening of cement clinker with exclusively near-mesh undersize material only, $0.8\,w < x < w$, on a circular-stroke vibrating screen ($e = 2.5$ mm, $\omega = 100\,\mathrm{s}^{-1}$, $\beta = -15\,°$) with a square mesh ($w = 3.15$ mm, $d = 1.25$ mm).
Top curve: relative underflow $c_F = \dot{m}_F/\dot{m}_0$ or relative overflow $c_G = \dot{m}_G/\dot{m}_0$
There are three regimes of screening: a) separated particle screening; b) thin-layer screening
Maximum undersize flow is realized in thin-layer screening with an efficiency $\eta \approx 30\,\%$ instead of the 100 % theoretically possible.

are designed for new products, extensive experimentation must be carried out. The results can be extrapolated to some extent for similar tasks or used as factors in the theories. Nevertheless, analyzing the process of screening and treating it analytically as far as possible is a way to understand the physics of screening. Ultimately it is possible to calculate and design a screen satisfactory by making use of theory and of data previously obtained by experiment.

In order to realize the basic process of screening as a comparison of particles with the meshes (Fig. 1) on a technical scale the feed should be a bed with a height of one coarse particle only. In this case nearly all particles are in contact with the screen, and only about ten comparisons are necessary to make the decision between fine or coarse due to the probability of passing the mesh. In the case of a vibratory screen, particles must only jump from one mesh to the next mesh. This results in slow transport along the screen and a very short screen. To achieve a high capacity, the screen must be very wide, but such a design is very unfavorable for feeding, especially for fine particles. Because of this, screens are usually about twice as long as they are wide with a feed bed about ten to a hundred coarse particles thick jumping over many meshes. Thus instead of ideal thin-layer screening, thick-layer screening is more the standard. It is this compromise of thick-layer screening that makes it so difficult to find a simple and universal theory of screening.

Basically, there are two resistances to be overcome: passing of fines through the meshes and, in thick-layer screening, additionally, segregation of fines out of the bed to the screen.

The latter process is the main problem. It can arise from segregation of fine particles due to overcoming of Coulomb friction and finally result in a layer of coarse particles floating on a layer of fines. Depending on the particle size distribution, the cohesion, and the friction in the bed the intensity of vibration may be sufficiently low to realize the phenomenon of percolation. In the case of high intensity, blending of the material occurs and results in a constant fines concentration across the bed height and a decreasing fines concentration along its length. Both mechanism of transport are a type of diffusion and can be treated as a first-order reaction (see Section 2.3).

To elucidate the complicated behavior of particulate material on a vibrating screen, Figure 4 shows as an example screening of exclusively near-mesh undersize particles. These particles, $0.8\,w < x < w$, are always critical for

the screening process and so govern it. Depending on the feed, three modes of operation occur: single-particle screening, thin-layer screening, and thick-layer screening. Based on such experiments, one of the many mathematical models for screening on vibrating screens with circular throw has been developed [1].

There are several ways to develop a theory of screening mathematically:

1) Free flow or trickling of particles through the meshes like in an hourglass
2) Probability of passage for a particle through a mesh
3) Applying a kinetic model by assuming a first-order reaction
4) Approximation by an *e*-function
5) Treating the process of screening by dimensional analysis

2.1. Free-Flow Screening, Scalping

Free-flow screening is analogous in operation to an hourglass. In scalping, it is frequently the dominant component process. The prerequisites are a free-flowing feed, a feed layer of sufficient thickness $H/w \gtrsim 2$, a ratio of mesh size to particle size of $w/x \gtrsim 4$, and only a small amount of coarse material. As shown in Figure 5, the feed forms an arch over the mesh, through which the particles fall. The configuration of the arch is similar for each feed – mesh combination [2]; thus h/w = constant.

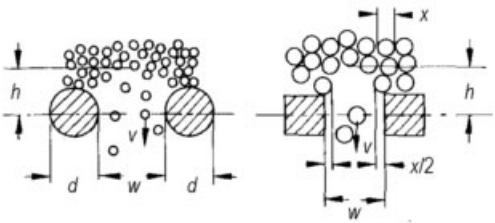

Figure 5. Arching over the screen mesh; *h/w* is constant according to similarity; trickle flow through the meshes

The screen capacity is controlled by the downward velocity through the plane of the screen. By using the formula for the velocity of a falling object $v = \sqrt{2gh}$ the proportionality

$$v \sim \sqrt{g\,w} \tag{3}$$

is obtained. The throughput per opening is

$$\dot{m}_{\mathrm{w}} \sim \varrho\, w^2\, v \tag{4}$$

From Equations (3) and (4) and by introduction of a constant *c*:

$$\dot{m}_{\mathrm{w}} = c\,\varrho\,\sqrt{g}\,w^{5/2} \tag{5}$$

Particles on the edge of the mesh hinder falling within a distance one-half of their diameter. This leads to a reduction in the effective mesh size *w* to $(w - x)$ (see Fig. 4). With this correction, the specific screen capacity, in kilograms per square meter per second, is finally obtained:

$$\dot{m} = c\,(w+x)^{-2}\,\varrho\,\sqrt{g}\,(w - x)^{5/2} \tag{6}$$

The correction factor *c* is about 0.2 – 0.5, so that the resulting important relationship for free flow is

$$\dot{m} \sim \sqrt{g}\,\sqrt{w} \tag{7}$$

2.1.1. Adhesive Forces between Particles.

In screening coarse- and intermediate-size feeds, the particles are affected only by inertial and frictional forces. For very fine particles, however, fluid flow and adhesive forces are dominant. If the former are ignored, the latter must be considered: liquid bridges between wet particles, van der Waals forces, and electrostatic attraction. The exact determination of these interparticle forces or forces between particles and screen is complicated, but it is known that the attractive or adhesive force between particles is proportional to the particle size. This gives

$$F_{\mathrm{H}} = c_{\mathrm{H}}\,x \tag{8}$$

For practical application, a rough estimate gives a cohesive stress of 1 bar for 1 µm particles, resulting in a coefficient of adhesion $c_{\mathrm{H}} \approx 0.1$ N/m. These forces hinder particle motion and, therefore, reduce screen capacity. With the assumption that the hindrance of flow is roughly proportional to the adhesive force, the following relationship is derived by using Equation (8); it applies to scalping for a very fine feed:

$$\dot{m} = c\varrho\sqrt{g}\sqrt{w}x \tag{9}$$

If $x \approx w$, then for very fine feed the relationship between specific screen capacity and separation size or mesh size is

$$\dot{m} = c\varrho\sqrt{g}w^{3/2} \tag{10}$$

with the constant $c \approx 2.8\,\mathrm{m}^{-1}$. Figure 6 shows that this model predicts the specific screen capacities observed in operation. The commonly given formula $m \sim w$ is an approximation for the range $0.1 < w < 10$ mm.

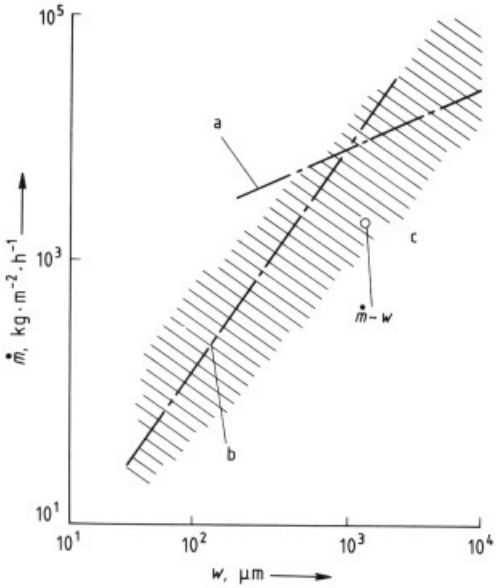

Figure 6. Relation between screening capacity $\dot{m}$ and mesh size w
a) Coarse material, $\dot{m} \sim w\sqrt{g}$; b) Fine products with interparticle bonding forces, $\dot{m} \sim w^{3/2}$; c) Practical relation for intermediate material, $\dot{m} \sim w$

2.1.2. Ultrasonic screening

In order to overcome the adhesional forces in ultrafine screening, an acceleration is necessary

$$a = e\omega^2 = e\,(2\pi n)^2 = 6c_{\mathrm{H}}x/\pi\varrho x^3 \tag{11}$$

This results in a frequency of vibration

$$n = (2c_{\mathrm{H}})^{1/2}/2\pi\left(ex^2\varrho\right)^{1/2} \tag{12}$$

Example: Amplitude $e = 10\,\mu$m, particle size $x = 20\,\mu$m, density $\varrho = 2500\,\mathrm{kg/m^3}$, and $c_{\mathrm{H}} = 10^{-1}\,\mathrm{kg/s^2}$; from Equation (12), the frequency required is $n \approx 22\,500$ Hz, which lies in the ultrasonic range [3].

2.2. Probability Approach to Screening

The probability P_{S} that a fine particle x going over the screen will pass through an opening of width w depends on the dimensionless parameters x/w and d/w, as well as on the geometric configuration. In the case of square mesh screens, the probability is

$$P_{\mathrm{S}} = (w - x)^2 / (w+d)^2 \tag{13}$$

The reciprocal $1/P$ of the probability gives the number of vibrations n necessary for fine particles to pass through the screen. Figure 7 shows the relationship for square-mesh screens. The passage of a bed of particles through a screen is impeded not only by the wire mesh but also by the volume fraction $V_{\mathrm{G}} = 1 - V_{\mathrm{F}}$ of coarse particles; it decreases steadily along the lenght of the screen. The resulting probability of passage is $P_{\mathrm{V}} \sim V_{\mathrm{F}}$. For particles that are not directly on the screen surface, the probability further decreases with increasing thickness H of the bed, so $P_{\mathrm{H}} \approx w/H$. This fact once again explains the effectiveness of thin-layer screening. For square-mesh screens, the total probability of fines passing through is therefore

$$P_{\mathrm{tot}} = P_{\mathrm{S}}P_{\mathrm{Y}}P_{\mathrm{H}} = \left[(w - x)^2 / (w + d)^2\right] wV_{\mathrm{F}}/H \tag{14}$$

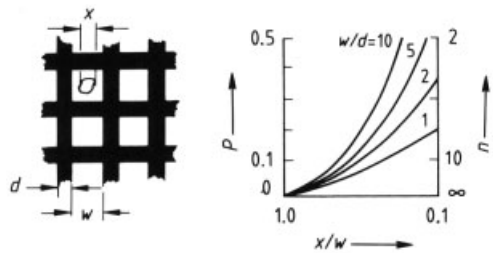

Figure 7. Probability P of a fine particle x passing through a square mesh w with wire diameter d in one vibration or the number of vibrations n needed to give $P = 1$

If, for example, in fines removal, $w = 200\,\mu$m, $x = 100\,\mu$m, $d = 100\,\mu$m, $V_{\mathrm{F}} = 10\,\%$ and $H = 5$ mm, then the resulting probability for passing the mesh at one impact is 0.00044. The number of vibrations required to yield a separation of 1 is 2250. On a screen vibrating at 50 Hz, this corresponds to a residence time $t = n/f$ of 45 s.

Because of the large number of factors that influence screening, applied statistics, like design of experiments, is a promising way to study this problem. An example for the use of regressional analysis is given in [4].

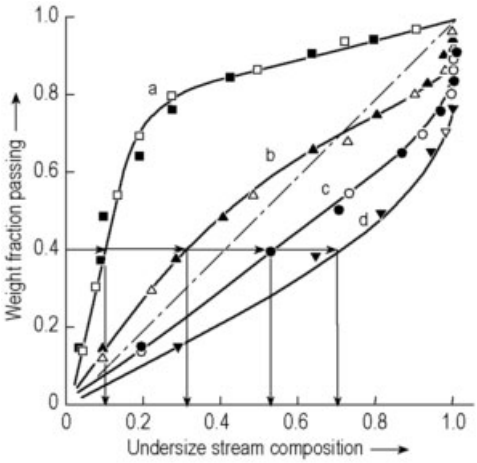

Figure 8. Sieving and screening as a first-order reaction Undersize stream composition vs. weight fraction passing through [7]
Example: for 40 % weight fraction passing through, 70 % of the fraction $x/w = 0.3$, but only 10 % of the fraction $x/w = 0.92$ has passed through
a) $x/w = 0.92$, $k = 0.2\,\mathrm{kg\,m^{-2}\,s^{-1}}$; b) $x/w = 0.77$, $k = 1.1\,\mathrm{kg\,m^{-2}\,s^{-1}}$; c) $x/w = 0.65$, $k = 2.1\,\mathrm{kg\,m^{-2}\,s^{-1}}$; d) $x/w = 0.3$, $k = 2.8\,\mathrm{kg\,m^{-2}\,s^{-1}}$

2.3. Screening as a First-Order Reaction

Most products show a continuous particle-size distribution. For this reason, effective segregation cannot take place in the bed. Therefore, the material on the screen is mixed or blended by intensive vibration to move the fines against the screen. In sieving this results in a constant fines concentration along the height of the bed at a certain time. Now the sieving and screening process can be described analytically with the aid of the mixing model [5] and [6]. It is identical with a first-order reaction:

$$\mathrm{d}m_\mathrm{F}/\mathrm{d}t = -kAm_\mathrm{F}m \tag{15}$$

The passage of fines at time t is proportional to the content m_F of fines in the material bed m on the sieve, to the area A of the screen, and to the sieving rate constant k.

If the sieving rate constants k_i, in kilograms per square meter per second, of the individual fractions i and the composition of the feed are known, then the sieving or screening capacity and efficiency can be calculated. This sieving rate constant k_i is specific for the fraction i of a certain material, the screen, and its motion.

Figure 8 shows the weight fractions passing through, as a function of the cumulative distribution of the fines, in weight percent. The sieving rate constants k_i are independent of sieving time and lenght. A decrease can be observed at the end of sieving and for near-mesh-size material.

The k values and the calculated probability according to Equations (13) and (14) are closely related. The probability approach theoretically also leads to an equation for the first-order reaction [7].

A more unified model of screening can be formulated by assuming a combination of two first-order processes occurring simultaneously; these are segregation of undersize particles onto the screen, and passage of undersize particles through the meshes. Screening results can thus be predicted for different systems provided the rate constants are evaluated for each size fraction [8].

Superposition of Mixing and Segregation Some screening processes are intermediate between mixing and segregation or diffusion. The theoretical treatment involves a superposition of these processes and leads to a differential equation, known as the Fokker – Planck – Kolmogorov equation; solutions are given in [9]. Further treatment is beyond the scope of this article.

2.4. Empirical Screening Function

In thick-layer screening the main problem is layer separation rather than passing of particles through the meshes. It is known from experience that in screening of this type an exponential function can be used to represent the fines concentration in the overflow after screening time t:

$$c_\mathrm{FG} = F\exp[-t/h] \tag{16}$$

The parameters F and h are a function of the feed [c_G0, c_F0 and $q(x)$], the size and geometry

Table 1. Data for calculation of misplaced particles c_{FG} in the overflow after screening time t on a vibrating screen according to Equation (14); mesh size $w = 1$ mm (0.4 mm for pit sand)

feed c_{G0}, %	Spherical *		Compact (pit sand)		Splittery (crushed marble)	
	F	h	F	h	F	h
30 %	649	6.7	716	14	668	19
50 %	453	7.7	452	18	448	23
70 %	205	17.5	261	25	250	29

* $e = 0.004$ m, $\omega = 90\,\mathrm{s}^{-1}$, $K = Fr = 3.3$

of the particles, type of screen and motion (circular, e, ω, Fr). To give an impression of applied screening, Table 1 shows such data as they are presented in tables and diagrams [10]. Equation (16) represents the profile of the material layer on the screen. The same authors have also published tables and diagrams for calculating the screen area necessary for a vibrating screen with near-mesh-size material in feed [11].

For continuous screening instead of screening time t the screen length L is relevant. It can be calculated by using the relation $t = L/v_T$, where v_T is the transport velocity along the screen. For linear stroke

$$v_{TS} = [g\cos(\alpha - \beta)/\omega\sin\alpha]/ \left[(Fr - 1)\left(Fr^2 - 1\right)/Fr\right]^{1/2} \tag{17}$$

and for circular stroke at $K \approx 2-3.3$ and $c \approx 0.25-1.5$

$$v_{TC} \approx cg/\omega\cos\beta \tag{18}$$

2.5. Applying Dimensional Analysis to Screening

All the screening theories up to now are patchwork in so far as they assume an ideal process, but at the same time their formulas make use of factors found by experiment. A scientific approach to treating the process of screening is the application of dimensional analysis. Starting from the parameters of influence dimensionless numbers are derived. The relations between these have to be found by experiment. The results are universal validity, assuming geometrical sismilarity.

2.5.1. Free-Flow Screening with Vibration

Often in screening, the fines do not run through the mesh voluntarily, like in free-flow screening; they must be set in motion. In practice, this is achieved by mechanical vibration. This results in a more complicated screening process that can no longer be treated analytically, but it is a starting point for a complete treatment of screening by dimensional analysis.

The most important parameters in free-flow screening of narrowly sized materials are particle size x, mesh size w; wire diameter d; stroke $= 2e$; frequency or angular velocity ω, gravitational acceleration g; feed $\dot{m}_0 \approx$ underflow $\dot{m}_F \approx \dot{m}$; density ϱ of the particles, and coefficient of friction μ. According to Buckingham's theorem, these 9 parameters result in $9 - 3 = 6$ dimensionless Π numbers for free-flow screening.

1) $\Pi_1 = x/w = (0.1-10)$
2) $\Pi_2 = d/w = (0.4-0.8 \approx \text{const.})$
3) $\Pi_3 = e/w$ (vibration: 1 – 100; sifting: 5 – 1000) $\approx$ const.
4) $\Pi_4 = K$ or $Fr = e\omega^2/g$ (vibration: 2 – 5; sifting: 1)
5) $\Pi_5 = \dot{m}^2/\varrho^2\, g\, x$
6) $\Pi_6 = \mu = (0.5-1.0 \approx \text{const.})$

To make things simpler, the number of Π numbers can be reduced further to three because Π_2, Π_4, and Π_6 can be regarded as constant. The relationship between Π_1, Π_2, and Π_5 must be determined experimentally. Exponential products can also be developed, as shown in Figure 9. Π_2 is introduced as the open area of the screen $\varphi = \Pi_2^* = w^2/(w+d)^2$ and combined with Π_5 to give Π_5/Π_2^*

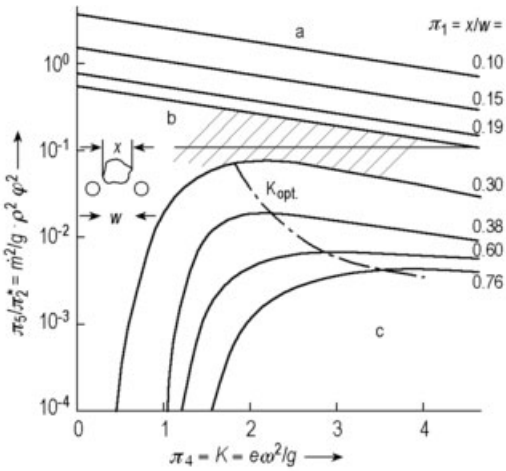

Figure 9. Relative capacity for scalping on vibrating screens with woven square mesh, presented as a relationship of dimensionless numbers.
a) Free-flow screening; b) Transition range; c) Vibratory screening

2.5.2. General Application of Dimensional Analysis

The most important parameters of influence for screening on a vibrating screen are included in Figure 3. In the following they are arranged according to (a) target, (b) geometry, (c) process, and (d) material data. For a circular-stroke vibrating screen this results in the parameters

1) Mass flow per unit width [kg m^{-1} s^{-1}] of undersize particles $\dot{m}_F$ or output $\eta = \dot{m}_F/\dot{m}_{F0}$
2) Mesh size w, wire diameter d, length of screen L, and slope β, (width $B = 1$ m)
3) Gravity g, amplitude e, frequency ω, and feed $\dot{m}_0$
4) Particle size distribution $q(x)$, represented by a RRSB distribution with mean particle size x and width of distribution n; particle density ϱ, and shape factor Ψ.

From these 13 Parameters of influence we obtain $13 - 3 = 10$ dimensionless Π numbers for the general task of screening on a vibration screen:

$\Pi_1 = x/w$ $\quad \Pi_2 = d/w$ $\quad \Pi_3 = e/w$ $\quad \Pi_4 = L/w$
$\Pi_5 = e\,\omega^2/g$ $\quad \Pi_6 = \dot{m}^2/g\,w^3\,\varrho^2$ $\quad \Pi_7 = \dot{m}_F/\dot{m}_0$ $\quad \Pi_8 = \beta$
$\Pi_9 = n$ $\quad \Pi_{10} = \Psi$

Note: These Π numbers hold for similar geometry only (e.g., woven square mesh or perforated plate). Parameters of influence not considered here are the elastic behavior of the particles and the screen, the surrounding air, coefficients of friction, roughness of particles and screen wires, and the blinding rate. The height of the material layer on the screen is not a parameter of influence; it varies with the feed and the velocity of transport along the screen. Figure 10 shows relations between four of the most important Π numbers as found by experiment in screening quartz sand on a circular vibrating screen with a mesh size $w = 1$ mm and constant slope $\beta = 15$ ° [12]. The other six Π numbers are regarded as being constant.

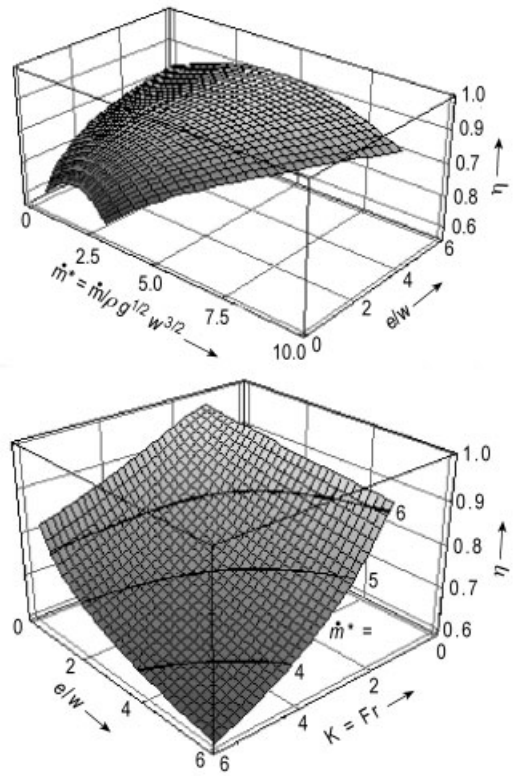

Figure 10. Screening efficiency η for screening quartz on a circular vibrating screen with mesh size $w = 1$ mm (top) as a function of the mass flow number $\dot{m}^*$, the relative stroke e/w, and $Fr = 3.3 =$ constant (bottom), depending on the Froude number $K = Fr$ and mass flow number $\dot{m}^*$. Slope $\beta = 15$ ° in both cases.

Evaluation of numerous experiments and mathematical analysis have resulted in an equation for screening efficiency:

$$\eta = 1 - \exp\left\{-0.277\,(e/w)^{-1/4}\,(K_V - 1)^{-1/2}\right.$$

$$[1 - (\beta/90)]^{-3}\,(L_{\mathrm{eff}}/w)^{1/3}\,[w/(w+d)]^{2/3}$$

$$\left.(x_{\mathrm{m}}/w)^{3/10}\,n^{3/2}\right\} \tag{19}$$

at optimum feed

$$\dot{m} = \varrho g w^{1.5}\left\{0.206\,(e/w)^{1/4}\,(K_V - 1)^{1/4}\right.$$

$$[1 - (\beta/90)]^{2.5}\,(L_{\mathrm{eff}}/w)^{1/2}\,[w/(w+d)]$$

$$\left.(x_{\mathrm{m}}/w)^{1/4}\,n^{-1/4}\right\}0.50 \tag{20}$$

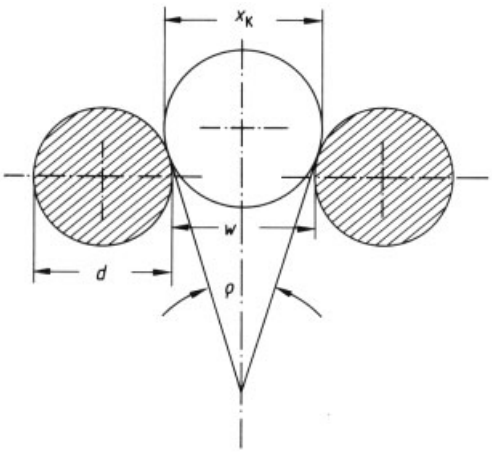

Figure 11. Particle diameter x_{K} for plugging is a function of the geometry of the mesh *w, d*, and the angle of friction ϱ

2.6. Particle – Mesh Interaction

Screening is influenced primarily by the relation of particle size to mesh size *x/w*. This is confirmed by Equation (13). For impact at 90 ° to the screen, a distinction must be made between near-mesh-size particles, $0.8\,w < x < 1.2\,w$, which hinder screening in general, and plugging particles, $w < x \lesssim 1.2\,w$, which prevent screening. Figure 11 illustrates how to calculate the largest particle plugging the mesh

$$x_{\mathrm{K}} = [(d+w)/\cos(\varrho/2)] - d \tag{21}$$

In reality, the plugging particle is often larger because of the tolerance Δw of the meshes (3 – 50 %), the tolerance of the coefficient of friction μ, and the particle form.

Although near-mesh-size particles strongly influence screening capacity, defining these particles is difficult. According to [5], they constitute the particle fraction in the sieving rate curve above the diagonal in Figure 8. Here the fraction nearest mesh size is always critical; the next one, only to some extent. Near-mesh-size characteristics ($0.8\,w < x < w$) depend also on the screen type, feed, and screen motion.

In general the plugging rate causes a decrease in the open screen area. This is equivalent to a reduction of the length *L* and width *B* of the screen and acts in the same way.

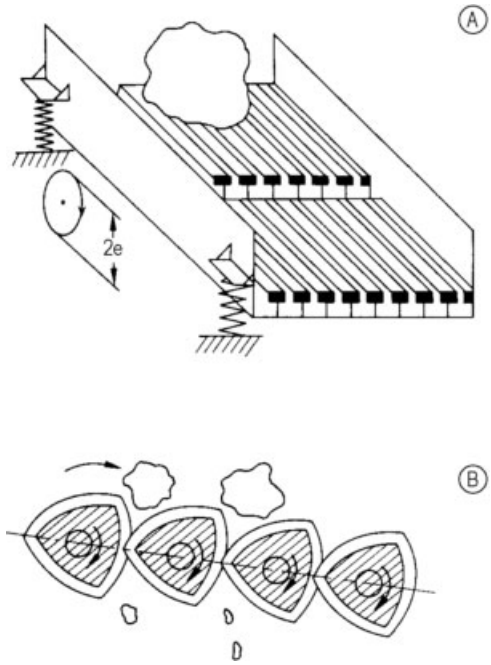

Figure 12. Inclined and possibly vibrating grating
A) Coarse screening (grizzly); B) Grid of power-driven rotating gratings (roll grizzly)

2.7. Oblique Flow to the Mesh

For oblique impact onto the screen the cut size becomes smaller than the mesh, $x_{\mathrm{T}} < w$. There is the geometric relation for the cut size

$$x_{\mathrm{T}} = (d + w)\cos\beta - d \tag{22}$$

Therefore, this is also a means to avoid blinding of the screen. In this case of oblique impact the probability P of passage also changes. It depends on the slope β of the screen, the relative particle size x/w, the surface rating, the relative diameter of the wires d/w, and a function $\Psi_M = f(x/w)$, [13], which for instance for $x/w = 0.5$ is $\Psi_M = 0.15$. The probability, then, is

$$P = \frac{[1 + \Psi_M (d/w) - (x/w)]\,[(1 + (d/w)) \cos\beta - (1 - \Psi_M)(d/w) - (x/w)]}{[1 + (d/w)]^2 \cos\beta} \tag{23}$$

The extreme situation is parallel flow of the material over the screen as on sifters with a usual relation $x_T \approx 0.8\,w$ (see Section 3.2). Another possibility is a sharp inclination of the decks, like the Mogensen Sizer (see Section 3.4).

2.8. Fluid-Flow Screening

In hydraulic or pneumatic transport of feed on the screen and through the meshes, velocities from 1 – 10 m/s are possible. This results in a very high capacity, independent of particle or mesh size. For the probability of particle passage through the mesh, Equation (13) is valid. If the direction of flow is perpendicular to the screen, problems arise with regard to the impact of the particles on the mesh, extremely strong blinding, and transport of oversize particles along the screen. Because of this, oblique flow is preferred. For wet screening, relationships of cut size to other parameters have been determined (see Section 3.6.2).

3. Screening Machines

The function of the screening machine is to move the feed, especially the coarse fraction, across the screen so that the fines segregate, encounter the mesh, and finally pass through. This is realized by vibration, which results in mechanical fluidization, overcoming Coulomb friction. Near-mesh particles are brought into repeated contact with the screen to increase the probability of their passing through.

In the simplest case, inclined and possibly vibrating gratings, so-called *grizzlies,* can perform coarse screening. A grid of rotating gratings can keep the openings clearer, as shown in Figure 12. Figure 13 surveys the types of screening machines manufactured.

3.1. Vibrating Screens

The vibrating screen is the classical screening machine. It operates on the principal of vibrational feed. The design is based on the ballistics of individual particles. These machines are classified by type of motion – linear, circular, or elliptical. The motion is described by two parameters, the amplitude e and the frequency f; the amplitude is half of the stroke. Angular velocity ω is related to frequency f by the relation $\omega = 2\,\pi f$, and to the number of revolutions n by $\omega = 2\,\pi\, n/60$. Linear or elliptical oscillation also includes the casting angle α. For the motion of the feed, the acceleration due to gravity g and the slope β relative to the horizontal must be considered as well. The process of vibrational screening is characterized by a dimensionless number, the screen number

$$K_v = a_n/g_n = e\omega^2 \sin(\alpha + \beta)/g\cos\beta \tag{24}$$

where a is vibrational acceleration. The machine itself is characterized by another dimensionless parameter, the acceleration number K, identical with the Froude number Fr,

$$K = e\omega^2/g \tag{25}$$

Figure 14 illustrates the relationship described above. The trajectories of the feed are shown in Figure 15. At a value of $K_v = 1$, the feed begins to jump. At $K_v = 3.3$, statistical resonance occurs, i.e., the time of flight is equal to the period of vibration. This is the starting point for adjustment of screening machines. At K_v between 3.3 and 6.6, the trajectories are irregular. At $K_v = 6.6$, the trajectories are twice as long as at $K_v = 3.3$. At $K_v = 1$, loosening and deagglomeration of the feed begin; however, mechanical stress begins too. The maximum velocity of feed

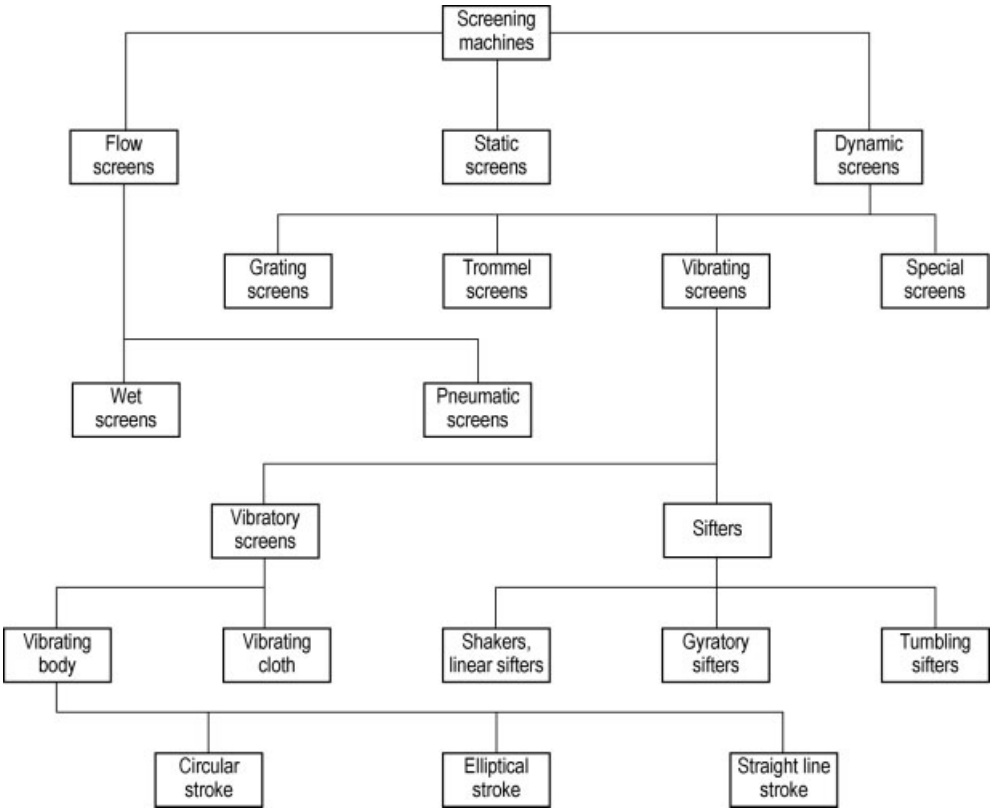

Figure 13. Survey of screening machines

for a horizontal screen occurs at $K_v = 3.3$ with a casting angle of $25-30\,°$.

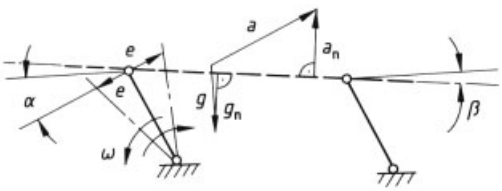

Figure 14. Derivation of the vibrational characteristic number, the throw number K_v for linear stroke $a = e\,\omega^2$;

$a_n = a \sin(\alpha + \beta)$; $g_n = g \cdot \cos\beta$

For horizontal screens with circular throw, the casting angle is relatively steep:

$$\alpha = \arcsin\left(g\cos\beta / e\omega^2\right) \tag{26}$$

Optimal settings for screening with vibratory screens are $K_v < 3$ for light, polishing screening; $K_v = 3-3.3$ for accurate separation when $\alpha > 45\,°$; and $K_v = 3.3-6.6$ for difficult feed. Screens with circular throw should be operated at a downward slope of $\beta = 10-15\,°$. The material bed must be mixed well. Therefore, the bed depth should be $H \lesssim 10\,w$ and the amplitude $e \sim w$.

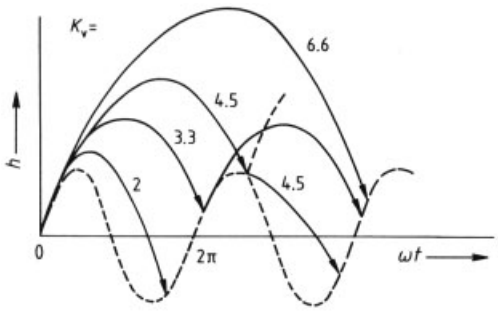

Figure 15. Idealized trajectories of a particle on the surface of a vibrating screen
Height h vs. $\omega\,t$ and throw number K_v; $K_v = 3.3$ first statistical resonance

Design of Vibration Screens There are innumerable designs of vibration screens to be found. The most important types can be distinguished by the motion: linear or straight-line, circular, and elliptic; and by the drive: forced and free drive (Fig. 16). Figure 17 gives an example of a vibration screen with circular throw and additional devices for improving the screening result.

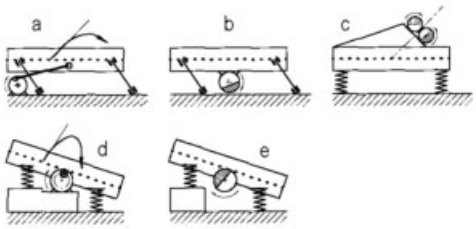

Figure 16. Most important types of vibration screens
a) Linear, forced drive, mounted on levers; b) Linear free drive, mounted on leaf springs; c) Linear, free drive with two rotating unbalances, rotating in opposite directions; d) Circular, forced drive with positive stroke; e) Circular, free drive unbalanced pulley.

To maintain operation in the favorable thin-layer screening regime (see Fig. 4, the transport velocity must decrease along the screen. This can be achieved by decreasing slope or by a differential angle deck. Such screens are sometime referred to as banana screens.

For fine and ultrafine classification on vibrating screens, the mesh can be driven directly as shown in Figure 18. Magnetic vibrators are attached at various points or act on the cloth via levers and support bars. Oscillating at 25 – 100 Hz, both types can provide the desired K_V value of up to 15. Because the screen is clamped at the rim, the casting or throw angle is approximately 90 °. For this reason, the screen is sharply inclined at an angle of 30 – 40 ° in order to move the feed alongside the screen. A thick, compact layer of feed is not possible. Actually, a thin layer moves rapidly across the screen. This is a very favorable range for operating a screen (see Figure 4) combined with some blinding (see Section 2.7). To achieve a high screening capacity, the screen must be wide, which leads to problems in distributing the feed evenly.

In ultrafine screening a pressure difference of 1 – 10 Pa causes an air flow that raises the capacity two- to fivefold.

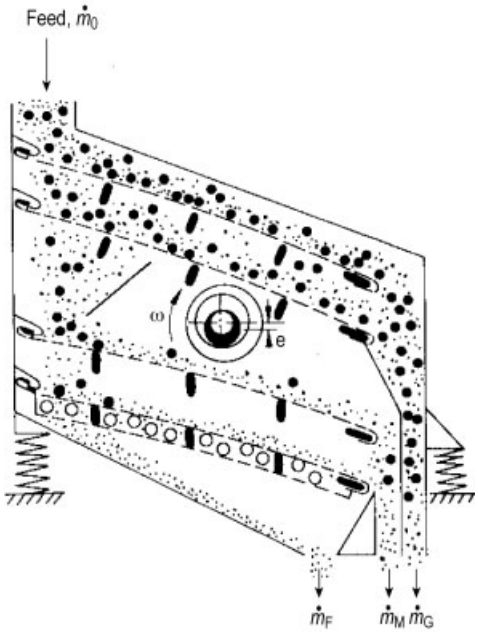

Figure 17. Circular stroke unit for fine screening with excenter drive in the center of gravity
Partial parallel configuration furnishes three fractions: $\dot{m}_G$, $\dot{m}_M$, and $\dot{m}_F$
Ball deck and balls are used on finest screen

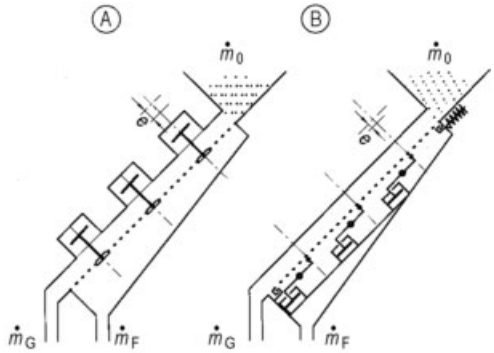

Figure 18. Vibrating screen with mesh driven directly by electromagnetic vibrators
A) Multipoint drive; B) Linear drive with screen tensioned lenghtwise
Rubber guards limit short-circuiting of the feed

3.2. Sifters

Sifting shows reciprocating or circular motion in the plane of the screen. In the first case it is referred to as shaking, and the second, as gyra-

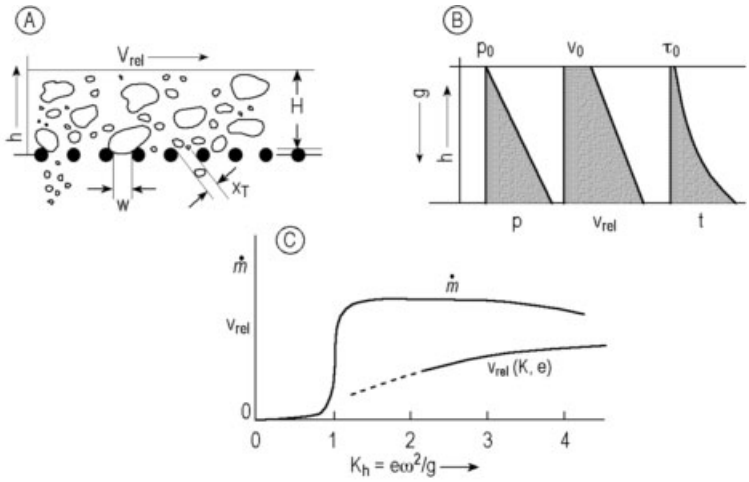

Figure 19. A) Process of sifting, near-mesh particle size $x_T < w$; B) Pressure p on the screen surface, relative velocity v_{rel}, shear stress τ; C) Capacity $\dot{m}$ and relative velocity v_{rel} vs. horizontal throw number K_{h} for $\mu_{\mathrm{s}} = \mathrm{const}$.

tory screening. Reciprocating motion is possible in the direction of feed flow or perpendicular to it if the screen is inclined. Gyrating sifters are becoming increasingly common. They are used mainly for fine and ultrafine screening.

The sifting process is the straining out, through the mesh, of the fines segregated on the screen. The direction of flow for the fines is not perpendicular to the mesh; instead, it is oblique to tangential. As a result, the separation size is smaller than the mesh size, $x_{\mathrm{T}} \approx 0.8\,w$. Because of this, few particles are jammed in the mesh. In comparison to vibrating screens, these devices show minor blinding of the screen for many products. To counter this kinetic narrowing of the mesh a screen can also be covered with a coarser mesh, $w \approx 1.2\,x_{\mathrm{T}}$.

The flow toward the mesh results in uniform separation conditions only if it occurs under constant flow velocity and if the vertical pressure remains constant. This is the case for gyratory, but not for linear, oscillation. In the latter, the thickness of the bed is crucial. As the content of fines decreases, the bed becomes thinner. However, it should not become so thin that the damping effect is lost and the particles jump around individually. In the feed bed itself, strong shear forces overcome friction and enable gravitational segregation to occur as the feed is deagglomerated. On the other hand, this results in high wear for sensitive materials, such as instant products (e.g., instant coffee). The shear stress due to Coulomb friction between the bed and wire cloth is

$$\tau = \varrho_{\mathrm{s}} \cdot g \cdot H \cdot \mu_{\mathrm{s}} \tag{27}$$

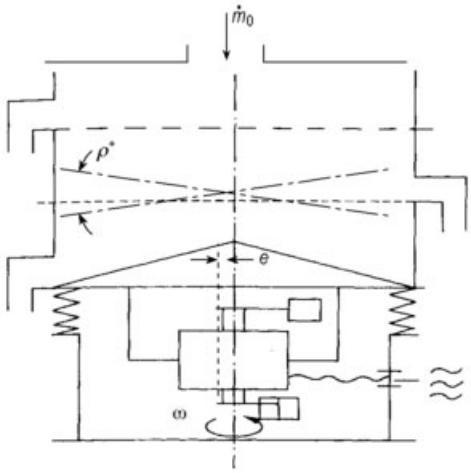

Figure 20. Tumbling sifter driven by a vibrator motor with a vertical shaft and two unbalanced weights
The twisting of the unbalanced weights against one another changes the movement of the feed across the screen: radial toward the outside, tangential to radial toward the inside

With a coefficient of friction $0.5 < \mu_{\mathrm{s}} < 2$. On a gyratory sifter there is a relative velocity between bed and screen

$$v_{\mathrm{rel}} = \sqrt{e^2\,\omega^2 - (\mu_{\mathrm{s}}^2\, g^2/\omega^2)} \tag{28}$$

where the angular velocity is

$$\omega = \sqrt{K_{\mathrm{h}} \cdot g/e} \tag{29}$$

The specific power input, in watts per square meter, is

$$P = \tau v_{\mathrm{rel}} \tag{30}$$

This applied power, mainly acting on one layer of particles, gives the power specific to volume

$$P/V = \tau v_{\mathrm{rel}}/x \tag{31}$$

Usually, a specific power of about 10 kW/m^3 exists in the boundary layer. This is even enough to accomplish grinding.

Figure 19 illustrates the dependence of cut size, pressure, horizontal velocity, and shear stress on the thickness of the feed bed, as well as on the resulting flow toward the screen. It also shows how the bed begins to slide over the mesh, after overcoming frictional forces, at values of $K_{\mathrm{h}} \approx 1$. This causes a sudden increase in screen capacity $\dot{m}$. As the K_{h} value or the relative velocity v_{rel} increases further, the capacity slowly decreases because the direction of flow toward the mesh becomes more oblique. For this reason sifters and similar types of screens rarely operate above $K_{\mathrm{h}} = 1$. An increase does not give any higher throughput. The relative velocity v_{rel} depends on the amplitude or the radius of gyratory motion for constant K_{h}.

Another important advantage of the sifter is the horizontal orientation of flat and elongated particles on the screen, such as wood chips, leaves, and straw. Wood chips, especially, do not get stuck in the mesh. Sifters are the only type of screen suitable for screening such products. However, this positive effect is disturbed at the rim. Each stroke produces an impact on the bed, which results in a mixing action. Thus, blinding of the screen is again possible. Screening of critical products should involve a free overflow for the oversize material, with no rim. Certain particles, such as metal powder or glass beads, often give rise to blinding when the sifter is started and stopped, because the relative velocity is not sufficient to cause oblique flow toward the mesh.

The gyratory sifter can be improved by controlling the residence time of the feed on the deck, instead of relying only on diffusion. This concept led to the development of the tumbling screen [14]. An example is shown in Figure 20 with simple construction operated by an eccentric drive. According to the relation $K_{\mathrm{h}} = e\,\omega^2/g \approx 1$, selection of a high frequency should be advantageous and favorable to the design.

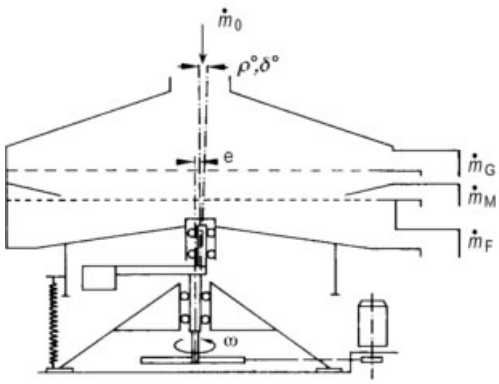

Figure 21. Tumbling sifter driven by a crankshaft with inclined trunnion; ϱ angle of radial inclination and θ angle of tangential inclination

However, for a frequency of 1500 rpm, an amplitude e of only 0.4 mm results. In this case, no flow toward the mesh occurs; rather, the particles dance around on it. But there is a relatively high angle ϱ of radial inclination that results in vibrational screening in the outer zone, instead of sifting as in the center. For actual screening, a high amplitude and, consequently, a lower frequency must be chosen. This consideration leads to a tumbling screen with exact construction requirements, driven by a driver motor (Figure 21). Optimal screening conditions can be obtained by adjusting the eccentricity, the radial and tangential inclination of the trunnion, and the rate of rotation. Good results can be obtained for most products.

Combinations of linear and circular motion are also used. These sifters are either inclined or driven by an inclined crankshaft (Fig. 22).

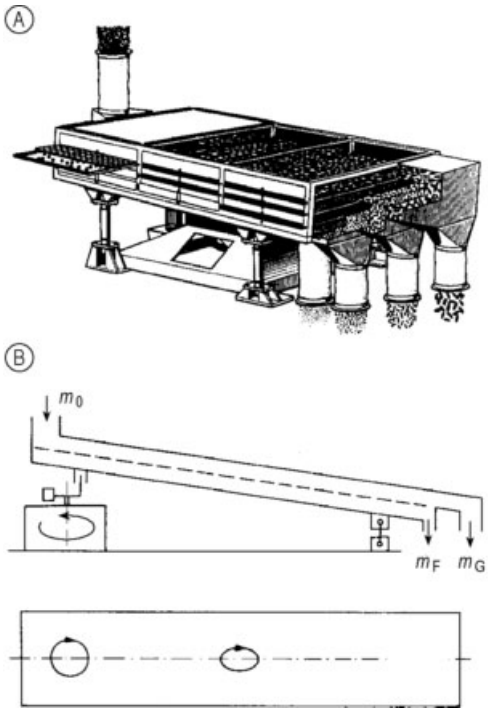

Figure 22. A) Inclined rectangular sifter driven by a crankshaft with vertical acceleration $a_v < 1\,g$; B) Inclined sifter with circular, elliptical, and linear motion along the screen

3.3. Trommel Screens

Trommel or revolving screens mix the feed thoroughly, without pushing it too much against the screen by centrifugal force. For this reason, the critical angular velocity ω_c should not be exceeded. It depends on the radius R of the trommel and its value is given by the formula

$$\omega_c = \sqrt{g/R} \tag{32}$$

Most of these devices are run at $\omega \approx 0.6\,\omega_c$. They are filled to less than 50 % of capacity. In the case of short trommels, axial transport of the feed is performed by diffusion or mixing. Longer trommels are inclined or have helical baffles which allow a definite residence time for the feed. Figure 23 shows that only a very small part of the screen surface is involved in the screening process.

Trommel screens have been used primarily for washing and simultaneously classifying gravel. They are also used as special screens for waste and scrap in recycling plants. One disadvantage of revolving screens is that to produce more fractions, the mesh size must increase along the screen, for example, screen surfaces in series. Parallel operation is also possible; this leads to concentric arrays of trommels with practical disadvantages in their construction.

Trommel screens offer the theoretical possibility of achieving very high screen capacities in the ultrafine range when used as centrifugal screens. Since $\dot{m} \sim \sqrt{a}$ and $a = R\,\omega^2$, the screen capacity can be increased by a factor of about ten.

3.4. Special Screens

For very fine particles, blinding of the screen is by far the most important factor impairing screen capacity. Most special screening machines are designed to keep the mesh open by special techniques.

The *centrifugal* or *cylinder screen* (see Fig. 24) is similar to the trommel screen. However,

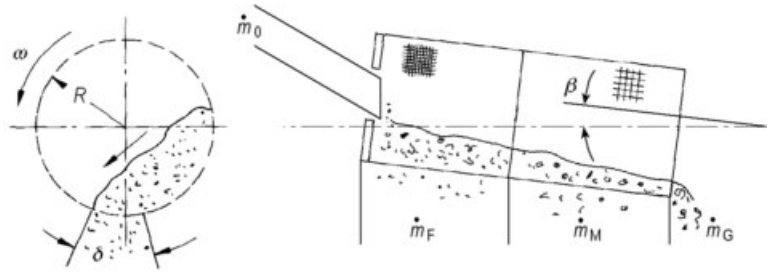

Figure 23. Trommel screen; δ region of active screening
Series arrangement of screen surfaces, $\beta = 0 - 10°$, $\omega \approx 0.5\,g/R\,\sqrt{g/R}$

in this case, the trommel does not move; vanes turning inside push the feed through the mesh instead. The screen cylinder is not under great tension and the mesh is cleaned because of a permanent slight deformation. These screens are used successfully for scalping in the range of 100 – 500 µm for nonabrasive products such as plastic powders [17]

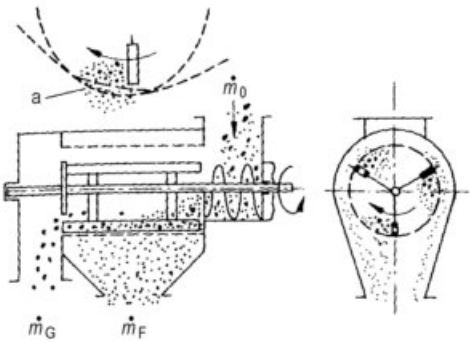

Figure 24. Centrifugal, cylindrical, or whirlpool screening machine
The mesh is bent at (a) by the vane and is cleaned by permanent deformation of the wire cloth

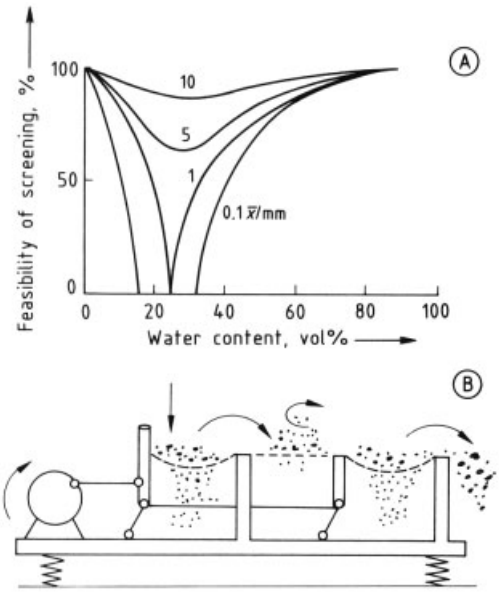

Figure 25. A) Feasibility of screening wet vs. dry products on conventional devices (feasibility plotted as function of average particle size x and water content); B) Principle of the stretching wave screening machine for screening wet products

Stretching wave screening machines, which provide for alternate stretching and sagging of the decks, are able to screen sticky products successfully. The mesh size must be greater than 2 mm because of the difficulty in fabricating more finely perforated sheets from rubber or polyurethane. Figure 25 shows the operation of one type of stretching wave screening machine. Also illustrated is the feasibility of screening wet feed, depending on the particle size and water content. At a critical moisture content of 20 – 30 vol %, separation can be performed with these screens. Stretching the screen by about 10 % leads to a high acceleration, with K_v values of approximately 100 [17].

The *Mogensen Sizer* (Fig. 26) employs a screenlike process. Generally, five sieve plates are stacked over one another. From top to bottom the mesh size decreases in a ratio 10:10:8:6:4. The slope of the sieve decks increases by 7 ° from plate to plate. In screening dry sand with a medium particle size $x = 2.5$ mm, for example, a sharpness of cut $\kappa_{25/75} = 0.6$ can be achieved. The capacity is $\dot{m} = 7.4\ \mathrm{kg\ m^{-2}\ s^{-1}}$, and the percentage of fines in the oversize $c_{FG} = 5$ %. However, the same amount of oversize particles are found in the undersize fraction. That is the crucial point in comparing the Mogensen Sizer with a conventional screen, for there are normally no misplaced particles in the undersize fraction.

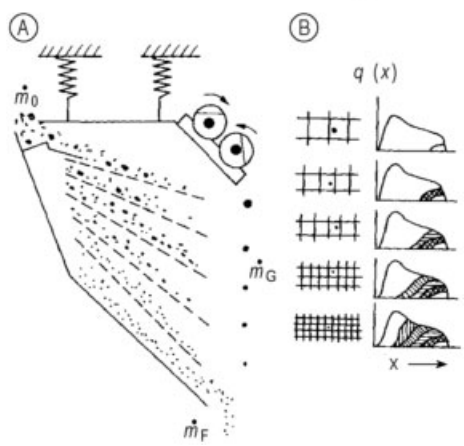

Figure 26. Mogensen Sizer
A) Operation; B) Composition of nonpassing fraction

Because there is no material bed on the screens and no near-mesh-size particles, blind-

ing does not occur. The vibration applied shows K values of 3 – 6 at high frequency and low amplitudes. It serves only to fluidize the material. This sizer is not sensitive to changes in capacity [18].

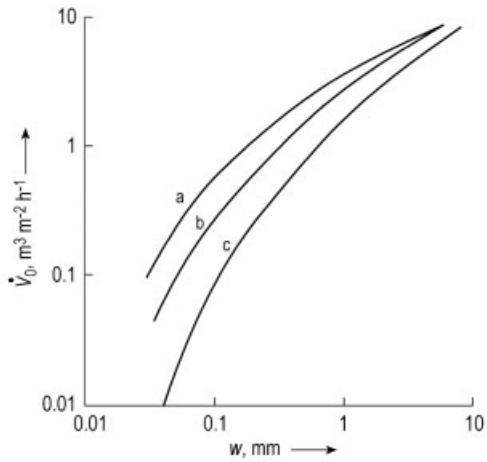

Figure 27. Rule of thumb diagram for calculating the screen surface for common products; specific volume flow $\dot{V}$ as a function of mesh size w and particle density
a) $\varrho = 10^4$ kg/m^3; b) $\varrho = 10^3$ kg/m^3; c) $\varrho = 10^2$ kg/m^3

3.5. Sizing of Screen Surfaces

The design of the required screen surface for a particular feed is the central problem of screening technology. The objective is to minimize the surface area by achieving the highest screen capacity. On the other hand, this depends on many parameters:

1) *Product:* composition, size, shape, particle size, particle size distribution, sharpness of cut
2) *Screen mesh size and type of screen:* vibrating, sifting, rolling, wet, pneumatic or other kinds
3) *Screening parameters:* amplitude, frequency, casting angle inclination, bed thickness

No generally applicable method exists for calculating the specific screen capacity and, consequently, the required screen surface area. For the special case of scalping, Equation (6) is used for coarse products and Equation (10) for fine ones. In segregation screening or fines removal, the area of the screen depends strongly on the adhesive forces in the feed. Thus, no calculation is possible for the screening area. In mixing screening, the area of the screen can be calculated by estimation of the sieving rate constant k experimentally. Then by using Equation (15) and the desired efficiency the screening area can be determined. For vibrating screens, the most accurate method is calculating the length of the screen from Equations (19) and (20).

Figure 27 serves as a rule of thumb for carrying out screening tests for a feed whose behavior

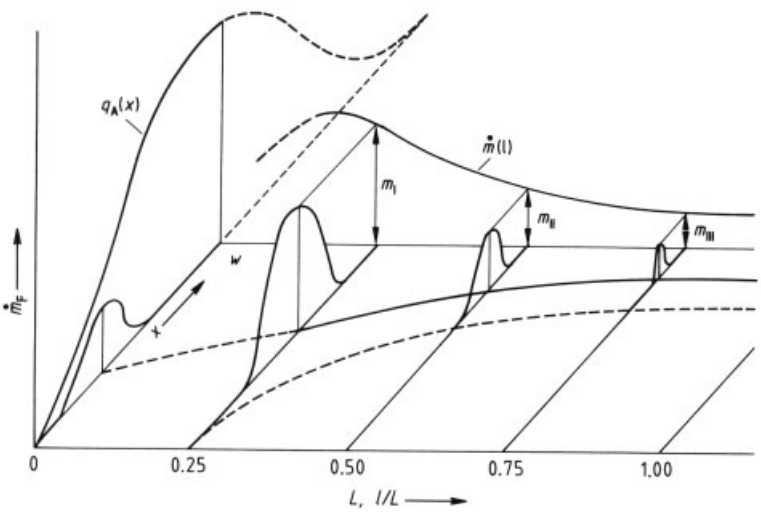

Figure 28. Amount and composition of the passing fraction along the length of the screen

is not known. It shows average values for the specific volume flow $\dot{V}$ for compact particles as a function of mesh size w.

Figure 28 shows qualitatively the yield of fines and the composition of the passing material over the length of a rectangular screen. As a consequence, the screen capacity can frequently be increased by placing a somewhat coarser screen over the first quarter, with $w \approx 1.2\,x_{\mathrm{T}}$. The sharpness of cut decreases only slightly [18].

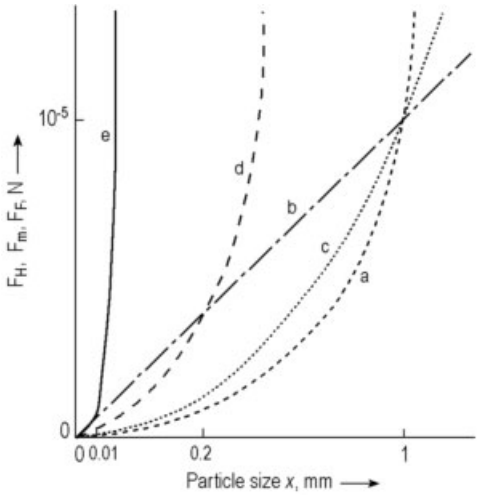

Figure 29. Inertial force F_m, adhesive forces F_{H}, and fluid forces F_{F} (F_{F} for air), respectively, as functions of particle size x
a) F_m; $a = g$; b) $F_{\mathrm{H}} \sim x$; c) F_{F}, $v = 6.3$ m/s; d) F_{F}, $v = 10$ m/s; e) F_{F}, $v = 50$ m/s

3.6. Flow Screens

The capacity of a mechanical screen reaches zero when the attractive forces become greater than the inertial forces that move the particles. Attractive or adhesive forces are roughly proportional to particle size x, and inertial forces are proportional to x^3. Consequently, the two functions intersect. If, as shown empirically, damp sand grains 1 mm in size just cling to each other under normal gravitational forces, then the limit of mechanical screening for this feed occurs at a vibrational acceleration $a = 10\,g$. The same situation applies to van der Waals forces and, roughly, to electrostatic forces. Only fluid forces are left to overcome adhesive forces. These are provided mainly by air or water. A rough estimation indicates that an airstream with a velocity of 50 m/s can overcome the adhesive forces of particles down to 15 µm in size, as shown in Figure 29.

3.6.1. Pneumatic Screens

By using fluid flow, flow screens can be constructed in which the feed is transported through the mesh hydraulically or pneumatically. In this case, the screen capacity is very high and is independent of mesh size. In practice, this type of screen has been used only for scalping inert materials such as plastic powder, as shown in Figure 30. These screens are prone to blinding, which occurs both by jamming of the near-mesh particles and by cementing due to fine dust. Nozzles that blow compressed air against the direction of flow through the screen are required to clean the screen [19].

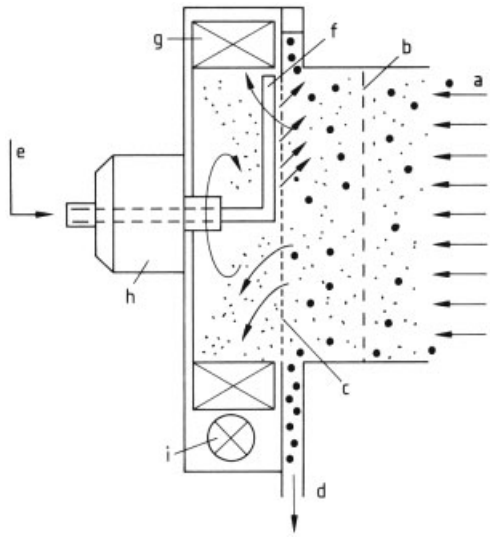

Figure 30. Principle of pneumatic screening
a) Two-phase flow composed of air and solids; b) Coarse protective screen; c) Screen; d) Nonpassing fraction; e) Compressed-air connection; f) Rotating compressed-air duct used to clean the screen and to remove nonpassing material; g) Blower; h) Drive motor; i) Exit for fines and air

In many cases, for fine screening, supplementation of mechanical screening by blowing air through the screen is sufficient. For vibrating screens, especially directly driven ones, a pressure difference of ca. 10 Pa is sufficient to measurably increase screen capacity and sharpness of cut. For sifters, a rotating nozzle arm blows

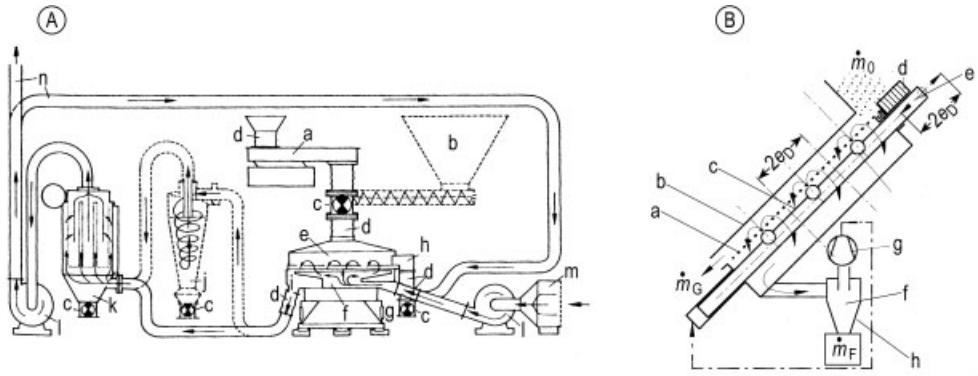

Figure 31. Combined mechanical and pneumatic screen array
A) Sifter based
a) Dispensing chute; b) Screw conveyor; c) Bucket-wheel valve; d) Flexible connector; e) Tumbling sifter; f) Air duct register; g) Air inlet; h) Coarse outlet; i) Fines exhaust; j) Cyclone separator; k) Filter; l) Blower; m) Air intake filter; n) Circulate – ventilate valve
B) Vibrating screen based with directly driven mesh
a) Screen surface; b) Body; c) Air duct register; d) Ultrasonic screen vibrator; e) Air duct register vibrator (low frequency); f) Fines collector; g) Blower; h) Closed air circuit (optional)

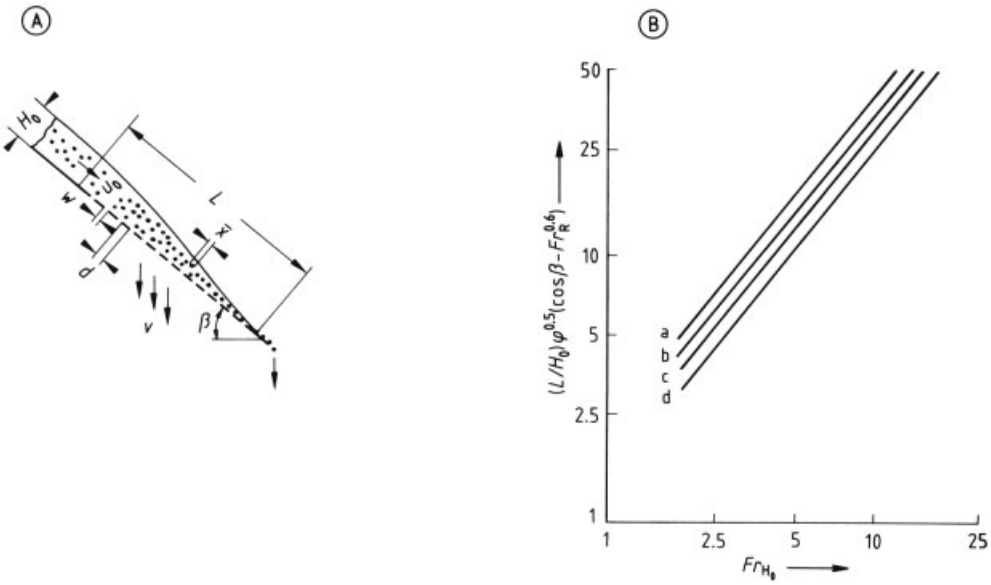

Figure 32. A) Minimum screen length L in wet screening, $\varrho_P/\varrho_F = 2.6$, $Fr = u_0/g\,R\,\sqrt{g}$ (sieve bend); B) Relation between dimensionless numbers for flat screens and sieve bends
a) $c_V = 0.2$; b) $c_V = 0.1$; c) $c_V = 0.05$; d) $c_V = 0$; $Fr = u_0/g\,H_0\,\sqrt{g}$, $\bar{x}/w = \text{const.} = 0.24$

air from below. In this case, supplementary traps for fines, such as cyclones or filters, are used. As a result, the machine evolves into an array of devices, as shown in Figure 31 A. Such air jet sieves of laboratory dimensions are frequently used for particle size analysis of small particles.

Occasionally, screens with directly driven mesh make use of air jet screening by applying a vibrating air duct register under the screen surface (Fig. 31 B).

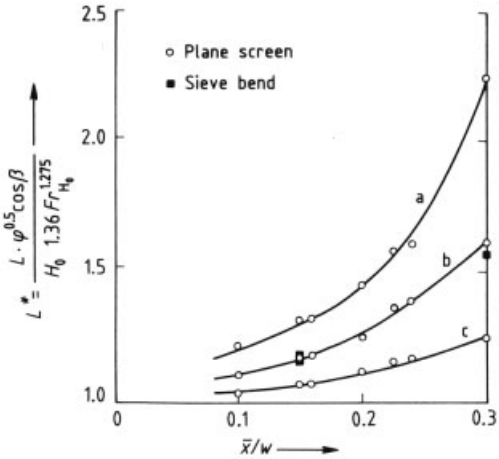

Figure 33. Minimum screen length L as a function of the ratio of average particle size to mesh size $\bar{x}/w$ and solids concentration c_V
Values shown for square-mesh flat screens and sieve bends
a) $c_V = 0.2$; b) $c_V = 0.1$; c) $c_V = 0.05$

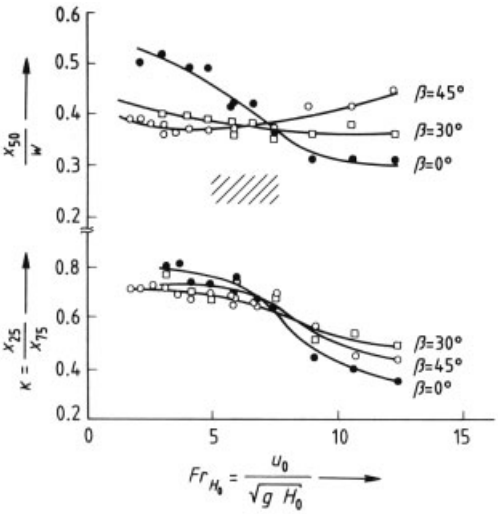

Figure 34. Specific cut point x_{50}/w and sharpness of cut $\kappa_{25/75}$ for wet screening on inclined screens vs. Fr number
$c_V = \text{const.} = 0.1$, hatched area shows the optimal Fr_{H0} range

3.6.2. Wet Screens

Wet screening may refer to underwater vibrational screening. For example, it can be used to improve the classification of a product obtained in water. In principal, it does not differ from the usual vibrational screening. In most cases, however, a suspension is passed over a screen surface. The objective is as much dewatering as it is classification or straining. The fundamental basis of this process is described by the relationship between characteristic dimensionless numbers, e.g., the Froude number Fr. Figure 32 shows the screen length necessary to remove free water from a suspension, as a function of the Froude number $Fr = u_0/\sqrt{g H_0}$ at various solids concentrations c_V. The influence of the particle size/mesh ratio and the solids concentration on the length of the screen is shown in Figure 33. In conclusion, Figure 34 shows the cut point and sharpness of cut that can be obtained on inclined wet screens as a function of Fr. Fundamentally, no difference exists between flat screens and sieve bends [20].

4. Screen Arrays

When more fractions of a material are wanted, use of a screening machine with more screen decks is not always helpful. In this case, the machines can be arrayed as shown in Figure 35. A combination with decreasing mesh size is generally used. However, increasing mesh size and various combinations are also possible. In this way, coarse material can clean finer meshes. When a high yield of a particular fraction is expected, screen surfaces can be arranged in parallel. In this case, splitting the feed into equal streams is often a problem.

5. Screen Surfaces

The most important element in a screen classifier is the screen surface. Figure 36 gives an overview of the important types. Most screens have a plain weave with square mesh and an aperture width w and are made from wire with uniform diameter d. Mesh sizes are standardized in ISO R 497, and the diameters of metal wire in ISO 4782. In the most general case, screens can have rectangular openings with different diameters of the wire in warp and weft. The free screen surface A_0 for this case is given by

$$A_0 = w_a\, w_b / (w_a + d_a)\,(w_b + d_b) \tag{33}$$

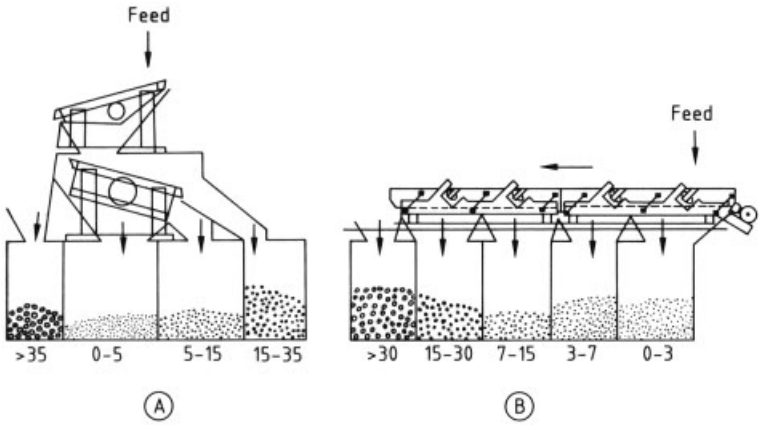

Figure 35. Screen arrays
A) Screening with decreasing mesh size and parallel configuration; B) Screening with increasing mesh size and series arrangement of screen surfaces

Twill weaving is also used for fine screens; in this case, the wire skips two or more meshes. The finest industrially produced screens have an opening of about 25 µm. Such screens are made of steel, spring steel, and stainless steel. Woven cloth is increasingly being made of synthetic fibers, mostly filaments, like polyester, polyamide, and polypropylene.

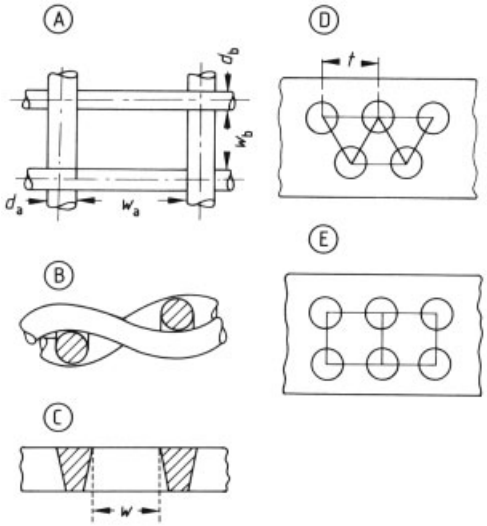

Figure 36. Screens
A) Woven rectangular mesh with opening sizes w_a and w_b, wire diameter d_a and d_b; B) Cross section of woven screen; C) Cross section of perforated plate with wider opening at the bottom; D) Perforated plate with triangular spacing; E) Perforated plate with square spacing

The following ISO standards apply to sieve plates:

- ISO 2194: wire screens and plate screens for industrial purposes – Nominal sizes of apertures
- ISO 4782: Metal wire for industrial screens and woven wire cloth
- ISO 4783: Industrial wire screens and woven wire cloth – Guide to the choice of aperture sizes and wire diameter
  - Part 1: Generalities
  - Part 2: Preferred combinations for woven wire cloth
  - Part 3: Preferred combinations for precrimped or pressure-welded metal wire screens
- ISO 9044: Industrial woven wire cloth – Technical requirements and testing.
- ISO 565: Test sieves – Woven metal wire cloth and perforated plate – Nominal sizes of apertures
- ISO 3310/1: Test sieves – Technical requirements and testing

Wire cloth is mostly woven 1.00 m wide; 2.00 m is often possible. The tolerances for woven screens are relatively large (Table 2).

Table 2. Examples for maximum permissible deviation of a single aperture measured in one direction. The deviation increases with decreasing size of the aperture. Test sieves show about half the deviation of industrial wire cloth

Aperture w, mm	Δw, % (ISO 9044)	Δw, % (ISO-DIS 3310-1)
16	12	6.2
1	28	14
0.5	36	17.8
0.25	46	23.2
0.125	61	30.4
0.063	83	41.3
0.045	100	48.9

The standard for wire cloth and sieve plates in the United States is given in

- ASTM E 437-85: Standard Specifications for Industrial Wire Cloth and Screens (Square Opening Series)
- USA Market Grades
- USA Bolting Cloth Specifications

Stainless steel bolting cloth is an extra-light wire cloth woven with highly durable wire. The mesh openings are in the range from mesh 14 ($w = 1.57$ mm) to mesh 230 ($w = 0.074$ mm). The wire is drawn with an extremely smooth surface to speed up the bolting action, which also increases the screening capacity. It is ideal for many industrial uses where high capacity and strength are required.

Perforated screens are generally stamped from sheet metal. The holes are round, square, or rectangular to slotlike. Special types also exist. Plates with fine and precise openings are made by electroplating. The precision of perforated plates, especially those made by electroplating, is much higher than that of woven screens. However, the free screen area is small. The probability of passing through a circular opening is lower than for a square mesh. For the same probability, the opening has to be about 1.25 – 1.42 larger than the mesh size w. Standards for perforated plate screens in the United States are:

- ASTM E 674: Standard Specification for Industrial Perforated Plate Screens (Round Opening Series)
- ASTM E 454: Standard Specification for Industrial Perforated Plate Screens (Square Opening series)
- ASTM E 161: Standard Specifications for Precision Electroformed Sieves ($w = 5\,500$ mm, $\Delta w = \pm 2$ mm)

German standards for screens made from natural and synthetic fibers are:

- DIN 4195: Screening surfaces; woven filament cloth of silk or man-made fibers
- DIN 4197: Screening surfaces; woven filament cloth of round man-made monofilaments; dimensions

Ready-to-use screens made of polyurethane have become increasingly important. These are the so-called system screens. They are available with slot openings down to $w \approx 100\,\mu$m. The advantage of this type of screen is its resistance to abrasion and the fact that it can be changed quickly. Because of the flexibility of the material, it is easy to clean.

Perforated plastic sheets are used on screening devices, that bend or stretch the mesh. Coarse mesh plastic screens with steel reinforcement are also available. Gratings and bars are used for very coarse screening.

For selection of screen surfaces, catalogues of wire-cloth manufacturers should be consulted. There are different edge preparations and hook strips available for fastening the screens with optimum tension.

5.1. Cleaning of Screens

Near-mesh-size material hinders screening by temporarily jamming meshes; it can even stop the process by jamming meshes permanently (see Fig. 11 and Eq. 21). Especially on rigid screens, such as perforated sheets with a mesh size less than 1 mm, screening rapidly comes to a standstill. The jamming force between glass beads and a square-mesh plastic wire cloth with $d = w$ has been measured as

$$F_{\mathrm{K}} = 0.01 E w^2 \tag{34}$$

where E is the modulus of elasticity of the wire cloth. To overcome such forces, acceleration of the order of $10^6\ g$ would be necessary; this cannot be realized. Reliance on the flexibility and deformation of the screen cloth is much easier. Thus about 50 % of the meshes can be kept clean because of the equilibrium between blinding and cleaning [21].

The simplest method of keeping the mesh free is to have or deliberately add coarse

material. Only about 10 – 30 % of coarse material ($1.5 < x/w < 2$), is needed to increase the throughput of near-mesh particles ($0.9 < x/w < 1.1$) by a factor of about 50. This leads to inverted screening with increasing mesh size.

Additional means of cleaning screens are the impact of bouncing rubber balls and the use of brushes, as shown in Figure 37. Roll brushes are better than flat ones because they can reach more deeply into the mesh of fine screens.

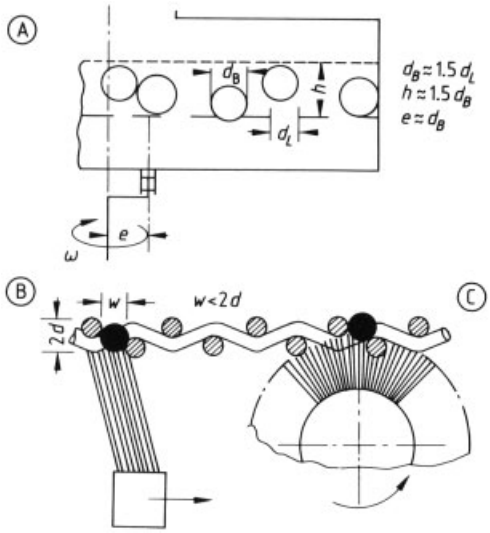

Figure 37. Screen cleaning
A) Rubber-ball, $d_B \approx 1.5\, d_L$, $h \approx 1.5\, d_B$, $e \approx d_B$; B) Flat brush; C) Roll brush

The latest method to prevent blinding is the application of high-energy ultrasonics (see Section 2.1.2). This is effective in screening ceramic and metal powders with mesh $\lesssim 500\,\mu$m. Using rectangular instead of square meshes is effective too. The ratio of the opening should be $w_a/w_b = 1.5$ or 2.5.

5.2. Electrostatics in Screening

The motion of particles in the bed results in shifting and separating of electrostatic charges. At the points of contact, opposite charges remain. Because of low conductivity, an electrostatic potential arises, resulting in an adhesive or attractive stress in the bed. Compared with van der Waals forces, electrostatic forces have a much greater range. Thus screening is hindered strongly or even prevented completely.

Experiments with plastic powder have shown that screening is possible for particles $> 50\,\mu$m by applying air jets or ionizing the surrounding air, e.g., by corona discharge. Screening of particles $< 50\,\mu$m can be improved only by raising the electrical conductivity or the humidity of the air. Up to 90 % R.H., no negative influence on flow behavior has been observed [23].

6. Safety Requirements

In the design of a screening machine, the danger of dust explosions must be considered; some necessary precautions are reducing mechanical friction, avoiding generation of sparks, grounding to prevent the buildup of static charges, and blanketing the space around the screen with nitrogen or argon, as is done in the screening of magnesium powder. The screen can also be placed inside a pressure vessel. Finally, an automatic extinguishing system can be installed which sprays an extinguishing fluid when triggered by a sensor.

Concerning environmental protection, dust removal with cyclone separators or fabric filters is usual. Absorption of vibration and noise is often necessary for vibrating screens operating at higher frequencies.

7. References

1. G. Tarjan, "Separation by Particle Size: Sizing on Screens," in *Mineral Processing,* **vol. 1**, Akademiai Kiada, Budapest 1981, pp. 434 – 497.
2. W. Beenken, "Method for the Mathematical Modelling of the Screening Process on Circular Vibrating Screens," *Aufber.-Tech* **31** (1990) 147 – 156.
3. K. Wieghardt, *Ing. Arch.* **20** (1952) 109 – 115.
4. M. Coppers: *Ultraschallsiebung,* Thesis Univ. Essen, 1997
5. N. Standish, "A Study of the Effect of Operating Variables on the Efficiency of a Vibrating Screen" *Powder Technol.* **48** (1986) 161 – 172.
6. N. Standish, "The Kinetics of Batch Screening," *Powder Technol.* **41** (1985) 57 – 67.

7. N. Standish, *Proc. Australas. Inst. Min. Metall.* **290** (1985) May, 59 – 66.
8. J. E. English, "A New Approach to the Theoretical Treatment of the Mechanics of Sieving and Screening," *Filtr. Sep.* **11** (1974) 195 – 203.
9. G. K. N. S. Subasinghe, W. Schapp, E. G. Kelly, "Modelling Screening as a Conjugate Rate Process," *Int. J. Min. Proc.* **28** (1990) 289 – 300.
10. R. B. Hudson, *Powder Technol.* **2** (1968/69) 229 – 240.
11. P. Andrejczak, P. Wodzinski, "Modell of Screening in the Layer," *Powder Handling and Processing* **6** (1994) 263 – 272.
12. P. Andrejczak, P. Wodzinski, "Screen Design by Means of Outflow Functions," *Powder Hand. and Proc.* **8** (1996) 259 – 265.
13. H. Lange: *Dimensionsanalytische Behandlung des Wurfsiebens,* Thesis Univ. Essen, 1997.
14. F. Mogensen, "A New Method of Screening Granular Material," *The Quarry Managers J.* (1965) 409 – 414.
15. T. Kanzleiter, "Die Taumelsiebmaschine, Maschinendynamik und Guttransport," *Aufbereit. Tech.* **12** (1971) 119 – 127, 380 – 390.
16. G. Tunicliff, *Powder Handl. Proc.* **6** (1994) 273 – 276.
17. H. Schmidt, *Aufbereit. Tech.* **18** (1977) 327 – 332.
18. U. Schmidt, J. Wessel, *Aufbereit. Tech.* **17** (1976) 467 – 470.
19. W. Beenken, "The Processing Significance of Thick Layer Screening for Reduction of Screen Size," *Aufbereit. Tech.* **30** (1989) 727 – 734.
20. P. Schmidt, *Aufbereit. Tech.* **26** (1985) 393 – 398.
21. A. Wignjosaputro, *Fortschr. Ber. VDI Z. Reihe 3,* **106** (1985).
22. F. J. Wesselbaum, *"Siebmaschinen mit verformenden Sieben,"* Dissertation Univ. Essen, 1982.
23. H. J. Smigerski, *Chem.-Ing.-Tech.* **54** (1982) MS 983/82.

Elutriation

See also → Solid – Solid Separation, Introduction. → Air Classifying (Air Elutriation) is a separate keyword.

YORAM ZIMMELS, Technion IIT, Haifa, Israel

1. Introduction 3
2. Scope 4
2.1. Historical Aspects 4
2.2. Extended Scope 4
2.2.1. Type of Force Field 4
2.2.2. Particle-Size Distribution 4
2.2.3. Mechanism of Elutriation 4
2.2.4. Range of Concentration 4
2.2.5. Elutriation in the Context of Physical Separation Processes 5
3. Theory 5
3.1. Hindered Sedimentation of Monodisperse Particles 5
3.2. Hindered Sedimentation of Polydisperse Particles 6
3.3. Effect of Aggregation 7
3.3.1. Monodisperse Aggregates 7
3.3.2. Polydisperse Aggregates 8
3.4. Composite Particulates 8
3.5. Particle Flux 8
3.6. Other $f(\varphi)$ Correlations for the Creeping-Flow Regime 10
3.7. Particle-Size Distribution 10
3.8. Performance Criteria 12
3.9. Sharpness Index 14
3.10. Effect of Flow Profiles 14
3.10.1. Counterflow System 14
3.10.2. Cross-Flow System 18
4. Elutriation Systems 19
4.1. Categories of Elutriation Systems 19
4.2. Counterflow Gravity Elutriators . 20
4.3. Centrifugal Elutriators 24
4.4. Axial Cross-Flow Elutriators . . . 27
4.5. Electromagnetic Elutriators 28
5. References 28

Symbols

Symbol	Name
A	area
a_1, a_2, a_3	constants used for evaluation of U_φ
B	normalization factor of the beta distribution
b', b_1, b_2	constants of distribution functions
C	constant
$C_\mathrm{D}, C_{\mathrm{D}\varphi}, C_{\mathrm{D}0}$	drag coefficient: general, extended to include the effect of particle concentration (i.e., φ), and at $\varphi = 0$, respectively
D	normalized sum of squares
d, d_i, d_j	particle size: general of ith particle, and of jth particle, respectively
d_s	Stokes diameter
d_0, d_m	average particle size and maximum size that exist in the beta distribution
E_c, E_f	total "coarse" and "fine" efficiencies
F_D	hydrodynamic drag force
f_p	force per unit volume
$f(\varphi)$	correlation function for dependence of U_p on φ
$f_1(\varphi)$	function of φ used to evaluate dependence of flux on φ
$G_\mathrm{c}(x_0), G_\mathrm{f}(x_0)$	"coarse" and "fine" cumulative grade efficiencies
g	gravitational acceleration
g'	apparent acceleration of particle set in motion by force field
$g_\mathrm{c}(x), g_\mathrm{f}(x)$	"coarse" and "fine" specific grade efficiencies
h	depth
J, J_i, J_m	volume flux: total value, value of ith size fraction, and total maximum value
$J_\mathrm{u}, J_\mathrm{d}$	"upward" and "downward" volume flux of particles
k	integer or a constant
L	distance along pipe axis; pool length
m	mass fraction; also used as an integer
m_e	value of mass fraction m that maximizes $f_1[\varphi(m)]$
N	size of sample
N_A	Avogadro constant
n	exponent; also used as integer
n_i	number of observed measurements that fall in the ith interval of a sample; hydration number

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_16

P_i probability evaluated from a known distribution

P_0, P_L pressure at reference position and at distance L downstream, respectively

Q_f, Q_u, Q_d flow rate: total, upstream, and downstream, respectively

$q(x), q'(x)$ particle flux densities

$q_u(x), q_d(x)$ upstream and downstream particle flux densities

R Rosin – Rammler cumulative distribution function; also inner radius of a pipe

r radial distance from axis of rotation; axis of a pipe; or parameter of distribution functions

r' $= t - r$, parameter of the beta distribution, or radial distance

r_e elutriation radius

Re, Re_φ Reynolds number and extended definition in which the effect of particle content (i.e., φ) is accounted for

Re_0 Reynolds number at $\varphi = 0$

$\tilde{S}$ cumulative sharpness index

s_1 cross section of particle taken perpendicular to direction of flow

s parameter, also specific sharpness index

t parameter of a distribution

t' parameter of a distribution

U_d average velocity of dispersion relative to the container

U_{da}, U_{dm} average and maximum flow velocities in a pipe

U_f average velocity of fluid which is displaced by motion of particles relative to the dispersion frame of reference

U_φ, U_p average velocity of particles, relative to the frame of reference of the fluid in which they are dispersed and relative to the frame of reference of the dispersion, respectively

$U_{\varphi}{}^h$ velocity of hydrated particles with respect to the free-flowing fluid

$U_{\varphi i}, U_{pi}$ average velocity of ith particles relative to the fluid and to the dispersion frame of reference, respectively

U_0 velocity of a particle at infinite dilution, i.e., $\varphi \to 0$

U'_p velocity of particle relative to the container

U'_{pi} velocity of ith particles with respect to the container frame of reference

$U_p(x_0)$ velocity of particles having size x_0 relative to the dispersion

$U'_{pa}(i, j)$ velocity of particles in i, jth aggregate relative to the container frame of reference

$U_\varphi(x), U_\varphi(x_0)$ velocities of particles having size x and x_0 relative to the fluid, respectively

$U_{\varphi}{}^a(i, j)$ velocity of i, jth aggregate relative to the fluid at φ_a

V_a, V_s, V_f volume of a single aggregate, of particles, and of fluid in the aggregate, respectively

V_h, V_{h0} volume of hydrated and unhydrated particles, respectively

v molar volume

W width

W_0, W_c, W_f mass of feed, coarse, and fine products respectively

w constant

x, x_0 independent variable of distributions; also a parameter

x_m, x_s variable denoting large and small particle size, respectively

X_φ independent variable in the Barnea – Mizrahi correlation, $(Re^2{}_\varphi C_{D\varphi})^{1/3}$

x_{25}, x_{75}, particle sizes used to evaluate the sharpness

x_{10}, x_{90} index; subindex denotes the corresponding value of the cumulative distribution function

x_{50+w}, x_{50-w} particle size used to evaluate the specific index of sharpness, $0 < w < 50$

x'_{50+w}, x'_{50-w} particle size used to evaluate cumulative index of sharpness, $0 < w < 50$

y $= \ln d$

Y_φ dependent variable in the Barnea – Mizrahi correlation, $(Re_\varphi / C_{D\varphi})^{1/3}$

y_0 $= \ln d_0$, the expectation of the log of normal size distribution

z variable of the standard normal distribution

Greek Symbols

α parameter in distributions; also used to denote an angle

β parameter (see Harris equation); also used to denote an angle

$\Gamma(r)$ gamma function

γ parameter (see Harris equation)

δ, δ_a particle – dispersion and aggregate–dispersion density difference, respectively
$\delta_a(i, j)$ i, jth aggregate – dispersion density difference
δ_+, δ_- variables having a value of 0 or 1 that are used to define up- and downstream flux densities
ε porosity
μ_f, μ_φ, μ_{φ^a} viscosity of fluid, of dispersion at φ, and of dispersion containing aggregates at φ_a, respectively
ξ_i variable denoting φ_i/φ; also a coordinate system
ϱ_f, ϱ_p, ϱ_φ density of fluid, particles, and dispersion, respectively
σ_y standard deviation of the logarithmic normal size distribution
φ, φ_i volume fraction occupied by particles: total volume and volume of ith particle-size fraction, respectively
φ_a total volume fraction occupied by monodisperse aggregates in dispersion
φ_f volume fraction occupied by fluid
φ_h volume fraction of hydrated particles
φ_s, φ_{sm} volume fraction occupied by particles within an aggregate and the maximum value of φ_s dictated by packing density considerations
φ_{sj} volume fraction occupied by particles in jth aggregate fraction
$\varphi(i, j)$ volume fraction occupied (in dispersion) by particles pertaining to the i, jth aggregate
$\varphi_h(i, j)$ volume fraction occupied (in dispersion) by hydrated particles
$\varphi_s(i, j)$ volume fraction occupied by particles in the i, jth aggregate
$\chi^2_{n-1,\,1-\alpha}$ parameter of the chi-square distribution with $n-1$ degrees of freedom and level of confidence $1-\alpha$
ψ_{LN} probability density function of the logarithmic normal distribution
$\psi(x)$, $\psi(z)$ probability density function of x and z, respectively
$\psi_0(x)$, $\psi_c(x)$, distribution functions characterizing the
$\psi_f(x)$ feed, coarse, and fine products, respectively
$\Psi(x)$, $\Psi(z)$ cumulative distribution functions
$\Psi'_c(x)$, $\Psi_c(x)$ Harris cumulative distribution functions
$\Psi_{GM}(x)$, Gaudin – Meloy, Gates – Gaudin
$\Psi_{GGS}(x)$, $\Psi_s(x)$ Schumann and S-shaped cumulative distribution functions
$\psi_{RR}(x)$ Rosin – Rammler distribution function
ω angular velocity

1. Introduction

Elutriation is a process whereby a polydisperse mixture of particles can be separated into distinct fractions by means of the combined action of a fluid stream and one or more external force fields. The drag force exerted by the fluid on particles that are smaller and less susceptible to the field deflects them (i.e., elutriates them) from trajectories of other, larger, more susceptible particles which tend to follow more closely trajectories dictated by the external force field. The term *elutriation* is sometimes used when the fluid stream is a liquid (wet classifying); with gas streams the operation is called *air classifying* (→ Air Classifying).

2. Scope

2.1. Historical Aspects

Elutriation and air classifying have traditionally been recognized as reverse sedimentation in which particles move against a current of fluid (usually water or air) that subsequently, by one or more process steps, effects separation and grading of the particle mixture. Elutriation has mostly been applied to fine particles in dilute dispersions that follow the Stokes sedimentation law. In this context, flows in elutriation systems were limited to the laminar regime. The force fields were traditionally of the acceleration type, i.e., gravitational or centrifugal, with the fluid stream originating from sources external to the dispersion.

2.2. Extended Scope

The extended scope that is used here under the title elutriation is in accordance with the definition given above [1] and includes the following elements:

2.2.1. Type of Force Field

No limitation is put on either the type of force field or its distribution in space. Forces may originate from potential and vector (i.e., kinetic) fields that are external to the system, e.g., from gravitational, centrifugal, electromagnetic, and hydrodynamic fields. Forces may also originate internally due to thermal and chemical potential fields, as in diffusion systems.

2.2.2. Particle-Size Distribution

Elutriation has been used traditionally for treatment of fine-particle distributions. Here the finer size fraction is separated from the rest of the distribution by the fluid stream. Elutriation has been applied generally to particles larger than a few micrometers. Under the extended definition used in this article, and in view of the mechanism of elutriation (see Section 2.2.3), submicron particles, colloids, and the large molecules encountered in diffusion are also included.

2.2.3. Mechanism of Elutriation

Elutriation can arise due to internally or externally driven flows:

1) internally counterflowing fluid which is displaced by motion of the particle mixture, or
2) counter- and cross-flowing fluid which is supplied from sources external to the dispersion.

The first mechanism of elutriation is inherent in cases where motion of the dispersed particles is set by external force fields other than hydrodynamic. Here, slower parts of the distribution are partially or fully elutriated counter to other parts that contain faster particles. This mechanism depends on the properties of the dispersion and can be manipulated only by changing them. Its importance increases for dispersions with a finer particle-size distribution.

The second mechanism is better known, but its applicability is limited by particle size. In contrast to internal flows, the external flow can be controlled readily to achieve desirable separation.

2.2.4. Range of Concentration

Elutriation is usually carried out in dilute dispersions. This requires the use of a large ratio of fluid volume to dispersed-phase volume. In this article, no limits are set on the concentration of the dispersion, with the understanding that extremes should be avoided for either economical or technological reasons. When the concentration increases, the effect of displaced fluid on elutriation is enhanced, as is the saving of fluid per unit dispersed phase treated. On the other hand, interparticle hindrance may adversely affect elutriation efficiency, so that an optimum concentration can be expected.

2.2.5. Elutriation in the Context of Physical Separation Processes

Elutriation can have a partial or full role in the separation of particles in flow fields. Thus, classification of particles can be achieved by an elutriation process in a separation zone, followed by auxiliary methods for dispersing and removing the particle stream. Internal flows result in the segregation of particles which, combined with external forces, can be used for density and electromagnetic separation.

3. Theory

3.1. Hindered Sedimentation of Monodisperse Particles

Two kinds of particle velocity can be distinguished: the average steady-state velocity of particles relative to the fluid that has been displaced by their motion, U_φ, and the average steady-state velocity of particles relative to the frame of reference of the whole dispersion, U_p.

In this case, φ is the total volume fraction occupied by particles in the dispersion. The fluid drag on particles depends on U_φ, whereas particle flux depends on U_p. The velocity of the displaced fluid relative to the dispersion frame of reference is referred to as U_f.

$$U_\mathrm{f} = U_\varphi \varphi \tag{1}$$

$$U_{\mathrm{p}}=U_{\varphi}-U_{\mathrm{f}}=U_{\varphi}\,(1-\varphi) \tag{2}$$

$$\varphi+\varphi_{\mathrm{f}}=1 \tag{3}$$

where φ_{f} is the volume fraction occupied by the fluid; φ_{f} is equivalent to the porosity ε. As shown below, available correlations facilitate the evaluation first of U_{φ} and then of U_{p}.

Figure 1 shows the generalized correlation of Barnea and Mizrahi [1] for hindered sedimentation of monodisperse spherical particles in Newtonian fluids. This plot obviates the need to resort to the well-known family of concentration-dependent plots $(Re/C_{\mathrm{D}})^{1/3}$ vs. $(Re^2\,C_{\mathrm{D}})^{1/3}$, which are traditionally used for calculation of velocity vs. diameter of particles. In the Barnea and Mizrahi plot, the Reynolds number Re is replaced by Re_{φ} and the drag coefficient C_{D} by $C_{\mathrm{D}\varphi}$; Re_{φ} and $C_{\mathrm{D}\varphi}$ are given by Equations (4) and (6), respectively.

$$Re_{\varphi}=\frac{U\varphi d\varrho_{\mathrm{f}}}{\mu_{\varphi}}=Re_0\frac{U_{\varphi}/U_0}{\exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right)} \tag{4}$$

where d is the particle size, ϱ_{f} is the density of the fluid and subindex 0 denotes here $\varphi=0$ (i.e., $U_0=U_{\varphi}\,(\varphi=0)$) and the extended viscosity μ_{φ} is given by

$$\mu_{\varphi}=\mu_{\mathrm{f}}\exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right) \tag{5}$$

with μ_{f} being the viscosity of the pure fluid.

$$C_{\mathrm{D}\varphi}=\frac{F_{\mathrm{D}}}{\frac{1}{2}\,\varrho_{\mathrm{f}}S_1U_{\varphi}^2\,\left(1+\varphi^{1/3}\right)}$$
$$=C_{\mathrm{D}0}\left(\frac{U_0}{U_{\varphi}}\right)^2\cdot\frac{(1-\varphi)}{\left(1+\varphi^{1/3}\right)} \tag{6}$$

where F_{D} is drag force, ϱ_{f} is fluid density, S_1 is the cross section of the particle perpendicular to the direction of flow and $(1+\varphi)^{1/3}$ is the particle-interaction equivalent of the wall-hindrance effect. The Barnea – Mizrahi correlation applies to a wide range of φ including the high concentrations encountered in distended beds, and also to the full range of flow regime encountered practically in particle sedimentation and fluidization. This correlation can be described by Equation (7 a) for the creeping-flow regime, and Equation (7 b) for the intermediate- and turbulent-flow regimes, respectively [2].

$$C_{\mathrm{D}\varphi}=24/Re_{\varphi} \tag{7a}$$

$$(\log Y_{\varphi}+2.563)^2+(\log X_{\varphi}-4.267)^2=18.775 \tag{7b}$$

$$Y_{\varphi}=\left(Re_{\varphi}/C_{\mathrm{D}\varphi}\right)^{1/3},\ X_{\varphi}=\left(Re_{\varphi}^2C_{\mathrm{D}\varphi}\right)^{1/3}$$

Note that Equation (7 a) corresponds to the straight line, in the creeping-flow regime, in the log – log plot of Y_{φ} vs. X_{φ}, see Figure 1.

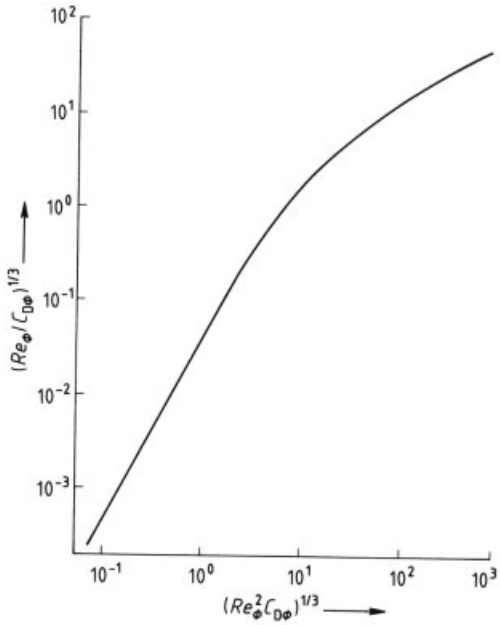

Figure 1. Barnea – Mizrahi correlation for hindered sedimentation of monodisperse spherical particles, $(Re_{\varphi}/C_{\mathrm{D}\varphi})^{1/3}$ vs. $(Re_{\varphi}^2\,C_{\mathrm{D}\varphi})^{1/3}$

With the results shown in Figure 1 and Equations (7 a) and (7 b), U_{φ} can be evaluated directly by Equation (8) for the creeping- and by Equation (9) for the general-flow regimes [2], [3]:

$$U_{\varphi}=\frac{(\varrho_{\mathrm{p}}-\varrho_{\mathrm{f}})\,g'd^2}{18\mu_{\mathrm{f}}}f\,(\varphi)\,,$$
$$f\,(\varphi)=\frac{(1-\varphi)}{\left(1+\varphi^{1/3}\right)\exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right)} \tag{8}$$

$$U_{\varphi}=\left(\frac{-a_2+\sqrt{a_2^2+4a_1a_3^{1/2}}}{2a_1}\right)^2 \tag{9}$$

$$a_1 = 0.63\,(d\varrho_\mathrm{f})^{1/2}$$
$$a_2 = 4.80\mu_\mathrm{f}^{1/2}\exp\left(\frac{5}{6}\cdot\frac{\varphi}{1-\varphi}\right)$$
$$a_3 = 4.0\,(\varrho_\mathrm{p}-\varrho_\mathrm{f})\,(1-\varphi)\,\mathrm{g}'d^2/\left[3\left(1+\varphi^{1/3}\right)\right]$$

where d is the particle diameter, ϱ_p the density of the particles, and g' the apparent acceleration of the particle in the driving force field:

$$g = f_\mathrm{p}/\delta\prime \tag{10}$$

where f_p is the force per unit volume of particle and δ is given by

$$\delta = \varrho_\mathrm{p} - \varrho_\varphi \tag{11}$$

$$\varrho_\varphi = \varphi\varrho_\mathrm{p} + (1-\varphi)\,\varrho_\mathrm{f} \tag{12}$$

For gravitational and centrifugal fields, $g' = g$ and $g' = \omega^2 r$, respectively, where g is gravitational acceleration, ω is angular velocity, and r is distance from the axis of rotation.

3.2. Hindered Sedimentation of Polydisperse Particles [2–4]

Elutriation is generally applied to polydisperse mixtures. In a mixture consisting of n size fractions,

$$\varphi = \sum_{i=1}^{n}\varphi_i \tag{13}$$

where φ_i is the volume fraction occupied by the ith size fraction of particles having size d_i. Equation (2) applies to each size fraction:

$$U_{\mathrm{p}i} = U_{\varphi i} - U_\mathrm{f} \tag{14}$$

where U_f is given by:

$$U_\mathrm{f} = \sum_{i=1}^{n} U_{\varphi i}\varphi_i \tag{15}$$

If an external source of fluid sets the whole dispersion in motion and the dispersion gains the velocity U_d relative to the dispersion container, then the velocity of particles with respect to the container, $U'_{\mathrm{p}i}$, becomes

$$U'_{\mathrm{p}i} = U_{\varphi i} - U_\mathrm{f} - U_\mathrm{d} \tag{16}$$

Particles will be elutriated (i.e., carried with the stream) if $U'_{\mathrm{p}i} < 0$; otherwise they will counterflow the stream and perhaps leave it because of external force fields if $U'_{\mathrm{p}i} > 0$. They will remain in the elutriation zone if $U'_{\mathrm{p}i} = 0$. Thus the cut size (for definition of cut size → Particle Size Analysis and Characterization of a Classification Process, Chap. 3.1.4.) is determined by

$$U_\mathrm{d} = U_{\varphi i} - U_\mathrm{f} = U_{\varphi i} - \sum_{i=1}^{n} U_{\varphi i}\varphi_i \tag{17}$$

At nonvanishing values of φ, the particle velocity $U_{\varphi i}$ and the particle volume fraction φ_i are functions of the size distribution and of the total volume fraction occupied by particles in the dispersion; the same is true for the cut point of the elutriation process.

Particles that are elutriated by internal fluid current (i.e., no external fluid supply is required to reverse their motion against the driving force field) are defined in this article as self-elutriated particles. Alternatively, "self-elutriation" implies that the system elutriates part of its distribution by internal hydrodynamic drag arising from the remainder of the distribution. The cut point under conditions of self-elutriation is determined simply by letting $U_d = 0$; then

$$U_{\varphi i} = U_\mathrm{f} \tag{18}$$

3.3. Effect of Aggregation [2], [3]

3.3.1. Monodisperse Aggregates

Aggregation involves the formation of rigid structures of particles that adhere to each other. Aggregates generally consist of particles, entrained fluid, and a binding material.

The aggregation of particles to form a new particulate (i.e., the aggregate) has the following effects:

1) Increase in the size of the new particulate, compared to the size of the particles from which it was formed
2) Decrease of the average density (magnetic susceptibility, etc.), due to entrainment of fluid, compared to nonaggregated particles
3) Increase of the volume fraction occupied by the particulates

4) Change in shape of the dispersed particulates
5) Change of distribution characteristics of the dispersed particulates
6) Increased probability of interparticle interaction
7) Increased probability of variation in distribution during motion, particularly where shear forces are involved

Aggregation can occur spontaneously, for example, due to existing electrostatic or van der Waals interactions, or as a result of deliberate changes (such as introduction of additives) in the composition of the dispersion. Its effect is important in fine-particle systems, where air elutriation (→ Air Classifying) cannot be applied because of the tendency of the particles to aggregate electrostatically. Wet elutriation facilitates the use of dispersants or flocculants (→ Flocculation) to prevent or enhance aggregation. The use of such agents becomes critically important in emulsions and colloidal systems. Although aggregation is generally believed to enhance sedimentation (see item 1 above), situations may exist in which the counteracting factors (see items 2 – 6) prevail. Let the volume V_a of a single aggregate contain particles and entrained fluid having volumes V_s, and V_f, respectively

$$V_\mathrm{a}=V_\mathrm{s}+V_\mathrm{f} \tag{19}$$

$$V_\mathrm{s}=\varphi_\mathrm{s}V_\mathrm{a},\ V_\mathrm{f}=(1-\varphi_\mathrm{s})\,V_a \tag{20}$$

where φ_s is the volume fraction occupied by particles within the aggregate. Let φ_a be the total volume fraction occupied by monodisperse aggregates in the suspension:

$$\varphi_\mathrm{a}=\varphi/\varphi_\mathrm{s},\ \varphi_\mathrm{s}\leq\varphi_\mathrm{sm},\ \varphi_\mathrm{a}\leq\varphi_\mathrm{sm} \tag{21}$$

where $\varphi_{\mathrm{s}m}$ is the maximum volume fraction that can be occupied by spherical particles in the aggregate, which here is assumed to be determined by the maximum packing density of spherical particles. Thus $\varphi\leq\varphi_\mathrm{sm}^2$, and the upper boundary of φ which is 0.6 – 0.7 in a nonaggregated dispersion decreases to 0.36 –0.49 in monodisperse spherically aggregated systems. The aggregate – suspension density difference δ_a and the apparent suspension viscosity $\mu_{\varphi\mathrm{a}}$ are given by

$$\delta_\mathrm{a}=(\varphi_\mathrm{s}-\varphi)\,(\varrho_\mathrm{p}-\varrho_\mathrm{f}) \tag{22}$$

$$\mu_{\varphi_\mathrm{a}}=\mu_\mathrm{f}\exp\left(\frac{5}{3}\cdot\frac{\varphi}{\varphi_\mathrm{s}-\varphi}\right) \tag{23}$$

Equations (22) and (23) show that the density difference (see Eq. 11) decreases and viscosity increases (see Eq. 5) compared to nonaggregated systems at the same φ.

Equations (8) and (9) can be used to evaluate U_{φ^a} of aggregates, where

$$f\,(\varphi_\mathrm{a})=\frac{\varphi_\mathrm{s}-\varphi}{\left[1+(\varphi/\varphi_\mathrm{s})^{1/3}\right]\exp\left(\frac{5}{3}\cdot\frac{\varphi}{\varphi_\mathrm{s}-\varphi}\right)} \tag{24}$$

a_1 remains the same; a_2 and a_3 are given by

$$a_2=4.80\mu_\mathrm{f}^{1/2}\exp\left(\frac{5}{6}\cdot\frac{\varphi}{\varphi_\mathrm{s}-\varphi}\right) \tag{25}$$

$$a_3=4.0\,(\varrho_\mathrm{p}-\varrho_\mathrm{f})\,(\varphi_\mathrm{s}-\varphi)\,\mathrm{g}'d^2/\left\{3\left[1+(\varphi/\varphi_\mathrm{s})^{1/3}\right]\right\}$$

3.3.2. Polydisperse Aggregates

In this section, aggregates are described which vary in size and packing density, i.e., in φ_s. If the aggregate mixture consists of *n* size fractions and *m* φ_s fractions, and each fraction denoted (*i*, *j*) is characterized by its size d_i and by $\varphi_{\mathrm{s}j}$,

$$\delta_\mathrm{a}\,(i,j)=[\varphi_\mathrm{s}\,(i,j)-\varphi]\,(\varrho_\mathrm{p}-\varrho_\mathrm{f}) \tag{26}$$

$$\varphi_\mathrm{a}=\sum_{i=1}^{n}\sum_{j=1}^{m}\varphi\,(i,j)\,/\varphi_\mathrm{s}\,(i,j)\,,$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m}\varphi\,(i,j)=\varphi \tag{27}$$

Equation (28) is the counterpart of Equation (16) for polydisperse aggregates:

$$U'_{\mathrm{p_a}}\,(i,j)=U_{\varphi_\mathrm{a}}\,(i,j)\,-$$
$$\sum_{i=1}^{n}\sum_{j=1}^{m}U_{\varphi_\mathrm{a}}\,(i,j)\cdot\varphi\,(i,j)\,/\varphi_\mathrm{s}\,(i,j)-U_\mathrm{d} \tag{28}$$

where $U_{\varphi_\mathrm{a}}(i,j)$ is calculated from Equation (8) or (9) in conjunction with Equations (24) and (25).

3.4. Composite Particulates

In contrast to aggregates, a composite particulate is defined here as a multicomponent entity subject to the condition that each component is continuous. This excludes the existence of separate particles or the formation of discontinuous islands which characterize aggregated systems. A simple illustration is the example of coated particles in which the core particle and its coating consist of continuous phases. Coated particles are important in solutions and colloidal systems, for example, hydrated species in aqueous solutions. Here, the volume V_{h} of the hydrated particle is given by [5]

$$V_{\mathrm{h}}=V_{\mathrm{h0}}+nv/N_{\mathrm{A}} \tag{29}$$

where $n, v,$ and N_{A} are the hydration number, the molar volume of the hydrating fluid, and the Avogadro constant, respectively. The new volume fraction φ_{h} of the hydrated particles becomes

$$\varphi_{\mathrm{h}}=\varphi\left(1+\left(nv/N_{\mathrm{A}}\right)/V_{\mathrm{h0}}\right) \tag{30}$$

To evaluate $U_{\varphi_{\mathrm{h}}}$ in monodisperse systems, φ is replaced by φ_{h} in Equations (8) and (9). In polydisperse hydrated particles which may vary in their initial size and also in aggregation number, $\varphi_{\mathrm{h}}\,(i, j)$ should be defined and used as indicated in Section 3.3.

This procedure applies also to collodial systems such as ferrofluids [6], i.e., magnetic fluids consisting of colloidal magnetic particles, in which ferromagnetic particles (size 5 – 10 nm) are prevented from aggregation by the steric stabilization that is imparted to them by a dispersant coating. Elutriation of larger composite particles can be applied to the separation of catalysts deposited on substrates and to coated magnetic and pharmaceutical particles.

3.5. Particle Flux

The volume flux J_i of the ith size fraction being elutriated in a fluid of velocity U_{d} is given by

$$J_i=U'_{\mathrm{p}i}\varphi_i=\left(U_{\varphi i}-U_{\mathrm{f}}-U_{\mathrm{d}}\right)\varphi_i \tag{31}$$

Let the fraction j be characteristic of the cut size. This means that particles of diameter d_j are neither lifted, nor do they sink, in the elutriation stream. Applying Equation (17) gives

$$U_{\mathrm{d}}=U_{\varphi j}-U_{\mathrm{f}} \tag{32}$$

hence

$$J_i=\left(U_{\varphi i}-U_{\varphi j}\right)\varphi_i \tag{33}$$

$d_i < d_{i+1}$ for $i = 1, 2, \ldots, n-1$, the "upward" J_{u} and "downward" J_{d} total fluxes are given by Equations (34) and (35), respectively.

$$J_{\mathrm{u}}=\sum_{i=1}^{j}\left(U_{\varphi i}-U_{\varphi j}\right)\varphi_i \tag{34}$$

$$J_{\mathrm{d}}=\sum_{i=j+1}^{n}\left(U_{\varphi i}-U_{\varphi j}\right)\varphi_i \tag{35}$$

Note that "upward" means that particles are carried by the elutriation stream and "downward" means the reverse. The total flux J is given by

$$J=J_{\mathrm{u}}+J_{\mathrm{d}} \tag{36}$$

The flux J is a property of the particle distribution and does not depend on the velocity of the elutriation stream. Next, the conditions for maximum flux are determined. If φ_i is expressed as a fraction ξ_i of φ, then

$$\varphi_i=\xi_i\varphi \tag{37}$$

The following equation can be used to evaluate J in the creeping-flow regime:

$$J=\frac{\left(\varrho_{\mathrm{p}}-\varrho_{\mathrm{f}}\right)g'}{18\mu_{\mathrm{f}}}f_1(\varphi)\sum_{i=1}^{n}\left|d_i^2-d_j^2\right|\xi_i \tag{38}$$

$$f_1(\varphi)=\varphi(1-\varphi)\Big/\left[\left(1+\varphi^{1/3}\right)\exp\left(\frac{5}{3}\cdot\frac{\varphi}{1-\varphi}\right)\right]$$

In deriving Equation (38), use was made of Equations (8), (36), and (37). At a given particle-size distribution, J is solely a function of φ. Figure 2 shows a plot of $f_1(\varphi)$ vs. φ; $f_1(\varphi)$ has a maximum of 66.8744×10^{-3} at $\varphi = 0.218$. This value of φ is higher than the one ($\varphi = 0.1792$) that maximizes the sedimentation flux of monodisperse particles in the creeping-flow regime [2]. The volume or mass flux can be expressed as a function of the total mass fraction m of particles in the dispersion:

$$m=\frac{\varphi\varrho_{\mathrm{p}}}{\varphi\varrho_{\mathrm{p}}+(1-\varphi)\varrho_{\mathrm{f}}} \tag{39}$$

Rearranging Equation (39) gives

$$\varphi = \frac{m}{\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}} - m\,(\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}} - 1)}$$

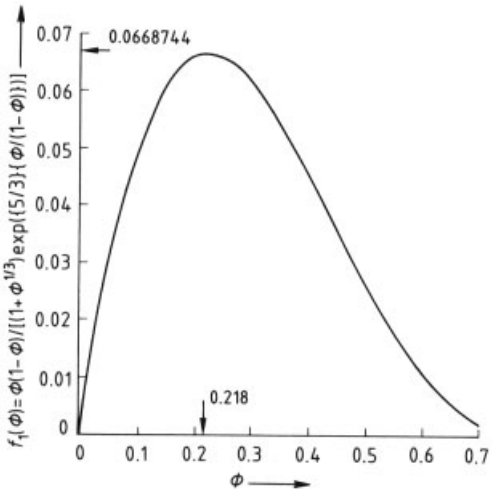

Figure 2. Plot of the flux volume concentration function $f_1(\varphi)$ vs. φ

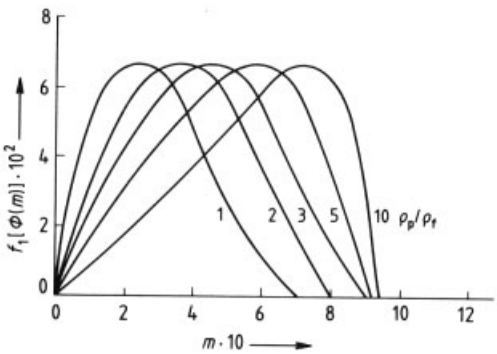

Figure 3. Plot of the flux mass concentration function $f_1[\varphi(m)]$ vs. m

Figure 3 shows a plot of $f_1\,[\varphi\,(m)]$ vs. m for different values of $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}}$. The curve $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}} = 1$ is identical to that of Figure 2. Figure 4 shows a plot of m_{e} vs. $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}}$, which was derived from Figure 3; m_{e} is the value of m that maximizes $f_1\,[\varphi\,(m)]$ (i.e., the flux at a given value of $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}}$). As expected, increasing $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}}$ shifts the optimum mass fraction toward higher values. For example, $\varphi = 0.218$ at $\varrho_{\mathrm{p}}/\varrho_{\mathrm{f}} = 3$, $m_{\mathrm{e}} = 0.455$; thus, the total maximum flux becomes

$$J_{\mathrm{m}} = 66.8744 \times 10^{-3}\,\frac{(\varrho_{\mathrm{p}} - \varrho_{\mathrm{f}})\,\mathrm{g}'}{18\mu_{\mathrm{f}}}\cdot \sum_{i=1}^{n} \left|d_i^2 - d_j^2\right| \xi_i \tag{40}$$

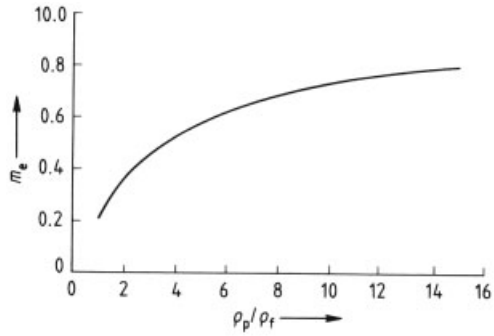

Figure 4. Plot of m values that maximize the flux m_{e} vs. $\varrho_p/\varrho_{\mathrm{f}}$

Elutriation is often performed in dilute suspensions. This is believed to increase the sharpness of separation. However, the costs in terms of the low throughput due to the low flux can be substantial. An example is the elutriation of particles with $\varrho_{\mathrm{p}} = 3000$ kg/m^3 in water with $\varrho_{\mathrm{f}} = 1000$ kg/m^3. If elutriation is performed at $m = 0.05$, then the flux is only 19.56 % of the maximum, i.e., at $m_{\mathrm{e}} = 0.455$. In other words, five times the amount of water is required compared to the demand had the elutriation been performed at the optimum point. This means larger investment in equipment and higher operating cost.

In summary, unless technical problems occur in working at the recommended values of φ or m, operating closer to the optimal values of these process variables should be considered.

3.6. Other $f(\varphi)$ Correlations for the Creeping-Flow Regime

Equation (41) gives the empirical correlation [7], [8] published by Richardson and Zaki:

$$f\,(\varphi) = (1-\varphi)^n \tag{41}$$

This correlation generally applies to the evaluation of U_{p} of monodisperse particles, which is given in the form

$$U_{\mathrm{p}} = U_{\mathrm{p0}} f\,(\varphi)$$

According to Equation (2), $n-1$ should be used instead of n when applying $f(\varphi)$ to U_φ. Thus, in creeping-flow elutriation where $n = 5.1$ fits the data of several investigations [9],

$$f(\varphi) = (1-\varphi)^{4.1} \tag{42}$$

The discrepancy between the two correlations is apparent at lower values of φ, with the one given by Equation (8) falling off faster at $\varphi < 0.2$; the reverse is true at $\varphi > 0.2$ [7]. A summary of available models for $f(\varphi)$ is given in [1] and [7]. Here, the model which led to Equation (8) is preferred, in view of the large body of experimental data that support it.

3.7. Particle-Size Distribution

Equations (17) and (18), which give the cut size, and Equations (34)–(40), which give particle flux, show that the outcome of elutriation depends on particle-size distribution. Thus, information regarding size distribution is necessary for analysis of elutriation systems. In the following, some useful size distributions are summarized [10–14] (→ Particle Size Analysis and Characterization of a Classification Process, Chap. 2.3.).

Logarithmic Normal Distribution. The logarithmic normal distribution in a standard form is given by

$$\psi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, z = \frac{y - y_0}{\sigma_y} \tag{43}$$
$$y = \ln d,\ y_0 = \ln d_0,\ d \geq 0$$

where $\ln d_0$ is the expectation, (i.e., weighted mean) of the random variable $y = \ln d$; σ_y is the standard deviation given in units of $\ln d$. The cumulative logarithmic normal distribution is

$$\Psi(Z \leq z) = \int_{-\infty}^{z} \psi_{\mathrm{LN}}(z)\,\mathrm{d}z \tag{44}$$

where LN indicates logarithmic normal; $\Psi(Z \leq z)$ is available in standard tables [11], [15].

Thus, being given a cut size d of particles that have a logarithmic normal size distribution facilitates evaluation of the fraction expected to be elutriated. For example, if $d_0 = 50 \times 10^{-6}$ m, $y_0 = \ln d_0 = -9.9035$, and $\sigma_y = 1$, then for a cut size of 40×10^{-6} m, $y = \ln(40 \times 10^{-6}) = -10.1266$; $z = -0.2231$. Using tables of standard normal distribution gives $\Psi(Z \leq z) = 0.4117$; i.e., 41.17 % of this distribution will be elutriated. The logarithmic normal distribution is useful for the approximate description of skewed distributions.

Rosin–Rammler Distribution. The following equation gives the Rosin–Rammler probability density distribution function $\psi_{\mathrm{RR}}(x)$:

$$\psi_{\mathrm{RR}}(x) = 100 n b' x^{n-1} \exp\left(-b' x^n\right), \quad x \geq 0 \tag{45}$$

The cumulative function R is given by

$$R = \int_{-\infty}^{x} \psi_{\mathrm{RR}}(x)\,\mathrm{d}x = 100 \exp\left(-b' x^n\right), \quad x \geq 0 \tag{46a}$$

$$\log\log(100/R) = n \log x + \text{constant},\ x \geq 0 \tag{46b}$$

where n and b' are constants, and b' is a measure of the range of particle size: high values of b' indicate a narrow distribution and vice versa; n is characteristic of the material under consideration. The size for which $R = 36.8$ % is used to characterize the degree of size reduction of the material. This distribution applies to broken, crushed, or ground material, such as broken coal and fan spray droplets [16].

Gates–Gaudin–Schumann Distribution [17], [18]. The cumulative Gates–Gaudin–Schumann distribution is given by

$$\Psi_{\mathrm{GGS}}(x) = (b_1 x)^n, x \geq 0 \tag{47}$$

where b_1 and n are constants.

Gaudin–Meloy Distribution [19]. The Gaudin–Meloy cumulative distribution is given by

$$\Psi_{\mathrm{GM}}(x) = 1 - (1 - b_2 x)^n, \quad x \geq 0 \tag{48}$$

where b_2 and n are constants.

S-Shaped Cumulative Distributions [20]. Equation (49) can be used to fit S-shaped cumulative data

$$\Psi_{\mathrm{s}}(x) = \frac{e^{\alpha x} - 1}{e^{\alpha x} + e^{\alpha} - 2}, x > 0 \tag{49}$$

The probability density function corresponding to $\Psi_{\mathrm{s}}(x)$ is given by

$$\psi_{\mathrm{s}}(x) = \frac{\alpha e^{\alpha x}(e^{\alpha} - 1)}{(e^{\alpha x} + e^{\alpha} - 2)^2} \tag{50}$$

Increase of α increases the steepness of $\Psi_{\mathrm{s}}(x)$ and, hence, decreases the standard deviation of $\psi_{\mathrm{s}}(x)$.

Gamma Distribution [11], [12]. The probability density function of the gamma distribution is given by

$$\psi(x) = \frac{\alpha}{\Gamma(r)}(\alpha x)^{r-1} e^{-\alpha x}, x > 0, r > 0 \tag{51}$$

where α and r are constants. If r is a positive integer, then

$$\Psi(x) = 1 - \sum_{k=0}^{r-1} e^{-\alpha x}(\alpha x)^k / k! \tag{52}$$

and $\Gamma(r)$ is the gamma function

$$\Gamma(r) = \int_0^{\infty} x^{r-1} e^{-x} \mathrm{d}x, \; r > 0 \tag{53}$$

If r is a positive integer, $\Gamma(r) = (r-1)!$. The gamma distribution presents a family of curves that includes the exponential distribution ($r = 1$) and the chi-square distribution ($r = n/2$ and $\alpha = 1/2$).

Beta Distribution. The probability density function of the beta distribution is given by

$$\psi(x) = \frac{1}{B} x^{r-1}(1-x)^{t-r-1}, \quad 0 \le x \le 1, r > 0, t - r > 0 \tag{54}$$

$$B = \Gamma(r)\Gamma(t-r)/\Gamma(t)$$

and if r and $t - r$ are integers,

$$B = (r-1)!(t-r-1)!/(t-1)!$$

The value of the beta distribution lies in the variety of shapes that it can be made to fit by varying the parameters r and t. For example, the beta distribution is skewed right if $r < t/2$ and left if $r > t/2$. It is bell-shaped if $r > 1$ and $t > r + 1$. Interchange of parameters $r' = t - r$ and $t' = t$ yields mirror images. Because $0 \le x \le 1$, normalized data must be used for particle size, such as $x = d/d_{\mathrm{m}}$, where d_{m} is the maximum size that occurs in the distribution.

Harris Equation [21], [22]. Harris suggested that most of the well-known particle-size distributions are special cases of Equation (55):

$$\psi(x) = cx^{\alpha}\left[1 - \left(\frac{x}{x_0}\right)^{\beta}\right]^{x_0^{\beta}\gamma} \tag{55}$$

where c, α, β, γ, and x_0 are parameters.

For example, if $\alpha = 0$, $\beta = 2$, and $x_0 \to \infty$, the normal distribution is obtained, whereas $\alpha = n - 1$, $\beta = n$, and $x_0 \to \infty$ yields the Rosin–Rammler distribution. If $\alpha = s - 1$, $x_0^{\beta}\gamma = r - 1$, and $\beta = s$, a three-parameter equation [23] results which, on integration and rearrangement, gives

$$\Psi'_{\mathrm{c}}(x) = 1 - \Psi_{\mathrm{c}}(x) = \left[1 - \left(\frac{x}{x_0}\right)^{s}\right]^{r} \tag{56}$$

where $\Psi'_{\mathrm{c}}(x)$ is the cumulative oversize distribution, x_0 is the maximum size in the sample, s and r are parameters concerned with the shape of the log–ln plot in the fine and coarse ends of the distribution. A convenient form of Equation (56) is given by

$$log\left\{ln\left[\Psi'_{\mathrm{c}}(x)\right]^{-1}\right\} = \log r + \log\left\{\ln\left[1 - (x/x_0)^s\right]^{-1}\right\} \tag{57}$$

Equations (56) and (57) are flexible and can be fitted to most unimodal distributions. Equation (57) has been used to evaluate results from scanning centrifugal sedimentation instruments for particle-size analysis.

Fitting Known Distributions to Data Obtained from Samples. Procedures to test whether a known distribution (such as the ones listed above) can be taken to describe a given population, by using data obtained from samples of this population, are outlined in standard statistics texts [10], [11]. This procedure involves evaluating the sum

$$D=\sum_{i=1}^{n}\frac{(n_i-NP_i)^2}{NP_i}$$

and testing if

$$D<\chi^2_{n-1,1-\alpha}$$

where n_i is the number of observed measurements that fall in a given interval, $N=\sum_{i=1}^{n} n_i$, and P_i is the probability evaluated from the known distribution that the result of the measurement will fall in the ith interval; $\chi^2_{n-1,1-\alpha}$ is the chi-square variable having $n-1$ degrees of freedom and a level of confidence $1-\alpha$. Usually, $1-\alpha$ is higher than 90 %.

If $D<\chi^2_{n-1,1-\alpha}$, the sample is assumed to be part of the tested known distribution at level of confidence $1-\alpha$. The known distribution that provides the lowest value of D is preferable.

The wealth of information concerning distributions, especially particle-size distributions, provides sound mathematical description of particulate populations and, hence, prediction of their motion in elutriation system.

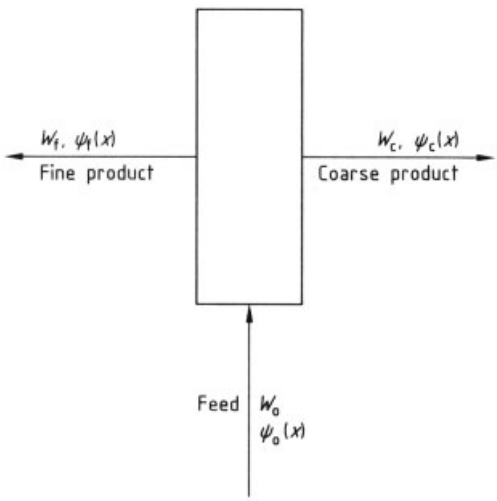

Figure 5. Mass balance around a single elutriation step

3.8. Performance Criteria

Figure 5 shows a single elutriation step that involves feed, as well as coarser and finer product streams. Let $\psi(x)$ and $\Psi(x)$ denote probability density and cumulative distribution functions, and subindexes 0, c, and f the feed, coarser, and finer product streams, respectively. Equations (58) and (59) give total and fractional mass balances, in terms of mass W and probability density functions $\psi(x)$, respectively

$$W_0=W_c+W_f \tag{58}$$

$$W_0\psi_0(x)=W_c\psi_c(x)+W_f\psi_f(x) \tag{59}$$

The total "coarse" and "fine" efficiencies are given by

$$E_c=W_c/W_0,\ E_f=W_f/W_0=1-E_c \tag{60}$$

The "coarse" and "fine" specific grade efficiencies, $g_c(x)$ and $g_f(x)$, are given by

$$g(x)=(W_c/W_0)\,\psi_c(x)/\psi_0(x)\,c$$
$$=E_c\psi_c(x)/\psi_0(x) \tag{61}$$

$$g(x)=(W_f/W_0)\,\psi_f(x)/\psi_0(x)\,f$$
$$=E_f\psi_f(x)/\psi_0(x) \tag{62}$$

Dividing Equation (59) by $W_0\,\psi_0(x)$ and comparing with Equations (61) and (62) give

$$g(x)=1-g_f(x)\,c \tag{63}$$

The grade efficiencies of Equations (61) and (62) are evaluated at x and thus refer to a specific size. The grade efficiencies of sizes coarser or finer than x_0 may be of interest. In this case, the cumulative grade efficiency $G(x)$ is defined as follows:

$$G_c(x_0)=E_c\frac{\Psi_c(x>x_0)}{\Psi_0(x>x_0)}=E_c\frac{1-\Psi_c(x_0)}{1-\Psi_0(x_0)} \tag{64}$$

$$G_f(x_0)=E_f\frac{\Psi_f(x_0)}{\Psi_0(x_0)} \tag{65}$$

Note that by definition $\Psi(x_0)=\Psi(x<x_0)$, hence, $\Psi(x>x_0)=1-\Psi(x_0)$. Figure 6 shows a plot [14] of $\Psi_c(x)$ vs. $\Psi_0(x)$ over a plot of $\Psi_f(x)$ vs. $\Psi_0(x)$ which was developed for calculation of $g_c(x)$ and $g_f(x)$ [24]; because $\psi(x)=d\Psi(x)\,dx$, Equation (61) can be written as

$$g(x)=E_c\,d\Psi_c(x)/d\Psi_0(x)=E_c(\tan\alpha)_x\,c$$

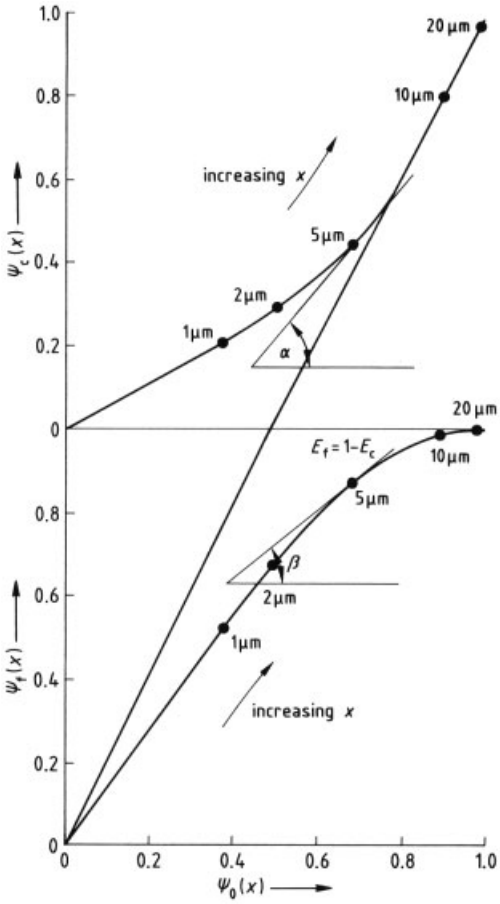

Figure 6. Plot of cumulative distribution
Oversize or coarse fraction $\Psi_c(x)$ appears directly above undersize or fine fraction $\Psi_f(x)$; both are plotted against size of feed $\Psi_0(x)$

The largest particles of the distribution (20 µm in Fig. 6) are expected to be completely transferred to the coarser fraction. Hence, for large x, (e.g., x_m) $g_c(x) = 1$ and $(\tan\alpha)_{x_m} = 1/E_c$. Thus the tangent to the $\Psi_c(x)$ vs. $\Psi_0(x)$ curve at $\Psi_c(x) = 1$ intersects the $\Psi_0(x)$ axis at a point $\Psi_0(x) = 1 - E_c$. Now $g_c(x)$ can be evaluated by using $(\tan\alpha)_x$ for different values of x (in Fig. 6 the tangent to the curve at 5 µm is shown). Consider the lower plot of Figure 6: the smallest particles are expected to be completely transferred to the finer fraction; hence, for sufficiently small x (e.g., x_s), $g_f(x) = 1$ and $(\tan\beta)_{x_s} = 1/E_f = 1/(1 - E_c)$. Now $g_f(x)$ can be evaluated by using $(\tan\beta)_x$ for different values of x. Distribution curves of feed, fines, and coarse fractions are evaluated elsewhere (→ Particle Size Analysis and Characterization of a Classification Process, Chap. 2.2.). For more details and numerical examples see [14] and [25].

3.9. Sharpness Index

Elutriation performance can also be characterized by the specific sharpness index s.

$$s = \frac{x_{(50+w)}}{x_{(50-w)}} \qquad 0 < w < 50 \tag{66}$$

where x_{50+w} and x_{50-w} are particle sizes defined by $g(x_{(50+w)}) = 50 + w$ and $g(x_{(50-w)}) = 50 - w$, respectively, and w is an arbitrary constant usually with a value $w = 25$ or $w = 40$.

The cumulative sharpness index S is defined by

$$S = \frac{x'_{(50+w)}}{x'_{(50-w)}} \qquad 0 < w < 50 \tag{67}$$

where $x'_{(50+w)}$ and $x'_{(50-w)}$ are defined by

$$G\left(x'_{(50+w)}\right) = 50 + w, G\left(x'_{(50-w)}\right) = 50 - w$$

The sharpness of cut obtained by the elutriator increases as either s or S approach 1, which is their lower bound that characterizes an ideal separation system.

3.10. Effect of Flow Profiles

3.10.1. Counterflow System

Figure 7 shows uniform and parabolic flow profiles in a tube. The uniform profile corresponds to nonviscous (i.e., potential) flow; the parabolic profile to laminar viscous flow in a circular tube. A flatter profile is expected to provide higher grade efficiencies and a sharpness of cut index closer to 1. Curved flow profiles introduce a position-dependent flow velocity $U_d = U_d(\xi_1, \xi_2, \xi_3)$, where ξ_i ($i = 1$, 2, and 3) is a coordinate system used to describe a steady flow. Different types of flow can be used to effect elutriation. For example, axial flow in circular or rectangular conduits, flow in channels, layer or film flow, and flow around objects such as matrices used

to capture particles by electromagnetic fields. In the following material, laminar flow in a circular tube and film flow are considered.

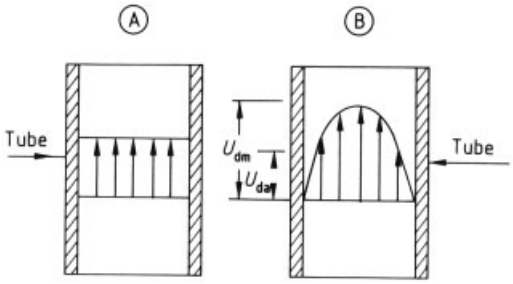

Figure 7. Flow profiles in a tube
A) Uniform velocity profile; B) Parabolic velocity profile

Steady Laminar Incompressible Flow in a Circular Tube. The flow profile is given by the Poiseuille equation [26]:

$$U_{\mathrm{d}}=\frac{(P_0-P_L)\,R^2}{4\mu_{\mathrm{f}}\,L}\left[1-\left(\frac{r}{R}\right)^2\right] \tag{68}$$

where R is the radius of the tube; r is radial distance from the tube axis; L is distance along the axis; and P_0 and P_L are the pressures at a reference position and at a distance L downstream from this position, respectively.

The flow rate Q and average velocity U_{da} are given by

$$Q_{\mathrm{f}}=\frac{\pi\,(P_0-P_L)\,R^4}{8\mu_{\mathrm{f}}\,L} \tag{69}$$

$$U_{\mathrm{da}}=\frac{(P_0-P_L)\,R^2}{8\mu_{\mathrm{f}}\,L} \tag{70}$$

The maximum velocity U_{dm} occurs at $r=0$; hence,

$$U_{\mathrm{dm}}=\frac{(P_0-P_L)\,R^2}{4\mu_{\mathrm{f}}\,L}=2U_{\mathrm{da}} \tag{71}$$

By using Equation (68),

$$r^2/R^2=1-U_{\mathrm{d}}/\,(2U_{\mathrm{da}}) \tag{72}$$

Combining Equations (72) and (17) gives the cut size d_i (via $U_{\varphi i}$) obtained at r with the average velocity U_{da}:

$$r_e^2/R^2=1-\left(U_{\varphi i}-\sum_{i=1}^{n}U_{\varphi i}\varphi_i\right)/\,(2U_{\mathrm{da}}) \tag{73}$$

Thus, if U_{da} is large enough, an elutriation radius r_{e} exists such that when $r<r_{\mathrm{e}}$, particles smaller than d_i (with settling velocities lower than U_{d}) will flow upstream. Outside r_{e} (i.e., for $r>r_{\mathrm{e}}$), particles having sizes smaller than d_i will still settle against the elutriation stream. Equation (73) also shows that if

$$U_{\mathrm{d}}=U_{\varphi i}-\sum_{i=1}^{n}U_{\varphi i}\varphi_i=0$$

then $r=R$ and no loss of $d<d_i$ particles is expected. If U_{da} is given, the fractional loss (i.e., relative to the whole distribution) of particles smaller than d_i which are not elutriated, is expected to maximize at infinite dilution (i.e., at $\varphi\to 0$). In this case, for any cutsize d_i, $U_d=U_{\varphi i}$ must be satisfied.

In the following material, grade efficiencies in elutriators characterized by nonuniform flow profiles are evaluated. Let the elutriator feed be characterized by a continuous distribution $\psi_0\,(x)$. The particle size is denoted by x, as before. The volume fraction occupied by particles having a size between x and $x+\mathrm{d}x$ is $\varphi\psi_0\,(x)\,\mathrm{d}x$, and the corresponding flux density $q'\,(x)$ is given by

$$q'\,(x)=U'_{\mathrm{p}}\varphi\psi_0\,(x)\,\mathrm{d}x \tag{74}$$

where U'_{p} is a function of x and of position across the flow stream.

$$U'_{\mathrm{p}}=U_{\mathrm{d}}-U_{\varphi}+\int_0^{\varphi}U_{\varphi}\mathrm{d}\varphi \tag{75}$$

The steady flux $q\,(x)$ of particles of size x consists of upstream, $q_{\mathrm{u}}\,(x)$, and downstream, $q_{\mathrm{d}}\,(x)$, fluxes.

$$q\,(x)=q_{\mathrm{u}}\,(x)+q_{\mathrm{d}}\,(x) \tag{76}$$

The upward and downward flux is given by the integral over the area A:

$$q_{\mathrm{u}}\,(x)=\int_0^{A}\delta_+\psi_0\,(x)\,\mathrm{d}xU'_{\mathrm{p}}\cdot\mathrm{d}A,$$

$$\delta_+=\frac{1}{2}\left(1+\frac{U'_{\mathrm{p}}}{|U'_{\mathrm{p}}|}\right) \tag{77}$$

$$q_{\mathrm{d}}(x) = \int_0^A \delta_- \psi_0(x)\,\mathrm{d}x U'_{\mathrm{p}} \cdot \mathrm{d}A,$$

$$\delta_- = \frac{1}{2}\left(1 - \frac{U'_{\mathrm{p}}}{|U_{\mathrm{p}}|}\right) \tag{78}$$

The expression $\psi_0(x)\,\mathrm{d}x$ is a constant and can be taken outside the integrals of Equations (77) and (78). If $U'_{\mathrm{p}} > 0$ is associated with dA, then $\delta_+ = 1$ and $\delta_- = 0$; hence, for the given size x, dA is characterized by $q_{\mathrm{u}}(x) > 0$ and $q_{\mathrm{d}}(x) = 0$. If $U'_{\mathrm{p}} < 0$, then $\delta_+ = 0$ and $\delta_- = 1$, and $q_{\mathrm{u}}(x) = 0$ and $q_{\mathrm{d}}(x) > 0$ characterize dA, as required. Dividing Equations (77) and (78) by Equation (76), with upstream and downstream fractions being the finer (f) and coarser (c) fractions respectively, gives

$$\mathrm{g_f}(x) = \frac{\int_0^A \delta_+ U'_{\mathrm{p}} \cdot \mathrm{d}A}{\int_0^A \delta U'_{\mathrm{p}} \cdot \mathrm{d}A} \tag{79}$$

$$\mathrm{g_c}(x) = \frac{\int_0^A \delta_- U'_{\mathrm{p}} \cdot \mathrm{d}A}{\int_0^A \delta U'_{\mathrm{p}} \cdot \mathrm{d}A} = 1 - \mathrm{g_f}(x) \tag{80}$$

where $\delta = \delta_+ - \delta_-$.

The respective cumulative grade efficiencies are given by

$$G_{\mathrm{f}}(x_0) = \frac{\int_0^{x_0}\int_0^A \delta_+ U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x}{\int_0^{x_0}\int_0^A \delta U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x} \tag{81}$$

$$G_{\mathrm{c}}(x_0) = \frac{\int_{x_0}^{\infty}\int_0^A \delta_- U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x}{\int_{x_0}^{\infty}\int_0^A \delta U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x} \tag{82}$$

where $G_{\mathrm{f}}(x_0) = G_{\mathrm{f}}(x \le x_0)$ and $G_{\mathrm{c}}(x_0) = G_{\mathrm{c}}(x > x_0)$.

The total upstream (Q_{u}) and downstream (Q_{d}) particle fluxes are obtained by integrating Equations (77) and (78) over the complete size range.

$$Q_{\mathrm{u}} = \int_0^{\infty}\int_0^A \delta_+ U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x \tag{83}$$

$$Q_{\mathrm{d}} = \int_0^{\infty}\int_0^A \delta_- U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x \tag{84}$$

Dividing Equation (77) by Equation (83) and Equation (78) by Equation (84) gives d$\Psi_{\mathrm{f}}(x)$ and d$\Psi_{\mathrm{c}}(x)$; division by dx yields $\psi_{\mathrm{f}}(x)$ and $\psi_{\mathrm{c}}(x)$, respectively.

$$\psi_{\mathrm{f}}(x) = \frac{\psi_0(x)\left(\int_0^A \delta_+ U'_{\mathrm{p}} \cdot \mathrm{d}A\right)}{\int_0^{\infty}\int_0^A \delta_+ U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x} \tag{85}$$

$$\psi_{\mathrm{c}}(x) = \frac{\psi_0(x)\left(\int_0^A \delta_- U'_{\mathrm{p}} \cdot \mathrm{d}A\right)}{\int_0^{\infty}\int_0^A \delta_- U'_{\mathrm{p}} \cdot \mathrm{d}A \psi_0(x)\,\mathrm{d}x} \tag{86}$$

Exampe of Evaluation of Grade Efficiencies. Let the size distribution of an elutriator feed and the expected cut size be $\psi_0(x)$ and x_0, respectively. The elutriation is performed with a uniform dilute dispersion in a vertical circular tube for which Equations (68) – (73) apply. The average elutriation flow velocity U_{da} is set so as to satisfy Equation (87):

$$U_{\mathrm{da}} = kU_{\mathrm{p}}(x_0) = k\left[U_\varphi(x_0) - U_{\mathrm{f}}\right],\ 0 < k < \infty \tag{87}$$

where k is an adjustable constant. For example, if particles of size x_0 are to have a settling velocity equal to U_{da}, then $k = 1$. However, if particles ten times larger than x_0 must have this velocity, then $k = \sqrt{10}$. In a dilute dispersion and under the assumption that $U_{\mathrm{f}} = \int_0^{\varphi} U_\varphi d_\varphi \ll U_\varphi(x_0)$, the approximation $U_{\mathrm{da}} = kU_\varphi(x_0)$ is justified. Combining Equations (68), (70), (75), and (87) (where U_{f} is neglected) gives

$$U'_{\mathrm{p}} = 2kU_\varphi(x_0)\left(1 - (r/R)^2\right) - U_\varphi(x) \tag{88}$$

By means of this equation, r' is now defined so that the following two ranges are formed: $r < r'$ where $U'_{\mathrm{p}} > 0$, $\delta_+ = 1$, $\delta_- = 0$; and $r > r'$ where $U'_{\mathrm{p}} < 0$, $\delta_+ = 0$, $\delta_- = 1$.

$$r' = R\left[1 - \frac{1}{2k} U_\varphi(x)/U_\varphi(x_0)\right]^{1/2} \tag{89}$$

$$g_{\mathrm{f}}(x)=\frac{\delta_+\int_0^{r'}\left\{2kU_\varphi(x_0)\left[1-(r/R)^2\right]-U_\varphi(x)\right\}2\pi r\mathrm{d}r}{\delta_+\int_0^{r'}\left\{2kU_\varphi(x_0)\left[1-(r/R)^2\right]-U_\varphi(x)\right\}2\pi r\mathrm{d}r-\delta_-\int_{r'}^{R}\left\{2kU_\varphi(x_0)\left[1-(r/R)^2\right]-U_\varphi(x)\right\}2\pi r\mathrm{d}r}$$

Integration over r and rearrangement give

$$g_{\mathrm{f}}(x)=\frac{\delta_+\left[2kU_\varphi(x_0)-U_\varphi(x)\right]\pi r'^2-\pi kU_\varphi(x_0)\,r'^4/R^2}{\delta_+\left\{\left[2kU_\varphi(x_0)-U_\varphi(x)\right]\pi r'^2-\pi kU_\varphi(x_0)\,r'^4/R^2\right\}-\delta_-\left\{\left[2kU_\varphi(x_0)-U_\varphi(x)\right]\pi\left(R^2-r'^2\right)-\pi kU_\varphi(x_0)\left(R^4-r'^4\right)/R^2\right\}} \tag{90}$$

Consider a simple case where $\psi_0(x)$ exists only if x satisfies creeping-flow sedimentation (i.e., Eq. 8 applies) otherwise $\psi_0(x)$ vanishes. Applying Equation (8) for $\varphi=0$, $f(\varphi)=1$ [hence, $U_\varphi(x)/U_\varphi(x_0)=(x/x_0)^2$] and rearranging yield

$$g_{\mathrm{f}}(x)=\frac{\delta_+\left[2k-(x/x_0)^2-k\,(r'/R)^2\right]}{\delta_+\left[2k-(x/x_0)^2-k\,(r'/R)^2\right]-\delta_-\left[\left(2k-(x/x_0)^2\right)\left((r'/R)^2-1\right)-k\left((r'/R)^{-2}-(r'/R)^2\right)\right]} \tag{91}$$

Substitution of Equation (88) in Equation (79), with $\mathrm{d}A=2\pi r\,\mathrm{d}r$ gives

$$g_{\mathrm{c}}(x)=1-\mathrm{g}_{\mathrm{f}}(x)\,,\ r'=R\left[1-\frac{1}{2k}\left(\frac{x}{x_0}\right)^2\right]^{1/2}$$

Equation (91) shows that $g_{\mathrm{f}}(x)$ and $g_{\mathrm{c}}(x)$ are functions of x/x_0 and of k. If $x/x_0<\sqrt{2k}$, then $g_{\mathrm{f}}(x)>0$ and $g_{\mathrm{c}}(x)<1$. However, if $x/x_0\geq\sqrt{2k}$, then $U'_{\mathrm{p}}\leq 0$, $\delta_+=0$, $\delta_-=1$, and $g_{\mathrm{f}}(x)=0$, $g_{\mathrm{c}}(x)=1$, as expected. Thus the size range that can be split in two product fractions by elutriation is $0\leq x\leq x_0\sqrt{2k}$. Outside this range, no splitting exists. Assume that $x<x_0\sqrt{2k}$; hence, in Equation (91), $\delta_+=\delta_-=1$, and $g_{\mathrm{f}}(x)$ can be reduced further to

$$g_{\mathrm{f}}(x)=\frac{\left[k-\frac{1}{2}\left(\frac{x}{x_0}\right)^2\right]^2}{\left[k-\frac{1}{2}\left(\frac{x}{x_0}\right)^2\right]^2+\frac{1}{4}\left(\frac{x}{x_0}\right)^4} \tag{92}$$

Figure 8 shows a plot of $g_{\mathrm{f}}(\mathrm{x})$ vs. $(x/x_0)\,k^{-1/2}$. By using this figure, the following typical index of sharpness is obtained: $x_{25}/x_{75}=1.29$, $x_{10}/x_{90}=1.76$. Figure 8 can be used in the following way. Suppose that the average particle size of a particle distribution is $x_0=30\times10^{-6}$ m. If $k=1$, then for the size fraction $x=x_0$, $(x/x_0)\,k^{-1/2}=1$; hence, from Figure 8, $g_{\mathrm{f}}(x_0)=0.5$. If $g_{\mathrm{f}}(x_0)\geq 0.9$, then from Figure 8, $k^{-1/2}\leq 0.7$ and $k\geq 2.041$. If particles twice as large as x_0 (i.e., $x/x_0=2$) are to satisfy $g(x)\geq 0.9$, then $0.7\,k^{1/2}\geq 2$, and $k\geq 8.163$. By using the following data, $\varrho_{\mathrm{p}}=2700$ kg/m^3, $\varrho_{\mathrm{f}}=1000$ kg/m^3, $\mu_{\mathrm{f}}=0.001$ kg m^{-1} s^{-1}, and $g=9.81$ m/s^2, then $U_{\mathrm{da}}=k\,U_\varphi(x_0)\geq 2.27\times10^{-4}$ m/s is obtained. In dilute dispersions, grade efficiencies are practically independent of particle-size distribution. This is not true for the size distribution of elutriation products (see Eqs. 85 and 86).

In a creeping-flow regime

$$r'^2=R^2\left[1-\frac{1}{2k}\left(\frac{x}{x_0}\right)^2\right],$$

thus using Equations (85) and (88), and canceling common terms in the numerator and denominator yields

$$\psi_{\mathrm{f}}(x)=\frac{\psi_0(x)\int_0^{r'}2kx_0^2\left[1-(r'/R)^2-x^2\right]2\pi r\mathrm{d}r}{\int_0^{\sqrt{2k}x_0}\int_0^{r'}2kx_0^2\left[1-(r^2/R)^2-x^2\right]2\pi r\mathrm{d}r\,\psi_0(x)\,\mathrm{d}x} \tag{93}$$

where $x\leq\sqrt{2k}x_0$

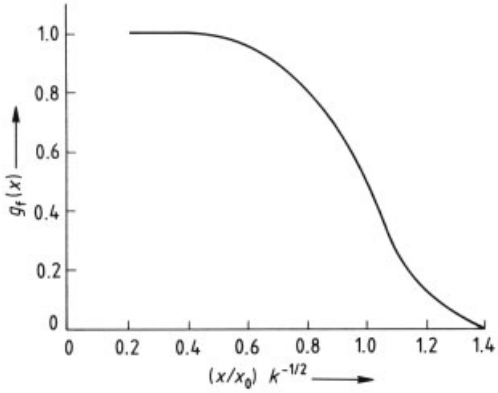

Figure 8. Plot of $g_f(x)$ vs. $(x/x_0)\,k^{-1/2}$ calculated from Equation (92)

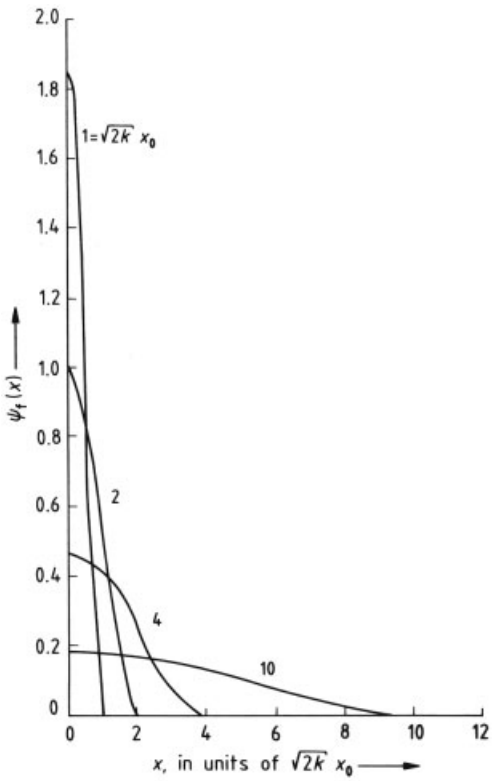

Figure 9. Plot of $\psi_f(x)$ vs. x calculated from Equation (94) for $2k \cdot x_0 = 1, 2, 4,$ and 10

Consider a simple case where $\psi_0(x)$ is uniform in a given size range $a < x < b$; hence, $\psi_0(x) = 1/(b-a)$. If, for example, $b = 40 \times 10^{-6}$ m and $a = 20 \times 10^{-6}$ m, then $\psi_0(x) = 10^6/20$. In this case, by integrating Equation (93) and performing some simple algebraic manipulations,

$$\psi_f(x) = \frac{\left[1 - \frac{1}{2k}\left(\frac{x}{x_0}\right)^2\right]^2}{(8/15)\sqrt{2k} \cdot x_0}, \quad 0 \le x \le \sqrt{2k} \cdot x_0 \qquad (94)$$

Figure 9 shows a plot of $\psi_f(x)$ vs. x for $\sqrt{2k} \cdot x_0 = 1, 2, 4,$ and 10. An increase in $\sqrt{2k} \cdot x_0$ increases the spread of $\psi_f(x)$ vs. x. Thus, to minimize the spread of $\psi_f(x)$ obtained from a uniform $\psi_0(x)$ in a circular-tube elutriator, $\sqrt{2k} \cdot x_0$ must be small, i.e., if the finer end of the distribution should be elutriated. By using a similar procedure, $\psi_c(x)$ can be evaluated.

Note that the tendency of particles to drift toward the elutriator center, due to velocity differentials across their diameter, is still not accounted for [14]. Thus, the results should be considered as applying to an idealized system in which this tendency is not important.

The effect of particle concentration on grade efficiencies and the expected size distributions in the elutriation products is not discussed in the above example; however, it can be deduced readily by following the same calculation procedures.

3.10.2. Cross-Flow System

Figure 10 shows schematically a two-dimensional steady cross-flow elutriation system. Here, a layer of fluid flows on a solid support set perpendicular to the force field. At position A, the dispersion is uniform and $\psi(x)$ is independent of the vertical dimension ξ_3; at position B, $\psi(x) = \psi(x, \xi_3)$ due to differential sedimentation of particles toward the solid wall. In this system, particles of different sizes are partially segregated because of different sedimentation velocities. This effect is intensified by internal elutriation, in which displaced fluid sweeps slower particles against the driving force. Evaluation of $\psi(x, \xi_3)$ can be made by applying a nonsteady sedimentation model [2] for polydisperse particle mixtures. At a given position, the system shown in Figure 10 appears steady to an observer at rest relative to the solid support, i.e., to the ξ_i $(i = 1, 2, 3)$ frame of reference. However, if the observer moves with the fluid, then he observes a nonsteady system characterized by a time-dependent particle-size

distibution which is equivalent to being position-dependent in the ξ_i frame of reference. Let the stream at position B be divided at $\xi_3 = c$ into fine (f) and coarse (c) substreams. Grade efficiency and size distribution for the fine stream are given by Equations (95) and (96), respectively.

$$g_{\mathrm{f}}(x) = \frac{\int_0^c (U_{\mathrm{d}} \cdot \mathrm{d}A)\, \psi(x,\xi_3)\, \mathrm{d}x}{\int_0^c (U_{\mathrm{d}} \cdot \mathrm{d}A)\, \psi(x,\xi_3)\, \mathrm{d}x + \int_c^h (U_{\mathrm{d}} \cdot \mathrm{d}A)\, \psi(x,\xi_3)\, \mathrm{d}x} \tag{95}$$

$$\psi_{\mathrm{f}}(x) = \frac{\int_0^c \psi(x,\xi_3)\, U_{\mathrm{d}} \cdot \mathrm{d}A}{\int_0^\infty \int_0^c \psi(x,\xi_3)\, (U_{\mathrm{d}} \cdot \mathrm{d}A)\, \mathrm{d}x} \tag{96}$$

$$U_{\mathrm{d}} = \frac{3}{2} U_{\mathrm{da}} \left[1 - \left(\frac{\xi_3}{h}\right)^2\right], \quad U_{\mathrm{da}} = \frac{2}{3} U_{\mathrm{dm}}$$

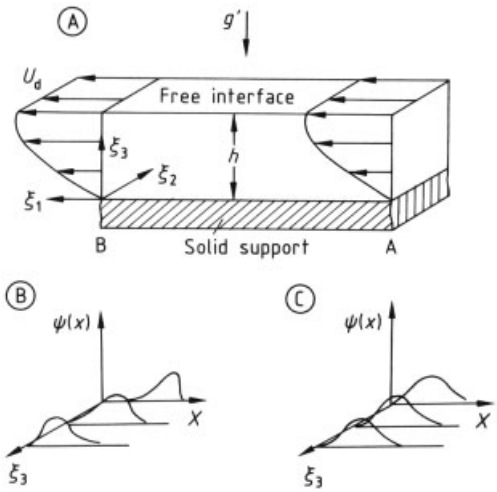

Figure 10. A) Cross-flow elutriation system
The dispersion flows in the form of a layer on a flat solid support. The force field producing accelerations g' is perpendicular to the flow. At position A, $\psi(x)$ is uniform across h; at position B, it is not.
B) Variable $\psi(x)$ and C) Uniform $\psi(x)$ at different position ξ_3

Consider the simple case of a settling pool, shown in Figure 11. The steady fluid flow and settling velocity of particles are assumed to be uniform. The time required for a particle to settle to the bottom of the pool, i.e., from the surface to depth h, is

$$t_h = h/U_0 \tag{97}$$

while the time required to travel the pool length L (i.e., to be elutriated) horizontally is

$$t_L = L/U_{\mathrm{da}} \tag{98}$$

Let Q be the volumetric flow rate of the suspension through the pool. Hence

$$U_{\mathrm{da}} = Q/(Bh) \tag{99}$$

where B is the pool dimension perpendicular to h and L. Setting $t_h = t_L$ means that the particle arrives simultaneously at the bottom and at the end of the pool. For a particle that satisfies this condition, by Equations (97) and (99),

$$U_{\mathrm{p}} = \frac{Q}{BL} \tag{100}$$

Equation (100) shows that U_{p} is independent of the pool depth and is a function only of Q and the pool area. This is the "area principle" of cross-flow elutriation settling pools, which remains valid also for nonuniform flows across the pool such as the one depicted in Figure 10 and described by Equation (96).

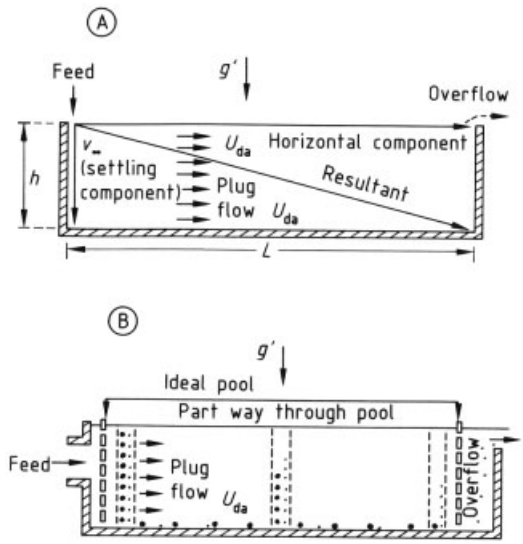

Figure 11. Cross-flow elutriator settling pools
A) Ideal cross-flow and settling velocities; B) Feed containing polydisperse particle mixture

If the feed consists of a polydisperse particle mixture as shown in Figure 11, then settling produces a spectrum of particle sediment along the flow path. Furthermore, the pool flow concept in conjunction with dispersion effects [27–30] has

been used to analyze the performance of classifiers. This is in accordance with the role of cross-flow separation in many modern elutriator classifiers and separators.

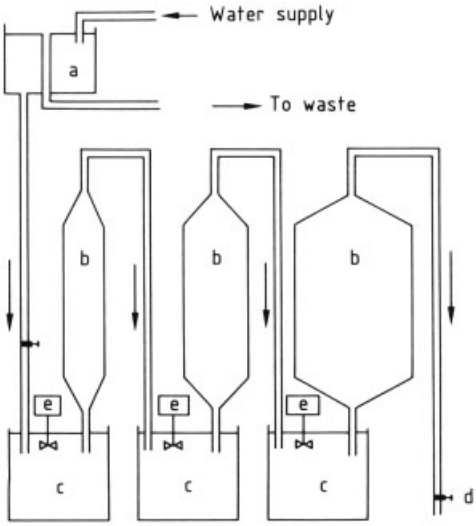

Figure 12. Three-stage Blythe elutriator
a) Constant head apparatus; b) Elutriation chamber; c) Sump; d) Screw clip; e) Stirrer

4. Elutriation Systems

4.1. Categories of Elutriation Systems

Elutriation systems can be classified according to the following categories of flow mechanism:

1) Steady counterflow
2) Steady transverse or cross flow
3) Nonsteady flow

Elutriators are also classified as vertical or horizontal and gas (e.g., air; → Air Classifying) or liquid (e.g., water) operated systems. A *single-stage elutriator* produces coarse and fine products characterized by cumulative oversize and undersize distribution functions. A *multistage elutriation system* can be used for grading into specific, relatively narrow size fractions. If, for example, steady counterflow (of fluid rising against settling particles) is used, then a battery of elutriators connected in series and increasing in diameter will produce progressively finer size fractions. Elutriation systems have been traditionally used for preparation of well-defined size fractions on both a laboratory and an industrial scale. Industrial elutriators are generally used as classifiers, laboratory devices for particle-size analysis and sorting. We shall also use the term "classifiers", along with a clear reference to the elutriation mechanism, when it applies.

4.2. Counterflow Gravity Elutriators

Counterflow elutriators are used for powders in the size range 10 – 200 µm, with densities greater than 2000 kg/m^3. This category includes the Schöne, Andrews, and Blythe elutriators. The Schöne [14], [31–33] and Andrews [14], [34] elutriators are early models of laboratory-size equipment [14], [31]. The Blythe elutriator [35] (Fig. 12) is designed to provide narrow size fractions. Each elutriating chamber, along with its exit tube and sump, is a link in a siphon chain. On-line sampling of each sump facilitates detection of the progress of elutriation. Increasing the diameter of consecutive chambers by a factor of 2 produces size grades that decrease by a factor of $\sqrt{2}$. The Blythe elutriator can be used to fractionate particles as small as 2 – 3 µm, a process that may take several days. Fractionation of particles as fine as a few micrometers in size requires elimination of convection currents that may interfere with the elutriation process. Other elutriators have been described in [14] and [36]. In the following material, several industrial devices are discussed.

Figure 13 shows the *Dorr – Oliver Monosizer* [37], [38] designed for sharp splitting of a feed into two fractions, for cleaning fine sand, of organic contaminants, e.g. coal, (sorting) and generally for desliming. The feed concentration should be about a 1 : 1 ratio of sand to water. Coarser particles settle to the bottom and form a uniform fluidized layer over the distribution pipes (teeter bed). This layer, which acts as a filter, enhances the sharpness of classification. Coarse product is discharged at the bottom, and fine product at the overflow launder. Capacities range up to 100 t/h for tank diameters up to 3 m.

The *Deister Superscalper* [39], which operates like the Monosizer, is shown in Figure 14. It can handle 35 to 50 t/h of coal feed < 25.4 mm, < 12.7 mm, or < 6.35 mm in diameter, containing 40 % or more rejects. The overflow, stripped of the coarser size fractions and

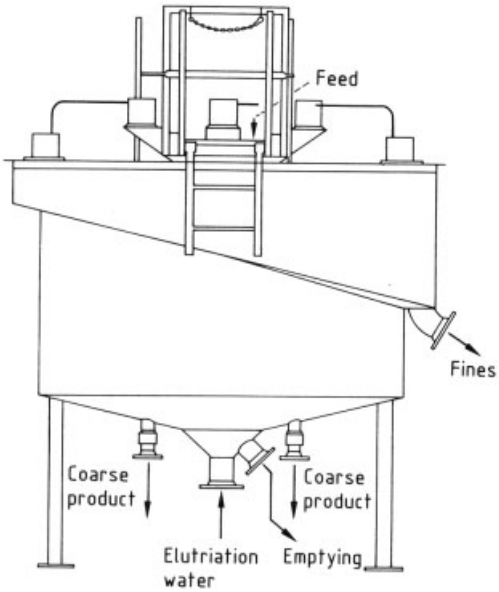

Figure 13. The Dorr – Oliver Monosizer

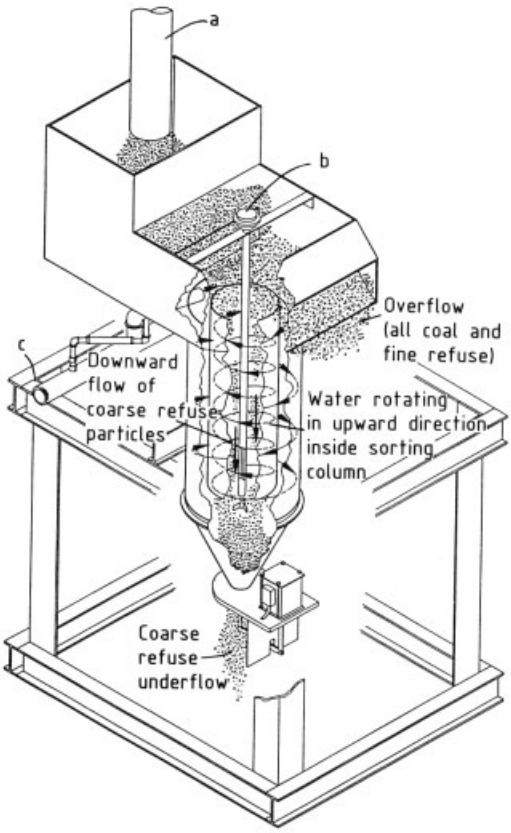

Figure 14. The Deister Superscalper
a) Feed inlet, b) Automatic level-sensing device; c) Water inlet

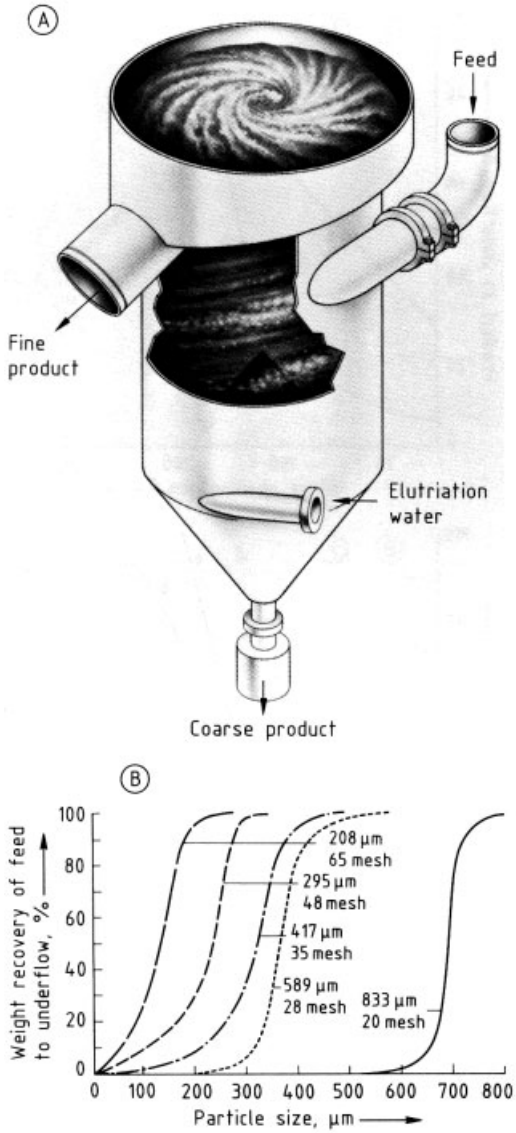

Figure 15. The Krebs C-H Whirlsizer
A) Principle of operation; B) Performance data

rejects, can normally be processed further by cyclones and shaking tables.

Figure 15 shows the *Krebs C–H Whirlsizer* [40–42], a counterflow elutriator incorporating a swirl that is superimposed on the rising water stream. The swirl action, in conjunction with internal design, obviates the need for the teeter bed sizing used in the Monosizer. The Whirlsizer is stable under variable feed conditions, and features compactness and high separation

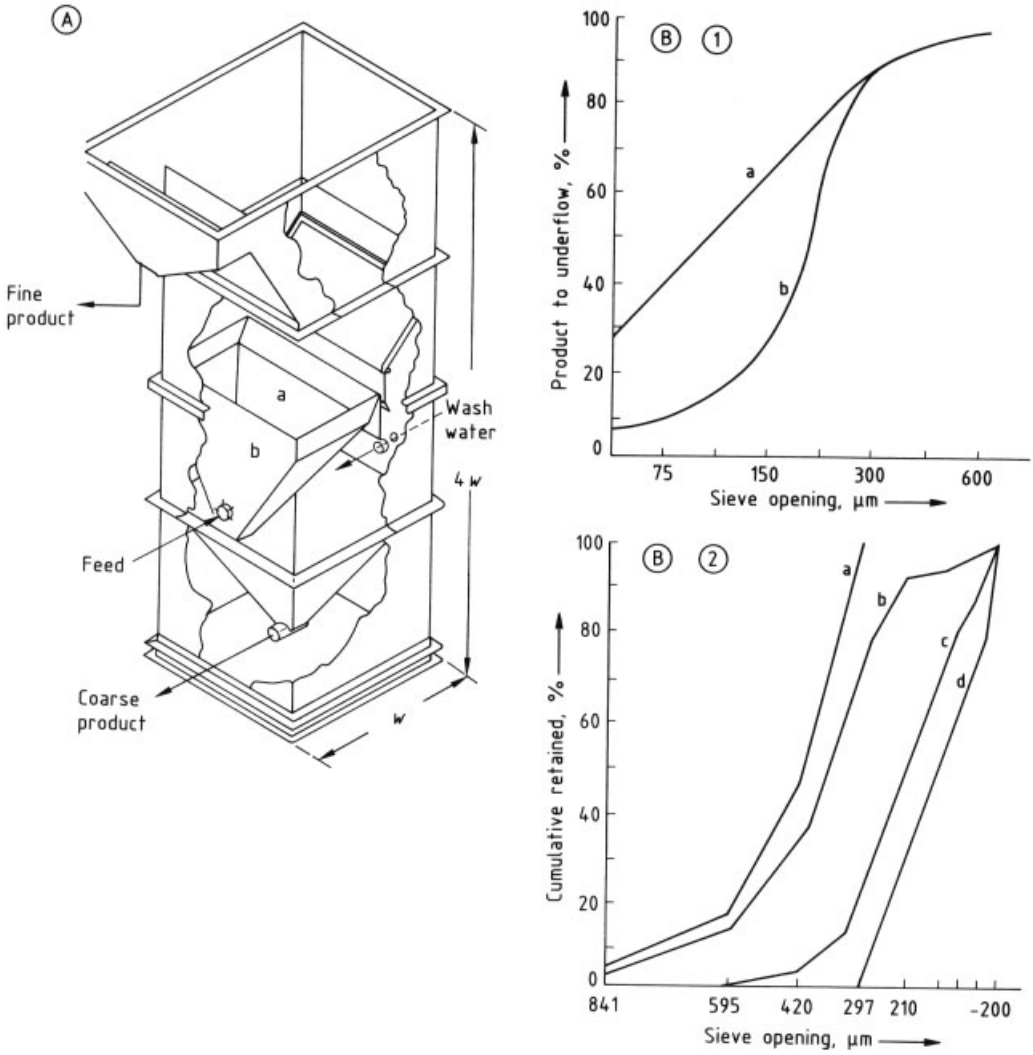

Figure 16. The Humphreys HydroSpec
A) Construction details of one unit: a) Inclined baffle; b) Density zone
B) Performance and comparative data:
1) a) Typical screw classifier, 400 % circulating load; b) HydroSpec, 100 % circulating load;
2) a) Screen oversize product (>297 µm); b) HydroSpec, Oversize product; c) HydroSpec undersize product; d) Screen undersize product (<297 µm)

rate per unit sizer area. The separation size range is from 37 to 2000 µm. Capacities per unit are ca. 100 t/h, but units up to 500 t/h have also been designed. Typical performance data are shown in Figure 15 B [41]. The curves are demarcated by size, above which recovery to the underflow is practically complete. The steep curves indicate a high sharpness index.

Figure 16 shows the *Humphreys HydroSpec* [43]. This elutriator – classifier is a commercial version of the Lewis hydrosizer [44]. The unit (Fig. 16 A) consists of an elongated box whose height is approximately three to four times greater than the base width *W*. Slurry is introduced via a feed inlet at the bottom of an inclined baffle (a) which generates a higher "density zone" (b) of circulating oversize parti-

cles. Coarse solids discharge over the top lip of the baffle into the settling chamber as they are displaced by new feed. Fines are mixed thoroughly in the density zone and pass uniformly across the interface of the density zone into the upper compartment where elutriation in a dilute slurry occurs. Figure 16 B shows typical performance curves that indicate the superiority of the HydroSpec over a typical screw classifier (chart 1) and the near screen efficiency of the HydroSpec for 297-µm (50-mesh) undersize product (chart 2). The HydroSpec is designed to replace screens for sizing particles finer than 1.41 mm. The unit provides two-stage washing of coarse fractions that are discharged at solid contents up to 75 %. By using this unit, circulating loads of rod mill discharge material (in closedcircuit grinding) were reduced from 400 to 100 % (Fig. 16 B, chart 2). A typical unit (1.22 m × 1.22 m × 3.05 m) can classify approximately 40 t/h of sand with a diameter of ≤ 420 µm.

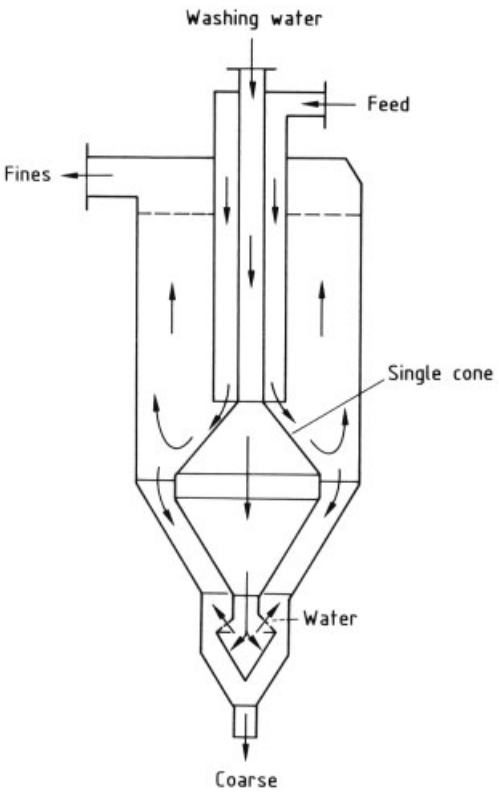

Figure 17. Single-cone elutriator [27], [42]

Details of a *single-cone elutriator* that incorporates top feed and wash water, as well as an internal cone surface for material distribution against rising water currents [27], [45], are shown in Figure 17. The feed descends through a central annular pipe (which encloses the downcoming wash-water pipe) and spreads on the cone surface, with coarse particles descending against the rising current that sweeps the fines toward their exit pipe.

Figure 18 shows a *multicone zigzag elutriator* [45]. The alternate cone surfaces and the enclosing guide walls form a three-dimensional zigzag pattern. Top feed and bottom wash water produce results similar to the single-cone elutriator.

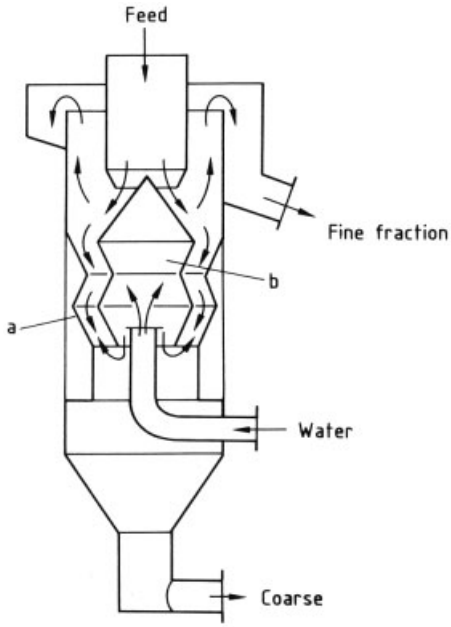

Figure 18. Multicone elutriator
a) Wall guides; b) Multicone surfaces

The *Dorr–Oliver Hydrosizer* [46] and the *Deister Constriction Plate Classifier* [47] (Fig. 19) are two examples of *multifraction hydrosizers.* Water flows across a row of compartments characterized by diminishing velocity of counterflowing water. The Deister elutriation classifier consists of a row of square cells. Each cell contains a pressure chamber at the bottom, a sorting column separated from the pressure chamber by a perforated constriction plate, and a launder which flares out above the sorting column. Secondary classification occurs at the vor-

Figure 19. Multifraction hydrosizer: Deister Constriction PlateClassifier

tex fitting through which coarse fractions leave the column via a spigot pipe.

A high ratio of solids to water feed, with solids content not lower than 30 % and preferably 50 – 60 %, is recommended. The constriction plates enhance formation of a teeter bed of coarse particles. This produces a desliming effect similar to the one obtained by the Dorr – Oliver Monosizer described above. The processing capacity for average sand (<6.35 mm in diameter) is generally rated at 2 – 6 t/h per cell. The number of cells (regularly 2 to 14) can be varied as required. Generally, makeup water does not exceed 4000 L per ton of feed.

Figure 20 shows the *Dorr – Oliver Bowl Desiltor* [48]. The elutriation bowl is connected to a rake classifier (c) that removes the dewatered coarse fraction. Fines are elutriated by water flowing upward and into the peripheral overflow launder. Settled grit is moved by rakes from the bowl into an inclined channel and then exits above the water level from its upper end. Feed concentrations up to 65 % solids in the size range < 12 mm, at capacities ranging from 5 to 120 t/h, can be handled. Normal solids content in the overflow is 1 – 15 %, with the underflow generally being 70 – 85 %. The Bowl Desiltor is used mainly in the recovery of relatively fine sand that overflows washing units in the steel (blast furnace scrubber water), metallurgical, concrete, foundry, glass, and limestone-processing industries. It also serves as a scalper used ahead of thickeners and clarifiers, without the need for a higher water head or the pumping capacities associated with other systems such as cyclones. The typical overall power required for the rake drive motors ranges from 2.6 to 5.2 kW.

4.3. Centrifugal Elutriators

In centrifugal elutriators, gravity is replaced by centrifugal force which enhances sedimentation of fine particles, thereby increasing flow capacity so as to satisfy process requirements. Because of design and technological constraints, the radial length of the sorting or elutriation chamber is relatively small; hence, actual classification takes place over a short (radial) distance.

The *hydrocyclone* is a well-known liquid-operated centrifugal elutriator. The conventional hydrocyclone is discussed in detail elsewhere (→ Hydrocyclones). The hydrocyclone incorporates complex elutriation mechanisms which range from the one described in Figures 10 and 11 to radial and axial countercurrent flow in streams that flow against the centrifugal and gravitational fields. The analysis of elutriation in hydrocyclones is outside the scope of this article. One drawback of hydrocyclones is that the short residence time of the particles may lead to incomplete sedimentation or elutriation. Alternatively, the internal circulation that characterizes novel air classifiers (→ Air Classifying) does not exist in conventional cyclones. To over-

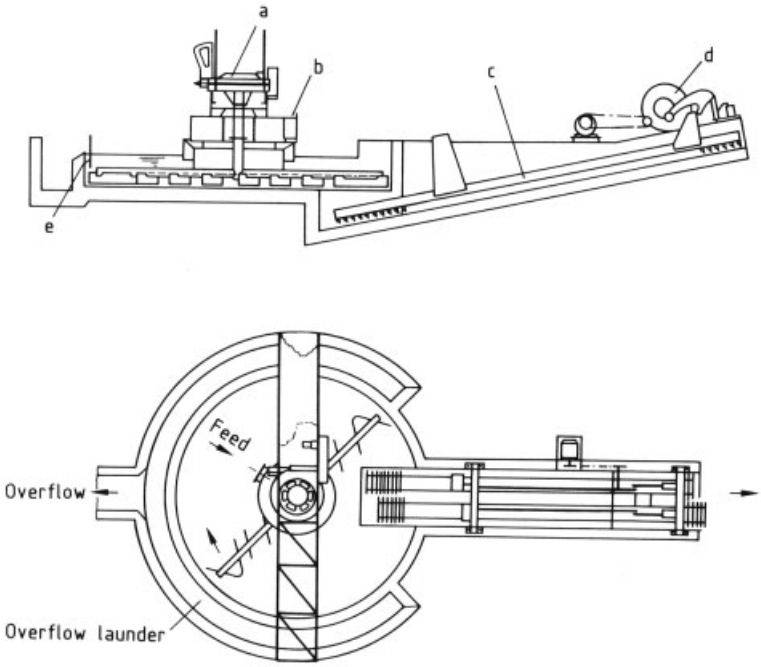

Figure 20. The Dorr Bowl Desiltor
a) Drive head; b) Overflow alarm; c) Rake classifier; d) Rake drive; e) Adjustable overflow weir

come this problem, an *inverted cyclone*, shown in Figure 21 A [49], was developed. Here, particles that have reached the walls, either because of closeness to them or because of sedimentation in the outer vortex, climb up toward the closed cylindrical collection box and are then recirculated into the conical section. This circulation facilitates repeated elutriation and classification which enhance the release of trapped fine particles (in the stream or in aggregates) that finally exit through the vortex finder.

Figure 21 C shows a diagram of the *Warman Cyclosizer* [27], [31], [50]; this consists of five inverted hydrocyclones connected in series, which are fed by a pump, and a sample container. The cyclones decrease progressively in size, no. 1 being the largest and no. 5 the smallest. The coarse product of each cyclone can be collected by opening a discharge valve at its otherwise closed collection box. An example of five size fractions, obtained for size analysis of ground silica sand after standard 30-min elutriation, is shown in Figure 21 B. Limiting sizes of quartz particles with ϱ_p of 2700 kg/m^3 for the cyclones are 44, 35, 23, 15, and 9 µm for cyclones no. 1, 2, 3, 4, and 5, respectively [31], [50]. Changing of ϱ_p to ϱ'_p varies the limiting size by a factor of $[(\varrho_p - \varrho_\varphi)/(\varrho'_p - \varrho_\varphi)]^{1/2}$. The Warman Cyclosizer is a laboratory apparatus capable of handling 100 g of < 149-µm (100-mesh) or < 74-µm (200-mesh) material in 10 – 30 min, depending on the precision required.

Figure 22 A is a schematic diagram of a *counterflow liquid centrifugal elutriator* [31], [51]. An elutriation chamber (a) is engraved in the body of the rotor (b). The elutriation liquid and feed are introduced into this chamber via a concentric inlet (c) liquid channel (d), and suspension feed channels (e) and (f). Liquid counterflowing the centrifugal force drags undersize particles to the fines drain (g), whereas oversize particles are centrifuged toward the oversize drain. Figure 22 B shows performance data in which the cumulative weight percent of particles $\Psi(d_s)$, characterized by size $d \leq d_s$, has been plotted as a function of the Stokes diameter d_s for the oversize (a) and undersize (b) fractions. The undersize fraction is sharp, with 90 % of particles being in the 1- to 3-µm range and the rest extending up to 5 µm. Loss of fines into

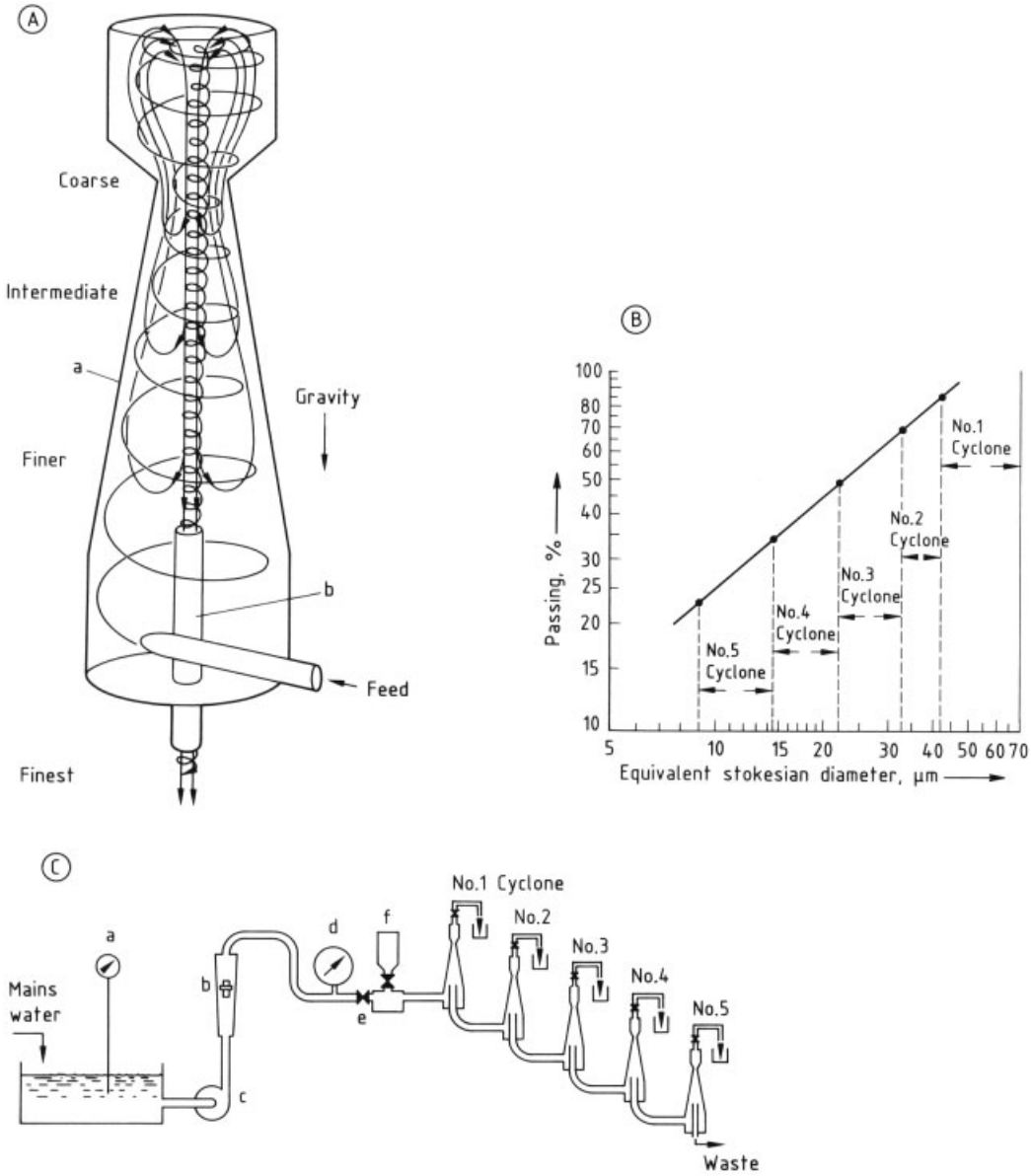

Figure 21.
A) Flow diagram in an inverted cyclone with closed collection box: a) Wall; b) Vortex finder
B) The Warman Cyclosizer–performance data
C) The Warman Cyclosizer–apparatus: a) Thermometer; b) Rotameter; c) Pump; d) Pressure gauge; e) Flow control; f) Sample container

the oversize fraction is probably related to the nonuniform flow profile of the fluid in the elutriation chamber.

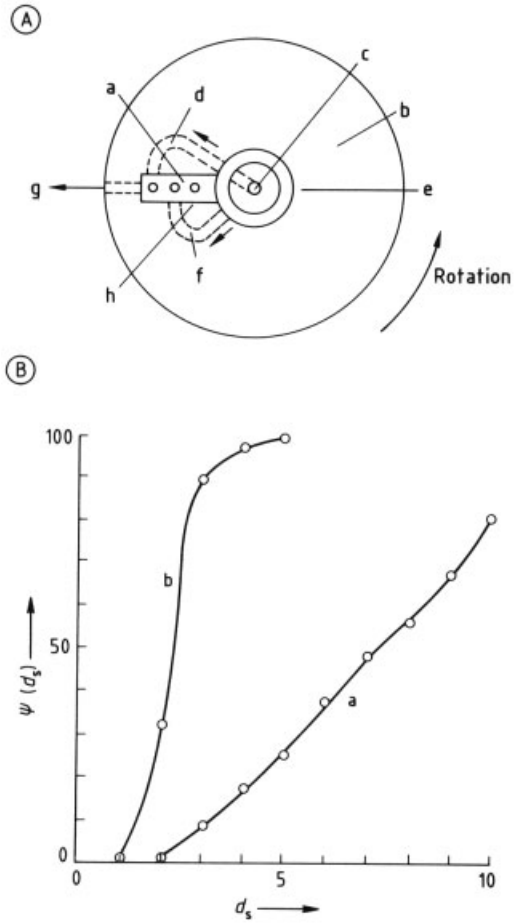

Figure 22. Counterflow liquid centrifugal elutriator
A) Schematic diagram: a) Elutriation chamber; b) Rotor; c) Elutriating liquid inlet; d) Liquid channel; e) f) Suspension feed channels; g) Oversize drain; h) Fines drain
B) Performance data ψ (d_s) vs. d_s: a) Oversize fraction; b) Undersize fraction

An improved elutriation chamber is shown in Figure 23 A; the chamber walls are shaped so as to maintain a constant ratio of drag to centrifugal force. The grade efficiency $g_f(x)$ vs, particle size x (of glass spheres at 2.6-µm cut size) is shown in Figure 23 B and gives an index of separation sharpness x_{25}/x_{75}= 0.61.

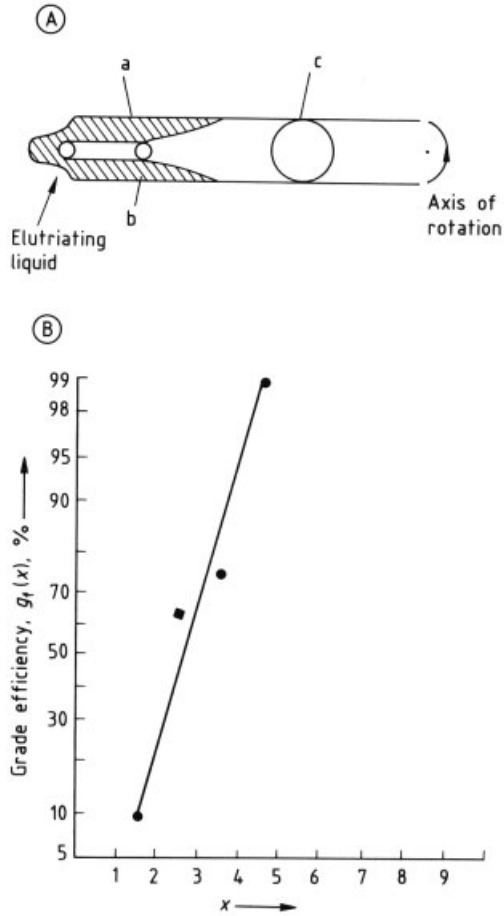

Figure 23.
A) An improved elutriation chamber of the counterflow liquid centrifugal elutriator: a) Elutriation chamber with shaped walls; b) Suspension feed hole; c) Fines overflow
B) Grade efficiencies $g_f(x)$ vs. particle size x of glass spheres at 2.6 µm cut size.

4.4. Axial Cross-Flow Elutriators

Axial cross-flow elutriators are designed to deposit particles axially, i.e., along the flow axis of the main fluid stream. The theory of selected cross-flow systems is discussed in Section 3.10.2. Cross-flow elutriation occurs in systems having non-colinear components of flow and of the force field acting on the particles. A simple example of gravitational cross-flow elutriators is a settling pool, where particles are elutriated horizontally as they settle toward the bottom of the pool.

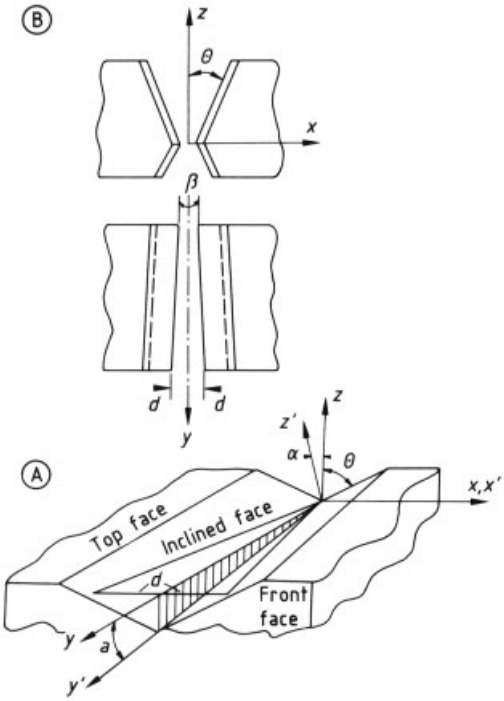

Figure 24. Magnetogravimetric horizontal elutriator
A) Wedge-type profiles, perspective view; B) Axially tapering wedge-type profile, side and top view

Axial cross-flow elutriation is not limited to separation according to size, it can also be used for fractionation of particles according to density and magnetic susceptibility. Figure 24 shows a *magnetogravimetric horizontal elutriator* [52]. In this system, the channel formed between two identical poles of an electromagnet is filled with magnetizable fluid. Particles elutriated axially experience a decreasing levitational force because of the axial tapering of the poles; thus, they settle at different points along the system axis. Figure 25 A shows typical results of separation tests with galena (ϱ_p 7500 kg/m^3) and copper (ϱ_p 8900 kg/m^3) at different levels of applied electromagnetic current. The axial distribution of these heavy particles is clear. For further details, see [52–54].

4.5. Electromagnetic Elutriators

Electromagnetic elutriators (generally referred to as magnetic and electrostatic separators, → Magnetic Separation; → Electrostatic Separation) are devices in which the force field is generated by an electromagnetic field.

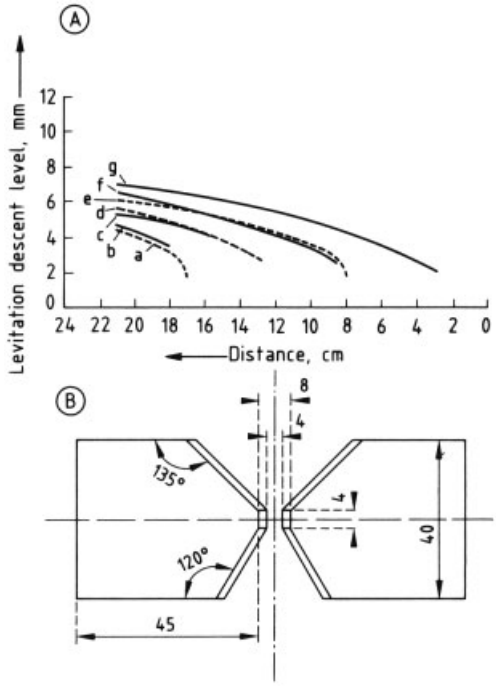

Figure 25.
A) Results of separation tests with galena (G —-) and copper (C · · · · ·) particles at different current levels; a) C (1.00 A); b) G (0.90 A); c) G (1.00 A); d) C (1.25 A); e) C (1.50 A); f) G (1.25 A); g) G (1.50 A)
B) Details of pole profiles used for the separation

In these devices, a strong diverging field is generated around matrix elements with two- or three-dimensional geometry (e.g., rods or spheres), which may be energized directly or indirectly (i.e., by a uniform field). Particles elutriated by a fluid stream passing in the vicinity of these elements [55] will experience a force which may or may not result in their removal from the stream, depending on flow, electromagnetic force, and residence time. This subject is outside the scope of this article.

5. References

1. E. Barnea, J. Mizrahi, *Chem. Eng. J. (Lausanne)* **5** (1973) 171 – 189.
2. Y. Zimmels: "Accelerated and Steady Particle Flows in Newtonian Fluids," *Encyclopedia of Fluid Mechanics,* Gulf Pub., Houston, 1986, pp. 93 – 153.

3. Y. Zimmels, *J. Appl. Phys.* **56** (1984) no. 7, 2133 – 2141.
4. Y. Zimmels, *AIChE J.* **29** (1983) no. 4, 669 – 676.
5. E. L. Cussler: *Diffusion, Mass Transfer in Fluid Systems,* Cambridge University Press, Cambridge 1984.
6. R. E. Rosensweig: *Ferrohydrodynamics,* Cambridge University Press, Cambridge 1985.
7. R. H. Davis, A. Acrivos, *Ann. Rev. Fluid Mech.* **17** (1985) 91 – 118.
8. J. F. Richardson, W. N. Zaki, *Trans. Inst. Chem. Eng.* **32** (1954) 35 – 53.
9. J. Garside, M. R. Al-Dibouni, *Ind. Eng. Chem. Process Des. Dev.* **16** (1977) 206 – 214.
10. H. Hald: *Statistical Theory with Engineering Applications,* Wiley International Edition, 1952.
11. P. L. Meyer: *Introductory Probability and Statistical Applications,* Addison Wesley, 2nd ed., Reading, Massachusetts 1970.
12. J. R. Benjamin, C. A. Cornell: *Probability, Statistics and Decision for Civil Engineers,* McGraw – Hill, New York 1970.
13. J. K. Beddow: *Particulate Science and Technology,* Chemical Publishing Co., New York 1980.
14. T. Allen: *Particle Size Measurement,* 3rd ed., Capman and Hall, London, 1981.
15. B. W. Lindgren: *Statistical Theory, Macmillan Publ. Co.,* New York 1960.
16. R. P. Fraser, P. Eisenklam, *Trans. Inst. Chem. Eng.* **34** (1956) 304.
17. A. O. Gates, *Trans. Amer. Inst. Mining Metall. Pet. Eng.* **52** (1915) 875 – 909.
18. R. Schumann, *Trans. AIME, Tech. Pub.* (1940) 1189.
19. A. M. Gaudin, T. P. Meloy, *Trans. Am. Inst. Min. Metall. Pet. Eng.* **223** (1962) 43 – 50.
20. A. J. Lynch: *Mineral Crushing and Grinding Circuits,* Elsevier, Amsterdam 1977, p. 116.
21. C. C. Harris, *Trans. SME* **241** (1968) 343 – 358.
22. L. Svarovsky in L. Svarovsky (ed.): *Solid Liquid Separation,* "Characterization of Particles Suspended in Liquids," Butterworths, London 1977, pp. 5 – 29.
23. C. C. Harris, *Trans. SME* **244** (1969) no. 6, 187 – 190.
24. K. R. Gibson, *Powder Technol.* **18** (1977) no. 2, 169 – 170.
25. L. Svarovsky in L. Svarovsky (ed.): *Efficiency of Separation of Particles from Fluids, in Solid Liquid Separation,* Butterworths, London 1977, p. 31 – 57.
26. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
27. E. G. Kelly, D. J. Spottiswood: *Introduction to Mineral Processing,* J. Wiley & Sons, New York 1982.
28. W. E. Dobbins, *Trans. ASCE* **109** (1944) 629 – 656.
29. T. R. Camp, *Trans. ASCE* **111** (1946) paper 2285, 895 – 958.
30. H. Schubert, T. Neesse, *10th Int. Miner. Process. Congr. London,* 1973, pp. 213 – 239. IMM (1974) .
31. B. H. Kaye: *Direct Characterization of Fine Particles,* Wiley-Interscience, J. Wiley & Sons, New York 1981.
32. C. J. Stairmand: "Some Practical Aspects of Particle Size Analysis in Industry," *Symposium on Particle Size Analysis* (1947). *Suppl. Transact. Inst. Chem. Eng.,* **25** (1947).
33. E. Schöne: *Über Schlammanalyse und einen neuen Schlammapparat,* Berlin 1867.
34. L. Andrews, *Proc. Inst. Eng. Inspection* **25** (1927 – 8) also Min. Mag. (1982), p. 301.
35. H. N. Blythe, E. J. Pryor, A. Eldridge (1953), *Symp. Recent Developments in Mineral Dressing, Inst. Min. Metall.,* London, 23 – 5 Sept., 1952, p. 11.
36. A. H. M. Andreasen, *Ber. Dtsch. Keram. Ges.* **11** (1930) 675.
37. Dorr – Oliver news release no. 002/72, Sand Beneficiation, Dorr – Oliver information 13/78 d.
38. W. Schwalbach: "The Processing and Decarbonization of Fine Sand," *Aufbereit. Tech.* **12** (1971) no. 3, 137 – 139.
39. Deister Superscalper, The Deister Concentrator Company Inc., Bulletin no. 220-B, 3/82.
40. P. Hedley, *Pit Quarry,* October 1984.
41. Krebs Engineering, Leaflets and News Bulletins.
42. M. Rukavina, *Rock Prod.,* Feb. 1987.
43. Coors, Humphreys Processing Equipment, Bulletin no. 42 (10/86) LB-1638.
44. R. M. Lewis: *Hydrosizing of Industrial Minerals,* a publication of Humphreys Mineral Industries Inc.
45. *Ullmann,* 4th ed., **2,** p. 70.
46. Dorr – Oliver Technologie Transfer T4, 8/81.
47. The Deister Concentrator Company Inc., Bulletin no. 200-B, 4/78.
48. Dorr – Oliver Europe publication, 05-IE-0884.

49. D. F. Kelsall, J. C. H. McAdam: "Design and Operating Characteristics of Hydraulic Cyclone Elutriator," *Trans. Inst. Chem. Eng.* **41** (1963).
50. Warman International Ltd., leaflet: Cyclosizer for Sub-Sieve Sizing.
51. F. J. Colon, J. W. Van Heuven, H. M. Van der Laan: "Centrifugal Elutriation of Particles in Liquid Suspension," in *Proc. Conf. Fine Particle Characterization,* Bradford, England, 1970, British Society for Analytical Chemistry, London, 1971.
52. Y. Zimmels, I. J. Lin, *IEEE Trans. Magn.* **18** (1982) no. 3, 921.
53. Y. Zimmels, I. J. Lin, *IEEE Trans. Magn.* **19** (1983) no. 1, 27.
54. Y. Zimmels, I. J. Lin, *Chem. Eng. Commun.* **21** (1983) 1 – 3.
55. Y. Zimmels, *IEEE Trans. Magn.* **20** (1984) no. 4, 597.

Embedding → Epoxy Resins
Emulsifiers → Emulsions
Embedding → Insulation, Electric
Emulsion Polymerization → Polymerization Processes

Air Classifying

See also → Solid – Solid Separation, Introduction

ROLAND NIED, Alpine Aktiengesellschaft, Augsburg, Federal Republic of Germany

HERBERT HORLAMUS, Alpine Aktiengesellschaft, Augsburg, Federal Republic of Germany

1. Introduction 1
2. Forces on the Individual Particle 2
3. Classifier Performance 4
3.1. Throughput 4
3.2. Cut-Point Range 5
3.3. Sharpness of Cut 5
3.4. Energy Requirement 5
4. Description of Individual Classifier Types 5
4.1. General 5
4.2. Countercurrent Classifiers . . . 7
4.2.1. Countercurrent Gravity Classifiers 7
4.2.2. Countercurrent Centrifugal Classifiers 8
4.3. Crosscurrent Classifiers 14
4.4. Recirculating-Air Classifiers . . 15
5. Operating Properties 16
5.1. Effect of Feed 16
5.2. Scaleup 17
5.3. Special Requirements 17
6. References 18

Symbols:

c_W	drag coefficient
F_M	mass force
F_W	drag force
g	acceleration due to gravity
$\dot{m}$	mass flow rate
P	passage
r	radius
Re	Reynolds number
v	velocity
$\dot{V}$	volume flow rate
W	mass fraction
W_c	mass fraction of coarse material
W_f	mass fraction of fines
x	particle diameter
x_{f97}	reference cut size
x_t	cut size (general)
x_{50}	Tromp's cut size
η	dynamic viscosity
κ	sharpness of cut
μ	load
ξ	efficiency
ϱ	density
φ	coefficient

Indices

A	gas
c	coarse material
F	feed material
f	fines
S	solid
v_r	radial component of velocity
v_p	peripheral component of velocity

1. Introduction

Air classification is a process for the dry separation of a disperse phase according to particle size, particle shape or density, or more precisely, settling rate. It is the analogue of hydraulic classification in wet-processing technology (→ Elutriation). The bulk material to be separated is introduced into a gas flow, and the particles are subjected primarily to two forces: (1) the resistance or drag force caused by the gas flowing around the particle, and (2) the mass force. The mass forces acting on the particle can be field forces (e.g., gravity) or inertial forces (e.g., centrifugal force). Depending on magnitude and direction of these forces, particles move along different trajectories and can thus be collected separately after leaving the classification zone.

Air classifiers are used partly as sorters but primarily as classifiers. In sorting, a heterogeneous bulk is separated into components of uniform material. The particles are separated ac-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_17

cording to density or particle shape. Examples of density sorting are the fractionation of household garbage, the separation of comminuted cable scraps into copper and plastic fractions, etc. Separation according to particle shape is possible for mica (plates)–quartz or mica–feldspar (cubic form) mixtures. In most cases, however, a homogeneous bulk is present, and the air separator operates as a classifier; i.e., it separates according to particle size. In this case, beginning with commercial or natural raw material, the following processes take place:

1) A product of the desired fineness is removed.
2) A relatively coarse-grained product is dedusted.
3) A small amount of oversize material is separated.
4) Fractions with a definite upper and lower limit are produced.
5) Particle size is analyzed.

Air classification is used predominantly in the fine and very fine range, i.e., for effective separation of particles from 300 µm down to 1 µm. It dovetails with screen classification, which operates more economically on coarse particles. Although air classifiers can also be used to separate particles in the millimeter-size range, in principle, their use then is primarily for sorting, not for classification.

2. Forces on the Individual Particle

An exact calculation of the forces acting on particles, and thus a calculation of trajectories and, ultimately, of the cut size, is not possible in most cases. The flow prevailing in the separation zone of an air classifier is too unsteady and inhomogeneous. The effects of feed (concentration, particle-size distribution), friction, and secondary flows preclude accurate description of temporal and local velocity patterns in the total classification zone. Theoretical considerations must, therefore, be based on simplified model flows. Nevertheless, even with this approach, some characteristic dependences and properties of air classifiers can be derived.

The *drag force* F_{W}, i.e., the force that a flow exerts on a particle moving at a relative velocity v_{rel} with respect to it, is calculated according to Newton's law as

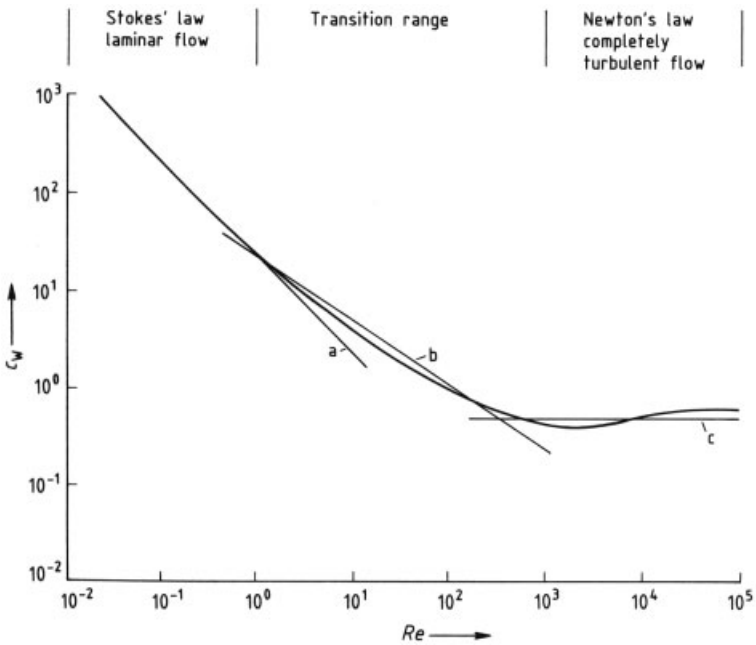

Figure 1. Drag coefficient of spherical particles
a) $c_{\mathrm{W}} = 24/Re$; b) $c_{\mathrm{W}} \approx 24/Re^{2/3}$; c) $c_{\mathrm{w}} \approx 0.45$

$$F_W = c_W \frac{\pi x^2}{4} \frac{\varrho_A}{2} \tag{1}$$

where x is the particle diameter, ϱ_A is the density of the gas, and c_W is the drag coefficient. The drag coefficient c_W is known to depend on the Reynolds number ($Re = v_{rel} \cdot x \cdot \varrho_A / \eta_A$, where η_A is the dynamic viscosity).

In the particle technique, Reynolds numbers between 10^{-2} and 10^5 occur. A generally valid relationship $c_W = f(Re)$ cannot be given for this large range, and it is therefore subdivided into three parts that characterize the different types of flow around a body (Fig. 1). For practical calculations, which are frequently restricted to only part of a range, more or less accurate approximations are used.

Air classification takes place predominantly in the transition range. Only analytical classifications in the rising-pipe classifier (see Section 4.2.1) and very fine classifications with effectivecut sizes x_t of a few micrometers advance into the Stokes range. Table 1 lists some Reynolds numbers for different classifiers under typical operating conditions.

The forces acting on a particle are considered first for the case of ideal flow in a rising pipe (Fig. 2). For the cut-size particle, which theoretically remains in the classification zone for an infinitely long time, the mass force F_M and drag force F_W are in equilibrium. The mass force F_M in this case is the force of gravity and is calculated as

$$F_M = \frac{\pi}{6} \varrho_S x^3 g \tag{2}$$

where ϱ_S is the particle density and g is the acceleration due to gravity.

By combining Equations (1) and (2) with $v_{rel} = v$, the equilibrium particle size or cut size x_t can be derived:

$$x_t = \frac{6}{8} c_W \frac{\varrho_A}{\varrho_S} \frac{v^2}{g} \tag{3}$$

For laminar flow around the particles, $c_W = 24/Re$ (Fig. 1). Equation (3) thus becomes

$$x_T = \sqrt{\frac{18 v \eta}{\varrho_S g}} \tag{4}$$

For the transition range, according to the approximate equation in Figure 1: $c_W = 24/Re^{2/3}$, the effective cut size is then calculated as

$$x_t = 5.66 g^{-0.6} \varrho_S^{-0.6} \varrho_A^{0.2} \eta^{0.4} v^{0.8} \tag{5}$$

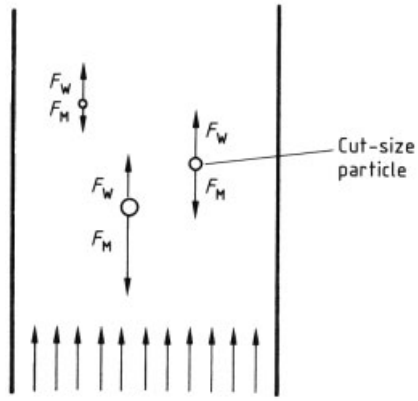

Figure 2. Forces on particles in the rising-pipe classifier

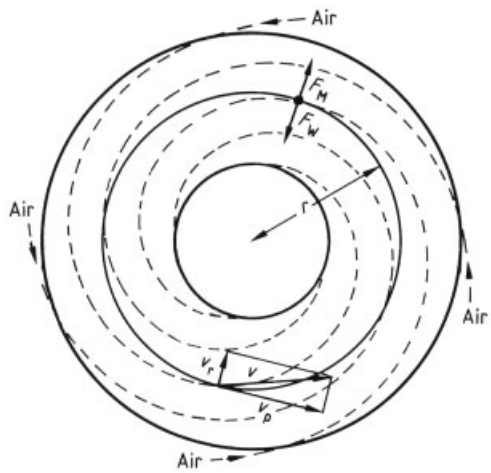

Figure 3. Equilibrium of forces on a particle in spiral flow

In the following material, the relationship of forces in an ideal spiral flow is investigated (Fig. 3). Spiral flow or at least similar forms of flow are found in the separation zones of most industrial air classifiers. Air flows along the trajectories indicated by broken lines; its velocity along the radius r is v, with the peripheral component v_p and the radial component v_r. For a particle of diameter x, if the mass force F_M and the drag force F_W are in equilibrium at the radius

Table 1. Reynolds numbers for different types of air classifiers

Classifier type	Pressure in classification space, kPa	v_{rel}, m/s	x_{t}, µm	*Re*
Rising-pipe analytical classifier	100	0.1	35	0.2
Classifier with high-speed rotor	80	3.0	2	0.4
Impeller classifier	90	6.5	10	0.4
Spiral classifier	95	6.5	25	10
Channel-wheel classifier	95	20	50	60
Recirculating-air classifier	100	10	100	60
Rising-pipe classifier	100	5.0	1000	300

r, the particle will describe a circular path. Theoretically, this particle will remain in the classification zone for an infinitely long time. Larger particles are thrown outward, while smaller particles are carried along by the air to the central outlet.

In this case, F_{M} is the centrifugal force and has a value

$$F_{\mathrm{M}} = \varrho_{\mathrm{S}} \frac{x^3 \pi}{6} \frac{\mathrm{v}_{\mathrm{p}}^2}{r} \tag{6}$$

From Equations (1) and (6) and $v_{\mathrm{rel}} = v_{\mathrm{r}}$, the following value is obtained for the equilibrium particle size x_{t}:

$$x_{\mathrm{t}} = \frac{6}{8} c_{\mathrm{W}} \frac{\varrho_{\mathrm{A}}}{\varrho_{\mathrm{S}}} r \frac{v_{\mathrm{r}}^2}{\mathrm{v}_{\mathrm{p}}^2} \tag{7}$$

Table 1 shows that calculation can be restricted primarily to the transition range; thus Equation (7) can be rearranged to

$$x_{\mathrm{t}} = 5.66 \varrho_{\mathrm{S}}^{-0.6} \varrho_{\mathrm{A}}^{0.2} \eta^{0.4} r^{0.6} \mathrm{v}_{\mathrm{r}}^{0.8} \mathrm{v}_{\mathrm{p}}^{-1.2} \tag{8}$$

The cut size or the cut-size range to be expected can be estimated roughly by means of these equations when ϱ_{S}, ϱ_{A}, and η are constant or given. The cut point is normally adjusted by means of v or by the ratio $v_{\mathrm{r}}/v_{\mathrm{p}}$. However, the equilibrium radius r, the so-called separation radius, is also used to adjust some spiral classifiers. The separation radius depends on the geometry of the classification zone and, of course, on the absolute size of the classifier. (For a comprehensive description of cut size, density distribution curves, and grade efficiency, see → Particle-Size Analysis and Characterization of a Classification Process).

3. Classifier Performance

The performance of a classifier is characterized by throughput, cut-point range, sharpness of cut, and energy requirements.

3.1. Throughput

The term throughput refers to the mass flow rate of the feed, fines, or coarse material, depending on which stream is considered the product. The capacity of a classifier is generally characterized by the quantity of fines that can be obtained. However, the quantity of fines is not a fixed value, because the output of fines depends on the particle-size distribution of the feed and on the set cut point. The finer the setting of the classifier, the lower its throughput will be. The throughput also has a decisive effect on the sharpness of cut (see Section 3.3). The throughput of both fines and feed material can vary greatly for an identical classifier setting, depending on the objective, e.g., clean coarse material (high sharpness of cut) or "milking" of fines (extraction of a small fraction of fines so that the distribution in the coarse material still corresponds roughly to the distribution in the feed material).

Throughput is dependent on the air flow rate and, thus, primarily on the size of the classifier. To eliminate the effect of size in a more general consideration, the load μ is used. This is defined as a ratio of flow rates: the mass flow rate of either the feed $\dot{m}_{\mathrm{F}}$ or the fines $\dot{m}_{\mathrm{f}}$ is related to the air flow rate $\dot{V}$:

$$\mu_{\mathrm{F}} = \frac{\dot{m}_{\mathrm{F}}}{\dot{V}} \text{ or } \mu_{\mathrm{f}} = \frac{\dot{m}_{\mathrm{f}}}{\dot{V}} \tag{9}$$

Most classifiers operate in the range of μ_{F} from 0.1 (for the finest separations) to 1.0 kg/m^3 (for coarser separations).

3.2. Cut-Point Range

The cut point can be varied within a smaller or larger range for all classifiers. The size and absolute location of this range characterize the performance: the largest range of adjustment and smallest possible cut point yield optimal efficiency.

Cut-point values must always be provided, along with the definition on which they were based (see → Particle-Size Analysis and Characterization of a Classification Process). The specific gravity of the material used must also be indicated; the most common test material is limestone with a density ϱ of 2.65 kg/dm^3. According to Equations (4), (5), and (8), products with a higher specific gravity produce finer effective cut sizes for the same classifier setting; those with a lower specific gravity result in more coarse effective cut sizes. The particle-size distribution of the feed material and the classifier size also affect the cut-point range (see Sections 5.1 and 5.2).

3.3. Sharpness of Cut

The different classifications of the sharpness of cut are discussed in detail in → Particle-Size Analysis and Characterization of a Classification Process. The sharpness of cut depends primarily on the throughput or on the load μ. It usually remains constant up to a certain throughput and decreases continuously above this.

The properties of the feed material, especially the particle-size distribution and agglomeration tendency, are also important. If a product with a steep distribution curve is fed to the classifier, then for an identical load, a significantly larger number of particles of near-cut size will be in the classification zone than with a flatter distribution of the feed material. The interaction between particles, and thus the danger of misclassified particles, increases and the sharpness of cut decreases.

The agglomeration tendency depends partly on particle size (smaller particles have greater adhesion relative to their mass), and partly on material properties such as conductivity, surface roughness, water or oil content, etc. The stronger the adhesion between individual particles, the more difficult it is to carry out a complete dispersion. However, complete dispersion is a basic requirement for a classification with a sharp cut.

3.4. Energy Requirement

Agglomeration is counteracted by the disintegrating power of the classifier, which is based on high local air velocity as well as frequent and intense particle impact. For very fine classifications, the disintegrating power of the classifier is so high in some cases that larger particles of the material are considerably comminuted.

The high air velocities and the acceleration of particles to high impact velocities consume energy. The specific energy requirement, i.e., the energy necessary to produce a defined quantity of end product with a specified fineness, is therefore an important parameter in comparing different classifiers. For a correct comparison, the following requirements must be met:

1) Both classifiers must receive identical feed materials.
2) Both classifiers must produce identical fines.
3) Both classifiers must give the same mass fraction of fines W_f.

In this case, the sharpness of cut will be identical.

The driving power of the fan required to produce classifying air is calculated on the basis of the pressure drop in the classifier or the system and on the quantity of air put through. In some cases, the driving power for moving classifier parts must also be considered. The total power, divided by the mass flow rate of the product, yields the specific energy requirement.

4. Description of Individual Classifier Types

4.1. General

Three standard types of air classifier, as well as a variety of combinations, are used.

Through-Air Classifier. Material is fed into the classifier with little or no additional air. The classifying air enters without any material and leaves it loaded with fines. It then passes through a dust separator into the open air. An advantage

of this system is that the coarse material is very clean because the coarse fraction is rinsed with pure air. One disadvantage is that a relatively large, effective, and therefore expensive, dust separator (filter) must be used.

Closed-Circuit Classifier. Feed is added as already described. Classifying air returns to the classifier after separation of the dust. This allows low-cost blanketing with nitrogen, if required, and a simple dust separator, usually a cyclone, is sufficient. The cyclone can be located outside the classifier (external air circuit), or it can be structurally combined with the classifier and fan (internal circuit). Often, a small amount of air must be removed from the circuit and passed through a filter (leakage-air, caused by rising-air etc.). Dedusting as a whole is, nevertheless, quite inexpensive.

A disadvantage of this method is that the fine dust remaining in the classifying air contaminates the coarse material and can form a deposit in the classifier. Furthermore, distinct heating of the classifying air occurs over the course of time.

Airstream Classifier. The bulk material, suspended in the classifying air, is fed into the classifier. Airstream classifiers are thus used advantageously in grinding–classifying circuits. Material extracted from the mill is fed pneumatically to the classifier, which makes inexpensive plant design possible. A disadvantage is the low purity of the coarse material, which is rinsed only with air loaded with feed material.

Many types of air classifiers are available on the market: some are named according to the mode of operation; others, according to characteristic machine parts or equipment. Therefore, a breakdown into different categories is difficult. In addition, some units have two classification zones (a fines zone and a coarse-material zone) which operate differently, thereby making systematic treatment even more difficult. For example, some air classifiers are distinguished by type of air supply, in others distinction is made between centrifugal and gravity classifiers; finally, division into crosscurrent and countercurrent classifiers is also widespread.

In a *crosscurrent classifier*, the direction of movement of the coarse material (i.e., the direction of mass forces), is at a right angle to the main gas flow or the direction of the drag force. The crosscurrent classifier is also aptly called a fanning-out classifier because separation of a product into any desired number of fractions is possible. In a *countercurrent classifier,* the direction of mass force is counter to the main gas flow. This type of classifier, also called an *accumulating* or *equilibrium classifier,* yields only two fractions.

Accumulating classifiers generally result in sharper separation because the fines and coarse material leave the classification zone in opposite directions, whereas in a *fanning-out classifier,* even the smallest deviation from the limiting trajectory leads to an incorrect assignment of material. Additionally, in an accumulating classifier, cut-size particles remain in the classifier for a longer time so that any incorrect separation is likely to be corrected.

In *elbow classifiers* the particles of the fines undergo a large change in direction (a deflection of ca. 180°) on their way through the classification zone.

The *recirculating-air classifier* is a type that does not fit well into the scheme developed above. This instrument represents a combination of several classification principles.

In this article, air classifiers are considered as follows:

Countercurrent classifiers (Section 4.2)
Countercurrent gravity classifiers (Section 4.2.1)
Countercurrent centrifugal classifiers (Section 4.2.2)
Crosscurrent classifiers (Section 4.3)
Recirculating-air classifiers (Section 4.4)

All the different classifiers available on the market cannot possibly be described because the number of manufacturers is too large. The selection made shows the variety of types developed this far.

Capacity figures are largely omitted because, in many cases, they are not generally valid; i.e., they are not applicable to other products (see also Chaps. 3 and 5). In order to find the suitable classifier for a certain problem, tests must always be conducted with the original material.

Where data on the cut-point range are given, these refer to Tromp's cut point x_{50}, unless otherwise explicitly noted. Furthermore, these data always refer to a product with a density ϱ of 2.65 kg/dm^3.

4.2. Countercurrent Classifiers

4.2.1. Countercurrent Gravity Classifiers

The *rising-pipe classifier* is the simplest form of countercurrent gravity classifier. Figure 4 is a schematic drawing of a model that is, in part, still used for particle-size analysis [1]. Air enters through a nozzle and stirs up the particles to be investigated. Separation into fines and coarse material takes place in the cylindrical pipe. All particles whose settling rate is lower than the upward air flow rate leave the pipe at its upper end, along with the air. The coarse material remains at the bottom of the pipe and can be removed at the end of the classification process.

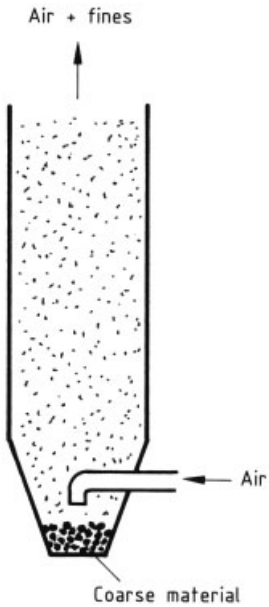

Figure 4. Diagram of a rising-pipe classifier for analytical purposes

An unequal velocity distribution over the cross section and deposition of material on the walls can impair classification. Material deposition can be largely prevented by the installation of knockers. The cut-point range used for analytical classification is located between ca. 10 and 100 µm. The sharpness of cut (κ) in this unit, which is operated in a batchwise manner, depends primarily on the time; for an appropriately long analysis period, κ values of 0.85 – 0.9 are achieved.

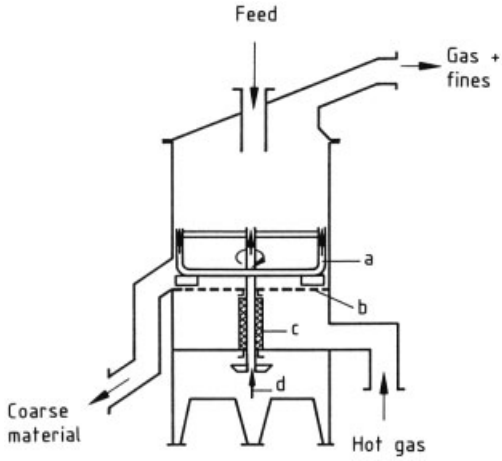

Figure 5. Combination rising-pipe classifier and suspension drier (Keller)
a) Stirring device; b) Screen plate; c) Insulation; d) Cooling air for bearing

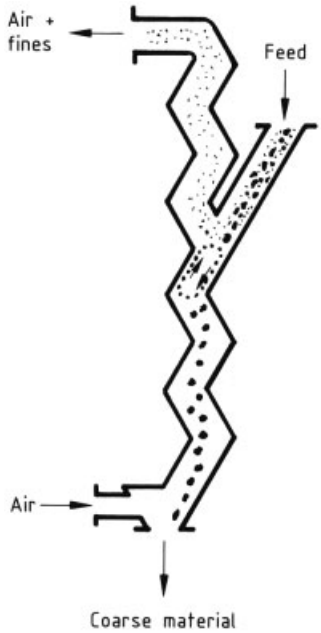

Figure 6. Zigzag classifier (Alpine)

Figure 5 shows a continuously operating gravity classifier in which the drying of moist material can be carried out simultaneously [2]. The material is fed to the center of a screen plate. The hot gas flowing through the screen from below carries the fines or dried product with it toward the top; the coarse material is moved by

means of a stirring device to the edge of the screen and discharged. This classifier is used for separating and drying of sawdust, wood chips, splintered wood, and pulpwood.

The *zigzag classifier* is a special variation of the gravity classifier (Fig. 6) [3]. A characteristic form of movement develops in the zigzag-shaped channel, in which predominantly near cut-size particles form a vortex cylinder in each zigzag section. These particles therefore pass through several classification stages, so that misdirected particles can be brought to the correct path. Accordingly, the sharpness of cut increases exponentially with the number of sections. Large classifiers are obtained by combining many classifying pipes.

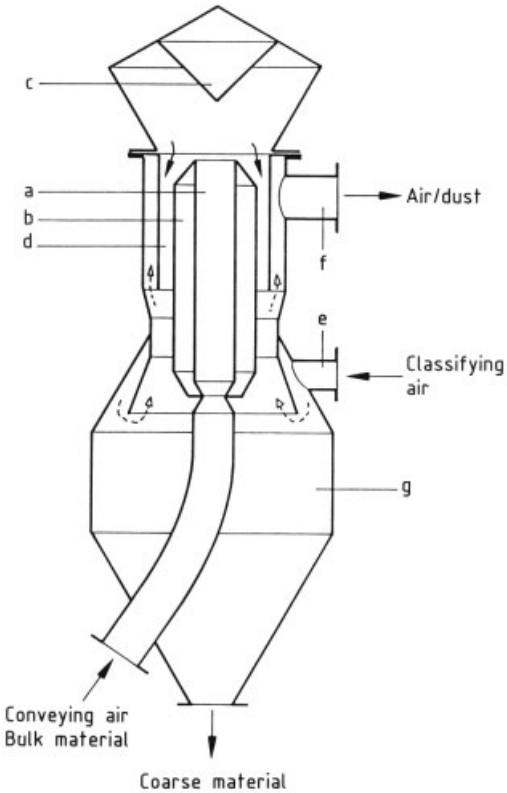

Figure 7. Elbow classifier (Waeschle)
a) Tube; b) Displacement body; c) Deflector cone; d) Annulus; e) Inlet for classifying air; f) Outlet; g) Container for coarse material

In the *elbow classifier* (Fig. 7) [4], feed material is fed pneumatically to the classifier. Before the material reaches the actual classification zone, it is preaccelerated in an annulus. The air flowing through the inlet nozzle slows down the smaller particles and carries them to the outlet. In this process, the fines are deflected by 180° as described in Section 4.1. The classifying air cannot carry along larger particles, which drop down into the coarse-material container. The classifier is used to separate powdery and fibrous components of plastic granulate.

Except for analytical purposes, countercurrent gravity classifiers are used primarily in the separation of coarse bulk material (0.3 – 10 mm). Even here, however, they are used less for actual classifying purposes (i.e., separation according to particle size) than for separation according to specific gravity or particle shape, as already mentioned. Although separation cuts in the fine range (10 – 100 µm; see rising-pipe analytical classifier in Table 1) are possible in principle, they are not used as operating classifiers in this case. The cross-sectional areas, and thus space requirements, would become unacceptably large as a result of the low air velocities.

4.2.2. Countercurrent Centrifugal Classifiers

Spiral Classifier. Spiral classification was first used and investigated systematically in the 1930s [5]. Most spiral classifiers have a flat cylindrical separation chamber in which a spiral flow, moving from the outside toward the inside, is produced. Adjustable guide vanes are located on the periphery. Through the setting angle of these vanes, the parameters of the spiral (i.e., the ratio of peripheral to radial velocity) are adjusted and the cut point is thus established. The cut-point range of spiral classifiers is located between ca. 5 and 80 µm. The fines leave the classification space with the air through a central opening. The different possibilities for input of feed material and discharge of coarse material are illustrated by the following examples.

In the *Mikroplex spiral air classifier* (Fig. 8), feed material is introduced through a tangential chute on the periphery of the classification zone. Coarse material flows in a kind of skein over the guide vanes (a) and is freed of adhering fines by the inflowing air. Finally, it is peeled off by a cutting edge (b) also located on the periphery of the classification zone and is removed by a screw (c). The fan (e) required for production of

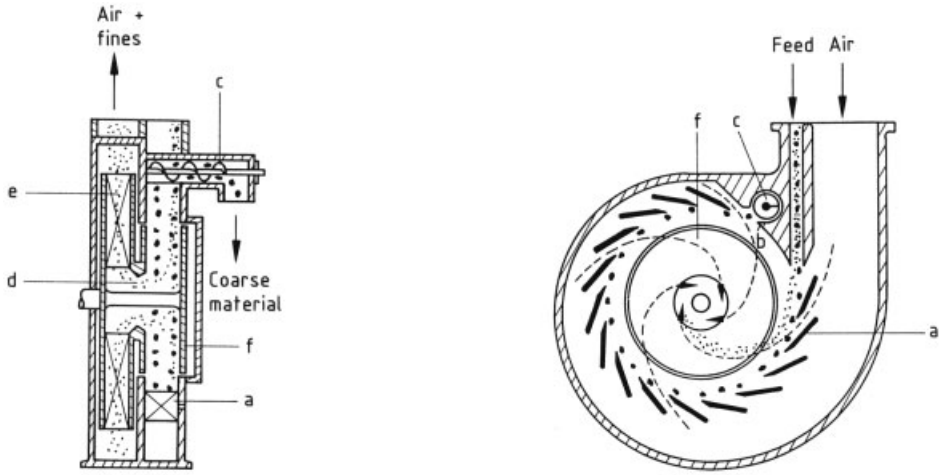

Figure 8. Mikroplex spiral air classifier (Alpine)
a) Adjustable guide vane; b) Coarse-material cutting edge; c) Coarse-material screw; d) Fines outlet; e) Fan; f) Rotating wall of classification chamber

the classifying air is integrated into this classifier. A special feature of the Mikroplex classifier is its rotating side walls (f), which largely eliminate the boundary layer close to the wall. This layer, which is characteristic of static classifiers, results in friction and in the transport of coarse particles to the fines outlet (d); this can lead, among other things, to so-called grit.

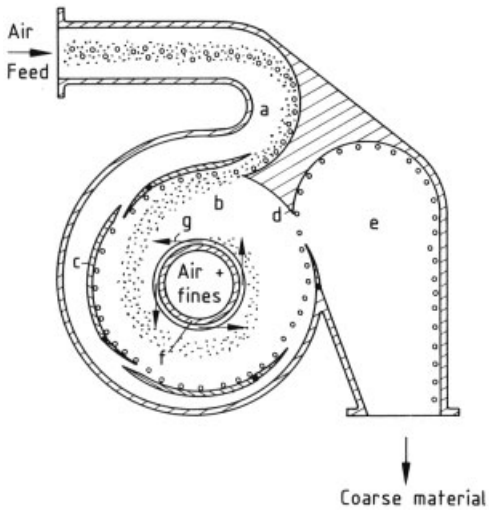

Figure 9. Stratoplex spiral air classifier (Alpine)
a) Intake bend; b) Classification zone; c) Guide vane; d) Coarse-material cutting edge; e) Coarse-material chamber; f) Fines outlet; g) Accelerating air

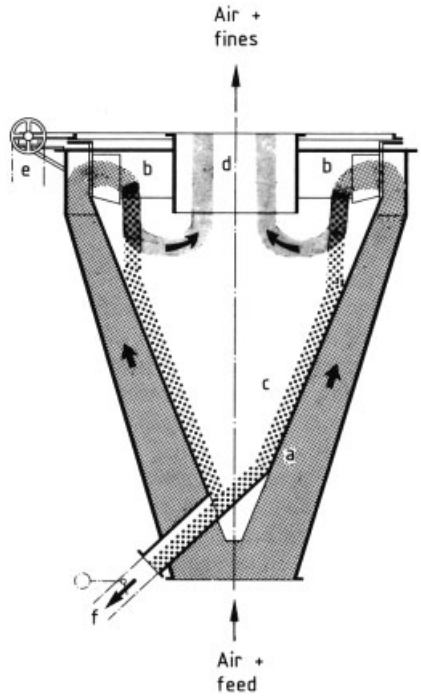

Figure 10. Statopol spiral classifier (Krupp–Polysius)
a) Outer cone; b) Guide-vane ring; c) Grit cone; d) Dip pipe; e) Adjusting device for guide-vane ring; f) Grit outlet

Figure 9 shows the *Stratoplex spiral air classifier.* This is an airstream classifier in which the feed is dispersed in the classifying air. The coarse material is again peeled off by a cutting

edge (d) on the periphery and drops into a coarse-material chamber (e). The ratio between depth and diameter of the classification space, which is of the order of 1 : 4 to 1 : 6, for most spiral classifiers was increased to 1 : 2 for the Stratoplex classifier in order to increase throughput. The air loaded with fines is, therefore, exhausted on both sides (f). To prevent the transport of coarse material in the boundary layer close to the wall, accelerating air (g) is blown in around the central pipes. For products with a tendency toward melt deposition, cooling of equipment parts vital to the operation, such as guide vanes (c) and the coarse-material cutting edge (d), is possible.

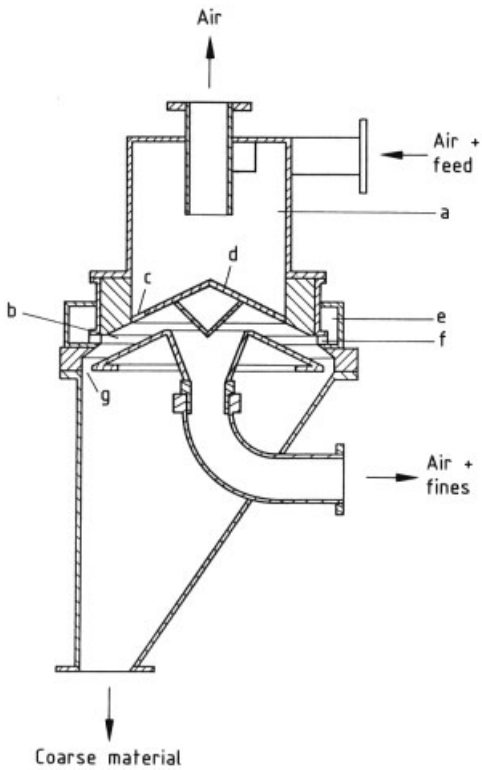

Figure 11. The DS Classifier (Nippon Pneumatic Mfg.)
a) Inlet chamber; b) Classification zone; c) Opening for passage of feed material; d) Guide cone; e) Chamber for classifying air; f) Guide vane; g) Discharge opening for coarse material

The *Statopol* is another airstream classifier (Fig. 10) [6]. The mixture of air and material is fed in over the guide-vane ring on the entire periphery. This classifier has a cyclone-like classification cone instead of a flat-cylindrical separation chamber. Coarse material falls down into this cone; air and fines leave the classification zone through a central pipe.

In the *DS classifier* (Fig. 11) [7], feed material is also introduced already dispersed in air. However, only part of this air is used for classifying. Above the classification zone is an inlet chamber, in which the material is separated from a major part of the transport air and then passed through an annulus into the actual classification zone. The classification zone, in turn, is surrounded by a guide-vane ring, through which the air flows in. Coarse material drops out through an annulus in the lower wall of the classification chamber.

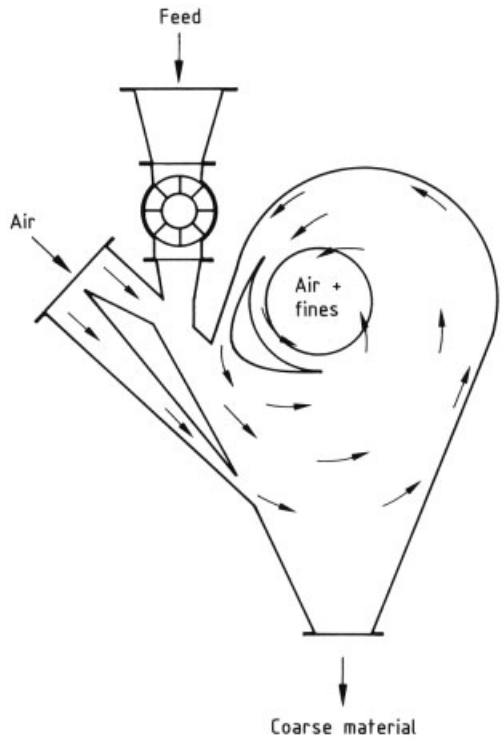

Figure 12. The EC Classifier (Larox)

Figure 12 shows another static spiral classifier, the *EC Classifier*, which is used predominantly in cement classification [8], [9]. The simple machine design makes operation with circulating air possible.

Deflector Wheel Classifier. The distinguishing feature of a deflector wheel classifier is a rotor equipped with deflecting blades, with air flow from the outside to the inside of this rotor.

However, deflection at the blades is limited, and classification takes place in the interior of the rotor in a spiral or spiral-helical flow. The cut point is set by the air flow rate and the rotor speed.

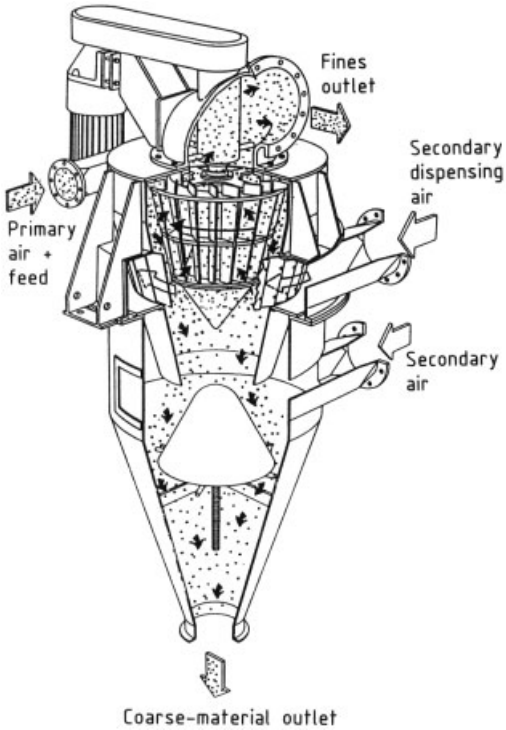

Figure 13. Micron Super Separator (Hosokawa–Nauta)

In conventional spiral air classifiers, the flow adjusts itself more or less freely ("free vortex") and, accordingly, can be influenced by several variables such as the mass flow rate of the feed. Contrary to this, significantly more stable classification conditions can be achieved with deflector wheel classifiers because of the bladed rotors ("forced vortex").

The blades, usually made of flattened irons, are attached axially at the rotor periphery; occasionally cylindrical rods are used instead. In most deflector wheel classifiers, the height of the rotor is approximately equal to its diameter. For reasons of strength, the speed of such rotors is limited. At maximum peripheral speed of ca. 65 m/s, a finest cut point of ca. $x_{50} = 3 - 6\ \mu m$ can be achieved, depending on classifier size. If an even finer classification is desired, flatter high-speed rotors with peripheral speed between 80 and 120 m/s must be used.

To perform classifications with a sharp cut, deflector wheel classifiers are generally combined with special classification zones for the coarse fraction. This type of classifier is frequently used in classifier mills.

The *Micron Super Separator* (Fig. 13) [10] has a conical rotor whose blades are set at an angle to the periphery. Feed material, dispersed in the primary air, is introduced into the classifier. Air and fines are sucked through the rotor and pass into the filter system located downstream. Air flows in through a ring of guide vanes located on the outer periphery of the classification zone, at the level of the impeller. Agglomerates rejected by the rotor are dispersed and again conveyed to the classifying wheel. Additional secondary air enters through the funnel-shaped coarse-material outlet.

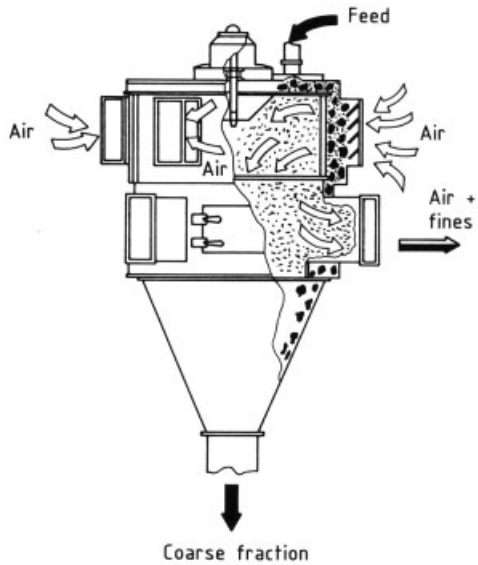

Figure 14. The SD Classifier (Sturtevant)

In the *SD Classifier* (Fig. 14), the rotor is equipped with cylindrical rods. The classification zone for coarse material consists of several conical deflector plates, which are superposed at the periphery of the classifier. Tangentially supplied air flows through these deflector plates. The classifier is loaded by means of gravity pipes [11], [12].

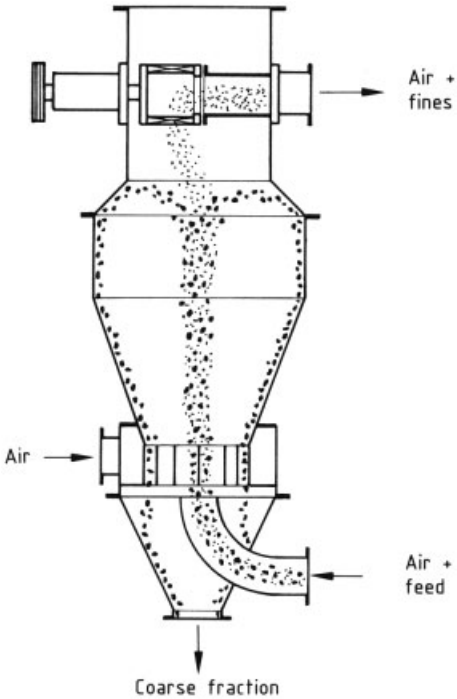

Figure 15. Turboplex high-efficiency air classifier (Alpine)

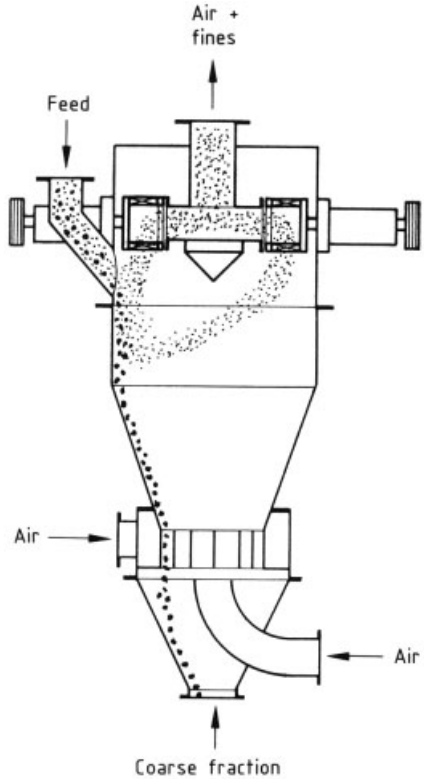

Figure 16. Turboplex multiwheel classifier (Alpine)

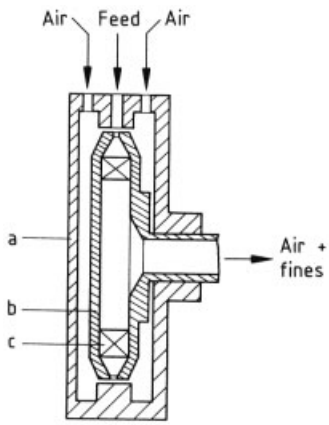

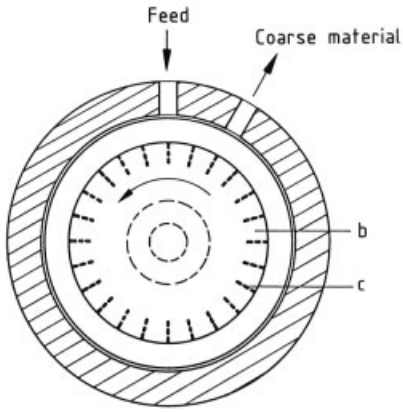

Figure 17. Centrifugal classifier (Donaldson)
a) Housing; b) Classifying wheel; c) Ribs

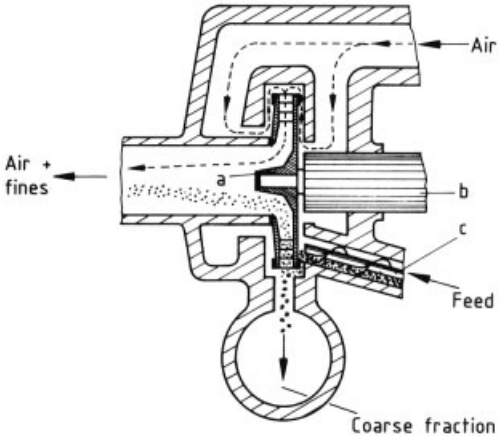

Figure 18. The MZR channel-wheel classifier (Alpine)
a) Rotor with zigzag channels; b) Bearing; c) Feeding screw

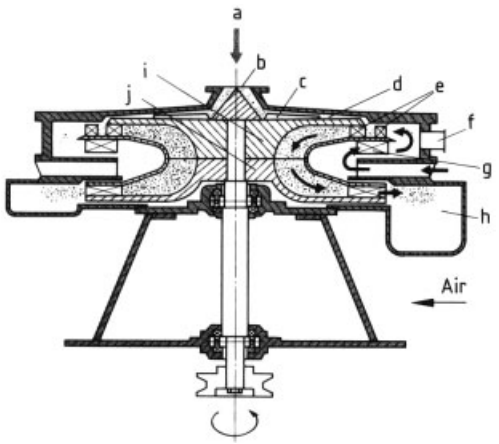

Figure 19. Turbo Classifier (Nishin Engineering)
a) Feed; b) Product inlet; c) Dispersing vane; d) Dispersing disk; e) Classifying vane; f) Coarse-material outlet; g) Auxiliary vane; h) Screw housing (air/fines outlet); i) Classifying rotor; j) Balancing rotor

The *Turboplex high-efficiency air classifier* (Fig. 15), in contrast to other classifiers, contains a rotor with a horizontal axis. Problems of wear or deposit in the fines outlet can thus be controlled easily. The classification zone for coarse material consists of a ring of guide vanes located at the lower end of the cylindrical–conical classifier housing. Figure 16 shows a multiwheel arrangement with which very fine separation can be achieved even with high throughput. The product can be fed to the machine either through gravity chutes or dispersed in a partial airstream.

Air classifiers with *high-speed rotors* use the kinetic energy of these rotors directly to disperse the feed material. However, this dispersing action based on shock and impact can lead to considerable comminution of larger particles and increased wear. Classifiers of this type are manufactured by Donaldson (Fig. 17) [13], Alpine AG (Fig. 18), Nishin Engineering Co. (Fig. 19) [14], and others.

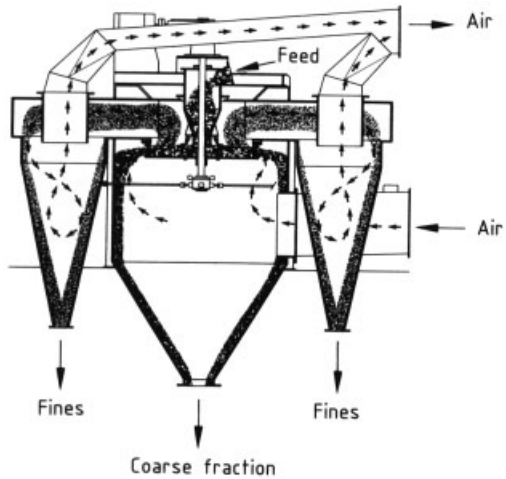

Figure 20. Caropol channel-wheel classifier (Krupp–Polysius)

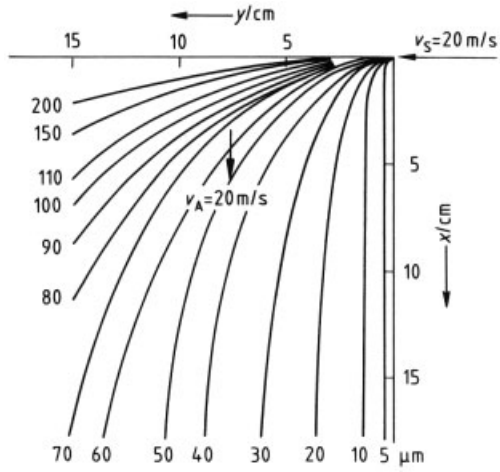

Figure 21. Trajectories of the crosscurrent jet arrangement

Channel-Wheel Classifier (*Elbow Classifier*). The principle of countercurrent elbow classification was described in [15] and [16]. was described in [15] and [16].

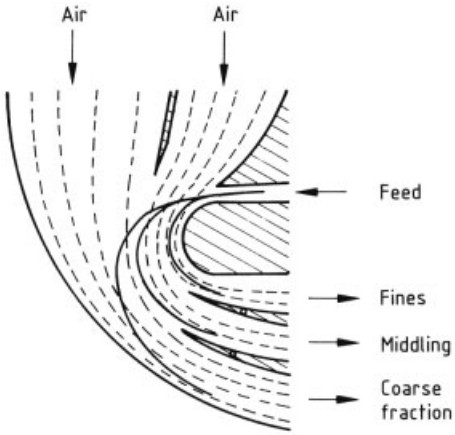

Figure 22. Crosscurrent elbow classifier

In a centrifugal field, this led to the so-called channel-wheel classifier, whose design and mode of operation are shown in Figure 20 [6]. The channel wheel is charged from above with the material to be classified. In the individual channels, particles are accelerated in the direction of the periphery and are thrown into the inward flowing classifying air. In this process, the fines are deflected (as described in Section 4.2.1) and exhausted with the air through openings in the upper part of the channel wheel. The adjustment of cut point is carried out by varying air quantity and speed of the channel wheel, as in deflector-wheel classifiers. The cut-point range is 10 – 150 µm.

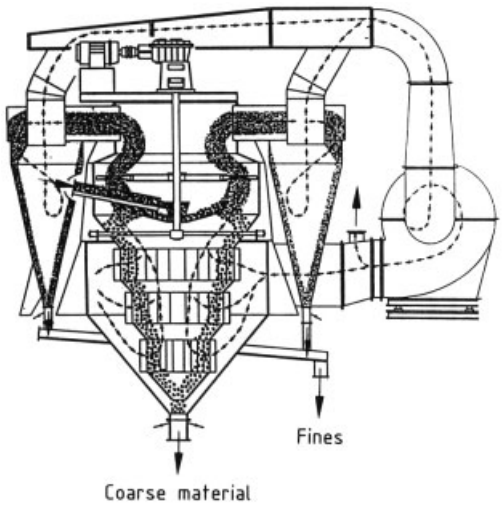

Figure 23. Zyklopol recirculating air classifier/cyclone air classifier (Krupp–Polysius)
External air circuit

4.3. Crosscurrent Classifiers

During the 1960s, the principle of crosscurrent classification was developed [17], [18]. By means of special dosing and dispersing devices, particles were fired at a right angle into a flow channel. The particles fanned out in the airstream move along almost fixed trajectories (Fig. 21). In a case of application, the fractions of fanned-out particles are measured quantitatively by photometric means. This unit serves as a fully automatic *particle-size analyzer* [19].

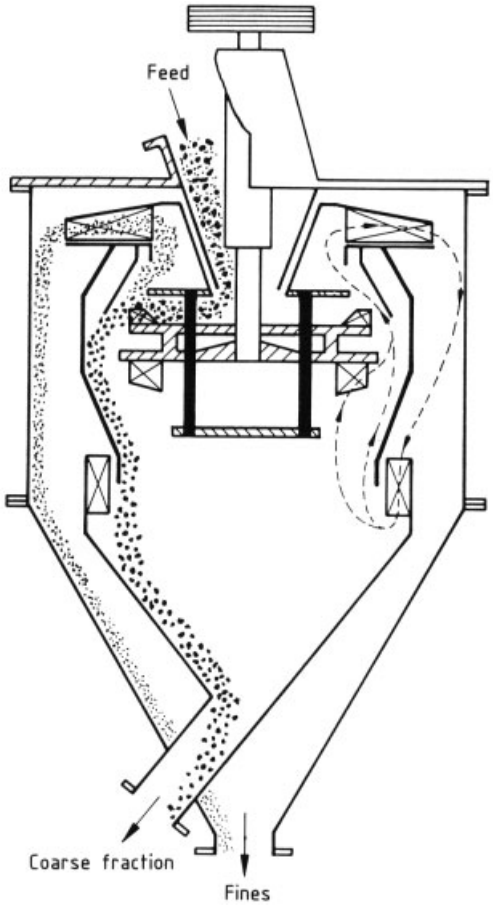

Figure 24. Ventoplex recirculating air classifier (Alpine)
Internal air circuit, main fan and dispersing disk are located on the same shaft

A *crosscurrent elbow classifier* has been developed for large-scale industrial purposes (Fig. 22) [20]. The fanned-out stream of material is divided into several partial streams of corresponding particle-size classes by means of cut-

ting edges. The finest particles again undergo a large deflection during this process.

4.4. Recirculating-Air Classifiers

Recirculating-air (or whizzer-type) classifiers are the oldest industrial air classifiers. The first one was built in 1885 by Pfeiffer in Kaiserslautern (Germany) according to a patent by Moodie and Mumford. This type is still the most frequently used classifier, especially in the production of cement, gypsum, mineral fillers, etc.

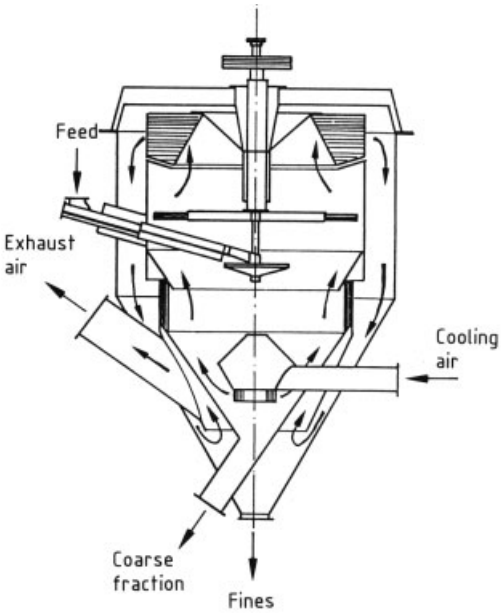

Figure 25. Heyd dispersing air classifier (Pfeiffer)–internal air circuit with supply of cooling air
Adjustable fan, dispersing disk, and main fan are on two separate shafts

The mode of operation is as follows: The material is sent by a dispersing disk into a rotating air flow moving in an upward direction (crosscurrent classification). The coarse material sliding down the wall of the classification chamber is reclassified as it passes over one or more louvers. Classification of the fines takes place in the conical screw-shaped flow, whose peripheral velocity, and therefore cut point, are adjusted by means of the speed or position of a rotor system. Classifying air may be produced by an external fan. With this *external air circuit,* the fines are separated in externally located cyclones (Fig. 23). If the fan is integrated into the head of the classifier, it is referred to as an *internal air circuit* (see Section 4.1). Here, the fines are separated from the air in a cyclone which envelops the classification space. In some designs, the fan, rotor, and dispersing disk are located on the same shaft (Fig. 24). However, two-shaft designs (i.e., a separate drive for the fan and rotor–dispersing disk) are used more often (Fig. 25).

For the classifier shown in Figure 25, the possibility of fresh-air input also exists. To cool the product, a departure from operation exclusively with circulating air is made [21]. The exhaust air must be dedusted in a filter connected on the downstream side.

Recirculating-air classifiers have a separation range of ca. 30 – 200 µm. Because these machines operate with relatively low air speed, they have a satisfactory wear behavior. On the other hand, because of the lack of disintegrating power, their sharpness of cut is not very good.

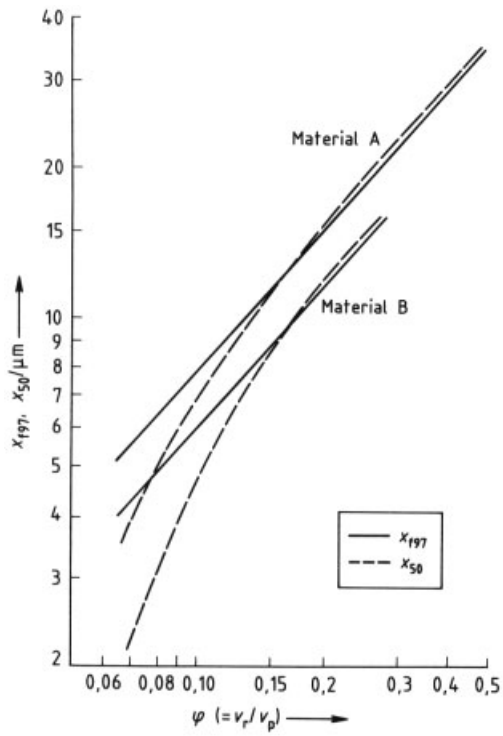

Figure 26. Effect of feed-material distribution on the cut point

5. Operating Properties

5.1. Effect of Feed

Classifiers can be more or less sensitive to changes in the particle-size distribution of the feed material and in the feed rate. If they are loaded with finer feed material with the same setting (speed, air flow rate, load, etc.), then air classifiers of almost all types produce a finer product or a finer cut point. Figure 26 shows this behavior for a deflector wheel classifier. In this figure, the reference cut point x_{f97} and Tromp's cut point x_{50} are plotted against the parameter φ, where φ is the ratio of the radial velocity v_r (which is proportional to air quantity) to the peripheral velocity v_p (which is proportional to rotor speed). The curves show the effect of two different feed materials A and B, in this case limestone powders produced by an impact grinding mill, whose particle-size distributions are shown in Figure 27.

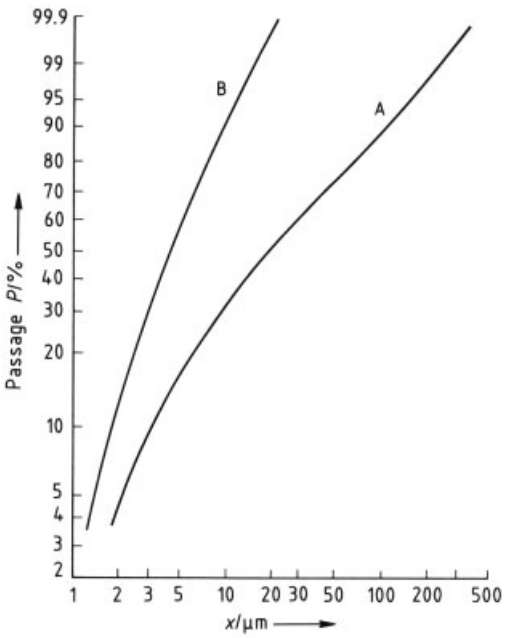

Figure 27. Cumulative passage curves of impact-ground lime stone powders

Individual classifier types exhibit different behavior when the feed rate is increased. In this connection, the terms *hard* and *soft characteristics* are used. The term "hard"designates a classifier that maintains a constant particle-size distribution of the fines with increasing overload, e.g., deflector wheel classifiers, zigzag classifiers, and recirculating-air classifiers. This behavior is illustrated in Figure 28 for a deflector wheel classifier. The fines even become slightly finer with increasing feed rate, and the coarse material becomes increasingly less pure, as manifest by a distinct decrease in the sharpness of cut.

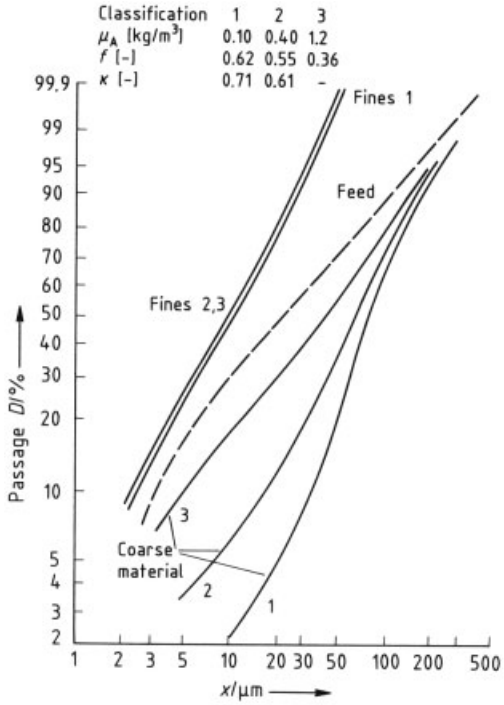

Figure 28. Behavior of a deflector wheel classifier under overload

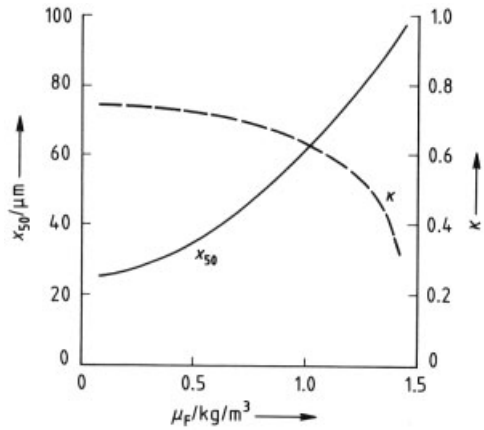

Figure 29. Behavior of a spiral air classifier under overload

Spiral classifiers, on the other hand (free vortex flow), have soft characteristics. When the classification zone is overloaded, they show a shift of the cut point, and thus of the fines distribution, toward the coarser range. The sharpness of cut also decreases continuously (Fig. 29) until the classification finally breaks down.

In closed grinding–classifying systems, this soft behavior is sometimes desirable. In the case of overloading, the classifier is meant to increase production of coarser fines; otherwise, the circuit will fill up continuously.

5.2. Scaleup

When the size of a classifier is increased, the change in throughput may be accompanied by a change in the sharpness of cut and cut point. A shift in the cut-point range occurs primarily with countercurrent classifiers in a centrifugal field and also with recirculating-air classifiers (i.e., machines in whose classification zones more or less ideal spiral flows predominate). The dependence of cut point x_t on $r^{0.6}$, a characteristic parameter of the classification zone, is shown in Equation (8) (see Chap. 2). This proportionality has been confirmed by investigations on different countercurrent centrifugal classifiers.

Smaller classifiers, therefore, achieve a finer cut point under identical classification conditions. The same cut point can usually be achieved by appropriate setting of larger machines, e.g., by increasing peripheral velocity. However, because the peripheral and radial velocities of a classifier are located within quite specific limits for various reasons (pressure drop, material strength, etc.), very fine classification carried out on a laboratory machine, for example, cannot be duplicated on a large-scale classifier.

5.3. Special Requirements

In the selection of a classifier, performance characteristics (Chap. 3) should be the deciding factor. However, other characteristics, such as operating behavior ("soft" or "hard" classification characteristics), are also involved. Often specific material properties or product requirements must be considered, e.g., deposition behavior, wear, product purity, explosion hazard, etc. In many cases, these requirements can be met only by means of special fittings or a special design.

Deposition Behavior. Very fine dusts have a tendency to agglomerate and form deposits on the walls of classification chambers and on parts of the equipment that are important for the operation. The result can be a shift of the cut point and, in the worst case, breakdown of the classification.

In the matter of material deposition, large classifiers have an advantage over smaller machines. Because of the larger dimensions (cross-sectional area), the deposit layer can grow to a certain thickness without significantly interfering with classification. On the other hand, the layer of deposit then becomes so heavy that it may fall off, be blown away, or be thrown off. High air speeds or elastic, "fluttering" linings can be used to address this problem. With products tending toward melt deposits, lowering the air speed, cooling the classifying air, or cooling critical equipment parts is recommended.

Wear. All materials with a Mohs hardness greater than 4 exert a wearing action on the classifier material. Damage to the classifier related to this becomes a problem primarily in machines with high-speed parts; the wear on classifier elements important to the operation can shift the cut point.

As a countermeasure, easily interchangeable parts and replaceable linings are used. No universally usable wear-resistant material exists as yet; depending on the bulk material and location in the classifier, soft or hard steel, hard metal, hard ceramic, soft rubber, or Vulkollan may be suitable.

Because wear increases exponentially with air velocity it can be minimized by throttling the air quantity. However, this measure decreases the sharpness of cut and the throughput.

Product Purity. In addition to the abrasion produced by wear, a product can also be contaminated by corrosion products and lubricants. Grease may pass easily from the bearings into the classification chamber especially where shafts penetrate. Conversely, the product may get into the bearings and destroy them prematurely. Pneumatic sealing has proved to be a suitable remedial measure. When frequent feed

change or a sterile mode of operation is necessary, adequate purification possibilities must be available.

Explosion Hazard. To carry out successful air classification, very fine dusts must be dispersed in large quantities of air. This involves the danger of dust explosions, even with normally harmless products.

Possible protective measures against dust explosions are:

1) avoidance of sparks and ignition sources (e.g., no high-speed machine parts),
2) inert gas blanket, or
3) shock-resistant design.

6. References

1. O. Lauer: *Feinheitsmessungen an technischen Stäuben,* Alpine AG, Augsburg 1962.
2. Keller GmbH, DE 1 059 977, 1959 (H. Ide).
3. F. Kaiser, *Chem. Ing. Tech.* **35** (1963) 273 – 282.
4. Waeschle Maschinenfabrik, DE 3 203 209, 1982 (W. Krambrock, H. Hoppe).
5. H. Rumpf, Dissertation, Technische Hochschule, Karlsruhe 1939.
6. Krupp-Polysius AG, Beckum (brochure).
7. Nippon Manufacturing Co., DE 2 949 618, 1979 (N. Nakayama, K. Yonezawa).
8. R. T. Hukki, *Zem. Kalk Gips Ed. B* (1977) no. 5, 199 – 205.
9. Larox News 2/1985, Larox Oy, Lappeenranta, Finland.
10. Hosokawa-Nauta Europe B.V., Haarlem, The Netherlands (brochure).
11. J. V. Klumpar, N. N. Zoubov, *World Cem.* **16** (1985) Oct., 302 – 305.
12. Sturtevant Inc., Boston, MA, USA (brochure).
13. Donaldson Co., DE 1 657 122, 1968 (C. E. Lapple).
14. Nishin Eng. Co., Tokyo, Japan (brochure).
15. J. Wessel, U. Schmidt, *Aufbereit. Tech.* **12** (1971) no. 7, 402 – 406.
16. J. Wessel, *Aufbereit. Tech.* **20** (1979) no. 9, 475 – 478.
17. H. Rumpf, K. Leschonski, *Chem. Ing. Tech.* **39** (1967) no. 21, 1231 – 1241.
18. K. Leschonski, *Chem. Ing. Tech.* **49** (1977) no. 9, 708 – 719.
19. K. L. Metzger, K. Leschonski in Proc. 1. Europ. Symp. "Partikelmeßtechnik," *DECHEMA Monogr.* **79** (1976) no. 1589/1615, Part B, 77 – 94.
20. L. C. Rumpf, DE 2 538 190, 1975 (H. Rumpf, K. Leschonski, K. Maly).
21. E. W. Hanke, *Zem. Kalk Gips* **28** (1975) no. 9, 105 – 114.
22. Alpine AG, unpublished results.

Air Pollution Control → Air
Alanine → Amino Acids
Albumin → Blood
Alcoholates → Alcohols, Aliphatic
Alcoholic Spirits → Spirits

Mineral Sorting

See also → Solid – Solid Separation, Introduction

HENRY E. COHEN, Royal School of Mines, London, United Kingdom

1. Introduction 1
2. Principles of Sorting 1
2.1. Ore Preparation 2
2.2. Feed Presentation 2
2.3. Inspection and Analysis 3
2.4. Separation 3
3. Performance 4
4. Economic Aspects 5
5. References 6

1. Introduction

Mineral sorting is the oldest method of solid – solid separation in the mining industry. Based upon visual recognition, hand picking was used by early man to separate desired metals (e.g., gold) or ore minerals (e.g., copper and lead sulfides) from associated gangue minerals and unwanted rocks. Hand sorting has persisted where inexpensive labor is available for recovering obvious values, or for removing visible waste ahead of more costly processing (e.g., fine grinding). For example, in 1952 all diamonds from the alluvial deposits worked by Consolidated Diamond Mines in South-West Africa (Namibia) were removed by hand at various stages in the concentration process [9]. This has been superseded by machine sorting. However, final grading according to quality is still carried out by hand sorting because the discrimination of the trained eye cannot be matched by any available machine.

Of around 36 major gold – uranium mines in the Witwatersrand area of South Africa, some 50 % use hand sorting of feeds coarser than 50 mm, or machine sorting (photometric or radiation) down to 25 mm. The main purpose is a substantial rejection of waste (up to 50 %) to reduce the cost of fine grinding. Additional benefits include a smaller chemical treatment plant and an increase in overall throughput capacity. In one uranium mine, radiometric sorting reduces the feed tonnage from 3200 t/d of run-of-mine ore to 2200 t/d of feed to the leach plant. This 31 % reduction is accompanied by a 40 % increase in grade. In most gold – uranium mines, the rejection of ca. 35 % of the feed mass is accompanied by ca. 95 % recovery of the values. Making low-grade deposits exploitable or reducing the need for selective mining is another economic benefit of preconcentration by sorting.

2. Principles of Sorting

Even primitive methods of hand sorting consisted of five consecutive process steps: *preparation, presentation, inspection, data analysis, and separation.* Clean ore must be prepared in relatively narrow size ranges. Generally, particle sizes suitable for sorting most minerals range from about 6 to 180 mm, but treatment has to be carried out on relatively narrow size ranges, usually not exceeding size ratios of about 2 : 1. Sorting tends to be most effective at the largest particle size permitted by liberation requirements (→ Size Reduction).

Each particle must be examined; hence, rates of throughput are clearly related to particle size. In principle, after preparation, either particles are aligned for individual inspection in single files along one or more paths (trajectories), or they are presented as a stream which is wide relative to particle dimensions.

Modern sorting processes are fully automated and use sophisticated systems for scanning, recognition, and analysis of information. Various sensors have been used or have been tested to detect differences between minerals in photometric transmission, reflection, or emission, including infrared, visible, or ultraviolet light; radiometric transmission, reflection or emission, including ultrasound, radio frequencies, X-rays, gamma rays, etc.; magnetic susceptibility, elec-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_18

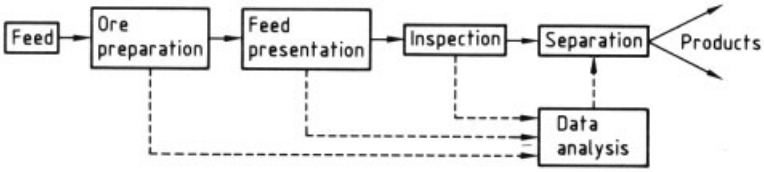

Figure 1. Operational sequence in ore sorting

trical conductivity [10], and other measurable properties. Examples of minerals and metals treated by sorting include asbestos, copper, diamonds, fluorspar, gold, limestone, magnesite, scheelite, silver, uranium, and wolframite.

A scanning or inspection system derives information which is analyzed with respect to selected parameters. The location of each particle is defined, and each is either accepted or rejected. The sorting action is applied preferably to the minority fraction, by use of either an air jet or some mechanical deflector.

The operational sequence in ore sorting is shown in Figure 1, with solid arrows indicating the movement of ore. Broken lines indicate the flow of information to the data-analysis unit and the control signals that activate separation.

2.1. Ore Preparation

Adequate liberation is an essential condition for successful sorting. The structural and textural characteristics of an ore determine the particle-size limits for adequate liberation of sufficiently barren gangue which can be rejected without undue loss of values (→ Size Reduction). Rejection of 25 – 30 % of the feed mass is usually a minimum requirement for sorting to be economically justifiable.

Normal preparation sequences include coarse crushing to, for instance, particles smaller than 180 mm, followed by washing to remove dust if surface characteristics are important for sensing. Washing is unnecessary for sorting by radiometry or magnetic susceptibility but essential for sorting by color, reflectance, or fluorescent emission. The feed must be screened into narrow size groups for separate sorting, because both presentation and inspection are very size dependent. Typical size groups are particles + 100 mm (i.e., > 100 mm), 100 – 50 mm, 50 – 25 mm, 25 – 12.5 mm, and − 12.5 mm (i.e., < 12.5 mm). However, specific size limits may have to be chosen in relation to the individual characteristics of a given feed material.

Surface conditioning can be applied at this stage. For example, an aluminum hydroxide flocculant can be made fluorescent with a dye and can be used to adhere selectively to asbestos [11].

2.2. Feed Presentation

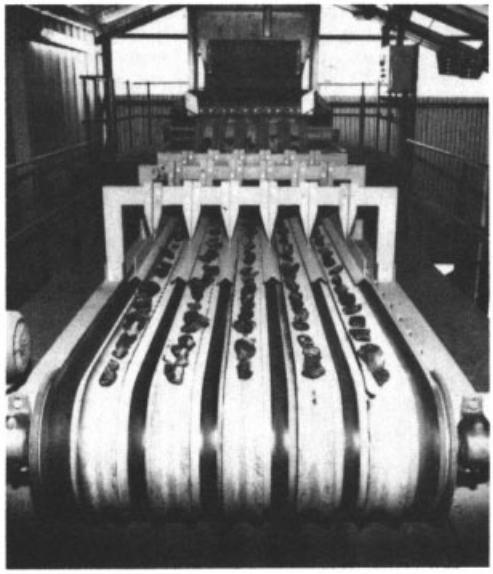

Figure 2. Feed belt of a 5 channel sorter

Feed presentation generally delivers the sized material from a surge bin via vibrating or belt feeders to chutes which channel the particles to a conveyor. Assistance may be provided by various types of accelerating rolls or belts to separate the particles. Different manufacturers adopt varying combinations of feeders, chutes, and belts to achieve closely spaced streams of discrete particles. Spaces between particles are

kept as small as possible in order to maximize the throughput rate. Typical feed presentation for a radiometric sorter is shown in Figure 2.

Fast movement is important for attaining high throughput rates, and linear speeds in the range of 3 – 5 m/s are common. Feed rates per channel vary from about 60 t/h for coarse material to about 6 t/h for fine sizes. A sectional overall view of a radiometric sorter installation is shown in Figure 3.

Some sorters are designed to inspect and sort particles in flight along a trajectory, rather than during transportation on a conveyor. For this purpose, particles must be aligned in a close-spaced single-file stream that follows an accurate trajectory. The in-line feeder consists basically of two closely spaced counterrotating rolls, whose axes are almost parallel and inclined to the horizontal. This arrangement simulates a V-shaped inclined chute but is able to achieve variable particle speeds by altering the speed of the rolls. Typically, the rolls are 760 mm in diameter and 4.3 m long. The sharp cusplike space between rolls breaks up groups of particles and aligns the particle lengths parallel to the rolls. The resulting in-line stream exposes maximum surface areas for inspection [11].

2.3. Inspection and Analysis

The inspection of particles must provide all the information on which decisions to accept or reject are based. This includes the following:

1) Quality – a diagnostic reaction to an applied radiation or signal, or detection of an inherent property such as radioactive emission. Quantitative measurement may be needed, but detection above a set threshold may suffice.
2) Size – a measure of particle size is necessary for quantifying the quality measurement.
3) Speed and position – the speed of the feed stream must be verified and the precise location of a particle in the stream must be defined to operate the ejector mechanism at the correct time lag after inspection.

These various inspection functions usually require complex arrays of sensing heads which repeatedly examine the stream of particles at millisecond intervals. Their signals are integrated and analyzed by the sorting microprocessor in relation to stored memory data. The microprocessor makes corrections (e.g., according to size), calculates the grade of particles, and compares the grade with preset values. A precisely timed decision signal is sent to the appropriate ejector if a particle is assessed as unacceptable.

Higher costs of more refined sensing and microprocessor systems should be balanced against possible gains in resolution: for example, in signal-to-noise ratio, surface detail, or spectral band-width. Resolving power imposes important limitations on the particle size that can be treated, the quality of separation, and the throughput rate.

A separate microprocessor for machine control usually interacts with the sorting microprocessor to ensure correct interlocking of machine parts and sequencing of process events. It also identifies malfunctions (e.g., of sensors, cameras, ejectors, timers, or power supplies) and automatically initiates emergency action such as stopping the feed and indicating the nature of a fault.

2.4. Separation

A highly controlled air jet for ejecting particles from a stream moving in a free flight trajectory is the most widely used method of separation. Both mechanical and hydraulic devices have been described, but they cannot match the speed of air-ejector valves. A blast of compressed air at about 500 – 700 kPa is triggered at a precise moment to deflect a particle from its trajectory into a separate receiver. The force of the airstream used to eject 30-mm particles is 6 kg/cm^2 [12]. The ejector valves have minimum open times of 3 – 4 ms and should generate close to "square-wave" impulses with the shortest possible rise and decay times. A typical sorter, for instance for 15 – 30-mm particles may have a presentation width of 800 mm and employ 80 valves, or one valve every 10 mm. Thus, the ejector resolution is 10 mm across and 12 – 16 mm in the direction of travel of the particles. Larger particles (up to 180 mm) require longer blasts, of ca. 7 – 8 ms with 20-mm width and at least 28-mm length in the direction of travel. Each valve can fire up to 120 times per second, with considerable noise emission. Soundproofing and remote operator controls are necessary.

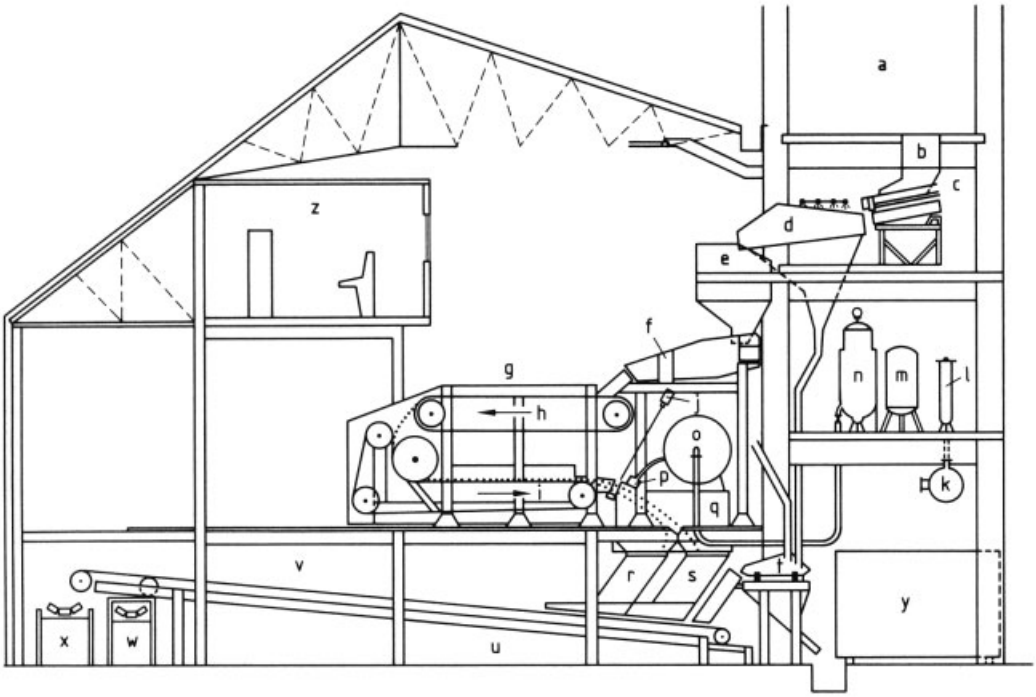

Figure 3. Section of a radiometric sorter plant
a) Main bin; b) Discharge chute; c) Vibrating feeder; d) Washing screen; e) Surge bin; f) Machine feeder; g) Radiometric sorting machine; h) Feeder belt (1.5 m/s); i) Main belt (5 m/s); j) Camera; k) Air manifold; l) Air filter; m) Oil absorber; n) Dryers; o) Air receiver; p) Ejector valve; q) Division plate; r) Accept; s) Reject; t) Screen; u) Accept and reject conveyors next to each other; v) Reef picking area; w) Accept to plant; x) Reject to dump; y) Compressors; z) Control room

Table 1. Open circuit operation (three months) at Vaal Reefs Gold Mine∗

Material	Total production, t	Split, wt %	Gold grade, g/t	Gold distribution, %	Uranium grade, kg/t	Uranium distribution, %
Feed	8136	100	2.40∗∗	100	0.125∗∗	100
Accept	4358	53.6	4.25	94.7	0.220	93.9
Reject	3778	46.4	0.28	5.3	0.017	6.1

∗ Total sorting time 260 h.
∗∗ Calculated feed grade.

3. Performance

Each sorting application has specific constraints that arise partly from the nature and particle size of the ore and partly from limitations of the equipment. For example, incomplete liberation always enforces an arbitrary grade – recovery trade-off. Contained values of an ore and the price of the end products limit acceptable process costs.

For example, radiometric sorting for scalping coarse waste from a gold – uranium ore at Vaal Reefs Mine in South Africa is shown in Table 1 [13].

Machine availability is quoted as 97.2 %. In this example, whether or not the gold loss of 5.3 % is acceptable, in return for removing 46.4 % of the feed mass prior to fine grinding, remains a purely financial decision.

Another typical performance in scalping coarse waste (25 mm) by radiometric sorting from a gold ore is shown in Table 2.

Table 2. Scalping of coarse waste from gold ore

	Mass, %	Gold content, g/t	Gold distribution, %
Feed	100	6.61	100
Product	63.5	10.02	96.3
Reject	36.5	0.67	3.7

Table 3 illustrates the performance in separating in separating magnesite from silicate gangue (particle size 40 – 70 mm) by optical sorting based on color difference at a throughput rate of about 60 t/h.

The products from all three examples proceed to further treatment with substantially reduced tonnage and relatively small loss of values.

In the recovery of diamonds, optical sorters remove diamonds because they are lighter in color than most waste minerals, while X-ray sorters remove them because of their fluorescent response. Sorting is applied to gravity concentrates in which diamonds form a small minority of particles. Feeds of particles 4 – 8 mm in size are treated at rates of 300 – 350 kg per channel per hour. Diamond recovery is 98 – 99 %, and the product is only 0.3 – 0.5 % of the feed mass. This product is sorted further by hand.

Table 3. Separation of magnesite from silicate gangue

	Mass fraction, %	Magnesite content, %	Distribution, %
Feed	100	30	100
Product	32	91	97
Reject	68	2	3

4. Economic Aspects

With wide ranges of particle sizes, different types of applications, and variations in local cost factors, general validity cannot be assigned to the somewhat scarce published costs of sorting. In 1985 capital costs of sorters appear to range from ca. \$ 11 000 to \$ 3000 per tonne hour capacity for 6-mm feed and 150-mm feed, respectively. Approximate costs of installation and commissioning range from 50 to 150 % of machine costs, depending on convenience (i.e., distance from supplier and ease of access) and climate (cold climate requires housing, heating, and lighting; hot climate requires airconditioning, dust extraction, etc.).

Among the factors contributing to operating costs, electric power ranges from 0.2 to 3.0 kW · h/t and compressed air from 5 to 30 m^3 per tonne of feed, or ca. 50 – 60 m^3 per tonne ejected. Costs of spares and maintenance are generally claimed to be very low, around \$ 0.07 – 0.14 per tonne treated. If the ore must be washed, water consumption appears to be around 100 L per tonne of feed. A quotation of operating costs for sorting limestone (70 – 130 mm) in Finland in 1970 gives \$ 0.05 per tonne for power, labor, and maintenance [14].

Operator costs probably constitute the largest and most variable component of operating charges. In general, operator attention appears to range from 0.1 to 2.0 min per tonne of feed, i.e., a throughput of 200 t/h could require one to six operators.

An interesting example of cost is available from the Daping tungsten mine in China [12]. The ore contains wolframite associated with

Table 4. Costs of preconcentration at Daping mine∗

Items	Cost basis, \$	Cost per month, \$	Cost per ton of ores, \$
Wages			
3 operators	16.6	49.65	0.04
18 hand-sorting workers	16.6	297.9	0.22
Electricity, 770 kW · h per month	0.025	57.75	0.04
Compressed air, 17.8 m^3 per ton of feed	0.0028	66.0	0.05
Depreciation charge for equipment, 5 % per year	525 (three sorters)	23.85	0.02
Consumption of materials, 4 % of investment per year		19.12	0.02
Total		514.27	0.38

∗ Plant works three-shift basis, 550 h per month, eleven months per year; 1350 t in 30 – 20-mm size range per month, 20 % of which is quartz.

quartz veins in metamorphic host rocks of generally dark color, and it is sorted for preconcentration in a size range of 20 – 30 mm. Magnetic detectors are used to eject wolframite, which is dark, but has higher magnetic susceptibility than the host rock. At the same time, optical detectors are used to eject white quartz (20 % of the feed) because this contains unliberated wolframite. Machine sorting (three machines with one operator each) is followed by hand sorting of both products (18 operators), which permits a substantial reduction in operators compared with hand sorting alone. The feed grade of 0.11 % wolframite is raised to 0.20 % by machine sorting and to 0.49 % by subsequent hand sorting, an enrichment ratio of 4.5 at 91.9 % recovery. The costs of this preconcentration during the period July 1981 to November 1982 are shown in Table 4.

5. References

General References

1. N. L. Weiss (ed.): *SME Mineral Processing Handbook,* Society of Mining Engineers, AIMMPE, New York 1985.
2. B. A. Wills: *Mineral Processing Technology,* 3rd ed., Pergamon Press, Oxford 1985.
3. G. Tarjan: *Mineral Processing,* Akademiai Klado, Budapest 1981.
4. P. Somasundaran (ed.): *Advances in Mineral Processing,* SME, Colorado 1986.
5. L. White (ed.): *E & MJ Second Operating Handbook of Mineral Processing,* McGraw-Hill, New York 1980.
6. R. P. King: "Automatic sorting of ores," *Miner. Sci. Eng.* **10** (1978) 198 – 207.
7. D. McLaughlin: "Operation of an optical ore sorting system," Am. Min. Congress, Los Angeles, 1979.
8. M. A. Schapper: "Beneficiation at large particle size using photometric sorting techniques,"*Aust. Min.* **69** (1977) April 44 – 53.

Specific References

9. R. G. Weavind: "The treatment and recovery of refractory diamonds," *J. Chem. Metall. Min. Soc. S. Afr.* (1952) 243 – 264.
10. A. Balint: "Ore sorting according to electrical conductivity," *J. S. Afr. Inst. Min. Metall.* Special Issue, October 1975, 40 – 42.
11. D. Collier et al.: "Ore sorters for asbestos and scheelite," 10th International Mineral Processing Congress, London 1973, pp. 1007 – 1022.
12. Shi Fengnian: "Magnetophotometric sorter and its application at a wolframite mine," Mineral Processing and Extractive Metallurgy IMM International Conference, Kunming 1984, pp. 589 – 597.
13. D. G. Kidd, N. P. G. Wyatt: "Radiometric sorting of ore," 12th CMMI Congress, Johannesburg 1982, pp. 645 – 651.
14. Gunson's Sortex Ltd., Personal Communication.

Magnetic Separation

See also → Solid – Solid Separation, Introduction

William J. Bronkala, Applied Magnetic Systems, Inc., Greenfield, Wisconsin 53 220, United States

1.	**History**	1
2.	**Principles of Magnetic Separation**	1
3.	**Types and Basic Application of Magnetic Separators**	3
3.1.	**Tramp Iron Removal**	3
3.2.	**Mineral Concentration and Product Purification**	6
3.2.1.	Wet Magnetic Separators	6
3.2.2.	Dry Magnetic Separators	7
4.	**Applications**	10
5.	**References**	11

1. History

In the mid-1800s, Faraday demonstrated in a series of experiments that any substance placed in a magnetic field will either increase or decrease, to a varying extent, the magnetic field passing through it. This variation in the magnetic susceptibilities of different materials makes possible the phenomenon of magnetic separation (see Chap. 2).

Based on its magnetic susceptibility, a material can be classified as ferromagnetic (strongly attracted magnetically), paramagnetic (attracted magnetically to some degree), or diamagnetic (repelled by a magnetic field).

All materials may be termed magnetic, although the relative value of the induced magnetism might be very small. *Ferromagnetic materials* have magnetic properties similar to iron, and some ferromagnetic materials will retain their magnetic properties in the absence of an applied magnetic field. This phenomenon is known as *remanence. Paramagnetic materials* have a positive magnetic susceptibility (showing a weak attraction to a magnet), and *diamagnetic materials* have a negative magnetic susceptibility (showing a weak repulsion from a magnet). Ferromagnetic materials are amenable to *low-intensity separation* and paramagnetic materials to *high-intensity separation*.

Since the work of Faraday, numerous magnetic devices have been developed that utilize differences in magnetic susceptibility to obtain physical separation of mineral products. The first successful commercial application of magnetic separators was the separation of iron from brass in the late 1860s, followed by the separation and concentration of the ferromagnetic mineral magnetite.

Low-intensity electromagnetic separators of various configurations were developed after 1880 for wet and dry magnetic separation of ferromagnetic materials; these were followed in the early 1900s by high-intensity magnetic separators and permanent magnet separators. The first *permanent magnet material* was an aluminum – nickel – cobalt – iron alloy designated ALNICO, which was followed in the 1950s by ceramic barium and strontium ferrite permanent magnets, and in the 1980s by rare-earth magnets (→ Magnets and Magnetic Materials).

High-intensity magnetic separators of several types were developed to separate paramagnetic materials of lower positive magnetic susceptibility. The first high-intensity separators were largely confined to dry separation, with subsequent development of high-intensity wet-type separators in the early 1960s and the high gradient magnetic separator (HGMS) in the mid

2. Principles of Magnetic Separation

Magnetic separation is a physical separation of discrete particles with different permeability or susceptibility, based on a three-way competition among tractive magnetic forces; gravitational, frictional, and inertial forces; and attractive interparticle forces. The feed to magnetic separators is split into two or more components. If the separator is to produce a magnetic concentrate, then a weakly paramagnetic or diamag-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_19

netic material in practice simply called nonmagnetic, constitutes the tailings (or reject) product. If liberation is incomplete, a less magnetic (middling) product can result. Each product must be transported into, through, and out of the magnetic separator. The magnetic and competing forces (gravitational, frictional, hydrodynamic, or inertial) tend to reduce the degree of separation.

Definitions (for a more comprehensive treatment, → Magnets and Magnetic Materials). A *magnetic field* is defined by its *magnetic field strength H* (sometimes also called *field intensity*). The field strength of a homogeneous magnetic field in a solenoid with *n* turns per meter is given by

$$H = nI$$

where *I* is the current per turn. (The SI unit for magnetic field strength is A/m.)

The *magnetic flux density B* is a measurement of the *magnetic forces* acting on particles in a magnetic field. In a vacuum, magnetic flux density and magnetic field strength are related by

$$B_0 = \Phi/A = \mu_0 H$$

where Φ is the *magnetic flux* (SI unit weber, $1\,\mathrm{Wb} = 1\,\mathrm{m^2\,kg\,s^{-2}\,A^{-1}}$), *A* is the cross section of the coil, and μ_0 is the *permeability of the vacuum* ($\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N/A^2}$). The SI unit for magnetic flux density is the tesla ($\mathrm{T} = \mathrm{kg\,s^{-2}\,A^{-1}}$), $1\,\mathrm{T} = 1\,\mathrm{Wb/m^2}$.

If a material is placed in a magnetic field, the flux density of this material is given by

$$B = \mu H$$

where μ (SI unit $\mathrm{N/A^2}$) is the permeability of the material.

Another method of calculating *B* for a material is

$$B = \mu_0 (H + M) \tag{1}$$

where *M* (SI unit A/m) is the *induced magnetization* of the substance observed whenever it is placed in a magnetic field.

The *magnetic susceptibility* of a material, χ (volume susceptibility), is dimensionless and is defined as the ratio of induced magnetization to magnetic field strength:

$$\chi = M/H \tag{2}$$

Substitution in Equation (1) gives

$$B = \mu_0 H (1 + \chi)$$

The specific magnetic susceptibility ψ is given by

$$\psi = \chi/\varrho$$

where ϱ is the density of the material. Table 1 lists the magnetic susceptibility of various minerals.

Substances with a positive susceptibility (corresponding to $\mu > 1$) are called paramagnetic; substances with a negative susceptibility (corresponding to $0 < \mu < 1$) are called diamagnetic. As can be shown in Equation (2), substances with a positive susceptibility have a positive magnetization and augment the flux density of a magnetic field; substances with a negative susceptibility have a negative magnetization and weaken the flux density.

A magnetic field exerts a force on each of the two poles of a *magnetic dipole*, making it align with the lines of the magnetic field. Because these forces are exerted in opposite directions, they are equal in a uniform magnetic field. Therefore, the net force on the dipole is zero. However, if the field has a *gradient*, i.e., it varies in space, the force on the dipole will be greater in the direction of the higher field and will be proportional to the magnetic dipole moment and the magnitude of the magnetic field gradient.

Process of Magnetic Separation. Magnetic separation relies on the different behavior of individual mineral particles under the influence of a magnetic field. Ferromagnetic or paramagnetic materials are attracted along the lines of the magnetic force from areas of lower magnetic field strength to higher field strength. Diamagnetic particles are repelled from areas of higher magnetic field strength to those of lower field strength.

The magnetization of various materials is directly dependent on their degree of magnetic susceptibility and the strength of the applied magnetic field (Eq. 2).

Table 1. Magnetic susceptibility of minerals ∗

Mineral	Susceptibility	Mineral	Susceptibility
Magnetite	0.12 – 3.07(a)	Graphite	2.2×10^{-6}
Franklinite	3.7×10^{-3}(a)	Fluorite	-2.85×10^{-7}(b)
Ilmenite	1.5×10^{-3}(a)	Aragonite	-3.92×10^{-7}
Magnetic pyrite	$0.337-5.75\times10^{-3}$(a)	Calcite	-3.63×10^{-7}
Siderite	8.4×10^{-4}	Ruby	4.7×10^{-7}(b)
Hematite	$0.11-1.1\times10^{-3}$(a)	Topaz	4.2×10^{-7}(b)
Zircon	-1.7×10^{-7}	Beryl	8.26×10^{-7}
Limonite	$7-8\times10^{-4}$(a)		3.86×10^{-7}(b)
Corundum	3.4×10^{-7}(b)	Epidote	2.38×10^{-5}
Pyrolusite	$6.2-7\times10^{-4}$(a)	Augite	2.66×10^{-5}
Manganite	4.9×10^{-4}(a)	Adularia	-4.27×10^{-7}
Garnet	3.75×10^{-4}(a)		-3.84×10^{-7}
Quartz	$1.75-4.38\times10^{-4}$(a)		-3.17×10^{-7}(b)
Rutile	1.96×10^{-6}	Diopside	8.8×10^{-6}(b)
Pyrite	$0.2-1.5\times10^{-4}$(a)	Sapphire	5.7×10^{-6}(a)
Zincblende	-2.64×10^{-7}(b)	Cobaltite	5.8×10^{-6}(a)
Dolomite	1×10^{-6}(b)	Feldspar	1.6×10^{-6}(a)
Apatite	2.64×10^{-6}(b)	Limestone	6×10^{-6}(a)
Willemite	1.9×10^{-4}(a)	Red serpentine	4.1×10^{-5}(a)
Chalcopyrite	8.5×10^{-7}(b)	Green serpentine	3.5×10^{-4}(a)
Spinel	0.2×10^{-7}(b)	Antimonite	-8.5×10^{-7}(a)
Galena	-3.5×10^{-7}(b)	Mica, transparent	0.8 – 1.2×10^{-5}(a)
Halite	-5.0×10^{-7}(b)		
Celestine	-3.42×10^{-7}		
	-3.14×10^{-7}		
	-3.59×10^{-7}(b)		
Tourmaline	1.12×10^{-6}(b)		

∗ (a) Volume susceptibility; (b) specific susceptibility.

Ferromagnetic materials are quickly saturated magnetically, and increasing the magnetic field strength will not increase the magnetization beyond a certain point. For paramagnetic materials that are difficult to magnetize, the induced magnetization is proportional to the magnetic field strength applied, and some materials, practically speaking, cannot be saturated.

The efficiency of magnetic separation may be expressed by both the recovery (i.e., the ratio of magnetic material in the magnetics relative to that in the feed) and the grade (i.e., the fraction of magnetic material in the magnetic concentrate).

3. Types and Basic Application of Magnetic Separators

Magnetic separators are used in two basic areas:

1) tramp iron removal–equipment protection, and
2) mineral concentration and material purification.

3.1. Tramp Iron Removal

Tramp iron magnetic separators are used to protect material handling and process equipment such as crushers, pulverizers, and screens. These applications usually involve dry material or material with only surface moisture. Iron coarser than 3 mm is usually defined as tramp iron.

The size and shape of the tramp iron, together with the material handling system in use or proposed, must be considered in selecting magnetic separators suitable for tramp iron removal.

Magnetic equipment that has been developed for tramp iron removal includes

1) magnetic head pulleys–permanent and electromagnetic pulleys used as the head pulley in belt conveyor systems;
2) suspended magnets–magnetic units installed over belt conveyors or feeders;

3) plate magnets, which are used at the bottom of chutes or launders; and
4) grate magnets employed in discharge hoppers.

Magnetic Pulleys. Magnetic pulleys are elongated cylinders supported by a shaft, used as the head pulley for conveyors that transport material in a plant. Magnetic pulleys have magnets installed in them to provide a magnetic field of the same strength around the pulley circumference, which removes any tramp iron contained in the product. Magnetic pulleys are easy to install, have low initial cost, and accomplish continuous and automatic removal of tramp iron. A typical installation is shown in Figure 1. Magnetic pulleys are available in diameters from 0.2 to 6.8 m and in widths to match the conveyor belt width.

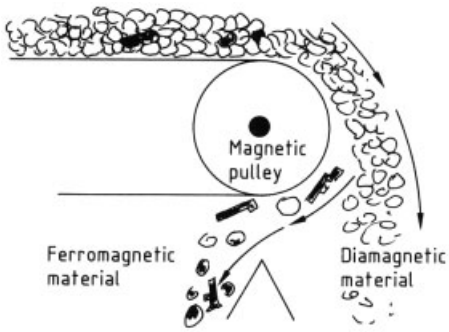

Figure 1. Operating principle of a magnetic pulley

Suspended Magnets. Suspended magnets are rectangular steel boxes that contain permanent or electromagnets and provide a magnetic field to remove tramp iron from material carried on belt conveyors.

These magnets are suspended over conveyor belts or feeders and develop a deep magnetic field through which the material on the conveyor must move. The term deep means that the magnetic field of a suspended magnet can extend to 760 mm (a pulley's effective magnetic field only extends to 130 mm). Any tramp iron contained in the material being transported will be intercepted and attracted to the suspended magnet face.

Suspended magnets are available in both permanent and electromagnetic types, with the electromagnetic unit applied to the deeper loads carried on belt conveyors. The depth of material carried on the conveyor belt, the belt speed, and the clearance required over the material on the belt will determine the size of the suspended magnet. Suspended magnets can be made self-cleaning by installing a conveyor belt that runs over the face of the magnet. A typical application of a suspended magnet installation is shown in Figure 2.

Tramp Iron Magnetic Drums. Magnetic drum separators are used for tramp iron removal when installation of magnetic pulleys or suspended magnets is not feasible. The tramp iron magnetic drum incorporates a magnet assembly held in a fixed position inside a rotating drum cylinder or a shell rotated around this magnet assembly. The active separation zone (within a few centimeters of the surface of the drum) draws the ferromagnetic particles onto the surface of the drum; here, the motion of the drum carries them to a region outside the influence of the magnetic assembly. The ferromagnetic particles then drop or are scraped off the edge of the drum.

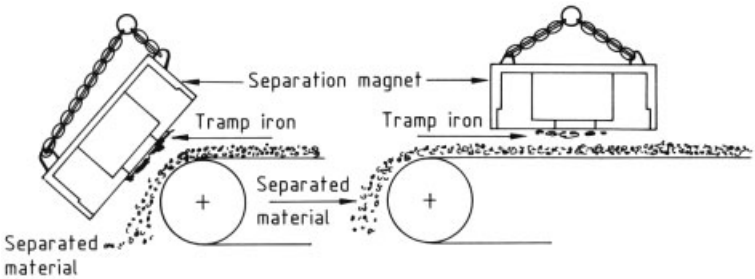

Figure 2. Two ways of mounting suspended magnets to remove core nails from foundry sand on a conveyor belt

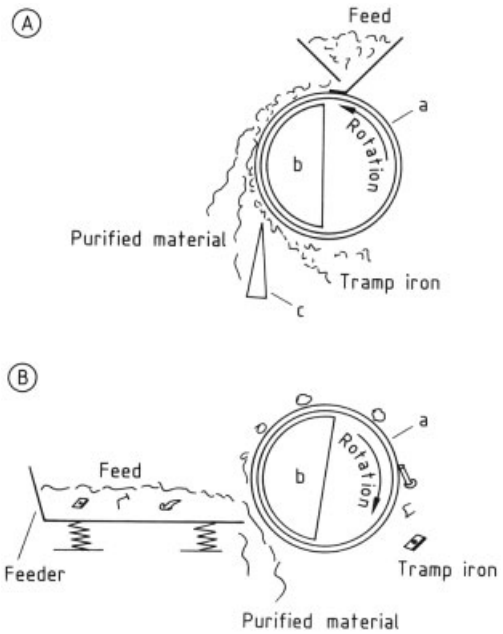

Figure 3. Feed arrangement for magnetic drums
A) Overfeed; B) Underfeed
a) Drum shell; b) Magnet; c) Adjustable nonmagnetic (diamagnetic) splitter

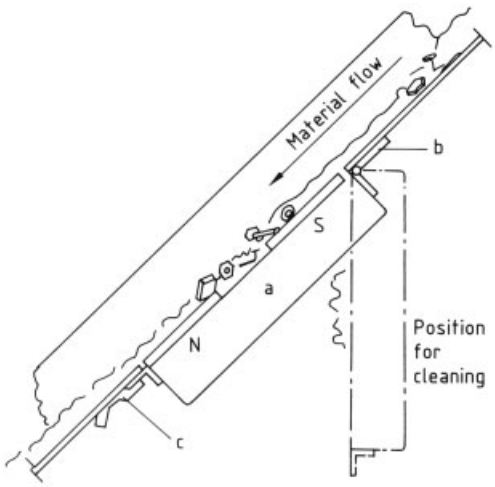

Figure 4. Operating principle of a plate magnet
a) Plate magnet; b) Hinge; c) Latch

Feed material can be introduced on the top vertical center line of the rotating shell (*overfeed* arrangement) or under the drum for clean pickup of magnetic material (*underfeed*). These two feed arrangements are illustrated in Figure 3. Overfeed magnetic drums provide the most complete magnetic removal because gravity is working with the magnetic field rather than against it as in the underfeed arrangement.

Plate Magnets. Plate magnets (see Fig. 4) are installed at the bottom of chutes or launders carrying raw materials. They develop a magnetic field which attracts and holds any tramp iron that might be present. Tramp iron is attracted and held on the face of the plate magnet from which it must be removed periodically by hand.

A chute angle of 45° or less is recommended, and the maximum depth of material flowing over the face of the plate magnet is 115 mm. The plate magnet should be installed as close to the feed point as possible.

Grate Magnets. A grate magnet consists of a set of magnetized tubes installed in a support frame through which feed material is allowed to flow. These tubes, containing the magnet elements, are installed on 25- or 37.5-mm centers, and the feed material is directed through these openings. The magnetic material collected on the magnetized tubes must be cleaned periodically by hand to remove collected ferromagnetics. Grate magnets can also be used to remove tramp iron from slurries or liquids. A typical grate magnet installation is shown in Figure 5.

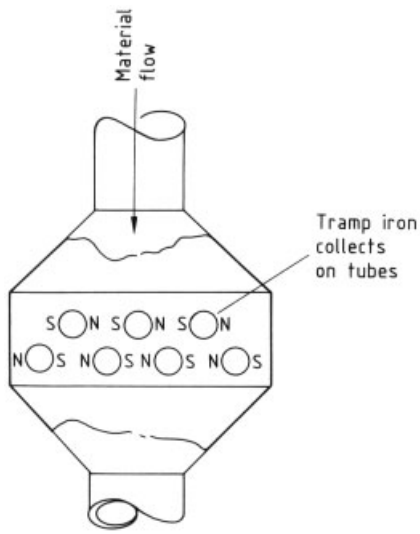

Figure 5. Operating principle of a grate magnet

3.2. Mineral Concentration and Product Purification

The magnetic responsiveness of minerals provides an effective means of concentrating naturally occurring ores.

Magnetic equipment used in *wet mineral concentration* includes

1) wet magnetic drum separators,
2) magnetic filters,
3) wet high-intensity magnetic separators (WHIMS), and
4) high-gradient magnetic separators (HGMS).

Magnetic separators used in *dry mineral concentration* include

1) alternating polarity magnetic drum separators,
2) induced roll magnetic separators,
3) high-intensity crossbelt magnetic separators, and
4) high-intensity disk-magnetic separators.

3.2.1. Wet Magnetic Separators

Wet magnetic drum separators incorporate a stationary electro- or permanent magnet assembly held in a fixed position within a revolving drum shell. The magnetic drum is mounted in a nonmagnetic tank arrangement.

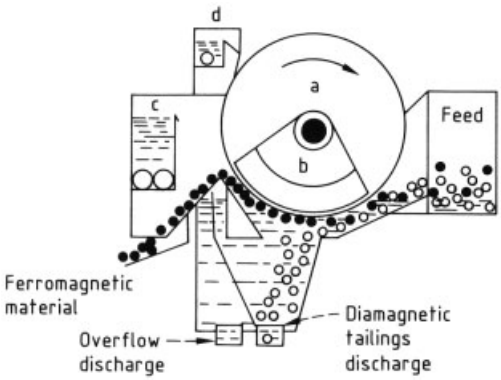

Figure 6. Wet drum magnetic separator
a) Rotation drum; b) Stationary magnet assembly; c) Repulping box; d) Drum wash

The tank arrangement has a feed entry point as well as collection hoppers for removal of the weakly paramagnetic or diamagnetic ("nonmagnetic") and ferromagnetic products of the magnet assembly. A typical arrangement is shown in Figure 6.

The feed to a wet drum separator is typically a slurry in which the products to be separated are carried, with water as a medium.

Wet drum separators have generally been used to separate ferromagnetic particles such as magnetite and ferrosilica from diamagnetic silica and coal fines.

Magnetic Filters. A magnetic filter is a simple device in which a direct-current electric coil or permanent magnets are used to inductively magnetize a steel grid. The steel grid is arranged in such a way that the feed slurry or liquid passes through it. The grid has a large number of magnetized edges that serve to collect any ferromagnetic particles present in the slurry or liquid feed.

Periodically, the filter assembly must be cleaned of the ferromagnetic particles collected, which means that the feed must be stopped or bypassed to permit washing or back flushing of the filter element. Magnetic filters are used to clean oil, paint, clay slip, and other liquids or slurries containing fine iron contamination.

Wet High-Intensity Magnetic Separators. The WHIMS magnetic separator employs a rotating carousel in which a matrix of vertically grooved plates, enclosed wedge-shaped bars, expanded metal mesh, or steel balls is fixed within a ring. Figure 7 shows the operation of a carousel WHIMS unit. At each feed point, the slurry to be separated is introduced and the magnetics are collected on the matrix with the diamagnetic "nonmagnetics" passing into a "nonmagnetic" collection hopper below the separator. As the matrix rotates, it is washed to remove any physically entrapped "nonmagnetics."

Continued rotation of the matrix brings the collected ferro- and paramagnetics into an essentially unmagnetized zone where they are washed off by a high-pressure water jet. These WHIMS separators have high magnetic field strength (intensity) and high magnetic field gradient and can separate very weakly paramagnetic particles.

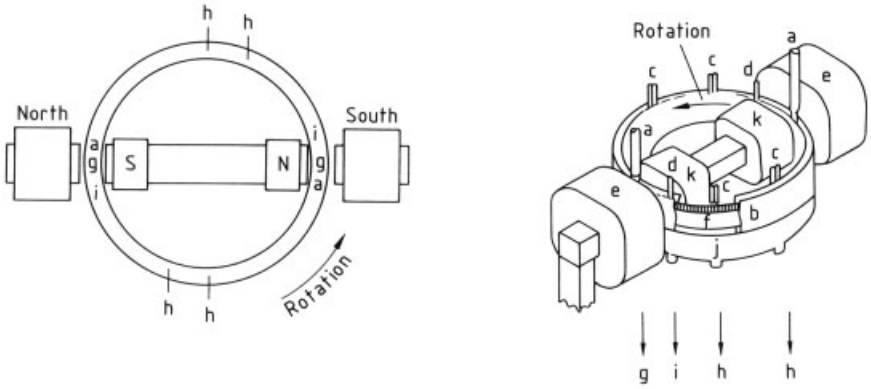

Figure 7. Wet high-intensity magnetic separator (WHIMS)
a) Feed pipe; b) Rotor; c) High-pressure water jet; d) Low-pressure water jet; e) Outer coil; f) Matrix; g) "Nonmagnetics" discharge; h) Magnetics discharge; i) Middlings discharge; j) Trough; k) Inner coil

High-Gradient Magnetic Separators. The HGMS magnetic separator also has a high magnetic field strength and a high magnetic field gradient, but it employs a static magnetized matrix canister through which the feed slurry is passed.

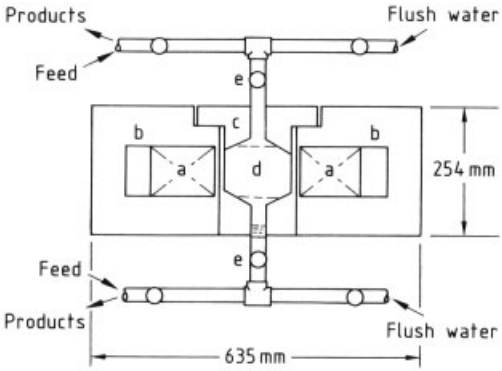

Figure 8. Typical flow through a high-gradient magnetic separator (HGMS)
a) Coil; b) Mild steel circuit; c) Nonmagnetic stainless steel matrix housing; d) Matrix area; e) Feed control for overfeed or retention time control for underfeed operation

Typically, this canister has a matrix of packed stainless steel wool which acts as a collecting element for the magnetic particles. Copper coils which contain hollow conductors energize the canister and are cooled by circulating water or liquid helium circulating through these hollow conductors to permit the development of magnetic flux densities as high as 5 T. This extremely high flux density will extract very weakly paramagnetic impurities. A typical application is in kaolin cleaning. The separation of paramagnetic biomaterials is described elsewhere (→ Biochemical Separations, Chap. 8.).

When the matrix is saturated with particles of a positive magnetic susceptibility, water is introduced to wash the matrix of diamagnetics. The magnet is then deenergized, and the matrix is flushed with high-velocity water to remove the paramagnetics. Processing of the slurry is resumed after reenergization of the magnet. Typical flow through an HGMS magnetic separator is shown in Figure 8.

3.2.2. Dry Magnetic Separators

Magnetic Drum Separators. Two types of magnetic drum are used in mineral concentration or product purification. In the first type, as illustrated in Figure 9, the poles of the stationary magnet assembly *alternate across the face width* of the drum. This presents a strong pattern for holding the ferromagnetics to be removed and permits use with coarse material (up to 305-mm diameter).

The second type of magnetic drum, as illustrated in Figure 10, uses a magnet assembly in which the magnetic poles *alternate around the circumference* of the drum. This type of magnet assembly usually has a greater length of magnet arc (that part of the drum cylinder–expressed in degrees–under which the magnet is located) and develops considerable agitation and reorientation of the ferromagnetic particles as they are

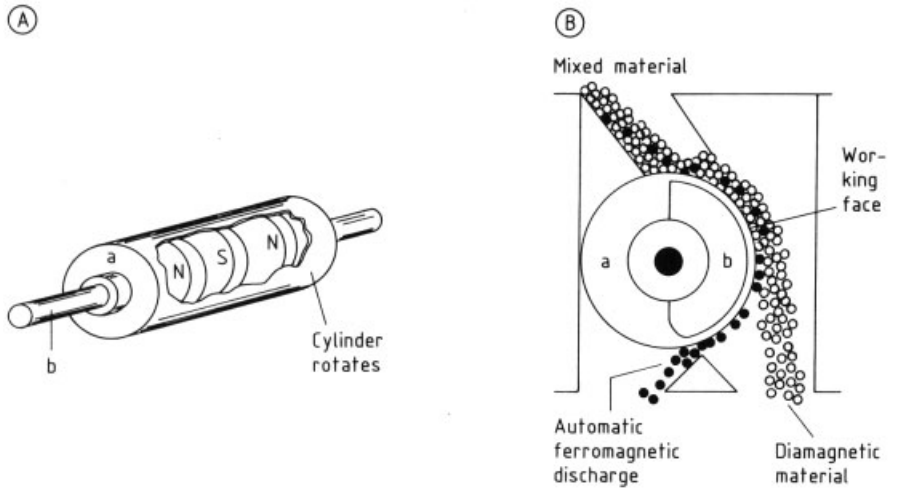

Figure 9. Magnetic drum separator with radial poles alternating across the face width of the drum
A) Arrangement of poles: a) Drum heads; b) Stationary shaft
B) Principle of operation: a) Revolving cylinder; b) Stationary magnet assembly

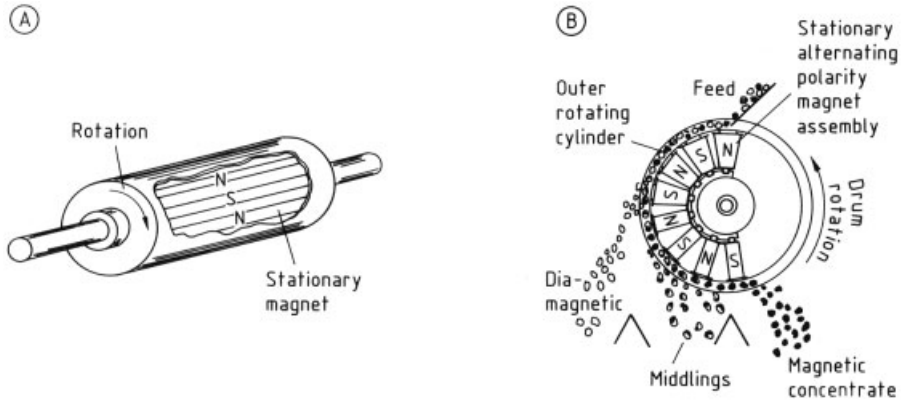

Figure 10. Magnetic drum separator with axial poles alternating around the circumference of the drum
A) Arrangement of poles; B) Principle of operation

carried from the feed to the magnetic discharge point. A magnetic drum is usually fed by a vibrating or other type of spreading feeder *on the top vertical center line* of the drum. Ferromagnetic particles are attracted and held to the drum shell until they pass beyond the end of the magnet assembly, where they are discharged into a magnetic product collection hopper.

In other applications where maximum purity of the ferromagnetic product is desired, such as in municipal waste treatment, an *underfeed* such as illustrated in Figure 11 is used. In this application, the ferromagnetics must be picked up against the influence of gravity to reach the magnetic drum face.

Induced Roll Magnetic Separators. The induced roll magnet separator is a high-field-strength magnetic separator that achieves a high field gradient and can remove paramagnetic minerals from a dry granular feed. Because of the narrow magnetic gaps used on this type of separator, the feed is usually less than 3 mm in diameter, and because of the surface activity of very fine material, all particles $\leq 74\,\mu m$ (≤ 200 mesh) in diameter should be removed.

The induced roll separator has one or more rotating rolls, made up of alternating steel and diamagnetic disks that are inductively magnetized by an electro- or permanent magnet source. The principle of operation is shown in Figure 12.

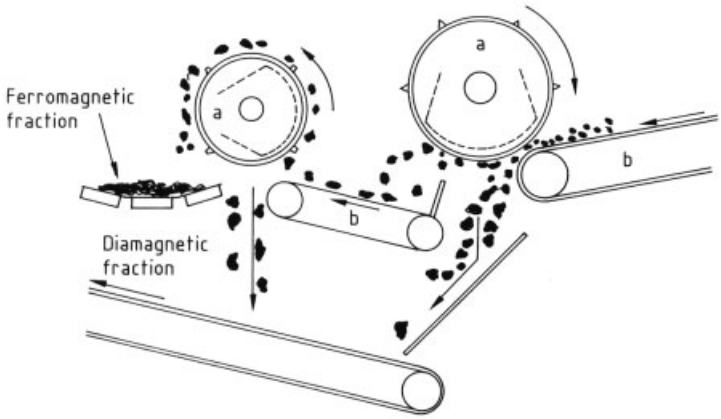

Figure 11. Underfeed magnetic drum separator
a) Magnet; b) Belt conveyor

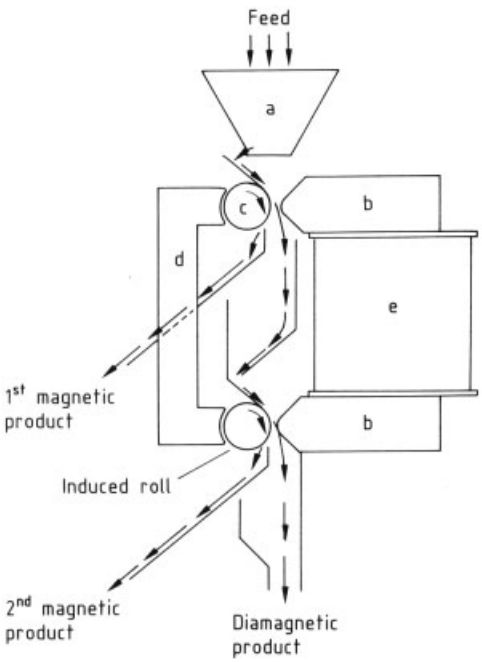

Figure 12. Induced roll magnetic separator
a) Feeding hopper; b) Pole piece; c) Induced roll; d) Bridge bar; e) Coil

Feed is introduced from a hopper or feeder on the top vertical center line of the roll, and the highly magnetized edges of the steel disks serve to attract and hold the weakly magnetic mineral to the roll surface. A change in polarity occurs as the roll rotates, and the collected magnetics are discharged after the roll reaches its null point.

Various adjustments, such as air gap, roll speed, splitter setting, feed rate, and field strength, can be used to control induced roll separators.

High-Intensity Crossbelt Magnetic Separators. For very selective concentration of weakly magnetic (paramagnetic) minerals, the high-intensity crossbelt separator has been used. This separator employs a feed belt on which a thin layer of the feed material is introduced to a high-strength magnetic field. Paramagnetic materials are lifted to the upper pole of this field and transferred by a crossbelt to a collecting hopper. To obtain the high field gradient required for separation of weakly magnetic minerals, the upper pole is shaped to a point (or series of points) for improved capacity, while the bottom pole is flat. When various weakly magnetic minerals are to be concentrated, a series of high-intensity poles is used, with coil strength increasing at each pole, or a variation in the field strength of the air gap may be employed. The operating principle of the high-intensity crossbelt magnetic separator is shown in Figure 13.

High-Intensity Disk-Type Magnetic Separators. The basic design and operation of a high-intensity, disk-type, rotor separator are shown in Figure 14. The induced magnet is a steel ring arranged to rotate between the feed belt and the primary magnet in such a way as to discharge para- and ferromagnetic products to the sides of the belt. This separator is

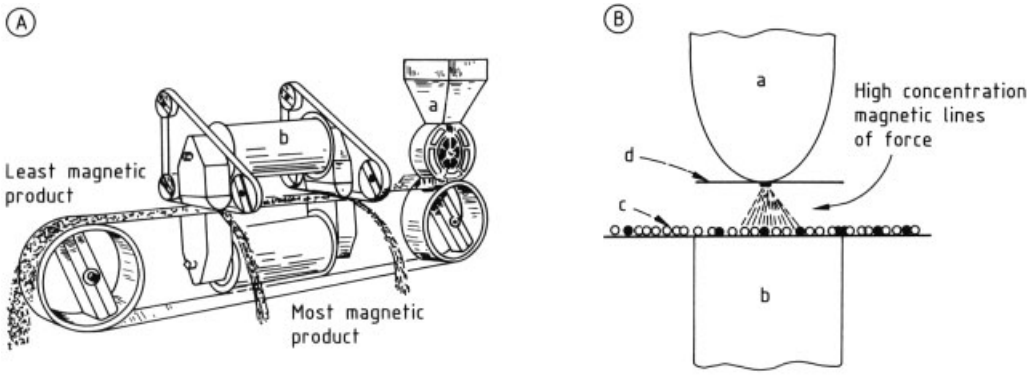

Figure 13. High-intensity crossbelt magnetic separator
A) Side view: a) Feeding hopper; b) Coil
B) Typical cross section and crossbelt arrangement: a) Parabolic curved nose of upper pole magnet; b) Flat lower pole; c) Feed belt; d) Crossbelt

used to very selectively concentrate such weakly paramagnetic materials as columbite, tantalite, wolframite, monazite, uexenite, and other high-value minerals. The use of a magnetized ring permits narrower separating air gaps and develops magnetic fields with a high flux density at each air gap.

4. Applications

The many types of magnetic separators result in a wide range of applications for these units. They are used in tramp iron removal, mineral processing, and iron recovery:

Removal of tramp iron:

Chemicals	Pharmaceuticals
Cooling fluids	Scrap metals
Food processing	Miscellaneous materials
Minerals	(water, glass, cork, textiles)

Mineral beneficiation:

Aluminum	Molybdenum
Barium	Nickel
Chromium	Niobium
Clay	Rare earths
Diamond	Rhenium
Garnet	Talc
Germanium	Tantalum
Hafnium	Tin
Iron	Titanium
Kyanite	Tungsten
Manganese	Yttrium

Recovery of iron:

Solid waste
Heavy media (ferrosilicon)

The newer, high-intensity, high-gradient magnetic separators open up other areas of magnetic treatment that have been pursued in varying degrees to date. Some of these potential applications are indicated in the following material:
Mineral beneficiation in addition to those systems listed above:

Antimony	Magnesium
Arsenic	Mercury
Asbestos	Mica
Beryllium	Platinum-group metals
Bismuth	Radium
Cadmium	Scandium
Coal	Silicon
Cobalt	Silver
Copper	Sulfur
Feldspar	Tellurium
Gallium	Thallium
Gold	Thorium
Graphite	Uranium
Indium	Vanadium
Lead	Zinc
Lithium	Zirconium

Water treatment (magnetic filtration):
- Diamagnetic suspended solids (by magnetic seeding and flocculation)
- Dissolved solids (by magnetic seeding and flocculation)
- Paramagnetic suspended solids
- Oils

Waste treatment for recovery of ferromagnetic materials:
- Coal and oil ash
- Ore tailings
- Smelter and furnace dust

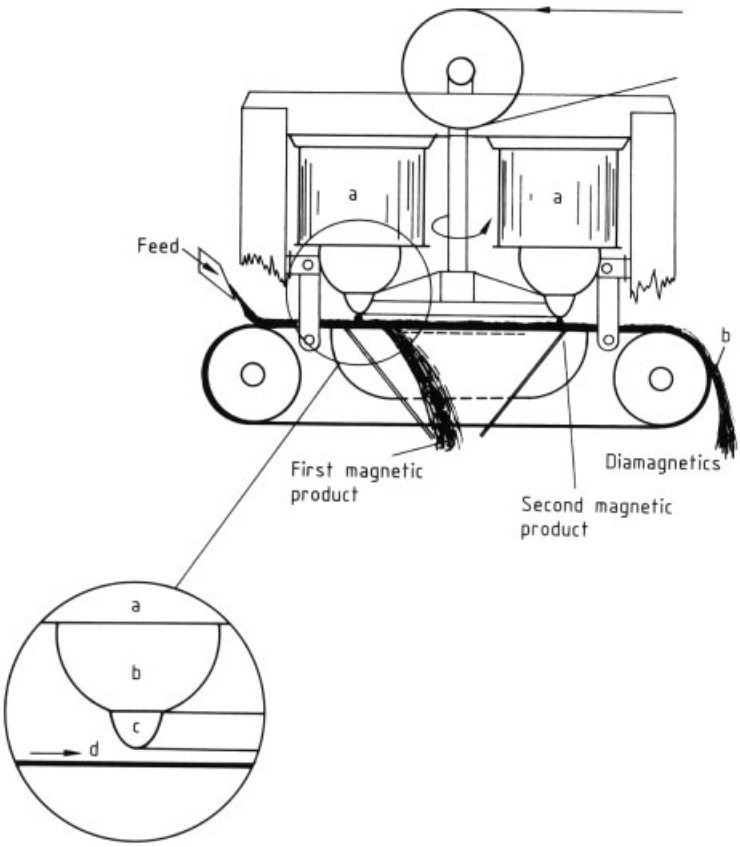

Figure 14. High-intensity disk-type magnetic separator
a) Disk-shaped induced magnet; b) Feed belt Enlargement: a) Coil; b) Pole; c) Ring; d) Belt

Removal of paramagnetic particulate impurities:
- Chemicals
- Fluids
- Minerals
- Pharmaceuticals

Chemical processing:
- Recovery of paramagnetic fine precipitates: deposition and recovery of substances on ferromagnetic particles.

The selection of a particular magnetic separator will be influenced by several factors such as

1) size of feed material,
2) tonnage or capacity to be handled,
3) relative magnetic responsiveness (i.e., susceptibility) of the material to be separated magnetically,
4) condition of feed material (wet, dry, or liquid slurry),
5) purity required in the ferro- or paramagnetic concentrate or the diamagnetic product,
6) temperature of the feed material and point of application,
7) material handling system involved,
8) operating cost parameters, and
9) level of magnetic strength required.

5. References

1. J. D. Kraus: *Electromagnetics,* McGraw-Hill, New York 1953.
2. D. M. Hopstock: "Fundamental Aspects of Design and Performance of Low Density Dry Magnetic Separators", *Trans Soc. Min. Eng. AIME* **258** (1975) 222 – 227.

3. *Permanent Magnet Handbook,* Crucible Steel Company of America.
4. R. L. Sanford: *Permanent Magnets,* National Bureau of Standards C448, 1944.
5. J. E. Lawver, D. M. Hopstock: "Wet Magnetic Separation of Weakly Magnetic Minerals", *Miner. Sci. Eng.* **6** (1974) no. 3, 154 – 172.
6. R. S. Dean, C. W. Davis: *Magnetic Separation of Ores,* Bulletin 425, US Bureau of Mines, 1942.
7. A. Nussbaum: *Electronic and Magnetic Behavior of Materials,* Prentice-Hall, Englewood Cliffs, N.J., 1967.
8. V. G. Derkatsch: *Die magnetische Aufbereitung schwachmagnetischer Erze,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1960.
9. J. A. Oberteuffer: "Magnetic Separation: A Review of Principles, Devices, and Applications", *IEEE Trans. Magn.* **MAG 10** (1974) no. 2, 223 – 238.
10. P. W. Selwood: *Magnetochemistry,* 2nd ed., Interscience, New York 1956.
11. J. E. Lawver, D. M. Hopstock: "Electrostatic and Magnetic Separation", *SME Mineral Processing Handbook,* **Sect. 6,** Society of Mining Engineers, New York 1985.
12. I. S. Wells: "Wet Separation of Paramagnetic Minerals", *Chem. Eng. (Rugby, England),* 1982.
13. G. Clark (ed.): "Magnetic Separation", *Ind. Miner.,* 1985.
14. J. E. Forciea, L. G. Hendrickson, O. E. Palasvirta: "Magnetic Separation for Mesabi Magnetite, Taconite", *Min. Eng.* (*Littleton, Colo.)* **10** (1958) 339 – 345.
15. J. E. Forciea, R. W. Salmi: "Primary Magnetic Separation Specifications", *Trans. Soc. Min. Eng. AIME* **232** (1965) 339 – 345.
16. B. Skold: "Progress in Magnetic Separation", *Mineral Processing Meeting of Swedish Mining Society,* Skellerftea, January 1970.
17. G. H. Jones, *Proc. Int. Miner. Process. Congr.,* Institute of Mining and Metallurgy, 1960, pp. 717 – 732.
18. J. Iannicelli, *Clays Clay Miner.* **24** (1976) 64 – 68.
19. E. J. Tenpas: "Magnetic Separators–Types and Applications", *Rock Prod. ,* 1971.
20. E. Laurilla: *An Approach to the Theoretical Treatment of Magnetic Concentration,* **ser. A, vol. 6,** Physica-Helsinki 1958.
21. H. W. Buus: "How to Select Magnetic Separation Equipment", *Foundry,* December 1960.
22. R. K. Singhal: "Magnetic Separators for Mineral Processing", *Min. Mag.,* **113** (1965) no. 5, 356 – 365.
23. E. J. Roberts, P. Stavenger, J. P. Bowersox, A. K. Walton, M. Melita: "Solids Concentration", *Chem. Eng.* (*N.Y),* June 29, 1970, pp. 52 – 68.
24. W. J. Bronkala: "Purification: Do It With Magnets", *Chem. Eng. (N. Y.)* **95** (1988) no. 3, 133 – 138.
25. P. B. Sherwood: "Cross Belt and In-Line Primary Magnets", *N.S.A. Operators Meeting,* Salt Lake City, Utah, February 4 – 6, 1976.
26. R. B. Jacob, J. A. Selvaggi: "Operating Characteristics of the Newly Designed High Gradient Magnetic Separator (HGMS)", *AIME Annual Meeting,* Atlanta, Ga, March 6 – 10, 1983.
27. H. Harley-Smith: "Magnetic Separation in Mineral Beneficiation", Metals and Minerals International. *Perry's Chemical Handbook*: "Solid – Solid Systems and Liquid – Liquid Systems", 6th ed., Chapter 21, McGraw Hill, New York 1984.
28. M. Gupta: "A New Method of Magnetic Separation of Ferro and Para-Magnetic Minerals from Feebly Magnetic and Non-Magnetic Minerals", *Proc Australas. Inst Min Metall,* No. 274, June 1980.
29. G. Clary, (ed.): "Magnetic Separation – 4", *Industrial Minerals,* May 1985.
30. L. A. Roc: "Magnetic Separation of Ores – History and New Foreign Developments", *Annual AIME Meeting,* New York, February 16 – 20, 1958.
31. R. F. Merwin: "How to Remove Tramp Iron from Burdens Carried on Conveyor Belt", *Annual AIME Meeting,* Tampa, Florida, October 13 – 15, 1966.
32. C. K. McArthur: P. R. Porath, "Concentrating Iron Ore in a Mobile Plant", *Min. Congr. J.* **33** (1965) October 28.
33. J. A. Oberteuffer: *Magnetic Separation: A Review of Principles, Devices and Applications,* Institute of Electrical and Electronics Engineers Inc., Copyright, 1977.

Magnetic Tape → Information Storage Materials
Magnets → Magnetic Materials

Electrostatic Separation

LUDWIG BRANDS, Kali und Salz GmbH, Kassel, Federal Republic of Germany

PETER-M. BEIER, Kali und Salz GmbH, Kassel, Federal Republic of Germany

INGO STAHL, Kali und Salz GmbH, Kassel, Federal Republic of Germany

1. Introduction 1
2. General Principles 1
2.1. Charging by Induction or Corona Discharge 3
2.1.1. Charging by Induction 3
2.1.2. Charging by Corona Discharge 3
2.1.3. Separators 4
2.2. Triboelectric Charging 6
2.2.1. Mechanism 6
2.2.2. Equipment 7
2.3. Dielectrophoresis 8
3. Applications 9
3.1. Mineral Processing 9
3.1.1. Potash Ores 9
3.1.2. Other Minerals 12
3.2. Purification of Coal 13
3.3. Recycling Processes 14
3.3.1. Inductive or Corona Charging 14
3.3.2. Triboelectrification 15
3.4. Others 15
4. References 17

1. Introduction

Electrostatic separation is the common term for all processing technologies utilizing electric forces acting on charged or polarized particles in a static electric field. The behavior of the particles depends on their electrical and geometrical properties. Therefore, a separation of different species is possible if the parameters of the electric field and the ambient conditions are set up in the right way.

Some of the basic principles of electrostatic separation have been well known since ancient times. THALES OF MILETUS (600 B.C.) observed that an amber rod attracts dust particles after being rubbed with a cloth. Electrostatic effects have first been utilized around 1880 for agricultural purposes in the separation of chaff from grain. Further applications became possible after the necessary electrical equipment for high-voltage charging of particles had been available. Nowadays numerous industrial installations, processing millions of tons of salts, iron ores, titanium ores, beach sands, and other minerals, are on stream, demonstrating the potential of this technology.

2. General Principles

An electrostatic separation process consists of the two steps: charging and separation. Charging is the key factor in all electrostatic separations due to its paramount importance for the process. Charging can be effected by either induction, corona discharging, triboelectrification, or polarization. The first three charging mechanisms involve charge transfer to or from a particle and every separation process on this basis is called *electrophoresis*. The fourth mechanism is not accompanied by a charge transfer, the corresponding separation process is called *dielectrophoresis*. In dielectrophoresis the particles remain neutral in contrast to electrophoresis where ionic particles are the species to be separated. *Electrophoresis* and *dielectrophoresis* can take place in fluids, liquids, and gases. In the narrower sense the term *electrophoresis*, however, is used for separation of organic molecules in liquids, especially for the separation of proteins and nucleotides (→ Electrophoresis). To avoid a misunderstanding this term will not be used further in this article. Charging of a particle will occur only on its surface, not within the entire particle. Therefore the factors rele-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_20

vant to charging are surface specific. The most important physical property is the surface conductivity of a given particle. On the one hand a high surface conductivity will facilitate charging of a particle. On the other hand discharging of a previously charged particle will occur much easier when its conductivity increases. Whether the conductivity of a given material can be utilized for an electrostatic separation or not has to be considered from case to case. In addition the surface work function of a given material is important for charging, especially for triboelectric charging. The polarity and maximum potential difference between particles of two materials to be separated are to be determined by the difference in their work functions. Another property of relevance is the dielectric constant of the particles, especially for dielectrophoresis. The higher the dielectric constant of a particle, the easier a separation will be. A high difference in the characteristics of the materials to be separated thus favors their separation.

The conductivity, the work function, and the dielectric constant of a given solid surface are not unique properties characteristic for a given material: They highly depend on ambient conditions such as temperature, moisture, and the presence of chemicals, that influence the structure and chemical composition of the particle surface. Attached to a particle surface is at least one layer of adsorbed molecules, water in most cases. Adsorbed water will lead to a rapid discharge of previously charged particles. Therefore in most cases heating in advance of an electrostatic separation step is necessary to remove water. At 100 to 150 °C only physically adsorbed water will be removed. But heating to higher temperatures usually is not economically feasible and in addition will impede sufficient charging. The surface properties often can be affected by treatment with chemicals, such as carboxylic or hydrofluoric acids. This will generate new electronic states on surface and increases or decreases the ability of a given material to become charged. If a carefully selected chemical is employed and the charging is carried out under appropriate conditions with respect to humidity and temperature, a selective charging can be achieved even in difficult cases (see Section 3.1.1).

Success of an electrostatic separation strongly depends on the particle size of the materials involved. To achieve an efficient separation by electrostatics, the upper limit for the particle size is 2 – 8 mm, depending upon the particle's density. Larger particles will be less affected by the electric field due to their higher weight and thus the higher gravitational force acting on them. When material with too broad a particle size distribution is fed, classification will interfere with the electrostatic effect, leading to a poor separation. Particles below 40 µm will adhere to the larger particles or to the electrodes and therefore cannot be processed by standard separators. For this kind of particles special separator types have to be employed (see Sections 2.2 and 3.4). An appropriate screening of the materials has thus to be carried out before feeding them to an electrostatic separator.

The separation itself can be effected either in a uniform or a nonuniform electric field as shown in Figure 1. Each of the field types finds special technical applications. Uniform electric fields are only employed as d.c. fields whereas nonuniform fields are also used as a.c. fields, the latter being the case in dielectrophoresis (Fig 1 B, see also Section 2.3).

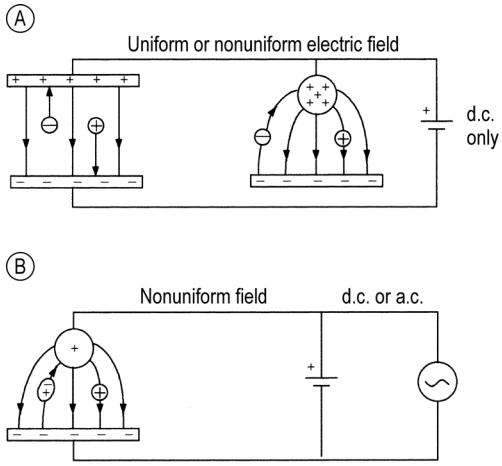

Figure 1. Principles of electrophoresis and dielectrophoresis
A) Electrophoretic force on charged particles; B) Dielectrophoretic force on neutral particles

There are numerous publications describing the general principles of electrostatics and its application for separation purposes in detail. Examples are given in [1,2,3,4,5,6,7,8].

2.1. Charging by Induction or Corona Discharge

Since these two charging mechanisms reveal great similarities with respect to the physical fundamentals, the apparatus, and the applications, they will be described in the same section.

2.1.1. Charging by Induction

A schematic image of an apparatus usually employed in electrostatic separation by inductive charging is shown in Figure 2. The mixture to be separated is fed onto a grounded rotational drum and passes through the electric field generated by a high voltage tubular electrode. In most cases the electrode is negatively charged, but a positive charging is in principle possible as well. The conductive particles get charged by induction for a short period of time, depending on their surface conductivity. The induced charge has the inverse polarity to that of the high voltage electrode. Therefore an attractive force towards the electrode will be imposed on the conductive particles. These particles are deflected from the surface of the rotor in the direction of the electrode and can subsequently be collected in an appropriately positioned bin. The nonconductive particles are not significantly charged during their residence on the rotor or their charge will be too small to affect a motion towards the high voltage electrode. Thus they remain on the rotor until they will fall into a collecting bin under the influence of gravitational force or will be scraped down by a brush.

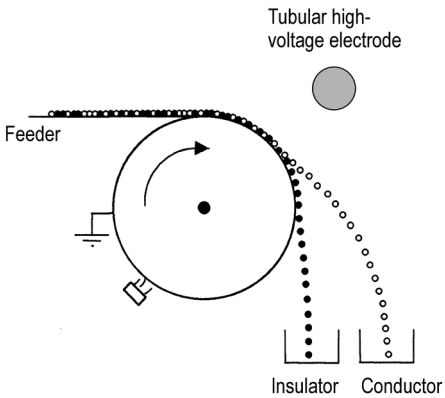

Figure 2. Rotational drum separator – inductive charging

Devices for electrostatic separation via induction – so-called electrostatic drum separators – can only be applied for separation of conductive particles from nonconductive ones. Although separation of particles having large differences in surface conductivities would be possible in principle, in practice this has never been done with success on a large scale. Electrostatic drum separators are especially employed for final cleaning of industrial products such as rutile, zircon, or food, e.g., tea leaves (see Section 3.4).

2.1.2. Charging by Corona Discharge

This kind of electrostatic separation works similar to the induction based process and is described schematically in Figure 3. In contrast to inductive charging all particles – conductive and nonconductive ones – are charged due to the high field strength produced by the ionizing corona electrode generating an ion stream. The corona electrode usually consists of a thin wire or a needle type electrode. The charge of the conductive particles and semiconductors rapidly migrates to the grounded rotor whereas the nonconductive ones remain charged for a longer time due to their higher surface resistivity. They are attracted to the rotor and follow its movement. Finally they are scraped down by a brush and collected in a bin. The neutralized conductive particles fall down into their corresponding bin.

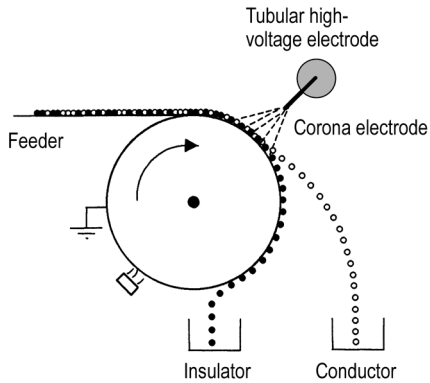

Figure 3. Rotational drum separator – corona charging

Electrostatic separation via corona charging – so-called corona separation – can be employed

for the separation of particles with different conductivities. In contrast to electrostatic drum separators corona separators provide a wider range of applicability due to the stronger field employed.

2.1.3. Separators

Drum separators are the most widely used separation devices. Examples for the use of this technology are the beneficiation of mineral sands [9, 10] and iron ores [11, 12], the production of iron superconcentrates [13, 14] and the processing of waste electric wire [8, 15, 16] (see also Chap. 3).

Extensive work has been done to understand the theoretical fundamentals leading to a better modeling of the electrostatic separation process. For designing an industrial installation, however, thorough experimental work on both laboratory and pilot plant scale has to be carried out. The key factors affecting the performance of a rotational drum separator, that have to be investigated for every electrostatic process plant, are

1) Conductivity of the particles
2) Dielectric constant of the particles
3) Density of the particles
4) Particle size
5) Particle shape
6) Rotary speed of the drum
7) Structure of the electric field.

Many of these topics have been discussed above. The dependence on the particle shape is shown in the shape separation process designed by Knoll et al. [17]. The specifically constructed separator and the controlled electric field led to an improvement of the recovery of coarse vermiculite from siliceous gangue. The flatter vermiculite is more efficiently attracted to the rotor than the gangue. The latter behaves like a conductor and can be separated from the vermiculite.

For electrostatic drum separators usually metallic tube electrodes are used. For enhancement of separation efficiency the applied voltage has to be increased. In order to prevent an electrical breakdown when the electrode voltage is increased, nonconductive coatings have been introduced. Vacuum discharge tubes made of glass are also widely used now [18].

The needle or wire electrodes of corona separators are often additionally equipped with tubular high voltage electrodes. This will enhance separation of the nonconductive or less conductive particles when they are deflected towards the electrode. For the arrangement of the corona electrode and the high voltage electrode in general two cases are possible as shown in Figure 4 [19]. The "pinning" geometry has become more prevalent especially in the mineral sands industry. A further improvement of separation can be carried out by combining the corona electrode with more than one high voltage electrode. As shown in Figure 5, depicting the separation of copper from PVC in a laboratory study, the copper content of the conductor fraction can be increased up to 99.6 % by the action of three additional high voltage electrodes. In industrial applications one electrode with elliptical cross section replaces the three tube electrodes, resulting in the same efficiency [8].

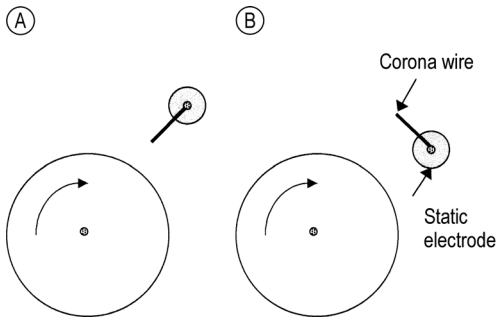

Figure 4. Combination of corona electrodes and static high voltage electrodes
A) "Pinning" geometry; B) "Lifting" geometry

Some new electrostatic separators have also been developed to improve the performance of drum separators. Some principle examples are shown in Figures 6, 7, 8. The combination of a high voltage electrode or corona electrode and a grounded counter electrode or moving device is a common feature of all these special machines. The structure of the specific separator depends on the characteristics of the materials, conductivity, particle size, amount of feeding, etc. Besides multiple drum machines the most important devices are belt-type and plate-type separators. For example, the separation of carrot and celery seed lots for selection of those with the best germination has been carried out on a

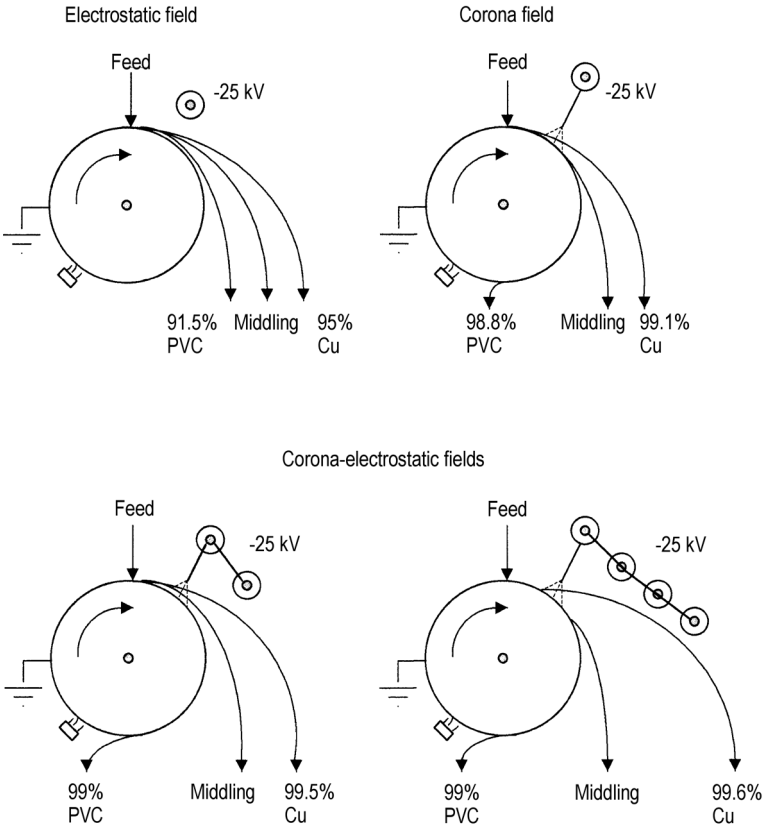

Figure 5. Influence of the electric field on electrostatic separation of metals and insulators [14]

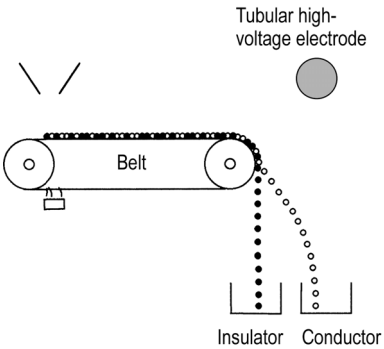

Figure 6. Moving belt separator – inductive charging

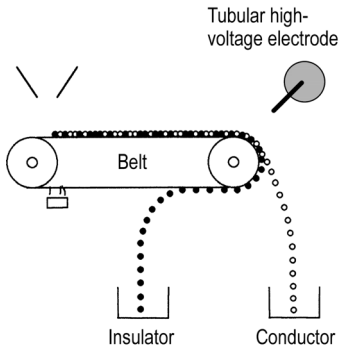

Figure 7. Moving belt separator – corona charging

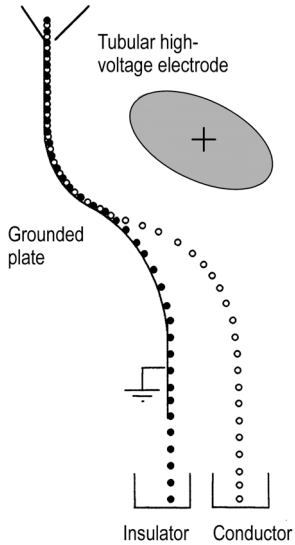

Figure 8. Plate-type separator – inductive charging

belt-type separator [20]. The *plate-type separator* has the advantage of absence of moving parts thus decreasing maintenance expense and stand-stills. It was developed primary for separation of low amounts of conductive minerals from zircon concentrates [3].

2.2. Triboelectric Charging

2.2.1. Mechanism

The well known phenomenon that two different materials will acquire different charges when rubbed against each other is the physical basis for this method of charging. Charging of particles can occur either on contact with particles of the other component in the mixture or on contact with the material of the charging device as demonstrated in Figure 9. When a mixture of the charged material is allowed to free-fall in an electric field formed by two oppositely charged electrodes, the materials will be separated as shown in Figure 10. The polarity of the charge of a specific material received upon rubbing depends on the particular material it is to be separated from, more precisely on the difference in the electron work functions of the materials

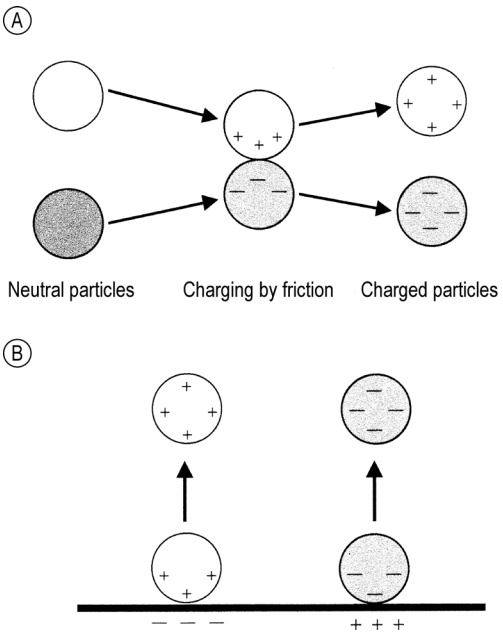

Figure 9. Triboelectric charging of particles
A) Particle – particle contact; B) Particle – material contact

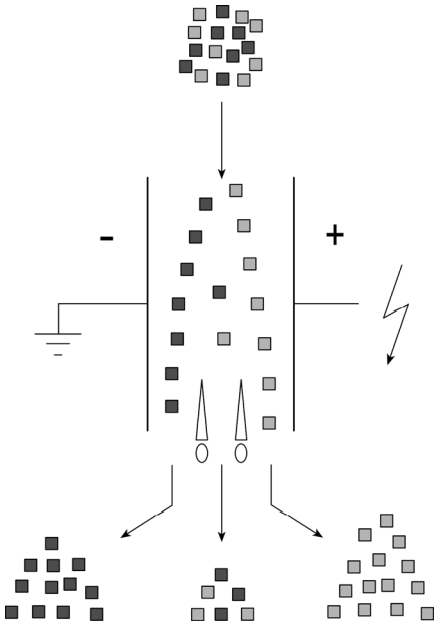

Figure 10. Separation in the electrostatic field

to be separated. All materials can be arranged in a series according to the polarity of charge acquired by that rubbing procedure, the *triboelectric series*. This matter has been the object of numerous investigations, because there is no unique series in which a specific material occupies a given position like, e.g., every chemical element in the periodic table. Depending upon the experimental conditions several triboelectric series exist [21]. The fundamentals of triboelectrification have been the object of many investigations and great advances have been made in the last years [7, 22, 23, 24, 25, 26, 27]. Nevertheless, a complete understanding has not yet been achieved. Therefore equipment design needs extensive laboratory tests and pilot plant studies as in the case of drum separators.

The effectivity of charging depends not only on the chemical composition of the materials to be separated but also markedly on the humidity of the ambient atmosphere, the particle size distribution, and the history of the particle surfaces. This can be understood by consideration of the fact that any solid surface is covered by a layer of adsorbed molecules, originating from the ambient atmosphere. Often a modification of fresh generated surfaces by adsorption of specific chemicals is necessary to obtain a sufficient charging upon friction.

2.2.2. Equipment

Electrostatic separation by triboelectrification can be employed for the separation of both coarse particles and fine powders. For many years industrial installations were restricted to the processing of potash ores only. Since 1995 some machines have been on stream for separation of fly ash (i.e., dust particles) in the USA. Also erection of some pilot plants for the separation of mixed plastics and purification of coal has been reported, but no industrial installation is known up to now.

In contrast to rotational drum separators, in triboelectric charging the charging and the subsequent separation are more distinct operations. Charging can be carried out, e.g., in a fluidized bed, in cyclones or tubes with helix-type internals or without any inner constructions. In potash processing (see Chap. 3) only fluidized beds are employed for charging. Plastics recovery pilot plants also use this technique in the most cases. Charging for coal purification often has been carried out with the other devices.

Also in the separation step many different devices are employed. In almost all installations, especially in the potash industry, so-called free-fall separators are utilized. In such an apparatus the charged particles fall "freely" due to gravitational forces in the electric field formed by appropriate electrodes. Plate electrodes are the most simple construction for this purpose. The field is formed by two metal plates arranged in parallel. The main problem of this kind of electrode was the dust adhering to the plates that decreased the electric field strength. Attempts to overcome this shortcoming by application of vibrational devices failed, so that further developments in separator design were necessary. A step forward was the construction of vertically or horizontally working belt separators. The moving belt consisted of rubber with a conductive coating. With this type of separator a throughput up to 20 t per hour and per meter was realizable but other shortcomings prevented the employment in process plants.

The tubular free-fall separator (Fig. 11), that was subsequently developed, fulfilled the requirements for an appropriate apparatus for potash processing. The electrodes remain dust-free on their entire height and the deflected salt between the electrodes can leave the apparatus on quite a short way. This separator was developed by the German potash industry on the basis of an American separator type with motionless tubular electrodes [28]. The electrodes developed in Germany consist of two rows of rotating steel tubes facing each other. The rows are permanently kept dust-free with brushes on the sides remote from the electric field. The tubes usually have a length of ca. 2 m. The longest separators can be more than 10 m in length. The throughput is about 30 t of salt per hour and per meter. Electrode distance is 250 mm, the field strength can reach values from 4 to 5 kV/cm. The voltage is 100 – 125 kV while the current is less than 1 mA. Consumption of current for the separation itself is therefore extremely low, only caused by brush-discharge loss and leak currents. Provided that the raw salt particles have been sufficiently liberated (→ Potassium Compounds, Chap. 5.4.), particle sizes up to 1.5 – 2 mm can be processed [29]. In some cases the extent of liberation of the minerals makes a separation of particles up to 4 mm possible. However, this requires a greater length of the separator due to the higher gravity force acting on the particles.

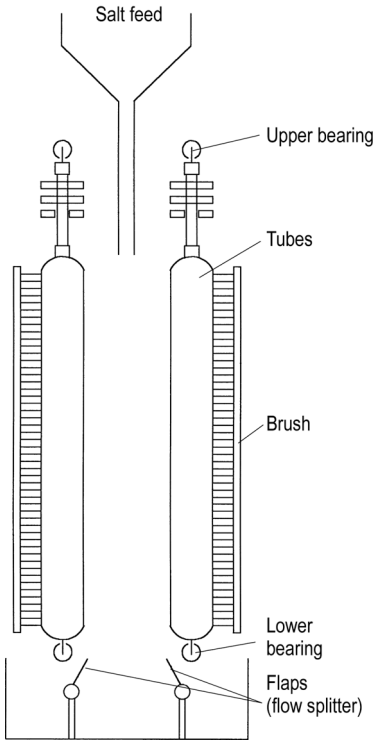

Figure 11. Tubular free-fall separator

Other types of separators have also been developed for special purposes. The electrodes can also be arranged in a nonparallel manner, e.g., in a plate-type separator utilized for separation of plastics on a laboratory scale [30, 31, 32]. Perforated electrodes incorporated in the fluidized bed utilized for charging have been constructed for purification of fine coal [33]. The coal particles are separated from the impurities by passing the electrode whereas the impurities are rejected. A sieve conveyor is another example of separation machines specially designed for this purpose (see Section 3.2) [34]. With the latter device purification of coal particles with diameters $< 40\,\mu m$ has been reported to be possible.

Most of the separators are only employed on a laboratory or pilot-plant scale. An exception is the new belt separator shown in Figure 12 [35]. This device is constructed for separation of finest particles after contact charging. The separator consists of two plate electrodes with a distance of up to 10 mm. A mesh belt of plastics between the electrodes is utilized as transport medium for the particles. The material is charged while it is fed into the separator by an air slide. The charged particles are deflected towards the oppositely charged electrodes and are conveyed by the mesh belt to the respective end of the separator, where a sampling will be carried out. The separator has been employed with success in fly ash processing in the United States (see Section 3.4).

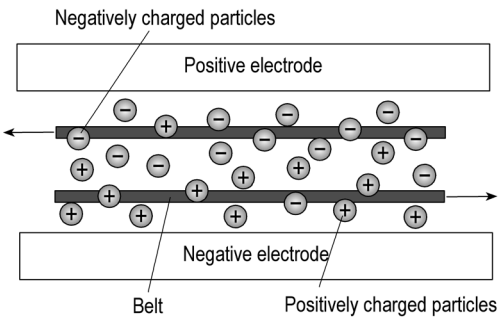

Figure 12. Belt separator for dust processing

2.3. Dielectrophoresis

Dielectrophoresis is the movement of neutral particles in a nonuniform electric field (Fig. 1B). It differs from the conventional electrostatic separation (electrophoresis) in the following features:

1) Dielectrophoresis produces motion of particles suspended in a fluid medium. The direction of motion is independent of the sign of the field. Accordingly, either a.c. or d.c. fields can be used. Electrophoresis on the other hand produces a motion of the suspended particles whereby the direction of the resultant path depends on the sign of the charge and the sign of the field direction.
2) Dielectrophoresis gives rise to an effect that is proportional to the particle volume and is, therefore, more easily observed on coarse particles. Contrary to that the effect of electrophoresis is observable with particles of any size.
3) Dielectrophoresis usually requires quite divergent fields for strong effect. Electrophoresis operates in both uniform and divergent fields.
4) A relatively high field strength is required in dielectrophoresis. In media of low dielectric constant, this usually amounts to some 10^4 V/m. In media of high dielectric constant, e.g., water, lower fields may be used. In contrast, electrophoresis with previously charged particles requires relatively low fields.
5) In dieletrophoresis usually a substantial difference is required between the relative dielectric constants of the particles and the surrounding medium. Electrophoresis can be appreciable even when the free charge per weight unit of the particle is quite small.

The properties of a dielectrophoretic separation vary with the medium in which separation occurs. In a dielectric field medium having an intermediate dielectric constant with respect to the feed constituents, a separation can be performed based on differences in the dielectric constants of the feed materials. Examples of intermediate dielectric fluids include nitrobenzene, kerosene, xylene, ethanol, and propanol. When air is used as a medium, separation is largely based on differences in shape, density, and dipole moment.

Although dielectrophoretic separation has been well known since the end of the 19th century [36], applications of this method are still rare (see Section 3.4). Widespread use of di-

electrophoresis has been hampered by, e.g., the environmental and health hazard of the utilized dielectric field media, difficulties in the employment of high-frequency a.c. fields in technical applied apparatus volumes, and, for a long time, the lack of understanding the mechanistic details of this technology. A review of the state-of-the-art including history is given in [37]. Dielectrophoresis is an emerging technology that will certainly provide new solutions in the area of physical separation processes.

3. Applications

Separation processes based on *inductive* or *corona charging* are the method of choice in cases of high differences in conductivity of the materials to be separated (see also Chap. 1). This encompasses processing of conductive minerals and recovery of values from electronic scrap and metal-containing manufacturing wastes. Separation of insulators such as potash and plastics however are typical cases where *tribolectrification* is employed. The applications of electrostatic separation known so far are listed in Tables 1, 2, 3. A summary of some developments is given in [38].

Table 1. Applications of inductive or corona charging

Equipment	Materials	References
Rotational drum (inductive/corona charging)	mineral sand	[9]
	boron minerals	[9]
	rutile – zircon – quartz	[11,39,40]
	iron ore, silica	[11,12,13,14]
	glass sands	[40]
	kassiterite, scheelite	[42]
	tantalite, columbite	[42]
	feldspar – quartz	[43,44]
	fly ash, carbon	[45]
	lunar soil	[46]
	aluminum dross	[47]
	Cu glass fiber, resin	[48,49]
	rare metals	[50]
	PET aluminum	[51]
	Cu PVC	[8,15,16,52,53,54]
	tea leaves – stalks	[55]
Slide plate (inductive charging)	mineral sand	[10]
	rutile – zircon	[51]
	lunar soil	[46]
Moving belt (inductive/corona charging)	onion seeds	[55]
	flowers	
	carrot – celery seeds	[20]

Table 2. Application of triboelectricity (free-fall separator)

Charging equipment	Material	References
Cyclone	coal	[56,57]
Rotating cone	coal	[58]
Tube	coal	[59,60]
Fluidized bed	potash	[28,29,61,62,63,64,65]
	coal	[66]
	plastics	[67,68,69,70]
Electrostatic sieve	coal	[34]
Crusher	coal	[71]
Rotating pipe	plastics	[30,31]
Vibrating feeder	plastics	[72]

Table 3. Applications of other methods

Method	Material	References
Electrofiltration	kaolin slurries	[73,74,75]
	oil cleaning	[76]
	water cleaning	[77]
Dielectrophoresis	food cleaning	
	oil purification	[78,79]
	biological materials	[80,81,82]
	diamonds/kimberlite rock	[37]

3.1. Mineral Processing

3.1.1. Potash Ores

In Germany electrostatic separation by tribolectrification is employed in the processing of potash ores since 1973 on an industrial scale. At present more than 15×10^6 t/a of potash ore are processed in this way.

Investigations on this separation technology began at the Minerals and Chemical Corporation (IMCC or IMC), Chicago [83,84]. Potassium chloride and sodium chloride were successfully charged by heating up to 300 – 700 °C. However, the energy demand of this method was quite high and separation results were not satisfactory. The employment of surfactants for the pretreatment of the feed made a temperature decrease to below 100 °C possible, which made the process economically feasible [65,85,86,87,88,89,90,91,92,93,94,95,96,97,98].

In Figure 13 (see next page) the amount of sylvine (KCl), kieserite ($MgSO_4 \cdot H_2O$), and halite (rock salt, NaCl) separated at the positive electrode is shown, depending on the relative humidity of the heating air. The raw salt is conditioned with aromatic monocarboxylic acids (Fig. 13A), aliphatic monocarboxylic

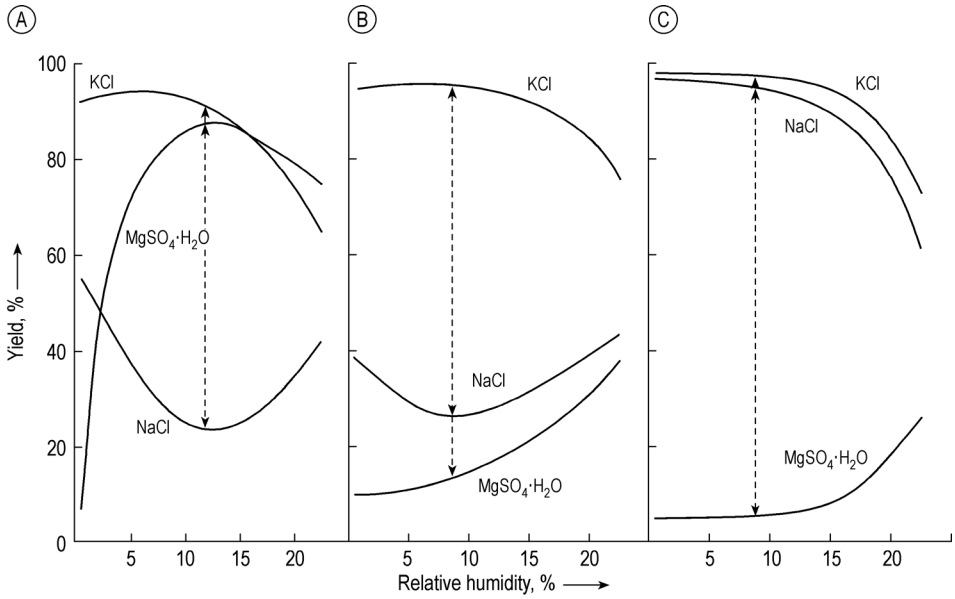

Figure 13. Fraction at the positive electrode
Conditioning with aromatic monocarboxylic acids (A); With aliphatic monocarboxylic acids (B); With aliphatic monocarboxylic acids and ammonium salts of aliphatic monocarboxylic acids (C)

acids (Fig. 13B), and aliphatic monocarboxylic acids in combination with the corresponding ammonium salts (Fig. 13C).

In principle any of the three aforementioned components can be separated at the positive electrode just by changing the relative humidity of the air. A better performance, however, is achieved by appropriate combination of the relative humidity and the conditioning agent. For an improvement of charging often a second agent in addition to the main conditioning agent is employed (Fig. 14).

As shown on the left-hand side of Figure 14 (conditioning with salicylic acid), a loss of $MgSO_4 \cdot H_2O$ has to be accepted to achieve a maximum yield of KCl and NaCl. Only by addition of a second agent the optimum yield of all three components is obtained. Meanwhile numerous reagents acting as conditioning agents are known. This will allow separation even of complex mineral mixtures into their components

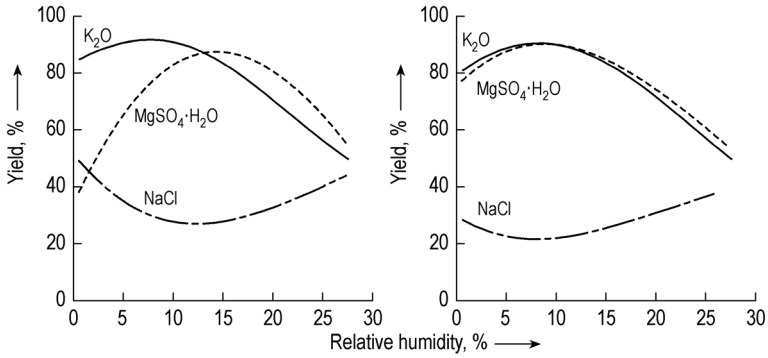

Figure 14. Yield of K_2O, kieserite, and rock salt at the positive electrode in dependence on relative humidity, conditioning with salicylic acid (left) and salicyclic acid plus chloroacetic or lactic acid (right)

[93,94]. According to the choice of conditioning agents and humidity of the air a selective charging of the different minerals against each other is achievable. A mixture consisting of more than two components therefore can simply be separated in several subsequent process stages, e.g., as follows:

1) Separation of sylvine in the first step [84, 86,95]. Then the residue is separated into kieserite and halite.
2) Kieserite is separated first and the potash component in the following step [97,98].
3) In the first step halite is separated [88,89,90] and the residue is separated subsequently into potash and kieserite concentrate.

All three methods are performed on an industrial scale. Method (3) is employed in the largest electrostatic potash processing units at the Werra works [61,62] of the German potash industry. A principal flow sheet of an electrostatic separation circuit is shown in Figure 15. The raw salt delivered from the mine is size-reduced by impact grinding and subsequently screened. The particle size of the feed has to be adjusted carefully. To achieve sufficient liberation of the individual components of the raw salt (i.e., halite, sylvite, kieserite) the upper limit should be $< 1.5 - 2$ mm. Grains > 2 mm cannot be processed due to insufficient liberation. Additionally, the gravitational force acting on them would require greater working lengths of the separators. Fine particles (< 0.1 mm) adhere to the larger particles and thus cannot be separated (see Chap. 2).

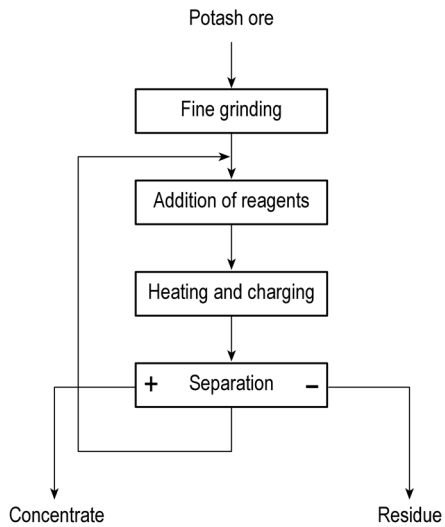

Figure 15. Electrostatic separation circuit

Conditioning is one of the key factors in electrostatic potash processing and consists of the following steps:

1) Application of the agent or agent combination onto the raw salt
2) Adjustment of the appropriate temperature and relative humidity of the air
3) Charging of the salt by friction in air with defined relative humidity.

The application of the agent – the so-called chemical conditioning – is carried out either in the solid, liquid, or gaseous state [92], but also on inert support materials, which possess large inner surfaces [87]. About 20 – 100 g of conditioning agent per ton of raw salt are employed. After conditioning the potash mineral is charged in a fluidized bed in air with defined humidity and the subsequent separation will be done as described in Section 2.2.2.

The flow sheet of one of the electrostatic separation plants at the Werra works [29,61] of the German potash industry is shown as an example in Figure 16. The raw salt delivered from the mine contains up to 70 % of rock salt. Therefore in the first separation step halite is separated and dumped. The other fraction contains the potash minerals sylvine and carnallite as well as kieserite. The following process steps can be carried out in considerable smaller devices. After renewed conditioning in the next step sylvine

and carnallite are obtained as a potash raw fraction at one electrode, the kieserite raw fraction is obtained at the other. Purification of each fraction leads to a potash enriched fraction and a magnesium enriched fraction. The first one contains sylvine and carnallite in a higher purity compared to the previous fraction whereas the latter one consists mainly of kieserite. The residues of these purification steps are subjected to the hot leaching process, the oldest industrial potash purification process (→ Potassium Compounds, Chap. 5.2.2.). The potash enriched fraction undergoes a wet process yielding KCl employed in potassium sulfate manufacturing. The kieserite is of fertilizer quality and can be shipped as such. Small amounts of langbeinite ($K_2SO_4 \cdot 2\,MgSO_4$) and anhydrite ($CaSO_4$) do not affect this application. Further treatment of kieserite to obtain $MgSO_4$ by dewatering, however, requires a material of higher purity. For this purpose kieserite is submitted to a subsequent electrostatic separation stage. After another conditioning separation gives the required pure kieserite and a langbeinite- and anhydrite-rich kieserite. The latter is employed in the manufacturing of potassium sulfate together with KCl.

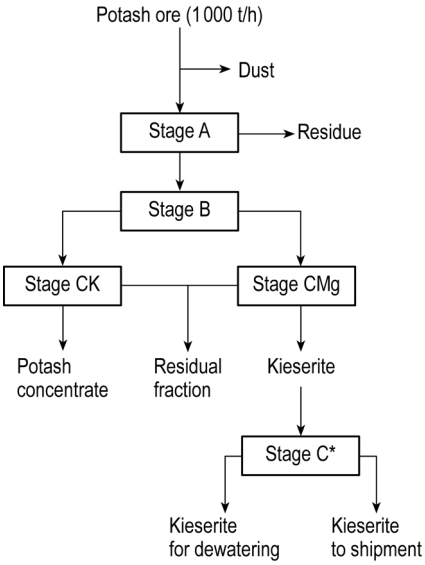

Figure 16. Flow sheet of electrostatic separation plant for potash processing

The dust (< 0.1 mm) separated from the fluidized beds is not recycled to the process. As pilot-plant tests have shown both the yield and the quality of the products will be lower when such a recycling is carried out [99, 100]. The dust will therefore be processed in the hot leaching process of the potash plant in the same way as the dust formed in the grinding step before conditioning.

3.1.2. Other Minerals

Electrostatic treatment of other minerals on an industrial scale is exclusively performed by drum separators and similar devices. Rotational drum separators were first employed for industrial applications in the 1920s. In the late 1940s this technology was first used for the extraction of titanium-bearing minerals from beach sands. The beneficiation of iron ores was first carried out in 1964 in the Wabush mines, Labrador, Canada, after extensive investigations on the laboratory scale and in a pilot plant. After successful operations of separation plants had been reported, many other applications of processes employing rotational drum separators were developed for materials such as, e.g., plastics, metals, foods, and agricultural products (see Sections 3.3, 3.4).

Iron Ore Beneficiation. In the Wabush installation the dry gravity concentrate of iron ore processing is fed into 36 rougher each having six rotational drums of 3.05 m diameter. The ore is separated in a conductor and a nonconductor fraction. The middling is retreated in 18 cleaner units. The conductor fraction leaves the circuit to form part of the final product. The nonconductor fraction is fed to the scavenger section. The conductor fraction of the cleaner units is combined with the conductor fraction of the rougher stage. The tailings of the cleaner step are retreated with the rougher tailings in the scavenger. The middling is recirculated to the rougher. The scavenger produces an iron ore fraction which is combined with the conductor fraction of the rougher and the cleaner step. The nonconductor fraction from the scavenger is completely iron-free and forms the final reject. The voltage at the electrostatic electrodes is + 25 kV. The iron product grade is reported to be 66.3 % with 98 % iron recovery [11].

Titanium Ore Separation. Rotational drum machines have been employed in processes for separation and beneficiation of titanium-bearing minerals, such as rutile, anatas, ilmenite, leucoxene, and perovskite. One of the first commercially utilized processes is the separation of rutile minerals from beach sand, of which the Sierra Leone operation will be described in the following. For an optimum separation efficiency the material circuit is divided in a fine (< 0.2 mm) and a coarse (> 0.25 mm) circuit. The heavy fraction from the wet gravity plant is fed to rougher drum separators yielding a conductor fraction and a nonconductor one. The first one is fed to the cleaner separators, the latter to scavenger separators. For the cleaner and the scavenger step also drum separators are employed. The middling of the rougher units together with the conductor fraction of the scavengers and the nonconductor fraction of the cleaner separators is added to the feed of the roughers. The conductor fraction of the cleaner step is submitted to a magnetic separation. The nonmagnetic fraction undergoes an additional electrostatic cleaning step in plate-type separators. The nonconductor fraction of the scavengers and the tailings of the magnetic separation step form the reject. The rutile concentrate leaving the final cleaning step was reported to have a TiO_2 content of 95 % (fine circuit) or 96 %, respectively (coarse circuit) with an overall rutile recovery of 88 % [11].

Separation via Triboelectrification. Besides potash processing no industrial mineral separation plant is known where triboelectrification is used. However, the potential of electrostatic separation by contact charging has been demonstrated in numerous laboratory tests and pilot plant separations. Examples are separations of feldspar – quartz, calcite – quartz, dolomite – quartz, brown coal – quartz sand. The purification of phosphate rock has also been established.

3.2. Purification of Coal

Coal contains impurities such as ash and pyrites and therefore has to be purified before combustion in order to lower the expense for flue gas desulfurization. Dry methods for purification of coal have been a focus of considerable interest for many years. Compared to well established wet cleaning processes they have the advantage of lower cost due to the fact that a drying step is unnecessary. Numerous electrostatic separators have been developed for this purpose. Among them charging via triboelectrification is most important [38]. Methods utilizing induction or corona charging have also been applied to the separation of coal from its impurities but, for the most part, these methods had only limited success [101].

One of the main characteristics of coal purification via triboelectrification is the small particle size of the coal compared to potash ore particles. In most cases particles of $< 100\,\mu m$ have to be processed, but separators exist that are able to process coal of particle size $< 40\,\mu m$. This particle size is required to ensure adequate liberation of the pyrite impurities. For particles of this size electrostatic separation in a gravitational field cannot be achieved. Instead the coal has to be conveyed by gas flow (pneumatic conveying) while passing through the electric field. The separators and installations shown in Figure 17 are examples for this technology [38]. The common feature is that finely ground coal is injected into a gas flow. The particles are charged by contacting the inner wall of the pipe. The wall consists of a material which is located in the triboelectric series between the two components (coal vs. ash/pyrite) to be separated.

Although many successful separations on the laboratory and pilot-plant scale are reported, no industrial installation is known until now.

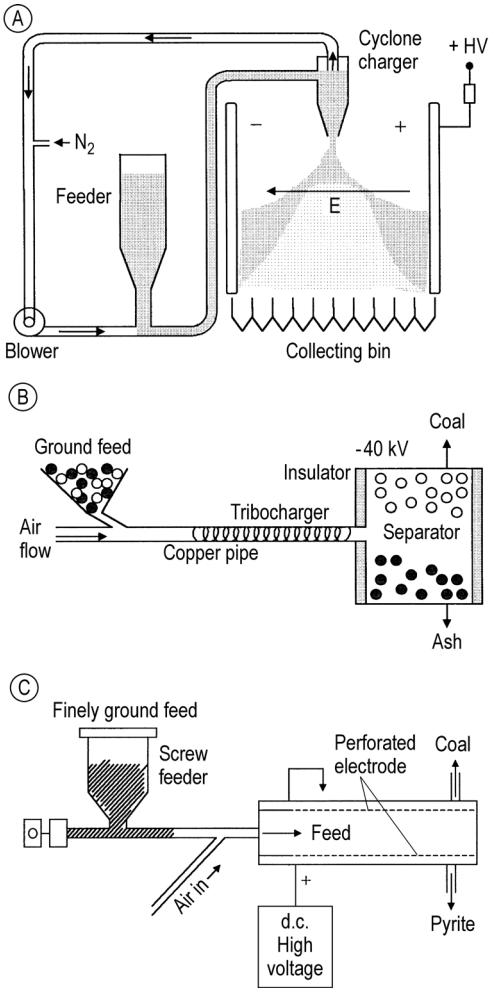

Figure 17. Examples of coal cleaning unit
A) Cyclone charger-plate electrodes [38,56,57]; B) Charging pipe [38,60]; C) Electrostatic sieve conveyor [34,38]

3.3. Recycling Processes

Electrostatic separation for recycling purposes is mainly employed in recovery of values from electronic scrap and separation of wastes from polymer manufacturing.

3.3.1. Inductive or Corona Charging

Due to increasing production of wiring boards or printed circuit boards, the amount of waste arising from the production process of the boards has also increased. Whereas the boards themselves mainly consist of epoxy resin and glass fiber, they also contain significant amounts of precious metals such as gold and silver besides copper. For example, in a process designed by NEC the electronic parts mounted on the boards are separated by heating, impacting, and shearing. The boards are then pulverized to obtain a sufficient liberation of the copper incorporated in the multi-layer resin – glass fiber construction. The mean diameters of the particles are reported to range from 100 to 300 µm. The separation process consists of two steps. First the separation of the glass fiber and resin powder is carried out in a gravimetric air separator. The subsequent separation of the heavier particles is achieved by employment of a rotational drum separator via corona charging. In this step the copper particles are separated from the nonconductive ones. The copper recovery rate was reported to be > 90 % with a copper content of > 60 % in the copper fraction. The copper rich powder is used as a copper resource while the recovered resin powder is utilized in filler applications [48,49].

The other area of employment of rotational drum separators is the recovery of values from wire and cable wastes [8,15,16,52,53,54]. Each year thousands of tons of such materials arising from off-specification production and maintenance procedures are to be processed. The metals – copper and aluminum – are usually recovered by granulation and subsequent air-gravity separation. In such process schemes electrostatic separation is a useful auxiliary technology. Rotational drum separators utilizing inductive and/or corona charging are employed. Metal recovery > 90 % and a metals fraction of 99 % metallics can be achieved. Plastics recycling from chopped wire and cable on the other hand also can be achieved, yielding fractions with 99.9 % plastics and < 0.1 % metallics.

Application of rotational drum separators to recycling purposes also led to interesting developments of special separator types and electrodes (see Fig. 5). However, the employment of these type of separators so far is restricted to process streams of a few hundred kilograms per hour at the most. Therefore in huge process streams a pretreatment by other technologies, especially air-gravity separation has to be carried out to reduce the feed stream to the drum separator.

3.3.2. Triboelectrification

In Figure 18 (see next page) the densities of the most common bulk plastics are shown. As can be derived from this figure, a heavy-media separation or separation by employment of a hydrocyclone can only be achieved in the case of separating of polyethylene (PE) and polypropylene (PP) from other plastics. The situation looks much better when utilizing the different charging behavior of plastics. The triboelectric series of plastics is shown in Figure 19. In principle all plastics in this series can be separated from each other [68, 69, 70, 102, 103, 104, 105]. When two different plastics come into contact, the one on the left side in the triboelectric series will acquire a negative charge, the one on the right side a positive one. The greater the distance of two plastics in the triboelectric series the easier a selective charging will be achieved. For example, in a mixture of poly(vinyl chloride) (PVC) and poly(ethylene terephthalate) (PET), PVC will be charged negatively, PET positively. However contacting of PET with polystyrene (PS) will result in negative charging of PET and positive charging of PS.

Thus electrostatic separation of plastics by triboelectrification will preferably be employed for separation of binary mixtures. A controlled charging of mixtures containing more than two different plastics is quite difficult. An exception to this is PVC. It will always be charged negatively against all other plastics, except poly(vinylidene fluoride) (PVDF). Therefore PVC can be separated from mixtures of numerous other plastics. The plastics that could be successfully separated are listed in Table 4.

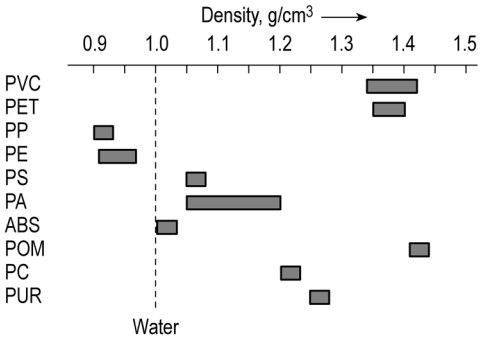

Figure 18. Densities of common plastics

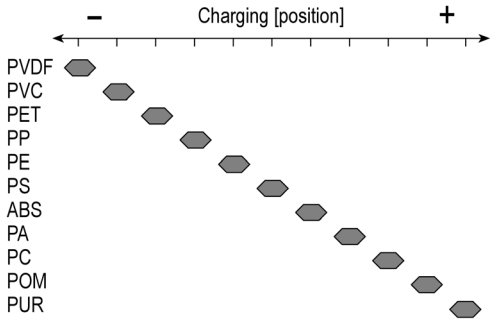

Figure 19. Triboelectric series of common plastics

However, the position of a particular plastic in the triboelectric series can shift. Responsible for this effect are different additives in otherwise chemically identical polymers. Thus one particular material can become charged positively or negatively, depending on the particular additive employed in the plastic. Additives can be dyes, plasticizer, etc. Impurities also can cause a shift of the position in the triboelectric series.

The process of electrostatic separation of plastics is carried out in a similar manner as potash processing. The plastics are first conditioned and the particles are then charged against each other in a fluidized-bed apparatus. As in the case of potash processing, surfactants are used as conditioning agents and added on a ppm level. The humidity of the warming air also has to be adjusted carefully. After charging the particles are separated in an electric field by gravitational forces. The construction of the separators, however, is somewhat different from that of potash separators. Thin tubes with 2 – 3 cm diameter are employed which are arranged in a staggered configuration [106]. This is to avoid that the particles rebound from the electrodes. The distance at the end of the electrodes from each other can be adjusted for every particular separation problem. This can result in a non-parallel arrangement at the electrodes. Another innovation are flow splitters with different length. As a result the amount of impurities in the product could be decreased down to the ppm level.

Some other charging devices and electrode configurations have been developed, but the most of them are still working on the laboratory scale [30, 31, 32]. Until now no large scale industrial separation plant exists.

3.4. Others

Fly Ash Processing. The separator described in Section 2.2.2 is successfully employed in the processing of fly ash. In power plants fly ash is separated into unburned coal and ash. The coal is recycled to the burner whereas the ash is utilized in the manufacturing of building material. A couple of these installations are on stream in power stations in the United States. Up to now these plants are the only examples for successfully running electrostatic dust separation processes on an industrial scale.

Tea Leaves. Many industrial electrostatic separation installations are employed in the tea industries of Japan, Taiwan, Sri Lanka, and India to separate stalks from tea leaves. The separators often consists of two rotational drum separators arranged in sequence. Charging is carried out via induction by a high voltage electrode. The diameter of the rolls in one installation is 150 mm, the distance from the high voltage electrode ranges from 30 to 50 mm. The voltage applied is − 20 kV. After steaming, coarse rolling, fine rolling, and drying the tea leaves with a water content of 7 – 9 % are fed to the finishing process. Here the leaves are classified into powder, sprout, and leaves by screening. The leaves containing many stalks are then fed to the first of two rotational drums from a hopper. They are separated into a stalk fraction attracted by the negative high voltage electrode and a leave fraction falling down to the next stage. Here the remaining stalks are separated, yielding a pure leaves fraction. Throughput often is about 300 kg/h [38].

Electrostatic Decaffeination. An apparatus for separation of charged substances from conductive fluids is reported to be applicable to decaffeination of coffee or tea. The device consists of two chambers connected by a porous wall. One chamber including the wall is charged negatively or grounded, the other is connected to the positive terminal of the field generating assembly. The caffeine fed into the device can be concentrated in one of the chambers whereas the fluid passes the porous wall thus resulting in a decrease of caffeine content. Several special em-

Table 4. Triboelectric separation of plastics

Feed	Polymers	Fedd composition, %	Concentrate purity, %	Yield, %
Production waste	PC/PBT	PBT > 50	PBT > 99	
		PC > 50	PC > 99	
		PBT and PC ca. 50	PC and PBT > 99	
	PVDF/POM	PVDF ca. 90	PVDF > 99.5	90
		POM ca. 10		
	PS/ABS	PS 85	PS 99.5	92
		ABS 15		
Beverage bottles	PET/PVC	PET 65	PET 99.7	80
		PVC 35		
	PET/PVC	PVC ca. 300 ppm	PVC < 20 ppm	95
Car plastics	PA/PP – EDPM	PP – EPDM ca. 10	PA > 99.5	97
		PA ca. 90		
	Noryl/PVC/PUR	Noryl 33.5	Noryl > 99	80
		PVC 39.5	PVC > 95	80
		PUR 27	PUR > 95	> 95
	PP/PUR/ABS	PP 30	PP > 99.5	80 – 90
	PPO/PUR/PVC		PPO > 99	85
			PVC > 95	90
			PUR 95	90
Cable plastics	PVC/EPDM/VPE	PVC 89	PVC 99.97	98
		EPDM 10	EPDM ca. 90	> 99
		VPE 1	VPE ≤ 0.01	
Furniture	PMMA/PC	PMMA 50	PMMA 99.5	> 70
		PC 50	PC 99.5	> 80
Refrigerators	ABS/HIPS/SAN/PVC/PE/ PP/PA/PC/Alu/PUR	ABS 42	ABS + HIPS > 99.5	85
		HIPS 42		
		PVC 6		
Syringes	PE/PP	PE 50	PP 95	
		PP 50		
Bottle tops	PE/PVC	PE 62	PVC > 99	90
		PVC 38	PE > 95	85

bodiments have been patented for continuous or discontinuous employment of the separator [109].

Electrofiltration is primary known as electrostatic precipitation of dust particles from gas streams (→ Dust Separation, Chap. 6.). But electric fields also can be applied in solid – liquid separation as electrostatically enhanced vacuum filtration. This technology has been developed since the 1970s [73, 74]. The unit is similar to an electrolytic cell as shown in Figure 20. Two parallel plate electrodes are placed in the slurry that is to be separated. When a potential is applied across these plates, the particles in the slurry, which normally carry a net charge, are attracted to the electrode with opposite polarity and collected. The cake is then removed as the solid phase product from the electrofiltration. The filtered liquid is drawn off from both the positive and negative electrodes by vacuum methods. This type of solid – liquid separation technology has first been installed in a Georgia kaolin plant for separation of clay.

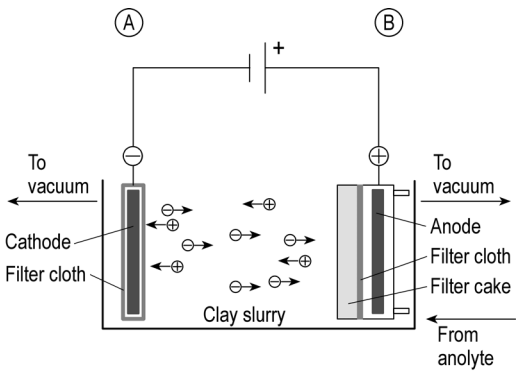

Figure 20. Electrofiltration unit

Electrofiltration can also be utilized for removing of ultrafine particles from liquids, such as, e.g., hydraulic oil [76], deionized ultrapure water for production of silicon integrated circuits [77], or kaolin-containing fats and oils from bleaching [75].

Dielectrophoresis. Dielectrophoretic separations are rarely used on an industrial scale at present and are restricted to removing catalyst fines from petroleum products, and separation of impurities (e.g., fibers) from foods. Dielectrophoresis is now being widely employed in separations of biological materials, such as, e.g., cells of yeast and *E. coli*, erythrocytes, or leucocytes [80, 81, 82]. However, these applications are laboratory methods for scientific or diagnostic work. The separation devices are quite small. Continuously working separators have been developed recently [107].

The prospect of a possible application of dielectrophoresis is provided by the effect of *selective chaining* of particles in an electric field. This means that particles with a dielectric constant higher or smaller than the dielectric constant of the medium will arrange in a linear chain along the force lines of the field. Each chain contains only particles of one species. This has been investigated for the separation of diamonds from heavy minerals of kimberlite rock [37, 108]. The separation takes place in a channel type apparatus where the chains are moving perpendicularly to the force lines of the field. In Figure 21A the cross-section of the feeding end of the channel is shown, the discharging end can be seen in Figure 21B. The high voltage electrode (a) generates the electrostatic force along the grounded flat electrode (b). The feed enters the channel through a single pipe at the mid-point of the cross-section. At the discharging end the material with the higher dielectric constant than the medium is falling through a pipe (g). The material with the lower dielectric constant is discharged through the central pipe (f). The screen (c) prevents particles from bridging the electrodes, which can lead to dielectric breakdown of the liquid medium. The walls (e) consist of insulating material. For vibrational movement of the particles the channel is mounted on a spring (d). In separation tests on diamond – kimberlite the average diamond loss is $< 0.25 - 0.3$ %. When the feed contains 3 % diamonds, the concentrate has a grade of 55 – 60 % with a recovery rate of almost 100 %.

4. References

1. A. D. Moore: *Electrostatics; Exploring, Controlling and Using Static Electricity, Including the Dirod Manual,* Laplacian Press, 1997.
2. H. Schubert: *Aufbereitung fester Stoffe, vol. II: Sortierprozesse,* Deutscher Verlag für Grundstoffindustrie, Stuttgart 1996.

3. E. G. Kelly, D. J. Spottiswood: *Introduction to Mineral Processing,* Wiley & Sons, New York 1982.
4. E. G. Kelly, D. J. Spottiswood, *Miner. Eng.* **2** (1989) 33 – 46.
5. E. G. Kelly, D. J. Spottiswood, "Particle Charging," *Miner. Eng.* **2** (1989) 193 – 205.
6. E. G. Kelly, D. J. Spottiswood, "The Separation of Particles," *Miner. Eng.* **2** (1989) 337 – 349.
7. B. A. Kwetkus, *Part. Sci. Technol.* **16** (1998) 55 – 68.
8. L. Dascalescu, R. Morar, A. Iuga, A. Samuila, V. Neamtu, *Part. Sci. Technol.* **16** (1998) 25 – 42.
9. M. S. Celik, E. Yasar, *Miner. Eng.* **8** (1995) 829 – 833.
10. A. D. Dance, R. D. Morrision, *Miner. Eng.* **5** (1992) 751 – 765.
11. F. S. Knoll, J. B. Taylor: *Design and Installation of Concentrating and Dewatering Circuits,* Soc. of Mining Engineers of AIME, Littleton, CO 1986, pp. 208 – 225.
12. R. Morar, I. Alexandru, I. Cuglesan, O. Muntean, L. Dascalescu, Proc. of the 1997 IEEE Industry Applications Conf. 32nd IAS Annual Meeting, (1997) Part 3, 1716 – 1723.
13. J. E. Lawver, R. M. Funk, *Min. Eng. (N.Y.)* **22** (1970) 71 – 73.
14. R. M. Funk, J. E. Lawver, *Trans. Amer. Inst. Min. Metall. Pet. Eng.* **247** (1970) 23 – 26.
15. A. Iuga, L. Dascalescu, R. Morar, I. Csorvassy, V. Neamtu, *J. Electrostat.* **23** (1989) 235 – 243.
16. S. Tenorio, J. Alberto, Extraction and Processing Division TMS Annual Meeting 1997, Minerals, Metals & Materials Soc. (TMS), Warrendale, PA 1997, pp. 505 – 509.
17. US 4 247 390, 1981 (F. S. Knoll).
18. F. Fraas: "Electrostatic Separation of Granular Materials," *US Bur. Mines Bull.* **603** (1962).
19. A. D. Dance, PhD Thesis, University of Queensland, 1992.
20. A. G. Berlage, D. B. Churchill, D. M. Bilsland, T. M. Cooper, *Trans. ASAE* (1990) 597 – 600.
21. H. Bauser, *DECHEMA Monogr.* **72** (1974) 11 ff.
22. J. Heidberg, S. Wiesner, *Glueckauf* **135** (1999) 123 – 127.
23. J. Heidberg, A. Rugullies, I. Stahl, M. Thomas, *Surf. Sci.* **428** (1999) 426 – 430.
24. G. Zhou, K. Moon, *Can. J. Chem. Eng.* **72** (1994) 78 – 84.
25. L. Ernst, *Kali Steinsalz* **9** (1986) 275 – 286.
26. L. Ernst, *Ber. Bunsenges. Phys. Chem.* **93** (1989) 857 – 863.
27. L. Ernst, *Ber. Bunsenges. Phys. Chem.* **94** (1990) 1435 – 1439.
28. Kali und Salz, DE 1 154 052, 1960 (H. Autenrieth, H. Dust).
29. P. M. Beier, I. Stahl, Proc. of the XX. International Mineral Processing Congress, Aachen 1997, **2** (1997) 657 – 665.
30. I. I. Inculet, G. S. P. Castle, J. D. Brown, *Part. Sci. Technol.* **16** (1998) 91 – 100.
31. I. I. Inculet, G. S. P. Castle, J. D. Brown, Proc. 29th IAS Annual Meet. (1994) Part 2, 1397 – 1399.
32. D. K. Yanar, B. A. Kwetkus, *J. Electrostat.* **35** (1995) 257 – 266.
33. A. Mukherjee, D. Gidaspow, D. T. Wasan, *ACS Fuel Chemistry* **32** (1987) 395 – 407.
34. D. Gidaspow, R. Gupta, A. Mukherjee, D. Wasan, *Processing and Utilization of High-Sulfur Coals II,* 1987, pp. 271 – 281.

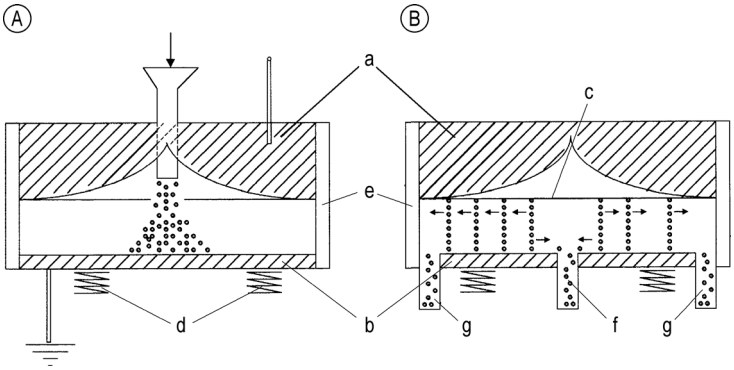

Figure 21. Schematic of the electrode system used in the dielectrophoresis of diamond
A) Feeding part; B) discharge end of the separator
a) High-voltage electrode; b) Grounded flat electrode; c) Screen; d) Spring; e) Wall; f) Central pipe; g) Pipe

35. Advanced Energy Dynamics, Int. Separations Systems, ADV Energy Dynamics, US 4 839 032, 1989 (D. R. Whitlock).
36. US 224 719, 1880 (J. T. Osborne)
37. U. Andres, *J. Electrostat.* **37** (1996) 227 – 248.
38. Y. Higashiyama, K. Asano, *Part. Sci. Technol.* **16** (1998) 77 – 90.
39. L. Yongzhi, Proc. 1st Int. Conf. Modern Process Mineral. Miner. Processing, Beijing 1992, 385 – 390.
40. F. S. Knoll, J. B. Taylor, *Trans. Amer. Inst. Min. Metallurg. Pet. Eng.* **278** (1985).
41. *Miner. Metallurg. Process.* 106 – 114.
42. W. P. Dyrenforth: "Electrostatic Separation," in A. L. Mular, R. B. Bhappu (eds.): *Mineral Processing Plant Design* Society of Mining Engineers, New York 1980, pp. 479 – 489.
43. E. Okur, G. Onal, Proc. 1st Int. Conf. Modern Process Mineral. Miner. Process., Beijing 1992, 391 – 396.
44. O. C. Ralston: *Electrostatic Separation of Mixed Granular Solids,* Elsevier, Amsterdam 1961.
45. Y. Higashiyama, I. Umetsu, K. Asano, *Proc. IEJ* **14** (1990) 234 – 240 (jap.).
46. W. N. Agosto, *Adv. Astronaut. Sci.* **53** (1983) 315 – 334.
47. K. Mah, J. M. Toguri, H. W. Smith, *Conserv. Recycl.* **9** (1986) 325 – 334.
48. M. Iji, S. Yokoyama, *Circuit World* **223** (1997) 10 – 15.
49. S. Yokoyama, M. Iji, Conf. Rec. 1997 IEEE Int. Symps. Electron. Environ., 1997, 109 – 114.
50. T. Harada, S. Owada, T. Otsuki, *Proc. Ann. Meet., Jpn. Mining Assoc.* (1987) 109 – 110 (jap.).
51. J. B. Taylor, *Conf. Rec. IAS Ann. Meet.* **23** (1988) 1741 – 1759.
52. J. Kojima, *Proc. IEJ* **6** (1982) 152 – 158 (jap.).
53. L. Dascalescu, A. Iuga, R. Morar, *Magn. Electr. Sep.* **4** (1993) 241 – 255.
54. L. Dascalescu, A. Iuga, R. Morar, *Ann. Rep. Conf. Electr. Insul. Dielectr. Phenom. 1989* (1989) 143 – 148.
55. S. Hayakawa, E. Sakaguchi, R. Yamashita, *J. Inst. Agric. Machinery* **47** (1985) 1, 107 – 112 (jap.).
56. S. Masuda, M. Toraguchi, T. Takahashi, K. Haga, *IEEE Trans. Ind. Appl.* **IA-19** (1983) 789 – 793.
57. J. Bouwma, *Conf. Rec. IAS Ann. Meet.* **19** (1984) 966 – 971.
58. S. Toraguchi, M. Nagasawa, Conf. Pap. IEJ, Tokyo (1982) 289 – 292 (jap.).
59. R. Gupta, D. Gidaspow, D. T. Wasan, *Powder Technol.* **75** (1993) 79 – 87.
60. D. Finseth, T. Newby, R. Elstrodt: "Dry Electrostatic Separation of Fine Coal", B. K. Parakeh, J. G. Groppo (eds.): *Processing Utilization High-Sulfur Coals V,* Elsevier, Amsterdam 1993, pp. 91 – 98.
61. D. Kunze, I. Stahl, *Kali Steinsalz* **12** (1997) 135 – 140.
62. G. Fricke, *Kali Steinsalz* **9** (1986) 287 – 295.
63. A. Singewald, G. Fricke, *Chem. Ing. Tech.* **55** (1983) 39 – 45.
64. A. Singwald, *Kali Steinsalz* **8** (1982) 252 – 260.
65. R. Bock, *Chem. Ing. Tech.* **53** (1981) 916 – 924.
66. G. Zhou, J. D. Brown, *Can. J. Chem. Eng.* **66** (1988) 858 – 863.
67. I. I. Inculet, G. S. S. Castle, Inst. Phys. Conf.-Ser. No. 118, Section 4, Electrostatics'91, Oxford 1991.
68. I. Stahl, P.-M. Beier, Proc. 20th Int. Mineral Processing Congress, Aachen, **5** (1997) 395 – 401.
69. I. Stahl, P.-M. Beier, *Erzmetall* **49** (1996) 1, 46 – 49.
70. I. Stahl, U. Kleine-Kleffmann in J. Brandrup, M. Bittner, W. Michaeli, G. Menges (eds.): *The Recycling of Plastics,* Carl Hauser Verlag, Munich – Vienna 1992, pp. 256 – 264.
71. C. W. Bouchiillon, W. G. Steele, *Particulate Sci. Technol.* **10** (1992) 73 – 89.
72. Y. Higashiyama, Y. Ujiie, K. Asano, *Proc. IEJ-ESA 1996 Joint Symp. Electrostatics, Tokyo* (1996) 211 – 217 (jap.).
73. R. A. Adams, *Proc. Int. Symp. Hydrometall. 3rd,* 1983.
74. M. P. Freeman, The Clay Minerals Society Meeting, Macon, GA, Aug. 1979.
75. W. Zschau, *Fett/Lipids* **100** (1998) 513 – 517.
76. S. G. Nash, *Iron Steel Eng.* **75** (1998) 38 – 39.
77. R. Mohindra, W. Kern, *Semicond. Int.* **20** (1997) 191 – 192.
78. P. J. Bailes, *Hydrometallurgy* **30** (1992) 417 – 430.
79. G. R. Fritsche, *Oil Gas J.* (1977) 73 – 81.
80. R. Pethik, Proc. 4th Toyota Conf. 21 – 24 October 1990, Aichi 1991 159 – 185.
81. Y. Huang, R. Pethik, *Meas. Sci. Technol.* **2** (1991) 1142 – 1146.
82. R. Pethik, Y. Huang, X. B. Wong, J. P. Bart, *J. Phys. D: Appl. Phys.* **25** (1992) 881 – 888.
83. Intern. Minerals and Chem. Corp., GB 723 477, FR 1 122 212, (J. E. Lawver).

84. I. M. Le Baron, W. C. Knopf, *Miner. Sci. Eng.* **10** (1958) 1081 – 1083.
85. Kali und Salz, DE 1 078 961, 1959 (R. Bachmann).
86. Kali und Salz, DE 1 076 593, 1957 (H. Autenrieth).
87. Kali und Salz, DE 1 180 331, 1963 (G. Peuschel).
88. Kali und Salz, DE 1 283 772, 1967 (H. Autenrieth, H. Wirries).
89. Kali und Salz, DE 1 792 120, 1968 (A. Singewald, G. Fricke).
90. Kali und Salz, DE 1 953 534, 1969 (A. Singewald, G. Fricke).
91. Kali und Salz, DE 1 056 551, 1957 (H. Autenrieth).
92. Kali und Salz, DE 2 619 026, 1976 (A. Singewald, G. Fricke, U. Neitzel, I. Geisler).
93. Kali und Salz, DE 1 129 903, 1958 (H. Autenrieth, G. Peuschel).
94. Kali und Salz, DE 1 142 802, 1961 (H. Autenrieth, G. Peuschel, G. Weichert).
95. Kali und Salz, DE 1 249 783, 1966 (A. Singewald, G. Fricke).
96. Winterschall, DE 1 966 154, 1969 (A. Singewald, G. Fricke).
97. Kali und Salz, DE 1 261 453, 1967 (A. Singewald, G. Fricke).
98. Kali und Salz, DE 1 667 814, 1968 (G. Fricke, A. Singewald).
99. Kali und Salz, DE 1 259 803, 1965 (A. Singewald, G. Fricke, H. Erbe).
100. Kali und Salz, DE 1 279 572, 1965 (A. Singewald, G. Fricke, H. Erbe).
101. I. I. Inculet et al., *CIM Bulletin* **73** (1980) 51 – 61.
102. Kali und Salz, DE 3 035 649, 1980 (G. Fricke, I. Geisler).
103. Kali und Salz, EP 0 553 315, 1992 (I. Stahl, A. Hollstein, U. Kleine-Kleffmann, I. Geisler, U. Neitzel).
104. Kali und Salz, EP 0 553 320, 1992 (I. Stahl, A. Hollstein, U. Kleine-Kleffmann, I. Geisler, U. Neitzel).
105. Kali und Salz, EP 0 553 319, 1992 (I. Stahl, A. Hollstein, U. Kleine-Kleffmann, I. Geisler, U. Neitzel).
106. Kali und Salz, DE 4 438 704, 1994 (N. Raschka, I. Stahl).
107. G. H. Markx, R. Pethik, *Biotechnol. Bioeng.* **45** (1995) 337 – 343.
108. GB 2 087 267, 1981 (U. Andres).
109. IMSCO Technologies, WO 9 853 895, 1997 (J. R. Crose, S. L. Berg).

Gravity Concentration

See also → Solid – Solid Separation, Introduction

RICHARD O. BURT, Tantalum Mining Corporation of Canada Ltd., Lac du Bonnet, Manitoba, Canada REO 1 AO

1. **Introduction** 1
2. **Principles** 1
3. **Choice of Equipment** 2
4. **Stratification Devices** 4
5. **Flowing-Film Devices** 5
6. **Shaking Separators** 6
7. **Slime Separators** 7
8. **Dry Gravity Concentration** 8
9. **Flow Sheet Example** 8
10. **References** 8

1. Introduction

Gravity concentration is the separation of one or more minerals, usually of different density, from others by relative movement in response to the force of gravity and other forces.

Next to handpicking, gravity concentration is the most ancient mineral-processing method, with early practices described over 2000 years ago [1]. One of the classic works on the subject is the 6th century book by AGRICOLA [2]. In the 20th century, the importance of gravity concentration declined somewhat as other processes, such as flotation or magnetic and electric separation methods, were developed (→ Flotation, → Magnetic Separation, → Electrostatic Separation). Nevertheless, it remains one of the major recovery processes in both mineral processing and, especially, coal preparation (→ Coal, Chap. 7.1.2.1.). Furthermore, the development of high-capacity equipment, along with heightened awareness of the negative environmental impact of many chemicals used in flotation and leaching, has resulted in renewed interest in gravity concentration for a wide range of products.

Today gravity concentration is used for a diverse range of minerals: from andalusite to zircon, from coal to diamond, from mineral sands to metal oxides, and from industrial minerals to gold.

2. Principles

Important factors in determining the relative movement of a particle in a fluid include relative density, mass, size, and shape of the particle.

If, in a hypothetical two-mineral separation, only one of these factors is significantly different, separation is relatively easy. In nearly every case, however, the variety of minerals present have different relative densities and a range of particle masses, sizes, and shapes. The ease or difficulty of separating one species from another depends on the relative differences between these factors and whether such differences assist or oppose separation. Some idea of the ease of separating two minerals can be obtained from the *concentration criterion* (CC), which is defined as follows:

$$\mathrm{CC} = \frac{d_\mathrm{h} - d_\mathrm{f}}{d_\mathrm{l} - d_\mathrm{f}}$$

where d_h is the relative density of the heavy mineral; d_l the relative density of the light mineral; and d_f the density of the suspending fluid.

In general, if $\mathrm{CC} > 2.5$, then separation is relatively easy; however, if $\mathrm{CC} < 1.25$ separation is not likely to be effective. For developmental test work, the concentration criterion is not sufficiently accurate, and tests involving heavy-liquid separation should be carried out (→ Dense-Medium Separation, Chap. 5.).

Gravity concentration is effected by the relative movement of minerals of different density in response to gravity and one or more other forces, which often include the resistance to motion offered by a viscous fluid. One characteristic of all separation devices is that, for the particles to

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_21

move relative to one another, at some stage of the process the particles must be kept slightly apart, or *dilated*, by the superimposed force.

No single mechanism can satisfactorily explain the behavior of any particular device; a combination of two or more mechanisms comes closer to describing the operation (but rarely does so completely). Conversely, however, various mechanisms have been proposed that are common to apparently diverse items of equipment and gravity concentration processes, and these are a useful means of classifying the various devices (see Fig. 1) [3].

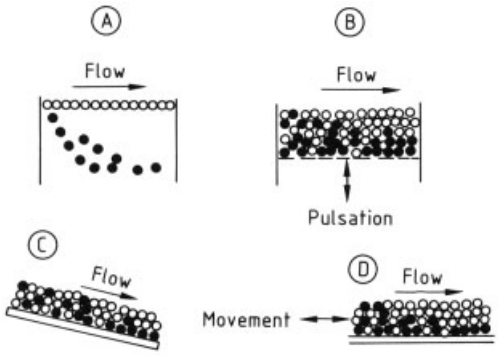

Figure 1. Mechanisms in gravity concentration
A) Density; B) Stratification; C) Flowing film; D) Shaking
○ = light; ● = heavy

Density (Fig. 1 A). The viscous fluid used has a density, or apparent density, between that of the minerals to be separated. Thus, one mineral(s) has a net positive buoyancy and floats, whereas the other(s) has a net negative buoyancy and sinks. This classification includes dense-medium separation (→ Dense-Medium Separation), one of the most important gravity concentration processes, and magnetohydrostatic separation.

Stratification (Fig. 1 B). An intermittent fluidization caused by the pulsation of the fluid in a vertical plane stratifies the various mineral constituents. This classification applies to a wide range of jigs.

Flowing Film (Fig. 1 C). The constituents are separated by their relative movement through a stream of slurry, which flows down a plane by the action of gravity. Flowing-film concentration is the oldest process used in gravity concentration and remains of major importance, with such units as the sluice, pinched sluice, Reichert cone, and, through inclusion of lateral movement caused by the addition of centrifugal force, the spiral.

Shaking (Fig. 1 D). The mineral components are stratified by superimposing a horizontal shear force on the flowing film. This force can be oscillating, as in a shaking table, or orbital, as in the Bartles – Mozley separator or crossbelt concentrator.

The theoretical aspects of these and other mechanisms involved in gravity separation are discussed in [4].

3. Choice of Equipment

A wide range of gravity concentration equipment is available. Factors that determine choice of the correct unit include the size range of the particles to be separated, the duty required, the desired throughput and efficiency, and–all other things being equal–the unit cost, both capital and operating, of the equipment.

Size Range. The overall particle-size range that can be separated by using gravity concentration is larger than for any other process. Particles can be separated from each other as long as there is a suitable density differential between them; however, the top size may be limited by the mechanical handling ability of the device, 500 mm is generally accepted as the largest size for separation. The practical lower cutoff limit for fine gravity concentration is ca. 6 µm.

No device is capable of treating the entire size range amenable to gravity concentration. In addition, few devices operating as a single unit can efficiently treat the complete range which they are designed to handle. Therefore, some form of feed preparation is normally required. Figure 2 indicates the optimum size range for some of the more common equipment types [4]. Jigs, Reichert cones, and sluices are essentially mass classification devices and are capable of treating a fairly wide range of sizes, but some loss of efficiency occurs at the top, bottom, and middle of the size range. Treatment of material in *series circuits* is practical with these units (Fig. 3).

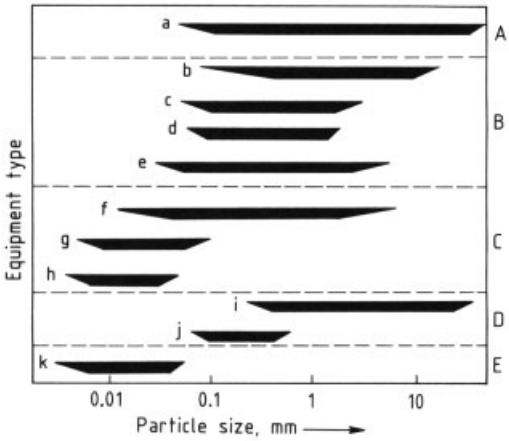

Figure 2. Operating range of gravity concentration equipment
A) Stratification devices: a) jig
B) Flowing-film devices: b) Sluice box; c) Reichert cone; d) Pinched sluice; e) Spiral
C) Shaking devices: f) Shaking table; g) Bartles – Mozley; h) Crossbelt
D) Air devices: i) pneumatic jig; j) Air table
E) Miscellaneous devices: k) centrifugal separators

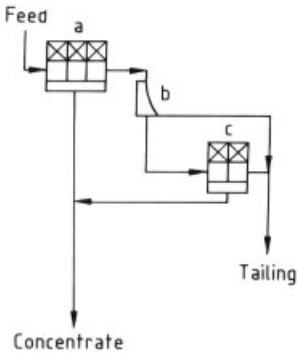

Figure 3. Series circuit
a) Jig; b) Screen; c) Jig

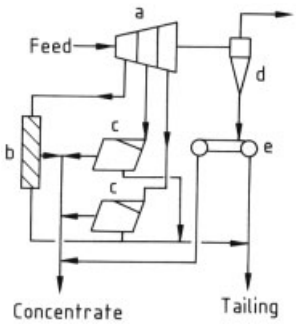

Figure 4. Parallel circuit
a) Hydrosizer; b) Spiral; c) Shaking table; d) Cyclone; e) Crossbelt concentrator

Other types of equipment (e.g., the table or Bartles – Mozley separator, concentrating by reverse classification, or the spiral concentrating by hindered settling) operate significantly better on sized or classified feed, and treatment in *parallel circuits* is required (Fig. 4).

Generally, the two types of circuit are combined. For example, a series circuit of jigs treating unsized feed may be followed by a sizing step, sized fractions being heated on a parallel circuit of tables and film sizing units.

Capacity. The capacity range of gravity concentration equipment is enormous (from a few kilograms per hour to ca. 1000 t/h per unit). Whereas sluices and jigs are manufactured in a variety of sizes suited for a range of capacity, other items are produced in one or two sizes only, and high throughputs require multiple units. Typical capacities for gravity concentration equipment are as follows (in tons per hour):

Jig	$\leq$800
Sluice	$\leq$500
Reichert cone	60 – 70
Spiral	1 – 12
Table (ore)	0.2 – 2
Table (coal)	8 – 12
Bartles – Mozley Separator	2 – 2.5
Crossbelt Concentrator	0.3 – 0.5
Centrifugal	1 – 15

Efficiency. Most devices can efficiently separate materials containing only very heavy and very light fractions. However, most materials consist of particles with a range of densities. The more efficient a device, the more capable it is of effectively separating near-density particles [5].

Equipment efficiency classified by this measure is shown in Table 1.

Table 1. Process applicability–density of separation

Wt % of feed within $\pm$0.10 kg/L of density of separating medium	Degree of difficulty expected	Type
0 – 7	simple	sluices, jigs
7 – 10	relatively simple	cones
10 – 15	moderately difficult	tables spirals, DMS *
>15	difficult	DMS *

* Dense-medium separation

4. Stratification Devices

Jigs sort by movement of a bed of particles, which are intermittently fluidized by pulsation of the fluid in a vertical plane. The particles become arranged in layers, with density increasing from top to bottom. This particle arrangement (stratification) results from several continuously varying forces which act on the particles. Although many types of jigs are manufactured, their construction principles are similar (Fig. 5).

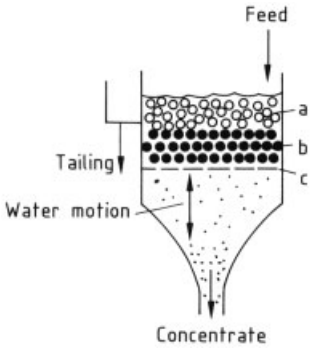

Figure 5. Basic construction of a jig
a) Jig bed; b) Ragging; c) Jig screen

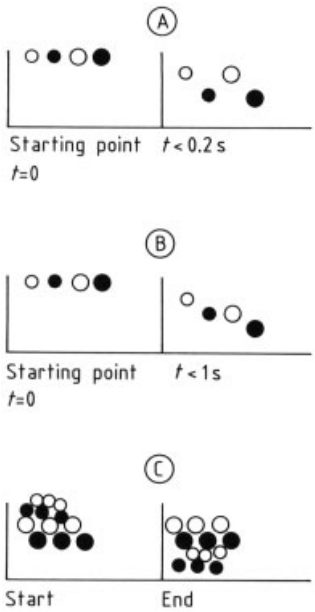

Figure 6. Three mechanisms of jigging
A) Differential acceleration; B) Hindered settling; C) Interstitial trickling
○ = light; ● = heavy

Theories postulated to describe the action of the jig include differential acceleration, hindered settling, and interstitial trickling (Fig. 6). Differential (initial) acceleration at the start of particle movement is related to particle density, rather than particle size or mass. Hindered settling occurs once the particles have reached their terminal velocity; movement is related to mass. In the interstitial trickling phase, movement is related to particle size. An alternative theory proposes that the mechanism of concentration is the attainment of minimum potential energy in the heterogeneous mass [6]. Each of these mechanisms undoubtedly plays some part in the stratification process; however, which (if any) is the most important is open to question. Because the design of the pulsation cycle (i.e., duration of the cycle and the intensity of both upward and downward pulsation) can be varied, different manufacturers have developed units that accentuate different mechanisms.

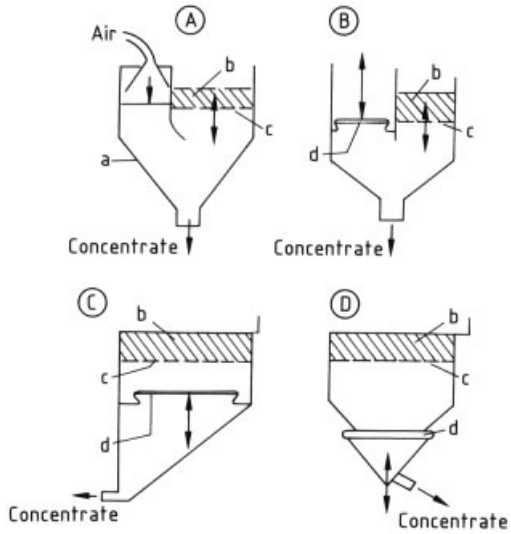

Figure 7. Typical jigs
A) Baum and OPM; B) Denver; C) Bendelari; D) IHC
a) Air box; b) Jig bed; c) Screen; d) Diaphragm

Jigs that rely on a long pulsation phase and a short suction phase, such as the Baum, OPM (Fig. 7 A), and Batac coal-washing jigs, tend to have a high capacity at the expense of recovery of finer particles. Although less efficient than heavy-medium separation, the high capacity and low operating cost of these jigs ensure that they

remain the primary means of concentrating thermal coal as well as metallurgical coal where the amount of near-density shale is not excessive.

Units with a longer suction phase, such as the Denver, Bendelari, and IHC (Fig. 7 B – D), exhibit a higher recovery of fines with some compromise to capacity. Therefore, these units are more commonly used for mineral recovery, such as alluvial tin or hard-rock gold ores.

5. Flowing-Film Devices

When a film of water flows down an inclined surface under laminar conditions, the velocity distribution is parabolic, being zero at the bottom of the film and maximal at the top. When mineral particles of different size and relative density are introduced into this laminar film, they arrange themselves in the general downslope sequence shown in Figure 8. A variety of devices use this mechanism for separation.

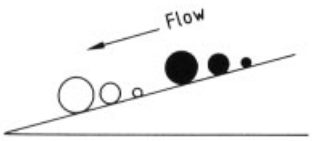

Figure 8. Particle arrangement on an inclined surface
○ = light; ● = heavy

Sluice Box. A sluice box is essentially an inclined trough or launder, usually of rectangular cross section, through which a wide size range of feed is washed by a large volume of water. A series of riffles is arranged at the bottom of the sluice box so as to provide some turbulence between each riffle and allow concentration of heavies.

These units are relatively inefficient and generally used only for alluvial mining in areas where power is not readily available.

Pinched Sluice and Reichert Cone. The principle of operation of a typical *pinched sluice* is shown in Figure 9.

Pulp, at a high solids density, enters the channel in a relatively thin stream at the upper or wider end. As this stream flows down the sluice, the finer and heavier particles concentrate (a) in the lower levels of the stream and increase in depth as the width of the sluice decreases. At the end of the channel, the slower moving lower level of the stream, enriched in heavy minerals, is peeled away from the main stream by a splitter (b).

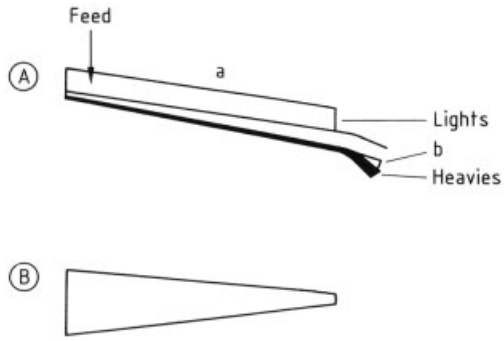

Figure 9. Pinched sluice
A) Cross section: a) Concentrating segment; b) Splitter
B) Plan view

A variety of pinched sluice designs are used, especially in Australia. Elsewhere, the more common variant is the *Reichert cone* (Fig. 10). The inverted cone concentrating surface of this device may be viewed as a number of pinched sluices placed adjacent to one another to form a circle, without any sidewalls to interfere with stratification [7]. The cone concentrator consists of a number of stages in series and in parallel; it can thus achieve multiple stages of separation in one unit. This is important because neither cone nor sluice is particularly efficient in a single stage.

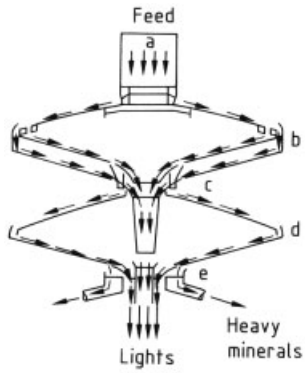

Figure 10. Reichert cone
a) Feed distributor; b) Double-cone assembly; c) Adjustable insert; d) Single-cone assembly; e) Variable insert

The advantages of the Reichert cone are its high capacity, low capital cost per unit throughput, light weight, compactness, and metallurgical performance. Cones have two disadvantages: one is the need for a feed with 60 % solids because the device is sensitive to variations in pulp density; the other is that cones are unsuitable for operations with a new feed rate of less than ca. 50 t/h.

Spiral Concentrator. Spiral concentrators are used widely in mineral processing and, more recently, in coal preparation. The basic design of the spiral has remained virtually unchanged for many years (Fig. 11).

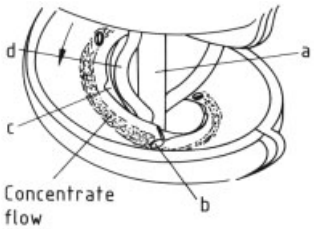

Figure 11. Spiral concentrator
a) Center column; b) Splitter; c) Wash-water channel; d) Concentrate pipe

Recently, however, spiral technology has developed rapidly, more so perhaps than that of any other gravity concentrator. Developments, mainly in Australia, the former Soviet Union, the Peoples Republic of China, and the United States, have included the introduction of a range of trough profile, pitch, and diameter; the introduction of limited offtake spirals; and the use of spirals that do not require wash water [8]. The correct choice of these parameters for a specific duty is probably the most important factor in efficient spiral flow sheet design.

Concentration in a spiral is a complex blend of a variety of forces: hindered settling classification, interstitial trickling, Bagnold shear, and stream cross-sectional rotation. Many gravity concentration devices use the first two forces, several employ the first three, but the addition of cross-sectional rotation makes the spiral unique as a self-compensating device.

6. Shaking Separators

Shaking Table. The shaking table (Fig. 12) remains one of the workhorses of the mineral-processing industry [9].

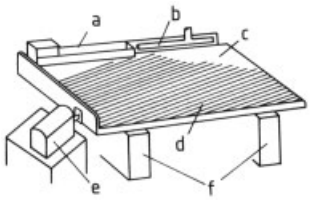

Figure 12. Shaking table
a) Feedbox; b) Wash water; c) Deck; d) Riffles; e) Shaking mechanism; f) Supports

The forces by which the shaking table separates the heavy particles from the lights are complex and, to some extent, opposing. Particles are stratified into discrete layers behind a series of riffles (Fig. 13) by superimposing, on the flowing film, a horizontal, asymmetrically reciprocating shear force acting laterally to the flow. The stratified layers are separated from each other by a lateral movement caused by the asymmetry of the shear force.

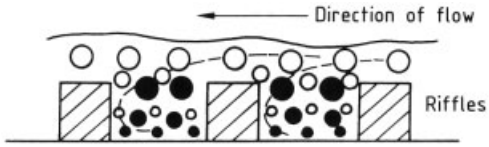

Figure 13. Stratification behind riffles

The multidecked Concenco (USA) and SKO (former states of USSR) units are ideally suited for roughing duties in larger plants and for coal preparation; the latter application is by far the most important in terms of volume. Single-deck units with mechanical head motions, such as the Holman (UK), Deister (USA), SKM (former states of USSR), Wilfley (UK and USA), and Yun Tin (Peoples Republic of China) types, are better suited for smaller plants or for cleaning duty, mainly in mineral processing. In such operations, careful feed preparation and treatment of table middlings on separate equipment are essential for optimum performance.

7. Slime Separators

The inherent difficulty of recovering very fine particles (< 50 µm) by gravity has led to the development and, in many cases, obsolescence of a more diverse range of equipment than for any other particle-size range.

The Endless Belt concentrator and Johnson barrel are used in South Africa to recover fine gold in milling circuits. Likewise, the Rocking – Shaking vanner is used in China for the treatment of tin, tungsten, and iron ores. The more widely used Bartles – Mozley separator and the Bartles Crossbelt concentrator superimpose, an orbital shear motion on a slightly inclined surface, and are capable of recoveries as fine as 6 µm. The former (Fig. 14) is a semibatch device that discharges concentrates discontinuously; the latter is a continuous device that uses a slowly moving belt to remove concentrates from the pulp stream.

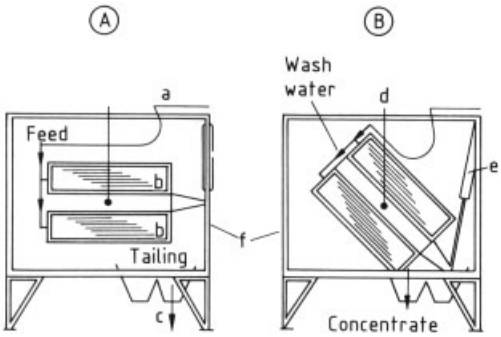

Figure 14. Bartles – Mozley separator
A) Collecting step: a) Feedwater valve assembly; b) Two sandwiches; c) Product launders
B) Discharging step: d) Suspension cables; e) Tilt cylinder; f) Main frame

Spiral sluices are popular in both the former Soviet Union and the Peoples' Republic of China. Similar in principle to the spiral separator, the cross section of the helix is much flatter, enabling recoveries as fine as 15 µm to be achieved.

The GEC (General Electric Co.) Duplex concentrator uses side shear rather than orbital shear. It is a semibatch device, with the decks tilting sideways rather than lengthways as on the Bartles – Mozley separator. Centrifugal separators are growing in popularity. The Knelson concentrator consists of a riffled basket rotating on a vertical axis with a continuous supply of backpressure water in the riffles to keep the bed of ore (usually gold) mobile. The Yunnan separators used in China for iron, gold, and metal oxide recovery are smooth-walled drums rotating on a horizontal axis. Both units are semibatch devices.

8. Dry Gravity Concentration

The majority of commercially available, dry gravity concentration processes are similar in principle to their wet gravity concentration counterparts. The latter include dense media separation, jigging, pinched sluices, and tabling. Although the mechanisms of separation of the wet and dry units may not be the same, the wet units already described make a reasonable point of reference for their dry counterparts. Other dry gravity concentration devices, which do not have a direct counterpart in wet concentration, have been developed. These include the fluidized-bed separator and the zigzag.

All dry concentrating equipment, except the dry jig, uses a constant controlled upflow of fluidizing air. The dry jig, however, uses a pulsating air flow; pulsation effects stratification in a manner comparable to its counterpart in wet concentration.

Dry separators generally require closely sized feed and are less efficient than wet separators. Consequently, they are used only where water is not available, or for "cleanup" of rough concentrates already dried for other processes.

9. Flow Sheet Example

Apart from sample separations, plants rarely incorporate only one type of device. This is especially true of metal oxide plants. Figure 15 (see next page) is an example of the complex flow sheet that may be required to maximize both recovery and purity of the valuable minerals.

Acknowledgment . Material from this section is drawn mainly from the book *Gravity Concentration Technology,* written by the author (assisted by C. Mills) and published by Elsevier Science Publishers BV, Amsterdam, who retain the copyright. Their permission to utilize this material is duly acknowledged.

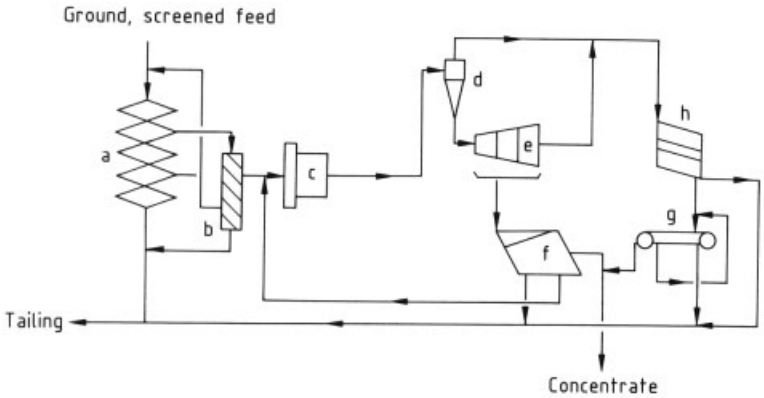

Figure 15. Example of gravity concentration circuit
a) Reichert cone; b) Spiral; c) Grinding mill; d) Cyclone; e) Classifier; f) Shaking table; g) Crossbelt concentrator; h) Bartles–Mozley separator

10. References

1. C. P. S. Pliny: *Natural History Book 33,* ca. 70 A.D.
2. G. Agricola: *De Re Metallica,* translated by H. C. Hoover, L. H. Hoover, Dover Publications, New York 1950, Book XIII.
3. E. G. Kelly, D. J. Spottiswood: *Introduction to Mineral Processing,* Wiley Interscience, 1982.
4. R. O. Burt (assisted by C. Mills): *Gravity Concentration Technology,* Elsevier, Amsterdam 1984.
5. J. W. Leonard, D. R. Mitchell: *Coal Preparation,* 3rd ed., AIME, New York 1968, Section 4, pp. 23–30.
6. F. W. Mayer: *Proc. 7th Int. Mineral Processing Cong.,* Gordon and Breach, New York 1964, pp. 75–97.
7. J. J. Ferree, *Min. Eng.* **25** (1973) no. 3, 29–31.
8. R. O. Burt, A. V. Yashin, *Mineral Processing & Extract. Metall.,* IMM. London 1984, pp. 117–128.
9. R. O. Burt, *XV. Int. Mineral Processing Cong.,* Cannes 1985, pp. 272–281.
10. A. V. Yashin, M. F. Anekin, V. A. Skrepko, *Spiral Concentrator Operation,* Nedra Press, Moscow 1984, pp. 54–65.

Grignard Reaction → Magnesium Compounds
Grinding → Abrasives
Grinding → Size Reduction
Grouting, Chemical → Cement and Concrete
Grouting, Chemical → Silicates
Guanamines → Melamine and Guanamines

Dense-Medium Separation

See also → Solid – Solid Separation, Introduction

RICHARD O. BURT, Tantalum Mining Corporation of Canada Ltd., Lac du Bonnet, Manitoba, Canada ROE 1AO

1. Introduction 1
2. The Process 1
3. The Separatory Vessel 2
4. The Medium 4
5. Testing for Dense-Medium Separation . 5
6. References 7

1. Introduction

Dense-medium separation (DMS), also known as heavy-medium separation or sink – float separation, is one of the most widely applied and most efficient gravity separation processes in both mineral processing and coal preparation. Essentially, it is a process in which particles of different relative densities are separated in a liquid or in a fairly stable suspension of predetermined density chosen such that the liquid density is higher than that of the lighter particles but lower than that of the heavier particles.

The range of operating density of separation is 1300 – 3800 kg/m^3. The medium can be any suitable fluid; however, in practice, it is invariably a suspension of fine, normally magnetic particles in water. Dense-medium separation effectively separates particles with density differentials of $<$100 kg/m^3 in the size range of 500 – 0.5 mm. The largest size is dictated by the mechanical considerations of the separating vessel; the smallest, by medium rheology.

Uses. Dense-medium separation has two main areas of application. It is used for preconcentration to obtain an economically acceptable waste product that does not warrant further treatment, such as in the processing of diamonds, sulfides, and metal oxide ores. This allows either (1) a significant reduction in the quantity of ore that requires grinding prior to other separation methods or (2) the treatment of a lower grade ore than would otherwise be economically viable. Alternatively, DMS is used to produce a salable end product, as in coal preparation and industrial mineral separation.

2. The Process

Unlike most other gravity concentration techniques, dense-medium separation is a system rather than a unit process [1]. To work effectively and economically, the system must consist of a series of interconnected phases:

1) feed preparation,
2) feed and medium presentation,
3) separation of heavies and lights,
4) product recovery and dewatering, and
5) medium recovery and recycling.

These stages are illustrated in a typical dense-medium separation circuit in Figure 1 (see next page).

In operation, the feed must be screened (a) to remove fine ore and slimes before it is fed, with reconstituted medium, to the separatory vessel (b) which separates the sinks (heavies) from the floats (lights). The vessel may be either static or dynamic; generally, the type of vessel chosen does not change the overall design concept of the circuit. Floats and sinks are withdrawn separately and are drained of most of the medium on static or vibratory screens (c) and (d). The medium is either returned directly to the system or cleaned prior to return. Next, the floats and sinks are washed on the vibratory screens to remove the remaining adhering heavy medium. The undersize products from the washing screens, consisting of medium, wash water, and fines, are too dilute and contaminated to be returned directly as medium to the separatory vessel. They are treated individually or together by magnetic separation (g) to separate the magnetic ferrosilicon or magnetite from the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_22

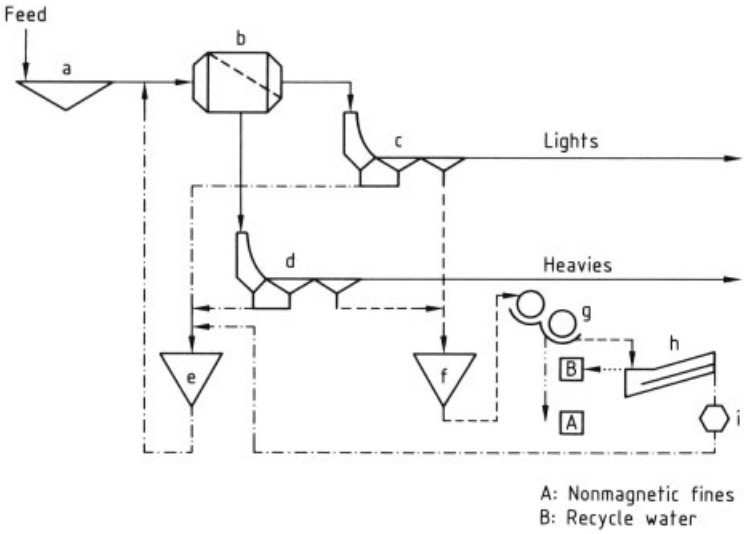

Figure 1. Typical dense-medium separation circuit
a) Feed preparation screen; b) Heavy-medium separation; c) Lights drain – wash screen; d) Heavies drain – wash screen; e) Heavy-medium sump; f) Dilute heavy-medium sump; g) Magnetic separator; h) Densifier; i) Demagnetizing coil
—— Ore, heavies, lights; – · – · – Heavy medium; · · · · · water

nonmagnetic fines. Reclaimed, cleaned medium is thickened to the required density by a suitable densifier (h), and continuously returned to the DMS circuit. The densified medium discharge passes through a demagnetizing coil (i) to ensure a nonflocculated, uniform suspension in the separatory vessel [2].

One of the essential components of the heavy-medium separation process is the medium itself (see Chap. 4). The correct choice of medium and its effective control, in terms of both consistency and physical parameters, are essential for efficient operation of the system.

3. The Separatory Vessel

Separatory vessels can be either static or dynamic. In *static vessels*, separation (generally of particles coarser than 2 mm) is carried out at normal gravity, whereas in *dynamic vessels*, finer particles (generally limited to the range of 20 – 0.5 mm) are separated at elevated gravitational force.

Static DMS Vessels. Static vessels can be subdivided into cone, drum, trough, and combination types (Fig. 2). The feed is generally introduced at or near the top of the separatory vessel. Lights float on the surface and are removed over a weir, with or without the assistance of paddles. Removal of the sinks varies with type of vessel.

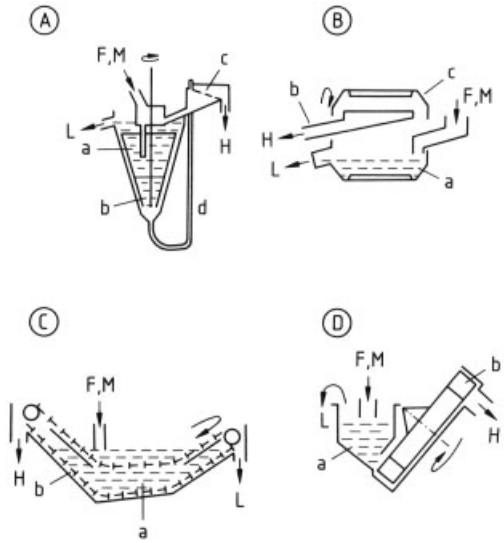

Figure 2. Types of static separatory vessels
A) Cone: a) Bath; b) Paddles; c) Screen; d) Airlift
B) Drum: a) Bath; b) Launder; c) Drum
C) Trough: a) Bath; b) Chain
D) Combination (Drewboy): a) Bath; b) Elevator
M = Medium; F = Feed; H = Heavies; L = Lights

Cone separators (Fig. 2A) contain the greatest depth of medium, which should be slightly agitated to minimize the density gradient from top to bottom of the vessel. Sinks are removed from the bottom by an internal air lift, external pump, or bucket elevator, with drained medium being recycled directly to the vessel. Cone separator circuits tend to be exceptionally stable because of the relatively large quantity of medium employed; they are especially suitable for separation of materials having similar densities.

Drum separators (Fig. 2B) consist essentially of a rotating cylindrical drum fitted with internal lifters, which elevate the heavies out of the medium bath into the sinks discharge launder. The lifters also cause some agitation of the medium, overcoming any tendency for it to settle. Drum separators are common in both coal preparation and mineral separation.

Trough separators (Fig. 2C) consist of a shallow bath or trough for the medium and a moving product removal system, which typically is a single-chain conveyor that moves the two products to opposite ends of the trough. Trough separators are used mainly in North American coal preparation plants.

Combination separating vessels (Fig. 2D), such as the Drewboy and Norwalt, are classified as shallow-bath vessels. Unlike other static vessels, the heavies removal system is not located within the medium bath itself but, like the deep cones, is a separate part of the system. These vessels, therefore, combine the advantages of the shallow bath (large heavies discharge capacity, low medium requirement, and low density gradient) with the advantages of the deep cone (division of the functions of product separation and removal). Therefore, these separators are among the most efficient available.

Dynamic DMS Vessels. Dynamic dense-medium separation was first studied by Driessen at the Dutch State Mines Laboratories with a standard hydrocyclone [3]. The higher forces involved in centrifugal separation permit much finer particles to be separated successfully. The high shear within the vessel enables finer medium, essential for stability, to be used without resulting in excessive viscosity. In addition to the cyclone, a variety of modified cyclone vessels have been developed (Fig. 3).

The smallest particle that can be effectively separated by dynamic dense-medium systems is rarely related to the separatory vessel used; rather, it is generally related to the medium recovery system. The accepted smallest size for efficient separation is 0.5 mm.

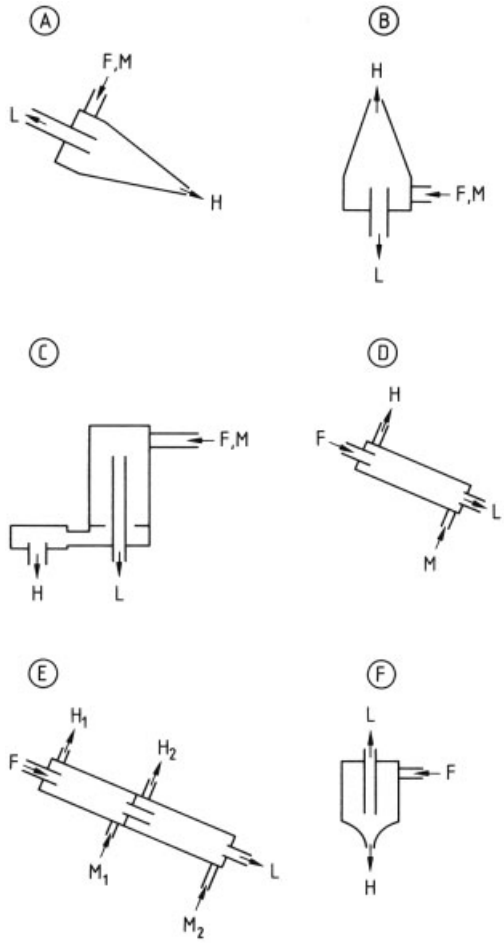

Figure 3. Types of dynamic dense-medium separating vessels
A) Cyclone; B) Swirl; C) Vorsyl; D) Dyna whirlpool; E) Triflo; F) Water-only cyclone
M = Medium; F = Feed; H = Heavies; L = Lights

The principle of operation of the dense-medium cyclone is similar to that of the conventional cyclone (→ Dust Separation , Chap. 3.) except that separation is based on particle density rather than size. Only a limited size classification effect occurs, on large or small particles

or those with a relative density close to the separating density. A typical cyclone separator is operated on an axis that is close to horizontal (Fig. 3A), whereas the Japanese swirl cyclone (Fig. 3B) operates upside down. The Vorsyl separator (Fig. 3C) is a modified cyclone with a cylindrical body; heavies leave the body via an annular throat to a secondary densifying cylinder.

In the separators shown in Figure 3A – C, medium and solids enter the separator together at high pressure (210 – 275 kPa) through the feed port; this results in high wear of feed pump and piping. In the Dyna whirlpool separator (Fig. 3D) and the Triflo separator (Fig. 3E) however, medium is pumped only into the lower tangential inlet; the feed enters by gravity at lower pressure through an axial pipe. Heavies pass through the cycloning medium and leave via the tangential sinks pipe. Lights remain on the axis of the vessel and exit through the axial discharge pipe. The Dyna Whirlpool is a single-stage unit. The Triflo separator has two stages; it can be set to produce two sink products at the same density of separation or at two different densities.

The water-only cyclone (Fig. 3F) can be regarded as an autogenous dense-medium cyclone, in that fines in the ore recirculate within the cyclone to form the medium. These cyclones are employed only in the coal preparation industry, and usually in series, because they are relatively inefficient.

4. The Medium

Although the correct choice of separatory vessel is obviously important in terms of capital and operating costs, as well as efficiency of separation, the control and efficiency of the operation are, to a large extent, dependent on the medium.

Ideally, the medium should be a true heavy liquid. However, although such liquids exist and some are used for laboratory simulations, their toxicity and cost preclude them from commercial application.

Practically, a commercial medium should

1) be available in quantity at or near the site,
2) be comparatively inexpensive at the site,
3) be noncorrosive,
4) be nontoxic,
5) form stable mixtures in water,
6) be physically and chemically stable
7) mix readily with water,
8) rinse from the product with high-pressure water sprays,
9) be capable of regeneration, and
10) be adjustable over a range of densities [4].

Dense media used for commercial separation are almost exclusively suspensions of fine insoluble particles in water. The most commonly utilized solids are *magnetite* [*1309-38-2*] ($d = 5.0 – 5.2$)and *ferrosilicon* [*8049-17-0*] ($d = 6.7 – 6.9$). Magnetite is most commonly used in coal preparation; ferrosilicon is generally chosen for preconcentration of metalliferous ores, although to reduce costs it is often mixed with magnetite. Magnetite and ferrosilicon are preferred because they can be regenerated almost completely in the medium recovery circuit by use of low-intensity magnetic separators as both recovery and densification units.

Ideally, the medium should have low yield stress, low viscosity, and high stability. Unfortunately, however, stable suspensions tend to have high yield stress and high viscosity, and vice versa. The classic work on the rheology of dense media was carried out by DeVaney and Shelton [5]. These workers showed that, for any suspension, *viscosity* (and hence apparent density) increases with increasing solids content. Above a critical value, the viscosity increases rapidly to a limiting point at a volumetric concentration of ca. 40 % solids for all suspensions (Fig. 4). Conversely, the *settling rate*, which is the reciprocal of stability, decreases rapidly with increasing volumetric solids content for equisized particles (Fig. 5, see next page).

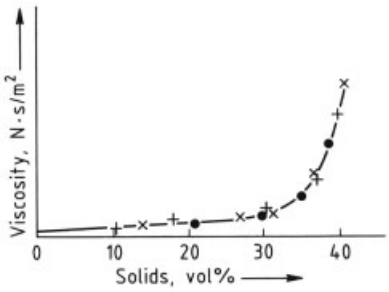

Figure 4. Effect of volumetric solids content on viscosity of three media
• Quartz; × Magnetite; + Ferrosilicon

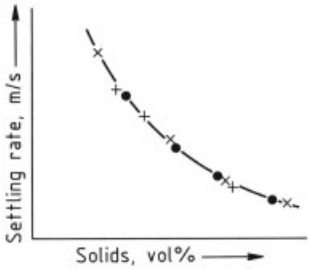

Figure 5. Effect of volumetric solids content on settling rate of three media
• Quartz; × Magnetite; + Ferrosilicon

The coarser the *particle size* of the medium, the faster the solids settle, producing a less stable slurry but a lower viscosity effect at higher concentration. Conversely, the finer the medium particle size, the more stable is the suspension, although the viscosity effect is higher at higher slurry concentrations.

A more stable, and consequently finer, medium is required in dynamic separation because the higher forces involved increase the tendency of the medium to segregate.

The more difficult the separation, the greater is the necessity to have a stable, low-viscosity medium at the operating density and to minimize variation in the medium density during the process.

Magnetohydrostatic separation , in which the apparent density of a magnetic fluid is controlled by a non-uniform magnetic field, is a potential extension of dense-medium separation. It has been studied in the Soviet Union, Israel, South Africa, Great Britain, and the United States.

5. Testing for Dense-Medium Separation

Complete separation, even of particles very close to the separating density, can be achieved in the laboratory by using true heavy liquids and allowing sufficient time for such separation. The most common heavy liquids used are carbon tetrachloride [*56-23-5*] ($\varrho = 1590\,kg/m^3$), 1,1,2,2-tetrabromoethane [*25167-20-8*] ($\varrho = 2970\,kg/m^3$), and thallium malonateformate (Clerici's solution) ($\varrho = 4200\,kg/m^3$). All can be diluted to the desired density, the first two with white spirits or acetone, the last with water. Both cost and toxicity increase with increasing density.

Sequential heavy-liquid separation, in which successive separations occur in a series of liquids of increasing (or decreasing) density, generates either washability curves or partition curves.

A *washability curve* is used to determine the response of an ore to gravity separation. A typical set of curves is shown in Figure 6. These curves allow prediction of the mass and assay of the sink and float; the anticipated recovery at any predetermined density of separation; and the ease or difficulty of the separation proposed. Alternatively, the curves can be used to determine the density of separation required to achieve a desired weight split, product assay, recovery, etc.

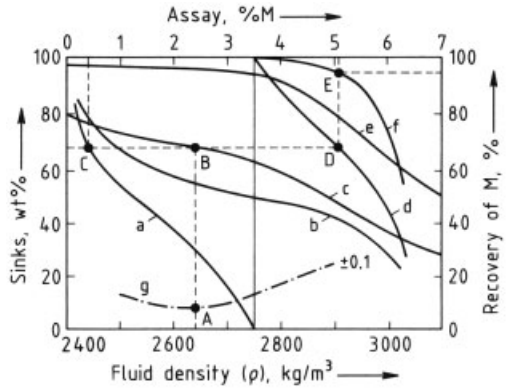

Figure 6. Typical set of washability curves
a) Cumulative assay of floats; b) Characteristic assay curve; c) Density curve – the weight percent of material reporting to the sinks at the density of separation; d) Cumulative assay of sinks; e) Recovery: density curve – the percent of heavy mineral M reporting to the sinks (i.e., recovery) at the density of separation; f) Grade: recovery curve – the recovery to sinks of heavy mineral M, at a specific assay of the sinks (%M); g) Tolerance curve – the weight percent of material within a specified density range at the density of separation

In the example shown, preconcentration of an industrial mineral (M) is to be carried out at a density of $2640\,kg/m^3$, equivalent to the minimum (point A) on the tolerance curve (g), thereby assuring maximum efficiency of a reasonably difficult separation. At this density, on the line AB, the weight split is 68 % to sinks, equivalent to a floats assay of 0.3 % (point C) and a sinks assay of 5.1 % (point D). Line DE then shows that recovery of M is 93 % (point E).

Partition curves are used to determine the efficiency of a particular separation process [6].

They are constructed by carrying out sequential heavy-liquid analyses of both products from the separation, and then plotting at each density the proportion of the calculated feed that reports to the sinks or floats. A typical partition curve is shown in Figure 7.

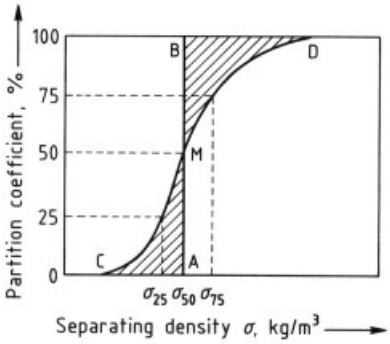

Figure 7. Partition curve

The separating density, σ_{50}, is the density at which a particle has an equal chance of sinking or floating or, conversely, at which it has an equal chance of being misplaced. This is equivalent to the 50 % point, M, on the partition curve. A perfect separation is represented by the vertical line AMB, i.e., all lighter material floats and all heavier material sinks. The actual separation is represented by the curve CMD. The difference between this curve and the vertical represents the amount of error of separation; hence, the total shaded area is the *error area*.

The most common guide to efficiency of separation is the *probable error* E_{p}. This is defined as half the differential in density between the 25 and 75 % partition coefficients on the partition curve:

$$E_{\mathrm{p}} = 0.5\,(\sigma_{75} - \sigma_{25})$$

When plotted on probability paper, the partition curve can approximate a straight line. However, significant "tails" occur at the top and bottom of the line. Modern computer-modeling techniques permit the shape of the generated curve to be described mathematically, although the model most likely to fit the data is still in question [7]. The partition curve, error area, probable error, the computer-generated curve can be used not only to monitor plant performance and acceptance testing, but also to design and predict dense-medium circuit operation. By starting with washability data generated as described earlier and applying the partition coefficients, related to E_{p} (which are known for a specific separator), the partition curve, and hence the characteristics of both products from the separator, can be predicted. A typical predicted separation of coal is shown in Table 1. In this example, at a cut point of 1600 kg/m^3 coal can be obtained with 5.59 % ash and a yield of >70 %.

Dense-medium separation processes have been modeled for many years by means of washability data, because separation is almost entirely dependent on the relative density of particles and medium. Recently, considerable developments have been made in the use of washability data to model most other types of gravity concentration devices; hence, the importance of heavy-liquid analysis in industry is growing.

Table 1. Anticipated separator results in treating any coal of size range 1.5 to 0.5 cm in the same DMS unit at density 1600 kg/m^3

Liquid density, kg/m^3	Heavy-liquid test on head sample of a raw coal			Partition coefficient at assumed cut point $d = 1.6$	Calculated sink–float analysis for separator					
					Clean coal (floats)			Rejects (sinks)		
	wt %	% ash	ash units		wt %	% ash	ash units	wt %	% ash	ash units
Floats 1300	52.8	2.6	1.373	100.0	52.80	2.6	1.373		2.6	
1300 – 1400	12.3	8.7	1.070	96.5	11.87	8.7	1.033	0.43	8.7	0.037
1400 – 1500	4.2	18.8	0.790	84.5	3.55	18.8	0.667	0.65	18.8	0.122
1500 – 1600	2.1	30.0	0.630	63.0	1.32	30.0	0.396	0.78	30.0	0.234
1600 – 1700	2.9	39.4	1.143	34.5	1.00	39.4	0.394	1.90	39.4	0.749
1700 – 1800	1.0	47.1	0.471	13.3	0.13	47.1	0.061	0.87	47.1	0.410
1800 – 1900	1.1	53.6	0.590	4.0	0.04	53.6	0.021	1.06	53.6	0.568
1900 – 2000	1.0	60.2	0.602	0.5	0.01	60.2	0.006	0.99	60.2	0.596
Sinks 2000	22.6	85.1	19.233			85.1		22.60	85.1	19.233
Total	100.0	25.9	25.902		70.72	5.59	3.951	29.28	75.0	21.949

Within a few years, sufficient data will likely be available for design of complete gravity concentration circuits by computer.

Acknowledgment. Material from this section is drawn mainly from the book **Gravity Concentration Technology**, written by the author (assisted by C. Mills) and published by Elsevier Science Publisher BV, Amsterdam, who retain the copyright. Their permission to utilize this material is duly acknowledged.

6. References

1. R. O. Burt, C. Mills: *Gravity Concentration Technology,* Elsevier, Amsterdam 1984, pp. 139 – 183.
2. R. J. Gochin, M. R. Smith, *Min. Mag.* (1983, Dec.) 453 – 460.
3. M. G. Driessen, *J. Inst. Fuel* **19** (1945) 33 – 45.
4. C. G. Wilson, Mineral Processing Plant Design, AIME 1978, pp. 520 – 538.
5. F. D. DeVaney, S. M. Shelton U.S. Bureau of Mines (USBM), Report of Investigations 3469R, 1940.
6. K. F. Tromp, *Colliery Guardian* **154** (1937) 995 – 999.
7. K. J. Reid, Lu Maixi, Zhang Shenggui, *Int. J. Miner. Process.* **14** (1985) 291 – 299.

Flotation

See also → Solid – Solid Separation, Introduction

Baki Yarar, Colorado School of Mines, Golden, Colorado 80401, United States

1. Introduction 2
1.1. Interfacial Energies and Contact Angle 2
1.2. Hydrophobic and Hydrophilic Solids 3
2. Flotation Reagents 3
2.1. Functional Classification of Flotation Reagents 3
2.2. Chemical Classification 4
2.3. Surface and Solution Chemistry . . 7
2.3.1. Solubility 7
2.3.2. Surface Activity 8
2.4. Interactions between Flotation Reagents and Solids 8
2.4.1. Adsorption – Desorption of Inorganic Reagents 10
2.4.2. Adsorption – Desorption of Organic Reagents 10
3. Electrical Phenomena 11
4. Bubble – Particle Interactions 14
5. Flotation Processes 15
5.1. The Gamma Flotation Process . . . 16
5.2. Design of Flotation Operations . . . 17
5.3. Flotation Machines 18
5.4. Industrial Applications 20
5.4.1. Sulfide Mineral Flotation 21
5.4.2. Nonsulfide Flotation 24
5.4.3. Coal Flotation 25
5.4.4. Flotation of Precipitates and Ions . . . 26
5.4.5. Use of Oil Droplets instead of Air Bubbles 27
6. Future Trends 27
7. References 28

Symbols

C = equilibrium concentration of surfactant in solution
E = electric field between electrodes applying the field
k = Boltzmann constant
p = pressure inside a bubble
p_o = pressure of the bulk of the liquid
P_a = probability of particle-bubble adhesion
P_c = probability of particle-bubble collision
P_f = probability of flotation
P_s = probability of formation of a stable particle-bubble aggregate
S = spreading coefficient for air on solid
T = temperature
V_e = electrophoretic velocity of particle
W = work of adhesion
Γ = surface excess concentration of surfactant measured in quantity per unit area (e.g., molecules/cm^2)
γ = specific surface energy of liquid film
γ_0 = specific surface energy of an infinitely thick film of a liquid
γ_{LG} = liquid – gas interfacial tension
γ_{SG} = solid – gas interfacial tension
γ_{SL} = solid – liquid interfacial tension
ε = dielectric constant
ζ = zeta potential
η = viscosity of aqueous medium
θ = contact angle
π_D = disjoining pressure

1. Introduction

General Aspects. Flotation is one of the well-established physicochemical separation processes, widely applied in mineral concentration and recovery operations; other applications include water treatment and purification, the recycling of secondary materials, and the recovery of ionic and colloidal materials from aqueous solutions.

Various forms of this process are known as froth flotation, ion flotation, precipitate flotation, piggyback flotation, flotoflocculation, skin flotation, column flotation, and two-liquid flotation [1–8]. Gamma flotation is based on the control of liquid – vapor interfacial tension to facilitate the particle – bubble adhesion process [9].

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_23

Because most flotation processes are applied with water as the principal liquid, and the widest area of applications is in the field of mineral processing, mostly systems containing water and minerals are discussed here.

1.1. Interfacial Energies and Contact Angle

The flotation process relies primarily on the fact that hydrophilic particles are wetted by water, whereas hydrophobic particles are wetted by oils and air bubbles; therefore, if air bubbles are introduced into an aqueous slurry, the bubbles adhere to the hydrophobic solid particles. As a result, air – solid aggregates are carried to the surface, forming a froth layer; this explains the name froth flotation. The froth layer can be removed manually or mechanically; the result is the separation of hydrophobic from hydrophilic particles. For further information on detergency, see also → Laundry Detergents, Chap. 2.1.

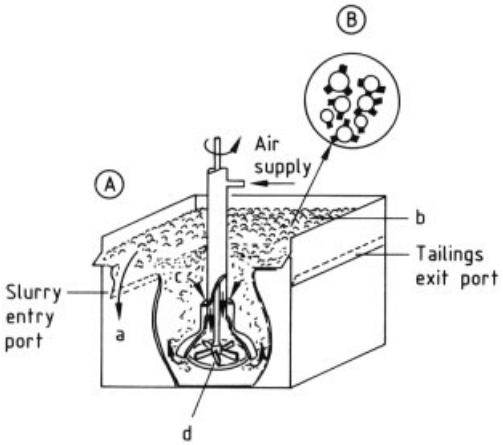

Figure 1. Processes occurring in a flotation cell (Schematic)
A) Flotation cell a) Froth overflow; b) Froth layer; c) Pulp; d) Rotor for pulp agitation
B) Mineralized air bubbles within flotation cell

The principal processes occurring in a slurry-containing vessel known as a *flotation cell* are shown in Figure 1. Equation (1), known as the Young–Dupré equation, and Equations (2) and (3) are used to describe the relationship of physicochemical variables to particle – bubble adhesion.

$$\gamma_{SG} - \gamma_{SL} = \gamma_{LG} \cos\theta \quad (1)$$

$$W = \gamma_{LG}\ (1 - \cos\theta) \quad (2)$$

$$S = \gamma_{SL} - (\gamma_{LG} + \gamma_{SG}) \quad (3)$$

where γ_{SG} = solid – gas interfacial tension, γ_{SL} = solid – liquid interfacial tension, γ_{LG} = liquid – gas interfacial tension, θ = contact angle, S = spreading coefficient for air on solid, and W = work of adhesion. For particle – bubble adhesion the conditions $W > 0$, $S > 0$, and $\theta > 0$ are satisfied.

Figure 2 describes the mechanical equilibrium among the three interfacial tensions indicated as vectors applied to point C, known as the three-phase contact point; the contact angle θ is measured in the liquid phase. The two variables $\cos\theta$ and γ_{LG} are readily accessible by experiment. Contact angles on minerals in various media are given in Table 1. The significance of the contact angle arises from the fact that it is a measure of surface wettability; $\theta > 0$ indicates a hydrophobic solid. Contact angle values of $\theta > 10°$ usually indicate that particles can form bubble – particle contacts strong enough to resist turbulence in a conventional flotation cell, enhancing the probability of flotation as decribed in Chapter 4.

Table 1. Contact angles of solids in water and various aqueous media

Solid	Contact angle (degrees)	Solution conditions
Colemanite	43	5×10^{-3} M oleate ∗ solution
Copper metal	93	in 50 mg/L oleate ∗ solution
Fluorite	91	in 10^{-5} M oleate ∗, pH = 8.1
Galena	60	in 10^{-3} M ethyl xanthate solution
Graphite	96	water
Ilmenite	80	treated with oleate ∗, pH = 8, T = 75 °C
Colorado oil shale with 28 % organic carbon	59.5	water
Paraffin wax	108	water
Silica	81	2.5 mg dodecylammonium chloride per L solution, pH = 10
Teflon	160	water
Teflon	0	methanol – water solution with surface tension <20 mN/m

∗ Sodium oleate.

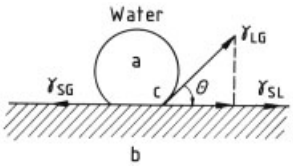

Figure 2. Equilibrium contact angle
a) Captive bubble in water; b) Hydrophobic solid
$\mathrm{Cos}\theta = \frac{\gamma_{SG} - \gamma_{SL}}{\gamma_{LG}}$

1.2. Hydrophobic and Hydrophilic Solids

Although most naturally occurring minerals are hydrophilic, some are hydrophobic. Examples of the latter include graphite, sulfur, antimonite (Sb_2S_3), molybdenite (MoS_2), talc, and high-rank coals such as anthracite. Many polymers, such as Teflon and Nylon, are also hydrophobic.

Hydrophilic materials can be made hydrophobic by the adsorption of chemicals. For example, calcite ($CaCO_3$) can be readily made hydrophobic by treating it with low concentrations of sodium oleate ($C_{17}H_{33}COONa$). Silica (SiO_2) can be made hydrophobic by treatment with dodecylamine ($C_{12}H_{25}NH_2$). A solid made hydrophobic by chemical treatment can revert to the hydrophilic state on further chemical change. For example, galena (PbS), in contact with 25 mg/L potassium ethyl xanthate (potassium *O*-ethyldithiocarbonate) [*140-89-6*] (C_2H_5OCSSK) solution, is hydrophobic at pH below 10.8, but hydrophilic at higher pH. Teflon is highly hydrophobic in most aqueous media, but behaves like a hydrophilic solid if the surface tension is reduced below 20 mN/m by the addition of methanol [10]. Chemicals used to affect the hydrophobic – hydrophilic properties of solids are known as flotation reagents.

2. Flotation Reagents

2.1. Functional Classification of Flotation Reagents

Collectors. These are used primarily to make solids hydrophobic and promote adhesion to air bubbles or oil droplets. Common examples are fatty acids, sulfonates, xanthates (dithiocarbonates), amines, and dithiophosphates.

Frothers. Frothers promote the formation of a metastable froth phase that facilitates the removal of particles carried by air bubbles to the top of the flotation cell. They also reduce the induction time; that is, they allow colliding particles and bubbles to establish contact more rapidly. The usual induction times in sulfide flotations are of the order of milli- or microseconds [12]. Examples of frothers are pine oil, long-chain alcohols, and polyoxypropylene derivatives.

Auxiliary Reagents. These reagents include *depressants,* which are used to prevent solids from becoming hydrophobic, and *activators*, which promote the adsorption of reagents onto selected solids. For example, high pH impedes the flotation of sulfide minerals by xanthates. Thus, reagents used to increase pH, e.g., NaOH or $Ca(OH)_2$, would be depressants in this context. Sulfide ions similarly impede the flotation of sulfides in the same system, and Na_2S used as a source of sulfide ions would act as a depressant. Other depressants include starch and quebracho.

On the other hand, sphalerite (ZnS) does not readily become hydrophobic in the presence of xanthates unless copper (II) ions are introduced into the aqueous pulp. In this case $CuSO_4$, used to supply copper (II) ions, is an activator. Other examples of activation are provided by divalent ions such as barium or calcium in the flotation of silica by fatty acids. In this case, silica is inert to these reagents unless a divalent cation is introduced into the pulp at alkaline pH. Thus, $CaCl_2$ or $Ba(NO_3)_2$ used as sources of these divalent cations and NaOH used to raise the pH would be activators of silica for flotation by fatty acids. The terms *collector* and *frother* do not precisely describe the action of such reagents in all systems.

Although oleic acid acts primarily as a collector in the flotation of solids such as scheelite ($CaWO_4$) it also exhibits excellent frothing properties when present in sufficient quantities. On the other hand, nonanol [*143-08-8*], a well-known frother, is the sole reagent required for the flotation of talc or high-rank coals, where it acts simultaneously as frother and collector. The same is true for activators and depressants. For example, sodium sulfide was cited previ-

ously as a depressant for sulfide minerals in the presence of xanthate reagents. In fact, heavily oxidized (tarnished) sulfide minerals are treated with sodium sulfide before the introduction of xanthates for flotation, making Na_2S an activator in this type of flotation pulp (see Fig. 3).

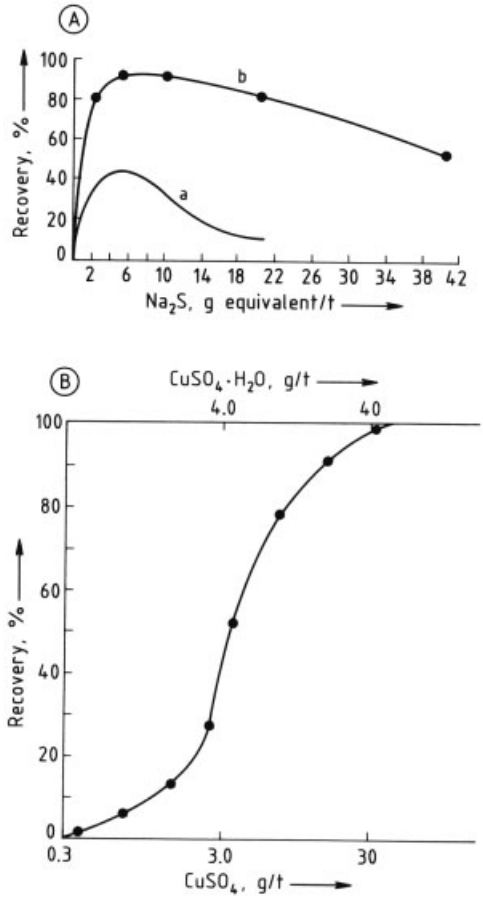

Figure 3. Activation Processes
A) Activation of malachite [$CuCO_3 \cdot Cu(OH)_2$] by sodium sulfide a) In the presence of 0.43 mol/L ethyl xanthate; b) In the presence of 0.43 mol/L isoamyl xanthate [a, b]
B) Activation of ZnS by $CuSO_4$ in the presence of 25 mg/L potassium ethyl xanthate [c]
[a] With permission of [7]. [b] Excessive quantities of sulfide reduce the recovery. [c] Redrawn from [11].

2.2. Chemical Classification

Flotation reagents may be inorganic or organic. The former are largely used as auxiliary reagents (Table 2), whereas the latter (Tables 3 and 4) fall into all categories of the functional classification. The organic reagents can be classified as follows:

Polar:
- anionic: polar group is an anion, e.g., oleate, $CH_3(CH_2)_7CH{=}CH(CH_2)_7COO^-$
- cationic: polar group is a cation, e.g., RNH_3^+
- amphoteric: anionic or cationic, depending on pH, e.g., alkylaminocarboxylic acids, $RNH(CH_2)_xCOOH$
- nonionic: polar group is not ionized, e.g.,*n*-hexanol, $CH_3(CH_2)_4CH_2OH$

Nonpolar: e.g., saturated hydrocarbons

Table 2. Inorganic auxiliary flotation reagents

Compound	Composition	Common Applications
Lime	CaO	pH regulator, depressant
Sodium carbonate (soda ash)	Na_2CO_3	pH regulator, dispersant
Sodium hydroxide (caustic soda)	$NaOH$	pH regulator, dispersant
Sodium sulfide	Na_2S	sulfide depressant and ore sulfidizer
Sodium bisulfide	$NaHS$	sulfide depressant and ore sulfidizer
Sulfuric acid	H_2SO_4	pH regulator
Sodium cyanide	$NaCN$	sulfide depressant
Calcium cyanide	$Ca(CN)_2$	sulfide depressant
Sodium dichromate	$Na_2Cr_2O_7$	PbS depressant
Cupric sulfate	$CuSO_4$	ZnS, FeAsS, Sb_2S_3 activator
Lead acetate	$Pb(CH_3COO)_2$	Sb_2S_3 activator
Sodium ferrocyanide	$Na_4Fe(CN)_6$	depressant in Cu – Mo sulfide circuits
Potassium permanganate	$KMnO_4$	FeS_2 depressant in FeAsS flotation
Sulfur dioxide	SO_2	activated ZnS depressant
Sodium thiosulfate	$Na_2S_2O_3$	SO_2 source in acid circuits
Sodium silicate	Na_2SiO_3	siliceous gangue dispersant
Sodium fluosilicate	Na_2SiF_6	depressant in iron-flotation circuits
Sodium polyphosphates	e.g., $(NaPO_3)_6$	dispersant
Sodium fluoride	NaF	activator in silicate flotation
Nokes reagent	complex mixture of P_2S_5, As_2O_3, Sb_2O_3, NaOH, etc.	general depressant in molybdenite flotation circuits except for MoS_2

Organic flotation reagents are also classified by functional group, including fatty acids, carboxylic acids, thiols, xanthates, sulfates, sulfonates, and amines.

Table 3. Organic reagents commonly used as flotation collectors

Compound	Active component(s)	Area of application
Primary amine salts	$RNH_3^+Cl^-$	silica, silicates, sylvite
Quaternary ammonium salts	$RN(CH_3)_3^+Cl^-$ $R = C_{10}-C_{16}$	silicates, oxides, clays
p-Tolylarsonic acid	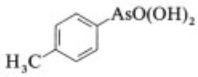	cassiterite
Sodium salts of carboxylic acids[a]	RCOONa	oxides, carbonates, apatite, iron ores, chromite, scheelite
Alkyldithiocarbamates	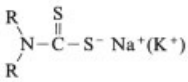	sulfides, metallic minerals
Dixanthogens	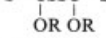	sulfides, metallic minerals
Hydrocarbon oils[b]	C_nH_{2n+2}	coal, molybdenite, colemanite with sulfonates
Sodium alkylhydroxamates	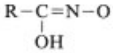	iron ores, wolframite, cassiterite
Naphthenic acids	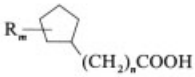	fluorapatite, colemanite
Oximes	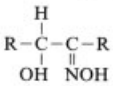	chrysocolla, cassiterite
Alkylsulfates and -sulfonates	$R-O-SO_3^-\ M^+$ $R-SO_3^-\ M^+$	iron ores, beach sand cleaning, borates, carbonates, fluorite
Sodium 2-(Methyloleylamino) ethylsulfonate	$C_{17}H_{33}-N(CH_3)-CH_2CH_2SO_3Na$	celestite
O-Ethyl isopropyl thionocarbamate	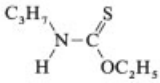	copper sulfides
Thionocarbanilide	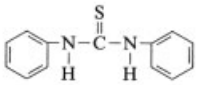	sulfide minerals
Alkyldithiocarbonates (xanthates)	$R-O-C(=S)-S^-Na^+(K^+)$	sulfide minerals, gold
Xanthogen formates[c]	$R-O-C(=S)-S-C(=O)-OR$	sulfide minerals
Dialkyl-dithiophosphates[d]	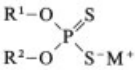	sulfide minerals, native gold, copper

[a] R may be saturated or unsaturated. [b] Vapor oils, kerosene, fuel oils. [c] Trade name: Minerec. [d] Trade name: Aerofloat.

Table 4. Organic chemicals used in flotation as depressants, slime depressants, and flocculants

Common name	Active components	Application
Sodium isopropyl-naphthalene sulfonate Trade name: Aerosol OS	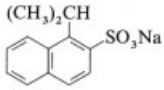	wetting agent, defoamer, dispersant, emulsifier
Sodium dioctylsulfo-succinate Trade name: Aerosol OT	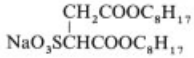	wetting agent, defoamer, dispersant, emulsifier
Poly(ethylene oxide)		flocculant, dewatering aid
Sodium polyacrylate	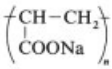	flocculant, dewatering aid
Starch	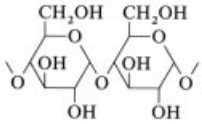	slime depressant, iron ore flocculant, slime control agent
Tannic acid	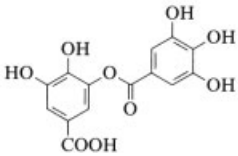	depressant for fluorite, carbonates, and non-sulfides
Quebracho (trade name) extracted from shinopsis trees	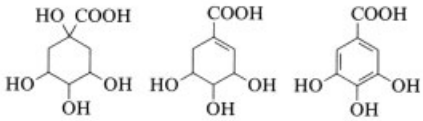 Quinic acid [77-95-2] Shikimic acid [138-59-0] Gallic acid [149-91-7]	depressant for carbonates, fluorite, and nonsulfides
	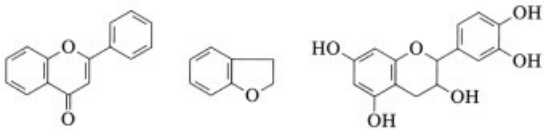 Flavone [525-82-6] Coumaran [496-16-2] Catechin [154-23-4]	

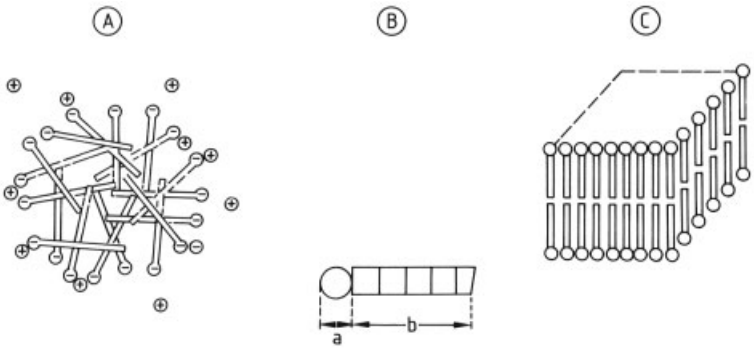

Figure 4. Micelle formation
A) Spherical micelle B) Straight-chain surfactant (amphipatic) molecule
a) Polar end; b) Nonpolar end C) Laminar micelle

The following organic reagents are commonly used as flotation frothers:

Long-chain aliphatic alcohols $CH_3(CH_2)_nCH_2OH$, $n = 3–6$

Camphor oil
Active component:
Safrole [*94-59-7*]

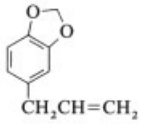

Dimethyl phthalate [*131-11-3*]

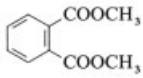

Dimethylphenylcarbinol
(2-Phenyl-2-propanol) [*617-94-7*]

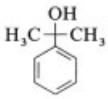

Eucalyptus oil
Active component:
cineole [*470-82-6*]

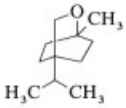

Methylisobutylcarbinol (MIBC)
(4-Methyl-2-propanol) [*108-11-2*]

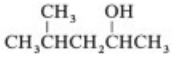

Poly(ethylene glycol)
[*25322-68-3*]
$R(OC_2H_4)_nOC_2H_4OH$, $n = 2–5$

Poly(propylene glycol)
[*25322-69-4*]
$H(OC_3H_6)_nOH$, $n = 2–5$

Poly(propylene glycol),
monoalkylether
Trade name: Dowfroth 250
[*37286-64-9*]
$R(OC_3H_6)_nOH$, $n = 2–5$,
$R = CH_3, C_4H_9$

Triethoxybutane
(2,2,3-Triethoxybutane)
[*38620-87-0*]

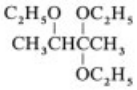

2.3. Surface and Solution Chemistry

2.3.1. Solubility

Inorganic reagents used in flotation are always highly soluble in water whereas organic reagents may not be. Nonpolar oils, for example, are insoluble in water and must be added to the flotation cell or the preceding conditioning tank, at high shear created by intense agitation of the pulp. Alternatively, these reagents can be prepared as emulsions or sometimes dissolved in an alcohol such as methanol before addition to the flotation system. Straight-chain alcohols containing more than eight carbon atoms are insoluble in water and are used as frothers.

Most ionic or polar flotation reagents are sufficiently soluble in water. Solubility is limited mainly by the tendency of such compounds to form molecular aggregates in the aqueous medium known as micelles when present in high concentrations (Fig. 4, see previous page). The concentration at which micelles form is known as the *critical micelle concentration* (cmc).

2.3.2. Surface Activity

The ability of reagents to alter the liquid – vapor interfacial tension, γ_{LG}, of aqueous solutions is called surface activity (see also → Surfactants). For example, although the surface tension of distilled water is about 72 mN/m at room temperature, the introduction of a flotation reagent such as sodium dodecyl sulfate at a concentration of 0.005 mol/L reduces this value to 48.5 mN/m. Such organic reagents contain two chemical groups: a nonpolar moiety, usually a saturated hydrocarbon, and a polar moiety, which may be ionic (e.g., $-COO^-$, $-SO_3^-$, $-SO_4^-$) or nonionic (e.g., $-OH$, $-NH_2$, $>C=O$).

The quantitative relationship between the concentration of an amphipatic molecule and its effect on the surface tension at concentrations below the cmc is described by the Gibbs adsorption equation:

$$\frac{d\gamma_{LG}}{dC} = -\Gamma k \frac{T}{C} \tag{4}$$

where γ_{LG} = solution surface tension, C = equilibrium concentration of surfactant in solution, k = Boltzmann constant, Γ = surface excess concentration of surfactant measured in quantity per unit area (e.g., molecules/cm^2), and T = temperature.

Based on Equation (4), the variation of solution surface tension can be discerned to be as shown in Figure 5, which predicts that some substances ought to increase surface tension, creating a surface deficiency, as opposed to surface excess concentration. This condition is met by a number of inorganic compounds shown in Figure 6.

Not all amphipatic flotation reagents exhibit pronounced surface activity, as demonstrated by

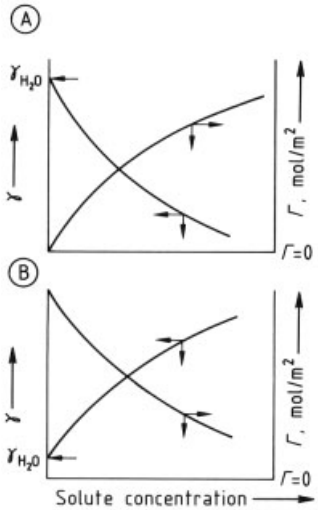

Figure 5. Variation of solution surface tension by solutes according to the Gibbs adsorption equation (Schematic)
A) Effect of organic surfactants on γ_{LG} and Γ. Condition of surface excess concentration
B) Effect of inorganic solutes on γ_{LG} and Γ. Condition for surface deficiency.
γ_{H_2O} = Surface tension of pure water; γ = Surface tension of solution; Γ = Surface excess concentration

Figure 6. Compounds that lower the surface tension of water are not necessarily surface active, although their structures may indicate amphipatic properties. Short-chain alcohols, acetone, and simple ethers fall into this category (see Fig. 7 on top of next page).

2.4. Interactions between Flotation Reagents and Solids

As a general rule, for alteration of the hydrophobic – hydrophilic properties of solids in a flotation pulp, adsorption or desorption at the surface is a prerequisite. Thus, for a solid to become hydrophobic or hydrophilic it must first take up ions or molecules from the medium or release them into it. Alternatively, a solid made hydrophobic by the adsorption of a reagent changes only by the action of other species which alter the surface properties. This may or may not require the desorption of the species originally present. For example, at a given pH, a given sulfide mineral requires a minimal concentration of flotation collector before stable bubble – particle contact (flotation) can occur [8] (Fig. 8, see next page).

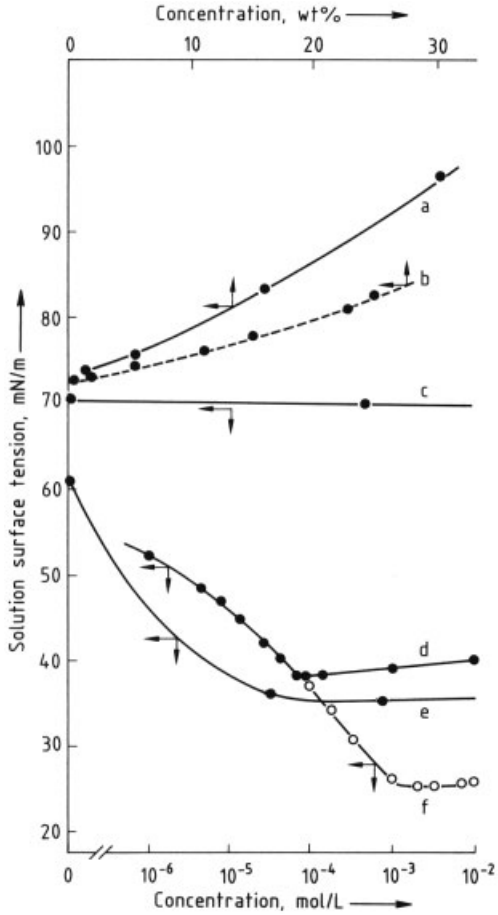

Figure 6. Effects of inorganic salts and organic flotation reagents on the surface tension of water
a) Sodium hydroxide [a]; b) Sodium chloride [a]; c) Potassium ethyl xanthate [b] d) Polyoxyethylene *n*-decanol with 30 ethylene oxide groups [b]; e) Cetyltrimethylammonium bromide [c]; f) Sodium oleate [d];
[a] Drawn from data in [13]. [b] With permission from [14]. [c] Drawn from data in [15]. [d] Redrawn from data in [16].

Silica, treated with trimethylchlorosilane ($Si(CH_3)_3Cl$) shows hydrophobic – hydrophilic transitions that depend on time, pH, and temperature [17]. In all the experimental cases reported for this system, whether the solid was hydrophilic or hydrophobic, the organic reagent was present at the solid surface. The variation of hydrophobic – hydrophilic properties was attributed to the tenacity of the aqueous layer at the solid – liquid interface; the desorption of species that confer hydrophobic characteristics is not always essential for a solid to exhibit hydrophilic properties.

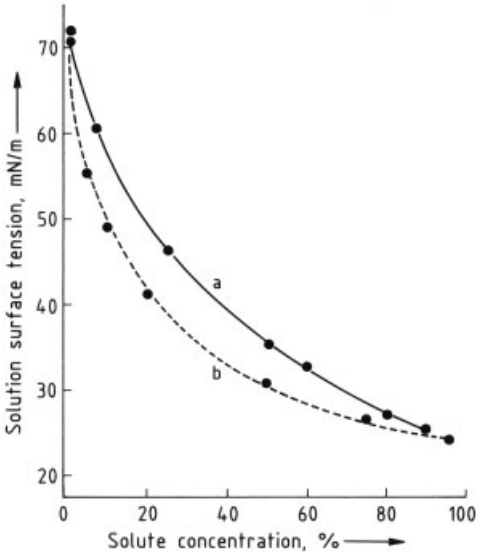

Figure 7. Surface tension of aqueous solutions of acetone and methanol
a) acetone; b) methanol

Naturally hydrophobic solids need to be considered a special case of this general rule, because they may require no surfactant uptake from the medium to become hydrophobic, but can be made hydrophilic by adsorbed reagents. For example, most wetting agents, e.g., Aerosol OT, (sodium dioctylsulfosuccinate) [*577-11-7*], eliminate the natural hydrophobicity of graphite.

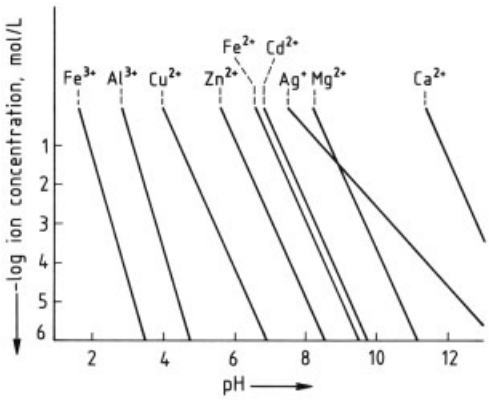

Figure 9. Cation concentrations derived from the dissolution of some metal hydroxides at different pH [21]
∗ Calculations made assuming that the solubility product expression given in Equation (6) applies; more elaborate calculations including other hydrolytic species can also be made [22, 23]

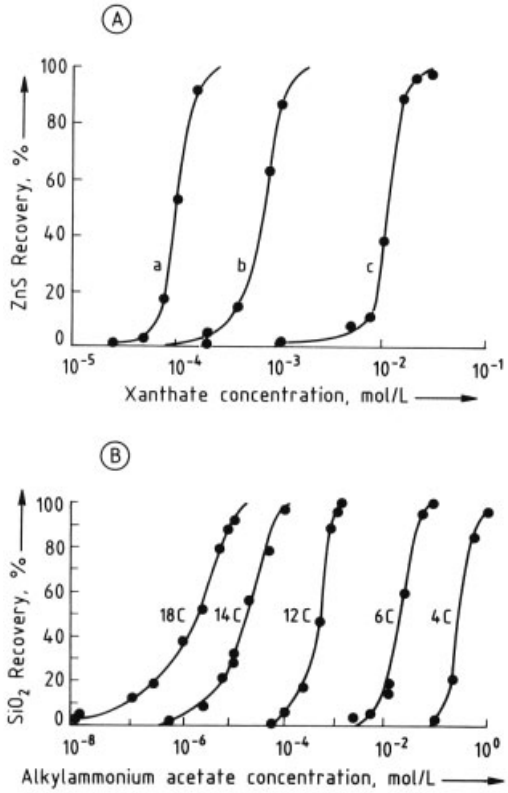

Figure 8. Role of carbon chain length in flotation
A) Sphalerite (ZnS) in the presence of various chain-length xanthates at pH = 3.5 ∗
a) 6 Carbons, hexyl xanthate; b) 4 Carbons, isopropyl xanthate; c) 2 Carbons, ethyl xanthate
B) Quartz (SiO_2) flotation by alkylammonium acetate solutions; chain lengths indicated at their natural pH ∗∗
∗ With permission [18]. ∗∗ With permission [19].

2.4.1. Adsorption – Desorption of Inorganic Reagents

In general, the chemical reactions of inorganic flotation species in aqueous solution can be described in terms of solubilities. The classical equilibrium constant (solubility product) applies:

$MA_2 \rightleftharpoons M^{2+} + 2\,A^-$ (5)

$$K_{sp} = [M^{2+}]\,[A^-]^2 / [MA_2] \qquad (6)$$

where M^{2+} indicates the cation (metal) concentration in solution, A^- the anion concentration, and K_{sp} the solubility product. Figure 9 shows the relationships of hydrolysis and precipitation that occur in the aqueous phase for various ions

in accordance with the solubility product principle. Traces of ions or their hydrolysis products many play important activator roles in flotation [20].

2.4.2. Adsorption – Desorption of Organic Reagents

Organic reagents and solids suspended in a flotation pulp interact by one or more of the mechanisms given below. Some processes, such as hydrogen bonding, are also important in the solubilization of polar substances in aqueous media.

Hydrogen Bond Formation. This bond (strength ca. 21 – 29 kJ/mol), usually associated with the physical properties of ice and water, is favored when a group contains hydrogen attached to a highly electronegative element such as F, O, S, or N. Thus, the adsorption of an amide group, present in proteins or nonionic polyacrylamide (a depressant), on fluorite or silica is due to this bond [24, 25]. The interaction of elemental sulfur and alcohols probably also takes place by hydrogen bond formation.

Electrostatic Interactions. Such interactions are nonspecific, occurring between groups that carry electrostatic charges of opposite sign. Bond formation occurs by charge attraction. A typical example in flotation systems is the adsorption of amines that carry cationic polar groups on silica, which carries a net negative electric charge at $pH > 2$. Many treatments of reagent – solid interactions in flotation systems tend to exaggerate the importance of electrostatic interactions, overlooking other mechanisms.

Chemical Bond Formation. Chemical bonds are characterized by their strength (typically greater than ca. 42 kJ/mol) and usually lead to the formation of saltlike structures at solid – liquid interfaces. A surface compound is not necessarily identical to a compound formed in the bulk of the solution [26, 27]. Examples of such bonds include the interaction of fatty acids, which carry the polar carboxylate group ($-COO^-$), with sparingly soluble solids such as calcite ($CaCO_3$) or fluorite (CaF_2). Infrared spectroscopy shows that the calcium oleate formed by the interaction of oleic acid with calcite or fluorite is not identical to the calcium salt of oleic acid, $Ca(C_{17}H_{33}COO)_2$, obtained from $CaCl_2$ and sodium oleate [28].

Hydrophobic Bonding. Association of the hydrocarbon ends of amphipatic substances leads to the formation of *micelles* (Sect. 2.3.1). Similarly, many saturated hydrocarbons, such as kerosene or fuel oils, are readily adsorbed on naturally hydrophobic solids, e.g., graphite or freshly cleaved bituminous coal. Such interactions are mainly a result of London – van der Waals forces. Furthermore, the hydrocarbon ends of surfactants, such as dodecylbenzenesulfonate, and dodecylbenzenesulfate, associate at the solid – liquid interfacial region, resulting in low-energy structures known as hemimicelles [29].

The strength of association is a function of the hydrocarbon chain length and is of the order of 1 kT per $-CH_2$-group (where k = Boltzmann constant and T = temperature). The unique bond that leads to such adsorptions and molecular associations is called the *hydrophobic bond*. Although the term has been criticized [30], it has been generally accepted.

Multiple Dipole Interaction Bonds (Crystal-field Interactions). The adsorption of tannins (depressants in flotation) or polyacrylamide on fluorite has been attributed to the formation of a bond in which an active group of the reagent is held in the crystal field of the adsorbing solid [31]. Little information on such bonds is available.

Collector Fitting into Solid Lattices. The separation of similar salts, such as NaCl and KCl, by the use of amines is a well-established industrial practice. Although a number of mechanisms have been suggested for the selective interaction between KCl and primary amines, the most widely accepted is that the amine fits into spaces on the KCl lattice vacated by K^+ because the ionic sizes of $-NH_3^+$ and K^+ are comparable. Thus, KCl is made selectively hydrophobic, whereas NaCl is not (Fig. 10).

In many systems more than one mechanism may act simultaneously and synergistically. The

term adsorption is used here for the accumulation of a dissolved species on a solid particle. The term sorption is preferred by some authors. The terms physical adsorption or physisorption and chemisorption are also used to distinguish between bonds with strengths less than 20 kJ/mol and those greater than 42 kJ/mol, respectively.

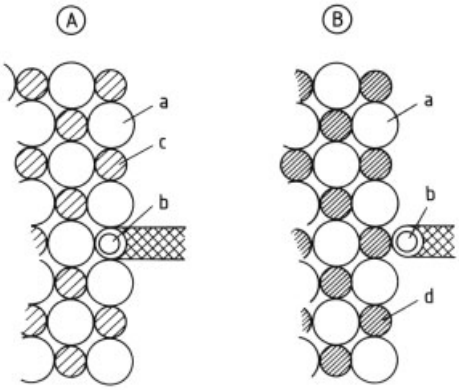

Figure 10. Schematic representation of the selective adsorption of amine ions on sylvite (KCl) [32]
A) Potassium chloride (KCl)
B) Sodium chloride (NaCl)
a) Cl^- b) NH_3^+ ; c) K^+ ; d) Na^+

3. Electrical Phenomena

The main phenomena to be considered in this connection are those associated with the solid – liquid interfacial region. On contact of the solid with the aqueous medium, ionization, ion dissolution, or ion adsorption occurs, resulting in charging of the particle surface. Hydrolysis in the aqueous medium may also alter the character of the ionic species passing into it. The predominant phenomena here relate to the fact that cations are hydrated more than anions and that thermal energy promotes ion mobilization in the system. Ruptured bonds and the lattice parameters also affect the properties of the solid – liquid interface. The cleavage of fluorite, graphite, and chalcopyrite are shown in Figure 11.

Within short periods of contact, a double layer (usually referred to as the electric double layer), consisting of the charged surface of the solid and counter ion, is established. Although earlier theories considered this charge arrangement to be in the form of a simple plate capacitor, colloid science showed that the double layer

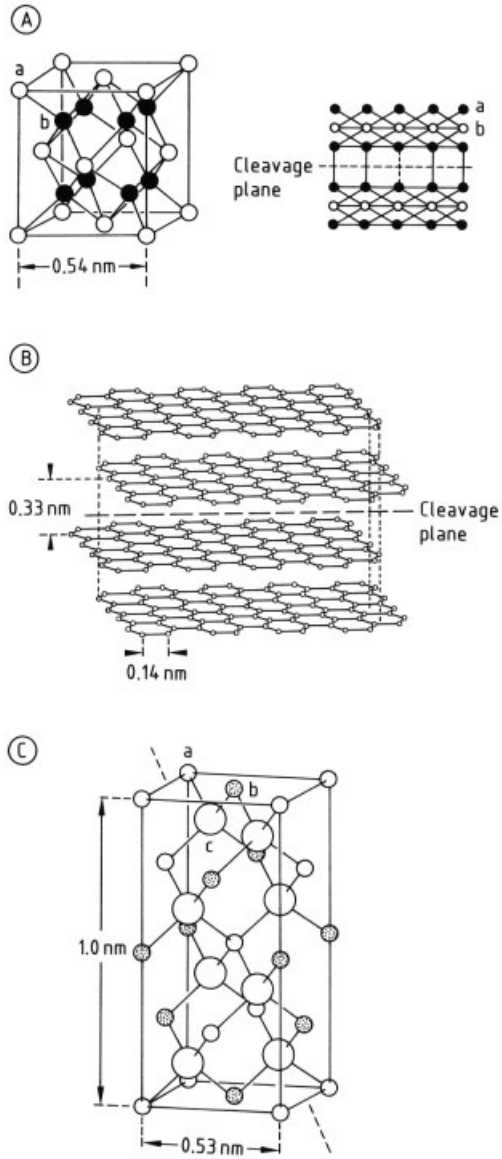

Figure 11. Crystal structures and mode of cleavage of minerals
A) Fluorite (CaF_2) and its cleavage plane
a) Calcium; b) Fluorine
B) Graphite
C) Chalcopyrite ($CuFeS_2$) which cleaves irregularly, sometimes at the (011) plane
a) Copper; b) Iron; c) Sulfur

consists of a surface with fixed charge whereas the ionic cloud countering it is diffuse [33]. The overall system, consisting of the solid and the ionic medium surrounding it, conforms to the principle of electroneutrality.

Zeta potential is widely used to describe the role played by the electrical double layer in flotation. It corresponds to the potential that exists at the so-called shear plane, which is created when a charged particle moves under the influence of an external electric field.

Zeta potential can be measured by electrophoresis, sedimentation potential, streaming potential, and electroosmosis. In the first two of these methods the shear plane is created by the motion of the particle; in the other two, the particle (more precisely, a porous bed consisting of many charged particles) is fixed [34, 35]. Equation (7) correlates the zeta potential with the electrophoretic mobility of the particles under the influence of an external electric field:

$$\zeta = \frac{4\pi\eta}{\varepsilon}\frac{V_e}{E} \tag{7}$$

where

ζ = zeta potential
η = viscosity of the aqueous medium
ε = dielectric constant
V_e = electrophoretic velocity of particle
E = electric field between electrodes applying the field

At 25 °C and $\eta = 0.895$ mPa · s, Equation (7) reduces to

$\zeta = 12.8\, V_e = 12.8\, u$ where u = electrophoretic mobility.

The significance of the zeta potential becomes apparent where electrostatic interactions play a major role in flotation reagent adsorption, as shown in Figure 12. The condition at which $\zeta = 0$ is known as the isoelectric point (iep), sometimes, less accurately, the zero point of charge (zpc). In principle, although an iep can be defined for each solid for numerous conditions, the variable most frequently used to describe the ionic conditions at which iep occurs is pH.

Adsorption mechanisms correlate with zeta potential; that is, physical adsorption affects the magnitude of the zeta potential. A change of sign in the presence of an appropriate concentration of adsorbate (see Fig. 13) is taken as an indication of a chemisorption-type reagent uptake [38]. As seen from Figure 13, the iep does not necessarily indicate zero potential at the surface of the particle, only that the potential at the shear plane is equal to zero.

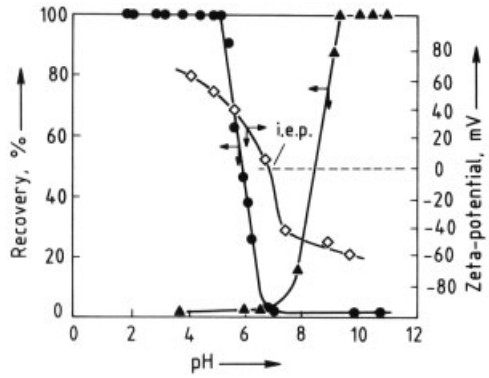

Figure 12. Relationship between iep and flotation properties of goethite (FeO (OH)) [36]
● = Sodium dodecyl sulfate or -sulfonate; ▲ = Dodecylammonium chloride; ◇ = iep
∗ Courtesy USBM; zeta potential measurements made in 10^{-4} M NaCl; note that minimum flotability occurs at the pH of iep and anionic reagents act as collectors at $\zeta > 0$ whereas cationic reagents collect goethite at $\zeta < 0$.

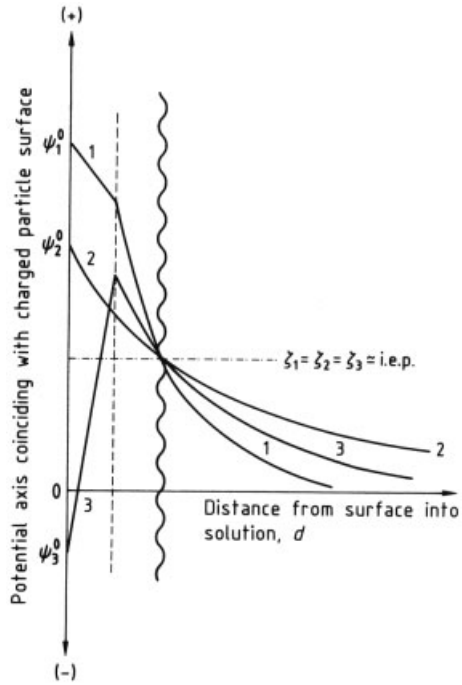

Figure 13. Representation of the fact that zero zetapotential does not necessarily indicate zero surface potential (ψ^o) [35, p. 221]
Dashed line = Inner Helmholz plane; wiggled line = Shear plane
∗ Although ψ_1^o and ψ_2^o are positive and ψ_3^o is negative, all conditions indicate $\zeta = 0$; inner Helmholz plane passes through the centers of unhydrated, specifically adsorbed ions.

Depending on the mechanism of reagent–solid interaction, many systems exhibit correlations between system variables and zeta potential (Fig. 14).

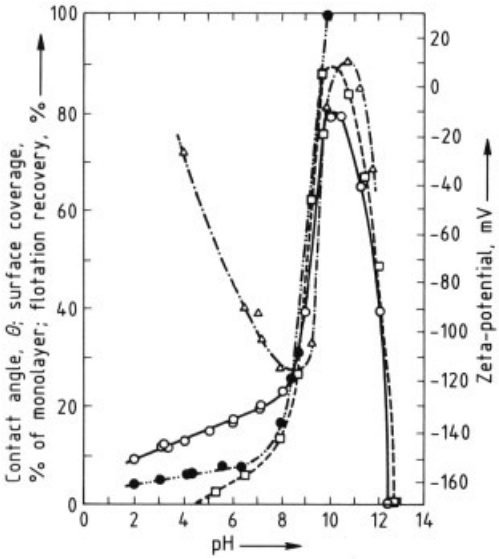

Figure 14. Correlation of commonly measured flotation and surface chemical variables in the quartz-dodecylammonium acetate system
□–□ = % Recovery; △–·–△ = Zeta potential; ○–○ = Contact angle; ●–··–● = Monolayer
∗ With permission [37].

The following isoelectric points ($\zeta = 0$) are observed in the absence of other substances.

AgCl	$pAg^+ = 4$
AgBr	$pAg^+ = 5.4$
Ag_2S	$pAg^+ = 10.2$
$CuSiO_3 \cdot 2H_2O$	$pCu^{2+} = 4$, at pH = 7
CaF_2	$pCa^{2+} = 3$
Fe_2O_3	pH = 6.5 and 8.2
FeS_2	pH = 7
FeS_2	pH = 5 (30 mg NaCN/L)
PbS	pH = 3.5
SiO_2	pH = 2 – 3.7
SiO_2	neutral pH in the presence of 3×10^{-3} mol dodecylammonium acetate/L

4. Bubble – Particle Interactions

The ultimate objective of a flotation process is the selective removal of solid particles from the aqueous medium, which is accomplished by the adhesion of air bubbles to the hydrophobic particles. Particle flotability can be treated as a probability [39]:

$$P_f = P_c \cdot P_a \cdot P_s \tag{8}$$

where

P_f = probability of flotation
P_c = probability of particle – bubble collision
P_a = probability of particle – bubble adhesion
P_s = probability of formation of a stable particle – bubble aggregate

In some methods, such as vacuum flotation, where dissolved gases become the bubble source, or in situ bubble-generation processes, where acids generate bubbles of carbon dioxide from carbonate-containing pulp, the probability treatment needs to be modified. Equation (8), however, is widely applied, because most flotation systems rely on extraneously introduced air bubbles. Particle – bubble aggregates are shown in Figure 15.

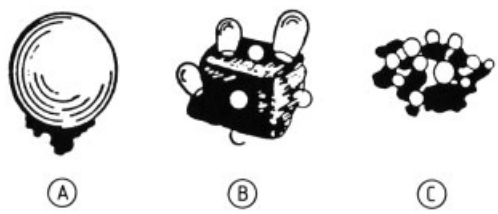

Figure 15. Different forms of particle – bubble aggregates [7,41]
A) Particle attached to bubble
B) Multiple bubble attached to large particle (common in vacuum flotation)
C) Particle – bubble flocculation in pulps containing fine particles

Solids to be floated carry an aqueous film, which may have a thickness as low as tens of nanometers [40]. The stability of this film determines whether a particle will adhere to a bubble on collision within the slurry: if the film thins, ruptures, and recedes, adhesion is facilitated. The time taken for the film to undergo these processes is known as the induction time.

Film stability has been quantified in terms as disjoining pressure [27].

$$\pi_D = p - p_o \tag{9}$$

$$\gamma = \gamma_0 \int_h \pi_D \, dh \tag{10}$$

where

π_D = disjoining pressure
p = pressure inside bubble
p_o = pressure of the bulk of the liquid
γ = specific surface energy of the liquid film
γ_0 = specific surface energy of an infinitely thick film of the same liquid

The conditions for film stability are shown in Figure 16.
For stable films: $\left(\frac{\partial \pi_D}{\partial h}\right)_T < 0$
For unstable films: $\left(\frac{\partial \pi_D}{\partial h}\right)_T > 0$

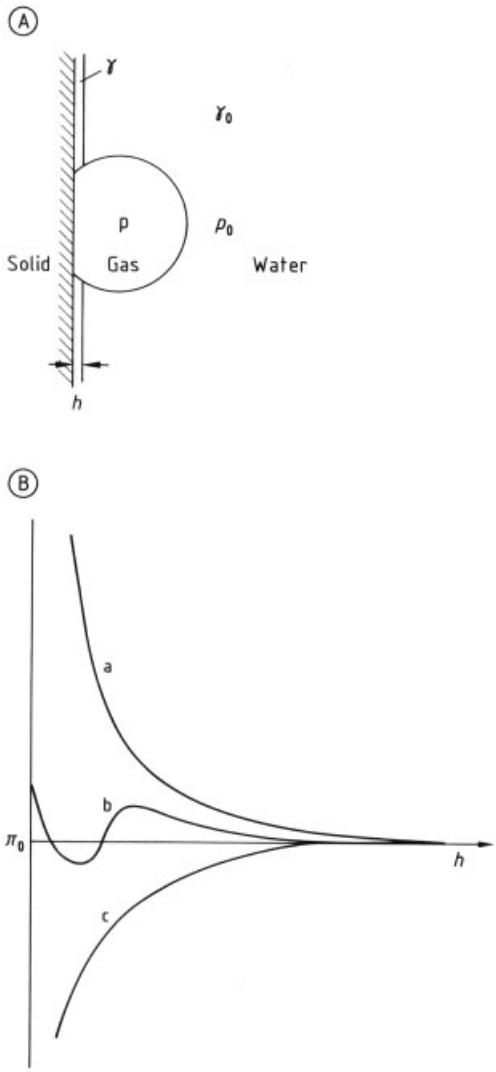

Figure 16. The concept of disjoining pressure [40]
A) Schematic representation of a disjoining film between a particle and a bubble
B) Relation of film stability to disjoining pressure
a) Stable film; b) Metastable film; c) Unstable film

Disjoining pressure depends on the summation of electrical double-layer effects (π_E), solvation effects (π_s), and London – van der Waals forces (π_v); i.e., $\pi_D = \pi_E + \pi_s + \pi_v$, as well as on the heterogeneity of the solid surface [42]. The role of frothers is to make π_D more positive, i.e., reduce film resistance to rupture.

5. Flotation Processes

The phenomena described so far relate to the froth flotation process. The following variations and techniques are also used:

Collectorless flotation:
froth flotation using the natural hydrophobicity of some minerals

Two-liquid flotation:
variant of froth flotation, using oil droplets instead of air bubbles and an oil layer instead of froth

Column flotation:
variant of froth flotation, employing columns 3 – 9 m high with 0.3 – 1.5 m cross-section

Microbubble flotation:
variant of column flotation, employing bubbles 10 – 50 µm in diameter

Dissolved-air flotation:
air bubbles are generated by applying vacuum to air-saturated slurries

Gamma flotation:
liquid – vapor surface-tension control is used to separate hydrophobic particles

Piggyback flotation:
also known as carrier flotation: uses the mutual coagulation of solids; when one is floated, the other (usually more difficult to float on its own) is also collected

Skin flotation:
hydrophobic particles are removed without froth, but at a layer approximately one particle thick

5.1. The Gamma Flotation Process

The gamma flotation process is characterized by emphasis on the control of solution surface tension. The critical surface tension of wetting of solids is based on a plot of the cosine of contact angle θ against the surface tension of the solution in contact with a solid [43]. Each hydrophobic solid exhibits $\cos\theta = 1$ (zero contact angle) at a given solution surface tension as shown in Figure 17. Difficulties in contact-angle measurements on powdered solids, as used in flotation practice, can be overcome by a technique where flotation recoveries are plotted against solution surface tension [10,44,45]. The γ_{LG} at which flotation recovery is equal to zero is taken as the critical surface tension of wetting, γ_c, of the solid under investigation. The two techniques [43–45] are equivalent [10]; the proof of this constitutes an independent demonstration of the thermodynamic basis of the flotation process [46].

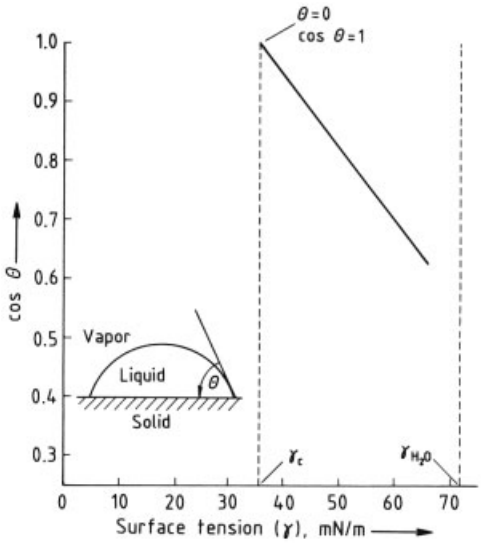

Figure 17. Schematic of critical surface tension of wetting by contact-angle measurement

Figure 18 shows the use of critical surface tension differences of two hydrophobic solids by controlling the surface tension of the solution in which they are suspended. This method is called gamma flotation [9]. The value of the process has been demonstrated for a number of flotation systems including oil shale, sulfide

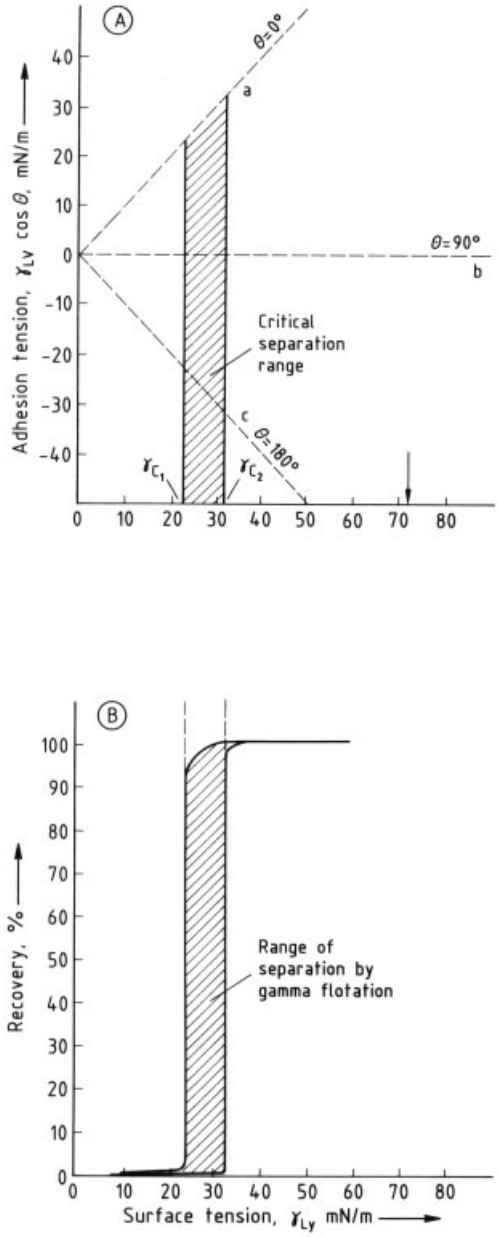

Figure 18. Schematic of physical basis of the gamma flotation process [9]∗
A) Adhesion tension vs. solution tension for two solids with different critical surface tension or wetting
B) Relation of flotation recovery to solution surface tension for two solids with different critical surface tension or wetting
γ_{c_1} = Critical surface tension of wetting of solid 1;
γ_{c_2} = Critical surface tension of wetting of solid 2;
∗ Lines a, b, and c obtained by substituting values into the adhesion tension equation, e.g., [$\gamma_{LG}\cos\theta$].
∗∗ In the shaded area solid 1 with γ_{c_1} floats, whereas solid 2 with γ_{c_2} is completely wetted by the aqueous solution.

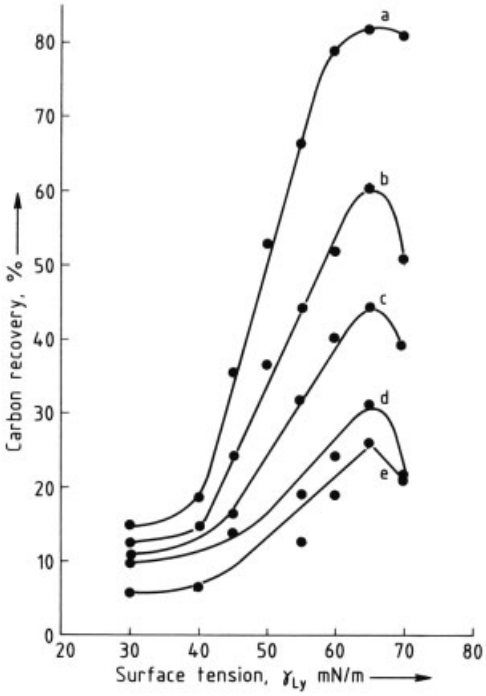

Figure 19. Separation of Colorado oil shale sizes fractions by exploiting the kinetic aspects of gamma flotation [47]
a) <13 µm; b) 18 – 13 µm; c) 26 – 18 µm; d) 38 – 26 µm; e) 48 – 38 µm

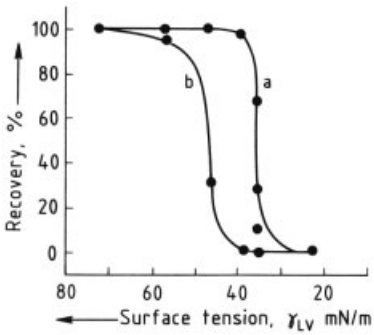

Figure 20. Separation of a mixture of silica and magnetite by gamma flotation ∗
a) SiO_2; b) Fe_3O_4
∗ Powders have been initially made hydrophobic by treatment with appropriate surfactants and mixed after filtration [49]

minerals, and others [48–50] (see Figures 19 and 20). Salt flotation, which has been used in coal processing, is a variant of gamma flotation in which the natural hydrophobicity of coal is exploited and the surface tension of the solution is controlled by salts. Figure 6 shows that some

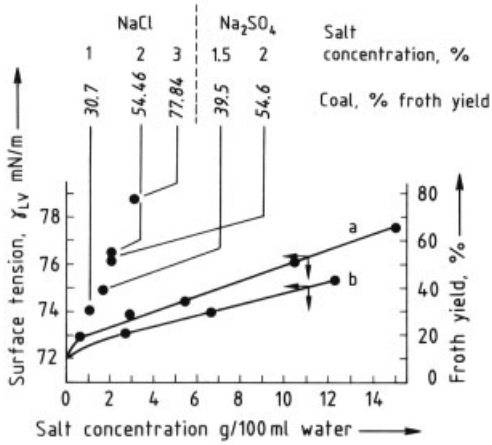

Figure 21. Flotation in high surface tension solutions ∗ [9]
a) NaCl; b) Na_2SO_4
∗ This method is also known as salt flotation [7, p. 338, 339], [13].

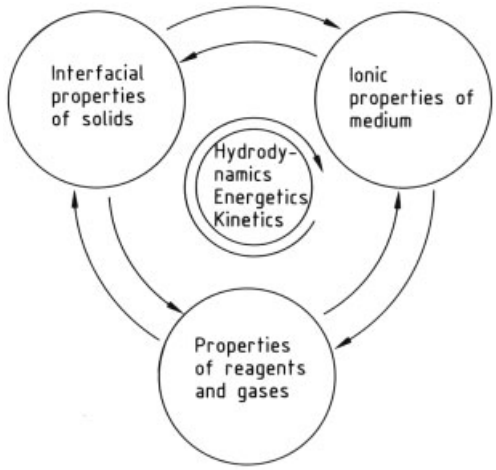

Figure 22. System variables that affect the flotation process [9]

salts raise the surface tension of water to values above the usual 72 mN/m. Figure 21 demonstrates the relationship between solution surface tension modified by salts and flotation results obtained with various coals [9].

5.2. Design of Flotation Operations

Flotation testing is generally preceded by geological, engineering, and mineralogical work. The overall flotation system is a complex combination of variables outlined in Figure 22, but

laboratory testing is based on mineral characteristics and reagent properties. Recovery is "the percentage of total values contained in the ore which is recovered in the concentrate"; grade refers to "assay of values in the concentrate." Recovery and grade tend to vary inversely.

Laboratory tests are designed to establish crushing and grinding conditions and optimize such variables as pulp density, reagent addition, pH, and auxiliary reagent concentration before the pilot plant stage. An important laboratory test is the "locked-cycle test", which simulates pilot plant conditions (Fig. 23) [1].

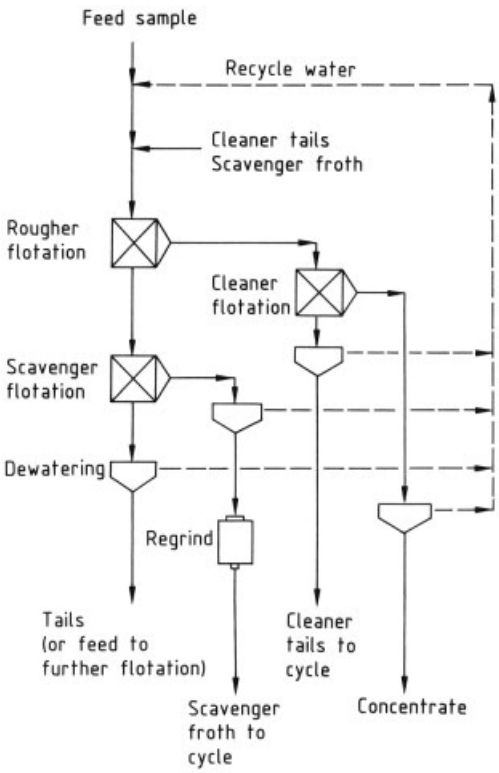

Figure 23. Schematic of locked-cycle test [51]

Flotation machine variables, such as volume, area, throughput, aeration rate, and cell layout are considered at the plant-design stage. Simulation studies using computer modeling techniques have gained increasing acceptance [1, 52]. The final plant layout provides for product handling, filtration, dewatering, tailings disposal, and plant-water recycling.

5.3. Flotation Machines

Flotation machines consist of a chamber (cell) equipped with entry and exit ports and an agitation assembly that keeps the pulp in constant motion during the separation. These machines fulfill the following functions: (1) mixing and agitation of the slurry, (2) aeration and promotion of particle – bubble collisions, (3) formation of a froth layer and its removal for product recovery, and (4) continuous discharge of tailings, consisting of components not collected in the froth layer [53].

The design of a flotation machine (or cell) is largely empirical and aims at maintaining high throughput subject to the previously stated criteria; the Wemco flotation machine is shown in Figure 24. Flotation cells can reach volumes of 20 – 25 m^3 [54]. Figure 25 is an outline of contemporary flotation machine types; Figure 26 shows cell geometries and paddle designs in mechanically agitated flotation machines. Flotation machines used in water treatment are not always mechanically agitated but may rely on the quieter conditions provided by the generation of air bubbles in the system.

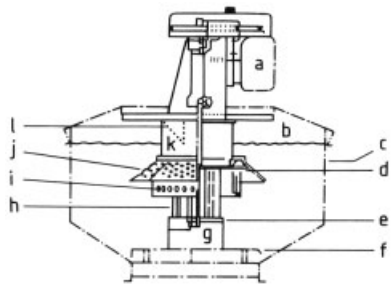

Figure 24. Wemco flotation machine ∗
a) Motor; b) Weir level; c) Rotor submergence; d) Rotor top; e) Adjustable collar; f) False bottom; g) Draft tube; h) Rotor; i) Disperser; j) Hood; k) Stand pipe; l) Air inlet duct
∗ With permission [53]

Such methods include vacuum flotation, in which air-saturated water is subjected to a vacuum for bubble generation. Pressure flotation, on the other hand, does not require the application of vacuum, since air bubbles are released after saturation of water under pressure. Electroflotation has also been used to a limited extent. In this method hydrogen and oxygen bubbles are generated by electrolysis [57].

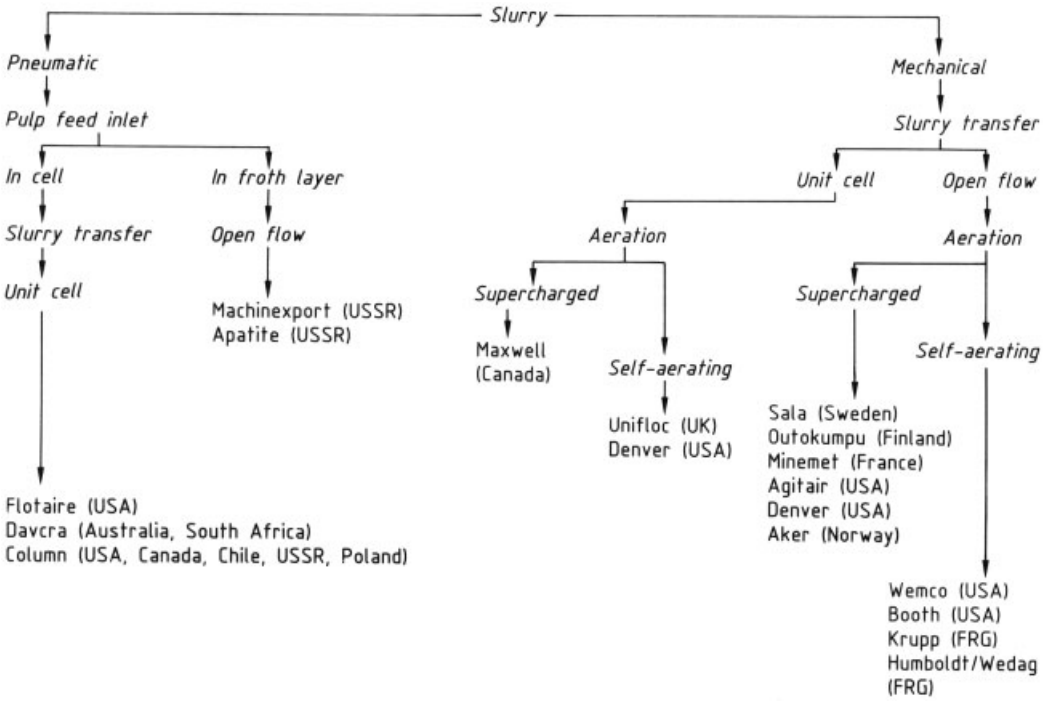

Figure 25. Types of flotation machines, their modes of operation, and representative manufacturers [54–56]

A new version of an old flotation cell has led to the development of column flotation (see Fig. 27) [58–60]. Column cells seem to perform best with highly hydrophobic solids. They are used for the recovery of sulfides and coal, and to recover valuable particles rejected in the tailings of conventional, mechanically agitated, batteries. Column flotation offers the following advantages:

1) reduction of gangue entrainment in froth,
2) increase of bubble residence times in the pulp,
3) improved selectivity,
4) energy economy per unit weight of values recovered, and
5) improved control and instrumentation.

The following relationships are used in the design of flotation machines [56]:

Geometrical
Effective volume: $V \propto L^{n}$ ($n \sim 2-3$)
Froth area: $A \propto L^{2}$
Impeller/tank aspect ratio: $D/L \propto D^{r}$ ($r \sim 0-0.5$)
Impeller speed
Peripheral speed: $S = \pi ND \propto D^{a}$ ($a \sim 0-0.5$)
Power (constant viscosity and density)
Net power consumption: $P \propto N^{3}D^{5}/N^{c}D^{2c}$ ($c \sim 0-0.33$)
Aeration
Air-flow number: $N(Q) = Q/AND \propto D^{f}$ ($f \sim 0$)
Impeller diameter is the parameter of machine size
Geometry: $L \propto D^{1-r}$; $A \propto D^{2(1-r)}$; $V \propto D^{n(1-r)}$
Power: $P \propto D^{(a+1)(3-c)-1}$
Aeration: $Q \propto D^{a+f+2(1-r)}$
Power intensity or specific power:
$P/V \propto D^{(a+1)(3-c)-n(1-r)-1} \propto D^{b}$
Nominal residence time of air:
$V/Q \propto D^{(2-n)(r-1)-(a+f)} \propto D^{q}$

N = Impeller rotational speed in rpm; D = Impeller diameter; L = Tank length. Exponents are experimental parameters characteristic of specific machine design.

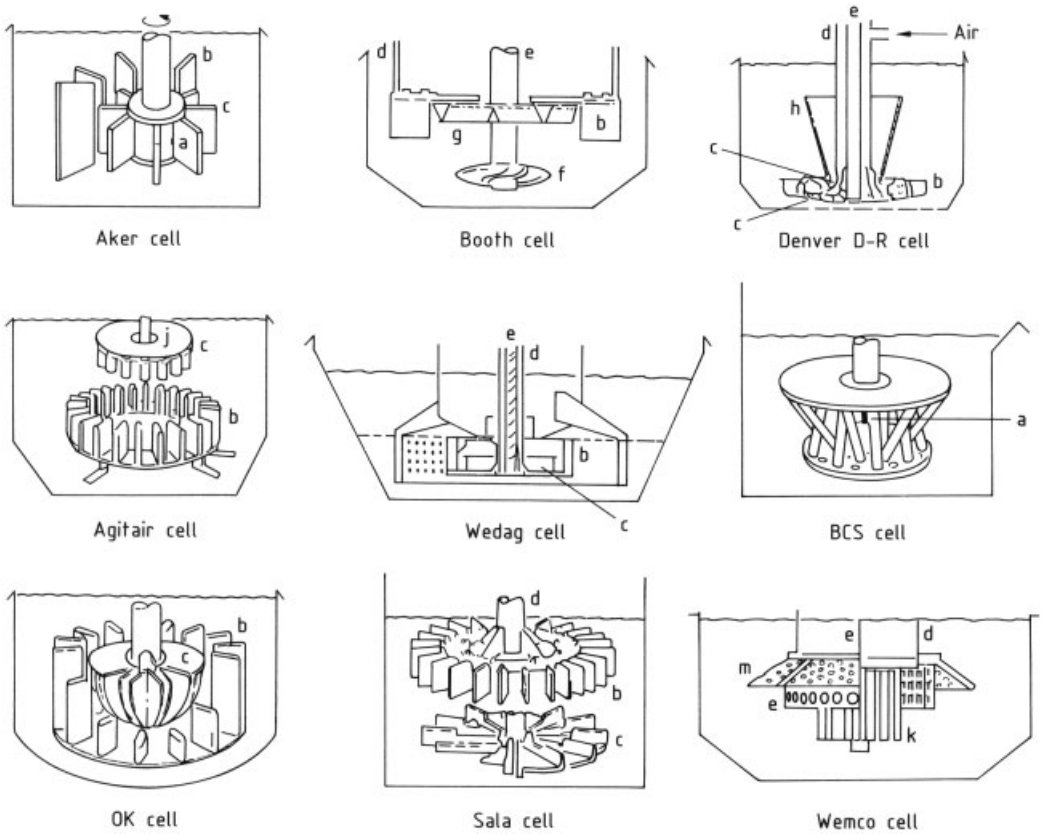

Figure 26. Schematic of impeller and tank geometries of mechanically agitated flotation cells [55]
a) Air port; b) Diffuser; c) Impeller; d) Standpipe; e) Shaft; f) Propeller; g) Aerating impeller; h) Recirculation well; j) Pulp entry; k) Rotor; m) Disperser hood

5.4. Industrial Applications

Various methods of separation exploit differences in size, relative density, and magnetic and electrostatic properties. Flotation, however, exploits two unique controllable properties of particulates (Table 5): (1) it operates at very fine particle sizes, and (2) it is based on inherent or modified surface properties of solids.

Since its discovery in the late 19th century, the flotation process has been used principally in the minerals industry. At present, it is used to treat several thousand million tonnes of ore annually. Most of the minerals-related applications aim at recovering ore values, rejecting the rest as tailings.

Applications of flotation in the minerals industry are diverse and include recovery of sulfide minerals, upgrading phosphates ores, ash and sulfur removal from coal, separation of soluble minerals such as halite (NaCl) and sylvite (KCl) from each other, and recovery of fine gold [6–8, 61].

Water-treatment applications include the removal of oily components and colloidal matter from industrial waters, water recycling, and the removal of color and metal ions.

Mineral Flotation. Minerals containing valuable elements can be recovered economically by flotation (Table 6) in the following stages:

Crushing and grinding. These steps reduce particle size and mechanically liberate the desired minerals. Some deposits, such as beach sands or placers, do not require crushing, which is usually effected by jaw or gyra-

tory crushers. For grinding, equipment such as ball mills and rod mills is employed.

Classification and sizing. As can be seen from Table 5, the optimal size range for flotation is 10 – 500 µm. Coal particles can be floated at the upper limit, and copper sulfide particles at the lower. Particles of nonsulfide minerals finer than 50 µm, however, interfere with flotation. Such fines are separated by hydrocyclones, sedimentation tanks, and other classification equipment; they are discarded as slimes (Figure 28).

Conditioning. In the conditioning operation the ore powder is slurried with water and flotation reagents, forming the flotation slurry. In the flotation step, air is introduced to transfer hydrophobic particles to a froth layer, which is skimmed away mechanically by paddles.

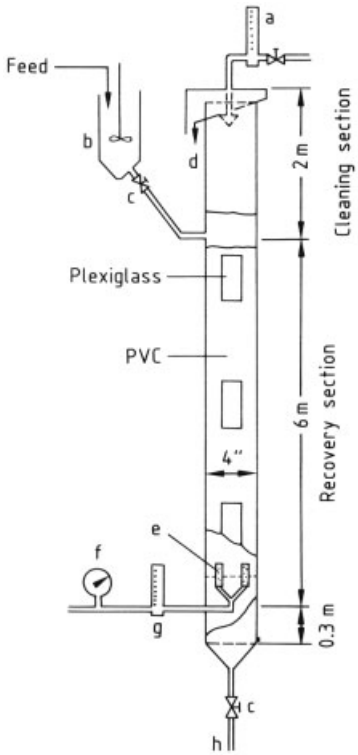

Figure 27. Pilot plant size flotation column
a) Wash water flow meter; b) Conditioner; c) Valve; d) Concentrate discharge; e) Porous glass diffuser; f) Air regulator; g) Air flow meter; h) Tails discharge

Table 5. Physical properties of particulates and size ranges suitable for separation technology

Physical property	Physical separation technique	Approximate particle size range, mm
Color or radiation	sorting	10 – 100
Size	dry or wet screening	0.01 – 10
Size, shape, density	wet or dry cyclones	0.005 – 1
	sedimentation and hydraulic classification	0.003 – 3
	centrifugation	0.001 – 0.05
Density	heavy media	1 – 10
Size, density	jigging	5 – 50
	tabling	0.005 – 3
Magnetic susceptibility	wet or dry magnetic separation	0.001 – 5
Surface or bulk conduction	electrostatic separation	0.05 – 5
Surface chemistry	flotation	0.001 – 0.5
	spherical agglomeration	0.01 – 0.5
	selective coagulation and selective flocculation	0.05 – 0.001

The solids content of the pulp, known as pulp density, averages 20 – 40 %, but may be 3 – 10 % in some systems, such as coal flotation. In the preparation of potable water, the solids content may be well below 1 %.

The mineral flotation process is carried out in interconnected batteries of flotation cells, where slurry flow occurs by gravity and the level of the slurry is controlled by gates or weirs.

The arrangement of flotation cells, sometimes with intermittent grinding of the products, leads to flotation circuits of which five basic types have been identified [62] (Fig. 29).

5.4.1. Sulfide Mineral Flotation

Sulfide minerals such as those of copper ($CuFeS_2$), lead (PbS), and zinc (ZnS) are commonly found in the same ore. Xanthates (dithiocarbonates) and dithiophosphates are used as collectors in sulfide mineral flotation; frothers include pine oil and synthetic commercial compounds such as Dowfroth 250. Some qualified generalizations apply to sulfide mineral flotation:

1) Xanthates and dithiophosphates are collectors for sulfides.
2) The chemisorption of xanthates and the presence of oxygen are essential for the flotation of sulfide minerals such as galena [63–65].

Table 6. Elements contained in minerals recovered by flotation

Element	Mineral	Composition	Element	Mineral	Composition
Al	bauxite	$Al_2O_3 \cdot xH_2O\,(Fe,Mn,Ti)$	Mn	pyrolusite	MnO_2
Al	kaolinite	$Al_2Si_2O_5(OH)_4$	Mo	molybdenite	MoS_2
As	arsenopyrite	$FeAsS_2$	Na	halite	NaCl
Au	native gold		P	fluorapatite	$Ca_5F(PO_4)_3$
B	colemanite	$Ca_2B_6O_{11} \cdot 5\,H_2O$	S	native sulfur	
Ba	barytes	$BaSO_4$	Si	quartz	SiO_2
C	diamond	carbon	Sn	cassiterite	SnO_2
C	graphite	carbon	Sr	strontianite	$SrCO_3$
Ca	calcite	$CaCO_3$	Th	fergusonite	$(Y,Ce,U,Th,Ca)\,(Nb,Ta,Ti)\,O_4$
Cr	chromite	$FeO \cdot Cr_2O_3$	Ti	rutile	TiO_2
Cu	chalcopyrite	$CuFeS_2$	U	carnotite	$K_2O \cdot 2\,U_2O_3 \cdot V_2O_5 \cdot 2\,H_2O$
F	fluorite	CaF_2	V	vanadinite	$Pb_5ClV_3O_{12}$
Fe	hematite	Fe_2O_3	W	scheelite	$CaWO_4$
Hg	cinnabar	HgS	Y	monazite	$(Ce,Y)\,PO_4$
K	sylvite	KCl	Zn	sphalerite	ZnS
Li	spodumene	$LiAl\,(SiO_3)_2$	Zr	zircon	$ZrSiO_4$
Mg	magnesite	$MgCO_3$			

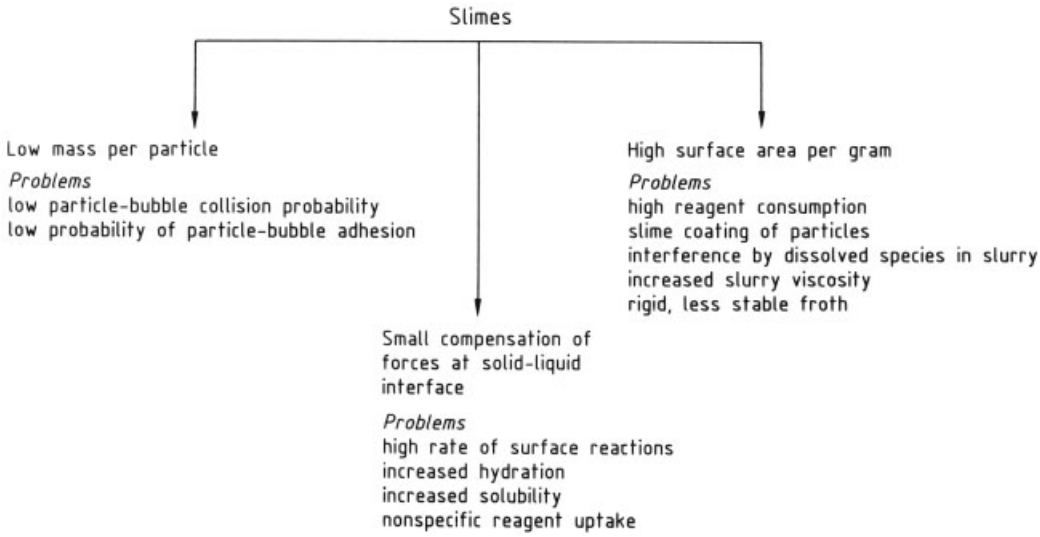

Figure 28. Characteristics of slimes [7]

This correlates well with the electrochemistry of the system and the stability of species that collect at the solid surface (Fig. 30).

3) Other variables being constant, an increase in chain length for a homologous series of reagents strengthens them as collectors; this rule also applies to other minerals; see Figure 14.
4) At a given constant collector concentration, a critical pH for each sulfide mineral determines the boundary of flotation – no flotation conditions. The ratio of concentration of xanthate to hydroxide for such a critical transition usually conforms to the Barsky relationship: $[X^-]/[OH^-] = \text{constant}$
5) The ions OH^-, S^{2-}, CN^-, $Cr_2O_7^{2-}$ are common sulfide mineral depressants for selectivity control in flotation; Cu^{2+} is an activator for ZnS in flotation by xanthates. Figure 31 is a flow diagram of a sulfide flotation plant.

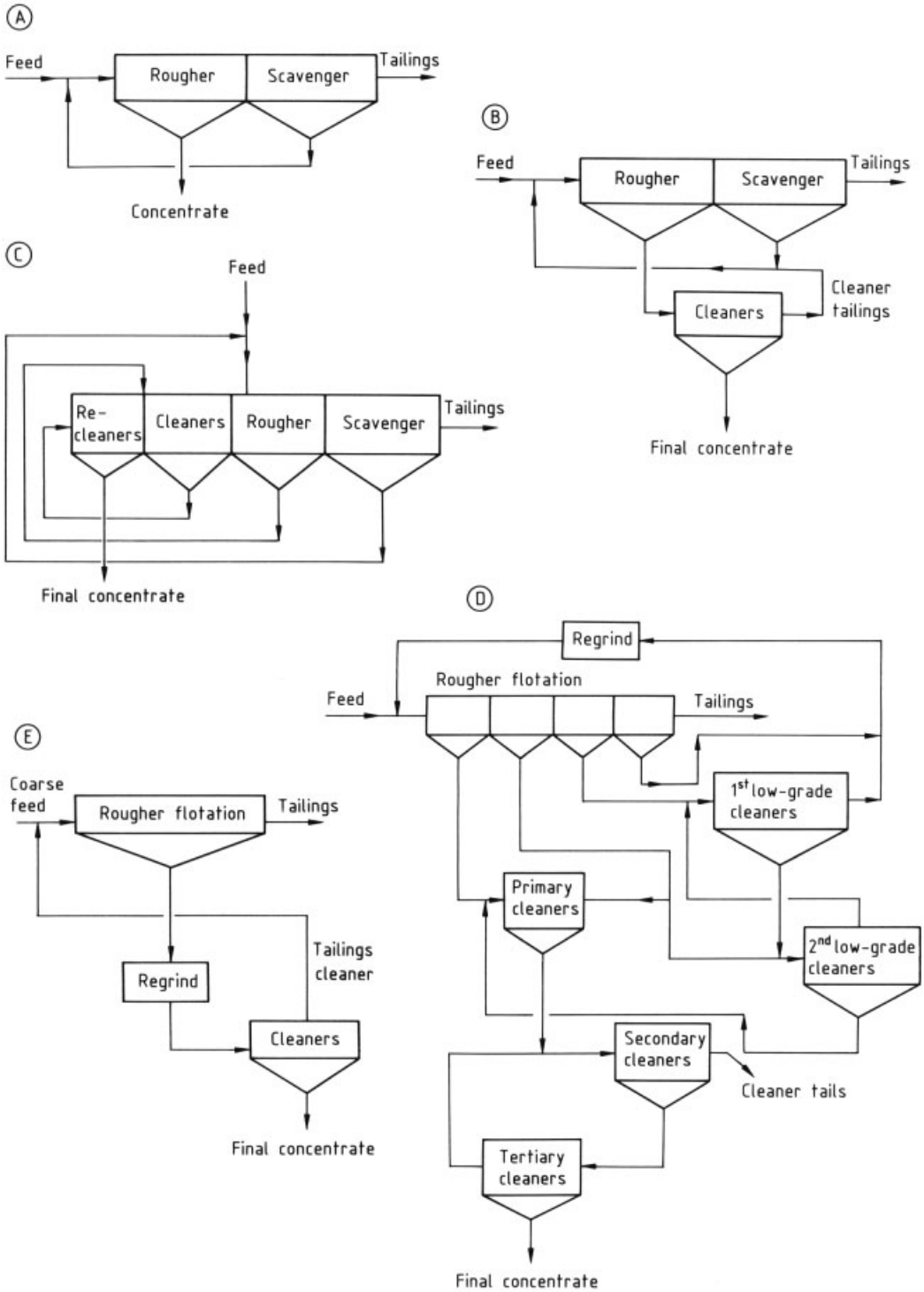

Figure 29. Flotation circuit configurations ∗
A) Type 1; B) Type 2; C) Type 3; D) Type 4; E) Type 5
∗ With permission [62].

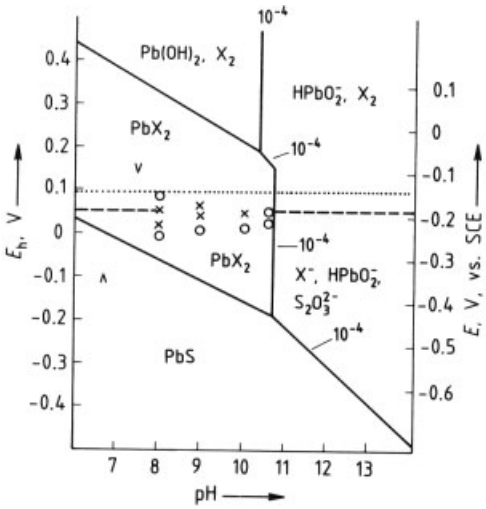

Figure 30. E_h-pH Diagram of thermodynamically stable species and associated experimental data in the galena – oxygen – xanthate flotation system [42]
x = Strong bubble contact; o = Feeble bubble contact; v = Oxidizing depression by potassium chromate; ∧ = Reducing depression by sodium sulfite-ferrous mixture; ····· = Xanthate (10^{-3} M) – dixanthogen (saturated); - - - - - = Galena – dixanthogen system in equilibrium with 10^{-3} M xanthate ion; E = Electrical potential

5.4.2. Nonsulfide Flotation

Nonsulfide ores include those of iron, phosphate, fluorite, magnesite, and beach sand. Some ores may contain sulfur-bearing minerals as secondary components. Collectors in these systems are generally fatty acids and their salts, sulfonates, and sulfates. Amines are also used, especially in iron-bearing ores and in silicate flotation. Figure 32 shows the separation of lead carbonate, barium sulfate, and calcium fluoride in a 100 t/d plant; the main gangue is siliceous material.

Phosphate-ore concentration plant flow sheets are shown in Figures 33 and 34. Their difference lies in the run-of-mine grades and the complexity of gangue minerals.

An interesting application of nonsulfide mineral flotation is the separation of soluble salts from brines (Fig. 35). Primary amines are well-established selective collectors and have long been used in the production of KCl from the Dead Sea. In some cases, clays interfere with

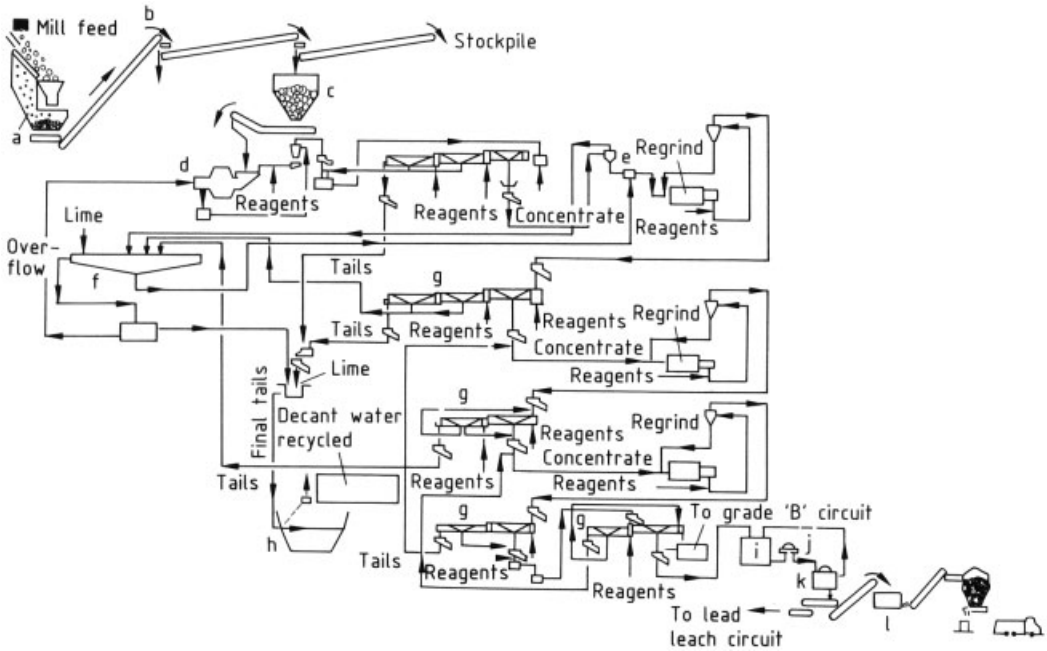

Figure 31. Flow sheet of plant treating 30 000 t/d of molybdenite ore, Henderson Mill, Colorado ∗
a) Gyratory crusher; b) Self-cleaning magnet; c) Ore storage bin; d) Semiautogenous mill; e) Pump box; f) Thickener; g) Cleaner; h) Tailing pond; i) Concentrate holding tank; j) Pump; k) Disc filter; l) Dryer
∗ Courtesy of Denver Equipment Co.

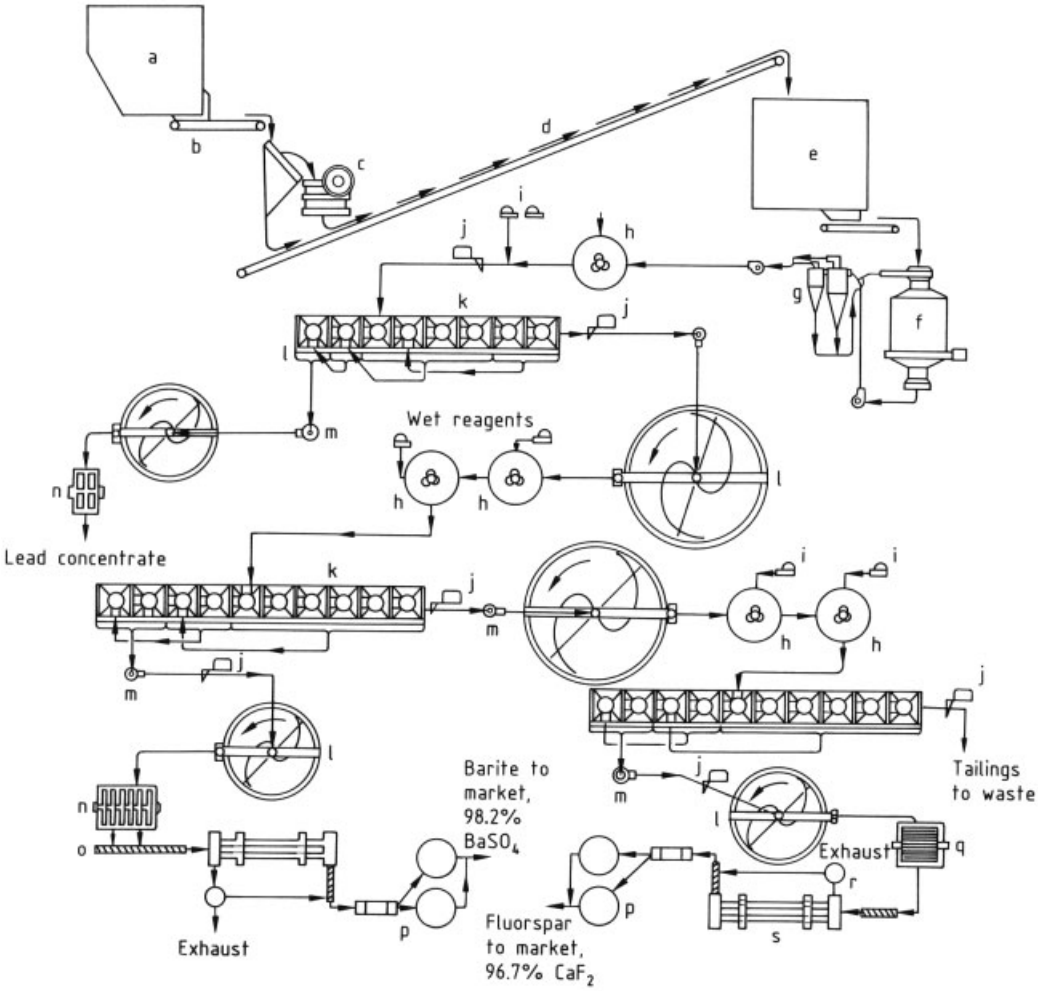

Figure 32. Flow sheet for $PbCO_3$, CaF_2, and $BaSO_4$ concentration ∗
a) Coarse ore bin; b) Apron feeder; c) Jaw crusher; d) Belt conveyor; e) Fine ore bin; f) Ball mill; g) Cyclone; h) Conditioner; i) Reagent feeders; j) Sampler; k) Flotation machine; l) Concentrate thickener; m) Pump; n) Disc filter; o) Screw conveyor; p) Elevator and storage bins; q) Fluorspar filter; r) Dust cyclone; s) Dryer
∗ Courtesy of Denver Equipment Co.

the flotation and need to be depressed by the addition of starch, lignin, or cellulose derivatives.

5.4.3. Coal Flotation

Coal is cleaned by gravity separation to reduce ash and pyrite content [69]. These methods include jigging, tabling, and heavy-media separation. The fines (ca. 400 µm, 35 mesh) are treated by flotation. To liberate ash and sulfur-bearing components, the run of mine coal is sometimes ground to ca. 600 µm (28 mesh) or finer, and the whole mass subjected to flotation without recourse to primary gravity separation. The natural hydrophobicity of freshly ground coal is an asset which minimizes the use of collectors. A small quantity of MIBC or pine oil (50 – 100 g per tonne of mineral) acts as both collector and frother.

On extended storage, or if the coal seam has been in contact with air for geological time periods, coal loses its hydrophobicity and is referred to as "weathered coal". In such cases, MIBC, low pH, and the addition of hydrocarbon oils such as

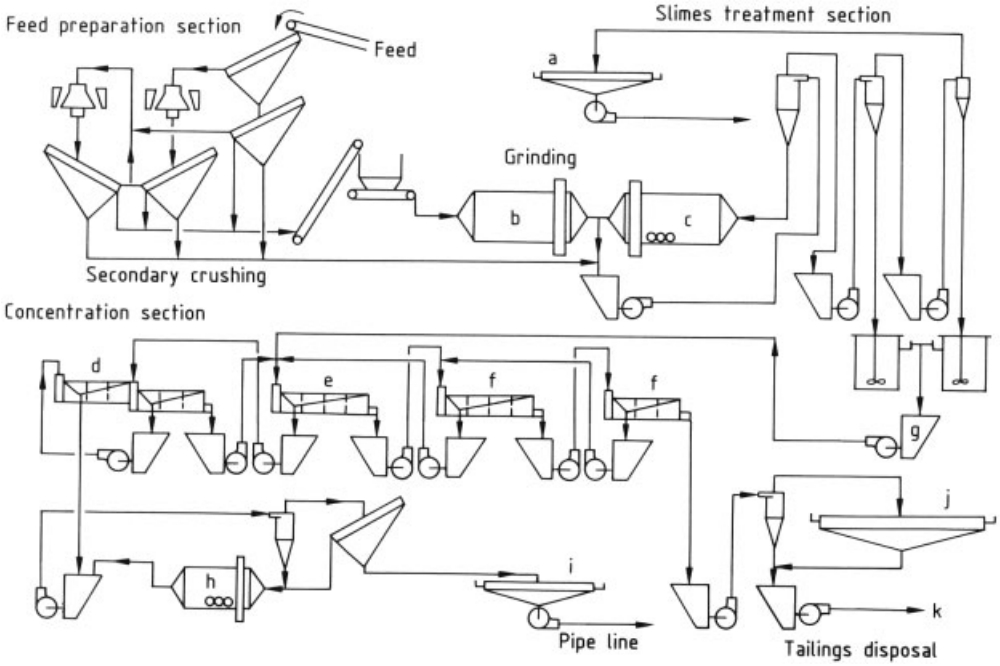

Figure 33. Flow sheet for phosphate ore flotation in Brazil ∗
a) Slime thickener; b) Rod mill; c) Ball mill; d) Cleaner; e) Rougher; f) Scavenger; g) Conditioner; h) Concentrate regrinder; i) Concentrate thickener; j) Tailing thickener; k) Tailing pond
∗ The main mineral of value is apatite; the gangue contains barytes, iron oxides, and silicates; fatty acid flotation generates concentrates assaying an average of 36 % P_2O_5, with recoveries of 62 – 63 %; with permission [66].

kerosene or fuel oil make coal matter floatable [42, 70].

Sometimes, pyrite is first removed by conventional sulfide mineral flotation using xanthates as collectors. This is followed by the flotation of coal matter and the rejection of ash-forming materials such as silica, silicates, and carbonaceous shale.

Unlike other minerals, coal exhibits a high degree of variation with respect to origin, age, rank, moisture content, degree of weathering, physical and chemical structure, and nature of gangue. As a result, it is difficult to develop a unified flotation strategy from sample to sample. Even samples of the same origin undergo surface chemical changes to different extents on prolonged storage.

Nonetheless, generalized strategies are possible (Table 7). Figure 36 shows a flow sheet of coal flotation [71].

5.4.4. Flotation of Precipitates and Ions

Precipitate flotation is based on the same principles as mineral flotation. For example, in the preparation of potable water, the adjustment of pH to precipitate ferric ions as hydroxides eliminates most of the iron and colloidal matter by a subsequent flotation step. In this case, the water contaminants render the hydrolysis product sufficiently hydrophobic for bubble – particle adhesion to permit flotation under quiescent conditions.

Alternatively, dissolved ions can be electrostatically attached to selective cations, which in turn can be collected in a foam layer (Fig. 37). The recovery of Au, Pd, Pt, and Ir by such foam fractionation methods, as well as the removal of uranium from aqueous solutions by the use of amines, has been demonstrated [73–75].

Figure 38 shows a foam fractionation (ion flotation) assembly. At times of increasing environmental concern, the removal of heavy metal ions and toxic substances such as PCBs

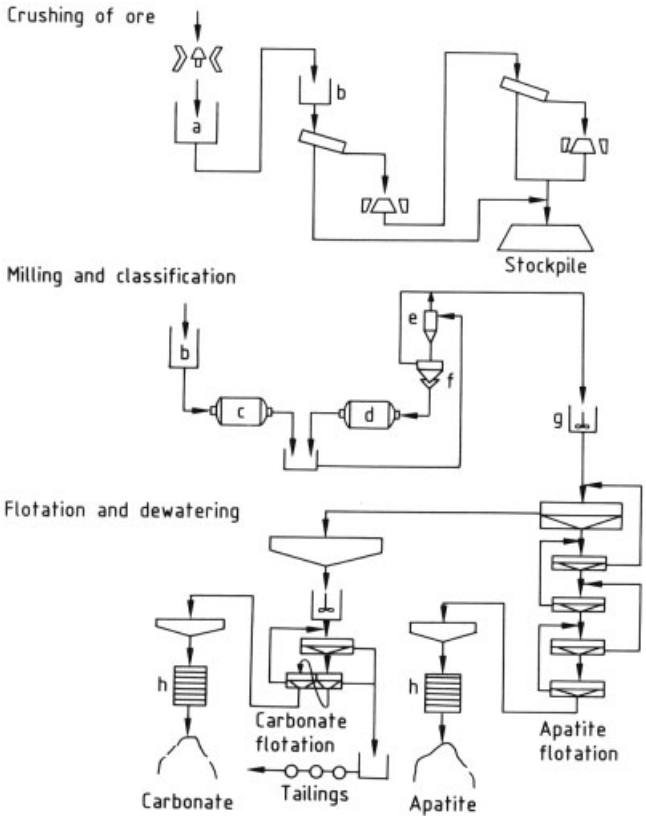

Figure 34. Flotation flow sheet for Finnish low grade apatite ore ∗ [67]
a) Silo; b) Feed bin; c) Rod mill; d) Ball mill; e) Hydrocyclone; f) Cone classifier; g) Conditioner; h) Pressure filter
Head ore = 3.9 % P_2O_5; Concentrate grade = 36.5 % P_2O_5; Recovery = 75 – 79 %
∗ Gangue minerals include calcite, dolomite phlogopite [$(KMg_3(AlSi_3)O_{10}(OH,F)_2$], and mica.

(polychlorinated biphenyls) seems a promising area for further application of ion flotation and foam fractionation techniques.

5.4.5. Use of Oil Droplets instead of Air Bubbles

Surfaces to which air bubbles can adhere are also readily wetted by hydrocarbon oil droplets in preference to water. This has been demonstrated thermodynamically and experimentally. Although micrometer-size solid particles may adhere nonselectively to air bubbles, the replacement of air by hydrocarbon oils provides better selectivity, as proved experimentally. These considerations have given rise to the development of two-liquid flotation for extraction of finely divided solids into an oil phase (oil extraction) in assemblies similar to those used in conventional flotation technology.

This method has been used for alumina (Al_2O_3), hematite (Fe_2O_3), cassiterite (SnO_2), and artificial diamonds, by first making the powders hydrophobic with an amine or a sulfonate, followed by extraction into an oil phase [76, 77]. The thermodynamics of the system apply equally well to the now well-established process of coal cleaning by oil agglomeration.

6. Future Trends

The decline in the mining and metallurgical industries has shifted emphasis away from conventional flotation technology. Nonetheless, understanding of the fundamentals of flotation is by no means complete and development of sophis-

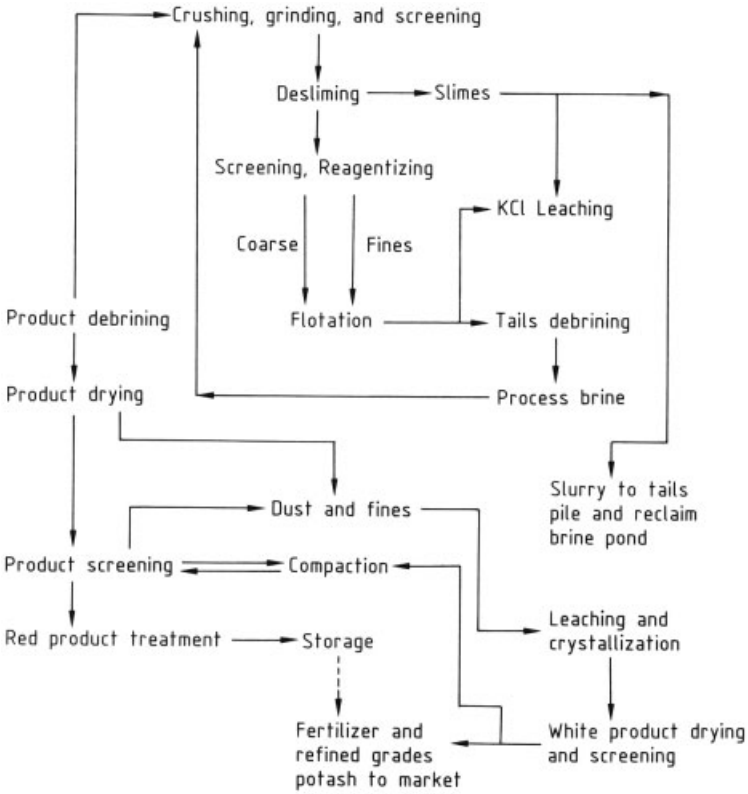

Figure 35. Potash treatment flow sheet as practiced by the Potash Corporation of Saskatchewan, Canada [68]

ticated instrumentation is leading to combined research on materials science and on surface fundamentals of flotation [78].

As new applications, such as control of the so-called pulp potential, control of redox conditions, E_h, use of critical surface tension of wetting, and flotation of ions and dissolved species, such as toxic and hazardous materials, become more widespread, froth flotation will be employed more and more.

Computer technology and instrumental control now make larger plants possible with higher-grade products and more complete recoveries. Furthermore, pyrometallurgy and hydrometallurgy are now used together with flotation for recovery of metals from slags or melts [79].

The flotation of precious metal-bearing ores and native gold is widely applied despite the apparent fate of sulfide mineral flotation. Less familiar systems such as rare-earth minerals are finding increased interest. Column flotation promises wider application. The custom-design of flotation chemicals is also under development and wider application of flotation in recycling industries is at its early stages.

7. References

1. B. Yarar, Z. Dogan (eds.): *Mineral Processing Design,* Martinus-Nijhof, Dordrecht, The Netherlands 1987.
2. B. Yarar: "Investigation of the Recovery of Cellulose from Difficult to Process Sources, Using Surface Chemical Principles," *Proc. 4th Recycling World Congress,* New Orleans, La., 1982.
3. B. Yarar, D. J. Spottiswood (eds.): *Interfacial Phenomena in Mineral Processing,* Engineering Foundation Publ., New York 1982.
4. B. Yarar, S. R. Rao: "Flotation," in J. J. McKetta (ed.): *Encyclopedia of Chemical Processing and Design,* vol. 23, Marcel Dekker, New York 1985, pp. 454 – 508.

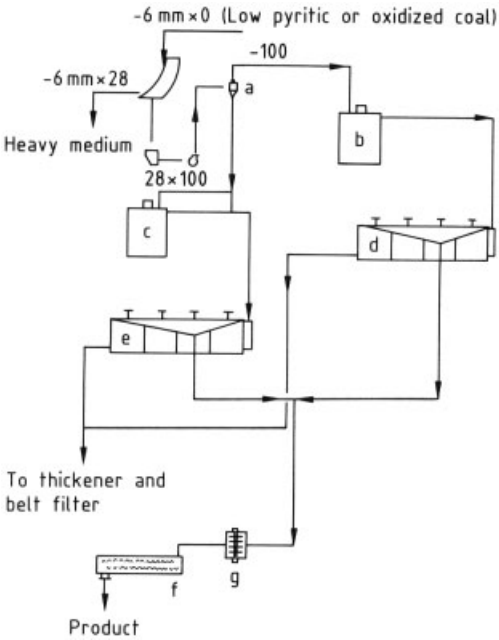

Figure 36. Flotation flow sheet applicable to oxidized coals or low pyrite-content coal ∗
a) Cyclone; b) Conditioner, 6 %; c) Conditioner, 25 %; d) Fine rougher float, 6 %; e) Coarse rougher float, 15 %; f) Drier; g) Disc filter
∗ With permission [71].

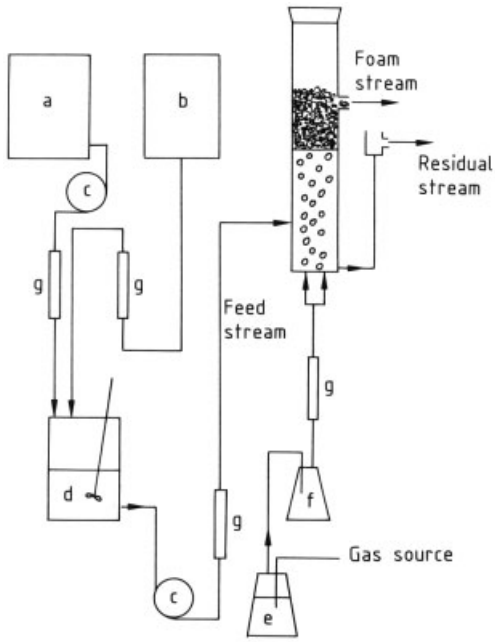

Figure 38. Schematic of continuous-flow foam fractionation column [72]
a) Ion storage; b) Surfactant storage; c) Pump; d) Mix tank; e) Saturator; f) Moisture trap; g) Flow meter

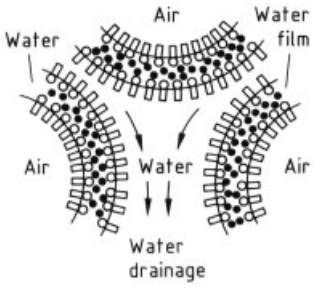

Figure 37. Scheme of selective ion flotation by foam fractionation
○ = Surfactant; ● = Ion removed

5. P. Somasundaran (ed.): *Advances in Mineral Processing,* Society of Mining Engineers (SME), New York 1986.
6. A. M. Gaudin: *Flotation,* McGraw-Hill, New York 1957.
7. V. I. Klassen, V. A. Mokrausov: *An Introduction to the Theory of Flotation,* (Transl. J. Leja, G. W. Poling), Butterworths, London 1963.
8. K. L. Sutherland, I. W. Wark: *Principles of Flotation,* Australian Institute of Minerals and Metals, Melbourne 1955.
9. B. Yarar: "Gamma Flotation: A New Approach to Flotation, Using Liquid – Vapor Surface Tension Control," *Proc. 2nd Latin American Flotation Congress,* **vol. 1,** Concepcion, Chile,Aug. 19 – 23, 1985, pp. 2.1 – 2.36.
10. B. Yarar, J. Kaoma, *Colloids and Surfaces* **11** (1984) no. 3, 429 – 437.
11. D. W. Fuerstenau, "Activation and De-activation in the Flotation of Sulfide Minerals," in P. E. Richardson et al. (eds.): *The Physical Chemistry of Mineral-Reagent Interactions in Sulfide Flotation,* USBM, Information Circular no. 8818, 1980, pp. 100 – 120.
12. J. Leja: *Surface Chemistry of Froth Flotation,* Plenum Publ., New York 1982, p. 587.
13. R. C. Weast (ed.): *Handbook of Chemistry and Physics,* 60th ed. CRC Press, New York 1979, p. F – 43.
14. H. Lange in M. J. Schick (ed.): *Nonionic Surfactants,* Marcel Dekker, New York 1967, p. 459.
15. M. H. Buckenham, J. H. Schulman, *Trans. Soc. Min. Eng. AIME* **226** (1963) 1 – 5.
16. Y. Zimmels: ref [3], p. 135.

Table 7.

Flotation circuit*		Yield, wt%	Product, ash%	Tailings, ash%
Single-stage flotation	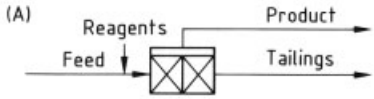	59.6	11.6	27.0
Desliming	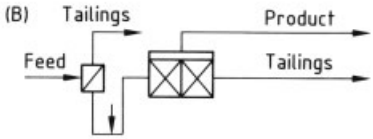	60.0	8.2	31.3
Separate conditioning of the coarse and fine size fractions	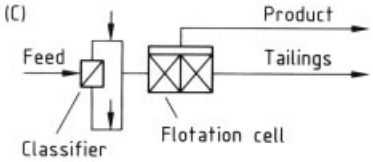	51.0	13.0	23.5
Two-stage conditioning	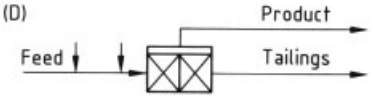	65.0	12.8	28.6
Split feed flotation	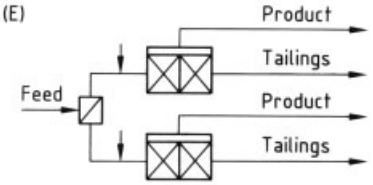	75.2	11.2	39.5
Two-stage reagent addition	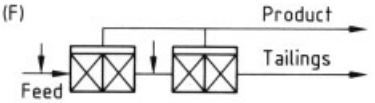	79.6	11.0	43.6
Reflotation of classified tailings	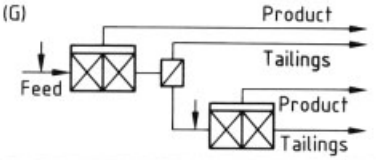	81.0	9.9	58.0

*Symbols defined in (c)

17. J. Laskowski, J. A. Kitchener, *J. Colloid Intface Sci.* **29** (1969) 670–679.
18. M. C. Fuerstenau, K. L. Clifford, M. C. Kuhn, *Int. J. Min. Process.* **1** (1974) 307–318.
19. D. W. Fuerstenau, T. W. Healy, P. Somasundaran, *Trans. Soc. Min. Eng. AIME,* **229** (1964) 321–325.
20. M. C. Fuerstenau, J. D. Miller, M. C. Kuhn: *Chemistry of Flotation,* Society of Mining Engineers (SME), New York 1985.
21. W. Stumm, J. J. Morgan: *Aquatic Chemistry,* J. Wiley & Sons, New York 1970, p. 164, p. 171.
22. R. M. Garrels, C. L. Christ: *Solutions Minerals and Equilibria,* Freeman, San Francisco 1965.

23. J. N. Butler: *Ionic Equilibrium, a Mathematical Approach,* Addison-Wesley, Reading, Mass., 1964.
24. O. Griot, J. A. Kitchener, *Trans. Faraday Soc.* **61** (1965) no. 509, 1026 – 1031.
25. B. Yarar: "Polymeric Flocculants and Selective Flocculation: An Overview," in K. L. Mittal, E. J. Fendler (eds.): *Solution and Surfactant Chemistry, Theoretical and Applied Aspects,* Plenum Publ. New York 1982, pp. 1333 – 1364.
26. B. Yarar: "The Surface Chemical Mechanism of Calcite-Colemanite Separation by Flotation" in J. M. Barker, S. J. Lefond (eds.): *Borates: Economic Geology and Production,* AIME, New York 1985, pp. 221 – 233.
27. O. Mellgren, *Trans. Soc. Min. Eng.-AIME* **235** (1966) 46 – 60.
28. A. S. Peck, M. E. Wadsworth: "Infrared Studies of Oleic Acid and Sodium Oleate Adsorption on Fluorite, Barite and Calcite," U.S. Bureau of Mines (USBM) Report of Investigation no. 6202, 1963.
29. S. Chander, D. W. Fuerstenau, D. Stigter: "On Hemimicelle Formation at Oxide – Water Interfaces," in R. H. Otewill et al. (eds.): *Adsorption from Solution,* Academic Press, London 1983, pp. 197 – 210.
30. H. E. Garrett: *Surface Active Chemicals,* Pergamon, London 1972, p. 150.
31. R. W. Slater, J. P. Clark, J. A. Kitchener: "Chemical Factors in the Flocculation of Mineral Slurries with Polymeric Flocculants," *Proc. 8th Int. Mineral Processing Congress,* Leningrad 1968.
32. J. Laskowski: *Physical Chemistry in Mineral Processing Technology, (Polish Text),* Wydawnictwo "Slask", Katowice, Poland 1969, p. 215.
33. H. R. Kruyt (ed.): *Colloid Science,* Elsevier, Amsterdam 1952, p. 126.
34. A. Sheludko: *Colloid Chemistry,* Elsevier, Amsterdam 1966, p. 141.
35. R. J. Hunter: *Zeta Potential in Colloid Science,* Academic Press, London 1981, p. 11.
36. I. Iwasaki, S. R. B. Cooke, A. F. Colombo: "Flotation Characteristics of Goethite," USBM, Report of Investigation, no. 5593, 1960.
37. D. W. Fuerstenau, *Trans. Soc. Min. Eng.-AIME,* **208** (1957) 834 – 836.
38. A. Yucesoy, B. Yarar, *Trans. Inst. Min. Metallo.* **83** (1974) C 96 – C 100.
39. J. Laskowski, *Miner. Sci. Eng.* **6** (1974) no. 4, 223 – 235.
40. A. D. Read, J. A. Kitchener: "The Thickness of Wetting Films," *Wetting, Soc. Chem. Ind. Monograph no. 25,* Gordon-Breach, New York 1967, pp. 300 –317.
41. J. H. Schulman, J. Leja, *Kolloid-Z.* **136** (1954) 107 – 119.
42. B. Yarar, J. Leja, *Trans. Soc. Min. Eng.-AIME,* **272** (1983) 1978 – 1983.
43. W. A. Zisman: "Relation of the Equilibrium Contact Angle to Liquid and Solid Constitution," in K. J. Mysels et al. (eds.): *20 Years of Surface and Colloid Chemistry, The Kendall Award Addresses,* Amer. Chem. Soc., Washington, D.C., 1973, pp. 109 – 157.
44. B. Yarar, J. Kaoma, *Trans. Soc. Min. Eng.-AIME* **276** (1985) 1875 – 1878.
45. J. Kaoma, B. Yarar: "Correlation of the Critical Surface Tension of Wetting and the Non-stoichiometric Composition of Molybdenum Sulfide," in H. F. Barry, P. C. H. Mitchell (eds.): *Proc. 4th Conference on the Chemistry and Uses of Molybdenum,* Climax Molybdenum Co., Ann Arbor, Mich., 1982, pp. 117 – 122.
46. J. Laskowski: "The Relationship between Flotability and Hydrophobicity," ref. [5], pp. 189 – 208.
47. B. Yarar, G. P. Hemphill: "Size-Surface Energy Relationships in the Flotation-Upgrading of Oil Shale," *Proc. 1983 Eastern Oil Shale Symp. Institute of Mining and Minerals Research-Kentucky,* Lexington, Ky., 1983, pp. 241 – 248.
48. G. P. Hemphill, B. Yarar: "Use of Critical Surface Tension of Wetting in the Beneficiation of Oil Shales," in F. A. Curtis (ed.): *Proc. Energy-84,* Pergamon, Toronto 1984, pp. 111 – 116.
49. J. A. Finch, G. W. Smith, *Can. Metall. Q.* **14** (1975) no. 1, 44 – 51.
50. D. T. Hornsby, J. Leja, *Colloids and Surfaces* **7** (1983) 339 – 349.
51. M. R. Smith, R. J. Gochin: ref. [1], pp. 166 – 201.
52. A. L. Mular, M. A. Anderson: *Design and Installation of Concentration and Dewatering Circuits,* Society of Mining Engineers (SME), Littleton, Colo., 1986, p. 588.
53. G. W. Poling: "Selection and Sizing of Flotation Machines," in A. L. Mular, R. B. Bhappu (eds.): *Mineral Processing Plant Design,* 2nd ed., Society of Mining Engineers (SME)-AIME, New York 1980, pp. 887 – 906.

54. I. S. Blagov et al.: "State and Development of Coal Flotation in the USSR," *Proc. 9th Int. Coal Preparation Congress,* New Delhi, India, 1982, pp. C 1 – C 5.
55. G. Barbery: "Engineering Aspects of Flotation in the Minerals Industry: Flotation Machines, Circuits and their Simulation," *Scientific Basis of Flotation, NATO-Advanced Study Institute on Scientific Basis of Flotation, Preprints,* Cambridge, UK, 1982.
56. C. C. Harris: ref. [35], p. 753.
57. A. E. Zhangozhina et al., *Chem. Abstr.* **102** (1985) 133410 e.
58. D. A. Wheeler: "Column Flotation: The Original Column," ref. [9], pp. c.1 – c.28.
59. G. S. Dobby, J. A. Finch: " Flotation Column Scale-Up and Simulation," *Proc. 17th Annual Meeting of Canadian Mineral Processors,* Jan. 22 – 24, Can. Inst. Min. Metall., Ottawa 1985, pp. 614 – 638.
60. K. S. Narasimhan et al., *Eng. Min. J.* **173** (1972) no. 5, 84 – 85.
61. M. C. Fuerstenau (ed.): *Flotation,* **(A. M. Gaudin Memorial vol.) AIME, Vols. 1 and 2,** New York 1976.
62. S. G. Malghan: "Typical Flotation Circuit Configurations," ref. [70], pp. 76 – 92.
63. B. Yarar, B. C. Haydon, J. A. Kitchener, *Trans. IMM (London)* **78** (1969) C 181 – C 184.
64. D. Toperi, R. Tolun, *Trans. IMM (London)* **78** (1969) C 191 – C 197.
65. R. Woods, P. E. Richardson: "The Flotation of Sulfide Minerals: Electrochemical Aspects," ref. [5], pp. 154 – 170.
66. E. W. Betz: "Beneficiation of Brazilian Phosphates" in J. Laskowski (ed.): *Proc. 12th International Mineral Processing Congress,* vol. 2 B, Elsevier, Amsterdam 1981, pp. 1846 – 1874.
67. K. Kiukkola et al.: "Selective Flotation of Apatite from Carbonatite Glimmerite Ore at the Silinjarvi Mine of Kemira Oy, Finland," *Proc. 14th Int. Mineral Processing Congress,* **vol. 5,** CIM, Toronto 1982, pp. 7.1 – 7.15. J. Li: "The Practice of Concentration and Ways to Increase Grade and Recovery of the Graphite Concentrate in Washu Graphite Mine, Shandong, China," ref. [75], pp. 9.1 – 9.10.
68. G. G. Strathdee et al.: "The Processing of Potash Ore by PCS," ref. [75], pp. 12.1 – 12.20.
69. J. W. Leonard (ed.): *Coal Preparation,* AIME, New York (1979) p. 9-1.
70. G. Ozbayoglu: "Coal Flotation," ref. [1], p. 76.
71. T. M. Plouf, *Min. Eng.* **32** (1980) 1218 – 1224.
72. R. B. Grieves: "Surfactant – Anion Interactions in Foam Fractionation," ref. [3], pp. 357 – 374.
73. R. Lemlich (ed.): *Adsorptive Bubble Separation Techniques,* Academic Press, London 1972.
74. D. W. Downey, E. W. Berg: "Ion Flotation of Platinum Group Metals," ref. [3], pp. 375 – 394.
75. D. J. Wilson: "Topics in Precipitate and Floc Flotation," ref. [3], pp. 395 – 414.
76. H. L. Shergold, O. Mellgren, *Trans. IMM (London)* **78** (1969) C 121 – C 132.
77. H. L. Shergold, R. Stratton-Crawley, *Colloids and Surfaces* **3** (1981) 253 – 265.
78. F. D. Schowengerdt, D. J. Spottiswood, R. Sen, B. Yarar: "Surface Behavior of Minerals in Flotation: An Auger Electron Spectroscopic Analysis," *Proc. 15th Int. Mineral Processing Congress,* vol. 1, Cannes 1985, pp. 49 – 58.
79. D. R. Gaskell et al. (eds.): *The Reinhardt Schuhmann International Symposium on Innovative Technology, and Reactor Design in Extraction Metallurgy,* The Metallurgical Society Inc., Warrendale, Pa., 1986, p. 159.

Flour → Cereals and Cereal Products
Flouranthene → Hydrocarbons

Mixing, Introduction

For unit operations in mixing see → Stirring; → Mixing of Highly Viscous Media; → Mixing of Solids.

Marko Zlokarnik, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Mixing of Homogeneous Miscible Substances 1
3. Mixing in Heterogeneous Systems 2
4. Mixing of Pastes and Granular Materials 2

1. Introduction

Mixing, stirring, and kneading have been used by man since time immemorial in the preparation of food and drink, and belong to the classical unit operations used in the most diverse branches of industry–production of building materials, glass, chemicals, food processing, etc. In the chemical industry, this technology is of special importance: to carry out a chemical reaction, the solutions and mixtures must be prepared in such a way that the reagents have a suitable form for reaction. During the reaction itself, the mixing operation must ensure that the reagents stay in close contact with each other; this is particularly important when the reagents have a different aggregate state.

Table 1. Classification of mixing operations according to the aggregate state of the predominant component

Aggregate state	Operation	Most common mixing equipment
Gaseous *	mixing	mixing chamber, nozzle
Liquid	stirring	stirrer
Paste	kneading	kneader, screw extruder
Solid (granular)	mixing	mixer

* The mixing of gases is treated under → Continuous Mixing of Fluids.

The operations used for mixing must first be classified according to the state of aggregation of the predominant component; as a rule, the resulting mixture has this state, too. From the process engineering point of view, stirring a gas into a liquid is not the same as mixing a liquid into a gas stream. Even the term liquid state must be made more specific, since low-viscosity liquids place drastically different demands on mixing equipment than viscous, pastelike substances. Bearing these considerations in mind, mixing operations can be divided into the groups listed in Table 1.

2. Mixing of Homogeneous Miscible Substances

A mixture that is homogeneous at the molecular level results only when mixing gases or miscible liquids. Here the mixing operation can be divided into two steps: (1) pre-mixing taking place through the large-scale convective transport of eddies and (2) elimination of concentration differences through purely diffusive transport. The former is governed by the physical properties as well as by the type of mixing device and the mixing energy expenditure; it is scale-dependent. The better the pre-mixing, i.e., the smaller the eddies produced, the faster the following mixing process can proceed. The latter is governed by those material properties of the system that influence diffusion (viscosity, diffusion coefficient) and is only rate-limiting if the pre-mixing was not intensive enough. Good mixing devices (nozzles, stirrers causing a pronounced bulk circulation) usually function so well that the process of diffusion has no effect on the homogenization time (no effect of the Schmidt number *Sc* on the homogenization characteristic, → Stirring, Chap. 2.

In the last few years, *micromixing* has received special attention; this refers to the production of extremely small eddies with the purpose of mixing the reaction components as fast as possible. For example, micromixing plays a key role in competitive reactions of the type

$$A + B \longrightarrow C \quad (1)$$
$$A + 2B \longrightarrow D \quad (2)$$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_24

In this case, the selectivity of the desired Reaction (1) can be increased through a fast reduction of the reagent concentrations. If the reaction, however, is of the competitive-consecutive type,

$$A + B \longrightarrow C \quad (3)$$
$$A + C \longrightarrow D \quad (4)$$

then it is also important to avoid backmixing of the reagents with the product C. In such cases a plug-flow reactor is used, and mixing is performed with a nozzle or an in-line mixer.

3. Mixing in Heterogeneous Systems

The main objective of the mixing operation in the spraying of liquids in a gas and in the dispersion of immiscible liquids or gases in a liquid is to create as large an interfacial area as possible. In these operations, however, fine primary droplets or gas bubbles often have the tendency to coalesce; i.e., a part of the work done in forming the interfacial area is wasted. During the production of fine bubbles or droplets in stirred vessels, the size distribution of the bubbles or droplets is the result of a continuous process of dispersion and coalescence.

The*coalescence process* depends to a great extent on the state of flow in the mixing device, thus requiring care in equipment scale-up. To allow reliable conclusions, design guidelines for mixing devices for these operations must be based on experiments in several different scales. The exact physical properties that determine coalescence are still not known: some substances suppress coalescence strongly, even in extremely small concentrations (a few ppm), while others increase the coalescence tendency. Especially in the dispersion of gases in liquids, the coalescence behavior of the system must be determined through appropriate experiments in the laboratory.

4. Mixing of Pastes and Granular Materials

Pastes and powders or granular materials are usually mixed to achieve the best possible blending. Here, however, only a stochastically homogeneous mixture is possible, a mixture that depends strongly on the physical properties of the substances to be mixed (viscosity and rheological behavior for pastes, size and morphology for granular solids). In the mixing of granular materials, *mixture separation,* determined by the state of flow in the device and therefore scale-dependent often presents a problem. Since the sampling, analysis, and the statistical evaluation of experimental data are problematic, there are comparatively few reliable design and scale-up guidelines for these mixing devices. In the future, significantly more intensive research activity in this field is necessary.

Stirring

See also → Mixing, Introduction

MARKO ZLOKARNIK, Grillparzer Strasse 58, A-8010 Graz

1. Stirring Operations, Stirring Vessels, and Stirrer Types 3
1.1. Stirring Operations 3
1.2. Stirring Vessels and Auxiliary Equipment 3
1.3. Operation of Individual Types of Stirrers 6
1.4. Mechanical Stress Exerted by the Stirrer 7
2. Homogenization of Liquid Mixtures 8
2.1. Problem Definition 8
2.2. Mixing-Time Characteristic 8
2.2.1. Definition and Measurement Method 8
2.2.2. Mixing-Time Characteristic at $\Delta\varrho \simeq 0$ and $\Delta\nu \simeq 0$ 8
2.2.3. Mixing-Time Characteristic at $\Delta\varrho \neq 0$ and $\Delta\nu \neq 0$ 10
2.2.4. Mixing-Time Characteristics in Non-Newtonian Liquids 10
2.3. Power Characteristic 11
2.3.1. Definition and Measurement Method 11
2.3.2. Power Characteristics in Homogeneous Systems 11
2.4. Optimization: Minimum Work of Mixing 12
2.5. Homogenization in Storage Tanks . 14
2.6. Liquid Throughput of Stirrers . . . 15
2.7. Vortex Depth in Unbaffled Stirred Tanks 15
3. Intensification of Heat Transfer . . . 16
3.1. Problem Definition 16
3.2. Heat-Transfer Characteristics 16
3.2.1. Definitions 16
3.2.2. Heat-Transfer Characteristics in the Range Re > 200 16
3.2.3. Heat-Transfer Characteristics in the Range Re < 200 17
3.3. Stirrer Power During Heat Transfer 18
3.4. Optimum Stirring Conditions for the Removal of Reaction Heat 19
4. Suspension of Solids in Liquids (S/L Systems) 20
4.1. Problem Definition 20
4.2. Classification of the Suspension Conditions 20
4.3. Suspension Characteristic 20
4.3.1. Suspension Characteristics with $d_{\mathbf{p}}$ as the Characteristic Particle Dimension 21
4.3.2. Suspension Characteristics with $w_{\mathbf{ss}}$ as the Characteristic Particle Property . . 22
4.4. Power Requirements Upon Suspension of Solids 22
4.5. Scale-up of the Solids Suspension Process 24
4.6. Effect of Geometric and Apparatus-Related Factors on the Suspension . 24
4.7. Mass Transfer in the Solid – Liquid (S/L) System 24
5. Dispersion in Liquid – Liquid (L/L) Systems 25
5.1. Lowest Stirrer Speed for Dispersion 25
5.2. Dispersion Characteristics 26
5.3. Stirrer Power for Dispersion 27
5.4. Scale-Up of the Dispersion Processes 27
5.5. Mass and Heat Transfer Upon Dispersion 27
6. Dispersion in Gas – Liquid (G/L) Systems 28
6.1. Physical Fundamentals of Mass Transfer 28
6.2. Mass-Transfer Characteristics in the G/L System 29
6.2.1. Sorption Characteristics in Coalescence-Bound Systems (Water) 30
6.2.2. Sorption Characteristics in Coalescence-Inhibited Systems 31
6.2.3. Sorption Characteristics in Non-Newtonian Material Systems . . 31
6.3. Stirrer Power in Gassed Liquids . . 31
6.4. Flooding Point 32
6.5. Special Gas – Liquid Contacting Techniques 33
6.5.1. Hollow Stirrers 33
6.5.2. Surface Aerators 34
6.5.3. Gas Spargers and Injectors 35
7. References 38

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_25

Symbols

a	interfacial area per unit volume, $a \equiv A/V$
a	thermal diffusivity; $a \equiv k/(\varrho C_p)$
A	area, interfacial area
c	concentration
c_s	saturation concentration
C_d	drag coefficient of a sphere in a fluid flow
c_p	heat capacity at constant pressure
d	stirrer diameter
d_b	bubble diameter, usually represented by d_{32}
d_{32}	mean gas bubble or liquid droplet diameter ("Sauter diameter")
d_p	particle or droplet diameter
D	inside tank diameter
D	diffusion coefficient
G	mass flow (rate of mass transfer, oxygen uptake)
h	stirrer distance from bottom of the vessel (bottom clearance); film heat transfer coefficient
H	total liquid depth (liquid height) in vessel
k	thermal conductivity
k_G	gas-side mass-transfer coefficient
k_L	liquid-side mass-transfer coefficient
$k_L a$	volume-related overall masstransfer coefficient
n	stirrer speed
P	stirrer power
q	volume throughput
q'	liquid throughput produced by the stirrer
Q	heat flow (rate of heat transfer)
R	heat of reaction
S	surface; cross-sectional area
u	tip velocity ($u = \pi\, n\, d$)
U	overall heat transfer coefficient
v	velocity; superficial velocity
v_G	superficial gas flow velocity ($v_G \propto q_G/D^2$)
V	liquid volume (ungassed)
$\dot{\gamma}$	shear rate
δ	thickness
ε	mixing power per unit mass $\varepsilon \equiv P/\varrho V$
θ	mixing time
μ	dynamic viscosity
ν	kinematic viscosity
ϱ	density
ϱc_p	heat capacity per unit volume at constant pressure
σ	interfacial, surface tension
φ_v, φ_m	volume or mass fraction

Dimensionless numbers

Ar	Archimedes number	$Ar \equiv Re^2/Fr'$
Fr	Froude number	$Fr \equiv n^2 d/g$
Fr'	Froude number, modified	$Fr' \equiv Fr\, \varrho/\Delta\varrho$
Ga	Galilei number	$Ga \equiv Re^2/Fr$
Ne	Newton number	$Ne \equiv P/(\varrho\, n^3\, d^5)$
Nu	Nusselt number	$Nu \equiv h_i\, D/k$
$n\theta$	Mixing number	
Pr	Prandtl number	$Pr \equiv \nu/a = c_p\, \mu/k$
Pe	Péclet number	$Pe \equiv RePr \equiv n\, d^2/a = n\, d^2\, \varrho c_p/k$
Q	Throughput number	$Q \equiv q/(n\, d^3)$
Re	Reynolds number	$Re \equiv n\, d^2\, \varrho/\mu$
Wi	Weißenberg number	$Wi \equiv N_1/\tau$
We	Weber number	$We \equiv \varrho\, n^2\, d^3/\sigma$
Sc	Schmidt number	$Sc \equiv \nu/\mathrm{D}$
Sh	Sherwood number	$Sh \equiv k_L\, d_p/\mathrm{D}$
St	Stanton number	$St \equiv Nu/RePr \equiv h/(v\, \varrho c_p)$
Vis	Viscosity number	$Vis \equiv \mu_w/\mu$
Y	Sorption number	see definition

1. Stirring Operations, Stirring Vessels, and Stirrer Types

1.1. Stirring Operations

When the liquid component predominates in the mixture (→ Mixing, Introduction), the mixing operation is termed *stirring*, and a stirrer is used to perform it. The following five stirring operations may be defined:

1) Homogenization, i.e., equalization of differences in concentration and temperature
2) Intensification of heat transfer between the liquid and the heat-exchange surface
3) Suspension of a solid in the liquid, or slurry formation
4) Dispersion of two immiscible liquids
5) Dispersion or sparging of a gas in the liquid (gas – liquid contacting)

Homogenization refers to the formation of a uniform phase out of several miscible liquids or to the elimination of concentration and temperature gradients during a chemical reaction in the liquid phase. (In the food industry, the same term is also used for emulsification under extreme shear conditions, e.g., the homogenization of milk.)

Intensification of heat transfer in mixing vessels – especially in highly viscous liquids – can be an important stirring operation, particularly in the case of a strongly exothermic chemical reaction, for example, block polymerization. In this case, stirring aims to reduce the boundary layer at the vessel wall, as well as to provide for the motion of the liquid to and from the heat-exchange surface.

When particulate materials are to be dissolved in a liquid, or when a solid-catalyzed reaction is to take place, the stirrer must make a *suspension* of the particles in the liquid to allow their entire surface to take part in the process. For continuous processes, a quasi-uniform distribution of the solids in the entire liquid volume is required; this allows the solid particles to be transferred from stage to stage together with the liquid (e.g., in crystallization cascades). In this case, the solids are exposed to high mechanical stress, which can cause their abrasion.

In order to enhance mass transfer between two immiscible liquids, the phase with the smaller volume is dispersed in the other, i.e., a *liquid – liquid dispersion* is made. When surfactants are added, a more or less stable emulsion can be produced.

Stirrers are also used to perform gas – liquid contacting to make *gas – liquid dispersions.* The process is also called *sparging.* Different stirring operations must often be performed simultaneously; an example is solid-catalyzed hydrogenation, in which the stirrer must both distribute the gas (hydrogen) in the liquid, as well as suspend the solid catalyst (e.g., Raney nickel).

1.2. Stirring Vessels and Auxiliary Equipment

Mixing operations are mainly performed in agitated vessels, but storage tanks and pipes can also be employed (cf. Section 2.6). The design of mixing vessels is governed by standards (DIN 28 136, ASME Code Section VIII). The interior auxiliary equipment consists of baffles, feed and drain pipes, coils, and probes (e.g., thermometers, level indicators). All of these inserts can affect stirring.

If an axially positioned stirrer is operated in a *vessel without inserts*, the liquid is set into rotation, and a vortex is formed. The vortex can reach the stirrer and cause gas to be entrained in the liquid. This is generally undesirable because it results in an extraordinarily high mechanical stress in the stirrer shaft, as well as in its bearings and seal, due to the absence of the liquid bearing, often leading to the destruction of the stirrer. Even when vortex formation causes no gas entrainment, bulk rotation of the liquid is always undesirable when a two-phase system with different densities is concerned since the centrifugal force counteracts stirring.

Bulk rotation of the liquid in cylindrical vessels is prevented by the installation of baffles. Fully effective baffling is achieved with four baffles having a width of $D/10$ (D = inside diameter of vessel), which are vertically arranged along the entire vessel wall. In order to avoid dead volume behind the baffles, the baffle width is reduced to $D/12$ and they are set at a clearance of $D/50$ from the wall. The baffles are usually attached to the vessel wall by means of welded brackets (Fig. 1 A). In enamel-coated vessels, they are attached to the cover. When this is not possible (glass vessels, wooden vats), they are made in the form of a basket with pressure-fitted rings (Fig. 1 B).

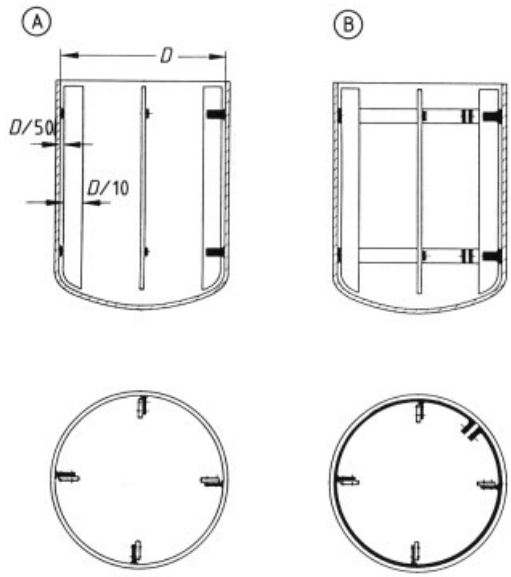

Figure 1. Baffle designs
A) Normal design; B) Design for glass and coated vessels (baffle basket with pressure-fitted ring)

Baffles are not necessary when stirring is carried out in vessels with rectangular cross section

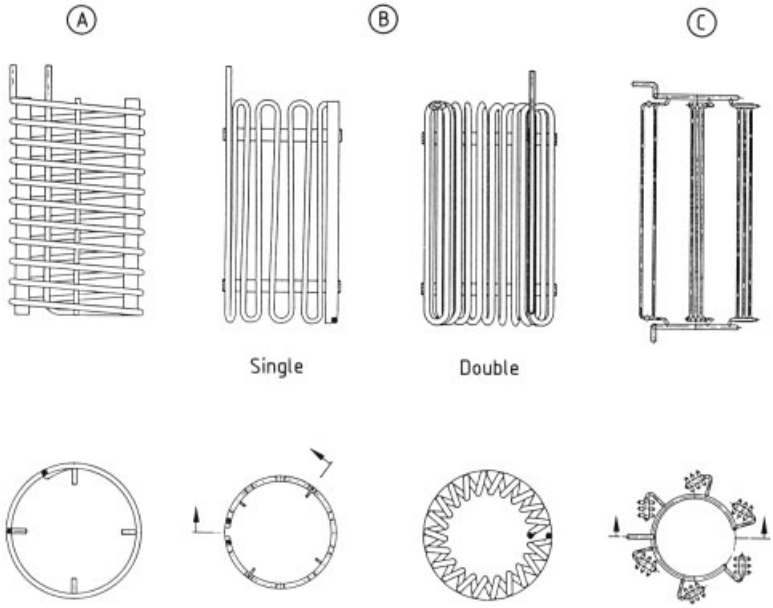

Figure 2. Tube coil designs
A) Spiral coil; B) Meander coils (cooling basket); C) Tube bundles

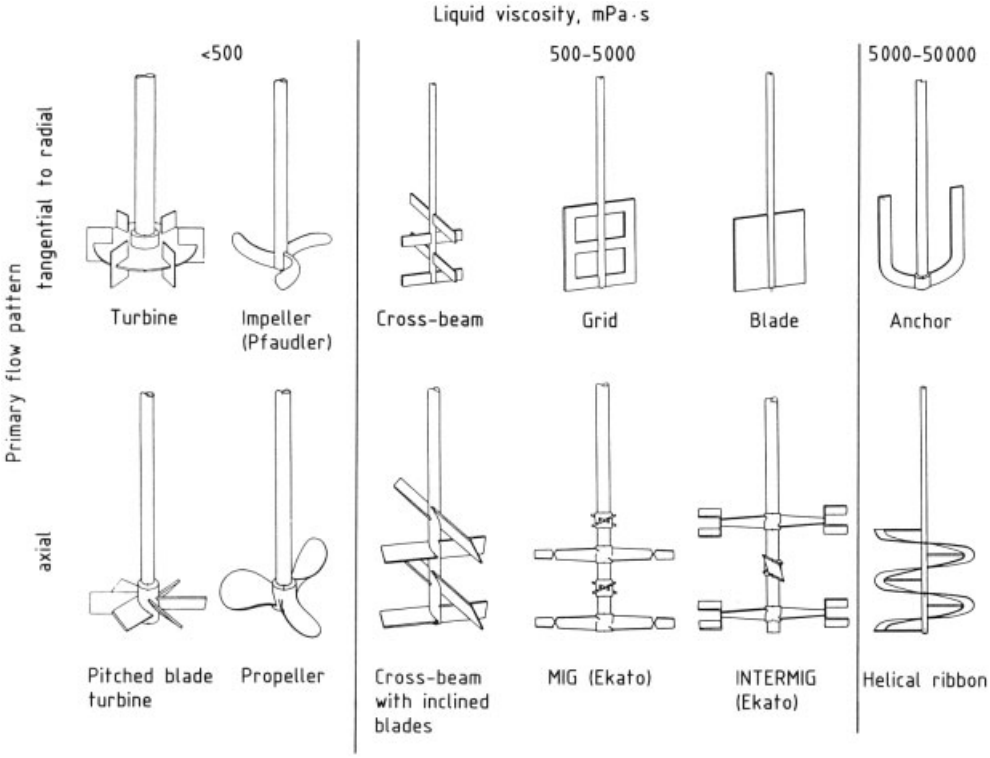

Figure 3. Common stirrer types

(pits), or when the stirrer is introduced into the vessel laterally. When the stirring is weak, liquid rotation can be prevented even in cylindrical vessels by installing the stirrer off-center and/or at an angle to the axis. In this case, however, uneven mechanical stress in the stirrer shaft must be accepted.

In order to supply or remove large amounts of heat, stirring vessels are equipped with *coils.* A spiral coil (Fig. 2 A) is only effective with stirrers that generate an axial flow pattern, since these produce good liquid circulation between the coil and the wall. In contrast, the liquid circulation produced by radial stirrers is strongly deflected by the spiral coil, so that the flow through the annulus between the coil and wall is insufficient. For these types of stirrers, it is better to arrange the coil vertically along the vessel wall (meander coil, Fig. 2 B). This arrangement does not deflect the radial flow pattern, but prevents bulk rotation of the liquid to such an extent that baffles are often unnecessary. The heat-exchange tubes can also be arranged into bundles and installed instead of baffles (Fig. 2 C).

1.3. Operation of Individual Types of Stirrers

The mixing operations described in Section 1.1 naturally cannot all be performed with a single stirrer type; therefore, a wide variety of stirrers exists. There are stirrers suitable to a specific mixing operation and a given material system. The following discussion is limited to those stirrer types which are most widely used in the chemical industry, and for which established design guidelines exist.

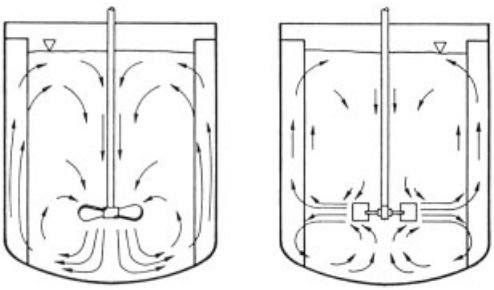

Figure 4. Flow patterns in a baffled tank, generated by an axial-flow propeller and a radial-flow turbine stirrer

These stirrer types are arranged in Figure 3 according to the predominant flow pattern they produce, as well as to the range of viscosities over which they can be effectively used. The flow patterns generated by stirrers causing radial and axial fluid motion are shown in Figure 4.

The *turbine stirrer* (Rushton turbine, six blades on a disk) is the only high-speed stirrer that sets the fluid in *radial motion* or, at higher viscosities, in *tangential motion.* This type is only effective with low-viscosity liquids and baffled vessels. In this case, the diameter ratio D/d (D = vessel diameter; d = stirrer diameter) ranges from 3 to 5. During rotation, the turbine stirrer causes high levels of shear and is well suited to dispersion processes. The *impeller stirrer* was developed for use in enamel-coated vessels, and thus has rounded stirring arms. It is used in conjunction with small clearances from the bottom and with a ratio $D/d = 1.5$, either with or without baffles. It can also operate with strongly fluctuating fill levels (e.g., during vessel discharging) because it mixes even small amounts of liquid very well. The *cross-beam*, *grid,* and *blade stirrers* belong to the group of low-speed stirrer types and are used with $D/d = 1.5 - 2$. They can operate with baffles or, for viscous fluids, without, and are especially well suited to homogenization. As a rule, the low-speed *anchor stirrer* is operated at very small clearances from the wall ($D/d \simeq 1.05$) and is only appropriate for enhancing heat transfer in highly viscous liquids.

The *paddle stirrer* with pitched blades (pitched blade turbine) and particularly the *propeller stirrer* belong to the group of high-speed mixers that generate an *axial flow pattern.* As a rule, both are used with low-viscosity liquids and in baffled vessels. They are well suited to homogenization and suspension of solids. In order to enhance the axial flow component in more viscous media and/or for $H/D > 1$ (e.g., in fermenters; H = liquid height) multistage stirrers with pitched stirring surfaces are required. This category includes the *cross-beam stirrer with pitched beams* and the *MIG* and *INTERMIG* stirrers of the Ekato company, Schopfheim, Germany. These stirrers are operated at low speeds, with $D/d \simeq 1.5$ (with baffles) or $D/d = 1.1$ (without baffles). Mixing operations for these types include homogenization, suspension of solids, and dispersion. The very low-speed *helical rib-*

bon stirrer is used with small wall clearances ($D/d \simeq 1.05$) and operated in such a way that it drives the liquid downwards along the wall. Under these conditions, it is the best-suited of all stirrer types for the homogenization of highly viscous liquids.

In addition to these often used types, a large number of special designs also exists. Of these, only three should be mentioned: the stirrer operating on the rotor – stator principle, the toothed disk, and the hollow stirrer.

In *rotor – stator stirrers* (Fig. 5), the rotor consists of a blade or paddle stirrer enclosed by a ring of baffles (stator). As a result, high levels of shear are exerted on an extremely small volume. When the stirrer consists of a planar *toothed disk* (Fig. 6), the liquid is accelerated radially in a thin ring away from the center, and then quickly decelerated. High levels of shear can thus be achieved, even without a stator ring or baffles. Both of these stirrer types are particularly well suited to emulsification and dispersion over a wide range of viscosities, e.g., in the production of paint pigments.

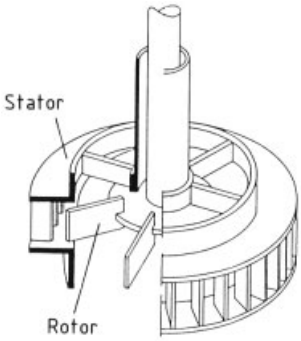

Figure 5. A rotor – stator stirrer

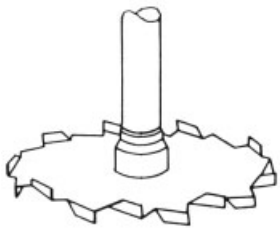

Figure 6. Disperser disk

In *hollow stirrers*, the stirrer head is hollow and is connected through a hollow shaft to the gas-filled space above the liquid surface. The suction generated behind the stirrer edges during rotation can thus be used to supply a gas to the liquid. As a single unit combining both stirrer and gas supply, hollow stirrers are well suited for enhancing mass transfer in gas-liquid systems. Figure 7 shows the so-called *tube stirrer*, a simple, effective hollow stirrer. All hollow stirrers operate at high speeds and are used in baffled vessels with $D/d = 3 - 5$.

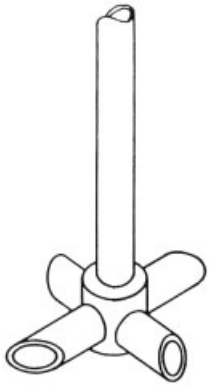

Figure 7. Hollow stirrer (tube type)

1.4. Mechanical Stress Exerted by the Stirrer

A distinction has to be made between the aspects which concern construction of a device and those which are of concern to users of the device. Information on the mechanical forces, which operate on the stirrer, the shaft, the shaft bearings, and the shaft seal, together with the critical stirrer speed etc., can be obtained from the brochures of the stirrer manufacturers. Information on the stress on baffles and on stirrer heads as well on the wear of stirrer heads is summarized in [1]. Here, only the shear stress on the particulate material being mixed is considered. Shear stress can be advantageous, e.g., in dispersion processes in gas – liquid (G/L) and liquid – liquid (L/L) systems. It can, however, also be undesirable, e.g., in many agglomeration and crystallization processes, particularly in biotechnological systems. Mammalian cells, for example, can be destroyed even at low shear forces. The morphological structure of cell clusters, important for product formation in fungal cultures (penicillin and citric acid), can be negatively influenced by

shear forces, as can the biofilm formation of microorganisms and enzymes on solid carriers (immobilization of biomaterial). It is therefore not surprising that an increasing number of papers have been published on this topic in the recent biotechnological literature [1].

It had been found that the destruction of clay flocks, as a representative model system for biological material, depended strongly on the power per unit volume P/V_s (in kW/m$_s^3$) where V_s is the volume swept by the stirrer, $V_s = \pi d^2 h$ (h = stirrer height), assuming that most of the power has been dissipated in this volume [2]. The following correlation has been obtained for six different stirrer types for the flock diameter d_f (μm):

$$d_f = 130\,(P/V_s)^{-1/3} \tag{1}$$

Propeller stirrers need higher rotational speeds than large-surface stirrers (cross-beam or anchor stirrers) to achieve the same P/V values. They therefore produce higher shear rates and should not be used for shear sensitive material.

Equation (1) is consistent with the Kolmogorov's theory of locally homogeneous and isotropic turbulence [1]. In the turbulent flow range, turbulence exists on the micro- and the macroscale. The microscale λ is responsible for the shear stress, which for a particle size, d_p, which is significantly smaller than λ (λ/d_p = 5 – 25), can influence the particle size of the dispersed phase.

According to Kolmogorov the microscale λ of turbulence is described by $\lambda = (\nu^3/\varepsilon)^{1/4}$ (ε is the power consumption per unit mass, $\varepsilon = P/\varrho V$, and ν is the kinematic viscosity of the liquid). The shear condition on a small scale is similar to that on a large scale when P/V, the physical properties (ν) and the geometry are kept constant:

$$\lambda = \left(\varepsilon/\nu^3\right)^{1/4} = \text{idem} \tag{2}$$

2. Homogenization of Liquid Mixtures

2.1. Problem Definition

Homogenization deals with the mixing of miscible liquids to obtain the desired degree of homogeneity, or with the maintenance of homogeneity for the purpose of carrying out a reaction under well-defined conditions. The question here is what type of stirrer and what mixing conditions to adopt, so that the mixing operation can be performed with the minimum effort (inexpensive stirrer, low mixing work). In order to answer this question, the mixing time and power characteristics must be known for the stirrer.

The term *characteristic* means a relationship between dimensionless groups that describes the dependence of a target quantity (mixing time, power, etc.) on geometrical, physical, and process parameters. In terms of fluid mechanics, mixing is a complicated operation. Process-related questions are almost exclusively answered by experiments, which are best evaluated using dimensional analysis. This allows the number of parameters to be reduced, and ensures a reliable scale-up in combination with the theory of models [3] → Scale-Up in Chemical Engineering.

2.2. Mixing-Time Characteristic

2.2.1. Definition and Measurement Method

The mixing time θ is the time required for the mixer to achieve the desired degree of homogeneity. A statement of the mixing time is therefore only meaningful in conjunction with the degree of homogeneity. The most common methods for determining the mixing time are the schlieren method and the chemical decolorization method. In the *schlieren method*, two liquids having different refractive indices are mixed; the complete homogenization of the liquids is indicated by the disappearance of the streaks (schlieren). In the *chemical decolorization method*, the first liquid, which contains one reaction component, is colored with an indicator; the second liquid, containing the other reaction component, is then added as the stirring begins; and the time at which the color disappears is measured. (Common reaction systems: sulfuric acid/alkali solution with phenolphthalein indicator; thiosulfate/iodine with starch indicator.) The degree of homogeneity at the point of decoloration depends on the excess of the added

reaction component (Danckwerts method [4]). The principles and applications of this method are described in [5]. Chemical decolorization methods fail for extremely short mixing times (micromixing). In such cases, a consecutive dye reaction (diazotization) is used instead; the degree of mixing can be determined from the selectivity. Details of both methods can be found in [1] and → Mixing, Introduction.

2.2.2. Mixing-Time Characteristic at $\Delta\varrho \simeq 0$ and $\Delta\nu \simeq 0$

In the homogenization of liquids having the same density ϱ and kinematic viscosity ν, the mixing time θ for a given stirrer type and vessel inserts depends on the speed n, and diameter d of the stirrer. In terms of dimensional analysis, this can be expressed as $n\theta = f\,(Re)$. In this equation, known as the mixing-time characteristic, $n\theta$ represents the mixing number and Re the Reynolds number, $Re = nd^2/\nu$.

Since each mixing-time or homogenization characteristic is only valid for the particular geometry, the dimensional proportions of the stirrer types treated are given in Figure 8.

Figure 9 shows the mixing-time characteristics for these stirrer types as determined by the chemical decolorization method. With an excess of ca. 0.01 N of acid, the decolorization indicated an essentially complete molecular homogenization [6]. Figure 9 allows the determination of the mixing conditions required to achieve a desired mixing time. It provides the number of stirrer rotations ($n\theta$) needed to homogenize the mixture under given flow conditions (Re). Depending on the range of Reynolds numbers, the following stirrers can produce homogenization with the minimum number of rotations: helical ribbon mixer, blade stirrer, and the MIG stirrer, the last two in baffled vessels.

As the ratio H/D increases, the mixing time becomes longer; for a baffled vessel with a cross-beam stirrer, the following relationship is valid over the range $10^3 < Re < 10^5$ [6]:

$$n\theta = 16.5\,(H/D)^{2.6} \tag{3}$$

(The number of stirrer beams is assumed to increase with vessel height in accord with Figure 8 A).

Mixing-time characteristics for 25 % and 10 % deviation from the final mixed state have been obtained by C. J. Hoogendoorn and A. P. den Hartog [7]. Homogenization and power characteristics for enamel-coated stirrers with rounded edges can be found in [8]; a number of other mixer types and insert configurations are treated in [9].

2.2.3. Mixing-Time Characteristic at $\Delta\varrho \neq 0$ and $\Delta\nu \neq 0$

In the homogenization of liquids with different densities and viscosities, coarse mixing is affected by the specific weight difference of the two components, $g\,\Delta\varrho$; however, the elimination of the final concentration gradients takes place under conditions characterized by the properties of the homogeneous mixture, $\bar{\varrho}$ and $\bar{\nu}$. In this case, the mixing time depends on the following parameters: $\theta = f\ (n, d, \bar{\varrho}, \bar{\nu}\ g\,\Delta\varrho)$; in terms of dimensional analysis, this can be expressed as $n\theta = f\,(Re, Ar)$, where $Re \equiv nd^2/\bar{\nu}$ is the Reynolds number and $Ar \equiv d^3\,g\,\Delta\bar{\varrho}/(\bar{\nu}^2\,\varrho)$ the Archimedes number. The mixing-time characteristic of a cross-beam stirrer, as in Figure 8 A, operating in a baffled vessel with $H/D = 1$ is as follows [10]:

$$n\theta = 51.6 Re^{-1}\left(Ar^{1/3} + 3\right) \tag{4}$$

This relationship is valid for $10 < Re < 10^5$ and $10^2 < Ar < 10^{11}$.

2.2.4. Mixing-Time Characteristics in Non-Newtonian Liquids

Mixtures of non-Newtonian liquids are homogenized considerably more slowly than Newtonian liquids in the laminar and transition flow regimes. This is also due to the shear rate field existing in the stirring tank, which causes strong viscosity differences.

In most studies on non-Newtonian liquids carboxymethyl cellulose (CMC), hydroxyethyl cellulose (HEC), polyacrylonitrile (PAN), and polyacrylamide (PAA) were used. All of them exert both pseudoplastic and viscoelastic viscosity behavior. Already at an early stage of these investigations, Opara [11] stated that it was not to be expected that a correlation between $n\theta$ and

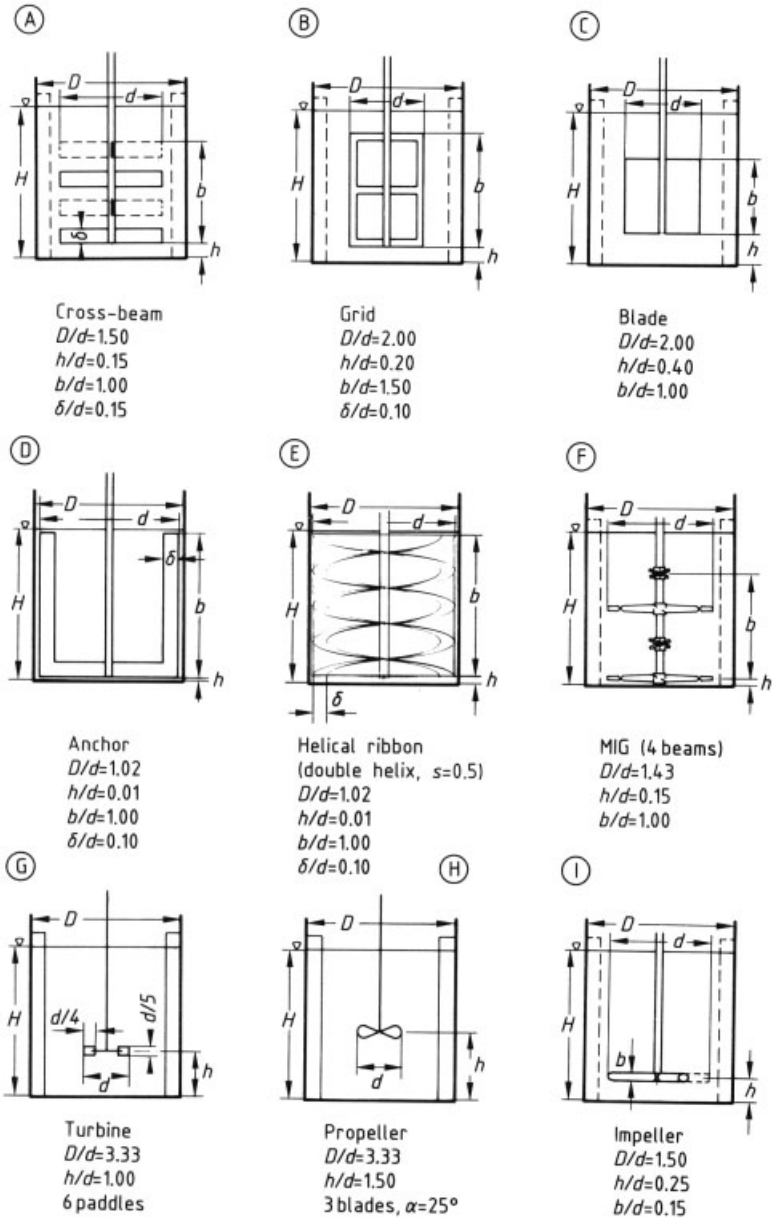

Figure 8. Dimensions and installation conditions of stirrer types for vessel with $H/D = 1$ (Baffles drawn with dotted lines indicate that the stirrer can be used in baffled or unbaffled vessels.)

$Re_{\text{eff}} \equiv n\ d^2 \varrho / \mu_{\text{eff}}$ could be realized, because μ_{eff} is determined from the shear rates $\dot{\gamma}$, which occur at the largest rate differences, whereas the regions close to the walls, in which the smallest rate differences are to be found, are decisive for homogenization times. It is therefore not surprising that in all research studies the Weissenberg number $We \equiv N_1/\tau$ had to be included (N_1 is the first normal stress difference and τ is the shear stress) to achieve a satisfactory correlation; see [12–14]. For the dimensional-analytical treat-

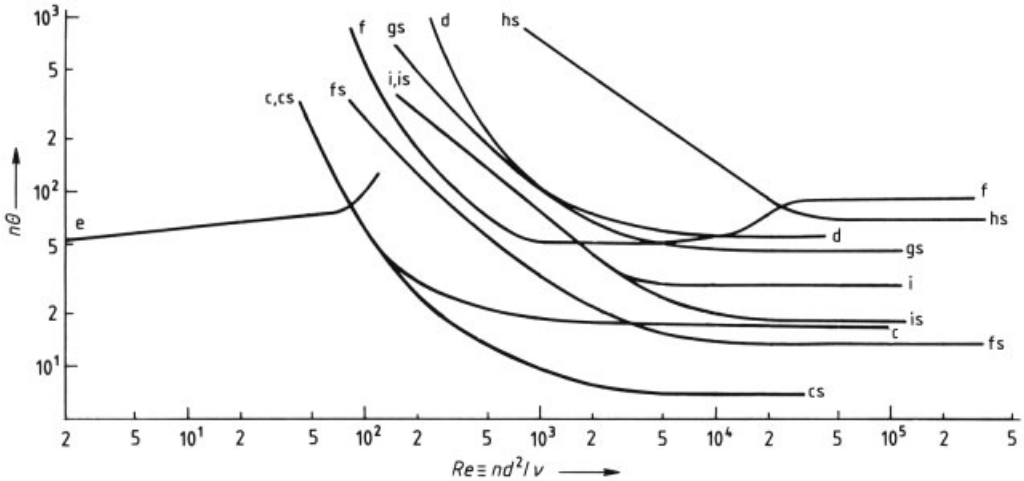

Figure 9. Mixing-time characteristics of the stirrer types in Figure 8 [6]
Cross-beam stirrer a = 1.8 · c (i.e., the mixing time is 1.8 times that of Curve c); as = 1.8 · cs; grid stirrer b = 1.25 · c, bs = 1.25 · cs; blade stirrer c, cs; anchor stirrer d; helical ribbon stirrer e; MIG-stirrer f, fs; turbine stirrer gs; propeller stirrer hs; impeller stirrer i, is. In each case s indicates the presence of baffles.
Example: What value of θ is obtained when a liquid with $\nu = 10^{-4}$ m^2/s is mixed in a baffled vessel using a cross-beam stirrer (as = 1.8 · cs) with $d = 1$ m and $n = 1$ s^{-1} ?
$Re = nd^2/\nu = 10^4$
$n\,\theta = 1.26 \times 10$
$\theta \approx 13$ s

ment of non-Newtonian liquids → Scale-Up in Chemical Engineering, Chap. 5.3. and [3].

2.3. Power Characteristic

2.3.1. Definition and Measurement Method

In addition to the stirrer (shaft) power P, the motor drive of a stirrer must also account for the power losses in the gearbox and the seals. In order to determine the mixing power P, the torque M_d, and the stirrer speed n of the must be known: $P = M_\mathrm{d}\omega$, with $\omega = 2\,\pi\,n$ as the angular velocity. The torque can be measured using, for example, a torsion shaft with strain gauges, electrically with eddy current torque transducers, or mechanically using a swiveling motor. The stirrer speed can be measured using mechanical, electrical (photocell), or optical (stroboscope) instruments.

2.3.2. Power Characteristics in Homogeneous Systems

The power required by a given stirrer type in a given vessel configuration depends on the speed of rotation n, the stirrer diameter d, the density ϱ, and the kinematic viscosity ν of the medium. In vessels without baffles, the liquid vortex, and therefore the acceleration due to gravity g, is immaterial, as long as no gas is entrained in the liquid. Thus, $P = f\,(n, d, \varrho, \nu)$ applies, and in accordance with dimensional analysis $Ne = f\,(Re)$. Here, $Ne \equiv P/(\varrho\, n^3\, d^{\,5})$ is the Newton or power number, and $Re \equiv n\, d^{\,2}/\nu$ the Reynolds number. The relationship $Ne = f\,(Re)$ is called the power characteristic of a stirrer.

The power characteristics of the stirrer types discussed in this section are presented in Figure 10. Three flow regimes can be identified:

1) In the *laminar regime* ($Re \leq 10$; for stirrers with very small wall clearances, such as the anchor or helical ribbon mixer, $Re \leq 100$), the effect of inertial forces (density) is suppressed by the viscous forces; therefore, baffles are unnecessary. In this regime, $Ne \propto Re^{-1}$, or $Ne\,Re =$ constant; the power is therefore given by

$$P = k_1 n^2 d^3 \mu \tag{5}$$

μ being the dynamic viscosity, and k_1 being the constant.

2) In the *turbulent flow regime*, the viscosity, and therefore the Reynolds number, ceases to have an effect. In this case Ne = constant applies, and the power is given by

$$P = k_2 n^3 d^5 \varrho \tag{6}$$

This regime begins at $Re \geq 10^2$ for baffled vessels and at $Re \geq 5 \times 10^4$ for vessels without baffles. In this regime baffles are fully effective and increase the required stirring power by up to a factor of twenty. For example, the power characteristics of the blade stirrer in a vessel with baffles (cs) and without (c) can be compared.

3) The *transition regime* arises only in vessels without baffles and for $Re \simeq 10 - 5 \times 10^4$; in this case, both viscosity and density have an effect. The relationship $Ne \propto Re^{-1/3}$ applies here.

The stirring power can be determined from Equations (5) and (6) using the constants k_1 and k_2 from Table 1. In general, the power can be calculated using Equation (7). The Reynolds number is determined first, and the corresponding value of Ne read from Figure 10; using this value, the power can now be evaluated by:

$$P = Ne\, n^3 d^5 \varrho \tag{7}$$

The power characteristic for vessels without baffles decreases monotonically over a wide range of flow conditions; this offers the possibility of determining the viscosity of a reaction mixture and therefore the conversion, or the degree of polymerization, as a function of time by measuring the mixing power.

If large differences in density or viscosity exist at the start of mixing, it is advisable to calculate the mixing power using the larger of the two values.

Of the extensive literature on stirring power, the basic works of W. Büche [15], J. H. Rushton and co-workers [16], as well as S. Nagata and co-workers [17] should be mentioned. Power characteristics for various designs of turbine and paddle stirrers are given in [18]. Many power characteristics can also be found in the literature on mixing time [6–9].

Table 1. The proportionality constants for Equations (5) and (6)

Stirrer type	Curve	k_1	k_2
		$Ne\,(Re = 1)$	$Ne\,(Re = 10^5)$
Cross-beam stirrer	a	110	0.4
	a s	110	3.2
Grid stirrer	b	110	0.5
	b s	110	5.5
Blade stirrer	c	110	0.5
	c s	110	9.8
Anchor stirrer	d	420	0.35
Helical ribbon stirrer	e	1000	0.35
MIG stirrer	f	100	0.22
	f s	100	0.65
Turbine stirrer	g s	70	5.0
Propeller stirrer	h s	40	0.35
Impeller stirrer	i	85	0.20
	i s	85	0.75

2.4. Optimization: Minimum Work of Mixing

The optimum stirrer type and stirring conditions for the homogenization of a liquid mixture are those that require the least work of mixing. These conditions can be found quickly with Figure 11, which shows the relationship between the modified power number $Ne\,Re^3\,(D/d) \equiv P\,D\,\varrho^2/\mu^3$ and the modified mixing number $n\theta\,Re^{-1}\,(D/d)^{-2} \equiv \theta\,\mu/D^2\varrho$. This representation is based on the information in Figures 9 and 10. The stirrer whose curve lies lowest in Figure 11 gives the desired mixing time with the least power. For the sake of clarity, Figure 11 includes only the curves for the optimum stirrers along with their corresponding Reynolds number scales.

The use of the figure is as follows: first the modified mixing number is determined, and then the best stirrer type can be read off the abscissa. The abbreviations are explained in Table 1. The point of intersection of this value with the curve provides the stirrer speed via the Reynolds number, $Re \equiv nd^2\varrho/\mu$. The ordinate of this point gives the modified power number, and through this the mixing power. A specific example is worked out in Figure 11.

To obtain a value of the mixing number lower than $\theta\,\mu/D^2\varrho \simeq 10^{-4}$ is practically impossible. For liquids with viscosity similar to that of water and for large vessels, the following approximation is valid:

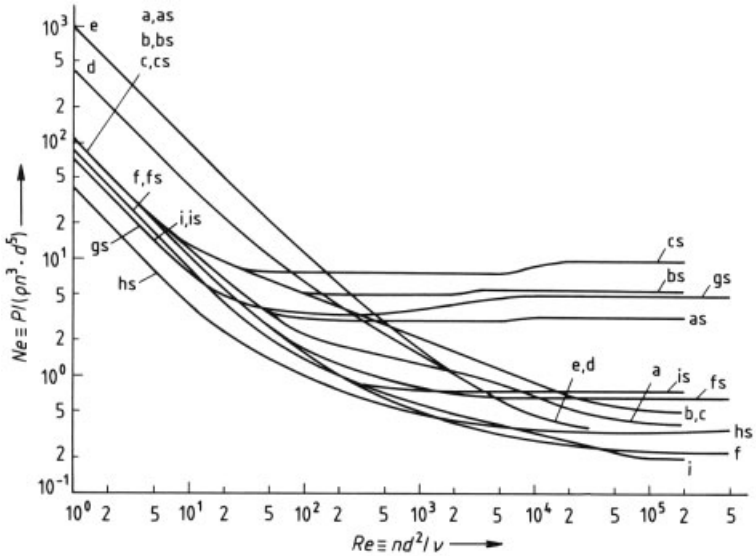

Figure 10. Power characteristics of the stirrer types in Figure 8 [6] (except g [18] and h [16])

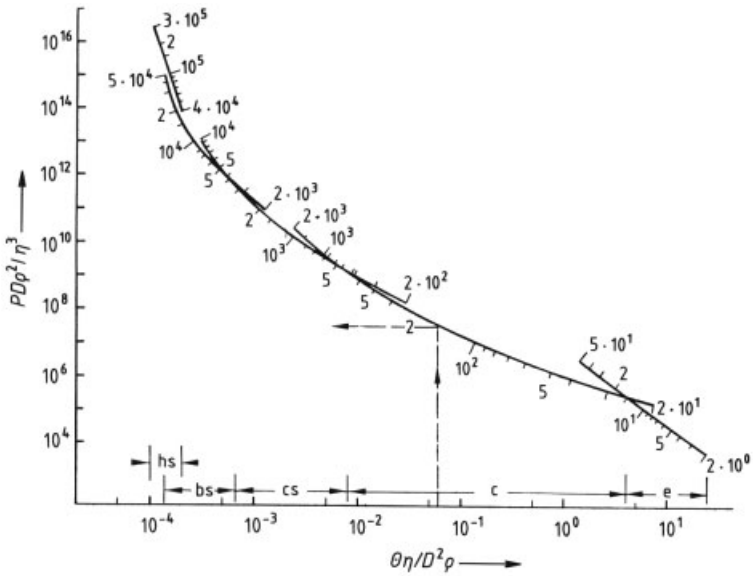

Figure 11. Curves for determining the optimum stirrer type and the operating conditions that minimize the mixing work [6]
The numbers along the curves are the Reynolds number, $Re = nd^2\ \varrho/\mu$
Example: $D = 1.0$ m, $\mu = 1$ kg m s^{-1} = 1000 mPa · s, $\varrho = 1000$ kg/m^3, $\theta = 60$ s. $\frac{\theta\mu}{D^2\varrho} = 6 \times 10^{-2}$
Blade stirrer, without baffles (Fig. 8 C)
$D/d = 2$; $d = 0.5$ m
$Re = 170$; $n \approx 41$ min^{-1}
$PD\ \varrho^2/\mu^3 = 3 \times 10^7$; $P = 30$ W
The dotted arrows show this example in the figure itself

$$\theta_{\min} \simeq 10^{-4} D^2 \varrho/\mu$$

Following a suggestion in [19], the pairs of values ($n\theta$, Re) can be selected from Figure 9 in which both $n\theta$ and Re are lowest; using the corresponding Ne values from Figure 10, the following relationship can be derived from Figure 11:

$$\frac{P D \varrho^2}{\mu^3} = a \left(\frac{\theta \mu}{D^2 \varrho} \right)^{-3} \text{ or } \frac{P \theta^3}{\varrho D^5} \triangleq \frac{P \theta^3}{V \varrho D^2} = \text{constant} \quad (8)$$

where $a \simeq 700$. Two observations may be drawn from this relationship, which is valid in the turbulent regime (μ irrelevant):

1) at constant P/V, θ increases as $D^{2/3}$
2) at constant D, θ decreases as $(P/V)^{-1/3}$

This confirms that the constancy of P/V in the turbulent regime does not constitute a scale-up criterion for the homogenization of liquid mixtures.

If the corresponding values for the helical ribbon mixer in the laminar regime are inserted in the same relationship [19], then

$$\frac{P \theta^2}{D^3 \mu} = 5.8 \times 10^5 \quad (9)$$

This indicates that the constancy of P/V is a scale-up criterion only in the laminar regime. It also shows that: $\theta \propto (P/V)^{-1/2}$.

2.5. Homogenization in Storage Tanks

Stirrers are rarely used for homogenization in storage tanks since the liquids involved are of low viscosity and relatively long mixing times are generally acceptable. In such cases, mixing is carried out with side-entering propeller stirrers, liquid jets, or rising gas bubbles.

In homogenization using *propeller stirrers* [20], [21], one or more are introduced laterally through the wall of the tank at an angle of 7 – 10° to the radius. At this angle, liquid circulation is intensive without significant rotation of the contents. Since the mixing time θ and the liquid throughput q' of the stirrer are related through $\theta \propto V/q'$ (V = liquid volume), the operating conditions are chosen to maximize the liquid throughput at a given power. This is achieved through large stirrer diameter and low stirrer speed (Section 2.6). At $H/D = 1$ and P = constant, the following is valid:

$$\theta \propto (D/d)^{2.3} \quad 10 < D/d < 30$$

At a given d, the mixing time is inversely proportional to the power, $\theta \propto P^{-1}$, and increases rapidly with density differences, $\theta \propto (\Delta\varrho/\varrho)^{0.9}$ [20].

Two conditions can generally arise during the mixing:

1) At high mixing intensities, homogenization mainly depends on convection and turbulent diffusion.
2) At lower mixing intensities, which is commonly the case, small density differences lead to the formation of layers, which are subsequently broken up through interfacial waves. This process is much slower and depends to a great extent on the tank geometry and physical properties of the system.

In the second case, the boundary between the two flow regimes is defined by the value of the critical Froude number, $Fr'_{\text{crit}} \equiv n^2\, d\varrho/g\, \Delta\varrho$ [22]:

$$\log Fr'_{\text{crit}} = 1.40 + 0.04\,(D/d) \quad (10)$$

At $Fr' \gg Fr'_{\text{crit}}$, $n\theta$ is independent of Fr', but depends strongly on D/d:

$$n\theta = 90\,(D/d)^{1.3} \quad (11)$$

On the other hand, at $Fr' \ll Fr'_{\text{crit}}$, the following applies:

$$n\theta = 90\,(D/d)^{1.3} \left(Fr'_{\text{crit}}/Fr'\right)^{1.5} \quad (12)$$

The most favorable installation geometry for the propeller stirrer (also according to [23]) is $\alpha = +10°$ and $\beta = +10°$ (where α is the angle between the stirrer shaft and the horizontal, and b the angle between the shaft projected on the horizontal plane and the radius). The power characteristic of side-entering propeller stirrers is approximately 1/3 lower than that of a vertical stirrer [23].

In the homogenization with *liquid jets*, the mixing time can be estimated relatively accurately under the assumption $\theta \propto V/q'$ [21]. The total liquid throughput q' consists of the jet

stream itself and the entrained liquid. For calculation of the jet stream, see [24], [25]; for the estimation of liquid entrainment, see [26].

The liquids with relatively low viscosity ($n < 100\,\mathrm{mPa} \cdot \mathrm{s}$) can be homogenized effectively with *rising gas bubbles* from a distributor at the bottom of the tank (sintered or perforated plate, two-fluid nozzle). Their effectiveness is due to their buoyancy, and therefore mixing ability, remaining constant over the entire height of the tank. For liquids similar to water and without significant density or viscosity differences, the following relationship applies [27]:

$$\theta\,(g/D)^{1/2} = 4.75\,Fr^{-1/4}\left(dH^2/D^3\right)^{1/4} \tag{13}$$

where $Fr \equiv q^2/(D^5/g)$. After transformation, this becomes

$$\theta/D = 4.75\,(H/q)^{1/2}\,(d/g)^{1/4} \tag{14}$$

q is the gas flowrate, and d the diameter of the gas distributor. For the estimation of the gas flowrate for lower surface tensions or higher viscosities than that of water, refer to [27].

Homogenization with rising gas bubbles is also effective for the mixing of water reservoirs or lakes. Model measurements [28] yield the following relationship for the mixing time θ,

$$\theta = 250\left[H^4/(qg)\right]^{1/3} \tag{15}$$

which can be expressed in dimensionless form as

$$\theta\,(g/D)^{1/2} = 250\,Fr^{-1/6}\,(H/D)^{4/3} \tag{16}$$

2.6. Liquid Throughput of Stirrers

The liquid flowrate that is displaced through the area swept by the stirrer is termed the liquid throughput of the stirrer or its pumping capacity q'. This area is given by $\pi\,d^2/4$ for axially operating stirrers, and by $\pi\,db$ for radial flow stirrers, where b is the paddle height. For axial flow stirrers, q' is determined from the circulation time of particles with $\Delta\varrho \simeq 0$; for radial flow stirrers, it is obtained from the integral value of the measured velocity distribution of the flow.

In continuously stirred tanks, homogenization can only be achieved if $q' \gg q$, where q is the liquid flowrate through the vessel. Under these conditions, the mixing time is $\theta \propto V/q'$, where V is the liquid volume. In the case of propeller stirrers, knowledge of q' enables calculation of the flow velocity along one or more spiral coils concentric to the stirrer. The installation of such coils may be unavoidable in the case of extremely exothermic reactions.

The throughput characteristic of a given stirrer is provided by $Q' = \mathrm{f}(Re, D/d, h/d)$, where $Q' \equiv q'/(nd^3)$ represents the dimensionless throughput number; $Re \equiv nd^2/\nu$ the Reynolds number; and D/d, h/d the geometric parameters of the stirrer installation. For propeller stirrers with pitch $s = 1$ (pitch s = tangent of the blade angle α) operating in baffled vessels with $H/D = 1$, the following relationship is valid in the turbulent regime ($Re > 10^3$) [35]:

$$Q' = 0.654\,(D/d)^{-0.16} \quad D/d > 3 \tag{17}$$

For paddle stirrers with six pitched paddles ($\alpha = 45°$, $b/d = 0.2$), the relationship is as follows [36]:

$$Q' = 1.29\,(D/d)^{-0.20}\,;\; 2 < D/d < 3.3 \tag{18}$$

For turbine stirrers with six paddles, as in Figure 8 G, the throughput is given by [37]:

$$Q' \simeq 0.75 \tag{19}$$

Other stirrer configurations and paddle numbers are treated in [38].

2.7. Vortex Depth in Unbaffled Stirred Tanks

All calculations for the design of stirrers in unbaffled vessels are valid only under the assumption that the vortex does not reach the stirrer head, otherwise, gas entrainment alters the physical properties of the system. The highest allowable rotation speed is given by the following relationships [39].

Propeller stirrer with $D/d = 3.33$ and $h/d = 1$ (Fig. 8 H):

$$Fr_{\max} = 0.072\left(H'/d\right)^{1.33}/\left(0.25 - Ga^{-0.10}\right) \tag{20}$$

Turbine stirrer with $D/d = 3.33$ and $h/d = 1$ (Fig. 8 G):

$$Fr_{\max} = 0.016\,(H'/d)^{1.16} / (0.10 - Ga^{-0.18}) \quad (21)$$

These expressions are valid over the range $2.3 < H'/d < 5.7$ and $10^6 < Ga < 10^{10}$. In these equations, $Fr_{\max} \equiv n^2_{\max}\, d/g$ is the Froude number; H' the height of liquid above the stirrer; and $Ga \equiv d^3 g/\nu^2$ the Galileo number. For other stirrer types and geometries, see [39–41].

3. Intensification of Heat Transfer

3.1. Problem Definition

Stirrers are often used to intensify the heat transfer between the liquid and the heat exchange surface in a vessel. The question is what type of stirrer to use and under what conditions so that the desired heat flux can be effected at a minimum expense (inexpensive stirrer, low power requirements). In the case of cooling, this question involves an optimization problem: high stirrer speeds result not only in an increase of the heat flux, but also in a higher heat of agitation.

3.2. Heat-Transfer Characteristics

3.2.1. Definitions

Heat transfer is described through the general heat-transfer equation $Q = UA\,\Delta T$, where Q is the heat flowrate, A the heat-exchange surface area, and ΔT the temperature difference. The overall heat-transfer coefficient U is composed of inner and outer film heat-transfer coefficients, h_i and h_o, the thermal conductivity of the wall material, k, and of the wall thickness, δ, according to the relationship

$$1/U = 1/h_o + \delta/\lambda + 1/h_i$$

Stirring affects only the inner-side heat-transfer coefficient h_i. This coefficient is defined through the relationship $h_\mathrm{i} \equiv Q/(A\,\Delta T)$, where the characteristic temperature difference ΔT is generally taken to be the difference between the average wall temperature T_w and the average temperature of the the bulk liquid T, i.e., $\Delta T = —T_\mathrm{w} - T—$.

For a given stirrer type and a fixed geometry, h_i is a function of the vessel diameter D, the stirrer diameter d, and the stirrer speed n, as well as the physical properties of the liquid (specific heat capacity c_p, heat conductivity k, density ϱ, dynamic viscosity in the bulk μ and at the wall μ_w): $\alpha_\mathrm{i} = \mathrm{f}\,(D, d, n, c_\mathrm{p}, \lambda, \varrho, \mu, \mu_\mathrm{w})$. Dimensional analysis (→ Scale-Up in Chemical Engineering) yields the following expression:

$$Nu = f\,(Re, Pr, Vis, D/d) \quad (22)$$

where $Nu \equiv h_\mathrm{i}\,D/\lambda$ is the Nusselt number, $Re \equiv nd^2\,\varrho/\mu$ the Reynolds number, $Pr \equiv c_\mathrm{p}\mu/\lambda$ the Prandtl number, $Vis \equiv \mu_\mathrm{w}/\mu$, and D/d the diameter ratio. The relationship between these dimensionless numbers is termed the heat transfer characteristic.

3.2.2. Heat-Transfer Characteristics in the Range Re > 200

In this flow regime, heat-transfer characteristics are generally of the form $Nu = \mathrm{c}\,Re^{2/3}\,Pr^{1/3}\,(\mu/\mu_\mathrm{w})^\mathrm{m}$, where the constant c is a function of the stirrer type and the nature of the heat exchange surface (baffled or unbaffled tank wall, spiral or meander coil). The exponent of the viscosity term, m, which enables both heating and cooling data to be presented together, has been found to be 0.14, 0.18, or 0.24 in several investigations. Studies, which separately investigated the effect of viscosity, revealed that the exponent m of Vis is not a constant, but a function of the sort: $m = -\,aVis^{-b}$. For details see [1], Section 7.4.

Table 2 lists the values of the constant c for various stirrer types for heat transfer through the tank wall (W) or a spiral coil (S); these values are taken from [42], in which 77 original works are evaluated and compared. A further collection of such expressions is given in [43].

The constant of the heat-transfer characteristic c has the following values:

1) Unbaffled vessels, $c = 0.35 - 0.4$, independent of stirrer type and diameter ratio
2) Baffled vessels, $c = 0.5$ for axial flow mixers (e.g., propeller stirrers) and $c = 0.75 - 0.80$ for radial ones (e.g., blade and turbine stirrers)
3) Tanks with spiral coils, $c = 0.70 - 0.90$ for radial flow stirrers, independent of whether baffles are present or not, whether inside or outside the coil.

Table 2. Values of c in the relationship $\mathrm{Nu} = c\,\mathrm{Pr}^{1/3}\,\mathrm{Re}^{2/3}\,(\mu/\mu_w)^{0.14}$

Stirrer Type and Installation	*c*	*Re* Range	Reference [d]
Cross-beam stirrer with one beam, $D/d = 1.67$, $\delta/D = 0.16$, no baffles, spiral coil	W 0.36 S 0.78 [a]	$3\times10^2 - 3\times10^5$	[44]
Two paddle stirrers with 6 paddles, arranged one above the other, $D/d = 2.5$, $\delta/d = 0.16$, no baffles, spiral coil	W 0.40 S 0.70 [a]	$2\times10^2 - 1\times10^6$	[39]
Blade stirrer $D/d = 2.0$, $b/d = 1.42$, no baffles, spiral coil	W 0.36 S 0.70 [a]	$1\times10^3 - 1\times10^6$	[45]
Blade stirrer, 4 baffles (D/15 wide), $D/d = 2.0$, $b/d = 1$, $d/D = 0.25 - 0.60$, $b/D = 0.10 - 0.60$	W 0.80 W 1.31 $(d/D)^{0.39}(b/D)^{0.34}$	$3\times10^2 - 5\times10^5$	[46]
Turbine stirrer, 4 baffles (D/10 wide) $D/d = 3.33$, $h/d = 1$ (Fig. 8 G) $h/D = 0.08 - 0.033$, $H/D = 0.66 - 1.0$	W 0.75 [b] W 1.15 $(h/D)^{0.4}(H/D)^{-0.56}$	$3\times10^1 - 4\times10^4$	[61]
Turbine stirrer, 4 baffles (D/10 wide) $D/d = 3.33$, $h/d = 1$ (Fig. 8 G) h/D = 0.05 − 0.70; $d/D = 0.16 - 0.75$	W 0.76 W 1.01 $(h/d)^{0.12}(d/D)^{0.13}$	$5\times10^4 - 8\times10^5$	[41]
Turbine stirrer (Fig. 8 G), spiral coil $d/D = 0.3$, $\delta/D = 0.03$, $D_s/S = 0.7$ $d/D = 0.25 - 0.58$, $\delta/D = 0.018$ -0.036 (D_s = coil diameter, δ = tube diameter) Baffles, within or outside the coil, have no effect	S 0.87 S 0.17 $(d/D)^{0.1}(\delta/D)^{-0.5}$	$4\times10^2 - 1\times10^6$	[47]
Propeller stirrer, 4 baffles (D/10 wide) pitch $s = d$, $h = d$ $s/d = 0.4 - \infty$, $h = d$	W 0.50 W 0.64 $(s/d)/(0.285 + s/d)$	$2\times10^5 - 9\times10^5$	[48]
Anchor stirrer (Fig. 8 D), D/d = 1.04	W 0.38	$3\times10^2 - 4\times10^4$	[49] , [72]
Anchor stirrer (Fig. 8 D), $D/d = 1.02$ heating cooling	W 0.40[c] W 0.26[c]	$2\times10^2 - 1\times10^5$	[50]

[a] Based on $Re^{2/3}$
[b] With (μ/μ_w)
[c] Without (μ/μ_w) term
[d] The reference numbers refer to the list of references in [42].
W wall; S spiral coil

3.2.3. Heat-Transfer Characteristics in the Range Re < 200

In the laminar flow range $Re < 10^2$ heat conduction dominates and the effect of density and viscosity disappears entirely, like in solid bodies. To account for this physical situation, Reynolds and Prandtl numbers have to be combined such as to eliminate ϱ and μ;. Then the heat transfer characteristic is given as:

$$Nu = \text{const}\, Pe^a Vis^{-b}\,(d/D)^c$$

where the Peclet number $Pe \equiv PePr = nd^2\varrho c_p/k$ (Note: ϱc_p denotes the heat capacity per unit volume!).

Ishibashi et al. [51] investigated the heat transfer behavior in a vessel with a single or double helix ribbon stirrer (with an inner and outer helix) and found that:

$$Nu = 2.0 Pe^{1/3} Vis^{-0.14} \qquad Re = 5 - 50$$

$$Nu = 1.0 Re^{1/2} Pr^{1/3} Vis^{-0.14} \quad Re = 50 - 500$$

and for a centrally mounted screw stirrer [51]:

$$Nu = 1.0 Re^{1/2} Pr^{1/3} Vis^{-0.14} \quad Re = 5 - 50$$

KURIYAMA et al. [52] proposed the following relationship for the double helix ribbon stirrer for cooling and heating and for different pitches (p) and clearances $c = (D - d)/D$:

$$Nu = 0.64 Pe^{1/3} (D/d)^{2/3} (p/c)^{1/3} Vis^{-0.14}$$

The experimental data of ZLOKARNIK [53] with an anchor stirrer with $d/D = 0.98$ can also be interpreted in the range $Re = 0.5 - 100$ as $Nu \propto Pe^{1/3}$. However, the following relationship was selected instead, which satisfactorily covers the whole Re range:

$$Nu = \text{const} \left(RePr^{1/2} + 4 \times 10^3 \right)^{2/3} Vis^b$$

The comparison between the anchor stirrer and a double helix ribbon stirrer with a pitch of 0.5 for the same $d/D = 0.98$ established that the helical ribbon stirrer for $Re \approx 1$ provided a Nu value which was approximately double that with the anchor stirrer. However, this superiority had already been lost at $Re \approx 100$ [53]. Apparently, here the radial flow component is entirely sufficient for heat transfer.

If considerable intensification of heat transfer is required for $Re < 100$, a stirrer with a very small wall clearance ($d/D \rightarrow 1.00$) is recommended. ZLOKARNIK found no effect of clearance in the whole Re range for anchor stirrers with $d/D = 0.91 - 0.98$, but an extremely large increase in heat flow for $d/D \approx 1.0$ (without wiper blades!) [53]. For cooling, the heat flow is approximately a factor of 4 higher, whereas for heating an increase of only about a factor of 2 was found. These results can be described by the relationship

$$Nu = a \left(RePr^{1/2} + b \right)^{2/3} Vis^c$$

For the range $Re = 10 - 200$, UHL and VOSNICK found the dependence [54]

$$Nu \propto Re^{0.5} Pr^{1/3}$$

whereas PENNEY and BELL confirmed the expected relationship (for $Re < 100$) [55]

$$Nu \propto Pe^{1/3}$$

Removal of the boundary layer from the inside wall of the tank to a considerable extent is only possible with stirrers with wiper blades. (For the numerical analysis of the three-dimensional flow and thermal behavior in a scraped-surface heat exchanger see [56].) A first review over such stirrers in [55] brought little of substance. NAGATA provided the following relationship for helical ribbon stirrers with wipers [57]:

$$Nu = 6 Pe^{1/3} Vis^{-0.20}$$

JSUDAT investigated the heat exchange behavior in a stirred tank ($D = 444$ mm, $H = 1080$ mm; $H/D = 2.43$) with helical ribbon stirrers ($d = 340$ mm; pitch 1), on which one, two or four wiper pairs (number of wipers $i = 2; 4; 8$) were so mounted on the shafts, that they pressed firmly on the inside wall of the tank upon rotation, as a result of hydraulic forces [58].

Figure 12 shows the evaluation of these measurements for $i = 4$. As can be seen, the correlation gives satisfactory results in the laminar range ($Re < 10^2$) and in the turbulent range ($Re > 10^4$):

$$Re < 10^2 : Nu = 10 Pe^{1/3}$$

$$Re > 10^4 : Nu = 0.5 Re^{2/3} Pr^{1/3}$$

In the creeping flow range of $Re < 10$, the insertion of wiper blades improves the heat transfer by a factor of 10!

In the transition range, spreading of measuring data with the stirrer speed appears, and indeed the more strongly, the higher the rotation speed. This phenomenon is satifactorily explained by the modified penetration theory of RAUTENBACH and BOLLENRATH [59].

3.3. Stirrer Power During Heat Transfer

The relationships in Section 2.3. also apply to the stirrer power in heat transfer. If wall clearances are small, the viscosity at the wall temperature should be used to calculate the Reynolds number.

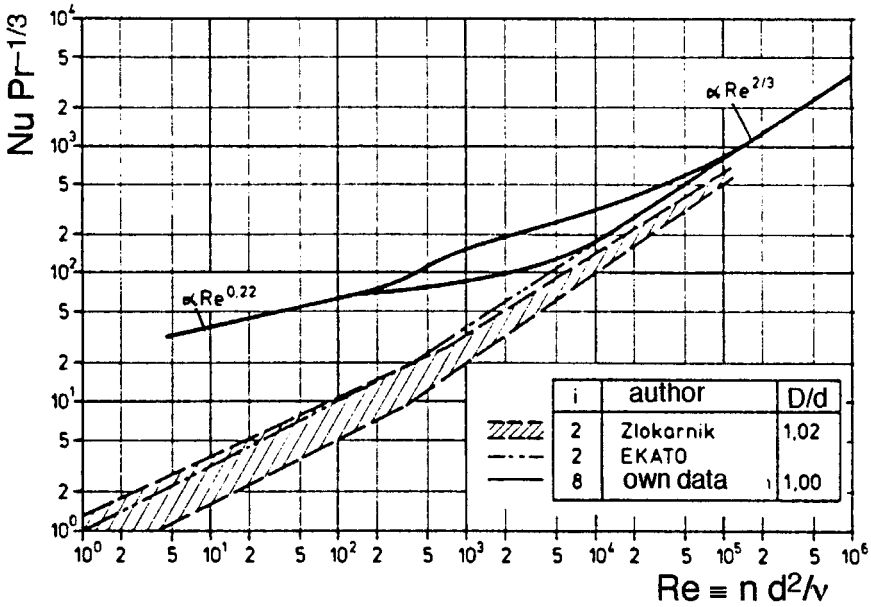

Figure 12. Heat transfer characteristic of a helical ribbon stirrer with hydraulic pressing wiper blades (2 wiper pairs, $i = 4$); according to [58]

3.4. Optimum Stirring Conditions for the Removal of Reaction Heat

Heat removal is a particularly severe problem in the case of strongly exothermic reactions in viscous mixtures. Figure 13 illustrates the problem with an example. The values of the rate of heat transfer Q through the wall, as well as the stirrer power P, are plotted as functions of the speed of rotation n. These values are calculated using the heat-transfer and power characteristics of the anchor stirrer (Fig. 10 and [53]). In the appropriate range of Reynolds numbers ($5 \leq Re \leq 50$), stirrer power increases much more rapidly with the stirrer speed ($P \propto n^2$) than the heat flux ($Q \propto h \propto n^{1/2}$). The thick curve represents the difference between the two straight lines on the log-log plot and corresponds to the heat of reaction $R = Q - P$ that can be removed through the wall: R goes through a maximum at $n \simeq 20\,\text{min}^{-1}$ and reaches zero at $n = 59\,\text{min}^{-1}$. At this point, the heat removed is that generated by the stirrer itself.

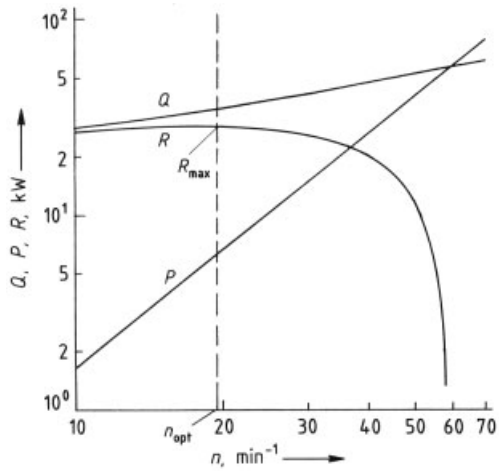

Figure 13. Graphical representation of the relationship between Q, P, $R = Q - P$, and the rotational speed n [60] Anchor stirrer in a vessel with $D = 2.0\,\text{m}$, $V = 5.86\,\text{m}^3$, $d = 1.8\,\text{m}$ ($D/d \simeq 1.10$), $\varrho = 1000\,\text{kg/m}^3$, $\mu = 50\,\text{kg/(m} \cdot \text{s)}$, $k = 2\,\text{W/(m} \cdot \text{K)}$, $c_p = 2000\,\text{J/(kg} \cdot \text{K)}$, $A = 14\,\text{m}^2$, $\Delta T = 25°\text{C}$, $h \approx U$

PCs and pocket calculators of today are useful tools for a quick determination of the optimum stirring conditions. The only thing to do is to insert the respective process equations for

$Ne(Re)$ and for $Nu(Re, Pr)$, then the dependence $R = Q - P = f(n)$ can be calculated. Thirty years ago, when these devices were not available, a nomogram was developed to facilitate this calculation [60].

The determination of the optimum stirring conditions has also been treated by PENNEY and KOOPMAN [61]. They found that in the laminar regime the maximum amount of reaction heat can be removed when the stirrer power is approximately 20 % of the heat of reaction; in the turbulent regime, this figure is 5 – 20 %.

4. Suspension of Solids in Liquids (S/L Systems)

4.1. Problem Definition

Suspension is connected with the presence of a dispersed solid phase, which is suspended ("whirled up") in the liquid. The material system solid – liquid (S/L) is thus involved. Occasionally a gas is distributed and dispersed in the liquid at the same time, in this case the material system concerned is a solid – liquid – gas (S/L/G) system.

For solids to be dissolved rapidly in liquids, their total surface must be accessible to the liquid, i.e., the solid particles must be completely suspended. In this case that critical stirrer speed n_{cr} is of primary interest, at which this state is just attained, and the other stirring conditions (stirrer type and installation), which ensure the maintenance of these flow conditions with the lowest stirrer power.

In crystallization or in solid catalyzed chemical reactions it is, on the other hand, generally necessary, that the dispersed phase is largely uniformly distributed over the whole liquid height. This condition naturally requires both a higher stirrer speed and power and occasionally also another stirrer type than suspension alone.

4.2. Classification of the Suspension Conditions

Two suspension criteria have become established, namely complete suspension and homogeneous suspension. A *complete suspension* is achieved, when no particle or no particle cluster remain on the bottom for more than 1 s This so-called 1-s criterion establishes the critical stirrer speed n_{cr} and is due to ZWEITERING (1958) [102]. This time criterion is easy to determine, but represents an inaccurate measured quantity with a low sensitivity.

The term *homogeneous suspension* describes a state in which the solid particles are homogeneously distributed over the whole liquid height. It concerns a suspension state, which is particularly aspired to, if the suspension must be continuously conveyed through the vessel. Since this state is difficult to achieve and to measure, in 1977 EINENKEL suggested that a "homogeneous suspension" should be referred to when the height of the suspended particle layer, h_p, attains 90 % of the liquid height [62]. Above this a usually clear, particle-free liquid layer is produced. The height h_p of the suspension zone can be measured with ± 5 % accuracy, particularly for small particles. The particle layer height criterion of $h* = 0.9$ is introduced, whereby $h* \equiv h_p/H$ represents the relative height, i.e., the particle layer height h_p related to the total liquid height H, n_{cr} coincides with the discontinuity in the $h*(n, \varphi_v)$ dependence, because for $n < n_{cr}$ the stirrer power also has to suspend an increasing number of particles from the bottom, whereas the additional power above $n = n_{cr}$ is utilized for increasing the particle layer height [62].

For the distribution of solids upon suspension see [103].

4.3. Suspension Characteristic

When different criteria for describing the suspension state are in use, these must also be taken into consideration in the formulation of the suspension characteristics. Depending on which criterion is used in the determination of the critical stirrer speed n_x, where the subscript x denotes the 1-s criterion or the particle layer height criterion $h* = 0.9$, the process numbers are correspondingly designated.

Initially, the average particle size d_p was taken as the characteristic length of the solid particle and the geometrical number d_p/d was thus formed. It was evident, however, that the

pi-spaces, which in addition to $g\Delta\varrho$ only considered this geometrical number, differentiated too strongly and led to contradictory scale-up rules.

On the basis of the analogy between $n_{0.9}$ and d_p, on the one hand, and the sinking velocity of the particle swarm w_{ss} and d_p, on the other, w_{ss} was later incorporated into the relevance list, because this leads to a better physical interpretation of the process.

4.3.1. Suspension Characteristics with d_p as the Characteristic Particle Dimension

This suspension characteristic is as follows:

$$\{Re_x, Fr'_x, d_p/d, \varphi_m\} \text{ or } \{Re_x, Ar, d_p/d, \varphi_m\}$$

where

$Re_x \equiv \frac{n_x d^2}{\nu}$	is the critical Reynolds number as target number,
$Fr'_x \equiv \frac{n_x^2 d \varrho}{g\Delta\varrho}$	is the Froude number extended by $\varrho/\Delta\varrho$, which is usually converted into
$Ar \equiv \frac{Re^2}{Fr'} \equiv \frac{d^3}{\nu^2}\frac{g\Delta\varrho}{\varrho}$	which is the Archimedes number and φ_m is the mass portion of solids in the liquid.

In a comprehensive study Kneule and Weinspach [63] have derived the suspension characteristics for the 1-s criterion with a number of stirrer types in baffled and unbaffled vessels with differently shaped bottoms (dished, hemispherical, truncated cone, flat) by using two tank sizes ($D = 0.37$ and 2.07 m) and different solid particles $\varrho_s = 2.630$ to 11.100 kg/m^3; $d_p = 1.1 - 10$ mm with mass fractions $\varphi_m = 0.02$ to 0.25. These characteristics can be divided into two parts. The turbulent Reynolds range begins at $Re_{1s} > 130\ (d/d_p)$ in which the liquid viscosity and therefore Re_{1s} is irrelevant and the process characteristic is hence given by

$$Fr'_{1s} = \text{const}_1\ \varphi_m^{0.5} \tag{23}$$

The corresponding numerical values in Equation (23) for propeller stirrers, which convey downwards, and turbine stirrers are 3.06 and 1.21, respectively. Therefore, in the same material system, the propeller stirrer must operate with a factor of $(3.06/1.21)^{0.5} = 1.59$ higher stirrer speed than an identically sized turbine stirrer to maintain the solids in suspension.

On the other hand, according to [63], for tests in the transition range $Re_{1s}\ (d/d_p) = 13 - 130$, the following applies:

$$Re_{1s} = \text{const}_2\ Ar^{0.4}\ [(d/d_p)\ \varphi_m]^{0.2} \text{ or}$$
$$Fr'_{1s} = \text{const}_2\ Re^{-0.5}\ [(d/d_p)\ \varphi_m]^{0.5} \tag{24}$$

The numerical values of const_1 and const_2 in the two process relationships (23) and (24) are summarized in tables for a number of stirrer types and the above-mentioned geometric ratios [63]. It emerges that the dished bottom is insignificantly less favorable than the hemispherical bottom for suspending solids. The shape of the bottom must be in such a condition that it conveys the solids to the center, where the stirring effect is the strongest. The D/d ratio exerts a very strong effect; D/d of ca. 3.3 is regarded as optimal. In the range $0.3 < h/d < 0.8$, the h/d ratio has practically no effect, a h/d ratio of ≈ 0.5 being recommended. As stirrer type, propeller stirrers, which convey axially downwards, are recommended as well as radially acting double turbines.

Lamadé [64] has evaluated the suspension characteristics for three different enameled stirrers (Pfaudler impeller, pitched-blade, paddle stirrers) in the turbulent flow range and found a process relationship of the form:

$$Fr'_{1s} \propto (d_p/d)^{0.42}\ \varphi_m^{0.3}\ (h/d)^{0.23(d/D)^{1.5}} \tag{25}$$

Fr'_{1s} increases with h/d to the power of $(d/D)^{1.5}$!

Einenkel [65] compared eight earlier studies dealing with the suspension of solid particles with propeller stirrer in the turbulent range and found for the process relationship of the type:

$$Re_{1s} \propto Ar^a\ (d_p/d)^b\ \varphi_m^c$$

the following exponents: $a = 0.45$ to 0.6; $b = -0.2$ to $+0.25$; $c = 0.25$ to 0.28.

There are still several studies to be mentioned, in which the experimental data was evaluated in the same way:

- twelve-turbine stirrer with paddles covered on both sides, different particles (sand, glass, siderite with $d_p = 0.1 - 1.25$ mm), kinematic viscosity of the liquid $\nu \times 10^6 = 1 - 160$ m^2/s [66]:

$$Fr_{1s} = 1.07 \times 10^{11}\ Re^{-2}\ \varphi_m\ (d/d_p)^{2.5}\ \ 10^{-2} < Ar < 10^6$$

- four- and six-pitched-blade stirrers in tanks with flat bottoms [67]:

$$Re_{1s} = a\,Ar^{0.5}\,\varphi^{0.25} \text{ applies for}$$
$$Ar\,(d_p/d)^2\,\varphi^{0.5} = 2 \times 10^3 - 3 \times 10^4$$

- six-pitched-blade stirrers, $D/d = 3$; $h/d = 0.5$; flat bottom [68]:

$$d_p/D < 7 \times 10^{-3}\text{: } Fr'_{1s} = 16.6\,(d_p/D)^{0.6}$$

$$d_p/D > 7 \times 10^{-3}\text{: } \mathrm{Fr}'_{1s} = 0.86$$

These relationships apply for $\varphi_v = 0.025$ and $d_p/D = 2 \times 10^{-4} - 3 \times 10^{-2}$.

- six-turbine stirrer at different installation conditions [69]:

$$Fr_{1s} = 0.61\,(d_p/d)^{0.38}\,(h/d)^{0.35}\,(D/d)^{2.5}\,\varphi^{0.32}$$

- six-blade and pitched-blade turbine at extremely low bottom clearances ($h/d = 0.021 - 0.25$) and different liquid viscosities [70].

4.3.2. Suspension Characteristics with w_{ss} as the Characteristic Particle Property

The sinking velocity of the swarm, w_{ss}, cannot be measured, but it can be calculated from the sinking velocity w_s of a single particle according to different expressions of the type:

$$w_{ss} = w_s\,f\,(\varphi_v),$$

where φ_v is the volume fraction of the solids in the liquid [1].

Because w_{ss}/w_s does not depend upon d_p/d, it cannot replace this parameter in the pi set. Whether w_{ss}/w_s can completely replace φ_v, can only be determined experimentally. The suspension characteristics with w_{ss} as the characteristic particle property therefore reads:

$$\{Re_x, Fr_x, w_{ss}/w_s, d_p/d, \varphi_v\} \text{ or}$$
$$\{Fr_x, Ar, w_{ss}/w_s; d_p/d, \varphi_v\}$$

There are many experimental studies which evaluated the results in this way, see [1]. In the following only the results of KRAUME and ZEHNER [71] and of EINENKEL [73] will be presented. In both cases, the quotient $P_{0.9}/P_{sink}$ was taken as the target quantity. $P_{0.9}$ is the stirring power needed to achieve $h* = 0.9$ and P_{sink} is the power consumption of the sinking particle swarm, which corresponds to the minimum stirring power to keep the particles in suspension. KRAUME and ZEHNER evaluated their experimental data in the pi space [71]

$$\{P_{0.9}/P_{sink},\, D^*\} \quad \text{with} \quad D^* \equiv$$
$$(Ar\,\varphi_v\,Re_{ss}/C_f)^{1/3} \equiv \frac{D}{\nu}\left(\frac{g\Delta\varrho}{\varrho}\,\frac{\varphi_v\,w_{ss}\,D}{C_f}\right)^{1/3}$$

The measurements with turbine and propeller stirrers were carried out both in laminar ($D = 0.1$ m, viscous medium, low w_{ss}) and turbulent ($D = 1.0$ m, low viscosity medium, high w_{ss}) ranges. The target quantity was $n_{0.9}$for which the Reynolds number of the stirrer must also be $\mathrm{Re} > 10^3$. The results are presented in Figure 14. The fitted curve for the turbine stirrer is:

$$\frac{P_{0.9}}{P_{sink}}\,\frac{d}{D} = 7.5 + 1.08 \times 10^4\,D^{*-3/4}$$

and for the propeller stirrer

$$\frac{P_{0.9}}{P_{sink}}\,\frac{d}{D} = 2.5 + 3.6 \times 10^3\,D^{*-3/4}$$

Experimental data from BREUCKER (1986) [71] and the correlation due to EINENKEL [72] are also inserted into the bottom figure. It can be seen that the relationship found by EINENKEL ($P_{0.9}/P_{sink} \propto D^{*-1/3}$) can indeed be included into the total picture, but that it only applies in a small $D*$ range. He later simplified this rather complicated dimensionless frame, but his approach is valid only for $\varphi_v \to 0$ [73].

4.4. Power Requirements Upon Suspension of Solids

WEINSPACH [104] has determined the power characteristics for the downwards conveying propeller stirrer and for the on both sides partly covered double turbine in suspension operation. Here the density and the dynamic viscosity of the suspension were defined as follows:

$$\varrho_{sus} = \varrho + \varphi_v\,(\varrho_{susp} - \varrho)$$

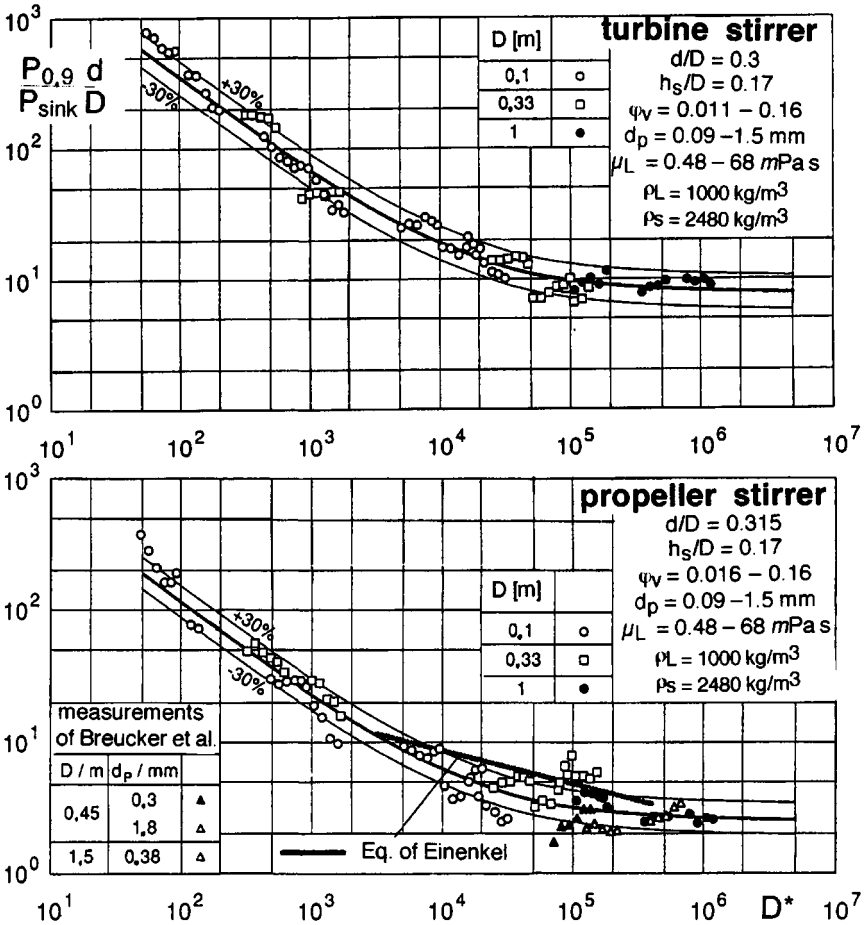

Figure 14. Dependence of the power ratios for $n_{0.9}$ as a function of the dimensionless tank diameter $D*$ for turbine and propeller stirrers; from [71]

$$\mu_{\text{sus}} = \mu \left(1 + \frac{1.25\,\varphi_{\text{v}}}{(1 - \varphi_{\text{v}}/\varphi_{\text{v,max}})}\right)^2$$

where φ_{v} is the average volume fraction of solids in the suspension, $\varphi_{\text{v,max}}$ is the maximum fraction of solids in the closest spherical packing ($\varphi_{\text{v,max}} = 0.74$) and μ is the dynamic viscosity of the homogeneous liquid; μ was varied in this study by adding tylose (methyl cellulose with Newtonian viscosity behavior).

It was found that up to $\varphi_{\text{v}} = 0.25 - 0.30$ suspensions behave as homogeneous mixtures, i.e., up to these φ_{v} values no dependence of Ne_{sus} on φ_{v} was found. If Ne and Re numbers are formulated with the material properties valid for suspensions and are represented as $Ne_{\text{sus}}(Re_{\text{sus}})$, the curves for $Re > 5 \times 10^2$ and $\varphi_{\text{v}} < 0.25 - 0.30$ coincide within experimental accuracy.

The stirrer which operates most favorably energetically is the one which requires the lowest stirrer power at the critical stirrer speed. In the turbulent flow range, for which $Fr' = \text{const}$ applies (see. Eq. 23), this condition is described by the relationship $NeFr'^{3/2} = \text{min}$. With $Ne = 0.35$ and 5.0 (see Table 1) for the propeller and turbine stirrers, respectively, and the above-mentioned constants for Equation (23) it can be calculated that the propeller stirrer only requires 28 % of the power of the turbine stirrer. From an energetic point of view stirrer types which convey the liquid axially downwards are clearly superior to axially operating ones.

4.5. Scale-up of the Solids Suspension Process

Studies in the 1980s and 1990s yielded different scale-up rules [1]. It should be recognized that always discrepancies occur between the individual studies. It can, however, be taken as proven, that two flow ranges exist, whereby in the turbulent range ($D \geq 1.0$ m, low viscosity medium, high w_{ss}) the scale-up rule P/V = idem can be seen as reliable.

4.6. Effect of Geometric and Apparatus-Related Factors on the Suspension

In this context four aspects are worth mentioning:

1) HERNDL [74] investigated the influence of the d/D ratio on the critical stirrer speed for suspension and found that it decreased approximately with the inverse of the square of the stirrer diameter:

$$n_{cr} \propto d^{-1.88}$$

This result shows that from an energetic standpoint it is most favorable to operate with a larger stirrer and correspondingly smaller stirrer speed, since the above expression corresponds to

$$P \propto d^{-0.64}$$

2) CHUDACEK [75], [76] investigated the influence of different bottom shapes on the suspension process. He used tanks with flat bottoms, with a completely profiled bottom (a) and a bottom provided with a cone in the middle and a sloping rim along the walls (b). As regards the dependence $P/V = f\ (h*)$, construction (a) always proved to be the best, if materials which are easily suspended are being dealt with. Bottom profile (b) is almost equally suitable to (a) in the case of rapidly sedimenting particles, otherwise it is intermediate in suitability between the profiled and flat bottoms. Only if slurry above the complete suspension has to be maintained in suspension, is (a) slightly more suitable. In summary, design (b) represents a considerable improvement over a flat bottom and is considerably simpler and cheaper than the fully profiled bottom (a).
3) BUURMAN et al. [77] have established through measurements (1-s criterion) in two differently sized tanks, equipped with a four-pitched blade stirrer ($a = 45°$, $d/D = 0.4$), that in the flow range investigated by them the scale-up rule P/V = const applies and that thin stirrer paddles are advantageous for saving power, because they produce eddies in the lower range of the turbulence spectrum, which is determining for suspension process. This corresponds to the Kolmogorov microscale $\lambda \approx 40\ \mu$m. BUURMAN et al. found:

$$Fr'\ (d/d_p)^{1/3} = \text{const.}\ (b/d)^{1/2}$$

where b is the paddle thickness.
Since for strength reasons b cannot be reduced without any limit, it is suggested that the paddle edges be slanted (for details see [1]). Measurements showed, that in this way the stirrer speed could be reduced by 5% and the stirrer power reduced by 17 %.
4) Turbine stirrers with larger paddle surfaces than the standardized (Rushton) turbine stirrer, but each provided with six perforations, which are sufficiently large so that they have an identical active paddle surface, exhibit a significantly lower power characteristic than standard turbine stirrer [78].

4.7. Mass Transfer in the Solid – Liquid (S/L) System

The enhancement of mass transfer in the solid – liquid system is a frequent stirring operation. Very often salts must be dissolved in the liquid, to prepare a salt solution or to initiate a chemical reaction. To accelerate the dissolution process, the whole surface of the solid particles must be wetted as completely as possible by the liquid and the liquid flow should be turbulent, so that the boundary layer on the liquid side is small and the transfer of the dissolved material to the bulk of the liquid proceeds rapidly.

The thermodynamic reason for the mass transfer in the solid – liquid system is the leveling out of the chemical potentials μ between the two phases. In the solid – liquid system, the change from $\Delta\mu$ to Δc is simpler to comprehend than in the G/L system, because the equilibrium concentration in the phase boundary S/L is given here by the saturation concentration $c* = c_s$, which immediately adjusts to the thermodynamic equilibrium in the phase boundary.

In comparison with mass transfer in gas-liquid (G/L) systems solid-liquid systems have two advantages: (a) the exchange surface is independent of the stirring intensity; (b) the saturation concentration c_s is only a function of temperature and the physical properties of the liquid phase.

The process characteristic of mass transfer in the S/L system reads

$$\{Sh, Ar, Sc, Re, d_p/d)\}$$

where

$Sh \equiv k_L d_p/D$	is the Sherwood number
$Ar \equiv d_p^3\, g\Delta\varrho/\varrho\, \nu^2$	is the Archimedes number
$Sc \equiv \nu/D$	is the Schmidt number
$Re \equiv n\, d^2/\nu$	is the Reynolds number
d_p/d	is the geometric ratio

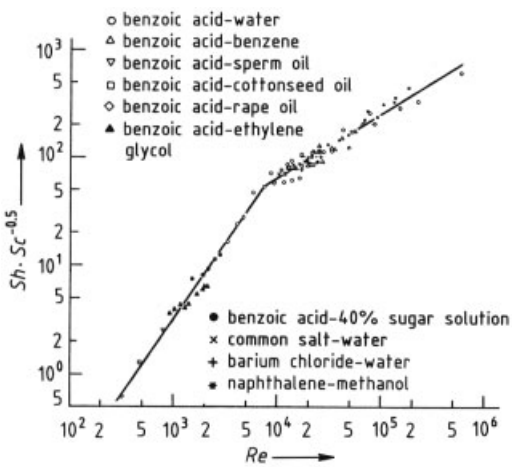

Figure 15. Dependence of the process characteristic of mass transfer in the S/L system upon the state of suspension; from [79]

As early as 1941 Hixon and Baum [79] pointed out that the process characteristic of this pi space depended essentially on the particular suspension condition concerned (see Fig. 15). Before the condition of complete suspension, the fraction of the particles, that are suspended and wetted by liquid, is particularly increased with increasing stirrer speed. Subsequently only the liquid boundary layer is further reduced. This is also expressed in the two process characteristics:

$$Re < Re_{cr}: Sh\, Sc^{-0.5} \propto Re^{1.44}$$

$$Re > Re_{cr}: Sh\, Sc^{-0.5} \propto Re^{0.58}$$

For the suspension conditions in the system S/L/G and the mass transfer involved see [1].

5. Dispersion in Liquid – Liquid (L/L) Systems

Dispersion in liquid – liquid (L/L) systems is associated with the enlargement of the interface area between two immiscible liquids, so that, e.g., an extraction process or a chemical reaction (saponification, nitration, etc.) can proceed rapidly or dispersions of particular droplet size are produced (bead and suspension polymerizations, etc.). In this chapter only dispersion by stirrers is considered. If this process is assisted by the addition of surface-active substances, it is termed emulsification, for which completely different laws generally apply.

In general the phase with the smaller volume is mixed into the other, continuous phase and broken up into droplets therein (dispersed phase). However, in the case of phases with comparable volume fractions, the possibility of inversion exists, which phase becoming the continuous one depends on the dispersion starting conditions (e.g., in which phase the stirrer is to be found) and on the physical properties of the system. For details see [1].

In dispersion in L/L systems the quantities of interest are the lowest stirrer speed at which one phase is completely mixed into the other, the achievable droplet size or droplet size distribution, the size of the interfacial area and the mass-transfer coefficient.

5.1. Lowest Stirrer Speed for Dispersion

In analogy with suspension in S/L systems, for which the lowest, critical stirrer speed at which

small particles are no longer to be found on the bottom is important, that stirrer speed at which a phase is completely stirred into the other is of interest in dispersion in L/L systems.

It is to be expected that this lowest stirrer speed $n_{\min}$ is dependent on the material system and the stirrer type. Many works treated this problem and produced similar process equations. These can be reduced to the relationship $n_{\min}\, d^{\mathrm{a}}$ = idem. Two statements are possible:

1) How does $n_{\min}$ behave upon scaling-up ($\mu \equiv d_{\mathrm{T}}/d_{\mathrm{M}}$ = scale factor)? Answer:

$$\frac{(n_{\min})_T}{(n_{\min})_{\mathrm{M}}} = \mathrm{const}\,\mu^a \qquad a = -2/3 \text{ to } -3/4$$

2) How does $(P/V)_{\min}$ behave upon scaling-up? Answer:

$$\frac{(P/\varrho V)_{\mathrm{T}}}{(P/\varrho V)_{\mathrm{M}}} = \mathrm{const}\,\mu^b\,;$$
$$b \equiv (2-3a) = -2.8 \text{ to } -3.1$$

It is evident, that both $n_{\min}$ and also the associated stirrer power per unit volume become smaller upon scaling-up.

5.2. Dispersion Characteristics

During dispersion a steady-state situation is established, in which the droplets formed coalesce to ever larger droplets and are broken up anew by the stirrer. The droplet size distribution of the dispersion produced by stirring is usually characterized by an average droplet size d_{32}, which is defined by the expression:

$$d_{32} \equiv \frac{\sum (n_i\, d_i^3/6)}{\sum n_i\, d_i^2}$$

where n_i and d_i are the number and diameter of the droplet of the class i.

The average droplet diameter d_{32} is obtained from the sum of the volumes of all the spherically shaped droplets, divided by their surface area. This definition is due to J. Sauter [80], who as early as 1926 introduced it in his study on fuel gasification in carburetors. Thus d_{32} is known in the English-language literature as the Sauter mean diameter.

The following relationship exists between d_{32} and the interfacial area a per unit liquid volume:

$$a \equiv \frac{A}{V} = \frac{\sum \pi\, d_i^2}{\sum (\pi\, d_i^3/6)} = \frac{6\,\varphi_{\mathrm{v}}}{d_{32}}$$

The process relationship for the target quantity d_{32} (or for a) can be formulated either with extensive or intensive quantities. The difference lies in the choice of the process parameter. The dispersion characteristic formulated with *extensive* quantities uses as the process parameter the extensive quantity stirrer speed and leads, assuming a given geometry (stirrer type; D/d, H/d, h/d = const), to the following dependence (the index d indicating the physical properties of the dispersed phase, no index indicating the physical properties of the continuous phase):

$$d_{32} = f\,(d;\, \varrho_{\mathrm{d}},\, \varrho,\, \mu_{\mathrm{d}},\, \mu,\, \sigma,\, \varphi_{\mathrm{v}};\, g\Delta\varrho,\, n)$$

From this follows the pi set:

$$\{d_{32}/d,\, We,\, Re,\, Fr',\, \varrho_{\mathrm{d}}/\varrho,\, \mu_{\mathrm{d}}/\mu,\, \varphi_{\mathrm{v}}\}$$

where $We \equiv \varrho\, n^2\, d^3/\sigma$ represents the Weber number.

Most process characteristics were represented in this pi space. For the six-blade turbine stirrer it was correspondingly found that:

$$d_{32}/d = \mathrm{const}\; We^{-0.6} f\,(\varphi_{\mathrm{v}}) \tag{26}$$

and most researchers have found that const = 0.05 – 0.08, see [1].

The relationship $(d_{32}/d)\; We^{0.6}$ = const can be derived with the help of the Kolmogorovs theory of isotropic turbulence (see Section 1.4). However, this relationship only applies, if the boundary conditions for the presence of isotropic turbulence are fulfilled, i.e., if the flow in the tank is turbulent ($Re > 10^4$), both liquids possess comparable viscosities, and the volume fraction of the dispersed phase is low.

Droplet disintegration is caused by the kinetic energy of the turbulence eddies. Eddies whose size λ corresponds approximately to the droplet diameter d_{32}, i.e. $\lambda \approx d_{32}$, have the strongest dispersing action. For $\lambda \gg d_{32}$ the droplet is carried along by the flow and thereby is hardly deformed, whereas for $\lambda \ll d_{32}$ the eddies possess insufficient kinetic energy to cope with the surface tension.

Details concerning the effects of coalescence and of φ_v on d_{32}, and of the effect of viscosity and of the stirring duration are given in [1]. Here, the evaluation of the droplet size distribution is also presented which uses the *intensive* quantity P/V instead of n.

5.3. Stirrer Power for Dispersion

In dispersing of L/L systems, the same power characteristics apply as with mixing in homogeneous material systems, if for the liquid density the average density of the dispersion

$$\varrho_m = \varrho + \varphi_v \, \Delta\varrho$$

and for the viscosity of the liquid the weighted geometric mean of the viscosities of both phases

$$\mu_m = \mu_d^{\varphi_v} + \mu^{(1-\varphi_v)}$$

or the physically more understandable expression

$$\mu_m = \frac{\mu}{1-\varphi_v}\left(1 + 1.5\,\varphi_v \frac{\mu_d}{\mu_d + \mu}\right)$$

is used.

5.4. Scale-Up of the Dispersion Processes

Starting from the dispersion characteristic $d_{32}/d \propto We^{-0.6}$, Equation (26), it follows that the scale-up criterion is:

$$(d_{32}/d)\, We^{0.6} = d_{32}\,(\varrho/\sigma)^{0.6}\,(n^3\, d^2)^{0.4} = \text{idem} \qquad (27)$$

From this the following scaling-up rule for the stirrer speed is obtained:

$$n_T = n_M\, \mu^{-2/3} \qquad (28)$$

If in the scale-up criterion the expression $n^3\, d^2$ is replaced by the equivalent expression for the turbulent flow range, $P/\varrho V$, the following expression is obtained

$$d_{32}\,(\varrho^{0.2}/\sigma)^{0.6}\,(P/V)^{0.4} = \text{idem} \qquad (29)$$

which states, that the average droplet size and thus the same volume-related interface area a is maintained, if, both on a small or large scale, the same stirrer power per unit volume is utilized.

5.5. Mass and Heat Transfer Upon Dispersion

Mass transfer in the L/L system proceeds via the boundary layer of the continuous phase, which surrounds the droplet. The mass-transfer coefficient k_c (index c denotes the continuous phase) depends on the flow conditions, the droplet size, and the physical properties of the surrounding continuous phase. It can be assumed, that mass transfer inside the droplet proceeds via diffusion.

In 1961 Calderbak and Moo-Young [81], carried out a comparative investigation on the mass and heat transfer in bubble columns and columns with sieve or sintered plate spargers in the G/L and L/L phase systems. They arrived at relationships, which apply for both phase systems ($k_L = k_c$). Both k_L and also the film heat-transfer coefficient h only depended upon the material properties and not upon the flow condition (relative velocity). Two process relationships were obtained, which apply for different diameter ranges for the dispersed phase. For particle sizes $d_p \leq 2.5$ mm, the droplet-bubble surface did not oscillate; the flow between the two phases was hindered. Thus so-called rigid spheres were involved. On the other hand for $d_p \geq 2.5$ mm, the dispersed phase moved freely under the influence of the acceleration due to gravity and free convection between the two phases could occur. The corresponding process relationships are as follows:

Bubble/droplet diameter $d_p \leq 2.5$ mm; rigid spheres:

$$Sh = 0.31\,(Ar\,Sc)^{1/3} \quad \rightarrow \quad k_L\left(\frac{\varrho^2}{\mu\, g \Delta\varrho}\right)^{1/3} \propto Sc^{-2/3}$$

$$Nu = 0.31\,(Ar\,Pr)^{1/3} \quad \rightarrow \quad \frac{h}{c_p\,\varrho}\left(\frac{\varrho^2}{\mu\, g \Delta\varrho}\right)^{1/3} \propto Pr^{-2/3}$$

Bubble/droplet diameter $d_p \geq 2.5$ mm; oscillating bubble surfaces:

$$Sh = 0.42\,Sc^{1/2}\,Ar^{1/3} \quad \rightarrow \quad k_L\left(\frac{\varrho^2}{\mu\, g \Delta\varrho}\right)^{1/3} \propto Sc^{-1/2}$$

$$Nu = 0.42\,Pr^{1/2}\,Ar^{1/3} \quad \rightarrow \quad \frac{h}{c_p\,\varrho}\left(\frac{\varrho^2}{\mu g \Delta\varrho}\right)^{1/3} \propto Pr^{-1/2}$$

This fundamental study, which also serves as an excellent example of the analogy between mass and heat transfer, only gives the k_L or k_c values for the flow conditions in tray columns, i.e., for flow processes, in which the final droplet or bubble size is achieved within or in the immediate neighborhood of the sparger.

Schindler and Treybal [82] investigated the mass transfer in the L/L system in continuously operated stirred tanks with and without baffles. They established that k_c increased according to $k_c \propto (P/V)^{0.55}$ and that in a tank without baffles the stirrer for the same power realized a ca. 2.5-fold k_c compared with the tank with baffles. This and the effect of the residence time of the droplets in the stirred tank are the reasons for the positive effect of droplet coalescence and redispersion, which renew the boundary layer of the surrounding phase.

Keey and Glen [83] used the same experimental apparatus as [82], but worked with a ternary material system and used other methods for determining the volume-related interfacial area. However, they could satisfactorily confirm the above results. If the dependence $k_c = f(P)$ is approximated with $k_c \propto (P)^{0.5}$, it follows:

$$k_c = 1.55 \times 10^{-4}\,P^{0.5}$$

or with respect to the tank volume used ($V = 11.0$ L):

$$k_c = 1.7 \times 10^{-2}\,(P/V)^{0.5}$$

where k_c is in cm, and P in W or P/V in W/m^3.

For theoretical approaches see [1].

6. Dispersion in Gas – Liquid (G/L) Systems

In this operation, generally termed gas – liquid contacting or simply gassing, the aim is to increase the interfacial area between the gas and the liquid. This important operation is utilized in the chemical industry in hydrogenation, chlorination, oxidation, etc. In fermentations and in aerobic wastewater purification, supply of oxygen is one of the most important process engineering operations. Stirred tanks and bubble columns are typically used for this operation in the chemical and pharmaceutical industries, whereas flat ponds and recently high cylindrical tanks ("towers") are being used in wastewater purification.

The target quantity of the gassing process is the absorption rate in the G/L system. The absorption rate is directly proportional to the interfacial area between the gas phase and the liquid phase. The limiting factor is the diffusion of the dissolved gas through the liquid side of the boundary layer, which can be affected by the thickness of the layer only to a limited extent. A substantial intensification of mass transfer is only possible by increasing the G/L interfacial area: gas sparging by means of stirrers, nozzles, sintered or perforated plates, etc., should therefore effect a dispersion of the gas into fine bubbles.

Unlike the mass transfer in the S/L system (see Section 5.5), the G/L interfacial area is not constant, but depends in addition to the process parameters also on specific physical properties of the liquid treated, which significantly contribute to the size of the gas bubbles.

6.1. Physical Fundamentals of Mass Transfer

The driving force in this process is the deviation from the thermodynamic equilibrium and thus the chemical potential μ. The mass transfer ceases, when the equilibrium is reached at the G/L interface.

With the two-film theory of Lewis and Whitman [84] two conditions are postulated for this transfer process:

1) On either side of the phase boundary two stationary laminar fluid layers (films) are formed with thicknesses δ_G and δ_L, through which the gas can only pass by diffusion.
2) At the interface a thermodynamic equilibrium immediately arises. This means that at the phase boundary the equilibrium concentration, i.e., the saturation concentration prevails:

$$c* = c_s = f\,(i, T, p)$$

The two-film theory therefore enables the chemical potential difference $\Delta\mu$ to be replaced by the concentration difference Δc of the gas between the gas phase and the phase boundary, as well the phase boundary and the liquid phase, respectively. However, the concentration change from one side of the phase boundary to the other is then discontinuous.

This discontinuity is eliminated using Henry's law

$$p = Hy\, c$$

where Hy is the Henry constant, a function of material conditions and temperature. Taking this into account, the concentration change from phase to phase is continuous [1].

According to the first postulate of the two-film theory, an overall mass transfer occurs. There are two mass transfer coefficients, k_G and k_L, which are defined by their corresponding film thicknesses δ and diffusion coefficients D:

$$k_i \equiv D_i/\delta_i$$

The mass transfer equation formed with the mass flux j_m then reads:

$$j_m \equiv G/A = k_G\,(p - p*) = k_L\,(c* - c) \tag{30}$$

Typical diffusion coefficients D for gases are approximately a factor of 10^4 to 10^5 higher than for liquids, hence $D_G \gg D_L$ applies and according to Equation (30) also $k_G \gg k_L$, i.e., resistance to gaseous transport $1/k_G$ is negligible compared with $1/k_L$.

The only important exception, in which $k_G \ll k_L$ has to be assumed, is the chemisorption of a gas component from a gas mixture in flue gas scrubbing, since here depletion of the gas to be adsorbed is involved and in addition the concentration of the gas is already zero at the interface due to the extremely fast chemical reaction. Since gas scrubbing is not normally carried out in stirred tanks, this exception will not be considered here.

Thus Equation (30) becomes:

$$j_m \equiv G/A = k_L\,(c* - c) \rightarrow G = k_L A\,\Delta c \tag{31}$$

The mass flow G of the gaseous component is directly proportional to the mass transfer coefficient on the liquid side, k_L, the mass exchange surface A and the concentration difference Δc.

The exchange surface A is unknown and with k_L is itself a function of the physical properties and the process-related parameters. This is the reason why the target quantity in mass transfer is, as a rule, a product of these two quantities. $k_L A$ is the target quantity in surface aeration:

$$k_L A = G/\Delta c \tag{32}$$

If gas – liquid contacting takes place in the liquid volume, e.g., in stirred tanks and in bubble columns, a quasi-uniform material system is usually assumed (same bubble density in each volume element) and the mass flow G is related to the (ungassed) volume of liquid V:

$$G/V = k_L a\,\Delta c \quad k_L a = \frac{G}{V\,\Delta c} \qquad a \equiv A/V \tag{33}$$

Therefore $k_L a$, the volume-related liquid-side mass-transfer coefficient, rather than k_L is the aim of most investigations in this area.

$k_L a$ is usually referred to in the English-language literature by the term "volumetric liquid-side mass-transfer coefficient", which is not strictly correct, because the mass transfer is not recorded by a volumetric measurement, but merely is related to the volume. Here, the term "volume-related liquid-side mass transfer coefficient" will be used.

For temperature dependence of $k_L a$ as well the saturation concentration c_s of the gas in different liquids, see [1]. This source also contains information on different measuring methods for $k_L a$: Unsteady-state methods (using oxygen electrodes or manometric method), steady-state methods (sulfite method, hydrazine method, hydrogen peroxide method).

6.2. Mass-Transfer Characteristics in the G/L System

Most of the measurements concerning mass transfer in G/L systems were carried out with the turbine stirrer, because very early it became apparent, that stirrers with resistance due to their shape are the most suitable for dispersing gas. (There are indeed measurements, which prove that turbine stirrers with twelve paddles are energetically more favorable than those with six paddles [94]). In several publications differently

designed hollow stirrers have been utilized. Hollow stirrers are also excellent gas dispersers in addition to their ability of sucking the gas in.

Under given geometric and material conditions the following quantities are involved:

Target quantity: $k_L a \equiv G/(V\,\Delta c)$

Process parameters: P/V and $v \equiv q/S$

where q is the gas throughput and S is the cross-sectional area of the working space.

Already in 1979 VAN'T RIET [85] evaluated and compared 76 publications in which the turbine stirrer had been utilized as gas disperser. He derived the following dimensionally formulated mass-transfer relationships:

a) Ion-free water (complete coalescence):

$$k_L a = 2.6 \times 10^{-2}\,(P/V)^{0.4}\,v^{0.5} \pm 20-40\,\% \qquad (34)$$

b) Ion-containing water (suppressed coalescence):

$$k_L a = 2.0 \times 10^{-3}\,(P/V)^{0.7}\,v^{0.2} \pm 20-40\,\% \qquad (35)$$

The boundary conditions in both cases were: $V = 0.002-2.6\,\mathrm{m}^3$; $P/V = 0.5-10\,\mathrm{kW/m}^3$. The constants applied for P/V in $\mathrm{W/m}^3$, v in m/s, and $k_L a$ in s^{-1}.

According to the above findings, the stirrer power must be formulated as P/V and the gas throughput as $v = q/S$. The following relevance list:

$$\{k_L a;\ \varrho,\ \mu,\ D,\ \sigma;\ P/V,\ \mathrm{v},\ g\}$$

is therefore the starting point from which the following set of pi numbers results:

$$(k_L a)^* \equiv k_L a \left(\frac{\nu}{g^2}\right)^{1/3};\ (P/V)^* \equiv \frac{P/V}{\varrho\,(\nu g^2)^{1/3}};$$
$$\mathrm{v}^* \equiv \frac{\mathrm{v}}{(\nu g)^{1/3}};\ Sc \equiv \frac{\nu}{D}\quad \sigma^* \equiv \frac{\sigma}{\varrho\,(\nu g^4)^{1/3}}$$

6.2.1. Sorption Characteristics in Coalescence-Bound Systems (Water)

JUDAT [86] compared the results of twelve publications in the area of coalescing systems. The measurements were carried out both under unsteady-state (nine publications) and under steady-state (hydrazine method; three publications) test conditions, in which the geometric parameters were varied within extraordinarily wide boundaries: $d = 0.05-3.1$ m; $D = 0.15-12.2$ m; $H = 0.15-6.1$ m; the volume of the test tank varied in the range $V = 2.5$ L to $906\,\mathrm{m}^3$! The geometric numbers ranged from $D/d = 2.0-7.3$; $h/d = 0.3-1.1$; $H/D = 0.5-1.0$ and the gas throughput number $B* = (\pi/4)\,\mathrm{v}* = 0.01-1.5$. This test material could be described with an accuracy of $\pm 30\,\%$ just with the two intensity quantities $(P/V)*$ and $B*$, see Figure 16:

$$(k_L a)* = 9.8 \times 10^{-5}\,(P/V)^{*0.40}\,(B*^{-0.6} + 0.81 \times 10^{-0.65/B*})^{-1} \qquad (36)$$

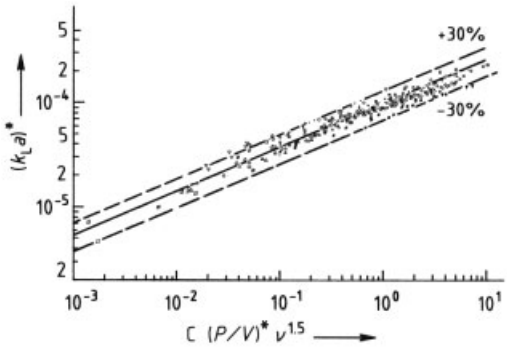

Figure 16. Sorption characteristic $k_L a* = f\ \{(P/V)*,\ B*\}$ for fully unsuppressed coalescence (water – air) [86]; Superficial velocity of the gas $B* \equiv (\pi/4)\,\nu*$

If the results obtained from Equation (36) are compared with those from Equation (34), it turns out that Equation (36) provides ca. 30 % higher values than Equation (34), the values from (34) being on the bottom edge of the $\pm 30\%$ spread. Thus the simpler Equation (34) can be used and 30 % added.

Measurements of $k_L a$ in low viscosity sugar, starch, and fermentation solutions performed in laboratory ($D = 0.2$ m; $V = 8$ L) and pilot-plant fermenters, for which $H/D = 1.3-1.7$ and one to three turbine stirrer on the same shaft were utilized [105] were evaluated within the same dimensionally analytical framework. It was found that:

$$(k_L a)* = 7.9 \times 10^{-5}\,(P/V)^{*0.34}\,\mathrm{v}^{*0.39} \pm 30\,\%$$

6.2.2. Sorption Characteristics in Coalescence-Inhibited Systems

The addition of some inorganic electrolytes (so-called "structure builders") or some organic compounds (e.g., aliphatic alcohols) to water changes the water structure in aqueous solutions, see [92]. This effect strongly influences the coalescence of primarily produced gas bubbles and is comprehensively treated in [1], Section 4.10.

Information about the sorption characteristics of coalescence-inhibited systems was obtained using hollow stirrers (type: three-edged stirrer) in geometrically similar stirred tanks in scales $\mu = 1-5$ with the material system 1 N sodium sulfite solution [88], [89]. In self-aspirating hollow stirrers v is not a freely selectable process parameter! The dependence $(k_L a)* = f\,(P/V)*$ is shown in Figure 17. The sorption characteristic was:

$$(k_L a)* = 1.1 \times 10^{-4}\,(P/V)^{*0.8} \tag{37}$$

In respect of P/V it corresponded well with the process relationships of VAN'T RIET (Eq. 35):

$$k_L a = 2.0 \times 10^{-3}\,(P/V)^{0.7}\,v_G^{0.2}$$

or from ROBINSON and WILKE [92]:

$$k_L a \propto (P/V)^{0.9}\,v^{0.39}$$

in which the gas throughput (i.e., the superficial velocity v) was a freely selectable process parameter.

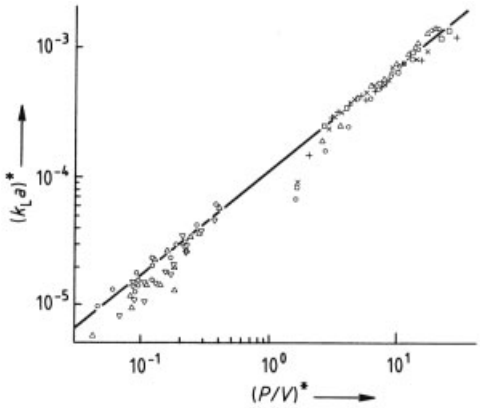

Figure 17. Sorption characteristic $(k_L a)* = f\,(P/V)*$ for self-aspirating hollow stirrers for full coalescence suppression, [88], [89]
$H* \equiv H - h$ = liquid height above the stirrer

6.2.3. Sorption Characteristics in Non-Newtonian Material Systems

With the advance of fermentation technology (production of antibiotics, citric acid, essential amino acids, etc.) interest has increased in the supply of atmospheric oxygen to more viscous and non-Newtonian liquids. Numerous publications are dealing with these material systems, see [1]. Here, only one by HENZLER [91] is presented. He correlated the same experimental results as JUDAT [86] by using a different dimensionless frame:

$$Y' \equiv \frac{k_L a}{v}\left(\frac{\nu_{eff}^2}{g}\right)^{1/3} = f\left(X' \equiv \frac{P}{Vv}\frac{1}{\varrho g}\right)$$

where Y' is the sorption number for bubble columns, see [100], while X' represents the ratio of the mechanical power of the stirrer to the hydraulic power of the gas throughput. In this frame also measurements with aqueous glucose and millet mash (Newtonian liquids) and with aqueous CMC solutions (non-Newtonian liquid, showing pseudoplastic and viscoelastic flow behavior) are presented (see Fig. 18).

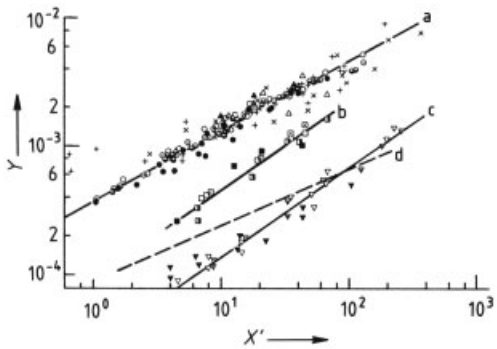

Figure 18. Sorption characteristics of some Newtonian and non-Newtonian liquids in the evaluation framework according to HENZLER [91]; the incorporation of the *Sc* number is not necessary. Tank with turbine stirrer ($d/D = 0.24-0.4$; $H/D = 1$)

6.3. Stirrer Power in Gassed Liquids

In mixing vessels gas is continuously introduced into the liquid either via a tube or by a sparger

centrally positioned under the stirrer. The task of the stirrer is to break up the gas flow into very small gas bubbles, i.e., to produce a G/L dispersion.

Knowledge of the stirrer power in a gassed liquid is necessary for dimensioning the stirrer drive. Otherwise, the stirrer power also determines the phase boundary per unit volume in the gas – liquid system and thereby the mass transfer itself.

The power P of a given stirrer under given installation conditions in an *ungassed* liquid depends upon the stirrer diameter d, the physical properties of the liquid (ϱ and ν) and the stirrer speed n. In a *gassed* liquid the following process parameters must be added to this list: the gas throughput q and the acceleration due to gravity g, since a material system with density differences $\Delta\varrho$ is only effective as $g\Delta\varrho$ under the influence of gravity. In contrast to those of the liquid phase the physical properties of the gas phase (ϱ' and ν') are certainly negligible, but not a priori the surface tension σ and the, up to now, unspecified physical properties S_i, which regulate the gas bubble coalescence and control the size of the gas bubbles. The smaller the stable gas bubble size in the system, the lower the tendency of gas bubbles to leave the system, thus the higher the gas content in the liquid and the lower the density of the material system. The relevance list is here, for given geometric conditions, much more extensive as compared with expression for the stirrer power in a pure liquid (see Section 2.3.2):

$$P;\ d;\ \varrho, \nu, \sigma, S_i;\ n, q, g\Delta\varrho \tag{38}$$

and is compressed by means of dimensional analysis to the following set of six pi numbers:

$$\{Ne, Q, Fr', Re, We, S_i*\} \tag{39}$$

The meaning of Ne and Re is the same as in the power characteristic for a homogeneous system (see Fig. 10) the other abbreviations stand for the following pi numbers:

Gas throughput number	$Q \equiv \frac{q}{n\,d^3}$
Froude number	$Fr' \equiv \frac{n^2\,d\,\varrho}{g\Delta\varrho}$
Weber number	$We \equiv \frac{n^2\,d^3\,\varrho}{\sigma}$
Coalescence numbers	S_i^*

At low process pressures the quotient $\Delta\varrho/\varrho$ in Fr' becomes $\cong 1$, whereupon the extended Froude number Fr' reduces to $Fr \equiv n^2 d/g$.

Zlokarnik [93] found no effect of We and S_i* in measurements with a turbine stirrer of conventional design in pure liquids and in glycerol-water mixtures (coalescing material systems!). Surprisingly, the Froude number also had no effect on Ne in the industrially interesting range of $Fr > 0.65$, so that for $Re \geq 10^4$ the power characteristic could just be described by $Ne(Q)$ (Fig. 19). The process equation is:

$$Ne = 1.5 + (0.5\,Q^{0.075} + 1600\,Q^{2.6})^{-1} \tag{40}$$

Validity range: $Q \leq 0.15$; $Fr \geq 0.65$; $Re \geq 10^4$ and $D/d \geq 5$.

For an extended D/d range ($D/d \geq 2.2$) and $Re > 10^3$ see [1].

Judat has comprehensively extended this work. In [94] $Ne(Q)$ relationships are given for different stirrer types, namely for perforated disk stirrers with twelve holes, for three-wane propeller stirrers with a pitch $\alpha = 26°$, for pitched-blade stirrer with a pitch $\alpha = 45°$ and for conventional turbine stirrers with different numbers of blades $i = 6$; 12; 18; 24. The measuring range of Q was extended for turbine stirrers to $Q \approx 1.0$. It was shown that in the Q range 0.2 to 1.0 Equation (40) did not apply; with increasing Q, Ne decreases further.

For the power characteristics of non-Newtonian fluids see [1], [95], [96].

6.4. Flooding Point

The most important boundary condition for the application of the proposed power characteristic $Ne(Q, Fr, D/d, h/d)$ is given by the upper boundary value of the gas throughput number Q. At a particular stirrer speed each stirrer can only disperse a particular maximum gas throughput in the liquid. Upon exceeding this value the stirrer is "flooded" by gas, i.e., it is completely enveloped by gas and is incapable of dispersing it.

In dispersion processes in gas – liquid systems gas cushions are developed in so-called suction regions behind the surfaces of the stirrer hit by the flow, whereby the pumping ability of the stirrer for the liquid (circulation flow)

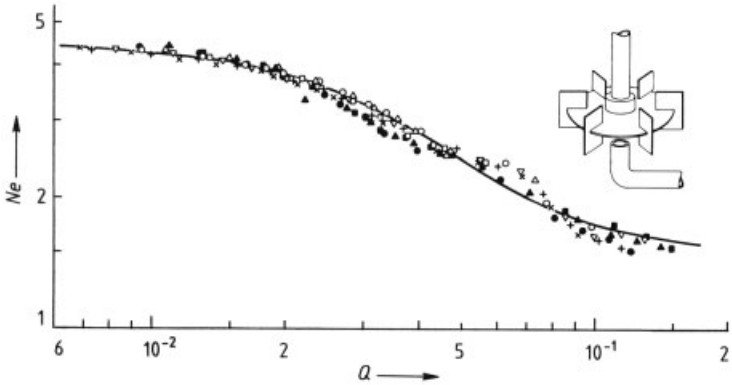

Figure 19. Power characteristic $Ne(Q)$ of a turbine stirrer in the air-water system ($H/D = 1$, $h/d = 1$) [93]; the curve corresponds to Equation (40)

ceases. Thereby the discharge of the gas dispersion is also diminished, the gas cushions becomes ever larger and at a particular value of the gas throughput number $Q_{\max}$ the stirrer is flooded with gas.

The $Q_{\max}(Fr)$ dependence for the turbine stirrer with six paddles (Rushton turbine) is given by the process equation

$$Q_{\max} = 0.19\, Fr^{3/4} \quad 0.1 < Fr < 2.0$$

For the $Q_{\max}(Fr)$ dependencies of the perforated disk stirrers with twelve holes, for three-wane propeller stirrers with a pitch $\alpha = 26°$, for pitched-blade stirrer ($\alpha = 45°$) and for conventional turbine stirrers with different numbers of blades $i = 6$; 12; 18; 24 see [94] or [1].

In order to obtain information on the efficiency $P/q_{\max}$ of these types of stirrer, the results $Ne(Q)$ and $Q_{\max}(Fr)$ were linked to one another by introducing two new dimensionless numbers:

$$\frac{Ne\,Fr}{Q_{\max}\,(D/d)} \equiv \frac{p}{q_{\max}\,\varrho\, g\, D} \quad \text{and}$$

$$\frac{Q_{\max}\,Fr^{1/2}}{(D/d)^{5/2}} \equiv \frac{q_{\max}}{g^{1/2}\,D^{5/2}}$$

The relationship between these numbers is shown in Figure 20 and proves that turbine stirrers perform best even with respect to the efficiency of gas dispersion. In this regard the twelve-blade turbine stirrer is clearly superior to the usual six-blade turbine stirrer.

In gassing viscous liquids, the flooding point was only clearly observable up to a viscosity of approx. $\nu = 3 \times 10^{-4}$ m^2/s. At higher viscosities a gradual change in the flow conditions occurred, whereby a clear definition of the maximum gas throughput was rendered impossible [96]. Measurements were performed with a turbine stirrer ($d/D = 0.3$) in aqueous glycerol solutions.

6.5. Special Gas – Liquid Contacting Techniques

In this section a few special stirrer types and nozzles will be shortly presented.

6.5.1. Hollow Stirrers

Hollow stirrers are designed such that they can utilize the suction, which develops behind the flowed against stirrer surfaces, for sucking gas into the liquid (see Fig. 7).

The use of hollow stirrers is gaining importance particularly in working under high system pressure, because in this way an isobaric reaction can be carried out without additional gas supplying device. For working with a pure gas (e.g., 99.9 % N_2 or with H_2 in hydrogenations) the manometric (pressure gauge) method can also be used for the determination of the overall reaction kinetics, see [89] and [1]. The pressure drop in a stirred tank as a result of gas absorption is compensated by further supply of reaction gas from a storage vessel, whose weight or pressure loss is monitored as a function of time.

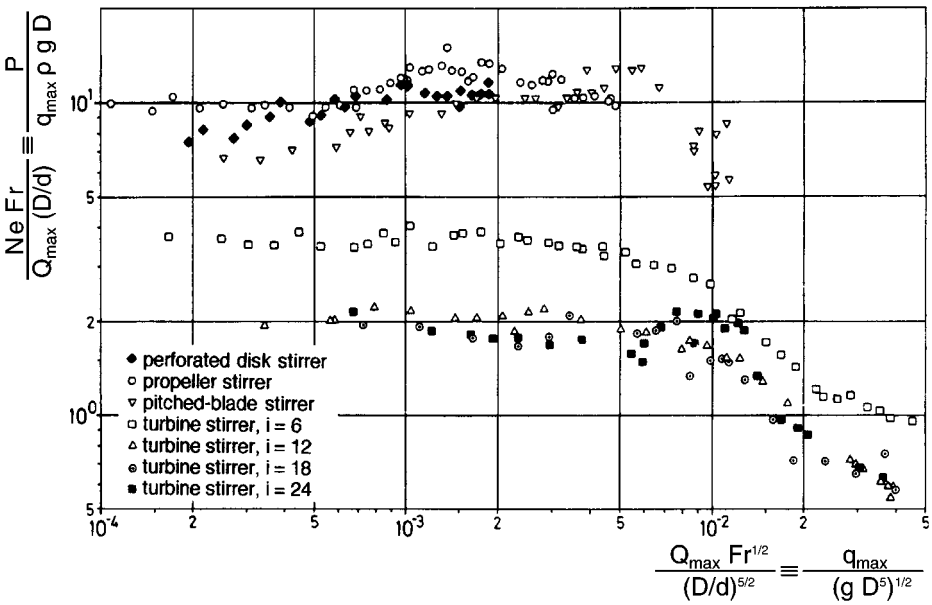

Figure 20. Efficiencies $(P/q_{\max})*$ for different stirrer types [94]

In [87], [88] the power and throughput characteristics of three-edged stirrers in the turbulent flow range (water-air system) are presented, which were determined under the geometric conditions in the scale range of $\mu = 1-5$. They show that the state of flow in the turbulent range ($Re > 10^4$) is unequivocally described by the Froude number:

$Ne = f_1(Fr)$ and $Q = f_2(Fr)$

There are stirring operations in chemical engineering in which a large gas throughput is involved (e.g., purging of volatile compounds from a liquid), for which a minimization of stirrer power input is desirable. The condition for which $P/q = \min$ can be obtained from the process characteristics $Q = f_1\,(Fr)$ and $Ne = f_2\,(Fr)$, if the Ne, Q, and Fr numbers are combined in a way that a dimensionless expression for P/q arises:

$$\frac{NeFr}{Q} \equiv \frac{P}{q\,d\,\varrho g} = f_3\,(Fr)$$

The course of this function shows that two operating ranges exist:

1) For $Fr \leq 10$, $Ne\ Fr/Q = \text{const}$ applies, i.e. $P/q \propto d$. In this range P/q increases directly proportionally with scale. (In case of a scale-up of $\mu = 1:10$, the full-scale device requires a ten-fold stirring power for the same gas throughput!)
2) For $Fr \geq 10$, hollow stirrer operate even less economically, since here $NeFr/Q$ increases approximately proportionally with Fr: i.e., $P/q \propto d^2$!

Here, the gas throughput and the stirrer power must be decoupled from one another, by feeding a fast rotating turbine stirrer from underneath by externally compressed gas.

However, there are many reactions in gas–liquid systems, in which a high gas throughput is not required e.g. because the microkinetics are rate-determining. In such circumstances the hollow stirrer due to its dual role as stirrer and gas pump is the stirrer type of choice, particularly in high-pressure reaction engineering.

6.5.2. Surface Aerators

Surface aeration is little used due to the unfavorable interfacial area volume ratio. For many years, the only important exception was the provision of oxygen for municipal sewage treatment

plants, until volume aeration with fine air bubbles became established in this area.

Turbine surface aerators are designed as liquid-conveying devices (turbines). They are arranged in the liquid surface. The design data were realized with classical turbine stirrers (Rushton turbine), which were arranged in the liquid surface as surface aerators [97].

The power characteristic of turbine stirrer as surface aerators shows that it has to be formulated as $Ne(Fr)$ as is to be expected. For a number of formerly common types of surface aerators these characteristics are found in [97].

The sorption characteristic of turbine stirrers as surface aerators in a given material system (water; Sc and $\sigma* = \text{const}$) is represented by the process equation [98]:

$$(k_L A)* = 1.41 \times 10^{-4}\, Fr^{1.205}\, Ga^{0.115}$$

Application range: $Fr = 0.02 - 0.34$; $Ga = (1.5 - 200) \times 10^9$; $\sigma* = 3317$; $h/d = 0.2$

The target number is:

$$(k_L A)^* \equiv \frac{G}{d^3\,\Delta c}\left(\frac{\nu}{g^2}\right)^{1/3}$$

and the process numbers are:

$$Fr \equiv \frac{n^2\,d}{g} \quad \text{and} \quad Ga \equiv \frac{Re^2}{Fr} \equiv \frac{d^3\,g}{\nu^2}$$

Through coupling the power and sorption characteristics information can be obtained as how the efficiency $E \equiv G/P$ (kg O_2/kWh) of this aerator class behaves upon scale-up to an industrial scale:

$$E^* \equiv (k_L A)^*\, Ne^{-1}\, Fr^{-1.5} \equiv \frac{G\,d^{0.5}}{P\,\Delta c}\,\varrho\,\left(\nu^2\,g^5\right)^{1/6} = f\,(Fr, Ga)$$

An evaluation of the comprehensive experimental data in [98] shows that the relationship $E* \propto Fr^{-0.35}$ exits, i.e., $E*$ decreased with increasing Fr. Therefore industrial-scale aeration should operate with the smallest possible Fr value.

If a turbine stirrer, operated as a surface aerator on a laboratory scale, was scaled up to an industrial scale and if a favorable Froude number was selected for the full scale realization, the effect of the scale-up on the efficiency E can be predicted from the above presented relation for $E*$:

$$Fr = \text{idem} \quad \rightarrow \quad E^*\,Ga^{-0.115} = \text{const} = \frac{G\,d^{0.155}}{P\,\Delta c} = \text{const}$$

For $\Delta c = \text{const}$ it follows that:

$$E_T = E_M\,\mu^{-0.144}$$

where the scale-up factor $\mu \equiv d_T/d_M$.

In a scale-up of $\mu = 4\,(10)$ the efficiency of the industrial turbine surface aerator will amount to 80 % (70 %) of that of the model; the efficiency of a stirrer surface aerator thus decreases slightly with increasing scale [98].

6.5.3. Gas Spargers and Injectors

Sintered glass or ceramics plates and perforated metal plates are classical gas spargers (dispersers) for bubble columns. Static mixers and nozzles became available with the emergence of biological wastewater purification. They all realize gas dispersion *directly* and the hydrodynamics in the processing space is solely determined by the rising bubble swarm. A stirrer, on the other hand, produces gas dispersion *indirectly* via an intensive shear flow of the liquid, which disperses the gas continuum as gas bubbles. In other words: for the stirrer to produce gas bubbles, the whole liquid content must be kept in motion!

There are numerous publications from the field of chemical reaction engineering, which deal with mass transfer in bubble columns [99]. They consistently report that the volume-related mass-transfer coefficient on the liquid side $k_L a$ at not too high superficial gas velocities ($v \leq 200$ m/h) is directly proportional to the superficial velocity v:

$$k_L a \propto v \rightarrow k_L a/v = \text{const} \tag{41}$$

This is an important information when one realizes that $k_L a \equiv \frac{G}{V\,\Delta c}$ and the superficial velocity is $v \equiv \frac{q}{S}$. Then the following expression is obtained:

$$k_L a/v = \frac{G}{V\,\Delta c} \times \frac{S}{q} = \frac{G}{q\,H\,\Delta c} = \text{const} \tag{42}$$

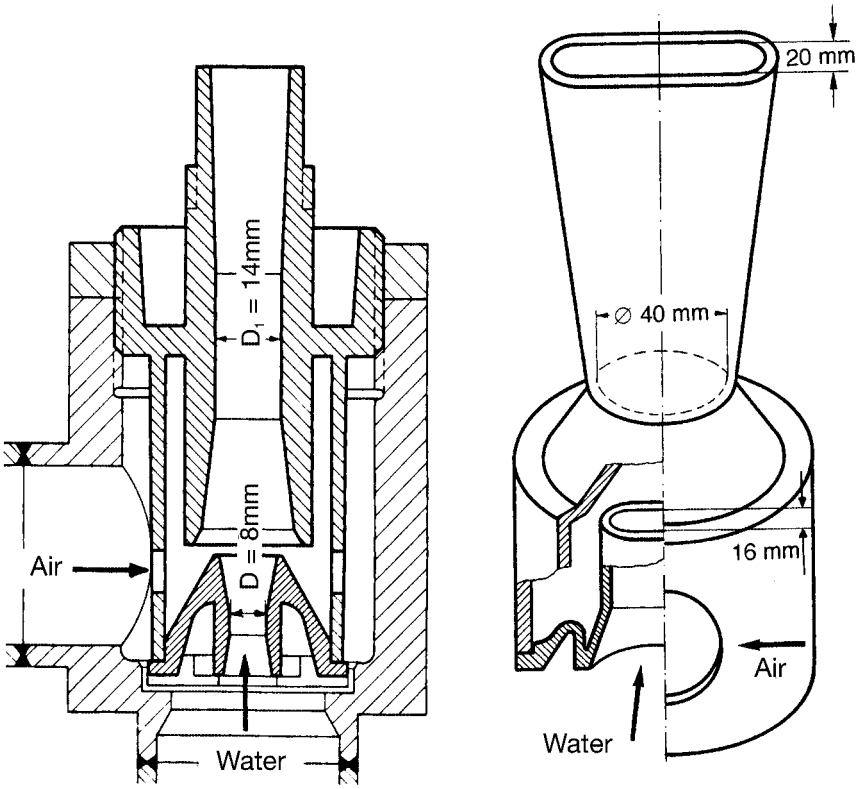

Figure 21. Bayer injectors [99] Left: injector 8/14; Right: slot injector

where the liquid volume V is the cross-sectional area S of the processing space multiplied by the liquid height H: $V = SH$.

In contrast to Equation (41), the proposition of Equation (42) is intellectually understandable, the rate of absorption G being directly proportional to the gas throughput q (number of gas bubbles), the liquid height (residence time of the gas bubbles) and the concentration difference Δc.

Spargers. For spargers with the gas throughput as the only process parameter, the target quantity $k_L a/\mathrm{v}$ should only depend upon the sparger type. In this case dimensional analysis provides only one target number which is called the sorption number Y:

$$Y \equiv \frac{k_L a}{\mathrm{v}} \left(\frac{\nu^2}{g}\right)^{1/3} = f\,(Sc,\, Si_i^*,\, \text{sparger type}) \tag{43}$$

The sorption number may depend on physical properties of the system (Sc, Si_i^*) and on the sparger type. For pipe and dome aerators it is possible to calculate roughly with the numerical value $10^5\,Y \approx 4-5$ [99]. For the Kenics aerator this value is higher, $10^5\,Y \approx 6$, because a gas – liquid dispersion is produced, whose momentum protects it against constriction as a result of the liquid entrainment of the bubble flue (airlift pump!).

Two-Component Nozzles. There are two types of two-component nozzles: ejectors and injectors. The *ejectors* are liquid jet aspirators, which operate according to the Venturi principle and with their propulsion liquid jet in the mixing space (diffuser) attain a substantial pressure drop on the gas side. Vacuum is commonly produced by water ejectors (vacuum pumps) in the laboratory and by steam ejectors on an industrial scale.

In an *injector* practically no pressure drop is produced, the power of the propulsion liquid jet

being utilized in the mixing chamber for producing an energy-rich gas – liquid dispersion. The free jet of the gas – liquid dispersion requires this energy to mix into the surrounding liquid, before a strong bubble coalescence takes place due to the high gas bubble density.

In contrast to spargers, the injectors possess a second freely selectable parameter in addition to the gas throughput q_G, the liquid throughput q_L, which would bring about the process characteristic q_L/q_G. In analogy to stirring, instead of q_L/q_G a dimensionless number was proposed [99], [101] which took the power of the propulsion jet into consideration: P_L/q_G. This number assumes knowledge of the pressure drop characteristics of the propulsion jet nozzle $Eu_L \equiv \frac{\Delta p_L\, d^4}{\varrho_L\, q_L^2}$ and is named the gas dispersion number, X:

$$X \equiv (P_L/q_G)^* \equiv \frac{P_L/q_G}{\varrho_L\,(\nu g)^{2/3}}, \text{ where } P_L = \Delta p_L\, q_L. \tag{44}$$

For injectors it therefore follows:

$$Y = f\,\{(P_L/q_G)*, Sc, Si_i^*, \text{type of injector}\}$$

Only two injectors of Bayer AG (Leverkusen, Germany) have been investigated in this framework (see Fig. 21). The larger of them, the so-called "slot injector" (see Fig. 21, right side) forms the basis of Bayer's tower biology for the aerobic wastewater treatment of chemically polluted effluents.

In Figure 22 the sorption characteristics of the slot injector are shown as a function of the degree of coalescence of the medium (water). This is varied via the addition of common salt.

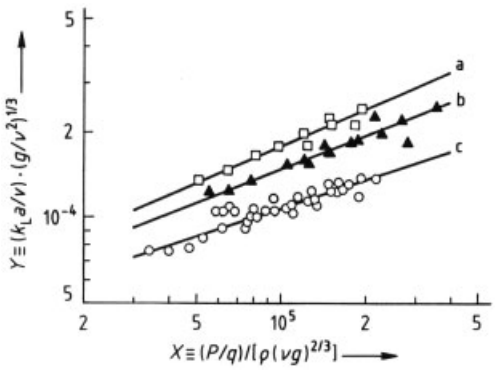

Figure 22. Dependence of sorption characteristics $Y = f\,(X)$ of the slot injector upon the system proneness to coalescence (NaCl concentration) [99]

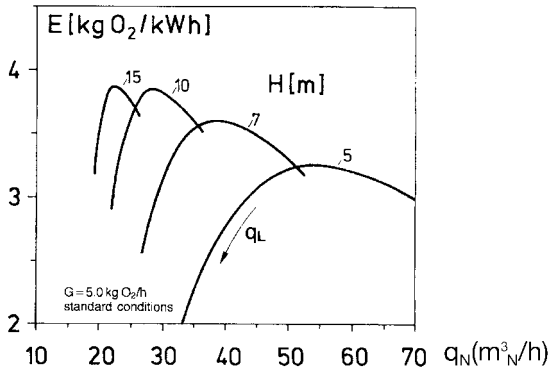

Figure 23. Efficiency E (kg O_2/kWh) of the oxygen uptake with a slot injector [99]
Boundary conditions: $G/V = 0.1$ kg O_2 m^{-3} h^{-1}, $T = 10\,°C$, $c\ (O_2) = 0$ mL/m^3; efficiencies of the air compressor $\eta_G = 0.60$ and the water pump $\eta_L = 0.75$ are taken into consideration

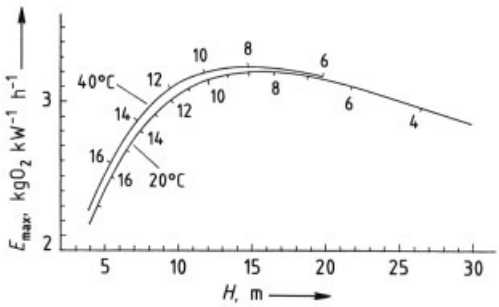

Figure 24. Maximum attainable efficiency E_{max} for slot injector
Operation conditions: $G/V = 0.1$ kg O_2 m^{-3} h^{-1}, $T = 20$ and 40 °C, $c\ (O_2) = 1$ mL/m^3. The scale on the curves gives the oxygen concentration in the off-gas in vol %

If the injectors are utilized for supplying activated sludge with atmospheric oxygen in a biological wastewater purification plant, it is of particularly interest that they realize the oxygen uptake G (kg O_2/h) with a high efficiency E (kg O_2/kWh). Since the injectors possess two degrees of freedom (q_G and q_L), an optimization procedure is used, which is shown in Figure 23 for the slot injector. The fact that the E value increases with increasing liquid height H, is caused by a longer residence time, during which the gas bubbles release oxygen to the liquid, thereby increasing the oxygen utilization. Thus for a given oxygen uptake G a smaller gas throughput and hence a lower compression power P_G is necessary.

Since the sorption characteristic $Y = f(P_L/q_G)*$ and the pressure drop characteristics Eu_L and Eu_G are known as a function of the throughput ratio q_L/q_G, an infinite number of value pairs of $\{q_L, q_G\}$ can be found for each required oxygen uptake G (e.g., 1 kg O_2/h) under given boundary conditions (liquid height H, concentration of dissolved oxygen to be maintained, e.g., 1 mL/m^3), which fulfill these requirements. However, only one of them satisfies this requirement with a minimum total power ($\sum P = P_G + P_L$) and hence maximum efficiency E. This is illustrated in Figure 24.

The plot of $E_{max}(H)$ in Figure 24 shows how the efficiency and the oxygen utilization behave with liquid height. The fact that 40 °C is more favorable than 20 °C, can be explained by the increase in Sc and hence k_L in this temperature interval not yet being nullified by the simultaneous decrease of Δc, due to $c_s(T)$.

7. References

1. M. Zlokarnik: *Rührtechnik – Theorie und Praxis,* Springer-Verlag, 1999; ISBN 3-540-64639-6. M. Zlokarnik: *Stirring – Theory and Practice,* Wiley-VCH, Weinheim 2001.
2. A. Biedermann, H.-J. Henzler, *Chem.-Ing. Tech.* **66** (1994) no. 2, 209 – 211.
3. M. Zlokarnik: *Scale-up – Modellübertragung in der Verfahrenstechnik,* Wiley-VCH, Weinheim 2000. M. Zlokarnik: *Scale-up in Chemical Engineering,* Wiley-VCH, Weinheim 2002.
4. P. V. Danckwerts: "Measurement of molecular homogeneity in a mixture", *Chem. Eng. Sci.* **7** (1957) 116 – 117.
5. J. W. Hiby: "Homogenisation", *Fortschr. Verfahrenstec. B* **17** (1979) 137 – 155.
6. M. Zlokarnik: "Eignung von Rührern zum Homogenisieren von Flüssigkeitsgemischen", *Chem. Ing. Tech.* **39** (1967) nos. 9/10, 539 – 548.
7. C. J. Hoogendoorn, A. P. den Hartog: "Model studies on mixers in the viscous flow region", *Chem. Eng. Sci.* **22** (1967) 1689 – 1699.
8. H. Gramlich, S. Lamadé: "Vergleich emaillierter Rührsysteme zum Homogenisieren von Flüssigkeitsgemischen", *Chem. Ing. Tech.* **45** (1973) no. 3, 116 – 122.
9. H.-J. Henzler: "Untersuchungen zum Homogenisieren von Flüssigkeiten oder Gasen", *VDI-Forschungsh.* **587** (1978).
10. M. Zlokarnik: "Einfluß der Dichte- und Zähigkeitsunterschiede auf die Mischzeit beim Homogenisieren von Flüssigkeitsgemischen", *Chem. Ing. Tech.* **42** (1970) no. 15, 1009 – 1011.
11. M. Opara, *Verfahrenstechnik* **9** (1975) no. 9, 446 – 449.
12. D. E. Ford, J. Ulbrecht, *AIChE J.* **21** (1975) no. 6, 1230 – 1233.
13. W. Müller, VDI-Forschungsberichte, Reihe 3 (1985), Nr. 103: "Untersuchung von Homogenisiervorgängen in nicht-Newtonschen Flüssigkeiten mit einem neuen bildanalytischen Verfahren"
14. R. Pörtner, U. Werner, *Chem.-Ing.-Tech.* **61** (1989) no. 3, 250 – 251.
15. W. Büche: "Leistungsbedarf von Rührwerken", *VDI-Z.* **81** (1937) no. 37, 1065 – 1069.
16. J. H. Rushton, E. W. Costich, H. J. Everett,: "Power characteristics of mixing impellers", *Chem. Eng. Prog.* **46** (1950) no. 8, 395 – 404, and no. 9, 467 – 476.
17. S. Nagata, K. Yamamoto, T. Yokoyama: "Studies on the power requirement of mixing impellers", *Mem. Fac. Eng. Kyoto Univ.* **19** (1957) no. 3, 274 – 290.
18. R. L. Bates, P. L. Fondy, R. R. Corpstein: "An examination of some geometric parameters of impeller power", *Ind. Eng. Chem. Process Des. Dev.* **2** (1963) no. 4, 310 – 314.
19. A. Mersmann et al.: "Auslegung und Maßstabsvergrößerung von Rührapparaten", *Chem. Ing. Tech.* **47** (1975) no. 23, 953 – 964.
20. J. Y. Oldshue, H. E. Hirschland, A. T. Gretton: "Blending of low-viscosity liquids with side-entering mixers", *Chem. Eng. Prog.* **52** (1956) no. 11, 481 – 484.
21. J. G. von de Vusse: "Vergleichende Rührversuche zum Mischen löslicher Flüssigkeiten in einem 12 000 m^3-Behälter", *Chem. Ing. Tech.* **31** (1959) no. 9, 583 – 587.
22. J. A. Wesselingh: "Mixing of liquids in cylindrical storage tanks with side-entering propellers", *Chem. Eng. Sci.* **30** (1975) 973 – 981.
23. H. Blasiński, J. Nowicki, E. Rzyski: "Mixing power and mixing times of propeller agitators introduced laterally", *Int. Chem. Eng.* **20** (1980) no. 1, 92 – 97.
24. H. Fosset, L. E. Prosser: "The application of free jets to the mixing of fluids in bulk", *J.*

Instn. Mech. Engr. London **160** (1949) 224 – 232.
25. L. S. Harris: "Jet eductor mixers", *Chem. Eng. N.Y.* **73** (1966) no. 21, 216 – 222.
26. R. G. Folsom, C. K. Ferguson: "Jet mixing of two liquids", *Trans. ASME* **71** (1949) 73 – 77.
27. M. Zlokarnik: "Homogenisieren von Flüssigkeiten durch aufsteigende Gasblasen", *Chem. Ing. Tech.* **40** (1968) no. 15, 765 – 768.
28. J. M. Smith, L. H. J. Goosens: "The mixing of ponds with bubble columns", *Proc. 4th Eur. Conf. Mixing,* NL-Leeuvenhorst 1982, 71 – 80.
29. J. W. Hiby: "Mischung in turbulenter Rohrströmung", *Verfahrenstechnik* (Mainz) 4 (1970) no. 12, 538 – 543.
30. O. Bolzern, J. R. Bourne: "Rapid chemical reactions in a centrifugal pump", *Chem. Eng. Res. Des.* **63** (1985) no. 5, 275 – 282.
31. H.-J. Henzler, "Eignung von kontinuierlich durchströmten Mischern zum Homogenisieren", *Chem. Ing. Tech.* **51** (1979) no. 1, 1 – 8.
32. F. Streiff: "Statische Mischer mit großer Anpassungsfähigkeit", *Chem. Ing. Tech.* **52** (1980) no. 6, 520 – 522.
33. M. H. Pahl, E. Muschelknautz: "Einsatz und Auslegung statischer Mischer", *Chem. Ing. Tech.* **51** (1979) no. 5, 347 – 364.
34. M. H. Pahl, E. Muschelknautz: "Statische Mischer und ihre Anwendung", *Chem. Ing. Tech.* **52** (1980) no. 4, 285 – 291.
35. I. Fort: "Pumping capacity of propeller mixer", *Collect. Czech. Chem. Commun.* **32** (1967) 3663 – 3678; **34** (1969) 1094 – 1097.
36. I. Fort, J. Podivinska, R. Baloun: "The study of convective flow in a system with a rotary mixer and baffles", *Collect. Czech. Chem. Commun.* 34 (1969) 959 – 974.
37. R. G. Cooper, D. Wolf: "Velocity profiles and pumping capacities for turbine type impellers", *Can. J. Chem. Eng.* **46** (1968) 94 – 100.
38. Y. Sano, H. Usui: "Interrelations among mixing time, power number and discharge flow rate number in baffled mixing vessels", *J. Chem. Eng. Japan* **18** (1985) no. 1, 47 – 52.
39. M. Zlokarnik: "Trombentiefe beim Rühren in unbewehrten Behältern", *Chem. Ing. Tech.* **43** (1971) no. 18, 1028 – 1030.
40. A. Le Lan, H. Angelino: "Etude du vortex dans les cuves agitées", *Chem. Eng. Sci.* **27** (1972) 1969 – 1978.
41. F. Rieger, P. Ditl, V. Novák: "Vortex depth in mixed unbaffled vessels", *Chem. Eng. Sci.* **34** (1979) 397 – 403.
42. R. Poggemann, A. Steiff, P.-M. Weinspach: "Heat Transfer in agitated vessels with single-phase liquids", *Ger. Chem. Eng.* (Engl. Transl.) 3 (1980) 163 – 174.
43. M. F. Edwards, W. L. Wilkinson: "Heat transfer in agitated vessels, Part I: Newtonian fluids", *Chem. Engr.* (London) (1972) no. 264, 310 – 319.
44. Y. Sano, N. Yamaguchi, T. Adachi: "Mass transfer coefficients for suspended particles in agitated vessels and bubble columns", *J. Chem. Eng. Japan* **7** (1974) no. 4, 255 – 261.
45. P. Sänger, W.-D. Deckwer: "Liquid-solid mass transfer in aerated suspensions", *Chem. Eng. J. (Lausanne)* **22** (1981) no. 3, 179 – 186.
46. C. A. Coulaloglou, L. L. Tavlarides: "Drop size distributions and coalescence frequencies of liquid – liquid dispersions in flow vessels", *AIChE J.* **22** (1967) no. 2, 289 – 297.
47. D. M. Levins, J. R. Glastonbury: "Particle-liquid hydrodynamics and mass transfer in a stirred vessel", *Trans. Inst. Chem. Eng.* **20** (1972) 132 – 146.
48. A. H. P. Skelland, R. Seksaria: "Minimum impeller speeds for liquid – liquid dispersion in baffled vessels", *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) no. 1, 56 – 61.
49. M. Zlokarnik, H. Judat: "Rohr- und Propellerrührer–eine wirkungsvolle Rührkombination zum gleichzeitigen Begasen und Aufwirbeln", *Chem. Ing. Tech.* **41** (1969) no. 23, 1270 – 1273.
50. M. Zlokarnik, H. Judat: "Rohr- und Scheibenrührer–zwei leistungsfähige Rührer zur Flüssigkeitsbegasung". *Chem. Ing. Tech.* **39** (1967) no. 20, 1163 – 1168.
51. K. Ishibashi, A. Yamanaka, N. Mitshuishi, *J. Chem. Eng. Japan* **12** (1979) no. 3, 230 – 234.
52. M. Kuriyama et al.: "Heat transfer and temperature distributions in an agitated tank equipped with helical ribbon impeller", *J. Chem. Eng. Japan* **14** (1981) no. 4, 323 – 330.
53. M. Zlokarnik: "Wärmeübergang an der Wand eines Rührbehälters beim Kühlen und Heizen im Bereich $10^{\circ} < \mathrm{Re} < 10^{5}$", *Chem. Ing. Tech.* **41** (1969) no. 22, 1195 – 1202.
54. V. W. Uhl, H. P. Voznick: "The anchor agitator", *Chem. Eng. Prog.* **56** (1960) no. 3, 72 – 77.
55. W. R. Penney, K. J. Bell, *Chem. Engng. Progr. Sympos. Ser.* **65** (1969) no. 92, 1 – 11.
56. M. Baccar, M. S. Abid, *Rev. gen. de Therm., Paris* **36** (1997) no. 10, 782 – 790.

57. S. Nagata, *Mixing – Principles and Application,* Kodansha Ltd. Tokyo & John Willey, New York 1975.
58. H. Judat, U. Johenneken, Jahrestreffen der Verfahrensingenieure, Stuttgart 1991: Wärmeübergang in Rührkesseln mit wandschabenden Rührelementen.
59. R. Rautenbach, F. M. Bollenrath, *Chem.-Ing.-Tech.* **50** (1978) 4.
60. J. Pawlowksi, M. Zlokarnik: "Optimieren von Rührern für eine maximale Abfuhr von Reaktionswärme", *Chem. Ing. Tech.* **44** (1972) no. 16, 982 – 986.
61. W. R. Penney, R. N. Koopman: "Prediction of net heat removal capabilities for agitated vessels", *AIChE Symp. Ser.* **68** (1972) no. 118, 62 – 73.
62. W.-D. Einenkel, A. Mersmann, *Verfahrenstechnik* **11** (1977) no. 2, 90 – 94.
63. F. Kneule, P. M. Weinspach: "Suspendieren von Feststoffpartikeln im Rührgefäß", *Verfahrenstechnik* (Mainz) 1 (1967) no. 12, 531 – 540.
64. S. Lamadé, *Verfahrenstechnik* **11** (1977) no. 2, 72 – 81.
65. W.-D. Einenkel, *Fortsch. d. Verfahrenstechnik* **16** (1978) 113 – 126.
66. F. Kneule, P.-M. Weinspach, *Chem.-Ing.-Tech.* **35** (1963) no. 10, 720 – 725.
67. H. Hraste, M. Husnjak, *Chem. Biochem. Eng. Q* **9** (1995) no. 2, 105.
68. F. Rieger, F. Sinevič, *Collect. Czech. Chem. Comm.* **60** (1995) 966 – 968.
69. B. Ayyappan, P. Kalaichelvi, T. Murugesan, *Bioprocess Eng.* **15** (1996) 205 – 208.
70. P. M. Armenante, E. U. Nagamine, *Chem. Eng,. Sci.; Oxford, Frankfurt* **53** (1998) no. 9, 1757 – 1775.
71. M. Kraume, P. Zehner, *Chem.-Ing.-Tech.* **67** (1995) no. 3, 280 – 288.
72. W.-D. Einenkel, *Ger. Chem. Eng.* **3** (1980) 118 – 124.
73. W.-D. Einenkel, *Chem.-Ing.-Tech.* **67** (1995) no. 8, 1000 – 1003.
74. G. Herndl, *Fortschr. d. Verfahrenstech.* **20** (1982) 137 – 162.
75. M. W. Chudacek, *Chem. Engng.* **Oct. 1** (1984).
76. M. W. Chudacek, *Chem. Eng. Sci.* **40** (1985) no. 3, 385 – 392,.
77. C. Buurman, G. Resoort, A. Plaschkes, *Chem. Eng. Sci.* **41** (1986) no. 11, 2865 – 2871.
78. R. V. Roman, R. Z. Todose, *Bioprocess Engng.* **15** (1996) 221 – 229.
79. A. W. Hixson, S. J. Baum: "Mass transfer coefficients in liquid – solid agitation systems", *Ind. Eng. Chem.* **33** (1941) no. 4, 478 – 485.
80. J. Sauter: "Die Größenbestimmung von Brennstoffteilchen",*Forschungsarbeiten,* **279** (1926).
81. P. H. Calderbank, M. B. Moo-Young: "The continuous phase heat and mass-transfer properties of dispersions", *Chem. Eng. Sci.* **16** (1961) no. 2, 39 – 54.
82. H. D. Schindler, R. E. Treybal, *AIChE J.* **14** (1968) no. 5, 790 – 798.
83. R. B. Keey, J. B. Glen, *AIChE J.* **15** (1969) no. 6, 942 – 947.
84. W. K. Lewis, W. G. Whitman, *Ind. Engng. Chem.* **16** (1924) 1215 – 1220 and W. G. Whitman, *Chem. Metall. Eng.* **29** (1923) 147.
85. K. Van't Riet: "Review of measuring methods and results in nonviscous gas – liquid mass transfer in stirred tanks", *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) no. 3, 357.
86. H. Judat: "Gas/liquid mass transfer in stirred vessels–a critical review", *Ger. Chem. Eng. (Engl. Transl.)* **5** (1982) 357 – 363.
87. M. Zlokarnik: "Auslegung von Hohlrührern zur Flüssigkeitsbegasung. Teil I: Bestimmung des Gasdurchsatzes und der Wellenleistung", *Chem. Ing. Tech,* **38** (1966) no. 3, 357 – 366.
88. M. Zlokarnik: "Auslegung von Hohlrührern zur Flüssigkeitsbegasung. Teil II: Ermittlung des erreichbaren Stoff- und Wärmeaustausches", *Chem. Ing. Tech.* **38** (1966) no. 7, 717 – 723.
89. M. Zlokarnik, *Adv. Biochem. Eng.* **8** (1978) 133 – 151.
90. C. W. Robinson, C. R. Wilke, *Biotechn. and Bioengng.* **15** (1973) 755 – 782.
91. H.-J. Henzler: "Verfahrenstechnische Auslegungsunterlagen für Rührbehälter als Fermenter", *Chem. Ing. Tech.* **54** (1982) no. 5, 461 – 476.
92. W. A. P. Luck, *Topics in Currrent Chemistry* **64** (1976) 113 – 180.
93. M. Zlokarnik: "Rührleistung in begasten Flüssigkeiten", *Chem. Ing. Tech.* **45** (1973) no. 10 a, 689 – 692.
94. H. Judat, *Chem.-Ing.-Tech.* **48** (1976) no. 3, 228 – 229.
95. H. Höcker, Dissertation Universität Dortmund 1978: Untersuchungen zum Leistungsbedarf und Stoffübergang in Rührreaktoren.
96. H.-J. Henzler, *Chem.-Ing.-Tech.* **52** (1980) no. 8, 643 – 652.
97. M. Zlokarnik: "Scale-up of surface aerators for waste water treatment", *Adv. Biochem. Eng.* **11** (1979) 157 – 180.

98. M. Zlokarnik: "Eignung und Leistungsfähigkeit von Oberflächenbelüftern für biologische Abwasserreinigungsanlagen", *Korrespondenz Abwasser* **27** (1980) no. 1, 14 – 21.
99. M. Zlokarnik: "Eignung und Leistungsfähigkeit von Volumenbelüftern für biologische Abwasserreinigungsanlagen", *Korrespondenz Abwasser* **27** (1980) no. 3, 194 – 209.
100. M. Zlokarnik: "Sorption characteristics of slot injectors and their dependency on the coalescence behaviour of the system", *Chem. Eng. Sci.* **34** (1979) no. 10, 1265 – 1271.
101. A. Steiff, P.-M. Weinspach: "Heat transfer in stirred and non-stirred gas – liquid reactors", *Ger. Chem. Eng. (Engl. Transl.)* **1** (1978) 150 – 161.
102. Th. N. Zwietering: "Suspending of solid particles in liquids by agitation", *Chem. Eng. Sci.* **8** (1958) 244 – 253.
103. W. D. Einenkel: "Fluiddynamik des Suspendierens", *Ger. Chem. Eng.* **3** (1980) 118 – 124.
104. P. M. Weinspach, *Chem.-Ing.-Tech.* **41** (1969) 5 + 6, 260 – 265.
105. J. Venus, K.-H. Wolf, *Chem.-Ing.-Tech.* **63** (1991) 2, 168 – 169.

Stoneware → Whitewares
Stoneware → Construction Ceramics

Mixing of Highly Viscous Media

See also → Mixing, Introduction

David B. Todd, APV Chemical Machinery Inc., South Plainfield, NJ 07080, United States

1. Introduction 1
2. Batch Mixers 3
3. Continuous Mixers 9
3.1. Motionless Mixers 10
3.2. Single Screw Mixers 11
3.3. Twin Screw Mixers 12
4. References 16

1. Introduction

Thick mixtures with viscosities >10 Pa s do not lend themselves to simple mixing with a small turbine or propeller stirrer in a large pot. The high viscosity may arise from a high concentration of solids in a slurry, the high viscosity of the matrix fluid itself, or by interactions between the ingredients.

With high viscosity fluids, the mixing Reynolds number ($Re = D^2N/\mu$) is likely less than 100. Thus, mixing occurs by viscous forces alone, and turbulence plays no part. The viscous forces cause either shear flow or elongational flow. Relative motion of an agitator stretches and deforms the material between itself and the vessel wall. As a layer of fluid gets stretched into thinner layers, and then undergoes reorientation, the striations become ever thinner until homogeneity is achieved. Likewise, the shear forces tear away at solid agglomerates and cause their disintegration.

Many mixers have been independently developed for specific process applications, and then their usage has been extended to other processes involving somewhat similar, but difficult mixing tasks. The agitated tubs used for kneading bread dough evolved into horizontal vessels that could more readily be tilted for emptying. The design evolved into double arm mixers, which in turn led to a variety of mixing blades tailored for other end uses as these double arm mixers began to be used for adhesives, gums, putties, rubber solutions, pigment dispersions, and a host of other products.

The very difficulty in defining precisely the consistency of these various doughs, pastes, and thick slurries has fostered a wide art of mixer design, rather than a sound scientific approach. Consequently, mixers are generally classified more by their mechanical features than by the process function for which they may best be suited.

With solids present, producing a homogeneous distribution is frequently only part of the mixing task. It may also be necessary to reduce the particle size of the solids, even down to submicrometer size, to achieve the desired end product, such as pigment dispersion. However, not all products need such fineness; for example, the fiber glass that is introduced into molten plastic should be uniformly distributed with minimum breakup.

High viscosity mixers usually have a limited high shear zone to minimize overall heat buildup, and depend upon the impeller also circulating all of the mixer contents past the high shear zone. Stagnant zones can exist, material may ride on blades and be kept out of the mixing region, and pockets of unmixed components may be carried through continuous process equipment since there is a sharp drop off of transmitted shear away from the shear creating device. High velocity impellers may be ineffective since they may create an isolated hole in the mass without producing any circulation, particularly with pseudoplastic materials.

When mixing a minor component (of either a solid or another viscous fluid) into a continuous phase viscous fluid, two kinds of mixing action may be required. The first is distributive mixing where interfaces are maintained intact, but the minor component is distributed uniformly throughout the mass. The second is dispersive mixing where the interfaces are broken up. With two-phase fluid systems, the differences bet-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_26

ween distributive and dispersive mixing has been described by OTTINO and coworkers [1–3].

ERWIN [4], [5] has measured distributive mixing in single screw extruders. Dispersive mixing has been evaluated for a variety of flow conditions by RUMSCHEIDT and MASON [6], [7], GRACE [8], and HAN [9]. ERWIN [10] concluded that extensional flow is the most efficient of plain strain laminar flows.

The degree of distributive mixing for a two-phase mixture can be characterized by the relative increase in interface area or number of striations per unit length. ERWIN and coworkers [2], [3] have shown that mixing increases linearly with the average total shear strain applied. They have also shown that mixing efficiency in single screw extruders was increased significantly by reorientation of the interface, such as with mixing pins. Also, poorer mixing occurs in laminar shear flows when the ratio of the viscosity of the disperse phase to the continuous phase increases.

For a given shear field, there is a maximum viscosity ratio beyond which a liquid droplet cannot be broken up. KARAM and BELLINGER [11] have shown that breakup is easier when the interfacial tension (σ) is low, the viscosity of the continuous phase (μ_c) is high, and the radius (r) of the droplet is large. Their plot of reduced shear rate vs. viscosity ratio ($\mu_r = \mu_d/\mu_c$) is reproduced here as Figure 1. Thus, when extremes in viscosity ratio are encountered, real fine dispersions cannot be achieved in simple shear fields alone.

For viscous fluid – fluid mixtures, the efficiency of a mixer in theory can be determined from a combination of three factors: critical capillary number at which breakup occurs, the time required for this breakup, and the number and size of the resulting drops. The capillary number is the ratio of the viscous forces to the interfacial forces. The critical capillary number in shear flow goes through a minimum at a viscosity ratio of 1, and above 4, simple shear flow is no longer effective.

With a hyperbolic flow, the critical capillary number is lowered. The time for breakup increases as the viscosity ratio is increased. As the capillary number is increased above the critical value, the number of fragments produced increases sharply at first, but then becomes constant. Thus, the relative dispersion efficiency of different mixing devices should be compared only under similar parameters of viscosity ratio, initial drop size, continuous phase viscosity, and interfacial tension. The number and size of the fragments produced versus time would be a reasonable measure of dispersion efficiency.

Equipment for viscous mixing is characterized by small clearances between impeller and vessel walls, high power per unit volume, and relatively small volume. Because of probable heat buildup, impeller speeds may have to be very slow. Intermeshing blades or stators may also be required to keep the material from cylindering on the rotating impeller.

A smearing blade profile may be required where dispersion is important, whereas a scraping profile may be desired if heat transfer is critical. A mixing action which causes material to work against itself can be used to advantage.

Ease of discharge, and ease of cleaning between batches are important requirements for paste mixers. Mere bottom drainage is frequently inadequate, in which case the vessel may need to be dumped.

Mixing in general has been treated in the two-volume edition *Mixing–Theory and Practice*, and in particular for viscous mixing in the chapter by IRVING and SAXTON [12]. Similar topics are also covered by PAHL [13]. The triennial European Conferences on Mixing [14], although concentrating more on multiphase dispersion and reaction in turbulent mileau, do also have some papers on viscous and non-Newtonian mixing phenomena.

Power. For viscous mixers with agitators that sweep near the vessel wall, the power drawn depends primarily on the viscous drag of the agitator rather than on the pumping effect incurred in circulating the material to the high shear zone. For scale up, the following approach may be taken:

Shear rate (DN/t)
Shear area (DL)
Shear stress = (shear rate) · (viscosity)

$$\alpha \left(\frac{DN}{t}\right) \mu$$

Force = (shear stress) · (shear area)

$$\alpha \left(\frac{\mu DN}{t}\right) (DL)$$

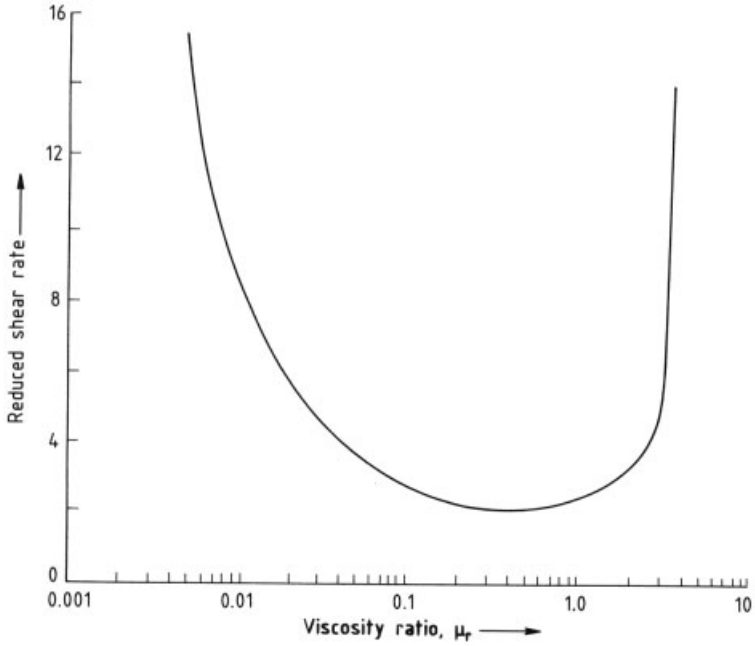

Figure 1. Effect of viscosity ratio on dispersion [11]
Reduced shear rate $\frac{\gamma_b \mu_c r}{\sigma}\left(\frac{1.19\mu_r+1}{\mu_r+1}\right)$ where
σ = interfacial tension; μ_c = viscosity, continuous phase; μ_d = viscosity, dispersion phase; $\mu_r = \mu_d/\mu_c$; r = drop radius; γ_b = shear rate at breakup

$$\text{Torque} = (\text{force}) \cdot (\text{radius})$$

$$\alpha \left(\frac{\mu D^2 L N}{t}\right) D$$

$$\text{Power} = (\text{torque}) \cdot (\text{speed})$$

$$\alpha \frac{\mu D^3 L N^2}{t}$$

for dimensional similarity, with both clearance (t) and agitator length (L) proportional to diameter D:

$$\text{Power} = \alpha \mu D^3 N^2$$

Actually, for most very viscous mixtures, the viscosity μ is not constant, and is usually shear thinning, with an effective viscosity depending upon the $(n-1)$ power of speed, where n is the power index number of the fluid (the slope of shear stress vs. shear rate on logarithmic coordinates).

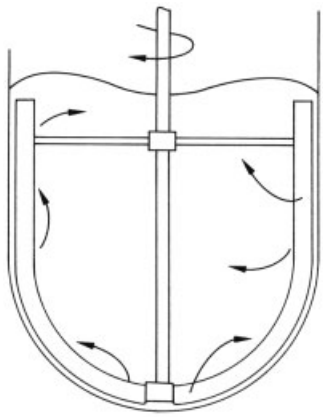

Figure 2. Anchor mixer

2. Batch Mixers

Single Stirrer Mixers. Thick pastes can be handled in a variety of simple batch mixers in which the agitator is in close proximity to the vessel wall. With an anchor blade (Fig. 2), the proximity of the blade to the vessel wall improves heat transfer while providing gentle ag-

itation. Use of single or double helical blades (Fig. 3) provides not only the scraping action at the wall but also an end-to-end turnover of vessel contents. A single helical auger may be used inside of a draft tube with reliance on flow back up the annulus caused by the pumping action of the central auger. Some mixers also employ stators attached to the vessel wall that are in close proximity to the multi-arm central agitator, such as shown in Figure 4.

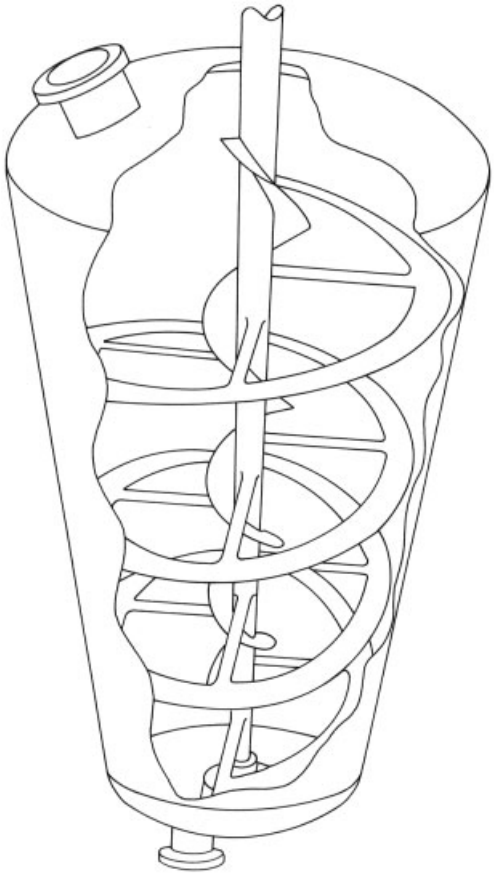

Figure 3. Helical blade mixer

Change Can Mixers. Change can mixers are vertical batch mixers in which the mixing vessel can readily be separated from the agitator. Agitation is usually provided by two planetary mixing blades (Fig. 5). Separate cans permit accurate weighing of the ingredients prior to mixing, easy transport of the mixture to the next process operation, and easier cleaning of the mixer between batches.

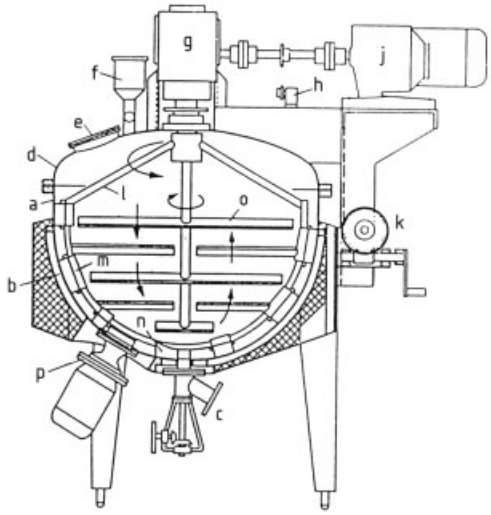

Figure 4. Multi-arm shearbar mixer
a) Body; b) Insulating heating mantel; c) Outlet; d) Cover; e) Observation port; f) Inlet; g) Counter-rotating drive; h) Tachometer; j) Motor; k) Cover mechanism; l) Outer stirrer; m) Wall scrapper; n) Discharge scrapper; o) Inner stirrer; p) Homogenization mill

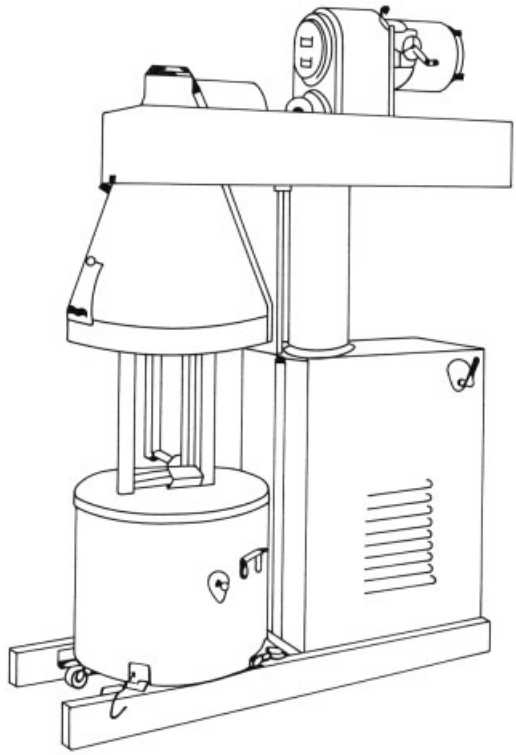

Figure 5. Change can mixer (Chas. Ross & Son)

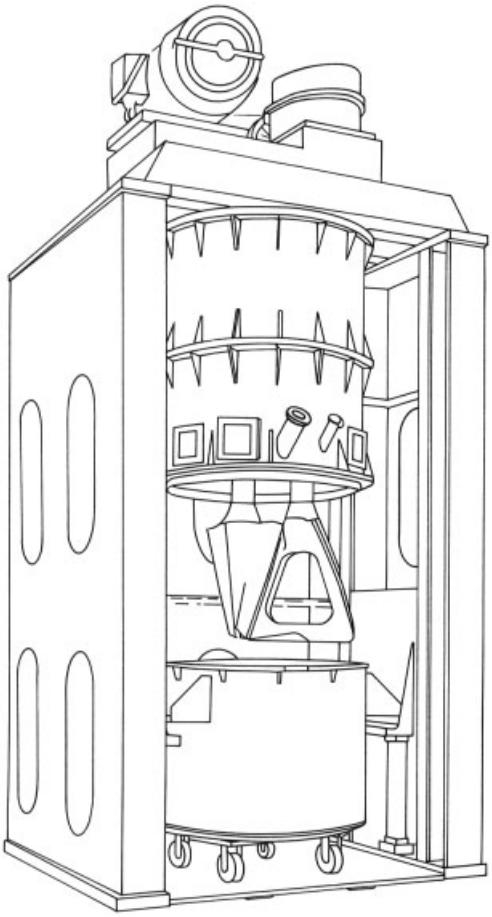

Figure 6. Vertical mixer (Baker Perkins)

With most smaller change can mixers, such as used in ink and pigment manufacture, the mixing head is raised from the can either by tilting or vertically. Change can mixers also have permanently mounted agitator elements with the can raised for mixing. The latter type (Fig. 6) is used extensively for mixing solid propellants.

A variety of blade shapes has evolved to provide for thorough mixing of the can contents. The blades are usually intermeshing and pitched to provide some axial movement. The can may be rotated on a turntable to bring all material past the mixing head, or the planetary gear mounting may rotate around a fixed can. The can may also tilt for dumping, as shown in Figure 7.

Change can mixers are available generally in the 1 – 500-L size, with the propellant mixers up to 4000 L.

Double Arm Kneading Mixers. These mixers consist of two tangential or overlapping horizontally mounted mixing shafts rotating in a rectangular trough with the bottom curved in two half cylinders to match the sweep of the mixing blades (Fig. 8). Several mixing blade shapes are available (Fig. 9). Most common is the Sigma Blade (Fig. 9A), with unequal arms angled and unequal rotating speeds to impart randomness to the mixing action. Material is swept arount the bottom and divided as it is brought down over the saddle ridge between the half cylinders.

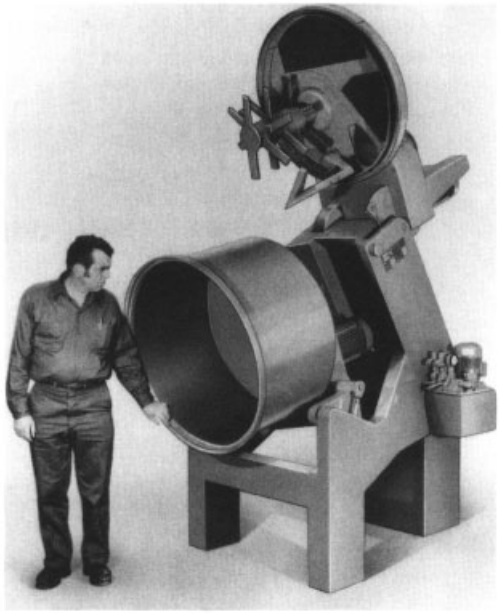

Figure 7. Intensive mixer with withdrawable agitator and tilting discharge (Eirich)

Dispersion of agglomerates is improved with blades that provide compressive shear forces against the trough shell with a wedged shape face (Fig. 9B) rather than a scraping blade profile. Heat transfer is promoted bvy using a cored blade and a scraping profile on the blade front (Sigma Blade, Fig. 9A). Shredded action can also be imparted (Fig. 9C), or cutting action (Fig. 9D).

If the mixed contents are free draining, or if a heel of the batch can be left behind without impairing the quality of succeeding batches, the mixer may be emptied through one or two gates or valves on the bottom. However, the usual pro-

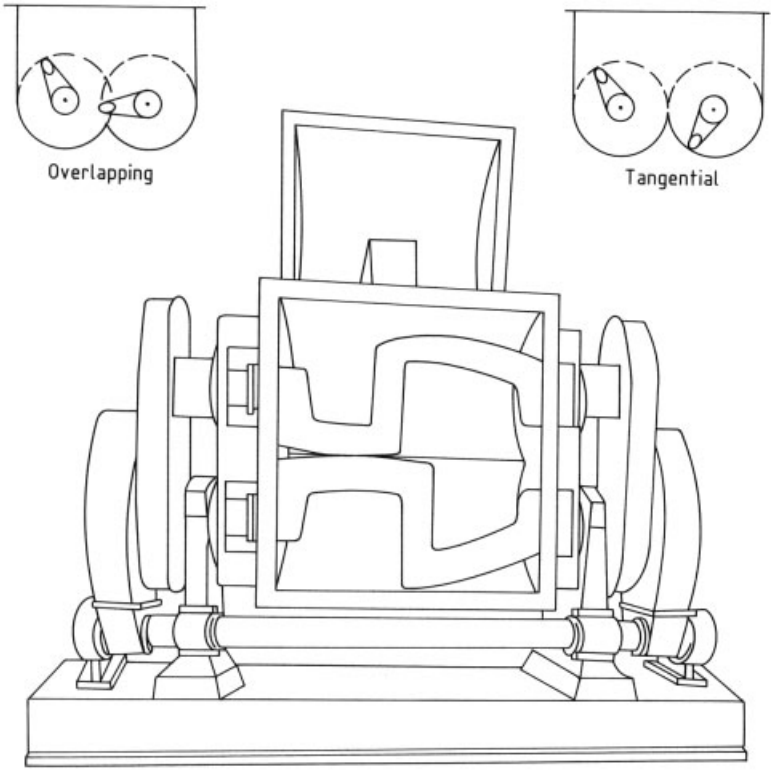

Figure 8. Double arm kneading mixer (Baker Perkins)

cedure is to tilt the mixer for emptying. The mixer can also be provided with a screw discharge mounted in the saddle section (Fig. 10). The screw is reversed during the mixing stage to assist mixing.

Double Arm mixers are available in size ranges from 1 L to 5 m^3. As size is increased, rotational speeds are generally reduced. Power inputs range from 0.02 – 0.5 kW/kg. A typical 500-L mixer might run with a front blade speed of 30 rpm and be supplied with a 40 kW drive motor.

Ram Mixers. The Banbury Type mixer (Fig. 11), is capable of extremely high power inputs, up to 6 kW/kg, in size ranges from small laboratory to 500 kg. A compression ram is used to confine the mixture between the tight fitting rotors and vessel walls. The masticating action is most useful for mixing elastomers and plastics. The mixed product is dumped through a front or bottom discharge door. The relatively large shafts permit very high torques. As with all very viscous mixing operations, batch temperature rise may be very rapid and may be the factor limiting cycle times. At 1 kW/kg, temperature rise per minute could readily reach 30°.

High Intensity Mixers (Fig. 12) utilize a high speed bottom mounted blade to impart high shear and to establish a vortex flow in the vessel. The bottom blade scoops the batch upward at peripheral speeds of 40 m/s. Blade impact is effevtive in reducing agglomerates and aids intimate dispersion. Energy input is high (up to 200 kW/m^3), and these high intensity mixers are frequently paired with a slower speed, larger batch cooler of similar design. These mixers are particularly suited for rapid mixing of powders and granules with minor amounts of liquids, such as when plasticizer is added to poly(vinyl chloride) (PVC) resin. The dry PVC granules are heated by the blade action within a few minutes, after which the plasticizer is added and rapidly penetrates the pores of the PVC particles to provide a homogeneous feed mixture for

downstream processing. Other uses are for removing volatiles from pastes under vacuum, and for dissolving solids in liquids. Sizes range from 1 – 1100 L.

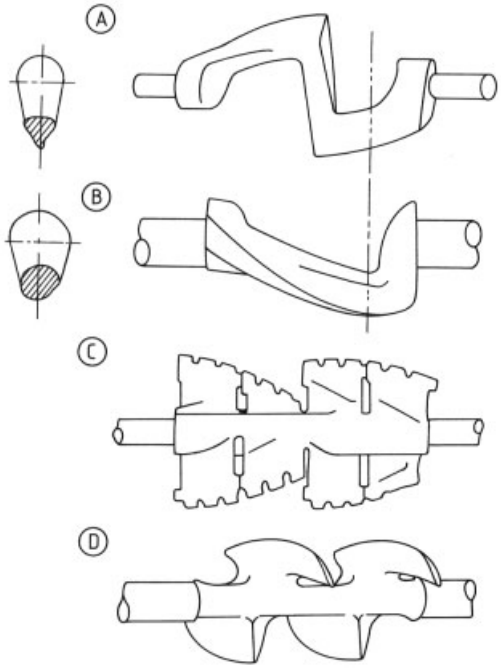

Figure 9. Agitator blades for double arm kneaders (Baker Perkins)
A) Sigma; B) Dispersion; C) Multiwing overlap; D) Double naben

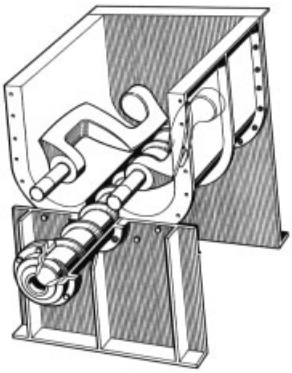

Figure 10. Screw discharge batch mixer (Werner & Pfleiderer)

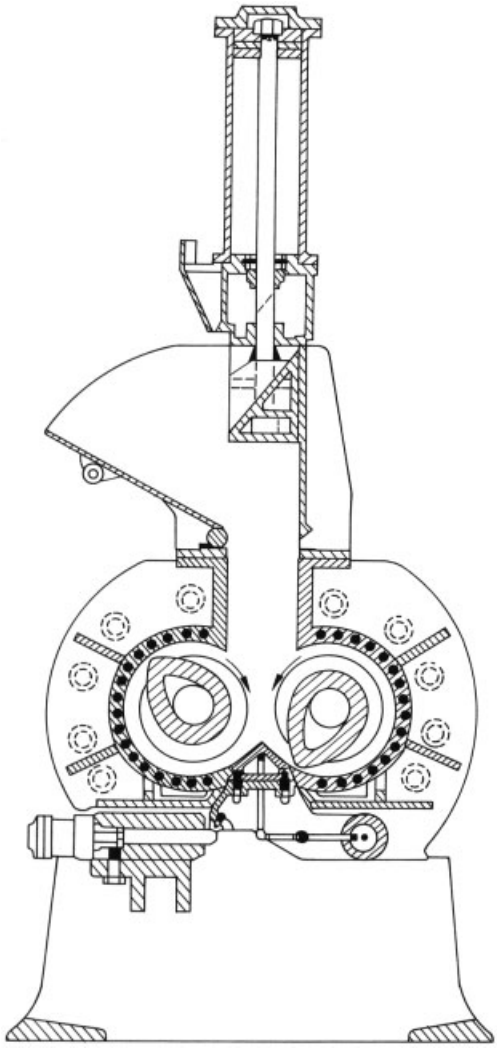

Figure 11. Banbury mixer (Farrel)

A somewhat similar mixers, shown in Figure 13, utilizes a horizontal impeller. Incorporated into the mixer is an optical pyrometer to indicate when the desired energy has been imparted by sensing the inside product temperature. Since a high level of energy can be supplied, the mix cycle time is quite short.

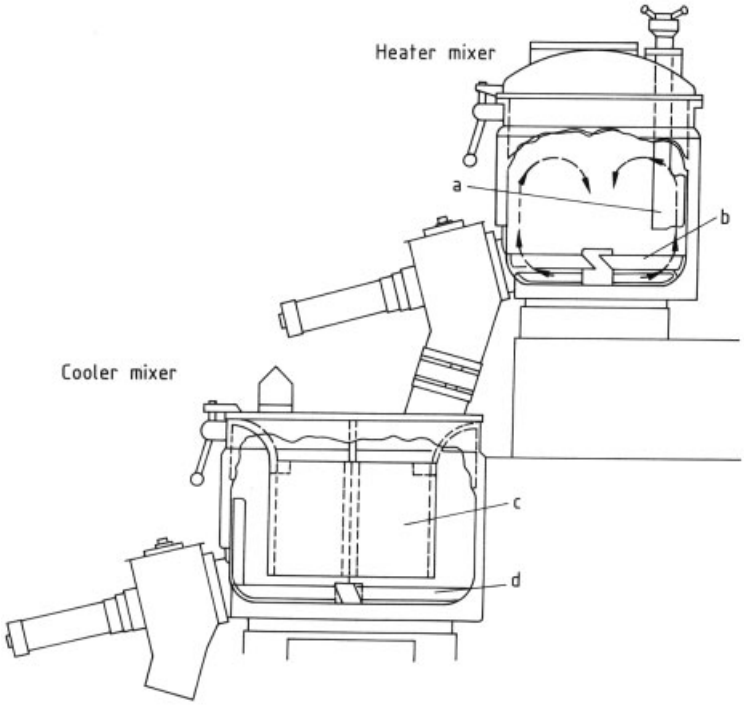

Figure 12. High density non-fluxing mixer (Baker Perkins)
a) Baffle, b) Impeller; c) Cooling ring; d) Impeller

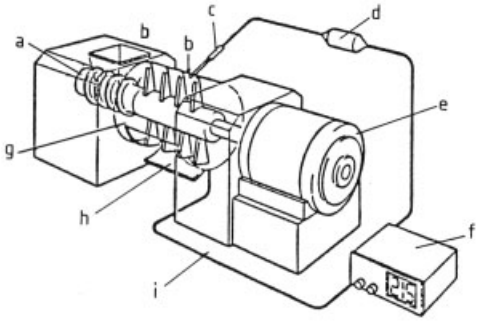

Figure 13. High intensity Gelimat fluxing mixer (Draiswerke)
a) Integrated screw feed; b) Sapphire lens; c) Optical fiber infrared sensor; d) Converter; e) Drive; f) Adjustable temperature control panel; g) Mixing chamber; h) Discharge door; i) Signal to discharge door

Plow Mixers have plow elements on a horizontal shaft rotating in a cylindrical vessel (Fig. 14). After charging, the smooth fitting cover is closed, and rapid shaft rotation hurls material into the free space. Addition blending is provided by the impellers plowing through the solids bed. If agglomerates are present, an additional small high speed chopper may be added.

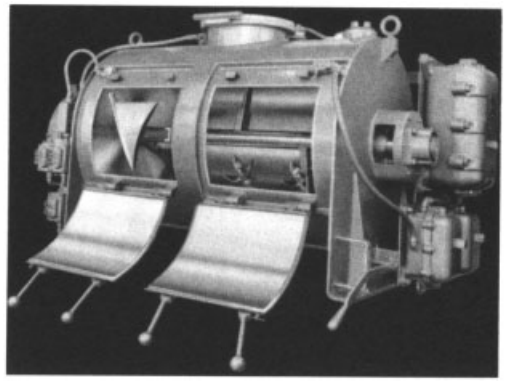

Figure 14. Plow mixer (Littleford Bros.)

Cone and Screw Mixers (Fig. 15) are used for a variety of low-energy blending applications ranging from dry powders to pasty materials. An orbiting screw rotates about its own axis and also around the center line of the cone shaped vessel

near the wall to provide bottom to top circulation. The screw is reversed to aid discharge of pasty materials. Screw and cone mixers, available in capacities from 40 to 20 m^3, work equally well when only partially filled.

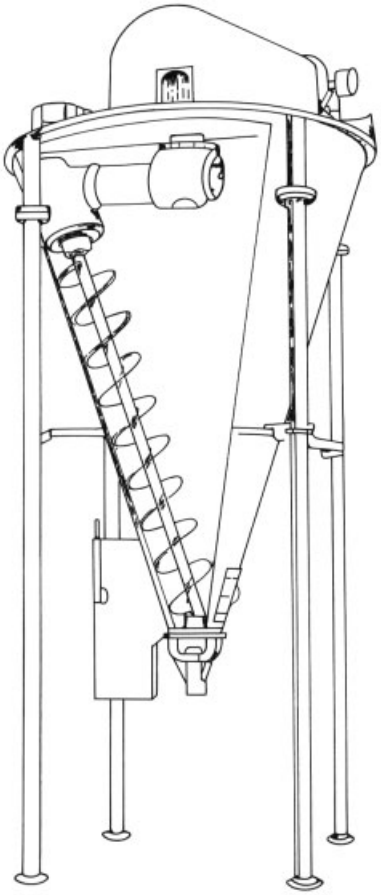

Figure 15. Cone and screw mixer (Day Mixing)

Mullers (Fig. 16) can be used for powders and pastes if the latter are not too soupy or too sticky. The mullers smash agglomerates as they rotate in a circular patch inside the round pan. An attached plow brings material progressively into the path of the mullers. Typical applications are for mixing putty and clay pastes, battery paste, and chocolate coatings. A continuous muller variant consists of two rotating sets of muller wheels operating with some overlap in a figure-eight pan.

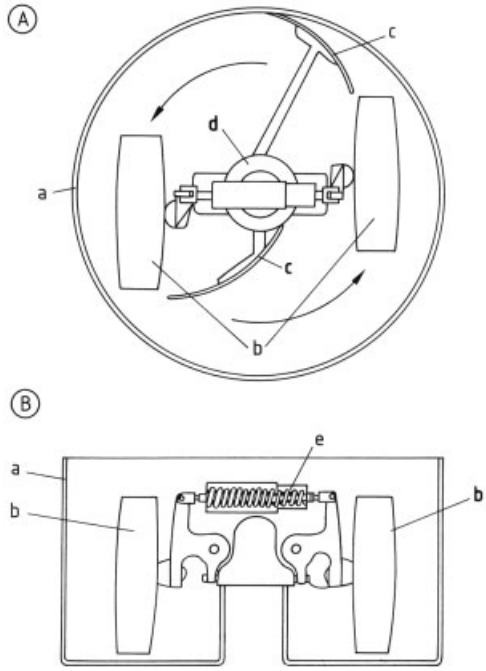

Figure 16. Muller (National Engineering)
a) Crib wall; b) Mullers; c) Plows (inner and outer); d) Stationary turret; e) Screw

Roll Mills provide a very high localized shear by drawing the mixture into the nip of closely spaced parallel rolls. The viscosity of the matrix can be held high to aid dispersion because of the extensive surface area for heat transfer. The material passing through the nip is returned to the feed point by rotation of the rolls. A temperature differential between the rolls frequently causes the material to selectively adhere to one of the rolls, as it does when blending rubber stocks. Although there is little distributive mixing as the material passes through the nip, there is a rotary motion of the bead of material immediately upstream of the nip. In addition to blending rubber stocks and dispersing carbon black, roll mills are used extensively in grinding and dispersing inks and pigment pastes.

3. Continuous Mixers

Although some of the batch mixers can be modified for continuous processing, the generally broad residence time is usually not conducive to

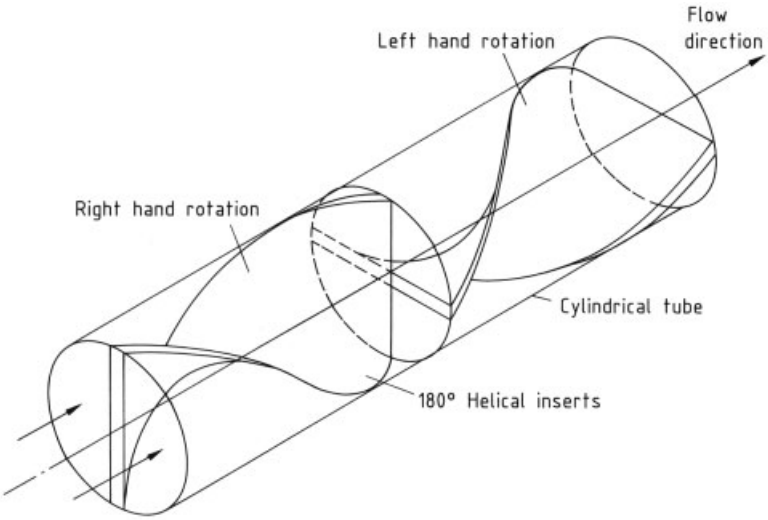

Figure 17. Kenics static mixing features

the best product uniformity. However where the component feed streams can be adequately and accurately metered, several mixers have been developed specifically for continuous processing. Continuous mixers generally consist of a closely fitted rotating element within a stationary housing with flow through the unit controlled to prevent by-passing.

3.1. Motionless Mixers

Longitudinal and transverse distribution of components can be obtained in flowing systems by placement of stationary flow redistribution devices, called motionless or static mixers, in the flow stream. Many proprietary designs have evolved, all of which divide and displace a laminar streamline from one portion of the conduit cross section to another. Examples of static mixers are shown in Figures 17, 18, 19. The Kenics mixer shown in Figure 17 consists of alternate right and left handed twisted ribbons arranged in sequence. The Ross ISG (Fig. 18) uses four drilled holes through solid sections between axially displaced cavities to redirect flow streamlines both radially and peripherally. The Sulzer SMV mixer (Fig. 19) has stacks of corrugated lamellae with open crossing channels. Most other static mixers involve variations of these three basic types with features that may permit easier fabrication or installation.

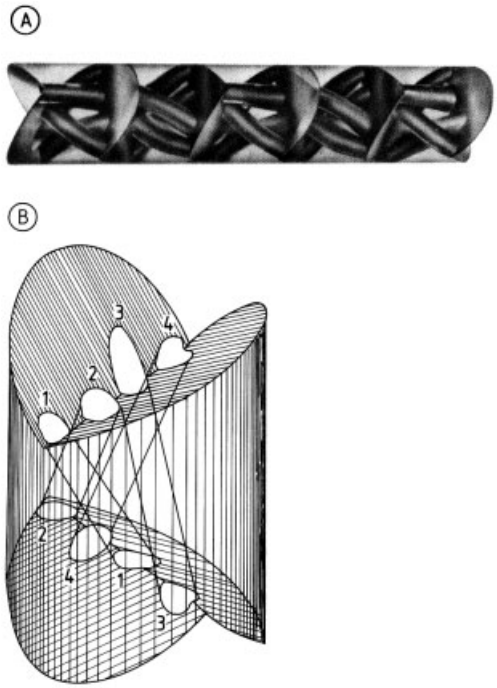

Figure 18. Ross ISG mixer
A) Assembly; B) Individual element

PAHL and MUSCHELKNAUTZ [15] have presented both mixing and pressure drop comparisons for these and other static mixers. In general, the more divisions, the better the homogenization. However, the smaller the passageways

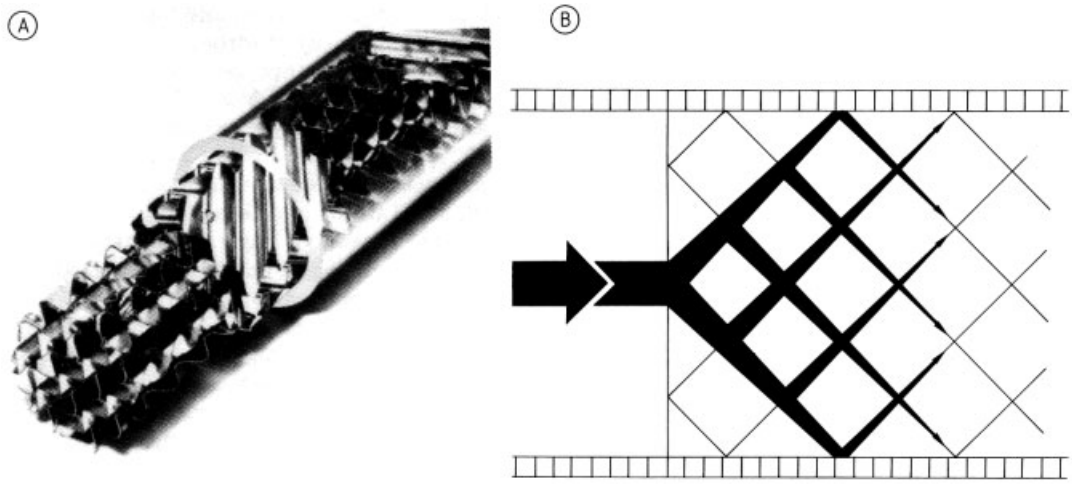

Figure 19. The Sulzer mixer
A) Sulzer mixing elements; B) Operating principle of the Sulzer mixing element

and the larger the fraction of conduit cross section occupied by metal, the higher the pressure drop. Choice of the best static mixer for a particular application involves a compromise based on energy and space available. Static mixers are of most value for distributive, rather than dispersive fluid mixing applications. With no moving parts, there are low investment, operating, and maintenance costs.

3.2. Single Screw Mixers

The use of continuous mixers has increased considerably with the expansion of the plastics industry. End products of greater usefulness depend upon the uniform incorporation of a multitude of additives, stabilizers, antioxidants, fillers, reinforcing agents, as well as alloys and blends of differing base polymers. The base polymer resin needs to be in the melt stage to permit uniform distribution of the additives. Since the thermal conductivity of polymers is generally very low, melting is best accomplished by work energy rather than by heat transfer through a vessel wall. The screw extruder shown in Figure 20 consists of a cylindrical barrel with a helical screw. The root diamater of the screw is smaller in the feed zone. The dry ingredients, sometimes having been premixed in a batch mixer, are conveyed by the rotation of the screw. As the root diameter of the screw is increased, the solids are compressed and caused to melt primarily by friction against each other and against the barrel wall. The homogenized product is forced through some shaping element, such as dies for film, sheet, profile, or strands for subsequent pelletizing. Since the final mixture is extremely viscous (10 – 5000 Pas effective viscosity) the pressure drop through the die is high; consequently, a pressure flow back toward the feed occurs down the screw channel, which partially offsets the drag flow caused by the conveying action of the screw. Cross-channel flow occurs, and most of the mixing is accomplished by the time the last bit of polymer has melted. A further homogenization can be achieved by using a mixing head at the end of the screw, such as shown in Figure 21. The mixing heads operate by forcing all of the product over confined flow paths, or reorienting the elongated laminar layers. The capacity of an extruder is determined primarily by the power which can be supplied. Single screw extruders can be equipped with large gears and thrust bearings to operate with high torque. Although the majority of extruders are in the 25 – 200 mm diameter range, large extruders do exist for tasks such as polyethylene homogenization: these are considerably larger and are equipped with up to 4000 kW drives. Comprehensive extruder texts are Rauwendaal [15], Bernhardt [17], and Tadmor and Gogos [18].

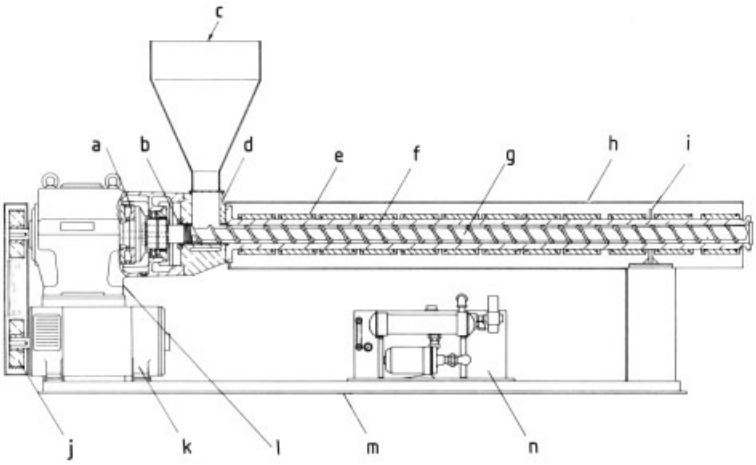

Figure 20. Extruder (Sterling Extruder)
a) Throat bearing assembly; b) Powder seal; c) Hopper; d) Throat; e) Water-cooled heaters; f) Barrel; g) Screw; h) Insulated barrel guard; i) Barrel support; j) Belt and sheave assembly; k) Motor; l) Reducer; m) Base; n) Closed-loop cooling system

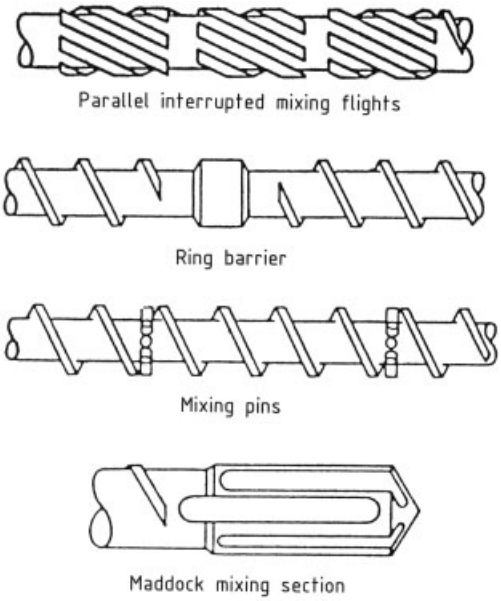

Figure 21. Mixing promoters for single screw extruders

Some variants of single screw extruders are offered which enhance mixing by greater reorientation of the product as it progresses down the barrel. The KoKneader (Fig. 22) achieves this by three rows of teeth projecting inward from the barrel wall, which intermesh with the interrupted screw flight as the screw reciprocates once each rotation. The passage of the teeth, as shown in Figure 23, provides an intimate kneading action [19].

Reorientation of the laminar striations to improve distributive mixing is accomplished in the Transfermix (Fig. 24) and the Cavity Transfer mixer (Fig. 25) as material being conveyed in the screw flights or pockets is repositioned into complementary flights or pockets in the barrel.

When starting with solid resin to be melted and mixed, the overall conversion of mechanical energy of a single screw extruder into useful work is improved by segregating the melt as it is formed by use of some type of barrier screw, as is shown in Figure 26. A channel of slightly decreased outer diameter and increasing width between the flights of the main channel enables the melt to be accumulated free of the unmelted solids and enables the solids to be more effectively compressed and fused between the decreasing width screw channel and barrel.

3.3. Twin Screw Mixers

With two screws inside a figure-eight barrel, a variety of mixing effects can be created depending on whether the screws co-rotate or counter-rotate, and whether they are tangential or they intermesh. Twin-screw extruders are used for a broad variety of products, but their most extensive usage is in polymer processing. They lend themselves to good control of temperature

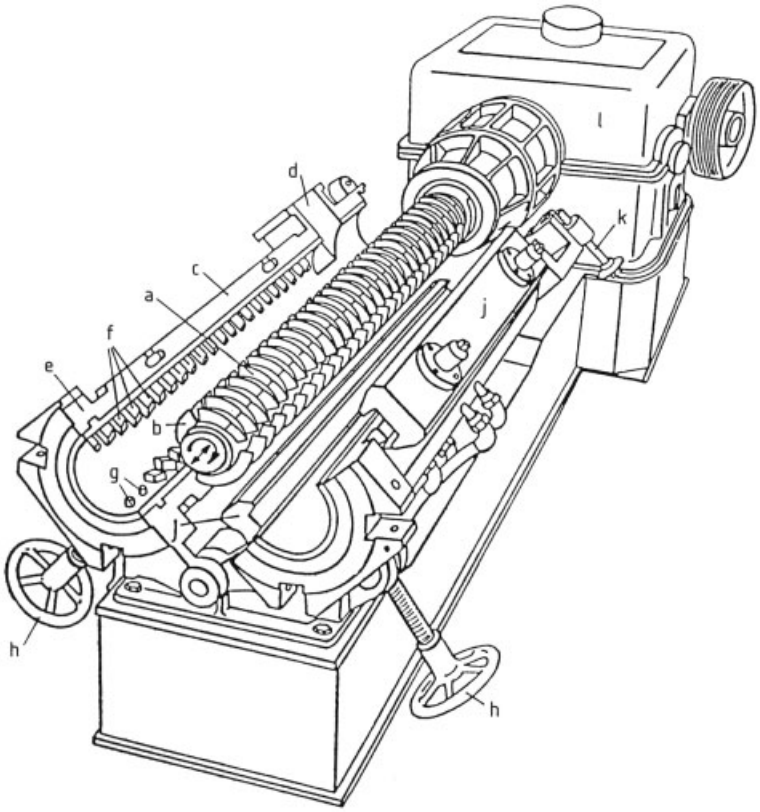

Figure 22. KoKneader mixer (Buss AG)
a) Screw; b) Three rows of teeth; c) Barrel, demountable; d) Inlet; e) Jet flange; f) Cylinder teeth; g) Mixing pins; h) Opening/closing screw; j) Clamps; k) Locking screws; l) Motor

and residence time, and for staging of sequential events such as degassing or further additive addition at the appropriate stage down the barrel. Like the single-screw extruders, twin-screw extruders are used for generating the pressure required for discharging the product through forming dies for profile or stranding.

With the co-rotating intermeshing twin screw depicted in Figure 27, there is complete wiping of all surfaces. Intense mixing action can be created with the use of kneading paddles of similar cross section, but arranged axially in an offset fashion. The kneading paddles create a series of compression – expansion zones [20] while forwarding the mixture down the barrel. Such twin-screw mixers have been built from 15 mm diameter laboratory size to huge commercial machines of 889 mm diameter.

With counter-rotating tangential screws (Fig. 28) the cross mixing between screws provides much distributive mixing and striation reorientation [21]. The large surface exposure is also of benefit in devolatilizing applications.

Robust rotors can be used in short barrels to provide a continuous variant (Fig. 29) of the Banbury-type intensive-fluxing mixer. The mixting action is a combination of intensive shear material between the rotors, and a rolling action of the material itself.

Counter-rotating intermeshing twin-screw extruders (Fig. 30) are very effective in pressure generation since axial back flow is severely restricted by the intermeshing.

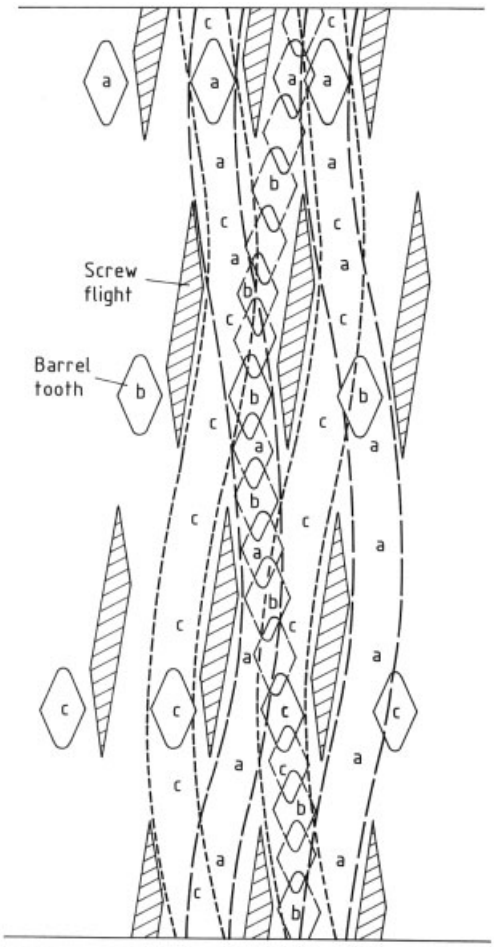

Figure 23. Paths of barrel teeth through KoKneader screw flights
a) – c) Barrel teeth

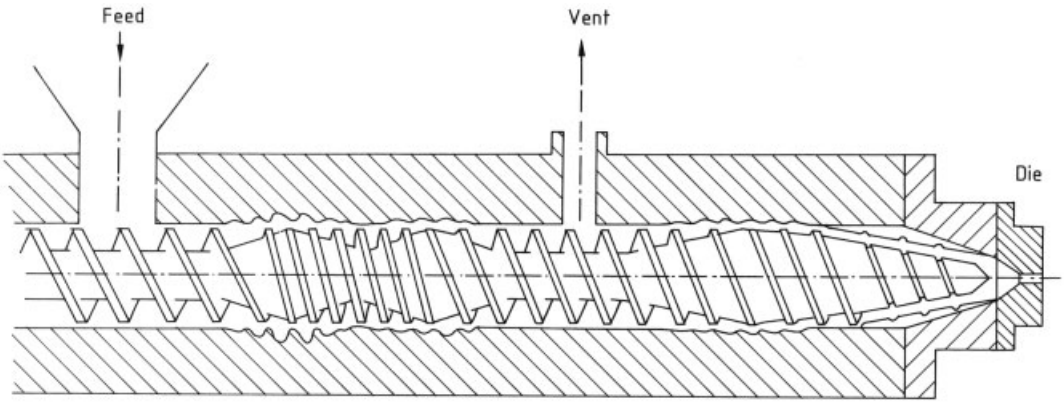

Figure 24. Transfermix (Sterling Extruder)

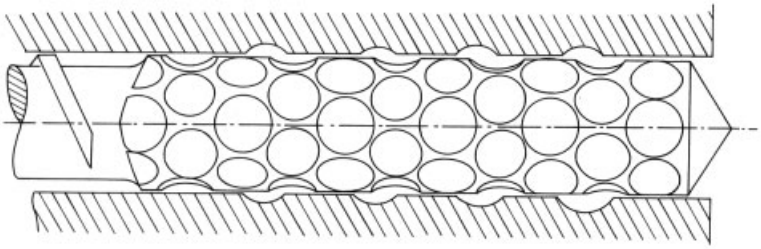

Figure 25. Cavity transfer mixer (RAPRA)

Figure 26. Barrier screw for single screw extruder (Sterling Extruder)

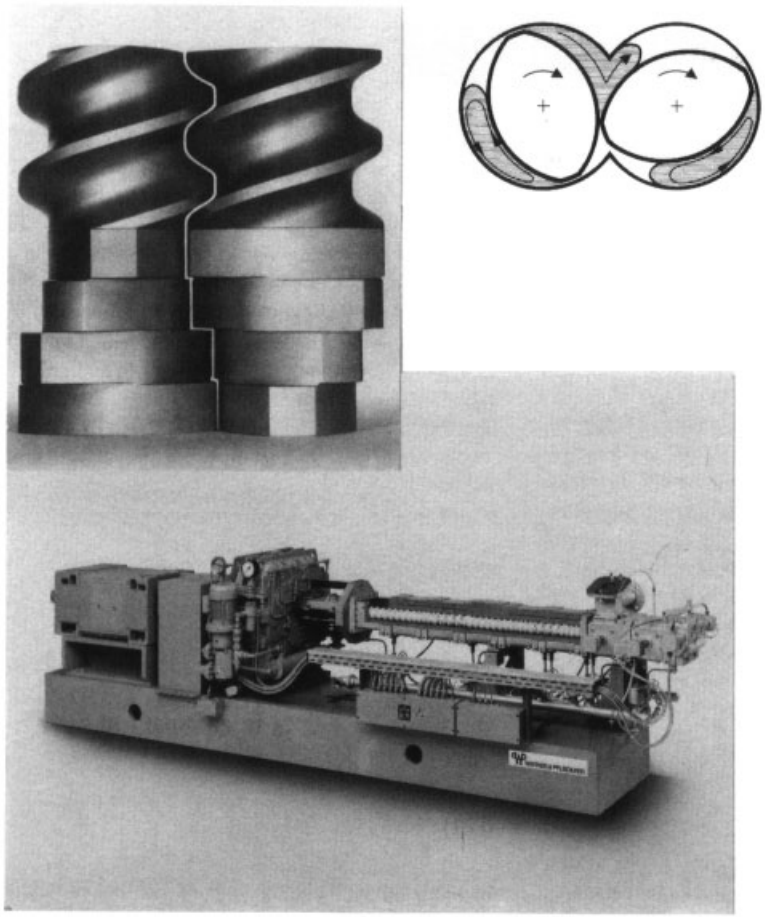

Figure 27. Co-rotating intermeshing twin screw extruder (Baker Perkins)

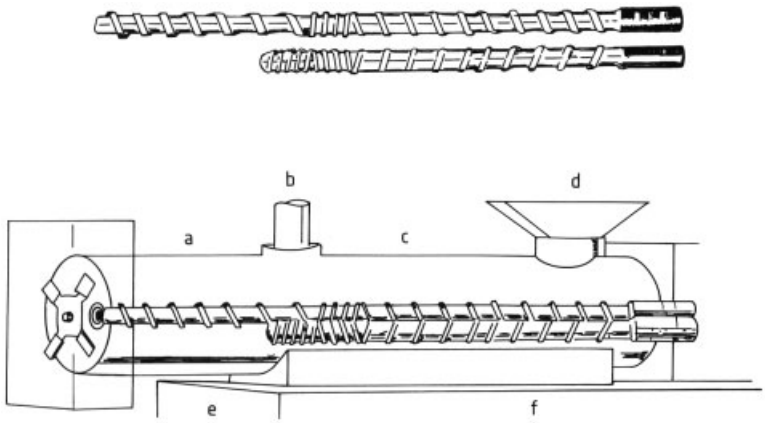

Figure 28. Counter-rotating tangential twin screw extruder (Welding Engineers)
a) Extrusion area; b) Vent area for volatile and air removal; c) Final mixing; d) Feed throat; e) Transfer area; f) Melt down and initial mixing

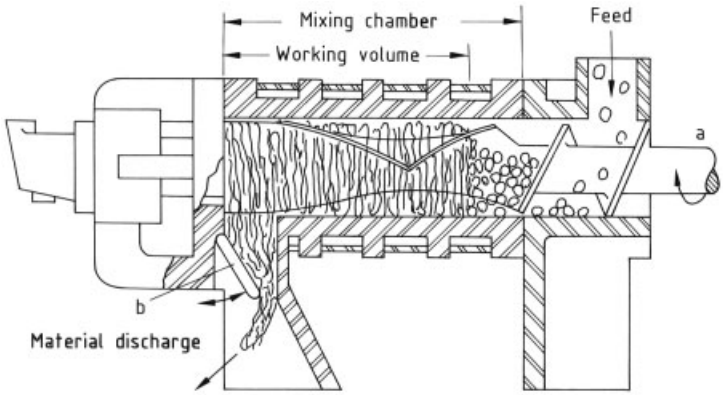

Figure 29. Farrel continuous mixer (Farrel)
a) Rotor drive; b) Discharge orifice gate

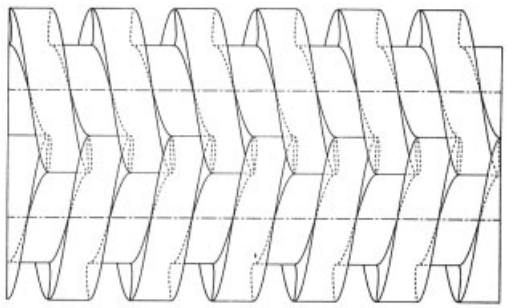

Figure 30. Counter-rotating intermeshing screws

4. References

1. J. M. Ottino, R. Chella, *Polym. Eng. Sci.* **23** (1983) 357.
2. R. Chella, J. M. Ottino, *Ind. Eng. Chem. Fundam.* **24** (1985) 170.
3. D. V. Khakhar, R. Chella, J. M. Ottino, *Proc. Int. Cong. on Rheology* Mexico 2 (1984) 81.
4. D. W. Gailus, L. Erwin, *Proceedings SPE ANTEC* **27** (1981) 639.
5. R. A. Strasser, L. Erwin, *Adv. Polym. Sci.* **4** (1984) 17.
6. F. D. Rumscheidt, S. G. Mason, *J. Colloid Sci.* **16** (1961) 238.
7. F. D. Rumscheidt, S. G. Mason, *J. Colloid Sci.* **17** (1962) 260.
8. H. P. Grace, *Chem. Eng. Commun.* **14** (1982) 738.

9. C. D. Han: *Multiphase Flow in Polymer Processing,* Academic Press, New York 1981.
10. L. Erwin, *Polym. Eng. Sci.* **18** (1978) 738.
11. H. J. Karam, J. C. Bellinger, *Ind. Eng. Chem. Fundam.* **7** (1968) 576.
12. H. F. Irving, R. L. Saxton in V. W. Uhl, J. B. Gray (eds.): "Mixing of High Viscosity Materials,"*Mixing – Theory and Practice,* **vol. 2, chap. 8,** Academic Press, New York 1967.
13. M. H. Pahl: *Mischen bei der Herstellung und Verarbeitung von Kunststoffen,* VDI-Verlag, Düsseldorf 1986.
14. *Proc. Eur. Conf. Mixing,* **4** (1982), 5 (1985).
15. M. H. Pahl, E. Muschelknautz, *Chem. Ing. Tech.* **52** (1980) 285; *Int. Chem. Eng.* **22** (1982) 197.
16. C. Rauwendaal: *Polymer Extrusion,* Hanser Verlag, Macmillan 1986.
17. E. C. Bernhardt (ed.): *Processing of Thermo-Plastic Materials,* Reinhold Publ. Co., New York 1967.
18. Z. Tadmor, C. G. Gogos: *Principles of Polymer Processing,* Wiley-Interscience, New York 1979.
19. D. B. Todd, *Proceedings SPE ANTEC* **33** (1987) 128.
20. D. B. Todd, E. A. Leach, 2nd Int. Conf., "Reactive Processing of Polymers", Pittsburgh 1982.
21. R. J. Nichols, J. C. Golba, B. C. Johnson, Proceedings Polymer Processing Society, Montreal 1986.

Mixing of Solids

See also → Mixing, Introduction

KARL SOMMER, Technische Universität München, Lehrstuhl für Maschinen- und Apparatekunde Weihenstephan, Freising – Weihenstephan, Federal Republic of Germany

1. Mixedness 2
1.1. Theoretical Variance 2
1.2. Variance of the Precision of Measurement 3
1.3. Variance of Random Mixing 3
1.4. Mixedness versus Time 3
2. Practical Determination of Mixedness 9
2.1. Sampling 9
2.2. Empirical Variance 11
2.2.1. Frequency Distribution and Chi-Square Distribution 11
2.2.2. Confidence Intervals 11
2.2.3. Determination of Mixing Time 12
3. Designs of Solid – Solid Mixers . . . 12
3.1. Tumblers 12
3.2. Screw Mixers 13
3.3. Paddle Mixers 13
3.4. Pneumatic Mixers 15
3.5. Silo Mixers 16
3.6. Pile Mixing 16
4. References 16

Symbols

d	diameter in the mixing drum, m
D	diffusion coefficient, m^2/s
D^*	dispersion coefficient, m^2/s
g	mean particle mass, kg
G	sample mass, kg
L	mixer length, m
m	degree of freedom (d.f.)
$\dot{M}$	flow, m/s
n	constant particle number
P	nominal concentration
r	given position, m
s^2	empirical variance
T	transport coefficient, m/s
X	property and sample concentration
ξ	integral variable for R
Δz	displacement, m
σ^2	variance of mixing
χ^2	chi-square variable

Mixing of solids is the distribution of at least one solid component through another. The mixing of granular solids is a widely employed process in chemistry, pharmacy, food manufacturing, the ceramics industry, and ore beneficiation.

Mixing solids is much more difficult than mixing liquids (stirring), where in most cases just a few material constants are enough to characterize the components of the mixture. The mixing process with solids is influenced by disperse properties such as grain shape, grain size, particle-size distribution (gradation), and friction in the bulk material. A comprehensive survey of equipment and methods was published by RIES [1].

Basic mixing mechanisms are not sufficiently well known. This is manifest in the variety of existing mixer types and in the continual appearance of new and novel mixers on the market. In principle, there are two limiting cases:

1) The mixing of powders depends so heavily on product qualities that every product requires its own mixer design.
2) Mixing is a universal process and can be carried out in any apparatus, provided that the material is placed in motion.

Limiting conditions are often more important in mixer selection than the actual mixing process. The following points are often crucial:

1) Explosion safety
2) Cleanability
3) Simultaneous use of the mixer as a silo or interim storage bunker
4) Controlled addition of liquid

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b02_27

1. Mixedness

To assess a mixing process, it is necessary to define a suitable measure of mixedness; and this measure must be kept in mind in all tests, also in regard to the sampling conditions.

Considerable space in the literature on solid – solid mixing has been devoted to the definition and determination of mixedness. Definitions differ so much that any attempt to comprehensively present the state of the art would necessarily end in confusion. The reviews on mixedness are, unfortunately, of little assistance in clarifying the situation [2–5].

One of the reasons for the differences is the measure of mixedness presented in a particular review is usually only appropriate for the mixing problem described in the review. Thus FAN et al. introduce elements of packing structure into the definition of mixedness by considering the number of contact points between particles of the two components [6–12]. Another reason is that theoreticians often find it easier to develop their own statistical measure of mixedness on the basis of some reasonable assumption than to follow up the lines that had been pursued by others.

The first critical evaluation of methods describing the state of a mixture appeared in 1965 [13]. An analysis of 37 statistical expressions describing the state of a mixture is presented in [14]. Although this description of mixedness is unchallenged, it frequently is too complex for the practicing chemical engineer or industrial chemist.

1.1. Theoretical Variance

All mixing operations, whatever their purpose, have in common that they attempt to even out a nonuniformity in the composition or other properties (e.g., color). Such mixing always involves relative motion between particles, for which energy must be expended.

Total equalization within a mixture leads to ideal mixing, which is symbolized by, say, a checkerboard. In powder mixtures, ideal mixing occurs only in the rarest of cases. The best possible result in technical powder mixing is "random mixing." Particles are randomly mixed when the probability with respect to composition or some property is equal in all regions of the mixture or samples from it. In real mixtures, the properties always differ from the ideal value, i.e., nominal concentration, P. This departure is used in the definition of mixedness. Since it is not critical whether the real values depart upward or downward from the nominal value, the mean absolute value of the deviation can be introduced as a measured value. But the absolute value is difficult to handle mathematically, and so the mean-square deviation, the *variance*, has come to be used.

The variance characterizes the mixture as a whole and thus requires a knowledge of the characteristic property everywhere in the mixture. Such knowledge can be gained only by making theoretical assumptions about the state of mixing or by determining the property everywhere in the mixture. Theoretical variances σ^2 must therefore be distinguished from variances determined on the mixture in practice, which are based on a finite number of samples. To prevent confusion, empirical variances are represented by s^2 (see Section 2.2).

The property X can normally be determined only by a measurement procedure of finite precision. Even when the properties really are completely identical in the samples, the measured values thus depart from the nominal value and contribute to the variance [15]:

$$\sigma^2_{\mathrm{tot}}(t) = \sigma^2_{\mathrm{M}} + \sigma^2_{\mathrm{Z}} + \left(1 - \frac{g}{G}\right) \cdot \sigma^2_{\mathrm{syst}}(t) \tag{1}$$

The variance is thus composed of three terms: σ^2_{M}, the variance of the precision of measurement; σ^2_{Z}, the variance of random mixing, which represents the optimal case for granular materials; and σ^2_{syst}, the variance characterizing the mixture, which itself can be used to assess the mixedness.

$$\sigma^2_{\mathrm{syst}}(t) = \frac{1}{1 - \frac{g}{G}} \left\{\sigma^2_{\mathrm{tot}}(t) - \sigma^2_{\mathrm{M}} - \sigma^2_{\mathrm{Z}}\right\} \tag{2}$$

The factor g/G is the ratio of the particle size (e.g., mean particle mass g) and the sample size (e.g., sample mass G). If the sample and particle sizes G and g are assumed to be equal, the measured property depends only on the particle component, not on the state of mixing. The parameter σ^2_{syst} can no longer be measured in the selection of a sampling method. Evaluation of Equation (2) requires a knowledge of σ^2_{M} and σ^2_{Z}.

1.2. Variance of the Precision of Measurement

The precision can usually be determined only through preliminary tests. The variance of the precision must be estimated on the basis of a sufficiently large number of measurements of the selected sample quantity G. Care must be taken that the measurement technique corresponds closely to reality; for example, preparation steps must be included. What is more, a check must be made to see how strongly the precision of measurement depends on the value of the property itself. Often all that is known for a measurement technique is the maximum error. In such cases, a rough estimate for use as a starting point is $\sigma_{\mathrm{M}} \approx \Delta X_{\max}/3$.

1.3. Variance of Random Mixing

If the property to be measured correlates directly with a concentration measure of the mixture, then the variance of random mixing can be calculated from the sampling condition [16], [17]. There are a number of special cases:
Constant particle number n; number concentrations x, p

$$\sigma_{\mathrm{Z}}^2(x) = \frac{p(1-p)}{n}$$

Constant particle number n; mass concentrations X, P

$$\sigma_{\mathrm{Z}}^2(X) = \frac{P(1-P)}{E(G)} \cdot (P \cdot g_{\mathrm{y}} + (1-P)\, g_{\mathrm{x}})$$

where g_{x} is the mean grain mass of component (X) weighted with the distribution q_3^{x} of component (X); g_{y} is the mean grain mass of component (Y) weighted with the distribution q_3^{y} of component (Y), and $E(G)$ is the mean value of sample mass G.

Constant sample property G, mass concentrations X, P

$$\sigma_{\mathrm{Z}}^2(X) = \frac{Q^2 P}{G} \cdot \int_0^1 \frac{g(Q_3^{\mathrm{x}})}{(Q - P Q_3^{\mathrm{x}})^2} \mathrm{d}Q_3^{\mathrm{x}}$$

$$Q = 1 - P$$

where $g(Q_3^{\mathrm{x}})$ gives the individual grain property of component (X) as a function of the undersize $Q^{\mathrm{x}}{}_3$. The smallest grain of component (X) with gradation function Q_3^{x} is larger than the largest grain of component (Y).

Constant sample property G, mass concentrations X, P

$$\sigma_{\mathrm{Z}}^2(X) = \frac{P}{G} \cdot \int_0^1 g(R)\, \mathrm{d}R - \frac{2P^2}{G} \int_0^1 (1-R) \int_0^R \frac{g(\xi)}{(1-\xi)^2} \mathrm{d}\xi \mathrm{d}R \tag{3}$$

where $g(R)$ gives the individual grain properties of components (X) and (Y) as functions of the oversize R. Components (X) and (Y) have the same gradation, $Q_3^{\mathrm{x}} = Q_3^{\mathrm{y}} = 1 - R$.

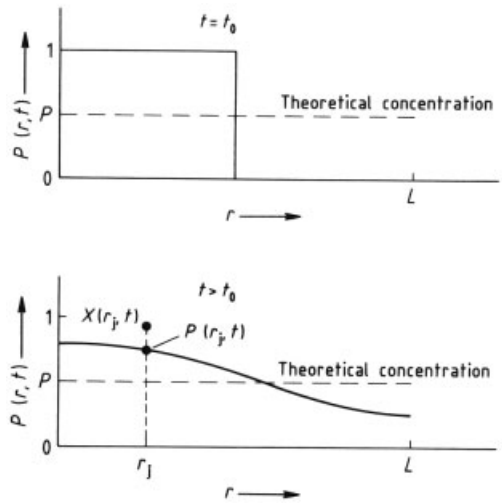

Figure 1. Axial concentration profile in a batch mixer at times t_0 and $t > t_0$. L = length of the mixer, P = theoretical concentration, $P(r_{\mathrm{j}}, t)$ = expected concentration in position r_{j} at time t

1.4. Mixedness versus Time

The random variance σ_{Z}^2 and, if the precision is dependent of concentration, the variance σ_{M}^2 are independent of time. The time dependence of the total variance $\sigma_{\mathrm{tot}}^2(t)$ is then given by the time variation of the systematic deviations alone.

At time $t = t_0$ let the components in a mixer be completely separated. With the passage of time, the spatial concentration distribution will become uniform. The changes in the system can be described in terms of variations along one

coordinate (Fig. 1). In a practical mixing device, the material particles sshould not undergo any predetermined changes of position. Changes of position should be stochastic. This type of motion is treated by the theory of random processes. MÜLLER [18] has shown that mixing, as a stochastic process without aftereffect, can be described by Kolmogorov's second differential equation. Equation (3) is valid if there is no selective transport of the constituents:.

$$\frac{\partial P(r,t)}{\partial t} = \frac{\partial^2}{\partial r^2}\left[D^*(r,t)\cdot P(r,t)\right] \tag{3}$$

where $D^*(r,t)$ is the dispersion coefficient, initially dependent on position and time. If we further assume that $D^*(r,\ t) = D^* =$ constant, then Equation (3) becomes Fick's second law of diffusion:

$$\frac{\partial P(r,t)}{\partial t} = D^*\cdot\frac{\partial^2 P(r,t)}{\partial r^2} \tag{4}$$

If only batch mixers are considered, the solution can be given as a Fourier series [18]:

$$P(r,t) = P + \frac{2}{\pi}\sum_{\mu=1}^{\infty}\frac{\sin\mu\cdot\pi}{\mu}\cos\mu\cdot\pi\frac{r}{L} \cdot\exp\left(-\mu^2\left(\frac{\pi}{L}\right)^2\cdot D^*\cdot t\right) \tag{5}$$

The system variance is again calculated as the average of the concentration deviations:

$$\sigma^2_{\text{syst}} = \frac{1}{L}\int_0^1\left[P(r,t) - P\right]^2\,\mathrm{d}r$$

and we obtain

$$\sigma^2_{\text{syst}} = \frac{2}{\pi^2}\sum_{\mu=1}^{\infty}\frac{\sin^2\mu\pi P}{\mu^2} \cdot\exp\left(-2\mu^2\left(\frac{\pi}{L}\right)^2 D^*\cdot t\right) \tag{6}$$

The series can be evaluated for two limiting cases:

1) At time $t = 0$ [19]:

$$\sigma^2_{\text{syst}}(t=0) = P(1-P) = \sigma_0^2 \tag{7}$$

The limiting case of the initial variance is em- bodied in Equation (6).

2) For long times, the higher-order terms ($\mu > 1$) can be neglected:

$$\sigma^2_{\text{syst}}(t) = \frac{2}{\pi^2}\cdot\sin^2(\pi\cdot P) \cdot\exp\left(-2\left(\frac{\pi}{L}\right)^2 D^*\cdot t\right) \tag{8}$$

Equation (8) shows the desired concentration dependence of the system variance. For $t\to\infty$, $\sigma^2_{\text{syst}}\to 0$ (Fig. 2). The limiting case of infinitely long times is likewise embodied in Equation (6).

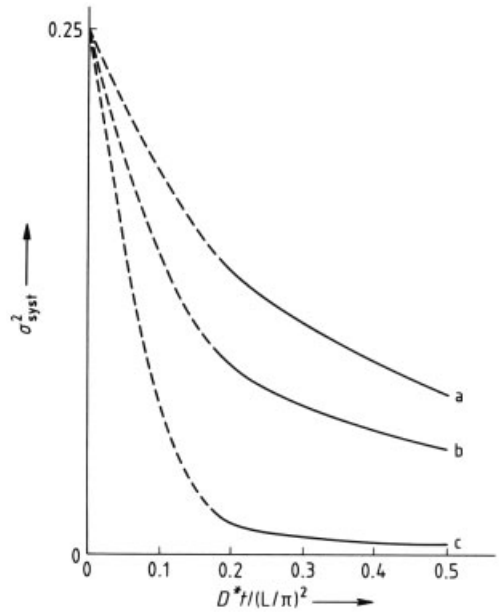

Figure 2. Systematic variance $\sigma^2_{\text{syst}}(t)$ as a function of dimensionless time, for pure dispersion. D^* is the dispersion coefficient; L, the mixer length. Parameter: nominal concentration P: a) 0.5; b) 0.3; c) 0.1

Limit of Low Concentrations *P*. If component (X) is present only in small quantities, so that $P \ll 1$, the sine in Equation (8) can be expanded into a series, and all but the first term dropped:

$$\frac{\sigma^2_{\text{syst}}(t)}{P^2} = 2\cdot\exp\left(-2\left(\frac{\pi}{L}\right)^2 D^*\cdot t\right) \tag{9}$$

Here σ_{syst}/P is called the system coefficient of variation. From Equation (9) follows that, for small admixtures $P \ll 1$, the system coefficient of variation is invariant with respect to both sample size and mixture composition, satisfies the requirements of SCHMAHL [20], and is suitable as a measure of mixedness.

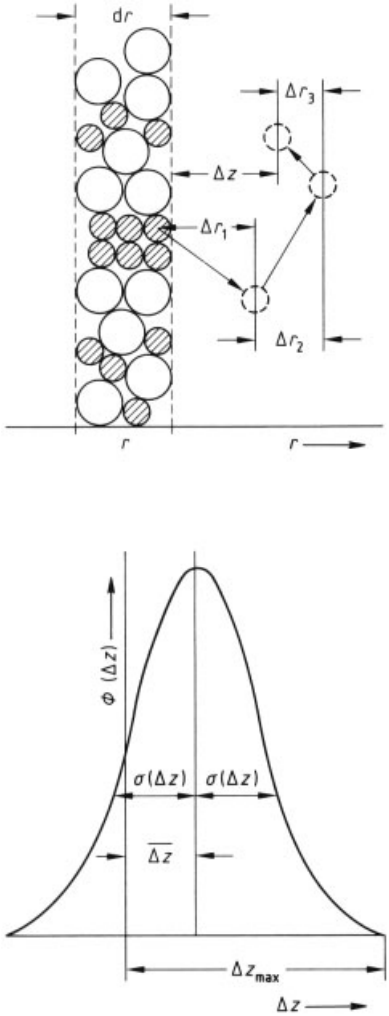

Figure 3. Schematic diagram illustrating the derivation of the Fokker–Planck differential equation: Δz is the displacement in time τ; $P(r)$, the concentration at point r; $\phi(\Delta z)$, the distribution function of Δz

Limit of Equal Proportions. If the two components to be mixed are present in roughly equal quantities, so that $P \approx 0.5$, Equation (8) simplifies to

$$\frac{\sigma^2_{\text{syst}}(t)}{P^2} = 2 \cdot \exp\left(-2\left(\frac{\pi}{L}\right)^2 D^* \cdot t\right) \tag{10}$$

In the region of almost equal proportions, the system variance itself is independent of sample size and mixture composition, satisfies the requirements of SCHMAHL, and is an appropriate measure of mixedness.

If the components of powder mixtures differ in grain shape, grain size, or density, stochastic mixing processes [21–28] often exhibit segregation that cannot be accounted for by the simple diffusion model with positive dispersion coefficient. The more fundamental MARKOV theories of random processes must be introduced in these cases; they lead to the so-called Kolmogorov equations. These equations were first written for statistical dynamics by FOKKER [29] and PLANCK [30] (Fig. 3).

Suppose there are a large number of spheres, mutually independent and having identical surface properties. A model system comprises small glass spheres in a matrix of large glass spheres. Let each of the spheres be set in motion. At a given time t, the concentration at a given position r has some value $P(r)$. Each of the small spheres experiences very small, rapid, stochastic collisions. Now the question is how the concentration $P(r)$ at point r changes in an interval of time. In this interval, each sphere experiences several collisions r in which the direction of motion may change. The sum of all these changes of position is called the displacement Δz. Each sphere will exhibit a different displacement Δz, even if the spheres are located at the same point r. The displacements Δz are statistically distributed with distribution function $\phi(\Delta z)$, where $\Delta z_{\max}$ is the greatest displacement that occurs. It should always be small in comparison with the mixer dimensions. Here $\overline{\Delta z}$ is the mean displacement of the spheres, that is, a mean convective transport.

Besides this convective motion, a broadening of the distribution density results from the stochastic nature of the process. The broadening is described by the mean square of the displacement $\overline{\Delta z^2}$:

$$\overline{\Delta z^2} = \sigma^2(\Delta z) + \left(\overline{\Delta z}^2\right) \tag{11}$$

where $\sigma^2(\Delta z)$ is the variance of the distribution density $\phi(\Delta z)$.

The quadratic quantity $\overline{\Delta z^2}$ is, of course, always positive. FOKKER [29] and PLANCK [30] calculated the change in concentration $P(r)$ in segment dr from the difference between the number of particles leaving dr in time τ and the number entering dr from adjacent layers in the same time. They stated the result in differential equation form:

$$\frac{\partial P(r)}{\partial t} = \frac{\partial}{\partial r}\left[P(r) \lim_{\tau \to 0} \frac{\overline{\Delta z}}{\tau}\right] + \frac{\partial^2}{\partial r^2}\left[P(r) \lim_{\tau \to 0} \frac{\overline{\Delta z^2}}{2\tau}\right] \tag{12}$$

As was already mentioned, $\overline{\Delta z}$ is the mean convective transport of the small spheres at position r in the time interval. Then $\overline{\Delta z}/\tau$ is the mean convective transport velocity. If the limit

$$\lim_{\tau \to 0} \overline{\Delta z}/\tau$$

exists, it is called the transport coefficient T and has the units of velocity. If the limit $\lim \overline{\Delta z}^2/2\,\tau$ exists, it is called the *dispersion coefficient* D^*; it has the units of area per unit time, the same as the diffusion coefficient D. The dispersion coefficient, however, is not a pure adjustment factor but a statistically derived quadratic quantity, and as such it is always positive. Then Equation (12) is written as

$$\frac{\partial P(r)}{\partial t} = -\frac{\partial}{\partial r}\left[P(r)\,T(x)\right] + \frac{\partial^2}{\partial r^2}\left[P(r)\,D^*(r)\right] \tag{13}$$

The form of this equation, which describes the time variation of concentration at positive r, matches Fick's second law of diffusion.

The analog of Fick's first law of diffusion is the mass flow rate through a unit cross-sectional area at position r at time t:

$$\dot{M}(r) = P(r)\,T(r) - \frac{\partial}{\partial r}\left(P(r)\,D^*(r)\right) \tag{14}$$

Thus the mass flow $\dot{M}$ is composed of the convective transport $\dot{M}_{\mathrm{T}}$ and the dispersive transport $\dot{M}_{\mathrm{D}}$.

Let us make the simplifying assumption that $D^*(r) = D^*$ independent of the coordinate r. If in this case the transport coefficient $T(r) = 0$, Equation (14) becomes Fick's first law of diffusion, and the dispersion coefficient D^* is identical with the diffusion coefficient D. If in the mixture under consideration the components are completely separate from each other at time t, the diffusive equalization already described takes place. The variance σ^2 of the sample concentration declines steadily from the initial state σ_0^2 until it reaches the value of the sampling error σ_z^2.

When a transport flow is present, there are two possibilities: (1) flows initially in the same direction (Fig. 4, third curve from top) and (2) flows initially in opposing directions (Fig. 4, bottom curve). If the two flows are initially in the same direction, transport at first aids the equalization of concentration, but then the diffusion flow reverses and opposes the transport until both flows are equal in magnitude. This state, which is steady, is a condition of partial mixing. At first the mixedness rapidly improves, but as demixing increases the mixedness becomes worse again. An experimental example of such qualitative segregation is found in the trials reported in [18] (see Fig. 5). The differing transport properties of iron (630 – 750 µm) and quartz (100 – 200 µm) and of iron (300 – 400 µm) and limestone (40 – 60 µm) give rise to segregation.

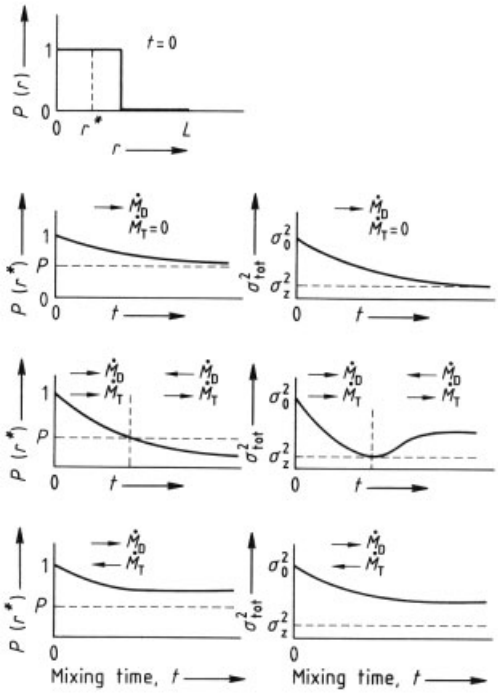

Figure 4. Concentration $P(r*)$ at position $r*$ and mixedness σ^2_{tot} as function of mixing time [18]

The empirical coefficients of variation reflect the prediction from the model; the diffusion and transport flows are initially in the same direction in Test 1, opposing in Test 2. For this reason, Test 1 (Fig. 5, Curve a) leads to homogeneous mixing at first but then to segregation. In Test 2, (Fig. 5,

Curve b) the stochastically homogeneous state is not reached at any time.

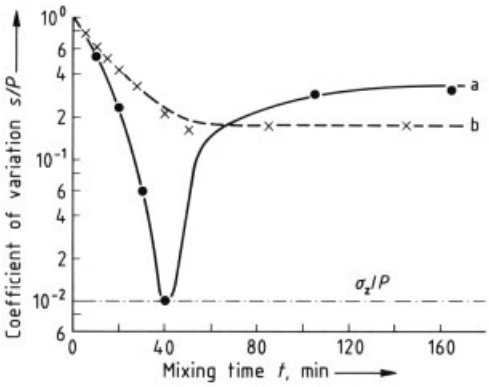

Figure 5. Coefficient of variation s/P versus mixing time t a) Quartz, 100 – 200 µm, and iron, 630 – 750 µm (indicator); b) Limestone, 40 – 60 µm, and iron, 300 – 400 µm (indicator), [18]

The segregations at constant dispersion coefficient, as described in the preceding section, always occur when movement of internal parts or inclination of the drum causes selective transport that opposes diffusive mixing. Segregations are seen in drum-type mixers even when no external convective transport processes are taking place. However, there is always an equalization of concentration when the transport coefficient vanishes and the dispersion coefficient is a constant. It should therefore be checked whether a position-dependent or concentration-dependent dispersion coefficient can account for the observed segregations without convective transport. For $T\ (r) = 0$, Equation (14) becomes

$$\frac{\partial P\,(r)}{\partial t} = \frac{\partial^2}{\partial r^2}\,[P\,(r)\,D^*\,(r)] \tag{15}$$

If the product is first differentiated with respect to concentration and the concentration is then differentiated with respect to r, a form comparable with Fick's first law of diffusion is obtained:

$$\dot{M}\,(r) = -\,\frac{\partial\,(P\,D^*)}{\partial P}\,\frac{\partial P\,(r)}{\partial r} \tag{16}$$

If the dispersion coefficient is taken to depend only on the concentration, the product rule can be used in differentiation:

$$\dot{M}\,(r) = -\left(D^* + P\frac{\partial D^+}{\partial P}\right)\frac{\partial P\,(r)}{\partial r} \tag{17}$$

The expression in parentheses corresponds exactly to the Fick diffusion coefficient:

$$D = \left(D^* + P\frac{\partial D^*}{\partial P}\right) \tag{18}$$

While the dispersion coefficient D^* is always positive, by hypothesis, the diffusion coefficient D can become negative when the concentration dependence of D^* supplies a sufficiently large negative derivative. Then there exists a material flow in the direction of higher concentrations; that is, the components separate.

To investigate these processes, glass spheres were mixed in a glass drum 205 mm long and 114 mm in diameter. First, dark and light spheres of the same size (d = ca. 2 mm) were used, with a fillage of 38 %; the drum rotated at 45 rpm. The mixture ratio was 1 : 1. Because the spheres were equal in size and made of the same material, the dispersion coefficient could be taken as constant independent of the concentration. The contents of the drum thus became mixed to stochastic homogeneity in a short time (≈20 min).

If small glass spheres (d = 1 mm) are mixed with large ones (d = 2 mm), the mobility of the small particles in the matrix of larger particles is estimated to be greater than in an environment of equal-sized spheres. Thus the dispersion coefficient decreases as the concentration increases. If the decline is great enough, the diffusion coefficient may become negative and the components, completely separate at first, will not mix. In the practical experiment, mixing was not observed even over very long mixing times. The contents of the drum remained separate.

If the small spheres are layered on top of the large ones, the concentration is initially uniform in the axial direction. Diffusion flow is theoretically impossible in this instance. But the equilibrium is labile if the diffusion coefficient is assumed to be negative. Because of irregularities in charging and the boundary conditions, violations can always occur, leading to concentration shifts and thus making segregation possible. Such upsets were seen close to the right and left end walls in the experiment. Radial disks near these walls showed an increased concentration of small spheres after five minutes of mixing, while the concentration in the other zones decreased. The observed segregations remained stable even over longer mixing times. The number of violations depends on the length of the mixing drum.

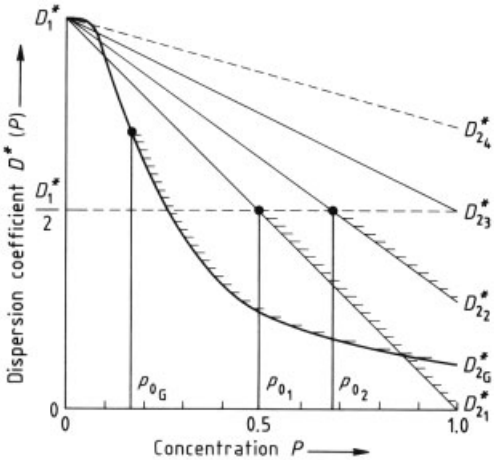

Figure 6. Estimated dispersion coefficient as a function of concentration P

Numerous strips could be generated by varying the drum geometry [21–23]. Symmetry in the stated form can be expected only when the initial conditions are symmetrical. If violations are imposed in the form of local increases on concentration, individual strips can even be programmed to appear at various places. Exact observation reveals that the quantity of small spheres concentrated in the stripes does not make up half the mixture contents, as at the beginning. Thus a large fraction of the small spheres must still be present in the matrix of the large spheres.

This fact can also be explained on theoretical grounds. Equation (17) shows that, for concentrations $P \to 0$, the diffusion coefficient becomes positive in every case, even when the derivative $\partial D^*/\partial P$ is large and negative. At low concentrations, then, the components are always miscible. If it is assumed, for example, that a linear relationship holds between the dispersion coefficients of the small spheres, D_1^* for concentration $P \to 0$ and $D_{2_i}^*$ at concentration $P = 1$, then the limit of segregation lies at $P_0 = 0.5$ in the first case (D_2^*) and at a higher concentration P_0 in the second case (D_2^*) (Fig. 6). If the concentration is below the limit P_0, the diffusion coefficient is positive, so that equalization of the concentration takes place by diffusion. At a higher concentration, segregation occurs. If the dispersion coefficients are in the ratio of $D_1^* : D_2^* = 1 : 2$, the concentration limit is near $P_0 = 1$; that is, at this ratio and all ratios nearer to unity, no segregation takes place.

The linear concentration dependence is merely an assumption, which could not be applied to the experimental system because of the high concentration limit of $P_0 = 0.5$. On the other hand, if the dependence resembled a normal distribution, the segregation limit P_0 lies at the steepest part of the curve, so that it may occur at much lower concentrations (Fig. 6).

Practical Consequences. The postulated concentration dependences of the dispersion coefficients are quite arbitrary. The first attempts to determine these dispersion coefficients for solids were made by Cahn and Fuerstenau [31], [32]. Unfortunately, these results cannot be applied generally. The analysis of mixing processes in terms of stochastic processes does, however, have qualitative implications. If there is a tendency to segregation because of convective transport, then according to [18], [24] the selective transport behavior can be altered by even slight changes such as the addition or removal of mixing elements.

Convective segregation tendencies should be offset, whenever possible, by a suitable choice of grain size and density [25]. If it is known that two components tend to segregate on account of negative diffusion coefficients, then the concentration dependence of the dispersion coefficient must be reduced. This is accomplished in practice by agitation in fluid-bed and centrifugal mixers. The mean free spacing between particles is large in comparison to the particle size, so that the mobility of individual particles is less dependent on the size of the surrounding particles. The success of powder mixers using the fluid-bed and centrifugal principles is due not only to the absolutely larger dispersion coefficients, but certainly also to the weaker concentration dependence. There is, however, a great danger that the ini-tially homogeneous mixtures may separate again during slow discharging and charging, because of the concentration dependence of the dispersion coefficient.

A second way to homogenize a mixture that tends to segregate is to add small amounts of liquid [33–35]. The liquid causes water to form bridges between the spheres. Because of the lesser mobility, the dispersion coefficients are much lower, but the difference in environment

and thus the concentration dependence no longer play a dominant role. Diffusive mixing is thus possible. The addition of about 1 wt % water to a charge of glass spheres 1 and 2 mm in diameter leads to a stochastically homogeneous final state after 10 min of mixing.

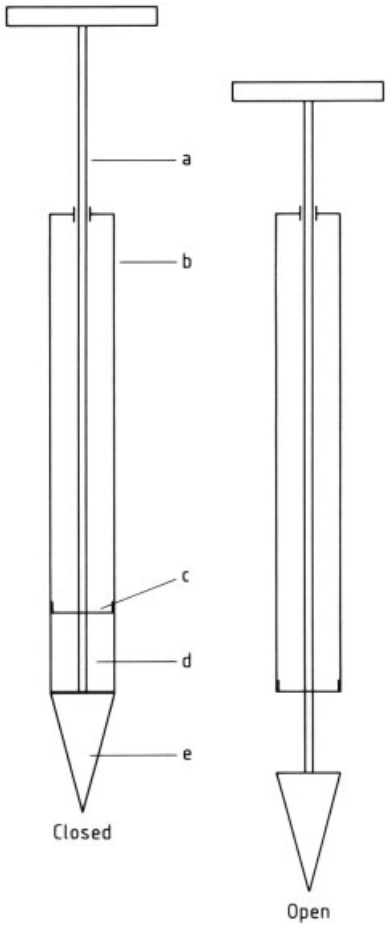

Figure 7. Thief for taking samples of free-flowing solids (Design: Lehrstuhl für Maschinen- und Apparatekunde [Department of Machinery and Apparatus], Freising – Weihenstephan, Federal Republic of Germany)
a) Guide rod; b) Tube; c) Cover; d) Sample chamber; e) Point

2. Practical Determination of Mixedness

For the theoretical mixedness (σ^2_{tot} or σ^2_{syst}) to be determined, the property must be known everywhere in the mixture. In practice, however, only a finite number of samples can be taken from the mixture. A variety of devices can be used for this purpose.

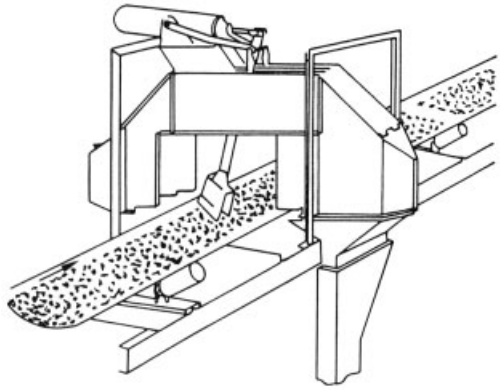

Figure 8. Sample stripper

2.1. Sampling

The simplest methods of sampling are manual ones employing sampling shovels or ladles. Sampling shovels have the top and one side open, and thus are suitable for pulling material out of a stationary bed. The sampling ladle is open only on top and is used for removing granular material from a falling stream of material. Sampling forks are good only for coarse materials such as coke.

Thieves and sampling drills are used for sampling from mixers and silos. These devices do not just sample the surface, as shovels do, but reach to depths of several meters. The thief, in simplest form, consists of a steel tube with an inside diameter of 30 – 50 mm, a wall thickness of 2 – 4 mm, and a length of 1.50 m (up to 4 m in special cases). The thief in Figure 7 is especially well suited to taking samples from free-flowing powders [36], [37]. The thief is closed and pushed into the material; sliding the tube axially along the guide rod opens and fills the chamber; the thief is then reclosed and withdrawn. For cohesive material such as moist soils, the sampling drill is appropriate. It consists of two fixed, curved plates mounted on a crosspiece. The drill is advanced into the pile by a screw motion.

The best-known sampling device for fluidized solids is the channel sampler. Normally, the channel sampler is advanced across the falling stream of material. Where the material under study is being conveyed by a belt, trough

or screw conveyor, hand sampling is more and more being replaced by automatic sampling. A simple way to take samples from a belt is to use a fixed stripper, which intercepts an exact cross-sectional area of material transverse to the conveying direction (Fig. 8).

Pendulum samplers (Fig. 9) and impact samplers (Fig. 10) find use in automatic sampling at the discharge from a belt or trough conveyor. Material is acquired at regular intervals, either transverse or parallel to the discharge direction. The pendulum sampler consists of a slotted vessel suspended from a pendulum arm. A drive rod causes it to move through the stream of material. When the sampler returns to its rest position, the material is dumped by opening the bottom gate [38], [39]. In the impact sampler, a concave impactor knocks the material out of the falling stream in the conveying direction.

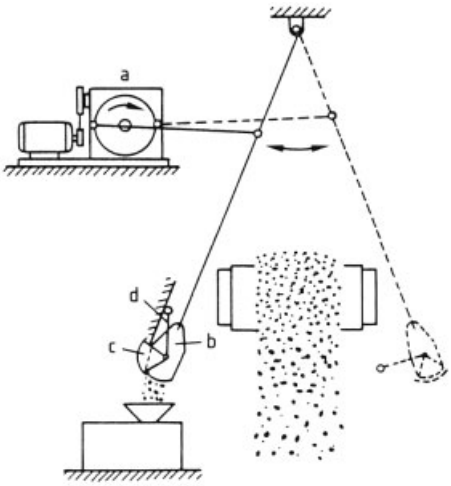

Figure 9. Pendulum sampler
a) Sabo timer drive; b) Slotted vessel; c) Bottom gate; d) Actuating lever with roller

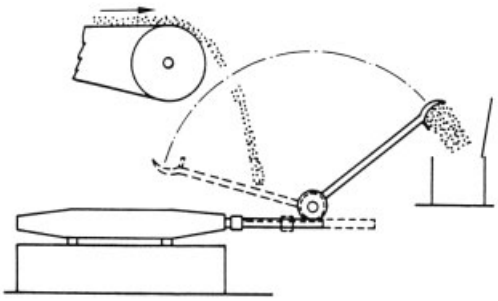

Figure 10. Impact sampler

A simple slot (Fig. 11) takes samples out of a trough conveyor. The upright slot diverts a substream of material, the sample, out of the middle of the trough. The substream remains proportional to the material flow rate as the layer depth and conveying speed vary. With a star-wheel device or a screw sampler with several slots and holes on top, fluidized material can be diverted from discharge ducts and conveying pipes [40]. The screw-type sampler can also remove a substream from a larger screw conveyor (Fig. 12). If the sampling screw rotates at a constant speed, the quantity sampled is independent of the quantity of material moved by the large screw.

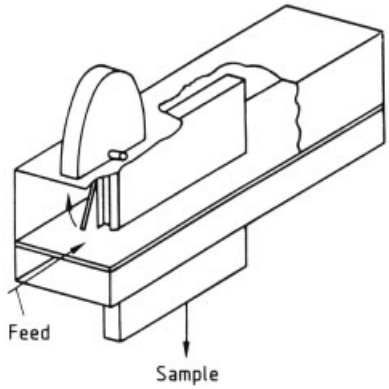

Figure 11. Slot arrangement for sampling from a trough conveyor

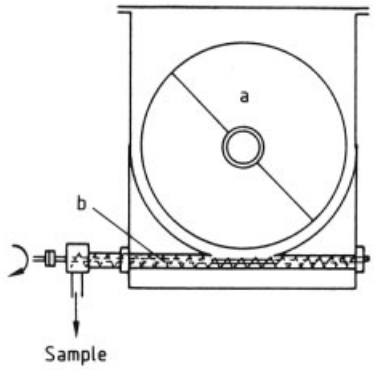

Figure 12. Screw sampler (b) mounted on screw conveyor (a)

In many cases, the samples thus acquired are too large for measurement. The samples are remixed in the laboratory, and a sample for analysis is then taken. This procedure should be used with caution. The laboratory mixing pro-

cess must itself be monitored, since there is always a danger of segregation. It is better to use a sample divider to reduce the sample to a size that can be analyzed. Errors are also possible with this technique, but they can be determined. If, however, the particle size is small in comparison with the sample for analysis, these errors are usually negligible. Otherwise, the starting material must be comminuted until this requirement is satisfied before the sample is divided. Given these conditions, the analytical value can be assigned to the starting sample.

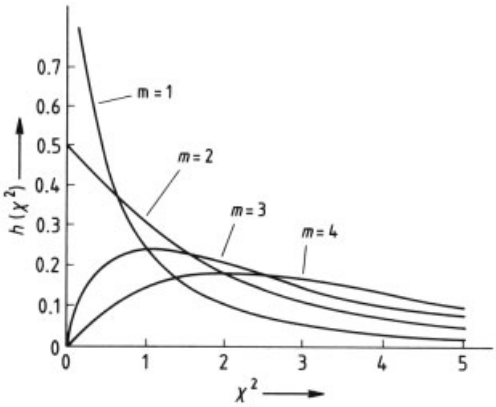

Figure 13. Chi-square distributions for degrees of freedom $m = 1, 2, 3$, and 4

2.2. Empirical Variance

2.2.1. Frequency Distribution and Chi-Square Distribution

If there are analytical values from k samples, the *mean-square deviation* or *variance* s^2 is

$$s^2(X) = \frac{1}{k-1} \sum_{i=1}^{k} (X_i - \bar{X})^2 \tag{19}$$

$$s^2(X) = \frac{1}{k} \sum_{i=1}^{k} (X_i - P)^2 \tag{20}$$

The second formula is used when the mixture composition P is known. If the analytical values X_i are normally distributed (an assumption that can always be made as a first approximation), the quantity $\lambda^2 = m\frac{s^2(x)}{\sigma^2(x)}$ has a *chi-square distribution*; the *number* of *degrees of freedom* (d.f.) is $m = k - 1$ when Equation (19) is used or $m = k$ when Equation (20) is used. The chi-square distribution is an asymmetrical one-dimensional probability distribution (Fig. 13). It allows a statistical assessment of the quality of the estimate s^2 in relation to σ^2, where σ^2 is the expectation of the statistically varying quantity s^2 (Fig. 14). The more d.f. and thus the greater the number of samples, the more frequently s^2 will be found close to the expected value σ^2, and the more symmetrical the frequency distribution will be.

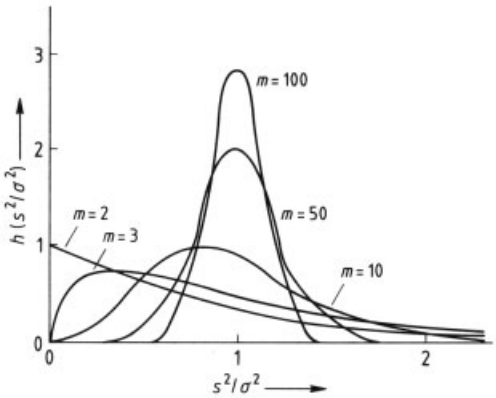

Figure 14. Frequency distributions of s^2/σ^2 for degrees of freedom $m = 2, 3, 10, 50$, and 100

2.2.2. Confidence Intervals

If a certain probability is stated (in engineering usually 95 – 99 %, but often higher in the pharmaceutical field), the variance can be found from the chi-square distribution (Fig. 15). For example, it is possible to say with 95 % confidence that

$$\frac{s^2}{\sigma^2} \geqq \frac{\chi_u^2}{m}$$

Conversely, it can be asserted with the same degree of confidence that the true value of σ^2 (the mixedness) satisfies the relation

$$\sigma^2 \leqq \frac{m}{\chi_u^2} \cdot s^2 \tag{21}$$

In the diagram of mixedness versus time, Equation (21) gives the confidence interval for the true mixedness (Fig. 16). With 95 % confidence, the true value is lower than the stated limit.

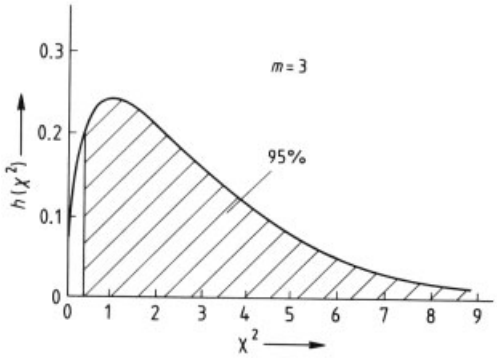

Figure 15. One-sided 95 % confidence interval of the chi-square distribution, $m = 3$, $\chi^2_u = \chi^2_{0.05}$

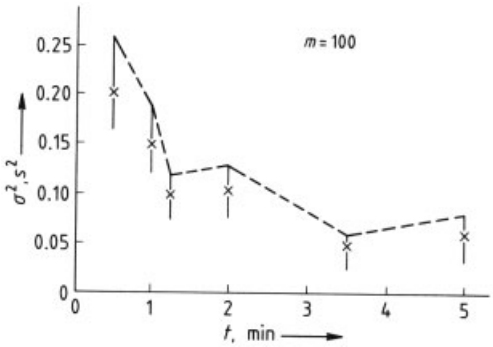

Figure 16. Example of a curve of mixedness versus time. Vertical bars represent confidence intervals

2.2.3. Determination of Mixing Time

More difficult to determine than the confidence interval is the mixing time required to achieve a given mixedness σ^2_{goal}. The frequency distributions of s^2 (Fig. 14) show that, for few d.f., the number of s^2 values smaller than the true value of σ^2 is greater than the number of s^2 values greater than the true σ^2. For $m = 2$ d.f., the frequency is actually greatest for $s^2 = 0$.

The measured values of s^2 thus give the illusion of far too good a result. For the determination of the mixing time, as in the case of the mixedness, it is more correct to use the upper limit of the confidence interval.

Two cases must be distinguished in practice:

1) The mixedness σ^2_E in the mixer is clearly better than the prescribed mixedness σ^2_{goal} (Fig. 17 A). The mixing curve distinctly intersects the goal value within two time intervals, and the mixing time can be determined accurately.
2) The mixedness in the mixer σ^2_E just reaches the prescribed mixedness σ^2_{goal} (Fig. 17 B).

When the final mixedness has been reached, the measured values of s^2 fluctuate about this value in accord with their frequency distribution. In most cases, the upper limit of the confidence interval is above the mixedness target, and no clear intersection can be expected. Only when a measured value is smaller than the value s^2_u from the chi-square distribution χ^2_u does the upper limit of the confidence interval lie below the target value. This happens $(1 - P)$ times, where P is the selected probability; thus, if the confidence interval is 95 %, this happens 5 times in 100 cases.

To keep from choosing too long a mixing time, the time intervals for sampling must be held short, even though this increases the cost.

If the final mixedness is prescribed not by the mixer (σ^2_{syst}) but by the material being mixed and the precision of measurement,

$$\sigma^2_E = \sigma^2_M + \sigma^2_Z$$

then minimizing the final variance σ^2_E is often a better way to reduce the cost in the determination of mixing time. This can be done by improving the sample preparation method and the precision of measurement. If σ^2_Z is the only possible approach for improvement, the ratio of particle size to sample size can be reduced (larger samples or smaller particles).

3. Designs of Solid – Solid Mixers

Coarse mixing in solids mixers is accomplished by the continuous movement of material in the process volume. It must be possible for adjacent particles to change places. Depending on the flowability of the powders being mixed, this can be achieved through a variety of mechanisms [41] and in a variety of geometries [42]. The large number of mixers on the market can be classified under several headings [43], [44].

3.1. Tumblers

A tumbler mixer is typically a vessel in which the mixture is rotated (Fig. 18). A tumbling motion

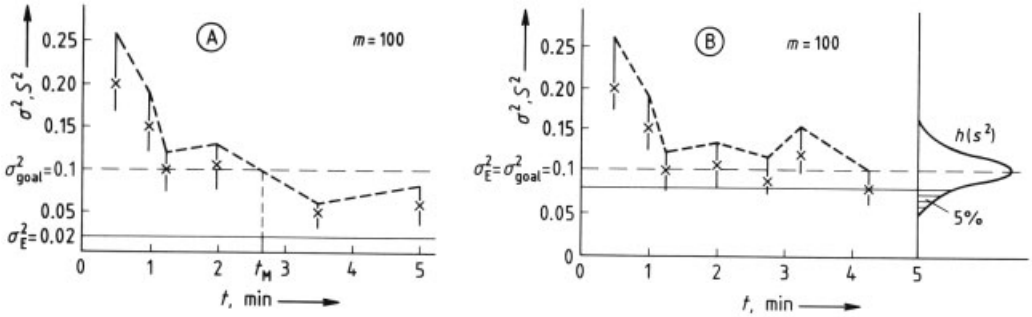

Figure 17. Mixedness versus time curves, where the final mixedness σ^2_{E} is (A) very much smaller than the goal value σ^2_{goal} or (B) roughly equal to σ^2_{goal}. Vertical bars represent confidence intervals.

may be superimposed on the rotational motion, or components mounted on the inside of the mixing drum (internals) may force the loose material into motion. A quite simple and well-known mixer of this type is the concrete mixer. It has a capacity of 2 m^3/h on average, with mixing times of some 1 – 3 min. When rotation speeds are low, the processes of importance for mixing occur in relatively thin layers at the surface of the material bed (*cascade motion*). The slow relative motion promotes any segregation effects that occur. As the rotation speed *n* increases, the fraction of solid material in motion becomes greater. Some individual particles escape from the bed and fall freely through the empty volume onto the bed below (*cataract motion*). If a critical rotation speed *n* is exceeded, the particles are held against the drum wall by centrifugal force, and free movement and thus mixing are prevented. The scaling law for model experiments is

$$Fr = n/n_{\mathrm{crit}} = n^2 d/\mathrm{g} = \text{constant}$$

where *d* is the diameter of the mixing drum. *Fr* is the dimensionless Froude number. Tumblers have drum volumes of 2 – 100 000 L and drive powers of 0.2 – 33 kW [43].

3.2. Screw Mixers

Another class of mixers operating at relatively low Froude numbers comprises screw mixers, in which the pan or drum is stationary and a screw or screw-shaped flights or ribbons move inside it (Fig. 19). The flights move through the bed of material at tool speeds of up to 2 m/s, loosening the material and moving portions of it along. This process creates voids, which can be filled by free-flowing material. Ribbon-type impellers, usually made of flat steel strip, rotate about a horizontal axis in horizontal drums or about one or more vertical axes in pans.

Another widespread form is the screw mixer in a conical vessel (Fig. 20). One or two orbiting screw conveyors cause the material not only to rotate in the cone but also to move upward at the same time. The dual motion causes continuous mixing. ENTROP [44] has developed design diagrams for this type of mixer, giving both the mixing time and the power consumption.

3.3. Paddle Mixers

Paddle or spreader mixers operate at higher rotation speeds; the peripheral speed of the mixing tool is 10 – 40 m/s. The shaft, usually vertical, bears paddles or shovels on short end pieces (Fig. 21); plowshares or partial screw flights are also used. The mixing elements separate particles from the bed of material, setting them in motion in the air space above the charge and allowing them to mix. Angled paddles or plowshares allow an axial movement that promotes axial mixing. Because the cylinder wall and the end surfaces are accessible, liquids can be metered in with little or no crust formation. Agglomerates can be broken either with cutter heads rotating at high speed in the bed of material or with pin devices. The Eirich mixer features a pan and shovels rotating in opposite directions, thus efficiently combining the advantages of the

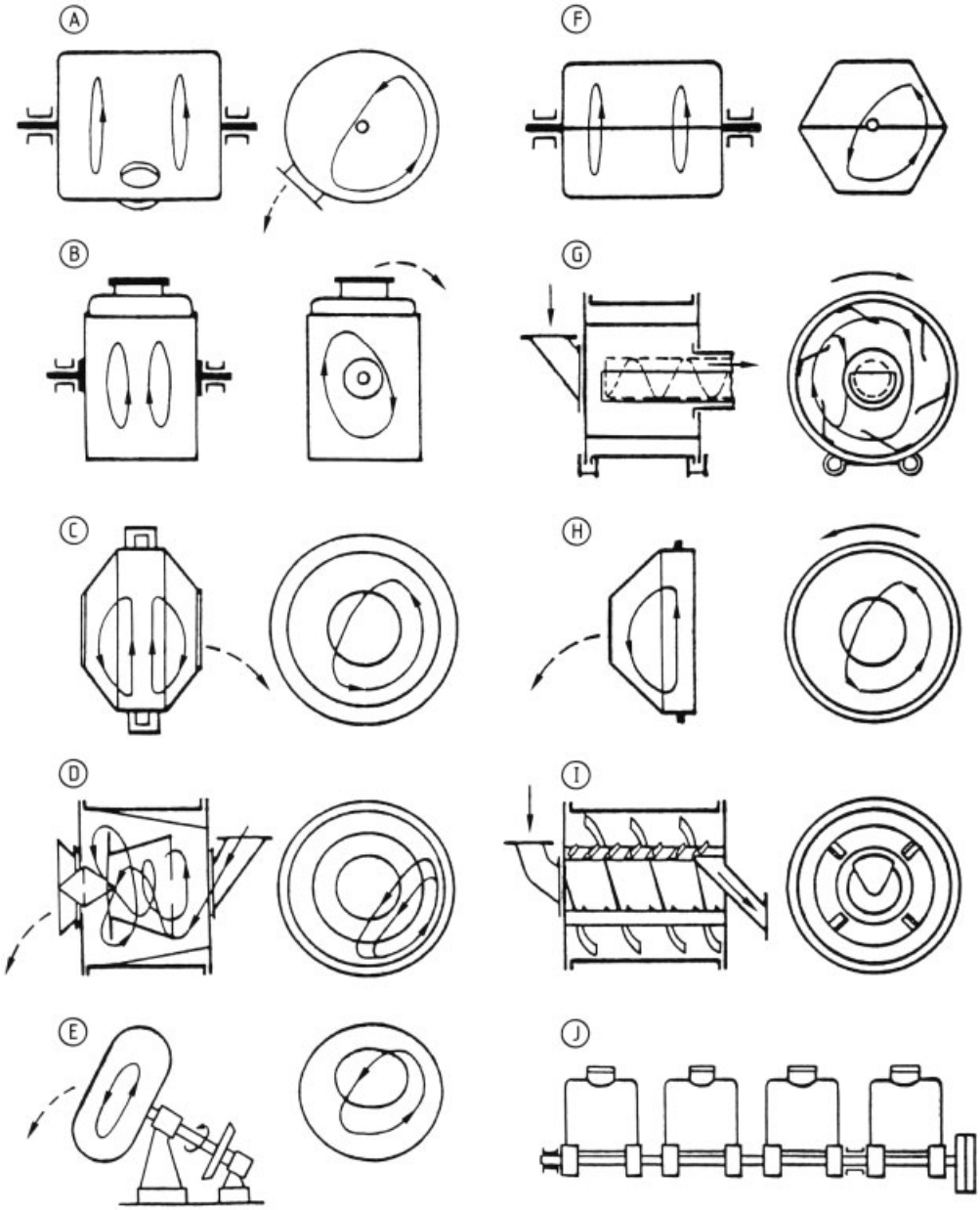

Figure 18. Tumbler mixers
A) Cylindrical drum; B) Barrel mixer; C) Cone mixer; D) Drum with adjustable baffles; E) Pear-shaped drum; F) Hexagonal drum; G) Drum with screw discharge; H) Single-cone mixer; I) Drum with spiral flights; J) Laboratory device for mixing vessels

drum mixer and the paddle mixer. The solid material moves in continually overlapping trajectories and is thus comminuted, mulled and, if necessary, moistened.

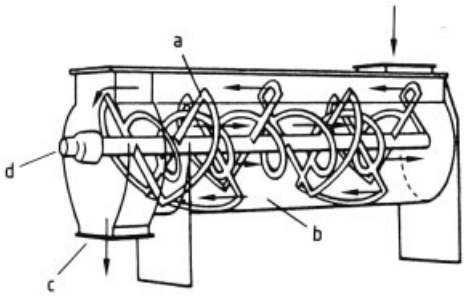

Figure 19. Ribbon mixer (Drais)
a) Helical ribbon; b) Vessel; c) Discharge; d) Drive

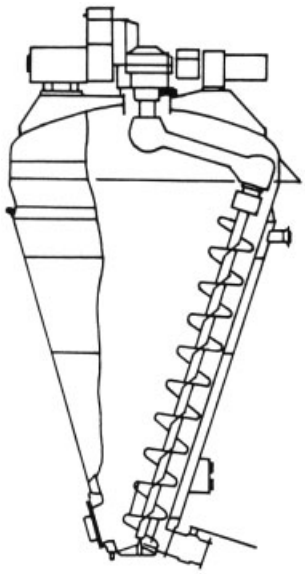

Figure 20. Hosokawa – Nauta mixer

3.4. Pneumatic Mixers

In pneumatic mixers, the mixing action is due to the injection of air. The simplest mixer of this type is the fluid-bed mixer.

The material bed is located above a horizontal or slightly inclined perforated plate. Air flows upward through the plate and bed until the gas velocity exceeds the value required to loosen and expand the bed. Gas bubbles form and carry the particles along as they rise. At the surface, the bubbles burst, distributing the particles in the air space above the fluidized bed.

Mixers now on the market differ in the air delivery pattern, the bottom plate being divided into quadrants or segments. As mixing proceeds, one quadrant or segment after another is supplied with extra air, and in this way the bed material is caused to circulate.

Fluid-bed mixers can be used only when the individual components are not too different in sinking velocity, since otherwise segregation will occur. Another class of pneumatic mixers includes air-jet mixers of free-flowing solids. These devices consist of vessels into which gas is injected through one or more nozzles at velocities up to supersonic (Fig. 22 A). The air-jet mixer has the advantage that the mixing time is short and there is not much danger of segregation. It has the disadvantages that a high-pressure compressor must be used and that the particles are subjected to high loads in the driving jet.

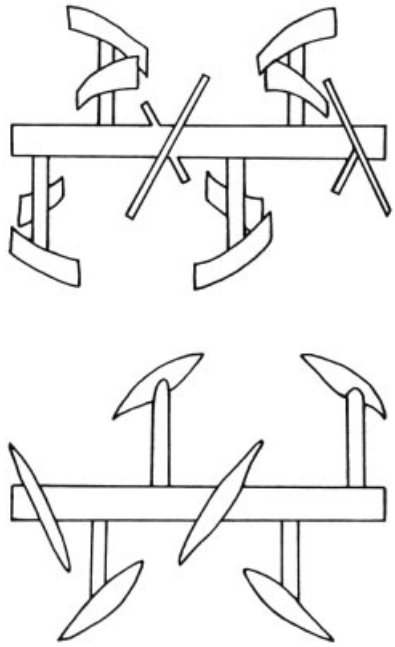

Figure 21. Paddle shafts

Much gentler in operation are rotating mixers (Fig. 22 B) with internal or external rotation. These are often used to homogenize granular resins. Air is blown into an ascending pipe and conveys the granular material upward to a sep-

arating zone. Material is admitted to the central ascender through, for example, mixing slots in various regions of the bed, so that compulsory mixing takes place.

3.5. Silo Mixers

The principle of the pneumatic rotating mixer, where material portions from various regions of the bed are brought together, is used so that large silos can be used as continuous mixers, especially for large-tonnage products. The gathering of material from various parts of the silo can be monitored with, for example, concentric floor discharge openings at various heights.

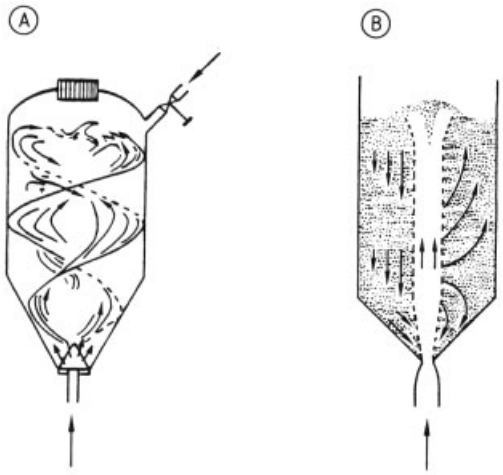

Figure 22. Pneumatic mixers
A) Airmix-system fluid-bed mixer; B) Rotating mixer

3.6. Pile Mixing

When the quantities to be mixed are so large that they can no longer be stored in a single silo, mixing must be done by appropriate piling and reclaiming techniques. This form of solid – solid mixing is important for the ore and beneficiation industries [45].

4. References

1. H. B. Ries, *Aufbereit. Tech.* **20** (1979) no. 1, 1 – 25, no. 2, 78 – 90.
2. L. T. Fan, S. J. Chen, C. A. Watson, *Ind. Eng. Chem.* **62** (1970) no. 7, 53 – 69.
3. H. B. Ries, *Aufbereit. Tech.* **17** (1976) no. 1, 16 – 27.
4. H. B. Ries, *Int. Chem. Eng.* **18** (1978) no. 3, 426 – 442.
5. L. T. Fan, *Powder Technol.* **24** (1979) 73 – 89.
6. Y. Akao, *Powder Technol.* **15** (1976) 207 – 214.
7. Y. Akao, *Powder Technol.* **21** (1976) 267 – 277.
8. H. Shindo, *Powder Technol.* **21** (1978) 105 – 111.
9. L. T. Fan, *Powder Technol.* **22** (1979) 205 – 213.
10. J. R. Too, *Powder Technol.* **23** (1979) 99 – 113.
11. S. H. Shin, L. T. Fan, *Powder Technol.* **19** (1978) 137 –146.
12. R. H. Wang, *Powder Technol.* **21** (1978) 171 – 182.
13. T. Yano, Y. Sano, *Kagaku Kogaku Abr. Ed. Engl.* **3** (1965) no. 2, 199 – 203.
14. J. Boss, *Bulk Solids Handling* **6** (1986) no. 6, 1207 –1215.
15. K. Sommer, *Aufbereit. Tech.* **23** (1985) no. 5, 266 –269.
16. K. Sommer: *Sampling of Powders and Bulk Materials,* Springer-Verlag, Heidelberg 1986.
17. K. Sommer: "Solid-Solid Mixing," *Fortschr. Verfahrenstech.* **19** (1981) 189 – 208.
18. W. Müller, *Untersuchungen über Mischzeit, Mischgüte und Arbeitsbedarf in Mischtrommeln mit rotierenden Mischelementen,* Dissertation, Karlsruhe 1966.
19. I. S. Gradshteyn, I. M. Ryzhik: *Table of Integrals, Series and Products,* Academic Press, New York-London 1965.
20. G. Schmahl, *VDI-Forschungsh.* **533** (1969).
21. M. B. Donald, B. Roseman, *Br. Chem. Eng.* **7** (1962) 749 – 753.
22. B. Roseman, M. B. Donald, *Br. Chem. Eng.* **7** (1962) 823 – 827.
23. M. B. Donald, B. Roseman, *Br. Chem. Eng.* **7** (1962) 922 – 923.
24. Fr. Müller, *Aufbereit. Tech.* **7** (1966) 274 – 285.
25. M. Ullrich, *Chem. Ing. Tech.* **41** (1969) 903 – 907.
26. H. Mattke, *Keram. Z.* **23** (1971) 282 – 286.
27. J. C. Williams, M. J. Kahn, *Chem. Eng. (N.Y.)* **80** (1973) 19 – 25.
28. R. N. Rowe, A. W. Nierrow, *Powder Technol.* **15** (1976) 141 – 147.
29. A. D. Fokker, *Ann. Physik (Leipzig)* **43** (1914) 810 – 820.

30. M. Planck, *Sonderber. preuß-Akad. Wiss. physik math. Kl.* (1917) 324 – 341.
31. D. S. Cahn, D. W. Fürstenau, *Powder Technol.* **1** (1967) 174 – 182.
32. D. S. Cahn, D. W. Fürstenau, *Powder Technol.* **2** (1968/69) 215 – 222.
33. N. Hoffmann, K. Schönert, *Aufbereit. Tech.* **12** (1971) 513 – 518.
34. A. Z. M. Abouzied, D.-W. Fürstenau, *Powder Technol.* **23** (1979) 261 – 271.
35. M. C. Coelho, N. Harnby, *Powder Technol.* **23** (1979) 209 – 217.
36. *Analyse der Metalle, Probenahme,* vol. 3, Springer Verlag, Berlin – Heidelberg – New York 1975.
37. L. T. Fan, *Powder Technol.* **12** (1975) 139 – 156.
38. R. Köhling, *Aufbereit. Tech.* **12** (1972) 689 – 704.
39. E. Block, *Glückauf* **101** (1965) no. 4, 255 – 264.
40. M. Hilbig, *Aufbereit. Tech.* **11** (1972) 705 – 712.
41. H. Schubert: *Mechanische Verfahrenstechnik II,* VEB-Verlag, Leipzig 1979.
42. Z. Sterbacek, P. Tausk: *Mixing in the Chemical Industry,* Pergamon Press, Oxford 1965.
43. M. Pahl: *Mischen beim Herstellen und Verarbeiten von Kunststoffen,* VDI-Verlag, Düsseldorf 1986.
44. W. Entrop: "Scaling up Solid-Solid Mixers," *Int. Symp. Mixing,* Mons 1978.
45. Ch. G. Schofield: *Homogenisation/Blending System Design and Central of Minerals Processing,* Trans. Tech. Publications, Clausthal 1980.

Modacrylic Fibers → Fibers, 4. Synthetic Organic

Separation Processes, Introduction

C. Judson King, University of California, Berkeley CA 94 720, United States

1. **Introduction** 1
2. **Definitions** 2
3. **Process Configurations** 2
4. **Separation Principles** 4
5. **Advantages and Disadvantages of Different Separation Methods** 4
6. **Selection of a Separation Process** 6
7. **Factors Influencing Phase Equilibria** . . 6
8. **Sources of Equilibrium Data** 7
9. **References** 7

1. Introduction

Separation processes are used to convert feed mixtures into two or more products differing in composition. Unit operations treated in the first part of this volume deal primarily with *diffusional separation processes* suitable for feeds that are homogeneous mixtures:

Evaporation
Distillation and Rectification
Sublimation
Liquid – Liquid Extraction
Absorption
Adsorption
Process-Scale Chromatography
Biochemical Separations (mainly diffusional processes)

These contrast with *mechanical separation* of heterogeneous mixtures, in which the product phases are already present on a microscale. These are discussed in the following articles:

Filtration
Centrifugation, Filtering
Centrifugation, Sedimenting
Hydrocyclones
Sedimentation
Dust Separation
Screening
Elutriation
Air Classification
Mineral Sorting
Magnetic Separation
Electrostatic Separation
Gravity Concentration
Dense-Media Separation
Flotation

Separation processes constitute more than half of the total equipment investment for the chemical and fuel industries. They are also widely used in pharmaceutical and food industries, in beneficiation of mineral ores and recovery of metals, in processing effluents, and in a diverse array of other industries. Separation processes may have a number of purposes, which can be loosely categorized as follows:

1) *Purification:* Removal of impurities, thereby enabling a desirable substance to be obtained at a higher level of purity; e.g., refining of sugar and treatment of drinking water.
2) *Concentration or Recovery:* Increasing the concentration of a desired substance in solution, usually by removal of a substantial fraction of solvent; e.g., production of fruit juice concentrates and recovery of metal values from effluents.
3) *Fractionation:* Separation of desired substances from one another; e.g., primary distillation of crude oil and chromatographic separations.

The first two categories imply separation of a small amount of one substance from a large amount of another. The desired products are the major and minor constituents, respectively, in the first and second processes.

2. Definitions

Most separation processes are based on the principle that different phases of matter have different compositions at equilibrium. Examples are vapor and liquid phases in distillation, and

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_01

immiscible liquid phases in liquid – liquid extraction. These processes achieve separation by allowing the phases to proceed toward equilibrium, hence the name *equilibration processes.*

Another category, known as *rate-governed processes*, achieves separation by means of differences in rates of transport of different species through a medium or barrier. Here the products are probably fully miscible with the feed; the separation would be destroyed by remixing them. Examples are ultrafiltration and gaseous diffusion.

A *separating agent* is added to achieve separation. In equilibration processes, this agent serves to create the second phase. The separating agent takes the form of energy or matter. Examples of *energy separating agents* are reboiler heat, which forms the vapor in distillation, and chilling or refrigeration, which causes ice to form in freeze concentration processes. Examples of *mass separating agents* are an absorbent used to separate gases and an ion-exchange resin used to desalt water. Typically, most of the utilities costs for a separation process are associated with the separating agent, e.g., the cost of steam to drive the reboiler of a distillation column or the cost of regenerating a mass separating agent for reuse. Economics and environmental restrictions usually dictate that a mass separating agent should be regenerated and reused. A separation process based upon a mass separating agent thereby requires a second, subsequent separation process.

The degree of separation achieved between two substances in different products relates to the *separation factor* between them. For equilibration processes, the separation factor α_{ij} is usually defined as the ratio of the partition coefficients (K_i and K_j) of the two species between the two phases at equilibrium ($\alpha_{ij} = K_i/K_j$). The constants K_i and K_j are the *partition coefficients* of species *i* and *j*, which are defined as the concentration of a particular species in phase 2 divided by that in phase 1. Any concentration units may be used for K_i and K_j in either phase, provided the same units are used for both species in a given phase. Phases 1 and 2 are usually chosen so that α_{ij} is greater than unity.

3. Process Configurations

Often, the species being separated is separated to an insufficient degree by a simple equilibration or by a single passage through a barrier. In these cases, the degree of separation between species can be improved by *staging* or *countercurrency*. Examples of each are shown in Figure 1. By sending the product from one contacting or stage to another and appropriately recycling intermediate products to previous stages (Fig. 1 A), the degree of separation between species in the ultimate products can be improved considerably. For a distillation operated at total reflux (infinitesimal feed and product flow), the separation factor for products from a single equilibrium stage is α_{ij}, as already noted, whereas that from *N* equilibrium stages joined countercurrently is α_{ij}^{N} (the Fenske equation) (→ Distillation and Rectification).

Continuous countercurrent contacting can be used to accomplish the same effect, as shown in Figure 1 B. Here, two phases flow countercurrent to each other through an open material such as a structured packing, which serves to create interface for mass transfer between phases. The action is analogous to that of a countercurrent heat exchanger, where the exit temperature of the hot stream can be lower than the exit temperature of the cold stream. Sometimes contactors are built with elements of both discretely staged and continuous countercurrent contacting, for example, rotating disk contactors used for liquid – liquid extraction (→ Liquid – Liquid Extraction).

Another way of improving the degree of separation achievable in a simple contacting is the *chromatographic mode.* Here a mobile phase flows along a thin, stationary phase (→ Process-Scale Chromatography).

The thinness of the stationary phase provides rapid rates of mass transfer, giving the effect of many stages or "transfer units" per unit length of the stationary phase. In elution chromatography, components of a mixture injected as a pulse at the inlet end of the contactor proceed along the stationary phase at different rates, determined by the different equilibrium distributions between phases, and emanate as a succession of isolated peaks (Fig. 2). Field-flow fractionations are rate-governed processes operated in an analogous, elution-chromatographic mode.

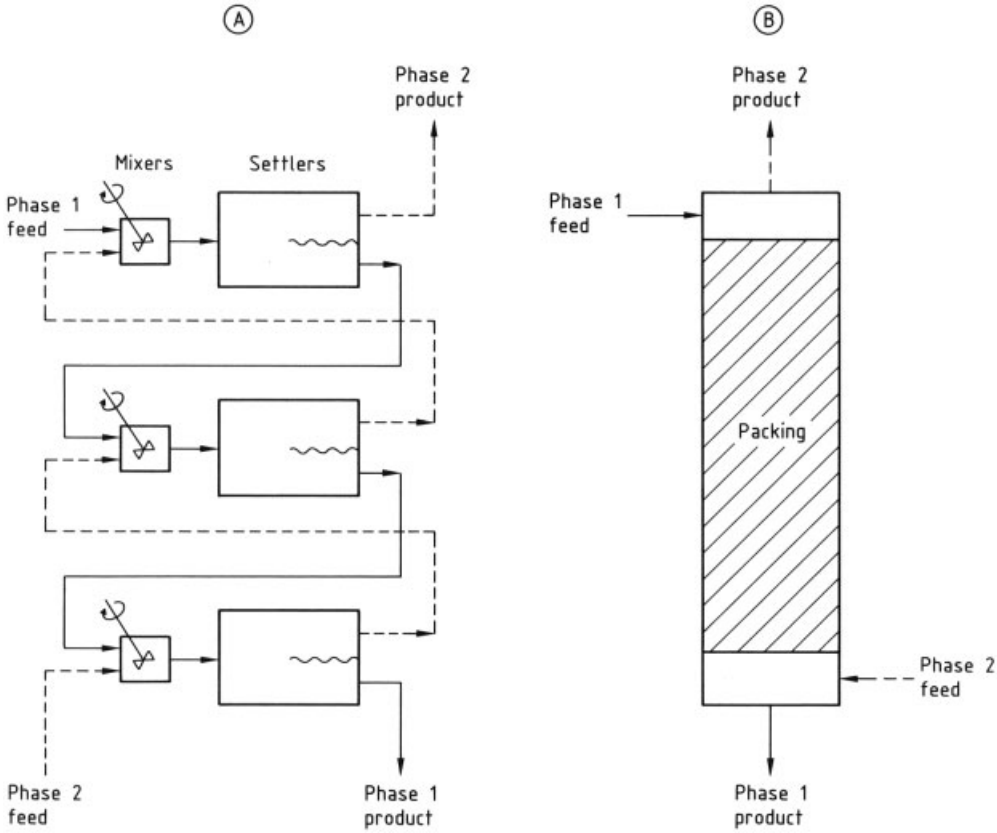

Figure 1. Staging and countercurrency
A) Staged mixer – settler system for solvent extraction; B) Continuous countercurrent extractor.

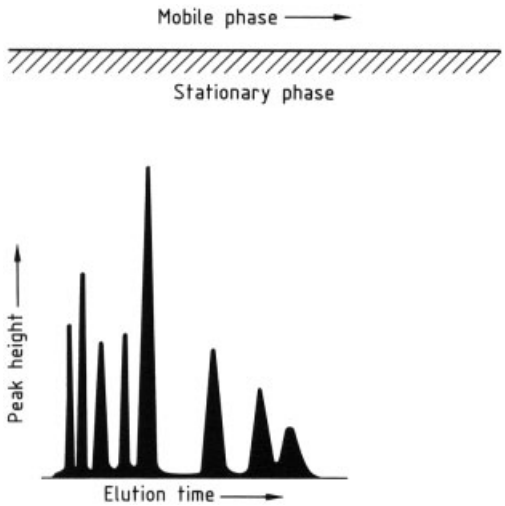

Figure 2. Typical output from a gas or liquid chromatograph (peak height, as sensed by any of various detection methods, vs. elution time).

A continuous staged or countercurrent contactor provides high feed capacity but generates only two products of high purity. (Intermediate products, or sidestreams, of lesser purity can also be present.) The elution-chromatography mode can separate a complex mixture into many pure peaks or products, but its inherent capacity is low because the feed is injected intermittently as pulses.

When a solid feed is used (e.g., leaching of roast and ground coffee to form coffee extract; → Liquid – Solid Extraction) or a solid mass separating agent is employed (e.g., an adsorbent; → Adsorption), movement of the solids should generally be avoided because attrition and undesirable mixing may result. *Fixed-bed processes* are commonly used (e.g., the home water softener filled with ion-exchange resin or a bed of activated alumina for drying air) when a solid feed or mass separating agent is used. A regenerated fixed-bed separation can handle a sub-

stantial feed rate but in its simplest form provides only one highly purified product, the initial effluent, before a second component leaves the bed. Among the more notable innovations in industrial separation during the past few decades have been effective ways of simulating countercurrency with fixed beds.

Yet another fundamentally different approach to separation involves migration of different species to equilibrium positions within a gradient field. One example is isoelectric focusing. These processes are inherently slow but can give separations that are not possible by other means.

4. Separation Principles

For a feed of any particlular phase condition, an equilibration separation process can be based upon formation of, or contact with, any immiscible second phase of matter. For a liquid feed, the second phase may be a gas (stripping), an immiscible liquid (extraction), or a solid (crystallization, adsorption). Equilibration may be with the bulk of the second phase or with a surface. Surface-based separation processes include adsorption, as well as foam, bubble, and emulsion fractionation (→ Flotation).

For rate-governed processes, differences in any form of transport can be exploited. These include rates of permeation through a solid membrane (→ Membranes and Membrane Separations), Knudsen diffusion in a porous medium (as used for separation of uranium isotopes in UF_6; → Isotopes, Natural), thermal diffusion, electrophoresis, or pressure diffusion.

Often, two separation principles used in concert can operate synergistically. Examples are the use of cross fields for rate-governed separation and the enhancement of relative volatility in a distillation by adding a substance that modifies the equilibrium between phases, as done in azeotropic and extractive distillation (→ Distillation and Rectification).

5. Advantages and Disadvantages of Different Separation Methods

When the same separation factors can be attained, *equilibration processes* using energy separating agents tend to be less costly than those using a mass separating agent, because of the need for circulating and regenerating the mass separating agent or else disposing of and replacing it. Similarly, when staging or countercurrency is needed, equilibration processes tend to be more attractive economically than *rate-governed* processes that give an equivalent separation factor, because equipment and energy can be utilized more efficiently in the former.

Separation processes involving a solid phase can suffer drawbacks associated with low or vanishingly small transport rates in the solid phase and with the desirability of keeping the solid phase stationary.

Certain methods of separation are better suited for certain ranges of feed concentration. In particular, *fixed-bed operations* are most effective for removing relatively dilute solutes, since the diluteness necessitates less bed volume and less frequent regeneration. Thus, adsorption and ion-exchange processes tend to be used to recover solutes from relatively dilute feeds or to remove impurities. Along this spectrum of feed concentration, extraction (with liquids) and absorption (with gases) are usually considered for a middle range of concentration, because the solvent flow rate required tends to be relatively independent of the feed solute concentration. Operational upper limits on feed solute concentration can come from the need to keep liquid phases immiscible in extraction or to avoid too large a percentage decrease of the flow rate of the feed phase. *Distillation* works well over a wide range of feed concentrations but can experience low stage efficiencies at very low solute concentrations.

Membrane separation can provide high selectivity for removing or concentrating (ultrafiltration) high molecular mass or macromolecular solutes, for removing salts from water or concentrating them (reverse osmosis, electrodialysis), and for fractionating solutes of high and low molecular mass (dialysis); (for further information, see → Membranes and Membrane Separations). Economical, high-capacity membranes that effectively remove polar organic solutes of lower molecular mass have yet to be developed, but are an area of research. The solute throughput capacity and selectivity provided by membranes can be enhanced by impregnation of the membrane with an appropriate extractant (solid-

supported liquid membranes, facilitated transport). Membrane processes are best suited for low molar solute concentrations, because providing the driving force for transport across the membrane is otherwise difficult and expensive.

Cost and Scaleup. Different methods of separation have inherently different costs. An inverse relationship tends to exist between the value of the product and the scale of production. For high-value substances produced on a small scale, a much wider range of separation techniques can be considered than for substances of lower value. Often the cost or value of a substance is influenced strongly by the difficulty of separation; therefore, for a high-value substance, utilization of a newer or less common means of separation may be necessary in order to perform the desired separation at all. Furthermore, certain separation processes (e.g., distillation, extraction) can be scaled up more readily than others. Those methods that rely on very thin phases or thin flow channels (chromatography), laminar flow (field-flow processes), or ready dissipation of heat (electrophoresis) are particularly difficult to scale up to large capacities.

Thermosensitive Product. In many cases, the feed or products are sensitive to thermal degradation, contamination, or changes due to a change in the chemical environment, such as denaturation of proteins. These constraints are particularly common in the food and pharmaceutical industries. Here, methods of separation that avoid these problems will have an advantage. To avoid thermal degradation, low temperatures and short residence times should be used. Processes that avoid heating the feed (extraction, sorption, crystallization, etc.) are advantageous; if vaporization must be used, operation under vacuum is helpful. Concerns about contamination and chemical environment limit the types of mass separating agents that can be used. For instance, in bioprocessing, many solvents have toxic effects, so precipitation and the use of solid sorbents are relatively advantageous. Similarly, proteins may denature if they are taken into a nonaqueous phase.

Processes Involving Reversible Chemical Reactions. Most industrial separations are carried out by processes that do not involve chemical reactions, because of the consumption of reagents required to accomplish the reaction, the cost of reagents needed to regenerate the original desired substance, and the need to dispose of unwanted reaction products. However, chemical complexation or association reactions (donor–acceptor, chelation, clathration, etc.) are much more readily reversible and can be used advantageously in separations to increase the selectivity among solutes or the capacity for a desired solute. Reversible chemical interactions can be implemented in extraction; sorption; ion exchange; azeotropic and extractive distillation; impregnated membranes; and foam, bubble, and emulsion fractionation. Finally, the common processes for concentrating a solute in solution by removal of solvent (evaporation, reverse osmosis, ultrafiltration, freeze concentration) incur expense in proportion to the amount of solvent that must be removed. Extraction or sorption of the solute(s) of interest can lead to concentration as well, because only a limited amount of the solvent will accompany the solute into or onto the extractant or sorbent. Reversible chemical interactions can be used effectively here as well.

6. Selection of a Separation Process

In seeking one or more appropriate methods of separating a particular mixture, the first consideration is the size of the *separation factor* likely to be attained by different methods of separation. Differences in volatility (distillation, evaporation, stripping), solubility (crystallization, extraction, absorption), charge-to-mass ratio (ion exchange, electrophoresis), molecular size and shape (adsorption with molecular sieves, crystallization, gel permeation, clathration, dialysis), and chemical reactivity can all be used for separations.

Next, whether *ordinary or extreme conditions* (very high or low pressure, very high or low temperature, etc.) are needed to obtain attractive separation factors must be determined. Methods that require excessive temperature–time combinations or result in contamination or chemical change may be ruled out by the nature of the feed or product. Also, as already noted, the *value of the substance* and the desired scale of operation can determine the number and types of alternatives to be considered, because of the varying cost of different separation methods and their suitability for scaleup. Finally, a very practical

consideration is the amount of previous *experience* with a particular process.

Among near equals, distillation has an advantage because it avoids solids, it is easy to stage and scale up, and a vast backlog of experience exists. Rate-governed processes are usually considered seriously only when the desired separation can be achieved in a single stage. Separation of uranium isotopes by gaseous diffusion is a notable exception to this generalization, however. With processes utilizing a mass separating agent, *ease of regeneration* becomes a dominant factor. Finally, as already noted, fixed-bed processes gain a compensating advantage when substances at relatively low concentration are to be removed.

7. Factors Influencing Phase Equilibria

Several general factors influence phase equilibria and can be utilized in the logic of selecting processes and choosing mass separating agents. A generalized equilibration separation process may be regarded as having a source (feed) phase and a second, receiving phase. For the solute(s) of interest to move in the desired direction, the chemical potential of the solute in the receiving phase must be less than that in the source phase.

The partition coefficient K_i for a solute between two bulk phases is independent of solute concentration when ideal solutions exist in both phases or, in most instances, when the solute is very dilute in both phases. The partition coefficient is simply the ratio of the activity coefficient in the source phase to that in the receiving phase, if the activity coefficients and the partition coefficient are based upon the same concentration units (moles or mass per volume, weight fraction, etc.). If *preferential interactions* occur between solute and solvent [defined as the major component(s)] in either phase, the distribution of the component toward that phase will be enhanced. Similarly, if less affinity exists between solute and solvent molecules than between solvent molecules themselves, the solute will be driven toward the other phase.

The values of K_i for many phase-distribution processes are dominated by the *nonidealities* in one of the two phases. This is particularly true for processes that remove nonpolar or low-polarity solutes from aqueous solution in which the aqueous phase is highly nonideal (very large activity coefficients) and the receiving phase is much more nearly ideal. Thus relationships among many organic compounds based upon a single set of chemical parameters related to dispersion forces, polarity and polarizability, and acidity and basicity, serve to correlate such seemingly diverse partitioning properties as solubility in water [2], adsorption from water onto activated carbon [3], and extraction from water into 1-octanol [4]. All of these situations are dominated by aqueous-phase nonidealities.

Dimerization or polymerization of the solute, or formation of solute molecules into micelles in one of the phases, serves to increase the distribution (i.e., increase K_i) of the solute in that phase at higher solute concentrations.

Another common equilibrium feature is the *saturation* of a phase. In adsorption processes, the finite amount of solid surface area present provides an upper limit on the amount of solute that can be taken up within the capacity of the selective layer adjacent to the surface. The same is true of gas – liquid surface capacity in foam or bubble fractionation. In these cases, the partition coefficient toward the surface phase decreases as saturation is approached. This is another reason why surface-based separation processes are more useful when the solute to be removed is present in low concentration.

A similar situation exists where *chemical reactions* are involved (e.g., ion exchange, reversible chemical complexation). The reactant in the receiving phase has a limited capacity, in that the stoichiometry of the underlying reaction(s) cannot be exceeded. Thus, a solvent phase containing a reactive organic extractant of molecular mass 400 at 40 wt % in an organic diluent can at most take up 5 wt % of a solute of molecular mass 50, with which the extractant forms a 1:1 complex. Complexation of additional solute molecules with a reactant molecule engenders additional uptake.

8. Sources of Equilibrium Data

Many measurements of phase-equilibrium properties exist, but they are widely scattered in the literature. For a particular system, a search of *Chemical Abstracts* may be the most useful

approach. *Complications of data* are given in [5–8] for vapor – liquid equilibria; [9] for liquid–liquid equilibria; [10] and [11] for solubilities of gases and solids in liquids; [12] and [13] for bulk solid – liquid equilibria; and [14] for phase diagrams in general.

Methods for prediction, correlation, and interpretation of vapor–liquid and liquid – liquid phase equilibria are treated in [15] and summarized in [16]. Prominent among these is the UNIFAC method, described for vapor – liquid equilibria in [17] and for liquid – liquid equilibria, in a more general context, in [18] and [19]. This method is suitable for mixtures with moderate degrees of nonideality but cannot yet be extended to situations involving simultaneous strong, specific chemical interactions. For further information, see → Estimation of Physical Properties.

9. References

General References

1. C. J. King: *Separation Processes,* 2nd ed., McGraw-Hill, New York 1980. R. W. Rousseau (ed.): *Handbook of Separation Process Technology,* Wiley, New York 1987. P. A. Schweitzer (ed.): *Handbook of Separation Techniques for Chemical Engineers,* McGraw-Hill, New York 1979.

Specific References

2. R. W. Taft, M. H. Abraham, R. M. Doherty, M. J. Kamlet, *Nature (London)* **313** (1985) 384.
3. M. J. Kamlet, R. M. Doherty, M. H. Abraham, R. W. Taft, *Carbon* **23** (1985) 549.
4. R. W. Taft, M. H. Abraham, G. R. Famini, R. M. Doherty, et al., *J. Pharm. Sci.* **74** (1985) 807.
5. I. Wichterle, J. Linek, E. Hala (eds.): *Vapor-Liquid Equilibrium Data Bibliography,* Elsevier, Amsterdam 1973 Suppl. 1, 1976. Suppl. 2, 1979. Suppl. 3, 1982. Suppl. 4, 1985.
6. M. Hirata, S. Ohe, K. Nagahama: *Computer Aided Book of Vapor-Liquid Equilibria,* Elsevier, Amsterdam 1975.
7. J. Gmehling, U. Onken, et. al.: *Vapor-Liquid Equilibrium Data Collection,* **vol. 1, 1 a, 2abcd, 3, 4, 5, 6abc,7, 8,** Dechema, Frankfurt 1977 ff.
8. H. Knapp, R. Döring, L. Oellrich, U. Plöckner, J. M. Prausnitz: *Vapor-Liquid Equilibria for Mixtures of Low-Boiling Substances,* Chemistry Data Ser., vol. 6, Dechema, Frankfurt 1982.
9. J. Wisniak, A. Tamir: *Liquid-Liquid Equilibrium and Extraction,* 2 vols., Elsevier, Amsterdam 1980 – 1981 Suppl. 1, 1985.
10. J. Wisniak, M. Herskowitz: *Solubility of Gases and Solids,* **2 vols.,** Elsevier, Amsterdam 1984.
11. A. Seidell, W. F. Linke: *Solubilities of Inorganic and Metal-Organic Compounds,* Van Nostrand, New York 1958.
12. J. Timmermanns: *The Physico-chemical Constants of Binary Systems,* Interscience, New York 1959 – 1960.
13. H. Stephen, T. Stephen (eds.): *Solubilities of Inorganic and Organic Compounds,* Pergamon, New York 1964.
14. J. Wisniak: *Phase Diagrams,* 2 vols., Elsevier, Amsterdam 1981.
15. J. M. Prausnitz, R. N. Lichtenthaler, E. Gomes de Azevedo: *Molecular Thermodynamics of Fluid-Phase Equilibria,* 2nd ed., Prentice-Hall, Englewood Cliffs NJ 1986.
16. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *Properties of Gases and Liquids,* 3rd ed., McGraw-Hill, New York 1977.
17. A. Fredenslund, J. Gmehling, P. Rasmussen: *Vapor-Liquid Equilibria Using UNIFAC,* Elsevier, Amsterdam 1977.
18. J. M. Sørensen, T. Magnussen, P. Rasmussen, A. Fredenslund, *Fluid Phase Equilib.* **2** (1979) 297; **3** (1979) 47; **4** (1980) 151.
19. T. Magnussen, P. Rasmussen, A. Fredenslund, *Ind. Eng. Chem. Process Des. Dev.* **20** (1981) 331.

Sera → Immunotherapy and Vaccines
Serine → Amino Acids
Sewer Pipes → Construction Ceramics
Sex Hormones → Hormones
Shampoos → Hair Preparations
Shaving Preparations → Hair Preparations
Shellac → Resins, Natural
Shielding Gas → Protective Gases

Heat Exchange

RAMESH K. SHAH, Harrison Radiator Division, General Motors Corporation, Lockport, New York 14094, United States

ALFRED C. MUELLER, Church Hill, Maryland 21623, United States

1. Introduction 6
2. Classification of Heat Exchangers 7
2.1. Classification According to Construction 7
2.1.1. Tubular Heat Exchangers 7
2.1.1.1. Shell-and-Tube Exchangers 7
2.1.1.2. Double-Pipe Heat Exchangers . . . 9
2.1.1.3. Drip-Type Heat Exchangers 9
2.1.2. Plate-Type Heat Exchangers 9
2.1.2.1. Plate Heat Exchangers 12
2.1.2.2. Spiral-Plate Heat Exchangers 14
2.1.3. Extended-Surface Heat Exchangers 15
2.1.3.1. Plate – Fin Heat Exchangers 15
2.1.3.2. Tube – Fin Heat Exchangers 17
2.1.4. Regenerators 19
2.2. Classification According to Process Function 19
2.2.1. Condensers 20
2.2.2. Liquid-to-Vapor Phase-Change Exchangers 23
2.2.2.1. Chemical Evaporators 23
2.2.2.2. Reboilers 26
2.2.2.3. Waste Heat Boilers 28
3. Heat Exchanger Design Methodology 30
4. Criteria for Heat Exchanger Selection 30
4.1. General Selection Criteria 30
4.1.1. Design Variables 30
4.1.2. General Selection Guidelines 32
4.2. Shell-and-Tube Exchangers 33
4.2.1. Tubes 33
4.2.2. Tube Pitch and Layout 34
4.2.3. Baffles 34
4.2.4. Shells 38
4.2.5. Front-End Heads 40
4.2.6. Rear-End Heads 40
4.2.7. Nozzles 42
5. Basic Heat-Transfer and Pressure-Drop Analysis 42
5.1. Basic Equations and Definitions of Heat-Transfer Analysis 42
5.2. Dimensionless Methods for Exchanger Heat-Transfer Analysis 45
5.3. Extensions of the Basic Heat-Transfer Theory 47
5.3.1. Longitudinal Wall Heat Conduction 47
5.3.2. Variable Overall Heat-Transfer Coefficients 56
5.4. Exchanger Pressure-Drop Analysis 57
5.4.1. Plate – Fin Exchangers 57
5.4.2. Tube – Fin Exchangers 58
5.4.3. Regenerators 59
5.4.4. Plate Heat Exchangers 59
6. Single-Phase Heat-Transfer and Pressure-Drop Correlations 59
6.1. Basic Concepts and Dimensionless Groups 59
6.2. Theoretical Solutions and Correlations for Simple Geometries . . 60
6.2.1. Fully Developed Flows 60
6.2.2. Hydrodynamically Developing Flows 64
6.2.3. Thermally Developing Flows 65
6.2.4. Simultaneously Developing Flows . 66
6.3. Empirical Correlations for Complex Geometries 66
6.3.1. Tubular Exchangers 67
6.3.2. Plate Exchangers 70
6.3.3. Extended Surface Exchangers 70
7. Two-Phase Heat-Transfer and Pressure-Drop Correlations 71
7.1. Flow Patterns 71
7.2. Two-Phase Pressure-Drop Correlations 74
7.2.1. Intube Pressure Drop 74
7.2.2. Shellside (Tube Bundle) Pressure Drop 76
7.2.3. Other Geometries 76
7.3. Heat-Transfer Correlations for Condensation 76
7.3.1. Condensation Inside a Horizontal Tube 77
7.3.2. Condensation Inside a Vertical Tube 78
7.3.3. Condensation Outside Horizontal Tube Bundles 79
7.3.4. Condensation over Finned Tubes . . 80
7.3.5. Condensation Outside Vertical Tube Bundles 80

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_02

7.3.6. Condensation of Mixtures 80
7.4. Heat-Transfer Correlations for Boiling and Evaporation 81
7.4.1. Intube Forced Convective Boiling . 81
7.4.2. Intube Critical Heat Flux 83
7.4.3. Shellside Forced Convective Boiling 84
7.4.4. Shellside Critical Heat Flux 84
8. Thermal Design for Single-Phase Heat Exchangers 84
8.1. Introduction 85
8.2. Shell-and-Tube Exchangers 85
8.2.1. Segmental Baffles 85
8.2.1.1. Shellside Heat Transfer 86
8.2.1.2. Shellside Pressure Drop 87
8.2.2. Disk-and-Doughnut Baffles 88
8.2.3. Rod Baffles 88
8.2.4. Design Procedure 88
9. Thermal Design for Multiphase Heat Exchangers 89
9.1. Condensers 89
9.1.1. Considerations in Condenser Design 89
9.1.2. Condenser Types and Their Characteristics 90
9.1.2.1. Shellside Condensers 90
9.1.2.2. Tubeside Condensers 91
9.1.2.3. Air-Cooled Condensers 93
9.1.3. Condensation of Mixtures 94
9.1.4. Mean Temperature Difference . . . 95
9.1.5. Desuperheating 95
9.1.6. Subcooling 95
9.2. Vaporizers 96
9.2.1. Kettle and Internal Reboiler Design 96
9.2.2. Horizontal Shellside Thermosiphon Reboiler Design 97
9.2.2.1. Pressure Balance in the Recirculation Loop 97
9.2.2.2. Shellside Heat Transfer and Pressure Drop 98
9.2.2.3. Design Methods and Special Considerations 98
9.2.3. Vertical Thermosiphon Reboiler Design 98
9.2.3.1. Pressure Balance in the Recirculation Loop 99
9.2.3.2. Flow Regimes and Heat Transfer . . 99
9.2.3.3. Approximate Design Methods . . . 100
9.2.4. Special Design Considerations . . . 101
9.2.5. Mean Temperature Difference . . . 101
9.2.6. Common Problems with Reboilers in Practical Applications 102
10. Flow-Induced Vibrations 103
10.1. Vibration Analysis 104
10.2. Design Guidelines for Vibration Prevention 104
10.2.1. Tube Vibration 104
10.2.2. Acoustic Vibration 105
11. Flow Maldistribution 105
11.1. Causes of Flow Maldistribution . 105
11.1.1. Mechanical Design, Basic Geometry, and Tolerances 105
11.1.2. Self-Induced Maldistribution 105
11.1.3. Mixed-Phase Flow 106
11.1.4. Fouling 106
11.2. Consequences of Maldistribution 106
11.2.1. Shell-and-Tube Exchangers 106
11.2.2. Other Exchangers 106
11.3. Recommendations for Shell-and-Tube Exchangers 107
12. Fouling 107
13. References 109

Symbols

A exchanger total heat-transfer area (primary plus secondary, if any) on one side, m^2 or ft^2

A_f fin or extended surface area on one side of the exchanger, m^2 or ft^2

A_{fr} frontal or face area of the exchanger, m^2 or ft^2

A_k total wall cross-sectional area for longitudinal conduction, m^2 or ft^2

A_o minimum free flow (or open) area, m^2 or ft^2

A_p primary surface area on one side of the exchanger, m^2 or ft^2

A_w total area for heat conduction from hot to cold fluid, m^2 or ft^2

B coefficient [Eq. (85)], dimensionless

b distance between two plates (fin height) in a plate-fin exchanger, m or ft

C flow stream heat capacity rate with a subscript c or h, $C = Wc_p$, W/K or Btu h^{-1} °F^{-1}

C^* heat capacity rate ratio $= C_{min}/C_{max}$, dimensionless

C_c flow-stream heat capacity rate of cold fluid $= (Wc_p)_c$, W/K or Btu h^{-1} °F^{-1}

C_h flow-stream heat capacity rate of hot fluid $= (Wc_p)_h$, W/K or Btu h^{-1} °F^{-1}

C_{max} maximum of C_c and $C_h = (W c_p)_{max}$, W/K or Btu h^{-1} °F^{-1}

C_{min} minimum of C_c and $C_h = (W c_p)_{min}$, W/K or Btu h^{-1} °F^{-1}

CHF critical heat flux

Co convection number [Eq. (127)], dimensionless

c constant, used in many equations with specific value at each place

c_p specific heat of fluid at constant pressure, J kg^{-1} K^{-1} or Btu lbm^{-1} °F^{-1}

D_e equivalent diameter, m or ft

D_h hydraulic diameter, m or ft

d_B tube bundle diameter $= D_i - \delta_{sb}$, where δ_{sb} is the shell-to-bundle diametral clearance, m or ft

d_e fin tip diameter of an individually finned tube, m or ft

d_i inside diameter of a circular tube, m or ft

d_o outside diameter of a circular tube, m or ft

F log-mean temperature difference correction factor, dimensionless

F_{fl} fluid dependent parameter, dimensionless

Fr Froude number $= G^2 g^{-1} d_i^{-1} \varrho_1^{-2}$, dimensionless

f Fanning friction factor $= \tau_w/(\varrho v^2/2 g_c)$, dimensionless

f' modified friction factor per tube row $= \Delta p/(4 N_r G^2/2 g_c \varrho)$, dimensionless

f_{app} apparent Fanning friction factor, defined in Table 8, dimensionless

G mass velocity based on minimum free flow area $= W/A_o$, kg m^{-2} s^{-1} or lbm h^{-1} ft^{-2}

Gr Grashof number, [Eq. (176)], dimensionless

g acceleration due to gravitation, m/s^2 or ft/s^2

g_c proportionality constant in Newton's second law of motion, $g_c = 1$ and is dimensionless in SI units, $g_c = 32.174$ lbm ft lbf^{-1} s^{-2}

(H1) thermal boundary condition referring to constant axial wall heat flux with constant peripheral wall temperature

(H2) thermal boundary condition referring to constant axial wall heat flux with constant peripheral wall heat flux

h heat-transfer coefficient, W m^{-2} K^{-1} or Btu h^{-1} ft^{-2} °F^{-1}

ID tube inside diameter, m or ft

i_{lg} latent heat of vaporization, J/kg or Btu/lbm

i'_{lg} $= i_{lg} + 3/8 c_{p,l} (T_{sat} - T_w)$, J/kg or Btu/lbm

J heat-transfer coefficient correction factors for shellside leakage and bypass flows, refer to specific definitions after Equation (144), dimensionless

j Colburn factor $= St\, Pr^{2/3}$, dimensionless

k thermal conductivity of the fluid if no subscript, W m^{-1} K^{-1} or Btu h^{-1} ft^{-1} °F^{-1}

k_f thermal conductivity of the fin material, W m^{-1} K^{-1} or Btu h^{-1} ft^{-2} °F^{-1}

k_w thermal conductivity of the wall material, W m^{-1} K^{-1} or Btu h^{-1} ft^{-1} °F^{-1}

L tube length or fluid flow (core) length on one side of the exchanger, m or ft

L_{hy} hydrodynamic entrance length, m or ft

LMTD log-mean temperature difference, °C

l_{ef} effective flow length between major boundary layer disturbances, distance between interruptions, m

l^*_{ef} $= l_{ef}/D_h Re\, Pr$

l_f fin height for individually finned tubes $= (d_e - d_o)/2$, m or ft

M molecular weight, M_v for vapor, M_m for mean gas phase, kg/kg mol or lbm/lbm mol

MTD mean temperature difference, °C

m fin parameter, defined in Table 4 for each fin geometry (m with this meaning used only in the fin analysis), m^{-1} or ft^{-1}

N number of passes

N_b number of baffles

N_f number of fins per unit length in a tube – fin exchanger, m^{-1} or ft^{-1}

N_p number of flow passage on one side in a plate – fin exchanger

N_r number of tube tows in the flow direction

N_t number of tubes

NTU number of heat-transfer units $= UA/C_{min}$, $NTU_1 = UA/C_1$, $NTU_2 = UA/C_2$, dimensionless

Nu Nusselt number $= h D_h/k$, dimensionless

n exponent, dimensionless

OD tube outside diameter, m or ft

P temperature effectiveness of the tube-side stream, defined in Table 5, dimensionless

$\tilde{P}$ fluid pumping power $= W \Delta p/\varrho$, W or hp

Pr Prandtl number $= \mu c_p/k$, dimensionless

p fluid static pressure, Pa or lbf/ft^2

p_v, p_g vapor and noncondensable gas partial pressure, Pa or lbf/ft^2

Δp fluid static pressure drop on one side of a heat exchanger core, Pa or lbf/ft^2

Δp_f frictional pressure drop in two-phase flow, Pa or lbf/ft^2

$\Delta p_{f,g}$ single-phase pressure drop considering gas (with its flow rate) in the two-phase mixture flowing alone in the tube across full cross-section, $\Delta p_{f,l}$ defined similarly for the liquid, Pa or lbf/ft^2

$\Delta p_{f,go}$ single-phase pressure drop considering total flow rate of two-phase mixture as gas flowing in the tube across full cross-section; $\Delta p_{f,lo}$ defined similarly for the liquid, Pa or lbf/ft^2

Δp_{hs} hydrostatic (graviational) pressure drop, Pa or lbf/ft^2

Δp_m momentum pressure drop, Pa or lbf/ft^2

p_r reduced pressure $= p/p_{cr}$, dimensionless

p_t tube pitch, center-to-center distance between tubes, m or ft

Q_T total heat removed from vapor to given point, W or Btu/h

q heat-transfer rate, heat duty, W or Btu/h

q'' heat flux, heat-transfer rate per unit area W/m^2 or Btu h^{-1} ft^{-2}

R heat capacity rate ratio, $R_1 = C_1/C_2$, $R_2 = C_2/C_1$, dimensionless

$\boldsymbol{R}$ thermal resistance based on area A, $R_0 = 1/UA$ (see text after Eq. (7) for the definitions of other thermal resistances, K/W or h °F/Btu)

R pressure drop correction factors for shell-side leakage and bypass flows, refer to specific definitions after Equation (148), dimensionless

Re Reynolds number based on the hydraulic diameter $= G D_h/\mu = \varrho v D_h/\mu$, dimensionless

Re_d Reynolds number based on the tube outside diameter $= G d_o/\mu = \varrho v d_o/\mu$, dimensionless

$Re_{\Gamma,L}, Re_{\Gamma,z}$ mean and local film Reynolds numbers, defined by Equations (112) and (104), dimensionless

St Stanton number $= h/(G c_p)$, dimensionless

s fin spacing or fin pitch $s = 1/N_f$; also transverse spacing for offset-strip fins, m or ft

s' distance between two fins $= s - \delta$, m or ft

T temperature relative to an arbitrary datum, °C

ΔT_{lm} log-mean temperature difference, °C

ΔT_m true mean temperature difference $= F\,\Delta T_{lm}$, °C

Ⓣ thermal boundary condition referring to constant wall temperature, both axially and peripherally

T_{sat} saturation temperature, °C

U overall heat-transfer coefficient, W m^{-2} K^{-1} or Btu h^{-1} ft^{-2} °F

V specific volume, m^3/kg or ft^3/lbm

v fluid mean axial velocity occurring at the minimum free flow area, m/s or ft/s

v_c fluid mean velocity for flow normal to a tube bank based on $(X_t - d_o)$ gap, m/s or ft/s

v_{gs} superficial gas velocity or volumetric gas flux, the rate of volumetric flow of gas of two-phase mixture divided by the total flow area, m/s or ft/s

v_{ls} superficial liquid velocity, defined similar to v_{gs}, m/s or ft/s

W fluid mass flow rate $= \varrho v A_o$, kg/s or lbm/s

X, X_{tt} Martinelli parameter, X defined in Equation (90); X_{tt} defined by Equation (95 b), dimensionless

X_d diagonal tube pitch $= (X_t^2 + X_l^2)^{1/2}$; m or ft

X_l longitudinal tube pitch, m or ft

X_t transverse tube pitch, m or ft

x vapor or gas mass quality, vapor weight fraction, or dryness fraction, dimensionless

Z ratio of vapor sensible heat removed to the total vapor enthalpy change

z Cartesian coordinate along the flow direction, m

z^* axial coordinate for the thermal entrance region $= z D_h^{-1} Re^{-1} Pr^{-1}$, dimensionless

z^+ axial distance $= z D_h^{-1} Re^{-1}$, dimensionless

Greek Symbols

α void fraction

$\boldsymbol{\Gamma}$ $= \Delta p_{f,go}/\Delta p_{f,lo}$, dimensionless

Γ condensate or liquid film flow rate per unit perimeter, kg s^{-1} m^{-1} or lbm h^{-1} ft^{-1}

δ wall or fin thickness, m or ft

ε exchanger effectiveness, dimensionless

η_f temperature effectiveness of fins or an extended surface, dimensionless
η_o total surface temperature effectiveness, defined by Equation (10), dimensionless
θ angle of inclination [Eq. (69)]
λ longitudinal wall heat conduction parameter $= k_w A_k L^{-1} C_{min}^{-1}$, dimensionless
μ dynamic viscosity, Pa s or lbm h^{-1} ft^{-1}
ϱ fluid density, kg/m^3 or lbm/ft^3
$\boldsymbol{\sigma}$ ratio of free flow area to frontal area $= A_o/A_{fr}$, dimensionless
σ surface tension, N/m or lbf/ft
τ shear stress, Pa or lbf/ft^2
Φ_B correction term, see Equation (166), dimensionless
φ_g, φ_l two-phase frictional pressure drop multipliers, defined by Equation (87), dimensionless
φ_{lo} two-phase frictional pressure drop multiplier, defined by Equation (76), dimensionless
ψ ratio of true mean temperature difference to inlet temperature difference $= \Delta T_m/(T_{h,i} - T_{c,i})$, dimensionless

Subscripts

B bundle
c cold fluid side
cf counterflow
cr critical
f fin (fluid in Section 10)
fd fully developed flow value
g gas
H1, H2 constant axial wall heat flux boundary conditions
h hot fluid side
hom homogeneous
i inlet to the exchanger
i interface
id ideal
L laminar flow
l liquid
lm logarithmic mean
m mean or bulk value
o outlet to the exchanger; overall when used with $\boldsymbol{R}$ or ε
s scale or fouling when used with $\boldsymbol{R}$ or h; otherwise, shell side
T constant wall temperature boundary condition, also turbulent flow in Equations (34), (66), and (67) only, otherwise total
t tube side
tp two-phase
v vapor
w wall or property at the wall temperature
z a local value along the flow length
1 one section (inlet or outlet) of the exchanger; one pass in a multipass exchanger; fluid 1
2 other section (outlet or inlet) of the exchanger; fluid 2

1. Introduction

The topic heat exchange includes a broad range of topics that cannot be treated here. This article describes heat exchangers, and would be better entitled Heat Exchangers.

A heat exchanger is a device that transfers thermal energy between two or more fluids that have differing temperatures. Typical applications involve heating or cooling of a fluid stream, evaporation or condensation of single or multicomponent fluid stream, and heat recovery or rejection from a system. In other applications, the purpose is to heat, cool, condense, vaporize, sterilize, pasteurize, fractionate, distil, concentrate, crystallize, or control process fluid. In some heat exchangers, the fluids transferring heat are in direct contact. In other exchangers, heat transfer between fluids takes place through a separating wall or into and out of a wall in a transient manner. In most process heat exchangers, fluids are separated by a heat-transfer surface, and ideally they do not mix. Such exchangers are referred to as *direct transfer type* or simply *recuperators.* In contrast, exchangers in which there is an intermittent flow of heat from the hot to cold fluid–via heat storage and heat rejection through the exchanger surface or matrix–are referred to as *indirect transfer type or storage type exchangers,* or simply *regenerators.*

In this article, comprehensive information is presented for design of single-phase and two-phase heat exchangers used in process, power, petroleum, air-conditioning and refrigeration, cryogenics, transportation, and other industries. Also, design information is provided on impor-

tant operating problems such as flow-induced vibration, flow maldistribution, and fouling.

2. Classification of Heat Exchangers

Heat exchangers can be classified according to transfer process, construction, flow arrangement, surface compactness, number of fluids, and heat-transfer mechanisms as shown in Figure 1 [1]. They can also be classified according to the process function (Fig. 2), industry, type of fluid, and site of installation in the plant.

2.1. Classification According to Construction

2.1.1. Tubular Heat Exchangers

Tubular heat exchangers are generally built of circular tubes, although elliptical and rectangular tubes have also been used. There is considerable design flexibility because the core geometry can be varied easily by changing the tube diameter, length, and arrangement. Tubular exchangers can be designed for any operating temperatures, pressures, and any temperature and pressure differences between the fluids, limited only by the materials of construction. They can be designed for special operating conditions: heavy fouling, highly viscous flow, corrosiveness, toxicity, radioactivity, etc. These exchangers may be further classified as shell-and-tube, double-pipe, spiral and tube – coil exchangers. Only the first two types, the most important, are described further here.

2.1.1.1. Shell-and-Tube Exchangers

Shell-and-tube exchangers are composed of round tubes mounted in a cylindrical shell with the tube axis parallel to that of the shell. One fluid flows inside the tubes, the other flows across and along the tubes. The major components of this exchanger are tubes (or tube bundle), shell, stationary or front-end head, rear-end head, baffles, and tubesheets. Detailed components for several shell-and-tube exchangers are shown in Figure 3.

A variety of internal constructions are used in shell-and-tube exchangers, depending upon the desired heat-transfer and pressure-drop performance and the methods employed to reduce thermal stresses, to prevent leakages, to provide for ease of cleaning, to contain operating pressures and temperatures, and to control corrosion. These exchangers are classified and constructed according to Tubular Exchanger Manufacturers Association (TEMA) standards [2] in the United States or modified TEMA standards in other countries.

TEMA has developed a notation system to designate major types of shell-and-tube exchangers. In this system, each exchanger is designated by a three letter combination, the first letter indicating the front-end head type, the second the shell type, and the third the rear-end head type. These are identified in Figure 4. Some of the common shell-and-tube exchangers are AES, AEP, CFU, AKT, and AJW. Other special types of shell-and-tube exchangers have front-end and rear-end heads different from those in Figure 3. These exchangers may not be identifiable by the TEMA letter designation.

The sizes of shells or tube bundles are designated by describing the shell/tube bundle nominal diameter and tube lengths (both in inches, 1 in = 2.54 cm, by TEMA). Except for the kettle reboiler, the nominal diameter is the shell inside diameter, rounded off to the nearest integer; for the kettle reboiler, the nominal diameter is the port diameter (the shell inside diameter (*ID*) near the tube fluid inlet) followed by the shell *ID*. The three letter designation of the type of the exchanger follows the size of shell. Some examples of Figure 3 are the following:

B) Fixed tubesheet exchanger with bonnet stationary head, single pass on shellside and tubeside, 23-1/4 in (59.1 cm) shell inside diameter, 18 ft (5.49 m) tubes. Size 23 – 216 type BEM.
C) Outside packed floating head exchanger with removable channel and cover, one shell pass, two tube passes, shell inside diameter 21 in (53.3 cm) with 14 ft (4.27 m) tubes. Size 21 – 168 type AEP.
D) U-tube exchanger with stationary head channel integral with tubesheet and removable cover, two-pass shell with longitudinal baf-

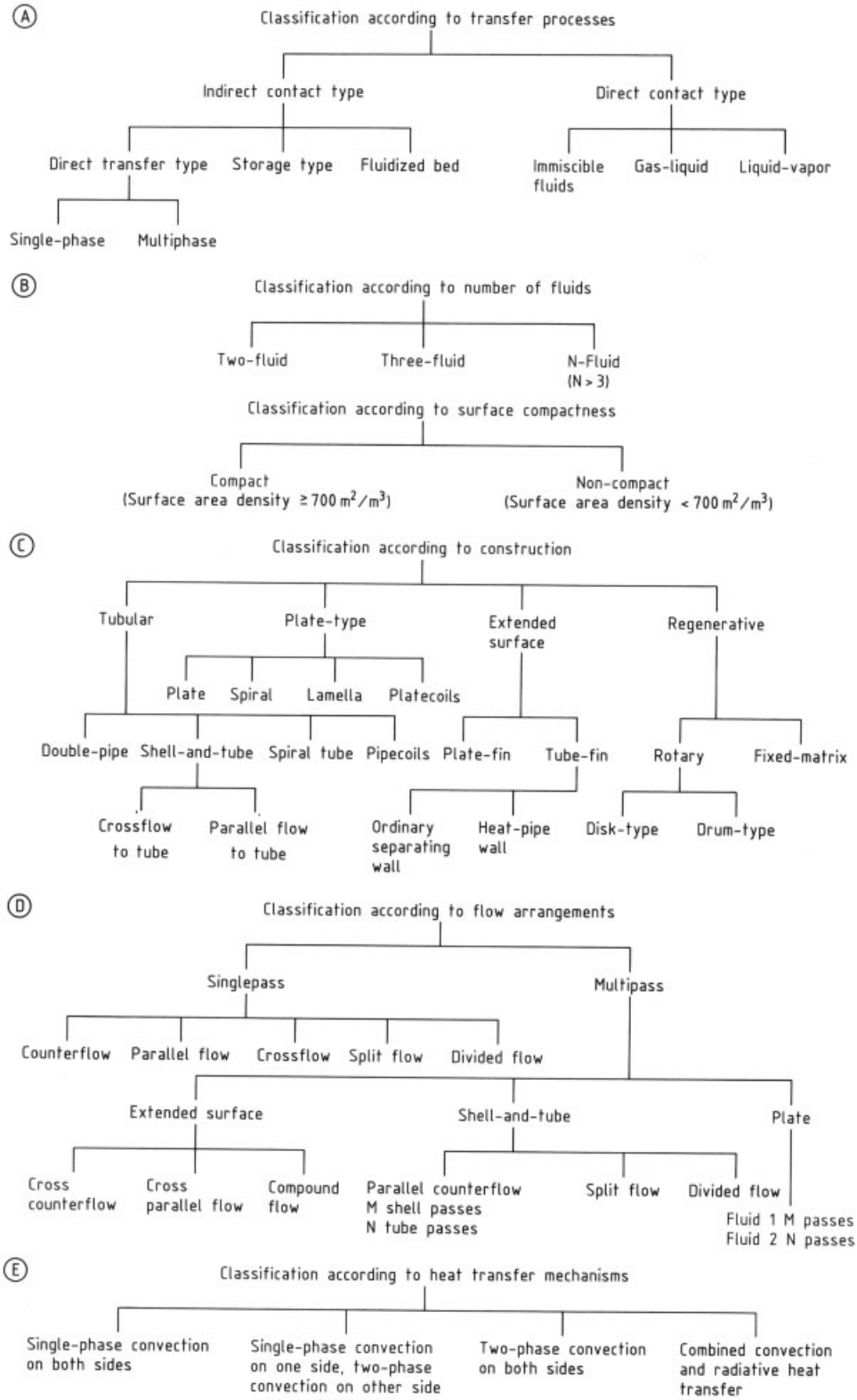

Figure 1. Classification of heat exchangers

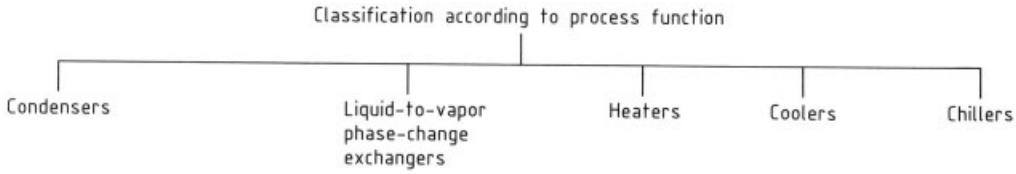

Figure 2. Classification according to process function

fle, 31 in (78.7 cm) shell inside diameter, 20 ft (6.1 m) tubes. Size 31 – 240 type CFU.

E) Pull-through floating head kettle-type reboiler with removable channel and cover, 25 in (63.5 cm) port diameter and 44 in (1.12 m) shell inside diameter with 20 ft (6.1 m) tubes. Size 25/44 – 240 type AKT.

TEMA has set up mechanical standards for design, fabrication, and materials for three classes of heat exchangers: R, C, and B. *Class R* designates the unfired shell-and-tube heat exchangers generally used for severe conditions in petroleum and related processing. *Class C* designates the unfired shell-and-tube heat exchangers generally used for moderate conditions of commercial and general processing. *Class B* designates the unfired shell-and-tube heat exchangers used for chemical process service. Generally, Class B exchangers employ nonferrous materials, while the other heat exchangers employ ferrous materials.

Further description of the individual components of shell-and-tube exchangers is provided in Section 4.2.

2.1.1.2. Double-Pipe Heat Exchangers

Double-pipe heat exchangers usually consist of concentric pipes. One fluid flows in the inner pipe and the other flows countercurrently in the annulus between the pipes. This is perhaps the simplest heat exchanger. Flow distribution is no problem, and cleaning is done easily by disassembly. This configuration is appropriate when one or both of the fluids are at very high pressure because containment in small diameter pipe or tubing is less costly than containment in large diameter shell. Double-pipe exchangers are generally used for the small capacity applications where the total heat-transfer surface area is 50 m^2 (500 ft^2) or less. Stacks of double-pipe or multitube-type heat exchangers are also used in some process applications with radial or longitudinal fins.

2.1.1.3. Drip-Type Heat Exchangers

Drip or trombone exchangers are serpentine or flattened helical coil tubes with water flowing over the horizontal tubes in a falling film. They are similar to double-pipe exchangers except that the outer pipe jacket is omitted. It has the advantages of a high water side heat-transfer coefficient, easy cleaning while in operation, and no possibility of leakage of water contaminating the process fluid. Additional cooling can take place by evaporation, but this is of minor importance. Major disadvantages are the large space requirement, the need for a sump pump or sewer line, and messy operation.

2.1.2. Plate-Type Heat Exchangers

Plate-type heat exchangers are usually built of thin plates (all prime surface), which are either smooth or corrugated and are either flat or wound in an exchanger. Generally, these exchangers cannot accommodate high pressure and temperature differences. These exchangers may be further classified as plate, spiral plate, lamella, and platecoil exchangers [1], the first two types being described further.

2.1.2.1. Plate Heat Exchangers

The plate and frame or gasketed plate heat exchanger (PHE) consists of a number of thin, rectangular metal plates sealed around the edges by gaskets and held together in a frame as shown in Figure 5. The frame usually has a fixed-end cover (headpiece) fitted with the connecting ports and a movable cover plate (pressure

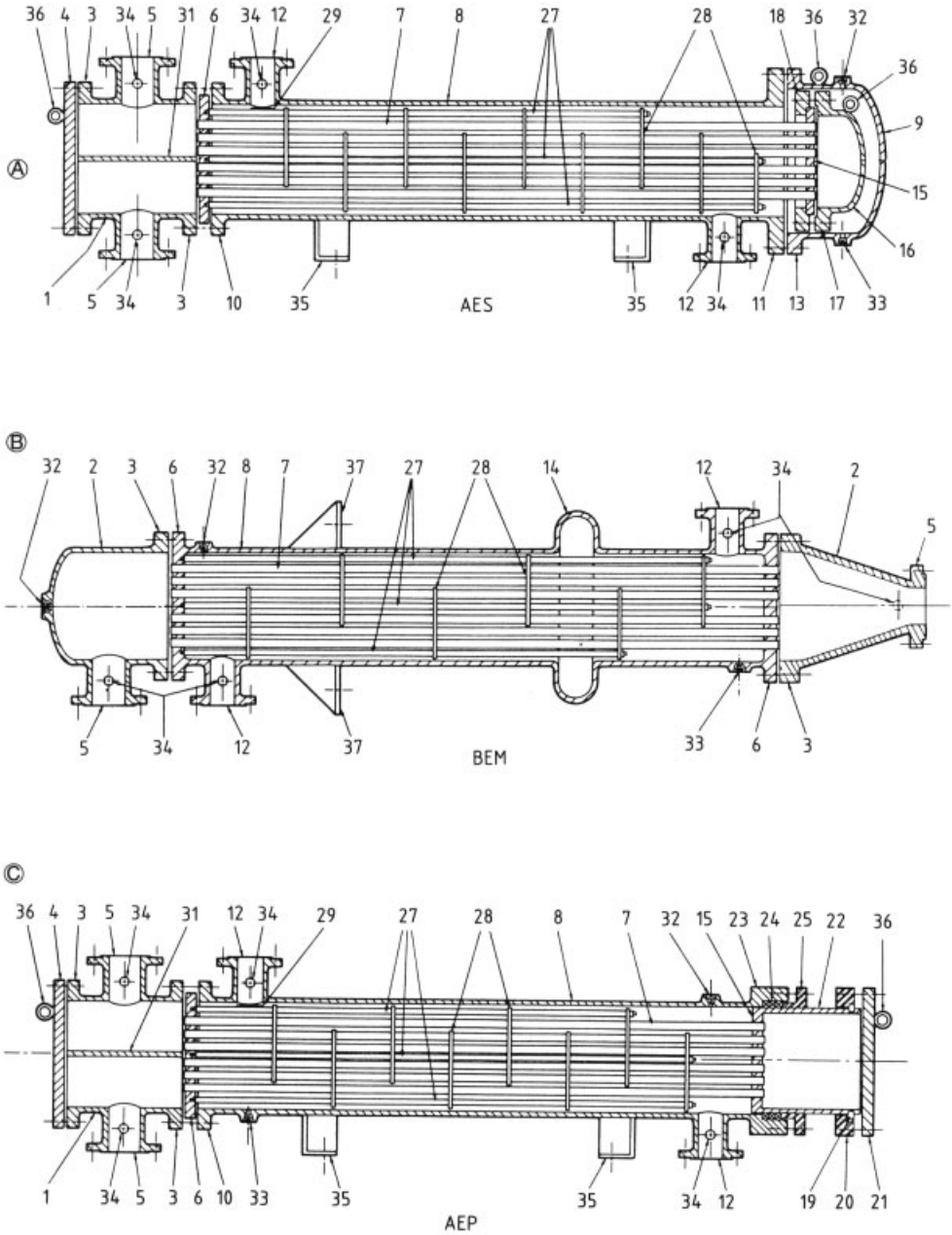

Figure 3. Shell-and-tube heat exchangers [2]
1 Stationary head–channel; 2 Stationary head–bonnet; 3 Stationary head flange–channel or bonnet; 4 Channel cover; 5 Stationary head nozzle; 6 Stationary tubesheet; 7 Tubes; 8 Shell; 9 Shell cover; 10 Shell flange–stationary head end; 11 Shell flange–rear-head end; 12 Shell nozzle; 13 Shell cover flange; 14 Expansion joint; 15 Floating tubesheet; 16 Floating head cover; 17 Floating head flange; 18 Floating head backing device; 19 Split shear ring; 20 Slip-on backing flange; 21 Floating head cover–external; 22 Floating tubesheet skirt; 23 Packing box; 24 Packing; 25 Packing gland; 26 Lantern ring; 27 Tierods and spacers; 28 Transverse baffles or support plates; 29 Impingement plate; 30 Longitudinal baffle; 31 Pass partition; 32 Vent connection; 33 Drain connection; 34 Instrument connection; 35 Support saddle; 36 Lifting lug; 37 Support bracket; 38 Weir; 39 Liquid level connection

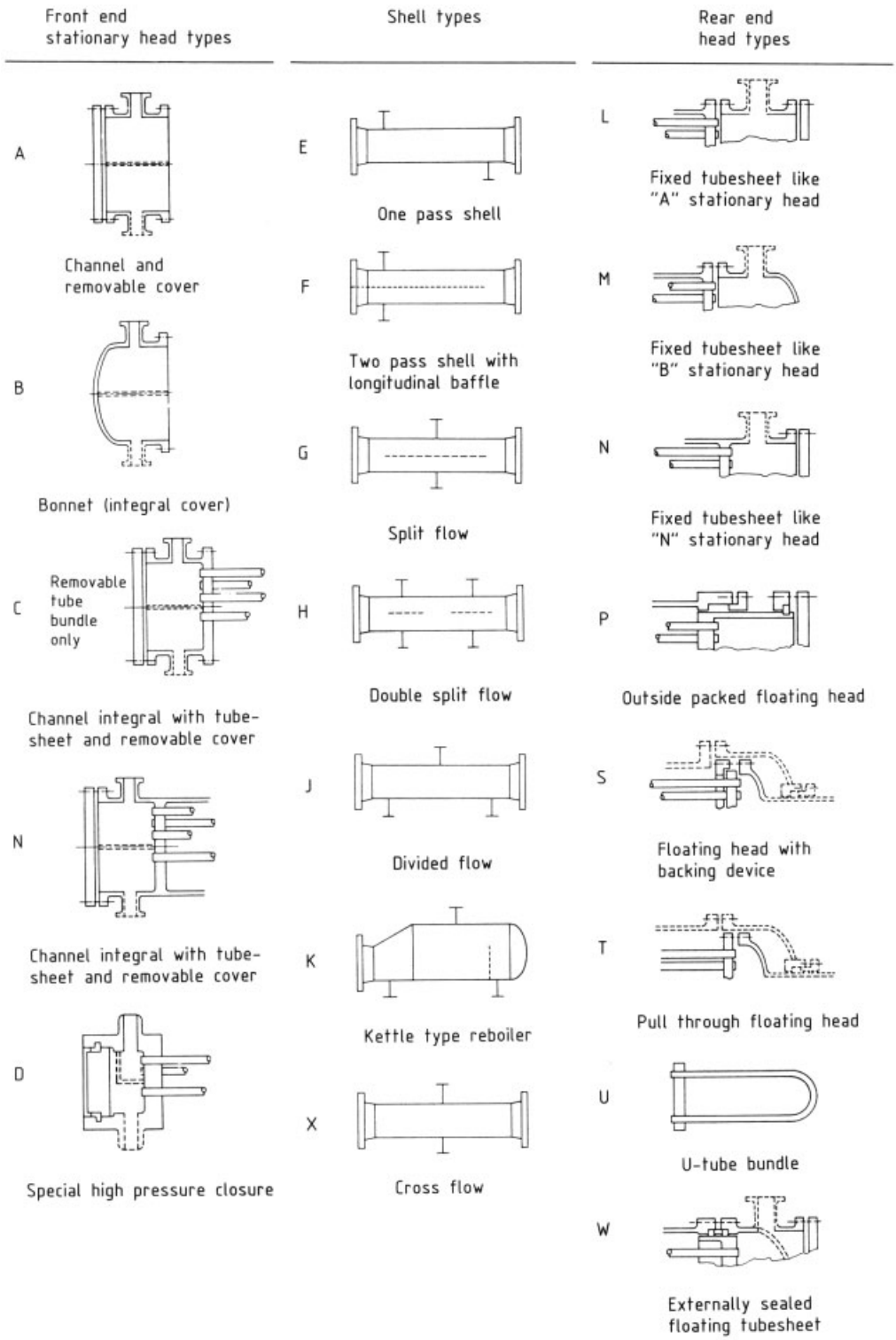

Figure 4. Standard shell types, front-end and rear-end head types [2]

plate, follower, or tailpiece). In the frame, the plates are suspended from an upper carrying bar and guided by a bottom carrying bar to ensure proper alignment. The plate pack with fixed and movable end covers is clamped together by lateral bolts. The carrying bars are longer than the compressed stack so that when the movable end cover is removed, plates may be slid along the support bars for inspection and cleaning.

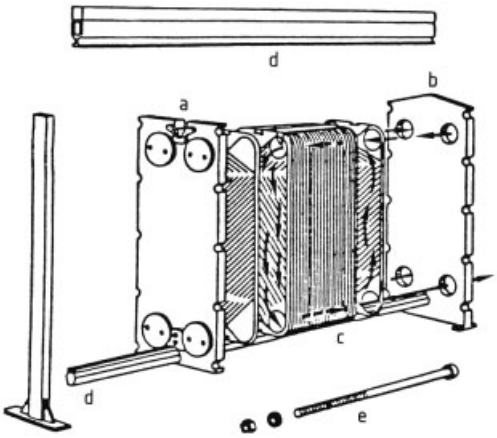

Figure 5. A plate heat exchanger (courtesy of Alfa-Laval AB, Sweden)
a) Movable end cover; b) Fixed end cover; c) Plate pack; d) Carrying bar; e) Compression bolt

Each plate is made by stamping or embossing a corrugated (or wavy) surface pattern on sheet metal. Typical plate geometries (corrugated patterns) are shown in Figure 6. The plates are assembled with the corrugations on successive plates abutting so as to form narrow interrupted or torturous flow paths. Since adjacent plates are in mutual contact, the mean plate spacing is well defined and determined by the geometry of the corrugations. The corrugations also improve the rigidity of the plates and provide mechanical support to the plate pack through multiple contact points. They also enhance the heat-transfer rate and decrease fouling resistance by increasing the level of turbulence. Plates are designed as *hard* or *soft*, depending upon whether they generate high or low intensity of turbulence.

Sealing is accomplished by elastomeric gasket (typically 5 mm, 0.2 in, thick) that are fitted in special grooves provided along the edges of the plates and around the ports as indicated by the dark lines in Figure 6. Some manufacturers offer special interlocking types to prevent gasket blowout at high differential pressures. Typical gasket materials and their range of applications are listed in Table 1 [3].

A large number of pass arrangements are possible in a plate heat exchanger [3], [4] depending on the required heat-transfer duty, available pressure drops, minimum and maximum velocities allowed, and the flow rate ratio of the two fluid streams. Whether the plate exchanger is a single-pass or multipass unit, the thermodynamically superior counterflow or overall counterflow arrangement is used exclusively.

Some advantages of plate heat exchangers are as follows:

They can easily be taken apart for cleaning, inspection, and maintenance. The heat-transfer surface area can be readily changed or rearranged for a different task or anticipated changing loads, through the flexibility of plate size, corrugation patterns, and pass arrangements. The high turbulence due to plates reduces fouling to ca. 10 – 25 % that of a shell-and-tube exchanger. Because of the high heat-transfer coefficients, reduced fouling, absence of bypass and leakage streams, and pure counterflow arrangements, the surface area required for a plate exchanger is one-half to one-third that of a shell-and-tube exchanger for a given heat duty, thus reducing the cost, overall volume, and maintenance space for the exchanger. The gross weight of a plate exchanger is about one-sixth that of an equivalent shell-and-tube exchanger. Leakage from one fluid to the other cannot take place unless a plate develops a hole. Since the gasket is between the plates, any leakage from the gaskets is to the outside of the exchanger. The residence time for fluid particles on a given side is approximately the same. This is desirable for uniformity of heat treatment in application such as sterilizing, pasteurizing, and cooking. There are no significant hot or cold spots in the exchanger that could lead to the deterioration of heat-sensitive fluids. The volume of fluid held up in the exchanger is small; this is important with expensive fluids, for faster transient response, and for a better process control. Finally, high thermal performance can be achieved in plate exchangers. The high degree of counterflow makes temperature approaches of 1 °C possible. The high thermal effectiveness (up to ca. 93 %) facilitates economical low-grade heat recovery. Flow-induced vibration, noise, thermal stresses, and entry im-

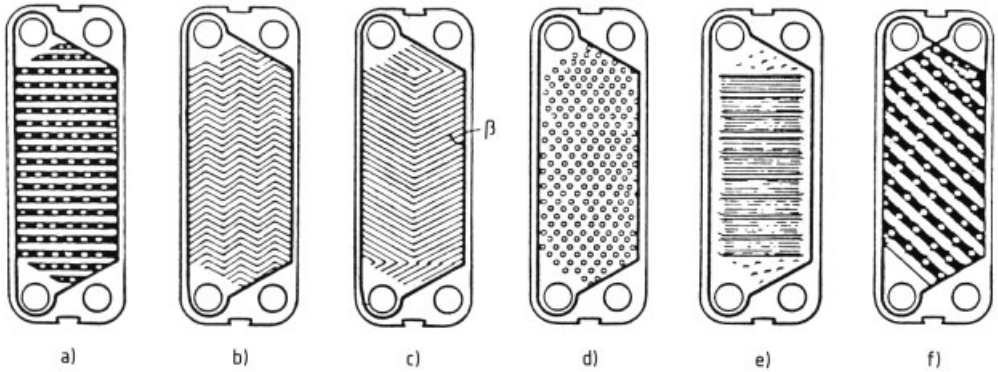

Figure 6. Plate patterns: a) Washboard; b) Zig-zag; c) Chevron or herringbone,d) Protrusions and depressions; e) Washboard with secondary corrugations; f) Oblique washboard

Table 1. Gasket materials used in plate heat exchangers [3]

Gasket material	Generic name	Maximum operating temperature, °C	Applications	Comments
Natural rubber	*cis*-1,4-polyisoprene	70	oxygenated solvents, acids, alcohols	
SBR (Styrene butadiene)		80	general purpose aqueous, alkalies, acids, and oxygenated solvents	poor fat resistance
Neoprene	*trans*-1,4-polychloroprene	70	alcohols, alkalies, acids, aliphatic hydrocarbon solvents	
Nitrile		100 – 140	dairy products, fruit juices, beverages, pharmaceutical and biochemical applications; oil, gasoline, animal and vegetable oils, alkalies, and aliphatic organic solvents	resistant to fatty materials; particularly suitable for cream
Butyl (resin cured)		120 – 150	alkalies, acids, animal and vegetable oils, aldehydes, ketones, phenols, and some esters	poor fat resistance; suitable for UHT milk duties; resists inorganic chemical solutions up to 150 °C
Ethylene propylene (EPDM) rubber		140	alkalies, oxygenated solvents	unsuitable for fatty liquids
Silicone rubber	polydimethylsiloxane	140	general low temperature use, alcohols, sodium hypochlorite	
Fluorinated rubber		175	high temperature aqueous solutions, mineral oils and gasoline, organic solvents, animal and vegetable oils	
Compressed asbestos fiber		200 – 260	organic solvents, high operating temperature applications	

pingement problems of shell-and-tube exchangers do not exist in plate heat exchangers.

Some inherent limitations of plate heat exchangers are caused by the plates and gaskets:

The plate exchanger can be used for a maximum pressure of about 2.5 MPa (360 psig), but usually below 1.0 MPa (150 psig). The gasket materials (except for the Teflon-coated type) restrict the use of plate exchangers in highly corrosive applications; they also limit the maximum operating temperature to 260 °C (500 °F), although operation is usually carried out below 150 °C (300 °F) to avoid the use of expensive gasket materials. For equivalent flow velocities, the pressure drop in plate exchangers is very high compared to shell-and-tube exchangers. However, the flow velocities are usually low and the plate lengths are short, so the resulting pressure drops are generally acceptable. Because of the long gasket periphery, plate exchangers are not suitable for high vacuum applications. Plate exchangers are not suitable for erosive duties [5]

or for fluids containing fibrous materials. In certain cases, suspensions can be handled, but to avoid clogging, the largest suspended particles should be at most one-third the size of the average channel gap. Viscous fluids can be handled, but extremely viscous fluids lead to flow maldistribution problems, especially on cooling. Plate exchangers should not be used for toxic fluids due to potential gasket leakage. Some of the largest units have a total surface area of about 2500 m^2 (27 000 ft^2) per frame. Large differences in fluid flow rates of two streams with reasonable pressure drops cannot be handled in a plate exchanger.

2.1.2.2. Spiral-Plate Heat Exchangers

The spiral-plate heat exchanger consists of two relatively long strips of sheet metal, provided with welded studs for plate spacing, and wrapped helically around a split mandrel to form a pair of spiral channels for two fluids, as shown in Figure 7. Thus, each fluid has a long single passage arranged in a compact package. The basic spiral element is sealed by either welding at each side of the channel or by providing a gasket (asbestos-based) at each end cover to obtain the following arrangements of the two fluids: (1) both fluids in spiral counterflow; (2) one fluid in spiral flow, the other in crossflow across the spiral; or (3) one fluid in spiral flow, and the other in a combination of crossflow and spiral flow. The whole assembly is housed in a cylindrical shell enclosed by two circular end covers, either flat or conical.

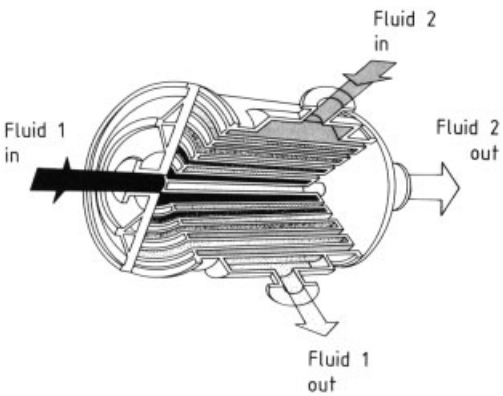

Figure 7. A spiral plate heat exchanger with both fluids in spiral counterflow

The spiral plate exchanger has a relatively large diameter because of the spiral turns. This is a noncompact heat exchanger with a maximum of about 185 m^2 (2000 ft^2) surface area for a maximum 1.47 m (58 in) diameter shell. The heat-transfer coefficients are not as high as in a plate exchanger if the plates are not corrugated. However, the heat-transfer coefficient is higher than that for a shell-and-tube exchanger because of curved rectangular passages. Hence, the surface area requirement is about 20 % lower than a shell-and-tube unit for the same heat duty.

The advantages of the spiral-plate heat exchanger are the following:

It can handle viscous, fouling liquids, and slurries more readily because of the *single passage*. The fouling rate is low compared to the shell-and-tube unit. It is more amenable to chemical, flush, and reversing fluids cleaning techniques, and thus maintenance is low compared to a shell-and-tube unit. The internal void volume is low compared to a shell-and-tube exchanger, and it is well suited as a condenser or reboiler.

The disadvantages of this exchanger are the following:

The maximum size is limited. The maximum operating pressure is limited to 1000 kPa (150 psig) for large units and the maximum operating temperature is limited to 500 °C (930 °F) with compressed asbestos gaskets. Field repair is difficult due to construction features.

2.1.3. Extended-Surface Heat Exchangers

The tubular and plate-type exchangers described previously are all prime surface heat exchangers, except for those having low finned circular tubes. The design thermal effectiveness is usually ≤ 60 % for a single-shell unit. The heat-transfer surface area density is usually $\leq 700\ m^2/m^3$ (200 ft^2/ft^3). In many applications, a much higher exchanger effectiveness (up to 98 %) is essential, and the box volume and mass are limited so that a more compact surface is mandated. When either a gas or a liquid having a low heat-transfer coefficient is the fluid on one or both sides, it results in a large heat-transfer surface area requirement. This is met by adding fins on the primary surface on one or both fluid sides. Plate – fin and tube – fin geometries are the

two most common types of extended surface exchangers.

2.1.3.1. Plate – Fin Heat Exchangers

Plate – fin heat exchangers have fins or spacers sandwiched between parallel plates (referred to as plates or parting sheets) as shown in Figure 8. Sometimes fins are incorporated in a flat tube with rounded corners (referred to as a formed tube), thus eliminating a need for the side bars. Other plate – fin constructions include drawn-cup or tube-and-center configurations. These are shown in Figure 9. The plates or flat tubes separate the two fluid streams and the fins form the individual flow passages. Alternate fluid passages are connected in parallel by suitable headers to form the two or more sides of the exchanger. Fins are die- or roll-formed and are attached to the plates by brazing, soldering, adhesive bonding, welding, mechanical fit, or extrusion. Fins are used on both sides in gas-to-gas heat exchangers. In gas-to-liquid applications, fins are usually used only on the gas side–if employed on the liquid side, they are used primarily for structural strength and flow mixing purposes. Fins are also sometimes used for pressure containment and rigidity. In Europe, the plate – fin exchanger is referred to as the *matrix heat exchanger*.

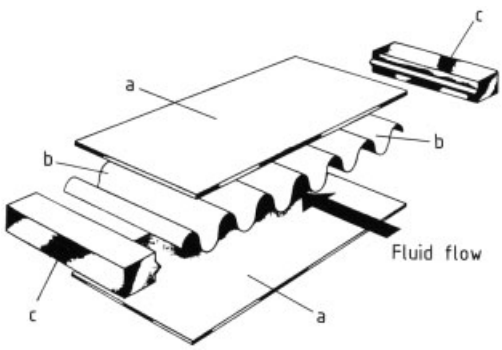

Figure 8. Basic components of a plate-fin heat exchanger
a) Plate or parting sheet; b) Fin; c) Side bar

Plate fins are categorized as (1) plain (i.e., uncut) and straight fins, such as plain triangular and rectangular fins, (2) plain but wavy fins, and (3) interrupted fins such as offset strip, louver, perforated, and pin fins. Examples of commonly used fins are shown in Figure 10. The velocity and temperature boundary layer thicken along the flow length on plain surfaces resulting in both a lower heat-transfer coefficient and a lower friction factor compared to interrupted fins. Plain fins are used when the pressure drop is critical, and interrupted or wavy fins cannot meet the pressure drop requirement together with a flow (or frontal) area constraint. For wavy fins, the boundary layers are either thinned or interrupted when the fluid follows waves, resulting in both a higher heat-transfer coefficient and a higher friction factor. Boundary layers can be disrupted further if the fin surface is made highly discontinuous. Examples are strip fins, louver fins, and perforated fins. Strip fins are also referred to as offset fins, lance-offset fins, serrated fins, and segmented fins. Many variations of interrupted fins are used in industry since they are more efficient than plain fins, and are therefore used when allowed by the design constraints.

Plate – fin exchangers are generally designed for moderate pressures (< 700 kPa (100 psig)), although plate – fin exchangers for operating pressures up to ca. 8300 kPa (1200 psig) are available commercially. The temperature limitation depends on the method of bonding and the material employed. Such exchangers have been used for temperatures up to ca. 840 °C (1550 °F). Plates and fins are made from a variety of materials.

Plate – fin exchangers have been built with a surface area density of up to 5900 m^2/m^3 (1800 ft^2/ft^3). There is total freedom in selecting the fin surface areas on each fluid side, as required by the design, by varying fin height and fin density. Although typically fin densities are 120 – 700 fins/m (3 – 18 fins/in), applications exist for as many as 2100 fins/m (53 fins/in). Common fin thicknesses range from 0.05 to 0.25 mm (0.002 – 0.01 in). Fin heights may range from 2 to 25 mm (0.08 – 1.0 in). A plate – fin exchanger with 600 fins/m (15.2 fins/in) provides about 1300 m^2 of heat-transfer surface area per cubic meter volume occupied by the fins (400 ft^2/ft^3). Plate – fin exchangers are manufactured in virtually all shapes and sizes.

Plate – fin exchangers are widely used in electric power plants (gas turbine, steam, nuclear, fuel cell, etc.), in thermodynamic cycles (heat pump, refrigeration, etc.), and in electronics, cryogenics, gas-liquefaction, air-conditioning, transportation, and waste-heat-recovery systems.

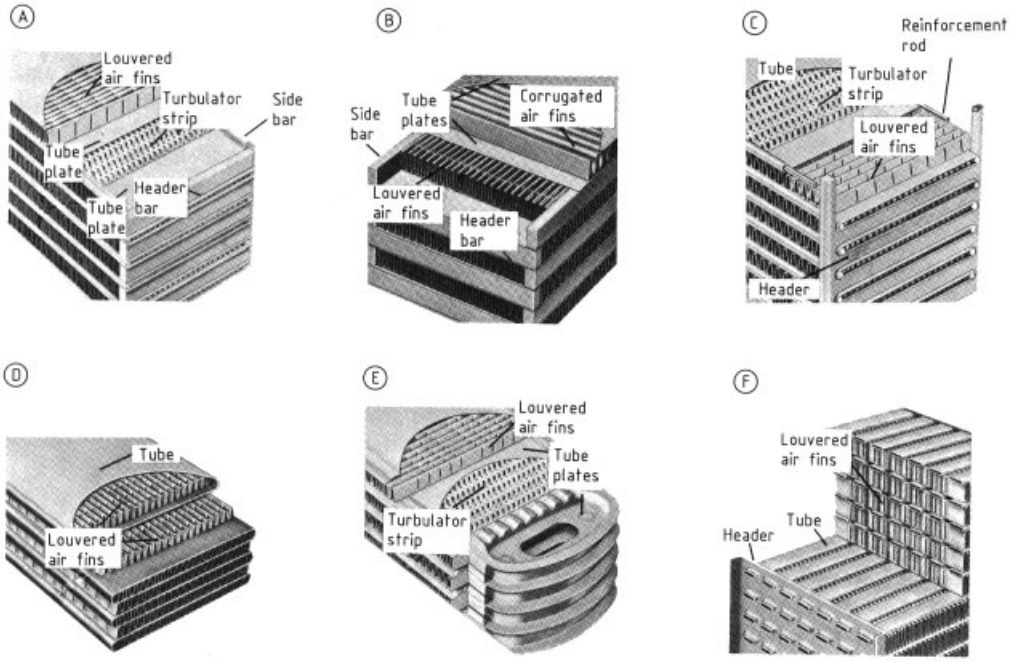

Figure 9. Plate – fin heat exchangers (courtesy of Harrison Radiator, GMC; Lockport, New York)
A) Bar and plate; B) Bar and plate; C) Formed plate – fin; D) Formed plate – fin; E) Drawn cup; F) Tube and center

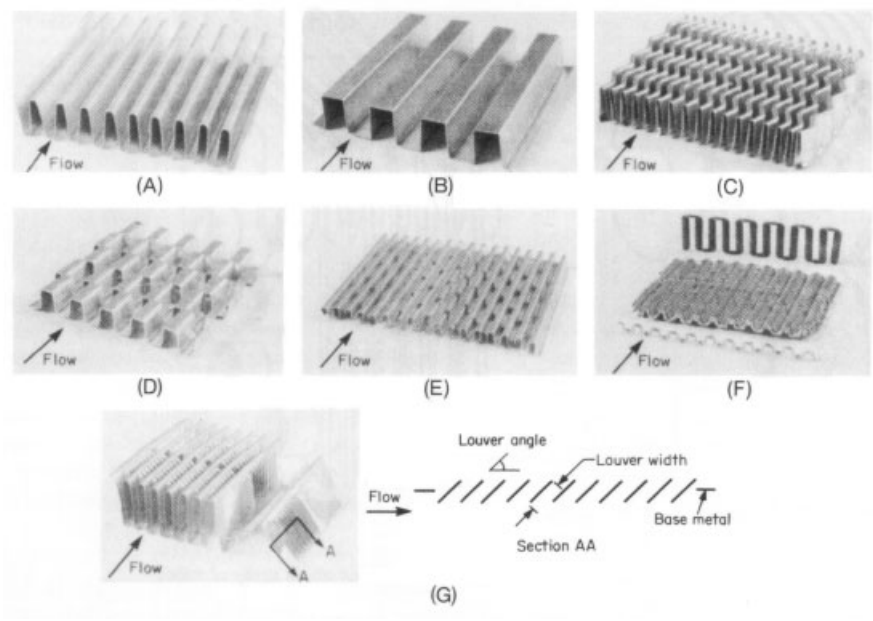

Figure 10. Fin geometries for plate – fin heat exchangers
A) Plain triangular fin; B) Plain rectangular fin; C) Wavy fin; D) Offset strip fin; E) Perforated fin; F) Pin fin; G) Multilouver fin

2.1.3.2. Tube – Fin Heat Exchangers

Tube – fin heat exchangers may be classified as conventional and specialized tube – fin exchangers. In a conventional tube – fin exchanger, heat transfer between the two fluids takes place through the tube wall by conduction. However, in a heat-pipe exchanger, a specialized type of tube – fin exchanger, tubes with both ends closed act as a separating wall, and heat transfer between the two fluids takes place through this separating wall (heat pipe) by conduction, and evaporation and condensation of the heat-pipe fluid [6].

In the gas-to-liquid exchanger, the heat-transfer coefficient on the liquid side is generally very high compared to that on the gas side. Hence, to have balanced thermal conductances (approximately equal hA) on both sides for an optimum heat exchanger, fins are used on the gas side to increase surface area A. This is also true with a condensing or evaporating fluid stream on one side and gas on the other. In addition, if the pressure is high for one fluid, it is generally economical to employ tubes.

In a tube – fin exchanger, round and rectangular tubes are the most common, although elliptical tubes are also used. Fins are generally used on the outside but in some applications they may be used on the inside of the tubes. They are attached to the tubes by a tight mechanical fit, tension winding, adhesive bonding, soldering, brazing, welding, or extrusion.

Fins on the outside of the tubes may be categorized as follows:

1) Normal fins on individual tubes, referred to as individually finned tubes or simply as finned tubes, as shown in Figure 11
2) Continuous (plain, wavy, or interrupted) flat fins on an array of tubes, as shown in Figure 12
3) Longitudinal fins on individual tubes as shown in Figure 13

The exchanger with continuous flat fins is usually less expensive on a unit heat-transfer surface area basis because of its simple and mass production type construction features. Individually finned tubes are more rugged and practical in large tube – fin exchangers. Longitudinal fins are generally used in condensing applications, and for viscous fluids in double pipe exchangers.

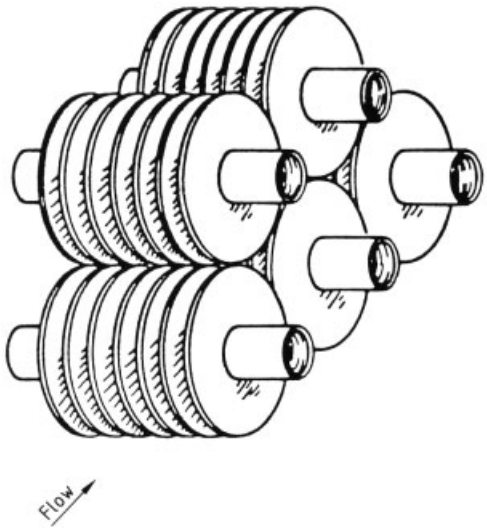

Figure 11. Individually finned tubes

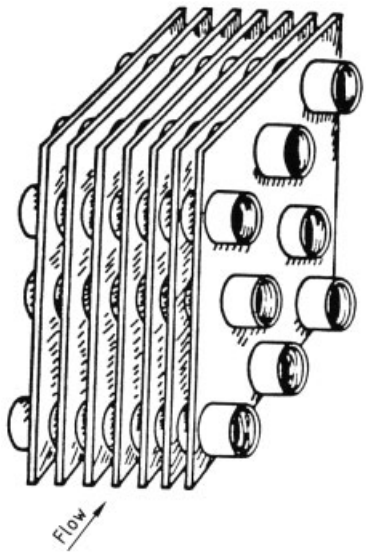

Figure 12. Flat fins on an array of tubes

Shell-and-tube exchangers sometimes employ low finned tubes to increase the surface area on the shellside when the shellside heat-transfer coefficient is low compared to the tubeside coefficient, such as with highly viscous liquids, gases, or condensing refrigerant vapors. The low finned tubes are generally helical or annular fins on individual tubes. Longitudinal fins on indi-

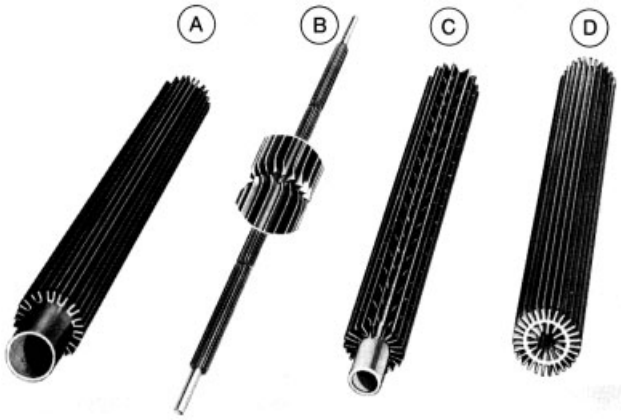

Figure 13. Longitudinal fins on individual tubes (courtesy of Brown Fintube Co., Tulsa, Oklahoma)
A) Continuous plain; B) Cut and twisted; C) Perforated; D) Internal and external longitudinal fins

vidual tubes are also used in shell-and-tube exchangers. Fins on the inside of the tubes are of two types: integral fins as in internally finned tubes, and attached fins.

Tube – fin exchangers can withstand high pressures on the tube side. The highest temperature is again limited by the type of bonding, materials employed, and material thickness. Tube – fin exchangers usually have lower compactness compared to the plate – fin units.

Tube – fin exchangers with an area density of about 3300 m^2/m^3 (1000 ft^2/ft^3) have been built. On the fin side, the desired surface area can be employed by using the proper fin density and fin geometry. The typical fin densities for continuous flat fins vary from 250 to 800 fins/m (6 – 20 fins/in), fin thicknesses vary from 0.09 to 0.25 mm (0.0035 – 0.010 in), and fin flow lengths from 25 to 250 mm (1 – 10 in). A tube – fin exchanger having continuous flat fins with 400 fins/m (10 fins/in) has a surface area density of ca. 720 m^2/m^3 (220 ft^2/ft^3).

Tube – fin exchangers are employed when one fluid stream is at a higher pressure or has a significantly higher heat-transfer coefficient compared to the other fluid stream. As a result, these exchangers are extensively used as condensers and evaporators in air-conditioning and refrigeration applications, as condensers in electric power plants, as oil coolers in propulsive power plants, and as air-cooled exchangers (see fin – fan exchangers, Section 9.1.2.3) in process and power industries.

2.1.4. Regenerators

The regenerator is a storage-type exchanger. In order to have continuous operation, either the matrix must be moved periodically into and out of the fixed streams of gases, as in a *rotary regenerator* (Fig. 14A), or the gas flows must be diverted through valves to and from the fixed matrices as in a *fixed-matrix regenerator* (Fig. 14B). The latter is also sometimes referred to as a *periodic-flow regenerator* or a *reversible heat accumulator.*

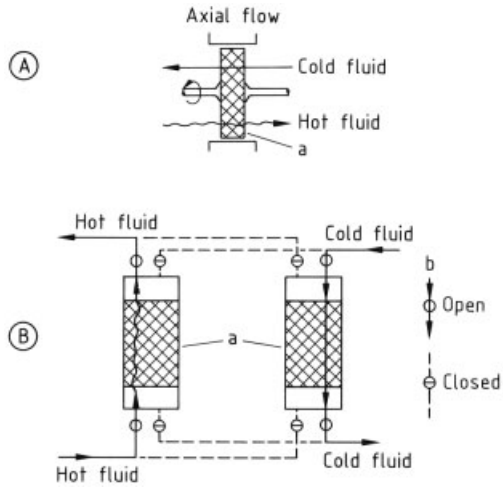

Figure 14. Regenerators
A) Rotary type; B) Valved type
a) Matrix; b) Valve positions

The rotary regenerator is usually a disk type in which the matrix (heat-transfer surface) is in a disk form and fluid flows axially. It is rotated by a hub shaft or a peripheral ring gear drive. Disk-type rotary regenerators are variously referred to as heat wheel, thermal wheel, Munter wheel, or Ljungstrom wheel, depending on their application. When the gas flow is laminar, the disk-type rotary regenerator is also referred to as a laminar flow wheel.

The thermodynamically superior counterflow arrangement is usually employed for storage-type heat exchangers. For some applications, a parallel flow arrangement may be used, but there is no counterpart of the singlepass or multipass crossflow arrangements common in recuperators. For a rotary regenerator, the design of seals to prevent leakages of hot to cold fluids and *vice versa* becomes a difficult task, especially if the two fluids are at significantly differing pressures. Rotating drives also pose a challenging mechanical design problem.

Major advantages of the regenerators are the following:

A much more compact surface may be employed compared to a recuperator. Most fin configurations of plate – fin exchangers, and any finely divided matrix material (high specific heat preferred) may be used providing they achieve high surface area density. Regenerators have been made from metals, ceramics, nylon, plastics, and paper, depending upon the applications. The cost of the regenerator surface per unit of transfer area is usually substantially lower than that for the equivalent recuperator. Compact regenerators usually have a smaller volume and are low in weight compared to an equivalent recuperator. The matrix surface has self-cleaning characteristics for low gas-side fouling because the hot and cold gases flow in alternatively opposite directions in the same fluid passage. Compact surface area density and the counterflow arrangement make the regenerator ideally suited for gas-to-gas heat exchanger applications requiring high exchanger effectiveness, generally exceeding 85 %.

A major disadvantage of the regenerator is an unavoidable carryover of a small fraction of the fluid trapped in the passage to the other fluid stream just after the periodic flow switching. Where fluid contamination is prohibited, a regenerator cannot be used. Hence, the regenerators are used exclusively for gas-to-gas heat or energy transfer.

2.2. Classification According to Process Function

The major process functions for industrial heat exchangers are heating, cooling, chilling, condensation (total or partial), evaporation and boiling, and waste heat recovery or heat rejection. Single-phase heating and cooling applications are relatively simple to design and understand, and are discussed in Section 8. More important but more difficult applications are the design of condensers and liquid-to-vapor phase exchangers (evaporators, boilers, etc.) because of the presence of two-phase phenomena as well as many different industrial uses. These exchangers are further classified next.

2.2.1. Condensers

Condensers are widely used in power generation, chemical process, air-conditioning, automotive, and many other industries. They are classified in Figure 15 (modified from [7]). Broadly speaking, they are either direct contact type or indirect contact type (stream separated).

In a direct contact condenser, the condensing process stream and coolant come into direct contact with each other and heat is transferred. These condensers are inexpensive, have high heat-transfer rates per unit volume, and fouling is not a problem. However, applications are limited due to mixing of the process stream with the coolant. In a pool condenser, the condensing vapor is injected into the pool of liquid coolant. The spray condenser is the most common direct contact condenser in which subcooled liquid is sprayed into the vapor as shown in Figure 16A. If the coolant is dirty, the spray nozzle would clog; in this case a tray condenser as shown in Figure 16B can be used.

Condensers in which the two fluid streams are separated by a wall are classified according to major exchanger construction types as shown in Figure 15. The most common type of condenser used in the power and process industry is the shell-and-tube exchanger. The coolant can be either water (or liquid) or a colder process

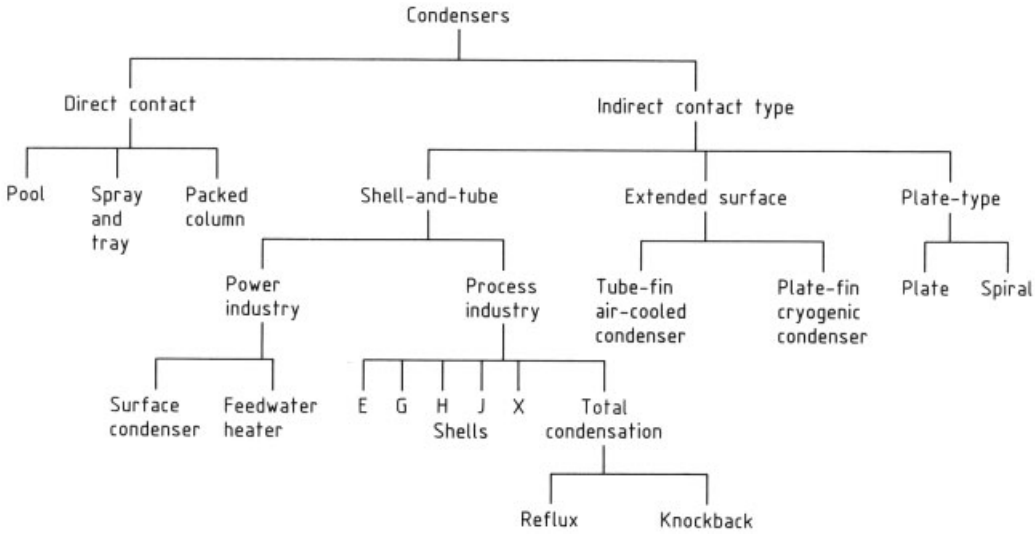

Figure 15. Classification of condensers

stream which requires heating. In power plants, the condenser is used to condense the turbine exhaust steam or preheat the feedwater with condensing steam bled from the turbine. The turbine exhaust steam condenser is referred to as the *surface condenser* or *vacuum condenser.* It has circular tubes and is usually very large; a circular shell is used when heat-transfer surface area is less than about 4650 m^2 (50 000 ft^2), otherwise a rectangular shell is used. The steam is condensed on the shellside. Good venting of noncondensables (to maintain high heat-transfer coefficients) and very low pressure drops on the steam side are the most important design criteria for surface condensers. Further construction details and thermal design are presented in [8].

In a feedwater heater, the boiler feedwater in tubes is heated with a superheated steam on the shell side. On the steam side, one or more of the three modes (desuperheating, condensing, and subcooling) may take place in the feedwater heater, and two- or three-zone feedwater requires special design features on the shellside of a shell-and-tube exchanger. The feedwater heater is usually a two-pass (U tube) exchanger mounted horizontally (Fig. 17), vertically channel-down, or vertically channel-up. Construction details, thermal performance relationships, and operating problems are dealt with in [9].

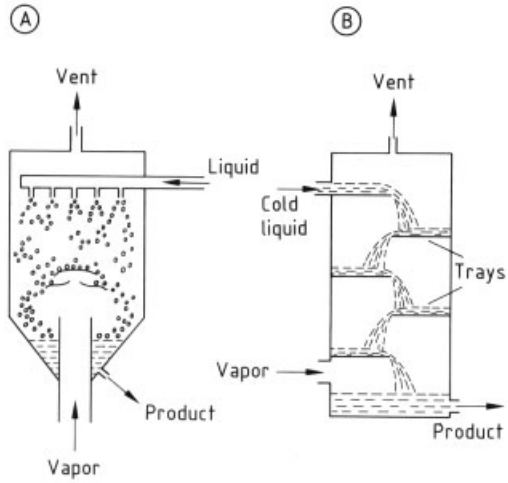

Figure 16. A) Spray condenser; B) Tray condenser

In the process industry, condensation occurs either inside or outside the tubes, the tubes being either horizontal or vertical, and the condensate may be a single component or multicomponent (miscible and/or immiscible) pure vapor with or without noncondensable gases. Condensation may be total (as with a pure vapor) or partial (single or multicomponent vapors with or without noncondensables). In addition to condensation, desuperheating and/or subcooling may also take place in a condenser. One of the most important design features of a condenser is to provide a vent for the removal of noncondensable gases

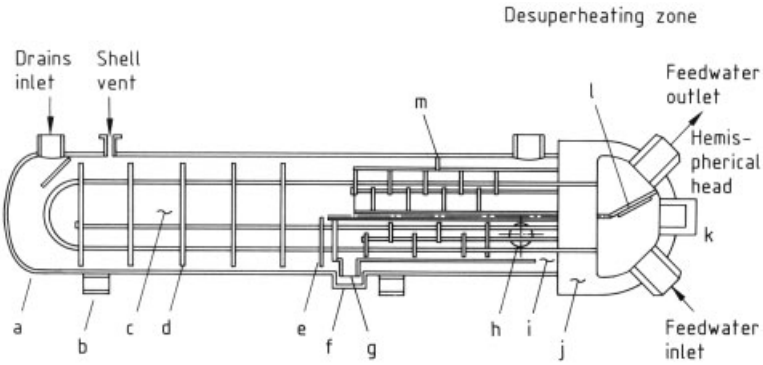

Figure 17. Typical horizontal three-zone feedwater heater
a) Shell head; b) Support; c) Condensing zone; d) Support plates; e) Anti-flashing (or water dam) baffle; f) Sump; g) Snorkel; h) Drains outlet; i) Subcooling zone; j) Tubesheet; k) Manway; l) Pass partition plate; m) Desuperheater closure ring

Table 2. Guide to the selection of type of shell-and-tube condenser [7]

Type	For subcooled condensate	For mixed vapors and/or noncondensables	Pressure range (for pressures between 4 kPa and 10^3 kPa, all types are satisfactory) ∗		If fouling or polymerization	For corrosive condensates or vapors
			below 4 kPa	above 10^3 kPa		
Inside tubes						
Vertical upflow (with reflux condensate)	not applicable	limited to small quantity of low boiler	very poor	good	not applicable	fair
Vertical downflow	good	excellent	poor	excellent	fair	good
Horizontal	very poor but possible with separate subcooling pass	good in annular flow, poor in stratified flow	poor	excellent	fair	good
Outside tubes						
Vertical	possible but not recommended	good	fair	poor–high shell cost	very poor	poor–alloy shell required
Horizontal	not recommended since predictions of cooling coefficient and effective area is unreliable	poor for wide condensing mixtures; good with noncondensables if baffle space is varied; heat and mass transfer prediction is unreliable	fair	poor–high shell cost	very poor	poor

∗ (0.04 – 10 bar)

(regardless of how small they may be in the vapor). The vent should be located near the coldest part of the condenser to avoid escape of vapor, but should be high enough so that it does not flood with condensate. An impingement plate should be placed under the vapor inlet nozzle to prevent tube erosion and possible flow-induced vibrations. A variety of shell types have been used depending upon the applications such as E (single- or multi-segmental baffles), G, H, J, and X, the X shell having the lowest pressure drop on the shell side. Table 2 provides guidelines for the selection of shell-and-tube condensers; further details are available in [7]. Two special types of total condensers, whereby all of the vapor is condensed in the condenser, used in specific applications are a total (with subcooling) condenser and a reflux or knockback condenser shown in Figure 18. These condensers are employed when the liquid in the reactor must be kept at boiling point, or when volatile liquid is stored in a tank (connected to the condenser). These condensers differ according to where the vapor enters the condenser, either at the top or bottom; if vapor entry is at the top, subcooling in the condenser is possible, and if entry is at the bottom, subcooling is not possible. The reflux

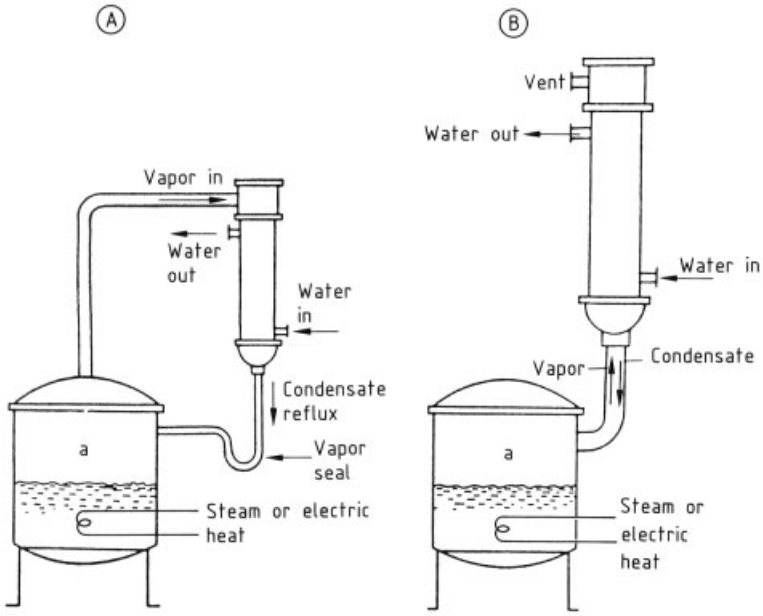

Figure 18. A) Total condenser; B) Reflux or knockback condenser
a) Reactor

condenser returns a hot condensate with most of the low boiling components stripped out.

Extended surface condensers are either tube – fin or plate – fin type and generally have air as the coolant. Tube – fin condensers are used in power and process industries, and in air-conditioning applications. Plate – fin condensers are used in the cryogenics industry.

Air-cooled condensers used in the process industry have a horizontal individually-finned tube bundle, and condensation takes place inside the tubes, air being blown across the tubes on the outside. The condenser usually has a large surface area and short air flow length in order to minimize the fan power consumption. The air is circulated either by forced or induced draft over the tubes. Further details are presented in [10]. Air-cooled condensers used in air-conditioning industry have flat fins (with waves or louvers for enhancement) on round tubes. They are usually smaller than those used in the process industry.

Plate – fin exchangers (see Fig. 8) are used in the cryogenics industry, where on one side condensation takes place (e.g., of air) and on the other side boiling or evaporation take place (e.g., of liquid nitrogen). These condensers are highly efficient (approach temperatures of 1 – 5 °C), compact, and of brazed aluminum construction. Flow passages are small, hence the fluids need to be clean or non-foulant. Heat-transfer and flow friction correlations and general design theory are dealt with in [11].

2.2.2. Liquid-to-Vapor Phase-Change Exchangers

Heat exchangers that involve a phase change from liquid to vapor on one side of a two-fluid exchanger are broadly classified as boilers and vaporizing exchangers, as shown in Figure 19. Boilers are fired exchangers in which fuel is burned to heat the air, which in turn vaporizes the fluid on the other side. Vaporizing exchangers are unfired; in most cases (except for steam generators), condensing steam is used to vaporize the fluid on the other side. The unfired vaporizing exchangers can be classified further– as steam generators, waste heat boilers, evaporators, reboilers, and vaporizers. *Steam generators* are used to generate saturated or superheated steam in power plants. Construction details and design theory of nuclear plant steam generators are covered in [12], [13]. *Waste heat boilers* are used to recover waste heat from a

variety of chemical processes. If the evaporation of water or an aqueous solution takes place in a vaporizing exchanger, it is referred to as an *evaporator.* A vaporizing exchanger (which supplies the heat required to generate vapors in a distillation process) located at the bottom of a distillating column is referred to as a *reboiler* regardless of whether or not the vapor generated is steam. A vaporizing exchanger is referred to as a *vaporizer* when it is neither used to generate the steam nor is it part of a distillation process. An evaporator is designated a *chemical evaporator* when it is used to concentrate a chemical solution by evaporation of part of the solvent. When it is used for the production of pure water or for any of the evaporative processes associated with power generation, it is referred to as a *power-plant evaporator.*

Because of their importance in the process industry, evaporators, reboilers, and waste heat boilers are dealt with in more detail.

2.2.2.1. Chemical Evaporators

Chemical evaporators are common in many industries such as the chemical, paper and pulp, pharmaceutical, food, dairy product, fruit juice, and beverage industries. The final product is either concentrate, crystal, or powder. They can be classified further according to the type of construction and means of energy supply, as shown in Figure 20A and B.

Classification According to Construction. *Shell-and-tube-evaporators* are the most commonly used chemical evaporators. In vertical orientation, the exchanger is singlepass and usually has support plates on the shellside. Steam is generally used on the shellside, while the feed which needs to be concentrated flows through the tubeside. This is because the tubeside can readily meet the cleaning requirements. Horizontal shell-and-tube exchangers are either singlepass or two tube-pass units. A U-tube bundle is the most common type of bundle. To ensure no contamination from the shellside, a double tubesheet design is used.

The shell-and-tube construction for evaporators is the most versatile. They are used for high evaporation duties, with very viscous liquids, for crystallizing feeds, and for liquids containing large suspended particles. The major disadvantages of these evaporators are that the liquids have a relatively large residence time and high liquid holdup, which may be undesirable for rapid startup and shutdown with minimum wastage.

The recirculation in the evaporator may be natural or forced. In natural circulation, the flow is induced due to the density difference between the entering liquid and the evaporating liquid. In forced circulation, the liquid is circulated by a pump. Natural circulation evaporators are used for simpler evaporation requirements. Forced circulation evaporators are used for viscous and scaling liquids.

Natural circulation evaporators are used with low viscosity and low fouling liquids. The liquid circulation is maintained by the density difference between the hot and cold liquids; a pump is not used for liquid circulation, hence control of the circulation rate is not easy. Generally, a high headroom is needed to obtain the required circulation rate. Five types of natural circulation evaporators are commonly used: short-tube vertical, basket, long-tube vertical, rising film, and falling film type. For further details → Evaporation or [14–16].

Forced circulation evaporators are used when concentration is needed for thermally degradable, heavy fouling, viscous liquids, or when crystallization is required. For these applications, the temperature rise across the exchangers in one cycle is kept 2 – 3 °C (3 – 5 °F) or lower, which results in circulation rates 200 – 300 times greater than the evaporation rate. Fouling tendency is lower due to high velocities in the tube. However, pumping power is required for forced circulation. The cost of a forced circulation evaporator is generally higher than that of a natural circulation evaporator. The overall space requirement could be significantly reduced compared to a natural circulation evaporator for a given duty because of higher heat-transfer coefficients with forced circulation. By circulating the liquid with a pump, the flow rate is controlled independently of the heat transfer in the evaporator (evaporation rate), and as a result, the desired flow control could be achieved. Many types of forced circulation evaporators are in use. The most common ones are the submerged-tube type with either a vertical or horizontal tube bundle (heating element) totally submerged in

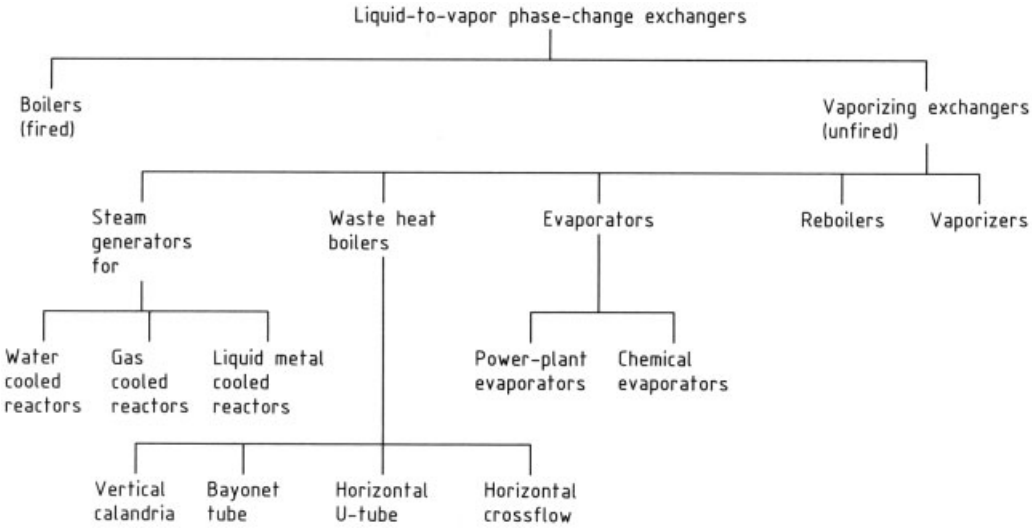

Figure 19. Classification of liquid-to-vapor phase-change exchangers

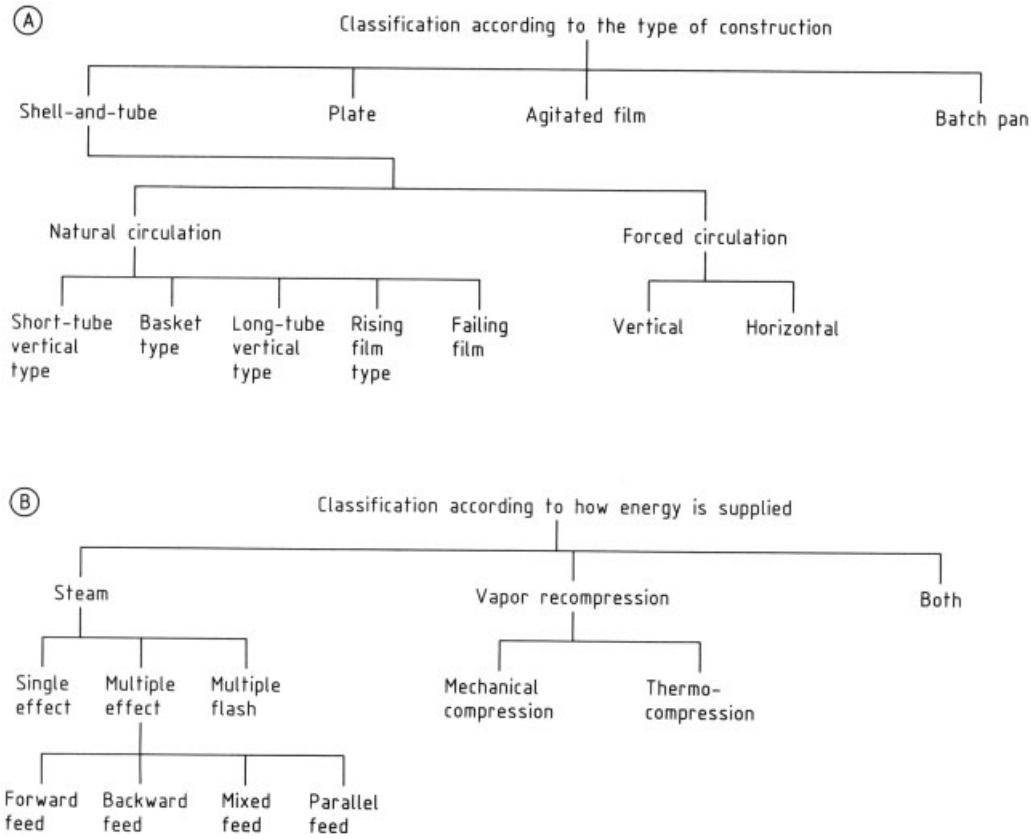

Figure 20. Classifiction of chemical evaporators

the liquid. For further details → Evaporation and [14], [15].

Plate evaporators (or plate and frame evaporators) are used when low to medium heat conditions are required. Four types of plate evaporators are available: falling film, rising/falling film, flash, and free-flow evaporators as described in [17].

The major advantages of plate evaporators are as follows: liquids have a low residence time as required for heat-sensitive liquids in pharmaceutical, fruit, and dairy products. Also, there is a low liquid holdup, which results in rapid startups and shutdowns with minimum wastage. They are relatively easy and fast to clean. Capacity can be increased or decreased by changing the number of plates.

The major disadvantage is that they cannot be used with very viscous liquids. Except for the flash evaporator, they are not suitable for liquids containing large suspended particles, or where crystallization may occur due to evaporation.

Agitated film-type evaporators, also referred to as wiped film or scraped film evaporators (see Fig. 21) are used when shell-and-tube or plate evaporators cannot be used. Some specific applications are: very viscous liquids, liquids with suspended particles, liquids with heavy scaling or foaming characteristics, and heat-sensitive liquids requiring a low residence time or high vacuums to reduce evaporation temperature.

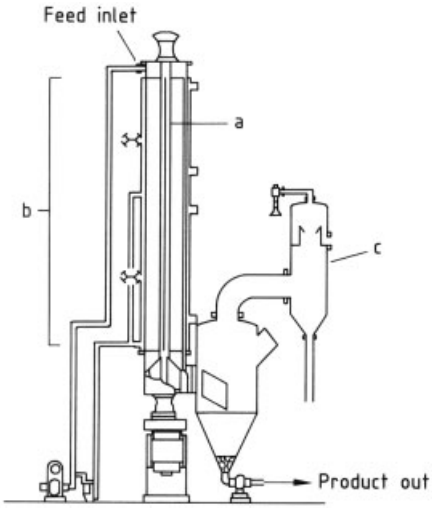

Figure 21. Wiped film evaporators
a) Rotor assembly; b) Vaporizing section; c) Condenser

Basically, this evaporator consists of a vertical cylindrical jacketed vessel which is heated primarily by steam (sometimes by oil), a rotor with blades, and a drive system to rotate the blades. A horizontal tapered cylindrical evaporator is used for special applications. The clearance between the vessel and rotor blade is either fixed (as in cylindrical units) or adjustable (as in tapered units). During operation, the liquid is fed from the top, and is transported down in a helical path by the action of the blade rotation and gravity. Evaporation of the solvent takes place in the thin film formed at the wall. The thin film yields high heat-transfer coefficients that are not possible with other types of evaporators. Agitated film evaporators are designed for operating pressures ranging from vacuums to 1 MPa (150 psig), operating temperatures up to 340 °C (650 °F), surface area up to 40 m^2 (430 ft^2), and liquid viscosities up to 50 Pa s (50×10^3 cp, although special units up to 1000 Pa s (1×10^6 cp) are available. Since the heat-transfer surface area is small, the heating medium is generally at a high temperature to obtain reasonable evaporation rates.

The major benefits are: extremely low residence time (in seconds), wide product viscosity range (outlet : inlet viscosity ratio = 1000 or higher), minimum fouling of heat-transfer surface due to wiping action, high recovery rates due to "squeezing" of residues, and minimum back-mixing.

Some of the major applications are: concentration of food products (dairy concentrates, fruit juices, tomato pastes, etc.), heat sensitive tobacco products, re-refining of waste lube oil, and recovery of wash solvents in paint and varnish manufacture.

The capital cost of the agitated film evaporator is high. Also the moving parts, blades, and rotor need periodic maintenance.

Batch pan is one of the oldest methods of concentration. A spherical or a spherical cylindrical vessel is either jacketed or has internal coils for heat transfer. Because of heat transfer by natural convection in the vessel, the heat-transfer coefficient is generally low; the surface area is also limited. Hence, the batch pan has limited evaporation capacity. The residence time of the product is generally large, in hours. Hence, for heat sensitive liquids, the batch must be either operated at low temperatures or in vacuum.

Batch pans are now used only in limited applications, such as for the concentration of jams and jellies that contain whole fruits, in the pharmaceutical industry, and in fermentation processes.

Classification According to How Energy is Supplied. Steam is generally used as the heating medium in the evaporator. The cost of steam or energy requirement of an evaporator may be significant depending upon the application. If the vapor produced in the evaporation process or the hot concentrated liquid could be used to evaporate the liquid to be concentrated, the amount of steam required could be significantly reduced, resulting in a considerable saving in the operating cost.

Steam economy is defined as unit mass of solvent evaporated per unit mass of steam used. The steam economy of the evaporator can be increased significantly in a number of ways. The most common method is the use of multiple evaporators arranged in series or alternatively; these are referred to as multiple effect evaporators. The vapor evolved from one stage of evaporation is thus used as the heating medium in the subsequent stage where boiling is carried out at reduced pressure. An alternative method of increasing steam economy requires that the steam generated in the first stage be recompressed to a sufficient pressure so that it could be used as the heating medium in the stage in which it is generated. The steam can be recompressed either by thermocompression or by mechanical compression. For further details → Evaporation or [14–16], [18].

2.2.2.2. Reboilers

A reboiler is a heat exchanger used to generate vapors (aqueous or nonaqueous) from a process liquid at the base of a distillation column. The vapors generated are used to provide the energy for vaporization in the distillation column. All reboilers considered here have a flow boiling process for the vaporizing fluid. As shown in Figure 22, they are classified according to how this flow is induced, whether by natural circulation or by forced circulation. *Natural circulation* can occur due to the hydrostatic pressure difference created by the difference in density between the liquid in the downcomer at the bottom of the distillation column and the two-phase mixture from the reboiler leading to the distillation column. *Forced circulation* is achieved by a pump in the downcomer line, and is used where the hydrostatic pressure difference is not adequate for the required circulation in a natural circulation reboiler.

Natural Circulation (Thermosiphon) Reboiler. In a *kettle reboiler,* the vapor at the exit of the boiling liquid side is dry or almost dry; in other thermosiphon reboilers, it is a two-phase (vapor – liquid) mixture. Proper design of the inlet and outlet piping to the reboiler is essential to maintain minimum pressure drop in the piping.

The kettle reboiler (TEMA K shell), as shown in Figure 23A, usually consists of horizontal bundle of heated U tubes (two tubepasses), circular in cross section, placed in an oversized shell; a multiple tube-pass floating head bundle may also be used. The tube bundle diameter ranges from 50 to 70 % of the shell diameter. The liquid to be vaporized enters from below and covers the tube bundle; the vapor occupies the upper space in the shell without the tubes. The large empty space in the shell acts as a vapor disengaging space, and almost dry vapor exits from the top nozzle. Depending upon the length of the kettle and a need to reduce liquid entrainment, one or more vapor nozzles are used. The liquid is fed at a rate greater than the vaporization rate so that low volatile components do not build up in the shell. The small holdup space beyond the weir is used to control the removal of excess liquid (which includes nonvolatiles); the nozzle in this space is used to drain the excess liquid. Typical weir heights exceed bundle heights by 50 – 150 mm (2 – 6 in). To maintain a high liquid circulation rate, the tube pitch-to-diameter ratio is kept between 1.5 and 2. Generally, the kettle reboiler is considered as a pool boiling device, however flow boiling prevails in the tube bundle.

The *internal* or *tube-in-column reboiler,* as shown in Figure 23B, is similar to the kettle reboiler without the shell, and is directly inserted into the bottoms reservoir of the distillation column. It is used for low heat duties and nonfouling liquids, thus eliminating the need for a shell and piping.

The *vertical thermosiphon reboiler,* as shown in Figure 23C, has a TEMA E shell with single-

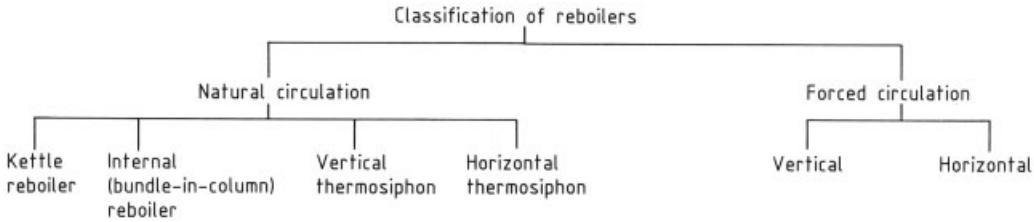

Figure 22. Classification of reboilers

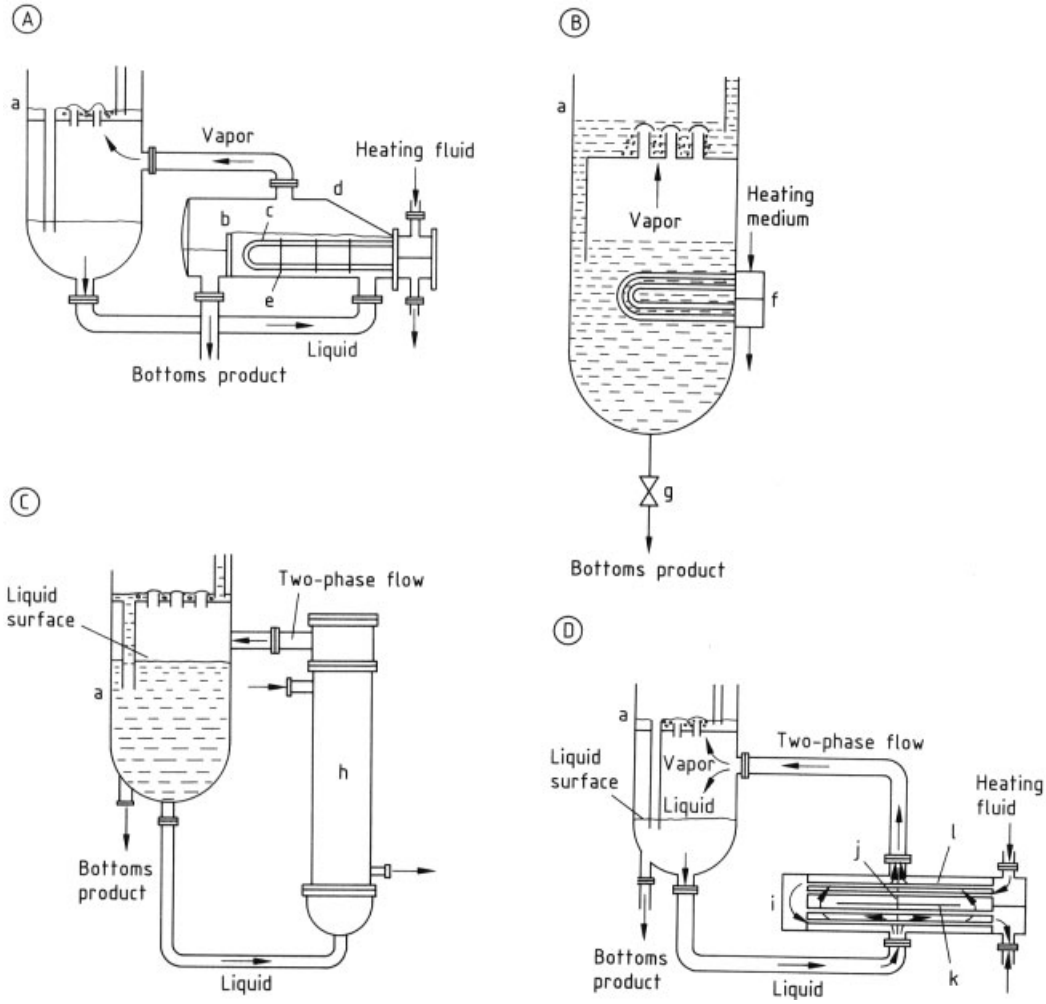

Figure 23. Reboiler installations
A) Kettle; B) Internal; C) Vertical thermosiphon; D) Horizontal thermosiphon
a) Distillation colomn; b) Weir; c) U-tube bundle; d) Kettle reboiler; e) Baffle support plates; f) Internal reboiler; g) Valve; h) Thermosiphon reboiler; i) Reboiler (G-type shell); j) Tube support plate; k) Horizontal baffle; l) Tube bundle

pass tubes in which boiling occurs. The tube diameter ranges from 25 to 50 mm (1 – 2 in) for low pressure operation with wide-boiling mixtures, and tube lengths are usually 2.4 – 3.7 m (8 – 12 ft). The outlet pipe is kept short with its flow area equal to the flow area of all tubes because of the high exit velocity. The inlet pipe area is kept small (up to 50 % of the tube area) to enable flow to be restricted in case of flow instability. The liquid level for non-vacuum operations is maintained to the upper (top) tubesheet. This means the liquid level in the base of the distillation column must be higher, and the column must be elevated above the ground. For vacuum applications, the liquid level may be dropped to 0.3 – 0.6 times the tube length. Vapor – liquid mixture leaves the reboiler and separation takes place in the column.

The horizontal thermosiphon reboiler, as shown in Figure 23D, has a TEMA X, G, H, or E shell, single or two-tube passes, and boiling takes place on the shell side. Flow is restricted in the inlet line for stable operation and good control of the reboiler.

Forced Circulation Reboilers. These reboilers are used for heavy fouling, viscous, or solid-bearing liquids, or when the vaporization rate is low. The pump in the liquid feed line to the reboiler is used to circulate the fluid through the system. The boiling fluid is on the tube side in most cases, but may be on the shell side in special applications. The reboiler may be horizontal or vertical. Generally, very little boiling (< 1 %) takes place within the reboiler due to the high circulation rate.

The foregoing four major types of reboilers are compared in Table 3 for several important process conditions [20]. This table may be used for prescreening the type of reboiler before going through a detailed design.

2.2.2.3. Waste Heat Boilers

Waste heat boilers are used in recovering heat from the flue gas, or for cooling and controlling of chemical process gases by generating steam. This steam can be used for industrial processes, power generation, or space heating. A combination boiler that operates both by firing fuel and by removing heat from a gas stream is also referred to as a waste heat boiler.

Waste heat boilers may be categorized as firetube and watertube units. In a *firetube boiler,* the hot gas flows through the tubes, and a water – steam mixture flows outside the tubes. In a *watertube boiler*, the water – steam mixture flows through the tubes and the hot gas flows outside the tubes. Firetube boilers are designed with natural circulation, and watertube boilers with natural or forced circulation.

Firetube boilers are used to recover or remove heat from relatively high pressure, high temperature process gases. For example, in the steam – methane reforming process, the high pressure gas leaves the reformer at several hundred kPa and at 870 – 980 °C (1600 – 1800 °F). The boiler resembles a shell-and-tube type heat exchanger. Since high pressure steam is generated on the shell side, the shell and tube wall thickness should be chosen properly because both the shell and tube are pressurized.

The firetube boilers are generally limited to steam pressure of under 6.9 MPa (1000 psig). They are usually more economical than the watertube boilers if the flow area required is less than 0.4 m^2 (4 ft^2) since construction is relatively uncomplicated. The heat-exchanger effectiveness of firetube boilers usually ranges from 65 to 75 % [21].

Watertube boilers are used to recover heat from flue gases at near-atmospheric pressure. Nearly all watertube boilers have finned tubes; plain tubes are used for those watertube boilers having gases above 760 °C (1400 °F). The heat-exchanger effectiveness for watertube boilers with fins usually ranges from 90 to 95 %, and that for watertube boilers without fins from 65 to 75 % [21].

In natural circulation boilers, thermosiphon action or natural circulation results due to the density difference between the steam – water mixture in the steaming area and the water in the downcomer. In this case, watertubes are either vertical or inclined. When horizontal tubes are used in the boiler with vertical gas flow, forced circulation is mandatory since buoyant force is not sufficient to sustain natural circulation. Forced circulation is achieved with pumps.

Table 3. Reboiler selection guide [20]

Process conditions	Kettle or internal	Reboiler type		
		Horizontal shell-side thermosiphon	Vertical tube-side thermosiphon	Forced flow
Operating pressure				
Moderate	E	G	B	E
Near critical	B – E	R	Rd	E
Deep vacuum	B	R	Rd	E
Design ΔT				
Moderate	E	G	B	E
Large	B	R	G – Rd	E
Small (mixture)	F	F	Rd	P
Very small (pure component)	B	F	P	P
Fouling				
Clean	G	G	G	E
Moderate	Rd	G	B	E
Heavy	P	Rd	B	G
Very heavy	P	P	Rd	B
Mixture boiling range				
Pure component	G	G	G	E
Narrow	G	G	B	E
Wide	F	G	B	E
Very wide, with viscous liquid	F – P	G – Rd	P	B

Category abbreviations: B, Best; G, Good operation; F, Fair operation, but better choice is possible; Rd, Risky unless carefully designed, but could be best choice in some cases; R, Risky because of insufficient data; P, Poor operation; E, Operable but unnecessarily expensive.

Natural circulation boilers cost less, are easier to construct, and require minimum maintenance. The circulation ratio (water mass flow : steam mass flow) in natural circulation boilers varies from 15 : 1 to 25 : 1; the circulation ratio for forced convection boilers is restricted to 5 : 1 – 7 : 1 to minimize pump size and power consumption.

Firetube boilers have either horizontal tubes or vertical tubes, and most of them have natural circulation of the steam – water mixture. Usually they are heated by process gas. A variety of firetube boilers are available; further design details are presented in [22], [23].

Watertube boilers with vertical and inclined tube have natural circulation, and those with horizontal tubes have forced circulation; the latter are less reliable and more expensive. For boilers with natural circulation, either unfinned or finned ordinary tubes are used or bayonet unfinned tubes are used. Again, many types of watertube natural circulation boilers are available; further design details can be found in [22], [23].

3. Heat Exchanger Design Methodology

A methodology for designing a new heat exchanger is illustrated in Figure 24 [24]. Major design considerations include:

1) Process and design specifications
2) Thermal and hydraulic design
3) Mechanical design including operation and maintenance considerations
4) Manufacturing considerations and cost
5) Trade-off factors and system-based optimization

There could be strong interactions and feedback among the aforementioned considerations as indicated by the double-sided arrows in Figure 24. One example of these interactions is as follows.

Consider a shell-and-tube exchanger design with heavy fouling. In the *process specification* module, geometry and material are properly selected to minimize fouling and corrosion, and to facilitate cleaning. The heat exchanger can be positioned and oriented such that on-location cleaning or easy removal for external cleaning is possible. In thermal – hydraulic design,

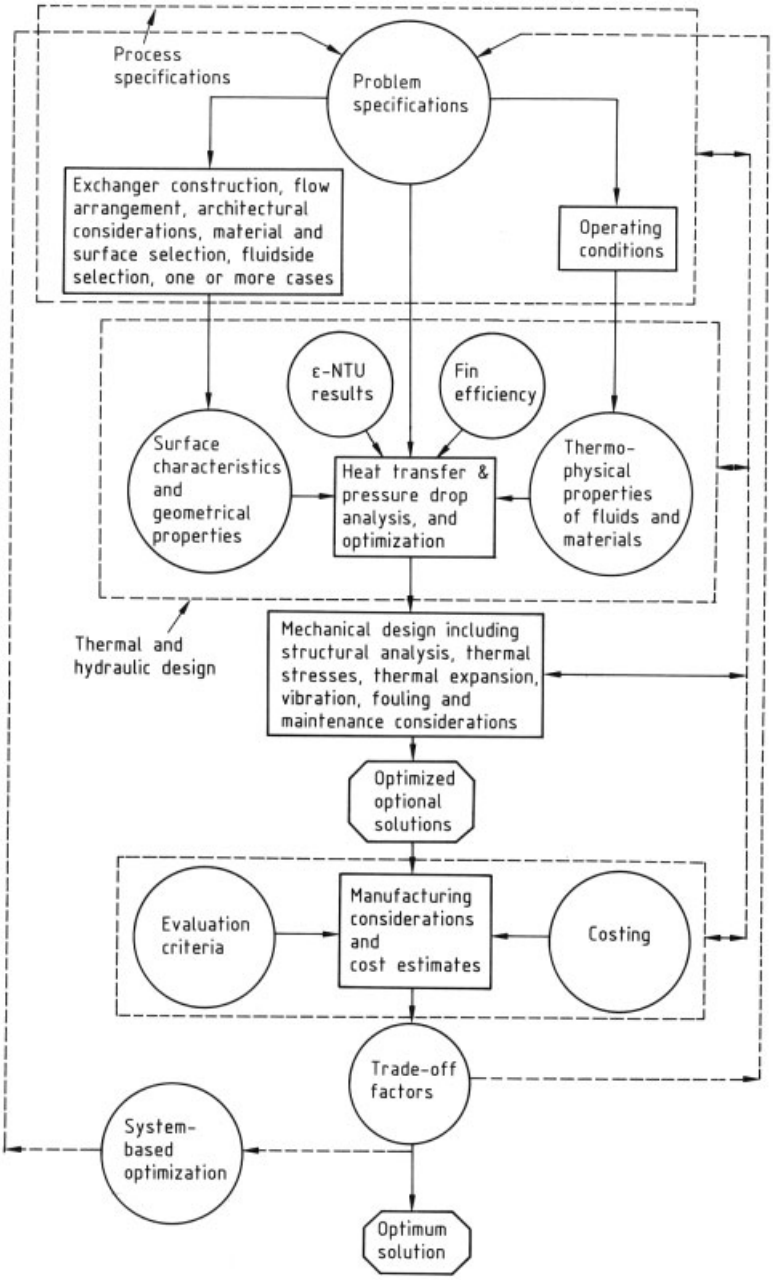

Figure 24. Heat exchanger design methodology

the fouling resistance, and hence the amount of increased surface area required, should be properly taken into account depending upon the cleaning cycle schedule. For the mechanical design, proper material gages should be selected for the desired life of the exchanger from the fouling and corrosion point of view. Also, the cleaning and maintenance schedule should be decided. The tube gage could, in turn, affect both heat transfer and pressure drop. The material and gage selection as well as the cleaning cycle will impact the fixed and operating costs of the exchanger. Two aluminum heat exchangers may be less expensive than one titanium exchanger for the same total service time when using ocean water for cooling. This affects the geometry, thermal-hydraulic design, and mechanical design.

The overall design methodology is quite complex because many qualitative judgments in addition to quantitative calculations must be introduced. Further details on this methodology can be found in [24]. A number of these quantitative and qualitative design aspects are covered in this article.

4. Criteria for Heat Exchanger Selection

The four basic heat exchanger construction types are tubular, plate-type, extended surface, and regenerative as shown in Figure 1C. Selection of a particular exchanger for a specific service depends upon the evaluation of many and often conflicting process requirements. A list of some important design variables and general selection criteria are summarized below for design evaluation of one or more types of exchangers.

4.1. General Selection Criteria

4.1.1. Design Variables

The following factors can affect the selection of a heat exchanger:

1) Pressure: very high or very low pressures will eliminate some types, e.g., plates.
2) Temperature: if the temperature limits of the usual construction materials are exceeded, special and expensive alloys may be required.
3) Fluid Characteristics:
   a) Corrosive fluids and those requiring minimum contamination will require special construction materials.
   b) Very toxic fluids must be prevented from leaking into the other fluid or to the surroundings thus requiring special design.
   c) Flammable or those fluids having explosive tendencies must be kept in leakproof designs.
   d) The phase condition of the fluid, i.e., whether liquid, vapor, or a mixture of phases, will affect the suitability of certain types.
   e) Physical properties such as viscosity affect the heat transfer and pressure drop, and can therefore affect the exchanger selection.
   f) Fouling characteristics of the fluid will also affect the selection of exchanger.
4) The potential size of the exchanger can eliminate certain designs.
5) Temperature difference and/or temperature approach influence the selection of shell type, number of passes or exchangers in series, or even exchanger types such as compact designs.
6) Pressure drop: when pressure drop is a critical parameter, those types that can be accurately calculated are chosen.
7) Historical: often the exchanger type is based on prior plant experience with regard to performance and maintenance. Certain types or plant standards may then dictate standard sizes and types.

4.1.2. General Selection Guidelines

Shell-and-tube exchangers are very versatile. Any combination of gas, liquid, condensing or evaporating fluid, or two-phase fluid can be used on one or both sides of a shell-and-tube two-fluid exchanger. Operating pressures can vary from high vacuums to very high pressures ($>$ 14 MPa, i.e., 2000 psig). Operating temperatures can vary from cryogenics to ca. 1000 °C (1800 °F), a limit for most special metals. Heavy fouling, corrosive, or viscous fluids can be accommodated. The cost of the heat-transfer surface is relatively high for shell-and-tube exchangers.

Plate exchangers are used with liquids and condensing or evaporating fluids. Because of relatively high pressure drops associated with the corrugated plates, these exchangers are seldom used for gases. The operating pressure is limited to ca. 2.5 MPa (360 psig) but pressures below 1.0 MPa (150 psig) are more usual. Similarly, the operating temperature is limited, due to the gaskets, to ca. 260 °C (500 °F) but operation is usually < 150 °C (300 °F). Heavy fouling fluids can be accommodated. Generally, the plate exchangers are competitive in price to shell-and-tube exchangers when stainless steel and more exotic materials are used.

Plate – fin exchangers are used with gas, liquid, condensing or evaporating fluid, or two-phase fluids. The operating pressures are usually limited to 700 kPa (100 psig), although some exchangers with operating pressures of ca. 8 MPa (1200 psig) are commercially available. These exchangers are available with operating temperatures of up to 840 °C (1550 °F). Because of the small flow-passage size, heavy fouling fluids and corrosive fluids cannot be used for this type of exchanger. The cost of heat-transfer surface per unit area is, however, considerably lower compared to the shell-and-tube and plate exchangers.

Tube – fin exchangers are used with gas, liquid, condensing or evaporating fluid, or two-phase fluids. The operating pressures on the tube side can be very high, as with shell-and-tube exchangers. The highest operating temperature is limited by the type of bonding, the materials employed, and material thickness; usually it ranges from 100 – 500 °C (200 – 900 °F). Heavy fouling fluids can be accommodated on the tube side. The cost of heat-transfer surface per unit area is generally somewhat higher than that for the plate – fin exchangers because of the higher cost of tubing.

Regenerators are exclusively used for gases (including humid air); liquids cannot be used because of the fluid carryover and mixing. Operating pressures are limited to ca. 400 kPa (60 psig). Operating temperatures are limited to ca. 800 °C (1450 °F) for metal rotary regenerators, and up to 2000 °C (3600 °F) or higher for fixed matrix ceramic regenerators. Very low fouling fluids are used for compact rotary regenerators. Heavy fouling and corrosive exhaust gases can be used in fixed matrix regenerators. The cost of heat-transfer surface per unit area is the lowest for compact rotary regenerators.

The above criteria are not stringent, because more than one type of heat exchanger can be designed for a given application. The sale price of a heat exchanger depends on many factors other than engineering material, labor, and tooling equipment costs. A relatively noncompact tubular heat exchanger may weigh more and occupy more space than a compact plate – fin exchanger for specified performance and durability considerations. However, if the tubular exchanger costs less than the plate – fin exchanger, and cost is the most important criterion, the tubular exchanger will probably be selected for that application.

4.2. Shell-and-Tube Exchangers

The major components of a shell-and-tube exchanger are the tubes, baffles, shell, front-end head, rear-end head, and tubesheets. A large number of geometrical variables or types are associated with each component and these may affect selection.

4.2.1. Tubes

Shell-and-tube exchangers have either plain or finned tubes. The outside diameter (*OD*) of the TEMA [2] standard plain tubes ranges from 6.4 to 50.8 mm (1/4 – 2 in). The most common sizes are 15.9, 19.1, and 25.4 mm (5/8, 3/4, and 1 in) tube *OD*. From the heat transfer point of view, smaller diameter tubes yield higher heat-transfer coefficients and result in a more compact exchanger. They are also more economical than large diameter tubes. However, larger diameter tubes are easier to clean and are more rugged: the common sizes available represent a compromise. For mechanical cleaning, the smallest practical size is 19.1 mm (3/4 in). For chemical cleaning, smaller sizes can be used provided that the tubes never plug completely.

Finned tubes are either low in height (up to 1.6 mm or 1/16 in) or high in height (15.9 – 19.1 mm or 0.63 – 0.75 in) with the fin density ranging from 630 to 1575 fins/m (16 – 40 fins/in). The surface area of such a fin tube is ca. 2.5 – 7 times that of a bare tube.

The tube wall thickness is generally identified by the Birmingham Wire Gage (BWG) or Stubs Gage. The nominal tube gages range from 10 to 21 BWG (3.40 – 0.81 mm or 0.134 – 0.032 in) [2]. Steel tubes are usually several gages thicker than the nonferrous or alloy tubes. Two types of wall thickness are available: minimum wall and nominal wall. The "minimal wall" has at least the thickness specified by the gage at the thinnest point (zero under-tolerance) and an over-tolerance of 20 %. The "nominal wall" has the wall thickness within ± 10 % of the specified gage. For the same specified gage, the minimum wall is thicker and heavier than the nominal wall by about one gage. The selection of the tube thickness depends upon many factors, such as pressure, corrosion, cost, tube-to-tubesheet joining process, and pressure drop through the tubes.

Tube lengths range from 1.83 to 6.10 m (6 – 20 ft) or longer, in 0.61 m (2 ft) increments. The TEMA [2] standard lengths are 2.44, 3.66, and 4.88 m (8, 12, and 16 ft) although longer exchangers are made with tubes up to 12.2 – 24.4 m (40 – 80 ft).

The number of tubes depends on the fluid flow rates and the available pressure drop. The number of tubes is selected so that the tubeside velocity for water and similar liquids ranges from 0.9 to 2.4 m/s (3 – 8 ft/s) and the shellside velocity from 0.6 to 1.5 m/s (2 – 5 ft/s). The lower velocity limit is to avoid excessive fouling, and the upper velocity limit to minimize erosion. When sand and silt are present, the minimum velocity is kept high enough to prevent settling.

The number of tube passes depends on the available pressure drop. Higher velocities in the tube result in higher heat-transfer coefficients, at the expense of increased pressure drop compared to that for low velocity. Therefore, if the increased pressure drop is acceptable, it is desirable to have fewer but longer tubes (reduced flow area and increased flow length). Long tubes are accommodated in a short shell exchanger by multiple tube passes. The number of tube passes in a shell generally ranges from one to eight. The standard design has one, two, or four tube passes. In multipass designs, an even number of passes is generally used; odd number passes are uncommon, and may result in mechanical and thermal problems in fabrication and operation. The open lanes between passes should be designed so that the shell fluid flows perpendicular to them to eliminate fluid bypass. Care should be also exercised to drain and vent the tube side in a multitube pass unit.

Tubes are fastened to tubesheets by mechanical rolling, butt welding, edge welding, explosion welding, or hydraulic expansion [25]. If one of the fluids is highly corrosive, toxic, or hazardous, leakage of shell fluid to tube fluid or vice versa is not acceptable, in which case a double tubesheet design would be used [25].

4.2.2. Tube Pitch and Layout

The selection of tube pitch is a compromise between a close pitch for increased shellside heat transfer and surface compactness, and an open pitch for decreased shellside plugging and ease in shellside cleaning. In most shell-and-tube exchangers, the ratio of tube pitch to tube outside diameter varies from 1.25 to 2. The minimum value is restricted to 1.25 because the tubesheet ligament may become too weak for proper rolling of the tubes and may cause leaky joints. The recommended width of the ligaments depends on the tube diameter, thickness, and pitch; the values are provided by TEMA [2]. Typically, the minimum ligament width is 6.4 mm (1/4 in) when mechanical cleaning is required, and 3.2 mm (1/8 in) when no fouling is expected, although in the latter case also 6.4 mm ligament width is desired to maintain sufficient strength.

Standard tube layouts are shown in Figure 25. For identical tube pitch and flow rates, the tube layouts in decreasing order of shellside heat-transfer coefficient and pressure drop are 30°, 45°, 60°, and 90°. Thus the 90° layout will have the lowest heat-transfer coefficient and the lowest pressure drop.

The square pitch (45° or 90°) is used when jet or mechanical cleaning is necessary on the shell side; here, a minimum cleaning lane of 6.35 mm (1/4 in) is provided. The square pitch is generally not used in the fixed tubesheet design because no cleaning is needed. The triangular pitch provides a more compact arrangement, usually resulting in a smaller shell, and the strongest header sheet for a specified shellside flow area. Hence, it is preferred when the operating pressure difference between two fluids is large. If properly designed, it can be cleaned from six di-

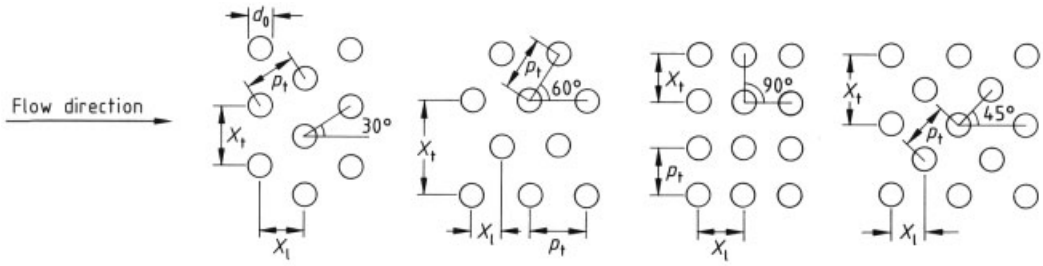

	30° triangular staggered array	60° rotated triangular staggered array	90° square inline array	45° rotated square staggered array
Transverse tube pitch X_t	p_t	$\sqrt{3}\ p_t$	p_t	$\sqrt{2}\ p_t$
Longitudinal tube pitch X_l	$(\sqrt{3}/2)p_t$	$p_t/2$	p_t	$p_t/\sqrt{2}$
Ratio of minimum free flow area to frontal area, $A_0/A_{fr}=\sigma$	$(p_t-d_0)/p_t$	$(\sqrt{3}\ p_t-d_0)/\sqrt{3}\ p_t$ for $p_t/d_0 \geq 3.732$ $2(p_t-d_0)/\sqrt{3}\ p_t$ for $p_t/d_0 \leq 3.732$	$(p_t-d_0)/p_t$	$(\sqrt{2}\ p_t-d_0)/\sqrt{2}\ p_t$ for $p_t/d_0 \geq 1.707$ $2(p_t-d_0)/\sqrt{2}\ p_t$ for $p_t/d_0 \leq 1.707$

Figure 25. Nomenclature and geometric properties of tube banks common in shell-and-tube exchangers

rections instead of four, as in the square pitch arrangement. Where mechanical cleaning is required, the 45° layout is preferred for laminar or turbulent flow of a single-phase fluid, and for condensing fluid on the shell side. If the pressure drop is constrained on the shell side, the 90° layout is used for turbulent flow. Where boiling is necessary, the 90° layout, which provides vapor escape lanes, is preferred. However, if mechanical cleaning is not required, the 30° layout is preferred for single-phase laminar or turbulent flow and condensing involving a high ΔT range (a mixture of condensibles). The 60° layout is preferred for condensing involving a high ΔT range (generally pure vapor condensation) and for boiling applications. Horizontal tube bundles are used for shellside condensation or vaporization.

4.2.3. Baffles

Baffles may be classified into transverse and longitudinal types. The purpose of longitudinal baffles is to control the overall flow direction of the shell fluid. For example, F, G, and H shells have longitudinal baffles (see Fig. 4). The transverse baffles may be classified as plate baffles and rod (or strip) baffles. The plate baffles are used to support the tubes, to direct the fluid in the tube bundle at approximately 90° to the tubes, and to increase the turbulence of the shell fluid. Two types of plate baffles are shown in Figure 26, segmented and disk-and-doughnut. The rod baffles are used to support the tubes and to increase the turbulence of the shell fluid and are shown in Figure 27. The flow in a rod-baffled heat exchanger is parallel to the tubes, and the problem of flow-induced tube vibration is virtually eliminated by the baffle support of the tubes. The choice of baffle type, spacing, and cut are largely determined by flow rate, allowable pressure drop, tube support, and flow-induced vibration.

Plate Baffles. The two types of plate baffles are either segmental or disk-and-doughnut (Fig. 26). The single- and double-segmental baffles are most frequently used. The *single-segmental baffle* is generally simply referred to as a segmental baffle. The practical range of

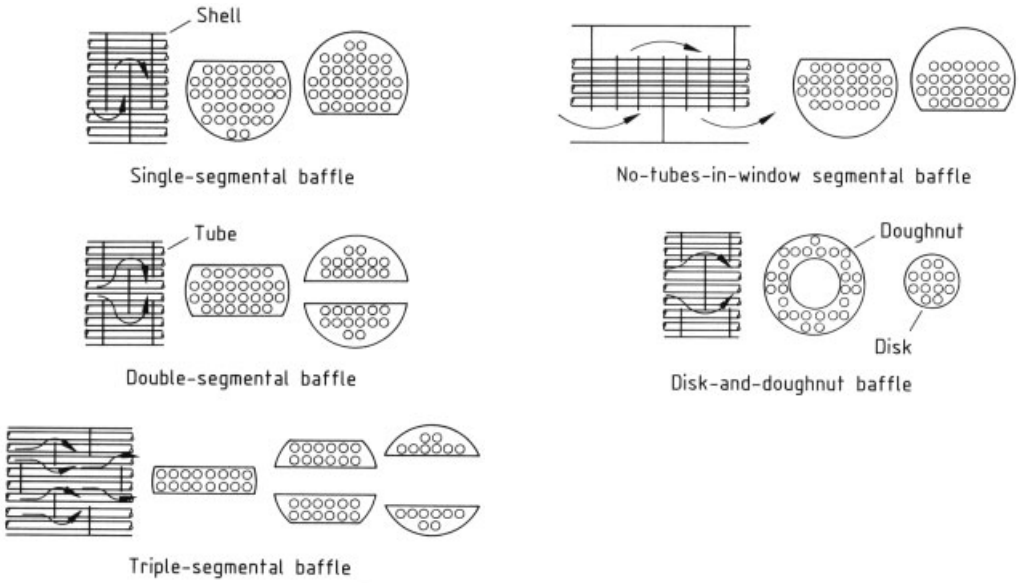

Figure 26. Plate baffle types

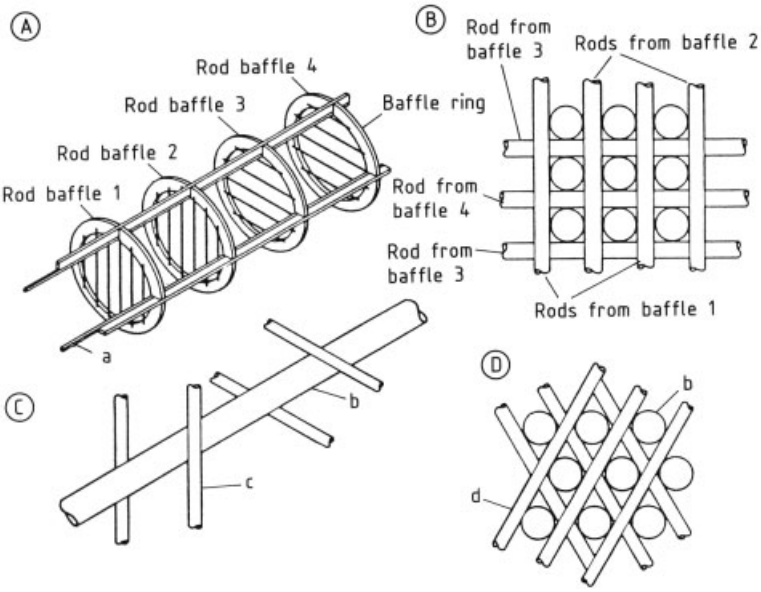

Figure 27. Rod baffle supports
A) Four rod baffles supported by skid bars (tubes not shown); B) A tube supported by four rods at 90 degree angle around the periphery; C) A square layout of tubes with rods; D) A triangular layout of tubes with rods
a) Skid bar; b) Tube; c) Rod; d) Rod or bar

single-segmental baffle spacing is one fifth to one shell diameter, though the optimum could be 40–50 % of the shell diameter. The minimum spacing is 50.8 mm (2 in) or one fifth of the shell diameter, whichever is larger. A spacing of 50.8 mm (2 in) is required for cleaning the bundle. Spacings closer than one fifth of the shell diameter provide added leakage (see Section 8.2.1) which nullifies the heat-transfer advantage of closer spacings. If the foregoing limits on the baffle spacing do not satisfy other design constraints such as $\Delta p_{\max}$ or tube vibration, no-tubes-in-window or pure crossflow design should be tried. The baffles are generally spaced between the nozzles, and do not cover the entire length of the tubes. The inlet and outlet baffle spacing is generally larger than the "central" baffle spacing due to the nozzle locations in the shell.

The segmental baffle is a circular disk (with baffle holes) with a segment removed. The baffle cuts, expressed as the percentage of the shell inside diameter, vary from 20 % to 49 %, the most common being 20 – 25 %. The baffle cut and spacing should be designed such that the flow velocity is approximately the same for the cross flow and window flow. At large baffle spacings, the percent cut may approach 45 – 49 % in order to avoid excessive pressure drop across the windows compared to the bundle. Large or small spacings coupled with large baffle cuts are undesirable because of the increased potential of fouling associated with stagnant flow areas. If fouling is of prime concern, the baffle cut should be kept < 25 % to maintain high velocity through the window zone. The direction of the baffle cut is selected as either horizontal (see Fig. 26) or vertical for a single-phase fluid (liquid or gas) on the shell side. For a very viscous liquid, the direction of the baffle cut should be horizontal for better mixing. The direction of the baffle cut is selected as vertical for the following shellside applications: condensation (for better drainage), evaporation/boiling (to promote more uniform flow), entrained solids in liquid (to provide the least interference for solids to fall out), and multishell pass exchangers (includes F shell).

Even though one of the major functions of the plate baffle is to include cross flow (flow normal to the tubes) for improved heat-transfer performance, this objective is only approximated. Various clearances are required for manufacturing. Fluid can leak through these clearance passages and reduce the heat-transfer effectiveness. Three clearances associated with a plate baffled shell-and-tube exchanger are tube-to-bundle hole clearance, baffle-to-shell clearance, and bundle-to-shell clearance.

The *tube-to-baffle hole* clearance should be kept at a minimum to reduce tube vibration and resultant damage as well as to minimize the A leakage stream (see Section 8.2.1). The clearance in terms of the difference between the baffle hole *ID* and tube *OD* is 0.40 mm (1/64 in) for unsupported tube length $L > 0.91$ m (36 in) or for the tube *OD* 31.8 mm (1.25 in) or smaller; this clearance is 0.80 m (1/32 in) for the unsupported tube length $L < 0.91$ mm (36 in) or for the tube $OD > 31.8$ mm (1.25 in). The *baffle-to-shell* clearance should be kept to a minimum to minimize the E leakage stream (see Section 8.2.1). The recommended maximum diametral clearance (shell *ID* minus baffle *OD*) varies from 2.5 mm (0.100 in) for 0.20 m (8 in) shell *ID* to 7.6 mm (0.300 in) for 1.52 m (60 in) shell *ID* [2]. From the strength and weldability viewpoints, the tube holes are kept some distance away from the edge of the tubesheet (shell *ID*) and hence from the baffle edge resulting in *bundle-to-shell* bypass stream C flowing around the edge (see Section 8.2.1). Sealing strips are used to block this space and force the bypass stream to flow through the tubes. The use of sealing strips is recommended every five to seven row of tubes in the bypass stream direction provided that the design is checked for flow-induced vibration of tubes near the sealing strips.

If the primary function of baffles is to support the tubes, they are referred to as "support plates". They may be thicker than the baffle, have less tube-to-baffle hole clearance, and provide greater stiffness to the bundle.

The *double-segmental baffle,* Figure 26, also referred to as a strip baffle, provides lower shellside pressure drop than that for the single-segmental baffle for the same unsupported tube span. Hence an exchanger with this type of baffle can handle larger fluid flows on the shell side. *Multi-segmental* baffles have a strong parallel flow component, provide a lower pressure drop, and permit closer tube support to prevent tube vibrations.

The lower pressure drop results in a large baffle spacing. Since the tubes in the window zone

are supported at a distance of two or more times the baffle spacing, they are most susceptible to vibration. To eliminate the susceptibility of tube vibrations and to reduce the shellside pressure drop, the tubes in the window zone are removed and support plates are used to reduce the unsupported span of the remaining tubes. The resultant design is referred to as the segmental baffle with no-tubes-in-window and support plates as shown in Figure 26. The low velocity regions in the baffle corners do not exist, resulting in good flow characteristics and less fouling. Thus the loss of heat-transfer surface in the window region is partially compensated for. However, the shell size must be increased to compensate for the loss in surface area in the window zone. If the shellside operating pressure is high, this no-tubes-in-window design is very expensive.

The *disk-and-doughnut baffle* is made up of alternate disk- and doughnut-shaped baffles as shown in Figure 26. Generally, the disk diameter is larger than the half shell diameter, and the diameter of the hole of the "doughnut" is smaller than the half shell diameter. This baffle design provides a lower pressure drop compared to a single-segmental baffle for the same unsupported tube span, and eliminates the tube bundle-to-shell bypass stream C. The disadvantages of this design are: (1) all tie rods to hold the baffles are within the tube bundle, and (2) the central tubes are supported by the disk baffles, which in turn are supported only by tubes in the overlap region of the larger diameter disk over the doughnut hole. This design is used in nuclear shell-and-tube exchangers.

Rod Baffles. A rod baffle is made up of parallel rods (a rod matrix) mounted on a baffle ring. The rod diameter is equal to the tube spacing (zero clearance), so that the baffle touches and constrains all tubes on four sides as shown in Figure 27. Thus, in a square layout, two vertical baffles constrain all tubes from two sides. The following two horizontal baffles constrain all tubes from the remaining two sides (90° to the first side). A set of four rod baffles (two vertical and two horizontal) is repeated in the exchanger. The rod baffle arrangements for both square and triangular tube layouts are shown in Figure 27. The baffle spacing is generally kept at 0.152 m (6 in) for single-phase applications and 0.305 m (12 in) for condensers and reboilers. Generally the tube diameters and tube pitches are selected so that the rod diameter is 4.8 mm (3/16 in) or 6.4 mm (1/4 in).

The rod baffle exchanger has several advantages over the plate-baffled exchanger, as follows: (1) It eliminates flow-induced tube vibrations since the tubes are rigidly supported at four successive points. (2) The pressure drop on the shell side is about one-half of that with a double-segmental baffle at the same flow rate and heat transfer rate. Alternatively, the rod baffle exchanger will result in a smaller unit for the same heat transfer and pressure drop. (3) There are no stagnant flow areas with the rod baffles, resulting in less fouling and corrosion, and improved heat transfer compared to that for a plate baffle exchanger. However, the rod baffle exchanger will require longer tubes of smaller diameter. If the tubeside fluid is controlling and has a pressure drop limitation, the rod baffle exchanger may not be applicable.

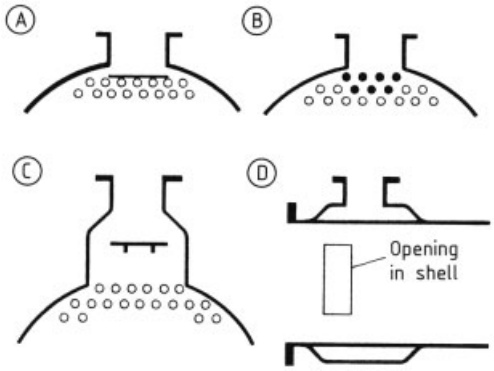

Figure 28. Impingement protection designs
A) Impingement plate; B) Impingement rods; C) Nozzle impingement baffle; D) Annular distributor [26]

Impingement Baffles. The most common cause of tube failure is the improper location and size of the impingement plate. Impingement baffles or plates are generally used in the shellside just below the inlet nozzle. They protect the tubes in the top row from erosion, cavitation, and vibration due to the impact of high velocity fluid jet from the nozzle to the tubes. One of the most common forms of this baffle is a solid circular plate located under the inlet nozzle just in front of the first tube row as shown in Figure 28A and C. The location of this baffle within the shell is critical in order to minimize

the associated pressure drop and high escape velocity. For this purpose, adequate areas should be provided both between the nozzle and plate and between the plate and tube bundle. This can be achieved by either omitting some tubes from the circular bundle as shown in Figure 28A or the nozzle contains an expanded section as shown in Figure 28C. Alternatively, impingement rods (Fig. 28B), dummy tubes, or thick-walled tubes are used. The annular distributor of Figure 28D is an excellent but expensive design. It permits higher velocities and any orientation of the nozzles.

4.2.4. Shells

Seven types of shells, as classified by the TEMA [2], are shown in Figure 4. The E shell is the most common as it is inexpensive and simple. In this shell, the shell fluid enters at one end of the shell and leaves at the other end, i.e., there is one pass on the shell side. The tubes may have single or multiple passes and are supported by transverse plate baffles. This shell is the most common for single-phase shell fluid applications. Multiple passes on the tube side and a single shell pass reduce the exchanger effectiveness or F factor over a single-pass arrangement. If it is too low, two E shells in series may be used to increase the mean temperature difference and the heat exchanger effectiveness ε.

The F shell has a longitudinal baffle, resulting in two shell passes and a pure counterflow arrangement. Although ideally this is a desirable flow arrangement, the F shell is rarely used in practice because there are many problems associated with the design. In particular, there will be heat transfer through the longitudinal baffle from the hot side to the cold side of the shell fluid; also if the longitudinal baffle is not continuously welded to the shell, or if seals are not provided between the longitudinal baffle and the shell, there will be leakage from the high pressure to the low pressure shell side. Both these factors will reduce the mean temperature difference and the exchanger effectiveness more than the gain by using the pure counterflow arrangement. It is also difficult to remove or replace the tube bundle. With the F shell, there are also additional problems of fabrication and maintenance. Hence, if one needs to increase the exchanger effectiveness, multiple shells in series are preferred over the F shell.

The G and H shells are related to the F shell, having variants of the longitudinal baffle. The double-split flow H shell is similar to the G shell, but with two inlet and two outlet nozzles (and two horizontal baffles) to accommodate high inlet velocities. The G and H shells are seldom used for shellside single-phase applications, since there is no advantage over E or X shells. They are used as horizontal thermosiphon reboilers, condensers, and other phase-change applications. The longitudinal baffle serves to prevent flashing out of the lighter components of the shell fluid, helps flushing out of noncondensables, provides increased mixing, and helps to distribute the flow. Generally, ΔT and Δp across the longitudinal baffles are small, and the heat transfer across the baffle and flow leakages from the sides have an insignificant influence on the performance. The H shell approaches the crossflow nozzle arrangement of the X shell, and it usually has low shellside Δp compared to that for the E, F, and G shells.

The divided flow J shell has two inlet and one outlet, or one inlet and two outlet nozzles. It has approximately one eighth the pressure drop of a comparable E shell, and is therefore used for low pressure drop applications such as a condenser in vacuum. For a condensing shell fluid, the J shell is used with two inlets for the gas phase and one central outlet for the condensates and leftover gases.

The K shell is used for partially vaporizing the shell fluid. It is used as a kettle reboiler in the process industry and as a flooded chiller in the refrigeration industry (see Section 2.2.2.2).

For a given flow rate and surface area, the crossflow X shell has the lowest shellside pressure drop compared to all other shell configurations except the K shell. Hence, it is used for gas heating and cooling with and without finned tubes and for vacuum condensation. It is also used when shell flows are large. No transverse baffles are used in the X shell; however, support plates are used to suppress the flow-induced vibrations. Flow distributions on the shell side can be a serious problem unless a proper provision has been made to feed the fluid uniformly at the inlet. This can be achieved by a single nozzle, multiple nozzles, or by keeping a clear lane

along the length of the shell near to the nozzle inlet as shown in Figure 29.

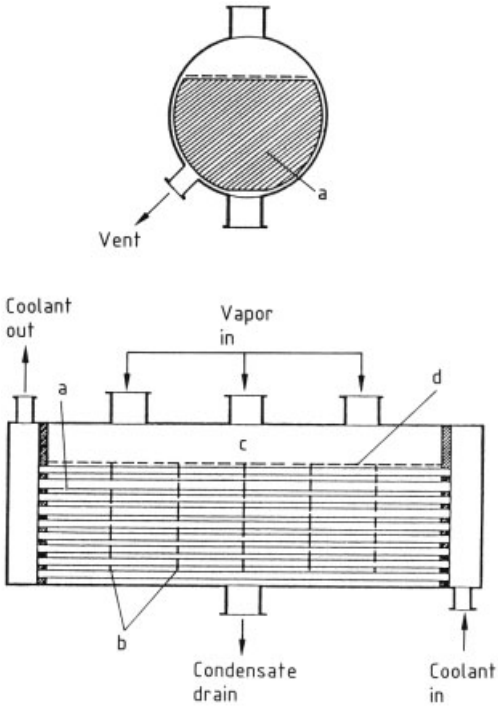

Figure 29. TEMA X shell with multiple nozzles with top tube rows omitted
a) Tube bundle; b) Tube support plates (baffles); c) Vapor distribution space; d) Perforated distributor plate

The type of shells described in Figure 4 has either one or two shell passes in one shell. The cost of the shell exceeds the cost of the tubes by far, hence a design should accommodate the required heat-transfer surface in one shell. Multipassing on the shell side with longitudinal baffles will reduce the flow area per pass compared to a single pass on the shell side in a single shell; this will result in higher shellside pressure drop. Hence instead, multiple shells in series are used for the following reasons: (1) For exchangers requiring high effectiveness, multipassing is the only alternative. (2) For part-load operations and where a spare bundle is essential, multiple shells (which may be of smaller size) will result in an economical operation. (3) Shipping and handling may dictate restrictions on the overall size or weight of the unit resulting in multiple shells for an application. In heat recovery trains and some other applications, up to six shells in series are commonly used. The limitation on the number of shells in such applications is the pressure drop limit on one of the streams.

4.2.5. Front-End Heads

The front-end head is stationary, while the rear-end head could be either stationary or floating depending upon the thermal stresses between the tubes and the shell. The major criteria for the selection of front-end and rear-end heads are the thermal stresses, operating pressures, cleanability, hazards, and cost.

The front-end heads are primarily of two types, the channels and the bonnet, as shown in Figure 4. The bonnet head B is cast in one piece and has either a side-entering or end-entering nozzle. Although the bonnet head is less expensive, the inspection and maintenance requires a breaking of the pipe joints and a removal of the bonnet. Hence, the bonnet head is generally used for clean tubeside fluids. The channel head can be removable as in the A head, or can be integral with the tubesheet as in C and N heads. There is a removable channel cover in these front-end heads to allow inspection and maintenance without disturbing the piping. The nozzles are side-entering in these types. The shell is welded onto the N head, but is flanged to the C head. In the N head, no joint exists between the channel and the tubesheet and between the tubesheet and the shell, thus eliminating any leakage between the shell and the tubes. The D head has a special high pressure closure and is used when pressures of 2100 kPa (300 psig) and higher are necessary.

4.2.6. Rear-End Heads

In a shell-and-tube exchanger, the shell is at a temperature different from the tubes because of the heat transfer between the shell and tube fluids. These would result in a differential thermal expansion and stresses in the shell, tubes, and the tubesheet. If proper provision is not made, buckling of the shell or tubes would result, or tubes would be pulled apart or out of the tubesheet. The provision for the differential thermal expansion is made in the rear-end heads. Basically, they may be categorized as fixed or floating rear-end heads depending on whether provisions are absent or present for differential thermal expansion, respectively.

A heat exchanger with fixed rear-end heads L, M, or N has a fixed tubesheet on that side. Hence, the overall design is rigid. The tube bundle-to-shell clearance is the smallest compared to other designs, thus minimizing the bundle-to-shell bypass stream C. Any number of tube passes can be employed. The major disadvantages of the fixed tubesheet exchanger are described as follows: since there is no relief of thermal stresses between the tubes and the shell, the fixed tubesheet exchangers are used where relatively low temperatures ($\leq 315\,^{\circ}$C or $600\,^{\circ}$F) coupled with low pressures (≤ 2100 kPa or 300 psig) are required, with the inlet temperature difference between the two fluids being $\lesssim 50$–$60\,^{\circ}$C ($100\,^{\circ}$F). With expansion bellows, the inlet temperature difference between two fluids can be increased to ca. 80–$90\,^{\circ}$C ($150\,^{\circ}$F). However, bellows are generally not available for all shell sizes and operating pressures. In fact, they are uneconomical for high pressures (> 4150 kPa or 600 psig). The fluid should be clean on the shell side or the shell side should be cleanable by the chemical cleaning methods. If it is essential to prevent possible leakage between the two fluid streams, double tubesheets are used. The fixed tubesheet exchanger is a low cost unit after the U-tube exchanger.

The differential thermal expansion can be accommodated by a floating rear-end head in which the tubes expand freely within the shell thus eliminating thermal stresses. The tube bundle can also be removed for cleaning the shell side. Basically, there are three types of floating rear-end heads: U-tube head, internal floating heads (pull-through or split ring heads), and outside packed floating heads.

In the *U-tube bundle,* the thermal stresses are almost eliminated due to free expansion of the U tubes, and the rear-end head has an integral cover which is low in cost compared to other rear-end heads. The exchanger construction is simple, having only one tubesheet and no expansion joints, and hence it is the lowest cost design. The tube bundle can be removed for shellside cleaning or replaced; however, it is difficult to remove a U-tube from the bundle except in the outer row, and it is also difficult to mechanically clean the tubeside bends. So a U-tube exchanger is used with clean fluids inside the tubes unless the tube side can be chemically cleaned. Flow-induced vibration can also be a problem for the tubes in the bend region in the outermost row because of a long unsupported span, particularly in large diameter bundles.

The next simplest floating head is the *pull-through head* T as shown in Figure 4 with its own bonnet head. The tube bundle can easily be removed from the shell by first removing the front-end head. Individual tubes or the tube bundle can also be replaced if required. Because of the floating head bonnet flange and bolt circle, many tubes are omitted from the tube bundle near the shell. This results in the largest bundle-to-shell clearance or a significant bundle-to-shell bypass stream C. In order not to reduce the exchanger performance, sealing strips (or dummy tubes or tie rods) in the bypass area are essential. However, localized high velocities near the sealing strips could cause flow-induced tube vibration; hence, proper care must be exercised for the design. Since this design has the least number of tubes in a bundle for a given shell diameter compared to other floating head designs, the shell diameter is somewhat larger to accommodate a required amount of surface area. One of the ideal applications of the T head design is in the kettle reboiler for which there is ample space on the shell side, and the flow bypass stream C is of no concern (see Fig. 3E, AKT).

The large bundle-to-shell clearance can be minimized by bolting the floating head bonnet to a split backing ring (flange) as shown in Figure 3A (AES). It is referred to as an S rear-end head. The shell cover over the tube floating head has a diameter larger than the shell. As a result, the bundle-to-shell clearances are reasonable and sealing strips are generally not required. However, both ends of the exchanger must be disassembled for cleaning and maintenance. In both S and T heads, the shell fluid is held tightly to prevent the leakage to the outside. However, internal leakage is possible due to the failure of an internal hidden gasket, and is not easily detectable. The T head has more positive gasketing between the two streams than the S head. Both S and T head configurations are used for the tubeside multipass exchangers; the single-pass construction is not feasible if the advantages of the positive sealing of S and T heads are to be retained. The cost of S and T head designs is relatively high compared to the U tube or fixed tubesheet units. The cost for S heads is higher than for T heads. The split backing ring

floating head is extensively used in the petroleum industry for moderate operating pressures and temperatures. For very high operating pressures and temperatures, the S head design has a special test ring [2].

In the *outside packed floating head P* design of Figures 3C (AEP) and 4, the stuffing box provides a seal against the skirt of the floating head and prevents shellside fluid leakage to the outside. This skirt (and the tube bundle) is free to move axially against the seal to account for thermal expansion. A split ring flange near the end of the skirt seals the back end of the chamber. Because of the specific design of this floating head, any leak (either from the shell side or the tube side) at the gaskets is to the outside. Hence, the P head is generally not used with very toxic fluids. In this design, the bundle-to-shell clearance is large (ca. 38 mm or 1.5 in); as a result, sealing strips are required. The P head exchanger is more expensive that the W head exchanger.

The *packed floating head with lantern ring* or W head is shown in Figures 3F (AJW) and 4. Here a lantern ring rests on the machine surface of the tubesheet and provides an effective seal between the shellside and tubeside flanges. Vents are usually provided in the lantern ring to help locate any leaks in the seals before the shellside and tubeside fluids are allowed to mix. Where severe fluctuations in the heat duty are expected, the W head is not used since the packing could be loosened. Although single-pass design on the tubeside is possible, generally an even number of tube passes are used. The W head exchanger is the least expensive design of all floating heads. Although its cost is higher than the U tube bundle, this is offset by the accessibility to the tube ends for cleaning and repair; consequently, this design is sometimes used in petrochemical and process industries.

A large number of combinations of the front-end and rear-end heads with different shell types of Figure 4 are possible depending upon the application and the manufacturer. Some common types of combinations are shown in Figure 3.

4.2.7. Nozzles

Nozzle sizes are related to the piping size, the exchanger shell design (Fig. 28), and the escape flow area into the tube bundle. This escape area should have $\varrho v^2 < 6000\,\mathrm{kg\ m^{-1}\,s^{-2}}$ ($4000\,\mathrm{lbm\ ft^{-1}\,s^{-2}}$). It is the total free area between a nozzle and the projected area on the tube bundle. Sometimes tubes may be removed to obtain sufficient area. This sizing can not be made until some mechanical drawings have been made. The location of the nozzle with respect to the shell flanges is determined by pressure vessel codes.

5. Basic Heat-Transfer and Pressure-Drop Analysis

5.1. Basic Equations and Definitions of Heat-Transfer Analysis

In order to determine heat transfer and inlet or outlet temperatures in a heat exchanger, the basic equations used are the energy (and mass) conservation and the rate equations, either in a differential or integrated form. Referring to Figure 30, the temperature drop $\mathrm{d}T_\mathrm{h}$ of the hot fluid and the temperature rise $\mathrm{d}T_\mathrm{c}$ of the cold fluid across a differential element $\mathrm{d}x$ for an exchanger having any arbitrary flow arrangement are given by

$$\mathrm{d}q = q''\,\mathrm{d}A = -C_\mathrm{h}\,\mathrm{d}T_\mathrm{h} = \pm C_\mathrm{c}\,\mathrm{d}T_\mathrm{c} \tag{1}$$

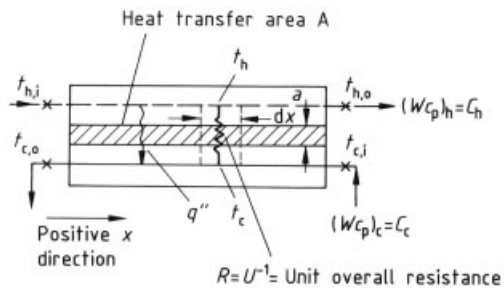

Figure 30. Nomenclature for heat exchanger variables

Here $\mathrm{d}q$ is the heat transfer from the hot fluid to the cold fluid across the surface area $\mathrm{d}A$; C_h and C_c are the heat capacity rates for the hot and cold fluids, and the $\pm$ sign depends upon whether $\mathrm{d}T_\mathrm{c}$ is increasing or decreasing with increasing $\mathrm{d}A$.

The overall rate equation in a differential form (on a local basis) is

$$\mathrm{d}q = q''\,\mathrm{d}A = U\,(T_\mathrm{h} - T_\mathrm{c})_\mathrm{local}\,\mathrm{d}A = U\,\Delta T\,\mathrm{d}A \tag{2}$$

Here U is the overall heat-transfer coefficient.

Equations (1) and (2) in an integrated form for a heat exchanger are

$$q = C_h\left(T_{h,i} - T_{h,o}\right) = C_c\left(T_{c,o} - T_{c,i}\right) \tag{3}$$

and

$$q = UA\Delta T_m = \frac{\Delta T_m}{\boldsymbol{R}_o} \tag{4}$$

The true temperature difference ΔT_m depends on the exchanger configuration (flow arrangement) and the degree of fluid mixing within each fluid stream. The reciprocal of the overall thermal conductance *UA* is referred to as the overall thermal resistance $\boldsymbol{R}_o$. The mean overall heattransfer coefficient is defined as

$$U_m = \frac{1}{A}\int_A U\,dA \tag{5}$$

Unless explicitly mentioned, U_m is used throughout this article simply as U. This is because a constant and uniform U is idealized in the simplified theory as well as experimentally obtained. Note that the integration of Equation (2) with the definition of U_m from Equation (5) will yield Equation (4) with the following correct definition of ΔT_m

$$\frac{1}{\Delta T_m} = \frac{1}{q}\int_q \frac{dq}{dT} \tag{6}$$

The overall thermal resistance $\boldsymbol{R}_o$ consists of component resistances in series as shown in Figure 31.

$$\begin{aligned}\frac{1}{UA} &= \boldsymbol{R}_o = \boldsymbol{R}_h + \boldsymbol{R}_{s,h} + \boldsymbol{R}_w + \boldsymbol{R}_{s,c} + \boldsymbol{R}_c \\ &= \frac{1}{(\eta_0 hA)_h} + \frac{1}{(\eta_0 h_s A)_h} + \boldsymbol{R}_w + \frac{1}{(\eta_0 h_s A)_c} + \\ &\quad \frac{1}{(\eta_0 hA)_c}\end{aligned} \tag{7}$$

where $\boldsymbol{R}_h$ = hot-side film convective resistance, $\boldsymbol{R}_{s,h}$ = hot-side scale (fouling) resistance, $\boldsymbol{R}_w$ = wall thermal resistance, $\boldsymbol{R}_{s,c}$ = cold-side scale (fouling) resistance, $\boldsymbol{R}_c$ = cold-side film convective resistance, η_o = the extended surface efficiency, and h_s = fouling coefficient (inverse of a fouling factor).

The wall thermal resistance $\boldsymbol{R}_w$ is

$$\boldsymbol{R}_w = \begin{cases} \dfrac{\delta}{A_w k_w} & \text{for a flat wall} \\[2ex] \dfrac{\ln(d_o/d_i)}{2\pi k_w L N_t} & \text{for a circular tube with a single-layer wall} \\[2ex] \dfrac{1}{2\pi L N_t}\left[\displaystyle\sum_j \frac{\ln(d_{j+1}/d_j)}{k_{wj}}\right] & \text{for a circular tube with a multiple-layer wall} \end{cases}$$

where δ is the plate thickness, A_w is the total wall area for heat conduction, k_w is the thermal conductivity of the wall material, d_o and d_i are tube outside and inside diameters respectively, L is the tube length, and N_t is the number of tubes.

For a flat (or plain) wall associated with a plate – fin or an all prime surface exchanger,

$$A_w = L_1 L_2 N_p \tag{9}$$

where L_1, L_2, and N_p are the length, width, and total number of separating plates.

The extended surface efficiency η_o is

$$\eta_0 = 1 - \frac{A_f}{A}\left(1 - \eta_f\right) \tag{10}$$

where A_f = fin surface area, A = total (fin + primary) surface area on one side of the exchanger, and η_f = fin efficiency.

The expressions for η_f are provided in Table 4 for some common types of fins. The fin efficiency for continuous flat fins on an array of circular tube is obtained by a sector method [27]. In this method, the rectangular or hexagonal fin (see Fig. 32A or B) is divided into N sectors. Each sector i is considered as a circular fin (see Table 4) with the radius $r_{e,i}$ equal to the length of the centerline of the sector i. The fin efficiency η_f of the whole fin is then given by

$$\eta_f = \frac{\sum_{i=1}^{N} \eta_{f,i} A_{f,i}}{\sum_{i=1}^{N} A_{f,i}} \tag{11}$$

Only a few segments N (e.g., 8 – 10) will suffice to provide η_f accurately.

UA in Equation (7) is defined as

$$UA = U_h A_h = U_c A_c = U_w A_w \tag{12}$$

Thus U is defined optionally in terms of the hot fluid surface area A_h, the cold fluid surface area A_c, or the wall conduction area A_w. For a plain

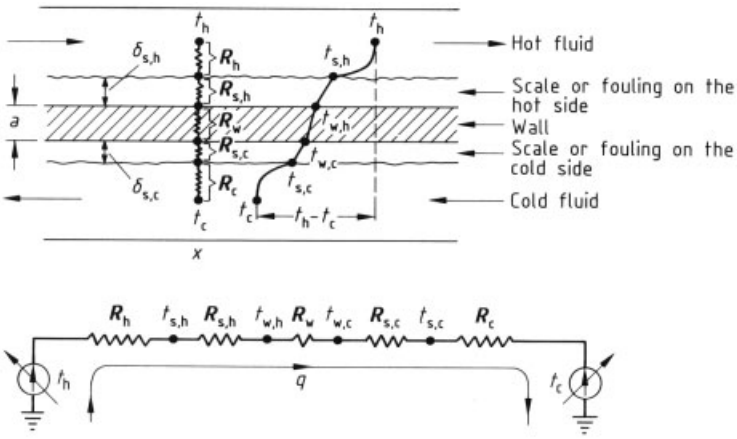

Figure 31. Thermal circuit for a heat exchanger section

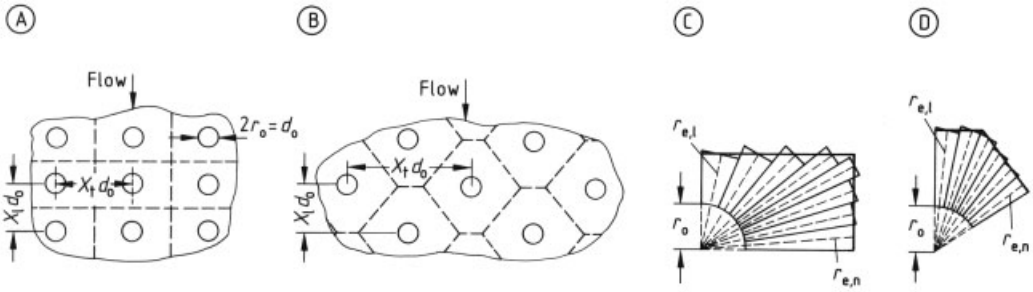

Figure 32. Flat fin over (A) an inline and (B) a staggered tube arrangement; the smallest representative segment of the fin for (C) an inline and (D) a staggered tube arrangement

tubular exchanger, U_o based on the tube outside surface area, from Equation (7), is given by

$$\frac{1}{U_o}=\frac{1}{h_0}+\frac{1}{h_{o,s}}+\frac{d_o \ln(d_o/d_i)}{2k_w}+\frac{d_o}{h_{i,s}d_i}+\frac{d_o}{h_i d_i} \quad (13)$$

Integration of Equations (1) and (2) for a heat exchanger configuration, when presented in a nondimensional form can be presented in four alternative methods referred to as ε – NTU, P – NTU, MTD, and ψ – P methods, see Section 5.2. The following idealizations are built into the results presented in Table 7.

1) The heat exchanger operates under steady-state conditions (i.e., constant flow rate, and thermal history of fluid particles independent of time).
2) Heat losses to the surroundings are negligible.
3) There are no thermal energy sources in the exchanger.
4) In counterflow and parallel-flow exchangers, the temperature of each fluid is uniform over every flow cross section. From the temperature distribution point of view, in crossflow exchangers, each fluid is considered mixed or unmixed at every cross section depending upon the specifications. For a multipass exchanger, the foregoing statements apply to each pass depending upon the basic flow arrangement of the passes; the fluid is considered mixed or unmixed between passes.
5) Either there are no phase changes (condensation or boiling) in the fluid streams flowing through the exchanger or the phase changes occur under one of the following conditions: (a) Phase change occurs at a constant temperature as for a single component fluid at constant pressure; the effective specific heat for the phase-changing fluid is infinity in this case, and hence $C_{\max} \to \infty$. (b) The temperature of the phase-changing fluid varies lin-

Table 4. Fin efficiency for plate – fin and tube – fin geometries of uniform fin thickness

Geometry	Fin efficiency formula $m_i = \left[\frac{2h}{k_f \delta_i}\left(1 + \frac{\delta_i}{l_{ef}}\right)\right]^{1/2}$ $E_i = \frac{\tanh(m_i l_i)}{m_i l_i}$ $i = 1, 2, 3, 4$
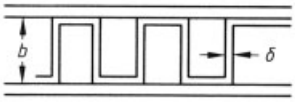 Plain, wavy, or offset strip fin of rectangular cross section	$\eta_f = E_1$ $l_1 = \frac{b}{2} - \delta_1 \quad \delta_1 = \delta$
Triangular fin heated from one side	$\eta_f = \frac{h A_1 (T_0 - T_a) \frac{\sinh(m_1 l_1)}{m_1 l_1} + q_e}{\cosh(m_1 l_1)\left[h A_1 (T_0 - T_a) + q_e \frac{T_0 - T_a}{T_1 - T_a}\right]}$
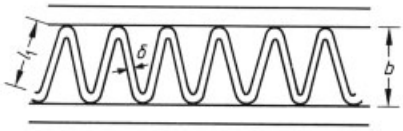 Plain, wavy, or louver fin of triangular cross section	$\eta = E_1$ $l_1 = \frac{l}{2}, \ \delta_1 = \delta$
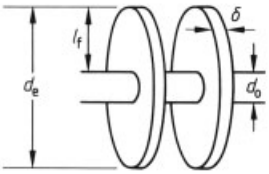 Circular fin	$\eta_f = \begin{cases} a(m l_e)^{-b} & \text{for } \Phi > 0.6 + 2.257(r^*)^{-0.445} \\ \frac{\tanh \Phi}{\Phi} & \text{for } \Phi \le 0.6 + 2.257(r^*)^{-0.445} \end{cases}$ $a = (r^*)^{-0.246} \quad \Phi = m l_e (r^*)^{\exp(0.13 m l_e - 1.3863)}$ $b = \begin{cases} 0.9107 + 0.0893 r^* & \text{for } r^* \le 2 \\ 0.9706 + 0.17125 \ln r^* & \text{for } r^* > 2 \end{cases}$ $m = \left(\frac{2h}{k_f \delta}\right)^{1/2} \quad l_e = l_f + \frac{\delta}{2} \quad r^* = \frac{r_e}{r_o}$
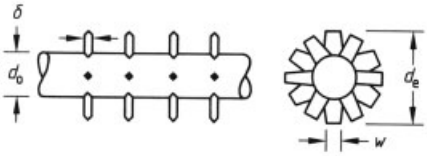 Studded fin	$\eta_f = \frac{\tanh(m l_e)}{m l_e}$ $m = \left[\frac{2h}{k_f \delta}\left(1 + \frac{\delta}{w}\right)\right]^{1/2} \quad l_e = l_f + \frac{\delta}{2} \quad l_f = \frac{(d_e - d_o)}{2}$
Rectangular fin over circular tubes	See the text.

early with heat transfer during the condensation or boiling. In this case, the effective specific heat is constant and finite for the phase-changing fluid.

6) The specific heat of each fluid is constant throughout the exchanger so that the heat capacity on each side is treated as constant.
7) The velocity and temperature at the entrance of the heat exchanger on each fluid side are uniform.
8) For an extended-surface exchanger, the overall extended-surface temperature effectiveness η_o is considered uniform and constant.
9) The overall heat-transfer coefficient between the fluids is constant throughout the exchanger, including the case of phase-changing fluid in idealization 5.
10) The heat-transfer area is distributed uniformly on each fluid side. In a multipass unit, heat-transfer surface area is equal in each pass.

11) For a plate-baffled shell-and-tube exchanger, the temperature rise per baffle pass is small compared to the overall temperature rise along the exchanger, i.e., the number of baffles is large.
12) The fluid flow rate is uniformly distributed through the exchanger on each fluid side in each pass. No stratification, flow bypassing, or flow leakages occur in any stream. The flow condition is characterized by the bulk (or mean) velocity at any cross section.
13) Longitudinal heat conduction in the fluid and in the wall is negligible.

Idealizations 1 to 4 are necessary in a theoretical analysis of steady-state heat exchangers. Idealization 5 essentially restricts the analysis to single-phase flow on both sides or on one side with a dominating thermal resistance. For two-phase flow on both sides, many of the foregoing idealizations are not valid, since mass transfer in phase change results in variable properties and variable flow rates of each phase, and the heat-transfer coefficients vary significantly. As a result, the two-phase heat exchanger cannot be analyzed using the theory presented in this section. The design theory of two-phase exchangers is presented in Section 9.

If idealization 6 is not valid, the exchanger is divided into small segments until the specific heats can be treated as constant. Violation of idealization 7 is treated in Section 11. Idealization 8 is essential since no information is available on local variations of the heat-transfer coefficient h. The influences of idealizations 9 and 13 are discussed in Section 5.3; the influences of idealizations 10 – 12 are covered in [28].

If any of these idealizations are not valid for a particular exchanger application, the best solution is to work directly with either Equation (1) and (2) or their modified form by including a particular effect, and to integrate them over a small exchanger segment in which all of the idealizations are valid.

5.2. Dimensionless Methods for Exchanger Heat-Transfer Analysis

The results of integration of Equations (1) and (2) for exchanger configurations under the idealizations made just above can be presented in terms of three dimensionless groups. Based on the choice of the dimensionless groups, four methods have been evolved: ε – NTU, P – NTU, MTD, and $\psi - P$. The dimensionless groups for each method, their functional relationship, and how to compute q are summarized in Table 5 [28]. The relationships among the dimensionless groups are summarized in Table 6 [28]. A brief description of the dimensionless groups follows.

Table 6. Relationships between dimensionless groups of the P-NTU$_1$, MTD, and ψ-P methods and those of the ε-NTU method

$$P_1 = \frac{C_{\min}}{C_1}\varepsilon = \begin{cases} \varepsilon & \text{for} \quad C_1 = C_{\min} \\ \varepsilon C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$R_1 = \frac{C_1}{C_2} = \begin{cases} C^* & \text{for} \quad C_1 = C_{\min} \\ 1/C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$\mathrm{NTU}_1 = \mathrm{NTU}\frac{C_{\min}}{C_1} = \begin{cases} \mathrm{NTU} & \text{for} \quad C_1 = C_{\min} \\ \mathrm{NTU}\, C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$F = \frac{\mathrm{NTU}_{\mathrm{ef}}}{\mathrm{NTU}} = \frac{1}{\mathrm{NTU}(1 - C^*)} \ln \frac{1 - C^*\varepsilon}{1 - \varepsilon} \xrightarrow[C^* = 1]{} \frac{\varepsilon}{\mathrm{NTU}(1 - \varepsilon)}$$

$$F = \frac{1}{\mathrm{NTU}_1(1 - R_1)} \ln\left[\frac{1 - RP_1}{1 - P_1}\right] \xrightarrow[R_1 = 1]{} \frac{P_1}{\mathrm{NTU}_1(1 - P_1)}$$

$$\psi = \frac{\varepsilon}{\mathrm{NTU}} = \frac{P_1}{\mathrm{NTU}_1} = \frac{FP_1(1 - R_1)}{\ln[(1 - R_1 P_1)/(1 - P_1)]} \xrightarrow[R_1 = 1]{} F(1 - P_1)$$

The *heat-exchanger effectiveness* ε is an efficiency factor and has a thermodynamic significance. It is the ratio of actual heat transfer from the hot to the cold fluid in any exchanger configuration to the thermodynamically limited maximum possible heat transfer [19]. It is also the temperature effectiveness of the $C_{\min}$ fluid. Its value ranges from 0 to 1.

The *number of transfer units* NTU is a ratio of the overall conductance UA to the smaller heat capacity rate $C_{\min}$ (minimum of C_{h} and C_{c}). It designates the dimensionless "thermal size" of the exchanger. It may be interpreted as the $C_{\min}$ fluid dimensionless residence time, a temperature ratio, or a modified Stanton number [29]. Its value can range from 0 to ∞. For most process exchangers, its value ranges from ca. 0.5 to 2.

The *heat capacity rate ratio* C^* is simply a ratio of the smaller to larger heat capacity rates of the two fluids such that $C^* \leq 1$.

The *temperature effectiveness* or *thermal effectiveness* P of a fluid in an exchanger is the

Table 5. General functional relationships and dimensionless groups for ε-NTU, P-NTU, MTD, and ψ-P methods [28]

ε-NTU method	P-NTU method
$q = \varepsilon C_{\min}(T_{h,i} - T_{c,i})$	$q = P_1 C_1 \lvert T_{2,i} - T_{1,i} \rvert$
$\varepsilon = \varphi(\text{NTU}, C^*, \text{flow arrangement})$	$P_1 = \varphi(\text{NTU}_1, R_1, \text{flow arrangement})$
$\varepsilon = \dfrac{C_h(T_{h,i} - T_{h,o})}{C_{\min}(T_{h,i} - T_{c,i})} = \dfrac{C_c(T_{c,o} - T_{c,i})}{C_{\min}(T_{h,i} - T_{c,i})}$	$P_1 = \dfrac{T_{1,o} - T_{1,i}}{T_{2,i} - T_{1,i}}$
$\text{NTU} = \dfrac{UA}{C_{\min}} = \dfrac{1}{C_{\min}} \int_A U\,dA$	$\text{NTU}_1 = \dfrac{UA}{C_1} = \dfrac{[T_{1,o} - T_{1,i}]}{\Delta T_m}$
$C^* = \dfrac{C_{\min}}{C_{\max}} = \dfrac{(Wc_p)_{\min}}{(Wc_p)_{\max}}$	$R_1 = \dfrac{C_1}{C_2} = \dfrac{T_{2,i} - T_{2,o}}{T_{1,o} - T_{1,i}}$
MTD method	**ψ-P method**
$q = UAF\Delta T_{lm}$	$q = UA\psi(T_{h,i} - T_{c,i})$
$\text{LMTD} = \Delta T_{lm} = \dfrac{\Delta T_1 - \Delta T_2}{\ln(\Delta T_1/\Delta T_2)}$	$\psi = \varphi(P_1, R_1, \text{flow arrangement})$
$\Delta T_1 = T_{h,i} - T_{c,o} \quad \Delta T_2 = T_{h,o} - T_{c,i}$	$\psi = \dfrac{\Delta T_m}{T_{h,i} - T_{c,i}}$
$F = \Phi(P_1, R_1, \text{flow arrangement}) \quad F = \dfrac{\Delta T_m}{\Delta T_{lm}}$	P_1 and R_1 are defined in the P-NTU method
P_1 and R_1 are defined in the P-NTU method	

ratio of the temperature rise (or drop) of that fluid stream to the inlet temperature difference of two fluid streams $P = \varepsilon$ for the $C_{\min}$ fluid, and $P = \varepsilon C^*$ for the $C_{\max}$ fluid. Thus $P \leq \varepsilon$ depending on the value of P for the $C_{\min}$ or $C_{\max}$ side.

The heat capacity rate ratio of the two fluid streams is designated as $R : R_1 = C_1/C_2$; $R_2 = C_2/C_1$; and R can range from 0 to ∞.

The *log-mean temperature difference correction factor* F, also referred to as the *exchanger configuration factor,* is a ratio of the true (actual) mean temperature difference in an exchanger to the log-mean temperature difference (LMTD). The LMTD is defined in Table 5 for all exchanger configurations except for a single-pass parallel flow exchanger for which $\Delta T_1 = T_{h,i} - T_{c,i}$ and $\Delta T_2 = T_{h,o} - T_{c,o}$. It must be emphasized that F does not represent the effectiveness or efficiency of an exchanger, i.e., $F = 1$ does not mean that the exchanger is 100 % efficient. It represents a degree of departure in the true mean temperature difference for a given exchanger from the counterflow log-mean temperature difference. The value of F ranges from 0 to 1; its value is maintained approximately at ≥ 0.8 in a good design.

The ψ factor is the ratio of the true mean temperature difference in an exchanger to the inlet fluid temperature difference. Its value ranges from 0 to 1.

The results for a variety of configurations for the $\varepsilon-$NTU, $P-$NTU, MTD, or $\psi-P$methods are available in the literature. The results for process industry configurations are provided in Table 7 for the $P-$NTU method. Similar results for the $\varepsilon-$NTU or $\psi-P$ methods are available in [28] and for the MTD and $\psi-P$ methods in [30].

In Table 7, those configurations that are stream symmetric are clearly indicated. For a stream symmetric configuration, reversing the flow direction of one or both fluid streams or interchanging the fluid streams (i.e., switching fluid 1 to the fluid 2 side and fluid 2 to fluid 1 side) will not change the expression for P_1. For example, $1-2$ shell-and-tube exchangers with shell fluid mixed (TEMA E shell) have stream symmetry. Hence, fluid 1 can flow either in the tubes or on the shellside; its temperature effectiveness P_1 is given by Equation (3.1) in Table 7. This is not true, for example, for the divided flow $1-2$ exchanger with the shell fluid mixed (TEMA G shell). Here fluid 1 must be the shellside fluid,

and P_1, NTU_1, and R_1 refer to the shellside fluid. For the tube fluid, P_2, NTU_2 and R_2 should then be computed from

$$P_2 = P_1 R_1, R_2 = 1/R_1, \mathrm{NTU}_2 = \mathrm{NTU}_1 R_1 \tag{14}$$

The relationship between the F factor, P_1, R_1, and NTU_1 is presented in Table 6 which is valid for all single-pass and multipass exchangers. However, when the exchangers are arranged in series, the F factors of individual exchangers are related to the overall F factor for all exchangers combined, as shown in Equation (4.2.4) of Table 7.

For a plate exchanger, when the number of thermal plates N is greater than ca. 40, its effectiveness approaches that of a plate exchanger with an infinite number of thermal plates [31]. Since most plate exchangers have the number of thermal plates $N > 40$, the asymptotic temperature effectiveness of single-pass and multipass exchangers are available [32].

5.3. Extensions of the Basic Heat-Transfer Theory

A number of idealizations mentioned in Section 5.1 have been relaxed and theoretical results are obtained which represent extensions to the basic theory. Two such results are summarized next. The influence of unequal pass areas, shellside baffle passes, and shell fluid bypassing is covered in [28].

5.3.1. Longitudinal Wall Heat Conduction

Longitudinal wall heat conduction in the opposite direction to fluid flow reduces the mean temperature difference in the exchanger and the exchanger effectiveness. This effect is primarily important for exchangers having a steep temperature gradient ($\Delta T/L$) in the flow direction, i.e., exchangers having high effectiveness ($> 75\,\%$) per pass and a short flow length L for the pass. The influence of longitudinal conduction is reported in the heat exchanger literature in terms of the $\varepsilon-\mathrm{NTU}$ method for single-pass counterflow and single-pass crossflow exchangers with both fluids unmixed. For the single-pass counterflow exchanger, the exchanger effectiveness, including the effect of longitudinal conduction, for $C^* = 1$ and $0.1 \le \eta_o hA \le 10$ is [33]:

$$1-\varepsilon = \frac{1}{1 + \mathrm{NTU}\,\frac{1+\lambda[\lambda\mathrm{NTU}/(1+\lambda\mathrm{NTU})]^{1/2}}{1+\lambda\mathrm{NTU}}} \tag{15}$$

where $\lambda = k_w A_k / L\, C_{\min}$ and A_k is the total wall cross-sectional area for longitudinal conduction.

The exchanger ineffectiveness $(1-\varepsilon)$ of Equation (15) is shown in [28], [29]. For $0.8 \le C^* \le 1$, the following correlation for ε is obtained [33]:

$$1-\varepsilon = \frac{1-C^*}{\Psi \exp(r_1) - C^*} \tag{16}$$

where

$$r_1 = \frac{(1-C^*)\,\mathrm{NTU}}{1+\lambda\mathrm{NTU}C^*} \tag{17}$$

$$\Psi = \frac{1+\gamma\Psi^*}{1-\gamma\Psi^*}; \Psi^* = \left(\frac{\alpha}{1+\alpha}\right)^{1/2} \frac{1+\gamma}{1/\alpha - \gamma - \gamma^2} \tag{18}$$

$$\gamma = \left(\frac{1-C^*}{1+C^*}\right)\left(\frac{1}{1+\alpha}\right); \alpha = \lambda\mathrm{NTU}C^* \tag{19}$$

Although Equation (16) is derived for $(\eta_o hA)^* = 1$, if $0.5 \le (\eta_o hA)^* \le 2$, the error introduced in the ineffectiveness $(1-\varepsilon)$ is within 0.8 % and 4.7 % for $C^* = 0.95$ and 0.8, respectively.

The effectiveness of a single-pass crossflow exchanger with both fluids unmixed, including the effect of longitudinal heat conduction, has been determined by Chiou [34]; the results are also available in [28].

For a multipass exchanger, the effect of longitudinal heat conduction is included in determining individual pass effectivenesses, as already mentioned. The individual pass effectivenesses are in turn are used to compute the effectiveness of a multipass exchanger.

5.3.2. Variable Overall Heat-Transfer Coefficients

The theory and results of Section 5.2 are valid for constant and uniform overall heat-transfer coefficient U. In fact, individual heat-transfer coefficients (on which U depends) could be strongly dependent on the flow Reynolds number, heat-transfer surface geometries, fluid physical properties, and operating temperatures and pressures. Particularly for viscous liquids and boiling or

Table 7. P_1-NTU_1 formulas and limiting values P_1 for R=1 and $NTU_1 \to \infty$ for various exchanger flow arrangements *

Flow arrangement	Eq. no.	General formula	Value for $R_1 = 1$	Value for $NTU_1 \to \infty$
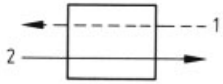	1.1.1	$P_1 = \dfrac{1 - \exp[-NTU_1(1-R_1)]}{1 - R_1 \exp[-NTU_1(1-R_1)]}$	$P_1 = \dfrac{NTU_1}{1 + NTU_1}$	$P_1 \to 1$ for $R_1 \le 1$ $P_1 \to 1/R_1$ for $R_1 \ge 1$
	1.1.2	$NTU = \dfrac{1}{(1-R_1)} \ln\left[\dfrac{1 - R_1 P_1}{1 - P_1}\right]$	$NTU_1 = \dfrac{P_1}{1 - P_1}$	$NTU_1 \to \infty$
Counterflow exchanger, stream symmetric	1.1.3	$F = 1$	$F = 1$	$F = 1$
	1.2.1	$P_1 = \dfrac{1 - \exp[-NTU_1(1+R_1)]}{1 + R_1}$	$P_1 = \dfrac{1}{2}[1 - \exp(-2NTU_1)]$	$P_1 \to \dfrac{1}{1+R_1}$
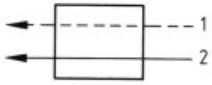	1.2.2	$NTU_1 = \dfrac{1}{1+R_1} \ln\left[\dfrac{1}{1 - P_1(1+R_1)}\right]$	$NTU_1 = \dfrac{1}{2} \ln\left[\dfrac{1}{1 - 2P_1}\right]$	$NTU_1 \to \infty$
Parallel flow exchanger, stream symmetric	1.2.3	$F = \dfrac{(R_1+1)\ln\left[\dfrac{1 - R_1 P_1}{1 - P_1}\right]}{(R_1 - 1)\ln[1 - P_1(1+R_1)]}$	$F = \dfrac{2P_1}{(P_1 - 1)\ln(1 - 2P_1)}$	$F \to 0$
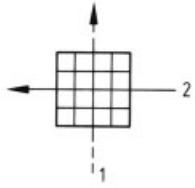	2.1	$P_1 = 1 - \exp(NTU_1) - \exp[-(1+R_1)NTU_1] \cdot \sum_{n=1}^{\infty} R_1^n P_n(NTU_1)$ $P_n(y) = \dfrac{1}{(n+1)!} \sum_{j=1}^{n} \dfrac{(n+1-j)}{j!} y^{n+j}$	same as Eq. (2.1) with $R_1 = 1$	$P_1 \to 1$ for $R_1 \le 1$ $P_1 \to \dfrac{1}{R_1}$ for $R_1 \ge 1$
Single-pass crossflow exchanger, both fluids unmixed, stream symmetric				

Table 7. Continued

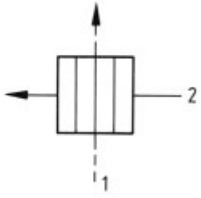 Single-pass crossflow exchanger, fluid 1 unmixed, fluid 2 mixed	2.2.1	$P_1 = [1 - \exp(-KR_1)]/R_1$ $K = 1 - \exp(-\mathrm{NTU}_1)$	$P_1 = 1 - \exp(-K)$	$P_1 \to \dfrac{1 - \exp(-R_1)}{R_1}$
	2.2.2	$\mathrm{NTU}_1 = \ln\left[\dfrac{1}{1 + \dfrac{1}{R_1}\ln(1 - R_1 P_1)}\right]$	$\mathrm{NTU}_1 = \ln\left[\dfrac{1}{1 + \ln(1 - P_1)}\right]$	$\mathrm{NTU}_1 \to \infty$
	2.2.3	$F = \dfrac{\ln[(1 - R_1 P_1)/(1 - P_1)]}{(R - 1)\ln\left[1 + \dfrac{1}{R_1}\ln(1 - R_1 P_1)\right]}$	$F = \dfrac{P_1}{(P_1 - 1)\ln[1 + \ln(1 - P_1)]}$	$F \to 0$
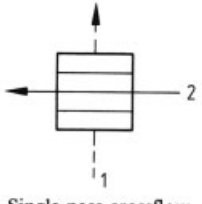 Single-pass crossflow exchanger, fluid 1 mixed, fluid 2 unmixed	2.3.1	$P_1 = 1 - \exp(-K/R_1)$ $K = 1 - \exp(-R_1\,\mathrm{NTU}_1)$	$P_1 = 1 - \exp(-K)$ $K = 1 - \exp(-\mathrm{NTU}_1)$	$P_1 \to 1 - \exp(-1/R_1)$
	2.3.2	$\mathrm{NTU}_1 = \dfrac{1}{R_1}\ln\left[\dfrac{1}{1 + R_1\ln(1 - P_1)}\right]$	$\mathrm{NTU}_1 = \ln\left[\dfrac{1}{(1 + \ln(1 - P_1)}\right]$	$\mathrm{NTU}_1 \to \infty$
	2.3.3	$F = \dfrac{\ln[(1 - R_1 P_1)/(1 - P_1)]}{(1 - 1/R)\ln[1 + R_1\ln(1 - P_1)]}$	$F = \dfrac{P_1}{(P_1 - 1)\ln[1 + \ln(1 - P_1)]}$	$F \to 0$
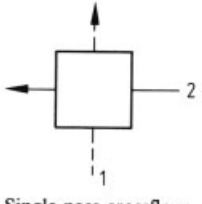 Single-pass crossflow exchanger, both fluid mixed, stream symmetric	2.4	$P_1 = \left[\dfrac{1}{K_1} + \dfrac{R_1}{K_2} - \dfrac{1}{\mathrm{NTU}_1}\right]^{-1}$ $K_1 = 1 - \exp(-\mathrm{NTU}_1)$ $K_2 = 1 - \exp(-R_1\,\mathrm{NTU}_1)$	$P_1 = \left[\dfrac{2}{K_1} - \dfrac{1}{\mathrm{NTU}_1}\right]^{-1}$	$P_1 \to \dfrac{1}{1 + R_1}$

Table 7. Continued

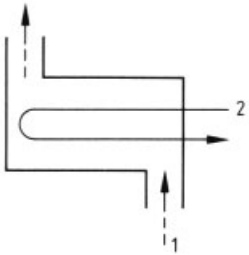 1–2 TEMA E shell-and-tube exchanger, shell fluid mixed, stream symmetric	3.1	$P_1 = \dfrac{2}{1 + R_1 + \varrho \coth(\varrho\,\mathrm{NTU}_1/2)}$ $\varrho = [1 + R_1^2]^{1/2}$	$P_1 = \dfrac{1}{1 + \coth(\mathrm{NTU}_1/\sqrt{2})/\sqrt{2}}$	$P_1 \to \dfrac{2}{1 + R_1 + \varrho}$
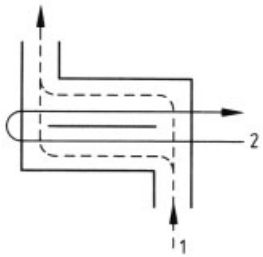 1–2 TEMA E shell-and-tube exchanger, shell fluid divided into two streams individually mixed	3.2	$P_1 = \dfrac{1}{R_1}\left[1 - \dfrac{(2 - R_1)(2A + R_1 B)}{(2 + R_1)(2A - R_1/B)}\right]$ $A = \exp(\mathrm{NTU}_1)$ $B = \exp(-\mathrm{NTU}_1 R_1/2)$ same as 1–1 J shell, Eq. (3.10)	$P_1 = \dfrac{1}{2}\left[1 - \dfrac{1 + A^{-2}}{2(1 + \mathrm{NTU}_1)}\right]$ for $R_1 = 2$	$P_1 \to \dfrac{2}{2 + R_1}$ for $R_1 \le 2$ $P_1 \to \dfrac{1}{R_1}$ for $R_1 \ge 2$

Table 7. Continued

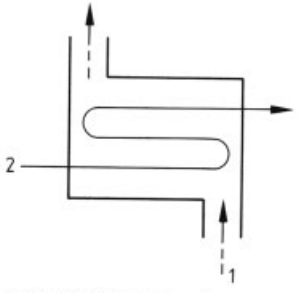 1–3 TEMA E shell-and-tube exchanger, shell and tube fluids mixed, one parallel-flow and two counterflow passes	3.3	$P_1 = \frac{1}{R_1}\left[1 - \frac{C}{AC + B^2}\right]$ $A = X_1(R_1 + \lambda_1)(R_1 - \lambda_2)/2\lambda_1 - X_3\delta$ $\quad - X_2(R_1 + \lambda_2)(R_1 - \lambda_1)/2\lambda_2 + 1/(1 - R_1)$ $B = X_1(R_1 - \lambda_2) - X_2(R_1 - \lambda_1) + X_3\delta$ $C = X_2(3R_1 + \lambda_1) - X_1(3R_1 + \lambda_2) + X_3\delta$ $X_i = \exp(\lambda_i \mathrm{NTU}_1/3)/2\delta, \quad i = 1,2,3$ $\delta = \lambda_1 - \lambda_2$ $\lambda_1 = -\frac{3}{2} + \left[\frac{9}{4} + R_1(R_1 - 1)\right]^{1/2}$ $\lambda_2 = -\frac{3}{2} - \left[\frac{9}{4} + R_1(R_1 - 1)\right]^{1/2}$ $\lambda_3 = R_1$	same as Eq. (3.3) with $R_1 = 1$ $A = -\exp(-\mathrm{NTU}_1)/18$ $\quad - \exp(\mathrm{NTU}_1/3)/2$ $\quad + (\mathrm{NTU}_1 + 5)/9$	$P_1 \to 1$ for $R_1 \le 1$ $P_1 \to \frac{1}{R_1}$ for $R_1 \ge 1$
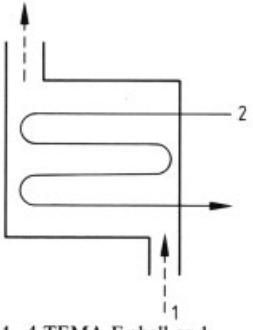 1–4 TEMA E shell-and-tube exchanger, shell and tube fluids mixed	3.4	$P_1 = 4[2(1 + R_1) + \varrho A + R_1 B]^{-1}$ $A = \coth(\varrho \mathrm{NTU}_1/4)$ $B = \tanh(R_1 \mathrm{NTU}_1/4)$ $\varrho = [4 + R_1^2]^{1/2}$	$P_1 = 4[4 + \sqrt{5}A + B]^{-1}$ $A = \coth(\sqrt{5}\mathrm{NTU}_1/4)$ $B = \tanh(\mathrm{NTU}_1/4)$	$P_1 \to \frac{4}{2(1 + R_1) + \varrho + R_1}$
Limit of 1-n TEMA E shell-and-tube exchanger, for large N. Shell and tube fluids mixed, stream symmetric	3.5	Eq. (2.4) applies in this limit	same as for Eq. (2.4)	same as for Eq. (2.4)

Table 7. Continued

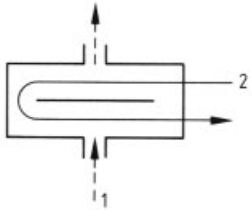 Overall counterflow 1–2 TEMA G shell-and-tube exchanger, shell and tube fluids mixed in each pass at a cross section	3.6	$P_1 = (B - \alpha^2)/(A + 2 + R_1 B)$ $A = -2R_1(1-\alpha)^2/(2+R_1)$ $B = [4 - \beta(2+R_1)]/(2-R_1)$ $\alpha = \exp[-\mathrm{NTU}_1(2+R_1)/4]$ $\beta = \exp[-\mathrm{NTU}_1(2-R_1)/2]$	$P_1 = \dfrac{1 + 2\mathrm{NTU}_1 - \alpha^2}{4 + 4\mathrm{NTU}_1 - (1-\alpha)^2}$ for $R_1 = 2$ $\alpha = \exp(-\mathrm{NTU}_1)$	$P_1 \to \dfrac{2 + R_1}{R_1^2 + R_1 + 2}$ for $R_1 \le 2$ $P_1 \to \dfrac{1}{R_1}$ for $R_1 \ge 2$
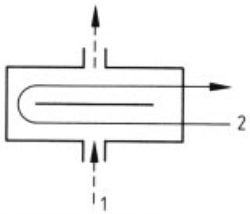 Overall parallel flow 1–2 TEMA G shell-and-tube exchanger, shell and tube fluids mixed in each pass at a cross section	3.7	$P_1 = (B - \alpha^2)/(A + 2 - R_1\alpha^2)$ $A = 2R_1(1-\alpha)^2/(2-R_1)$ $B = [4 - \beta(2-R_1)]/(2+R_1)$ $\alpha = \exp[-\mathrm{NTU}_1(2-R_1)/4]$ $\beta = \exp[-\mathrm{NTU}_1(2+R_1)/2]$	$P_1 = \dfrac{1 + 2\mathrm{NTU}_1 - \beta}{4 + 4\mathrm{NTU}_1 + \mathrm{NTU}_1^2}$ for $R_1 = 2$ $\beta = \exp(-2\mathrm{NTU}_1)$	$P_1 \to \dfrac{2 - R_1}{2 + R_1}$ for $R_1 \le 2$ $P_1 \to \dfrac{R_1 - 2}{R_1^2}$ for $R_1 \ge 2$
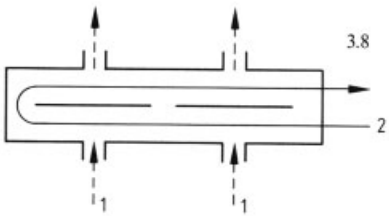 Overall parallel flow 1–2 TEMA H shell-and-tube exchanger, shell and tube fluids mixed in each pass at a cross section	3.8	$P_1 = \dfrac{1}{R_1}\left[1 - \dfrac{B + 4G/R_1}{(1-D)^4}\right]$ $B = (1+H)(1+E)^3$ $G = (1-D)^2(D^2+E^2) + D^2(1+E)^2$ $H = [\exp(-2\beta) - 1]/(1 + 4/R_1)$ $E = [\exp(-\beta) - 1]/(1 + 4/R_1)$ $D = -[\exp(-\alpha) - 1]/(1 - 4/R_1)$ $\alpha = \mathrm{NTU}_1(4 - R_1)/8;\ \beta = \mathrm{NTU}_1(4 + R_1)/8$	Same as Eq. (3.8) with $D = -\mathrm{NTU}_1/2$ for $R_1 = 4$	$P_1 \to \dfrac{(4 - R_1)(R_1^2 + 16)}{(R_1 + 4)^3}$ for $R_1 \le 4$ $P_1 \to (R_1 - 4)/R_1^2$ for $R_1 \ge 4$

Table 7. Continued

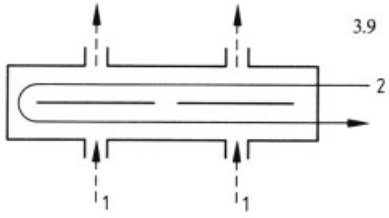 Overall counterflow 1–2 TEMA H shell-and-tube exchanger, shell and tube fluids mixed in each pass at a cross section	3.9	$P_1 = \frac{1}{R_1}\left[1 - \frac{(1-D)^4}{B - 4G/R_1}\right]$ $B = (1+H)(1+E)^2$ $G = (1-D)^2(D^2+E^2) + D^2(1+E)^2$ $H = [1-\exp(-2\beta)]/(4/R_1 - 1)$ $E = [1-\exp(-\beta)]/(4/R_1 = 1)$ $D = [1-\exp(-\alpha)]/(4/R_1 + 1)$ $\alpha = \mathrm{NTU}_1(4+R_1)/8$ $\beta = \mathrm{NTU}_1(4-R_1)/8$	same as Eq. (3.9) with $H = \mathrm{NTU}_1$ $E = \mathrm{NTU}_1/2$ for $R_1 = 4$	$P_1 \to \left[R_1 + \frac{(4-R_1)^3}{(4+R_1)(R_1^2+16)}\right]^{-1}$ for $R_1 \le 4$ $P_1 \to \frac{1}{R_1}$ for $R_1 \ge 4$
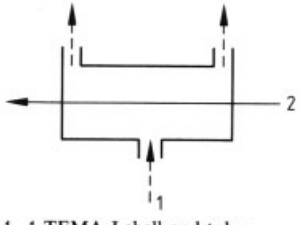 1–1 TEMA J shell-and-tube exchanger, shell and tube fluids mixed	3.10	$P_1 = \frac{1}{R_1}\left[1 - \frac{(2-R_1)(2A+R_1B)}{(2+R_1)(2A-R_1/B)}\right]$ $A = \exp(\mathrm{NTU}_1)$ $B = \exp(-\mathrm{NTU}_1 R_1/2)$ same as Eq. (3.2)	$P_1 = \frac{1}{2}\left[1 - \frac{1+A^{-2}}{2(1+\mathrm{NTU}_1)}\right]$ for $R_1 = 2$	$P_1 \to \frac{2}{2+R_1}$ for $R_1 \le 2$ $P_1 \to \frac{1}{R_1}$ for $R_1 \ge 2$
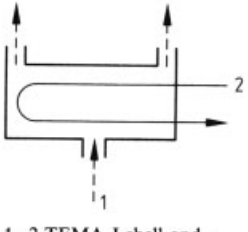 1–2 TEMA J shell-and-tube exchanger, shell and tube fluids mixed	3.11	$P_1 = [1 + \frac{R_1}{2} + \lambda B - 2\lambda C D]^{-1}$ $B = (A^\lambda + 1)/(A^\lambda - 1)$ $C = \frac{A^{(1+\lambda)/2}}{\lambda - 1 + (1+\lambda)A^\lambda}$ $D = 1 + \frac{\lambda A^{(\lambda-1)/2}}{A^\lambda - 1}$ $A = \exp(\mathrm{NTU}_1)$ $\lambda = (1 + R_1^2/4)^{1/2}$	same as Eq. (3.11) with $R_1 = 1$	$P_1 \to [1 + \frac{R_1}{2} + \lambda]^{-1}$

Table 7. Continued

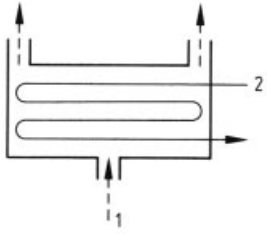 1–4 TEMA J shell-and-tube exchanger, shell and tube fluids mixed	3.12	$P_1 = \left[1 + \frac{R_1}{4}\left(\frac{1+3E}{1+E}\right) + \lambda B - 2\lambda C D\right]^{-1}$ $B = \frac{A^\lambda + 1}{A^\lambda - 1}$ $C = \frac{A^{(1+\lambda)/2}}{\lambda - 1 + (1+\lambda)A^\lambda}$ $D = 1 + \frac{\lambda A^{(\lambda-1)/2}}{A^\lambda - 1}$ $A = \exp(\mathrm{NTU}_1)$ $E = \exp(R_1 \mathrm{NTU}_1/2)$ $\lambda = (1 + R_1^2/16)^{1/2}$	same as Eq. (3.12) with $R_1 = 1$	$P_1 \to [1 + \frac{3R_1}{4} + \lambda]^{-1}$
Limit of 1–n TEMA J shell-and-tube exchangers for large N. Shell and tube fluids mixed	3.13	Eq. (2.4) applies in this limit	same as for Eq. (2.4)	same as for Eq. (2.4)
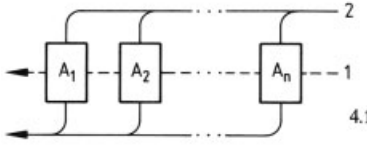	4.1.1	$P_1 = 1 - \prod_{i=1}^{n} (1 - P_{1A_i})$	same as Eq. (4.1.1)	
	4.1.2	$\frac{1}{R_1} = \sum_{i=1}^{n} \frac{1}{R_{1A_i}}$	$1 = \sum_{i=1}^{n} \frac{1}{R_{1A_i}}$	
Parallel coupling of n exchangers. Fluid 2 split arbitrarily into n streams	4.1.3	$\mathrm{NTU}_1 = \sum_{i=1}^{n} \mathrm{NTU}_{1A_i}$	same as Eq. (4.1.3)	

Table 7. Continued

	4.2.1	$P_1 = \dfrac{\prod_{i=1}^{n}(1 - R_1 P_{1A_i}) - \prod_{i=1}^{n}(1 - P_{1A_i})}{\prod_{i=1}^{n}(1 - R_1 P_{1A_i}) - R_1 \prod_{i=1}^{n}(1 - P_{1A_i})}$	$P_1 = \dfrac{\sum_{i=1}^{n} \dfrac{P_{1A_i}}{1 - P_{1A_i}}}{1 + \sum_{i=1}^{n} \dfrac{P_{1A_i}}{1 - P_{1A_i}}}$	same as Eq. (1.1.1) (counterflow)
	4.2.2.	$R_1 = R_{1A_i}, \quad i = 1, \ldots, n$	$1 = R_{1A_i}, \quad i = 1, \ldots, n$	same as Eq. (4.2.2)
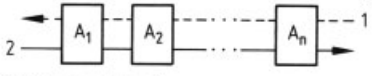	4.2.3	$\mathrm{NTU}_1 = \sum_{i=1}^{n} \mathrm{NTU}_{1A_i}$	same as Eq. (4.2.3)	$\mathrm{NTU}_1 \to 0$ $i = 1, \ldots, n$
Series coupling of n exchangers, overall counterflow arrangement.	4.2.4	$F = \dfrac{1}{\mathrm{NTU}_1} \sum_{i=1}^{n} \mathrm{NTU}_{1A_i} F_{A_i}$	same as Eq. (4.2.4)	same as Eq. (4.2.4)
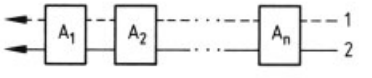	4.3.1	$P = \dfrac{1}{1 + R_1}\left\{1 - \prod_{i=1}^{n}[1 - (1 + R_1)P_{1A_i}]\right\}$	$P_1 = \dfrac{1 - \prod_{i=1}^{n}[1 - 2P_{1A_i}]}{2}$	same as Eq. (4.3.1)
Series coupling of n exchangers, overall parallel flow arrangement; stream symmetric if all A_i are stream symmetric	4.3.2	$R_1 = R_{1A_i}, \quad i = 1, \ldots, n$	$1 = R_{1A_i}, \quad i = 1, \ldots, n$	same as Eq. (4.3.2)
	4.3.3	$\mathrm{NTU}_1 = \sum_{i=1}^{n} \mathrm{NTU}_{1A_i}$	same as Eq. (4.3.3)	$\mathrm{NTU}_1 \to \infty$

* Table condensed from R. K. Shah and A. Pignotti: *Basic Thermal Design of Heat Exchangers*, National Science Foundation Report, Int-8601771, 1988.

In this table, all variables, except P_1, R_1, NTU_1, and F, are local or dummy variables not necessarily related to similar ones defined in the list of symbols given at the beginning of this article.

condensing fluids, h and U may vary strongly in the exchanger. The influence of some specific variations in U are now summarized.

For a single-pass counterflow exchanger, if U varies linearly with a temperature of either fluid or if it varies linearly with local T_x, the mean overall heat-transfer coefficient U_m is [35]:

$$\frac{q}{A}=U_m \Delta T_{lm}=\frac{U_2 \Delta T_1 - U_1 \Delta T_2}{\ln\left[(U_2 \Delta T_1)/(U_1 \Delta T_2)\right]} \tag{20}$$

where U_1 and U_2 are the values on the ends of the exchanger having temperature differences of ΔT_1 and ΔT_2.
U_m in Equation (20) is a good approximation for very viscous liquids and partial condensation.

If U varies linearly with the surface area A, integration of Equation (5) will yield

$$U_m=(U_1+U_2)/2 \tag{21}$$

If $1/U$ varies linearly with T_x, U_m is given by [7]:

$$\frac{1}{U_m}=\frac{1}{U_1}\frac{\Delta T_{lm}-\Delta T_2}{\Delta T_1-\Delta T_2}+\frac{1}{U_2}\frac{\Delta T_1-\Delta T_{lm}}{\Delta T_1-\Delta T_2} \tag{22}$$

If U_1 and U_2 do not vary significantly, Equation (21) is a good approximation. However, if they do vary considerably for condensers, Equation (22) is recommended. If a question arises as to which of the three Equations (20), (21), or (22) should be used, the exchanger should be divided into a number of sections, and for each section U should be treated as either a constant or an arithmetic mean.

For a single-pass crossflow exchanger, SIEDER and TATE [36] proposed to modify Equation (20) by the F factor when U varies linearly with T_c, T_h, or T_x:

$$q=U_m A F \Delta T_{lm} \tag{23}$$

where U_m is given by Equation (20). This equation provides the results within about 10 % accuracy in the practical ranges of F [37]. ROETZEL [38] has provided specific results and correlations for several cases of variable coefficients.

For multipass shell-and-tube exchangers with variable U, GARDNER and TABOREK [37] recommend the use of Equation (23) with appropriate values of F for the specific exchanger. Some specific results for the 1 – 2 exchanger with shell fluid mixed (TEMA E) are presented graphically in [28] in terms of $\psi - P$ charts.

5.4. Exchanger Pressure-Drop Analysis

Usually a fan, blower, or pump is used to promote fluid flow through individual sides of a heat exchanger. The required fluid pumping power is important for gases and viscous liquids. The fluid pumping power $\tilde{P}$ is approximately related to the core pressure drop in the exchanger [27]:

$$\tilde{P}=\frac{W\Delta p}{p}\approx\begin{cases}\dfrac{1}{2g_c}\dfrac{\mu}{\varrho^2}\dfrac{4L}{D_h}\dfrac{W^2}{D_h A_o} f\,Re & \text{for laminar flow} \quad (24a)\\[2ex] \dfrac{0.046}{2g_c}\dfrac{\mu^{0.2}}{\varrho^2}\dfrac{4L}{D_h}\dfrac{W^{2.8}}{A_o^{1.8}D_h^{0.2}} & \text{for turbulent flow} \quad (24b)\end{cases}$$

It is clear from Equations (24 a) and (24 b) that the fluid pumping power is strongly dependent upon the fluid density (particularly for low density fluids) in laminar and turbulent flows, and upon the viscosity in laminar flow. In addition, the pressure drop itself can be an important consideration when blowers and pumps are used for the fluid flow since they are head-limited. Also for condensing or evaporating fluids, the pressure drop affects the saturation temperature which in turn affects the heat-transfer rate. Hence, determination of the pressure drop in the exchanger is important.

The pressure drop associated with a heat exchanger consists of core (matrix or a portion with active heat-transfer surface) pressure drop and the pressure drop associated with the fluid distribution devices such as inlet and outlet manifolds, headers, tanks, nozzles, ducting, etc.

The core pressure drop is determined separately on each fluid side. It consists of one or more of the following components depending on the exchanger construction: (1) frictional losses associated with fluid flow over the heat-transfer surface (this usually consists of skin friction plus form drag), (2) momentum effect (pressure drop or rise due to the fluid density changes in the core), (3) pressure drop associated with sudden contraction and expansion at the core inlet and outlet, and (4) gravity effect due to the change in elevation between the inlet and outlet of the exchanger. The gravity effect is generally negligible for gases. For vertical liquid flows through

the exchanger, the pressure drop or rise ("static head") due to the elevation change is given by

$$\Delta p = \pm \frac{\varrho_m g L}{g_c} \tag{25}$$

where the "+" sign denotes vertical upflow (i.e., pressure drop), the "−" sign denotes vertical downflow (i.e., pressure rise), g is the gravitational acceleration, and L is the exchanger length.

The first three components of the core pressure drop are now discussed for several major exchangers. The pressure drop on the shellside of a shell-and-tube exchanger is presented in Section 8.2.1.

5.4.1. Plate – Fin Exchangers

For the plate-fin exchanger, all three components are considered in the core pressure drop evaluation [39]:

$$\frac{\Delta p}{p_1} = \frac{G^2}{2 g_c} \frac{1}{p_1 \varrho_1} \Bigg[\underbrace{(1 - \sigma^2 + K_e)}_{\text{entrance effect}} + \underbrace{f \frac{L}{r_h} \varrho_i \left(\frac{1}{\varrho}\right)_m}_{\text{core friction}} + \underbrace{2\left(\frac{\varrho_i}{\varrho_o} - 1\right)}_{\text{flow acceleration}} - \underbrace{(1 - \sigma^2 - K_e) \frac{\varrho_i}{\varrho_o}}_{\text{exit effect}} \Bigg] \tag{26}$$

where f is the Fanning friction factor (see Table 8), K_c and K_e are flow contraction (entrance) and expansion (exit) loss coefficients [39], G is fluid mass velocity (W/A_o), and σ is a ratio of minimum free flow area to frontal area.

K_c and K_e for four different entrance flow passage geometries are presented in [27], [39]. The entrance and exit losses are important at low values of σ and L (short cores), high values of Re, and for gases; they are negligible for liquids. The values of K_c and K_e presented in [27], [39] apply to long tubes for which flow is fully developed at the exit. For partially developed flows, K_c is lower and K_e is higher than that for fully developed flows. For interrupted surfaces, flow is never fully developed. For highly interrupted fin geometries, the entrance and exit losses are generally small compared to the core pressure drop, and the flow is well mixed; hence, K_c and K_e for $Re \to \infty$ should represent a good approximation.

The mean specific volume V_m or $(1/\varrho)_m$ in Equation (26) is given as follows: For liquids with any flow arrangement, or for a perfect gas with $C^* = 1$ and any flow arrangement (except for parallel flow),

$$\left(\frac{1}{\varrho}\right)_m = V_m = \frac{V_i + V_o}{2} = \frac{1}{2}\left(\frac{1}{\varrho_i} + \frac{1}{\varrho_o}\right) \tag{27}$$

where V is the specific volume in m^3/kg (ft^3/lbm).

For a perfect gas with $C^* = 0$ and any flow arrangement,

$$\left(\frac{1}{\varrho}\right)_m = \frac{\tilde{R}}{p_{ave}} T_{lm} \tag{28}$$

Here $\tilde{R}$ is the gas constant in $J\,kg^{-1}\,K^{-1}$, $p_{ave} = (p_i + p_o)/2$, and $T_{lm} = T_{const.} + \Delta T_{lm}$ where $T_{const.}$ is the mean average temperature of the fluid on the other side of the exchanger; the log-mean temperature difference ΔT_{lm} is defined in Table 5.

The core frictional pressure drop in Equation (26) may be approximated as

$$\Delta p \approx \frac{4 f L G^2}{2 g_c D_h} \left(\frac{1}{\varrho}\right)_m \tag{29}$$

5.4.2. Tube – Fin Exchangers

The pressure drop inside the tube is computed using Equation (26) with proper values of f (see the equations in Table 10) and K_c and K_e from [27], [39] for circular tubes.

For flat fins on an array of tubes (see Fig. 12), the components of the core pressure drop (such as those in Eq. (26)) are the same with the following exception: the core friction and momentum effect take place within the core with $G = W/A_o$, where A_o is the minimum free flow area within the core, and the entrance and exit losses occur at the leading and trailing edges of the core with the associated flow area A'_o so that

$$W = G A_o = G'_o A'_o \text{ or } G' \sigma' = G \sigma \tag{30}$$

where σ is the ratio of free flow area to frontal area at the fin leading edges.

The pressure drop for flow normal to a tube bank with flat fins is then given by

Table 8. Important dimensionless groups for internal flow forced convection heat transfer and flow friction, useful in heat exchanger design

Dimensionless groups	Definitions and working relationships	Physical meaning and comments
Reynolds number	$Re = \frac{\varrho v D_h}{\mu} = \frac{G D_h}{\mu}$	a flow modulus, proportional to the ratio of flow momentum rate ("inertia force") to viscous force
Fanning friction factor	$f = \frac{\tau_w}{(\varrho v^2/2 g_c)}$ $f = \Delta p^* \frac{r_h}{L} = \frac{\Delta p}{(\varrho v^2/2 g_c)} \frac{r_h}{L}$	the ratio of wall shear (skin frictional) stress to the flow kinetic energy per unit volume commonly used in heat transfer literature
Apparent Fanning friction factor	$f_{app} = \Delta p^* \frac{r_h}{L}$	includes the effects of skin friction and the change in the momentum rates in the entrance region (developing flows)
Incremental pressure drop number	$K(z) = (f_{app} - f_{fd}) \frac{L}{r_h}$ $K(\infty)$ = constant for $z \to \infty$	represents the excess dimensionless pressure drop in the entrance region over that for fully developed flow
Darcy friction factor	$f_D = 4f = \Delta p^* \frac{D_h}{L}$	four times the Fanning friction factor commonly used in fluid mechanics literature
Euler number	$Eu = \Delta p^* = \frac{\Delta p}{(\varrho v^2/2 g_c)}$	the pressure drop normalized with respect to the dynamic velocity head commonly used in the Soviet literature
Dimensionless axial distance for the fluid flow problem	$z^+ = \frac{z}{D_h Re}$	the ratio of the dimensionless axial distance (z/D_h) to the Reynolds number useful in the hydrodynamic entrance region
Nusselt number	$Nu = \frac{h}{k/D_h} = \frac{q'' D_h}{k(T_w - T_m)}$	the ratio of the convective conductance h to the pure molecular thermal conductance k/D_h
Stanton number	$St = \frac{h}{G c_p}$ $St = \frac{Nu}{Pe} = \frac{Nu}{Re Pr}$	the ratio of convection heat transfer (per unit duct surface area) to amount virtually transferable (per unit of flow cross-sectional area) no dependence upon any geometric characteristic dimension
Colburn factor	$j = St\, Pr^{2/3} = (Nu\, Pr^{-1/3})/Re$	a modified Stanton number to take into account the moderate variations in the Prandtl number for $0.5 \lesssim Pr \lesssim 10.0$ in turbulent flow
Prandtl number	$Pr = \frac{\nu}{\alpha} = \frac{\mu c_p}{k}$	a fluid property modulus representing the ratio of momentum diffusivity to thermal diffusivity of the fluid
Péclet number	$Pe = \frac{\varrho c_p v D_h}{k} = \frac{v D_h}{\alpha} = Re Pr$	proportional to the ratio of thermal energy convected to the fluid to thermal energy conducted axially within the fluid the inverse of Pe indicates relative importance of fluid axial heat conduction
Dimensionless axial distance for the heat transfer problem	$z^* = \frac{z}{D_h Pe} = \frac{z}{D_h Re Pr}$	useful in describing the thermal entrance region heat transfer results
Graetz number	$Gz = \frac{W c_p}{k L} = \frac{Pe P}{4L} = \frac{P}{4 D_h} \frac{1}{z^*}$ $Gz = \pi/(4z^*)$ for a circular tube	conventionally used in the chemical engineering literature related to z^* as shown when the flow length in Gz is treated as a length variable

$$\frac{\Delta p}{p_i} = \frac{G^2}{2 g_c} \frac{1}{p_i \varrho_i} \left[f \frac{L}{r_h} \varrho_i \left(\frac{1}{\varrho} \right)_m + 2 \left(\frac{\varrho_i}{\varrho_o} - 1 \right) \right] + \frac{G'^2}{2 g_c} \frac{1}{p_i \varrho_i} \left[\left(1 - \sigma'^2 + K_c \right) - \left(1 - \sigma'^2 - K_c \right) \frac{\varrho_i}{\varrho_o} \right] \quad (31)$$

For individually finned tubes shown in Figure 11, flow expansion and contraction take place along each tube row, and the magnitude is of the same order as that at the entrance and exit. Hence, the entrance and exit losses are generally lumped into the core friction factor. Equation (26) for individually finned tubes then reduces to

$$\frac{\Delta p}{p_i} = \frac{G^2}{2 g_c} \frac{1}{p_i \varrho_i} \left[f \frac{L}{r_h} \varrho_i \left(\frac{1}{\varrho} \right)_m + 2 \left(\frac{\varrho_i}{\varrho_0} - 1 \right) \right] \quad (32)$$

5.4.3. Regenerators

For regenerator matrices having cylindrical passages, the pressure drop is computed using Equation (26) with appropriate values of f, K_c, and K_e. For regenerator matrices made up of any porous material (such as checkerwork, wire mesh, spheres, copper wools, etc.), the pressure drop is calculated using Equation (32) in which the entrance and exit losses are included in the friction factor f.

5.4.4. Plate Heat Exchangers

Pressure drop in a plate heat exchanger consists of three components: (1) pressure drop associated with the inlet and outlet manifolds and ports, (2) pressure drop within the core (plate

passages), and (3) pressure drop due to the elevation change. The pressure drop in the manifolds and ports should be kept as low as possible (generally $< 10\,\%$, but may be as high as $25-30\,\%$ in some designs). Empirically, it is calculated as approximately 1.5 times the inlet velocity head per pass. Since the entrance and exit losses in the core (plate passages) cannot be determined experimentally, they are included in the friction factor for the given plate geometry. The pressure drop (rise) caused by the elevation change for liquids is given by Equation (25). Hence, the pressure drop on one fluid side in a plate heat exchanger is given by

$$\Delta p = \frac{1.5G^2N}{2g_c\varrho_i} + \frac{4fLG^2}{2g_cD_e}\left(\frac{1}{\varrho}\right)_m + \left(\frac{1}{\varrho_o} - \frac{1}{\varrho_i}\right)\frac{G^2}{g_c} \pm \frac{\varrho_m gL}{g_c} \tag{33}$$

where N is the number of passes on the given fluid side, and D_e is the equivalent diameter of flow passages (usually twice the plate spacing).

6. Single-Phase Heat-Transfer and Pressure-Drop Correlations

6.1. Basic Concepts and Dimensionless Groups

Heat-transfer characteristics of an exchanger surface are presented in terms of the Nusselt number, Stanton number, or Colburn factor vs. the Reynolds number, z^*, or Graetz number. Flow friction characteristics are presented in terms of the Fanning friction factor vs. Re or z^+. These and other important dimensionless groups used in presenting and correlating internal flow forced convection heat transfer are summarized in Table 8 with their definitions and physical meanings. Where applicable, the hydraulic diameter D_h is used as a characteristic dimension in all dimensionless groups. A number of different definitions are used in literature for some of the dimensionless groups; the user should pay particular attention to the specific definitions used in any research paper before using specific results. This is particularly true for the Nusselt number where many different temperature differences are used in the definition of h, and for f, Re, and other dimensionless groups having characteristic dimensions different from D_h. As an example of these correlating groups, basic heat-transfer and flow friction characteristics for air flow in a long circular tube are presented in Figure 33. There is a parallel behavior of j vs. Re and f vs. Re curves, but no such parallelism exists between Nu and f vs. Re curves. The parallel behavior of j and f vs. Re is useful for: (1) identifying erroneous test data for some specific surfaces for which parallel behavior is expected but indicated otherwise by test results; (2) identifying specific flow phenomena in which the friction behavior is different from the heat transfer behavior (such as flow over rough surfaces in turbulent flow); and (3) predicting the f factors for an unknown surface when the j factors are known by some predictive method. It should be remembered that j vs. Re can readity be converted to Nu vs. Re curves or vice versa because $j = (Nu\,Pr^{-1/3})/Re$ by definition.

Figure 33 shows three flow regimes: laminar, transition, and turbulent. This is characteristic of fully developed flow in continuous flow passage geometries such as a long circular tube, triangular tube, etc. Generally, the compact interrupted surfaces do not have a sharp "dip" in the transition region ($Re \sim 1500-10\,000$) as shown for the circular tube.

6.2. Theoretical Solutions and Correlations for Simple Geometries

Flow passages in most compact heat exchangers are complex with frequent boundary layer interruptions; some heat exchangers (particularly the tube side of shell-and-tube exchangers and highly compact regenerators) have continuous flow passages. The velocity and temperature profiles across the flow cross section are generally fully developed in the continuous flow passages, whereas they develop at each boundary layer interruption in an interrupted surface. The heat-transfer and flow friction characteristics are generally different for fully developed flows and developing flows. Analytical results for developed and developing flows for simple flow passage geometries follow. For complex surface geometries, the basic surface characteristics are primarily obtained experimentally [27]; the pertinent correlations are presented in Section 6.3.

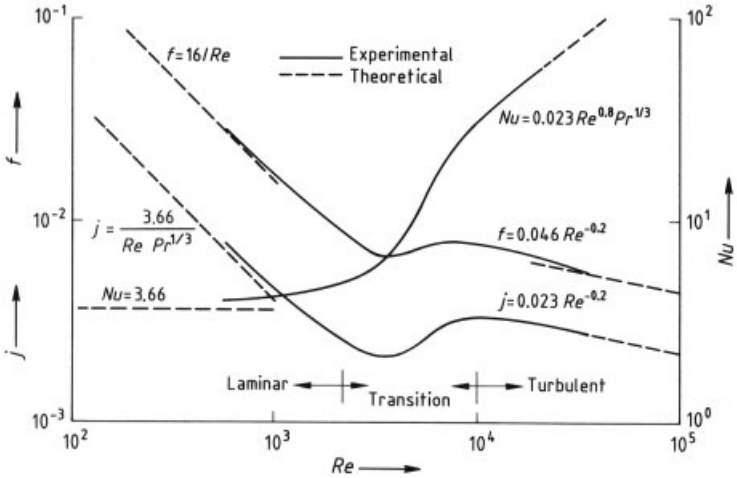

Figure 33. Basic heat transfer and flow friction characteristics for air flow through a long circular tube

Three important thermal boundary conditions for heat exchangers are Ⓣ, Ⓗ1, Ⓗ2. The Ⓣ boundary condition refers to constant wall temperature, both axially and peripherally throughout the passage length. The wall heat-transfer rate is constant in the axial direction while the wall temperature at any cross section is constant in the peripheral direction for the Ⓗ1 boundary condition. The wall heat-transfer rate is constant in the axial direction as well as in the peripheral direction for the Ⓗ2 boundary condition. The Ⓗ1 boundary condition is realized for highly conductive materials where the temperature gradients in the peripheral direction are at a minimum; the Ⓗ2 boundary condition is realized for very poorly conducting materials for which temperature gradients exist in the peripheral direction. For intermediate thermal conductivity values, the boundary condition will be in between that of Ⓗ1 and Ⓗ2. In general, $Nu_{\mathrm{H1}} > Nu_{\mathrm{T}}$, $Nu_{\mathrm{H1}} \geq Nu_{\mathrm{H2}}$, and $Nu_{\mathrm{H2}} \lessgtr Nu_{\mathrm{T}}$.

The heat-transfer rate in laminar duct flow is very sensitive to the thermal boundary condition. Hence, it is essential to carefully identify the thermal boundary condition in laminar flow. The heat-transfer rate in turbulent duct flow is insensitive to the thermal boundary condition for most common fluids ($Pr \gtrsim 0.7$) except for liquid metals ($Pr < 0.03$). Hence, there is generally no need to identify the thermal boundary condition in turbulent flow for all fluids except liquid metals.

6.2.1. Fully Developed Flows

Laminar Flow. Nusselt numbers for fully developed laminar flow are constant, but depend on the flow passage geometry and thermal boundary conditions. The product of the Fanning friction factor and the Reynolds number is also constant, but is dependent on the flow passage geometry. Fully developed laminar flow problems are analyzed extensively in [40], [41]; most of the analytical solutions are also presented in closed-form equations in [41]. Solutions for some technically important flow passages are presented in Table 9.

A number of observations may be made from the results of Table 9 for fully developed laminar flow through straight constant cross-sectional ducts:

1) Flow passage geometry strongly influences the values of Nu and fRe. Rectangular flow passages which approach a small aspect ratio exhibit the highest Nu and fRe.
2) The thermal boundary conditions Ⓗ1, Ⓗ2, and Ⓣ have a strong influence on the Nusselt number. Depending on the flow geometry, j factors for the Ⓗ1 boundary condition may be of the order of 50 % greater than that for the Ⓗ2 boundary condition, and about 20 % greater than that for the Ⓣ boundary condition.

Table 9. Solutions for heat transfer and friction for fully developed laminar flow through specified ducts [40]

Geometry ($L/D_h > 100$)		Nu_{H1}	Nu_{H2}	Nu_T	fRe	$\frac{j_{H1}}{f}$ [a]	$K(\infty)$ [b]	L_{hy}^{+} [c]
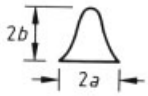	$\frac{2b}{2a} = \frac{\sqrt{3}}{2}$	3.014	1.474	2.39	12.630	0.269	1.739	0.04
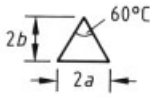	$\frac{2b}{2a} = \frac{\sqrt{3}}{2}$	3.111	1.892	2.47	13.333	0.263	1.818	0.04
	$\frac{2b}{2a} = 1$	3.608	3.091	2.976	14.227	0.286	1.433	0.090
		4.002	3.862	3.34	15.054	0.299	1.335	0.086
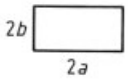	$\frac{2b}{2a} = \frac{1}{2}$	4.123	3.017	3.391	15.548	0.299	1.281	0.085
		4.364	4.364	3.657	16.000	0.307	1.25	0.056
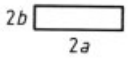	$\frac{2b}{2a} = \frac{1}{4}$	5.331	2.94	4.439	18.233	0.329	1.001	0.078
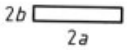	$\frac{2b}{2a} = \frac{1}{6}$	6.049	2.93	5.137	19.702	0.346	0.885	0.070
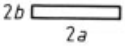	$\frac{2b}{2a} = \frac{1}{8}$	6.490	2.94	5.597	20.585	0.355	0.825	0.063
	$\frac{2b}{2a} = 0$	8.235	8.235	7.541	24.000	0.386	0.674	0.011

[a] $j_{H1}/f = Nu_{H1}\, Pr^{-1/3}/(fRe)$ with $Pr = 0.7$. Similarly, values of j_{H2}/f and j_T/f may be computed.
[b] $K(\infty)$ for sine and equilateral triangular channels may be too high [40]; $K(\infty)$ for some rectangular and hexagonal channels is interpolated based on the recommended values in Ref. [40].
[c] L_{hy}^{+} for sine and equilateral triangular channels is too low [40], so use with caution. L_{hy}^{+} for rectangular channels is based on the faired curve drawn through the recommended value in [40]. L_{hy}^{+} for a hexagonal channel is an interpolated value.

3) As $Nu = h\,D_h/k$, a constant Nu implies a convective coefficient h independent of the flow velocity.
4) An increase in h is best achieved by reducing D_h, or by a change in the type of geometry for a specified fluid.
5) Since $fRe = \text{const.}$, $f \propto 1/Re \propto 1/v$. In this case, it can be shown that $\Delta p \propto v$.

Transition Flow. The lower limit of the critical Reynolds number Re_{cr} at which transition to turbulent flow occurs in smooth passages is a function of the passage inlet configuration. With a sharp square inlet configuration, Re_{cr} is ca. 10 – 15 % lower than that with a rounded inlet configuration. The lower limits of Re_{cr} for various passages with a sharp square inlet configuration vary from about 2000 to 3100 [42]. The upper

limit of Re_{cr} may be taken as 10^4 for most practical purposes.

BHATTI and SHAH [42] developed a Fanning friction factor formula for laminar, transition, and turbulent flow regimes in a circular passage; this is presented in Table 10. The transition flow f data for noncircular passages are rather sparse; the equation for the circular tube in Table 10 may be used to obtain fair estimates of f for noncircular flow passages (having no sharp corners) using the hydraulic diameter as the characteristic dimension.

The heat transfer results for transition flow are rather uncertain in view of the large number of parameters necessary to characterize the heat-affected transition flow. The following Nusselt number correlation [42] for $0 < Pr < \infty$ and $2100 \leq Re \leq 10^6$ spanning laminar, transition, and turbulent flow regimes in a circular passage, is recommended for calculating the transition flow $j = Nu\,Pr^{-1/3}/Re$:

$$Nu = \left\{ Nu_{\mathrm{L}}^{10} + \left[\frac{\exp\left[(2200 - Re)/365\right]}{Nu_{\mathrm{L}}^{2}} + \frac{1}{Nu_{\mathrm{T}}^{2}} \right]^{-5} \right\}^{1/10} \tag{34}$$

where

$$Nu_{\mathrm{L}} = \begin{cases} 3.657 & \text{for (T) boundary condition} \\ 4.364 & \text{for (H) boundary condition} \end{cases} \tag{35}$$

$$Nu_{\mathrm{T}} = Nu_0 + \frac{0.079\,(f/2)^{1/2}\,RePr}{\left(1 + Pr^{4/5}\right)^{5/6}} \tag{36}$$

$$Nu_{\mathrm{o}} = \begin{cases} 4.8 & \text{for (T) boundary condition} \\ 6.3 & \text{for (H) boundary condition} \end{cases} \tag{37}$$

Since no j factors are available for transition flow for noncircular passages, Equations (34) – (37) may be used to obtain fair estimates of j for noncircular passages (having no sharp corners) using the hydraulic diameter as the characteristic dimension.

Turbulent Flow. A compendium of available f and Nu correlations for circular and noncircular flow passages are presented in [42]. Table 10 is condensed from [42] summarizing the most accurate f and Nu correlations for smooth circular and noncircular passages.

It is generally accepted that the hydraulic diameter correlates Nu and f for fully developed turbulent flow in circular and noncircular ducts. This is true for the results accurate to within $\pm 15\,\%$ for most noncircular ducts. Exceptions are for those having sharp or acute-angled corners in the flow passage or concentric annuli with inner wall heating. In these cases, Nu and f could be lower than 15 % compared to the circular tube values. Table 10 can be used for more accurate correlations of Nu and f for noncircular ducts.

Roughness on the surface causes local flow separation and reattachment. This generally results in an increase in the friction factor as well as the heat-transfer coefficient. A roughness element has no effect on laminar flow, unless the height of the roughness element is not negligible compared to the flow cross section size. However, it exerts a strong influence on turbulent flow. Specific correlations to account for the influence of surface roughness are presented in [42], [43].

6.2.2. Hydrodynamically Developing Flows

Laminar Flow. Based on the solutions for laminar boundary layer development over a flat plate and fully developed flow in circular and some noncircular ducts, $f_{\mathrm{app}}Re$ can be correlated by the following equation:

$$f_{\mathrm{app}}Re = 3.44\left(z^{+}\right)^{-0.5} + \frac{K(\infty)/\left(4z^{+}\right) + fRe - 3.44\left(z^{+}\right)^{-0.5}}{1 + c'\left(z^{+}\right)^{-2}} \tag{38}$$

where $K(\infty)$, fRe, and c' are given in Table 11.

Table 10. Fully developed turbulent flow friction factors and Nusselt numbers ($Pr > 0.5$) for technically important smooth-walled ducts [41]

Duct geometry and characteristic dimension	Recommended correlations[a]
 Circular $D_h = 2a$	friction factor correlation for $2300 \le Re \le 10^7$ [41]: $f = A + \frac{B}{Re^{1/m}}$ where $A = 0.0054$, $B = 2.3 \times 10^{-8}$, $m = -2/3$ for $2100 \le Re \le 4000$ and $A = 1.28 \times 10^{-3}$, $B = 0.1143$, $m = 3.2154$ for $4000 \le Re \le 10^7$ [41]. Nusselt number correlation by Gnielinski for $2300 \le Re \le 5 \times 10^6$: $Nu = \frac{(f/2)(Re - 1000)\,Pr}{1 + 12.7\,(f/2)^{1/2}\,(Pr^{2/3} - 1)}$
 Flat $D_h = 4b$	use circular duct f and Nu correlations. Predicted f are up to 12.5% lower and predicted Nu are within $\pm 9\%$ of the most reliable experimental results.
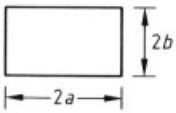 Rectangular $D_h = \frac{4ab}{a+b}$, $\alpha^* = \frac{2b}{2a}$ $\frac{D_l}{D_h} = \frac{2}{3} + \frac{11}{24}\alpha^*(2 - \alpha^*)$	f factors: (1) substitute D_l for D_h in the circular duct correlation, and calculate f from the resulting equation. (2) Alternatively, calculate f from $f = (1.0875 - 0.1125\,\alpha^*)\,f_c$ where f_c is the friction factor for the circular duct using D_h. In both cases, predicted f factors are within $\pm 5\%$ of the experimental results. Nusselt numbers: (1) With uniform heating at four walls, use ciruclar duct Nu correlation for an accuracy of $\pm 9\%$ for $0.5 \le Pr \le 100$ and $10^4 \le Re \le 10^6$. (2) With equal heating at two long walls, use circular duct correlation for an accuracy of $\pm 10\%$ for $0.5 < Pr \le 10$ and $10^4 \le Re \le 10^5$. (3) With heating at one long wall only, use circular duct correlation to get approximate Nu values for $0.5 < Pr < 10$ and $10^4 \le Re \le 10^6$. These calculated values may be up to 20% higher than the actual experimental values.
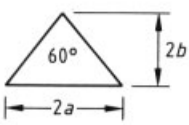 Equilateral triangular $D_h = 2\sqrt{3}\,a = 4b/3$ $D_l = \sqrt{3}\,a = 2b/3\sqrt{3}$	use circular duct f and Nu correlations with D_h replaced by D_l. Predicted f are within $+3\%$ and -11% and predicted Nu within $+9\%$ of the experimental values.
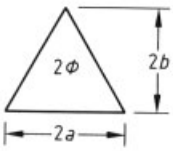 Isosceles triangular $D_h = \frac{4ab}{a + \sqrt{a^2 + b^2}}$ $\frac{D_g}{D_h} = \frac{1}{2\pi}\left[3 \ln \cot \frac{\theta}{2} + 2 \ln \tan \frac{\varphi}{2} - \ln \tan \frac{\theta}{2}\right]$ where $\theta = (90° - \varphi)/2$	for $0 < 2\varphi < 60°$, use circular duct f and Nu correlations with D_h replaced by D_g; for $2\varphi = 60°$, replace D_h by D_l (see above); and for $60° < 2\varphi \le 90°$ use circular duct correlations directly with D_h. Predicted f and Nu are within $+9\%$ and -11% of the experimental values. No recommentations can be made for $2\varphi > 90°$ due to lack of the experimental data.

Table 10. Continued

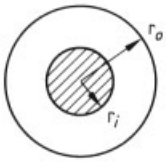

Concentric annular

$D_h = 2(r_o - r_i)$,

$r^* = \frac{r_i}{r_o}$

$\frac{D_1}{D_h} = \frac{1 + r^{*2} + (1 - r^{*2})/\ln r^*}{(1 - r^*)^2}$

f Factors: (1) Substitute D_1 for D_h in the circular duct correlation, and calculate f from the resulting equation. (2) Alternatively, calculate f from $f = (1 + 0.0925\, r^*)\, f_c$ where f_c is the friction factor for the circular duct using D_h. In both cases, predicted f factors are within $\pm 5\%$ of the experimental results.

Nusselt Numbers: In all the following recommendations, use D_h with a wetted perimeter in Nu and Re: (1) Nu at the outer wall can be determined from the circular duct correlation within the accuracy of about $\pm 10\%$ regardless of the condition at the inner wall. (2) Nu at the inner wall cannot be determined accurately regardless of the heating/cooling condition at the outer wall.

[a] The friction factor and Nusselt number corrlations for the circular duct are the most reliable and agree with a large amount of the experimental data within $\pm 2\%$ and $\pm 10\%$ respectively. The correlations for all other duct geometries are not as good as those for the circular duct on an absolute basis.

Table 11. $K(\infty)$, fRe, and c' for use in Equation (38)

	$K(\infty)$	fRe	$10^3\, c'$	% error
α^*	*Rectangular ducts*			
1.00	1.43	14.277	0.29	± 2.3
0.50	1.28	15.548	0.21	± 1.9
0.20	0.931	19.071	0.076	± 1.7
0.00	0.674	24.000	0.029	± 2.4
2φ	*Equilateral triangular duct*			
60°	1.69	13.333	0.53	± 2.4
r^*	*Concentric annular ducts*			
0	1.25	16.000	0.212	± 1.9
0.05	0.830	21.567	0.050	± 2.0
0.10	0.784	22.343	0.043	± 1.9
0.50	0.688	23.813	0.032	± 2.2
0.75	0.678	23.967	0.030	± 2.1
1.00	0.674	24.000	0.029	± 2.4

The $f_{\mathrm{app}}\, Re$ factors for four passage geometries calculated from the foregoing formulas are compared in [40], [41].

Turbulent Flow. $f_{\mathrm{app}}\, Re$ for turbulent flow depends on Re in addition to z^+. A closed-form formula for $f_{\mathrm{app}}\, Re$ is given in [42], [43] for developing turbulent flow.

The hydrodynamic entrance length for developing laminar and turbulent flow are given by [43]:

$$L_{\mathrm{hy}}/D_{\mathrm{h}} = 0.056 Re$$

$$\text{for laminar flow } (Re \leq 2100) \tag{39}$$

$$L_{\mathrm{hy}}/D_{\mathrm{h}} = 1.359 Re^{1/4}$$

$$\text{for turbulent flow } \left(Re \geq 10^4\right) \tag{40}$$

6.2.3. Thermally Developing Flows

Laminar Flow. Thermal entry length solutions with developed velocity profiles are summarized in [40], [41] for a large number of practically important flow passage geometries with extensive comparisons.

SHAH and LONDON [40] proposed the following correlations for thermal entrance solutions for circular and noncircular ducts having laminar developed velocity profiles and developing temperature profiles.

$$Nu_{\mathrm{z,T}} = 0.427\, (f\, Re)^{1/3}\, (z^*)^{-1/3} \tag{41}$$

$$Nu_{\mathrm{m,T}} = 0.641\, (f\, Re)^{1/3}\, (z^*)^{-1/3} \tag{42}$$

$$Nu_{\mathrm{z,H1}} = 0.517\, (f\, Re)^{1/3}\, (z^*)^{-1/3} \tag{43}$$

$$Nu_{\mathrm{m,H1}} = 0.775\, (f\, Re)^{1/3}\, (z^*)^{-1/3} \tag{44}$$

where f is the Fanning friction factor for fully developed flow, Re is the Reynolds number, and $z^* = z/(D_{\mathrm{h}} Re\, Pr)$. For interrupted surfaces,

$z = l_{\mathrm{ef}}$. Equations (41) – (44) are recommended for $z^* < 0.001$.

The following observations may be made from Equations (41) – (44) and solutions for laminar flow surfaces having developing temperature profiles given in [40], [41].

1) The influence of thermal boundary conditions on the convective behavior appears to be of the same order as that for fully developed flow.
2) Since $Nu \propto (z^*)^{-1/3} = [z/(D_{\mathrm{h}} Re\, Pr)]^{-1/3}$, then $Nu \propto Re^{1/3} \propto v^{1/3}$. Therefore h varies as $v^{1/3}$.
3) Since the velocity profiles is considered fully developed, $\Delta p \propto v_{\mathrm{m}}$ as noted earlier.
4) The influence of the duct shape on thermally developed Nu is not as great as that for the fully developed Nu.

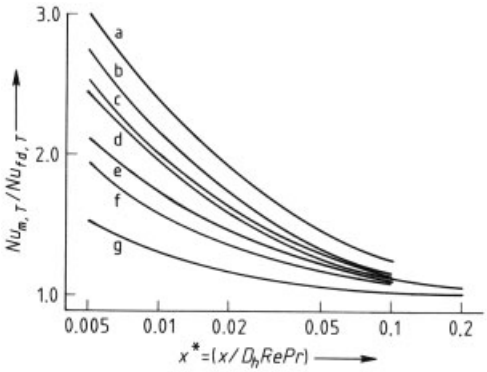

Figure 34. The ratio of laminar developing to developed flow Nu for different ducts; the velocity profile is developed for both Nu's
a) Equilateral triangular; b) Square duct; c) $\alpha^* = 1/2$ Rectangular duct; d) Circular duct; e) $\alpha^* = 1/4$ Rect. duct; f) $\alpha^* = 1/6$ Rect. duct; g) Parallel plates

The theoretical ratio Nu_{m}/Nu_{fd} is shown in Figure 34 for several passage geometries having constant wall temperature boundary conditions. Several observations may be made from this figure:

1) The Nusselt numbers in the entrance region and hence the heat transfer coefficients could be 2 – 3 times higher than the fully developed values depending on the interruption length $l^*_{\mathrm{ef}} = z^*$.
2) At $z^* \approx 0.1$, the local Nusselt number approaches the fully developed value, but the value of the mean Nusselt number can be significantly higher for a channel of length $l^*_{\mathrm{ef}} = z^* = 0.1$.
3) The order of increasing $\mathrm{Nu}_{\mathrm{m}}/Nu_{\mathrm{fd}}$ as a function of channel shape at a given z^* is the opposite of Nu_{fd} in Table 9. For a highly interrupted surface, a basic inferior passage geometry for fully developed flow (such as triangular) will not be penalized in terms of low Nu or low h in developing flow.
4) A higher value of $Nu_{\mathrm{m}}/Nu_{\mathrm{fd}}$ at $z^* = 0.1$ means that the flow channel has a longer entrance region.

Turbulent Flow. The thermal entry length solutions for smooth ducts of several cross-sectional geometries have been summarized [42]. As for laminar flow, the Nusselt numbers in the thermal region are higher than those in the fully developed region. However, unlike laminar flow, $Nu_{z,\mathrm{T}}$ and $Nu_{z,\mathrm{H1}}$ are very nearly the same for turbulent flow. The local and mean Nusselt numbers for a circular tube with Ⓣ and Ⓗ boundary conditions are [42]:

$$\frac{Nu_x}{Nu_\infty} = 1 + \frac{c}{10\,(z/D_{\mathrm{h}})}\,; \quad \frac{Nu_{\mathrm{m}}}{Nu_\infty} = 1 + \frac{c}{z/D_{\mathrm{h}}} \tag{45}$$

where Nu_∞ stands for the fully developed Nu_{T} or Nu_{H} derived from the formulas in Table 10, and

$$c = \frac{(z/D_{\mathrm{h}})^{0.1}}{Pr^{1/6}} \left(0.68 + \frac{3000}{Re^{0.81}}\right) \tag{46}$$

This correlation is valid for $z/D_{\mathrm{h}} > 3$, $3500 < Re < 10^5$, and $0.7 < Pr < 75$. It agrees within $\pm 12\,\%$ with the experimental measurements for $Pr = 0.7$.

6.2.4. Simultaneously Developing Flows

Laminar Flow. In simultaneously developing flow, both the velocity and temperature profiles develop in the entrance region. The available analytical solutions are summarized in [40], [41]. The theoretical entrance region Nusselt numbers for simultaneously developing flow are higher than those for thermally developing and hydrodynamically developed flow. These theoretical solutions do not take into account the wake effect or secondary flow effect which are present in flow over interrupted heat-transfer

surfaces. Experimental data indicate that the interrupted heat-transfer surfaces do not achieve higher heat-transfer coefficients predicted for the simultaneously developing flows. The results for thermally developing flows (and developed velocity profiles) are in better agreement with the experimental data for interrupted surfaces and hence are recommended for design purpose.

Turbulent Flow. The Nusselt numbers for simultaneously developing turbulent flow are practically the same as the Nusselt numbers for the thermally developing turbulent flow [42]. However, the Nusselt numbers for simultaneously developing flow are sensitive to the passage inlet configuration.

Table 12 summarizes the dependence of Δp and h on v for developed and developing laminar and turbulent flows. Although these results are for the circular tube, the general functional relationship should be valid for noncircular ducts as a first approximation.

Table 12. Dependence of pressure drop and heat-transfer coefficient on the flow mean velocity for internal flow in a constant cross-sectional duct

Flow type	$\Delta p \propto v^{\mathrm{p}}$		$h \propto v^{\mathrm{q}}$	
	Laminar	Turbulent	Laminar	Turbulent
Fully developed	v	$v^{1.8}$	v^{0}	$v^{0.8}$
Hydrodynamically developing	$v^{1.5}$	$v^{1.8}$	–	–
Thermally developing	v	$v^{1.8}$	$v^{1/3}$	$v^{0.8}$
Simultaneously developing	$v^{1.5}$	$v^{1.8}$	$v^{1/2}$	$v^{0.8}$

6.3. Empirical Correlations for Complex Geometries

The results presented in the preceding sections are for well-defined flow passages with unobstructed constant area and essentially unidirectional flow. The flows encountered in actual heat exchangers are very complex, involving such processes as separation, reattachment, recirculation (i.e., backflow), and vortex action. These flow processes significantly influence the hydraulic and thermal performance of a heat exchanger surface. However, the detailed manner in which these processes influence j and f is difficult to treat analytically. Hence, f and j factors for complex heat exchanger surfaces have to be determined experimentally.

There is only a limited amount of information in the literature available for j and f data for complex surfaces used in present-day commercial devices. Such information is frequently retained as business-sensitive proprietary information by heat exchanger manufacturers. The primary source of j and f data for complex surfaces is found in [39]: this literature source contains information for over 100 surfaces, many of which, however, are manufactured with technologies that are now obsolete.

Correlations have been developed in literature for tubular, plate, and extended surface heat exchangers which are summarized next.

6.3.1. Tubular Exchangers

The principal geometric parameters that influence fluid flow and heat transfer of a tube bank are: (1) tube diameter, (2) longitudinal and transverse tube pitches, (3) number of tube rows, (4) tube arrangement, and (5) fin geometry and type, if any. Figure 25 shows the two most common tube arrangements–the inline and staggered arrangements in four-tube layouts.

Flow conditions within the tube bank are dominated by boundary layer separation effects and wake interactions; therefore, the heat-transfer coefficients and friction factors associated with a tube are determined by the tube location in the bank. Due to the turbulence caused by the first few rows, the inner rows of a bare tube bank will have higher heat-transfer coefficients and friction factors. For bare or finned tube banks, the heat-transfer coefficient will decrease with an increasing number of tube rows in an inline bank due to the flow bypass effect; it will, however, increase with the number of tube rows in a staggered bank, due to the shedding of turbulent eddies, thereby providing good fluid mixing for downstream tube rows. In bare tube banks, the dimensionless heat-transfer coefficients and friction factors approach asymptotic values beyond about 16 tube rows, and for finned tube banks beyond about 4 tube rows [44].

Bare Tube Banks. Many correlations are published for friction factors and Nusselt numbers for bare tube banks. One of the most com-

prehensive correlations by ZUKAUSKAS [44] is now summarized. He provided the modified friction factor f'/χ vs. Re_d correlations in graphical form for (a) inline arrangement (90° tube layout), and (b) staggered arrangement (30° tube layout) as shown in Figures 35 and 36 for specific values of X_t and X_l. For other inline and staggered tube arrangements, a correction factor χ is obtained from the insets of these figures. Here f' is defined by

$$\Delta p = 4f' \frac{G^2}{2g_c \varrho} N_r \tag{47}$$

where f' is an empirical Fanning friction factor per tube row, N_r is the number of tube rows in flow direction, and G is the mass velocity based on the minimum free flow area.

The Nusselt number for flow normal to the tube banks is also dependent upon the tube arrangements, tube pitch-to-diameter ratio, number of tube rows, Reynolds number, and Prandtl number. ZUKAUSKAS [44] also presented correlations for the mean Nusselt number $Nu_m = h_m d_o/k$ as

$$Nu_m = F_c \, (Nu_m)_{16\text{rows}} \tag{48}$$

Values of Nu_m for 16 and more tube rows are presented in Table 13 for inline (90° tube layout) and staggered (30° tube layout) arrangements. For all expressions in Table 13, fluid properties in Nu, Re_d, and Pr are evaluated at the bulk mean temperature and for Pr_w at the wall temperature. The tube row correction factor F_c is presented in Figure 37 as a function of the number of tube rows N_r for inline and staggered arrangements.

Table 13. Mean Nusselt number correlations for inline and staggered tube bundles [44]

Range of Re	Recommended correlations
	Inline tube bundles
$10^0 - 10^2$	$Nu = 0.9\, Re^{0.4}\, Pr^{0.36}\, (Pr/Pr_w)^{0.25}$
$10^2 - 10^3$	$Nu = 0.52\, Re^{0.5}\, Pr^{0.36}\, (Pr/Pr_w)^{0.25}$
$10^3 - 2\times10^5$	$Nu = 0.27\, Re^{0.63}\, \mathrm{Pr}^{0.36}\, (Pr/Pr_w)^{0.25}$
$2\times10^5 - 2\times10^6$	$Nu = 0.033\, Re^{0.8}\, Pr^{0.4}\, (Pr/Pr_w)^{0.25}$
	Staggered tube bundles
$10^0 - 5\times10^2$	$Nu = 1.04\, Re^{0.4}\, Pr^{0.36}\, (Pr/Pr_w)^{0.25}$
$5\times10^2 - 10^3$	$Nu = 0.71\, Re^{0.5}\, Pr^{0.36}\, (Pr/Pr_w)^{0.25}$
$10^3 - 2\times10^5$	$Nu = 0.35\, (X_t/X_l)^{0.2}\, Re^{0.6} \cdot Pr^{0.36}\, (Pr/Pr_w)^{0.25}$
$2\times10^5 - 2\times10^6$	$Nu = 0.031\, (X_t/X_l)^{0.2}\, Re^{0.8} \cdot Pr^{0.36}\, (PrPr_w)^{0.25}$

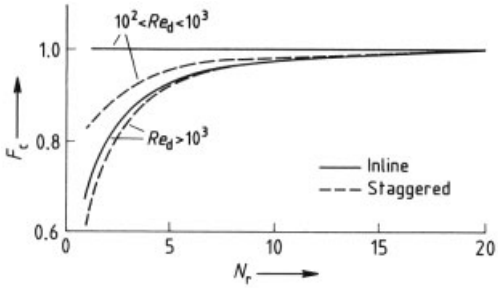

Figure 37. A correction factor to account for the tube-row effect for heat transfer for flow normal to bare tube banks [44]

Finned Tube Banks. Heat-transfer and flow friction correlations are summarized separately for two types of tube – fin surfaces: individually finned tubes, and flat fins on a tube array.

A staggered tube arrangement is preferred for multirow finned tube banks. The heat-transfer coefficient of an inline finned tube bank is substantially lower than that for a staggered arrangement because of the flow bypass through the gap if the fins are not touching. No correlations are available for inline tube banks. Therefore, staggered tube arrangements are dealt with in this section.

Individually Finned Circular Tubes. These fins are quite commonly used on the gas side. Many different interrupted circular fins have been used by industry. However, the data are mostly proprietary. Multiple regression-type correlations are only published for the plain fins, as follows.

For heat transfer, BRIGGS and YOUNG [45] recommend the following correlation for equilateral triangular staggered tube banks.

$$j = 0.134\, Re_d^{-0.319} \left(s'/l_f\right)^{0.2} \left(s'/\delta_f\right)^{0.11} \tag{49}$$

where $s' = (1/N_f) - \delta_f$ is the distance between adjacent fins. Equation (49) is valid for the following ranges: $1100 \le Re_d \le 18\,000$, $0.13 \le s'/l_f \le 0.63$, $1.01 \le s'/\delta_f \le 6.62$, $0.09 \le l_f/d_o \le 0.69$, $0.011 \le \delta_f/d_o \le 0.15$, $1.65 \le X_t/d_o \le 8.23$, fin root diameter d_o between 11.1 and 40.9 mm, and fin density N_f between 246 and 768 fins/m.

For isothermal pressure drops, based on data for an equilateral triangular and two isosceles triangular staggered tube banks, the following correlation is recommended [46]:

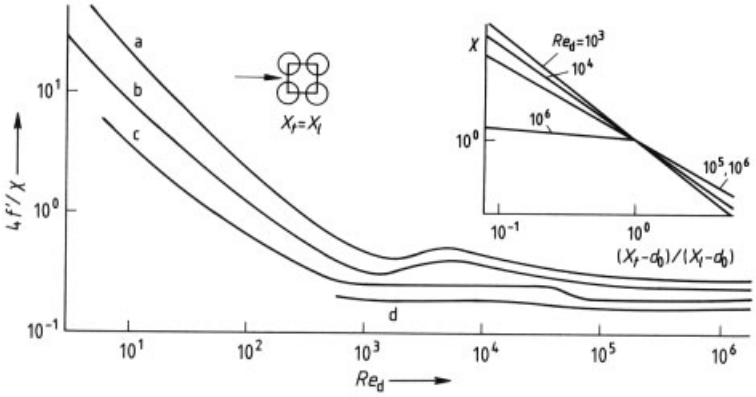

Figure 35. Friction factors for inline tube arrangement [44]
a) $X_l = 1.25$; b) $X_l = 1.5$; c) $X_l = 2.0$; d) $X_l = 2.5$

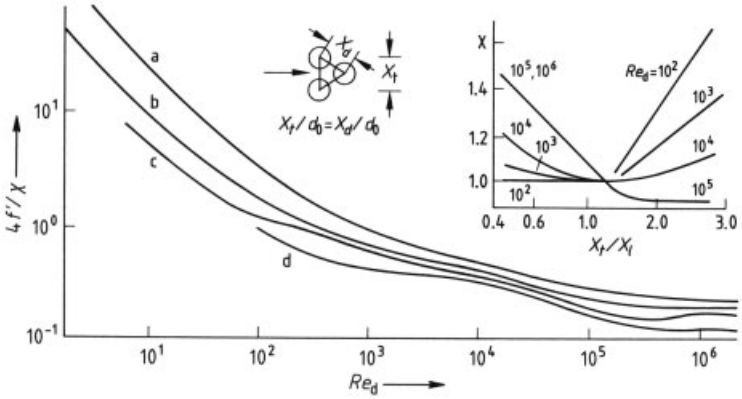

Figure 36. Friction factors for staggered tube arrangement [44]
a) $X_l = 1.25$; b) $X_l = 1.5$; c) $X_l = 2.0$; d) $X_l = 2.5$; $X_t^* = X_t/d_o$

$$f' = 9.465 Re_d^{-0.316} (X_t/d_o)^{-0.927} (X_t/X_d)^{0.515} \tag{50}$$

where $X_d = (X_t^2 + X_l^2)^{1/2}$ is the diagonal pitch. The correlation is valid for the following ranges: $2000 \le Re_d \le 50\,000$, $0.15 \le s'/l_f \le 0.19$, $3.75 \le s'/\delta_f \le 6.03$, $0.35 \le l_f/d_o \le 0.56$, $0.011 \le \delta_f/d_o \le 0.025$, $1.86 \le X_t/d_o \le 4.60$, $18.6 \le d_o \le 40.9$ mm, and $311 \le N_f \le 431$ fins/m.

Alternative correlations for (high) finned tubes are given in [44].

Extensive work has been reported on low-finned tubes and is assessed in [47–49]. A simple but accurate correlation for heat transfer is given in [48] as

$$j = 0.255 Re_d^{-0.3} (d_e/s')^{-0.3} \tag{51}$$

A more accurate correlation for heat transfer is given in [47]. Based on the comparison of existing data in literature and various correlations, Chai [49] has arrived at the best correlation for friction factors:

$$f' = 1.748 Re_d^{-0.233} \left(\frac{l_f}{s'}\right)^{0.552} \cdot \left(\frac{d_o}{X_t}\right)^{0.599} \left(\frac{d_o}{X_l}\right)^{0.1738} \tag{52}$$

This correlation is valid for $895 \le Re_d \le 713\,000$, $20 \le \theta \le 40$, $X_t/d_o < 4$, $N_r \ge 4$, and θ is the tube layout angle. It predicts 89 literature data points within a mean absolute error of 6 %; the range of actual error is from − 16.7 to + 19.9 %.

The design data for a number of augmented circular fins have been summarized in [50].

Flat Fins on a Tube Array. GRAY and WEBB [51] recommend the following heat-transfer correlation for plain flat fins on staggered tube arrays valid for four or more tube rows.

$$j_4 = 0.14 Re_{\mathrm{d}}^{-0.328} \left(\frac{X_{\mathrm{t}}}{X_{\mathrm{l}}}\right)^{-0.50} \left(\frac{s}{d_{\mathrm{o}}}\right)^{0.031} \tag{53}$$

For the number of rows N_{r} from 1 to 3, the j factor is lower and is given by

$$\frac{j_{N_{\mathrm{r}}}}{j_4} = 0.991 \left[2.24 Re_{\mathrm{d}}^{-0.092} \left(\frac{N_{\mathrm{r}}}{4}\right)^{-0.031}\right]^n \tag{54}$$

where $n = 0.607\,(4 - N_{\mathrm{r}})$.

Equations (53) and (54) correlated 89 % of the test data for 16 heat exchangers within $\pm 10\,\%$.

Pressure drop on the fin side in a tube and flat fin exchanger (staggered layout) consists of pressure drop associated with fins and that associated with tubes. Hence, GRAY and WEBB [51] defined the tube side effective friction factor f' as a sum of the individual friction factors associated with the fin and the tube, weighted with respect to surface areas

$$f' = f'_{\mathrm{f}} \frac{A_{\mathrm{f}}}{A} + f'_{\mathrm{t}} \left(1 - \frac{A_{\mathrm{f}}}{A}\right) \left(1 - \frac{\delta_{\mathrm{f}}}{s}\right) \tag{55}$$

Here f'_{f} is the friction factor associated with fins and f'_{t} is the friction factor associated with bare bank of tubes. While f'_{t} is obtained from the Zukauskas correlation (see Fig. 36), f'_{f} was correlated as

$$f'_{\mathrm{f}} = 0.508 Re_{\mathrm{d}}^{-0.521} \left(\frac{X_{\mathrm{t}}}{d_0}\right)^{1.318} \tag{56}$$

Equation (55) correlated 90 % of the data for 19 heat exchangers within $\pm 20\,\%$. The range of dimensionless variables of Equations (53) – (55) are: $500 < Re_{\mathrm{d}} < 24\,700$, $1.97 < X_{\mathrm{t}}/d_{\mathrm{o}} < 2.55$, $1.7 < X_{\mathrm{l}}/d_{\mathrm{o}} < 2.58$, and $0.08 < s'/d_{\mathrm{o}} < 0.64$.

Published data on other flat interrupted fins on a staggered tube array are presented in [39], [50].

6.3.2. Plate Exchangers

Most heat-transfer and flow friction characteristics (correlations) for various plate designs are proprietary; very little useful information has been published in the open literature. A comprehensive survey of correlations which are available in the open literature is given in [52].

Heat-Transfer Correlations. Forced convection heat-transfer coefficients for fully developed turbulent liquid flow in plate exchangers are usually correlated as

$$Nu = c_1 Re^{c_2} Pr^{c_3} \left(\frac{\mu_{\mathrm{m}}}{\mu_{\mathrm{w}}}\right)^n \tag{57}$$

where the empirical constants c_1 and c_2 are dependent upon the plate pattern and geometrical parameters. Typical values for transition and turbulent flow regimes are [53]:

$$c_1 = 0.15 - 0.40 \quad c_2 = 0.65 - 0.85$$
$$c_3 = 0.30 - 0.45 \quad n = 0.05 - 0.20 \tag{58}$$

Since liquids have $Pr > 1$ and are exclusively used in plate exchangers, the Nusselt number of Equation (57) is independent of thermal boundary conditions in turbulent flow.

In the laminar regime, Lévêque's correlation has been proposed for plate exchangers [54]

$$Nu = c_4 \left(z^*\right)^{-1/3} \left(\frac{\mu_{\mathrm{w}}}{\mu_{\mathrm{m}}}\right)^{-0.14}$$
$$= c_4 \left(\frac{L}{D_{\mathrm{e}} Re Pr}\right)^{-1/3} \left(\frac{\mu_{\mathrm{w}}}{\mu_{\mathrm{m}}}\right)^{-0.14} \tag{59}$$

where c_4 will depend both on the thermal boundary condition and the plate geometry, and is found to have values from 1.2 to 4.5. The dependence of the plate length L in the foregoing equation is not correct and should be replaced by a function of the corrugation pitch. JACKSON and TROUPE [55] proposed a exponent of 0.38 on the Reynolds number in Equation (59), and this has been confirmed in simulation studies.

Friction Factor Correlations. Based on the analogy to packed beds, the Fanning factor is presented in the form

$$f = c_5 + c_6 / Re \tag{60}$$

where the first and second terms on the right-hand side represent the contributions of turbulent and laminar effects respectively. This type

of equation could therefore be expected to correlate data in both the laminar and turbulent regimes.

For most plate designs, flow is turbulent and the following equation correlates pressure drop data better:

$$f = c_7 Re^{c_8} \tag{61}$$

where c_7 and c_8 are empirical constants. As an illustration of Equations (60) and (61), the correlations for the chevron plate (Fig. 6c) with $\beta = 30°$ are [56]:

$$f = \begin{cases} 0.093 + 57.5/Re & \text{for } 260 < Re < 3000 \\ 0.898 Re^{-0.26} & \text{for } 3000 < Re < 50000 \end{cases} \tag{62}$$

A very limited amount of other published correlations (design data) for plate exchangers are summarized in [3].

6.3.3. Extended Surface Exchangers

Compact exchanger surface basic data are usually presented in terms of j and f vs. Re. In these groups, the hydraulic diameter D_h is used, directly or indirectly, arbitrarily but consistently as a characteristic dimension. Although j and f data are applicable to any geometrically similar surface, as soon as the basic geometry is changed (such as an offset strip-fin geometry changed from 400 to 600 fins/m), the original j and f data are no longer applicable. This is because D_h is not a universal characteristic dimension. A great variety of compact surface geometries are available, and generally these different geometries will have different j and f data. Such surface basic data have been provided in [39] for over 60 plate-fin surface geometries: plain fins, wavy fins, offset strip fins, louver fins, perforated fins, and pin fins. Correlations for offset strip fin geometry are now presented as an illustration. Some additional surface basic data are provided in [27].

Offset Strip Fins. This is one of the most widely used enhanced fin geometries in compact heat exchangers. In addition to the fin spacing and fin height, the major variables are the fin thickness and fin strip length in the flow direction. The situation is complicated by burrs formed on leading and trailing fin edges during the manufacturing operation.

The following multiple regression power-law correlations of Joshi and Webb [58] are recommended for design use. They are based on the test data for 21 test cores with a wide variation in geometrical dimensions of the offset strip fin. Flow visualization measurements of Joshi and Webb show that the Reynolds number ($G D_h/\mu$) at which the wake departs from laminar flow is given by

$$Re^* = 257.2 \left(\frac{l_{ef}}{s'}\right)^{1.23} \left(\frac{\delta_f}{l_{ef}}\right)^{0.58} \cdot \left[\frac{\delta_f}{D_h} + 1.328 \left(\frac{l_{ef}}{D_h}\right)^{0.5} Re^{-0.5}\right]^{-1} \tag{63}$$

The j and f correlations for $Re < Re^*$ are given by

$$f_L = 8.12 Re^{-0.74} \left(\frac{l_{ef}}{D_h}\right)^{-0.41} \left(\frac{s}{b}\right)^{-0.02} \tag{64}$$

$$j_L = 0.53 Re^{-0.56} \left(\frac{l_{ef}}{D_h}\right)^{-0.15} \left(\frac{s}{b}\right)^{-0.14} \tag{65}$$

The correct exponent on Re is -0.56; the exponent in [58] is not correct.

The correlations for $Re > Re^* + 1000$ are given by

$$f_T = 1.12 Re^{-0.36} \left(\frac{l_{ef}}{D_h}\right)^{-0.65} \left(\frac{\delta_f}{D_h}\right)^{0.17} \tag{66}$$

$$j_T = 0.21 Re^{-0.40} \left(\frac{l_{ef}}{D_h}\right)^{-0.24} \left(\frac{\delta_f}{D_h}\right)^{0.02} \tag{67}$$

The j factor in the transition region between $Re^* < Re < Re^* + 1000$ is a log-linear interpolation between j_L (at Re^*) and j_T (at $Re^* + 1000$). A similar method is used for the friction factor. In the above equations, s is the fin pitch, l_{ef} the strip length, b the plate spacing, and δ_f the fin thickness.

The correlation predicted 82 % of the f data and 91 % of the j data within ± 15 %. The geometric parameters of the test arrays on which Equations (64) – (67) are based are $0.13 < s/b < 1.0$, $0.012 < \delta_f/l_{ef} < 0.048$, and $0.04 < \delta_f/s < 0.20$.

Published correlations on many plate – fin type surface geometries are presented in [57], [59].

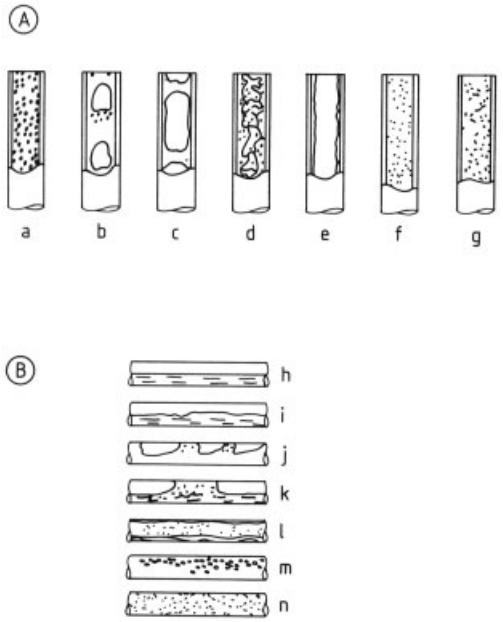

Figure 38. Two-phase flow patterns in (A) vertical tubes and (B) horizontal tubes [62]
a) Bubble flow; b) Plug; c) Slug flow; d) Froth or churn turbulent flow; e) Annular flow; f) Annular mist flow; g) Mist flow; h) Stratified smooth flow; i) Stratified wavy flow; j) Plug flow; k) Slug flow; l) Annular flow; m) Dispersed bubble flow; n) Mist flow

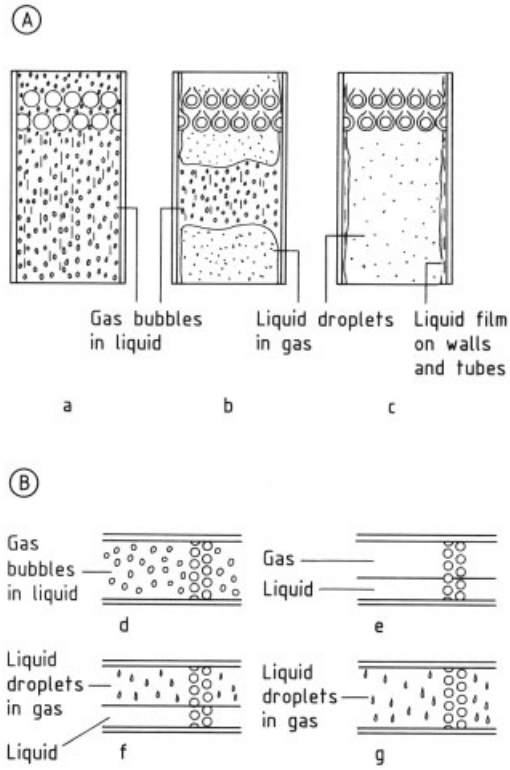

Figure 39. A) Vertical shellside two-phase upflow pattern; B) Horizontal shellside two-phase crossflow patterns [63]
a) Bubbly flow; b) Slug flow; c) Spray flow; d) Bubbly flow; e) Stratified flow; f) Stratified spray flow; g) Spray flow

Table 14. Property and pipe diameter correction for Weisman and Kang flow pattern maps [67]

Flow Orientation		φ_1	φ_2
Horizontal, vertical, and inclined flow	transition to dispersed flow	1.0	$\left(\frac{\varrho_l}{\varrho_{sl}}\right)^{-0.33}\left(\frac{d_i}{d_{i,s}}\right)^{0.16}\left(\frac{\mu_{sl}}{\mu_l}\right)^{0.09}\left(\frac{\sigma}{\sigma_s}\right)^{0.24}$
Horizontal, vertical, and inclined flow	transition to annular flow	$\left(\frac{\varrho_{sg}}{\varrho_g}\right)^{0.23}\left(\frac{\Delta\varrho}{\Delta\varrho_s}\right)^{0.11}\left(\frac{\sigma}{\sigma_s}\right)^{0.11}\left(\frac{d_i}{d_{i,s}}\right)^{0.415}$	1.0
Horizontal and slightly inclined flow	intermittent–separated transition	1.0	$\left(\frac{d_i}{d_{i,s}}\right)^{0.45}$
Horizontal flow	wavy–stratified transition	$\left(\frac{d_{i,s}}{d_i}\right)^{0.17}\left(\frac{\mu_g}{\mu_{gs}}\right)^{1.55}\left(\frac{\varrho_s}{\varrho}\right)^{2.55}\left(\frac{\Delta\varrho}{\Delta\varrho_s}\right)^{0.69}\left(\frac{\sigma_s}{\sigma}\right)^{0.69}$	1.0
Vertical and inclined flow	bubble–intermittent transition	$\left(\frac{d_{i,s}}{d_i}\right)^{n}(1-0.65\cos\alpha)$ $n = 0.26\exp[-0.17(V_{ls}/v_{ls}^s)]$	1.0

s denotes standard conditions: $d_{i,s} = 1.0$ in ($= 2.54$ cm); $\varrho_{gs} = 0.0013$ kg/1; $\varrho_{s,l} = 1.0$ kg/1; $\mu_{ls} = 1$ centipoisė ($= 1$ m Pa s); $\sigma_s = 70$ dynes/cm ($= 70 \cdot 10^{-5}$ N/cm); $v_{ls}^s = 1.0$ ft/s ($= 0.305$ m/s)

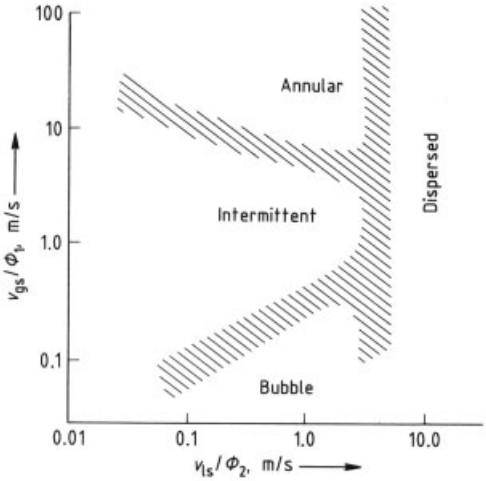

Figure 40. Overall flow pattern map for vertical and inclined tubes [67]

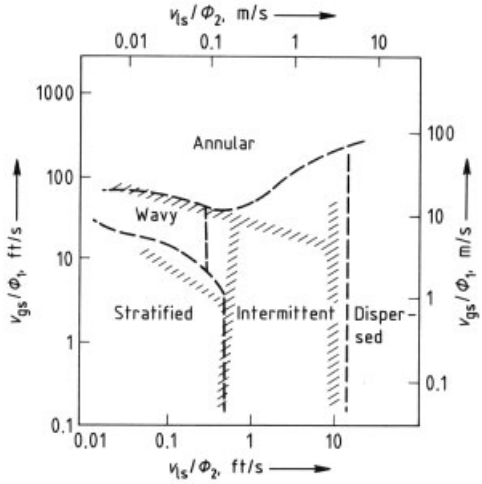

Figure 41. Overall flow pattern map for horizontal tubes based on air – water data for 2.5 cm (1 in) diameter tube, [66]

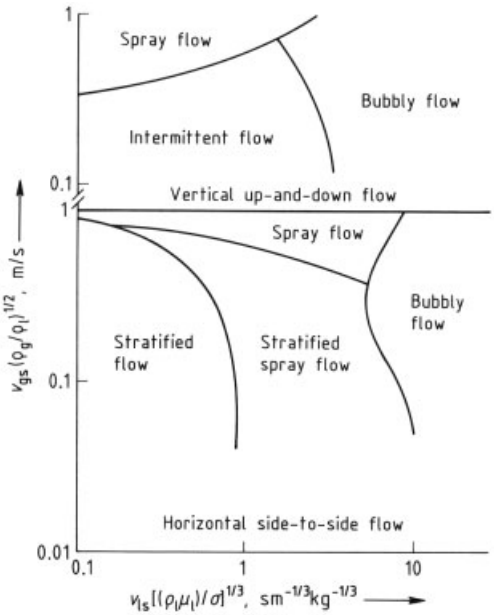

Figure 42. Shellside flow pattern maps [68]

7. Two-Phase Heat-Transfer and Pressure-Drop Correlations

Since two-phase flows are much more complex, it is essential to have the basic understanding of the following topics before conducting a heat exchanger design: flow patterns, two-phase pressure-drop correlations, condensation correlations, and pool boiling and forced convection boiling correlations, which are discussed in this chapter.

7.1. Flow Patterns

Flow regimes for a single-phase fluid flow are "laminar", "transition", and "turbulent". In a sharp-cornered pipe, all three regimes may exist simultaneously. In addition, secondary flows may exist such as in a curved pipe or in a straight pipe with a twisted tape. Flow separation and reattachment may also take place such as in flow over an interrupted heat-exchanger surface or over a tube bank. Knowledge of these characteristics is important for heat-transfer and pressure-drop correlations and analyses for single-phase exchangers.

These characteristics are also important for two-phase correlations. However, the geometry

or topology of two phases is even more important for pressure-drop and heat-transfer correlations and analyses. The liquid and gas or vapor flowing in or over a tube exhibit a variety of configurations known as *flow patterns.* A particular flow pattern is dependent upon the pipe geometry, relative magnitudes of gas and liquid flows, fluid properties, and operating conditions (flow rates, pressure, temperature, boundary conditions, heat flux, etc.). Many flow patterns have been identified [60–62]. Typical flow patterns in vertical (upflow) and horizontal pipes are shown in Figure 38 [62], and those for vertical (upflow) and horizontal shellside are shown in Figure 39 [63].

For flow regime-based correlations (which are more accurate) and understanding of the operating characteristics of a system, knowledge of transitions between flow patterns is very important. These transitions are presented in the form of a two-dimensional flow pattern map. Two axes of this map represent dimensional or dimensionless groups of flow parameters and fluid properties. Over 30 such maps have been presented in literature. The transition boundaries are represented either by lines or short regions. Usually, the transition is not abrupt; to indicate the uncertainty associated with the flow pattern changes, the region is more representative of the transition boundaries. TAITEL et al. developed semi-empirical models for transition between the flow patterns [64], [65]. WEISMAN et al. [66], [67] developed the flow pattern maps empirically, and these are in reasonably good agreement with other popular flow pattern maps. Their flow pattern maps are presented in Figures 40 and 41 with the functions φ_1 and φ_2 in Table 14. GRANT and MURRAY empirically developed the flow pattern maps for the shellside or flow normal to a tube bank (both vertical and horizontal flow) as shown in Figure 42 [68]. The flow pattern transition boundaries, shown by solid lines in this figure, indicate a mean of the transition regions. Note that v_{gs} in Figures 40, 41, and 42 and Table 14 represents the superficial gas velocity (the rate of volumetric flow of the gas phase in the two-phase mixture divided by the total flow area); similarly v_{ls} represents the superficial liquid velocity.

7.2. Two-Phase Pressure-Drop Correlations

Two-phase pressure drop in tubes consists of mainly three components: friction, momentum change (acceleration/deceleration), and gravity.

$$\Delta p = \Delta p_{\mathrm{f}} + \Delta p_{\mathrm{m}} + \Delta p_{\mathrm{hs}} \tag{68}$$

The entrance and exit pressure losses, usually considered in compact heat exchanger application, are neglected because (1) no two-phase data exist for these pressure losses, and (2) they represent a small fraction of the total pressure drop and are well within the uncertainty involved in the two-phase Δp prediction.

7.2.1. Intube Pressure Drop

Expressions and correlations for pressure-drop components of Equation (68) for two-phase flow in a circular tube are now summarized.

The pressure drop or rise caused by an elevation change (hydrostatic effect) is given by

$$\Delta p_{\mathrm{hs}} = \pm \frac{g}{g_{\mathrm{c}}} \sin\theta \int_0^L [\alpha \varrho_{\mathrm{g}} + (1-\alpha)\, \varrho_{\mathrm{l}}]\, dz \tag{69}$$

where the angle of inclination θ is measured from the horizontal, the "−" sign stands for a downflow (pressure rise) and the "+" sign for an upflow (pressure drop); α is the void fraction.

A void fraction correlation based on steam–water data obtained by THOM has been fitted to a curve by BUTTERWORTH as [69]

$$\alpha = \left[1 + \left(\frac{1-x}{x}\right)\left(\frac{\varrho_{\mathrm{g}}}{\varrho_{\mathrm{l}}}\right)^{0.89}\right]\left(\frac{\mu_{\mathrm{l}}}{\mu_{\mathrm{g}}}\right)^{0.18} \tag{70}$$

For the homogeneous model, α is given by

$$\alpha = \left[1 + \left(\frac{1-x}{x}\right)\left(\frac{\varrho_{\mathrm{g}}}{\varrho_{\mathrm{l}}}\right)\right]^{-1} \tag{71}$$

When Δp_{hs} is a dominant term in Equation (68) (i.e., 1/2 to 2/3 of the total Δp), such as for low velocity and low pressure applications, Equation (69) should be used for calculating Δp_{hs}. For other cases, a homogeneous model is fairly accurate; Equation (69) then reduces to

$$\Delta p_{\mathrm{g}} = \pm \frac{g}{g_{\mathrm{c}}} \sin\theta \int_0^L \varrho_{\mathrm{hom}}\, dz \tag{72}$$

where

$$\frac{1}{\varrho_{\mathrm{hom}}}=\frac{x}{\varrho_{\mathrm{g}}}+\frac{1-x}{\varrho_{\mathrm{l}}} \tag{73}$$

The momentum pressure drop or rise from cross section 1 to 2 is given by

$$\Delta p_{\mathrm{m}}=\frac{G^2}{g_{\mathrm{c}}}\left[\left(\frac{x^2}{\alpha\varrho_{\mathrm{g}}}+\frac{(1-x)^2}{(1-\alpha)\varrho_{\mathrm{l}}}\right)_2-\left(\frac{x^2}{\alpha\varrho_{\mathrm{g}}}+\frac{(1-x)^2}{(1-\alpha)\varrho_{\mathrm{l}}}\right)_1\right] \tag{74}$$

which for a homogeneous model reduces to

$$\Delta p_{\mathrm{m}}=\frac{G^2}{g_{\mathrm{c}}}\left(\frac{1}{\varrho_{\mathrm{hom},2}}-\frac{1}{\varrho_{\mathrm{hom},1}}\right) \tag{75}$$

The pressure drop due to friction is given by

$$\Delta p_f=\Delta p_{f,\mathrm{lo}}\varphi_{\mathrm{lo}}^2=f_{\mathrm{lo}}\frac{4L}{D_{\mathrm{h}}}\frac{G^2}{2g_{\mathrm{c}}\varrho_{\mathrm{l}}}\varphi_{\mathrm{lo}}^2 \tag{76}$$

where $\Delta p_{f,\mathrm{lo}}$ is the frictional pressure drop for liquid flowing alone, in the same tube, with a mass flow rate equal to the total mass flow rate of the two-phase flow. φ_{lo}^2 is the two-phase friction multiplier. Extensive comparisons and assessments of 25 000 data points (with and without heat transfer) have been made in [70–72] to arrive at specific models for φ_{lo}^2 (18 different available models for φ_{lo}^2 have been reported in the literature). f_{lo} is the single-phase friction factor for liquid (with total flow rate) flowing in the tube. f_{lo} is given by

$$f_{\mathrm{lo}}=0.079Re_{\mathrm{lo}}^{-0.25} \tag{77}$$

where $Re_{\mathrm{lo}}=GD_{\mathrm{h}}/\mu_{\mathrm{l}}$.

For high reduced pressures (p/p_{cr}), the homogeneous model yields accurate results, and is given by

$$\varphi_{\mathrm{lo}}^2=1+x\left(\frac{\varrho_{\mathrm{l}}}{\varrho_{\mathrm{g}}}-1\right) \tag{78}$$

Hewitt [73] recommends three different correlations for different ranges of viscosities and mass velocities as summarized next. Friedel [72] and Chisholm [74] developed correlations for many fluids for a wide range of variables. Friedel's correlation, more accurate for $\mu_{\mathrm{l}}/\mu_{\mathrm{g}}<1000$, is given by

$$\varphi_{\mathrm{lo}}^2=S_1+3.24S_2S_3Fr^{-0.045}We^{-0.035} \tag{79}$$

where

$$S_1=(1-x)^2+x^2\varrho_{\mathrm{l}}f_{\mathrm{go}}/\varrho_{\mathrm{g}}f_{\mathrm{lo}} \tag{80}$$

$$S_2=x^{0.78}(1-x)^{0.224} \tag{81}$$

$$S_3=\left(\frac{\varrho_{\mathrm{l}}}{\varrho_{\mathrm{g}}}\right)^{0.91}\left(\frac{\mu_{\mathrm{g}}}{\mu_{\mathrm{l}}}\right)^{0.19}\left(1-\frac{\mu_{\mathrm{g}}}{\mu_{\mathrm{l}}}\right)^{0.7} \tag{82}$$

$$Fr=\frac{G^2}{gd_{\mathrm{i}}\varrho_{\mathrm{hom}}^2} \tag{83}$$

$$We=\frac{G^2d_{\mathrm{i}}}{g_{\mathrm{c}}\sigma\varrho_{\mathrm{hom}}} \tag{84}$$

and ϱ_{hom} is given by Equation (73).

Chisholm's correlation, more accurate for $\mu_{\mathrm{l}}/\mu_{\mathrm{g}}>1000$ and $G>100\,\mathrm{kg/m^2\,s}$, is given by

$$\varphi_{\mathrm{lo}}^2=1+\left(\boldsymbol{\Gamma}^2-1\right)\cdot\left[Bx^{(2-n)/2}(1-x)^{(2-n)/2}+x^{2-n}\right] \tag{85}$$

where

$$\boldsymbol{\Gamma}^2=\frac{\Delta p_{\mathrm{f,go}}}{\Delta p_{\mathrm{f,lo}}} \tag{86}$$

$n=$ exponent in Blasius type equation
$B=$ coefficient as given in Table 15.

Table 15. Values of B in Equation (85) for smooth tubes [74]

Γ	G, kg/m^2 s	B
≤ 9.5	<500	4.8
	$500<G<1900$	$2400/G$
	≥ 1900	$55/G^{0.5}$
$9.5<\Gamma<28$	≤ 600	$520/(\Gamma G^{0.5})$
	≥ 600	$21/\Gamma$
≥ 28		$15\,000/(\Gamma^2 G^{0.5})$

Finally, the Lockhart–Martinelli correlation [75], more accurate for $\mu_{\mathrm{l}}/\mu_{\mathrm{g}}>1000$ and $G<100\,\mathrm{kg/m^2\,s}$, is given in terms of $\Delta p_{\mathrm{f,l}}$ or $\Delta p_{\mathrm{f,g}}$ by

$$\Delta p_{\mathrm{f}}=\Delta p_{\mathrm{f,l}}\varphi_{\mathrm{l}}^2=\Delta p_{\mathrm{f,g}}\varphi_{\mathrm{g}}^2 \tag{87}$$

where $\Delta p_{\mathrm{f,l}}$ is the frictional pressure drop for liquid flowing alone with actual liquid flow (and not total flow) rate in the total cross-sectional

area of the tube; φ_l^2 is the corresponding two-phase friction multiplier. $\Delta p_{f,g}$ and φ_g^2 are defined similarly for gas flowing alone. The correlations for φ_l^2 and φ_g^2 are given by [60]:

$$\varphi_l^2 = 1 + \frac{c}{X} + \frac{1}{X^2} \tag{88}$$

$$\varphi_g^2 = 1 + cX + X^2 \tag{89}$$

where

$$X^2 = \frac{\Delta p_{f,l}}{\Delta P_{f,g}} \tag{90}$$

and the value of c is dependent on the four possible single-phase regimes for liquid and gas streams in the tube.

$$c = \begin{cases} & \text{liquid} \quad \text{gas} \\ 20 & \text{turbulent} - \text{turbulent} \quad (\text{tt}) \\ 10 & \text{turbulent} - \text{viscous} \quad (\text{tv}) \\ 12 & \text{viscous} - \text{turbulent} \quad (\text{vt}) \\ 5 & \text{viscous} - \text{viscous} \quad (\text{vv}) \end{cases} \tag{91}$$

The foregoing correlations for predicting the adiabatic two-phase pressure drop are one of the best accurate correlations which have been reported in literature. Usually, the predicted two-phase pressure drop will have an uncertainty of ca. $\pm 20\,\%$ under ordinary operating conditions and with ordinary fluids; the error can be as high as $\pm 50\,\%$ or more for unusual operating conditions and fluids.

During the condensation process, since mass transfer takes place at the liquid – vapor interface, a correction needs to be made to the foregoing value of Δp_f in Equation (76). This correction is done on a local basis [76].

$$\frac{\Delta p_f}{\Delta z} = \frac{4}{d_i} \tau_i c_f \tag{92}$$

where

$$c_f = \frac{\gamma}{1 - \exp(-\gamma)} \text{ and } \gamma = \frac{W \Delta x}{\Delta A} \frac{v_g - v_i}{\tau_i} \tag{93}$$

Here τ_i is the interfacial shear stress, c_f is the correction factor for the frictional pressure gradient for the effect of mass transfer to the vapor – liquid interface, and Δx is the change in quality over surface area ΔA. τ_i is determined using the Martinelli parameter [76].

Equations (92) and (93) are solved iteratively to determine Δp_f for each increment Δz with an initial assumption that c_f is 1, and the sum of these Δp_f increments could result in the total Δp_f for the tube. At high condensing rates, the foregoing cumulative correction c_f to Δp_f can be as much as a factor of 2.

The frictional pressure drop for condensation or boiling must be calculated in a stepwise manner for small increments of Δz for which changes in quality, fluid properties, and other conditions are only moderate.

7.2.2. Shellside (Tube Bundle) Pressure Drop

A review on two-phase pressure drop for horizontal and vertical crossflow on the shellside is presented in [77]. Crossflow in a baffled horizontal heat exchanger with horizontal or vertical baffle cuts is referred to as "horizontal crossflow". Such flow was investigated by Grant and Chisholm [78] using an air – water system. Their correlation is given by Equation (85), with the values of the parameter B given in Table 16 [63], [78] for the crossflow zone. For the window zone, $n = 0$ for Equation (85) with $B = 0.25$ for vertical up-and-down flow, and $B = 2/(\Gamma + 1)$ for horizontal side-to-side flow.

For vertical crossflow (upflow) in a horizontal tube bundle, Schrage's correlation is recommended for φ_l^2 in Equation (87) as follows [79]:

$$\varphi_l^2 = 1 + \frac{c}{X_{tt}} + \frac{c_5}{X_{tt}^2} \tag{94a}$$

where

$$c = c_1 Fr_l^{c_2} \ln X_{tt} + c_3 Fr_l^{c_4} \tag{94b}$$

Here constants $c_1 - c_5$ are given in Table 17. For slug and spray flows, Equations (94 a and b), i.e., the constants $c_1 - c_5$ are valid for $Fr_l > 0.15$. For $Fr_l \leq 0.15$, the following correlation can be used [80]:

$$\varphi_l^2 = 1 + \frac{8}{X_{tt}} + \frac{1}{X_{tt}^2} \tag{95a}$$

where

$$X_{tt} = \left(\frac{1-x}{x}\right)^{0.9} \left(\frac{\varrho_g}{\varrho_l}\right)^{0.5} \left(\frac{\mu_l}{\mu_g}\right)^{0.1} \tag{95b}$$

Equation (95 a) is valid for $X_{tt} < 0.2$.

Table 16. Values of B and n of Equation (85) for the two-phase pressure drop for flow normal to a tube bank [63], [67]

Flow type	B	n
Vertical up-and-down spray and bubble	1	0.37
Horizontal side-to-side spray and bubble	0.75	0.46
Horizontal side-to-side stratified and stratified spray	0.25	0.46

Table 17. Coefficients for the two-phase friction multiplier correlation of Equation (94)

Flow pattern [a]	c_1	c_2	c_3	c_4	c_5
Bubbly	0.036	1.51	7.79	−0.057	0.774
Slug [b]	2.18	−0.643	11.6	0.233	1.09
Spray [b]	0.253	−1.50	12.4	0.207	0.205

[a] Flow pattern evaluated with Grant and Chisholm [68] flow pattern map
[b] If $Fr_1 \leq 0.15$, use Ishihara et al. [78] correlation

7.2.3. Other Geometries

Information on the two-phase pressure drop in compact heat exchanger surfaces is given in [11].

7.3. Heat-Transfer Correlations for Condensation

The process of converting vapor into liquid by removing heat from a system is defined as condensation. This occurs when the vapor is cooled below its saturation temperature causing nucleation of droplets. Such nucleation may occur homogeneously within the vapor (fog formation) or heterogeneously on the walls of the equipment. The latter may be categorized as dropwise or filmwise condensation. A number of different condensation types exist:

1) *Film Condensation.* The condensate forms a continuous liquid film on the cooled surface. This is the most common type of condensation in industrial equipment.
2) *Dropwise Condensation.* Here the condensate forms droplets on the cooled surface. High heat-transfer coefficients are obtained with dropwise condensation but it is difficult to maintain in industrial exchangers. Dropwise condensation occurs only with liquids having a high surface tension, such as water on specially treated surfaces.
3) *Condensation of Vapor Mixtures Forming Immiscible Liquids.* Here the condensate consists of a film of one component in which drops of the second component float and/or adhere to the cooled surface.
4) *Direct Contact Condensation.* This occurs when the vapor is brought directly into contact with a cold liquid.
5) *Homogeneous Condensation.* Here the vapor condenses in very small droplets forming fog. This occurs when a vapor and a noncondensable gas are cooled below the saturation temperature. While fogging does occur occasionally in industrial exchangers, it is not the desired method to condense the vapor.

In film condensation, the condensate wets the surface and forms a continuous liquid film. The behavior of this liquid film is governed by the normal fluid dynamic laws with gravity, vapor shear, and surface tension forces, either singly or together, acting to promote the removal of the condensate film. Two major condensation mechanisms are the gravity-controlled (vapor space) condensation and vapor shear-controlled (forced convective) condensation. Resistance to heat transfer during condensation is shown in Figure 43. For a pure saturated vapor, the resistance at the vapor – liquid interface is small, and may be neglected for a first approximation. It can, however, be significant for noncondensable gases or a multicomponent vapor mixture.

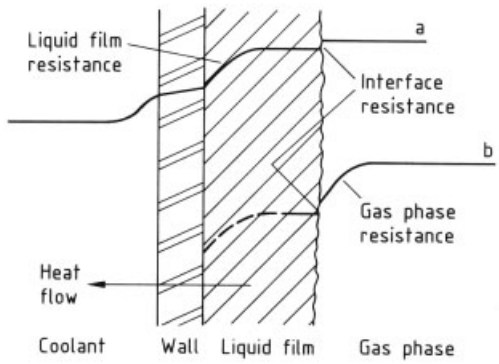

Figure 43. Heat-transfer resistances during condensation a) Temperature profile for pure saturated vapor; b) Temperature profile for superheated vapor or noncondensable gas

Heat must also be transferred through the condensate liquid film to the wall and then to

the coolant. In the case of a pure saturated vapor, this temperature drop across the liquid film represents the prime resistance to heat transfer. Techniques that reduce the condensate film thickness or promote a higher "effective" conductivity will therefore increase the heat-transfer coefficient on the condensing side. In the case of a vapor containing noncondensables or with a multicomponent vapor mixture, the gas phase resistance often dominates the overall condensing side coefficient; increase in the condensate film coefficient may not promote a comparable increase in overall condensation rate.

Condensation coefficients can vary significantly depending on the condensation rates, flow patterns, tube geometry, etc. During the condensation process, flow varies from being mostly vapor at the inlet to mostly liquid at the outlet in many processes, or from shear-dominated to gravity-dominated flows, and goes through a number of flow regimes. During this process, the condensing heat-transfer coefficient and temperature difference are evaluated as functions of the local quality and other parameters or variables of importance. Subsequent stepwise numerical or graphical integration would then yield the required surface area or the heat duty whichever is unknown. In only a few simple condensers would a single calculation suffice for a design.

At high-vapor shear, the condensate layer is turbulent and the heat-transfer coefficients are independent of the tube orientation. At low shear rates, gravity controls the effect of tube orientation. At intermediate shear rates where wavy, slug, or bubble-type flow patterns occur, the local heat-transfer coefficient can fluctuate widely. Currently, this intermediate region is calculated by prorating the high and low shear coefficients based on a parameter used to define these limits on two-phase flow maps. The most accurate condensation correlations available in the open literature are now summarized. A detailed study on the subject is given in [81–84].

7.3.1. Condensation Inside a Horizontal Tube

The two most common flow patterns during condensation within a horizontal tube are annular (vapor shear-dominated) flow and stratified (gravity-dominated) flow. Breber et al. [85] have provided the following criteria to identify these flow patterns:

$$\text{Annular flow}: j_g^* > 1.5 \tag{96a}$$

$$\text{Stratified flow}: j_g^* < 0.5 \tag{96b}$$

For intermediate values of j_g^*, a linear interpolation of the heat-transfer coefficient calculated for annular and stratified flows must be used. Here the Wallis parameter j_g^* is the dimensionless gas-phase mass velocity defined as

$$j_g^* = \frac{xG}{[g d_i \varrho_g (\varrho_l - \varrho_g)]^{1/2}} \tag{97}$$

Equation (96) is valid when the Martinelli parameter X_{tt} in Equation (95 b) is less than one.

For annular flows, the Nusselt-type laminar flow models predict heat-transfer coefficients that are too low; hence, turbulent flow models must be used to enable better prediction. The following empirical correlation based on 473 experimental data points for steam, refrigerants, and other organics, predicts condensation coefficients with a mean deviation of 15 % [86]:

$$Nu_z = Nu_l \left[(1-x)^{0.8} + \frac{3.8 x^{0.76} (1-x)^{0.04}}{p_r^{0.38}} \right] \tag{98}$$

where $p_r = p/p_{cr}$ is the reduced pressure, and Nu_l is the Nusselt number assuming that the total flow rate is the liquid flow rate in the tube, and is calculated from

$$Nu_l = 0.023 Re_{lo}^{0.8} Pr_l^{0.4} \tag{99}$$

Alternatively, there are a number of analytical correlations available for annular flow condensation inside a tube based on heat-transfer – momentum-transfer analogy, which are summarized in [81–84].

For stratified flows, the condensate forms a thin film on the top portion of the tube walls, then drains toward the bottom of the tube and forms a thick film. The stratified layer at the bottom flows axially due to the action of shear forces. The Nusselt equation is used for the upper condensate film portion, and the heat transfer from the condensate layer flowing along the bottom of the tube is neglected, correcting for the tube area covered by the stratified layer. In order to determine the peripheral mean heat-transfer coefficient based on the entire perimeter, the

following modified Nusselt equation applies for stratified flow [81], [83]:

$$Nu_z = \frac{h_z d_\mathrm{i}}{k_\mathrm{l}} = \Omega \left[\frac{\varrho_\mathrm{l} (\varrho_\mathrm{l} - \varrho_\mathrm{g}) g i'_\mathrm{lg} d_\mathrm{i}^3}{k_\mathrm{l} \mu_\mathrm{l} (T_\mathrm{sat} - T_\mathrm{w})} \right]^{1/4} \tag{100}$$

where the correction factor Ω is based on the void fraction:

$$\Omega = 0.728 \left\{ 1 + \left(\frac{1-x}{x} \right) \left(\frac{\varrho_\mathrm{g}}{\varrho_\mathrm{l}} \right)^{2/3} \right\}^{-3/4} \tag{101}$$

and

$$i'_\mathrm{lg} = i_\mathrm{lg} + \frac{3}{8} c_\mathrm{p,l} (T_\mathrm{sat} - T_\mathrm{w}) \tag{102}$$

7.3.2. Condensation Inside a Vertical Tube

The condensation-coefficient correlations for vertical tubes depend on whether the vapor flow is downward or upward.

For downward vapor shear-controlled flow (see criteria of Eq. (96 a)), an annular flow pattern exists, and the correlation of Equation (98) can be used. For downward gravity-controlled flow, the following correlations for the local heat-transfer coefficient are recommended which are dependant on the film (laminar, wavy, or turbulent) Reynolds number $Re_{\Gamma,z}$ [81], [82]:

$$\frac{h_z}{k_\mathrm{l}} \left[\frac{\mu_\mathrm{l}^2}{\varrho_\mathrm{l} (\varrho_\mathrm{l} - \varrho_\mathrm{g}) g} \right]^{1/3} = \begin{cases} 1.1 Re_{\Gamma,z}^{-1/3} & \text{for } Re_{\Gamma,z} \leq 33 \\ 0.76 Re_{\Gamma,z}^{-0.22} & \text{for } 33 < Re_{\Gamma,z} < 1600 \\ 0.023 Re_{\Gamma,z}^{1/4} Pr_\mathrm{l}^n & \text{for } Re_{\Gamma,z} \geq 1600 \end{cases} \tag{103}$$

where

$$Re_{\Gamma,z} = 4\Gamma_\mathrm{z}/\mu_\mathrm{l} \tag{104}$$

$$\Gamma_\mathrm{z} = \frac{\varrho_\mathrm{l} (\varrho_\mathrm{l} - \varrho_\mathrm{g}) g \delta^3}{3\mu_\mathrm{l}};$$

$$\delta = \left[\frac{4\mu_\mathrm{l} k_\mathrm{l} z (T_\mathrm{sat} - T_\mathrm{w})}{\varrho_\mathrm{l} (\varrho_\mathrm{l} - \varrho_\mathrm{g}) g i'_\mathrm{lg}} \right]^{1/4} \tag{105}$$

Although the Pr_l exponent $n = 0.5$ is recommended [81], [82], it appears to be high for the turbulent film, and a value of 0.33 may be a better choice [81]. Expressions for the mean heat-transfer coefficients (counterpart of Eq. (103)) are given in [82].

For upward vapor flows, the interfacial shear will retard the condensate drainage, the film will thicken, and if the vapor velocity is high, it will cause flooding (preventing downflow of condensate). Based on the air – water system, WALLIS [87] has provided the following criterion for the onset of flooding:

$$(v_\mathrm{g}^*)^{1/2} + (v_\mathrm{l}^*)^{1/2} = c \tag{106}$$

where

$$v_\mathrm{g}^* = \frac{v \varrho_\mathrm{g}^{1/2} \mathrm{g}}{[g d_\mathrm{i} (\varrho_\mathrm{l} - \varrho_\mathrm{g})]^{1/2}} \tag{107}$$

$$v_\mathrm{l}^* = \frac{v_\mathrm{l} \varrho_\mathrm{l}^{1/2}}{[g d_\mathrm{i} (\varrho_\mathrm{l} - \varrho_\mathrm{g})]^{1/2}} \tag{108}$$

The gas and liquid phase velocities v_g and v_l should be calculated at the bottom of the tube where they are at a maximum. While WALLIS assumes the value of c as 0.725, BUTTERWORTH suggests a conservative value of 0.6 and a more appropriate value as follows [88]:

$$c^2 = 0.53 F_1 F_2 \tag{109}$$

$$F_1 = (\sigma_\mathrm{l}/\sigma_\mathrm{w})^{0.5} \tag{110}$$

where σ_w is the surface tension of water at room temperature. F_2 is unity if the tube is cut at a right angle to the tube axis; otherwise, it is greater than unity.

The condensing heat-transfer coefficient for upward vapor flow is reduced by a factor of 0.7 to 1.0 compared to that calculated from the Nusselt theory depending upon the vapor velocity. It is given by

$$\frac{h_\mathrm{m}}{k_\mathrm{l}} \left[\frac{\mu_\mathrm{l}^2}{\varrho_\mathrm{l} (\varrho_\mathrm{l} - \varrho_\mathrm{g}) g} \right]^{1/3} = 1.47 Re_{\Gamma,\mathrm{L}}^{-1/3} c \tag{111}$$

where

$$Re_{\Gamma,\mathrm{L}} = 4\Gamma_\mathrm{L}/\mu_\mathrm{l};\ \Gamma_\mathrm{L} = q''_\mathrm{ave} L / i'_\mathrm{lg} \tag{112}$$

where the value of c ranges from 0.7 to 1.0 [81], and $L = \pi d_\mathrm{i}$.

7.3.3. Condensation Outside Horizontal Tube Bundles

The mean transfer coefficient for gravity-controlled laminar film condensation over a single horizontal tube is given by Nusselt as

$$Nu_{\mathrm{m}}=\frac{h_{\mathrm{m}}d_0}{k_{\mathrm{l}}}=0.725\left[\frac{d_{\mathrm{o}}^3\varrho_{\mathrm{l}}\left(\varrho_{\mathrm{l}}-\varrho_{\mathrm{g}}\right)g\,i'_{\mathrm{lg}}}{k_{\mathrm{l}}\mu_{\mathrm{l}}\left(T_{\mathrm{sat}}-T_{\mathrm{w}}\right)}\right]^{1/4} \tag{113}$$

For crossflow over a tube bundle, dripping of condensate from row to row and the vapor velocity influence the heat-transfer coefficient for subsequent rows. Empirically, the mean heat-transfer coefficient for N rows is given by

$$h_{\mathrm{m}}=h_{\mathrm{m},1}N_{\mathrm{r}}^{-1/6} \tag{114}$$

where $h_{\mathrm{m},1}$ for the first tube row is given by Equation (113).

For a circular tube bundle (such as in X shells), the effective number of tubes in a vertical path is given empirically by

$$N_{\mathrm{r}}=0.78d_{\mathrm{B}}/X_{\mathrm{l}} \tag{115}$$

where d_{B} is the bundle diameter and X_{l} is the longitudinal vertical tube pitch.

At moderate vapor shear, the following equations are recommended:

$$Nu=c\left(\chi^4Re^2+Nu_{\mathrm{m}}^4\right)^{1/4} \tag{116}$$

where

$$c=\begin{cases}1.0 & \text{for staggered tubes}\\ 0.8 & \text{for inline tubes}\end{cases} \tag{117}$$

$$\chi=0.9\left[1+\frac{1}{S_4S_5}\right]^{1/3} \tag{118}$$

$$S_4=\left[\frac{\varrho_{\mathrm{l}}\mu_{\mathrm{l}}}{\varrho_{\mathrm{g}}\mu_{\mathrm{g}}}\right]^{1/2} \tag{119}$$

$$S_5=\frac{c_{\mathrm{p,l}}\left(T_{\mathrm{sat}}-T_{\mathrm{w}}\right)}{Pr_{\mathrm{l}}i_{\mathrm{lg}}} \tag{120}$$

$$Re=\frac{W\,d_0\,\varrho_{\mathrm{l}}}{A_{\mathrm{f_1}}\,\mu_{\mathrm{l}}\varrho_{\mathrm{g}}} \tag{121}$$

Nu_{m} in Equation (116) is defined in Equation (113). The first term in the parenthesis in Equation (116) represents the velocity effect and the second term is the conventional Nusselt condensation equation. Equation (116) is used for either horizontal or vertical downflow.

TABOREK [89] suggested a simpler and perhaps equally good method. For an ideal bundle,

$$Nu=0.3Re^{0.6}Pr^{0.4}\left(1+\varrho_{\mathrm{l}}/\varrho_{\mathrm{g}}\right)^{1/2} \tag{122}$$

In a baffled exchanger, the effect of leakage streams on the ideal flow through the bundle must be allowed for.

At very high vapor shear the coefficients appear to decrease. This has been attributed to (1) hydrodynamic trapping of condensate between tubes, which then reduces the condensing area, or (2) local pressure reductions and, hence, lower temperature difference. No criteria are available to enable prediction of this phenomenon.

7.3.4. Condensation over Finned Tubes

Finned tubes are widely used in refrigeration and air-conditioning systems. The short fins, ca. 1.6 mm high and with 630 – 1050 fins/m, are extruded from the tube walls, resulting in an approximately three- to fourfold increase in external area. BEATTY and KATZ [90] correlated data by assuming that the fins act as vertical surfaces and the tube between the fins as a horizontal tube. They derived an equivalent diameter expression which, when substituted in a theoretical equation, (or empirical correlation, Eq. (98)) predicted the coefficients for a single tube in quiescent vapor. This equivalent diameter D_{e} is

$$D_{\mathrm{e}}=\left[\frac{1.3A_{\mathrm{f}}\eta_{\mathrm{f}}}{A_{\mathrm{ef}}\,l^{1/4}}+\frac{A_{\mathrm{bt}}}{A_{\mathrm{ef}}\,d_{\mathrm{o}}^{1/4}}\right]^{-1/4} \tag{123}$$

where

$A_{\mathrm{f}}=$ actual fin area
$A_{\mathrm{bt}}=$ horizontal bare tube (base) area
$\eta_{\mathrm{f}}=$ fin efficiency
$A_{\mathrm{ef}}=A_{\mathrm{f}}\,\eta_{\mathrm{f}}+A_{\mathrm{bt}}$
$l=A_{\mathrm{f}}/2\,d_{\mathrm{e}}$

The fin efficiency η_{f} is essentially 1.0 for copper tubes but it needs to be taken into account when stainless steel tubes are used (since $\eta_{\mathrm{f}}<1.0$). The heat transfer of low finned tubes is affected by the surface tension of condensate. Liquids with high surface tension, such as water, flood the spaces between the fins and result in a substantial reduction of heat transfer.

The lower portion of the tubes tend to be flooded with a condensate holdup, and poor heat transfer would be expected. However, it appears that for the upper portion, the effect of surface tension is to thin the condensate layer on the fins and to improve the heat transfer; the net result seems to be a cancellation of these effects and Equation (123) still holds.

In process exchangers, there will be an effect of tube rows and vapor shear. Much of the information for these condensers is proprietary but it seems that using Equation (123) for a single tube coefficient and then applying the row and shear corrections given above for the plain tubes should produce reasonable coefficients.

Condensation inside or outside of other enhanced tubes and surfaces is covered in [11], [83].

7.3.5. Condensation Outside Vertical Tube Bundles

For parallel flow outside vertical tubes, the condensation heat-transfer coefficient is computed using the correlations presented earlier for condensation inside a circular tube with alteration: the circular tube diameter is replaced by the hydraulic diameter of the tube bundle for parallel flow outside the tube bundle.

7.3.6. Condensation of Mixtures

While very high heat-transfer coefficient can be achieved during the condensation of a pure vapor, the presence of a small amount of noncondensable gas or a more volatile vapor can introduce a significant thermal resistance to the heat flow at the interface, and reduce the heat-transfer rate significantly. A schematic of the condensation of a vapor in the presence of a noncondensable gas is shown in Figure 43.

Four main approaches have been used in the development of design procedures for various mixtures of condensable vapors and noncondensable gases. These are approximate techniques, film models, penetration models, and boundary layer models. Specific methods and design procedures for condensers with mixtures are summarized in Section 9.1.3 and further details are given in [91].

7.4. Heat-Transfer Correlations for Boiling and Evaporation

Forced convective vaporization (boiling and evaporation) in confined channels (circular tubes, annuli, ducts of other cross sections, and flow over tube bundles) is a very important industrial problem. The fluid can be a liquid and its vapor, or a mixture of fluids in liquid and/or vapor phases. Because of the heat addition along the flow length, the liquid will go through a series of flow regimes, such as shown in Figure 44 for a vertical tube. The most important mechanisms are nucleate boiling (subcooled and saturated), evaporation, and film boiling. In *nucleate boiling,* wall conditions are such that the vapor bubbles start in the liquid at the nucleation sites on the wall of the heated tube. Usually, the wall superheat (i.e., the wall temperature minus the saturation temperature of the liquid) is large, the liquid film thickness is also large, and vapor mass is low in nucleate boiling. Depending on whether the liquid is subcooled or saturated, it is referred to as subcooled or saturated nucleate boiling. Bubble, slug, and plug flow patterns exist during nucleate boiling. When the liquid film near the wall is thin, and the wall superheat is relatively low, no nucleation occurs at the wall. Heat is carried away from the wall by forced convection in the film to the liquid film – vapor core interface where evaporation occurs. Since the nucleation is completely suppressed, the phenomenon can no longer be called "boiling". The *evaporation* region is also referred to as the two-phase forced convective region. The vapor volume is usually high in this region with the quality ranging from 0.2 to 0.6. Annular and mist annular flow exist during evaporation. This is the most important region for the design of evaporators. If the temperature difference exceeds the critical temperature difference, a continuous vapor film then blankets the tube wall and results in a significant decrease in the heat-transfer rate. This phenomenon is referred to as *film boiling.* For nuclear reactor fuel elements, radiantly heated furnaces, and other applications for which the heat flux is an independent variable, there will be a sharp increase in the wall temperature as the critical heat flux value is reached. This could burn out the tubes. Film boiling should be avoided in the design of heat-transfer equipment.

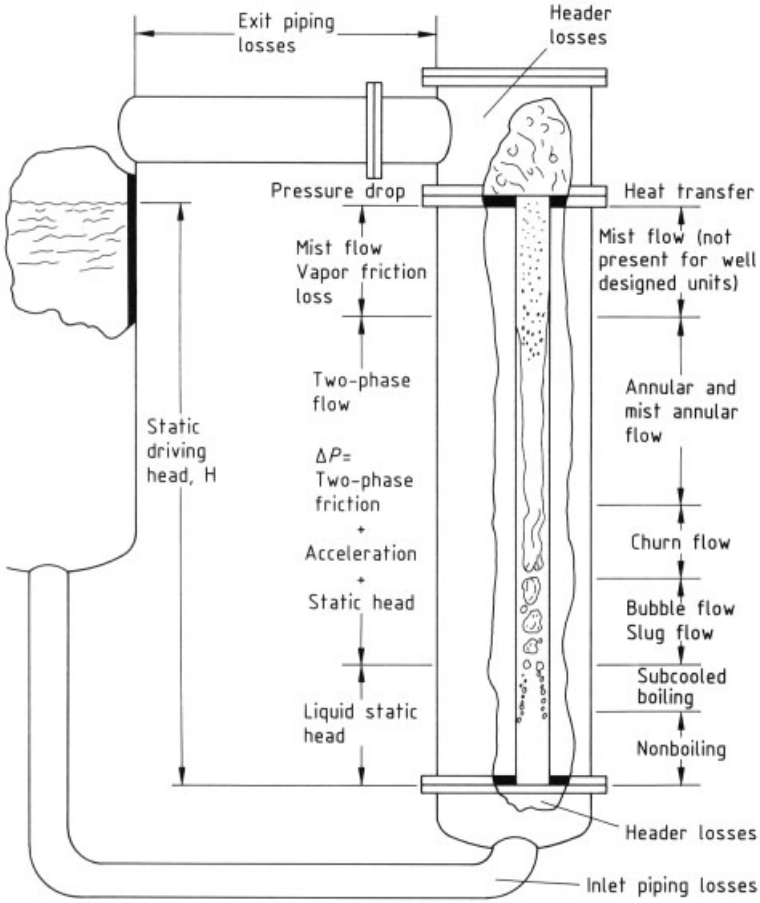

Figure 44. Heat-transfer regions in convective boiling

Saturated flow boiling is referred to as the region of two-phase flow where the liquid is at its saturation temperature. This would include both saturated nucleate boiling and evaporation mechanisms and will encompass bubble, slug, plug, churn, annular, and annular mist flow regimes. Specific design correlations are summarized in this chapter for saturated flow boiling and critical heat flux for vaporization in tubes and over tube bundles. Correlations for other flow regimes and further details are given in [92], [93].

7.4.1. Intube Forced Convective Boiling

The vaporization of a liquid flowing inside a tube is a complex process. Typical flow patterns for vaporization in vertical tubes are shown in Figure 44. Similarly, the flow patterns in horizontal tubes are given in [60] among others. Of the various flow regimes shown, the most important regime is the annular which exists for 80 – 90 % of the tube length.

Saturated flow boiling refers to both nucleate boiling (dominant at low qualities) and evaporation (dominant at high qualities) processes. In 1966, Chen [94] was the first to correlate the heat-transfer coefficients for saturated flow boiling in a vertical tube. He conjectured that the microconvection of nucleate boiling and macroconvection of single-phase heat transfer are additive, and that the contributions of one mechanism over the other is dependent on the two-phase Reynolds number. During the last 21 years, many attempts have been made to refine

Table 18. Constants in the saturated flow boiling correlation of Equation (124)

	$Co < 0.65$ Convective region	$Co > 0.65$ Nucleate boiling region
c_1	1.1360	0.6683
c_2	− 0.9	− 0.2
c_3	667.2	1058.0
c_4	0.7	0.7
c_5	0.3	0.3
$c_5 = 0$ for vertical tubes and for horizontal tubes with $Fr_l > 0.04$		

Table 19. Fluid dependent parameter F_{fl} of Equation (124)

Fluid	F_{fl}
Water	1.00
R-11	1.30
R-12	1.50
R-13B1	1.31
R-22	2.20
R-113	1.10
R-114	1.24
R-152 a	1.10
Nitrogen	4.70
Neon	3.50

this correlation which is valid for many different fluids [61], [95]. One of the most accurate correlations, based on 5246 data points of ten fluids, is given in [95]. The local two-phase heat-transfer coefficient is correlated as

$$\frac{h_{tp}}{h_l} = c_1 Co^{c_2} \left(25 Fr_l\right)^{c_5} + c_3 Bo^{c_4} F_{fl} \tag{124}$$

where

$$h_{tp} = q'' / \left(T_w - T_{sat}\right) \tag{125}$$

$$h_l = 0.023 Re_l^{0.8} Pr_l^{0.4} \left(k_l / d_i\right) \tag{126}$$

$$Co = \left(\frac{1-x}{x}\right)^{0.8} \left(\frac{\varrho_g}{\varrho_l}\right)^{0.5} \tag{127}$$

Co = Convection number

$$Fr_l = \frac{G^2}{g d_i \varrho_l^2} \tag{128}$$

Fr_l = Froude number with all flow as liquid

$$Bo = q'' / \left(G i_{lg}\right) \tag{128a}$$

Bo = Boiling number

Constants $c_1 - c_5$ are given in Table 18, the fluid-dependent parameter F_{fl} is given in Table 19, and Bo is defined in Equation (130). For fluids other than those listed in Table 19, F_{fl} could be estimated by the method given in [96].

The correlation in Equation (124) is applicable to vertical upflow and downflow, and horizontal flow. It was used to correlate water data with a mean deviation of ± 16 % and all refrigerant data with a mean deviation of ± 19 %.

7.4.2. Intube Critical Heat Flux

The prediction of critical heat flux (CHF) is of great practical importance in design and analysis of heat-transfer equipment. Either the heat-transfer coefficient and rate could decrease sharply and/or the tube wall temperature could increase sharply with a possible catastrophic failure. Many of the investigations have centered on the critical heat flux for the uniformly heated vertical tubes. Many literature reviews have been published on the subject to scrutinize an estimated 200 – 400 correlations [61], [97]. The most accurate and versatile correlation is by Shah [97] which is based on the test data of 23 fluids from 62 sources. It is derived for vertical upflow with uniform heat flux condition. It would also be applicable to vertical downflow and horizontal flow as long as a high flow rate is maintained. This correlation predicts q_{cr} for all test data with a mean deviation of ± 16 %. This correlation consists of two subcorrelations: the first one depends upon the upstream conditions (inlet subcooling and distance from inlet); the second depends upon the local quality, except for very short tubes. The correlations are presented first followed by the procedure to use them.

The upstream condition correlation is given by

$$Bo = 0.124 \left(\frac{d_i}{L_{ef}}\right)^{0.89} \left(\frac{10^4}{Y}\right)^n \left(1 - x_{i,ef}\right) \tag{129}$$

where

$$Bo = q''_{cr} / \left(G i_{lg}\right) \tag{130}$$

$$Y = Pe_l Fr_l^{0.4} \left(\mu_l/\mu_g\right)^{0.6}, \; Pe_l = G d_i c_{p,l}/k_l \tag{131}$$

$$n = \begin{cases} 0 & \text{for all fluids with } Y \le 10^4 \\ (d_i/L_{ef})^{0.33} & \text{for helium with } Y > 10^4 \\ (d_i/L_{ef})^{0.54} & \text{for all other fluids with } 10^4 < Y \le 10^6 \\ 0.12/(1 - x_{i,ef})^{0.5} & \text{for all other fluids with } Y > 10^6 \end{cases} \tag{132}$$

$$L_{ef} = \begin{cases} L_{cr} & \text{for } x_i \le 0 \\ L_b & \text{for } x_i > 0 \end{cases} \tag{133}$$

$$x_{i,ef} = \begin{cases} x_i & \text{for } x_i \le 0 \\ 0 & \text{for } x_i > 0 \end{cases} \tag{134}$$

Here L_{ef} is the effective length of the tube and $x_{i,ef}$ is the effective quality. L_{cr} is the axial distance from the tube inlet to the location x_{cr} of the critical heat flux. L_b is the boiling length, or the axial distance between $x = 0$ and x_{cr}, and is given by the following expression for uniformly heated tubes:

$$\frac{L_b}{d_i} = \frac{x_{cr}}{4Bo} = \frac{L_{cr}}{d_i} + \frac{x_i}{4Bo} \tag{135}$$

The local condition correlation is given by

$$Bo = F_e F_x Bo_o \tag{136}$$

where

$$F_e = \max\left[1, 1.54 - 0.032\left(L_{cr}/d_i\right)\right] \tag{137}$$

$$F_x = \begin{cases} F_3 \left[1 + \frac{\left(F_3^{-0.29} - 1\right)(p_r - 0.6)}{0.35}\right]^c & \text{for } x_{cr} > 0 \\ F_1 \left[1 - \frac{(1 - F_2)(p_r - 0.6)}{0.35}\right]^c & \text{for } x_{cr} < 0 \end{cases} \tag{138}$$

$$c = \begin{cases} 0 & \text{for } p_r \le 0.6 \\ 1 & \text{for } p_r > 0.6 \end{cases} \tag{139}$$

$$F_3 = \left(\frac{1.25 \times 10^5}{Y}\right)^{0.833 x_{cr}} \tag{140}$$

$$F_1 = \begin{cases} 1 + 0.0052\left(-x_{cr}^{0.88}\right) Y^{0.41} & \text{for } Y \le 1.4 \times 10^7 \\ 1 + 4.425\left(-x_{cr}^{0.88}\right) & \text{for } Y > 1.4 \times 10^7 \end{cases} \tag{141}$$

$$F_2 = \begin{cases} F_1^{-0.42} & \text{for } F_1 \le 4 \\ 0.55 & \text{for } F_1 > 4 \end{cases} \tag{142}$$

$$Bo_o = \max\left[15 Y^{-0.612}, \; 0.082 Y^{-0.3}\left(1 + 1.45 p_r^{4.03}\right), \; 0.0024 Y^{-0.105}\left(1 + 1.15 p_r^{3.39}\right)\right] \tag{143}$$

Here the function max means the maximum value of 2 or 3 quantities within the brackets.

The choice between Equation (129) and (136) is made as follows:

Always use equation (129) for helium
For $Y \le \pi\, 10^6$, use Equation (129) for all other fluids
For $Y > \pi\, 10^6$, and $L_{ef} > 160/p^{1.14}$, use Equation (129) for all other fluids
For $Y > \pi\, 10^6$ and $L_{ef} < 160/p_r^{1.14}$, determine *Bo* from Equations (129) and (136), and use the lower value.

7.4.3. Shellside Forced Convective Boiling

Heat-transfer coefficients for forced convective boiling over a tube bundle are significantly higher than those for a single tube due to circulation and turbulence caused by the rising vapor bubbles. Nucleation and convection govern the heat-transfer behavior of boiling in a tube bundle. Convection effects are important at low heat fluxes and low pressures, and nucleation effects dominate at high heat fluxes and high pressures. Increasing quality adds to the convective effects only at low heat fluxes and mass velocities. The heat-transfer coefficient increases with increasing mass velocity, but is insensitive to the quality. It is dependent upon the clearance between the tubes, but not so dependent upon the tube layout, staggered or inline. Within limits, the heat-transfer coefficient increases with increasing height in the bundle [77]. In large bundles, the upper tubes may be starved of liquid.

Because of the trade-off between nucleation and convective effects, many attempts have been made in literature to devise a Chen-type correlation. These efforts have been summarized in [77], [79], [98]. They conclude that the Chen-type correlation is not quite accurate for the tube

bundle since the whole tube surface is not swept by the flow, and the form drag effects over tubes are present which were not considered with the two-phase friction multipier in the original Chen formulation. JENSEN [77] recommends to use the Chen-type correlation until a better technique is developed with an accuracy of the order of $\pm 40\,\%$; one such correlation is given in [79], [98]. Another correlation is given later by Equation (160).

7.4.4. Shellside Critical Heat Flux

Critical heat flux for a tube bundle immersed in a pool is considerably lower than that for a single tube in pool boiling. The pool boiling curve for a tube bundle has a flat plateau at the maximum heat flux in contrast to a peak for the single tube. The critical heat flux condition for the tube bundle is defined as the condition when the integrated effect of all tubes begins to adversely affect the tube bundle performance. Different tubes in a bundle reach the critical heat flux condition at different times. Increasing the temperature driving force yields higher heat-transfer coefficients on those tubes that have not yet reached the CHF condition, and are counterbalanced by lower heat-transfer coefficients on the CHF-affected tubes [77]. Only a very limited amount of data are available in literature for the tube bundle CHF. Hence, no generalized accurate correlation can be recommended. Palen and Small's [99] correlation for the CHF is often quoted and is conservative; it is given by Equation (166).

8. Thermal Design for Single-Phase Heat Exchangers

Two of the most common heat-exchanger design problems are the rating and sizing.

The *rating problem,* also referred to as the performance problem, is analyzed for predicting the performance of an exchanger which has already been built or designed. The objective here is either to verify the manufacturer's specifications or to determine the performance at off-design conditions. The following quantities need to be specified: the exchanger construction type, flow arrangement, overall core dimensions, complete details on the material and surface geometries on both sides including their heat-transfer and pressure-drop characteristics, fluid flow rates, inlet temperatures, and fouling factors. The designer's task is to predict the fluid outlet temperatures, heat-transfer rate, and pressure drop on each fluid side.

The *sizing problem*, also referred to as a design problem, involves the design of a new heat exchanger for a specified performance within known constraints and maximum/minimum objective function. In this problem, the fluid inlet and outlet temperatures and flow rates are generally specified as well as the pressure drop on each fluid side. The designer's task is to select the appropriate construction type, flow arrangement, materials and surfaces on each side, and to determine the exchanger dimensions to meet the required heat transfer and pressure drop.

Computers are widely used to perform the exchanger rating and sizing calculations. Many proprietary programs of varying accuracy and usefulness are available. Major computer programs for shell-and-tube exchangers are summarized in [100]. The better programs are those supported by research organizations specializing in the field. The output of any computer program needs to be carefully analyzed especially if used in the design mode, since its internal logic affects the final design.

Very detailed step-by-step solution procedures are presented for plate – fin and tube – fin single-phase exchangers in [101], for plate exchangers in [3], for regenerators in [102], and heat pipe heat exchangers in [6]; hence, they are not covered here. In shell-and-tube exchangers, various leakage and bypass streams would affect both the heat-transfer coefficient and pressure drop on the shell side.

8.1. Introduction

Tubeside flow is through straight or U tubes. Hence, the correlations given in Chapter 7 on heat-transfer coefficients and friction factors for circular tubes are applicable. For U tubes, pressure drop associated with a 180° bend should be included in the analysis. For multipass exchangers, allowance must be made for the turnaround losses in the front and/or rear end heads.

In contrast, the shellside fluid flow is very complicated. It is a function of the type of baffling used, i.e., segmental, disk-and-doughnut, rod, etc., which will determine the method of design and the specific heat-transfer and pressure-drop correlations to be used. Segmental baffles are the type most widely used, especially in the United States, and have been extensively researched. Disk-and-doughnut baffles are reported to be frequently used in Europe but the design correlations are based on limited data. The rod baffling is a patented design and other similar designs are also available, but design data are again limited with respect to quantity and sources.

8.2. Shell-and-Tube Exchangers

8.2.1. Segmental Baffles

The segmental baffle is formed from a circular disk by cutting off a segment which is in the range of 0.2 – 0.45 of its diameter. These baffles are placed so that the cut segment is on alternate ends, thus causing the fluid to flow at approximately 90° to the tube bundle. Manufacturing clearances and tolerances are required for an assembly of the exchanger; they provide paths for the fluid to leak past the baffle thus reducing the fluid flowing across the bundle. The tube bundle does not completely fill the shell thus some fluid bypasses the bundle by flowing through the clearance between the shell and the bundle outside diameter. This clearance is a function of the rear-end head design. Figure 45 shows these leakage and bypass flow streams as originally described in [103] and later modified in [104] for a segmentally baffled exchanger. Various streams are described in order of decreasing thermal effectiveness as follows:

B Stream: Crossflow stream flowing over the tubes (and fins if any) between successive windows. This stream is considered fully effective for both heat transfer and pressure drop.

A Stream: Tube-to-baffle hole leakage stream through the annular clearance between the tubes and the baffle holes of a baffle. This stream is created by the pressure difference on the two sides of the baffle. As heat-transfer coefficients are very high in the annular spaces, this stream is considered fully effective.

C Stream: Bundle-to-shell bypass stream through the annular spaces (clearances) between the tube bundle and shell. It flows between successive baffle windows. This stream is only partially effective for heat transfer as it contacts only those tubes near the circumference.

E Stream: Shell-to-baffle leakage stream through the clearance between the edge of a baffle and the shell. This stream is the least effective for heat transfer because it may not come in contact with any tubes (i.e., heat-transfer surface area).

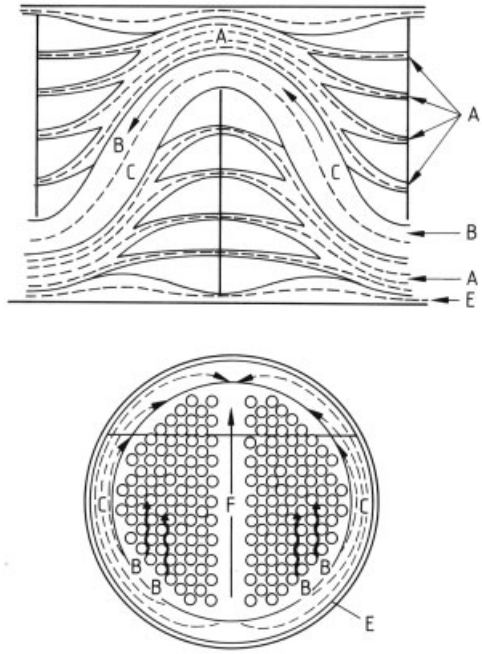

Figure 45. Various flow streams for shellside flow

F Stream: Tube-pass partition bypass stream through open passages created by tube layout partitions (when placed in the direction of the main crossflow stream) in a multipass unit. This stream is less effective than the A stream because it comes into contact with less heat-transfer area per unit volume; however, it is slightly more effective than the C stream.

Based on extensive test data for well-designed exchangers, Taborek [105] reported flow fractions for various streams as shown in Table 20. It is surprising to note from this table that the B stream may represent only 25 % of the total flow for some exchangers in laminar flow. Hence, the predicted performance based on the

conventional LMTD method will not be accurate in general. As a result, there is no need to compute very accurate values of F factors for various exchanger configurations.

Table 20. Flow fractions for various shellside flow streams [105]

Stream	Turbulent	Laminar
Main cross flow (B)	40 – 70 %	25 – 50 %
Bundle bypass (C + F)	15 – 20 %	20 – 30 %
Shell-to-baffle leakage (E)	6 – 20 %	6 – 40 %
Tube-to-baffle leakage (A)	9 – 20 %	4 – 10 %

By defining the pressure drops across the various leakage and bypass paths, TINKER [103] determined the amount of fluid flow in each path by a pressure balance analysis. This calculation procedure is best done with computers since iterations are required as the fluid physical properties influence the flow resistance values of the several paths. Tinker's basic method has been modified and improved, and is used in the better proprietary computer programs.

8.2.1.1. Shellside Heat Transfer

All methods use heat transfer in ideal tube data banks as a basis. An ideal tube bank is a specially designed rectangular bank having no leakage or bypass streams. The fluid velocity is based on the minimum area between tubes which is a function of the tube layout, triangular or square. For circular shells, the velocity is based on the minimum area at the shell diameter.

The best method to predict the shellside heat-transfer coefficient is given in [106], referred to as Bell – Delaware method. The heat-transfer coefficient h_{id} for the ideal crossflow is modified by five correction factors as follows.

$$h_s = h_{id} J_c J_l J_b J_s J_r \tag{144}$$

where

h_{id}= coefficient for pure crossflow in an ideal tube bank assuming the entire shellside stream flows across the tube array at the centerline of the exchanger.

J_c= correction factor for the baffle cut. This correction includes the effect of window and bundle heat transfer. Its value is 1.0 for an exchanger with no tubes in the window and increases to 1.15 for small baffle cuts and decreases to 0.65 for large baffle cuts. For a typical well-designed heat exchanger, its value is near to 1.0.

J_l= correction factor for baffle leakage effects including both tube-to-baffle and baffle-to-shell leakage (A and E streams) with a heavy weight given to the latter. It is a function of the ratio of leakage area to crossflow area and a function of the clearances. A typical value of J_l is in the range of 0.7 – 0.8.

J_b= correction factor for bundle bypass (C and F stream). For a fixed tubesheet construction, $J_b \approx 0.7$, which can be increased to approximately 0.9 by the use of side strips.

J_s= correction factor for baffle spacing at the inlet and outlet sections. The nozzle locations result in larger baffle spacing and lower velocities and thus lower heat-transfer coefficients. J_s usually ranges from 0.85 to 1.0.

J_r= correction factor for any adverse temperature gradient buildup in laminar flows. This correction applies only for shellside Reynolds numbers < 100.

The calculation of these factors requires determination of leakage and bypass areas and involves considerable calculations best done with the aid of computers. Refer to [26], [105], [106] for manual determination of these factors. In a well-designed exchanger, the product of five correction factors is ca. 0.6, but can be as low as 0.4 [106].

The leakage between the baffle and shell (E stream) and the bypass stream between the bundle and shell (C stream) both tend to bypass the heat-transfer surface, and thus affect the effective temperature difference in the exchanger. Based partially on theory and experimental data, some proprietary programs have an additional empirically-derived correction factor that is applied to the log-mean temperature difference. Based on the model of some mixing within the shell, FISHER and PARKER [107] have presented a few of the many possible charts for the temperature difference correction factor F for multipass shell-and-tube exchangers. They show a significant effect on the number of baffles and the bypass ratio especially for less than 16 baffles.

The best predictions are totally computer dependent; they involve the use of a stream analysis method [100], [104] to determine the actual flow through the tube bundle which is then used with ideal tube bank data to determine the heat-transfer coefficient.

8.2.1.2. Shellside Pressure Drop

The pressure drop evaluation on the shell side of a segmentally baffled heat exchanger is complicated due to the presence of bypass and leakage streams in addition to the crossflow stream. Basically, the pressure drop is evaluated first for an ideal crossflow section and an ideal window section. Correction factors are then applied for the leakage and bypass streams. The total pressure drop is then the sum of the pressure drops for each window section and each crossflow section. Excluding the pressure drop associated with the nozzles, it is given by

$$\Delta p_s = \Delta p_c + \Delta p_w + \Delta p_e \tag{145}$$

where Δp_c = pressure drop associated with crossflow zones

$$\Delta p_c = \Delta p_{b,id}\,(N_b - 1)\,R_l R_b \tag{146}$$

Δp_w = pressure drop associated with window zones

$$\Delta p_w = N_b \Delta p_{w,id} R_l \tag{147}$$

Δp_e = pressure drop associated with end (inlet and outlet) sections

$$\Delta p_e = 2\Delta p_{b,id}\left(1 + \frac{N_{cw}}{N_c}\right) R_b R_s \tag{148}$$

Here N_b is the number of baffles, N_c is the number of tube rows in one crossflow section (between baffle tips), and N_{cw} is the number of effective crossflow rows in each window. $\Delta p_{b,id}$ and $\Delta p_{w,id}$ are ideal pressure drops in crossflow and window zones, respectively. The correction factors are defined as:

R_l = correction factor for leakage (A and E) streams. This factor is based on the same factor as J_l but is of different magnitude. Usually $R_l \approx 0.4 - 0.5$ although lower values are possible with small baffle spacing.

R_b = correction factor for bypass flow (C and F) streams. It is different in magnitude from J_b and ranges from 0.5 to 0.8.

R_s = correction factor for inlet and outlet sections having a baffle spacing different from the central section.

Specific values of these correction factors and the computation of ideal crossflow and window zone pressure drops are dealt with in [26], [105], [106].

8.2.2. Disk-and-Doughnut Baffles

The disk-and-doughnut baffles as shown in Figure 26 consist of an alternating series of doughnuts followed by disks with a diameter larger than the doughnut hole. Fluid flows in radially toward the hole, then radially out to the annular space between the disk and the shell. While there is leakage through the clearances between the tube and baffle and the baffle and shell, there is no bypass stream such as the C stream designated by Tinker (Fig. 45). This baffle has better thermal performance per unit pressure drop compared to the segmental baffle due to the absence of the bypass stream C. No research has been reported on the effect of the leakage streams on the heat-transfer and pressure-drop performance of this baffle. Only limited design information is available on this type of baffle [108–112].

8.2.3. Rod Baffles

The rod baffle consists of parallel spacer rods or bars welded to a closely fitted baffle ring as shown in Figure 27. A series of these baffles are rotated to support the tubes on several sides depending on the type of tube layout used. The fluid flows parallel to the tubes except at the ends when entering and leaving the bundle. There are no leakage areas between tubes and baffles but there is a bypass between the bundle outside diameter and shell that passes through the clearance between the baffle ring and shell. Some design data for rod-baffled exchangers are given in [113].

8.2.4. Design Procedure

The solution to the rating problem for a shell-and-tube exchanger is straightforward once all necessary input data are known and the influences of leakage and bypass flow are properly evaluated. The solution to the sizing problem is done iteratively and the steps required for design are as follows:

1) Calculation of the required duty
2) If not already specified, determination of the unknown temperatures or fluid flow rates
3) Estimation of an overall coefficient based on experience, from tables such as Table 21, or estimate the individual coefficients and then calculate the overall coefficient
4) Calculation of the heat-transfer surface area
5) Based on the heat-transfer surface area, flow rates, and process conditions, selection of a suitable type of exchanger
6) For a multipass shell-and-tube exchanger, determination of the correction factor *F,* and whether or not multiple shells are required
7) Selection of tube size, pitch, and layout
8) Selection of tubeside fluid. Selection of number of passes, shell size, and baffle spacing
9) Proceeding to the rating calculations, evaluation of the results, modification of the design and reiteration of these calculations until an acceptable design is obtained.

More detailed solution procedures to the rating and sizing problems are presented in [26], [105] and [106].

9. Thermal Design for Multiphase Heat Exchangers

9.1. Condensers

9.1.1. Considerations in Condenser Design

The principal factors involved in selecting a suitable type of condenser depends on whether condensation is total or partial, whether the vapors are single or multicomponent, and whether some components are noncondensable. The coolant can impose further restrictions, particularly if it vaporizes. The additional requirement for condensate subcooling or the need for vapor–condensate contact within the condenser also influences the condenser type.

The decision of tubeside, shellside, or direct contact condensation depends on the following factors in addition to those discussed in Section 4:

1) *Freezing.* If the condensate can freeze, shellside condensation with large tube pitches is preferred as it is less susceptible to plugging.
2) *Condensate control.* When condensing multi-component mixtures having a substantial boiling or dew point range, or when there are soluble gases present, it is necessary to control the condensate and vapor flow so as to enable the low boilers to condense, or to prevent their condensation or absorption when stripping. The best control is with tubeside condensation, since on the shellside the condensate can drop away from the cooling surfaces. For good condensate drainage, sufficient flow area should be provided at the condensate nozzle; so that the tube bundle does not block the nozzle opening.
3) *Noncondensables.* The accumulation of a small percent of noncondensable gases in condensers can significantly reduce the condensing coefficients. Venting for in-tube condensation is simple as the flow path is fixed; however, for shellside condensation, the gases can segregate in pockets at the cold end, and are difficult to remove unless sufficient pressure drop (vapor velocity) is used to force these gases to the vent outlet. The venting problem is most severe in the typical crossflow condenser with either horizontal or vertical tubes. Some studies on venting are presented in [114–116].

9.1.2. Condenser Types and Their Characteristics

9.1.2.1. Shellside Condensers

Horizontal Tubes. Horizontal condensers can be baffled or crossflow in design. Baffles and/or support plates are usually arranged with vertical cuts, thus causing the vapor to flow from side to side. A horizontal cut is occasionally used under special conditions.

Table 21. Range of overall heat-transfer coefficients in two-fluid heat exchangers [26]

Hot fluid	Cold fluid	Overall heat-transfer coeffcient	
		Btu/h · ft^2 · °F	W/m^2 · K
	Coolers		
Water	water	250 – 500	1220 – 2440
Ammonia	water	250 – 500	1220 – 2440
Aqueous solutions	water	250 – 500	1220 – 2440
Light organics [a]	water	75 – 150	370 – 730
Medium organics [b]	water	50 – 125	240 – 610
Heavy organics [c]	water	5 – 75 [d]	25 – 370
Gases	water	2 – 50 [e]	10 – 240
Water	brine	100 – 200	490 – 980
Light organics	brine	40 – 100	200 – 490
	Heaters		
Steam	water	200 – 700	980 – 3400
Steam	ammonia	200 – 700	980 – 3400
Steam	aqueous solutions		
	< 2 cP	200 – 700	980 – 3400
	> 2 cP	100 – 500	490 – 2940
Steam	light organics	100 – 200	490 – 980
Steam	medium organics	50 – 100	240 – 490
Steam	heavy organics	6 – 60	30 – 300
Steam	gases	5 – 50 [e]	25 – 240
	Exchangers		
Water	water	250 – 500	1220 – 2440
Aqueous solutions	aqueous solutions	250 – 500	1220 – 2440
Light organics	light organics	40 – 75	190 – 370
Medium organics	medium organics	20 – 60	100 – 290
Heavy organics	heavy organics	10 – 40	50 – 200
Heavy organics	light organics	30 – 60	150 – 300
Light organics	heavy organics	10 – 40	50 – 200

[a] Fluids with viscosities < 0.5 centipoise ($= 0.5$ m Pa s)
[b] Viscosities between 0.5 – 1.0 centipoise ($= 0.5 - 1.0$ m Pa s)
[c] Viscosities > 1.0 centipoise ($= 1.0$ m Pa s)
[d] Pressure drop 20 – 30 lbf/in^2 ($= 9 - 13.6$ kg/2.5×10^{-2} m)
[e] These rates are influenced by operating pressures [26]

The lower edges of the baffles are notched to permit draining. The baffle spacing may be made variable if a high vapor velocity needs to be maintained, such as when noncondensables are present. A few close-spaced baffles near the vent gas outlet will promote cooling of these gases. Side-strip baffles are desirable, especially when condensing mixtures, but are also useful with single vapors to improve venting. A significant amount of vapor can bypass the bundle even with modest clearances. With the impingement plate partially blocking the vapor from entering the bundle, it is important to allow sufficient clearance around the edges to reduce pressure drop and possible tube vibration. If the pressure drop becomes limiting, then double segmental baffles, split-flow (J shells), or crossflow (X shells) are used.

For partial condensation of a wide-boiling-range mixture, vertical baffle cuts are acceptable if the flow pattern is maintained in the spray region. A horizontal baffle cut allows the vapor condensate mixture to flow up and down and will give good contact between the vapor and condensate, but requires high pressure drop and a spray flow pattern. The shellside horizontal condenser is an unacceptable design for total condensation of a wide-boiling-range mixture, especially if the dew-point temperature of the lowest-boiling component approaches the temperature of the coolant discharge.

Low (1 – 2 mm) fin tubes are often used in horizontal shellside condensers when the condensing coefficient is low; e.g., refrigerants. The surface tension of the condensate must be low ($< 4 \times 10^{-4}$ N/cm or 40 dynes/cm); otherwise,

the condensate will fill the fins, resulting in loss of effective surface and capacity.

Subcooling is sometimes attempted by partially flooding the condenser with a dam or using an external level control in order to establish a cooling zone. However, the pressure gradient within the shell will affect the actual liquid levels, and hence, the cooling area.

The advantages of horizontal shellside condensers are that the pressure drop can be kept low and the fouling fluid is on the tube side. The tube side can be multipassed to improve its coefficients, reduce fouling, and obtain greater temperature rises. The baffle spacing can be varied to maintain vapor velocities. Finned and enhanced tubes are available for improving heat transfer and reducing size.

The disadvantages are high shell cost if alloy or high-pressure shells are required. Subcooling is poor and not predictable. Venting of crossflow shells is a problem. The vapor and condensate tend to separate and stratify, thus creating a problem in condensing wide-boiling-range mixtures. Tube vibration may occur, and providing adequate supports may be difficult.

Vertical Tubes. Condensation outside of vertical tubes is most often encountered in vaporizers and heaters. The shell side can be baffled, but most often only support plates are used. If the shell is baffled, condensation coefficients are calculated using the same equations as for a horizontal baffled condenser. For unbaffled shells or with support plates, the conservative approach is to ignore the effect of velocity and the effect of the support plate in removing the condensate.

Vertical shell side condensers normally have downflow of vapor. An upflow is rarely used, and if used, should not have any baffles or support plates.

9.1.2.2. Tubeside Condensers

Horizontal Tubes. Horizontal in-tube condensers are made in single, multipass, and U-tube arrangements and in a wide range of tube sizes and lengths. Tube layouts are governed by the coolant requirements. With straight-tube multipass arrangements, the succeeding passes should be below one another (vertically downward). The condensate drop-out between passes will cause maldistribution, and its effect on the condensation process must be considered.

Vapors enter the tubes at high velocity, and the initial condensate is an annular film. If the shear rate is high enough, some condensate may be ripped off the walls and is entrained as a mist. As condensation proceeds, the ratio of condensate to vapor increases and a thick layer of condensate builds up on the tube bottom. Waves will form due to vapor shear and can become high enough to reach the top of the tube, thus forming slugs of vapor and condensate. Under some conditions, these slugs may stop and reverse direction temporarily, thus causing an unsteady or oscillating flow. Finally, as the vapor velocity approaches zero, the condensate will drain from the tube under its hydrostatic head. For large amounts of condensate, the tube may become completely filled, but this condition should be avoided since, as mentioned above, the oscillations may become violent and have been known to destroy tube bundles. Thus, the important key to both heat transfer and pressure drop is the flow pattern.

Horizontal in-tube condensers are most commonly employed as air-cooled condensers. These condensers need little head room, but much floor space. However, they are useful in partial condensation with noncondensables since the path length can be made longer by multipassing, the velocities can be kept up by reducing the number of tubes per pass, and the condensate can drop out between passes to travel in separate lower tubes, thus reducing pressure drop. In this case, the condensate and vapor contact is poor; therefore, the arrangement is unsuitable for total condensation of wide-boiling-range mixed vapors. Subcooling is poor, as the condensate only partially fills the tube. Most condensers are single pass, but two-pass and U-tubes are very common. More than two passes is unusual.

Vertical Tubes – Downflow. A vertical downflow condenser is shown in Figure 46 as a shell-and-tube unit with an outside packed head and a separating head. A fixed tubesheet construction can be used if shellside cleaning is unnecessary or can be done chemically. The space between the upper tubesheet and the up-

per shell nozzle can entrap air; therefore, special vents are drilled into the upper tubesheet and arranged to discharge continuously into an open drain. This assures that the tubesheet is wetted and cooled, and that the water is kept from stagnating, thus reducing fouling and corrosion at the air – water interface. The lower separating head has either a funnel or a baffle designed to minimize reentrainment of condensate into the vent gas stream. The condensate level is kept below the baffle or funnel.

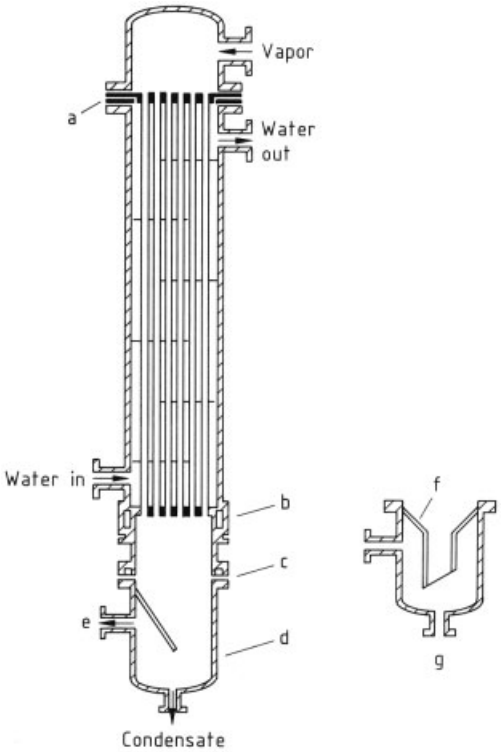

Figure 46. In-tube vertical downflow condenser
a) Tube sheet vent; b) Packed head; c) Slip-on flange with split ring; d) Baffle plate separating head; e) Vapor vent; f) Funnel separating head; g) Alternate head

The vapor enters the top head usually through a radial nozzle, although an axial nozzle can also be used. For axial nozzles, the nozzle entrance velocity head pressure should be very low compared to the condenser tube pressure drop to ensure that maldistribution is not severe. If needed, a perforated impingement with 5 – 10 % hole area placed 0.5 – 1.0 nozzle diameter downstream will help.

The vapors condense on the tube wall as an annular film and drain to the bottom. Tube diameters are usually 19 or 25 mm, although for low pressures larger tubes up to 50 mm are used to reduce pressure drop. Occasionally small diameters of 16 mm are used. At the end of the condensing zone, there is a vapor – noncondensable gas interface, and below this interface the condensate is subcooled as a falling film. As the load on the condenser varies, the vapor – gas interface (and hence, the ratio of condensing to subcooling lengths) will change. System pressure control is obtained by controlling the vent gas pressure. If the noncondensable gases in the incoming vapor are insufficient to maintain pressure control for a reduction in load, then an innert gas is bled in for pressure operation, or air is bled into the vacuum line for vacuum operation. For atmospheric operation, the vent line exhausts to a stack.

It is possible to operate without the inert gas by flooding the bottom of the tubes; however, this is not recommended because (1) subcooling is poorer; (2) there is no venting, and noncondensables can accumulate and reduce condensing and/or subcooling capacity; (3) pressure control is more difficult; and (4) there is a possibility of a slugging effect.

The shellside is baffled, and the coolant flows in a single pass countercurrent to the vapor. If control of the condensate temperature is necessary, a cocurrent flow of the coolant will give better subcooling control, but more coolant may be used.

There are many advantages to the vertical downflow condenser as the condensate is always in contact with a cold wall and with the vapors, thus promoting the condensation of low boilers from a wide-boiling-range mixture. The condensate washes all the surfaces, which is an advantage under certain corrosive conditions. Sprays can be readily placed in the upper head to supply additional liquid for absorption of low boilers or to aid in washing the tubes when condensing dusty vapors. The vapor flow path is well defined, and venting of noncondensables is positive. The subcooling coefficients are high and predictable.

The disadvantages are that the coolant, which may have a greater fouling tendency, is on the shell side and may require a packed head design as well as the need for a continuous vent through the tubesheet. At low absolute pressures

(< 3.3 kPa or 25 mm Hg), the tubeside pressure drops become high and require use of large (38 – 50 mm) tubes in order to have a reasonable condensing length.

Vertical Tubes – Upflow. A typical in-tube vertical condenser with vapor flowing upward and the condensate returning by gravity is called a *knockback* or *reflux condenser*. Special features are that the lower ends of the tubes that extend beyond the bottom tubesheet are cut at an angle of about 60 – 75 degrees to provide drip points for the returning condensate. The shellside should be vented through the upper tubesheet. Depending on coolant fouling characteristics, a fixed or removable bundle can be used. Vapors enter the bottom head through a radial nozzle, but frequently the condenser sits directly on top of the distillation column without a nozzle. Reflux condensers usually are short (2 – 3 m) with large diameter (≥ 25 mm) tubes.

Reflux condensers are used principally when a hot condensate return is necessary and when small amounts of low-boiling components are to be stripped out or prevented from condensing. However, the high boilers entering the condenser will raise the condensing temperature and may thus avoid the use of low-temperature coolants. The effectiveness of separation is not as good as the addition of more packing or plates in a distillation tower with a product take-off below the top plate. As pointed out in [117], it is not possible to specify fractionation, heat requirements, and coolant temperatures independently, because only two of the three can be selected arbitrarily. Also, as shown in [118], the fractionation is affected by throughput rate. Based on a number of simplifying idealizations, TROSTER [118] suggested a two-part approach to the design of reflux condensers.

The major disadvantage is the capacity limitation due to flooding. Flooding occurs when the upward vapor velocity prevents the free return of condensate and the tubes periodically load with condensate and then dump it. At higher velocities, all the condensate can be blown out of the top of the condenser, but then only a small fraction of vapor is condensed. Because of the relatively low velocities in a reflux condenser, they are not suitable for condensation of mixtures containing large amounts of noncondensable gases.

A number of studies have been made on flooding velocities in tubes resulting in different equations or graphs. The English et al. [119] equation below gives results close to HOLMES's [120] curves for beveled tubes.

$$G_{\mathrm{f}} = \frac{80 d_{\mathrm{i}}^3 \varrho_{\mathrm{l}}^{0.46} \sigma_{\mathrm{l}}^{0.09} \varrho_{\mathrm{g}}^{0.5}}{\mu_{\mathrm{l}}^{0.14} (\cos\theta)^{0.32} \Lambda^{0.07}} \tag{149}$$

Here, G_{f} is the mass velocity at the start of flooding (kg m^{-2} s^{-1}), d_{i} is the tube *ID* (mm), ϱ_{l}, ϱ_{g} are densities (kg/m^3), σ_{l} is the surface tension (dynes/cm), μ_{l} is the dynamic viscosity (cP), θ is the taper of the tube end from horizontal (degrees), and Λ is the weight ratio of condensate to vapor at the bottom of the tube. Flooding correlations are reviewed in [121].

9.1.2.3. Air-Cooled Condensers

An air-cooled condenser consists of a shallow bank of finned tubes with the condensing vapor inside the tubes frequently in a single-pass arrangement. With a uniform distribution of vapor, the condensing length of the first tube row is considerably shorter than in the last tube row. As pointed out in [122], any excess vapor leaving the last tube row will not backflow into the first row due to the accumulation of noncondensable gases at the end of the condensing zone. A number of methods to reduce the backflow and inert gas blanketing in these condensers are reviewed in [123].

Air coolers are designed on an air temperature that is usually exceeded by only 2 – 5 % of the annual hours based on the recorded daily temperatures of the plant (or closest weather station). The condenser will have excess capacity for > 90 % of the time; hence, few additional safety factors, other than allowances for fouling, are added to the theoretical surface.

The air face velocity ranges from 2 to 4 m/s (6 – 12 ft/s) depending upon the number of tube rows and fan design. The fan is designed to handle large volumes of air at pressure drops of approximately 200 – 300 Pa (0.8 – 1.2 in H_2O) and its speed is limited by the noise it generates.

Either forced or induced draft coolers are used depending on the local preference or the designer's judgement of the relative advantages and disadvantages of these types. For forced

draft, the fan is in the cool inlet air resulting in better fan performance and heat-transfer coefficients. However the upper exposed tube bundle is more susceptible to sun, wind, or rain, and control of condenser pressure can be more difficult. Induced draft, with its high discharge velocity avoids air recirculation problems and is less influenced by surrounding buildings. Air coolers must be carefully integrated with the plant layout of buildings and equipment, and prevailing winds, otherwise the performance will suffer. The cooler must also be designed for, and be capable of operating over, the full range of air temperatures. Further details are given in [10].

9.1.3. Condensation of Mixtures

The two classes of mixtures are: (1) mixtures in which some of the components are not condensable, and (2) mixtures that are totally condensable or where the low boiling components are soluble in the condensate. Each is a nonisothermal condensation process and has a vapor-phase diffusional resistance. Design equations for binary mixtures have been developed but approximate design methods are used for more complex mixtures, as the theories of complex mixtures are too cumbersome.

For a mixture with a noncondensable gas the Colburn – Hougen [124] equation is used to find the heat flux at a given point along the condenser:

$$U'(T_i - T_c) = h'_g (T_v - T_i) + K_m i_{lg} (p_v - p_i) \tag{150}$$

where U' = the overall coefficient from the condensate interface to the coolant, and

$$h'_g = \left(\frac{a}{1 - e^{-a}}\right) h_g \tag{151}$$

$$a = \frac{N_v M_v c_p}{h_g} \tag{152}$$

$$h_g = j c_p G_v / Pr^{2/3} \tag{153}$$

$$K_m = j \left(\frac{G_v}{M_m}\right) Sc^{-2/3} \tag{154}$$

Here, N_v is the molal rate of condensation (mol m^{-2} s^{-1} or mol ft^{-2} h^{-1}), M_v is the molecular weight of the vapor, G_v is the vapor mass velocity, and M_m is the mean molecular weight of noncondensables and vapor.

The Colburn – Edison [125] equation is used to determine the condensation path and the conditions at the next point downstream in the condenser:

$$\frac{dp_v}{dT_v} = \frac{(p_v - p_i)(p - p_v)}{(T_v - T_i)\,\Delta p_{lm}} \left(\frac{e^a - 1}{a}\right) \left(\frac{Pr}{Sc}\right)^{2/3} \tag{155}$$

where

$$\Delta p_{lm} = \frac{(p - p_v) - (p - p_i)}{\ln\left[(p - p_v)/(p - p_i)\right]} \tag{156}$$

and p = total pressure, p_i = interface pressure, and p_v = vapor pressure.

The design procedure for a condenser with noncondensables is as follows:

1) At the vapor inlet, the vapor partial pressures and vapor and coolant temperatures are known. The partial overall coefficient U' can be determined from the condensate, wall, coolant, and fouling coefficients with an equation similar to Equation (7). The heat- and mass-transfer coefficients are determined for the inlet Reynolds number using an appropriate $j - Re$ curve.
2) Assume a value of T_i and find the corresponding value of p_i from the vapor pressure curve.
3) Solve Equation (150). If both sides of Equation (150) balance, then proceed to Step 4, otherwise assume another value of T_i and repeat Steps 2 and 3.
4) For the balanced condition of Step 3 solve Equation (155).
5) Assume a step change in either the vapor temperature or partial pressure, using Equation (155) to determine the step size of the other variable. If the value of Equation (155) is very small, make the assumed step change in the partial pressure; if large, in the vapor temperature.
6) By heat- and mass-balance calculations, the conditions at the next step are determined. Steps 1 to 5 are repeated until the vapor exit conditions are reached: the calculated coolant temperature at the vapor exit should equal the available inlet temperature.
7) Plot $1/(U' \Delta T)$ versus q'', and graphically integrate the curve to determine the required heat-transfer surface area A.

For a condensable binary mixture, the Colburn – Drew [126] equation is

$$U'(T_i - T_c) = h'_g(T_v - T_i) + F'\Delta i \ln\left(\frac{z - y_{Ai}}{z - y_{Av}}\right) \quad (157)$$

where

$F' = j(G_v/M_m)(Sc)^{-2/3}$
$z =$ $\tilde{N}_{Ai}/\tilde{N}_t$ mole fraction of A_i in the total $\tilde{N}_t$ condensing moles
$\Delta i =$ molal heat of condensation
$y =$ mole fraction of component A at $i =$ interface, $v =$ vapor

For multicomponent (3 or more) mixtures, the theoretical equations are unwieldy, and approximation methods such as that by Bell – Ghaly [127] can be used. Here it is assumed that the condensation path follows the boiling point curve.

The design procedure for the Bell – Ghaly method is:

1) Calculate condensing curves as a function of the vapor temperature:
   a) Total condensing stream enthalpy assuming thermodynamic equilibrium between phases
   b) Vapor weight flow rate
   c) Total heat removed from the vapor inlet to a given point Q_T
   d) Coolant temperature
2) For a number of steps along the Q_T curve, calculate the ratio Z of vapor sensible heat removed to the total vapor enthalpy change and plot Z as a function of Q_T.
3) Calculate the condensing film coefficient and the vapor film coefficient as a function of condensate and vapor flows, respectively, and both as functions of the total stream enthalpy Q_T.
4) Calculate U' as a function of Q_T. U' is the partial overall coefficient consisting of the condensate, wall, coolant, and fouling coefficients.
5) Calculate $(1 + ZU'/h_g)/U'(T_v - T_c)$ as a function of Q_T, where T_v and T_c are the local vapor and coolant temperatures, respectively.
6) Calculate A by integrating the function of step 5) numerically or graphically from $Q_T = 0$ to Q_T^* where Q_T^* is the total heat removed in the condenser.

A comparison of the above methods is given in [128] and further information on condenser design and operation in [7], [81–84].

9.1.4. Mean Temperature Difference

The calculation of a single mean temperature difference (usually a log-mean) based on terminal temperatures for condensers is frequently done; however, it is theoretically wrong as the condensing coefficients vary throughout the exchanger. In some cases, the error is not large but the best approach is to calculate the local overall coefficients together with the local temperature differences and integrate the products along the condensation path.

9.1.5. Desuperheating

Vapor may enter the condenser with a large amount of superheat, and, depending upon the local conditions, the tube wall may be dry or may have a condensate layer. If the wall is dry, its temperature must be above the vapor saturation temperature, and the vapor heat-transfer coefficient is that of a gas, and the temperature difference is based on the superheat value. Note that the heat flux will be higher than for the condensing value, as the tube wall temperature is above the saturation temperature. The only exception is where the coolant coefficient is a function of the wall temperature, such as could occur with boiling coolant in the film boiling region. If the wall is below the saturation temperature, a condensate layer is formed whose interface temperature is at the saturation temperature; hence, the temperature difference is based on the saturation temperature. The amount of condensate formed is, however, based on the total heat released which is the latent heat plus the amount of superheat.

9.1.6. Subcooling

All condensate in contact with the condenser tube wall is subcooled because of the temperature gradient across the condensate film. The

amount of subcooling depends on the temperature difference and on whether the film is in laminar or turbulent flow. Often further subcooling is required before the condensate is discharged.

For vertical tube condensers with downward vapor flow, the condensate film below the condensing section will be subcooled and the condensate coefficient is for the laminar region:

$$h\left(\frac{\mu}{k^3\varrho^2 g}\right)_1^{1/3} = 0.78\left[\left(\frac{c_p u}{k}\right)_1\left(\frac{\mu^2}{\varrho^2 g}\right)_1^{1/3}\left(\frac{1}{L}\right)\right]^{1/3}\left(\frac{4\Gamma}{\mu_l}\right)^{1/9} \tag{158}$$

In the turbulent region, $4\Gamma/\mu_l > 2100$, and the following equation should be used:

$$h\left(\frac{\mu}{k^3\varrho^2 g}\right)_1^{1/3} = 5.7\times 10^{-3} Re_l^{0.4} Pr_l^{0.34} \tag{159}$$

where Γ is the condensate mass flow rate per unit perimeter.

Inside the horizontal tubes, the condensate flows in a layer whose thickness, area, and coefficient cannot be predicted with any reasonable accuracy. Subcooling is poor, and this arrangement is not recommended.

Subcooling on the outside of horizontal tube bundles is frequently attempted by creating a dam near the condensate outlet in an effort to create a pool partially submerging the bundle. Although some subcooling is accomplished, it cannot be predicted accurately, as the actual liquid level may be substantially different from the static level because of pressure gradients in the shell. Likewise, the velocity cannot be predicted adequately. A conservative estimate can be made by using a natural convection coefficient with the area calculated on the basis of the static liquid level. Predictability of subcooling is poor, temperature control is almost impossible, and thermal stress problems become important due to differential expansion of the tubes. Separate coolers should be used if subcooling is critical with shellside condensation.

9.2. Vaporizers

9.2.1. Kettle and Internal Reboiler Design

Flow Circulation Mechanism. In a kettle reboiler, the tube bundle is submerged only a few centimeters below the stagnant liquid level. When boiling takes place, the vapor generated within the bundle reduces the local density. Rising bubbles create an upward flow of a vapor–liquid mixture through the bundle and create a froth mound above the bundle. The vapor escapes from the froth and the liquid spills to the bundle sides and returns to the bottom portion through the eccentric annular space between the bundle and the shell. Liquid enters the bundle in the lower half of its periphery. The complex circulation and velocity pattern in the bundle makes meaningful experimental measurements difficult and the effect of circulation on heat transfer has to be deduced from overall performance by means of assumed mathematical models.

The two-phase flow generated within the bundle will affect the heat-transfer rate. In large bundles with small tube pitches, the fraction of vapor can become high enough to vapor blanket (i.e., the mist flow regime) upper bundle portions and reduce performance. Enlarged tube pitches and/or provision of lanes for vapor escape are used in very large bundles.

Boiling Coefficient and Maximum Heat Flux. Since a tube bundle is an assembly of tubes, the basic approach to heat transfer in bundles is to use the single tube pool boiling coefficient modified for the influence of neighboring tubes. Early experiments indicated little difference between single tubes and bundles but later experiments indicated an enhanced performance for bundles.

The bundle average boiling heat-transfer coefficient $h_{\mathrm{tp,\,m}}$ is empirically modeled as

$$h_{\mathrm{tp,m}} = h_{\mathrm{nb1}} F_{\mathrm{B}} F_{\mathrm{m}} + h_{\mathrm{nc}} \tag{160}$$

where h_{nb1} = single-tube nucleate boiling heat-transfer coefficient

$$h_{\mathrm{nb1}} = 0.00417 p_{\mathrm{cr}}^{0.69}\left(q''\right)^{0.7} F_{\mathrm{p}} \tag{161}$$

$$F_{\mathrm{p}} = 1.8\left(\frac{p}{p_{\mathrm{cr}}}\right)^{0.17} + 4\left(\frac{p}{p_{\mathrm{cr}}}\right)^{1.2} + 10\left(\frac{p}{p_{\mathrm{cr}}}\right)^{10} \tag{162}$$

F_{B} = factor for bundle geometry ≈ 1.5 as a conservative value [20], but can range from 1 to 3

F_m = factor for mixtures ($F_m = 1$ for single component mixtures)

$$F_m = \frac{1}{1 + 0.027\,(q'')^{0.1}\,\Delta T_{BR}^{0.7}} \tag{164}$$

form [129]

$$h_{nc} = \text{natural convection heat} - \text{transfer coefficient for a single tube.} \tag{165}$$

In the foregoing equations, p_{cr} is the critical pressure in kPa, q'' is the bundle heat flux in W/m^2, and ΔT_{BR} is the boiling range (= dew point – bubble point temperature) in °C. The last two terms in Equation (162) should be used only when $p/p_{cr} > 0.2$, otherwise they should not be included in the calculation. For mixtures, h_{nb1} in Equation (161) is based on the fluid properties of the light component. The lower limit of F_m is 0.1. Newer developments for F_b involve relating it to the circulation velocities through the bundle but much of this information is based on proprietary data.

The bundle coefficient $h_{tp,m}$ does not show sensitivity to surface conditions; however, finned surfaces do give better performance but mostly due to the increased surface.

The maximum (critical) flux for a bundle is different and lower than the single tube maximum flux. The bundle maximum flux is given as

$$q''_{B,max} = q''_{1,max}\,\Phi_B \tag{166}$$

where

$$q''_{1,max} = 367\,p_{cr}\left(\frac{p}{p_{cr}}\right)^{0.35}\left(1 - \frac{p}{p_{cr}}\right)^{0.4} \tag{167}$$

where p and p_{cr} are in kPa and q'' in W/m^2. Based on some plant experience, Palen and Small [99] proposed a bundle correction model for the vapor blanketing effect. They developed a correction term, Φ_B in Equation (166), which can be simplified to

$$\Phi_B = 2.2\left(\frac{\pi d_B L}{A}\right) = \frac{K p_t}{d_0\,(d_B/p_t)^{1.1}} \tag{168}$$

where

$$K = \begin{cases} 4.12 & \text{for square pitch} \\ 3.56 & \text{for triangular pitch} \end{cases} \tag{169}$$

p_t = tube pitch

and

$$0.1 < \Phi_B < 1 \tag{170}$$

L and A in Equation (168) are the bundle tube length and total outside tube surface area. Equation (168) is reported to be conservative [130].

9.2.2. Horizontal Shellside Thermosiphon Reboiler Design

The horizontal shellside thermosiphon in an X shell arrangement is similar to a kettle except that the circulation through the bundle is controlled by the piping arrangement. In kettles, the circulation pattern and rates are unknown. With the use of baffled shells, E, G, etc., the velocities across the tubes can be increased and the effective depth of a crossflow bundle is increased. With baffled shells, the flow pattern approaches that for vertical in-tube thermosiphons in that the heat-transfer coefficient and pressure drops must be calculated incrementally along the bundle due to the influence of changing vapor quality.

9.2.2.1. Pressure Balance in the Recirculation Loop

A pressure balance is illustrated by Figure 47. The available pressure drop across the exchanger, $p_i - p_e$, is created by the static head (H) differences less other losses in the inlet and outlet piping. Here

$$p_i = p_0 + \varrho_l H_i \frac{g}{g_c} - \Delta p_l - \Delta p_m \tag{171}$$

and

$$p_e = p_0 + \varrho_{tp} H_e \frac{g}{g_c} + \Delta p_{f,tp} + \Delta p_{m,tp} \tag{172}$$

where Δp_l and Δp_m are the pressure drops due to the frictional and momentum change for liquid entering the exchanger, and $\Delta p_{f,tp}$ and $\Delta p_{m,tp}$ are the pressure drops caused by frictional and momentum change for the two-phase mixture leaving the exchanger.

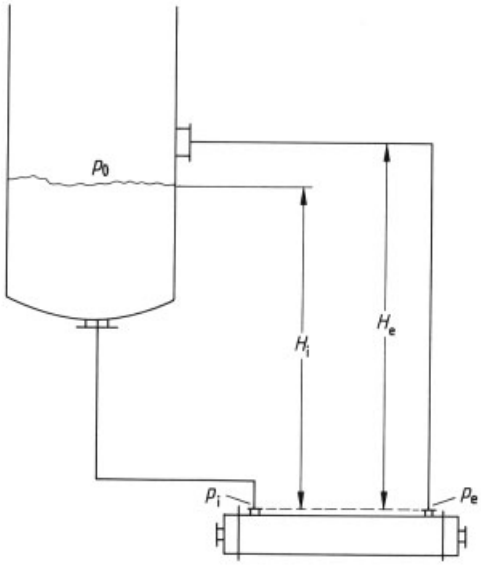

Figure 47. Horizontal thermosiphon reboiler installation with various pressure and static head terms

9.2.2.2. Shellside Heat Transfer and Pressure Drop

There is little published heat-transfer data for two-phase shellside boiling; a summary is given in [77]. The basic approach is similar to two-phase flow in tubes. The various bypass and leakage streams that exist for shellside flow and their effect for single-phase and two-phase flows complicate the analysis. The single-phase leakage effects are known but the bypass and leakages may be different for two-phase flows. The shellside heat transfer can be determined as:

$$\frac{h_{\mathrm{tp}}}{h_{\mathrm{lo}}} = \left(\frac{\Delta p_{\mathrm{f,tp}}}{\Delta p_{\mathrm{l}}}\right)^n \tag{173}$$

where n seems to range from 0.4 to 0.5 and an average of 0.45 can be used. The heat-transfer coefficient h_{lo} and pressure drop Δp_{l} are calculated according to methods given earlier (Chapter 8) where allowances are made for flow bypass and leakage, and in this manner some correction is also made for the two-phase flow leakages.

9.2.2.3. Design Methods and Special Considerations

Horizontal shell thermosiphons have used many types of shells, X, E, G, etc., and their selection is influenced by heating medium flows, temperatures, etc. as well as the physical restraints of available heights and ground area. This selection process is beyond the present scope and a simple E shell is used for basic discussion.

Tube diameter and length are mostly based on the heating medium requirements. Tube pitch ratio is usually 1.25 – 1.33 and a square arrangement, either staggered or inline, is used to allow for cleaning. Segmental baffles with somewhat larger baffle cuts (ca. 35 %) are used and placed on proper spacings to avoid vibration problems. The baffle cuts may be horizontal or vertical. Horizontal cuts promote better mixing but have higher pressure drops, while vertical cuts may have phase separation and possible different tube wall temperatures in each phase zone.

The vaporization fraction should be kept low to achieve higher circulation rates, lower fouling and higher heat-transfer coefficients. The available head will limit the allowable pressure drop and circulation. The discharge piping for the vapor – liquid mixture must also be sized to keep the frictional pressure drop low yet keep the mixture out of the slugging flow regime for the expected range of process operations. The shellside flow patterns and flow pattern maps are shown in Figures 39 and 42.

The heat-transfer and pressure drop calculations are made on an incremental basis by calculating a preheating zone and then by selected increments of vaporization for the boiling zone. Since the design process starts with an assumed arrangement and is then modified by iteration, the use of computers is essential although hand calculations are possible but tedious.

The designer should at all times be aware of the limitations of these methods and flow pattern maps.

9.2.3. Vertical Thermosiphon Reboiler Design

The hydrostatic head on the liquid entering the tubes results in a subcooling with respect to the local pressure, thus there is a liquid preheating zone required to bring the liquid to the boiling point as it rises through the tube. The boiling zones can consist of several types: subcooled

boiling, nucleate boiling modified by a superimposed two-phase convective flow, mist flow, or film boiling also modified by the liquid core velocity. Heat-transfer and pressure-drop relations for each of these zones are different, hence each zone is calculated separately. For the two-phase flow zones, further subdivision is necessary to account for the changing ratio of liquid to vapor. Various zones and the effects on wall temperature and heat transfer are shown in Figure 44.

9.2.3.1. Pressure Balance in the Recirculation Loop

Vertical thermosiphons require the simultaneous solution of flow and heat-transfer rates in the complete recirculation system. As shown in Figure 44, the driving force for the circulation is the static head of liquid in the column base or reservoir which must equal the various losses. That is

$$\Delta p_{\mathrm{T}} = \Delta p_{\mathrm{i}} + \Delta p_{\mathrm{t}} + \Delta p_{\mathrm{e}} \tag{174}$$

where the total pressure drop Δp_{T} is the available head in the reservoir. The inlet piping and header losses, Δp_{i}, are easily calculated for the liquid flow. The pressure losses in the tubes, Δp_{t}, due to the changing ratio of the two phases, is a difficult calculation. It requires an incremental determination along the tube of the flow pattern and the associated hydrostatic, acceleration, and frictional losses. The exit head and piping losses, Δp_{e}, must be kept at $< 30\,\%$ of Δp_{T} to avoid stability problems and can only be approximated due to lack of data.

In the following discussion, the heat-transfer and pressure-drop methods for each flow regime are considered together.

9.2.3.2. Flow Regimes and Heat Transfer

Non-Boiling Zone. The liquid entering the tube is at a higher pressure than its saturation pressure leaving the tubes due to the static head; hence, sensible heat is required to bring the liquid to its local boiling point plus any superheat to activate the bubble nuclei. The liquid zone heat-transfer coefficients are calculated as follows: For the laminar flow where $L/d_{\mathrm{i}} > 50$ and $Re < 2300$, the empirical modification of the laminar flow equation is [60]:

$$Nu_{\mathrm{l}} = 0.17 Re_{\mathrm{l}}^{0.33} Pr_{\mathrm{l}}^{0.43} \left(\frac{Pr_{\mathrm{l}}}{Pr_{\mathrm{w}}}\right)^{0.25} Gr_{\mathrm{l}}^{0.1} \tag{175}$$

where

$$Gr_{\mathrm{l}} = \frac{d_{\mathrm{i}}^3 \varrho_{\mathrm{l}}^2 g \beta (\mathrm{T_w} - T_{\mathrm{l}})}{\mu_{\mathrm{l}}^2} \tag{176}$$

Here the natural convection effect in laminar flow is accounted for by the Grashof number term ($Gr_{\mathrm{l}}^{0.1}$); β is the coefficient of thermal expansion in 1/°C (or 1/°F).

For the transition and turbulent flow regions, $Re > 2300$, and the Gnielinski correlation, given in Table 10, should be used.

The pressure drop for the preheat zone is

$$\Delta p = \left[\frac{4 f G^2}{2 g_{\mathrm{c}} d_{\mathrm{i}} \varrho_{\mathrm{l}}} + \varrho_{\mathrm{l}}\right] L_{\mathrm{ph}} + 0.5 \frac{G^2}{2 g_{\mathrm{c}} \varrho_{\mathrm{l}}} \tag{177}$$

where L_{ph} is the preheat zone length. The last term of Equation (177) represents the pressure drop due to acceleration.

Subcooled Boiling Zone. A subcooled boiling takes place when the local tube wall temperature is sufficiently high to initiate bubble formation on the wall nuclei, and when the local bulk liquid temperature is below the local boiling point. The formation of bubbles at the nuclei and the bubble collapse due to the subcooled state of the bulk of the liquid causes an increase in heat transfer but the rapid collapse of the bubble has very little effect on acceleration of the fluid or its density, thus liquid properties are used.

The agitation due to bubble formation and collapse increases the heat flux as an additional heat flux q''_{nb} to the heat flux q''_{l} due to the convective flow of the fluid. Their sum is the total heat flux q''. In terms of the boiling coefficient, this is

$$h_{\mathrm{tp}} = h_{\mathrm{nb}} \left(\frac{T_{\mathrm{w}} - T_{\mathrm{sat}}}{T_{\mathrm{w}} - T_{\mathrm{l}}}\right) + h_{\mathrm{l}} \tag{178}$$

where h_{l} is the same as the liquid single-phase coefficient in the preheat zone, Equation (175) or the Gnielinski correlation (see Table 10), and h_{nb} is determined from some suitable predictive equation for nucleate boiling. Usually h_{nb} is based on equations for single tube nucleate

boiling in pools (see subsection on the Boiling Zone below).

The pressure drop in this zone is calculated as in the preheat zone since it is assumed that the rapid collapse of bubbles results in a negligible vapor fraction and neglible acceleration.

Boiling Zone. Once the bulk temperature reaches the local saturation temperature, the bubbles accumulate and the vapor fraction increases; a variety of two-phase flow patterns are found such as bubble, slug, or annular. The boiling heat-transfer zone extends to the location where the mist zone begins. As a result of the changing vapor fraction, the local heat-transfer rates vary and it is necessary to calculate the local heat fluxes for incremental steps in vapor fraction or in incremental lengths of the tube.

The boiling heat-transfer coefficient h_{tp} is again a function of the convective and nucleate boiling coefficients (see Eq. (124)).

The saturated flow boiling heat-transfer coefficient can not be predicted with a high degree of accuracy as its value is dependent not only on the physical properties and heat flux, but also on the surface characteristics, flow regimes and flow rates, and on mass transfer in both phases. The nucleating characteristics of surfaces are presently impossible to specify and, furthermore, they change due to deposits or corrosion.

The maximum heat flux has been empirically correlated and given by Equation (129) or (136).

Mist Zone. The mist zone is not often encountered in vertical tube thermosiphons as the fraction vaporized per pass is normally too small.

The small volume of liquid in mist flow is entrained as droplets and the tube wall is essentially dry thus the coefficient approaches that for gas flow. In this regime, sensible heat is transferred to the gas which in turn transfers some of the heat to the droplets until they are completely evaporated after which only sensible heat transfer takes place. The main problem is the determination of the vapor temperature, and hence the temperature difference. The two extreme conditions are: (1) no heat is transferred to the droplets so that the vapor temperature rises rapidly; and (2) heat is rapidly transferred to the droplets until they disappear and the vapor is at the saturation temperature during this evaporation phase. Condition 1 is approached at low pressures and velocities, and condition 2 at high pressures and velocities. The actual case is somewhere inbetween. An intermediate result can be obtained by assuming half or more of the heat is transferred to the droplets and the remaining heat raises the vapor temperature. The LMTD and heat flux can then be calculated.

The mist region can be determined from the following equation which is based on the Fair map [131]

$$G_{\mathrm{m}} = cX_{\mathrm{tt}} \tag{179}$$

where G_{m} is the maximum mass velocity before mist flow begins and $c = 2.44 \times 10^3\ \mathrm{kg\,m^{-2}\,s^{-1}}$ or $1.8 \times 10^6\ \mathrm{lbm\ ft^{-2}\,h^{-1}}$.

Film Boiling. This type of boiling should be avoided due to control problems, possible fouling, and lack of data on pressure drop calculations. For unavoidable high temperature differences, the film boiling heat-transfer coefficient can be calculated by the Glickstein and Whiteside correlation [132]:

$$\frac{hd_{\mathrm{i}}}{k_{\mathrm{v}}} = 0.106 Re_{\mathrm{v}}^{0.64} Pr_{\mathrm{v}}^{0.4} \left(\frac{\varrho_{\mathrm{m}}}{\varrho_{\mathrm{v}}}\right)^{0.5} \tag{180}$$

where the bulk mean density on a no-slip basis is

$$\varrho_{\mathrm{m}} = \frac{\varrho_{\mathrm{l}}}{1 + x\left(\frac{\varrho_{\mathrm{l}}}{\varrho_{\mathrm{v}}} - 1\right)} \tag{181}$$

In film boiling inside a tube, a core of liquid is surrounded by an annular layer of a low viscosity gas, thus the frictional pressure drop will be low. The major losses are due to hydrostatic head and momentum losses. Since the liquid core velocities can easily increase by a factor of 2 or more, the losses in external piping due to the higher circulation also become significant.

9.2.3.3. Approximate Design Methods

The maximum heat flux in thermosiphons with essentially no inlet and exit nozzle and piping losses is limited by a surging of flow through the exchanger. This surging instability is a function of the boiling fluid and the operating conditions, and is dependent upon the hydraulic resistances in the piping circuit. It is very sensitive

to pressure drop in the exit vapor-mixture lines, which should be kept $< 30\,\%$ of the total available head. It is strongly recommended that the cross-sectional flow area of the exit pipe be equal to the total cross-sectional area of the tubes. Increasing the flow resistance in the inlet pipe, however, has a stabilizing effect, thus the use of a valve or orifice should be anticipated. The inlet line usually has a cross-sectional area of $20-50\,\%$ of the total tube cross-sectional area.

If the available temperature difference is sufficient to reach $q''_{\mathrm{B,\,max}}$ (generally $> 45\,^{\circ}\mathrm{C}$ or $80\,^{\circ}\mathrm{F}$), then the thermosiphon design is a selected fraction ($\lesssim 0.7$) of $q''_{\mathrm{B,\,max}}$, where the fraction is based on the effect of potential process upsets. When the maximum available temperature difference is insufficient to reach $q''_{\mathrm{B,\,max}}$, the tube must be divided into a nonboiling preheat zone and a boiling zone for a preliminary and quick calculation. An average vapor fraction is used to calculate the boiling zone. The tube inlet velocity is assumed to be $0.3-1$ m/s ($1-3$ ft/s), and the heat transfer and pressure drops are calculated. Several trials may be needed to reach a solution where the pressure drop equals the available pressure head.

A better calculation can be made by dividing the tube into a number of sections and calculating the heat transfer and pressure drop by the method above. Such calculations are incorporated into computer programs due to the complexity and number of calculations involved. Although the computer results may give a better answer than the approximation methods, the predictability of boiling heat transfer is poor and substantial errors ($> 30\,\%$) are still possible.

9.2.4. Special Design Considerations

Circulation through the thermosiphon is an important and necessary characteristic of the vaporizer. Many of the piping features and operating limits have already been discussed; however, those as well as methods and experimental tests pertain to narrow boiling range liquids. Wide boiling range mixtures, especially those where the maximum boiling point exceeds the available heating medium temperature, may cease to circulate unless the liquid level is sufficiently high to provide a complete liquid circuit, or if vapor sparging beneath the lower tubesheet is used to initiate circulation.

Thermosiphons operate best in the pressure range of moderate vacuum where the boiling point elevation due to the static head is not too large to about half the critical pressure. Beyond these limites the preheating zone becomes too large and the available driving force for circulation is insufficient.

As in all vaporizers, the maximum and minimum temperature differences need to be determined for clean and fouled conditions and for the range of process flows and pressures in order to acertain that the vaporization is controllable; e.g., in steam heated systems is it possible for condensate to be removed and for the inerts to be vented?

9.2.5. Mean Temperature Difference

For boiling outside of tubes, the determination of a mean temperature difference is an arbitrary choice based on the amount of subcooling in the feed, the difference between the boiling points of entering and leaving liquid, and to some extent, the geometry of the reboiler. The problem is the degree of mixing that occurs in the shell which depends on the circulation rate. As described in [130], the bulk temperature within the tube bundle varies with the length and the effective ΔT differs substantially from the ΔT based on exit temperatures; thus, depending upon the idealizations, several methods of calculating the mean temperature difference can be used. Some of the choices are:

1) For a pure component or a narrow boiling range mixture and a constant temperature heating medium, the difference between the condensing and boiling temperatures is used. If the heating medium is transferring sensible heat, a logarithmic mean of the vapor saturation temperature and the heating medium terminal temperatures is used.
2) For a wide boiling range mixture or where the sensible heat load is a substantial fraction of the total heat load, a counterflow logarithmic mean temperature difference (LMTD) gives optimistic results [130] and a LMTD based on the exit vapor temperature is recommended.

3) When the effect of static head on the boiling point is significant, e.g., in large bundles and/or vacuum service, the boiling temperature should be based on the mean pressure in the bundle including the imposed static head.
4) When a horizontal thermosiphon reboiler is used, a counterflow LMTD is very optimistic and use of a parallel flow LMTD is suggested.

Experimentally measured boiling coefficients and any resulting correlations are dependent on the temperature differences used to calculate these coefficients from the original test data. The same MTD should be used in the design of reboilers when using these data or correlations. However, the generalized correlations given above have such a spread of data that the temperature difference determination is a minor factor in the data spread and the MTD suggestions in the above paragraph should be followed.

9.2.6. Common Problems with Reboilers in Practical Applications

Dryout. If the surfaces are not wetted, it is obvious that no boiling can take place. A nonwetted condition or its equivalent can occur due to: (1) insufficient liquid level; (2) too high a vaporization, i.e., $> 50\,\%$ by weight of vapor where a mist flow regime occurs and the surface dries; 3) too high a temperature difference so that film boiling occurs; and 4) an equivalent effect occurs when some of the low boiler is stripped from the feed, thus increasing the boiling point of the remaining solution so that the temperature difference required for boiling decreases to zero. This phenomenon occurs when the natural circulation path from the reboiler to the distillation column is incomplete, e.g., in vertical tube thermosiphons, the liquid level is at the upper tubesheet but the connecting vapor pipe is above the tubesheet. As a result, the concentrated solution in the thermosiphon can not circulate and mix with the feed solution in the column. The remedy is to increase the liquid level to the midpoint of the horizontal vapor pipe or to sparge a vapor beneath the lowest tubesheet to promote circulation.

Overdesign is very common due to conservative design methods and to fouling allowances. Control of boiling rate then becomes difficult since the required steam pressure may be too low to permit discharging condensate to the condensate return system. The condensate level can build up in the exchanger to blanket the excess surface and may block the vents so that venting the inerts may cease. When the control valves increase the steam pressure sufficiently to discharge the condensate and the inert gases, the increased surface and thus boil-up will result in a cycling effect of the boil-up rate.

Fouling in evaporators can occur by decomposition reactions of the process fluids, corrosion of tube surfaces, or deposition of solids in the feed. Prediction of the fouling rates is impossible due to lack of knowledge of this process, thus a fouling coefficient is just an arbitrarily assumed value. Since the boiling coefficients are very sensitive to its film temperature difference, the effect of even a small temperature drop across the fouling layer has a greatly magnified effect on the heat flux. Being overly conservative in the assumed fouling coefficient can increase the problem of controlling the boil-up as explained for oversized surfacing.

Instability and Its Remedies. Instability results from a feedback between the rate of boiling and the rate of circulation. The phenomenon is mostly encountered in the thermosiphon types of reboilers and is characterized as a surging or oscillating flow. Several theoretical analyses have been made [133], [134], and it usually occurs under certain conditions of boil-up and circulation. Surging is easily suppressed by adding a resistance, a valve or orifice, in the circulation line to the evaporator. Such a resistance has the effect of having a higher resistance in back flow than in forward flow and thus has a stabilizing effect. A similar effect may occur in kettle reboilers where a sudden evolution of vapor forces out the liquid which then later rushes back in when the vapor escapes and the cycle repeats. The only recourse here is to try changing the liquid level otherwise tubes would be removed to allow for easier escape of the vapor.

Vapor Separation for the thermosiphon type of reboilers takes place in the reservoir vessel or more commonly in the base of the distillation column. Here some space is allowed

between the inlet nozzle and the bottom plate to allow for vapor separation. The column vapor velocity will then control the amount of entrainment.

Kettle reboilers with a bundle diameter 60 % of the shell diameter provides a minimum space for vapor separation. Some small amount of entrainment is expected. If a dry vapor is required, additional separating devices should be employed and if the boiling liquid has any tendency to foam or froth, special separators will be required as the normal designed (40 % of shell diameter) vapor space is inadequate. Increasing the number of vapor nozzles in order to reduce the horizontal flow velocities also helps in reducing the entrainment.

For the internal reboiler, considerable space is needed between the liquid level and bottom plate as the vapor – liquid mixture can shoot up several feet above the tube bundle. The height of throw will be a function of bundle diameter and vapor mass velocity.

Others. Location of a fresh feed line to a circulating type reboiler can influence its operation. In refrigeration systems, a mixture of vapor and liquid from the expansion valve is fed to the bottom of the horizontal tube bundle. Feeding close to the low temperature difference end of the tube bundle can improve the circulation and heat transfer at that end. For wide boiling range mixture with a low concentration of low boiler, the circulation may stop if the stripped low boiler can escape without promoting circulation. Here sparging of a vapor beneath the lower tubesheet will promote and maintain a circulation.

Operating across the maximum flux point should not be attempted as crossing this point reverses the effect of control instruments, e.g., an increase in steam pressure will result in less boil-up when in the film boiling zone.

10. Flow-Induced Vibrations

There are two basic types of flow-induced vibrations in heat exchangers: (a) tube vibration, and (b) acoustic vibration (for gas flows only) that can exceed 160 dB. While the reported frequency of such problems with heat exchangers is < 1 % of the installed exchangers, the costs involved to correct such failures justify checking the thermally designed exchanger for potential vibration. The methods to be described below are based on idealized tube bundles and single-phase flows (little information exists for two-phase flow); they only approximate the conditions existing within an exchanger. Better predictions can be made using computer programs that calculate the flow distribution within the shell, but even here some uncertainty exists due to the lack of knowledge of tube stresses formed during fabrication and during operation.

Tube vibration occurs when an exciting forcing frequency is within $\pm$ 20 % of the tube natural vibration frequency and coupling takes place. Tubes may then fail or damage as a result of: (1) Collision with adjacent tubes or shell walls due to a large amplitude of vibration. The collision produces a diamond-shaped wear pattern at the midspans of the tubes and could eventually rupture the tubes. (2) Fretting wear. Because of the clearance between the baffle hole and the tube, the tubes are free to move. Hence, the vibrating tubes can be cut by the baffles if the tubes are softer than the baffle, or the baffle ligament is cut. (3) Cutting of the tubes at the tubesheet due to maximum stress as a result of a lateral deflection. (4) Fatigue. Sustained tube vibration for a long period can produce failure due to fatigue of either the tubes or of the tube-to-tubesheet joints; any tube material defect can propagate and result in tube failure. Oxidation of the tube material and erosion and corrosion can also accelerate fatigue failures.

Acoustic vibration occurs for gas flows when the forcing frequency (flow or heat transfer induced) matches the natural frequency of the shellside gas column. The resulting acoustic resonance is intense and low frequency, and causes severe vibration and fatigue damage to the shell, connecting pipes, and floor. If the frequency of the standing wave coincides with the tube natural frequency, tube failure may occur.

A procedure follows which outlines whether or not there will be a flow-induced vibration problem, and some methods are described to prevent or reduce tube vibration and acoustic vibration. Further details and correlations are given in [135].

10.1. Vibration Analysis

For a thermally designed exchanger, the following is a procedure to determine if there is a vibration problem [135]:

1) Compute the tube natural frequency for straight plain or finned tubes and U-tube bundles [2], [25], [135], [136]. The tube natural frequency should be calculated for tubes in the end regions and window zone, and corrected for axial loads if any.
2) Calculate the crossflow (gap) velocity in the end regions and central region of the tube bundle, and compute the axial velocity in the window zone.
3) Determine the vortex shedding frequency and turbulent buffeting frequency in the end regions and window zone of the tube bundle [137–140]. If either of these frequencies is within 20 % of the tube natural frequency, then resonance is likely to occur. In that case, compute the vibration amplitude, which should be limited to 2 % of the tube outside diameter, *OD,* to avoid damage to the tubes.
4) Estimate the critical velocity for fluid-elastic whirling in the end regions and in the central region of the bundle [141–143]. Compare these velocities with the crossflow velocities computed in step 2 above. If the crossflow velocity is higher than the estimated critical velocity, tube vibration will occur due to fluid elastic whirling. It is currently believed that fluid-elastic whirling, and not vortex shedding or turbulent buffeting, is mostly responsible for tube vibration.
5) Compute the acoustic frequency on the shell side. Acoustic vibration is likely to occur if it is within 20 % of the vortex shedding frequency, turbulent buffeting frequency, or tube natural frequency [135], [144], [145]. Acoustic vibrations can also be generated by the heat-transfer process itself [146–148].

If the above analysis indicates that tube vibration or acoustic vibration is possible, or remedied, the shellside design must be modified as discussed in the next section.

10.2. Design Guidelines for Vibration Prevention

Vibration problems are best prevented rather than corrected. After the heat exchanger has been designed for thermal and hydraulic performance, it should be checked for flow-induced vibrations. Once the sources and frequency of vibration have been identified, either the lowest natural frequency of the tubes should be increased sufficiently so that it is higher than the highest excitation mechanism or the sources of vibration mechanisms should be eliminated or minimized. For acoustic vibrations, detuning baffles can be installed. Further details and discussion are presented in [25], [135].

10.2.1. Tube Vibration

Tube vibration can be reduced within limits by either increasing the tube natural frequency, or reducing or eliminating the crossflow velocity component. There are, however, two designs that have no tube vibration problems: (1) The no-tube-in-window design (Fig. 26), where tubes in the window zone are eliminated and sufficient intermediate support plates are used to prevent tube vibration. (2) The rod or honeycomb type of tube supports (Fig. 27), which stiffen the tube bundle and prevent tube vibration, but here the shell fluid flows parallel to the tubes.

The tube natural frequencies can be increased by reducing the baffle spacing or by adding intermediate partial support plates. Tests have shown that the critical tube rows are those just beyond the baffle tips and that partial support plates (strip baffles) as shown in Figure 48 are effective [149]. Field fixes are lacing the tubes if the tube layout permits it or removing some tubes.

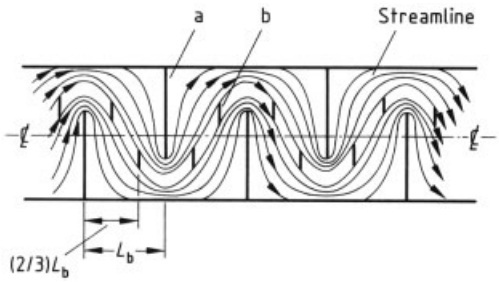

Figure 48. Strip baffles to protect tubes in the window zone a) Segmental baffle; b) Strip baffle

The crossflow velocity, which is the source of the excitation forces, can be reduced by increasing baffle spacing, increasing tube pitch, using multiple segmental baffles, or changing to X or J shells.

10.2.2. Acoustic Vibration

The acoustic vibration can be eliminated by disturbing or destroying the standing wave or changing the character of vortex shedding by one of the following methods:

1) Addition of a detuning baffle close to the pressure node. The detuning baffle should not span the whole tube field in order to minimize the flow disturbance on the shell side.
2) The damping of standing waves could also be accomplished by the use of a porous baffle in place of the solid detuning baffle. While there are a number of detuning baffles required for a tube bundle with the inline arrangement, only one porous baffle is required to detune multiple resonant acoustic modes [150].
3) Selective removal of a few tubes at the displacement antinode of the standing wave will eliminate the vortex shedding and hence the coupling between acoustic frequency and vortex shedding frequency, thus avoiding the acoustic resonance [151].
4) Installation of an acoustic damper such as a tuned Helmholtz cavity resonator on the outside of the shell or acoustic damping material in the exchanger side walls.

None of the above methods of avoiding tube or acoustic vibrations is an ideal solution to the problem, and each will have some disadvantages that apply to the specific exchanger.

Developments of flow-induced vibrations in heat exchangers are dealt with in [152–155].

11. Flow Maldistribution

Heat exchangers are designed on the premise that the flow through the exchanger is steady and uniformly distributed, conditions which almost never exist in industrial heat exchangers.

Flow maldistributions are caused by:

1) Mechanical design, basic geometry and tolerances
2) Self-induced as a result of the heat-transfer process
3) Mixed-phase flows, e.g., gas – liquid
4) Fouling

11.1. Causes of Flow Maldistribution

11.1.1. Mechanical Design, Basic Geometry, and Tolerances

In shell-and-tube exchangers, the design of the entry region (heads, nozzles, manifolds), due to space limitations and cost, results in some flow maldistribution into the tube bundle. Some measurements on large exchangers are reported in [156]. The head design limits the ability to place the tubes close to the shell, thus resulting in a bypass clearance between the tube bundle and shell. This space may or may not be partially closed by the use of side strip baffles. To assemble the exchanger, some clearances are required between the tubes and baffles, and between the baffle and shell and the clearances along with their mechanical tolerances result in leakage areas.

In plate heat exchangers, the manifolds are used to distribute the fluids into and out of the exchanger. Because of the different pressure drops associated with different amounts of fluid flowing axially in the manifolds, flow maldistribution occurs; it becomes more severe as the number of plates is increased per pass.

In compact plate – fin and regenerative surfaces, the flow passages have a small hydraulic diameter. Manufacturing tolerances become very critical in order to maintain uniform flow through highly compact flow passages in laminar flow.

11.1.2. Self-Induced Maldistribution

Flow uniformity can become upset due to heat transfer, which causes viscosity changes, two-phase flows, or density differences. These maldistributions can be either steady state maldistributions or a transient or oscillating flow.

In laminar liquid flow, there can be a large increase in viscosity through the exchanger, such

as cooling of a viscous fluid. As a result, some of the tubes can "freeze" with the high viscosity fluid while other tubes flow with a lower viscosity. This condition can be predicted and avoided [157], [158]. Density changes can also result in maldistribution. Thermoacoustic oscillations occur under certain combinations of flow, geometry, and heat transfer to gases. These vibrations can be sufficiently intense to affect heat transfer, be unpleasantly loud, and can even damage the equipment. Gas flows across tube banks can also result in acoustic and tube vibration problems.

11.1.3. Mixed-Phase Flow

Uniform distribution of a two-phase mixture into a tube bundle (so that each tube gets the same quantity and quality of fluid) is almost impossible. For shellside flow, the possibility of separation due to gravity must be considered especially for horizontal flow. In boiling systems, there are two types of flow maldistributions: (1) Different flow rates can exist between parallel tubes having the same terminal pressures due to the variation of vapor quality in the tubes, sometimes known as the Ledinegg instability; (2) The second type is the more complex oscillating flow and pressure drop known as a density-wave instability (also described as a "steam" or "water hammer").

In condensers, the density-wave instability can also occur under the right conditions. Flow maldistribution can also occur due to the possible accumulation of noncondensable gases in the colder regions of the condenser.

11.1.4. Fouling

Any nonuniformity of fouling within the exchanger affects the local flow rates and will in turn have the feedback effect of increasing the fouling rates.

11.2. Consequences of Maldistribution

11.2.1. Shell-and-Tube Exchangers

Maldistribution in exchangers affects their thermal performance and may damage the exchanger. Thermal performance is most seriously affected when the fluid bypasses the heat-transfer surfaces. Normally flow nonuniformity is not very large and has a small effect on the thermal performance unless the required NTU is > 5 or the secondary fluid conditions result in a temperature pinch. Damage can occur due to the kinetic energy in the liquid slugs of the density-wave oscillation, the vibrations induced by the acoustic phenomenon, or thermal stresses due to large temperature differences in the tube walls in various sections of the bundle, e.g., a horizontal condenser with the lower tubes in a cold condensate pool and the upper tubes in a hot vapor. A fluctuating wall temperature can result in fluctuating stress and metal fatigue, e.g., when hot surface is quenched by waves of fluid.

Systems having radiant, electrical, or nuclear sources of energy are very susceptible to damage due to maldistribution. Here the rate of heat flow is unaffected by the flow or temperature of the fluid. Hence, fluid and/or tube wall temperatures can readily exceed safe limits.

Any flow maldistribution also results in a temperature pattern maldistribution depending on the flow and temperatures of the secondary fluid. Studies indicate that it is this change in the temperature difference that has the greatest effect on the thermal performance rather than variations in the heat-transfer coefficient. In analyzing data taken on exchangers with some maldistribution for comparison with literature correlations, it is very important to correct the temperature difference calculated from the measured terminal temperatures for the effect of internal maldistribution.

Both fluids in an exchanger are subject to maldistribution and the combined effect can result in an excessive thermal stress.

Sixteen papers on the subject of flow maldistribution are contained in [159]. More detailed information into the rather extensive literature on this subject, especially that dealing with instabilities, is to be found in [160] which contains 102 references.

11.2.2. Other Exchangers

Manifold-induced flow maldistribution can be significant and can have a significant influence

on reduction of heat transfer. For further information, refer to [161–165]. Guidelines for the design of manifolds are summarized in [27].

Gross maldistribution at the core inlet is a nonuniformity on a macroscopic level. It is a result of either inlet header shape, upstream flow conditions, or a gross blockage in the heat exchanger core (plugged tubes or flow passages); this may be due to brazing, soldering, or other manufacturing considerations. This nonuniformity causes a significant increase in pressure drop with some reduction in heat-transfer rate [27].

In highly compact heat exchangers, the small flow passages are not identical as a result of manufacturing tolerances. Differently sized and shaped passages exhibit different flow resistances resulting in nonuniform flow distribution, even when the flow approaching the core face is uniform. Thus, this passage-to-passage maldistribution occurs within a core on a microscopic level. It reduces heat transfer significantly, and also reduces the pressure drop slightly [27].

11.3. Recommendations for Shell-and-Tube Exchangers

Normally the maldistribution due to entrance flow does not have a large (usually $< 10\,\%$) effect on thermal performance; however, an impingement baffle is recommended to improve distribution. The internal bypassing and leakages are to some extent allowed for in the stream analysis methods or the Delaware method of calculation. However, an additional allowance should be made for the bypass effect on temperature differences using either the Fisher and Parker curves [107] or the semiempirical methods in some proprietary computer programs.

Any laminar flow exchanger having a large viscosity increase should be checked for maldistribution [157], [158].

The instabilities in vaporizers can be controlled by restricting the liquid circulation. In condensers, most problems are resolved through good venting and good condensate drainage.

The acoustical vibration problems are usually cured by proper placement of detuning baffles.

12. Fouling

Fouling occurs over a period of time with the deposition of solids on heat-transfer surfaces. This solid layer adds an additional resistance to heat flow and can also increase the resistance to fluid flow. The accumulation of this layer can be classified as:

Precipitation Fouling. This is the deposition of dissolved substances on the surface when the surface concentration exceeds the solubility limit. If the deposited layer is hard and tenacious, it is often referred to as "scaling". If it is porous and mushy, it is called "sludge".

Particulate Fouling. Here finely divided solid particles settle on the surface as a result of gravity or other attractive forces. If settling occurs due to gravity, it is referred to as *sedimentation fouling.*

Chemical Reaction Fouling. In this mechanism, solids are formed as a result of reactions, such as coking or polymerization, at the heat-transfer surface in which the surface material does not participate in the chemical reaction.

Corrosion Fouling. The heat-transfer surface reacts with the fluids and the resulting corrosion products form an insulating layer.

Biological Fouling. Various organisms (micro and/or macro) attach themselves to the surface and grow in place.

Solidification Fouling. This results from the freezing of the fluid or a component of the fluid on a subcooled surface.

Very often several of the above classes of fouling occur simultaneously and frequently in a synergistic manner.

Fouling is very costly since it (1) increases capital costs due to the need to oversurface the heat exchanger and for cleaning; (2) increases maintenance costs resulting from cleaning, chemical additives, or troubleshooting; (3) results in loss of production due to shutdown or reduced capacity; and (4) increases energy losses due to reduced heat transfer, increased pressure drop, and dumping of dirty gas streams it present. Fouling promotes corrosion, plugging, and failure of uncleaned exchangers. Gas-side fouling can also be a potential fire hazard.

Fouling significantly reduces heat transfer with a relatively small increase in pressure drop in systems with liquid flows and high heat-transfer coefficients. For systems having low heat-

Table 22. Influence of operating variables on liquid-side fouling

Variables	Precipitation	Solidificiation	Particulate	Chemical reaction	Corrosion	Biological
Temperature	↑↓	↓	↑↓–	↑↓	↑↓	↑↓–
Velocity	↓–	↑↓	↓	↓	↑↓–	↑↓
Supersaturation level	↑	↑				
pH	↑		↑↓		↑↓	↑↓
Impurities		↓				
Concentration	↑	↑	↑			
Surface roughness	↑	↑	↑–		↑–	↑
Pressure	–	–		↑	↑	↑↓
Oxygen	–	–		↑	↑	↑↓

transfer coefficients such as with gases, fouling increases the pressure drop significantly with some reduction in the heat transfer [27].

The rate of fouling is a complex function of time and can be linear, falling rate, or asymptotic depending on the mechanism of formation and solid removal, if any, by an erosion or spalling mechanism. Often there may be an induction period before fouling begins. Many factors affect fouling rates such as temperature, fluid velocity, concentration, surface material and its condition, in addition to the mechanisms influencing several classes of fouling involved. The qualitative effect of some of the operating variables on various fouling mechanisms is shown in Table 22 for liquid-side fouling and in Table 23 for gas-side fouling. In these tables, when the value of a variable is increased, it increases (arrow going up), decreases (arrow going down), or has no effect (a dash) on the specific fouling mechanism listed. Blank blocks indicate that no influence of these variables has been investigated and reported in the literature.

Table 23. Influence of operating variables on gas-side fouling

Variables	Solidification	Particulate	Chemical reaction	Corrosion
Temperature	↓	↓↑	↑	↑↓–
Velocity	↓	↓↑–	↑↓–	↑–
Impurities	↓			
Concentration	↑	↑		↑
Fuel – air ratio		↑	↑	
Surface roughness		↑–		↑–
Oxygen	–	–	↑	
Sulfur			↑	↑

Some progress in understanding the nature of fouling has taken place during the past twenty years as documented in [166–170], but except for cooling waters, very little has been quantitatively correlated for the design of heat exchangers. Much of the fouling research needs to be confirmed by plant operation before being acceptable for design use. In addition, these prediction methods require much more information concerning the process and the exchanger environment than is usually available to the designer.

Present practice is to arbitrarily select a fouling coefficient (or its inverse, the fouling resistance) based on TEMA tables, or if available, based on plant experience for similar process fluids. This is equivalent to specifying an unknown variable coefficient as a fixed constant in Equation (7). Ideally the fouling coefficient selected should involve economic factors such as the cost of excess (over the clean theoretical) area, cost of cleaning, lost production due to cleaning, etc. It is not unusual to encounter designs where the assumed fouling resistance is greater than the process fluid resistance that was determined by detailed and lengthy calculations.

The TEMA tables were developed in the early 1940s by a consensus of opinions and not on any documented test data. The original tables have been slightly revised and extended and owe their continued existance to the fact that exchangers designed and based on these values may not represent the actual fouling but may cover other errors, such as the effect of maldistribution on the effective temperature difference that were not recognized when the tables were developed. Until better analysis of the fouling problem is developed, the fouling factors should be selected from Tables 24 and 25, which are abstracted from the TEMA tables [2].

Table 24. TEMA recommended fouling factors for waters[2]

Types of water	Fouling factor $\times 10^{-4}$ m^2 K/W			
	Heating-Medium Temp. $\leq 115\,^\circ$C Water Temp. $\leq 50\,^\circ$C		115 – 205 °C $> 50\,^\circ$C	
	Velocity ≤ 0.9 m/s	> 0.9 m/s	Velocity ≤ 0.9 m/s	> 0.9 m/s
Seawater	0.88	0.88	1.76	1.76
Brackish water	3.52	1.76	5.28	3.52
Cooling tower				
Treated makeup	1.76	1.76	3.52	3.52
Untreated	5.28	5.28	8.81	7.04
City or well water	1.76	1.76	3.52	3.52
River water:				
Minimum	3.52	1.76	5.28	3.52
Average	5.28	3.52	7.04	5.28
Muddy or silty	5.28	3.52	7.04	5.28
Hard ($>$ 15 grains/gal) (= 257 g/m^3)	5.28	5.28	8.81	8.81
Engine jacket	1.76	1.76	1.76	1.76
Distilled or closed-cycle condensate	0.88	0.88	0.88	0.88
Treated boiler feedwater	1.76	0.88	1.76	1.76
Boiler blowdown	3.52	3.52	3.52	3.52

Table 25. TEMA recommended fouling factors for fluids [2]

Fluid	Fouling factor $\times 10^{-4}$ m$^2 \cdot$ K/W
Industrial Fluids	
Fuel oil	8.81
Tranformer oil	1.76
Engine lube oil	1.76
Quench oil	7.04
Refrigerant liquids	1.76
Hydraulic fluid	1.76
Industrial organic heat-transfer media	1.76
Molten heat-transfer salts	8.81
Chemical Processing streams	
MEA and DEA solutions	3.52
DEG and TEG solutions	3.52
Caustic solutions	3.52
Vegetable oils	5.28
Natural gas-gasoline processing streams	
Lean oil	3.52
Rich oil	1.76
Natural gasoline and liquified petroleum gases	1.76
Crude and vacuum liquids	
Gasoline	1.76
Naptha and light distillates	1.76
Kerosene	1.76
Light gas oil	3.52
Heavy gas oil	5.28
Heavy fuel oils	8.81
Asphalt and residuum	1.76

Table 25. (Continued)

Fluid	Fouling factor $\times 10^{-4}$ m$^2 \cdot$ K/W
Cracking and coking unit streams	
Light crude oil	3.52
Heavy cycle oil	5.28
Light coker gas oil	5.28
Heavy coker gas oil	7.04
Light liquid products	3.52
Light ends and lube oil processing streams	
Liquid products	1.76
Absorption oils	3.52
Alkylation trace acid streams	3.52
Reboiler streams	5.28
Feed stock	3.52
Solvent	1.76
Wax slurries	5.28
Conversion factors	1 m$^2 \cdot$ K/W = 5.678 hr $\cdot$ ft$^2 \cdot\,^\circ$F/Btu 1 m/s = 3.281 ft/s $T_F = 1.8\,(T_C) + 32$

13. References

1. R. K. Shah in S. Kakaç, A. E. Bergles, F. Mayinger (eds.): "Classification of Heat Exchangers," *Heat Exchangers:*

Thermal-Hydraulic Fundamentals and Design, Hemisphere Publishing Co., New York 1981, pp. 9 – 46.
2. Tubular Exchanger Manufacturers Association (TEMA): *Standards of TEMA,* 6th ed., New York 1978.
3. R. K. Shah, W. W. Focke in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Plate Heat Exchangers and Their Design Theory," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 227 – 254.
4. S. G. Kandlikar, R. K. Shah: "Multipass Plate Heat Exchangers–Effectiveness-NTU Results and Guidelines for Selecting Pass Arrangements."
5. C. T. Cowan: "Choosing Materials of Construction for Plate Heat Exchangers," *Chem. Eng. (N.Y.)* **82** (1975) June 9, 101 – 103; **82** (1975) July 7, 102 – 104.
6. R. K. Shah, A. D. Giovannelli in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Heat Pipe Heat Exchanger Design Theory," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 609 – 653.
7. D. Butterworth in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Condensers and Their Design," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 779 – 828.
8. K. P. Singh in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Power Plant Surface Condenser Technology–An Overview," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 513 – 526.
9. K. P. Singh in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Closed Feedwater Heaters–An Overview," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 595 – 608.
10. A. C. Mueller in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Air Cooled Heat Exchangers," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 179 – 189.
11. V. P. Carey, R. K. Shah in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Design of Compact and Enhanced Heat Exchangers for Liquid – Vapor Phase-Change Applications," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 909 – 968.
12. S. M. Cho in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Nuclear Heat Exchangers," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 325 – 336.
13. J. G. Collier in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Nuclear Steam Generators and Waste Heat Boilers," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 659 – 682.
14. P. E. Minton: *Handbook of Evaporation Technology,* Noyes Data Corp., Park Ridge, NJ, 1986.
15. J. G. Collier in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Evaporators," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 683 – 705.
16. R. A. Smith in E. U. Schlünder (ed.): "Evaporators," *Heat Exchanger Design Handbook,* Section 3.5, Hemisphere Publishing Co., New York 1983, pp. 3.5.3-1,2.
17. APV Equipment and System for Food and Beverage Processing, Bulletin FP-178, APV Co., Tonawanda, New York 1980.
18. D. Q. Kern: *Process Heat Transfer,* McGraw-Hill, New York 1950.
19. A. E. Bergles, J. G. Collier, J. M. Delhaye, G. F. Hewitt, F. Mayinger: *Two-Phase Flow and Heat Transfer in the Power and Process Industries,* Hemisphere Publishing Co., New York 1981.
20. J. W. Palen in E. U. Schlünder (ed.): "Shell-and-Tube Reboilers," *Heat Exchanger Design Handbook,* Section 3.6, Hemisphere Publishing Co., New York 1983.
21. V. L. Eriksen: "Waste Heat Recovery Exchangers," in *Compact Heat Exchangers–History, Technological Development and Mechanical Design Problems,* Book no. G00183, HTD-vol. 10, ASME, New York 1980, pp. 181 – 185.
22. D. Csathy: "Latest Practice in Industrial Heat Recovery," presented at *The Energy-Source Technology Conference and Exhibition,* New Orleans, Feb. 4, 1980.
23. P. Hinchley: "Specifying and Operating Reliable Waste-Heat Boilers," *Chem. Eng.* **86** (1979) Aug. 13, 120 – 134.
24. R. K. Shah in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Heat Exchanger Design Methodology," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 17 – 22.

25. K. P. Singh, A. I. Soler: *Mechanical Design of Heat Exchangers and Pressure Vessel Components,* Arcturus Publishers, Cherry Hill, N.J., 1984.
26. A. C. Mueller in W. M. Rohsenow, J. P. Hartnett, E. N. Ganić (eds.): "Process Heat Exchangers," *Handbook of Heat Transfer Applications,* 2nd ed., **Chap. 4, Part 2,** McGraw-Hill, New York 1985.
27. R. K. Shah in W. M. Rohsenow, J. P. Hartnett, E. N. Ganić (eds.): "Compact Heat Exchangers," *Handbook of Heat Transfer Applications,* 2nd ed., **Chap. 4, Part 3,** McGraw-Hill, New York 1985.
28. R. K. Shah, A. C. Mueller in W. M. Rohsenow, J. P. Hartnett, E. N. Ganić (eds.): "Heat Exchanger Basic Thermal Design Methods," *Handbook of Heat Transfer Applications,* 2nd ed., Chap. 4, Part 1, McGraw-Hill, New York 1985.
29. R. K. Shah in S. Kakaç, R. K. Shah, A. E. Bergles (eds.): "Heat Exchanger Basic Design Methods," *Low Reynolds Number Flow Heat Exchangers,* Hemisphere Publishing Co., New York 1983, pp. 21 – 72.
30. J. Taborek in E. U. Schlünder (ed.): "Charts for Mean Temperature Difference in Industrial Heat Exchanger Configurations," *Heat Exchanger Design Handbook,* Section 1.5, Hemisphere Publishing Co., New York 1983.
31. R. K. Shah, S. G. Kandlikar: "The Influence of the Number of Thermal Plates on Plate Heat Exchanger Performance," *Current Researches in Heat and Mass Transfer, A Compendium and a Festschrift for Professor Arcot Ramachandran,* Hemisphere Publishing Co., New York 1988, pp. 227 – 241.
32. S. G. Kandlikar, R. K. Shah: "Asymptotic Effectiveness–NTU Formulas for Multipass Plate Heat Exchangers," *J. Heat Transfer* **109** (1989).
33. P. G. Kroger: "Performance Deterioration in High Effectiveness Heat Exchangers Due to Axial Heat Conduction Effects," *Adv. Cryog. Eng.* **12** (1967) 363 – 372.
34. J. P. Chiou: *The Advancement of Compact Heat Exchanger Theory Considering the Effects of Longitudinal Heat Conduction and Flow Nonuniformity,* ASME, Book No. G00183, HTD-vol. 10, New York 1980, pp. 101 – 121.
35. A. P. Colburn: "Mean Temperature Difference and Heat Transfer Coefficient in Liquid Heat Exchangers," *Ind. Eng. Chem.* **25** (1933) 873 – 877.
36. E. N. Sieder, G. E. Tate: "Heat Transfer and Pressure Drop of Liquids in Tubes," *Ind. Eng. Chem.* **28** (1936) 1429 – 1435.
37. K. Gardner, J. Taborek: "Mean Temperature Difference: A Reappraisal," *AIChE J.* **23** (1977) 777 – 786.
38. W. Roetzel: "Heat Exchanger Design With Variable Transfer Coefficients for Crossflow and Mixed Flow Arrangements," *Int. J. Heat Mass Transfer* **17** (1974) 1037 – 1049.
39. W. M. Kays, A. L. London: *Compact Heat Exchangers,* 3rd ed., McGraw-Hill, New York 1984.
40. R. K. Shah, A. L. London: "Laminar Flow Forced Convection" in *Ducts, Supplement 1 to Advances in Heat Transfer,* Academic Press, New York 1978.
41. R. K. Shah, M. S. Bhatti in S. Kakaç, R. K. Shah, W. Aung (eds.): "Laminar Convective Heat Transfer in Ducts," *Handbook of Single-Phase Convective Heat Transfer,* **Chap. 3,** J. Wiley & Sons, New York 1987.
42. M. S. Bhatti, R. K. Shah in S. Kakaç, R. K. Shah, W. Aung (eds.): "Turbulent and Transition Flow Convective Heat Transfer in Ducts," *Handbook of Single-Phase Convective Heat Transfer,* **Chap. 4,** J. Wiley & Sons, New York 1987.
43. R. K. Shah, M. S. Bhatti in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Assessment of Correlations for Single-Phase Heat Exchangers," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 81 – 122.
44. A. Zukauskas in S. Kakaç, R. K. Shah, W. Aung (eds.): "Convective Heat Transfer in Cross Flow," *Handbook of Single-Phase Convective Heat Transfer,* **Chap. 6,** J. Wiley & Sons, New York 1987.
45. D. E. Briggs, E. H. Young: "Convection Heat Transfer and Pressure Drop of Air Flowing Across Triangular Pitch Banks of Finned Tubes," *Chem. Eng. Prog. Symp. Ser.* **59** (1963) no. 41, 1 – 10.
46. K. K. Robinson, D. E. Briggs: "Pressure Drop of Air Flowing Across Triangular Pitch Banks of Finned Tubes," *Chem. Eng. Prog. Symp. Ser.* **62** (1966) no. 64, 177 – 184.
47. T. J. Rabas, J. Taborek: "Survey of Turbulent Forced-Convection Heat Transfer and Pressure Drop Characteristics of Low-Finned Tube Banks in Cross Flow," *Heat Transfer Eng.* **8** (1987) no. 2, 49 – 62.

48. A. Ganguli, S. B. Yilmaz: "New Heat Transfer and Pressure Drop Correlations for Crossflow over Low-Finned Tube Banks," *AIChE Symp. Ser.* **83** (1987) no. 257, 9 – 14.
49. H. C. Chai: "A Simple Pressure Drop Correlation Equation for Low Finned Tube Crossflow Heat Exchangers," *Int. Commun. Heat Mass Transfer* **15** (1988) 95 – 101.
50. R. L. Webb: "Air-Side Heat Transfer in Finned Tube Heat Exchangers," *Heat Transfer Eng.* **1** (1980) no. 3, 33 – 49.
51. D. L. Gray, R. L. Webb: "Heat Transfer and Friction Correlations for Plate Fin-and-Tube Heat Exchangers Having Plain Fins," *Heat Transfer 1986* **6** (1986) 2745 – 2750.
52. K. S. N. Raju, J. C. Bansal in S. Kakaç, R. K. Shah, A. E. Bergles (eds.): "Design of Plate Heat Exchangers," *Low Reynolds Number Flow Heat Exchangers,* Hemisphere Publishing Co., New York 1983, pp. 913 – 932.
53. J. Marriott: "Where and How to Use Plate Heat Exchangers," *Chem. Eng. N.Y.* **78** (1971) no. 8, 127 – 134.
54. M. P. Bond: "Plate Heat Exchangers for Effective Heat Transfer," *Chem. Eng. (London)* **367** (April 1981) 162 – 166.
55. B. W. Jackson, R. A. Troupe: "Laminar Flow in a Plate Heat Exchanger," *Chem. Eng. Prog.* **60** (1964) no. 7, 62 – 65.
56. W. W. Focke, J. Zachariades, I. Olivier: "The Effect of the Corrugation Inclination Angle on the Thermohydraulic Performance of Plate Heat Exchangers," *Int. J. Heat Mass Transfer* **28** (1985) 1469 – 1479.
57. R. L. Webb in S. Kakaç, R. K. Shah, W. Aung (eds.): "Enhancement of Single-Phase Heat Transfer," *Handbook of Single-Phase Convective Heat Transfer,* Chap. 17, J. Wiley & Sons, New York 1987.
58. H. M. Joshi, R. L. Webb: "Heat Transfer and Friction in the Offset Strip Fin Heat Exchanger," *Int. J. Heat Mass Transfer* **30** (1987) 69 – 84.
59. R. K. Shah, R. L. Webb in J. Taborek, G. F. Hewitt, N. Afgan (eds.): "Compact and Enhanced Heat Exchangers," *Heat Exchangers: Theory and Practice,* Hemisphere Publishing Co., New York 1983, pp. 425 – 468.
60. J. G. Collier: *Convective Boiling and Condensation,* 2nd ed., McGraw-Hill, New York 1981.
61. M. K. Jensen in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Boiling Heat Transfer and Two-Phase Flow in Tubes and Tube Bundles," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 413 – 438.
62. S. Z. Rouhani, M. S. Sohal: "Two-Phase Flow Patterns: A Review of Research Results," *Prog. Nucl. Energy* **11** (1983) 219 – 259.
63. D. Chisholm: *Two-Phase Flow in Pipelines and Heat Exchangers,* George Godwin Publishing Co., London 1983.
64. Y. Taitel, A. E. Dukler: "A Model for Predicting Flow Regime Transitions in Horizontal and Near Horizontal Gas-Liquid Flow," *AIChE J.* **22** (1976) 47 – 55.
65. Y. Taitel, D. Bornea, A. E. Dukler: "Modelling Flow Pattern Transitions for Steady Upward Gas-Liquid in Vertical Tubes," *AIChE J.* **26** (1980) 345 – 354.
66. J. Weisman, D. Duncan, J. Gibson, T. Crawford: "Effects of Fluid Properties and Pipe Diameters on Two-Phase Flow Patterns in Horizontal Tubes," *Int. J. Multiphase Flow* **5** (1979) 437 – 462.
67. J. Weisman, S. Y. Kang: "Flow Pattern Transitions in Vertical and Upwardly Inclined Tubes," *Int. J. Multiphase Flow* **7** (1981) 271 – 291.
68. I. D. R. Grant, D. Chisholm: "Two-Phase Flow on the Shell-Side of a Segmentally Baffled Shell-and-Tube Heat Exchanger," *J. Heat Transfer* **101** (1979) 38 – 42.
69. D. Butterworth: "A Comparison of Some Void-Fraction Relationships for Co-Current Gas-Liquid Flow," *Int. J. Multiphase Flow* **1** (1975) 845 – 850.
70. W. Idsinga, N. Todreas, R. Bowring: "An Assessment of Two-Phase Pressure Drop Correlations for Steam-Water Systems," *Int. J. Multiphase Flow* **3** (1977) 401 – 413.
71. D. G. Reddy, C. F. Fighetti, M. Merilo: "Evaluation of Two-Phase Pressure Drop Correlations for High Pressure Steam-Water Systems," Proceedings of ASME-JSME Thermal Engineering Joint Conference 1 (1983), ASME, New York, pp. 251 – 259.
72. L. Friedel: "Improved Friction Pressure Drop Correlations for Horizontal and Vertical Two-Phase Pipe Flow," *European Two-Phase Flow Group Meeting,* Ispra, Italy, Paper 2.E (1979).
73. G. F. Hewitt in G. Hetsroni (ed.): "Gas-Liquid Systems," *Handbook of Multiphase Systems,* Hemisphere Publishing Corp., New York 1982, pp. 2-1 –2-136.
74. D. Chisholm: "Pressure Gradients Due to Friction During the Flow of Evaporating

Two-Phase Mixtures in Smooth Tubes and Channels," *Int. J. Heat Mass Transfer* **16** (1973) 347 – 358.
75. R. W. Lockhart, R. C. Martinelli: "Proposed Correlations of Data for Isothermal Two-Phase Two-Component Flow in Pipes," *Chem. Eng. Prog.* **45** (1949) 39 – 48.
76. R. G. Sardesai, R. G. Owen, D. J. Pulling: "Pressure Drop for Condensation of a Pure Vapour in Downflow in a Vertical Tube," *Heat Transfer 1982* **5** (1982) 139 – 145.
77. M. K. Jensen in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Boiling on the Shellside of Horizontal Tube Bundles," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 707 – 746.
78. I. D. R. Grant, D. Chisholm: "Horizontal Two-Phase Flow Across Tube Banks," *Int. J. Heat Fluid Flow* **2** (1980) no. 2, 97 – 100.
79. D. S. Scharge, J.-T. Hsu, M. K. Jensen: "Void Fractions and Two-Phase Friction Multipliers in a Horizontal Tube Bundle," *AIChE Symp. Ser.* **83** (1987) no. 257, 1 – 8.
80. K. Ishihara, J. W. Palen, J. Taborek: "Critical Review of Correlations for Predicting Two-Phase Flow Pressure Drop Across Tube Banks," *Heat Transfer Eng.* **1** (1980) no. 3, 23 – 32.
81. A. C. Mueller in E. U. Schlünder (ed.): "Condensers," *Heat Exchanger Design Handbook,* Section 3.4, Hemisphere Publishing Co., New York 1983.
82. W. M. Rohsenow in W. M. Rohsenow, J. P. Hartnett, E. N. Ganić (eds.): "Film Condensation," *Handbook of Heat Transfer Fundamentals,* 2nd ed., Chap. 11, Part 1, McGraw-Hill, New York 1985.
83. P. J. Marto in S. Kakaç, A. E. Bergles, E. O. Fernandes (eds.): "Fundamentals of Condensation," *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Kluwer Academic Publ., Dordrecht, The Netherlands, 1988, pp. 221 –291.
84. G. Breber in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Intube Condensation," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 451 – 473.
85. G. Breber, J. W. Palen, J. Taborek: "Prediction of Horizontal Tubeside Condensation of Pure Components Using Flow Regime Criteria," *J. Heat Transfer* **102** (1980) 471 – 476.
86. M. M. Shah: "A General Correlation for Heat Transfer During Film Condensation Inside Pipes," *Int. J. Heat Mass Transfer* **22** (1979) 547 – 556.
87. G. B. Wallis: "Flooding Velocities for Air and Water in Vertical Tubes," UKAEA Report No. AEEWR123 (1961).
88. D. Butterworth in E. U. Schlünder (ed.): "Film Condensation of Pure Vapor," *Heat Exchanger Design Handbook,* Section 2.6.2, Hemisphere Publishing Co., New York 1983.
89. J. Taborek in N. Afgan, E. U. Schlünder (eds.): "Design Method for Heat Transfer Equipment–A Critical Survey of the State-of-the-Art," *in Heat Exchangers: Design and Theory Source Book,* Hemisphere Publishing Co., New York 1974, pp. 45 – 74.
90. K. O. Beatty, D. L. Katz: "Condensation of Vapors on Outside Finned Tubes," *Chem. Eng. Prog.* **44** (1948) 55 – 70.
91. M. K. Jensen in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Condensation with Noncondensables and in Multicomponent Mixtures," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 497 – 512.
92. W. M. Rohsenow in W. M. Rohsenow, J. P. Hartnett, E. N. Ganić (eds.): "Boiling," *Handbook of Heat Transfer Fundamentals,* 2nd ed., **Chap. 12,** McGraw-Hill, New York 1985.
93. J. G. Collier in E. U. Schlünder (ed.): "Boiling and Evaporation," *Heat Exchanger Design Handbook,* Section 2.7., Hemisphere Publishing Co., New York 1983.
94. J. C. Chen: "A Correlation for Boiling Heat Transfer to Saturated Fluids in Convective Flow," *Ind. Eng. Chem. Process Des. Dev.* **5** (1966) 322 – 329.
95. S. G. Kandlikar: "A General Correlation for Saturated Two-Phase Flow Boiling Heat Transfer Inside Horizontal and Vertical Tubes," HTD-vol. 85, ASME, New York 1987, pp. 9 – 19; a revised version to be published in *J. Heat Transfer.*
96. S. G. Kandlikar: *An Improved Correlation for Predicting Two-Phase Flow Boiling Heat Transfer Coefficient in Horizontal and Vertical Tubes,* Book no. H00271, HTD-vol. 27, ASME, New York 1983, pp. 3 – 10.
97. M. M. Shah: "Improved General Correlation for Critical Heat Flux During Upflow in Uniformly Heated Vertical Tubes," *Heat Fluid Flow* **8** (1987) no. 4, 326 – 335.

98. M. K. Jensen, J.-T. Hsu: "A Parametric Study of Boiling Heat Transfer in a Tube Bundle," Preceedings of the 1987 ASME-JSME Thermal Engineering Joint Conference 3 (1987), ASME, New York, pp. 133 – 139.
99. J. W. Palen, W. M. Small: "New Way to Design Kettle and Internal Reboilers," *Hydrocarbon Process.* **43** (1963) no. 11, 199 – 208.
100. G. Breber in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Computer Programs for Design of Heat Exchangers," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 167 – 177.
101. R. K. Shah in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Plate-Fin and Tube-Fin Heat Exchanger Design Procedures," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 255 – 266.
102. R. K. Shah in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Counterflow Rotary Regenerator Thermal Design Procedures," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 267 – 296.
103. T. Tinker: *Shell Side Heat Transfer Characteristics of Segmentally Baffled Shell and Tube Exchangers,* ASME Paper No. 47-A-130, 1947.
104. J. W. Palen, J. Taborek: "Solution to Shell Side Flow Pressure Drop and Heat Transfer by Stream Analysis Method," *AIChE Symp. Ser.* **65** (1969) no. 92, 53 – 63.
105. J. Taborek in E. U. Schlünder (ed.): "Shell-and-Tube Heat Exchangers: Single-Phase Flow," *Heat Exchanger Design Handbook,* Section 3.3., Hemisphere Publishing Co., New York 1983.
106. K. J. Bell in S. Kakaç, A. E. Bergles, F. Mayinger (eds.): "Delaware Method for Shell Side Design," *Heat Exchangers: Thermal-Hydraulic Fundamentals and Design,* Hemisphere Publishing Co., New York 1981, pp. 581 – 618.
107. J. Fisher, R. O. Parker: "New Ideas on Heat Exchanger Design," *Hydrocarbon Process.* **49** (1969) no. 7, 147 – 154.
108. B. Slipcevic: "Designing Heat Exchangers With Disk and Ring Baffles," *Sulzer Tech. Rev.* **58** (1976) no. 3, 114 – 120.
109. B. Slipcevic: "Shell-Side Pressure Drop in Shell and Tube Heat Exchangers With Disk and Ring Baffles," *Sulzer Tech. Rev.* **60** (1978) no. 1, 28 – 30.
110. M. A. Founti, J. H. Whitelaw: *Shell-side Flow in a Model Disc-and-Doughnut Heat Exchanger,* AERE-R 10831, United Kingdom Atomic Energy Authority, Harwell 1983.
111. K. P. Goyal, B. K. Gupta: "An Experimental Performance Evaluation of a Disk and Doughnut Type Heat Exchanger," *J. Heat Transfer* **106** (1984) 759 –765.
112. F. Ratnasamy: "Exchanger Design Using Disc and Donut Baffles," *Hydrocarbon Process.* **66** (1987) no. 4, 63 – 65.
113. C. C. Gentry, R. K. Young, W. K. Small: "RODbaffle Heat Exchanger Thermal-Hydraulic Predictive Methods for Bare and Low-Finned Tubes," *AIChE Symp. Ser.* **80** (1984) no. 236, 104 – 109.
114. R. A. Khan: "Effect of Vent Rate, Noncondensable Loading and Vent Location on Heat Transfer in a Large Vertical Tube Bundle of a Sea Water Evaporator," *Symp. on Special Problems in Process Heat Transfer,* Paper 14G, AIChE 71st Natl. Mtg., Dallas, Texas, Feb. 1972.
115. "Steam Turbine Condensers," NEL Report 619, East Kilbride, Glasgow, Aug. 1976.
116. T. J. Rabas, A. C. Mueller: "The Effect of Noncondensable Gases and Vent Flow Rate on the Thermal Performance of Single Pass "X" Shell Condensers," *Heat Transfer Eng.* **6** (1986) 2703 – 2708.
117. E. R. Kent, R. L. Pigford: "Fractionation During Condensation of Vapor Mixtures," *AIChE J.* **2** (1956) 363 – 371.
118. E. Toster: "Berechnung der Trennwirkung von Teilkondensatoren," *Chem. Ing. Tech.* **22** (1960) 525 – 531.
119. K. G. English, W. T. Jones, R. C. Spillers, V. Orr: "Criteria of Flooding and Flooding Correlation Studies With a Vertical Updraft Partial Condenser," *AIChE Preprint 9,* 6th Natl. Heat Transfer Conf., Boston, 1963.
120. R. C. Holmes in J. H. Perry (ed.): *Chemical Engineers Handbook,* 3rd ed., McGraw-Hill, New York 1950, p. 686.
121. K. J. Bell in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Design of Reflux Condensers," *Heat Transfer Equipment Design,* Hemisphere Publishing Co., New York 1988, pp. 527 –534.
122. W. F. Berg, J. L. Berg: "Flow Patterns for Isothermal Condensation in One-Pass Air-Cooled Heat Exchangers," *Heat Transfer Eng.* **1** (1980) no. 4, 21 –31.
123. G. Breber in R. K. Shah, E. C. Subbarao, R. A. Mashelkar (eds.): "Design Methods that Reduce Backflow and Inert-Blanketing in Crossflow Condensers," *Heat Transfer*

Equipment Design, Hemisphere Publishing Co., New York 1988, pp. 535 – 546.
124. A. P. Colburn, O. A. Hougen: "Design of Cooler Condensers for Mixtures of Vapors with Noncondensing Gases," *Ind. Eng. Chem.* **26** (1934) 1178 – 1182.
125. A. P. Colburn, A. G. Edison: "Prevention of Fog in Cooler Condensers," *Ind. Eng. Chem.* **33** (1941) 457 – 458.
126. A. P. Colburn, T. B. Drew: "The Condensation of Mixed Vapors," *AIChE Trans.* **33** (1937) 197 – 215.
127. K. J. Bell, M. A. Ghaly: "An Approximate Generalized Design Method for Multicomponent/Partial Condensers," *AIChE Symp. Ser.* **69** (1972) no. 131, 72 – 79.
128. B. C. Price, K. J. Bell: "Design of Binary Vapor Condensers Using the Colburn-Drew Equations," *AIChE Symp. Ser.* **70** (1974) no. 138, 163 – 171.
129. J. W. Palen, C. C. Yang: "Circulation Boiling Model for Analysis of Kettle and Internal Reboiler Performance," in *Heat Exchangers for Two-Phase Applications,* Book no. H00271, HTD-vol. 27, ASME, New York 1983, pp. 55 – 61.
130. J. W. Palen, A. Yarden, J. Taborek: "Characteristics of Boiling Outside Large-Scale Horizontal Multitube Bundles," *AIChE Symp. Ser.* **68** (1972) no. 118, 50 – 61.
131. J. R. Fair: "What You Need to Know to Design Thermosiphon Reboilers," *Pet. Refiner* **39** (1960) no. 2, 105 – 123.
132. M. R. Glickstein, R. H. Whitesides: *Forced Convection Nucleate and Film Boiling of Several Aliphatic Hydrocarbons,* ASME paper 67-HT-7, 1967.
133. J. Boure, A. E. Bergles, L. S. Tong: "Review of Two-Phase Flow Instability," *Nucl. Eng. Des.* **25** (1973) 165 – 192.
134. A. Blumenkrantz, J. Taborek: "Application of Stability Analysis for Design of Natural Circulation Boiling Systems and Comparison With Experimental Data", *AIChE Symp. Ser.* **68** (1972) no. 118, 136 – 146.
135. R. K. Shah: *Flow-Induced Vibration and Noise in Heat Exchangers,* Proceedings of the Seventh National Heat and Mass Transfer Conference, Kharagpur, India, Paper No. HMT-A-32-83, 1983, pp. 89 – 108.
136. U. Bolleter, R. D. Blevins: "Natural Frequencies of Finned Heat Exchanger Tubes," *J. Sound Vibration* **80** (1982) 367 – 371.
137. J. M. Chenoweth, J. H. Kissel: "Flow-Induced Tube Vibration in Shell-and-Tube Heat Exchangers, A Short Course," *AIChE Today Series,* AIChE, New York 1979.
138. P. M. Moretti: "Flow-Induced Vibrations in Heat Exchangers," Workshop Notes, School of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, Oklahoma, October 1982.
139. P. R. Owen: "Buffeting Excitation of Boiler Tube Vibration," *J. Mech. Eng. Sci.* **7** (1965) 431 – 439.
140. Y. N. Chen, M. Weber: "Flow-Induced Vibrations in Tube Bundle Heat Exchangers with Cross and Parallel Flow," *Flow-Induced Vibration in Heat Exchangers,* Proceedings of a Symposium, ASME, New York 1970, pp. 57 – 77.
141. H. J. Conners, Jr.: "Fluid Elastic Vibration of Tube Arrays Excited by Cross Flow," *Flow-Induced Vibration in Heat Exchangers, Proceedings of a Symposium,* ASME, New York 1970, pp. 42 – 56.
142. M. P. Paidoussis in E. Naudascher, D. Rockwell (eds.): "Flow-Induced Vibrations in Nuclear Reactors and Heat Exchangers-Practical Experiences and State of Knowledge," *Practical Experience with Flow-Induced Vibrations,* Springer, New York 1980, pp. 1 – 81.
143. B. M. H. Soper: ''The Effect of Tube Layout on the Fluid-Elastic Instability of Tube Bundles in Crossflow," *J. Heat Transfer* **105** (1983) 744 – 750.
144. T. A. Cowell, C. J. Davenport: "Acoustic Resonance in Air Cooled Heat Exchangers," *Inst. Chem. Eng. Symp. Ser.* **86** (1984) no. 2, 1223 – 1234.
145. R. Parker: "Acoustic Resonances in Passages Containing Banks of Heat Exchanger Tubes," *J. Sound Vibration* **57** (1978) 245 – 260.
146. K. T. Feldman: "Review of the Literature of Rijke Thermoacoustic Phenomena," *J. Sound Vibration* **7** (1968) 83 – 89.
147. J. A. Liburdy, J. L. Wofford: *A Stability Criterion for the Occurance of Thermally Induced Oscillations in Steady Laminar Flow,* ASME Paper 79-HT-74 (1974).
148. M. M. Friedlander, T. J. B. Smith, A. Powell: "Experiments on the Rijke Tube Phenomenon," *J. Acoust. Soc. Am.* (1964) 1737 – 1738.
149. H. Halle, J. M. Chenoweth, M. W. Wambsganass in S. S. Chen, J. C. Simonis, Y. S. Shin (eds.): "Shellside Flow-Induced Tube Vibration in Typical Heat Exchanger

Configurations: Overview of a Research Program," *Flow-Induced Vibration–1986,* Book No. G00355, PVP-vol. 104, ASME, New York 1986, pp. 161 – 169.
150. K. P. Byrne: "The Use of Porous Baffles to Control Acoustic Vibrations in Crossflow Heat Exchangers," *J. Heat Transfer* **105** (1983) 751 – 758.
151. G. Baylac, J. P. Gregoire: "Acoustic Phenomena in a Steam Generating Unit," *J. Sound Vibration* **42** (1975) 31 – 48.
152. M. P. Paidoussis, J. M. Chenoweth, M. D. Bernstein (eds.): *Symposium on Flow-Induced Vibrations,* vol. 3: "Vibration in Heat Exchangers," Book No. G00269, ASME, New York 1984.
153. S. S. Chen, J. C. Simonis, Y. S. Shin (eds.): *FlowInduced Vibration–1986,* Book No. G00355, PVP-vol. 104, ASME, New York 1986.
154. S. S. Chen: *Flow-Induced Vibration of Circular Cylindrical Structures,* Hemisphere Publishing Corp., New York 1987.
155. D. S. Weaver, J. A. Fitzpatrick: "A Review of Cross-Flow Induced Vibrations in Heat Exchanger Tube Arrays," *J. Fluids Structures* **2** (1988) 73 – 93.
156. H. Gotoda, S. Izumi: "A Modern Condenser Design," *Proceedings of the Institute of Marine Engineers,* March 1977, pp. 253 – 273.
157. A. C. Mueller: "Criteria for Maldistribution in Viscous Flow Coolers," *Heat Transfer 1974* **5** (1974) 170 – 174.
158. G. R. Putman, W. M. Rohsenow: "Viscosity Induced Non-Uniform Flow in Laminar Flow Heat Exchangers," *Int. J. Heat Mass Transfer* **28** (1985) 1031 – 1038.
159. J. B. Kitto, Jr., J. M. Robertson (eds.): *Maldistribution of Flow and Its Effect on Heat Exchanger Performance,* Book No. H00394, HTD-vol. 75, ASME, New York 1987.
160. A. C. Mueller, J. P. Chiou: "Review of Various Types of Flow Maldistribution in Heat Exchangers," *Heat Transfer Eng.* **9** (1988) no. 2, 36 – 50.
161. R. A. Bajura, E. H. Jones, Jr.: "Flow Distribution Manifolds," *J. Fluids Eng.* **98** (1976) 654 – 666.
162. M. K. Bassiouny, H. Martin: "Flow Distribution and Pressure Drop in Plate Heat Exchangers–I: U-Type Arrangement," *Chem. Eng. Sci.* **39** (1984) no. 4, 693 – 700.
163. M. K. Bassiouny, H. Martin: "Flow Distribution and Pressure Drop in Plate Heat Exchangers–II: Z-Type Arrangement," *Chem. Eng. Sci.* **39** (1984) no. 4, 701 – 704.
164. M. K. Bassiouny, H. Martin: "Temperature Distribution in a Four Channel Heat Exchanger," *Heat Transfer Eng.* **6** (1985) no. 2, 58 – 72.
165. A. K. Singhal, L. W. Keaton, A. K. Majumdar, T. Mukerjee: "An Improved Mathematical Formulation for the Computations of Flow Distributions in Manifolds for Compact Heat Exchangers," *Advances in Heat Exchanger Design,* Book No. H00373, HTD-vol. 66, ASME, New York 1986, pp. 105 –112.
166. E. F. C. Somerscales, J. G. Knudsen (eds.): *Fouling of Heat Transfer Equipment,* Hemisphere/McGraw-Hill, Washington, D.C., 1981.
167. R. W. Bryers (ed.): *Fouling and Slagging Resulting from Impurities in Combustion Gases,* Engineering Foundation, New York 1983.
168. R. W. Bryers (ed.): *Fouling of Heat Exchanger Surfaces,* Engineering Foundation, New York 1983.
169. B. A. Garrett-Price, S. A. Smith, R. L. Watts, J. G. Knudsen, W. J. Marner, J. W. Suitor: *Fouling of Heat Exchangers,* Noyes Data Corp., Park Ridge, N.J., 1985.
170. W. J. Marner, J. Suitor in S. Kakaç, R. K. Shah, W. Aung (eds.): "Fouling With Convective Heat Transfer," *Handbook of Single-Phase Convective Heat Transfer,* Chap. 21, J. Wiley, New York 1987.

Evaporation

REINHARD BILLET, Ruhr-Universität Bochum, Bochum, Federal Republic of Germany

1. **Separating Substances by Evaporation** . . . 2
2. **Types of Evaporators and their Energy Consumption** . . . 2
2.1. **Single Stage Evaporators** . . . 2
2.2. **Multistage Evaporation** . . . 4
2.3. **Evaporation with Vapor Recompression** . . . 7
2.4. **Optimizing Evaporation Systems** . . . 11
2.5. **Evaporation Using Pressure Reduction** . . . 12
3. **Designing Evaporators for Fabrication** . . . 14
4. **Design** . . . 15
4.1. **Heating Equipment** . . . 15
4.2. **The Vapor Separation Chamber** . . . 27
5. **Evaporator System Ancillary Components** . . . 30
5.1. **Indirect Heat Exchangers** . . . 30
5.2. **Equipment for Direct Heat Exchange** 30
5.3. **Pumps** . . . 33
5.4. **Control Instrumentation** . . . 34
5.5. **Insulation** . . . 35
6. **Evaporator System Operation** . . . 36
7. **Crystallization by Evaporation and by Evaporation with Reduced Pressure Flashing** . . . 36
8. **References** . . . 37

Symbols

V_H(kg/h)	heating steam flow rate
h_H(kJ/kg)	enthalpy of vaporization of the heating steam
t_H(°C)	condensing temperature of the heating steam
Δt_H(°C)	temperature drop for the heating steam condensate
p_H(Pa)	pressure of the heating steam
F (kg/h)	fresh solution feed rate
h_F(kJ/kg)	enthalpy of the fresh solution feed
t_F(°C)	inlet temperature of the fresh solution feed
ξ_F(wt %)	dissolved solids content of the fresh solution feed
D (kg/h)	solvent vapor flow rate
p (Pa)	pressure in the evaporator vapor separation chamber
h_D(kJ/kg)	enthalpy of the evaporated solvent
t_C(°C)	temperature in the evaporator vapor separation chamber
t (°C)	boiling point of pure solvent at pressure p
τ (°C)	boiling point rise for the solution at evaporator conditions
τ' (°C)	boiling point rise for salt solutions
τ^+ (°C)	boiling point rise due to hydrostatic head and pressure drop
ξ_C(wt %)	dissolved solids content of the concentrated solution
h_C(kJ/kg)	enthalpy of the concentrated solution
Q_L(kJ/h)	heat loss rate
c_H(kJ kg^{-1} K^{-1})	specific heat of the heating steam condensate
t_{WH}(°C)	temperature of the wall surface under the condensate film
z	number of ions into which a dissolved molecule dissociates
a	fraction of dissociated molecules in a solution with concentration ξ at the boiling point t of the pure solvent at operating pressure p
μ_s(kg/kgmol)	molecular mass of the solute
r_a(kJ/kg)	heat of vaporization of the solution
r_L(kJ/kg)	heat of vaporization of the pure solvent

1. Separating Substances by Evaporation

Solutions consisting of a solvent and dissolved solid substance may undergo simple thermal separation by vaporizing the solvent, thus concentrating the residual solid. This is a result of the vanishingly small vapor pressure of most solid components. This method plays an impor-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_03

tant part in the recovery of many common materials, for example, polymers, salts and the like. Evaporation may also serve the purpose of separation and recovery of the pure solvent. It is also possible to use evaporation to concentrate emulsions and suspensions. Usually, the term *vaporization* is used when the predominant purpose is recovering the solvent and the term *evaporation* when the task is separating the solid or increasing its concentration.

Solvent vaporization takes place by reducing the pressure above a hot boiling solution. This principle finds broad industrial application in evaporating highly corrosive solutions, particularly in separation processes using evaporative crystallization. For solutions of crystallizable solids, vaporizing the solvent leads inevitably to supersaturating the solute and precipitating crystals from the solution. Solvent vaporization through pressure reduction or flash vaporization finds frequent use in industrial vacuum crystallization processes.

In designing evaporation plants, particularly in the choice and design of the equipment, a series of points needs careful consideration: chief among these is the operational behavior of the materials being processed, for example, any tendency toward forming deposits and fouling the heating surfaces, or the residence time of thermally unstable products.

2. Types of Evaporators and their Energy Consumption

2.1. Single Stage Evaporators

Energy consumption is determined by preparing the material and energy balances. To assist in illustrating and explaining the relationships, Figure 1 shows an equipment sketch of a calandria-type evaporator together with the symbols most often used for schematic evaporator representation. The evaporator consists of a vertically arranged cylindrical vessel with two horizontal tube sheets built into the lower half. The tube sheets support vertical heat exchange tubes. Steam condensing on the outside of the tubes brings the liquid in the tubes to boiling. Vapor bubbles formed within the tubes discharge, mixed with liquid, from the top of the tubes, promoting a natural recirculating fluid flow regime. A large vertical pipe is located in the center of the tube bundle to provide for the necessary downflow of unvaporized solution.

The vapors produced separate from liquid in the upper section of the evaporator and, when necessary, flow through special internal separating devices to remove entrained liquid droplets. A condenser or preheater condenses the vapors to liquid.

Dry saturated or slightly superheated steam is introduced outside the tubes below the upper tubesheet into the "stream chest". If the increase in the boiling point of the solution at the bottom of the tubes caused by the hydrostatic pressure due to the column of liquid is neglected, then the temperature difference between the heating steam and boiling liquid is practically constant along the length of the tube wall. Heat exchange in this evaporating equipment is thus characterized by practically no change in temperature of the evaporating and condensing media. This is not true for other evaporator constructions employing once-through flow. In these, the boiling temperature increases along the length of the vaporizing surface due to increase in solvent concentration. Once-through designs include falling film and thin film evporators. Also, vaporizers heated through the sensible heat of hot liquids show varying temperature differences between the cold and hot fluids along the length of the vaporizing surface.

The vapor pressure of the solvent in equilibrium with a solution is less than that of the pure solvent at the same temperature. This increases the boiling point of the solution. The boiling point rise τ is defined as the difference between t_C, the temperature in the evaporator vapor separation chamber, and t, the boiling point of pure solvent at pressure p. The heating steam condensate suffers a drop in temperature in the steam chest. It leaves the steam chest at the temperature $(t_H - \Delta t_H)$. The concentrated solution leaves the evaporator at $(F-D)$ flow rate, enthalpy h_C, temperature $(t+\tau=t_C)$, and ξ_C solids content. Equation (1) gives the solvent vapor rate evaporated from a feed, F, at initial composition, ξ_F, when concentrated to ξ_C:

$$D=F\left(1-\frac{\xi_F}{\xi_C}\right) \tag{1}$$

The heat balance Equation (2) gives the necessary heating steam rate:

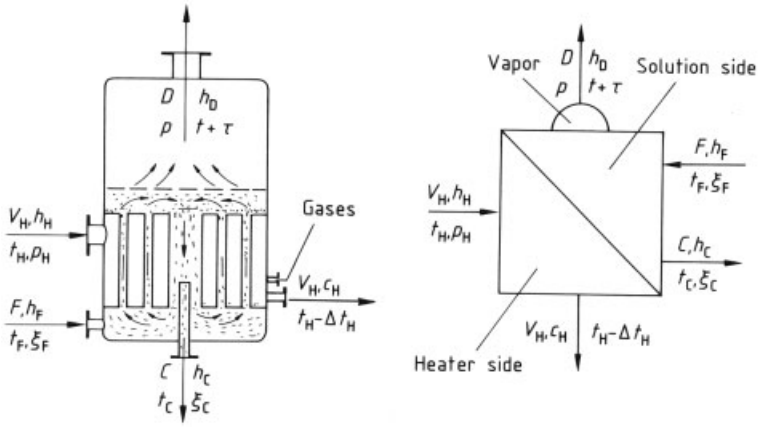

Figure 1. Calandria type evaporator and evaporator symbolic diagram

$$V_{\mathrm{H}} = \frac{D\,(h_{\mathrm{D}} - h_{\mathrm{c}}) + F\,(h_{\mathrm{c}} - h_{\mathrm{F}}) + Q_{\mathrm{L}}}{h_{\mathrm{H}} - c_{\mathrm{H}}\,(t_{\mathrm{H}} - \Delta t_{\mathrm{H}})} \tag{2}$$

The quantity Δt_{H} may be determined from a relation by KIRSCHBAUM applying to steam heated equipment with vertical heating surfaces so long as the steam condenses on the surface in the form of a liquid film:

$$\Delta t_{\mathrm{H}} = 0.55\,(t_{\mathrm{H}} - t_{\mathrm{WH}}) \tag{3}$$

Equation (4) may be used to calculate the boiling point rise τ' of a salt solution, insofar as it results from the solute content:

$$\tau' = 8.3\,(1 + (z-1)\,a)\,\frac{\xi}{100-\xi}\,\frac{(273+t)^2}{\mu_{\mathrm{s}} r_{\mathrm{a}}} \tag{4}$$

For dilute solutions, r_{a} is set approximately equal to r_{L}, the heat of vaporization of the solvent. Depending on the tube length and hydraulic behavior, a boiling point increase τ^+ results due to the hydrostatic head of the solution as well as the pressure drop arising from flow through the tube. The total boiling point rise τ results as the sum:

$$\tau = \tau' + \tau^+ \tag{5}$$

From Equation (1) and Figure 2 follows that, in simple, single-stage evaporators, most of the heat content of the vaporized solvent is lost when condensing the residual vapors not used for feed preheating, in a separate condenser. Therefore, in the absence of measures for improving efficiency, the consumption of heating steam is comparatively high. In evaporating aqueous solutions, the ratio of heating steam to vapors generated amounts to about 1.2. Efficiency improvement measures can reduce this ratio considerably. Such measures may include advantageous local operating circumstances, integrating an evaporation stage into the energy balance for a complete process, or so-called *heat pumping* schemes to be described below.

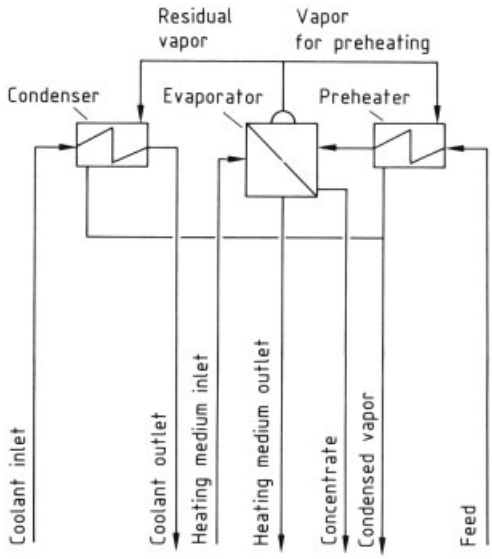

Figure 2. Single-stage evaporator system

A small efficiency improvement, of the order of 10 – 15 %, is possible through preheating the fresh feed solution by condensing a small

portion of the vaporized solvent. Figure 2 is a schematic diagram of a process for achieving this result. In Figure 2, a small portion of the vaporized solvent preheats the feed by condensing in the preheater. A cooling medium condenses the residual vapors in the condenser.

Because of their low thermal efficiency, simple single-stage evaporators are seldom used. Their application is limited to systems of small capacity or those in which the product requires short solution residence time due to temperature sensitivity.

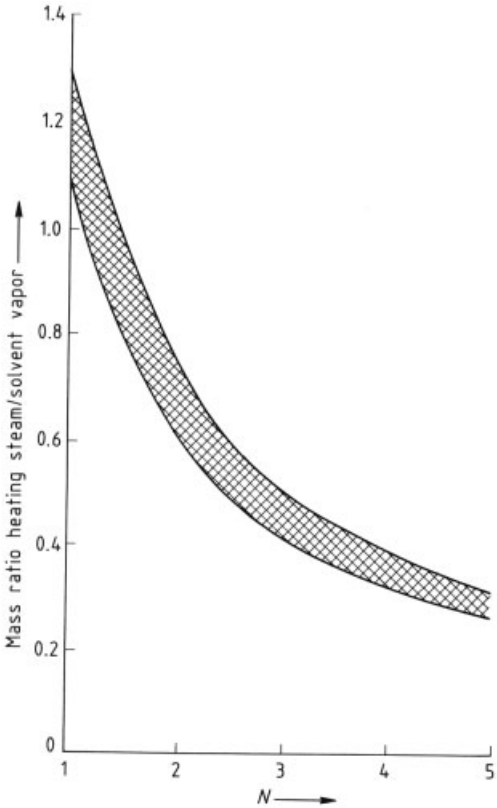

Figure 3. Heating steam consumption as a function of the number of stages *N*

2.2. Multistage Evaporation

Multistage evaporation is an effective means for reducing the steam consumption. A number of evaporator stages are connected together in such a manner that the solvent vapors generated in one stage serve as the heating medium for a succeeding stage. Also, only the vapors generated in the last stage must be condensed by a cooling medium, thus reducing cooling-water usage. Figure 3 is a representation of a potential efficiency improvement as a result of using multistage evaporation. The plant capital cost increases approximately linearly with the number of stages. Thus, for a given set of economic and operating circumstances, there exists an optimum number of evaporator stages for a given difference between the heating steam temperature in the first stage and the boiling temperature in the last stage. This optimum results in a minimum in the sum of costs due to capital charges and operating expenses. Moreover, the difference between the heating steam temperature in the first stage and the boiling temperature in the last stage sets a practical limit to the number of evaporator stages. For example, in the case of evaporators hydraulically dependent on natural circulation as shown in Figure 1, too small a difference in temperature between the hot condensing side and the boiling liquid side may inhibit uniform, steady state circulation. The boiling point rise tends to reduce even more the temperature difference between hot and boiling sides.

Three fundamental variants of multistage evaporator arrangements are possible. The relative direction of flow of steam or solvent vapors versus that of the solution serves to differentiate them. The method most frequently used in practice is so-called cocurrent flow feeding. In this, the feed solution is pumped into the stage at highest pressure first. Then the solution flows from stage to stage one after another without pumping, in the same direction as the solvent vapor using the stage to stage difference in pressure. The concentrated product leaves the system from the last, lowest pressure stage. In contrast, for countercurrent flow feeding, the fresh feed solution enters the stage at lowest pressure and is pumped from stage to stage to that at highest pressure. Solvent vapors flow in the opposite direction. Concentrated solution leaves the system from the evaporator at highest pressure and temperature. In the third arrangement, the fresh feed solution is divided into separate streams, with each stream feeding directly to a stage in parallel with the others.

Accordingly, the cocurrent flow feeding arrangement needs only a single solution pump. In operation, control is the simplest. The last

evaporator stage operates at the lowest temperature. This is a special advantage when evaporating solutions that become more temperature sensitive with increasing concentration. On the other hand, countercurrent flow feeding offers the advantage that the highly concentrated solution vaporizes in the stage at the highest temperature. There its viscosity is lower and the heat transfer rate is thereby better than would be the case in cocurrent flow feeding. For countercurrent flow feeding several liquid pumps are required. This means plant operations might be more prone to upsets than would be the case for cocurrent flow feeding. In practice, the third arrangement, splitting the feed, is used in those cases where, on evaporation, the solubility limit is exceeded, thus crystallizing out the solid material. It may also be used when concentrating a large feed flow for which exposure for only a short residence time is permitted.

Cocurrent flow feeding of steam and solution is the method predominantly used in practice. Figure 4 shows details of one possible arrangement for this process concept and the nomenclature for it. The nomenclature is an extension of Figure 1, and in the nomenclature listing the subscript n designates the stage number. Figure 4 shows condensate removal from the steam chest of each stage. The scheme provides several preheaters for multistage preheating of the fresh feed solution, each preheater using a portion of the solvent vapors generated in the respective evaporator stage.

The derivation of the relationships for determining the steam consumption result from material and energy balances across the thermodynamic boundaries of the total system and the subsystems comprising its parts. Since the assumption of steady-state operation applies, the net summations of mass and energy flows across system boundaries equal zero, i.e., total inflow of mass or heat across a boundary is equal to total outflow. From such balances around stages $1-n$, Equation (6) results for the solvent vapors from the nth stage:

$$\begin{aligned} &D_{\mathrm{n}}\left(h_{\mathrm{D}n}-h_{\mathrm{L}n}\right) \\ &=V_{\mathrm{H}}\left\{h_{\mathrm{H}}-c_{\mathrm{H}}\left(t_{\mathrm{H}}-\Delta t_{\mathrm{H}}\right)\right\} \\ &-D_1\left\{c_1\left(t_1-\Delta t_1\right)-h_{\mathrm{L}n}\right\} \\ &-D_{n-1}\left\{c_{n-1}\left(t_{n-1}-\Delta t_{n-1}\right)-h_{\mathrm{L}n}\right\} \\ &-V_1\left\{h_{\mathrm{D}1}-c_1\left(t_1-\Delta t_1\right)\right\} \\ &-V_{n-1}\left\{h_{Dn-1}-c_{n-1}\left(t_{n-1}-\Delta t_{n-1}\right)\right\} \\ &-F\left(h_{\mathrm{L}n}-h_{\mathrm{F}}\right)-Q_{\mathrm{L}1}\ldots-Q_{\mathrm{L}n} \end{aligned} \tag{6}$$

Using the nomenclature from Figure 4, the required withdrawal of solvent vapor for preheating the nth preheater by Equation (7) is calculated. In Equation (7), Δt_n designates the drop in condensate temperature on the heating side of preheater n:

$$V_n = F\frac{h_{\mathrm{F}n-1}-h_{\mathrm{F}n}}{h_{\mathrm{D}n}-c_n\left(t_n-\Delta t_n\right)} \tag{7}$$

Analogous to Equation (1), Equation (8) permits calculating the total solvent vapor removed from the solution in the n stages to achieve a desired final concentration:

$$\begin{aligned} &\sum_{n=1}^{n=N} D_n = D_1+D_2+\ldots\ldots D_n+\ldots\ldots D_N \\ &= F\left(1-\frac{\xi_{\mathrm{F}}}{\xi_{\mathrm{C}}}\right) \end{aligned} \tag{8}$$

On assuming a set of temperature and corresponding concentration distributions for the individual stages, substituting the numerical values of enthalpies, temperatures, concentrations, and flows into Equations (6) and (7) allows calculation of D_n as a function of V_{H}:

$$D_n = f_n\left(V_{\mathrm{H}}\right) \tag{9}$$

Then, Equation (8) permits determination of V_{H} for the selected distributions. This procedure may be applied to other multistage evaporator configurations in an analogous manner.

Equation (10) defines the effective temperature difference, the difference between the first stage heating medium (the stage operating at highest pressure) and the boiling solution in the last stage (the stage operating at lowest pressure) reduced by the sum of all boiling point elevations according to Equation (5):

$$\Delta t_{\mathrm{eff}} = t_{\mathrm{H}} - t_{\mathrm{N}} - \sum_{n=1}^{n=N} \tau_n \tag{10}$$

This effective temperature difference obviously sets limits to the usable number of stages N in a multistage evaporator plant. Since this total temperature difference must be allocated over the n stages and since the boiling point elevations reduce the temperature differences effective for heat transfer, a large increase in the total

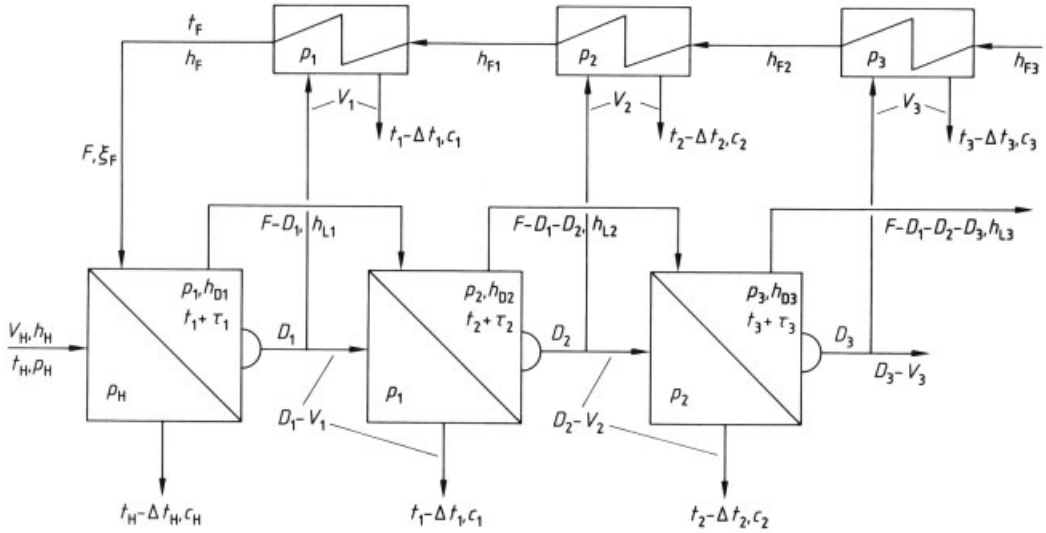

Figure 4. Three-stage evaporator system with cocurrent flow of liquid and vapor

heat transfer area for the evaporator calandria becomes necessary. Equation (11) expresses the relationship between the heat transferred in a stage Q_n, the transfer coefficient k_n, the area A_n, and the temperature differences:

$$Q_n = k_n \cdot A_n \left(t_{\mathrm{H}} - t_{\mathrm{N}} - \sum_{n=1}^{n=N} \tau_n \right) / N \tag{11}$$

A large number of stages leads to small temperature differences. For evaporators operating with natural circulation arising from a thermosiphon effect, a certain minimum temperature difference is necessary in every stage to achieve stable hydraulic conditions. If not, vapor bubble formation becomes unstable and natural circulation ceases. Thus, in multistage systems with natural circulation evaporators, the maximum possible number of stages is determined by the effective total temperature difference and the minimum permissible temperature difference in a stage. The latter lies roughly between 10 – 20 °C. For highly viscous liquids, it may take on even higher values. If a pump forces the liquid circulation in the evaporator equipment, then the temperature difference in a stage may be held as small as is desired. Then, its optimal value, and thus also the optimum number of stages, must be determined by detailed economic studies. This becomes necessary due to the fact that energy costs do indeed decrease as the number of stages increases. However, the capital costs are higher.

For equally sized heat-exchange surfaces A_n, in all stages, Equation (12) shows the relationships for allocating the total temperature difference among the stages. Equation (13) for the scheme in Figure 4 applies for Q_n, the heat transferred in a stage:

$$t_{n-1} - (t_n + \tau_n) = \frac{\frac{Q_n}{k_n}}{\sum\limits_{n=1}^{n=N} \frac{Q_n}{k_n}} \cdot \left(t_{\mathrm{H}} - \left(t_N + \sum_{n=1}^{n=N} \tau_n \right) \right) \tag{12}$$

$$\begin{aligned} Q_n &= (D_{n-1} - V_{n-1}) \\ &\{h_{\mathrm{D}n-1} - c_{n-1}(t_{n-1} - \Delta t_{n-1})\} \\ &+ (D_1 + D_2 + \ldots + D_{n-2}) \\ &\{c_{n-2}(t_{n-2} - \Delta t_{n-2}) - c_{n-1}(t_{n-1} - \Delta t_{n-1})\} \end{aligned} \tag{13}$$

Apportioning the total temperature difference among the stages to result in equal heat-transfer surface in each stage does not, however, always lead to the most economical system. For example, if it is necessary to use expensive materials of construction, then it may be more advantageous to choose the temperature differences so that the total heat transfer surface for the evaporator system results in a minimum. This leads to Equation (14), which, like Equation (12), can only be solved iteratively:

$$t_{n-1} - (t_n + \tau_n) = \frac{\sqrt{\frac{Q_n}{k_n}}}{\sum_{n=1}^{n=N} \sqrt{\frac{Q_n}{k_n}}} \left[t_{\mathrm{H}} - \left(t_N + \sum_{n=1}^{n=N} \tau_n \right) \right] \tag{14}$$

Frequently, in fixing the boiling temperatures, the thermal instability of the feed requires consideration. Or, allowance must be made for the fact that, with increasing temperature, many products agressively corrode the heat-transfer surfaces.

Equation (15) is recommended for making a rough estimate of the heating steam consumption:

$$V_{\mathrm{H}} = \frac{\sum_{n=1}^{n=N} D_n}{\Delta h_{\mathrm{H}}} \left(\Delta h_{\mathrm{m}} - \Delta h_1 \cdot \frac{N-1}{N} \right) \tag{15}$$

$$\Delta h_{\mathrm{m}} = \frac{F\,(h_{\mathrm{L1}} - h_{\mathrm{F}})}{\sum_{n=1}^{n=N} D_n} + \frac{1}{1 + 0.01 \cdot N} \Delta h_1 \tag{16}$$

where

Δh_{H} = heat of condensation of the heating steam

Δh_{l} = heat of condensation of the solvent vapors from the first stage

Δh_m = average heat of vaporization according to Equation (16)

h_{F} = enthalpy of the feed solution to the first stage

h_{L1} = enthalpy of the solution boiling in the first stage

2.3. Evaporation with Vapor Recompression

Vapor-recompression evaporation is a heat-pumping technique. A portion, or the total flow of solvent vapor leaving the evaporator vapor separation chamber, is compressed to a higher pressure and condensed in the steam chest of the same evaporator. In many cases, this processing scheme achieves specific heating steam efficiencies that can only be attained by multistage evaporator systems having as many as three to five stages. One may use steam jet ejectors operated by high pressure steam to compress the solvent vapors. Or this can be done mechanically with centrifugal compressors, or, in exceptional cases, with positive displacement machines. In all cases, the steam or energy consumption for operating the compressor increases greatly with an increase in the temperature difference between the condensing compressed vapors and the boiling solution being evaporated. Solutions having high boiling point elevation cannot economically be concentrated by vapor recompression.

Setting the temperature difference required for heat transfer in the evaporator results from economic studies. These determine the minimum sum of operating costs and capital charges for the evaporator system. The most economical temperature difference between condensing heating vapors and boiling liquid usually lies in the range of about 10 – 20 °C. In general, the capital cost of vapor recompression systems using electrically driven compressors is higher than for multistage systems. Occasionally, a combination of both schemes may be used.

A single-stage evaporator permits selecting a relatively low boiling temperature, one attainable, usually, only in the last stage of multistage systems. Thus evaporation with heat recovery through vapor recompression often represents the sole possibility for concentrating temperature-sensitive solutions economically at the lowest possible boiling temperature. The field of application for vapor recompression is thus definitely in the area of products requiring lower boiling temperatures.

Figure 5 schematically shows compressing the solvent vapors by using thermal compression. A steam jet ejector driven by high pressure inlet steam draws in the evaporated solvent vapor and exhausts at a pressure sufficiently high that it supplies the necessary evaporator heating steam. The steam jet ejector consists of the body with inlet flange and steam nozzle and the mixing nozzle with diffusor. Since it is simple and has no moving parts, its operation is free of breakdowns and its purchase cost is low. In addition, the choice of materials of construction is not limited. However, the efficiency of new units is low and, in operation, usually becomes lower due to steam nozzle erosion, which is caused by

water droplets in the drive steam. Figure 5 also defines the nomenclature for the symbols used in the following description.

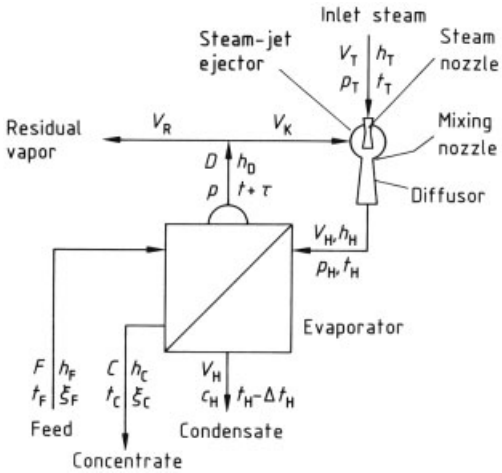

Figure 5. Evaporator with steam jet ejector

Drive steam at a pressure greater than that in the evaporator enters the steam nozzle. It expands isentropically to the pressure in the mixing nozzle, drawing in and mixing with evaporated solvent vapors. The combined drive steam and solvent vapor together flow through the diffusor, which converts the kinetic energy of velocity back again into the potential energy of pressure. Finally, combined drive steam and solvent vapor leave the diffusor and enter the evaporator steam chest at a pressure below that of the drive steam but above that of the evaporator vapor separation chamber.

To determine the fraction of the total heating steam V_{H} represented by drive steam V_{T} and by solvent vapor V_{K}, a performance chart for the particular ejector is used–Figure 6 represents a typical example. Figure 6 shows the pumping behavior of a steam jet ejector. The dependence of the entrainment ratio

$$\varphi = \frac{V_{\mathrm{K}}}{V_{\mathrm{T}}} \tag{17}$$

is plotted as a function of solution boiling point (or evaporator pressure) at fixed steam chest pressure (or heating steam condensing temperature) with parameters of drive steam pressure. Then, by material balance, Equation (18) gives the flow of heating steam, Equation (19) the drive steam for the steam jet ejector or the net heating steam, and Equation (20) the compressed solvent vapor rate:

$$V_{\mathrm{H}} = V_{\mathrm{T}} + V_{\mathrm{K}} = V_{\mathrm{T}}\,(1+\varphi) \tag{18}$$

$$V_{\mathrm{T}} = V_{\mathrm{H}}\,\frac{1}{1+\varphi} \tag{19}$$

$$V_{\mathrm{K}} = V_{\mathrm{H}}\,\frac{\varphi}{1+\varphi} \tag{20}$$

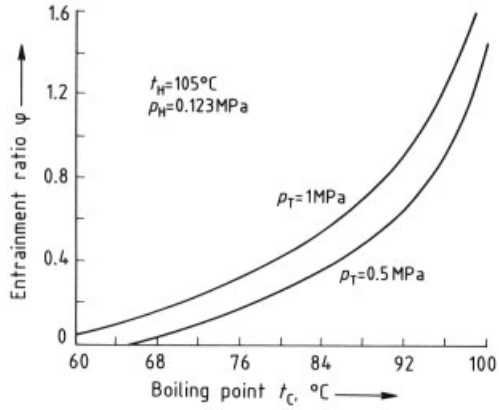

Figure 6. Entrainment ratio as a function of the boiling point

The flow of heating steam required for the specified evaporation is known from Equation (2). Then the entrainment ratio is read from Figure 6 for the evaporator temperature, and the drive steam rate is calculated from Equation (19) and the solvent vapor flow from Equation (20).

From energy efficiency considerations, steam jet ejectors are more favorable for vapor recompression in the area of lower working pressures. The higher flow rates, above 5000 $\mathrm{Nm^3/h}$, are the field of application for centrifugal compressors. Not only electric motors but also steam turbines serve for the drive equipment. Figure 7 shows a process schematic along with nomenclature for a vapor recompression process using an electrically driven compressor. The auxiliary heating is necessary for startup. However, it is also needed during operation for the case where the compressed solvent vapors do not suffice. This occurs when feeding relatively cold solution.

The schematic enthalpy – entropy diagram shown in Figure 7 assists in defining the compressor performance. Simplifying assumptions

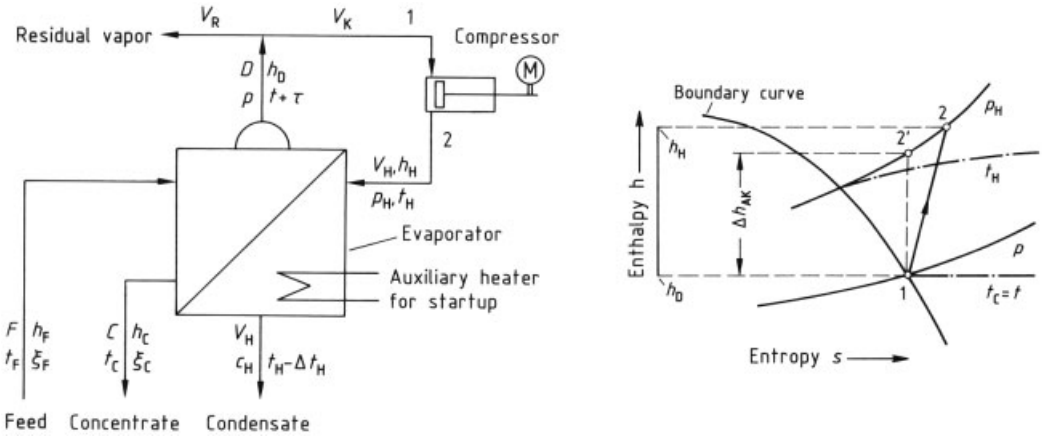

Figure 7. Evaporator with vapor compressor

are that the solvent vapors are dry and saturated and that the compression requires only a single stage. A theoretically reversible isentropic compression would follow the dotted path 1 – 2′ with isentropic enthalpy change Δh_{AK}. The real compression follows the solid line polytropic path 1 – 2 with discharge enthalpy h_{H} resulting from Equation (21):

$$\Delta h_{\mathrm{AK}} = (h_{\mathrm{H}} - h_{\mathrm{D}}) \cdot \eta_{\mathrm{iK}} \tag{21}$$

where η_{iK} is the indicated isentropic gas efficiency. Designating η as the total efficiency (including mechanical inefficiencies) allows expressing the compressor total power consumption N_{K} (kW) by equation (22):

$$N_{\mathrm{K}} = \frac{1}{3600} \cdot \frac{\Delta h_{\mathrm{AK}} \cdot V_{\mathrm{H}}}{\eta} \tag{22}$$

Mechanical vapor recompression with an electrically driven compressor is then economical only if the temperature difference ϑ between the condensing compressed vapor and the boiling solution can be kept small, and a single-stage compressor suffices. Usually, this means that ϑ must be less than 20 °C, which is the case only for solutions having low boiling-point elevation.

Figure 8 shows a schematic diagram for vapor recompression using a steam turbine driven compressor. The compressed solvent vapors and the turbine drive steam exhaust to the same pressure, that of the steam chest. The heating steam flow V_{H} is the sum of the two flows. Also given in Figure 8 is a schematic Mollier enthalpy – entropy diagram, which qualitatively shows the changes in state for the steam flows. The evaporated vapors are assumed to be practically dry and saturated, permissible when the boiling point rise is small. Moreover, both expansion in the turbine as well as compression are assumed to be single stage.

Theoretically reversible isentropic expansion of the superheated drive steam from p_{T} to p_{H} would follow the dotted path a – b′ with isentropic enthalpy change Δh_{AT}. Likewise, theoretically reversible isentropic compression of the evaporated vapors from p to p_{H} would follow the dotted path c – d′ with isentropic enthalpy change Δh_{AK}. Actually, both expansion and compression are polytropic, following the paths a – b with enthalpy drop Δh_{T} and c – d with enthalpy rise Δh_{K}, respectively. For indicated isentropic gas efficiencies η_{iT} for the turbine and η_{iK} for the compressor, the enthalpy of the expanded drive steam h_{b} is:

$$h_{\mathrm{b}} = h_{\mathrm{T}} - \eta_{\mathrm{iT}} \cdot \Delta h_{\mathrm{AT}} \tag{23}$$

and that for the compressed vapor h_{d} is:

$$h_{\mathrm{d}} = h_{\mathrm{D}} + \frac{\Delta h_{\mathrm{AK}}}{\eta_{\mathrm{iK}}} \tag{24}$$

From these, Equation (25) gives the enthalpy of the mixed vapor:

$$h_{\mathrm{H}} = \frac{V_{\mathrm{T}} h_{\mathrm{b}} + V_{\mathrm{K}} h_{\mathrm{d}}}{V_{\mathrm{H}}} \tag{25}$$

Once again, Equation (2) gives the flow of heating steam V_{H}, composed of drive steam

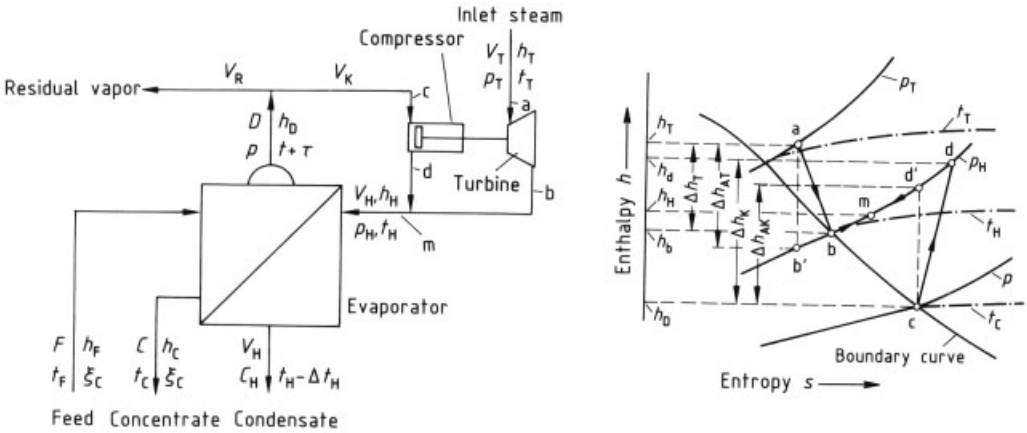

Figure 8. Evaporator with steam turbine driven vapor compressor

and compressed evaporated vapors, required to achieve the specified evaporation. Equations (25) and (26) are then solved simultaneously giving equations (27) and (28) for calculating V_{T}, the drive steam flow, and V_{K}, the compressed vapor flow. η_{K} and η_{T} are the respective total efficiencies of compressor and turbine determined as the products of the respective indicated gas and mechanical efficiences:

$$V_{\mathrm{K}} - \frac{\Delta h_{\mathrm{AK}}}{\eta_{\mathrm{K}}} = V_{\mathrm{T}} \eta_{\mathrm{T}} \Delta h_{\mathrm{AT}} \tag{26}$$

$$V_{\mathrm{T}} = \frac{V_{\mathrm{H}} \Delta h_{\mathrm{AK}}}{\Delta h_{\mathrm{AT}} \eta_{\mathrm{T}} \eta_{\mathrm{K}} + \Delta h_{\mathrm{AK}}} \tag{27}$$

$$V_{\mathrm{K}} = V_{\mathrm{H}} \frac{\eta_{\mathrm{T}} \eta_{\mathrm{K}} \Delta h_{\mathrm{AT}}}{\eta_{\mathrm{T}} \eta_{\mathrm{K}} \Delta h_{\mathrm{AT}} + \Delta h_{\mathrm{AK}}} \tag{28}$$

Given values of η_{T} and η_{K}, and with the help of the Mollier diagram, Equation (29) allows determination of h_{H}, the enthalpy of the heating steam:

$$h_{\mathrm{H}} = \frac{\Delta h_{\mathrm{AK}}}{\eta_{\mathrm{T}} \eta_{\mathrm{K}} \Delta h_{\mathrm{AT}} + \Delta h_{\mathrm{AK}}} \cdot \left(h_{\mathrm{b}} + h_{\mathrm{d}} \eta_{\mathrm{T}} \eta_{\mathrm{K}} \frac{\Delta h_{\mathrm{AT}}}{\Delta h_{\mathrm{AK}}} \right) \tag{29}$$

Then, Equation (2) is used to calculate the required flow of heating steam V_{H}; Equation (27), the drive steam flow for the turbine V_{T}; and Equation (28) is used to calculate V_{K}, the flow of compressed evaporated vapors.

Vapor recompression results in significant energy savings. Specific consumption of drive steam d, the net steam required, may be as low as 0.3.

$$d = V_{\mathrm{T}} / D$$

Economically justifiable operating temperatures can be as low as about $t_{\mathrm{C}} = 50\,^{\circ}\mathrm{C}$, and temperature differences between condensing heating steam and boiling solution ϑ is $10 - 20\,^{\circ}\mathrm{C}$.

According to the definition of the problem and the operating conditions, evaporator systems using vapor recompression may require auxiliary condensers or heaters. To the extent that the specified evaporation requirements produce more total solvent vapor than can be compressed and condensed in the steam chest, given the turbine and compressor efficiencies and initial pressure and superheat of the drive steam, an auxiliary condenser is needed. Likewise, if the total compressed solvent vapor plus drive steam does not suffice for the specified evaporation requirements, then the deficiency must be met with auxiliary heating.

For a plant with motor driven compressor, in the limiting case where neither auxiliary condensing or heating is required, the compressed evaporated solvent vapor flow V_{K} must equal the required steam flow rate D. Equation (30) applies to this limiting case:

$$(h_{\mathrm{F}})_{D=V_{\mathrm{K}}} = h_{\mathrm{C}} - \left(1 - \frac{\xi_{\mathrm{F}}}{\xi_{\mathrm{C}}}\right) \cdot \left[\{h_{\mathrm{H}} - c_{\mathrm{H}} (t_{\mathrm{H}} - \Delta t_{\mathrm{H}})\} \left(1 + \frac{1}{\varphi}\right) - (h_{\mathrm{D}} - h_{\mathrm{C}}) \right] \tag{30}$$

The probability of needing an auxiliary condenser is substantially greater if the solvent vapor compression is effected with a steam jet ejector or a steam turbine driven compressor than when an electrically driven unit is used.

With electrically driven compressor units, auxiliary heating is often necessary. For this case, one may calculate the auxiliary steam V_z needed with Equation (31):

$$V_Z = V_H \left\{ 1 - \frac{h_H - c_H\,(t_H - \Delta t_H)}{h_D - h_C + \frac{h_C - h_F}{1 - \xi_F/\xi_C}} \right\} \tag{31}$$

If the calculation results in a negative value for V_z, then auxiliary heating is unnecessary, but an auxiliary vapor condenser is required. In that case, the portion of the evaporated solvent that must be condensed, V_R, is:

$$V_R = D - V_H = -V_Z \tag{32}$$

However, in both the above cases, and all other cases when needed, evaporator startup requires auxiliary heating until a sufficient flow of solvent vapor for the compressor is produced. In the case of evaporated vapor compression using a steam jet ejector or steam turbine driven compressor, where it is not possible to dispense with auxiliary heating, the relationship giving the addition flow of steam required is Equation (33) with the pumping ratio $\varphi = V_K/V_T$ given by Equation (34):

$$V_Z = V_H \left\{ 1 - \left(1 + \frac{1}{\varphi}\right) \frac{h_H - c_H\,(t_H - \Delta t_H)}{h_D - h_C + \frac{h_C - h_F}{1 - \xi_F/\xi_C}} \right\} \tag{33}$$

$$\varphi = \eta_T \eta_K \frac{\Delta h_{AT}}{\Delta h_{AK}} \tag{34}$$

2.4. Optimizing Evaporation Systems

An economically optimal evaporator system is the result of a project engineering effort that always strives to minimize the total cost, which is made up of investment costs and operating costs. In this way, the optimal number of stages N_{opt} is found for a project-engineering multistage evaporator system. If the design of a multistage evaporator system requires expensive materials of construction to avoid corrosion, this may then, in the design, fix a need to minimize total heating surface (see Section 2.2). This additional optimization parameter must then be taken into consideration in the aforementioned evaluation procedures for determining the optimal number of stages.

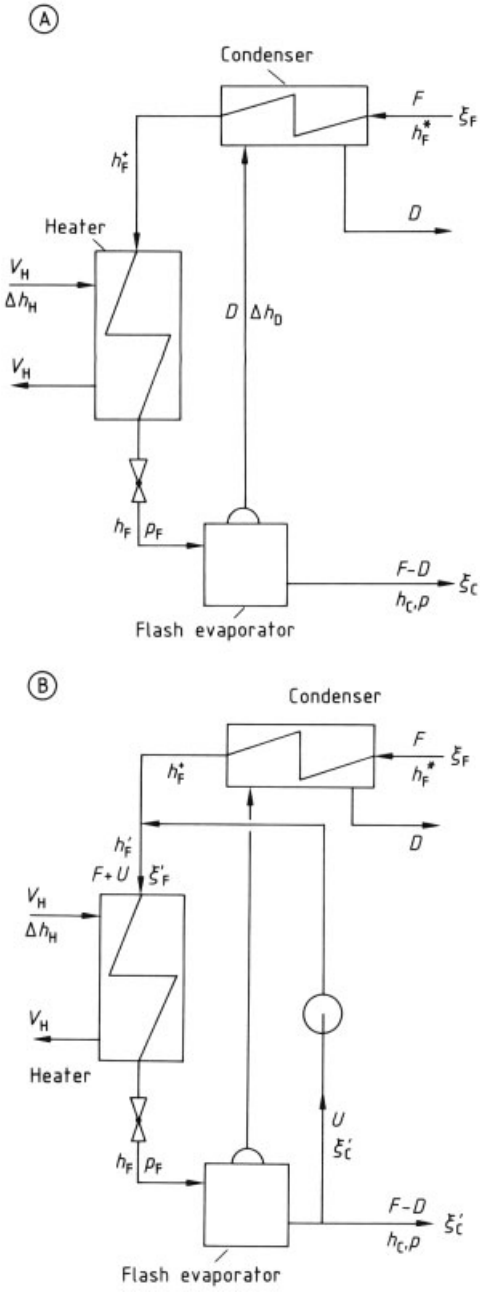

Figure 9. Evaporation by reduced pressure flashing
A: Without liquid circulation; B: liquid circulation

If the temperature difference in a particular stage is insufficient to provide a natural, thermosiphon circulation, then the liquid circulation in the evaporators can be forced by installing pumps. This forced circulation leads to considerable improvement in heat transfer with increasing liquid velocity, and thus reduction in heat transfer surface. The optimal circulation velocity is that resulting in minimum total cost.

In the case of vapor recompression with an electrically driven compressor, the necessary energy for compression increases with increasing difference in temperature between the condensing heating vapors and the boiling solution, $\vartheta = t_H - t_C$. At the same time, however, the evaporator heating surface decreases. Consequently, there is an optimal temperature difference at which a minimal value in the total cost results.

The local conditions and specific energy costs at the erected location of an evaporation system may materially influence the results of optimization studies. For example, the cost of cooling water in itself may be a decisive factor.

2.5. Evaporation Using Pressure Reduction

Figure 9 A shows a flash vaporizer system consisting of a liquid heater, a separator drum for the flash vapors formed on expansion, and a vapor condenser. Frequently, this type of evaporation scheme is employed for solutions inclined toward forming deposits on or otherwise scaling or corroding the heat transfer surfaces. The nomenclature and units for the following discussion of the single-stage flash vaporizer in Figure 9 A generally follows previous usage.

The enthalpy of fresh feed at h_F* increases to h_F^+ in the vapor condenser and to h_F in the steam-heated liquid preheater. The throttle valve maintains the required pressure difference between the liquid preheater outlet and the flash separator drum, $(p_F - p)$. The vapors D produced in the flash separator drum condense in the vapor condenser. If the temperature drop for the condensed vapors or steam in the heat exchangers as well as any boiling point rise in the preheater is neglected, then only the latent heats Δh_H of the heating steam in the preheater and that of the vapors Δh_D in the condenser enter into consideration. Consequently, Equations (35) and (36) express the relationships for calculating the specific heating steam consumption from the enthalpies fixed by the job specification:

$$\frac{V_H}{D} = \frac{\Delta h_D}{\Delta h_H} \cdot \frac{h_F - h_F^*}{h_F^+ - h_F^*} = \frac{\Delta h_D}{\Delta h_H} \cdot \frac{h_F - h_F^*}{h_F - h_C} \tag{35}$$

or

$$\frac{V_H}{D} = \frac{\Delta h_D}{\Delta h_H} \cdot \frac{h_C - h_F^*}{h_F - h_C} \tag{36}$$

Equation (37) fixes the enthalpy h_F^+ to which the fresh feed can be warmed in the evaporated vapor condenser:

$$h_F^+ = h_F^* + h_F - h_C \tag{37}$$

The ratio of inlet steam consumption to evaporated vapor produced is substantially higher than that for a single-stage calandria heated evaporator. The reason is that the evaporated vapor D raised in the flash separator drum at relatively high feed liquid preheat, even above the boiling temperature itself, makes up only a fraction of the available liquid flow. Equation (38) gives this ratio:

$$\frac{D}{F} = \frac{h_F - h_C}{\Delta h_D} = \frac{h_F^+ - h_F^*}{\Delta h_D} \tag{38}$$

The concentration ratio

$$\frac{\xi_F}{\xi_C} = 1 - \frac{h_F - h_C}{\Delta h_D} \tag{39}$$

is comparatively unfavorable. Its best value is set by the maximum attainable feed liquid enthalpy h_F. Also, the feed liquid pressure upstream of the throttle valve depends on h_F.

To improve the concentration ratio for a single-stage flash vaporizer, an arrangement like in Figure 9 B is used. A pump is installed to recycle a partial flow of flashed liquid to mix with the fresh solution preheated in the vapor condenser. Thus, liquid flow through the preheater increases by the factor f:

$$f = \frac{F + U}{F} \tag{40}$$

By this means the flow of vapor evaporated by flashing is greater by the factor f than that without liquid recycle:

$$\frac{D}{F} = f \cdot \frac{h_F - h_C}{\Delta h_D} \tag{41}$$

Also, the fresh feed solution is heated in the vapor condenser to a higher enthalpy:

$$h_F^+ = h_F^* - f\,(h_F - h_C) \tag{42}$$

Equation (43) also defines the characteristic factor f, allowing calculation of the recycle flow by Equation (44) as well as the heating steam consumption by Equation (45)

$$\frac{V_H}{D} = \frac{\Delta h_D}{\Delta h_H} \cdot \frac{1}{f} \cdot \frac{h_C - h_F^*}{h_F - h_C} \tag{43}$$

$$U = F\,(f-1) \tag{44}$$

$$\frac{V_H}{D} = \frac{\Delta h_D}{\Delta h_H} \cdot \frac{1}{f} \cdot \frac{h_C - h_F^*}{h_F - h_C} \tag{45}$$

Corresponding to the greater flow of evaporated vapor in operation with liquid recycle, the concentrated solution has a higher solids concentration ξ'_C than that in operation without recycle, Equation (46), as well as achieving a more favorable concentration ratio, Equation (47)

$$\xi_C = \frac{\xi_C}{f - \frac{f-1}{\xi_F/\xi_C}} \tag{46}$$

$$\frac{\xi_F}{\xi'_C} = 1 - f \cdot \frac{h_F - h_C}{\Delta h_D} \tag{47}$$

The multistage principle may also be applied to flash vaporization to achieve better heat utilization. A series of evaporated vapor condensers preheats fresh feed solution to progressively higher temperatures by condensing evaporated vapors at progressively higher pressures. Inlet steam completes solution preheating, followed by successive stages of pressure reduction in flash separator drums, generating vapors at the pressures used in preheating. This brings the solution to progressively higher solids concentration. Recovery of potable water from sea water to a large extent makes use of this process of multistage flash vaporization. In contrast to multistage calandria evaporators, multistage flash vaporizers are notable for having no lower limit with respect to the temperature difference between two neighboring evaporator stages. Thus, at otherwise equal conditions, it is possible to build a flash vaporizer system with a substantially greater number of stages.

Relationships derived from heat and material balances also permit engineering energy calculations for multistage flash vaporizers. Figure 10 shows the unmistakably strong dependence of the energy requirement on the number of stages N, preheat temperature t_F, as well as the temperature difference ϑ_e between a flash vapor condensing temperature and the outlet feed liquid temperature of the respective condenser. With respect to energy conservation, then, systems with a large number of stages and thus higher seawater preheat are especially suited for economically producing potable water.

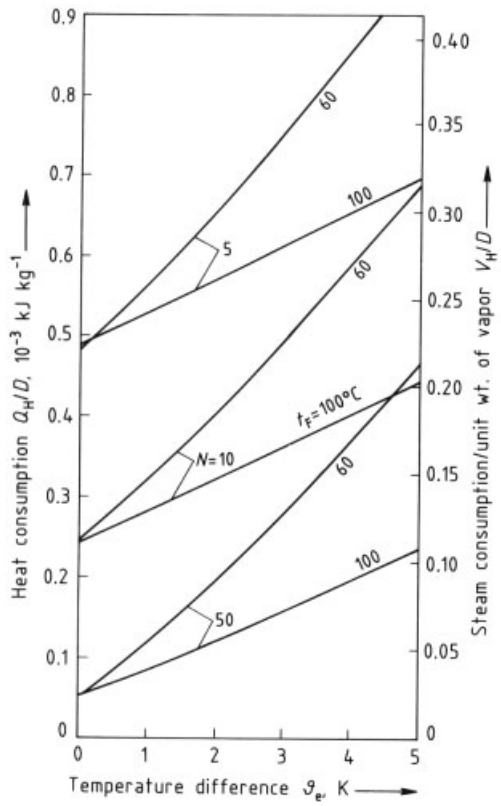

Figure 10. Heat consumption of reduced pressure flash evaporators as a function of number of stages, preheat temperature, and temperature difference

There are yet a number of other possible industrial applications of flash vaporizers. Thus, vacuum cooling units used in the potash industry employ this principle of operation. These serve to cool hot salt solutions by removing heat through partial autovaporization in a series of vacuum vessels operating at stagewise decreasing pressures.

3. Designing Evaporators for Fabrication

The decisive factors governing the selection of an evaporator design are the system capacity and the behavior of the feedstock from an operational viewpoint. For physically stable solutions with little tendency toward scaling and corrosion, few operational upsets should be expected, and comparatively simple and inexpensive equipment design would normally be employed. But for heat-sensitive materials that are easily decomposed, a relatively expensive unit having a short product residence time and designed to operate at as low vacuum as possible would be necessary. The properties of the substance, which are dependent on the pressure and temperature, as well as the vaporization behavior, in turn dependent on the concentration of the solution, determine the equipment design concept. Figure 11 shows some design concepts frequently employed:

11 A) Units with vertical evaporator tubes, heated on the outside, and having a central downcomer for natural thermosiphon liquid circulation, often referred to as calandria;
11 B) Units with vertical evaporator tubes, heated on the outside, contained in an external heat exchanger, also with external downcomer for natural thermosiphon or forced liquid circulation;
11 C) Units with sloping, or in the limiting case, horizontal evaporator tubes, heated on the outside, contained in an external heat exchanger, also with external downcomer for natural thermosiphon or forced liquid circulation;
11 D) Units with horizontal, internally heated evaporator tubes, immersed in the liquid, commonly referred to as kettles;
11 E) Units with vertical evaporator tubes, heated on the outside, for once-through liquid flow from bottom to top;
11 F) Units with vertical evaporator tubes, heated on the outside, for liquid flow from top to bottom, in accordance with the falling film principle;
11 G) Agitated film evaporators, vertical units equipped with rotating wiper blades for spreading the downward flowing liquid in an evenly distributed thin film on the inside surface of a large diameter, externally heated evaporator pipe. The pipe may be tapered. Sometimes, the units are horizontal;
11 H) Units with a rotating, externally heated, conical vaporizing surface, with liquid flowing upward in a thin film, evenly distributed by centrifugal force, often referred to as molecular stills of the Hickman type;
11 I) Units with no heating surface, consisting of a simple flash separator drum, for flash vaporization of preheated liquids.

While Types 11 A – C employ the same operating principle wherein the evaporation phenomenon is coupled with liquid circulation, the remainder of the units function differently, and there are fundamental differences among them. In Types 11 A – E, the vapor forms in the presence of a large flow of liquid. In contrast, in Types 11 F – H, vaporization occurs from a liquid film or a relatively thin layer. In Type 11 I, vaporizing the solvent is achieved by reducing the pressure on a preheated fresh solution.

The indicated differences relative to construction and manner of operation lead, in part, to diverse operating conditions. Thus, for example, an agitated thin film evaporator, Type 11 G, must operate with a higher temperature difference between heating medium and boiling liquid to achieve high vaporization capacity. The reason is a limitation of this construction in that a unit of given total volume can only accomodate a comparatively small heat exchanger surface.

Table 1 presents a brief overview of the suitability of the various vaporizer operating concepts for performing the tasks met in evaporation technology.

4. Design

4.1. Heating Equipment

In Types 11 A – E, the energy Q given up by the heating medium is imparted to the solution by heat transfer through the exchanger tube walls. For this, based on an expression for steady-state transfer of heat through circular walls with height H, inner diameter d_i, and outer diameter d_a, the following relationship applies:

$$Q = \frac{\pi H \vartheta}{\frac{1}{\alpha_\mathrm{V} d_\mathrm{a}} + \frac{1}{2\lambda_\mathrm{W}} \ln \frac{d_\mathrm{a}}{d_\mathrm{i}} + \frac{1}{a_\mathrm{L} d_\mathrm{i}}} \tag{48}$$

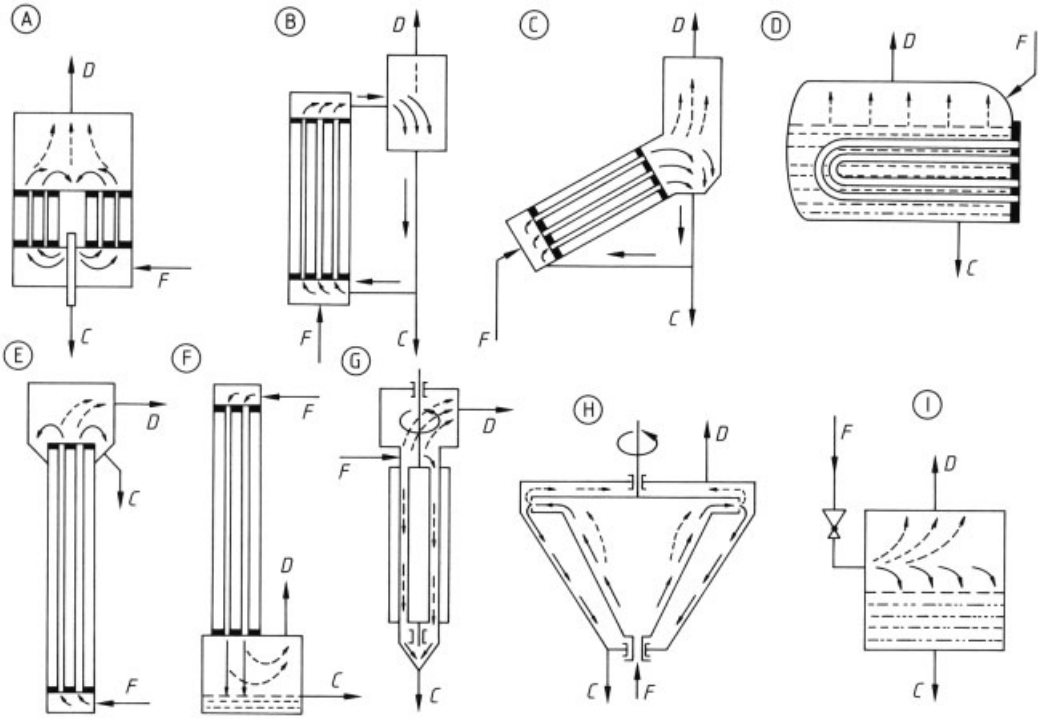

Figure 11. Evaporator types
See the text for the description of the various types.

Table 1. Operating characteristics of various types of evaporators

Type of evaporator	Viscosity, Pa s	Evaporative capacity	Tendency to salting and scaling	Holdup time
Natural circulation	0.05	small – large	great	very long
Forced circulation	1.0	small – large	less	very long
Climbing film	0.5	small – large	great	medium
Falling film	0.1	small	medium – great	short
Thin film	10 – 100	medium – large	slight	short
Flash	0.01	small – medium	slight	short – long

where

Q (J/s) = heat transferred
H (m) = tube length
d_i(m) = tube inner diameter
d_a(m) = tube outer diameter
ϑ (°C) = effective overall temperature difference between heating medium and solution on either side of the tube wall
λ_W(W m^{-1} K^{-1}) = coefficient of thermal conductivity
α_V(W m^{-2} K^{-1}) = film coefficient of heat transfer for the tube outer surface
α_L(W m^{-2} K^{-1}) = film coefficient of heat transfer for the tube inner surface
K (W m^{-2} K^{-1}) = overall coefficient of heat transfer based on the average tube surface
s_W(m) = tube wall thickness

For thin-walled heating surface of thickness s_W, the following relationship results from Equation (48):

$$Q = \frac{1}{\frac{1}{\alpha_V} + \frac{s_W}{\lambda_W} + \frac{1}{\alpha_L}} \cdot A \cdot \vartheta \tag{49}$$

For the case of an exchanger tube bundle with R tubes, Equation (50) gives the total heat transfer surface:

$$A = R \cdot \frac{d_i + d_a}{2} \cdot \pi \cdot H \tag{50}$$

Then the expression for the overall heat transfer coefficient K is:

$$K = \frac{1}{\frac{1}{\alpha_V} + \frac{s_W}{\lambda_W} + \frac{1}{\alpha_L}} \tag{51}$$

Equation (51) applies precisely for flat surfaces only. But it may be used for round, thin-walled evaporator tubes with sufficient accuracy for equipment design.

If scaling or severe salting of the tube surface is expected, a thermal conductivity resistance in the term s_W/λ_W must be incorporated in determining the overall heat transfer coefficient K. Scaling or salting of the tube surface usually causes severe deterioration in the heat transfer.

The sum of the thermal conductivity resistances s_W/λ_W and the reciprocals of the film coefficients $1/\alpha_V$ and $1/\alpha_L$ makes up the total thermal resistance $1/K$. Tables of values for λ_W are available for the usual materials used in exchanger fabrication. In general, however, one must calculate the film coefficients α_V and α_L, since they are dependent on the mechanical design of the particular unit and on the operating conditions. For rough estimations, a variety of average K values may be used.

Steam is the medium most often used for heating evaporators. The physical phenomenon is then a matter of virtually static, saturated steam condensing as a liquid film on a vertical heating surface with the height H. If this flows continuously as a laminar film on the wall, then one may calculate the film heat-transfer coefficient α_V for the steady state in accordance with the Nusselt film theory:

$$\alpha_V = K \sqrt[4]{\frac{\Delta h_V \cdot \varrho_K^2 \cdot \lambda_K^3 \cdot g}{\eta_K \cdot \vartheta_H \cdot H}} \tag{52}$$

where

Δh_v(J/kg) = latent heat of condensation of steam at the pressure in question
ϱ_K(kg/m^3) = density of the condensate film
λ_K(W m^{-1} K^{-1}) = thermal conductivity of the condensate film
η_K(kg m^{-1} s^{-1}) = dynamic viscosity of the condensate film

The value of the constant k in the equation is 0.943.

Frequently, however, the steam condenses on the wall in the form of droplets. For example, this may occur when the steam or the heat-transfer surface is contaminated with oil or grease. Then, the film heat-transfer coefficients may be about six to twenty times higher than those with pure film condensation. If the assumption that the steam is practically static is not admissible, then Equation (52) gives low numerical values. According to the Nusselt theory, the equation for the film heat-transfer coefficient α_V for flowing condensing saturated steam with the velocity u_V then changes in accordance with the relationship:

$$\alpha_V = K \sqrt[3]{\frac{C \cdot \Delta h_V \cdot \varrho_K \cdot \varrho_V \cdot \lambda_K^2 \cdot u_V^2 \cdot g}{\eta_K \cdot \vartheta_V \cdot H}} \tag{53}$$

Here, the constant is $C = 10.5 \times 10^{-4}$.

The saturated steam available in industry is usually contaminated with air or inert gases. This air impurity can lead to a severe adverse effect on the film heat transfer coefficient for condensation in heating equipment. Therefore, an important requirement in fabricating the evaporator heating element is an absolutely reliable means for venting inert gases from the steam chest.

The basic types of industrial evaporation equipment represented in Figure 11, in general, show varied high heat transfer coefficients for the boiling liquid. They vary in accordance with the operating conditions and the principles governing the particular boiling behavior. Without doubt, the nature and conditions for the hydraulic flow regime are determining factors for the film heat-transfer coefficient on the liquid side of the exchanger. Thus, in the case of recirculating and once-through vaporizers, the proportion of vapor admixed in the boiling liquid is a decisive factor. Then, according to whether nucleate or convection boiling takes place, distinct laws of physics and physical chemistry determine the film heat-transfer coefficient. The film heat-transfer coefficient α_L of the vaporizing solution can be represented as a function of characteristic design parameters of the equipment, values for the physical properties of the fluids, and the operating conditions. These are combined into various dimensionless numbers well known in the general heat-transfer and fluid-flow theory.

For the most part, industry is concerned with vaporizing flowing or moving liquids. For va-

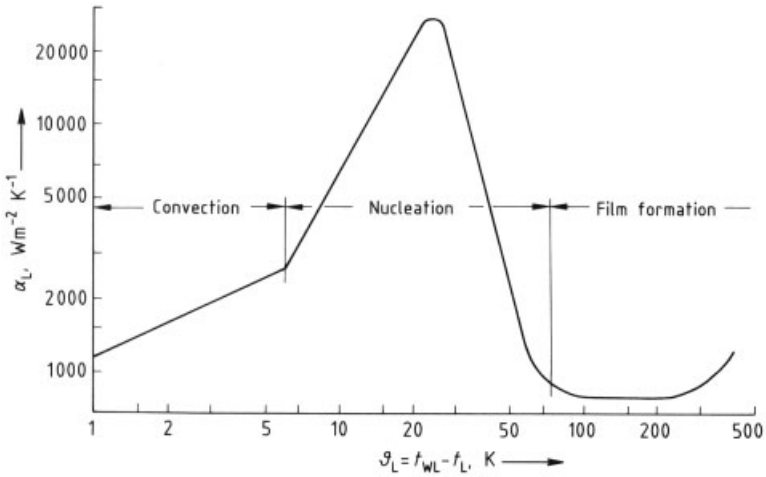

Figure 12. The film heat-transfer coefficient for boiling water α_L as a function of temperature difference ϑ_L

porizing liquids at rest, the film heat-transfer coefficient of the boiling liquid α_L characteristically depends on the temperature difference ϑ_L between the wall temperature on the liquid side t_{WL} and the boiling temperature t_C. Figure 12 shows, for example, the behavior for boiling water at 0.1 MPa. Only occasionally is this behavior found in vaporizing flowing liquids in industrial equipment.

Representing heat transfer is made more difficult for vaporization in that one must deal, in general, with a combined single phase – two phase flow condition. Therefore, predictive calculation of the film heat-transfer coefficient according to the natural laws governing single-phase flow is limited only to special cases. For the most part, semi-empirical relationships are used. Equation (54) is a dimensionless relationship from KIRSCHBAUM, making possible predictive calculation of the film heat-transfer coefficient α_L for boiling liquids in natural circulation vaporizers, as in Figures 11 A and 11 B. It agrees well with measured values.

$$\frac{\alpha_L \cdot \sigma_L}{\lambda_L \cdot p} = c \cdot \frac{\vartheta_L \cdot c_L}{\Delta h} \left(\frac{\varrho_L}{10^3 \varrho_V}\right)^{0.5} \left(\frac{\eta_{H_2O}}{\eta_L}\right)^{0.25} \quad (54)$$

The factor c depends on the apparent liquid level h. The parameter h is the liquid holdup or liquid inventory in an exchanger tube, which is defined as the height of a column of clear liquid, expressed as a percentage of the tube length measured upward from the lower tube sheet of the exchanger, which would appear in a manometer connected between the lower channel of the exchanger and the liquid free vapor separation chamber. Empirically, one finds $c = 0.24$ at an apparent level $h = 75$ % and $c = 0.37$ at an apparent level $h = 40$ %.

Other nomenclature in Equation (54) is as follows:

σ_L(kg/m) = surface tension of the boiling liquid in contact with its vapor
λ_L(W m^{-1} K^{-1}) = thermal conductivity of the boiling liquid
c_L(J kg^{-1} K^{-1}) = liquid specific heat
Δh (J/kg) = latent heat of vaporization of the liquid
ϱ_L(kg/m^3) = density of the liquid
ϱ_V(kg/m^3) = density of the vapor in equilibrium with the liquid
η_L (kg m^{-1} s^{-1}) = dynamic viscosity of the fluid

In predictive calculation of the film heat-transfer coefficient α_L using Equation (54) the selected temperature difference ϑ_L should not be smaller than the least temperature difference necessary to stimulate natural circulation. This limiting temperature difference necessary to stimulate natural circulation may be taken from empirically derived limit curves. These curves are also necessary for design of multistage evaporators as they identify the relationships that limit the number of stages.

Figure 13 reproduces examples of such limit curves. They show the dependence of the appar-

ent liquid level h on the temperature difference ϑ_L between the inside tube wall temperature t_{WL} and the boiling temperature t_C with t_C as a parameter. These limit curves are typical of those showing operating points at which natural circulation just ceases. At operating conditions lying in the area below the curves, natural circulation will not automatically initiate nor continue if initiated at other conditions.

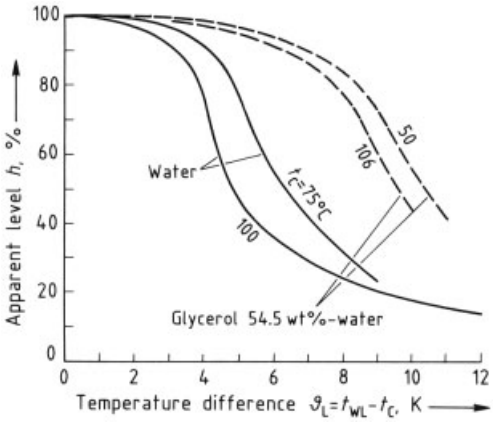

Figure 13. Natural circulation limit curves at various boiling temperatures

For evaporating solutions having severely corrosive characteristics, systems using comparatively expensive materials of construction for the heat-exchange surfaces must be designed. This sets a requirement for designing relatively small evaporator heat-exchanger elements. In turn, this may lead to a need for increasing the velocity of the recirculating, vaporizing liquid.

The Nusselt – Kraussold equation may be used for calculating the inside film heat transfer coefficient α_L for boiling liquid in tubes in forced circulation evaporators:

$$\alpha_L = 0.032 \frac{\lambda_L}{d_i} \left(\frac{u_L \cdot d_i \varrho_L}{\eta_i} \right)^{0.8} \left(\frac{c_L \cdot \eta_L \varrho_L}{\lambda_L} \right)^{0.37} \left(\frac{H}{d_i} \right)^{0.054} \tag{55}$$

Equation (55) applies when the liquid-side temperature difference $\vartheta_L = (t_{WL} - t_L) \leq 6\,°C$ and the Reynolds number exceeds 10 000; H denotes the tube length and d_i its inner diameter. Table 2 contrasts ranges of film heat-transfer coefficients commonly experienced in natural and forced circulation evaporators with one another.

Table 2. Overall coefficient of heat transfer for some modes of operation

Overall coefficient of heat transfer	k, W m^{-2} K^{-1}
Natural circulation, viscous solution	350 – 900
Natural circulation, mobile solution	600 – 1800
Forced circulation	900 – 2900

As compared to evaporators with vertical tubes, units with inclined tubes have the advantage that the liquid holdup can be held relatively low in proportion to the evaporator capacity, achieving the desired concentration for the recycle liquid in short residence times. Therefore, evaporation of temperature-sensitive solutions or materials inclined to discolor or precipitate deposits even at still higher temperatures is made possible. A further advantage over evaporators with vertical tube arrangements is that the construction results in a lesser liquid hydrostatic pressure in the tubes. In turn, a much lower boiling point rise results, which causes a substantially lesser deterioration in the total temperature difference in the evaporator system. Heat-transfer measurements have been made on operating evaporators not only with vertical but also inclined vaporizer tube arrangements. These show that inclined tube evaporators have substantially higher heat-transfer coefficients than those with vertical tubes. Figure 14 reproduces these measurements.

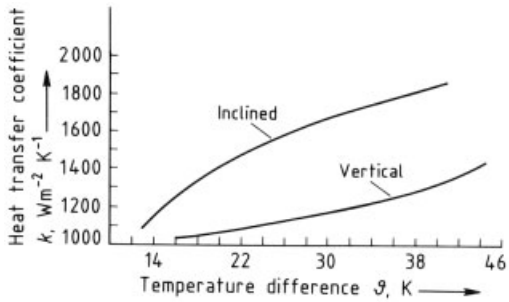

Figure 14. Heat-transfer coefficient in vertical and inclined tubes

Figure 11 D shows that kettle-type units, with horizontal heat exchanger tubes immersed in the liquid, have an advantage that a still lesser level of liquid suffices to envelop the tubes and wet them completely. As a result, the vaporizing liquid is under a low hydrostatic pressure and at a boiling temperature only slightly above

the saturated vapor temperature obtaining at the pressure in the vapor separation chamber. This type of unit is also able to operate comparatively well in vaporizing solutions that foam severely. The concept generally permits greater area for vapor – liquid release and allows better separation of entrained liquid droplets resulting from the bubbling and foaming than evaporators with vertical tube arrangements.

At small temperature differences, in kettle units, free convection is the heat-transfer mechanism on the liquid side. This initiates vaporization at the surface with only slight superheating of the liquid. Therefore, the laws governing heat transfer by free convection apply. At higher temperature differences, nucleate boiling begins. For both boiling regimes, the film heat-transfer coefficient α_L depends on the temperature difference ϑ_L and is characterized by the relationship:

$$\alpha_L = C \cdot \vartheta_L^n \tag{56}$$

In the free convection regime, $n = 0.33$. For nucleate boiling, the value of n is in the range 2 – 5. The liquid film heat-transfer coefficients for submerged tube evaporators, however, are not as favorable as, for example, for circulating vaporizers with vertical tubes. Figure 15 shows this comparison.

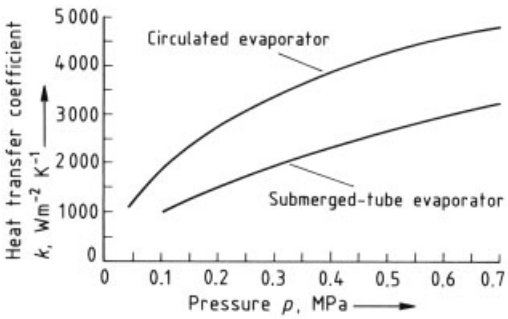

Figure 15. Heat-transfer coefficient in recirculating and submerged tube evaporators

In addition to the construction shown in Figure 11 D, horizontal heat exchangers having two tube sheets are also in use. This style of evaporator also finds application as a trickle film vaporizer, for example, for seawater desalting or in the refrigeration industry. In these, there is a liquid sump underneath the tube bundle. A pump removes solution from the sump and conveys it to a device that distributes it over the outer surface of the tube bundle. Here it vaporizes from a thin liquid film. This principle of operation totally avoids the disadvantage of any elevation of the boiling point by liquid static pressure. Beyond that, is the advantage that the liquid inventory required for operating the evaporator is comparatively small.

Additionally, related also to a low liquid inventory, the residence time of the feed material in the unit can be a deciding factor in the choice of evaporator type. Many evaporating solutions incline toward undesirable chemical or physical changes if they are exposed too long to temperatures too high. To be able to vaporize even such solutions gently requires evaporators having low residence time for the evaporating material. At present, with the normal recirculating evaporators, to achieve the desired concentration the liquid undergoes repeated recycling through the equipment. Moreover, it is exposed to the evaporating temperature for a comparatively long time. Thus, for concentrating thermally unstable solutions, a type of evaporator permitting more gentle treatment of the solution should be installed: that is, one in which the liquid reaches its specified final concentration after only one pass through the vaporizer tubes. Designs are available in which the product stream flows not only from bottom to top, Figure 11 E, but also from top to bottom, Figure 11 F. If the space available does not permit the required height, then, in any case, the desired final concentration may be attained by linking several units together in series each having shorter vaporizing tubes.

For units with upward product flow, if the superficial velocities (based on the inside cross-sectional area of a tube of diameter d_i) not only of the liquid phase, index L, but also of the vapor phase formed by boiling, index V, are known, the film heat-transfer coefficient for liquid boiling may be calculated by a relation developed by COULSON and MCNERLY:

$$\alpha_L = \frac{1.3 + 128 d_i}{d_i} \cdot \lambda_L \cdot \left(\frac{c_L \eta_L \varrho_L}{\lambda_L}\right)^{0.9} \cdot \left(\frac{u_V d_i \varrho_V}{\eta_V}\right)^{0.34} \cdot \left(\frac{u_L d_i \varrho_L}{\eta_L}\right)^{0.23} \cdot \left(\frac{\varrho_L}{\varrho_V}\right)^{0.25} \frac{\eta_V}{\eta_L} \tag{57}$$

A principal field for once-through evaporators is the concentration of solutions of alkalis and

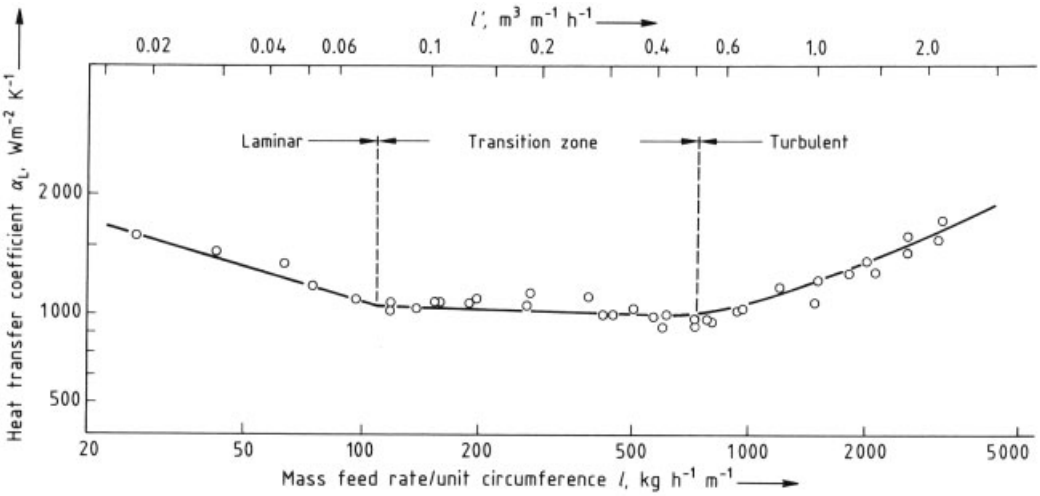

Figure 16. Heat-transfer coefficient for surface evaporation of refrigerant R11 in a falling film

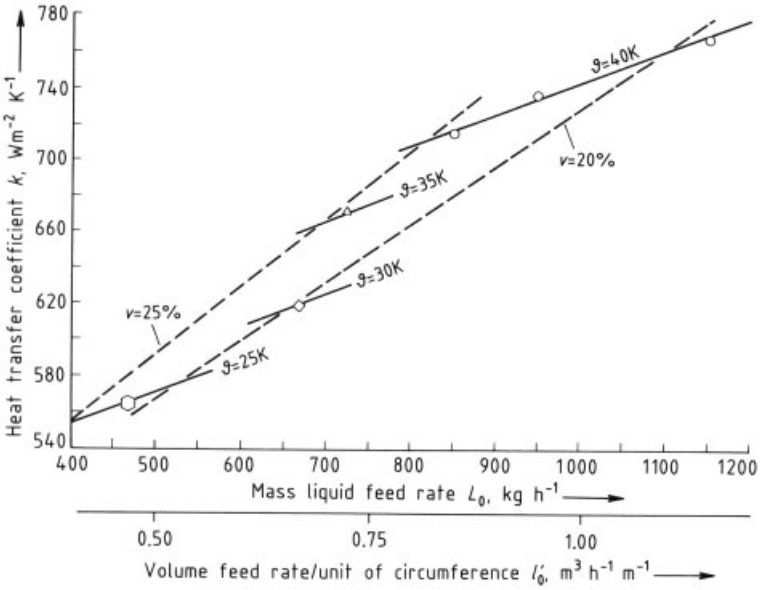

Figure 17. Heat-transfer coefficient for a falling film evaporator as a function of the feed rate
v = feed evaporated, % and ϑ = temperature difference between the steam and the boiling liquid

salts: NaOH, KOH, Na_2S, and $CaCl_2$. In these applications, frequently, two-stages systems are installed with total utilization of the solvent vapors. The first stage is a once-through vaporizer, preconcentrating the solution from about 50 – 75 % by weight salts, while the main thickening to 99 % solids takes place in a falling film evaporator.

The evaporating liquid flows downward in a falling film evaporator. Liquid, usually preheated, enters the unit above the upper tubesheet. Suitable devices distribute the liquid uniformly among the outside heated evaporator tubes. The liquid then flows downward with partial vaporization as a thin film on the inner tube walls. The momentum of the likewise downward flowing evaporated vapors assists in accelerating liquid downflow, reducing liquid film thickness. This means of operation results in a substantially lesser liquid residence time even than that

for the once-through vaporizer. In single-stage systems, the residence time is less than a minute, in comparison to the recirculating evaporators, an advantage in evaporating heat-sensitive solutions.

In determining the film heat-transfer coefficients on the liquid side of falling film evaporators, the physical laws valid for trickling films are valid if conservative. In so doing, the purely laminar region, the pseudo laminar or wavy laminar region, the transition region, and, finally, the turbulent region are differentiated. In these differing regions, heat-transfer phenomena obey different laws for the trickling film since the average film thickness, the film configuration, and the discharge velocity are different. The measured results shown in Figure 16 illustrate this. The data apply to vaporizing a trickling film of refrigerant R11 on a clean surface with fully developed flow. Figure 16 shows the dependence of the film heat-transfer coefficient of the vaporizing, trickling film α_L on the specific liquid loading l (kg m^{-1} h^{-1}) or l' (m^3 m^{-1} h^{-1}), the amount of liquid that flows down 1 m of tube circumference in 1 h. Since the principal liquid side heat-transfer resistance itself lies in the trickling film (for purely laminar flow, the region on the left of Figure 16) Equation (58) permits determination of α_L:

$$\alpha_L = \frac{\Delta h \varrho_L^2}{3\eta_L \vartheta_L H}\left(s_0^3 - s_u^3\right) \tag{58}$$

The film thickness, so at the feed entry position, i.e., the upper end of the tube, for laminar film velocity distribution, is given by the Hagen – Poiseuille law when the quantity of liquid feed L_o(kg/h) fed to each evaporator tube with inside diameter d_i is known:

$$s_0 = \sqrt[3]{\frac{3\eta_L L_0}{\varrho_L^2 \pi d_i 3600}} = \sqrt[3]{\frac{3\eta_L l_0}{\varrho^2 \cdot 3600}} \tag{59}$$

Equation (60) allows calculation of the film thickness s_u at the lower end of the tube with length H(m), i.e., at the outlet position:

$$s_u = \sqrt[4]{s_D^4 - \frac{4\lambda_L \eta_L \vartheta_L H}{\Delta h \varrho_L^2}} \tag{60}$$

In Equations (58) – (60), the liquid-film material properties are taken at the average temperature of the trickling film, taking into account the temperature drop ϑ_L between the tube inner wall and the surface of the film.

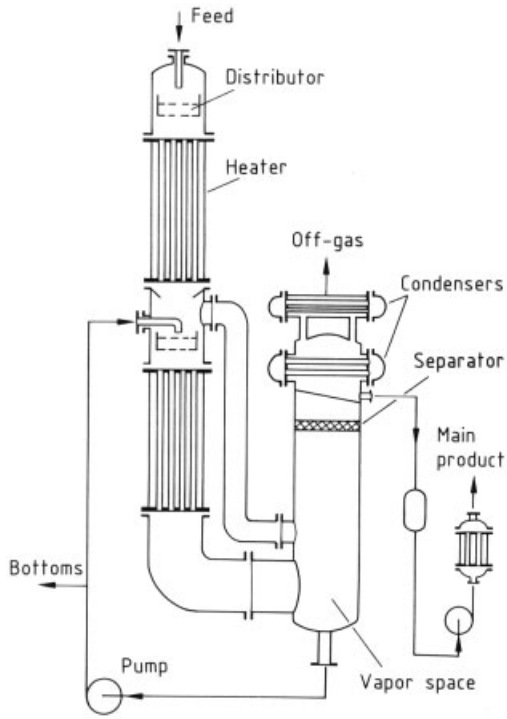

Figure 18. Falling film evaporator with circulating pump

An important prerequisite for smooth operation of a falling film evaporator and for effective heat transfer is complete wetting of the heating surface. This means that, in design, the specific liquid loading l'_0 (m^3 h^{-1} m^{-1} = m^2/h) must be chosen with care so that it is not too small. To achieve sufficiently high film heat-transfer coefficients, Figure 17 shows that the specific feed quantity at the upper end of the tube l'_0 should be:

$$l'_0 > 1.0 - 1.5\,\mathrm{m^3 h^{-1} m^{-1}} = \mathrm{m^2 h^{-1}} \tag{61}$$

In industrial systems, in special cases, dealing with evaporative degassing, for example, smaller liquid loading values are chosen as well, as low as $l'_0 = 0.3$. But otherwise, l'_0 values frequently lie above 2 m^3 h^{-1} m^{-1}.

Using a recycle pump for recycling liquid to achieve complete tube surface wetting is common, for example, for falling film evaporators in the edible oils industry. Figure 18 shows a scheme for thermally separating free fatty acids

from nonvolatile residue. Since such systems operate at high vacuum, less than 1 kPa, one must take care in design that the vapors developed do not cause an inadmissably high pressure drop. By dividing the heat-exchange surface into more than one bundle and removing the device for separated vapor withdrawal from the exchangers, the vapor hydraulic losses can be held correspondingly low.

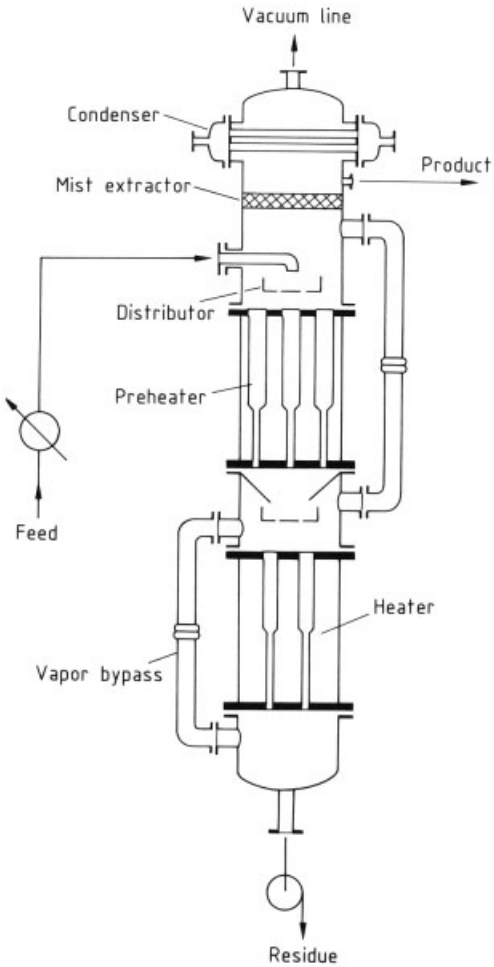

Figure 19. Falling film evaporator with partial countercurrent and partial cocurrent flow of liquid and vapor

In principle, it is also possible to design units in which vapor and liquid flow partially cocurrently and partially countercurrently to one another, as in Figure 19. By varying the number of tubes in the respective separate exchanger bundles or changing the tube diameter along the direction of liquid flow, the specific liquid loading on the tube circumference l'_0 can be maintained high enough that the system can operate without a recycle pump. This design has proven to be advantageous as the final stage in fatty acid evaporation systems.

Table 3 lists parametes, including heat-transfer coefficients, for some falling film evaporators from operating industrial systems.

Relatively small temperature differences between heating medium and evaporating liquid, as little as 4 °C, are possible in operating falling film evaporators. Moreover, the design excludes superheating the liquid due to static liquid head. Thus, at a given operating pressure, the boiling temperature is lower than in conventional evaporators. At a liquid viscosity of about 1 kg/ms, the lower limit of pressure is an operating vacuum of about 4 kPa.

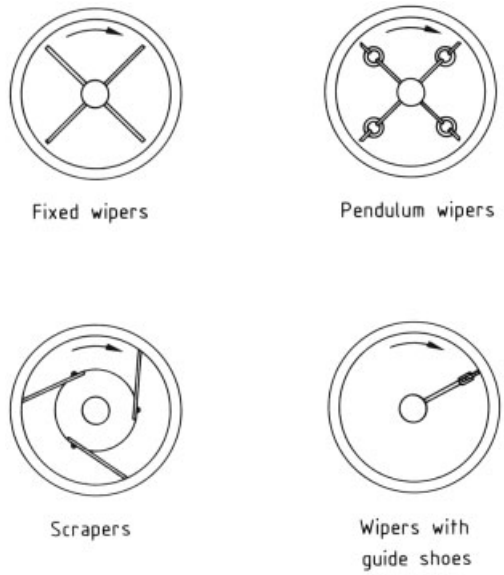

Figure 20. Wiper systems for thin film evaporators

Figure 11 G illustrates a thin film evaporator with rotating internal element. These are suited for heat-sensitive products requiring low evaporator operating pressures, down to 0.1 kPa, and having liquid viscosities far above 1 Pa s. Mechanical wipers rotate inside an externally heated cylindrical shell, as shown schematically in Figure 20. They produce sufficiently intensive agitation to develop thin films even with high viscosity liquids. The film thickness varies in the range 0.5 – 2 mm depending on the design and the speed of rotation. The product residence

Table 3. Coefficient of heat transfer for some industrial applications of falling film evaporators

Falling film evaporator	Styrene distillation	Fatty acid distillation	Fatty acid fractionation
Tube dimensions, mm *	38.1×2.64×3000	76.1×2.9×3000	76.1×2.9×3000
Area, m^2	214.9	43.6	35
Steam temperature, °C	112.7	251	266
Boiling point, °C	98.0	230	242
Heat-transfer coefficient, $W\ m^{-2}\ K^{-1}$	670	400	400

* inner diameter (d_i) × wall thickness (s_w) × height (H)

time, depending on the loading conditions, may be even less than that for falling film evaporators.

The liquid to be vaporized enters the unit above the heated wall and distributes uniformly around the circumference. The wiper system provides for developing a thin liquid film which flows downward along the hot wall under the influence of gravity while evaporating to a higher concentration. Designs using self-aligning, pendulum wiper systems and those using sliding guide shoes have potential for thickening solutions of solids to the point that the end product leaves the evaporator in powder form.

Inherent in the design concept is that these evaporators must operate with a comparatively large temperature difference. Now, since the thin films produced by the rotating wiper systems result in residence times less than 1 min for the material being concentrated, substantially higher temperatures may be used in evaporating solutions of temperature-sensitive substances than, for example, with circulating vaporizers.

Equation (62), a result of the experimental investigations of DIETER, permits calculating approximate values of the film heat-transfer coefficients α_L for thin-film evaporators with rotating internal elements:

$$\alpha_L = K \cdot \sqrt[3]{d_i \cdot n_r} \cdot \frac{\lambda_L}{\eta_L^{1/3}} \tag{62}$$

The nomenclature for Equations (62) – (64) is as follows:

K = a constant with the numerical value 500
d_i(mm) = internal diameter of the heating cylinder wall
n_r(min^{-1}) = rotor rotational frequency
λ_L($W\ m^{-1}\ K^{-1}$) = thermal conductivity of the liquid film
η_L($kg\ m^{-1}\ s^{-1}$) = dynamic viscosity of the liquid film
u_w(m/s) = wiper circumferential speed
d_w(m) = wiper diameter
f_w(s^{-1}) = wiper frequency
H (m) = length of the heated cylinder wall
D (kg/h) = evaporator capacity in terms of solvent evaporated
d ($kg\ m^{-2}\ h^{-1}$) = specific evaporator capacity

Equation (63) gives the rotor rotational frequency given the wiper circumferential speed:

$$\eta_r = \frac{60}{\pi} \cdot \frac{u_w}{d_w} \tag{63}$$

Recommended values for u_w are 8 – 9 m/s for fixed and 3.5 – 4 m/s for movable wipers. For thin film evaporators of industrial size, wiper frequencies f_w from 15 to 20 have shown favorable operating experience. Then, the wiper frequency results from u_w, f_w, and the wiper diameter d_w.

Figure 21 makes it possible to estimate the expected heat-transfer coefficients for some uses of industrial interest. Equation (64) defines the specific evaporator capacity d. Equation (65) shows the relationship between d, the average heat transfer coefficient k, and the assigned total temperature difference between the heating medium and the vaporizing liquid ϑ. Given the solvent latent heat of vaporization Δh and temperature difference ϑ sets the specific evaporator capacity d:

$$d = \frac{D}{d_i \pi H} \tag{64}$$

$$d = k \cdot \frac{\vartheta}{\Delta h} \tag{65}$$

Thin film evaporators with rotating internal elements are especially suited for temperature-sensitive materials that decompose easily or change chemically during evaporation in conventional equipment. Condensation and polymerization are among possible chemical

changes. Applications include concentrating substances whose taste and active ingredients are not permitted to degrade. Vaporizing fatty acids or softeners, for example, requires low vacuum on the product side and high temperature on the heating medium side. For such cases, the heating medium often is Dowtherm, an azeotropic mixture of diphenyl and diphenyl oxide, or HT-OIL-C. Further fields of application are concentration of fruit juices, milk, yeast, pectin solutions, dyestuff solutions, gelatin solutions, glues, gums, plant extracts, recovery of drug extracts and serums, distilling solutions of organic solvents or mixtures of such with water, and partial separation of multicomponent mixtures.

However, based on industrial experience, it is known that thin film evaporators with rotating wipers are not successful in all cases for all thermally unstable materials. The reason is that even product residence times of one minute are too long. If product residence times of only one second are required for liquids of high viscosity, 20 kg/ms and more, for example, then conventional evaporators are not applicable or acceptable only with limitations on product yield and quality. A completely different design concept (Fig. 11 H) enables successful solution of even these vaporizing problems. The advantage inherent in this design is extremely mild vaporizing behavior. For example, the residence time in the actual vaporizing zone on the heat-transfer surface may even amount to only fractions of a second.

In this equipment, as in many designs of molecular distillation apparatus, the vaporizing surface rotates. Under the influence of high centrifugal forces resulting from correspondingly high rotational frequency, the liquid spreads out in an extremely thin film moving at high velocity. Varying the feed rate and the rotor rotational frequency up to 1600 revolutions per minute permits stepfree adjustment of the film thickness within defined limits. On the average, under normal operating conditions, even with comparatively viscous liquids, the film thickness amounts to only ca. 0.1 mm. With rotational speed adjusted to feed rate conditions and liquid viscosity, the film thickness can successfully be held low in such equipment, even at relatively high liquid rates. This leads to comparatively high radial film velocities and to relatively high heat-transfer coefficients at the particular position considered on the vaporizer rotor at the varying cone diameter d_k.

In addition, this principle of operation also creates favorable conditions for heat transfer on the heating medium side of the vaporizer unit. If steam is that heating medium, then it condenses predominantly in the form of droplets whose size decreases with increasing rotational frequency. Now, since centrifugal force flings the condensate drops off the condensing surface immediately after they are formed, the heat-transfer resistance is substantially less than in conventional condensing systems. Consequently, the heat-transer coefficients on the condensing steam side are about an order of magnitude higher than for pure film condensation of the heating steam.

Operating thin film vaporizers having the design concept shown in Figure 11 H best fulfills the prerequisites for good heat transfer on both sides of the heat-transfer wall of the vaporizer unit. Heat-transfer coefficients k are a factor of about two to four times higher, depending on the thermal conductivity resistance of the vaporizer wall, than, for example, those for conventional falling film vaporizers. This results in less heating surface or even using a comparatively small temperature difference between heating medium and product and thus a lesser wall temperature on the product side. The high realizable heat transfer coefficients result in high specific evaporating capacity and thus comparatively small equipment volume.

The heat-transfer coefficients vary over a broad range in accordance with the vaporizer-rotor frequency, the physical property values of the feed material at the operating conditions, as well as the state of the heating medium, heating surface loading, and capacity. Numerical values for k range between 500 and 10 000 W m^{-2} K^{-1}. Total temperature differences ϑ may lie between 5 and 80 °C. Whereas the high heat-transfer coefficients given in Table 4 are more typical of dilute solutions, heat-transfer coefficients calculated for vaporizing liquids with high solids concentration would be an order of magnitude lower. But even for such evaporation problems, the figures given in Table 5 show the unmistakable advantages of the vaporizer with rotating heating surface, attesting convincingly to its far superior heat-transfer performance.

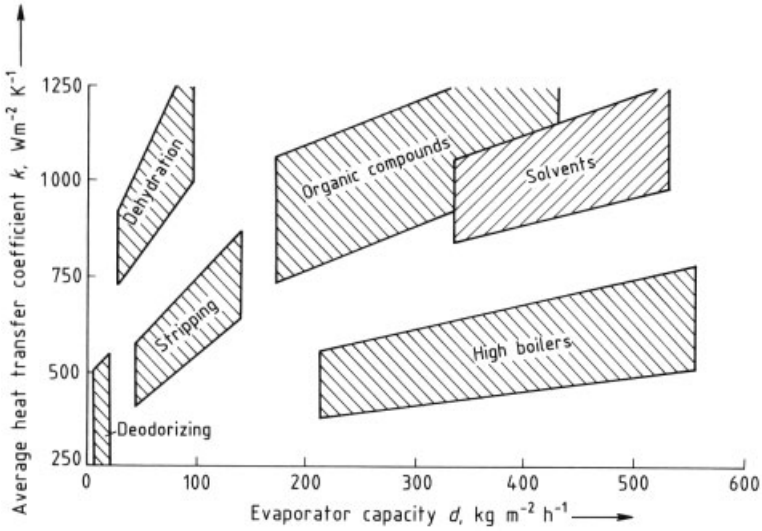

Figure 21. Heat-transfer coefficients for thin film evaporators for some applications as a function of the specific evaporator

Table 5. Comparative design parameters for falling film and thin film evaporators for liquids with high solids content

Magnitude	Falling-film evaporator		Thin-film evaporators		
			with rotating wipers		with rotating heaters
			Luwa	Rotafilm	
A, m^2	0.54		0.125	0.2	0.2
p, kPa	2.0	1.3	0.26	0.26	0.53
ϑ, °C	37	31	74	40	40
k, W m^{-2} K^{-1}	68	57	145	175	530

Table 4. Comparative design parameters for falling film and thin film evaporators

Magnitude	Falling film	Rotating heater surfaces
s_W, m		0.002
λ_W, W m^{-1} K^{-1}		16.3
t_L, °C		99
k, W m^{-2} K^{-1}	730	2 900
ϑ, °C	40	10
α_H, W m^{-2} K^{-1}	5 800	17 500
α_L, W m^{-2} K^{-1}	925	6 100
t_{WL}, °C	130.4	103.8

Installation of a thin film evaporator with rotating heating surface is advantageous in all cases for applications which require, in operation, the lowest possible wall temperature on the product side. It is a relatively versatile device finding application in the chemical and pharmaceutical industries, the edible oil industry, and not the least in the foodstuff industry, for small to medium capacities.

At fixed operating conditions, the transfer-rate coefficient k and the difference between the temperature of the heating medium t_H and the boiling temperature $(t+\tau)$ set the size of the heat-transfer surface A of the evaporator heating element for the heat-transfer duty Q through the equation:

$$A = \frac{Q}{k\,\{t_H - (t+\tau)\}} \tag{66}$$

Correspondingly, Equation (67) determines the size of the required heat-transfer surface for a given evaporator stage of a multistage system, where Q_n is the stage duty. The temperature difference $\{t_{n-1} - (t_n + \tau_n)\}$ refers to the difference between the solvent saturation temperature corresponding to the pressure in the previous stage and the boiling temperature in the stage under consideration:

$$A_n = \frac{Q_n}{k_n \{t_{n-1} - (t_n + \tau_n)\}} \tag{67}$$

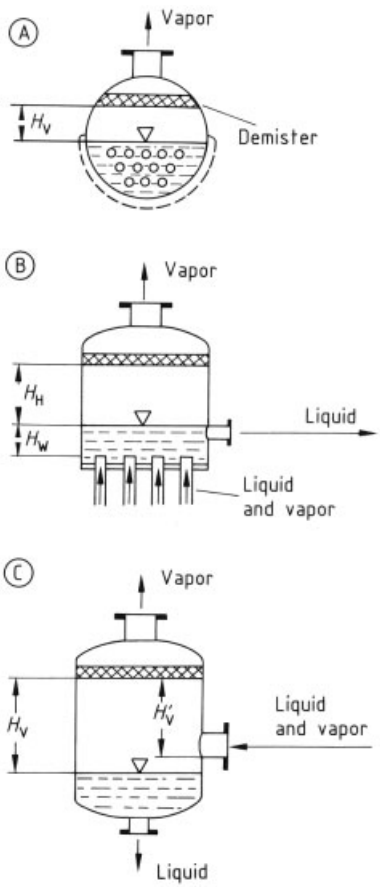

Figure 22. Arrangement of the vapor separation chamber in various types of evaporators

4.2. The Vapor Separation Chamber

The vapor separation chamber, or vapor head, has the job of separating the two phase, vapor – liquid mixture that flows into it from the evaporator heater element into its components. The separated vapor should be free of entrained liquid particles. This will, insofar as is possible, prevent fouling the condensing surface of the vapor condenser in single-stage systems or of the vaporizing surface of the succeeding stage in mul-tistage systems. For its part, the liquid leaving the vapor separation chamber should be as completely free of vapor as is possible.

The orientation of the vapor separation head relative to the evaporator heating element, its design features, and the way in which the vapor – liquid mixture is introduced, determine the degree to which the unit will fulfill the aforementioned requirements. Thus, among the basic conceptual evaporator designs shown in Figure 11, there are merely three principal vapor separation head configuration variations. Figure 22 illustrates these.

The arrangement in Figure 22 A refers to submerged tube vaporizers called kettles or for simple evaporator drums heated through the walls. In these constructions, the vaporizing element and the vapor separation chamber form a single unit. The design in Figure 22 B illustrates the vapor separation chamber arrangement corresponding to the Robert or calandria evaporator, recirculating evaporators with external liquid downcomers, as well as to once-through vaporizers. In these, the vapor – liquid mixture enters the vapor separation head from below, directly under a pool of accumulated liquid. Finally, the design in Figure 22 C shows the configuration for all recirculating evaporators with externally arranged heat exchanger elements, falling film evaporators with vapor separation chamber below the exchanger, as well as vaporizers flashing preheated solutions. This last arrangement has the feature that the vapor – liquid mixture enters the vapor head above the pool of accumulated liquid. This allows the level of accumulated liquid to be set comparatively low.

Thin film evaporators with rotating internal elements, in general, have no vapor separation chamber in the same sense as the designs shown in Figures 22 A, B, and C. However, they do possess special internal devices for removing entrained liquid droplets from the vapor.

The vapor separation head is so dimensioned that liquid contained in the separated vapor phase is limited to just the minimum physically attainable. This is in the form of small, entrained liquid particles corresponding in size to equilibrium among the gravity, inertial, surface tension, viscous, and drag forces extant. Installing liquid separators as illustrated in Figure 23 achieve further purification of the exiting vapor from residual liquid. Setting the dimensions of the vapor separation chamber determines the specific vapor volume per hour per cubic meter of

chamber volume that a given head can handle at the least liquid content. Empirical Equations (68) – (70) serve to calculate this flow of vapor v in accord with the three characteristic vapor separation chamber configurations A, B, and C of Figure 22. These apply to a vapor separation head pressure of 100 kPa.

Equation (68) applies to the case shown in Figure 22 A where H_v (m) is the height of the vapor space measured from the liquid surface to the demister pad or droplet separator. The most favorable vapor separation capacity v_a ($m^3\ m^{-3}\ h^{-1} = h^{-1}$) is:

$$v_a = 500 + \frac{40}{0.01 + (H_V)^{3/2}} \tag{68}$$

Extending Equation (68) gives Equation (69) applying to Case B where H_w (m) is the height of the accumulated liquid level in the vapor head and H_v and v_a are as above. Then the most favorable vapor separation capacity v_b ($m^3\ m^{-3}\ h^{-1} = h^{-1}$) is:

$$v_b = v_a + \frac{1}{(1 + 1.1 \cdot H_W)^4}\,(4500 - v_a) \tag{69}$$

For Case C, the most favorable vapor separation capacity v_c ($m^3\ m^{-3}\ h^{-1} = h^{-1}$) is constant:

$$(v_c)_{H_v > H'_v} = 7300 \tag{70}$$

For operating pressures different from 100 kPa, Equation (71), with the help of Equations (68) – (70) and the pressure factor f read from Figure 24, permits converting the most favorable vapor separation capacities v_a, v_b, and v_c to the revised pressure basis:

$$(v_{a,b,c}) = f \cdot (v_{a,b,c})_{p=1} \tag{71}$$

For a given hourly evaporated vapor flow D (kg/h) with density ϱ_D (kg/m^3), Equation (72) determines the required height of the vapor separation head H_v, where H_v is the distance measured from the liquid surface to the demister pad or droplet separator for removing entrained liquid particles from the vapor stream:

$$H_V = \frac{1}{d_B^2 \pi / 4} \cdot \frac{D}{\varrho_D\,(v_{a,b,c})_p} \tag{72}$$

For the vapor separation chamber-free crosssectional area and water vapor at atmospheric pressure, allowable vapor velocities u_D of 1 m/s are the usual in practice. At lower pressure, the equations give even higher vapor velocities. Then Equation (73) sets the diameter d_B of the vapor separation chamber:

$$d_B = \sqrt{\frac{4D}{\pi u_D \varrho_D \cdot 3600}} \tag{73}$$

Since the liquid density is practically independent of pressure, geometrically similar vapor separation heads are then about equal in their separation effect when operating at equal velocity head. When the pressure differs from 100 kPa, the following relationship applies for setting the vapor velocities u_D in the vapor separation chamber:

$$(u_D^2 \cdot \varrho_D)_{p_1} = (u_D^2 \cdot \varrho_D)_{p_2} = (u_D^2 \cdot \varrho_D)_{p_n} \tag{74}$$

Operation of industrial evaporator systems shows that, even at comparatively low vapor velocities in the vapor separation chamber, some liquid is always entrained by the flowing vapor if no special measures are taken for their separation. Even using liquid separators, it is not always possible to remove the smallest liquid droplets. In general, relying on the force of gravity alone achieves a separation effect up to 70 %.

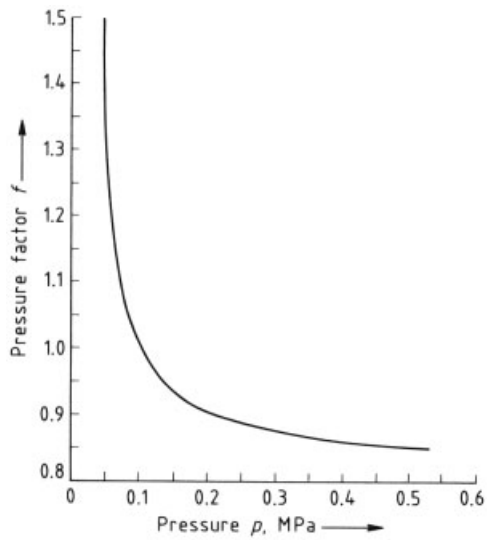

Figure 24. Pressure factor f for Equation (71)

The stability of the liquid drops is a determining factor in designing liquid separators. The We-ber number is the principal criterion for drop

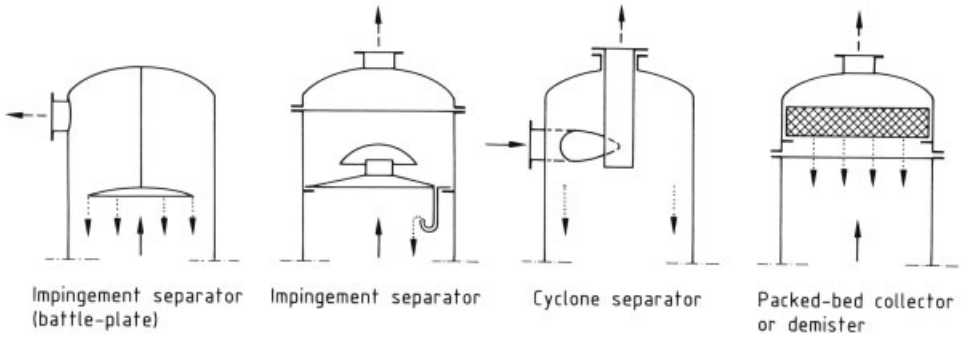

Figure 23. Liquid separators

stability. The Weber number is equal to the quotient of the velocity head arising from the vapor flow external to the drop p_s and the internal excess pressure of the drop p_u caused by the surface tension along the phase boundary between the drop and vapor. In designing liquid separating devices, the critical Weber number of 22 should not be exceeded so that the larger drops do not break up. From this condition, the relationship expressed by Equation (75) results for the highest allowable velocity in the separator. This equation shows the dependence of the highest allowable velocity in the separator on the surface tension of the liquid σ (kg/m), the vapor density ϱ_D (kg/m^3), and the diameter d_T (m), of a liquid droplet.

$$(u_D)_{max} = 4.7\sqrt{\frac{\sigma}{\varrho_D \cdot d_T}} \tag{75}$$

In industrial evaporator systems, the desired degree of purity of the vapors determines the type of separator device. For the most part, separators use reversal of the vapor flow and liquid flow or both to create impingement force or the many varieties of centrifugal force for liquid drop separation. The designs used in practice, then, range from simple baffle and reversing plates to cyclone type internals as well as packed beds and demister pads in the vapor separation chambers. Figure 23 shows a sampling of design concepts. A simple baffle plate suffices as a liquid separator if a vapor flow no more than 1000 – 2000 m^3/h escapes from each square meter of liquid surface in the vapor separation drum.

Consequently, in the free space of a vapor separation drum, predominantly the gravity force determines the effectiveness of separation of the liquid particles entrained by the vapor. With an internally installed separator device, impingement or centrifugal force or both are controlling. Therefore, the degree of separation depends on the extent to which use of these mechanical forces permits smaller drops to coalesce into larger ones. It is self-evident that physical properties like large density difference between the phases and high surface tension of the separated liquid exert a favorable influence. Then for any separator a minimum limiting drop size exists such that drops of smaller size will no longer be separated.

Beds of packing are low-pressure drop devices especially well suited for separation of liquid drops measuring down to about 300 µm. For ring or saddle-shaped packings, the preferred sizes to use are those with principal dimension 15 – 25 mm. Because of their relatively thin wall thickness, use of metal packings results in a lesser pressure drop than that for those manufactured from ceramic materials. Supported on expanded metal mesh, these packings should form a bed of ca. 200 – 300 mm height across the total vapor head cross-sectional area under the vapor head cover.

If very small liquid droplets must be removed from the vapor stream, the recommended device to install is a demister pad, comparatively tightly woven wire mesh packings. At vapor velocities 2 – 5 m/s, they can separate drops down to 3 µm in diameter. Moreover, for these materials, the pressure drop amounts to 15 – 40 mm water column per meter of packed height. Such demisters make possible almost 100 % removal of liquid.

5. Evaporator System Ancillary Components

A number of additional engineered instrument or equipment items are necessary for operating an evaporator system. Thus, in most cases, fresh feed solution requires preheating in a heat exchanger. These are heated either with evaporated vapors, the hot concentrated product, or steam condensate from the evaporator heating element. Not uncommonly, it is also necessary to degas the fresh feed solution before it enters the evaporator to avoid corrosion. Frequently, it may also be necessary to free the concentrated product from the evaporator of residual solvent components with a degassing step. Additionally, pumps are required for feeding the systems, to produce recirculating flow, for product disposition, as well as for producing vacuum. Finally, it is necessary to control the total system, consisting of evaporators and auxiliary equipment, to ensure an automatic and essentially upset-free, safe operation.

5.1. Indirect Heat Exchangers

Preheating, cooling, condensing, and degassing require installing heat-exchange equipment. These may employ direct or indirect heat transfer as their principle of operation. In the indirect method, the streams involved are sepa-rated from one another by a wall through which the heat flows as a result of the difference in temperature. The direct method brings the streams into intimate contact. Normally, most equipment for feed preheating or product cooling employs the method of indirect heat transfer. For preheaters heated by evaporated vapors as well as by product concentrate and condensate cooling, units built as tubular exchanger bundles are usually preferred. Depending on the operating conditions and location factors, one may select either a vertical or horizontal arrangement of the tubes. Inclined tube arrangement applications remain limited to rare cases only.

Surface condensers for evaporated vapors have an important position among heat exchangers using indirect heat transfer. Insofar as vapor generated in single-stage evaporator equipment is not used for preheating or degassing the fresh feed solution, it is condensed in a condenser. Likewise, if the condensation heat of vapors from the last stage of a multistage evaporator is not utilized in some other manner, then they must be condensed in a condenser. The same applies to the so-called residual vapors from vapor recompression evaporators.

The first objective of vapor condensors is to maintain the boiling pressure required in the evaporator stage by continuously condensing the vapors developed and removing the noncondensable gases. Therefore, since often pressures must be held lower than one atmosphere, the methods named are the simplest and most economic to produce the desired lower pressures. To the extent that the condenser operates in the negative pressure domain, the vacuum pump connected downstream serves to remove leakage air or other noncondensable gases. If the main purpose of the evaporation is recovery of the evaporated solvent when it is a valuable product, or to recover the evaporated solvent in the pure form, then condensing the vapors in a surface condenser is the optimum industrial solution. This applies to that vapor not condensed in the heating elements of the stages heated by evaporated vapors in multistage evaporator systems.

5.2. Equipment for Direct Heat Exchange

Heat exchangers in which tube walls separate the streams being cooled from the streams being heated are not always the optimum conceptual solution to a heat-transfer problem. If the substances taking part in the heat transfer are permitted to come in contact with one another, then it may be possible to install equipment for heat transfer directly between them. These units may be less expensive than tubular heat exchangers. The so-called barometric condensers for condensing vapor by direct heat exchange between vapors and cooling medium belong to this field of evaporator design; likewise, the direct degassing equipment for removing gases dissolved in the fresh feed solution to be evaporated. If not degassed, corrosion or impairment in heat transfer could then result in the evaporator. Lastly, the units for degassing the product concentrate from the last evaporator stage by direct contact steam stripping are also in this group.

Comparatively low installed cost distinguishes the barometric condensers using cooling water to condense vapors in industrial systems for evaporating aqueous solutions. The flow direction of the two media differentiates cocurrent (Fig. 25) from countercurrent (Fig. 26) barometric condensers. Based on economic considerations, the latter are most frequently used.

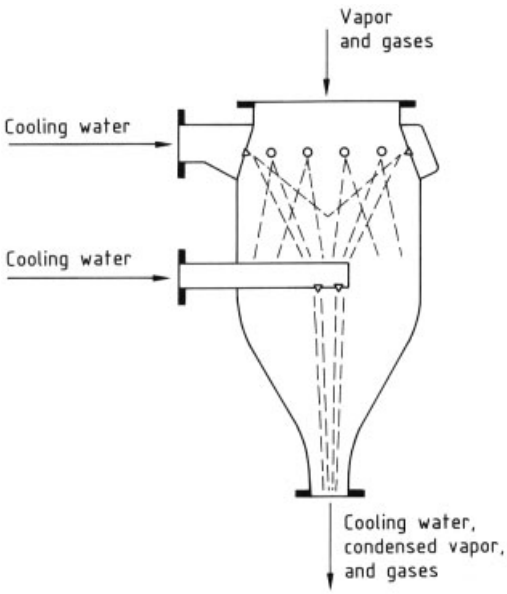

Figure 25. Cocurrent flow barometric condenser

In barometric condensers, a direct heat exchange takes place between vapor and cooling medium through intimate mixing of both materials. In barometric condensers operating countercurrently, the vapors D to be condensed with the enthalpy h enter the condenser at the bottom. Of necessity, they rise countercurrent to the cooling water W flowing from above. The condensed vapor leaves the unit together with the cooling water through the barometric leg. The cooling water temperature rises from the inlet value t_{W} to the common end temperature t_{K}. In general, the vapors entering the condenser contain noncondensable gases G. Especially when operating under vacuum, this gas consists predominantly of air leakage. But, in addition, depending on the nature of the solution to be evaporated, this gas may also contain small quantities of carbon dioxide or ammonia, for example. In the condenser, these gas flows are cooled from their inlet temperature, equal to the vapor temperature t to the final temperature t_{G}. A vacuum pump removes the noncondensable gases after separation of still entrained liquid droplets by a liquid separator located immediately downstream of the condensing chamber.

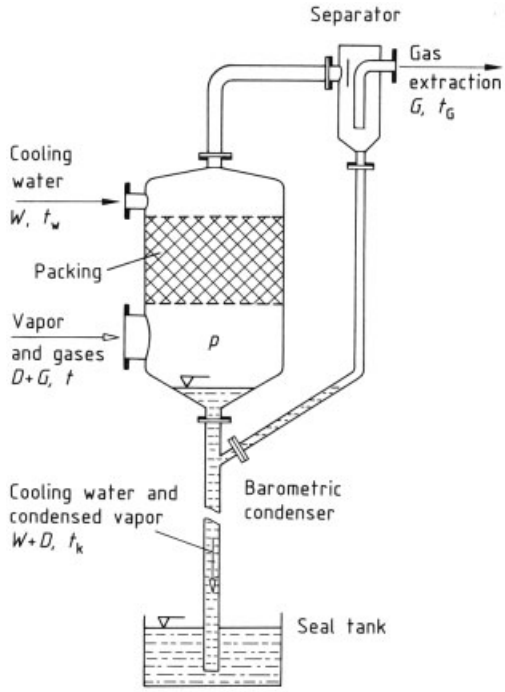

Figure 26. Countercurrent flow barometric condenser

Through direct heat exchange, under otherwise the same conditions, using a barometric condenser permits achieving a usefully lower pressure than with a surface condenser. Equation (76) gives the required cooling water flow for a countercurrent barometric condenser:

$$W = \frac{D\,(h_{\mathrm{D}} - h_{\mathrm{K}}) + G c_{\mathrm{G}}\,(t - t_{\mathrm{G}})}{h_{\mathrm{K}} - h_{\mathrm{W}}} \tag{76}$$

In Equation (76), c_{G} (J kg^{-1} K^{-1}) represents the specific heat of the gas, and h_{W} or h_{K} (J/kg) are the enthalpies of the cooling water and of the condensate. In this balance, heat losses are neglected.

The cooling water consumption necessary for the required condensing capacity depends critically on the temperature level of the exit cooling water – condensate mixture. A high exit water temperature t_{K} results in lower cooling water consumption. On the other hand, the noncondensable gas outlet temperature t_{G} also in-

creases with an increasing value of t_K leading to an increase in the suction volume to the vacuum pump. One may estimate the value of t_G from the empirical equation:

$$t_G = t_W + 4 + 0.1\,(t_K - t_W) \tag{77}$$

In practice, with countercurrent barometric condensers, an exit water temperature t_K that is 6 – 8 °C less than the vapor temperature at the condenser inlet has proven economical.

A barometric condenser must be designed so that it achieves total vapor condensation. This requires a correspondingly large contact surface area between the vapor and cooling water and a sufficiently long residence time in the equipment. Obtaining sufficient contact surface between evaporated vapors and cooling water is a matter of installing internal devices in the usually cylindrically shaped condensation chambers. Suited for this, aside from packings, are also contact trays as well as disc-and-doughnut baffles or shower decks. Moreover, in addition, designs are conceivable which spray the cooling water into the condensing chamber through atomizing nozzles to produce sufficiently high contact surface.

With countercurrent barometric condensers, if available equipment layout space permits sufficient elevation of the condensation system, the mixture of cooling water and vapor condensate leaves the condensing chamber through a simple pipe downcomer or barometric leg. This simple construction concept makes the condensate pump otherwise required, unnecessary or simplifies its design, control, or operation. In sizing the barometric leg, the water column existing in the leg must compensate for the negative pressure in the condensing chamber, in accordance with the principles governing ducts connecting to the surrounding atmospheric pressure. The minimum height of this water column H_{min} (m) is given by the following equation:

$$H_{min} \simeq \frac{p_u - p}{\varrho_W} \cdot 10^4 \tag{78}$$

In Equation (78), p_u (Pa) represents the ambient pressure, p (Pa) the negative pressure prevailing in the condenser, and ϱ_W the density of water, which, as a practical matter, may be set equal to 1000 kg/m^3. At a flow velocity of ca. 1 m/s, an additional liquid head of a maximum of ca. 0.8 m must be added to this minimum water column height. This takes into account the pressure drop arising from the cooling water and condensate flow in the downcomer. The total height of the liquid column thus determined, $H = H_{min} + 0.8$ (m), ensures positive removal of cooling water and condensate. Simultaneously, it maintains the operating pressure in the condenser.

The barometric leg empties into a receiver installed under it. The receiver must be so sized that its capacity for liquid, beyond that for discharge from the leg, is at least 1.5 times the volume of the leg.

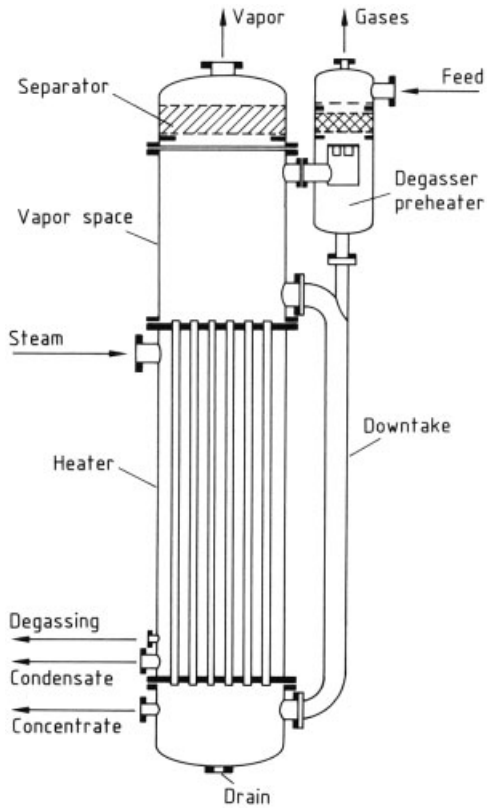

Figure 27. Liquid degasser

Degassing of liquids plays an important part in engineering evaporator systems. Liquids have the tendency to dissolve gases, for example, air, carbon dioxide, ammonia, oxygen and the like, in small quantities. Then, when the liquid warms to the boiling point, the gases are almost completely eliminated. If they are not removed in a timely manner, they may concentrate in the system and cause problems, e.g., deterioration in

heat transfer in steam heated equipment, or perhaps corrosion, or may damage the product by oxidation. Thus, degassing enjoys a special significance, not only for the fresh feed solution, but also, in many cases, for the intermediate or end products. In general, degassing is accomplished through direct heating with evaporated vapor. This warms the liquid practically to the boiling point prior to vaporization. The degassing pressure may be chosen to be the same as that of the first evaporation stage, but lower pressures may also be desirable. Ordinary degasser designs are conceptually similar to countercurrent barometric condensers. The liquid to be degassed flows countercurrent to steam or evaporated vapors entering from below for heating.

Figure 27 shows a compact assembly of liquid degasser integrated with an evaporator. This degasser design uses packing to achieve the required contact surface between liquid and vapor. Other designs employ shower decks or the well-known trays used in distillation or a combination of both.

Degassing a liquid or solution is thus synonymous, in the process engineering sense, with directly preheating it with steam or vapors almost to its boiling point. This direct condensation of steam or vapors results in mixing the condensate formed with the liquid being degassed. Then, in general, these degassing methods may be used only when dealing with the same kind of materials for both vapors and liquids. This is the case when the fresh feed solution is degassed by direct preheating with vapors produced from its own evaporated solvent. Degassing of liquids may be required for reasons of maintaining or improving product quality. Figure 28 is a sketch of an equipment arrangement applied, among others, with success in industrial installations for such cases. This example schematically shows combined degassing, dewatering, and deodorizing fatty acids in a falling film degasser. The key equipment item in the system is an externally heated falling film vaporizer charged at the bottom with stripping steam. Preheated raw acid enters the upper section. It then flows downward as a thin film on the inner walls of the vaporizer tubes countercurrent to the stripping steam, thus not only degassed but also dewatered and deodorized. By previously removing the oxygen in the upper portion of the vaporizer tubes, no damage to the product occurs by oxidation in the lower portion even above 175 °C. The purified fatty acid exits from the bottom of the unit. The condensate dropping out in the condenser flows to a downstream separator. There it separates into a heavy aqueous phase and a small quantity of light fatty acid phase enriched in the color and odor carriers.

5.3. Pumps

In designing evaporator systems, special care must be given to the selection, arrangement, and operation of pumps for transferring liquids and for producing vacuum. Pumps recycle the liquid in forced circulation evaporators. Suitable pumps withdraw the concentrated product from the evaporator system. The vacuum pump unit must operate at the highest level of reliability as pumps withdraw the concentrated product from an indispensable condition for ensuring freedom from upsets in vacuum evaporator system operation.

Pumping capacity, materials of construction, and principle of operation are determining factors in pump selection. Pumps normally available on the market can fulfill the requirements for the preponderant number of applications. In special circumstances, for example, for oxygen-sensitive products in vacuum operations, only pumps without seals and having submerged electric motors, so-called “canned pumps”, or pumps equipped with magnetic couplings, suffice.

Evaporator systems operating under partial or full vacuum, with pressure in the condenser under one bar (0.1 MPa) require a vacuum unit. This removes the noncondensable gases which, in most cases, consist predominantly of air, from the condensing chamber. In general, this unit connects directly to the evaporated vapor condenser. Aside from mechanically operating vacuum pumps, steam jet vacuum pump installations have also proven successful. In their principles of construction and method of operation, these are comparable to steam jet compressors. A single jet is able to produce only a limited compression ratio. To obtain higher vacuums requires connecting several jets in series with one another. A single jet can reach a vacuum of about 10 kPa absolute pressure; two jets can attain about 4 kPa; and three down to 1.3 kPa.

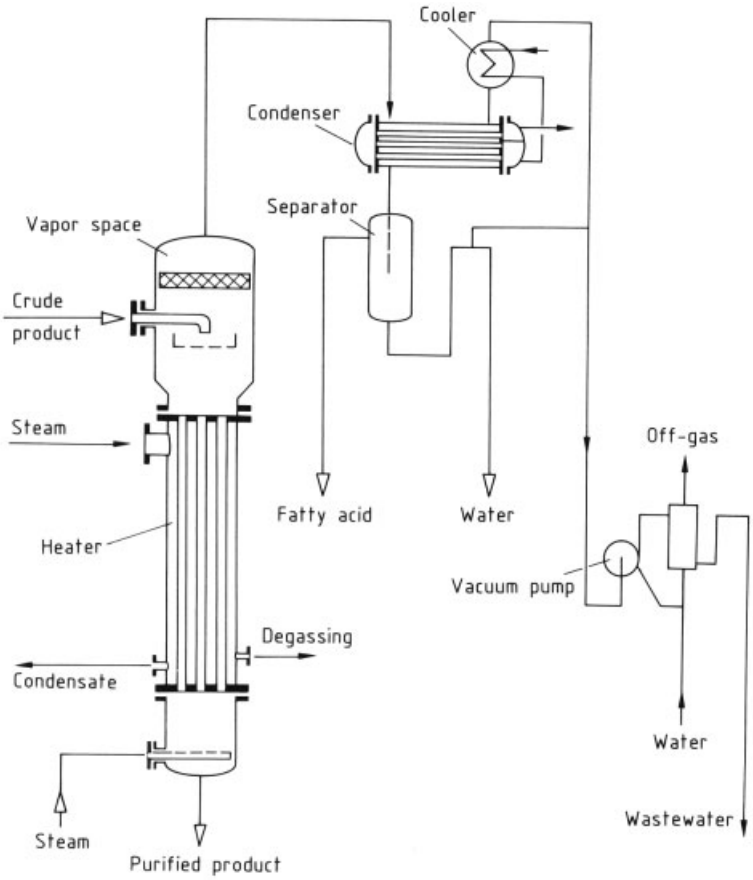

Figure 28. Degassing fatty acids in a falling film evaporator

A high steam consumption results if the vapors compressed in a jet stage of a multistage system pass on to the succeeding jet. To avoid this, intercondensers are installed between the jets to condense the vapors as well as the motive steam. Figure 29 schematically shows this mode of operation.

5.4. Control Instrumentation

Control and surveillance enables determining acceptable values of the characteristic process and operating variables. The multitude of process variants, as well as the variations in operating conditions and desired product purities, do not permit a generally valid standard control concept applicable for all or any evaporation processes. Rather, for best control of a particular evaporator system, a new control concept is developed for that application.

As an example, in multistage recirculating evaporator systems, controllers for the liquid level in the vapor separation chambers, in reality, control the liquid flow through all units by means of throttle valves in the feed liquid piping to each stage. Figure 30 shows this concept. A call for a change in level in any given stage propagates through the upstream stages.

In principle, the evaporating vapor temperature in the last stage of a multistage evaporator system is the control variable regulating the heating steam flow to the first stage. In the concept shown in Figure 30, holding constant pressure in the last stage vapor separation chamber permits using the temperature difference between heating steam and last stage vapors as the control variable.

A density measurement loop in the product drawoff can be used to achieve uniformity of the product concentrate. A controller picking up

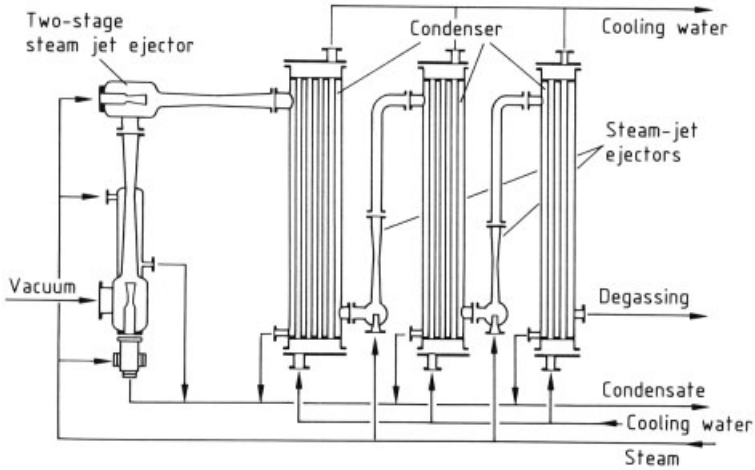

Figure 29. Multistage steam jet vacuum pump

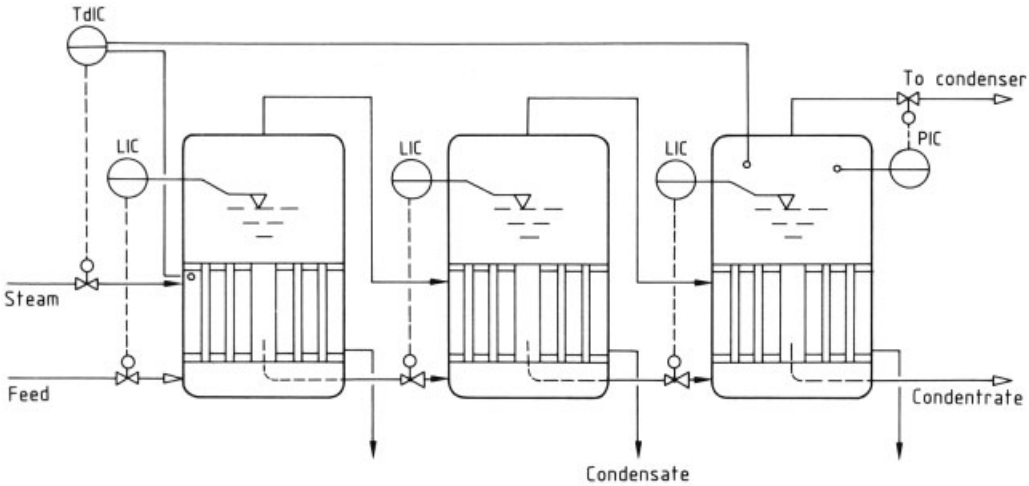

Figure 30. Throughput control in a multistage evaporator system
ISO codes: LIC, level indicating controlling; PIC, pressure indicating controlling; TdIC, temperature difference indicating controlling

this signal can adjust a three-way valve located downstream of the product drawoff pump. Then, deviations from the setpoint density change a partial product recycle flow back to the last evaporator stage. This makes it possible to achieve a correction in the product concentration. In this case, the level controllers serve to readjust the operation of the upstream evaporator stages.

5.5. Insulation

Well-conceived and maintained insulation on vaporizer equipment and piping in an evaporator system serves to reduce local heat losses extensively. This prevents partial refluxing of the evaporated vapors, and rather beyond that, especially avoids overloading the evaporator heating element also.

The insulation normally used depend for their effectiveness on their extremely low thermal conductivity. Low thermal conductivity follows from their low bulk density, characteristic of their porous or wooly nature.

An optimum insulation thickness is usually determined. For this, the annual costs of both the insulation and the expense resulting as a consequence of the heat losses are needed. The annual

cost of insulation is the depreciation. Both costs depend on the insulation thickness. When the sum of the annual costs is plotted versus the insulation thickness, the minimum cost represents the optimum thickness. Determining the optimum insulation thickness is an especially important task in design of evaporation systems with open-air construction, frequently the case in chemical plants. During the winter months, heat loss can be high.

6. Evaporator System Operation

Today, most commercial-scale evaporation processes operate continuously. This is advantageous for system control since steady-stage operation, by definition, requires no changes in process and operating parameters and conditions during the entire duration of the operation. However, if the task involves evaporating small quantities of solution, then, in many cases, batchwise operation may be referred over the continuous. An evaporation system for batchwise operations differs from a continuous evaporator design essentially in that a comparatively large equipment volume is necessary to hold the initial charge of solution to be evaporated. However, due to the often more favorable heat-transfer conditions, the heat-exchange surface in the evaporator heating element proves to be less than that for a comparably operating continuous evaporator. In intermittent operations, the final concentration and, with it, a low heat-transfer coefficient, is the case only at the end of the operating run. For continuous operation, the same conditions for low heat-transfer coefficient apply during the entire operating run. Moreover, in batchwise operations, the time-average increase in boiling point is less than that obtaining for continuous operation at the same end concentration. Thus, if the heating steam temperature is the same for both cases, a greater effective temperature difference is available for batch operation. For these reasons, batch operation is still common even for medium-sized evaporator capacities.

In general, precise process engineering calculations prove to be somewhat more difficult for batch evaporators than those for the steady-state mode. For batch operation, namely, the concentration and with it the boiling temperature and also the film heat-transfer coefficient for the boiling liquid change continuously during the course of a run. This is true also for the effective temperature difference between heating steam and boiling liquid when heating steam temperature is constant. Further, with simple batch evaporation, the inventory of solution in the system decreases with time.

7. Crystallization by Evaporation and by Evaporation with Reduced Pressure Flashing

The enrichment of solids in solutions by evaporating the solvent covers a broad scope among thermal separation processes. In many cases, however, just thickening to higher concentrations alone does not suffice. More often, the objective is to recover the dissolved solid material in virtually pure form, isolating it completely from the solvent. This job cannot readily be done by adhering strictly to the evaporator equipment arrangements so far discussed. These have been conceived for a concentrated product that flows and is pumpable. The design features may need modification, and the method of operation may need alternation especially when processing solutions whose solubility characteristics make crystallization a possible method for separating materials. In principle, all salt solutions fall into this classification in the higher concentration ranges.In the field of evaporator engineering, evaporative crystallization (Fig. 31) and crystallization by evaporation with reduced pressure flashing (Fig. 32) have a special position among separation processes. In practice, frequently the latter is also called vacuum cooling crystallization. With intelligent application, both processes can lead to an economical system design concept. Now, if the solubility of the substance to be crystallized depends only slightly on the temperature, then, in general, the evaporative crystallization process (Fig. 31) would be used. However, crystallization by evaporation with reduced pressure flashing (Fig. 32) is a suitable separation process for preparing crystalline substances whose solubility in the solvent is strongly dependent on temperature.

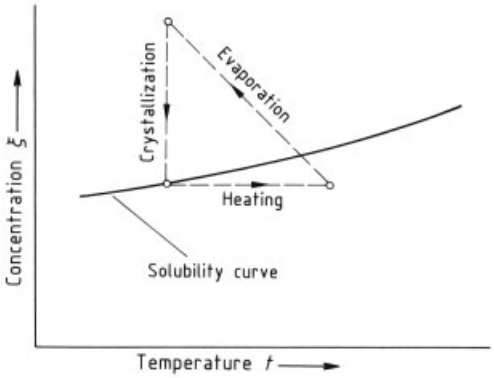

Figure 31. Process steps in evaporative crystallization

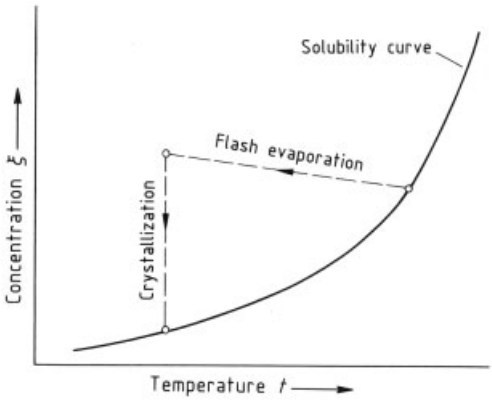

Figure 32. Process steps in crystallization by reduced pressure flash evaporation

Figure 33 shows the scheme for a crystallizer installed mainly in evaporative crystallization systems in combination with an independent external evaporator heating element. Recycling the clear solution and controlling product drawoff hold the degree of supersaturation within the desired limits. Depending on the nature of the feed material, the crystal reaches a size in the range 0.5 – 3 mm.

In designing systems for evaporative crystallization and crystallization by evaporation with reduced pressure flashing, the information in the above chapters, relative to sizing the heat exchangers and vapor separation chambers as well as the ancillary equipment, can be helpful. However, for a good critical examination of the processes occurring in the crystallization characterizing zones and their influence on equipment sizing, a thorough study is absolutely necessary. This is especially the case for troublesome products which show a tendency toward severe salting of the walls of the equipment. For this, the specialized technical literature offers valuable advice.

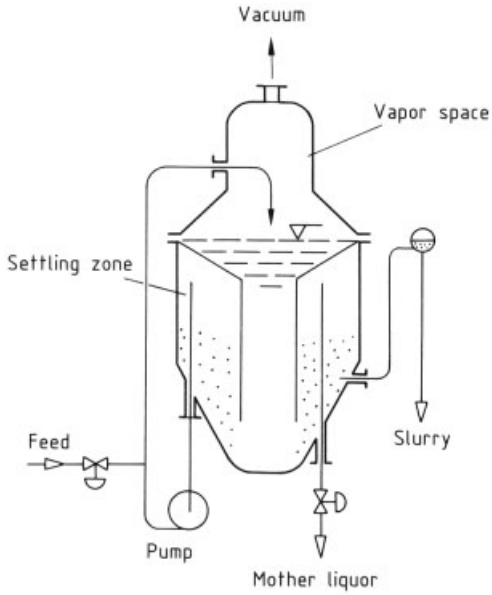

Figure 33. Crystallizer for evaporative crystallization

8. References

1. Reinhard Billet: *Evaporation and Its Technical Applications,* VCH Verlagsgesellschaft, Weinheim-New York 1988.
2. Reinhard Billet: *Verdampfung und ihre technischen Anwendungen,* VCH Verlagsgesellschaft, Weinheim-New York 1981.
3. Paul E. Minton, *Handbook of Evaporation Technology,* Noyes Publications, Park Ridge, N. J., 1986.

Evaporators → Evaporation
Exhaust Control → Automobile Exhaust Control
Expectorants → Cough Remedies
Explosive Cladding → Corrosion
Explosive Cladding → Welding and Cutting

Distillation and Rectification

JOHANN STICHLMAIR, Universität Gesamthochschule Essen, Essen, Federal Republic of Germany

1. Introduction 5
2. Vapor – Liquid Equilibrium 6
2.1. Empirical Findings in Vapor – Liquid Equilibrium 6
2.1.1. Binary Mixtures 6
2.1.2. Ternary Mixtures 8
2.2. Empirical Laws of Vapor – Liquid Equilibrium 13
2.2.1. Equilibrium Constant 14
2.2.2. Equilibrium Curve 14
2.2.3. Distillation Curve 14
2.3. Calculation of Phase Equilibrium from Excess Enthalpy 15
2.4. Calculation of Phase Equilibrium from Equations of State 16
2.5. Practical Determination of Phase Equilibria 18
3. Distillation 18
3.1. Continuous Distillation 19
3.1.1. Continuous Distillation of Binary Mixtures 19
3.1.2. Continuous Distillation of Multicomponent Mixtures 20
3.2. Batch Distillation 21
3.2.1. Batch Distillation of Binary Mixtures 21
3.2.2. Batch Distillation of Multicomponent Mixtures 22
3.3. Semicontinuous Distillation 23
3.3.1. Semicontinuous Distillation of Binary Mixtures 23
3.3.2. Semicontinuous Distillation of Multicomponent Mixtures 23
4. Continuous Rectification (Multiple Distillation) 24
4.1. Principles 24
4.1.1. Equilibrium-Stage Concept 24
4.1.2. Transfer-Unit Concept 26
4.1.3. Comparison of Equilibrium-Stage and Transfer-Unit Concepts 28
4.2. Rectification of Binary Mixtures . 28
4.2.1. Prediction of Rectification Based on Material Balance 29
4.2.2. Prediction of Rectification Based on Material and Enthalpy Balances . . 31
4.2.3. Rectification of Binary Mixtures at Infinite Reflux Ratio 32
4.2.4. Rectification of Binary Mixtures at Minimum Reflux Ratio 33
4.2.5. Heat Requirements for Rectification of Binary Mixtures 34
4.3. Rectification of Ternary Mixtures 36
4.3.1. Rectification of Ternary Mixtures at Infinite Reflux Ratio 36
4.3.2. Rectification of Ternary Mixtures at Minimum Internal Loads 37
4.3.2.1. Minimum Reflux Ratio and Minimum Reboil Ratio for Preferred Separation 37
4.3.2.2. Minimum Reflux Ratio for Separation of the Low Boiler 39
4.3.2.3. Minimum Reboil Ratio for Separation of the High Boiler 40
4.3.3. Rectification of Ternary Mixtures at Finite Reflux Ratio 41
4.3.4. Problems Arising in Column Calculations 42
4.4. Rectification of Multicomponent Mixtures 43
5. Batch Rectification 45
5.1. Batch Rectification of Binary Mixtures 45
5.1.1. Operation at Constant Reflux Ratio 45
5.1.2. Operation at Constant Distillate Composition 46
5.1.3. Operation with Minimum Energy Requirement 47
5.2. Batch Rectification of Multicomponent Mixtures 47
6. Industrial Rectification Processes 49
6.1. Binary Rectification 49
6.2. Multicomponent Separation . . . 49
6.3. Separation of Heterogeneous Azeotropic Mixtures 51
6.4. Separation of Binary Azeotropic Mixtures by Changing System Pressure 52
6.5. Separation of Azeotropic Mixtures with an Entrainer (Azeotropic and Extractive Distillation) 53
6.5.1. Process Description 53
6.5.2. Choice of Entrainer 54

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_04

6.5.3. Entrainment Rectification of Azeotropic Mixtures in Combination with Other Separation Methods 55
7. Energy Economization in Rectification 60
7.1. Single Columns 61
7.1.1. Reduction of Heat Requirement . . 61
7.1.2. Reduction of Exergy Losses 61
7.2. Optimal Separation Sequences for Multicomponent Mixtures 62
7.3. Column Coupling 66
7.3.1. Direct Coupling of Columns 66
7.3.2. Indirect Coupling of Columns . . . 68
7.3.2.1. Binary Mixtures 69
7.3.2.2. Ternary Mixtures 69
7.4. Design of Heat-Exchanger Networks 72
8. Design and Dimensioning of Mass-Transfer Equipment 72
8.1. Types of Construction 72
8.1.1. Plate Columns 73
8.1.2. Packed Columns 74
8.1.3. Criteria for Use of Plate or Packed Columns 76
8.2. Dimensioning of Plate Columns . 77
8.2.1. Operating Region of Plate Columns 77
8.2.2. Two-Phase Flow in Plate Columns . 80
8.2.3. Mass Transfer in the Two-Phase Layer on Column Plates 84
8.3. Dimensioning of Packed Columns 86
8.3.1. Operating Region of Packed Columns 89
8.3.2. Two-Phase Flow in Packed Columns 90
8.3.3. Calculation of Mass Transfer in Packed Columns 95
9. References 97

Symbols

a low boiler
a relative area of liquid – vapor interface, m^{-1}
a ratio of surface area to volume of column packing, m^{-1}
$\bar{a}$ coefficient (equation of state), $N\,m^4\,mol^{-2}$
a_e interface effective in mass transfer, m^{-1}
A area, mass-transfer area, m^2
A Wilson coefficient
A – 1 absorber
b medium boiler/high boiler
$\bar{b}$ coefficient (equation of state), m^3/mol
B amount of bottom product, kmol
B second virial coefficient, kmol/m^3
$\dot{B}$ bottom stream (flow rate), kmol/s
c medium boiler/high boiler
c exponent of Reynold's number in Equation (8.56)
C constant
C third virial coefficient, $(\text{kmol/m}^3)^2$
C_G capacity factor, m/s
C_h constant in Equation (8.62)
$C_{\beta G}$ factor in Equation (8.80)
C – 1 column
c.p. critical point
d high boiler
d diameter, m
d_N nominal diameter, m
D diameter, m
D diffusion coefficient, m^2/s
D amount of overhead product, kmol (distillate)
$\dot{D}$ overhead stream (flow rate), kmol/s
D_E dispersion coefficient, m^2/s
e entrainer
E efficiency
$\hat{E}$ exergy, kJ/kmol
E – 1 heat exchanger
f fugacity, Pa
F factor in Equation (8.3), $kg^{1/2}\,m^{-1/2}\,s^{-1}$
F amount of feed, kmol
$\dot{F}$ feed stream (flow rate), kmol/s
F_a factor in Equation (8.79)
Fr Froude number
g acceleration due to gravity, m/s^2
g specific free enthalpy, kJ/kg
$\dot{G}$ gas or vapor stream (flow rate), kmol/s
$\dot{G}$ gas or vapor stream (flow rate) below the feed point, kmol/s
h height, m
h holdup
h specific enthalpy, kJ/kg
h_p height of pressure drop, m
h_0 holdup below the loading point
H_{ij} Henry coefficient for component i in solvent j, Pa
H plate spacing, height of packing, m
H_{OG} height of a transfer unit with reference to the vapor side, m
k binary interaction coefficient

K equilibrium constant
K mass-transfer coefficient, kmol m^{-2} s^{-1}
l length, path length, m
L amount of liquid, kmol
$\dot{L}$ liquid stream (flow rate), kmol/s
$\dot{L}$ liquid stream (flow rate) below the feed point, kmol/s
m coefficient (equation of state), slope of equilibrium curve
$\dot{M}$ mixing stream (flow rate), kmol/s
$\hat{M}$ molar mass, kg/kmol
n mole quantity, kmol
n number of equilibrium stages; exponent in Equation (8.46)
N number of transfer units, number of nodes
$\dot{N}$ molar flow rate, kmol/s
p pressure, Pa
p_i partial pressure of component i, Pa
$p_i{}^0$ vapor pressure of pure component i, Pa
Pe Péclet number
Q heat, kJ
$\dot{Q}$ heat flow rate, kJ/s
$\hat{Q}$ molar heat, kJ/kmol
r molar latent heat of vaporization, kJ/kmol
R external reflux ratio
R gas constant, kJ kg^{-1} K^{-1}
$\hat{R}$ universal gas constant 8.31451, J mol^{-1} K^{-1}
$\bar{R}$ external reboil ratio
Re Reynold's number
s plate thickness, m
S number of saddle points
$\dot{S}$ molar flow rate after the decanter, kmol/s
S – 1 decanter
t spacing, m
t temperature, °C
t time, s
T absolute temperature, K
v specific volume, m^3/kg
V volume, m^3
$\dot{V}$ volumetric (flow rate), m^3/s
$\hat{V}$ molar volume, m^3/kmol
w superficial velocity, m/s
W molar work, kJ/kmol
We Weber number
x mole fraction (liquid phase)
$\bar{x}$ mole fraction (liquid phase) below the feed point
y mole fraction (vapor phase)
$\bar{y}$ mole fraction (vapor phase) below the feed point
z mole fraction (vapor – liquid mixture), number of spheres
Z compressibility factor

Greek Symbols

α contraction number, relative volatility, coefficient (equation of state)
β mass-transfer coefficient, m/s
ε porosity, holdup
ζ orifice coefficient
γ activity coefficient
Δg^{E} free excess enthalpy, kJ/kmol
Δg^{id} free enthalpy change for mixing of ideal fluids, kJ/kmol
Δp pressure drop, N/m^2
$\Delta \varrho$ density difference, kg/m^3
η dynamic viscosity, $N \cdot s/m^2$
θ contact angle
ν kinematic viscosity, m^2/s
π pole on the enthalpy – concentration diagram, kJ/mol
ϱ density, kg/m^3
σ surface tension, N/m
τ contact time, s
τ_{L} liquid residence time in the in the two-phase layer, s
φ fugacity coefficient, caloric factor, relative free hole area
Φ empirical packing constant
ω acentric factor

Subscripts and Superscripts

a low boiler
ac active
b medium boiler/high boiler
B bottom product
c medium boiler/high boiler, column
C condenser (cooler)
cl clearance under downcomer (skirt clearance)
cr critical
d downcomer, high boiler, dry
D overhead product
E entrainment
f froth layer
F feed
G vapor (gas)
h hole
i component i
id ideal

irr	irrigated
j	component j
k	highest boiling component
l	component l
L	liquid
m	intermediate
max	maximum
min	minimum
n	number of plates or steps
N	nominal
o	orifice, single particle
OG	overall gas-phase state
OL	overall liquid-phase state
R	reboiler
s	solid matter, particle swarm
sp	sphere
t	plate (tray)
V	valve
w	weir
0	initial state
∞	infinity
$*$	equilibrium state
$'$	liquid state, irrigated packing
$''$	vapor state
$\hat{}$	molar
–	average, below the feed point

1. Introduction

Separation of individual substances in a homogeneous liquid mixture or complete fractionation of such mixtures into their components is an important step in many production processes. Different separation procedures can be used for this purpose, but distillation is the most important industrial method.

Distillation utilizes a very simple separation principle based on the development of intimate contact between the homogeneous mixture and a second phase, which thereby allows mass transfer to occur between phases. The thermodynamic conditions are chosen so that only the component to be separated enters the second phase. The phases are subsequently separated; one of them contains the desired substance, and the other consists of a mixture that is largely free of this substance.

Three steps are always involved in industrial implementation of this separation principle:

1) Creation of a two-phase system
2) Mass transfer between phases
3) Separation of the phases

A large number of separation techniques utilize this very effective principle or modifications thereof. Absorption, desorption, evaporation, condensation, and distillation involve a gaseous and a liquid phase; solvent extraction uses two liquid phases. Separating techniques that utilize a fluid phase and a solid phase are adsorption, crystallization, drying, and leaching.

In most of these separations, the necessary two-phase system is created by adding an auxiliary phase to the mixture; the diluted substances to be separated collect with the auxiliary agent. However, in distillation, the second phase is produced by partial vaporization of the mixture. Hence, the use of an auxiliary substance, which usually requires laborious recovery, can be avoided, and the components to be separated can be recovered as pure substances. Indeed, distillation requires energy only in the form of heat, which can subsequently be removed from the system, this is another important advantage.

Historical. Distillation is a very old separation method; chemists pioneered its use in Alexandria in the first century A.D. [7]. Around the eleventh century, distillation was used for the first time in northern Italy to produce alcoholic beverages. The development of distillation equipment has been influenced tremendously by this field of application. Distillation equipment taken from "The Alchemy of Andreas Libavius" dating from 1597 is illustrated in Figure 1 [8]; it was used for the batch distillation of alcohol. Heat was applied to the liquid contents of the boiler (a) built into the oven (b), and the vapor formed was allowed to condense in two coolers (c). Cooling water was changed periodically. The only visible process was the condensate dripping into the receiver (d). Hence, this separation technique was named after the Latin word *destillare*, which means dripping or trickling down.

Even in early times, the fact was known that an even higher alcohol content could be attained by using a second distillation step. In the apparatus shown in Figure 1 two distillations could be carried out simultaneously. Condensate from the first distillation was returned to the headpiece (e), the so-called *Rectificatorium,* which

was heated with vapor rising from the boiler. The vapor thus produced was also condensed in the two coolers (c). A liquid with a higher alcohol content was then collected in a second receiver (f). The term *rectification* is derived from this process which, especially in Europe, is used to describe multistage distillation. The Latin words *recte facere* mean to rectify or improve. Indeed, to this day, rectification refers to the process by which a further concentration change is achieved after the first evaporation step.

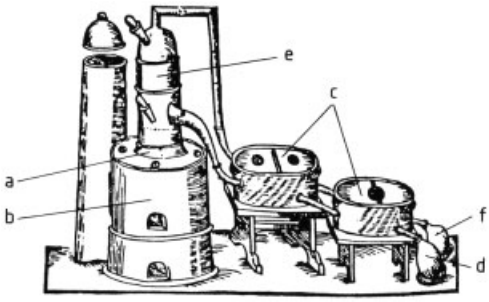

Figure 1. Distillation and rectification equipment taken from "The Alchemy of Andreas Libavius," 1597 [8]
a) Boiler; b) Oven; c) Coolers; d) Receiver; e) Headpiece; f) Receiver

The vapor and liquid are brought into intimate contact by countercurrent flow and mass exchange occurs because the two phases are not in thermodynamic equilibrium. The phases produced during rectification are formed by evaporation and condensation of the initial mixture. The separation of a liquid mixture into its pure components can be controlled solely via the heat supply.

The basis for planning rectification is a knowledge of the *vapor – liquid equilibrium.* The separation achieved depends primarily on the concentration of the individual substances in the vapor and liquid phases. The principles of vapor – liquid equilibrium are discussed in Chapter 2; special attention is paid to the equilibria of ternary and higher mixtures.

Thermodynamic analysis of distillation and rectification is essential to establish the optimal conditions for mass transfer. The decisive factor is the driving force for mass transfer, i.e., the difference between the actual concentrations of the substances and their equilibrium concentrations. Operating conditions should ensure that this difference is sufficiently large. Appropriate relationships and methods are described in Chapters 3, 4, and 5. Examples of industrially important separation processes and energy requirements are discussed in Chapters 6 and 7, respectively.

In distillation and rectification, separation is achieved by bringing two phases into intimate contact. Hence, in practice, the problems caused by multiphase flow and mass transfer between phases must be confronted. Although the mathematical interrelationships of multiphase flow have been studied systematically from the viewpoint of fluid mechanics, often only empirical results are available for practical equipment design, as described in Chapter 8.

2. Vapor – Liquid Equilibrium

If vapor and liquid are in intimate contact for a long period of time, equilibrium is attained between the two phases. This means that no net flow of heat, mass, or momentum occurs across the phase boundary. The complete formulation for a system at equilibrium is as follows:

Thermal equilibrium	$T'' = T'$
Mechanical equilibrium	$p'' = p'$
Composition equilibrium	$y_i* = f(x_i \ldots)$

where T is the absolute temperature, p the pressure, y_i* the mole fraction of component i in the vapor phase at phase equilibrium, and x_i the mole fraction of i in the liquid phase. Single and double primes denote the liquid and vapor phases, respectively. The mass equilibrium between vapor and liquid is of primary importance in separation by distillation and rectification.

2.1. Empirical Findings in Vapor – Liquid Equilibrium

The most important empirical findings pertaining to the vapor – liquid equilibrium of binary and ternary mixtures are discussed below.

2.1.1. Binary Mixtures

Experimental phase equilibria measured at moderate pressure for binary systems composed of substances a and b are shown in Figure 2 [9].

In the upper row, total and partial pressures are plotted vs. the mole fraction of substance a in the liquid (x_a) at constant temperature. Substance a is more volatile than substance b and thus has a higher vapor pressure and a lower boiling point. The middle row shows boiling points (boiling point or bubble point line) and condensation temperatures (dew point line) plotted against the liquid and vapor compositions at constant pressure. The lower row is the concentration of a in the vapor phase (y_a) plotted vs. its concentration in the liquid phase (x_a) at constant pressure.

Significant differences in the behavior of these binary systems at equilibrium are due to the varying forces of interaction between the two molecular species a and b. Mixtures in which molecules of a and b repel each other are shown on the left side of Figure 2. Both substances undergo an increase in partial pressure p' of the liquid phase. According to the Clausius – Clapeyron relationship, this increase in vapor pressure leads to a decrease in boiling point. In the 1-butanol – water system, repulsive forces between the molecules are so strong that separation into two liquid phases occurs at intermediate concentration ranges.

Binary mixtures shown on the right side of Figure 2 exhibit exactly the opposite behavior. Molecules of substances a and b attract each other and tend to form complexes. This leads to a decrease in partial pressure p' and thus to an increase in boiling point.

In the binary mixture benzene – toluene shown in the center of Figure 2, forces of interaction between the molecular species a and b are negligible. A molecule a does not "know" whether it is colliding with an identical molecule or with a molecule b of the other substance. Such mixtures are known as *ideal mixtures*; their phase equilibria can be represented by very simple equations (see Section 2.2).

The conditions at which the vapor and liquid compositions are equal (i.e., $y = x$) are very important. Such mixtures, known as *azeotropic mixtures,* have a minimum or a maximum boiling point. In a minimum-boiling azeotrope the molecular species repel each other, whereas in a maximum-boiling azeotrope they attract each other. If the azeotrope occurs in the composition range in which the two liquids are immiscible, phase splitting occurs and a heterogeneous azeotrope is formed. A heterogeneous azeotrope is always minimum boiling.

A mixture of close-boiling components tends to form an azeotrope when the chemical structures of the components are dissimilar. Binary mixtures with two azeotropes (one minimum and one maximum) have been observed occasionally, e.g., perfluorobenzene – benzene (cf. Fig. 6 H).

The influence of pressure on the phase equilibrium of a mixture of nitrogen and methane is shown in Figure 3 A. At higher pressures, differences between the concentrations of the vapor y and the liquid x decrease markedly. When the critical pressure of the pure substances is exceeded, they can no longer be liquefied in the supercritical region. However, mixtures of supercritical vapors can split into two phases, as shown in Figure 3 B. The region in which this can occur is restricted by the convergence pressure curve, which represents the critical points of mixtures of the two substances and may pass through a maximum.

2.1.2. Ternary Mixtures

The phase equilibrium of ternary mixtures is depicted conveniently on a triangular diagram, in which the corners represent the pure substances a, b, and c.

In Figure 4, the mole fraction of component i in the liquid, x_i, is represented by the basic grid and the vapor mole fraction y_i* in equilibrium with the liquid is shown as parameter curves. For ideal mixtures with constant relative volatilities α_{ik} (see Section 2.2.2), these parameter curves are straight lines. However, for nonideal systems plots of y_i* do not yield straight lines, and the entire representation becomes confusing. In addition, the azeotropic points, which are very important in rectification, are difficult to recognize.

The use of *distillation curves* is more effective. If an initial liquid concentration of component i, x_{i0}, is assumed, the equilibrium vapor concentration $y_{i0}*$ is determined. The vapor is then assumed to be totally condensed. Thus, a new liquid of composition $x_{i1} = y_{i0}*$ is obtained and marked on the triangular diagram. The vapor state $y_{i1}*$ in equilibrium with this liquid state x_{i1} is, in turn, determined and so on. A series of liquid states $x_{i1}, x_{i2}, x_{i3}, \ldots$, is finally obtained; the

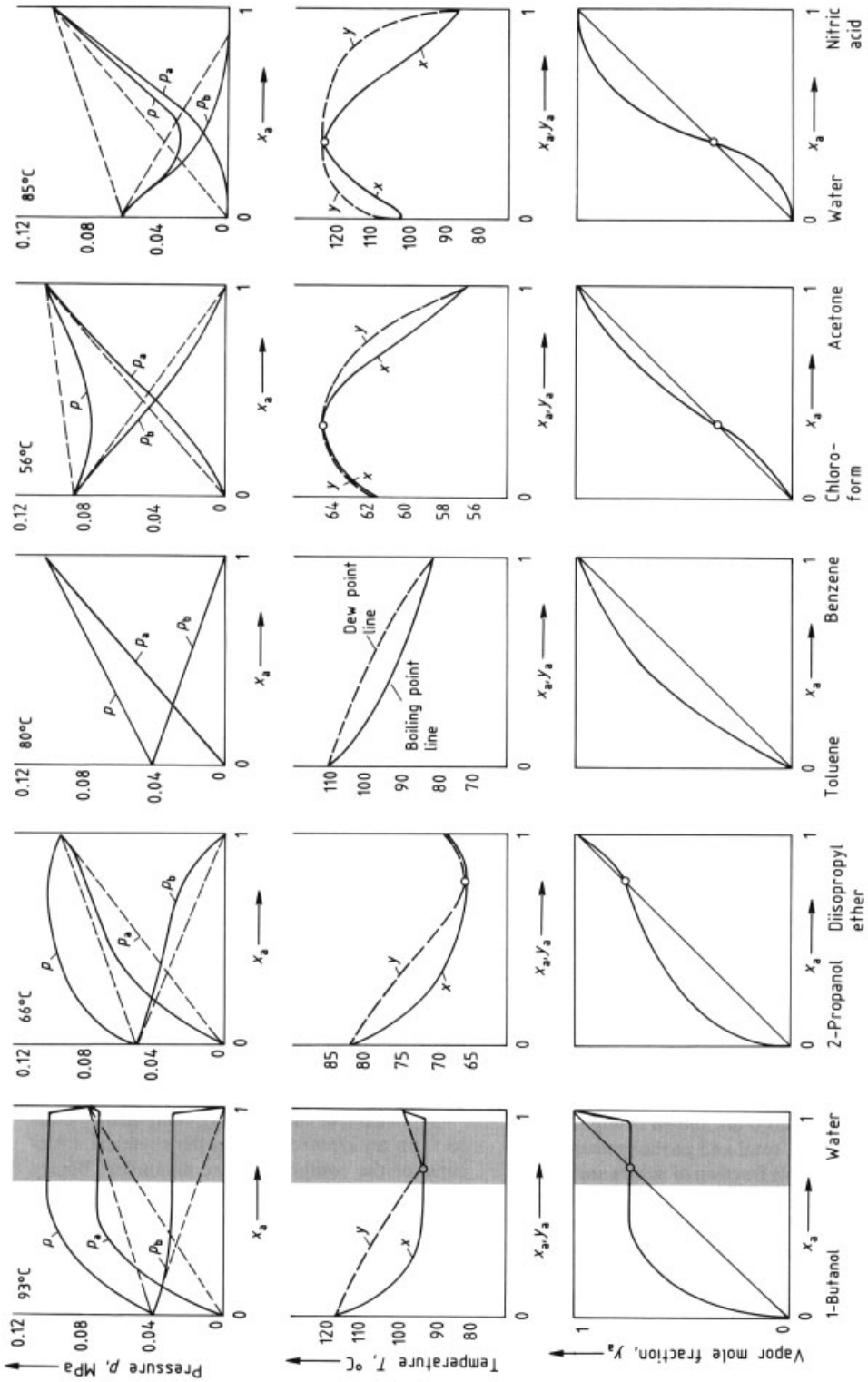

Figure 2. Phase equilibria of five binary mixtures composed of substances a and b [9]
Substance a is more volatile than substance b. The dashed lines in the upper row represent ideal behavior. Temperature and vapor concentration equilibria (middle and bottom rows) were measured at constant pressure (0.1 MPa). Shaded area represents the liquid immiscibility composition range. Azeotropic points are indicated by open circles.
p_a = partial pressure of a; p_b = partial pressure of b; x_a = mole fraction of a in liquid phase; y_a = mole fraction of a in vapor phase

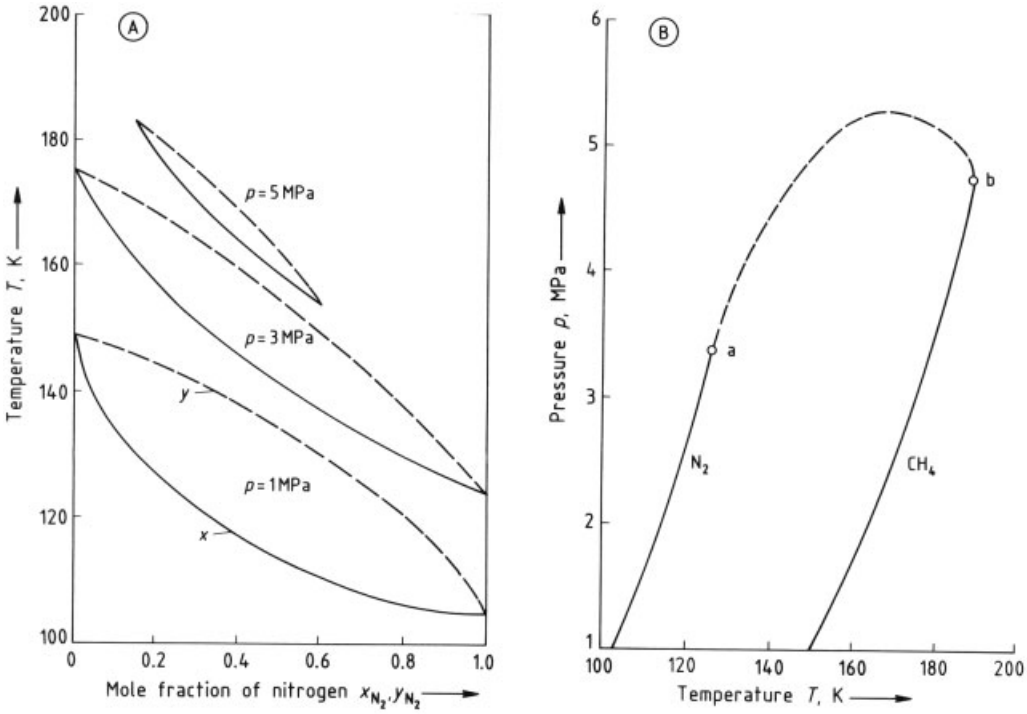

Figure 3. Influence of pressure p on the phase equilibrium of nitrogen and methane
A) Boiling point lines (——) and dew point lines (– – –) of nitrogen – methane mixtures; x_{N_2} = mole fraction of nitrogen in the liquid phase; y_{N_2} = mole fraction of nitrogen in the vapor phase; B) Critical points of nitrogen (a), methane (b), and their mixtures (convergence pressure curve – – –)

composition difference in each case represents an equilibrium step. These liquid states can be joined together to give a distillation curve. This process leads to step-by-step enrichment of the more volatile component, which is accompanied by a decrease in boiling point.

Several distillation curves are shown in Figure 5. The direction of boiling point decrease is indicated by an arrow. The sides of the triangle, which represent binary mixtures, can also be interpreted as distillation curves. The dots on one of the curves represent the sequence of distinct liquid states from one equilibrium stage to the next. Large spaces between points depict regions with large differences in concentration between the liquid and the corresponding vapor.

Distillation curves for nine ternary systems are shown in Figure 6. In each case, the lowest boiling substance (a) is at the top of the triangle, the medium boiler (b) is at the right side, and the high boiler (c) is at the left.

Mixture A exhibits almost ideal behavior. The distillation curves proceed from the high-boiling substance c to the low boiler a. At the same time, the medium-boiling component b exhibits a more or less pronounced concentration maximum.

A binary minimum azeotrope between the low- and medium-boiling substances is formed in mixture B. Indeed, the lowest boiling temperature of the system is at this azeotrope. As a result, all distillation curves starting from the high boiler c tend toward this binary azeotrope. Analogous behavior is observed with mixture C in which a binary maximum azeotrope occurs between the medium-boiling substance b and the high boiler c. The highest boiling temperature of the system is at this azeotrope. All distillation curves start with this temperature maximum and end at the temperature minimum, i.e., at the low boiler a.

Mixture D shows a binary a – b maximum azeotrope. The distillation curves, starting from

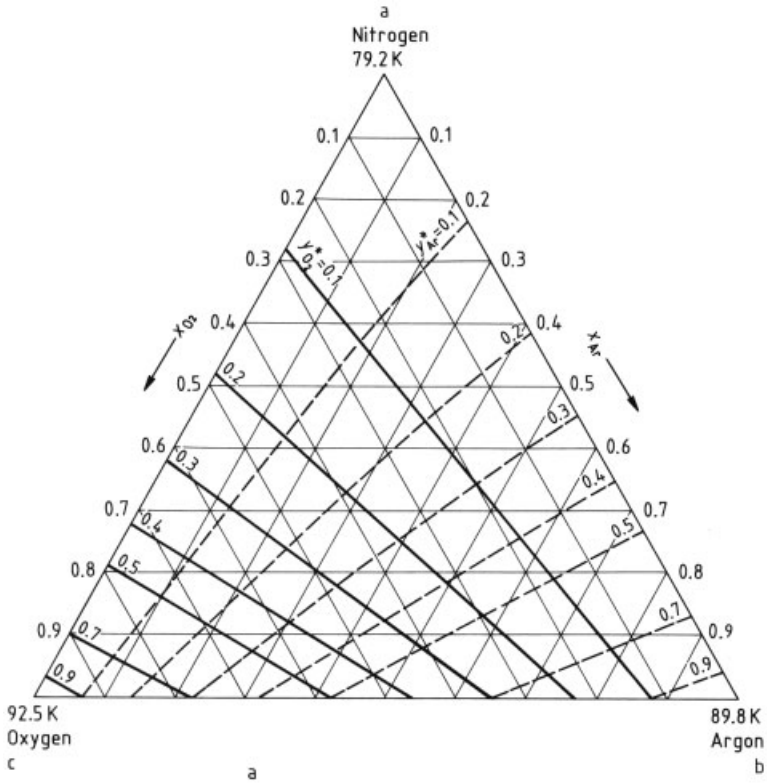

Figure 4. Representation of the phase equilibrium of the ternary mixture nitrogen – argon – oxygen at 131 kPa
The uniformly divided grid of the triangular diagram represents the liquid composition x_{Ar}, x_{O_2}, and x_{N_2}. The bold and dashed lines indicate equilibrium vapor mole fractions of oxygen ($y*_{O_2}$) and argon ($y*_{Ar}$), respectively. Temperatures indicate boiling points of the pure components.

the high boiler c, first proceed toward this azeotrope and then turn toward the low boiler a or the medium boiler b. Thus, this ternary system is divided into two areas whose distillation curves have different end points. The distillation curve that forms the boundary between these two areas is called the *distillation boundary.* In this case, the boundary runs from the pure high-boiling substance c to the binary maximum azeotrope.

A distillation boundary also arises when a binary mixture of a medium-boiling substance b and a high boiler c produces a minimum azeotrope (mixture E). In this case, a heterogeneous azeotrope is formed because it lies within the liquid immiscibility range. The distillation curves for mixture E have two starting points, but they all end at the low-boiling substance a. The distillation boundary runs from the heterogeneous azeotrope to the low boiler a.

Two binary minimum azeotropes occur in mixture F where the distillation boundary runs between the two azeotropes. This ternary system is thus divided into two areas, the distillation curves in each area having different starting points.

In mixtures G and H in Figure 6, a saddle-point ternary azeotrope is formed in addition to the two azeotropes. Distillation curves do not start or end at the saddle point. Instead, they proceed from different directions toward this point and then turn away. Distillation boundaries divide both ternary systems into four distillation fields. The binary mixture of the low boiler a and the medium-boiling substance b in mixture H forms a minimum- and a maximum-boiling binary azeotrope.

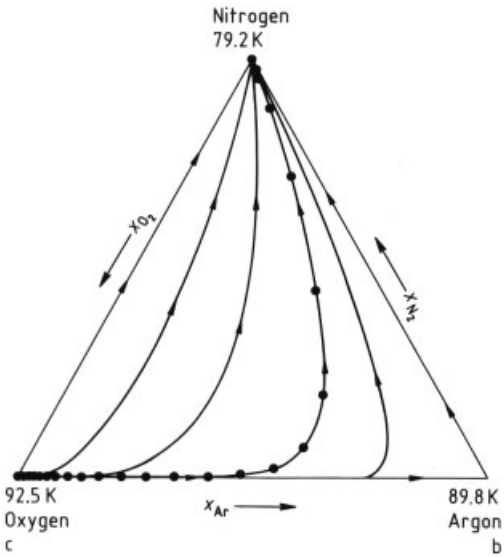

Figure 5. Representation of the phase equilibrium of the ternary mixture nitrogen – argon – oxygen at 131 kPa by distillation curves
The distillation curves start at the component with the highest boiling point and end at the component with the lowest boiling point. The direction of decreasing temperature is indicated by arrows. Distinct equilibrium states are shown on one curve as dots.

An immiscibility range with a heterogeneous azeotrope and two binary minimum azeotropes occur in system I. In addition, a ternary minimum azeotrope with the lowest boiling temperature of the entire system is formed. Thus, all distillation curves originate at the pure substances and converge at the ternary azeotrope. The azeotropes are connected by distillation boundaries giving rise to three distillation fields.

Distillation curves facilitate clear representation of the phase equilibria of ternary systems. Reliable vapor – liquid equilibrium data are essential for planning distillation and rectification processes. As shown in Figure 5, determination of the distance from one equilibrium stage to the next is also very useful. However, knowledge of the boiling temperatures of the points indicated (pure substances and azeotropes) and of the directional arrows on the distillation curves is usually sufficient for process design. The position of the distillation boundaries, which divide the ternary system into different fields, is of considerable importance. These boundaries always connect points of minimum or maximum temperature, which are usually the high boiler, the low boiler, and binary or ternary azeotropes. Knowledge of these points alone may be sufficient to estimate the approximate positions of distillation boundaries; phase equilibrium data are not always required.

The boiling points of a ternary mixture can be represented three-dimensionally where the distillation boundaries form either a ridge or a valley. Graphical representation of the contours of the temperature surface as a function of liquid composition in the system acetone – chloroform – methanol (G in Fig. 6) is shown in Figure 7. The distillation boundaries intersect at a saddle-point indicating a ternary azeotrope.

The existence of ternary azeotropes is of considerable importance and can be checked by using the following equation [10]:

$$2\,(N_3 - S_3) + N_2 - S_2 + N_1 = 2 \tag{2.1}$$

where N is the number of nodes at which distillation curves either start or end, and S is the number of saddle points at which distillation curves either meet or split, usually combined with a sudden change of direction. The subscripts 1, 2, and 3 indicate a pure substance, a binary mixture, and a ternary mixture, respectively. For example, mixture A in Figure 6 exhibits only two nodes for pure substances. Thus $N_1 = 2$ and Equation (2.1) balances. The distillation curves in mixture I start at all three pure substances; thus $N_1 = 3$. All three binary azeotropes are saddle points at which two distillation curves meet and then break away by changing direction. Therefore, $S_2 = 3$ and Equation (2.1) becomes

$$2\,(N_3 - S_3) + 0 - 3 + 3 = 2 \tag{2.2}$$

In other words, ternary azeotropes must exist. The simplest solution to this equation is $N_3 = 1$ and $S_3 = 0$; i.e., one ternary azeotrope exists. Several ternary azeotropes can occur in this system only if ternary saddle points exist at the same time, which however is not the case.

Analogously, for mixture G in Figure 6, $N_1 = 1$ and $N_2 = 3$; the following equation is obtained:

$$2\,(N_3 - S_3) + 3 - 0 + 1 = 2 \tag{2.3}$$

The simplest solution to this equation is $S_3 = 1$. In other words, one ternary saddle point

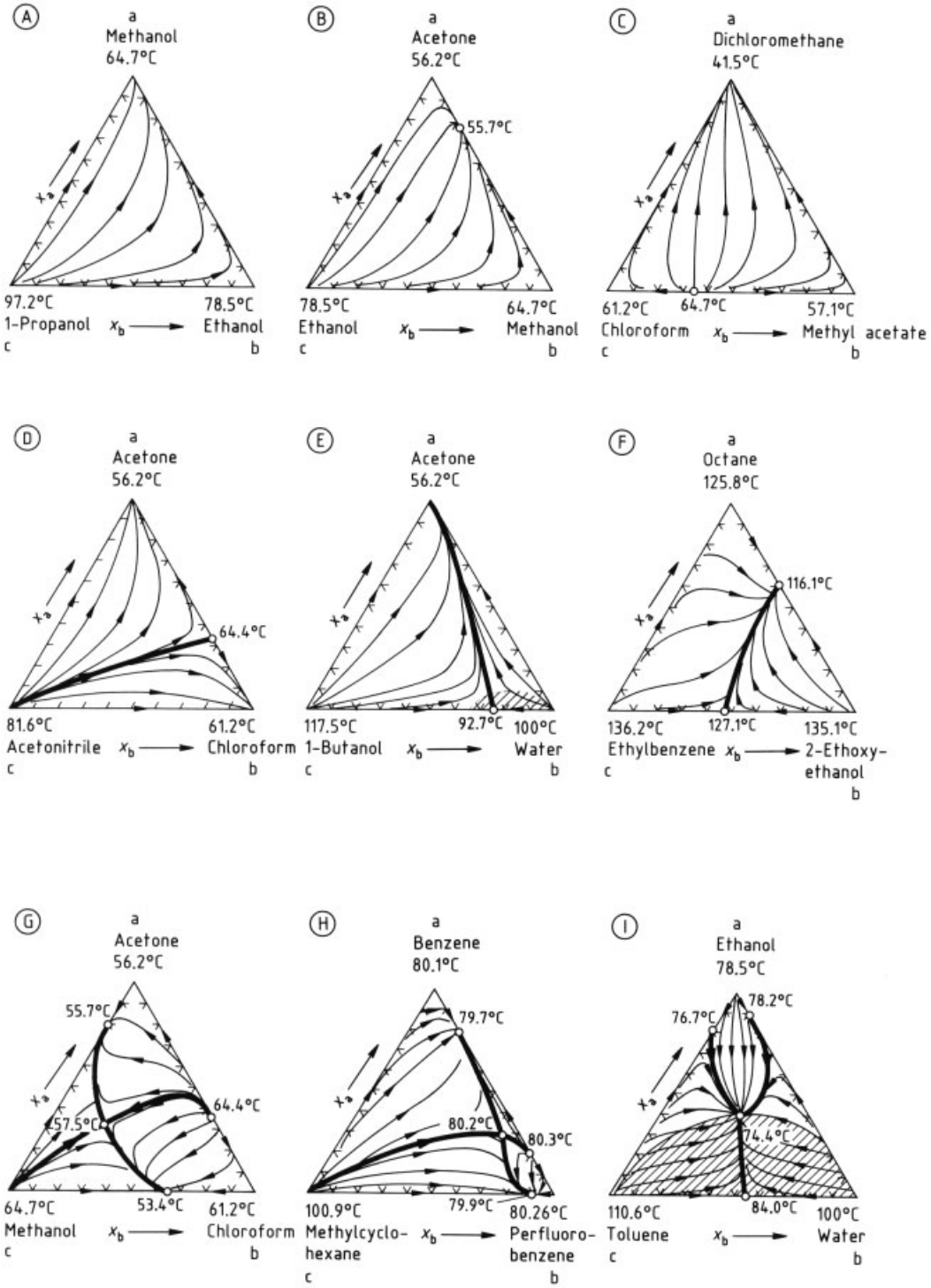

Figure 6. Phase equilibria of nine ternary mixtures at 101.3 kPa
The distillation curves proceed from a temperature maximum to a temperature minimum. Distillation boundaries are represented by bold lines. Shaded areas indicate the presence of two liquid phases. Open circles denote azeotropic points.

exists, which corresponds to the experimental findings.

The preceding examples show that Equation (2.1) can be employed to predict the existence of ternary azeotropes by using a knowledge of binary marginal mixtures. Although several solutions may exist, the simplest is probably correct.

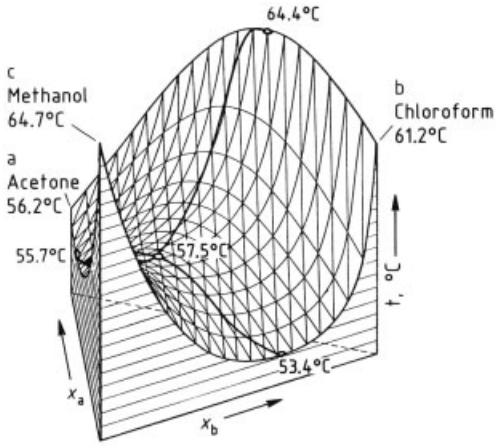

Figure 7. Boiling point surface as a function of liquid composition for the ternary acetone – chloroform – methanol system
Temperatures indicate boiling points of pure substances and azeotropic points.

2.2. Empirical Laws of Vapor – Liquid Equilibrium

Vapor – liquid equilibria are best understood with the help of simple, empirical laws which may, however, only apply within certain limits.

Dalton's law describes vapor systems:

$$p_i''/p''=n_i''/n''=V_i''/V'' \tag{2.4}$$

where p'' is the pressure, n'' the number of moles of substance, and V'' the volume of the vapor. The subscript i denotes component i. Because $n''_i/n''=y_i$ (where y_i is the mole fraction of component i in the vapor), this law can be expressed as:

$$p_i''=y_i\cdot p'' \tag{2.5}$$

Dalton's law generally holds at low pressure ($p''\ll p_{\mathrm{cr}}$, where p_{cr} is the critical pressure). At higher pressure, however, it may require modification. Instead of using pressure p, the equation is written in terms of fugacity f:

$$f_i''=y_i\cdot f'' \tag{2.6}$$

The fugacity f can be determined from the fugacity coefficient $\varphi=f/p$ (Section 2.4).

According to *Raoult's law,* the partial pressure p_i' of component i from a liquid mixture is

$$p_i'=x_i\cdot p_i^0 \tag{2.7}$$

Hence, the partial pressure p_i' depends only on the vapor pressure $p_i{}^0$ of pure component i and its liquid mole fraction x_i; it is not affected by the nature of other substances in the liquid mixture. Raoult's law holds only as long as the forces of interaction between different types of molecules are equal (ideal mixture). Nonideal liquid-phase behavior (i.e., unequal interaction forces) is described by using the activity coefficient γ_i, which is defined as follows:

$$p_i'=\gamma_i\cdot x_i\cdot p_i^0 \tag{2.8}$$

The activity coefficient γ_i is highly concentration-dependent. Methods used to determine activity coefficients are described in Section 2.3.

Equations (2.7) and (2.8) apply only at temperatures below the critical temperature ($T<T_{\mathrm{cr}}$) because the substance does not exert a vapor pressure at higher temperatures. When $T>T_{\mathrm{cr}}$, *Henry's law* is applied instead of Raoult's law:

$$p_{\mathrm{i}}'=H_{ij}\cdot x_i \tag{2.9}$$

where H_{ij} is the Henry coefficient for substance i in solvent j. Again, a correction must be made for nonideal mixtures by using the activity coefficient γ_i:

$$p_i'=H_{ij}\cdot\gamma_i\cdot x_i \tag{2.10}$$

This relationship is analogous to Raoult's law, except that the proportionality constant is not the vapor pressure $p_i{}^0$, but Henry's coefficient H_{ij}.

2.2.1. Equilibrium Constant

The equilibrium constant for component i, K_i, is commonly used to represent phase equilibrium data, especially in multicomponent mixtures. It is defined as

$$K_{\mathrm{i}} = y_i^*/x_i \tag{2.11}$$

Because at phase equilibrium

$$p_i'' = p_i' \tag{2.12}$$

then from Equations (2.5) and (2.7),

$$K_i = p_i^0/p \text{ or } K = p_i^0 \cdot \gamma_i/p \tag{2.13}$$

Similarly, at temperatures above the critical point,

$$K_i = H_{ij}/p \text{ or } K = H_{ij} \cdot \gamma_i/p \tag{2.14}$$

Both the vapor pressure ${p_i}^0$ and Henry's coefficient H_{ij} are strongly dependent on temperature. Hence, the equilibrium constant K_i can be employed only at known temperatures.

2.2.2. Equilibrium Curve

For binary systems consisting of components a and b,

$$y_{\mathrm{a}}^* = \frac{p_{\mathrm{a}}'}{p_{\mathrm{a}}' + p_{\mathrm{b}}'} = \frac{x_{\mathrm{a}} \cdot p_{\mathrm{a}}^0}{x_{\mathrm{a}} \cdot p_{\mathrm{a}}^0 + x_b \cdot p_{\mathrm{b}}^0} \tag{2.15}$$

Since the vapor pressures of the pure components p_{a}^0 and p_{b}^0 are highly temperature-dependent, this equation applies only at constant temperature. However, the ratio $p_{\mathrm{a}}^0/p_{\mathrm{b}}^0$ is often independent of temperature; therefore, vapor – liquid equilibria are frequently expressed in this form as the relative volatility α:

$$\alpha_{\mathrm{ab}} = p_{\mathrm{a}}^0/p_{\mathrm{b}}^0 \tag{2.16}$$

From the equation $x_{\mathrm{b}} = 1 - x_{\mathrm{a}}$,

$$y_{\mathrm{a}}^* = \frac{\alpha_{\mathrm{ab}} \cdot x_{\mathrm{a}}}{1 + (\alpha_{\mathrm{ab}} - 1) \cdot x_{\mathrm{a}}} \tag{2.17}$$

This relationship is called the *equilibrium curve* and is valid when the temperature is constant. If α_{ab} does not vary with temperature, Equation (2.17) can also be employed at other temperatures at constant pressure; this is particularly important in distillation and rectification.

Similarly, for ternary systems consisting of components a, b, and c:

$$y_{\mathrm{a}}^* = \frac{p_{\mathrm{a}}'}{p_{\mathrm{a}}' + p_{\mathrm{b}}' + p_{\mathrm{c}}'} = \frac{x_{\mathrm{a}} \cdot p_{\mathrm{a}}^0}{x_{\mathrm{a}} \cdot p_{\mathrm{a}}^0 + x_{\mathrm{b}} \cdot p_{\mathrm{b}}^0 + x_{\mathrm{c}} \cdot p_{\mathrm{c}}^0} \tag{2.18}$$

Substituting $\alpha_{\mathrm{ac}} = p_{\mathrm{a}}^0/p_{\mathrm{c}}^0$, $\alpha_{\mathrm{bc}} = p_{\mathrm{b}}^0/p_{\mathrm{c}}^0$, and $x_{\mathrm{c}} = 1 - x_{\mathrm{a}} - x_{\mathrm{b}}$ gives

$$y_{\mathrm{a}}^* = \frac{x_{\mathrm{a}} \cdot \alpha_{\mathrm{ac}}}{1 + (\alpha_{\mathrm{ac}} - 1) \cdot x_{\mathrm{a}} + (\alpha_{\mathrm{bc}} - 1) \cdot x_{\mathrm{b}}} \tag{2.19}$$

$$y_{\mathrm{b}}^* = \frac{x_{\mathrm{b}} \cdot \alpha_{\mathrm{bc}}}{1 + (\alpha_{\mathrm{ac}} - 1) \cdot x_{\mathrm{a}} + (\alpha_{\mathrm{bc}} - 1) \cdot x_{\mathrm{b}}}$$

The general equation for multicomponent mixtures is

$$y_i^* = \frac{\alpha_{ik} \cdot x_i}{1 + \sum_{j=1}^{k-1} (\alpha_{jk} - 1) \cdot x_j} \tag{2.20}$$

for component 1 to $k - 1$ where k refers to the highest boiling substance. In nonideal systems, the activity coefficient γ_i must be incorporated into the equations.

2.2.3. Distillation Curve

The distillation curves of ideal mixtures are very important in representing the phase equilibria of ternary mixtures and can be expressed by a simple equation. Rearrangement of Equation (2.17) gives the equilibrium curve of ideal mixtures in the following form:

$$\frac{y_0^*}{1 - y_0^*} = \alpha \cdot \frac{x_0}{1 - x_0} \tag{2.21}$$

where the subscript 0 denotes the initial state.

Total condensation of the vapor arising from the liquid is assumed in determining the paths of distillation curves (see Section 2.1.2). A new liquid with the concentration $x_1 = y*_0$ is obtained. Thus:

$$\frac{x_1}{1 - x_1} = \alpha \cdot \frac{x_0}{1 - x_0} \tag{2.22}$$

Similarly, $x_2 = y*_1$ is valid for the second step, giving:

$$\frac{x_2}{1 - x_2} = \alpha \cdot \frac{x_1}{1 - x_1} = \alpha^2 \cdot \frac{x_0}{1 - x_0} \tag{2.23}$$

For the nth step,

$$\frac{x_n}{1-x_n}=\alpha^n\frac{x_0}{1-x_0} \tag{2.24}$$

Rearrangement gives

$$x_n=\frac{\alpha^n\cdot x_0}{1+(\alpha^n-1)\cdot x_0} \tag{2.25}$$

This equation is similar to the equilibrium relationship (Eq. 2.17), except that here the term α^{n} replaces α. Thus, for ternary mixtures,

$$x_{\mathrm{a}n}=\frac{\alpha_{\mathrm{ac}}^n\cdot x_{\mathrm{a0}}}{1+(\alpha_{\mathrm{ac}}^n-1)\cdot x_{\mathrm{a0}}+(\alpha_{\mathrm{bc}}^n-1)\cdot x_{\mathrm{b0}}} \tag{2.26}$$

$$x_{\mathrm{b}n}=\frac{\alpha_{\mathrm{bc}}^n\cdot x_{\mathrm{b}}}{1+(\alpha_{\mathrm{ac}}^n-1)\cdot x_{\mathrm{a0}}+(\alpha_{\mathrm{bc}}^n-1)\cdot x_{\mathrm{b0}}}$$

Similarly, for multicomponent mixtures, the distillation curve is represented by

$$x_{in}=\frac{\alpha_{ik}^n\cdot x_{i0}}{1+\sum_{j=1}^{k-1}\left(\alpha_{jk}^n-1\right)\cdot x_{j0}} \tag{2.27}$$

The distillation curves of ideal mixtures with a constant relative volatility α_{ik} can easily be calculated by using these equations.

2.3. Calculation of Phase Equilibrium from Excess Enthalpy

Deviations from ideal behavior occur in the liquid phase rather than in the gas phase. As a result of smaller intermolecular distances, the forces of interaction between molecules in the liquid are considerably stronger. In contrast, the vapor phase can be assumed to be ideal at moderate pressure.

Real liquid-phase behavior is usually described by means of the activity coefficient γ (Eq. 2.8). However, the activity coefficients of individual components *i* of a liquid mixture are not independent of each other, as formulated by the *Gibbs – Duhem equation* [11]:

$$\sum x_i\cdot \mathrm{d}\ln\gamma_i=0 \tag{2.28}$$

For binary mixtures, the Gibbs – Duhem equation requires that both activity coefficients are either <1 or >1 and that they exhibit oppositely directed gradients when plotted as a function of the concentration.

Equation (2.28) is a fundamental thermodynamic equation which must always be satisfied and has long been used to check the thermodynamic consistency of vapor – liquid equilibrium data [11].

To obtain a relationship for the concentration dependency of the activity coefficient γ_i, starting with one value of state of the entire mixture is useful. In this way, a thermodynamically consistent formulation of phase equilibrium is attained automatically. Mixing processes can be described most simply by using the free enthalpy g [11]. The free enthalpy change for the mixing of ideal liquids, Δg^{id}, is

$$\Delta g=\hat{R}\cdot T\cdot\sum x_{\mathrm{i}}\cdot\ln x_{\mathrm{i}}id \tag{2.29}$$

where $\hat{R}$ is the universal gas constant. Analogously, for nonideal mixtures,

$$\Delta g=\hat{R}\cdot T\cdot\sum x_i\cdot\ln x_i+\hat{R}\cdot T\cdot\sum x_i\cdot\ln\gamma_i \tag{2.30}$$

The second term accounts for the nonideal behavior of the mixture and is called the free excess enthalpy Δg^{E}:

$$\Delta g=\hat{R}\cdot T\cdot\sum x_i\cdot\ln\gamma_i E \tag{2.31}$$

Hence,

$$\Delta g=\Delta g^{\mathrm{id}}+\Delta g^{\mathrm{E}} \tag{2.32}$$

Differentiation of the equation for the free excess enthalpy Δg^{E} gives relationships for the activity coefficients γ_i of the individual components of a mixture. In general,

$$\ln\gamma_i=\frac{1}{\hat{R}\cdot T}\left(\Delta \mathrm{g}^{\mathrm{E}}+\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_i}-\sum_{k=1}^{n}x_k\cdot\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_k}\right) \tag{2.33}$$

For a binary mixture of substances a and b,

$$\ln\gamma_{\mathrm{a}}=\frac{1}{\hat{R}\cdot T}\left(\Delta g^{\mathrm{E}}+\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_{\mathrm{a}}}-x_{\mathrm{a}}\cdot\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_{\mathrm{a}}}\right) \tag{2.34}$$

Using $x_{\mathrm{b}}=1-x_{\mathrm{a}}$ finally gives

$$\ln\gamma_{\mathrm{a}}=\hat{R}\cdot T\left(\Delta \mathrm{g}^{\mathrm{E}}+x_{\mathrm{b}}\cdot\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_{\mathrm{a}}}\right) \tag{2.35}$$

Similarly,

$$\ln\gamma_{\mathrm{b}}=\hat{R}\cdot T\left(\Delta g^{\mathrm{E}}-x_{\mathrm{a}}\cdot\frac{\partial\Delta g^{\mathrm{E}}}{\partial x_{\mathrm{a}}}\right) \tag{2.36}$$

Relationships for the free excess enthalpy Δg^{E}, which satisfy the condition $\Delta g^{\mathrm{E}}=0$ for $x_i=1$, are required to calculate the activity coefficient γ_i of a mixture from Equation (2.35). Empirical correlations containing two or three correlating parameters are generally used for binary mixtures [11]. However, improved theoretical formulations, usually based on the concept of local composition or on the increment method, are being employed increasingly. An example is the *Wilson equation* for a binary mixture [11]:

$$\frac{\Delta g^{\mathrm{E}}}{\hat{R}\cdot T}= -\left[x_{\mathrm{a}}\cdot\ln\left(x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}\right)+x_{\mathrm{b}}\cdot\ln\left(x_{\mathrm{b}}+A_{\mathrm{ba}}\cdot x_{\mathrm{a}}\right)\right] \quad (2.37)$$

where A_{ab} and A_{ba} are the Wilson coefficients for the binary pair a – b. Hence, the activity coefficients can be represented by

$$\ln\gamma_{\mathrm{a}}= 1-\ln\left(x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}\right)-\frac{x_{\mathrm{a}}}{x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}}-\frac{A_{\mathrm{ba}}\cdot x_{\mathrm{b}}}{x_{\mathrm{b}}+A_{\mathrm{ba}}\cdot x_{\mathrm{a}}}$$
$$\ln\gamma_{\mathrm{b}}= 1-\ln\left(x_{\mathrm{b}}+A_{\mathrm{ba}}\cdot x_{\mathrm{a}}\right)-\frac{x_{\mathrm{b}}}{x_{\mathrm{b}}+A_{\mathrm{ba}}\cdot x_{\mathrm{a}}}-\frac{A_{\mathrm{ab}}\cdot x_{\mathrm{a}}}{x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}} \quad (2.38)$$

The forces of interaction between molecules of a and b can be determined from a few measured values and are described by the coefficients A_{ab} and A_{ba}; these coefficients depend only slightly on pressure.

The Wilson equation for a ternary mixture is

$$\frac{\Delta g^{\mathrm{E}}}{\hat{R}\cdot T}= -x_{\mathrm{a}}\cdot\ln\left(x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}+A_{\mathrm{ac}}\cdot x_{\mathrm{c}}\right)-$$
$$-x_{\mathrm{b}}\cdot\ln\left(A_{\mathrm{ba}}\cdot x_{\mathrm{a}}+x_{\mathrm{b}}+A_{\mathrm{bc}}\cdot x_{\mathrm{c}}\right)-$$
$$-x_{\mathrm{c}}\cdot\ln\left(A_{\mathrm{ca}}\cdot x_{\mathrm{a}}+A_{\mathrm{cb}}\cdot x_{\mathrm{b}}+x_{\mathrm{c}}\right) \quad (2.39)$$

Therefore, the activity coefficient of substance a is

$$\ln\gamma_{\mathrm{a}}= 1-\ln\left(x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}+A_{\mathrm{ac}}\cdot x_{\mathrm{c}}\right)-$$
$$-\frac{x_{\mathrm{a}}}{x_{\mathrm{a}}+A_{\mathrm{ab}}\cdot x_{\mathrm{b}}+A_{\mathrm{ac}}\cdot x_{\mathrm{c}}}-$$
$$-\frac{A_{\mathrm{ba}}\cdot x_{\mathrm{b}}}{A_{\mathrm{ba}}\cdot x_{\mathrm{a}}+x_{\mathrm{b}}+A_{\mathrm{bc}}\cdot x_{\mathrm{c}}}-$$
$$-\frac{A_{\mathrm{ca}}\cdot x_{\mathrm{c}}}{A_{\mathrm{ca}}\cdot x_{\mathrm{a}}+A_{\mathrm{cb}}\cdot x_{\mathrm{b}}+x_{\mathrm{c}}} \quad (2.40)$$

The activity coefficients of substances b and c are obtained analogously.

The Wilson equation for a multicomponent mixture consisting of n components is

$$\frac{\Delta g^{\mathrm{E}}}{\hat{R}\cdot T}= -\sum_{i}^{n} x_i\cdot\ln\left(\sum_{j}^{n} A_{ij}\cdot x_j\right);$$
$$\text{where } A_{ii}=1 \quad (2.41)$$

Hence, the activity coefficients are given by

$$\ln\gamma_i= 1-\ln\left(\sum_{j}^{n} A_{ij}\cdot x_j\right)-$$
$$-\sum_{l}^{n}\frac{A_{li}\cdot x_l}{\sum_{j}^{n} A_{lj}\cdot x_j} \quad (2.42)$$

Even in ternary and higher systems, the Wilson equation contains only the interaction parameters A_{ij} for the binary mixtures involved. Hence, prediction of the phase equilibria of multicomponent systems from binary data is possible. The Wilson equation has proved its worth in practice; the majority of distillation curves presented in Figure 6 were calculated by using this equation. However, unlike the Uniquac (UNiversal QUAsiChemical) equation [12], it is not applicable to systems that undergo phase splitting. Systems E and I in Figure 6 were calculated with the Uniquac equation.

2.4. Calculation of Phase Equilibrium from Equations of State

The method for calculating phase equilibrium from the excess enthalpy, described in Section 2.3, can be applied to systems that are nonideal in the liquid phase only. However, because Dalton's law does not hold at higher pressures, the vapor phase may also deviate from ideal behavior. An effective way of calculating phase equilibria at high pressure is based on equations of state for the fluid, which describe both the vapor and the liquid phases with sufficient accuracy. Partial fugacities f_i are used instead of partial pressures p_i. At equilibrium,

$$f_i''=f_i' \quad (2.43)$$

where f_i'' and f_i' are the fugacities of component i in the vapor and liquid phases, respectively. By analogy with Equation (2.6) and using the fugacity coefficients $\varphi_i=f_i/p_i$, the following equation is obtained:

$$\varphi_i'' \cdot y_i \cdot p = \varphi_i' \cdot x_i \cdot p \tag{2.44}$$

Hence, the equilibrium constant K_i is

$$K_i = \varphi_i' / \varphi_i'' \tag{2.45}$$

Thus, the equilibrium constant can be calculated from fugacity coefficients determined in the two phases. The following rigorous relationship is obtained by using the laws of thermodynamics [11]:

$$\hat{R} \cdot T \cdot \ln\varphi_i = \int_0^p \left(\hat{V}_i - \hat{R} \cdot T/p\right) \cdot \mathrm{d}p \tag{2.46}$$

where the partial molar volume

$$\hat{V}_i = (\partial V / \partial n_i)_{p,T,n_{j \neq i}}$$

and the total volume

$$V = \hat{V}_\mathrm{i} \cdot \sum n_i$$

The molar volume $\hat{V}$ and the derivative $\partial V/\partial n1$ are determined by using a suitable equation of state.

The real behavior of pure fluids can be described by the compressibility factor Z:

$$Z = p \cdot \hat{V} / \left(\hat{R} \cdot T\right) \tag{2.47}$$

For ideal vapors, $Z = 1$. Deviations from this value are directly proportional to the degree of nonideal behavior.

Two different types of equations of state describe the real behavior of pure fluids. One is based on the van der Waals equation:

$$Z = \hat{V} / \left(\hat{V} - \bar{b}\right) - \bar{a} / \left(\hat{R} \cdot T \cdot \hat{V}\right) \tag{2.48}$$

The coefficients $\bar{a}$ and $\bar{b}$ can be interpreted as representing interaction forces and molecular volume, respectively.

The second type of expression is based on the virial equation

$$Z = 1 + B/\hat{V} + C/\hat{V}^2 + \ldots \tag{2.49}$$

Intermolecular forces are described by the virial coefficients B, C, etc., which for pure fluids are a function only of temperature.

Both types of equations have been improved by numerous modifications to obtain a pressure – volume – temperature equation that is applicable to both vapor and liquid phases. Equations of state used to determine phase equilibria must correctly represent mixtures of substances and must include equations for the pure components as special cases. Equations of state for mixtures are generally derived from equations of state for the pure substances by using concentration-dependent coefficients. However, theoretical methods are not available, and empirical rules of combination are used. In principle, equations for excess enthalpy (see Section 2.3) can be applied successfully to the formulation of mixing rules [13].

The procedure can be illustrated by using the Soave modification of the Redlich – Kwong equation (Eq. 2.48) [14]:

$$Z = \hat{V} / \left(\hat{V} - \bar{b}\right) - \bar{a} / \left(\hat{R} \cdot T \cdot \left(\hat{V} + \bar{b}\right)\right) \tag{2.50}$$

The coefficients $\bar{a}$ and $\bar{b}$ can be determined from the critical values p_cr, T_cr, and the acentric factor ω of the pure substances.

$$\begin{aligned} \bar{a} &= 0.42747 \cdot \alpha \cdot \hat{R}^2 \cdot T_\mathrm{cr}^2 / p_\mathrm{cr} \\ \text{where } \alpha &= \left[1 + m\left(1 - \sqrt{T/T_\mathrm{cr}}\right)\right]^2 \\ m &= 0.48 + 1.574 \cdot \omega - 0.176\omega^2 \\ \bar{b} &= 0.08664 \cdot \hat{R} \cdot T_\mathrm{cr} / p_\mathrm{cr} \end{aligned} \tag{2.51}$$

The following mixing rules apply to mixtures of components i and j:

$$\bar{a} = \sum_i \sum_j x_i \cdot x_j \cdot \bar{a}_{ij}$$

Here,

$$\bar{a}_{ij} = \sqrt{\bar{a}_{ii} \cdot \bar{a}_{jj}}\,(1 - k_{ij}), \text{ where } k_{ii} = 0 \tag{2.52}$$

k_{ij} for $i \neq j$ can be found in tables [15].

$$\bar{b} = \sum x_i \cdot \bar{b}_i$$

Substituting Equation (2.50) in Equation (2.46) gives

$$\begin{aligned} ln\varphi_i = &\left[\bar{b}_i / \left(\bar{b} \cdot \hat{R} \cdot T\right)\right] \cdot \left[p \cdot v - \hat{R} \cdot T\right] - \\ &- \ln\left[p / \left(\hat{R} \cdot T\right)(v - \bar{b})\right] - \\ &- \left[a / \left(\bar{b} \cdot \hat{R} \cdot T\right)\right]\left[2/\bar{a} \cdot \Sigma\left(x_j \cdot \bar{a}_{ij}\right) - \bar{b}_i/\bar{b}\right] \cdot \\ &\cdot \ln\left[\left(v + \bar{b}\right)/v\right] \end{aligned} \tag{2.53}$$

Trial values for the concentration and molar volume of each phase must be estimated to solve this equation for each phase. Equation (2.50) is used to improve the values during iteration. The equilibrium constant K for each component can then be calculated by using Equation (2.45) to check the assumed phase concentrations. This iteration method often converges to trivial solutions. However, other available procedures require good starting values close to those of the solution [16].

2.5. Practical Determination of Phase Equilibria

Reliable phase equilibrium data are essential for the design of distillation and rectification operations. Despite the highly advanced state of mixed-phase thermodynamics, good experimental data should be used when possible. Several thousand publications on vapor – liquid equilibria are available which include complete descriptions of the p, v, T, x behavior of a system. However, the majority provide equilibrium concentrations only over a certain range of temperature and pressure. Useful data on azeotropic compositions are readily available [17], mainly in the form of data collections [15], [18–23]. Older data collections consist merely of compilations of measured values; however, newer data are subjected to a consistency test. In addition, coefficients for several equations of state or for equations for excess enthalpy are presented in [15], [23].

Data for many binary mixtures can be found in the above collections. Available data for ternary or higher mixtures over the concentration range of interest are seldom adequate. At the very best, data are provided for binary mixtures present in the system. Usually, equilibrium data are far from complete. In all these cases, the methods of thermodynamic representation of phase equilibria described in Sections 2.3 and 2.4 must be used.

Equations for excess enthalpy are useful at low pressure and when nonideal behavior occurs, mainly in the liquid phase. Equations that use binary interaction parameters to calculate the phase equilibrium of multicomponent mixtures are especially useful. These are primarily equations based on the concept of local composition. Furthermore, the incremental method permits approximation of the interaction parameters from the chemical structure of the substances involved [24].

In high-pressure systems, the liquid phase and especially the vapor phase exhibit nonideal behavior. The convergence pressure method was often used in the past [25] but has now been superseded by the more effective equations of state. Equations of state based on the virial equation have the advantage of being theoretically based for mixtures. However, these equations are valid only at reduced pressure, i.e., $p/p_{\mathrm{cr}} < 0.5$. Equations of the van der Waals type are preferred at higher pressure. Apart from empirical mixing rules, almost all equations of state contain a further binary correlating parameter (k_{ij} in Eq. 2.52), which greatly influences phase equilibrium.

Equations of state are preferred to equations for excess enthalpy, because the former permit calculation not only of phase equilibria but also of other values of state, e.g., density and enthalpy. Empirical mixing rules are not completely satisfactory because they are usually applied to both phases in the same way; this does not correspond fully to real conditions in multicomponent mixtures. The vapor – liquid equilibrium cannot yet be reliably predicted. Progress in mixed-phase thermodynamics permits only a thermodynamically consistent representation of equilibria based on relatively few experimental data.

3. Distillation

As discussed in Chapter 2, the composition of the vapor usually differs from that of the liquid from which it originates. This concentration difference between the two phases is the basis of distillation, which is a simple method of separating substances. The method involves partial evaporation of the liquid followed by condensation of the vapor. The resulting condensate is enriched in the more volatile components, in accordance with the phase equilibrium for the system at hand. Thus, the starting mixture or feed is separated into two fractions with different compositions.

3.1. Continuous Distillation

Figure 8 represents the continuous distillation of a liquid mixture. The *feed* (molar flow rate $\dot{F}$) is separated into an *overhead product* or *distillate* (molar flow rate $\dot{D}$) and a *bottom product* (molar flow rate $\dot{B}$). The composition difference between these two products depends on the phase equilibrium.

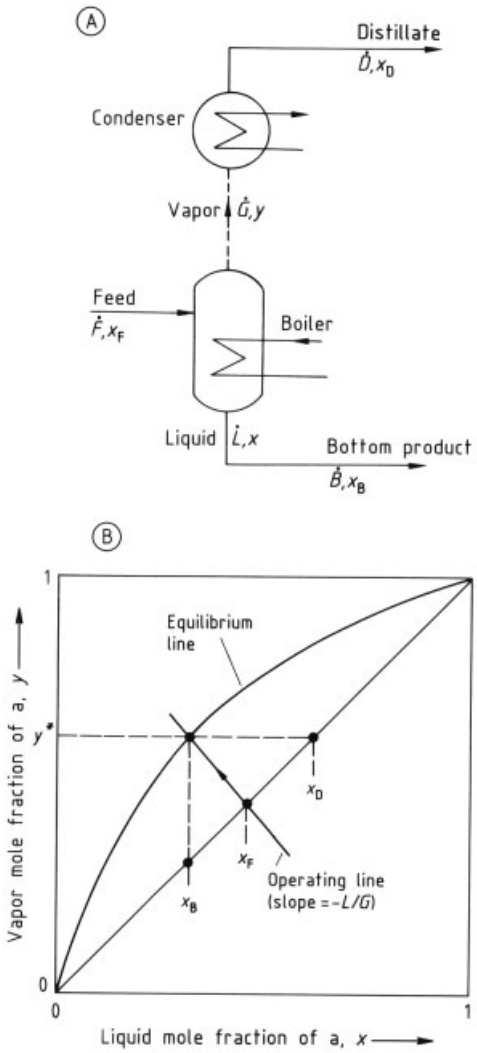

Figure 8. Continuous distillation of a binary mixture a – b A) Flow diagram showing symbols for total molar streams or flow rates ($\dot{F}$, $\dot{L}$, etc.) and mole fractions of the more volatile component a (x_F, x_D, etc.); B) A y – x diagram showing equilibrium and operating linesThe y – x diagram can be used to determine the composition of the overhead and bottom product streams ($\dot{D}$ and $\dot{B}$, respectively).

3.1.1. Continuous Distillation of Binary Mixtures

The material balance around the boiler (see Fig. 8) yields

$$\dot{F} = \dot{G} + \dot{L} \tag{3.1}$$

where $\dot{F}$, $\dot{G}$, and $\dot{L}$ are the molar flow rates of the feed, vapor, and bottom liquid, respectively. Similarly, the material balance for the more volatile feed component a of a binary mixture a – b gives

$$\dot{F} \cdot x_F = \dot{G} \cdot y + \dot{L} \cdot x \tag{3.2}$$

where x_F is the mole fraction of a in the feed, y is the mole fraction of a in the vapor, and x is the mole fraction of a in the bottom liquid. Thus,

$$y = -\dot{L}/\dot{G} \cdot x + \left(1 + \dot{L}/\dot{G}\right) \cdot x_F \tag{3.3}$$

This is the equation for a straight line with slope $-\dot{L}/\dot{G}$. When plotted on the $y-x$ diagram, it furnishes the *operating line.* The operating line intersects the $y = x$ diagonal at $x = x_F$ (Fig. 8 B).

With the closely met assumption that equilibrium exists between the vapor $\dot{G}$ and the liquid $\dot{L}$ in the boiler, (i.e., $y = y*$), the following equation holds for ideal mixtures (see Section 2.2.2):

$$y^* = \frac{\alpha \cdot x}{1 + (\alpha - 1) \cdot x} \tag{3.4}$$

where α is the relative volatility.

The parameters x and y can be determined by equating relationships (3.3) and (3.4) to obtain a quadratic equation which can be solved easily. However, a graphical solution based on the $y-x$ diagram is usually preferred because it is a visual method that can also be easily applied to nonideal mixtures. The desired concentrations of vapor y and liquid x can be determined from the point of intersection of the operating lines with the equilibrium line. Subsequent total condensation of the vapors is not accompanied by a change in concentration; i.e., the distillate concentration x_D is the same as the vapor concentration y. This is represented graphically in the $y-x$ diagram by the intercept of the horizontal line between $y*$ and x_D (i.e., y is constant) with the diagonal. Figure 8 B clearly shows the large

concentration difference between the heads x_D and the bottoms x_B. The decisive factor here is the position of the equilibrium curve, i.e., for ideal systems, the relative volatility α. Generally, only a limited difference between the concentrations of the two fractions can be obtained by simple continuous distillation.

3.1.2. Continuous Distillation of Multicomponent Mixtures

The calculations for multicomponent systems are identical with those used for binary mixtures. The material balance for the apparatus gives the operating line for each substance *i*:

$$y_i = -\dot{L}/\dot{G} \cdot x_i + \left(1 + \dot{L}/\dot{G}\right) \cdot x_{Fi} \tag{3.5}$$

where x_{Fi} is the mole fraction of *i* in the feed. The phase equilibrium is effectively defined by using the equilibrium constant K_i. Thus,

$$y_i^* = K_i \cdot x_i \tag{3.6}$$

For ideal mixtures, the equilibrium constant K_i is a function only of absolute temperature *T*, which is at first unknown. Combining Equations (3.5) and (3.6) gives the following relationship for each component *i*:

$$y_i = \frac{x_{Fi} \cdot \left(1 + \dot{L}/\dot{G}\right)}{1 + \frac{\dot{L}/\dot{G}}{K_i(T)}} \tag{3.7}$$

From the mole fraction constraint $\sum y_i = 1$,

$$\sum_{i=1}^{k} \frac{x_{Fi} \cdot \left(1 + \dot{L}/\dot{G}\right)}{1 + \frac{\dot{L}/\dot{G}}{K_i(T)}} = 1 \tag{3.8}$$

This set of equations must be solved iteratively. A trial value for the boiling temperature is first assumed, and the equilibrium constant K_i for each component is then determined. The estimation is checked and corrected by using Equation (3.8). The individual summation terms of Equation (3.8) give the vapor concentration y_i of each component involved.

Iterations involving the boiling temperature can be avoided by using the relative volatility α_{ik}, instead of the equilibrium constant K_i. Equation (3.7) is then solved for K_i, the equilibrium constant, as follows:

$$K_i = \frac{\dot{L}/\dot{G}}{x_{Fi}/y_i \cdot \left(1 + \dot{L}/\dot{G}\right) - 1} \tag{3.9}$$

This relationship applies to each component of the mixture; it also applies to the highest boiling substance *k*.

$$K_k = \frac{\dot{L}/\dot{G}}{x_{Fk}/y_k \cdot \left(1 + \dot{L}/\dot{G}\right) - 1} \tag{3.10}$$

Using the relationship $\alpha_{ik} = K_i/K_k$ gives

$$\alpha_{ik} = \frac{x_{Fk}/y_k \cdot \left(1 + \dot{L}/\dot{G}\right) - 1}{x_{Fi}/y_i \cdot \left(1 + \dot{L}/\dot{G}\right) - 1} \tag{3.11}$$

Solving for y_i and employing the mole fraction constant $\sum y_i = 1$ yield

$$\sum_{i=1}^{k} \frac{\alpha_{ik} \cdot x_{Fi} \left(1 + \dot{L}/\dot{G}\right)}{x_{Fk}/y_k \cdot \left(1 + \dot{L}/\dot{G}\right) + \alpha_{ik} - 1} = 1 \tag{3.12}$$

This relationship is equivalent to Equation (3.8) but employs the gas concentration y_k of the highest boiling component as the iteration variable instead of the boiling temperature *T*. Calculation of this method is identical with that of Equation (3.8).

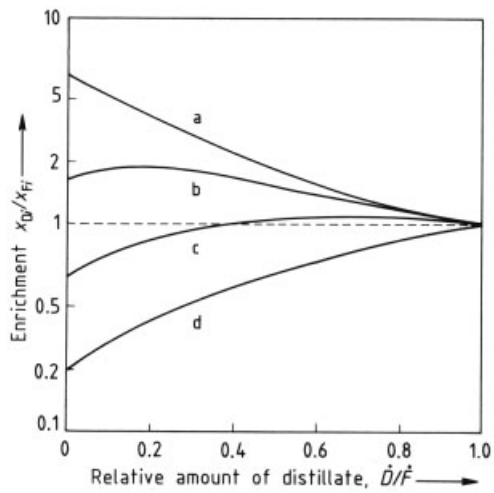

Figure 9. Continuous distillation of a mixture of butane (a), pentane (b), hexane (c), and heptane (d)
x_{Di} = liquid mole fraction of i in distillate; x_{Fi} = liquid mole fraction of i in feed; $\dot{D}$ = distillate stream; $\dot{F}$ = feed stream

Analysis of a quaternary system (a, b, c, d) is illustrated in Figure 9. The ratio of the liquid distillate concentration to the feed concentration

x_{Di}/x_{Fi} (i.e., enrichment) is plotted against the relative amount of distillate $\dot{D}/\dot{F}$. The best separation is achieved at low $\dot{D}/\dot{F}$ values. Obviously separation diminishes as the concentration ratio $\dot{D}/\dot{F}$ approaches unity. Both medium boilers b and c pass through a concentration maximum during distillation.

3.2. Batch Distillation

A schematic of batch distillation is shown in Figure 10. The vapor stream $\dot{G}$ produced by heating L moles of liquid is continuously removed and condensed. The distillate (D) is usually collected in several receivers.

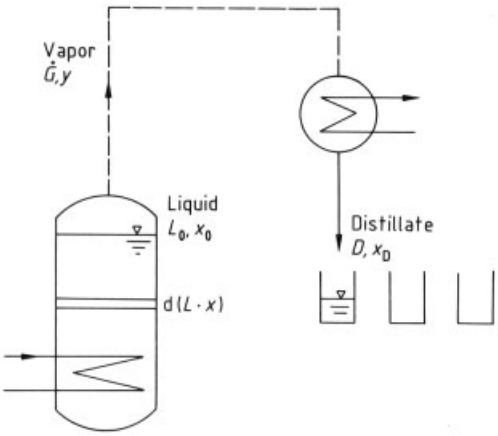

Figure 10. Schematic of batch distillation

3.2.1. Batch Distillation of Binary Mixtures

Material balance around the apparatus permits calculation of the batch distillation of a binary mixture of substances a and b. For a differential time element dt, the following holds:

$$\dot{G} \cdot \mathrm{d}t + \mathrm{d}L = 0 \tag{3.13}$$

and

$$\dot{G} \cdot y \cdot \mathrm{d}t + \mathrm{d}\,(L \cdot x) = 0$$

where the vapor and liquid mole fractions x and y refer to the more volatile component a. By using $\dot{G} = \mathrm{d}G/\mathrm{d}t$, the Rayleigh equation is obtained [26]:

$$\frac{\mathrm{d}L}{L} = \frac{\mathrm{d}x}{y^* - x} \tag{3.14}$$

For ideal mixtures, the equilibrium concentration of a in the vapor phase, $y*$, can be represented by

$$y^* = \frac{\alpha \cdot x}{1 + (\alpha - 1) \cdot x} \tag{3.15}$$

An analytical solution to the differential equation is obtained by substitution into Equation (3.14):

$$\frac{L}{L_0} = \left[\frac{x}{x_0}\right]^{1/(\alpha-1)} \cdot \left[\frac{1-x_0}{1-x}\right]^{\alpha/(\alpha-1)} \tag{3.16}$$

where the subscript 0 represents the liquid in its initial state.

Typical changes in the concentrations of the more volatile component a in the vapor and in the liquid are plotted against the relative amount of distillate $D/L_0 = 1 - L/L_0$ in Figure 11. During the operation, both phases are increasingly depleted of substance a. The intermediate distillate concentration x_{Dm} ($= y_m$) is given by

$$x_{Dm} = x + \frac{x_0 - x}{D/L_0} \tag{3.17}$$

This equation can be interpreted visually with the help of Figure 11. Even for nonideal mixtures, the intermediate distillate concentration x_{Dm} can be determined easily.

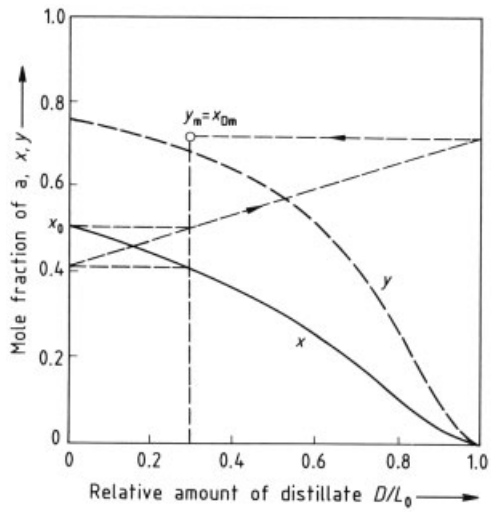

Figure 11. Batch distillation of a binary mixture a – b showing concentrations of the more volatile component a in the boiler (x) and in the vapor or distillate (y) as a function of the relative amount of distillate D/L_0
The intermediate distillate concentration x_{Dm} can be determined graphically (see Eq. 3.17).

3.2.2. Batch Distillation of Multicomponent Mixtures

The Rayleigh equation (Eq. 3.14) can be applied directly to multicomponent mixtures. The following holds for each component *i* of a system:

$$\frac{\mathrm{d}L}{L} = \frac{\mathrm{d}x_i}{y_i^* - x_i} \tag{3.18}$$

Numerical methods are generally used to solve this set of equations because the equilibrium concentration $y*_i$ is usually influenced by all the substances present in the mixture.

An analytical solution is obtained under ideal conditions by employing the equilibrium constant $K_i = y*_i/x_i$ which is then independent of the other components of the system. Hence, the following equation applies:

$$\frac{\mathrm{d}L}{L} = \frac{\mathrm{d}x_i}{x_i \cdot (K_i - 1)} \tag{3.19}$$

Rearrangement gives

$$(K_i - 1) \cdot \mathrm{d}\ln L = \mathrm{d}\ln x_i \tag{3.20}$$

Integration of each component gives

$$(K_i - 1) \cdot \ln (L/L_0) = \ln (x_i/x_{i0}) \tag{3.21}$$

where the index 0 denotes the starting condition. Solving for the concentration x_i gives

$$x_i = x_{i0} \cdot (L/L_0)^{K_i^{-1}} \tag{3.22}$$

For the mole fraction constraint $\Sigma x_i = 1$, the following is obtained:

$$\sum_{i=1}^{k} x_{i0} \cdot (L/L_0)^{K_i^{-1}} = 1 \tag{3.23}$$

To solve Equation (3.23) for given L/L_0 values, the boiling temperature of the mixture must be estimated. The vapor pressure of each substance $p^0{}_i$, and consequently the equilibrium constant $K_i = p^0{}_i/p$, can then be determined. The temperature estimation is checked and corrected iteratively by employing the above relationship, whereby convergence is usually attained rapidly.

Alternatively, temperature iteration can be avoided by using the relative volatility α_{ik} instead of the equilibrium constant K_i. By analogy with Equations (3.9) – (3.12):

$$\sum_{i=1}^{k} \left[x_{i0} \cdot (L/L_0)^{\alpha_{ik}-1} \cdot (x_k/x_{k0})^{\alpha_{ik}} \right] = 1 \tag{3.24}$$

In this case, instead of absolute temperature *T*, the mole fraction of the high boiler x_k is the iteration variable. Equations (3.23) and (3.24) describe the changes in concentration of the liquid in the boiler. Thus, the distillate composition can be determined with the help of the equilibrium constant K_i.

Variation of liquid and vapor concentration with the relative amount of distillate D/L_0 for a quaternary mixture is shown in Figure 12. The concentration of the most volatile component (a) decreases first because it is released preferentially into the vapor phase. Concentrations of the higher boiling substances in the boiler consequently increase. When most of substance a is separated ($D/L_0 \approx 0.7$ in Fig. 12), substance b is the most volatile component, so its concentration decreases similarly. Both medium-boiling substances b and c pass through a concentration maximum in the course of distillation. The distillate is collected in several receivers, which enable the recovery of fractions of varying composition. However, complete separation of a liquid mixture into pure components is seldom achieved. Batch rectification must usually be employed for this (see Chap. 5).

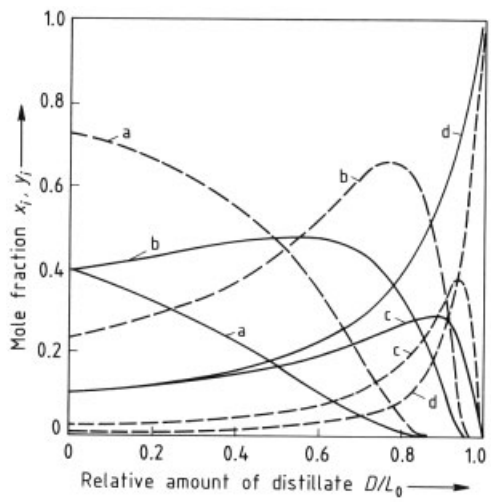

Figure 12. Course of concentration changes during the batch distillation of a mixture of butane (a), pentane (b), hexane (c), and heptane (d)
—— Mole fraction of component in the liquid in the boiler, x_i; – – – Mole fraction of component in the distillate, y_i

3.3. Semicontinuous Distillation

Semicontinuous distillation can often be used to advantage with liquids containing small amounts of impurities with low volatility. As shown schematically in Figure 13, feed is fed continuously into the boiler at a rate $\dot{F}$, and head product is withdrawn continuously at a rate $\dot{D}$. During the course of distillation, low-volatility impurities collect in the boiler and must be drained off periodically.

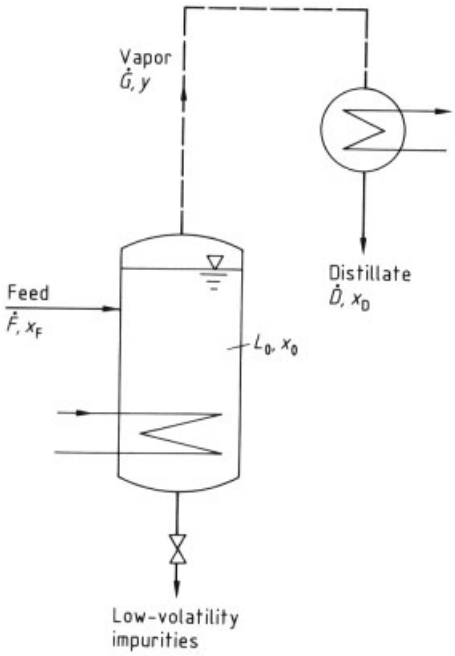

Figure 13. Schematic of semicontinuous distillation

3.3.1. Semicontinuous Distillation of Binary Mixtures

As in Equation (3.14), a material balance for a binary mixture yields

$$\frac{\mathrm{d}F}{L_0} = \frac{\mathrm{d}x}{x_\mathrm{F} - y^*(x)} \tag{3.25}$$

where mole fractions refer to the more volatile component a. For ideal mixtures, integration gives

$$\frac{F}{L_0} = \frac{\alpha - 1}{x_\mathrm{F}(\alpha - 1) - \alpha} \cdot (x - x_\mathrm{F}) - \\ - \frac{\alpha}{[x_\mathrm{F}(\alpha - 1) - \alpha]^2} \cdot \\ \cdot \ln \left\{ \frac{x[x_\mathrm{F}(\alpha - 1) - \alpha] + x_\mathrm{F}}{x_\mathrm{F}[x_\mathrm{F}(\alpha - 1) - \alpha] + x_\mathrm{F}} \right\} \tag{3.26}$$

Hence, the concentration changes of impurities in the boiler and in the distillate can be determined. Variation with time t of the concentration is given in the relationship $F = \dot{F} \cdot t$. In the special case of nonvolatile impurites (i.e., $\alpha \to \infty$), Equation (3.26) is simplified as follows:

$$F/L_0 = \frac{x - x_\mathrm{F}}{x_\mathrm{F} - 1} \tag{3.27}$$

Rearrangement gives

$$x = x_\mathrm{F} + F/L_0 \cdot (x_\mathrm{F} - 1) \tag{3.28}$$

Thus, the concentration x of the more volatile substance in the boiler decreases linearly with the amount of feed, while that of the less volatile substance $(1 - x)$ increases linearly.

3.3.2. Semicontinuous Distillation of Multicomponent Mixtures

In combination with a suitable equilibrium relationship, Equation (3.25) can also be used to describe multicomponent mixtures. In general, a set of differential equations is obtained, which are coupled by means of the equilibrium concentrations. However, this does not apply to ideal mixtures because the equilibrium constant K_i of each component is independent of other substances in the mixture. Individual equations are related to each other only via temperature.

Thus, for ideal solutions,

$$\frac{\mathrm{d}F}{L_0} = \frac{\mathrm{d}x_i}{x_{\mathrm{F}i} - y_i^*} = \frac{\mathrm{d}x_i}{x_{\mathrm{F}i} - K_i \cdot x_i} \tag{3.29}$$

Rearrangement gives

$$\int_{F=0}^{F} \frac{\mathrm{d}F}{L_0} = -\frac{1}{K_i} \int_{x_i = x_{\mathrm{F}i}}^{x_i} \cdot \mathrm{dln}\,(K_i \cdot x_i - x_{\mathrm{F}i}) \tag{3.30}$$

After integration, the following equation is obtained:

$$\frac{F}{L_0} = -\frac{1}{K_i} \cdot \ln \frac{K_i \cdot x_i - x_{\mathrm{F}i}}{x_{\mathrm{F}i} \cdot (K_i - 1)} \tag{3.31}$$

Solving for the concentration of component i and using the mole fraction constraint $\Sigma x_i = 1$ yields

$$\sum_{i=1}^{K} \frac{1}{K_i} \cdot \left[x_{\mathrm{F}i} + x_{\mathrm{F}i} \cdot (K_i - 1) \cdot \exp\left(-\frac{F}{L_0} \cdot K_i \right) \right] = 1 \quad (3.32)$$

This equation can be solved iteratively, the temperature T being the iteration variable. The individual summation terms give the concentrations x_i in the boiler.

4. Continuous Rectification (Multiple Distillation)

4.1. Principles

As discussed in Chapter 3, only limited separation of components is possible in simple distillation. The composition difference between the fractionation products is often very small; therefore, pure substances are seldom obtained.

Figure 14 represents a multiple distillation, i.e., rectification. The overhead product of the first distillation is subjected to a second distillation, whose overhead product, in turn, is subjected to a third distillation and so on. Like Figure 8, the vapor and liquid concentrations at each step can be determined easily on the $y-x$ diagram (Fig. 14 B).

This process can yield a very pure fraction, but it has two serious disadvantages:

1) Only a small part of the starting mixture is recovered in a highly concentrated state.
2) Byproducts are obtained at each stage, which often cannot be processed further.

Both disadvantages can be avoided if the liquid from each distillation step is returned to the previous step, as shown in Figure 15 A. The vapor does not have to be condensed before each step nor does the liquid have to be subsequently revaporized (Fig. 15 B). On the contrary, vapor and liquid can be brought into direct contact. In fact, intensive countercurrent contact between the vapor and liquid streams at each stage is important so that intensive mass transfer can take place. The concentrations attainable at each step can be determined on the $y-x$ diagram by means of material balances around the top of each column, as shown in Figure 16.

Two methods are usually employed to evaluate countercurrent separation efficiency: the equilibrium-stage and the transfer-unit concepts.

4.1.1. Equilibrium-Stage Concept

The equilibrium-stage concept assumes that the vapor and liquid phases are indistinct mass-transfer elements, where they are brought into such intimate contact that they are in equilibrium with each other at each stage. The phases are subsequently separated; vapor flows up to the next stage and liquid runs down to the stage below (countercurrent phase flow), as shown in Figure 16. A material balance for the more volatile component around the top of the column gives the following relationship for the vapor and liquid concentrations (x, y) and streams ($\dot{G}$, $\dot{L}$) between the individual stages:

$$y_{n-1} = \frac{\dot{L}_n}{\dot{G}_{n-1}} \cdot x_n + \frac{\dot{G}_0}{\dot{G}_{n-1}} \cdot y_0 - \frac{\dot{L}_0}{\dot{G}_{n-1}} \cdot x_0 \quad (4.1)$$

The subscripts 0, n, and $n-1$ refer to the stages from which the mass stream originates, 0 being the initial stage. When Equation (4.1) is plotted on the $y-x$ diagram, it furnishes the operating line (Fig. 16 B). The phase equilibrium can also be represented by a line on the $y-x$ diagram. The liquid concentration x_n is connected to the vapor concentration y_{n-1} by a balance line; the concentrations x_n and y_{n-1} define a point on the operating line. Equilibrium exists between y_n, the vapor above stage n, and x_n the liquid concentration in the same stage. Therefore, a vertical line at the liquid concentration x_n must intersect the equilibrium curve at point y_n. Likewise, a horizontal line at the vapor concentration y_n must pass through the point x_{n+1} on the operating line. The point of intersection of x_{n+1} with the equilibrium curve gives the vapor concentration y_{n+1} of the stage above and so on. Alternate use of the equilibrium curve and the operating line to "step" from one equilibrium stage to the other is illustrated on the $y-x$ diagram in Figure 16 B. The operating line is the locus of states between the individual stages.

The number of equilibrium stages n required to achieve a desired concentration change is a good index of the difficulty of separation. The number of equilibrium stages, represented by steps, can easily be determined graphically

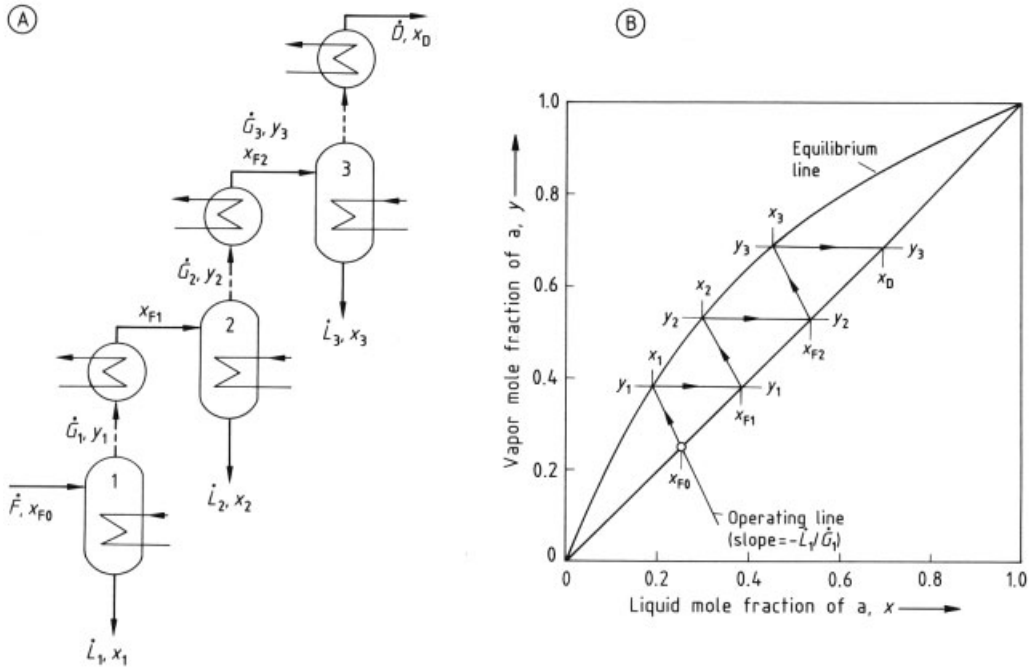

Figure 14. Multiple distillation of a binary mixture a – b
A) Flow diagram; B) Determination of vapor and liquid mole fractions of component a at each stage on a y – x diagram

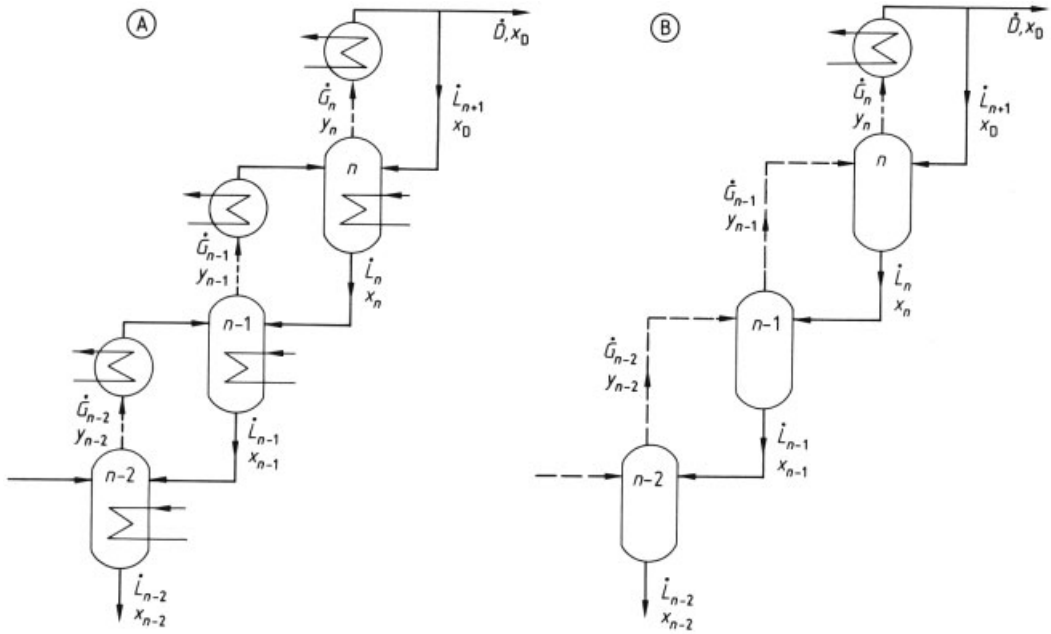

Figure 15. Improvement of multiple distillation by returning liquid from each distillation step to the previous step with (A) or without (B) condensation

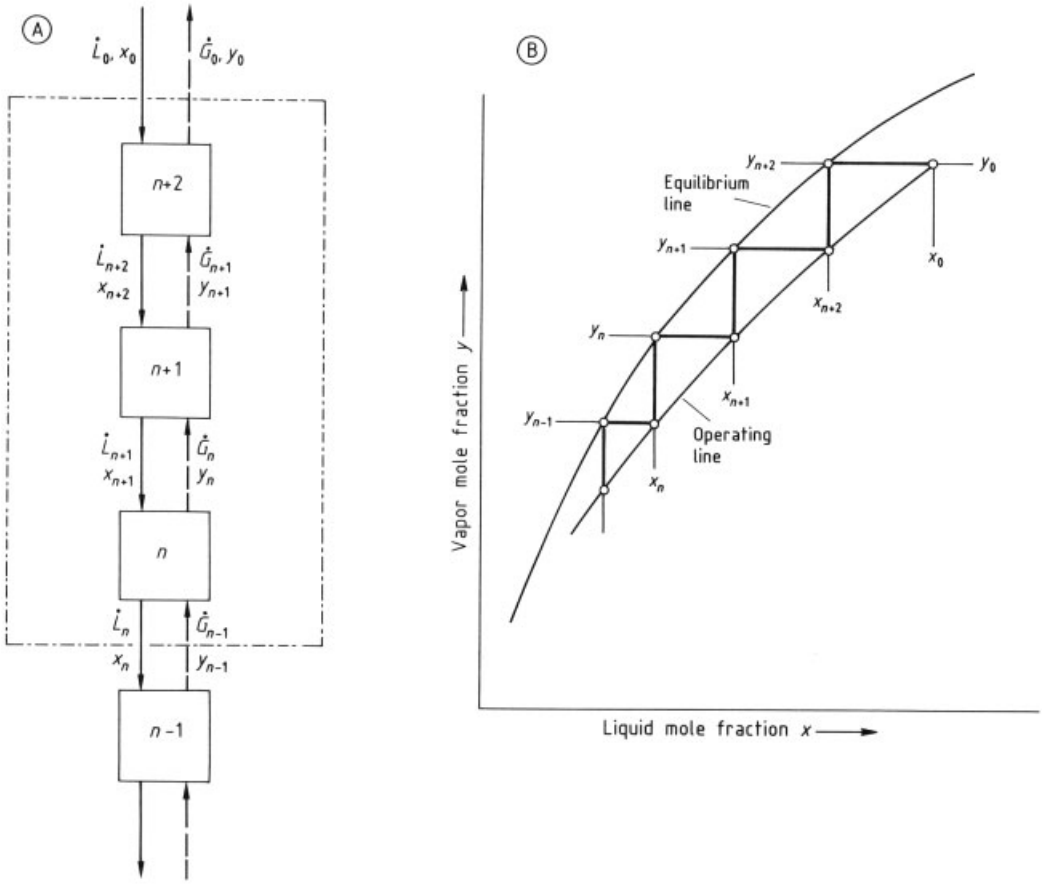

Figure 16. Equilibrium-stage concept for determining countercurrent separation efficiency
A) Schematic showing equilibrium stages and associated parameters: dotted – dashed line indicates the boundary for the material balance described in Equation (4.1).
B) McCabe – Thiele diagram: concentrations at each equilibrium stage and number of stages can be determined by stepping off the stages between the equilibrium line and the operating line.

on the $y-x$ diagram; this is often called the McCabe – Thiele method. Analytical solutions are available for a few special cases. A particularly important case occurs when both the equilibrium curve and the operating line are straight lines; the number of equilibrium stages n is given by [27]

$$n = -\frac{1}{\ln \frac{m}{\dot{L}/\dot{G}}} \ln \left[\left(\frac{m}{\dot{L}/\dot{G}} - 1 \right) \frac{y - y_0}{y_0^* - y_0} + 1 \right] \tag{4.2}$$

where m is the slope of the equilibrium line. When $m/(L/G) = 1$,

$$n = -\frac{y - y_0}{y_0^* - y_0} \tag{4.3}$$

Equation (4.2) can often be used to determine the number of stages in the separation of minute impurities.

4.1.2. Transfer-Unit Concept

The concept of transfer units, conceived by Chilton and Colburn, considers a differential transfer area dA of the column [28]. It is illustrated schematically in Figure 17 A. Mass transfer can be formulated as follows:

$$d\dot{N} = K_{\mathrm{OG}} \cdot (y^* - y) \cdot \mathrm{d}A \tag{4.4}$$

where K_{OG} is the overall mass transfer coefficient. The molar mass stream d$\dot{N}$ causes a change in vapor composition dy:

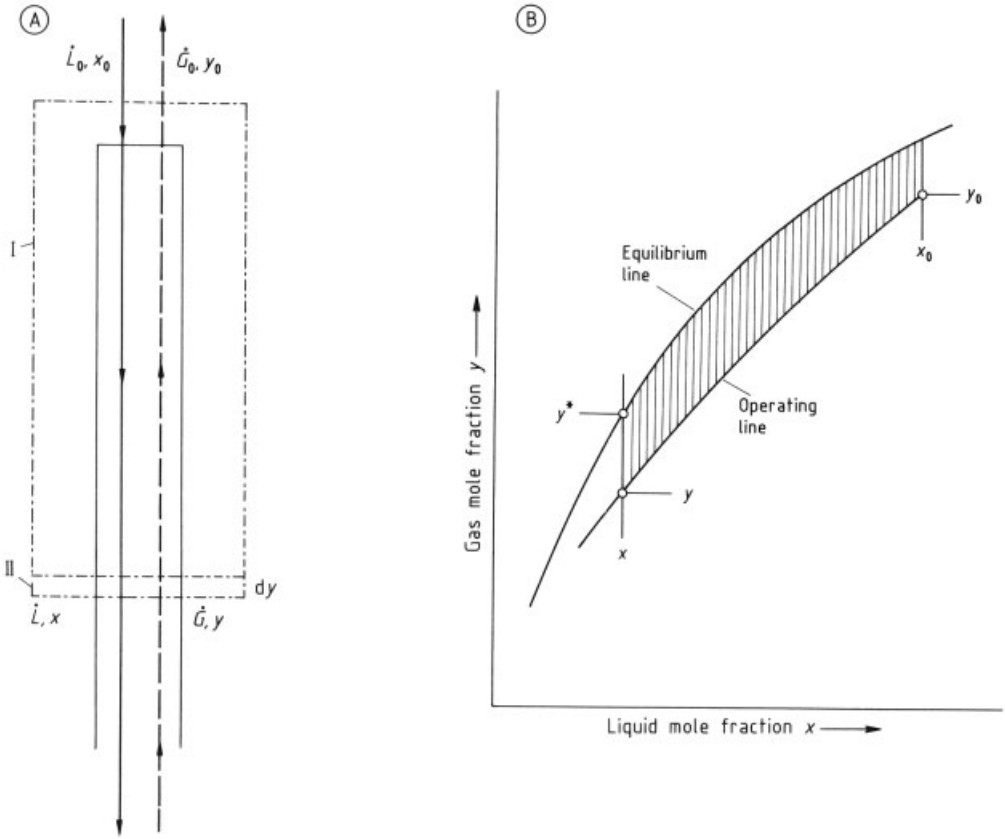

Figure 17. Transfer-unit concept for determining countercurrent separation efficiency
A) Flow diagram showing important parameters and material balance boundaries (I, II)
B) A y – x diagram: the number of transfer units N_{OG} is determined by integrating between the equilibrium line and the operating line (shaded area)

$$\dot{G} \cdot \mathrm{d}y = \mathrm{d}\dot{N} \tag{4.5}$$

Thus,

$$\dot{G} \cdot \mathrm{d}y = K_{OG} \cdot (y^* - y) \cdot \mathrm{d}A \tag{4.6}$$

Separating variables gives

$$\frac{\mathrm{d}y}{y^* - y} = \frac{K_{OG} \cdot \mathrm{d}A}{\dot{G}} \tag{4.7}$$

The term on the right-hand side of Equation (4.7) contains factors pertaining mainly to the apparatus, which are of interest only for column design (see Chap. 8). The difficulty of separation is characterized sufficiently by the term on the left-hand side, which contains factors pertaining mainly to the process itself. It is defined as the overall number of gas-phase transfer units N_{OG}:

$$\mathrm{d}N_{OG} = \frac{\mathrm{d}y}{y^* - y} \tag{4.8}$$

This term can be interpreted as the ratio of the desired concentration change dy to the driving force ($y* - y$) available for mass transfer. At a given liquid state x, y lies on the operating line and $y*$ on the equilibrium curve. The shaded area in Figure 17 B represents the driving force ($y* - y$).

The difficulty of the separation to be accomplished is characterized in terms of transfer units N_{OG}. If N_{OG} is very large a very tall column is required.

The number of transfer units N_{OG} is usually determined by integrating between the operating line and the equilibrium curve, either graphically

or numerically. Analytical solutions are available in a few special cases. Again, the solution when both the equilibrium curve and the operating line are straight lines on the $y-x$ diagram is important [27]:

$$N_{\mathrm{OG}} = -\frac{1}{\frac{m}{\dot{L}/\dot{G}} - 1} \cdot$$

$$\cdot \ln\left[\left(\frac{m}{\dot{L}/\dot{G}} - 1\right) \cdot \frac{y - y_0}{y_0^* - y_0} + 1\right] \tag{4.9}$$

When $m/(L/G) = 1$,

$$N_{\mathrm{OG}} = -\frac{y - y_0}{y_0^* - y_0} \tag{4.10}$$

Equation (4.9) is comparable to Equation (4.2) for the number of equilibrium steps n.

4.1.3. Comparison of Equilibrium-Stage and Transfer-Unit Concepts

The equilibrium-stage concept for characterization of the difficulty of separation provides an adequate measure of separability for distinct mass-transfer elements in countercurrent columns, e.g., in a highly efficient plate column. On the other hand, the transfer-unit concept is preferred when vapor and liquid are in continuous contact, e.g., in packed columns and plate columns which have an efficiency well below 1. The equilibrium-stage concept is used almost exclusively in practice because graphical determination of the number of equilibrium steps on the $y-x$ diagram is much easier than integration with respect to the driving force.

The equilibrium curve and the operating line can almost always be linearized in narrow concentration ranges. The following relationship is then obtained:

$$N_{\mathrm{OG}} = n \cdot \frac{\ln \frac{m}{\dot{L}/\dot{G}}}{\frac{m}{\dot{L}/\dot{G}} - 1} \tag{4.11}$$

In the special case when $m/(\dot{L}/\dot{G}) \to 1, N_{\mathrm{OG}} \approx n$. This condition is often satisfied to a close approximation in difficult separations that require many equilibrium stages or transfer units.

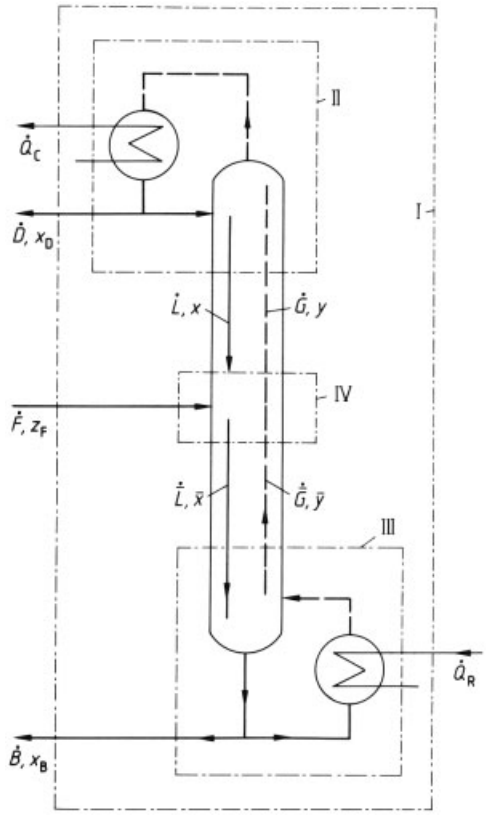

Figure 18. Schematic of a continuously operated rectification column used to separate a binary mixture
Part of the overhead product and part of the bottom product are returned to the column to maintain countercurrent flow. Boundaries for material balances discussed in the text are indicated by dotted – dashed lines (I – IV).

4.2. Rectification of Binary Mixtures

A continuously operated rectification column is shown schematically in Figure 18. The binary feed material (molar flow rate $\dot{F}$) is introduced into the middle of the column. The distillate (molar flow rate $\dot{D}$), which mainly contains the lower boiling component a, is withdrawn from the top of the column. The bottom product (molar flow rate $\dot{B}$), which contains the higher boiling component b in greater concentration, is removed from the bottom of the column. Some overhead and bottom products are returned to the column to maintain countercurrent flow of vapor and liquid in the column. As discussed in Chapter 3, this is a prerequisite for the separation of feed

into substances a and b with maximum purity. The sections of the column above and below the feed are called the *rectification section* and the *stripping section,* respectively.

4.2.1. Prediction of Rectification Based on Material Balance

Rectification can be predicted simply with the help of material balances by assuming that the vapor and liquid streams ($\dot{G}$ and $\dot{L}$, respectively) in the column do not change (see Section 4.2.2).

The overall column balance (boundary I in Fig. 18) yields

$$\dot{F}=\dot{D}+\dot{B}$$
$$\dot{F}\cdot z_{\mathrm{F}}=\dot{D}\cdot x_{\mathrm{D}}+\dot{B}\cdot x_{\mathrm{B}} \tag{4.12}$$

where z_{F}, x_{D}, and x_{B} are the mole fractions of the more volatile substance a in the feed, distillate, and bottom product, respectively. Rearrangement gives

$$\dot{D}=\dot{F}\,\frac{z_{\mathrm{F}}-x_{\mathrm{B}}}{x_{\mathrm{D}}-x_{\mathrm{B}}} \text{ and } \dot{B}=\dot{F}\cdot\frac{x_{\mathrm{D}}-z_{\mathrm{F}}}{x_{\mathrm{D}}-x_{\mathrm{B}}} \tag{4.13}$$

These values can be used to determine the quantities of distillate $\dot{D}$ and bottom product $\dot{B}$ for given product specifications. A material balance for the top of the column (boundary II in Fig. 18) gives a relationship for the states in the column:

$$\dot{G}=\dot{L}+\dot{D}$$
$$\dot{G}\cdot y=\dot{L}\cdot x+\dot{D}\cdot x_{\mathrm{D}} \tag{4.14}$$

where $\dot{L}$ is the molar stream of the reflux liquor returned to the column. Thus,

$$y=\frac{\dot{L}}{\dot{G}}\cdot x+\left(1-\frac{\dot{L}}{\dot{G}}\right)\cdot x_{\mathrm{D}} \tag{4.15}$$

The exact position in the column to which balance II applies is not defined. Hence, Equation (4.15) gives the relationship between the vapor concentration y and the liquid concentration x at any point in the column above the point of feed entry, i.e., in the rectifying section. Because $\dot{L}/\dot{G}$ is constant, Equation (4.15) represents a straight line–the rectifying operating line–with a slope of $\dot{L}/\dot{G}$ on the $y-x$ diagram, which intersects the $y=x$ diagonal at x_{D} (Fig. 19).

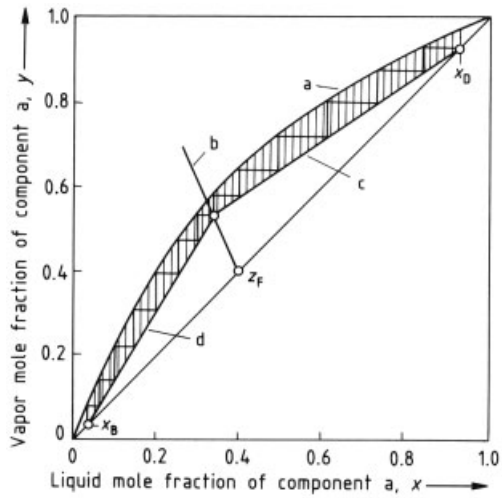

Figure 19. McCabe – Thiele diagram for a binary mixture a – b
Shaded area between the operating line and the equilibrium curve is decisive for operation. Equilibrium stages are also shown.
a) Equilibrium line; b) Feed or φ line; c) Rectifying operating line; d) Stripping operating line

The ratio $\dot{L}/\dot{G}$ is known as the *internal reflux ratio.* The use of the *external reflux ratio R* is more advantageous:

$$R=\frac{\dot{L}}{\dot{D}} \tag{4.16}$$

Thus Equation (4.15) becomes

$$y=\frac{R}{R+1}\cdot x+\frac{1}{R+1}\cdot x_{\mathrm{D}} \tag{4.17}$$

Analogously, material balance III around the bottom of the column yields the following relationship:

$$\bar{y}=\frac{\bar{\dot{L}}}{\bar{\dot{G}}}\cdot\bar{x}+\left(1-\frac{\bar{\dot{L}}}{\bar{\dot{G}}}\right)\cdot x_{\mathrm{B}} \tag{4.18}$$

where the horizontal bar ($\bar{y}$, $\bar{x}$, etc.) denotes parameters below the feed point, i.e., in the stripping section. Use of the *external reboil ratio* $\bar{R}$, according to the definition $\bar{R}=\bar{\dot{G}}/\dot{B}\bar{R}=\bar{G}\cdot/\dot{B}$, gives

$$\bar{y}=\frac{\bar{R}+1}{\bar{R}}\cdot\bar{x}-\frac{1}{\bar{R}}\cdot x_{\mathrm{B}} \tag{4.19}$$

Equation (4.19) defines the relationship between the vapor concentration $\bar{y}$ and the liquid concentration $\bar{x}$. If the flow rates $\bar{G}\cdot$ and $\bar{L}\cdot$ are assumed to be constant, it represents a straight line, called

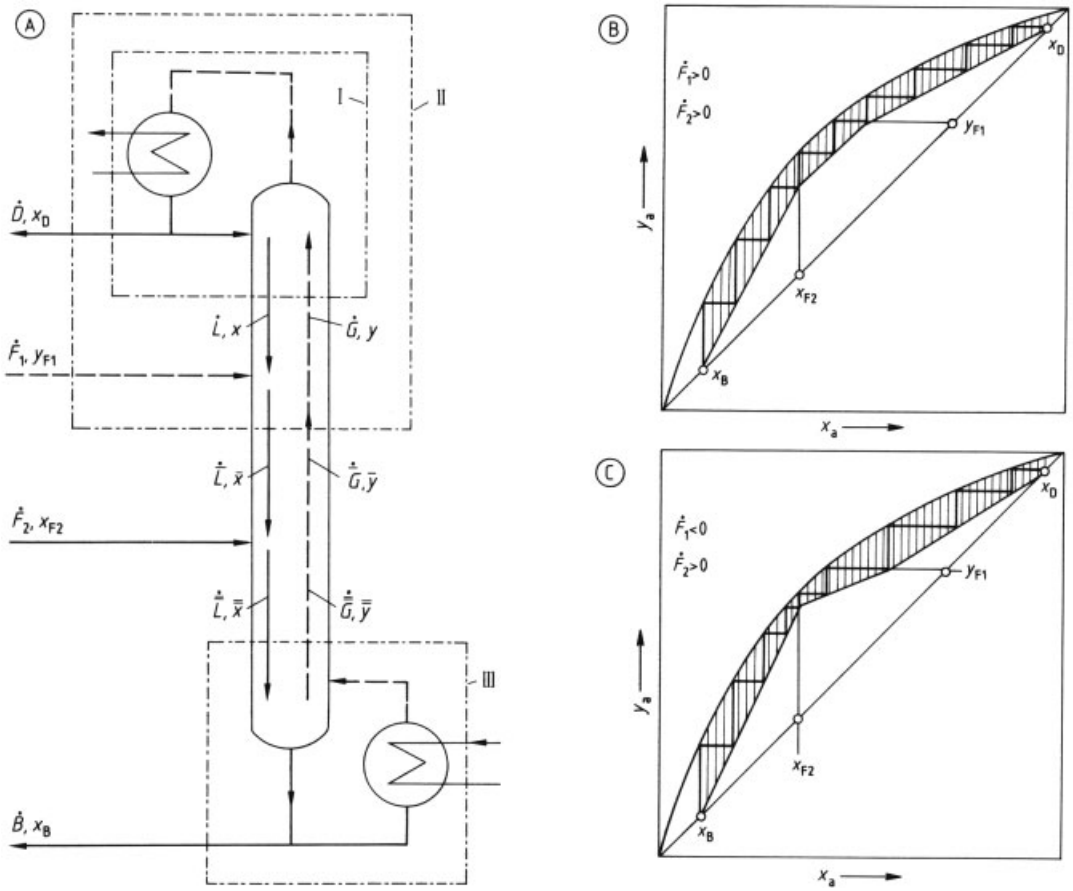

Figure 20. Rectification column and McCabe – Thiele diagram for multiple feed entries or side-stream withdrawal
A) Flow diagram; B) McCabe – Thiele diagram for two feeds; C) McCabe – Thiele diagram for one feed and one side stream

the stripping operating line, on the $y-x$ diagram (Fig. 19). Feed is introduced into the column at the point of intersection of the rectifying and stripping operating lines defined by Equations (4.17) and (4.19) (see Fig. 19). Material balance IV around the feed point gives the following equation for the point of intersection [29]:

$$y=\frac{\varphi}{\varphi-1}\cdot x-\frac{1}{\varphi-1}\cdot z_{\mathrm{F}} \tag{4.20}$$

The caloric factor φ characterizes the thermal condition of the feed and is defined by

$$\varphi=\frac{\text{Enthalpy required to vaporize 1 mol of feed}}{\text{Molar latent heat of vaporization of feed}}$$

If the feed is introduced as a boiling liquid, the value of φ is 1 and Equation (4.20) gives a vertical line on the $y-x$ diagram; i.e., the operating lines intersect at $x=z_{\mathrm{F}}$. However, if the feed is introduced as a saturated vapor, $\varphi=0$ and Equation (4.20) yields a horizontal line; the two operating lines intersect at $y=z_{\mathrm{F}}$. If the feed is introduced into the column as a vapor – liquid mixture, the point of intersection of the two operating lines will be between the two extreme cases described above, as shown in Figure 19. The φ line can also represent a subcooled liquid or a superheated vapor feed.

According to Equation (4.20), the φ line or feed line defines the locus of all points of intersection of the two operating lines in the rectifying and stripping sections of the column. Thus, if the feed conditions and the composition of overhead and bottom products are specified, and one operating line is established, the location of the second operating line is also known. This means

that a relationship exists between the external reflux ratio R and the external reboil ratio $\bar{R}$. If the feed enters at the boiling point, i.e., $\varphi = 1$,

$$\bar{R} = \frac{\dot{D}/\dot{F}}{1-\dot{D}/\dot{F}}(R+1) \tag{4.21}$$

Analogously, for a saturated-vapor feed,

$$\bar{R} = \frac{\dot{D}/\dot{F}}{1-\dot{D}/\dot{F}}(R-1) \tag{4.22}$$

The term $\dot{D}/\dot{F}$ is the relative quantity of overhead product, which can be determined by using Equation (4.13) after the product concentrations have been established.

The calculation of a rectification column with multiple feed entry and side stream withdrawal is conducted in an analogous manner (Fig. 20). The operating line bends at each entry or withdrawal point. Figure 20 B shows the course of the operating lines when $\dot{F}_1$ is a vapor feed stream ($\dot{F}_1 > 0$); Figure 20 C shows the corresponding lines for a vapor side stream ($\dot{F}_1 < 0$).

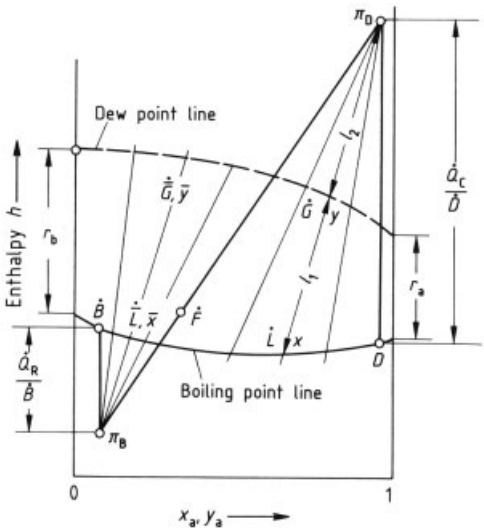

Figure 21. Enthalpy – concentration diagram of binary mixture a – b
The points of intersection of the lines through poles π_D and π_B with the dew point line and the boiling point line give the concentrations x and y on the operating line.

If the positions of the operating lines and the equilibrium curve on the $y-x$ diagram are known, the difficulty of separation can be determined by using the number of equilibrium stages n (shown as steps in Fig. 19, cf. Section 4.1.1) or the number of transfer units N_{OG} (shaded area in Fig. 19, cf. Section 4.1.2). The desired separation is possible only when the operating lines neither touch nor intersect the equilibrium curve. The smaller the space between the equilibrium curve and the balance lines, the more difficult is the separation.

4.2.2. Prediction of Rectification Based on Material and Enthalpy Balances

The calculation of rectification based on material balance alone is possible only if the vapor and liquid streams ($\dot{G}$ and $\dot{L}$, respectively) in the individual column sections are constant. If this is not the case, the enthalpy balance must also be taken into consideration. The enthalpy – concentration ($h-x$) diagram, shown in Figure 21 for a binary mixture a – b, is very convenient in this respect.

Overall enthalpy balance I (see Fig. 18) for the column gives the following relationship:

$$\dot{F}\cdot h_F + \dot{Q}_R = \dot{D}\cdot h_D + \dot{B}\cdot h_B + \dot{Q}_C \tag{4.23}$$

where h_F, h_D, and h_B are the specific enthalpies of the feed, distillate, and bottom product, respectively; $\dot{Q}_R$ is the molar heat flow supplied by the reboiler; and $\dot{Q}_C$ is the molar heat flow removed by the condenser. Thus,

$$\dot{F}\cdot h_F = \dot{D}\cdot \pi_D + \dot{B}\cdot \pi_B \tag{4.24}$$

where $\pi_D = h_D + \frac{\dot{Q}_C}{\dot{D}}$ and $\pi_B = h_B - \frac{\dot{Q}_R}{\dot{B}}$

Equation (4.24) represents a straight line in Figure 21 that extends from point π_D to point π_B and passes through $\dot{F}$. The amounts of heat $\dot{Q}_R$ and $\dot{Q}_C$ can be determined easily as shown in the diagram when the compositions of the feed and of the desired overhead or bottom products are known (i.e., $\dot{F}$, $\dot{D}$, and $\dot{B}$). Both the distillate ($\dot{D}$) and the bottom product ($\dot{B}$) are assumed to leave the column in the boiling state.

Heat balance II in Figure 18 at the top of the column (rectifying section) gives

$$\dot{G}\cdot h_G = \dot{L}\cdot h_L + \dot{D}\cdot \pi_D \tag{4.25}$$

Here, h_G and h_L are the specific enthalpies of the vapor and liquid, respectively. This equation also represents a straight line through the point π_D on the enthalpy – concentration diagram. If

the vapor in the column is always at the dew point and the liquid always at the boiling point (an assumption that is met closely), the compositions of the vapor $\dot{G}$ and liquid $\dot{L}$ streams as well as their concentrations y and x can be established as shown in Figure 21. Concentrations y and x are points on the operating line on the $y-x$ diagram (Fig. 19). Analogously, if the balance lines in the stripping section are allowed to pass through π_B, the vapor and liquid states are obtained from the points of intersection with the dew point and the boiling point lines, respectively, as shown in Figure 21. The operating lines in the rectifying and stripping sections are obtained by transferring the vapor and liquid concentrations to the $y-x$ diagram. The operating lines are generally slightly curved. The quantities of vapor and liquid can be determined by using the lever arm rule:

$$\dot{G}\cdot l_2=\dot{L}\cdot(l_1+l_2) \text{ or } \frac{\dot{L}}{\dot{G}}=\frac{l_2}{l_1+l_2} \tag{4.26}$$

where l_1 is the distance between the boiling point and dew point lines and l_2 is the distance between the pole π_D and the dew point line.

In the special case in which the dew point line runs parallel to the boiling point line on the enthalpy – concentration diagram, Equation (4.26) produces a constant value for $\dot{L}/\dot{G}$. Thus, the criteria for constant vapor and liquid loads can be formulated:

1) The enthalpies of vaporization of the two substances a and b must be equal.
2) No heat of mixing should occur.
3) The boiling points of the substances should be as close as possible.

The first criterion is the most important. Equal enthalpies of vaporization can often be achieved by using molar enthalpies of vaporization, which in turn require that the composition be given in mole fractions. Thus, rectification calculations generally employ these units.

If the criteria for a constant liquid – vapor ratio $\dot{L}/\dot{G}$ are not satisfied, the following approach is recommended:

1) The poles π_D and π_B are etablished on the enthalpy – concentration diagram.
2) The vapor and liquid concentrations y and x are determined, in each case, by drawing straight lines that radiate from these poles. The values thus obtained are plotted on a McCabe – Thiele diagram and the curved operating line is obtained, point by point.
3) The number of equilibrium steps n or the number of transfer units N_{OG} is determined in the usual manner on the $y-x$ diagram.

The number of equilibrium steps can also be determined on the $h-x$ diagram [30], but this method is so complex that it has little practical use.

4.2.3. Rectification of Binary Mixtures at Infinite Reflux Ratio

If the internal streams $\dot{G}$ and $\dot{L}$ are much larger than those of the feed ($\dot{F}$) and the products ($\dot{D}$ and $\dot{B}$), the internal reflux ratio $\dot{L}/\dot{G}$ in all parts of the column approaches 1. Thus, according to Equations (4.15) and (4.17),

$$R\rightarrow\infty \tag{4.27}$$

In this special case (i.e., at infinite reflux ratio), the energy input needed for separation is very large, but a minimum number of equilibrium stages n or transfer units N_{OG} is required. As shown in Figure 22, the operating line coincides with the diagonal and is represented by the simple equation

$$y=x \tag{4.28}$$

The equilibrium state $y*$ is determined by starting with the liquid state x_B. The new liquid concentration x_1 is now calculated from the vapor concentration according to Equation (4.28). The vapor state y_1* in equilibrium with this liquid concentration x_1 is, in turn, determined and so on. This procedure is identical with the determination of the distillation curve described in Section 2.2.3. Equation (2.25) thus applies to ideal binary mixtures with constant relative volatility α:

$$x_D=\frac{\alpha^n\cdot x_B}{1+(\alpha^n-1)\cdot x_B} \tag{4.29}$$

After rearrangement, the minimum number of equilibrium steps n_{min} is obtained:

$$n_{min}=\frac{1}{\ln\alpha}\cdot\ln\left(\frac{x_D}{x_B}\cdot\frac{1-x_B}{1-x_D}\right) \tag{4.30}$$

Analogously, the minimum number of transfer units $N_{\mathrm{OG,\,min}}$ is given by

$$N_{\mathrm{OG,min}} = \frac{1}{\alpha - 1} \cdot \ln \left[\frac{x_{\mathrm{D}}}{x_{\mathrm{B}}} \cdot \left(\frac{1 - x_{\mathrm{B}}}{1 - x_{\mathrm{D}}} \right)^{\alpha} \right] \tag{4.31}$$

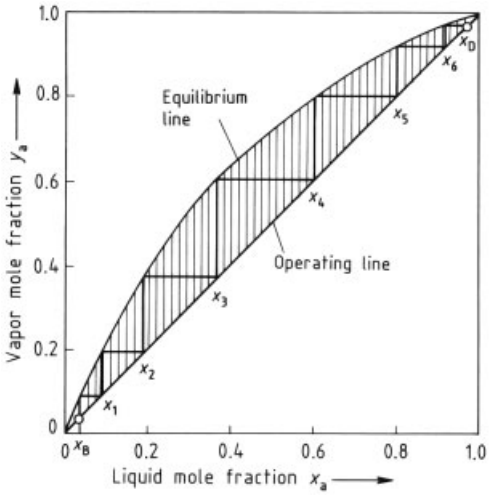

Figure 22. McCabe – Thiele diagram of a binary mixture at infinite reflux ratio

4.2.4. Rectification of Binary Mixtures at Minimum Reflux Ratio

Operation with minimum column loads (i.e., minimum reflux ratio) has considerable practical importance because it results in separation of a mixture by using the smallest possible column diameter and the lowest energy input.

Operation with minimal internal loads is characterized by intersection of the rectifying and stripping operating lines at a single point on the equilibrium curve. This point is called a *pinch point,* and an infinite number of equilibrium stages n or transfer units N_{OG} is required to reach this state. Figure 23 shows the $y-x$ diagram for a binary system with minimal internal loads. If the vapor and liquid flow rates ($\dot{G}$ and $\dot{L}$, respectively) in the individual sections of the column are constant, the operating lines are straight lines. The minimum slope of the rectifying operating line for the more volatile component a can be calculated as follows:

$$\left(\frac{\dot{L}}{\dot{G}} \right)_{\min} = \frac{x_{\mathrm{D}} - y_{\mathrm{F}}^{*}}{x_{\mathrm{D}} - x_{\mathrm{F}}^{*}} \tag{4.32}$$

By use of the relationship $\dot{L}/\dot{G} = R/(R+1)$, the minimum reflux ratio becomes

$$R_{\min} = \frac{x_{\mathrm{D}} - y_{\mathrm{F}}^{*}}{y_{\mathrm{F}}^{*} - x_{\mathrm{F}}^{*}} \tag{4.33}$$

Concentrations $y*_{\mathrm{F}}$ and $x*_{\mathrm{F}}$ are related by the phase equilibrium. By using Equation (2.17), the following relationship is obtained for ideal mixtures:

$$R_{\min} = \frac{1}{\alpha - 1} \left(\frac{x_{\mathrm{D}}}{x_{\mathrm{F}}^{*}} - \alpha \frac{1 - x_{\mathrm{D}}}{1 - x_{\mathrm{F}}^{*}} \right) \tag{4.34}$$

If the overhead product is required as pure substance a (i.e., $x_{\mathrm{D}} = 1$), Equation (4.34) can be simplified:

$$R_{\min} = \frac{1}{(\alpha - 1) \cdot x_{\mathrm{F}}^{*}} \tag{4.35}$$

The point of intersection of the feed line (Eq. 4.20) with the equilibrium line (Eq. 2.17) gives the concentration $x*_{\mathrm{F}}$. A quadratic equation is obtained which can be solved as follows:

$$x_{\mathrm{F}}^{*} = -\frac{B}{2} + \sqrt{\frac{B^2}{4} + C} \tag{4.36}$$

where

$$B = \frac{1}{\varphi} \left(\frac{\alpha}{\alpha - 1} - \varphi - z_{\mathrm{F}} \right)$$

$$C = \frac{z_{\mathrm{F}}}{\varphi \cdot (\alpha - 1)}$$

z_{F} is the mole fraction of substance a in a liquid – vapor feed and φ is the thermal state of the feed (caloric factor) defined in section 4.2.1.

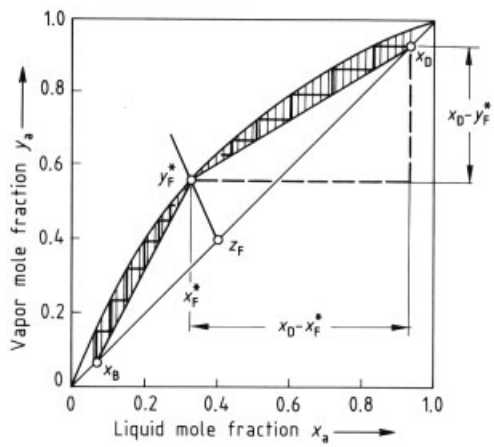

Figure 23. McCabe – Thiele diagram of a binary mixture at minimum reflux ratio and minimum reboil ratio
The operating lines intersect the equilibrium line at the pinch point.

If the feed is a boiling liquid with composition x_F, then

$$x_F^* = x_F \tag{4.37}$$

If the feed is a saturated vapor with composition y_F, then

$$x_F^* = \frac{y_F/\alpha}{1-(\alpha-1)\,y_F/\alpha} \tag{4.38}$$

Analogously, laws can be derived for the minimum reboil ratio $\bar{R} = \dot{\bar{G}}/\dot{B}$. The slope of the stripping operating line can be calculated by using Figure 23 as follows:

$$\left(\frac{\dot{L}}{\dot{\bar{G}}}\right)_{max} = \frac{y_F^* - x_B}{x_F^* - x_B} \tag{4.39}$$

From the relationship $\dot{\bar{L}}/\dot{\bar{G}} = \left(\bar{R}+1\right)/\bar{R}$, the minimum reboil ratio is

$$\bar{R}_{min} = \frac{x_F^* - x_B}{y_F^* - x_F^*} \tag{4.40}$$

The relationship $y*_F = f(x*_F)$ can be formulated easily for ideal mixtures:

$$\bar{R}_{min} = \left[\frac{1}{1-x_B/x_F^*}\cdot\left(\frac{\alpha}{1+(\alpha-1)\cdot x_F^*}-1\right)\right]^{-1} \tag{4.41}$$

An important case exists when pure substance b is to be separated as the bottom product (i.e., $x_{B_a} = 0$):

$$\bar{R}_{min} = \left[\frac{\alpha}{1+(\alpha-1)\cdot x_F^*}-1\right]^{-1} \tag{4.42}$$

The concentration $x*_F$ can generally be calculated from the composition z_F and the thermal state φ of the feed (Eq. 4.36). Again, the special cases in which the feed is a boiling liquid or a saturated vapor are very important. The relationship between the minimum reflux ratio and the minimum reboil ratio can then be expressed by using Equations (4.21) and (4.22). Separation using a rectification column with a minimum reflux ratio requires an infinite number of equilibrium steps n or transfer units N_{OG}. However, even a slight increase in the reflux ratio of ca. 5 – 10 % can result in economical operation. Thus, the special case and laws discussed in this section closely approximate actual rectification and have immediate practical importance.

4.2.5. Heat Requirements for Rectification of Binary Mixtures

In rectification columns, a quantity of heat $\dot{Q}_R$ is supplied in the reboiler at the bottom of the column and a quantity $\dot{Q}_C$ is removed by the condenser at the top. The following equations hold:

$$\dot{Q}_R = r\cdot\dot{G} \; or \; \dot{Q}_C = -\,r\cdot\dot{G} \tag{4.43}$$

where r is the molar latent heat of vaporization for the substance in question. Approximately equal molar heats of vaporization of the substances involved are assumed.

Important relationships can be established for the quantities of heat $\dot{Q}_R$ and $\dot{Q}_C$. The heat required by the reboiler $\dot{Q}_R$ is formulated by using the reboil ratio $\bar{R} = \dot{\bar{G}}/\dot{B}$. The heat removed by the condenser $\dot{Q}_C$ is determined by using the reflux ratio $R = \dot{L}/\dot{D}$. Thus,

$$\dot{Q}_R = r\cdot\bar{R}\cdot\dot{B} \text{ or } \dot{Q}_C = -r\cdot(R+1)\cdot\dot{D} \tag{4.44}$$

The term $(R+1)$ arises because both the reflux liquid and the overhead product must be condensed at the top of the column.

The product flow rates $\dot{B}$ and $\dot{D}$ in binary mixtures can be expressed in terms of the feed stream $\dot{F}$ and the composition of the products by using Equation (4.13).

$$\frac{\dot{Q}_R}{\dot{F}\cdot r} = \bar{R}\cdot\frac{x_D - z_F}{x_D - x_B}$$

and

$$\frac{\dot{Q}_C}{\dot{F}\cdot r} = -\,(R+1)\cdot\frac{z_F - x_B}{x_D - x_B} \tag{4.45}$$

Thus, $\dot{Q}_R$ and $\dot{Q}_C$ are directly proportional to the reboil ratio $\bar{R}$ and the reflux ratio R, respectively. These ratios cannot fall below a limiting value (for details, see Section 4.2.4); the minimal quantities of heat are calculated as follows:

$$\frac{\dot{Q}_{R,min}}{\dot{F}\cdot r} = \bar{R}_{min}\cdot\frac{x_D - z_F}{x_D - x_B}$$

$$\frac{\dot{Q}_{C,min}}{\dot{F}\cdot r} = -\,(R_{min}+1)\cdot\frac{z_F - x_B}{x_D - x_B} \tag{4.46}$$

If fractionation into a pure product a is required during rectification of a binary mixture (i.e.,

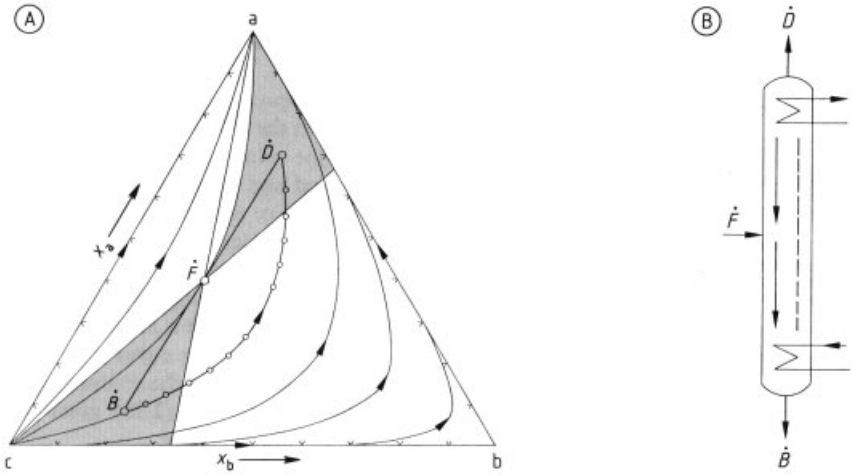

Figure 24. Separation of an ideal ternary mixture a – b – c in a single rectification column
A) Triangular phase diagram showing calculated distillation curvesEquilibrium stages are shown as open circles on one of the curves. Compositions of the overhead and bottom products $\dot{D}$ and $\dot{B}$ lie on a straight line that passes through the feed $\dot{F}$ and must be located within the shaded areas
B) Schematic of rectification column

$x_{Da} = 1$ and $x_{Ba} = 0$), use of the appropriate relationships for the reboil and reflux ratios gives

$$\frac{\dot{Q}_{R,min}}{\dot{F}\cdot r} = \frac{1}{\alpha - 1}\left(1+\alpha\frac{x_F^*}{1-x_F^*}\right)\cdot(1-z_F)$$

$$\frac{\dot{Q}_{C,min}}{\dot{F}\cdot r} = -\,z_F\left[\frac{1}{(\alpha-1)\cdot x_F^*}+1\right] \tag{4.47}$$

The concentration x_F* can be determined from the concentration of the feed by using Equation (4.36). In the special case of a boiling-liquid feed having composition x_F, the following equations are obtained:

$$\frac{\dot{Q}_{R,min}}{\dot{F}\cdot r} = \frac{1}{\alpha - 1} + x_F \text{ and}$$

$$\frac{\dot{Q}_{C,min}}{\dot{F}\cdot r} = -\left(\frac{1}{\alpha-1}+x_F\right) \tag{4.48}$$

The heat requirement $\dot{Q}_{R,min}$ consists of two terms. The first term accounts for the heat required to maintain the countercurrent in the column and is independent of feed composition. The second term, however, is directly proportional to the composition of the feed and represents the heat required for vaporization of the overhead product; this quantity of heat would also be required for simple distillation.

For a saturated-vapor feed of composition y_F, Equation (4.47) becomes

$$\frac{\dot{Q}_{R,min}}{\dot{F}\cdot r} = \frac{1}{\alpha - 1} \text{ and } \frac{\dot{Q}_{C,min}}{\dot{F}\cdot r} = \frac{1}{\alpha - 1} + 1 \tag{4.49}$$

Both enthalpies are independent of the composition of the feed. If the distillate is withdrawn in the liquid state, more heat must of course be removed from the top of the column than supplied at the bottom.

In general, either the quantity of heat $\dot{Q}_R$ to be added or the quantity of heat $\dot{Q}_C$ to be removed is determined by using the above equations. The remaining quantity of heat is established by using an enthalpy balance around the column. This enthalpy balance can be carried out especially well on the enthalpy – concentration diagram in Figure 21.

If the feed is introduced in the boiling state and the distillate and bottom products are removed in the boiling state, then the following equation holds:

$$\dot{Q}_R \approx -\dot{Q}_C \tag{4.50}$$

This relationship is fully valid only when the boiling point line on the enthalpy – concentration diagram, on which points $\dot{F}$, $\dot{B}$, and $\dot{D}$ are

located, is a straight line. The latent heats of vaporization of components a and b, r_a and r_b, do not have to be equal in this case.

4.3. Rectification of Ternary Mixtures

The binary mixtures discussed in Section 4.2 represent a special case in rectification. Generally, multicomponent mixtures are separated. Rectification of a ternary mixture of substances a, b, and c can be represented relatively clearly on a triangular diagram. Relationships derived from ternary mixtures can be applied directly to multicomponent mixtures.

4.3.1. Rectification of Ternary Mixtures at Infinite Reflux Ratio

A very simple approach to ternary rectification is possible when the internal mass vapor and liquid streams in the column, $\dot{G}$ and $\dot{L}$, are much larger than the feed and product streams $\dot{F}$, $\dot{D}$, and $\dot{B}$ (i.e., $R \to \infty$). For the special case $R \to \infty$, the operating line coincides with the diagonal on the $y-x$ diagram, which means that the vapor concentration y of a given component is always equal to the liquid concentration x. As described in Section 4.2.3, the distillation curve represents the changes in concentration in the column. Thus for ideal mixtures, Equation (2.26) gives the course of concentration changes:

$$x_{na} = \frac{\alpha_{ac}^{n} \cdot x_{Ba}}{1 + (\alpha_{ac}^{n} - 1) \cdot x_{Ba} + (\alpha_{bc}^{n} - 1) \cdot x_{Bb}} \tag{4.51}$$

$$x_{nb} = \frac{\alpha_{bc}^{n} \cdot x_{Bb}}{1 + (\alpha_{ac}^{n} - 1) \cdot x_{Ba} + (\alpha_{bc}^{n} - 1) \cdot x_{Bb}}$$

$$x_{nc} = 1 - x_{an} - x_{bn}$$

where the subscripts a, b, and c denote the three components of the mixture; n denotes the number of equilibrium stages; and B, the bottom product. Starting with a defined liquid composition (e.g., the bottom state $\dot{B}$), the changes in concentration in the column can be calculated as a function of the number of equilibrium stages n. These are represented by the distillation curves shown in Figure 24 A. The arrows indicate the direction in which the boiling temperatures decrease, i.e., toward the top of the column.

A material balance around the column (Fig. 24 B) gives

$$\dot{F} = \dot{D} + \dot{B} \tag{4.52}$$

$$\dot{F} \cdot z_{Fa} = \dot{D} \cdot x_{Da} + \dot{B} \cdot x_{Ba}$$

$$\dot{F} \cdot z_{Fb} = \dot{D} \cdot x_{Db} + \dot{B} \cdot x_{Bb}$$

These equations produce a straight line through the feed $\dot{F}$ on the triangular diagram, which also passes through the points representing the distillate $\dot{D}$ and the bottom product $\dot{B}$.

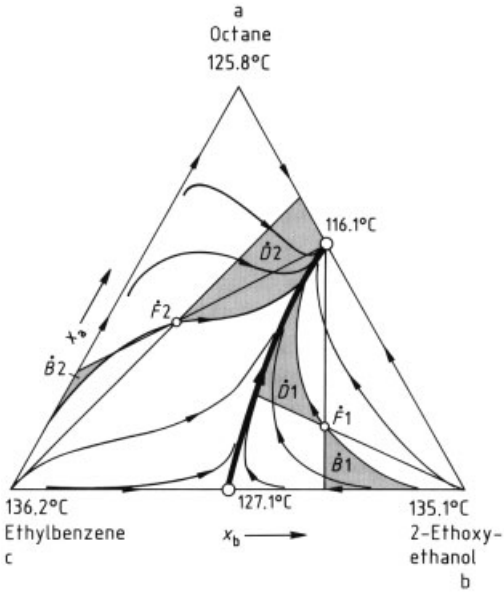

Figure 25. Separation of a ternary mixture with two minimum azeotropes (p = 101.3 kPa)
Azeotropic points are represented by open circles. The bold line is a distillation boundary which cannot be crossed and divides the phase diagram into two sectors. Shading denotes the separation areas for two different feed compositions F1 and F2.

Both Equations (4.51) and (4.52) must be valid at the ends of the column. Thus, the states of products $\dot{D}$ and $\dot{B}$ must lie both on the balance line through $\dot{F}$ and on a distillation curve. As shown in Figure 24 A, the balance line forms a chord of the distillation curve. Equations (4.51) and (4.52) obviously permit an infinite number of different overhead $\dot{D}$ and bottom $\dot{B}$ products. Limits for $\dot{D}$ and $\dot{B}$ are reached when either pure substance a or pure substance c is obtained at the top or bottom of the column, respectively. All other solutions of Equations (4.51) and (4.52) lie within these extremes; thus, the overhead and bottom products ($\dot{D}$ and $\dot{B}$, respectively) can be

located only in the shaded areas shown on Figure 24 A. This is generally applicable irrespective of the number of equilibrium stages and the mode of operation, i.e., also at a finite reflux ratio R. Thus, a ternary mixture can be separated in a single rectification column to only a very limited extent. Medium-boiling substance b can be recovered in the pure state as a side-stream product only in the special case when $R \to \infty$ and $n \to \infty$; generally it can merely be made more concentrated. Only substances at which distillation curves either start or end can be separated efficiently (see Chap. 6).

Distillation curves for a highly nonideal mixture which forms two binary minimum azeotropes are shown in Figure 25. Shaded regions represent possible separation areas for two different feeds $\dot{F}1$ and $\dot{F}2$. Two sectors exist in which separation into $\dot{D}$ and $\dot{B}$ products occurs. The distillation boundary separating these two sectors connects the two minimum azeotropes and cannot be crossed during rectification. Thus the distillation boundaries of ternary mixtures are as important as the azeotropic points of binary mixtures.

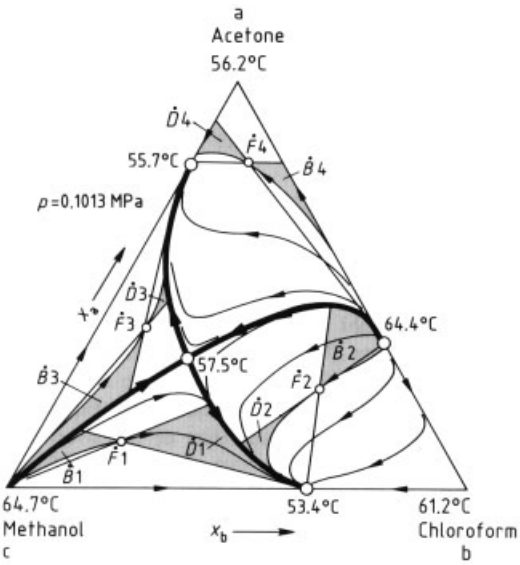

Figure 26. Separation areas of a ternary mixture with three binary azeotropes and one ternary azeotrope

An extreme example of a ternary mixture is shown in Figure 26. The mixture forms three binary azeotropes and one saddle-point ternary azeotrope. Two distillation boundaries exist: one connects the two minimum azeotropes, and the other runs from the high boiler c to the maximum binary azeotrope; they intersect at the saddle point. Hence, the three-component mixture is divided into four distillation fields. Because separation boundaries cannot be crossed during rectification, fractionation is possible only within one field. As illustrated in Figure 26, within a given distillation area, only limited separation is possible.

A knowledge of the paths of distillation curves and the possible positions of distillation boundaries of ternary mixtures is valuable for process design. Details about expected rectification fractions can then be given without extensive column calculation (see Chap. 6). Exact knowledge of the phase equilibrium is not absolutely necessary. The course of the distillation curves and of the separation boundaries can be estimated solely from the boiling temperatures of the marked points, i.e., of the pure substances and the azeotropes.

4.3.2. Rectification of Ternary Mixtures at Minimum Internal Loads

Industrial rectification columns are preferentially operated at a low reflux ratio R. Hence, the special case of the minimum reflux ratio $R_{\min}$ is of considerable interest. A method described in [31] permits determination of the corresponding minimum reflux ratio for any overhead product at a given feed composition. However, the equations can only be solved iteratively. This section shows that the minimum reflux ratio can be calculated directly, i.e., without iteration, in three special cases. Many separations are possible but this method yields an especially favorable, "preferred" separation.

4.3.2.1. Minimum Reflux Ratio and Minimum Reboil Ratio for Preferred Separation

The relationships presented in Section 4.2.4 for the minimum reflux ratio of binary mixtures can be applied directly to ternary mixtures. Hence, as in Equation (4.33), the following is valid for each component i:

$$R_{\min} = \frac{x_{\mathrm{D}i} - y^*_{\mathrm{F}i}}{y^*_{\mathrm{F}i} - x^*_{\mathrm{F}i}} \tag{4.53}$$

This equation represents a straight line through $\dot{F}$, as shown in Figure 27. The minimum reflux ratio for all compositions of the distillate $\dot{D}$ on this line can be determined by using Equation (4.53).

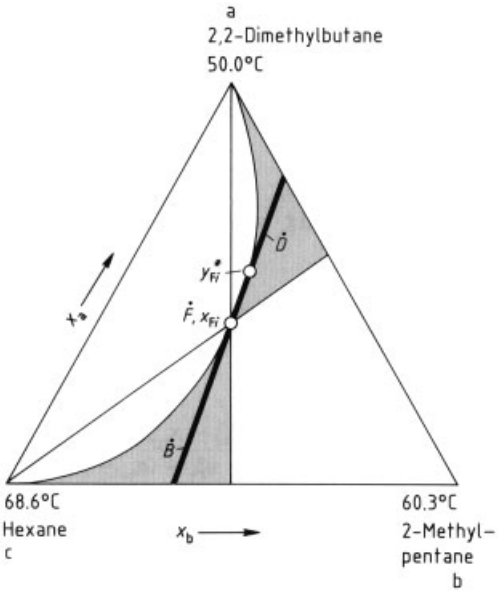

Figure 27. Preferred separation of a ternary mixture at minimum reflux ratio (p = 101.3 kPa)
Both separation products $\dot{D}$ and $\dot{B}$ can lie only on the line through x_{Fi} and $y*_{Fi}$ (feed at boiling point). Thus, only one "preferred" separation exists in the shaded separation area.

If complete removal of the high-boiling substance c from the overhead product is required, then $x_{Dc} = 0$. Thus,

$$R_{min} = -\frac{y^*_{Fc}}{y^*_{Fc} - x^*_{Fc}} \tag{4.54}$$

Concentrations of the other two substances in the distillate can be calculated from Equation (4.53):

$$x_{Da} = (R_{min}+1) \cdot y^*_{Fa} - R_{min} \cdot x^*_{Fa}$$
$$x_{Db} = (R_{min}+1) \cdot y^*Fb - R_{min} \cdot x^*_{Fb} \tag{4.55}$$

These relationships are generally applicable; i.e., they also apply to nonideal mixtures as long as the corresponding straight lines do not intersect the equilibrium curve above the feed point.

For ideal mixtures, the relationship between $y*Fi$ and $x*Fi$ can be formulated in general terms and inserted into Equation (4.54) to give

$$R_{min} = \frac{1}{(\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}} \tag{4.56}$$

The composition of the head product $\dot{D}$ can then be expressed as:

$$x_{Da} = \frac{(\alpha_{ac}-1) \cdot x^*_{Fa}}{(\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}} \tag{4.57}$$
$$x_{Db} = \frac{(\alpha_{ac}-1) \cdot x^*_{Fb}}{(\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}}$$
$$x_{Dc} = 0$$

An analogous equation can be obtained for the minimum reboil ratio:

$$\bar{R}_{min} = \left[\frac{\alpha_{ac}}{1 + (\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}} - 1\right]^{-1} \tag{4.58}$$

The corresponding composition of the bottom product is given by

$$x_{Ba} = 0 \tag{4.59}$$
$$x_{Bb} = \frac{x^*_{Fb} \cdot (\alpha_{ac} - \alpha_{bc})}{\alpha_{ac} - \left[1 + (\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}\right]}$$
$$x_{Bc} = \frac{x^*_{Fc} \cdot (\alpha_{ac}-1)}{\alpha_{ac} - \left[1 + (\alpha_{ac}-1) \cdot x^*_{Fa} + (\alpha_{bc}-1) \cdot x^*_{Fb}\right]}$$

Concentrations $x*_{Fi}$ can be determined from the point of intersection of the feed with the equilibrium relationship, as in Section 4.2.4.

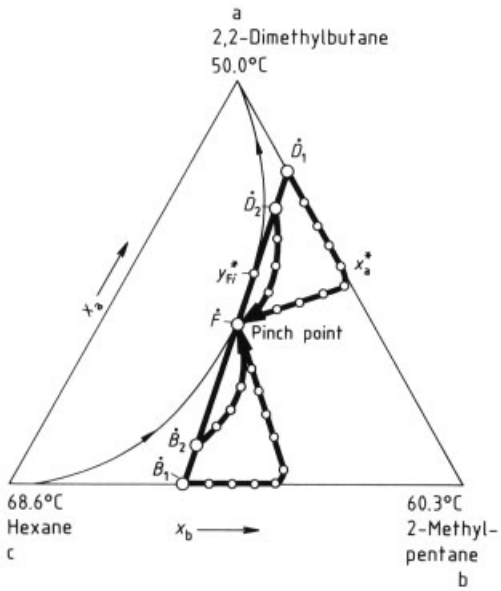

Figure 28. Course of the concentration changes in a column for the "preferred" separation of a ternary mixture at minimum reflux ratio (p = 101.3 kPa)
Two different separations are shown. A pinch point is attained in both column sections at the feed composition for the "preferred" separation. This separation is totally analogous to binary rectification at minimum reflux ratio.
—— Balance around the column; –○– Changes in liquid concentration in the column

In the especially important case of a boiling-liquid feed with composition $x_{\mathrm{F}i}$, the following holds:

$$x^*_{\mathrm{F}i} = x_{\mathrm{F}i} \tag{4.60}$$

Introduction of a saturated-vapor feed with composition $y_{\mathrm{F}i}$ represents another important case. Here,

$$x^*_{\mathrm{F}i} = \frac{y_{\mathrm{F}i}/\alpha_{ik}}{1 - \sum\left(\alpha_{jk} - 1\right) \cdot y_{\mathrm{F}j}/\alpha_{jk}} \tag{4.61}$$

The material balance around the column also gives a straight line through the feed point $\dot{F}$. Both products $\dot{D}$ and $\dot{B}$ must be located on this line. Consequently, the bottom state $\dot{B}$ also lies on the straight line established by Equation (4.53). The relative quantities of the overhead product $\dot{D}/\dot{F}$ and of the bottom product $\dot{B}/\dot{F}$ can be calculated by using the lever arm rule.

The changes in concentration within the column obtained at $R_{\min}$ are shown in Figure 28. Case 1 represents a very sharp separation into two fractions $\dot{D}_1$ and $\dot{B}_1$ in which the concentrations of the high boiler in the distillate x_{Dc} and the low boiler in the bottoms x_{Ba} are zero (Eq. 4.54). Case 2 represents a moderate separation of the same feed, also at minimum (still lower) reflux ratio. The concentration changes in both the stripping and the rectification sections lead to, and end at, the feed composition. Thus, the point of intersection of the operating lines with the equilibrium curve (pinch point) coincides with the feed. Equation (4.53) (i.e., the straight line in Fig. 27) represents a certain natural, preferred course of separation, which is achieved with a minimum energy input in an infinitely long rectification column. Here, the reflux ratio $R_{\min}$ has the lowest value. Of the large number of separations possible, separation of the feed into an overhead product $\dot{D}$ and a bottom product $\dot{B}$ at this value represents a very important case. This separation usually lies in the middle of the fractionation region, as shown in Figure 27. This preferred position for the overhead and bottom fractions of a rectification column is extremely important. As described in Chapter 7, it is usually the most favorable separation path for multicomponent rectification.

4.3.2.2. Minimum Reflux Ratio for Separation of the Low Boiler

Often, the most volatile substance of a ternary mixture must be separated from the other components in as pure a state as possible. The minimum reflux ratio can also be calculated directly. In the separation of a "c-free" overhead product, only a binary mixture of substances a and b is present in the upper part of the column (Fig. 28). Enrichment in the high-boiling substance c occurs only when the liquid equilibrium concentration $x*_{\mathrm{a}}$ is reached. The $y-x$ diagram (Fig. 29) plotted in terms of the low boiler (a) clearly indicates the importance of the concentration $x*_{\mathrm{a}}$. The operating line intersects the equilibrium curve, represented by the relative volatility α_{ab}, at the point $x*_{\mathrm{a}}$. Thus, enrichment in the high-boiling component c occurs only after the binary mixture in the column reaches the pinch point. Equation (4.34) can be modified for this state as follows:

$$R_{\min} = \frac{1}{(\alpha_{\mathrm{ab}} - 1)}\left[\frac{x_{\mathrm{Da}}}{x^*_{\mathrm{a}}} - \alpha_{\mathrm{ab}}\frac{1 - x_{\mathrm{Da}}}{1 - x^*_{\mathrm{a}}}\right] \tag{4.62}$$

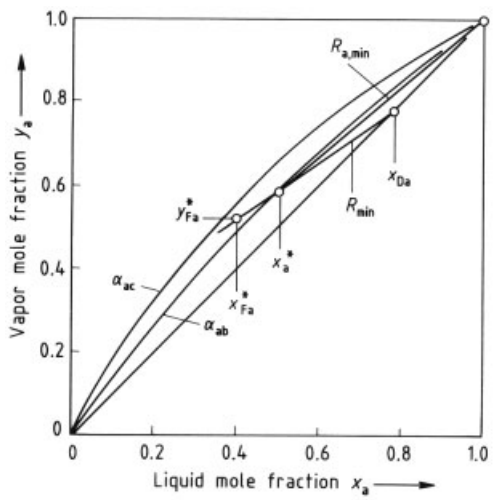

Figure 29. Derivation of $R_{\mathrm{a,\,min}}$, the minimum reflux ratio for the separation of pure overhead product a from the ternary mixture a – b – c
α_{ab} = relative volatility of a and b; $R_{\min}$ = minimum reflux ratio required for the preferred separation (Eq. 4.34).

If the desired overhead product is pure substance a, then a larger reflux ratio $R_{\mathrm{a,\,min}}$ is required, as illustrated in Figure 29. The new operating line also intersects the binary equilibrium curve at $x*_{\mathrm{a}}$. Since $x_{\mathrm{Da}} = 1$, Equation (4.62) becomes

$$R_{\mathrm{a,min}} = \frac{1}{(\alpha_{\mathrm{ab}} - 1)\cdot x_{\mathrm{a}}^{*}} \tag{4.63}$$

This equation can be solved for $x*_{\mathrm{a}}$ and inserted into Equation (4.62). A quadratic equation is obtained for $R_{\mathrm{a,\,min}}$ which is solved in the following way:

$$R_{\mathrm{a,min}} = -\frac{B}{2} + \sqrt{\frac{B^2}{4} - C}$$

$$\text{where } B = 1 - \frac{1}{x_{\mathrm{Da}}}\left(R_{\mathrm{min}} + \frac{\alpha_{\mathrm{ab}}}{\alpha_{\mathrm{ab}} - 1}\right) \tag{4.64}$$

$$C = \frac{R_{\mathrm{min}}}{(\alpha_{\mathrm{ab}} - 1)\cdot x_{\mathrm{Da}}}$$

The values of R_{min} and x_{Da} are determined by using Equations (4.56) and (4.57). The minimum reflux ratio $R_{\mathrm{a,\,min}}$ required for separation of pure substance a as the overhead product is always larger than the reflux ratio R_{min} for the preferred separation described in Section 4.3.2.1.

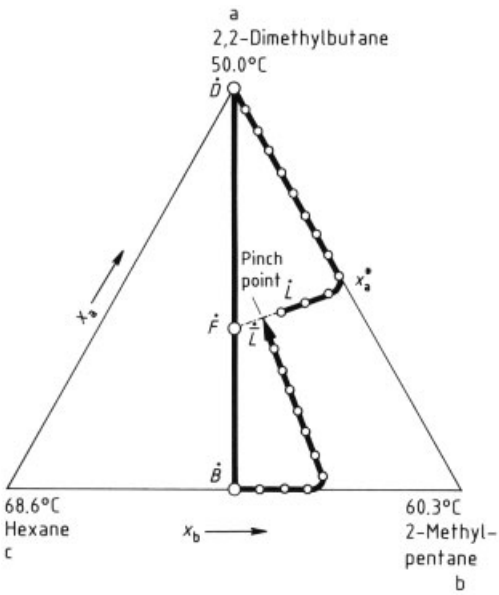

Figure 30. Separation of the low boiler (a) from ternary mixture a – b – c as an overhead product at minimum reflux ratio (p = 101.3 kPa)
The pinch point is located directly below the feed point.
—— Balance around the column; –o– Changes in liquid concentration in the column; – · – · – Mixing line at the feed point

The course of concentration changes in the column during separation of the low boiler a are shown in Figure 30. Again, a binary mixture of substances a and b is present in the rectifying part of the column. Enrichment in the high boiler c occurs only at the binary pinch point $x*_{\mathrm{a}}$. The concentration changes in the rectifying section lead to the feed. However, a ternary pinch point is not attained here. The pinch point of the ternary mixture lies in the stripping section of the column at $\dot{\bar{L}}$, i.e., directly below the feed point.

4.3.2.3. Minimum Reboil Ratio for Separation of the High Boiler

An approach similar to that described in Section 4.3.2.2 can be used in the separation of pure bottom product from a ternary mixture. The minimum reboil ratio $\bar{R}$ can be determined according to the definition $\bar{R} = \dot{\bar{G}}/\dot{B}$. The minimum reflux ratio R can then be calculated by using Equation (4.21). The following is analogous to Equation (4.64), where $\bar{R}_{\mathrm{c,\,min}}$ is the minimum reboil ratio required for separation of the high boiler c as a pure bottom product.

$$\bar{R}_{\mathrm{c,min}} = \left(-\frac{B}{2} - \sqrt{\frac{B^2}{4} - C}\right)^{-1} \tag{4.65}$$

where

$$B = -\frac{1}{\bar{R}_{\mathrm{min}}}\left[1 + x_{\mathrm{Bb}}\,(\alpha_{\mathrm{bc}} - 1)\right] - (\alpha_{\mathrm{bc}} - 1)$$

$$C = \frac{(1 - x_{\mathrm{Bb}})\cdot(\alpha_{\mathrm{bc}} - 1)}{\bar{R}_{\mathrm{min}}}$$

In this case, $\bar{R}_{\mathrm{min}}$ is the minimum reboil ratio for the preferred separation according to Equation (4.58). The value x_{Bb} is determined by use of Equation (4.59).

Thus, $\bar{R}_{\mathrm{c,\,min}}$ for the separation of a pure bottom product consisting of the high boiler c can then be determined. The quantities $x*_{\mathrm{F}i}$ are calculated from Equations (4.60) and (4.61), depending on the feed condition. The minimum reboil ratio $\bar{R}_{\mathrm{c,\,min}}$ for separation of pure high boiler c is always larger than the minimum reboil ratio $\bar{R}_{\mathrm{min}}$ required for the preferred separation according to Equation (4.58).

Figure 31 shows the concentration profiles in the column during separation of the high boiler c. The concentrations in the stripping section lead to, but do not end at, the feed. The ternary pinch point, represented by an arrow, lies directly above the feed point.

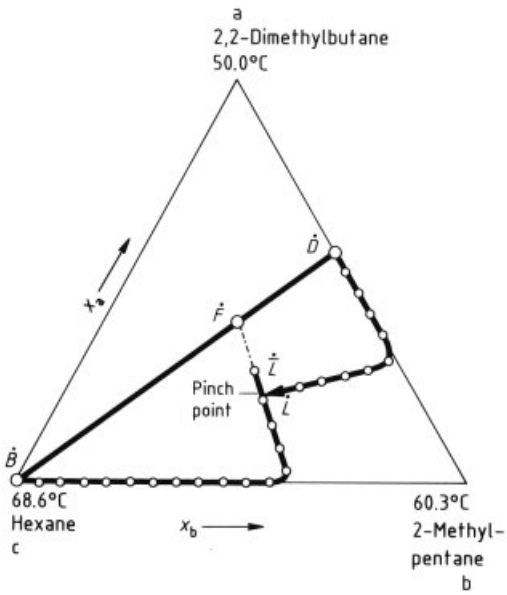

Figure 31. Separation of the high boiler (c) from ternary mixture a – b – c as bottom product (p = 101.3 kPa) at minimum reflux ratio
The pinch point is located directly above the feed point.
—— Balance around the column; –○– Changes in liquid concentration in the column; – · – · – Mixing line

4.3.3. Rectification of Ternary Mixtures at Finite Reflux Ratio

Rectification at a finite reflux ratio can be calculated by first specifying the column and operating conditions and then determining concentrations at the top and bottom of the column. If the required product specifications are not achieved, either the operating conditions or the column configuration must be altered and the calculation repeated.

The number of equilibrium steps n and the position of the feed entry point are required for column specification. The operating conditions are characterized sufficiently by the reflux ratio R and the relative amount of overhead product $\dot{D}/\dot{F}$. The calculation is best carried out in an analogous manner to the stepwise method used for binary mixtures described in Section 4.1.1 (McCabe – Thiele diagram). By starting with a known liquid composition (e.g., at the bottom), the equilibrium vapor-phase state is calculated. For ideal mixtures,

$$\begin{aligned} y_a^* &= \frac{\alpha_{ac} \cdot x_a}{1 + (\alpha_{ac} - 1) \cdot x_a + (\alpha_{bc} - 1) \cdot x_b} \\ y_b^* &= \frac{\alpha_{bc} \cdot x_b}{1 + (\alpha_{ac} - 1) \cdot x_a + (\alpha_{bc} - 1) \cdot x_b} \\ y_c^* &= 1 - y_a - y_b \end{aligned} \tag{4.66}$$

For nonideal mixtures, the equilibrium state can be determined as described in Chapter 2.

The next step involves elucidation of the new liquid state from the vapor state by using the operating line. Material balance around the bottom of the column gives the following equation for the operating line:

$$\begin{aligned} x_a &= \frac{\bar{R}}{\bar{R}+1} \cdot y_a^* + \frac{1}{\bar{R}+1} \cdot x_{Ba} \\ x_b &= \frac{\bar{R}}{\bar{R}+1} \cdot y_b^* + \frac{1}{\bar{R}+1} \cdot x_{Bb} \\ x_c &= 1 - x_a - x_b \end{aligned} \tag{4.67}$$

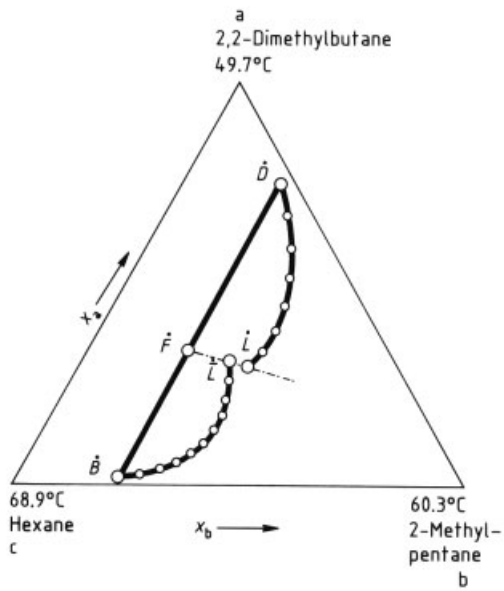

Figure 32. Separation of a ternary mixture at a finite reflux ratio (p = 101.3 kPa)
A concentration jump occurs at the feed point due to admixing of feed.
—— Balance around the column; –○– Changes in liquid concentration in the column; – · – · – Mixing line

The alternate use of Equations (4.66) and (4.67) produces a mathematical stepwise construction similar to the graphical construction for binary mixtures on the $y-x$ diagram (see chapter 4). In Figure 32, the concentration changes observed in the stripping section of the column are represented by dots.

When the feed point has been reached, admixing of the feed must be considered in the calculation process. The following relationship is obtained from a material balance for the liquid composition directly above the feed point by assuming a boiling-liquid feed and negligible temperature differences at the feed entry point:

$$x_{\mathrm{a}}=\frac{(\bar{R}+1)\cdot\dot{B}/\dot{F}}{(\bar{R}+1)\cdot\dot{B}/\dot{F}-1}\cdot(\bar{x}_{\mathrm{a}}-x_{\mathrm{Fa}})+x_{\mathrm{Fa}}$$
$$x_{\mathrm{b}}=\frac{(\bar{R}+1)\cdot\dot{B}/\dot{F}}{(\bar{R}+1)\cdot\dot{B}/\dot{F}-1}\cdot(\bar{x}_{\mathrm{b}}-x_{\mathrm{Fb}})+x_{\mathrm{Fb}}$$
$$x_{\mathrm{c}}=1-x_{\mathrm{a}}-x_{\mathrm{b}} \tag{4.68}$$

Equation (4.68) represents a straight line (mixing line) through $\dot{F}$ on the triangular diagram. The states of the liquid below ($\dot{\bar{L}}$) and above ($\dot{L}$) the feed point must lie on the mixing line, as indicated in Figure 32. The position of the operating line changes above the feed point because the quantity of liquid is altered. The following holds:

$$x_{\mathrm{a}}=\frac{\bar{R}\cdot y_{\mathrm{a}}^{*}+x_{\mathrm{Ba}}-x_{\mathrm{Fa}}\cdot\dot{F}/\dot{B}}{\bar{R}+1-\dot{F}/\dot{B}}$$
$$x_{\mathrm{b}}=\frac{\bar{R}\cdot y_{\mathrm{b}}^{*}+x_{\mathrm{Bb}}-x_{\mathrm{Fb}}\cdot\dot{F}/\dot{B}}{\bar{R}+1-\dot{F}/\dot{B}} \tag{4.69}$$
$$x_{\mathrm{c}}=1-x_{\mathrm{a}}-x_{\mathrm{b}}$$

The calculation is continued by alternating Equations (4.66) and (4.69) and working upward step-by-step to the top of the column. According to Equation (4.52), when starting with bottom state $\dot{B}$, the overhead stream $\dot{D}$, determined by stepwise calculation, must fall on a straight line through $\dot{F}$ and $\dot{B}$ (balance line around the column).

The concentrations of components a, b, and c calculated in this way are plotted as a function of the number of equilibrium stages n, as shown in Figure 33. The medium-boiling substance b can pass through a concentration maximum in the column. Furthermore, a concentration jump occurs at the feed point within the column.

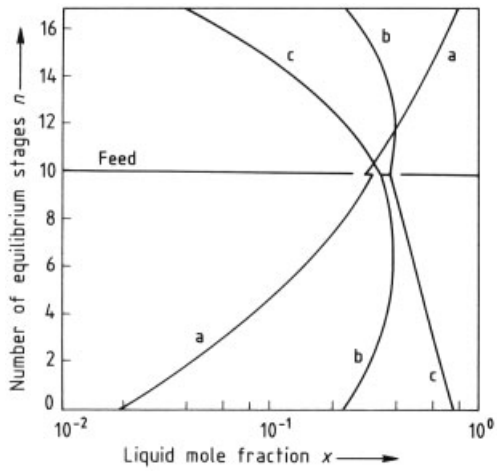

Figure 33. Rectification of a ternary mixture showing the liquid concentration of individual components a, b, and c as a function of the number of equilibrium stages n
a) 2,2-Dimethylbutane; b) 2-Methylpentane; c) Hexane

4.3.4. Problems Arising in Column Calculations

Calculation of a ternary rectification, discussed in Section 4.3.3, is basically very simple and can be done with a personal computer. However, a few problems often complicate the calculation.

Choice of Reflux Ratio. The values used in practice for the reflux ratio R and the reboil ratio $\bar{R}$ are ca. 10 – 30 % higher than the minimum values. The relationships for calculating $R_{\min}$ and $\bar{R}_{\min}$ in the most important cases in Section 4.3.2 apply to ideal mixtures and to weakly nonideal mixtures. However, mixtures that deviate more strongly from nonideal behavior can produce a pinch point above or below the feed point. The reflux ratio can then be determined only by iterative column calculation.

Choice of Bottom State. The bottom composition $x_{\mathrm{B}i}$ is the iteration variable in the calculation process described in Section 4.3.3. A poorly estimated value of $x_{\mathrm{B}i}$ often gives unreasonable concentrations in the first run, e.g., concentrations >1 or <0. A useful approach is limitation of the possible bottom state as described in Section 4.3.1.

Even if the choice of bottom composition is satisfactory, unreasonable concentrations can still be obtained in the rectifying section of the column. This can be avoided by starting with estimated values of the distillate composition $x_{\mathrm{D}i}$ for calculation of the rectifying section. The calculation begins at the top and bottom of the column, and continues step-by-step until the feed point is reached. Estimated values are checked with the mixing calculation at the feed point. For real mixtures, this procedure is effective; however, iterations are required at each step of the calculation in the rectifying section because equilibrium relationships cannot be solved for liquid concentrations.

Improvement of Estimated Values. The material balance around the column is generally not fulfilled after the first calculation. A new bottom composition must be chosen which complies better with the balance around the column. No physically based criterion is available to improve the estimated values, which is the most severe handicap associated with column calculations. Indeed, serious convergence problems often occur at this point, especially in the case of highly nonideal systems.

Choice of Feed Point. For binary mixtures, feed is introduced at that point along the column at which the liquid concentration in the column is identical with the composition of the feed. This is not possible in ternary rectification because the compositions in the column usually differ from that of the feed. As a rule, the concentration of only one substance can be made equal to that of the feed. Other substances undergo irreversibe changes of state during mixing which energetically impair separation. A series of criteria are described in [32] for optimal choice of feed point; most aim at decreasing the irreversibilities.

Consideration of Enthalpy Balance. The calculation process described in Section 4.3.3 is based on the assumption of constant vapor and liquid loads in each section of the column. If these criteria (presented in Section 4.2.2) are not adequately fulfilled, the enthalpy balance must be considered in addition to the material balance. This complicates the calculation.

4.4. Rectification of Multicomponent Mixtures

The methods and equations used in ternary rectification and described in Section 4.3 can be applied directly to multicomponent mixtures. Changes in liquid concentration in the column at very high reflux ratio, $R \to \infty$, are represented by distillation curves. For ideal mixtures,

$$x_{\mathrm{D}i} = \frac{\alpha_{ik}^{n} \cdot x_{\mathrm{B}i}}{1 + \sum_{j=1}^{k-1} \left(\alpha_{jk}^{n} - 1\right) \cdot x_{\mathrm{B}j}} \tag{4.70}$$

The material balance around the column gives

$$\dot{F} \cdot x_{\mathrm{F}i} = \dot{D} \cdot x_{\mathrm{D}i} + \dot{B} \cdot x_{\mathrm{B}i} \tag{4.71}$$

These two sets of equations do not fully describe the separation. Further data are necessary, such as the relative amount of overhead product $\dot{D}/\dot{F}$ and the concentration of a component at one end of the column, i.e., the value $x_{\mathrm{D}i}$ or $x_{\mathrm{B}i}$. One extreme value for composition of the distillate is the requirement $x_{\mathrm{D}a} = 1$ with $x_{\mathrm{D}i \neq a} = 0$; the other extreme is the requirement $x_{\mathrm{B}k} = 1$ with $x_{\mathrm{B}i \neq k} = 0$, i.e., a component is separated as a pure overhead or bottom product. The area within which fractionation products $\dot{D}$ and $\dot{B}$ can be located is limited by these two extremes. However, unlike three-component mixtures (Figs. 24, 25, 26), clear representation of the separation area of multicomponent mixtures is not possible.

Relationships pertaining to operation at minimum reflux ratio $R_{\min}$ can also be applied directly to multicomponent mixtures. The requirement that the overhead product should not contain the highest boiling component k leads to, e.g.,

$$R_{\min} = \frac{1}{\sum_{j=1}^{k-1} \left(\alpha_{jk} - 1\right) \cdot x_{\mathrm{F}j}^{*}} \tag{4.72}$$

A single separation is selected from the entire separation region (see Fig. 27). The distillate composition obtained for this separation is

$$x_{\mathrm{D}i} = \frac{\left(\alpha_{ik} - 1\right) \cdot x_{\mathrm{F}i}^{*}}{\sum_{j=1}^{k-1} \left(\alpha_{jk} - 1\right) \cdot x_{\mathrm{F}j}^{*}} \tag{4.73}$$

After the relative amount of overhead product $\dot{D}/\dot{F}$ has been established, the composition of the bottom product $\dot{B}$ can be determined from the material balance around the column. Even the relationships determined for separation of the low boiler and the high boiler can be applied directly to multicomponent mixtures.

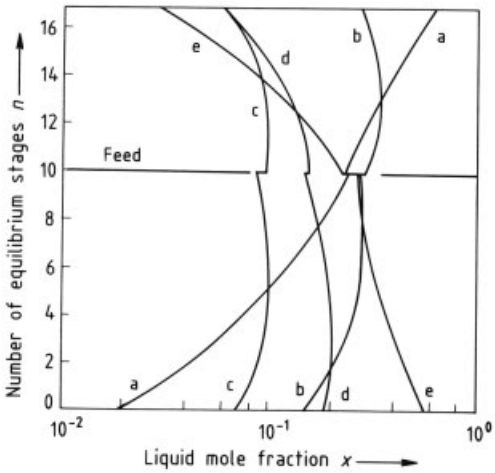

Figure 34. Rectification of a five-component mixture showing changes in liquid concentration as a function of the number of equilibrium stages
a) Butane; b) Pentane; c) Hexane; d) Heptane; e) Octane

In principle, the step-by-step calculation described in Section 4.3.3 for a ternary mixture can also be applied to multicomponent systems. The result of a calculation for a five-component system is illustrated in Figure 34. However, the more components present in the mixture, the more serious is the convergence problem. The chosen starting values of concentration at one end of the column must be close to the correct solution. Specification of the concentration of the more volatile components in the bottom product or of the less volatile components in the overhead product is especially difficult; estimated values may differ from the correct solution by several orders of magnitude. Hence, step-by-step calculation, which uses the concentration at one end of the column as the iteration variable, is unsuitable.

Several more effective methods for column calculation are available [33–37]. These produce accurate results without the use of simplifying assumptions, such as ideal mixtures or constant $\dot{L}/\dot{G}$ ratios [33]. They are based on MESH equations (i.e., *M*aterial balance, *E*quilibrium relationship, *S*ummation of all substances, and enthalpy balance *H*). The MESH equations are formulated for each equilibrium step in such a way that a tridiagonal matrix is obtained. The nonlinear set of equations is solved iteratively. The iteration variable is not the composition of the overhead or the bottom product, but usually the temperature profile or the loads in the column. The theta convergence method [34] and the Newton – Raphson method [35] have been employed successfully to improve estimated values.

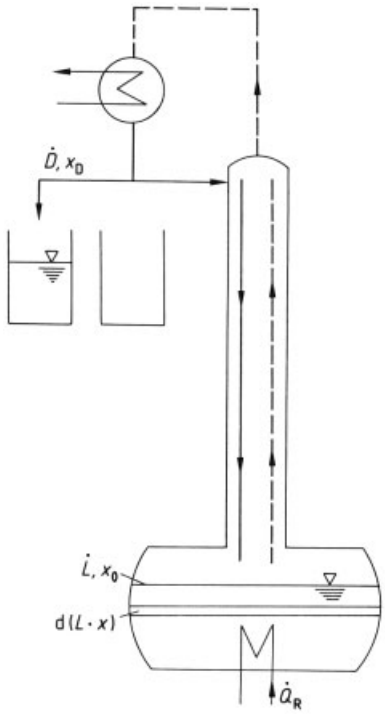

Figure 35. Schematic of batch rectification

These calculation methods are highly developed. They have been applied to systems with two liquid phases [36] and to systems with superimposed chemical reactions [37] in the liquid.

Exact multicomponent column calculation, without simplifying assumptions, is mathematically very complex. Programs of this type are

commercially available and it is not generally advisable to develop them oneself.

5. Batch Rectification

Batch rectification of liquid mixtures is of great industrial importance; it is used extensively in the laboratory and in small production units. The main advantage of batch rectification over continuous operation is that a single apparatus can process many different liquid mixtures. The mixture at hand and different product requirements can be taken into account by simply changing the reflux ratio. This technique can also be extended to a multicomponent mixture; components are separated by collecting the overhead products in separate receivers.

Figure 35 represents a setup for batch rectification. The boiler is filled with the liquid mixture (L) and heated. Vapor flows up the column and is condensed at the top. Initially, the entire condensate is usually returned to the column as reflux (i.e., $R \rightarrow \infty$). This countercurrent considerably improves separation (cf. Section 4.1). After some time, a portion of the overhead condensate is withdrawn continuously as distillate. The liquid in the boiler is increasingly depleted of the more volatile components, whereas the vapor at the top becomes enriched in these substances. The volume of liquid in the boiler simultaneously decreases, resulting in a highly increased concentration of less volatile substances in the boiler.

5.1. Batch Rectification of Binary Mixtures

The calculation of batch rectification in binary mixtures is based on the Rayleigh equation, derived in Section 3.2.1 for batch distillation. Application of this equation to the rectification of L moles of liquid requires the use of x_{D}, the mole fraction of the more volatile component in the mass stream, leaving the balance (i.e., the distillate) instead of $y*$ in Equation (3.14):

$$\frac{\mathrm{d}L}{L} = \frac{\mathrm{d}x}{x_{\mathrm{D}} - x} \tag{5.1}$$

The concentration of distillate x_{D} depends on the phase equilibrium, the separation efficiency of the column (number of equilibrium steps n or number of transfer units N_{OG}), and the reflux ratio R:

$$x_{\mathrm{D}} = f\,(x, \alpha, n, R)$$

For binary mixtures, the course of the function f is usually determined graphically on the $y-x$ diagram. Calculations depend on the operating mode.

5.1.1. Operation at Constant Reflux Ratio

For operation at constant reflux ratio, the separation efficiency of the column (e.g., number of equilibrium steps n), the reflux ratio R, and the starting concentration x_0 are specified. The amount of product D, the mean composition x_{Dm} of the product, and the energy requirement for rectification are determined.

Integration of Equation (5.1) gives

$$\ln \frac{L}{L_0} = \int_{x_0}^{x} \frac{\mathrm{d}x}{x_{\mathrm{D}} - x} \tag{5.2}$$

where L_0 is the initial amount of liquid in moles. Integration of the right-hand side of this equation requires determination of the correlation $x_{\mathrm{D}} = f\,(x, \alpha, n, R)$. In a binary system, this is usually done graphically on the $y-x$ diagram, as shown in Figure 36. A composition x_{D1} is assumed, and an operating line is drawn, followed by the usual determination of the concentration x_1 in the boiler, e.g., by stepping off. Subsequently, another distillate concentration x_{D2} is assumed, and the corresponding x_2 is determined in an analogous manner and so on. The operating lines are all parallel because R is assumed to be constant. In this way, the correlation between distillate concentration x_{D} and boiler concentration x can be determined point by point. The integral in Equation (5.2) is solved by using an auxiliary diagram in which the term $1/(x_{\mathrm{D}} - x)$ is plotted on the ordinate and x on the abscissa. The area A under the curve, determined point by point, from x_0 to x represents the solution to the integral. The following equation is obtained:

$$L = L_0 \cdot e^{A} \tag{5.3}$$

When $D = L_0 - L$,

$$D = L_0 \cdot \left(1 - e^A\right) \tag{5.4}$$

The mean distillate concentration x_{Dm} is calculated from the mass balance:

$$x_{\mathrm{Dm}} = \frac{L_0 \cdot x_0 - L \cdot x}{L_0 \cdot (1 - e^A)} \tag{5.5}$$

The energy (heat) requirement for the separation, Q_{R}, is calculated from

$$Q_{\mathrm{R}} = L_0 \cdot \left(1 - e^A\right) \cdot (R+1) \cdot r \tag{5.6}$$

where r is the latent heat of vaporization.

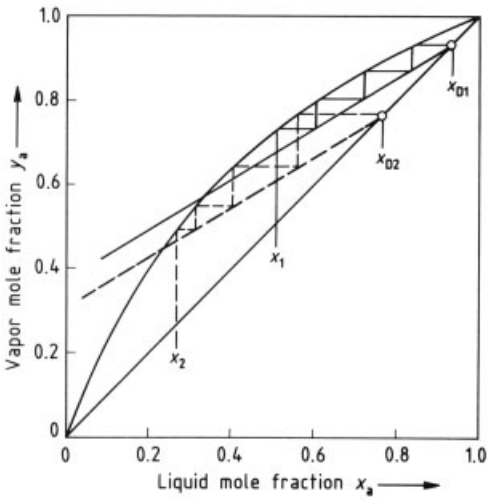

Figure 36. Correlation between the concentration x of liquid in the boiler and the concentration x_{D} of distillate at constant reflux ratio R (batch rectification of binary system a – b)

Operation at constant reflux ratio R is relatively simple. However, the distillate concentration x_{D} decreases with time, as illustrated in Figure 36. Thus, the desired product composition is achieved by having a higher starting purity than required. The reflux ratio is often increased one or more times during separation; this prevents a steep drop in product purity.

5.1.2. Operation at Constant Distillate Composition

Deterioration of the distillate concentration x_{D} over time can be avoided by continuously increasing the reflux ratio R throughout the run. The correlation between boiler concentration x and corresponding reflux ratio R is, in turn, determined graphically on the $y - x$ diagram. A reflux ratio R, and thus the position of the operating line, are specified, and the corresponding boiler concentration x is determined by stepping off. This is illustrated in Figure 37 for two different cases. The correlation obtained depends on the separation efficiency of the column and is similar to that shown in Figure 38.

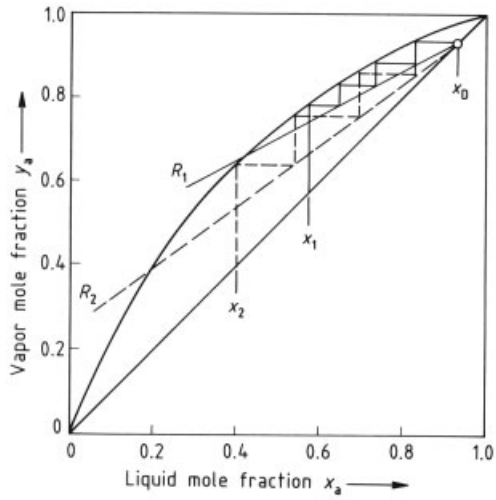

Figure 37. Correlation between reflux ratio R and concentration x of liquid in the boiler at constant distillate composition x_{D} (batch rectification of binary system a – b)
Two different cases are shown and indicated by subscripts 1 and 2.

Operation at constant distillate concentration x_{D} permits direct integration of the Rayleigh equation (Eq. 5.1) to give

$$L = L_0 \cdot \frac{x_{\mathrm{D}} - x_0}{x_{\mathrm{D}} - x} \tag{5.7}$$

The quantity of distillate is calculated from

$$D = L_0 \cdot \frac{x_0 - x}{x_{\mathrm{D}} - x} \tag{5.8}$$

The energy requirement depends on the reflux ratio R, which is variable. Hence, only a differential equation can be given:

$$dQ_R = dD \cdot (R+1) \cdot r \tag{5.9}$$

By using the correlation $dD = -dL$, the following relationship can finally be obtained from Equations (5.7) and (5.9):

$$dQ_R = L_0 \cdot r \cdot (x_D - x_0) \cdot \frac{R+1}{(x_D - x)^2} \cdot dx \tag{5.10}$$

Integration gives

$$Q_R = L_0 \cdot r \cdot (x_D - x_0) \cdot \int_{x_0}^{x} \frac{R+1}{(x_D - x)^2} \cdot dx \tag{5.11}$$

The term on the right-hand side of Equation (5.11) is integrated graphically by using an auxiliary diagram. The boiler concentration x is plotted on the abscissa vs. $(R+1)/(x_D - x)^2$ on the ordinate. The area A under the curve between x_0 and x gives the value of the integral. Hence, the heat requirement is

$$Q_R = L_0 \cdot r \cdot (x_D - x_0) \cdot A \tag{5.12}$$

In practice, operation at constant distillate concentration x_D is laborious because the reflux ratio R changes (Fig. 38) and must constantly be readjusted. On the other hand, this occurs automatically during the rectification of mixtures such as 1-butanol – water which show an immiscibility region and form a heterogeneous azeotrope (Fig. 39). In this case, part of the separation efficiency is furnished by the decanter.

5.1.3. Operation with Minimum Energy Requirement

The two operating modes described in Sections 5.1.1 and 5.1.2 require different quantities of heat Q to obtain the same amount and composition of product. Generally, less energy is required for the operation at constant distillate concentration. However, in the energetically optimal operational mode, the distillate concentration usually decreases slowly and the reflux ratio increases steadily. Pontryagin's maximum principle can be used to calculate the correlation between reflux ratio R and boiler concentration x [38]. Energy savings of up to ca. 20 % can be achieved, which justify this more laborious method only in special cases.

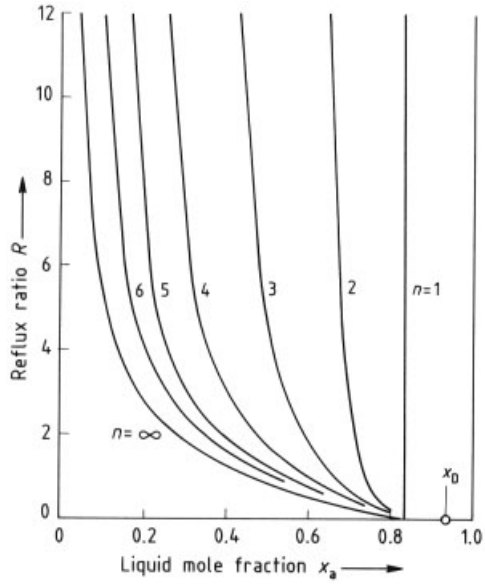

Figure 38. Correlation between boiler concentration x and reflux ratio R at constant distillate concentration x_D as a function of the number of equilibrium steps n

5.2. Batch Rectification of Multicomponent Mixtures

Multicomponent batch rectification is also represented by the Rayleigh equation. For each component i in the mixture,

$$\frac{dL}{L} = \frac{dx_i}{x_{Di} - x_i} \tag{5.13}$$

This set of equations can only be solved numerically.

Figure 40 shows the result of a calculation with a constant reflux ratio R for a quaternary system. The changes in distillate concentration x_{Di} clearly indicate that as rectification progresses, the individual components of the starting mixture are successively obtained as distillate in highly pure states. The pure substances can be collected in order of decreasing volatility by changing the receiver at the appropriate time. The boiler is gradually depleted of the low-boiling component (a) early in rectification; concentrations of the higher boiling substances b, c, and d increase simultaneously. After most of

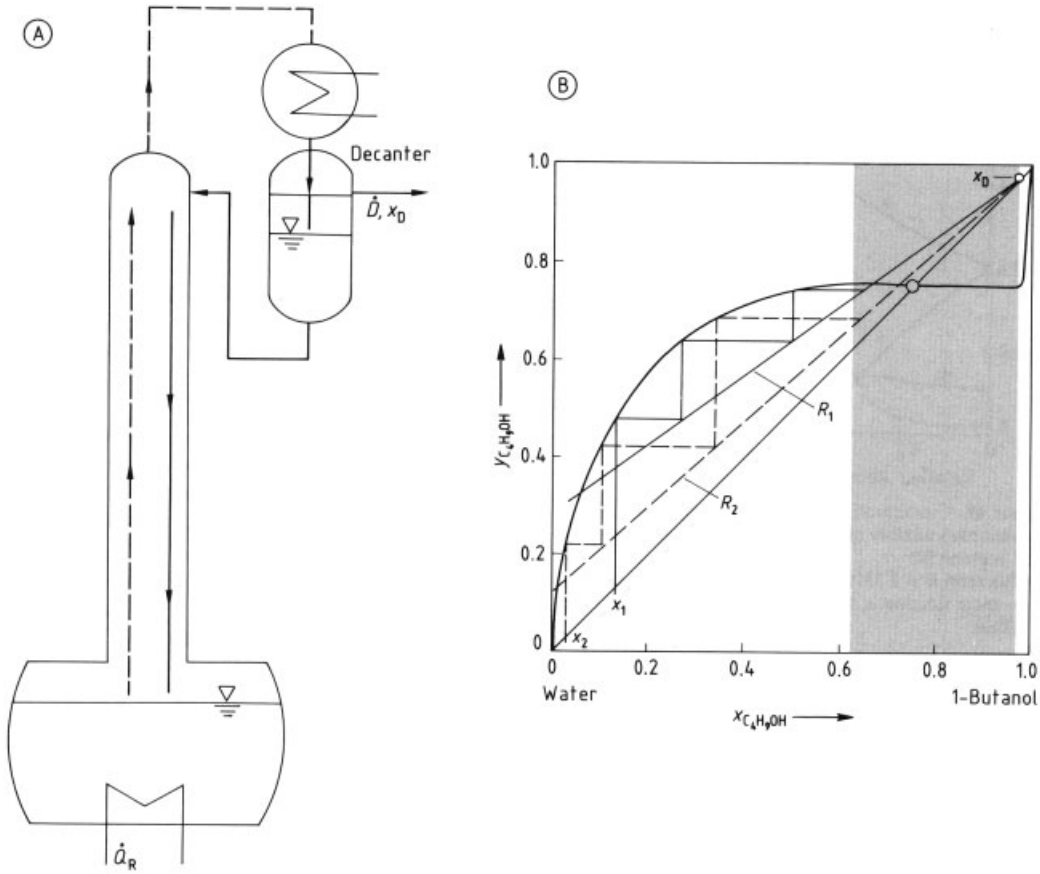

Figure 39. Batch rectification of a mixture of 1-butanol – water with heterogeneous azeotrope formation
A) Flow diagram; B) McCabe – Thiele diagram for two different reflux ratios R_1 and R_2 (the shaded area denotes the immiscibility region)

substance a has been separated, the boiler is depleted of substance b, followed by substance c. Finally, the highest boiling substance d is obtained in pure form in both the boiler and the distillate.

Batchwise distillation of the same mixture is shown in Figure 12. Comparison of Figures 12 and 40 shows clearly that rectification produces a sharper separation than distillation. Rectification can be improved considerably by using a larger number of equilibrium stages in combination with a higher reflux ratio.

When the internal loads are much larger than the distillate load (i.e., $R \to \infty$), the equation derived for batch distillation (Eq. 3.24) can be applied to batch rectification. The relative volatility α_{ik} is replaced by the term α_{ik}^{n}, where n is the number of equilibrium stages. Using the mole fraction constraint $\Sigma x_i = 1$ gives

$$\sum_{i=1}^{k} \left[x_{i0} \cdot (L/L_0)^{\alpha_{ik}^{n}-1} \; (x_k/x_{k0})^{\alpha_{ik}^{n}} \right] = 1 \qquad (5.14)$$

This relationship can be used to determine the enrichment of individual substances as a function of the number of equilibrium stages n at a very high reflux ratio $R \to \infty$.

The operational mode described in Section 5.1.3 with minimum energy requirements can, in principle, also be extended to multicomponent mixtures. The calculation is described in

[39]; in addition, a chemical reaction in the liquid is taken into account.

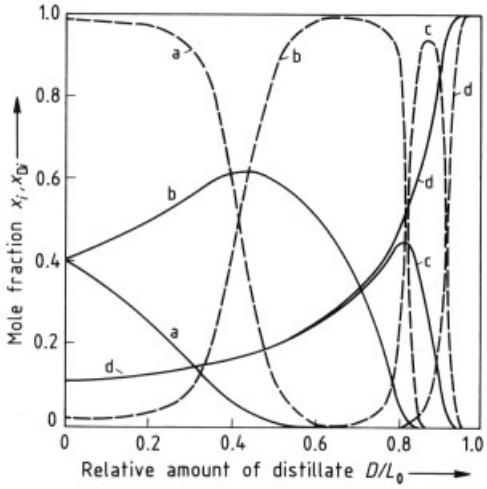

Figure 40. Concentration profile for batch rectification of a quaternary mixture of butane (a), pentane (b), hexane (c), and heptane (d)
Reflux ratio R = 2.33; number of equilibrium stages $n = 5$; —— mole fraction x_i in boiler; – – – mole fraction x_{Di} in distillate.

The calculation procedures described hitherto contain a number of simplifications. For instance, a straight operating line was assumed in each case, but this is valid only when the vapor and liquid loads in the column are constant (see Section 4.2.2). Furthermore, liquid holdup in the transfer elements in the column has not been considered. The real, constantly present liquid holdup in the column can greatly influence the purity of the individual substances at the top of the column [40].

6. Industrial Rectification Processes

Rectification is a very important separation process used widely in industry. Years of commercial use and adaptation to important separations have resulted in the development of more complex and efficient processes.

6.1. Binary Rectification

A very sophisticated binary rectification process has been developed for the recovery of oxygen and nitrogen from air, which is basically the separation of an almost ideal two-component mixture. The separation is conducted in two steps (Fig. 41 A). Pure nitrogen is recovered as overhead product from column C-1 and fed as reflux to a second separating column C-2. Actual separation into the products nitrogen and oxygen takes place in the second column C-2.

If the feed is introduced as a gas – liquid mixture and pure gaseous nitrogen is withdrawn from C-2, then only one condenser is required for column C-1 and one reboiler for column C-2.

If column C-2 is operated at a lower pressure than column C-1, the bottom temperature of C-2 is lower than the overhead temperature of C-1; i.e., the waste heat from column C-1 can be used to heat column C-2. In the Linde double column, C-2 is placed on top of C-1 in such a way that the condenser and reboiler are combined (Fig. 41 B) [41]. The column is operated without external heat input. Energy is supplied by the fall in pressure from 0.55 to 0.13 MPa, provided the feed contains as much vapor as is released in the form of gaseous product.

6.2. Multicomponent Separation

Complete separation of a multicomponent mixture into individual components requires several rectification steps.

As discussed in Chapter 4, a ternary mixture can be separated to only a limited extent into an overhead product stream $\dot{D}$ and a bottom product $\dot{B}$. Figure 24 shows the possible position of the two fractions on a triangular diagram. In special cases, pure overhead product a or pure bottom product c can be separated. The remaining binary fraction must subsequently be separated into the pure substances in a second column.

Figures 42 A and 42 B show the two separation paths a and c, respectively. In an alternative approach (Fig. 42 C, path a – c), the first column produces an overhead product without the high boiler c and a bottom product without the low boiler a. Two additional columns are then required to separate these binary fractions. The medium boiler b appears, in one case, as the bottom product and, in the second case, as the overhead product. As discussed in Chapter 7, this alternative is energetically the most favorable at high concentrations of medium boiler b.

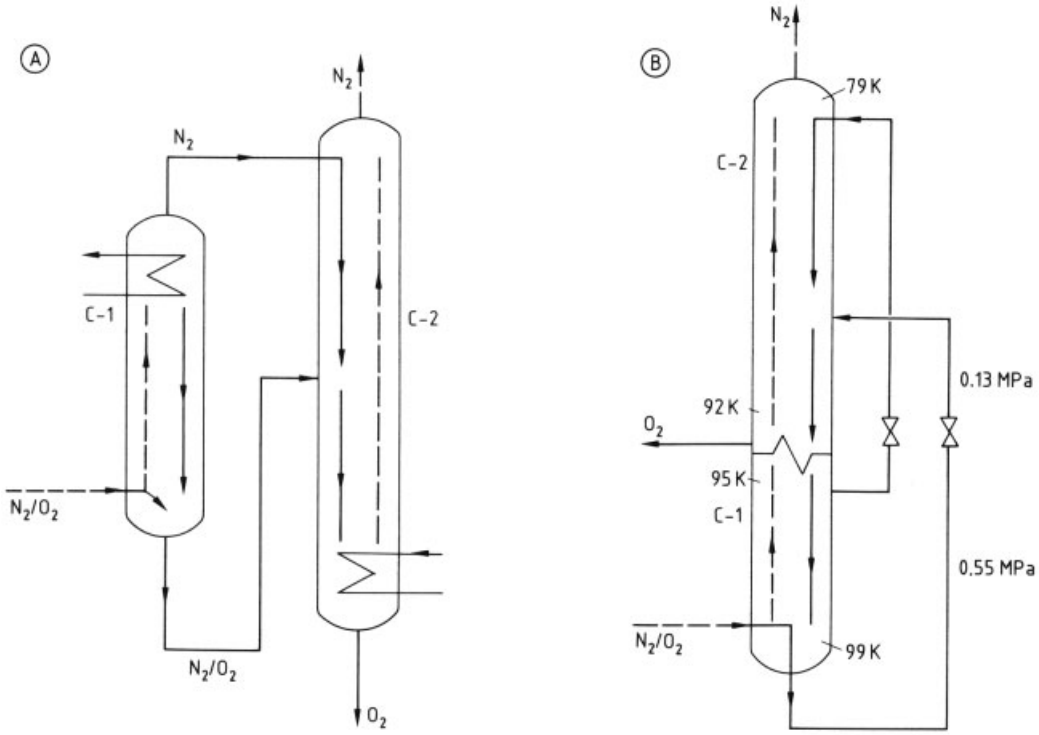

Figure 41. Recovery of oxygen and nitrogen from air by rectification
A) With two separate columns C-1 and C-2; B) With Linde double column

The separation of ternary mixtures often involves side columns. A side column does not represent another separating path, merely a different equipment design for path a or c. The lower section of the second column is combined with the first column, as depicted in Figure 43 for path a, thereby eliminating a heat exchanger. This type of design offers no energetic advantages. On the contrary, heat $\dot{Q}_{\mathrm{R1}}$ must be supplied at a higher temperature, which is an exergetic disadvantage. In addition, internal loads in the shaded part of the column are increased. This design is advantageous only when the amount of the medium-boiling substance b is relatively low.

For the complete separation of a four-component mixture, 21 different separation paths are possible. However, as discussed in Chapter 7, the energetically optimal separating sequence is usually the following:

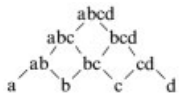

Analogously, 651 different separation paths are available for the separation of a five-component mixture! Here too, the energetically optimal path can easily be found by using the methods presented in Chapter 7.

In practice, the separation sequence of multicomponent mixtures is often determined by other limiting conditions, e.g., stability or corrosiveness of individual substances, danger of explosion, toxicity, tendency to stick to column internals, etc. In all these cases, the difficult substance should, if possible, be separated first. Thus, the energy requirement is only one of many aspects determining the choice of separating sequence.

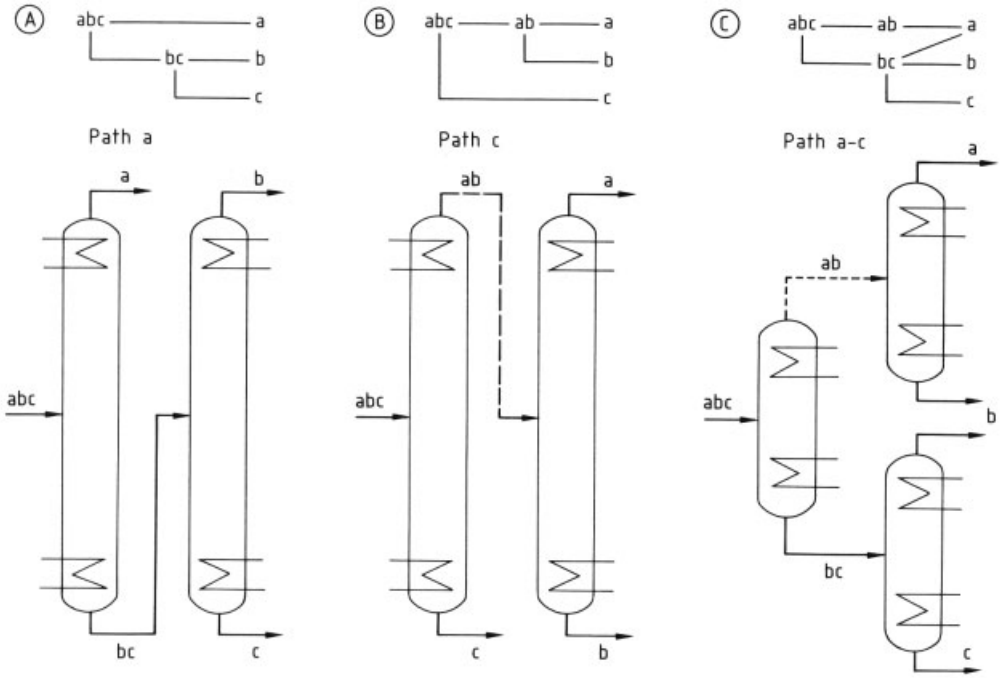

Figure 42. Paths for complete separation of a ternary mixture

6.3. Separation of Heterogeneous Azeotropic Mixtures

Azeotropic points constitute a barrier to separation by rectification. A heterogeneous azeotrope can be overcome by use of a decanter. For example, in Figure 44, separation of the binary nitromethane – water mixture requires two columns, C-1 and C-2, but both pure substances are removed as bottom products (Fig. 44 A). The vapor composition at the top of each column is approximately that of the azeotrope (Fig. 44 B). Hence, both vapor streams can be combined and condensed in the same condenser. The resulting two-phase mixture goes through a decanter S-1. The nitromethane-rich phase x_{F2} is fed into column C-2, and the aqueous phase is returned to column C-1. Because water is the bottom product in one of the columns (C-1), steam used for heating can be fed directly into the column and no reboiler is required.

In ternary mixtures, several distillation fields can occur which are separated from each other by distillation boundaries (see Sections 4.3 and 6.4). As in binary azeotropes, these distillation boundaries cannot be crossed during rectification. However, if phase splitting occurs, the distillation boundary can be crossed by separation in a decanter.

Separation of an acetone – water – butanol mixture illustrates this procedure (Fig. 45) [42]. The water – butanol mixture forms a heterogeneous azeotrope; the distillation boundary runs from the azeotrope to the low boiler acetone (Fig. 45 B). Separation is carried out in two columns (Fig. 45 A). Acetone is withdrawn overhead from the first column C-1. The bottom fraction $\dot{B}1$ lies in the shaded two-phase range and is separated into a water-rich phase $\dot{S}1$ and a butanol-rich phase $\dot{S}2$ in a decanter S-1. Because $\dot{S}2$ is located in a different distillation area from $\dot{S}1$, butanol can be withdrawn from column C-2 as bottom product. The composition of the overhead fraction from C-2 is approximately that of the heterogeneous azeotrope and can also be fed into the decanter.

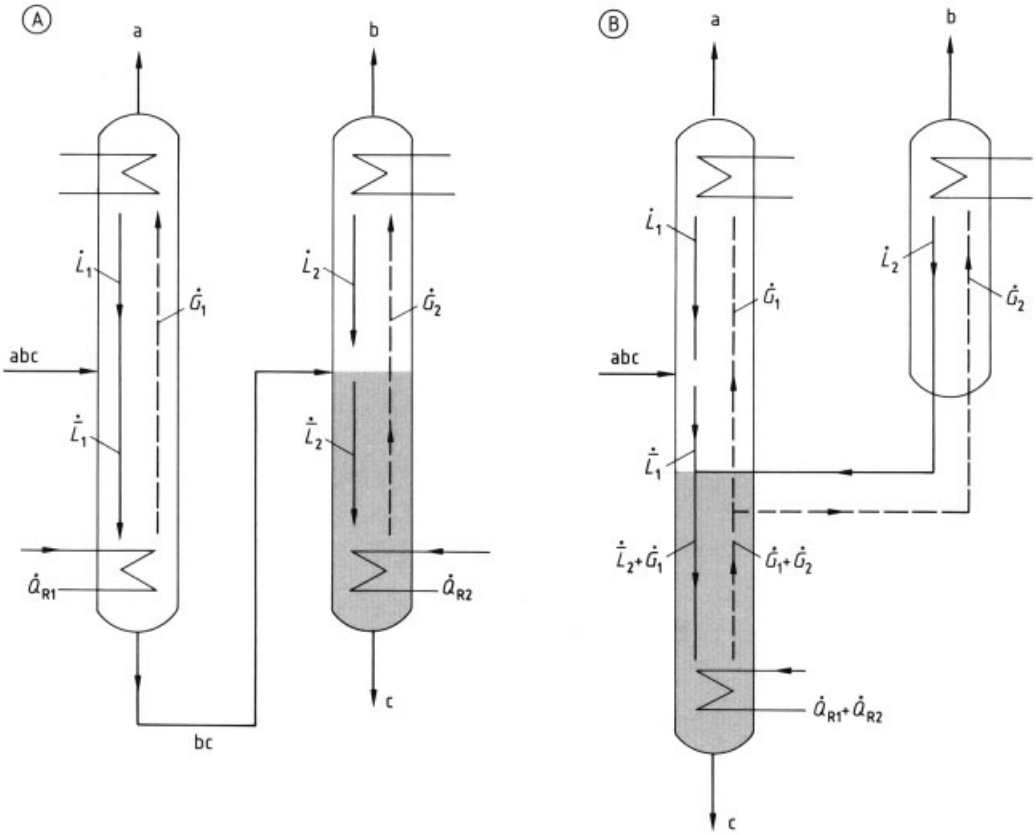

Figure 43. Use of a side column in the separation of a ternary mixture
Shaded portion of the second column in the conventional two-column design (A) can be combined with the first column (B). The remainder of the second column is called the side column and may be either a rectifying or a stripping section. In this figure, it is a rectifying section.

6.4. Separation of Binary Azeotropic Mixtures by Changing System Pressure

Many mixtures contain a minimum-boiling azeotrope which allows the position of the azeotrope to be shifted substantially by changing system pressure. An important example is the mixture tetrahydrofuran – water (Fig. 46) [43]. As shown in Figure 46 B, this mixture forms a binary minimum azeotrope, which shifts toward the low-boiling substance (tetrahydrofuran) at lower system pressure.

The binary azeotrope can be crossed by first separating the high-boiling substance (water) at low pressure (0.1 MPa) in column C-1 (Fig. 46 A). The composition of the overhead product $\dot{D}1$ should be as close as possible to that of the azeotrope at this pressure. After the pressure has been increased to 0.8 MPa, $\dot{D}1$ is fed into a second column C-2. At the higher pressure in this column, the azeotrope forms at a lower concentration of low boiler (tetrahydrofuran), which can then be removed as bottom product. The overhead product of column C-2 ($\dot{D}2$), preferably in the vapor state, is returned to column C-1 after pressure reduction. Hence, column C-1 has a liquid feed $\dot{F}$ and a vapor feed $\dot{D}2$; its operating line shows two bends (Fig. 46 B).

The overhead temperature of column C-2 is higher than the boiling temperature of column C-1 [43]. Therefore, the requirements for an indirect (thermal) column coupling are fulfilled (see Section 7.3).

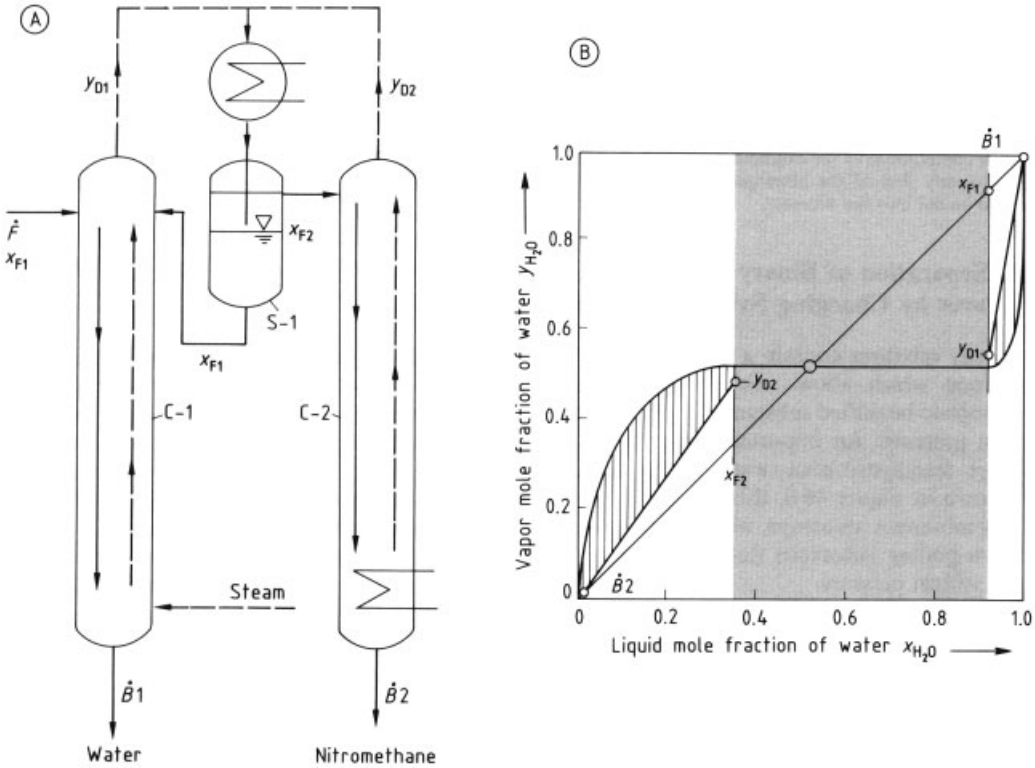

Figure 44. Flow diagram (A) and phase diagram (B) for separation of the binary mixture water – nitromethane
The water – nitromethane mixture forms an immiscibility region (shaded area) with one heterogeneous azeotrope. Complete separation is possible by using two columns C-1 and C-2 and a decanter S-1. The vertically hatched region on the y – x diagram represents the area between the operating line and the equilibrium curve in which the number of equilibrium steps n or transfer units N_{OG} is determined.

6.5. Separation of Azeotropic Mixtures with an Entrainer (Azeotropic and Extractive Distillation)

Binary azeotropic mixtures can be separated into pure substances by the addition of an auxiliary substance called an entrainer. The choice of entrainer must permit easy separation of the resulting ternary mixture (a, b, and entrainer e) into the starting components a and b as overhead or bottom product. Ternary mixtures possessing this property have at least one distillation boundary; i.e., the ternary mixture has several distillation fields (see Section 4.3).

6.5.1. Process Description

Figure 47 A shows a general flow diagram and Figure 47 B the corresponding triangular diagram for the separation of a binary maximum-boiling azeotrope with an entrainer. Substance b is separated overhead in column C-1 and lies in the same distillation field as the feed $\dot{F}$. The separation boundary is crossed by admixing the entrainer e with the bottom fraction $\dot{B}1$ of column C-1. Because the mixing point $\dot{M}2$ lies in the other distillation area of the triangular diagram, the second substance (a) can be separated as an overhead product in column C-2. The bottom fraction of column C-2 ($\dot{B}2$) is fed into column C-3, where the entrainer is recovered and returned to column C-2 in fraction $\dot{B}3$. The overhead fraction of column C-3 ($\dot{D}3$) still contains a considerable amount of substance b, which is separated by returning this fraction to the feed. Thus, the distillation boundary is recrossed and the mixture is again in the distillation field where substance b can be separated. This system can also be used for minimum azeotropes, but pure

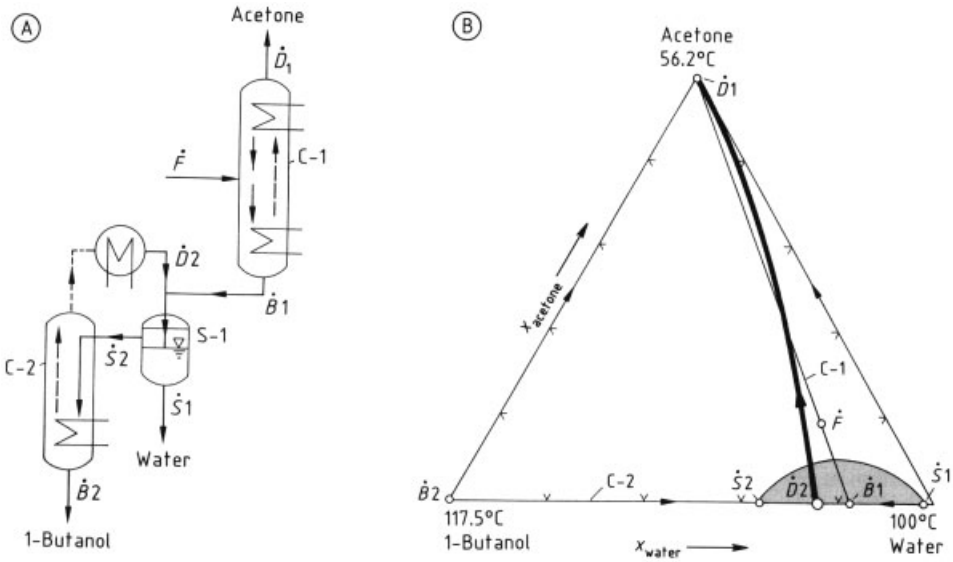

Figure 45. Flow diagram (A) and phase diagram (B) for separation of the ternary mixture acetone – water – 1-butanol (p = 101.3 kPa)
The existence of an immiscibility region (shaded area) permits complete separation of the mixture in two columns C-1 and C-2 with an intermediate separator S-1.

substances a and b are then taken from the bottom of the columns.

This generalized separation process contains three separating columns and two recycling steps. Six fractions are obtained from the three separating columns. Three of the fractions are products a and b and the entrainer e; two lie on the separation boundary. The remaining fraction of column C-2 lies in the same separation field as the entrainer. The two recycling steps allow the separation boundary between the distillation fields to be crossed twice. This is possible only when the entrainer and the feed are located in different separation areas, which is always achieved by using an entrainer-rich fraction instead of a pure entrainer in the process. Such processes are generally known as *azeotropic distillation.*

6.5.2. Choice of Entrainer

The entrainer must permit the recovery of products a and b at the top or bottom of the rectification columns. This is possible only if the distillation curves on the triangular diagram either start or end at a and b (see Section 4.3). Figure 48 shows the minimum qualitative requirements that must be met by the entrainer.

In the case of a maximum-boiling binary azeotrope (Fig. 48 A – C), components a and b in the starting mixture are low-boiling substances; i.e., the distillation curves terminate at points a and b. The entrainer used must ensure that the distillation curves of both binary mixtures a – e and b – e terminate at a and b, respectively. This condition is fulfilled when entrainer e constitutes the high boiler of the system (Fig. 48 A). When the entrainer is a medium-boiling substance, component a can be separated, but component b can be recovered only if the mixture b – e forms a maximum azeotrope (Fig. 48 B). With a low-boiling entrainer (Fig. 48 C), mixtures b – e and a – e must both form a maximum azeotrope. Products a and b leave the rectification column overhead.

If a minimum-boiling binary azeotrope occurs (Fig. 48 D – F), a and b are high-boiling substances; i.e., the distillation curves start at a and b. This must also be the case in a ternary system. This condition is always fulfilled when the entrainer is a low-boiling substance (Fig. 48 F). If the entrainer has a boiling temperature between that of a and b, pure b can be separated;

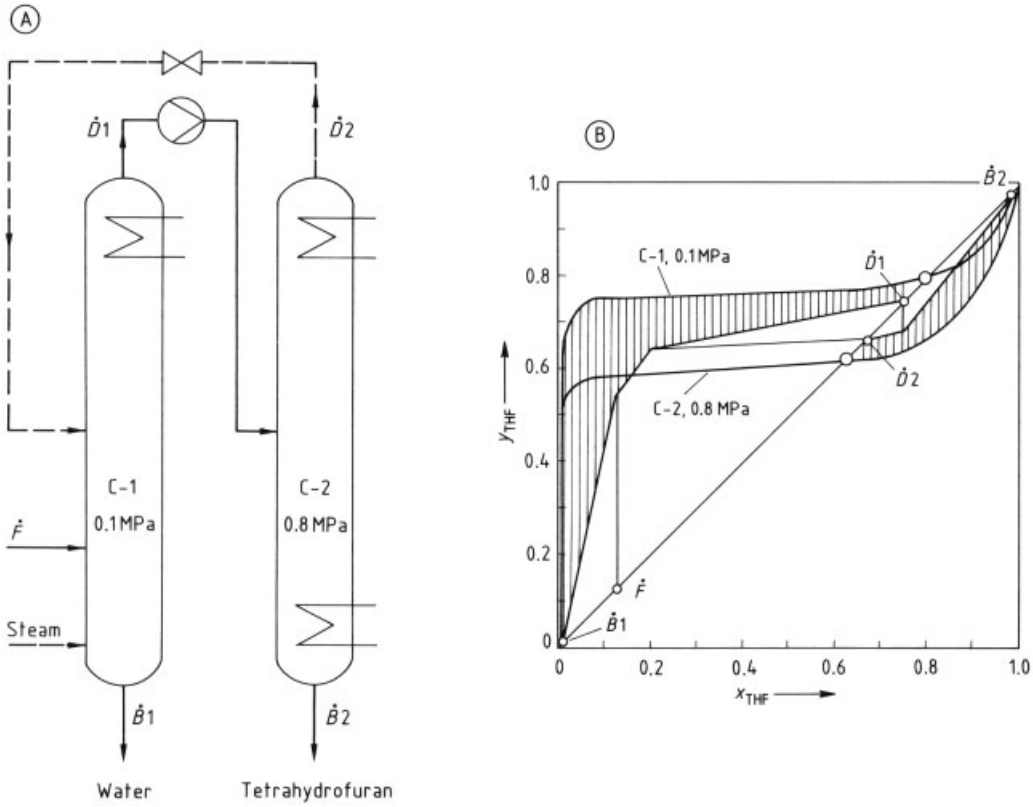

Figure 46. Flow diagram (A) and phase diagram (B) for separation of a binary tetrahydrofuran – water mixture in two columns, C-1 and C-2, operating at different pressures
This mixture contains a minimum azeotrope, whose position can be shifted by changing system pressure. Hatched region represents area between the equilibrium curves and the operating lines in which the number of equilibrium steps n or transfer units N_{OG} is determined.

however, substance a can be recovered only if the mixture a – e forms a minimum azeotrope (Fig. 48 E). If the entrainer is a high-boiling substance (Fig. 48 D), a and b can be separated only if the entrainer forms a minimum azeotrope with both substances. Substances a and b are obtained as bottom products.

Thus, a binary mixture with a maximum azeotrope can generally be separated into its two components only if the entrainer is higher boiling than the pure substances or if it forms high-boiling azeotropes (i.e., maximum azeotropes). On the other hand, a mixture with a minimum azeotrope requires a low-boiling entrainer or the formation of low-boiling azeotropes. In other words, only azeotropes of the same type occur in one process.

6.5.3. Entrainment Rectification of Azeotropic Mixtures in Combination with Other Separation Methods

The generalized rectification scheme shown in Figure 47 A requires three steps in separate rectification columns. However, inclusion of other operations is sometimes advantageous, e.g., decantation, desorption, absorption, or extraction.

Decantation. Decantation can be used in place of a rectification column if the entrainer forms a heterogeneous azeotrope; this represents a simplification of the general rectification scheme. Because heterogeneous azeotropes are always minimum azeotropes and only azeotropes of the same type can occur in one process, decanters can be used only with mixtures that exhibit a minimum azeotrope.

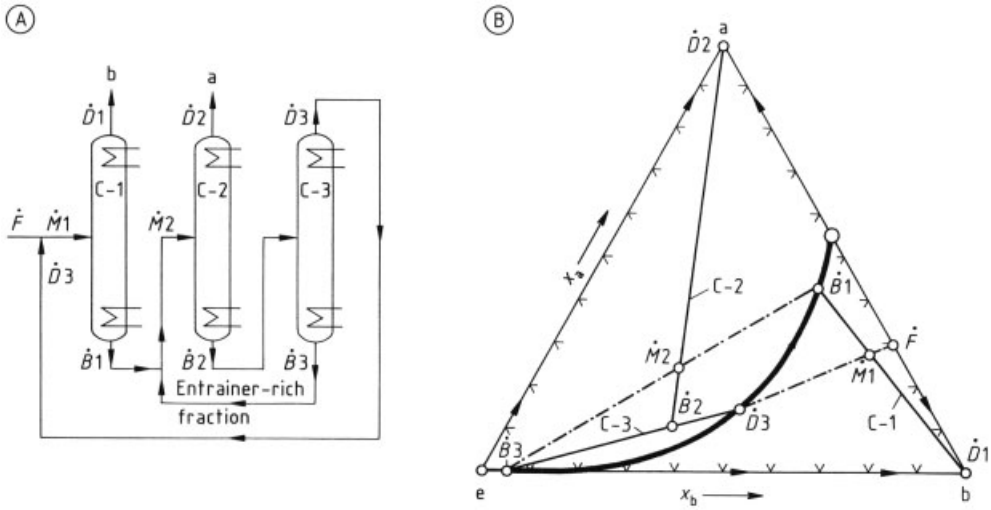

Figure 47. Flow diagram for the separation of binary mixture a – b with a maximum azeotrope by using an entrainer e
A) Flow diagram: components a and b are obtained as pure overhead products from columns C-2 and C-1, respectively. In mixtures with a minimum azeotrope, pure substances a and b are taken from the bottom of the columns. B) Triangular phase diagram: the ternary mixture is divided into two distillation fields by a distillation boundary running from the entrainer to the maximum azeotrope. Rectification separation is possible only within one distillation field. The distillation boundary is crossed by the admixing of entrainer e.
—— Distillation boundary; —— Balance around the rectification column; – · – · – Mixing of two liquids

An important example is the production of absolute ethanol by using toluene as entrainer (Fig. 49) [44]. Ethanol and water form a minimum azeotrope at about 90 mol% of ethanol. According to Section 6.5.2, a low-boiling entrainer is required. However, the entrainer toluene is the highest boiling substance in the three-component mixture. Toluene forms a minimum azeotrope with ethanol and a minimum heterogeneous azeotrope with water. Hence, the entrainer complies with the separation requirements described in Section 6.5.2. In addition, a ternary azeotrope exists which has the lowest boiling temperature in the system.

Water is first separated as a bottom product in column C-1. The overhead fraction $\dot{D}1$ lies on the separation boundary, which is crossed by the admixing of a toluene-rich phase. In column C-2, ethanol is removed as bottom product. The temperature of the overhead product $\dot{D}2$ is reduced to approximately 25 °C, which brings it into the two-phase region. The toluene-rich phase $\dot{S}2$, used as entrainer, is recovered as the light phase in a decanter S-1. The water-rich phase $\dot{S}1$ is taken as the heavy phase from the decanter and combined with the feed. Separator S-1 replaces separating column C-3 in the generalized scheme shown in Figure 47.

A decanter is also used in separating pyridine – water mixtures which form a minimum azeotrope [45]. According to Figure 48, a mixture with a minimum azeotrope requires a low-boiling entrainer. In this example, however, toluene is employed, which has a boiling point between that of pyridine and water. Toluene forms a minimum azeotrope with water, so the requirements of Figure 48 E are satisfied. In addition, toluene forms a minimum azeotrope with pyridine, but this is of no importance.

The separation is shown schematically in Figure 50. Feed $\dot{F}$ is mixed with the entrainer and introduced into column C-1. Pure pyridine ($\dot{B}1$) is taken as bottom product from this column. The overhead fraction $\dot{D}1$ has the approximate composition of the heterogeneous water – toluene azeotrope. This fraction is cooled to 25 °C and separated into two liquid phases in the decanter S-1. The heavy phase $\dot{S}2$ contains mainly water, which is removed as product. The toluene-rich phase $\dot{S}1$ is returned to the process as entrainer.

In contrast to the generalized scheme shown in Figure 47 A, only one rectification column and one separator are required to separate an

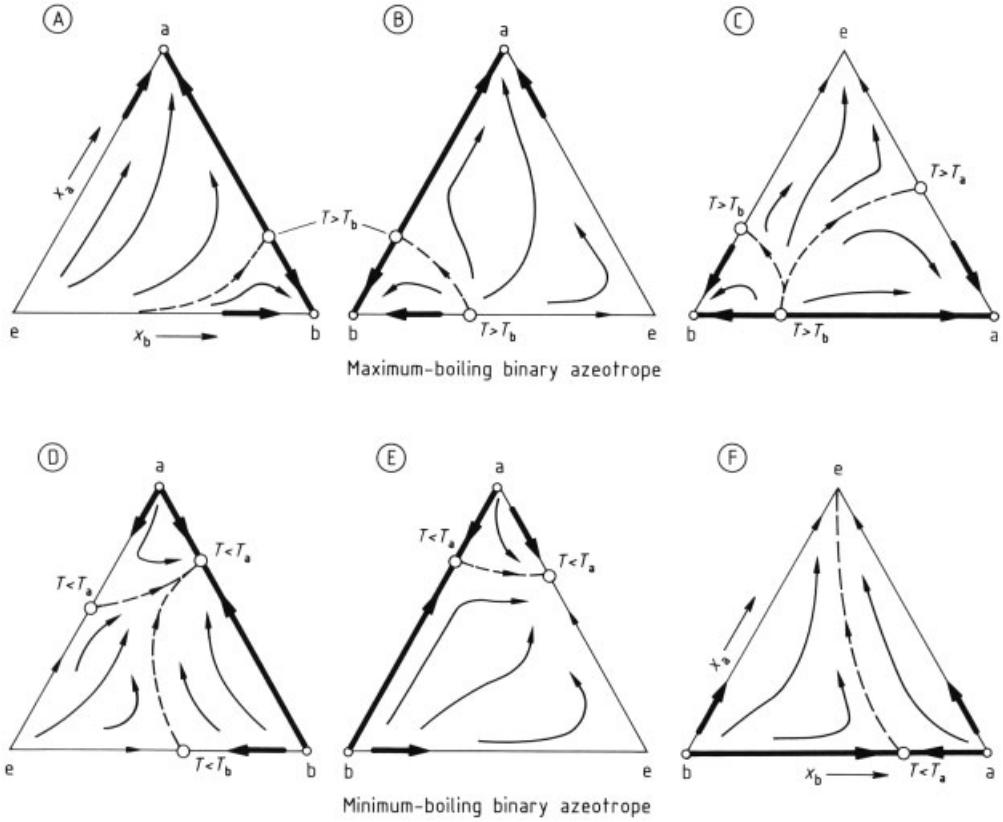

Figure 48. Minimum requirements to be met by entrainer e for separation of binary mixture a – b with a maximum azeotrope (A – C) or a minimum azeotrope (D – F)
Separation into pure a and b is possible only when the distillation curves either start or terminate at a and b. If a maximum azeotrope is present, either the entrainer must be a high-boiling substance (A), or it must form new maximum azeotropes (B, C). If a minimum azeotrope is present, the entrainer must be a low-boiling substance (F) or it must form new minimum azeotropes (D, E).
Azeotropic points are represented by the large open circles.
T = azeotrope boiling point; T_a = boiling point of component a; T_b = boiling point of b; – – – distillation boundary; —— distillation curve

azeotropic mixture. The favorable position of the immiscibility region permits simultaneous performance of two separating tasks, namely, separation of water as a product and recovery of the entrainer.

Desorption is used in the recovery of nitric acid from an aqueous mixture (→ Nitric Acid, treated in the A series) [46]. The system nitric acid – water forms a maximum azeotrope at a nitric acid concentration of 39 mol%. Accordingly, a high-boiling entrainer should be selected. Although sulfuric acid forms a maximum azeotrope with water at 96.8 mol% and is used as entrainer, this has no importance in the separation procedure.

The process is illustrated in Figure 51. The separation boundary is crossed by mixing approximately 46 mol% of sulfuric acid with the feed. The separation of nitric acid as overhead product $\dot{D}1$ is possible in the separation field in which the mixing point $\dot{M}$ is located. After air and steam stripping, the bottom product $\dot{B}1$ is practically free of nitric acid, giving fraction $\dot{\bar{B}}1$. The entrainer-rich fraction $\dot{B}2$ is recovered from $\dot{\bar{B}}1$ by removing water as fraction $\dot{D}2$ in a vacuum distillation column C-2.

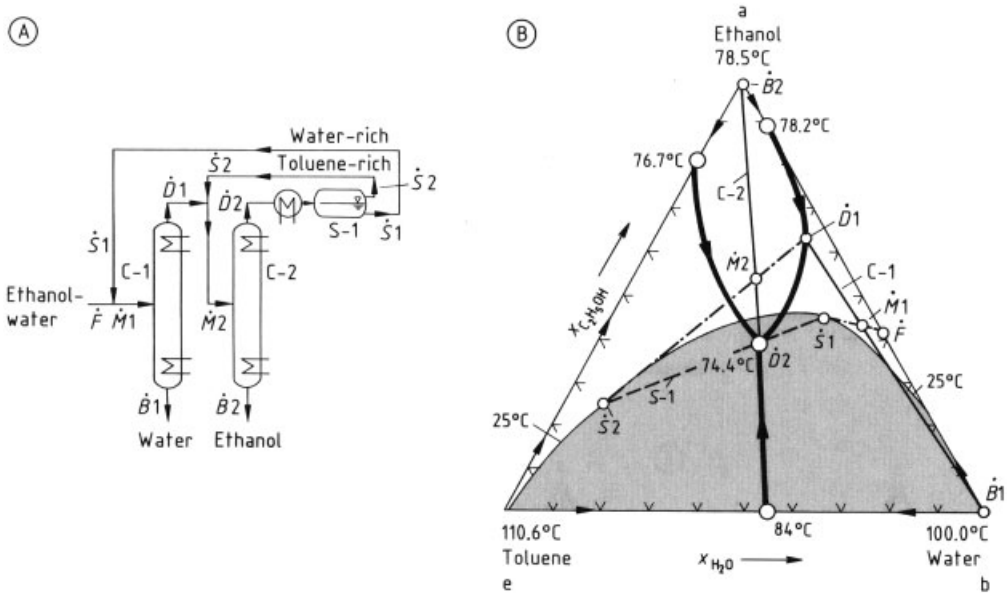

Figure 49. Separation of a binary ethanol – water mixture using toluene as entrainer
A) Flow diagram showing rectification columns C-1 and C-2 and decanter S-1
B) Triangular phase diagram (p = 101.3 kPa): because the mixture contains three binary minimum azeotropes and one ternary azeotrope, it is divided into three distillation fields. Rectification separation is possible only within a distillation field; distillation boundaries can be crossed by the mixing of different liquids.
—— Distillation boundaries; —— Balance around the rectification column; – · – · Mixing of liquids; – – – Separation in the decanter. Shaded area represents the immiscibility range at 25 °C.

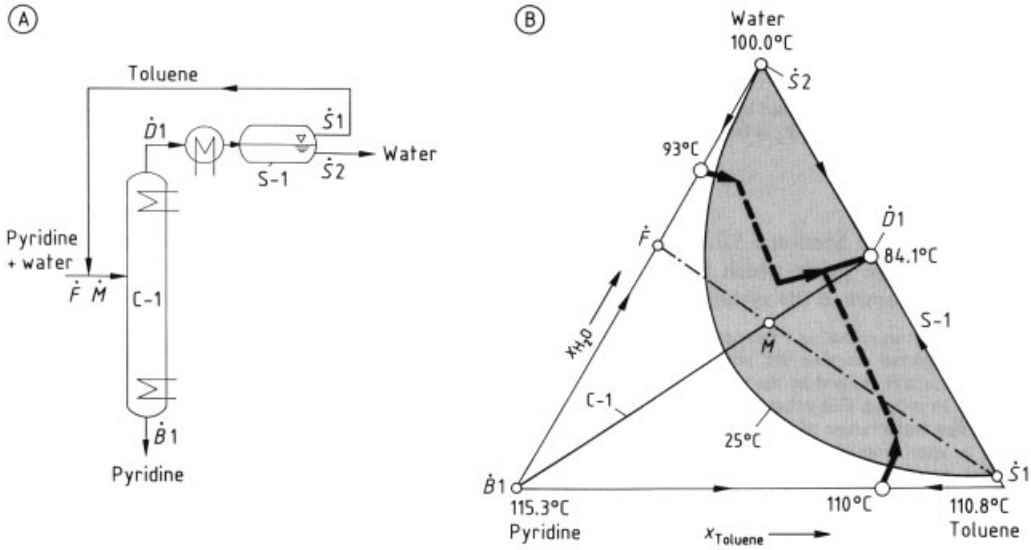

Figure 50. Separation of a water – pyridine mixture by using toluene as entrainer
A) Flow diagram showing rectification column C-1 and decanter S-1
B) Triangular phase diagram (p = 101.3 kPa)
—— Distillation boundary; —— Rectification; – · – · Mixing of liquids. Shaded area represents the immiscibility range at 25 °C.

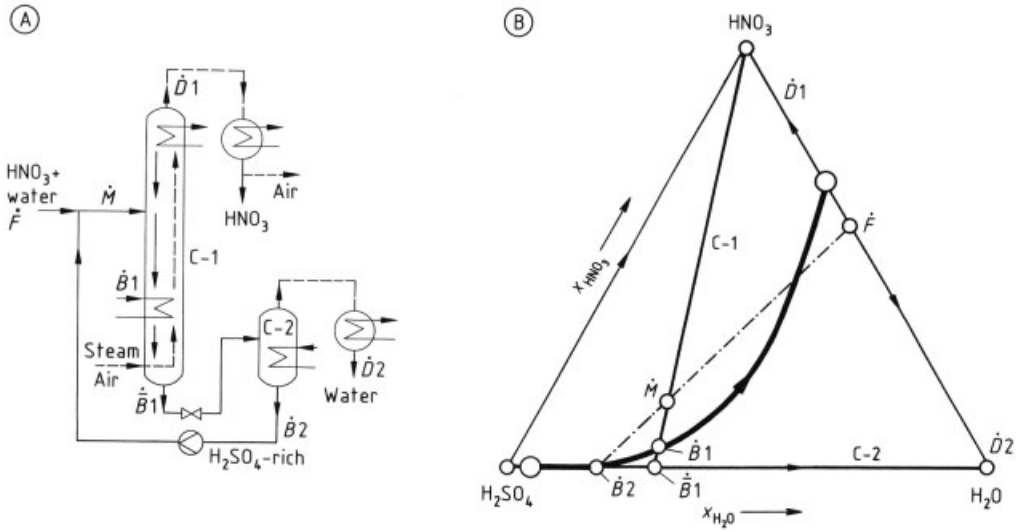

Figure 51. Concentration of nitric acid and water by using sulfuric acid as entrainer
A) Flow diagram showing one rectification column C-1 and one distillation column C-2
B) Triangular phase diagram
—— Distillation boundary; —— Rectification; – · – · Mixing of liquids

The unusual feature of this process is the desorption step in the lower part of column C-1, which permits crossing of the separation boundary between the two distillation fields, as shown in Figure 51 B. Consequently, the separation of water and the simultaneous recovery of entrainer are possible. Thus only one recycling step and two separation columns are required; column C-2 is used for simple distillation. Furthermore, the entrainer and the feed are located in the same distillation field, which should not be permitted according to the general process scheme. However, the separation boundary is curved to such an extent that the mixing of $\dot{B}2$ and $\dot{F}$, both of which lie in the same distillation field, results in crossing of the separation boundary. However, the composition of the feed must then be almost the same as that of the binary azeotrope.

A similar process is used to concentrate dilute hydrochloric acid [47] with sulfuric acid as entrainer.

Absorption can be conveniently used to separate mixtures containing a minimum azeotrope, such as ethanol – water. The high boiler is separated in the first column C-1 (see Fig. 52 A). The overhead fraction has approximately the same composition as the binary azeotrope. The vapor mixture is fed into an absorber A-1, in which the entrainer is used to wash out primarily the higher boiling component b. Pure substance a is thus obtained at the top of the absorber. The loaded washing agent from the absorber is introduced into column C-2 where entrainer e is recovered as bottom product. The overhead product of C-2 contains substances a and b, and is returned to column C-1.

The entrainer does not have to satisfy the criteria in Figure 48 because the azeotrope distillation boundary is crossed in the absorption column. However, the entrainer must have a greater affinity for the high boiler b than for the low boiler. In the dehydration of ethanol, a hygroscopic liquid (e.g., ethylene glycol) is used for this purpose.

The equipment required can be simplified, as shown in Figure 52 B, by placing the absorption column A-1 on top of the rectification column C-1, thereby eliminating a reflux condenser. The heavy entrainer also flows through the rectification column C-1. The high boiler b is withdrawn from the top of column C-2. A relatively large number of separations are carried out in this way. Important examples are the separation of ethanol – water with diethylene glycol, of benzene – cyclohexane with ani-

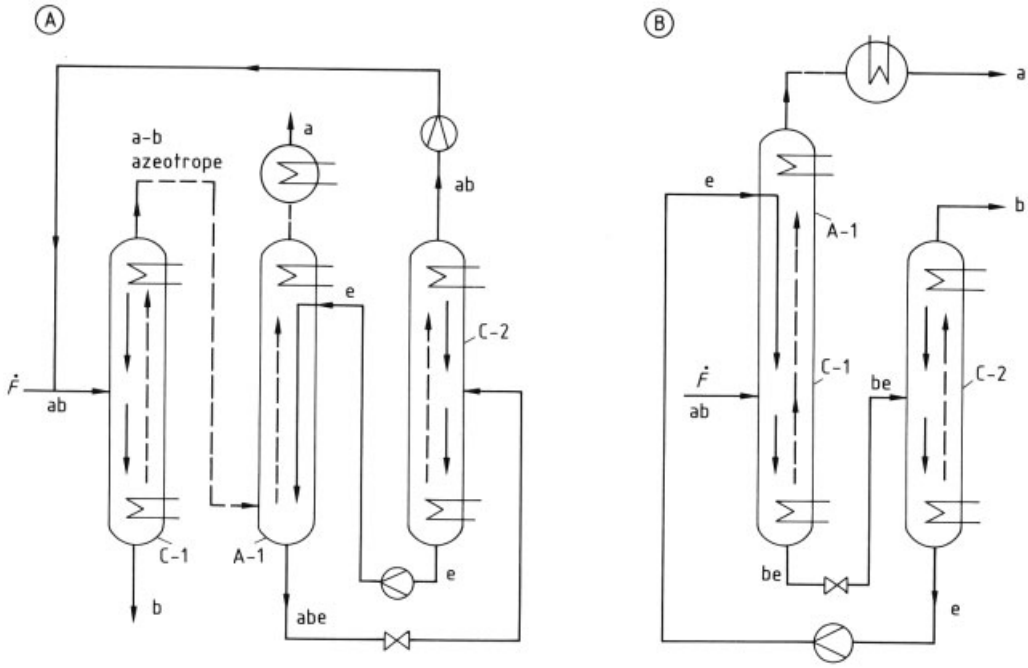

Figure 52. Separation of binary mixture a – b with one minimum azeotrope by using an absorption step
A) Flow diagram showing two rectification columns C-1 and C-2 and an absorption column A-1
B) Flow diagram showing the absorption column A-1 on top of the rectification column C-1This process is generally known as extractive distillation although it really involves absorption and not extraction.

line, and of butadiene – butene with N-methyl-2-pyrrolidone [48], [49].

The process shown in Figure 52 B is generally known as *extractive distillation,* although it actually involves absorption and not extraction. A more correct name would be absorptive distillation.

7. Energy Economization in Rectification

The energy requirement $\hat{W}$ for separation of a mixture into its pure components i can be calculated [50] from

$$\hat{W} = -\hat{R}\cdot T\cdot \sum x_{\mathrm{F}i}\cdot \ln\left(\gamma_i \cdot x_{\mathrm{F}i}\right) \tag{7.1}$$

This relationship is independent of the separation method and merely describes the reversal of the mixing procedure (see Eq. 2.30).

Energy required for rectification is supplied in the form of heat. The exergy E describes the ability of heat to be converted into work:

$$E = Q\cdot\left(T_1 - T_0\right)/T_1 \tag{7.2}$$

where Q is the amount of heat, T_1 the heat input temperature, and T_0 the ambient temperature. The work performed W is calculated from the exergy difference between the heat input temperature T_1 and the heat output temperature T_2:

$$W = Q\cdot T_0\left[\frac{1}{T_2} - \frac{1}{T_1}\right] \tag{7.3}$$

Thus, energy consumption depends not only on the quantity of heat Q, but also on the temperature.

In a rectification column, a quantity of heat $\dot{Q}_{\mathrm{R}}$ is required at the bottom of the column at temperature T_{R} and a quantity of heat $\dot{Q}_{\mathrm{C}}$ is removed at the top at temperature T_{C}. The following relationship holds:

$$\dot{Q}_{\mathrm{R}} + \dot{Q}_{\mathrm{C}} \approx 0, \text{ where} T_{\mathrm{R}} - T_{\mathrm{C}} > 0 \tag{7.4}$$

The temperature difference $T_R - T_C$ is usually ca. 5 – 20 °C, so exergy losses are normally small. However, the heat required for rectification is usually supplied in the form of steam at temperature $T_1 \gg T_R$ and is removed by cooling water at $T_2 \ll T_C$. In this way, larger amounts of exergy are lost during separation than are actually necessary.

Means of saving energy during rectification include reducing the heat requirement $\dot{Q}$ and minimizing the temperature difference $(T_1 - T_2)$ between heat input and output (see also → Energy Management in Chemical Industry, Chap. 6.

7.1. Single Columns

7.1.1. Reduction of Heat Requirement

The heat requirement $\dot{Q}$ of a single rectification column can be minimized best by effective heat exchange between the warm products and the cold feed, and also by efficient insulation of the column. Furthermore, the operational reflux ratio R should be kept as small as possible; optimum ratios are in the range of $R = 1.05 - 1.1\, R_{min}$ [51].

The thermal state of the feed (caloric factor φ) greatly influences the heat requirement of a column. The theoretical minimum values $\dot{Q}/(\dot{F} \cdot r)$ for ideal binary mixtures are given in Figure 53.

The heat requirement of a liquid feed (Fig. 53 A) is greater by the additive term x_{Fa} than that of a vapor feed (Fig. 53 B). In a vapor feed, the latent heat is used only in the top of the column. Indeed, the vapor feed should be condensed in the reboiler before it is introduced into the column (Fig. 53 C). In this way, latent heat is used in both parts of the column, which leads to a lower external heat-input requirement $\dot{Q}_R$. The column can then be operated without an external heat source if the feed concentration y_{Fa} is low and the relative volatility α is high. This mode of operation is often employed in low-temperature processes, but the feed pressure must be higher than the operating pressure in the column. Hence, a vapor feed is preferable to a liquid feed. However, a liquid should not be prevaporized because then the total energy requirement is higher than with a liquid feed (Fig. 53 D).

When the feed is composed of several streams with different compositions, multiple feeding should be considered (see Fig. 20). Optimal choice of feed point and careful column control are very important.

7.1.2. Reduction of Exergy Losses

The use of intermediate reboilers and condensers has been proposed frequently. Although they do not lower the heat-input requirements of a column, part of the heat required can be supplied at a lower temperature and removed at a higher temperature (i.e., exergy losses are reduced). This method offers practical advantages only if variable external heat sources (e.g., steam with variable pressure) or heat sinks are available, or if thermal coupling is possible (Section 7.3.2).

Exergy losses can be effectively reduced by utilizing waste heat obtained at the top of the column to produce low-pressure steam (Fig. 54 A) [52]. No steam is "consumed" during rectification, and high-pressure steam is converted into low-pressure steam. Thus, exergy, and not energy, is consumed. If a low-pressure steam system is not available, steam can be returned to the high-pressure system after compression (Fig. 54 B).

Systematic development of this operational mode leads to incorporation of a heat pump with supplementary circulation (Fig. 54 C). The selection of a suitable working medium for the heat pump always arises. The substances separated in the rectification column can often be used to advantage as a working medium for the heat pump. Consequently, the pump can be integrated into the rectification [53]. Utilization of the overhead product as a heat pump medium (Fig. 55 A) is usually more favorable than utilization of the bottom product (Fig. 55 B).

The integrated heat pump is used in many industrial low-temperature plants (e.g., ethylene plants) [54]. In high-temperature processes, a problem occurs because the high molecular mass substances present have a very small adiabatic exponent $\kappa = c_p/c_v$. Therefore, the vapor temperature barely rises on compression; it may fall below the dew point and cause machine damage due to condensation [55].

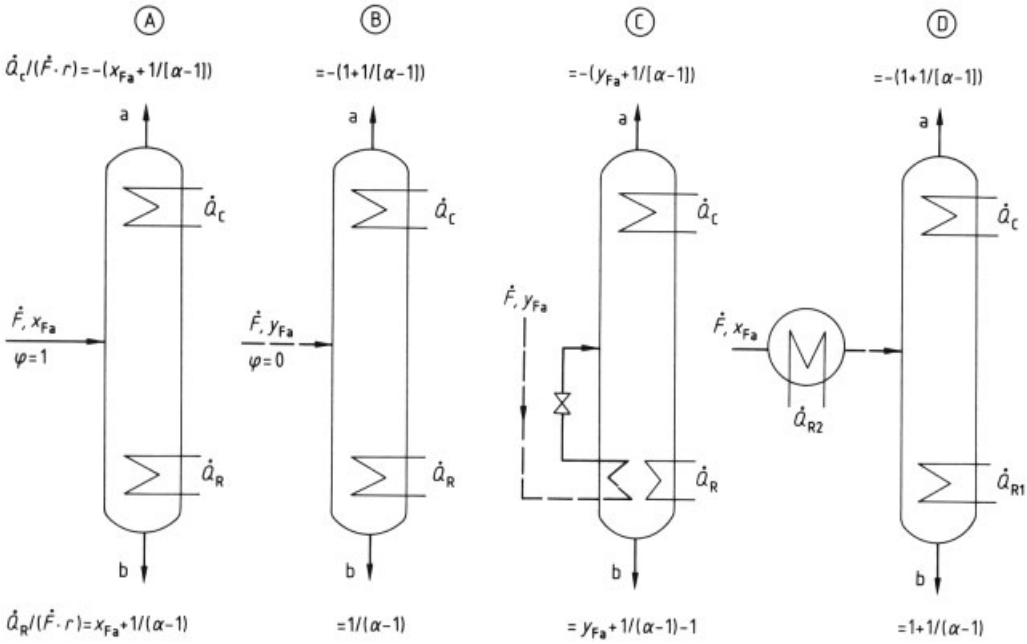

Figure 53. Influence of thermal state of the feed on minimal heat turnover during complete rectification of an ideal binary system a – b in a single column
A) Liquid feed; B) Vapor feed; C) Condensation of vapor feed in the reboiler; D) Prevaporization of liquid feed
$\dot{F}$ = molar feed stream; r = latent heat of vaporization of feed; $\dot{Q}_C$ = heat removed by condenser; $\dot{Q}_R$ = heat supplied to reboiler; x_{Fa} = liquid mole fraction of a in feed; y_{Fa} = vapor mole fraction of a in feed; α = relative volatility of a; φ = thermal state of feed

7.2. Optimal Separation Sequences for Multicomponent Mixtures

As described in Section 6.2, a large number of different separating sequences exist for complete fractionation of multicomponent mixtures. Investment and operating costs depend on the separating path selected. Operating costs are determined primarily by energy costs. Heat requirements for the different separating paths in a ternary mixture are presented below.

The minimal heat-input requirements for separation of an ideal ternary mixture a – b – c into pure substances by the a, c, and a – c paths can be formulated in general terms as described in Chapter 4 and are illustrated in Figure 56.

In path a (Fig. 56 A) the low-boiling substance a is removed first. The remaining binary mixture b – c is fed into column C-2 and separated into b and c. If the feed is a boiling liquid, then according to Section 4.2.5 the following expression holds for the minimum heat requirement of column C-1:

$$\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r} = \left(R_{\mathrm{a,min}} + 1\right)\cdot x_{\mathrm{Fa}} \tag{7.5}$$

where r is the molar latent heat of vaporization of the feed and $R_{\mathrm{a,\,min}}$ is defined in Equation (4.64).

The minimum heat requirement of column C-2 is given by

$$\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r} = x_{\mathrm{Fb}} + \frac{1 - x_{\mathrm{Fa}}}{\alpha_{\mathrm{bc}} - 1} \tag{7.6}$$

The heat requirement for the entire process is

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r} = \frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r} + \frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r} \tag{7.7}$$

Values of $\dot{Q}_{\min}/(\dot{F}\cdot r)$ for the mixture 2,2-dimethylbutane – 2-methylpentane – hexane are in the range of 3.4 to 6.3. In Figure 56 A, they are plotted in parametric form against the feed composition of the mixture on a triangular diagram; they increase more rapidly as the concentration of medium-boiling substance b increases.

In path c, the high boiler c is removed from column C-1 (Fig. 56 B). The overhead product

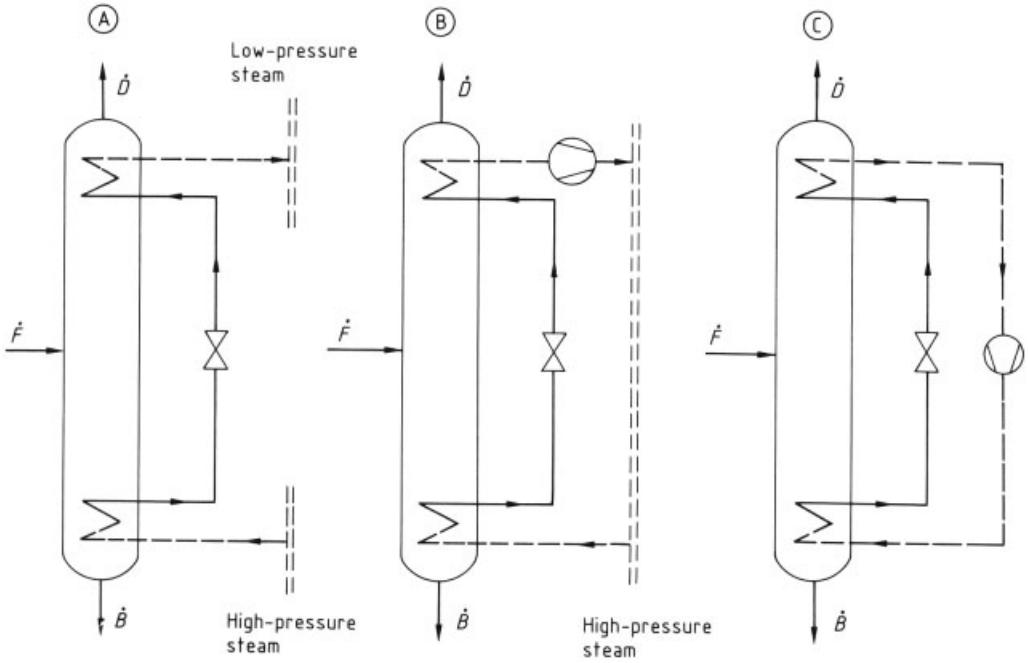

Figure 54. Reduction of exergy losses during rectification
A) Production of low-pressure steam; B) Compression and recycling of low-pressure steam; C) Heat pump with supplementary circulation

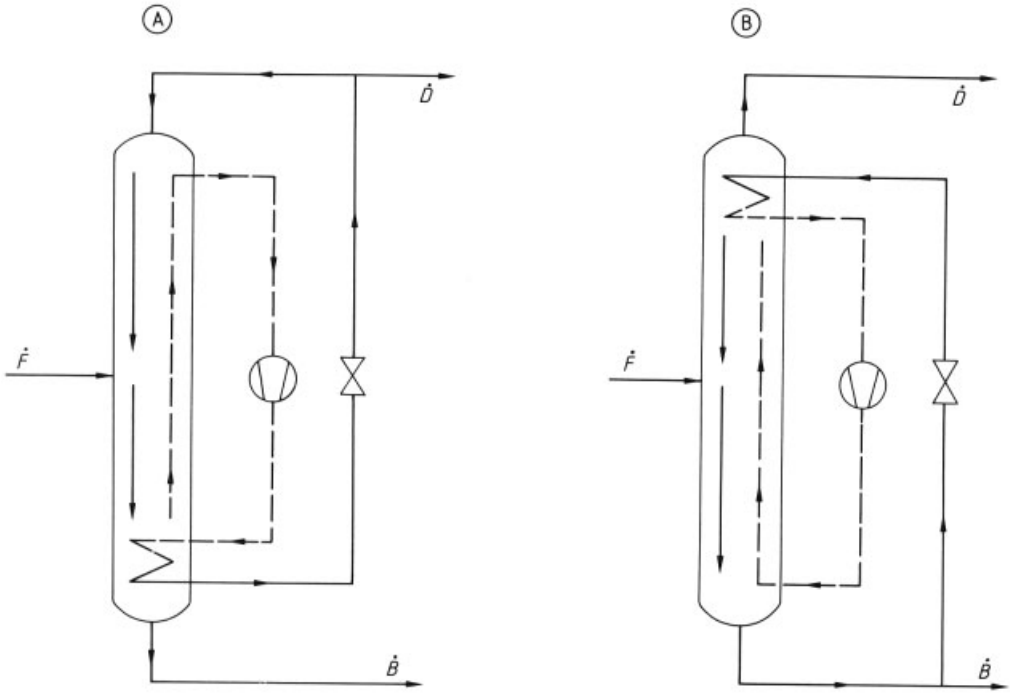

Figure 55. Schematic of integrated heat pumps using the overhead product (A) or the bottom product (B) as the pump medium

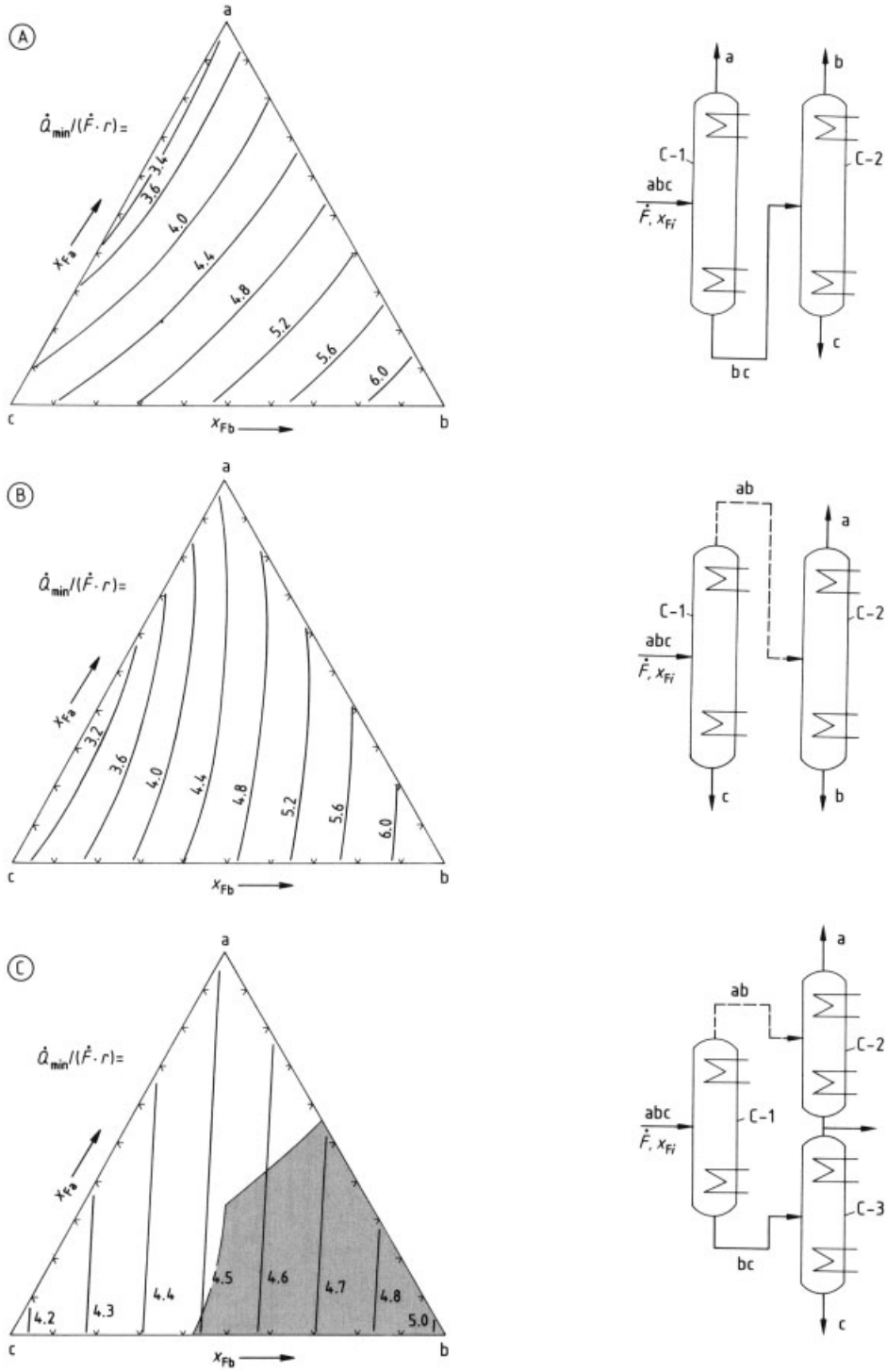

Figure 56. Minimum heat requirement $\dot{Q}/(\dot{F}\cdot r)$ as a function of the feed composition for separation of a ternary mixture of 2,2-dimethylbutane (a), 2-methylpentane (b), and hexane (c)
A) Path a; B) Path c; C) Path a – c
Relative volatility: $\alpha_{ac} = 1.887$; $\alpha_{bc} = 1.329$. Shaded area in C represents the concentration range in which path a – c is more favorable than paths a and c.

of column C-1 (ab) is fed in the vapor state into column C-2, where pure products a and b are separated. For a boiling-liquid feed, the heat required by column C-1 is

$$\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}=\bar{R}_{\mathrm{c},\min}\cdot x_{\mathrm{Fc}} \tag{7.8}$$

where $\bar{R}_{\mathrm{c},\,\min}$ is obtained from Equation (4.65). For a liquid feed, the heat required by column C-2 is

$$\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}=x_{\mathrm{Fa}}+\frac{1-x_{\mathrm{Fc}}}{\alpha_{\mathrm{ab}}-1} \tag{7.9}$$

For the more favorable vapor feed, the heat required by column C-2 is

$$\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}=\frac{1-x_{\mathrm{Fc}}}{\alpha_{\mathrm{ab}}-1} \tag{7.10}$$

The heat requirement of the entire process is expressed by

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r}=\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}+\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r} \tag{7.11}$$

The values of $\dot{Q}_{\min}/(\dot{F}\cdot r)$ for the mixture 2,2-dimethylbutane – 2-methylpentane – hexane are in the range of 3.0 to 6.3; here again, the highest values are obtained at high concentrations of the medium boiler b (Fig. 56 B).

As discussed in Section 4.3.1, in the large number of possible rectification products from a three-component mixture, separation of the low boiler a or the high boiler c represents a special case.

In path a – c, three separating columns are required (Fig. 56 C). To obtain pure substances, the overhead product from column C-1 must be free of the high boiler c. Correspondingly, the bottom product of column C-1 must be free of the low boiler a. The overhead product should preferably be fed in the vapor state into column C-2.

In the following discussion, the assumption is made that the "preferred" separation is carried out in column C-1 (cf. Section 4.3.2.1). Thus, the reflux ratio and the heat requirement of this column reach an absolute minimum. If the feed is a boiling liquid, Equations (4.44), (4.56), and (4.57) give the following relationship for column C-1:

$$\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}=\frac{1+(\alpha_{\mathrm{ac}}-1)\cdot x_{\mathrm{Fa}}+(\alpha_{\mathrm{bc}}-1)\cdot x_{\mathrm{Fb}}}{\alpha_{\mathrm{ac}}-1} \tag{7.12}$$

The heat requirement of column C-2 for a liquid feed is

$$\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}=x_{\mathrm{Fa}}+\frac{(\alpha_{\mathrm{ac}}-1)\cdot x_{\mathrm{Fa}}+(\alpha_{\mathrm{bc}}-1)\cdot x_{\mathrm{Fb}}}{(\alpha_{\mathrm{ac}}-1)\cdot(\alpha_{\mathrm{ab}}-1)} \tag{7.13}$$

For a vapor feed, the following equation is obtained for column C-2:

$$\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}=\frac{(\alpha_{\mathrm{ac}}-1)\cdot x_{\mathrm{Fa}}+(\alpha_{\mathrm{bc}}-1)\cdot x_{\mathrm{Fb}}}{(\alpha_{\mathrm{ac}}-1)\cdot(\alpha_{\mathrm{ab}}-1)} \tag{7.14}$$

The heat requirement of column C-3 is

$$\frac{\dot{Q}_{3,\min}}{\dot{F}\cdot r}=x_{\mathrm{Fb}}+\frac{\alpha_{\mathrm{ac}}-\alpha_{\mathrm{bc}}}{\alpha_{\mathrm{ac}}-1}+ \\ +\frac{\alpha_{\mathrm{ac}}-[1+(\alpha_{\mathrm{ac}}-1)\cdot x_{\mathrm{Fa}}+(\alpha_{\mathrm{bc}}-1)\cdot x_{\mathrm{Fb}}]}{(\alpha_{\mathrm{ac}}-1)(\alpha_{\mathrm{bc}}-1)} \tag{7.15}$$

For the entire process,

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r}=\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}+\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}+\frac{\dot{Q}_{3,\min}}{\dot{F}\cdot r} \tag{7.16}$$

In Figure 56 C, the minimal heat requirement for the system 2,2-dimethylbutane – 2-methylpentane – hexane is plotted against the composition of the starting mixture and gives a series of parallel lines. Values for $\dot{Q}_{\min}/(\dot{F}\cdot r)$ are in the range of 4.2 to 5.0; they increase only slightly with increasing concentration of the medium-boiling substance b.

Comparison of heat requirements for the separating paths of a ternary mixture shows large variations depending on the composition of the feed material. If the feed contains a high concentration of the low boiler a, path a (Fig. 56 A) is the most favorable. However, if the feed contains an excess of the high boiler c, path c is preferred, as shown in Figure 56 B. Path a – c is favored when the feed contains an excess of medium-boiling substance b. The concentration range in which path a – c is most useful is indicated by the shaded area in Figure 56 C.

Path a – c requires three separating columns; i.e., it requires more equipment than the other two paths. However, this sequence is exergetically advantageous because heat is added or removed at many different temperature levels; therefore its energy-saving potential is considerable. If exploited properly, path a – c gives a more efficient separation than the other two sequences (see Section 7.3.2).

7.3. Column Coupling

Column coupling may be advantageous in processes involving several rectification columns. In *direct coupling,* columns are coupled by means of their product streams [56]. *Indirect* (*thermal*) *coupling* exploits the waste heat from one column to heat a second column. Indirect coupling is a very effective way of substantially reducing the heat requirement in rectification processes and is used more often than direct coupling. However, it increases equipment requirements; additional pumps and compressors are usually required. Direct coupling saves not only energy but also lowers equipment requirements.

7.3.1. Direct Coupling of Columns

Direct coupling of columns C-2 and C-3 is possible in separating the ternary mixture shown in Figure 56 C. Vapors of substance b are condensed at the top of column C-3 to obtain the required reflux. However, substance b must also be vaporized at the bottom of column C-2 to maintain the countercurrent in this column. Columns C-2 and C-3 can be combined by omitting a condenser and a reboiler to give column $\overline{\text{C-2}}$ (Fig. 57). The amount of heat converted in the two columns generally differs. This heat difference can be compensated with an intermediate heat exchanger. However, the equipment requirement is lower if the heat supplied at the bottom of column $\overline{\text{C-2}}$ corresponds to the maximum heat requirement of one of the two original columns C-2 or C-3, whichever is larger. Thus, the heat requirement of the entire process in Figure 57 is

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r}=\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}+\text{Max}\left[\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}\text{or}\frac{\dot{Q}_{3,\min}}{\dot{F}\cdot r}\right] \quad (7.17)$$

The individual amounts of heat are calculated by using Equations (7.12), (7.14), and (7.15). Direct coupling reduces the heat requirements, at best, to approximately two-thirds of the original amount.

In Figure 57, Equation (7.17) is applied to the mixture 2,2-dimethylbutane – 2-methylpentane – hexane. The heat requirement is presented in parametric form in terms of the feed composition. The values for $\dot{Q}_{\min}/(\dot{F}\cdot r)$ are in the range of 3.1 to 4.4, i.e., substantially lower than the values obtained without direct coupling in Figure 56 C. The lowest heat requirements are obtained in the intermediate concentration range in which $\dot{Q}_{2,\min}$ and $\dot{Q}_{3,\min}$ are almost equal; hence, omission of the smaller value results in the largest energy saving.

As a result of the direct coupling of columns C-2 and C-3, only two columns are required for path a – c. Thus, equipment requirements for all three paths are roughly the same. With regard to the heat requirement, however, path a – c is more efficient over a wider concentration range than the other two paths. The shaded part of Figure 57 represents the optimal region. Comparison with Figure 56 C shows clearly that direct coupling is particularly useful for increasing the optimal area of this path.

Another variation in the direct coupling of columns for path a – c is shown in Figure 58 A. However, the heat requirement is not lowered because the steam for column C-1 must also be produced in column $\overline{\text{C-2}}$. Nevertheless, omission of two heat exchangers reduces the equipment requirement.

The column construction shown in Figure 58 B is still more efficient. Here, columns C-1 and $\overline{\text{C-2}}$ are combined to give a single column in which the middle part is divided by a vertical wall [57]. This construction can be applied easily to packed columns and involves lower investment costs. However, its disadvantage is that all the heat is added at the highest temperature and removed at the lowest. In addition, the different separations on the right and left sides of the partition must be carried out at the same packing depth. Thus, the amounts of liquid in both parts of the column must be controlled carefully.

Direct coupling of columns can also be employed in the separation of quaternary mixtures. Of the 21 different possible separation sequences, Figure 59 shows the path in which the overhead product and the bottom product from each column are free of the high boiler and the low boiler, respectively. If each column is operated at the lowest possible reflux ratio, the preferred separation according to Section 4.3.2.1 is obtained. The amounts and compositions of all material streams are clearly fixed. If a liquid feed is introduced into column C-2, the composition of the bottom product of C-2 is identical with the overhead product of C-3 (for a vapor

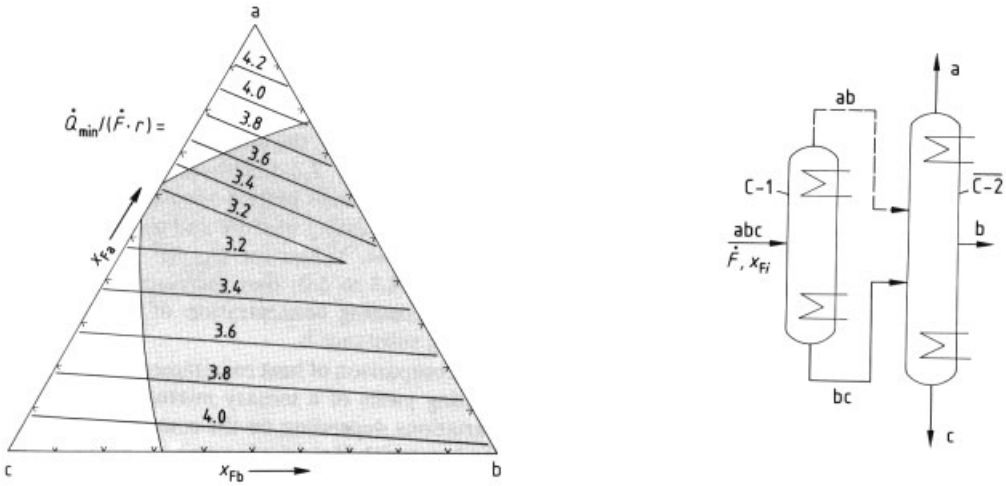

Figure 57. Heat requirement $\dot{Q}/(\dot{F}\cdot r)$ as a function of the feed composition for the path a – c with direct column coupling in the separation of a ternary mixture of 2,2-dimethylbutane (a), 2-methylpentane (b), and hexane (c)
The shaded portion represents the concentration range in which this path is energetically more favorable than paths a and c (cf. Fig. 56).

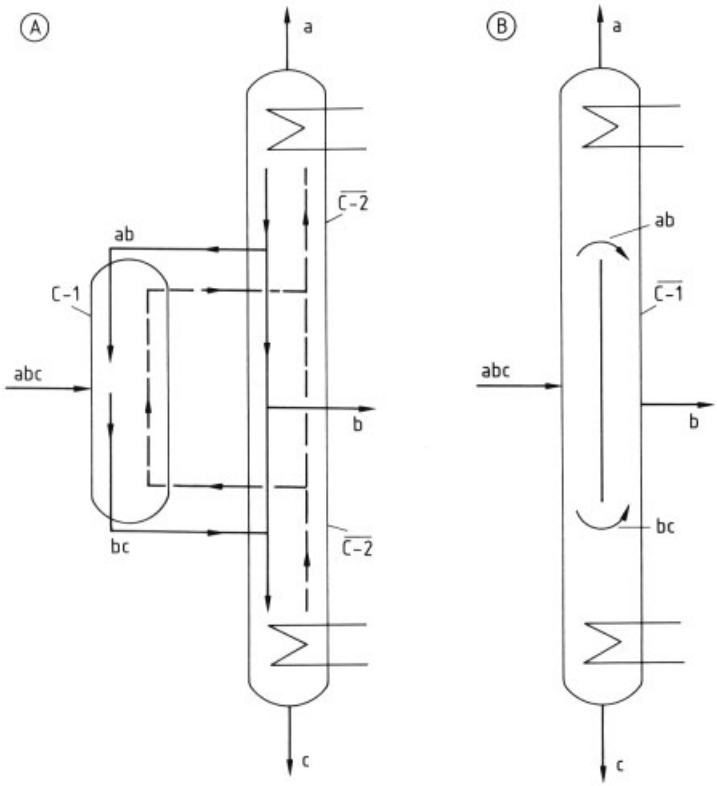

Figure 58. Reduction of the equipment requirement for separation of a ternary mixture by direct column coupling
A) Omission of the heat exchanger in the first column; B) Combination of two columns to give a single column in which the middle part is divided by a vertical wall

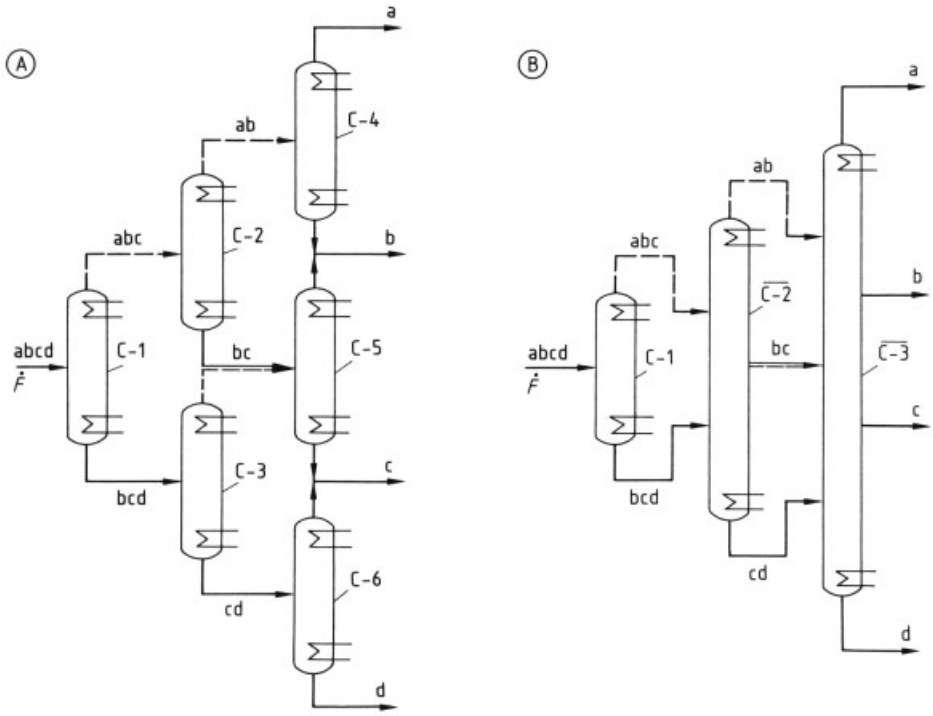

Figure 59. Path a – d in the separation of a quaternary mixture
A) Without direct coupling; B) With direct coupling

feed the products are in equilibrium). Thus, the product streams can be combined and fed into column C-5. Six columns are required for complete separation into pure substances.

Direct coupling can again be carried out between columns C-2 and C-3, C-4 and C-5, and C-5 and C-6 (Fig. 59 B). Hence, only three columns are required for the process. The heat requirement is

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r}=\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}+\mathrm{Max}\left[\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}\,\mathrm{or}\,\frac{\dot{Q}_{3,\min}}{\dot{F}\cdot r}\right]+$$
$$+\mathrm{Max}\left[\frac{\dot{Q}_{4,\min}}{\dot{F}\cdot r},\frac{\dot{Q}_{5,\min}}{\dot{F}\cdot r}\,\mathrm{or}\,\frac{\dot{Q}_{6,\min}}{\dot{F}\cdot r}\right] \qquad (7.18)$$

Direct coupling reduces the heat requirement of a four-component mixture by as much as one-half of the original value.

Another example of the direct coupling of two columns is the separation of an acetone – water – 1-butanol mixture by using the separator shown in Figure 45. The vapor at the bottom of column C-1 has approximately the same concentration as the vapor at the top of column C-2. Hence, the two columns can be combined [58], as shown in Figure 60. A condenser and a reboiler can be omitted, and the heat requirement of the new column $\overline{\text{C-1}}$ corresonds to the value for either column C-1 or column C-2, whichever is larger.

The examples show that direct coupling is particularly useful and always leads to substantial reduction of the heat requirement. It also simplifies equipment because considerably fewer columns and heat exchangers are required. This, in turn, lowers investment costs.

7.3.2. Indirect Coupling of Columns

In indirect (thermal) coupling, waste heat from one column is used to heat another column. Generally, the columns must be operated at different pressures to obtain the necessary temperature gradient. Because the vapor pressure of liquids increases exponentially with temperature, relatively large changes in pressure are needed to produce a significant temperature change.

Hence, indirect coupling is of particular interest when the feed contains close-boiling species.

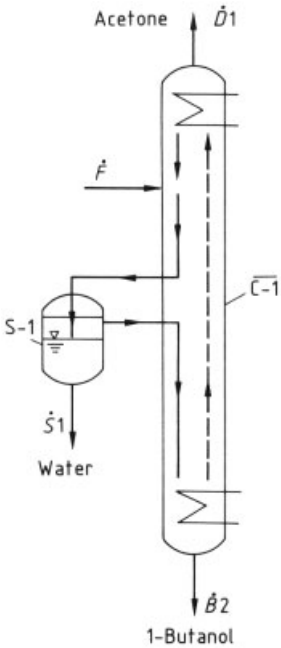

Figure 60. Direct column coupling in the separation of acetone – water – 1-butanol
For schematic without direct coupling, see Figure 45 A.

7.3.2.1. Binary Mixtures

If a process contains several columns, indirect coupling should always be considered. Sometimes separation in two thermally coupled columns is preferable to a single column. Figure 61 A shows two columns operating in parallel: each completely separates the binary starting mixture. However, the pressure p_2 of column C-2 is much lower than the pressure p_1 of column C-1. Hence, the overhead temperature of column C-1 is higher that the bottom temperature of column C-2, so waste heat from column C-1 can be utilized to heat column C-2. Thus, the heat requirements of rectification can be halved. The feed, however, must be compressed to the higher pressure p_2.

A variation of this process permits indirect coupling of the two columns at a lower pressure difference (Fig. 61 B). Pure substance b is taken from the bottom of column C-1, and an a-rich fraction containing a significant amount of b is taken from the top of the column. The vapor is used to heat column C-2. The condensate thus obtained is partially returned to column C-1 as reflux and partially fed into column C-2, where it is completely separated into pure substances a and b.

In Figure 61 C, column C-1 functions as a prefractionator from which an a-rich overhead fraction and a b-rich bottom fraction are fed to column C-2. Separation takes place in C-2. The heat requirement of column C-2 is lowered by this double feeding. Indirect coupling of the columns permits prefractionation without additional heat requirement.

Modifications of the process in Figure 61 are employed in industry. For instance, the air-separation process described in Figure 41 is comparable to the process shown in Figure 61 C. Exceptional features of air separation also permit omission of the reboiler in column C-1 and the condenser in column C-2.

Requirements for the indirect coupling of columns are always fulfilled in the separation of azeotropic mixtures by pressure changes. Figure 62 shows indirect coupling in the process used to separate a mixture of water and tetrahydrofuran. Separation of this mixture in two separate columns without indirect coupling is discussed with Figure 46; the temperature at the top of column C-2 is 136 °C, but the temperature at the bottom of column C-1 is only 100 °C. Therefore, indirect coupling is possible, and the heat requirement is reduced to ca. 65 % of the original value [59]. Equipment requirements are also substantially lower.

7.3.2.2. Ternary Mixtures

Separation of ternary mixtures requires at least two rectification columns; therefore, indirect column coupling is possible.

Indirect coupling of the two columns of path a is shown in Figure 63 A. Column C-2 operates at a higher pressure, so its waste heat can be used to heat column C-1. Indirect coupling reduces the heat requirement of the entire process to a value equivalent to the larger value of the two original columns. Thus,

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r} = \mathrm{Max}\left[\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}\,\mathrm{or}\,\frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}\right] \tag{7.19}$$

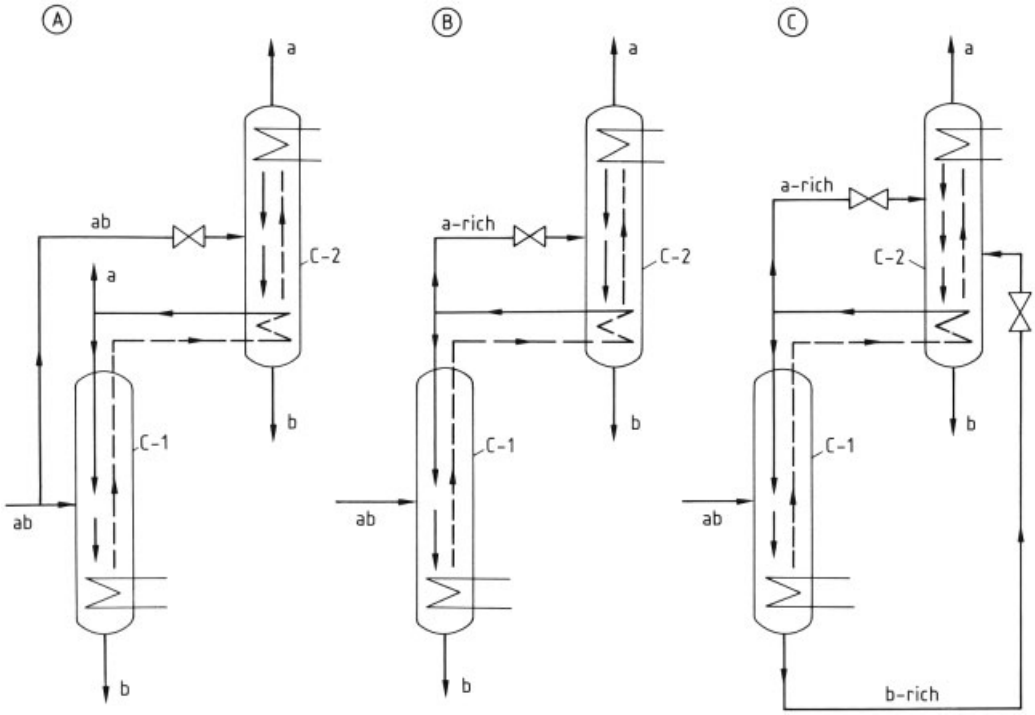

Figure 61. Indirect (thermal) column coupling in the separation of binary mixtures
A) Two columns operated in parallel with complete separation in each column; B) Separation of the a-rich overhead product from C-1 in C-2; C) Prefractionation to give enriched products in C-1 followed by complete separation in C-2. Colum C-1 operates at a lower pressure (p_1) than column C-2 (p_2).

The quantities $\dot{Q}_{1,\min}$ and $\dot{Q}_{2,\min}$ can be calculated by using Equations (7.5) and (7.6). Indirect coupling reduces the heat requirement of the process by as much as a half.

Figure 63 A shows the values calculated for $\dot{Q}_{\min}/(\dot{F} \cdot r)$ from Equation (7.19) for the mixture 2,2-dimethylbutane – 2-methylpentane – hexane. The heat requirement is plotted in parametric form against the feed stream $\dot{F}$ and is in the range of 1.7 – 4.0, i.e., considerably lower than the values obtained without indirect coupling (Fig. 56 A). In the intermediate concentration range, the heat required by each of the two columns is about the same; therefore, omission of one quantity of heat results in the largest energy saving. Here, indirect coupling halves the heat requirement.

Analogously, the indirectly coupled columns used in path c are shown in Figure 63 B. The liquid overhead product of column C-1, which operates at a higher pressure, is introduced into column C-2. The heat requirement of the entire process is

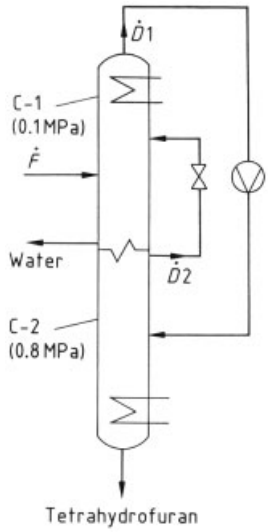

Figure 62. Indirect (thermal) column coupling to separate an azeotropic mixture of water and tetrahydrofuran by changing system pressure
For schematic without indirect coupling, see Figure 46 A.

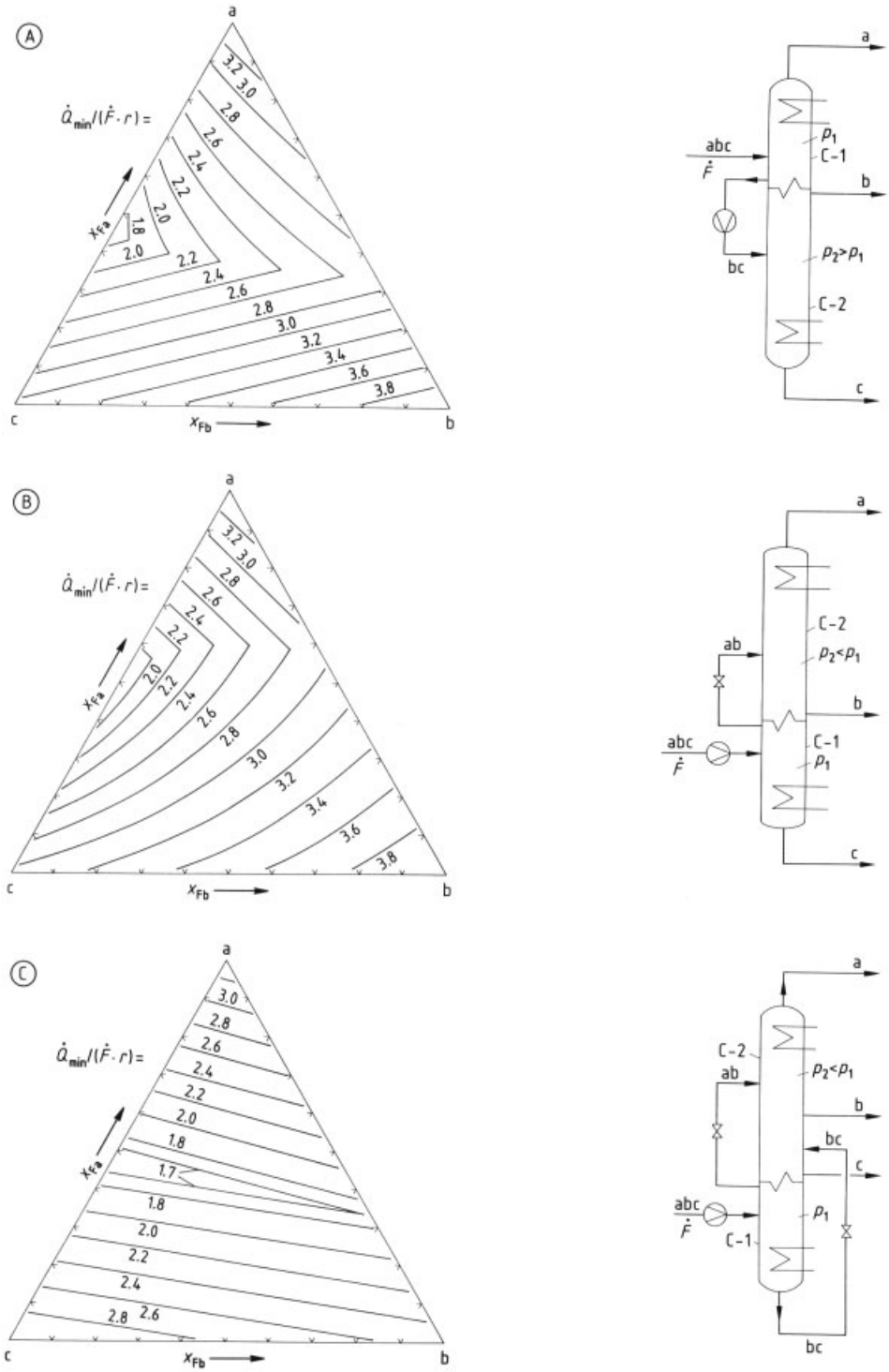

Figure 63. Minimum heat requirement $\dot{Q}/(\dot{F}\cdot r)$ as a function of feed composition for separation paths with indirect coupling in the separation of a ternary mixture of 2,2-dimethylbutane (a), 2-methylpentane (b), and hexane (c)
A) Path a; B) Path c; C) Path a – b (direct and indirect coupling)
Relative volatility: $\alpha_{ac} = 1.887$; $\alpha_{bc} = 1.329$.

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r} = \text{Max}\left[\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r} \text{ or } \frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r}\right] \tag{7.20}$$

The individual quantities of heat are calculated from Equations (7.8) and (7.9).

The results of Equation (7.20) are shown on the triangular diagram in Figure 63 B. The parametric representation clearly shows the transition between the two quantities of heat. Energy savings are at a maximum in the transition range. However, a 50 % reduction is not achieved because the feeding of vapor into column C-2 represents a coupling (see Fig. 56 B), which is, however, less effective than that shown in Figure 63 B.

The indirect coupling of path a – c is presented in Figure 63 C. Here, too, overhead product from the high-pressure column C-1 is fed into column C-2 in the liquid state. The minimum heat requirement of the process is

$$\frac{\dot{Q}_{\min}}{\dot{F}\cdot r} = \text{Max}\left[\frac{\dot{Q}_{1,\min}}{\dot{F}\cdot r}, \frac{\dot{Q}_{2,\min}}{\dot{F}\cdot r} \text{ or } \frac{\dot{Q}_{3,\min}}{\dot{F}\cdot r}\right] \tag{7.21}$$

The individual quantities of heat are calculated from Equations (7.12), (7.13), and (7.15).

The results of Equation (7.21) for the mixture 2,2-dimethylbutane – 2-methylpentane – hexane are presented in parametric form on the triangular diagram shown in Figure 63 C. A series of parallel lines are obtained in each region. The value $\dot{Q}_{1,\min}$ is always smaller than $\dot{Q}_{2,\min}$ or $\dot{Q}_{3,\min}$ so that only two regions are obtained on the triangular diagram.

Values for $\dot{Q}_{\min}/(\dot{F}\cdot r)$ range from 1.6 to 3.3 and are lower than those obtained for paths a and c (Figs. 63 A and 63 B). Only if the mixture has a very low content of medium boiler b are the three processes shown in Figure 63 equivalent. In all other cases, path a – c with direct and indirect coupling is the most favorable sequence.

7.4. Design of Heat-Exchanger Networks

In thermal processes, different quantities of heat must generally be added and removed at many points and at different temperatures. Optimal coupling of the heat sources and heat sinks of the process allows minimization of external heat transfer to and from the process. A temperature – heat flow diagram showing the values, temperatures, and signs (+ or −) of all heat quantities is extremely useful; it indicates clearly if thermal coupling is possible [60].

Figure 64 A represents the separation of a ternary mixture by using path a – c with ten heat exchangers (E-1 – E-10). The signs and temperatures of the quantities of heat for all exchangers are presented on the temperature – heat flow diagram (Fig. 64 B). The heat exchanger E-1 can be coupled thermally only with E-8, E-9, or E-10 because only there is the temperature at which the heat is supplied higher than that at which it is removed. However, the quantities of heat converted are relatively small; thus, almost the entire amount of heat (146 MJ/kmol) must be added from an external heat source (heating steam) and removed by an external heat source (cooling water).

The same path, but with direct and indirect coupling, is shown in Figure 65. The feed is heated in heat exchanger E-10 by heat transfer from the separated products. It is subsequently brought to boiling with heating steam in exchanger E-1. External heat is added to the column via E-2 and removed via E-5 at a much lower temperature. The heat of condensation from column C-1 is transferred to column C-2 via intermediate heat exchangers E-3 and E-6. This is possible only if the temperature of E-3 is higher than that of E-6, which is guaranteed by increasing the operating pressure of column C-1 to ca. 0.25 MPa.

Comparison of Figures 64 B and 65 B shows that the total amount of heat converted is higher in the process with direct and indirect coupling. However, only about one-third of the heat must be supplied from an external source.

The design of heat-exchanger systems is of considerable importance in many processes and has been perfected [61].

8. Design and Dimensioning of Mass-Transfer Equipment

8.1. Types of Construction

The basic requirements for mass-transfer are given by

$$\dot{N} = K_{OG}\cdot A\cdot (y^* - y) \tag{8.1}$$

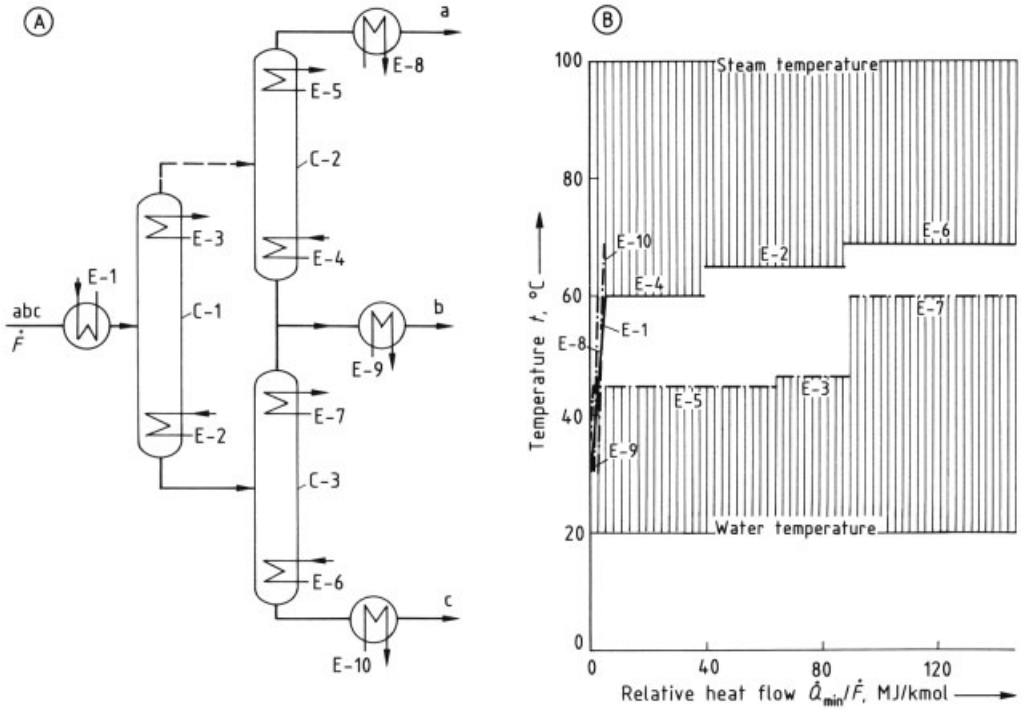

Figure 64. Heat flow in the separation of a mixture of 2,2-dimethylbutane (a), 2-methylpentane (b), and hexane (c) by using path a – c without direct or indirect coupling
A) Flow diagram showing columns (C-1, C-2) and heat exchangers (E-1 – E-10); B) Temperature – heat flow ($T - \dot{Q}$) diagram: the shaded areas represent the driving force, i.e., temperature difference for heat transfer.

The maximum mass-transfer rate $\dot{N}$ is obtained when all three terms of the above equation are as large as possible:

The mass-transfer coefficient K_{OG} depends on the flow condition of the liquid and vapor phases; it increases proportionally to the relative velocity. The situation is improved if new fluid elements constantly reach the phase boundary and produce a nonsteady mass transfer with short transfer times. Columns that allow constant regeneration of the contact area are particularly effective.

The contact area A for mass transfer depends on the geometry of the apparatus and the local velocities of the two phases. It is strongly affected by the column internals.

The largest driving force ($y* - y$) for mass transfer between phases is achieved when the overall flow pattern at all points in the apparatus allows countercurrent contact between equal streams of vapor and liquid without remixing. Thus, both phases must be distributed uniformly over the entire flow area.

Plate columns and *packed columns* are most often used for rectification; they guarantee excellent countercurrent flow and permit a large overall height. The internals provide a large mass-transfer contact area, which is constantly regenerated, especially in plate columns.

The use of mass-transfer machines has been proposed repeatedly for rectification. Although such devices have proved useful for extraction, they have not been accepted in rectification [62].

8.1.1. Plate Columns

Figure 66 shows the most important features of a plate column. The vapor moves up the column through perforations in the horizontal plates, and the condensed liquid flows downward as a countercurrent. However, the two phases on the individual plates exhibit cross flow relative to each other. The ascending vapor passes through

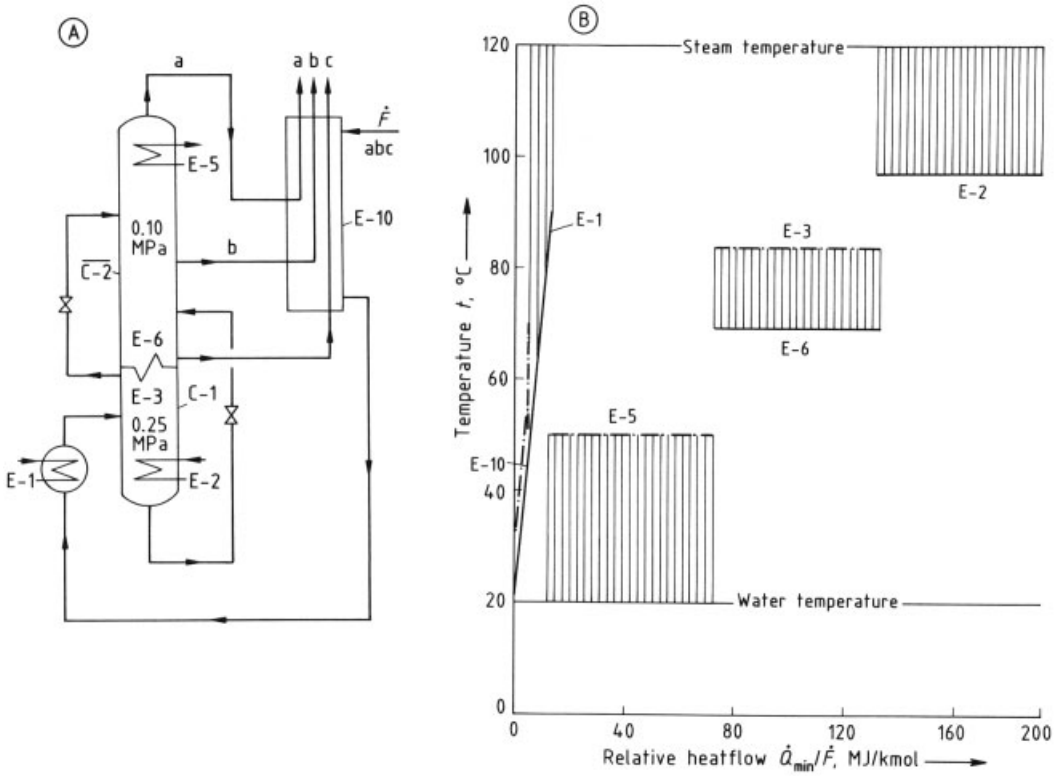

Figure 65. Heat flow in the separation of a mixture of 2,2-dimethylbutane (a), 2-methylpentane (b), and hexane (c) with direct and indirect coupling
A) Flow diagram showing columns (C-1, C-2) and heat exchangers (E-1 – E-3, E-5, E-6, E-10); B) Temperature – heat flow ($T - \dot{Q}$) diagram

the perforations into the liquid flowing across the plate, to form a large two-phase interface. The two phases subsequently separate; the vapor rises to the plate above, whereas the liquid flows over the exit weir into the downcomer, which is usually segmental in shape. Some of the liquid accumulates in each downcomer; this compensates for the pressure drop caused by the vapor as it passes through the tray. The liquid then flows through an inlet onto the plate below. The orifices in the tray floor can have various designs.

Bubble-cap plates were used almost exclusively for many years. Their construction prevents liquid from leaking downward (weeping) through the plate. However, complex bubble caps are relatively expensive and have a higher pressure drop than other designs.

Sieve plates have become increasingly important. They are inexpensive, have a high separation efficiency, and produce a low pressure drop across the plate. Previously, sieve plates were used with holes of ca. 1-mm diameter, which were easily blocked by dirt particles. However, hole diameters of 12 and sometimes 25 mm are now employed.

Valve plates are a relatively new, but widely used, development. The cross section of the orifices in the plate is adjusted to prevent the liquid from leaking through the plate at low vapor loads and to avoid excessive pressure increase at high vapor loads.

Many other variations of the sieve, valve, and bubble-cap plates are also available, e.g., Performkontakt [63].

8.1.2. Packed Columns

Figure 67 shows the most important features of a packed column. Vapor enters at the bottom of the column and flows upward through the

free cross-sectional area of the internals. Liquid flows downward countercurrently to the ascending vapor. Thus, a packed column provides countercurrent contacting of the two phases, which is crucial for good mass transfer.

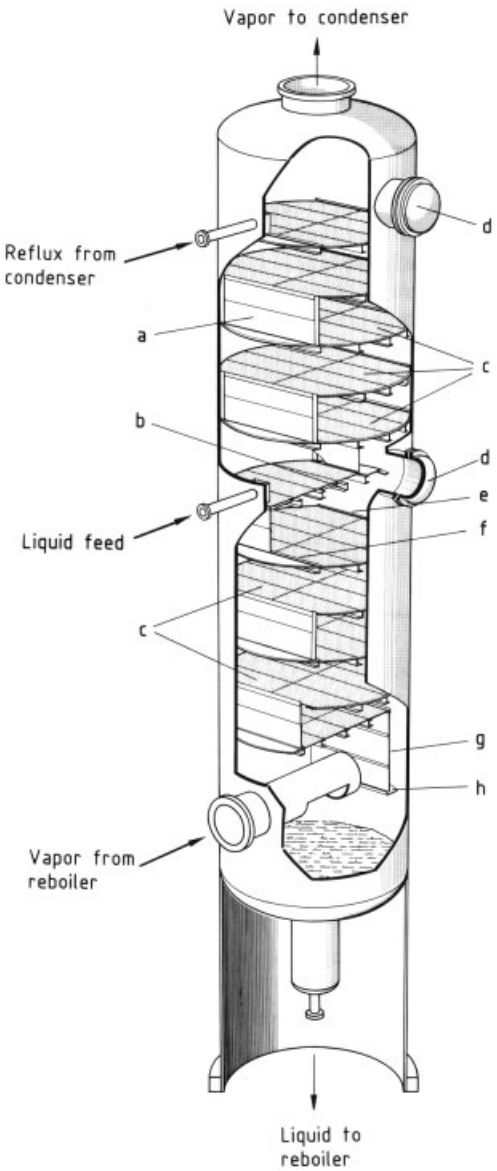

Figure 66. Cutaway section of a plate column
a) Downcomer; b) Tray support; c) Sieve trays; d) Manway; e) Outlet weir; f) Inlet weir; g) Side wall of downcomer; h) Liquid seal

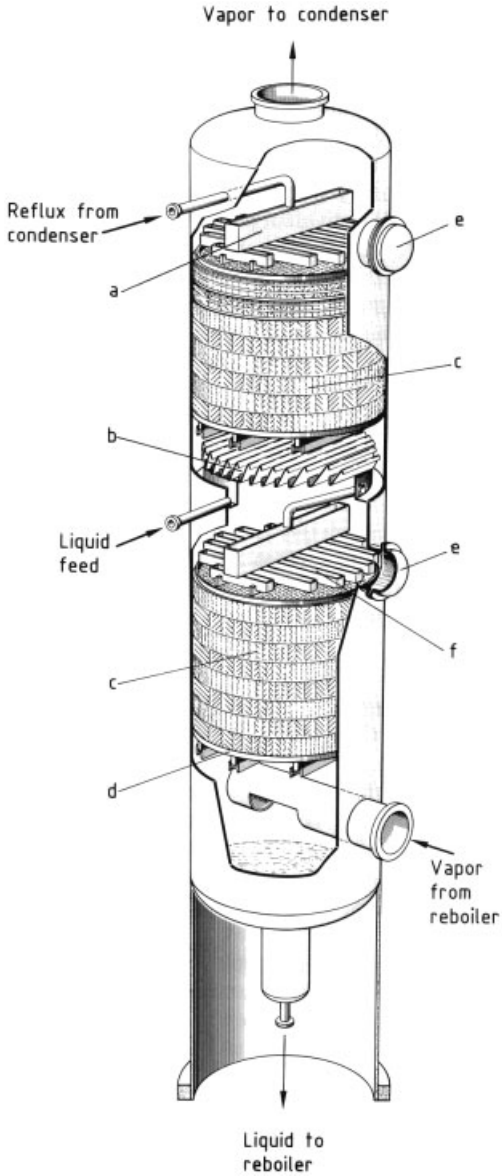

Figure 67. Cutaway section of a packed column with a structured wire gauze packing (Sulzer) [64]
a) Liquid distributor; b) Liquid collector; c) Structured packing; d) Support grid; e) Manway; f) Liquid redistributor

The liquid should wet the column packing (random or structured) as completely as possible to maximize the mass-transfer contact area between phases. The type of packing greatly influences the efficiency of the column. *Random dumped packings* made of specially formed particles (rings, saddles, cf. Fig. 80) are often used: their surface area should be as large as possi-

ble; they should have a low flow resistance; and they should be distributed homogeneously in the column (no large cavities). *Structured packings* comply more readily with these requirements and are used widely.

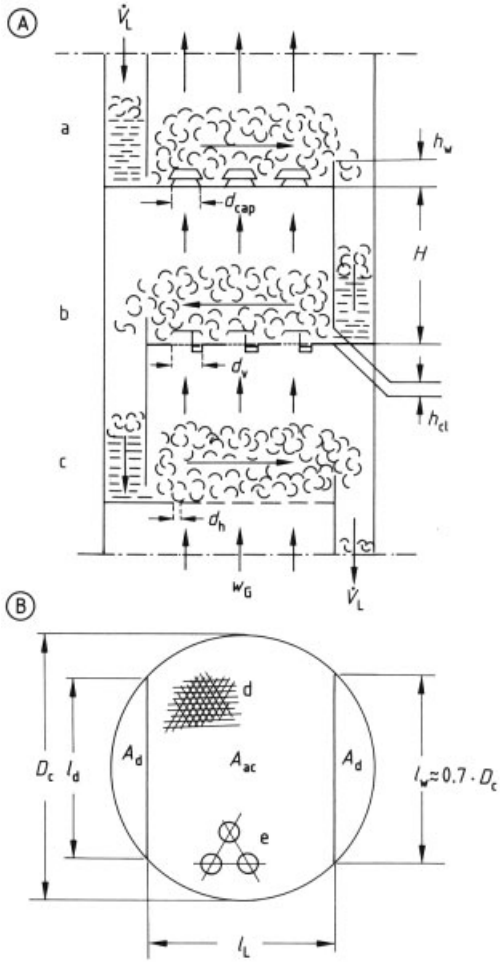

Figure 68. Schematic of a plate column showing important parameters
A) Vertical section
a) Bubble-cap plate; b) Valve plate; c) Sieve plate
d_{cap} = cap diameter; d_v = valve diameter; d_h = hole diameter; h_{cl} = height of skirt clearance of downcomer; h_w = weir height; H = plate (tray) spacing; $\dot{V}_L$ = volumetric flow rate of liquid
B) Cross section
d) Small holes with narrow spacing; e) Large holes with wide spacing
A_d = cross-sectional area of downcomer; A_{ac} = active cross-sectional area; D_c = column diameter; l_{cl} = skirt clearance length of downcomer; l_L = length of liquid path; l_w = weir length

Uniform distribution of liquid over the entire cross-sectional area is very important. In Figure 67, a cascade distributor is used for this purpose. Although the initial liquid distribution is uniform, liquid maldistribution usually occurs as the height of the packing increases; this can greatly impair separation efficiency. Therefore, the height of individual packed beds is limited to 3 – 6 m. Liquid leaving the first packing bed is collected and led into an annular channel, where it is thoroughly mixed. It is then redistributed by a second cascade distributor. Smaller columns generally use side deflectors to redistribute the liquid.

The internals of the column must offer minimum resistance to vapor flow. Modern packings have a relative free cross-sectional area of more than 90 %. The liquid collector and distributor should also reach this value. The support grating for the packing is often a barrier to flow, especially when its free cross-sectional area is blocked by the lowest particle layer.

8.1.3. Criteria for Use of Plate or Packed Columns

Proper choice of equipment is very important for effective and economical separation. An unambiguous choice is not always possible because both plate and packed columns can be modified to fulfill different requirements.

Plate columns are generally employed when large-diameter (> 1 m) towers are required; several downcomers are often necessary. The vapor load in plate columns must be kept within a relatively narrow range; only valve plates allow greater operational flexibility. The liquid load, however, can be varied over a very wide range.

Plate columns can also be operated efficiently at very low liquid loads, e.g., in vacuum operation. However, the relatively high pressure drop involved is a disadvantage, especially when the separation requires a large number of plates. Plate columns have a relatively large liquid holdup in both the two-phase layer on the plate and the downcomer. This compensates for fluctuations in feed composition. However, the high holdup results in a high liquid residence time in the column, which may lead to the decomposition of thermally unstable substances. The

holdup is also significant when a chemical reaction occurs during rectification. Special cap designs allow the liquid on the tray to be retained for short shutdown periods so that column operation can be resumed quickly. A further advantage of plate columns is that heat-exchanger coils incorporated above the plate allow heat to be added or removed easily. Plate columns are also relatively insensitive to impurities in the liquid.

Packed columns are used almost exclusively when small-diameter (e.g., < 0.7 m) towers are required. However, the development of structured packings permits the use of packed columns with very large diameters.

Packed columns are extremely flexible as far as vapor load is concerned, but they require a minimum liquid load, which often excludes their use in vacuum operation. Newly developed structured packings, however, have brought substantial progress to this area too.

The small pressure loss in packed columns, in comparison with plate columns, is a tremendous advantage. The danger of decomposition of thermally unstable substances is also less in packed columns because liquid holdup is very small.

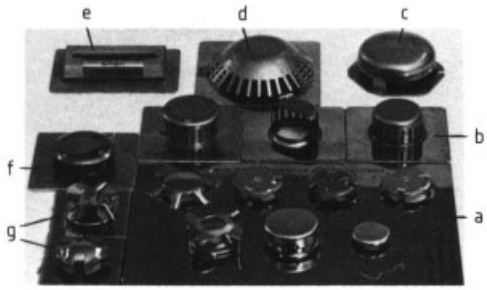

Figure 69. Typical bubble-caps and valves
a) Glitsch valves; b) Bayer bubble-cap; c) Linde bubble-cap; d) ACV bubble-cap; e) Montz KSG bubble-cap; f) Stahl valve; g) Koch valves

Packed columns are not suitable for the treatment of liquids that contain particulate contaminants or tend to crystallize. They are, however, less sensitive to foaming mixtures than plate columns. Ceramic random packings are resistant to corrosion.

8.2. Dimensioning of Plate Columns

Schematic cross sections of bubble-cap, valve, and sieve plates together with their most important dimensions are shown in Figure 68. Typical examples of bubble caps and valves are shown in Figure 69. Standard geometrical dimensions, developed on the basis of practical experience, are given in Table 1.

8.2.1. Operating Region of Plate Columns

Careful column dimensioning requires determination of the entire operating region and the design parameters so that a safety margin can be maintained with respect to the limits for liquid and vapor loads.

The operating region is represented in Figure 70, where the vapor load $\dot{V}_G$ is plotted against the liquid load $\dot{V}_L$ [65]; loads are expressed relative to the active plate area A_{ac}. Several factors limit the flow rate. The upper limits are absolute and cannot be exceeded. However, the lower limits are not absolute; although the column can be operated below these limits, separation efficiency deteriorates.

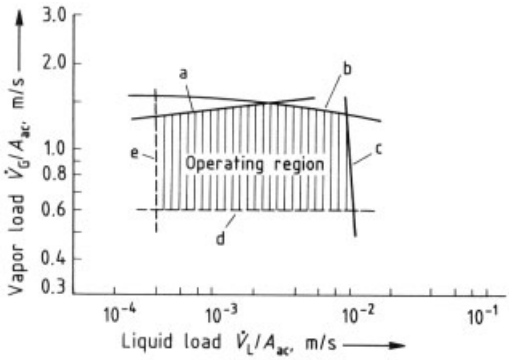

Figure 70. Operating region of a plate column
Operating boundaries due to: a) Entrainment; b) Froth height; c) Downcomer capacity; d) Weeping; e) Minimum crest over weir
A_{ac} = active plate area; $\dot{V}_G$ = volumetric flow rate of vapor; $\dot{V}_L$ = volumetric flow rate of liquid

Maximum Vapor Load. The vapor load in a plate column can be increased to the point where the vapor blows the liquid off the plate in droplets. The liquid then no longer moves countercurrently to the vapor, and column operation stops.

Table 1. Standard values for the construction of plate columns

	Vacuum rectification	Rectification at atmospheric pressure	High-pressure rectification/ absorption
Plate spacing H, m	0.4 – 0.6	0.4 – 0.6	0.3 – 0.4
Weir length l_w, m	(0.5 – 0.6) $D*_c$	(0.6 – 0.75) D_c	0.85 D_c
Weir height h_w, m	0.02 – 0.03	0.03 – 0.07	0.04 – 0.1
Skirt clearance height of downcomer h_{cl}, m	0.7 h_w	0.8 h_w	0.9 h_w
Bubble-cap diameter d_{cap}, m	0.08 – 0.15	0.08 – 0.15	0.08 – 0.15
Bubble-cap spacing t_{cap}, m	1.25 d_{cap}	(1.25 – 1.4) d_{cap}	1.5 d_{cap}
Valve diameter d_v, m	0.04 – 0.05	0.04 – 0.05	0.04 – 0.05
Valve spacing t_v, m	1.5 d_v	(1.7 – 2.2) d_v	(2 – 3) d_v
Hole diameter d_h, m	0.004 – 0.013	0.004 – 0.013	0.004 – 0.013
Hole spacing t_h, m	(2.5 – 3) d_h	(3 – 4) d_h	(3.5 – 4.5) d_h
Relative free hole area φ, %	15 – 10	10 – 6	7.5 – 4.5

* D_c = column diameter.

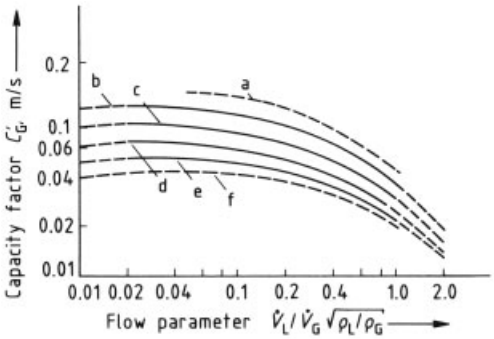

Figure 71. Correlation of the capacity factor for sieve plates according to FAIR [69]
Plate spacing H: a) 0.91 m; b) 0.61 m; c) 0.46 m; d) 0.30 m; e) 0.23 m; f) 0.15 m
Values for C'_G are shown for weir height $h_w \leq 0.15$ H, hole diameter $d_h \leq 6.35$ mm, and liquid surface tension = 0.02 N/m. The capacity factor is calculated from the chart value C'_G with the formula: $C_G = C'_G \left(\frac{\sigma}{0.02}\right)^{0.2}$
Corrections must be made for the relative free hole area φ as follows: $\varphi \geq 0.101$, $C_G = C'_G$
$\varphi = 0.08$, $C_G = 0.9 \cdot C'_G$
$\varphi = 0.06$, $C_G = 0.8 \cdot C'_G$

The maximum possible vapor load is calculated by using the Souders and Brown [66] equation:

$$w_G = C_G \sqrt{(\varrho_L - \varrho_G)/\varrho_G} \tag{8.2}$$

where w_G is the superficial vapor velocity, C_G is the capacity factor, ϱ_L is the density of the liquid, and ϱ_G is the density of the vapor. If the F factor, defined by $F = w_G \cdot \sqrt{\varrho_G}$, is used as a measure of vapor load, the above equation becomes

$$F = C_G \sqrt{\varrho_L - \varrho_G} \tag{8.3}$$

The capacity factor C_G must be determined experimentally. Several correlations are described in [67], [68]. The most frequently used correlation is that of FAIR, who found that C_G is proportional to $\sigma^{0.2}$; it is presented in Figure 71 [69].

A theoretical relationship assumes that the vapor on the plate forms the continuous phase while the liquid is dispersed into individual droplets [70]. The diameter d of the drop is expressed in terms of the critical Weber number We_{cr}, which is computed by using the vapor velocity through the orifices (holes, slots) in the plate.

$$d = We_{cr} \cdot \sigma / F_h^2 = We_{cr} \cdot \sigma \cdot \varphi^2 / F^2 \tag{8.4}$$

where σ is the surface tension of the liquid, φ is the relative free hole area, F is the vapor loading factor defined in Equation (8.3), and $F_h = F/\varphi$.

A balance of forces between the resistance to flow and the weight minus the buoyancy is obtained for the diameter of the drops that are suspended in the vapor stream above the plate:

$$d = \frac{3}{4} \frac{\zeta_{sp}}{(\varrho_L - \varrho_G) \cdot g} \cdot F^2 \tag{8.5}$$

where ζ_{sp} is the frictional coefficient of a sphere and g is the acceleration due to gravity. Eliminating the drop diameter from both equations yields

$$F_{max} = \left(\frac{4 \cdot We_{cr}}{3 \cdot \zeta_{sp}}\right)^{0.25} \cdot \left[\varphi^2 \cdot \sigma \cdot (\varrho_L - \varrho_G) \cdot g\right]^{0.25} \tag{8.6}$$

According to [71], the critical Weber number is approximately 12. If the frictional factor ζ_{sp} is assumed to be 0.4 (the value obtained for a sphere in a turbulent region), then

$$F_{max} = 2.5\left[\varphi^2 \cdot \sigma\left(\varrho_L - \varrho_G\right) \cdot g\right]^{0.25} \tag{8.7}$$

This purely theoretical equation corresponds not only to the measured values, but also to Fair's correlation [69]. A proportionality to $\sigma^{0.25}$ is found in this equation, whereas FAIR empirically obtained $\sigma^{0.2}$. The dependence on the relative free hole area φ also corresponds to the $\varphi^{1/2}$ relationship in Equation (8.7).

Equation (8.7) describes the process of blowing off the plate and is applicable only to plate towers with very large plate spacings. When plate spacings are smaller, the height of the two-phase layer, and not the blowing-off process, is the limiting factor that determines maximum vapor load. This height must always be less than the plate spacing for safe operation.

Minimum Vapor Load. The vapor load on the column plate can be decreased to the point where the plate is not uniformly exposed to the vapor or the liquid leaks through the plate. Both conditions are undesirable because they lower plate efficiency. According to MERSMANN [72], a sieve plate is uniformly exposed to vapor when the Weber factor exceeds the critical value 2 at the holes; thus,

$$F_h \geq \sqrt{2 \cdot \sigma / d_h} \tag{8.8}$$

where d_h is the hole diameter.

If the Froude number Fr exceeds a critical value, the liquid can be prevented from leaking through a sieve plate. According to RUFF [73],

$$F_h \geq \sqrt{0.37 \cdot d_h \cdot g \cdot \left(\varrho_L - \varrho_G\right)^{1.25} / \varrho_G^{0.25}} \tag{8.9}$$

Equations (8.8) and (8.9) are evaluated for three different systems in Figure 72. At small hole diameters up to 2 mm, uniform vapor distribution is the limiting factor, and the minimum vapor load required decreases with increasing hole diameter. At hole diameters greater than 2 – 3 mm, weeping of liquid through the plate becomes the decisive factor; the minimum vapor load then increases with increasing hole diameter.

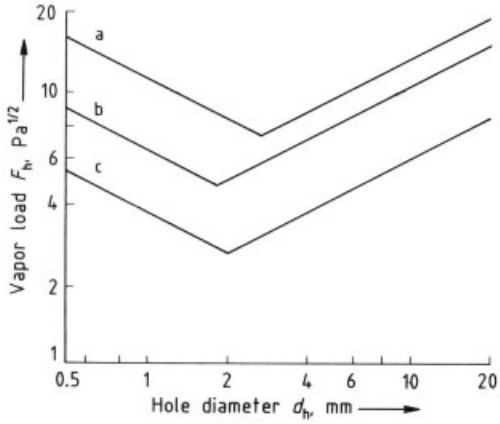

Figure 72. Minimal vapor load of sieve plates for three different systems
a) Air – water; b) Benzene – toluene; c) Isobutane – *n*-butane

Equations (8.8) and (8.9) do not take into account the influence of the liquid load on the lower vapor load limit, although large liquid loads require a somewhat higher minimum vapor load to prevent weeping of liquid through the plate. However, if weeping does occur, it does not lower separation efficiency significantly.

Maximum Liquid Load. Liquid flows down the column through the downcomers due to the influence of gravity. This driving force permits a limited liquid flow rate $\dot{V}_L$ only. The four following empirical criteria are generally used to determine the maximum liquid flow rate [74]:

1) The weir load $\dot{V}_L/l_w$ (l_w = weir length) should be less than 60 $m^3\,m^{-1}\,h^{-1}$.
2) The liquid velocity in the downcomer should not exceed ca. 0.1 m/s.
3) The liquid residence time in the downcomer should be greater than 5 s.
4) The height of the clear liquid in the downcomer should not exceed half the plate spacing H.

If the entire downcomer is filled with a vapor – liquid mixture with a relative liquid holdup $\bar{\varepsilon}_{Ld}$, the following equation is obtained [75]:

$$\frac{\dot{V}_L}{l_{cl}} = 0.61 \cdot \bar{\varepsilon}_{Ld} \cdot h_{cl} \cdot \sqrt{2 \cdot g \cdot H \cdot \left(\varrho_L - \varrho_G\right) / \varrho_L \cdot \left(1 - \frac{h_p + h_L}{\bar{\varepsilon}_{Ld} \cdot H}\right)} \tag{8.10}$$

where l_{cl} is the skirt clearance length of the downcomer, h_{cl} is the skirt clearance height, h_p is $\Delta p/\varrho_L g$, and h_L is the clear liquid height on the plate. The parameter $\bar{\varepsilon}_{Ld}$ is unknown; an approximate value of 0.4 can be employed for nonfoaming systems. Further details are given in [76].

Minimum Liquid Load. In theory, a column plate can be operated even at very small liquid loads because the height of the two-phase layer on the plate is guaranteed by the exit weir. However, at extremely low liquid loads, liquid flows unevenly across the plate, which decreases the separation efficiency (see Section 8.2.3). Accordingly, the minimum height of the weir overflow is usually set at ca. 5 mm; this corresponds to a minimum weir load of $\dot{V}_L/l_w \approx 2\,m^3\,m^{-1}\,h^{-1}$. In small-diameter columns, however, this value may be considerably lower.

Entrainment of liquid by the vapor also determines the minimum liquid load.

8.2.2. Two-Phase Flow in Plate Columns

The behavior and properties of the two-phase layer on the plates are important for the mode of action of plate columns. The two-phase layer has three basic structures:

1) *Bubble Regime.* The liquid forms the continuous phase; the vapor rises in the liquid in the form of discrete bubbles.
2) *Drop Regime.* The liquid is dispersed in the form of drops in the continuous vapor phase.
3) *Froth Regime.* This regime represents a transition from the bubble to the drop system. The vapor – liquid layer is agitated intensively, so no clearly dispersed phase is formed.

The bubble regime can exist only at very high pressures for the following reasons: At the minimum gas load, the vapor load at the holes in the plate floor should not fall below $F_h \approx 5$ (see Fig. 72); therefore, the vapor velocity at the holes is expressed by $w_{Gh} = 5/\sqrt{\varrho_G}$. However, the bubbles rise at a velocity of ca. 0.1 – 0.2 m/s. Only if the density of the vapor ϱ_G is very high (i.e., at high pressure) does the vapor velocity in the holes reach the velocity at which the bubbles rise, as is necessary in a bubble regime.

At low pressures, the vapor velocity is much higher than the velocity at which the bubbles rise. Thus, bubbles build up at the orifices, forming a region of very high vapor content, which breaks through the two-phase layer suddenly and rises upward. This jerky mode of operation is responsible for the two-phase froth regime.

The vapor load at which the transition to the drop regime (phase inversion) occurs is of special interest [75]. When the liquid level on the plate h_L is high, phase inversion occurs at $F/F_{max} \approx 0.65$. However, if h_L is very low, phase inversion can occur much earlier. Hence, column dimensioning requires a minimum vapor penetration depth of ca. 20 mm.

Plate columns are usually operated in the froth region. Operating conditions approach the drop regime only in vacuum rectification.

The *relative liquid holdup* ε_L is a very important parameter of the two-phase layer. Published values can be summarized by the following equation:

$$\varepsilon_L = 1 - (F/F_{max})^{0.28} \tag{8.11}$$

This relationship is valid only if $\varphi \cdot \varrho_L/\varrho_G > 50$. Determination of the relative liquid holdup for smaller values of this parameter is described in [75].

Expressions for the *height of the two-phase (froth) layer* h_f given in the literature are not dimensionally correct, nor do they adequately represent the limiting conditions. The following equation summarizes most of the known measured values [75]:

$$h_f = h_w + \frac{1.45}{g^{1/3}}\left(\frac{\dot{V}_L/l_w}{\varepsilon_L}\right)^{2/3} + \left(\frac{125}{g\cdot(\varrho_L - \varrho_G)}\right)\cdot \cdot\left(\frac{F - 0.2\sqrt{\varrho_G}}{1-\varepsilon_L}\right)^2 \tag{8.12}$$

where h_w is weir height, l_w weir length, and g the acceleration due to gravity. The influence of the liquid load on the height of the two-phase layer is covered primarily by the second term, which represents the overflow height of the exit weir. The last term takes into account upward spraying of liquid caused by the vapor, which occurs only in the froth and drop regimes. Therefore, it should be considered only when

$$F - 0.2\sqrt{\varrho_G} > 0.$$

Equation (8.12) applies only when vapor is allowed to pass through the liquid up to the exit weir. If the exit weir is preceded by a calming zone, vapor – liquid separation can take place, with the result that clear liquid flows over the exit weir. The height of the foaming two-phase layer is then expressed by

$$h_f = \left[h_w + \frac{1.45}{g^{1/3}} \cdot \left(\dot{V}_L / l_w \right)^{2/3} \right] / \varepsilon_L \tag{8.13}$$

Equation (8.13) usually gives a larger value for h_f than Equation (8.12).

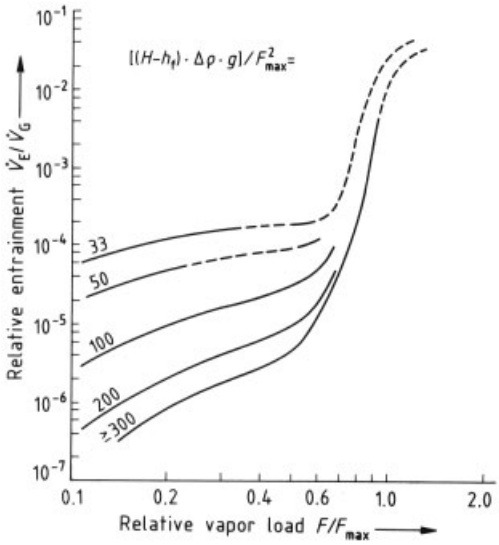

Figure 73. Dependence of entrainment on the vapor load [75]
Curves are plotted at indicated values of the parameter $(H - h_f) \cdot \Delta\varrho \cdot g/F^2_{max}$ for a hole diameter of less than 10 mm (H = plate spacing, h_f = height of two-phase froth layer, $\Delta\varrho$ = density difference, g = acceleration due to gravity).

Liquid Entrainment. Vapor flowing through the two-phase layer always entrains droplets of liquid. Entrainment is undesirable because it disturbs the countercurrent flow of vapor and liquid. In the drop regime, relative entrainment $\dot{V}_E/\dot{V}_G$ ($\dot{V}_E$ is the volume flow rate of entrained liquid) can be expressed as a function of the relative vapor load F/F_{max} in an entrainment diagram. Published data are correlated in Figure 73 [75]. This diagram applies to all three plate designs: sieve, valve, and bubble-cap plates. Entrainment depends strongly on vapor load. A tenfold increase in vapor load increases entrainment by a factor of ca. $10^4 - 10^5$. A sudden increase in entrainment occurs at a relative vapor load F/F_{max} of ca. 0.65; this is due to the change from the froth to the drop regime.

Liquid Mixing in the Two-Phase Layer. In plate columns, liquid flows across the plate to the exit weir; i.e., the liquid exhibits cross flow relative to the vapor. Mass transfer between phases causes a concentration profile in the liquid, which greatly influences plate efficiency. The two-phase layer on the plate is agitated intensely, particularly in the froth and drop regimes. This results in partial mixing of the liquid on the plate, which in turn reduces the differences in liquid concentration. According to the dispersion model, the degree of liquid mixing on the plate can be represented by the dimensionless Péclet number *Pe:*

$$Pe = \frac{l_L^2}{D_E \cdot \tau_L} \tag{8.14}$$

where l_L is the path length of the liquid, D_E is the dispersion coefficient, and τ_L is the residence time of the liquid in the two-phase layer on the plate. The liquid is totally mixed when $Pe = 0$; the opposite case (i.e., no mixing) is called plug flow and is characterized by $Pe \to \infty$.

The dispersion coefficient D_E is obtained by modification of an equation derived by Weisshuhn [77]:

$$D_E = 1.06 \cdot h_f \cdot F \left(\frac{\varepsilon_L}{\varrho_L - \varrho_G} \right)^{1/2} \tag{8.15}$$

Equation (8.15) applies only to the froth and the drop regimes. The degree of liquid mixing in the bubble regime is usually so small that it can be neglected.

Liquid Maldistribution. To achieve optimal mass-transfer efficiency, equal amounts of vapor and liquid must be brought into contact at each site on the plate. However, uniform distribution of the liquid stream on the plate area is almost impossible to achieve in practice.

Results of tests conducted on a bubble-cap plate, 2.3 m in diameter, are presented in Figure 74 [78]. The plate was operated with a warm water – air system. As a result of the favorable mass-transfer conditions in the two-phase layer,

the air is virtually saturated with water vapor, which in turn leads to a temperature drop in the liquid. The temperatures in the froth layer on a plate are relatively easy to determine experimentally and are shown as liquid isotherms in Figure 74. These isotherms can be interpreted as lines of constant liquid residence time. In the case of plug flow (no mixing), the liquid isotherms are represented by lines parallel to the liquid inlet and outlet. The characteristic deviations of the isotherms in Figure 74 from a straight line are due to the characteristic design of the plate.

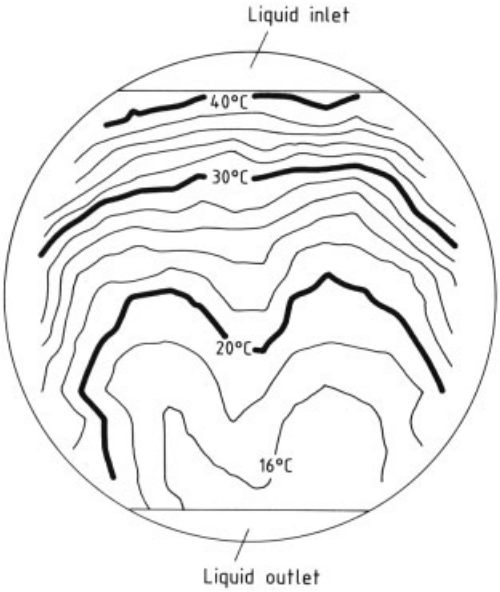

Figure 74. Liquid maldistribution on column plates represented in the form of isotherms [78]
Data were determined on a bubble-cap plate with a diameter of 2.3 m, at a vapor load F of 1.61 $\mathrm{Pa}^{1/2}$ and a relative liquid flow rate $\dot{V}_\mathrm{L}/l_\mathrm{w}$ of 10 $\mathrm{m}^3\,\mathrm{m}^{-1}\,\mathrm{h}^{-1}$.

The most important factors influencing liquid flow on the plate are the design of the inlet and outlet weirs, a highly uniform distribution of the bubble caps over the plate area, and absolutely horizontal installation of the plate [78].

Liquid – Vapor Interfacial Area. The interfacial area of the two-phase layer on the plate is very important and is conveniently expressed in a ratio to the volume of the layer. No satisfactory correlation is available for determination of the relative interfacial area a. Various procedures have yielded values that differ by a factor of 10 [75]. However, the order of magnitude of the interfacial area and its influencing factors can be estimated by using the following equation, which is valid for spherical dispersed particles [75]:

$$a = 6\,(1-\varepsilon)\,/d \tag{8.16}$$

where ε is the porosity and d is the particle diameter. Equation (8.16) can be applied to the bubble and drop regimes, but not to the froth regime. The volume fraction occupied by the dispersed phases can be computed by using Equation (8.11) for the liquid holdup ε_L. The bubble diameter d is approximately

$$d = \left[\frac{6\cdot\sigma}{(\varrho_\mathrm{L}-\varrho_\mathrm{G})\cdot \mathrm{g}}\right]^{1/2} \tag{8.17}$$

The interfacial area in the bubble regime is given by

$$a = 6\left[\frac{(\varrho_\mathrm{L}-\varrho_\mathrm{G})\cdot g}{6\cdot\sigma}\right]^{1/2}\cdot\left(F/F_\mathrm{max}\right)^{0.28} \tag{8.18}$$

Analogously, an equation for the drop regime can be obtained by using Equation (8.4) to determine the diameter of the droplet:

$$a = \frac{F^2}{2\cdot\sigma\cdot\varphi^2}\cdot\left[1-\left(F/F_\mathrm{max}\right)^{0.28}\right] \tag{8.19}$$

where φ is the relative free hole area, i.e., ratio of total hole area to active area.

Equations (8.18) (bubble regime) and (8.19) (drop regime) are evaluated for three systems in Figure 75. The froth regime cannot be calculated because a clear dispersed phase does not exist. However, the interpolated values shown in Figure 75 should approximate the real course. The interfacial area in aqueous systems is ca. 500 $\mathrm{m}^2/\mathrm{m}^3$, but in organic systems values of 1000 $\mathrm{m}^2/\mathrm{m}^3$ or more can be reached. The interfacial tension σ has a considerable effect on the area of the interface; analysis shows that

$$a \approx \sigma^{-0.57} \tag{8.20}$$

This calculation permits a rough estimation of the mass-transfer area and the most important influencing factors. The values obtained are only partially applicable to calculation of the mass-transfer efficiency of a plate column. This is because it is not known whether all parts of the interface contribute to the same extent to mass transfer between phases (cf. Section 8.2.3).

Pressure Drop of Vapor on the Plate. The pressure drop Δp of the vapor as it passes through the plate can be expressed as follows:

$$\Delta p = \Delta p_{\mathrm{d}} + \Delta p_{\mathrm{L}} + \Delta p_{\mathrm{R}} \tag{8.21}$$

The first term Δp_{d} accounts for the pressure drop through a dry plate. The second term Δp_{L} represents the pressure drop resulting from the height of liquid h_{L} on the column plate and can be formulated by

$$\Delta p_{\mathrm{L}} = h_{\mathrm{L}} \cdot \varrho_{\mathrm{L}} \cdot \mathrm{g} = h_{\mathrm{f}} \cdot \varepsilon_{\mathrm{L}} \cdot \varrho_{\mathrm{L}} \cdot \mathrm{g} \tag{8.22}$$

The last term of Equation (8.21), Δp_{R}, accounts for several residual factors including bubble formation, liquid mixing, and vertical acceleration of the liquid. This term is much smaller than the first two and can generally be neglected.

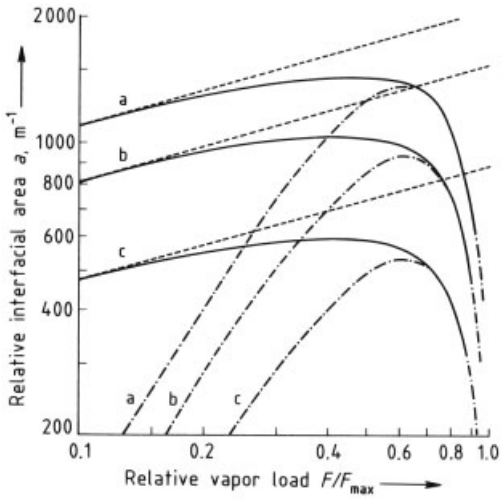

Figure 75. Calculated values for the interfacial area-to-volume ratio a of the two-phase layer on column plates
a) Isobutane – n-butane; b) Acetone – benzene; c) Air – water
- - - Bubble regime (Eq. 8.18); – · – · – Drop regime (Eq. 8.19); —— Presumed real values

Pressure Drop in Dry Sieve Plates. The pressure drop of the vapor as it passes through a dry sieve plate Δp_{d} is given by

$$\Delta p_{\mathrm{d}} = \zeta \cdot \frac{\varrho_{\mathrm{G}}}{2} \cdot w_{\mathrm{Gh}}^2 = \frac{\zeta}{2} \cdot F_{\mathrm{h}}^2 \tag{8.23}$$

A relationship for the orifice coefficient ζ can be derived from the law of conservation of energy and the law of momentum. For very thin perforated plates (i.e., when the ratio of plate thickness to hole diameter $s/d_{\mathrm{h}} \to 0$), the following formula applies:

$$\zeta = \left(\frac{1}{\alpha} - \varphi\right)^2 \tag{8.24}$$

where α is the contraction number and φ is the relative free hole area. For thick perforated plates (i.e., when $s/d_{\mathrm{h}} \gg 0$), frictional losses can be neglected and the following equation is obtained:

$$\zeta = \left(\frac{1}{\alpha} - 1\right)^2 + (1-\varphi)^2 \tag{8.25}$$

At high Reynold's numbers, a value of 0.61 is sufficiently accurate for the contraction number α.

The normal situation lies between the extreme cases described in Equations (8.24) and (8.25), and the orifice coefficient must be determined experimentally. Figure 76 shows measured values of the orifice coefficient ζ_0, i.e., for a relative free hole area $\varphi \to 0$, plotted as a function of the ratio of plate thickness to hole diameter [75]. Equation (8.24) gives a constant theoretical limit of $\zeta_0 = 2.69$ for very thin plates. A limit of $\zeta_0 = 1.41$ is obtained from Equation (8.25) for thick plates.

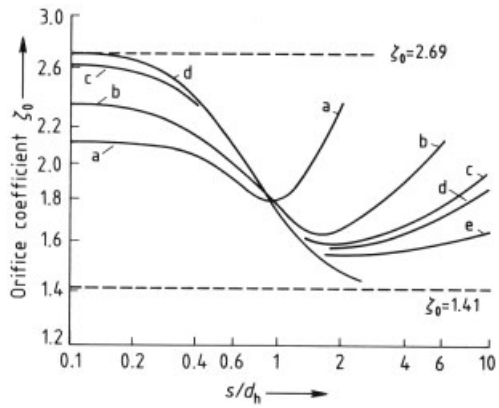

Figure 76. Dependence of the orifice coefficient of sieve trays (sharp-edged holes) on the ratio of plate thickness to hole diameter (s/d_{h})
Value of Reynold's number: a) 400; b) 1000; c) 5000; d) 10^4; e) 10^6
Dashed lines indicate theoretical limits.

The relationship between the orifice coefficient ζ and the relative free hole area φ differs for thin and thick plates. For thin plates, from Equation (8.24)

$$\zeta = \zeta_0 + \varphi^2 - 2 \cdot \varphi \sqrt{\zeta_0} \tag{8.26}$$

For thick plates, the following is obtained from Equation (8.25):

$$\zeta = \zeta_0 + \varphi^2 - 2 \cdot \varphi \tag{8.27}$$

In practice, the differences between Equations (8.26) and (8.27) are very small because φ does not exceed 0.2. Therefore, Equation (8.27) can be used with sufficient accuracy for most plate thicknesses.

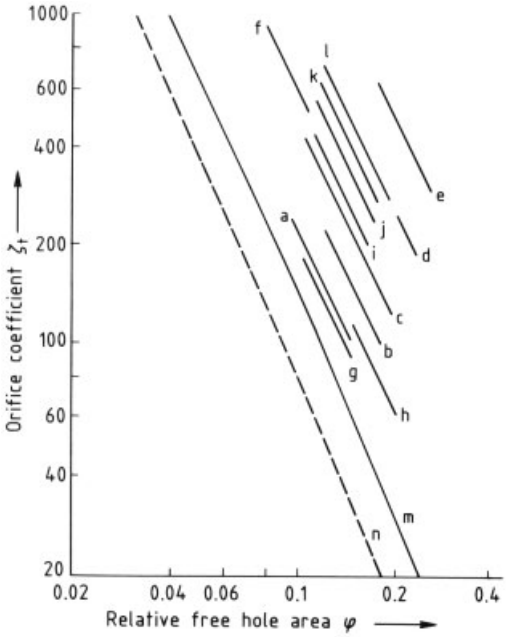

Figure 77. Dependence of the orifice coefficient of column plates on the relative free hole area φ
Bubble-cap plates: a) Bayer; b) Montz KSG; c) Linde; d) ACV; e) Thormann 1963; f) Kirschbaum 1951
Valve plates: g) Koch TK8; h) Stahl VV 12/2; i) Koch AT; j) Glitsch V-4; k) Glitsch V-1 0.25″; l) Nutter type 1972
Sieve plates: m) $\zeta_0 = 1.7$; n) $\zeta_0 = 1.0$

Pressure Drop in Dry Valve and Bubble-Cap Plates. Bubble caps and valves are so complex that prediction of the pressure drop is almost impossible. Generally, the pressure drop must be measured. Both the orifice coefficient ζ of each element and the number of elements that can be accommodated per unit area depend on the prevailing construction; therefore, comparison is made with a plate orifice coefficient ζ_t:

$$\zeta_t = 2 \cdot \Delta p_d / F^2 \tag{8.28}$$

A compilation of orifice coefficients for different types of plates is shown in Figure 77 [75]. The orifice coefficient in valve plates and bubble-cap plates is always larger than in sieve plates. However, the relative free hole area φ in valve plates can be larger than in sieve plates because the danger of liquid leaking through the plate is much less. Thus, differences between the three types of plates can be minimized.

8.2.3. Mass Transfer in the Two-Phase Layer on Column Plates

In practice, mass transfer between vapor and liquid in the two-phase layer on the plate floor is usually defined by means of an efficiency factor.

In *point efficiency* E_{OG}, the concentration change along an individual stream of vapor is considered; it is defined by

$$E_{OG} = \frac{y - y_{n-1}}{y^*(x) - y_{n-1}} \tag{8.29}$$

The subscript OG means that the overall resistance has been formally placed in the gas (vapor) phase. The actual change in concentration of vapor $(y - y_{n-1})$ as it flows through the two-phase layer is related to the maximum concentration change attainable; i.e., $y*(x) - y_{n-1}$. Here, $y*$ is the vapor concentration in equilibrium with liquid concentration x at the point on the plate where the vapor stream passes through the two-phase layer.

Point efficiency is a very useful theoretical factor but has no immediate practical importance. It can reach a maximum value of 1 provided that the liquid is completely mixed throughout the head of the two-phase layer.

In practice, *plate efficiency* E_{OGM} is used and is defined by

$$E_{OGM} = \frac{y_n - y_{n-1}}{y^*(x_n) - y_{n-1}} \tag{8.30}$$

Here, the overall vapor flow through the column is considered. The actual concentration change of the total vapor $(y_n - y_{n-1})$ is related to the maximum possible change in concentration $y*(x_n) - y_{n-1}$. Here, $y*$ is the vapor concentration in equilibrium with liquid concentration x_n leaving plate n. This definition is arbitrary because the concentration usually differs at each point on the plate, and efficiencies of more than 100 % can be obtained. Although the theoretical definition of plate efficiency E_{OGM} is

unsatisfactory, it is of practical interest because it describes the separation efficiency of a real plate.

Equations (8.29) and (8.30) define the gas transfer efficiency; analogous equations are also sometimes used for the liquid transfer efficiency. However, they have no practical importance because the main resistance to mass transfer in rectification is attributed to the vapor.

The relationship between point and plate efficiency can be formulated easily when the liquid flows across the plate without mixing (plug flow). According to [79],

$$\frac{E_{\mathrm{OGM}}}{E_{\mathrm{OG}}} = \frac{\dot{L}/\dot{G}}{m} \cdot \frac{1}{E_{\mathrm{OG}}} \left[\exp\left(\frac{m}{\dot{L}/\dot{G}} \cdot E_{\mathrm{OG}} \right) - 1 \right] \tag{8.31}$$

The vapor under each plate is assumed to be completely mixed.

As a rule, the liquid in the two-phase layer on column plates is at least partially mixed. For complete mixing,

$$E_{\mathrm{OGM}}/E_{\mathrm{OG}} = 1 \tag{8.32}$$

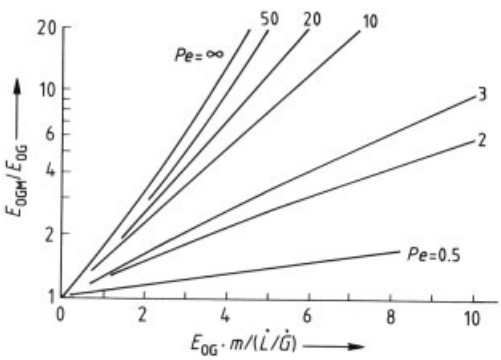

Figure 78. Correlation between the ratio of overall plate efficiency E_{OGM} to point efficiency E_{OG} and the liquid mixing represented by the Péclet number *Pe*

The normal condition lies between the two extremes described by Equations (8.31) and (8.32). The degree of liquid mixing, which is expressed in terms of a Péclet number *Pe* (see Section 8.2.2), must then be considered. A solution [80] for this is represented graphically in Figure 78. In plug flow (no mixing), $Pe = \infty$; when $Pe = 0$, complete liquid mixing takes place. When both the Péclet number and the abscissa $E_{\mathrm{OG}} \cdot m/(\dot{L}/\dot{G})$ are large, the ratio of the plate efficiency to the point efficiency $E_{\mathrm{OGM}}/E_{\mathrm{OG}}$ becomes very large. However, for the abscissa to be very large, the liquid flow $\dot{L}$ must be small. The high residence time at low flow rates causes considerable remixing on the plate so that the parameter *Pe* in Figure 78 becomes very small. In practice, plate efficiency is also affected adversely by other processes such as liquid leakage through the plate, entrainment of liquid to the next higher plate, and maldistribution. In careful column design, the plate efficiency can be adjusted so that it is only slightly higher than the point efficiency.

Point Efficiency. The equations of mass transfer must be examined before theoretically defining the point efficiency. The following relationship is valid:

$$E_{\mathrm{OG}} = 1 - \exp\left(-N_{\mathrm{OG}}\right) \tag{8.33}$$

The overall number of gas-phase transfer units N_{OG} can be calculated from the number of transfer units in the gas N_{G} and in the liquid N_{L} by using the following equation:

$$\frac{1}{N_{\mathrm{OG}}} = \frac{1}{N_{\mathrm{G}}} + \frac{m}{\dot{L}/\dot{G}} \cdot \frac{1}{N_{\mathrm{L}}} \tag{8.34}$$

where $m/(\dot{L}/\dot{G})$ is known as the stripping factor. The number of transfer units in the vapor phase is given by

$$N_{\mathrm{G}} = \frac{\beta_{\mathrm{G}} \cdot a \cdot h_{\mathrm{f}} \cdot A_{\mathrm{ac}}}{\dot{V}_{\mathrm{G}}} \tag{8.35}$$

where β_{G} is the vapor mass-transfer coefficient, a is the relative interfacial area (Fig. 75), h_{f} is the height of the foaming layer, A_{ac} is the active transfer area, and $\dot{V}_{\mathrm{G}}$ is the volumetric vapor flow rate. Analogously, for the liquid phase,

$$N_{\mathrm{L}} = \frac{\beta_{\mathrm{L}} \cdot a \cdot h_{\mathrm{f}} \cdot A_{\mathrm{ac}}}{\dot{V}_{\mathrm{L}}} \tag{8.36}$$

Substitution of Equations (8.35) and (8.36) into Equation (8.34) gives

$$N_{\mathrm{OG}} = \frac{a \cdot \beta_{\mathrm{G}} \cdot h_{\mathrm{f}}/w_{\mathrm{G}}}{1 + m \cdot \beta_{\mathrm{G}}/\beta_{\mathrm{L}} \cdot \hat{M}_{\mathrm{L}}/\hat{M}_{\mathrm{G}} \cdot \varrho_{\mathrm{G}}/\varrho_{\mathrm{L}}} \tag{8.37}$$

where w_{G} is the superficial vapor velocity; $\hat{M}_{\mathrm{L}}$ and $\hat{M}_{\mathrm{G}}$ are the molar masses; and ϱ_{L} and ϱ_{G} the densities, of the liquid and vapor, respectively. The numerator in this equation represents mass

transfer in the vapor phase, while the denominator takes into account the additional resistance to mass transfer in the liquid.

Equation (8.37) is generally applicable and contains four unknowns that are specific to the apparatus: the relative interfacial area *a,* the height of the two-phase layer h_f, and the mass-transfer coefficients in the vapor β_G and in the liquid β_L.

Determination of the relative interfacial area *a* and the mass-transfer coefficients β_G and β_L, in particular, is extremely difficult. Thus, Equation (8.37) is seldom used directly; instead empirical correlations are given for the transfer units [80].

Evaluation of Equation (8.34) for different values of N_G and N_L is shown in Figure 79. The number of vapor transfer units N_G is in the range of ca. 0.5 to 2, whereas the number of liquid transfer units N_L varies more widely from 0.5 to 15. Because, in rectification, the term $m/(\dot{L}/\dot{G})$ is usually in the order of 1, point efficiencies of approximately 0.6 to 0.9 are obtained.

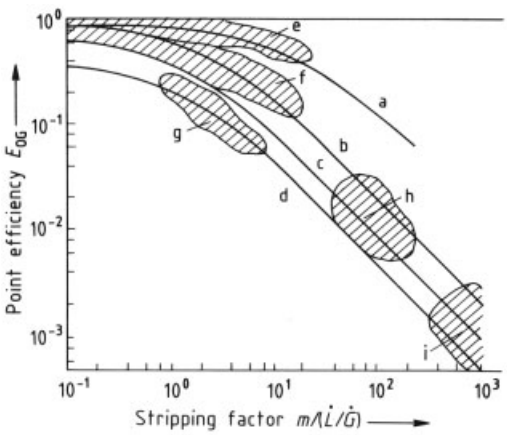

Figure 79. Standard values for the gas point efficiency E_{OG}
Values calculated from Equations (8.35) and (8.36): a) $N_G = 2$, $N_L = 15$; b) $N_G = 2$, $N_L = 2$; c) $N_G = 1$, $N_L = 1$; d) $N_G = 0.5$, $N_L = 0.5$
Experimental data: e) Distillation; f) Vacuum distillation; g) Absorption $\eta_L = 10-20$ cP; h) Absorption of carbon dioxide in water; i) Desorption of oxygen from water

Mass-Transfer Coefficients. The mass-transfer coefficients in the vapor, β_G, and in the liquid, β_L, can be estimated by assuming nonsteady mass transfer:

$$\beta_G = \sqrt{\frac{4}{\pi} \cdot \frac{D_G}{\tau}} \text{ and } \beta_L = \sqrt{\frac{4}{\pi} \cdot \frac{D_L}{\tau}} \tag{8.38}$$

where D_G and D_L are the diffusion coefficients of the vapor and liquid, respectively. Determination of the contact time τ is difficult. According to [75], it is equivalent to the lifetime of the interface in the two-phase layer. Because the interface is produced by the vapor penetrating the liquid, it is simultaneously the residence time of the vapor in the two-phase layer. Thus the same value is obtained for the contact time τ in both the liquid and the vapor phases. Hence, the mass-transfer coefficient in the vapor is

$$\beta_G = \sqrt{\frac{4}{\pi} \cdot \frac{D_G}{h_f} \cdot \frac{w_G}{\varepsilon_G}} \tag{8.39}$$

where w_G is the superficial gas velocity, h_f is the height of the froth layer, and ε_G is the vapor porosity. Analogously, the mass-transfer coefficient in the liquid is

$$\beta_L = \sqrt{\frac{4}{\pi} \cdot \frac{D_L}{h_f} \cdot \frac{w_G}{\varepsilon_G}} \tag{8.40}$$

These relationships apply to the froth regime and, to a close approximation, to the drop regime, but not to the bubble regime. Substitution of Equations (8.39) and (8.40) into Equation (8.37) gives

$$N_{OG} = \frac{a\sqrt{4/\pi \cdot D_G/w_G \cdot h_f/\varepsilon_G}}{1 + m\sqrt{D_G/D_L} \cdot \hat{M}_L/\hat{M}_G \cdot \varrho_G/\varrho_L} \tag{8.41}$$

The denominator in this equation takes into account additional mass-transfer resistance in the liquid phase; it contains only system properties.

Whereas in the basic relationship (Eq. 8.37), the overall vapor mass-transfer units N_{OG} are directly proportional to the height of the two-phase froth layer h_f and inversely proportional to the vapor velocity w_G, a proportionality to the square root of these parameters is obtained here; this agrees well with observed behavior.

8.3. Dimensioning of Packed Columns

The majority of packed columns use random packings composed of individual particles. Some of the more important packings are shown in Figure 80). The flow rate and the separation efficiency of the column are directly dependent on particle shape. Randomly dumped packings

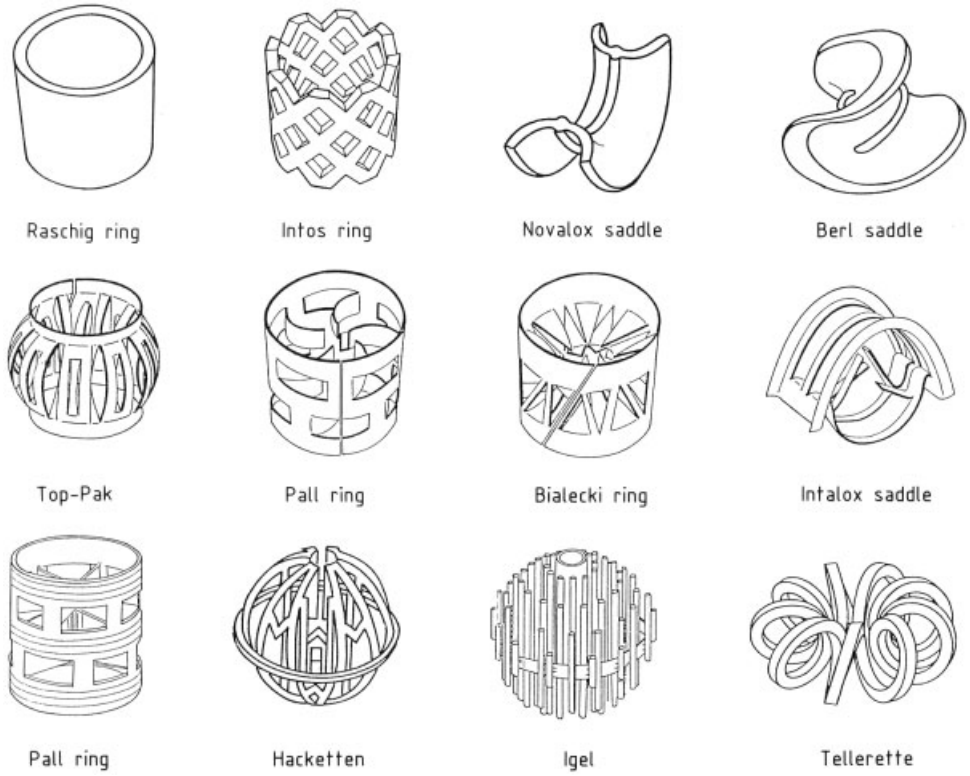

Figure 80. Common random column packings
Ceramic packings are shown in the top row, metal packings in the middle row, and plastic packings in the bottom row.

should be as homogeneous as possible; bridging and large cavities between particles should be avoided. The particles should be sufficiently stable and should form a large packing surface with a low pressure drop. The particle size used depends on column diameter; depending on particle shape, it should not exceed 1/10 to 1/30 of the column diameter. Most particles are made of ceramics or metals, but plastics are being used increasingly. Characteristics of the most important packings compiled in Table 2.

The use of structured column packing dates from the 1960s. These packings provide a homogeneous bed structure (Fig. 81). They are especially especially useful, above all in vacuum columns. Sheet metal and plastic packings are being used increasingly. Characteristics of the most important structured packings are given in Table 3.

Table 3. Manufacturers' specifications for structured column packings

Packing	Designation	Relative area a, m^2/m^3	Porosity ε
Corrugated metal gauze (Sulzer)	CX	700	0.85
	BX	500	0.90
	AX	250	0.96
Corrugated metal sheet (Montz)	B1 300	300	0.97
	B1 200	200	0.98
	B1 100	100	0.99
Corrugated metal sheet (Gempak)	4A	525	0.91
	3A	394	0.92
	2A	262	0.93
	1A	131	0.96
Corrugated metal sheet (Flexipak)	1	558	0.91
	2	250	0.93
	3	135	0.96
	4	69	0.98
Wire mesh (Rombopak)		754	

Table 2. Manufacturers' specifications for random column packings

Particle	Diameter, mm	Ceramic		Metal		Plastic	
		Relative area a, m^2/m^3	Porosity ε	Relative area a, m^2/m^3	Porosity ε	Relative area a, m^2/m^3	Porosity ε
Spheres	2	1800	0.4				
	4	900	0.4				
	6	600	0.4				
	8	450	0.4				
	10	360	0.4				
	15	240	0.4				
	20	180	0.4				
	25	144	0.4				
	30	120	0.4				
	35	100	0.4				
	50	72	0.4				
Raschig rings	5	1000	0.56	1000	0.87		
	10	440	0.65	500	0.89		
	15	330	0.70	350	0.92	350	0.86
	20	240	0.72			290	0.87
	25	195	0.73	220	0.92	220	0.87
	35	140	0.76	150	0.93	150	0.90
	50	98	0.77	110	0.95	110	0.91
	80	60	0.77	65	0.96	65	0.91
Pall rings	10			515	0.92	350	0.88
	15			360	0.93	220	0.91
	25	220	0.73	215	0.94	160	0.93
	35	165	0.76	145	0.94	110	0.93
	50	120	0.77	105	0.95		
	80	75	0.77	78	0.96		
Bialecki rings	25			225	0.95		
	35			155	0.95		
	50			110	0.96		
	80			68	0.97		
Torus saddles	15	450	0.71				
	20	355	0.72				
	25	255	0.74				
	35	166	0.76				
	50	120	0.79				
	75	92	0.80				
Hiflow	20	285	0.76				
	25			185	0.96	218	0.92
	35	140	0.83				
	50	97	0.81	93	0.98	114	0.93
	75	61	0.85				
Tellerettes	no. 1					180	0.87
	no. 2					125	0.93
	no. 3					98	0.92
Cascade mini-rings	no. 0			338	0.97	321	0.89
	no. 1			249	0.96	183	0.92
	no. 2	89		114	0.97	141	0.94
	no. 3	66	0.76	105	0.98	75	0.96
	no. 4	49	0.78	75	0.98		
	no. 5			49	0.99		
	no. 7	46	0.78				

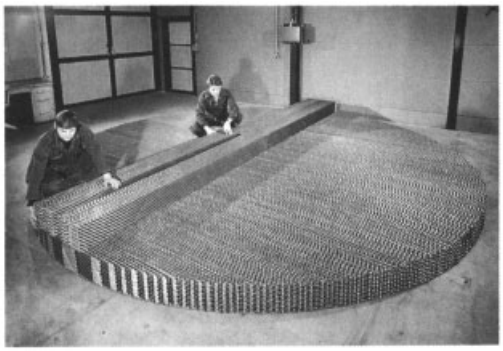

Figure 81. Element of a structured packing (Sulzer Mellapak) constructed to fit through the manway of a large-diameter column

Column calculations are carried out by representing the different packings with simple geometric models, which are usually parallel tubes or beds of spheres. The model structure is chosen so that it has the same ratio of surface area to volume (a) and the same porosity ε as the actual packing. If a bed-of-spheres model is assumed,

$$a = \frac{A}{V} = \frac{z \cdot d^2 \cdot \pi}{V} \tag{8.42}$$

where A is the surface area of the packing, V the bed volume, z the number of "model" spheres, and d their diameter. Furthermore,

$$\varepsilon = \frac{V - V_{\mathrm{sp}}}{V} = 1 - \frac{z}{V} \cdot \frac{d^3 \cdot \pi}{6} \tag{8.43}$$

where V_{sp} is the volume occupied by the spheres.

By eliminating the number of spheres z, the following equation is obtained:

$$d = \frac{6 \cdot (1 - \varepsilon)}{a} \tag{8.44}$$

Because a (the surface area relative to the volume) and ε (the porosity) are known for the actual model, the diameter of the model sphere can be determined by using Equation (8.44). The porosities of modern packings are so large that the spheres in the model structure do not touch each other.

8.3.1. Operating Region of Packed Columns

The permitted operating region of packed columns is depicted in Figure 82; the vapor load $\dot{V}_{\mathrm{G}}/A_{\mathrm{c}}$ is plotted against the liquid load $\dot{V}_{\mathrm{L}}/A_{\mathrm{c}}$ (A_{c} is the cross-sectional area of the column). The operating region of packed columns differs considerably from that of plate columns (cf. Fig. 70). In packed columns, the vapor load can be reduced to extremely small values, but the liquid load must be maintained within a defined range. This difference influences the choice of apparatus.

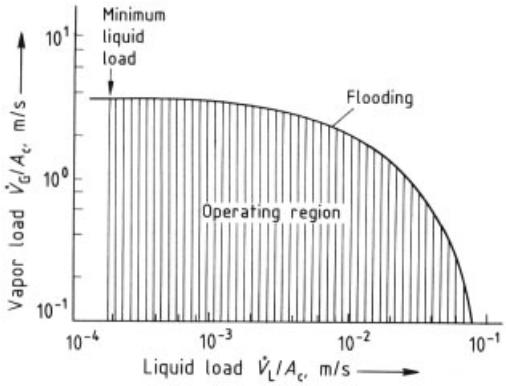

Figure 82. Operating region of a packed column

Two principal mechanisms limit the gas and liquid flow rates: flooding and wetting.

Flooding. The *flooding boundary* is an absolute boundary that cannot be crossed under any circumstances. At the flooding point, the countercurrent flow of vapor and liquid in the column breaks down. The liquid builds up in the packing and is pushed upward by the vapor. A froth layer can arise above the packing after the flood point has been reached. The pressure drop of the vapor increases tremendously at the flood point. Determination of the flood point is discussed in Section 8.3.2.

Wetting. In contrast to plate columns, packed columns require a minimum liquid load to ensure sufficient mass transfer. Below this minimum value, only a very small part of the column is wetted, and liquid and vapor are no longer in intimate contact. This results in a considerable drop in separating efficiency. The minimum liquid load $w_{\mathrm{L,\,min}}$ can be calculated from

$$w_{\mathrm{L,min}} = 7.7 \times 10^{-6} \left(\frac{\varrho_{\mathrm{L}} \cdot \sigma^3}{\eta_{\mathrm{L}}^4 \cdot g} \right)^{2/9} \cdot \left(\frac{g}{a} \right)^{1/2} \tag{8.45}$$

where ϱ_L is the liquid density, σ the surface tension, g the acceleration due to gravity, η_L the liquid viscosity, and a the relative surface area of the packing [81]. In practice, minimum liquid loads for dumped packings are $10\,m^3\,m^{-2}\,h^{-1}$ for aqueous systems and $5\,m^3\,m^{-2}\,h^{-1}$ for organic systems. Columns with structured packings can be operated in an organic medium with liquid amounts as small as $0.2\,m^3\,m^{-2}\,h^{-1}$ [64].

8.3.2. Two-Phase Flow in Packed Columns

A packed column allows a pure countercurrent flow of liquid and vapor streams. The liquid trickles downward over the surface of the packing in the form of rivulets or films. Considerable drop formation can also occur within the bed if a large packing size is used.

Pressure Drop. Flow through a particle bed is a basic step of multiphase flow processes that occur, for example, in fixed beds, pneumatic and hydraulic conveying, and fluidized beds. A common principle is involved, namely, flow around a single particle in the presence of neighboring particles. Fluidized beds are more suitable for investigating this basic process than fixed beds because the porosity of the fluidized bed can be varied more widely.

RICHARDSON and ZAKI have developed the following relationship for homogeneous fluidized beds [82]:

$$w_s/w_0 = \varepsilon^n \tag{8.46}$$

where w_0 is the terminal velocity of a single particle, w_s is the superficial velocity of the vapor required to fluidize a swarm of particles, and ε is the porosity. The exponent n is a function of the Reynold's number of a single particle Re_0, as shown in Figure 83. If the influence of porosity is expressed in terms of the friction factors for a swarm of particles ζ_s and a single particle ζ_0, the following expression is obtained:

$$\zeta_s/\zeta_0 = \varepsilon^{-4.65} \tag{8.47}$$

This equation is equivalent to Equation (8.46) for the entire range of Reynold's numbers, as shown in Figure 83.

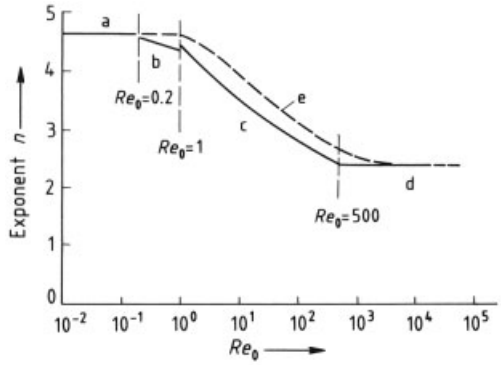

Figure 83. Correlation of the exponent n as defined by RICHARDSON and ZAKI (Eq. 8.46) with the Reynold's number of a single particle Re_0 in a homogeneous fluidized bed [82]
$Re_0 = w_0 d/\nu$, where w_0 is the superficial particle velocity, d its diameter, and ν the kinematic viscosity.
a) $n = 4.65$; b) $n = 4.35/Re_0^{0.03}$; c) $n = 4.45/Re_0^{0.1}$; d) $n = 2.39$; e) $\zeta_s/\zeta_0 = \varepsilon^{-4.65}$ (Eq. 8.47)

The following expression holds for the pressure drop Δp in the fluidized bed:

$$\Delta p \cdot A_c = H \cdot A_c \cdot (1-\varepsilon) \cdot \Delta\varrho \cdot g \tag{8.48}$$

where H is the height of the packing, A_c the cross-sectional area of the column, $\Delta\varrho$ the density difference, and g the acceleration due to gravity. For a single particle in the bed,

$$\frac{d^2\pi}{4} \cdot \zeta_s \cdot \frac{\varrho_G}{2} w_G^2 = \frac{d^3\pi}{6} \cdot \Delta\varrho \cdot g \tag{8.49}$$

where w_G is the superficial vapor velocity. Elimination of $\Delta\varrho \cdot g$ and use of Equation (8.47) give

$$\frac{\Delta p}{H} = \frac{3}{4} \cdot \zeta_0 \cdot \frac{1-\varepsilon}{\varepsilon^{4.65}} \cdot \varrho_G \cdot \frac{w_G^2}{d} \tag{8.50}$$

This equation can also be applied to fixed beds. It differs considerably from the commonly used Ergun equation in the improved porosity term $(1-\varepsilon)/\varepsilon^{4.65}$ [83]:

$$\frac{\Delta p}{H} = \zeta_E \left(\frac{1-\varepsilon}{\varepsilon^3}\right) \varrho_G \cdot \frac{w_G^2}{d} \tag{8.51}$$

The improved term in Equation (8.50) enables determination of the pressure drop in a particle bed from the friction factor ζ_0 of a single particle. The Ergun equation requires utilization of a different packing friction factor ζ_E which bears no relation to the friction factor of the single particle.

Figure 84 shows a correlation of the friction factor ζ_0 and the Reynold's number for the most

important packings. The curves can be correlated by using the following equation:

$$\zeta_0 = \frac{C_1}{Re} + \frac{C_2}{\sqrt{Re}} + C_3 \tag{8.52}$$

where C_1, C_2, and C_3 are constants; for values, see Table 4.

Table 4. Constants for calculating the constants in Equation (8.52)

Packing material	C_1	C_2	C_3
Spheres	24	4	0.28
Raschig rings (metal)	48	8	2.0
Raschig rings (ceramic)	60	1	7.5
Pall rings	33	7	1.4
Bialecki rings	50	7	2.5
Saddles	32	7	1.0
Structured packings	18	4	0.2

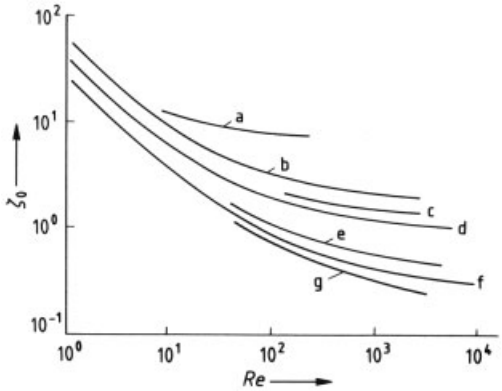

Figure 84. Dependence of the friction factor ζ_0 of a single particle on the Reynold's number *Re*
a) Raschig rings, metal; b) Raschig rings, ceramic; c) Pall rings; d) Saddles; e) Cylinders; f) Spheres; g) Structured packings

Pressure Drop. As shown in Figure 85, the pressure drop is always larger in an irrigated column (curves b – f) than in a dry column (curve a). The liquid flowing through the column changes the bed structure due to holdup liquid in the bed, which is defined as

$$h = V_L/V \tag{8.53}$$

where V_L is the liquid volume and V the bed volume. The holdup decreases the porosity ε', increases the particle diameter d', and changes the friction factor ζ_0' (the prime indicates the irrigated column).

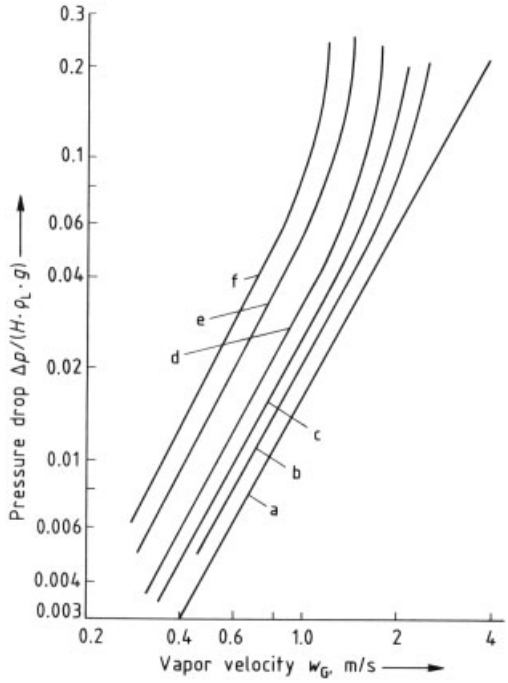

Figure 85. Pressure drop Δp in a stacked packing composed of 25-mm metal Bialecki rings (porosity 0.94) as a function of vapor (w_G) and liquid (w_L) loads [84]
Liquid velocity w_L, m/h: a) 0; b) 10; c) 20; d) 40; e) 60; f) 80
Measurements were made using air – water at 20 °C and 0.1 MPa in a column with a packing height of 1.5 m and a diameter of 0.15 m.

The porosity ε' of the irrigated bed can be expressed as

$$\varepsilon' = \varepsilon - h \text{ or } \varepsilon' = \varepsilon \cdot (1 - h/\varepsilon) \tag{8.54}$$

The change in particle diameter can be formulated by assuming that the liquid is present mainly as a film or rivulet on the particles. If the number of particles is the same in irrigated and dry beds, then

$$\frac{1-\varepsilon'}{d'^3} = \frac{1-\varepsilon}{d^3} \tag{8.55}$$

Substituting Equation (8.54) yields

$$d' = d\left[\frac{1-\varepsilon\cdot(1-h/\varepsilon)}{1-\varepsilon}\right]^{1/3} \tag{8.56}$$

Liquid holdup in the packing has a negligible effect on particle shape. Hence, Equation (8.52) for the friction factor can also be applied to irrigated packings. However, the increased particle diameter d' leads to an increased Reynold's

number for the vapor and thus a lower friction factor ζ_0 (Fig. 84). If the Reynold's number is held within a narrow range, Equation (8.52) can be simplified to

$$\zeta_0 \approx Re^c \tag{8.57}$$

where $c = \frac{1}{\zeta_0}\left(-\frac{C_1}{Re} - \frac{C_2}{2\sqrt{Re}}\right)$. The exponent c lies between the limits 0 (fully turbulent flow) and 1 (laminar flow). Thus,

$$\zeta_0' = \zeta_0 \cdot (d'/d)^c \tag{8.58}$$

Using Equation (8.56) gives

$$\zeta_0' = \zeta_0 \left[\frac{1-\varepsilon\cdot(1-h/\varepsilon)}{1-\varepsilon}\right]^{c/3} \tag{8.59}$$

Thus, the pressure drop in the irrigated packing Δp_{irr} can be calculated from Equation (8.50) which applies to a nonirrigated packing:

$$\frac{\Delta p_{\mathrm{irr}}}{H} = \frac{3}{4}\cdot\zeta_0'\cdot\frac{1-\varepsilon'}{\varepsilon'^{4.65}}\cdot\varrho_{\mathrm{G}}\cdot w_{\mathrm{G}}^2/d' \tag{8.60}$$

where H is the packing height, ϱ_{G} the vapor density, and w_{G} the superficial vapor velocity.

A comparison of experimental values obtained for two dumped packings and one structured packing with values calculated by using the above equation is shown in Figure 86.

Equation (8.61) does not contain a packing-specific parameter. The first term accounts for porosity-dependent changes, a fact that has been corroborated by experimental data.

The second term of Equation (8.61) is very similar to the relationship derived by BEMER and KALIS [85]:

$$\frac{\Delta p_{\mathrm{irr}}}{\Delta p_{\mathrm{d}}} = \left(1 - \frac{h/\varepsilon}{2\cdot\Phi\cdot x^{5/3}}\right)^{-5} \tag{8.62}$$

where Φ and x are empirical constants that depend on the shape of the packing. Equation (8.62) is often used with an empirical constant C_{h} which also depends on the shape of the packing and must be determined empirically [86], [87]:

$$\frac{\Delta p_{\mathrm{irr}}}{\Delta p_{\mathrm{d}}} = (1 - C_{\mathrm{h}}\cdot h)^{-5} \tag{8.63}$$

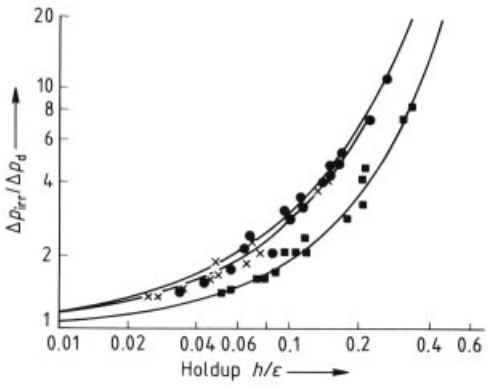

Figure 86. Pressure drop in irrigated packings Δp_{irr} as a function of liquid holdup h
Curves indicate values calculated with Equation (8.61). Symbols indicate experimental values [84], [86]: Bialecki rings (●), porosity $\varepsilon = 0.94$; Gempak 2A (×), $\varepsilon = 0.93$; and Raschig rings (■), $\varepsilon = 0.70$.

The pressure drop in the irrigated packing can be conveniently related to the pressure drop in the nonirrigated (dry) packing Δp_{d}:

$$\frac{\Delta p_{\mathrm{irr}}}{\Delta p_{\mathrm{d}}} = \left[\frac{1-\varepsilon\cdot(1-h/\varepsilon)}{1-\varepsilon}\right]^{(2+c)/3} [1-h/\varepsilon]^{-4.65} \tag{8.61}$$

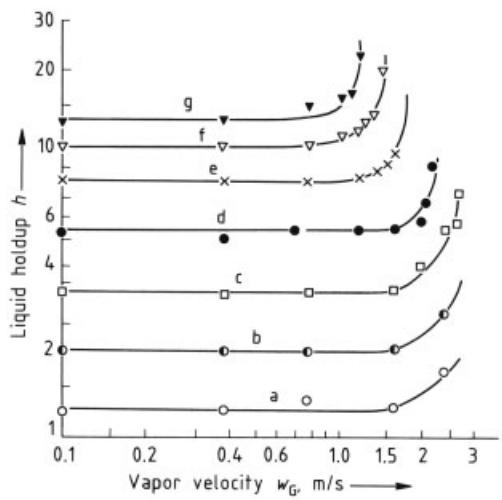

Figure 87. Liquid holdup h in a dumped packing composed of 25-mm Bialecki rings, as a function of vapor and liquid loads [84]
Liquid velocity w_{L}, m/h: a) 2.8; b) 5; c) 10; d) 20; e) 40; f) 60; g) 80
Measurements were made using air – water at 20 °C and 0.1 MPa in a column with a packing height of 1.5 m and a diameter of 0.15 m.

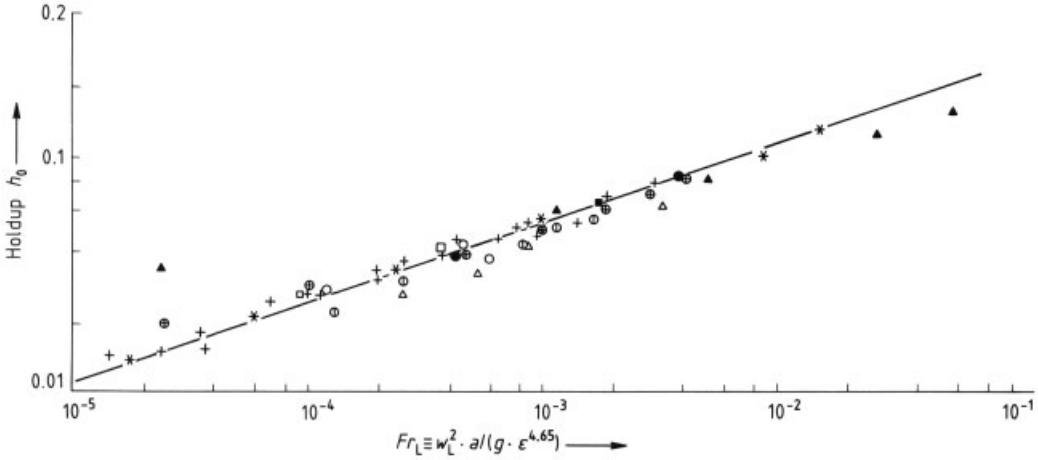

Figure 88. Dependence of the liquid holdup h_0 below the loading point on the Froude number of the liquid Fr_L for random and structured packings

The straight line indicates values calculated with Equation (8.64), $h_0 = 0.555 \cdot Fr_L^{1/3}$. The porosities ($\varepsilon$) and relative surface areas (a) of the different packings are as follows:
Random ceramic packings: Intalox saddles (△), $\varepsilon = 0.73$, $a = 255\,m^{-1}$ [88]; Pall rings (⊕), $\varepsilon = 0.78$, $a = 120\,m^{-1}$ [88]; Raschig rings (▲), $\varepsilon = 0.70$, $a = 300\,m^{-1}$ [86].
Random metal packings: Bialecki rings (∗), $\varepsilon = 0.94$, $a = 255\,m^{-1}$ [84]; Pall rings (+), $\varepsilon = 0.935 - 0.952$, $a = 360 - 100\,m^{-1}$ [88].
Random plastic packings: NSW rings (□), $\varepsilon = 0.96$, $a = 90\,m^{-1}$ [88]; Pall rings (○—), $\varepsilon = 0.92$, $a = 110\,m^{-1}$ [88]; Hiflow (○), $\varepsilon = 0.94$, $a = 110\,m^{-1}$ [88].
Structured metal packings: Mellapack (●), $\varepsilon = 0.96$, $a = 250\,m^{-1}$ [88]; Gempack 230 A2T (■), $\varepsilon = 0.977$, $a = 202\,m^{-1}$ [88].

Liquid Holdup. An important factor for two-phase flow in packed columns is the liquid holdup h (defined in Eq. 8.53) in the wet packing. Typical experimental curves for the holdup h in Bialecki ring packing at various gas and liquid loads are shown in Figure 87 [84]. Two regions can be distinguished: (1) a low vapor-load region in which the holdup h depends solely on the liquid load and (2) a high vapor-load region in which h depends on both the liquid and the vapor loads. The transition between these two regions occurs at the loading point.

Experimental data for h_0, the holdup below the loading point, are presented in Figure 88. These data can be correlated by using the following equation:

$$h_0/\varepsilon = 0.555 \cdot Fr_L^{1/3}, \text{ with } Fr_L = \frac{w_L^2 \cdot a}{g \cdot \varepsilon^{4.65}} \qquad (8.64)$$

where Fr_L is the Froude number of the liquid, w_L the superficial liquid velocity, a the relative surface area of the packing, and ε its porosity. This correlation does not account for any liquid properties and is applicable only to air and water. However, the holdup also depends on the liquid viscosity η_L [89] and the interfacial tension σ [90], [91]. These results have been correlated, but the correlation applies only to Raschig rings with a maximum diameter of 25 mm [92].

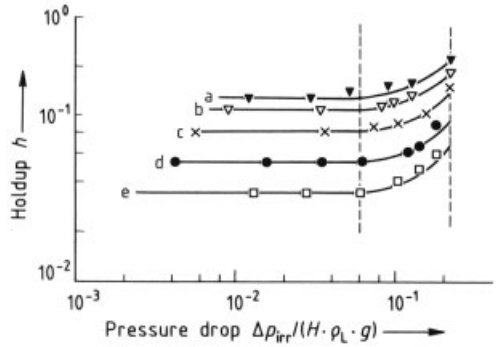

Figure 89. Liquid holdup in a dumped packing composed of 25-mm Bialecki rings (metal) as a function of the pressure drop in the irrigated packing Δp_{irr}
Liquid velocity w_L, m/h: a) 80; b) 60; c) 40; d) 20; e) 10

The liquid holdup h above the loading point is further influenced by both the liquid and the vapor loads. Figure 87 shows that as the absolute values of the liquid holdup h increase, increas-

ingly smaller vapor loads become effective. If the liquid holdup h is plotted against the pressure drop of the vapor, $\Delta p_{\mathrm{irr}}/(H \cdot \varrho_{\mathrm{L}} \cdot g)$, instead of against the vapor velocity w_{G}, a uniform increase is obtained (Fig. 89). The following relationship is deduced from Figure 89:

$$h = h_0 \left[1 + 20 \left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \right)^2 \right] \tag{8.65}$$

where h_0 is the holdup below the loading point.

Flooding in Packed Columns. The flooding point is an important parameter in column dimensioning. It represents the column load at which the liquid can no longer flow countercurrently to the rising vapor. Liquid flow is obstructed by the high vapor load to such an extent that it collects in the bed. The flooding point is clearly indicated (e.g., in Fig. 85) by a very steep increase in the pressure drop.

Empirical correlations are available for determining the flooding point in the form of diagrams [93–96]. An example is shown in Figure 90.

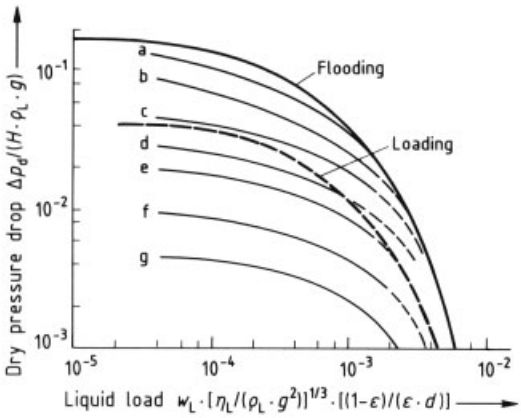

Figure 90. Mersmann correlation of flooding point and pressure drop in an irrigated packing [96]
Calculated curves are shown for different values of the pressure drop in an irrigated packing, $\Delta p_{\mathrm{irr}}/(H \cdot \varrho_{\mathrm{L}} \cdot g)$: a) 0.2; b) 0.1; c) 0.05; d) 0.03; e) 0.02; f) 0.01; g) 0.005

The pressure drop in irrigated packings described above can be used to determine the flooding point. By using Equations (8.61) and (8.65), the pressure drop based on the holdup h above the loading point can be expressed as

$$\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} = \frac{\Delta p_{\mathrm{d}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \cdot \frac{\left(1 - \varepsilon \cdot \left\{ 1 - \frac{h_0}{\varepsilon} \cdot \left[1 + 20 \left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \right)^2 \right] \right\} \right)^{(2+c)/3}}{1 - \varepsilon} \cdot \left\{ 1 - \frac{h_0}{\varepsilon} \left[1 + 20 \left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot \mathrm{g}} \right)^2 \right] \right\}^{-4.65} \tag{8.66}$$

The flooding point is reached when an infinitely steep increase in Δp_{irr} occurs with increasing vapor load [97]. Thus, if the dry pressure drop is used as a measure of the vapor load

$$\frac{\partial \Delta p_{\mathrm{irr}}}{\partial \Delta p_{\mathrm{d}}} = \infty \text{ or } \frac{\partial \Delta p_{\mathrm{d}}}{\partial \Delta p_{\mathrm{irr}}} = 0 \tag{8.67}$$

After differentiation of Equation (8.66),

$$\left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \right)_{\mathrm{f}}^{-2} - \frac{40 \frac{2+c}{3} \cdot h_0}{1 - \varepsilon + h_0 \left[1 + 20 \left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \right)_{\mathrm{f}}^2 \right]} - \frac{186 \cdot h_0}{\varepsilon - h_0 \left[1 + 20 \left(\frac{\Delta p_{\mathrm{irr}}}{H \cdot \varrho_{\mathrm{L}} \cdot g} \right)_{\mathrm{f}}^2 \right]} = 0 \tag{8.68}$$

The pressure drop at which flooding occurs can be calculated from the above equation. In general, values of $\Delta p_{\mathrm{irr}}/(H \cdot \varrho_{\mathrm{L}} \cdot g)_{\mathrm{f}}$ in the range of 0.1 to 0.3 are obtained, which agree well with experimental results.

The results calculated from the pressure drop equation (Eq. 8.66) and the flooding point equation (Eq. 8.68) for Bialecki rings are plotted as a function of relative holdup h_0/ε in Figure 91. Agreement with experimentally determined flooding points is excellent. The representation in Figure 91 is very similar to the flooding point diagram shown in Figure 90; the ordinate and parameters are identical, but in Figure 90 the term $(Fr^2/Re)^{1/3}$ is used as the abscissa. This term gives a relatively good correlation for the holdup h_0, within a certain range of liquid properties. However, the surface tension σ can also be expected to have some effect [90], [91].

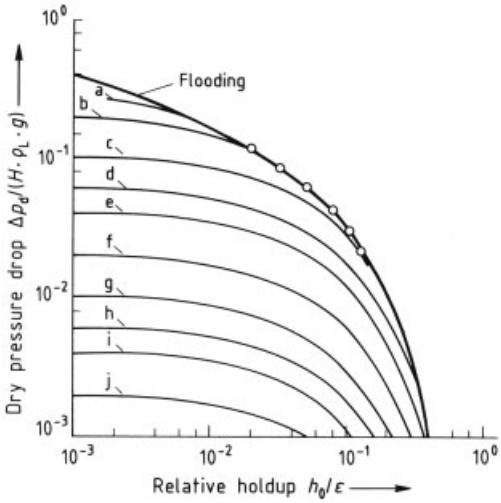

Figure 91. Flooding boundary and pressure drop in an irrigated packing composed of 25-mm Bialecki rings (metal) Open circles indicate experimental values [84]. Calculated curves (Eqs. 8.66 and 8.68) are shown for different values of the pressure drop in an irrigated packing, $\Delta p_{\mathrm{irr}}/(H\cdot\varrho_{\mathrm{L}}\cdot g)_{\mathrm{f}}$: a) 0.3; b) 0.2; c) 0.1; d) 0.06; e) 0.04; f) 0.02; g) 0.01; h) 0.006; i) 0.004; j) 0.002

The decisive factor in calculating the flooding point and the pressure drop in an irrigated column packing is the liquid holdup h_0. Published data (usually for air – water) are not sufficient to provide a reliable equation for this; data at low interfacial tensions and viscosities, which are typical of rectification, are lacking.

8.3.3. Calculation of Mass Transfer in Packed Columns

Principles. The calculation of mass transfer in a packed column is based on the assumption that both vapor and liquid move through the column bed in a type of plug flow. This means that back-mixing in the axial direction is not allowed. Furthermore, both phases must be uniformly distributed over the mass-transfer area A and throughout the packed height H. By using the transfer-unit concept, the following relationship is obtained (cf. Section 4.1.2):

$$\mathrm{d}N_{\mathrm{OG}}=K_{\mathrm{OG}}/\dot{G}\cdot\mathrm{d}A \tag{8.69}$$

When $\mathrm{d}A=\mathrm{d}H\cdot A_{\mathrm{c}}\cdot a$ (A_{c} is the cross-sectional area of the column and a is the relative surface area of the packing), the differential height of the packing is

$$\mathrm{d}H=\mathrm{d}N_{\mathrm{OG}}\cdot\frac{\dot{G}}{K_{\mathrm{OG}}\cdot A_{\mathrm{c}}\cdot a} \tag{8.70}$$

The second term of this equation is defined as the height of an overall gas-phase transfer unit H_{OG}:

$$H_{\mathrm{OG}}=\frac{\dot{G}}{K_{\mathrm{OG}}\cdot A_{\mathrm{c}}\cdot a} \tag{8.71}$$

The overall gas-phase mass-transfer coefficient K_{OG} is calculated from the coefficients of the phases involved, K_{G} and K_{L}:

$$\frac{1}{K_{\mathrm{OG}}}=\frac{1}{K_{\mathrm{G}}}+\frac{m}{K_{\mathrm{L}}} \tag{8.72}$$

where m is the slope of the equilibrium curve on the $y-x$ diagram. Furthermore,

$$K_{\mathrm{G}}=\beta_{\mathrm{G}}\cdot\varrho_{\mathrm{G}}/\hat{M}_{\mathrm{G}}\ \text{and}\ K_{\mathrm{L}}=\beta_{\mathrm{L}}\cdot\varrho_{\mathrm{L}}/\hat{M}_{\mathrm{L}} \tag{8.73}$$

Simple rearrangement gives

$$\frac{\dot{G}}{A_{\mathrm{c}}}=w_{\mathrm{G}}\cdot\varrho_{\mathrm{G}}/\hat{M}_{\mathrm{G}} \tag{8.74}$$

By using Equations (8.72) – (8.74), the height of a mass transfer unit can be expressed as

$$H_{\mathrm{OG}}=\frac{w_{\mathrm{G}}}{a\cdot\beta_{\mathrm{G}}}\cdot\left(1+m\frac{\beta_{\mathrm{G}}}{\beta_{\mathrm{L}}}\cdot\frac{\hat{M}_{\mathrm{L}}}{\hat{M}_{\mathrm{G}}}\cdot\frac{\varrho_{\mathrm{G}}}{\varrho_{\mathrm{L}}}\right) \tag{8.75}$$

Equation (8.75) contains three unknown terms specific to the apparatus: the volumerelated interfacial area a, the mass-transfer coefficient β_{G} in the gas, and the mass-transfer coefficient β_{L} in the liquid. Standard values for the height of one transfer unit are in the range $H_{\mathrm{OG}}=0.2-0.3$ m for structured packing. Values for randomly dumped packing depend strongly on the column diameter D_{c} because the particle size and, consequently, the surface area of the packing bear a definite relation to the column diameter. The following value is often employed:

$$H_{\mathrm{OG}}\approx(0.5-1)\cdot D_{\mathrm{c}} \tag{8.76}$$

Thus, small laboratory columns produce excellent separation, whereas industrial columns with large diameters are much less effective.

Determination of Unknowns. Determination of the unknown factors in Equation (8.75) is discussed in the following material.

The *mass-transfer area A* in a packing is related to the bed volume V to give the volume-related mass-transfer area of the packing a:

$$a = A/V \tag{8.77}$$

The liquid generally moves in rivulets over the surface of the bed, but complete wetting of the particles usually does not occur. Hence, the mass-transfer area is generally smaller than the surface area of the packing. Experiments on the movement of rivulets on plates have provided the following relationship for the wet surface area a_{irr} [98], [99]:

$$a_{\mathrm{irr}} = 0.76 \cdot w_{\mathrm{L}}^{0.4} \cdot \nu_{\mathrm{L}}^{0.2} \cdot \left(\frac{a^2}{\varepsilon}\right)^{0.6} \cdot \left(\frac{\varrho_{\mathrm{L}}}{\sigma \cdot g}\right)^{0.15} / (1 - 0.93 \cdot \cos\theta) \tag{8.78}$$

The ability of different liquids to wet various packing materials is expressed in terms of the contact angle θ. Values of θ are given in [99]. The effective interface area for mass transfer a_{e} is calculated as follows:

$$a_{\mathrm{e}} = F_{\mathrm{a}} \cdot a_{\mathrm{irr}} \tag{8.79}$$

where $F_{\mathrm{a}} = 52 \cdot d^{1.1}$ for ceramic materials
$F_{\mathrm{a}} = 138 \cdot d^{1.1}$ for plastic materials
The proportionality factor F_{a} takes into account the additional influences of packing size and material.

The *gas-phase mass-transfer coefficient* β_{G} can be determined with the following equation [100]:

$$\beta_{\mathrm{G}} = C_{\beta\mathrm{G}} \left(\frac{w_G^2 \cdot D_{\mathrm{G}}^2}{d_{\mathrm{N}} \cdot \nu_{\mathrm{G}}}\right)^{1/3} \tag{8.80}$$

where $C_{\beta\mathrm{G}}$ is a constant, w_{G} is the superficial gas velocity, D_{G} is the vapor diffusion coefficient, d_{N} is the nominal diameter of the packing, and ν_{G} is the kinematic viscosity. This relationship is applicable when

$$Re_{\mathrm{G}} = \frac{\varrho_{\mathrm{G}} \cdot w_{\mathrm{G}} \cdot d_{\mathrm{N}}}{\eta_{\mathrm{G}} \cdot (1 - \varepsilon)} > 1000 \tag{8.81}$$

Values of the constant $C_{\beta\mathrm{G}}$ for ceramic and metal packings are 0.673 (beads), 0.417 (Raschig rings), and 0.73 (saddles); this is purely a shape factor that is independent of particle size.

According to a model developed by Zech [101], a nonsteady mass-transfer mechanism exists in the liquid. A defined liquid element is located on the phase interface as long as the film or rivulet flows over a single packing element. At the point of contact with the next element, the liquid is completely mixed and a new liquid surface is formed. This new interface, in turn, exists only as long as the liquid flows over this particle and so on. An equation for the *liquid mass-transfer coefficient* based on this model has been derived [99]:

$$\beta_{\mathrm{L}} = 0.86 \left(\frac{6 \cdot D_{\mathrm{L}}}{\pi \cdot d_{\mathrm{N}}}\right)^{1/2} \cdot \left[\frac{w_{\mathrm{L}}^{1.2} \cdot g \cdot \sigma_{\mathrm{L}}^{0.3} \cdot \varepsilon^{1.2} 1/3}{\nu_{\mathrm{L}}^{1.4} \cdot \varrho_{\mathrm{L}}^{0.3} \cdot a^{2.4}} (1 - 0.93 \cdot \cos\theta)^2\right]^{1/6} \tag{8.82}$$

Problems. The formulas for calculating the mass-transfer efficiency of packed columns, described above, are not totally satisfactory. One reason is that the phase interface is difficult to determine. The large number of equations developed for this purpose do not show a uniform trend with respect either to the order of magnitude of the interface or to the dependence on operational, material, or geometric parameters. Even the determination of mass-transfer coefficients, especially in the liquid, is somewhat uncertain.

The main reason for the persistent inaccuracy in calculating the mass-transfer efficiency of packed columns is the assumption that plug flow occurs within the packing. The associated assumption of uniform liquid flow rate over the column cross section is certainly not valid in large-scale industrial columns. Uniform distribution of liquid over the entire flow area becomes increasingly difficult to achieve at large column diameters (maldistribution).

Liquid maldistribution in a column of 3.6-m packed height and 0.63-m diameter is shown in Figure 92, which employs the technique described for plate columns in Figure 74 [102]. The longitudinal liquid isotherms can be interpreted as lines of equal liquid residence time. These isotherms should be horizontal straight lines if plug flow occurs. Figure 92 A shows that despite excellent initial liquid distribution, significant

deviations from plug flow occur with increasing column length. Uneven liquid distribution is observed not only at the edges of the column, but also within the column; this is caused primarily by a slight inhomogeneity of the packing.

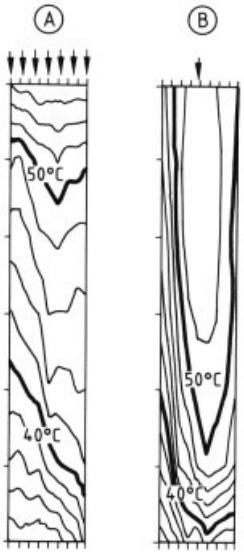

Figure 92. Liquid maldistribution in a column of 3.6-m packed height and 0.63-m diameter packed with 35-mm Pall rings [102]
The indicated liquid isotherms should be horizontal straight lines for uniform liquid flow through the cross-sectional area of the column.
A) Multiple feed entry points, $\bar{H}_{OG} = 0.44$ m; B) Single feed entry point, $\bar{H}_{OG} = 0.67$ m

Deviations from plug flow are especially marked when liquid is fed in at a single point (Fig. 92 B). The initial liquid distribution is poor and changes only slightly along the length of the column. Thus, unequal amounts of vapor and liquid come in contact with each other along the cross-sectional area of the column. In regions of small liquid load, phase equilibrium can be reached relatively rapidly and these zones no longer contribute to the separation efficiency of the column (pinch point).

Detailed theoretical studies have shown that even if the initial liquid distribution is good, a substantial part (up to 50 %) of the possible separation efficiency is lost as a result of maldistribution [102]. Maldistribution is especially troublesome in purely countercurrent columns because the effects accumulate over the entire column height. In plate columns, however, these effects are restricted to a single plate because the liquid is mixed in the next downcomer. Thus, plate columns are especially suitable for large column diameters. Large-diameter packed columns can be employed only if special precautions are taken to ensure good liquid distribution over the flow area.

9. References

General References

1. E. Kirschbaum: *Destillier- und Rektifiziertechnik,* Springer Verlag, Berlin–Heidelberg–New York 1969.
2. E. J. Henley, J. D. Seader: *Equilibrium-Stage Separation Operations in Chemical Engineering,* J. Wiley & Sons, New York 1981.
3. B. D. Smith: *Design of Equilibrium Stage Processes,* McGraw-Hill, New York 1963.
4. R. E. Treybal: *Mass-Transfer Operations,* McGraw-Hill, New York 1968.
5. R. Billet: *Industrielle Destillation,* Verlag Chemie, Weinheim 1973.
6. R. H. Perry, D. W. Green, J. O. Maloney: *Perry's Chemical Engineers Handbook,* McGraw-Hill, New York 1984, pp. 13-1 – 13-97.

Specific References

7. R. J. Forbes: *Short History of the Art of Distillation,* E. J. Brill, Leiden 1948.
8. A. Libavius: *Die Alchemie des Andreas Libavius, ein Lehrbuch der Chemie aus dem Jahre 1597,* Verlag Chemie, Weinheim 1964.
9. A. Mersmann: *Thermische Verfahrenstechnik,* Springer Verlag, Berlin 1980.
10. M. F. Doherty, J. D. Perkins, *Chem. Eng. Sci.* **34** (1979) 1401 – 1414.
11. J. M. Prausnitz: *Molecular Thermodynamics of Fluid-Phase Equilibria,* Prentice-Hall, Englewood Cliffs, N.Y. 1969.
12. J. M. Prausnitz, T. F. Anderson, E. A. Grens, C. A. Eckert, R. Hsieh, J. P. O'Connel: *Computer Calculations for Multicomponent Vapor – Liquid and Liquid – Liquid Equilibria,* Prentice-Hall, Englewood Cliffs, N.Y. 1980.
13. W. Wagner: *Berechnung von Hochdruckphasengleichgewichten azeotroper Gemische,* Dissertation Universität Stuttgart 1985.
14. G. Soave, *Chem. Eng. Sci.* **27** (1972) 1197.
15. H. Knapp, R. Döring, L. Oelbrich, V. Plöcker, J. M. Prausnitz: *Vapor – Liquid Equilibria for*

Mixtures of Low Boiling Substances, Chemistry Data Series vol. VI, Dechema, Frankfurt/Main 1982.
16. L. Asselineau, G. Bogdanic, J. Vidal, *Fluid Phase Equilib.* **3** (1979) 273 – 290.
17. L. H. Horsley: *Azeotropic Data,* Amer. Chem. Soc., Washington, D.C. 1973.
18. *Landolt – Börnstein: Zahlenwerte und Funktionen, Lösungsgleichgewichte I,* vol. II 2 b, Springer Verlag, Berlin 1962.
19. E. Hala, J. Pick, V. Vilim: *Vapor – Liquid Equilibrium,* Pergamon Press, Oxford 1967.
20. M. Hirata, S. Ohe, K. Nagahama: *Computer Aided Data Book of Vapor-Liquid Equilibria,* Kodanska/Elsevier, Tokyo 1975.
21. R. C. Reid, J. M. Prausnitz, T. K. Sherwood: *The Properties of Gases and Liquids,* McGraw-Hill, New York 1977.
22. A. Fredenslund, J. Gmehling, P. Rasmussen: *Vapor – Liquid Equilibria Using UNIFAC,* Elsevier, Amsterdam 1977.
23. J. Gmehling, U. Onken, W. Arlt: *Vapor – Liquid Equilibrium Data Collection,* Dechema Chemistry Data Series, Frankfurt/Main from 1977.
24. J. Gmehling, P. Rasmussen, A. Fredenslund, *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 118 – 127.
25. W. C. Edmister: *Applied Hydrocarbon Thermodynamics,* Gulf Publ., Houston 1961.
26. C. J. King: *Separation Processes,* McGraw-Hill, New York 1980, p. 116.
27. C. J. King: *Separation Processes,* McGraw-Hill, New York 1980, p. 360 ff.
28. T. H. Chilton, A. P. Colburn, *Ind. Eng. Chem.* **27** (1935) no. 3, 255 – 260.
29. C. J. King: *Separation Processes,* McGraw-Hill, New York 1980, p. 220.
30. R. E. Treyball: *Mass-Transfer Operations,* McGraw-Hill, New York 1968.
31. A. J. V. Unterwood, *Chem. Eng. Prog.* **44** (1948) no. 8, 603 – 614.
32. A. Vogelpohl, *Chem. Ing. Tech.* **47** (1975) no. 21, 895.
33. J. C. Wang, G. E. Henke, *Hydrocarbon Process.* **45** (1966) no. 8, 155.
34. C. D. Holland: *Multicomponent Distillation,* Interscience, New York 1963.
35. R. P. Goldstern, R. B. Stanfield, *Ind. Eng. Chem. Process Des. Dev.* **9** (1970) 78.
36. U. Block, B. Hegner, *AIChE J.* **22** (1976) no. 3, 582 – 589.
37. V. Block, B. Hegner, *Vt Verfahrenstechnik* **11** (1977) no. 2, 101 – 104.
38. I. Coward, *Chem. Eng. Sci.* **22** (1967) 503 – 516.
39. H. Egly, V. Ruby, B. Seid, *Chem. Ing. Tech.* **54** (1982) no. 9, 842 – 843.
40. U. Block, H. Schoenmakers, D. Wolf, *Vt Verfahrenstechnik* **12** (1970) no. 10, 678 – 681.
41. M. Streich, *Vt Verfahrenstechnik* **9** (1975) no. 5, 240 – 243.
42. A. Pucci, P. Mihitenho, L. Asselineau, *Chem. Eng. Sci.* **41** (1986) no. 3, 485 – 494.
43. S. I. Abu-Eishuh, W. L. Luyben, *Ind. Eng. Chem. Process Des. Dev.* **24** (1985) 132 – 140.
44. C. J. King: *Separation Processes,* McGraw-Hill, New York 1980, p. 346.
45. L. Berg, J. M. Harrison, C. W. Montgomery, *Ind. Eng. Chem.* **37** (1945) no. 6, 585 – 587.
46. D. Geriche: "Konzentrieren von Salpetersäure mit Schwefelsäure," *Schott Inf.* **3** (1973).
47. Th. Grewer, *Chem. Ing. Tech.* **43** (1971) no. 11, 655.
48. E. J. Hoffmann: *Azeotropic and Extractive Distillation,* Interscience, New York 1964.
49. D. P. Tassios: "Extractive and Azeotropic Distillation," *Adv. Chem. Ser.* **115** (1972).
50. C. J. King: *Separation Processes,* McGraw-Hill, New York 1980, p. 660.
51. H. Leideker: *Wirtschaftliche Einsatzbereiche von Wärmepumpen und Mehrstufenschaltungen in der Rektifiziertechnik,* Dissertation, TU Berlin 1985.
52. K. Kleinhenz, *VDI Ber.* **383** (1980) 43 – 45.
53. H. Lüdeker, H. Gelbe, *Chem. Ing. Tech.* **58** (1986) no. 11, 847 – 855.
54. W. Kreuter, *Linde Reports on Science and Technology* **18** (1973) 23 – 26.
55. K. Kleinhenz, *VDI Ber.* **442** (1982) 45 – 51.
56. H. Schoenmakers, *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) 199 – 203.
57. G. Kaibel, DE 3302525 A1, 1983.
58. T. Umeda, K. Niida, K. Shirdo, *AIChE J.* **25** (1979) no. 3, 423 – 429.
59. B. Linnhoff, E. Hindmarsh, *Chem. Eng. Sci.* **38** (1983) no. 5, 745 – 763.
60. B. Linnhoff, D. R. Vredeveld, *Chem. Eng. Prog.* **7** (1984) 33 – 40.
61. B. Linnhoff, *Chem. Eng. Res. Des.* **61** (1983) 207 – 223.
62. A. Mersmann, H. Voit, R. Zeppenfeld, *Chem. Ing. Tech.* **58** (1986) no. 2, 87 – 96.
63. K. Hoppe, G. Krüger, A. Bahr, M. Hentschel, *Chem. Ing. Tech.* **21** (1969) no. 9, 536 – 542.
64. Prospekt der Fa. Sulzer, No. d/22.13.06–V. 82-45, Winterthur, Switzerland.
65. M. Molzahn, R. Schmidt, *Vt Verfahrenstechnik* **9** (1975) no. 8, 388 – 395.

66. M. Souders, G. G. Brown, *Ind. Eng. Chem.* **26** (1934) no. 1, 98 – 103.
67. C.-J. Huang, J. R. Hodson, *Pet. Refin.* **37** (1958) no. 2, 104 – 118.
68. J. R. Fair, R. L. Matthews, *Pet. Refin.* **37** (1958) no. 4, 153 – 158.
69. J. Fair, *Petro/Chem Eng.* **33** (1961) no. 10, 45 – 52.
70. J. Stichlmair, *Vt Verfahrenstechnik* **8** (1974) no. 11, 323 – 326.
71. G. Wallis: *One-Dimensional Two-Phase-Flow,* McGraw-Hill, New York 1960, p. 377.
72. A. Mersmann, *Chem. Ing. Tech.* **35** (1963) no. 2, 103 – 107.
73. K. Ruff, *Chem. Ing. Tech.* **46** (1974) no. 18, 769 – 771.
74. K. Hoppe, M. Mittelstrass: *Grundlagen der Dimensionierung von Kolonnenböden,* Theodor Steinkopf, Dresden 1967.
75. J. Stichlmair: *Grundlagen der Dimensionierung des Gas/Flüssigkeit-Kontaktapparates Bodenkolonne,* Verlag Chemie, Weinheim–New York 1978.
76. M. Volpert, J. Stichlmair, *Chem. Ing. Tech.* **58** (1986) no. 7, 606 – 607.
77. E. Weisshuhn: *Untersuchungen zur Flüssigkeitsvermischung auf Siebböden,* Dissertation, TU München 1976.
78. J. Stichlmair, S. Ulbrich, *Chem. Eng. Technol.* **10** (1987) 33 – 37.
79. W. K. Lewis, *Ind. Eng. Chem.* **28** (1936) no. 4, 399 – 402.
80. J. A. Gester, A. B. Hill, N. N. Hochgraf, D. G. Robinson: *Tray Efficiencies in Distillation Columns,* Final Report from the University of Delaware (1958) Amer. Inst. Chem. Eng., New York 1958.
81. R. Schmidt: *Instit. Chem. Eng., Symp. Ser.* **56** (1979), pp. 3.1/1.
82. J. F. Richardson, W. N. Zaki, *Trans. Inst. Chem. Eng.* **32** (1954) 35 – 53.
83. S. Ergun, *Chem. Eng. Prog.* **48** (1952) 89 – 94.
84. R. Billet, J. Mackowiak, *Fette Seifen Anstrichm.* **80** (1984) no. 9, 349 – 358.
85. G. G. Bemer, G. A. J. Kalis, *Trans. IChemE.* **56** (1978) 200 – 204.
86. J. E. Buchanan, *Ind. Eng. Chem. Fundam.* **8** (1969) no. 3, 502 – 511.
87. J. L. Bravo, J. R. Fair, *Hydrocarbon Process.* **65** (1986) no. 3, 45 – 49.
88. R. Billet: *Proceedings of the First Glitsch Packed Columns Workshop,* Dallas TX 1985.
89. J. E. Buchanan, *Ind. Eng. Chem. Fundam.* **6** (1967) 401 – 407.
90. E. Blaß, R. Kurz, *Vt Verfahrenstechnik* **10** (1976) no. 11, 721 – 724; *Vt Verfahrenstechnik* **11** (1977) no. 1, 44 – 48.
91. H. Gelbe: "Der Flüssigkeitsinhalt und die Rektifizierwirkung beim Vakuumbetrieb in Füllkörperschüttungen,"*Fortschr. Ber. VDI Z. Reihe 3,* **23** (1968).
92. A. Mersmann, A. Deixler, *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) no. 5, 265 – 276.
93. T. K. Sherwood, G. H. Shipley, F. A. L. Holloway, *Ind. Eng. Chem.* **30** (1938) 765 – 769.
94. W. E. L. Lobo, L. Friend, F. Hushmall, F. A. Zenz, *Trans. Am. Inst. Chem. Eng.* **41** (1945) 693 – 710.
95. J. S. Eckert, *Chem. Eng. Prog.* **66** (1970) no. 3, 39 – 44.
96. A. Mersmann, *Chem. Ing. Tech.* **37** (1965) no. 3, 218 – 226.
97. J. E. Buchanan, D. C. Dixon: *Correlation of Flooding in Packed Columns and the Generalized Pressure Drop Relation,* 1987 in press.
98. J. B. Zech, A. Mersmann, *Chem. Ing. Tech.* **50** (1978) no. 7, 604.
99. M. G. Shi, A. Mersmann, *Ger. Chem. Eng. (Engl. Transl.)* **8** (1985) 87 – 96.
100. H. Brauer: *Stoffaustausch einschließlich chemischer Reaktion,* Sauerländer, Aarau 1971.
101. J. B. Zech: *Flüssigkeitsströmung und Stoffaustausch in berieselten Füllkörperschüttungen,* Dissertation, TU München 1978.
102. J. Stichlmair, A. Stemmer: *Inst. Chem. Eng. Symp. Ser.* **104** (1987) B213 – B224.

Sublimation

JOHN W. MULLIN, University College London, Torrington Place, London WC1E 7JE, United Kingdom

1. Introduction 1
2. Phase Equilibria 2
3. Vaporization and Condensation . . 4
4. Sublimation Processes 5
4.1. Simple and Vacuum Sublimation . . 5
4.2. Entrainer Sublimation 5
4.2.1. Process Description 5
4.2.2. Product Yield 6
4.2.3. Comparison of Entrainer Sublimation and Crystallization 8
4.3. Fractional Sublimation 8
5. Sublimation Equipment 9
5.1. Vaporizers 9
5.2. Condensers 10
6. References 11

Symbols

d	diameter, m
D	diffusivity, m^2/s
G	mass flow rate of inert gas, kg/h
M	molecular mass
p	partial pressure, Pa
P	pressure, vapor pressure, Pa
S	mass sublimation rate, kg/h
u	vapor velocity, m/s
v	molar volume, m^3/mol
V	vaporization rate, $kg\,m^{-2}\,s^{-1}$
x	effective film thickness, m
Re	Reynolds number ($=ud/\nu$)
Sc	Schmidt number ($=\nu/D$)
α	evaporation coefficient
λ	latent heat, J/mol
ν	kinematic viscosity, m^2/s

Subscript

f	fusion
g	gas
l	liquid
s	solid, sublimation, surface
v	vaporization

1. Introduction

The term sublimation refers strictly to the phase change:
Solid → Vapor
when the transformation occurs without intervention of the liquid phase. In industrial applications, however, the term sublimation usually includes the reverse process, i.e., condensation (desublimation):
Solid → Vapor → Solid
In practice, vaporization of a substance from the liquid state is sometimes desirable; therefore, the complete series of phase changes in an industrial sublimation process can be:
Solid → Liquid → Vapor → Solid
Appearance of the liquid phase is specifically prohibited on the condensation side of the process: the supersaturated vapor must condense directly to the crystalline solid state.

Common organic compounds that can be purified by sublimation include:

2-aminophenol	isophthalic acid
anthracene	naphthalene
anthranilic acid	2-naphthol
anthraquinone	phthalic anhydride
benzanthrone	phthalimide
benzoic acid	pyrogallol
1,4-benzoquinone	salicylic acid
camphor	terephthalic acid
cyanuric chloride	thymol

Amenable elemental and inorganic substances include:

aluminum chloride	iron(III) chloride
arsenic	magnesium
arsenic(III) oxide	molybdenum trioxide
calcium	sulfur
chromium(III) chloride	titanium tetrachloride
hafnium tetrachloride	uranium hexafluoride
iodine	zirconium tetrachloride

In addition, the sublimation of ice in freeze drying has become an important operation in the biological and food industries (→ Drying of Solid Materials, Chap. 2.5.; → Foods, 2. Technology, Chap. 2.4.2.).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_05

Reviews of the industrial applications of sublimation techniques are given in [1–6], of the principles underlying vaporization and condensation in [7], and of techniques for growing crystals from the vapor phase in [8, 23] (see also → Crystal Growth, Chap. 8.).

2. Phase Equilibria

A sublimation process is controlled primarily by the conditions under which phase equilibria occur in a single-component system, i.e., in the sublimable substance. The phase diagram of a simple one-component system shown in Figure 1 depicts the equilibria between solid, liquid, and gaseous states. The diagram is divided into three zones by (1) the *sublimation curve* AT, which records the vapor pressure of the solid, (2) the *vaporization curve* TB, which records the vapor pressure of the liquid, and (3) the *fusion curve* TC, which records the effect of pressure on the melting point. In virtually all one-component systems, the fusion curve slopes to the right, i.e., in the direction of increasing temperature, indicating that the melting point of the solid phase increases with pressure. However, water – an extremely important one-component system – is very unusual in that increasing pressure decreases the melting point of ice.

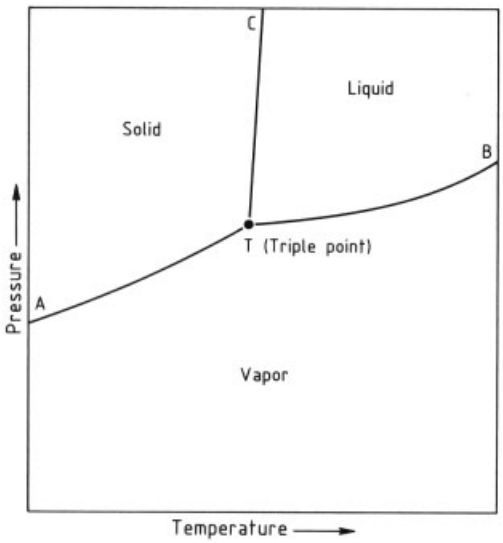

Figure 1. Phase diagram for a single-component system

The slopes of the sublimation, vaporization, and fusion curves in Figure 1 can be expressed quantitatively by the Clapeyron equation:

$$\left(\frac{\mathrm{d}P}{\mathrm{d}T}\right)_{\mathrm{sub}} = \frac{\lambda_{\mathrm{s}}}{T\,(v_{\mathrm{g}} - v_{\mathrm{s}})} \tag{1}$$

$$\left(\frac{\mathrm{d}P}{\mathrm{d}T}\right)_{\mathrm{vap}} = \frac{\lambda_{\mathrm{v}}}{T\,(v_{\mathrm{g}} - v_{\mathrm{l}})} \tag{2}$$

$$\left(\frac{\mathrm{d}P}{\mathrm{d}T}\right)_{\mathrm{fus}} = \frac{\lambda_{\mathrm{f}}}{T\,(v_{\mathrm{l}} - v_{\mathrm{s}})} \tag{3}$$

where P is the vapor pressure, T the absolute temperature, and v_{s}, v_l, and v_{g} are the molar volumes of the solid, liquid, and gas phases, respectively. The latent heats (enthalpies) of sublimation, vaporization, and fusion (λ_{s}, λ_{v}, and λ_{f}, respectively) are related at a given temperature by

$$\lambda_{\mathrm{s}} = \lambda_{\mathrm{v}} + \lambda_{\mathrm{f}} \tag{4}$$

Analysis of sublimation–desublimation is often hindered by the scarcity of published physical data. If, however, only two reliable values of vapor pressure are available, a considerable amount of information can be derived from the Clausius–Clapeyron equation:

$$\ln\left(\frac{P_1}{P_2}\right) = \frac{\lambda_{\mathrm{v}}\,(T_1 - T_2)}{RT_1T_2}$$

where the gas constant $R = 8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}$.

For example, the vapor pressures of naphthalene at 190 and 160 °C are 780 and 220 Pa, respectively. Equation (5) can be used to estimate a mean value of λ_{v}. If the assumption is made that λ_{v} does not vary greatly over the temperature range considered, λ_{v} can then be used to estimate the vapor pressure at some other temperature, e.g., 120 °C. Thus,

$$\ln\left(\frac{780}{220}\right) = \frac{\lambda_{\mathrm{v}}\,(463 - 433)}{8.314 \times 463 \times 433}$$

which gives $\lambda_{\mathrm{v}} = 70\ 340$ J/mol

and

$$\ln\left(\frac{220}{P}\right) = \frac{70340\,(433 - 393)}{8.314 \times 433 \times 393}$$

which gives $P = 30$ Pa at 120 °C

The three curves in Figure 1 intersect at the *triple point* T; this represents the temperature and pressure at which the solid, liquid, and gas

phases coexist in equilibrium. The position of the triple point in the phase diagram is of the utmost importance in sublimation–desublimation processes. If it occurs at a pressure above atmospheric, the solid cannot melt under normal atmospheric conditions, and true sublimation (i.e., the transition solid → vapor) is easily achieved. For example, the triple point for carbon dioxide occurs at − 57 °C and 500 kPa; thus, liquid CO_2 is not formed when solid CO_2 is heated at atmospheric pressure (101.3 kPa), and the solid simply vaporizes. If the triple point occurs at a pressure less than atmospheric, however, certain precautions are necessary if the phase changes solid → vapor (sublimation) and vapor → solid (desublimation) are to be controlled. For example, because the triple point for water is 0.01 °C and 0.6 kPa, ice melts when it is heated above 0 °C at atmospheric pressure. For ice to sublime, both the temperature and the pressure must be kept below the triple-point values.

From the industrial point of view, desublimation (crystallization) is the most important step; this means that the equipment responsible for condensation must operate below the triple point. If the solid → liquid stage is permitted before vaporization, the operation is often called *pseudosublimation.*

Both true sublimation and pseudosublimation cycles are depicted on the phase diagram in Figure 2. For a substance with a triple point (T) at a pressure greater than atmospheric, true sublimation (i.e., solid → vapor) occurs. The complete cycle is given by path ABCDE. The original solid, represented by point A, is heated to a temperature represented by point B. The increase in vapor pressure of the substance is traced along the sublimation curve from A to B. The condensation part of the process is represented by the broken line BCDE. As the vapor passes from the vaporizer to the condenser, it may cool slightly and be diluted with an inert gas such as air. Therefore, point C, which represents a temperature and partial pressure slightly lower than point B, can be taken as the condition at the condenser inlet. After entering the condenser, the vapor is mixed with more inert gas, and the partial pressure of the substance and its temperature drop to a point D. Thereafter, the vapor cools essentially at constant pressure to the conditions represented by point E, the temperature of the condenser.

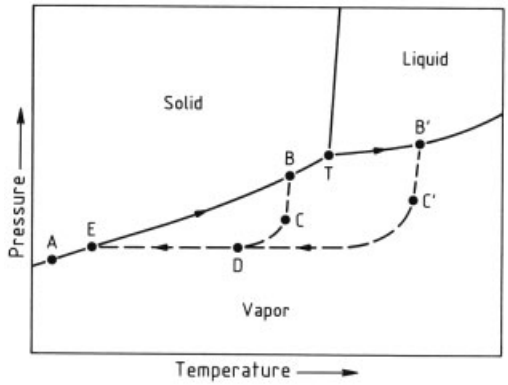

Figure 2. Phase diagram showing true sublimation (ABCDE) and pseudosublimation (AB′C′DE) cycles

When the triple point of the substance occurs at a pressure lower than atmospheric, heating may easily result in the temperature and vapor pressure of the solid exceeding triple-point conditions. The solid will then melt in the vaporizer; path AB′ represents such a process. Great care must be taken in the condensation stage; the partial pressure of the substance in the vapor stream entering the condenser must be reduced below the triple-point pressure to prevent initial condensation to a liquid. The required partial pressure reduction can be brought about by diluting the vapor with an inert gas, but the frictional pressure drop in the vapor line is often sufficient. Point C′ represents the conditions at the point of entry into the condenser; the condensation path is represented by C′DE.

Fractional Sublimation. If two or more sublimable substances form true solid solutions, their separation by fractional sublimation is theoretically possible. The phase diagram for a binary solid-solution system at a pressure below the triple-point pressures of the two components is shown in Figure 3. Points A and B represent the equilibrium sublimation temperatures of pure components A and B, respectively, at a given pressure. The lower curve represents the sublimation temperatures of mixtures of A and B, while the upper curve represents the solid-phase condensation temperatures, generally called *snow points.* Figure 4 shows that if a solid solution (e.g., point S) is partially sublimed (e.g., heated to some temperature represented by point X), the resulting vapor phase (point Y) and residual solid (point Z) contain different propor-

tions of the original components, quantified by the *mixture rule* (→ Crystallization and Precipitation). The sublimate and the residual solid may each be resubjected to this procedure; therefore, the possibility of fractionation exists. However, the practical difficulties in performing such a separation may be considerable.

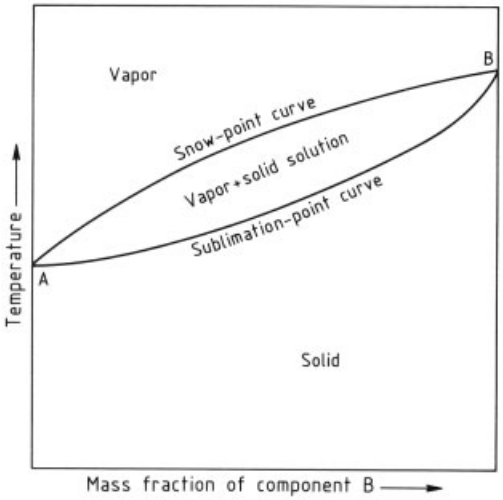

Figure 3. Phase diagram for a two-component solid-solution system at a pressure below the triple points of the two components A and B

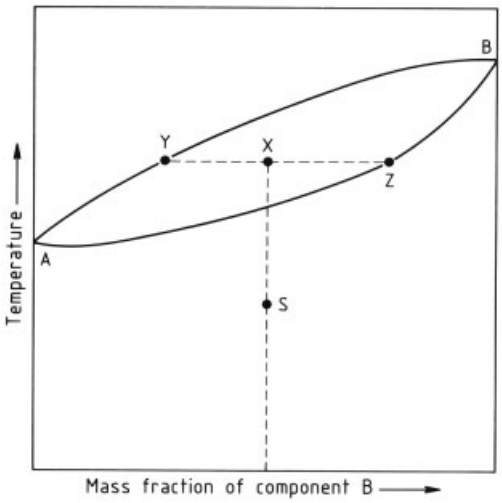

Figure 4. Fractional sublimation of two components A and B demonstrated on a solid-solution phase diagram

Experimental studies of fractional sublimation are described in [9–11]; nucleation and growth rates of organic condensates, as well as their application to the fractionation of mixed vapors, are reported in [12].

3. Vaporization and Condensation

Vaporization. The maximum theoretical vaporization rate V (kg m^{-2} s^{-1}) from the surface of a pure liquid or solid under its own vapor pressure is given by the Hertz–Knudsen equation [13], which can be derived from the kinetic theory of gases:

$$V = P_s\,(M/2\pi RT_s)^{1/2} \tag{6}$$

where P_s is the vapor pressure (in Pa), at the surface temperature T_s (in K), M is the molecular mass, and R is the gas constant (8.314 J mol^{-1} K^{-1}).

In practice, however, the actual vaporization rate may be slower than predicted by Equation (6), and a correction factor α (≤ 1), generally referred to as an evaporation coefficient, is conventionally included:

$$V = \alpha P_s\,(M/2\pi RT_s)^{1/2} = 52.2\alpha P_s\,(M/T_s)^{1/2} \tag{7}$$

A laboratory technique used to measure values of α for sublimable solid materials is described in [14]. Sublimation rates of pure solids into turbulent air streams have been successfully correlated by the Gilliland–Sherwood equation [11]:

$$d/x = 0.023Re^{0.38}Sc^{0.44} \tag{8}$$

where d is a characteristic dimension of the vaporization chamber, x is the thickness of the effective film for mass transfer at the vapor–solid interface, and Re and Sc are the dimensionless Reynolds and Schmidt numbers, respectively.

Condensation (desublimation) is generally a transient operation, in which simultaneous heat and mass transfer are further complicated by the effects of spontaneous condensation in the bulk gaseous phase [15]. Several steps, which may or may not be mutually dependent, can be involved in the condensation of a solid phase from a vapor. The first step after the creation of supersaturation in the vapor phase is nucleation. Nucleation may be homogeneous but is predominantly heterogeneous under most circumstances; it is followed by both crystal growth

and agglomeration in the formation of the final crystal product (→ Crystallization and Precipitation, → Crystallization and Precipitation, → Crystallization and Precipitation). As a rate process, the condensation of solids from vapors is less well understood than vaporization [7]. A simple laboratory technique for measuring kinetics in subliming systems is described in [16]; it has been used to compare the solid evaporation and crystal growth rates of benzophenone under comparable conditions.

The two most common ways of creating the supersaturation necessary for crystal nucleation and subsequent growth are (1) cooling via a metal surface (which can lead to either a glassy or a multicrystalline deposit that necessitates mechanical removal) and (2) dilution (e.g., with an inert gas) which, under suitable conditions, can produce a loose crystalline mass that is easy to handle.

The measurement and correlation of heat-transfer rates during condensation in a pilot plant connected to an industrial phthalic anhydride desublimation unit are described in [17]. A comprehensive study of the condensation of several sublimable materials in a fluidized bed is reported in [18]; an extensive study on heat- and mass-transfer processes in a fluidized-bed desublimation unit is summarized in [19]. A desublimation mathematical model to predict optimum cooling conditions for smooth- and fin-tube heat exchangers is proposed in [24] and comparisons are made with industrial-scale data for phthalic anhydride.

4. Sublimation Processes

Sublimation techniques can be classified conveniently into three types: simple, vacuum, and entrainer.

4.1. Simple and Vacuum Sublimation

Simple sublimation is a batchwise process in which the solid material is heated and vaporized. The resulting vapor diffuses toward a condenser, the driving force for diffusion being the partial-pressure difference between the vaporizing and condensing surfaces. The vapor path between the vaporizer (sublimer) and the condenser should be as short as possible to reduce resistance to flow. Simple sublimation has been practiced for centuries, often in very crude equipment, for the commercial production of ammonium chloride, iodine, and "flowers" of sulfur.

Vacuum sublimation is a natural development of simple sublimation, particularly if the triple-point pressure of the substance is lower than atmospheric. The transfer of vapor from the vaporizer to the condenser is enhanced by reducing the pressure in the condenser and, consequently, increasing the partial-pressure driving force. Iodine, pyrogallol, and many metals have been purified by vacuum sublimation. Exit gases from the condenser usually pass through a cyclone or scrubber to protect the vacuum equipment and minimize product loss.

4.2. Entrainer Sublimation

4.2.1. Process Description

In entrainer sublimation, an inert gas (entrainer) is blown into the vaporization chamber of a sublimer to increase the vapor flow rate to the condensing equipment and, thus, increase the yield. Air is the most commonly used entrainer, but superheated steam can be employed for substances that are relatively insoluble in water such as anthracene. If steam is used, the vapor may be cooled and condensed by direct contact with a spray of cold water. Efficient recovery of the sublimate is thus obtained, but the product is wet.

The use of an entrainer gas in a sublimation process has many desirable features. As already mentioned, it enhances the flow of vapor from the sublimer to the condenser. It also provides the heat needed for sublimation and, thus, an efficient means of temperature control. The gas can provide the dilution, if necessary, for fractional condensation at the desublimation stage. Entrainer sublimation, whether by gas flow over a static bed of solid particles or through a fluidized bed, is ideally suited to continuous operation.

The purification of salicylic acid [*69-72-7*] is a good example of the industrial use of entrainer sublimation. Air may be used as the carrier gas, but because salicylic acid can be decarboxylated in hot air, a mixture of air and CO_2 is often preferred. A batchwise process is shown in Figure 5.

A 5 – 10 vol% mixture of CO_2 in air is recycled through the plant; it passes over heater coils (b) and then over the containers (c) (bins or trays) of impure salicylic acid in the vaporizer (d). The vapor is then conducted through a series of air-cooled condensers (e), where the sublimed salicylic acid is deposited. A trap (f) removes any entrained sublimate from the gas stream before it is returned to the heater. Makeup CO_2 and air are introduced as required. The process continues until the containers are emptied of all volatile matter.

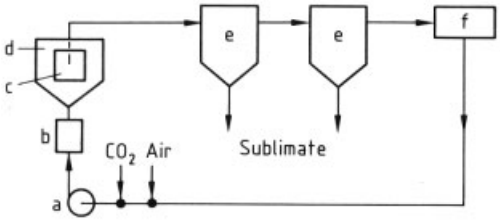

Figure 5. Entrainer batch sublimation process for the purification of salicylic acid [3]
a) Pump; b) Heater; c) Salicylic acid container; d) Vaporizer; e) Condenser; f) Trap

A general-purpose, continuous entrainer sublimation plant is shown in Figure 6. The impure feedstock is pulverized in a mill (b), if necessary, a suitable entrainer gas (e.g, hot air) blows the fine particles, which volatilize readily, into a series of separators (c) (e.g., cyclones) where nonvolatile solid impurities are removed. A filter (d) may also be located in the vapor line to remove final traces of inert, solid impurities. The vapor then passes to a series of condensers (e) from which the sublimate is subsequently discharged. Exhaust gases can be recycled or passed to the atmosphere through a cyclone or wet scrubber to recover any entrained product.

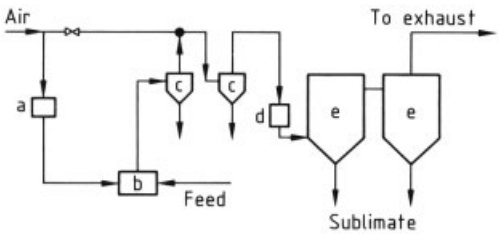

Figure 6. General-purpose continuous sublimation unit [3]
a) Heater; b) Mill; c) Separator; d) Filter; e) Condenser

The application of fluidization techniques to sublimation–desublimation processes was first proposed by MATZ with the objective of enhancing both heat and mass transfer [20]. The technique has not yet been widely adopted for large-scale commercial use, but a fluidized-bed desublimation unit operating in the United States has been reported to produce aluminum chloride at a rate of ca. 10 t/h [21].

4.2.2. Product Yield

The yield from an entrainer sublimation process can be estimated as follows. The mass flow rate G of the inert gas and the mass sublimation rate S are related by

$$G/S = \varrho_g p_g / \varrho_s p_s \tag{9}$$

where p_g and p_s are the partial pressures of the inert gas and vaporized substance, respectively, in the vapor stream, and ϱ_g and ϱ_s are their respective vapor densities. The total pressure P of the system is the sum of the partial pressures of the components:

$$P = p_g + p_s \tag{10}$$

Thus, Equation (9) can be written as

$$S = G\left(\frac{\varrho_s}{\varrho_g}\right)\left(\frac{p_s}{P - p_s}\right) \tag{11}$$

or, in terms of the molecular masses of the inert gas, M_g, and of the material being sublimed, M_s

$$S = G\left(\frac{M_s}{M_g}\right)\left(\frac{p_s}{P - p_s}\right) \tag{12}$$

The theoretical maximum yield from an entrainer sublimation process is the difference between the calculated sublimation rates corresponding to the conditions in the vaporization and condensation stages.

For example, salicylic acid ($M_s = 138$) is to be purified by entrainer sublimation with air ($M_g = 29$) at 150 °C. The vapor is fed at atmospheric pressure (101.5 kPa) to a series of condensers, the internal temperature and pressure of the last condenser being 40 °C and 100 kPa, respectively. The air flow rate is 2000 kg/h, and the pressure drop between the vaporizer and the last condenser is 1.5 kPa. The vapor pressures of salicylic acid at 150 and 40 °C are 1.44 and 0.0023 kPa, respectively. Therefore, under saturated conditions:

Vaporization stage ($P = 101.5$ kPa, $p_s = 1.44$ kPa):

$$S_v = 2000\left(\frac{138}{29}\right)\left(\frac{1.44}{101.5 - 1.44}\right) = 137\text{kg/h}$$

Condensation stage ($P = 100$ kPa, $p_s = 0.0023$ kPa):

$$S_c = 2000\left(\frac{138}{29}\right)\left(\frac{0.0023}{100 - 0.0023}\right) = 0.22\text{kg/h}$$

In this example, therefore, the loss from the condenser exit gases is only 0.22 kg/h, and the theoretical maximum yield is ca. 137 kg/h. This maximum yield is obtained, however, only if the air is saturated with salicylic acid vapor at 150 °C, and saturation is approached only if the air and salicylic acid are in contact for a sufficient period of time at the required temperature. A fluidized-bed vaporizer may allow realization of these optimum conditions, but if air is simply blown over bins or trays containing the solid, saturation will not be reached and the actual rate of sublimation will be lower than calculated. In some cases, the degree of saturation achieved may be as low as 10 % of the calculated value.

The calculated loss of product in the condenser exit gases is only a minimum value. Other losses resulting from entrainment of solids depend on the condenser design and cannot be calculated. An efficient exit-gas scrubber can, of course, minimize these losses.

4.2.3. Comparison of Entrainer Sublimation and Crystallization

An analysis of the commercial purification of naphthalene [*91-20-3*] suggests that sublimation is potentially more economical than conventional melt crystallization [6] (for a description of melt crystallization, see →Crystallization and Precipitation). A flow sheet for entrainer sublimation is given in Figure 7, and this process is compared economically with crystallization in Table 1. In the proposed sublimation method, which has not yet been scaled up to commercial size, the solid feedstock is completely va-

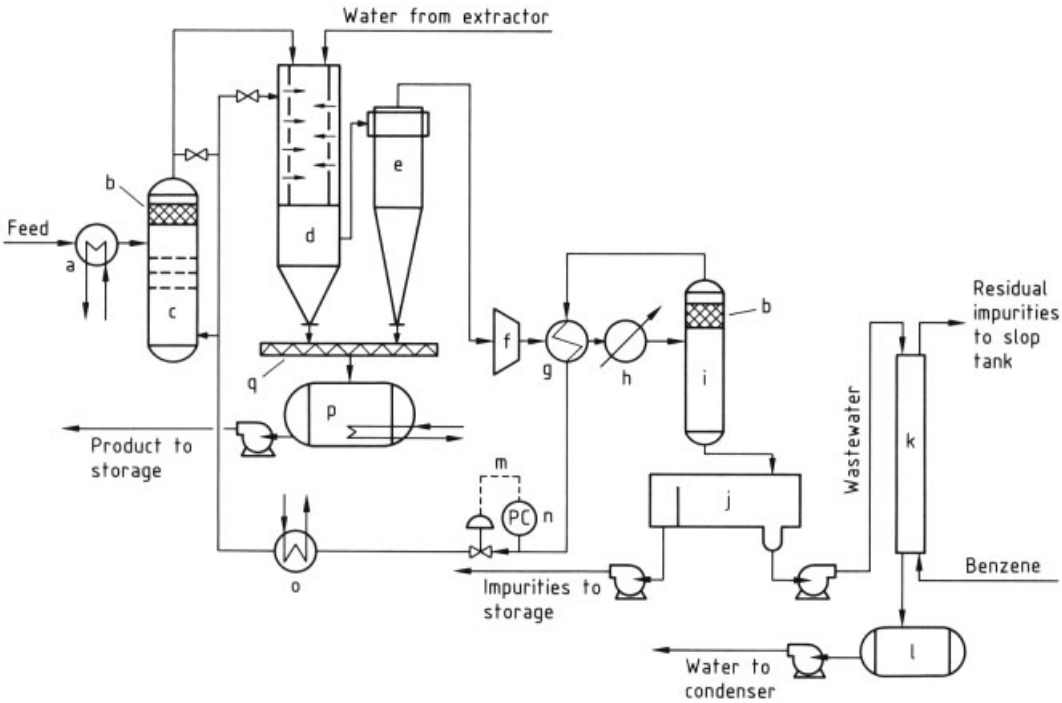

Figure 7. Proposed entrainer sublimation process for the purification of naphthalene [6]
a) Feed heater; b) Mist eliminator; c) Sublimer; d) Condenser; e) Cyclone; f) Circulation compressor; g) Heat exchanger; h) Cooler; i) Liquid–gas separator; j) Impurity–water separator; k) Extractor with rotating disks; l) Tank for injected water; m) Instrument line; n) Pressure controller; o) Entrainer-gas heater; p) Product tank; q) Screw conveyor

porized in an entrainer gas (nitrogen) stream and then partially condensed. Heat is removed by vaporizing water at the top of the condenser (d). To prevent deposition of sublimate on vessel walls, the inner wall of the condenser is permeable and allows some of the entrainer gas to pass through. Impurities remain in the vapor stream and are subsequently condensed in the cooler (h) located after the entrainer circulation compressor (f). The impurity–wastewater stream leaving the separator (j) is washed with benzene in an extractor (k).

Table 1. Comparison of entrainer sublimation with conventional melt crystallization for the production of naphthalene [6]

Process parameter	Crystallization	Sublimation
Naphthalene content of feedstock, wt %	92.5	90.0
Naphthalene content of product, wt %	99.6	99.6
Yield of naphthalene, %	78.0	93.7
Naphthalene production rate, kg/h	1000	1000
Steam used per tonne of naphthalene, kg	200	320
Electricity used per tonne of naphthalene, kW · h	70	119
Cooling water used per ton of naphthalene, m^3	200	14
Space requirement, m^3	7500	500
Investment cost (1983), 10^6 \$	1.75	0.62

As shown in Table 1, sublimation gives a higher yield of naphthalene at a lower cost, and with a smaller space requirement, than crystallization. Although steam and electricity consumption is higher for sublimation, this is offset by a much lower cooling-water requirement.

4.3. Fractional Sublimation

As mentioned in Chapter 2, the separation of two or more sublimable substances by fractional sublimation is theoretically possible if the substances form true solid solutions. However, large-scale commercial exploitation of fractional sublimation has not yet been reported. A laboratory-scale process known as thin-film fractional sublimation has been developed and is shown schematically in Figure 8 [9]. This process has been applied successfully to the separation of volatile solid mixtures such as hafnium and zirconium tetrachlorides, 1,4-dibromobenzene and 1-bromo-4-chlorobenzene, and anthracene and carbazole. A stream of inert, nonvolatile solids (e.g., glass beads or sand) is fed to the top of a vertical fractionation column (b) and falls countercurrently to the rising supersaturated vapor. Vapor movement is facilitated by an upwardly flowing entrainer gas stream. The temperature of the incoming solids is maintained well below the snow-point temperature of the vapor. Consequently, the solids become coated with a thin film (10 μm) of sublimate, which acts as a reflux for the enriching section of the column above the feed entry point.

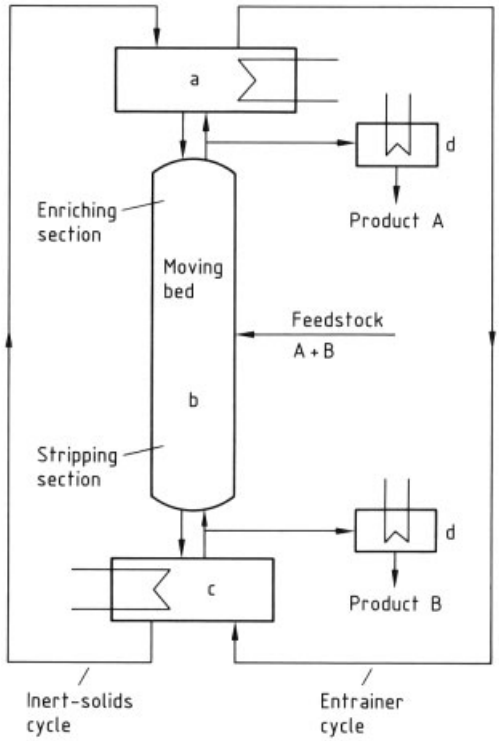

Figure 8. Thin-film fractional sublimation column [9]
a) Reflux condenser; b) Fractionation column; c) Vaporizer; d) Condenser

5. Sublimation Equipment

Very few standard forms of sublimation or desublimation equipment are in common use. Most industrial units, particularly on the condensation side of the process, have been developed on an ad hoc basis for a specific substance

and purpose. The most useful source of information on sublimation equipment is the patent literature, but this has the severe drawback of not indicating whether a process has been, or even can be, put into practice [4,5].

5.1. Vaporizers

A variety of vaporization units have been used or proposed for large-scale operation, depending on the manner in which the solid feedstock is to be vaporized. Examples are [4]:

1) a bed of dry solids without entrainer gas (the simplest arrangement);
2) dry solids suspended in a dense nonvolatile, liquid;
3) solids suspended in a boiling (entrainer) liquid where the entrainer gas is formed in situ;
4) entrainer gas flowing through a fixed bed of solid particles;
5) entrainer gas bubbling through molten feedstock; i.e., vaporization takes place above the triple-point pressure;
6) entrainer gas flowing through a dense phase of solid particles in a fluidized bed;
7) entrainer gas flowing through a dilute phase of solid particles (e.g., in a transfer-line vaporizer where the solid and gas phases are in concurrent flow, or in a raining solids unit where the solids and entrainer may be in countercurrent flow);
8) vacuum sublimation where the vapour phase consists essentially of the sublimate; e.g., vaporization takes place below the triple-point pressure without entrainer gas in an agitated vaporizer with large self-cleaning heat exchange surfaces [25].

5.2. Condensers

Sublimate condensers are usually large, air-cooled chambers which tend to have very low heat-transfer coefficients (generally 5–10 $W\,m^{-2}\,K^{-1}$). This is expected because sublimate deposits on the condenser walls act as an insulator. In addition, vapor velocity in the chambers is generally very low. Quenching the vapor with cold air in the chamber may increase the rate of heat removal, but excessive nucleation is likely and the product crystals will be very small. Condenser walls may be kept free of solid by using internal scrapers, brushes, and other devices. All vapor lines in sublimation units should have a large diameter, be adequately insulated, and if necessary, be provided with supplementary heating to minimize blockage due to the buildup of sublimate.

One of the main hazards of air-entrainment sublimation is the risk of fire; many substances that are considered quite safe in their normal state can produce explosive mixtures with

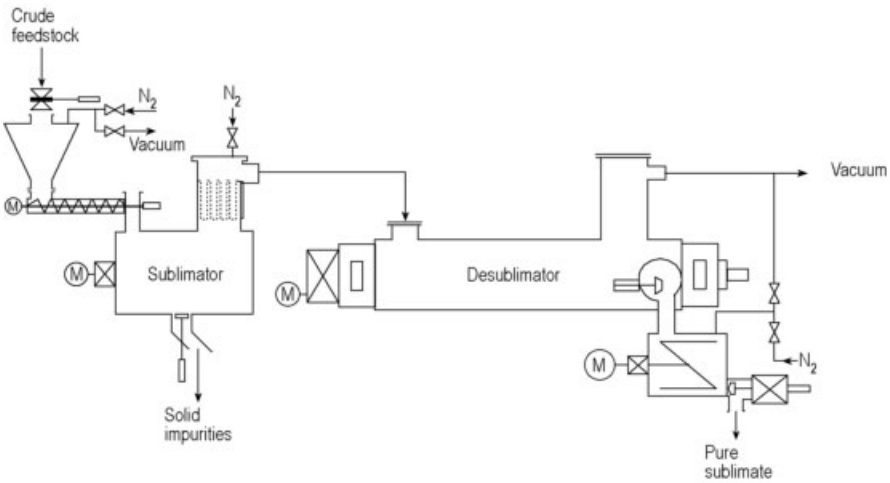

Figure 9. Semi-continuous DISCOTHERM B vacuum sublimation unit [26]

air. All electrical equipment should be flameproof, and all parts of the plant should be efficiently grounded (earthed) to avoid buildup of static electricity. Vacuum operation after nitrogen purging can provide a much safer environment.

The totally enclosed production unit (Fig. 9) is provided with self-cleaning heat exchange surfaces and operates semi-continuously under reduced pressure without a carrier gas. Accumulated impurities are discharged from the sublimator periodically. Batch and continuous modifications of this unit are available [26]. Sublimation units as shown in Figure 9 have been successfully applied industrially for the purification of dyestuff intermediates, anthraquinone, metal-organic compounds and pharmaceutical products with capacities from 300 – 10 000 t/a.

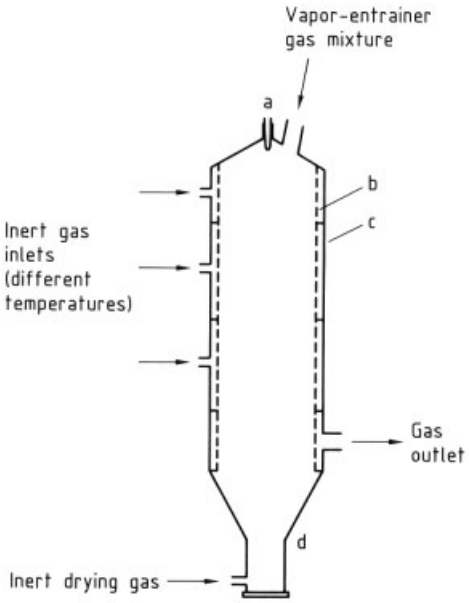

Figure 10. Sublimation chamber with gas-permeable walls [10]
a) Water jet; b) Inserted porous cylinder; c) Jacket; d) Product collection zone

Calculation of the density of deposited layers of sublimate, and of associated variables, as an aid towards the optimization of sublimate condenser design, is discussed in [22]. The starting point of the analysis is the assumption that the growth of sublimate layers is governed mainly by heat and mass transfer; the model is based on conditions in the diffusion boundary layer. The main process-determining factors (growth rate, mass transfer, and gas concentration) are accounted for. The derived theoretical relationship is shown to fit experimental data.

In a variant of the large-chamber desublimation condenser, the crystallization chamber is fitted with gas-permeable walls (Fig. 10) [10]. Vapor and entrainer gas are cooled by evaporation of water dispersed in the cooling space formed by the porous walls. An inert gas passes through the porous walls into the cooling space and protects the internal walls from solid deposits. Crystallization take place in the bulk vapor–gas mixture by direct contact with the dispersed water. This system has been used, for example, for partial separation of a mixture of phthalic anhydride and naphthalene by using nitrogen as the entrainer.

Fluidized-bed condensers (desublimation units) have been considered for large-scale application, although most of the published reports are concerned with laboratory-scale investigations [19].

6. References

1. G. Matz: *Kristallisation,* 2nd ed., Springer Verlag, Berlin 1969.
2. *Ullmann,* 4th ed., **2,** pp. 664 – 671.
3. J. W. Mullin: *Crystallization,* 4th. ed., Butterworths-Heinemann, Oxford2001, pp. 323 – 332.
4. C. A. Holden, H. S. Bryant, *Sep. Sci.* **4** (1969) no. 1, 1 – 13.
5. S. D. Kemp in H. W. Cremer, T. Davies (eds.): *Chemical Engineering Practice, "Sublimation and Vacuum Freeze Drying,"* **vol. 6**, Butterworths, London 1958, pp. 567 – 600.
6. L. Kudela, M. J. Sampson, *Chem. Eng. (N.Y.)* **93** (1986) no. 12, 93 – 98.
7. E. Rutner, P. Goldfinger, J. P. Hirth (eds.): *Condensation and Evaporation of Solids,* Gordon and Breach, New York 1964.
8. M. M. Faktor, I. Garrett: *Growth of Crystals from the Vapour,* Chapman and Hall, London 1974.
9. J. Gillot, W. M. Goldberger, *Chem. Eng. Prog. Symp. Ser.* **65** (1969) no. 91, 36 – 42.
10. J. Vitovec, J. Smolik, J. Kugler, *Collect. Czech. Chem. Commun.* **43** (1978) 396 – 400.

11. H.-H. Eggers, D. Ollmann, D. Heinz, D. W. Drobot, A. W. Nikolajew, *Z. Phys. Chem. (Leipzig)* **267** (1986) 353 – 363.
12. M. Matsouka in S. J. Jancic, E. J. de Jong (eds.): *Rates of Nucleation and Growth of Organic Condensates of Mixed Vapours, Industrial Crystallization* **84**, Elsevier, Amsterdam 1984, pp. 357 – 360.
13. T. K. Sherwood, C. Johannes, *AIChE J.* **8** (1962) 590 – 593.
14. A. C. Plewes, J. Klassen, *Can. J. Technol.* **29** (1951) 322.
15. H. Ueda, Y. Takashima, *J. Chem. Eng. Japan* **10** (1977) 6 – 12.
16. R. F. Strickland-Constable: *Kinetics and Mechanism of Crystallization,* Academic Press, London 1968.
17. J. Bilik, R. Krupiczka, *Chem. Eng. J.* **26** (1983) 169 – 180.
18. J. Ciborowski, S. Wronski, *Chem. Eng. Sci.* **17** (1962) 481 – 489.
19. M. Knuth, P.-M. Weinspach, *Chem. Ing. Tech.* **48** (1976) 893.
20. G. Matz, *Chem. Ing. Tech.* **30** (1958) 319 – 329.
21. V. Cedro, letter to *Chem. Eng.* (*N.Y.)* **93** (1986) no. 21, 5.
22. K. Wintermantel, H. Holzknecht, P. Thoma, *Chem. Eng. Technol.* **10** (1987) 205 – 210.
23. E. Krabbes in J. P. van der Eerden, O. S. L. Bruinsma (eds.): *Science and Technology of Crystal Growth,* Kluwer, Dordrecht 1995, pp. 123 – 136.
24. R. Krupiczka, J. Pyschny, *Chem. Eng. and Processing* **28** (1990) 29 – 34.
25. W. Schwenk, G. Raouzeos, *Chem. Technol. Europe* **2** (1995) no. 2, 14 – 17.
26. List AG., company brochure, Switzerland.

Substitute Natural Gas → Gas Production
Succinic Acid → Dicarboxylic Acids, Aliphatic

Liquid – Liquid Extraction

Eckart Müller, Frankfurt a. Main, Germany (Chaps. 1 and 2)

Rosemarie Berger, Ludwigshafen, Germany (Chaps. 3 and 4)

Eckhart Blass, Technische Universität München, München, Germany (Chaps. 1 – 4, Sections 5.1 and 5.2)

Domien Sluyts, Bayer AG, Antwerpen, Belgium (Section 5.3)

1. Fundamentals and Fields of Application . . . 2
2. Thermodynamic Fundamentals . . . 4
2.1. Graphic Representation of Phase Equilibria . . . 4
2.1.1. Triangular Diagrams . . . 4
2.1.2. Other Graphic Representations . . . 6
2.2. Measuring Methods of Phase Equilibria . . . 7
2.3. Empirical Correlations of Phase Equilibria . . . 9
2.4. Selection of Solvent . . . 10
2.5. Calculation of Theoretical Stages . 11
2.5.1. Selection of Calculation Methods . . . 11
2.5.2. Specific Methods . . . 11
2.5.3. Evaluation of Stage-Number Calculation for Process Design . . . 13
3. Apparatus . . . 15
3.1. Survey . . . 15
3.1.1. Columns without Energy Input . . . 15
3.1.2. Pulsed Columns . . . 17
3.1.3. Columns with Rotating Internals . . . 20
3.1.4. Mixer – Settlers . . . 21
3.1.5. Centrifugal Extractors . . . 23
3.2. Fluid-Dynamic Fundamentals . . . 24
3.2.1. Problems and Process Strategy . . . 24
3.2.2. Operating Characteristics of Pulsed Columns and Columns with Rotating Internals . . . 25
3.2.3. Fluid Dynamic Calculation Methods . 28
3.3. Apparatus Design . . . 30
3.3.1. Internals and Operating Conditions . . 30
3.3.2. Column Diameter . . . 31
3.3.3. Column Height . . . 32
3.4. Criteria for Equipment Selection . . 32
4. Phase-Separation Equipment . . . 36
4.1. Gravity Settlers without Inserts . . 36
4.2. Settlers with Coalescing Aids . . . 37
5. Liquid – Liquid Extraction Processes . . . 41
5.1. General . . . 41
5.2. Combined Processes of Extraction and Distillation . . . 43
5.3. Reactive Extraction . . . 44
5.3.1. Introduction . . . 44
5.3.2. Extraction Mechanism of Different Types of Solvent . . . 45
5.3.3. Uses . . . 48
5.3.4. Setting up an Extraction System . . . 49
5.3.5. Diluents and Modifiers . . . 50
6. References . . . 51

Symbols

$2a$	Pulsing amplitude, m
A	key component of the phase from which the solute is extracted
B	key component of the phase into which the solute is extracted
c_i	constants
C	transferred component
d_i	hole or packing diameter, m
d_p	particle diameter, m
d_{32}	Sauter mean drop diameter, m
D	column diameter, m
D_R	rotor diameter, m
D	distribution coefficient
$E = K\dot{m}_s/\dot{m}_F$	extraction factor
f	pulsing frequency, s^{-1}
g	acceleration due to gravity, m/s^2
G_i	coefficients in empirical correlations of the distribution coefficient

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_06

h	compartment height, m
H	height of column, m
HETS	height equivalent of a theoretical stage, m
$k = y_c/x_c$	distribution ratio of transferred component (in terms of mass fraction)
$K = Y_C/X_C$	distribution ratio of transferred component (in terms of loading)
K_p	plait point
$m_{j\,k}$	mass of component j in phase or flow k (j = A, B, C; k = F, R, S, E), kg
$m_k = m_{Ak} + m_{Bk} + m_{Ck}$	total mass of phase or flow k (k = F, R, S, E), kg
$\dot{m}$	mass flow, kg/h
n_s	revolutions of stirrer, s^{-1}
n	number of stages
NTS	number of theoretical stages
s	column cross section, m
v	superficial velocity, $m^3\ m^{-2}\ h^{-1}$ or m/s
$v = \dot{m}_S/\dot{m}_F$	extractant ratio
v_p	individual drop velocity, m/s
v_r	slip velocity (Eq. 9), m/s
x_C	mass fraction of C in A-rich phase
$X_C = x/(1-x)$	loading of A-rich phase with component C, kg/kg
y_C	mass fraction of C in B-rich phase
$Y_C = y/(1-y)$	loading of B-rich phase with C, kg/kg

Greek symbols

ε	holdup, m^3/m^3
η	dynamic viscosity, $Pa \cdot s \times 10^{-3}$
ϱ	density, kg/m^3
σ	interfacial tension, $N/m \times 10^{-3}$
φ	free area

Dimensionless Groups

$Ar = d_P^3 g \varrho_c \Delta\varrho / \eta_c^2$	Archimedes number
$Fr_R = N^2 D_R / g$	Froude number
$Re_p = v_p d_p / v_d$	Particle Reynolds number
$K_L = \varrho_c^2 \sigma^3 / \left(\Delta\varrho g \eta_c^4\right)$	Characteristic number of liquid

Subscripts

A	key component of the phase from which the solute is extracted
B	key component of the phase into which the solute is extracted
C	transferred substance (solute)
c	continuous phase
d	dispersed phase
E	extract
f	flooding
F	feed
i	internals (characteristic size)
o	organic
p	drop (particle)
R	raffinate
S	solvent
1	pilot scale
2	large scale

1. Fundamentals and Fields of Application

Liquid – liquid extraction or *solvent extraction* is a separation process which is based on the different distribution of the components to be separated between two liquid phases. It depends on the mass transfer of the component to be extracted from a first liquid phase to a second one.

Definitions. The phases form definite contact areas and cannot or can only to a limited extent mix with one another on account of their properties. The phases are composed of individual substances, so-called components, and flow as liquid streams through the extractor (see Fig. 1). Phases are liquid streams that are in equilibrium with each other.

Generally, more than three components are involved in an extraction. To make the description of the problem simpler, *key components* are

defined. The *feed* to a liquid – liquid extraction is the solution that contains the components to be separated. The key component of this liquid phase (the phase from which the solute is extracted) is called A, the transferred substance (the *solute*) C. The (*extraction*) *solvent* is the liquid added to the process to extract the substance C from the feed. The solvent can be the pure component, B, but usually contains small quantities of C and A, because the solvent is usually recycled from a recovery system. The solvent phase leaving the *extractor* is the *extract.* The extract contains mainly B and the extracted component C but also small quantities of A. The *raffinate* is the liquid phase left from the feed after being contacted by the extraction solvent; it is composed mainly of A, but generally small quantities of the extracted component C and solvent B.

Comparison with Distillation. Liquid – liquid extraction has similarities with distillation. In both cases the substances to be separated have concentration ratios $\neq 1$ in the two phases. In the case of distillation the second phase (the vapor phase) is formed exclusively from the components of the initial (liquid) phase, with extraction a new substance, the (extraction) solvent, must be added to form the second (liquid) phase. The solvent enters into the material balance of the process and its behavior towards the substances to be separated is decisive for the separation process.

Generally, the extracted substance is separated from the solvent by distillation and the solvent B is recycled to the extractor. Sometimes the selective action of the solvent is used in the distillation which generates the reflux for the extraction, e.g., *extractive distillation* (extraction of aromatic compounds, see Section 5.2) or *azeotropic distillation* (extraction of acetic acid). In the extraction of metal salts (see Section 5.3) the purity is often achieved by adjusting the pH. Extraction is an isothermal process, normally carried out at ambient temperature and pressure.

Comparison with Absorption. In both absorption and extraction solvents are used which are recycled and thereby take up the required components selectively and then release them again. Selection of the most suitable solvents therefore plays an important role in the design of the two separation processes.

Applications of Liquid – Liquid Extraction. Liquid – liquid extraction is primarily applied where direct separation methods such as distillation and crystallization cannot be used or are too costly. Liquid – liquid extraction is also employed when the components to be separated are heat-sensitive (e.g., antibiotics) or relatively nonvolatile (e.g., mineral salts).

Liquid – liquid extraction is used in industry for the following purposes:

1) Separation of systems with similar boiling points (e.g., separation of aromatics from aliphatic hydrocarbons);
2) separation of highboilers and low concentrated solutes from aqueous solutions (e.g., phenol);
3) separation of mixtures with high boiling points (e.g., vitamins);
4) separation of temperature-sensitive compounds (e.g., acrylates, biotechnology);
5) separation of azeotropic mixtures (extraction of acetic or formic acid from aqueous media using, e.g., MTBE as solvent);
6) extraction of organic compounds from salt solutions (e.g., caprolactam);
7) extraction of salts from polymer solutions (e.g., ketone resins, polyols);
8) extraction of metal salts from low-grade ores (e.g., copper);
9) extraction of metal salts from wastewater (e.g., copper);
10) recovery of nuclear fuels (Purex process).

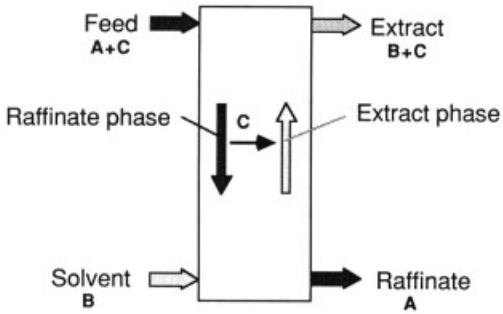

Figure 1. Principles of extraction – streams, phases and components

Processes applied in the petroleum industry, in metal-salt extraction and in chemical and biochemical processes are described in Chapter 5.

Only a general outline is given in this article. For further details, the reader is referred to books [1–7] and the Proceedings of the International Solvent Extraction Conference (ISEC) [8–17]. Data of liquid – liquid equilibria can be found in [18], [19] and [20].

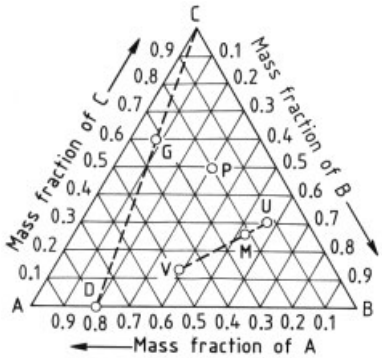

Figure 2. Triangular diagram for representation of a three-component system

2. Thermodynamic Fundamentals

A comprehensive treatment of the thermodynamic fundamentals of phase equilibria is given elsewhere (→ Absorption; → Distillation and Rectification).

2.1. Graphic Representation of Phase Equilibria

2.1.1. Triangular Diagrams

Triangular diagrams (Fig. 2) are used for representing three-component systems. Every possible composition of the ternary mixture corresponds to a point in the diagram. Each corner of the triangle represents a pure component and its designation is marked at this corner. On the side opposite to this corner the mass fraction of this component is zero. The sides of the triangle represent the corresponding two-component systems. The proportions of the individual components in a mixture are obtained as follows: The mass fraction of each component is given by lines parallel to the side opposite to the corner which represents the pure component, from 0 kg/kg at this opposite side to 1 kg/kg at the corner. The numbering can be placed at the edges of the triangle, in which case it is advisable to extend the lines. The numbers can also be inserted in the middle of the lines. This makes the diagram easier to use but some of the numbers are always skew and inverted. The point P has the composition 0.5 kg/kg C, 0.2 kg/kg A, and 0.3 kg/kg B.

The following rules apply in the triangular diagram: If component C is added to a mixture of 0.8 kg/kg A and 0.2 kg/kg B, represented by the point D, then the composition of the new mixture follows the path of the straight line DC. On this straight line, and on all straight lines which pass through a corner, the ratio between the two other components which do not form the corner is constant.

If a mixture of composition U and one of composition V are combined then the resulting mixture lies on the straight connecting line UV. The composition of the final mixture is obtained by dividing the connecting line between the two original mixtures in inverse ratio to the quantities mixed (lever rule). The mixture M, for example, comprises three parts U and one part V, and the mixture G comprises three parts C and two parts D.

The concentration data obtained from the triangular diagram (e.g., point G: 0.6 kg/kg C, 0.08 kg/kg B, 0.32 kg/kg A; B : A = 1 : 4) apply to the total quantity of resulting mixture regardless of whether it is homogeneous or separated into several phases.

As every point in a triangular diagram corresponds to a composition there is no coordinate free for another reference variable. It is only possible to plot parameters such as lines of equal density, boiling points or other physical data, lines of equal activity coefficients or lines of phase boundaries.

Only ternary systems with miscibility gaps are suitable for extractions, and the boundary line between the liquid single-phase region(s) and the two-phase region is particularly important. This boundary line is called the *binodal curve* (Fig. 3). Every point on the binodal curve (except the plait point in systems with a complete binodal curve, see below) is in equilibrium with another binodal point. The lines which connect points in equilibrium with one another are called *tie lines.* All mixtures which correspond to a point on the tie line separate into two phases. The composition of these phases is given by the

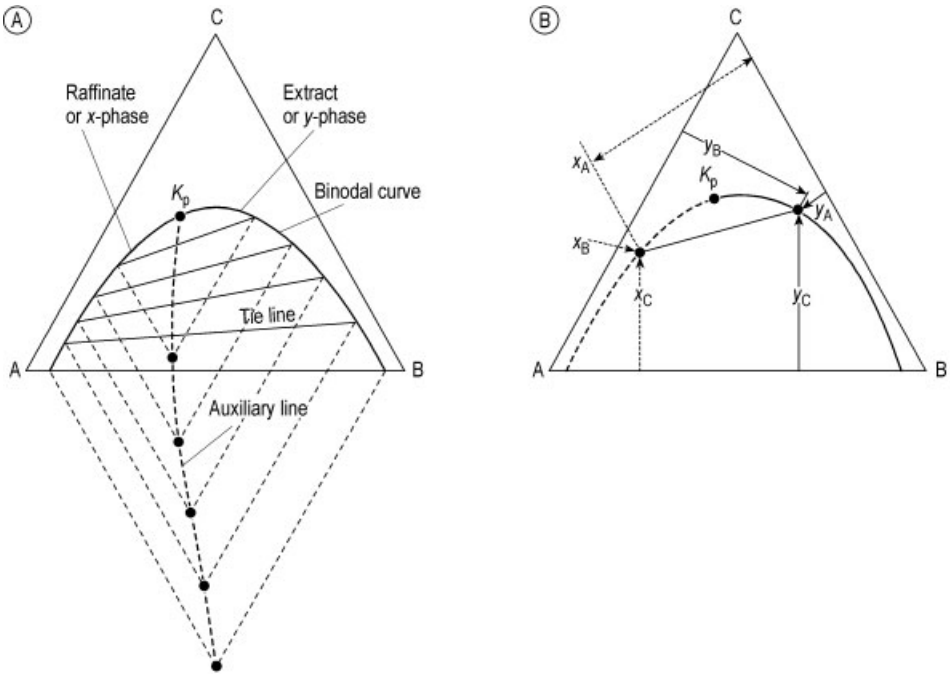

Figure 3. Definitions of phase equilibria and mass fractions in a triangular diagram
A) Phase equilibria; B) Mass fractions

end points of the tie lines and their quantities are given by the lever rule. The equilibrium data have to be determined experimentally in each individual case. Lines parallel to the triangle sides CB and AC respectively, drawn through related points of equilibrium on both branches of the binodal curve, produce an auxiliary line, formed by the intersections of the parallels, and enable to construct any tie line as demonstrated in Figure 3 A.

Several types of triangular diagrams are shown in Figure 4. In every system with a complete binodal curve (*closed diagram,* Type II in Fig. 4) the end points of the tie lines draw closer and closer together until they finally coincide at the *plait point.* The plait point on the binodal curve represents a single phase which does not split into two phases.

Diagram I is called an *open diagram* in which B is not fully miscible with the substance to be dissolved. The capacity of such a solvent is correspondingly low and its selectivity is usually high (see Section 2.4).

In diagram II the substance to be extracted (C) is totally miscible with both A and B. This form of extraction diagram is often found in extractions where solute C and A are present in approximately the same quantities. If in this case the concentration of C in the AB-mixture is too high (higher than the concentration at the plait point) the system forms only one phase and extraction is not possible.

Diagram III shows a system where C has a miscibility gap with A. Its concentration in the B-rich phase is then substantially higher than in the feed. The diagram is typical for wastewater extractions.

Diagram IV is an open triangular diagram in which three liquid phases are formed in the middle concentration range. This behavior is often found in systems consisting of inorganic salts in aqueous solutions and organic solvents. The appearance of a third liquid phase can cause considerable difficulties during metal-salt extraction (see Section 5.3). The influence of pressure on the binodal curve and equilibria is negligible below values of several bar. The influence of

temperature is strong. The lower the temperature the more will the triangular diagram normally be filled by the miscibility gap.

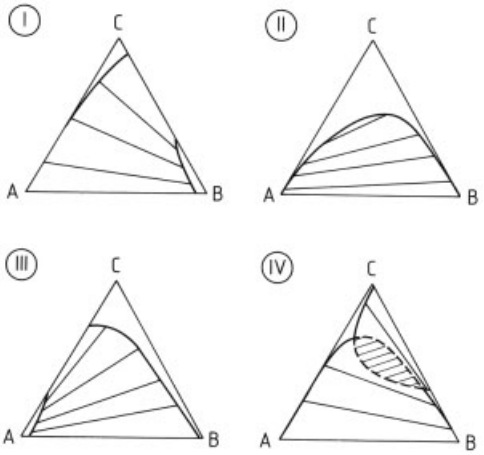

Figure 4. Various types of phase equilibria in ternary systems
Type I: Open system, miscibility gap between the key component of the solvent (B) and solute C
Type II: Closed system, C is completely miscible with the key component of the feed (A) and the key component of the extraction solvent (B)
Type III: Open system, miscibility gap between A and C
Type IV: System with formation of three phases

2.1.2. Other Graphic Representations

Liquid two-phase systems comprising three components can also be plotted on rectangular coordinates. In many cases these diagrams are preferred because the measured values can be adjusted better and the configuration of the calculations is simpler. The most important forms of representation are shown for three mixtures in Figures 5, 6, 7. A comprehensive review of forms of representation is given in [27].

In the equilibrium diagram (Figs. 5, 6, 7 II) the mass fraction of C in the B-rich phase y_C (ordinate) is plotted as a function of this mass fraction in the A-rich phase x_C (abscissa). This diagram is analogous to the McCabe – Thiele diagram in distillation (→ Distillation and Rectification, Chap. 4.1.1.). It shows the relationship between equilibrium and operating conditions.

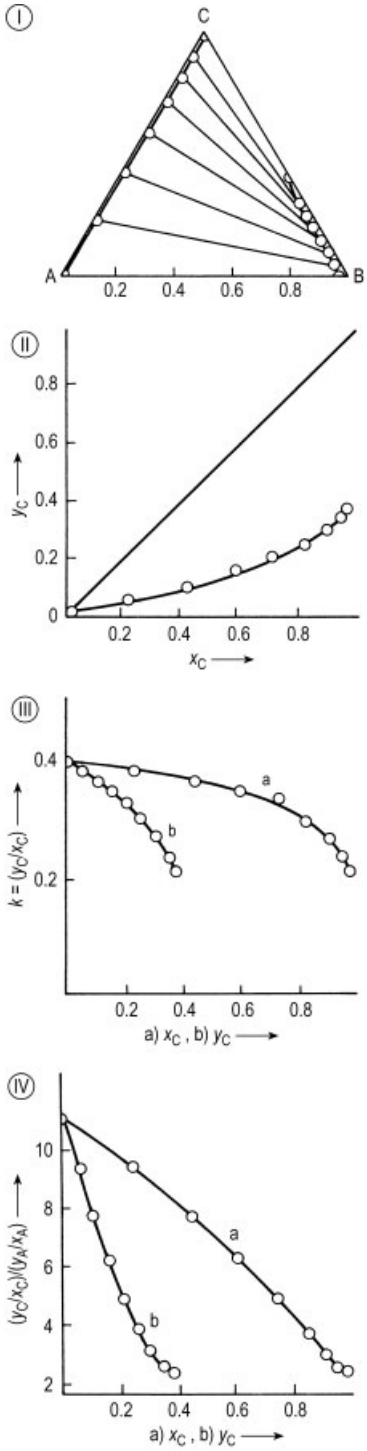

Figure 5. Representation of phase relationship of the ternary system heptane (A), diethylene glycol (B), benzene (C), at 20 °C. All data refers to mass fractions.
I) Triangular diagram; II) Equilibrium diagram; III) Distribution diagram; IV) Selectivity diagram

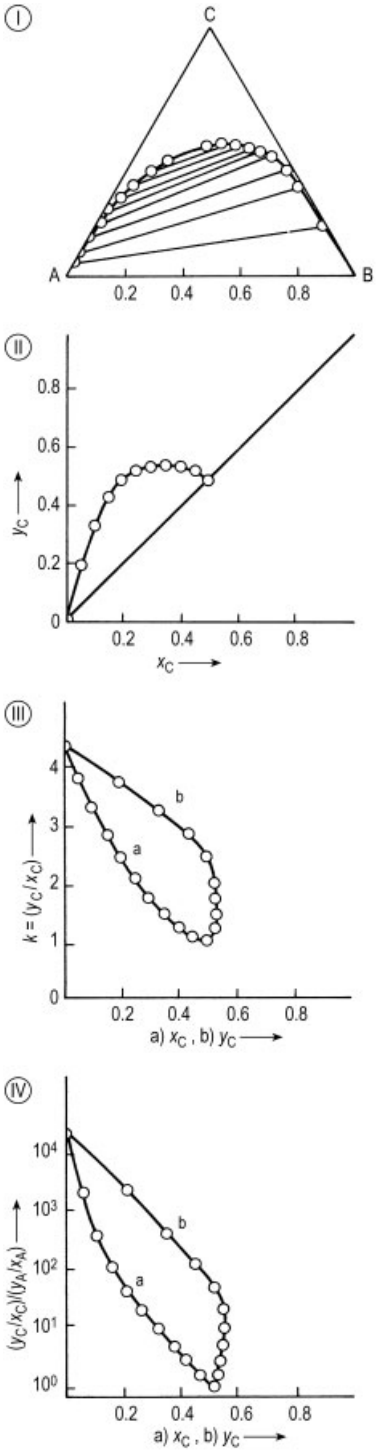

Figure 6. Representation of phase relationship of the ternary system trichloroethylene (A), water (B), caprolactam (C), at 20 °C. All data refers to mass fractions.
I) Triangular diagram; II) Equilibrium diagram; III) Distribution diagram; IV) Selectivity diagram

In the example shown in Figures 5, 6, 7 III the distribution ratios of the extracted components (benzene, caprolactam or phenol) are plotted as functions of their mass fractions in the A-rich phase (curve a) and in the B-rich phase (curve b). The distribution ratio k indicates how much better or worse the extracted component is soluble in the B-rich than in the A-rich phase; the minimal solvent ratio can be evaluated directly from Figures 5, 6, 7 III. The distribution ratio therefore is an important criterion for the selection of a solvent. The mass fraction in the A-rich phase is easier to visualize as a reference variable. At the feed stage it corresponds approximately to the initial feed concentration. It governs the yield at the raffinate outlet. On the other hand, showing k as a function of y_C (curve b) has the advantage that it permits comparisons with respect to solvent loading. It is always advisable to plot both curves because their suitability for extrapolation and interpolation varies considerably in the different concentration ranges.

In the selectivity diagram (Figs. 5, 6, 7 IV) the quotient of two distribution ratios is plotted as a function of the mass fractions of component C in the A-rich (curve a) and in the B-rich phase (curve b). The quotient of the distribution ratios for the extracted component C (y_C/x_C) and for A (y_A/x_A) corresponds to the relative volatility in distillation (see Section 5.1). It is a measure of the selectivity of a solvent.

2.2. Measuring Methods of Phase Equilibria

Whenever possible a ternary system is used as the starting point for measurement of liquid – liquid equilibria. It is first necessary to establish which of the three possible binary systems have a miscibility gap and which are completely miscible. For this purpose it is sufficient in each case to mix the two substances in ratios of 3 : 1, 1 : 1 and 1 : 3. If no formation of two liquid phases occurs in any of the cases then the system is fully miscible and not suitable for extraction.

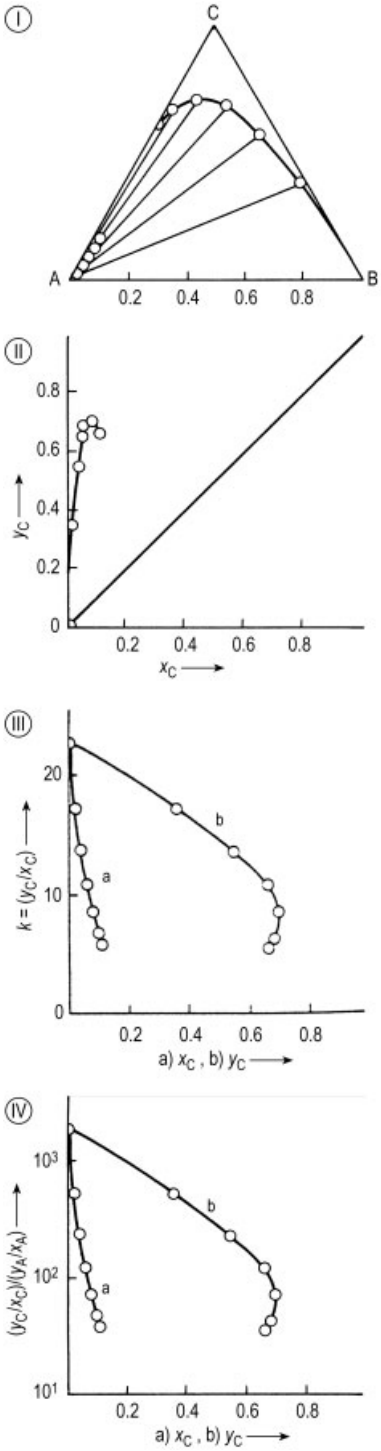

Figure 7. Representation of phase relationship of the ternary system water (A), butyl acetate (B), phenol (C), at 20 °C. All data refers to mass fractions.
I) Triangular diagram; II) Equilibrium diagram; III) Distribution diagram; IV) Selectivity diagram

If, for example, A and B are not miscible while the third, C, is completely miscible with the two others (closed triangular diagram) then the binodal curve can be determined as follows. Equal amounts of A and B are placed in a controlled-temperature vessel which can be shaken (e.g., an Erlenmeyer flask in a thermostat) and substance C is added until the mixture becomes clear. In two further experiments pure A or B is placed in the vessel and in each case the second substance is then added until turbidity occurs. Three points have now been obtained in the triangular diagram and the approximate shape of the binodal curve is known.

For precise determination of the binodal curve, mixtures of A and C containing 0.9, 0.8, 0.7 (etc., for as far as appears necessary) kg/kg A are prepared and titrated with B at constant temperature. When turbidity is reached titration is ceased and a small quantity of the original AC mixture is added until the solution is clear again. Then a characteristic variable such as refractive index, density, dielectric constant, freezing point, boiling point, etc. of this mixture can be determined. In the same way mixtures of C and B are titrated with A. This measurement gives the precise shape of the binodal curve and all the compositions lying on this curve can be defined by a physical parameter.

To determine phase equilibrium and therewith the tie lines, i.e., the points which are in equilibrium with one another, a mixture from the two-phase region is placed in a controlled-temperature, shaking cylinder. Equilibrium is established by shaking. The phases are allowed to settle and are separated by draining off into different vessels. Then their quantities are determined and the physical variable measured during determination of the binodal curve is then measured for both phases. This immediately gives the two tie line points. Furthermore, both phases can also be analyzed by well known chemical methods, such as gas or liquid chromatography and titration. The correctness of the measurement can be checked by connecting the two tie line points in the triangular diagram. The connecting line must pass through the point of the original mixture, and this point must divide the length of the tie line in inverse ratio to the quantities of the two phases.

If a system of more than three components has to be measured then several components are combined into one group in order to arrive at a ternary system again. Determination of a binodal curve would still be possible in such a case but the two phases into which the test mixture separates are no longer identical with the mixtures which were obtained during the determination of the binodal curve. It is therefore better to allow the mixture to reach equilibrium conditions in a controlled-temperature, shaking cylinder, separate it and then analyze the phases using some preplanned method.

The time consuming step in the determination of equilibria is the preparation of the experiment, such as purification of the substances, setting up the calibration curves, etc. Therefore it is senseless to restrict the number of measured points and use interpolation methods. To increase the accuracy a material balance is set up for each component and the measured points are adjusted by drawing the curves from Figures 5, 6, 7.

2.3. Empirical Correlations of Phase Equilibria

In ternary systems, the use of empirical correlations to represent the binodal curves and distribution equilibria of component C is of advantage [28]. The mutual solubilities of the liquid phases are usually determined by a laboratory turbidity titration. The binodal curve is then determined by extrapolating the measured values (see Section 2.2). In Type I systems with an open miscibility gap, the binodal curve can be simply described by parabolic equations:

$$x = D' + E' x'_{\mathrm{A}} + F' x'^2_{\mathrm{A}}$$

$$y = D'' + E'' x''_{\mathrm{B}} + F'' x''^2_{\mathrm{B}}$$

where x'_{A} is the mass fraction of A in the A-rich phase, x''_{B} the mass fraction of B in the B-rich phase and D', D'', E', E'', F' and F'' are coefficients.

In systems with a closed miscibility gap (Type II), the parabolic equations are only valid for the lower parts of the curve; they cannot be used to represent the entire binodal curve. The best equation for the closed binodal curve is that proposed by Hlavaty [29]. The correlation is derived from a polynomial equation and has the advantage that the binodal curve ends exactly at the binary solubility points. Its disadvantage is, however, that the lower, in practice virtually linear, regions of the binodal curves are represented in wave form as a result of the polynomial equation. The Hlavaty equation should, therefore, be used with caution.

The equilibrium distribution of solute C in the A- and B-rich phases is generally determined independently from the binodal curve by means of laboratory experiments. The mass fraction x_{C} and y_{C} of the two phases in thermodynamic equilibrium are thus obtained and determine the positions of the end points of the tie lines on the binodal curves (see Section 2.2). Equations for correlating and interpolating the end point of the tie lines are described in the literature, they are primarily exponential equations such as those proposed by Othmer and Tobias, and Hand [1]. However, all exponential equations generally fail when the concentration of C is very low. To allow a more accurate estimation of distribution equilibria and extraction processes at the low raffinate concentrations that are of industrial importance, Bender and Block [28] proposed the following correlation for the distribution ratio K:

$$K = G_1 + G_2 x'_{\mathrm{C}} + G_3 x'^2_{\mathrm{C}} + G_4 x'^3_{\mathrm{C}}$$

The coefficients G_i must be fitted to the measured values of the tie lines x and y. Although the equation does not generally permit extrapolation of the curve above the measured values to the plait point, the distribution coefficient K can be obtained accurately up to x and $y \to 0$.

The empirical correlations are simple and clear. They can be calculated quickly with appropriate computer programs and give an accurate representation of the measured liquid – liquid equilibria. However, these advantages are accompanied by several disadvantages. The correlations can only be used with ternary systems and do not allow extrapolation to higher multicomponent systems. Extensive laboratory work is required to determine the mutual solubilities of components A and B as a function of the concentration of C as well as the equilibrium distribution of component C. Systematic measurement errors cannot be detected because the experimental values are empirically correlated.

2.4. Selection of Solvent

The selection of the most suitable solvent is one of the hardest problems of the extraction, and the decision made has a great influence on the economy of the process. The solvent must fulfill two basic requirements:

1) it must have a miscibility gap with the A-rich phase, and
2) it must be a better solvent for component C than for A.

A great number of methods have been recommended for discovering suitable solvents [30], [31]. The more empirical strategy of former times is being replaced more and more by a systematic research with the aid of data banks [26]. The concept of polarity is used for characterizing suitable parameters [32].

The greater the difference in polarity between two liquids, the greater is the probability that they have a miscibility gap. The substance to be extracted will have a higher concentration in the liquid to which it is closer in polarity.

However, the relationships are not so simple that only a scale of polarities has to be set up and the difference in polarity at which miscibility gaps occur read off the scale. For example, aniline, phenol, and furfural have miscibility gaps with both water and hexane. Acetone, ethanol, and dioxan which have a similar polarity are miscible with both water and hexane. The molecular form and the size of the molecule also contribute to solvent behavior. The behavior of solvents towards hydrocarbons has been investigated particularly intensively [33], [34]. It is also possible to obtain effects from mixing two solvents which could not be predicted from the properties of the mix components [35], [36].

The most important property of a solvent is its *selectivity.* This indicates how much better the solute is dissolved than the A-rich phase. As can be seen from Figures 5, 6, 7 III the selectivity is strongly dependent on concentration.

In addition to selectivity the *capacity* of a solvent is also important. This is defined by the distribution coefficient of C between the B-rich and A-rich phases and determines the quantity of solvent which has to be recycled. When testing different solvents for a specific extraction task it is often found that solvents with higher selectivities have lower capacities and vice versa. The greater the throughput of a plant the more important the capacity of the solvent is for the economy of the process. If there are several possible solvents for a process then their behavior with respect to selectivity and capacity can be tested with a gas chromatograph without starting with time-consuming equilibrium measurements [37].

In addition to these basic requirements there are also other criteria which must be taken into consideration when selecting a solvent and which sharply reduce the number of possible compounds. Most important is the availability of the solvent in commercial quantities at an economically acceptable price. It is then necessary to find data concerning boiling points, density, viscosity and – where available – information about solubility [38], [39].

The *boiling point* of the solvent is important because the extracted component is usually separated from the solvent by distillation and the boiling point difference determines the reflux ratio in this distillation. If the solvent boils at a higher temperature than the solute, this solute is distilled overhead. If it is necessary to recycle a great deal of solvent then the solvent has to be purified by distillation in a subsequent step; the specific heat which is required for heating the solvent to its boiling temperature must also be taken into account. Too great a temperature difference is then a disadvantage.

The *density* of the solvent is important for obtaining a sufficiently large difference in densities between the two liquid phases for the separation process. Chlorinated compounds are suitable if a solvent with a high density is to be used, and particularly low-density compounds are found among the saturated hydrocarbons and ethers.

The *thermal stability* of the solvent can be highly significant. The researcher hardly ever finds the answer to this question in the literature and is directed to carry out his own measurement. It is necessary to know (1) how much solvent make-up will be needed in the process; (2) whether a regeneration process for continuous removal of decomposition products needs to be installed. Very careful tests must be carried out to find out whether decomposition products can contaminate the extract or have a corrosive effect on the construction materials of the plant. Sometimes the toxicological properties and biological

degradation of a solvent have to be considered [40].

The *viscosity* should be as low as possible as this favors mass transfer and phase separation. However, this property is less significant than those mentioned above. If the possible solvents have been reduced to a manageable number after testing all the aspects then the individual properties can be evaluated with a point system to discover the best solvent.

The *surface tension* is another important parameter because it influences the energy demand necessary to produce a suitable drop dispersion. High surface tension gives rise to formation of stable drops and needs a high energy input. A solvent with a high surface tension decreases the danger of producing an emulsion. The surface tension is dependent on the concentration of the transfer component. Further properties such as suitability for regeneration, disposability, costs, combustibility, toxicity and corrosiveness also have to be taken into account to select a suitable solvent.

Bound up with solvent selection is the choice of the dispersed and the continuous phase, respectively. Direction of mass transfer and phase ratio have to be taken into account. The direction of mass transfer influences the surface tension and in this way can promote or hamper the coalescence of drops. The phase to be dispersed should be the one which impedes the drop coalescence to produce stable drops, and the one with the higher flow rate to obtain a mass transfer area as large as possible.

2.5. Calculation of Theoretical Stages

(→ Absorption, Chap. 4.; → Distillation and Rectification, Chap. 4.)

Two possibilities exist to calculate the mass transfer performance of extractors. The first one uses the concept of theoretical stages and is based on the assumption that concentration equilibrium is established in every section of an extractor. This assumption can be fulfilled only in the mixer – settler with long contact times of the phases involved. In this concept the total height of an extraction column is given by the product of the number of theoretical stages (NTS) and the height equivalent of a theoretical stage (HETS). NTS follows from a material balance in combination with the distribution equilibrium; HETS depends on fluid dynamics and mass transfer in the columns and comprises all real conditions as for instance the conditions of phase mixing. Therefore this concept is preferred only for rough estimations in practice because it is easy to handle and sufficient for preparation of laboratory experiments. In the following sections, the method will be presented by aid of equilibrium diagrams.

The second possibility is a stream concept. The column height required for a given throughput and concentration gradient of phases along the column can be calculated from height integration of a set of differential equations, derived from a differential material balance and the local mass transfer between the both phases. The stream concept supplies the course of concentrations along the column and is suitable for analysis of the dynamic behavior of a column and for taking into account the drop population inside the column. It is preferred until now for scientific investigations, e.g., for stationarily operated pulsed sieve tray columns [45], [47–49], [51], [62] and for instationarily operated rotating extractors [65], [68].

2.5.1. Selection of Calculation Methods

The conditions for an applicable method of calculation are set by the problem to be solved, by whether simplified assumptions are permissible and by the main influencing variables of an extraction problem. They are, therefore, different. The following points should be considered when selecting a method of calculation:

1) Do changes in mass flows have to be considered?
2) Is the distribution ratio constant?
3) Is the concentration gradient large or small?
4) Does only one component have to be extracted, or are there several with different distribution coefficients?

2.5.2. Specific Methods

Method 1. If the following conditions are satisfied:

1) the component A and solvent B are virtually insoluble in each other,

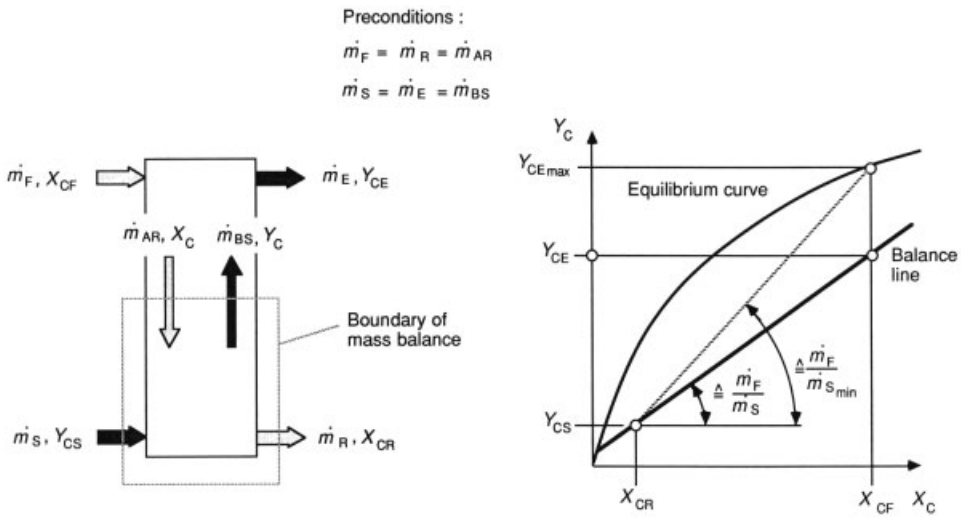

Figure 8. A) Material balance of a column cross section; B) Corresponding operating diagram

2) the changes in mass flow rates of the extract and raffinate phase are negligible (e.g., for a low content of C in the feed)

then the compositions of the phases can be expressed as loads. If a material balance is made as shown in Figure 8, the two loads Y_c and X_c in a given column cross section can be related as follows

$$Y_C - Y_{C,S} = \frac{\dot{m}_F}{\dot{m}_s} \cdot (X_C - X_{C,R})$$

When plotted on the equilibrium diagram, this relation gives a straight line, the balance or operating line. The number of theoretical stages required to obtain a desired decrease in the concentration of C from $X_{C,F}$ to $X_{C,R}$ can be determined graphically by constructing a series of steps as shown in Figure 9. The result of the use of the steps often means that the load obtained from C in the raffinate ($X_{C,R}$) does not correspond to the required load ($X_{CR,req}$).

Figure 8 also shows the position of the balance line for the minimum extractant ratio V_{min}

$$V_{min} = \dot{m}_{s,min} / \dot{m}_F$$

The desired enrichment of the feed with C using the minimum amount of extractant $\dot{m}_{s,min}$ can only be achieved with an infinite number of stages.

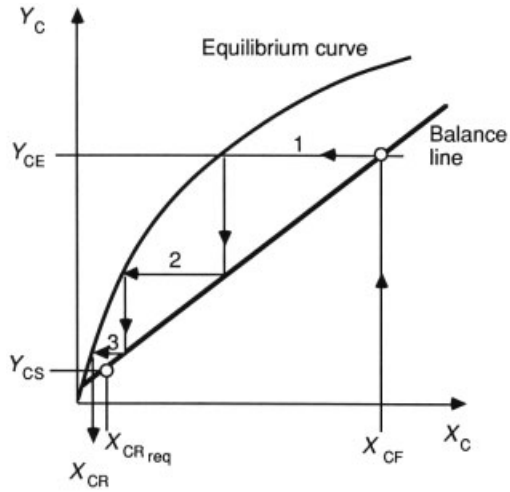

Figure 9. Construction of steps in the operating diagram to obtain the number of theoretical stages (NTS)

Method 2. Figure 10 shows the equilibrium diagram for the system toluene – ethylene glycol – water as a double logarithmic plot. This type of plot is useful if the distribution ratio varies considerably, despite low loading, and the concentration range considerably exceeds the desired raffinate purity. In this case the operating line is at an angle of 45°.

$$\log(Y_C - Y_{C,S}) = \log\left(\frac{\dot{m}_F}{\dot{m}_S}\right) + \log(X_C - X_{C,R})$$

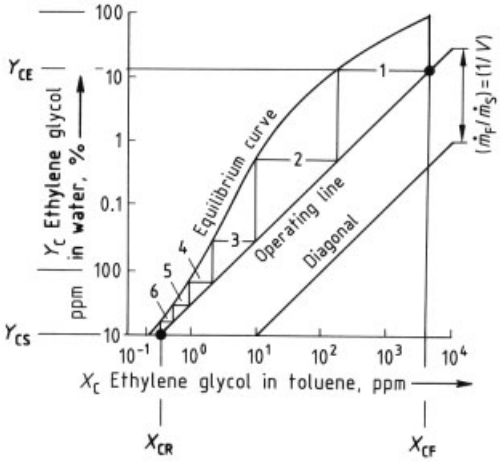

Figure 10. Determination of the number of stages for a variable distribution ratio and a high degree of depletion

The position of the straight lines is determined by the extractant ratio V. The number of required theoretical stages can be determined by constructing a series of steps.

Method 3. If the distribution ratio K is a constant, further simplification is possible; in analogy with the Colburn equation (→ Absorption, Chap. 4.2.), the number of theoretical stages NTS can be calculated:

$$\mathrm{NTS} = \frac{\log\left[1+(E-1)\left(\frac{X_{\mathrm{C,F}}}{X_{\mathrm{C,R}}}+\frac{Y_{\mathrm{C,S}}}{X_{\mathrm{C,R}}}\right)\right]}{\log E} - 1$$

where the extraction factor E is given by

$$E = K \cdot \dot{m}_{\mathrm{s}}/\dot{m}_{\mathrm{F}}$$

If NTS is known and the degree of depletion $X_{\mathrm{C,R}}/X_{\mathrm{C,F}}$ is desired, the following equation is obtained:

$$\frac{X_{\mathrm{C,R}}}{X_{\mathrm{C,F}}} = \frac{(E-1)\left(1+\frac{Y_{\mathrm{C,S}}}{X_{\mathrm{C,F}}}\right)}{E^{(\mathrm{NTS}+1)}-1}$$

This equation can only be solved iteratively; a graphical solution is shown in Figure 11.

Method 4. If none of the above-mentioned conditions is satisfied, and the extraction problem can be limited to consideration of the three key components, the calculation can be carried out with the aid of the triangular mass fraction diagram. This method is based on the mixing rules described in Section 2.1.1.

The specified compositions of the feed F and solvent S streams are plotted on the phase equilibrium diagram for the ternary mixture (Fig. 12). For material balance of the feed and product streams, the column can be assumed to be a simple mixing vessel. The composition M of the mixture lies on the tie line FS; the position of M on the tie line is determined by the ratio $\dot{m}_{\mathrm{S}} : \dot{m}_{\mathrm{F}} = \mathrm{MF} : \mathrm{MS}$. The compositions of the raffinate R and extract E streams leaving the column are obtained from the intercepts of the tie line with the binodal curve. The raffinate composition is determined by the desired mass fraction of C in the raffinate ($x_{\mathrm{C,R}}$). The composition of the extract E is given by the point at which the extended line RM intersects the binodal curve. The ratio $\dot{m}_{\mathrm{R}} : \dot{m}_{\mathrm{E}}$ is obtained by applying the lever rule. The quantities and the compositions of the incoming and outflowing streams are thus determined.

The required number of stages in a countercurrent column are obtained by using the *pole point procedure.* The procedure can be explained by imaging a stream P to maintain the material balance. Mixing of this stream with the raffinate stream R gives the composition of the solvent stream S. To maintain the material balance the extract on the other hand must be composed of the feed stream F and the imaginary stream P. The imaginary stream P can thus be regarded as a common mixing partner for F and R to maintain the compositions of E and S. The composition of the imaginary stream P is represented by a point outside the triangular diagram, the pole point. The imaginary mass fractions of the three components in the pole point can be > 1 or negativ. If the material balance boundaries for the individual stages are taken as shown in Figure 13, the above procedure gives the stage extracts E_2, E_3, etc. from the mixture of the imaginary stream P and the individual stage raffinates in the proceeding stages R_1, R_2, etc. The composition of the raffinates R_1, R_2, etc. are linked to the corresponding extracts E_1, E_2, etc. via the distribution ratios (tie lines).

2.5.3. Evaluation of Stage-Number Calculation for Process Design

In order to determine the operating conditions and the dimensions of the equipment used for

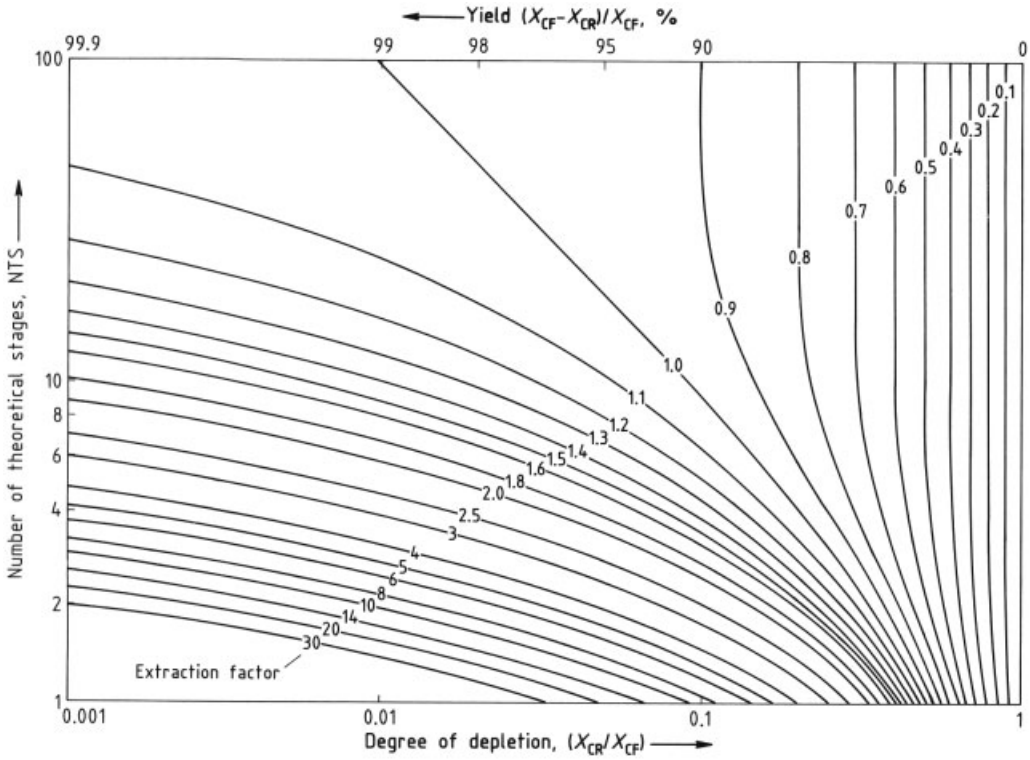

Figure 11. Dependence of the theoretical number of stages, the degree of depletion, and the yield on the extraction factor *E*

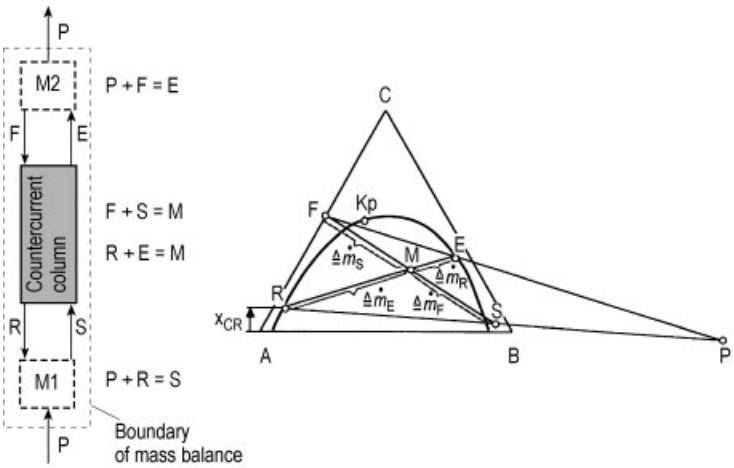

Figure 12. Pole point construction – balance diagram

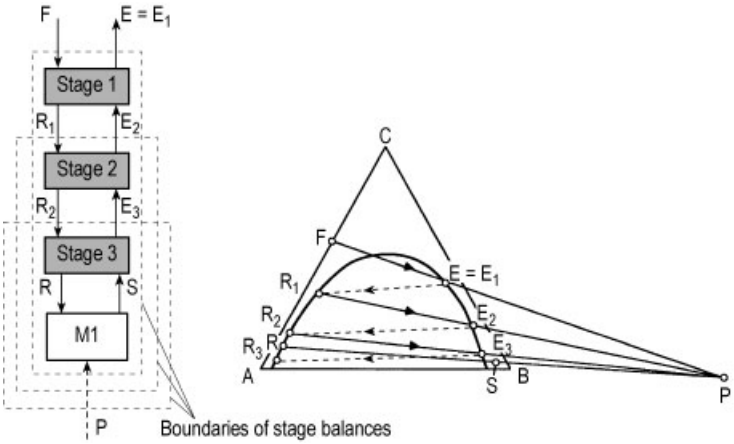

Figure 13. Pole point construction – operating diagram

extraction, the results of several stage-number calculations must be plotted in a diagram showing the relationship between the solvent circulation, the number of stages, and the residual content of solute in the raffinate. Such a diagram is shown in Figure 14 for mixture II (see also Fig. 6). It represents the extraction of a 10 % solution of caprolactam in trichloroethylene with water as the extraction solvent. The favorable operating region is usually located in the region which shows the most pronounced curvature. Other variables (e.g., the effect of temperature and the residual loading of the solvent with solute) may have to be taken into consideration.

3. Apparatus

3.1. Survey

Extraction apparatus can be classified into countercurrent columns, centrifugal extractors, and mixer – settlers. All industrial equipment designs use the principle to disperse one of the two liquids into the other in order to enlarge the contact area for mass transfer and to influence the drop generation by energy input such as mixing or pulsing. After the resulting droplets have been contacted with the continuous phase long enough to allow the extractable component(s) to migrate to the dispersed phase, both liquids have to be separated again. Densities will define the flow directions of the phases: The heavier liquid flows from top to bottom, the lighter one from bottom to top. Accordingly, the fluids have to be transported to the apparatus and removed by pumps.

In the following chapter, a few typical designs have been selected from the large number of equipment designs actually used in industry. A more detailed treatment can be found in literature, especially in [5] and [6].

3.1.1. Columns without Energy Input

The simplest extractor is the *spray column* (Fig. 15 A). The spray column achieves only very small numbers of separation stages and is rarely used nowadays. If the column of Figure 15 A is filled mainly with the heavier liquid, the lighter liquid breaks at the feeding point into drops that rise in the continuous liquid. Drops reaching the interface are kept at this position by controlling the bottom discharge (LIC). There, they will then gradually coalesce and form a continuous layer floating on top of the continuous liquid.

The reverse procedure is possible as well, as shown in Figure 15 B. Here, the distributor is located at the top and the water interface, called the principle interface, is controlled at the bottom of the column by filling it mainly with the lighter liquid. In this case, the heavier liquid is the dispersed one, broken into droplets descending in the lighter liquid. The principle interface can be adjusted at any height (between top and

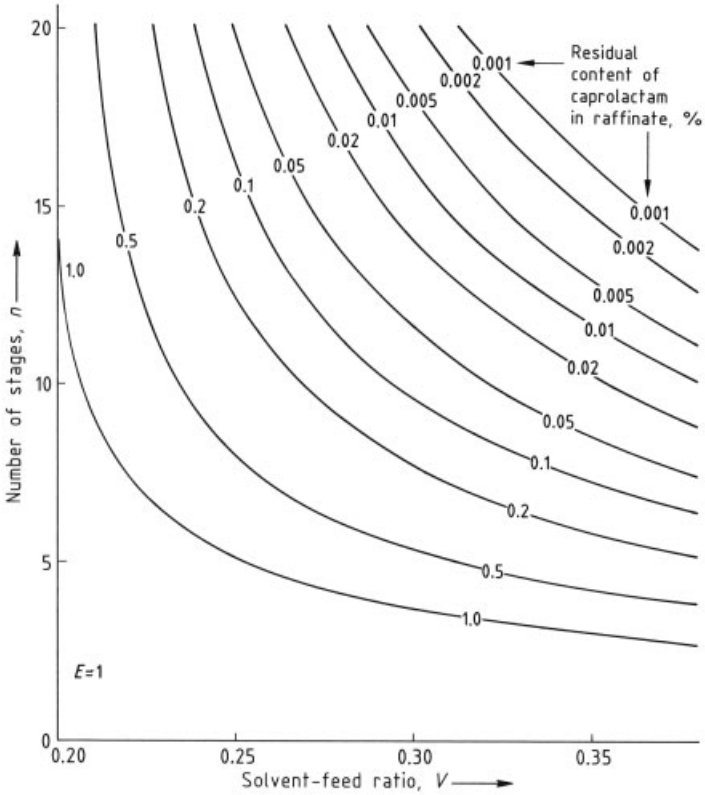

Figure 14. Extraction of caprolactam from a 10 wt % solution in trichloroethylene with water as solvent

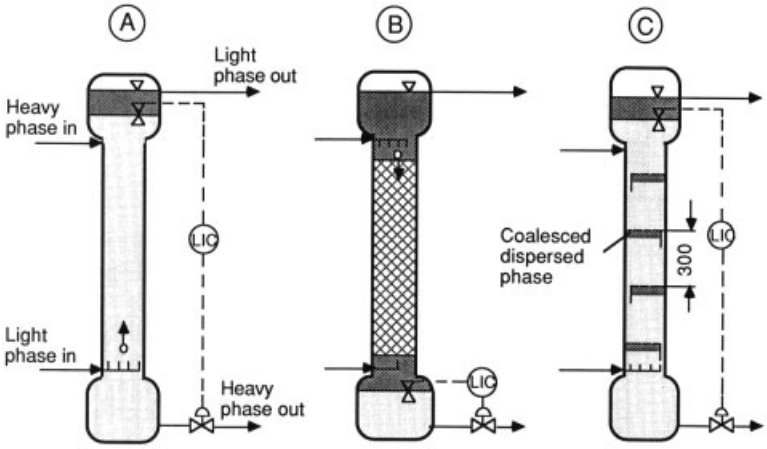

Figure 15. Columns without energy input (LIC = level indicating controller)
A) Spray column (light phase dispersed); B) Packed column (heavy phase dispersed); C) Sieve plate column (light phase dispersed)

bottom) of the column as needed, thus the heavier liquid being dispersed above and the lighter liquid below the principle interface.

Countercurrent columns are equipped with settling zones at the top and the bottom, with coalescing aids provided if necessary. The liquids should never be fed into the column in these settling zones because this would cause hydrodynamic instability of the column thus reducing the operational performance. The dispersed liquid often cannot be selected from theoretical considerations, but only by aid of experiments in a pilot column using the real material system.

Packed columns (structured and unstructured) and *sieve plate columns* have been adopted from distillation technology (→ Distillation and Rectification, Chap. 8.). In *packed columns* drop formation occurs on entry of the dispersed phase in the packed bed (Fig. 15 B). The drops rise through the packed bed without significant further dispersion. However, mass transfer takes place mainly during the formation of a new interfacial area. This explains the low efficiency of packed columns [6, Chap. 8].

In a *sieve plate column* with downcomers, as shown in Figure 15 C, the lighter dispersed phase coalesces beneath each plate to give a continuous layer. Its buoyancy presses the liquid through the holes of the sieve plates, producing jets which disintegrate into drops. The continuous liquid flows through the downcomers from stage to stage and then across the plates. In the case of low interfacial tension the layers of coalesced phase are too small for a symmetrical operation of the plates. Then it is advantageous to use dual flow plates without downcomers [57].

Sieve plate columns have the advantage over packed columns that new drops are formed at each tray. However, the drop load of the plate is often irregularly distributed, especially if the column diameter is large (more than 1 m). This is the reason why sieve plate columns can reach efficiencies of 10 – 15 % only in rare favored cases. Moreover, since columns are generally built with a plate spacing of 0.3 m, even simple separations require considerable overall heights.

Mass transfer between the two liquids can be substantially improved by application of additional mechanical energy in form of superimposed agitation or pulsation, resulting in higher turbulence near the drops and enlarged mass transfer area due to smaller drop sizes.

3.1.2. Pulsed Columns

The simplest design of pulsed column is the *pulsed sieve plate column* (Fig. 16 A). A piston pump with an electrical or an air-pressure drive causes the entire liquid content of the column to vibrate. Both phases are alternately forced through the holes in the rhythm of the pulsation. The downcomers used in the conventional sieve plate column are thus superfluous. The dispersed phase does not coalesce above or below the plates to give a continuous layer. The pulsation shock flings the drops off the plate faster than would be possible as a result of only the density difference. This leads to a fine, uniform drop distribution and hence enlarges the mass transfer area. Applications are limited by the physical properties of the substances concerned and high throughputs. Systems which emulsify easily cannot be processed in pulsed columns just as sticky and greasy liquids. For very large throughputs (column diameter > 0.6 m), the pulsed volume and pulsing power become very high; however, pulsing units for large column diameters have been developed.

The separation effect of the column decreases with increasing diameter, this has not yet been fully investigated. The plate spacing is 50 – 100 mm. Larger plate spacings favor backmixing and hence reduce the efficiency of the column. Smaller plate spacings lead to a strong decrease of the throughput. Sieve plates have cylindrical holes and a triangular pitch. The relationship between the fractional free area φ, the hole diameter d_i and the pitch p is given by

$$\varphi = \pi d_i^2 / \sqrt{12p^2}$$

For usual liquid – liquid systems sieve plates with hole diameters from 2 to 4 mm resulting in free areas of 23 to 40 % are recommended [6, Chap. 10], the lower value of 23 % is used for high and the higher value of 40 % for low interfacial tensions. Free areas less than 20 % are used for liquids with higher viscosities [58].

The plate material has a pronounced influence on drop formation and hence on the separation efficiency. The plate material and phase selection must be matched in such a way that the

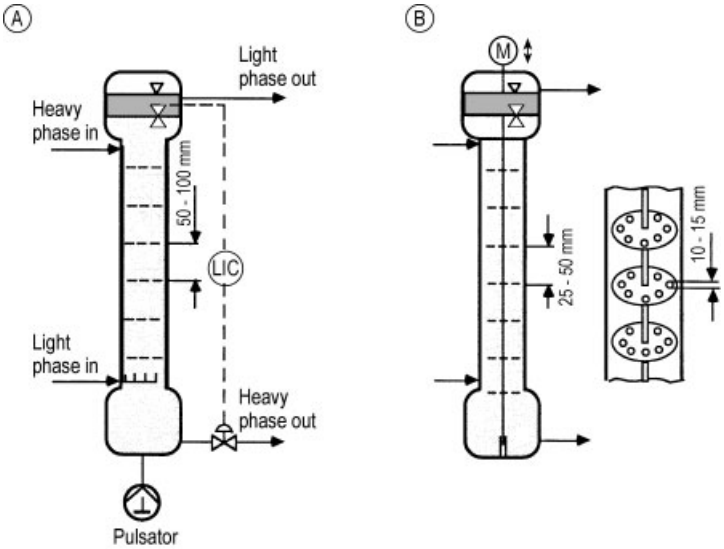

Figure 16. Pulsed and reciprocating-plate columns (Light phase dispersed)
A) Pulsed sieve-plate column; B) Reciprocating-plate column (Karr)

plates are wetted by the continuous phase but not by the dispersed phase, because this would lead to strong coalescence effects. If, for example, an aqueous phase is to be dispersed into drops, metal trays are unsuitable, but plastic plates are satisfactory. The sharp edged side of the sieve plate, which is due to manufacturing, should exactly face the flow direction of the dispersed liquid so that drops will shear at the edges of the holes. Gaps between the sieve plate edge and the column wall should be avoided by precise construction or by applying edge-sealing elements made of materials which are wetted by the continuous liquid.

In most cases, piston pumps or diaphragm pumps are used as the pulsators. The pulsing frequencies are between 30 and 150 min^{-1}, with pulsing amplitudes from 5 – 15 mm (relative to the column cross section). (Pulsing amplitude is defined as the distance between upper and lower center of the pulsing motion in the active part of the column.) The dispersed phase has to be fed into the column via a distributor. The mean drop diameter of the predispersed swarm should be larger than the mean drop size in the active part of the column, otherwise the hydrodynamics of the column would mainly be affected by the distributor. The holes in the distributor should be very sharp-edged to avoid very fine drops. Even in packed columns, the separation effect can be substantially improved by pulsing of the continuous phase, at the expense of a reduced throughput. The increased performance results from greater shear forces causing a reduction in drop size and an increase in interfacial area [6, Chap. 8]. As in distillation columns, the phenomenon of maldistribution in randomly packed columns of larger diameter can be overcome by use of structured packings. These packings lead to a good cross-sectional mixing and have a higher capacity, typically a total flow rate of both phases of approximately 100 $m^3m^{-2}h^{-1}$ compared to 30 $m^3m^{-2}h^{-1}$ for a randomly packed column. Structured packings are particularly suitable for low interfacial tension systems or for small density differences; separation efficiency is nearly independent of scale-up. Hydrodynamics and mass transfer performance are well investigated [6, Chap. 8], [59].

Pulsing Units. The use of pulsed columns is currently limited to small and medium throughputs. The column diameter is limited by the pulsing unit. For very large throughputs, the pulsed volume becomes so large that the use of a piston pump as the pulsing unit causes difficulties. Consequently, pulsing units have been developed which operate without a piston pump. The pulsator developed by Misek (Fig. 17) is based on

a centrifugal pump. The pump (e) is connected on both the suction side and the delivery side to pressure vessels which are alternately connected to the column by means of a rotating valve (c). With the valve setting shown in Figure 17 A, the pump delivers the liquid into the column via the discharge drum (b), taking the rate delivered from the suction drum (a). After changeover of the valve (Fig. 17 B), the pump delivers from the column via vessel (a) into drum (b). The pump thus always delivers continuously in one direction. Reversal of the direction of flow in the pulse line is effected by the change-over valve (c) in accordance with the pulse frequency. A pulsator based on this principle has been developed by Bronswerk/DSM (Fig. 17 C), and is now in operation for columns with a maximum diameter of 2.7 m.

For large pulsed volumes, the hydraulic pulsator from Wepuko [60] is also suitable (Fig. 18). It is flanged directly to the lower bottom of the column end and acts in the direction of the column axis. The pulsing piston is connected via a common piston rod to a working piston, the hydraulic drive of the piston is a rapidly reversible radial piston pump.

Reciprocating-Plate Columns. In another approach for solving the problem of large throughputs, the plates, and not the liquid, are made to reciprocate, so that the pulsing pump is eliminated. This principle led to the development of the reciprocating-plate column which can be regarded as the converse of the pulsed sieve plate column. In the pulsed sieve plate column, the trays are fixed and the entire liquid content of the column is caused to vibrate; in the reciprocating-plate column, the plates are moved. In both cases, the liquid undergoes an oscillating motion relative to the plate.

In the *Karr column* (Fig. 16 B), sieve plates with a free area of about 60 %, a 10 – 15 mm hole diameter, and a 25 –50 mm plate spacing, are mounted on a central shaft. The complete pack is moved up and down, driven by an eccenter.

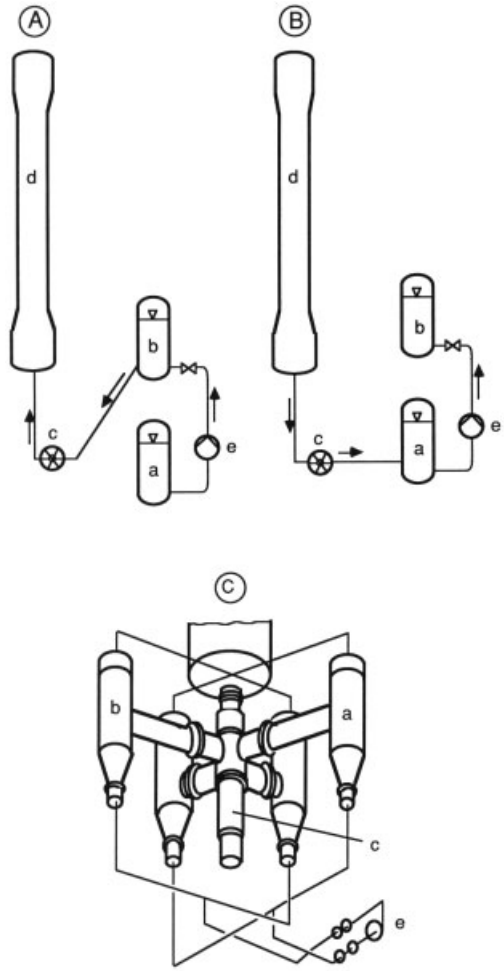

Figure 17. Pulsator with rotating valve
A) Principle diagram of compression stroke; B) Principle diagram of suction stroke; C) Technical design (Bronswerk)
a) Suction drum; b) Discharge drum; c) Rotating valve; d) Column; e) Centrifugal pump

The relatively large free area of reciprocating plate columns allows much greater throughputs than conventional pulsed columns. However, recent investigations showed that pulsed sieve plate columns can be equipped with larger free areas, too. Both types of pulsed columns show very similar performances and operational behavior if they are equipped with the same free areas [61]. Research on the established types of pulsed columns has been very intensive since the 1970s.

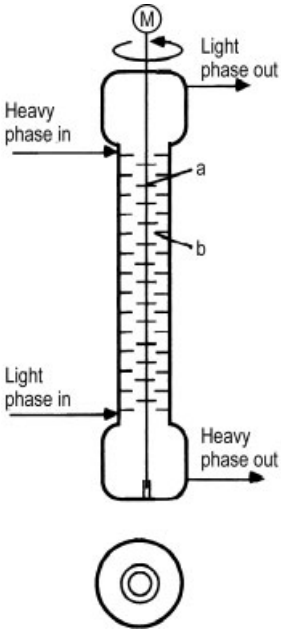

Figure 18. Hydraulic pulsator
a) Pulsing piston; b) Working piston; c) Rinsing; d) Hydraulic aggregate with radial piston pump

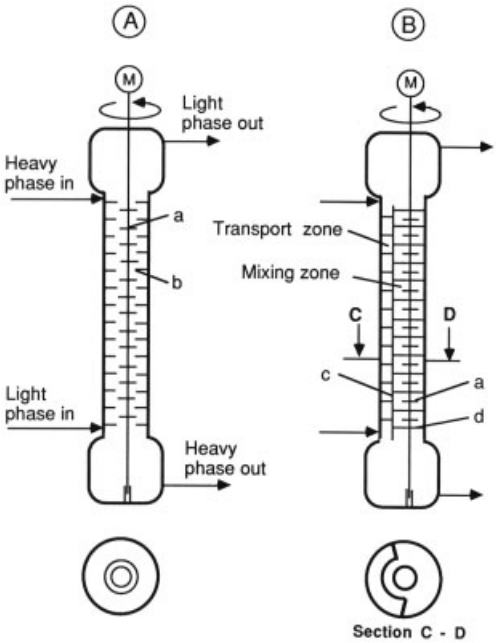

Figure 19. Rotating disk contactor (RDC)
a) Rotating disk; b) Stator ring

3.1.3. Columns with Rotating Internals

Rotating Disk Contactor (RDC). In the RDC (Fig. 19 A), stator rings are mounted at close intervals. Disks are fitted between the rings to a rotating shaft. The rotating disks fling the drops of the dispersed phase off centrifugally, onto the stator rings where they are braked and deflected. The drop distribution is excellent and can reach the emulsification limit as the rotational speed increases.

The RDC has a relatively simple construction with a large free cross-sectional area. On the one hand, it reaches large throughputs, on the other hand it favors backmixing of phases causing a decrease of efficiency. A large range of throughputs is possible without remarkable decrease of efficiency. The loading limit is generally lower than in pulsed columns. The RDC is unsuitable for mixtures which tend to form emulsions and for products containing dirt. The separation efficiency decreases sharply with increasing diameter. Nevertheless, successful RDCs of up to 3500 mm in diameter have been constructed. They are thus suitable for greater throughputs than pulsed columns and are preferably used to operate with large product flows when the demands on separation performance are only low as for instance in the oil industry or in wastewater treatment [42].

As in pulsed columns, the wetting behavior also influences drop formation in the RDC, and should therefore be taken into account when selecting a material for the column internals.

Kühni Extractor. The Kühni extractor (Fig. 20 A) uses a turbine as the mixing element. Perforated partition disks separate the extraction stages and reduce backmixing.

This apparatus is also used for performing reactions between two liquid phases; the residence time can be adjusted via the free area of the perforated disk.

QVF-Extractor. The QVF extractor RZE (Fig. 20 B) uses blade stirrer as mixing elements. The partition disks have meander-formed metal stripes at the inner circumference to reduce the phase dispersion by channeling of both phases. The operational behavior is similar to that of the Kühni extractor (see Fig. 20 A).

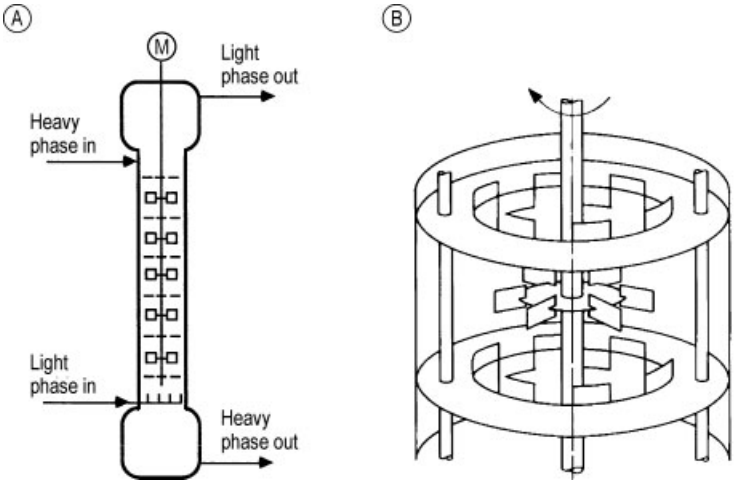

Figure 20. Columns with rotating internals
A) Kühni extractor; B) QVF extractor

Graesser Contactor. The Graesser contactor (Fig. 21) is another type of rotating equipment. Here, not only the internals but the entire apparatus rotate. Internal half-tubes are fitted as scooping elements in a horizontal rotating cylinder. Due to the rotation, the half-tubes transport part of the heavy phase into the overlying light phase, and vice versa. To reduce backmixing, the extractor length is divided into chambers by partition disks. Because of its low dispersing effect, this equipment is particularly suitable for systems which tend to emulsify. The throughput is low compared with other extractors.

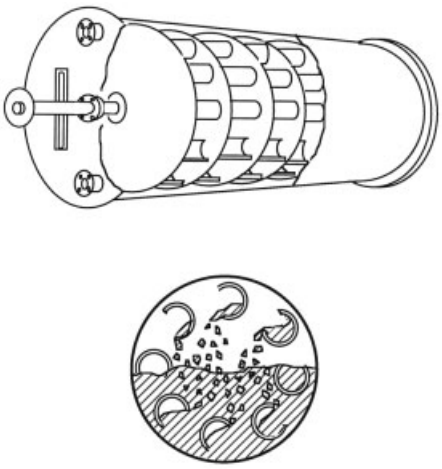

Figure 21. Graesser contactor

3.1.4. Mixer – Settlers

Mixer – settlers consist of single or several stages in series (Fig. 22 A), each stage consisting of a stirred vessel and a settling vessel (Fig. 22 B). Various arrangements are possible such as cocurrent, countercurrent, or crosswise flow of phases. The mixer – settler is the only extractor that reaches stage efficiencies of nearly 100 % and is almost insensitive for load fluctuations.

This extractor has the advantage that backmixing can be completely precluded due to the complete phase separation in each stage. As a result, there are no transfer problems in scale-up. The favorable efficiency is almost completely retained even for large throughputs. Any number of separation stages can be built, and, moreover, throughput does not appear to be limited. These advantages account for the widespread use of mixer – settlers in industrial processes for many decades. However, early mixer – settler equipment was very complex. Each separation stage consisted of a stirred vessel and a settling vessel. The two phases had to be pumped between the individual stages. Each settler needed an interface level control. The horizontal arrangement resulted in a large area requirement.

In an effort to overcome these disadvantages of the mixer – settler batteries without foregoing any of their advantages, a number of different designs were developed. Mixer – settlers are used

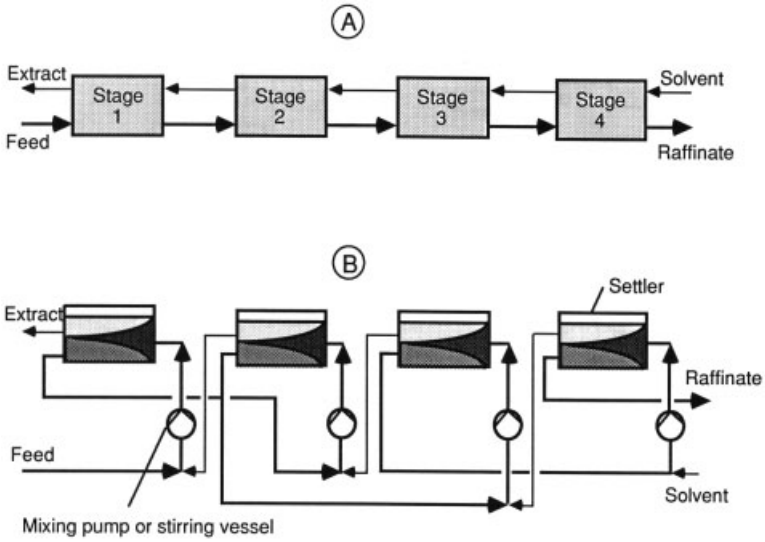

Figure 22. Principle of a mixer – settler
A) Flow Sheet; B) Arrangement

in two different areas. Industrial-scale equipment is used to obtain extremely high throughputs when a small number of stages is needed, e.g., in the extraction of metal salts. Laboratory-scale devices are employed in process development to check the number of calculated stages as a function of the ratio solvent.

Large-scale mixer – settlers were originally developed to extract copper from dilute aqueous solutions. The basic design is a *box-type mixer – settler* according to Denver (Fig. 23). The mixing and separating zones are combined in a single unit and are separated by an overflow weir for the light phase and an underflow slit for the heavy phase. The individual stages are arranged so that the outlet of the settler also acts as the inlet of the mixer. The agitator has two functions, it disperses and conveys the liquid phases. These functions may conflict with one another. The rate of rotation necessary to overcome the conveying height may be too high when unstable systems must be dispersed, causing problems in phase separation, e.g., a high degree of entrainment. Mixing pumps with low shearing forces, and thus a low mixing intensity, have been developed to avoid these difficulties. Examples are the Pumpmix impeller used with the *Davy McKee mixer – settler* (Fig. 24 A) and the *IMI turbine pump-mix unit* (Fig. 24 B).

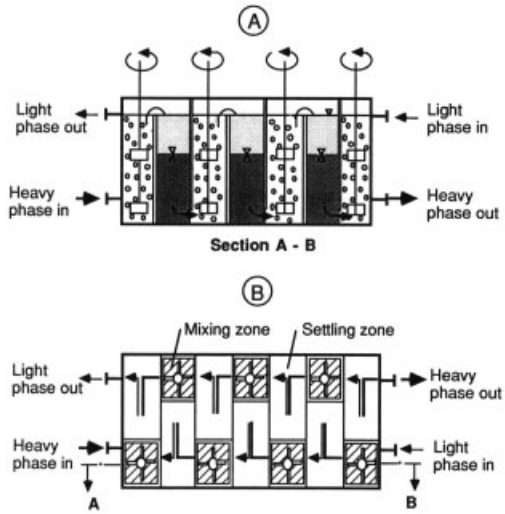

Figure 23. Box-type mixer – settler (Denver)
A) Side view; B) Ground plain

A further development of the Davy McKee mixer – settler has been described [43]. Each stage consists solely of a vessel in which mixing and separation occur. The product inlet and the agitator are located in the middle and create a dispersion zone. The light phase separates out at the top and the heavy phase at the bottom.

In the *Lurgi tower extractor,* the battery of adjacent settlers is replaced by a space-saving

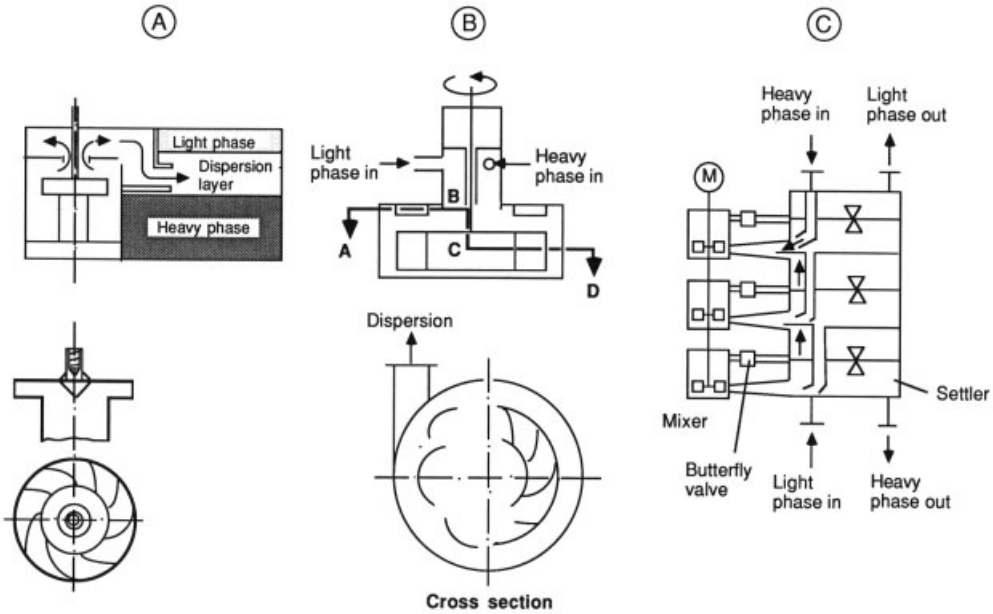

Figure 24. Industrial mixer – settlers
A) Davy McKee mixer – settler; B) IMI turbine pump-mix unit (cross section A – B – C – D shown in the lower part); C) Lurgi tower extractor

tower design in which the settlers are arranged on top of each other. External centrifugal pumps are used for mixing (Fig. 24 C). Like all mixer – settlers, this device is suitable for very large throughputs provided that the liquid system can withstand the high energy input of the mixing pump. Backmixing to a certain degree is unavoidable since the stages have hydrodynamic connections.

Multistage mixer – settlers are a valuable aid in process development on a laboratory scale. They can be integrated in "mini plants"; at this stage of development no decision should be made which type of equipment is suitable for the large-scale plant. Several examples of laboratory mixer – settlers are shown in Figure 25.

3.1.5. Centrifugal Extractors

Centrifugal extractors are used when the density difference between the two phases is insufficient to produce a countercurrent under gravity. The best known of these devices is the *Podbielniak centrifugal extractor* (Fig. 26 A), in which cylindrical perforated sheets are concentrically arranged as a compact rotor on a high-speed shaft. The heavy phase is fed in via the shaft and moves outwards due to the centrifugal force. The light phase moves in the opposite direction. Similar to columns, either the light phase or the heavy phase can be dispersed into drops, as desired. If the light phase is to be dispersed, the main interface must be located near the shaft. Depending on the system the number of stages which can be combined in one unit is limited to 3 to 4 theoretical stages.

The *Robatel centrifugal extractor* is based on a totally different principle and allows greater numbers of stages (Fig. 26 B). Disks serving as the mixing element are centered on a vertical stator. The housing is divided by annular disks into a number of stages and rotates around the central shaft. Each stage consists of a mixing chamber and a settling chamber. The disks each dip into the mixing chamber. An extensive review on operation and calculation of centrifugal extractors is given in [6, Chap. 14].

Centrifuges are more complex than columns, their investment and maintenance costs are correspondingly high. They can therefore only stand comparison with columns in cases where

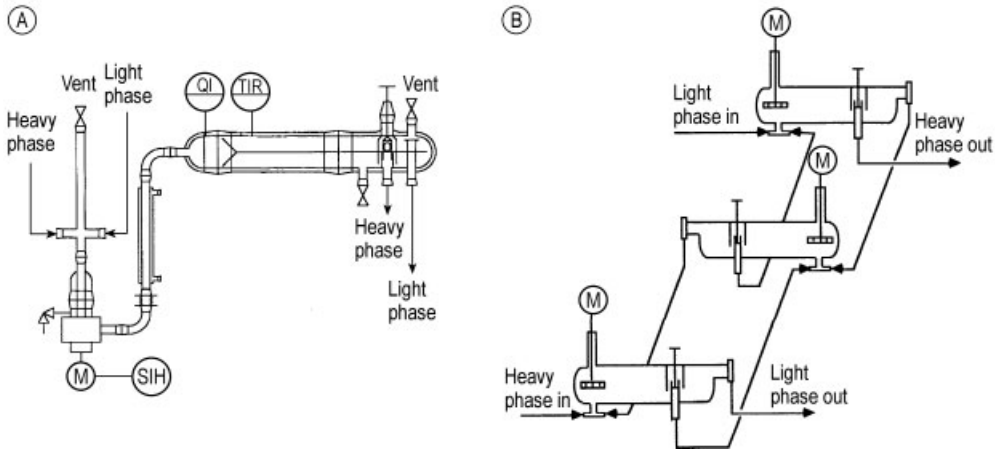

Figure 25. Laboratory mixer – settlers
A) Normag mixer – settler; B) Schott/QVF mixer – settler

they have advantages which no column can provide. Phase separation is possible even in systems having a density difference as low as 15 kg/m^3. Systems with large differences in viscosity can be processed, since the high shear forces effect intensive mixing.

As a result of its compact design, the centrifugal extractor is a space-saving apparatus with an extremely small operating volume and a short residence time. These advantages can be desirable if expensive solvents or substances of limited stability are used. The steady state is reached soon after starting. This extractor is therefore suitable for batchwise operation and frequent changes of products. Extraction under pressures of up to 5 MPa is possible.

3.2. Fluid-Dynamic Fundamentals

A comprehensive treatment of fluid-dynamic fundamentals is given elsewhere (→ Fluid Mechanics).

3.2.1. Problems and Process Strategy

In Section 2.5, it was shown how the number of theoretical stages and the flow diagram of the unit operation "extraction" are calculated on the basis of equilibria. The selection and design of apparatus are based on fluid dynamics and are not carried out until the thermodynamic design is completed. At present, model experiments are still necessary for the design of extraction apparatus. For this reason, a knowledge of the operating characteristics of the various types of apparatus is of great interest so that the number of experiments can be substantially reduced, and preselection and preliminary sizing can be carried out at a relatively early stage of process development.

The fluid-dynamic design of an apparatus is divided into the following areas:

1) Selection of suitable internals with respect to material of construction, shape, and size.
2) Determination of the operating conditions with respect to phase selection, pulsing, and loading.
3) Sizing of the column diameter in accordance with the required throughputs of feed and solvent.
4) Sizing of the column height in accordance with the actual separation problem.

However, these four problems cannot be considered in isolation. Thus, both the internals and the operating conditions affect (1) the loading limit and hence the column diameter and (2) the separation efficiency and thus the column height. Measures for increasing the loading limit have

an adverse effect on the separation efficiency, and vice versa.

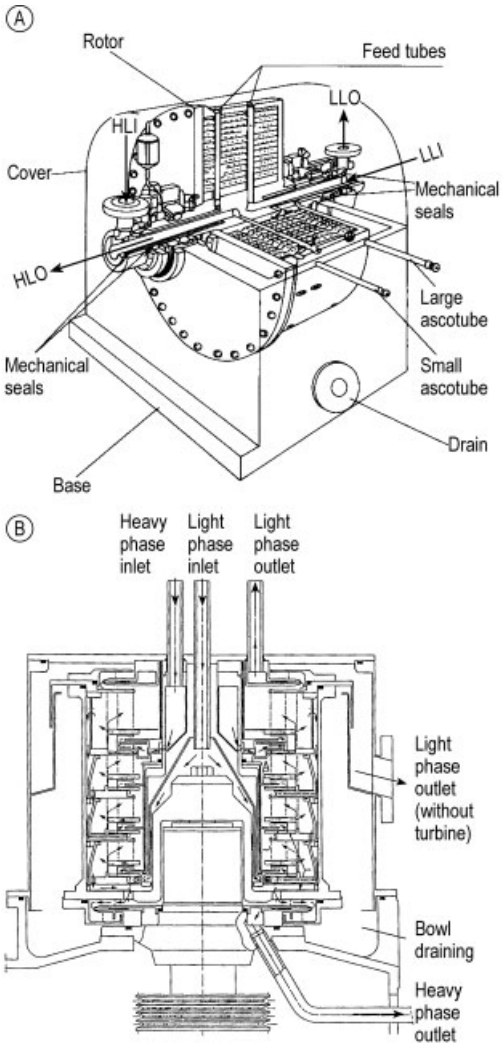

Figure 26. Centrifugal extractors
A) Podbielniak extractor; B) Robatel four-stage countercurrent extractor

The sizing of the column diameter depends on the *flooding point*, i.e., the maximum loading limit at which a column can be operated. On flooding, accumulation of the dispersed phase occurs at individual points within the column and blocks countercurrent flow. Consequently, the light phase appears in the discharge of the heavy phase, and vice versa. This phenomenon is detectable by holdup measurements. On flooding, the holdup increases continuously and indicates an unstable operating state.

The flooding limit depends on the drop size. Small drops rise less rapidly than large drops and therefore cause a greater holdup. The more energy is supplied to the system – either by pulsing or by rotating internals – the smaller are the drops, so that the flooding limit decreases.

Internals also effect dispersing. The smaller their open area, the more the flooding limit decreases.

The sizing of the column height is determined by the mass transfer, which is described by the well-known mass-transfer equation: the rate of mass transferred is equal to the product of the mass-transfer coefficient, the mass-transfer area and the concentration difference between the two phases. A great progress in scientific research has been achieved as regards measurement and calculation of the quantities of influence on mass transfer (see for instance [6], [45–49], [51], [62–65], [68]. But in practice, the various influences on mass transfer are best summarized in the expression " height equivalent of a theoretical stage" (HETS), as a measure for the separation efficiency of the column, as mentioned in Section 2.5.

Measurements in pilot-plant columns will be used below to demonstrate the way in which various parameters affect the loading limit and separation efficiency. The test systems recommended by the European Federation of Chemical Engineering (EFCE) are used for this demonstration [44].

3.2.2. Operating Characteristics of Pulsed Columns and Columns with Rotating Internals

Loading Limit. Figure 27 shows the loading limit $v_C + v_D$ as a function of the pulsing frequency for plates of small free area φ for the toluene – water system, i.e., for high interfacial tension. v_c and v_d are the volume flows of continuous (c) and dispersed (d) phase related to the total cross-sectional area of the column, the so-called superficial velocities. This flooding point curve starts at the origin of the coordinates; in other words, without pulsing, no throughput is possible. With increasing pulsing, the throughput also rises up to a maximum. This part of the curve is called the lower flooding limit; here, pulsing has mainly a transport func-

tion. With increasing frequency, however, the dispersing function predominates, the drops become smaller and the loading limit falls. The interesting region in practice is that in which dispersing predominates.

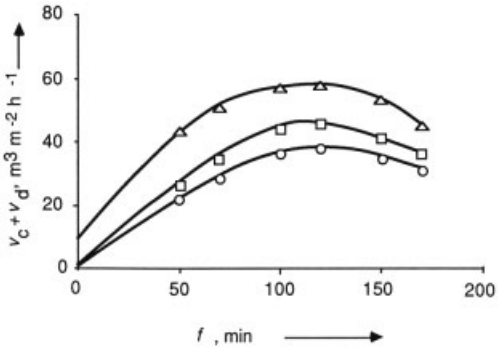

Figure 27. Loading limit as a function of the pulsing frequency at different free area for toluene – water [50]
$a = 6$ mm; $v_d/v_c = 1$
△ $\varphi = 0.60$; $d = 6.5$ mm
□ $\varphi = 0.40$; $d = 4.0$ mm
○ $\varphi = 0.23$; $d = 2.5$ mm

As is to be expected, the loading limit rises with increasing free area over the entire pulsing range, but the position of the maximum remains unchanged. At the largest free area of $\varphi = 0.6$, a modest throughput is possible even without pulsing.

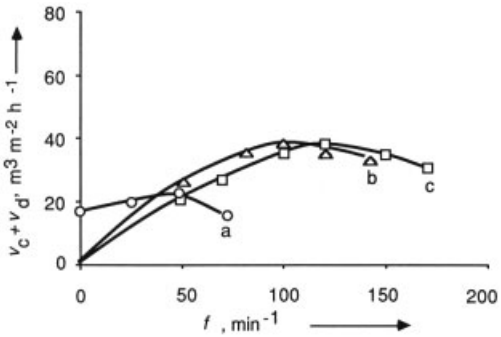

Figure 28. Loading limit as a function of pulsation frequency at different interfacial tension [50]: $\varphi = 0.23$; $v_d/v_c = 1$
a) Butanol – water: $\sigma = 1.8$ mN/m; b) Butyl acetate – water: $\sigma = 13.5$ mN/m; c) Toluene – water: $\sigma = 36.1$ mN/m

The effect of the geometry of the internals can also be shown for pulsed packed columns. There is no lower loading limit. At frequency $f = 0$, i.e., the unpulsed column, the loading limit has its maximum value, and decreases with increasing frequency [53].

Of all the physical properties, the *interfacial tension* has the greatest effect on the fluid dynamics of an extraction column. Figure 28 shows a comparison of three test systems with different interfacial tensions. As can be seen, the throughput optimum is shifted to higher pulsing frequencies with increased surface tension, i.e., high surface tensions need a high energy supply.

In addition to the physical properties, the phase ratio is another parameter determined by the process. Decisions can only be made as to which of the two phases is to be dispersed (the organic or the aqueous phase) and whether in the greater or smaller quantity. Figure 29 shows the effect of the parameter v_d/v_c [52]. As the phase ratio increases, the throughput decreases.

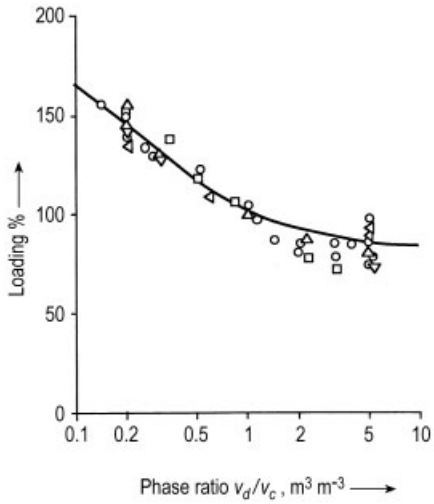

Figure 29. Relative loading of a pulsed sieve plate column as a function of flow ratio [52]
○ Toluene – water; □ MIBK – water; △ Butyl acetate – water; ◁ *n*-Butanol – water

A further parameter affecting the loading limit and separation efficiency is the direction of the mass transfer. Curve b in Figure 30 A shows the flooding point curve of the pure two-component mixture toluene – water, where no mass transfer takes place [54]. Curve a for mass transfer from the dispersed organic phase to the continuous aqueous phase lies substantially higher by comparison. The reason for this is an increased tendency to coalescence and hence formation of larger drops. In the converse case (mass transfer from the continuous to the dispersed phase), the interfacial tension and hence the drop size are reduced by the component

being transferred, leading to a lowering of the flooding point [53], [55]. The effect of the direction of mass transfer on the drop size also causes differences in the separation efficiency. In the case of mass transfer from the continuous phase to the dispersed phase (into the drop), smaller HETS values result, i.e., better separation efficiency is obtained (Fig. 30 B).

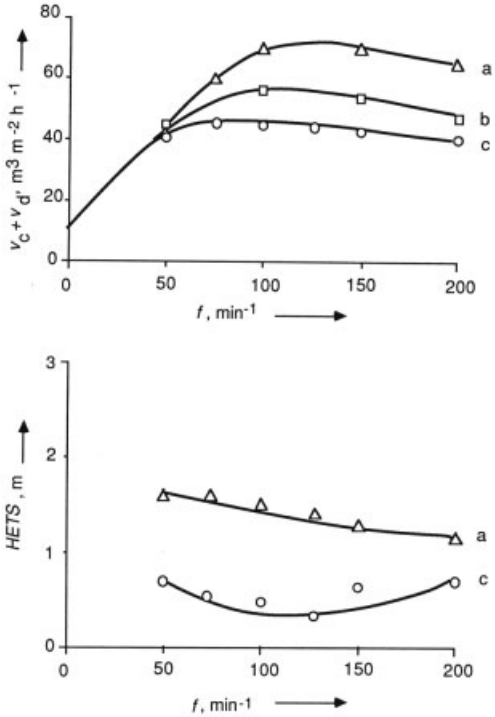

Figure 30. Influence of mass transfer on loading limit and separation efficiency for the toluene – acetone – water system ($\varphi = 0.60$)
△ Mass transfer from the disperse to the continuous phase; □ System without mass transfer; ○ Mass transfer from the continuous to the disperse phase

PILHOFER [56] has shown that the loading limit of extraction columns depends only on the drop size. Accordingly, high loading limits can be obtained only with relatively large drops.

Separation Efficiency. However to obtain a high separation efficiency a large mass transfer area, i.e., the greatest possible holdup of dispersed phase, is required. The holdup rises with decreasing drop size. Any input of energy has the purpose of producing the optimal drop dispersion. The interrelation of separation efficiency and pulsing can be demonstrated by measurements on the toluene – water – acetone system, which has a high interfacial tension [41]. Figure 31 shows the measured height of a theoretical stage (HETS) for this system as a function of the pulsation frequency, for the three different plate types (curves a – c). In all cases, the HETS decreases as expected with increasing pulsing intensity, i.e., pulsing enhances the separation efficiency. The curves show an important influence of the plate geometry: with increasing free area, the separation efficiency decreases sharply.

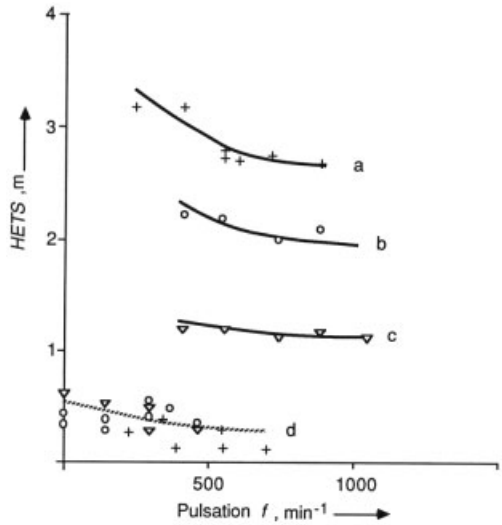

Figure 31. Separation efficiency measured as a function of interfacial tension [41]
a) – c) Toluene – water – acetone;
d) Butanol – water – succinic acid
+ $\varphi = 0.60$; ○ $\varphi = 0.40$; ▽ $\varphi = 0.23$

The butanol – succinic acid – water system shows very favorable HETS values because of its extremely low interfacial tension (curve d). The effect of free area is negligible. This means that, for systems of low interfacial tension, the free area can be increased in favor of higher throughputs, without a significant loss in separation efficiency.

A comparison of these two systems clearly shows the pronounced influence of the interfacial tension on separation efficiency. This fluid-dynamic aspect should also be taken into account when selecting solvents for extraction processes. Hitherto, this selection has been made almost exclusively under thermodynamic criteria.

The relationship between plate geometry, interfacial tension, and separation efficiency, as described here for a pulsed sieve plate column, qualitatively applies also to reciprocating-plate columns and pulsed packed columns. Generally speaking, in systems with a high interfacial tension, particular attention must be paid to the choice of the internals, a higher energy input is also necessary.

3.2.3. Fluid Dynamic Calculation Methods

The calculation methods for determining the loading limit can be classified in two categories. The first category consists of empirical correlations which are based on physical and geometrical data, and energy input. The second category is based on fluid-dynamic models and presupposes knowledge of the drop size. Both method classes are applicable only to two-component systems and thus neglect the effect of mass transfer, as shown in Figure 30. The effects of wetting and of drop coalescence are likewise disregarded. Moreover, the empirical correlations in particular are frequently restricted to certain geometries and cannot be extrapolated to other designs, they can therefore only be used to obtain an approximate estimate. For final equipment sizing, an experimental determination of the loading limit in the particular type of equipment with the original process streams is absolutely necessary.

Packed Column. Unpulsed packed columns have no industrial significance any longer. A review on empirical equations to calculate holdup, drop size, flooding, mass transfer, and axial dispersion is given in [6, Chap. 8]. Performance can be improved by mechanical pulsation of the continuous phase at the expense of a reduced throughput. For a rough estimate, pulsing within reasonable ranges may be assumed to cause the load limits to fall by 20 – 30 % as compared with the unpulsed column. Still, no equations for calculation are known.

Sieve Plate Column. The conventional sieve plate column has been investigated by Pilhofer [66] for toluene – water systems. Hirschmann [69] expanded the investigations of Pilhofer to liquid systems with lower surface tensions and density differences than toluene – water and to dual flow plates. For calculation see [69].

Sieve Plate and Rotating Columns. Special empirical equations are given in the literature for each type of extraction columns in order to calculate the flooding condition, so, e.g., for sieve plate columns in [66] and [67], for pulsed sieve plate columns in various chapters of [6], for rotating columns in [6] and [5]. However, since important effects are disregarded as mentioned before, here it will be sufficient to introduce a rough method based on a fluid dynamic model [66]. The model, originally developed for unpulsed sieve tray columns, proved to be applicable to all types of columns. It is assumed that the liquids occupy a share of the cross-section according to their holdup ε and flow toward each other as layers. If their superficial velocity v, i.e. their volumetric flow rate V related to the complete column cross section S, is

$$v_\mathrm{c} = \frac{V_\mathrm{c}}{S}\ ;\ v_\mathrm{d} = \frac{V_\mathrm{d}}{S} \tag{1}$$

they have an effective velocity $v_\mathrm{c}/(1-\varepsilon_\mathrm{d})$ or $v_\mathrm{d}/\varepsilon_\mathrm{d}$. That yields the relative velocity v_r, also called *slip velocity*, between the two liquids

$$v_\mathrm{r} = \frac{v_\mathrm{d}}{\varepsilon_\mathrm{d}} + \frac{v_\mathrm{c}}{1-\varepsilon_\mathrm{d}} \tag{2}$$

Either v_r or ε_d can be calculated from Equation (2) if one or more relationships between the two terms are known. In [45] a simple empirical equation is given to calculate v_r

$$\frac{v_\mathrm{r}}{v_\mathrm{p}} = (1-\varepsilon_\mathrm{d})^k \tag{3a}$$

$$k = 4.6 - 0.13 \ln Ar \tag{3b}$$

when $1 < Ar < 10^7$ and $2.5 < k < 4.6$

v_P is the individual drop velocity given by [66]

for *circulating drops*

$$Re_p = K_L^{0.15}\left(Ar^{0.523} K_L^{-0.1435} - 0.75\right) \tag{4a}$$

if

$$1.83\, K_\mathrm{L}^{0.275} < Ar < 3.94\, K_\mathrm{L}^{0.275}$$

for *oscillating drops*

$$Re_p = K_L^{0.15}\left(4.18 Ar^{0.218} K_{\mathrm{L}}^{-0.00773} - 0.75\right) \quad (4b)$$

if

$$Ar > 372.9 K_L^{0.275} \quad (4c)$$

The dimensionless numbers are defined as follows:
particle Reynolds number

$$Re_{\mathrm{p}} = \frac{v_{\mathrm{p}} d_{\mathrm{p}}}{\nu_{\mathrm{d}}} \quad (5)$$

Archimedes number

$$Ar = \frac{d_{\mathrm{p}}^3 g \varrho_{\mathrm{c}} \Delta\varrho}{\eta_{\mathrm{c}}^2} \quad (6)$$

characteristic number of liquid

$$K_{\mathrm{L}} = \frac{\varrho_{\mathrm{c}}^2 \sigma^3}{\Delta\varrho g \eta_{\mathrm{c}}^4} \quad (7)$$

density difference

$$\Delta\varrho = |\varrho_{\mathrm{c}} - \varrho_{\mathrm{d}}| \quad (8)$$

Mathematical formulation implies the fact that an increase of the flow rate at flooding cannot increase the drop holdup:

$$\frac{\mathrm{d}v_{\mathrm{d}}}{\mathrm{d}\varepsilon_{\mathrm{d}}} = 0 \quad (9a)$$

when v_{c} = constant,

$$\frac{\mathrm{d}v_{\mathrm{c}}}{\mathrm{d}\varepsilon_{\mathrm{d}}} = 0 \quad (9b)$$

when v_{d} = constant.

With Equation (2) for the slip velocity v_{r} the equations for calculating the superficial velocities v_{c}, v_{d} at the flooding point are obtained in a dimensionless form:

$$v_{\mathrm{c}}\left(\frac{\varrho_{\mathrm{c}}}{\Delta\varrho v_{\mathrm{c}} g}\right)^{1/3} = Ar^{-1/3}(1-\varepsilon_{\mathrm{d}})^2\left(Re_{\mathrm{r}} + \frac{\varepsilon_{\mathrm{d}}\,\mathrm{d}Re_{\mathrm{r}}}{\mathrm{d}\varepsilon_{\mathrm{d}}}\right) \quad (10a)$$

$$v_{\mathrm{d}}\left(\frac{\varrho_{\mathrm{c}}}{\Delta\varrho v_{\mathrm{c}} g}\right)^{1/3} = Ar^{-1/3}\varepsilon_{\mathrm{d}}^2\left(Re_{\mathrm{r}} - (1-\varepsilon_{\mathrm{d}})\frac{\mathrm{d}Re_{\mathrm{r}}}{\mathrm{d}\varepsilon_{\mathrm{d}}}\right) \quad (10b)$$

Equations (10a) and (10b) together with Equation (3) for v_{r} and Equation (4) are the basis for flooding point diagrams according to Figure 32. Ordinate and abscissa contain the superficial velocities v_{d} and v_{c}, i.e., the phase flow rates, modified by physical data groups to a dimensionless form. The diagram shows the calculated flooding curves of *circulating* drops as solid lines with the parameter Ar. Dashed lines of constant phase ratio $v_{\mathrm{d}}/v_{\mathrm{c}}$ and the total throughput as the sum of v_{d} and $v_{\mathrm{c}}$are also shown.

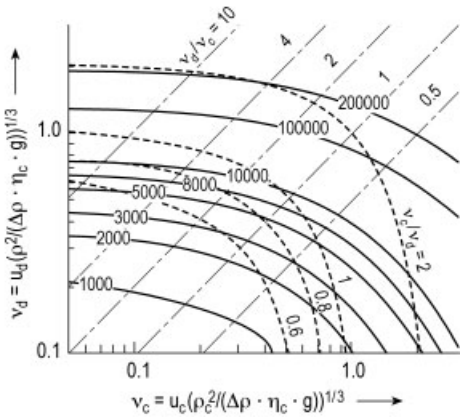

Figure 32. General loading diagram [52], [56]

It can be seen from Equation (4b) that Re_{P} is approximately proportional to $Ar^{0.3}$ and therefore to d_{P}. That means that the drop velocity v_{P} is nearly independent of drop diameter d_{P} in the region of oscillating drops. Therefore Figure 32 can be used also for *oscillating drops* if Ar is replaced by K_{L} using Equation 4c. Otherwise the diagram is valid for a certain K_{L} value in the region of circulating drops, in case of Figure 32, $K_{\mathrm{L}} = 10^{11}$.

The dependence of drop diameter d_{P} upon parameters of influence must be known in order to use Figure 32 and to calculate flooding curves. Many scientific investigations on drop disintegration in various extractors are found in the literature, but often without considering the influence of mass transfer and the interaction of dispersing and coalescence. Furthermore, particularly for higher interfacial tension systems a considerable height of column is required to reach a steady-state drop size, and this height varies with liquid system and inlet arrangement for the dispersed phase. Since population balances are not used in this introductory article,

the concept of mean drop sizes is applied – normally reported as the Sauter mean or surface percentage mean drop size – defined by

$$d_{32} = \frac{\sum_i n_{\mathrm{p}i} d_{\mathrm{p}i}^3}{\sum_i n_{\mathrm{p}i} d_{\mathrm{p}i}^2} \tag{11}$$

where $n_{\mathrm{p}i}$ represent the numbers of drops of the class i of diameter $d_{\mathrm{p}i}$.

Finally, discussion will be restricted to empirical equations by KUMAR and HARTLAND [6, Chap. 17] which give the best fit to experimental data and have the broadest range of validity:

Pulsed sieve tray and reciprocating extractor

$$\frac{d_{32}}{\sqrt{\frac{\sigma}{\Delta\varrho g}}} = C_1 e^{0.74} \left(\frac{h}{0.05\mathrm{m}}\right)^{0.10} \cdot \left[\exp\left(-3.00\frac{af\Delta\varrho^{1/4}}{g\sigma^{1/4}1/4}\right) + \exp\left(-28.65\frac{af\Delta\varrho^{1/4}}{g\sigma^{1/4}1/4}\right)\right] \tag{12}$$

The best values of the parameter C_1 are 1.51, 1.36, and 2.01 for no mass transfer, c → d, and d → c direction of transfer, respectively. The product $a \cdot f$ is considered as the agitation variable in the equation, since the fit could not be improved if a and f were treated separately. The average absolute value of the relative deviation in the predicted values of d_{32} from the experimental points is 16.3 %.

Rotating disk extractors

$$\frac{d_{32}}{D_{\mathrm{R}}} = \left(\frac{C_1}{0.07 + \sqrt{Fr_{\mathrm{R}}}}\right)\left(\frac{\eta_{\mathrm{c}}}{\sqrt{\sigma\varrho_{\mathrm{c}}D_{\mathrm{R}}}}\right)^{-0.12} \left(\frac{\varrho_{\mathrm{d}}}{\varrho_{\mathrm{c}}}\right)^{0.16} \cdot \left(\frac{D_{\mathrm{R}}^2\varrho_{\mathrm{c}}g}{\sigma}\right)^{-0.59} \left(\frac{h}{D_{\mathrm{C}}}\right)^{0.25}\left(\frac{D}{D_{\mathrm{R}}}\right)^{0.46} \tag{13}$$

The optimized values of C_1 are 0.63, 0.53, and 0.74 for no mass transfer, c → d and d → c directions of transfer, respectively. The effect of holdup is ignored due to lack of data. Equation (13) predicts the drop size with an average absolute value of the relative deviation of 23 %.

Kühni columns

$$\frac{d_{32}}{D_{\mathrm{R}}} = C_1 e^{0.37} n_{\mathrm{S}}^{-0.11} \left[0.14 + \exp\left(-18.73 Fr_{\mathrm{R}}\right)\right] \cdot \left(\frac{\eta_{\mathrm{c}}}{\sqrt{\sigma\varrho_{\mathrm{c}}D_{\mathrm{R}}}}\right)^{-0.20}\left(\frac{D_{\mathrm{R}}^2\varrho_{\mathrm{c}}g}{\sigma}\right)^{-0.24} \tag{14}$$

The values of the constant C_1 are 9.81×10^{-2} for no mass transfer and c → d transfer and 0.31 for d → c transfer. The stage number n_{S}, which varies from 2 to 17 in the present set of data, shows a rather weak effect on drop size. Equation (14) predicts the drop diameter with an average absolute value and relative deviation of 17.6 %.

3.3. Apparatus Design

Extraction apparatus can be approximatively designed purely by calculation using fluid-dynamic models [49], [70]. For reliable design, however, pilot-scale tests are indispensable. If possible, the test should be carried out with the original operating mixture in the type of apparatus which is also envisaged for the large-scale unit. Desirably, the geometry of the internals should be the same in the pilot-scale and industrial apparatus. However, this condition cannot always be satisfied. Generally, a column diameter of at least 60 – 80 mm is necessary for design tests. Tests in smaller columns cannot be used as the basis for equipment sizing (for criteria on equipment selection, see Section 3.4). The number of tests can, however, be substantially reduced if knowledge of the operating characteristics is utilized.

The physical data and the phase ratio are determined by the process used, the following parameters are available for optimization: pulsing or rotation, geometry of the internals, material of the internals, and phase selection, i.e., the direction of mass transfer. Energy input and the geometry of the internals are selected on the basis of the physical data.

3.3.1. Internals and Operating Conditions

The following guidelines apply for the selection of the free area in pulsed columns. An increase in the free area is recommended in systems with a low interfacial tension where it leads to a great reduction in the column volume. At moderate interfacial tension, the column diameter can be reduced by a greater free area, but at the expense of the column height. In the case of extremely

Case		Dispersed phase	Location of interface	Arrangement	Plate material
1 Aqueous feed	$\rho_F > \rho_S$	Org. solvent	Head	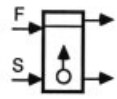	Metal
2 Aqueous feed	$\rho_F < \rho_S$	Org. solvent	Bottom	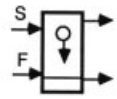	Metal
3 Organic feed	$\rho_F > \rho_S$	Water	Head	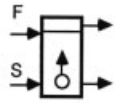	Plastics
4 Organic feed	$\rho_F < \rho_S$	Water	Bottom	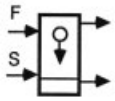	Plastics

Figure 33. Selection of phases and plate materials

high interfacial tension, there is no point in increasing the free area.

The material of the internals should be selected so that the continuous phase, but not the dispersed phase, wets the plates. As a rule of thumb, metal plates can be used when the organic phase is dispersed; if the aqueous phase is dispersed, the internals should be of plastic, both in pulsed columns and in rotary-agitated columns.

In deciding which phase should be dispersed, the direction of mass transfer must be taken into account. In accordance with Figure 30, mass transfer should be from the continuous phase to the dispersed phase, if possible. In the final phase selection, however, fluid-dynamic considerations frequently conflict with operating and safety requirements. The position of the interface in the column determines the phase which is dispersed. Figure 33 summarizes the possible combinations.

The separation efficiency determines the column height and the load limit determines the column diameter. For optimization, the two parameters are related by the ratio $\theta = \text{HETS}/(v_c + v_d)$ which can be used to optimize the column internals and energy input. A minimum value of this expression is equivalent to a minimum column volume. Figure 34 shows that optimum pulsing is obtained between the maximum of the flooding point curve and the minimum of the HETS curve. An analogous procedure is used for rotary agitated columns. The design point for industrial-scale columns is chosen to be 75 – 80 % of the load limit. The operating point with respect to load and pulsing (rotation) is fixed by this procedure.

3.3.2. Column Diameter

Assuming the same column geometry, the specific throughput of the test column is adopted for the large-scale column.

$$(v_c + v_d)_2 = (v_c + v_d)_1$$

with

$$(v_c + v_d)_1 = 0.75 \cdot (v_c + v_d)_{1\max}$$

Index 1 = pilot scale
Index 2 = large scale

This gives the column diameter:

$$D_2 = \sqrt{\frac{4\left(\frac{\dot{m}_F}{\varrho_F} + \frac{\dot{m}_S}{\varrho_S}\right)_2}{\pi \cdot (v_c + v_d)_2}}$$

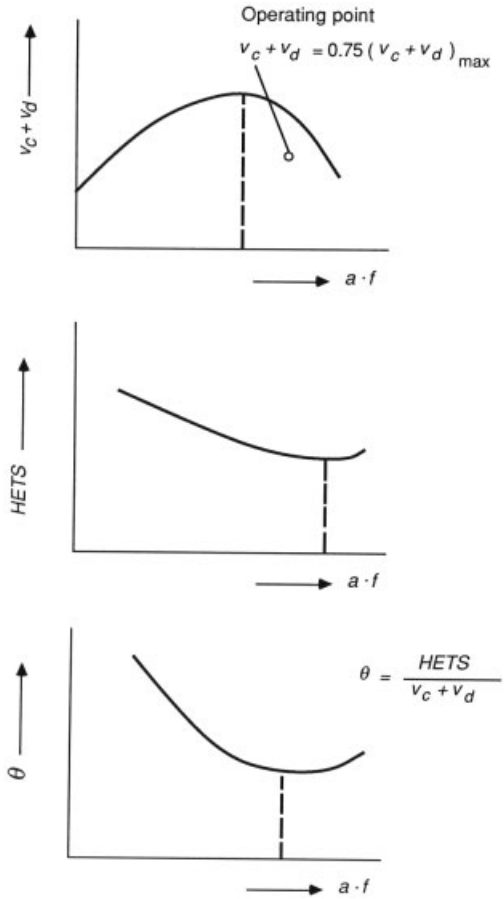

Figure 34. Optimization of pulsation

3.3.3. Column Height

The separation efficiency of all columns decreases with increasing column diameter, because the residence time distributions of the phases increase. Therefore values of the height of a theoretical stage, measured for a laboratory column, need to be corrected by empirical scale-up equations to be applicable for industrial-scale columns. However, scientific investigations of this problem are rather poor and mainly empirical knowledge of industrial engineers exists. The different chapters of [5] list formulae from experience with countercurrent extractors as described here, including

Pulsed sieve tray columns [71]

$$\frac{\mathrm{HETS}_2}{\mathrm{HETS}_1} = \exp\left[1.64\,(D_2 - D_1)\right] \tag{15}$$

(in Equation 15 the factor 1.64, given by THORNTON, is only valid for small columns. Figure 35 shows an empirical dependence of the Thornton factor from the column diameter published by REISSINGER [49])

Reciprocating columns [72]

$$\frac{\mathrm{HETS}_2}{\mathrm{HETS}_1} = \left(\frac{D_2}{D_1}\right)^{0.38} \tag{16}$$

When the plate geometry and amplitude of the vibration are left unchanged in the large-scale column, the frequency should be reduced, according to:

$$\frac{f_2}{f_1} = \left(\frac{D_1}{D_2}\right)^{0.14}$$

The following applies to stirred columns [73]:

$$\frac{\mathrm{HETS}_2}{\mathrm{HETS}_1} = \left(\frac{D_2}{D_1}\right)^{0.33}$$

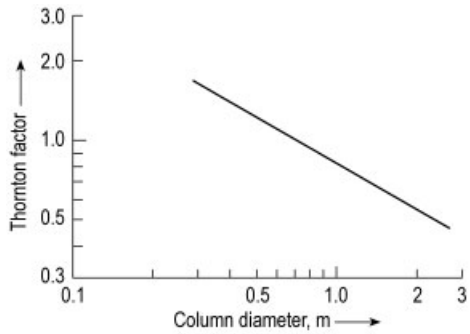

Figure 35. Dependence of Thornton factor on column diameter

3.4. Criteria for Equipment Selection

For a given separation problem, equipment is selected essentially according to three criteria: physical properties, required number of separation stages, and throughput. Further operating aspects are frequently important, for example, flexibility regarding load variations or product quality, sensitivity to small amounts of solids, suitability for batch operation.

The physical properties of interest in extraction are the density difference between the two

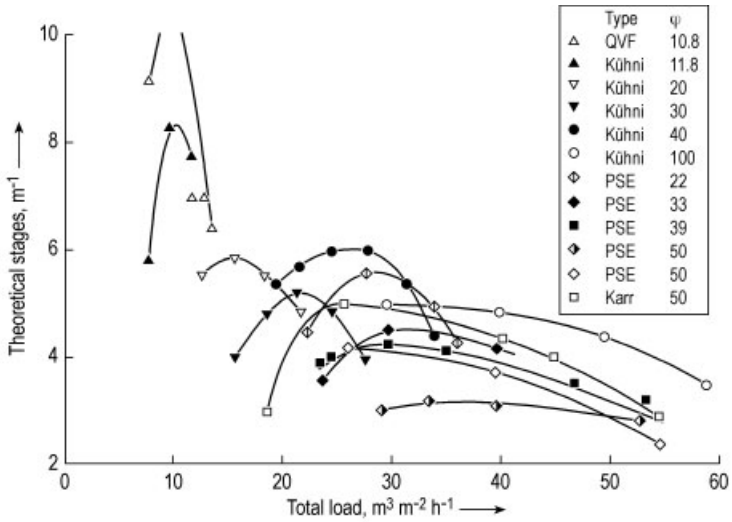

Figure 36. Comparison of the number of theoretical stages NTS in dependence of total load for Kühni, QVF, Karr and pulsed sieve plate extractors (PSE) [64] (system butyl acetate – water – acetone, phase ratio $v_d/v_c \approx 1.12$, measurements of various authors cited in [64])

phases, their viscosity, and interfacial tension. These three parameters affect dispersibility and phase separation behavior.

High density difference, viscosity, and interfacial tension make dispersing more difficult; systems with such physical properties require a higher dissipation of energy than in the converse case. Phase separation is favored by a high density difference and interfacial tension, whereas a high viscosity makes phase separation more difficult.

The lower limit for the density difference can be taken to be 50 kg/m^3. Mixtures with higher density differences can generally be processed in columns and mixer – settlers. If the density difference is smaller, a centrifugal extractor must be used.

A low interfacial tension leads to phase separation problems. In industrial units, approximately $\sigma = 1 - 2$ mN/m can be taken as the lower limit. Therefore, mixer – settlers can certainly not be used for mixtures of relatively low interfacial tension; because of the shear field generated, centrifugal extractors also have a dispersing effect which should not be underestimated. In such cases, the energy input must be very carefully controlled. Rotary-agitated columns and pulsed columns can be used, in which the energy input can be adapted to the physical properties by the speed of rotation or the pulse frequency.

The selection of equipment is also affected by the number of theoretical stages (NTS) required for a given separation problem. Only small numbers of theoretical stages can be achieved in columns without energy input, because of their poor separation efficiency. Pulsed columns and rotary agitated columns give higher numbers of separation stages. They are limited by the pulsing units or by the lengths of the continuous shaft. In centrifugal extractors, the Robatel design provides more separation stages than, for example, the Podbielniak extractor. Mixer – settlers are constructed in modules so that there is virtually no limit to the number of separation stages.

A first estimate of the maximum possible throughputs in the individual extractor types can be obtained by considering the largest diameters (D_{max}) which have hitherto been installed:

	D_{max}, m
Reciprocating-plate column	1.5
Pulsed column	2.7
Kühni extractor	2.5
QVF extractor	2.8
Rotating Disk Extractor	3.0
Mixer – settler tower	8.0

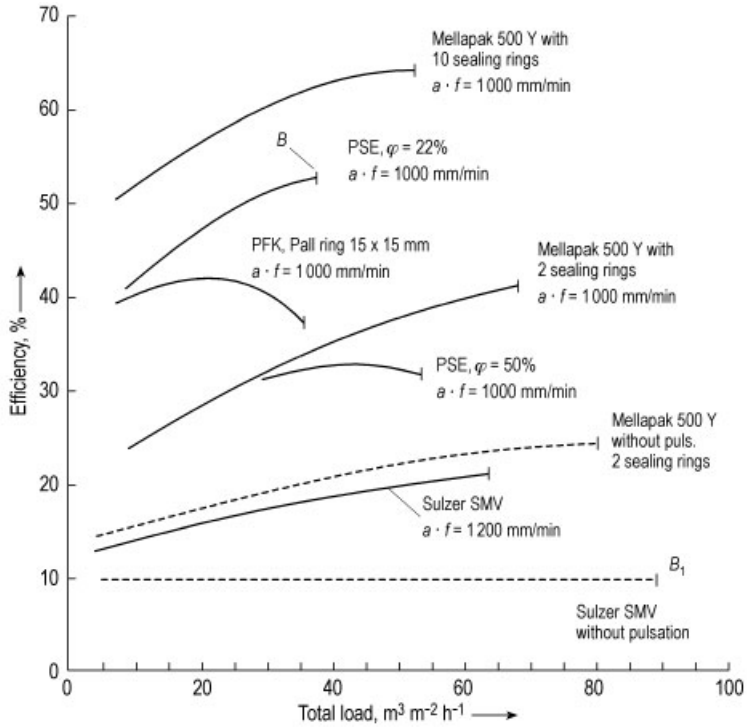

Figure 37. Comparison of efficiency versus load of packed columns with and without pulsing and pulsed sieve tray extractors $a \cdot f$ = pulsation intensity, B_f = flooding load, φ = relative free area

Accordingly, the highest throughputs can clearly be achieved with the mixer – settler. In pulsed columns, the diameter is limited by the pulsing unit. In rotation columns, the separation efficiency at large diameters is no longer satisfactory. However, informations about maximal cross-sectional loads which are necessary to obtain definitive throughput data of a type of extractor are not given in the list.

The literature contains some publications which recommend criteria for the selection of the best apparatus. However, most of these recommendations are only qualitative, and some are even contradictory; no studies of apparatus have so far allowed an objective comparison. Only the standardization of test systems for extraction has allowed fluid-dynamic studies to be carried out under comparable conditions [44]. The following systems have been recommended:

Toluene – water – acetone
Butyl acetate – water – acetone
Butanol – water – succinic acid

These systems provide an objective basis for comparison of equipment. Nevertheless, investigations by different authors are not always mutually comparable with respect to the absolute HETS values. Evidently, the column length and the phase ratio have an unexpectedly large effect.

For the toluene system, Pilhofer compared the different operating characteristics of various groups of equipment. Analogous studies on the liquid system butyl acetate – water – acetone were carried out by Goldmann [64]. Figure 36 shows a diagram of Goldmann with the number of theoretical stages NTS per meter of column height, as a function of the load. The figure contains data for QVF and Kühni columns – the latter with five different free areas of the perforated partition disks – and for pulsed sieve plate extractors (PSE) with five different free areas φ and one reciprocating column (Karr). The rotary agitated columns of QVC and Kühni with the smallest free areas φ clearly achieve the highest NTS values, but allow only low loads. Also, the other Kühni columns show higher NTS values than the

pulsed columns but smaller ranges of load. Generally speaking, all measures which improve the separation efficiency reduce the load capacity.

A comparison between pulsed sieve trays and structured packings with the test system toluene – water – acetone was published [74]. Figure 37 shows the diagram with the extractor efficiency in dependence on the load. It is obvious that structured packings have the broadest range of load capacity and reach the same or higher efficiencies as pulsed sieve trays. From the results for two structured packings without pulsation the positive influence of pulsing can clearly be seen: The pulsed unstructured packing, Pall rings 15 × 15 mm, has a clearly smaller load range.

In Figure 38, the range of loads is compared for sieve trays and structured packings with pulsation [75]. According to that, packed columns have significant advantages. As shown in [75], the investment costs of pulsed structured packed columns are 30 % lower than the ones of pulsed sieve tray columns and 40 to 50 % lower than pulsed unstructured packed columns.

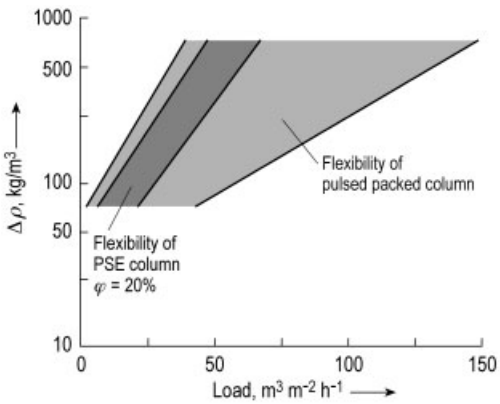

Figure 38. Range of load of pulsed structured packed columns in comparison to pulsed sieve tray columns in dependence of the mass density difference for aqueous systems and systems similar to water

In [76] a systematic study on extractor selection is given.

Adaptation of the Column Geometry to the Separation Problem. In the early days of extraction, various types of apparatus were restricted to clearly defined geometries and were thus sharply distinguished from one another. Recent apparatus developments allow adaptation of the column geometry to physical data, both for pulsed columns and rotary-agitated columns. The boundaries are thus no longer distinct and a general fluid-dynamic description of related types of apparatus is feasible. The optimization of operating conditions and the geometry of the internals is as equally important as the selection of a suitable type of apparatus.

In practice the final selection of equipment is often based on product-specific conditions or plant requirements which include sensitivity to solids, tendency of the product to polymerize, and safety aspects, for example with regard to solvent holdup.

Untypical Operating Conditions. Usually extractors operate with nearly the same volume flow for both liquids and for liquids with viscosities similar to that of water. Therefore, the equations given before are only valid for such conditions. Some approaches for untypical operating conditions are given below.

Rauscher [62] investigated pulsed sieve tray extractors with phase ratios between 0.06 and 10 and the liquid system *n*-butyl acetate (o) – water (d) – methyl isopropyl ketone (MIPK), which enriches MIPK in equilibrium preferable in the organic phase. He confirmed the experience that the separation efficiency declines for phase ratios unequal 1 and found the following causes: By dispersion of the smaller volume flow (phase ratio $\ll 1$), stronger backmixing of droplets by the continuous phase will occur, leading to a broader drop size distribution. By dispersion of the larger volume flow (phase ratio $\gg 1$), the wake behind the drops causes remarkable backmixing of continuous phase. These phenomena have to be taken into account by an adequate calculation via drop population balances.

G. Wagner [58] investigated the fluid dynamics of pulsed sieve tray and Kühni rotating extractors using liquid systems with viscosities up to 50 mPa · s, mixed from paraffin oil and *n*-hexane as organic phase and of glycerol and water as aqueous phase. I. Wagner [77] continued the work with mass transfer investigations using the system tridecanol as drop phase ($\eta = 42$ mPa · s), water, and propanol. The experiments showed that the mean Sauter drop diameter increases with increasing drop viscosity; holdup is decreased, while the flooding curve is shifted to higher throughputs. In case the

continuous phase is the more viscous one, the drop swarm contains a larger fraction of small droplets causing increased droplet entrainment by the continuous phase. In stirred columns, the drop size can be influenced in a broader range than in pulsed columns; yet, more energy will be required for a certain drop dispersion. Intensive stirrers as in Kühni and QVF extractors should be preferred to stirrers in RDC extractors.

In the case of a more viscous continuous phase, the drop dispersion is nearly independent of drop viscosity and is hardly influenced by the rotating or pulsating internals. Here, lower intensity dispersers are preferable, such as RDC extractors.

The mass transfer to drops is distinctly decreased by increasing drop viscosity, developing towards the state of transfer to rigid spheres, and depends on the contact time. It can be improved by suitable internals and by supply of energy.

4. Phase-Separation Equipment

4.1. Gravity Settlers without Inserts

Each extraction stage consists of two steps: (1) dispersing and (2) phase separation by coalescence of drops, thus generating a homogeneous layer of drop phase. Figure 39 A shows a vertical settler at the top of a column for the case that the drop phase has a lower density than the continuous phase. The drops rise against the sinking continuous liquid and are blocked by the upper level of the continuous phase, which is the principal interface. There, coalescence of the dispersed phase will occur both between adjacent drops and at the principal interface. The coalesced drop phase leaves the extractor at the top of the column.

In a horizontal settler, as shown in Figure 39 B, both phases flow in the same direction. Drops of lighter dispersed phase rise to the interface where they coalesce to form a homogeneous layer. In the lower part of the settler a layer of the continuous phase remains, growing in flow direction due to decreasing drop numbers. The wedge-shaped zone of dispersion should end before the placement of phase outlets.

As the accumulation of impurities in technical liquid systems changes important properties such as viscosity and surface tension, an exclusively theoretical modeling, as developed by many authors (see for instance [80]), is insufficient. However, the length of a horizontal settler can be estimated by a combined procedure of calculations and simple and cheap batch settling experiments [79]. The calculation is based on a model of the physical processes in a settler, while experiments supply one empirical parameter to include surface phenomena that could not be modeled so far (see Figs. 40 and 41).

A glass cylinder as shown in Figure 40 A, filled with about 1 L of the original liquid – liquid system, is vigorously shaken for 1/2 to 1 min. After placing it in an upright position, the heights of sedimentation and coalescence fronts are measured in dependence of time as shown in Figure 40 B for drops being lighter than the continuous phase. The drops are banked up in a dense layer if sedimentation is faster than the drop coalescence. The curves of sedimentation and coalescence meet each other at time t_e when all drops are coalesced. For a sufficient reproducibility of measurements it should be made sure that the glass flask is washed with deionized water, that the distribution of drop sizes is narrow and the temperatures of the liquid – liquid system in the glass cylinder and the continuous settler are equal [78]. Figure 37 C shows the height H of the wedge-shaped zone between the curves of sedimentation and coalescence.

The concept of settler dimensioning is shown in Figure 41. All equations mentioned in the following are given in [79]. Settler dimensioning starts with evaluation of settling experiments according to Figure 41 A. The linear curve of sedimentation supplies the velocity of sedimentation; a model of drop swarm sedimentation supplies the diameter of the starting drop swarm [66]. The thickness of the dense layer of drops follows from a balance of sedimenting and coalescing drops in time steps. The next stage is calculation of drop deformation in vertical steps [80]. This serves to calculate coalescence times between drops and between drops and their homophase, the former to determine the drop growth, the latter to calculate the drop homophase formed by coalescence [80]. The coalescence model contains a fitting parameter F_p to fit the calculated and measured coalescence curve in Figure 41 B. This parameter is the link between the settling experiment and the settler

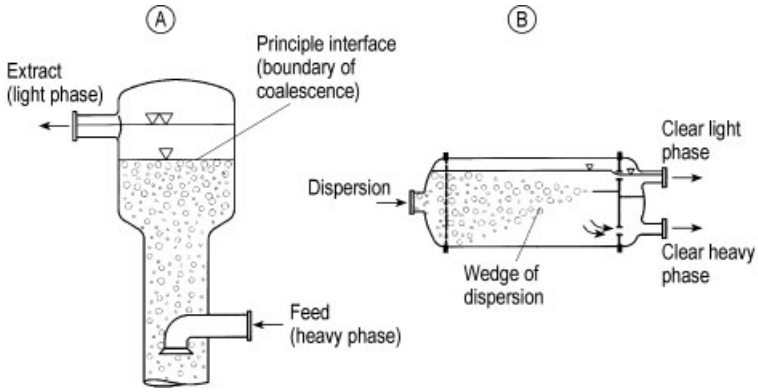

Figure 39. Vertical (A) and horizontal (B) gravity settlers for drop separation from a continuous liquid

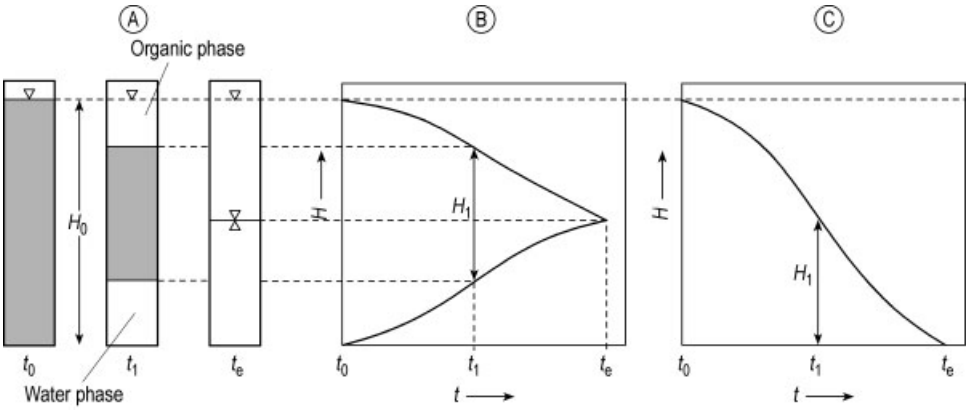

Figure 40. Experiments on batch settling

dimensioning; F_p can be used for the settler calculation, given in Figure 41 B.

The settler needs an inlet length L_{in} which can be calculated by a model equation in dependence on settler diameter, operation parameters, and the mean drop size at the settler inlet, coming from the extractor outlet. For calculation of coalescence in the dense drop layer the same equations that were modeled for the settling experiment can be applied using the fitting parameter F_p derived from these experiments. The result is the geometry of the wedge of dispersion, i.e., its height and length, which has to be taken into account for dimensioning the settler.

This combined experimental and theoretical procedure yields good results although the drop distributions in the settling experiment and the real settler are not the same. The extensive equation system needs computer aided solving. A computer program is offered by [81]. A simplified method is given in [79].

4.2. Settlers with Coalescing Aids

Gravity settlers are often equipped with coalescing aids to fit a settler to special conditions of the liquid system or of the operational behavior. Industrially proven coalescing aids are inclined or corrugated plates, fixed beds of small packings, fiber beds, porous sintered materials, and membranes. In some cases, when the drop sizes to be settled are smaller than 1 µm or if surfactants render the settling more difficult, centrifuges [6, Chap. 8] or electrical fields must be applied [82]. The most important coalescing aids are inclined packages of plain or corrugated plates and fiber bed filters.

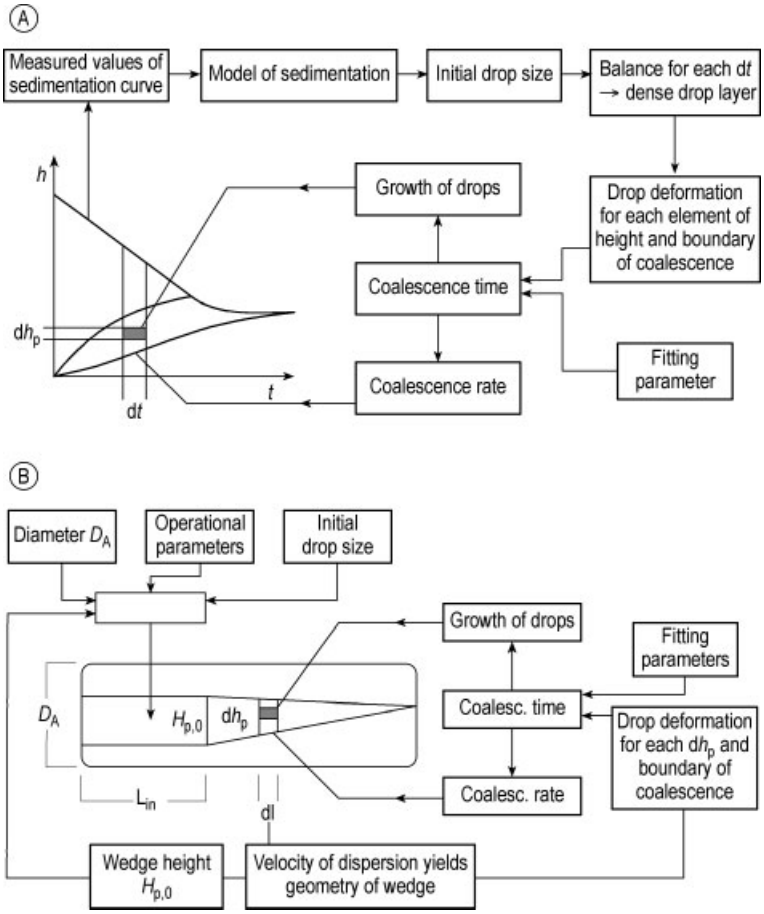

Figure 41. Dimensioning a continuous settler [79]
A) Evaluation of settling experiments; B) Fitting of measured and calculated curve

Inclined Plates. Figure 42 shows one possible arrangement of inclined plates in a horizontal settler. Other plate configurations are possible. In any case, the feed has to be distributed by a sieve plate at the settler entry over the cross-sectional area, filled with inclined plates. The plates shorten the sedimentation distance of drops and provide a larger interfacial area for the coalescence. The drops coalescing at the plates form trickling films which flow off to the horizontal, principle interface due to a hydrostatic pressure difference. A plate material wettable by the drop phase is a prerequisite. Several scientific investigations and experiments with liquid systems of a wide range of properties were made during the 1970s and 1980s on the hydrodynamics of inclined films and the coalescence of drops at the film surface, these are summarized in [83–86]. The main results are:

1) Thin laminar films should flow on the plates, because the coalescence of drops is promoted by the shear stress of film flow near the plate; the larger the volume flow of cocurrent continuous phase, the smaller the film thickness and the better the coalescence.
2) The plate should be wettable by the drop phase (advancing contact angle $< 70°$, receding contact angle $> 15°$).
3) The drop swarm has to be distributed equally over the width of plates.
4) Constructions that can thin the trickling film increase the settler performance.

5) Short lengths of plates of about 400 mm are sufficient for coalescence of drops riding on the film.
6) The width b_p of a plate can be calculated by the empirical equation of the Reynolds number of trickling film

$$Re_f \leq 80 K_{L,d}^{0.19}$$

$$b_p = 0.03 \frac{V_d}{(Re_f k v_d)^{3/2}} \left[\frac{\Delta \varrho g \sin\alpha}{\eta_d} \right]$$

with the liquid number of trickling film

$$K_{L,d} = \frac{\sigma^3 \varrho_d}{g \eta_d^4} \frac{\varrho_d}{\Delta \varrho}$$

where α is the plate inclination, $V_{d,1} = V_d/n$ the volume flow of dispersed phase on a plate, and n the number of plates.
7) The optimum of plate inclination is within the range $10° < \alpha < 15°$.
8) The distance of adjacent plates should not be smaller than 30 to 40 mm in order not to disturb film flow and drop coalescence.
9) The trickling film continuously carries surfactants to the outflow of a plate and thus prevents their accumulation. Therefore surfactants hardly influence coalescence and settling performance up to medium flow rates of drop phase.
10) Under certain conditions, the drop coalescence leaves very small droplets, so-called secondary drops, due to partial coalescence, which generate a droplet entrainment at the settler exit. Its occurrence is primarily a question of liquids involved, i.e., their viscosity ratio [86].

The coalescence time of secondary drops is significantly higher than that of primary drops, due to interfacial effects. Settling of secondary drops requires other coalescers, for instance membrane coalescers or centrifuges.

Fiber bed filters are built from short fibers stuffed to rectangular mats or to disks of 15 to 25 mm thickness, held in a suitable support, or from fiber fleeces as shown in Figure 43.

They have a definite, large voidage of about 95 %, nevertheless they are sensitive to solid dirt particles in the liquid. The droplets couple to the fibers while flowing through the fiber bed, succeeding drops coalesce with them and enlarge the size of adhering drops until the liquid flow pulls them out of the bed due to increasing flow resistance. A fiber bed is able to enlarge the sizes of incoming drops at least by one order of magnitude so that they can coalesce in a subsequent settler with their homophase. Fiber bed filters are restricted to drop concentrations < 10 vol % to avoid redispersion.

Rebelein [87], [88] and Magiera [89], [90] defined heuristic rules for the design of fiber beds on the basis of careful experiments with a broad variation of parameters of influence:

1) Fiber materials as glass or stainless steel with high energetic surfaces provide different interaction sites (e.g., hydrogen bonds and ion – dipole interaction) for the liquid and improve the fiber bed performance compared to materials with low energetic surfaces as for instance teflon. High energetic surfaces improve the wetting properties of the fibers.
2) The drop enlargement in a fiber bed decreases with increasing flow rate, i.e., with superficial velocities > 1 cm/s.
3) Small fiber diameters of about 15 µm are favorable.
4) The bed depth required for coalescence increases with the fiber diameter.
5) The concentration of droplets in the incoming liquid hardly influences the bed performance.
6) Best efficiency means a separation of about 95 % of incoming drops.
7) The larger the surface tension of the liquid system, the better the separation performance.
8) Smallest drop sizes that can be separated are about 10 µm.
9) Presence of surfactants and mud and a higher viscosity of the continuous phase reduce the separation performance, the viscosity of the drop phase has no influence within 50 and 5000 mPa · s.

All attempts to calculate the bed performance in advance via filter coefficients were fruitless [87–90]. Magiera developed a mathematical model of physical mechanisms inside a fiber bed and was able to simulate the bed behavior and drop enlargement by introduction of two experimental fitting parameters. Therefore, a reliable dimensioning of fiber beds needs experiments

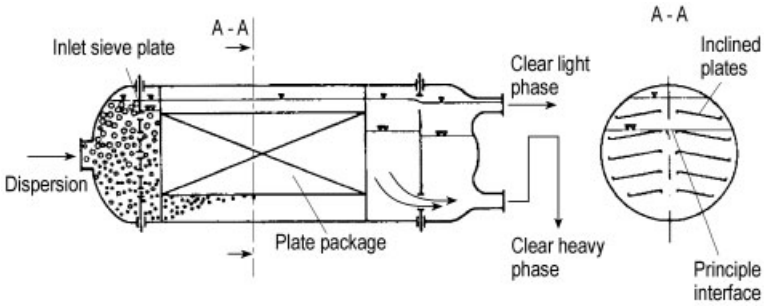

Figure 42. Horizontal gravity settler with two inclined packages of plane, parallel plates. The height of principal interface is best situated, when both phases have the same exit velocity

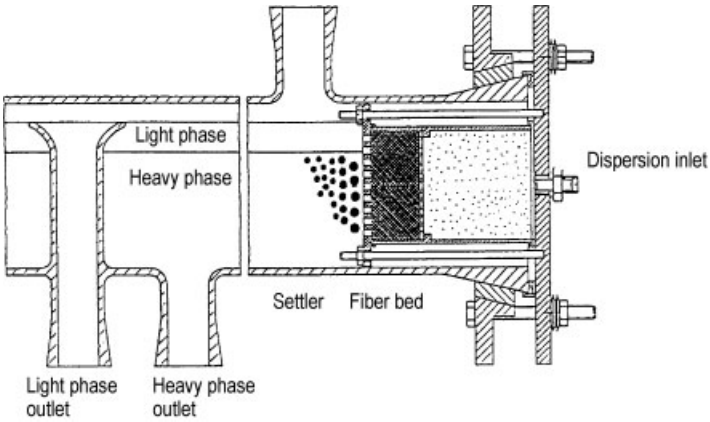

Figure 43. Combined fiber bed coalescer and gravity settler

with the liquid dispersion which is to be separated.

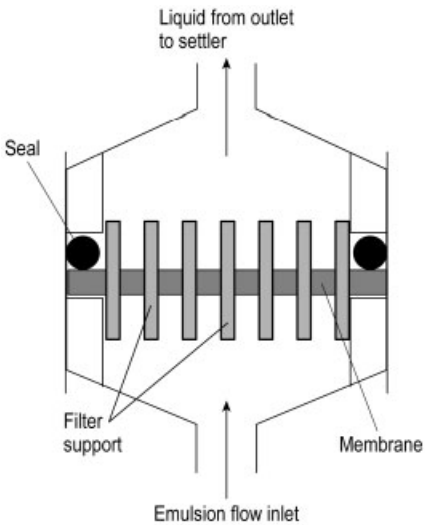

Figure 44. Membrane coalescer [91]

Membrane Coalescence. A new and very promising method to separate fine droplets down to 1 μm from emulsions, stabilized by surfactants, was discovered and investigated by HOFFMANN [91] (Fig. 44).

Membrane coalescers are applicable in a broad range of volume concentrations down to a lower concentration limit of the organic phase (0.1 vol %). A thin, hydrophobic, stretched membrane, for instance polytetrafluoroethylene (PTFE), held in place by a support, is permeated by two-phase flow of emulsion. The emulsion leaves the membrane as a liquid foam with large oil domains which will immediately coalesce at the principle interface in a succeeding settler. A superficial velocity through the membrane is possible within the range of 0.1 to 19 cm/s, the pressure drop is comparable with that in fiber beds. The main advantages are huge separation performances of about 98 %, in combination with a large range in throughput; the

ability to separate droplets one order of magnitude smaller than in fiber beds; and the insensitivity against hydrophobic surfactants. Hydrophilic surfactants are without influence, too, if their amount is clearly lower than the critical micelle concentration.

Selection of Appropriate Settler. Most important and the first step in selection of a settler is a careful analysis of separation problems [78], [92]. The flow sheet of the complete process can give hints on origin and kind of the dispersion to be separated. Sometimes the preceding steps could be altered in order to simplify the phase separation problem. Measurements of properties of the liquid system and settling experiments in a glass cylinder show, if and why a phase separation can be problematic. Drop sizes smaller than 100 µm need suitable coalescing aids which are supposed to be well wettable by the drop phase. When the drop settling in a glass cylinder is fast, but the liquids appear cloudy, this indicates partial coalescence, resulting in very fine drops, so that a fiber bed or a membrane coalescer will be necessary.

The presence of solid particles and mud requires filters before the settler. The wetting behavior of coalescing aids can significantly change the operation of the process. This should be accounted for in the procces. The phase ratio can be changed by recycling of one phase to the settler. By this, the type of dispersion can be inverted. In case the organic phase has a higher viscosity than the water phase, an o/w dispersion separates much better than a w/o dispersion, and vice versa.

5. Liquid – Liquid Extraction Processes

5.1. General

The quality of the *raffinate* in liquid – liquid extraction can be controlled easily by changing the ratio of solvent to feed and the extraction temperature. The quality of the *extract* is governed only by the conditions of feed location and no direct control is possible.

This is especially true because most extractions are carried out in a closed system to minimize solvent requirements. In such a system, a maximum extract purity is obtainable and some components that should end up in the raffinate leave with the extract.

Additional steps are, therefore, required to enhance extract purity and thus raffinate yield. If the extract is the product desired, improving extract purity is the most important – in fact, the only – objective. Different solutions to this problem are applied in commercial operation.

Change in Temperature. Extract purity can be increased by cooling the extract from the main system and adding one more stage (see Fig. 45 A). In the triangular diagram, this means that the binodal curve is changed from a to b (Fig. 45 B).

Because the composition of the first extract is now inside binodal curve b, it separates in a settler b in a purer final extract phase and a small amount of intermediate raffinate phase. The latter is returned to the main extraction system.

The selectivity of separation in the additional stage is not very great because the loading of the first extract phase is rather high and separation takes place near the plait point. This system can be used when the main system is operated at high temperature for other reasons or when an increase in extract purity and the resultant increase in raffinate yield are desirable.

Addition of Water. The addition of water at or below the feed point generally changes the system from a closed to an open one (see Section 2.1.1.). In this way, extract purity can be increased to the required level. Water is soluble, at least to some extent, in most polar solvents and therefore leaves the system with the final extract phase. In this way, no unwanted influence is exerted on the main extraction system.

However, because water must be separated from the solvent by distillation before its reintroduction in the main extractor, the heat requirements are high. Addition of water is, therefore, not regarded as a generally attractive solution. Also, water can only be added when its boiling point is much lower than those of the solvent and the feedstock.

Use of a Countersolvent. As shown in Figure 46, a second solvent can be fed into one end of the extraction system, on the opposite side of

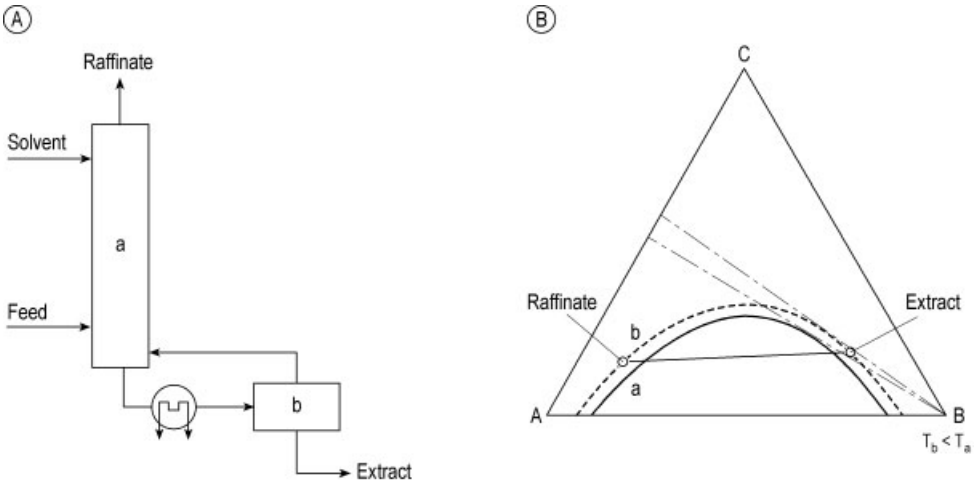

Figure 45. Use of cooling stage
A) Process flow scheme: a) Extractor; b) Settler
B) Triangular diagram: a) Binodal curve at higher temperature; b) Binodal curve at lower temperature

the solvent entry point. The feed in this case is introduced somewhere in the middle of the system.

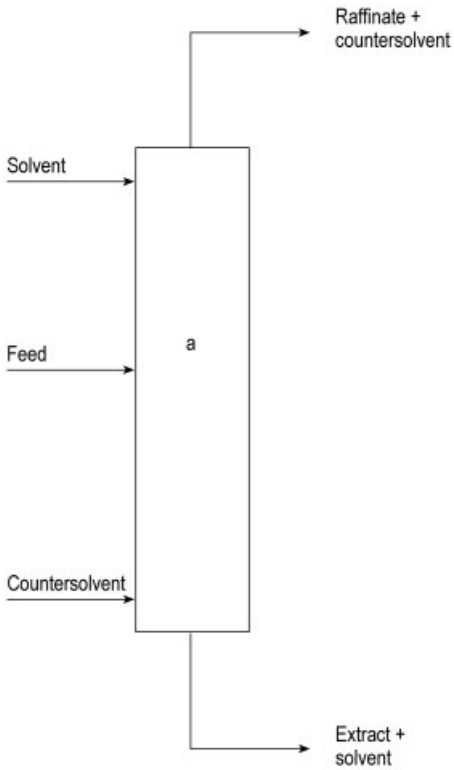

Figure 46. Extractor for separation using a countersolvent
(a) Extractor

The countersolvent should be only slightly soluble in the solvent and, therefore is generally similar in chemical structure to the key component of the raffinate phase. The boiling point of the countersolvent should be such that it can easily be separated by distillation from the solvent as well as from the raffinate and the extract (i.e., either a much lower or a much higher boiling point). Because the countersolvent does not contain any raffinate components when it enters, it can wash out the raffinate components in the solvent at the feed point.

Additionally, some extract components are washed back by the countersolvent, thus increasing the load in the top section. In this way, a very pure final extract can be produced, depending on the ratio of countersolvent to feed used.

A disadvantage of this system is that the countersolvent enters the top section. Consequently, the phase ratio of B-rich to A-rich phase is reduced, resulting in a less pure raffinate (this also means a lower yield of extract). To restore raffinate purity, more B-rich phase must be applied. Such a system is thus costly and sensitive to changes in feed composition, solvent-to-feed ratio, and ratio of countersolvent to feed.

Some improvement can be obtained by carrying out the extraction in two parts (Fig. 47).

In the first extractor (a), the feed is extracted with the primary solvent. The intermediate extract phase is fed to a second extractor (c) and

brought in contact with the countersolvent. The mixture of countersolvent and intermediate raffinate is fed to a column (b) in which the intermediate raffinate and the countersolvent are separated by distillation. The pure intermediate raffinate is then fed to the first extractor, while the countersolvent is charged to the second extractor. In this way, some of the disadvantages can be overcome.

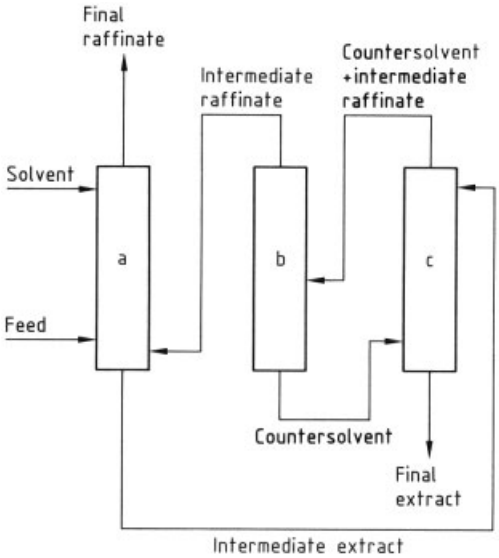

Figure 47. Use of countersolvent with intermediate recovery
a) First extractor; b) Distillation column; c) Second extractor

5.2. Combined Processes of Extraction and Distillation

Azeotropic Distillation. Usually it is necessary to combine extraction and distillation processes for recovering the solvent from the extract phase and also from the raffinate phase, if the solvent is soluble in the raffinate.

Figure 48 shows as an example the recovery of acetic acid from an aqueous solution with methyl *tert*-butyl ether (MTBE) as solvent. The system of water and acetic acid is a highly nonideal azeotropic mixture that could be separated by simple distillation. However, for dilute feed compositions solvent extraction is more advantageous because of the very high energy demand of a water-rich feed distillation. The triangular diagram of Figure 48 shows that the binary system of solvent MTBE and water has a heterogeneous azeotrope. As a consequence, extraction has to be combined with azeotropic distillation as shown in the flow sheet of Figure 48 B. By extraction in the extractor E extraction feed F is changed to raffinate R and the solvent L1 to extract phase F1, which on the other hand is the feed of distillation column C-1. In C-1 the extract is split into the bottoms acetic acid B1 and an acetic acid-free distillate D1 connected by the main balance line of distillation in the triangle diagram. The acetic acid-free raffinate R is fed into column C-2 where it is separated into the bottoms, pure water B2 and water – MTBE azeotrope D2. Both distillates D1 and D2 lie in the immiscible region between MTBE and water. This makes simple decantation (S) of the condensed phases possible. The solvent-rich fraction L1 is not further purified since MTBE as solvent needs not to be pure. So L1 can be recycled directly to the extractor E, a part of it is used as reflux for C-1. In column C-2 a mixture of water and MTBE has been separated into pure water and MTBE – water azeotrope. Therefore, L2, the water-rich phase from the settler, can also be fed into column C-2.

The economy of the hybrid process depends on a careful optimization of operational and constructional parameters, besides the careful selection of the solvent. Recently, mixed integer nonlinear programming (MINLP) has been successfully introduced in optimization of such processes [93], [94]. MINLP can be used as a process synthesis tool. It is necessary to generate a so-called superstructure of the hybrid process which contains a variety of alternative flow sheets. Then MINLP is able to choose the best solution for given parameters and also from competitive unit operations like extractive distillation, processes with more than one solvent or processes with reextraction for solvent recovery.

Extractive Distillation. Another possibility to separate azeotropic mixtures is extractive distillation. A solvent with a much higher boiling point than the feed is fed at the top of a distillation column to entrain one component of the distillation feed that enters the column at the bottom (→ Distillation and Rectification, Chap. 6.5.). The combined system of liquid – liquid extraction and extractive distillation is applied in nearly all processes used to manufacture

very pure light aromatics. *N*-Formylmorpholine is a suitable solvent to separate the light aromatics benzene and toluene used as feedstock for the petrochemical industry [95–100]. Due to the selectivity of *N*-formylmorpholine (NFM), extremely high yields of very pure products are obtained.

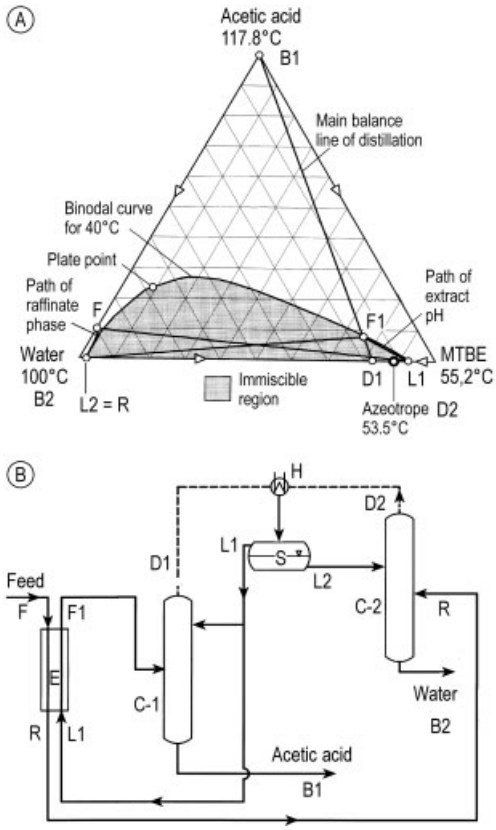

Figure 48. Triangular diagram (A) and flow sheet (B) of a hybrid extraction – distillation process for separation of water and acetic acid with MTBE as solvent [93]
E = extraction; C-1, C-2 = distillation; S = settling; H = condenser

A simplified process flow scheme for *N*-formylmorpholine extraction is given in Figure 49. Feed enters the extractor at an appropriate point in the bottom half of the column. The exact location depends on the composition of the feed and can be optimized for each feedstock. Solvent enters at the top of the extractor at the selected temperature. Because NFM has a high solvent power, this must often be reduced by addition of water. In general, 4 – 10 % water is sufficient to ensure a low solubility of saturated hydrocarbons while still maintaining a high solubility of aromatics. The raffinate phase leaving the top of the extractor is saturated with solvent. Washing with water has proved to be a simple and effective means of recovering dissolved solvent. The water – solvent mixture from this washing operation is charged to the bottom of the main extractor to increase the water content of the solvent phase as discussed above.

The solvent – aromatics mixture from the bottom of the extractor still contains some alkanes. This stream is routed to an extractive distillation column. In the lower part of this column, saturated hydrocarbons are stripped off, which results in high-purity aromatics at the bottom. Because the volatility of alkanes in the presence of NFM is much higher than that of aromatics, the need for sharp separation is not very high. At the top of the extractive distillation column, some solvent is added. This solvent stream serves to wash back any aromatics evaporated at the feed point. The resulting top product, therefore, has a very low aromatic content and is very effective as a reflux in the bottom part of the extractor, washing out the heavier alkanes.

The bottom product of the extractive distillation column, consisting of solvent and very pure aromatics, is fed to a distillation column where the aromatics are evaporated along with water. The top product of this column is condensed and fed to a settler. The extracted aromatics are available as product or can be routed to a distillation train to separate the individual compounds. Some of the aromatics are returned to the top of the distillation column as reflux. In this way, any ascending solvent vapor is condensed and the top products are solvent-free. The water layer from the settler is used for raffinate washing. Sometimes, part of the water is used as reflux instead of extract. In this combination, ultimate use is made of the solvent characteristics in both the extraction and the extractive distillation steps.

5.3. Reactive Extraction

5.3.1. Introduction

If the solvent reacts with the extracted substance during the extraction, the whole process is called reactive extraction. This reaction is normally

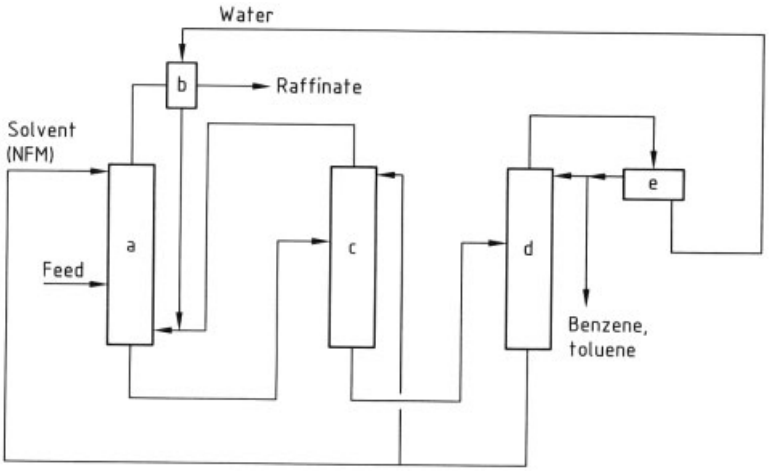

Figure 49. Morpholine solvent process
a) Extractor; b) Washing; c) Extractive distillation column; d) Distillation column; e) Settler

used to alter the properties of inorganic cations that are to be extracted from an aqueous solution to make them soluble in the nonpolar organic solvent. The reaction mechanism can best be compared with a *cation exchange* process. The solvent acts like a liquid ion exchanging the ion to be extracted for protons.

Table 1. Commercial chelating extraction reagents

Name	Producer	Chemical composition
LIX 63	Henkel	5,8-diethyl-7-hydroxy-6-dodecanone oxime
LIX 64	Henkel	2-hydroxy-5-dodecylbenzophenone oxime
LIX 64N	Henkel	LIX 65N with added LIX 63 (promoter)
LIX 65N	Henkel	2-hydroxy-5-nonylbenzophenone oxime
LIX 70	Henkel	chlorinated LIX 65 N with added LIX 63 (promoter)
LIX 71	Henkel	mixture of LIX 64 N and LIX 70
LIX 73	Henkel	mixture of LIX 64 N and LIX 70
LIX 34	Henkel	8-(*p*-dodecylphenyl sulfonamido)quinoline
LIX 54	Henkel	acetyl-*p*-dodecyl acetophenone
LIX 622	Henkel	
LIX 6022	Henkel	
SME 529	Shell	2-hydroxy-5-nonylacetophenone oxime
SME 530	Shell	SME 529 with kinetic modifier
P 17	Acorga	phenyl benzyl ketone oxime
P 50	Acorga	substituted salicyl aldoxime
P 5100	Acorga	one part P 50 with equal part nonylphenol
P 5200	Acorga	one part P 50 with two parts nonylphenol
P 5300	Acorga	one part P 50 with three parts nonylphenol
P 5050	Acorga	two parts P 50 with one part tridecanol
Kelex 100	Sherex	substituted 8-hydroxyquinoline with 10 % isodecanol
Kelex 120	Sherex	similar to Kelex 100 but with 20 % isodecanol

The use of this process for extraction of anions is far less known but is of the same importance. The mechanism of this type of reaction proceeds most of the time through ion pair formation and can best be compared with an *anion exchange process*.

A third possible mechanism is the *solvating and/or chelating reaction*. In a reaction of this type, one or more molecules of water in the hydrate layer of the ions are replaced by the active solvent. The new complex will be insoluble in water but soluble in the solvent.

Generally, the formation of a complex is a slow reaction. Therefore, in many cases, it is not possible to use columns for this type of extraction. In many applications mixer – settler are used to make the extraction possible.

A reactive extraction will need a re-extraction to isolate the extracted substance from the solvent. The active solvents used in reactive extractions are water insoluble compounds. They are used undiluted or in solution in an inert solvent.

5.3.2. Extraction Mechanism of Different Types of Solvent

Classical Complexing and/or Chelating Agents. The active solvents are usually large, polyfunctional molecules. Most of them are designed to extract specific cations. Dioximes, hydroxyoximes, quinolines, β-diketones, etc., are rather expensive but widely used solvents to extract metal ions such as copper and nickel from the aqueous phase.

Metal complexes are formed in the water phase in a substance-specific pH range and extracted into an inert solvent. After extraction the metal ion is liberated from the complex by changing the pH, and re-extracted into the water phase. The complexing/chelating agents are designed such that they are chemically stable, nontoxic, and insoluble in water to minimize losses. The liquid ion exchange (LIX) [5] agents used in the production of copper are typical for this category.

Henkel, Hoechst, Shell, Sherex, and Acorga are the largest producers of complexing/chelating extractants (see Table 1).

Phosphoric, Sulfonic, and Carboxylic Acids. This class of products acts typically as cation exchangers in solution. The products have one or more protons that can be exchanged for metal ions.

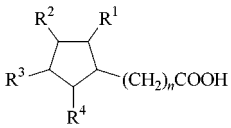

Naphthenic acids

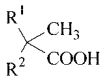

Versatic acids

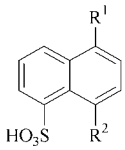

Sulfonic acids

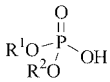

Phosphoric acids

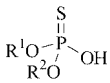

Thiophosphoric acids

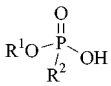

Phosphonic acids

The metal salts of these acids are soluble in most nonpolar solvents and, depending on the pH of the extraction and the pK_a of the acid, the cation will migrate to the organic phase or stay in the water phase. The dependence of the distributions coefficients on pH is shown in Figure 50.

The different curves give the variation of the distribution coefficient D in dependence of pH for different cations. These solvents are not specific for certain cations, but are very suitable to concentrate and recuperate trace elements from effluent streams [109], [110].

The major producers of this type of compounds are Bayer, Hoechst, Daihachi, Shell, and Albright & Wilson.

Ethers and Thioethers.

$$R_2O + HX \rightleftharpoons R_2OH^+ + X^-$$

$$R_2S + HX \rightleftharpoons R_2SH^+ + X^-$$

where HX stands for a strong acid, e.g., H_2SO_4 or HCl, and X is the anionic rest of the acid. Most metal ions that are capable of forming anionic complexes are extractable with ethers and thioethers. Fe, Au, In, Ga, etc. can be extracted with these solvents from an aqueous HCl solution (HCl concentration > 6 M) and re-extracted from the organic phase with water. Most of these anionic complexes form strong acids which protonate with the ether group to form oxonium (thionium) salts. These salts, however, are soluble in ether or thioether. It is also possible to extract Cl^- from water with $(R_2O)_2 \cdot H_2SO_4$, the sulfate ion being replaced by the chloride. This type of solvent is normally used in undiluted form [106].

Ketones extract the same anionic complexes as ethers, but in most cases at lower acid concentration. Complexes with weaker acids can also be extracted. Methyl isobutyl ketone and cyclohexanone are industrially used in the separation of Nb – Ta as F^- complexes [112] (→ Niobium and Niobium Compounds, Chap. 4.3.1.) and in the separation of Zr – Hf as CNS^- complexes (→ Zirconium and Zirconium Compounds, Chap. 1.4.2.) [116]. The solvents are used in undiluted form. A great disadvantage of this and of the former type of solvents is that they are not completely insoluble in water. A distillation is needed to remove the dissolved solvent from the water phase.

Trialkylphosphine oxides, Trialkylphosphinates, -phosphonates, and -phosphates.

$$R_3PO + HX \rightleftharpoons R_3POH^+ + X^-$$

$$ROR_2PO + HX \rightleftharpoons ROR_2POH^+ + X^-$$

$$(RO)_2RPO + HX \rightleftharpoons (RO)_2RPOH^+ + X^-$$

$$(RO)_3PO + HX \rightleftharpoons (RO)_3POH_+ + X^-$$

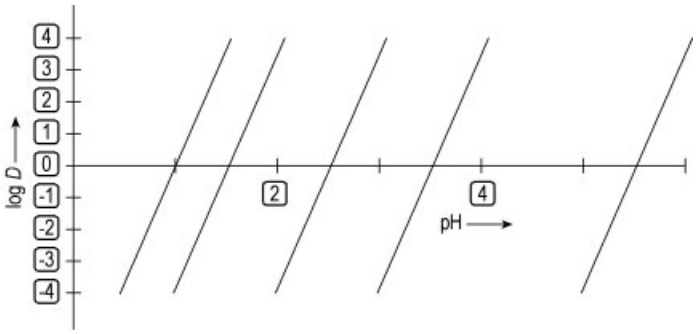

Figure 50. Plots of log D vs. pH for extraction of metals

With this type of extractants most metals from group 3 to group 8 can be extracted, even if only weak complexes are formed with the solvent. It is for example possible to extract $TiOSO_4$ [114] from a sulfuric acid solution with dibutyl butylphosphonate (DBBP) as the solvated complex $2\,DBBP \cdot TiOSO_4 \cdot H_2SO_4 \cdot 5\,H_2O$. These organophosphorus solvents are, due to their outstanding chemical stability and good physical properties, most widely used in reactive extraction [107], [108].

Free acids can also be extracted, the basicity of the solvents decreases from phosphine oxide to trialkylphosphate. The distribution coefficients are rather small so that large volumes of solvent are needed in an extraction process [104], [105], [111]. These extractants are used as such or diluted in an inert solvent (petroleum fraction) with or without modifier. The distribution coefficients in metal extraction are usually very high and depend on the acid concentration. The products are rather expensive.

Bayer, Hoechst, Daihachi, Shell, and Albright & Wilson are the major producers.

Trialkylsulfonium (Selenium, Tellurium), Tetraalkylammonium, and Tetraalkylphosphonium (Arsonium, Stibonium) Salts.

$$[R_4N^+A^-]_{Solvent} + [B^-]_{Raffinate} \rightleftharpoons [R_4N^+B^-]_{Solvent} + [A^-]_{Raffinate}$$

These onium compounds form ions both in acidic and basic medium. The mechanism of extraction is practically always anion exchange [106]. The onium compounds are regenerated in the same way as anion exchanger resins by re-extraction with a base. These solvents are used in the production of Cr, V, Re, Tc, Mo, and W as a solution in a petroleum fraction.

Producers are Henkel, Sherex, Hoechst, and BASF.

Amine and Phosphine Salts.

$$R_3N + HX \longrightarrow R_3NH^+ + X^-$$

$$R_3P + HX \longrightarrow R_3PH^+ + X^-$$

Primary, secondary, and tertiary amines must be protonated, thus converted to the ionic form before they can be used as an ion exchanger. Thus acids can be extracted with free amines, metals have to be transformed to the corresponding acids: vanadium can be extracted as HVO_3, chromium as $H_2Cr_2O_7$, and manganese as $HMnO_4$. Most used are the tertiary amines because they are chemically inert to most reagents.

In the system trilaurylamine – citric acid a reversible extraction can be set up, due to the difference in distribution coefficient at different temperatures. At low temperature (20 °C), the distribution coefficient solvent/raffinate is > 10. At high temperature (130 °C), the coefficient for the same system is < 1. Thus, the acid can be extracted from water at low temperature and re-extracted with water at high temperature. Much less accounted for is the fact that these amines have very different distribution coefficients for different anions.

$$[R_3NH^+A^-]_{Solvent} + [B^-]_{Raffinate} \rightleftharpoons [R_3NH^+B^-]_{Solvent} + [A^-]_{Raffinate}$$

For trioctylamine in an aromatic solvent the order of extractability is

$$I^- \gg Br^- > NO_3^- > Cl^- \gg F^- > S_2O_3^{2-} > PO_4^{3-} > SO_4^{2-}$$

This series is dependent on the type of amine, the diluent, and the temperature and must be determined for any given system [5]. This series can be enlarged for organic compounds, capable of forming ion pairs with the amine. Phenols and substituted phenols can be extracted with better distribution coefficients than chlorides. In this series thiophenols and some mercaptans are situated between bromide and nitrate and all sulfonic acids are better extracted than phosphate. This makes the sulfate salts of amines a very good choice for new extraction systems.

$(R_3N)_2 \cdot H_2SO_4$ (R = octyl) is capable of extracting Cl^- and/or NO_3^- [113] from a mixture of both anions with SO_4^{2-}. The value of the distribution coefficient of the anions is surprisingly high.

$$D_{Cl}/D_{SO_4} \approx 100$$

$$D_{NO_3}/D_{SO_4} \approx 400$$

A disadvantage of these solvents is the fact that per mole amine one mole of salt is formed after regeneration. However, most of the time, the anions to be removed are present in the solution only as trace impurities, so that this drawback is not significant. Another disadvantage of amines is the poor solubility of ammonium salts in aliphatic hydrocarbons. To avoid third phase formation, the use of aromatic solvents (smell) or aliphatics with modifiers is necessary. Amines can also act as solvating agents in an acid medium. Iron, zinc, and uranium salts can be extracted in this way [106].

5.3.3. Uses

1) A bleed from a $(NH_4)_2SO_4$ crystallization unit in a caprolactam plant contains 40 % $(NH_4)_2SO_4$, 10 % NO_3^-, 50 ppm Cl^- and 3 – 4 % organic sulfonic acids. The solution is extracted by trioctylammonium sulfate, the added amount is stoichiometric to $[NO_3^-] + [Cl^-] +$ [sulfonic acid]. In three stages > 99 % of the nitrate, 90 % of the chloride, and 85 % of the sulfonic acids are extracted by the trioctylamine in solution (40 %) in Shellsol AB (aromatic hydrocarbon). The raffinate is returned to the crystallization plant for $(NH_4)_2SO_4$recovery. The re-extract contains 35 % of nitrate, 0.016 % of chloride, the sulfonic acids, and only 0.1 % of sulfate and is treated separately. Compared with a conventional stripping, the use of chemicals in this system is strongly reduced. 80 % less NaOH is consumed because the base is only used to liberate the amine and not to neutralize the NH_3 from $(NH_4)_2SO_4$ [113].
2) Traces of phenol (20 – 200 ppm) and bisphenol (20 – 100 ppm) can be extracted from the wastewater of a polycarbonate plant to values < 1 ppm, thus avoiding biological treatment of the water. The re-extract containing sodium phenolate and bisphenolate (300 – 500 times more concentrated) can be returned to the polycarbonate plant. The solvent used is 5 % trioctylamine in Shellsol AB [117].
3) From a plant for the production of vulcanization accelerators 2-mercaptobenzothiazol (100 – 300 ppm) and its oxidation products are completely extracted in two stages and re-extracted as a 20 – 30 % solution. This concentrated waste can be easily removed. The solvent is 20 % trioctylamine in Shellsol AB [113].
4) Bromide ions can be extracted selectively from sea water (concentration in the Dead Sea ≈ 4 %) with amine hydrochlorides and concentrated in a re-extraction step, thus avoiding the energetically unfavorable stripping of sea water. The extraction can be carried out with various types of amines as HCl or H_2SO_4 salt [106].
5) Mg^{2+} can be removed from a saturated solution of $(NH_4)_2SO_4$ by extraction with the ammonium salt of diethylhexyl phosphonic acid (DEHPA) at pH > 5. It can be recovered by lowering the pH to < 3 [115].
6) Traces of uranium are extracted from phosphate ores. The uranium content in these ores is only 30 – 300 ppm. The reserves of those ores are so huge that the quantity of uranium from phosphates many times exceeds that of uranium found in other uranium ores.
7) Lanthanides and actinides have only recently become available in higher quantity and high purity. This was enabled by the development of the extraction technology (→ Rare Earth Elements).

8) Worldwide about 40 % of the copper is produced by reactive extraction. The plants are an order of magnitude larger then any other type of solvent extraction presently in operation (→ Copper, Chap. 5.8.). With an aqueous feed of 200 – 3000 m^3/h a whole new technology had to be developed. Copper can be extracted after acid or ammoniacal leaching. The aqueous stripping phase is used directly in the electrowinning of copper. To avoid the build-up of impurities in the electrolysis of copper, Fe, Zn, and other metals are removed from this phase by extraction with other solvents [5], [102], [111].
9) More than 90 % of the world production of cobalt is performed by reactive extraction. Separation of Co and Ni can be achieved from alkaline, neutral, or from acidic media. Most of the impurities (Sn, Bi, Fe, Pb, Cu, Cd, and Zn) have to be removed from the solution by precipitation or extraction. From a medium concentrated aqueous chloride solution (100 – 200 g/L), Co can be extracted by amines together with only traces of Ni. After scrubbing with 4 – 5 M HCl, most of the Ni is removed from the solvent. The cobalt is stripped by water.
   From neutral or weakly acidic media (pH 4.8 – 5.5) Co can be extracted by the Na^+ or NH_4^+ salt of di-2-ethylhexylphosphoric acid (D2EHPA) (to avoid acidification of the aqueous solution). Some Ni is co-extracted and is removed by scrubbing with a Co solution thus replacing any extracted Ni by Co. Co is finally stripped by HCl, H_2SO_4, or HNO_3 depending on the use.
   The use of monooctyl octylphosphonic acid (MOOP) to replace D2EHPA [101] permits the same separation, but at lower pH and with enhanced distribution constants for Co.
   In alkaline solution Co and Ni can be extracted by LIX agents from ammoniacal solution. Prior to the extraction, cobalt is oxidized to the Co^{3+} state. Nickel is stripped first with dilute acid, Co remaining in the solvent. Cobalt is then reduced to the Co^{2+} state and stripped with diluted acid [5], [102], [104], [105].

All these examples are used in industry or in pilot plants. Elaborate process and flow control are essential to use this technology in a economical and meaningful way.

5.3.4. Setting up an Extraction System

A solution R contains three components: acid, metal A and metal B. Both metals and the acid are extractable with a solvent S and metal A is the desired component that is to be obtained in acid solution. The procedure to set up a new extraction system with these data is typical for any reactive extraction system [114], [118].

The distribution coefficients are to be measured at different acid concentrations at a temperature suitable for the extraction. A typical measurement curve is shown in Figure 51.

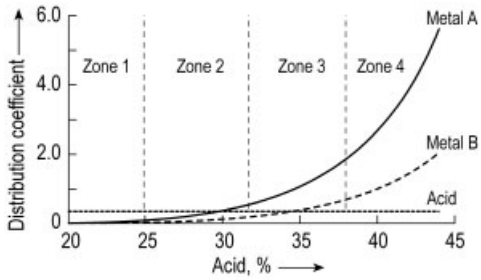

Figure 51. Distribution coefficient in dependence on acid concentration

The diagram shows a poor, practically constant distribution coefficient *D* for the acid and an increasing one for both metals A and B at increasing acid concentration.

This means that the acid extraction is in fact a normal physical extraction and that both metals form solvation complexes. At any acid concentration, all three components are extracted together. The extraction coefficients *E* are related in the same way as are the distribution coefficients,

$$E = D \cdot S / R.$$

To extract metal A completely and in a reasonable number of steps extraction has to be carried out in zone 4 of Figure 51, the *extraction zone*. At the same time, metal B and the acid are only partially extracted into the solvent. The number of steps can be calculated with the Kremser formula. The Kremser formula calculates the efficiency *Z* in the extraction

$$Z = \left(E^{n+1} - E\right) / \left(E^{n+1} - 1\right)$$

over n stages. For this equation the number of stages can easily be calculated

$$n = \log \{[1/(1-Z)\cdot(1-1/E)] + 1/E\} / \log E$$

To remove metal B from the solvent the solvent has to be washed with water or acid in such a way that most of metal A stays in the solvent, while metal B is washed out as completely as possible. This is done in zone 3 of Figure 51, the *scrubbing zone.* The amount of water to be used is given again by the distribution coefficients of both acid and metal B. If too much acid is removed, metal A will also be washed out. Therefore it is more adequate to use an acid solution. The concentration of the dilute acid should be such that the acid solution can only take up the amount of acid to be removed from the solvent. By scrubbing with pure water only a limited amount of scrubbing fluid can be used because of the low distribution coefficient of the acid. By adding a calculated amount of acid the amount of scrubbing fluid can be increased and more component B can be removed without removing component A. The raffinate, leaving the scrubbing zone contains too much of metal A to be discarded and is sent back to one of the steps in the extraction. The number of steps can again be calculated with the Kremser formula.

In zone 2 of Figure 51, the *stripping zone*, metal A is washed out of the solvent with water or a dilute acid solution. The amount of water is chosen such as to obtain the desired concentration of metal A in the re-extract. With the given distribution coefficient, it is not possible to remove metal A completely form the solvent. At any acid concentration, traces of metal A will remain in the solvent.

In zone 1 of Figure 51, the *washing zone*, the acid and the rest of metal A are washed out of the solvent. After this step, the solvent is completely regenerated and can be reused.

General Flow Sheet. Figure 52 shows a general flow sheet of a reactive extraction system.

Solvent Make-Up. Due to the poor distribution coefficient of the acid several extraction steps are needed to load the solvent with sufficient acid to make extraction of the metals possible. In the given example the acid comes from the initial solution. To avoid losses of metal A, the solvent must be completely free of A. At the low acid concentration in the first steps of the solvent make-up any metal A present in the solvent will be re-extracted in the raffinate.

Reactive Extraction. If sufficient acid is present in the solvent, both metals will be extracted from the feed solution. To limit the number of steps in the extraction, concentrated acid is added. The high concentration is necessary for the metal extraction and in the solvent make-up.

Solvent Scrubbing. To make sure that most of metal B is washed out from the solvent, the acid solution containing metal A and acid from the solvent washing step is used, so that we stay in the extraction zone for metal A. The second component, with a lower distribution coefficient, is already in its stripping zone and will be washed out of the solvent. To avoid losses, this stream is sent back to one of the steps in the extraction.

Solvent Stripping. After scrubbing, the solvent contains only metal A and acid. Using water or a dilute acid solution, the purified metal A can now be re-extracted into a water phase.

Solvent Washing. The solvent must be completely regenerated before it can be recycled to the solvent make-up. This is done by adding some supplementary washing steps to the system. The amount of water is chosen according to the requirements in the scrubber and the number of steps is calculated as a function of that amount.

5.3.5. Diluents and Modifiers

Diluents are used to reduce the solvent viscosity, improve phase separation, and most of all, when properly chosen, to reduce the loss of solvent. The solubility of tri-*n*-butyl phosphate (TBP, one of the most frequently used extractants) in pure water is 420 mg/L at room temperature; a 50 % dilution of TBP in *n*-hexane lowers this solubility to 175 mg/L. The choice of diluent is very important. Higher amines are completely soluble in aliphatic diluents. Their salts, however, are only slightly soluble in the same solvents. Therefore it is essential to test the solutions in all steps of the process. If third-phase formation occurs in one of the steps, the diluent must be changed or a *modifier* is to be added. These modifiers are normally long chain alcohols, monoesters of polyhydric alcohols, or phosphate

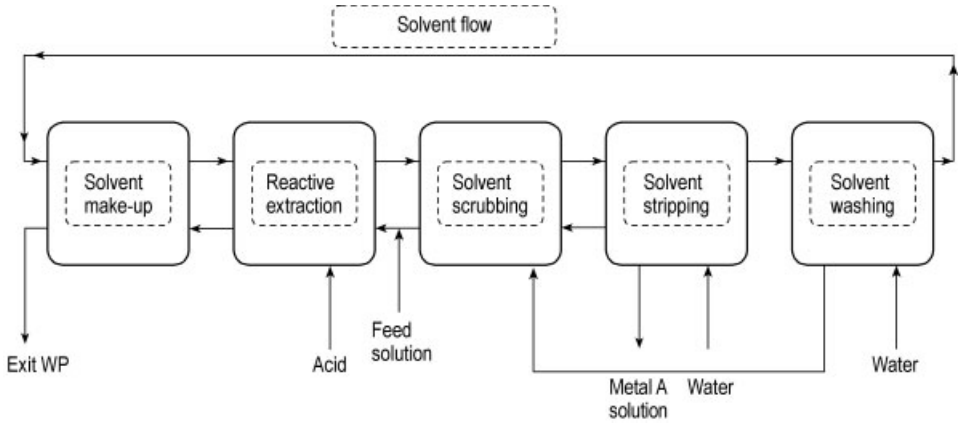

Figure 52. General flow sheet of a reactive extraction system

esters. These modifiers may considerably influence the kinetics of the extraction. Especially in the extraction of copper the choice of modifier is very important. The different types of solvent are sold mixed with a suitable modifier which greatly enhances the extraction performance of the solvent.

6. References

1. R. E. Treybal: *Liquid Extraction,* McGraw-Hill, New York – Toronto – London 1963.
2. S. Hartland: *Counter Current Extraction,* Pergamon Press, Oxford 1970.
3. J. Rydberg, C. Musikas, G. R. Choppin (eds.): *Principles and Practises of Solvent Extraction,* M. Dekker, Inc., New York 1992.
4. C. Hanson: *Recent Advances in Liquid – Liquid Extraction,* Pergamon Press, Oxford 1971.
5. T. C. Lo, M. H. I. Baird, C. Hanson (eds.): *Handbook of Solvent Extraction,* Wiley-Interscience, New York 1983.
6. J. D. Godfrey, M. J. Slater (eds.): *Liquid – Liquid Extraction Equipment,* J. Wiley & Sons, New York 1994.
7. J. Wisniak, A. Tamir: *Liquid – Liquid Equilibrium and Extraction. A Literature source Book.* Physical Sciences Data, 7, Part A 1980, Part B 1981, Physical Sciences Data, 23, Supplement 1 1985, Elsevier, Amsterdam.
8. Proceedings ISEC '88, Vernadsky Institute of Geochemistry and Analytical Chemistry of the USSR Academy of Sciences, 4 vols., NAUKA, Moscow, 1988.
9. Proceedings ISEC '90, Chemical Society of Japan, Elsevier, New York 1992.
10. Proceedings ISEC 71, The Hague, 2 vols., Society of Chemical Industry, London 1971.
11. Proceedings ISEC 74, Lyon, 3 vols., Society of Chemical Industry, London 1974.
12. Proceedings ISEC 77, Toronto, 2 vols., CIMM Special Publication, no. 21, Can. Inst. Min. Met., 1979.
13. Proceedings ISEC 80, Liège, 3 vols., Assocn. des Ingenieurs sortis de l'Université de Liège, Liège 1980.
14. Proceedings ISEC 83, Denver AIChE 1983.
15. Proceedings ISEC 86, München, 3 vols., DECHEMA, Frankfurt 1986.
16. Proceedings ISEC '93, Solvent Extraction in the Process Industries, 2 vols., Elsevier, New York 1993.
17. Proceedings ISEC '96, Value Adding Through Solvent Extraction, 2 vols., The University of Melbourne, Parkville, Victoria, Australia 1996.
18. J. M. Sorensen, W. Arlt: "Liquid – Liquid Equilibrium Data Collection," vol. V, part 1 – 3, DECHEMA Chemistry Data Ser., Frankfurt 1979.
19. Landolt-Börnstein: **II 2 c,** Springer Verlag, Berlin 1980.
20. O. Madelung: *Landolt-Börnstein directory disk 1.1,* Springer-Verlag, Berlin 1994.
21. A. Seidel, F. W. Linke: "Solubilities of Inorganic and Organic Compounds," in A. W. Francis (ed.): *Ternary Systems Separating into Two Liquid Layers,* van Norstrand, London 1952.
22. A. W. Francis: *Liquid – Liquid Equilibria,* Wiley, New York 1963.

23. H. Stephen, T. Stephen: *Solubilities of Inorganic and Organic Compounds,* **4 vols.,** Pergamon Press, Oxford 1964.
24. F. Ruiz, D. Prats, A. F. Marcilla: "Liquid – Liquid Extraction: a Graphical Method for Equilibrium Stage Calculations for Quaternary Systems," *Fluid Phase Equilib.* **15** (1984) no. 3, 257 – 265.
25. W. Rückl, R. Marr: "Multicomponent Liquid – Liquid Extraction," *Ger. Chem. Eng. (Engl. Transl.)* **8** (1985) no. 1, 27 – 31.
26. M. J. Hampe: "Selection of Solvents in Liquid – Liquid Extraction According to Physico-Chemical Aspects," *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) no. 4, 251 – 263.
27. *Ullmann,* 4th., **vol. 2,** 550 – 551.
28. E. Bender, U. Block, "Thermodynamische Berechnung der Flüssig-Flüssig-Extraktion," *Verfahrenstechnik* (Mainz) 9 (1975) no. 3, 106 – 111.
29. K. Hlavaty: "Correlation of the Binodal Curve in a Ternary Liquid Mixture with One Pair of Immiscible Liquids," *Collect. Czch. Chem. Commun.* **37** (1972) 4005 – 4007.
30. E. Hecker: "Auswahl von Lösungsmittelsystemen zur multiplikativen Verteilung," *Chimia* **8** (1954) 229 – 236.
31. C. M. Hanson: "The Universality of Solubility Parameter," *Ind. Eng. Chem. Prod. Res. Dev.* **8** (1969) no. 1, 2 – 11.
32. C. Reichardt, K. Dimroth: "Lösungsmittel und empirische Parameter zur Charakterisierung ihrer Polarität," *Fortschr. Chem. Forsch.* **11** (1968) no. 1, 1 – 73.
33. R. Anderson, R. Cambio, J. M. Prausnitz: "Physical and Chemical Forces in Solvent Selectivity for Hydrocarbons," *AIChE J.* **8** (1962) 66 – 69.
34. C. H. Deal, E. L. Derr: "Selectivity and Solvency in Aromatics Recovery," *Ind. Eng. Chem. Prod. Res. Dev.* **3** (1964) 394 – 399.
35. R. A. Grieger, C. A. Eckert: "Solvent Mixtures for Separation Processes," *Ind. Eng. Chem. Prod. Res. Dev.* **6** (1967) 250 – 255.
36. E. Müller, G. Höhfeld: "Aromatenextraktion mit Lösungsmittelgemischen," *Erdöl Kohle Erdgas Petrochem.* **24** (1971) 573 – 578.
37. D. Tassios: "GLC Screens Extraction Solvents," *Hydrocarbon Process* **49** (1970) no. 7, 114 – 118.
38. I. Mellan: *Industrial Solvents Handbook,* Noyes Data Corp., New Jersey 1985.
39. H. Gnamm, W. Sommer: *Lösungsmittel und Weichmachungsmittel,* Wissenschaftliche Verlags-anstalt, Stuttgart 1980.
40. G. Dave, M. Lidman: "Biological and Toxicological Effects of Solvent Extraction Chemicals," *Hydrometallurgy* **3** (1978) no. 3, 201 – 216.
41. R. Berger, W. Leuckel, D. Wolf, *Chem. Ing. Tech.* **50** (1978) no. 7, 544 – 545.
42. R. Marr, G. Husung, F. Moser, *Chem. Ing. Tech.* **47** (1975) no. 5, 203.
43. G. A. Rowden, M. Dilley, C. F. Bonney, G. A. Gillet, Lecture at Hydrometallurgy 81, Manchester 1981.
44. T. Misek, R. Berger, J. Schröter: *Standard Test Systems for Liquid Extraction,* The Institution of Chemical Engineers 1985.
45. D. Niebuhr, Ph. D. Thesis, Technische Universität Clausthal (Fed. Rep. of Germany) 1982.
46. E. Aufderheide, Ph. D. Thesis, Technische Universität Clausthal (Fed. Rep. of Germany) 1985.
47. W. Pietzsch, Ph. D. Thesis, Technische Universität München (Fed. Rep. Germany) 1984.
48. H. Haverland, Ph. D. Thesis, Technische Universität Clausthal (Fed. Rep. of Germany) 1988.
49. K. H. Reissinger, Ph. D. Thesis, Technische Universität Graz (Austria) 1985.
50. R. Berger, K. Walter, *Chem. Eng. Sci.* **40** (1985) no. 12, 2175 – 2184.
51. H. Zimmermann, Ph. D. Thesis, Technische Universität München (Fed. Rep. of Germany) 1981.
52. Th. Pilhofer, *Chem. Ing. Tech.* **48** (1976) no. 3, 237.
53. E. Bender, R. Berger, W. Leuckel, D. Wolf, *Chem. Ing. Tech.* **51** (1979) no. 3, 192 – 199.
54. R. Berger, Lecture GVC annual meeting, Düsseldorf 1981.
55. J. Komasawa, J. Ingham, *Chem. Eng. Sci.* **33** (1978) no. 3, 341 – 347, no. 4, 479 – 485.
56. Th. Pilhofer, *Chem. Ing. Tech.* **51** (1979) no. 3, 231.
57. K. Hirschmann, E. Blass, *Ger. Chem. Eng.* **7** (1984) 280 – 287.
58. G. Wagner, Ph. D. Thesis, Technische Universität München, 1995.
59. B. Hoting, A. Vogelpohl, *Chem. Ing. Tech.* **68** (1981) 105 – 112.
60. H. W. Brandt, K. H. Reissinger, J. Schröter, *Verfahrenstechnik* **9** (1975) no. 8, 383 – 387.
61. G. Dichtl, T. Pilhofer, *Verf. Tech.* **15** (1981) 615 – 617.
62. H. Rauscher, Ph. D. Thesis, Technische Universität München, 1992.

63. M. Lorenz, Ph. D. Thesis, Technische Universität Clausthal, 1990.
64. G. Goldmann, Ph. D. Thesis, Technische Universität München, 1986.
65. H. Hufnagl, Ph. D. Thesis, Technische Universität München, 1992.
66. Th. Pilhofer, D. Mewes: *Siebboden-Extraktionskolonnen,* Verlag Chemie, Weinheim – New York 1979.
67. Th. Pilhofer, *Ger. Chem. Eng. (Engl. Transl.)* **2** (1979) 69 – 76.
68. G. Zamponi, Ph. D. Thesis, Technische Universität München, 1996.
69. K. Hirschmann, Ph. D. Thesis, Technische Universität München, 1984.
70. K. H. Reissinger R. Marr, *Chem. Ing. Tech.* **58** (1986), 7, 540 – 547.
71. J. D. Thornton, *Brit. Chem. Eng.* **32** (1958) 247 – 251.
72. A. E. Karr, T. C. Lo, *Proc. ISEC 71.*
73. R. Bauer, Ph. D. Thesis, ETH Zürich (Switzerland) 1976.
74. Th. Pilhofer, J. Schröter, *Chem. Ing. Tech.* **56** (1984) 883 – 890.
75. W. Bäcker, J.-P. Schäfer, J. Schröter, *Chem. Ing. Tech.* **63** (1991) 1008 – 1011.
76. E. Blass, W. Goettert, M. J. Hampe: "Selection of Extractors and Solvents," in [6, Chap. 18].
77. I. Wagner, Ph. D. Thesis, Technische Universität München, 1999.
78. R. Berger, *Chem. Ing. Tech.* **58** (1986) no. 6, 499 – 456.
79. M. Henschke, Ph. D. Thesis, RWTH Aachen, 1995.
80. S. Hartland, S.A.K. Jeelani in [6, Chap. 13].
81. Computer Aided Settler Design "CASEDY" Rauschert Verfahrenstechnik GmbH, Postfach 20, D-96347 Steinwiesen.
82. A. Kriechbaumer, R. Marr, *Chem. Ing. Tech.* **55** (1983) 700 – 707.
83. W. Rommel, W. Meon, E. Blass, *Separat. Sci. Technol.* **27** (1991) 129 – 159.
84. W. Rommel, E. Blass, W. Meon, *Chem. Engng. Sci.* **47** (1992) 555 – 564.
85. W. Meon, W. Rommel, E. Blass, *Chem. Engng. Sci.* **48** (1993) 159 – 168.
86. W. Rommel, E. Blass, W. Meon, *Chem. Engng. Sci.* **48** (1993) 1735 – 1743.
87. F. Rebelein, Ph. D. Thesis, Technische Universität München, 1989.
88. F. Rebelein, E. Blass, *Filtration & Separation* **27** (1990) 360 – 363.
89. R. Magiera, Ph. D. Thesis, Technische Universität München, 1995.
90. R. Magiera, E. Blass, *Filtration & Separation* **34** (1997) 369 – 376.
91. S. Hoffmann, Ph. D. Thesis, Technische Universität München, 1999.
92. Abwassertechnischer Verein e.V. (ed.): *Industrieabwässer,* Verlag Ernst & Sohn, Berlin 1999.
93. S. Glanz, Ph. D. Thesis, Technische Universität München, 1998.
94. S. Glanz, J. Stichlmair in R. Darton (ed.): "Distillation and Absorption '97," *Ichemt, Symp. Ser. No. 142,* 1997.
95. F. Trefny, *Erdöl Kohle Erdgas Petrochem.* **23** (1970) 337 – 340.
96. E. Cinelli et al., *Hydrocarbon Process.* **51** (1972) no. 4, 337 – 340.
97. *Hydrocarbon Process.* **51** (1972) no. 4, 141 – 144.
98. G. Precesser et al., *Oil Gas J.* July 16 (1973) 114 – 118.
99. Krupp Koppers, *Hydrocarbon Process.* **52** (1973) no. 4, 139 – 141.
100. *Hydrocarbon Process.* (1982) 182.
101. G. M. Ritcey, A. W. Ashbrook: *Solvent Extraction,* Part 1, Elsevier, Amsterdam 1984.
102. G. M. Ritcey, A. W. Ashbrook: *Solvent Extraction,* Part 2, Elsevier, Amsterdam 1979.
103. *Solvent Extraction in Metallurgical Processes*, Internat. Symp. Antwerpen 1972.
104. W. W. Schulz, J. D. Navratil: *Science and Technology of Tributyl Phosphate,* **vol. I,** CRC Press, Boca Raton, FL 1984.
105. W. W. Schulz, J. D. Navratil, A. S. Kertes: *Science and Technology of Tributyl Phosphate,* **vol. VI,** CRC Press, Boca Raton, FL 1991.
106. A. K. De, S. M. Khopar, R. A. Chalmers: *Solvent Extraction of Metals,* Van Nostrand Reinhold, London 1970.
107. Bayer AG: Geschäftsbereich Anorganische Chemikalien, Baysolvex – TBP für die Flüssig-Flüssig Extraktion 1993.
108. Bayer AG: Geschäftsbereich Anorganische Chemikalien, Baysolvex – Dialkan-phosphate für die Flüssig-Flüssig Extraktion 1993.
109. Bayer AG: Geschäftsbereich Anorganische Chemikalien, Baysolvex – D2EHPA für die Flüssig-Flüssig Extraktion 1993.
110. Bayer AG: Geschäftsbereich Anorganische Chemikallien, Baysolvex – D2EHTPA für die Flüssig-Flüssig Extraktion 1993
111. Bayer AG: Geschäftsbereich Anorganische Chemikalien, Baysolvent for your Solvent-extraction 1993.
112. F. Fairbrother: *The Chemistry of Niobium and Tantalum,* Elsevier Amstrad, London – New York 1967, pp. 8 – 10.

113. Bayer Antwerpen N.V., EP 0 396 790 A1, 1989.
114. Bayer Antwerpen N.V., DE 3 643 711 A1, 1986.
115. Bayer Antwerpen N.V., EP 638 515-A2, 1993.
116. W. D. Jamrack: *Rare Metal Extraction by Chemical Engineering Technology,* Pergamon Press, Oxford 1963.
117. M. L. Wang, B. L. Liu: "Extraction Equilibrium of Phenol by Sulphuric Acid Salts of Triisooctylamine," *Chem. Eng. Comm.* **156** (1996) 131 – 146.
118. Bayer Antwerpen N.V., DE 3 524 053 A1, 1985.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Liquid–Solid Extraction

Standard Article

Theodor Voeste[1], Klaus Weber[2], Brent Hiskey[3], Gerd Brunner[4]

[1]Lurgi GmbH, Frankfurt, Germany
[2]Consultant, Hamburg, Germany
[3]University of Arizona, Tucson, Arizona, USA
[4]Technische Universität Hamburg – Harburg, Hamburg, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b03_07.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Extraction without Chemical Reaction
2.1. General Comments
2.2. Technical Principles and Process Design
2.3. Extraction Processes and Typical Equipment
2.3.1. Batch Extraction
2.3.2. Continuous Extraction
3. Chemical Leaching in Hydrometallurgy
3.1. Methods
3.2. Principles
3.3. Typical Processes and Equipment
3.3.1. Agitation Leaching of an Oxidized Gold Ore
3.3.2. Pressure Leaching of Zinc Sulfide Concentrates
3.3.3. Bioleaching of Refractory Precious-Metal Concentrates
4. Extraction with Supercritical Fluids
4.1. Principles
4.2. Process Design
4.3. Equipment

[Top of Page]

1. Introduction

Theodor Voeste and Klaus Weber

Liquid – solid extraction, also called *leaching* or *lixiviation*, is the process of totally or partially dissolving a solid in a suitable solvent. Lixiviation is derived from the Latin word *lixivius* which means made of lye.

Liquid – solid extraction covers a variety of processes such as the extraction of oilseeds, bone, fuller's earth, and catalysts. Further examples are the production of sugar from beet and sugar cane and the extraction of bitumen from oil-impregnated rock or sand. Laundries, dry-cleaning, and the making of coffee from ground coffee beans are everyday examples of automated leaching processes.

Extraction is often accompanied by a chemical reaction which converts the extract into soluble form; examples of chemical leaching are the hydrolysis of wood, cellulose digestion, and many leaching and decomposition processes used in metallurgy and the pigment industries.

The first chapter of this article describes extraction without chemical reactions which depends solely on the solubility or miscibility of the extract in the solvent. Oilseed extraction is primarily described because its characteristics and objectives are typical for the physical processes and equipment used in other applications.

Chemical leaching in hydrometallurgy is discussed in the second chapter. The third chapter deals with supercritical-fluid extraction, also called gas extraction which is a relatively modern technology. Here extraction is achieved with a gas or gaseous mixture under near-critical conditions (pressure and temperature). Near the critical point, the differences between the liquid and gas phases disappear and extraction occurs via the resulting fluid phase. This separation process bears some resemblance to distillation and is sometimes referred to as destraction.

For detailed information on specific extraction processes, see ⟶ Coffee, ⟶ Coffee – Brewed and Instant (Soluble) Coffee; ⟶ Fats and Fatty Oils – Extraction ; ⟶ Sugar; ⟶ Tar Sands.

[Top of Page]

2. Extraction without Chemical Reaction

Theodor Voeste and Klaus Weber

2.1. General Comments

Definitions. Some of the technical terms used in this chapter are listed below together with their meanings:

Feed material	(prepared) raw material which is to be extracted
Extract	substance which is to be separated from the feed material
Solvent	medium used to carry out the extraction
Miscella	the conventional expression in the oilseed crushing industry for the extract solution
Degree of extraction	(extract in feed material – extract in extraction residue)/extract in feed material

A cost-effective extraction process must fulfill the following requirements:

1. The feed material must be prepared in such a way that the solvent can dissolve the extract quickly. This is achieved by crushing, grinding, or flaking.
2. Only the desired extract should be dissolved and leached out. This is governed by the solvent selectivity and the temperature.
3. The extract should be recovered at the highest possible concentration. Therefore, plants in which the solid and liquid pass countercurrently to one another are widely used.
4. Economic separation of the solvent from the extract solution and in many cases, from the extraction residue must be possible (e.g., by means of distillation or solvent extraction).

The complete extraction process usually includes the following additional stages: preparation of the feed material, as well as separation and recovery of the solvent from both the extract and the extraction residue (see Fig. 1).

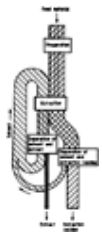

Figure 1. Schematic diagram of liquid – solid extraction

Comparison of Extraction with Pressing. Many substances can be obtained by liquid – solid extraction (e.g., quinine, coffee, tea, as well as drugs, and alkaloids), but vegetable oils and some other substances can also be produced by mechanical pressing (expelling). If the feed material has a high oil content, mechanical pressing is the more economic method. However, pressing gives a residual oil content of approx. 5 wt % at best, and is accompanied by undesirable heating of the solid expeller cake, and possible deterioration of the oil. For this reason, prepressing is normally used in the oil industry to obtain a residual oil content of 15 – 20 wt %: the remaining oil is obtained by extraction with light hydrocarbons. The mechanical wear in the presses and the ever-increasing cost of routine maintenance certainly favor total extraction, especially as the capital cost of pressing increases approximately in proportion to the throughput whereas the relative capital cost of leaching decreases.

Oil is obtained from oilseeds with relatively low oil content (e.g., soybeans) solely by extraction. In modern countercurrent plants, liquid – solid extraction can achieve residual oil contents of 0.3 wt %. Much higher throughputs can be achieved with one and the same plant and the same expenditure if residual oil contents of 0.6 – 1 wt % are acceptable. Most industrial plants now operate in this range, the more so because such oils are of improved quality and contain less unwanted contaminants.

The dependence of the degree of extraction on the throughput is typical of all leaching plants. A very high degree of technical effort is required to remove the final residues of extract; this can easily be demonstrated theoretically [1].

Sometimes extraction can yield different extracts depending on the conditions used. In the case of rice bran, for example, an oil is extracted at low temperature and a wax at high temperature. A polar solvent can be used to extract the water from

bones in a cold stage and the bone fat in a hot extraction step [2], [3].

2.2. Technical Principles and Process Design

Structure of the Feed Material. The extract may be present in the feed material in solid or liquid form. It may, like the oil in oilseeds, be enclosed in cells, or, as in coffee, be attached to the carrier material as a fine dispersion.

The feed material is not homogeneous but porous and contains many capillaries. At the start of extraction, the solvent penetrates into the material through the capillaries and dissolves the extract forming a highly concentrated solution. During the extraction process, diffusion produces virtual equalization of the concentrations of the extract solution in the feed material and of the extract solution surrounding the solid. After completion of the process, a certain quantity of solvent is retained by the residue as a result of adhesive forces. This adherent solvent contains a quantity of extract which remains in the extraction residue after the solvent has been expelled. In practice there is no such thing as total extraction.

Extraction Rate. Three main theories have been proposed to account for the factors governing the rate of extraction; to some extent, these can be substantiated experimentally and described by mathematical expressions. The diffusion theory [4] implies that the extraction rate depends primarily on the rate of diffusion. The soaking theory [5], [6] assumes that the extraction rate is determined not only by the rate of diffusion but also by the rate of dissolution of the extract in the solvent. A third theory interprets the extraction rate as a function of the flow rate of the extract solution in the capillaries [7].

The following requirements for a high extraction rate can be deduced from these theories:

1. The extraction should be carried out at the highest possible temperature because the viscosity and solubility of many extracts are temperature-dependent.
2. The capillary paths must be short so that the distance over which diffusion occurs is as short as possible. Short capillary paths are achieved in materials which are extracted using the displacement system (see Section Extraction Processes and Typical Equipment) by fine grinding of the feed material. If the extraction is carried out on the percolation principle, the feed material must be prepared in such a way that the solvent can percolate freely. For this reason, materials such as oilseeds are rolled to thin flakes to produce a feed material with short capillary paths in one direction and good percolation properties (Fig. 2) [8]. Flaking should disrupt the oil-containing cells to allow direct contact between the solvent and the extract. Solids such as fish meal, which are very finely divided, have short capillary paths, but the rate of percolation is so low that pelletizing must be used to produce a granular material with adequate percolation properties.
3. The percolation rate must be as high as possible in order to wash away the extract solution that has diffused to the surface of the solid particle. This ensures that a high concentration gradient always exists between the extract solution in the capillaries and the extract solution surrounding the particle. In the case of powdered solids, concentration equilibrium between the concentrated extract solution at the particle surface and the bulk extract solution can be attained more rapidly by stirring.
4. To achieve good efficiency at each stage in multistage extraction, only a small amount of solvent should be retained by the feed material. This is achieved industrially by using drainage zones. Pressing or centrifuging is often not justifiable because each of these techniques requires a high degree of technical effort.

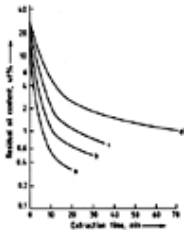

Figure 2. Extraction of soybean flakes of differing thickness during percolation with hexane

Flake thickness, mm: a) 0.22; b) 0.35; c) 0.43; d) 0.55

Number of Theoretical Extraction Stages. In order to design and dimension equipment for extraction and distillation of the solvent from the extract solution and the extraction residue, the solvent requirements and the number of theoretical extraction stages must be known. In this context—as in distillation—the number of theoretical extraction stages denotes the number of theoretical solvent equilibria that are required to achieve a given final state from a specified initial state. They are either calculated by drawing up the mass balances for the individual stages or determined graphically [1], [9]. The extract contents of the residue and the solution are specified. The composition of the feed material and the quantity of solvent taken up by it are measured experimentally as a function of the concentration.

The following numerical example of a simplified method for calculating the number of theoretical extraction stages is taken from [1]. This publication also contains the derivation of the equation for the operating curve used in the calculation.

The required number of theoretical stages is to be determined for an oilseed extraction with a light hydrocarbon such as hexane.

Initial oil content of the oilseed	20.0	wt %
Permissible residual oil content in the meal	0.5	wt %
Required oil content of the micella (extract solution)	30.0	wt %

Table 1 shows the results of the experimental determination of the amount of adherent solvent l_n and its content of extract e_n at full equilibrium.

Table 1. Experimental results of an oilseed extraction used to determine the number of theoretical extraction stages for 1000 g of carrier material

Oil concentration of the adherent solution x_n, wt %	Amount of adherent solution l_n, g	Amount of adherent extract e_n, g
50	874	437
30	600	180
20	530	106
10	480	48
0	450	0

The following equation is used to establish the operating curve:

$$y_n = \frac{(E_{end} - e_a) + e_n}{(L_{end} - l_a) + l_n}$$

where the symbols and subscripts in the equation have the following meanings:

e	amount of extract in the adherent solution
E	amount of extract in the free solution
l	amount of adherent solution
L	amount of free solution
$x = e/l$	extract concentration in the adherent solution
$y = E/L$	extract concentration in the free solution
a	material entering the system
end	material leaving the system
n	an arbitrary intermediate stage

The individual parameters in the equation are calculated as follows:

e_a: If the oil content of the seed is 20 %, then 1000 g of carrier material (as the reference quantity) corresponds to an oil content of 250 g, i.e., e_a= 250 g

E_{end}: For oil extraction giving meal with a residual oil content of 0.5 %, i.e., 5 g of oil per 995 g (≈ 1000 g) of carrier

$E_{end} = e_a - e_{end} = 250 - 5 = 245$ g

L_{end}: Since the required oil content of the miscella is 30 %, 245 g of oil corresponds to 30 % of the final amount of solution L_{end}, thus

$L_{end} = \frac{100}{30} \times 245 = 817\,\mathrm{g}$

l_a: As the fresh oilseed does not contain any solvent, the oil is present as a 100 % solution, i.e., $l_a = e_a$= 250 g

The equation then becomes

$$y_n = \frac{(245 - 250) + e_n}{(817 - 250) + l_n} = \frac{(-5) + e_n}{(567) + l_n}$$

Substitution of the experimental values for e_n and l_n listed in Table 1 give the values for y_n (Table 2).

Table 2. Calculation of the operating curve from the experimental results given in Table 1

x_n, wt %	e_n, g	l_n, g	y_n, wt %
0	0	450	–0.5
10	48	480	4.1
20	106	530	9.2
30	180	600	15

50 437 847 30

The points x_n and y_n are plotted in a coordinate system and then joined together to produce the operating curve. The equilibrium curve is the 45° line where $y = x$ (Fig. 3).

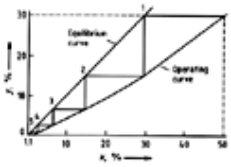

Figure 3. Graphical determination of the number of theoretical extraction stages

The steps between $x_{a,\ total}$= 50 % (corresponding to y = 30 %) and y = 0 % (corresponding to x_{end}= 1.1 %) produce five points, i.e., theoretical stages, on the equilibrium curve.

In this method, the extract is assumed to be completely dissolved in the solvent which means that extraction is not limited by diffusion. This is usually the case in industrial oil extraction where residence times are also determined by charging, discharging, and drainage. Thus, the number of theoretical extraction stages is virtually the same as the number of extraction stages required in practice. If dissolution equilibrium is not established quickly enough or is not achieved at all, the time dependence of the extraction rate must be determined experimentally and incorporated in the calculation [10].

Diffusional limitation in cellular structures is discussed in detail in [11-13].

Solvent Requirements. An “ideal” solvent must satisfy the following requirements:

1. High, selective dissolving capacity: only the required extract and no other constituents of the feed material are dissolved
2. Low specific heat, relative density, and heat of vaporization
3. Nonflammable, no formation of explosive mixtures with air
4. Nontoxic
5. Noncorrosive action on the extraction equipment
6. Good ability to penetrate the feed material, easy and complete removal from the extract and extraction residue without affecting their odor or taste
7. Chemically defined, homogeneous, and stable; constant boiling point which is not too high

No solvent fulfills all of these conditions, and the optimum solvent must be found for each extraction process. The most important solvents are: hydrocarbons (extraction naphtha), benzene, carbon disulfide, diethyl ether, acetone, chlorinated hydrocarbons, 2-propanol, ethanol, and water. Special effects can be achieved with mixtures of different solvents.

2.3. Extraction Processes and Typical Equipment

2.3.1. *Batch Extraction*

Batch extraction is based on the principle of displacement or enrichment.

Displacement Method. The displacement method is the oldest and simplest method of extraction. A given quantity of fresh solvent is added to the feed material. After thorough mixing, the extraction residue is allowed to settle and the extract solution is drawn off. If the liquid and solid phases cannot be separated by sedimentation, extraction pressing or centrifugation can be employed. The procedure is repeated until the required degree of extraction has been achieved. This system in which successive batches are treated with pure solvents is known as *multiple or cross-flow extraction.*

The disadvantage of this method is that, as the number of extraction steps increases, the concentration of the extract solution continually decreases and recovery of the solvent becomes more expensive.

The cross-flow method must, however, be used when a large concentration difference is required or when the mechanical flow properties of the feed material (slurry or powder) do not permit the use of any other method of processing. In principle, cross-flow extraction can also be performed continuously.

Enrichment Method. If the feed material has the required percolation characteristics, several extractors can be connected to form a battery. The solvent is then passed countercurrently through the extraction system. This mode of operation is called *counterflow battery extraction* or *enrichment process* (Fig. 4)

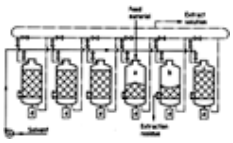

Figure 4. Schematic drawing of a battery extraction plant

a) Filling; b) Emptying; c) Separation of solvent and extraction residue; d) Drive

The solvent flows into the extractor containing the almost fully extracted material, and becomes increasingly enriched with extract as it flows through the following stages. It is withdrawn in a highly concentrated state from the extractor that contains the untreated material.

Investment cost and utilities consumption are lower for the enrichment process than for the displacement process.

Extractors. Batch (pot) extractors and rotating extractors are used in batch extraction. They serve both for the actual extraction process and for distillation of the solvent from the extraction residue.

Batch extractors generally comprise an upright cylindrical steel vessel of 2 to 10 m^3 capacity (Fig. 5). In processing by the enrichment method, the feed material rests on a perforated tray fitted into the vessel. The tray is not required in displacement extraction. Filling and emptying take place through the top of the vessel by loading the feed material into perforated trays. Separate connections are provided for solvent, steam, and vapor. The solvent is distilled off by direct heating. Steam is introduced through a purpose-designed distributor pipe.

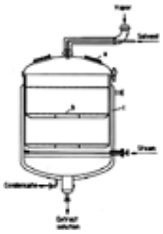

Figure 5. Pot extractor

a) Removable lid; b) Perforated tray; c) Steel vessel

Rotating extractors consist of a horizontal steel vessel which is supported on bearings at both ends so that it can rotate (Fig. 6). It is filled and emptied through connections in the shell. The connections for the solvent, steam, and vapor lead to or from the hollow axle, which allows admission and discharge of liquids and gas during rotation. The use of heating coils and a heated double-jacket permits distillation of the solvent by indirect heating; the heating coils also act as agitators. Direct steam can also be introduced through built-in steam pipes.

Figure 6. Rotating extractor with double jacket

a) Filling inlet for feed material; b) Drive

Batch extractors are simple and robust. Their disadvantages are their limited capacity and batchwise output of product. Periodic filling and emptying make high solvent losses almost unavoidable. Due to the batchwise mode of operation, the vapor produced during distillation of solvent from the residue can seldom be used for heating the distillation plant in which the solvent is separated from the extract.

2.3.2. *Continuous Extraction*

Continuous extractors operate primarily on the percolation system. The percolation method requires a feed material which is highly permeable to the solvent. This procedure has the advantage that the feed is undisturbed and is therefore exposed to little mechanical stress. Self-sfiltration takes place during extraction when the liquid passes through the material. The extract solution has a minimal content of fine material.

Many extractor designs are available, only a few characteristic types are described here. The first extractor operating on the percolation principle to have proved itself in practice is the Bollmann *bucket-elevator extractor* (Fig. 7).

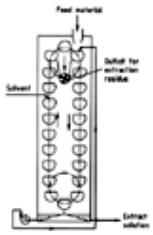

Figure 7. Bollmann bucket-elevator extractor

The extractor comprises an airtight casing in which buckets are fixed to an endless circulating chain. The feed material is fed into the buckets by means of a filling device. The solvent (which is already enriched with extract) is sprayed onto the buckets as they move downward; it then flows through the individual buckets, and is drawn off at the bottom of the extractor in highly concentrated form. On the downward side, extraction is cocurrent and self-filtration occurs. On the upward side, fresh solvent is fed in and flows countercurrently through the feed material, leaching out the extract. This extract solution is collected in a separate chamber at the base and is fed as the extracting medium onto the downward-moving buckets. After one cycle the buckets are tipped up and the extraction residue is discharged.

Continuous countercurrent percolation extractors with a capacity of up to 8000 t/d are widely used in modern large-scale plants to process bulk materials such as oilseeds. These types of extractor are also used for feed materials such as tea and drugs, which are processed in quantities of a few tonnes per day. The feed material is conveyed through the extractor by translation or rotation.

The material to be processed is fed continuously into the individual chambers (as in a box conveyor). In the extractor shown in Figure 8, the boxes move from right to left. Solvent or extract solution of differing concentrations is sprayed on in several stages. The liquid flow (from left to right in Fig. 8) is countercurrent to the material flow. Fresh solvent is fed in just before the feed material outlet; it passes through the material, is collected in a trough, and then pumped to the next extraction stage. This process is repeated in the subsequent stages and the solvent becomes increasingly enriched with extract. The highly

concentrated extract solution is withdrawn from the collecting trough of the final stage.

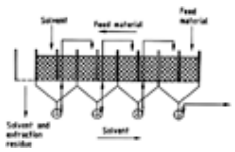

Figure 8. Schematic diagram of a continuous extractor

Higher liquid loads and, in many cases improved extraction for a given quantity of pure solvent, can be obtained if the extract solution from a given extraction stage is not passed immediately to the next stage but is recycled within the same stage. The associated collecting trough is then filled with extract solution. The amount of liquid that overflows into the next collecting trough corresponds to the inflow of extract solution from the previous collecting trough and the relevant extraction stage. In extractors operating on this principle, the height of the partitions between the collecting troughs decreases in the flow direction of the extract solution. Thus, the flow of extract solution is clearly defined and the liquid always overflows to the stage with the next highest concentration. This system is often used only in the first part of the extraction process, final leaching being a countercurrent process.

Figure 9 shows a *rotation extractor* or *carousel extractor* as supplied by Harburg-Freudenberger Maschinenbau GmbH. A rotor with sector-shaped cells rotates in an airtight housing over a fixed, slotted base. A sector of the base is open. The feed material is introduced continuously with a filling screw and extracted countercurrently. The extraction process is completed after a single revolution. The extraction residue is discharged through the open segment of the slotted base [14], [15].

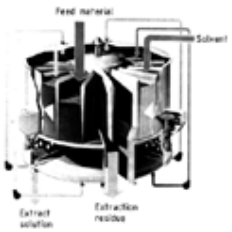

Figure 9. Harburg-Freudenberger carousel extractor

The *De Smet Reflex extractor* operate in a similar manner to the carousel extractor.

The *Lurgi sliding-cell extractor* is shown in Figure 10. An endless belt composed of U-shaped cells runs on roller tracks and pushes the oilseed over stationary screen plates. The feed material is introduced evenly into the cells of the upper belt by a filling device. After approximately half the extraction time has elapsed, the feed is then dumped onto the lower belt. When extraction is complete, the material is passed through a drainage zone and discharged. The solvent passes through the feed material countercurrently to the direction of the belt and becomes enriched with extract; it finally leaves the extractor after a material filter stage.

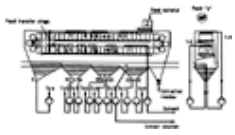

Figure 10. Lurgi sliding-cell extractor

In the *De Smet LM extractor* (Fig. 11), the feed material is conveyed on a perforated belt and the height of the material bed is regulated by an adjustable slide. The extraction residue is discharged through a discharge gate.

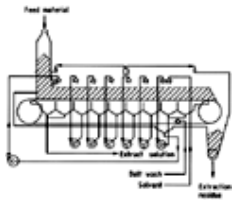

Figure 11. De Smet LM extractor system

The *Crown Iron extractor* is based on the principle of a loop-trough chain conveyor (Fig. 12). It has one cocurrent extraction stage and two countercurrent extraction zones comprising several countercurrent stages.

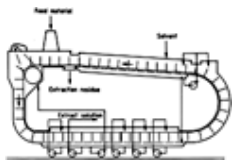

Figure 12. Crown Iron loop-trough chain conveyor extractor

Continuous processes have a series of important advantages over batch processes. Large quantities of feed material can be processed in equipment of relatively small dimensions and compact design. High concentration of the extract solution is achieved with a low residual content of extract in the extraction residue; this allows a considerable reduction of solvent usage. The extraction times are shorter than those of batch methods and this has a beneficial effect on the quality of the end products. Economic, quantitative solvent recovery from the extraction residue and the extract solution constitutes an important part of liquid – solid extraction but is discussed elsewhere (⟶ Evaporation; ⟶ Distillation and Rectification – Energy Economization in Rectification).

[Top of Page]

3. Chemical Leaching in Hydrometallurgy

Brent Hiskey

History. PLINY described the extraction of salt from wood ash by the solvent action of brine [16].

AGRICOLA in his *De Re Metallica* (1556) indicates that the separation of silver and gold with nitric acid was well established by his time [17]. In 1661, a license was granted for the recovery of copper from mine waters at Rio Tinto in Spain; this probably involved the recirculation of leach solution through waste heaps [18].

An important development in liquid – solid extraction was the cyanidation of gold and silver ores. As early as 1783, aqueous solutions of potassium cyanide were known to exhibit a solvent action on gold [19]. However, it was not until the pioneering work of MACARTHUR and the FORREST brothers (1887, 1889) that cyanidation became a commercial process for treating gold ores [20]. Their contribution included two important advances: (1) the use of dilute cyanide solutions to selectively dissolve gold (stronger lixiviants, tend to dissolve numerous impurities), and (2) a simple method for recovering gold from the cyanide solutions by precipitation with zinc shavings.

3.1. Methods

Chemical leaching is customarily the first step in a hydrometallurgical sequence, but it may also be applied whenever the dissolution of a solid phase is desired. Figure 13 shows a general hydrometallurgical flow sheet. A broad range of raw materials can be treated: some are primary materials such as ores, mineral concentrates, and certain intermediate products (i.e., calcines, mattes); others are secondary materials such as scrap, sludge, and residues. Selective leaching can be used to concentrate a valuable constituent in the solid phase, as in the leaching of metallic copper from refinery anode slimes to concentrate gold and silver for subsequent recovery. Alternatively, leaching is often used to obtain an extract, called a pregnant leach solution, from which the valuable constituent may be recovered; the solid residue or gangue is discarded.

Figure 13. Major components of hydrometallurgical flow sheets

Leach solutions frequently require further processing to concentrate the desired components and remove impurities. A number of separation processes are used, including crystallization, adsorption, selective precipitation, cementation, ion exchange, and liquid – liquid extraction.

Product recovery is the final step in the hydrometallurgical sequence; the valuable component is then converted into a marketable product. Three techniques commonly employed for product recovery are electrolysis, gaseous reduction, and precipitation.

Leaching methods vary depending on the amenability of the solid to the specific leaching reaction and on the content or grade of the material. High-grade materials generally justify the added cost of prior comminution, sizing, and in some cases, concentration (i.e., flotation, gravity concentration, selective flocculation). Figure 14 illustrates different processing routes for leaching low-, intermediate-, and high-grade ores [21]. Ore is usually transported from mineral deposits for leaching; however, when the mineral chemistry and geologic conditions permit, it can be leached in situ. Low-grade ore, which does not justify any beneficiation, is typically leached on dumps and heaps. Intermediate-grade ores are often crushed or ground and sized for direct leaching in vats and agitated vessels. High-grade ore is normally of sufficient value to warrant the preparation of mineral concentrates. The concentrates may be suitable for leaching directly or may require pretreatment.

Figure 14. Processing routes for leaching various grades of ore

In situ leaching is performed on a mineral deposit in its natural or original geologic position. Liquid – solid contact may occur under either unsaturated or saturated flow conditions. Permeability enhancement may be required to improve solution flow through the rock mass and within rock pores and fractures. This can be accomplished by explosives, hydrofracturing, or chemical dissolution. To achieve high recoveries, in situ leaching often requires long leach cycles (usually years). Recovery

of copper from oxide and sulfide deposits, uranium from low-grade sandstone deposits, and soda ash from deep trona deposits are excellent examples of in situ leaching.

Heap and dump leaching are used to treat low-grade waste materials and some moderate- to high-grade ores not suitable for beneficiation. Leaching solutions are usually sprayed or sprinkled on the surface of the heap or dump and allowed to percolate through the material by gravity. The pregnant leach solution containing valuable dissolved components is collected and delivered to a metal-recovery circuit.

Figure 15 shows a heap-leaching system in which the run-of-mine (i.e., untreated) or crushed ore is placed on a specially prepared, impervious pad. Heap leaching has been used to recover copper and uranium, and has been widely applied to the treatment of low-grade gold and silver ores.

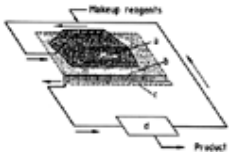

Figure 15. Schematic of a heap-leaching operation

a) Heap; b) Collection channel; c) Impervious leaching pad; d) Product recovery plant

Dump leaching is usually used to treat waste rock and, as such, does not justify crushing. Most dumps are deposited directly on the ground with little if any base preparation. Dump leaching is usually performed over a period of years, whereas heap leaching involves shorter periods (days to months). The dump leaching of low-grade copper waste from open-pit porphyry operations in the western United States is an important example of this technique [22].

Vat Leaching. In some cases, the valuable components desired may be finely dispersed in a nonporous solid matrix. Crushing and grinding may be required to permit adequate contact between the leaching solutions and the solid phase. The material can be leached in vats or tanks if the permeability of the bed permits percolation and circulation of the leaching solution. This system allows good liquid – solid contact, as well as washing and recovery of dissolved valuable components from the leaching residue. When compared with heap or dump leaching, overall recovery is usually higher, leach cycles are shorter (days), pregnant leach solution concentrations are higher, and soluble losses are lower. The vat leaching of copper oxide ores in Arizona is an important example of this technique [22].

Agitation Leaching. Fine grinding may be necessary to expose enclosed particles or to generate sufficient surface area to maximize extraction rates. Suspensions of fine particles can be leached in a tank or vessel by mechanical or air agitation in batch, continuous cocurrent, or continuous countercurrent circuits. Agitation (i.e., mixing) has two important effects: (1) the suspension and dispersion of solid particles in the solvent and (2) the promotion of efficient mass transfer between the solid and liquid phases. Gold and silver cyanidation is commonly accomplished by agitation leaching in staged circuits with residence times of 24 – 48 h. Other examples are the leaching of copper oxides or uranium ores, and the purification of molybdenite concentrates by dissolving impurities [23].

Pressure Leaching. In pressure or autoclave leaching, hydrometallurgical reactions are performed at high pressure above the normal boiling point of the leaching solvent. Other high-pressure systems use over-pressures of gaseous reactants to increase the solubility of the gas. Higher temperatures can increase both the solubility of dissolved species and their reaction rates. Autoclaves typically used for pressure leaching contain a mixer – impeller for agitation of the pulp, which is allowed to flow under pressure through the reactor. A classic example of pressure leaching is the Bayer process for the production of alumina (⟶ Aluminum Oxide – Bayer Process). Another important example is the pressure leaching of base-metal sulfides. Pressure oxidation is also applied to the treatment of refractory gold ores and concentrates [24], [25].

3.2. Principles

Selection of the leaching agent or *lixiviant* is very important and depends on the following factors:

1. Physicochemical nature of the solid to the extracted
2. Selectivity and stability of the leaching agent
3. Regeneration of the agent
4. Ability to recover the desired components from the leaching solution
5. Cost, toxicity, and availability of the agent

Water is the simplest leaching agent. A metal salt MX dissolves in water to yield its ions:

$$MX \rightleftharpoons M^+ + X^-$$

The thermodynamic equilibrium is established by the solubility product

$$K_{so} = [M^+][X^-]$$

Table 3 lists the solubility products and the total solubilities of selected metal sulfates [26-28].

Table 3. Solubility of metal sulfates in water

Metal sulfate	K_{so} *	Solubility, g/L **

$SrSO_4$	3.24×10^{-7}	0.14
$PbSO_4$	1.58×10^{-8}	0.045
$BaSO_4$	1.1×10^{-10}	0.00223
Ag_2SO_4	1.70×10^{-5}	5.7 (0 °C)
$CuSO_4$	4.43×10^{2}	225
K_2SO_4	0.0193	120
Na_2SO_4	0.645	338
Hg_2SO_4	7.41×10^{-7}	0.6
$CaSO_4$	9.12×10^{-6}	2.09 (0 °C)
$Al_2(SO_4)_3$	7.66×10^{16}	313 (0 °C)

* Calculated from data in [26] (25 °C).
** Solubility values at 25 °C except where otherwise indicated [28].

Some solids undergo hydrolysis when they dissolve in water, for example,

$$CaO + H_2O \rightarrow Ca(OH)_2$$

$$As_2O_3 + 3\,H_2O \rightarrow 2\,H_3AsO_3$$

In the presence of air or oxygen, water reacts with metal sulfides to produce soluble metal sulfates. The aqueous oxidation of the iron disulfide-containing mineral pyrite [1309-36-0] is an important example:

$$FeS_2 + 7/2\,O_2 + H_2O \rightarrow FeSO_4 + H_2SO_4$$

This reaction is useful in the dump leaching of low-grade, waste-containing copper sulfide [29]. It provides a source of sulfuric acid and dissolved iron, both necessary in the leaching of copper sulfide minerals (i.e., chalcopyrite, chalcocite, bornite): the iron(III) ions act as an oxidant and the sulfuric acid stabilizes the iron(III) ions in solution. Heat generated by the oxidation of sulfide promotes air convection and thus transport of oxygen through the rock mass. Bacteria use the oxygen to oxidize iron (II) ions to iron(III) ions, which constitute the primary oxidant.

To achieve the desired results in the leaching step, the leaching solution must be formulated to suit the problem at hand. This may involve adjusting the pH, setting a specific redox potential, or introducing complexing ligands.

Thermodynamics. Thermodynamic relationships of solids in contact with aqueous solutions of defined composition are best illustrated by redox potential (E_h) vs. pH, or Pourbaix, diagrams [30]. Although developed for metal corrosion equilibria, these diagrams are equally appropriate for geochemical and hydrometallurgical relationships.

Figure 16 represents the oxidation potential – pH diagram for the Cu – O_2– H_2O system. The broken lines indicate the limits of water stability; expressed as reduction reactions, they are

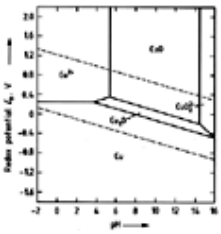

Figure 16. Redox potential – pH diagram showing stability regions for the copper – water system at 25 °C and 101.3 kPa

Concentration of dissolved copper species is 1×10^{-3} mol/L. Broken lines represent the limits of water stability.

$$2\,H^+ + 1/2\,O_2 + 2\,e^- \rightarrow H_2O(\text{upper limit})$$

$$2\,H^+ + 2\,e^- \rightarrow H_2(\text{lower limit})$$

Hydrogen is evolved below the lower line and oxygen above the upper line.

Figure 16 shows that copper oxide in its highest oxidation state may be soluble in a strong acid (to the left of the vertical line at pH 5.5, the stable phase is Cu^{2+} and not CuO)

$$CuO + 2\,H^+ \rightarrow Cu^{2+} + H_2O$$

or a strong base

$$CuO + 2\,OH^- \rightarrow CuO_2^{2-} + H_2O$$

Copper oxide in its lowest oxidation state may require oxidizing conditions for solution:

$$Cu_2O + 1/2\,O_2 + 4\,H^+ \rightleftharpoons 2\,Cu^{2+} + 2\,H_2O$$

However, some metal oxides in the highest oxidation state require reducing conditions to accomplish dissolution. An example of reductive leaching is the dissolving of manganese dioxide in sulfurous acid:

$$MnO_2 + 2\,H^+ + H_2SO_3 \rightleftharpoons Mn^{2+} + H_2SO_4 + H_2O$$

Manganese is reduced from the 4^+ to the 2^+ oxidation state in solution.

The role of a complexing agent is best illustrated with the gold – cyanide system. Figure 17 shows the redox potential – pH diagram for the Au – O_2– H_2O system. Metallic gold occupies a large area that includes the entire domain of water stability. The gold(I) ion (Au^+) does not appear on the diagram because it disproportionates spontaneously to the gold(III) ion and metallic gold

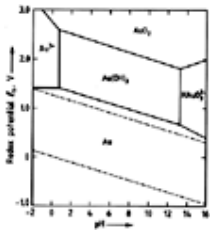

Figure 17. Redox potential – pH diagram showing stability regions for the gold – water system at 25 °C and 101.3 kPa

Concentration of dissolved gold species is 1×10^{-4} mol/L. Broken lines represent the limits of water stability.

$$3\,Au^+ \rightarrow Au^{3+} + 2\,Au$$

The gold(III) ion (Au^{3+}) and other oxidized forms of gold occur only at potentials above the upper stability limit of water. However, at these potentials, water is oxidized to oxygen, and the oxidized gold species is reduced simultaneously to the metallic state. Thus, metallic gold cannot be oxidized in strong acid or alkali in the absence of complexing ligands.

Gold dissolves readily in the presence of a complexing ligand and an oxidizing agent. Figure 18 represents the redox potential – pH diagram for the Au – CN – H_2O system. The presence of cyanide as a complexing agent results in a relatively large region of $Au(CN)_2^-$ stability, which makes cyanidation a practical process for leaching gold ores (⟶ Gold, Gold Alloys, and Gold Compounds).

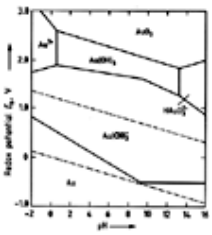

Figure 18. Redox potential – pH diagram showing stability regions for the gold – cyanide – water system at 25 °C and 101.3 kPa

Concentration of dissolved gold species is 1×10^{-4} mol/L and of cyanide 1×10^{-3} mol/L. Broken lines represent the limits of water stability.

Reaction Kinetics. Reaction rate data are of practical importance for two reasons: (1) full-scale commercial leaching operations can be designed from kinetic data collected from bench-scale and pilot-plant tests, and (2) kinetic data can be used to predict processing under various conditions.

Liquid – solid extraction is a multistep, heterogeneous process that includes [31]:

1. diffusion of reactants to the particle surface,
2. adsorption,
3. surface chemical reaction,
4. desorption, and
5. diffusion of products to the bulk solution.

If a solid product layer surrounds an unconverted core, both the reactants and the soluble products must diffuse through the pores of this layer. The overall reaction rate is determined by the slowest step(s); in some cases, diffusion is rate-limiting.

Diffusion. Transport of a reactant to the solid – solution interface involves diffusion through a viscous boundary layer in the vicinity of the solid. Figure 19 depicts the concentration profile of the diffusion boundary layer for inward flux; the solid line represents the true solute concentration, and the dashed line is the assumed linear concentration [32]. Fick's first law expresses mass transfer from the solution phase to the solid surface:

$$J = -D\frac{\partial C}{\partial x}$$

where J is the mass flux to the surface (mol cm^{-2} s^{-1}), D is the coefficient of diffusion (cm^2/s), C is concentration (mol/cm^3), and x is the position coordinate perpendicular to the reference plane (cm). Thus,

$$J = \frac{1}{A}\frac{dn}{dt}$$

where *n* is the amount of material diffusing (mol), *t* is time (s), and *A* is the area of the reference plane (cm^2). Under steady-state conditions (*J* = constant), the integrated form of Fick's first law yields

$$\frac{dn}{dt} = -DA\frac{\Delta C}{\Delta x}$$

The term $D/\Delta x$ is often used as a mass-transfer coefficient. The thickness of the diffusion boundary layer (Δx) depends on the degree of agitation (i.e., flow velocity) and the viscosity. Vigorous agitation reduces the diffusion boundary-layer thickness and results in higher mass-transfer coefficients. The minimum value of Δx (i.e., the limiting boundary-layer thickness) has been calculated to be ca. 1×10^{-3} cm for violent agitation [33]. The value of the diffusion coefficient for aqueous solutions is ca. 1×10^{-5} cm^2/s for all species (simple cations, simple anions, neutral molecules, gases, etc.). Therefore, the mass-transfer coefficient should be in the range of 1×10^{-2} cm/s for diffusion through a limiting boundary layer. Besides minimizing Δx, the rate of leaching (dn/dt) can be increased by increasing the interfacial area or the concentration of the reacting species.

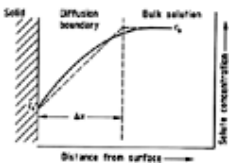

Figure 19. Concentration profile of the diffusion boundary layer for inward flux

C_b= concentration in bulk phase; C_s= concentration in solid phase.

Temperature. The temperature dependence of the reactions involved in liquid – solid extraction is best analyzed in terms of the Arrhenius law [34]:

$$k = ve^{-E_a/RT}$$

where *k* is the specific rate constant, *v* is the frequency factor, and E_a is the activation energy. The diffusion coefficient is also temperature-dependent and obeys an Arrhenius-type equation. However, the activation energy for diffusional processes is usually low: 8 – 13 kJ/mol for aqueous diffusion and 16 – 21 kJ/mol for viscous boundary-layer diffusion. In contrast, a process controlled by a surface chemical reaction exhibits a much higher activation energy (42 kJ/mol). Near ambient temperature, the reaction rate approximately doubles for a 10 °C increase in temperature when E_a= 50 kJ/mol.

Particle geometry is another important factor in liquid – solid processes because the reaction rate depends on the total surface area available:

$$\frac{dn}{dt} = Ak'C$$

where *A* is the surface area, *k´*is the rate constant, and *C* is the concentration. Sample geometry influences the way in which the surface area varies as the reaction proceeds. Shapes such as flat plates and disks exhibit a minimum variation in area, whereas samples with a cubic, spherical, or cylindrical wire-type geometry exhibit the greatest variations in area. The effect of sample geometry and the “shrinking core model” are treated in detail in [35-37].

Electrochemical Processes. In liquid – solid extraction (leaching reactions), electrochemical dissolution occurs at the surface of a conducting solid (i.e., metal M):

$$\text{Solid}(M) \rightleftharpoons \text{Ions in solution}(M^{n+}) + \text{Electrons in solid}(ne^-)$$

Figure 20 depicts the flow of electrons in the solid and the transfer of ions into solution. The electron acceptor is the oxidant used to dissolve the materials. Most metal sulfides and certain oxides (UO_2, MnO_2, SnO_2) are semiconductors and, therefore, their electronic properties favor electrochemical dissolution [38]. This involves the establishment of an internally short-circuited system, in which the solid assumes a mixed potential determined by the spatially separated anodic and cathodic processes.

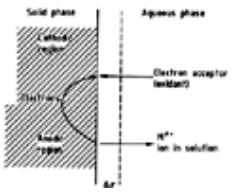

Figure 20. Electrochemical dissolution at a solid surface

When conducting or semiconducting minerals in local electrical contact are subjected to leaching, a galvanic interaction may occur which causes enhanced corrosion of one or more members of the electrochemically coupled mineral regime and cathodic protection of other mineral constituents. Cathodic protection occurs because one material dissolves “sacrificially” and protects the other against dissolution by making it negative with respect to its equilibrium potential. The mineral with a lower “rest” potential will react anodically (dissolution), whereas cathodic behavior (passivation) may be imposed on the

mineral possessing the higher “rest” potential.

The cyanidation of gold and silver involves an electrochemical mechanism [39]. The dissolution of uranium dioxide in both acidic and alkaline carbonate solution is also electrochemically controlled [40-42]. Chalcopyrite [1308-56-1] ($CuFeS_2$) leaching in oxygenated ammoniacal solution has been explained in terms of an electrochemical model [43]. Galvanic coupling between pyrite (FeS_2) and chalcopyrite in a leaching environment results in enhanced dissolution of $CuFeS_2$ and passivation of the FeS_2 phase [44].

Biological Processes. The role of bacteria in leaching sulfidic deposits has long been recognized and is well documented, particularly in the leaching of copper (western United States and the former Soviet Union) and of uranium (Canada) [45].

The autotrophic aerobe Thiobacillus ferrooxidans has been associated with the leaching of base-metal sulfide minerals [46], [47]. It can utilize energy obtained by the metabolic oxidation of reduced inorganic sulfur and iron(II) and, simultaneously, catalyze the leaching of metal sulfides.

The oxidative leaching of metal sulfides (MS) with iron(III) ions is represented by

$$MS + 2\,Fe^{3+} \rightarrow M^{2+} + S + 2\,Fe^{2+}$$

The elemental sulfur produced forms a dense, tenacious coating on the partially reacted metal sulfide. This product layer is nonporous and makes the leaching rate extremely slow because it hinders diffusion. Oxidation of elemental sulfur is catalyzed by the bacteria present:

$$S + 3/2\,O_2 + H_2O \rightarrow H_2SO_4$$

The iron(II) ions produced during the oxidation of metal sulfides are also reoxidized with the acid of the bacteria:

$$2\,Fe^{2+} + 1/2\,O_2 + 2\,H^+ \rightarrow 2\,Fe^{3+} + H_2O$$

The mechanisms involved in bacterial leaching are still a matter of dispute; they may catalyze the reaction passively or may actively promote dissolution [48]. Extremely complex interactions occur among microorganisms, mineral substrates, and nutrients. *Thiobacillus ferooxidans* and *T. thiooxidans* show an amazing ability to adapt to the leaching environment. They can function over a wide range of pH and solution composition; some strains have adapted to tolerate high concentrations of metal ions under very acidic conditions. For example, *T. ferrooxidans* has been able to adapt to iron and arsenic concentrations of 50 and 20 g/L, respectively, at a pH of 0.5 [49]. The temperature range for optimum *T. ferrooxidans* and *T. thiooxidans* activity is limited and lies between 25 and 45 °C [50]. A chemoautotrophic, thermophilic microorganism (*Sulfolobus*) has been isolated and is effective above 60 °C [51].

Bioleaching is attracting attention for the treatment of refractory precious-metal ores and concentrates [52-54]. Some gold ores and concentrates containing sulfide minerals (i.e., pyrite and arsenopyrite) are refractory to (i.e., withstand) conventional cyanidation. Biooxidation of the sulfide prior to cyanidation has proved to be effective for gold and silver recovery from refractory materials. Bioleaching techniques have been developed for the heap leaching of ores and the leaching of concentrates in stirred tanks [49].

3.3. Typical Processes and Equipment

Typical examples illustrating the use of liquid – solid extraction (leaching) are (1) agitation leaching of an oxidized gold ore with alkaline cyanide solutions, (2) pressure leaching of a zinc sulfide concentrate and (3) biological leaching of a refractory gold ore.

3.3.1. *Agitation Leaching of an Oxidized Gold Ore*

Selection of the processing methods for gold ores depends on economic, metallurgical, and engineering considerations. For a detailed description of gold production, see ⟶ Gold, Gold Alloys, and Gold Compounds. Conventional milling (agitation leaching) is generally preferred to heap leaching for higher grade deposits that justify the added costs of crushing and grinding [55]. Agitation leaching also maximizes gold recovery by increasing the exposure of metal surfaces to the leaching solution. The size of the agitation reactor depends on the feed rate and the time necessary to achieve the desired degree of extraction. Usually, a stagewise (semibatch) operation is used to balance fluctuations in the feed composition and to control the continuity of output.

Figure 21 shows a simplified flow sheet for a leaching circuit consisting of six agitation vessels. The ore is ground to give a particle size that yields acceptable extraction levels (>90 %). A ball mill is linked in closed circuit with a hydrocyclone classifier. The ground and sized product is thickened to between 45 and 50 % solids prior to leaching. The overall gold-leaching reaction is

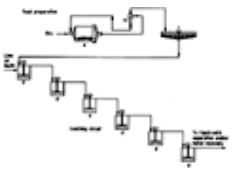

Figure 21. Flow sheet of an agitation leaching circuit for precious-metal ores

a) Ball mill; b) Cyclone; c) Thickener; d) Agitated leaching vessel

$$4\,Au + 8\,CN^- + O_2 + 2\,H_2O \rightleftharpoons 4\,Au(CN)_2^- + 4\,OH^-$$

Lime (or caustic), sodium cyanide, and air are introduced in the pulp to achieve the desired dissolution. In each leaching stage, air is mixed vigorously into the pulp by air agitation or mechanical impellers.

The draft-tube circulator – aerator shown in Figure 22 combines efficient oxygen transfer and effective mixing [56]. The agitator shaft and impeller are located inside a hollow tube, and compressed air is introduced immediately below the impeller. These systems are used in deep tank mixing.

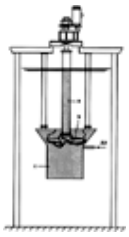

Figure 22. Schematic of a draft-tube submerged turbine aerator

a) Agitator shaft; b) Impeller; c) Draft tube

Courtesy of Philadelphia Gear Corp.

The pulp leaving the last leaching vessel proceeds to a liquid – solid separation step to prepare the leach solution for metal recovery and to wash the solids free of dissolved gold. This can be accomplished, for example, by countercurrent decantation (CCD) washing or filtration which, however, require a high capital investment. Certain pulps do not respond readily to liquid – solid separation. New carbon-in-pulp (CIP) technology overcomes these limitations and has gained widespread acceptance in the processing of gold ores [57], [58]. It is effective for low- and variable-grade solutions.

A typical CIP extraction circuit is shown in Figure 23. Ore leached as shown in Figure 21 is fed into the CIP circuit. The objective is to adsorb the dissolved aurocyanide complex onto activated carbon in a countercurrent fashion; the process consists of the following steps:

1. Prescreening
2. Multistage, countercurrent carbon adsorption
3. Carbon elution and reactivation
4. Gold reduction (electrowinning)

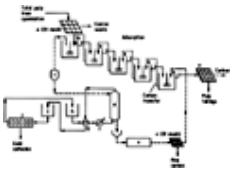

Figure 23. Flow sheet for a carbon-in-pulp (CIP) gold—recovery circuit

a) Screen; b) Interstage screen (20 mesh); c) Carbon transfer vessel; d) Carbon stripping column; e) Carbon reactivation kiln; f) Gas-fired exchanger; g) Heat exchanger; h) Fresh stripping solution; i) Pregnant stripping solution; j) Electrowinning

Before the pulp enters the adsorption circuit, it is prescreened to remove coarse material that might otherwise contaminate the loaded carbon and to eliminate waste that might block the interstage screens (b). During subsequent adsorption, carbon moves stepwise in the opposite direction to the pulp. The grains of carbon are coarser than the ore particles and can, therefore, be separated from the pulp by simple screening on interstage screens. Fresh carbon is added to the last adsorption vessel; when sufficiently loaded with gold, this is conveyed batchwise from tank to tank by airlifts or recessed impeller pumps. The carbon in the first adsorption vessel is highly loaded with gold and is transferred to a stripping column (d) where gold is eluted from the carbon by using high-pressure, high-temperature NaOH/NaCN solutions, or by alcohol stripping. The carbon is reactivated in a kiln (e) and returned to the last adsorption vessel. Gold is obtained from the pregnant stripping solution by electrowinning (j). Carbon is separated from the pulp by screening.

3.3.2. *Pressure Leaching of Zinc Sulfide Concentrates*

The world's first commercial zinc sulfide pressure leaching plant was built by Cominco in Canada [59]. Zinc sulfide [1314-98-3] concentrates are leached oxidatively in an autoclave reactor to produce a zinc sulfate solution and elemental sulfur, according to the following reaction:

$$ZnS + H_2SO_4 + 1/2\,O_2 \rightarrow ZnSO_4 + S + H_2O$$

Figure 24 shows a flow sheet of the zinc pressure leaching process. Reground zinc concentrates, typically containing 49 % Zn, 11 % Fe, 4 % Pb, and 33 % S, are fed continuously to the autoclaves with preheated spent electrolyte containing 150 g/L of sulfuric acid. The autoclave reactor contains four compartments and measures 3.7 m in diameter and 15.2 m in length. It consists of a mild-steel outer shell lined with lead, a refractory cloth, and acid-resistant brick (Figure 25). Pulp enters at one end and is discharged at the other after passing through the four compartments, each of which is equipped with a mechanically driven impeller to provide agitation.

Figure 24. Flow sheet for the zinc pressure leaching process

a) Grinding; b) Autoclave; c) Heat exchanger; d) Flash tank; e) Flash tank; f) Conditioning tank; g) Flotation cells; h) Tank for melting sulfur; i) Filter press

Courtesy of Cominco.

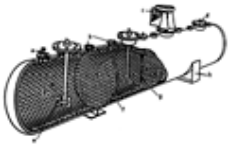

Figure 25. Schematic of a lead- and brick-lined autoclave reactor with a carbon steel shell

a) Service cluster; b) Service nozzles; c) Motor drive assembly; d) Agitator nozzle, e) Agitator; f) Compartment divider; g) Dip pipe; h) Support saddle

Courtesy of Sherritt Gordon Mines Ltd.

The slurry is pumped to the first compartment of the autoclave at the delivery pressure up to 1.85 MPa. Preheated spent electrolyte from the tank house is introduced at ca. 1.5 MPa and 70 °C. Heat of reaction elevates the temperature to 145 – 155 °C. The oxygen partial pressure is maintained at ca. 0.75 MPa with a total overpressure of 1.3 MPa. The system is designed for a leach retention time of about 100 min and for 98 % zinc extraction. The zinc is extracted into solution as zinc sulfate; sulfide is oxidized to elemental sulfur in the molten state, and iron and lead are rejected into the residue as insoluble plumbojarosite [$PbFe_6(SO_4)_4(OH)_{12}$]. The complete process chemistry is described in [60].

Oxidative pressure leaching has been applied to the treatment of sulfidic precious-metal oresand concentrates which are refractory to conventional cyanidation [24], [25]. Pressure oxidation is an effective way of “chemically” liberating precious metals and offers an attractive alternative to roasting.

A pressure oxidation process for pretreatment of a refractory gold ore has been developed by the Homestake Mining Company. The commercial plant can treat 3000 t/d of ore by pressure oxidation and carbon-in-leach cyanidation [61].

3.3.3. *Bioleaching of Refractory Precious-Metal Concentrates*

One of the most promising applications of bioleaching technology is the pretreatment of refractory precious-metal ores and concentrates. Continuous biological preoxidation of sulfidic ores and concentrates is being developed to compete with pressure oxidation and roasting. Bioextractive processes have been proposed for leaching concentrates of lead sulfide [62], copper – nickel complex sulfides [63], and chalcopyrite [64].

Many gold ores and concentrates that contain sulfidic minerals such as pyrite (FeS_2) and arsenopyrite [1303-18-0] (FeAsS) are refractory to conventional cyanidation, partially because the gold is finely disseminated within the sulfide matrix in a submicroscopic form or in solid solution, and as such may not be liberated even with fine grinding. Chemical dissolution of the sulfide matrix exposes the gold for subsequent extraction by cyanidation. Certain bacteria can metabolize and dissolve pyrite and arsenopyrite under acidic conditions according to the following exothermic reactions:

$$2\,FeS_2 + 15/2\,O_2 + H_2O \rightarrow 2\,Fe_3^{+} + 2\,H^{+} + 4\,SO_4^{2-}$$

$$FeAsS + 7/2\,O_2 + H^{+} + H_2O \rightarrow Fe^{3+} + SO_4^{2-} + H_3AsO_4$$

Bacterial strains (e.g., *Thiobacillus ferrooxidans*) have been adapted which can tolerate high concentrations of iron and arsenic at low pH [49]. However, temperature continues to have a regulating role in determining the activity of *T. ferrooxidans.* The heat generated by the exothermic bacterial oxidation of sulfides is normally removed via cooling coils placed in the bioleaching reactor.

Figure 26 is a flow sheet of a biological preoxidation circuit for processing sulfidic material to improve gold and silver recovery by cyanidation.

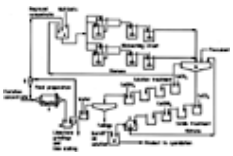

Figure 26. Flow sheet of a bioleaching process for treating refractory precious-metal concentrates

a) Conditioning tank; b) Thickener; c) Filter; d) Repulp tank; e) Sump; f) Ball mill; g) Cyclone

Courtesy of Albert Bruynestyne.

The bioleaching section consists of two parallel streams, each with four reactors divided into three stages; total retention time is 96 h in each line. The first stage employs two reactors in parallel (retention time 48 h). The second and third stages use a single reactor each (retention time, 24 h per reactor).

Leaching efficiency is improved by feeding reground concentrate into the bioleaching circuit. The feed will run 90 % <37 μ (400 mesh) and will be adjusted to 20 % solids. Because the oxygen required for leaching pyrite and arsenopyrite is fairly high, the reactors must be designed for efficient mass transfer of oxygen. Air is sparged into the reactors, and mechanical agitation ensures adequate oxygen uptake.

The leached slurry leaving the last reactor is concentrated to provide a thickened solid which is neutralized prior to cyanidation. A series of limestone – lime neutralization steps delivers a product slurry at about pH 10 for feed to cyanidation. A portion of the bioleach residue thickener overflow is returned to the bioleaching circuit. Another portion is bled to the solution treatment section to eliminate iron, arsenic, and sulfate, from the system. Another series of limestone – lime neutralization steps is employed to treat the bleedstream. A tailing containing primarily gypsum, jarosite, and ferric arsenite is thickened for disposal. The resulting neutral solution is added to the ground feed.

[Top of Page]

4. Extraction with Supercritical Fluids

Gerd Brunner

Gaseous material of high density can be used industrially in the same way as liquid solvents [65]. At moderate pressure (30 – 50 MPa) and close to the critical temperature T_{cr}, the densities of gases are comparable to those of liquids. These dense gases ($T > T_{cr}$) or compressible fluids ($T < T_{cr}$) display a high solvent capacity. Their mean intermolecular distance is so small that molecular interactions are of importance. The properties of a compressed substance in the region of the critical temperature can be modified by changes in pressure and/or temperature in such a way that the substance behaves either as a normal low-density gas or as a liquid.

A large difference in volatility exists between the compressed gas solvent and the substances dissolved in it. After expansion, almost all of the solvent can be separated. Thus, the properties of gases in the supercritical state (i.e., $T > T_{cr}$; $p > p_{cr}$) and liquids in the subcritical state close to the critical temperature (i.e., $0.95 \cdot T_{cr} < T < 1 \cdot T_{cr}$; $p < p_{cr}$), largely correspond to those of a solvent. These substances are of potential use in (extraction) processes employing solvents. The operating temperature of such a process is largely determined by the critical temperature. Carbon dioxide (T_{cr} = 31 °C, p_{cr}= 7.3 MPa) is a readily available gas with favorable properties. The process is known as *gas extraction* or *supercritical-fluid extraction* and is mainly of interest for the separation of temperature-sensitive mixtures of low volatility or mixtures in which liquid solvent residues must be avoided.

4.1. Principles

Compressed gases can be used as solvents in many different industrial processes. The most common application of high-density compressible fluids is as physical solvents in one-stage processes. However, multistage processes and countercurrent operations are also possible, and compressed gases can be used as solvents for chemical reactions.

Types of Processes. In single-stage processes (Fig. 27 A), one or more components are extracted from the feed mixture in an extraction vessel (a) with the aid of the compressed gas. The separation factor α for the components and the residue approaches infinity. The separation factor for components A and B of a mixture (e.g., the extract and the residue) is given by

$$\alpha_{AB} = \frac{x_A^{I}}{x_A^{II}} \cdot \frac{x_B^{II}}{x_B}$$

where x denotes the mole fraction of the substance and the indices I and II indicate different phases or product streams. The extract product is obtained from the compressed gas by means of a change of state or by absorption/adsorption in a separator (b). The gas is then readjusted to the extraction conditions in the compressor (c) and recycled. Examples of such processes are the extraction of alkaloids [66], hops [67], and edible oils [68]; the deoiling of lecithin; and the extraction of flavors [69] from solids that are insoluble in the compressed gas. The extraction of lignin also takes place in a single step but involves a chemical reaction.

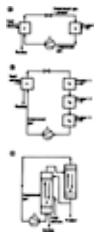

Figure 27. Schematic diagrams of gas extraction

A) Single-stage extraction; B) Multistage extraction; C) Continuous multistage contercurrent extraction

a) Extraction vessel; b) Separator, c) Compressor

If the separation factor is too small to permit single-stage extraction, *multistage extraction* can be used (Fig. 27 B). Either the extraction or separation conditions can be modified in several steps so that different components or mixtures can be consecutively extracted or separated. Examples are the refining of used oil [70], the fractionation of edible oils [71], and the deasphalting of oil [72]. In the refining of used oil, the fluid fraction of the oil is first separated from the solid contaminants with ethane. The pressure is then reduced and the temperature increased in several successive stages to separate fractions, according to either their pressure-dependent solubility in ethane or their volatility.

Supercritical gaseous solvents have also been studied for the production of synthetic fuels from coal, oil shale, and heavy oil [73]. Solvents such as supercritical toluene, tetralin, water vapor, and hydrogen can dissolve soluble components and pyrolysis products. The extract can then be fractionated by means of changes of state.

In the case of low separation factors, extraction may be performed as a *continuous multistage countercurrent process* (Fig. 28 C). This type of gas extraction is comparable to liquid – liquid extraction in countercurrent columns and can be treated analogously (⟶ Liquid–Liquid Extraction – Apparatus).

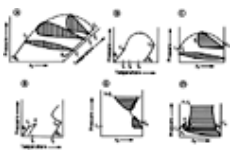

Figure 28. Phase diagrams of binary systems with a continuous critical curve (A – C) or an interrupted critical curve (D – F)

A_1= gaseous phase; A_2= liquid phase (rich in component 1); A_3= liquid phase (rich in component 2); B_1= triple point of component 1; B_2= triple point of component 2; C_1= critical point of component 1; C_2= critical point of component 2; K = critical end point of three-phase line (liquid – liquid – gas); x_2= mole fraction of component 2; ⇒vapor-pressure curve of pure component; – – – critical curve; hatching indicates two-phase areas

Examples of such separations are the laboratory-scale concentration of ethanol from dilute aqueous solutions [74] and the separation of monoglycerides from diglycerides [75].

Solvent Properties. The solvent capacity of compressed gases is primarily determined by their density. In the region of the critical temperature, small changes of temperature or pressure can be used to vary the density within wide limits. Close to the critical point, the latent heat of vaporization is very low; this means that separation of the extract from the solvent is simple and consumes very little energy. Furthermore, the properties of the compressed gas are favorable for transport, i.e., low viscosity, high diffusion coefficient, and high thermal conductivity (Table 4).

Table 4. Properties of solvents used for gas extraction

State of solvent	Density ϱ, g/cm^3	Viscosity η, g cm^{-1} s^{-1}	Diffusion coefficient D, cm^2/s
Gas (0.1 MPa, 25 °C)	$(0.6 – 2)\times10^{-3}$	$(1 – 3)\times10^{-4}$	0.1 – 0.4
Compressed gas			
p_{cr}, T_{cr}	0.2 – 0.5	$(1 – 3)\times10^{-4}$	0.7×10^{-3}
4 p_{cr}, ≈ T_{cr}	0.4 – 0.9	$(3 – 9)\times10^{-4}$	0.2×10^{-3}
Liquid (organic solvent, water at 0.1 MPa, 25 °C)	0.6 – 1.6	0.002 – 0.003	$(0.2 – 2)\times10^{-5}$

In general, the solvents used for gas extraction are either supercritical or contain a supercritical substance as their major component. A number of gases are available with critical temperatures below 100 °C:

ethylene	9.5 °C
carbon dioxide	31 °C
ethane	35 °C
nitrous oxide	36.5 °C
propane	96.8 °C

Mixtures of these gases yield solvents with intermediate critical temperatures. Relatively low concentrations (1 – 15 wt %) of polar substances, such as ethanol, can be used as entrainers. They cause specific interactions which improve the solvent capacity for polar substances considerably [76].

A compressed gas separates a mixture of substances primarily according to their molecular mass or volatility. Molecular interactions between the gas and the dissolved substances are of secondary importance. The gas solvents are generally only slightly polar; highly polar substances such as water are much less soluble in them than nonpolar substances. For a given solvent, the solubility decreases with increasing polarity of the solute. Another important factor affecting separation is the solubility of the gas solvent in the liquid phase formed. Incorporation of solvent molecules into the liquid phase causes considerable structural changes. These changes include a decrease in the attractive interactions between the molecules that are to be dissolved because of an increase in their mean separation.

4.2. Process Design

The most important data required for the design of a gas extraction (pressure, temperature, type and quantity of solvent, energy requirements) can be obtained from phase equilibria, mass transfer data, and a thermodynamic analysis of solvent circulation.

Phase Equilibria. The binary phase equilibria between a supercritical component (solvent) and the dissolved substances are the most important factors in gas extraction. The equilibria can usually be divided into two types, according to whether the critical curve is continuous (Fig. 28 A – C) or interrupted (Fig. 28 D – F). The equilibrium concentrations can be determined directly from the *p, x* curves at constant temperature. The solvent capacity of the gaseous solvent increases exponentially with pressure over a relatively wide range (Fig. 29). The solvent capacity of a gas generally increases with its density. Thus, at constant pressure, the solubility of a substance in a compressed gas can increase or decrease with increasing pressure, depending on whether the influence of the decrease in density or the increase in vapor pressure dominates.

Figure 29. Solubility of oleic acid in ethylene (v_i = molar volume of ethylene) as a function of temperature (⇒) or pressure (— — —)

a) 20 °C; b) 54.8 °C; c) 102.2 °C; d) 126.7 °C; e) 10 MPa; f) 15 MPa; g) 20 MPa; h) 25 MPa; i) 30 MPa

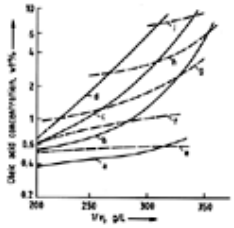

Although substances of low volatility are more soluble in compressed gases than is to be expected on the basis of their vapor pressure, their absolute concentrations are relatively low. For example, at 70 °C and 30 MPa the concentration of dissolved triglycerides (the main constituent of edible oils) in carbon dioxide only reaches 1 wt %. Solubility can be effectively increased to achieve complete miscibility by choosing an appropriate solvent (Fig. 30 A) or by adding an entrainer to the compressed gas (Fig. 30 B).

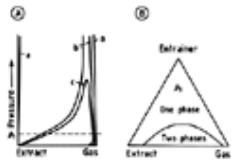

Figure 30. Solubility increase in solvent mixtures

A) Miscibility of extract and gas (solvent) mixtures as a function of pressure

Curves indicate phase boundaries for an extract which is (a) only slightly soluble, (b) moderately soluble, or (c) very soluble in the gas

B) Ternary phase diagram showing the increase in solubility produced by an entrainer at pressure p_1

Binary solubility of extract in gas according to case a, b, and c

The equilibrium solubility in fluid phases can be correlated over a wide range of conditions of state by using three-parameter equations of state of the Van der Waals type. Modified forms of the Redlich – Kwong equation and the equation of state according to Peng – Robinson are most commonly used. A general view of the phase behavior of a one-component system can be obtained from a small number of measurements by using these equations. To describe mixtures, mixing rules for binary mixtures are introduced which take into consideration the interactions between the two different types of molecules. These interaction parameters are determined by fitting with experimental equilibrium values. This method is inadequate for describing the critical curves or phase equilibria of associating molecules.

Mass Transfer. Mass transfer under the conditions used for gas extraction has not been as intensively investigated as phase equilibria. However, to a first approximation, it can be assumed that if the physical characteristics of the appropriate mixtures are used, the dimensionless correlations for liquid solvents are applicable. The flow rates of the compressed gas in the extraction vessel are ca. 10^{-2} m/s, whereas flow rates of 10 – 15 m/s are found in pipes because the viscosity is low. Extraction of solids is usually governed by the diffusion of solvent and solute within the solid. Accordingly, the extraction rate decreases exponentially with time (Fig. 31). If the material is comminuted (e.g., by grinding) and a high concentration of the substance to be extracted is present, this extraction phase may be preceded by a phase with a constant extraction rate. The resistance to transport in the solvent phase and the stationary mass transfer in a bulk solid can be calculated for a constant extraction rate with the following mass-transfer equation:

$$Sh = 2 + 1.1 Sc^{1/3} \cdot \mathrm{Re}^{0.6}\,;\; Sh = \left(\beta_a \cdot d\right) / D_G$$

where *Sh* is the Sherwood number, *Sc* the Schmidt number, *Re* the Reynolds number, β_a the outside mass-transfer coefficient, *d* the particle diameter, and D_G the diffusion coefficient of the gas.

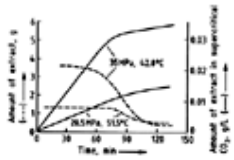

Figure 31. Time course of the extraction of rapeseed oil from prepressed rapeseed using supercritical carbon dioxide

Quasistationary and nonstationary approximations can be calculated for the nonstationary fraction (decreasing extraction rate) [77]. The quasistationary approximation is

$$\frac{c_m - c_\infty}{c_A - c_\infty} \approx \exp\left(\frac{-kA}{M_s c_A} \cdot t\right)$$

where

A = surface area for mass transfer

c_A = initial concentration of extract (solute) in solid

c_m = concentration of extract in the solid at time t

c_∞ = concentration of extract in the bulk solvent

k = mass-transfer coefficient
M_s = relative molecular mass of extract (solute)
t = time

The nonstationary approximation is

$$\frac{m}{m_c} = c_q[1 - \exp(-EFo)]$$

where

c_q = tabulated values of the approximation [78]
E = tabulated values of the approximation [78]
Fo = Fourier number
m = extract content of solid
m_c = initial extract content of solid

The entire process can be represented by a model analogous to that used for drying [79]. Measurements of mass transfer have been made on glycerides in continuous countercurrent extraction with fluid phases [80]. At a flow rate of 10 mm/s in an empty tube, two or three separation stages per meter are required for bubble-cap plates or a Sulzer packing, respectively.

The circulation of gaseous solvent depends primarily on whether separation of the extract is achieved by expansion or by absorption/adsorption. In absorption/adsorption the energy consumption for gas circulation can be calculated from the pressure loss of ca. 1 – 2 MPa at an extraction pressure of 30 MPa. The energy requirement for obtaining the product corresponds essentially to that needed to separate the adsorbent or absorbent from the extract.

The use of expansion for product separation can extend into the two-phase region. The subsequent pressure increase then occurs when the solvent is in the liquid state. However, expansion may also remain within the supercritical region, the subsequent pressure increase then results in gas compression (gas state) [81]. Both types of circulation are shown in Figure 32 in the form of temperature – entropy diagrams.

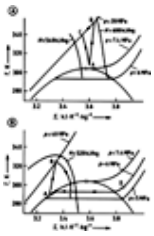

Figure 32. Solvent circulation in gas extraction (temperature – entropy) diagrams

Curves show values for constant pressure or enthalpy.

A) Compressor processA ⟶ B = compression; B ⟶ C = cooling; C = extraction; C ⟶ A = expansion; A = separation of extract/solvent;

B) Pump process

A ⟶ B = compression with pump; B = extraction; B ⟶ C = expansion; C ⟶ D = evaporation; D = separation of extract/solvent; D ⟶ A = condensation

4.3. Equipment

The temperatures and pressures used in gas extraction plants lie within the ranges commonly used in industry. Special features are the large amounts of circulating gas, the increasing strain imposed on the pressurized containers used in batch processes, and the feed devices. Various quick-sealing systems are used in solid extraction for the rapid and complete filling and emptying of the extraction vessel. Examples are shown in Figure 33. Devices such as screw presses and lock – hopper systems have been proposed for continuous feeding and product removal. Lock – hopper systems have been used successfully in industrial plants. However, screw presses still have to be implemented on an industrial scale as feeding devices because leakage of gas is difficult to avoid.

Figure 33. Examples of rapid-sealing devices for pressure containers used in gas extraction

A) Segmented-ring seal; B) Clamp seal; C) Uhde – Bretschneider lid seals

a) Pressure-vessel wall; b) Cover; c) Retaining ring composed of several segments; d) Ring for fixing retaining ring or cover; e) Sealing gasket

Courtesy of R. Eggers.

Commercial plants for solid extraction have so far only been used in the food industry and employ solely carbon dioxide as a solvent. The first plant was opened in 1981 by HAG (General Foods) in Bremen (Federal Republic of Germany) for the decaffeination of coffee (see also ⟶ Coffee – Decaffeination); a second plant followed in 1988 in Houston (United States). Plants for the production of hop extract are operated by Barth, Wolnzach, and SKW in Münchsmünster (Federal Republic of Germany). A plant for the decaffeination of black tea was installed in 1988 by SKW.

Hop extract is also produced with liquid carbon dioxide in a virtually critical state (⟶ Beer – Processing of Hops and Hop Products); such plants can be found, for example, in England and Australia. Critical liquefied gases have also been used as solvents to separate liquid mixtures. Examples include deasphalting with propane, removal of triglycerides from fish oils with propane in the Solexol process, and more recently, the residuum oil supercritical extraction (ROSE) process for extracting residual oil with butane and/or pentane (⟶ Asphalt and Bitumen – Solvent Separation Methods). In a process for methyl ethyl ketone used by Idemitsu (Japan), the intermediate product 2-butanol is extracted from an aqueous solution with 1-butene under supercritical conditions.

[Top of Page]

References

1. K. Schönemann, T. Voeste, “Zur Berechnung von Feststoff-Extractionsanlagen,” *Fette Seifen Anstrichm.* **54** (1952) 385.
2. R. Lüde: *Die Gewinnung von Fetten und fetten Ölen*, Gewinnung durch Extraktion, Steinkopff Dresden – Leipzig 1948.
3. R. N. Rickles, *Chem. Eng.* March 15 (1965) 157 – 172.
4. D. F. Boucher, J. C. Brier, J. O. Osborn, “Extraction of Oil from a Porous Solid,” *Trans. Am. Inst. Chem. Eng.* **38** (1942) 967.
5. G. Karnofsky, “The Mechanics of Solvent Extraction,” *J. Am. Oil Chem. Soc.* **26** (1949) 564.
6. H. B. Coats, G. Karnofsky, “Solvent Extraction II. the Soaking Theory of Extraction,” *J. Am. Oil Chem. Soc.* **27** (1950) 51.
7. D. F. Othmer, W. A. Jaatinen, *Ind. Eng. Chem.* **51** (1959) 543.
8. W. Depmer, “Überblick über die kontinuierlichen Extraktionsanlagen,” *Fette Seife* **53** (1951) 699 - 703.
9. J. F. Figlmüller, “Die quantitativen Grundlagen der Extraktionssysteme Fest-Flüssig,” *Oesterr. Chem. Ztg.* **56** (1955) 221.
10. G. Karnofsky, “Design of Oilseed Extractors,” *JAOCS J. Am. Oil Chem. Soc.* **63** (1986) 1011 – 1014.
11. J. O. Osburn, D. L. Katz, “Structure as a Application of Diffusion Theory to Extraction,” *Trans. A.I.Ch.E.* **40** (1944) 511.
12. F. H. Schneider, “Zur Mikrostruktur von Ölsaaten unter dem Aspekt der extraktiven Trennung,” *Fette Seifen Anstrichm.* **80** (1978) 265.
13. F. H. Schneider, “Zur extraktiven Lipidfreisetzung aus pflanzlichen Zellen,” *Fette Seifen Anstrichm.* **82** (1980) 16 – 23.
14. W. Kehse, “Fest-Flüssig-Extraction mit dem Karusell-Extrakteur,” *Chem. Ztg.* **94** (1970), 56 – 62.
15. K. Weber, “Entwicklungsstand und Tendenzen in der Fest-Flüssig-Extraktion,” *Maschinenmarkt MM Industriejournal* **80** (1974) 1733 – 1736.
16. Pliny: *The Natural History of Pliny*, translation by J. Bostock, H. T. Riley, vol. 5, George Bell and Sons, London 1900, pp. 503 – 504.
17. G. Agricola: *De Re Metallica*, translation by H. C. Hoover, L. H. Hoover, Dover Publication Inc., New York 1950, p. 443.
18. J. H. Taylor, P. F. Whelan, *Bull. Inst. Min. Metall.* **457** (1942) 1 – 36.
19. F. Habashi: *Principles of Extractive Metallurgy*, “Hydrometallurgy,” vol. **2**, Gordon and Breach Science Publishers, New York 1970, p. 25.
20. J. S. MacArthur, R. W. Forrest, W. Forrest, GB 14 174, 1887; US 403 202, 418 137, 1889.
21. M. E. Wadsworth in K. Osseo-Asare, J. D. Miller (eds.): *Hydrometallurgy-Research*, Development and Plant Practice, TMS-AIME, New York 1982, p. 8.
22. H. W. Sheffer, L. G. Evans: *Copper Leaching Practices in the Western United States*, U.S. Bureau of Mines, IC 8341 1968, p. 57.
23. G. J. C. Young: “A Review of Some of the Developments in the Gold Recovery Process Used in South Africa,” presentation at AIME Meeting, New York, Feb. 24 – 28, 1985. C. Kovacs-Figueroa, “Morenci Tailings Leach Plant,” *Trans. Soc. Min. Eng. AIME* **260** (1976) 170 – 172.
24. R. M. G. S. Berezowsky, D. R. Weir, *Minerals Metall Process.* **1** (1984) 1 – 4. D. R. Weir in J. F. Spisak, G. V. Jergensen (eds.): *Front. Technol. Miner. Process.*, SME-AIME, New York 1985, p. 14.
25. P. G. Mason, F. D. Wicks, J. C. Gathje, US 4 522 589, 1985.
26. W. M. Latimer: *The Oxidation States of the Elements and Their Potentials in Aqueous Solutions*, 2nd ed., Prentice-Hall, Englewood Cliffs, N.J. 1952.
27. R. C. Weast (ed.). *CRC Handbook of Chemistry and Physics*, CRC Press, Inc., Boca Raton, Florida, 1986, pp. B-68 – B-161.
28. W. F. Linke (ed.): *Solubilities – Inorganic and Metal Organic Compounds*, vol. **1**, D. van Nostrand Company, Inc., Princeton, N. J. 1958, vol. 2, Amer. Chem. Soc., Washington, D.C. 1965.
29. W. J. Schlitt, J. S. Jackson, *In Situ* **5**, (1981) no. 2, 103 – 131. Links
30. M. Pourbaix: *Atlas of Electrochemical Equilibria*, translation by J. A. Franklin, Pergamon Press, New York 1966. R. M. Garrels, C. L. Christ: *Solutions, Minerals, and Equilibria*, Harper and Row, New York 1965.
31. O. Levenspiel: *Chemical Reaction Engineering*, 2nd ed., J. Wiley and Sons, New York 1972, p. 361.
32. M. E. Wadsworth in H. Y. Sohn, M. E. Wadsworth: (eds.) *Rate Processes of Extractive Metallurgy,* Plenum Press, New York 1979, p. 137.
33. G. F. Kortum, J. O'M. Bockris, *Textbook of Electrochemistry*, vol. **2**, Elsevier, Amsterdam 1951, pp. 403 – 405.
34. S. Glasstone: *Textbook of Physical Chemistry*, 2nd ed., MacMillan and Co. Ltd., London 1948, p. 1090.
35. W. D. Specer, B. Topley, *J. Chem. Soc.* **50** (1929) 2633 – 2650.
36. C. Y. Wen: *Ind. Eng. Chem.* **60** (1968) 34 – 54.
37. M. E. Wadsworth in H. Y. Sohn, M. E. Wadsworth (eds.): *Rate Processes of Extractive Metallurgy*, Plenum Press, New York 1979, pp. 141 – 157.
38. J. B. Hiskey, M. E. Wadsworth in M. C. Kuhn (ed.): *Process and Fundamental Considerations of Selected Hydrometallurgical Systems*, SME-AIME, New York 1981, pp. 303 – 325.
39. V. Kudryk, H. H. Kellogg, *J. Met.* **6** (1954) 541 – 548.
40. C. R. S. Needs, M. J. Nicol, N. P. Finkelstein in A. R. Burkin (ed.): *Leaching and Reduction in Hydrometallurgy*, IMM, London 1975, pp. 12 – 19.
41. M. J. Nicol, C. R. S. Needes, N. P. Finkelstein: South African National Institute of Metallurgy Report PIN 171 (1973)

p. 37.

42. J. B. Hiskey. *IMM Trans. Sect. C,* **88** (1979) 145 – 152.
43. L. W. Beckstead, J. D. Miller, *Met. Trans. B*, 8 B (1977) 19 – 29.
44. V. K. Berry, L. E. Murr, J. B. Hiskey, *Hydrometallurgy* **3** (1978) p. 309 – 326.
45. J. V. Beck, *Biotechnol. Bioeng.* **9** (1967) 487 – 497. A. E. Torma in T. K. Ghose, A. Feichter, N. Blakebrough (eds.): *Biochemical Engineering*, Springer, Berlin 1977, p. 37.
46. L. C. Bryner, J. V. Beck, D. B. Davis, D. G. Wilson, *Ind. Eng. Chem.* **46** (1954) no 12, 2587–0. L. C. Bryner, R. Anderson, *Ind. Eng. Chem.* **49** (1957) no. 10, 1721–0.
47. E. E. Malouf, J. D. Prater, *J. Met.* no., **5** (1961) 353 –356.
48. L. E. Murr, V. K. Berry in J. C. Yannopoulos, J. C. Agarwal (eds.): *Extractive Metallurgy of Copper*, vol **2**, TMS-AIME, New York 1976, p. 671.
49. R. P. Hackl, F. Wright, A. Bruynesteyn: Presentation at “Biotechnology for Gold Industry Symposium,” 25th Annual Conference of Metallurgists, Toronto. Ontario, August 17 – 20, 1986, p. 21.
50. N. Tomizuka, M. Yagisawa in L. E. Murr, A. E. Torma, J. A. Brierley (eds.): *Metallurgical Application of Bacterial Leaching and Related Microbiological Phenomena*, Academic Press, New York 1978, p. 327.
51. C. L. Brierley, J. A. Brierley, *Can. J. Bacteriol.* **19** (1973) 183 – 188.
52. R. W. Lawrence, A. Bruynesteyn: *CIM Bull.* **76** (1983) no. 857, 107 – 110. Links
53. R. W. Lawrence, and J. D. Gunn in J. F. Spisak, G. V. Jergensen (eds.): *Front. Technol. Miner. Process.*, SME-AIME, New York 1985, pp. 13 – 17.
54. P. B. Marchant: *Presentation at Symposium on Complex Sulfides*, San Diego, California, November 10 – 13, 1985.
55. R. M. Nendick in J. B. Hiskey (ed.): *Au and Ag Heap and Dump Leaching Practice*, SME-AIME, New York 1984, pp. 85 – 91.
56. Philadelphia Mixers Technical Brochure: “Energy Conservation in Fluid Mixing,” Philadelphia Gear Corp., 1980, p. 15.
57. S. J. Hussey, H. B. Salisbury, G. M. Potter: U.S. Bureau of Mines RI 8368 (1979), p. 22.
58. P. A. Laxen: *Carbon-in-Pulp technology for the Extraction of Gold*, The Australasian Institute of Mining and Metallurgy, Victoria, Australia 1982, pp. 23 – 47.
59. E. G. Parker, D. R. McKay, H. Salomon-De-Friedberg in K. Osseo-Asare, J. D. Miller (eds.): *Hydrometallurgy-Research, Development and Plant Practice*, TMS-AIME, New York 1983, pp. 927 – 940.
60. E. G. Parker, *CIM Bull.* **74** (1981) no. 829, 145 – 150. Links
61. R. S. Kunter, J. R. Turney, R. D. Lear: Presentation at “International Precious Metals Symposium,” TMS-AIME, Los Angeles, California, February 26 –March 1, 1984.
62. A. E. Torma in L. E. Murr, A. E. Torma, J. A. Brierley (eds.): *Metallurgical Applications of Bacterial Leaching and Related Microbiological Phenomena*, Academic Press, New York 1978, pp. 375 – 387.
63. H. M. Tsuchiya in [47] , pp. 365 – 373.
64. R. O. McElroy, A. Bruynesteyn in [47] , pp. 441 – 462.
65. L. G. Randall, *Sep. Sci. Technol.* **17** (1982) no. 1, 1 – 18. M. E. Paulaitis, V. J. Krukonis, R. T. Kurnik, R. C. Rid, *Rev. Chem. Eng.* **1** (1983) no. 2, 179 – 250. Links
66. Studiengesellschaft Kohle, Mülheim, DE 2 005 293, 1970 (K. Zosel)HAG AG, DE 2 212 281 1972(O. Vitzthum, P. Hubert); DE 2 737 793, 1977 S. Peter, G. Brunner). E. Stahl, K. W. Quirin, A. Glatz, D. Gerard, G. Rau, *Ber. Bunsenges. Phys. Chem.* **88** (1984) 900 – 907.
67. O. Vitzthum, P. Hubert, W. Sirtl, DE 2 127 618, 1971.
68. R. Eggers, U. Sievers, W. Stein, *JAOCS J. Am. Oil Chem. Soc.* (1985) no. 8, 1222 – 1230. G. Brunner, *Fette Seifen Anstrichm.* **88** (1986) no. 12, 464 – 474.
69. O. G. Vitzthum, P. Hubert, *Angew. Chem. Int. Ed. Engl.* **17** (1978) 710 – 715. H. Coenen, E. Kriegel, *Chem. Ing. Tech.* **55** (1983) 890 – 891.
70. H. Coenen, P. Rinza, *Tech. Mitt. Krupp, Werksber.* **9** (1981) no. 1, 21 – 31.
71. K. Zosel, DE-AS 2 322 038, 1973. H. Coenen, E. Kriegel, DE-OS 2 443 920, 1978.
72. S. R. Nelson, R. G. Roodman, *Chem. Eng. Prog.*, May 1985, 63 – 68.
73. A. Wilhelm, K. Hedden, *Erdöl Kohle Erdgas Petrochem.* **33** (1983) 269 – 274. R. Hoffmann, K. Künstle, G. Brunner, *Prepri. Int. Symp. High Pressure Chemical Engineering*, Erlangen 1984, 155 – 162.
74. G. Brunner, K. Kreim, *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) 246 – 250.
75. S. Peter, G. Brunner, R. Riha, *Fette Seifen Anstrichm.* **78** (1976) 45 – 50; US 4 345 976, 1982.
76. G. Brunner, in J. A. Marinsky, Y. Marcus, (eds.): *Ion Exchange and Solvent Extraction*, vol. **10**, Marcel Dekker, New York 1988, pp. 105 – 140.
77. G. Brunner, in J. M. L. Penninger, M. Radosz, M. A. McHugh, V. J. Krukonis, (eds.): *Supercritical Fluid Technology*, Elsevier, Amsterdam 1985, pp. 245 – 261.
78. U. Grigull, *Temperaturausgleich in einfachen Körpern*, Springer, Heidelberg 1964.
79. E. A. Lack, Dissertation, TU Graz, 1985.
80. H. Kohmann, Dissertation, Univ. Erlangen-Nürnberg, 1981. C. Tiegs, S. Peter, *Fette Seifen Anstrichm.* **87** (1985) 231 – 235.
81. R. Eggers, *Angew. Chem. Int. Ed. Engl.* **17** (1978) 751 – 754.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Absorption

JOHANN SCHLAUER, Lurgi GmbH, Frankfurt, Federal Republic of Germany (Chaps. 1 – 2)

MANFRED KRIEBEL, Lurgi GmbH, Frankfurt, Federal Republic of Germany (Chaps. 3 – 5)

1. Equilibrium of Gas Solubility 3
1.1. Introduction 3
1.2. Gas-Phase Fugacity 4
1.3. Liquid-Phase Fugacity and Activity . 5
1.4. Physical Absorption 6
1.5. Chemical Absorption 8
1.6. Enthalpy and Absorption Equilibrium 13
2. Mass Transfer 14
2.1. Introduction 14
2.2. Mass Transfer Coefficients 15
2.3. Effect of Solute Concentration 16
2.4. Correlation of Mass Transfer Coefficients 17
2.5. Mass Transfer and Chemical Reaction 19
2.6. Modeling of Mass Transfer with Chemical Reaction 21
3. Design of Absorption Systems 21
3.1. Methods of Absorption 21
3.2. Methods of Desorption 22
3.3. Absorption and Desorption Equipment 23
3.4. Columns with Mass-Transfer Plates . 23
3.5. Columns with Random Packing 23
3.6. Columns with Structured Packing . . 24
3.7. Columns with Special Internals 24
4. Design of Absorption Equipment . . . 25
4.1. Physical Absorption in Plate Columns 25
4.2. Chemisorption in Packed Columns . . 28
4.3. Nonisothermal Absorption 29
5. Design of Desorption Equipment . . . 30
5.1. Physical Desorption in Flash Columns 30
5.2. Physical Desorption in Stripper Columns 31
5.3. Physical Desorption in Reboiler Columns 33
5.4. Chemical Desorption in Reboiler Columns 34
6. References 35

Symbols

a effective interfacial mass transfer area per unit volume of tower or apparatus, m^2/m^3
A absorption factor
A concentration of alkali or amine, $kmol/m^3$
B parameter
[B] bulk-liquid reactant concentration, $kmol/m^3$
c concentration, $kmol/m^3$
c^b solute concentration in bulk liquid, $kmol/m^3$
c^i liquid solute concentration at gas – liquid interface, $kmol/m^3$
c_k liquid solute concentration (head), $kmol/m^3$
c_p specific heat capacity, $kJ\ mol^{-1}\ K^{-1}$
c_s liquid solute concentration (sump), $kmol/m^3$
C dimensionless constant
d characteristic length appropriate to the geometry of the system under consideration, m
D desorption factor
D_{AL} diffusion coefficient of reactant A in liquid phase, m^2/s
D_{AB} gas-phase diffusion coefficient of solute A in inert gas B, m^2/s
D_B diffusion coefficient of reactant B in liquid phase, m^2/s
D_L diffusion coefficient of inert solute in liquid phase, m^2/s
E enhancement factor
E^A reaction activation energy, kJ/mol
E_A absorption rate
f fugacity, kPa
F reaction frequency factor, $m^3\ kmol^{-1}\ s^{-1}$
$F(\ldots)$ function of . . .
ΔG molar Gibbs energy, kJ/mol
G^E molar excess Gibbs energy, kJ/mol
G gas-phase flow rate, m^3/h
G_M gas-phase mass velocity, $kg\ s^{-1}\ m^{-2}$
h height of packed column, m
ΔH molar enthalpy, kJ/mol
H Henry's law constant, kPa
H_i^c Henry's law constant of solute i (molarity based), $kPa\ m^3\ kmol^{-1}$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_08

H_i^{m} Henry's law constant of solute i (molality based), kPa kg mol^{-1}
H_{c} Henry's law constant of carbon dioxide in water, kPa m^3 kmol^{-1}
$H_{i,\,\mathrm{L}}$ Henry's law constant of solute i in solvent L, kPa
H_{s} Henry's law constant of hydrogen sulfide in water, kPa m^3 kmol^{-1}
Ha Hatta number
j_{M} Chilton – Colburn j factor for mass transfer
k equilibrium constant, kmol m^{-3} bar^{-1}
k' equilibrium constant, mol m^{-3} Pa^{-1}
k_{c} rate constant of (pseudo) first-order reaction, s^{-1}
k_{g} gas-phase mass transfer coefficient, m/s
k_{G} gas-phase mass transfer coefficient, kmol s^{-1} m^{-2}
$k_{\mathrm{G}}\,a$ volumetric gas-phase mass transfer coefficient, kmol s^{-1} m^{-2}
k_{l} liquid-phase mass transfer coefficient, m/s
k_{l}' liquid-phase mass transfer coefficient with reaction, m/s
k_{L} liquid-phase mass transfer coefficient, kmol s^{-1} m^{-2}
$k_{\mathrm{L}}a$ volumetric liquid-phase mass transfer coefficient, kmol s^{-1} m^{-3}
k_{11} rate constant of a second-order reaction, m^3 s^{-1} kmol^{-1}
K vapor – liquid equilibrium K value ($\mathrm{d}y*/\mathrm{d}x$)
K_{a} dissociation constant of amine, kmol/m^3
K_b dissociation constant of bicarbonate, kmol/m^3
K_{c} dissociation constant of carbamate, kmol/m^3
K_{G} overall gas-phase mass transfer coefficient, kmol s^{-1} m^{-2}
$K_{\mathrm{G}}a$ overall volumetric gas-phase mass transfer coefficient, kmol s^{-1} m^{-3}
K_h dissociation constant of hydrogen sulfide, kmol/m^3
K_{k} dissociation constant of carbonate, kmol/m^3
K_{L} overall liquid-phase mass transfer coefficient, kmol s^{-1} m^{-2}
$K_{\mathrm{L}}a$ overall volumetric liquid-phase mass transfer coefficient, kmol s^{-1} m^{-3}
K_s dissociation constant of sulfide, kmol/m^3
K_w ion product of water, kmol^2/m^6
L liquid-phase flow rate, m^3/h
L_{M} liquid-phase mass velocity, kg s^{-1} m^{-2}
m mass, kg
m slope of equilibrium curve ($\mathrm{d}y*/\mathrm{d}x$)
m_i molality, mol/kg
M_r molecular mass, kg/kmol
n_i moles of solute i, mol
n_{th} theoretical plate number
n number of transfer units
N_A mass transfer rate of solute A per unit interfacial area, kmol s^{-1} m^{-2}
N_{F} number of transfer units in the liquid phase
N_{G} number of transfer units in the gas phase
p total system pressure, kPa
$p(\ldots)$ solute partial pressure in bulk gas, kPa
p^{r} reference pressure, kPa
p_{A}^0 pure-component vapor pressure of solute A, kPa
p_{c} critical pressure, kPa
p_{i} solute partial pressure, kPa
Pr Prandtl number ($c_p\mu/k$)
Q enthalpy, kJ
r volumetric reaction rate, kmol s^{-1} m^{-3}
R gas constant, 8.314 J mol^{-1} K^{-1}, 8.314 m^3 kPa kmol^{-1} K^1
Re Reynolds number ($G_{\mathrm{M}}\,d/\mu_G$)
s volume of liquid on plate divided by total area of plate, cm
S_{s} sulfur dioxide concentration, mol/100 mol H_2O
Sc Schmidt number ($\mu_{\mathrm{G}}/\varrho_{\mathrm{G}}D_{\mathrm{AB}}$ or $\mu_{\mathrm{L}}/\varrho_{\mathrm{L}}D_{\mathrm{L}}$)
Sh Sherwood number ($k_G RTd/D_{\mathrm{AB}}p$)
St Stanton number ($k_{\mathrm{G}}M_{\mathrm{G}}/G_{\mathrm{M}}$ or $k_{\mathrm{L}}M_{\mathrm{L}}/L_{\mathrm{M}}$)
t temperature, °C
t time, s
T temperature, K
T_{c} critical temperature, K
u superficial gas velocity, cm/s
v_i^{G} molar volume of solute i in gas phase, m^3/kmol
v_i^{L} molar volume of solute i in liquid phase, m^3/kmol
v_i^0 molar volume of solute i in gas phase at 0 °C, 101.3 kPa, m^3/kmol
V volume of gas phase, m^3
V_i volume of solute i in gas phase, m^3
V_i^0 volume of solute i in gas phase at 0 °C, 101.3 kPa, m^3
V_{L} volume of solute-free solvent, m^3
x mole fraction in liquid phase

$x*$ mole fraction in equilibrium with bulk mole fraction y
X_c moles of carbon dioxide in liquid per mole of alkali or amine
X_s moles of hydrogen sulfide in liquid per mole of alkali or amine
y mole fraction of solute in gas phase
$y*$ mole fraction in equilibrium with bulk mole fraction x
z distance from interface, m
z compressibility factor

Greek Symbols

α Bunsen solubility coefficient (old dimension, $m^3\,m^{-3}\,atm^{-1}$)
β Ostwald solubility coefficient (old dimension, $m^3\,m^{-3}\,atm^{-1}$)
γ activity coefficient
δ film thickness, m
ε void fraction of packing available for gas flow
λ solubility coefficient (SI-units), $mol\,kg^{-1}\,kPa^{-1}$
μ viscosity, $kg\,m^{-1}\,s^{-1}$
ν specific solvent demand (L/G)
ϱ density, kg/m^3
σ relative supersaturation
φ percentage of absorption
Φ fugacity coefficient
ω acentric factor

Superscripts

$*$ equilibrium with the second phase
b bulk
G gas phase
i interface
L liquid phase
0 pure component

Subscripts

A, B components
A, D absorption, desorption
c critical
g gas phase
G gas phase
h head
i component
l liquid phase
L liquid phase
m mean
r reduced condition
s sump

1. Equilibrium of Gas Solubility

1.1. Introduction

Absorption is the uptake of gases by a solvent. Adsorption means accumulation on the surface of a solid (→ Adsorption). The equilibrium of solubility describes the distribution of absorbed materials between the vapor and liquid phases (VLE i.e., vapor – liquid equilibrium). Depending on its volatility, the solvent may appear in the vapor phase; this phenomenon will only be mentioned tangentially (see Section 5.3 and 5.4).

The interaction of absorbed materials with the solvent can be physical or chemical in nature (van der Waals forces, dissociation, neutralization, oxidation). In *physical absorption*, the gas molecules are polarized but remain otherwise unchanged. In *chemical absorption*, they are also chemically converted. The type of interaction critically affects the solubility equilibria and the shape of the absorption isotherms.

Chemical absorption entails greater changes in enthalpy than physical absorption, in which the enthalpy change is about equivalent to the enthalpy of condensation. Depending on which interaction predominates, the terms physical or chemical solubility, and even physical or chemical solvent, are used. The latter designation can be misleading because one and the same solvent exhibits chemical solubility for some gases (e.g., carbon dioxide or hydrogen sulfide) and physical solubility for others (e.g., hydrogen or methane).

Solubility equilibria are governed by the general thermodynamic condition that the entropy is at its maximum in a closed system. Thus, at equilibrium the molar Gibbs energy G is

$$\delta\left(\sum_i n_i G\right)/\delta n_i = 0 \quad [T,p = \text{const.}] \tag{1}$$

in all phases, and the fugacity is

$$f^{G} = f^{L} \text{ and } f_i^{G} = f_i^{L} \tag{2}$$

in the gas and liquid phase. This chapter deals first with the fugacity of the gas in the gas and

liquid phase, as well as with the activity of the liquid phase. The laws governing physical and chemical solubility equilibria are then discussed.

1.2. Gas-Phase Fugacity

The fugacity f_i of component i in the gas phase is defined by

$$dG_i = RT\,\mathrm{d}\ln f_i = v_i^{\mathrm{G}}\,\mathrm{d}p \tag{3}$$

and a boundary condition

$$f_i = y_i p, \text{ at } p \to 0 \tag{4}$$

Fugacity cannot be measured directly but is determined from an equation of state that describes the gas mixture in question.

$$\ln\frac{f}{p} = \int_0^p \left(\frac{V^{\mathrm{G}}}{RT} - \frac{1}{p}\right)\mathrm{d}p = \int_0^p \frac{z-1}{p}\,\mathrm{d}p \tag{5}$$

To solve this, either the molar volume v^{G} of the gas mixture from the equation of state or the compressibility factor z is substituted. The fugacity of component i is calculated from

$$\ln f_i/y_i = \ln f - \sum_{k\neq i} y_k\,(\delta \ln f/\delta y_k) \tag{6}$$

In the partial derivation, all components except the kth and ith are kept constant. Equations (5) and (6) have been solved and published for many equations of state. Solutions for the van der Waals equation, the virial equation with one or two constants, the cubic (Redlich – Kwong, Soave and Peng – Robinson), and the complex [Benedict – Webb – Rubin (BWR) and BWR – Starling] equations of state will be found in [1].

The generalized method of calculation according to Hayden and O'Connell, based on the virial equation, appears in [2]. All thermodynamic data and many binary interaction parameters required for the calculation are listed there for many gases. While the applicability of this method is limited, the calculations are not too time-consuming. The method is not suitable for the critical region, and its application is confined to the following pressures:

$$p/p_{\mathrm{c}} \leq 0.5T/T_{\mathrm{c}} \tag{7}$$

For gas mixtures, a weighted average is used for the critical pressure p_{c} and the critical temperature T_{c}

$$p_{\mathrm{c}} = \sum_i y_i p_{c,i} \text{ and } T_{\mathrm{c}} = \sum_i y_i T_{c,i} \tag{8}$$

In the model considered most efficient at the moment, the fugacity is calculated from two terms [3]:

$$\log(f/p) = [\log(f/p)]^{(0)} + \omega\,[\log(f/p)]^{(1)} \tag{9}$$

where ω is the component-specific acentric factor. The logarithmic terms in square brackets have been published as a function of the reduced temperature and the reduced pressure [1], [4]. Unfortunately, this method can be used only for single-component gases; computer programs are required to calculate the fugacity in gas mixtures. A modification of the method has been published along with the corresponding rules of mixing [5]. For polar components, an additional polar factor is introduced.

Simulation programs calculate the fugacity by several methods. The user can select a suitable method but must be mindful of the areas of application to avoid gross errors.

The *virial equation* offers a simple method of calculation and is very popular. However, it is not suitable in the critical region or for liquids. Outside the critical region, predicted compressibility factors are reproduced with deviation <2 % for pure components <3 % for weakly polar mixtures, and <5 % for highly polar mixtures. This accuracy can be achieved in the following region of application:

$$p_{\mathrm{r}} < T_{\mathrm{r}} - 0.65 \text{ for } T_{\mathrm{r}} < 1 \tag{10}$$

$$p_{\mathrm{r}} < 1.05T_{\mathrm{r}} - 0.35 \text{ for } T_{\mathrm{r}} > 1 \tag{11}$$

For the critical data of gas mixtures weighted averages according to Equation (8) are used for the reduced pressure ($p_{\mathrm{r}} = p/p_{\mathrm{c}}$) and for the reduced temperature ($T_{\mathrm{r}} = T/T_{\mathrm{c}}$). Among the *cubic equations* of state, Soave and Peng – Robinson are widely employed. They can also be used sucessfully in the critical region, but binary interaction parameters are required for gas mixtures. Benedict – Webb – Rubin is suitable for hydrocarbons, and Lee – Kesler – Plöcker is recommended for broad ranges of pressure and temperature ($0 < p_{\mathrm{r}} < 10$ and $0.3 < T_{\mathrm{r}} < 4$) [4].

In the commercial programs Process, Pro II, and Aspen, one method is offered as default for the fugacity calculation, and the user can select another. If the correct method is in doubt, the selected method must be checked against experimental data in the area of interest.

1.3. Liquid-Phase Fugacity and Activity

The *fugacity of dissolved component i* in the liquid phase is given by

$$f_i^{\mathrm{L}} = \gamma_i x_{\mathrm{i}} f_i^0 \exp \int_{p^{\mathrm{r}}}^{p} \left(v_i^{\mathrm{L}} / RT \right) \mathrm{d}p \tag{12}$$

The exponential term represents the effect of gas-phase pressure and is called the *Poynting factor*. Here, v_i is the molar volume of component i in the liquid phase, and the integration is carried out from the reference pressure p^{r} to the pressure of the system p. The *activity coefficients* γ_i are needed to correct the real liquid phase; they are related to the molar excess Gibbs energy G^{E} as follows:

$$G^{\mathrm{E}} = RT \sum_i x_i \ln \gamma_i \tag{13}$$

There are many G^{E} models for calculating the activity coefficients γ_i in the liquid phase. A key advance was made by Wilson in 1964, who introduced the local concentration that led to the development of the powerful NRTL (nonrandom two liquid), ASOG (analytical solution of groups), UNIQUAC (universal quasi-chemical equation) and UNIFAC (UNIQUAC functional group activity coefficient) models [6–9].

The following two conventions are preferred for normalizing the activity coefficients γ_i and the standard fugacity f_i^0:

Pure Liquid Component *i*. The activity coefficients γ_i are

$$\gamma_i = 1 \text{ if } x_i \to 1 \tag{14}$$

and the standard fugacity is

$$f_i^0 = p_i^0 \varphi_i^0 \exp \left[\left(p^{\mathrm{r}} - p_i^0 \right) \bar{v}_i^{\mathrm{L}} / RT \right] \tag{15}$$

where $\bar{v}_i^{\mathrm{L}}$ is the average molar volume of the dissolved component i in the liquid phase, p_i^0 is the saturated vapor pressure of the pure component i at temperature T and the fugacity coefficient of this vapor $\varphi_i^0 = f/p$ is calculated according to Equation (5). This *symmetrical normalization* can be used as long as the pure component i forms a liquid, hence, in the temperature range of its triple point up to the critical temperature. At low to moderate pressure,

$$\varphi_i^0 = 1, \text{ so that } f_i^0 = p_i^0 \tag{16}$$

and the Poynting factor $\to 1$ so that Raoult's law follows from Equations (12) – (15):

$$p_i = x_i p_i^0 \ (\text{if } x_i \to 1) \tag{17}$$

The standard fugacities of many liquids have been determined for the reference pressure $p^{\mathrm{r}} = 0$ and generalized by using the theorem of corresponding states as a function of the reduced temperature T_{r} and the acentric factor ω.

$$f_i^0 / p_{c,i} = F\left(T_{\mathrm{r},i}, \omega_i\right) \tag{18}$$

By extrapolation of Equation (18), the utility of symmetric normalization has been expanded beyond the critical temperature [2].

Infinitely Dilute Solutions. The activity coefficients are chosen as

$$\gamma_i = 1 \text{ if } x_i \to 0 \tag{19}$$

and the standard fugacity as

$$f_i^0 = H_{i,\mathrm{L}} \tag{20}$$

$H_{i,\mathrm{L}}$ is Henry's law constant of component i in the solvent L because, at low pressure, Equation (12) becomes Henry's law

$$p_i = x_i H_{i,\mathrm{L}} \ \ (x_i \to 0) \tag{21}$$

For gaseous component i in solvent L, the Henry's law constant depends only on temperature and can be determined from the solubility equilibria (see Section 1.4). No temperature limit exists to this so-called unsymmetric normalization because for gases, Henry's law constant can also be measured in the supercritical region. Henry's law constant is determined from solubility data. To do so, the ratio f_i/x_i is plotted vs. x_i and extrapolated to $x \to 0$. Henry's law

constant holds exactly only if the total pressure $p = p_L^0$ = vapor pressure of the solvent.

At higher pressure of the dissolved component, Henry's law constant is corrected similar to the Poynting factor according to the Krichevsky – Kasarnovski equation.

$$\ln H_{i,\mathrm{L}}(p) = \ln H_{i,\mathrm{L}}\left(p_{\mathrm{L}}^0\right) + v_i^\infty\left(p - p_{\mathrm{L}}^0\right)/RT \quad (22)$$

At infinite dilution the molar volume v_i^∞, is assumed to be constant over the entire range of pressure. The Krichevski – Illinskaya equation applies to a larger range of concentration:

$$\ln H_{i,\mathrm{L}}(p) = \ln H_{i,\mathrm{L}}\left(p_{\mathrm{L}}^0\right) + v_i^\infty\left(p - p_{\mathrm{L}}^0\right)/RT + C\left(x_{\mathrm{L}}^2 - 1\right) \quad (23)$$

The constant C is fitted to the experimental data. Henry's law applies at low loading and is used frequently. Raoult's law is used only rarely, for example, in the region of condensation if loading becomes extremely high.

Both laws are obeyed over the entire range of concentration $0 < x_i < 1$ only by ideal solutions. The deviation from Henry's and Raoult's laws in real solutions between the boundary values is demonstrated in Figure 1 by the solubility of carbon dioxide in methanol and in Figure 2 by the solubility of hydrogen sulfide in N-methylpyrrolidone. The ratio p_i/x_i is shown in the figures as a function of the mole fraction x_i for several temperatures; in real solutions, considerable deviations from ideality occur.

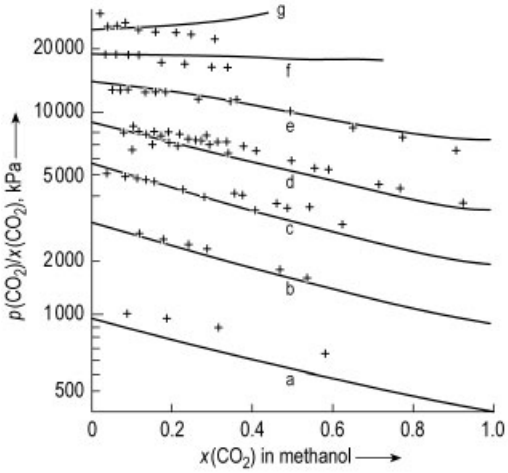

Figure 1. Solubility of carbon dioxide in methanol
Data from [14]; lines calculated with UNIQUAC
a) $t = -60\,°C$; b) $t = -40\,°C$; c) $t = -20\,°C$; d) $t = 0\,°C$; e) $t = 25\,°C$; f) $t = 50\,°C$; g) $t = 75\,°C$

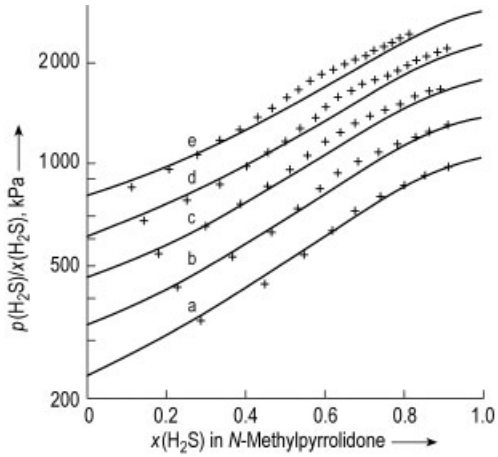

Figure 2. Solubility of hydrogen sulfide in N-methylpyrrolidone
Data from [10]; lines calculated with UNIQUAC
a) $t = 0\,°C$; b) $t = 10\,°C$; c) $t = 20\,°C$; d) $t = 30\,°C$; e) $t = 40\,°C$

The calculated lines show that measured values can be reproduced quite accurately once both phases have been corrected in the following manner.

The fugacity of the gas phase is determined from the virial equation, and the activity coefficients in the liquid phase are calculated with UNIQUAC; the temperature-dependent binary UNIQUAC parameters are then fitted to the measured data (for details, see [2]).

1.4. Physical Absorption

In physical absorption, the concentration of dissolved gases in the solvent increases in proportion to the partial pressure of these gases; i.e., it follows Henry's law (Eq. 21).

Henry's law constant $H_{i,\,\mathrm{L}}$ has the same dimensions as the vapor pressure p_i^0 and is numerically equal to it in ideal solutions. In real solutions, the numerical value of Henry's law constant depends on the solvent. While Henry's law constant has physical meaning even at supercritical temperatures, the vapor pressure exists only in the subcritical region. In the literature, Henry's law constant is often specified in different units if the concentration of the dissolved gas is not expressed as the mole fraction x_i. Some examples are given in Table 1.

Table 1. Different expressing of Henry's law constant

Constant	Dimension	Concentration
$H_{i,\mathrm{L}}=p_i/x_i$	bar	mole fraction
$H_i^{\mathrm{m}}=p_i/m_i$	bar kg mol^{-1}	molality
$H_i^{\mathrm{c}}=p_i/c_i$	bar m^3 kmol^{-1}	molarity

The following relations are used to recalculate the various dimensions:

$$H_{i,\mathrm{L}}/H_i^{\mathrm{m}}=1000/M_{\mathrm{L}}+\sum_j m_j \tag{24}$$

$$H_{i,\mathrm{L}}/H_i^{\mathrm{c}}=\varrho/M_{\mathrm{L}}+\sum_j c_j\,(1-M_j/M_{\mathrm{L}}) \tag{25}$$

where M is the molar mass in g/mol; m is the molality in mol/kg; c is the molarity in kmol/m^3; and ϱ is the density of the solution in kg/m^3. Indices represent the following: L solvent; i dissolved gaseous component; and j all dissolved components.

At low loading, the concentration-dependent contribution can be neglected and the recalculation can be limited to merely the first term of Equations (24) and (25).

A lot of solubility data have been published in the literature as *solubility coefficients*, defined as the ratio of the amount of gas dissolved at the partial pressure $p_i = 1$ to the amount of solvent. Many solubility coefficients differ in the dimensions used to describe the amounts of gas and solvent or the pressure. Some examples are given in Table 2.

Table 2. Examples of different solubility coefficients

Name	Definition	Dimension
Bunsen	$\alpha_i = V^0{}_i/V_{\mathrm{L}}p_i$	m^3 m^{-3} atm^{-1}
Ostwald	$\beta_i = V_i/V_{\mathrm{L}}\,p_i$	m^3 m^{-3} atm^{-1}
SI	$\lambda_i = n_i/m_{\mathrm{L}}p_i$	mol kg^{-1} bar^{-1}

∗ Conversion factor: 1 atm = 101.325 kPa = 1.013 bar.

In the Bunsen solubility coefficient, the gas volume is given at standard conditions (0 °C and 1 atm); in the Ostwald coefficient, at the measuring temperature. The value m_{L} designates the solvent mass.

A review of the old solubility coefficients with appropriate conversion factors can be found in [11]. Because Henry's law constants are defined in terms of mole fraction and the solubility coefficients in terms of mass ratio, the conversion of solubility coefficients to Henry's law constants depends on the loading or on the partial pressure:

$$H_{i,\mathrm{L}}=\mathrm{factor}/x_{\mathrm{L}}=\mathrm{factor}+p_i \tag{26}$$

where $x_{\mathrm{L}}=1-\sum_i x_i=\mathrm{solvent\,mole\,fraction}$. Factors for converting the most common solubility coefficients to Henry's law constants are listed in Table 3. At low gas loading, $\sum x_i \to 0$, and $x_{\mathrm{L}} \to 1$ or $p_i \to 0$, so that the conversion is independent of the concentration of dissolved gas.

Table 3. Factors for converting solubility coefficients to Henry's law constants

Solubility coefficient	Factor ∗
α_i Bunsen	$1.01325\ v_i^0\,\varrho_{\mathrm{L}}/M_{\mathrm{L}}\ \alpha_i$
β_i Ostwald	$T v^0{}_i\,\varrho_{\mathrm{L}}/269.58\,M_{\mathrm{L}}\ \beta_i$
λ_i SI	$1000/\mathrm{M}_{\mathrm{L}}\ \lambda_i$

∗ $H_{i\ \mathrm{L}}$ is the Henry's law constant of gas i in solvent L in bar; v_i^0 is the molar volume of gas i under standard conditions (0 °C and 101.35 kPa) in m^3/kmol; ϱ_{L} is the density of gas-free solvent in kg/m^3; and T is the temperature in K.

In the design of natural gas and petroleum refineries, the so-called K value

$$K_i=y_i/x_i \tag{27}$$

is often used for the equilibrium of gas solubility. The K value is particularly useful for representing and calculating the solubility equilibria of hydrocarbons (e.g., methane or ethane) in petroleum or gasoline fractions. In these chemically related materials the liquid phase behaves ideally (Raoult's law holds) so that the K values do not depend on the composition of the solvent but only on the temperature and total pressure. The Henry's law constant of component i is calculated from the K value and the pressure

$$H_i=K_i p \tag{28}$$

For an ideal liquid phase, the K values and Henry's law constants are independent of the solvent so the index $_{\mathrm{L}}$ is omitted. Solubility data of gases in liquids, as well as overviews of published solubility equilibria, can be found in [1], [2], [4], [11], and [12].

The IUPAC Solubility Data Series publishes critically selected measured values of gas solubilities [13]; 100 volumes are planned in all.

Many of the published solubility equilibria are stored in data banks [e.g., the Dortmund Data

Bank (DDB), the Berlin Thermodynamic Data Bank (BDBT)] or in Dechema Data Series; they have been published in part or can be accessed via commercial data markets.

A data reduction by the G^E model was illustrated in Section 1.3 for the systems carbon dioxide – methanol and hydrogen sulfide – *N*-methylpyrrolidone. The systems were correlated by using the Peng – Robinson or the Redlich – Kwong – Soave equation of state with the temperature-dependent binary interaction parameters determined by fitting experimental data [14], [15], too.

To represent the solubility of gases in water, in concentrated aqueous solutions of electrolytes, and in hydrocarbons, the hard-sphere model and the Lennard – Jones potential 6 – 12 are used with the dipole – dipole and dipole – induced dipole contributions as the intermolecular interactions [16]. An overview of the older models appears in [11].

1.5. Chemical Absorption

Chemical absorption differs from physical absorption mainly in the higher enthalpy of absorption and the nonlinearity of the absorption isotherms. The high absorption enthalpy is due to chemical reactions of the absorbed gases in the solvents. This is also the reason for the greater temperature dependence of chemical solubility (see Section 1.6). The enthalpy of chemical absorption decreases much more rapidly with increased loading than the enthalpy of physical absorption. Especially after consumption of the chemical capacity (e.g., if amines in the solvent are neutralized by the absorbed acid gases), the chemical enthalpy of absorption decreases precipitously, approaching the physical absorption enthalpy.

The chemical absorption isotherm is nonlinear and Henry's law does not hold. At low loading, the relation $\Delta p/\Delta(\text{load})$ of the dissolved gas increases with increasing loading.

Several isotherms have been proposed for quantitative representation of chemical solubility equilibria; however, they do not have general validity but describe only certain systems. The most important ones are listed in the following material.

Carbon Dioxide in Alkali Carbonate – Bicarbonate. An absorption isotherm for carbon dioxide in an aqueous solution of alkali carbonate – bicarbonate was derived by MacCoy [17] as early as 1903 from the mass action law of the following reactions:

Reaction	Equilibrium	
$CO_2(g) \rightleftharpoons CO_2(l)$	$p(CO_2) = H_c[CO_2]$	(29)
$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$	$[CO_2]K_b = [HCO_3^-][H^+]$	(30)
$HCO_3^- \rightleftharpoons CO_3^{2-} + H^+$	$[HCO_3^-]K_k = [CO_3^{2-}][H^+]$	(31)

The formulas in square brackets denote the concentration of the reaction partners; H_c is Henry's law constant for carbon dioxide in water; and K_b and K_k are the dissociation constants of bicarbonate and carbonate, respectively. Thus, from Equations (29) – (31)

$$p(CO_2) = \frac{H_c K_k \left[HCO_3^-\right]^2}{K_b \left[CO_3^{2-}\right]} \tag{32}$$

The concentration of alkali A (alkalinity) follows from the condition of electroneutrality

$$A = \left[HCO_3^-\right] + 2\left[CO_3^{2-}\right] \tag{33}$$

If a specific loading is introduced,

$$X_c = \left[HCO_3^-\right]/A \tag{34}$$

from Equations (32) – (34) follows,

$$p(CO_2) = \frac{H_c K_k 2AX_c^2}{K_b(1-X_c)} = \frac{1}{k} \cdot \frac{2AX_c^2}{1-X_c} \tag{35}$$

The new equilibrium constant k combines all previous constants; like Henry's law constant and the dissociation constant, it is temperature dependent. This temperature dependence is represented by the following empirical formula in the temperature range from 0 to 190 °C [11]:

$$\log k = 4990/(558+t) - 6.265 \tag{36}$$

The constant k is calculated in kmol m^{-3} bar^{-1} if the temperature t is in degrees Celsius. The absorption isotherm (Eq. 35) shows that the partial pressure of carbon dioxide does not grow linear at low loading but with the square of the loading X_c, if $X_c \ll 1$.

The solubility of CO_2 in cold potassium carbonate can be calculated from the following equation [18]:

$$p\,(CO_2) = \frac{0.0338 A^{1.29} X_c^2}{(150-t)\,\lambda_c\,(1-X_c)} \tag{37}$$

which is valid for $0 < t < 40\,°C$ and for $1 < A < 2$ kmol/m^3. The solubility of carbon dioxide in sodium carbonate can be calculated from a similar formula [19]

$$p\,(CO_2) = \frac{0.1015 A^{1.29} X_c^2}{(185-t)\,\lambda_c\,(1-X_c)} \tag{38}$$

which is valid for $18 < t < 65\,°C$ and for $0.5 < A < 2$ kmol/m^3. The solubility constant λ_c in Equations (37) and (38) is calculated as follows:

$$\log\lambda_c = 169/\,(94-t) - 2.89 \tag{39}$$

An empirical formula for CO_2 pressure (atm) over carbonate solution using the relationship molarity $m_a = 2A$ is given by the equation of Astarita et al. (Eq. 40) [71]:

$$p\,(CO_2) = 1.95 \times 10^9 m_a^{0.4} \exp\,(-8160/T)\; X_c^2/\,(1-X_c) \tag{40}$$

In Equations (30)–(40), $p(CO_2)$ is the equilibrium pressure of carbon dioxide in bar, A is the concentration of alkali (alkalinity) in kmol/m^3; X_c is the carbon dioxide loading ($X_c = [HCO_3^-]/A$), t is the temperature in °C; and λ_c is the solubility constant of carbon dioxide in water in kmol m^{-3} bar^{-1}.

In practice, the absorption of carbon dioxide by hot potassium carbonate solution (Hot-Pot of Benfield) is applied widely. Sodium carbonate and cold potassium carbonate are no longer used industrially.

The solubility equilibria of carbon dioxide in a potassium carbonate solution with additives (KH_2AsO_3, KH_2BO_3, K_2SeO_3, $KHTeO_4$, K_2HPO_4), which serve as activators to accelerate the rate of absorption of carbon dioxide, are described in [20].

Sulfur Dioxide in Alkali Sulfite–Bisulfite. By using Equations (29)–(33), with SO_2, HSO_3^-, and SO_3^{2-} replacing CO_2, HCO_3^-, and CO_3^{2-} a formally similar isotherm can be derived for the absorption of sulfur dioxide [21], [22]. The definition of loading, however, differs from that in Equation (34),

$$S_s = [HSO_3^-] + [SO_3^{2-}] \text{ so that}$$
$$[HSO_3^-] = (2\,S_s - A) \text{ and } [SO_3^{2-}] = (A - S_s) \tag{41}$$

are used for the bisulfite and sulfite concentrations, respectively. The partial pressure of sulfur dioxide is then

$$p\,(SO_2) = F\,(T\;)\,(2\,S_s - A)^2/(A - S_s) \tag{42}$$

The temperature-dependent Henry's law and dissociation constants are collected in the function $F\,(T)$. The following empirical formulas for $F\,(T)$ have been proposed for three bases [21], [22]:

Sodium hydroxide	$\log F\,(T\,) = 1.644 - 1987/T$
Methylamine	$\log F\,(T\,) = 2.515 - 2308/T$
Ammonia	$\log F\,(T\,) = 2.990 - 2368/T$

These formulas are valid for $308 < T < 363\,K$ and for the following alkalinities A in mol per 100 mol H_2O:

Sodium hydroxide	$4.0 < A < 7.8$
Methylamine	$7.3 < A < 22.0$
Ammonia	$5.8 < A < 22.4$

The partial pressure of sulfur dioxide is calculated in bar if the loading S_s is expressed in moles per hundred moles of water.

Hydrogen Sulfide and Carbon Dioxide in Alkali Carbonate–Bicarbonate. By the absorption of hydrogen sulfide in aqueous solutions of alkali carbonate–bicarbonate, the following reactions and equilibria are taken into account:

Reaction	Equilibrium	
$H_2S(g) \rightleftharpoons H_2S(l)$	$p(H_2S) = H_s[H_2S]$	(43)
$H_2S \rightleftharpoons H^+ + HS^-$	$K_h[H_2S] = [H^+]\,[HS^-]$	(44)

H_s is Henry's law constant and K_h is the dissociation constant of hydrogen sulfide. From Equations (31), (43), and (44), the partial pressure of hydrogen sulfide is

$$p\,(H_2S) = \frac{H_s K_k\,[HS^-]\,\left[HCO_3^-\right]}{K_h\,\left[CO_3^{2-}\right]} \tag{45}$$

Analogous to Equation (34), the hydrogen sulfide loading X_s is defined as

$$X_s = [HS^-]/A \quad (46)$$

and, from electroneutrality,

$$A = [HS^-] + [HCO_3^-] + 2[CO_3^{2-}] \quad (47)$$

By substitution in Equation (45), the partial pressure of hydrogen sulfide is obtained:

$$p(H_2S) = \frac{H_s K_k 2AX_sX_c}{K_h(1-X_s-X_c)} = \frac{1}{k'}\frac{2AX_sX_c}{(1-X_s-X_c)} \quad (48)$$

All previous constants are combined in the new constant k'. The following empirical formula has been proposed [23] for its temperature dependence:

$$\log k' = -5.81488 + 895.3/T \quad (49)$$

The constant k' is calculated in moles per kilogram and pascal if the temperature T is expressed in kelvin. Absorption isotherms (Eqs. 45 and 48, respectively) show that the partial pressure of hydrogen sulfide increases in direct proportion to its loading X_s. However, during the absorption of 1 mol of H_2S 1 mol H^+ and 1 mol HS^- are generated. By the reversal of Equation (31) 1 mol bicarbonate is formed by the reaction of 1 mol of H^+-ions while 1 mol of carbonate is consumed, so that the partial pressure of H_2S increases more rapidly than the loading.

Many experimental solubility data of hydrogen sulfide and carbon dioxide in aqueous potassium carbonate solution can be found in [24] and [25]. The design of carbon dioxide and hydrogen sulfide absorption with hot potassium carbonate can be calculated with the aid of a published simulation program written in FORTRAN [26].

Equilibria of Carbon Dioxide and Hydrogen Sulfide in Aqueous Ammonia. A model of the solubility equilibria of carbon dioxide and hydrogen sulfide in aqueous ammonia solution is based on the following reactions and equilibria: deprotonization of ammonium ion and reaction of carbamate ion with water [27].

Reaction	Equilibrium	
$NH_4^+ \rightleftharpoons NH_3 + H^+$	$K_a[NH_4^+] = [NH_3][H^+]$	(50)
$NH_2COO^- + H_2O \rightleftharpoons NH_3 + HCO_3^-$		
Carbamate	$K_c[NH_2COO^-] = [NH_3][HCO_3^-]$	(51)

as well as the equilibria for bicarbonate (Eq. 30), carbonate (Eq. 31), and hydrogen sulfide (Eq. 44). In addition, Henry's law (Eqs. 29 and 43) was taken into account. The mathematical description of the system is based on the material balance of ammonia, carbon dioxide, and hydrogen sulfide, as well as on electroneutrality. The equations can be solved by an approximation method, and the equilibria can thus be calculated. A better representation of solubility equilibria over a broad range of concentration is achieved by introducing activity coefficients. Remarkable progress was made during the 1970 s in calculating activity coefficients of electrolytes [28], [29].

A calculation of activity coefficients by the semiempirical method of Bromley and Pitzer, by using the TIDES program, successfully reproduces experimental solubility data of carbon dioxide in concentrated ammonia solutions [30]. An additional program, DELTAS, is available for calculating equilibria in a multicomponent system containing ammonia, phenol, carbon dioxide, hydrogen sulfide, hydrogen cyanide, sulfur dioxide, and water; it was used for calculating solubility equilibria in sour water strippers (i.e., desorbers for the removal of hydrogen sulfide, carbon dioxide, hydrogen cyanide, and ammonia from waste water).

This model was checked experimentally by measuring hydrogen sulfide – ammonia solubilities at extremely high ammonia concentrations; it reproduced the total pressure with less than ± 10 % error up to about 30 mol/kg [31]. This error becomes larger with further increasing concentration.

The fugacity of the vapor phase has been calculated according to the Peng – Robinson equation of state, and the effect of water vapor pressure on solubility equilibria at temperatures above 150 °C is demonstrated in [32]. Commercial simulation programs also make use of other models, e.g., the Chen model in Aspen Plus.

The vapor – liquid equilibria of ammonia and hydrogen sulfide in aqueous solutions are given by Newman [33] in the form of charts. They can be used for fast estimations without computer, but a trial-and-error method is required.

Experimental data on the simultaneous solubility of ammonia and carbon dioxide in concentrated aqueous solutions containing dissolved salts are reported by Maurer et al. [34]; for solutions containing sodium sulfate, see [35]. A data-correlation model treats the gas phase by

means of second virial coefficients, and the activity coefficients in the liquid are calculated by a modified Pitzer method for the excess Gibbs energy of an aqueous electrolyte solution. Selected binary and some ternary interaction parameters were fitted to the experimental data.

Carbon Dioxide – Hydrogen Sulfide in Alkanolamine. A well-known method of KENT and EISENBERG for representing solubility equilibria of carbon dioxide and hydrogen sulfide in aqueous alkanolamine solutions is based on nine reactions with their accompanying equilibrium constants [36]:

Reaction	Equilibrium	
$RRNH_2^+ \rightleftharpoons H^+ + RRNH$	$K_a[RRNH_2^+] = [H^+][RRNH]$	(52)
$RRNCOO^- + H_2O \rightleftharpoons RRNH + HCO_3^-$	$K_c[RRNCOO^-] = [RRNH][HCO_3^-]$	(53)
$H_2O \rightleftharpoons H^+ + OH^-$	$K_w = [H^+][OH^-]$	(54)

as well as Equations (29) – (31) and (40) – (41).

Together with the mass balance of amine, carbon dioxide, and hydrogen sulfide, as well as electroneutrality in the solution, a system of equations is obtained which can be solved by an approximation method. No explicit formula for the isotherm of the entire system of equations is known, but the equations can be solved by computer and the solubility equilibria calculated. The numerical values needed to calculate the temperature-dependent Henry's law constants and the dissociation constants of carbonate, bicarbonate and carbamate, sulfide and hydrogen sulfide, as well as two alkanolamines [monoethanolamine (MEA) and diethanolamine (DEA)], are given in [36].

With the aid of this model, nomograms can also be calculated [37], by means of which the simultaneous solubility equilibria of carbon dioxide and hydrogen sulfide in MEA or DEA can be determined even without a powerful computer. Significant deviations from the model, which occur at high loading, are corrected by using empirical factors.

Subprograms are available for calculating solubility equilibria by use of professional simulation programs, for example, an "Amine Data Package" for PROCESS. Similar subprograms are offered for ASPEN, SOFTEC, and ChemShare. The simulation programs TSWEET and AMSIM (Amine Gas Treating Plant Simulator) are also available [38], [39].

In a simplified solubility model for a single gas (CO_2 or H_2S) reacting with an amine in stoichiometric ratios of 1 : 1 and 1 : 2, the effect of amine concentration is described by an activity coefficient of the solvent [40]. Numerical values of Henry's law and reaction constants are given for three nonaqueous solvents and their mixtures.

A formally similar model has been used for two acid gases [41]. In this case, only a single reaction is considered per gas: the dissociation of H_2S to HS^- and H^+ and the conversion of carbon dioxide to carbamate. In both solubility models, a system of equations must be solved by an approximation method.

To represent the solubility of carbon dioxide in MEA solutions explicitly [42], [43], an isotherm has been derived from the following reaction

$$CO_2 + 2\,RNH_2 \rightleftharpoons RNHCOO^- + RNH_3^+ \tag{55}$$

From the equilibrium constant of this reaction and the amine balance, it folllows that

$$p\,(CO_2) = A\,[X_c/\,(1 - 2X_c)]^2 \exp\,(B) \tag{56}$$

The dependence of parameters A and B on the temperature T and specific loading X_c is represented by empirical equations

$$\ln A = -3.70 - 4.83X_c \tag{57}$$

$$B = \left(9.71 + 8.18X_c - 25.2X_c^2\right)(3.0 - 1000/T) \tag{58}$$

which can be used in the temperature range 0 – 140 °C, at specific loadings $X_c < 0.45$ mol/mol and MEA concentrations from 0.5 to 9.5 kmol/m^3. Another explicit isotherm of carbon dioxide is derived in [44]. The derivation is based on the assumption that the carbonate concentration in an alkanolamine solution is so low it can be neglected in the mass balance. The solution of the equations for physical solubility (Eq. 29), bicarbonate (Eq. 30), and carbamate (Eq. 53), and for the mass balance of amine and carbon dioxide as well as electroneutrality (without H^+ and OH^-),

$$A = [RRNH] + \left[RRNCOO^-\right] + \left[RRNH_2^+\right] \tag{59}$$

$$AX_c = \left[RRNCOO^-\right] + \left[HCO_3^-\right] \tag{60}$$

$$\left[\mathrm{RRNH_2^+}\right] = \left[\mathrm{RRNCOO^-}\right] + \left[\mathrm{HCO_3^-}\right] \tag{61}$$

leads to the following isotherm

$$p\left(\mathrm{CO_2}\right) = \frac{H_\mathrm{c} K_\mathrm{a} A}{K_\mathrm{b}} \cdot \frac{X_\mathrm{c}\left[X_\mathrm{c} - F\left(X_\mathrm{c}\right)\right]}{1 - X_\mathrm{c} - F\left(X_\mathrm{c}\right)} \tag{62}$$

The function $F\left(X_\mathrm{c}\right)$ can be replaced by

$$F\left(X_\mathrm{c}\right) = Sa - \left[Sa^2 - X_\mathrm{c}\left(1 - X_\mathrm{c}\right)\right]^{1/2} \tag{63}$$

where Sa is a factor proposed by SARTORI and SAVAGE which is defined as

$$Sa = \left(1 + K_\mathrm{c}/A\right)/2 \tag{64}$$

Equation (62) shows the effect of carbamate formation on the solubility isotherm. Carbamates can be formed only from primary and secondary amines. Tertiary amines do not form carbamates, and sterically hindered amines do so to only a minor extent.

For the simultaneous solubility equilibria of carbon dioxide and hydrogen sulfide in aqueous MDEA solution, absorption isotherms have been derived from the equilibria of bicarbonate (Eq. 30), hydrogen sulfide (Eq. 44), and amine hydrolysis (Eq. 52), as well as from Henry's law constant, electroneutrality, and amine balance [45].

$$p\left(\mathrm{CO_2}\right) = \frac{H_\mathrm{c} K_\mathrm{a} A}{K_\mathrm{b}} \frac{X_\mathrm{c}\left(X_\mathrm{c} + X_\mathrm{s}\right)}{1 - X_\mathrm{c} - X_\mathrm{s}} \tag{65}$$

$$p\left(\mathrm{H_2S}\right) = \frac{H_\mathrm{s} K_\mathrm{a} A}{K_\mathrm{h}} \frac{X_\mathrm{s}\left(X_\mathrm{c} + X_\mathrm{s}\right)}{\left(1 - X_\mathrm{c} - X_\mathrm{s}\right)} \tag{66}$$

OTTO et al. and MATHER et al. have published extended experimental data on solubility equilibria for carbon dioxide and hydrogen sulfide in aqueous solutions of alkanolamine (see Table 4).

Table 4. Solubility equilibria of carbon dioxide and hydrogen sulfide in aqueous alkanolamine solutions

Amine *	Gas	Reference
DEA	$H_2S + CO_2$	[46], [47]
DIPA	H_2S, CO_2	[48]
DIPA	$H_2S + CO_2$	[49]
DGA	H_2S, CO_2	[50]
MEA	$H_2S + CO_2$	[51]
MDEA	H_2S, CO_2	[52], [55], [56]
MEA + MDEA	CO_2	[57], [58]
DEA	H_2S, CO_2	[41], [54]
AMP	H_2S, CO_2	[59]
AMP	$H_2S + CO_2$	[60]
TEA	$H_2S + CO_2$	[53]

* AMP = 2-amino-2-methyl-1-propanol, DEA = diethanolamine, DGA = diglycolamine, DIPA = diisopropanolamine, MDEA = methyl diethanolamine, MEA = monoethanol amine, and TEA = triethanolamine

A model of vapor – liquid equilibria for aqueous solutions of acid gases in alkanolamines in which the activity coefficients are represented by the electrolyte NRTL equation was published by AUGSTEN et al. [61], [62]. The NRTL molecule – ion pair binary interaction parameters are fitted to experimental data with the data-regression system of the Aspen Plus process simulator.

A correlation of the solubility of CO_2 and H_2S in a mixed alkanolamine solutions by using the Pitzer electrolyte model and the Margules expansion for excess Gibbs energy was published by LI and MATHER [63], [64]. One binary molecule – molecule, as well as one binary and two ternary ion – moleule interaction parameters, are fitted to experimental data.

A new absorption isotherm for the solubility of acid gases in basic solutions was published in 1993 by SCHLAUER [65]. The chemistry of the reactions is established by the kinetics in Chapter 2.5. With the amine R_2NH as the base B, $R_2N^+HCO_2^-$ as the zwitterion, and $R_2NH_2^+$ as the protonated base, the equilibria of Equations (150) and (151) are given by Equations (67) and (68).

$$K_\mathrm{z}\left[\mathrm{R_2NH}\right]\left[\mathrm{CO_2}\right] = \left[\mathrm{R_2N^+HCO_2^-}\right] \tag{67}$$

$$K_\mathrm{d}\left[\mathrm{R_2N^+HCO_2^-}\right]\left[\mathrm{R_2NH}\right] = \left[\mathrm{R_2NCOO^-}\right]\left[\mathrm{R_2NH_2^+}\right] \tag{68}$$

with reaction constants K_z and K_d for the generation of the zwitterion and its decomposition to the carbamate R_2NCOO^-, respectively. Such reactions generally describe stoichiometric ratios of acid gas to amine of 1:1 and 1:2.

$$A = \left[\mathrm{R_2N^+HCO_2^-}\right] + \left[\mathrm{R_2NH}\right] + \left[\mathrm{R_2NCOO^-}\right] + \left[\mathrm{R_2NH_2^+}\right] \tag{69}$$

$$X_\mathrm{c} A = \left[\mathrm{CO_2}\right] + \left[\mathrm{R_2N^+HCO_2^-}\right] + \left[\mathrm{R_2NCOO^-}\right] \tag{70}$$

$$\left[\mathrm{R_2NCOO^-}\right] = \left[\mathrm{R_2NH_2^+}\right] \tag{71}$$

Eq. (29):

$$p_\mathrm{c} = H_\mathrm{c}\left[\mathrm{CO_2}\right]$$

Equations (67) and (68), the mass balances of the base (Eq. 69) and of the acid gas (Eq. 70), the electroneutrality condition (Eq. 71), and Henry's law (Eq. 29) can be solved analytically to give

$$X_c = \frac{p_c}{H_c A} + \frac{p_c K_z/H_c + (p_c K_d K_z/H_c)^{1/2}}{1 + p_c K_z/H_c + 2\,(p_c K_d K_z/H_c)^{1/2}} \quad (72)$$

$$X_c = \frac{p_c}{H_c A} + \frac{p_c C_c + (p_c B_c)^{1/2}}{1 + p_c C_c + 2\,(p_c B_c)^{1/2}} \quad (73)$$

where

$$C_c = K_z/H_c \text{ and } B_c = K_d K_z/H_c \quad (74)$$

whereby the parameters H_c, C_c, and B_c are fitted to the experimental data. The isotherm also applies to the solubility of other sour gases in alkaline solvents, for example, H_2S in aqueous MDEA (Fig. 3). The H_2S fugacity was calculated from the experimental partial pressure by a truncated virial equation of state, and the second virial coefficients were calculated according to PRAUSNITZ et al. [2]

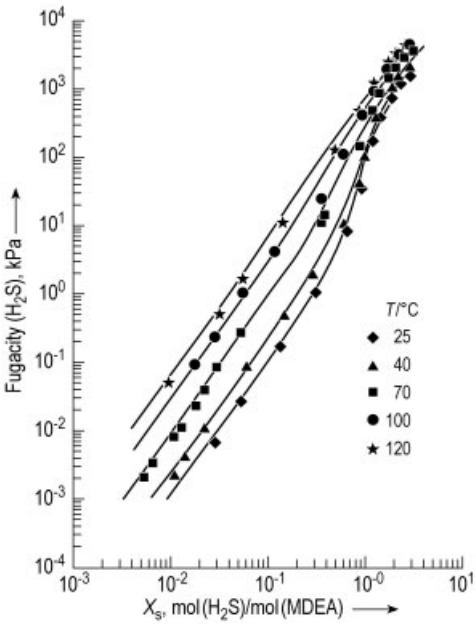

Figure 3. Comparison of predicted and experimental equilibrium data for H_2S in 1 M aqueous MDEA

The isotherm can be expanded to the simultaneous solubility of more sour gases (Eq. 75).

$$X_i = \frac{p_i}{H_i A} + \frac{p_i C_i + p_i B_i / \left(\sum_j p_j B_j\right)^{1/2}}{1 + \sum_j p_j C_j + 2\left(\sum_j p_j B_j\right)^{1/2}} \quad (75)$$

Experimental solubility data of the single gases carbon dioxide and hydrogen sulfide in aqueous solutions of MDEA were correlated by MAURER et al. [55] with a model already described for aqueous solutions of ammonia. Selected binary parameters and one ternary interaction parameter were fitted to the experimental data.

1.6. Enthalpy and Absorption Equilibrium

The temperature dependence of the fugacity f_i of a dissolved component i is related to the molar enthalpy of absorption $\Delta H_i^{\mathrm{Abs}}$ of this component by the Clausius – Clapeyron equation

$$\delta \ln f_i/\delta T = -\,\Delta H_i^{\mathrm{Abs}}/RT^2 \Delta z_i \quad (76)$$

where Δz_i is the difference between the compressibility factors of the vapor and liquid phases

$$\Delta z_i = z_i^{\mathrm{G}} - z_i^{\mathrm{L}} = \left(v_i^{\mathrm{G}} - v_i^{\mathrm{L}}\right) p/RT \quad (77)$$

For $z_i^{\mathrm{G}} \approx 1$ in the vapor phase and $z_i^{\mathrm{L}} \ll 1$ in the liquid phase follows: $\Delta z_i \approx 1, f_i \approx p_i$,

$$\Delta H_i^{\mathrm{Abs}} = R\,[\delta \ln p_i/\delta\,(1/T)]_x \quad (78)$$

Equation (78) is used to calculate the *differential enthalpy of absorption* at a loading x_i from known solubility equilibria. An *average enthalpy of absorption* for a loading from x_i (1) to x_i (2) is determined by the appropriate integration

$$\Delta H_i^{\mathrm{Abs}} = \int_{x(1)}^{x(2)} \Delta H_i^{\mathrm{Abs}} \mathrm{d}x_i / [x_i\,(2) - x_i\,(1)] \quad (79)$$

The enthalpy of absorption can be calculated by using UNIQUAC [2] or UNIFAC [66]. The temperature dependence of Henry's law constant $H_{i,\,\mathrm{L}}$ is determined from Equations (21) and (78):

$$\delta \ln H_{i,\mathrm{L}}/\delta\,(1/T) = \Delta H_i^{\mathrm{Abs}}/R \quad (80)$$

The temperature dependence of Henry's law constants can be estimated roughly by using the following formula [9]:

$$H_{i,\mathrm{L}} = \gamma_{i,\mathrm{L}}^{\infty} f_i^0 \tag{81}$$

where $\gamma_{i,\mathrm{L}}^{\infty}$ is the activity coefficient of gas component i in the solvent L at infinite dilution and f_i^0 is the standard fugacity. The standard fugacity is independent of the solvent, and its generalized temperature dependence is given by

$$\ln\left(f_i^0/p_{\mathrm{c},i}\right) = 7.224 - 7.534/T_{\mathrm{r},i} - 2.598 \ln T_{\mathrm{r},i} \tag{82}$$

for a range of temperatures $0.6 < T_{\mathrm{r},i} < 16$. The activity coefficient at infinite dilution $\gamma_{i,\mathrm{L}}^{\infty}$ is specific for each pair (gas component i and solvent L) but only weakly dependent on the temperature.

At low temperatures, Henry's law constant increases with increasing temperature until a maximum at $T_{\mathrm{r},i} = 2.8$ is reached. At higher temperatures Henry's law constant decreases slightly, i.e., the solubility again increases.

The enthalpy of absorption increases steadily with increasing temperature (i.e., it decreases in absolute value) and reaches zero at $T_{\mathrm{r},i} = 2.8$. At lower temperatures, heat is liberated during absorption; at higher temperatures, it is consumed. In *chemical absorption* the enthalpy of absorption is strongly dependent on loading [52]. Once the chemical capacity has been saturated, the absorption enthalpy drops substantially.

The temperature dependence of the solubility equilibria and the enthalpy of absorption is of paramount importance in designing an absorption installation. In the absorber, the gas component to be washed out must be highly soluble. In the regenerator where the absorbed gas component from the solvent is removed, the solubility must be as low as possible. The strong increase in solubility at low temperatures is utilized particularly in physical absorption processes, the absorber being operated at as low a temperature as possible and the regenerator at a high temperature. Because the enthalpy of absorption becomes negative and, in absolute terms, always increases with decreasing temperature, provision must be made to remove the liberated heat of absorption from the absorber to improve the efficiency of the process. To do so, solvent is removed from the absorber, cooled, expanded, and then recirculated to the absorber.

To obtain as pure a treated gas as possible, the regenerated solvent is cooled to a very low temperature before it is brought to the absorber head. The lowest temperature is obtained by adiabatic expansion of the loaded solvent (possibly under vacuum).

In processes involving chemical absorption, reaction kinetics limit the use of low temperatures in an absorber. Although the partial pressure of a dissolved component drops as the temperature decreases, this fact cannot be used to advantage in an industrial absorber because the rates of chemical reaction are retarded by low temperatures to such an extent that mass transfer is hindered and a long residence time (i.e., a larger absorber) is required. Reaction kinetics of chemical absorption and the use of kinetic effects are dealt with in Chapter 2.

2. Mass Transfer

A comprehensive treatment of mass transfer phenomena is given elsewhere (→ Transport Phenomena).

2.1. Introduction

This chapter deals with mass transfer during absorption occurring in a gas – liquid system not in equilibrium. The nonequilibrium produces concentration gradients at the interface and mass transfer across this boundary, which tends to equilibrate the system.

During absorption or desorption, the absorbed component cannot spread spontaneously over the total volume of a phase. In the gas phase, mass transfer during absorption is hindered by the gaseous components that are not absorbed and remain as purified gas. During desorption, it is hindered by the stripping gas or stripping steam. In the liquid phase, mass transfer is hindered by the solvent.

Mass transfer of the absorbed components near the interface occurs mainly by diffusion and, in the bulk of the phase, mainly by turbulence. At equilibrium, mass transfer across the interface is equal in both directions so that no concentration gradient is established near the phase boundary. This chapter deals only with mass transfer that occurs before equilibrium has been established between the two phases, not with mass transfer at equilibrium.

2.2. Mass Transfer Coefficients

Mass transfer across the interface of a component to be absorbed is described quantitatively by the mass transfer coefficients. The definition of these coefficients is based on two assumptions:

1) At the phase boundary the gas and liquid phases are at equilibrium. Significant deviations from equilibrium at the interface occur only if mass transfer is extremely rapid.
2) Mass transfer is proportional to the difference between the concentration in the bulk of the phase and the concentration at the interface.

In the stationary state, mass transfer is equal on both sides of the interface so that

$$N_{\mathrm{A}}=k_{\mathrm{G}}\left(y^{\mathrm{b}}-y^{\mathrm{i}}\right)=k_{\mathrm{L}}\left(x^{\mathrm{i}}-x^{\mathrm{b}}\right) \tag{83}$$

The mass transfer rate N_{A} is the number of moles of an absorbed component A being transferred across a unit area of the interface per time unit; y^{b} and x^{b} are the mole fractions of component A in the bulk of the gas and liquid phases; y^{i} and x^{i} are the mole fractions at the interface. The proportionality constants k_{G} and k_{L} are the mass transfer coefficients in the gas and liquid phase, respectively. The mole fractions at the interface can be calculated from known mass transfer coefficients and solubility equilibria by using the following approximation method:

1) Mole fractions $x*$ and $y*$ are calculated which are in equilibrium with mole fractions y^{b} and x^{b} in the bulk of the other phases

$$y^{*}=F\left(x^{\mathrm{b}}\right) \text{ and } y^{\mathrm{b}}=F\left(x^{*}\right) \tag{84}$$

where $F\left(x^{\mathrm{b}}\right)$ and $F\left(x*\right)$ are the solubility equilibrium relationships.

2) An intermediate value C is calculated

$$C=\frac{k_{\mathrm{G}}\left(y^{\mathrm{b}}-y^{*}\right)}{k_{\mathrm{L}}\left(x^{*}-x^{\mathrm{b}}\right)} \tag{85}$$

3) As a first approximation, the equilibrium concentration at the interface is

$$y_1=\left(Cy^{\mathrm{b}}+y^{*}\right)/(1+C) \tag{86}$$

$$x_1=x^{\mathrm{b}}+\left(y^{\mathrm{b}}-y_1\right)k_{\mathrm{G}}/k_{\mathrm{L}} \tag{87}$$

4) The question of whether these values satisfy the solubility equilbrium relation

$$y_1^{*}=F\left(x_1\right) \tag{88}$$

by meeting the following condition

$$Abs\left[y_1^{*}-y_1\right]<\mathrm{Abs}\left[y_1-y^{\mathrm{b}}\right]/100 \tag{89}$$

must be checked. If condition (89) is met, the relative deviation is less than 1 %. If not, the calculation is repeated beginning with step 1, but the new values x_1 and y_1 in Equation (87) are used in place of the mole fractions x^{b} and y^{b}. Figure 4 shows that this approximation method converges rapidly, even if the absorption equilibrium is strongly nonlinear.

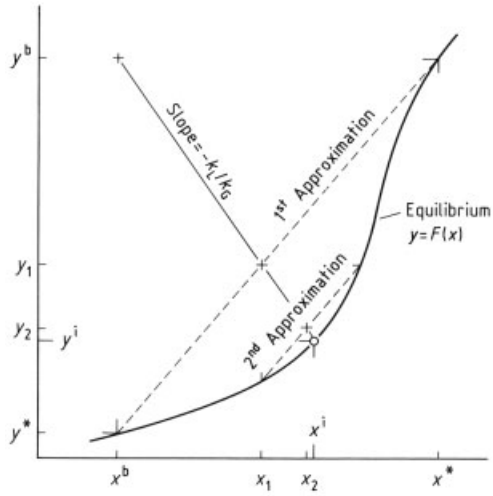

Figure 4. Prediction of the concentration at the interface: graphical demonstration of the approximation

If the absorption equilibrium in the concentration region employed can be represented by a straight line

$$y=n+mx \quad (n,m=\text{ const.}) \tag{90}$$

the mass transfer is also proportional to the difference between the concentration in the bulk of one phase and the equilibrium value of the other phase

$$N_{\mathrm{A}}=K_{\mathrm{G}}\left(y^{\mathrm{b}}-y^{*}\right)=K_{\mathrm{L}}\left(x^{*}-x^{\mathrm{b}}\right) \tag{91}$$

These proportionality constants K_{G} and K_{L} are the overall mass transfer coefficients in the gas and liquid phase, respectively. For recalculation,

$$1/K_L = 1/k_L + 1/mk_G \tag{92}$$

$$1/K_G = 1/k_G + m/k_L \tag{93}$$

$$K_L = mK_G \tag{94}$$

The inverse of the mass transfer coefficients can be considered the resistance to mass transfer, and the overall resistance consists of the resistance in the gas and liquid phases. If the resistance is substantially greater in one phase than in the other, it is called a one-sided resistance. For one-sided resistance in the gas phase,

$$k_L \gg mk_G \;\; x^i \to x^b,\; y^i \to y^* \tag{95}$$

If resistance in the liquid phase predominates, analogously,

$$k_L \ll mk_G \;\; x^i \to x^*,\; y^i \to y^b \tag{96}$$

The concentrations at the interface can only be between these extremes (cf. Fig. 4).

$$x^b \le x^i \le x^* \text{ and } y^* \le y^i \le y^b \tag{97}$$

In industrial absorption devices, where the surface of the interface cannot be determined accurately, volume-related mass transfer coefficients are used

$$K_L a, K_G a, k_L a, \text{and } k_G a \tag{98}$$

where a is the unknown surface of the interface per unit volume of the absorption device. For recalculation,

$$1/K_L a = 1/k_L a + 1/mk_G a \tag{99}$$

$$1/K_G a = 1/k_G a + m/k_L a \tag{100}$$

These relationships are especially useful in a physical solubility equilibrium, where a dimensionless Henry's law constant $H_{i,\,L}/p$ or the K value can be substituted for m. The surface a decreases with decreasing liquid flow rate L and with increasing surface tension (Marangoni effect) because streams are formed and the surface of the packing is not completely wetted [74]. The effective phase surface and the mass transfer coefficients has been measured in contact devices [75] for gases with polar and nonpolar liquids as well as with highly viscous hydrocarbons.

Numerical values of mass transfer coefficients $K_G\,a$ in packed columns for several industrially important absorption systems can be found in [67], [73] and [75].

2.3. Effect of Solute Concentration

The relations given in Section 2.2 are valid only for low concentration of the absorbed component A, as long as $x_A \ll 1$ and $y_A \ll 1$. At higher concentration, the mass flow created by diffusion must be taken into account [68]. If two components A and B in a gas film at the interface are considered, a concentration gradient is formed in which

$$N_A = -cD_{AB}\frac{dy_A}{dz} + y_A\,(N_A + N_B) \tag{101}$$

Here, c is the overall concentration of components A and B in the gas phase, z the distance from the interface, and D_{AB} the diffusion constant of component A in the binary gas mixture. In the steady state, the components diffuse in opposite directions so that only the adsorbed component A flows through the gas film; the nonadsorbed component B remains in the gas film. Hence, $N_B = 0$ and from Equation (101),

$$N_A = -\frac{cD_{AB}}{1-y_A}\frac{dy_A}{dz} \tag{102}$$

The mass transfer of component A increases in proportion to the concentration gradient dy_A/dz and to $1/(1-y_A)$, hence with increasing concentration of y_A. From the continuity of mass transfer,

$$\frac{dN_A}{dz} = 0 \text{ then } \frac{d}{dz}\left(\frac{1}{1-y_A}\cdot\frac{dy_A}{dz}\right) = 0 \tag{103}$$

After the first and second integration,

$$[1/(1-y_A)]\,dy_A/dz = C_1 \tag{104}$$

$$-\ln(1-y_A) = zC_1 + C_2 \tag{105}$$

The constants C_1 and C_2 can be calculated from the boundary conditions that exist at the interface and at the boundary between the gas film and the bulk of the gas phase

$$y_A = y_A^i \text{ if } z = 0 \tag{106}$$

$$y_A = y_A^b \text{ if } z = \delta_G \tag{107}$$

where δ_G is the thickness of the gas film or the distance from the interface where the concentration in the gas phase attains the value y_A^b. From Equations (104) – (107),

$$C_2 = -\ln\left(1-y_A^i\right) = -\ln y_B^i \tag{108}$$

$$C_1 = \frac{1}{z}\ln\frac{1-y_A^i}{1-y_A^b} = \frac{1}{z}\ln\frac{y_B^i}{y_B^b} \tag{109}$$

and, finally, Equations (103), (104), and (105)

$$N_A = \frac{cD_{AB}}{\delta_G}\ln\frac{1-y_A^i}{1-y_A^b} = \frac{cD_{AB}}{\delta_G}\ln\frac{y_B^i}{y_B^b} \tag{110}$$

This equation describes the mass transfer at all concentrations of adsorbed component A. At a lower concentration,

$$\ln(1-y_A) \approx -y_A \quad (y_A \to 0) \tag{111}$$

and Equation (110) becomes

$$N_A = \frac{cD_{AB}}{\delta_G}\left(y_A^b - y_A^i\right) \tag{112}$$

By comparison with the definition of N_A (Eq. 83),

$$k_G = cD_{AB}/\delta_G \tag{113}$$

In an ideal gas phase, the concentration c is

$$c = p/RT \tag{114}$$

and the mass transfer coefficient is

$$k_G = pD_{AB}/\delta_G RT \tag{115}$$

which defines the thickness of the gas film δ_G; it is strongly dependent on turbulence in the gas phase. Similar dependencies on concentration apply in the liquid, but as a rule, the concentration of the absorbed component in the liquid phase is low.

2.4. Correlation of Mass Transfer Coefficients

Empirical relationships from dimensionless analysis are used to correlate mass transfer coefficients in the gas phase. One criterion for the *mass transfer coefficient in the gas phase* is the Sherwood number

$$Sh = k_G RT d/pD_{AB} \tag{116}$$

As characteristic length d is usually taken the dimensions of the adsorption column packing, the diameter of the gas channel in a perforated plate, or the space between the bubble cap and the plate. Turbulence in the gas phase is characterized by the Reynolds and Schmidt numbers:

$$Re = G_M d/\mu_G \tag{117}$$

$$Sc = \mu_G/\varrho_G D_{AB} \tag{118}$$

The following relationship has been established from experimental mass transfer data in an absorption unit:

$$Sh = St \cdot Re \cdot Sc = f(Re, Sc) \tag{119}$$

This empirical relation can then be used to scale up absorbers of the same type. The dimensionless Stanton number is defined as

$$St = k_G M_G/G_M \tag{120}$$

The following empirical formula has been proposed for mass transfer in packed columns with Raschig rings and Berl saddles [76]:

$$j_M = St \cdot Sc^{2/3} = 1.195 Re^{-0.36} \tag{121}$$

where j_M is the j factor according to Chilton–Colburn for the modified Reynolds analogy among mass, heat, and impulse transfer. Here, the Reynolds number is modified somewhat to

$$Re = G_M d/\mu_G (1-\varepsilon) \tag{122}$$

where d is the diameter of a sphere having the same surface as a piece of packing and ε is the void fraction of the packing available for gas flow.

In Equations (117)–(122), k_G is the mass transfer coefficient in kmol s^{-1} m^{-2}; G_M is the specific gas flow rate in kg s^{-1} m^{-2}; d is the characteristic length in m; μ_G is the viscosity of the gas phase in kg s^{-1} m^{-1}, ϱ_G is the density of the gas phase in kg/m^3; D_{AB} is the diffusion constant of component A in m^2/s in inert gas B; and M_G is the molar mass of the gas phase $= \sum y_i M_i$ in kg/kmol.

The following empirical formula can be used for mass transfer in bubble-cap tray absorption columns [77]:

$$k_g = 7\left(u^{1/2} D_{AB}/s\right)^{1/2} \tag{123}$$

Here k_g is the mass transfer coefficient in cm/s; u is the gas velocity in cm/s; s is the liquid level on the tray in cm; and D_{AB} is the diffusion coefficient in cm^2/s.

To convert from k_g to k_G (SI units),

$$k_G = k_g/100V \tag{124}$$

or

$$k_G = k_g p/100RT \tag{125}$$

where V is the molar volume of the gas phase in cubic meters per kilomole. Equation (125) applies to an ideal gas phase. The *mass transfer coefficients* k_L *in the liquid phase* for Raschig rings and Berl saddles can be calculated with the following empirical formula [76]; similar correlations also apply to packed columns [78]:

$$St = 25.1 Re^{0.45} \cdot Sc^{0.5} \tag{126}$$

The Stanton, Reynolds, and Schmidt numbers in the liquid are

$$St = k_L M_L / L_M \tag{127}$$

$$Re = L_M d/\mu_L \tag{128}$$

$$Sc = \mu_L / p_L D_{AL} \tag{129}$$

where k_L is the mass transfer coefficient in kmol s^{-1} m^{-2}; M_L the molar mass of liquid in kg/kmol; d the characteristic length in m; D_{AL} the diffusion constant of component A in the liquid phase in m^2/s; ϱ_L the density of liquid phase in kg/m^3; L_M the specific liquid flow rate in kg s^{-1} m^{-2}; and μ_L the viscosity of the liquid phase in kg s^{-1} m^{-1}.

Additional correlations of mass and heat transfer can be found in [79–84].

Another relation holds for the mass transfer of carbon dioxide from a *supersaturated aqueous solution* in a container with a stirrer [85]. As long as no bubbles are formed, mass transfer through the surface of the liquid can be described by the following empirical formula:

$$Sh = 0.322 Re^{0.7} \cdot Sc^{1/3} \tag{130}$$

In this case, the Sherwood, Reynolds, and Schmidt numbers are defined as follows

$$Sh = k_l d / D_{AL} \tag{131}$$

$$Re = d_L^2 n_L \varrho_L / \mu_L \tag{132}$$

$$Sc = \mu_L / \varrho_L D_{AL} \tag{133}$$

where d_L is the stirrer diameter in m; n_L is the rotation rate of the stirrer in s^{-1}; d is the container diameter in m; and k_l is the mass transfer coefficient in m/s.

The mass transfer coefficient k_l is expressed in meters per second and like Equation (83), is defined by the difference in concentrations

$$N_A = k_l \left(c_A^i - c_A^b \right) \tag{134}$$

The concentrations of component A at the interface c_A^i and the bulk of the liquid phase c_A^b, are expressed in $kmol/m^3$.

To convert k_l to k_L (mass transfer coefficient referred to mole fraction)

$$k_l = k_L / M_L \varrho_L \tag{135}$$

is used. In *highly supersaturated solutions*, bubble formation occurs which increases the phase surface and intensifies turbulence. Two regions of spontaneous bubble formation of this type exist which differ in the intensity of bubble formation depending on the supersaturation σ:

$$\sigma = \left(c_A^b - c_A^* \right) / c_A^* \tag{136}$$

where c_A^* is the concentration in equilibrium with the gas phase (bulk). Bubble formation is weak up to critical supersaturation σ_c, and the volume-related mass transfer coefficient in the bulk of the liquid phase is

$$k_l^b a = 6.77 10^{-6} Re^{0.5} \cdot \sigma^{0.78} \tag{137}$$

Once the critical supersaturation has been exceeded and $\sigma > \sigma_c$, bubble formation becomes intense, and the mass transfer coefficient is

$$k_l^b a = 2.45 10^{-6} Re^{0.98} \cdot \sigma^{2.5} \tag{138}$$

For the critical supersaturation σ_c between weak and intense bubble formation, the range of validity is defined as follows:

$$\sigma_c = 1.81 Re^{-0.25} \tag{139}$$

$$2190 < Re < 14000 \tag{140}$$

$$0.038 < \sigma < 0.80 \tag{141}$$

2.5. Mass Transfer and Chemical Reaction

If the absorbed component reacts in the liquid phase and is thus transformed chemically, its mass transfer in the liquid phase is accelerated. If this reaction is irreversible, the absorbed component reacts until its concentration in the liquid phase is practically zero. Thus, for example, in the absorption of carbon dioxide in aqueous potassium hydroxide solution, carbonate (CO_3^{2-}) and traces of bicarbonate (HCO_3^-) can be found in solution, but practically no carbon dioxide as long as an excess of alkali is present. Mass transfer is then independent of loading and proportional to the concentration at the interface c_A^i because the concentration in the bulk of the liquid phase $c_A^b = 0$.

For an irreversible first-order reaction, the mass transfer coefficient in the liquid phase depends on the Hatta number [86]. According to Danckwert's surface renewal theory, the mass transfer coefficient including a reaction of component A with reactant B is calculated according to the following relation [87]:

$$k_l' = \left(k_l^2 + D_{AL} k_{11} c_B^b\right)^{1/2} = k_l \left(1 + Ha^2\right)^{1/2} = k_l E \tag{142}$$

where Ha is a dimensionless Hatta number, E is the enhancement factor, k_{11} is the reaction rate constant, and k_l is the mass transfer coefficient in the liquid phase without a reaction.

$$Ha = \left(D_{AL} k_{11} c_B^b\right)^{1/2} / k_l \tag{143}$$

$$E = \left(1 + Ha^2\right)^{1/2} \tag{144}$$

For large Hatta numbers beginning with $Ha > 3$, the enhancement factor $E \approx Ha$. These parameters describe the contribution of the reaction to mass transfer if a chemical reaction occurs. This equation can also be used for a pseudo-first-order reaction, as long as an excess of B is present. For very rapid reactions, the Hatta number is large and mass transfer is controlled increasingly by the mass transfer coefficient of the gas phase.

The reaction rate constants k_{11} of the most frequently encountered reactions are known, and their temperature dependence is represented according to ARRHENIUS by

$$k_{11} = F \exp\left(-E^A / RT\right) \tag{145}$$

where F is the frequency factor and E^A the activation energy of the reaction; numerical values of some *second-order reactions* are given in Table 5.

Table 5. Frequency factors and activation energies of several second-order reactions

Reaction	ln F, $m^3\,k\,mol^{-1}\,s^{-1}$	E^A, kJ/mol	References
$CO_2 + OH^-$	31.43	55.42	[88], [89]
$CO_2 + H_2O$	26.13	70.98	[89]
$CO_2 + NH_3$	25.63	48.43	[89]
CO_2 + MEA	25.84	41.92	[90]
	25.31	41.20	[91]
CO_2 + MIPA	26.43	43.84	[90]
CO_2 + TEA	24.68	51.46	[91]
CO_2 + MDEA	16.83	37.21	[92]
	18.63	42.48	[93]
	29.80	71.71	[94]
NO + Co – TE	31.00	22.40	[95]
$O_2 + SO_3^{2-}$	30.8	58.1	[96]
(at pH 8.0; $[Co^{2+}] = 0.1$ mmol/L as catalyst)			

∗ MIPA = monoisopropanolamine;
TE = tetraethylenepentamine; other abbreviations are defined in Table 4.

The reaction rate constant k_{11} is calculated in $m^3\,kmol^{-1}\,s^{-1}$; from equation (145) by using $R = 0.008314$ kJ mol^{-1} K^{-1} for the gas constant and expressing the temperature T in Kelvin. In Table 5, Co – TE denotes a complex of cobalt (II) with tetraethylenepentamine, which forms a nitrosyl compound with nitric oxide. Numerical values of the reaction rate constants of nitric oxide with similar cobalt and iron complexes in aqueous solution have also been published [95], [97], and [98]. BRØNSTED has given a relationship between the reaction rate constant k_{11} of carbon dioxide with alkanolamines and the dissociation constand K_a of these alkanolamines, which suggests a homogeneous base catalysis, the following relation is recommended for primary and secondary alkanolamines [93]:

$$\ln k_{11} = \mathrm{p}K_a + 16.26 - 7188/T \tag{146}$$

This empirical formula applies from 293 to 303 K. For tertiary alkanolamines at 293 K,

$$\ln k_{11} = \mathrm{p}K_a - 14.24 \tag{147}$$

is recommended [93]. Some industrially important reactions are of a higher order, e.g., the reaction of carbon dioxide with the alkanolamines

DEA or DIPA, shown in Figures 5 and 6. For higher order reactions,

$$r = k_{mn}\,[\mathrm{A}]^m\,[\mathrm{B}]^n \tag{148}$$

the Hatta number according to HIGBIE is

$$Ha = k_1^{-1} \cdot \left[2 D_{\mathrm{AL}}\, k_{mn} \left(c_{\mathrm{A}}^{\mathrm{i}}\right)^{m-1} \left(c_{\mathrm{B}}^{\mathrm{b}}\right)^n / (m+1)\right]^{1/2} \tag{149}$$

These reactions probably occur according to the following mechanism [88], [100]. Carbon dioxide forms a zwitterion with an alkanolamine which, in a subsequent reaction, donates a proton to a base and is converted to the carbamate:

$$CO_2 + R_2NH \underset{k_1}{\overset{k_{11}}{\rightleftharpoons}} R_2NHCO_2 \tag{150}$$

$$R_2NHCO_2 + B \overset{k_2}{\rightleftharpoons} BR_2NHOO^- + BH^+ \tag{151}$$

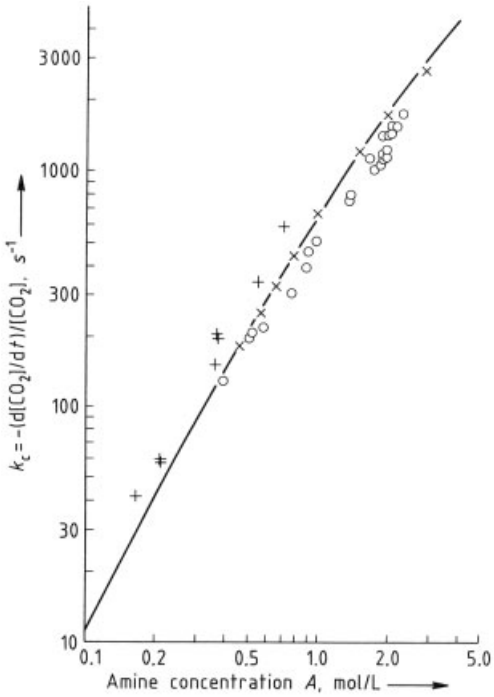

Figure 5. Rate constant for the reaction of carbon dioxide with diethanolamine – effect of amine concentration
Temperature = 298 K.
+ Data from [91]; o Data from [101]; × Data and calculation [99]: $k_c = A/(1/1410 + 1/1200\,A)$

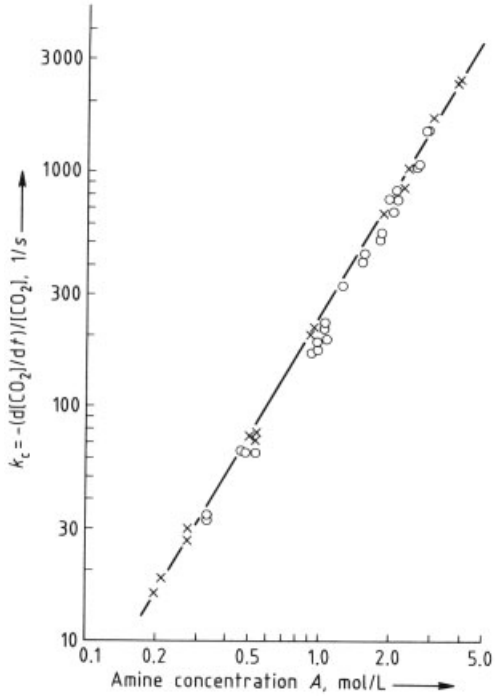

Figure 6. Rate constant for the reaction of carbon dioxide with diisopropanolamine – effect of amine concentration
Temperature = 298 K.
o Data from [101]; × Data from [93]

In the second reaction (Eq. 151) the proton can be accepted by various bases for example, OH^- ($\rightarrow H_2O$), water ($\rightarrow H_3O^+$), or an amine ($\rightarrow$ amine H^+). After the concentration of the zwitterion $[R_2NHCO_2]$ is eliminated mathematically, from the kinetics of Equations (150) and (151),

$$r = \frac{\mathrm{d}\,[\mathrm{CO_2}]}{\mathrm{d}t} = \frac{k_{11}\,[\mathrm{CO_2}]\,[\mathrm{R_2NH}]}{1 + k_{-1}/\sum k_{\mathrm{B}}\,[\mathrm{B}]} \tag{152}$$

If the increase in concentration of bases, [B], is proportional to the amine concentration

$$[B] = K_{\mathrm{B}}\,[\mathrm{R_2NH}] \tag{153}$$

and the denominator of Equation (152) is

$$1 \ll k_{-1}/\sum k_{\mathrm{B}}\,[\mathrm{B}] \tag{154}$$

the reaction tends toward third order and the reaction rate is proportional to the square of the amine concentration

$$r \approx \left(\sum k_{\mathrm{B}} k_{11} K_{\mathrm{B}}/k_{-1}\right)[\mathrm{CO_2}]\,[\mathrm{R_2NH}]^2 \tag{155}$$

Numerical values for the rate constants and energies of activation of reactions of carbon dioxide with aqueous and alcoholic solutions of DEA, MEA, and DIPA are given in [93], [101], and [102].

2.6. Modeling of Mass Transfer with Chemical Reaction

An analytical solution of the differential equations for mass transfer with a reaction of the absorbed component in the liquid phase is known only for a (pseudo) first-order reaction. This reaction must be either irreversible or coupled with a rapid secondary reaction of the products so that re-formation of the absorbed components need not be considered. Formulas derived for the following three models are known.

1) *Penetration Theory* (**Danckwerts**, **Higbie**):

$$k_1' = k_1 Ha \left(1 + 1/8Ha^2\right) \operatorname{erf}\left(2Ha/\pi^{1/2}\right) + 2Ha \exp\left(-4Ha^2/\pi\right) \quad (156)$$

2) *Two-Film Theory* (**Sherwood** and **Pigford**):

$$k_1' = k_1 Ha/\tanh Ha \quad (157)$$

3) *Surface-Renewal Theory* (**Danckwerts**):

$$k_1' = k_1 \left(1 + Ha^2\right)^{1/2} \quad (158)$$

Here, Ha is the Hatta number $(D_{\mathrm{AL}} k_{\mathrm{c}}/k_1^2)^{1/2}$, k_1' is the mass transfer coefficient with chemical reaction in the liquid phase in m/s; k_1 is the mass transfer coefficient in the liquid phase (with no reaction) in m/s; k_{c} is the reaction rate constant of the (pseudo) first-order reaction in s^{-1}, and D_{AL} is the diffusion coefficient of absorbed component A in liquid phase L in m^2/s.

A reaction of higher order

$$A + n\mathrm{B} \rightarrow \text{Product} \quad (159)$$

and its reaction rate r

$$r = k_{1n} [\mathrm{A}] [\mathrm{B}]^n \quad (160)$$

can be considered pseudo-first order with respect to A as long as the reaction partner B is not consumed to any great extent. Until then,

$$r = k_{\mathrm{c}} [\mathrm{A}], \text{ where } k_{\mathrm{c}} = k_{1n} [\mathrm{B}]^n \quad (161)$$

All three models lead to similar results [103]. No analytical solution of the differential equations is known for reversible reactions: other models or numerical methods are used to solve differential equations.

To represent the absorption kinetics of carbon dioxide and hydrogen sulfide in aqueous solutions of DIPA and MDEA the two-film theory with two layers is used [104]. The assumption is that the reaction is an irreversible first-order reaction and that the concentration profiles are linear [105]. If the boundary values from Danckwert's model are used for the concentration gradient of both layers, the differential equations can be replaced by nonlinear algebraic equations and solved numerically. This model was used by Shell Oil in the design of commercial installations for the selective absorption of hydrogen sulfide from gases containing carbon dioxide.

With the aid of a computation model based on the penetration theory, the differential equations for mass transfer, including complex reversible reactions in the liquid phase, can be solved numerically [106]. The concentration gradients are calculated by using a grid, the differential equations being replaced by a system of linear equations. Further computational models can be found in [107] and [108].

A professional simulation program dealing with kinetics and absorption is available under the name Aspen plus, which offers the method Ratefrac.

3. Design of Absorption Systems

3.1. Methods of Absorption

The use of *physical absorption* processes is preferred whenever

1) feed gases are present in large amounts at high pressure and the amount of the component to be absorbed is relatively large;
2) the purified gas (the fraction of gas that is not absorbed) must be of high purity;
3) one or more selectively absorbed components are to be obtained in enriched or pure form, possibly still at elevated pressure;
4) the absorbent is also the product of a subsequent synthesis, e.g., methanol (→ Gas Production).

On the other hand, absorption processes with simultaneous chemical reaction (*chemisorption*) are always preferred when

1) the components to be separated from feed gases are present in small concentrations and at low partial pressures;
2) the purity requirements of the purified gas are not too high;
3) the components to be separated from feed gases are strongly acidic and undergo a thermally reversible reaction with the chemical component of the absorbent;
4) low-cost waste heat is available for thermally regenerating the absorbent.

Absorption equipment for both types of absorption generally consists of a column with internals for heat and material exchange (plate or packing) in which the feed gas is brought into countercurrent contact with the regenerated absorbent.

The purified gas is taken off the head of this column. The absorbent loaded with the removed component, the absorbate, leaves the bottom of the column and is regenerated by desorption. The basic scheme thus consists of an absorption and a desorption stage with the absorbent cycling between them as it takes up and again gives off the absorbate (Fig. 7).

3.2. Methods of Desorption

Desorption (→ Gas Production) is the reverse process of absorption in which the absorbate is again removed from the solvent. Desorption is achieved with the following methods:

1) Flashing the solvent to lower the partial pressure of the dissolved components, to vacuum if necessary;
2) Reboiling the solvent, to generate stripping vapor by evaporation of part of the solvent;
3) Stripping with an (extraneous) inert gas;
4) A combination of methods 1 to 3.

Organic solvents can also be dried during desorption, when the water vapor absorbed with the feed gas is again given off. This leads to a dry clean gas with reduced vapor dew point.

Desorption occurs by the same basic mechanism as absorption. Each reduction in partial pressure of the absorbed component over the loaded solvent produces a desorption effect. Wide use is made of desorption by *pressure reduction* because the energy requirements are low and cover energy can even be recovered. In the second regeneration step after depressurization, stripping with inert gas is often more economic than thermal regeneration.

The desired purity of the gas determines the final cost of desorption. The necessary difference between the partial pressure of the absorbed

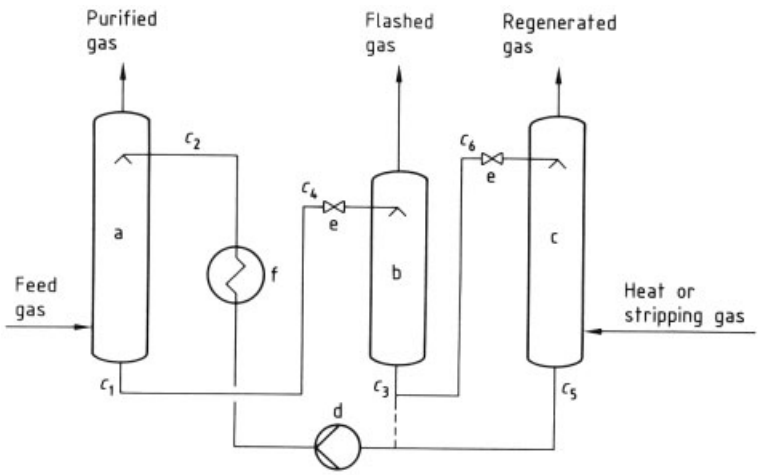

Figure 7. Schematic of absorption installation with desorption for regeneration
a) Absorber; b) Depressurization (flash) stage; c) Desorber (regenerator); d) Pump; e) Reducing valve; f) Cooler

key componentover the regenerated solution (1) and in the purified gas (2) serves as a criterion for determining the dimensions of the absorption and desorption equipment.

3.3. Absorption and Desorption Equipment

Because reactions and conversions between gaseous and liquid phases are substantially slower than those between equal states of aggregation, relatively large reaction volumes are required in gas absorption installations. Mechanical accessories which create the largest possible phase interface serve to improve mass exchange in sorption equipment. Relatively large gas and liquid velocities produce high turbulences as the hydraulic flooding point is approached.

Appropriate internals direct the liquid and gas streams and increase the contact area between the two phases, thereby substantially improving mass transfer. Various designs have proved suitable, especially absorption plates, randomly poured packings and structured packings (→ Distillation and Rectification, Chap. 8.3.).

The internals must be designed to withstand high hydraulic loads, produce minimum pressure loss and achieve good mass transfer with high separation efficiency. Various types of internals are described in [126–133].

3.4. Columns with Mass-Transfer Plates

Columns with mass-transfer plates have proved to be as suitable in absorption as in distillation (→ Distillation and Rectification, Chap. 8.1.3.). In comparison with most other internals, they are distinguished by high hydraulic flexibility, especially under extremely low load. Common characteristics of the most diverse designs of such plates are overflow weirs and downcomers for hydrostatic holdup of the liquid. Another type are dynamic plates without downcomers, holding up the liquid by the kinetic energy of the rising gases.

Valve plates (*ballast plates*) are widely used [130]. The *Varioflex Plate* (Stahl) has proved itself in practice. It is distinguished particularly by its lower hydraulic loading limit of 20 % and its very small pressure drop. High flexibility is achieved by the movable plate inside the fixed cap, whose lifting motion is always synchronized with the gas loading. In more recent designs the valve tray is perforated so that an even wider loading range can be achieved [131].

The *Cross-Stream Plate* (Montz), a valve plate without moving parts, is intended for especially high gas loading. The rectangular valves are arranged so that the gas streams intersect with neighboring streams and head-on collisions of opposing streams are thus prevented. In this manner, the otherwise inevitable upward entrainment of liquid is avoided [131].

Sieve plates are suitable for high gas loadings. They are needed to prevent liquid weeping through the plate (→ Distillation and Rectification, Chap. 8.2.). Good mass transfer occurs due to the bubble layer formation on the plate. The sieve plate can also be designed to have several downcomers. Except for special low-temperature separation plants, it has thus far not been widely used in absorption technology.

The *Kittel Polygonal Plate* (Stahl) is especially suited for high liquid loadings. The downcomers can be positioned inside or outside the plate. Some types have an internal tube in the center of the plate which is used to transport liquid or gas [132].

The *Performkontact Plate* (CLG) is also a cross-flow plate with downcomers for high liquid loading. Its active surface consists of expanded metal arranged so that the gas slots in the mesh are at an acute angle to the flow direction of liquid flows [133], [134].

3.5. Columns with Random Packing

Packings can take high liquid loads and have a low pressure drop. In addition to the well-known and well-tested Raschig rings, Pall rings, Berl saddles and Intalox packings [135–139] (→ Distillation and Rectification, Chap. 8.3.) new developments contribute to further reduction of the pressure drop occurring with increased mass transfer. Thus, for example, *Cascade mini rings* (Mass Transfer, Inc.) are used in a wide range of applications and excel by virtue of high performance [140].

Intalox metal saddles also have special advantages [141–144]. *Super-intalox* (R) packings of plastic and ceramic materials (Norton) are used often in practice and are distinguished by high capacities and good separation performance [145]. Compared to steel packings, thermoplastic packings are not only lighter but frequently also more cost-effective. In contrast to ceramic packings, they do not exhibit abrasion. Stabilizing additives ensure that their temperature resistance is relatively high. Wetting of plastic packings can be improved substantially by hydrophilic surface treatment which eliminates this disadvantage compared with metal packings [146].

The new *VSP packing* (Vereinigte Füllkörper Fabriken) also shows high performance. This packing has better properties than the classical Pall rings and may also be somewhat more cost-effective [148], [149].

3.6. Columns with Structured Packing

Structured packings for columns have come into increasing use. Their efficiency often considerably exceeds that of conventional random packings.

Thus, packings of metal, synthetic woven fabrics and ceramics such as *Mellapack* (metal) and *Kerpak* (ceramics) have been used (Sulzer) (→ Distillation and Rectification). Existing columns can be retrofitted with such packings to increase their capacity [150], [151].

Special requirements tailored to specific separation problems can be met with *Pyrapack* packings G and F [152], [153]. The elements are fabricated from expanded metal (CLG). Type F packing has a greater specific surface and thus achieves a high specific separation performance. Variant G is coarser and is mainly used for high gas and liquid loading.

Increased packing surfaces in connection with additional turbulences in the liquid film are achieved with *Rombapak* (Kühni) by means of finely grooved surfaces. The distribution of liquid is significantly improved, and a separation performance of 3.5 – 4 separation stages per meter of packing height is achieved [154].

3.7. Columns with Special Internals

Atomizing absorbents in spraying devices, e.g., through nozzles, produces a suspension of finely divided drops in a stream of gas. Such absorption methods are suitable for highly soluble components if resistance to mass transfer occurs primarily in the gas phase. The ratio of absorbent to gas can then often be very small and contact times can be short.

Scrubbers with several series of nozzles and short contact times are successfully employed for these tasks. The size of the droplets and the residence time for a given gas throughput can be adjusted appropriately by a suitable choice of nozzles shapes [118].

Cyclone scrubbers can be used whenever a single absorption stage is sufficient. The liquid drops are injected, entrained by the high-velocity gas, loaded with absorbate and separated at the wall.

This group also includes *Venturi scrubbers* (→ Dust Separation, Chap. 4.2.) in which the liquid is separated from the gas at the outlet. In addition to their original application for separating such particles as dust, mist, or aerosols, they are increasingly used in absorption technology (e.g., of sulfur dioxide in flue gas scrubbers) or in chemisorption (e.g., of hydrogen sulfide in oxidative scrubbers).

However, Venturi scrubbers normally reach only moderate absorption efficiencies of about 95 % and their pressure drop of 2.0 – 5.0 kPa is relatively high.

Absorption equipment with rotating internals to disperse the solvent (e.g., "*High Gee*") has so far not been widely accepted. Practical applications were limited to special cases [155], [156]. Spraying devices generally require too much energy to be cost-effective.

Film scrubbers are used for absorption particularly when large absorption enthalpies have to be removed at the same time, e.g., for the absorption of hydrogen chloride in water. Tube walls on which the absorbent trickles down can easily be cooled with water on the other side. In the case of an "adiabatic absorber" the reaction enthalpy, e.g., from hydrogen chloride absorption, is also removed by water evaporating from the upper end of the tube [114].

The absorption efficiency can be improved by installing "mechanical stirrers" which increase

the path of the gas bubbles and mix the liquid more thoroughly. Stirrer vessels are normally employed for systems in which the resistance to mass transfer occurs in the liquid phase. This may be the case for aeration basins of oxidative scrubbers or open wastewater treatment basins if oxygen has to be injected into the liquid.

4. Design of Absorption Equipment

Graphical and analytical methods are used to design absorption equipment [157]. Usually, a material and heat balance is drawn up first. The absorbed component in each case should preferably be referred to the pure absorbent at a predetermined temperature (e.g., 20 °C). In this way, stationary mass flows are defined.

4.1. Physical Absorption in Plate Columns

For component i to be absorbed, the material balance for the absorber with cross-section a – a (see Fig. 8) is, for the lower section:

$$G_1 \cdot y_1 + L \cdot c = G \cdot y + L \cdot c_1 \tag{162}$$

for the upper section:

$$G \cdot y + L \cdot c_2 = G_2 \cdot y_2 + L \cdot c \tag{163}$$

with:

$$G\,(1-y) = G_1\,(1-y_1) = G_2\,(1-y_2) \tag{164}$$

where G_1 is the flow rate of gas entering the absorber; G_2 the flow rate of gas leaving the absorber; y_1 the gas phase mole fraction of absorbate entering the absorber; y_2 the gas phase mole fraction of absorbate leaving the absorber at the heat; L the flow rate of the solvent; c_1 the concentration of absorbate leaving the absorber at the bottom; and c_2 the residual load of absorbate in the regenerated solvent entering the absorber.

Solving for c, the concentration after loading, gives:

$$c = c_2 + \frac{G_1}{L} \cdot \frac{1-y_1}{1-y_2} \cdot \frac{y-y_2}{1-y} \tag{165}$$

with

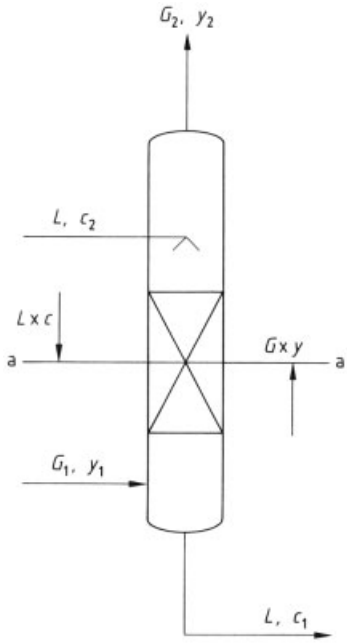

Figure 8. Absorption by a solvent in an absorption column

$$\frac{G_1}{L} \cdot \frac{1-y_1}{1-y_2} = \text{constant},$$

the balance or operating line of the absorber is

$$c = c_2 + \text{const.}\,\frac{y-y_2}{1-y} \tag{166}$$

If the operating line and the equilibrium line are plotted on a concentration-loading diagram the curves shown in Figure 9 are obtained. For constant streams of liquid and gas, the balance line is a straight line. The equilibrium line is obtained by the method described in Chapter 2.

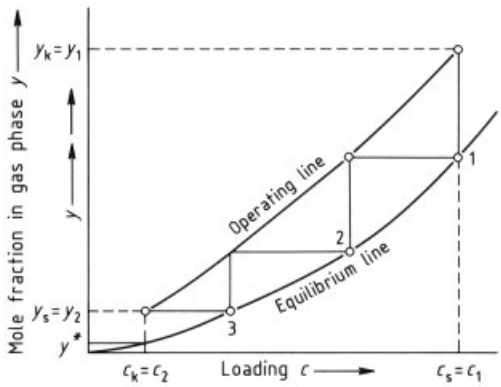

Figure 9. Loading diagram for determining theoretical number of separation stages for isothermal absorption

The equilibrium line in Figure 9 is plotted for a constant temperature, which is permissible only for isothermal absorption. With the aid of a set of steps drawn between the equilibrium line and the operating line from y_1 to y_2, the number of theoretical separation stages can be determined. The number of theoretical separation stages depends on the absorbent flow rate L. The less absorbent enters the absorber, the smaller will be the slope of the operating line and the more stages are needed for a given separation problem. If the operating line and the equilibrium line intersect, the number of separation stages is infinite. To increase the slope of the operating line it is necessary to increase the absorbent flow rate.

A theoretical separation stage is characterized by:

1) perfect phase equilibrium between gas and absorbent,
2) complete mixing of the liquid on the plate of the separation stage, the concentration of liquid above the plate being constant, and
3) flow of gas to the next plate without entrained absorbent droplets.

Practically, phase equilibrium is never reached completely with the plates in commercial installations. Also, droplets are frequently entrained with the gas and a concentration gradient occurs above the plate. The combined effect of these "imperfections" is reflected by the degree of plate efficiency.

Although different values are obtained, depending on the principle of plate construction and the pairing of materials, the degree of efficiency in absorption technology is generally about 18 – 25 % for poorly soluble components and 30 – 40 % for readily soluble components. This is clearly below the degrees of efficiency attainable in distillation technology (→ Distillation and Rectification).

Once the number of theoretical separation stages is known, the practical number of plates has to be defined for each theoretical stage, using this degree of efficiency. The practical number of absorption plates required is extremely important for the economy of an absorption process. Therefore, an optimum balance must be achieved between excess of absorbent and minimum number of plates in the column.

The minimum amount of absorbent required is obtained from the slope of the operating line for an infinite number of theoretical separation stages. It is given by the ratio of absorbent to the total gas rate:

$$\frac{L_{\min}}{G}=\frac{y_2-y_1}{c_2-c_1}=\frac{\Delta y}{\Delta c}$$

$$L_{\min}=\frac{G}{\alpha_i \cdot P_{\mathrm{g}}} \tag{168}$$

where α_i is the solubility coefficient of component i and

$$c_{\mathrm{i}}=\alpha_i \cdot p_i=\alpha_i \cdot y_i \cdot P_{\mathrm{g}} \tag{169}$$

To determine the true amount of absorbent, $L_{\min}$ is multiplied by an absorption factor A for the excess of solvent

$$L=A \cdot L_{\min} \tag{170}$$

The absorption factor A must always be chosen such that the operating line is steeper than the equilibrium line, i.e., it must be greater than 1. If the equilibrium line is a straight line, the absorption factor is the ratio of the slope of the operating line to that of the equilibrium line. For a mixture of gases, the theoretical stage number can be determined graphically for each individual component to be absorbed. When this is done, different numbers of required theoretical stages are usually obtained. In such a case, A is calculated by using the critical component for the absorption process (key component).

Next, the rest content of each individual component at the head of the absorber is determined with this selected absorption factor A and the appropriate equilibrium line, starting with the loading at the bottom. If the purity requirements for a component are not met, the operating line and/or the absorption factor A have to be changed.

The design of columns for the absorption of multiple components can be calculated by a number of methods, largely supported by computers. Plate-to-plate calculations have been successfully carried out with variable gas and liquid streams and with variable absorption and desorption factors [158–161].

Concepts which have been introduced to good advantage in absorption technology are the *absorption rate* E_{A} and the *absorption coefficient* φ_{A} (see Figs. 10 and 11, next two pages).

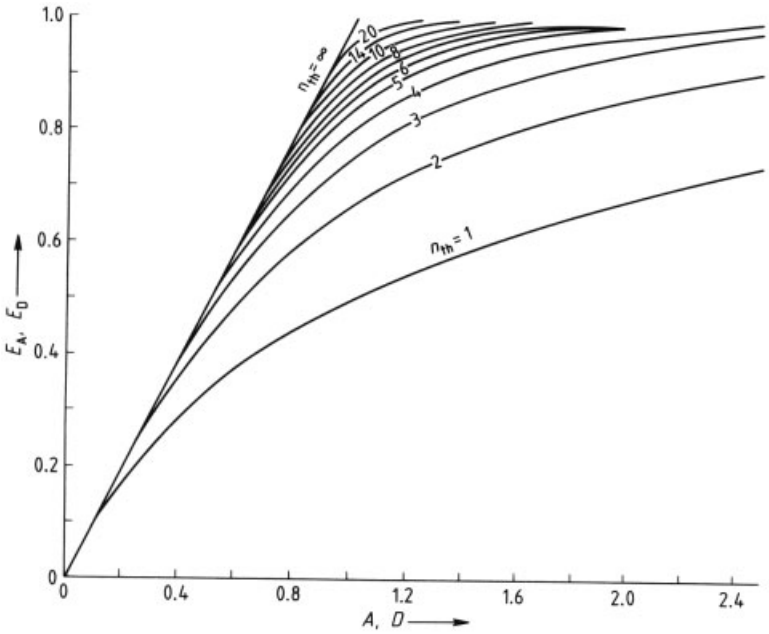

Figure 10. Degree of absorption or desorption as functions of the absorption and desorption factors A and D, respectively, and of the theoretical plate number n_{th} according to Equation (177)

The (percent) absorption coefficient φ_{A} of a component i from a gas which can be attained in case of incomplete regeneration is defined as

$$\varphi_{\mathrm{A}}=\frac{y_2-y_2^*}{y_1-y_2^*}=\frac{y_{\mathrm{h}}-y_{\mathrm{h}}^*}{y_{\mathrm{s}}-y_{\mathrm{h}}^*} \tag{171}$$

Substituting the relation

$$y_{\mathrm{h}}^*=\frac{c_{\mathrm{h}}^*}{\alpha_i^*\cdot P_{\mathrm{g}}} \tag{172}$$

for the equilibrium content of the gas phase $y*_{\mathrm{h}}=y*_2$ and designating the slope of the equilibrium line as

$$\frac{1}{\alpha_i^*\cdot P_{\mathrm{g}}}=m \tag{173}$$

the absorption rate E_{A} is obtained as

$$E_{\mathrm{A}}=\frac{y_1-y_2}{y_1-m\cdot C_2^*}=\frac{y_{\mathrm{s}}-y_{\mathrm{h}}}{y_{\mathrm{s}}-m\cdot c_{\mathrm{h}}^*} \tag{174}$$

If the amount of component i in the absorber head approaches the equilibrium value, than

$$y_{2i}=y_{2i}^* \text{ or } y_{\mathrm{h}i}=y_{\mathrm{h}i}^* \tag{175}$$

the absorption rate $E_{\mathrm{A}}=1$, i.e., the maximum possible amount of component i has been absorbed from the gas. The absorption factor A, defined as the ratio of the slopes between the operating line and the equilibrium curve, can also be written as

$$A=\frac{y_1-y_2}{y_1^*-y_2^*}=\frac{y_{\mathrm{s}}-y_{\mathrm{h}}}{y_{\mathrm{s}}^*-y_{\mathrm{h}}^*} \tag{176}$$

Combining the relations for the absorption rate (Eq. 174) and the absorption factor (176) for arbitrary numbers of theoretical stages n_{th} yields [158]

$$E_{\mathrm{A}}=\frac{A^{n_{\mathrm{th}}+1}-A}{A^{n_{\mathrm{th}}+1}-1} \tag{177}$$

Equation (177) is plotted in Figure 10 for various theoretical stage numbers between 1 and 20. The limiting tangent of this family of curves is the line for infinitely large theoretical stage numbers.

With the aid of Figure 10, the required theoretical plate numbers can be determined both for absorption, by using the absorption factor A, and for desorption, by using the desorption factor D, if

1) gas and liquid loading in the column are constant,

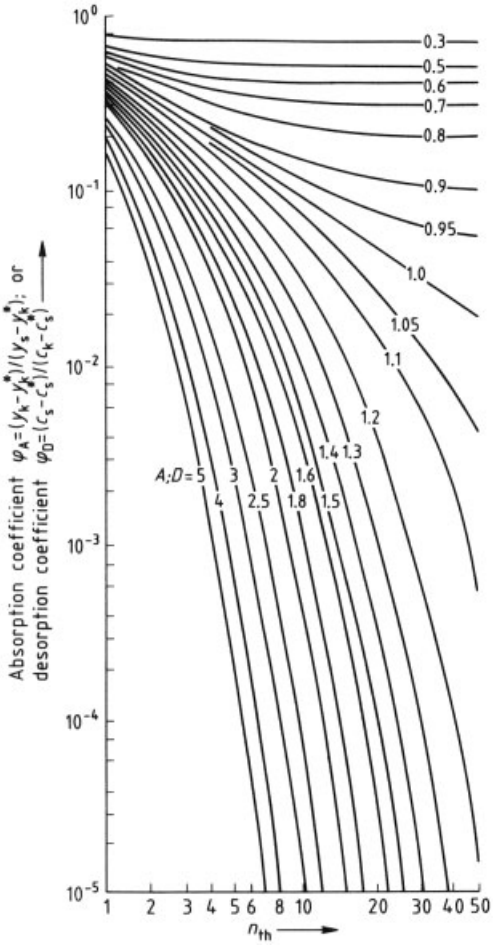

Figure 11. Absorption and desorption coefficient φ_A and φ_D as a function of the theoretical stage number n_{th} depending on the absorption and desorption factor *A* and *D* (Eq. 179, 180)

2) absorption or desorption is isothermal,
3) the absorbate is present at low concentrations, and
4) the solubility coefficient α_i is constant.

Equation (178) shows the dependence of the theoretical stage number n_{th} on the absorption factor *A*:

$$n_{th}=\frac{\ln\left[\frac{y_1-y_2^*}{y_2-y_2^*}\cdot\left(\frac{A-1}{A}\right)+\frac{1}{A}\right]}{\ln A} \tag{178}$$

Here, $y*_2$ is the equilibrium concentration of the gas phase pertaining to the liquid with residual loading c_2 in the absorber head. Analogous to the absorption factor *A*, the desorption or stripping factor *D* can be used to determine the theoretical stage number n_{th} of the desorber.

By relating the difference between actual and equilibrium concentration in the absorber head for the gas phase (y_2-y_2*) and for the liquid phase (c_2-c*_2) to the change in concentration at the bottom of the absorber for the gas phase (y_1-y*_1) and for the liquid phase (c_1-c*_1), the following relation is obtained for the absorption coefficient:

$$\varphi_A=\frac{y_2-y_2^*}{y_1-y_2^*}=\frac{y_2-m\cdot c_2^*}{y_1-m\cdot c_2^*}$$

or

$$\varphi_A=\frac{y_h-y_h^*}{y_s-y_h^*}=\frac{y_h-m\cdot c_h^*}{y_s-m\cdot c_h^*} \tag{179}$$

Analogously, for desorption

$$\varphi_D=\frac{c_1-c_1^*}{c_2-c_1^*}=\frac{c_1-\frac{1}{m}\cdot y_1^*}{c_2-\frac{1}{m}\cdot y_1^*}$$

or

$$\varphi_D=\frac{c_s-c_s^*}{c_h-c_s^*}=\frac{c_s-\frac{1}{m}\cdot y_s^*}{c_h-\frac{1}{m}\cdot y_s^*} \tag{180}$$

This correlation is illustrated graphically in Figure 11 using φ_A and φ_D respectively as ordinate and the theoretical stage number n_{th} as abscissa and the absorption or desorption factor *A, D* as parameter. The assumptions already mentioned for Figure 10 apply also to this diagram.

4.2. Chemisorption in Packed Columns

Chemisorption normally uses columns with packings or similar internals. Mass transfer occurs along the trickling film of the packing and calculating the required column height becomes difficult. Hence, packed columns are often calculated according ot the concept of theoretical separation steps. However, experimental data are required as a rule to get the true number of theoretical separation stages per meter of packing height for given packings and material systems.

Colburn [162] introduced a method in which the region of separation is subdivided into a number of consecutive mass exchange units, the so-called number of transfer units (NTU). The sum of the transfer units is called N_{tot}. The transfer unit *N* must be related either to the gas phase or to the liquid phase, N_G or N_F, respectively.

The transfer unit N is a pure thermodynamic quantity representing the driving force of the equilibrium process. It is a measure of the deviation from thermodynamic equilibrium and depends on mass transfer.

A simple relation for determining the sum of the transfer units in the gaseous phase $N_{\mathrm{G,tot}}$ is obtained by analogy to the corresponding relation (178) for plate columns. For absorption

$$N_{\mathrm{G,tot}} = \frac{A \cdot \ln\left[\frac{y_1 - y_2^*}{y_2 - y_2^*}\left(\frac{A-1}{A}\right) + \frac{1}{A}\right]}{(A-1)}$$

or

$$N_{\mathrm{G,tot}} = \frac{A \cdot \ln\left[\frac{y_{\mathrm{s}} - y_{\mathrm{h}}^*}{y_{\mathrm{h}} - y_{\mathrm{h}}^*}\left(\frac{A-1}{A}\right) + \frac{1}{A}\right]}{(A-1)} \tag{181}$$

A corresponding relation can be derived for desorption as a function of loading c.

Figure 12 shows the absorption coefficient φ_{A} or the desorption coefficient φ_{D}, respectively, plotted as a function of $N_{\mathrm{G,tot}}$. The absorption and desorption rates are defined analogous to those for plate columns (Fig. 10). Other calculation methods for determining transfer units will be found in the literature [163].

In addition to the number of required (theoretical) transfer units N_{tot}, another practical equivalence value corresponding to the transfer unit $N_{\mathrm{G,tot}} = 1$ is needed for the calculation of packed columns. This is the *height of transfer unit* (HTU) which can be referred to either the gas or the liquid phase. The practical total packing height h of a packed column is then

$$h = h_{\mathrm{G,tot}} \cdot N_{\mathrm{G,tot}} = h_{\mathrm{F,tot}} \cdot N_{\mathrm{F,tot}} \tag{182}$$

The transfer units $h_{\mathrm{G,tot}}$ or $h_{\mathrm{F,tot}}$ are best determined on actual working columns. If such practical data are not available, packed columns can also be calculated by a number of methods found in the literature [157].

Nevertheless, despite the increasing accuracy of these often highly complex calculations, it has always been necessary to subject them to a practical test.

4.3. Nonisothermal Absorption

Only if the component to be absorbed from the gas mixture (key component) is present in low concentrations will the thermal effect be so small that the absorption can be regarded as being virtually isothermal. If the key component is present in larger amounts, the solvent will be heated by taking up enthalpy of absorption because physical processes of absorption are as a rule exothermic. Part of the enthalpy of absorption is then transferred to the gas.

In chemisorption, this effect is produced by the enthalpy of the reaction which, as a rule, entails greater heat effect. This temperature increase has such a strong effect on the equilibria along the height of the column that it must be taken into account in the calculation.

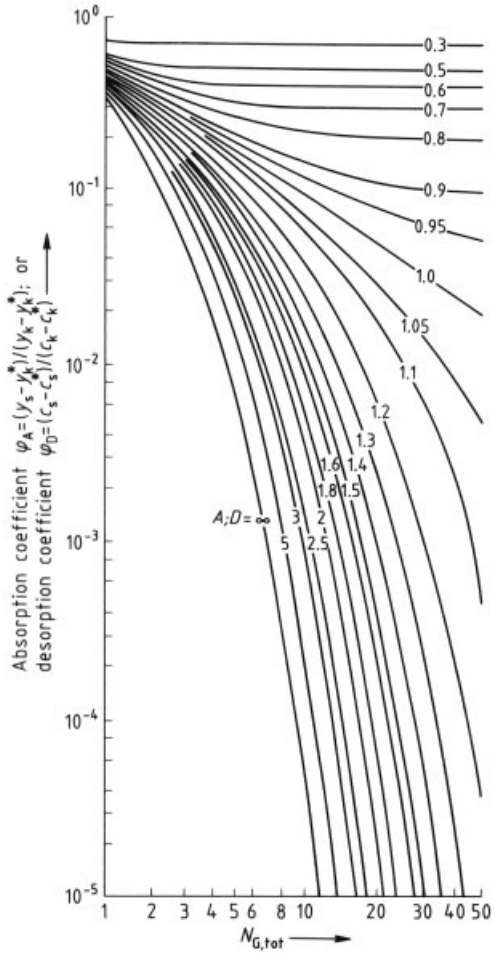

Figure 12. Absorption and desorption coefficient φ_{A} and φ_{D} as a function of the number of transfer units $N_{\mathrm{G,tot}}$ with absorption and desorption factor A and D as parameters (Eq. 181)

The enthalpy of absorption is approximately equal to the enthalpy of condensation only for components of the gas present in sub-critical concentrations. For super-critical components it must be calculated from known equilibrium values (see Section 1.6).

In nonideal mixtures, the enthalpy of mixing (excess enthalpy) appears in addition to the enthalpy of condensation. The total enthalpy of absorption consists of the enthalpy of condensation and the differential enthalpy of solution.

The enthalpy of absorption depends on the concentration but often only slightly on temperature. To determine the equilibrium line for nonisothermal absorption, a temperature profile along the column must be given or calculated. In the simplest case, a linear relation with the loading of the solvent is assumed:

$$\frac{\Delta c}{\Delta t} = \text{const.} \tag{183}$$

To determine the equilibrium line for constructing the McCabe – Thiele diagram (→ Distillation and Rectification,, Chap. 3.1.1.), the equilibria at different temperatures must be known.

For given loadings and known average enthalpies of solution and specific heats, the temperatures of the solvent at the various equilibria can then be determined from the material and heat balances for adiabatic absorption. A characteristic of the equilibrium line is then obtained as shown in Figure 13.

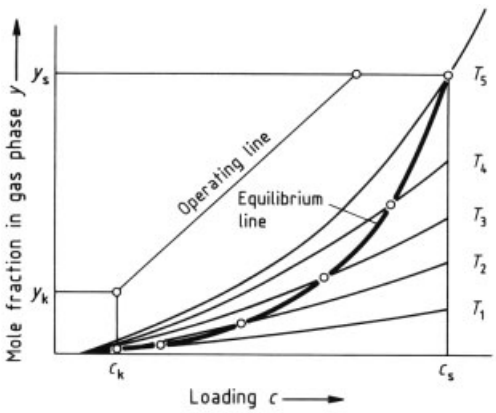

Figure 13. Equilibrium curve for nonisothermal absorption superimposed on equilibrium curves at constant temperatures T_1 to T_5

5. Design of Desorption Equipment

[164–169]

5.1. Physical Desorption in Flash Columns

When the total pressure of physical solvents containing dissolved gases is reduced, the gases desorb until a new equilibrium is established. This is by far the most economical method for regenerating loaded solvents.

This pressure reduction occurs predominantly isenthalpic, i.e., without performing work, by way of butterfly valves. If the quantities of solvent are large and the pressure differential great, using flash turbines, that perform work (Francis or Pelton turbines, or even a reverse running pump) is recommendable.

To calculate the amount of gas that has to be flashed per stage in order to reach a specified pressure it may generally be assumed that the rate of desorbed gas is in equilibrium with the regenerated solvent. A flash stage thus corresponds to a theoretical separation stage.

The new equilibrium at the reduced pressure is again calculated via the material and enthalpy balance for the dissolved component *i*, with the nomenclature used in Figure 14.

$$L \cdot c_4 = L \cdot c_3 + y_4 \cdot G_4$$
$$L \cdot (c_4 - c_3) = y_4 \cdot G_4 \tag{184}$$

where c_4 is the concentration of component *i* in the absorbent entering the flash column, c_3 is the concentration of *i* in the absorbent leaving the column, G_4 is the flow rate of flash gas, and y_4 is the mole fraction of component *i* in the flash gas with

$$\gamma = \frac{G}{L} \text{ (relative gas volume)} \tag{185}$$

then

$$y_4 = \frac{c_4 - c_3}{\gamma} = \frac{c_4}{\gamma} - \frac{c_3}{\gamma} \tag{186}$$

The absorption equilibrium after flashing is:

$$c_3 = \alpha_i^* \cdot p_3^* \tag{187}$$

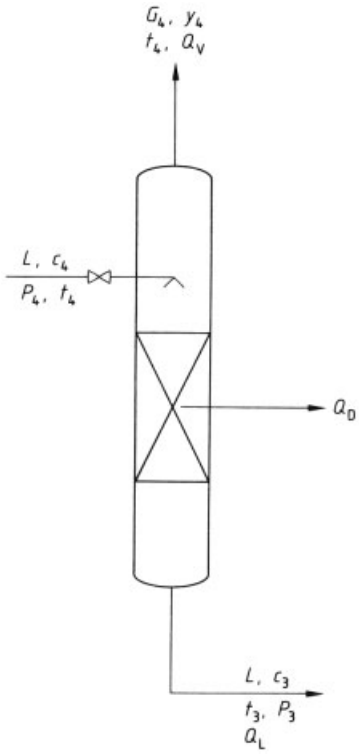

Figure 14. Isenthalpic depressurization of a loaded solvent in a flash column

By combining equations (186) and (187), the concentration of component i in the flashed gas is:

$$y_4 = \frac{c_3}{\gamma} - \frac{\alpha_i^*(t_4) \cdot P_4^*}{\gamma} \tag{188}$$

Strictly speaking, this relationship is valid only for ideal gas behavior, as is usually the case at reduced flash pressure. At higher total pressure the real gas behavior must be taken into account by introducing fugacity coefficients (see Chap. 1).

Equation (177) serves for iterative flash calculations. The flash temperature t_3 and the flash concentration y_4 must initially be estimated for each component in a first pass. Here, the gas volume referred to L must be chosen such that the boundary condition

$$\Sigma y_i = 1 \tag{189}$$

is met for all desorbed components. For the enthalpy balance about the flash stage

$$Q_{\mathrm{L},4} = Q_{\mathrm{G},4} + Q_{\mathrm{L},3} + Q_{\mathrm{D}} \tag{190}$$

where Q_{D} is the enthalpy of desorption, or $L \cdot c_{\mathrm{p,\,m}} \cdot t_4 = G_4 \cdot c_{\mathrm{p,\,m}} \cdot t_4 + L \cdot c_{\mathrm{p,\,m}} \cdot t_3 + Q_{\mathrm{D}}$ The flash temperature t_4 can be eliminated by combination with Equation (185) to obtain

$$t_4 = t_3 = \frac{C_{\mathrm{P,m}} \cdot t_4 - \frac{Q_{\mathrm{D}}}{L}}{v \cdot C_{\mathrm{P,m,G}} + C_{\mathrm{P,m,L}}} \tag{191}$$

Equation (180) provides a relation between the flash temperature $t_4 = t_3$ and the flow rate of flash gas G_4 at a given flash pressure. For physical systems, the enthalpy of desorption Q_{D} is generally not very temperature-dependent. Equations (188) and (191) are conveniently solved by iteration on computers. In doing so, the time required for the calculation can be reduced by using suitable initial values from experience with known material systems.

If the solvent is desorbed by flashing to atmospheric pressure, the partial pressure of the key component that can be attained is about 0.11 to 0.12 MPa. With direct recycling to the absorber, this permits only moderate gas purities so that this method is adequate only for bulk removal of gaseous components or as a prestage for fine regeneration. Another flash stage operating under vacuum may be provided in order to further reduce this partial pressure. Further reduction of the partial pressure can be achieved also by bringing the absorbent into contact with an inert gas. This is more economical even in most cases. Also a combination of vacuum regeneration and stripping with inert gas is frequently a good method.

5.2. Physical Desorption in Stripper Columns

Desorption of a loaded absorbent by stripping with an inert gas has the advantage that the absorbent usually does not have to be heated or evacuated. This method of regeneration can be used also at elevated temperatures, thereby reducing the amount of stripping gas required.

To design such a stripper column, the operating line is again determined from a balance in addition to the equilibrium line. The material balance for the stripper column is calculated in the cross-section a – a according to Figure 15

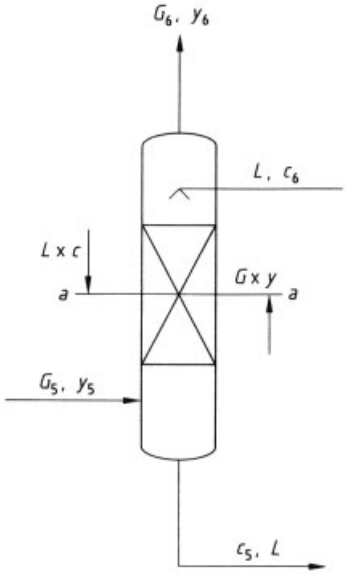

Figure 15. Desorption of a loaded solvent by stripping in a stripper column

for a given component *i*, by following the same principle as for absorption (Chap. 4).

For the *lower part*

$$L \cdot c + G_5 \cdot y_5 = L \cdot c_5 + G \cdot y \tag{192}$$

and for the *upper part*

$$L \cdot c_6 + G \cdot y = L \cdot c + G_6 \cdot y_6 \tag{193}$$

where G_5 is the flow rate of strip gas entering the stripper column, G_6 is the flow rate of gas leaving the stripper column, y_5 is the mole fraction of desorbed component *i* in the stripping gas, y_6 is the mole fraction of *i* in the gas leaving the stripper column, c_5 is the concentration of *i* in the absorbent leaving the stripper column, and c_6 is the concentration of *i* in the absorbent entering the column. The stripping gas balance in the lower part is

$$G\,(1-y) = G_5\,(1-y_5) \tag{194}$$

For each loading *c*, the operating line of the desorber is obtained as a function of the mole fraction *y* of the key component in the stripping gas

$$c = c_5 - \frac{y_5 \cdot G_5}{L} + \frac{y \cdot G_5}{L} \cdot \left[\frac{1-y_5}{1-y}\right] \tag{195}$$

If a pure inert stripping gas is used that does not contain fractions of the component to be desorbed $y_5 = 0$ and equation (195) simplifies to

$$c = c_5 + \frac{G_5}{L}\left[\frac{y}{1-y}\right] \tag{196}$$

Plotting the operating line determined according to Equation (184) on a concentration-loading diagram leads to the curve shown in Figure 16.

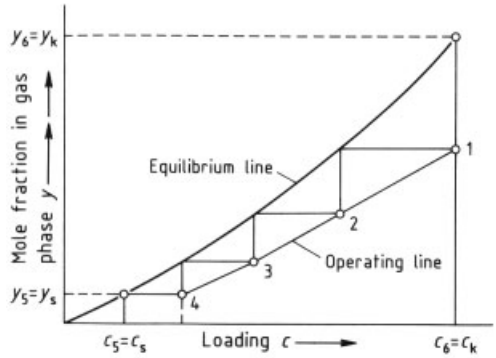

Figure 16. Loading diagram for determining the theoretical number of separation stages for isothermal desorption by stripping with inert gas

For constant liquid and gas streams, the balance or operating line for desorption, as for absorption, is a straight line. However, the operating line for desorption is below the equilibrium line. By using the same method employed in absorption, the number of theoretical separation stages required for nonisothermal desorption in a stripper column is obtained by drawing a series of steps between y_{h} and y_{s}. The theoretical number of separation stages for isothermal desorption can also be determined according to [158].

From the ratio of the slopes between the operating and the equilibrium line, the stripping factor *D* can be determined as shown in Section 4.1. Its relation to the absorption factor is

$$A = \frac{1}{D} \tag{197}$$

The isothermal desorption conditions represented in Figure 16 apply, strictly speaking, only to the low-loading common when stripping with inert gas. At high loading and without the use of an inert stripping fluid, desorption proceeds like distillation if the stripping vapor is generated by boiling and evaporation of the solvent.

5.3. Physical Desorption in Reboiler Columns

Here, nonisothermal desorption occurs by boiling the solvent at the bottom of the column. Temperature increases from the solvent inlet to the bottom of the column where it reaches the boiling point of the solvent. The component to be driven off, as a rule, boils at a lower temperature than the solvent. The rising vapor stream, consisting mostly of solvent, acts as a stripping fluid, supported, if necessary, by simultaneously desorbed inert gases.

Thermodynamically, this type of desorption is a distillation or rectification process (→ Distillation and Rectification). It is the most energy-consuming method for the regeneration of a loaded solvent.

To desorb dissolved hydrocarbons, for example, superheated (live) steam is often blown directly into the bottom of the column, producing an additional stripping effect that lowers the partial pressure similar to vacuum regeneration.

If the solvent has a high boiling point (a so-called high-boiler), boiling under vacuum is often required to attain high degrees of regeneration. If compared to stripping with inert gas, this method has the advantage that the component to be desorbed, usually hydrogen sulfide, is not unnecessarily diluted and is easier to recover in subsequent process steps (e.g., Claus units → Sulfur).

The design of reboiler columns follows the same principles as in distillation (→ Distillation and Rectification). Here, too, the use of equilibrium constants with K_i for the key component to be driven off and K_L for the solvent is advantageous.

For a binary mixture, the following equations apply to the key component *i* to be desorbed and the solvent L (see also Chapter 1):

$$K_i = \frac{y_i}{x_i} = \frac{y_i}{c_i}; K_L = \frac{y_L}{x_L} = \frac{y_L}{c_L}$$

$$\frac{K_i}{K_L} = \frac{y_i}{c_i} \cdot \frac{c_L}{y_L} = \frac{y_i\,(1-c_i)}{c_i\,(1-y_i)} \tag{198}$$

$$K_i\,K_L = \frac{P_i^s}{P_L^s} = \beta \tag{199}$$

$$y_i = \frac{\beta \cdot c_i}{1 + (\beta - 1) \cdot c_i} \tag{200}$$

In an equilibrium diagram, Equation (189) is a hyperbola (Fig. 17).

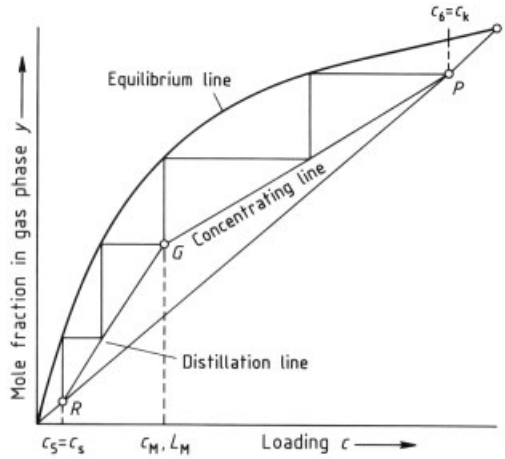

Figure 17. Loading diagram for determining the theoretical number of separation stages in nonisothermal desorption by reboiling

As in distillation, the reboiler column may consist of a distilling and a concentration section. The solvent loaded with load C_M is fed to the head of the column. In this case, the reboiler column consists only of a distillation section, the operating line being the distillation line R – G.

Only when the vapor pressures p_i^s of the components to be distilled and of the solvent p_L^s are very close, i.e., when β is relatively small, will a concentrating section G – P be needed. This section will then be placed above the solvent inlet. With the solvents used today, this is generally not necessary.

The theoretical number of separation stages in desorption is determined graphically by constructing a sequence of steps between head and bottom. The vapor rate to be produced by the reboiler in the column bottom is defined by the slope of the operating line. The minimum vapor rate for concentrations between $c_5 = c_s$ and c_M is given by the slope of the distillation line R – G.

The gaseous mixture leaving the uppermost plate of the column still contains a considerable amount of solvent vapor which is condensed in an overhead condenser. The condensate is generally used as reflux.

In addition to the production of stripping steam, the heat requirements of the reboiler column consist of the fraction needed to raise the temperature of the solvent if the latter is not supplied to the column at boiling temperature. If any low-boiling component, which desorb eas-

ily, are present in the solvent in addition to the key component, the heat requirements are even higher. In this case, their enthalpy of desorption must be taken into account.

Any higher boiling components in the solvent (e.g., hydrocarbons) are (1) either left in the solvent, because they desorb with difficulty, or (2) considerably increase the steam required for regeneration if they have to be distilled. In addition to the material balances, enthalpy balances are also calculated for the reboiler column in the usual way to finally determine the performance of the reboiler.

The spent solvent leaving the absorber is often cooled further in the subsequent flash regeneration stage. To minimize the energy required for the production of stripping steam, the solvent is heated in heat exchangers in counterflow to the hot solvent leaving the reboiler column. The design of these heat exchangers has to allow particularly for the inevitable degassing of easily desorbable gases. Otherwise, a so-called "back pressure" has to be applied to the solvent by means of an additional pump to avoid degassing of dissolved components.

5.4. Chemical Desorption in Reboiler Columns

Because of the parabolic form of the equilibrium curves in chemisorption, only a relatively small amount of gas can be desorbed from chemical absorbents by flashing. Regeneration occurs by increasing the temperature and reducing the partial pressure in a regenerator. Stripping vapor is normally generated by reboiling the solvent with steam.

Nevertheless, a physical flash stage frequently precedes this thermal regenerator, particularly if the absorption pressure is extremely high and a large fraction of valuable gases such as carbon monoxide, hydrogen and hydrocarbons, is coabsorbed in the water content of the solvent.

Chemically loaded absorbents are thermally regenerated in counterflow columns in the same way that is used for stripping regeneration of physical solvents (see Section 5.3).

Assuming that desorption in the regenerating column is isothermal, and neglecting the fact that the reduction in pressure leads to an increase in temperature, the conditions during chemical desorption are reflected by the loading diagram (Fig. 18). The theoretical number of separation stages is established as previously described by a sequence of steps.

The partial pressure of the component to be expelled is reduced by reboiling the solvent or by directly injecting live steam into the bottom of the regenerating column.

While the stripping vapor rises to the column head, if condenses partly, providing the desorption enthalpy for the components to be expelled. If the temperature of the spent chemical solvent is lower than the boiling temperature in the regenerator bottom, the heat required for heating it to the boiling point in order to produce stripping vapor must be supplied as well.

The steam demand of the reboiler is derived from an accurate regenerator heat balance.

Sufficient stripping vapor must remain at the regenerator head to sufficiently reduce the partial pressure of the component to be expelled relative to the total pressure. The resulting slope of the operating line is generally such that thermal chemical desorption requires a low number of theoretical plates, as shown in Figure 18.

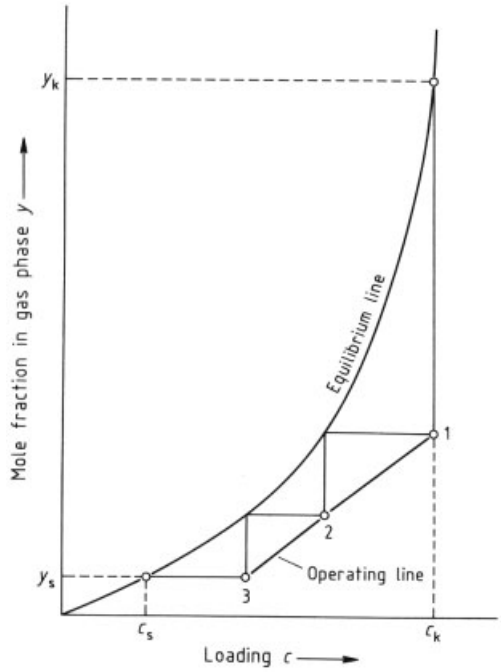

Figure 18. Loading diagram for the isothermal desorption of a chemically loaded solvent by reboiling

Due to the inflection of the equilibrium curve, the difference in partial pressure in the regenerator bottom is considerably less than in the head.

As the stripping vapor rate G_h at the column head is required to make up for the difference between the total pressure p_g and the partial pressure of the key component p_1 as illustrated by point 1 in Fig. 18, the required stripping vapor rate G_h at the column head can be described by

$$G_h = \frac{L \cdot (c_h - c_s) \cdot (P_g - p_1)}{p_1} \tag{201}$$

The stripping vapor rate for the column bottom can then be derived from the regenerator heat balance according to Equation (190). If the residual partial pressure of the expelled component above the regenerated solvent is limited by the vapor requirement of the regenerator, the gas purity increases with higher pressure in the absorber. Chemical regeneration systems provide for heat exchange between the spent and the regenerated solvent in order to reduce the demand for stripping vapor.

Also in chemical desorption systems the stripping effect can be supported by an inert gas. If a dry inert gas is used, it takes up steam until it is saturated. The cooling effect in the regenerator due to this resaturation must be considered in the heat balance.

Whenever live steam is used for stripping, the solvent temperature cannot drop below the condensation temperature because the solvent is consistently reheated by the condensing steam. Stripping with live steam has the advantage that the pressure can be much lower than that of reboiler heating vapor because of the required temperature differential. However, because of an equalized water balance the live steam rate to be used for stripping is limited to that rate that leaves the regenerator and/or absorber together with the gases. Experience showed that the steam demand to regenerate chemically loaded solvents can also be reduced by a split-flow arrangement [114], [115].

6. References

1. S. W. Walas: *Phase Equilibria in Chemical Engineering,* Butterworth, Stoneham, MA 1985.
2. J. M. Prausnitz, T. F. Andersen, E. A. Grens, C. A. Eckert, R. Hsieh, J. P. O'Connell: *Computer Calculations for Multicomponent Vapor-Liquid Equilibria,* Prentice Hall, Englewood Cliffs, N.J. 1980.
3. B. I. Lee, M. G. Kesler: "A Generalized Thermodynamic Correlation Based on Tree-Parameter Corresponding States," *AIChE J.* **21** (1975) 510 – 527.
4. S. Westmeier, W. H. Hanthal, G. Hecht, R. Kümmel, D. Lempe, F. Rösler, M. Rotter, I. Westmeier, W. Wettengel: *"Stoffwerte, Verfahrenstechnische Berechnungsmethoden,"* Part 7, VCH Verlagsgesellschaft, Weinheim, New York 1986.
5. U. Plöcker, H. Knapp, J. M. Prausnitz: "Calculation of High-Pressure Vapour-Liquid Equilibria from a Corresponding States Correlation with Emphasis on Asymmetric Mixtures," *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) 324 – 332.
6. H. Renon, J. M. Prausnitz: "Local Compositions in Thermodynamic Excess Functions for Liquid Mixtures," *AIChe J.* **14** (1968) 135 – 144.
7. J. Gmehling, B. Kolbe; *Angewandte Chemische Thermodynamik für Chemiker, Verfahrenstechniker and Chemieingenieure,* Thieme Verlag, Stuttgart 1988.
8. A. Trendenslund, P. Rasmussen: "From UNIFAC to SUPERFAC and back ?" *Fluid Phase Equilib.* **24** (1985) 115 – 150.
9. Dechema Data Series (1977 ff).
10. N. L. Yarym-Agaev, V. G. Matvienko, N. V. Povalyaeva: "Solubility of Hydrogen Sulfide in N-Methylpyrrolidone," *Zh. Prikl. Khim. (Leningrad)* **53** (1980) 2456 – 2461.
11. *Landolt-Börnstein,* **IV, 4 c,** Springer Verlag Berlin, 1976.
12. *Landolt-Börnstein,* **IV, 42 c,** Springer Verlag Berlin, 1980.
13. A. S. Kertes (ed.): *Solubility Data Series,* Pergamon Press, Oxford 1979 ff.
14. W. Weber, S. Zeck, H. Knapp: "Gas Solubilities in Liquid Solvents at High Pressures: Apparatus and Results for Binary and Ternary Systems of N_2, CO_2 and CH_3OH," *Fluid Phase Equilib.* **18** (1984) 253 – 278.
15. T. Chang, R. W. Rousseau, J. K. Ferrell: "Use of the Soave Modification of the Redlich – Kwong Equation of State for Phase Equilibrium Calculations. System Containing Methanol," *Ind. Eng. Chem. Process Des. Dev.* **22** (1983) 462 – 468.

16. Y. Hu, Y. Xu, J. Prausnitz: "Molecular Thermodynamics of Gas Solubility," *Fluid Phase Equilib.* **23** (1985) 15 – 40.
17. MacCoy: *J. Am. Chem. Soc.* **29** (1903) 437 – 462. (M. S. Litvinenko: "Equilibrium in the System Hydrogen Sulfide and Carbon Dioxide and Solutions of Sodium or Potassium Carbonate," *Zh. Prikl. Khim.* **25** (1952) 579 – 595.)
18. A. Sievers, A. Fritzsche: "Pottasche Lösung und CO_2," *Z. Anorg. Allg. Chem.* **133** (1924) 1 – 16.
19. C. Harte, E. M. Backer, H. H. Purcell, *Ind. Eng. Chem.* **25** (1933) no. 5, 528 – 531. (cit. Ref. [11] p. 376).
20. M. Schingnitz, G. Franke, R. Gall: "Absorptive Entfernung von Kohlendioxid aus Synthesegasen," *Freiberg. Forschungsh. A,* **A 464** (1969) 37 – 53.
21. H. F. Johnstone, "Recovery of Sulfur Dioxide from Waste Gases," *Ind. Eng. Chem.* **27** (1935) 587 – 593.
22. H. F. Johnstone, H. J. Read, H. C. Blankmeyer: "Recovery of Sulfur Dioxide from Waste Gases," *Ind. Eng. Chem.* **30** (1938) 101 – 109.
23. K. Y. Park, T. F. Edgar: "Stimulation of the Hot Carbonate Process for Removal of CO_2 and H_2S from Medium BTU Gas," *Energy Prog.* **4** (1984) no. 3, 174 – 181.
24. J. S. Tosh, J. H. Field, H. E. Benson, W. P. Haynes: "Equilibrium Study of the System Potassium Carbonate, Potassium Bicarbonate, Carbon Dioxide, and Water," *Rep. Invest. U.S. Bur. Mines* **5484** (1959).
25. J. S. Tosh, J. H. Field, H. E. Benson, R. B. Anderson: "Equilibrium Pressures of Hydrogen Sulfide and Carbon Dioxide over Solutions of Potassium Carbonate," *Rep. Invest, U.S. Bur. Mines* **R 5622** (1959),
26. P. Wellman, S. Katell: "Hot Carbonate Purification Computer Program," *Inf. Circ. U.S. Bur. Mines* **8366** (1968).
27. D. W. van Krevelen, P. J. Hoftijzer, F. J. Huntjens: "Composition and Vapor Pressure of Aqueous Solutions of Ammonia, Carbon Dioxode and Hydrogen Sulfide," *Recl. Trav. Chim. Pays Bas* **68** (1949) 191 – 216.
28. G. Maurer: "Electrolyte Solutions," *Fluid Phase Equilib.* **13** (1983) 269 – 296.
29. F. X. Ball, W. Fürst, H. Renon: "An NRTL Model for Representation and Prediction of Deviation from Ideality of Electrolyte Solutions Compared to the Models of Chen (1982) and Pitzer (1973)," *AIChE J.* **3** (1985) no. 3, 392 – 399.
30. E. M. Pawlikowski, J. Newmann, J. M. Prausnitz: "Phase Equilibria for Aqueous Solutions of Ammonia and Carbon Dioxide," *Ind. Eng. Chem. Process Des. Div.* **21** (1982) 764 – 770.
31. R. Stimming, D. Pape, S. Stöck, W. Wettengel, H. Freydank: "Dampf-Flüssigkeit-Gleichgewichte im System Ammoniak-Schwefelwasserstoff-Wasser," *Chem. Tech.* (Leipzig) 37 (1985) no. 1, 27 – 30.
32. B. E. Roberts, P. R. Tremaine: "Vapour Liquid Equilibrium Calculations for Dilute Aqueous Solutions of CO_2, H_2S, NH_3, and NaOH to 300 °C," *Can. J. Chem.* **63** (1985) 294 – 300.
33. A. Newman: "Sour Water Design by Charts", *Hydrocarbon Processing* **70** (1991) no. 9, 145 – 150; **70** (1991) no. 10, 101 – 106; **70** (1991) no. 11, 139 – 141.
34. B. Rumpf, F. Kurz, G. Maurer: "Vapor – Liquid – Solid Phase Equilibria in the System $NH_3 - CO_2 - H_2O$ from Around 310 to 470 K: New Experimental Data and Modeling", *Fluid Phase Equilib.* **104** (1995) 261 – 275.
35. V. Bieling, F. Kurz, B. Rumpf, G. Maurer: "Simultaneous Solubility of Ammonia and Carbon Dioxide in Aqueous Solution of Sodium Sulfate in the Temperature Range 313 – 393 K and Pressures up to 3 MPa", *Ind. Eng. Chem. Res.* **34** (1995) 1449 – 1460.
36. K. L. Kent, B. Eisenberg: "Better Data for Amine Treating," *Hydrocarbon Processing* **55** (1976) no. 2, 87 – 90.
37. C. C. Chen, A. Ng: "Determine Equilibrium Absorption Rates," *Hydrocarbon Process.* **59** (1980) no. 4, 122 – 126.
38. R. A. Tomcej, F. D. Otto: "Computer Simulation and Design of Amine Treating Units," *Energy Process. Can.* (1983) 27 – 32.
39. D. L. Erwin: ChemCalc 11 AMSIM (Amine Gas Treating Plant Simulator) Gulf Publishing Comp. (1987).
40. O. R. Rivas, J. M. Prausnitz: "Sweeting of Sour Natural Gas by Mixed-Solvent Absorption: Solubilities of Ethane, Carbon Dioxide, and Hydrogen Sulfide in Mixtures of Physical and Chemical Solvents," *AIChE J.* **25** (1979) no. 6, 975 – 984.
41. D. Lal, F. D. Otto, A. E. Mather: "Solubility of H_2S and CO_2 in a Diethanolamine Solution at Low Partial Pressures," *Can. J. Chem. Eng.* **63** (1985) 681 – 685.

42. R. H. Weiland, M. Rawal, R. G. Rice: "Stripping of Carbon Dioxide from Monoethanolamine Solutions in a Packed Column," *AIChE J.* **28** (1982) no. 6, 963 – 973.
43. P. V. Danckwerts, McNeil: "The Absorption of Carbon Dioxide into Aqueous Amine Solutions and the Effects of Catalysis," *Trans. Inst. Chem. Eng.* **45** (1967) T32 – T49.
44. G. Sartori, D. W. Savage: "Sterically Hindered Amines for CO_2 Removal from Gases," *Ind. Eng. Chem. Fundam.* **22** (1983) 239 – 249.
45. D. W. Savage, E. W. Funk, W. C. Yu, G. Astarita: "Selective Absorption of H_2S and CO_2 into Aqueous Solutions of Methyldiethanolamine," *Ind. Eng. Chem. Fundam.* **25** (1986) 326 – 330.
46. J. I. Lee, F. D. Otto, A. E. Mather: "Design Data for Diethanolamine Acid Gas Treating Systems," *Gas Processing/Canada* **65** (1973) March-April, 26 – 34.
47. J. I. Lee, F. D. Otto, A. E. Mather: "Solubility of Hydrogen Sulfide in Aqueous Diethanolamine Solutions at High Pressure," *J. Chem. Eng. Data* **18** (1973) 71 – 73.
48. E. E. Isaacs, F. D. Otto, A. E. Mather: "Solubility of Hydrogen Sulfide and Carbon Dioxide in an Aqueous Diisopropanolamine Solution," *J. Chem. Eng. Data* **22** (1977) 71 – 73.
49. E. E. Isaacs, F. D. Otto, A. E. Mather: "The Solubility of Mixtures of Carbon Dioxide and Hydrogen Sulfide in an Aqueous DIPA Solution," *Can. J. Chem. Eng.* **55** (1977) 210 – 212.
50. J. L. Martin, F. D. Otto, A. E. Mather: "Solubility of Hydrogen Sulfide and Carbon Dioxide in a Diglycolamine Solution," *J. Chem. Eng. Data* **23** (1978) 163 – 164.
51. E. E. Isaacs, F. D. Otto, A. E. Mather: "Solubility of Mixtures of H_2S and CO_2 in a Monoethanolamine Solution at Low Partial Pressures," *J. Chem. Eng. Data* **25** (1980) 118 – 120.
52. F.-Y. Jou, F. D. Otto, A. E. Mather: "Solubility of H_2S and CO_2 in Aqueous Methyldiethanolamine Solutions," *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 539 – 544.
53. F.-Y. Jou, F. D. Otto, A. E. Mather: "Equilibria of H_2S and CO_2 in Triethanolamine Solutions," *Can. J. Chem. Eng.* **63** (1985) 122 – 125.
54. M. L. Kennard, A. Meisen: "Solubility of Carbon Dioxide in Aqueous Diethanolamine Solutions at Elevated Temperatures and Pressures," *J. Chem. Eng. Data* **29** (1984) 309 – 312.
55. G. Kuranov, B. Rumpf, N. A. Smirnova, G. Maurer: "Solubility of Single Gases Carbon Dioxide and Hydrogen Sulfide in Aqueous Solutions of A-Methydiethanolamine in the Temperature Range 313 – 413 K at Pressures up to 5 MPa", *Ind. Eng. Chem. Res.* **35** (1996) 1959 – 1966.
56. F.-Y. Jou, J. J. Carroll, A. E. Mather, F. D. Otto: "The Solubility of Carbon Dioxide and Hydrogen Sulfide in a 35 wt% Aqueous Solution of Methydiethanolamine", *Can. J. Chem. Eng.* **71**(1993) 264 – 268.
57. K.-P. Shen, M.-H. Li: "Solubility of Carbon Dioxide in Aqueous Mixtures of Monoethanolamine with Methyldiethanolamine", *J. Chem. Eng. Data* **37** (1994) 96 – 100.
58. F.-Y. Jou, F. D. Otto, A. E. Mather: "Vapor – Liquid Equilibrium of Carbon Dioxide in Aqueous Mixtures of Monoethanolamine and Methydiethanolamine", *Ind. Eng. Chem. Res.* **33** (1994) 2002 – 2005.
59. B. E. Roberts, A. E. Mather: "Solubility of CO_2 and H_2S in a Hindered Amine Solution", *Chem. Eng. Comm.* **64** (1988) 105 – 111.
60. T. T. Teng, A. E. Mather: "Solubility of CO_2, H_2S and their Mixtures in an AMP Solution", *Can. J. Chem. Eng.* **67** (1989) 846 – 850.
61. D. M. Augsten, G. T. Rochelle, X. Peng, C. C. Chen: "Model of Vapor – Liquid Equilibria for Aqueous Acid Gas – Alkanolamine Systems Using the Electrolyte – NRTL Equation", *Ind. Eng. Chem. Res.* **28** (1989) 1060 – 1073.
62. D. M. Augsten, G. T. Rochelle, C. C. Chen: "Model of Vapor – Liquid Equilibria for Aqueous Acid Gas – Alkanolamine Systems. 2. Representation of H_2S and CO_2 Solubility in Aqueous MDEA and CO_2 Solubility in Aqueous Mixture of MDEA with MEA or DEA", *Ind. Eng. Chem. Res.* **30** (1991) 543 – 555.
63. Y. G. Li, A. E. Mather: "Correlation and Prediction of the Solubility of Carbon Dioxide in a Mixed Alkanolamine Solution", *Ind. Eng. Chem. Res.* **33** (1994) 2006 – 2015.
64. Y. G. Li, A. E. Mather: "Correlation and Prediction of the Solubility of CO_2 and H_2S in Aqueous Solutions of Triethanolamine." *Ind. Eng. Chem. Res.* **35** (1996) 4804 – 4809.
65. J. Schlauer: "Eine neue Absorptions-Isotherme zur Korrelation von simultanen Löslichkeitsgleichgewichten von saueren

Gasen in basischen Lösungsmitteln", presented at the GVC meeting, Bamberg, Germany, May 7, 1993.
66. D. Dang, D. P. Tassios: "Prediction of Enthalpies of Mixing with a UNIFAC Model," *Ind. Eng. Chem. Process Des. Dev.* **25** (1986) 22 – 31.

General References
Chapter 2
67. R. H. Perry, D. W. Green, J. O Maloney: Chemical Engineers' Handbook. 5th ed., McGraw-Hill, New York 1984, pp. 14-1 – 14-40.
68. R. B. Bird, W. E. Steward, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1965 ff.
69. G. Astarita: *Mass Transfer with Chemical Reaction* Elsevier, Amsterdam 1967.
70. P. V. Danckwerts: *Gas-Liquid Reactions,* McGraw-Hill, New York 1970.
71. G. Astarita, D. W. Savage, A. L. Bisio: *Gas Treating with Chemical Solvents,* Wiley & Sons, New York 1983.
72. R. Taylor, R. Krishna: *Multicomponent Mass Transfer,* Wiley, New York 1993.

Specific References
73. P. V. Danckwerts, M. M. Sharma: "The Absorption of Carbon Dioxide into Solutions of Alkalis and Amines," *The Chemical Engineer* (1966) 244 – 280.
74. K. Sridharan, M. M. Sharma: "New Systems and Methods for the Measurements of Effective Interfacial Area and Mass Transfer Coefficients in Gas-Liquid Contactors," *Chem. Eng. Sci.* **31** (1976) 767 – 774.
75. G. E. H. Joosten, P. V. Danckwerts: "Chemical Reactions and Effective Interfacial Areas in Gas Absorption," *Chem. Eng. Sci.* **28** (1973) 453 – 461.
76. H. L. Schulman, C. F. Ulbrich, A. Z. Proulx, J. O. Zimmermann: "Performance of Packed Columns III: Wetted and Effective Interfacial Areas, Gas and Liquid Phase Mass Transfer Rates," *AIChE J.* **1** (1955) 253.
77. S. P. S. Andrew: "Aspects of Gas/Liquid Mass Transfer," *Acta Technologia Chimica* (1961) 153.
78. K. Onda, H. Takeuchi, Y. Okumoto: "Mass Transfer Coefficients between Gas and Liquid Phases in Packed Columns," *J. Chem. Eng. Japan* **1** (1968) 56 – 62.
79. L. T. Thuy, R. H. Weiland: "Mechanisms of Gas Desorption from Aqueous Solution," *Ind. Eng. Chem. Fundam.* **15** (1976) 286 – 293.
80. R. H. Weiland, M. Rawal, G. Rice: "Stripping of Carbon Dioxide from Monoethanolamine Solutions in a Packed Column," *AIChE J.* **28** (1982) 963 – 973.
81. R. D. Scheffe, R. H. Weiland: "Mass-Transfer Characteristics of Valve Trays," *Ind. Eng. Chem. Res.* **26** (1987) 228 – 236.
82. B. Pocher, K. Schmok, G. Wünsch: "Experimentelle Untersuchungen zum Studium des CO_2-Stofftransportes bei der Gasreinigung mittels der physikalischen Tieftemperatur-Absorption," *Freiberg. Forschungsh. A,* **A 698** (1984) 5 – 78.
83. M. L. Yüksel, E. U. Schlünder: "Heat and Mass Transfer in Non-Isothermal Absorption of Gases in Falling Liquid Film. Part I and Part II," *Chem. Eng. Process.* **22** (1987) 193 – 213.
84. B. Hörner, U. Viebahn, K. Dialer: "Mass Transfer in Turbulent Liquid during Absorption," *Chem. Eng. Sci.* **41** (1986) 1723 – 1733.
85. H. Hikita, Y. Konishi: "Desorption of Carbon Dioxide from Supersaturated Water in an Agitated Vessel," *AIChE J.* **30** (1984) 945 – 951.
86. D. W. van Krewelen, P. J. Hoftijzer: "Kinetics of Gas-Liquid Reactions. Part I. General Theory," *Rec. Trav. Chim.* **67** (1948) 563.
87. P. V. Danckwerts: "Significance of Liquid-Film Coefficients in Gas Absorption," *Ind. Eng. Chem.* **43** (1951) 1460.
88. P. V. Danckwerts: "The Reaction of CO_2 with Ethanolamines," *Chem. Eng. Sci.* **34** (1979) 443 – 446.
89. B. R. W. Pinsert, L. Pearson, F. J. W. Roughton: "The Kinetics of Combination of CO_2 with OH^-," *Trans. Faraday Soc.* **52** (1956) 1512.
90. M. M. Sharma: "Kinetics of Reaction of Carbonyl Sulfide and Carbon Dioxide with Amines and Catalysis with Brønsted Bases of the Hydrolysis of COS." *Trans. Faraday Soc.* **61** (1965) 681 – 688.
91. H. Hikita, S. Asai, H. Ishikawa, M. Honda: "The Kinetics of Reaction of Carbon Dioxide with Monoethanolamine, Diethanolamine and Triethanolamine by Rapid Mixing Method," *Chem. Eng. J.* (Lausanne) 13 (1977) 7 – 12.
92. W. C. Yu, G. Astarita, D. W. Savage: "Kinetics of Carbon Dioxide Reaction in Solutions of Methyldiethanolamine," *Chem. Eng. Sci.* **40** (1985) 1585 – 1590.

93. G. F. Versteeg: *Mass Transfer and Chemical Reaction Kinetics in Acid Gas Treating Processes,* Twente University Enschede, NL, Thesis 1987.
94. N. Haimour, A. Didarian, O. C. Sandall: "Kinetics of the Reaction Between Carbon Dioxide and Methyldiethanolamine," *Chem. Eng. Sci.* **42** (1987) 1393 – 1398.
95. W. Weisweiler, R. Blumhofer, T. Westermann: "Absorption of Nitrogen Monoxide in Aqueous Solutions Containing Sulfite and Transition-Metal Chelates such as Fe(II)-EDTA, Fe(II)-NTA, Co(II)-Trien and Co(II)-Treten," *Chem. Eng. Process.* **20** (1986) 155 – 166.
96. R. G. Rice, E. L. Benoit: "An Experimental Pressure–Response Method to Measure Gas–Liquid Kinetics," *Chem. Eng. Sci.* **41** (1986) 2629 – 2638.
97. D. B. Hobson, P. J. Richardson, P. J. Robinson, E. A. Hewitt, I. Smith: "Kinetics of the Oxygen–Sulfite Reaction at Waterflood Concentrations: Effect of Catalyst and Seawater Medium," *Ind. Eng. Chem. Res.* **26** (1987) 1818 – 1822.
98. W. Weisweiler: "Naßabsorption von Stickoxiden: Laboruntersuchungen zur Ermittlung der Absorptionskinetik," *Haus Tech. Vortragsveröff.* **500** (1986) 75 – 81.
99. S. S. Laddha, P. V. Danckwerts: "Reaction of CO_2 with Ethanolamines Kinetics from Gas-Absorption," *Chem. Eng. Sci.* **36** (1981) 479 – 482.
100. M. Caplow, *J. Am. Chem. Soc.* **90** (1968) 6795.
101. P. M. M. Blauwhoff, G. F. Versteeg, W. P. M. van Swaij: "A Study on the Reaction between CO_2 and Alkanolamines in Aqueous Solution," *Chem. Eng. Sci.* **38** (1983) 1411 – 1429, 39 (1984) 207 – 225.
102. E. Sada, H. Kumazawa, Z. Q. Han, H. Matsuyama: "Chemical Kinetics of the Reaction of Carbon Dioxide with Ethanolamine in Nonaqueous Solvents," *AIChE J.* **31** (1985) 1297 – 1303.
103. P. V. Danckwerts, A. M. Kennedy: "Kinetics of Liquid-Film Processes in Gas Absorption, Part 1: Models of the Absorption Process," *Trans. Inst. Chem. Eng.* **32** (1954) 49.
104. C. Ouverkerk: "Design for Selective H_2S Absorption," *Hydrocarbon Process.* **57** (1978) 89 – 94.
105. A. E. Cornelissen: "Simulation of Absorption H_2S and CO_2 into Aqueous Alkanolamines in Tray and Packed Columns," *Trans. Inst. Chem. Eng.* **58** (1980) 242 – 250.
106. R. Cornelisse, A. A. C. M. Beenackers, F. P. H. van Beckum, W. P. M. van Swaij: "Numerical Calculation of Simultaneous Mass Transfer of Two Gases accompanied by Complex Reversible Reactions," *Chem. Eng. Sci.* **35** (1980) 1245 – 1260.
107. R. M. Counce, J. J. Perona: "Scrubbing of Gaseous Nitrogen Oxides in Packed Towers," *AIChE. J.* **29** (1983) 26 – 32.
108. W. C. Yu, G. Astarita: "Selective Absorption of Hydrogen Sulfide in Tertiary Amine Solution," and "Design of Packed Towers for Selective Chemical Absorption," *Chem. Eng. Sci.* **42** (1987) 419 – 433.

General References

Chapter 3

109. R. K. Schlünder, F. Thurner: *Destillation, Absorption, Extraktion,* Thieme Verlag, Stuttgart 1986
110. P. Grassmann: *Physikalische Grundlagen der Verfahrenstechnik,* Verlag Sauerländer Salle, Aarau 1982
111. Köge, Moser; *Grundlagen der Verfahrenstechnik,* Springer Verlag, Berlin – Heidelberg – New York 1959.
112. Schubert: *Mischphasenthermodynamik polynärer Systeme,* VCH Verlagsgesellschaft, Weinheim 1986.
113. *Kirk-Othmer,* 3rd ed., **vol. 1,** p. 53 – 96.
114. R. H. Perry, P. W. Green, J. O. Maloney: *Chemical Engineers Handbook,* McGraw-Hill, New York 1984, pp. 14-1 – 14-40.
115. A. L. Kohl, F. C. Riesenfeld: *Gas, Purification* 2nd ed., Gulf Publishing Company, Houston 1974.
116. K. Thormann: *Absorption,* Springer Verlag Berlin-Göttingen-Heidelberg 1959.
117. S. A. Newman: *Acid and Sour Gas Treating Processes,* Gulf Publishing Company, Houston 1985.
118. K. Dieter, W. Hübner: *Gasabsorption, Fortschritte der Verfahrenstechnik,* **vol. 8,** 1966/67, Verlag Chemie, Weinheim 1969.
119. H. Rober et al.: *Lehrbuch der chemischen Verfahrenstechnik,* VEB Verlag für Grundstoffindustrie, Leipzig 1983.
120. P. A. Schweitzer: *Handbook of Separation Techniques for Chemical Engineers,* McGraw-Hill, New York 1979.
121. C.-J. King: *Separation Processes,* McGraw-Hill, New York 1980.

122. J. C. Charpentier: "What's new in Absorption with Chemical Reaction," *Trans. Chem. Eng.* **60** (1982) 131 – 156.
123. G. A. Morris, J. Jackson: *Absorption Towers,* Butterworths, London 1963.
124. K. Hoppe, M. Mittelstrass: *Grundlagen der Dimensionierung von Kolonnenböden,* Verlag Theodor Steinkopf, Dresden 1967.
125. J. Stichlmair, A. Mersmann: "Dimensionierung von Bodenkolonnen für die Absorption und Rektifikation," *Chem. Ing. Tech.* **49** (1977) 106 – 117.
126. R. Billet: *Industrielle Destillation,* Verlag Chemie, Weinheim 1973.
127. E. Kirschbaum: *Destillier- und Rektifiziertechnik,* Springer Verlag, Berlin-Heidelberg-New York 1969.

Specific References

128. K. Hoppe, G. Krüger, A. Bahr, M. Hentschel, *Chem. Ing. Tech.* **21** (1969) no. 9, 536 – 542.
129. Glitsch-Bulletin 4900, Glitsch Inc. Dallas, Tex. 1974.
130. Leaflet Fa. Stahl, Apparate- und Gerätebau GmbH, Viernheim, Federal Republic of Germany.
131. Leaflet Fa. J. Montz GmbH, Hilden, Federal Republic of Germany.
132. Leaflet Fa. Stahl, Apparate- und Gerätebau GmbH, Viernheim, Federal Republic of Germany.
133. Performkontakt Leaflet CLG, Leipzig-Grimma, German Democratic Republic.
134. K. Hoppe, J. Müller: "Zur Frage der Ablaufschachtdimensionierung hochbelasteter Stoffaustauschböden, insbesondere Perform-Kontakt Böden," *Chem. Tech. (Leipzig)* **31** (1979) no. 10, 515 – 519.
135. A. Mersmann, A. Deixsler: "Packungskolonnen," *Chem. Ing. Tech.* **58** (1986) no. 1, 19 – 31.
136. W. Reichelt, E. Blaß: "Berechnung von Füllkörperapparaten bei Gegenstrom von Gas und Flüssigkeit," *Chem. Ing. Tech.* **46** (1974) no. 5, 171 – 180.
137. R. Billet, J. Maćkowiak: "Wirksamkeit von Kunststoff-Füllkörpern bei der Absorption, Desorption und Vakuum-Rektifikation, *VT Verfahrenstech.* **16** (1982) no. 2, 67 – 74.
138. R. Billet, J. Maćkowiak: "Neuartige Füllkörper aus Kunststoffen für thermische Trennverfahren," *Chem. Tech. (Heidelberg)* **9** (1980) no. 5, 219 – 226.
139. M. Kolev: "Wirkungsweise von Füllkörperschüttungen," *Chem. Ing. Tech.* **48** (1976) no. 2, 1105 – 1110.
140. Leaflet P63 Mass Transfer Ltd. Houston, USA.
141. Norton, Akron, Ohio, USA Bulletin IM-82.
142. R. F. Strigle, jr., F. Rukovena: Packed Distillation Column Design, CEP, March (1979), S. 86 – 91
143. M. J. Dolan, R. F. Strigle jr.: Advances in Distillation Column, *Chem. Eng. Prog.* Nov. 1980 78 – 83.
144. R. F. Strigle jr., Kazumichi Hiramatsu: Replacing Trays with Metal Packing saves Energy in an Ethylene plant's hot Section, OGJ Report," *Oil Gas J.* Sept. 26 (1986) 63 – 65.
145. Norton, Akron, Ohio, USA, Bulletin Si-72.
146. Raschig GmbH, Ludwigshafen, West Germany: *Hydrophilierte Kunststoff-Füllkörper.*
147. H. Harttig, K. Hüttinger: "Hydrophilierte Füllkörper," *Chem. Ing. Tech.* **54** (1982) no. 10, 909 – 912.
148. E. Seebach: "Packringe fordern Ablösung," *Chem. Ind.* (1987) no. 2, 2 – 3.
149. R. Billet, R. Chromik: "Hochleistungsfüllkörper für Gas/Flüssig-Systeme," *Chem. Tech.* (Heidelberg) (1987) no. 2 – 5.
150. W. Meier, R. Hunkeler, W.-D. Stöker: "Sulzer Mellapack–Eine neue geordnete Packung für Stoffaustausch-Apparate," *Chem. Ing. Tech.* **51** (1979) no. 2, 119 – 122.
151. *Trennkolonnen für Destillation und Absorption,* 1979, Sulzer AG, Winterthur (Switzerland).
152. K. Hoppe, R. Kunze: "Fortschritte auf dem Gebiet apparativer Lösungen für Gas/Dampf/Flüssigkeits-Kontaktierung," *Chem. Ing. Tech.* **57** (1985) 686 – 687.
153. *Technische Information zur Kolonnenpackung Pyrapak,* VEB Chemieanlagenbaukombinat Leipzig-Grimma, German Democratic Republic (1985).
154. U. Bühlmann: "Rombapack, ein neuer geordneter Packungskörper für Stoffaustauschböden," *Chem. Ing. Tech.* **55** (1983) no. 5, 379 – 381.
155. H. Short: "New mass-transfer find is a matter of gravity," *Chem. Eng.* Febr. 21 (1983) 23 – 29.
156. R. Fowler: *Potential Applications of Higee offshore,* brochure Glitsch Inc., Dallas/Texas, USA. Chapter 4
157. S. Weiß et al.: *Verfahrenstechnische Berechnungsmethoden,* "Thermisches

Trennen," Part 2, VCH Verlagsgesellschaft, Weinheim 1986.
158. Souders, G. G. Brown, *Ind. Eng. Chem.* **24** (1932).
159. W. C. Edmister, *Ind. Eng. Chem.* **35** (1943) 837.
160. G. Horton, W. B. Franklin, *Ind. Eng. Chem.* **32** (1940) 1384.
161. A. Kremser, *Nat. Pet. News* **22** (1930) no. 21, 42.
162. A. P. Colburn, *Trans. Am. Inst. Chem. Eng.* **35** (1939) 211.
163. J. S. Eckert: "No Mystery in Packed Design,"*Oil Gas J.* Aug. 24 (1970) 60 – 64. Chapter 5
164. G. S. Gottschink, H. J. Wehner: "Hydraulische Auslegung des Durchmessers von Füllkörperkolonnen," *CAV* (1983) 23 – 26.
165. R. M. Kelly, R. W. Rousseau, J. K. Ferrel: "Physical Absorption of Acid Gases in Methanol," *Ind. Eng. Chem. Process Des. Dev.* **23** (1984) no. 1, 102 – 109.
166. R. Semmelbauer: "Die Berechnung der Schütthöhe bei Absorptionsvorgängen in Füllkörperkolonnen," *Chem. Eng. Sci.* **22** (1967) 1237 – 1255.
167. S. Y. Diab, R. N. Maddox: "Here are the Techniques for Choosing a Solvent and Calculating the Column Size required for a Specific Process in which a Gas is absorbed into a Liquid," *Chem. Eng.* Dec. 27 (1982) 38 – 56.
168. J. L. Bravo, J. R. Fair: "Generalized Correlations for Mass Transfer in Packed Distillation Columns," *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 162 – 170.
169. P. M. M. Blawhoff, B. Kamphuis, W. P. M. Swaaiz, K. R. Westerterp: "Absorber Design in Sour Natural Gas Treatment Plants: Impact of Process Variables on Operation and Economics," *Chem. Eng. Process.* **19** (1985) 1 – 25.

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Adsorption

Standard Article

Hans-Jörg Bart[1] and Ulrich von Gemmingen[2]

[1]Technische Universität Kaiserslautern, Lehrstuhl für Thermische Verfahrenstechnik, Kaiserslautern, Germany

[2]Linde AG, Höllriegelskreuth, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.b03_09.pub2
Article Online Posting Date: January 15, 2005

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Definition and History
1.2. The Development of Adsorption Technology
1.3. Adsorbents and Processes for Separation of Gases and Liquids
2. Adsorption Apparatus
2.1. Gas Phase Adsorption
2.2. Liquid-Phase Adsorption
3. Adsorbents
3.1. Oxidic Hydrophilic Adsorbents
3.2. Carbon-Containing Adsorbents
3.3. Chemical Complexing Sorbents
4. Thermodynamics, Equilibria, and Heat of Adsorption
4.1. Explicit Equilibria
4.1.1. Single-Component Pure-Gas Equilibria
4.1.2. Temperature Dependence of the Saturation Loading
4.2. Adsorption Enthalpy
4.2.1. Differential Adsorption Enthalpies
4.2.2. Integral Adsorption Enthalpies
4.3. Mixture Equilibria
4.3.1. Homogeneous Surfaces
4.3.2. Mixture BET Equation
4.3.3. Ideal Adsorbed Solution Theory
4.3.3.1. Spreading Pressure
4.3.3.2. Examples for Binary Mixture Calculations
4.3.3.2.1. Given Gas-Phase Concentration
4.3.3.2.2. Given Loading
4.3.4. Statistical Thermodynamics Model
5. Kinetics
5.1. External/Internal Transport
5.1.1. Internal Resistances
5.1.2. Linear Driving Force and Other Models
6. Adsorber Dynamics
6.1. Modeling the Isothermal Fixed-Bed Adsorber
6.2. Modeling the Adiabatic Fixed-Bed Adsorber
6.2.1. Loading and Breakthrough Curves
6.2.2. Simplified Models
6.2.3. Coefficients
7. Regeneration of Adsorbents
7.1. Regeneration by Temperature Swing
7.2. Regeneration by Pressure Swing
7.3. Regeneration by Displacement
7.4. Regeneration by Extraction
7.5. Reactivation of Adsorbents
7.6. Liquid Phase Desorption
8. Gas Phase Processes
8.1. Adsorptive Separation of Gas Mixtures
8.2. Pressure Swing Processes
8.3. Separation of Nitrogen from Oxygen
8.4. Adsorptive Purification of Air
8.5. Removal of Radioactive Nuclides from Exhaust Gas
8.6. Removal of Organic Components from Exhaust Air
8.7. Adsorptive Desulfurization Processes
8.8. Adsorptive Purification of Methane
8.9. Adsorptive Purification of Hydrogen
8.10. Separation of Isomers
9. Liquid-Phase Processes
9.1. Purification of Drinking Water
9.2. Wastewater Treatment
9.3. Separation of Nonaqueous Substances

[Top of Page]

1. Introduction

1.1. Definition and History

Physical adsorption commonly is understood as reversible binding of molecules and atoms from the gaseous or liquid phase on surfaces, mostly of highly porous adsorbent media. The reverse process is called desorption. In adsorption, the *adsorptive* accumulates on the *adsorbent* which is then loaded with *adsorbate.* During desorption, the adsorbate present in the condensed phase passes from the surface of the adsorbent into the fluid phase. Ad- and desorption are similar to condensation and sublimation, but fluid and solid are heterogeneous and the solid remains inert. Linguistically, “sorption” originates from the latin “sorbere” (to devour) and indicates an event of a certain vehemence. The suitable physical description is phase thermodynamics with an additional degree of freedom attributed to the surface. The basic statistical approach for adsorption phenomena is different to that for condensation and pore condensation: while surface adsorption is governed by individual statistics with occupation indices for characteristic adsorption sites of distinct binding energies, condensation is related to collective statistics with sharp phase transitions due to the loss of the surface degree of freedom. Adsorption events are accompanied by intensive heat production, volumetric contraction, and space confinement leading to saturation limits. The most prominent effect on adsorption, the volumetric contraction, has been known for more than two thousand years. Examples are water vapor adsorption by freshly sheared lamb's wool and the purification of contaminated water with charcoal. In both cases, finely structured surfaces are involved. Since the 1400s recovery of adsorbed water from air was used. Around 1780 the famous Genovian universalist HORACE BÉNÉDICT DE SAUSSURE measured the adsorption capacity of charcoal for a series of gases and showed that they could achieve loadings of about 25 % of modern activated carbons.

1.2. The Development of Adsorption Technology

The technical applications of adsorption cover a wide range, such as separation, synthesis, storage, purification of gases and liquids, and air conditioning, and reveal a multifacetted discipline within thermal and chemical engineering science. It is increasingly becoming established in chemical processes as an alternative to the conventional separation techniques of rectification, absorption, condensation, drying, extraction, membrane processes, and catalysis, and it profits from various and optimally tailored adsorbents.

The overlap with the thermal separation techniques mentioned above and with surface science and nanotechnology is considerable. To avoid confusion, the term “technical adsorption” can be restricted to processes involving highly porous media. The development and characterization of such media is one of the most challenging tasks of adsorption technology for the time being and in the future. The advent of molecular sieves in 1965 opened new dimensions for highly effective separations by providing highly versatile and advantageous technical adsorbents.

Our knowledge on adsorption is owed to exceptional scientists: SIR WILLIAM THOMSON, later LORD KELVIN, found the relation between adsorption energy and pore dimensions in 1850; JOSIAH WILLIARD GIBBS introduced the thermodynamic description of adsorption phenomena in 1876, WILHELM OSTWALD developed an isotherm equation at heterogeneous surfaces that is still valid today in 1885. MICHAEL POLANYI elucidated the relation between adsorption potential and adsorptive compression [1], and IRVING LANGMUIR derived the exact equilibrium equation under homogeneous conditions from ad- and desorption rates [2, 3].

With the beginning of the technical use of adsorption for the first industrial separation units in 1920, multilayer adsorption was described by BRUNAUER, EMMET, and TELLER in the form of the BET equation [4], which is still in use for adsorbent characterization. DUBININ proposed a correlation for predicting isotherms from liquid properties [5, 6]. The frontal analysis of breakthrough curves showed a complex interaction of heat and mass fronts in fixed beds, which first was interpreted by E. WICKE [7]. The important relation between isotherm form and the adsorption energy distribution at energetic heterogeneous surfaces was pioneered by SIPS [8]. Parallel to the new multicomponent separation technology LEWIS [9] and later MYERS and PRAUSNITZ [10] developed a prediction scheme for mixtures from pure gas equilibria, which is still a standard method.

1.3. Adsorbents and Processes for Separation of Gases and Liquids

Three principle mechanisms in adsorptive separation processes can be found: steric, kinetic, and equilibrium effects. The steric effect occurs with zeolite and carbon molecular sieves (⟶ Zeolites, ⟶ Ion Exchangers). These have a very narrow distribution of micropores. Only components with critical diameter smaller than the opening of the micropores will be adsorbed. In this way, mixtures can be separated by adsorption. An example is the separation of *n*-paraffins from isoparaffins, despite its being mathematically treated as an equilibrium separation. Kinetic separation has its major application in air separation. Zeolites have a good N_2 equilibrium adsorption, but carbon molecular sieves exhibit a 30 times faster diffusion of oxygen than nitrogen. Despite the considerably lower capacities compared to zeolites the overall economics are better. However, for equilibrium separations the most vital point is sorbent selection.

Most technical-scale adsorption units typically are fixed-bed processes with multilayers. The stratified bed arrangements can be adapted well to the working ranges of several mass fronts of the various components. Therefore, the knowledge of the working capacities of the different adsorbents with respect to the various components under the given conditions of pressure and temperature are the key issue for all process design.

Three major types of processes have been established. Their names refer to the method of regeneration. The temperature swing process (TSA), the pressure swing process (PSA) or (VSA), and the concentration swing process (CSA).

Historically the TSA variant, which regenerates under elevated temperatures, was developed first in 1920. It is especially suitable where strongly bound traces must be removed from a feedstock. Compared with other desorption techniques, temperature elevation produces the highest saturation of these components in the residual gas. The high adsorption capacity for the strongly bound traces results in adsorption times that are long enough to heat and cool the adsorbers in the regeneration step of cyclic processes. Typical TSA applications are the removal of volatile organic compounds (VOC) from air (air protection) and prepurification units (PPU) in front of cryogenic air separators removing H_2O, CO_2, N_2O, and C_{3+} hydrocarbonsdown to the vol ppb level with activated alumina and molecular sieves. The total worldwide air streams to be treated with TSA PPU is about 5×10^7 m^3/h (STP) using about 10^4 t of adsorbent.

Since 1970 PSA processes have been developed that regenerate only by depressurization and low-pressure purge steps. The TSA technique is mainly suitable if more or less volatile components of balanced concentrations have to be separated at ambient temperature. With PSA, the effective saturation of the waste gas with desorbed components is acheived simply by expansion without external heat. Multidesorber systems with up to 16 vessels with various pressure-equalization steps can provide product recoveries up to 92 % for hydrogen plants, and even ultrahigh purities such as 99.999 % H_2. Examples of PSA applications are the fine purification of hydrogen from steam-reformed natural gas, refinery, and coke oven off-gas, as well as methane recovery form natural gas blends. Another application is effective separation of CO_2 and CO in steel production facilities. The worldwide production of refined H_2 of ca. 10^7 m^3/h (STP) requires 10^4 t of silica, activated carbon, and molecular sieves. When the desorption pressure is in the vacuum range, then PSA is known as the vacuum swing process (VSA). A typical application is N_2/O_2 separation from air.

If neither heat nor pressure energy is available or liquid are to be separated, desorption usually uses the CSA technique. Typical examples are VOC removal with activated carbons, hydrophobic zeolites, or adsorption polymers.

In practice, TSA, PSA, and CSA operations are not carried out in a pure form. Real processes typically use all mechanisms simultaneously. The adiabaticity of technical-scale plants always produces distinct temperature effects with ad- and desorption, even in PSA plants, and most TSA plants have additional pressure changes and pure-gas purge steps.

VOC-removal plants and PPU units can have ultrahigh throughputs of up to 5×10^5 m^3/h (STP). Then the construction is subject to extreme flow conditions, which require considerable effort for flow distribution, pressure drop, heat dissipation to and from walls, and thermal expansion of bed-limiting grids. With PPU units three major adsorber types have been established for low, medium, and high air flows: vertical, horizontal, and radial adsorbers, all which of have numerous constructive finesses to conform to actual industrial needs.

Liquid adsorption processes are expanding, with new tasks in the areas of preparative chromatography, bioadsorption, and life sciences. Here well-tailored adsorbents and customized separation units are operative. The adsorption strengths of these materials often could be conditioned by ion doping and pH control of the eluents. Relevant topics are excess isotherms, frontal analysis, computational fluid dynamics (CFD), and surface characterization. In addition to the classical fixed- and fluidized-bed processes, variants such as micellary, capillary, reactive, extractive, and gradient-oriented chromatography have been investigated for separating pharmaceutical substances, for example, enantiomers. Constructions with plates and annular adsorbers have resulted from research in this field.

The commonly applied adsorbents for technical process design of gas and liquid separations are K, Na, and Ca zeolites and Li-exchanged zeolites of the types A and X for O_2 VSA, activated carbons of various activation levels, carbon molecular sieves, activated alumina, silica of distinct porosities, and adsorber polymers. While molecular sieves, zeolites, activated alumina, and silica preferably bind polar or polarizable components, carbons and polymers prefer nonpolar molecules like hydrocarbons. By varying the inner porosity of the adsorbents they can be adapted to the volatility of the components, e.g., less volatile components need materials with larger pore radii to be well desorbable. Separation processes therefore mainly are designed as stratified beds with adsorbent layers adapted to the properties of the various components.

Activated carbon has a hydrophobic character and is especially suitable for the removal of organic substances contained in liquids (aqueous solutions) and gases. The even more hydrophobic polymeric resins are preferably used in liquid adsorption processes. New developments led to activated carbon fibers, carbon molecular sieve membranes,

and carbon nanotubes, the application potential of which is not fully exploited yet.

In general, adsorption is performed at ambient temperatures and can be tailored to any separation or purification problem with commercially available adsorbents. This makes a process very attractive and economical, even for low or nonstationary solute concentrations.

[Top of Page]

2. Adsorption Apparatus

Adsorption processes have a minimum of two steps, ie., adsorption and desorption, which define an adsorption cycle. This is true of liquid and gas adsorption despite the different physical mechanisms involved. In gas adsorption, parallel adsorption and condensation result in an adsorptions – desorption hysteresis which is not seen in liquid adsorption. Also, temperature effects in gas adsorption are more pronounced due to the additional heat of condensation and the low heat capacity. Nonisothermal operation and even safety problems (hot spots, ignition) are not seen with liquid operations, where the high heat capacity of the liquid phase guarantees isothermal operation.

A further marked difference is in the diffusivities and solubilities of the fluids involved. Diffusivities in the liquid phase are about 1000 times smaller than in the gas phase. If the solute (e.g., ethanol) is totally miscible with the process fluid (e.g., water or air) the mass of solute per cubic meter is 1000 times higher in liquid than in gas systems. However, there is a wide span of solubilities in the liquid phase, e.g., nonvolatile pharmaceuticals may have solubilities in the ppm range. Liquid adsorption and chromatography (⟶ Basic Principles of Chromatography) use similar concepts as gas phase adsorption to describe sorption in an appropriate apparatus. Chromatography also uses steric effects (gel permeation chromatography) or equilibrium solubility (partition chromatography) or chemical interactions (ion exchange or affinity chromatography) for separation.

Chromatography separates components that are injected as pulses of feed. The more weakly retained component dissolves better in the eluting mobile phase and separation is thus achieved (Fig. 1A). When the feed is pushed continuously through the column until breakthrough of the strongly adsorbed component occurs, the mode of operation is referred to as adsorption instead of chromatography (Fig. 1B)

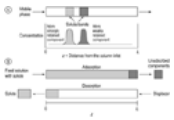

Figure 1. Chromatographic (A) and adsorptive (B) steps

.

2.1. Gas Phase Adsorption

In gas-phase adsorption, *fixed beds* are used almost exclusively. In only a few cases are *moving* or *fluidized* beds employed. All types of bed can likewise be used in liquid-phase adsorption, although suspension mixers are also utilized in the case of finely dispersed adsorbents (e.g., powdered coal).

Fixed-Bed Adsorbers. Figure 2 shows an adsorption unit with two fixed beds. While the mixture to be purified or separated flows through the adsorber on the left, the other adsorber is being regenerated. At the start of the adsorption cycle, the incoming fluid encounters regenerated, highly active adsorbent. The adsorptive diffuses into the pores of the adsorbent. In a certain layer of the fixed bed, mass is transferred from the fluid to the solid phase. Once this layer has been saturated, succeeding mixture entering the bed flows past it into downstream layers, which then become saturated. In this way, the *mass-transfer zone* (MTZ) moves through the fixed bed with time. In this zone, the concentration of adsorptive in the fluid decreases from the feed concentration (c_0 or Y_0) to the equilibrium concentration (c_1^* or Y_1^*).

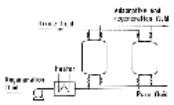

Figure 2. Adsorption unit with two fixed beds (adsorber and desorber)

Breakthrough Curves. Figure 3 shows how the effluent concentration Y_i changes with time or with the total fluid volume throughput. The mass-transfer zone separates the fixed bed into two equilibrium zones, one upstream and one downstream. Frequently, the loading curve is almost symmetrical and S-shaped in the mass-transfer zone. This means that adsorbent utilization in the MTZ is only 50 %. When the adsorptive breaks through at the end of the loading cycle, the *length of unused bed* (LUB) is just half that of the MTZ. The breakthrough curve is obtained by continuing to operate the bed until the concentration of the mixture leaving the adsorber is equal to the feed concentration. In some cases, the shape of the breakthrough curve changes little as it moves through the bed (constant pattern). In favorable cases (equilibrium favorable for adsorption, small adsorbent particle diameter), the LUB is only one-fifth to one-fourth the length of the bed; in other cases it can be substantially longer (see Chap. Kinetics).

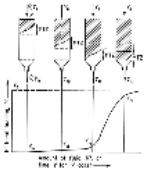

Figure 3. Breakthrough curves in fixed bed showing mass-transfer zone (MTZ) The diagram gives effluent loading as a function of the amount of fluid or of time

Breakthrough Time. If the volume flux $\dot{V}$ of the fluid phase with feed concentration $c_{0,i}$ of adsorptive *i* is completely depleted and the unloaded mass of adsorbent *S* is thereby enriched to a loading X_i (in kilograms of adsorbate per kilogram of adsorbent), the breakthrough time t_B* is obtained from the following material balance:

$$t_B^* = \frac{SX_i}{\dot{V}c_{0,i}} \tag{2.1}$$

This breakthrough time, however, is valid only for the case of a rectangular loading profile in the bed, which is technically unattainable. Depending on the slope of the breakthrough curve, the actual breakthrough time is shorter than t_B^*. Temperature increases in the bed due to release of the heat of adsorption further decrease the breakthrough time.

Moving-Bed Adsorbers. While the mass-transfer zone moves downstream in a fixed bed, it remains stationary in a continuously operated moving bed. Because a bed height corresponding to the mass-transfer zone is sufficient, the mass of adsorbent required is less than in the fixed bed. This advantage is offset by both the higher cost of moving the solids and the concomitant abrasion. Occasionally, moving beds are operated in cross-current (Fig. 4) or cocurrent modes.

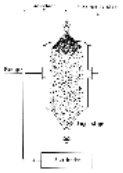

Figure 4. Single-step, crosscurrent moving-bed adsorber

Fluidized-Bed Adsorbers. Fluidized-bed adsorbers can also be operated continuously. Figure 5 shows a multistage countercurrent adsorber with an adsorption zone at the top and a regeneration zone at the bottom. Such an apparatus, for example, is suitable for drying moist air that enters at the lower end of the adsorption zone. Dry air leaves the adsorber at the upper end. The adsorbent moves continuously from top to bottom over individual trays on which fluidized beds are formed.

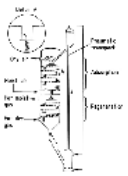

Figure 5. Continuous countercurrent adsorber for drying air

The loaded adsorbent, moving countercurrent to the rising hot gas, reaches the regeneration zone at the bottom and is regenerated in this way. Subsequently, solids are returned to the top pneumatically via a vertical line.

Figure 6 shows the so-called Hypersorber, which permits continuous separation of a gas mixture into its two components. The feed is injected at the middle of the column. The adsorbent moving downward from the top adsorbs one gaseous component, while the other component is drawn off as the head product. The adsorbed component leaves the adsorbent in the evaporation or regeneration zone, where the adsorbent is stripped with a hot gas or steam. Above this zone, the bottom product (the adsorbed component) is drawn off. The adsorbent leaves the apparatus at the bottom and is returned to the top pneumatically. After being recooled, it is again delivered to the adsorption zone.

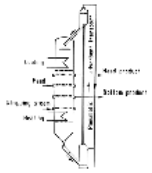

Figure 6. Continuous countercurrent adsorber for separating a gas mixture

Adsorption Wheels. Vertical and horizontal adsorption wheel systems operate continuously. The adsorbent is contained in basket segments or coated onto a metal or ceramic honeycomb substrate. These structures rotate on a vertical or horizontal axis for removing the adsorbate, and a small portion of the wheel is used for thermal regeneration (Fig. 7). The advantage is the low pressure drop, whereas due to the short residence time and mechanical leakage of the seals efficiency is low.

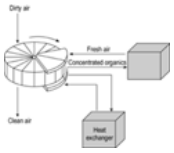

Figure 7. Rotary adsorber

The adsorbent is regenerated by temperature or pressure swing, displacement, or extraction. The temperature swing process, which is the most common, can be achieved by indirect heating (external double jacket) or directly with hot gas or steam. If steam is used, an adsorption cycle then consists of the steps adsorption, desorption, drying, and cooling prior to fresh adsorption.

2.2. Liquid-Phase Adsorption

In liquid-phase adsorption of suspensions, fluidized beds and mixers are used in addition to fixed beds. Figure 8 shows a mixer in which mass is transferred from the liquid to the adsorbent particles suspended in it; the suspension is then separated by filtration into the purified liquid and the loaded adsorbent. A new concept for continuous adsorption are carousel adsorbers, which have their origins in ionic separations (ISEP) and liquid adsorption or chromatographic separations (CSEP) [11]. Up to 30 absorber units are installed on a rotating support disk, and flexible valve switching provides true continuous operation (Fig. 9).

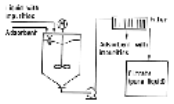

Figure 8. Adsorption unit for liquid processing

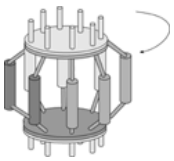

Figure 9. Carousel adsorber

Figure 10 shows a three-component separation. Three or more products can be obtained, and the unit is very flexible with regard to installation of additional sanitation (usually with NaOH), regeneration (usually with an acid), or washing steps (usually with water) prior to reuse. Scrubbing, i.e., product reflux prior to desorption, is also possible to remove impurities when high-purity products are demanded. The same principle is applied in annular adsorption or chromatography. Instead of slowly rotating columns the sorbent is packed between two concentric cylinders. The feed and the eluent are continuously fed into the top of the annular bed. While the eluent is uniformly fed to the entire circumference, the feed mixture is introduced into only one sector of the annular bed. The column assembly is slowly rotated, while the feed nozzle remains stationary. Rotation causes the separated components to appear as helical bands, each of which has a characteristic, stationary exit point (see Fig. 11) [12, 13]. The result is quite similar to the ISEP principle in Figure 10, and all the features, e.g., sanitation, washing, scrubbing etc., can be applied here in a similar manner [14].

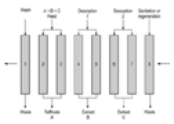

Figure 10. Carousel adsorber

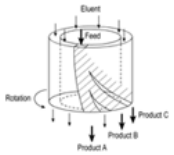

Figure 11. Annular chromatography

The advantage of the ISEP principle is high flexibility. All the colums can be connected in series to obtain a simulated moving bed (SMB) configuration, as shown in Figure 12. Like in distillation, the feed is fed into the middle of the column assembly resulting in a pure component A in the enriching and B in the stripping section. In contrast to a true moving bed there is no packing attrition due to shear forces, column packing is uniform, and flow control is precise. To achieve countercurrent flow of the packing all inlet and

outlet flows are periodically shifted in the same direction to a new zone (Fig. 13). The simulated solid flow is thus directly linked to the shift period, and a triangular operating diagram has been developed [15] that allows a quasi-steady-state SMB process to be designed with flow control that depends on the sorption equilibria involved. Without this, the concentration fronts, as shown in Figure 14, will move, so either no pure A or B or even no separation would be achieved.

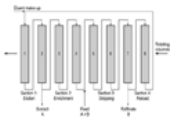

Figure 12. SMB carousel adsorber

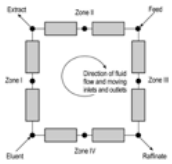

Figure 13. Simulated moving bed

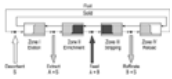

Figure 14. Simulated moving bed concentration profile

Zone I, desorption of A, is between the point of desorbent injection and extract withdrawal. The adsorbent entering Zone I carries only A and S. The liquid entering the top of the zone consists of pure S. As the liquid stream flows downward, component A in the pores is displaced by S. A portion of the liquid leaving the bottom of Zone I is withdrawn as extract; the remainder flows downstream into Zone II as reflux.

Zone II, desorption of B, is between the point of feed injection and extract withdrawal. At the fresh feed point, the upward flowing solid adsorbent contains the quantity of component B that was adsorbed in Zone III. However, the pores will also contain a large amount of A, because the adsorbent has just been in contact with fresh feed. The liquid entering the top of Zone II contains no B, only components A and S. Thus, component B is gradually displaced from the pores by A as the adsorbent moves up through Zone II. At the top of Zone II the pores of the adsorbent contain only A and S.

Zone III, adsorption of A, is between the point of feed injection and raffinate withdrawal. As the feed flows down through Zone III, countercurrent to the solid adsorbent flowing upward, component A is selectively adsorbed from the feed into the pores of the adsorbent. At the same time, the desorbent, component S, is desorbed from the pores of the adsorbent to the liquid stream to make room for A in the pores.

Zone IV, the isolation zone, is where the feed components in Zone III are segregated from extract in Zone I: At the top of Zone I, the adsorbent pores are completely filled with S. The liquid entering the top of Zone IV consists of B and S. By properly regulating the flow rate [15] of Zone IV it is possible to prevent the flow of component B into Zone I and avoid contamination of the extract.

In practice, it is very difficult to actually move a solid bed of adsorbent. In the Sorbex process, the countercurrent flow of liquid feed and solid adsorbent is accomplished without physical movement of the solid. Instead, countercurrent flow is simulated by periodically changing the points of liquid injection and withdrawal along a stationary bed of solid adsorbent (Fig. 15). In this simulated moving bed (SMB) technique, first patented by UOP [16], the concentration profile shown in Figure 14 actually moves down the adsorbent chamber. As the concentration profile moves, the points of injection and withdrawal of the net streams are moved along with it. This movement of the net streams is performed with a unique rotary valve developed by UOP specifically for the Sorbex family of processes.

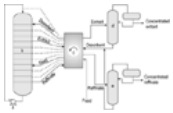

Figure 15. Sorbex process (UOP)

a) Pump; b) Adsorbent chamber; c) Rotary valve; d) Extract column; e) Raffinate column

The flow distribution at each stage is achieved with specialized internals and grids, which are used to inject or withdraw liquid via the bed lines connecting the beds to the rotary valve. A typical unit has 12 beds and, in order to double the number of stages, two colums are then in series. At any given time, only four of the bed lines are active in carrying the net streams into and out of the adsorber. The rotary valve is used to periodically switch the position of the liquid feed and withdrawal points as the composition profile moves down the apparatus. It also sends the diluted extract and raffinate to the distillation columns and reprocesses the desorbent. A circulating pump provides a liquid flow from the bottom to the top of the column.

[Top of Page]

3. Adsorbents

The most important industrial adsorbents are oxygen-containing compounds such as silica gel, activated alumina, zeolites and molecular sieves, clay minerals, and pillared clays with hydrophilic properties, and adsorbents with carbon as the major component (activated carbon, activated coke, carbon molecular sieves, carbon nanotubes, and polymeric resins), which are hydrophobic in nature (see Table 1).

Table 1. Industrially important adsorbents

	Adsorbent	Major component	Shape* and size	True density (kg/m^3)	Apparent density (kg/m^3)	Packing density (kg/m^3)	Micropore volume (cm^3/g)	Macropore volume (cm^3/g)	Specific surface (m^2/g)	Specific heat capacity (J/kg K)	Thermal conductivity (W/m K)	Porosity ∈ (-)
Oxygen-containing adsorbents	Silica gel	SiO_2	P, S (1–8 mm)	2200	750 – 1250	300 – 850	0.30 – 1.15	< 0.10 – 0.25	100 – 850	0.92 – 1.0	0.14 – 0.20	0.45 – 0.65
	Activated alumina	Al_2O_3	G, S (2–10 mm)	3000 – 3100	1200 – 2400	700 – 950	0.40 – 0.60	0.10 – 0.15	100 – 400	0.88 – 1.05	0.12	0.13 – 0.6
	Zeolites and molecular sieves	SiO_2 Al_2O_3	P, S (1–5 mm), HC (6–7 mm), CP (2–4 mm)	2100 – 2600	1100 – 1500	400 – 900	0.20 – 0.30	0.30 – 0.70	350 – 1100	0.80 – 1.05	0.13 – 0.58	0.5 – 0.6
	MCM	SiO_2 Al_2O_3	HC		550 (min)	170 – 450	Mesopore: 0.21 – 1.00	450 – 1200			0.20 – 0.52	
	Clay mineral	SiO_2 Al_2O_3	P, G	2200	1600 – 1700	500 – 700	~ 150	0.19 – 0.40	120 – 500			0.20 – 0.35

	and pillared clays											
Carbon adsorbents	Activated carbon	C	P, G (1–10 mm), CP	1880 – 2200	440 – 900	250 – 550	0.20 – 0.50	0.40 – 1.5	200 – 4000	0.76 – 0.92	0.65 – 0.1	0.45 – 0.77
	Activated coke	C	P, G	~ 1900	~ 900	450 – 600	0.01 – 0.10	0.20 – 0.30	15 – 400			~ 0.35
	Carbon molecular sieves	C	P, S, CP	1800 – 2100	900 – 1290	600 – 900	~ 0.18	0.38 – 0.60	100 – 1500	0.84 – 1.00	0.65	0.35 – 0.57
	Carbon nanotubes	C	P				Mesopore: 0.05 – 0.45	150 – 650				
	Polymeric resins	C, H	S, G	1000 – 1300	400 – 900	300 – 750	< 0.10 – 0.43	0.37 – 1.50	80 – 1500	0.33 – 1.5**	0.06 – 0.07**	0.35 – 0.65
Π - Complexation sorbents		SiO_2, C, Cu, Ag	P, G, CP					0.20 – 1.60	50 – 670			

*P: fine powder; S: spheres; G: granules; HC: hollow cylinder (external diameter); CP: cylindrical pellets (diameter)
**In dry state

Adsorbate – Adsorbent Interactions. The adsorbate – adsorbent interactions are of threefold nature: dispersion, electrostatic, and chemical bonding. Polarizability, magnetic susceptibility, permanent dipole moment, and quadrupole moment of the solute decide the selection of the adsorbent. With activated carbon nonspecific van der Waals (dispersion) interactions dominate, and thus they preferentially accumulate organic substances and nonpolar adsorptives with high polarizability and magnetic susceptibility and low polarity. If the solute has a high dipole or quadrupole moment highly polar (hydrophobic) surfaces are needed. For ultrapurification a strong (chemical) bond results in a high Henry constant. Pore size and surface characteristics determining the interactions are created in the sorbent preparation process, which is discussed in detail in [17].

The selection of a sorbent from a practical point of view are determined by the following criteria:

1. Capacity of the sorbent
2. Product purity or selectivity
3. Regeneration method (pressure or thermal swing)
4. Length of unused bed (LUB)
5. Price

Capacity and selectivity are a matter of the equilibrium isotherms discussed in Chapter Thermodynamics, Equilibria, and Heat of Adsorption. The equilibrium isotherms of all constituents in the mixture must be considered, and in a simplified approach the isotherms of pure components may be regarded as additive to obtain the multicomponent adsorption behavior. The LUB is approximately one-half of the span of the mass transfer zone and is primarily determined by the equilibrium isotherm. A sharp concentration front (short LUB) results from a high sorbent productivity and gives a high product purity. This is associated with a higher price, which can be paid for bulk separations and is more critical for purification processes. Here bulk separation refers to separations of a mixture with adsorptive concentrations higher than 10 %.

Particle size influences mass transfer, pressure loss, and operability. Small particles have a shorter intraparticle diffusion path and thus promote diffusion and mass transfer, but they result in a higher pressure loss. Particle sizes for gas-phase processes range from 1 to 10 mm, while liquid-phase applications use smaller diameters due to lower liquid diffusivities. Generally, adsorbents are used in the form of spheres, rods, or moldings. Here spherical particles are more expensive than angular ones resulting from crushing and sieving of larger blocks. The latter tend to produce fines which affect pressure drop and result in a more troublesome operation. Adsorbents for moving or fluidized beds must exhibit high abrasion resistance, which depends on particle shape and hardness.

For determination of a particle size distribution established methods are available [18]. However, care should be taken in their measurement and presentation. In a number distribution the smaller particles will be dominant, whereas in a volume or mass distribution the larger ones will prevail.

Density and Porosity. A distinction is made between the solid particle density (the true density of the material), the apparent density of a grain with pores, and the bulk density of the packing in a fixed-bed adsorber (see Table 1).

The true density is the quotient of mass of dry adsorbent and the volume of the solid phase.

$$\rho = \frac{m_p}{V_m} \tag{3.1}$$

The true density can be estimated by pycnometric methods using helium or methanol [19, 20].

The apparent density includes the volume of the pores

$$\rho_s = \frac{m_p}{V_m + V_{Pore}} \tag{3.2}$$

and can be determined by mercury pycnometry [20] or by weighing [21]. The porosity of a particle is then [19]

$$\varepsilon_p = 1 - \frac{\rho_s}{\rho} \tag{3.3}$$

The density of a packed bed [19, 22] determines the voidage or bed porosity:

$$\varepsilon_B = 1 - \frac{\rho_B}{\rho_s} \tag{3.4}$$

It can be determined from a residence-time experiment with a tracer which cannot enter the pores. In liquid applications the retention time $t_{\in}$ of, e.g., Dextran blue (molecular weight ca. 2×10^6), the flow rate V, and the total column volume V_C, give:

$$\varepsilon_B = \frac{t_s \cdot V}{V_C} \tag{3.5}$$

In general, $\in_B$ lies between 0.33 and 0.5. If a nonretained (nonadsorbable) tracer is available that is small enough to diffuse into the pores, the inner porosity $\in_p$, can also be determined. From the peak maximum of the small tracer the total porosity is determined

$$\varepsilon_t = \frac{t_p \cdot V}{V_C} \tag{3.6}$$

and the inner porosity can be calculated:

$$\varepsilon_t = \varepsilon_B + (1 - \varepsilon_B) \cdot \varepsilon_p \tag{3.7}$$

The most important characteristics of a sorbent is its high *specific internal surface area a*. It is inversely proportional to the mean pore diameter. Thus, the molecular size of the adsorptive determines the minimal pore diameter accessible for penetration. The standard technique to evaluate the surface area of a sorbent is the BET method [23]. The BET equation is used to calculate the amount of N_2 for monolayer coverage at 77 K. For activated carbon a simpler method using iodine [24] results in an equivalent iodine number (mg/g), which is usually in good accordance with BET measurements [25].

Pore Size Distribution. A grain is permeated by many pores of different width. The system of pores in an adsorbent particle may contain *submicropores* (< 0.4 nm) *micropores* (< 2 nm), *mesopores* (between 2 and 50 nm), and *macropores* (> 50 nm) [26, 27]. Some adsorbents (e.g., zeolitic molecular sieves) have a well-developed bimodal pore-size distribution so that the total pore volume consists of the volume of micropores and macropores.

The total pore volume and its distribution over the pore diameter characterizes the pore volume. Macro- and mesopores are determined by mercury porosimetry [28] or nitrogen adsorption [29], while micropores are estimated by molecular sieving [30]. All these techniques are widely discussed elsewhere [23, 31, 32]. Many adsorptives have broad pore size distributions, with macro- and mesopores for transport, and micropores for adsorption. A large micropore volume leads to a high specific internal surface available for accumulation of adsorptive. Specific surface areas are between 200 and 1800 m^2/g. A survey from simple to fractal pore models is given in [33]. There is an influence on sorption equilibria, hysteresis, and capillary condensation (see Chap. Thermodynamics, Equilibria, and Heat of Adsorption) and a further impact is on intraparticle diffusion, depending on pore size and tortuosity (see Chap. Kinetics).

Specific heat capacity and thermal conductivity (Table 1) are important if the bed of an adsorber is heated during loading by the heat of adsorption or if it is to be regenerated by heating in a temperature swing process.

Thermal stability is very important in industrial applications. Zeolite molecular sieves can withstand a maximum temperature of 800 °C, but with silica gel and aluminum oxide, the risk of damage begins above 300 °C.

Chemical Resistance. Small amounts of acid can damage aluminum oxide, but this adsorbent is insensitive to alkali. Adsorbents containing carbon are highly resistant to acid, except for oxidizing acids above ca. 400 °C. Chemically neutral silica gel is unaffected by acids except for aqueous hydrogen fluoride, but it is attacked by alkali. Polymerizable materials may plug the pores (see Section Separation of Nonaqueous Substances). Zeolitic molecular sieves are chemically resistant in the pH range 5 – 12.

Porous glass of various types can be used as adsorbents if organic acids are to be dried by adsorption, even though they exhibit neither the narrow micropore distribution nor the adsorption capacity of zeolites. Both their selectivity and their capacity are therefore poorer.

Aging of adsorbents may occur due to irreversible changes in structure at elevated temperature or plugging of pores by impurities and reaction products, (e.g., polymers). Such aging reduces both the adsorption capacity and the pore diffusion coefficient. Reactivation (see Section Reactivation of Adsorbents) involves a loss of capacity.

3.1. Oxidic Hydrophilic Adsorbents

Oxidic hydrophilic adsorbents such as silica gel, activated alumina (activated aluminum oxide), and zeolitic molecular sieves preferably adsorb polar adsorptives.

Activated Alumina. The surface and pore structure of aluminates or activated alumina can be controlled during the activation process and by using dopants [34]. The surface acidity is due either to Brønsted acid sites (i.e., OH groups that release protons) or Lewis acid sites (sites which can accept electrons). Due to the attractive price (ca. 1.5 €/kg) they are used in wastewater treatment and water purification. Removal of arsine, phosphate, chloride, and fluoride in water treatment has been reported [35]. They are very hydrophilic and are used in the liquid phase for drying of organics such as gasoline, cyclohexane, refrigerants, and lubricants [36].

In the gas-phase activated alumina is widely used for desiccation since it has higher capacity for water than zeolites. However, there is major competition with molecular sieves, which have high capacities at low vapor pressures. A further application is to remove acid gases such as HCl and HF from air and COS, CO_2, H_2S, and CS_2 from hydrocarbons. They are attractive for polar Lewis bases in the following order: aromatic and halogenated hydrocarbons < ethers, esters, ketones < amines, alcohols < carboxylic acids [37]. Impregnation of activated alumina with alkali metal oxides (i.e. K_2O, Na_2O, NH_4OH) yields alkalized alumina, which is used commercially for flue gas desulfurization [38]. CO_2 removal down to the 1 ppm level [39] and NO/NO_2 separation [40] are novel areas, whereas a traditionally wide application is as support material in catalysis.

Silica gel (⟶ Silica – Colloidal Silica) is composed of silicon dioxide and water (up to 40 %). Once water has been removed from silica gel, it exhibits a high adsorption capacity for water and other polar substances. Commercial silica gels are mesoporous with a pore size larger than 2 nm. The structure is a network of silicon atoms connected by oxygen atoms. The surface is saturated with hydroxyl groups, and the silanol number refers to the OH density (nm^{-2}). The high hydrogen-bonding capability of water with the OH groups makes silica gel an ideal desiccant. Silica gel doped with $CoCl_2$ changes color from blue to pink with increasing moisture content. Substances other than water, such as esters, ketones, ethers, and nitrogen-containing compounds (NH_3, amines, pyridines, etc.) that have hydrogen-bonding abilities can be adsorbed in a similar way. Silica gels can be easily regenerated by temperature swing (at about 150 °C) and are rather cheap materials (0.5 – 2 €/kg). To make the originally polar surface apolar, it is modified by grafting with a monomolecular layer of an organic ligand. The resulting materials have wide application in reversed-phase chromatography [41, 42], where most of the interactions are then reversed compared to pure silica.

Grafting agents for silica include various silanes [17, 43], of which 3-aminopropyltriethoxysilane is the most commonly used. Amine-grafted silicas adsorb CO_2 and H_2S rapidly with high capacities and are promising sorbents for selective adsorption in fields such as catalysis and environmental protection [17]. Molecularly imprinted microporous silica is applied in biosciences, e.g., for separation of racemates, but suffers from very low capacities [44].

Zeolites (⟶ Zeolites) are natural or synthetic crystalline aluminosilicates of alkali (x=2) or alkaline earth metals (x=1) with the stoichiometry: $M_xO{\cdot}Al_2O_3{\cdot}y\ SiO_2{\cdot}z\ H_2O$.

Zeolite crystals consist of SiO_4 and AlO_4 tetrahedra. By linking these polyhedra via four-, six-, eight-, and twelve-membered oxygen rings, cavities are formed which extend over the entire zeolite crystal. Figure 16 shows widely used zeolites, type A, type X, and type Y. In principle, a distinction is made between the larger α-cages and the smaller β-cages.

Each aluminum atom contributes a negative charge to the crystal structure, which must be offset by a cation. The cations (K^+, Li^+, Na^+, Ca^{2+}) can be exchanged during the manufacturing process and determine the channel diameter of the zeolite cage.

Type 5A (Ca^{2+} cation) has a channel diameter of about 0.43 nm, while the channel diameter of type 10X (Na^+ cation) is ca. 0.74 nm. Depending on the type of zeolite and the diameter of the exchanged ions, the channel diameter of the cages can be tailored to permit only molecules with smaller diameters to enter. This produces a sieve effect, hence the name “molecular sieve”. Table 2 gives an overview of some industrially important zeolites and their effective channel diameters. In addition, some areas of application are listed.

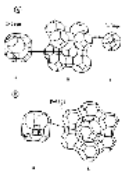

Figure 16. Cage structure of zeolitic molecular sieves

A. Type A zeolite: a) α-cage, diameter 1.14 nm, opening ca. 0.43 nm; b) Structure of type A; c) Truncated octahedron, β-cage, diameter 0.66 nm, opening 0.25 nm
B. Type X and Y zeolite: a) Largest cavity, diameter 1.16 nm, opening 0.74 nm; b) Structure of types X and Y

Table 2. Industrially important zeolites and their effective channel diameters [133]

Framework	Critationic form	Formula of typical unit cell	Window	Effective channel diameter, nm	Application
A	Na^+	$Na_{12}[(AlO_2)_{12}(SiO_2)_{12}]$	8-ring (obstructed)	0.38	desiccant, CO_2 removal from natural gas
A	Ca^{2+}	$Ca_5Na_2[(AlO_2)_{12}(SiO2)_{12}]$	8-ring (free)	0.44	linear alkane separation, air separation
A	K^+	$K_{12}[(AlO_2)_{12}(SiO_2)_{12}]$	8-ring (obstructed)	0.29	drying of cracked gas containing C_2H_4 , etc.
X	Na+	$Na_{86}[(AlO_2)_{86}(SiO_2)_{106}]$	12-ring	0.84	pressure-swing H_2 purification
X	Ca^{2+}	$Ca_{40}Na_6[(AlO_2)_{86}(SiO$	12-ring	0.80	removal of mercaptans from natural gas

		$_2)_{106}]$			
X	Sr^{2+}, *Ba^{2+}	$Sr_{21}Ba_{22}[(AlO_2)_{86}(SiO_2)_{106}]$	12-ring	0.80	xylene separation
Y	Na+	$Na_{56}[(AlO_2)_{56}(SiO_2)_{136}]$	12-ring	0.80	xylene separation
Y	K^+	$K_{5+}[(AlO_2)_{56}(SiO_2)_{136}]$	12-ring	0.80	xylene separation
Mordenite	Ag^+	$Ag_8[(AlO_2)_8(SiO_2)_{40}]$	12-ring	0.70	iodine and krypton removal
Mordenite	H^+	$H_8[(AlO_2)_8(SiO_2)_{40}]$			from nuclear off-gases
Silicalite		$(SiO_2)_{96}$	10-ring	0.60	removal of organic chemicals from water
ZSM-5	Na +	$Na_3[(AlO_2)_3(SiO_2)_{93}]$	10-ring	0.60	xylene separation

*Also K·BaX, where X = zeolite anion.

Thus, the effective channel diameter of the zeolite cages determines, during adsorptive separation on molecular sieves, whether or not a molecule having a certain critical molecular diameter can diffuse into the cage and be adsorbed there. Critical molecular diameters for some inorganic and organic substances are given in Table 3.

Table 3. Critical molecular diameters and molecular sieves suitable for adsorption on the basis of pore size

Molecule	Critical diameter, nm	Adsorption by 3 A	4 A	5A	10 X or 13X
He	0.2	x	x	x	x
Ne	0.32		x	x	x
Ar	0.38		x	x	x
Kr	0.39		?	x	x
Xe	0.47			x	x
H_2	0.24	x	x	x	x
O_2	0.28	x	x	x	x
N_2	0.30	x	x	x	x
Cl_2	0.82				x
Br_2	0.39		?	x	x
H_2O	0.26	x	x	x	x
CO	0.32 (0.28)		x	x	x
CO_2	0.28		x	x	x
NH_3	0.38	(x)*	x	x	x
$H2_S$			x	x	x
SO_2			x	x	x
CH_4	0.40		x	x	x
C_2H_6	0.44		x	x	x
C_3H_8 and higher *n*-alkanes	0.49			x	x
C_2H_2	0.22 (0.24)	x	x	x	x
C_2H_4	0.43		x	x	x
C_3H_6	0.50		x	x	x
C_4H_8	0.51			x	x
C_4 and higher isoalkanes	0.56				x
Cyclopropane	0.48			x	x
Cyclopentane, cyclohexane	0.61				x
Benzene	0.57				x
CH_3OH	0.30		x	x	x
C_2H_5OH			x	x	x
n-C_3H_7OH			(x)*	x	x
CCl_4	0.69				x
CS_2				x	x

*(x) = suitable

The manufacturing process for zeolites leads to a bimodal distribution of pore radii typical of oxidic adsorbents. The differential distribution of pore radii in oxidic adsorbents is shown in Figure 17. Molecular sieves exhibit two clearly developed maxima: one for micropores with a radius of 0.2 – 0.3 nm and another in the macropore range at ca. 300 nm. The macropores are interspersed between individual microcrystals. They have a very small specific surface and hence a negligible adsorption capacity. Nevertheless, diffusion in the macropores often determines the rate of sorption (see Chap. Kinetics).

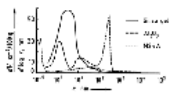

Figure 17. Distribution of pore radii in oxidic adsorbents

The most commonly used zeolites and molecular sieves in adsorption and ion exchange are the types A, X, and Y. However, by substituting other elements for Al and/or Si an infinite number of molecular sieves (which are formally not zeolites) can be created. This is of interest for catalysis or petrochemical industry, when for instance $[TiO_h]$ provides oxidation sites for redox reactions [17, 45]. While Linde's interest was in the development of new sorbents (zeolites A, X, and Y), Mobil was more interested in catalysis when developing the ZSM zeolites, the most useful of which is ZSM-5, prepared by using amine templates to produce zeolite crystals of a specific channel diameter [45]. Linde independently developed a similar zeolite, called silicalite [46], in which both channel sizes are comparable to the size of many important solutes such as *n*-paraffins, isoparaffins, aromatics, and their derivatives.

A high aluminum content promotes adsorption of polar components. Dealuminated Y zeolites (DAY) are more hydrophilic and have a higher affinity for nonpolar organic

substances from wet exhaust gases [47]. Zeolites do not burn and are stable up to 700 – 1000 °C, which can be used for cleaning purposes [48]. They are not inert against acidic attack, but a higher silicon content improves resistance against strong acids but make it worse against strong bases [49]. Generally speaking, the relatively high price of zeolites of 1 – 4€/kg makes them largely suitable for applications in special gas separation processes, and their use as desiccants is not really promising due to the high regeneration temperatures of about 350 °C.

By using quaternary ammonium surfactants as templates, it was possible to synthese a new family of ordered, mesoporous silicates or aluminosilicates. MCM-41 is the best known and has a simple structure due to the rod-type hexagonal array of cylindrical micelles. The silicate precursors condense on the wall of the templates, which are subsequently removed by oxidation to leave behind a honeycomblike structure. Despite the uniform pore structure and pore sizes of up to 30 nm, application in adsorption is not likely owing to the high price and low capacities. However, in modified forms it may be interesting for niche applications. Metal recovery from waste waters with thiol-grafted MCM-41 has been reported [50]; regeneration is possible with HCl.

Pillared Clays. Pillared interlayered clays (PILCs) are aluminosilicates with cations and a two-dimentional layer structure. The basis is montmorillonite with a ratio of silicic acid to clay of 4:1, like bentonite. Each layer consists of a core alumina layer surrounded above and below by silica (Fig. 18). These layers are then sandwiched with pillars of metal oxide. The pore sizes in the PILCs are not limited by the interlayer spacing but by the interpillar spacing [51]. Pore sizes are near 0.4 and 0.6 nm and can be controlled and tailored with potential applications as molecular sieves [36]. They are considerably less hydrophobic than zeolites, silica gel, and activated alumina [52]. The attractive price of 1 – 1.5 €/kg promotes application in raffination of sweet oil, mineral oil, oil residues, and wastewaters.

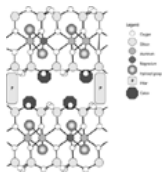

Figure 18. Montmorillonite crystal structure

3.2. Carbon-Containing Adsorbents

Carbon-containing adsorbents such as activated carbon, activated coke, and carbon molecular sieves (⟶ Carbon – Activated Carbon) are highly porous solids which, because of their surface properties, preferentially accumulate organic substances and nonpolar adsorptives. Their amorphous skeleton consists of microcrystallites with a graphite lattice. Carbon adsorbents are produced from organic materials such as wood, peat, coal, and petroleum coke. During manufacture, the particles are first formed, then carbonized at low temperature, and finally activated (⟶ Carbon – Activating Furnaces). Activated carbon is used in the form of grains (usually for adsorption in the gaseous phase) or powder (usually in liquid-phase adsorption). Carbon-containing granular adsorbents are generally produced in extruders and sold in cylindrical shape. Pellets and crushed granules are also available.

Activated carbon fibers (ACFs) have been commercially available for gas-phase applications since the 1960s. They have high surface areas (1000 – 2000m^3/g), narrow and uniform pore size distribution, and are graphitic and elastic. Their advantage is in a high mass-transfer rate and the possibility of electrothermal regeneration. The graphitic nature makes ACFs heat resistant, so they are inflammable up to 1000 °C. Also the lack of fines and the high tensile strength of the fibers is an advantage in use. However, so far only mats are available due to problems in final tailoring, and the high price (ca. 200 €/kg) only allows special applications. Comprehensive reviews on ACFs are given in [53, 54].

Activated carbon is characterized by large surface areas of between 300 and 4000 m^2/g, which exceed those of all other sorbents (see Table 1). It has a polymodal pore size distribution, whereby microporosity (1 – 2.5 nm) prevails in gas-phase, and mesoporosity (> 3 nm) in liquid-phase applications. The specific micropore volume of activated carbon, i.e., the specific volume of pores smaller than 2 nm, is about 0.25 – 0.4 cm^3/g (see Table 1). This shows clearly that the adsorption capacity of activated carbon is determined almost exclusively by the micropores and their volume.

Figure 19 shows the distribution of pore radii in activated carbon, activated coke, and carbon molecular sieves. The carbon-containing adsorbents are distinguished by different pore sizes in the micropore range. For activated carbon, the maximum is at about 0.8 – 1 nm; for activated coke, between 0.5 and 0.7 nm, and for carbon molecular sieves, between 0.2 and 0.3 nm. This means that neither activated carbon nor activated coke is capable in practice of adsorbing only molecules below a certain molecular diameter.

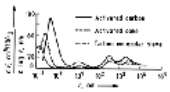

Figure 19. Distributions of pore radii in carbon adsorbents

The surface of activated carbon is nonpolar or slightly polar as a result of acidic surface oxide groups and inorganic impurities. They exhibit van der Waals interactions (i.e., dispersion and expulsion forces) with the solute. The electrical charges are too weak to contribute significantly to the adsorption process. The low heat of adsorption makes stripping of adsorbed molecules relatively easy, and hence energy requirements for regeneration are low. Aromatic adsorptives are much better adsorbed than aliphatic ones, due to higher van der Waals interactions of the π-electron system. In general, adsorption capacity for carboxylic acids, alcohols, and esters increases with increasing polarity.

The hydrophobic nature of activated carbon makes it as an ideal adsorbent for treatment of wet gaseous feed streams. The adsorption of moisture is generally low, but at humidities higher than 40 % water sorption due to capillary condensation is found.

The applications of activated carbon can be extended by impregnating it with Ag, Cu, S, Zn, or I for specific recovery of a certain solute, e.g., H_2S from an CS_2/H_2S mixture with I impregnation [55]. A disadvantage of activated carbon is its low mechanical and abrasion stability. Operation in moving or fluidized beds therefore needs the harder activated coals. Besides this, SO_2 in a gas stream may oxidize carbon resulting in H_2SO_4 and “glued” activated carbon pellets, which results in inhomogeneous gas flow in the bed. Generally, inhomogenities hinder the convective transport of the heat released by adsorption and thus result in hot spots which, when they exceed the carbon ignition temperature, cause fires [56, 57].

The broad applicability of activated carbons and their attractive price of 0.25 – 2 €/kg make them the leader in worldwide sales [58]. They also have broad applications in adsorption from liquid solutions, e.g., decolorization of sugar, fruit juice, chemicals, treatment of potable water and wastewater, gold recovery in the mining industry, etc. Adsorption from liquids is more complicated than from the gas phase, since the ternary solute – solvent – adsorbent interaction is now involved. For a series of environmentally harmful substances (phenolic, aromatic, and chlorinated aromatic compounds, organic pesticides, carbon tetrachloride and other volatile organic compounds (VOC), and metal ions), isotherms are available in [59].

Carbon molecular sieves (CMS) have a mean pore diameter of 0.3 – 0.9 nm, on the same order of the magnitude of the critical molecule size of small molecules (Table 1). The distribution of the micropore radii in a CMS is not as narrow as in a ZMS, but they have the same properties towards water as activated carbons. They are made from anthracite and hard coal by partial oxidation and subsequent thermal treatment. Thermal treatment, in this case, means the thermal decomposition of synthetic polymers [e.g., polyethylene, poly(vinyl chloride), poly(vinylidene chloride)]. This treatment creates micropores with a relatively narrow distribution of pore radii. The pore structure can be modified to a certain extent by repeated thermal treatment, e.g., controlled cracking of hydrocarbons in the micropores and partial gasification.

The main use of CMS is in separation processes involving wet gas streams. Since they have no cations they are even more hydrophobic than activated carbon. The dominant use is in PSA processes (see Chap. Gas Phase Processes) where kinetic separation of N_2/O_2, ethylene/ethane or CH_4/CO_2 are performed. The last-named is is used for landfill gas with about 50 % CO_2 and tertiary oil recovery with around 80 % CO_2 [60]. They show excellent separation properties (selectivity, permeability) and high stability (chemical and thermal) and often have the advantage of operating in a harmful environment [61]. However, research in this field is just beyond the initial stages [62] and not all the drawbacks have been solved yet [63].

Further promising types of new sorbents are fullerenes and carbon nanotubes [64] with potential applications in hydrogen storage. The surface of the nanotubes contains a high density of π electrons and is highly aromatic. The nanotubes are therefore highly graphitic and it is expected that carbon nanotubes will adsorb molecules more strongly than activated carbons [63].

Polymeric resins are synthetic porous polymers with a large internal surface; their main application is in liquid adsorption. The raw materials are styrene, divinylbenzene, and acrylates. The macroporous polymeric resins can be further treated to attach functional groups to generate resins for ion exchange and chromatography (⟶ Ion Exchangers).

Polymeric resins have similar adsorption characteristics to activated carbon and are of high mechanical stability, but they swell in nonaqueous solvents [65, 66]. They are chemically stable at high pH values, and despite their 3 – 5 times higher price relative to activated carbon, they are dominant in liquid sorption processes. They can remove aromatic or chlorinated hydrocarbons, phenols, and pesticides from wastewater [67] and find a broad application in the food and pharmaceutical fields [37] for the purification of vitamins, antibiotics, and food products (sugars, fruit juices, dairy products, amino acids).

3.3. Chemical Complexing Sorbents

Traditional adsorption of solutes uses van der Waals or/and electrostatic interactions. For separation of ultrapure components, chemical bonds offer stronger and more specific interactions. A large variety of stationary sorbent phases are available serving different needs with respect to capacity and selectivity [17]. In addition, (monolayer-) impregnated porous substrates offer further applications. Supported monolayer salts and metal oxides, such as CuCl on activated carbon or alumina, have been commercially used for CO_2 recovery and separation [62], and Cu^+- and Ag^+-impregnated zeolites are, e.g,. used to remove dienes from olefins [68]. A further approach is to impregnate porous support materials with liquid ion exchangers, which are cheap and readily available. Cation, anion, chelating, and neutral liquid ion exchangers can be used; the chemistry is discussed elsewhere [69, 70]. Despite the fact that the liquid ion exchangers are only physically adsorbed on the support material, bleeding in liquid adsorption processes is negligible [71]. The main application is to remove traces of heavy metals such as Zn, Cu, Pb, and Cd from effluents, [72, 73]. However, large-scale applications have not been reported so far, which gives prospects to applications in new fields [74].

[Top of Page]

4. Thermodynamics, Equilibria, and Heat of Adsorption

The basic approach to adsorptive equilibria, i.e., loading versus partial pressure, must provide a suitable description of both the equation of state for the fluid phase and the adsorption energy levels due to the various adsorption sites involved. The knowledge of the distribution of the molecules in the fluid phase and on the adsorption sites is the key issue in understanding and describing adsorptive equilibria. Statistical considerations lead to an occupation index for each individual site, dependent on the chemical potential of the gas phase, i.e., its pressure and temperature. However, the statistical approach is different for surface adsorption with only one molecule per energy level (Fermi – Dirac case, isotherm type I), with a restricted number of adsorbed molecules per energy level (BET isotherm types II, IV) and with pore condensation (Bose – Einstein case, isotherm types III, V). Each approach exhibits a characteristic picture for adsorption isotherms, as is apparent in the IUPAC classification of types (Fig. 20).

An ideal pure gas is defined by two parameters: pressure p and temperature T. The energetically homogeneous surface is described by two adsorption parameters, the adsorption energy level μ_a and saturation loading q_0 (equivalent to the specific surface area). The equilibrium condition between fluid and adsorbed phase reduces the four independent parameters to three (Gibbs –Duhem relation). The appropriate homogeneous three-parameter equation was derived by LANGMUIR [2]. Equilibria with real gases have two further fugacity parameters (critical pressure and temperature), while real surfaces with energetic heterogeneities are best approached by a suitable energy distribution of sites in connection with a further parameter characterizing the distribution energy width. Hence, typical adsorption equilibria for an ideal gas phase but a heterogeneous adsorbent are four-parameter equations like the Toth, Unilan, Langmuir – Freundlich, and Dubinin relations [6, 8, 75, 76]. To take account of condensation effects, parameters for liquid properties are necessary, as in, e.g., the BET and the Ruthven equations [4, 77].

In practice some caution is advisable if condensation can not be excluded in an adsorption processes because of conditions in the vicinity of the dew point. Then effective regeneration is thwarted by hysteresis effects due to the loss of the surface degree of freedom.

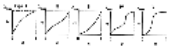

Figure 20. Various types of adsorption isotherms

4.1. Explicit Equilibria

The basic variables to describe adsorption equilibria are the fluid partial pressure p_i (or fugacity f_i for real-gas conditions) for each component i, the temperature T, the fractional surface (monolayer) coverage $\Theta_i=q_i/q_{0i}$ and the adsorptive vapor pressure Φ_i at the standardized condition Θ_i=1/2 (adsorptive saturation pressure). Here q_i and q_{0i} are equilibrium and saturation loading. With condensable components also the liquid saturation pressure Φ_{Li} is important, as well as the relation of liquid to adsorptive gas phase saturation pressure $b_i= \Phi_{Li}/ \Phi_i$. The fractional volumetric loading V_i/V_{is} accounts for the relative pore filling with respect to the entire pore volume available for wetting, which typically is 3 – 5 times the volume surface coverage. Figure 21 shows an isotherm exhibiting surface coverage, pore filling, and condensation. Figure 22 presents water and n-hexane isotherms on several porous adsorbents.

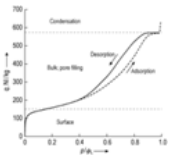

Figure 21. N_2 loading on silica gel at 77.4 K

For p/ Φ_L < 0: surface adsorption; for p/ Φ_L > 0: pore filling (hysteresis); near p/ Φ_L=condensation

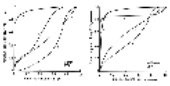

Figure 22. Relative sorbate volume of water and n-hexane on different adsorbents

a) Molecular sieve; b) Silica gel; c) Activated aluminum oxide; d) Activated carbon

4.1.1. *Single-Component Pure-Gas Equilibria*

Homogeneous Surfaces. Under the condition that all adsorption sites are energetically homogeneous with adsorption enthalpy H_a and each site can be occupied by one or no molecule the Langmuir equation [3] holds

$$\Theta = \frac{q}{q_0} = \frac{1}{1+\phi/p} \tag{4.1}$$

where the surface adsorption saturation pressure

$$\phi = p_0 e^{A-H_a/(RT)} \tag{4.2}$$

is of *Arrhenius* form. This reference pressure at the condition Θ=1/2 normally is about three orders of magnitude smaller than the corresponding liquid saturation Φ_L at pore condensation with vapors. The surface adsorption saturation pressure Φ can be calculated from Φ_0, given for a reference temperature T_0 as

$$\phi = \phi_0 e^{(H_a/R)(1/T_0-1/T)} \tag{4.3}$$

The Langmuir equation can be transformed into

$$\frac{\Theta}{1-\Theta} = \frac{p}{\phi} \tag{4.4}$$

with the point of symmetry Θ=1/2 and $p= \Phi$. In the low-coverage range the Langmuir equation approaches the Henry equation

$$q = hp \tag{4.5}$$

where $h=q_0/ \Phi$ is the Henry constant obeying the van't Hoff equation

$$R\frac{\partial \ln(h/h_0)}{\partial(1/T)} = H_a \tag{4.6}$$

where h_0 is the Henry constant at a reference temperature T_0. The isosteric equilibrium pressure p corresponds to the Clausius – Clapeyron relation

$$\left(\frac{\partial p}{\partial T}\right)_{q=\text{const}} = \frac{pH_a}{RT^2} \tag{4.7}$$

The Langmuir equation has three independent parameters: q_0, Φ_0, and H_a.

Heterogeneous Surfaces. If, instead of a narrow distribution of adsorption site energies peaking at H_a, the adsorbent has a broadened distribution with variance $\in$ the isotherms are flattened near Θ=1/2. The equilibrium isotherm then must be calculated by an integral, weighting the homogeneous equilibria for each energy level with the energy distribution function. From the manifold of known approaches, each related to a certain energy distribution function, the discussion is restricted here to two major representations: the uniform and the hyperbolic energy distributions, which result in the UniLan and the GM equations.

The *UniLan equation* [76] is related to a rectangular energy distribution within the range of $H_a \pm \in$

$$\Theta = \frac{RT}{2\varepsilon} \ln \frac{1 + \left[p/\phi\right] e^{\varepsilon/(RT)}}{1 + \left[p/\phi\right] e^{-\varepsilon/(RT)}} \tag{4.8}$$

At higher temperatures $T > \in/R$ the UniLan equation approaches exactly the Langmuir equation for the homogeneous case.

The second practical equation for heterogeneous surfaces is based on the smoother hyperbolic distribution $1/[\in(2 + e^{\omega} + e^{-\omega})]$ where $\omega = (H - H_a)/\in$, presenting a Gaussian-like shape of broadness $4\in$, but with an exponential decay. The approximately related *GM isotherm* [78, 79] is of the form

$$\Theta = 1/\left\{1 + [\phi/p]^{a(T)}\right\} \tag{4.9}$$

where

$$a(T) = RT/\varepsilon \quad \text{if } RT < \varepsilon \quad \text{and } a(T) = 1 \quad \text{elsewhere} \tag{4.10}$$

This equation is of the Langmuir – Freundlich type but includes distinct temperature dependence of the exponent $a(T)$. Again, for $T > \in/R$ the equation approaches the Langmuir equation. High-accuracy modifications of Equation (4.10) use $a(T) = 1/[m + \in/(RT)]$ where m is an adjustable parameter [78]. Thus, homogeneity or heterogeneity is not simply a question of the absolute broadness of the distribution $\in$ of the adsorption energy level, but is defined by the ratio of $\in$ and the thermal fluctuation RT. At higher temperatures the adsorbent's heterogeneity is overridden by thermal fluctuations filling all surface sites simultaneously. The result then resembles a homogeneous situation, where the adsorption energy distribution is narrow with respect to RT.

With the UniLan and GM equations, the homogeneous isotherm parameters q_0, Φ_0, and H_a are supplemented by a fourth independent parameter: the energetic heterogeneity $\in$ of the site distribution. Table 4 lists some examples for molecular sieve 5A. Figures 23-25 present methane, ethane, and propane isotherms on activated carbon. They show an increasing curvature (nonideality) with increasing C-number. Therefore the appropriate approximations are the UniLan or GM-equation. Figures 26 and 27 show the strongly nonideal water and CO_2 isotherms on molecular sieve 5A.

Table 4. Equilibrium parameters for several components at molecular sieve 5A using the UniLan approximation (Eq. 4.8) at T_0 = 273.15 K. For comparison: RT_0 = 2.27 kJ/mol.

Component	q_0 (0 °C), mol/kg	H_a, kJ/mol	$\in$, kJ/mol	Φ_0 (0 °C), bar
CO_2	4.91	38.6	14.0	0.007
C_2H_4	3.25	30.7	14.0	0.029
CO	3.35	27.4	5.2	0.937
CH_4	3.66	17.6	3.5	4.70
N_2	2.95	17.9	3.5	4.13
O_2	3.93	13.0	2.2	21.6
H_2	2.01	5.6	1.5	101.0

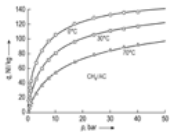

Figure 23. Methane isotherms at 0, 30, and 70 °C on activated carbon (Norit). Approximation: GM isotherm (Eq. 4.9).

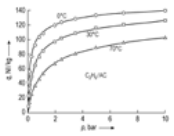

Figure 24. Ethane isotherms at 0, 30, and 70 °C on activated carbon (Norit) Approximation: GM isotherm (Eq. 4.9).

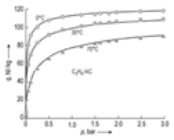

Figure 25. Propane isotherms at 0, 30, and 70 °C on activated carbon (Norit) Approximation: GM isotherm (Eq. 4.9).

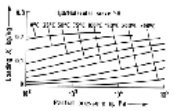

Figure 26. Adsorption isotherms of water on molecular sieve 5 A

Figure 27. Adsorption isotherms of carbon dioxide on molecular sieve 5 A

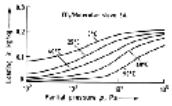

Homogeneous Surface with Additional Multilayer Pore Condensation. If the first (surface) layer is adsorbed with uniform enthalpy H_a and the second to the *N*th layer are condensed with liquefication enthalpy H_L the equilibrium is represented by the BET equation [4] by using $\varphi = p/\Phi_L$, $b = \Phi/\Phi_L$, and $\Phi_L = p_0 e^{A-H_L/(RT)}$

$$\Theta = \frac{b\varphi}{1-\varphi}\,\frac{1-(N+1)\varphi^N + N\varphi^{N+1}}{1+(b-1)\varphi - b\varphi^{N+1}} \tag{4.11}$$

Here $\Theta = q/q_{0mon}$ is related to monolayer saturation loading q_{0mon}. For N=1 the equation results in the Langmuir formula. The equation is widely used to determine the BET surface area.

Vapors. For condensable components DUBININ found a correlation of adsorptive equilibria at various temperatures and for similar components via the Polanyi binding potential [1]

$$\mu_b = \mu_a - \mu_L = -RT \ln (p/\phi_L)\,, \tag{4.12}$$

where μ_a and μ_L are adsorption and liquefaction energies. With the pore filling volume

$$V = q v_m \tag{4.13}$$

where v_m is the liquid molar volume (cm^3/mol) under standard conditions (e.g., vapor pressure of 1 bar), the general Lewis correlation is

$$V = F(\mu_b/\beta) \tag{4.14}$$

with an experimentally fitted function *F* and an affinity coefficient β. The affinity coefficient was taken by *Lewis* as $\beta = v_m$ [9]. From that, DUBININ assumed a Gaussian-like energy distribution, leading to an isotherm [5, 6]

$$\Theta = \exp\{-(\mu_b/\varepsilon)^a\} = \exp\{[(RT/\varepsilon)\ln(p/\phi_L)]^a\} \tag{4.15}$$

with an adjustable energy spread $\in$ and an constant exponent *a* of 1 –3, preferably 2.

As mentioned above, liquid-phase adsorption is different to gas phase adsorption. In the latter one there are distractive adsorption sites and — a defined interface between adsorbed/condensed phase and gas phase. Liquid adsorption always involves replacement of a solvent molecule by the adsorptive, but there is no distinct border between the liquid adsorbate and liquid bulk. Also heat effects are virtually negligible. However, the isotherms used for liquid adsorption are the same as those for the gas phase if the influence of the solvent phase is neglected.

According to the Dubinin – Polanyi theory, a characteristic temperature-independent relation exists between the adsorption potential and the adsorbed volume for each adsorbent – adsorbate system. Such characteristic curves for various adsorptives are shown in Figures 28-30 for molecular sieves, silica gel, and activated carbon adsorbents. Strictly speaking, such temperature-independent characteristic curves can be valid only for nonpolar systems, because only van der Waals forces are independent of temperature. The characteristic curve is thermodynamically inconsistent with Henry's law because the adsorption isotherms for small partial pressures do not end up in a Henry line. However, the Dubinin – Polanyi theory can also be applied to many polar systems with sufficient accuracy.

The Dubinin equation is suitable mainly for organic adsorptives on activated carbon. The exponent *a*=1 often results for wide-pore activated carbons. For this special case, the Dubinin equation can be transformed into the Freundlich relation. The exponent *a*=2 is often valid for narrow-pore activated carbons, while *a*=3 frequently holds for molecular sieves.

The affinity coefficient β characterizes the wetting behavior of an adsorptive on an adsorbent. Thermodynamic considerations result in the affinity coefficient being interpreted as the free enthalpy *G*° of wetting [80]. If *G*° is known, the adsorption isotherms of various adsorptives can be represented in one diagram (see Fig. 31). In this diagram, the binding potential μ_b referred to the free enthalpy *G*°, i.e., the quantity $\mu_b/G°$ is plotted against the degree of pore filling V_i/V_{is} for benzene, cyclohexane, *n*-heptane, and 1,2-dichloroethane on silica gel at 30 °C.

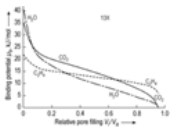

Figure 28. Characteristic curves for water, carbon dioxide, and ethylene on molecular sieve 13X

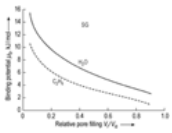

Figure 29. Characteristic curves for water and ethane on silica gel

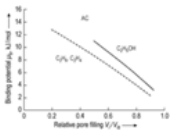

Figure 30. Characteristic curves for ethanol, ethylene, and ethane on activated carbon

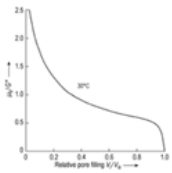

Figure 31. Ratio of binding potential $\mu_b = RT\ln(\varphi)$ to free enthalpy of immersion G^o with pore filling for benzene, cyclohexane, *n*-heptane and 1,2 dichloroethane on silica gel

Capillary condensation results in adsorption – desorption hysteresis. This effect occurs at higher sorptive concentrations and results in an almost vertical slope of the isotherm (e.g., type II isotherm).

In capillary condensation from the gas phase the binding potential μ_b is correlated with the pore radius *r*, the surface tension $\sigma = \sigma_0 \cos\delta$ with contact angle δ, and the molar

volume v_m by using the Gibbs – Thomson equation

$$\mu_b = 2\sigma v_m / r_0. \tag{4.16}$$

For each relative saturation $\varphi_L = p/\Phi_L$, all pores with $r_0 < r_\varphi$ are considered to be filled completely, where

$$r_\varphi = \frac{2\sigma v_m}{RT \ln(1/\varphi_L)}. \tag{4.17}$$

From measured $\varphi_L(\Theta)$ values the pore radius distribution can be evaluated by inserting the values in Equation (4.17). The smaller the pores, the more dramatic is the effect, and it cannot be neglected for pores smaller than 100 nm. During adsorption the pore wall is dry and δ is smaller compared to the wet pore wall in desorption, where δ higher. The difference in vapor pressure Φ_L is then seen in the form of hysteresis, as depicted in Figure 32. Type H1 hysteresis is observed with cylindrical capillaries and zeolites, type H2 mainly with silica gel, type H3 with wide pores having small openings, and H4 with microporous active carbon having mesopores.

The affinity coefficient β_j of adsorptive *j* can be recalculated from the corresponding coefficient of adsorptive *i* with the aid of Equation (4.14 ff) and the Dubinin equation:

$$\beta_i/\beta_j = v_{mi}\sigma_i/(v_{mj}\sigma_j) \tag{4.18}$$

This procedure is successful for the system activated carbon – organic adsorptive [81]; narrow-pore (pore volume 0.45 cm^3/g) and wide-pore (pore volume 0.63 cm^3/g) activated carbon were used in these studies. The reference adsorptive was benzene.

In Figure 33, the specific adsorbed volume V_i in cubic centimeters of fluid per gram of adsorbent is plotted as a function of $[\log(p/\Phi_L)]^a$ (exponent *a*=2 in the Dubinin equation) for the adsorptives phenol, benzene, acetone, and methanol for both measured values and values recalculated according to Equation (4.15). The agreement between values is quite good at high degrees of pore filling (V_i=0.45 cm^3/g) or at very high relative saturation φ_L; it becomes worse as loading decreases. A similar result was found for wide-pore activated carbon.

Table 5 lists measured and calculated affinity coefficients β/β_B for several organic adsorptives, each related to the value β_B of benzene. It shows that adsorption isotherms of other adsorptives can indeed be estimated to a first approximation via the surface tension σ and the molar volume v_m if the isotherm of a reference adsorptive is known. According to Table 5, the affinity coefficient of benzene for the activated carbon used is β_B=18 kJ/mol, but it depends in principle on the type of carbon and especially on its saturation volume. The lower the saturation volume, the higher is the affinity coefficient; for benzene, 15 < β < 25 kJ/mol [82]. This tendency can also be expected from Equations (4.16-4.18). For a given degree of saturation φ_L and for adsorbents with the same structural parameter *C*, from the Dubinin equation

$$\mu_b/\beta = \text{const} \tag{4.19}$$

This means that the affinity coefficient increases with increasing surface tension σ of the adsorptive and decreasing radius r_f of the concave fluid surface in the pore.

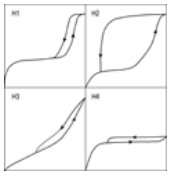

Figure 32. IUPAC classification of hysteresis adsorption curves

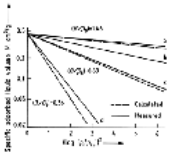

Figure 33. Comparison of measured and calculated isotherms of narrow-pore activated carbon in the Dubinin equation (reference adsorptive: benzene)

a) Phenol, β/β_B=1.46; b) Benzene, β/β_B; c) Acetone, β/β_B=0.69; d) Methanol, β/β_B=0.36

Table 5. Relative affinity coefficients β/β_B of different organic adsorptives on wide-pore activated carbon (V_s = 0.63 cm^3/g) [81]

Adsorptive	β/β_B (exptl)	β/β_B (calcd)
Phenol	1.48	1.46
Toluene	1.12	1.17
Carbon tetrachloride	1.02	1.01
Benzene (β_B= 18 kJ/mol)	1	1
Cyclohexane	0.96	1.05
Ethyl acetate	0.93	0.92
Chloroform	0.88	0.86
Acetone	0.72	0.69
Methylene chloride	0.70	0.71
Methanol	0.38	0.36

4.1.2. *Temperature Dependence of the Saturation Loading*

Real adsorption equilibria exhibit a distinct dependence of the available saturation loading on temperature. While this effect is not fully understood yet, it can be explained by thermal expansion of the adsorbed layer reducing its density. Therefore, the saturation loading near the reference state T_0, q_0 is approximated as [78]

$$q_0(T) = q_0(1 - [T - T_0]\alpha_T) \tag{4.20}$$

where α_T is an appropriate adsorbate heat expansion coefficient. At constant loading *q* the fractional coverage Θ in Section Differential Adsorption Enthalpies is taken to be temperature dependent, by inserting $q_0(T)$ instead of q_0.

4.2. Adsorption Enthalpy

4.2.1. *Differential Adsorption Enthalpies*

Even for energetically heterogeneous surfaces the mean adsorption enthalpy H_a is constant, as it is for homogeneous surfaces, if the temperature is high enough to compensate the energetic heterogeneity $\in$ by the thermal fluctuation *RT*. Then all energetic levels are occupied nearly simultaneously. However the spread of the energy levels $H_a \pm \in$ produces distinct selectivites of the various components at each level. The reason for this is the nonlinear response of the selectivity via its Boltzmann factors leading to deviations from the selectivity based only on mean adsorption enthalpies. At low temperatures the adsorption enthalpy shows a distinct coverage dependence for

heterogeneous surfaces which allows the recalculation of the energy level distribution. Formally, the differential isosteric heat of adsorption H_a may be combined as the constant heat of liquefication H_L for subcritical components and the heat of binding H_b to

$$H_a = H_L + H_b \tag{4.21}$$

where each part obeys the related Clausius – Clapeyron equations for chemical equilbria at constant coverage (isosteric conditions)

$$H_x = -R \frac{\partial \ln (p_x/p_o)_{q=\text{const}}}{\partial(1/T)}. \tag{4.22}$$

Here the index x stands for a,L,b and the corresponding pressures p_x are: p_a=adsorptive equilibrium pressure, $p_L = \Phi_L$=liquid saturation pressure, and $p_b = \varphi = p_a / \Phi_L$. The isosteric heat of adsorption can be derived from the experimental slopes of the equilibria as $\ln(p/p_0)$ against T^{-1} at constant adsorptive loading. Examples of mean adsorption and binding enthalpies on molecular sieve 13X are presented in Table 6 for several components.

On heterogeneous surfaces the differential isosteric heat of adsorption is generally coverage-dependent at lower temperatures. For three of our isotherm examples, Langmuir (L; Eq. 4.1), UniLan (UL; Eq. 4.8) and GM; (Eq. 4.9) they are calculated using Equation (4.22) with q_0=const.

$$H_a(\Theta) = \left\{ \begin{array}{c} H_a = \text{const} \\ H_a + \varepsilon - 2\varepsilon\Theta \\ H_a - \varepsilon \ln (\Theta/[1-\Theta]) \end{array} \right\} \begin{array}{c} (L) \\ (UL) \\ (GM) \end{array} \tag{4.23}$$

The GM equation has weak singularities at Θ=0 and Θ=1 which vanish for the integral adsorption enthalpy. Using a temperature-dependent saturation coverage gives an additional configuration enthalpy [83].

Determination of heats is demonstrated by using the system propane–activated carbon [84]. Figure 25 shows the propane loading of activated carbon as a function of the partial pressure at different temperatures. The logarithm of the partial pressure p_i or of the relative saturation is plotted against the reciprocal of absolute temperature [see Figs. 34A and B for isosteres, i.e., lines of equal loading (x_i=const.)]. Such isosteres are straight lines over a somewhat limited temperature range. If the slope of the lines is multiplied by the gas constant R, the heats of adsorption and binding, respectively, are obtained according to Equations (4.22); these heats are shown in Figure 35 as a function of loading.

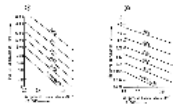

Figure 34. Partial pressure p_i (A) and relative saturation φ (B) as a function of the reciprocal of absolute temperature 1/T for propane on activated carbon Loading of adsorbent X, kg/kg is given for each line.

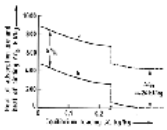

Figure 35. Specific heat of adsorption and heat of binding as a function of loading X for propane on activated carbon

a) Specific heat of adsorption Δh; b) Specific heat of binding Δh_B

Table 6. Heats of adsorption ΔH and binding ΔH_B (estimated) of various materials on zeolitic molecular sieve 13X [85]

Adsorptive	Dipole moment, D	ΔH_i, kJ/mol	$\Delta H_{B,i}$, kJ/mol
Argon	0.0	12	5.5
Oxygen	0.0	14	7.2
Krypton	0.0	15	6.0
Nitrogen	0.0	19	13.4
Carbon monoxide	0.1	24	17.9
Carbon dioxide	0.0	46	20.7
Ammonia	1.46	67	43.3
Water	1.8	95	54.3

4.2.2. *Integral Adsorption Enthalpies*

Evaluating heat balances $Q_{s,e}$ due to changes in the adsorptive loading $\Delta q = q_e - q_s$ requires the use of the integral adsorption enthalpy $I(\Theta)$

$$Q_{s,e} = q_0 m_a [I(\Theta_s) - I(\Theta_e)] \tag{4.24}$$

where $Q_{s,e} = q_{s,e}/q_0$ and m_a is the amount of adsorbent. For q_0=const. the integral adsorption enthalpy is

$$I(\Theta') = \int_0^{\Theta'} d\Theta H_a(\Theta) \tag{4.25}$$

and can again be evaluated for the three equilibria (Eq. 4.23; L, UL, and GM)

$$I(\Theta) = \left\{ \begin{array}{c} \Theta H_a \\ \Theta H_a + \varepsilon\Theta(1-\Theta) \\ \Theta H_a - \varepsilon\{\Theta \ln (\Theta) + (1-\Theta) \ln (1-\Theta)\} \end{array} \right\} \begin{array}{c} (\text{L}) \\ (\text{UL}) \\ (\text{GM}). \end{array} \tag{4.26}$$

4.3. Mixture Equilibria

In industrial practice, fluids usually contain more than one adsorbable component. Even for air – adsorptive or general carrier gas – adsorptive mixtures, the adsorption of air or carrier gas at higher pressure must be considered. However, it can be neglected at low total pressure and thus also at low partial pressure of air. Water vapor has a high interaction energy with oxidic adsorbents and is therefore strongly adsorbed by them. Because many gaseous mixtures contain water vapor in addition to the adsorptive to be separated, the adsorption capacity for the latter is often substantially reduced.

As an example, Figure 36 shows the loading of carbon dioxide at 25 °C as a function of its partial pressure at various equilibrium loadings of water. If a 5A molecular sieve (MS) is preloaded with 0.09 g of H_2O per gram of MS, the equilibrium loading of carbon dioxide drops to a fraction of the loading without water vapor. Therefore, low residual water content of the adsorbent during regeneration must be ensured (see Section Removal of Organic Components from Exhaust Air). Figure 37 shows the dew point of the purge gas used for regeneration as a function of the temperature of regeneration with the residual water content of MS 5 A as parameter.

Water vapor also participates as an adsorptive in the adsorption of organic substances on activated carbon. Because at low saturation the energies of interaction of organic substances with activated carbon are much higher than those of water vapor with activated carbon, water vapor does not reduce the capacity for organic substances. This factor is the basis for the ability of active carbon to adsorb organic substances. With increasing relative saturation of water vapor, capillary condensation can finally occur. The capacity of activated carbon for organic components is then substantially reduced, as described in Section Removal of Organic Components from Exhaust Air.

The closer the adsorption isotherms of two materials are at the same temperature, the more competitive adsorption of both components must be considered.

A useful graphic representation of binary adsorption equilibria employs the methanol – water – silica gel system as an example (see [86]). Figure 38 shows the loading n_M of methanol as a function of the partial pressure p_M of methanol for various partial pressures p_W of water at 50 °C. At a given value of p_M, the loading of methanol decreases as the partial pressures p_W of water increases. Figure 39 shows, also at 50 °C, how increasing partial pressures of methanol reduce the loading of water at a given p_W. Such diagrams at first say nothing about the relationship between the concentration of adsorptive in the gas phase and in the sorbate. Therefore, expressing the loading of adsorbent q_{tot} as a function of the mole fraction y_i can be useful. In Figure 40, the loadings of methanol and water are plotted versus the mole fraction y_M of methanol in the gas phase for various active pressures $p_M + p_W$ at 50 °C. The value y_M=1 represents pure methanol vapor, and y_M=0, pure water vapor. For these limiting cases, at 50 °C the methanol loading (y_M=1) is higher than the water loading (y_W=0).

Methanol loading increases with increasing mole fraction y_M of methanol, while water loading decreases. Above 50°C, all curves $q_i=f\,[y_i,(p_i + p_j)]$ would be displaced downward; at lower temperature, displacement would be upward.

Such diagrams permit determination of the partial loading q_i of individual components, as well as the total loading q_{tot}; however, they say nothing about the composition of the sorbate. The mole fraction x_i of component i in the adsorbate can now be determined from the partial loading q_i according to $x_i=q_i/q_{tot}$, where $q_{tot}= \sum q_j$.

From this, equilibrium diagrams $y_i=f(x_i)$ can be calculated. In Figure 41, the mole fraction x_M of methanol in the sorbate is presented as a function of the mole fraction y_M of methanol in the gas phase for different active pressures $p_M + p_W$, again at 50 °C. Such equilibrium curves may exhibit different shapes at other temperatures and at other active pressures.

Mixture equilibria with n components usually are represented as $q_j=q_j(p_1,p_2,p_3, \ldots T)$. In the case of binary mixtures the representations preferably are $x – y$ and q_{tot} diagrams, e.g., $x_1(y_1)$ and $q_{tot}(y_1)$, where $x_1=q_1/q_{tot}$, $y_1=p_1/p$, and $q_{tot}=q_1 + q_2$ is the total loading.

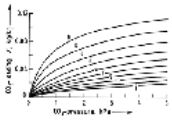

Figure 36. Working diagram for determining reduced carbon dioxide capacity as a function of residual water loading on MS 5A at 25°C

a) X_{H2O}=0 g/g; b) X_{H2O}=0.01 g/g; c) X_{H2O}=0.02 g/g; d) X_{H2O}=0.03 g/g; e) X_{H2O}=0.04 g/g; f) X_{H2O}=0.05 g/g; g) X_{H2O}=0.06 g/g; h) X_{H2O}=0.07 g/g; i) X_{H2O}=0.08 g/g; j) X_{H2O}=0.09 g/g

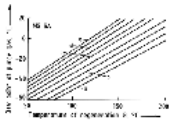

Figure 37. Working diagram for determining residual water loading after thermal regeneration with humid ambient air

Residual water loading: a) 0.01 g/g; b) 0.03 g/g;c) 0.05 g/g; d) 0.07 g/g; e) 0.09 g/g

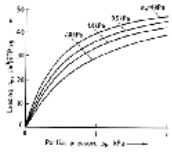

Figure 38. Methanol loading as a function of the partial pressure of methanol (p_M) for various partial pressures of water (p_W)

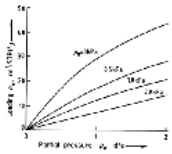

Figure 39. Water loading as a function of the partial pressure of water (p_W) for various partial pressures of methanol (p_M)

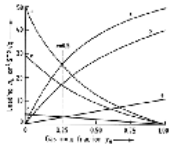

Figure 40. Loading as a function of mole fraction of methanol (y_M) in the gas phase for various active pressures at 50 °C

a) Loading of methanol at p_t=0.1 kPa; b) Loading of methanol at p_t =1.0 kPa; c) Loading of methanol at p_t=3.0 kPa; d) Loading of water at p_t=0.1 kPa; e) Loading of water at p_t=1.0 kPa; f) Loading of water at p_t =3.0 kPa

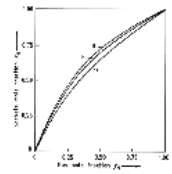

Figure 41. Mole fraction x_M of methanol in the sorbate as a function of mole fraction y_M in the gas for various pressures at 50 °C (system methanol–water–silica gel)

a) p_W+p_M=0.1 kPa; b) p_W+p_M=1 kPa;c) p_W+p_M=3 kPa

4.3.1. *Homogeneous Surfaces*

The Langmuir equation has an extension for mixtures which is exact under the hypothesis of a perfect homogeneous surface exhibiting an identical saturation loading for all components [87]. Then the fractional coverage for component i in a mixture of n components is

$$\Theta_i = \frac{q_i}{q_0} = \frac{p_i/\phi_i}{1+\sum_{j=1}^{n} p_j/\phi_j}, \tag{4.27}$$

and the total loading is

$$q = q_0 \sum_{j=1}^{n} \Theta_j. \tag{4.28}$$

For an approximation, the saturation loading may be inserted as component- and temperature-dependent, e.g., q_0 is replaced by $q_{0i}(T)$.

4.3.2. *Mixture BET Equation*

The BET equation (Eq. 4.11) can be extended to mixtures by using $\varphi_i=p_i/\,\Phi_{Li}$, $b_i= \Phi_i/\,\Phi_{Li}$,

$$\phi_{Li} = p_0 e^{A_i - H_{ti}/(RT)}, \psi = \Sigma\varphi_j, B = \left(\Sigma b_j \varphi_j\right)/\psi, S_i = b_i/B \tag{4.29}$$

$$\Theta_i = \frac{B\varphi_i}{1-\psi}\frac{1+(S_i-1)(1-\psi)+\psi^N[N+S_i]}{1+(B-1)\psi-B\psi^{N+1}}$$

where $\Theta_i = q_i/q_{0i,mon}$ is related to the monolayer saturation loading $q_{0i,mon}$. For N=1 the equation results in the extended Langmuir equation (4.27).

4.3.3. *Ideal Adsorbed Solution Theory*

The basic n-component equations of the real and ideal adsorbed solution (RAST/IAST) with $y_i = p_i/p$ and $x_i = q_i/\Sigma q_j$ [10] are

$$q_i P = \gamma_i x_i P_j(\pi, T) \tag{4.30}$$

$$q_j = x_j/\Sigma[x_j/Q_j(\pi, T)] \tag{4.31}$$

Summation (j=1 . . .n) and the boundary conditions Σy_i=1 and Σy_i=1 are used to solve Equations (4.31) iteratively. To fit real conditions, activity coefficients $\gamma_i \neq 1$ have been introduced (RAST). For the ideal solution (IAST) approach they are unity (γ_i=1). The key issue of Equations (4.30) and (4.31) are $P_i(\pi,T)$ and $Q_i(\pi,T)$, which are the pure gas equilibrium pressure and loading Q related to the same spreading pressure, i.e., surface energy π for all components.

4.3.3.1. *Spreading Pressure*

The spreading pressure concept uses the Gibbs – Duhem relation for adsorbates

$$S\mathrm{d}T - a\mathrm{d}\pi + N\mathrm{d}\mu = 0 \tag{4.32}$$

which simplifies to the Gibbs isotherm at constant temperature

$$\mathrm{d}\pi = (N/a)\mathrm{d}\mu \tag{4.33}$$

where N is the amount adsorbed, a the surface area, and S the entropy. Equating N/a with $\Theta q_0/\sigma$, where $q = \Theta q_0$ and σ is specific surface area, and the ideal gas chemical potential is equal to the adsorption potential at equilibrium

$$\mu = \mu_0 - RT\ln(p/p_0) \tag{4.34}$$

the spreading pressure is derived as

$$\pi(P, T) = (RTq_0/\sigma)\int_0^P \mathrm{d}p(\Theta(p)/p) \tag{4.35}$$

by using the pure gas fractional coverage $\Theta(p)$. For practical applications it is convenient to work with the reduced spreading pressure π', summarizing all component independent parameters:

$$\pi'(P, T) = \pi\sigma/(RT) = q_0\int_0^P \mathrm{d}p(\Theta(p)/p) \tag{4.36}$$

Examples are displayed in Figure 42. Evaluating the integral of Equation (4.36) for three isotherm relations, Langmuir (L, Eq. 4.1), UniLan (UL, Eq. 4.8), and (GM, Eq. 4.9) gives analytical expressions only for L and GM, while the UniLan equation has to be solved analytically. The analytical solutions are

$$\pi'(P_i, T) = \left\{\begin{matrix} q_{0i}\ln(1+p_i/\phi_i) \\ \left\{q_{0i}/a_i(T)\right\}\ln\left(1+[p_i/\phi_i]^{a_i(T)}\right)\end{matrix}\right\}\begin{matrix}(\mathrm{L})\\(\mathrm{GM})\end{matrix} \tag{4.37}$$

where the exponent $a_i(T)$ is defined in Equation (4.10). The inverse function $P_i(\pi,T)$ to be inserted into the first IAST equation (4.31) is

$$P_i(\pi', T) = \left\{\begin{matrix} \phi_i\left(e^{\pi'/q_{0i}}-1\right) \\ \phi_i\left(e^{\pi' a_i(T)/q_{0i}}-1\right)^{1/a_i(T)}\end{matrix}\right\}\begin{matrix}(\mathrm{L})\\(\mathrm{GM}).\end{matrix} \tag{4.38}$$

Using the isotherm relations we derive $Q_i(\pi,T)$ for insertion into the second IAST equation (4.31) as

$$Q_i(\pi, T) = \left\{\begin{matrix} q_{0i}\left(1-e^{-\pi'/q_{0i}}\right) \\ q_{0i}\left(1-e^{-\pi' a_i(T)/q_{0i}}\right)\end{matrix}\right\}\begin{matrix}(\mathrm{L})\\(\mathrm{GM}).\end{matrix} \tag{4.39}$$

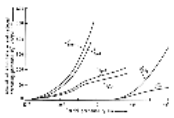

Figure 42. Isotherms and reduced spreading pressure curves ($\pi'_i = \pi_i\, a/RT$) for carbon dioxide and water (pure components on MS 5A at 25 °C) (isotherms according to [132])

—— Both α- and β-cages; – – – Only α-cages.

4.3.3.2. *Examples for Binary Mixture Calculations*

4.3.3.2.1. Given Gas-Phase Concentration

Find the componentwise adsorptive loading of a 30/70 % gas mixture of CH_4/N_2 on an activated carbon at T=25 °C and p=20 bar with the IAS model and alternatively with the extended Langmuir equation of Markham and Benton

Pure-Gas Parameters. Isotherm parameters of CH_4 as first component:

$$q_{01} = 1.55\ \mathrm{mol/kg} \quad H_{a1} = 18.3\ \mathrm{kJ/mol} \quad \phi_{01} = 2.02\ \mathrm{bar}$$

related values:

$$H_{a1}/R = 2201\ K \quad \phi_1\left(25^\circ C\right) = \phi_{01}e^{2201(1/273-1/298)} = 4.00\ \mathrm{bar}$$

$$y_1 = 0.3 \quad \varphi_1 = y_1 p/\phi_1 = 1.50$$

Isotherm parameters of N_2 as second component:

$$q_{02} = 4.57\ \mathrm{mol/kg} \quad H_{a2} = 13.1\ \mathrm{kJ/mol} \quad \phi_{02} = 7.37\ \mathrm{bar}$$

related values:

$$H_{a2}/R = 1585\ K \quad \phi_2\left(25^\circ C\right) = \phi_{02}e^{1585(1/273-1/298)} = 12.00\ \mathrm{bar}$$

$y_2 = 0.7 \quad \varphi_2 = y_2 p / \Phi_2 = 1.17$

IAS Calculation. The IAS calculation with $\gamma_1 = \gamma_2 = 1$ and the reduced spreading pressure π' for constant T and σ yields from Equation 4.30 by inserting Equation (4.39) for $P_i(\pi, T)$ from the Langmuir case

$$x_i = y_i p / P_i(\pi, T) = (y_i p / \phi_i) / (e^{\pi'/q_{0i}} - 1) = \varphi_i / (e^{\pi'/q_{0i}} - 1) \tag{4.40}$$

leading to the equation

$$x_1 + x_2 = 1 = 1.50 / (e^{\pi'/5.55} - 1) + 1.17 / (e^{\pi'/4.57} - 1)$$

which is iteratively solved by variation of π (the same π' for both components). The result is

$$\pi' = 6.67 \text{ mol / kg}$$

The spreading pressure is $\pi RT\, \pi' / \sigma = 17$ J/m^2 for a typical specific surface area σ of 10^6 m^2/kg.

Summarizing

$$x_1 = 1.50 / (e^{6.67/5.55} - 1) = 0.645 \quad x_2 = 1.17 / (e^{6.67/4.57} - 1) = 0.355$$

The total mixture loading follows by using $Q_i(\pi, T)$ from Equation 4.39 (L)

$$Q_i(\pi', T) = q_{0i}[1 - e^{-\pi'/q_{01}}]$$

and Equation (4.31) as

$$q_{tot} = 1 / [x_1 / Q_1(\pi', T) + x_2 / Q_2(\pi', T)] \tag{4.41}$$

and evaluated as

$$q_{tot} = 1 / \{0.645 / [5.55(1 - e^{-6.67/5.55})] + 0.355 / [4.57(1 - e^{-6.67/4.57})]\} = 3.74 \text{ mol / kg}$$

where $q_1 = x_1 q_{tot} = 2.41$ mol/kg and $q_2 = x_2 q_{tot} = 1.33$ mol/kg.

The selectivity is $S_{12} = (x_1/y_1)/(x_2/y_2) = (0.645/0.3)/(0.355/0.7) = 4.24$.

Markham – Benton Calculation. With Equation 4.27 and the related φ_i from the above pure-gas calculation

$$q_i = q_{0i} \varphi_i / (1 + \Sigma \varphi_j) \tag{4.42}$$

and inserting

$$q_1 = 5.55[1.50/(1 + 1.50 + 1.17)] = 2.27 \text{ mol / kg}$$

$$q_2 = 4.57[1.17/(1 + 1.50 + 1.17)] = 1.46 \text{ mol / kg}.$$

$$q_{tot} = 3.73 \text{ mol / kg} \quad x_1 = 0.609 \quad x_2 = 0.391 \quad S_{12} = 3.63$$

experimentally q_{tot} was 3.75 mol/kg, x_1 was 0.6, and hence S_{12}=3.5.

4.3.3.2.2. Given Loading

Find the componentwise partial pressures for the example of Section Given Gas-Phase Concentration but with total loading q_{tot} of 3.0 mol/kg and a CH_4 loading of q_1=1.2 mol/kg for the adsorbed CH_4/N_2 mixture on activated carbon at T=25 °C.

IAS Calculation. The calculation steps are now of inverse order compared to a given gas-phase content. Taking the isotherm parameters from the pure-gas calculation of Section Given Gas-Phase Concentration and using the related values

$$x_1 = 0.4 \quad \Theta_1 = x_1 q_{tot} / q_{01} = 0.216$$
$$x_2 = 0.6 \quad \Theta_2 = x_2 q_{tot} / q_{02} = 0.394$$

in the IAS calculation with $\gamma_1 = \gamma_2 = 1$ starting with the second IAS equation we obtain with Equation (4.41)

$$q_{tot} = 1 / \{x_1 / Q_1(\pi', T) + x_2 / Q_2(\pi', T)\}.$$

inserting $x_i = q_i / q_{tot}$ and $\Theta_i = q_i / q_{0i}$ we derive the equation

$$\Theta_1 / (1 - e^{-\pi'/q_{01}}) + \Theta_2 / (1 - e^{-\pi'/q_{02}}) = 1$$

or with Θ_1 = 1.2/5.55 = 0.216 and Θ_2 = 1.8/4.57 = 0.394

$$0.216 / (1 - e^{-\pi'/5.55}) + 0.394 / (1 - e^{-\pi'/4.57}) = 1$$

which is iteratively solved by variation of π'. The result is

$$\pi' = 4.63 \text{ mol / kg}$$

Inserting into Equation (4.30) using Equation (4.39)

$$p_1 = y_1 P = x_1 P_1(\pi', T) = x_1 \phi_1 (e^{\pi'/q_{02}} - 1) = 0.4 \cdot 4.0 (e^{4.63/5.55} - 1) = 2.1 \text{ bar}$$

$$p_2 = y_2 P = x_2 P_2(\pi', T) = x_2 \phi_2 (e^{\pi'/q_{02}} - 1) = 0.6 \cdot 12.0 (e^{4.63/5.57} - 1) = 12.6 \text{ bar}$$

Therefrom y_1=0.143 and p=14.7 bar. The selectivity is

$$S_{12} = (x_1/y_1)/(x_2/y_2) = (0.4/0.143)/(0.6/0.857) = 4.00$$

Markham – Benton Calculation. Equation (4.27) can be rearranged to

$$p_i = \phi_i \Theta_i / (1 - \Sigma \Theta_i) \tag{4.43}$$

where $\Theta_i = q_i / q_{0i}$. Inserting Θ_1=0.216, and Θ_1=0.394 from the above IAS calculation

$$p_1 = 4.0[0.216/(1 - 0.216 - 0.394)] = 2.2 \text{ bar}$$

$$p_2 = 12.0[0.394/(1 - 0.216 - 0.394)] = 12.1 \text{ bar}$$

and the total pressure is 14.3 bar, y_1=0.154 and the selectivity turns out as S_{12}=3.65. The small deviations of the different approaches from each other and the experimental data mainly reflect the (poor) pure-gas approximation as ideal equilibria neglecting all contributions from the heterogeneous surface.

4.3.4. *Statistical Thermodynamics Model*

The statistical thermodynamics (ST) model [88] is suitable for adsorbents with a cage structure (zeolitic molecular sieves). The configuration integrals are approximated by simple expressions that contain the Henry coefficient and the molar volume. The model has been extended to binary adsorption equilibria [89]. The most important assumption made in this model is that the adsorbed molecules are enclosed cages, with practically no exchange of molecules between individual cages. Further, the molecules are assumed to be free to move within the cage; i.e., attractive forces can be neglected. Repulsion of the molecules is taken into account by reducing the cage volume. Because of these assumptions, the model is suitable only for the adsorption of nonpolar molecules of approximately equal size in adsorbents with a cage structure.

Therefore, a simplified statistical thermodynamics model was proposed in 1985, which takes into account both the interactions between molecules and their localization on the adsorbent. The single-component adsorption isotherm can then be described by Equation (4.44) [88, 89].

$$n_i^k = \frac{h_i p_i + (h_i p_i)^2 \; R_{1i} + (h_i p_i)^3 \; \frac{R_{2i}}{2!} + \; \ldots \; + (h_i p_i)^{m_i} \; \frac{R_{mi}}{(m_i - 1)}}{1 + h_i p_i + (h_i p_i)^2 \; \frac{R_{1i}}{2!} + (h_i p_i)^3 \; \frac{R_{2i}}{3!} + \; \ldots \; + (h_i p_i)^{m_i} \; \frac{R_{mi}}{(m_i - 1)!}} \tag{4.44}$$

or

$$n_i^k = \frac{h_1 p_1 + \sum_j \sum_i \frac{(h_1 p_1)^i (h_2 p_2)^j}{i!j!} \left(R_{1s}^i R_{2s}^j \right)^{\frac{1}{i+j}}}{1 + h_1 p_1 + \sum_j \sum_i \frac{(h_1 p_1)^i (h_2 p_2)^j}{i!j!} \left(R_{1s}^i R_{2s}^j \right)^{\frac{1}{i+j}}} \tag{4.45}$$

where $i + j = s$ and $iV_{M1} + jV_{M2} \leq V_K$

The constants R_{1i}, R_{2i}, . . .R_{mi} describe the configuration integrals of statistical thermodynamics and must be determined from single-component adsorption isotherms. Because of this, the number and quality of these constants are of paramount importance. Only two or three constants should be determined, and only after the Henry coefficient h_i has been fitted carefully to reliable measured data, especially at low partial pressure p_i. The two-component adsorption isotherm for component 1 (index 1), for example, then follows from Equation (4.44).

A cage thus contains *s* molecules, i.e., *i* molecules of component 1 and *j* molecules of component 2. The volume occupied by the molecules ($iV_{M1} + jV_{M2}$) must, of course, be smaller than the cage volume V_K. The number n_2^k of molecules of component 2 in the cage is given by an analogous relation.

This simplified statistical thermodynamic model is capable of calculating two-component adsorption isotherms for the following systems from single-component isotherms: *n*-heptane – cyclohexane and ethylene – ethane on molecular sieve 13X, as well as ethylene – propane, carbon dioxide – ethylene, and cyclopropane – ethylene on molecular sieve 5A. This model is probably useful only if the deviation of the adsorbate components from ideal behavior is small. This is often nearly the case when the total loading is less than one molecule per cage. For nonideal behavior and especially for azeotropic behavior (i.e., the equilibrium curve $x=f(y)$ intersects the diagonal line, at this point $x=y$), different molecular volumes and varying interactions of dissimilar molecules must be accounted for by additional coefficients which can be determined only from the equilibrium of the mixture, not from single-component equilibria alone.

KRISHNA showed that configurational entropy can be matched with CBMC (configurational-bias Monte Carlo). In a 50:50 mixture of *n*-hexane (nH) and 3-methyl pentane (3MP) the latter has a maximum loading at 100 Pa (362 K) which decreases at higher pressures virtually to zero. The nH molecules fit nicely in both straight and zig-zag channels, whereas the 3MP molecules are preferentially located at the intersections (cages) between the straight and zig-zag channels in the zeolite. At higher pressure 3MP is then replaced by nH. This corresponds to a sharp increase of sorption selectivity beyond a total loading of four molelcules per unit cell, corresponding to a situation in which all intersections are occupied [90]. Kinetic Monte Carlo and related techniques [33] can also be used to describe diffusion in ordered structures such as zeolites and carbon nanotubes for separation of, e.g., ethane isomers [91].

[Top of Page]

5. Kinetics

The term kinetics generally refers to the rate at which thermodynamic equilibrium is established in a system that is not in equilibrium. In the case of adsorption, the fluid concentration c_i or the partial pressure p_i is greater than the value that is in equilibrium with the loading of the adsorbent. Adsorption continues until the equilibrium loading associated with the fluid concentration is reached.

The reverse process occurs during desorption. As a result of a temperature increase for the adsorbent (regeneration by temperature swing) or a reduction in (partial) pressure (regeneration by pressure swing), the fluid concentration c_i or the partial pressure p_i is smaller than the value in equilibrium with the loading of the adsorbent. Again, desorption continues until equilibrium is established.

The transport of heat and mass in an industrial adsorber takes place in the interstitial bulk phase between the solid particles and is coupled to the balances inside each pellet. When considering mass transfer the adsorptive must overcome the following resistances:

1. Transport from the bulk phase through the boundary layer around the pellet
2. Transport in the fluid and adsorbed phases within the pore system
3. Adsorption

The desorption process occurs in the reverse manner; i.e., it begins with desorption and ends with transport from the concentration boundary layer around the pellet into the fluid volume. In exothermic adsorption, heat is released which is transferred to the adsorbent and the fluid. The adsorbent heats up. In endothermic desorption, heat is required, and the adsorbent cools. Therefore, additional equations for the conservation of energy must be formulated for individual balance volumes. Because of the enormous computational expense involved in solving these equations of conservation, the kinetics are often described only for the isothermal case. The results thus obtained are valid only if the concentration of adsorptive and the heat of adsorption are low or if the system is not very sensitive to temperature.

A similarity exists between fixed-bed adsorbers and heterogenous catalytic reactors. If the reaction rate is faster than the transport rates, then the models are identical; otherwise a chemical reaction adds terms. Thus, much information can be drawn from studies of reactors [92-97]. The same is true for drying, where similar information can be retrieved from the literature, especially in regard to mass transport onto solids [98, 99].

5.1. External/Internal Transport

The transfer of heat and mass between a bulk phase (c_B) and a solid interphase (c_s) takes through a laminar film according to the film theory [100]. For mass transfer this is then

$$n = \beta(c_S - c_B) \tag{5.1}$$

and heat transfer

$$q = \alpha(T_S - T_B) \tag{5.2}$$

However, the dominant mass transfer resistances are usually inside the pellet, whereas the dominant heat transfer resistance is mostly in the film. This is true for liquid- and gas-phase systems, whereby heat effects in liquid adsorption are often negligible.

5.1.1. *Internal Resistances*

Viscous Flow. Within the pellet the transport mechanism of the solute is either in the gas or liquid adsorbed phase. If the pore diameter is considerably (e.g., ten times) larger than the mean free diffusion path, then there is viscous flow. This occurs especially when there is a pressure gradient as in pressure swing adsorption. The mass flow is then determined by the Hagen – Poisseuille law, and the laminar flow in a pore is described by:

$$\dot{v} = -\frac{d_M^2}{32\eta}\frac{\partial P}{\partial z} \tag{5.3}$$

where P is the total pressure, d_M the diameter of the macropore, and η the dynamic viscosity. In a porous system the tortuosity of the pores must be considered as

$$z = \mu_{\mathrm{p,lam}} \cdot r \tag{5.4}$$

The mass flux density for a compound is proportional to the local gas density:

$$\dot{m}_i = \dot{v} \cdot \rho_i \tag{5.5}$$

or

$$\dot{m}_1 = \frac{d_M^2}{32\mu_{\mathrm{p,lam}}\eta}\frac{\partial P}{\partial r}\rho_i \tag{5.6}$$

The influence of the pore size geometry is expressed in the factor $d_M^2/\mu_{\mathrm{p,lam}}$. The smaller the pores, the higher the transport resistance, and larger pores will contribute to a laminar mass flow. An empirical estimate for the tortuosity is given by [99].

$$\mu_{\mathrm{p,lam}} \approx \mu_D^{2.6} \tag{5.7}$$

Diffusional Flow. Besides viscous flow the dominant transport mechanism in the gas phase in smaller pores is free diffusion or Knudsen diffusion. A criterion to distinguish between the two mechanisms is the Knudsen number *Kn*:

$$Kn = \frac{\Lambda}{d} \tag{5.8}$$

where the mean free diffusion path Λ can be derived from the kinetic gas theory:

$$\Lambda = \frac{kT}{\sqrt{2}\pi\sigma^2 P} \tag{5.9}$$

where k is the Boltzmann constant, T the temperature, and σ is the collision diameter. It can be calculated from transport properties or the second virial coefficient and is available from various sources [101].

Knudsen diffusion prevails if $Kn > 10$ (or $d < 0.1\ \Lambda$) where

$$\dot{m} = -\frac{4}{3}\frac{d}{\mu_{\mathrm{p},Kn}}\sqrt{\frac{M_i}{2\pi RT}}\frac{\partial p_i}{\partial r} \tag{5.10}$$

The diffusivity is then given by:

$$D_{Kn} = \frac{4}{3}\frac{d}{\mu_{\mathrm{p,Kn}}}\sqrt{\frac{M_i}{2\pi RT}} \tag{5.11}$$

The empirical tortuosity relation after [99] is:

$$\mu_{\mathrm{p},Kn} = \mu_{\mathrm{p}}^{1.7} \tag{5.12}$$

If $Kn < 0.1$ (or $d > 10\ \Lambda$) then *free diffusion* prevails, as described by Fick's law

$$\dot{m} = -\frac{D_{ij}}{\mu_{\mathrm{D}}}\frac{M_i}{RT}\frac{\partial p_i}{\partial r} \tag{5.13}$$

where D_{ij} is the binary diffusion coefficient, and the tortuosity μ_D must be estimated experimentally or obtained elsewhere [22]. In the slip flow region ($0.1 < Kn < 10$) neither mechanism can be neglected.

Surface Diffusion. As there is a concentration gradient in the pores in the gas phase there is a similar one in the adsorbate on the pore wall. This driving force is referred to as surface diffusion and is low in a monolayer and may be considerable at high coverage. The transport mechanism is considered as a hopping of molecules between adsorption sites [102], which could be also simulated by molecular dynamics simulations [103]. The transport equation is similar to that in the gas phase

$$\dot{m} = -\rho_\theta D_{\mathrm{S}}\frac{\partial X}{\partial r} \tag{5.14}$$

where X is the local sorptive loading, ρ_θ the sorbate density, and D_S the surface diffusivity.

In principle there are two limiting cases [22]:

1. A monomolecular adsorption layer will occur as $D_S \longrightarrow 0$, since the adsorptive has a strong interaction with the active sites.
2. When the capillary is completely filled with condensate then $D_S \longrightarrow \infty$, which means there is a capillary flow without a gradient. The differences in the local capillary pressure are then the reason for liquid flow in an unsaturated porous medium [104-106]. Clearly, in liquid adsorption this phenomenon is more likely than in gas adsorption. However, activated carbon and silica gels will show surface diffusion if the relative humidity is higher than 40 %.

Intracrystalline Diffusion. Intracrystalline diffusion in the micropores, also known as activated slit diffusion [107], can play a limiting role in adsorption and desorption [22, 63]. Here the sorptive diameter is on the order of magnitude of the diameter of the pores, so it is difficult to distinguish from surface diffusion. The transport equation is then analogous to Equation (5.14)

$$\dot{m} = -D_0^0\rho_\theta\frac{\partial X}{\partial r} \tag{5.15}$$

where D_0^0 is the micropore diffusion coefficient. In liquid adsorption, intercrystalline diffusion is lumped together with surface diffusion and not considered separately.

5.1.2. *Linear Driving Force and Other Models*

In chemical catalysis the dusty-gas model is used as a standard to describe diffusion in a porous media. The pore walls or the solid matrix are considered as giant gas molecules ("dust") in an ideal gas mixture. In a formal way the mole fraction of dust can be adapted to account for all diffusional effects from free to Kundsen diffusion [33, 108-110]. In structured geometries like those of zeolites it is nowadays possible to describe surface diffusion and diffusion in narrow pores with statistical mechanic methods such as molecular dynamics or Monte-Carlo simulations [33]. However, in industrial practice, coupling of several mass-transfer resistances with the fixed-bed absorber balance results in a mathematically tedious enterprise. Additionally, in industrial practice a detailed knowledge of the rate-determining steps is not always available, so the linear driving force model (LDF) is often applied. This is why it is generally used in liquid adsorption. All the intraparticle resistances are neglected, and an effective mass transfer resistance in the laminar film is rate-determining

$$\frac{\partial q}{\partial t} = k_{\mathrm{eff}}\left(q_B - q_s\right) \tag{5.16}$$

Here q_S is the average overall amount in the particle and q_B is the equilibrium loading with respect to the bulk concentration c_B. There is an approximation for k_{eff} if the Fourier number $D_{\mathrm{eff}}t/r_{\mathrm{p}}^2 < 0.1$ [111]:

$$k_{\mathrm{eff}} = \frac{15D_{\mathrm{eff}}}{r_{\mathrm{p}}^2} \tag{5.17}$$

However, the validity should be checked prior to use. The diffusion time constant $D_{\mathrm{eff}}t/r_{\mathrm{p}}^2$ is in the order of 10^{-1} s^{-1} for activated carbon and 10^{-3} s^{-1} for zeolite crystals. This makes the application of LDF in the initial periods of uptake invalid, especially for zeolite crystals [63]. Results obtained with the Glueckauf formulation are often in good agreement with the more accurate macropore diffusion model while requiring less calculation time [112].

An alternative approach used with the desorption of water [113] considered the desorptive transfer process as a first-order kinetics rate [82]:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = k_{\mathrm{des}} \cdot q_{\mathrm{B}} \tag{5.18}$$

With the temperature dependency using the isosteric activation energy E_a

$$k_{\mathrm{des}} = k^0 e^{-\frac{E_a}{RT}} \tag{5.19}$$

and a frequency factor [114]

$$\ln k^0 = a + b e^{-cq_{\mathrm{B}}^2} \tag{5.20}$$

where *a*, *b*, *c* are adjustable parameters.

The kinetics of nonisothermal ad- and desorption are more complex and discussed in detail in [22], and multicomponent adsorption has so far not been well investigated. According to [83] one can assume that the molecule – molecule interactions are ten times smaller than molecule – wall interactions. However, multidiffusional effects play a vital role in pressure-swing adsorption. In this regard, diffusion and friction in multicomponent mixtures in porous systems using the Maxwell – Stefan diffusion law has been discussed by several authors [115, 116].

[Top of Page]

6. Adsorber Dynamics

Mass and heat transport in an adsorber bed result in a breakthrough curve which can be simulated appropriately. Depending on the scope of the results the models are then more or less sophisticated.

The simplest approach is an *equilibrium model* considering a convective gas phase in equilibrium with the stationary phase. This is a good approach for a rough estimation of the bed behavior. To account for backmixing, i.e., the deviation from plug flow, a tank-in-series model can be used. This is equivalent to the dispersion or back-flow models, in which one parameter accounts for all flow nonidealities [94, 117].

Nonequilibrium models additionally consider the mass (heat) transfer from the bulk gas phase to the adsorbent. With more sophisticated models a balance inside the particle is made with consideration of the kinetics behavior (see Chap. Kinetics). However, both gas and liquid adsorption models are usually simplified, and the LDF model is considered appropriate and efficient to account for the major effects.

6.1. Modeling the Isothermal Fixed-Bed Adsorber

Figure 43 shows a fixed-bed adsorber of cross-sectional area *A* with a volume flow $\dot{V}$ passing through it. The volume flow density $\dot{v}$ is then $\dot{v} = \dot{V}/A$. The adsorber contains monodisperse particles of adsorbent with radius R_a. Let the porosity of the packing be Ψ, which represents the ratio of void volume to packing volume. If the adsorbent exhibits a bimodally dispersed pore structure with macropores and cage-type micropores, as is the case with zeolitic molecular sieves, the micropore system is then located within the zeolite crystals, which are assumed to be monodisperse and have the radius r_a.

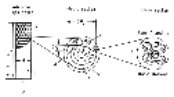

Figure 43. Balance volumes in molecular sieve 5A fixed bed

a) Pellet; b) Adsorber; c) Zeolite crystals; d) Macropore; e) Outer boundary layer; f) Micropore

Material balances in the fixed-bed adsorber which correspond to these assumptions can then be formulated for three balance volumes:

1. Balance in the adsorber with differential bed height d*z* (fixed-bed absorber):

$$\begin{gathered}\frac{\partial c_i}{\partial t} + \frac{1-\psi}{\psi}\rho_{\mathrm{p}}\frac{\partial \overline{X}_{m,i}}{\partial t} = \frac{\dot{v}}{\psi}\frac{\partial c_i}{\partial z} + D_{\mathrm{ax}}\frac{\partial^2 c_i}{\partial z^2} \\ c_i(z, t=0) = 0; \quad c_i(z, t=0) = c_{0,i}; \quad \frac{\partial c_i}{\partial z}(z = Z, t) = 0; \\ \underbrace{c_i(z=0, t) = c_{0,i,\mathrm{ad}} + \frac{D_{\mathrm{ax}}\psi}{\dot{v}}\frac{\partial c_i}{\partial z};}_{\text{Adsorption}} \quad \underbrace{c_i(z=0, t) = c_{0,i,\mathrm{des}} + \frac{D_{\mathrm{ax}}\psi}{\dot{v}}\frac{\partial c_i}{\partial z}}_{\text{Desorption}}\end{gathered} \tag{6.1}$$

2. Balance in the macropores of the particle (radius R_a):

$$\frac{\partial c_{M,i}}{\partial t} + \frac{\rho_{\mathrm{p}}}{\varepsilon_{\mathrm{p}}}\frac{\partial \overline{X}_{m,i}}{\partial t} = \underbrace{\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{D_{M,i}}{\mu_{\mathrm{p}}}\frac{\partial c_i}{\partial R}\right)}_{\text{Fluid phase}} + \underbrace{\frac{1}{R^2}\frac{\partial}{\partial R}\left(R^2\frac{\rho_{\mathrm{p}}}{\varepsilon_{\mathrm{p}}}D_{s,i}\frac{\partial \overline{X}_{m,i}}{\partial R}\right)}_{\text{Sorbate phase}} \tag{6.2}$$

Transport only in fluid phase:

$$\begin{gathered}\underbrace{c_{M,i}(R, t=0) = 0;}_{\text{Adsorption}} \quad \underbrace{c_{M,i}(R, t=0) = c_{0,i};}_{\text{Desorption}} \quad \frac{\partial c_{M,i}}{\partial R}(R=0, t) = 0 \\ \beta_i(c_i - c_{M,i})\Big|_{R=R_a} = \frac{\varepsilon_{\mathrm{p}}}{\mu_{\mathrm{p}}}D_{M,i}\frac{\partial c_{M,i}}{\partial R}\Big|_{R=R_a}\end{gathered} \tag{6.3}$$

3. Balance in the zeolite cages of the crystal (radius r_a):

$$\begin{gathered}\frac{\partial X_{m,i}}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 D_{m,i}\frac{\partial X_{m,i}}{\partial r}\right) \\ \underbrace{X_{m,i}(r, t=0) = 0;}_{\text{Adsorption}} \quad \underbrace{X_{m,i}(r, t=0) = X_i^*(c_{0,i});}_{\text{Desorption}} \quad \frac{\partial X_{m,i}}{\partial r}(r=0, t) = 0; \\ X_{m,i}(r = r_a, t)\,|_R = X_i^*(c_{m,i}, t)\,|_R\end{gathered} \tag{6.4}$$

In principle, the adsorption and desorption steps can also affect the rate of sorption. This can be described analogously to first-order reactions:

$$\text{Adsorption:} \quad \frac{\mathrm{d}c}{\mathrm{d}t} = k_{\mathrm{ad}}c\left(n^\infty - n\right) \tag{6.5}$$

$$\text{Desorption:} \quad \frac{\mathrm{d}n}{\mathrm{d}t} = k_{\mathrm{des}}n = k_{\mathrm{des}}^0 \exp\left(-\frac{\Delta H}{RT}\right)n \tag{6.6}$$

Numerous investigations have shown that in the physical sorption of small molecules, the time required for the adsorption step can be neglected. The same almost always holds true for the desorption step, too. This means that fluid and adsorbate phase are in mutual equilibrium.

The initial and boundary conditions for simple cases are shown below the balance equations. The cases include a bed completely unloaded at the beginning of adsorption, as well as a bed completely loaded uniformly at the beginning of desorption by purging, in which the loading $X_{0,i}^*$ is in equilibrium with the concentration of adsorptive $c_{0,i}$.

Balances must be formulated for each component in multicomponent adsorption and can be applied to both adsorption and desorption. When steam regeneration is applied then additional drying and cooling slips have to be considered, too [118]. The equations contain the changes with time $\partial c/\partial t$ in the concentration of adsorptive and $\partial X/\partial t$ in the loading of adsorbent, as well as various coefficients which are explained next.

6.2. Modeling the Adiabatic Fixed-Bed Adsorber

In principle, fixed beds for adsorption and desorption can be operated *isothermally* (low concentration of adsorptive or low heat of adsorption), *nonisothermally* (removal or supply of heat), or *adiabatically*. In larger diameter industrial adsorbers without internal cooling, the surface-to-volume ratio is so small that heat losses through the outside jacket are small. In such cases, the apparatus is operated approximately adiabatically. Formulas for the *energy balance* of the adsorbent and of the gas for a differential bed height d*z* during adiabatic operation follow.

Adsorbent:

$$\left(1-\psi\right)\rho_s c_s \frac{\partial T_s}{\partial t} = \bar{\alpha} a_{Sch}\left(T_G - T_s\right) + \left(1-\psi\right)\lambda_s \frac{\partial^2 T_s}{\partial z^2} + \left(1-\psi\right)\rho_s \Delta h \frac{\partial X}{\partial t} \tag{6.7}$$

Gas:

$$\psi \rho_s c_G \frac{\partial T_G}{\partial t} = \bar{\alpha} a_{Sch}\left(T_s - T_G\right) \dot{v} \rho_G c_G \frac{\partial T_G}{\partial z} + \psi \lambda_G \frac{\partial^2 T_G}{\partial z^2} \tag{6.8}$$

$$\text{If } Bi = \frac{\alpha d_p}{6\lambda_s} < 1 \tag{6.9}$$

the temperature gradient within the adsorbent particle is small. Then, $\bar{\alpha} = \alpha$ [119], where the heat-transfer coefficient of the particle can be calculated as follows:

$$Nu = \frac{\alpha d}{\lambda_s} = \left[1 + 1.15\left(1-\psi\right)\right] Nu_b \tag{6.10}$$

$$Nu_b = 2 + \sqrt{Nu_{lam}^2 + Nu_{turb}^2} \tag{6.11}$$

$$Nu_{lam} = 0.664\sqrt{Re_\psi} Pr^{1/3} \tag{6.12}$$

$$Nu_{turb} = \frac{0.037 Re_\psi^{0.8} Pr}{1 + 2.44 Re_\psi^{-0.1}\left(Pr^{2/3} - 1\right)} \tag{6.13}$$

To calculate the temperature and concentration profiles, as well as the breakthrough curve, the equations for the conservation of mass (Eqs. 6.1-6.4) and energy (Eqs. 6.7-6.9, 6.10-6.13) must be solved simultaneously. Not many results are available, but qualitative statements are possible in principle. The heat released during adsorption at low temperature is transported through the fixed bed mainly by *conduction* and *convection*. The temperature front thus created migrates through the bed in the direction of flow and is coupled to the concentration front via the temperature-dependent equilibrium and the likewise temperature-dependent kinetics of sorption. Figure 44 shows the situation in which a temperature front is getting rapidly ahead, while Figure 45 applies to a combined front in which concentration and temperature increase simultaneously at the end of the bed. In both figures, the profiles of gas concentration and temperature in the bed are plotted on the left; the breakthrough curves (i.e., the profiles of concentration and temperature at the end of the adsorber) are plotted as a function of time on the right.

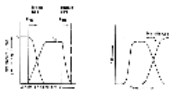

Figure 44. Schematic representation of temperature and concentration curves during adsorption in which the temperature front moves rapidly ahead of the concentration front

—— Concentration; – – –Temperature; – · – · – Width of transfer zones.

Figure 45. Schematic representation of temperature and concentration curves during adsorption—combined wave fronts

—— Concentration; – – – Temperature.

According to [120], the temperature front moves rapidly ahead of the concentration front in the case of a linear adsorption isotherm if the condition

$$\frac{c_s}{c_p} < \frac{\Delta X}{\Delta Y} \tag{6.14}$$

is met. If, on the other hand,

$$\frac{c_s}{c_p} \geq \frac{\Delta X}{\Delta Y} \tag{6.15}$$

a combined wave front exists, i.e., concentration and temperature break through at the same time. Because the ratio of the specific heats of gas and adsorbent is approximately one in the case of oxidic adsorbents and with air as the carrier gas, the slope $\Delta X/\Delta Y$ of the adsorption isotherms will be the key to which case actually occurs. If the temperature front moves ahead, the maximum temperature increase $\Delta\theta_{max}$ and hence the plateau temperature θ_{pl} are obtained from a heat and mass balance over the migrating mass-transfer zone:

$$\Delta\theta_{max} = \theta_{pl} - \theta_0 = \frac{\Delta X \Delta h / c_p}{\left(\Delta X/\Delta Y\right) - \left(c_s/c_p\right)} \tag{6.16}$$

The temperature effects are particularly strong with a large difference in loading ΔX of an adsorptive having a high average heat of adsorption Δh and with a low specific heat capacity c_p of the gas. Temperature increases during adsorption and temperature decreases during desorption can be considerable. Figure 46 shows both the concentration ratio c/c_0 and the temperature difference $\Delta\theta = \theta - \theta_0$ versus time for the adsorption of CO_2 on MS 5A. At a feed concentration of 1 vol % CO_2, the temperature moves ahead of the concentration front (curve a); at higher values, a combined front exists. However, the maximum temperature increase is determined not by the kinetics of adsorption but rather by the course of the temperature increase in the mass transfer zone. While any higher temperatures that occur here produce a minor increase in the diffusion coefficients, they nevertheless lead to lower rates of adsorption, because at a given concentration of adsorptive, the adsorbent cannot be loaded as much. For this reason, the type of isotherm field is important.

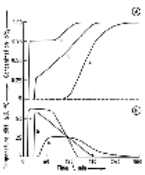

Figure 46. Adsorption System: CO_2–molecular sieve 5A; c/c_0; $\Delta\theta = f(t)$; θ_0=25°C; v=0.2 m/s; d=2.0 mm; D=100 mm; Z=1.0 m

A) Effect of feed concentration of adsorptive on concentration curve; B) Effect of feed concentration on temperature curve

a) c_0=1.0 vol%; b) c_0=3.0 vol%; c) c_0=10.0 vol%

During *desorption by purging*, the temperature drop increases as the loading of the fixed bed increases due to the preceding adsorption; this is shown in Figure 47. Initially, a combined mass- and heat-transfer zone occurs with a steep drop in concentration and simultaneous cooling of the fixed bed. This is followed by an equilibrium or *plateau*

zone in which the temperature level is so low that desorption stops. Only then does the bed warm up again; the adsorbent is desorbed further, now in a combined wave front. In desorption by purging, a considerable amount of purge gas is necessary in the case of an equilibrium that favors adsorption. In such a case, thermal desorption with a heated gas is recommended. Thereby, the plateau temperature will be slightly higher than the initial temperature and will remain constant until the adsorbent is completely desorbed. Then a steep temperature increase follows. As an example, Figure 48 illustrates the thermal desorption of the system CO_2 – MS 5A at a temperature of 200 °C for the regenerating gas. The concentration of carbon dioxide and the temperature at the end of the bed are plotted for z=0.5 m and z=1 m. Only when the concentration drops does the temperature increase steeply. This behavior shows clearly that equilibrium zones occur in the fixed bed when large deviations from isothermal conditions arise. The adsorption or desorption process can then become heat-transport-limited; i.e., it depends only on the supply and removal of heat, respectively.

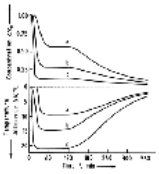

Figure 47. Desorption by purging: effect of feed concentration of adsorptive on the concentration and temperature curves System: CO_2–molecular sieve 5 A; c/c_0; $\Theta f(t)$; Z=1.0 m; Θ_0=25°C; d=2.0 mm; v=0.2 m/s; D=100 mm

a) c_0=1.0 vol%; b) c_0=3.0 vol%, c) c_0=10.0 vol%

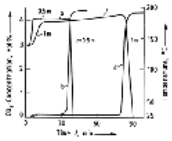

Figure 48. Thermal desorption: temperature and concentration curves in adiabatic fixed bed

System: CO_2–molecular sieve 5 A; Θ_0 =25°C; Θ_ω=200°C; v=0.2 m/s

a)+c) CO_2 concentration at end of bed as a function of time; b)+ d) Time dependence of temperature

6.2.1. *Loading and Breakthrough Curves*

Many literature references verify that Equations (6.1-6.4) describe the rates of both adsorption and desorption on single pellets as well as the breakthrough curves of single- and multicomponent mixtures in isothermally operated fixed beds.

As an example of processes occurring in single pellets during adsorption, as well as during desorption by purging, Figure 49 shows the *change in loading* $\Delta X/\Delta X_{max}$ of carbon dioxide as a function of time for pellets of zeolitic molecular sieve MS 5A with radii R_a of 1.34 and 2.54 mm [119]. The curves apply to pressure jumps from 0 to 3 vol % (adsorption) and from 3 to 0 vol % (desorption). The rate of sorption dX/dt increases as the particle radius decreases. Adsorption is faster than desorption because the adsorption isotherms are strongly curved in a sense that favors the adsorption process. In the calculations, only macropore diffusion in the fluid phase is taken into account. If breakthrough curves for isothermal fixed beds are calculated according to Equations (6.1-6.4), in some cases, transport in the sorbate phase of macropores in planar adsorbents or micropore diffusion in zeolites with cage structure can be neglected.

Figure 50 shows breakthrough curves for a gaseous methanol – water mixture adsorbed at 50 °C on silica gel from nitrogen as carrier gas [86]. An evident hump occurs for the more weakly adsorbed component (water); it leaves the fixed-bed adsorber with a higher concentration than the feed concentration; this roll-over becomes more pronounced, the closer together the adsorption isotherms of the two pure components are. In calculating this curve, only macropore diffusion in the fluid phase is considered in addition to the balance of the fixed-bed adsorber. By contrast, during adsorption of mainly organic materials from the gas [121, 122] and fluid phases [123-125], diffusion in the sorbate phase can contribute substantially to transport in the macropores.

For zeolitic molecular sieves, micropore diffusion in the crystals must sometimes be taken into account during modeling. As an example, Figure 51 shows the relative effluent concentrations of an ethylene – carbon dioxide mixture as a function of adsorption and desorption time at 40 °C [126]. Because the molecular diameters (C_2H_4: 0.43 nm, CO_2: 0.28 nm) are only somewhat smaller than the cage opening (MS 5A: 0.43 nm), and the particle radius (R_a=0.5 mm) is relatively small compared to the radius of the crystal (r_a=0.5 µm), micropore diffusion offers considerably more resistance to transport than macropore diffusion.

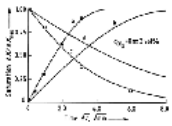

Figure 49. Effect of pellet diameter on course of sorption (according to calculations)

System: Carbon dioxide–molecular sieve 5 A; Θ_0=25°C; v=0.2 m/s; μ_p=4.3

a) Adsorption curve, R_a=1.34 mm; b) Adsorption curve, R_a=2.54 mm; c) Desorption curve, R_a=1.34 mm; d) Desorption curve, R_a=2.54 mm; •, □ measured

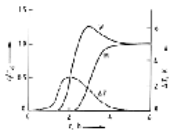

Figure 50. Measured concentration and temperature curves in the system water–methanol–silica gel

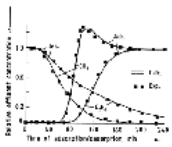

Figure 51. Comparison of calculated and measured ethylene – carbon dioxide adsorption and desorption curves on MS 5A

D=0.03 m; z=0.3 m; v=0.15 m/s; T=313 K; $y_{a,i}$=0.15 vol%

6.2.2. *Simplified Models*

Simplified models to describe the dynamic behavior of an adsorber are illustrated by a system with only a single adsorbable component.

Micropore diffusion is important only if

$$\frac{R_a^2 \mu_p D_m}{r_a^2 D_M}\left(\frac{\rho_p X}{\varepsilon_p c}\right) < 10 \qquad (6.17)$$ [126]

. This means that micropore diffusion can be neglected above all when the radius R_a of the particle of adsorbent is large and the radius r_a of the zeolite crystals is small. By using local equilibrium

$$\frac{\partial X}{\partial t} = \frac{\partial X}{\partial c}\frac{\partial c}{\partial t} \qquad (6.18)$$

the law of conservation of mass (Eqs. 6.2 and 6.3) can be simplified for $D_{S,i}$=0 as follows:

$$\frac{\partial c_M}{\partial t} = \underbrace{\frac{D_M/\mu_p}{1+\frac{\rho_p \partial X_M}{\varepsilon_p \partial c_M}}}_{D_{eff}}\left(\frac{2}{R}\frac{\partial c_M}{\partial R} + \frac{\partial^2 c_M}{\partial R^2}\right) \qquad (6.19)$$

This equation defines an “effective” pseudomacropore diffusion coefficient D_{eff} which, however, now contains the slope $\partial X/\partial c$ of the adsorption isotherms. The definition of the

effective diffusion coefficient does not have a strictly physical meaning; it is, however, very convenient for practical calculations because the literature gives solutions for nonstationary diffusion in spheres under various boundary and initial conditions. Nevertheless, the effective diffusion coefficient is independent of loading only if the adsorption isotherms are linear. For this case, ROSEN provides the following solution to the system of equations (isothermal, external mass transfer, and macropore diffusion in the fluid phase, linear isotherms) for the breakthrough curve of the fixed bed [127]:

$$\frac{c(t,z)}{c_\alpha} = \frac{1}{2}\left\{1 + \mathrm{erf}\left[\frac{\frac{\dot{v}(t - z(\psi/\dot{v}))}{Kz} - 1}{2\sqrt{\frac{1+5D_{\mathrm{eff}}(K/(R_a\beta))}{15D_{\mathrm{eff}}Kz\left(R_a^2\dot{v}\right)}}}\right]\right\} \tag{6.20}$$

where $K = \Pi_p\,(dX/dc)/\epsilon_p$ and z is the axial distance in the fixed bed.

Several breakthrough curves calculated in this way are shown in Figure 52. The desired steep slope of the breakthrough curves is obtained if

1. The mass-transfer coefficient is so large that mass transport through the outer concentration boundary layer is nonlimiting
2. The effective diffusion coefficient is large, and hence the coefficients of free and Knudsen diffusion are large, and the tortuosity factors μ_p are small (curve c)
3. The radius of the adsorbent pellet is small (curve e)
4. The volume flow density of the fluid is low and its concentration is high
5. The slope $\partial X/\partial c$ and K for the adsorption isotherms are large

In some cases, external mass transfer can be neglected in comparison to macropore diffusion. This is generally the case when the Biot number is much greater than 1. This occurs frequently in gas-phase adsorption. In that case, the term

$$Bi = \frac{R_a\beta}{3\varepsilon_p D_M/\mu_p} \tag{6.21}$$

$D_{\mathrm{eff}}K/(R_a\beta)$ can be neglected in the Rosen equation.

Instead of nonstationary pore diffusion according to Equations 6.2 and 6.3, the kinetic formulation of GLUECKAUF [111] is often used (see Eqs. 5.16 and 5.17).

If the adsorption isotherms are not linear, the system of equations for the conservation laws of the fixed bed and the macropores must be solved numerically. In so doing, the rates of adsorption and desorption are different. This is due to the curvature of the isotherms.

The example of adsorption on a single pellet illustrates this (see Fig. 53). If the isotherms are curved upward (Fig. 53, top), the equilibrium is favorable for adsorption but unfavorable for desorption. The diagrams at the bottom illustrate an isotherm which is favorable for desorption but not for adsorption. The diagrams on the right show qualitatively the loading profiles versus particle radius that result from calculations based on the macropore diffusion model. Large changes in partial pressure, and hence in the loading of a relatively thin spherical shell of the adsorbent pellet, occur both in adsorption with an isotherm favorable to it (upper left) and in desorption with its favorable isotherm (lower right). In these cases, the maximum difference in partial pressures in the pellet—the driving difference in concentration—is available over a longer period of time. The profiles in the upper right (desorption with an isotherm unfavorable to desorption) and in the lower left (adsorption with an isotherm unfavorable to adsorption) exhibit no steep profiles for partial pressure. This means that over longer periods of time, only a partial pressure difference $\Delta p_i < \Delta p_{i,\max}$ is effective in pore diffusion. With highly curved isotherms, the rates of adsorption and desorption therefore differ substantially. For practical purposes, two cases are distinguished:

1. The isotherm is curved upward; hence, it is favorable for adsorption and unfavorable for desorption. In this case, the rate of adsorption is greater than the rate of desorption.
2. The isotherm is curved downward; hence, it is favorable for desorption but unfavorable for adsorption. In this case, the rate of desorption is greater than the rate of adsorption.

The difference in rates of adsorption and desorption increases with the deviation of the isotherms from linearity. This shows clearly that in processes of adsorptive separation (adsorption of one component with subsequent desorption), the useful shape of the isotherms represents a compromise between two divergent requirements.

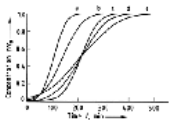

Figure 52. Breakthrough curves calculated according to the Rosen equation

Standard case: K_{st}=2540; v_{st}=0.12 m/s; D_{eff}=1.9×10^{-10} m^2/s; $R_{a,s}$=4×10^{-3} m; z=1.5 m; Ψ=0.4

a) $K=K_{st}/2$; b) $v=1.5\,v_{st}$; c) $D_{eff}=2\,D_{eff,st}$; d) Standard case; e) $R_a=1.5\,R_{a,st}$

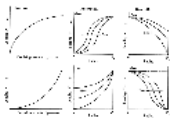

Figure 53. Change of loading of a single pellet with time for an isotherm favorable to adsorption but unfavorable to desorption (top) and for one favorable to desorption but unfavorable to adsorption (bottom)

6.2.3. *Coefficients*

The balance equations contain the axial dispersion coefficient D_{ax} of the fixed bed, the mass-transfer coefficient β between fluid and adsorbent particle, the diffusion coefficient D_M in the macropores, and the diffusion coefficient D_m in the zeolite cages.

Axial Dispersion Coefficient. In a fixed-bed adsorber, plug flow is desirable. Deviation from this flow is described by the axial dispersion coefficient D_{ax} or a dimensionless number K of the fixed bed

$$K = \frac{\dot{v}Z}{\psi D_{\mathrm{ax}}} \tag{6.22}$$

This dimensionless number is a function of D_{ax}, the volume flow density $\dot{v}$, the porosity Ψ, and the bed length Z. Two limiting cases can be distinguished.

1. $$\text{Pure plug flow}: \quad \frac{\dot{v}Z}{\psi D_{\mathrm{ax}}} \to \infty \tag{6.23}$$

2. $$\text{Complete mixing}: \quad \frac{\dot{v}Z}{\psi D_{\mathrm{ax}}} \to 0 \tag{6.24}$$

The axial dispersion coefficient is a function of the dimensionless Péclet number Pe

$$D_{\mathrm{ax}} = \frac{\dot{v}d}{\psi Pe} \tag{6.25}$$

where $d=2R_a$ is the particle diameter. The Péclet number depends on the product of the Reynolds number of the particle and the Schmidt number of the fluid:

$$Re\,Sc = \frac{\dot{v}d}{D_{\mathrm{AB}}} \tag{6.26}$$

where D_{AB} is the diffusion coefficient of the fluid. This relationship can be read from Figure 54 [128]. For small characteristic numbers (K < 100), which occur in vacuum and

liquid-phase adsorption, axial dispersion can lead to an earlier breakthrough of the adsorptive compared with plug flow and hence to a reduced usable capacity. For $K > 100$ and $D_{ax} \longrightarrow 0$, the corresponding term in the adsorber balance equation (Eq. 6.1) can be neglected. This is sometimes the case in gas-phase adsorption.

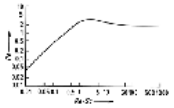

Figure 54. Péclet number *Pe* as a function of *Re·Sc*

Range of validity: 0.008 < *Re* < 400; 0.28 < *Sc* < 2.2

Mass-Transfer Coefficient. The mass-transfer coefficient β between fluid and adsorbent particle surface depends on the Reynolds number of the particle and the Schmidt number $Sc = \nu/D_{AB}$ of the fluid, where ν is the kinematic viscosity. The mass-transfer coefficient can be determined with the aid of Equations 6.27-6.30.

$$Sh_{fb} = \frac{2R_a\beta}{D_{AB}} = \left[1 + 1.5\left(1 - \psi\right)\right] Sh_b \tag{6.27}$$

$$Sh_b = 2 + \sqrt{Sh_{lam}^2 + Sh_{turb}^2} \tag{6.28}$$

$$Sh_{lam} = 0.664\sqrt{Re_\psi}\,Sc^{1/3} \tag{6.29}$$

$$Sh_{turb} = \frac{0.037\,Re_\psi^{0.8}\,Sc}{1 + 2.44\,Re_\psi^{-0.1}\left(Sc^{2/3} - 1\right)} \tag{6.30}$$

The Sherwood number Sh_{fb} in a fixed or moving bed depends on the Sherwood number Sh_b (see Eq. 6.28) of individual beads, which in turn is a function of the Schmidt and Reynolds numbers of the fluid (Eqs. 6.29 and 6.30). Here, the fluid passes the adsorbent particle with an effective velocity $\dot{v}/\psi$. In liquid-phase adsorption, the mass-transfer coefficient can be calculated from similar relations [129].

Macropore Diffusion Coefficient and Tortuosity Factor.
Macropore Diffusion Coefficient. In the macropores, material can be transported by convection and diffusion in the fluid and in the sorbate phases. If a drop of total pressure occurs in gas-phase adsorption, flow will result. Depending on the Knudsen number *Kn* (ratio of the mean free path of molecules to the pore diameter), this may be laminar flow ($Kn < 0.01$), slip flow ($0.01 < Kn < 1$), or molecular or Knudsen flow ($Kn > 1$). If the influence of rapid deflation or pressurization in the adsorber is neglected, and only slow adsorption or desorption processes are considered, mass transport contributions from viscous flow in the pores can often be neglected. Mass is then transported in the fluid phase on the basis of *free diffusion* with diffusion coefficient D_{AB}, as well as *Knudsen diffusion* with diffusion coefficient D_{Kn}. In planar adsorbents, *transport* in the sorbate phase (*surface diffusion*), with diffusion coefficient D_S, may also contribute to the overall mass-transfer rate. If the adsorbent is loaded to only half the monomolecular coverage, transport in the sorbate phase can be neglected because the molecules are almost immobile due to the high energy of binding. Mass is then transported only by free and Knudsen diffusion in the macropores. The coefficients for free diffusion in a gaseous or fluid phase D_{AB} and for Knudsen diffusion D_{Kn} can be calculated from Equations 6.31-6.34. The macropore diffusion coefficient D_M can be determined from them by superposition of the coefficients D_{AB} and D_{Kn} [119, 130]. At high total pressures, $D_M \approx D_{AB}$; in a vacuum, $D_M \approx D_{Kn}$.

Free diffusion:

$$D_{AB} = \frac{2}{3}\sqrt{\left(\frac{k}{\pi}\right)^3}\sqrt{\frac{1}{2m_A} + \frac{1}{2m_B}}\,\frac{\sqrt{T^3}}{p[(d_A + d_B)/2]^2} \quad \text{(Gases)} \tag{6.31}$$

$$D_{AB} = 7.4 \cdot 10^{-12}\frac{\sqrt{\psi_B M_B T}}{\eta(V_A)^{0.6}}\left[\frac{m^2}{s}\right] \quad \text{(Liquids)} \tag{6.32}$$

(M_B in kg/kmol, T in K, V_A in cm^3/mol, η in mPa·s)

Knudsen diffusion:

$$D_{Kn} = \frac{4}{3}\sqrt{\frac{RT}{2\pi}}\,d_M \tag{6.33}$$

Superposition of free gas diffusion and Knudsen diffusion:

$$D_M = \frac{1}{\frac{1-(1+(\dot{n}_B/\dot{n}_A))y_A}{D_{AB}} + \frac{1}{D_{Kn}}} = \frac{1}{\frac{1}{D_{AB}} + \frac{1}{D_{Kn}}} \tag{6.34}$$

(for equimolar counterdiffusion)

Tortuosity Factor. Because macropores are not straight cylinders, but rather tortuous with changing and partly slitlike cross sections, a tortuosity or detour factor μ_p, also called the diffusion resistance factor, must be introduced. On the basis of this definition, the factor should depend only on the structural characteristics of the adsorbent, not on operating conditions such as pressure, temperature, type and concentration of adsorptive, or kind of operation (adsorption or desorption). Such is, however, not the case for many values given in the literature. Often, the quantity μ_p contains either the porosity ϵ_p of the adsorbent particle or the reduction of total diffusion by Knudsen diffusion [131]. In other cases, the coefficient of Knudsen diffusion is calculated for a cumulative undersize passage of the macroporosity of 0.5 [$\epsilon(r_p)/\epsilon_{macro}=0.5$], which may not weight correctly the contributions of pores of various sizes to Knudsen diffusion [119, 126], [132]. Because tortuosity factors are almost always curve-fitting factors resulting from a comparison of measured and calculated breakthrough curves, they are defined by the corresponding calculation model and cannot be applied to other models. Tortuosity factors that do not contain any effects of Knudsen or micropore diffusion (e.g., determined during the adsorption of small molecules in wide micropores at higher pressures) and are corrected for the porosity of the adsorbent should be between 3 and 6.

Diffusion coefficients D_S in the sorbate phase can be determined experimentally only by recording loading curves for a single pellet. This loading curve should be described as accurately as possible by means of the macropore diffusion model. The diffusion coefficient D_S in the sorbate phase is then a curve-fitting parameter. Coefficients determined in this way are often in the range 10^{-7} $m^2/s < D_S < 10^{-12}$ m^2/s [133] and are strongly dependent on temperature because diffusion in the sorbate phase is an activated process.

Micropore Diffusion Coefficient. The micropore or zeolitic diffusion coefficient D_m depends on the coefficient D_0^0, on temperature, and on loading:

$$D_m = D_0^0 \exp\left(-\frac{E_a}{RT}\right)\frac{\partial \ln c}{\partial \ln \overline{X}} = D_0^0 \exp\left(-\frac{E_a}{RT}\right)\frac{\overline{X}\,\partial c}{c\,\partial \overline{X}} \tag{6.35}$$

where $\overline{X}$ is the mean mass loading of adsorbent. The factor D_0^0 and the activation energy E_a are functions of micropore geometry, as well as shape, size, and polarizability of the adsorptive molecules. The larger the critical diameter of the molecules compared to the micropore diameter, the smaller is D_0^0. The activation energy E_a is often between 10 and 40 kJ/mol and increases with molecular diameter [133]. Micropore or zeolitic diffusion coefficients cannot be precalculated; they must be determined experimentally. They are frequently in the range 10^{-16} $m^2/s < D_m < 10^{-11}$ m^2/s and depend, among other things, on the type of adsorptive and zeolite.

[Top of Page]

7. Regeneration of Adsorbents

Loaded adsorbents can be regenerated by temperature or pressure swing processes and by displacement. In some cases, the adsorptive can be extracted from the adsorbent. If the pores of an adsorbent become plugged, reactivation is necessary.

Regeneration is favored at low pressures, high temperatures, and large differences in concentration, which results in pressure swing adsorption (PSA), temperature swing adsorption (TSA), and composition swing adsorption (CSA). Sometimes several effects are used simultaneously, as in steam regeneration, where steam acts as displacer and increases the temperature. In vacuum swing adsorption (VSA) desorption is carried out below atmospheric pressure.

Figure 55 shows how the loading of adsorbent is reduced by an increase in temperature (*temperature swing*, TSA) or by a (partial) pressure reduction (*pressure swing*, PSA). The more the adsorption isotherms favor adsorption, the higher are the heats of adsorption and binding, and the greater is the use of thermal regeneration. Desorption by pressure swing is often economical if the heat of adsorption is less than about 30 kJ per mole of adsorptive. Finally, costs (investment cost, cost of supplying thermal energy during temperature swings, cost of electrical energy for compressors or vacuum pumps during pressure swings) will dictate the choice of regeneration method.

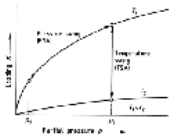

Figure 55. Schematic representation of temperature swing and pressure swing processes

7.1. Regeneration by Temperature Swing

Figure 56 shows an installation in which the temperature-swing process is used. To remove adsorbate, the adsorbent is heated either indirectly via an external double jacket and an internal heating surface or directly via a hot gas. The remaining residual loading depends on the final temperature of the adsorbent and, possibly, the adsorptive loading of the hot gas. The temperature-swing process can be used efficiently only if the loading of the adsorbent decreases substantially as temperature increases. This is the case, for example, for the system water – molecular sieve. If the regenerating gas at a temperature of 200 °C (or 300 °C) has a partial pressure $p(H_2O)$=1.5 kPa and if it is used to heat to 200 °C (or 300 °C) a molecular sieve bed with an equilibrium loading X=0.22 kg/kg at 25 °C, the loading decreases to X=0.04 kg/kg (or 0.015 kg/kg at 300 °C). Regeneration costs are determined primarily by the course of the adsorption isotherms and the temperature difference between adsorption and desorption. In addition, the heating of a larger fixed bed may take hours because the adsorbents have low thermal conductivity.

In direct heating, the heated regeneration gas often flows countercurrently, i.e., in the direction opposite to that of the gas treated in the preceding adsorption step. Thus, a very clean zone at the end of the fixed bed, with a very low residual loading of adsorbent for the following adsorption step, results. In some applications, the required bed length decreases, so that investment and operating costs are reduced.

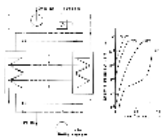

Figure 56. Adsorption unit with regeneration of adsorbent by the temperature swing process

Equilibrium diagram shows adsorption isotherms of water vapor on a 0.5 nm molecular sieve.

a) Heating; b) Cooling

Closed-circle regeneration is advisable when the gas to be treated and the heated regenerating gas must not be mixed. A blower circulates the regenerating gas—often an inert gas, cleansed of adsorptive—through a heater and a cooler. The heater warms up the gas before it enters the adsorber to be regenerated, while the adsorptive that has accumulated in the regenerating gas is condensed in the cooler. If the adsorber is operated at elevated pressure and the gas is heated during compression, the heat of compression can be used for regeneration by appropriate connections. To do so, part of the hot gas is sent to the adsorber to be regenerated. The gas is then loaded with adsorptive and rejoins the suitably throttled main stream of gas. In this case, the adsorptive is also condensed and removed in a cooler.

7.2. Regeneration by Pressure Swing

Regeneration by pressure swing is carried out at the approximate temperature of the adsorption step but at lower pressure. If adsorption is carried out at elevated pressure, a *release of pressure* is often sufficient. On the other hand, *vacuum* must be applied if adsorption is carried out at ambient pressure. The pressure-swing process also requires a favorable equilibrium. As much adsorptive as possible should be removed for a given pressure drop. Figure 57 shows an installation in which the adsorbent is regenerated by the pressure swing process. The equilibrium diagram contains 25 °C isotherms for both methane and carbon dioxide. If, for example, adsorption occurs at 0.1 MPa and the regeneration pressure is reduced to 25 kPa, the loading of methane drops from 0.075 to 0.025 kg/kg, but that of carbon dioxide drops only from 0.186 to 0.169 kg/kg. Although carbon dioxide can be removed at very low regeneration pressure, energy consumption by the vacuum pump is increased. The course of the adsorption isotherms along with the pressure difference during adsorption and desorption, essentially determines the cost of regeneration. Pressure reduction is rapid, requiring only minutes or even seconds.

Figure 57. Adsorption unit with regeneration of adsorbent by the pressure-swing process

Equilibrium diagram shows adsorption isotherms for methane and carbon dioxide on 0.5 nm molecular sieve.

7.3. Regeneration by Displacement

During regeneration by displacement, the adsorbate is removed from the adsorbent by displacement from the active sites due to adsorption of a third material (the displacing agent). Examples are the removal of organic substances from active carbon by steam (see Section Removal of Organic Components from Exhaust Air) and the displacement of long-chain by short-chain alkanes from zeolitic molecular sieves (see Section Separation of Isomers). The choice of a suitable displacement agent depends both on the equilibrium of the system and on the kinetics of adsorption and desorption.

The most common displacing agent is steam, especially when water and adsorptive are immiscible. To prevent early condensation at the top of the column supersaturated steam at about 130 °C and 1 bar is preferentially used [134-136]. Characteristic is also the occurrence of a water front within the adsorber [137]. This should not be confused with the mass-transfer zone, which is not very sharp during desorption. At higher residence times the superheated steam desorbs water and leads to a drying effect behind the water front. This results in a distinct maximum water loading in the bed (Fig. 58). Steam flow in the same direction as the flow direction during adsorption is thus advantageous [137, 138].

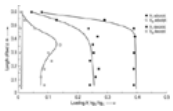

Figure 58. Axial loading profiles of toluene and water after adsorption and desorption

The commonly applied drying in the same flow direction as in the previous adsorption step gives almost the same results as if no drying was done. It has been shown [136] that an internal local desorption/adsorption process, due to pronounced differences in the rate-determining step of mass transfer on the local particle, is the cause of this inefficiency. To achieve a long breakthrough time with an environmentally acceptable low emission level, a long part of the bed must be regenerated towards too low loading levels. If this low-loaded part is too short, the activated carbon cannot adsorb the solute in the necessary amount. That means the exit gas concentration of the column is too high for environmental purposes. If this limit has to be fulfilled, the length of the well-cleaned part of the bed must be long enough to enable a sufficient mass transfer, even when only a small driving force between the gas and the carbon phase exists. Otherwise, if the part of the column which has a sufficient low residual carbon loading is too short, early breakthrough will occur. This is why a drying countercurrent to the adsorption step is always of benefit [136].

To save energy and costs the adsorbent is not fully dried and there is also a residual loading profile of the adsorbate. The experimental investigation of sorption cycles in a small technical column [136, 138] revealed that the regeneration conditions influence the following adsorption much more than the type of adsorbent used.

7.4. Regeneration by Extraction

In regeneration by extraction, the adsorptive is picked up by an extracting agent which is brought into contact with the loaded adsorbent. In this manner, adsorbed sulfur can be extracted with carbon disulfide or a solution of caustic soda (see Section Adsorptive Desulfurization Processes). Adsorbents loaded with sulfuric acid can be regenerated by extraction with water or ammonia.

7.5. Reactivation of Adsorbents

Reactivation may be necessary when the adsorbate is not volatile or soluble enough to be desorbed by increasing the temperature, reducing the concentration, or extracting

with solvents. This is the case for adsorbates on granular *activated carbon* during water purification or during decolorization of sugar-beet juices. Adsorbed organic impurities coke during heating and permanently damage the adsorption agent. New pores for adsorption can then be created by gas activation with carbon dioxide and gas containing water vapor at ca. 900 °C (⟶ Carbon – Carbonaceous Adsorbents).

This type of regeneration always entails a loss of activated carbon because it involves a chemical reaction between activated carbon and the regenerating gas. Additional losses occur due to mechanical abrasion during transport and in the furnace. In general, a loss of ca. 10 % is necessary to regain 90 – 95 % of the original adsorption capacity. Higher capacity entails higher losses and vice versa.

Powdered coal can be reactivated as a sludge by atomization in a combustion chamber at 850 – 900 °C and at a residence time of about 1 s [139].

Thermooxidative regeneration may also be required for *molecular sieves* if polymers or carbon deposits have formed on the adsorbent. The impurities must then be burned off under controlled conditions at low oxygen concentration and about 500 °C.

Fuller's earth, which is used in the treatment of edible oil or mineral oil products, is also regenerated thermally or oxidatively. First, the adhering oil is extracted with a solvent; then the fuller's earth is heated in a stream of oxygen-containing gas to 350 – 550 °C. During this treatment, carbon and organic substances are burned off. This regeneration can be carried out in a fluidized bed of low bed height, placed in a vibrating channel, at residence times of a few minutes. Temperature treatment is followed by an acid wash and drying.

7.6. Liquid Phase Desorption

Adsorbents used in liquid-phase processes are regenerated similarly as described above for gas-phase processes. However, thermal activation will cause losses of the adsorbent of about 5 – 15 %. This is feasible with activated carbon, alumina, and pillared clays but not economical with expensive sorbents like zeolites or polymeric adsorbents.

However, highly volatile adsorbates are easily removed by a temperature swing with hot purge gas, for instance, when drying organic liquids or compressed gases and then removing water by TSA. Desorption with steam is similar in price to extractive desorption, whereas the greater flexibility of the latter leads to broader application. The efficiency of chemical regeneration is higher when small-molecule solvents which can enter the micropores are used [140, 141]. In many cases it can be performed in situ and there is then a need to distillate the extract prior reuse. Also the bed has to be backwashed or steam-stripped to remove entrained solvents [66, 142].

Other methods are to change the ionic strength or pH value. The latter is applied with dissociating species, since the undissociated form is better adsorbed. Thus, organic acids, phenols, polyphenols, and their derivatives are more easily desorbed at a high pH. High salt concentrations promote a salting-out effect, and regeneration can then be performed with pure water [143]. These salt effects are seldom used, but the general rule is that small ions act as salting-out agents, whereas large ones tend to salt-in [144]. Alternative solvents, like supercritical fluids (liquid CO_2, etc.) or micellar surfactant solutions, are not commonly applied for reasons of cost.

[Top of Page]

8. Gas Phase Processes

Adsorption processes in the gaseous or liquid phase are used either to remove certain components from a fluid phase (e.g., air, water) or to separate a mixture.

The diameter of fixed-bed adsorbers depends on the permissible superficial velocity of the fluid when pressure loss is considered. The so-called *F* factor $\dot{v}\sqrt{\rho}$ is often between 0.05 and 0.5 m s^{-1} $kg^{1/2}$ $m^{-3/2}$. The pressure drop Δp of a fluid flowing through a fixed bed of spherical particles with height *Z* and porosity Ψ is

$$\frac{\Delta P}{Z} = \xi \frac{(1-\psi)}{\psi^3} \frac{\dot{v}\rho}{2} \frac{1}{2R_a} \tag{8.1}$$

where the drag number ξ depends on the Reynolds number *Re* ψ (see Fig. 59).

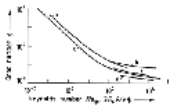

Figure 59. Drag number ξ as a function of Reynolds number *Re* ψ for various packings

a) Coke and ore packings; b) Raschig rings; c) Berl saddles; d) Irregular sphere packings; e) Regular sphere packings

The length of the adsorber primarily depends on the length of the mass-transfer zone and the length of the unused bed (LUB) as well as the cycle time.

In addition to fixed beds, moving and fluidized beds are utilized for gases and liquids. Liquids can also be treated adsorptively by being brought into contact with adsorbents in mixers. After adsorption, the suspension is separated mechanically by filtration, sedimentation, or centrifugation.

Industrially important processes generally involve the separation of gas mixtures, often using the pressure swing process. These may include the removal of water from various gases (CO_2, SO_2, H_2S, NO_x) and organic or radioactive materials from air, exhaust gas, or flue gas. The adsorptive purification of natural gas, hydrogen purification, the removal of CH_4, CO_2, CO, and N_2, and the separation of isomers are also important.

8.1. Adsorptive Separation of Gas Mixtures

Mixtures in the gas or vapor phase can be separated by rectification (⟶ Distillation and Rectification), absorption (⟶ Absorption), membrane separation processes (⟶ Membranes and Membrane Separation Processes), partial condensation or desublimation, or adsorption. Adsorptive separation may be based on *steric, equilibrium,* or *kinetic effects*.

With the *steric effect*, only molecules whose critical diameter is less than the diameter of the micropores are adsorbed in the latter. A mixture is thus separated by molecular size. Both zeolitic molecular sieves with cage openings of 0.38, 0.43, 0.5, 0.8, or 0.9 nm and carbon molecular sieves with micropore widths of 0.3, 0.35, or 0.5 nm are utilized in adsorptive separation based on the steric effect. Examples are the separation of iso- and *n*-alkanes by using molecular sieve 5A (0.43 nm) and the separation of nitrogen and oxygen by using a carbon molecular sieve.

In addition, mixtures of gases can be separated by a *kinetic effect* if one component is adsorbed much faster than the other(s). This is generally due to the higher micropore diffusion coefficient of the first component. Depending on the average size and distribution of micropores, the steric or the kinetic effect may predominate. If the critical molecular diameter of an adsorptive is almost the same size as the diameter of the micropores, a *steric – kinetic effect* can occur because adsorptives with smaller molecular diameters are adsorbed much more rapidly.

Separation of mixtures by the *equilibrium effect* is based on one component being bound more strongly than the other(s). A large distance between single-component adsorption isotherms favors this effect (see Fig. 60). Highly bulging equilibrium curves are thus obtained, as shown in Figure 61. Thus, for example, methane – ethane mixtures can be separated with activated carbon [145] and nitrogen – oxygen mixtures with molecular sieve 5A based on the equilibrium effect.

In adsorptive separation processes, loaded adsorbents can be regenerated by temperature swing, pressure swing, or displacement by an additional component. Pressure swing processes are widely used for a number of gaseous mixtures whose components are adsorptively bound at only moderate strength.

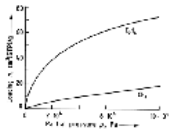

Figure 60. Single-component adsorption isotherms for methane and ethane on activated carbon at 20°C [66]

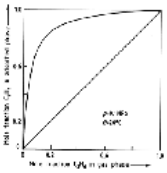

Figure 61. Mixed adsorption equilibrium during adsorption of methane and ethane on activated carbon at 20°C [90]

8.2. Pressure Swing Processes

In pressure swing processes, the following steps occur successively: 1) pressure buildup, 2) adsorption at high pressure and hence production of a pure component, 3) depressurization, and 4) purging at low pressure or under vacuum (see Fig. 62). Step 4 is not always necessary.

In an installation that is run continuously with constant feed of a mixed stream, several units are placed side-by-side. The number of adsorbers is at least two. Because the four steps are of unequal duration, a larger number of adsorbers may be more cost efficient.

During depressurization, a portion of the bound adsorptive desorbs, due partly to the pressure swing effect and partly to the increased temperature of the adsorbent resulting from the heat liberated in adsorption (temperature swing effect). When the bed is at low pressure, it is purged usually with pure product component. In a sense, this corresponds to reflux during rectification (⟶ Distillation and Rectification – Continuous Rectification (Multiple Distillation)), the reflux ratio being dictated by the desired purity of the product and being at least one for a linear adsorption isotherm [146]. During adsorption, the component that is more weakly bound is generally obtained in fairly pure form. The more strongly bound component becomes the desorbate during regeneration. If its purity is not adequate, an additional adsorptive separation can follow.

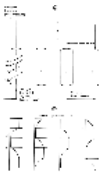

Figure 62. Pressure-swing process with four beds

A) Adsorptive separation of different impurities from hydrogen; B) Pressure change in individual beds as a function of time: At t=0, adsorption occurs in bed 1 at the highest pressure of the process; bed 2 is being partially depressurized; bed 4 partially pressurized; the flash gas of bed 2 is used as stripping gas for bed 3; bed 4 is pressurized by part of the pure gas and is used as an adsorber in the next step

a) Pressurization; b) Depressurization – pressurization; c) Depressurization – purging

If several components are to be removed adsorptively, a bed with zones containing different adsorbents (mixed bed) can be used. For example, if both water and carbon dioxide are to be removed from air, water can be separated on activated aluminum oxide as soon as the gas enters; then, carbon dioxide can be removed by using a zeolitic molecular sieve. Because the curvature of the adsorption isotherms is slight in both cases, pressure swing regeneration presents no problems (see Section Regeneration by Pressure Swing).

The choice of adsorbent depends primarily on the separation effect employed. If separation is carried out on the basis of the equilibrium effect, the adsorption isotherms should be approximately linear so that the resulting rate of desorption by depressurization is not too low. Although the capacity of the adsorbent decreases with increasing temperature, the linearity of an isotherm favorable to adsorption increases. The choice of a proper pressure ratio p_{ad}/p_{des} is particularly important and depends on the position and shape of the isotherms. Table 7 lists examples of adsorptive separation by the pressure-swing process [147-153]. Several of these are very important in industry and are explained in more detail.

Table 7. Examples of pressure-swing proceses

Mixture	Adsorbent	P_{ad}/P_{des}, MPa/MPa	T, °C	Ref.
H_2 isotopes	Pd on AO	0.79/0.15	25	[147]
$H_2 – CO_2$	CMS, ZMS, AAO			
$H_2 – N_2$	CMS, ZMS, AAO			[148]
$H_2 – CH_4$	CMS, ZMS, AAO			
$H_2 – H_2S$	CMS, ZMS, AAO			
$N_2 – O_2$	CMS, ZMS			[148]
	CMS	0.6, 0.7, 0.8/0.1		[149]
	CMS	0.35, 0.18/0.01		[149]
$CO_2 – CO$	CMS			[148]
CO_2 – air	ZMS “Zeosorb” 13 X	10/0.013	5 – 20	[150]
$N_2 – CH_4$	ZMS 5 A	0.14/0.006	25	[151]
$N_2 – C_2H_4$	CMS 5 A	0.34/0.1	30	[152]
$O_3 – O_2$	ZMS, SiO2			[153]

*AAO = activated aluminum oxide, AO = aluminum oxide, CMS = carbon molecular sieve, ZMS = zeolitic molecular sieve.

8.3. Separation of Nitrogen from Oxygen

Nitrogen – oxygen mixtures can be separated adsorptively (see also ⟶ Nitrogen – Air Separation by Adsorption) by both the equilibrium and the steric effect, depending on the type of adsorbent used.

Separation by the Equilibrium Effect. The adsorption of oxygen on zeolitic molecular sieves is weaker than that of nitrogen (see Fig. 63) [145, 154]. Based on the equilibrium effect, the gas leaving the adsorption step is rich in oxygen. The desorbate resulting from pressure-swing regeneration consists largely of nitrogen. The purity of oxygen can be as high as 95 %, which is satisfactory for some applications. Adsorption is carried out conveniently at a pressure of 0.6 MPa and at ambient temperature. For desorption, the pressure is lowered to 0.1 MPa.

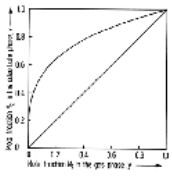

Figure 63. Equilibrium curve of nitrogen–oxygen system on molecular sieve 5 A at –130°C and 0.1 MPa

Separation by the Steric Effect. Nitrogen – oxygen mixtures can also be separated on carbon molecular sieves with micropores of 0.3 nm, based on the steric effect [145, 155]. The critical molecular diameters of oxygen and nitrogen are 0.28 and 0.3 nm, respectively. This means that nitrogen molecules can force their way into the micropores only by micropore or crack diffusion. For that reason, oxygen is adsorbed much more rapidly than nitrogen, as shown in Figure 64, where the relative loading X/X_{max} of a pellet is plotted as a function of time for oxygen and nitrogen. Within a short time, the pellet is almost loaded with oxygen while the adsorption of nitrogen is considerably

lower. The purity of nitrogen is as high as 99 %, and that of oxygen, ca. 80 – 98 %, depending on the process steps and the operating procedure. The bed loaded with oxygen is regenerated either by evacuation to 10 kPa or by purging with a gas component at atmospheric pressure [155].

Separation on carbon molecular sieves based on the equilibrium effect is much less favorable because the adsorption isotherms of oxygen and nitrogen on carbon molecular sieves are relatively close together.

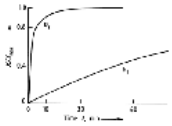

Figure 64. Loading as a function of time for oxygen and nitrogen on a carbon molecular sieve at 20°C

8.4. Adsorptive Purification of Air

Drying of Air. Industry frequently requires air with a very low *dew point.* Examples are control air for pneumatic controllers, air for ventilating storage tanks of organic liquids, and process air for the chemical industry.

Air can be dried with the aid of zeolitic molecular sieves, silica gel, and activated alumina. These adsorbents are usually regenerated by the temperature swing process. The regeneration temperature depends on the dew point desired and hence on the residual loading of dried air. Some regeneration temperatures and heat consumption per kilogram of water to be removed are listed below:

Silica gel	120 – 150 °C	ca. 8000 kJ/kg
Aluminum oxide	160 – 190 °C	ca. 11000 kJ/kg
Molecular sieve	200 – 300 °C	ca. 12000 kJ/kg

With silica gel, dew points of dried air of ca. –60 °C can be attained; with aluminum oxide, ca. –70 °C; and with molecular sieves, ca. –80 °C. The relation between the saturation concentration of water vapor in air ($c^* = p_s/RT$) and the dew point is shown in Figure 65.

If the air to be dried is compressed to more than 0.5 MPa, the adsorbent can also be regenerated by pressure swing. To do so, part of the stream of dried gas is expanded and conveyed countercurrently into the previously depressurized loaded adsorber. During the adsorption step, the pressure is at least 0.5 MPa, while desorption is carried out at 0.1 MPa.

Removing Traces of Oil. If the air contains traces of oil in the form of droplets or vapor, these can be efficiently removed by activated carbon beds. Loaded beds are usually regenerated with steam.

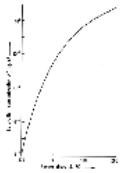

Figure 65. Saturation concentration $c^*=p_s/RT$ of water as a function of temperature

8.5. Removal of Radioactive Nuclides from Exhaust Gas

Exhaust gases of nuclear power stations contain gaseous radioactive nuclides such as tritium, nitrogen, oxygen, argon, krypton, and xenon in concentrations far below 10^{-6} vol % [156]. Hydrogen-containing tritium is burned catalytically, and the water is purified. Short-lived nitrogen, oxygen, krypton (except ^{85}Kr), and xenon nuclides (half-lives from a few minutes to a few days) are adsorbed on activated carbon, the bed providing lag time for radioactive decay. Tests have shown that activated carbon with a micropore diameter of about 0.65 nm has the highest retention factor for krypton.

Since ^{85}Kr has a half-life of 10.7 years, the principle of lag time cannot be used. It is therefore adsorbed on activated carbon or molecular sieves at low temperature (–40 to –196 °C). The loaded adsorbent can be regenerated by heating to 20 – 100 °C, by lowering the pressure, or by purging with inactive gas.

Radioactive iodine is separated on activated carbon impregnated with potassium iodide. In so doing, isotope exchange takes place. Radioactive iodine, in the form of silver iodide and iodate, can also be bound by a chemical reaction to catalyst carriers based on ceramics that have been coated with silver nitrate.

8.6. Removal of Organic Components from Exhaust Air

Organic materials are most frequently separated from exhaust air by adsorption on activated carbon, the loaded adsorbent being regenerated by displacement desorption with steam. This is particularly advantageous if the organic substance is insoluble or sparingly soluble in water, because the mixture formed during regeneration (steam + vapor of organic component) can then be separated readily into two liquid phases after condensation.

Occasionally, molecular sieves are used as adsorbents because they can be regenerated by purging with a heated gas. In principle, this type of regeneration can also be employed with activated carbon. The adsorbent is then purged with an inert gas (e.g., nitrogen) which circulates through a condenser and a heater. The organic materials are condensed in the cooler. The heater serves to preheat the purge gas before it enters the adsorber.

Figure 66 depicts an installation for purifying exhaust air in which organic components are adsorbed on activated carbon and the loaded adsorbent is regenerated by steam. The exhaust air, which contains organic materials, flows from the bottom to the top through the bed of activated carbon (a). When the exit concentration reaches permissible values (between 20 and 150 mg/m^3, according to the 1986 clean air regulations in the Federal Republic of Germany (TA Luft) [157]), the stream of exhaust air is directed to the regenerated adsorber. Superheated steam flows countercurrently, from top to bottom, through the loaded fixed bed, displacing the organic materials. The mixture of steam and organic vapors is liquefied in the condenser (c). The regenerated adsorber is then dried by means of a hot gas and later cooled with a cold gas.

The concentration of organic materials attainable in the purified gas depends on the conditions of bed regeneration (temperature and residual loading of water), especially of the top layer. Other important factors are the water vapor content of the air and thus its relative saturation with water vapor [158]. This can be understood by considering the adsorption isotherms of water, methanol, and the aromatic compounds benzene and toluene (see Fig. 67). Although activated carbon strongly adsorbs benzene and toluene at low relative saturation φ because of its hydrophobic character, and only weakly adsorbs water, it still accumulates considerable amounts of water by capillary condensation at $\varphi > 0.5$.

Figure 66. Schematic of an installation for purifying exhaust air

a) Adsorber; b) Blower; c) Condenser; d) Pump; e) Separator; f) Stripping column; g) Evaporator

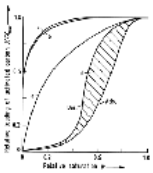

Figure 67. Adsorption isotherms on activated carbon at 20°C

a) Benzene (nonpolar); b) Toluene (nonpolar); c) Methanol (weakly polar); d) Water (highly polar)

As a result, the lower the water vapor content of the air, the greater is the adsorption capacity for organic compounds [158, 159]. Figure 68 shows the effluent loading with toluene as a function of time for air at 21 °C, initially loaded with about 1 g of toluene per cubic meter of air, at relative water saturation φ between 0.30 and 1 [158]. These studies and others on methylene chloride and methyl isobutyl ketone have shown that moisture in the gas does not impair complete purification at very low concentration. However, insufficient evaporation of organic vapors and insufficient removal of condensing water vapor from the pores of activated carbon lead to higher contents of organic components in the purified gas. The amount of steam that condenses should just suffice to heat the activated carbon so as to desorb the organic materials [158]. In this case, water loading will be less than about 0.1 kg per kilogram of activated carbon, and very little capillary condensation will occur. During the adsorption step, the molecules of organic substances can then penetrate the pores rapidly.

Tests have shown that using superheated steam and preheating the activated carbon to be regenerated both favorably affect the attainable concentration of purified gas, as well as the adsorption capacity of the bed [159]. This is illustrated in Figure 69, in which the effluent loading is shown as a function of the loading time for various temperatures of regenerating steam, with and without preheating of the bed, and for two different amounts of regenerating steam, 3 and 6 kg of steam per kilogram of organic material. Almost equal purification can be obtained when the amount of steam is lowered from 6 to 3 kg/kg if the steam is superheated to 150 °C and the bed is preheated to about 125 °C.

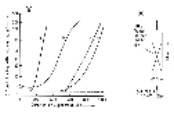

Figure 68. Separation of ca. 1 g/m^3 of toluene from moist gas

A. At inlet
a) Temperature 21°C, toluene 0.9 g/m^3, R. H. 100%, abs. humidity 20 g/m^3; b) Temperature 21°C, toluene 0.8 g/m^3, R. H. 70%, abs. humidity 14 g/m^3; c) Temperature 21°C, toluene 0.9 g/m^3, R. H. 50%, abs. humidity 10 g/m^3; d) Temperature 21°C, toluene 0.9 g/m^3, R. H. 30%, abs. humidity 6 g/m^3 (effluent loading monitored with flame ionization detector)
B. Adsorber which was investigated (adsorption on activated carbon Supersorbon Td IV)

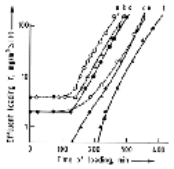

Figure 69. Effluent concentration of toluene as a function of loading time for various states of regeneration

a) 3 kg of steam per kilogram of organic material, temperature of regenerating steam (ϑ) 100°C, without preheating; b) 3 kg/kg, ϑ=150°C, without preheating; c) 3 kg/kg, ϑ=100°C, with preheating; d) 3 kg/kg, ϑ=150°C, with preheating; e) 6 kg/kg, ϑ=150°C, without preheating; f) 6 kg/kg, ϑ=150°C, with preheating

After steaming, care must be taken to dry and cool the bed of activated carbon as much as possible, because this favors complete purification in the subsequent adsorption step. The lower the temperature of the regenerated bed, the better is the purification.

Investigations of the sorption kinetics for toluene and 1-butanol adsorptives at concentrations up to 5 g/m^3 (STP) [160] on activated carbon showed only minor temperature increases within the pellets. Adsorption can therefore be regarded as isothermal. Checks of measured loading curves according to the pore diffusion model (Eq. 6.2) showed that the diffusion coefficients determined from the measured values (i.e., 2.9×10^{-5} m^2/s for toluene and 1.5×10^{-5} m^2/s for 1-butanol) were higher than those during molecular diffusion in air. This suggests that additional transport mechanisms (e.g., mass transport in the sorbate phase) contribute substantially to total mass transport in the particle. The model takes into account the curvature of the adsorption isotherms. Besides macropore diffusion in both the fluid and the adsorbed phase, external mass transfer to the pellet and thus resistance in the outer boundary film, must be taken into account to calculate the measured adsorption curves with sufficient accuracy.

8.7. Adsorptive Desulfurization Processes

Sulfur-containing compounds in synthesis and flue gases, especially hydrogen sulfide, carbon disulfide, carbonyl sulfide, sulfur dioxide, and mercaptans can be removed adsorptively. For synthesis gases, sulfur-containing components must be removed down to the last few parts per million to protect the catalyst.

At higher concentration, hydrogen sulfide can be separated on activated carbon from gases that are completely free of oxygen. Loaded activated carbon is regenerated by temperature swing with a hot gas, the result being a gas rich in hydrogen sulfide.

If the concentration of hydrogen sulfide is low (<50 g/m^3) and the exhaust gas contains oxygen, elemental sulfur separates on activated carbon. Hydrogen sulfide and sulfur dioxide can react at 125 – 150 °C to give elemental sulfur according to the Claus reaction:

$$SO_2 + 2\,H_2S \Leftrightarrow 3\,S + 2\,H_2O$$

Elemental sulfur accumulates on the adsorbent. This sulfur can be recovered in very pure form by hot-gas desorption. The catalytic adsorbents used are aluminum oxide and impregnated, activated carbon. The total conversion of sulfur-containing compounds is between 95 and 98.6 %, depending on the adsorbent; this can be increased to 99.4 % in a two-stage unit. Gases with low concentrations of hydrogen sulfide can be purified at room temperature with activated carbon impregnated with potassium iodide. Thus, hydrogen sulfide can be removed to a residual concentration of less than 1 ppm. Carbon loaded with elemental sulfur, can be regenerated by extraction with carbon disulfide or caustic soda.

For *exhaust air from viscose staple and yarn plants* which may contain hydrogen sulfide and carbon disulfide, adsorption on two layers of activated carbon (Sulfosorbon process) is available (see Fig. 70). The lower layer of wide-pore activated carbon (b) serves for the catalytic conversion of hydrogen sulfide to elemental sulfur according to the reaction

$$2\,H_2S + O_2 \Leftrightarrow 2\,S + 2\,H_2O$$

Carbon disulfide is adsorbed in the upper layer (a) and is displaced from time to time by steam to regenerate the activated carbon. When, after a longer period of time, the lower layer of activated carbon is loaded with sulfur, the latter is removed by extraction with hydrogen sulfide (d). The sulfur – hydrogen sulfide mixture can be separated by distillation (f). The regenerated carbon is separated from the condensate by steaming. Through various purification steps, high purity sulfur and hydrogen sulfide are finally obtained. Off-gases containing small amounts of carbon disulfide, carbonyl sulfide, and the lower mercaptans can be purified by alkalized (e.g., potassium carbonate) activated carbon in the presence of steam and oxygen. First, hydrogen sulfide is formed by hydrolysis; then, it is transformed by oxygen and alkali to the sulfate, which can be leached out with water. After this, the carbon must be reimpregnated.

Figure 70. Schematic of a Sulfosorbon plant

a) Carbon disulfide adsorption; b) Hydrogen sulfide reaction; c) Carbon disulfide desorption; d) Sulfur extraction; e) Separation; f) Distillation; g) Carbon disulfide to extraction

In *flue-gas desulfurization* with activated coke, sulfur dioxide is bound to coke in the form of sulfuric acid, according to the equation [161]:

$$SO_2 + H_2O + 1/2\,O_2 \Leftrightarrow (H_2SO_4)_{ads}$$

A suitable installation is shown in Figure 71. Regenerated activated coke moves through the moving bed adsorber and is loaded with sulfuric acid. Flue gas flows through the bed in cross flow. This results in a moderate pressure drop of the gas due to the relatively short flow length, a desirable situation in view of the large volume streams in conventional power stations. Chemisorption is carried out at ca. 120 °C and at slight excess pressure. At degrees of desulfurization greater than 85 %, sulfur dioxide loading X=0.10 – 0.15 kg/kg is achieved. The activated coke loaded with sulfuric acid is regenerated at about 600 °C by the temperature-swing process, in which coke is heated either by a hot inert gas or by hot sand (750 °C). In the schematic shown in Figure 71, activated coke is heated with an inert gas. Sulfur dioxide is desorbed as the temperature increases according to

$$2H_2SO_4 + C \Leftrightarrow 2\,SO_2 + CO_2 + 2\,H_2O$$

and coke and sand are separated by screening. The regenerated coke is cooled and returned to the adsorber pneumatically. The sand, which serves as a heat carrier, is also recirculated. In addition, loaded coke can be regenerated by washing with water, which yields dilute sulfuric acid.

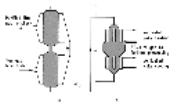

Figure 71. Flow diagram of BF–Uhde-Mitsui process for simultaneous removal of SO_2 and NO_x from flue gases (NO_x is convered catalytically with NH_3 to N_2 and H_2O)

a) Purification of flue gas in the adsorber; b) Regeneration of activated coke

8.8. Adsorptive Purification of Methane

Methane must often be separated from methane-containing gases such as biogas, mine gas, and deposit gas. In addition, components like water, carbon dioxide, hydrogen sulfide, sulfur dioxide, carbonyl sulfide, and mercaptans may have to be removed from natural gas. Pressure-swing processes are particularly well suited for these purposes.

Purification of Biogas, Mine and Deposit Gases. Biogas, mine gas, and deposit gases are mixtures of mainly methane, carbon dioxide, and often air. After preliminary separation of hydrogen sulfide and organic components, carbon dioxide and, to some extent, air can be separated from methane adsorptively [162]. Methane, carbon dioxide, and lower hydrocarbons are adsorbed more strongly than air on carbon molecular sieves (see Fig. 72), and separation is based on the equilibrium effect. Upon desorption, a gas rich in methane is obtained.

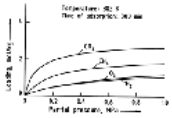

Figure 72. Adsorption isotherms of oxygen, nitrogen, methane, and carbon dioxide on carbon molecular sieve

During *depressurization and purging*, the gas obtained is moderately rich in methane, having a heat of combustion similar to illuminating gas. Upon *evacuation*, a gas is obtained whose methane content is comparable to that of natural gas (> 90 %). The adsorption pressure for the separation process is between 0.2 and 0.8 MPa. The yield of methane is >95 %, with less than 8 % air [163].

These separation processes can also be carried out on zeolitic molecular sieves, silica gel, and activated aluminum oxide [163]. Figure 73 shows adsorption isotherms of oxygen, nitrogen, methane, and carbon dioxide on zeolites. If an adsorbent with a very low micropore width is chosen, small molecules like carbon dioxide, nitrogen, and oxygen are adsorbed much more rapidly than methane. This is based on the steric – kinetic effect. In this way, a methane-rich gas can be obtained in the adsorption step.

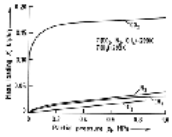

Figure 73. Adsorption isotherms of oxygen, nitrogen, methane and carbon dioxide on molecular sieve 5 A (Linde)

Purification of Natural Gas. Besides its major component methane, natural gas contains materials such as water, carbon dioxide, hydrogen sulfide, carbonyl sulfide, and mercaptans which are frequently removed on molecular sieves. During drying of natural gas, dew points near –100 °C are reached if the water-loaded adsorbent is regenerated thermally. Carbon dioxide is separated on zeolitic molecular sieve 4A or 5A. At high pressure, considerable coadsorption of methane occurs, as shown in Figure 74.

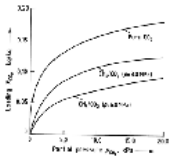

Figure 74. Effect of methane on the change in loading of carbon dioxide on molecular sieve 5 A (*p* is the active pressure)

The breakthrough curves in fixed-bed adsorbers become flatter with increasing pressure in the region between 2 and 8 MPa [164]. This results from decreasing diffusion coefficients, because (see Eq. 6.31) these coefficients are inversely proportional to pressure. The height of the mass-transfer zone (HTZ) and the length of unused bed (LUB) therefore increase linearly with pressure, as shown in Figure 75. Solid lines apply to the ternary mixture hydrogen sulfide – carbon dioxide – methane, while dashed lines refer to the binary mixture hydrogen sulfide – methane.

Experimentally determined breakthrough curves can be checked by calculation according to a model that considers external mass transfer and macropore diffusion only in the gaseous phase. Optimization calculations show that an operating pressure of approximately 6 MPa leads to maximum breakthrough times if only hydrogen sulfide is to be separated from methane to a purity of 5 mg of H_2S per cubic meter [164]. The presence of 1 mol % CO_2 leads to substantially shorter breakthrough times in the ternary than in the binary mixture.

Mercaptans can be separated by using zeolitic molecular sieve 10A with a cage opening of 0.74 nm.

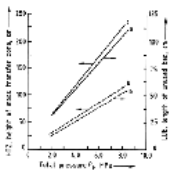

Figure 75. HTZ and LUB of the hydrogen sulfide component as a function of total pressure at different concentrations of carbon dioxide

a) H_2S–CO_2–CH_4, average mole fraction of CO_2 at the inlet y_{0,CO_2}=1 mol%, average mole fraction of H_2S at the inlet y_{0,H_2S}=0.1 mol%, average superficial velocity $\dot{v}$ =0.113 m/s; b) H_2S–CH_4, y_{0,H_2S}=0.1 mol%, $\dot{v}$=0.154 m/s

8.9. Adsorptive Purification of Hydrogen

Hydrogen from the production of synthesis gas contains nitrogen, carbon monoxide, methane, carbonyl sulfide, hydrogen sulfide, and other materials. These components can be removed by adsorption on zeolitic and carbon molecular sieves [148, 165]. Because practically no adsorption of hydrogen occurs, it can be obtained in highly pure form with very few parts per million of impurities. Adsorption is carried out at high pressures between 2 and 5 MPa and at ambient temperature. The adsorbents are regenerated by depressurization to 0.1 MPa and subsequent purging. Several adsorbers are often arranged in parallel to increase the yield of hydrogen [166] because the regeneration processes of depressurization and purging, together with repressurization, take longer than adsorption. Some plants contain as many as twelve adsorbers with mixed beds (see Section Pressure Swing Processes). Depending on the amount of the components to be separated and their affinity for the adsorbent, individual adsorbers are filled with layers of different adsorbents.

Pressure-swing technology is highly economical when the crude gas contains more than 50 % hydrogen. If the gas contains higher hydrocarbons or polymerizable materials, these should be separated beforehand. This will increase the lifetime of the adsorbent.

8.10. Separation of Isomers

Normal (straight-chain) and isoalkanes can be separated by the steric effect [167, 168]. The critical molecular diameter of *n*-alkanes is 0.49 nm, while that of isoalkanes is 0.56 nm. Hence, only *n*-alkane molecules will be adsorbed on zeolitic molecular sieve 5A. The loaded adsorbent can be regenerated by pressure swing and purging or by displacement of the adsorbate with a displacing agent.

Figure 76 illustrates schematically a pressure-swing installation in which molecular sieve 5A is used as adsorbent. The feed mixture, which consists of approximately equal parts of normal and isoalkanes, is evaporated in a circulation reboiler (a) at 580 kPa. It then reaches an adsorber (c) operated at 510 kPa at a temperature of 300 °C. In the

adsorption step, a product highly enriched in isoalkanes can be obtained. During regeneration by pressure swing, the pressure is lowered to 7.0 kPa with the aid of a vacuum pump. Desorption is also carried out at about 300 °C. The desorbate, which is rich in *n*-alkanes, is returned to the adsorber via the isomerizing reactor (f). In this way, normal alkanes can be almost completely converted to isoalkanes in this plant. According to [167], the energy costs of adsorptive separation are less than half those of separation by rectification.

Longer straight-chain alkanes may have a low rate of desorption, so regeneration can require a large consumption of purge gas. Therefore, displacing the component to be desorbed by a suitable material may be more economical [169]. Figure 77 shows an installation for the separation of normal and isoalkanes according to this principle. The feed is supplied to the adsorber (a) on the left. Again, normal alkanes are adsorbed selectively based on the steric effect, while the isoalkane-enriched product reaches a rectification column (b) via a heat exchanger. Before the normal alkanes break through in the adsorption step, the adsorber is regenerated by displacing the C_{10} – C_{20} alkanes with an *n*-pentane – *n*-hexane mixture (shown on the right). The desorbate reaches a second rectification column (c) where the high-boiling C_{10} – C_{20} alkanes leave at the bottom, and the displacement agent consisting of *n*-pentane and *n*-hexane is distilled overhead. In the column (b) on the left, the displacement agent, which is itself displaced by the long-chain normal alkanes in the adsorption step, is separated from the isoalkanes and withdrawn from the bottom.

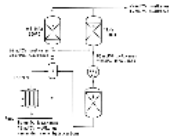

Figure 76. Setup for separating a mixture of normal and isoalkanes by pressure-swing adsorption

a) Circulation reboiler; b) Separator; c) Adsorption; d) Desorption; e) Vacuum compressor; f) Isomerization reactor

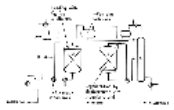

Figure 77. Installation for adsorptive separation of a mixture of normal and isoalkanes

a) Adsorber with 5 A molecular sieve; b) Rectification column; c) Second rectification column

[Top of Page]

9. Liquid-Phase Processes

Adsorption processes in the liquid phase are used to remove impurities from drinking water, wastewater, and aqueous solutions. Beside this, there are major applications in the food and beverage as well as the chemical industry. In principle all the adsorbents (see Chap. Adsorbents) can be used, as with gas-phase processes. However, the regeneration may be different since the adsorbates are usually less volatile.

Organic impurities dissolved in water may be removed on activated carbon. Either carbon pellets (particle diameter 0.5 – 0 mm) or powdered carbon can be used in the suspension process. Powdered carbon can be separated from water by sedimentation, filtration, or centrifugation. The direction of flow in fixed beds is from top to bottom. If the resistance of the bed becomes very high because of dirt deposits, the bed can be cleaned by back purging and loosening of the layer. Loaded carbon is either reactivated or burned. Rotary kilns, moving beds, and fluidized beds are used in thermal reactivation [139]. Such reactors are operated between 700 and 800 °C. in order to destroy all the adsorbate. Carbon pellets to be reactivated are dried between 200 and 300 °C before they reach the reaction section of the regenerator.

9.1. Purification of Drinking Water

In the preparation of drinking water, chlorine, ozone, manganese, and iron can be removed in addition to dissolved organic compounds. Activated chlorine or ozone added as a bactericide reacts on activated carbon to form Cl^- or O_2 after the water has been treated. Dissolved iron(II) or manganese(II) ions are oxidized catalytically on the surface of activated carbon and separated as the insoluble oxide or hydroxide.

All these waste and odour components can be removed by activated carbon [170]. Also fulvic and huminic acids, which are the major components of naturally colored colloids, can be removed from surface waters (lakes, ponds, and reservoirs) effectively [140]. Furthermore, water condensate containing oil can be purified efficiently with activated carbon. An oil content of less than 1 mg per liter of pure condensate can be attained, whereby the carbon takes up 10 – 15 % of its own mass.

9.2. Wastewater Treatment

Wastewater from the chemical industry often contains organic substances that are difficult or impossible to degrade biologically. These include chlorinated and nitrated hydrocarbons, especially aromatic and polycyclic compounds. Such substances can be removed with activated carbon because their isotherms favor adsorption.

Figure 78 shows adsorption isotherms of phenol and 4-nitrophenol [124]. Isotherms of 4-chlorophenol, 2,4-dichlorophenol, benzoic acid, phenylacetic acid, dodecyl benzenesulfonic acid, and their binary mixtures can be found in [125].

The single-component isotherms often obey the Freundlich equation. By using the “ideal dilute solution” theory [171], the adsorption equilibrium for mixtures of substances can be predicted simply from the isotherms of the individual components in the pore liquid.

Figure 79 shows breakthrough curves of 4-nitrophenol on activated carbon for an initial concentration around 0.1 mmol/L. The diffusion coefficients necessary for calculating such curves can be determined experimentally by establishing the decrease in concentration of the solution as a function of time in the presence of activated carbon as it is being loaded [125].

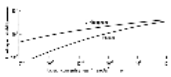

Figure 78. Adsorption isotherms of phenol and 4-nitrophenol on activated carbon at 20°C

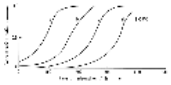

Figure 79. Breakthrough curves for 4-nitrophenol on activated carbon at an initial concentration of ca. 0.1 mmol/L at 20°C

a) *Z*=2 cm; b) *Z*=4 cm; c) *Z*=6 cm; d) *Z*=8 cm

The removal of ionic compounds with activated carbon is not as effective as arsenic removal with activated alumina [59]. Also iron-based adsorption media are reported to be very effective in that respect [37]. Fluoride removal is also accomplished with activated alumina [172] and pesticide removal is either by activated carbon or clay minerals [59, 170]. Ultrahigh-purity water in the semiconductor industry is produced with polymeric resins and activated carbon.

The recovery of gold in the mining industry [170, 173] is a special applications of activated carbon (⟶ Gold, Gold Alloys, and Gold Compounds – Recovery of Gold with Carbon). Nowadays, secondary mining residues with gold on the ppm level are economically processed by adsorbing gold onto activated carbon.

If fixed beds of activated carbon are used to purify wastewater containing several organic substances, the individual components can displace each other, which leads to rollover (see Section Modeling the Isothermal Fixed-Bed Adsorber). After such beds have been operated for some time, the concentration of certain displaced organic substances may be higher in the purified water than in the feedwater.

To describe such fixed bed adsorbers, the balances of various materials in the bed and in the pellet must be formulated, and Equations 6.1-6.4 must be solved.

As examples, Figure 80 shows breakthrough curves of the binary mixture 4-nitrophenol – phenol [125]. A marked rollover of the less strongly bound component, phenol, occurs. Such breakthrough curves can be calculated by using a model that includes transport in both the fluid and the adsorbed phases in the macropores, as well as external mass transfer.

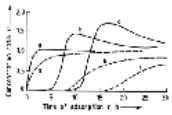

Figure 80. Breakthrough curves for a mixture of 4-nitrophenol (- - -) and phenol (—) on activated carbon at 20°C

a) *Z*=4 cm; b) *Z*=12 cm; c) *Z*=20cm

One of the oldest applications of activated carbon in the food industry is decolorization of sugars. In the 1970s ca. 25 % of activated coal in the USA was consumed in the sugar industry [174]. Wine, sweet oils, and nutrient fats can be decolorized and deodorized with mixtures of activated carbon (up to 20 %) and pillared clays [175]. A similar application is the decolorization of acids. In the case of crude phosphoric acid, activated carbon is used to remove humic acids and related compounds. Citric and lactic acids from fermentation broths are deodorized and decolorized by using activated carbon, as are natural products such as quinine, caffeine, gelatine, and glycerin for use in the food industry. Debittering of citrus juices [176], color control of juices and beer, and final purification of pharmaceuticals, vitamins, and amino acids are further applications of activated carbon in liquid adsorption [174]. Polyphenols, which form complexes with proteins and thus give rise to undesirable dullness in beverages such as beer, wine, and clear fruit juices, can be adsorbed with polymeric adsorbents [177]. MCM-41 has been studied as protein adsorbent in the food industry [178], and sugar separation/removal can be achieved with zeolites. To remove short-chain sugars a specific interaction between the cation in the zeolite and the sugar is necessary. By using a hydrophobic zeolite, long-chain sugars, which are hydrophobic, can be removed [179].

9.3. Separation of Nonaqueous Substances

The recovery of *p*-xylene from a mixed xylene feedstock was performed in the first patented simulating moving bed arrangement of UOP [16] using zeolitic adsorbents. The first commercial Parex unit came on stream 1971, and UOP has licenced 72 units resulting in a nameplate capacity of 13.8×10^6 t/a. The same principle (see Fig. 15) is applied to separate *n*-paraffins from kerosene in the Molex process using molecular sieve adsorbents. Other Sorbex technologies are the separation of ethylbenzene or *m*-xylene from mixed C_8 aromatic isomers, of olefins from paraffins, *p*-cresol or *m*-cresol from other cresol isomers, *p*-cymene or *m*-cymene from other cymene isomers, and fructose from mixed sugars [180].

Organic materials of all kinds, especially solvents and liquefied petroleum gases that contain traces of water, can be dried by adsorption. Zeolites, silica gel, and activated aluminum oxide are used. If the organic liquid contains a polymerizable olefinic component, molecular sieve 3A is employed to prevent polymerization. The residual water content is below 1 ppm. As a rule, the adsorbents are regenerated thermally.

Components such as water, carbon dioxide, hydrogen sulfide, and mercaptans can be removed adsorptively from liquefied natural gas. The main bed uses 4A or 5A zeolites, which adsorb water, carbon dioxide, and hydrogen sulfide. Mercaptans are removed, in a bed connected to the outlet side, by using molecular sieve 10A. If efficient thermal desorption of hydrogen sulfide is desired, a temperature of 250 °C is required.

Mercury is a dangerous contaminant which is often present in natural gas, petrochemicals, and some refinery feed streams. Typical adsorbents used for mercury removal are activated carbon or coke [181]. A newer challenge is the desulfurization of transport fuels, the success of which will depend on the development of a highly selective sorbent with a high sulfur capacity, because zeolites and activated alumina used so far are not promising [182-186]

[Top of Page]

References

1. M. Polanyi: “Über die Adsorption vom Standpunkt des dritten Wärmesatzes”, *Verh. Dtsch. Phys. Ges.* **16** (1914) 1012–1016.
2. I. Langmuir: “The Evaporation, Condensation and Reflection of Molecules and the Mechanism of Adsorption”, *Phys. Rev.* **8** (1916) 149–176.
3. I. Langmuir: “The Adsorption of Gases on Plane Surfaces of Glass, Mica and Platinum”, *J. Am. Chem. Soc.* **40** (1918) 1361–1402. Links
4. S. Brunauer, P. H. Emmet, E. Teller: “Adsorption of Gases in Multimolecular Layers”, *J. Am. Chem. Soc.* **60** (1938) 309–319. Links
5. M. M. Dubinin, V. A. Astakov: “Description of Adsorption Equilibriums of Vapors on Zeolites”, *Adv. Chem.* **102** (1971) 69–85.
6. M. M. Dubinin, L. V. Radushkevich: “The Equation of the Characteristic Curve of Activated Charcoal”, *Dokl. Akad. Nauk SSSR* **55** (1947) 327–329.
7. E. Wicke: “Empirische und theoretische Untersuchungen der Sorptionsgeschwindigkeit von Gasen an porösen Stoffen I, II”, *Koll.-Z.* **86** (1938) 167–186.
8. R. Sips: “On the Structure of a Catalyst Surface I”, *J. Chem. Phys.* **16** (1948) 490–495. Links
9. W. K. Lewis, E. R. Gilliland, B. Chertow, W. H. Hoffmann: “Vapor–Adsorbate Equilibrium. I. Propane–Propylene on Activated Carbon and on Silica Gel”, *J. Am. Chem. Soc.* **43** (1950) 1153–1157.
10. A. L. Myers, J. M. Prausnitz: “Thermodynamics of Mixed-Gas Adsorption”, *AIChE J.* **11** (1965) no. 1, 121–127.
11. Calgon, “*Continuous Ion Exchange and Chromatography Technology*”, 2003, www.ionexchangeseparation.com.
12. J. B. Fox, R. C. Calhoun, W. J. Eglinton: “Continuous Chromatography Apparatus: I. Construction”, *J. Chromatogr. A*, **43** (1969) 48–54. Links
13. H.-J. Bart, R. C. Messenböck, C. H. Byers, A. Prior, J. Wolfgang: “Continuous Chromatographic Separation of Frucose, Mannitol and Sorbitol”, *Chem. Eng. Process.* **35** (1996) 459–471. Links
14. A. Prior, 2003, www.priorsep.com.
15. G. Sorti, M. Mazzotti, M. Morbidelli, S. Carra: “Robust Design of Binary Countercurrent Adsorption Separation Processes”, *AIChE J.* **39** (1993) 471–492.
16. US 2985589, 1961 D. B. Broughton, C. G. Palatine, D. Plaines).
17. R. T. Yang: *Adsorbents. Fundamentals and Applications*, Wiley-Interscience, New York, 2003.
18. R. H. Perry, D. W. Green: *Perry's Chemical Engineer's Handbook*, 7th ed., McGraw-Hill, New York, 1997.
19. DIN-EN, 12901, Anorganische Filterhilfs- und Filtermaterialien–Definitionen, 2000.
20. R. Kümmel, E. Worch: *Adsorption aus wässrigen Lösungen*, 1st ed., VCH, Weinheim, Germany, 1990.
21. H. Sontheimer, B. R. Frick, J. Fettig, G. Hörner, C. Hubele, G. Zimmer: *Adsorptionsverfahren zur Wasserreinigung*, 1st ed., DVGW-Forschungsstelle am Engler-Bunte-Institut der Universität Karlsruhe, Karlsruhe, 1985.
22. W. Kast: *Adsorption aus der Gasphase: ingenieurwissenschaftliche Grundlagen und technische Verfahren*, VCH, Weinheim, Germany, 1988.
23. S. J. Gregg, K. S. W. Sing: *Adsorption, Surface Area and Porosity*, 2nd ed. 1982, Academic Press, London, 1982.
24. DIN 66131, Bestimmung der spezifischen Oberfläche von Feststoffen durch Gasadsorption nach Braunauer, Emmett und Teller (BET), 1991.
25. M. A. Bayati: “Know the Three C's of Carbon”, *Water Technol.* **9** (1993) 35–38.
26. IUPAC, *Manual of Symbols and Terminology for Physicochemical Quantities and Units,* Appendix II, Part I. Pure and Applied Chemistry, Oxford, 1972. p. 576–638.
27. VDI-Richtlinie 3674, Abgasreinigung durch Adsorption, Prozessgas- und Abgasreinigung VDI Verlag, Düsseldorf, 1998.
28. DIN 66133, Bestimmung der Porenvolumenverteilung und der spezifischen Oberfläche von Feststoffen durch Quecksilberintrusion, 1993.
29. DIN 66134, Bestimmung der Porengössenverteilung und der spezifischen Oberfläche mesoporöser Feststoffe durch Stickstoffsorption-Verfahren nach Barret, Joyner und Halenda (BJH), 1998.
30. H. F. Stoeckli, F. Kraehenbuehl, D. Morel: “The Adsorption of Water by Active Carbons in Relation to the Enthalpy of Emmersion”, *Carbon* **21** (1983) 589. Links
31. E. P. Barrett, L. G. Joyner, P. H. Halenda: “The Determination of Pore Volume and Area Distributions in Porous Substances. I. Computations from Nitrogen Isotherms”, *J. Am. Chem. Soc.* **73** (1951) 373–380. Links
32. R. L. Bond: *Porous Carbon Solids*, Academic Press, London, 1967, Chaps. 1, 5, and 6.
33. F. Kreil: *Diffusion und Chemische Reaktionen in der Gas/Feststoff-Katalyse*, Springer, Berlin, 1999.
34. K. P. Goodboy, H. L. Fleming: “Trends in Adsorption with Aluminas”, *Chem. Eng. Prog.* **11** (1984) 63–68.
35. D. Clifford: “Ion exchange and Inorganic Adsorption”, in R. D. Letterman (ed.): *Water Quality and Treatment: A Handbook of Community Water Supplies*, 5th ed., McGraw-Hill, New York, 1999.
36. L. K. Hudson, C. Misra, K. Wefers: “Aluminum Oxide”, *Ullmann*, 6th ed. (electronic release), Wiley-VCH, Weinheim, Germany, 1998.
37. S. A. Gembicki, A. R. Oroskar, J. A. Johnson: “Adsorption, Liquid Separation”, *Kirk-Othmer*, **4** (1991) 573–600.
38. US 3557025, 1971 (R. B. Emerson, W. A. Brian).
39. US 4493715, 1985 (J. P. Hogan, C. R. Rease).
40. C. H. Nelli, G. T. Rochelle: “Nitrogen Dioxide Reaction with Alkaline Solids”, *Ind. Eng. Chem. Res.* **35** (1996) 999–1005. Links
41. R. K. Iler: *The Chemistry of Silica*, John Wiley & Sons, New York, 1979, p. 478.
42. K. K. Unger: *Porous Silicia, Its Properties and Use as a Support in Column Liquid Chromatography*, Elsevier, Amsterdam, 1979.
43. E. F. Vansant, P. van der Voort, K. C. Vrancken: *Characterization and Chemical Modification of the Silica Surface*, Elsevier, Amsterdam, 1995.

44. B. Sellergren: “Enantiomer Separations Using Designed Imprinted Chiral Phases”, in G. Subramanian (ed.): *Chiral Separations Techniques*, 2nd ed., Wiley-VCH, Weinheim, Germany, 2001, pp. 153–186.
45. US 3702886, 1972 (R. J. Argauer, G. R. Landolt).
46. E. M. Flanigen, J. M. Bennett, J. P. Grose, J. P. Cohen, R. L. Patton, R. M. Kirchner, J. V. Smith: “Silicalite, a New Hydrophobic Crystalline Silica Molecular Sieve”, *Nature* **271** (1978) 512–516. Links
47. W. Otten, E. Gail, T. Frey: “Einsatzmöglichkeiten hydrophober Zeolithe in der Adsorptionstechnik”, *Chem. Ing. Tech.* **64** (1992) no. 10, 915–925.
48. U. Eiden, J. Ciprian “Auswahl von Sorbenzien für die Gasphasenadsorption”, *Chem. Ing. Tech.* **65** (1993) no. 11, 1329–1336.
49. E. Roland, P. Kleinschmit: “Zeolites”, *Ullmann*, 6th ed. (electronic release), Wiley-VCH, Weinheim, Germany, 1998.
50. X. Feng, G. E. Fryxell, L.-Q. Wang, A. Y. Kim, J. Liu, K. M. Kemner: “Functionalized Monolayers on Ordered Mesoporous Supports”, *Science*, **276** (1997) 923–926. Links
51. M. S. Baksh, E. S. Kikkinides, R. T. Yang: “Characterization by Physisorption of a New Class of Microporous Adsorbents: Pillared Clays”, *Ind. Eng. Chem. Res.* **31** (1992) 2181–2189. Links
52. P. B. Malla, S. Komarnenei: “Synthesis of Highly Microporous and Hydrophilic Alumina-Pillared Montmorillonite: Water-Sorption Properties”, *Clays Clay Miner.* **38** (1990) 363–372.
53. M. Suzuki: “Activated Carbon Fiber: Fundamentals and Applications”, *Carbon*, **32**(4) (1994) 577–586. Links
54. T. J. Mays: “Activated Carbon Fibers”, in T. D. Burchell (ed.): *Carbon Materials for Advanced Technologies*, Elsevier Science, Oxford, 1999.
55. VDI-Richtlinienentwurf 3452, Auswurfbegrenzung; Viskoseherstellung und -verarbeitung; Schwefelwasserstoff und Schwefelkohlenstoff, VDI Verlag, Düsseldorf, 1977.
56. U. Seifert, J. Guderian, C. Wagner: “Schadensanalytik und Sicherheitsempfehlungen zu Aktivkoks-Adsorbern für die Rauchgas-Endreinigung”, *VGB Kraftwerkstechnik* **2** 1998, 77–80.
57. J. Guderian, U. Seifert, C. Wagner: “Modellgestütze Sicherheitsbetrachtungen zu Aktivkoksadsorbern für die Rauchgasendreinigung in thermischen Abfallbehandlungsanlagen”, *VGB Kraftwerkstechnik* **3** (1999) 68–72.
58. J. L. Humphrey, G. E. I. Keller: *Separation Process Technology*, McGraw-Hill, New York, 1997.
59. S. D. Faust, O. M. Aly: *Adsorption Process for Water Treatment*, Butterworths, Boston, 1987.
60. G. V. Baron: “Industrial Gas Separation Using PSA”, in E. V. Vansan (ed.): *Separation Technology*, Elsevier, Amsterdam, 1994, p. 201.
61. C. W. Jones, W. J. Koros: “Carbon Molecular Sieve Gas Separation Membranes I. Preparation and Characterization Based on Polyimide Precursors”, *Carbon* **32** (1994) 1419–1425. Links
62. C. M. A. Golden, M. B. Rao, S. Sircar in F. Meunier (ed.): *Fundamentals of Adsorption*, Elsevier, Amsterdam, 1998, p. 1083.
63. R. T. Yang: *Gas Separation by Adsorption Processes*, Imperial College Press, London, 1997.
64. S. Iijima: “Helical Microtubes of Graphitic Carbon”, *Nature* **354** (1991) 56–58. Links
65. C. R. Fox: “Industrial Wastewater Control and Recovery of Organic Chemicals by Adsorption”, in F. L. Slejko (ed.): *Adsorption Technology*, Marcel Dekker, New York, Basel, 1985, pp. 167–212.
66. Fox, C. R. and D. C. Kennedy: “Conceptual Design of Adsorption Systems”, in F. L. Slejko (ed.): *Adsorption Technology*, Marcel Dekker, New York, Basel, 1985, pp. 91–165.
67. T. Augustin: “Neue Entwicklungen von Adsorberharzen und deren Einsatzgebiete für die Abreicherung von organischen Stoffen aus gewerblichen und industriellen Abfällen”, *Gewässerschutz Wasser Abwasser* **147** (1994) 1–14.
68. H. Y. Huang, J. Padin, R. T. Yang: “Comparison of π-Complexation of Ethylene and Carbon Monoxide with Cu^+ and Ag^+”, *Ind. Eng. Chem. Res.* **38** (1999) 2720–2725. Links
69. H.-J. Bart: *Reactive Extraction*, Springer, Berlin, 2001.
70. H.-J. Bart, G. Stevens: “Reactive Solvent Extraction”, in M. Kertes, A. K. Sengupta (eds.): *Ion Exchange and Solvent Extraction*, Marcel Dekker, New York, 2004, **vol. 17**, pp. 37–82.
71. M. Traving, H. -J. Bart: “Recovery of Organic Acids Using Ion-Exchanger-Impregnated Resins”, *Chem. Eng. Technol.* **25** (2002) 997–1003.
72. A. Warshawsky: “Extraction with Solvent-Impregnated Resins”, in J. A. Marinsky, Y. Marcus (eds.): *Ion Exchange and Solvent Extraction*, Marcel Dekker, New York, 1981, pp. 229–281.
73. A. Warshawsky, J. L. Cortina, M. Aguilar, K. Jerabek: “New Developments in Solvent Impregnated Resins. An Overview”, in M. Cox, M. Hidalgo, M. Valiente (eds.): *Solvent Extraction for the 21st Century*, Society of Chemical Industry (SCI), London, 2001, pp. 1267–1272.
74. A. Kostova, H.-J. Bart: “Reactive Sorption of Amino Acids”, in T. Sapundzhiev and R. Popov (eds): 10th Int. Summer School Chem. Eng, 24–31.5.04, Varna, Bulgaria, Bulgarian Academy of Science Sofia, pp. 230–233.
75. J. Toth: “State Equations of the Solid-Gas Interface Layers”, *Acta Chim. Acad. Sci. Hung.* **69** (1971) 311–328.
76. R. Sips: “On the Structure of a Catalyst Surface II”, *J. Chem. Phys.* **18** (1950) 1024–1026. Links
77. D. M. Ruthven: “Simple Theoretical Adsorption Isotherm for Zeolites”, *Nature Phys. Sci.* **232** (1971) no. 29, 70–71.
78. U. v. Gemmingen: “A new Approach to Adsorption Isotherms”, *Gas Sep. Purif.* **7** (1993) 175–181.
79. U. v. Gemmingen: “Adsorptive Gasphasengleichgewichte bei kleinem Bedeckungsgrad”, *Chem. Eng. Technol.* **73** (2001) 352–357.
80. S. Sircar, A. L. Myers: “Characteristic Adsorption Isotherm for Adsorption of Vapors on Heterogeneous Adsorbents”, *AIChE J.* **32** (1986) 650–656.
81. H. Jüntgen: “Grundlagen der Adsorption”, *Staub Reinhalt. Luft* **36** (7) (1976) 281–288.
82. H. Seewald, H. Jüntgen: “Kinetik der nicht-isothermen Desorption organischer Stoffe von Aktivkohlen”, *Ber. Bunsenges. Phys. Chem.* **81** (1977) 638–645.
83. A. Mersmann, U. v. Gemmingen, W. Sievers, P. Schweighard: “Adsorption” in *Verfahrenstechnische Berechnungsmethoden*, Deutscher Verlag für Grundstoffindustrie, Stuttgart, 1996, Vol. 6.
84. L. Szepesy, V. Illes: “Adsorption of Gases and Gas Mixtures, II. Measurement of the Adsorption Isotherms of Gases on Active Carbon under Pressures of 1 to 7 atm”, *Acta Chim. Hung.* **35** (1963) 53–60.
85. H. Seewald: “Technisch verfügbare Adsorbentien”, *Haus Tech. Vortragsveröff.* **404** (1977) 24–34.
86. H. Dreher, *“Gleichgewicht und Kinetik bei der Adsorption mehrerer Komponenten aus der Gasphase in einem isothermen Adsorber am Beispiel der Systeme Wasser–Benzol–Methanol–Silicagel*”, Dissertation, TH Darmstadt, 1979.
87. E. C. Markham, A. F. Benton: “The Adsorption of Gas Mixtures by Silica”, *J. Am. Chem. Soc.* **53** (1931) 497–507. Links
88. D. M. Ruthven: “Multicomponent Sorption Equilibrium in Molecular Sieve Zeolites”, *Chem. Eng. Sci.* **28** (1973) 701–709. Links
89. D. M. Ruthven: “Generalized Statistical Model for the Prediction of Binary Adsorption Equilibria in Zeolites”, *Ind. Eng. Chem. Fundam.* **24** (1985) 27–32.
90. B. Smit, R. Krishna: “Simulating Adsorption of Alkanes in Zeolites”, in S. M. Raunbach, K. A. Carrado, P. K. Dutta (eds.): *Handbook of Zeolite Science and Technology*, Marcel Dekker, New York, 2003, pp. 317–330.
91. R. Krishna: “Reactive Separations, Teaching an Old Dog New Tricks” in *Int. Max-Planck Symp. Integrated Chemical Processes*, 22.–24.3.2004, Magdeburg.
92. O. Levenspiel: *Chemical Reaction Engineering*, 2nd ed., Wiley, New York 1972.
93. O. Levenspiel: *The Chemical Reactor Omnibook*, Corvallis, Oregon, 1989.
94. H. S. Fogler: *Elements of Chemical Reaction Engineering*, Prentice Hall, Engelwood Cliffs, 1992.
95. G. F. Froment, K. B. Bischoff: *Chemical Reactor Analysis and Design* Wiley, New York, 1979.
96. H. H. Lee: *Heterogeneous Reactor Design*, Butterworth, Boston, 1985.
97. M. Baerns, H. Hofmann, A. Renken: *Chemische Reaktionstechnik*, Thieme, Stuttgart, 1987.
98. K. M. Waananen, J. B. Lichfield, M. R. Okos: “Classification of Drying Models for Porous Solids”, *Drying Technol.* **11** (1993) 1–40.
99. O. Krischer, W. Kast: *Die wissenschaftlichen Grundlagen der Trocknungstechnik*, Springer, Berlin, 1978.
100. W. Nernst: “Theorie der Reaktionsgeschwindigkeit in heterogenen Systemen”, *Z. Phys. Chem.* **23** (1897) 52–55.
101. R. B. Bird, W. Stewart, E. N. Lightfoot: *Transport Phenomena*, Wiley, New York, 1988.
102. P. C. Carman: *Flow of Gases through Porous Media*, Academic Press, New York, 1956.
103. D. Paschek, R. Krishna: “Diffusion of Binary Mixtures in Zeolites: Kinetik Monte Carlo vs Molecular Dynamics Simulations”, *Langmuir* **17** (2001) 247–254. Links
104. E. Buckingham: “Studies on the Movement of Soil Moisture”, *U.S. Dept. Agr. Bureau Soils Bul.* **38** (1907).
105. W. Büssing, H.-J. Bart, R. Germerdonk: “Isothermal Liquid Transport in Porous Media—Capillary Hysteresis Effects”, *Int. J. Heat Mass Transfer* **39** (1996) 1925–1934.
106. B. Steinweg: *Über den Stofftransport bei der Ad- und Desorption von Schadstoffen an/von Aktivkohle im Bereich umweltrelevanter Gaskonzentrationen*, Shaker Verlag, Aachen, 1996.
107. W. Peters, H. Jüntgen: “Einfluß der Hohlraumstruktur auf Gas/Feststoff-Reaktionen an Steinkohle”, *Chem. Ing. Tech.* **40** (1968) 1039–1044.

108. R. Jackson: *Transport in Porous Catalyst*, Elsevier, Amsterdam, 1977.
109. R. E. Cunningham, R. J. J. Williams: *Diffusion in Gases and Porous Media*, Plenum Press, New York, 1980.
110. E. A. Mason: *Gas Transport in Porous Media: The Dusty Gas Model*, Elsevier, Amsterdam, 1983.
111. E. Glueckauf: "Theory of Chromatography, Part 10. Formulae for Diffusion into Spheres and their Application to Chromatography", *Trans. Faraday Soc.* **51** (1955) 1540–1551.
112. A. P. Damaso, "*Über die Adsorption im isothermen Festbett am Beispiel der Systeme Wasser/Aktivtonerde und Tetrachlorkohlenstoff/Aktivtonerde*", Dissertation, TH Darmstadt, 1976.
113. R. Erpelding, H.-J. Bart: "Modelling and Pilot-Scale Experiments of the Steam Desorption Process in a with Organic Solvents Loaded Activated Carbon Column", in F. Meunier (ed.): *Fundamentals of Adsorption*, Elsevier, Paris, 1998, pp. 1029–1034.
114. G. Hörner, "*Adsorption von Chlorkohlenwasserstoffen an Aktivkohlen und deren Regeneration durch thermische Desorption*", Dissertation, Universität Karlsruhe, 1985.
115. J. A. Wesselingh, R. Krishna: *Mass Transfer in Multicomponent Mixtures*, Delft University Press, Delft, 2000.
116. R. Krishna, J. A. Wesseling: "The Maxwell-Stefan Approach to Mass Transfer", *Chem. Eng. Sci.* **52** (1997) 861–911. Links
117. H. R. C. Pratt, G. W. Stevens: "Axial Dispersion", in J. D. Thornton, (ed.): *Science and Practice of Liquid-Liquid Extractions*, Clarendon Press, Oxford, 1992, pp. 416–491.
118. J. Gu: *Simulation eines periodischen Adsorptionsverhaltens*, Shaker Verlag, Aachen, 2000.
119. J. C. Schadl: "Vergleich der Kinetik der Ad- und Desorption am Einzelpellet und im adiabaten Festbett am Beispiel von H_2O und CO_2 an Molekularsieb 5A", Dissertation, TU München, 1986.
120. C. Y. Pan, P. D. Basmadjian: "An Analysis of Adiabatic Sorption of Single Solutes in Fixed Beds: Pure Thermal Wave Formation and its Practical Implications", *Chem. Eng. Sci.* **25** (1970) 1653–1664. Links
121. F. Jokisch, "*Über den Stofftransport im hygroskopischen Feuchtebereich kapillarporöser Stoffe am Beispiel des Wasserdampftransports in technischen Adsorbentien*", Dissertation, TH Darmstadt, 1975.
122. E. M. Reed, J. B. Butt: "Surface Diffusion of Single Sorbates at Low and Intermediate Surface Coverage", *J. Phys. Chem.* **75** (1971) no.1, 133–141.
123. K. Knoblauch, H. Jüntgen, W. Peters: "Adsorption in Wasser gelöster Substanzen am Aktivkoksen, Teil II, Kinetische Messungen im Differentialreaktor bei konstanter Außenkonzentration", *Chem. Ing. Tech.* **41** (1969) no. 14, 789–805.
124. W. Fritz, "*Konkurrierende Adsorption von zwei organischen Wasserinhaltstoffen auf Aktivkohlekörnern*", Dissertation, TH Karlsruhe, 1978.
125. W. Merk, "*Konkurrierende Adsorption verschiedener organischer Wasserinhaltstoffe in Aktivkohlekörnern*", Dissertation, TH Karlsruhe, 1978.
126. W. M. Polte: "Zur Ad- und Desorptionskinetik in bidispersen Adsorbentien", Dissertation, TU München, 1987.
127. J. B. Rosen: "General Numerical Solution for Solid Diffusion in Fixed Beds", *Ind. Eng. Chem.* **46**(8) (1954) 1590–1594.
128. C. Y. Wen, L. T. Fan: *Models for Flow Systems and Chemical Reactors*, Marcel Dekker, New York, 1975.
129. A. Mersmann: *Stoffübertragung*, Springer Verlag, Berlin, 1986.
130. C. N. Satterfield: *Mass Transfer in Heterogeneous Catalysts*, MIT Press, Cambridge, 1970.
131. W. Otten, W. Kast: "Ein vereinfachtes Modell zur Beschreibung nichtisothermer Sorptionsvorgänge am Einzelkorn als Grundlage der Berechnung technischer Adsorber", *Chem. Ing. Tech.* **58** (1986) no. 11, 910–911.
132. U. Münstermann, "*Adsorption von CO_2 aus Reingas und aus Luft am Molekularsieb-Einzelkorn und im Festbett. Berücksichtigung von Wärmeeffekten und H_2O-Präadsorption*", Dissertation, Technische Universität München, 1984.
133. D. M. Ruthven: *Principles of Adsorption and Adsorption Processes*, J. Wiley & Sons, New York, 1984.
134. M. Figueira, "*Über die Wasserdampfregeneration von Aktivkohle*", Dissertation, Universität Kaiserslautern, 1987.
135. J. Küntzel, R. Ham, T. Melin: "Regeneration hydrophober Zeolithe mittels Wasserdampf", *Chem. Ing. Tech.* **71** (1999) 508–511.
136. H. J. Zumkeller, H. J. Bart: "Einfluß des Vorheizverfahrens auf die Aktivkohlebeladung bie der Wasserdampfdesorption", *Chem. Ing. Tech.* **68** (1996) no. 8, 946–949.
137. R. Erpelding: *Untersuchung der Wasserdesorption in halbtechnischen Aktivkohleschüttungen*, Shaker Verlag, Aachen, 1996.
138. G. Börger: "Emissionen am Aktivkohle-Adsorbern durch verbesserte Dampf-Regeneration", *Chem. Ing. Tech.* **64** (1992) 200–201.
139. W. Poggenburg: "Einsatz von Aktivkohle in Trinkwasserwerken-Verfahren der Adsorption und Reaktivierung", *Haus Tech. Vortragsveröff.* **404** (1977) 50–59.
140. *APEC Virtual Center for Environmental Technology Exchange*, 2003, www.apec-vc.or.jp.
141. R. J. Martin, W. J. Ng: "Chemical Regeneration of Exhausted Activated Carbon II", *Water Res.* **19** (1985) 1527–1535. Links
142. H. S. Mc Laughlin: "Regenerate Activated Carbon using Organic Solvents", *Chem. Eng. Prog.* **91** (1995) 45–53. Links
143. V. V. Podlesnyuk, R. M. Marutovsky, L. A. Savchina, K. H. Radeke: "Einfluss von organischen Salzen auf die Adsorption organischer Schadstoffe am Adsorberpolymer Wofatit EP 63", *Chem. Tech.* **46** (1994) 343–345.
144. V. F. Sergeeva: "Salting-in and Salting-out of Non-Electrolytes", *Russ. Chem. Rev.* **34** (1965) 309–318.
145. A. Mersmann, U. Münstermann, J. C. Schadl: "Trennung von Gasgemischen durch Adsorption", *Chem. Ing. Tech.* **55** (1983) no. 6, 446–458.
146. D. Tondeur, P. C. Wankat: "Gas Purification by Pressure Swing Adsorption", *Sep. Purif. Methods* **14** (1985) 157–212.
147. K. Weaber, C. E. Hamrin: "Separation of Hydrogen Isotopes by Heatless Adsorption", *Chem. Eng. Sci.* **29** (1974) 1873–1882.
148. E. Richter, K. -B. Harder, K. Knoblauch, H. Jüntgen: "Neue Entwicklungen zur Druckwechsel-Adsorption", *Chem. Ing. Tech.* **56** (1984) no. 9, 684–691.
149. K. Knoblauch: "Inertgaserzeugung aus Luft unter Einsatz von Kohlenstoffmolekularsieben", *Wärme Gas Int.* **32** (1983) 78–80.
150. U. Harms, W. Höse, W. Fetkenheuer: "Erzeugung kontrollierter Atmosphären mit zeolithischen Molekularsieben", *Chem. Tech. Leipzig* **33** (1981) 427–429.
151. P. M. Turnock, R. M. Kadlek: "Separation of Nitrogen and Methane via Periodic Adsorption", *AIChE J.* **17** (1971) 335–342.
152. R. L. Jones, G. E. Keller: "Pressure Swing Parameter Pumping - A New Adsorption Process", *J. Sep. Process Technol.* **2** (1981) no. 3, 17–23.
153. E. Lassmann, C. Benkmann: "Wirtschaftliche Ozonerzeugung aus Sauerstoff durch die Linde-Druckwechseladsorptionsanlage", *Linde Ber. Tech. Wiss.* **49** (1981) 9–13.
154. H. H. Lee, D. E. Stahl: "Oxygen Rich Gas from Air by Pressure Swing Adsorption Processes", *AIChe Symp. Ser.* **134** (1973) 1–8.
155. K. Knoblauch, H. Jüntgen, W. Peters: "Gastrennung mit Kohlenstoffmolekularsieben", *Erdöl Kohle Erdgas Petrochem.* **32** (1979) no. 12, 551–556.
156. H. J. Schröter: "Adsorptive Abscheidung radioaktiver Abgase", *Haus Tech. Vortragsveröff.* **404** (1977) 119–124.
157. Technische Anleitung zur Reinhaltung der Luft, Carl Heymanns, Köln, 1986.
158. G. Börger, A. Jonas: "Bessere Abluftreinigung in Aktivkohle durch weniger Wasser", *Chem. Ing. Tech.* **58** (1986) no. 7, 610–611.
159. M. E. Minelli-Figueira de Greco, R. Gemerdonk: "Versuche zum Regenerieren von Aktivkohle mit Damf zum Entfernen besonders gefährlicher/belästigender Stoffe aus merklich <20 mg/m^3_N Abluft", *Chem. Ing. Tech.* **58** (1986) no. 4, 336–337.
160. S. Ripperger, R. Gemerdonk: "Kinetik der Adsorption organischer Stoffe an umströmten Aktivkohle-Körnern", *Chem. Ing. Tech.* **56** (1984) no. 2, 131–132.
161. E. Richter, K. Knoblauch, H. Jüntgen: "Adsorptionsverfahren zur Abscheidung von SO_2 und NO_x aus Feuerungsabgasen", *Glückauf* **121** (1985) 1561–1566.
162. E. Richter, K.-D. Henning: "Erdgasgewinnung aus Deponie- und Faulgasen", *WLB Wasser Luft Betr.* **11/12** (1984) 45–48.
163. E. Richter, K. Knoblauch, H. Jüntgen: "Möglichkeiten zur adsorptiven Methan-/Kohlendioxid-Trennung aus Biogasen", *GWF + Gas Wasserfach Erdgas* **126** (1985) 10–14.
164. G. Groningen, K. Hedden, B. R. Rao: "Kinetik der adsorptiven Entfernung von Verunreinigungen aus Gasen bei hohen Drücken", *Chem. Ing. Tech.* **57** (1985) no. 9, 794–795.
165. L. Lancelin, E. Rudelstorfer, M. Schöller, "*Wasserstoffreinigung mit Druckwechseladsorption*, Paper at Presentation at Association Française des Techniciens du Petrole, Paris, 1978.
166. D. Werner: "Gasreinigungsverfahren für große Wasserstoff-Mengen", *Chem. Ing. Tech.* **53** (1981) no 2., 73–81.
167. G. J. Griesmer, H. B. Rhodes, K. Kiyonaga: "Union Carbide-Molekularsieb-Verfahren für die Trennung von Kohlenwasserstoffen", *Erdöl Kohle* **13** (1960) 650–656.
168. W. C. J. Quick, J. Collins: "Isomerisierung von leichten Kohlenwasserstoffen und anschliessender Trennung mit Molekularsieben", *Erdöl Kohle Erdgas Petrochem.* **25** (1972) 706–711.
169. E. Ruhl: "Temperaturwechsel-, Drusckwechsel- und Verdrängungssorptions- Verfahren", *Chem. Ing. Tech.* **43** (1971) no. 15, 870–876.
170. CPL Carbon Link, 2003, http://www.activated-carbon.com.
171. C. J. Radke, J. M. Prausnitz: "Thermodynamics of Multi-Solute Adsorption from Dilute Liquid Solutions", *AIChE J.* **18** (1972) 761–768.
172. UNICEF, "*Fluoride in Water. An Overview*", 2003, www.unicef.org.
173. S. Ubaldini, R. Massidda, C. Abbruzzese, F. Veglio: "A Cheap Process for Gold Recovery from Leached Solutions," *6th Italian Conference on Chemical and Process Engineering*, 8.–11.6.03, Pisa, Italy.

174. H. von Kienle, E. Bäder: *Aktivkohle und ihre industrielle Anwendung*, F. Enke Verlag, Stuttgart, 1980.
175. G. Biernoth, H. E. Rost: "Occurence of Polycyclic, Aromatic Hydrocarbons in Edible Oils and their Removal", *Arch. Hydr.* **152** (1968) 238.
176. P. E. Shaw: "Citrus Juice Debittering—Current Status Worldwide", *Citrus Ind.* **71** (1990) no. 6, 54–55.
177. *Divergan F. Speciality for Stabilizing Beer*, 1995, www.basf.com.
178. J. M. Kisler, G. W. Stevens, A. J. O'Connor: "Adsorption of Proteins on Mesoporous Molecular Sieves", *Mater. Phys. Mech.* **4** (2001) 89–93.
179. *Adsorptive Separation and Removal of Sugars*, www.mep.tno.nl.
180. UOP, 2003, www.uop.com.
181. G. Corvini, J. Stilner, K. Clark, *Mercury Removal from Natural Gas and Liquid Streams*, 2002, www.uop.com.
182. A. Takahashi, F. H. Yang, R. T. Yang: "New Sorbents for Desulfurization by pi-Complexation: Thiophene/Benzene Adsorption", *Ind. Eng. Chem. Res.* **41** (2002) 2487. Links
183. US 5730860, 1998 (R. L. Irvine).
184. A. B. S. H. Salem: "Naphta Desulfurization by Adsorption", *Ind. Eng. Chem. Res.* **33** (1994) 336–340. Links
185. US 4188285, 1980 (M. J. Michlmayr).
186. A. J. Hernández-Maldonado, R. T. Yang: "Desulfurization of Liquid Fuels by Adsorption via pi-Complexation with Cu(I)-Y and Ag-Y Zeolites", *Ind. Eng. Chem. Res.* **42** (2003) 123–129.

[Top of Page]

Symbols

***a*:** exponent in Freundlich equation

***a*:** specific internal surface of adsorbent, m^2/kg

***a*:** thermal diffusivity, m^2/s

a_{sch}: volume-related surface area of packing, m^2/m^3

***A*:** factor in Freundlich equation, mol/g

***A*:** cross-sectional area, m^2

***b*:** factor in Langmuir and BET equations

c_B: bulk concentration, kg/m^3, mol/m^3

c_S: surface concentration, kg/m^3, mol/m^3

c_i: concentration of component *i*, kg/m^3, mol/m^3

c_i^*: concentration of component *i* at equilibrium, kg/m^3, mol/m^3

$c_{M,i}$: concentration of component *i* in the macropore, kg/m^3, mol/m^3

c_p, c_s: specific heat capacity of fluid or adsorbent, $J\ kg^{-1}\ K^{-1}$

c_0, *i*: concentration of component *i* at entrance to fixed bed, kg/m^3, mol/m^3

***C*:** structural parameter in Dubinin–Radushkevich equation

***d*:** gas-kinetic molecular diameter, m

***d*:** particle diameter (= $2\ R_a$), m

d_A: gas-kinetic molecular diameter of component A, m

d_B: gas-kinetic molecular diameter of fluid B, m

d_M: macropore diameter, m

***D*:** bed diameter, m

D_{ax}: axial dispersion coefficient, m^2/s

D_{AB}: diffusion coefficient of component A in fluid B, m^2/s

D_{eff}: "effective" pseudomacropore diffusion coefficient, m^2/s

D_{Kn}: Knudsen diffusion coefficient, m^2/s

D_m: diffusion coefficient in micropores, m^2/s

D_M: diffusion coefficient in macropores, m^2/s

D_0^0: basic micropore diffusion coefficient, m^2/s

D_S: surface diffusion coefficient, m^2/s

***e*:** exponent in Dubinin - Radushkevich equation

***E*:** energy, J

E_a: activation energy, J/mol

$F = \dot{v}\sqrt{\rho}$: F factor, $(\mathrm{m/s})\sqrt{\mathrm{kg/m^3}}$

ΔG: molar Gibbs free energy, J/mol

Δh: specific heat of adsorption, J/kg

Δh_B: specific heat of binding, J/kg

Δh_{LG}: specific heat of evaporation, J/kg

ΔH: molar heat of adsorption, J/mol

ΔH_B: molar heat of binding, J/mol

ΔH_{LG}: molar heat of evaporation, J/mol

h_i: Henry coefficient, mol m^2 g^{-1} N^{-1}

k: Boltzmann constant ($k = 1.38 \times 10^{-23}$ J/K)

k: factor in binary Dubinin equation

k: number of adsorbable components

k_{ad}: rate constant for adsorption, g mol^{-1} s^{-1}

k_{des}: rate constant for desorption, s^{-1}

L: liquid mass, kg

m: mass of one molecule, kg

m_A: mass of molecule of component A, kg

m_B: mass of molecule of component B, kg

m_P: mass of dry adsorbent, kg

M: molar mass, kg/kmol

n: molar flux density, mol m^{-2} s^{-1}

n_A: molar flux density of component A, mol m^{-2} s^{-1}

n_B: molar flux density of fluid B, mol m^{-2} s^{-1}

n_i: molar loading of component *i*, mol/g, cm^3 (STP)/g

$n_{i,\,mon}$: monomolecular loading mol/g, cm^3 (STP)/g

$n_{i,\,s}$: molar loading at complete pore filling, mol/g, cm^3 (STP)/g

n_i^{∞}: molar loading at saturation, mol/g, cm^3 (STP)/g

n^0: molar loading of a pure component, mol/g

N: number of molecular layers

p: active pressure, N/m^2

P: total pressure, N/m^2

p_i: partial pressure of component *i*, N/m^2

$p_{i,}$: saturation pressure of component *i* , N/m^2

$(p_{i,}\ s)_r$: saturation pressure of component i at a curved surface, N/m^2

p_0: partial pressure of a pure component, N/m^2

p_t: total pressure, N/m^2

q: volumetric loading of adsorbent, mol/m^3

q: heat flux density, W/K

Q: integral heat of adsorption, J

r: radial distance in microcrystal, m

r_a: external radius of microcrystal, m

r_o: radius of curvature of liquid surface, m

r_p: pore radius, m

R: radial distance in adsorbent particle, m

R: special gas constant, J kg^{-1} K^{-1}

***R*:** universal gas constant (R = 8.31 J mol^{-1} K^{-1})

R_a**:** external particle radius, m

***S*:** solid mass of adsorbent, kg

***S*:** selectivity

***t*:** time, s

***T*:** absolute temperature, K

$u\left(\frac{c}{c_0}, \frac{y}{y_0}\right)$**:** concentration ratio

***U*:** molar internal energy, J/mol

ΔU**:** isochoric heat of adsorption, J/mol

$\dot{v}$**:** volume flux density of fluid, m/s

***V*:** specific adsorbed (liquid) volume, cm^3/g

v_m**:** molar volume, m^3/mol

$\dot{V}$**:** volume flux of fluid, m^3/s

V_c**:** total column volume, m^3

V_m**:** volume of solid phase, m^3

V_s**:** total pore volume, cm^3/g

***x*:** mole fraction of sorbate, mol/mol

***X*:** mass loading of adsorbent, g/g

$X_{m,i}$**:** mass loading of a zeolite crystal with component *i*, g/g

X_i^***:** equilibrium loading of component *i*, g/g

$\overline{X}$**:** mean mass loading of adsorbent, g/g

$\overline{X}_M$**:** mean mass loading of adsorbent in macropores, g/g

***y*:** mole or mass fraction of fluid, mol/mol, g/g

Y_0**:** mass loading of fluid at entrance to fixed bed, g/g

Y^***:** mass loading of fluid at equilibrium, g/g

***Y*:** molar or mass loading of fluid, mol/mol g/g

***z*:** axial distance, m

***Z*:** length of adsorber bed, m

Greek Letters

α**:** heat-transfer coefficient, W m^{-2} K^{-1}

$\overline{\alpha}$**:** total heat-transfer coefficient, W m^{-2} K^{-1}

α_{iv}**:** parameter of nonideality

β**:** affinity coefficient, J/mol

β**:** mass-transfer coefficient, m/s

γ**:** activity coefficient

ϵ**:** adsorption potential, J/mol

ϵ_B**:** voidage

ϵ_p**:** particle porosity

ϵ_t**:** total porosity

η**:** dynamic viscosity, kg m^{-1} s^{-1}, mPa s

θ**:** degree of coverage

θ**:** temperature, °C

λ**:** thermal conductivity, W m^{-1} K^{-1}

Λ**:** free path of molecules, m

μ_p**:** tortuosity factor

ν**:** kinematic viscosity, m^2/s

ξ**:** drag number

π: spreading pressure, J/m^2

π': reduced spreading pressure, cm^3/g

ρ: true density, kg/m^3

ρ_B: bed density, kg/m^3

ρ_p: particle density, kg/m^3

ρ_s: solid density, kg/m^3

σ: surface tension, J/m^2

σ: collision diameter, m

φ: relative saturation

Φ: surface adsorption saturation pressure

Ψ: association parameter

Ψ: porosity of bed, m^3/m^3

τ: time of adsorption, h

Characteristic Numbers

$Bi = \frac{\alpha d}{6\lambda_s}$: Biot number of heat transfer

$Bi' = \frac{R_a \beta \mu p}{3\epsilon_p D_M}$: Biot number of mass transfer

$Fo = \frac{Dt}{r_p^2}$: Fourier number

$K = \frac{\dot{v} Z}{\psi D_{ax}}$: dimensionless number

$Kn = \frac{\Lambda}{2r}$: Knudsen number

$Nu = \frac{\alpha d}{\lambda_s}$: Nusselt number

$u_b = 2 + \sqrt{Nu_{\text{lam}}^2 + Nu_{\text{turb}}^2}$: Nusselt number of single bed

$Nu_{\text{lam}} = 0.664\sqrt{Re_\psi} Pr^{1/3}$: Nusselt number of laminar flow

$Nu_{\text{turb}} = \frac{0.037 Re_\psi^{0.8} Pr}{1 + 2.44 Re_\psi^{-0.1}\left(Pr^{2/3} - 1\right)}$: Nusselt number of turbulent flow

$Pe = \frac{\dot{v} 2R_a}{\psi D_{ax}}$: Péclet number of particle

$Pr = \frac{\nu}{a}$: Prandtl number

$Re_\psi = \frac{\dot{v} 2R_a}{\psi \nu}$: Reynolds number of particle

$Sc = \frac{\nu}{D_{AB}}$: Schmidt number of fluid

$Sh_{fb} = \frac{2R_a \beta}{D_{AB}}$: Sherwood number of fixed bed

$Sh_{\text{lam}} = 0.664\sqrt{Sh_{\text{lam}}^2 + Sh_{\text{turb}}^2}$: Sherwood number of single bead

$Sh_{\text{lam}} = 0.664\sqrt{Re_\psi} Sc^{1/3}$: Sherwood number of laminar flow

$Sh_{\text{turb}} = \frac{0.037 Re_\psi^{0.8} Sc}{1 + 2.44 Re_\psi^{-0.1}\left(Sc^{2/3} - 1\right)}$: Sherwood number of turbulent flow

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Process-Scale Chromatography

Standard Article

Günter Jagschies[1]

[1]Pharmacia LKB GmbH, Freiburg, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b03_10
Article Online Posting Date: June 15, 2000

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Chromatographic Techniques
2.1. Adsorptive Methods
2.1.1. Ion-Exchange Chromatography
2.1.2. Affinity Chromatography
2.1.3. Hydrophobic Interaction Chromatography
2.1.4. Reversed Phase Chromatography
2.2. Gel Filtration
3. Downstream Strategy
3.1. Number of Preparative Steps
3.2. Preparation of an Optimal Raw Material
3.3. Development of a Chromatographic Purification System
3.3.1. Screening of Chromatographic Methods (Method Scouting)
3.3.1.1. The Gel Bank
3.3.1.2. Experimental Data
3.3.2. Logical Sequence of Chromatographic Steps
3.3.2.1. Selection Criteria
3.3.2.2. The Starting Step
3.3.2.3. Subsequent Steps
4. Optimization
4.1. Purpose of Optimization
4.2. Optimization Parameters
4.2.1. Resolution
4.2.2. Throughput
4.3. Optimization of Ion-Exchange Chromatography
4.4. Optimization of Affinity Chromatography
4.5. Optimization of Hydrophobic Interaction Chromatography
4.6. Optimization of Gel Filtration
5. Scaling Up
5.1. Transfer Protocol
5.2. Practical Examples and Case Studies
5.3. Dimensioning of a Purification Cycle
6. Process-Scale Chromatography Plants
6.1. Basic Design Requirements
6.2. Plant Design
6.2.1. The Chromatography Column
6.2.2. Process Monitoring
6.2.3. Pumps and Valves

6.2.4. Automation and Safety Devices
6.3. Effects of Apparatus on Separation
7. Quality Assurance
8. Developments

[Top of Page]

1. Introduction

The extraction of a protein or peptide from biological raw materials, i.e., *downstream processing*, can be divided into two main steps: *isolation* and *purification*. In the isolation step, a crude product is obtained by techniques such as cell digestion, microfiltration, ultrafiltration, precipitation, or adsorption. The objective is to obtain a fully “clarified” material from which all particulate components have been removed (⟶ Biochemical Separations).

This article is concerned with the second step, i.e., the purification of biomolecules from crude material by means of high-resolution chromatography. It is intended as a guide for the development of an economic purification scheme. A systematic approach is used that simplifies process planning and helps avoid errors. Planning consists of three main steps:

1. development of a logical downstream strategy in the laboratory,
2. optimization of process parameters, and
3. scaling up from the laboratory to the production scale.

Since the development of ion-exchange materials based on cellulose [1] and gel filtration media based on cross-linked dextrans [2], liquid chromatography has become the most important technique in the purification of proteins and peptides.

The introduction of additional separation principles (e.g., affinity chromatography [3]) and of new media based on cross-linked agarose [4] or organic polymers [5] has greatly increased the potential and performance of liquid chromatography. Many biologically active products can be purified chromatographically on a commercial scale. Their number continues to increase, partially as a result of the possibilities offered by genetic engineering. These substances include proteins from natural sources (e.g., plasma albumin) and products of modern biotechnology (e.g., monoclonal antibodies, interferon, vaccines, insulin, growth hormone, and tissue plasminogen activator).

Biological products are often highly active, relatively labile substances that occur in nature in very low concentrations. They must be separated from numerous impurities and are difficult to characterize fully.

The use of high-resolution liquid chromatography is, therefore, justified principally in the “high value – low volume” market, i.e., for the purification of products which are used in small amounts because of their efficacy, but which command high prices (Fig. 1). These are often enzymes, proteins, or peptides with potential use in clinical therapy or diagnosis; some are industrially used proteins with high purity requirements (e.g., proteases, glucose isomerase).

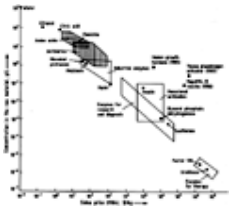

Figure 1. Products separated by process-scale chromatography (adapted from [11] with permission)

Experience with chromatography as a production method for therapeutic agents was gained primarily with insulin [6] and serum proteins [7]; chromatography has been used for their industrial production since the early 1970 s. A number of other substances produced by the human body are potential drugs with good prospects of becoming commercially successful (Table 1). With these highly effective substances, side effects are expected to be largely excluded. Licensing authorities already consider at least one chromatographic step essential for their purification [8]. However, implementation of chromatography on an industrial scale depends on the success of products purified in the pharmaceutical and diagnostic sectors.

Table 1. Chromatographically purified biological products used as human therapeutic agents

Product	Use
Albumin	plasma substitution
Epidermal growth factor (EGF)	wound healing
Erythropoetin	anemia in dialysis patients
Factor VIII	hemophilia
Human growth	growth disorders

hormone (HGH)	
Human insulin	diabetes
Interferons	treatment of cancer and infectious diseases
Interleukin-2	cancer treatment
Monoclonal antibodies (MAb)	cancer treatment and diagnosis, general diagnosis, affinity chromatography
Superoxide dismutase (SOD)	myocardial infarction therapy
Tissue plasminogen activator (tPA)	dissolution of blood clots
Tumor necrosis factor (TNF)	cancer treatment and diagnosis

Additional factors must be taken into account, especially in the development stage. Economic considerations require thorough optimization of parameters such as yield or the dimensioning of a process step. Factors that are not primarily chromatographic are also important, e.g., process hygiene. Finally, detailed process documentation for the control of biological product safety is almost as important as the preparative process itself.

Introduction of these additional aspects means that all sectors of the production process must be analyzed thoroughly and considered in relation to each other. Isolation and purification (downstream processing) can no longer be separated from one another. In addition, selection and treatment of the biological raw material (upstream processing) have far-reaching consequences. Process development thus creates an *integrated process* with a *logical sequence* of coordinated steps.

Because chromatographic purification is always the final step of this integrated process, it is greatly affected by decisions concerning the raw material and earlier steps (isolation and conditioning). The feasibility and elegance of the final purification, together with economic and technical requirements, thus best reflect the standards of the entire process. Criteria for evaluating a production process are listed in Table 2.

Table 2. Criteria for the evaluation of a production process

Criterion	Comments and requirements
Cost	investment costs are governed by product and market; production costs should be as low as possible
Productivity	should be high, with low effort
Product purity	depends on application; high purity has a market advantage
Process time	should be short because of low product stability and minimization of process costs
Execution	should be technically simple and clear (improved process safety, compliance with working time regulations)
Flexibility	necessary because of raw material variations, use of investment for different processes
Documentation	should be as complete as possible for product registration

[Top of Page]

2. Chromatographic Techniques

Various chromatographic methods available for the purification of biological products are summarized in Table 3. (For a detailed description, see ⟶ Basic Principles of Chromatography; ⟶ Enzymes – Chromatography.) Their management on a laboratory scale is described in many textbooks (e.g., [9]).

Table 3. Chromatographic separation methods

Method	Separation criteria	Important parameters
Gel filtration	particle size	column length
Ion exchange	particle charge	pH, ionic strength
Hydrophobic interaction and reversed phase chromatography	hydrophobicity	polarity, ionic strength
Affinity chromatography	biospecific interaction	ligand, eluent

Ion-exchange chromatography (IE) is used in ca. 75 % of all protein purifications; affinity chromatography (A) in ca. 60 %; gel filtration in ca. 50 %; and hydrophobic interaction chromatography (HIC) in ca. 30 % [10]. Reversed phase chromatography (RPC) is of interest virtually only for peptides and small organic molecules. In chromatographic purification sequences, the sample should be passed from step to step without conditioning. This can be achieved if the elution conditions for a given stage correspond to the starting conditions for the next stage. Logical sequences are given in Table 4.

Table 4. Logical sequence of chromatographic steps

Technique	Starting conditions	Elution conditions	Logical sequence	
			before	after
Gel filtration (G)	optional	optional	optional	optional
fractionation or desalting mode *				
Ion-exchange chromatography (IE)	low ionic strength	increased	HIC	HIC, A
fractionation or concentration mode		ionic strength		
Hydrophobic interaction	high ionic	low ionic	IE, A, salt	IE, A
chromatography (HIC)	strength	strength	precipitation	
Affinity chromatography (A)	low or high ionic strength	increased ionic strength, competitive eluent, reduced pH	HIC, IE	IE

* The fractionation mode should only be used as a final step.

2.1. Adsorptive Methods

All chromatographic methods apart from gel filtration utilize the selective adsorption of the sample on the surface of the chromatographic medium. This allows both *fractionation* of the sample and *concentration* of the desired component.

Separation of a mixture of substances involves two steps. First, a fraction of the sample is adsorbed and the nonadsorbed fraction is washed off. The bound components are then eluted by changing the eluent composition. Continuous, usually ionic strength gradients are often used for elution in the laboratory. In large-scale separations, stepwise gradients are generally preferred. However, programmable linear gradients are also being introduced. Their use allows direct scaleup, thereby eliminating a developmental step; fewer elution solutions are required and equipment costs may be lower.

Short columns only 5 – 20 cm long but with a large diameter (5 – 180 cm) are used. This allows very high flow rates of 200 – 400 cm/h, practically complete utilization of capacity, and rapid processing of large volumes. The process cycle time is unnecessarily long with longer columns, which are required only when the separation is carried out under isocratic conditions, i.e., when the composition of the elution buffer remains constant as in reversed phase or, more rarely, ion-exchange chromatography.

The fraction of the gel bed utilized for adsorption varies from 10 % for the separation of two substances with very similar behavior to nearly 100 % for particularly selective conditions.

2.1.1. *Ion-Exchange Chromatography (* ⟶ Ion Exchangers*)*

Many biological molecules possess a net charge whose size and sign depend on pH. This property is utilized in ion-exchange chromatography for selective adsorption onto a carrier derivatized with charged groups. Distinction is made between anion exchangers with positively charged groups (e.g., diethylaminoethyl or quaternary ammonium) and cation exchangers with negatively charged groups (e.g., carboxymethyl, sulfopropyl). The exchangers may be either strong or

weak. Examples are listed in Table 5.

Table 5. Examples of media for ion-exchange chromatography.

Trade name *	Manufacturer
Strong ion exchangers	
Q Sepharose Fast Flow	Pharmacia (Sweden)
S Sepharose Fast Flow	Pharmacia
Fractogel TSK SP-650	Merck (FRG)
QA Trisacryl LS	IBF biotechnics (France)
SP Trisacryl LS	IBF biotechnics
Q Sepharose High Performance **	Pharmacia
S Sepharose High Performance **	Pharmacia
Weak ion exchangers	
DEAE Sepharose Fast Flow	Pharmacia
CM Sepharose Fast Flow	Pharmacia
Fractogel TSK DEAE-650	Merck
Fractogel TSK CM-650	Merck
DEAE Trisacryl LS	IBF biotechnics
CM Trisacryl LS	IBF biotechnics

* Abbreviations: DEAE = diethylaminoethyl; CM = carboxymethyl; Q, QA = quaternary ammonium; S, SP = sulfopropyl; Sepharose Fast Flow = highly cross-linked agarose matrix; Fractogel = hydrophilic vinyl polymer matrix; Trisacryl = hydrophilic acrylic polymer matrix.
** High performance, particle size 34 µm.

Sample constituents bind to the exchanger at low ionic strengths. Between 10 and 100 mg of protein can be adsorbed per milliliter of exchanger. Numerous sample components usually compete for adsorption onto binding sites. Therefore, the binding capacity of an exchanger for the target product may differ very greatly from its theoretical capacity for a pure substance. The strength and selectivity of binding for a given protein are controlled by the initial pH and ionic strength. The actual *dynamic capacity* used under working conditions (i.e., the binding capacity of a given exchanger for a product with given impurities) is governed not only by the pH and ionic strength, but also by the flow rate and matrix properties (see Section Experimental Data). Bound substances are eluted selectively by increasing the ionic strength or changing the pH. Examples of cation- and anion-exchange separations are shown in Figures 2 and 3, respectively. Separation performance can be controlled by the shape of the gradient used.

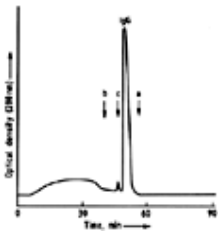

Figure 2. Purification of immunoglobulin G (IgG) from plasma by cation-exchange chromatography

Elution buffers: a) 0.02 mol/L sodium acetate, pH 6.5; b) 0.01 mol/L glycine, pH 7; c) 0.1 mol/L glycine and 0.15 mol/L NaCl, pH 9.0

Chromatography medium: CM Sepharose Fast Flow; column diameter: 80 cm; bed height: 5 cm; productivity: 1.3 kg of IgG per hour

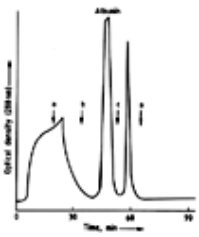

Figure 3. Purification of albumin from plasma by anion-exchange chromatography

Sodium acetate elution buffers: a) 0.020 mol/L, pH 5.2; b) 0.025 mol/L, pH 4.5; c) 0.15 mol/L, pH 4.0

Chromatography medium: DEAE Sepharose Fast Flow; column diameter: 80 cm; bed height: 5 cm; productivity: 1.7 kg of albumin per hour

Hydrophobic interaction chromatography, affinity chromatography, or gel filtration is best used after ion-exchange chromatography (Table 4).

2.1.2. *Affinity Chromatography*

In affinity chromatography, the product to be purified is specifically adsorbed on an affinity ligand coupled to a matrix. This is analogous to the specific binding processes that are necessary for the biological function of proteins, but with the difference that the immobilized ligands do not undergo subsequent conversion.

Well over 100 different ready-to-use media are available, some examples are given in Table 6. Activated media for in-house production are also being marketed. Many techniques exist, with different starting and elution conditions; their range of applications is wide [12]. They can be divided into three groups according to use:

1. *specific ligands* (e.g., monoclonal antibodies, protein A, enzyme inhibitors) for single proteins [13];
2. so-called *general or group-specific ligands* such as immobilized metal ions [14], and
3. *artificial ligands* such as textile dyes [15].

Table 6. Examples of affinity chromatography media

Trade name	Manufacturer
Group-specific ligands	
Chelating Sepharose Fast Flow	Pharmacia (Sweden)
Lysine Sepharose Fast Flow	Pharmacia
Specific ligands	
Protein A Sepharose Fast Flow	Pharmacia
Heparin Sepharose	Pharmacia
Artificial ligands	
Avidin Ultrogel A4R	IBF biotechnics (France)
Blue Sepharose Fast Flow	Pharmacia
Matrex Gel Blue A	Amicon (USA)
Blue Trisacryl M	IBF biotechnics
Fractogel TSK AF Red	Merck (FRG)
Fractogel TSK AF Phenylboronate	Merck

Because of the high selectivity of affinity media, very large volumes of dilute product solutions can be processed. The capacity of the gel is not reduced by the binding of numerous foreign proteins, which occurs with ion-exchange or hydrophobic interaction chromatography.

Purification factors of well over 1000 can be achieved with affinity chromatography, although the mean enrichment obtained in practice is only ca. 100-fold [10]. Higher values can often be achieved only with relatively impure samples.

The sample mixture is generally applied to the column at low ionic strength and in the neutral pH range. After nonbound impurities are washed out, the product can be eluted by using a gradient of increasing ionic strength or by reducing the pH.

A further increase in selectivity can be achieved with product-specific eluents which compete with the product for binding to the matrix. In favorable cases, a product can be purified to homogeneity in one or two steps with the aid of affinity chromatography. Figure 4 shows the one-step purification of monoclonal antibodies with Protein A Sepharose. (Protein A is a bacterial protein that contains high-affinity sites for immunoglobulin G and is used as a ligand for purifying antibodies.)

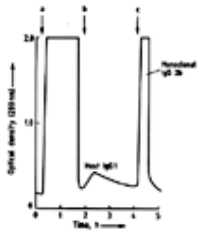

Figure 4. Purification of monoclonal antibody 2 b from mouse ascites fluid by affinity chromatography

a) Sample: 50 mL of mouse ascites fluid diluted 1 : 2 with phosphate-buffered saline (PBS), pH 8.0; b) Eluent: PBS, pH 6.0; c) Eluent: PBS, pH 5.0

Chromatography column (1.6×5 cm) containing 10 mL of Protein A Sepharose CL-4 B. Results published courtesy of A. LARSSON, Pharmacia.

2.1.3. *Hydrophobic Interaction Chromatography*

Most proteins (not only “hydrophobic” proteins) possess hydrophobic regions on their surface which allow interaction with a hydrophobically derivatized matrix in aqueous surroundings. The forces responsible for adsorption to a carrier material of this type depend on the increase in entropy arising from the displacement of hydrophobic molecular regions from the polar surroundings [16]. These forces are increased at high salt concentrations (salting out).

Phenyl, butyl, and octyl residues are the common ligands used in HIC (Table 7). The hydrophobic properties of phenyl matrixes are suitable for purifying most proteins under mild conditions.

Table 7. Examples of media for hydrophobic interaction chromatography

Trade name	Manufacturer
Phenyl Sepharose Fast Flow	Pharmacia (Sweden)
Octyl Sepharose CL-4B	Pharmacia
Fractogel TSK Butyl	Merck (FRG)

Hydrophobic interaction is affected by the action of specific salts or the polarity of the solvent [9], [17]. Proteins can be adsorbed by a hydrophobic matrix at high salt concentrations, and nonbound material can be washed away. The bound substances are then eluted selectively with decreasing polarity in a second step. Decreasing salt gradients are used generally, but further reduction of polarity by addition of nonpolar organic solvents such as ethanediol may be necessary (Fig. 5).

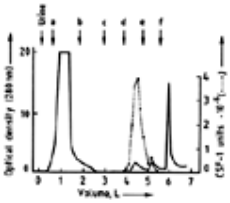

Figure 5. Purification of colony stimulating factor CSF-1 from human urine by hydrophobic interaction chromatography (published with permission [47])

Elution buffers: a) – e) Phosphate-buffered saline containing decreasing concentrations of NaCl (mol/L): a) 4; b) 3; c) 2; d) 1; e) 0.5; f) 4 mol/L guanidine · HCl in 10 mmol/L NaOH containing 4 mol/L 20 vol % ethanol

Sample: 500 mL of concentrated human urine containing 4 mol/L NaCl; chromatography medium: Phenyl Sepharose CL-4 B; column dimensions: 5×22 cm; flow rate: 15 cm/h

Chromatographic techniques using low ionic strength as the starting condition can be utilized after HIC (Table 4).

2.1.4. *Reversed Phase Chromatography*

Hydrophobic interactions are also utilized in reversed phase chromatography (RPC). However, whereas HIC is a nondenaturing method, proteins are usually denatured by the nonpolar solvents used for elution from the highly hydrophobic RPC matrices [18]. Examples of RPC media are Lichrosorb and Lichroprep (Merck, Federal Republic of Germany), and Matrex silica gels (Amicon, United States). The preparative use of RPC in biotechnology is thus limited to the fractionation of substances such as antibiotics (Fig. 6) and synthetic or physiological peptides [19] up to about the size of insulin, M_r 5800. The use of RPC together with gel filtration and ion exchange has already proved to be of value in the production of peptides such as oxytocin and vasopressin [20].

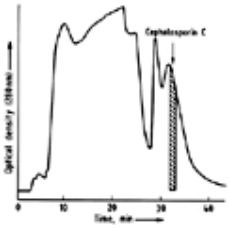

Figure 6. Reversed phase chromatography of a cephalosporin C fermentation broth (published with permission [11])

Chromatography medium: $PREPC_{18}$ column: 6×2 ft segment; mobile phase 97.5 parts of 20 mmol/L ammonium acetate (pH 5.2), 2.5 parts of acetonitrile; flow rate: 1 L/min; injection volume: 16 L containing 4.5 mg/mL cephalosporin C

Because of the particularly good selectivity of RPC, very high resolution can be obtained. Future developments may produce RPC materials which, unlike conventional silica-based materials, are more suitable for industrial production processes. Silica materials are not alkali-stable and often have poor batch-to-batch reproducibility.

2.2. Gel Filtration

In gel filtration, a porous matrix is used to separate a mixture of substances according to their molecular mass and shape. The size range in which resolution can be achieved is determined by the pore size of the matrix. Large molecules do not enter the pores and are eluted quickly. Small molecules enter the pores and move through the column more slowly. Molecules are thus eluted in order of decreasing molecular mass. Ideally, adsorption of sample constituents does not occur.

Gel filtration is used for *fine fractionation* of individual substances, e.g., separation of monomers and oligomers. The fine fractionation of immunoglobulin G (IgG) is shown in Figure 7. Media with a separation range that includes the molecular masses of the substances to be separated are employed (Table 8).

Table 8. Examples of gel filtration media

Trade name	Separation range, M_r	Manufacturer	Gel matrix polymer
Sephacryl S-100 HR	1000 – 100 000	Pharmacia (Sweden)	dextran – acrylamide
Sephacryl S-200 HR	5000 – 250 000	Pharmacia	dextran – acrylamide

Sephacryl S-300 HR	10 000 – 1.5×10^6	Pharmacia	dextran – acrylamide
Sephacryl S-400 HR	20 000 – 8×10^6	Pharmacia	dextran – acrylamide
Fractogel TSK HW-55	1000 – 1×10^6	Merck (FRG)	hydrophilic vinyl
Fractogel TSK HW-65	50 000 – 5×10^6	Merck	hydrophilic vinyl
Trisacryl GF 2000 LS	10 000 – 15×10^6	IBF biotechnics (France)	hydrophilic acrylic
Sephadex G-25 *	1000 – 5000	Pharmacia	cross-linked dextran
Bio-Gel P-2 *	100 – 1800	Bio-Rad (USA)	acrylamide
Matrex Cellufine GH 25 *	200 – 2000	Amicon (USA)	silica based
Trisacryl GF 05 LS *	200 – 2500	IBF biotechnics	hydrophilic acrylic
Sepharose 12 prep grade **	1000 – 300 000	Pharmacia	highly cross-linked agarose
Superdex 75 **	3000 – 70 000	Pharmacia	agarose – dextran

* Media for desalting – rebuffering,
** High-performance media, particle size 34 µm.

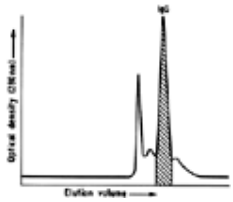

Figure 7. Separation of immunoglobulin G (IgG) oligomers from a previous ion-exchange step by gel filtration (final step in purification of monoclonal antibodies)

Chromatography medium: Sephacryl S-200 HR; column diameter: 11.3 cm; bed height: 100 cm; eluent: 0.05 mol/L Tris – HCl, 0.15 mol/L NaCl, pH 7.5; flow rate: 7.5 cm/h

When gel filtration is used to separate salts from macromolecules (*desalting* or *rebuffering* of a protein sample), only *group fractionation* is achieved. This technique is an alternative to ultrafiltration or diafiltration. Gel filtration media suitable for desalting have a fractionation range for molecules with molecular masses below 5000 (Table 8).

In gel filtration, only the pore volume of the particles packed in the column is available for separation, so sample volumes are limited. The upper protein concentration limit depends on the viscosity, the viscosity of the sample should not be more than twice that of the elution solvent.

The possible sample volume per cycle is restricted to about 5 % of the column volume in fine fractionation and to ca. 15 – 30 % in desalting. Column lengths ranging from 60 cm (desalting mode) to 180 cm (fractionation mode) and diameters of 5 – 180 cm are used. Flow rates are generally 3 – 15 cm/h for the fractionation mode and 100 – 200 cm/h for the desalting mode.

Gel filtration always leads to rebuffering if the compositions of the sample and the eluent differ. Consequently, irrespective of sample composition, gel filtration can be combined with all other chromatographic techniques or can be used as a conditioning process for the next step (Table 4).

Gel filtration in the desalting mode can be applied with high productivity at any processing stage. In the fractionation mode, it can be used economically only as a final, fine purification step (separation of polymers) due to its relatively low volume capacity. Formulation of the product is also achieved this way.

[Top of Page]

3. Downstream Strategy

All preparative steps must be logically coordinated, and subsequent processing must be taken into account in selecting raw materials (cf. Chap. Introduction). The present chapter describes the correct strategy for an integrated preparative process. The main factors considered are:

1. minimization of the number of process steps required,
2. optimal preparation of the raw material, and
3. development of a chromatographic purification system.

3.1. Number of Preparative Steps

Product yield decreases with the number of preparative steps (Fig. 8). Each additional step also increases production time and, thus, processing cost. Therefore, the number of steps must be kept as small as possible. The target product can generally be purified in two or three chromatographic steps.

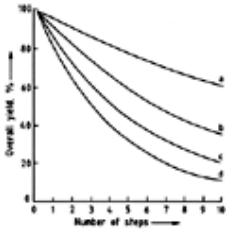

Figure 8. Overall product yield as a function of the number of process steps

Percentage yields for individual steps: a) 95; b) 90; c) 85; d) 80

However, an important exception to this rule may occur when a very complex, perhaps more expensive, step replaces several simple ones. In this case, preference might be given to another rule: the process should be kept as simple as possible.

A number of ways are available for simplifying processing and for minimizing the number of steps (Table 9). The raw material should be selected and optimized in such a way that the product is easy to isolate and impurities are kept at a minimal level (Section Preparation of an Optimal Raw Material). The crude product should be optimally conditioned for the first purification step. Subsequent chromatographic steps should be arranged in a logical sequence (Section Logical Sequence of Chromatographic Steps). Each step should be optimized to make full use of the separation performance of the method used (Chap. Optimization). Only necessary processing should be performed. The degree of purity required does not always have to be the maximum purity achievable by chromatography. In some cases, characterization of the residual impurities and demonstration of their “harmlessness” are sufficient.

Table 9. Integration of chromatography and accompanying operative steps

Operative step	Parameters affected by operative step
Selection of raw material	1) product concentration
	2) product accessibility (is cell disruption necessary?)
	3) number and type of impurities (can protease content be decreased?)
	4) batch-to-batch variation
Genetic engineering of raw material	1) product concentration (can it be increased?)
	2) product accessibility (can it be improved?)
	3) number and type of impurities (can protease content be lowered?)
	4) reproducibility (can stability of cloned gene be guaranteed?)
	5) product safety (effort required for removal of foreign DNA)
Cell culture	1) scale of final purification (continuous culture allows a reduced scale)
	2) number and type of impurities (can they be reduced by modifying the culture medium?)
Isolation, conditioning	1) lifetime of gel media (clarification of product solution)
	2) efficiency of first chromatographic step (adjustment of pH, ionic strength, protein concentration)
	3) product yield (minimization of mechanical stress, protease content, number of steps)

3.2. Preparation of an Optimal Raw Material

As a result of advances in genetic engineering, proteins and peptides are increasingly being produced by genetically modified cells that can be cultivated in fermentors. The cost of fermentation is a highly significant factor in the calculation of production costs. Thus, the product must be isolated from the fermentation broth in maximum yield. The resultant product solution should be fully clarified, particle-free, and preferably sterile-filtered, i.e., a suitable raw material for chromatographic purification [21].

Choice and Treatment of the Raw Material. Raw materials may be either natural or tailor-made.

Natural materials have specific properties that cannot be changed: availability, product content, contamination with secondary materials, and risk factors. An example is human serum used for the production of albumin, clotting factors, and other plasma proteins [7]. The preparation of natural materials must be adapted to the given conditions.

Tailor-made materials can be modified easily. Availability, product concentration, and impurities can be controlled within wide limits. Genetic engineering [22], the hybridoma technique [23], or combinations of both (chimeric antibodies) can be used to modify the chemical structure of a product to improve its action or eliminate side effects. Raw materials include genetically engineered bacteria, yeasts, or mammalian cells, as well as hybridoma cells.

Table 10 is a survey of the potential advantages and disadvantages of various producer cells and organisms (see also [41]).

Table 10. Advantages and disadvantages of productive cells [24]

Cells	Advantages	Disadvantages
Bacteria	1) cell culture well mastered	1) pathogenicity
	2) many genetic "tricks" possible [25], [26]	2) protein glycosylation not possible *
	3) inclusion bodies protect against proteolysis [27]	3) product contamination by foreign DNA and viruses
	4) high productivity [28]	4) inclusion bodies difficult to process
Yeasts [29]	1) cell culture well mastered	1) different glycosylation pattern from mammalian cells
	2) hardly any endogenous protein secretion [30]	
	3) secretion can be partially controlled genetically [31]	
	4) no endotoxins, safe organisms	
Fungi [32], [33]	1) cell culture well-mastered, simple culture media	1) different glycosylation pattern from mammalian cells
	2) safe organisms (*Aspergillus niger*)	
	3) high productivity, used for industrial enzymes [34]	
Insect cells [35]	1) baculovirus as vector with good expression system	1) posttranslational modification not clear
	2) high productivity [36]	2) costs similar to mammalian cells [36]
Mammalian cells [37], [38]	1) posttranslational modification, correct glyco-sylation [39]	1) genetic instabilty, oncogenicity
		2) sensitivity of cells
	2) cell immobilization, e.g., with microcarriers [40]	3) slow growth, low cell density
	3) continuous perfusion, removal of proteases	4) relatively low productivity, high cost

* Many proteins require correct posttranslational glycosylation to perform their functions.

Isolation of Cell Products. The target product must be isolated from its natural surroundings to allow further purification. The steps required for this are governed largely by the nature of the raw material and the localization of the product (intra- or extracellular).

Intracellular products occur in the cytoplasm (perhaps in the form of inclusion bodies) or in the periplasmic space. Harvested cells must be disrupted and particulate components removed by one or two centrifugation steps and microfiltration. In the case of inclusion bodies, refolding of the product must be achieved. Conditioning of the fermentor broth by addition of flocculants or filtration aids, by heat treatment, or by pH adjustment may also be necessary.

Table 11 summarizes the techniques used for preliminary isolation of intracellular products. Many isolation methods are problematic due to the mechanical stress imposed on the product and difficulty in scaling up.

Table 11. Isolation of cell products

Objective	Technique	Advantages	Disadvantages
Product release	cell disruption [9], [42]	none	mechanical stress
Clarification	centrifugation [38]	none	aerosol formation, scaleup difficult
	microfiltration [43]	simultaneous removal of bacteria	short membrane lifetime
Concentration	ultrafiltration [44],	simultaneous	high costs, system

	[45]	sterilization	difficult to optimize
Preliminary purification (salts, proteases, nucleic acids)	precipitation	none	high costs, environmental contamination
	adsorption [46]	simultaneous use of selectivity for removal of contaminants	maintenance can be laborious
Lipid removal	defatting	none	low product yields, environmental contamination

Techniques for greatly simplified isolation have also been developed for nonsecretory systems. Proteins have been successfully extracted in two-phase systems used for processing cell homogenates (⟶ Biochemical Separations [48]). If Phenyl Sepharose is added to the system, the proteins can be adsorbed and simultaneously protected against the action of proteases [21]. With this combination, the crude material can quickly be prepared for final purification.

Extracellular products are secreted into the medium so that the cells need only be filtered off, possibly continuously. This offers advantages in terms of time and yield. A secretion system also seems to be useful in avoiding the technical problems listed in Table 11. However, special regulatory aspects must be considered when running a continuous cell culture [49].

In favorable cases, continuous adsorption onto a chromatographic medium can be used, which leads to rapid separation of critical contaminants such as proteases that might digest the secreted product. Unter certain circumstances, classical isolation steps can be avoided completely (see Chap. Developments). As a result of the simultaneous concentration effect, all further steps can be undertaken on a significantly smaller scale saving both time and costs.

Product Stabilization. The isolated product requires specific conditions if it is to remain stable for long periods. Optimal ionic strength, pH range, and concentration of added stabilizers (e.g., trace elements) must be determined and adjusted in the crude product. Protease activity can be controlled to some extent by keeping the temperature as low as possible. Organic solvents or special additives such as protease inhibitors should be avoided because their later removal causes problems. Foaming should be inhibited to prevent denaturation.

3.3. Development of a Chromatographic Purification System

An individual purification system must be determined for each biological product. A screening process is used to select appropriate chromatographic methods. The choice of a chromatographic material is determined not only by the "chromatographic criteria" applied during screening (Section Screening of Chromatographic Methods (Method Scouting)), but also by "non-chromatographic marginal conditions." Examples of the latter are process hygiene measures and the scale of production, which dictate requirements such as the physical and chemical stability of gel media. Chromatographic techniques are combined to give a logical sequence with only a few steps (Section Logical Sequence of Chromatographic Steps).

All developmental work must be carried out on the smallest suitable scale in the laboratory to save time. Usually, each step is then rechecked and conditions are optimized (fine tuning). This may be done after the process is implemented in the pilot plant where significant amounts of the substance are produced for toxicological or clinical studies. Nevertheless, optimization should be continued to achieve the smallest scale of production possible. Optimization is discussed in Chapter Optimization. Table 12 gives a scheme of tasks allocated to each operational unit involved in process development.

Table 12. Allocation of tasks in the development of an industrial chromatographic purification process

Operational unit	Tasks listed in chronological order
Research and development	1) development of analytical methods
	2) screening of possible purification methods
	3) choice of suitable separation media according to separation performance, dynamic capacity, chemical stability (lifetime, process hygiene), mechanical stability (handling on a production scale), and current documentation *
	4) development of a purification plan
	5) optimization of each step according to productivity, required purity, product safety (removal of DNA, toxins, etc.), and adaption to changes after scaling up the cell culture

	6) full documentation of procedure
Pilot plant	1) scaling up process: requires modification due to column features and longer storage time for intermediate products
	2) manufacture of experimental amounts of product
	3) process automation
	4) development of manufacturing instructions including handling of separation media, sterilization, pyrogen removal, quantitative data for dimensioning
Production plant	1) planning and construction of a production plant fulfilling the requirements
	2) production and on-site analysis work
	3) process validation, batch documentation

* Instructions given in the literature can generally be extrapolated to the equivalent separation media suitable for industrial use (e.g., instructions for DEAE cellulose can be used for DEAE Sepharose Fast Flow).

3.3.1. *Screening of Chromatographic Methods (Method Scouting)*

The “chromatographic” criteria for gel selection are studied by testing a number of gels and buffer solutions for ion-exchange, affinity, hydrophobic interaction, or gel filtration chromatography. Parameters such as the nature of the groups coupled to the matrix or the gel porosity are selected together with provisional conditions such as pH, loading, and flow rate. Precise conditions are determined during subsequent optimization when the gel and the sequence of processing steps have been selected (see Chap. Optimization).

3.3.1.1. *The Gel Bank*

Methodological screening is best carried out with the aid of a *gel bank*, which is a collection of different chromatographic media. The media are stored ready-for-use in small, perfectly packed (1 – 50 mL) columns; their packing quality can be tested. Only materials with physical and chemical stabilities suitable for production processes should be used, examples are given in Table 13. The preliminary selection of gel media is based on the best possible linear flow rate and on nonchromatographic criteria such as expected lifetime and process hygiene.

Table 13. Examples of chromatographic media suitable for industrial processes

Chromatographic method	Medium	CIP * using sodium hydroxide, mol/L
Gel filtration	Sephacryl S-100 HR, S-200 HR, S-300 HR	0.5
	Sephadex G-25 c, m, f, sf	0.5
	Fractogel TSK HW-55, HW-65	1
Ion exchange	Q Sepharose Fast Flow	2
	S Sepharose Fast Flow	2
	DEAE Sepharose Fast Flow	1
	CM Sepharose Fast Flow	1
	Fractogel TSK SP-650, DEAE-650, CM-650	1
Hydrophobic interaction	Phenyl Sepharose Fast Flow	1
	Fractogel TSK Butyl	no data available
Affinity	Chelating Sepharose Fast Flow	1
	Arginine Sepharose 4 B	0.5
	Protein A Sepharose Fast Flow **	no data available
	Fractogel TSK AF Red, AF Phenylboronate	no data available

* Cleaning in place.
** Medium with a labile ligand, maintenance using other washing procedures [50].

Modern media with better mechanical and chemical stability, and higher productivity than those required in the laboratory are expensive but may allow considerable savings when used on a production scale. Although cheaper materials can also give the desired separation, they may incur considerable costs, mainly as a result of more laborious regeneration and much shorter lifetime.

The major requirements of the gel material can be deduced from the criteria listed in Table 2.

A *high dynamic capacity* contributes to faster production at lower cost and is favored by the following factors:

1. low linear flow rate,
2. large pore diameter,

3. low sample concentration,
4. high diffusion constant of the sample,
5. high binding rate constant,
6. small gel particle diameter, and
7. low competitive binding of other sample constituents.

The last aspect implies high selectivity, which is the most important factor in designing a high dynamic capacity. Aspects 1 and 3 of course compromise productivity. Thus, a balance has to be found. It is inadvisable to choose a gel material with a high dynamic capacity at only a very low flow rate or sample load.

Good physical stability allows easy packing of the gel in large columns, without deterioration of the separation performance and packing quality achieved in the laboratory. The high flow rates (200 – 400 cm/h) obtained in the laboratory can thus also be used on a production scale.

High chemical stability, especially to sodium hydroxide (0.5 – 2 mol/L, Table 13), is a prerequisite for sterilization and the removal of pyrogens and other contaminants from the packed column (cleaning in place, CIP). Chemical stability is also the best general indicator of the lifetime of a gel under processing conditions.

Modern gel media with *a long predictable lifetime* can be considered part of the column and thus are capital investments rather than consumables. Far more than 100 cycles can be expected of ion exchangers, even after repeated regeneration and cleaning in place (Fig. 9). Some affinity media that utilize proteins as ligands are exceptions; they can be used for only a limited period because they are not stable to rigorous cleaning procedures.

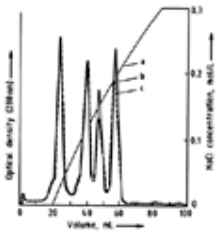

Figure 9. Separation performance of an ion-exchange column after 100 CIP cycles

a) NaCl concentration of eluent; b) Initial chromatogram; c) Chromatogram after 100 cycles;

Chromatography medium: Q Sepharose Fast Flow; sample: 2.5 mg transferrin, 5 mg ovalbumin, 5 mg β-lactoglobulin in 0.5 mL; column diameter: 1 cm; bed height: 4.5 cm; gradient volume: 83 mL; eluent: 20 mmol/L piperazine, pH 6.0; flow rate: 24 mL/h

The *long-term availability* of large amounts of chromatographic media with little batch-to-batch variation guarantees that the process can be carried out reproducibly for as long as required.

Extensive documentation, i.e., Product Master Files, on the chromatographic media and analytical methods supplied by the manufacturer for initial control should facilitate both product registration and regular control checks for the user.

General requirements also apply to the *gel matrix properties:*

Low nonspecific interaction of the gel matrix with sample constituents increases yields and simplifies handling of the material. For proteins, the matrix should have hydrophilic properties; suitable gels are based on cross-linked agarose, dextran, hydrophilic vinyl polymers, or polyacryl-amide.

High matrix porosity or decreased gel bead diameter improves separation kinetics and shortens the total process. Materials are available with bead diameters of ca. 10, 30, and 90 µm (see Tables 5 and 8).

High-Performance Media in the Gel Bank. Particle diameter is misused as a criterion to distinguish high-performance liquid chromatography from conventional liquid chromatography. High-performance gels are considered to have a particle diameter of 30 µm or less. Column efficiency (measured as the number of theoretical plates *N*) and separation kinetics can indeed be improved by using smaller particles. However, in process-scale chromatography, use of the largest possible particle diameter for the desired separation is preferred because it minimizes the resulting counterpressure and thus reduces running costs [51]. Above a certain level, very little is gained by a further increase in column efficiency [52].

Thus, high performance in preparative chromatography is defined by increased productivity that is obtained by optimizing the selectivity of each step (Fig. 10), rather than by increasing the efficiency *N* (selectivity is described in Selectivity). This is achieved by choosing the correct pH in ion exchange, the correct ligands in affinity chromatography, or the correct pore distribution in gel filtration [52].

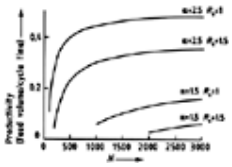

Figure 10. Relationship of selectivity (α), resolution (R_S), efficiency (*N*, number of theoretical plates), and productivity (published with permission [52])

Productivity is improved more by increasing the selectivity than by increasing the efficiency.

In addition, because of their relatively low matrix content, media with large particle diameters (>30 µm) often show higher porosity, less nonspecific adsorption, and a greater capacity for product loading than do media with small particle diameters (<30 µm). High-performance media often have a lower capacity than conventional media because less room is available on the highly cross-linked matrix for derivatization with charged groups or affinity ligands, and the porosity is not optimal for use with macromolecules.

The gel bank should thus include primarily materials with particle diameters around 90 µm. Particles of 30 µm are useful and, in some cases, indispensable for final polishing or for difficult separation problems. They require more complex apparatus if high flow rates are used because a high counterpressure is produced. A gel bank that includes 30 and 90 µm particles can solve almost all separation problems.

However, particles of 10 µm are used in very small columns for particularly rapid process development. Other materials with larger particle diameters must then be available to which the results can later be transferred as soon as the scale is increased. The appropriate conversion factors must be known. With ion exchangers, the gels must all have the same charged groups, and the influence of different charge densities must be determined in advance experimentally (see Section Transfer Protocol). This is the manufacturer's task. For gel filtration media, the selectivity curves must be identical.

Columns from the gel bank can be run in an automated medium-pressure chromatography system that works with all types of media and particle sizes (Fig. 11). The testing of materials and their related conditions is facilitated by the small scale and automation.

Figure 11. An automated medium-pressure chromatography system

3.3.1.2. *Experimental Data*

Many biological substances of interest for industrial production are described in the literature and have been purified or enriched by chromatography. Extensive reference lists are available [53]. Published methods may be suitable for initial experiments but are often inappropriate for a production process.

Sample Material. For a systematic approach, the sample material must be thoroughly characterized to determine which techniques should be used in screening. The molecular masses, isoelectric points, and titration curves of the product solution obtained in each purification step can be determined by rapid, automated electrophoresis (Fig. 12). Chromatographic titration curves can also be obtained. These data often provide important information for choosing a technique and for determining suitable separation conditions [54]. It is worthwhile to characterize impurities such as proteases, lipids, DNA, and endotoxins carefully. Additional help is obtained by considering the logical sequence of chromatographic steps (Section Logical Sequence of Chromatographic Steps).

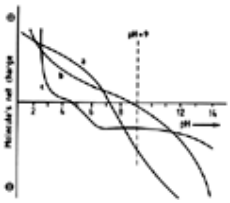

Figure 12. Titration curves of three proteins (a, b, and c)

Optimal pH for separation is 9.0 because the difference in net charge is greatest at this point.

Product yield and degree of purification must be determined at each step to evaluate results and compare alternative methods.

The viscosity, primarily a result of high protein and nucleic acid concentrations, should not be greater than ca. 40 N s m^{-2}. If higher, it must be reduced by dilution to avoid increased pressure in the column. Operations should be carried out at room temperature if possible because the viscosity increases and the flow rate decreases at cold-room temperatures of 4 –10 °C.

Gel Media. Guidelines regarding parameters such as flow rate range or height of gel bed can be obtained from the gel manufacturer's product specifications.

Dynamic capacity is a measure of the degree to which adsorbents can be loaded under process conditions and must, therefore, be determined with the original sample in each gel used, as described in [55], [56]. However, guidelines for loading are available for gel filtration (see Section Gel Filtration), and published data may also be usable for adsorption techniques.

The pH of the starting buffer in ion-exchange chromatography should not be more than one pH unit higher or lower than the isoelectric point of the desired product. This ensures that the greatest possible number of impurities either are not bound or can be removed by selective elution. The dynamic capacity for product adsorption is thus maximal. For most applications the pH selected should be kept constant during the separation process; adsorption and elution are then controlled by adjusting the conductivity. A pH value is chosen at which the product just binds but impurities are largely washed away.

In affinity chromatography, screening is often tedious because the choice of gel is difficult if published data are not available. Materials with a broad spectrum of applications are recommended as starting materials, e.g., lysine, chelating, or dye gels. Preactivated gels for in-house synthesis are available if a highly specific affinity medium is required. Optimized affinity media with special matrices, ligands, or ligand concentrations can also be obtained as tailor-made products.

Gel filtration is used for fine fractionation at the end of a purification scheme (see Section Logical Sequence of Chromatographic Steps). The main purpose of screening is thus to find the gel with the best selectivity in the desired separation range of molecular masses. Particular attention should be paid to adsorption effects. The separation range for the desalting mode is predetermined, and the number of possible gel materials is restricted. The choice is based on the productivity desired and on general production requirements.

3.3.2. *Logical Sequence of Chromatographic Steps*

Requirements relating to purification techniques and gel materials are based on the experimentally determined composition of the sample. Conversely, each chromatographic technique requires that the sample fulfills specific, known requirements. Furthermore, each separation step produces a new sample whose properties are determined by the chromatographic technique used and are thus partly predictable.

The interdependence of the sample and the separation techniques means that each purification scheme can be planned to a large degree to develop an effective, integrated process. Individual steps should be combined in a logical, coordinated sequence; the logical position and function of the various steps are summarized in Table 14.

Table 14. Logical position of chromatographic separation methods in purification sequences

Method	Operating mode	Volume capacity	Productivity *	Logical position	Function
Gel filtration	fractionation	low	low **	final step	final purification and formulation
	desalting	moderate	high	optional	conditioning (but a logical se-quence avoids frequent condi-tioning)
Ion exchange	concentration	very high	very high	starting step	reduction of volume, removal of critical contaminants
	fractionation	high	high	optional	fine purification, final purifi-cation
Hydrophobic interaction	fractionation	high	high	starting step, intermediate steps	removal of critical contami-nants, fine purification
Affinity	fractionation	very high	very high	optional, depends on technique	fine purification, final purifi-cation, removal of critical contaminants

* See Section Purpose of Optimization.
** But see Section Optimization of Gel Filtration.

3.3.2.1. *Selection Criteria*

A biological product must almost always be purified in several steps. A sequence of chromatographic methods utilizing several different separation principles (e.g., hydrophobicity, charge, and biological specificity of the product) is advisable.

Individual techniques should follow each other so that the sample can pass from one purification step to the next without intermediate conditioning steps such as concentration by ultrafiltration or rebuffering by gel filtration (Table 4). Appropriate starting conditions for each subsequent step should be achieved by simple means such as dilution, addition of salts, or pH adjustment. Separate concentration steps can easily be avoided because large volumes can be purified and concentrated in one stage with adsorptive techniques (Table 14).

Sample volumes should be reduced quickly during the early steps. Impurities that reduce yield or shorten gel life should be removed at an early stage. An example is the preparation of monoclonal antibodies from cell perfusion systems [57].

If intermediate conditions adversely affect product stability, the step in question must be carried out quickly and the product must be restabilized, e.g., in affinity chromatography of monoclonal antibodies using Protein A Sepharose.

For reasons of cost, high performance techniques using gel materials with particle diameters of ≤30 µm (Section The Gel Bank) should be used only at the end of a purification scheme. Small columns can be employed in the last step in which contaminants harmful to the gel are no longer present.

The definitive choice of a technique is made after coordination of sample properties, chromatographic method, and gel characteristics. Important properties of a sample are its total volume, protein and product contents, ionic strength (conductivity), pH, and viscosity (Table 15). Important properties of a chromatographic method (Tables 4 and 14) are the

binding capacity for the sample in terms of both mass and volume; the ionic strength required for effective loading; and gel properties such as flow rate or stability to stress caused by the sample.

Table 15. Preparation of sample for the first chromatographic step

Parameter	Numerical value	Starting step	Conditioning required
Product concentration	>1 g/L	1) adsorption (fractionation mode)	adjustment of conductivity (e.g., by using a desalting step)
	<1 g/L	2) adsorption (concentration mode)	adjustment of conductivity (e.g., by further dilution)
		3) ultrafiltration	none
Viscosity (total protein)	>40 N s/m^2	1) adsorption	dilution
Conductivity	>5 mS/cm	1) ion exchange	dilution (product concentration ca. 10 mg/L)
		2) gel filtration (desalting mode)	desalting (product concentration 1 g/L)
		3) ultrafiltration	none
		4) hydrophobic interaction chromatography	salt addition, e.g., after precipitation or ion exchange
pH		1) ion exchange	pH adjustment (also stabilizes biological activity)

Choice of the initial step is especially complex and is, therefore, discussed separately (Section The Starting Step) from subsequent steps (Section Subsequent Steps).

3.3.2.2. *The Starting Step*

The choice of the starting step has a decisive effect on process economics because product losses are greatest at this stage.

Sample preparation for the first chromatographic step is summarized in Table 15. Two methods exist for preparing crude products for ion exchange and other techniques requiring low conductivity (including affinity chromatography):

1. If the product is present at very low concentration (≤10 mg/mL), the sample can simply be diluted further.
2. If the product is present at high concentration (ca. 1 g/L), gel filtration can be used for desalting [58]. This is advantageous because not only is conductivity decreased, but small charged molecules that would later reduce the available capacity of an ion exchanger can also be separated [46].

The largest volumes and amounts of protein must be processed in the first purification step. This step should be completed quickly to remove destructive proteases. A large decrease in total volume should be obtained so that the dimensions of subsequent steps and thus costs can be reduced. Consequently, only adsorption methods are usually considered. In some cases, however, desalting or rebuffering by gel filtration may be advantageous (see above); however, this leads to slight dilution and an adsorptive concentration step must follow.

The sample often contains components (e.g., lipids, nucleic acids, or pyrogens) that contaminate the gel material and may reduce its capacity for the product. In affinity media with protein ligands, dangerous contaminants also include proteases. Consequently, the gels used at the start are usually robust materials that can be regenerated even after the heaviest contamination.

Affinity chromatography offers the maximum selectivity and a high capacity for the product, with little interference from foreign proteins. Therefore, it is often chosen as the starting step. Some affinity matrices, such as Chelating Sepharose [59], Lysine Sepharose, or Arginine Sepharose, possess a broad spectrum of applications and are relatively inexpensive. They have been used successfully at the start of purification procedures [60]. The simple ligand in these gels is chemically stable and coupled to the matrix in such a way that the material can be regenerated and sterilized even after heavy sample loading. Samples with high salt content can also be processed by chromatography with immobilized metal ions (Chelating Sepharose).

However, many affinity media are relatively labile and thus unsuitable for the initial step. Prior ion-exchange or hydrophobic interaction chromatography then offers some advantages, despite a possibly lower selectivity.

Ion-exchange chromatography with anion exchangers allows lipids, nucleic acids, and even bacterial endotoxins (lipopolysaccharides) to be bound more strongly than many proteins. These contaminants can thus often be removed from large volumes of crude material and the product can be concentrated simultaneously. Despite the resultant heavy contamination, modern ion exchangers have prolonged lifetimes. Proteases and other impurities can then be removed in the subsequent step, possibly even with the same ion exchanger but with different conditions, e.g., fractionation mode.

Thus, ion-exchange chromatography is by far the most common starting method for purification. The ionic strength of material from cell cultures or natural products is usually so high that conductivities markedly above 5 mS/cm are measured in the crude product (Table 16). The binding of products with conductivities above 5 mS/cm to an ion exchanger presents problems if, as is desirable for selective separation, the working pH is close to the isoelectric point (Section Experimental Data). The ionic strength is thus usually reduced (see above).

Table 16. Approximate conductivity of cell cultures and natural products [61]

Material	Conductivity, mS/cm
Tissue, plasma, body fluids, animal cell culture medium	15
Bacterial cultures	30
Yeast cultures	20

Depending on selectivity, distinction is made between the concentration mode and the fractionation mode in ion-exchange chromatography. If product concentration is very low (≤10 mg/L), the *concentration mode* is used; here, conditions are chosen such that the sample binds strongly and rapidly to the carrier even at very high flow rates (200 – 400 cm/h) where selectivity may be reduced [62]. The product can then be concentrated to a small volume at a lower flow rate (10 – 50 cm/h). Reversing the direction of flow is often advisable to achieve a higher concentration [63].

With higher product concentrations (>10 mg/L), significant purification can be achieved even in the starting step. The *fractionation mode* is then used in which the best possible selectivity is obtained by fine adjustment of pH and ionic strength.

The protein content is greatest in the starting step, although the ratio between total protein and product can vary greatly (1 : 1 to 100 : 1). If only a small proportion of product is present, conditions (pH, etc.) should be adjusted so that it can be preferentially adsorbed and selectively eluted. If the product and foreign substances are present in similar proportions, tests should be carried out to see whether adsorption of impurities rather than adsorption of product would be preferable. The product is then almost immediately available for the next step; the conductivity and pH remain unchanged, and the same buffer can be used in the second step (see Section Subsequent Steps).

Hydrophobic interaction chromatography provides a suitable introductory purification step if the crude product has a particularly high conductivity, e.g., after addition of more than 40 % ammonium sulfate for precipitation. The ammonium sulfate concentration should be chosen so that the product remains in the supernatant. The supernatant is then used directly as the sample for selective HIC adsorption and further purification of the product (e.g., [64]).

With a suitable combination of temperature and ionic strength, proteases can be separated selectively from other proteins and peptides if a hydrophobic matrix is used as the adsorbent [65]. Because hydrophobic interaction chromatography can operate with very large volumes and amounts of protein, and concentration can be achieved, chemically stable HIC media offer a good alternative to ion exchange as the initial step.

3.3.2.3. *Subsequent Steps*

In the starting step, the crude product is concentrated, prepurified, or merely rebuffered and freed from low molecular mass impurities. This provides an intermediate product with properties that depend on the initial purification rather than on the raw material. Subsequent steps are thus governed by this factor. Figure 13 shows some logical purification sequences; see also Tables 4 and 14.

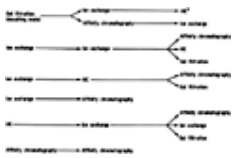

Figure 13. Logical sequences of chromatographic purification steps

* HIC = hydrophobic interaction chromatography.

The choice is simplest if purification starts with desalting by gel filtration. In this case, 20 – 25 % of the column volume is applied as a sample. The intermediate product is diluted by a factor of ca. 1.4 so that a subsequent concentration step is required. Because ionic strength or pH can be adjusted as desired with the aid of gel filtration, the correct starting conditions for the next step can be obtained automatically, as in the purification of monoclonal antibodies shown in Figure 14 [46].

Figure 14. Purification of monoclonal antibodies from serum in three steps

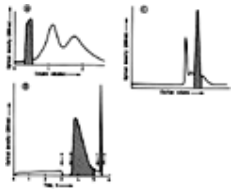

Hatching denotes fractions containing monoclonal antibodies.

A) Desalting with Sephadex G-25: buffer 20 mmol/L sodium citrate, pH 5.3

B) Ion exchange with S Sepharose Fast Flow; bed height: 3 cm; linear flow rate: 300 cm/h (loading and washing), 5 – 10 cm/h (elution)

Buffers: a) 20 mmol/L sodium citrate, pH 5.3; b) 20 mmol/L sodium citrate, 140 mmol/L NaCl, pH 5; c) 100 mmol/L NaOH

C) Gel filtration with Sephacryl S-200 HR; bed height: 100 cm; linear flow rate: 7.5 cm/h; buffer 50 mmol/L Tris – HCl, 150 mmol NaCl, pH 7.5

If the total sample volume is greatly reduced in the first step by ion exchange in the concentration mode, the second ion-exchange step should be used for fine purification. In some cases, the same exchanger is suitable; its pH must then be modified to give greater selectivity. A small change in ionic strength determines adsorption and elution of the product. Because the concentrated intermediate product has a high ionic strength, it can also be purified by HIC: if necessary, the salt concentration can be increased even further.

If the crude product is relatively concentrated and good purification is obtained by ion exchange in the selective fractionation mode, a highly enriched product can be obtained in the second step. In the case of human interleukin 1 (IL-1), for example, anion and cation exchangers can be used consecutively [66]. The culture supernatant is first purified on a strong cation exchanger which binds IL-1. Elution is achieved by using the buffer needed for the second step. The intermediate product can be applied directly to a DEAE exchanger, and IL-1 is again bound. Elution is carried out selectively by increasing the ionic strength.

In all the above examples, a highly selective affinity chromatography step can also be included, generally after dilution to reduce ionic strength. If proteases, lipids, and other interfering constituents have already been removed, hardly any restrictions are imposed by the lability of the affinity medium. Two consecutive affinity steps often give a homogeneous end product. This route has been selected for the purification of tissue plasminogen activator (tPA) [60].

In the case of therapeutic agents, product oligomers or polymers often have to be separated. This is achieved by using gel filtration in the fractionation mode. In all cases in which particle size is an indispensable separation criterion, gel filtration should come at the end of purification, after all foreign substances have been removed. A buffer in which the product can be brought to its marketable form (formulation) should be used for this purpose.

[Top of Page]

4. Optimization

After individual methods and gel materials have been screened and a logical sequence of purification steps has been developed, the process should be optimized.

4.1. Purpose of Optimization

The product should be prepared as simply and economically as possible in a specific amount dictated by the market volume.

Costs must be optimized to ensure the economical utilization of chromatography in commercial use. In analytical chromatography, this is done by increasing the number of analyses per unit time and reducing the amount of sample required. The trend in this field is toward more efficient column materials with small particle sizes that give the maximum number of peaks, i.e., maximum resolution.

In contrast, the main purpose of optimization in preparative chromatography is to increase the productivity of each step, i.e., to optimize the amount of product obtained per unit time with the desired purity and biological activity. This is achieved by increasing the amount of product processed per run (throughput) and by increasing the proportion of product recovered (yield). Throughput is increased by optimizing both sample loading and linear flow rate, i.e., dynamic capacity in the case of adsorptive techniques. To increase yield, buffer solutions must be optimized because their composition and pH affect stability. Gel materials without adverse side effects (e.g., nonspecific adsorption) must be selected. Care should be taken to achieve early elimination of proteases that may affect product stability. Yields are also improved by shortening storage time, i.e., processing as fast as possible. Thus, using the correct dimensions in each step (see Chap. Scaling Up) and minimizing the number of steps by maintaining a logical sequence (Section Logical Sequence of Chromatographic Steps) are vitally important.

In contrast to analytical chromatography, the aim of preparative chromatography is not to resolve the sample into as many peaks as possible, but to exploit the selectivity of a chromatographic medium to separate groups of substances. A preparative chromatogram often shows only two or three distinguishable peaks. As illustrated in Figure 15, one or two peaks containing as much foreign material as possible are separated from a single product peak. Only in difficult separation problems requiring a high final degree of purity is the resolution of single substances desired as is the case in analysis. However, production of several substances from the same run may be required.

Figure 15. Preparative chromatograms for adsorptive techniques (A, B) and gel filtration (C)

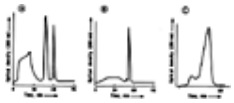

These considerations accelerate processing and simplify the management (i.e., automation) and design of a production plant.

4.2. Optimization Parameters

Despite extensive investigation of aspects of chromatography such as thermodynamics, mass transfer, and kinetics, the chromatography of proteins remains largely empirical and is based on the experience of the user. This is mainly due to the complex properties of macromolecules and their interactions with chromatography media.

In addition, simple theoretical adsorption – desorption models cannot be employed in process-scale chromatography because permanent overloading occurs [52]. Nevertheless, certain parameters (resolution and throughput) can be used to give a theoretical description of optimization and to deduce a practical approach. Some of these factors are summarized in Table 17.

Table 17. Dependence of achievable resolution and throughput on various parameters (adapted from [67])

Parameter	Resolution	Throughput
Column length, *L*	$\sim \sqrt{L}$ under isocratic conditions (i.e., constant composition of elution buffer as in gel filtration)	~1/L
Column radius, *r*	depends on design (i.e., sample distribution at top of column)	$\sim r^2$
Temperature, *T*	higher temperature mass transfer favors	~*T*
Viscosity, η	mass transfer lower viscosity favors	~1/ η
Sample volume, *V*	has an optimum under isocratic conditions (e.g., gel filtration)	~ *V*
Flow rate, *J*	has an optimum, limited by the rate of achieving equilibrium (molecular size) and peak tailing	~ *J*

4.2.1. *Resolution*

The resolution (R_s) describes how well two substances can be separated from each other. In a chromatogram, resolution is defined as the distance between two peak maxima, i.e., retention volumes V_{e2} and V_{e1}, divided by the mean of the peak widths W_1 and W_2 at the base line (Fig. 16):

$$R_s = \frac{V_{e1} - V_{e2}}{1/2(W_1 + W_2)}$$

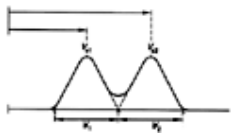

Figure 16. Symbols used in the definition of resolution R_s

The numerator is a function of selectivity, which depends on the separation chemistry. The denominator is a function of column efficiency, which depends on the particle size of the gel.

In the threshold region for overloading, the capacity of the gel bed also becomes significant:

$$R_s = \frac{1}{4}\sqrt{N}\frac{(\alpha-1)}{\alpha} \cdot \frac{k'}{(k'+1)}$$

where α is the selectivity and *N* is the number of theoretical plates (efficiency). The parameter *k′* is the capacity factor

$$k' = (V_e - V_0)/V_0$$

where V_e is the elution volume and V_0 the void volume.

In a purification scheme, each step has minimal resolution. Better resolution can increase the yield if fractionation is improved and less (contaminated) product has to be discarded.

In practice, the purification factor (i.e., the increase in the ratio of product to total protein obtained in a given step) is used,

rather than resolution, as a control parameter for optimization.

Selectivity describes how well a method distinguishes between two different substances. In preparative chromatography, increased resolution is achieved primarily by increasing the selectivity (Fig. 10); this can be recognized in chromatograms by a greater distance between peaks, i.e., $V_{e2} - V_{e1}$ in Figure 16. The selectivity of a given gel is usually controlled by optimizing both the starting conditions (e.g., pH and conductivity) and the elution conditions (e.g., gradient shape). However, in gel filtration, selectivity is dictated by the choice of gel.

Several substances compete for the binding sites on the adsorbent in less specific adsorption methods; therefore, capacity also increases with increasing selectivity.

Efficiency (*N*, number of theoretical plates) plays a minor role in optimizing resolution in preparative chromatography. It describes how effectively the equilibrium necessary for separation is achieved on the column. The number of theoretical plates and the height equivalent to a theoretical plate (HETP) are defined in Figure 17.

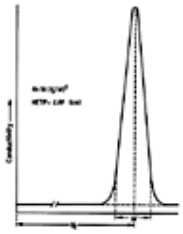

Figure 17. Calculation of the height equivalent of a theoretical plate (HETP)

Symbols: L = column length, cm; N = number of theoretical plates; V_e = elution volume, L; W = peak width, L

Efficiency is determined mainly by the particle diameter of the gel. Above a given level of efficiency, relatively little increase is obtained in the productivity of a purification step (see Section The Starting Step and Fig. 10). However, high efficiency is essential for achieving the required resolution in many complex separation problems [68].

The variable quality of gel bed packing can have a marked effect on efficiency; therefore, optimized packing instructions must be followed in preparative chromatography, and packing quality must be checked regularly. The number of theoretical plates *N* and, thus, the separation efficiency increase linearly with the square root of the column length. Therefore, with techniques such as gel filtration in which the composition of the elution buffer remains constant (i.e., isocratic techniques), resolution can be improved by using a longer column. In practice, however, column length is limited by band spreading caused by diffusion.

Capacity. The required resolution limits the available capacity since increased loading leads to peak broadening. This is especially apparent in isocratic methods, where the sample volume rarely exceeds 10 % of the column volume. In practice, therefore, a high resolution can be used to optimize capacity.

Flow Rate. Full use of the separative performance of a column with given selectivity, efficiency, and capacity can only be made if the maximal flow rate for the diffusion-dependent equilibrium is not exceeded. This optimal flow rate depends on the size of the molecules being separated, as well as on the particle size: the smaller they are, the greater is the permissible flow rate. Peak broadening due to tailing is observed if the maximum is exceeded.

4.2.2. *Throughput*

Optimization of throughput is a highly complex procedure which starts with methodical screening and gel selection. The final conditions chosen depend on economic calculations. For example, certain losses in yield may have to be tolerated to obtain higher throughput, or resolution may be improved at the expense of lower throughput to obtain the marketing advantage of a purer product.

The most effective way of increasing throughput is to increase column radius (throughput is proportional to the square of the radius, Table 17). This is the path taken in scaling up (Chap. Scaling Up). The throughput of a given column cross section is optimized by modifying the dynamic capacity and linear flow rate. Dynamic capacity denotes the possible loading under process conditions; it depends to a large extent on a linear flow rate, but also on properties of the chromatographic medium (ligand concentration on the matrix, porosity, particle diameter, and particle-size distribution) and on the compositions of the sample and the solvent. The sample composition is fixed, but the gel material, selectivity (governed by the conditions set up in the solvent), and flow rate are variables.

The gel material selected during screening (Section Screening of Chromatographic Methods (Method Scouting)) should provide the greatest throughput, i.e., the best combination of loading and flow rate. Among ion exchangers, for example, gel selectivity differs considerably (Fig. 18); throughput can thus be optimized by correct choice of the gel. After the gel has been selected, all the conditions that significantly affect capacity (pH, ionic strength, etc.) are then finely adjusted during the optimization process. Optimal operating conditions for one gel may give a lower dynamic capacity when applied to another medium. However, by modifying these conditions, the apparently poorer material may well yield significantly better results. A series of experiments to optimize dynamic capacity must be carried out on each basically suitable material (Section Selection Criteria). Agarose adsorbents have clear advantages in this respect (curve a, Fig. 19).

Figure 18. Selectivity of chromatographic ion-exchange media for the separation of transferrin (left-hand peak) and albumin (right-hand peak) at the same flow rate

A) DEAE ion-exchanger beads, diameter 60 µm; B) DEAE Sepharose Fast Flow, particle diameter 100 µm; C) Q Sepharose Fast Flow, particle diameter 100 µm; D) Mono Q, particle diameter 10 µm

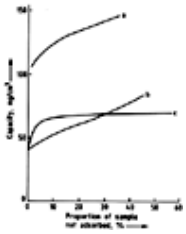

Figure 19. Dynamic capacities of DEAE anion exchangers

Results observed are due to the combined effect of matrix properties and charge density. (published with permission [69])

a) DEAE Sepharose CL-6B (highly cross-linked agarose); b) DEAE Trisacryl M (hydrophilic acrylic polymer); c) DEAE Fractogel 650 (M) (hydrophilic vinyl polymer)

In practice, the mode of operation of the purification step (e.g., concentration or fractionation) may have to be established first (i.e., is water the main contaminant?). This indicates an appropriate flow rate range (Table 18). The dynamic capacity of ion exchangers and affinity media is determined within this range on small laboratory columns by using various flow rates. In the method of CHASE and ARNOLD, a sample solution is applied continuously to the column [55], [56]. After some time, increased product concentration at the column outlet shows that the component of interest is no longer being bound (Fig. 20). The slope of the curve indicates the efficacy of adsorption; its position (elution volume) is a measure of the binding capacity under the given conditions (pH, ionic strength, etc.). Both slope and position depend greatly on flow rate (Fig. 21).

Table 18. Typical flow rates for chromatography

Chromatographic technique	Mode	Linear flow rate, cm/h
Adsorption (ion exchange,	concentration	100 – 300 *
affinity, hydrophobic interaction)	fractionation	10 – 75
Gel filtration	desalting	100 – 200
	fractionation	3 – 50

* These values are also valid for all cleaning and regeneration procedures.

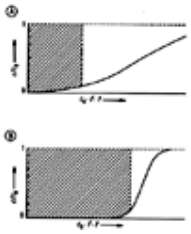

Figure 20. Calculation of dynamic capacity according to the Chase method (published with permission [55])

Low dynamic capacity results in a shallow breakthrough curve (A), high dynamic capacity gives a sharper curve (B). The ratio of protein concentration in the column outlet (c) to that in the column inlet (c_0) is plotted against the amount of protein applied ($c_0 \cdot f \cdot t$, where f = flow rate and t = time). Hatched areas represents the amount of protein bound if loading terminates when c/c_0= 0.1.

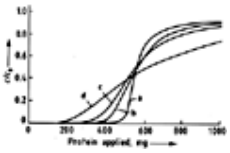

Figure 21. Dependence of the breakthrough profile on the flow rate (published courtesy of H. A. CHASE)

Flow rate, cm/h: a) 30; b) 60; c) 120; d) 240

Symbols: c = concentration of product in column outlet; c_0= concentration of product in column inlet

The objective of the purification step also determines the required selectivity, which affects not only resolution but also dynamic capacity. For a given gel material, the composition of the eluent should be determined before the dynamic capacity is examined to reduce the number of experiments required. With ion exchangers, for example, the best pH can be determined with a simple batch experiment [70].

4.3. Optimization of Ion-Exchange Chromatography

Sample load, linear flow rate, and slope or volume of the elution gradient are the most important variables in the optimization of ion-exchange chromatography. The required resolution and the degree of purity of the (intermediate)

product are fixed parameters.

If a laboratory test with a relatively small sample has given more than adequate resolution, the load can be increased until the required resolution is still just reached. The individual substances are then eluted earlier at lower ionic strengths due to displacement processes. To increase the load further, the gradient volume must be increased. Resolution increases steadily with the volume of the elution gradient [71]. Of course, this approach can be continued only until the dynamic capacity of the column material is exhausted.

With high loading, (Fig. 22 B), resolution becomes much less dependent on flow rate than with analytical loading (Fig. 22 A). Flow rate is then limited only by the gel properties.

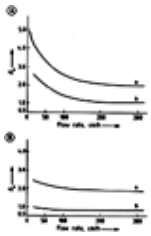

Figure 22. Optimization of ion-exchange chromatography: dependence of resolution on flow rate

A) Column loaded with 0.75 mg of protein per milliliter of gel; B) Column loaded with 12 mg of protein per milliliter of gel

Resolution was determined by using a chromatogram with three peaks: a) resolution of peaks 1 and 2; b) resolution of peaks 2 and 3. In a column with a higher load, the resolution of two substances is less dependent on flow rate.

However, because a higher flow rate increases the peak volume [71], the gradient volume must be increased further to maintain the required resolution. In the above example with a load of 0.75 mg, a flow rate of 200 cm/h requires 8.5 column volumes instead of the 5 column volumes needed with a flow rate of 100 cm/h.

Under process conditions with high loading levels, resolution should be optimized by increasing gradient volume rather than by reducing flow rate [61]. The best results are obtained by using maximal flow rate and adjusting gradient volumes to the required resolution.

4.4. Optimization of Affinity Chromatography

Selectivity, and thus both the resolution and the capacity of affinity chromatography, are determined largely by the correct choice of ligands coupled to the matrix. Optimization is therefore achieved during method screening. Because the best ligands are not always available commercially, optimization requires in-house synthesis of the medium. For economic reasons, the manufacture of material for production-scale work should be handed over to a gel manufacturer who has the equipment and experience to produce the material to specification. Optimization of the dynamic capacity of this type of material is basically similar to that described in Section Throughput.

4.5. Optimization of Hydrophobic Interaction Chromatography

As in ion-exchange chromatography, the most important variables for optimizing hydrophobic interaction chromatography are the loading conditions (especially ionic strength), the linear flow rate, and the volume of the gradient or elution step. The resolution can be improved initially by decreasing the flow rate (Fig. 23) and increasing the gradient volume (Fig. 24). Typical values for good resolution are a flow rate of 50 – 100 cm/h and a gradient of 10 column volumes. Column loading is then optimized by selecting the lowest possible ionic strength for sample application. The protein to be purified should still bind to the column, but as much of the foreign protein as possible should be eluted to increase the product-binding capacity of the column. This strategy is especially favorable for highly hydrophobic proteins that can be eluted from the column only at very low polarities.

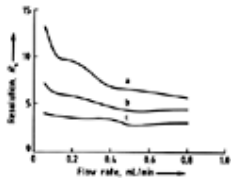

Figure 23. Optimization of hydrophobic interaction chromatography: dependence of resolution on flow rate

Calculated resolution of three pairs of proteins: a) Lysozyme – chymotrypsinogen A; b) Ribonuclease – lysozyme; c) Myoglobin – ribonuclease

Chromatography medium: Phenyl Superose HR5/5; gradient volume: 15 mL/mL of gel; sample: 50 μl of standard protein mixture containing 1 mg/mL each of lysozyme, chymotrypsinogen A, and myoglobin, and 2 mg/mL of ribonuclease; gradient buffer A: 0.1 mol/L potassium phosphate, pH 7.0, and 1.7 mol/L ammonium sulfate; gradient buffer B: 0.1 mol/L potassium phosphate pH 7.0

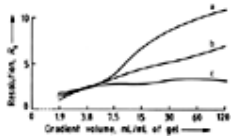

Figure 24. Optimization of hydrophobic interaction chromatography: dependence of resolution on gradient volume

Calculated resolution of three pairs of proteins: a) Lysozyme – chymotrypsinogen A; b) Ribonuclease – lysozyme; c) Myoglobin – ribonuclease

Conditions are the same as in Figure 23 except that flow rate is 0.5 mL/min.

Ammonium sulfate is commonly employed to adjust the ionic strength in hydrophobic interaction chromatography, but in some cases the use of other salts provides another means of optimizing separation. The salting-out effect of a salt, and thus its effect on protein binding and elution, can be estimated from the Hofmeister series:

(Ions are listed in the order of increasing salting-out effect)

Anions: = PO_4^{3-}, SO_4^{2-}, CH_3COO^-, Cl^-, Br^-, NO_3^-, ClO_4^-, I^-, SCN^-

Cations: = NH_4^+, Rb^+, K^+, Na^+, Cs^+, Li^+, Mg^{2+}, Ca^{2+}, Ba^{2+}

Other parameters such as temperature and pH have different effects on the hydrophobic interaction of proteins with the gel matrix and can also be varied to improve resolution. Some proteins are eluted more readily if the temperature is decreased because this weakens hydrophobic interaction.

4.6. Optimization of Gel Filtration

The most important variables in gel filtration are the linear flow rate and the column length. The required resolution and the selectivity of the chosen gel are fixed. Although flow rate has no effect on selectivity, it does affect resolution via efficiency, i.e., plate height (Fig. 25). The effect of flow rate on separation is different for substances of differing molecular mass (Fig. 26). The slow mass transfer of macromolecules leads to peak broadening and poor resolution, which can be overcome to some degree by reducing the flow rate. The minimum value for linear flow rate is about 2 – 3 cm/h, but if molecular mass differences in the sample are large (e.g., in desalting), separation can be effective even with flow rates of 100 – 200 cm/h. In the example shown in Figure 27, α-chymotrypsinogen (peak c) should be separated if the minimum flow rate of 2 cm/h (Fig. 27 A) is increased 12-fold (Fig. 27 B). Further optimization can be achieved easily by increasing the load.

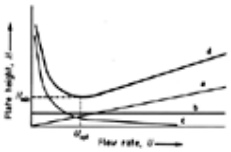

Figure 25. Optimization of gel filtration: dependence of plate height on flow rate

a) Dispersion due to incomplete mass transfer; b) Eddy diffusion effects; c) Axial diffusion effects; d) Resulting profile

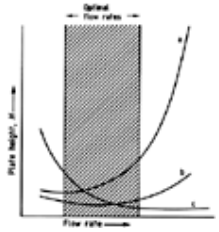

Figure 26. Optimization of gel filtration: dependence of plate height on flow rate and influence of molecular mass

a) Medium-sized protein, $M_r \approx$ 70 000; b) Small protein, $M_r \approx$ 17 000; c) Amino acid, $M_r \approx$ 200

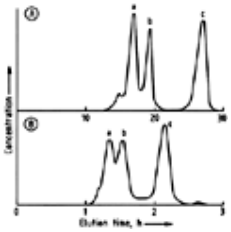

Figure 27. Gel filtration of a mixture of IgG (a), transferrin (b), and α-chymotrypsinogen (c) on Sephacryl S-200 Superfine at flow rates of 2 cm/h (A) and 25.5 cm/h (B)

Column diameter: 2.6 cm; bed height: 85 cm; eluent: 0.1 mol/L Tris – HCl buffer, pH 8.0, containing 0.5 mol/L NaCl

Throughput can be optimized further by increasing column length. Mechanically stable gel materials are required because in this strategy productivity can be increased even more by increasing the flow rate. Figure 28 shows how productivity can be increased by doubling the column length from 90 cm to 180 cm. In a 90-cm column, a resolution of ca. 1.3 is achieved at a flow rate of 10 cm/h. Doubling the column length allows twice the sample load. However, productivity remains constant provided that the flow rate is kept at 10 cm/h. (The separation on the 180-cm column takes twice as long.) An increase in linear flow rate to 40 cm/h restores the resolution to the originally desired value with a fourfold increase in productivity.

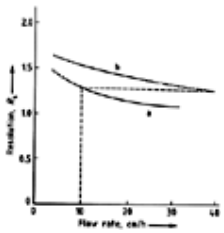

Figure 28. Flow rate and resolution in columns with a length of 90 cm (a) and 180 cm (b) (published by courtesy of J. BERGLÖF, Pharmacia)

To obtain a resolution of 1.3 in a 90-cm column (a), the flow rate may not exceed 10 cm/h. In the 180-cm column (b), the same resolution can be obtained with a flow rate of ca. 40 cm/h.

Alternatively, improved resolution (on a 180-cm column at 10 cm/h) may also be used to increase yield as a result of better fractionation.

If the mechanical stability of the separation medium limits the flow rate, the gel bed can be divided into segments to eliminate stresses due to high flow rates. Zone broadening caused by the additional column segments is partially offset by the improved packing quality of the shorter column segments.

[Top of Page]

5. Scaling Up

Scaling up denotes an increase in the volume handled in a production cycle. It applies not only to all stages of chromatographic purification but also to the fermentation and isolation stages. An essential feature of scaleup is the introduction of new, larger pieces of apparatus, whose effects on chromatographic separation must be analyzed and balanced out.

The first important factor in scaling up is the decision to invest large sums in chromatographic media and the plant for the product. Processing enters a new and critical economic phase. The correct dimensions of the full-scale plant (Section Dimensioning of a Purification Cycle) and its optimal operation are essential factors in process economics. The scaleup phase starts as soon as large amounts of product are required for test purposes, e.g., clinical trials. These amounts must be manufactured under conditions similar to those required later for full-scale production. This involves factors such as process hygiene and opens up a new dimension, namely quality assurance (see Chap. Quality Assurance).

Scaling up of chromatography is relatively simple: “only” the column diameter has to be increased, the sample load then increases proportionally. However, a number of “nonchromatographic” effects, (e.g., the introduction of new apparatus and changes in sample composition occurring when fermentation is scaled up) must be counterbalanced to ensure that the separation results achieved with the large-scale equipment are the same as those obtained with the laboratory unit [72]. Typical parameters for the preparation of mouse 2 a monoclonal antibodies in the laboratory, pilot plant, and production plant are compiled in Table 19.

Table 19. Scaling up and optimization of ion-exchange chromatography for mouse 2 a monoclonal antibodies on S Sepharose Fast Flow [a]

Parameter	Research laboratory	Pilot plant [c]	Production plant
Column volume, mL	1	16	1500
Bed height, cm	1.2	3.0	3.0
Linear flow rate, cm/h	300	300	300
Column loading, L	0.05	3.79	425
Scaleup factor [b]	1	75	8500
Elution volume, L	0.006	0.025	2.3
Yield, %	95	95	95
Efficiency, %	80 – 85	80 – 85	80 – 85
Process time, min	25	160	180

[a] Sample: culture supernatant desalted by gel filtration on Sephadex G-25.
[b] Calculated on the basis of column loading.
[c] Conditions from laboratory changed as a result of the optimization of capacity and concentration of eluted product. Note that a laboratory-scale column is still used.

5.1. Transfer Protocol

Scaling up is often carried out not by the development department but by the pilot-plant staff. All relevant processing instructions and parameters should be available in a *transfer protocol.* This ensures that the results achieved during research and development are obtained in practice without costly, time-consuming repeat studies or modification.

The process that is handed over must meet the requirements of a production process; this necessitates close communication between research and pilot-plant staff (see Table 12). The purification scheme must be complete before it is handed over to the pilot plant.

Fixed Parameters. The choice of gel is an especially important fixed parameter because a change of gel means extensive work to determine the precise separation conditions required. In some cases, all the important parameters such as dynamic capacity, flow rate, and pH would have to be redetermined. For economic reasons, this work should be carried out only on a laboratory scale. Therefore, the task of the development department is to determine, before scaling up, which media are suitable for all production stages and to hand only these materials over to the pilot plant. Only these media ensure that the other fixed parameters remain constant in practice. This strict rule for gel selection applies even if optimization of the process has not been finalized, e.g., when the precise linear flow rate still has to be established.

Scaling up in this sense, does not include cases in which separation conditions have been determined with a “high-performance” material and transition to a production medium with larger particle size is envisaged (cf. Section The Gel Bank). This transition is still part of the research and development work for gel selection. Media should be found which allow achievement or improvement of attainable throughput with simple apparatus and minimal expense (Table 19).

The height of the gel bed and the linear flow rate should also be kept constant. With adsorptive methods, the ratio of the mass of the sample to the mass of the adsorbent remains constant. In gel filtration, however, the ratio of sample zone height in the column to column length is kept constant.

Matters are simplified if the sample composition (ratio of individual constituents to total protein concentration) has been optimized for separation and does not have to be changed during scaleup. In practice, this means that definitive

optimization of a chromatographic purification scheme can be achieved only when the production conditions for cell culture or fermentation have been established. Changes in upstream processes may have a significant effect on sample composition and thus on chromatographic separation conditions [72]; they create further optimization work.

If process development under these ideal conditions is not possible, proceeding with optimization to the utmost limits of separative performance may be inadvisable. Variations in sample properties, which are common with biological materials, should be taken into account by a certain tolerance range in loading and resolution rather than by repeated optimization. Tolerance of this type is important for product reproducibility with different batches of raw material.

Variable Parameters. The major variable in scaling up is the *column cross section.* With a given bed height, column volume and thus load increase proportionally to cross section. In gradient elution, the gradient volumes must also be increased in proportion to column volume.

An increase in column cross section causes other changes. Figure 29 shows the effect of column diameter on the *linear flow rate* at a constant pressure drop through the gel bed. A significant decrease in flow rate (30 – 40 %) occurs on transition from laboratory columns 1 – 5 cm in diameter to pilot plant columns 10 – 25 cm in diameter. Flow rate decreases because the mechanical support of the gel bed by the column walls rapidly becomes insignificant as the column diameter increases. In practice, this means that during optimization, flow rate should not be increased to the threshold of pressure stability of the gel medium. However, modern Fast-Flow media allow such high flow rates even in large columns that the rate no longer has to be reduced during scaleup. The maximum possible linear flow rate for the gel is not used for real processing.

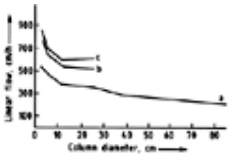

Figure 29. Dependence of attainable flow rate on column diameter at a constant pressure

a) 0.1 MPa; b) 0.2 MPa; c) 0.3 MPa

If mechanically less stable media are used (as in gel filtration), the optimal flow rate should be determined in segmented columns in the laboratory. Segmentation of the gel bed (e.g., in 15-cm segments) reduces the mechanical strain imposed on the “stacked” gel particles by the throughflow and allows high productivity even with media of low mechanical stability. This approach has been of value in the production of albumin and insulin [6]. Resolution hardly changes because the packing quality obtained in each segment is better than in a long column. A segmented arrangement also gives high productivity with materials of better mechanical stability such as Sephacryl S-200 HR (see Table 20).

Table 20. Linear flow rate and productivity in a segmented gel filtration column used for albumin production

Gel medium	Total bed height, * cm	Sample volume, % of total	Flow rate, cm/h	Cycle time, h	Productivity, g h^{-1} 1000 cm^{-2}
Sephacryl S-200 sf	90	8	15	6	72 **
Sephacryl S-200 HR	90	4	75	1.2	180 **

* Each segment has a bed height of 15 cm.
** With automation three separations can run directly after each other in one column volume.

Other variables in scaling up are the storage times of the intermediate products, and of the materials with which they come in contact. These factors influence the choice of apparatus and the dimensioning of the plant (see Section Practical Examples and Case Studies and Chap. Process-Scale Chromatography Plants). Therefore, the transfer protocol should also include information on possible product-stabilizing measures and the materials (e.g., types of steel) that may come in contact with the product.

Reliable tolerance limits for the preparation of buffers should also be established [73].

Zone Broadening. The large apparatus used in scaling up causes additional peak dilution in the sample distribution system at the top and bottom of the column, as well as in the tubing and measurement cells outside the column. This dilution or zone broadening leads to poorer, possibly intolerable resolution [74]. If laboratory results cannot be reproduced in the pilot plant, the quality of the gel bed packing should be checked first. Satisfactory results can be achieved with modern production columns. Thus, for example, an HETP value (efficiency) of 0.034 has been measured in a column of 100-cm diameter. This correlates well with the values of 0.039 and 0.033 found in smaller preparative columns (diameter 37 cm) and laboratory columns (diameter 5 cm) shown in Figure 30.

Figure 30. Zone spreading on columns with a diameter of 5 cm (· · · · ·) and 37 cm (–) packed with Sephacryl S-200 Superfine to a bed height of 15 cm (published with permission [74])

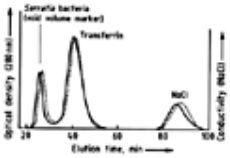

Sample volume: 2 % of total bed volume, v_t; flow rate: 10 cm/h

Peak broadening can be avoided by using suitably coordinated columns and equipment (measurement cells, etc.) in an optimal system configuration.

5.2. Practical Examples and Case Studies

Numerous examples prove that direct scaling up is possible. The production scale depends on the product use. The commonly used phrase “large-scale purification” thus includes amounts from a few grams to several kilograms of biological product per production run [75].

Figure 31 shows the results of a process development study for purifying a filtered superoxide dismutase homogenate. Economic process criteria (Tables 2 and 14) were taken into consideration, and the demands for high product purity were satisfied [58]. The purification scheme comprised three steps. The crude product was first desalted on a gel filtration column, small molecules that would affect the capacity of subsequent ion-exchange chromatography were removed simultaneously. A strong anionic exchanger, Q Sepharose Fast Flow, was used for concentration and further purification in the second step. The final fine purification by gel filtration on Sephacryl S-200 HR yielded the product in the required purity. Scaleup was achieved in a total of four steps up to a level at which 4 g of superoxide dismutase could be obtained each day.

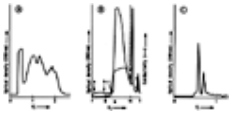

Figure 31. Development of a three-step process for purifying superoxide dismutase (published courtesy of J. BERGHOF, Pharmacia)

A) Desalting on Sephadex G-25; flow rate: 180 cm/h; buffer: 10 mmol/L Tris – HCl, pH 7.5; sample: 3500 mL; column: 25.2×43 cm (diameter × bed height)

B) Anion exchange on Q Sepharose Fast Flow; sample: 3780 mL, 12 mg/mL gel; column: 11.3×5 cm (diameter × bed height)

a) Sample; b) 10 mmol/L Tris – HCl, pH 7.5, 125 mmol/L NaCl; c) 10 mmol/L Tris – HCl, pH 2.0, 125 mmol/L NaCl; d) 10 mmol/L Tris – HCl, pH 2.0, 0.5 mol/L NaCl; e) 0.5 mol/L NaOH; f) Flow rate: 300 cm/h; g) Flow rate: 30 cm/h

C) Gel filtration on Sephacryl S-200 HR; flow rate: 30 cm/h; sample: 320 mL; column: 11.3×77.5 cm (diameter × bed height)

A similar purification scheme was developed for an anti-tPA monoclonal antibody [46]. The sequence included a desalting step, cation-exchange chromatography, and gel filtration for final purification (Fig. 14). The antibody was obtained in 95 % purity, with a yield of 75 %; ca. 2.2 mg of product was obtained from 35 L of culture supernatant. Affinity chromatography on Protein A Sepharose is an alternative to the process shown in Figure 14. Available capacity data show that a 5-L column of Protein A Sepharose would produce 100 g of antibody per cycle (i.e., in 3 h) [76]. Figure 4) shows the laboratory-scale separation. A survey of possible purification strategies for monoclonal antibodies is given in [76].

Plasma proteins have been purified on a large scale since the early 1970s [7], and processes have meanwhile been optimized for modern gel media. Figure 3) shows an example of process development for the production of albumin, together with data required for scaling up and optimal dimensioning. In the anion-exchange step, 350 g of albumin can be processed per hour per 1000 cm^2 of column cross section. In a typical production column of 80-cm diameter and 15-cm bed height, about 2.6 kg of albumin is produced per 90-min cycle. A column of the same diameter with a bed height of only 5 cm, packed with the cation exchanger CM Sepharose Fast Flow, can produce about 1.3 kg of IgG per hour (Fig. 2).

Finally, all biotechnological products such as human growth hormone or tissue plasminogen activator are purified using at least one chromatographic step.

5.3. Dimensioning of a Purification Cycle

The instructions for production processes should be developed during scaleup in the pilot plant, when considerable amounts of material are produced for testing purposes. Column dimensions must be calculated so that the planned production volume can be processed within the time available.

Basic data for this calculation are the volume of a batch of crude or intermediate product, the time interval between two batches, column length, linear flow rate, capacity of the gel bed, and the cycle time required for complete separation. This time covers all work, including any gel bed regeneration or sterilization that may be necessary. The minimum column diameter required to process the given batch volume within the available time can be calculated easily [77]. Because columns are not available in all diameters, at least the next largest size above the minimum column volume must be chosen. The actual cycle volumes and number of cycles can then be derived from the column data determined in this way.

This calculation is a simple way of obtaining an approximate value for minimum investment. The assumption is made that the plant is automated and can operate virtually on a 24-h basis. If this is not the case, correction can be made, e.g, by the arbitrary stipulation of available time.

[Top of Page]

6. Process-Scale Chromatography Plants

Planning and constructing a plant for the industrial purification of biomolecules require expert knowledge and experience in both biological sciences (biochemistry, microbiology) and engineering (process engineering, measurement, and control). Whereas experience in the chromatography of biological macromolecules stems mainly from laboratories, engineering experience is derived almost exclusively from the production areas of classical microbiology (fermentation) or chemical technology [78]. Biochemists must learn to handle construction materials such as stainless steel, which they do not normally regard as a matter for scientific consideration. Conversely, engineers must become accustomed to the relative "smallness" of process-scale chromatography systems and adapt themselves to the hygienic operation of a plant with pipework of only $\frac{1}{4}$-inch (0.64-cm) diameter or less.

6.1. Basic Design Requirements

The requirements of the development laboratory, the pilot plant, and the production plant differ.

A number of chromatography units and systems are available for laboratory operation. Some are specially designed for the purification of biological macromolecules. Examples are the Fast Protein Liquid Chromatography and BioPilot (Fig. 11) systems in which the materials used (borosilicate glass, titanium, fluorinated plastics) and the specifications (flow rate, pressure range, gradient programming, reproducibility) are matched to the requirements of the separation principles and media suitable for protein separation [79], [80]. However, transfer of a laboratory purification process to the pilot plant produces additional requirements that are not always met by these systems (see Chap. Scaling Up).

Pilot Plant. In the pilot plant, apparatus for scaling up and preparing the product for testing purposes is required [80]. Units with a throughput of between 1 – 10 L/h and 400 – 500 L/h are used in pilot studies (smaller units can be used in a laboratory). They should be able to meet the demands of different chromatographic methods (e.g., isocratic or gradient elution, monitoring parameters such as pH and conductivity). A relatively large number of fractions must be obtainable if product fractionation has not been accurately defined. In production, the purification process is usually so well-defined that only two or three fractions are taken in a very reproducible manner.

The precise dimensions of the individual process steps and their automation should be developed in the pilot plant. During the experimental phase of automation, each modification and intervention must be readily achievable. However, because the work is carried out with relatively large amounts of valuable product, safety measures must be emphasized in the event of malfunction. Disturbing external effects (e.g., failure of the power supply) should be excluded, and the control unit should be separated spatially from areas in contact with the product. Good hygiene must also be observed where necessary. All results should be documented. For these reasons, the plant for the pilot phase must be highly flexible and easy to operate.

Production Plant . Because of the special properties of each biological product and the related production process, an individual solution is generally required for the production plant. The controlled, reproducible execution of each run (i.e., plant reliability) is extremely important. Flexibility becomes less important than in the pilot plant because intervention in the process is undesirable and should be necessary only for safety reasons. However, if only a few production batches are required per year, the production unit may be a multipurpose plant with built-in flexibility for different products.

Together, the development laboratory and pilot plant must define precisely the following process parameters:

- separation media and bed height
- productivity, g h^{-1} L^{-1} (column volumes)
- linear flow rate, cm/h
- counterpressure from gel bed
- buffer types, permissible tolerance range (pH, conductivity)
- stability of sample and buffer
- material in contact with solutions
- product sensitivity (shear forces)
- yield, detection procedure
- cleaning in place protocols
- lifetime of the separation media
- volumes (product and buffers)
- number of fractions
- process control parameters
- safety measures

The requirements for flexibility, process hygiene, and the production area (e.g., buildings) form another group of criteria which must be satisfied in plant design. The plant is generally a closed system from which contamination cannot reach the surroundings and into which outside impurities cannot penetrate. In pharmaceutical processes, the plant with its packed columns can be kept sterile and pyrogen-free by cleaning in place (CIP). Sterilization in place with steam (SIP) may also be required. For this, the gel material is removed from the column and sterilized separately. Sensors such as the UV measuring cell should also be removed from the unit before sterilization.

The suitability of the plant for a purification process must finally be demonstrated by validation (Chap. Quality Assurance).

6.2. Plant Design

Large expenditure of time and money is involved in planning and selecting the apparatus for a preparative chromatography plant. Conventional builders of chemical production plants do not generally have the relevant know-how. Effective cooperation of the pharmaceutical industry primarily with manufacturers of chromatography media who already have experience in general production (especially in the pharmaceutical sector) may be required to find optimal solutions.

A chromatographic production plant can be divided into seven functional modules whose equipment can be planned more or less independently of each other (Fig. 32).

1. The first *tank module* consists of tanks for storing the sample, the elution buffers, and the regeneration buffers, and for preparing buffer solutions.
2. The *delivery module* comprises tubing to the columns, valves controlling the choice of elution buffer, pumps delivering process fluids, and safety devices that protect columns from excess pressure and the penetration of particles or air.
3. The *control module* is essentially an automated unit that is suitably programmed for the process. It controls the valves, the adjustment and regulation of flow rate, and the response to signals from monitoring devices.
4. The *separation module* is the heart of the plant and contains the packed chromatography columns in which purification occurs. The column dimensions and the linear flow rate determine the design of the delivery module and the rest of the plant. The separation module also includes the connector tubes and valves which allow flow reversal or the hygienic isolation of columns when the plant is not in operation.
5. The *monitoring module* is connected to the columns with the shortest possible tubing. This module consists of devices that monitor the course of the process, usually via UV absorbance. The module may also contain flow-through cells for measuring conductivity and pH.
6. The product is isolated in the *fractionation module*, which contains valves and supply leads to the product tanks. In automated processes, the control module regulates fractionation by evaluating UV signals and switching valves accordingly.
7. The product is collected in the second *tank module.*

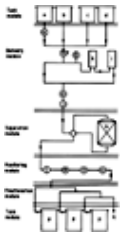

Figure 32. Schematic of a modular chromatography plant

a) Raw material tank; b) Elution buffer tank; c) Regeneration buffer tank; d) Equilibration buffer tank; e) Air sensor; f) Pump; g) Pressure indicator; h) Air trap; i) In-line filter; j) Flow indicator; k) Column; l) Flowmeter; m) pH meter; n) Conductivity meter; o) UV meter; p) Product tank; q) Waste

Control and documentation of the process are performed in a fully automatic control module (not shown).

6.2.1. *The Chromatography Column*

The column and the packed gel bed form the most important part of a chromatographic plant. Production columns made of stainless steel, glass, or plastic can be obtained in various designs, with adjustable adapters or fixed endpieces to isolate the gel bed. Diameters range from 5 to 180 cm and lengths from 5 to 100 cm. In long gel beds, problems may arise due to mechanical stress; segmented “stack” columns can then be used [74, pp. 87 – 89]. The choice is made according to process data regarding bed height, linear flow rate, and required productivity (diameter). The pressure gradient across the gel bed and the chemical specifications of the process determine the choice of construction material for the column. Glass possesses some advantage as a biocompatible, chemically resistant material. Visual inspection of packing quality, separation performance, and precipitated material on the gel bed is desired, as is possible in laboratory columns. However, only the separation of colored substances can be followed visually. Moreover, with columns more than 10 – 20 cm in diameter, only an insignificant portion of the gel bed can be inspected optically. In this case, the only effective controls of packing quality are the determination of the HETP value (Fig. 17) and the elution curve (UV absorbance). For safety reasons, especially when working at high pressure, a high-grade steel column is often preferred.

An adjustable adapter in the column head facilitates packing and allows correction if the bed volume changes. It also increases the flexibility of the column for use in different processes. The adapter should have a central device to allow adjustment and fixation for easy handling. In production columns, however, the use of fixed column ends is often more economical and hygienic. Less dead space is available in which microorganisms can accumulate. Columns can still be packed by relatively simple procedures without loss of convenience.

An essential factor in column operation is a description of the procedure for packing the gel bed with different chromatographic media; this should be issued as a set of instructions in the same way as for other plant equipment. The gel and column form an inseparable functional unit. Modern, mechanically stable media can be packed at very high flow rates in large columns just as well as in laboratory columns; they give fully analogous separation results provided the column is of optimal design (see Section Transfer Protocol). A larger, separate pump may be planned for packing. Not more than 80 % of the flow rate used for packing should be employed during the separation process.

6.2.2. *Process Monitoring*

After the columns, the devices used to monitor separation are the most important components of a chromatography plant. Sample components are generally detected in the column eluate by means of their *UV absorbance* (e.g., at 280 nm). In an automated system, these signals also control fractionation and guide the product into the correct tank. Malfunction of the UV monitor can be a very expensive disaster. Although the typical operating life of UV lamps is around 8000 h, in planning the plant continuous monitoring of the lamps should be considered [73].

The *linear flow rate* is also important for process control; it affects both resolution and productivity. The flow rate should,

therefore, be recorded continuously with flow-through cells. Units based on inductive or turbine principles are used most commonly. The signal supplied by the measurement converter can be used by a control unit to compare the actual flow rate with the programmed value. In case of deviation, the pumps can be adjusted. Deviations may, for example, arise when a sample component is deposited gradually on the gel bed or on an in-line filter. After the flow-through signal has been integrated to give the total flow volume, it can also be used to control process volumes. A fixed volume can be allocated to each process step: sample application, washing, or elution. These volume segments provide a basis for automation (Section Automation and Safety Devices). The fixed scheme can be overridden by signal-dependent steps, i.e., the valve switch can be controlled by monitor signals such as those provided by UV absorbance during fractionation.

Particularly in adsorption (ion-exchange or affinity) chromatography, *pH* and *ionic strength* are control parameters that govern reproducibility and the general success of separation. Regeneration of the gel bed can also be controlled. Both elution with a continuous gradient and fresh sample application after regeneration can be automated by evaluating the signals from flow-through pH or conductivity cells.

The required pH and ionic strength ranges in which measurements should be made must be defined. The fluids that come in contact with the cells must be known before a suitable item of equipment can be selected. To operate reproducibly, pH sensors in particular should be recalibrated before each run.

6.2.3. *Pumps and Valves*

Biological macromolecules are often sensitive to strong mechanical or thermal effects, which may arise in pumps and valves. The choice of a suitable pump is one of the least predictable tasks in plant design. Experience with the product determines the type of pump drive used. For example, gear pumps cause problems in the production of human albumin. Pumps suitable for process-scale chromatography include peristaltic pumps, double-membrane pumps, and some rotor pumps. In addition to product-specific restrictions, the required flow rate, pressure range, degree of suction, throughput, and viscosity of the pumped medium are the major selection criteria. Because of the chemical composition of the liquids, special materials must be selected for pump components such as ball valves, membranes, and seals.

In closed systems, the tanks are connected to the pumps and column outlets via valves. Solenoid or pneumatically controlled valves are generally used. Three-way valves have the advantage of reducing the total volume of the tubing system; this is particularly important in the fractionation module (see below). However, valves of this type are not available, which meet the hygienic demands of a pharmaceutical process. Butterfly valves with flaps or plug valves are used. Two-way valves, especially membrane valves, are preferable from a hygienic point of view. The disadvantage of a rather larger tubing volume is of little importance in the supply leads and can be offset by appropriate design measures. Apart from in-line filters and flow-through cells, valves create the greatest flow resistance in the plant and thus may need to be optimized [74].

Malfunction is highly undesirable in valves that direct the product into the appropriate tank during fractionation. Therefore, monitoring of the proper position of the valve and appropriate reaction of the control unit may be required in case of valve malfunction.

6.2.4. *Automation and Safety Devices*

Automation of the chromatographic process is particularly desirable when numerous cycles must be run in succession to achieve high productivity and plant utilization (Fig. 33). Because automation allows several process parameters to be monitored simultaneously, it increases reliability and reproducibility. The fact that variations in the quality of biological materials make human supervision indispensable does not contradict this. Electronic data processing affords numerous solutions for the automation of production processes, some of which differ greatly in relation to performance, user-friendliness, start-up time, and cost [81].

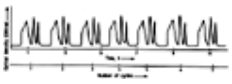

Figure 33. Increasing productivity through automation in the anion-exchange chromatography of plasma on DEAE Sepharose Fast Flow

Column diameter: 11.3 cm; bed height: 15 cm; flow rate: 200 mL/min at 20 kPa; sample: 2.9 L of desalted bovine plasma (87 g of protein per cycle)

Chromatographic separation is less troublesome to control than fermentation because relatively few parameters must be monitored. A proven concept for the automation of process chromatography thus uses control devices that have fixed, specially designed, built-in software and hardware. Programming of these units can be learned and mastered quickly even by laymen. Process data such as flow rate or volume breakpoints for process control (Section Process Monitoring) are simply entered into an existing program. Programming is user-friendly and is based on a menu driven dialogue procedure. The hardware of these “dedicated controllers” [81] can control a chromatographic plant with the aid of 10 – 20 programmable outputs for valve switching and for the power supply to peripheral apparatus. A pump can be regulated and signals from up to four monitors can be evaluated for control purposes. In addition, sensors can be included which, for example, give an alarm signal that switches off the process if the pressure rises, the pH exceeds a preset level, or a valve malfunctions. These alarms may also produce more flexible responses such as reduction of flow. Table 21 summarizes some points to be considered in process control and monitoring, including the necessity for an alarm signal.

Table 21. Crucial parameters in process control and monitoring of process-scale chromatography

Parameters	Precautions	Alarm function required

pH	continuous monitoring, defined limits	+
Conductivity	continuous monitoring, defined limits	+
Feedstock concentration	off-line analysis, defined limits	
Air presence in the column	continuous monitoring, in-line air trap	+
Tank levels	level sensors	+
Flow rate	continuous monitoring	+
Pressure	continuous monitoring	+
Column	operation within given specifications, in-line filter, feedstock preparation	
Product monitors	signal level control, off-line calibration	+
Pump	continuous monitoring of flow rate and pressure	+
Valves	continuous monitoring of position	+

Despite automation, provision must be made for manual operation for reasons of safety and economy if the control apparatus fails. Because all parameters of a production process need documenting as well as monitoring, the control unit should be able to store data or direct it to a superior computer located elsewhere. Documentation can be performed with a personal computer, suitable software is offered by equipment manufacturers.

6.3. Effects of Apparatus on Separation

Zone broadening before separation or remixing after successful separation reduces resolution. Such effects may occur in the tubing, at the column inlet and outlet, in the flow-through monitoring cells, and in the leads of the fractionation module; they must be minimized by appropriate design. Apparatus effects are discussed in detail in [74].

The use of narrow tubing in which turbulent flow predominates reduces the peak tailing caused by laminar flow in wide tubes. Uniform distribution of flow across the column cross section then becomes the critical factor in zone spreading. In practice, $\frac{1}{4}$-inch (0.64 cm) tubes are commonly used for throughputs up to 40 L/h, ½-inch (1.27 cm) tubes up to 400 L/h, and 1-inch (2.54 cm) tubes at greater throughput. Tubing may be made from plastic (e.g., polyethylene) or stainless steel.

The sample distributing system at the column head usually consists of the inlet tube and a fine net covering the gel. Here, the axial counterpressure opposing entry of the sample into the gel must be considerably greater than the radial pressure; a high radial pressure would prevent the sample from spreading over the column cross section. The pressure gradient across the mesh covering the gel surface must be uniform. Critical factors in zone spreading at the column inlet are depicted in Figure 34. The mixing chamber at the bottom of the column must be as small as possible to prevent remixing. For the same reason, the total volume of the tubing and measurement cells between the columns and fractionation unit must be minimized. Dead spaces from which contaminants can reach the purified product must be avoided, especially after the column.

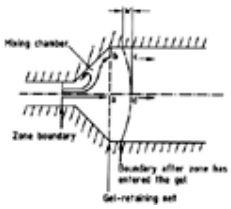

Figure 34. Zone spreading at the column inlet (published with permission [74])

To minimize zone spreading, transport times to points a and b should be the same; pressures at a and b should also be the same so that equal flow rates are obtained at points c and d. The mixing chamber should be as small as possible to prevent eddying and back-mixing. The boundary width *w* should be small compared to the peak width caused by dispersion in the column.

Another cause of zone broadening may be pump pulsation, which can largely be eliminated by a pulsation-dampening device such as a bubble trap. Optimal results can then be achieved even with pumps that exhibit relatively marked pulsation (e.g., double-membrane pumps).

[Top of Page]

7. Quality Assurance

Chromatography is an effective purification method for routine production of a number of hormones and therapeutic enzymes from both conventional and genetic engineering sources [7], [75]. The quality assurance of these products is one of the most important tasks to be met before drugs produced by modern biotechnology are generally accepted. Some of the aspects relevant to process-scale chromatography are summarized in Figure 35.

Figure 35. Factors involved in assuring end-product quality

Quality assurance is achieved by the purification process itself (i.e., by working under aseptic conditions), by appropriate design of the production plant, and by the properties of the chromatographic media. Together, these features should ensure reproducible operation. Furthermore, tests of biochemical and biological purity are carried out on the product.

Good Manufacturing Practice. Like other process steps, process-scale chromatography must comply with the extensive safety requirements of organizations such as the Food and Drug Administration (FDA) in the United States. Development and production must constantly be adapted to conform to new regulations, such as those specified in “current Good Manufacturing Practice” (cGMP) and other official recommendations [49], [82-90]. Good manufacturing practice means that documentary evidence is kept to ensure reliable, safe operation. Quality standards are defined and converted into measurable parameters. Records must be kept to show that standards are being upheld. Despite the expense and effort involved, the production process must remain economical.

Regulations for GMP are flexible when new, generally applicable results may lead to the adoption of novel procedures and when special product properties must be considered. End-product control and in-process control of critical steps in production are required. A great deal of attention is paid to processing aids, e.g., components of the culture medium in cell culture or the gel medium in chromatography. Their specifications, processing behavior, and suitability must be documented if possible with the help of the supplier. However, the user (i.e., the producer) bears ultimate responsibility. The use of substances with intrinsic biological activity, toxicity, or pyrogenicity should be avoided.

In process-scale chromatography, the purification process must be described in detail. Reproducibility with respect to the removal of special contaminants (e.g., pyrogens) or the composition and activity of the end product must be demonstrated by appropriate analytical methods. The chromatographic medium must not release harmful contaminants into the product. If foreign substances can be tolerated in the end product, their maximum limits should be defined and their levels monitored. With DNA and viruses, purification factors for the entire process and for its individual steps should be given.

Chromatography is performed aseptically. The process description must, therefore, include CIP measures which ensure that contaminants can be removed and that a predefined standard of hygiene is maintained. Hygiene in process-scale chromatography starts with careful, regular maintenance of the gel materials. The methods employed depend on the nature of the sample responsible for contamination. The use of sodium hydroxide solution is the method of choice not only for gel maintenance but also for the inactivation of viruses, microorganisms, and pyrogens. The solution sterilizes the system and washes out contaminants.

Suitability of the Chromatographic Plant. After specifying all the necessary requirements, the chromatographic purification plant is planned, built, and finally installed. The first phase of suitability testing begins with the commissioning of the plant, which is carried out by the user together with the manufacturer. Criteria for testing the equipment, the calibration of inbuilt instruments, and suitability must be established. Suitability tests may, for example, include the introduction of malfunctions and testing the emergency stop function. When this hardware-oriented testing phase is complete, validation of the entire process starts, i.e., “the systematic testing of the essential operating steps and equipment in development and production, including the control of pharmaceutical products, with the aim of ensuring that the product can be manufactured reliably and reproducibly in the required quality when the stipulated production and control processes are applied” [91]. Documentation is required to prove that the equipment is designed to maintain a predetermined level of process hygiene and allow the process to function reproducibly at a predetermined level. Challenge tests (i.e., setting the processing pH to the limiting value defined for reproducible separation) and “worst case” assumptions are made during validation.

Suitability of the Chromatographic Media. Media for process-scale chromatography are one of the most critical components in a purification process; they must function reproducibly in the plant and be compatible with CIP measures. Validation of their suitability must always be undertaken with regard to the system as a whole.

Like plant evaluation, characterization of the medium (i.e., the determination of process-relevant specifications) is very important. The integrity of the gel matrix should be tested in the presence of typical chemicals used for cleaning and sterilization.

A test program for the Pharmacia ion-exchange media based on the Sepharose Fast Flow matrix follows:

1. Autoclave five times for 30 min at 121 °C and pH 7.0
2. Maintain at 40 °C for one week in
   - Phosphate buffer, pH 7.0, 50 mmol/L
   - Acetic acid, 1 mol/L
   - Sodium hydroxide, 1 mol/L
   - Urea, 8 mol/L
   - Guanidine hydrochloride, 8 mol/L
   - Ethanol, 24 vol %

After treatment, all media must pass quality control and function tests.

Studies of this type should already have been carried out by the gel manufacturer. Deterioration of separative performance should not occur even after prolonged exposure (at least one week) to the test substances. This can be expressed in weeks of constant exposure or in number of cleaning cycles (see Fig. 9).

A small-scale separation should be performed under the original processing conditions before a new batch of separation medium is used in production. This step has to be taken by the user of the media.

Sterile and Aseptic Operation. For genetic engineering products, the raw material is often of microbial origin and the

product possesses very limited stability under sterilizing conditions. Therefore, the process is often aseptic rather than fully sterile. However, measures must be available to allow regular sterilization of the plant and ensure a sterile end product. Relevant FDA requirements are given in [83].

Good process hygiene not only reduces the risk of product contamination but also improves the economic efficiency of the process, for example by increasing the lifetime of gel media. Separation media employed in plasma fractionation have been in use for as long as five years. With immunoadsorbents, 200 production cycles have been achieved [92]; modern, rigid ion exchangers have undergone as many as 600 cycles in the purification of albumin [61].

Good process hygiene is achieved by the use of pyrogen-free water of the same quality as “water for injection purposes” (United States Pharmacopeia, USP XX), which can be obtained by distillation or reverse osmosis. All buffer and cleaning solutions and, if possible, the sample should be sterile-filtered (filter pore diameter 0.22 µm). The production room should have a clean air supply, and personnel must wear clean protective clothing and caps. The columns and tubing system should be rinsed from time to time when the plant is not in use. Slow flow rates can be used, and the volume of rinsing liquid required can be reduced by recycling and filtration. Columns should be flushed with a cleaning solution (e.g., 1 mol/L sodium chloride, 0.1 mol/L sodium hydroxide) at regular intervals during prolonged downtimes [93].

Maintenance of Gel Materials. Process hygiene in chromatography starts with thorough, regular care of the gel materials by methods that depend on the nature of the contaminants. In general, the frequency of cleaning increases the “closer” the purification step in question is to the crude material. The properties of various crude materials give important hints for planning a maintenance protocol [94].

Plant extracts normally contain little lipid material but large amounts of charged polyphenols with varying molecular mass as well as colored hydrophobic substances. Polyphenols can cause decomposition processes to occur on the surface of anion exchangers where they accumulate. Samples of animal origin, including plasma, are rich in lipids which may bind to proteins. The proteins and lipids may be adsorbed on the matrix surface and can, in turn, bind hydrophobic proteins. This effect is sometimes observed as a gray zone in the gel bed and leads to an increase in column counterpressure. Secretory microbial systems yield a sample containing few lipids and proteins, but they can still cause problems if an antifoaming agent has been added to the culture medium. Nonsecretory microbial systems furnish a crude, lipid-rich product after the breakdown of cell membranes. In mammalian cell expression systems, the product is usually secreted into the medium. Because of the relatively low concentration of product, the components of the culture medium constitute the major impurities in cell cultures. Cultures of animal cells cause relatively few problems due to decomposition or other effects in chromatography columns.

Individual chromatographic procedures have different resistances to the stresses described above. The maintenance protocol shown in Table 22 generally removes both lipids and proteins from the gel to a satisfactory degree. Ion exchangers are frequently used as the first step in a purification procedure. The most serious problem in this situation is caused by the binding of phospholipids to charged groups; anion exchangers are most affected because the phospholipids are generally negatively charged. The bound lipids in turn bind neutral fats and hydrophobic proteins.

Table 22. Basic protocol for cleaning in place (CIP) of chromatography columns

Step	Aim	Washing solution *	Column volumes used
1	removal of proteins bound by strong ionic bonds	filtered sodium chloride, 2 mol/L	0.5
2	removal of proteins or lipids bound by strong hydrophobic bonds	sodium hydroxide, 1 mol/L	4
3	removal of precipitated protein	acetic acid, 75 vol %	2
4	removal of lipids and highly hydrophobic proteins	ethanol, 40 vol %	10

* Column is washed with water or buffer between steps. Flow rate is ca. 50 cm/h but depends on the gel matrix.

Among hydrophobic separation media, phenyl and butyl gels with their relatively low hydrophobicity present few maintenance problems. Octyl Sepharose and similar gels are more difficult to clean. Temperature reduction may help by reducing binding forces by 20 – 30 %. Because gel filtration concludes the purification scheme, samples are normally very clean and cause no problems. In general, a small amount of precipitated protein can be removed with 1 – 2 mol/L sodium chloride solutions. Lipids can generally be removed by combining temperature elevation to 40 – 60 °C with other measures such as rinsing with sodium hydroxide solution.

Sterilization and Removal of Pyrogens. The fear of contamination of biological products derived from plasma or mammalian cell cultures is considerable because these sources may be infected, for example, with hepatitis or AIDS (acquired immune deficiency syndrome) viruses.

Fortunately, nucleic acids and other hazardous contaminants such as lipopolysaccharides bind very effectively to anion exchangers [95], [96]. Thus, they can be removed by an early ion-exchange step. Effective decontamination methods for inactivating adsorbed contaminants are an essential prerequisite in the manufacture of modern biotechnical products for therapeutic use. Autoclaving, a bacteriostatic agent, or sodium hydroxide is generally used; their advantages and disadvantages are listed in Table 23. Sodium hydroxide is the best agent for inactivating viruses, bacteria, and their products (spores and endotoxins) in chromatography [97]. For example, HTLV is inactivated by 0.1 mol/L sodium

hydroxide within 10 min at room temperature [98]. The entire chromatographic system can be sterilized and cleaned simultaneously with sodium hydroxide solution. Pyrogens are degraded and flushed out. Unlike bacteriostatics, for example, no risk of product contamination exists.

Table 23. Advantages (+) and disadvantages (–) of decontamination methods for inactivating adsorbed contaminants

Autoclaving	Bacteriostatics	Sodium hydroxide solution
+ gives sterility	+ result in a low bacterial level (high sterility)	+ sterilizes complete column
– has no cleaning effect	– only possible with certain affinity media	systems
– does not destroy endotoxins	– do not remove endotoxins	+ has good cleaning effect
– is not possible with packed	– have no cleaning effect	+ destroys pyrogens
columns	– lead to risk of contaminating traces in the	+ cannot contaminate the
	product	product
		– requires careful handling

The action of sodium hydroxide depends on its concentration and temperature. Results for *Pseudomonas aeruginosa* cells and *Bacillus subtilis* spores are shown in Table 24. Other microorganisms tested include *Staphylococcus aureus, Escherichia coli*, the yeast *Candida albicans*, and the fungus *Aspergillus niger* [61]. These studies concluded that 0.1 mol/L sodium hydroxide is an effective disinfectant for bacteria and yeast; a concentration of only 0.01 mol/L is required for gram-negative bacteria. However, 0.5 – 1 mol/L alkali is required to reliably kill fungi [61]. Endotoxins can also be degraded by sodium hydroxide in a relatively short time (Fig. 36).

Table 24. Concentration and time required for sodium hydroxide to inactivate microorganisms

Microorganism	NaOH concentration, mol/L	Temperature, °C	Time, h
Pseudomonas	0.15	ca. 20	16
aeruginosa[*]	0.25	ca. 20	16
	0.5	ca. 20	5
	1.0	ca. 20	4
Bacillus subtilis	0.5	4	76
spores [99]	0.5	20	8
	1.0	4	22
	1.0	20	2

[*] Data courtesy of F. HASKO, Institute of Haematology, Budapest, Hungary.

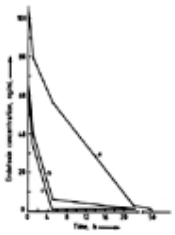

Figure 36. Inactivation of *Escherichia coli* endotoxin by sodium hydroxide at room temperature

Sodium hydroxide concentration, mol/L: a) 0.1; b) 0.5; c) 1.0

The above discussion shows that the selection of separation media which can withstand relatively high concentrations of sodium hydroxide is important. However, affinity media are an exception because the appropriate CIP procedure is dictated by their ligands. Alternative cleaning methods involve killing bacteria with 70 vol % ethanol and solubilizing pyrogens with 4 mol/L urea. Formaldehyde or phenol has also been used to clean immunoadsorbents without loss of activity of the separation medium.

[Top of Page]

8. Developments

Process chromatography is a relatively young branch of biotechnology. Significant developments can, therefore, be expected.

One particularly weak point of all currently available chromatographic methods in the initial purification step is the need to free the crude product from particulate components. The isolation steps used for this (⟶ Biochemical Separations – Product Isolation) are in some cases difficult to operate on a large scale (e.g., centrifugation with its risk of aerosol formation) or otherwise unsuitable for simple, direct scaling up (e.g., many membrane techniques). In addition, this step is time-consuming and decreases the yield of product.

A “chromatographic link” between cell cultures and final purification is obtained by using adsorbents (ion exchange) with very large particle diameters (>150 µm) that are based on highly cross-linked agarose. They allow isolation and concentration of products at very high flow rates (13 – 18 m/h) even with only partially filtered crude product solutions. Conventional chromatographic equipment can be used; columns do not become blocked; and the gel bed can be regenerated relatively easily because of the high chemical stability of matrices such as cross-linked agarose.

The technology of gel manufacture is advancing. Tailor-made separation media will be available for specific applications, e.g., special affinity media, ion exchangers with a high capacity for small molecules, and gel filtration media with special separation ranges suitable for monoclonal antibodies or recombinant DNA products.

The chemical – mechanical stabilities and capacities of many affinity media will be increased by the use of highly cross-linked matrices.

The oldest, dextran-based gel filtration media still have the highest selectivity but relatively poor mechanical and chemical stability. Attempts to produce highly cross-linked macroporous matrices for gel filtration have only improved stability; therefore, a compromise must be made between selectivity and adsorption behavior. A new generation of gel filtration media based on agarose – dextran mixed polymers combines the stability required for process media with the very high selectivity of the old materials.

In contrast with the trend toward tailor-made separation media for specific processes, an alternative approach modifies individual products for a standardized purification scheme; the aim is to use a single, highly specific affinity material to purify many products. This is conceivable on the basis of work with fusion proteins. A gene coding for the desired product is fused with the gene for a fragment of protein A and expression yields a fusion protein with a protein A marker that can bind to IgG Sepharose Fast Flow. This affinity medium would be suitable for all products fused with protein A [100], [101]. However, a number of problems (e.g., subsequent cleavage of the marker in each case) must still be solved.

[Top of Page]

References

1. E. A. Peterson, H. A. Sober, *J. Am. Chem. Soc.* **78** (1956) 751 – 755.
2. J. Porath, P. Flodin, *Nature (London)* **183** (1959) 1657 – 1659.
3. R. Axén, J. Porath, S. Ernback, *Nature (London)* **214** (1967) 1302 – 1304.
4. J. Porath, J.-C. Janson, T. Laas, *J. Chromatogr.* **60** (1971) 167 – 177.
5. L. Söderberg *et al.*: “Protides of the Biological Fluids,” *Proc. Colloq. AMPERE* **30** (1983) 629 – 634.
6. J.-C. Janson, *J. Agric. Food Chem.* **19** (1971) 581 – 588.
7. J. M. Curling: *Methods of Plasma Protein Fractionation*, Academic Press, London 1980.
8. Hybridoma Committee of the office of Biologics (FDA): *Points to Consider in the Manufacture and Testing of Monoclonal Antibody Products for Human Use*, June 1, 1987, pp. 3 – 4.
9. R. Scopes: *Protein Purification, Principles and Practice*, Springer Verlag, New York 1982.
10. J. Bonnerjea, S. Oh, M. Hoare, P. Dunnill, *Bio/Technology* **4** (1986) 954 – 958.
11. J. L. Dwyer, *Bio/Technology* **2** (1984) 957 – 964.
12. Y. D. Clonis, *Bio/Technology* **5** (1987) 1290. J.-C. Janson, *Trends Biotechnol.* **2** (1984) 31 – 38.
13. B. Vaks, Y. Mory, J. U. Pederson, O. Horovitz, *Biotechnol. Lett.* **6** (1984) 621 – 626.
14. E. Sulkowski, *Trends Biotechnol.* **3** (1985) 1 – 6.
15. F. Quadri, *Trends Biotechnol.* **3** (1985) 7 – 11.
16. H. A. Artmann, S. Shaltiel, *Eur. J. Biochem.* **70** (1976) 269.
17. P. Steffens: “Hydrophobe Interaktionschromatographie,” *Pharmacia Biotechnology aktuell (2)*, Pharmacia, Freiburg (1987) 2 – 5.
18. J. L. Fausnaugh, L. A. Kennedy, F. E. Regnier, *J. Chromatogr.* **317** (1984) 141 – 155.
19. P. Steffens: “Reversed Phase Chromatography (RPC),” *Pharmacia Biotechnology aktuell* (1), 1986, 3 – 8.
20. K. B. A. Larsson: *5th International Symposium on HPLC of Proteins, Peptides and Polynucleotides*, Toronto, Ontario, Canada 1985.
21. P. Hedman, *Int. Biotechnology Lab.*, May/June 1984, pp. 36 – 42.
22. E.-L. Winnacker: *Gene and Klone*, VCH Verlagsgesellschaft, Weinheim, Germany 1985 ; From Genes to Clones, VCH Verlagsgesellschaft, Weinheim 1987.
23. J. H. Peters, H. Baumgarten, M. Schulze: *Monoklonale Antikörper*, Springer Verlag, Berlin – Heidelberg – New York – Tokyo 1985.

24. A. Pickett, J. Hardy, “Comparison of Hosts for rDNA,”. Applied Molecular Genetics, *Proc. 3rd Eur. Congr. Biotechnol.* **84** (1984) 48.
25. B. R. Glick, G. K. Whitney, *J. Industrial Microbiology* **1** (1987) 277 – 288.
26. T. Moks, L. Abrahmsén, B. Österlof, S. Josephson, *et al.*, *Bio/Technology* **5** (1987) 379 – 382.
27. P. A. Lowe, *Proceedings of Biotech.* **85** (1985) 127 – 130.
28. J. P. Gilligan, T. G. Warren, J. A. Koehn, S. D. Young, *et al.*, *Bio Chromatography* **2** (1987) 20 – 27.
29. S. M. Kingsman, A. J. Kinsman, J. Mellor, *TIBTECH* **5** (1987) 53 – 57.
30. R. C. Das, J. L. Schultz, *Biotechnol. Prog.* **3** (1987) 43 – 48.
31. P. J. Kniskern *et al.*, *Gene* **46** (1986) 135 – 141.
32. D. Cullen, S. Leong, *TIBTECH* **4** (1986) 285 – 288.
33. J. van Brunt, *Bio/Technology* **4** (1986) 505 – 510.
34. Bio/Technology Conference: *Production Systems, Alternatives for Commercial Production of Recombinant Proteins*, New York, May 6, 1987.
35. M. D. Summers, G. E. Smith: *A Manual of Methods for Baculovirus Vectors and Insect Cell Culture Procedures*, Texas Agricultural Experiment Station Bulletin No. 1555, 1984.
36. A. Klausner, *Bio/Technology* **3** (1985) 193 – 194.
37. P. C. Brown: “Considerations in the Use of Mammalian Cells for the Expressions of Heterologous Proteins,” Bio/Technology Conference, *Production Systems, Alternative for Commercial Production of Recombinant Proteins,* New York, May 6, 1987.
38. T. Cartwright, *TIBTECH* **5** (1987) 25 – 30.
39. A. Pjavirani *et al.*, *Bio/Technology* **5** (1987) 389 –392.
40. J. van Brunt, *Bio/Technology* **4** (1986) 1057 – 1062.
41. H. Bialy, *Bio/Technology* **5** (1987) 883 – 890.
42. M.-R. Kula, H. Schütte, *Biotechnol. Prog.* **3** (1987) 31 – 42.
43. G. D. Ball in R. E. Spier, J. B. Friffiths (eds.): *Animal Cell Biotechnology*, Academic Press, London 1985, pp. 87 – 125.
44. C. G. Rosén, R. Datar: “Primary Separation Steps in Fermentation Processes,” *Proceedings of Biotech 83*, Online Publishers, London 1983, pp. 201 – 224.
45. M. Cheryan: *Ultrafiltration Handbook*, Technomic Publishing Co., Lancaster, Pennsylvania, 1986.
46. Ch. Östlund, P. Borwell, B. Malm, *Dev. Biol. Stand.* **66** (1987) 367 – 375.
47. P. V. Byrne, *J. Chromatogr.* **283** (1984) 459 – 461.
48. K. A. Kroner, M.-R. Kula, *Proc. Biochem.* **13** (1978) 7 – 9. K. H. Kroner, H. Schütte, W. Stack, M.-R. Kula, *J. Chem. Technol. Biotechnol.* **32** (1982) 130 – 137.
49. *Notes to Applicants for Marketing Authorization on the Production and Quality Control of Medicinal Products Derived by Recombinant DNA Technology*, Commission of the European Communities, Committee for Proprietary Medicinal Products, Ad Hoc Working Party on Biotechnology/Pharmacy, Brussels, Belgium 1987.
50. S. S. Block in J. F. Gardner, K. G. Grey (eds.): *Disinfection and Preservation*, Lea & Febiger, Philadelphia 1983, pp. 251 – 270.
51. R. D. Johnson, *Dev. Ind. Microbiol.* **27** (1987) 77 –83.
52. J.-C. Janson, P. Hedman, *Biotechnol. Prog.* **3** (1987) 9 – 13.
53. *Pharmacia Literature References*, published annually, Pharmacia LKB, Munzingerstr. 9, D-7800 Freiburg.
54. *Ion Exchange & Chromatofocussing, Principles & Methods*, FPLC, Pharmacia, Laboratory Separation Division, 75182 Uppsala, Sweden 1985.
55. H. A. Chase, *J. Chromatogr.* **297** (1984) 179 – 202.
56. F. H. Arnold, J. J. Chalmers, M. S. Saunders, M. S. Croughan *et al.*, *ASC Symp. Ser.* **271**, (1985) 113.
57. R. W. Scott, S. A. Duffy, B. J. Moellering, C. Prior: “Purification of MAb from Large Scale Mammalian Cell Culture Perfusion Systems,” *Biotechnol. Progr.* **3** (1987) no. 1, 49 – 56.
58. A. I. Daniels, N. T. Pettersson, J. H. Berglöf, H. Scandella: *Purification of Recombinant Human Superoxide Dismutase Produced in Yeast, a Process Development Study*, 3rd Preparative Scale Liquid Chromatography Symposium, Washington D.C., May 4 – 5, 1987.
59. J. Porath, B. Olin, *Biochemistry* **22** (1983) 1621 – 1630.
60. I. Dodd, S. Jalalpour, W. Sowthwick, P. Newsome, M. J. Browne, J. H. Robinson, *FEBS Lett.* **209** (1986) 13 – 16.
61. G. Sofer, L. E. Nyström: *Process Chromatography, A Practical Guide*, Academic Press, London 1988.
62. H. D. J. Johansson, N. T. Pettersson, J. H. Berglöf: *Development of a Process for Large Scale Purification of S. aureus Cell Culture Supernatants*, 3rd Preparative Scale Liquid Chromatography Symposium, Washington D.C., May 4 – 5, 1987.
63. T. Kristiansen in O. Hoffman – Ostenhof (ed.): *Affinity Chromatography,* Pergamon Press, Oxford 1978, pp. 191 – 206.
64. R. Neumair, H. R. Maurer, *Hoppe Seyler's Z. Physiol. Chem.* **363** (1982) 1493 – 1500.
65. P. Hedman, J.-G. Gustaffson: *Verminderung der Proteaseaktivität in Zellhomogenaten aus Mikroorganismen*, Pharmacia, Uppsala, Sweden 1983 (Abstract for JABS/WHO-Meeting, Standardization and Control of Biologicals Produced by Recombinant DNA Technology, Geneva 1983).
66. S. R. Kronheim, C. J. March, S. K. Erb, P. J. Conlon, D. Y. Mochizuki, T. P. Hopp, *J. Exp. Med.* **161** (1985) 490 – 502.

67. G. Sofer, Ch. Mason, *Bio/Technology* **5** (1987), 239 – 244.
68. Gr. Guichon, H. Colin: *Theoretical Concepts and Optimization in Preparative Scale Liquid Chromatography*, Chromatography Forum, Sept. – Oct. 1986, pp. 21 – 28.
69. H. A. Chase in M. S. Verall (ed.): *Discovery and Isolation of Microbial Products*, Ellis Horwood, Chichester, England 1985, p. 129.
70. *Ion Exchange Chromatography, Principles & Methods*, Pharmacia, Uppsala, Sweden 1983.
71. "Scale Up Factors for Ion Exchange," *Separation News 13/2*, March 1986, Pharmacia, Uppsala, Sweden.
72. H. A. Chase, *Chem. Eng. Sci*, **39** (1984) 1099 – 1125.
73. H. Johansson, L. O. Lindquist, D. Low: *Design Criteria for the Purification of Biopolymers for the Pharmaceutical Industry*, Pharmacia, Uppsala, Sweden 1985.
74. J.-C. Janson, P. Hedman in A. Fiechter (ed.): *Advances in Biochemical Engineering*, vol. **25**, Springer Verlag, New York 1982, pp. 87 – 89.
75. J.-Ch. Janson, *Process Scale Purification of Biological Macromolecules*, 4th European Conference of Biotechnology, Amsterdam, June 14 – 19, 1987.
76. Ch. Östlund, *TIBTECH* **4** (1986) 288 – 293.
77. L. O. Lindquist: *Freiburg Lectures*, Pharmacia LKB, Freiburg, Federal Republic of Germany 1984.
78. D. Behrens, *Swiss Biotech* **3** (1985) 13 – 24.
79. R. Bilder, *Int. Labmate* **8** (1983) 5.
80. C. Blanche, A. Johannsson, P. Collins, *International Biotechnology Laboratory*, February 1988, pp. 8 – 15.
81. H. Johannson, G. Jagschies, *GIT Fachz. Lab.* **30** (1986) 854 – 858.
82. Code of Federal Regulations, Title 21, *Current Good Manufacturing Practice and Finished Pharmaceuticals*, Office of the Federal Register, General Services Admin., Washington DC, Part 211.
83. *Guideline on Sterile Drug Products Produced by Aseptic Processing*, Div. of Drug Quality Compliance, Center for Drugs and Biologics and Office of Regulatory Affairs, FDA, Rockville, MD, USA 1985.
84. *Points to Consider in the Manufacture and Testing of Monoclonal Antibody Products for Human Use*, Office of Biologics Research and Review, Center for Drugs and Biologics, Food and Drug Administration, USA 1987.
85. E. M. Fry: "FDA Update on Aseptic Processing Guidelines," *J. Parenter. Sci. Technol.* **41** (1987) 56 – 60.
86. *Guideline on General Principles of Process Validation*, **3/1983 (Draft 1), 3/1984 (Draft 2), 9/1985 (Draft 3), 3/1986 (Draft 4)**, National Center for Devices and Radiological Health, Rockville, MD, USA.
87. R. F. Tetzlaff: "Regulatory Aspects of Aseptic Processing," *Pharm. Technol.* **19** (1984) 38 – 44.
88. E. M. Fry: "Process Validation: The FDA's Viewpoint," *D & CI* (1985) 46 – 51.
89. *Richtlinien zum Schutz vor Gefahren durch in-vitro neukombinierte Nukleinsäuren*, 5th revised version, Bundesminister für Forschung und Technologie, Bundesanzeiger Verlagsges. mbH, Köln 1986.
90. *Considerations for the Standardization and Control of the New Generation of Biological Products* (1985), The National Institute for Biological Standards and Control, Holly Hill, Hampstead, London NW3 6RB, England.
91. Kommittee für Laboratorien und offizielle Medikamentkontrolldienste und der Sektion der Industrieapotheker der FIP: *Richtlinien für die gute Validierungs-Praxis*, FIP Kongress, Madrid 1980.
92. W. van der Loo, R. Hamfers in H. V. Peeters (ed.): *Protides Biol. Fluids*, **23** (1976) 603.
93. J. M. Curling, J. M. Cooney, *J. Parenter. Sci. Technol.* **36** (1982) pp. 59 – 63.
94. Downstream, News and Views for Process Biotechnologists 3, Pharmacia, Uppsala, Sweden 1987.
95. R. W. Grabner, US 3 897 309, 1975.
96. T. Shibatani, T. Kakimoto, I. Chibata, *Thromb. Haemostasis* **49** (1983) 91.
97. Downstream, News and Views for Process Biotechnologists 2 (1986), *Process Hygiene in Industrial Chromatographic Processes,* Pharmacia, Uppsala, Sweden.
98. L. S. Martin *et al.*, *J. Infect. Dis.* **152** (1985) 400 – 403.
99. R. L. Whitehouse, L. F. L. Clegg, *J. Dairy Res.* **30** (1963) 315 – 322.
100. M. Uhlèn *et al.*, *J. Biol. Chem.* **259** (1984) 1695.
101. B. Nilsson *et al.*, *Nucleic Acids Res.* **13** (1985) 1151.

[Top of Page]

Abbreviations

A: affinity chromatography

AIDS: acquired immune deficiency syndrome

CIP: cleaning in place

CM: carboxymethyl

CSF: colony stimulating factor

DEAE: diethylaminoethyl

DNA: deoxyribonucleic acid

EGF: epidermal growth factor

GMP: good manufacturing practice

HETP: height equivalent to a theoretical plate

HGH: human growth hormone

HIC: hydrophobic interaction chromatography

HTLV: human T-cell leukemia virus

IE: ion-exchange chromatography

IgG: immunoglobulin G

IL-1: interleukin 1

MAb: monoclonal antibody

PBS: phosphate-buffered saline

RPC: reversed phase chromatography

SIP: sterilization in place

SOD: superoxide dismutase

TNF: tumor necrosis factor

tPA: tissue plasminogen activator

Tris: tris(hydroxymethyl)aminomethane

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Biochemical Separations

Joseph A. Shaeiwitz, West Virginia University, Morgantown, West Virginia, 26506-6101, United States

Joseph D. Henry, Jr., West Virginia University, Morgantown, West Virginia, 26506-6101, United States

1. Introduction 1
2. Protein Recovery Considerations . . 2
3. Cell Disruption 4
4. Solids Removal 4
4.1. Filtration and Ultrafiltration 4
4.1.1. Traditional Filter Operations 5
4.1.2. Theory of Ultrafiltration and Cross-Flow Filtration 7
4.1.3. Biological Applications of Ultrafiltration and Cross-Flow Filtration 7
4.1.4. Additional Methods for Reducing Fouling and Cake Formation 8
4.2. Centrifugation 9
4.3. Coagulation and Flocculation 10
5. Product Isolation 10
5.1. Evaporation 10
5.2. Extraction 10
5.2.1. Organic Solvents 10
5.2.2. Aqueous Two-Phase Systems 12
5.3. Adsorption Including Ion Exchange 13
5.3.1. Sorption Processes 13
5.3.2. Affinity Adsorption 13
5.3.3. Ion-Exchange Adsorption 14
5.3.4. Elution 15
5.3.5. Quantitative Aspects 15
5.4. Precipitation 16
5.5. Solubilization 17
6. Purification 17
6.1. Chromatography 17
6.1.1. Gel Filtration Chromatography 18
6.1.2. Affinity Chromatography 18
6.1.3. Quantitative Aspects 18
6.2. Electrical Separations 20
6.2.1. Electrophoresis 20
6.2.2. Isoelectric Focusing and Isotachophoresis 21
6.2.3. Electrodialysis 21
6.2.4. Dielectrophoresis 22
6.3. Magnetic Separations 22
7. Polishing 22
7.1. Crystallization 22
7.2. Drying 25
8. Bioseparation Strategies 26
9. References 26

1. Introduction

The separation and purification of fermentation products can be the major economic factor in commercializing a biochemical process. Although few rigorously documented studies exist, numerous unpublished reports suggest that separation can account for 75 % of the total cost of the process. This is particularly true when the product is a protein, enzyme, or antibody.

The major reason for the high cost of purifying fermentation products is the nature of both the processes and products involved. Many products, especially proteins, are fragile molecules. Products required for clinical purposes must be ultrapure. The products of fermentation processes are formed in a very dilute aqueous solution, which contains solids (cells or cell fragments) that must be removed. The need for solids removal, product concentration, and subsequent purification yields bioseparation schemes consisting of numerous processing steps.

In general, a bioseparation sequence is composed of four main classes of separation steps [1], [8]:

1) Removal of insolubles (i.e., cells and cell fragments)
2) Product isolation (product is concentrated; process volume is decreased)
3) Product purification (a specific, high-purity product is obtained)
4) Polishing (the product is made acceptable for the consumer)

Table 1 shows quantitative product information for these four general steps for antibiotics.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_11

Table 1. Typical processing profile for an antibiotic ∗

Step	Product concn., g/L	Quality, %
Harvest broth	0.1 – 5	0.1 – 1.0
Filtration	0.1 – 5	0.1 – 2.0
Primary isolation	5 – 10	1 – 10
Purification	50 – 200	50 – 80
Crystallization		90 – 100

∗ Reproduced from [1] with permission of Academic Press.

Fermentation products include low molecular mass organic molecules (e.g., acetone, acetic acid, and ethanol), amino acids, antibiotics, proteins, antibodies, enzymes, lipids, hormones, polysaccharides, and vitamins and are summarized in → Biotechnology, Chap. 3.6. Many of these substances are described in detail under separate keywords. Proteins, enzymes, antibodies, and some hormones are fragile molecules; most of the other products are robust. Enzymes and antibodies require a specific three-dimensional structure for their biological activity. This structure can be irreversibly altered (denaturation) by extremes in pH, temperature, and shear or by exposure to air – water interfaces or nonpolar solvents. However, most of the other molecules listed above are not damaged by extreme conditions; thus their purification is somewhat simpler.

Categorizing the scale of bioprocesses is somewhat imprecise. A method has been described that is based on the type of product and the capacity of the process [2]. Large-scale processes are considered to involve products such as ethanol, organic acids, and polysaccharides. Process volumes can exceed $100\,m^3$ per batch and recovery strategies are well-established. Medium-scale processes involve production of robust products such as penicillin or streptomycin. Conventional separation processes may be used, and their cost depends heavily on the cost of the initial volume reduction steps. Small-scale processes involve production of specialty pharmaceuticals and enzymes, some of which may be fragile, thus increasing the cost of the separation. The separation steps involved are usually scaled-up laboratory techniques, even though the process volumes can be $1 - 100\,m^3$ per batch. These small-scale biochemical processes benefit the most from development of new separation techniques or strategies.

The emphasis of this article is placed on separation techniques used in the purification of proteins, including enzymes and antibodies (see also → Enzymes, Chap. 3.2.). These are the most difficult and most costly biochemical separations; many of their details are quite novel. The operations used to purify low molecular mass products, including antibiotics, tend to be conventional separation processes (e.g., distillation and extraction), which are discussed elsewhere. The steps used for separating antibiotics have been developed by the industries involved and are reviewed in [2] (see also → Antibiotics, Chap. 6.).

Table 2 lists separation operations used to recover biomolecules. General reviews of bioseparation operations are also available [3–18].

2. Protein Recovery Considerations

Proteins are composed of sequences of ca. 20 different amino acids, and their molecular masses are typically $10^4 - 5\times10^5$. They have four levels of structure. The *primary structure* refers to the amino acid sequence. The *secondary structure* refers to the short-range conformation of the molecules (e.g., α-helices or β-sheets). The *tertiary structure* refers to the complex three-dimensional structure created by the folding of the long chains of amino acids. The tertiary structure of enzymes and antibodies is responsible for their biological activity. This folding often isolates hydrophobic portions of the protein from the aqueous environment, leaving the charged amino and carboxylic acid groups on the surface. Disruption of this tertiary structure often appears irreversible, resulting in a permanent loss of biological activity and a worthless product. Renaturation is possible in principle, but is not yet fully understood. The *quaternary structure* refers to the manner in which multiple polypeptide chains (subunits) with tertiary structure interact; the best known example is hemoglobin, which is composed of four subunits, each of which is bound to an iron atom.

The tertiary structure of a protein can be disrupted by any change in the environment that alters the interaction between the hydrophilic portions of the protein and the solvent (water). This change may be reversible, as in the case of

Table 2. Summary of biochemical separation operations

Operation	Application	Comments	References
Solids removal			
Filtration	removal of cells or cell debris	rotary vacuum filter most common, often with precoat	[1], [14]
	dewatering of cells prior to lysis	cross-flow configuration reduces cake formation	[19], [20]
Centrifugation	removal of cells or cell debris	size of centrifuge is limited, rpm limited if centrifuge is not continuous flow	[14]
Coagulation/flocculation	removal of cells or cell debris	polyelectrolytes used as flocculants	[14]
Product isolation (volume reduction)			
Evaporation	volume reduction, product must be heat-stable	vacuum drying used to keep temperature low	[12], [21], [22]
Extraction	removal of cells or cell debris (whole-broth processing)	aqueous two-phase polymer systems	[23–26]
	product concentration (proteins, antibiotics, etc.)	aqueous two-phase for proteins, mild organic solvents sometimes containing complexing agents for antibiotics	[18], [23–26]
	affinity separation (proteins)	polymer-bound ligands	[23–26]
Adsorption including ion exchange	whole-broth processing (antibiotics)	adsorbent contacted with fermentation	[1], [16], [17],
		broth without solids removal	[27], [28]
	product concentration (proteins, antibiotics, etc.)	nonspecific adsorption, elution required, gradient elution is popular	[1], [3]
	affinity adsorption (proteins)	specific adsorption, elution required	[1], [3]
Precipitation	volume reduction with partial purification (proteins)	salt, poly(ethylene glycol) added, additive not coprecipitated	[29]
Solubilization	isolation of membrane-bound proteins, which are usually insoluble in water	an empirical laboratory technique, nonionic surfactants usually used	[2], [30–34]
	solubilization of inclusion bodies		[35–37]
Purification (fractionation)			
Chromatography	gel permeation, separation according to molecular mass (proteins)	scaleup problems	[8]
		less specific separation	[3], [16]
	affinity purification (proteins)	highly specific purification	[3], [8]
	other chromatographic methods	any method resulting in differential migration through column	[38]
Electrical separations (proteins)	electrophoresis	small-scale method, thermal convection problems	[3], [39], [40]
	isoelectric focusing	small-scale method, gel process	[3], [41]
	electrodialysis	advantageous for salt removal	[42]
	dielectrophoresis	cell separations	[43]
Ultrafiltration (proteins)	protein concentration, salt removal	cake formation (fouling) is troublesome, cross-flow configuration in hollow fiber devices is popular, diafiltration also possible	[44]
Magnetic separations	separation of proteins adsorbed to magnetic support	magnetic affinity beads used, really an adsorption technique	[45], [46]
	cell separations	red blood cells are paramagnetic	
Polishing			
Crystallization	product obtained as pure solid	from solution, supersaturation involved	[1]
Drying	product obtained as pure solid	solvent must be recovered, protein activity is preserved, temperature must be kept low if proteins involved, freeze-drying (lyophilization) often used	[1], [47], [48]

precipitation with mild organic solvents, ammonium sulfate, or poly(ethylene glycol). Extremes in pH, temperature, or shear, as well as exposure to an air – water interface or a nonpolar solvent, can cause complete unfolding of the chain that usually appears to be irreversible. Therefore, many traditional separation processes, such as extraction into an organic solvent, distillation, and evaporation, often cannot be applied to protein recovery processes.

Although several methods are available for removing solids and concentrating protein solutions, relatively few methods exist for separating proteins from each other. Available methods exploit differences in solubility, charge, and molecular mass. However, the most specific property of a protein used in purification is its tertiary structure and, thus, its biological affinity for certain molecules.

Enzymes, for example, have affinity for the substrates and cofactors in the reactions they catalyze. Since nucleotide cofactors (e.g., nicotinamide – adenine dinucleotide, NAD) are involved in many enzymatic reactions, use of nucleotides in affinity purification is not highly specific for a particular protein. However, a monoclonal antibody raised against a specific enzyme is extremely specific and has affinity for only that enzyme (→ Monoclonal Antibodies).

3. Cell Disruption

Small bioproducts are usually secreted from the cell, making the first separation step (i.e., solids removal) reasonably straightforward. Larger molecules, such as proteins, are not always secreted from the cell. If the product is intracellular, it generally must be released by rupturing (lysing) the cell [12]. The solids (cells) may be concentrated prior to lysis, but another step to remove solids must follow lysis.

Lysis involves the rupture of either a cell wall or cell membrane, or both. Animal cells are surrounded by a lipid – lipoprotein cell membrane and are reasonably easily ruptured. Bacteria and plant cells (including *Escherichia coli* and yeasts, which are commonly used in genetic engineering as hosts for the production of foreign products) have an additional cell wall that is usually composed of polysaccharide.

The cell wall is very difficult to rupture, especially in the case of the yeast *Saccharomyces cerevisiae.* In some cases, rupture poses problems because the desired proteins are located in the periplasmic space between the cell membrane and the cell wall.

Techniques used to lyse microorganisms are summarized in Table 3; see also → Enzymes, Chap. 3.2. Some lysis methods involve mechanical breakage resulting from shear or friction with solid objects. Friction may produce heat, which can denature the protein. Other techniques use a detergent, an enzyme, or a strong base to dissolve the cell wall. However, these treatments may denature certain proteins and also liberate intracellular proteins, including proteases. Proteases attack and inactivate other proteins, causing product loss. Therefore, the lysis technique must be chosen carefully, and initial product isolation should proceed rapidly.

4. Solids Removal

The three main methods of solids removal are filtration, centrifugation, and coagulation/flocculation. Some cells can be separated by magnetic methods (Section 6.3). Solids removal by aqueous two-phase extraction and whole-broth processing are discussed in Sections 5.2.2 and 5.3.1, respectively.

Ultrafiltration is used primarily for the purification and concentration of macromolecules (Chap. 6) but is discussed in this chapter together with filtration because there is extensive overlap between the two techniques (e.g., theory, cake formation, and membrane fouling).

4.1. Filtration and Ultrafiltration

Filtration is a solid – liquid separation in which a pressure difference is applied to force liquid through a filter medium (membrane) that is impermeable to the solid. The product may be a concentrated solid or a clarified liquid. In *ultrafiltration,* a higher pressure is required to separate dissolved macromolecules by forcing the solvent and small solutes through a membrane that is impermeable to the macromolecules. An even higher pressure is required for *reverse osmosis* (hyperfiltration) in order to

Table 3. Lysis of microorganisms ∗

Method	Description	Advantages	Disadvantages
Physical Methods			
Agitation with abrasives (glass beads)	rapid mixing with glass beads used for Brewers yeast, *Saccharomyces cerevisiae*	effective against tough cell walls enzyme activity compares favorably with that obtained by other methods suitable for large-scale enzyme production	removal of glass beads from the disrupted cells
Solid shear (freeze pressing)	pressure extrusion of frozen material through a narrow hole disruption is probably caused by deformation of cells embedded in the ice	particularly advantageous for heat-sensitive enzymes high enzyme yields are obtained high percentage of the cells are lysed	large-scale equipment not commercially available freezing of cells prior to disintegration
Liquid shear	passage of a cell suspension (10 – 17 wt/vol%) through a restricted orifice at pressures of up to 200 MPa	effective for bacterial cells and to a lesser extent fungal cells very popular suitable for large-scale enzyme production	heat generation single-pass operations usually insufficient irregular particle size (filament or pellets) reduces the efficiency of cell breakage
Freeze–thawing	freeze-thawing the suspension at − 20 to − 35 °C ice crystals break cell membrane		most intracellular enzymes defy release most cells are resistant to disruption slow and inefficient
Chemical Methods			
Enzyme treatment	lysis with egg-white lysozyme large-scale lysis with *Micrococcus lysodeikticus* without a strong cell wall, the resulting protoplast or spheroplast lyses because of the high osmotic pressure of the cell contents (5 – 25 kg/cm^2)	gentle and selective	possible destruction of enzyme by cellular proteases expensive enzymes are mixed with the whole cell lysate, which adversely affects subsequent purification steps
Detergent treatment	ionic detergents: sodium dodecyl sulfate and cetyl trimethyl ammonium bromide nonionic detergents: Tween or Triton X-100	effective for the commercial production of cholesterol oxidase	may inactivate enzymes presence of detergents affects subsequent purification
Alkali treatment	pH: (11.5 – 12.5) for 20 – 30 min	inexpensive used to extract the therapeutic enzyme L-asparaginase from *Erwinia chrysanthemi*	may inactivate enzymes reduces pyrogen content
Osmotic shock	rapid change in salt concentration	only low levels of soluble proteins are released, facilitating subsequent purification	tough cell walls are less susceptible to this method cells from solid – liquid recovery steps often contain much occluded salt

∗ Reproduced from [12] with permission of CRC Press.

overcome the osmotic pressure that is created across the membrane due to the impermeability of the membrane to the solute. The applied pressure forces the solvent through the membrane and reverses the direction of osmotic solvent transfer (→ Membranes and Membrane Separations).

Only filtration and ultrafiltration are applicable to biological processes. Filtration is discussed in detail elsewhere (→ Filtration), as is ultrafiltration (→ Membranes and Membrane Separations).

4.1.1. Traditional Filter Operations

Filtration is sometimes used to concentrate whole cells prior to lysis, but most often to re-

move cells from an extracellular product. Two types of filtration are employed in the pharmaceutical industry [1], [14]. The plate and frame arrangement, a batch operation, requires periodic removal of the filter cake (impermeable solids); it is applicable only to small-scale processes. The rotary vacuum (rotary drum) filter, the workhorse of the industry, comprises a filter cloth on a rotating drum so that the cake may be removed continuously and the membrane reused.

Filter cakes tend to be compressible, causing the filtrate flow to decrease with increasing pressure. Even though filtrate flow may drop due to cake formation, increasing the pressure drop is counterproductive for a compressible filter cake. In a rotary vacuum filter, accumulated cake is removed in the latter part of the drum cycle, most commonly by the string-discharge or knife-edge method. The string-discharge method is applicable only to very porous, usually incompressible filter cakes. The knife-edge method is more suitable for denser, compressible cakes [14].

When cake formation is particularly difficult, as in almost all biological applications, *filter aids* can be employed to facilitate cake removal but they also perform several other functions. They help to retain finer particles than the membrane would otherwise allow. Their most common function is to offset the clogging properties of the cake. Cells are deposited on the cake, but reasonable filtrate flow is maintained due to the porosity of the filter aid. A thin layer of the filter aid is removed during cake removal; removal is thus more efficient than scraping the cells from the membrane. The most common filter aids for biological applications are diatomaceous earth, which consists of the skeletal remains of tiny aquatic plants, and perlite, which is a volcanic rock processed into an expanded form. Coagulants and flocculants may also be used to create larger particle aggregates that do not clog the pores of the membrane to the same extent as single particles.

Filter aids must be used with caution, however [14]. Waste cells that are contaminated with filter aids are no longer suitable for animal feeds. If the product is an extracellular enzyme, it may adsorb onto the filter aid. Furthermore, filter aids are costly. Thus, the advantage of a filter aid over decreased filtrate flow should be carefully considered.

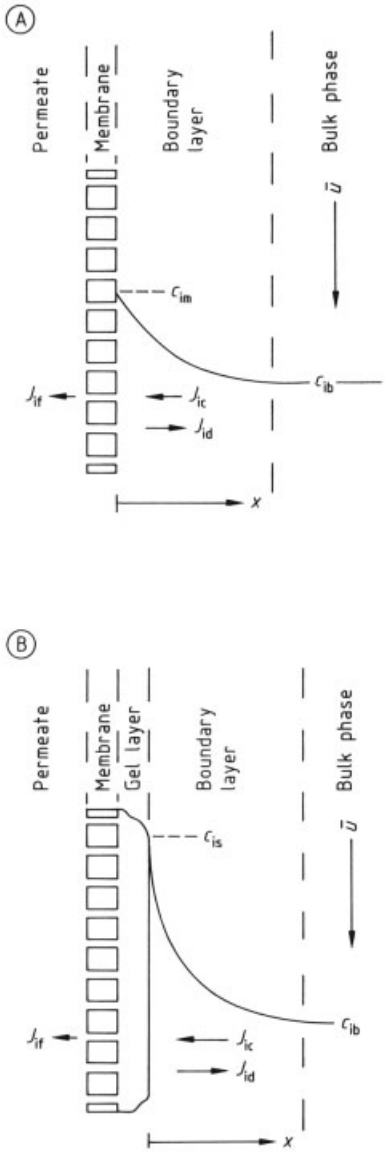

Figure 1. Schematic representation of concentration polarization without gel formation (A) and with gel formation (B) ∗
c_{ib} = bulk feed concentration of macromolecule i
c_{im} = concentration of macromolecule at the membrane
c_{is} = concentration of macromolecule at the edge of the gel
J_{ic} = molar flux of macromolecules toward the membrane due to flow
J_{id} = molar diffusion flux of macromolecules from the membrane
J_{if} = molar flux of macromolecules in the permeate
$\bar{u}$ = velocity of bulk flow x = distance from membrane (A) or gel (B)
∗ Reproduced from [44], with permission of Springer Verlag.

4.1.2. Theory of Ultrafiltration and Cross-Flow Filtration [19], [20], [44]

In ultrafiltration, a pressure difference across a membrane that is impermeable to dissolved macromolecules causes solvent and small solutes to pass through the membrane while the macromolecule is retained. Macromolecular adsorption at the membrane surface fouls the membrane and decreases the permeate flux. This phenomenon is known as *concentration polarization* because diffusion of the macromolecule back into the solution creates a concentration gradient. The macromolecule concentration at the membrane surface often exceeds its solubility limit, and an insoluble gel layer forms. Proteins often exhibit this behavior, which is illustrated in Figure 1. The flow of feed is tangential to the membrane filter (i.e., cross-flow), which minimizes fouling.

Conventional, "dead end" filtration of cells, i.e., filtration involving bulk flow toward the membrane, is associated with the problem of cake formation. Tangential flow (cross-flow) filtration can be employed to reduce cake formation [19], [20]. In principle, this is similar to the ultrafiltration illustrated in Figure 1B, the gel layer representing the filter cake. However, in cross-flow filtration the back-diffusion phenomenon is more correctly termed *particle polarization*.

In both ultrafiltration and cross-flow filtration, the permeate flux J (the volumetric flux of filtrate, $m^3\,m^{-2}\,s^{-1}$), is given by

$$J = \Delta P / R_T$$

where ΔP is the pressure drop and R_T is the total resistance to permeate flow. The parameter R_T is the sum of three effects caused by the membrane, the gel (or cake) layer, and the diffusion layer. In ultrafiltration, the numerator is really $\Delta P - \Delta \Pi$, where $\Delta \Pi$ is the osmotic pressure difference; however, for dilute macromolecular systems, $\Delta \Pi$ is negligible. If quantitative retention of the macromolecule (or particle) and gel (or cake) formation are assumed, then

$$J_{if} = 0;\ \text{i.e.},\ J_{ic} + J_{id} = 0$$

where

J_{if}= molar flux of macromolecules (or particles) in the permeate (given by Fick's law)
J_{ic}= molar flux of macromolecules (or particles) toward the membrane due to flow
J_{id}= molar diffusion flux of macromolecules (or particles) from the membrane

With $J_{ic} = Jc_i$, $J_{id} = -D_i \mathrm{d}c_i/\mathrm{d}x$, and $k_d = D_i/\delta$, then

$$J = k_d \ln (c_{is}/c_{ib})$$

where

c_i= concentration of macromolecules (or particles), g/cm^3
D_i= macromolecule (or particle) diffusivity, cm^2/s
x= length coordinate from membrane or gel surface, cm
k_d= mass transfer coefficient, cm/s
δ= boundary layer thickness, cm
c_{is}= concentration of macromolecules (or particles) at the edge of the gel (or cake) layer, g/cm^3
c_{ib}= bulk feed concentration of macromolecules (or particles), g/cm^3

The parameter k_d is easily predicted by the Graetz solutions for laminar flow or by the Dittus – Boelter correlation for turbulent flow [44].

The above result implies that the gel (or cake) properties, not the characteristics of the membrane, control the resistance to permeate flow. Therefore, the permeate flow can be increased only by reducing the cake (or gel) resistance; this is accomplished by using high shear rates that are produced by operating at high velocities in narrow channels, as predicted by the theoretical dependence of k_d [44]. Furthermore, the permeate flux is independent of pressure. Increasing the pressure results in a brief, transient increase in permeate flux, but the gel (or cake) thickness increases concomitantly, yielding an identical steady-state flux. Finally, the value of k_d must be considered in the decision of whether to use cross-flow filtration or ultrafiltration in a batch or continuous mode [19], [20].

4.1.3. Biological Applications of Ultrafiltration and Cross-Flow Filtration

Cross-flow filtration of whole cells provides high filtration rates by minimizing cake formation

without the need for filter aids or flocculants [19]. The cost of increased filtrate flow is determined by the energy necessary to maintain the cross-flow in excess of that necessary for the transmembrane pressure drop. Cross-flow filtration has been used to separate blood plasma from red blood cells [4]. It should also be suitable for processing cell lysates such that cell debris is retained, but intracellular protein passes into the filtrate [49]. However, cake resistance then appears to be more of a problem because protein is adsorbed onto the cell debris in the cake. Even with large volumes of wash water, protein retention is still a problem. For this process to become economically viable, cake formation must be further minimized. Cross-flow processes are generally carried out in rectangular channels. Attention must be paid to possible shear denaturation of proteins. The design and choice of cross-flow filtration equipment are discussed in [19], [20].

Ultrafiltration is an established procedure for concentrating biological macromolecules. These applications are reviewed elsewhere [44] and include the fractionation of blood and plasma [4] (see also → Membranes and Membrane Separations). Hollow fiber modules are the most popular configurations for ultrafiltration, but other designs are also possible. Important considerations in designing and choosing equipment are described in [50].

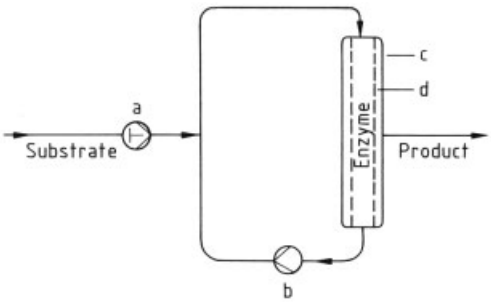

Figure 2. Schematic of a continuously operated ultrafiltration membrane reactor ∗
a) One-way valve; b) Pump; c) Ultrafiltration module; d) Ultrafiltration membrane
∗ Reproduced from [44], with permission of Springer Verlag.

A novel concept involving ultrafiltration reactors is illustrated in Figure 2 [44], where substrate is fed continuously into a closed loop containing an enzyme. The enzymatic reaction occurs and the product is removed through the ultrafiltration membrane; the enzyme is retained to react with more substrate. However, shear enzyme denaturation and nonquantitative enzyme retention have been reported in this system [44].

Another novel variation is affinity ultrafiltration, in which the membrane is permeable to free proteins, but impermeable to proteins that are bound to affinity ligands. This allows retention of specific proteins while others pass into the permeate [51].

Diafiltration combines aspects of dialysis and ultrafiltration. Dialysis is a rate process in which macromolecules enclosed by a membrane diffuse more slowly through the membrane than do solvent and small solute molecules. Membranes that are impermeable to macromolecules are available. The solution outside the membrane is exchanged several times to allow effective removal of small solute molecules. Dialysis is a useful laboratory technique. In diafiltration, the filtrate formed in ultrafiltration is continuously replaced with fresh solvent (water) to enhance small solute removal [44].

4.1.4. Additional Methods for Reducing Fouling and Cake Formation

Cross-flow filtration or ultrafiltration does not entirely eliminate cake formation or fouling. Other methods, such as the following, are available.

Promotion of Turbulence. When laminar flow occurs in an entrance region, e.g., in tangential flow processes, the mass-transfer coefficient for particle (or concentration) polarization is proportional to the cube root of the Reynolds number, i.e., $k_d \sim Re^{1/3}$. (The Reynolds number $Re = dV\varrho/\mu$, is a dimensionless velocity, where d is the flow channel diameter or thickness, V is the average or bulk velocity, ϱ is the fluid density, and μ is the fluid viscosity.) In turbulent flow, $k_d \sim Re^{0.8}$. Therefore, caking and fouling in cross-flow filtration and ultrafiltration equipment are often avoided by promoting turbulence, e.g., with static mixer wires placed near the membrane or by fluidization of glass particles [52], [53]. However, any method that induces turbulence also increases the shear rate and can irreversibly denature labile proteins.

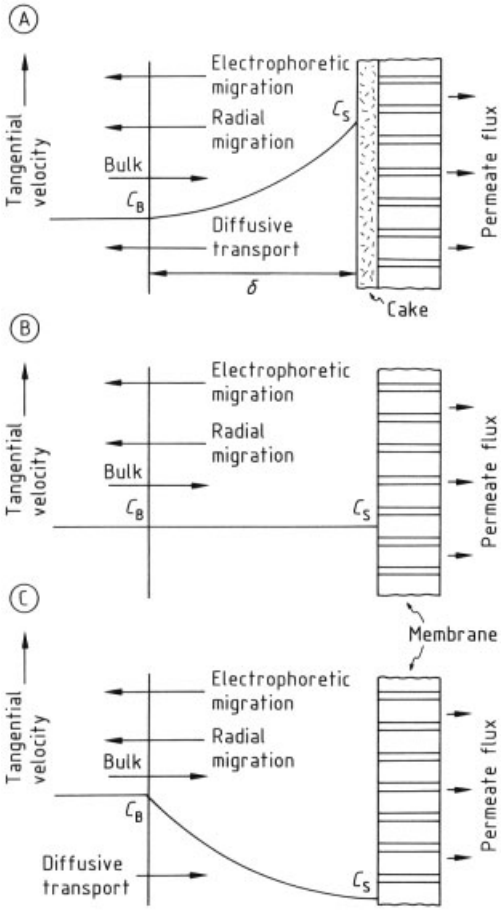

Figure 3. Regimes of operation of cross-flow electrofiltration ∗
A) Voltage < critical voltage; B) Voltage = critical voltage; C) Voltage > critical voltage.
C_B = bulk feed concentration of macromolecule
C_S = concentration of macromolecule at the membrane
δ = thickness of region of particle polarization (boundary layer)
∗ Reproduced from [55], with permission of McGraw-Hill.

Electrophoresis. Since macromolecules and particles almost always bear a charge that is usually negative, an electric field of appropriate magnitude and polarity applied perpendicular to the membrane should reduce or remove the cake or fouling [54], [55]. This phenomenon is known as electro(ultra)filtration and is illustrated in Figure 3. A critical voltage exists below which electrophoresis is too slow to offset the net flow toward the membrane caused by other mechanisms (Fig. 3A). Above the critical voltage, macromolecules or particles concentrate away from the membrane, resulting in a diffusive flux in the direction of the membrane (Fig. 3C). In this case, increasing the flow decreases the permeate flux by inducing cake formation. At the critical voltage there is no concentration gradient and, hence, no diffusive transport (Fig. 3B). Cake formation and fouling do not occur and increased flow increases permeate flux. Cross-flow electrofiltration has been demonstrated, but not in biological systems. Electro-ultrafiltration has been demonstrated in both nonbiological and biological systems [56].

Radial Migration. The dissolved macromolecules involved in ultrafiltration are too small to exhibit radial migration. However, particles migrate radially across streamlines due to inertia. Migration reduces cake formation and fouling and occurs from regions of high shear to regions of low shear. High-shear regions are found close to solid boundaries, whereas regions of lowest shear are always farthest from a boundary. Radial migration is reviewed in [57].

4.2. Centrifugation

Centrifugation is a well-developed unit operation and is discussed elsewhere (→ Centrifugation, Filtering, → Centrifugation, Sedimenting). Biochemical applications include cell separation from broths, removal of cell debris, separation of protein precipitates, adsorbent recovery, and, if ultracentrifugation is included, separation of dissolved macromolecules. Although standard centrifugation is a batch operation in the laboratory, it may also be a continuous, large-scale operation [14]. Continuous ultracentrifugation is also possible [58], [59].

Several different types of equipment are available for industrial centrifugation, including the tubular bowl, scroll and disk-nozzle arrangements. Basket centrifugation may be used to combine filtration and centrifugation in the same unit. The relative merits of these designs have been evaluated [14].

Centrifugation is limited to removal of particles with significantly higher densities than that of the solvent ($\Delta\varrho > 10^{-3}$ g/cm^3). Larger par-

ticles sediment more rapidly, requiring smaller, lower-speed centrifuges.

Although centrifugation is a reasonably large-scale operation, additional scaleup to handle very large throughputs is limited by the mechanical strength of the materials used to construct centrifuges. Thus, size and speed limitations on large-scale devices exist that do not apply to laboratory-scale devices [14].

4.3. Coagulation and Flocculation

(→ Flocculants)

Coagulation and flocculation are similar and their definitions vary. However, here coagulation is defined as the agglomeration of suspended particles due to the reduction of the repulsive forces caused by surface charge that keeps them separate and suspended in a liquid medium. The repulsive forces are reduced either by addition of inorganic electrolytes, which shield the repulsive surface charges, or by addition of polyelectrolytes that bind to and neutralize the surface charge.

Flocculation is the agglomeration of particles due to the bridging effect exerted by polymers that are adsorbed to more than one particle. Often, both mechanisms occur simultaneously when polymeric polyelectrolytes are involved. Flocculated aggregates tend to be more porous than coagulated aggregates.

Coagulation and flocculation have been applied mainly to removal of whole cells, although, in principle, they can be applied to cell debris and proteins. Particles are often coagulated or flocculated prior to filtration to reduce passage of small noncoagulated particles through the filter cloth (breakthrough) and to produce a more porous cake. For important operating parameters for these largely empirical operations, see [1], [14].

5. Product Isolation

5.1. Evaporation

Evaporation may be used as a volume reduction step in a bioprocess. The principles and equipment that are involved are discussed elsewhere in this volume (→ Evaporation). To minimize protein denaturation, vacuum evaporation may be used. Operating temperatures $< 40\,^{\circ}\mathrm{C}$ are desirable [12]. Foaming often causes equipment fouling and protein denaturation at the air–water interface. Ultrafiltration (Section 4.1.) is a much gentler method of volume reduction for proteins, and has to some extent replaced evaporation.

The most common examples of evaporation operations in bioprocesses are fruit juice concentration and essence recovery processes [60]. The aim of fruit juice concentration is to remove only the water, while retaining the flavor and nutritional components intact. Subsequent addition of water then yields the desired product. Since essential oils are usually more volatile than water, they are removed first and remixed with the product after evaporation of water [21], [22].

5.2. Extraction

Liquid – liquid extraction is a traditional separation operation that can be applied to purification of bioproducts; both organic and aqueous extraction media are used. Antibiotics and other small biomolecules (e.g., amino acids and vitamins) may be soluble in organic phases under certain conditions. Proteins, in general, either are insoluble in organic phases or irreversibly denature upon dissolution in an organic phase. However, extraction between aqueous, two-phase polymer systems provides a method of protein separation.

The equipment used in extraction of biomolecules does not differ substantially from extraction of other compounds and is described in detail elsewhere in this volume (→ Liquid – Liquid Extraction) [23]. However, in aqueous two-phase systems, the closeness in density of the two phases may cause some operating problems.

5.2.1. Organic Solvents

Small biomolecules may partition into organic solvents, such as straight-chain or branched aliphatic alcohols in the C_4–C_6 range and their acetate esters. Subsequent reextraction into an aqueous phase results in significant volume reduction. Many antibiotics contain carboxylic acid groups so that their solubilities change

significantly with pH [61]. Most steroids are more soluble in organic solvents than in aqueous media and, hence, are easily extracted into oil phases. Partition coefficients for these small biomolecules are not readily available however.

The solubility of a small biomolecule (e.g., a vitamin or antibiotic) can be modified by changing its counterion. In these *ion-pair extractions,* the counterion (e.g., a proton) is exchanged with a univalent cation (e.g., a quaternary ammonium ion, RNH_3^+) [18, Chap. 5]. If R is aliphatic, the solubility of the molecule in the organic phase increases as the length of R increases.

Employing *reverse micelles* may permit extraction of proteins into an organic phase, as shown in Figure 4 [62]. Micelles are aggregates of amphiphilic surfactant molecules. In aqueous solutions, the nonpolar tails of the surfactant molecules constitute the core of a usually spherical aggregate, whereas the polar heads face outward. In reverse micelles, the nonpolar tails orient outward in an organic solvent and the polar groups form the core. Proteins can be extracted with some degree of selectivity into the aqueous environment in the core of a reverse micelle, where they are stable. Reextraction into an aqueous phase follows, resulting in significant volume reduction.

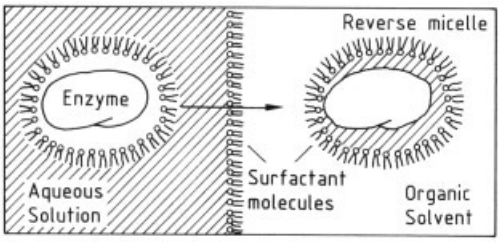

Figure 4. Protein solubilization in a reverse micelle ∗
The hydrophilic interior of the reverse micelle shields the enzyme from the organic solvent.
∗ Reproduced from [62], with permission of the American Institute of Chemical Engineers.

Liquid-emulsion membranes permit extraction into an organic phase and subsequent reextraction into an aqueous phase in a single step, as shown in Figure 5. The technology for these membranes is well-established in nonbiological systems and should be applicable to biological systems. Transport through the membrane occurs by means of diffusion and a carrier that is soluble only in the organic phase. Facilitated transport is also possible; cotransport and countertransport mechanisms can move solutes against their concentration gradients, providing a significant increase in concentration. This technology has been used successfully in the laboratory to extract phenylalanine, one of the components of Nutrasweet (aspartame [*2239-47-0*]) [63].

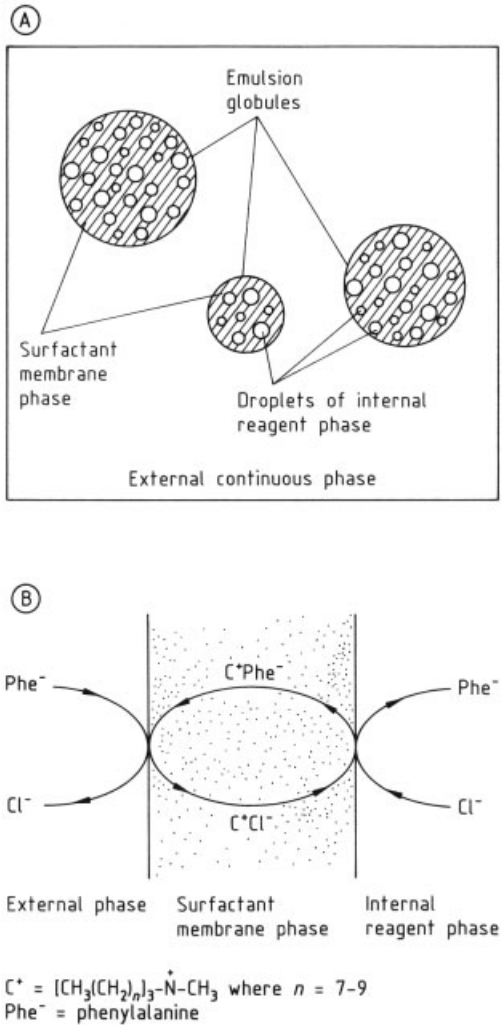

Figure 5. Use of liquid emulsion membranes for amino acid separation ∗
A) Composition of the phasesIn this water-in-oil-in-water emulsion, the oil phase acts as a membrane between the two phases.
B) Mechanism for the separation of phenylalanine
C^+ is a carrier that is retained within the membrane phase due to its solubility properties. Transport of Cl^- counter to Phe^- provides the driving force for Phe^- transport.
∗ Reproduced from [63], with permission of the American Chemical Society.

5.2.2. Aqueous Two-Phase Systems [24–26]

Certain water-soluble polymers are mutually incompatible, as are certain polymers and inorganic salts [24], [26]. If these substances are added to water, partitioning occurs to produce a multiphase mixture in which each phase contains a majority of one solute. As many as 18 phases have been documented [24], [26]. Of particular interest for protein separations are two-phase systems involving poly(ethylene glycol) (PEG) with either dextran or a salt. These systems partition into two aqueous phases (water content ≥ 90 %) such that each contains a majority of one component. Poly(ethylene glycol) is always found in the less dense phase. Proteins are soluble in the aqueous phases, and no significant denaturation occurs in the presence of the polymers. Proteins, proteins and cells, or cell debris partition differently between the two phases, providing a basis for separation [24–26].

Removal of Insolubles. In most cases, cells, cell debris, and cell organelles partition preferentially into the more dense phase, and the proteins remain in the less dense poly(ethylene glycol) phase. Therefore, extraction can also be used as a whole-broth processing step (see Section 5.3.1), which eliminates the need for a separate, costly step for removing solids. Ionic strength, polymer type, molecular mass, molecular mass distribution, and the characteristics of the cell or organelle surface are considered to be important for this type of extraction. Prediction of the partition coefficient is not yet possible; experimental measurements are required. Although not often cited, the polymer concentration is an important variable because the length of a tie line is directly related to the polymer concentration and determines the phase volume ratio, which ultimately affects purification. Table 4 presents some data illustrating the degree of partitioning and the yield of some enzymes obtained from cell homogenates.

Several two-phase systems may be suitable for a particular separation. Retention of insolubles at the interface has occasionally been observed; however, this occurs most frequently with whole cells, not cell debris.

Protein Partitioning. Two polymer phases [dextran and poly(ethylene glycol)] are usually used for protein separations rather than a polymer phase and a salt phase. Proteins partition preferentially into the dextran phase in most cases. However, increasing the salt concentration favors partitioning into the poly(ethylene glycol) phase. The exact magnitude of the partition coefficient depends on the type of polymer, its molecular mass, and the ionic strength. For example, increasing the dextran molecular mass favors partitioning into the poly(ethylene glycol) phase. Therefore, as for insolubles, the partition coefficients of proteins must be obtained empirically. Table 5 shows some measured partition coefficients. Partitioning into the poly(ethylene glycol) phase is favored due to the relatively high phosphate concentration.

Table 5. Enzyme partition coefficients in an aqueous two-phase system consisting of poly(ethylene glycol) 4000 (9 wt %), dextran T-500 (2 wt %), and potassium phosphate (0.3 mol/L, pH 7.5) [a]

Enzyme	Volume ratio [b]	K [c]
α-Amylase	8.16	3.64
Glucoamylase	8.60	1.22
Isoleucyl tRNA synthetase	5.86	2.17
α-Glucosidase	6.31	2.84
Glucose 6-phosphate dehydrogenase	8.20	0.62

[a] Adapted from [24], with permission of Academic Press.
[b] Equilibrium volume ratio of less dense to more dense phase.
[c] Enzyme concentration in less dense phase divided by concentration in more dense phase.

The expense of dextran has been a major hindrance to large-scale commercialization of aqueous two-phase systems for protein separations. However, use of cheaper crude dextran rather than purified dextran does not appreciably alter protein partition coefficients.

Since poly(ethylene glycol) is also a precipitating agent for proteins (Section 5.4), care must be taken not to exceed the concentration at which precipitation occurs. Use of poly(ethylene glycol) with a lower molecular mass not only increases the allowable concentration before precipitation occurs, but often increases the partition coefficient in favor of the glycol phase [23].

Affinity and Ion-Exchange Partitioning. Affinity partitioning and ion-exchange partitioning between two phases are alternative protein separation techniques. Both phases should be polymer phases because a salt phase may compromise the binding strength of the protein. Ion-exchange groups can be readily incorporated

Table 4. Partitioning of enzymes from cell homogenates [a]

Microorganism	Enzyme	Two-phase system [c]	K_{enzyme} [b]
Candida boidinii	catalase	PEG 4000–crude dextran	2.95
	formate dehydrogenase	PEG 4000–crude dextran	7.0
		PEG 1000–potassium phosphate	4.9
Sacharomyces cerevisiae	α-glucosidase	PEGH 4000–dextran T-500	2.5
	glucose 6-phosphate dehydrogenase	PEG 1000–potassium phosphate	4.1
Saccharomyces carlsbergensis	α-glucosidase	PEG 4000–dextran T-500	1.5
Escherichia coli	isoleucyl tRNA synthetase	PEG 6000–potassium phosphate	3.6
	phenylalanyl tRNA synthetase	PEG 6000–potassium phosphate	1.7
	fumarase	PEG 1550–potassium phosphate	3.2
	aspartase	PEG 1550–potassium phosphate	5.7

[a] Adapted from [23], with permission of Springer Verlag.
[b] K_{enzyme} is defined as the enzyme concentration in the less dense phase divided by the concentration in the more dense phase.
[c] PEG = poly(ethylene glycol)

into both poly(ethylene glycol) and dextran, but affinity ligands are usually added to the terminal hydroxyl groups of poly(ethylene glycol). Addition of these functionalities does not affect the polymer partitioning as long as the concentration of the derivatized polymer is small [23].

Advantages of affinity partitioning are that a higher degree of ligand use is possible and binding equilibration is more rapid than in fixed-bed processes. Due to the expense of monoclonal antibodies, general affinity ligands such as triazine dyes and nucleotides have been used. The problem of ligand confinement (i.e., recovery and reuse of the ligand) has not been investigated thoroughly, but it depends on the cost of the ligand and the value of the product being purified.

Ion-exchange partitioning is analogous to affinity partitioning. There are only a few documented studies of this phenomenon [23].

5.3. Adsorption Including Ion Exchange

Adsorption is a very common operation in the purification of biological products. The details of adsorption unit operations are presented elsewhere in this volume (→ Adsorption). In this article, only the aspects pertinent to biological separations are discussed, and ion-exchange and affinity adsorption are considered to be specific types of adsorption. These techniques are often termed ion-exchange or affinity chromatography, even though they are not strictly chromatographic operations, but rather adsorption followed by elution.

5.3.1. Sorption Processes

Conventional adsorption (sorption) processes are commonly used as an isolation step in antibiotic purification. Activated carbon, the adsorbent found throughout the chemical process industry, is the traditional adsorbent [1]. Ion-exchange resins, particularly the macroreticular variety, are becoming more popular [64]. Adsorption isotherms usually follow the Freundlich model. Although this type of adsorption is nonspecific, elution yields a smaller volume of solution with a higher product concentration. Fixed-bed operations are most common, but stirred-tank operations are also possible [1].

Whole-Broth Processing. In certain cases, adsorption can be used to combine removal of insolubles and product isolation by adding an adsorbent directly to the unclarified fermentation broth. Screening is used to separate the adsorbent from other insolubles. The advantage of this method is that filtration of the insolubles, an expensive processing step, is not needed. Applications of this concept to ionic antibiotics using ion-exchange adsorbents have been documented [27], [28].

5.3.2. Affinity Adsorption

Proteins may be adsorbed biospecifically on the basis of their interaction with a complementary tertiary structure. This process is termed affinity adsorption and is usually depicted by the "lock

and key" model shown in Figure 6. Since adsorption occurs at specific sites, affinity adsorption usually follows the Langmuir isotherm. An affinity adsorbent may be an antibody (monoclonal), an enzyme cofactor, substrate, inhibitor, or compound with a structure similar to any of these substances. For example, the nucleotide cofactor NAD is involved in many enzymatic reactions and can act as an affinity ligand for any of the associated enzymes. Cibacron Blue F3G-A [*12236-82-7*], a triazine dye with a similar ring structure to a portion of the NAD molecule, is also a good, widely used affinity ligand for these enzymes. Other triazine dyes are also useful in certain cases. Although these adsorbents are not very specific, nonspecific affinity adsorption is still a powerful technique because additional purification can be effected during the elution step. Affinity ligands for various proteins are listed in [65–67]. Many well-developed procedures exist for coupling affinity ligands to solid supports [68]. Some common ligands that are already bound to supports are available commercially.

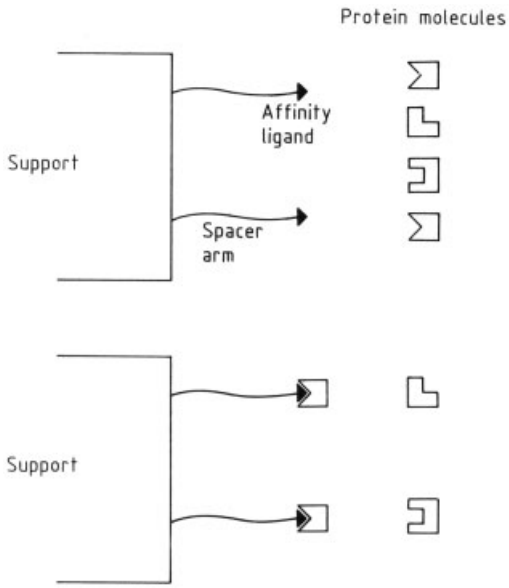

Figure 6. Schematic of "lock and key" mechanism of affinity binding
The affinity adsorbent retains only those molecules that have a complementary tertiary structure.

Monoclonal antibodies represent the ultimate in specificity, but monoclonal antibody technology is still in its infancy (→ Monoclonal Antibodies). The expense of producing these highly specific affinity ligands has limited their application to certain very expensive pharmaceuticals [69].

Affinity adsorption is most often accomplished in a fixed bed, hence the use of the chromatography nomenclature. However, stirred-tank operations with suspended adsorbent are also possible [3].

5.3.3. Ion-Exchange Adsorption (→ Ion Exchangers)

The adsorbents used in ion-exchange adsorption have ionic groups with easily dissociable counterions. Ionic products, such as proteins, amino acids, or certain antibiotics that carry the same charge as the counterion, adsorb by "exchanging" with the counterion. Clearly, the tendency to bind or elute is a function of the isoelectric point of the protein or the p*K* of the antibiotic, and can be manipulated by altering the pH or ionic strength of the solution.

Table 6. Some commercially available ion-exchange products [e]

Type of product	Ion-exchange groups	
	Anionic	Cationic
Cellulose-based	DEAE [a, b] TEAE [a] QAE [a]	CM [a, b] P [a, b]
Sephacel (spherical cellulose beads)	DEAE [c]	
Sephadex (dextran beads)	DEAE [c] QAE [c]	CM [c] SP [c]
Agarose-based	DEAE [a, c] PEI (polybuffer exchanger)	CM [a, c]
Synthetic polymer based (Trisacryl)	DEAE [d]	CM [d]

Suppliers: [a] Bio-Rad; [b] Whatman; [c] Pharmacia; [d] LKBAbbreviations: CM = carboxymethyl; DEAE = diethyl-aminoethyl; P = phospho; PEI = polyethyleneimino; QAE = diethyl(2-hydroxypropyl)aminoethyl; SP = sulfopropyl; TEAE = triethylaminoethyl; [e] Reproduced from [3] with permission of Springer Verlag.

Ion-exchange adsorption is a very useful product isolation step, and high-capacity ion-exchange media are available commercially. Both fixed-bed and stirred-tank processes are possible [1], [3]. Table 6 lists some of the ion-exchange products traditionally used in the purification of biological molecules. A problem common to all large-scale column operations is

bed compression due to the deformation of the packing material [70]. A new generation of supports now exists that are suitable for all column operations; they are significantly more rigid and resist deformation [71]. Examples are listed in Table 7.

Table 7. Rigid supports for large-scale ion-exchange operations in columns

Support	Suppliers
Porous silica	Pharmacia (Uppsala, Sweden; Piscataway, NJ, USA); Syn Chrom, Inc. (Linden, IN, USA); Toyo Soda (Tokyo, Japan); E. Merck (Darmstadt, FRG); Waters Associates (Milford, MA, USA)
Fractogel (cross-linked vinyl polymer)	Toyo Soda (Tokyo, Japan)
Spheron, Separon (hydroxylated methacrylate polymer)	Lachema (Brno, Czech Rep.)

5.3.4. Elution

The adsorbed product must be removed (eluted) from the adsorbent. This is not difficult in conventional adsorption processes where volume reduction is the goal. However, given the relatively low specificity of ion-exchange and many affinity adsorption processes, ingenious elution methods can provide additional purification. However, in any elution process, care must be exercised to avoid increasing the processing volume.

Gradient Elution. Gradient elution is a very common method in which the concentration of a component of the eluent, usually a salt, is increased with time [3]. The salt concentration can be increased stepwise or continuously, most often linearly. The binding strength in ion exchange is a function of the pH and the salt concentration of the eluent, as well as the isoelectric point of the protein. Thus, proteins with different pI values may be eluted separately by subjecting the adsorbent to a salt gradient. Since a portion of affinity adsorption may be ionic in nature, gradient elution may also be useful in this context. Gradient elution is commonly used if the product contains numerous impurities.

Affinity Elution. Affinity elution employs dissolved ligands to preferentially displace adsorbed proteins [3]. Inexpensive ligands must be used here because they are not immobilized.

Magnetic Fluidization. This elution method, the affinity adsorbent is coated on magnetic beads, so that the beads may be fluidized with a magnetic field [72]. Both adsorption and elution occur by countercurrent solid and liquid flow; the affinity ligand is confined to a closed loop within the system. Magnetic supports have also been proposed as methods for separating solid and liquid phases [45], [46].

5.3.5. Quantitative Aspects

Quantitative analysis suggests that three variables affect the fraction of a species α that is adsorbed in affinity or ion-exchange adsorption:

1) the dissociation equilibrium constant, K_p
2) the total concentration of adsorbed and nonadsorbed substrate, p_t
3) the total concentration of free and bound binding sites, m_t [3].

Table 8 shows results of this analysis. Strong binding requires that α approaches unity.

Table 8. Adsorption parameters *

K_p, mmol/L	p_t, mol/L	m_t, mmol/L	α
0.0001	0.01	0.02	0.99
0.001	0.01	0.02	0.92
0.01	0.01	0.02	0.59
0.1	0.01	0.02	0.16
1.0	0.01	0.02	0.02
0.1	0.1	0.1	0.38
0.1	0.5	0.1	0.16
0.1	0.1	0.5	0.81
0.1	0.1	1.0	0.90
0.1	0.5	1.0	0.85
0.1	1.0	1.0	0.73
0.1	0.1	10.0	0.99

* K_p = dissociation equilibrium constant; p_t = total substrate concentration; m_t = total concentration of binding sites; α = fraction of substrate adsorbed. Adapted from [3], with permission of Springer Verlag.

In affinity adsorption beds, $m_t \approx 0.02$ mmol/L and $p_t \approx 0.01$ mmol/L; hence, for strong binding (i.e., $\alpha \to 1$), $K_p \leq 0.001$ mmol/L [3]. Other analyses confirm this [73].

However, strong binding can also result from increasing m_t and/or p_t. A large portion of the binding strength in ion exchange is probably due to the large concentration of binding

sites. Therefore, even though K_p may exceed 0.001 mmol/L in nonspecific affinity binding, binding can be very strong if the ligand concentration m_t is increased or if, as in the case of a genetically altered culture, p_t is increased.

5.4. Precipitation [29]

Under mild conditions, protein precipitation is reversible and subsequent redissolution restores total activity. Protein precipitation may be either a product isolation step or a purification step. In the former case, precipitation and subsequent redissolution in a smaller volume of water not only reduce processing volume, but the resulting solution also contains mostly dissolved protein, free of other soluble contaminants. In the latter case, the differential solubility of proteins may be used for fractionation.

Proteins can be precipitated by changing the pH or temperature or by adding a mild organic solvent, a salt, a multivalent metal ion, or a nonionic polymer [3], [29].

The precipitation method chosen depends on whether the protein precipitate that is formed can be redissolved without loss of activity, the expense of the precipitating agent and its recovery, and the effects of precipitating agent impurities in the precipitate. Models for aggregation and precipitation, as well as precipitation equipment considerations, are described in [29].

Salting Out. High salt concentration promotes protein aggregation and precipitation. Although the mechanism is not well understood, the salt is thought to remove the water of solution from the protein, thereby reducing protein solubility. The Hofmeister series represents decreasing anion effectiveness: citrate > phosphate > sulfate > acetate ≈ chloride > nitrate > thiocyanate. The salts at the low end of this series cause structural damage to proteins. The high solubility of ammonium sulfate [*7783-20-2*] in water and the position of sulfate in the Hofmeister series make it the most popular choice for salting out proteins.

Metal ions such as Mn^{2+}, Fe^{2+}, Ca^{2+}, Mg^{2+}, Ag^{+} also precipitate proteins. These ions bind to different functional groups of the proteins. They act at much lower concentrations than the ions of the Hofmeister series and are easily removed by ion-exchange adsorption or chelating agents.

pH Adjustment. Proteins are soluble in water due to the interaction of their charged groups with water. Adjustment of the pH to the isoelectric point (pI), i.e., the point at which the protein bears zero net charge, yields minimum solubility. Most proteins have a pI < 7, and the relatively low cost of acids makes pH adjustment with acid a popular method of protein precipitation. However, too much acid or base can cause irreversible denaturation.

Organic Solvents. Addition of a mild organic solvent to an aqueous protein solution reduces the solvent dielectric constant, thereby inducing protein precipitation. The solvent must be completely water miscible; ethanol [*64-17-5*] and acetone [*67-64-1*] are the two most common choices. Solvent precipitation is typically performed at low temperature (< 10 °C) because conformational rigidity then hinders irreversible denaturation.

Temperature. Proteins precipitate and denature at different rates when the temperature is increased. However, some robust proteins are resistant to heat denaturation. Therefore, by subjecting an impure mixture to an elevated temperature for an appropriate period of time, a protein can be purified in solution due to the irreversible denaturation and precipitation of the impurities.

Heat denaturation is a first-order rate process and the rate constant obeys the Arrhenius law. Empirical measurement of the activation energies of the proteins involved is necessary to determine the optimal time and temperature conditions [29].

Polymers and Polyelectrolytes. Nonionic, water-soluble polymers induce protein precipitation by excluding water from the solvation structure of a protein. Poly(ethylene glycol) [*25322-68-3*] is the most widely studied and widely used polymer, but dextrans are also employed. The effect of polymers as precipitating agents can be considered to be similar to partitioning in aqueous two-phase polymer systems (cf. Section 5.2.2).

High poly(ethylene glycol) concentrations are required to precipitate low molecular mass

proteins, whereas low concentrations are required for high molecular mass proteins. Poly(ethylene glycol) with a high molecular mass induces precipitation at a lower concentration than does poly(ethylene glycol) with a low molecular mass. Poly(ethylene glycol) with a maximum molecular mass of ca. 8000 is desirable; concentrations do not usually exceed 20 wt %.

Polyelectrolytes such as poly(acrylic acid), carboxymethyl cellulose, and polyethyleneimines precipitate proteins at a much lower concentration (usually < 0.1 wt %) than nonionic polymers. They act more like flocculants and adsorb to the protein. Thus, polyelectrolytes, unlike poly(ethylene glycol), coprecipitate with the protein and can cause irreversible denaturation. Polyelectrolyte precipitation is used mainly for food proteins, which are quite robust [29].

Affinity Precipitation. Affinity precipitation is used mainly as a laboratory technique [74], although one industrial patent has been reported [75]. This method employs a ligand that is bound to the two ends of a spacer arm consisting of carbon chains that sometimes include nitrogen atoms. Most enzymes are multimeric and usually contain two or four identical subunits, each subunit containing an active site. The spacer arm and ligands can form a large network with the multimeric enzyme, which precipitates due to its large molecular mass. This method has been used to precipitate lactate dehydrogenase with a bis(NAD) ligand [74]:

$$\mathrm{NAD{-}CH_2\,CONHNHCO(CH_2)_4\,CONHNHCOCH_2{-}NAD}$$

5.5. Solubilization

Solubilization is the incorporation of an otherwise insoluble species into solution and may be achieved by using a surfactant (detergent). Surfactants contain polar and nonpolar regions and form aggregates (micelles) to minimize unfavorable hydrocarbon – water interactions. In aqueous solutions, insoluble nonpolar species are incorporated into the hydrophobic core of the micelle and are thus solubilized. In hydrocarbon phases, reverse micelles form, and insoluble polar species can be analogously solubilized by incorporation into the hydrophilic core of the reverse micelle. Reverse-micelle solubilization is used as a product isolation step (see Section 5.2.1) [62].

Solubilization is used in the separation of membrane-bound proteins, which are inherently not very soluble in water [30–34]. After cell lysis, addition of an appropriate surfactant solubilizes the membrane-bound proteins. This technique is employed in the laboratory; no large-scale commercial operations are documented. Selection of surfactants is highly empirical but nonionic surfactants are mostly used, Triton X-100 [*9002-93-1*] being the most common.

A further application of solubilization is in the purification of proteins from insoluble inclusion bodies [35]. Inclusion bodies are formed in recombinant DNA technology when microorganisms, especially *Escherichia coli,* overproduce a particular protein. The nonionic detergent Triton X-100 solubilizes the protein granules in the presence of high urea concentrations (5 – 8 mol/L) [36], [37].

6. Purification

The primary methods for purification of biochemical products are chromatography and electrical separations; magnetic separations are less important. Ultrafiltration is also used to concentrate macromolecules and remove small, dissolved solutes and is discussed in Section 4.1.

6.1. Chromatography

Chromatography is the separation of species due to different migration rates through a stationary medium, which is usually a fixed bed. Gel permeation (filtration) and affinity chromatography are most commonly used for protein separations. Although ion exchange is often associated with the term chromatography, it is most often an adsorption – elution operation and is discussed in Section 5.3.

General principles of chromatography are presented elsewhere (→ Chromatography) and industrial-scale chromatography is described in → Process-Scale Chromatography. Biochemical applications are discussed in this section.

In chromatographic separations, a pulse of solution is applied to the column by injecting it into a large volume of flowing carrier liquid. Separation occurs due to the different migration rates of the dissolved species. As a result of dispersion (zone spreading), each species leaves the fixed bed as a bell-shaped peak. Optimal resolution requires minimal zone spreading and significant differential migration. If there is little differential migration and significant zone spreading, resolution is very small, i.e., separation is poor.

The laboratory workhorse is high-pressure liquid chromatography (HPLC), in which high pressure enhances speed and resolution but throughputs are very small. Fast protein, polypeptide, polynucleotide chromatography (FPLC, Pharmacia) performs rapid separations are possible at moderate pressure, again on a very small scale.

Chromatography is inherently a batch operation and, hence, is not very amenable to scaleup. Attempts have been made at achieving continuous operation, most notably, the continuous annular chromatograph [76]. To avoid problems caused by pressure drop and bed compression effects, industrial scaleup procedures inevitably yield short, fat beds, that are often stacked together [70].

6.1.1. Gel Filtration Chromatography

In gel filtration (gel permeation) chromatography, differential migration of globular proteins is achieved due to differences in molecular mass. The column packing is completely inert; only its pore structure is involved in the separation. Lower molecular mass proteins permeate the pores of the packing material, while larger molecular mass proteins do not and, thus, are eluted first.

Gel filtration does not afford precise separation, however. The best possible separation is exclusion of 90 % of the macromolecules that are 5 – 6 times larger than the macromolecules entering the pores [3]. Therefore, this method is most suitable for separating very large from very small macromolecules.

Gel filtration in chromatographic beds is described in [8]. Table 9 summarizes traditional gel filtration media. More rigid media resist deformation [77], a common problem in large-scale column operations (cf. Table 7).

6.1.2. Affinity Chromatography

In affinity chromatography, differential migration is achieved due to the affinity of proteins for surface-bound ligands. The binding strength (cf. Section 5.3.5) need not be very high due to the large size of the bed. Chromatography beds often have $10^3 - 10^6$ theoretical plates; therefore, even a small difference in affinity can yield good resolution as long as it is not compromised by zone spreading. If binding is too strong the operation ceases to be chromatography and becomes an adsorption – elution operation (cf. Section 5.3).

Affinity ligands must be attached to a support medium that usually contains large pores to avoid gel filtration effects and to maximize usable surface area. More than 75 % of all affinity chromatography is performed on agarose supports (e.g., Biogel A series, Pharmacia Sepharoses, LKB Agaroses) [70]. This is due to the high permeability, low nonspecific adsorption, and chemical stability of agarose. Procedures for coupling affinity ligands to agaroses are well established [68]. The most common has been activation using cyanogen bromide (CN-Br) [*506-68-3*], however, instability of the bond formed with polysaccharides has spawned numerous other techniques [67], [78]. Furthermore, the problem of bed compression in large-scale operations has led to increased use of silica as an affinity support [79].

The choice of an affinity ligand is based on the chemistry involved and the cost as discussed in Section 5.3.2. Protein – ligand pairs are tabulated in the literature [65–67]. Monoclonal antibodies are very expensive to produce and may bind their substrates so strongly that chromatography is not feasible; elution under different conditions may be required [69]. Triazine dyes are inexpensive, but less selective. Nucleotides and other cofactors or coenzymes are more expensive, but more selective.

6.1.3. Quantitative Aspects

Although chromatography is largely an empirical science, quantitative prediction of elution

Table 9. Gel filtration media ∗

Company	Gel code	Gel type	Useful working range for globular molecules, M_r
Bio-Rad Biogels	P-2 ∗∗	polyacrylamide	100 – 1800
	P-4 ∗∗	polyacrylamide	800 – 4000
	P-6 ∗∗	polyacrylamide	1 000 – 6000
	P-10 ∗∗	polyacrylamide	1 500 – 20 000
	P-30 ∗∗	polyacrylamide	2 500 – 40 000
	P-60	polyacrylamide	3 000 – 60 000
	P-100	polyacrylamide	5 000 – 100 000
	P-150	polyacrylamide	15 000 – 150 000
	P-200	polyacrylamide	30 000 – 200 000
	P-300	polyacrylamide	60 000 – 400 000
	A-0.5 m	agarose	$1\,000 - 0.5 \times 10^6$
	A-1.5 m	agarose	$2\,000 - 1.5 \times 10^6$
	A-5.0 m	agarose	$4\,000 - 5 \times 10^6$
	A-15 m ∗∗	agarose	$60\,000 - 15 \times 10^6$
	A-50 m ∗∗	agarose	$200\,000 - 50 \times 10^6$
	A-150 m ∗∗	agarose	$1 \times 10^6 - 150 \times 10^6$
LKB Ultrogels	AcA22	agarose–polyacrylamide	$60\,000 - 1 \times 10^6$
	AcA34	agarose–polyacrylamide	20 000 – 400 000
	AcA44	agarose–polyacrylamide	$12\,000 - 1.3 \times 10^6$
	AcA54	agarose–polyacrylamide	6 000 – 70000
LKB Agaroses	A2 ∗∗	agarose	$120\,000 - 20 \times 10^6$
	A4 ∗∗	agarose	$55\,000 - 9 \times 10^6$
	A6 ∗∗	agarose	$25\,000 - 2.4 \times 10^6$
Pharmacia Sephadex	G-10 ∗∗	dextran	50 – 700
	G-15 ∗∗	dextran	50 – 1500
	G-25 ∗∗	dextran	1 000 – 5000
	G-50	dextran	1 500 – 30 000
	G-75	dextran	3 000 – 70 000
	G-100	dextran	4 000 – 150 000
	G-150	dextran	5 000 – 300 000
	G-200	dextran	5 000 – 600 000
Pharmacia Sepharoses	6B	agarose	$10\,000 - 4 \times 10^6$
	4B	agarose	$60\,000 - 20 \times 10^6$
	2B ∗∗	agarose	$70\,000 - 40 \times 10^6$
	CL-6B	cross-linked agarose	$10\,000 - 4 \times 10^6$
	CL-4B	cross-linked agarose	$60\,000 - 20 \times 10^6$
	CL-2B ∗∗	cross-linked agarose	$70\,000 - 40 \times 10^6$
Pharmacia Sephacryls	S-200	dextran-bisacrylamide	5 000 – 300 000
	S-300	dextran-bisacrylamide	10 000 – 800 000
	S-400	dextran-bisacrylamide	$20\,000 - 2 \times 10^6$
	S-500	dextran-bisacrylamide	$> 10^6$ (multienzyme complexes)
	S-1000 ∗∗	dextran-bisacrylamide	

∗ Reproduced from [3] with permission of Springer Verlag.
∗∗ Not ideally suited for protein fractionation.

times and resolution is possible. Two models are usually used: (1) the fixed-bed adsorption theory involving the continuum conservation equations [80], and (2) the equilibrium stage model, in which the fixed bed is modeled as a succession of equilibrium stages [60]. The equilibrium stage model is simpler, but neither predicts the number of theoretical stages required nor is sensitive to the effect of operating conditions. The fixed-bed model is completely rigorous, but difficult to solve analytically and requires numerical solutions.

The interrelationship between the affinity desorption coefficient, ligand concentration, and protein concentration is as important in predicting chromatography performance as it is in adsorption (Section 5.3.5); for a detailed quantitative discussion, see [73].

6.2. Electrical Separations

(→ Electrophoresis)

Electrophoresis and dielectrophoresis form the basis of the electrical separation of biological species. *Electrophoresis* is the migration of a charged species toward an electrode of opposite charge in an applied electric field; monopoles, i.e., permanent charges, define the direction of motion. *Dielectrophoresis* is the migration of a particle toward regions of field nonuniformity in a nonuniform electric field; migration is independent of the sign or existence of a permanent charge and involves uncharged species.

6.2.1. Electrophoresis

Theory. The electrophoretic migration velocity v_E of a particle is directly proportional to the applied electric field:

$$v_E = u_E E$$

where

E = electric field, V/cm
u_E = electrophoretic mobility, $cm^2\,V^{-1}\,s^{-1}$.

The value of u_E for spherical particles depends on (1) the radius of the particle, R and (2) the Debye parameter, which is given by

$$\kappa = \left(8\pi n e^3/\varepsilon k T\right)^{1/2}$$

where $n = 1/2\,\Sigma\, z_i^2\, c_i$

n = solution ionic strength
z_i = valence number of species i
c_i = molar concentration of species i
e = charge on an electron
ε = solvent dielectric constant
k = Boltzmann's constant
T = absolute temperature

When $\kappa R \ll 1$, the Hückel theory applies and $u_E = \varepsilon\xi/4\pi\mu$, where μ is the solvent viscosity and ξ is the zeta potential of the particle. The zeta potential is the potential at the surface of shear of the particle and, for practical purposes, may be considered to be equal to the surface potential. When $\kappa R \gg 1$, the Smoluchowski theory applies and $u_E = \varepsilon\xi/6\pi\mu$. The Hückel theory disregards the deformation of electric field lines by the particle and is only valid for small particles. The Smoluchowski theory assumes that electric field lines are always uniform and parallel to the particle surface; it is valid only for large particles. When $\kappa R \approx 1$, u_E passes through a minimum due to the combination of the retardation and relaxation effects. Analytical solutions are not available for $\kappa R \approx 1$, but computer-generated values exist [81]. For all aqueous systems at room temperature, $\kappa \approx 3 \times 10^8\, n^{1/2}$ cm, n being expressed in mol/L. For proteins in solution at typical ionic strengths ($\geq 10^{-3}$ mol/L), $\kappa R \leq 1$, but the value may be sufficiently close to unity to require a computer-generated result for u_E. For cells at typical ionic strengths ($\geq 10^{-3}$ mol/L), $\kappa R \gg 1$.

Applications. From a practical standpoint, precise values for u_E are not needed for protein separations. It is usually sufficient to know the difference between the isoelectric point (p I) of the protein involved and the pH, and to proceed empirically; (the p I is the pH value at which the protein has zero net charge). Since proteins acquire their charge by ionization of amino and carboxylic acid groups, the difference pI − pH represents both the magnitude of the charge on the protein and its sign. If the pH is less than pI, i.e., pI − pH > 0, then the protein is positively charged. If pI − pH < 0, the protein is negatively charged.

Typical useful values for the applied electric field are in the range of $10^3 - 10^4$ V/m, which can cause significant resistance heating in the solution. Heating creates convection currents, which compromise protein resolution. The magnitude of convection due to heating is quantified by the Rayleigh number *Ra:*

$$Ra = \frac{\varrho\beta\Delta T L^3}{\mu\alpha}$$

where

ϱ = average solution density, g/mL
β = coefficient of volume expansion, K^{-1}
ΔT = temperature difference, K
L = length, cm
μ = solution viscosity, $g\,cm^{-1}\,s^{-1}$
α = thermal diffusivity, cm^2/s

If *Ra* is small, convection is unimportant. Small *Ra* values are obtained if μ is large, as in gel electrophoresis. This technique is a popular analytical tool in the laboratory, but is not

a practical industrial process. A second method is to reduce L by, for example, performing electrophoretic separations through a membrane. In this case L corresponds to the membrane pore diameter, not the overall size of the system. However, systems utilizing this concept are not currently used in industry.

Considerable research has been aimed at developing a device for continuous-flow electrophoresis, a desirable operation for commercial processing, but convection problems have still not been resolved. Practical aspects of electrophoresis are reviewed in [39], [40]. A novel concept combines electrophoresis and gel filtration chromatography to focus components of a mixture of proteins [41].

6.2.2. Isoelectric Focusing and Isotachophoresis

In isoelectric focusing, a pH gradient is established with polymeric ampholytes, which migrate in the direction of the electric field. Proteins migrate until they reach their pI (isoelectric pH), where they carry no net charge. Subsequent protein diffusion in either direction is counteracted by migration back toward the pI, hence the term "focusing". This high-resolution separation method is applicable only to proteins that remain soluble at their pI and do not interact with the ampholytes used to establish the pH gradient [3].

The ampholytes migrate faster than the proteins; thus the pH gradient is established more rapidly than the rate of protein migration.

Isoelectric focusing is not a popular large-scale process because the ampholytes are expensive; separation of the product and the ampholytes also presents a problem. Improvements to materials and methods for establishing the pH gradient have been suggested [82].

Heating is also a problem, and, as in electrophoretic separations, gels are often used. A continuous-flow isoelectric focusing device has been developed for preparative-scale operations, that has minimal convection and good heat dissipation characteristics [83]. A membrane sandwich device, which has discrete pH values in separate compartments, has also been suggested for potential commercial-scale applications [40], [82].

In *isotachophoresis,* the electric field varies as well as the pH. The theory of the method is complex, but it is a potentially valuable preparative separation technique. A few successes have been documented in the laboratory [3].

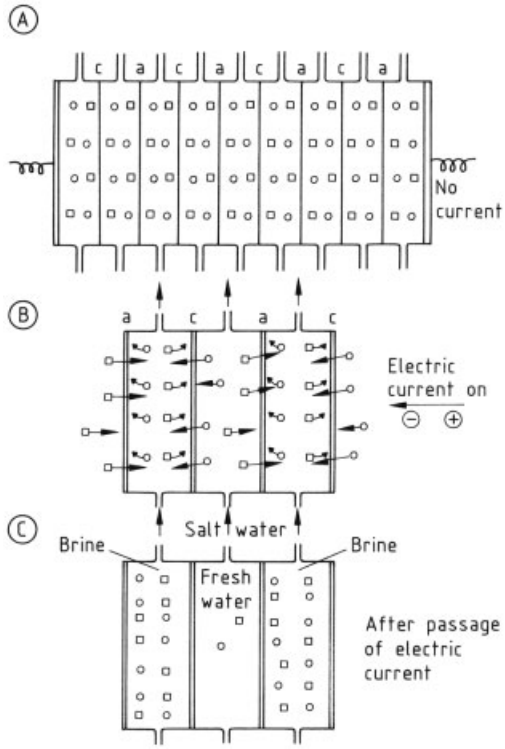

Figure 7. Principle of electrodialysis ∗
(A) Apparatus: a) Anion-permeable membrane; c) Cation-permeable membrane; ○ = positive ion (e.g., sodium); □ = negative ion (e.g., chloride)
Ion migration under action of electrical current (B) causes salt depletion in alternate compartments and salt enrichment in adjacent ones (C).
∗ Reproduced from [42], with permission of Plenum Press.

6.2.3. Electrodialysis (→ Membranes and Membrane Separations)

In electrodialysis, the electrolyte concentration of a solution is changed by the migration of the electrolytes in an electric field through ion-exchange membranes. Anion-exchange membranes are sandwiched alternately with cation-exchange membranes. Figure 7 illustrates the principles of electrodialysis. The dissociation of water into H^+ and OH^- ions provides the coun-

terions, so that the cation and anion of the electrolyte can migrate in opposite directions toward the anode and cathode, respectively. Due to the arrangement of the membranes, alternate compartments become depleted and enriched in electrolyte.

The major application for electrodialysis is the desalination of water (→ Water) [42], but it can also be used to remove salt from protein solutions, possibly after precipitation. In the arrangement shown in Figure 7C, protein solution is simply filled into the "fresh water" compartment. The ion-exchange membrane is impermeable to the protein, and the salt is then removed from the protein solution as described above.

6.2.4. Dielectrophoresis

Dielectrophoresis is the migration of polarizable material in a nonuniform electric field toward the region of highest field nonuniformity. Unlike electrophoresis, permanent charges are not required and an alternating field can still provide unidirectional migration. Dielectrophoresis is discussed comprehensively in [43].

Dielectrophoresis is still mainly a laboratory curiosity; no industrial applications are documented. However, extensive research on cell migration by means of dielectrophoresis, may lead to applications in cell sorting operations for preparative or characterization purposes.

6.3. Magnetic Separations

In magnetic separations, species with high magnetic susceptibilities migrate to regions of high magnetic field strength. Therefore, magnetic separations are analogous to dielectrophoresis in electrical separations. High magnetic field strengths are required, and the term high gradient magnetic separations (HGMS) is often used.

Three different classes of magnetic properties exist. *Ferromagnets* contain unpaired electrons that create a strong magnetic susceptibility. These include iron, cobalt, nickel, and their oxides, such as magnetite (Fe_3O_4). *Paramagnets* contain unpaired electrons that create a low magnetic susceptibility. Examples include molecules, such as hemoglobin, containing transition metals.

Dimagnets, such as silicon dioxide, graphite, and most organic compounds, contain no unpaired electrons, and in a strict sense are repelled by a magnetic field, although in practice the degree of repulsion may be insignificant [46].

The most common application of magnetic separation for biological species involves coating ferromagnetic particles, usually magnetite, with a polymer that can act as an affinity support. The affinity supports may then be used to separate affinity-bound proteins from impurities by isolating the support magnetically [84]. L-Asparaginase has been isolated in this way [85]. Adsorption beds may also be fluidized magnetically. Collection of red blood cells based on the paramagnetic properties of hemoglobin has also been reported [86]. Reviews on the biotechnological applications of magnetic separations are available [45], [46].

7. Polishing

7.1. Crystallization

Crystallization is often used as a final purification step to produce high-purity antibiotics and other solid consumer products. Products are usually crystallized by lowering the temperature. The goal is usually the dewatering of a pure product. Fractional crystallization is also possible, thereby purifying one material while leaving contaminants in solution. However, phase diagrams necessary for fractional crystallization are not readily available, especially for biological products. Thus, crystallization in the biochemical process industry is very empirical. Advantages of crystallization include operation at a comparatively low temperature thus avoiding protein denaturation, and high product concentrations which keep the cost low [1].

In precipitation, solubility is decreased by a chemical additive, which may coprecipitate. The goal is usually volume reduction by redissolution in a smaller volume of water, or fractionation.

A more detailed description of crystallization and precipitation is presented elsewhere (→ Crystallization and Precipitation).

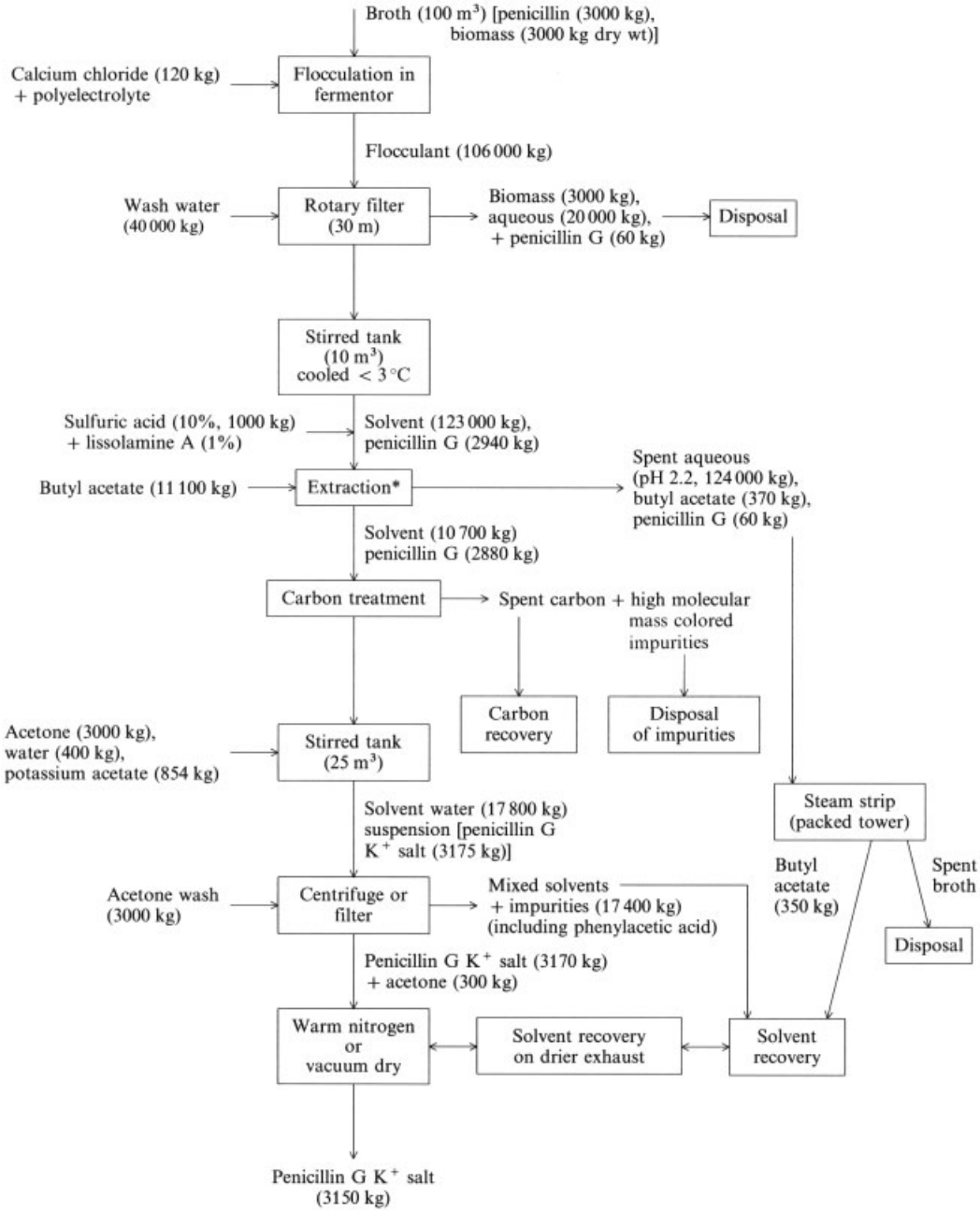

Figure 8. Extraction of penicillin G potassium salt, based on a 100 m^3 batch giving 3150 kg of penicillin G potassium salt ∗∗
∗ Two or more stages involved. Two Podbielniak D 36 machines can process 100 m^3 in 12 h including cleaning time. Lissolamine A [*57-09-0*], a quaternary ammonium surfactant, acts as an emulsion breaker.
∗∗ Reproduced from [2], with permission of Macmillan Publishers, London and Basingstoke.

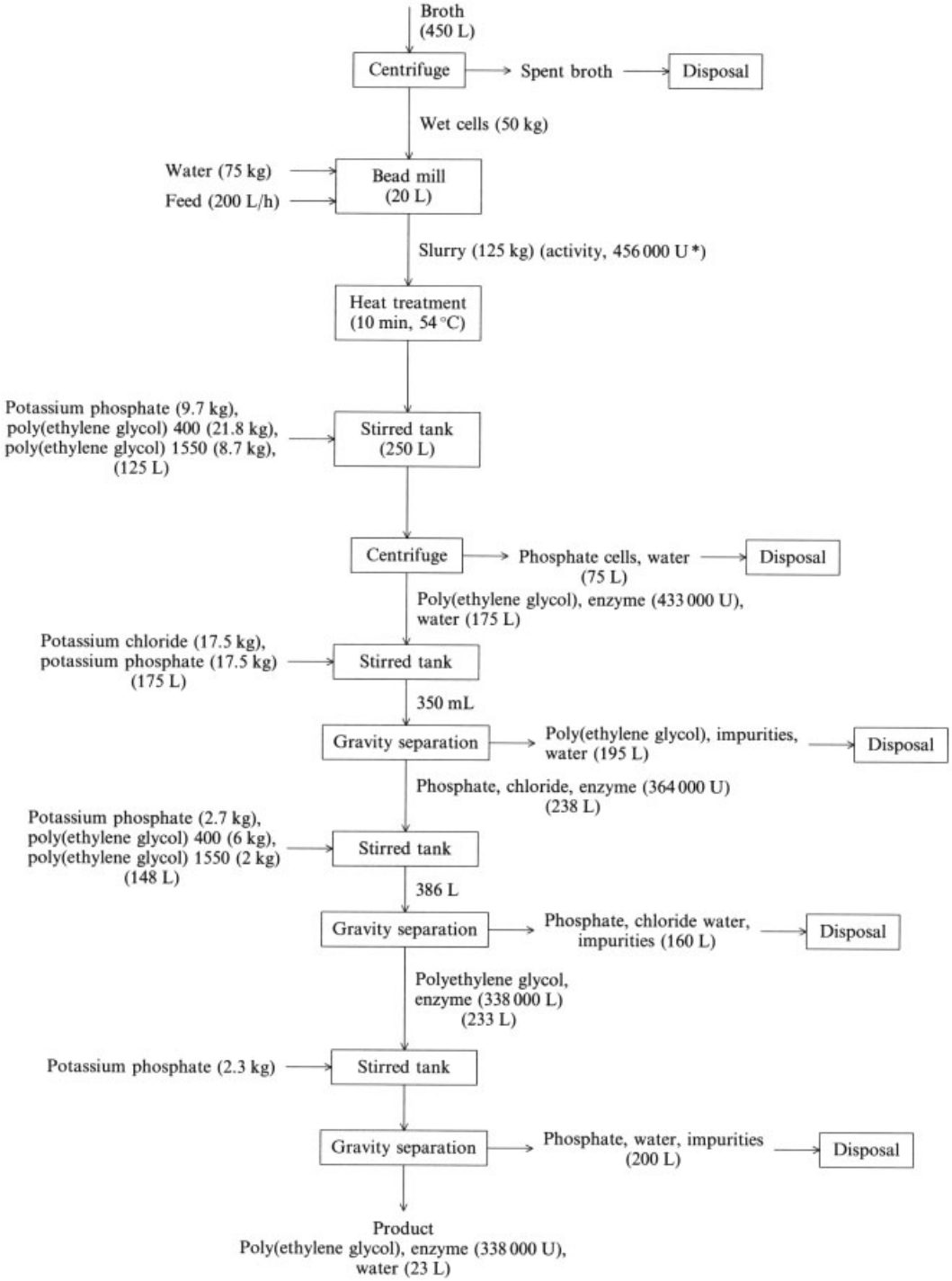

Figure 9. Semitechnical-scale example of an aqueous two-phase extraction process for an intracellular enzyme based on a 0.5 m^3 fermentor producing 111 g of enzyme protein ∗∗
The enzyme is separated from poly(ethylene glycol) in the product at the point of use by ultrafiltration.
∗ 1 U = 1 μmol of substrate converted per minute.
∗∗ Reproduced from [2], with permission of Macmillan Publishers, London and Basingstoke.

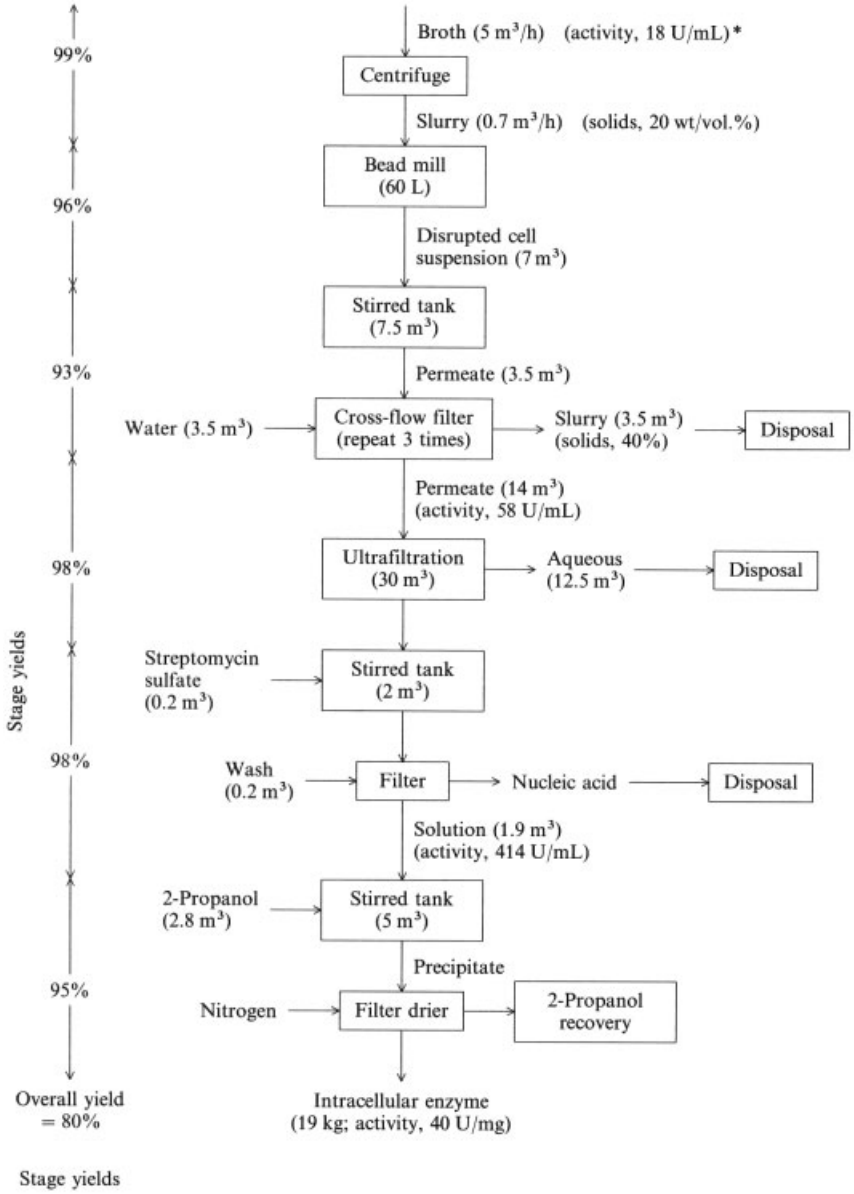

Figure 10. Typical flowsheet for production of an intracellular enzyme based on a 50 m^3 broth ∗∗
Processing time = 12 h (operating time = 10 h, cleaning time = 2 h). Activity figures are appropriate for amidase.
∗ 1 U = 1 µmol of substrate converted per minute.
∗∗ Reproduced from [2], with permission of Macmillan Publishers, London and Basingstoke.

7.2. Drying

Drying, if used, usually is the final purification step and often follows crystallization. The principles of drying are discussed elsewhere (→ Drying of Solid Materials), and few, if any, features are unique to drying bioproducts. Methods are applicable to pharmaceuticals, food products, or enzymes. Although many enzymes cannot withstand the high temperature associated with spray drying [1], this technique is sometimes applicable, because the ex-

posure time to the increased temperature is usually short. Lyophilization (freeze drying) is used when a high temperature must be avoided [47], [48].

8. Bioseparation Strategies

Biological separations involve numerous separation steps, the processing of large volumes of dilute product, the handling of fragile molecules, the need for a high-purity consumer product, and the need to purify a single component from hundreds of similar components. General strategies for selecting bioseparation processes are treated in [60], [87]. Furthermore, methods for classifying separation operations, which suggest ways of conceiving new separation methods, are available [88], [89].

In planning a bioseparation strategy, the general sequence of operations, (1) removal of insolubles, (2) product isolation, (3) product purification, and (4) polishing, should be maintained. However, if necessary, a step may be eliminated or steps may be combined as in whole-broth processing (Section 5.3.1). One of the major rules in selecting separation sequences is that difficult separations should be performed last. This rule is especially important in view of the large processing volumes involved in bioseparations and the difficulty of protein fractionation.

Most future novel separations and separation sequences will be concerned with proteins (i.e., antibodies, enzymes, and hormones). For example, potential production quantities for bovine growth hormone are so large that improvements in its purification would have a major impact on process economics. Important variables that must be considered when designing a protein separation sequence are as follows:

1) protein characteristics
   - molecular mass
   - isoelectric point (charge)
   - biological affinity
   - solubility
2) process characteristics
   - anticipated impurities and their characteristics
   - denaturation sensitivity to heat or shear
   - product concentration
3) cell characteristics
   - difficulty of cell wall rupture
   - intra- or extracellular product
   - if product is intracellular, is it membrane-bound or located in an inclusion body?

This list is not exhaustive because considerations particular to a given protein may introduce other variables.

Three examples of bioseparations are presented. Figure 8 is a flowsheet for penicillin production. Although the technology involved may be outdated, it illustrates the separation sequence concept. The general processing sequence described above is followed: filtration is used to remove insolubles, extraction and adsorption are used for reducing the processing volume and for purification, and the solid product is finally recovered by centrifugation or filtration. For a high-purity product, subsequent processing is required [2].

Figures 9 and 10 are flowsheets for the production of intracellular enzymes. Both obey the general processing sequence and briefly illustrate the variety of processing sequences from which to choose for biological product recovery. In both processes, centrifugation is used to concentrate the cells and a bead mill is used to rupture them. In Figure 9, aqueous two-phase extraction is used for cell debris removal, volume reduction, and purification. In Figure 10, crossflow filtration is used for cell debris removal, followed by ultrafiltration for volume reduction and precipitation for purification. Further examples are described in [2].

9. References

1. P. Belter in H. J. Peppler, D. Perlman (eds): *Microbial Technology,* 2nd ed., **vol. 2,** Academic Press, New York 1979, Chap. 16, p. 403.
2. B. Atkinson, F. Mavituna: *Biochemical Engineering and Biotechnology Handbook,* Macmillan, London-Basingstoke 1983, Chap. 12, 13.
3. R. Scopes: *Protein Purification,* 2nd ed., Springer Verlag, New York 1987.
4. J. M. Curling (ed.): *Methods of Plasma Protein Fractionation,* Academic Press, New York 1980.
5. J. A. Asenjo, J. Hong (eds.): *Separation, Recovery and Purification in Biotechnology,*

ACS Symp. Ser. no. 314, Am. Chem. Soc., Washington, DC, 1986.
6. D. LeRoth, J. Shiloach, T. J. Leahy: *Purification of Fermentation Products,* ACS Symp. Ser. no. 271, Am. Chem. Soc. Washington, DC, 1985.
7. A. Fiechter (ed.): *Reaction Engineering, Advances in Biochemical Engineering,* vol. 24, Springer Verlag, New York 1982.
8. A. Fiechter (ed.): *Chromatography, Advances in Biochemical Engineering,* vol. 25, Springer Verlag, New York 1982.
9. A. Fiechter (ed.): *Downstream Processing, Advances in Biochemical Engineering/Biotechnology,* vol. 26, Springer Verlag, New York 1983.
10. W. C. McGregor (ed.): *Membrane Separations in Biotechnology,* Marcel Dekker, New York 1986.
11. T. Becker, J. R. Ogez, S. E. Builder, *Biotechnol. Adv.* **1** (1983) 247.
12. B. Volesky, J. H. T. Luong, *CRC Crit. Rev. Biotechn.* **2** (1984) 119.
13. H-C. Janson, *Trends Biotechnol.* **2** (1984) 31.
14. D. I. C. Wang, C. L. Cooney, A. L. Demain, P. Dunnill, A. E. Humphrey, M. D. Lilly: *Fermentation and Enzyme Technology,* J. Wiley & Sons, New York 1979, Chap. 12, p. 238.
15. J. E. Bailey, D. F. Ollis: *Biochemical Engineering Fundamentals,* 2nd ed., McGraw-Hill, New York 1986, Chap. 11, p. 726.
16. W. B. Jakoby (ed.): *Enzyme Purification and Related Techniques, Part C, Methods in Enzymology,* vol. 104, Academic Press, New York 1984.
17. J. D. Stowell, P. J. Bailey, D. J. Winstanley (eds.): *Bioactive Microbial Products 3, Downstream Processing,* Academic Press, London 1986.
18. P. A. Belter, E. L. Cussler, W.-S. Hu: *Bioseparations, Downstream Processing for Biotechnology,* Wiley-Interscience, New York 1988.
19. J. D. Henry, Jr., R. C. Allred, *Dev. Ind. Microbiol.* **13** (1972) 177.
20. J. D. Henry in N. N. Li (ed.): *Recent Developments in Separation Science,* **vol. 2,** CRC Press, Boca Raton, FL, 1972, p. 205.
21. D. K. Tressler, M. A. Joslyn: *Fruit and Vegetable Juice Processing Technology,* 2nd ed., AVI, Westport, Conn., 1971.
22. J. L. Bomber, S. Bruin, H. A. C. Thijssen, R. L. Merson in : G. F. Stewart, E. Mrats, C. O. Chichester (eds.): *Advances in Food Research,* **vol. 20,** Academic Press, New York 1973, p. 1.
23. M.-R. Kula, K. H. Kroner, H. Hustedt in Reference 7, p. 73.
24. K. H. Kroner, H. Hustedt, M.-R. Kula, *Biotechnol. Bioeng.* **24** (1982) 1015.
25. H. Walter, D. E. Brooks, D. Fisher (eds): *Partitioning in Aqueous Two-Phase Systems,* Academic Press, New York 1985.
26. P.-A. Albertsson: *Partition of Cell Particles and Macromolecules,* 3rd ed., Wiley-Interscience, New York 1986.
27. C. R. Bartels, G. Kleinman, J. N. Korzum, D. B. Frish, *Chem. Eng. Prog.* **54** (1958) no. 8, 49.
28. P. Belter, *Biotechnol. Bioeng.* **15** (1973) 533.
29. D. J. Bell, M. Hoare, P. Dunnill, in Reference [9], p. 1.
30. A. Helenius, K. Simons, *Biochim. Biophys. Acta* **415** (1975) 29.
31. L. M. Hjelmeland, A. Chrambach, *Methods Enzymol.* **104** (1984) 305.
32. J. van Renswoude, C. Kempf, *Methods Enzymol.* **104** (1984) 329.
33. H. S. Penefsky, A. Tzagoloff, *Methods Enzymol.* **22** (1971) 204.
34. A. Tzagoloff, H. Penefsky, *Methods Enzymol.* **22** (1971) 219.
35. T. Atkinson, D. A. Barstow, J. Court, N. P. Minton, R. J. Sharp, R. Sherwood, in Reference 17, p. 27.
36. F. A. O. Marston, P. A. Lowe, M. T. Doel, J. M. Schoemaker, S. White, S. Argal, *Bio/Technology* **2** (1984) 800.
37. R. G. Schoner, L. F. Ellis, B. E. Schoner, *Bio/Technology* **3** (1985) 151.
38. D. C. Preive, E. Ruckenstein in M. Kerker (ed.): *Colloid and Interface Science,* **vol. 4,** Academic Press, New York 1976, p. 73.
39. C. J. van Oss, *Sep. Purif. Methods* **8** (1980) 119.
40. H. Hirai (ed.): *Electrophoresis '83',* De Gruyter, New York 1984.
41. P. O'Farrell, *Science* **227** (1985) 1586.
42. K. S. Spiegler, *Salt Water Purification,* 2nd ed., McGraw-Hill, New York 1980.
43. H. A. Pohl, *Dielectrophoresis,* Cambridge Press, New York 1978.
44. E. Flaschel, Ch. Wandrey, M.-R. Kula, in Reference 9, p. 73.
45. E. H. Dunlap, W. A. Feiler, M. J. Mattione, *Biotechnol. Adv.* **2** (1984) 63.
46. B. L. Hirschbein, D. W. Brown, G. M. Whitesides, *Chemtech.* **12** (1982) 172.

47. C. J. King: *Freeze Drying of Foods,* CRC Press, Boca Raton, FL, 1971.
48. D. W. Moore, *Pharm. Technol.* **7** (1983) no. 3, p. 84.
49. R. Gabler, M. Ryan, in Reference 6, p. 1.
50. W. C. McGregor, in Reference 10, p. 1.
51. B. Mattiasson, T. G. I. Ling, in Reference 10, p. 99.
52. J. J. S. Shen, R. S. Probstein, *Ind. Eng. Chem. Process Des. Dev.* **18** (1979) 547.
53. E. W. Pitera, S. Middleman, *Ind. Eng. Chem. Process Des. Dev.* **12** (1973) 52.
54. J. D. Henry, Jr., L. F. Lawler, C.-H. A. Kuo, *AIChE J.* **23** (1977) 851.
55. J. D. Henry, Jr., in R. H. Perry, D. Green (eds.): *Perry's Chemical Engineers Handbook,* 6th ed., McGraw-Hill, New York 1984, Chap. 17.
56. J. M. Radovich, N. S. Mason, R. E. Sparks, *Sep. Sci. Technol.* **15** (1980) 1491.
57. H. Brenner in T. B. Drew, J. W. Hooper, Jr., T. Vermeulen (eds.): *Advances in Chemical Engineering,* **vol. 6,** Academic Press, New York 1966, p. 287.
58. G. B. Cline, R. B. Ryel, *Methods Enzymol.* **22** (1971) 168.
59. J. J. Round, R. A. Liptak, W. C. McGregor, *Ann. NY. Acad. Sci.* **369** (1981) 265.
60. C. J. King: *Separation Processes,* 2nd ed., McGraw-Hill, New York 1977.
61. M. Sounders, G. J. Pierotti, C. L. Dunn, *AIChE Symp. Ser.* **66** (1979) no. 100, 37.
62. K. E. Goklen, T. A. Hatton, *Biotechnol. Progress* **2** (1985) 69.
63. M. P. Thien, T. A. Hatton, D. I. C. Wang, in Reference 5, p. 67.
64. B. Boothroyd, in Reference 17, p. 59.
65. I. Parikh, P. Cuatrecasas, *Chem. Eng. News* **63** (1985) no. 34, 17.
66. C. M. Sturgeon, J. F. Kennedy, *Enzyme Microb. Technol.* **8** (1986) 125.
67. M. Wilchek, T. Miron, J. Kohn, *Methods Enzymol.* **104** (1984) 3.
68. J. Porath, *Methods Enzymol.* **34** (1974) 13.
69. H. A. Chase, *Chem. Eng. Sci.* **39** (1984) 1099.
70. J.-C. Janson, P. Hedman, in Reference 8, p. 43.
71. F. Regnier, in Reference 16, p. 170.
72. M. A. Burns, D. J. Graves, *Biotechnol. Progress* **2** (1985) 95.
73. D. J. Graves, Y.-T. Wu, *Methods Enzymol.* **34** (1974) 140.
74. P.-O. Larsson, K. Mosbach, *FEBS Lett.* **98** (1979) 333.
75. M. Schneider, C. Buillota, B. Lami, US 4 066 505, 1978.
76. J. M. Begovich, W. G. Sisson, *AIChE J.* **30** (1984) 705.
77. K. Unger, in Reference 16, p. 154.
78. S. W. May, L. M. Landgraff in N. N. Li (ed.): *Recent Developments in Separation Science,* **vol. 5,** CRC Press, West Palm Beach, FL, 1979, p. 227.
79. P.-O. Larsson, in Reference 16, p. 212.
80. C.-M. Yang, G. T. Tsao, in Reference 8, p. 1.
81. R. W. O'Brien, L. R. White, *J. Chem. Soc. Faraday Trans. 2* **72** (1978) 1607.
82. M. Bier, in Reference 5, p. 185.
83. M. Bier, in Reference 40, p. 547.
84. P. J. Haling, P. Dunnill, *Enzyme Microb. Technol.* **2** (1980) 2.
85. P. Dunnill, M. Lilly, *Biotechnol. Bioeng.* **16** (1974) 481 – 490.
86. C. S. Owen in T. G. Pretlow, T. P. Pretlow (eds.): *Cell Separation: Methods and Selected Applications,* **vol. 2,** Academic Press, New York 1983, p. 127.
87. D. F. Rudd, G. J. Powers, J. J. Siirola: *Process Synthesis,* Prentice-Hall, Englewood Cliffs, NJ, 1973, Chap. 5, p. 155.
88. H. L. Lee, E. N. Lightfoot, J. F. G. Reis, M. D. Waissbluth, in N. N. Li (ed.): *Recent Developments in Separation Science,* **vol. 3,** part A, CRC Press, Boca Raton, FL, 1977, p. 1.
89. J. C. Giddings, *Sep. Sci. Technol.* **13** (1978) 3.

Energy Management in Chemical Industry

COLIN D. GRANT, University of Strathclyde, Glasgow, United Kingdom

1. Introduction . . . 1
2. Areas and Levels for Energy Management . . . 3
2.1. Areas . . . 3
2.2. Levels . . . 3
3. Good Housekeeping . . . 3
4. Audits, Monitoring, and Control . . . 4
4.1. Energy Audits . . . 4
4.2. Monitoring and Control . . . 5
4.3. Energy Usage and Production Levels 5
4.4. Batch Processes . . . 6
4.5. Advanced Monitoring and Control Systems . . . 6
5. Heat Recovery . . . 6
5.1. Process Design for Heat Recovery . . . 7
5.2. Types of Heat Exchanger . . . 7
5.3. Equipment Design . . . 8
5.4. Example: Optimum Duty of a Waste-Heat Boiler . . . 8
6. Distillation . . . 10
6.1. Column Design . . . 11
6.2. Methods of Increasing Energy Efficiency . . . 11
6.3. Control and Operation of Columns . . . 11
6.4. Column Improvements–Internals . . . 12
6.5. Column Improvements–Control Systems . . . 13
6.6. New Column Designs . . . 13
7. Drying . . . 14
7.1. Energy Efficiency of Dryers . . . 14
7.2. Operation and Control of Convective Dryers . . . 14
7.3. Heat Recovery from Convective Dryers . . . 15
7.4. Microwave and Radio-Frequency Drying . . . 15
7.5. Hybrid Dryers for Batch Processes . . . 16
8. Process Design . . . 16
8.1. Less Energy-Intensive Process Conditions . . . 16
8.2. Improved Conversions . . . 17
9. References . . . 17

1. Introduction

Objectives of Energy Management. The chemical industry uses energy for *fuel* to provide primary heat and *power* for the operation of various processes. Energy management is concerned with the efficient use of this energy; its primary objective is to reduce operating costs, rather than simply to reduce energy consumption. The purpose of energy management is, therefore, to reduce the total energy requirement of a process per unit of output while reducing the total cost of production (or at least holding it constant). Energy usage for space heating and lighting is insignificant compared to process energy usage, except for some sectors of the fine-chemical industry where all operations are carried out indoors.

Many methods for improving the energy efficiency of processes require an expenditure of either capital investment or manpower or both. The return on such investment through reduced operating costs must be acceptable and worthwhile to the company, and investment in energy-saving measures will normally compete with other investment opportunities. The *payback period* is a simple criterion (although it has many limitations) widely used for a preliminary analysis of investment in energy-saving measures. Payback period is defined as:

$$\text{Payback period} = \frac{\text{Capital required}}{\text{Annual savings}}$$

Most of the methods discussed in this article for improving the energy efficiency of chemical processes are based on established technologies and are realistic in terms of being financially attractive and yielding an acceptable return on investment. Whether or not they should be implemented depends on a detailed technical and economic study of each potential application.

Energy for Feedstocks. The chemical industry uses substantial quantities of energy, mainly in the form of oil and gas, as feedstock for conversion to other products. The efficient chemical conversion of energy feedstocks to higher value products is of great importance, but it is normally considered separately from the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_12

efficient management of energy used to operate processes. Energy usage in the form of feedstocks is not considered in this article.

Energy Supply and Distribution to Processes. The energy used in processes, either as heat or as power, can be supplied and distributed in several ways:

Heat	Power
Primary	Electricity
Steam systems	Direct drives
Direct firing	(steam turbines)
Electric heating	
Secondary	
Heat recovery	

Energy-Using Operations. Every operation in a chemical process requires energy in some form or other, but for the purposes of energy management, the most important energy-using and energy-intensive operations must be identified. An approximate analysis of the main energy-using operations for the chemical industry as a whole was made by GRANT [1]. A more recent and detailed analysis for the chemical industry in the United Kingdom [2] has identified the main energy-using operations in each sector of the industry and for the industry as a whole; the results are given in Table 1. A wide range of processes was analyzed in each sector, and a similar breakdown of energy usage could be expected for the same sectors in the chemical industry worldwide.

For the industry as a whole, around 71 % of the energy is used to provide heat, and around 29 % to provide power. Heat usage is dominated by three operations: (1) heating of process streams, (2) distillation, and (3) drying. Methods for improving the energy efficiency of these operations are considered in this article.

The main power-using operations involve fluid flow through pumps, compressors, fans, and blowers. The energy efficiency of these operations can be improved with the use of more efficient equipment, better matching of equipment capacity to throughput, and more energy-efficient flow control such as variable-speed drives. These techniques are not considered in detail in this article. The methods are more applicable to large pumps, motors, and other items of equipment. In many plants, energy usage is so widely distributed that the actual use in each item of equipment is small, and in such cases, energy saving through improved equipment is difficult to achieve.

Significant power savings are possible if processes are improved to reduce the flow of process streams (for example, by reducing recycles) or to operate at reduced pressure. An example of the latter approach is given (see Section 8.2).

2. Areas and Levels for Energy Management

2.1. Areas

Energy management is involved in two main areas:

Table 1. Energy use in the chemical industry in the United Kingdom [2]

Operation	Energy use in each sector, %						
	Pharmaceuticals	Dyestuffs and Pigments	Plastics and rubber	Inorganic	Organic	Miscellaneous	Total, all sectors, %
Process heating	36	40	42	52	35	38	40
Evaporation	0	1	3	7	3	3	4
Distillation	15	4	20	2	23	5	13
Drying	3	14	16	17	4	14	10
Space heating	23	3	5	1	1	9	4
Fans and blowers	2	2	3	3	1	2	2
Drive motors	1	3	2	4	1	2	2
Refrigeration	6	3	2	2	8	2	5
Compression	6	4	3	3	16	18	10
Mixing	6	5	2	1	1	2	2
Pumping	2	3	2	7	7	4	6
Comminution	0	18	0	1	0	1	2
Total	*100*	*100*	*100*	*100*	*100*	*100*	*100*

1) energy production and distribution, and
2) energy usage within processes.

The area of energy production and distribution deals with the operation of boilers, furnaces, and electrical generators; with compressed air and nitrogen generation; and with the distribution of steam, electricity, and other services. These matters are important, and good energy management should ensure that services are generated and distributed as efficiently as possible. The methods for this are well-known and are not discussed here.

Once the maximum practical efficiency has been achieved in the supply and distribution of services, the only remaining area for energy saving involves reducing the energy requirements of the process. The greatest potential for more efficient use of energy in the chemical industry will almost always lie in decreasing the process energy requirements and usage, rather than in supplying energy more efficiently to processes. Energy usage within processes should, therefore, be the most important consideration for energy management.

2.2. Levels

Energy management is involved at three main levels:

1) Efficient operation of the existing plant (*good housekeeping* measures)
2) Major improvements to the existing plant (*retrofits* or *revamps*)
3) New plant or process designs

Level 1 is primarily the concern of plant management, and the energy management measures involved in operating plants efficiently are often referred to as good housekeeping measures. Level 2 involves plant management as well as process and plant design. Level 3 involves principally research and development, together with process and plant design. The potential for substantial improvement in energy efficiency increases in going from Level 1 to Level 3, but the necessary capital costs and the time scale to achieve the improvements also increase.

3. Good Housekeeping

Good housekeeping measures cover a wide range of relatively minor measures which generally represent good plant management and operating practice. The difference between good and bad housekeeping typically represents around 10 % of the total process energy usage, but it might be more in extreme cases. A checklist of the most important measures follows:

1) *Steam Systems*
   Reduce leaks
   Improve operation of steam traps
   Increase condensate recovery
   Increase flash steam recovery
   Use lower steam pressure if possible
   Use direct steam for heating if possible
   Improve maintenance of steam ejectors
2) *Insulation*
   Check optimum thickness; increase if necessary
   Ensure flanges and manholes are lagged
   Improve protection from rain and water
   Replace after maintenance
3) *Heat Exchangers*
   Monitor fouling; reduce if possible
   Increase cleaning frequencies
4) *Pumping and Compression*
   Avoid oversizing pumps
   Consider replacing undersize lines and valves
   Consider if flows can be reduced
   Switch off when not required
5) *Electrical Systems*
   Improve power factor; minimize other losses
6) *Compressed Air Systems*
   Use minimum operating pressure
   Use minimum inlet air temperature
   Reduce pressure drops in pipework
   Reduce leaks
7) *Cooling Water Systems*
   Minimize flow
   Consider scope for "cascade" flow through units in series
   Switch off when not required
8) *Vacuum Systems*
   Reduce leaks
   Replace steam ejectors by liquid ring pumps

Although these measures may require relatively little capital expenditure, they can involve significant investments of time and manpower. Most of the measures are not "once-and-for-all," in the

sense, for example, that installing a new heat exchanger for heat recovery is a once-and-for-all measure. Good housekeeping measures, therefore, require regular attention if energy savings are not to be eroded.

4. Audits, Monitoring, and Control

The measurement and control of energy use within processes are interrelated, essential aspects of good energy management.

4.1. Energy Audits

An energy audit is an analysis of energy usage which identifies the main areas, operations, and types of energy used in a process. It should be based on actual plant measurements and data, rather than on whatever process design data might be available. Plants are frequently modified so that the operation is very different from the original design. Also, design data may be inaccurate or optimistic with regard to energy requirements.

In many cases, an audit may involve primarily obtaining accurate heat and mass balances for the process, and the major problems that arise are usually concerned with inadequate instrumentation, particularly for steam flow. For example, electrical power consumption is often better metered and easier to quantify than heat consumption.

An energy audit, normally based on a single set of typical operating conditions, is the essential first step in identifying methods of improving the energy efficiency of a process. The results of the audit should, therefore, be analyzed carefully in order to:

1) develop a better understanding of the process,
2) establish a measure of energy efficiency for the process, and
3) establish targets for improving energy efficiency.

If the results of an energy audit are in close agreement with the process design data, they may provide little further understanding of the process. However, if substantial differences exist, the reasons should be investigated because they will often indicate possible methods of improving energy efficiency.

Figure 1 shows two approaches to establishing targets for improving energy efficiency. In the first, the *top-down approach,* targets are often set in an arbitrary manner based on current energy use. For example, a target might be set as a 10 % reduction in energy usage within two years. This is an easy way to set targets, but it provides little insight into the real potential for energy savings or the manner in which the targeted reduction might be achieved.

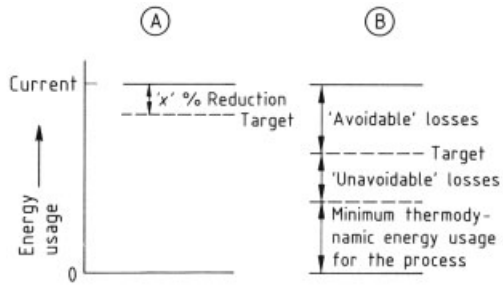

Figure 1. Targets for energy savings
A) Top-down approach; B) Bottom-up approach

The second method, the *bottom-up approach,* is to try to identify the minimum practicable energy requirement for the process. This provides an absolute target for improving energy efficiency. This approach requires a thermodynamic analysis to identify first the minimum theoretical energy required for the process route (which may be negative for an exothermic reaction). Then the unavoidable additional energy required to operate a real process with the best available technology must be identified. This may require a mixture of thermodynamics and process engineering techniques.

The latter, more rigorous approach is not yet widely applied to complete chemical processes. However, the rigorous thermodynamic approach, in the form of process integration or *pinch technology* [3], is now being used with considerable success for identifying targets in heat recovery systems (→ Pinch Technology).

4.2. Monitoring and Control

Monitoring refers to the regular, systematic measurement of energy use in relation to production

(rather than the one-shot analysis of usage in an energy audit). A good monitoring and control system requires:

1) good instrumentation,
2) a short time period between measurements (otherwise excessive energy use continues for too long before remedial action is taken),
3) clear norms and targets with which to compare the measured energy usage,
4) knowledge of control action to be taken if energy use exceeds the norm, and
5) control systems to implement the action.

4.3. Energy Usage and Production Levels

In establishing norms and targets with which to compare the measured energy usage, problems may arise in quantifying the effects of (1) variations in product mix, (2) variations in weather, and (3) variations in throughput. Of these, the effect of variations in throughput is often the most important.

A typical relationship between energy consumption and production level is shown in Figure 2. Many chemical processes have relationships of this form. In many cases, plants cannot operate over a wide range of throughputs–particulary single-stream continuous processes–and operation at very low production levels is not possible.

The data can show considerable scatter, and the temptation often arises to fit a regression straight line to all the data points to establish a norm. This should be avoided if possible; a detailed analysis of the data may reveal a more complex and accurate relationship between energy use and production levels.

With relationships between energy usage and production of the form shown in Figure 2, both *fixed* and *variable,* or *output-related, components* exist. This can be seen clearly by extrapolating the energy use to zero production. The fixed component can range from 30 % of the total energy consumption at full output to as to as much as 60 % for some batch processes.

With energy consumption vs. output relationships of this form, the energy consumption per unit of output–the normal measure of energy efficiency–decreases as output increases, as shown in Figure 3. Operation of the plant at high output levels, therefore, usually represents an important method of improving energy efficiency. If required production levels are significantly lower than capacity, then a variety of operating strategies, including campaign operation, should be considered [4].

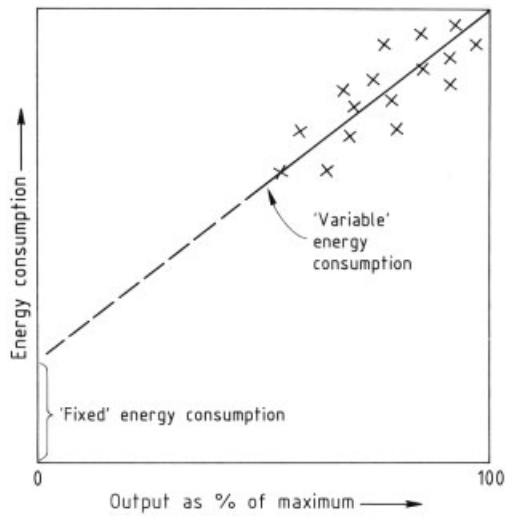

Figure 2. Energy consumption versus output for a typical process

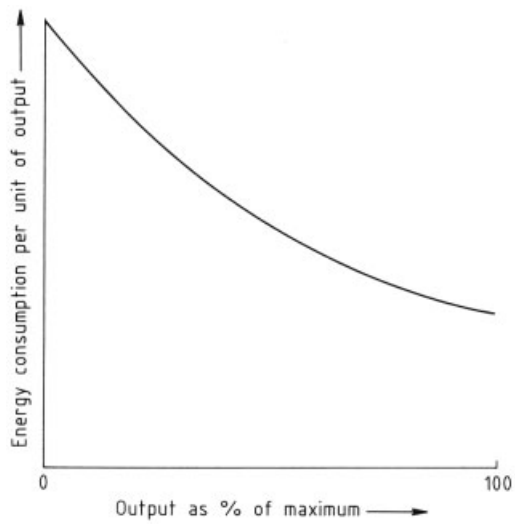

Figure 3. Energy consumption per unit of output versus output for a typical process

4.4. Batch Processes

Increasing the energy efficiency of batchwise processes through improved technology or process modifications can often be more difficult than for continuous processes. As a simple example, the potential for heat recovery is often more limited in batch processes. The control and operation of batchwise processes is, therefore, particularly important for good energy management. Improved sequencing and control of operations, which reduce batch times and increase throughputs, usually significantly reduce the energy consumption per unit of output [5].

4.5. Advanced Monitoring and Control Systems

The monitoring methods discussed above can be applied by using conventional and existing control systems. The use of modern computer-based monitoring and control systems, which are designed to operate on a plant-wide basis, can yield further major improvements in energy efficiency [6–8]. These systems are designed to supervise the individual programmable controllers that operate the control loops on the plant. The system will extract data from the programmable controller networks and then output set points and controller configuration information. Systems are available which are independent of programmable controller manufacture and can link up with most types of instruments and computers.

The data collected can be manipulated and presented in a form useful to plant managers and operators. For example, the data could be production rates, or energy use and waste. These can be projected on a real or historic time basis, and displayed to show how a plant is performing against targets. Data can be presented in the form of bar charts, graphs, and tables or as values superimposed on a mimic diagram.

The system can be used in a passive role as a monitoring and management tool to assist skilled plant operation by bringing together groups of readings, calculating totals and averages, and indicating trends and optimum operating conditions. The system can also be used as a supervisory control system, for example, by changing set points on programmable controllers.

5. Heat Recovery

Heat recovery is the single most widely used technology for improving energy efficiency in chemical industries. This is to be expected because process heating is by far the most significant energy-consuming operation, accounting for 40 % of the energy used by the industry as a whole and as much as 52 % in the inorganic chemical sector (Table 1). In addition, evaporation, distillation, and drying together account for another 27 % of the energy used by the industry as a whole, and heat recovery is an established method of improving the energy efficiency of these operations (Chaps. 6 and 7).

Sometimes, in attempting to consider heat recovery in a process, the potential for reducing the heat requirements of the process is ignored. Improving a process to use less heat is almost always better than modifying it to recover more heat, and the possibility of reducing heat requirements should always be considered first.

Two interrelated aspects of applying heat recovery to chemical processes are (1) process design and (2) equipment design and selection.

Process design involves determining the process streams between which heat is to be transferred and the amount of heat that is to be transferred; this is also called the "duty" of the heat exchanger. *Equipment design* involves the design or selection of the best heat exchanger for the duty. The criteria for the best exchanger involve capital costs, installation costs, and considerations of all aspects of operability.

The basic design equation for any heat exchanger is an overall heat balance of the form

$$\begin{aligned} Q &= \text{rate of heat transfer from hot stream} \\ &= \text{rate of heat transfer to cold stream} \\ &= UA\Delta T_{\mathrm{m}} \end{aligned} \tag{1}$$

where Q is duty (W), U is the overall heat-transfer coefficient (W m^{-2} K^{-1}), A is the heat transfer area (m^2), ΔT_{m} is the mean temperature difference between the fluids (K).

5.1. Process Design for Heat Recovery

The first stage in the design of a heat recovery system is to identify potential heat sources and heat sinks. Heat sources can be either process streams (e.g., the outlet stream from a reactor) or waste streams (e.g., a flue-gas stream). Heat sinks can be process streams (e.g., a reactor feed) or other streams such as combustion air or boiler feedwater.

The stream data necessary to specify heat sources or heat sinks include the following:

1) stream phase (vapor, liquid, mixture),
2) mass flow rate,
3) temperature,
4) pressure,
5) contaminants and fouling characteristics,
6) corrosion and hazardous characteristics, and
7) variations in flowrate and temperature.

The last parameter is important for continuous processes with significant variations in throughput and for any attempt at heat recovery in batch or semibatch processes.

The next stage in process design is to consider matching heat sources and heat sinks, while taking account of the above parameters and the physical proximity of the streams. (Installed costs of heat exchangers can be many times greater than the purchase costs of the unit, and long piping runs between units are undesirable.)

If a number of hot and cold streams exist between which heat exchange is possible, then the resulting heat-exchanger network can be designed in many different ways. In such cases, the techniques of *process integration* (or *pinch technology*) should be applied to achieve the optimum heat recovery network design, by taking into account the balance between energy savings and capital costs, as well as operability considerations for the network (→ Pinch Technology) [3].

Even for a heat recovery system with only a single heat source and sink, consideration must be given to the optimum economic duty for the exchanger. A preliminary economic analysis should, therefore, be carried out before any detailed design calculations are made.

5.2. Types of Heat Exchanger

When a feasible match between hot and cold streams has been identified, the most appropriate type of heat exchanger for the required duty must be selected. The main types of exchanger are listed in the following according to type of construction (→ Heat Exchange). *Tubular heat exchangers:*

Double pipe∗
Shell-and-tube∗

Plate heat exchangers:

Gasketed$^+$
Spiral$^+$

Extended surface heat exchangers:

Plate – fin$^+$
Tube – fin∗

Regenerative heat exchangers:

Rotary$^+$
Fixed matrix∗

A comprehensive guide to heat-exchanger selection and costs can be found in [3]. The types of heat transfer that can occur in heat recovery and a guide to the most appropriate types of exchanger are given in Table 2.

Table 2. Exchanger selection for types of heat transfer

Type of heat transfer	Type of heat exchanger
Liquid – liquid	tubular, plate
Gas – liquid	extended surface
Gas – boiling liquid	shell-and-tube (waste-heat boiler)
Gas – gas	tubular, plate, regenerative

5.3. Equipment Design

Once a heat recovery duty has been specified and the type of exchanger selected, two interrelated aspects of the detailed equipment design must be considered: (1) thermal design and (2) mechanical design.

Some types of exchanger are normally designed by specialist manufacturers for a duty specified by the purchaser (marked by $^+$ in preceeding list); other types are more normally designed by the user to standard methods and codes (marked by ∗ in preceeding list).

Thermal design involves calculating mean temperature differences and heat-transfer coefficients, and hence the required heat-transfer area. The pressure drop for each stream is usually also calculated at this stage.

Mechanical design involves specifying the required heat-transfer area in terms of equipment details, at least to the extent of providing sufficient detail for exchanger fabrication to appropriate codes and standards.

For example, for a shell-and-tube exchanger, this would involve specifying the diameter, length, layout, and number of tubes; the diameter of the shell; the type and spacing of the shell-side baffles; etc.

The two aspects of thermal and mechanical design are interrelated. Heat-transfer coefficients (or pressure drops) cannot be calculated without a knowledge of the basic mechanical features that affect the geometry of the heat transfer area and the velocity of the fluids. An iterative design procedure is, therefore, necessary.

A preliminary mechanical design is established, based on assumed values of the overall heat-transfer coefficient. The thermal design of this exchanger is then checked by detailed calculation. If the thermal design is not suitable, the mechanical design is amended and the procedure repeated. This is known as a *rating procedure* for heat exchanger design. Comprehensive design procedures can be found in [9].

5.4. Example: Optimum Duty of a Waste-Heat Boiler

Process Description. An effluent gas stream from a reactor at a temperature of 550 °C can be used to generate steam at 0.5 MPa (5 bar) in a waste-heat boiler. The waste-heat boiler can be a vertical shell-and-tube exchanger with the hot gases on the tube side and the steam on the shell side. The following data pertain:

Gas flow rate 10 t/h
Gas specific heat 1.63 kJ kg^{-1} K^{-1}
Estimated overall heat-transfer coefficient in the waste-heat boiler 0.12 kW m^{-2} K^{-1}
For steam at 0.5 MPa (5 bar) raised from boiler feed water at 15 °C: saturation temperature, T_{sat} = 150 °C and increase in enthalpy h_{fg} = 2686 kJ/kg
Operating period 4000 h/a

Cost. The value of the steam raised in the waste-heat boiler is ca. 26 $/t. Approximate purchased costs of shell-and-tube exchangers are available, e.g., in [3]. The installed cost of the exchanger is approximately twice the purchased cost.

Method of Analysis. The optimum economic duty depends on the value of the steam raised in relation to the installed cost of the exchanger. Both of these depend on the gas outlet temperature T. The payback period can be used as a first guide for evaluating the investment.

A range of values of T is chosen, starting with a "base case" in which a short payback period might be expected. Successively larger exchangers are then priced, and the payback period for each incremental investment is calculated until the return on the additional investment is unattractive.

Calculations of Heat-Transfer Areas. The temperature profile in the waste-heat boiler is shown in Figure 4 (for explanation of symbols, see Table 3). From Equation (1):

Table 3. Heat transfer areas for waste-heat boiler*

T, °C	Q, kW	M, kg/s	A, m^2
200	1590	0.589	79
180	1680	0.623	98
170	1720	0.640	113
160	1770	0.657	139

*
T = gas outlet temperature;
Q = duty (rate of heat transfer);
M = rate of steam generation;
A = heat-transfer area.

$$Q = 4.53\,(550 - T)$$
$$= 2.69 \times 10^3 M = 0.12 A \Delta T_{\mathrm{m}}$$

The values M and A have been calculated for a range of T, as shown in Table 3.

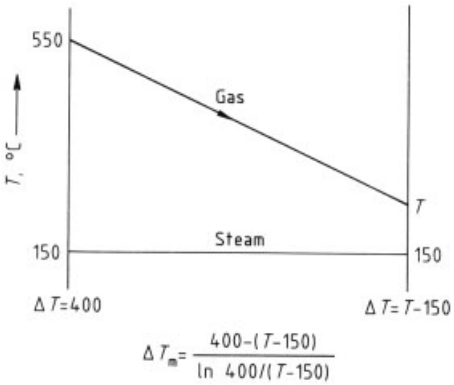

Figure 4. Temperature profile in a waste-heat boiler

Calculations of Costs. Based on one year's operation (4000 h), the value of the steam generated is ca. $3.81 \times 10^5 \cdot M$ (in \$/a).

The value of steam generated and the installed cost of the exchanger have been calculated for all four values of T. The payback period for the base-case exchanger, area 79 m^2, was calculated first and then the payback period for the incremental investment in the next three sizes, as shown in Table 4.

Comments. Preliminary analysis showed that heat recovery with a waste-heat boiler is economically attractive. The base-case exchanger has a payback period of only 0.3 year. If, for example, the company criterion for investment in energy saving measures required a payback period of about two years (a fairly common requirement), then the optimum economic gas outlet temperature would be between 170 and 160 °C. Detailed design calculations would be concentrated on conditions in this region.

6. Distillation

Distillation is the most widely used operation in the chemical industry for the separation of mixtures of chemical species (→ Distillation and Rectification). It accounts for around 13 % of the energy usage of the industry as whole, and up to 23 % of the energy usage in some sectors such as organic chemicals (Table 1). A summary of 130 of the most important separations in the chemical industry which are carried out by distillation is given in [10].

The basic process of distillation is the equilibrium flash where a mixture of components is separated into a vapor phase and a liquid phase, with the components partitioned selectively between them. The exact equilibrium compositions of the phases are determined by temperature, pressure, initial composition, and relative volatility of the components. Because the equilibrium concentrations for a single flash are not usually pure enough, many successive equilibria are required to achieve good separation. Thus, distillation is carried out in columns having multiple vapor – liquid stages in succession.

A typical distillation column is shown in Figure 5. Feed enters at some intermediate position. Vapor rising up the column comes in contact with liquid falling down the column. At the bottom, a portion of the liquid is vaporized in a reboiler. Vapor leaving the top of the column is condensed, and a portion is returned to the column as reflux. A temperature gradient is created in the column by the addition of heat in the reboiler and the removal of heat in the condenser – reflux system. The concentration of the components tends to follow the temperature profile, with the lighter components concentrating in the upper (cooler) section of the column

Table 4. Costs and payback period for waste-heat boiler

Exchanger area, m^2	Exchanger cost, 10^3 \$	Value of steam, 10^3 \$	Payback period, a	
79	66.2	224.5	66.2/224.5	= 0.29
98	77.7	237.5	11.5/13.0	= 0.88
113	85.2	243.9	7.5/6.4	= 1.17
139	99.4	250.4	14.2/6.5	= 2.18

and the heavier components concentrating in the lower (hotter) section.

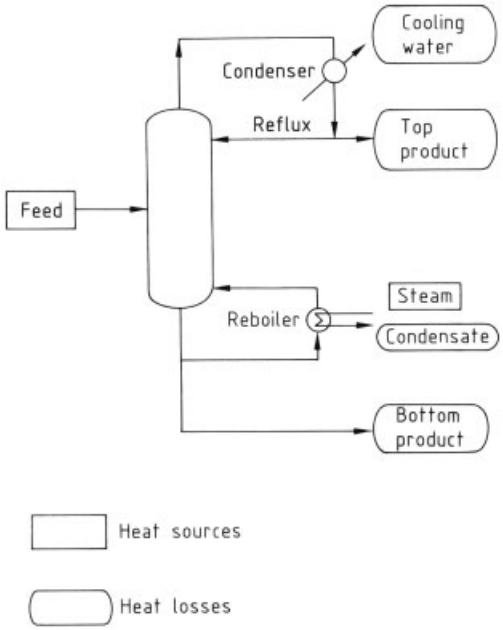

Figure 5. Heat flow in a simple distillation column

The column is filled with either trays or packing material to promote intimate vapor – liquid contact. The effectiveness of the stage contacting equipment is measured by tray efficiency or by the *height of a transfer unit* for packed columns. These can be regarded as the fractional approach to equilibrium that occurs in any one physical tray or packing length.

Distillation is a heat-driven separation process in which the heat input at the reboiler (often steam-heated) flows up the column and is rejected at a lower temperature in the overhead condenser. Methods of improving the energy efficiency of distillation are almost always directed toward reducing the reboiler heat demand, because this represents the major energy requirement. Electrical energy is also necessary for pumping cooling water through condensers and for reflux pumps, but this is usually of minor significance compared with the heat requirements.

6.1. Column Design

For a given separation, the reflux ratio and hence the required heat input decrease as the number of theoretical plates (equilibrium stages) in a column increases. The bounding limits are the minimum number of theoretical plates at total reflux and the minimum reflux with an infinite number of theoretical plates, as shown in Figure 6. Column design represents an optimum balance between the relative cost of energy (operating costs) and the cost of the column and internals (capital cost). Current practice is to design for reflux ratios between 10 and 20 % above the minimum.

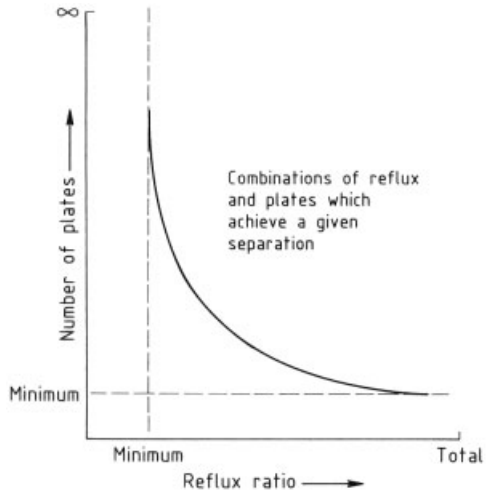

Figure 6. Relationship between reflux ratio and number of plates

Regardless of the design reflux ratio, columns are frequently operated at much higher reflux ratios, with a corresponding increase in energy consumption. Control of the reflux ratio represents a major operating parameter.

6.2. Methods of Increasing Energy Efficiency

The methods that can be used to increase the energy efficiency of distillation may be divided into three levels:

1) Better operation and control of existing columns (requiring relatively low costs)
2) Major improvements to existing columns (e.g., new internals, advanced control systems)
3) New designs (e.g., split-tower (multi-effect) distillation)

The first two levels are generally applicable to the existing plant, while the third level is usually worth considering only for new plants.

In addition to these methods which are specific to the distillation column itself, heat recovery from the associated hot streams is an important energy-efficient measure. For example, preheating feed streams with both product streams is common practice.

To evaluate and analyze methods for energy efficiency in distillation, detailed calculations of column operation, including stage-to-stage calculations, are usually necessary (→ Distillation and Rectification, Chap. 7.). Computer programs are now available for the design and simulation of distillation columns, and these greatly facilitate the detailed calculations.

6.3. Control and Operation of Columns

The main variable controlling the energy consumption on an existing column is the *reflux ratio.* A number of other variables can be manipulated to allow a reduction in reflux:

product specification
feed plate location
operating pressure

Product specification should be examined critically. The column product specification is often set slightly higher than the ultimate specification for sale or further processing. This minimizes production of off-specification material during normal operating conditions, at the expense of increased energy usage.

The correct feed-plate location should be used. The feed should enter the column at the point where the composition on the tray closely matches the composition of the feed. If it enters anywhere else, efficiency is lost through blending. Many columns are designed with a number of optimal feed points, and if necessary the feed point can easily be relocated.

For many mixtures, separation is easier at lower pressure because the relative volatility increases with decreasing pressure. Lower pressure operation can, therefore, be used to achieve the same separation at lower reflux ratios. The operating pressure of most columns is set to allow condensation of the overhead vapors with cooling water (or air). Because this limits overhead temperatures to a minimum of ca. 35–40 °C, relatively high pressures may be necessary to condense some of the more volatile components. Opportunities may exist for reducing column pressure when cooling capacity is high due to low ambient temperature. Advanced control systems for such pressure reduction are available, but action can also be taken by alert process operators.

6.4. Column Improvements–Internals

New columns are designed with reflux ratios close to the minimum, but many columns still in use were originally designed for higher reflux ratios. In such cases, energy consumption can be reduced by revamping the column to increase the number of stages. Table 5 is a very rough guideline to the economic feasibility of this approach.

Table 5. Economic feasibility of column improvement

Ratio of actual to minimum reflux	Potential cost savings
> 1.3	good
1.3 to 1.1	possible, worth studying
< 1.1	doubtful

The performance of a distillation column depends on the actual number of stages and the stage efficiency (or the total height of packing and the height of the transfer unit for packed columns). Therefore, for stage columns the number of theoretical stages can be increased by (1) increasing the number of trays by increasing the column height or reducing the tray spacing or (2) replacing existing trays with a more efficient type.

Increasing column height can be very expensive. Reducing tray spacing may be expensive because of the cost of modifying tray supports, and problems due to entrainment may result. In many cases, retraying a column with more efficient contacting devices is the most attractive approach [11]. For packed columns, changing to a packing with improved liquid distribution may be attractive.

Generalizations about tray efficiency (or height of a transfer unit for packing) are difficult

to make because of the dependence on the specific application. A preliminary analysis can be made of the energy savings available at various levels of increased tray efficiency. If the results look attractive, tray and packing suppliers can be approached to determine the maximum feasible increase in efficiency, and a detailed economic analysis is then possible.

6.5. Column Improvements–Control Systems

The control system for a distillation column has two basic functions: to stabilize and to optimize the operation. When applied properly, the control system should ensure the production of products to specification at the minimum cost.

Conventional distillation control systems are designed only for stabilization, relying on operator intervention for any optimization. A conventional system for a two-component fractionator is shown in Figure 7. Product flows are level-controlled. The reflux is set independently at a fixed rate, which is usually conservatively high. Product quality is controlled inferentially by varying the reboiler heat input to hold a constant temperature on some critical tray. This type of system is susceptible to variations in heat balance, which lead to changes in product quality.

Advanced control systems are available that use direct material-balance control, feed-forward control, on-stream product analyzers, and digital computers for optimization. Such systems can give more stable operation, control the column at the optimum operating point, and reduce the need for operator action. They offer considerable potential for reducing energy requirements and operating costs [8, 10].

6.6. New Column Designs

In conventional distillation, each column has its own reboiler and its own condenser, normally water- or air-cooled. Heat supplied to the reboiler is rejected without further use in the condenser. Many methods of making better use of these heat flows are possible, involving new approaches to column design; these include:

heat pumping
split-tower operation

A *heat pump system* can upgrade the heat content of the overhead vapors so that they may be used in the reboiler of the same column. A system based on vapor recompression, using the overhead vapors as the heat pump working fluid, is shown in Figure 8. Although heat pumping can produce significant energy savings, only a limited range of applications offer any prospect of

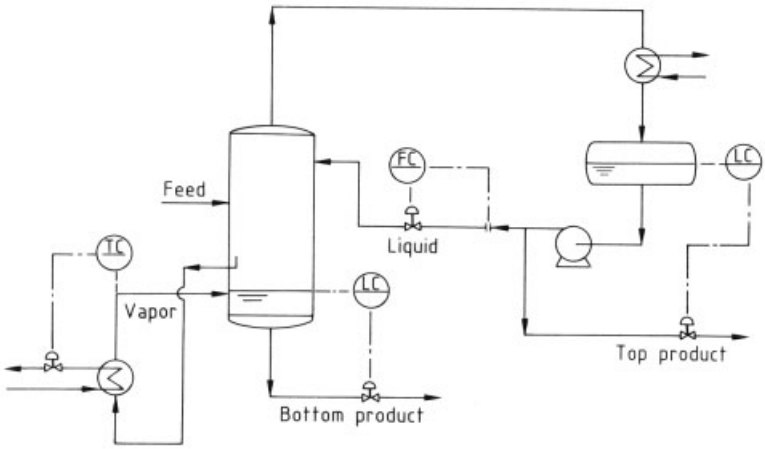

Figure 7. Conventional control system for a distillation column
TC = Temperature control;
LC = Level control;
FC = Flow control

economic viability. The major criterion is the temperature range in the column, and potential applications are restricted to relatively close-boiling components.

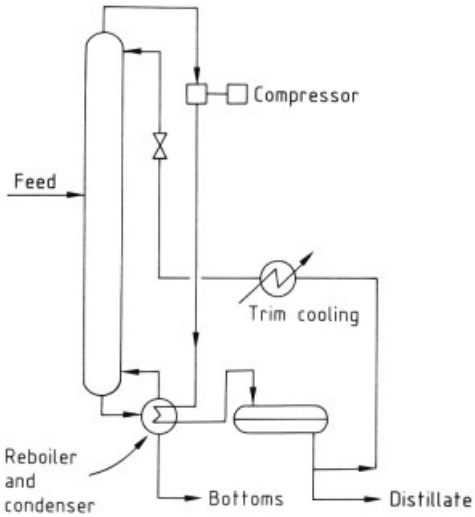

Figure 8. Distillation with vapor recompression

The basic principle of *split-tower distillation* is to increase the thermodynamic efficiency by repeated use of a given quantity of energy (analogous to the repeated use of heat in multi-effect evaporation). One possible application is shown in Figure 9. A series of N towers is operated in parallel with approximately equal feed rates, but each at a different (decreasing) pressure. All the heat is applied to the first column, and its overhead vapors are condensed by reboiling the next column, and so on. The energy requirement for the series is theoretically about $1/N$ times the single-column energy requirement. Although this method offers substantial energy savings, it has many practical constraints and has not yet been applied widely.

7. Drying

Thermal drying of solids accounts for around 10 % of the energy usage in the chemical industry as a whole, and up to 17 % of the energy used in sectors such as inorganics, dyes and pigments, and polymers (Table 1). For further information, see → Drying of Solid Materials.

The three main classes of dryers, depending on the method of heat transfer are

1) *Direct contact* (*convective*) *dryers:* heat transfer by direct contact of the wet solid and a hot gas stream
2) *Indirect contact* (*conduction*) *dryers:* heat transfer to the wet solid through a solid wall

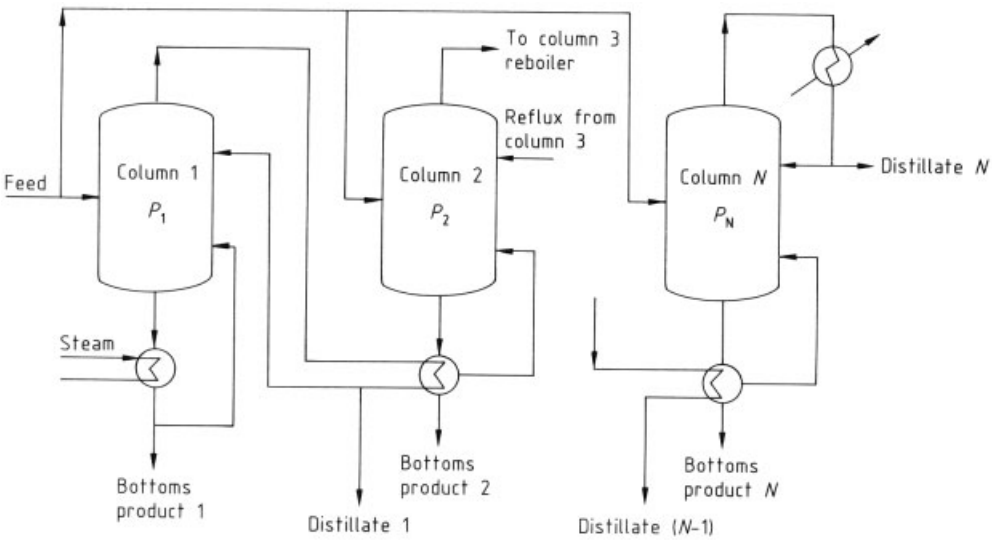

Figure 9. Multi-effect distillation

3) *Microwave or radio-frequency dryers:* heat transfer throughout the wet solid by the action of a high-frequency electric field.

Convective dryers are the most widely used type in the chemical industry. The most common application is the evaporation of water vapor, although dryers that evaporate solvents are used in some pharmaceutical and fine-chemical processes. Dryers evaporating organic solvents are less significant energy users because of the lower latent heat of vaporization of organic solvents compared to water.

7.1. Energy Efficiency of Dryers

The thermal efficiency of dryers is usually defined as the ratio of the heat required for vaporization of the water removed to the heat actually used. For indirect-contact dryers, the efficiency can be close to 100 %, but the efficiency of many convective dryers may be as low as 30 %. With this definition, an efficiency of 100 % does not represent the minimum energy use because efficiencies greater than 100 % are theoretically possible with heat recovery from the exhaust gases.

The primary methods of improving the thermal efficiency of convective dryers are

1) improved operation and control
2) heat recovery from exhaust gases

7.2. Operation and Control of Convective Dryers

The flow rate and inlet temperature of the hot gas stream (normally air) are the main parameters in dryer operation. The heat required for a given duty can be reduced by controlling these parameters, i.e., reducing the air flow rate and increasing the inlet temperature, to minimize the exhaust gas enthalpy per unit mass of water vapor [12, 13].

Dryers that operate at a range of throughputs should have correct conditions throughout the range. When running at less than full capacity, many dryers are operated at excessive air flow rates. Spray dryers, fluidized-bed dryers, and rotary dryers may offer little scope for altering the air flow rate without affecting particle transport characteristics within the dryer, and the air inlet temperature may be the only variable that can be controlled.

The minimum level of instrumentation and control for convective dryers should involve temperature measurement of inlet and outlet air, as well as the use of dampers for flow control. A better approach to optimum operation can be achieved with on-line measurement of exhaust air humidity, and hygrometers are now available which can operate up to 100 °C in dust-free gases.

Further improvements in operation, including the avoidance of overdrying, may be possible with direct on-line measurement of product moisture content. Several techniques are available for this, such as microwave absorption [14], but reliable methods are not yet possible for all applications.

7.3. Heat Recovery from Convective Dryers

Recirculation of a fraction of the exhaust gas stream to the inlet air for the heater represents a form of heat recovery that can yield significant energy savings. While this is a simple concept, the increased moisture content of the hot air entering the dryer means that a higher temperature is necessary to achieve the same drying rate. The capital costs of the ducting and control for air recirculation may be high in some cases.

Heat recovery from the exhaust gas stream by means of heat exchangers can result in substantial energy savings. Preheating the inlet air is the most common use of recovered heat, although in principle the heat can be used for other process streams.

Low outlet gas temperatures and heat-exchanger fouling are practical limitations (apart from the cost of the heat exchanger and associated ductwork) which may restrict heat recovery via heat exchangers.

Fouling of heat-exchanger surfaces by particles in the gas stream can often be a major restriction to heat recovery. For example, exhausts from spray, fluidized bed, and rotary dryers may contain too much dust for heat recovery if a cyclone is the only method of particle removal from the gas stream. More efficient gas

cleaning (e.g., by filtration) may be uneconomical if required only for heat recovery, but may be justifiable for high-value products if significant additional product recovery occurs.

The conventional approach to heat recovery from dryer exhaust gases is aimed primarily at recovering sensible heat, i.e., the temperature of the hot gas stream is decreased without condensing any vapor. The potential for heat recovery is greatly increased if the latent heat of the vapor can also be recovered by cooling the exhaust gas below its dew point. This can be achieved in heat pump systems where the recovered latent heat is upgraded to higher temperatures and used to preheat the inlet air. Although the application of heat pumps to drying is the subject of much research and development work, as yet no evidence exists of proven applications in the chemical industry.

7.4. Microwave and Radio-Frequency Drying

When a product containing water is treated with microwave or radio-frequency electromagnetic radiation, the energy is absorbed mainly by the water molecules. Drying is, therefore, possible without the use of large quantities of heated air, which suggests that energy savings might be possible in many applications. However, power for microwave and radio-frequency drying can be generated at an efficiency of only ca. 17 % on a primary fuel basis [14], so that applications with net energy savings are very limited. The capital cost of microwave or radio-frequency drying is high and has not been justified so far on the basis of energy savings alone. Factors that justify microwave or radio-frequency drying are more likely to involve improved product quality, ease of operation, and increased throughput on existing convective dryers.

7.5. Hybrid Dryers for Batch Processes

The drying of pharmaceuticals and fine chemicals is affected by some special factors, compared to the drying of bulk chemicals. These may include the following:

1) High product value
2) Low throughput
3) Difficulty of handling wet sticky material
4) Moisture to be evaporated includes solvents
5) Use of plant to process different materials

For these duties, the trend is toward the use of hybrid solids-processing equipment, in which more than one operation apart from drying can be performed [15]. Examples are combined filter – dryers and equipment in which crystallization, filtration, washing, and drying operations can all be carried out.

The aim of hybrid equipment is to reduce material losses (e.g., in dryer exhausts), handling problems, and hazards, and to increase throughput and flexibility of plant use. Improved energy efficiency is also a factor, but it is not the only consideration. However, even if energy efficiency is not a primary consideration, changes to equipment which increase the throughput of batch processes usually also increase the energy efficiency of the process as a whole (discussed in Section 4.4).

8. Process Design

For many chemical processes, changes in overall process design or operating conditions offer the greatest potential for reducing energy requirements. Even in "mature" processes, which are well-established and have undergone extensive development, substantial improvements in energy efficiency may still be possible. A good example of this is the ammonia process, which by 1960 was regarded as a mature process. However, since then, considerable efforts in process development have substantially reduced the energy requirements from ca. 70 GJ per ton of ammonia to 30 GJ per ton at present (Fig. 10, see next page); see also → Ammonia, Chap. 4.6. The energy requirements are now beginning to approach the thermodynamic minimum for this process route, which is an absolute limitation on further energy savings. Most other chemical processes still have energy requirements that are far in excess of the theoretical minimum.

Processes can be improved in many ways, and generalizations about how to achieve this are not very meaningful. Each process must be considered in detail to identify potential improvements. However, the following examples are given to

show the improvements that can be achieved through (1) less energy-intensive process conditions and (2) improved conversions in the reactor.

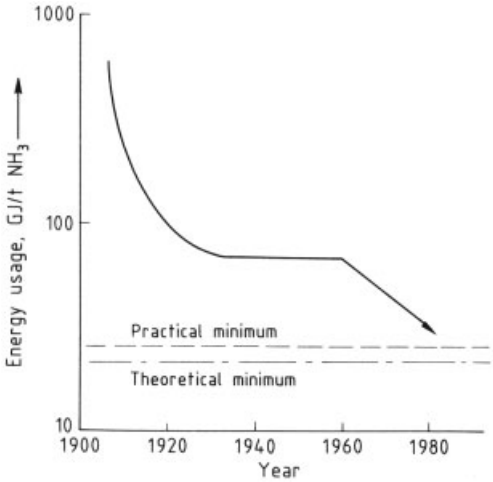

Figure 10. Energy requirements for the ammonia process

8.1. Less Energy-Intensive Process Conditions

If processes can be operated at less extreme conditions of temperature or pressure, then energy requirements may be substantially reduced. An example of this is the production of low-density polyethylene. For many years, standard processes involved polymerization at pressures in the range of 100 – 300 MPa (1000 – 3000 bar), with very high power requirements for compression. Typical total energy requirements for the process were around 15 MJ per kilogram of polymer [16]. Modern processes now use low-pressure gas-phase polymerization at about 0.7 – 2 MPa (7 – 20 bar). The total energy requirement for these processes is ca. 3 MJ per kilogram of polymer (→ Polyolefins) [17].

8.2. Improved Conversions

In many processes, the major energy requirement is for separation. Separation also frequently dominates the capital costs of the process plant. Increased conversions in the reactor, for example, by better and more selective catalysts, may substantially reduce energy requirements for downstream separations.

An example of this is the production of styrene monomer from ethylene and benzene. The reaction has two stages: (1) alkylation to give ethylbenzene, followed by (2) dehydrogenation to styrene. The dehydrogenation does not go to completion, and the products from the reactor are condensed to give a mixture of ethylbenzene, styrene, and other hydrocarbons. Separation of this mixture by distillation–in particular, the styrene – ethylbenzene column–requires most of the energy for the entire process. Improvements in operating conditions and catalysts have increased the conversion to styrene per pass from around 40 % to over 75 %. The effect of this on the styrene – ethylbenzene separation is to reduce column steam requirement from ca. 1.8 kg of steam per kilogram of styrene produced to less than 0.5 kg of steam per kilogram of styrene (→ Styrene) [18].

9. References

1. C. D. Grant: *Energy Conservation in the Chemical and Process Industries,* IChemE/George Godwin Ltd., Rugby 1979.
2. Energy Management Focus, no. 9, Focus on the Chemicals Industry, Dept. of Energy, London 1986.
3. B. Linnhoff et al.: *User Guide on Process Integration for the Efficient Use of Energy,* IChemE, Rugby 1982.
4. J. Whiston, G. R. Taylor, *Proc. Eurochem. '83,* IChemE Symp. Ser. no. 79, Rugby 1983, pp. 335 – 343.
5. M. G. Kemp, P. Smith, in [4] pp. 344 – 353.
6. J. Springell, *Proc. Conf. Energy and the Process Industries,* IMechE, London 1985.
7. J. Redmond, *Chem. Eng.* (1987, Jan.) 10.
8. H. Bozenhardt, M. Dybeck, *Chem. Eng.* (1986, Feb. 3) 99 – 102.
9. *Heat Exchanger Design Handbook,* Hemisphere Publishing Corporation, 1983.
10. *Technology Applications Manual:* "Energy Conservation in Distillation," DOIE/CS/4431-T2, US Department of Energy, 1980.
11. A. Jenkins, A. Challinor, J. Jones, *Proc. Eurochem. '83,* IChemE Symp. Ser. no. 79, Rugby 1983, pp. 364 – 375.

12. D. Reay, *Chem. Eng.* (1976, Jul/Aug.) 507.
13. J. C. Ashworth, *Processing* (1985, May) 12.
14. *Drying, Evaporation and Distillation,* Energy Technology Support Unit Market Study no. 3, Energy Efficiency Office, London 1985.
15. J. Ashworth, *Chem. Eng.* (1986, Sept.) 30 – 34.
16. "Process Flowsheets," *Hydrocarbon Process.,* Nov. 1979.
17. *Eur. Chem. News* (1980, Sept. 8) 8.
18. L. F. Albright: *Processes for Major Addition Type Plastics and their Monomers,* McGraw-Hill, 1974.

Engineering Plastics → Plastics, General Survey
Engineering Plastics → Plastics, Properties and Testing

Pinch Technology

Bodo Linnhoff, University of Manchester Institute of Science and Technology, Manchester, United Kingdom
Vimal Sahdev, Linnhoff March Ltd., Manchester, United Kingdom

1. **Introduction** 1
2. **Composite Curves and Energy Targets** 1
3. **Pinch Principle** 3
4. **Energy – Capital Tradeoff** 3
5. **Process Modifications and Applications** 4
5.1. **Batch Processes** 5
5.2. **Cogeneration** 5
5.3. **Flexibility** 5
6. **Specific Examples** 5
6.1. **Modification of a Continuous Process** 5
6.2. **Integration of a Batch Operation** . . . 6
7. **References** 7

1. Introduction

The discovery of the pinch principle by Linnhoff has provided engineers with a scientific design methodology which has achieved outstanding results across the range of process industries [1]. The way in which industrial processes are designed and retrofitted has been revolutionized. Using the pinch principle, engineers can obtain a better understanding of the fundamentals of a particular process. Some of the original applications in ICI [2] and Union Carbide [3] showed energy savings and capital savings in new designs. The technology has evolved to incorporate cogeneration schemes [4], to improve plant flexibility [5], [6], increase production capacity [5], and reduce the effect of fouling [7] in both new designs and existing plants.

The early applications were in the oilrefining, petrochemical, and bulk chemical plants, but the technology has also been proved across a wide range of process industries, including pharmaceuticals, food, pulp and paper, cement, brewing, coffee making, and ice-cream and dairy products. Pinch technology has been used in various kinds of processes including batch, semicontinuous, and continuous operations incorporating various operating parameters, such as different feedstocks, seasonal demand fluctuations, multiple utilities, quality constraints, and environmental constraints.

2. Composite Curves and Energy Targets

All processes consist of hot and cold streams. A hot stream is defined as one that requires cooling, and a cold stream as one that requires heating. For any process, a single line can be drawn on a temperature – enthalpy plot which represents either all the hot streams or all the cold streams of the process. A single line representing all the hot streams and a single line representing all the cold streams are called the *hot composite curve* and the *cold composite curve*, respectively.

The construction of a composite curve is illustrated in Figure 1. Two hot streams are shown on a temperature – enthalpy diagram.

Stream 1 is cooled from 200 °C to 100 °C. It has a *CP* (i.e., mass flow rate times specific heat capacity) of 1; therefore, it loses 100 kW of heat. Stream 2 is cooled from 150 °C to 50 °C. It has a *CP* of 2; therefore, it loses 200 kW of heat.

The hot composite curve is produced by simple addition of heat contents over temperature ranges. Between 200 °C and 150 °C, only one stream exists and it has a *CP* of 1. Therefore, the heat loss across that temperature range is 50 kW. Between 150 °C and 100 °C, two hot streams exist, with a total *CP* of 3. The total heat loss from 150 °C to 100 °C is 150 kW. Since the total *CP* from 150 °C to 100 °C is greater than the *CP* from 200 °C to 150 °C, that portion of the hot composite curve becomes flatter in the second temperature range from 150 °C to 100 °C. Be-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_13

tween 100 °C and 50 °C, only one stream exists, with a *CP* of 2. Therefore, the total heat loss is 100 kW.

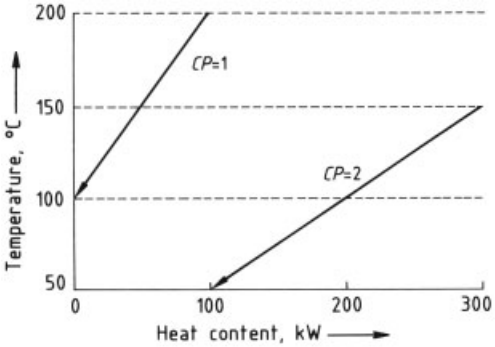

Figure 1. Two hot streams plotted on a temperature – enthalpy diagram

Figure 2 shows the hot composite curve for the problem. The cold composite curve is constructed in the same way. In practical applications the number of streams is generally much greater, but these streams are constructed in exactly the same way. Figure 3 shows the hot and cold composite curves plotted on the same temperature – enthalpy diagram. The diagram represents the total heating and cooling requirements of the process.

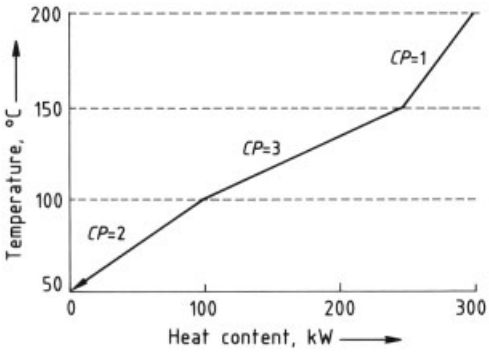

Figure 2. Hot composite curve

Along the enthalpy axis the curves overlap. The hot composite curve can be used to heat up the cold composite curve by process-to-process heat exchange. However, at either end an overhang exists such that the top of the cold composite curve needs an external heat source ($Q_{\mathrm{H,\ min}}$) and the bottom of the hot composite curve needs external cooling ($Q_{\mathrm{C,\ min}}$). These are known as the hot and cold *utility targets.*

The point at which the curves come closest to touching is known as the *pinch.* At the pinch the curves are separated by the minimum approach temperature ΔT_{min}. For that value of ΔT_{min}, the region of overlap shows the maximum possible amount of process-to-process heat exchange. Furthermore, $Q_{\mathrm{H,\ min}}$ and $Q_{\mathrm{C,min}}$ are the minimum utility requirements.

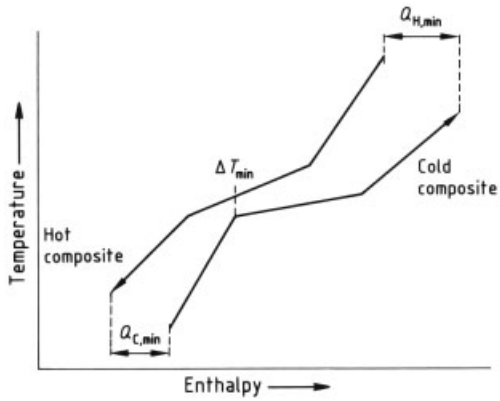

Figure 3. Composite curves showing the pinch and the energy targets

3. Pinch Principle

Once the pinch and utility targets of a process have been identified, the three "golden rules" of pinch technology can be applied. The process can be considered as two separate systems (Fig. 4) – a system above the pinch and a system below the pinch. The system above the pinch requires only residual heat and is, therefore, a heat sink, whereas the system below the pinch has heat to reject and is, therefore, a heat source. The three rules are as follows:

1) Heat must not be transferred across the pinch.
2) There must be no outside cooling above the pinch.
3) There must be no outside heating below the pinch.

If the amount of heat traveling across the pinch is α, then an extra amount (α) of hot utility must be supplied and an extra amount of cold utility α is required (Fig. 5). Similarly, any outside cooling of the heat sink and any outside heating of the heat source increases the energy requirements. Thus:

$$T = A - \alpha$$

where

T = target energy consumption,
A = actual energy consumption
α = cross-pinch heat flow

To achieve the energy targets, cross-pinch heat flows must be eliminated.

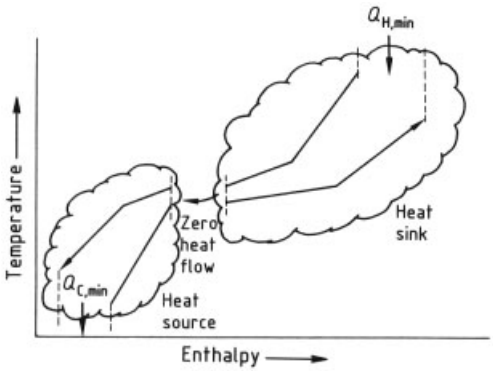

Figure 4. Schematic representation of the systems above and below the pinch

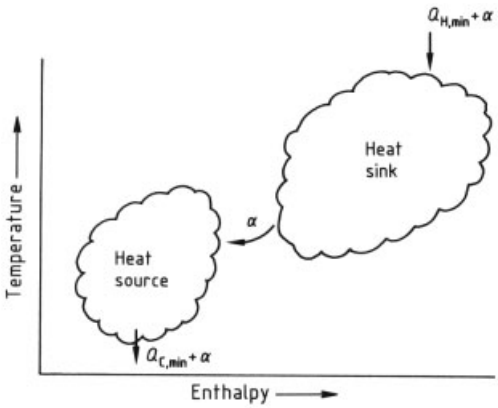

Figure 5. Heat transfer across the pinch from heat sink to heat source

4. Energy – Capital Tradeoff

The composite curves can be moved anywhere parallel to the enthalpy axis, because the composite curves represent changes in enthalpy between certain temperatures, and not absolute values. Thus, $\Delta T_{\min}$ can be increased or decreased simply by shifting one or both of the composite curves horizontally relative to the other. Vertical movement of the composite curves is not allowed since temperatures are governed by the process.

If $\Delta T_{\min}$ is increased, the overlap decreases and the energy target increases. However, capital cost decreases.

If $\Delta T_{\min}$ decreases, capital cost increases but energy cost decreases. Therefore, a tradeoff between energy and capital and, thus, a cost optimum exists. Figure 6 shows the effect of $\Delta T_{\min}$ on energy, capital, and total cost.

Pinch technology provides the value of $\Delta T_{\min}$ with the minimum total cost prior to design. This is determined by a procedure known as *supertargeting* [8].

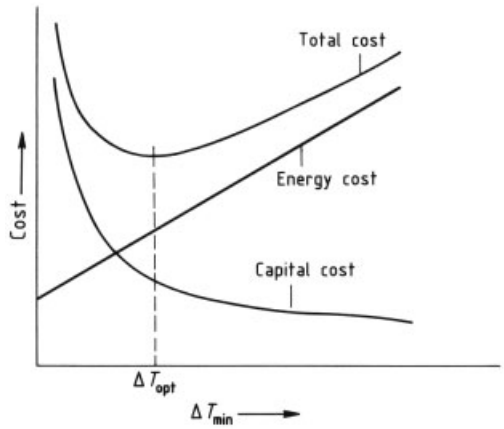

Figure 6. Energy cost, capital cost, and total cost plotted against $\Delta T_{\min}$

5. Process Modifications and Applications

The insight into the thermodynamics of the process provided by pinch technology enables an experienced practitioner to identify potentially advantageous process changes. For example, by alteration of the temperatures or pressures, it may be possible to reduce substantially the energy bill without an adverse effect on the product or process. Proper modification of the process also leads to other benefits, such as the removal of bottlenecks and the improvement of process operability.

A systematic search of the data for such opportunities is an essential element in any pinch technology survey. This can be an iterative procedure: once any beneficial changes are highlighted, the composite curves must be redrawn and the minimum energy needs recalculated, which may then open the way to further process changes [3], [9].

Table 1. Savings from some applications of pinch technology ∗

Process description	Savings
Crude oil unit	savings of ca. \$ 1.75×10^6 at 1.6 year payback
Large petrochemical complex manufacturing ethylene, butadiene, HDPE, LDPE, and polypropylene	savings over ca. \$ 7.00×10^6 with paybacks from 12 to 20 months
Tailor-made chemicals, batch process with 30 reactors and over 300 products	savings of ca. \$ 0.45×10^6 at paybacks of 3 months to 3 years
Sulfur-based speciality chemicals, batch and continuous	30 % savings to total site energy bill (worth ca. \$ 0.18×10^6 at paybacks of 9 – 16 months
Edible oil refinery, batch operation, wide range of feedstocks	savings of 70 % of process energy equivalent to ca. \$ 0.79×10^6 with paybacks from 12 to 18 months and debottlenecking equivalent to 15 % increased capacity
Batch processing of dairy products and dried beverages	savings of 30 % (equivalent to ca. \$ 0.20×10^6) with payback of less than 1 year
Brewery	savings from 12 % to 25 % of energy costs with paybacks from 9 months to 2 years
"State-of-the-art" whisky distillery	significant debottlenecking and savings of ca. \$ 0.35×10^6 with paybacks from 18 months to 2 years
Paper mill	savings of 8 – 20 % of energy bill at paybacks from 1 to 3 years
Continuous cellulose acetate processing	savings of ca. \$ 0.28×10^6 at 1 year payback
Continuous dry cement process	large energy savings

∗ Savings mentioned above are concerned primarily with energy costs. The majority of the companies also benefited from increased throughput and improved process flexibility and operability; the economic value of these benefits is not included in the table above.

Since 1983, pinch technology has achieved energy savings in the range of 10 – 70 %, reductions in the capital cost of new plants by up to 25 %, increased process capacity by the removal of bottlenecks, and greater flexibility and operability. The projects summarized in Table 1 indicate the variety of industries in which pinch technology has been successfully applied: food, paper, dairy products, whisky, cement, speciality and bulk chemicals, and oil refining. To date, the projects identified by pinch technology have been implemented, in whole or part, in all instances mentioned in Table 1.

5.1. Batch Processes

In the early days pinch technology achieved its most impressive results in continuous processes, but Table 1 includes a number of processes with exceptionally successful studies in batch operations. The key to applying pinch technology in noncontinuous processes lies in the data extraction. There are no shortcuts; detailed measurements and timings of all the process streams are essential if cost-saving opportunities are to be found. Where the technology has been applied to existing operations, there have frequently been wider process benefits, as well as lower energy bills. By improved integration of the plant, capacity has often been increased, without adversely affecting operability [10].

5.2. Cogeneration

The principles governing the correct siting of cogeneration systems were established several years ago [11], [12]. A pinch study uncovers opportunities for introducing such equipment or optimizing existing units.

In the pulp and paper industries, for example, power generation is an important element, and pinch technology has been applied successfully to improve performance and to evaluate alternatives. Some of the examples in Table 1 feature the use of cogeneration schemes.

5.3. Flexibility

A well-integrated plant need not be less flexible [5], [6]. In fact, the opposite is the case, with techniques having been developed to ensure that flexibility is incorporated into the process design in the most cost-effective way. For example, it is possible to allow for changes in feedstock and product specifications, turndown requirements, and seasonal variations.

In a study for Shell, UK, where flexibility was of paramount importance, an integrated design was produced that included twelve operating modes [13]. Throughput was also increased

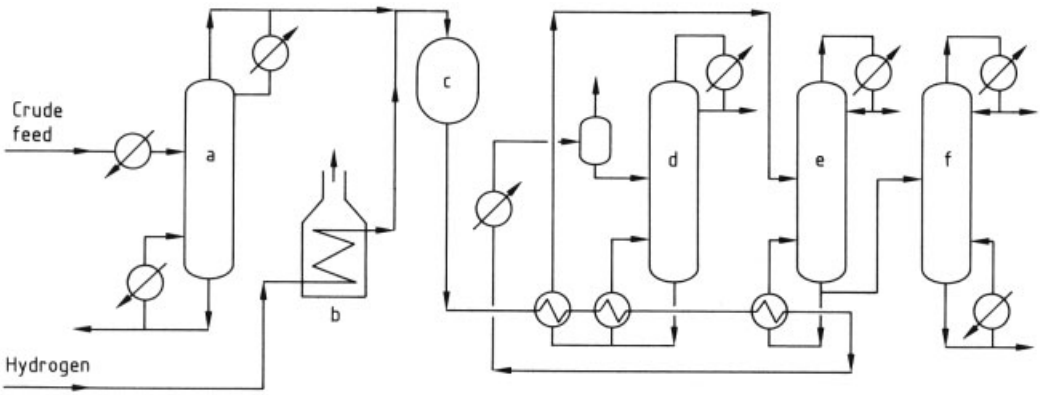

Figure 7. Example of a process flow diagram
a) Fractionation column; b) Furnace; c) Reactor; d) Stabilizer column; e) Benzene column; f) Toluene column

and energy savings of ca. $\$ 2 \times 10^6$ were achieved within a 1.6 year payback period.

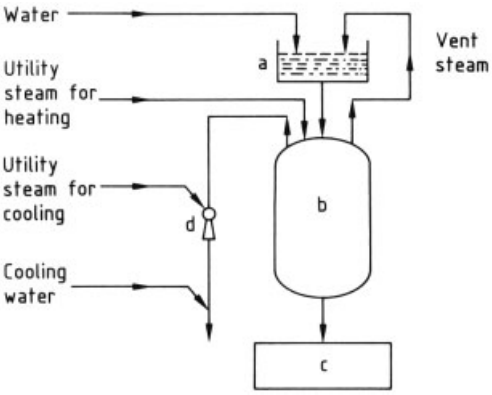

Figure 8. Original flow diagram of distillery cooker system
a) Water tank; b) Cooker; c) Fermenter; d) Ejector

6. Specific Examples

The two examples that follow give some indication of the way in which pinch technology is applied to specific problems and help to show why it has had such an impact on every industry that has used it.

6.1. Modification of a Continuous Process

Figure 7 shows a simplified flow sheet of an organic chemicals plant. The crude feed is separated into residue and light aromatic fractions, which are mixed with a preheated hydrogen-rich stream before entering the main reactor. The reactor product is cooled and partially separated before being fed to the first of a series of three distillation columns, each fitted with a reboiler and a condenser.

The pinch analysis without process modifications showed that the utility energy target was 22 % lower than the actual use. The pinch point occurred at 150 °C on the hot curve.

When the aspect of process modification was included, it was obvious that on all four of the columns (the initial fractionation column and the three distillation columns), the reboiler temperature was above the pinch and the condenser temperature was below it. If the possibility existed to reduce the reboiler temperature to below 150 °C, where excess heat is readily available, then the requirement for heating above the pinch would be correspondingly reduced, thus offering significant savings. In fact, practical considerations limited the scope for this change, but change was achieved on one of the columns by a reduction of the operating pressure.

The next step was to examine the furnace, which supplied most of the process heating. The flue gas exit temperature was well above 150 °C; therefore, recovery of this heat above 150 °C would reduce the heating target. After composite curves had been redrawn and process modifications incorporated, the revised energy target was less than half the former energy use.

The implementation of these proposals produced an integrated design that used 45 % less energy than before, with the investment that was involved being paid for in less than one year.

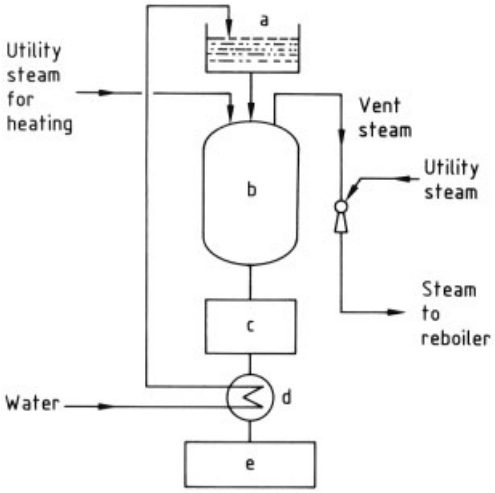

Figure 9. Distillery cooker system after integration
a) Water tank; b) Cooker, c) Buffer tank; d) Interchanger; e) Fermenter

6.2. Integration of a Batch Operation

The batch cooking operation in a distillery is shown in Figure 8. Preheated water and grain are cooked by steam injection; cooling begins by venting steam at above atmospheric pressure to the water preheat tank and is then speeded by a steam ejector that evaporates water from the cooker under vacuum. The cooled batch then moves to the fermentation stage.

The pinch analysis of the distillery as a whole had identified the pinch point at 90 °C, which meant that a substantial cross-pinch heat transfer existed within the batch cooking process. Vent steam hotter than the pinch was heating the process water below it, and the ejector exhaust steam was being condensed by cooling water.

The cross-pinch heat transfer was eliminated by a series of modifications. By introduction of a buffer tank between the cooker and the fermenter, the process water could be preheated from the cooling cooked grain, which was below the 90 °C pinch, thus eliminating the inappropriate steam ejector. The vent steam, no longer needed for preheating the process water, was utilized elsewhere on the site to reduce the need for steam to be brought in. The resulting energy savings were only part of the benefit; the cooking cycle was shortened because the cooling was faster and, thus, a process bottleneck was removed and the production capacity of the plant increased. The revised design of the batch cooker is shown in Figure 9.

7. References

1. B. Linnhoff., D. W. Townsend, D. Boland, G. F. Hewitt, B. E. A. Thomas, A. R. Guy, R. H. Marsland: *User Guide on Process Integration for the Efficient Use of Energy,* IChemE (Rugby, UK) 1982.
2. B. Linnhoff, J. A. Turner: "Heat Recovery Networks – New Insights Yield Big Savings," *Chem. Eng.* November 2, 1981 pp, 56 – 70.
3. B. Linnhoff, D. R. Vredeveld: "Pinch Technology Has Come of Age," *Chem. Eng. Prog.* July 1984, pp. 33 – 40.
4. B. Linnhoff, A. R. Eastwood: *Combined Heat and Power and Process Integration,* Institution of Gas Engineers, 51st Autumn Meeting, London, November 1985.
5. E. Kotjabasakis, B. Linnhoff: "Sensitivity Tables for the Design of Flexible Processes," *Chem. Eng. Res. Des.* **64** (1986) 197 – 211.
6. E. Kotjabasakis, B. Linnhoff: *Flexible Heat Exchanger Network Design: Comments on the Problem Definition and on Suitable Solution Techniques,* paper presented at IChemE Symposium "Innovation in Process Energy Utilisation" Bath, 16 – 18 September 1987.
7. E. Kotjabasakis, B. Linnhoff: "Better System Design Reduces Heat Exchanger Fouling Costs," *Oil Gas J.* September 1987, pp. 49 – 56.
8. B. Linnhoff, S. Ahmad: *Supertarget: Optimal Synthesis of Energy Management Systems,* presented at the ASME Winter Meeting, Anaheim, December 1986.
9. B. Linnhoff, H. Dunford, R. Smith: "Heat Integration of Distillation Columns into Overall Process," *Chem. Eng. Sci.* **38** (1983) no. 8, 1175 – 1188.
10. B. Linnhoff, G. J. Ashton, E. D. A. Obeng. *Process Integration of Batch Processes,* paper presented at the 79th AIChE Annual Meeting, New York November 1987, pp. 15 – 20.
11. D. W. Townsend, B. Linnhoff: "The Preliminary Design of Networks in Process Design," Part 1: 'Criteria for placement of heat

and engines and heat pumps in process networks', *AIChE J.* **29** (1983) pp. 742 – 771.
12. D. W. Townsend, B. Linnhoff: "Heat and Power Networks in Process Design," Part 2: 'Design Procedure for Equipment Selection and Process Matching', *AIChE J.* **29** (1983) no. 5, 748 – 771.
13. T. N. Tjoe, B. Linnhoff: "Using Pinch Technology for Process Retrofit," *Chem. Eng.* April 28 (1986) pp. 47 – 60.

Pine Oil → Turpentines
Pipelines → Natural Gas
Piperazine → Amines, Aliphatic
Piperidine → Amines, Aliphatic
Piping Systems → Chemical Plant Design and Construction
Pivalic Acid → Carboxylic Acids, Aliphatic

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Combustion

Standard Article

Selim Senkan[1] and Marco Castaldi[2]
[1]Department of Chemical Engineering, Los Angeles, CA, USA
[2]Precision Combustion Inc., North Haven, CT, USA

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2003 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b03_14.pub2
Article Online Posting Date: March 15, 2003

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
2. Conventional Combustion
2.1. Transportation
2.1.1. Fuels for Transportation
2.1.2. Combustion in the Transportation Sector
2.1.3. Emissions from Combustion in the Transportation Sector
2.2. Power Generation
2.2.1. Fuels for Power Generation
2.2.2. Combustion in the Power-Generation Sector
2.2.3. Emissions in the Power-Generation Sector
2.3. Industrial, Commercial, and Residential Combustion
2.3.1. Fuels for Industrial, Commercial, and Residential Use
2.3.2. Combustion in the Industrial, Commercial, and Residential Sectors
2.3.3. Emissions in the Industrial, Commercial, and Residential Sector
3. Nonconventional Combustion
3.1. Catalytic Combustion
3.2. Fuel Cells

[Top of Page]

1. Introduction

Combustion is the major source of the energy which drives our technological societies as well as being the principal source of air pollution. Direct combustion is a vast global activity and constitutes the basis for a large number of industries, including transportation, power generation, and heating. Collectively these activities represent a significant fraction of the world gross product.

Technically, combustion is the self-sustaining, exothermic interaction of a fuel with an oxidant through a series of chemical and physical events, ultimately resulting in the formation of water and carbon dioxide, which are the most stable reaction products. In this article, conventional combustion is defined as oxidation of fuel with formation of flames. Our everyday encounters with combustion are mostly subsonic events or deflagrations, in which combustion reaction waves propagate below the speed of sound. However, supersonic combustion events or detonations, in which wave propagation speeds exceed the speed of sound, also exist [1]. Probably the most extensive source of information on combustion is the *Proceedings of the International Symposium on Combustion*, which is regularly published by the Combustion Institute [2]. Information on combustion processes can also be found in the general references. Flameless or nonconventional combustion reactions are accomplished by means of catalysts, such as in catalytic combustion and in fuel cells.

Fuels for combustion are primarily based on fossil hydrocarbons such as natural gas (methane, ethane, and propane), petroleum products (gasoline, diesel, jet fuel, fuel oil), and coal and coal products (synthetic liquid and gaseous fuels). The oxidant is generally oxygen from the air because of its abundance, but there are exceptions. For example, hydrogen and pure oxygen are used in rocket propulsion and in fuel cells. Alternatively, the fuel and the oxidant can be part of the same molecule, as in explosives (e.g., TNT) and solid propellants [1].

The availability and importance of fossil fuels has changed with time. Coal was the dominant fuel in the early 1900s, replacing wood. The discovery and processing of petroleum significantly changed this picture, so that today, for example, transportation almost entirely depends on petroleum products. However, electric power generation by coal combustion continues to dominate. Natural gas, although initially dismissed as an energy resource, is now recognized to be one of the most abundant fuels; it is used largely for residential heating and cooking.

In Figure 1, the role of fossil fuels and therefore combustion in the global energy portfolio are shown. Combustion of coal, oil, and gas collectively provide nearly 80 % of world energy demand. Based on the projections in Figure 2, these fossil fuels, and hence combustion, are expected to maintain their dominant position for the next several decades.

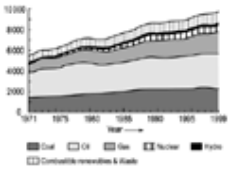

Figure 1. Total primary energy supply in Mtoe units (1 Mtoe unit = 16.3 GJ). ** Other includes geothermal, solar, wind, etc. Source: International Energy Agency, www.iea.com.

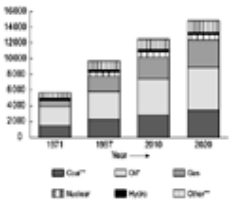

Figure 2. Outlook for total primary energy supply in Mtoe units (1 Mtoe unit = 16.3 GJ). ** Other includes geothermal, solar, wind, etc. Source: International Energy Agency, www.iea.com.

Emissions from combustion processes are the result of many factors, ranging from the type and composition of fuels to the designs of combustion systems to the additives that may be introduced. These emissions can be categorized into nonpolluting (desired) and polluting (undesired). The desired emissions are carbon dioxide (CO_2) and water (H_2O), whereas the undesired emissions include unburned fuel; nitrogen oxides (NO, NO_2, and N_2O, collectively referred to as NO_x), sulfur oxides (SO_2 and SO_3, or SO_x); and products of incomplete combustion (PIC), which include carbon monoxide (CO) and unburned hydrocarbons (UHC). Other PICs are volatile organic compounds (VOC, such as ethane, ethylene, propane, acetylene and solvents), aromatics, polycyclic aromatic hydrocarbons (PAH), and particulates (solid carbon or soot). Other combustion byproducts include formaldehyde, halogenated compounds, and metals. Although the relative amounts of these pollutants formed are small, their impact on urban environments can be significant due to the large-scale utilization of combustion. In fact, most air-pollution episodes in urban areas today are directly attributable to the pollutant emissions from combustion devices and their subsequent atmospheric reactions [3]. In the USA, National Ambient Air Quality Standards (NAAQS) have been set for a number of pollutants generated by combustion processes (Table 1).

Table 1. Major air pollutants and U.S. national ambient air quality standards (NAAQS)

Pollutant	Major Sources	NAAQS[a] & comments
Carbon dioxide (CO_2)	All combustion sources	Possibly injurious to health at concentrations greater than 5000 ppm over 2 – 8 h; atmospheric levels have increased from a preindustrial level of about 280 ppm to over 350 ppm in the 1990s, and the rate of increase has accelerated.
Carbon monoxide (CO)	Motor-vehicle exhaust; some industrial processes	NAAQS for CO: 9 ppm (10 mg/m^3) over 8 h (P), 35 ppm (40 mg/m^3) over 1 h (P), no more than once per year (for both)
Nitrogen oxides (NO, NO_2)	Motor-vehicle exhaust; heat and power generation; nitric acid; explosives; fertilizer plants	NAAQS for NO_2: 0.053 ppm (100 µg/m^3) over a year (P & S); react with hydrocarbons and sunlight to form photochemical oxidants
Sulfur oxides (SO_2, SO_3)	Heat- and power-generation facilities that use oil or coal containing sulfur; sulfuric acid plants	NAAQS for SO_2: 0.03 ppm (80 µg/m^3) over a year (P), 0.14 ppm (365 µg/m^3) over 24 h no more than once per year (P), 0.5 ppm (1300 µg/m^3) over 3 h (S)
Volatile organic compounds (VOC; include ethane, ethylene, acetylene, propane, butanes, pentanes, aldehydes, ketones,	Motor-vehicle emissions; solvent evaporation; industrial processes; solid-waste disposal; fuel combustion; petroleum refineries; gasoline stations;	React with nitrogen oxides and sunlight to form photochemical oxidants. Some VOCs, such as formaldehyde and benzene, are also hazardous to health.

solvents)	dry cleaners; printing; many paints	
Photochemical oxidants (primarily ozone (O_3); also peroxyacetyl nitrate (PAN) and aldehydes)	Formed in the atmosphere by reactions of nitrogen oxides, hydrocarbons, and sunlight	NAAQS for O_3: 0.12 ppm (235 μg/m^3) over 1 h[b] (P & S) no more than one day per year, 0.8 ppm (157 mg/m^3) over 8 h (P & S)
Air toxics (includes benzene, PAHs, vinyl chloride, formaldehyde, beryllium, inorganic arsenic, mercury, asbestos)	Specific types of chemical processes; coke ovens; building demolition	Typically cause cancer or are mutagenic or are hazardous to unborn infants, or cause neurological disorders. Current levels of emission may result in 1000 – 3000 cancer deaths each year.
Suspended particulate matter (PM10: particulate matter less than 10 μm in diameter; PM 2.5: particulate matter less than 2.5 μm in diameter)	Motor-vehicle exhaust; industrial processes; refuse incineration; heat and power generation; reaction of pollutant gases in the atmosphere	NAAQS for PM10: 50 μg/m^3 over a year (P & S), 150 μg/m^3 over 24 h no more than once per year (P & S); NAAQS for PM2.5: 15 μg/m^3 over a year (P & S), 65 μg/m^3 over 24 h no more than once per year (P & S); composed of carbon, nitrates, sulfates, and many elements including lead, copper, iron, and zinc
Lead (Pb)	Motor-vehicle exhaust; lead smelters; battery plants	NAAQS for Pb: 1.5 μg/m^3 over 3 months (P & S)

[a] P: primary (to protect public health), S: secondary (to protect public welfare).
[b] The ozone 1 h standard applies only to areas designated as nonattainment areas when the new ozone 8 h standard was adopted in July 1997.

Although carbon dioxide is a major desired product of combustion of fossil fuels, it is also a recognized greenhouse gas. Carbon dioxide together with other greenhouse gases, such as methane, absorb the infrared radiation reflected by the earth, resulting in a net increase of the energy deposited on earth by the sun and an increase in atmospheric temperature. Therefore, CO_2 emissions have become a global issue, especially after the 1997 Kyoto Protocol [4]. In Figure 3, combustion-generated CO_2 emissions are presented for each of the major fossil fuels. The CO_2 emissions are region-specific, with OECD nations generating nearly 50 % of global CO_2 emissions (Fig. 4). It is likely that in the future, CO_2 emissions will decrease from OECD nations and increase from non-OECD nations.

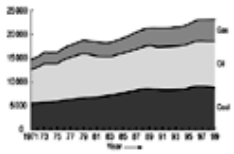

Figure 3. World CO_2 emissions by fuel source (in 10^6 t). Source: International Energy Agency, www.iea.com.

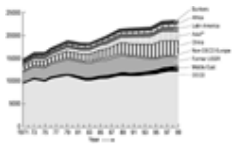

Figure 4. World CO_2 emissions by region (in 10^6 t). ** Excluding China. Source: International Energy Agency, www.iea.com.

CO is a flammable, colorless, odorless, and tasteless gas that is produced as a result of incomplete combustion of carbon-containing materials. CO is formed whenever carbon or substances containing carbon are burned with an insufficient oxygen supply. Even when the amount of oxygen is theoretically sufficient, the reaction is not always complete, so that the combustion gases contain some free oxygen and some CO. Incomplete combustion is especially probable when it takes place quickly, as in an automobile engine. CO is a major source of indoor air pollution, and it is categorized as a toxic gas because exposure to CO reduces the ability of the blood to carry oxygen.

NO_x, a mixture of nitric oxide (NO) and nitrogen dioxide (NO_2), is produced by most flame-combustion processes. Both molecular nitrogen in the atmosphere and chemically bound nitrogen in materials being burned can react with oxygen to form thermal NO_x and fuel NO_x, respectively. In general, the formation of NO increases with increasing temperature, while the formation of NO_2 decreases with increasing temperature. At common combustion temperatures, most of the primary emissions of NO_x are in the form of NO. As it is emitted into the atmosphere, smaller amounts of NO_2 are formed by further oxidation of NO. However, more than 90 % of the exhaust NO_x is still NO. In the atmosphere, a mixture of NO_x (mainly NO) and hydrocarbons (mainly VOCs) undergo photochemical reactions in ultraviolet light from the sun to form such secondary pollutants as ozone. Ozone in the upper atmosphere is beneficial because the O_3 layer helps protect us from the sun's dangerous ultraviolet rays. However, O_3 at ground level is quite the opposite because of its reactivity.

SO_x emissions are produced by combustion of sulfur-containing materials, such as coal. More than 90 % of the exhaust SO_x from combustors is in the form of sulfur dioxide (SO_2). In the atmosphere, SO_2 can be oxidized by O_3 and hydrogen peroxide (H_2O_2) with sunlight to form sulfur trioxide (SO_3) as a secondary pollutant. Both SO_2 and SO_3 have high solubilities in water. Therefore, in the atmosphere SO_x dissolves in water droplets to form sulfuric acid (acid rain).

Volatile organic compounds (VOCs), defined as any volatile compounds of carbon that include ethane, ethylene, acetylene, propane, butanes, and pentanes, but exclude methane, carbon monoxide, carbon dioxide, carbonic acid, metal carbides and carbonates, and ammonium carbonate, are produced by incomplete combustion of carbon-containing fuels or unburned fuels themselves. VOCs react with NO_x in sunlight to form ground-level ozone and photochemical smog.

Polycyclic aromatic hydrocarbons (PAHs) are trace products formed by the incomplete combustion of fossil fuels or other organic substances (e.g., tobacco) and are believed to be precursors for soot formation. PAHs are a group of more than 100 different compounds that include at least two aromatic rings in their structures. Some PAHs are potent carcinogens. Atmospheric PAHs have various forms; for example, highly volatile naphthalene (two-ring) exists in the gas phase, benzo[*a*] pyrene (five-ring) exists as solid absorbed on the surfaces of fine particles, and semivolatile phenanthrene (three-ring), pyrene (four-ring), and fluoranthene (four-ring) are distributed between the gas and solid phases. The mechanisms of formation and abatement of various combustion-generated pollutants are discussed further below.

Controlled conventional combustion processes occur in specially designed reactors or “chambers” that can withstand the high temperatures produced. Some examples of these chambers are cylinders in automotive internal combustion engines, and high-temperature alloy tubes or “cans” in energy-producing systems such as gas turbines and furnaces. It is the task of combustion engineers to design these chambers that can operate under demanding thermal conditions for extended periods. Nonconventional combustion processes operate at temperatures significantly lower than flames. For example, hydrogen-air fuel cells operate at about 80 °C while hydrogen-air flames can achieve temperatures of nearly 2000 °C. Catalytic combustors operate at about 800 °C and allow the processing of fuels over a broader temperature window than flames.

[Top of Page]

2. Conventional Combustion

Conventional combustion is defined as a gas-phase reaction involving the creation of a flame in which a fuel reacts with oxygen according to the stoichiometry given in Equation (1).

$$C_xH_y + (x + y/4)O_2 \rightarrow xCO_2 + y/2\,H_2O \qquad (1)$$

The source of O_2 is generally air, which contains about 21 % O_2 and 79 % N_2. This nitrogen dilution reduces the attainable flame temperature and also serves as the source of pollutants such as NO_x. Conventional uses of combustion are broad and range from production of energy and materials to waste-disposal processes. In manufacturing, the major use of combustion is to liberate the energy stored in the reactants to generate heat. The heat is used directly or indirectly. Some direct uses of heat include home/residential heating and drying processes such as for the paper industry, petroleum refining, and iron smelting. Indirect uses are for power generation, where the heat is converted to mechanical or electrical energy by specially designed systems. Combustion of gasoline produces heat, which is converted into mechanical energy by expanding gases in the automotive engine cylinder, which in turn produces a force on a piston. The piston then turns a drive shaft and thus heat is converted to mechanical energy for propulsion of a car. For production of electrical energy, a common system is a gas turbine, in which natural gas or diesel fuel is combusted to generate heat to expand gases, which are then passed trough a turbine producing work to turn a shaft connected to an electrical generator, thus converting the heat to mechanical energy and then to electrical energy. Combustion reactions are also exploited to manufacture materials such as carbon black, metal oxides, and ceramic powders [5-7].

Destructive uses of combustion besides as explosives are primarily limited to the field of incineration, where the combustion process is used to break down toxic materials into less toxic or benign ones so that the reaction products can be more easily disposed of [8], [9]. A significant complication encountered in incineration is the presence of heteroatoms. Unlike hydrocarbon fuels, which primarily consist of C, H, and O atoms, process waste streams can contain heteroatoms such as halogens, nitrogen, sulfur, phosphorus, metals, etc. The ultimate fate of these elements, i.e., their partition among different combustion products, depends on the combustion conditions, and will influence the subsequent air pollution control equipment needed to meet emission standards.

Flames can be categorized into two major groups: premixed flames and diffusion flames. Examples of these flames are the blue flames used in domestic gas-stove cooking and candle flames, respectively. In premixed flames the fuel and oxidizer are mixed at the molecular level, and the relative amounts of the reactants are described by the equivalence ratio Φ, defined by Equation (2).

$$\phi = (\text{fuel amount/oxidant amount})_{\text{actual}}/(\text{fuel amount/oxidant amount})_{\text{stiochiometric}} \qquad (2)$$

Based on this definition $\Phi > 1$ represents fuel-rich and $\Phi < 1$ fuel-lean, conditions. Premixed systems generally have lower and upper flammability limits, i.e., equivalence ratios within which self-sustaining flame combustion occurs. As expected, flame temperatures T_f first increase and then decrease as a function of increasing equivalence ratio, exhibiting a maxima around the stoichiometric conditions, i.e., $\Phi = 1$ (Fig. 5). The presence of excess air and excess fuel at lower and higher equivalence ratios, respectively, limits the maximum attainable flame temperature. Premixed systems also exhibit a unique combustion wave propagation velocity V_f which varies with equivalence ratio in a similar manner to temperature. Since increased flame temperature and velocity correspond to increased intensity of combustion, the formation and emission of

UHC mirror these variables, as shown in Figure 5. The NO_x quantities produced in flames generally follow the temperature pattern for a combination of thermodynamic and kinetic reasons discussed below.

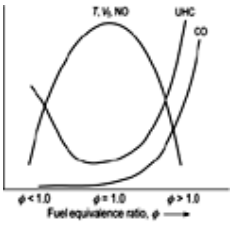

Figure 5. Effects of equivalence ratio on flame temperature T_f, flame velocity V_f, NO, UHC, and CO

In many practical applications that exploit diffusion flames, the fuel and the oxidant are brought together just before combustion for safety reasons, i.e., large-scale premixing is avoided to prevent the possibility of destructive explosions. In diffusion flames, combustion reactions occur at an interface to which the fuel and oxygen diffuse from bulk sources in accordance with the overall combustion stoichiometry. Thus, there is no unique equivalence ratio that describes a diffusion flame. Although one can define an overall equivalence for a diffusion-flame burner, local equivalence ratio Φ vastly varies from infinity (pure fuel) to 0 (pure air). Because a flame is produced at the stoichiometric interface, the maximum flame temperatures reached in diffusion flames are high, e.g., close to adiabatic flame temperatures.

For stable gas-phase (homogeneous) combustion, flame temperatures generally must exceed 1400 °C, and typically range between ca. 1500 and 1900 °C or more. The high temperatures preheat the incoming fuel and air through conduction, convection, and to a lesser extent by radiation. In premixed systems, the extent of air dilution can be used to control the maximum flame temperature (see Fig. 5). In diffusion flames, however, because stoichiometric conditions prevail at the flame front, the resulting maximum temperatures are considerably higher, e.g., about 2000 °C for natural gas and 2200 °C for diesel fuel flames, respectively. Consequently, diffusion flames are easier to stabilize than premixed flames. On the other hand, the higher temperatures associated with diffusion flames promote the formation of certain pollutants, such as nitrogen oxides and soot.

The pictures of premixed and diffusion methane flames in Figure 6 show that flame luminosity increases with increasing equivalence ratio due to increased production of soot. Increasing luminosity also increases radiative heat transfer, which is desirable to extract as much thermal energy as possible from combustion processes. Also shown in Figure 6 are probes to withdraw samples from within the flames, which subsequently can be analyzed by techniques such as gas chromatography, mass spectrometry, or both. Indeed, these types of measurements have been the primary sources of information on flame chemistry and have contributed significantly to our understanding of the chemistry and mechanisms of combustion. For example, it is now known that flames can contain hundreds of intermediate species. Figure 7 shows concentration profiles of some selected species in a fuel-rich methane flame, such as that shown in the middle picture in Figure 6, as a function of distance along the flame or height above the burner surface.

Figure 6. Laminar premixed flames of methane, slightly fuel rich (Φ = 1, left), fuel-rich and sooting (Φ > 1, center), and diffusion flame (right). Luminosity increases with increasing equivalence ratio.

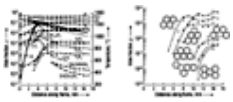

Figure 7. Concentration profiles of major species (left) and of some of polycyclic aromatic hydrocarbons (PAH, right) formed in a fuel-rich, premixed laminar methane flame. The formation of a large number of intermediates and byproducts is evident [10]. Carcinogenic benzo[a]pyrene is the third PAH from the bottom (right).

Flame Chemistry. As can be seen in Figure 7, even the chemistry of combustion of a simple hydrocarbon fuel such as methane is complex. The fuel molecule, before it ultimately produces CO_2 and H_2O, undergoes a complex series of reaction steps forming a large number of intermediates. The general sequence of reactions that lead to CO_2 production follows:

$$CH_4 \rightarrow CH_3^{\bullet} \rightarrow CH_2O \rightarrow CO \rightarrow CO_2 \tag{3}$$

CO oxidation represents the last sequence in CH_4 combustion, as well as in the combustion of all other hydrocarbons [10]. CO oxidation mainly proceeds by reaction with $OH^{\bullet}$ (Eq. 4).

$$CO + OH^{\bullet} \rightarrow CO_2 \tag{4}$$

$OH^{\bullet}$ is one of the important free radicals in flame-combustion processes, together with $H^{\bullet}$ and $O^{\bullet}$. Collectively $H^{\bullet}$, $O^{\bullet}$, and $OH^{\bullet}$ establish the radical pool in flames through the following chain-branching reactions:

$$H^{\bullet} + O_2 \rightarrow OH^{\bullet} + O^{\bullet} \tag{5}$$

$$O^{\bullet} + H_2 \rightarrow OH^{\bullet} + H^{\bullet} \tag{6}$$

These reactions also constitute the basis for H_2 combustion, an essential submechanism in the combustion of all hydrocarbon fuels. In fact, the destruction of the fuel and the formation and destruction of intermediates are all governed by free-radical reactions. For example, methane in flames is destroyed as a consequence of attack by $H^{\bullet}$, $OH^{\bullet}$, and $O^{\bullet}$ radicals (e.g., Eqs. 7-14).

$$CH_4 + H^\bullet \rightarrow CH_3^\bullet + H_2 \quad (7)$$

$$CH_4 + OH^\bullet \rightarrow CH_3 + H_2O \quad (8)$$

$$CH_4 + O^\bullet \rightarrow CH_3^\bullet + OH^\bullet \quad (9)$$

Once formed, $CH_3^\bullet$ radicals can then undergo recombination and dehydrogenation reactions to form a variety of C_2 species, the concentrations of which can reach substantial levels in fuel-rich flames (see Fig. 7 and Eqs. 10-14).

$$CH_3^\bullet + CH_3^\bullet \rightarrow C_2H_6 \quad (10)$$

$$C_2H_6 + H^\bullet \rightarrow C_2H_5^\bullet + H_2 \quad (11)$$

$$C_2H_5^\bullet \rightarrow C_2H_4 + H^\bullet \quad (12)$$

$$C_2H_4 + H^\bullet \rightarrow C_2H_3^\bullet + H_2 \quad (13)$$

$$C_2H_3^\bullet \rightarrow C_2H_2 + H^\bullet \quad (14)$$

Based on detailed chemical kinetic modeling studies, vinyl radicals appear to play an important role in flame chemistry. In the presence of abundant oxygen, vinyl radicals are rapidly oxidized to CO and CO_2 (Eqs. 15-17).

$$C_2H_3^\bullet + O_2 \rightarrow CHO^\bullet + CH_2O \quad (15)$$

$$CHO^\bullet \rightarrow CO + H^\bullet \quad (16)$$

$$CO + OH^\bullet \rightarrow CO_2 + H^\bullet \quad (17)$$

However, under fuel-rich conditions, further molecular growth becomes likely by competing polymerization reactions that lead to aliphatic C_3 – C_6, aromatic, and polyaromatic species [11] (e.g., Eqs. 18-22).

$$C_2H_3^\bullet + CH_3 \rightarrow C_3H_6^\bullet \rightarrow C_3H_4 + H_2 \quad (18)$$

$$C_2H_3^\bullet + C_2H_2 \rightarrow C_4H_5^\bullet \rightarrow C_4H_2 + H_2 + H^\bullet \quad (19)$$

$$C_4H_5^\bullet + C_2H_2 \rightarrow C_6H_7^\bullet \rightarrow C_6H_6 + H^\bullet \quad (20)$$

$$C_3H_4 + C_3H_3^\bullet \rightarrow C_6H_6 + H^\bullet \quad (21)$$

$$C_6H_6 + C_2H_3^\bullet \rightarrow C_6H_6C_2H_3^\bullet(+C_2H_2) \rightarrow C_{10}H_8 + H_2 + H^\bullet \quad (22)$$

Subsequent reactions of benzene and naphthalene ($C_{10}H_8$) then produce even higher PAHs, examples of which are presented in Figure 7 [11]. The interest in PAH formation and emission arises from the potential toxicity of some of the isomers of PAH. For example, benzo[*a*]pyrene $C_{20}H_{12}$ is an established potent carcinogen in laboratory animals [12]. Further growth of the hydrocarbon byproducts forms species with extremely low vapor pressures. These larger species then condense and undergo dehydrogenation to produce heterogeneous (i.e., liquid) nuclei for soot formation, an important universal characteristic of fuel-rich flames. This sequence of events is shown in Figure 8 [13]. As is evident from the foregoing discussion, the discharge of CO and hydrocarbon intermediates involved in fuel combustion processes can represent the products of incomplete combustion (PIC) emitted from combustion systems.

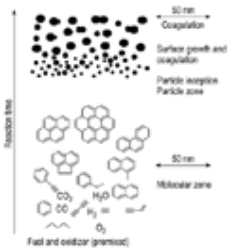

Figure 8. Schematic of soot growth from fuel molecules via PAH [13]

The combustion chemistry of higher molecular fuels such as propane, butane, gasoline (e.g., heptane, octane) also involves the reactions discussed above. That is, these fuels first undergo decomposition to reactive hydrocarbon intermediates and lower molecular mass fragments, such as methane and ethylene. Thus, it is the combustion of these decomposition products that determines the flame chemistry and PIC emissions from combustion devices burning high molecular fuels. This aspect of combustion chemistry has significantly helped in the development of detailed kinetic mechanisms of combustion of all types of hydrocarbon fuels [10].

Although the nature of intermediates formed in diffusion flames are similar to those in premixed ones, the contact pattern of the fuel and oxidant are significantly different. Figure 9 shows the species concentration profiles across a laminar ethylene diffusion flame front. The fuel arriving on the left and the oxygen on the right diffuse towards each other and undergo virtual annihilation within the flame zone with concomitant establishment of a peak temperature of about 1600 °C, which is

significantly higher than that of a premixed methane flame.

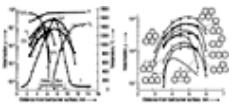

Figure 9. Species concentration profiles versus distance from fuel burner surface for major species (left) and PAHs (right) generated in an opposed-flow ethylene diffusion flame [14].

Similar to premixed flames, a large variety of hydrocarbon intermediates, including PAHs and CO are formed in diffusion flames (Fig. 9). These intermediates are also believed to form by reaction mechanisms similar to those of premixed systems. Since premixed systems provide a better control of combustor performance, and many practical combustion devices operate with separate fuel and air feed conditions, considerable effort has been expended to ensure rapid mixing of fuel and oxygen in combustion chambers and approach premixed conditions. Operation under turbulent-flow conditions is one such approach. In fact, most practical combustion systems today exploit turbulent mixing to dramatically reduce the mixing times of fuel and air.

Although higher flame temperatures are desirable for improved fuel combustion, flame stability, and for increased thermodynamic energy conversion efficiency, there are also drawbacks associated with high flame temperatures. Most significantly, the reaction of atmospheric molecular nitrogen with oxygen to form NO_x increases with increasing temperature, and this creates practical limits on the maximum flame temperatures that can be tolerated in combustion systems. This is clearly illustrated in Figure 10, where equilibrium NO levels are shown to increase with increasing temperature and decreasing fuel equivalence ratio in methane-air mixtures [15]. In flames, molecular nitrogen and oxygen interact with one another in accordance with the following Zeldovich or thermal NO mechanism (Eqs. 23 and 24) [16].

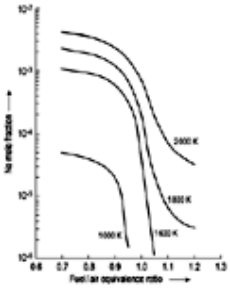

Figure 10. Equilibrium NO concentration as a function of equivalence ratio for various flame temperatures [15].

$$O^{\bullet} + N_2 \rightarrow NO^{\bullet} + N^{\bullet} \tag{23}$$

$$N^{\bullet} + O_2 \rightarrow NO^{\bullet} + O^{\bullet} \tag{24}$$

The kinetics of these reactions are such that their role becomes significant only above 1500 °C. Consequently, diffusion flames are particularly prone to higher levels of thermal NO production because of their higher peak flame temperatures. Besides the Zeldovich mechanism, NO formation can also occur via the prompt NO mechanism [17] and from fuel nitrogen sources. In the promp-NO mechanism, the reactions of CH radicals, produced by sequential degradation of hydrocarbon fuels with N_2, are responsible for NO production (Eqs. 25-30).

$$CH^{\bullet} + N_2 \rightarrow HCN + N^{\bullet} \tag{25}$$

$$HCN + O^{\bullet} \rightarrow NCO^{\bullet} + H^{\bullet} \text{or } NH^{\bullet} + CO \tag{26}$$

$$NCO + H^{\bullet} \rightarrow NH^{\bullet} + CO \tag{27}$$

$$NH^{\bullet} + O^{\bullet} \rightarrow NO + H^{\bullet} \tag{28}$$

$$NH^{\bullet} + H^{\bullet} \rightarrow N^{\bullet} + H_2 \tag{29}$$

$$N^{\bullet} + O_2 \rightarrow NO + O^{\bullet} \tag{30}$$

These and related reactions that are believed to be involved in prompt-NO production are presented in Figure 11 [18]. This reaction mechanism also plays an important role in the thermal deNOx process to remove NO from combustion products by using NH_3.

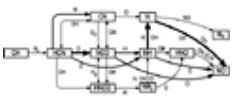

Figure 11. NO production from fuel reactions [18]

Coals and some heavy petroleum products contain nitrogen bound in organic molecules in amounts ranging from 0.1 to 2.5 % and thereby contribute to formation of nitrogen oxides. The rates of NO_x production from fuel-bound nitrogen strongly depend on combustion conditions, such as the local stoichiometry and temperature. The primary mechanism of NO_x formation in this case is conversion of fuel-bound nitrogen to HCN and HNCO, followed by the conversion of these species to

NH_x and then NO by a reaction mechanism identical to that for prompt-NO described in Figure 11.

In addition to NO, other nitrogen oxides, such as NO_2, and N_2O, can also form in combustion processes. NO_2 is produced from NO in the cooler parts of the combustion chamber (e.g., Eq. 31).

$$NO + HO_2^{\bullet} \rightarrow NO_2 + OH^{\bullet} \quad (31)$$

N_2O can be produced catalytically or by homogeneous recombination (Eq. 32).

$$O^{\bullet} + N_2 \rightarrow N_2O \quad (32)$$

NO_x Abatement. Lowering the amount of NO_x in combustion off-gases is one of the most significant challenges in environmental protection. There are two primary approaches for NO_x control: modifying the combustion process to minimize NO_x production, and postcombustion NO_x abatement. Usually, the nature of the fuel influences the selection of the most suitable approach. For low-nitrogen fuels such as natural gas, thermal NO is the major contributor, and temperature control is the most important parameter. Thus modifications to the combustion process that allow the use of excess air and its rapid mixing with fuel would be most effective to reduce NO emissions. Excess air serves as a diluent which reduces the maximum flame temperature, thereby resulting in lower thermal-NO production. For fuels that contain significant amounts of fuel nitrogen, both the thermal and the fuel NO formation processes become operative. In this case, the most effective strategy is to carry out the combustion process in stages. The first stage operates in a fuel-rich regime, and a second stage is fuel-lean. In the first stage, fuel nitrogen is released primarily as molecular nitrogen as opposed to oxides. In the second stage, unburned hydrocarbons and CO are converted into final combustion products. A variant of the staged-combustion method is reburning [19], [20]. In this case, the fuel – air mixture is first burned with excess air in the first stage, creating substantial levels of NO but relatively little unburned hydrocarbons. The products of the first stage are then introduced into a second stage, where a reburning fuel, preferably natural gas, is also introduced. In the fuel-rich reburning environment, both NO_x formation and removal reactions occur. However, the net effect is a significant reduction in the concentration of NO.

Methods to control NO_x emissions in post-combustion gases can be noncatalytic or catalytic. Noncatalytic methods involve the injection of nitrogen-containing additives at concentrations comparable to those of NO_x into the combustion exhaust gases over a well-defined temperature window to induce NO_x reduction. The additives used and their effective temperature windows are ammonia (NH_3, 825 – 1125 °C) in the thermal DeNOx process [21], cyanuric acid [$(HOCN)_3$, 725 – 1075 °C) in the rapreNOx process [22], and urea [$(NH_2)_2CO$, 725 – 1075°C) in the NOxOUT process [23]. IThe mechanisms responsible for NO reduction in each of these processes are similar. The principal reaction responsible for NO consumption is given in Equation 33.

$$NH_2^{\bullet} + NO \rightarrow N_2 + H_2O \quad (33)$$

A large number of methods have been developed for the selective catalytic reduction (SCR) of NO_x in combustion products [24]. At present, the SCR methods can be categorized into four groups: 1) Direct conversion of NO to N_2 and O_2; 2) SCR of NO with ammonia, primarily for industrial facilities and stationary power plants; 3) SCR of NO with CO and H_2, primarily for automotive applications; and 4) SCR of NO with hydrocarbons, again primarily for automotive applications. Modern auto exhaust emission control technologies largely depend on the utilization of three-way catalysts (TWC) to simultaneously reduce the emissions of CO, hydrocarbons and NO_x [25] (⟶ Automobile Exhaust Control). TWCs primarily utilize 0.1 – 1 wt % precious metals such as Pt/Rh combinations or Pd deposited on a high surface area carrier material primarily composed of γ-Al_2O_3 and CeO_2/ZrO_2 to serve as an oxygen-storage component during transient fuel-rich phases of cyclic engine operation. The γ-Al_2O_3 carrier material is also stabilized by oxides of rare earth and alkaline earth metals such as La and Ba to maintain high surface area and structural integrity under the influence of high temperatures and water vapor associated with engine exhaust gases. New catalyst formulations that can be used in lean-burn gasoline and diesel engines need to be developed to better utilize the higher efficiency of these engines [24].

Sulfur. Coal and some petroleum products also contain sulfur, which forms SO_2 and SO_3 (SO_x) on combustion. Over 60 % of anthropogenic emissions of SO_x are due to coal combustion, 30 % due to petroleum combustion, and the balance due to smelting operations. Anthropogenic emissions contribute over 75 % of the total atmospheric sulfur content, and the sources of these emissions almost exclusively are concentrated in the northern hemisphere. Together, NO_x and SO_x are contributors to acid rain by transformation into HNO_3 and H_2SO_4 in the atmosphere. These emissions, besides their obvious adverse environmental and health effects, also result in the erosion of structural and ornamental stonework and lead to corrosion of equipment. It is generally agreed that SO_2 formation occurs by sequential reactions of sulfur (Eq. 34) [26].

$$\text{Fuel-S} \rightarrow RS \rightarrow SO \rightarrow SO_2 \rightarrow SO_3 \quad (34)$$

The oxidation of SO_2 to SO_3 is thermodynamically favored at lower temperatures. However, since the homogeneous reaction kinetics are slow at low temperatures, SO_2 emissions dominate SO_x emissions from most combustion systems. On the other hand, significant SO_3 formation can occur through heterogeneously catalyzed reaction pathways when the fuel contains transition metals, especially vanadium, which subsequently appear as particulate matter. The most effective way to reduce SO_x emissions is to decrease the use of sulfur-containing fuels. In the absence of this, sulfur oxides can readily be removed from combustion flue gases by adsorption on a sorbent [27]. For example, calcium oxide (CaO), which can be obtained from

abundant limestone, can be used to efficiently remove SO_2 from combustion effluents (Eq. 34).

$$CaO + SO_2 + 1/2O_2 \rightarrow CaSO_4 \quad (35)$$

The $CaSO_4$ formed is stable and thus can safely be stored. However, the management of massive quantities of $CaSO_4$ that accumulate over a period of time can be a problem. Other, regenerative methods of SO_2 removal have also been developed [27].

Combustion of Coal. The combustion of coal is significantly more complicated than with gaseous and liquid fuels due to the presence of inorganic mineral matter (e.g., ash) [41-44]. Nevertheless, coal combustion occurs primarily through the establishment of diffusion flames. In this case, oxygen diffuses towards the solid fuel surface and reacts with either the products of gasification and/or volatilization at the flame zone or with the residual solid carbon. The combustion process can significantly be slowed down by the diffusion barrier created by the mineral matter. Thus, it is necessary to pulverize coal to affect efficient combustion. The solid fuel particles can shrink during combustion, ultimately leaving behind an ash particle. If the resulting ash particles are small, with terminal velocities smaller than the gas velocities, they will be carried over with the hot combustion gases in the form of fly ash. If combustion temperatures exceed the softening temperature of the ash, ash particles can form agglomerates. In any case, ash particles, together with organic particulate matter (i.e., soot) must be removed from the exhaust gases in coal combustors. (See General References on Power Generation)

Incineration is a special category of combustion in which the volume and potential toxicities of hazardous chemical wastes are reduced [8]. Incineration of halogenated hydrocarbons, and in particular chlorinated hydrocarbons (CHCs), have received particular attention since they are present in many consumer and industrial wastes, and their combustion is often associated with the formation of trace byproducts such as chlorinated aromatics, polychlorinated dioxins (PCDD), and furans (PCDF). The combustion stoichiometry of CHCs, designated by the formula $C_xH_yCl_z$, can be described by Equation 35.

$$C_xH_yCl_z + [x + (y - z)/4]O_2 \rightarrow xCO_2 + zHCl + [(y - z)/2]H_2O \quad (35)$$

That is, the thermodynamically preferred combustion product is HCl for $y > z$. However, for CHC compounds in which $y < z$, the formation of molecular chlorine must also be considered (Eq. 36).

$$C_xH_yCl_z + xO_2 \rightarrow xCO_2 + yHCl + [(z - y)/2]Cl_2 \quad (36)$$

Since the formation of Cl_2 is undesirable as it is difficult to scrub from combustion effluents, an auxiliary fuel with sufficient hydrogen content is used to increase the H/Cl ratio and suppress Cl_2 formation.

The combustion kinetics of CHCs also exhibit several distinguishing features when compared to the reactions of analogous unsubstituted hydrocarbons [28]. Most importantly, CHCs decompose easily as a consequence of the weaker C-Cl bond dissociation energies compared to the C—H bond in hydrocarbons, i.e., they start reacting at lower temperatures and begin forming combustion intermediates. Second, chlorine and chlorinated compounds also effectively compete with the important chain-branching combustion reaction of $H^{\bullet}$ radicals $H^{\bullet} + O_2 \longrightarrow OH^{\bullet} + O^{\bullet}$ (e.g., Eqs. 37 and 38).

$$H^{\bullet} + HCl \rightarrow H_2 + Cl \quad (37)$$

$$H^{\bullet} + CH_3Cl \rightarrow CH_3 + HCl \quad (38)$$

Thus, CHCs inhibit hydrocarbon combustion processes. This inhibition manifests itself in the reduction in laminar combustion velocities of hydrocarbon flames and delays in the onset of flame formation. In fact, these features of halogenated hydrocarbons have been exploited to develop effective flame retardants. For example, CCl_4 was used in earlier flame extinguishers, until its adverse health effects were discovered. Subsequently, CF_3Br (halon) was used in fire fighting. However, in spite of its effectiveness, CF_3Br is also banned due to its high ozone-depletion potential. Today, most halogen-based flame suppressants contain hydrogen in their molecules to increase their reactivity and therefore their destruction in the atmosphere by $OH^{\bullet}$ radicals [29].

Combustion of chlorinated compounds is also associated with the potential formation of toxic byproducts such as C_6H_5Cl and chlorinated dioxins and furans (see Fig. 12), as well as PAH and soot. The issue of byproducts in CHC combustion continues to be a highly controversial. This controversy arises from the fact that the 2,3,7,8 isomer of tetrachlorodibenzodioxin (commonly referred to as the TCDD, see Fig. 12) is highly toxic to certain laboratory animals [12]. At present two mechanisms exist to explain the formation of PCDD and PCDF in the combustion of CHC: 1) homogeneous synthesis from aromatic precursors such as chlorinated benzenes and phenols [30], and 2) heterogeneous and/or de novo synthesis from carbonaceous materials with particulate carbon and carbon species with functional groups believed to be the best starting materials [31], [32]. However, with proper design and operating procedures, the formation and emission of dioxins and furans can be minimized, and incineration can effectively be used as a powerful treatment technology in the management of toxic and hazardous wastes.

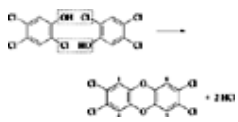

Figure 12. Formation of tetrachlorodibenzodioxin from 2,4,5-trichlorophenol

From the above discussions, it is clear that intimate mixing of the fuel and air are necessary to achieve clean combustion. However, energy recovery and the need to minimize the formation of NO_x bring additional constraints. For example, operation under fuel-rich conditions, such as those implemented in staged combustion, enhances radiative heat transfer and minimizes thermal NO_x formation. However, fuel-rich conditions promote PAH and soot formation. Thus, practical systems must be designed and operated to achieve the combined objectives of efficient energy recovery and environmental protection.

Modeling Combustion Processes. The objective of combustion modeling is to predict macroscopic phenomena, e.g., the pollutant species concentrations in and emitted from a combustor, from the knowledge of fundamental chemical and physical parameters that are process or device independent, together with a mathematical model for the process [10]. It is now possible and feasible to develop predictive models of combustion processes as a consequence of a number of advances made in related areas. First, with the availability of fast computers together with efficient numerical algorithms to handle stiff differential equations, an increasing complexity of combustion problems can be coded and solved rapidly without making any fundamental assumptions. Second, the quality and quantity of data on the thermochemistry of chemical species and the rates and mechanisms of their reactions relevant to combustion, have improved significantly to such a point that the detailed chemical kinetic mechanisms of combustion can readily be assembled. Third, with advances made in theoretical and computational quantum chemistry, accurate thermochemical information can be obtained via electronic structure calculations.

Some of the fundamental chemical parameters of interest include the thermochemistry of species, e.g., enthalpies of formation, entropies and specific heats, necessary to calculate equilibrium constants, and the reaction rate parameters such as the activation energies (E_a), and pre-exponential coefficients (k_o) for the rate coefficient expression $k_o\exp(-E_a/RT)$ to calculate combustion reaction rates. Relevant fundamental physical parameters of interest include the Lennard-Jones parameters, dipole moments, polarizabilities, and rotational relaxation numbers that are necessary for the calculation of transport parameters such as the viscosity, the thermal conductivity and species diffusion coefficients. These fundamental data, together with their associate uncertainties, are then used in modeling the macroscopic behavior of combustion processes. The Models are also subjected to sensitivity analysis to identify their components that are most important in influencing predictions. Combustion modeling efforts to date has been extremely successful in predicting relevant combustion phenomena in a large number of practical applications, including flames, shock tubes, soot formation, internal combustion engines, detonations [10].

Propellants and explosives also represent a special case of solid combustion, where the combustion reactions are exploited to create an impulse, e.g., to propel a missile or to cause fragmentation and blast [1, 2]. Propellants and explosives can be categorized into two basic groups. In the first category, the fuel and oxidizer parts are separate, such as in case of blackpowder which consists of 75 % potassium nitrate as the oxidizer, with 15 % charcoal and 10 % sulfur as fuels. In the second and more widely used category, the propellant or the explosive molecules combine the properties of fuel and oxidizer. Examples of these monopropellants, include trinitrotoluene (TNT), nitroglycerine (NG), ammonium nitrate (AN), cyclotrimethylene-trinitramine (RDX), and tetramethylene-tetranitramine (HMX). These monopropellants can be used either alone or in various combinations with each other and with other compounds for specific applications. An important aspect of these energetic materials is that, depending on the nature of initial ignition, they can exhibit either deflagrations or detonations. Detonation speeds of RDX and HMX can be as high as 8500 and 9100 m/s, respectively. The mechanisms of combustion of propellants and explosives are quite complex. However, the ultimate combustion products include the thermodynamically most stable compounds, i.e., CO_2, H_2O, N_2, SO_2. When the fuel and oxidant are separate entities, the proximity of fuel and oxidant, i.e., solid particle mixing, is important. In this case, the volatilization and the subsequent transport of at least one of the components would be required to sustain combustion. Volatilization can be achieved through heat transfer from the reaction zone. Consequently, the burning rate depends on gas pressure and particle size. In monopropellants, the mixing of fuel and oxidizer is no longer a problem. In this case, the volatilization and reaction can occur simultaneously. In principle, solid monopropellants, when heated by conduction, undergo surface decompositions and melting. Further heating and volatilization then creates a “fizz” reaction zone, which is followed by the generation of a luminous flame after a brief delay. The burning rates of monopropellants also depend on the pressure. Development of detailed chemical kinetic mechanisms describing the combustion of energetic materials is also being pursued today by a large number of investigators [1, 2].

2.1. Transportation

2.1.1. *Fuels for Transportation*

Transportation can be defined as any process that generates thrust for movement. Other than recently developed fuel-cell and electrically powered vehicles or rockets, all transportation relies on combustion of fossil fuels, primarily petroleum products. Nearly one quarter of total energy in the USA is consumed in the transportation sector, and of that 97 % is accounted for by petroleum products [33]. These fuels, overwhelmingly liquids, comprise gasoline, aviation (jet) fuels, and diesel fuels with only a small percentage being made up of natural gas. The liquid fuels derived from crude oil contain various combinations of different hydrocarbons and can be generally classified by the average molecular weights of the fuels or their carbon number, as shown in Figure 13.

Figure 13. Carbon numbers of various fuels used in the transportation sector

Gasoline (⟶ Automotive Fuels) is a complex mixture of over one hundred hydrocarbons. The hydrocarbons vary by class (paraffins, olefins, naphthenes, and aromatics) and, within each class, by molecular weight. The mixture of hydrocarbons (and added oxygenates, such as methyl *tert*-butyl ether or MTBE) in a gasoline determines its physical properties and combustion performance. Gasoline contains the lightest end of the liquid hydrocarbon fuels and thus is the most volatile, then the aviation fuels are next in terms of molecular weight and are less volatile, and finally there are the diesel fuels which are the least volatile and are considered to be fuel oils. Gasoline is used chiefly as a fuel for internal-combustion engines that

operate in a spark-ignited four-stroke mode. It is produced by the fractional distillation of petroleum; by condensation or adsorption from natural gas; by thermal or catalytic decomposition of petroleum or its fractions; by the hydrogenation of producer coal or gas; or by the polymerization of hydrocarbons of low molecular weight and from CO and H_2 through the Fischer-Tropsch process. The range of temperatures in which gasoline boils is roughly between 38 and 205 °C. There are many different types of gasoline grades from high-octane to reformulated blends to those that contain additives that assist in ignition, complete combustion, and reducing pollutant emissions. The U.S. Clean Air Act Amendments of 1990 required oil companies to make available cleaner gasoline with a higher oxygen content. This higher oxygen content is primarily added in molecules that contain oxygen in order to affect the combustion process to achieve complete oxidation of the gasoline to carbon dioxide and water while trying to eliminate the incomplete byproducts such as carbon monoxide (CO) and unburned hydrocarbons (UHC). Table 2 shows a representative composition of motor gasoline fuel.

Table 2. Sample composition of gasoline

Compound	Weight %
C_4	5
C_5	12
C_6 (nonaromatic)	15
C_7 (nonaromatic)	10
C_8 (nonaromatic)	5
C_9 (nonaromatic)	3.2
C_{10} (nonaromatic)	6.1
Benzene	3.2
Toluene	16
Xylenes	12
Styrene	0.7
C_9 aromatics	8
Other minor components	5.9

Aviation fuels (⟶ Aviation Turbine Fuels) are a subset of the fuel oil or diesel category and are comprised of a very narrow range of hydrocarbon types. These fuels conform to specific guidelines that control the ability of the fuel to flow at extreme temperatures as well as its ignition characteristics. The lower freezing point jet fuels are particularly suitable for long international flights, especially over polar routes during the winter. An aviation jet fuel's carbon number range is restricted by distillation, freezing point, and sometimes amount of naphthenes and/or smoke point requirements for the product. Kerosene-type jet fuel has a carbon number distribution between about 8 and 16; wide-cut jet fuel, between about 5 and 15 carbon numbers. Most of the hydrocarbons in jet fuel are members of the paraffin, naphthalene, or aromatic classes. When jet fuels of the same type differ, it is because they contain different proportions of compounds from these three classes.

Diesel fuels (⟶ Automotive Fuels) are also classified on the basis of various properties such as boiling point, freezing point, pour and cloud point, cetane number, density, viscosity and heating value. The reason for the existence of more specifications for the diesel fuels is their broader use in transportation, aviation, residential, and commercial purposes. A typical carbon number distribution for no. 2, diesel fuel, which is the dominant transportation fuel is shown in Figure 14. As with the other liquid fuels, diesel is a complex mixture of hundreds of individual compounds, most with carbon numbers between 10 and 22. Additives are also incorporated into diesel fuels to assist in ignition and combustion performance, and to visually distinguish various grades.

Environmental regulations have had a significant effect on the formulation of diesel fuels. The introduction of new fuels, coupled with the rapid changes in engine design to meet new emission regulations, has created the need to address several fuel properties to ensure proper performance, while also minimizing engine maintenance problems. The fuel with these enhancements is often called premium diesel, although other names are also used. The premium diesel concept is not the same as that of premium gasoline. Gasoline engines of certain types and compression ratios require a higher octane fuel to avoid knocking. Therefore, premium gasoline is required for these vehicles and is provided by all fuel suppliers. Premium diesel, on the other hand, is related to varying one or more fuel properties. Some of the these properties are cetane number, low-temperature operability, stability, lubricity, detergency, and heating value. Figure 14 shows the relative differences and classifications and cuts of gasoline, diesel, and aviation fuels with regard to carbon number distribution.

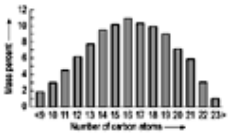

Figure 14. Typical distribution of carbon-containing compounds in no. 2 diesel fuel

2.1.2. *Combustion in the Transportation Sector*

Internal combustion (IC) engines are the primary combustion technology used in the transportation sector. They have many

uses: transport vehicles (motorcycles, automobiles, trains, aircraft, boats/ships); electric generators; pumps, gas compressors; construction equipment, agriculture and gardening equipment (mowers, chain and rotary saws, blowers). There are three major different types of IC engines employed today. The first is the spark-ignited (SI) four-stroke gasoline or Otto cycle engine, named after the inventor Nikolaus August Otto, which most passenger automobiles employ, in which combustion is best approximated by a premixed turbulent flame. The second type is the diesel engine, named after Christian Karl Diesel, which operates with direct injection, compression ignition (DICI) where the combustion process is very similar to a turbulent diffusion flame. Diesel engines are employed in trucks, buses, trains and in some automobiles, as well as in electricity-generating and marine-power plants. Both Otto-cycle and diesel engines are manufactured in two-stroke and four-stroke models. The third type is the rotary or turbine engine where combustion occurs in a chamber or “can”, where the combustion can be in the form of turbulent premixed or diffusion flame, depending on the application and fuel used.

The Gasoline Engine. In the four-stroke engine, first the intake valve opens and the piston moves down, creating a vacuum, and a mixture of air and vaporized fuel (gasoline) is drawn into the cylinder through the intake valve port (Fig. 15). When the piston reaches the lower limit of its travel, it begins to move upward. As this happens, the intake valve closes, as does the exhaust valve, sealing the cylinder. As the piston moves upward, the air-fuel mixture is compressed. In some small, high-compression engines, by the time the piston reaches the top of its travel, the mixture is compressed to as little as one-tenth of its original volume. Thus, the compression of the air-fuel mixture increases the pressure in the cylinder. The compression process also raises the temperature of the air-fuel mixture by adiabatic compression, yet the temperature is not high enough to ignite the mixture. As the piston reaches the top of its travel on the compression stroke, an electric spark is produced by the spark plug, which ignites the air-fuel mixture. The mixture burns rapidly, and cylinder pressure increases to as much as 4 MPa in a very short time and forces the piston down in the cylinder. The power impulse is transmitted down through the piston, through the piston rod (connecting rod), and to the crankshaft, which rotates. As the piston reaches the bottom of its travel, the exhaust valve opens. Now, as the piston moves up on the exhaust stroke, it forces the burned gases out of the cylinder through the exhaust port. When the piston reaches the top of its travel, the exhaust valve closes, and the intake valve opens. The cycle repeats again with the intake stroke. The four strokes are continuously repeated during the operation of the engine.

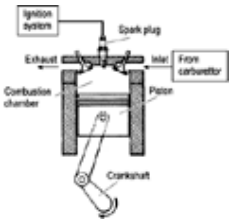

Figure 15. Spark-ignited gasoline engine [1]

During combustion there is a volumetric expansion of gases due to the greater number of moles of products than reactants. This is one of means of providing force on the piston to turn the drive shaft. Another means of creating a volume expansion is by the temperature rise caused by the change in enthalpy from reactants to products, i.e., the heat of combustion. Because the heat of combustion cannot be completely converted to mechanical energy, all engines must be equipped with some type of cooling system. Some aircraft and automobile engines, small stationary engines, motorcycles, and outboard motors for boats are cooled by air. In this system the outside surfaces of the cylinders are designed in an array of fins to expedite the dissipation of heat. Other engines are liquid-cooled and have their cylinders enclosed in an external liquid jacket. In automobiles, the coolant is circulated through the jacket by means of a liquid pump and the heated coolant subsequently is cooled by passing through the finned coils of a radiator.

As noted above, the power stroke is when the combustion energy is released. Since it is desirable to rapidly release of the chemical energy stored in the fuel, to prevent engine knock, significant effort has been spent to design fuel-delivery systems to generate highly premixed conditions. The premixing allows the flame front to progress rapidly through the mixture at a flame velocity determined by the equivalence ratio and the temperature attained during the compression process, with the maximum flame velocity attained under stoichiometric conditions. Interestingly, all present day SI engines operate nearly at stoichiometric conditions. However, there is a growing interest to operate internal combustion engines with excess air to improve fuel economy [25].

While there are many parameters that effect the final outcome of the combustion process, such as cold surfaces, level of unmixedness, turbulence, and pressure variations, an understanding of premixed flame combustion processes (see Chapter 2) have led to significant improvements in SI engine design and operation. For example, the combination of chemical kinetic mechanisms of combustion, determined from well-defined studies on premixed flames, with fluid dynamic models that represent the piston-cylinder systems have provided crucial insights into the origins of pollutants that are emitted from SI engines, leading to improved cylinder and piston-ring designs [10]. The majority of emissions caused by gasoline SI combustion systems are CO, NO_x, and UHC, which are greatly reduced by three-way catalytic converters [25].

The Diesel Engine. The four cycles of the gasoline engine are not much different in the diesel engine. However, in the diesel engine the fuel is not mixed with the air entering the cylinder during the intake stroke. Instead, the air alone is compressed during the compression stroke, and the diesel fuel is injected or sprayed into the cylinder at the end of the air-compression stroke. Compression ratios can be as high as 25:1 and generate pressures of to 3.5 MPa and temperatures as high as 550 °C. This temperature is high enough to spontaneously ignite the fuel as it is injected into the cylinder. The high pressure of the explosion forces the piston down, as in the case of gasoline engine. Since the fuel is injected as a liquid into the cylinder and forms droplets, the combustion occurs primarily through a diffusion-flame process in diesel engines. There are a number of complicating issues in diesel engines that are not present in SI engines. First, there is a need to sequentially vaporize, mix, and burn the liquid fuel with air within the time period of a compression/expansion stroke. If this is not achieved, products of incomplete combustion are emitted from the engines. In fact, diesel engines are well known for their significantly higher CO, hydrocarbon, PAH, and soot emissions when compared to gasoline engines. Second, because of the presence of diffusion

flames, diesel engines also exhibit higher local flame temperatures and thus generate higher levels of thermal NO_x.

Jet Engine. Jet engines, which exploit combustion-generated thrust, are associated with all modern aircraft. Jet engines enable the rapid transportation of large masses of people inexpensively over long distances. The gas turbine engine is widely used in both civilian and military sectors. However, regardless of the field of application, the main components of jet engine are similar (Fig. 16). The jet engine uses air as the working fluid to provide thrust. To achieve this, a compressor is used to increase the pressure and temperature of the inlet air to about 200 – 550 °C. Fuel, such as jet fuel or kerosene, is then sprayed into the pressurized air and ignited. Heated air and combustion products then pass through a turbine, which converts some of the combustion energy released into rotational energy to drive the compressor. The hot gases are then expanded through a nozzle producing high-velocity jet exhaust, which is used to propel the aircraft. The amount of fuel added to the compressed air depends on the temperature rise needed. However, the thermal properties of the turbine blades and other engine parts generally determine the maximum temperature that can be tolerated, which is in the range 850 – 1700 °C. Modifications of the gas-turbine engines include the use of multiple chambers and burners.

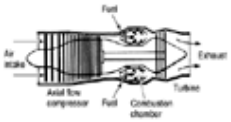

Figure 16. Schematic of a gas-turbine engine [1]

2.1.3. *Emissions from Combustion in the Transportation Sector.*

UHC, CO, and NO_x. Gasoline-powered vehicles tend to be the greatest emitters of CO, and substantial emitters of UHCs and NO_x, but only modest emitters of PAH and particulate matter (PM), while diesel-powered vehicles are substantial emitters of PAH, PM, and NO_x, as well as CO and UHCs.

UHC and CO emissions are incomplete combustion products that arise primarily because of lack of mixing of the fuel and air, i.e., innappropriate equivalence ratio. Secondary reasons for the emissions of these pollutants from combustion systems include the short residence times and lower local temperatures. According to the estimates by the EPA, gasoline engines are responsible for almost 75 % of anthropogenic CO emissions. One explanation of this is the difficulty in controlling emissions in periodically operated IC engines, where transient conditions are the norm. In contrast, stationary combustion facilities that operate continuously under nearly steady-state conditions are easier to control. The EPA estimates also indicate that gasoline engines generate about 24 % of the anthropogenic NO_x emissions. As discussed in the Introduction, CO and UHC are the primary products of incomplete combustion of fuel, and the main factor influencing their formation is the equivalence ratio. Stoichiometric mixtures are generally desirable to minimize UHC. On the other hand, NO_x production is primarily determined by reaction temperature, and thus stoichiometric conditions increase its production (Fig. 5). Therefore, a proper engine design involves a careful balance that simultaneously minimizes CO and NO_x emissions while maintaining stable operation.

PAH are generated in trace quantities as incomplete combustion products by all internal combustion engines. Mobile sources such as internal combustion engines burning gasoline and diesel fuel have been implicated as major contributors to PAH emissions in urban areas [34]. Quantitative data has been generally lacking for PAH emissions from aircraft, ships, heavy machinery, and two-stroke engines such as motorcycles, outboard motor powered boats, and lawnmowers. For example, vehicles using unleaded gasoline produce about 173 µg/km of naphthalene and 1.4 µg/km of pyrene [35]. However, because of their accumulation, even trace amounts emitted can pose a risk to human health and the environment over a period of time. Different PAH are also partitioned and react differently in the environment. For example, the phase distribution for naphthalene has been estimated to be 73 % in air, 8.5 % in water, and 17 % in soil, whereas the distribution of pyrene is 99 % in soil, 0.72 % in water, and 0.13 % in air. Significant PAH is also known to be associated with particulate matter emitted from combustion sources. Some PAH undergo atmospheric reactions and form products that are even more toxic than the parent molecules. For example, the formation of nitroaromatics (some of which are potent carcinogens) has been demonstrated for atmospheric reactions involving ozone and NO_x [36]. Clearly, the most effective strategy to protect the environment and human health from PAH is to minimize their emission from combustion devices in the first place. This indeed is the focus of intense research and development in combustion science and technology.

To avoid the emission of PAH from combustion devices, the use of gaseous fuels such as natural gas is expected to increase in the future. Because of the ease of premixing gaseous fuels with air prior to flame formation, combustion reactions proceed under conditions where the local equivalence ratio never reaches the sooting limits. Methane is the largest constituent of natural gas, present at concentrations of 85 % and above, followed by ethane at 10 % and below, then propane at 3 % and below. Smaller concentrations of butane and inert substances, such as carbon dioxide, nitrogen, and water are also present in natural gas. This mix gives natural gas an average hydrogen to carbon ratio (H/C) of about 3.5, which is considerably higher than those of petroleum-derived liquid fuels or coal. The use of natural gas as a combustion fuel is also desirable from the global warming point of view. First, because of its high hydrogen to carbon ratio, natural gas produces more water vapor and less carbon dioxide than liquid fuels or coal with the equivalent heat of combustion. Second, the combustion of methane is preferable to its direct venting to the atmosphere, which often is done at oil wells and remote gas wells, as methane is a more potent greenhouse gas than carbon dioxide.

Liquid fuels are generally expected to generate higher quantities of PAH than gaseous fuels due to their constitution. For example, gasoline and diesel fuels contain significant concentrations of aromatics and PAH, such as benzene and naphthalene, which serve as initial templates for further polymerization to form even larger PAH. Gaseous fuels, on the other hand, do not have such intrinsic templates, and as such are expected to produce less PAH than liquid fuels. Among the gaseous fuels, a similar trend is also observed. That is, the propensity for PAH formation increases with increasing molecular weight and decreasing H/C ratio. However, methane has been noted to be an exception. In a study with premixed flames,

higher concentrations of in-flame gas-phase aromatics and PAH were noted to exist in methane flames than in ethane flames [37], although soot formation was higher in the latter. This surprising finding was attributed to the presence of a larger set of reactions leading to PAH formation in the methane flame, as well as to the adsorptive removal of PAH from the gas phase in the sootier ethane flame.

Particulates. Particulate matter (PM) is solid or liquid particles in air that are larger than single molecules (ca. 0.0002 µm in diameter) but smaller than 500 µm. Combustion-generated PM is mostly soot. PM can be directly emitted from combustion systems or formed in the atmosphere through a series of chemical and physical events. Some PM is large or dark enough to be seen as soot or smoke, but mostly an electron microscope is necessary to detect it. Because PM originate from a variety of mobile and stationary sources (diesel trucks, woodstoves, power plants, etc.), its chemical composition and physical properties can vary. PM has lifetimes in suspension in air that vary from a few seconds (large particles) to several months (submicron particles). Particles less than 0.1 µm in diameter undergo random Brownian motion resulting from collision with individual molecules and can remain suspended in air indefinitely. Particles between 0.1 and 2 µm in diameter have small settling velocities in still air, and thus can be transported with prevailing winds. These particles tend to be removed by rain or washout. Particles larger than 2 µm have significant but small settling velocities. Particles above 20 µm in diameter have larger settling velocities and are removed from the air by gravity and other inertial processes, such as impaction. Since Brownian motion influences smaller particles, and inertial forces the larger ones, there is usually a stratification of the type of PM with distance from a given source. For example, PM emitted from a moving truck or car would deposit most of the largest particles on the roads or right-of-way, whereas the smaller particles would be carried significantly further throughout the environment and are more likely to be ingested.

Particulate emissions from all engines are of increasing concern due to growing awareness of their impact on human health. The chemical and physical characteristics of engine PM emissions depend on fuel composition, the engine design, whether it is carbureted or fuel-injected, and the cycle, two- or four-stroke. With leaded gasolines used earlier, lead compounds dominated the particulate emissions. With unleaded fuels, the PM comprise primarily soot particles. In addition, inorganic compounds such as ash, rust, and sulfates (from sulfur present in the fuel) can also form as PM. In the common four-stroke gasoline engines, PM emissions are typically on the order of 1 µm. The particles largely are formed and grow in the exhaust system by vapor-phase condensation. Soot forms in fuel-rich environments in flames when the fuel molecules are first pyrolyzed, then polymerize and condense into PM.

Diesel PM consists mostly of carbonaceous material generated from pyrolysis, incomplete combustion, or both. In addition, diesel fuel has a very high content of aromatics and is typically injected into the combustion cylinder as a liquid. The liquid is sprayed through a nozzle that generates a fine mist to assist in vaporization. This mist consists of many tiny liquid particles and must penetrate the combustion cylinder. The fuel-injection system parameters that determine the fuel injection rate, droplet size distribution and penetration into the combustion chamber, fuel vaporization, fuel – air mixing, and air temperature strongly influence the properties of the PM formed in diesel engines. Over 90 % of the particles released from two stroke diesel engines are less than 0.1 µm in size [38]. Similar to the gasoline engines, the formation of the particulates occurs via agglomeration of and surface deposition on very small particles originating from hydrocarbons in the diesel fuel. There is some evidence that the soot and PM emissions from combustion processes can have an impact on climate change [39].

As discussed earlier, methods are being developed to reduce the amount of pollutants that are emitted from combustion devices. Widely successful catalytic converters are one example of such a technology. Similarly the use of oxygenated fuel additives (e.g., MTBE) and oxygenated fuels (e.g., ethanol) provide benefits in decreasing pollutant emissions from combustion devices. These technologies, together with improvements in engine design and operation, have helped and will continue to help reduce the emissions of pollutants into the atmosphere. Of course, a variety of new technologies are also emerging that will dramatically reduce pollutant emissions from the transportation sector. Examples of these technologies include the hybrid gasoline/electric and the gasoline/hydrogen fuel cell vehicles, and ultimately the totally electric vehicle.

2.2. Power Generation

2.2.1. *Fuels for Power Generation*

Fossil fuels account for more than 70 % of power generated, of which coal has a contribution of 56 %, followed by natural gas with 10 %, and liquid petroleum products with 3 %. Nuclear plants provide about 20 % and hydroelectric sources about 10 % of our electricity demand. In spite of larger coal reserves, which are expected to last for several hundred years, there is a push towards increased use of natural gas combustion in power generation today because of increasing concerns over excessive CO_2 emissions and their potential impact on global warming. The resistance to nuclear power and the abundance of coal and natural gas reserves ensure their use for probably the next 100 years. Based on projections until 2020 (see Fig. 2), coal is expected continue to be the dominant fuel but to provide only about 42 % of electricity needs, followed by natural gas at 30 % [40]. The same study also predicts that nuclear plants will contribute about 10 % of our electricity demand, with the remaining 18 % coming from oil and hydro and other renewable resources. Since the combustion of liquid and gaseous fuels are discussed in Section Transportation, the focus of this section is on coal combustion.

Coal (⟶ Coal) is a solid, black carbonaceous material that comprises many different compounds in varying amounts. The composition varies significantly with source (mine), and even differs from location to location at the same source. While there is a significant variation in composition, the typical make up can be considered as follows: 60 – 95 % carbon, 2 – 7 % hydrogen, 1 – 30 % oxygen, usually about 0.5 – 10 % sulfur, and 0.25 – 2 % nitrogen. In addition, inorganic material such as aluminum, silicon, iron, thorium, uranium, and titanium typically exist in coal in amounts of 5 – 15 % and become ash when coal is combusted. The ash consists of inorganic oxides, such as silicon dioxide (SiO_2) or aluminum oxide (Al_2O_3) and consequently possesses a high melting temperature and therefore can form an outer layer or shell around the coal particle as it is burning. This presents special problems in that the layer is, at the very least, a resistance to diffusion of oxygen to reach the fresh fuel or for the products to diffuse away. The ash layer that forms may significantly arrest the combustion process, therefore leaving a large amount of the energy in the coal particle unused. A considerable amount of water is often incorporated in coal, usually in the range 2 – 20 %. This wide variation in composition makes it very difficulty to accurately predict the combustion characteristics of this fuel. Nevertheless, the combustion of coal can be considered as diffusion-flame-like, with the added complication of the ash layer. Table 3 lists compositions of some typical ranks of coals, together

with their energy content. Lignite is the lowest rank of coal with high contents of moisture and volatile matter and low heating value, especially on wet basis. Bituminous and subbituminous coals are next in rank, with lower contents of moisture and volatile matter and higher carbon content and heating value. Anthracite, the most metamorphosed coal, has low moisture and the highest carbon content. An excellent source of information on the classification, characterization, processing, and use of coals is *Chemistry of Coal Utilization* [41].

Table 3. Proximate and elemental analysis of some coal ranks, in wt % [42]

Type	Moisture	Fixed carbon	Volatile matter	C	H	N	S	O	Ash	Dry Heating Value, kJ/g
Anthracite	4.3	81.0	5.1	79.7	2.9	0.9	0.8	6.1	9.6	34 940
Bituminous	2.1	67.4	24.4	81.6	5.0	1.4	1.0	4.9	6.1	35 910
Lignite	15.2	43.1	40.7	70.1	5.1	1.1	1.6	21.4	5.9	28 280

2.2.2. *Combustion in the Power-Generation Sector*

Coal combustion is a complicated process which is not adequately understood [43-45]. However, it involves essentially two sequential steps: coal volatilization followed by char oxidation. When coal is subjected to high temperatures, it softens, and water is evaporated first. This is followed by the release of volatile organic material consisting of gases and tarry substances. The remaining mass, depleted in oxygen and hydrogen, and enriched in carbon, nitrogen, sulfur and mineral matter, is referred to as char. Volatilization products react with oxygen in the vicinity of the coal particle in a diffusion-flame mode, and this further heats the char. The char particles have cracks and holes made by escaping gases, can swell to a larger size, and thus can be considerably porous. Char combustion is believed to be a heterogeneous reaction process in which oxygen diffuses through the growing ash layer, is adsorbed on the char, and reacts to form carbon oxides, which diffuse back into the bulk gas stream. This heterogeneous process is much slower than the devolatilization step and thus limits the completion of coal combustion. In addition, this time scale is proportional to the original size of the coal particle. Consequently, all modern coal-combustion facilities utilize powders as opposed to granules or lumps of coal. Any residual mass that is not fully oxidized or cannot be transferred into the gaseous state or both (e.g., Al_2O_3, TiO_2, or C) can act as a seed particle to encourage further agglomeration or growth and becomes fly ash or slag. That is why coal combustion is very susceptible to particulate matter emissions, either carbon-based or inorganic.

Char combustion can occur in three distinct operational regimes. At low temperatures, the diffusion of oxygen to the surface is very fast compared to the rate of surface reaction, thus intrinsic reaction kinetics (R_{rxn}) limits the rate of combustion. The rate expression for R_{rxn}, can be given by Equation 39.

The rate expression for R_{rxn}, can be given by Equation 39.

$$R_{rxn} \approx S\rho d[O_2]^n A e^{-E/RT} \tag{39}$$

Where S is the surface area, ρ the density of the char, d the diameter of the char particle, $[O_2]$ the oxygen concentration of the bulk gas, n the order of the surface reaction, A and E are the Arrhenius parameters of the surface reaction, and T is the temperature. At high temperatures, surface reaction rates become very high, so much so that the rate of diffusion R_{diff} of oxygen to the char surface becomes the limiting process for combustion. Under these external diffusion limited conditions the oxygen reacts as soon as it reaches the char surface. The combustion reaction rate expression is then given by Equation 40.

$$R_{diff} \approx \left(\frac{D_o}{d}\right)\rho_o\left(\frac{T}{T_o}\right)^{0.75} \ln F \tag{40}$$

Where D_o is the average bulk gas diffusion coefficient, d is the characteristic char particle diameter, T is the reaction temperature, ρ_o and T_o are the reference gas density and temperature, and F is a function of the oxygen mass fraction in the bulk gas far from the particle (i.e., inlet O_2 concentration). From Equation 37, it can be seen that the combustion rate is inversely proportional to the particle diameter and moderately dependent on the temperature. Lastly, since the process is diffusion-controlled, the reaction rate is independent of the coal composition. In some applications, combustion can proceed at intermediate temperatures, so that observed combustion rates will be between R_{rxn}, and R_{diff}. Under these conditions, the oxygen is transported rapidly to the external surface of the char, but diffuses relatively slowly through the pores.

The amount of heat generated in coal combustion is related to its composition or rank. Carbon, the major component of coal, is the principal source of heat, generating about 32 000 kJ/kg at full combustion. Although hydrogen combustion can generate about 136 000 kJ/kg of heat, its presence in coal as fuel is limited to less than 5 – 6 %. Consequently, the heat of combustion of anthracite is closer to that of pure carbon (see Table 4). On the other hand, higher the oxygen content of coal, the lower its heating value. This is due the fact that oxygen in coal is chemically bound to the carbon, which is therefore already partially oxidized. The amount of heat contributed by the combustion of sulfur in coal is also relatively small, because of its low concentration in coal and low heating value.

Table 4. The role of combustion in the industrial sector in the USA (% total energy from

Industry	Steam	Heat	Combustion

Petroleum Refining	29.6	62.6	92.2
Forest Products	84.4	6.0	90.4
Steel	22.6	67.0	89.6
Chemicals	49.9	32.7	82.6
Glass	4.8	75.2	80.0
Metal Casting	2.4	67.2	69.6
Aluminum	1.3	17.6	18.9

Source: U.S. Department of Energy, Washington, D.C., www.doe.gov)

Most of the power-generation facilities in the USA today utilize fluidized-bed combustors (FBC), which not only allow the efficient operation of the combustion and associated heat-transfer processes, but also enable the control of pollutant emissions without external emission control devices (e.g., gas scrubbers). Fluidized beds suspend solid fuels on upward-blowing jets of air during the combustion process, as shown in Figure 17. The result is a turbulent mixing of gas and solids, and the heat-transfer tubes embedded into the bed allow the generation of steam. Cyclone separators are used to remove particulate matter from the flue gas prior to its entry into a gas turbine. The latter can be designed to accept a moderate amount of particulate matter, or ruggedized. The tumbling action, much like a bubbling fluid, provides more effective chemical reactions and heat transfer. In fluidized beds, coal burns at temperatures of 750 to 950 °C, well below the threshold for formation of nitrogen oxides (ca. 1400 °C). The mixing action provided by the fluidized bed also brings the flue gases into intimate contact with a sulfur-absorbing chemicals, such as limestone or dolomite, thereby removing more than 95 percent of the sulfur pollutants inside the reactor. Pressurized fluidized-bed combustion (PFBC) is a newer technology, and builds upon the earlier work in atmospheric fluidized-bed combustion technology. Power plants utilizing PFBC are significantly smaller and more efficient.

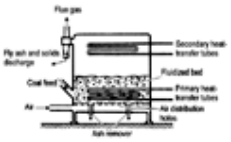

Figure 17. Schematic of an atmospheric-pressure fluidized-bed coal-fired boiler [17]

Another important reason for the popularity of fluidized-bed combustion technology is its fuel flexibility. Almost any solid fuel can efficiently be burned in fluidized-bed combustors, from low-rank lignites to anthracites. Even municipal waste can be burned as fuel in FBCs. This fuel flexibility, as well as the ability to meet sulfur dioxide and nitrogen oxide emission standards without the need for expensive add-on measures, has made FBC as the technology of choice in the power generation industry.

The majority of coal-fired power plants use steam turbines to produce electricity. However, as natural gas becomes more prevalent, the gas turbine engine will be used more for power generation. Currently, the primary fuel for gas turbines is natural gas, which is combusted in a premixed-flame mode. Some turbine combustors have small natural-gas diffusion flames as pilots to ensure the main premixed flame is stable. Other, usually older, designs employ a diffusion-flame combustor. These flames are described in Chapter Conventional Combustion. Natural gas is increasingly becoming the fuel of choice for new electricity projects worldwide. Improvements in gas turbine technology, the fact that natural gas burns cleaner than other fossil fuels, and confidence in its availability are factors that favor power generation using natural gas fueled generators. In fact, natural gas is currently (ca. 2002) the fuel for about 90 % of proposed new generation in the USA and is expected to gain shares relative to coal and nuclear power. Gas consumption is also expected to grow in South America, Asia, and the Middle East. Western Europe is moving away from reliance on nuclear power in favor of natural gas, and Eastern Europe is also expected to expand its natural gas capacity and depend less on coal. Unless rising natural gas prices stifle the forecast growth, world consumption of natural gas as a fuel will grow from 16 % in 1996 to 25 % in 2020 [46], mostly at the expense of coal (Fig. 2).

2.2.3. *Emissions in the Power-Generation Sector*

While NO_x emissions are on par with the transportation sector, coal-fired power plants also emit significant amount of radioactive material and sulfur oxides. Coal ash is composed primarily of oxides of silicon, aluminum, iron, calcium, magnesium, titanium, sodium, potassium, arsenic, mercury, and sulfur plus small quantities of uranium and thorium. Fly ash is primarily composed of noncombustible silicon compounds (glass) melted during combustion. Small glass spheres form the bulk of the fly ash. Trace quantities of uranium in coal range from less than 1 ppm in some samples to around 10 ppm in others. Generally, the amount of thorium contained in coal is about 2.5 times greater than the amount of uranium. For a large number of coal samples, according to EPA figures released in 1984, average uranium and thorium contents were determined to be 1.3 and 3.2 ppm, respectively. Using these values along with reported and projected consumption of coal by utilities provides a means of determining the amounts of potentially radioactive elements emitted by coal combustion. The main sources of radiation released from coal combustion include not only uranium and thorium but also their daughter products produced by the decay of these isotopes, such as radium, radon, polonium, bismuth, and lead. Although not a decay product, naturally occurring radioactive ^{40}K is also a significant contributor.

Electric power generation contributes nearly 50 % of all nitric oxide emissions, with the transportation sector accounting for the remaining 50 %. The electric utility sector is also the major source of sulfur oxides SO_x because of the dominant role of coal in power generation. Over the next decade, power plant operators will face significant requirements to reduce emissions of SO_2, NO_x, CO_2, and Hg. At present, neither the future reduction requirements nor the complete timetable is known for any

of these airborne emissions, and compliance planning is difficult. However, in response to the U.S. Clean Air Act Amendments of 1990 (CAA90), power plant operators are now reducing emissions of SO_x and NO_x.

The SO_x emissions can be controlled by the use of low-sulfur coal or other fuels such as natural gas and diesel that do not have sulfur to start with or can be processed to remove sulfur by passing it through sorbent beds. However, NO_x is still a problem because it comes from the nitrogen in the air and from the significant nitrogen concentrations in the coal matrix. A major effort therefore is focused on NO_x reduction. Those techniques vary from aftertreatment processes which prevent NO_x formed in the combustor from escaping into the exhaust to preventing the formation of NO_x in the first place. Some important NO_x abatement technologies are discussed in Chapter 1.

Techniques to prevent NO_x formation include lean premixed combustion, staged combustion, stoichiometric combustion with water injection, exhaust gas recirculation, and catalytic combustion. While significant progress has been made, with the exception of catalytic combustion, there is still reluctance to adopt these NO_x prevention techniques because of the possibility of destabilizing the combustion process.

2.3. Industrial, Commercial, and Residential Combustion

2.3.1. *Fuels for Industrial, Commercial, and Residential Use*

Combustion is universally used in the industrial, commercial, and residential sectors to meet steam and heat requirements. In addition to the basic premixed and diffusion-flame combustion modes, the combustion devices used in these sectors can be categorized into confined (indirect) and unconfined (direct) types. In the confined or standard “can” combustors, the combustion process is kept in a separate chamber remote from the point of application, such as in heaters, steam generators and boilers, where the products of combustion are exhausted. In other applications, the combustion process remains directly in contact with the application, as is the case with incinerators, some kilns, some space heaters, driers, and cooking stoves.

The fuels used in these sectors are predominantly gas (natural gas or refinery fuel gas) to about 70 – 80 % and to a lesser degree liquid petroleum products, such as diesel fuel, at 20 – 30 %. Residential applications also include fireplaces as space heating devices, and thus wood can also be considered as a fuel for this sector. Coal, once important to residential and commercial consumers, has almost totally been replaced by natural gas in most industrialized countries.

2.3.2. *Combustion in the Industrial, Commercial, and Residential Sectors*

Industry heavily relies on combustion for the production of a diverse array of products (Table 4). For example, boilers, furnaces and other process heaters consume roughly two-thirds of the total energy used by the manufacturing industries. Boilers are used in every major industry, and natural gas is the most common fuel. Exceptions are the petroleum and paper industries, where waste materials such as the refinery fuel gas and wood and pulp are available as fuels. Two primary boiler types are used in industry: water tube and firetube. In the former, tubes carrying water and steam are exposed to the products of combustion. In the firetube boilers, the products of combustion are carried inside the tubes, which heat the water and steam passing over them in the boiler vessel. Fluidized beds are also used in steam generation. In this case a fuel is burned inside a bed of granulated particles (sand) which are suspended by the upward flow of air and combustion products. Tubes carrying water are located inside the fluidized bed of particles.

Process heating equipment include furnaces, ovens, reactors, kilns, incinerators and dryers. Process heaters operate at low to medium temperatures, while furnaces, incinerators, and kilns operate at higher temperatures. The largest number of furnaces are found in the metals, metals fabricating, and heat-treating industries, followed by the glass and cement industries. The variety of process heating functions include changing the properties of materials, such as oxidation, reduction, decomposition, curing, baking, and drying. The nature of the material and the desired change determine the process temperature and the composition of the atmosphere employed in furnaces. Furnaces can be directly or indirectly fired. In directly fired furnaces, the flame and combustion products contact the material being processed. Examples of these types of applications include cement manufacturing and waste incineration. In indirectly fired furnaces, the combustion zone is confined to a separate chamber to prevent the contact of combustion products with the charge.

Commercial and residential applications of combustion are space heating, water heating, cooling, and cooking. Most commercial and residential combustion systems are designed for and operate at much less than optimal efficiencies to facilitate reliability and ease of use by inexperienced operators, and for esthetic reasons.

2.3.3. *Emissions in the Industrial, Commercial, and Residential Sector*

Pollutant emissions from these sectors are highly variable and sector-specific. Commercial and residential sectors, burning predominantly gaseous and liquid fuels, can produce the typical hydrocarbon-combustion emissions that include CO, NO_x, SO_x, particulates, and polycyclic aromatic hydrocarbons (PAH). Historically, residential wood combustion contributed the largest quantity of CO and VOC emissions in the stationary-source sector. However, with the switch to gaseous fuels, the control of trace levels of pollutants now dominate the development of new technology. The industrial sector contributed about 10 % of total U.S. NO_x emissions in 1998, the residential and commercial sectors approximately 5 %, the balance coming from the transportation and power-generation sectors.

Stationary sources such as residential heating that employ fireplaces and coal or diesel furnaces, meat cooking operations, open burning, coal-fired power plants, and coke ovens are other contributors to PAH emissions.

Relatively large peak exposures of PAH can be expected in local environments when charcoal grilling of meat products occurs. Although the grilling of meat is not likely to contribute a major fraction of the total PAH emissions to the atmospheric environment, studies have shown that local PAH levels are approximately 1000 times higher when charcoal grilling of meats occurs relative to PAH levels found in heavy road traffic [47]. This high local level of PAH is likely due to liquid grease

droplets falling into the gas flame or onto heated charcoal briquets. Vaporization or splattering of these grease droplets influences combustion, which likely follows a similar pattern to liquid-diffusion combustion.

Because of their broad use, there is significant interest in the development of burners that are robust, energy-efficient, compliant with pollutution emission regulations, and process friendly. In addition, multifuel capability and advanced flame-management systems are desirable to render them reliable and safe. The resulting furnaces are expected to process products uniformly and efficiently, and be cost effective with minimal fuel consumption and environmental impact. Boilers similarly are expected be fuel-flexible, cost effective, reliable, and safe.

[Top of Page]

3. Nonconventional Combustion

Nonconventional combustion processes do not involve flames or high temperatures in converting fuels into carbon dioxide and water. Significant attributes of some of the nonconventional combustion processes include the generation of little or no pollution, and higher energy conversion efficiencies compared to conventional combustion. Examples of nonconventional combustion processes include catalytic combustion and fuel cells. In addition, it is also possible to view biological processes, in particular cellular respiration, as a combustion process that converts carbohydrates such as sugars, into CO_2 and H_2O. However, the discussion of these biological processes is beyond scope of this chapter.

3.1. Catalytic Combustion

Catalytic combustion represents an attractive alternative technology to conventional combustion in which the formation and emission of pollutants such as NO_x, CO, UHC, PAH, and soot are significantly diminished. Gaseous fuels such as natural gas are most commonly used in catalytic combustion. The technology is applied broadly, from large-scale industrial gas-turbine combustors and boilers to small-scale heat generators such as residential boilers, heaters, and stoves. Besides low pollutant emissions, and hence high fuel conversion efficiencies, catalytic combustors also have the ability to operate over a wide range of equivalence ratios and power modulations, i.e., turndown ratios. Consequently, they are also used as afterburners to clean up pollutants in conventional combustion effluents.

Fuel oxidation in catalytic combustion is achieved through the participation of a solid catalyst, usually at temperatures below 1000 °C (Fig. 18). Combustion catalysts generally are metals supported on high surface area ceramic substrates, such as Al_2O_3, ZrO_2, MgO, or mixtures thereof. Active metals used as catalysts include Pt, Pd, Cr, and Rh. Because the catalyst provides a large thermal inertia, stable catalytic combustion can be achieved over a very broad range of conditions. The fuel and oxygen first diffuse to and then are adsorbed on the catalyst surface, followed by surface-catalyzed oxidation processes. The reaction products then desorb from the surface, regenerating the catalytic sites. Because of the close proximity of the fuel and oxygen on the catalyst surface, reactions achieve a higher degree of completion in catalytic combustion. As is the case with char combustion, catalytic combustion rates also exhibit different regimes of operation. At low temperatures (e.g., 100 – 300 °C), surface reaction rates are low relative to diffusion rates, and thus surface kinetics dominate the overall rate of combustion. However, at temperatures above several hundred degrees celcius, the rates of surface reaction become so large that the diffusion of reactants to the catalyst surface becomes the rate-limiting step in the combustion process. Generally, practical catalytic combustors are operated in the diffusion-controlled regime below 1000 °C. Consequently, thermal NO_x formation is virtually avoided in catalytic combustion. In addition, both prompt and fuel NO_x formation is reduced because fuel radicals are not released from the surface of the catalyst until they are completely oxidized to CO_2.

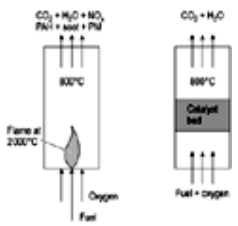

Figure 18. Catalytic combustion produces less pollutants than flames over a broader range of operating conditions

In addition to being utilized in a stand-alone device, catalytic combustion can also be used in conjunction with conventional combustion systems to stabilize premixed flames at lower temperatures. This is achieved by catalytically prereacting, and thus preheating, a portion of the fuel – air mixture entering a gas-phase combustion device. Preheating allows the stabilization of flames at equivalence ratios lower than otherwise possible, and thus extremely fuel lean and stable operation becomes feasible. The creation of reactive intermediates such as partially converted hydrocarbons or radicals in the prereaction process also stabilize the downstream flame. Depending upon the method of operation, flames can be catalytically stabilized at temperatures below 1300 °C. Thus, catalytic flame stabilization has an advantage over conventional lean premixed systems, in which combustion stability and increasing CO/UHC emissions become problematic at the low flame temperatures required for low NO_x emissions over a wide turn-down range (see Fig. 5). Additionally, premixed combustion is prone to combustion-driven acoustic oscillations or vibrations, which can affect flame stability and combustor life. By stabilizing the rate of gas-phase energy release in the combustor, a catalytic pre-reactor can significantly improve the performance of a lean premixed combustion system with respect to combustion dynamics.

The catalysts can be located either directly in the combustor itself or upstream of the combustion zone. In either case, the catalyst is located downstream of the fuel and air injection point, so that premixed fuel and air react as they pass through the catalyst bed. Some degree of gas-phase reaction, including CO burnout, generally is permitted downstream of the catalyst, so that the catalyst bed can remain small in size and 100 % catalytic oxidation is not required.

Durability is the overriding issue in catalytic combustion, because many metals exhibit good total oxidation characteristics. Since exposure to high temperatures, e.g., 1300 – 1800 °C, is a potential problem, innovative reactor designs become necessary to maintain the catalyst at more reasonable temperatures (e.g., < 1000 °C) where the catalyst can remain operational for tens of thousands of hours without deactivation.

Another significant challenge for catalytic combustion is the potential for severe thermal shock during rapid startup and shutdown processes. For example, the temperature of the catalyst can decrease at a rate of thousands of degrees per second when the fuel flow is abruptly stopped and air continues to flow over the catalyst, as can be the case during an emergency process shut down. Alternatively, the catalyst temperature can rapidly rise by 500 – 1000 °C when the system starts up or when additional fuel is rapidly injected. A catastrophic failure of the catalyst substrate is not acceptable because of the possibility of damage to the combustion hardware components, such as gas turbine blades, downstream of the catalytic reactor. Therefore, it is generally preferred to use metallic substrates or special packaging for ceramic substrates that accommodates their thermal fragility. The tradeoff of using metallic substrates is their lower thermal resistance in the demanding environment of a combustor. This requires the use of high temperature metal alloys containing Fe, Ni, Cr, Al. Other substantial issues with catalyst stability include sintering, adhesion, and volatilization [48-50].

3.2. Fuel Cells

Fuel cells (⟶ Fuel Cells) Fuel Cells convert the chemical energy of fuels directly into electrical energy using a catalytic electrochemical process. They provide a clean and highly efficient source of electrical energy, potentially to power vehicles and replace conventional combustion in power generation applications. Although fuel cell research dates back at least 30 years and was discovered nearly 100 years ago, automakers only recently have begun taking serious steps to develop fuel-cell-powered vehicles. Fuel cells are one of the cleanest and most efficient technologies for generating electricity. Since there is no flame formation, fuel cells do not suffer from any of the pollution problems associated with conventional combustion devices such as internal combustion engines, boilers, furnaces, and gas turbines. For fuel cells designed to utilize hydrogen directly, the only products are electricity and water, and small quantities of heat. Fuel cells also have very high energy conversion efficiencies. For example, they are 2 – 3 times more efficient than an internal combustion engine in converting fuel to power since they are not limited by the Carnot cycle.

When natural gas or other hydrocarbons are used, fuel cells produce some carbon dioxide. Advanced fuel cells using natural gas, for example, could potentially reduce carbon dioxide emissions by 60 % compared to a conventional coal plant and by 25 % compared to natural gas plants because of their high efficiencies. Moreover, the carbon dioxide is emitted in a purer form which makes its capture and sequestration much easier. Water is the only emission when hydrogen is the fuel.

In a fuel cell, hydrogen gas from the fuel reacts electrochemically (catalytically) at one of the electrodes (anode) and is transformed into protons and electrons, as shown in Figure 19. The protons move through an electrolyte to the other electrode (cathode), where they combine with oxygen from the air and with electrons to form water, which is expelled from the cell as vapor. The involvement of hydrogen and oxygen in the two reactions - one releasing electrons and the other consuming them - yields electrical current that can be tapped across the electrodes for power, as shown in Figure 19. Typically platinum is used as electrochemical catalyst both at the anode and cathode. Since fuel cells utilize highly selective catalysts and low temperatures (about 80 °C), the formation of NO_x is totally avoided. Since fuel cell catalysts are susceptible to sulfur poisoning, the fuels used must be processed to remove even trace of sulfur compounds. Therefore, fuel cells also are free from SO_x emissions. Finally, the production and emission of PAH, soot, and PM are negligible in fuel cells because of the mild operating conditions.

The largest barrier to the wide spread use of fuel cells is their cost of production and daily usage. First, fuel cells utilize substantial amounts of expensive platinum as catalyst, for which there is an uncertainty with regard to its worldwide supply. For example, according to one study, the displacement of conventional transportation by fuel-cell-powered vehicles in 66 years would require nearly 50 % of the world production of platinum, which must also be increased by 2.5-fold [51].

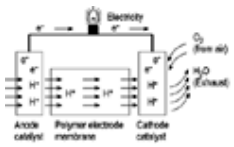

Figure 19. Operating principle of the hydrogen polymer electrode membrane fuel cell.

Second, the lack of readily available hydrogen, both as an inexpensive resource and distribution infrastructure, will also significantly add to the routine use of fuel cells in society. In fact, recognizing the infrastructure problem, current fuel cell research also addresses the on board creation of hydrogen from liquid fuels, such as gasoline, liquified natural gas, and alcohols. This requires the implementation of a catalytic reforming technology to convert these hydrocarbon fuels into hydrogen for fuel cell use and carbon dioxide, which must be vented. Thus, fuel cells using hydrocarbon fuels would continue to emit greenhouse gases.

In conclusion, only when inexpensive, nonhydrocarbon-based sources of hydrogen can be created, and ways to safely store and transport are developed, will fuel cells become viable enough to challenge conventional combustion systems. Even then, the question of availability of platinum must still be addressed.

[Top of Page]

References

General combustion

J. A. Barnard, J. N. Bradley: *Flame and Combustion*, Chapman and Hall, New York, 1985.

W. Bartok, A. Sarofim (eds.): *Fossil Fuel Combustion: A Source Book*, Wiley, New York, 1991.

H. Bockhorn (ed.): *Soot Formation in Combustion*, Springer Verlag, Berlin, 1994.

R. M. Fristrom, A. A. Westenberg: *Flame Structure*, McGraw-Hill, New York, 1965.

W. C. Gardiner, Jr.: *Gas-Phase Combustion Chemistry*, Springer Verlag, New York, 2000.

I. Glassman: *Combustion*, San Diego, California, Academic Press, 1996.

B. Lewis, G. von Elbe: *Combustion, Flames and Explosions of Gases*, Academic Press, New York, 1961.

C. Vovelle: *Pollutants from Combustion: Formation and Impact on Atmospheric Chemistry*, NATO Science Series C., Kluwer, Dordrecht, T2000.

Z. Warhaft: *An Introduction to Thermal-fluid Engineering: The Engine and the Atmosphere*, New York, Cambridge University Press, Cambridge, 1997.

J. Warnatz: *Combustion: Physical and Chemical Fundamentals, Modelling and Simulation, Experiments, Pollutant Formation*, Springer, New York, 1996.

Transportation

J. B. Heywood, E. Sher: *The Two-Stroke Cycle Engine: Its Development, Operation, and Design, Society of Automotive Engineers*, Taylor & Francis, London, 1999.

J. B. Heywood: *Internal Combustion Engine Fundamentals*, McGraw-Hill, New York, 1988.

Power Generation:

M. A. Elliott (ed.): *Chemistry of Coal Utilization*, Wiley-Interscience, New York, 1981.

A. H. Lefebvre, *Gas Turbine Combustion, Washington*, Hemisphere Publishing Corp., Washington, D.C., 1983.

J. P. Longwell, E. S. Rubin, J. Wilson, *Prog. Energy Combust. Sci.* **21**, (1995) 269.

K. L. Smith, L. D. Smoot, T. H. Fletcher, R. J. Pugmire: *The Structure and Reaction Processes of Coal*, Plenum Press, New York, 1994.

Reviews

J. M. Beer: “Combustion Technology Developments in Power Generation in Response to Environmental Challenges,” *Prog. Energy Combust. Sci.* **26** (2000) 301.

S. C. Hill, S. C. Smoot, L. Douglas: “Modeling of Nitrogen Oxides Formation and Destruction in Combustion Systems,” *Prog. Energy Combust. Sci.* **26** (2000) 417.

A. M. Mastral, M. S. Callen: “A Review on Polycyclic Aromatic Hydrocarbon (PAH) Emissions from Energy Generation,” *Environ. Sci. Tech.* **34** (2000) 3051.

S. S. Penner, A. J. Appleby, B. S. Baker, J. L. Bates, L. B. Buss, W. J. Dollard, P. J. Farris, E. A. Gillis, E. A. Gunsher, A. K. Khandkar, M. Krumpelt, J. B. O'Sullivan, G. Runte, R. F. Savinell, J. R. Selman, D. A. Shores, P. Tarman, “Commercialization of Fuel Cells,” *Prog. Energy Combust. Sci.* **21** (1995) 145.

M. Placet, C. O. Mann, R. O. Gilbert: “Emissions of Ozone Precursors from Stationary Sources: A Critical Review” *Atmos. Environ.* **34** (2000) 2183. Links

H. Richter, J. B. Howard, “Formation of Polycyclic Aromatic Hydrocarbons and their Growth to Soot – A Review of Chemical Reaction Pathways,” *Prog. Energy Combust. Sci.*, **26** (2000) 565.

S. X. Shi, “Recent Progress in Combustion Technologies for Automotive Engines” *Combust. Sci. Technol.* **7** (2001) 1.

L. D. Smoot, “A Decade of Combustion Research,” *Prog. Energy Combust. Sci.* **23** (1997) 232.

A. Williams, M. Pourkashanian, J. M. Jones, “Combustion of Pulverised Coal and biomass,” *Prog. Energy Combust. Sci.* **27** (2001) 587.

F. Zhao, M-C. Lai, D. L. Harrington, “Automotive Spark-Ignited Direct-Injection Gasoline Engines,” *Prog. Energy Combust. Sci.* **25** (1999) 437.

Specific References

1. J. A. Barnard, J. N. Bradley: *Flame and Combustion*, Chapman and Hall, New York, 1985.
2. *Proceedings of the International Symposium on Combustion*, The Combustion Institute, Pittsburgh, PA, 1961 to date.
3. J. H. Seinfeld: *Atmospheric Chemistry and Physics of Air Pollution*, Wiley-Interscience, New York 1986.
4. Kyoto Protocol, December 1 – 11, 1997; www.climatechange.org/kyoto.htm.
5. Z. A. Munir: *Combustion and Plasma Synthesis of High-Temperature Materials*, Wiley, New York 1990.
6. A. Makino, *Prog. Energy Combust. Sci.* **27** (2001) 1; Links S. E. Pratsinis, *Prog. Energy Combust. Sci.* **24** (1998) 197. Links
7. A. Varma, A. S. Rogachev, A. S. Mukasyan, S. Hwang, *Adv. Chem. Eng.* **24** (1998) 79.
8. C. R. Dempsey, E. T. Oppelt, *J. Air Waste Mgmt. Assoc.* **43** (1993) 25.
9. S. C. Saxena, C. K. Jotshi, *Prog. Energy Combust. Sci.* **22** (1996) 401. Links
10. V. V. Lissiannski, V. M. Zamansky, W. C. Gardiner, Jr., in W. C. Gardiner, Jr. (ed.): “Gas-Phase Combustion Chemistry”, Springer Verlag, New York 2000. C. K. Westbrook, F. A. Dryer, *Prog. Energy Combust. Sci.* **10** (1984) 1. N. M. Marinov, W. J. Pitz, C. K. Westbrook, A. M. Vincitore, M. J. Castaldi, S. Senkan, *Combust. Flame* **114** (1998) 192. Links S. Senkan, *Adv. Chem. Eng.* **18** (1992) 95.
11. M. Frenklach, H. Wang in H. Bockhorn (ed.): *Soot Formation in Combustion*, Springer Verlag, Berlin, 1994, p. 162.
12. C. D. Klaassen, M. O. Amdur, J. Doull: *Casarett and Doull's Toxicology*, 3rd ed., Macmillan, New York, 1986.
13. H. Bockhorn (ed.): *Soot Formation in Combustion*, Springer Verlag, Berlin, 1994.
14. N. Olten, N. Senkan, *Combust. Sci. Tech.* **159** (2000) 1.
15. C. T. Bowman in W. Bartok, A. Sarofim (eds.): *Fossil Fuel Combustion: A Source Book*, Wiley, New York, 1991.
16. Y. B. Zeldovich, *Dok. Akad. Nauk.* **51** (1946) 213.

17. C. P. Fenimore, *13th Symposium (Int'l) on Combustion*, 1971, p. 373.
18. C. T. Bowman in C. Vovelle (ed.): *Pollutants from Combustion: Formation and Impact on Atmospheric Chemistry*, NATO Science Series C., Kluwer, Dordrecht, 2000.
19. J. O. S. Wendt, C. V. Sterling, M. A. Matovich, *14th Symposium (Int'l) on Combustion*, 1973, p. 897.
20. L. D. Smoot, S. C. Hill, H. Xu, *Prog. Energy Combust. Sci.* **24** (1998) 385. Links
21. US 3 900 554, 1975 (R. K. Lyon).
22. US 4 731 231, 1988 (R. A. Perry).
23. US 4 208 386, 1982 (J. K. Arand, L. J. Muzio, J. G. Sotter).
24. V. I. Parvulescu, P. Grange, B. Delmon, *Catal. Today* **46** (1998) 233. Links
25. R. J. Farrauto, R. M. Heck, *Catal. Today* **55** (2000) 179. Links
26. A. J. Hynes, P. H. Wine in W. C. Gardiner Jr. (ed.): *Combustion Chemistry*, Springer Verlag, New York, 2000.
27. R. C. Flagan, J. H. Seinfeld: *Fundamentals of Air Pollution Engineering*, Prentice Hall, Englewood Cliffs, New Jersey, 1988.
28. S. Senkan in W. C. Gardiner Jr. (ed.): *Combustion Chemistry*, Springer Verlag, New York, 2000.
29. W. M. Pitts, M. R. Nyden, R. G. Gann, W. G. Mallard, W. Tsang, US National Institute of Standards and Technology, NIST Technical Note: 1279, 1990; R. G. Gann (ed.): *Halogenated Fire Suppressants*, American Chemical Society, Washington, D.C., 1975.
30. W. M. Schaub, W. Tsang, *Enviro. Sci. Tech.* **17** (1983) 721. Links E. R. Ritter, J. W. Bozzelli, *Combust. Sci. Tech.* **101** (1994) 153. Links
31. R. Addink, K. Olie, *Enviro. Sci. Tech.* **29** (1995) 1425.
32. L. Stieglitz, *Enviro. Eng. Sci.* **15** (1998) 5.
33. H. Miller, N. Mui, A. Pasternak, U.S. Energy Flow 1998, U.S. Department of Energy, Lawrence Livermore National Laboratory, UCRL-ID-129990-99, December 1999.
34. B. A. Benner, Jr., G. E. Gordon, S. A. Wise, *Environ. Sci. Technol.* **23** (1989) 1269. Links
35. W. O. Siegl, E. Chladek in *Symposium on the Mechanisms and Chemistry of Pollutant Formation and Control from Internal Combustion Engines, ACS.*, Washington, D.C., Aug. 1992.
36. R. Atkinson, *Atmos. Environ.* **3** (2000) 2063.
37. M. J. Castaldi, S. M. Senkan, *Combust. Sci. Tech.* **107** (1996) 141.
38. J. B. Heywood, E. Sher: *The Two-Stroke Cycle Engine: Its Development, Operation, and Design, Society of Automotive Engineers*, Taylor & Francis, London, 1999.
39. *National Academy of Sciences, Climate Change Science: An Analysis of Some Key Questions, Committee on the Science of Climate Change*, Division on Earth and Life Studies, National Academy Press, 2001.
40. *International Energy Agency*: “World Energy Prospects to 2020” Prepared for the G8 Energy Ministers' Meeting, Moscow 1998.
41. M. A. Elliott (ed.): *Chemistry of Coal Utilization*, Wiley Interscience, New York, 1981.
42. P.-C. Wu, H. C. Hottel in W. Bartok, A. F. Sarofim (eds.): *Fossil Fuel Combustion*, Wiley-Interscience, New York 1991, p. 783.
43. L. D. Smoot in W. Bartok, A. F. Sarofim (eds.): *Fossil Fuel Combustion*, Wiley-Interscience, New York 1991, p. 653.
44. K. L. Smith, L. D. Smoot, T. H. Fletcher, R. J. Pugmire: *The Structure and Reaction Processes of Coal*, Plenum Press, New York 1994.
45. A. Williams, M. Pourkashanian, J. M. Jones, *Prog. Energy Combust. Sci.* **27** (2001) 587.
46. U.S. Department of Energy, Energy Information Administration: International Energy Outlook 2001, DOE/EIA-0484, March 2001.
47. A. T. Dyremark, R. Westerholm, E. Overvik, J. Gustavsson, *Atmos. Environ.* **29** (1995) 1553. Links
48. S. T. Kolaczkowski, *Trans. IChemE* **73** (1995) 168.
49. J. H. Lee, D. L. Trimm, *Fuel Proc. Technol.* **42** (1995) 339.
50. R. A. Dalla Betta, *Catal. Today* **35** (1997) 129. Links
51. R. H. Borgwardt, *Trans. Res. D* **6** (2001) 200.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Electrically Generated Heat

Robert M. Perkin, Electricity Council Research Centre, Capenhurst, Chester, United Kingdom

1. Introduction 1
2. Resistance Heating 2
2.1. Direct Resistance Heating of Electrically Conducting Materials 2
2.2. Indirect Resistance Heating of Materials 2
3. Infrared Heating 4
4. Induction Heating 5
4.1. Direct Induction Heating of Metals . . 5
4.2. Indirect Induction Heating of Materials 6
5. Dielectric Heating of Nonmetals 7
6. Thermal Plasma Heating 9
7. Electron Beam Heating 11
8. Laser Heating 11
9. References 12

Symbols

A	area, m^2
I	current, A
P	power, W
R	resistance, Ω
V	voltage, V
f	frequency, Hz
l	length, m
δ	skin depth, m
ε''	loss factor
μ_r	relative magnetic permeability
ϱ	resistivity, Ωm

1. Introduction

The use of electricity for thermal heating offers a number of techniques applicable to a wide range of materials (Table 1). Direct heating involves applying a voltage, or equivalently an electric field, across the material to be heated (workpiece) causing an electric current to pass through the product. Energy is dissipated as heat due to the resistance that the material offers to the current flow. With indirect heating, energy is transferred from an electrically heated element to the product by thermal conduction, convection, or infrared radiation. The techniques to be discussed are distinct from those where energy is absorbed directly by intermolecular bonds to cause specific reactions without the body as a whole heating up. Applications where electricity is used indirectly in a heating system to provide the motive force, for example, electrically driven heat pumps, also are not discussed.

Electrical heating is commonly perceived to be expensive in comparison with other methods. While this can certainly be true, electrical techniques should not be rejected before an appraisal has been made: their advantages can radically improve a process and can far outweigh initial considerations of capital and operating costs [1, 2]. The general advantages of electrical heating are:

1) Rapid start up

Table 1. The range of electrical heating techniques

Technique	Type of material treated	Applications
Direct resistance	good electrical conductors	heating metals, melting glass, sterilization and cooking of foodstuffs
Indirect resistance	all materials	all forms of heat treatment, trace heating, ovens and furnaces
Infrared	all materials that absorb IR	drying and curing of surface finishes, drying of sheet material
Direct induction	metals, good conductors	metal melting and surface treatment
Indirect induction	all materials	heating materials in metal vessels, drying and calcining
Dielectric	lossy dielectrics, very poor electrical conductors	drying, food processing
Plasma arcs and torches	metals, nonmetals	welding, cutting, metal melting and refining
Electron beam	metals, nonmetals	welding, cutting, drilling
Lasers	metals, nonmetals	cutting, drilling, surface treatment

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_15

2) Capability of automation and accurate control of the process parameters
3) Clean working environment with no products of combustion
4) Large power densities, hence smaller equipment
5) Improved product quality

2. Resistance Heating

2.1. Direct Resistance Heating of Electrically Conducting Materials

Principles. The simplest way to heat an electrical conductor is to connect the leads from a line (mains)-frequency power supply across the ends of the product and use Joule or Ohmic heating arising from the electrical resistance R (Fig. 1). The time-averaged power dissipation is given by

$$P = 0.5 I^2 R = 0.5 V^2 / R \tag{1}$$

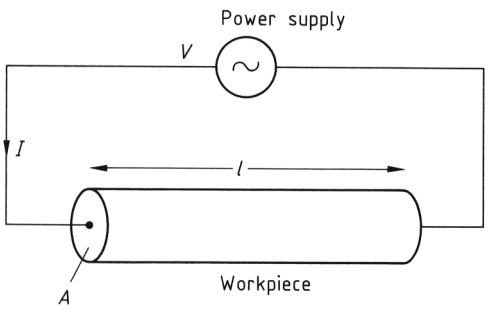

Figure 1. Principles of direct resistance heating (schematic); I current, V voltage, A cross-sectional area, l length of conductor (workpiece)

The second relation is obtained from Ohms law relating the applied voltage V to the current I flowing through the resistance R:

$$R = V / I \tag{2}$$

where V and I refer to peak values. At line frequencies the power is distributed uniformly throughout the volume of the material. Large powers can be obtained by increasing I and R. While the currents used can be as high as 10^5 A, considerations of economy and safety regarding the voltage impose an upper limit on the permissible value of R. The resistance is related to the resistivity of the material ϱ, length of the conducting path l, and cross-sectional area for current flow A by

$$R = \varrho l / A \tag{3}$$

Ideally, for a given volume the workpiece should be long in proportion to its cross-sectional area to maximize the value of R. Good electrical contact must be made between the supply leads and the workpiece; otherwise, localized overheating at the contacts can result, and the voltage required for heating is greater than really necessary.

Applications. This form of heating is used extensively in the metals industry for preheating and annealing. A less common application is in the direct heating of metal pipelines to improve the transport of viscous materials [1, pp. 303 – 321].

For heating liquids, these methods are employed in the melting of glass, the treatment of metal components in a heated molten salt bath (typically 100 L in volume with up to 100 kW power input), and the rapid generation of relatively small quantities of hot water or steam in compact electrode boilers (40 – 2500 kW power input). Another application is the sterilization and cooking of pumpable, semiliquid foodstuffs in continuous flow, with powers of up to 0.5 MW envisaged [3]. Such foodstuffs and industrial water supplies usually have enough dissolved salts present to make the product sufficiently electrically conducting for Ohmic heating purposes. Alternating currents and voltages must be used to avoid unwanted electrolytic effects.

2.2. Indirect Resistance Heating of Materials

Principles. If a directly heated metallic element is placed in contact with a workpiece, then heat is transferred to the latter by thermal conduction. Alternatively, if the element is mounted in an enclosure, heat transfer can occur by natural or forced convection, and infrared radiation. Only a small number of materials are suitable for the actual electrical resistor: nickel and iron alloys, for example, Nichrome, the so-called refractory metals, and at operating temperatures

above 1000 °C, some nonmetals such as silicon carbide and graphite, which are electrically conducting at these temperatures.

In use, the electrical conductor (element) may be left exposed, as in some air heaters, but more often it is held on a former within a metal sheath with powdered electrical insulation between the element and sheath. Flexible heaters can be formed from carbon loaded rubber and plastics, some "self-regulating" semiconductors, insulated wire ribbons, or meshes woven into or held on a fabric support. Power densities are typically 2 – 60 kW/m^2, depending upon the application. However, larger values are possible with specially designed cartridge heaters.

Applications. Indirect resistance heating is the most widely used form of electrical heating with loads ranging from a few watts up to a megawatt. A large proportion of applications however require less than 10 kW. In these circumstances, operating costs are often not important, and the advantages of low capital cost, good control, absence of combustion products, a high degree of flexibility, and safe operation in flammable areas can be exploited. The applications are too diverse to list individually. This technique is used in all forms of heat treatment, e.g., annealing, forging, stress relieving, and hardening of metals; firing, glazing, enameling, sintering, drying, baking, cooking, and melting of nonmetals [1, pp. 321 – 419]. Such versatility is possible because most materials can be raised to the required temperature if heated in a hot air stream or infrared beam for long enough. That is not to say that this method is the most appropriate in any given situation. Indirect resistance heating may be classified according to the method of application: cartridge, surface and immersion heaters, batch and continuous ovens, and furnaces.

Cartridge heaters are used to provide high-intensity localized heating by thermal conduction to dies, nozzles, platens, and blades. This is achieved by constructing the heater so that both the internal and external thermal resistance are minimized. In this manner power densities of 1 MW/m^2 can be obtained.

Surface heaters allow uniform heating, with large overall power inputs, but at modest power densities so that there is no danger of hot spots. Sheathed heating elements can be attached to the sides of metal vessels, or in the case of cylindrically shaped containers they can be arranged on a metal former, which can then be clamped on. Flexible resistance heaters are used to provide contact (trace) heating of valves and pipes in chemical processing plants to compensate for heat losses or assist the transport of viscous materials. Self-regulating heating tapes, which have the ability to decrease their power output as the temperature rises, are available for some trace heating applications.

Immersion heaters, commonly using metal sheathed elements, are used to heat liquids, gases, and soft metals with melting points below 650 °C at efficiencies close to 100 %. The power densities that can be achieved are generally limited by the thermal resistance at the sheath interface with the material. The heater elements in a U or W shape can be flange-mounted into the side of the holding vessel or, alternatively, may be hung from the top. Heaters in the form of a spiral with additional metal fins to increase the available heat transfer area can be mounted in ducts to heat air and other gases to high temperatures.

Electric ovens and furnaces are competitive with comparable fuel fired systems because of the general advantages mentioned above and their efficiency of electrical energy utilization: 40 – 60 % for an oven, 70 % for a tunnel furnace. The term furnace applies to equipment operating above 500 °C. An oven consists essentially of a double skinned metal chamber with thermal insulation sandwiched between the skins. Electric heating elements are mounted either in the oven, with suitable baffles and shields to direct the air flow over them, or externally in connecting ducts. The distribution of heaters and baffles is arranged to obtain the desired heating pattern. Fans are used to provide a forced draft of air. This type of construction can be used up to ca. 700 °C, but at higher temperatures refractory materials are required to line furnaces to provide the necessary insulation.

Heat transfer occurs by conduction through the workpiece supports, natural or forced convection, and radiation. While natural convection ovens are the simplest in terms of construction, forced convection is often to be preferred since greater heat transfer and better temperature control are possible. Radiation becomes the dominant mode of heating at temperatures above

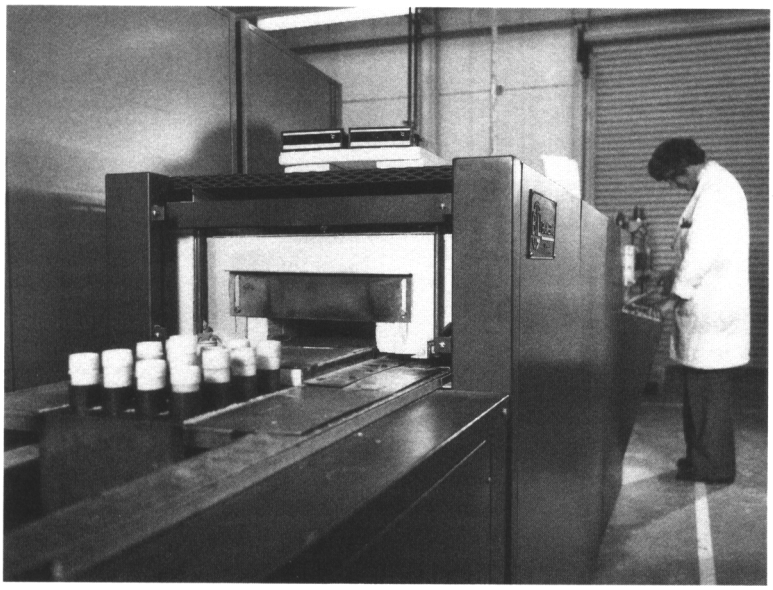

Figure 2. Batch furnace operating at 950 °C

500 °C for natural convection and 700 °C for forced convection. Where radiation is the principal mode of heat transfer, the furnace is built with the heating elements distributed so that the temperature of the inner furnace lining varies as little as possible to enable uniform radiant heating. Many designs of oven and furnace are possible depending upon the requirements: batch (Fig. 2) or continuous operation; front, top, or bottom loading; horizontal or vertical forced convection; transportation of the product by trolley, rollers, or various types of belt conveyor; operation under vacuum or in a specially controlled atmosphere, or the need for a well-defined processing temperature.

3. Infrared Heating

Principles. Any hot body will emit infrared (IR) radiation. The hotter the body, the smaller the wavelength at which the maximum energy is emitted and the greater the energy density (→ Radiation Heating). Electric infrared power sources utilize a heated electrical element, often sheathed within another material, to produce radiant heat. Details of typical sources are shown in Table 2. Elements are classified as short, medium, or long wave according to the element temperature, or equivalently the peak wavelength. The higher the operating temperature of the element, the greater the proportion of the electrical energy that is dissipated as infrared radiation. The remainder heats the surrounding air by convection. Emitters can be used individually, or more commonly, in modules with built-in reflectors. The modules in turn can be arranged to give the required energy profiles for the particular product concerned. Infrared radiation can be focused and reflected in the same manner as light waves, which enables power to be concentrated and transmitted relatively simply. Conversely, with irregularly shaped materials heating can be uneven due to portions of the product lying in shadow. The technique is better suited to continuous processes, with flat or regularly shaped products.

The efficiency of heating a product with infrared radiation depends on the ability of the material to absorb rather than reflect radiation at the peak wavelength of the emitter. Consequently, a knowledge of the absorption as a function of wavelength is desirable. As a rough guide, metals absorb radiation at wavelengths below ca. 1 μm, while nonmetals absorb at wavelengths above this. For most practical purposes the IR absorption can be regarded as occurring at the surface of the material. With some food prod-

Table 2. Commercial infrared emitter characteristics

Emitter type	Class [a]	Maximum temp., °C	Peak energy wave-length, µm	Maximum[b] intensity, kW/m^2	Maximum process temp., °C	Radiant heat, %	Convective heat, %	Heating/ cooling period	Expected life
Reflector heat lamp	SW	2200	1.2	10	300	75	25	1 s	5000 h
Tubular fused quartz	SW	2200	1.2	80	600	80	20	1 s	5000 h
Tubular fused quartz	MW	950	2.6	60	500	55	45	30 s	years
Flat quartz	MW	950	2.6	60 [c]	500	60	40	5 – 15 min	years
Metal sheathed	LW	800	3.0	40	500	50	50	$\leq$ 2 min	years
Ceramic	LW	700	4.0	40	400	50	50	$\leq$ 2 min	years
Focused spot radiator	SW	2300	1.0	–[d]	1600	98	2	1 s	50 h [e]

[a] SW Short wave; MW Medium wave; LW Long wave.
[b] Measured close up to element.
[c] Forced air cooling preferable above 40 kW/m^2.
[d] 5 – 30 W/mm^2.
[e] Life is longer if gun is used at reduced voltage.

ucts, however, the radiation can penetrate several millimeters before all is absorbed.

The choice of emitter in a given application is determined not only by the absorption characteristics of the workpiece but also by the process conditions, e.g., the presence of mechanical vibration, the possibility of surface contamination, and the need for fast shut-down in the event of an interruption to the process. Some compromise between these various factors is often required. Most materials, excluding metals, cannot absorb more than 15 – 20 kW/m^2 without sustaining damage. Consequently an installed power density of not more than 20 – 25 kW/m^2 is advisable.

Applications. There are a wide variety of uses, which can be grouped into three main areas: drying and curing of paints, inks, resins, and other finishes; mass heating ranging from plastics molding to sterilization; and moisture removal from, for example, webs and granular material. The design of the equipment ranges from relatively simple modules to complex computer controlled installations. Examples are belt-conveyor and tunnel ovens, double sided panels for on-machine mounting, and static panels for even heating of large areas [4].

An important development is the manufacture of elements consisting of short-wave lamps and panels of very low thermal mass material. These panels can absorb the short-wave IR radiation reflected from the product and re-emit radiation at longer wavelengths. This results in an efficient, fast response emitter. Such advances, together with improvements in sensing, enable much more precise control of power levels throughout the process and consequently new opportunities for this form of heating are appearing.

4. Induction Heating

4.1. Direct Induction Heating of Metals

Principles. Induction techniques enable a voltage to be applied directly across a workpiece without the need for connecting leads from the generator [5]. This is achieved by utilizing a basic result of electromagnetism, namely, that a magnetic field that changes with time generates a voltage in a circuit. The technique is similar to that used in a transformer and can be explained with reference to Figure 3A (see next page). A time-varying magnetic field is produced in the primary coil by an alternating current from the power supply. A second coil is positioned so that the changing magnetic field passes through it to produce a voltage, which in turn gives rise to a current *I,* often referred to as an *eddy current,* in the secondary circuit. Energy is dissipated in the resistor R by Ohmic heating (I^2R). Taking the heating of a metal cylinder as an example (Fig. 3B), the workpiece forms the secondary circuit and is placed within the primary coil to ensure good penetration (linkage) by the magnetic field. The metal cylinder corresponds to a single-turn coil with the resistance distributed throughout.

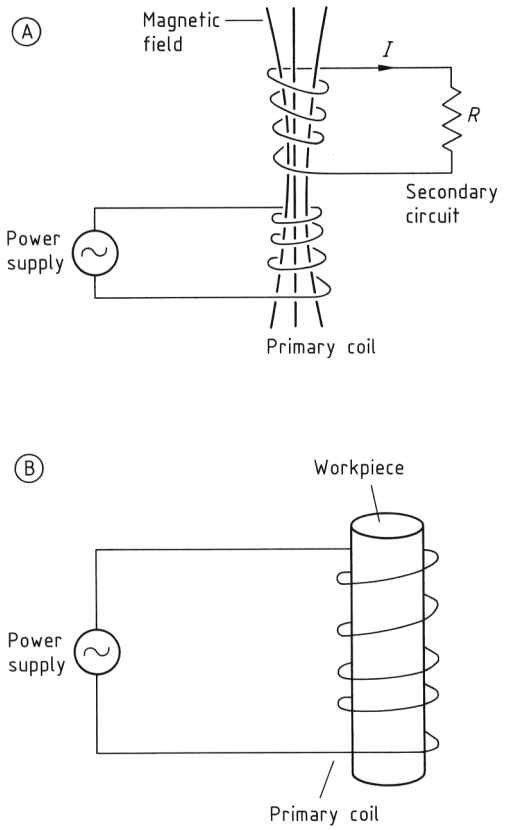

Figure 3. Direct induction heating
A) Principles of operation; B) Application to a cylindrical workpiece; *I* current produced in workpiece, *R* resistance of workpiece

An important characteristic parameter is the skin depth δ (in m) given by

$$\delta = 503\sqrt{\varrho/\mu_r f} \tag{4}$$

where f = frequency of the alternating current, Hz; ϱ = the resistivity of the material, Ωm; μ_r = the relative magnetic permeability of the material. The significance of δ lies in the fact that most of the induced current in the workpiece, and hence the power dissipation, is found within a region extending from the surface of the material to about one skin depth below. In addition, the power dissipation per unit surface area of the material is inversely proportional to δ. Two of the major objectives when designing induction heating equipment are to ensure that (a) the energy is dissipated in the desired manner, that is, surface or volumetric heating and (b) that linkage of the magnetic field to the workpiece is adequate to give efficient energy transfer. As the physical properties of the workpiece are fixed, the only available variable is the operating frequency. For efficient operation, as the diameter of the workpiece decreases, the frequency used must increase. The three frequency bands employed are:

Line frequencies	50 – 60 Hz
Medium frequency	500 – 10 000 Hz
High frequency	100 kHz – 5 MHz

Applications. Induction techniques are used extensively in metal melting and volumetric and surface heat treatment of metals to modify the physical properties, e.g., surface hardening. Localized heating is used in brazing, soft soldering, and welding, where the fast heating rates can minimize oxidation and distortion. These applications are covered thoroughly in references [5, 6].

4.2. Indirect Induction Heating of Materials

Principles. Induction heating is used to heat a metal surface, which in turn heats the product through thermal conduction. The lack of electrical contacts and the absence of a flame or products of combustion can be used to advantage to heat nonmetallic materials.

Applications. The heating of material, e.g., paint, within a large metal vessel can be achieved simply by winding the primary coil around the vessel, which acts as the secondary coil [5]. The power densities required are usually modest and line frequencies can be used. Such heaters in the range 10 – 250 kW are available and are to be preferred for power densities greater than 10 kW/m^2. Below this, resistance and steam heating are more economical. A related application is the use of induction methods for heating the surface of metal drum containers.

The rapid but controlled heating obtainable with induction heating can be applied to molding and extrusion processes. The induction heaters are built into mold platens or wrapped around the barrels of extrusion machines.

Indirect induction methods provide a convenient source of heat for direct contact drying. A

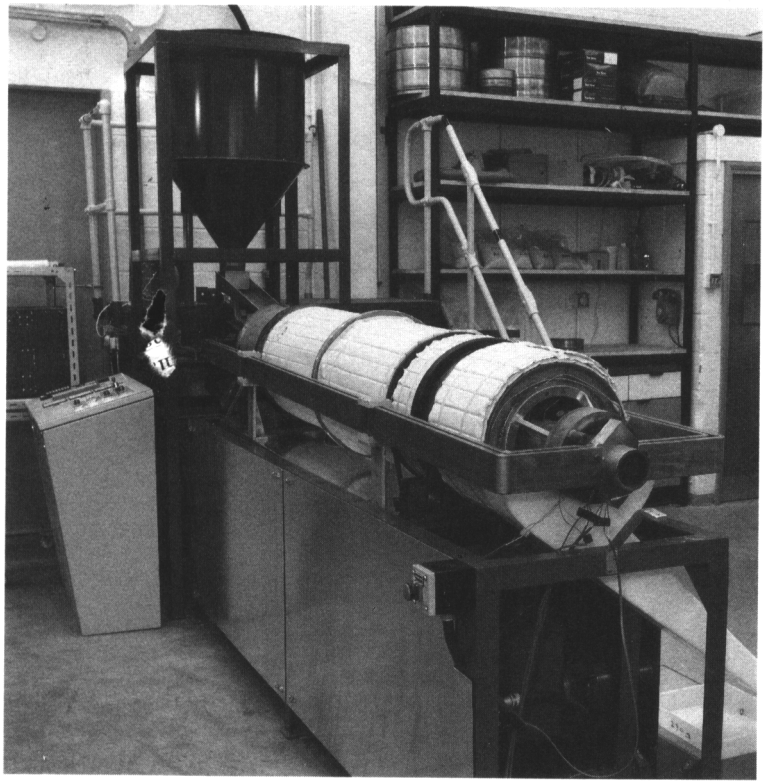

Figure 4. 100-kW rotary induction heater for contact heating of granular material

most dramatic example of this was the drying of paint on the metal surface of an entire railway carriage [6]. The attraction of drying paints and varnishes in this way is that the heat comes from below so that a surface skin does not form and volatiles can escape. The lack of electrical contacts has been exploited in a range of rotary dryers and calciners (Fig. 4) [7]. The material is transported and heated by rotating fins in the usual way. The fins themselves, however, form part of the secondary coil and are inductively heated at line frequency. Large power densities, up to 16 kW/m^2, and fine control of the fin temperature are possible, enabling temperature sensitive products to be dried.

5. Dielectric Heating of Nonmetals

Principles. Dielectric heating provides a method of contactless, direct, volumetric heating of suitable lossy dielectrics (dielectric materials capable of absorbing electromagnetic energy) and very poor electrical conductors [8]. This is achieved by the application of a time varying electric field generated either at radio frequencies (rf) or microwave frequencies.

In dielectric materials, which are nonconductors, there are no conduction currents. Instead there is a so-called displacement current, related to the time varying electric field and the dielectric properties of the material. For many dielectrics that are good electric insulators, such as some ceramics, this current gives rise to negligible heating. However, significant energy dissipation in lossy dielectrics can occur by several physical mechanisms.

Of most practical interest is the behavior of polar dielectrics: the permanent electric dipoles attempt to align themselves with the changing electric field, so gaining energy, which is then lost as heat through collisions with neighboring molecules. Maximum energy dissipation occurs at a particular resonant frequency and is directly

proportional to a dimensionless quantity, again related to the dielectric properties of the material, called the loss factor ε''. For polar liquids such as water and many solvents this resonant frequency occurs in the microwave band. For solids, the resonant frequency is lower, and the energy dissipation is usually much less since the movement of the dipoles is constrained by the solid structure.

A poor electrical conductor of particular importance is industrial water, which generally contains sufficient dissolved salts to make it weakly conducting. In addition to the displacement current, a conduction current can flow and dissipation by essentially direct resistance heating can take place. In the present context, however, the practical techniques used to apply the electric field are those of dielectric rather than resistance heating: contactless capacitive coupling at high frequencies rather than direct electrical contacts. Consequently, the effects due to any dissolved salts are usually incorporated within the framework of dielectric heating and described in terms of an effective loss factor. It is the presence of dissolved salts, in fact, which makes possible the use of rf techniques for moisture removal.

Dielectric heating equipment can only be operated at certain frequencies to ensure that there is no interference with telecommunications. These permitted frequencies vary from country to country but the most common are

13.56 MHz ± 0.06 %
27.12 MHz ± 0.6 %
896/915 MHz ± 10/13 MHz
2450 MHz ± 50 MHz

The essence of radio frequency techniques is that an rf voltage is applied across the plates of a capacitor to give an alternating electric field (Fig. 5). The product is held or passed between the plates and heated by the mechanisms described above. The physical size and shape of the capacitor can be varied to suit the particular process subject to certain constraints related to the rf energy generation. The rf power output available from typical generators ranges from a few kilowatts to ca. 100 kW, although units up to 600 kW are available. Between 60 – 80 % of the rf output power reaches the workpiece, while the efficiency of energy utilization, in terms of the energy drawn from the line supply to that dissipated in the product, is between 45 – 55 %.

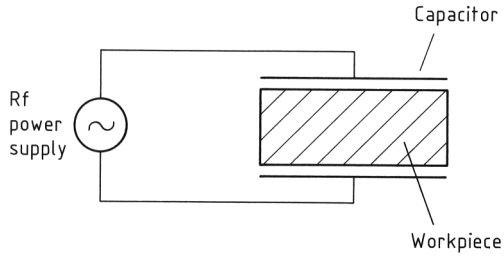

Figure 5. Principle of dielectric heating at radio frequencies

A common form of industrial microwave heater is basically an enlarged version of the familiar domestic microwave oven with product ports to allow continuous processing. The power sources, known as magnetrons, generate a microwave electromagnetic wave that excites the molecules, giving rise to energy dissipation as the wave passes through the product. Other forms of ovens, employing microwave horn antennas, single mode cavities, and slotted waveguides, are possible. At the operating frequency of 2450 MHz, the maximum power output available from a single magnetron is 6 kW with an overall efficiency (line input to product) of ca. 65 – 70 %. Improvements in magnetron design at 896/915 MHz have resulted in the availability of 60 kW magnetrons of high overall efficiency of the order of 80 %.

Applications. Despite the high capital cost of these techniques, they are widely used in certain areas, notably the textile and food industries. Their appeal lies in the direct volumetric heat transfer, which can eliminate limitations due to thermal conduction within the material and, in the case of moisture removal by evaporation, can give rise to internal evaporation and enhanced internal moisture flow. The most important consideration in existing commercial installations appears to be the improved product quality arising from this unique form of heat and mass transfer.

In the past, rf techniques have been used more for plastics welding and moisture removal applications, while microwaves have been used mainly for cooking and food processing, such as melting, thawing, and blanching, and more recently vacuum drying of high-value materials.

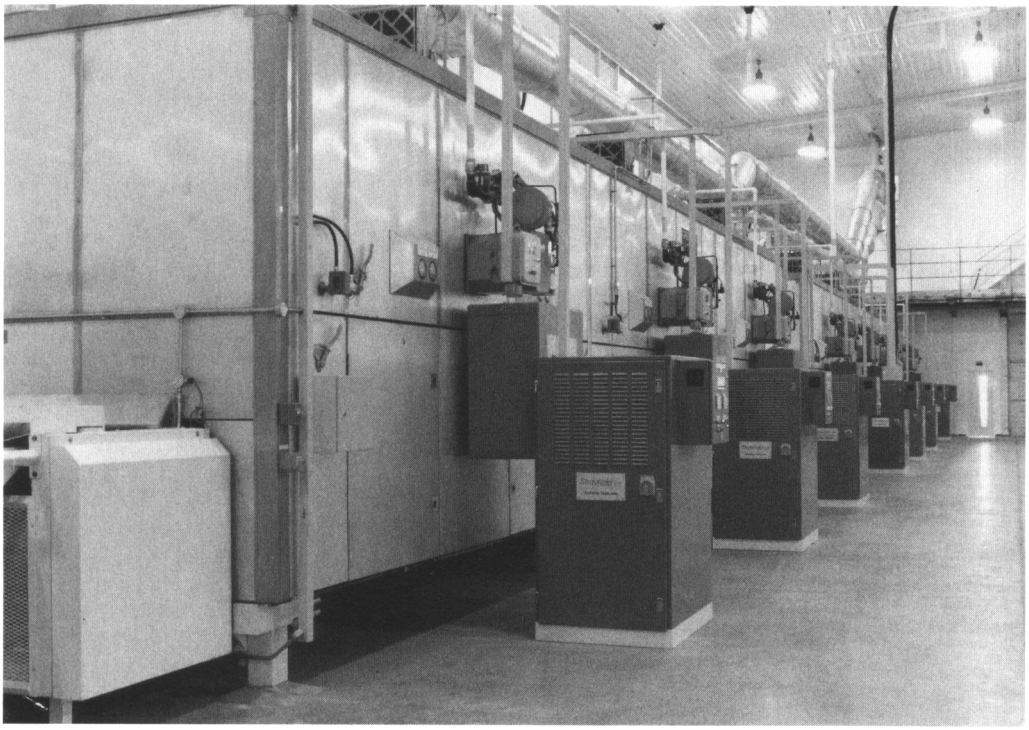

Figure 6. Radio frequency assisted convection band dryer as used in the food industry
Courtesy of Greenbank Engineering

The loss factor of wet materials increases with the moisture content. This variation is used to great effect to place the greatest amounts of heat into the initially wettest parts of the material. This leads to increased local evaporation and a leveling out of moisture profiles, giving improved product quality. This feature is used extensively in textile, paper, and pasta drying and postbaking of biscuits.

Another approach, applicable to certain capillary porous products, is to combine rf or microwave heating with convective drying (Fig. 6) [8,9]. The internal evaporation and enhanced moisture flow due to the dielectric heating are used in such a way that the convective drying remains effective for a longer proportion of the drying cycle, as compared to convective drying alone. The benefits of this approach are smaller sized equipment, improved product quality, cheaper running costs, and a more favorable capital cost, since the rf or microwave component is kept to a minimum. These factors make this combinational technique attractive to a wider range of applications where pure dielectric heating alone might not be considered owing to its high capital cost.

6. Thermal Plasma Heating

Principles. A gaseous plasma is a mixture of ions, electrons, and non-ionized neutral gas capable of conducting an electrical current when a voltage is applied across it. The characteristics of the plasma depend very much on the operating regime, especially the pressure. For thermal processing, operation at atmospheric pressure is common [10]. The kinetic energy of the electrons is increased by the influence of the applied electric field on their motion. This energy is dissipated as heat as a result of the frequent collisions between the gaseous species, which are in thermal equilibrium with each other, hence the name *thermal plasma.* Temperatures of the order of thousands of degrees centigrade can be obtained. Unlike conventional oxy – gas flames, combustion does not occur: the

heat in the plasma "flame" is derived from the electrical power source. Such plasmas may be used as a source of heat, sometimes with one of the components taking part in chemical reactions; as an indirect source of radiation and resistance heating, the classical plasma arc; and as a direct source of hot gas, the classical plasma torch. The first application, while of immense interest, is, with few exceptions, in the preindustrial stage [11].

Applications–Direct Heating with a Plasma Torch. The plasma torch consists of an anode electrode in the form of a metal tube, shaped into a nozzle at one end, and an inner cathode electrode (Fig. 7A). Gas at slightly above atmospheric pressure passes between the electrodes, is ionized to form a thermal plasma, and issues from the nozzle as a high-velocity, high-temperature plasma "flame." This basic form of operation is known as the *nontransferred* mode with the majority of the current flow between the two electrodes. The other, *transferred mode,* is applicable to metal workpieces that are used as a third electrode in such a way that the workpiece rather than the nozzle acts as the anode (Fig. 7B). In the first case the maximum amount of input energy appearing in the plasma flame is ca. 65 %, with the rest dissipated as heat at the anode. With the transferred plasma torch the electrical energy utilization is 65 – 75 %, and powers of several hundred kilowatts can be used.

Welding with a transferred plasma torch has a number of advantages over tungsten and metal inert gas techniques (TIG and MIG): greater energy densities, more stable operation, higher flame heat content and velocity. These enable faster and deeper welds to be achieved [12]. The nontransferred torch can be used to spray materials, initially in powder form, onto surfaces to form coatings. Again, due to the high temperature and velocity of the plasma, this technique is particularly attractive for high melting point materials and those which would be oxidized or contaminated with flame spraying.

Transferred plasma arcs are capable of cutting any metal, unlike oxy – fuel cutting, but up to now have been used principally for stainless steel over 6 mm thick. High cutting speeds and good quality cuts can be obtained using air or most diatomic gases. A variant of cutting is plasma machining. A nontransferred torch is used to soften hard metals in front of the cutting tool to allow faster metal removal rates in turning operations.

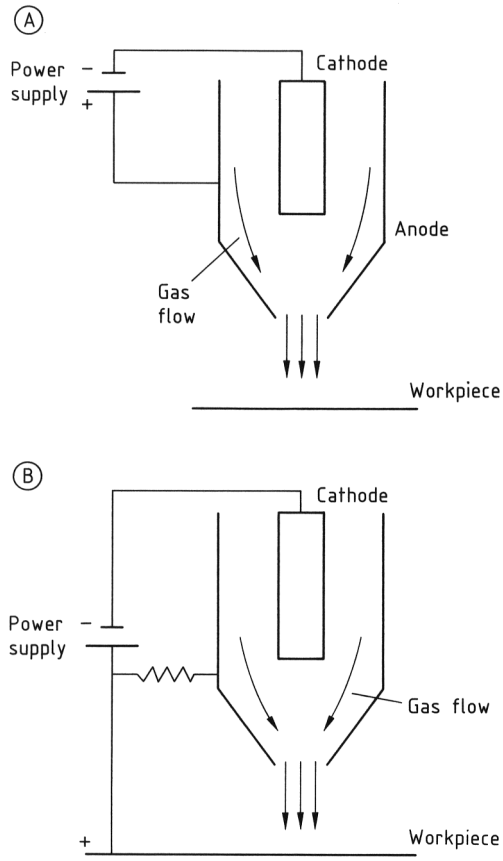

Figure 7. Principle of a plasma torch
A) Nontransferred mode; B) Transferred mode

The need for the workpiece to act as a third electrode has limited the cutting applications of the transferred arc to metals. A nontransferred cutting torch potentially extends of this method to a vast range of nonmetallic materials [13].

Applications–Indirect Heating with an Electric Arc. Direct and indirect arcs are used for metal melting and refining. In the indirect form an atmospheric plasma is struck between two electrodes and the infrared radiation obtained from the high-temperature arc is used to melt scrap metal held in a refractory-lined container below. In the direct form the metal to be

melted forms an intermediate electrode and the plasma is struck between it and the other two electrodes. Heating arises from the IR radiation as before and from the resistance heating of the metal, which occurs due to the current flowing in the intermediate metal electrode. A variation on this form of arc is where the electrodes are actually submerged in the molten metal.

7. Electron Beam Heating

Principles. In a typical electron beam gun device, electrons from a heated cathode are accelerated in a vacuum by a large voltage (20 – 150 kV) applied to the anode. The beam of electrons is directed, and for some applications finely focused, on to the workpiece, which is heated by absorbing the kinetic energy of the bombarding electrons. Typically, the vacuum required is $10^{-2} - 10^{-3}$ Pa ($10^{-4} - 10^{-5}$ Torr), the electron currents range from tens of milliamperes to amperes, and the power in the beam varies from a few kilowatts to megawatts, with power densities up to 10^9 MW/m^2 [1, pp. 745 – 782]. An alternative source of electrons can be found in the cathode region of a glow discharge operating at higher pressures (ca. 10 Pa or 0.1 Torr), an anode voltage in the range 1 – 10 kV, and power densities of the order of 1 MW/m^2. These glow discharge guns produce a broad electron beam, typically 100 mm^2 in cross section and are intended for heat treatment.

Applications. The technique generally requires the workpiece to be contained within a vacuum chamber. Partly for this reason and the high capital cost, applications with the electron gun are confined to specialist high-quality welding, and to a lesser extent cutting and drilling, in the aerospace, nuclear, electronic, and automobile industries. Welds with deep penetration, narrow width, and minimal distortion due to the reduced heat input are characteristic of these techniques. Electron beams have been used to melt or refine metals requiring high temperature and high purity. Heat treatment, annealing, and sintering using electron beams have exploited effects due to the high power densities and rapid rate of localized cooling on removal of the beam, to give auto-quenching at the surface.

Electron beams are in use for curing inks, coatings, and glues by selective energy absorption, rather than thermal effects, with systems that allow the workpiece to be handled in a normal environment rather than under vacuum.

8. Laser Heating

Principles. Lasers require electrical energy to excite the atoms or molecules of the lasing medium into a thermodynamically nonequilibrium state, from which they are induced to decay in an ordered manner, emitting a near parallel, monochromatic beam of intense light [14]. For industrial purposes, interest lies in carbon dioxide and argon ion gas lasers and solid-state neodymium glass lasers with characteristics as shown in Table 3. As with conventional IR radiation the suitability of heating a product depends on its ability to absorb radiation at the frequency of the laser.

Applications. While the high capital cost limits these techniques at present to specialized cutting, drilling, profiling, welding, and surface treatment operations, their use is growing. In some respects, laser processing can be seen as an alternative to electron beam techniques, with greater flexibility because a vacuum environment is not needed. For some cutting operations, the laser is in competition with the plasma torch, which generally has a much smaller capital cost.

Table 3. Industrial lasers for heating applications

Laser	Operating wavelength, nm	Mean power, W	Application
Argon	351 – 529	2 – 20 cw	scribing microcircuits
Neodymium	1060	20 – 1000 cw 0.5 J at 50 Hz	cutting, drilling
Carbon dioxide	10 600	500 cw 20 000 cw	nonmetal cutting welding metal, cutting

cw continuously working

9. References

General References

1. K. Kegel (ed.): *Elektrowärme, Theorie und Praxis,* Verlag W. Girardet, Essen 1974.

Specific References

2. P. Schmidt: *Electricity and Industrial Productivity,* Pergamon Press, Oxford 1984.
3. P. Skudder, C. Biss, *Chem. Eng. (Int. Ed.)* no. 433 (1987) 26 – 28.
4. M. La Toison: *Infra-red and Its Thermal Applications,* Philips Technical Library, Eindhoven 1964.
5. J. Davies, P. Simpson: *Induction Heating Handbook,* McGraw-Hill, London 1979.
6. M. G. Lozinskii: *Industrial Applications of Induction Heating,* Pergamon Press, Oxford 1969.
7. "Greater Potential for Industrial Plant Design," *Eng. Mater. Des.* **29** (1986) 13.
8. R. M. Perkin in R. J. Wakeman (ed.): *Progress in Filtration and Separation,* **vol. 3,** Elsevier, London, (1983) 205 – 206.
9. G. Swift, P. L. Jones, *Pap. Technol. Ind.* **24** (1983) no. 6, 224 – 226.
10. C. W. Chang, J. Szekely, *J. Met.* **34** (1982) no. 2, 57 – 63.
11. S. M. L. Hamblyn, *Min. Sci. Engng* **9** (1977) no. 3, 151 – 176.
12. "Progress in Plasma Technology," *Weld. Met. Fabr.* **38** (1970) 46 – 50.
13. "Plasma Jet Cutter Rivals the Laser," *Electr. Rev.* (London) 215 (1984) no. 6, 24.
14. J. E. Harry: *Industrial Lasers and Their Applications,* McGraw-Hill, London 1974.

Electric Cells → Batteries

Radiation Heating

ROBERT M. PERKIN, The Electricity Council Research Centre, Capenhurst, Chester, United Kingdom

1. **Introduction** 1
2. **Principles** 1
3. **Applications** 2
4. **References** 3

Symbols

c_1, c_2	Planck's constants
$E_{\lambda,T}$	emitted power density per unit wavelength, W/m^3
e_λ	emissivity
T	temperature, K
α_λ	absorptivity
λ	wavelength, m
σ	Stefan – Boltzmann constant, W m^{-2} K^{-4}

1. Introduction

Radiation is a broad term that includes ionizing X-rays, ultraviolet, infrared (IR), and visible light waves, radio frequency (rf) waves, and microwaves–all part of the electromagnetic spectrum shown in Figure 1. While X-rays and ultraviolet rays can be used for selective energy absorption, and rf and microwaves are used for volumetric heating, radiation heating is generally taken to mean heat transfer by thermal infrared radiation with wavelengths of a few micrometers. The principal means of obtaining such radiation is by heating an element to a temperature of the order of a thousand degrees centigrade using either a gas flame or an electric heating element (→ Electrically Generated Heat).

2. Principles

In contrast to all other forms of heat transfer, radiation heating is a two-way process: the material to be heated (the workpiece) absorbs infrared radiation from the heating element while simultaneously and independently emitting a smaller amount of radiative energy. Generally, not all the electromagnetic radiation incident on the surface of the workpiece is absorbed: some can be reflected and some transmitted through the material with further internal absorption, depending on the physical properties of the material.

A convenient concept is that of a notional body that absorbs all the radiant energy incident upon it–a so-called *black body.* Real materials can only approximate such a body but for initial engineering calculations the idea is useful. The radiative power emitted from a black body is a function of its absolute temperature T and wavelength λ. For a wavelength interval λ_1 to λ_2, this power in W/m^2 is given by

$$\int_{\lambda 1}^{\lambda_2} E_{\lambda,T}\,\mathrm{d}\lambda$$

where

$$E_{\lambda,T} = c_1\lambda^{-5} / [\exp(c_2/\lambda T) - 1] \tag{1}$$

and

$$c_1 = 3.742 \times 10^{-16}\ \mathrm{W\,m^2}$$
$$c_2 = 1.4388 \times 10^{-2}\ \mathrm{m\,K}$$

Curves of the function $E_{\lambda,T}$ and the above integral can be found in standard texts, for example [1].

The total emitted radiant power is found by integrating the emissive power $E_{\lambda,T}$ over all wavelengths to give the Stefan – Boltzmann equation

$$E_\mathrm{b} = \sigma T^4 \tag{2}$$

where $\sigma = 5.6697 \times 10^{-8}\ \mathrm{W\,m^{-2}\,K^{-4}}$

At a particular temperature, the wavelength corresponding to the peak power density is obtained from Wien's law:

$$\lambda_\mathrm{max} T = 2.8979 \times 10^{-3}\,\mathrm{mK} \tag{3}$$

For a real material, both the emissive power and the absorbed power are less than that of a

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_16

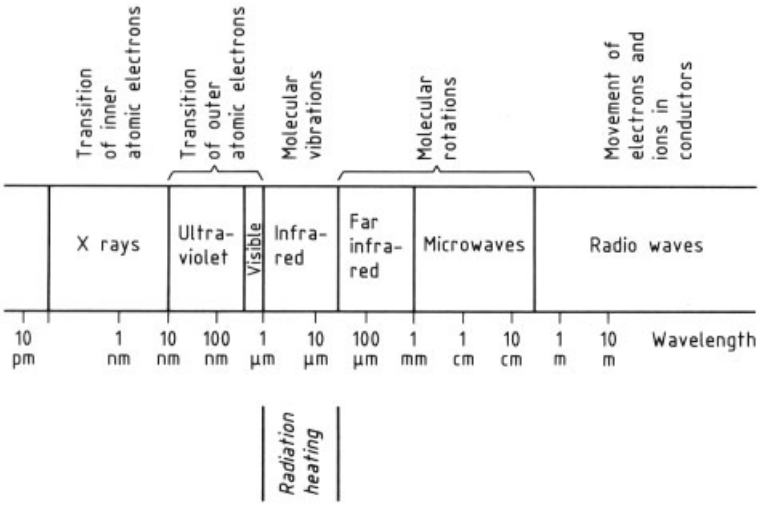

Figure 1. The electromagnetic spectrum, showing the physical mechanisms for emission and absorption of radiation over the spectrum and the range of radiation heating

black body. It is useful to express these powers as a proportion of that which would be emitted or absorbed by a black body at the same wavelength and temperature–the emissivity e_λ and absorptivity α_λ, respectively, where $e_\lambda = \alpha_\lambda$ and $e_\lambda,\ \alpha_\lambda \leq 1$. These parameters can be strong functions of wavelength and temperature, and for practical heating applications, the absorption bands of interest lie in the infrared micrometer wavelength range. The object in designing efficient heating systems is to match, as closely as practicable, the peak emissive power of the heating source with the absorptivity of the material to be heated.

In order to gain some estimate of the net heat transfer to a body, it is usual to first consider the case of two infinite, parallel black bodies at temperatures T_1 and T_2 taking, for example, T_2 to be the temperature of the heating source. Then, the net heat transfer per unit area (in W/m^2) to the body at temperature T_1 is given by

$$\sigma\left(T_2^4 - T_1^4\right) \tag{4}$$

For other shapes it is necessary to multiply the above expression by a viewing factor F, where $F \leq 1$, which accounts for the fact that not all the radiation from one body falls onto the other. Values of F for different geometries are tabulated in standard works [1,2].

Sometimes it is possible to assume that the absorptivity and emissivity are independent of wavelength–a so-called *gray body.* An expression for the net power transfer analogous to Equation (4) is available in the literature [2]. However, this concept of a gray body is of limited use for many practical applications where the wavelength dependence of the parameters cannot be ignored, e.g., in moisture removal.

3. Applications

For most purposes, absorption of infrared radiation occurs at or near the surface of the product and thereafter relies on thermal conduction to transmit the heat. The heating elements used are classified as short, medium, and long wave, with peak emissive power at approximately 1, 2.5, and 3 micrometers, respectively. The power density obtained from these elements ranges from 10 – 400 kW/m^2 as measured close to the element and diminishes with distance from the source [3]. Many materials, with the exception of metals, cannot sustain more than 15 – 20 kW/m^2 at their surface without damage [4]. Present uses lie mainly in three areas: drying and curing of paints, inks, resins, and other finishes; mass heating; and moisture removal from webs and shallow granular beds. At the moment, electric rather than gas-fired elements are more common owing to their cheaper capital cost, simpler construction, and ease of application (→ Electrically Generated Heat). The main

obstacles in the use of infrared radiation in some areas where it should be the preferred method, e.g., in the heat treatment of textile webs, have been the slow response of the emitters, that is, the time to heat up or cool down from their operating temperature, and the uneven heating obtained. Developments in both gas and electric infrared elements, combined with the use of sensors and computer control, are overcoming these limitations so that greater exploitation of these techniques should be possible.

Radioactive Labeling → Radionuclides
Radioactivity → Radionuclides
Radio-Immuno Assay → Diagnostic Reagents

4. References

1. D. R. Pitts, L. E. Sisson: *Theory and Problems of Heat Transfer,* Schaum's Outline Series, McGraw-Hill, New York 1977.
2. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960, pp. 426 – 451.
3. M. La Toison: *Infra-red and its Thermal Applications,* Philips Technical Library, Eindhoven 1964.
4. M. Bauman in P. N. Cheremisinoff (ed.): *Radio Frequency/Radiation and Plasma Processing,* Technomic Publishing Corp., Lancaster, Pa., 1985, pp. 79 – 85.

Cooling Tower Technology

Robert Burger, Burger Associates, Inc., Dallas, Texas 75 234, United States

1. **Introduction** 1
2. **Design** 2
3. **Cooling Tower Fundamentals** 2
4. **Component Parts** 5
5. **Other Evaporative Cooling Devices** . . . 9
6. **References** 9

1. Introduction

A common denominator runs through all cooling towers whether they be large units in chemical plants circulating 76 000 L/min [20 000 gallons per minute (gpm)] of water, huge electric power plant natural-draft towers cooling 380 000 L/min (100 000 gpm), or small retail-store 5-ton refrigeration units with a 60-L/min (15-gpm) circulation rate. The common denominator that cannot be ignored in all of these systems is *waste heat*.

This problem is enormous because industry uses ca. 2.0×10^{12} L (500×10^{9} gallons) of water a day, which if not recirculated would seriously deplete the world's available water supply. Unwanted excess heat from the water must be released to the environment without creating ecological havoc or adding unnecessary expenses.

Cooling towers solve the problem of possible ecological damage from returning hot water to the environment and enable industry to recycle the greater part of the water it requires at reasonable capital investment outlays. Operating procedures, costs, and upgrading potential for greater efficiency must continually be investigated to provide economic returns for the installation. To achieve this intelligently, an understanding of the fundamental principles of cooling tower performance, plus the identification and comprehension of its various elements, are necessary to be able to upgrade them.

After the cooling tower has been erected, proper operation and maintenance are required to yield the design conditions of cooling. "By not maintaining towers adequately, they can lose appreciable performance as soon as they are started up" [2].

Too many engineers pay a minimum of attention to the cooling tower and are aware only of its being that "box in the back where we send the hot water and it comes back cold." Many managers are not concerned about the role of the cooling tower in their profit and loss [3].

The major profit area to be analyzed is utilization of the cold water. In a refrigeration or air-conditioning system, the colder the water, the lower is the energy required to power the equipment because of reduced head pressures and temperatures. An additional benefit of colder water is the extension of service life due to easier running of the machinery:

The cooling tower plays one of the key roles in the efficiency of your air conditioning machinery. The efficiency of the equipment is measured in kilowatt usage. If your tower does not create the proper heat transfer, your machine will work harder to compensate for the loss in efficiency [4].

In chemical manufacturing plants, colder water yields more efficient condensation, as well as greater volume and lower cost of salable condensate product. In power plants, colder water reduces the cost of the electricity used to produce salable or usable power (the electric generation penalty). The second area of profit to be considered is the electrical or steam dollar expenditure that is purchased to power the systems. The cooling tower is a high-energy-utilization machine as well as a mass- and heat-transfer device.

With regard to cooling towers, the manager and the engineer should have two basic goals: (1) understanding their cooling tower and (2) obtaining *colder water*.

2. Design

The criteria of cooling tower performance are outlined in the *design conditions* (Fig. 1) specified when the cooling tower is purchased or rebuilt; these designate the cooling of a specific

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_17

quantity of circulating water from a particular entering temperature to a definite leaving temperature. The difference between these two temperatures is called *the range* or *delta T.* This cooling of the circulating water also must be performed at the stated wet-bulb temperature noted in the specifications. The difference between the cold discharged water and the wet-bulb temperatures is called the *approach* to the wet bulb.

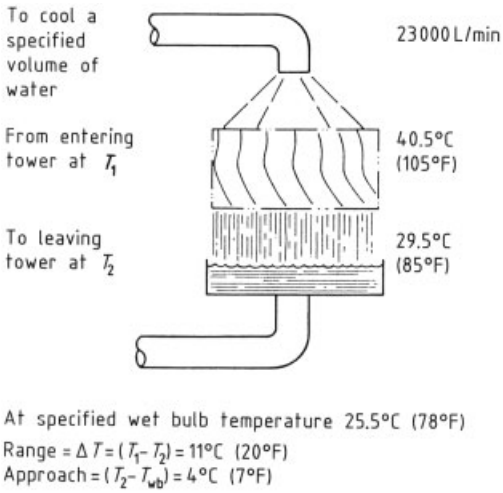

Figure 1. Design conditions, the criteria for cooling tower performance specifications

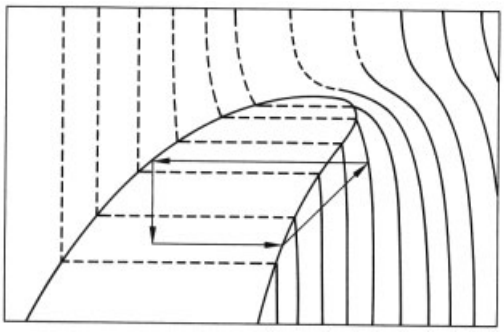

Figure 2. Pressure – enthalpy diagram: its construction, use, and value [5]

A 1.0 °C colder water returned to the compressors and condensers in air-conditioning and refrigeration equipment, calculated from enthalpy charts (Fig. 2), amounts to ca. 3 % savings in electrical energy input to these machines (1 °F corresponds to 2 % savings) [6]. Therefore, water from the tower, which is colder by a little less than 2 °C, can save more than 6 % in electrical energy and resulting charges. A 1800-tonne (2000-ton) refrigeration system circulating 23 000 L/min (6000 gpm) could use $ 350 000 in electrical power a year. A 6 % savings of $ 21 000 is a significant factor in rapid payback for upgrading the tower.

3. Cooling Tower Fundamentals

The basic principles of cooling tower operation are evaporative cooling and exchange of sensible heat. The degree of evaporative cooling is determined by the difference between ambient and wet-bulb temperatures, which takes into consideration the temperature and dryness of the air: the wet bulb is a measurement of the ambient temperature and relative humidity or the ambient temperature and dew point. In practice, these relations are obtained by the use of a psychrometric chart. A portion of the water exposed to the moving air of the atmosphere evaporates, and as the water changes to vapor, heat is absorbed which amounts to ca. 2.3×10^6 J per kilogram of water (1000 Btu per pound of water) evaporated. The heat is taken from the water that remains, thereby lowering its temperature. This is evaporative cooling, an effect based on release of the latent heat of vaporization.

However, a penalty is involved, namely, the loss of water that is discharged from the cooling tower into the atmosphere as hot water vapor. Under normal operating conditions, this amounts to ca. 2.2 % for each 10 °C (1.2 % for 10 °F) of cooling. A tremendous quantity of water is lost through evaporation, which must be added to the system via makeup water: a 38 000-L/min (10 000-gpm) industrial 10 °F tower operating 24 h per day for ca. 300 days of the year would require 1.6×10^8 L (43×10^6 gallons) of water per year. By comparison, a small 4000-L/min (1000-gpm), 333-ton refrigeration system operating 12 h a day for six months of the year would require a makeup volume of 4.9×10^6 L (1.3×10^6 gallons) of water per year.

Sensible heat that changes temperature is also responsible for part of the cooling tower's operation. Sensible heat is heat that can be felt. Sensible heat exchange is the transfer of heat from the hot water to the cold air going through the tower. In a typical case, this represents ca. 20 % of the heat exchange in the cooling tower. The

Figure 3. Discharge plume: an indication that the evaporative cooling process is working

remaining 80 %, or the hard work performed in the cooling tower, is due to the evaporative cooling effect.

Hot always goes to cold, so if water is warmer than air, the air gets hotter as it picks up sensible heat from the water, and the water is cooled as its heat is transferred to the moving air.

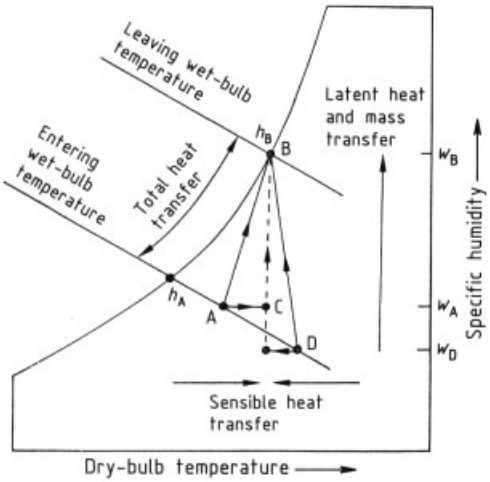

Figure 4. Psychrometric analysis of air passing through a cooling tower: sensible heat exchange and latent heat of vaporization cooling [7]
Reprinted by permission of ASHRAE

Cooling due to evaporation amounts to ca. 75 % of the heat exchange in the cooling tower, while 25 % is sensible heat transfer. This ratio varies with the wet-bulb temperature, which is a measure of the evaporative capability available at a particular ambient temperature and dew point or relative humidity. Discharged moisture is visible as a plume or fog (Fig. 3) when the ambient temperature and dew point rise to 100 % saturation. Below that point, moisture is still being discharged, but it is invisible. Fogging can be predicted by drawing a straight line on a psychrometric chart (Fig. 4) from the tower's entering conditions to a point representing the discharge conditions. The greater the area of intersection to the left of the 100 % saturation curve, the more intense is the plume.

A cooling tower is "the heat rejection solution of the chemical process" or "correction of the heat penalty generation of compression equipment." The cooling tower is a machine that moves heat from point A to point B and ultimately discharges this heat into the atmosphere, which thermal engineers call a heat sink.

Psychrometric Analysis. The thermal performance of a cooling tower is affected by the entering wet-bulb temperature (WBT). The wet bulb is an important measurement of cooling tower performance. It determines the evaporative capacity available in the ambient temperature: 100 % wet bulb is 100 % saturation, i.e., fog or rain. The lower the wet bulb, the better the cooling tower functions. No cooling tower can cool water below the wet-bulb temperature.

In 1921, Merkel, a German physicist, developed the integration method of analyzing the driving force of a cooling tower, i.e., the difference in enthalpy between air saturated at the water temperature and the main airstream:

$$\frac{KaV}{L} = Me/c = \int_{T_2}^{T_1} \frac{\mathrm{d}T}{H'' - h}$$

where

$\frac{KaV}{L}$ = tower characteristic
Me = Merkel number

c = specific heat of water
T = bulk water temperature
T_1 = temperature in
T_2 = temperature out
H'' = air – water enthalpy at bulk water temperature
h = air – water enthalpy at wet-bulb temperature

The KaV/L, or cooling tower characteristic, is a measure of the order of difficulty for the cooling tower heat rejection requirements, independent of the fill performance. It is plotted logarithmically as the ordinate (vertical axis) on a design characteristic curve, with the liquid-to-gas (L/G) ratio as the abscissa (horizontal axis).

As an example, statements such as "this tower is not operating efficiently because it was designed for a 20 °F range and is only cooling an 18 °F range" and "this tower is doing quite well because it is rated for a 20 °F range and is cooling at a 22 °F delta T" may both be incorrect because neither statement mentions the WBT.

In the psychrometric schematic, Figure 4, when air enters at ambient point A, it absorbs heat and moisture from the water, exiting at point B in a saturated condition. The amount of heat transferred from the water to the air is proportional to the difference in the enthalpy of the air between the entering and leaving conditions ($h_B - h_A$). Vector AB may be separated into component AC, which represents sensible heat exchange, and component CB, which represents latent heat exchange.

If, at a higher entering air dry bulb, the condition is point D at the same wet-bulb temperature, but a different dry-bulb temperature, the total heat transfer remains the same, but the sensible and latent components have changed. The amount of evaporation depends upon the amount of sensible heating or cooling of the air, which in turn depends on the dew point.

The ratio of latent to sensible heat is important in analyzing the water usage of a cooling tower. Mass transfer (evaporation) occurs only in the latent portion of the heat-transfer process and is proportional to the change in specific humidity. Because the dry-bulb temperature or relative humidity of the entering air affects the latent : sensible heat-transfer ratio, it also affects the rate of evaporation. The rate of evaporation is less for vector AB ($w_B - w_A$) than for vector DB ($w_B - w_D$) because the latent heat transfer (mass transfer) represents a smaller portion of the total. At higher entering wet-bulb temperatures, the evaporative components are reduced because the specific humidity of the air is greater with less potential for evaporative cooling.

Counterflow vs. Crossflow. Cooling towers are available in two basic operational designs, counterflow and crossflow (Fig. 5). The main difference between them is that in counterflow towers, the air moves vertically through the fill, while in crossflow towers, it flows horizontally through the fill. In both cases, the water drops vertically through the fill after leaving the nozzle or water distribution area. In counterflow, the air and water mix at 180°, or counter to each other. In crossflow, the air – water mix is at 90°, or cross to each other.

4. Component Parts

All cooling towers, except small, packaged assembly-line air-conditioning models, are custom-made, and the manufacturer varies the shape, size, configuration, and input to meet the particular set of thermal parameters or design conditions required by the customer.

In both crossflow and counterflow towers, uniformity is the key to success: uniformity of water distribution, of static pressure, of fill configuration, and of air velocity (even though in a crossflow tower, the pressure differential will vary with the height). After installation, any operational deficiencies, such as clogged or broken nozzles and arms, broken or out-of-position decking fill, or drift eliminators with sections missing, will seriously affect optimization of the cooling tower's performance.

Wet Decking Fill. Generally the most significant improvements in thermal performance can be made simply by changing the wet decking fill. This, however, may not be done capriciously. Heat transfer must be investigated from a thermal engineering point of view, in conjunction with the fill characteristics. The manufacturer's performance curves are developed painstakingly by trial, error, and experimentation and are expressed as KaV/L or the heat-transfer characteristic [7]. Precise thermal, hydraulic, and

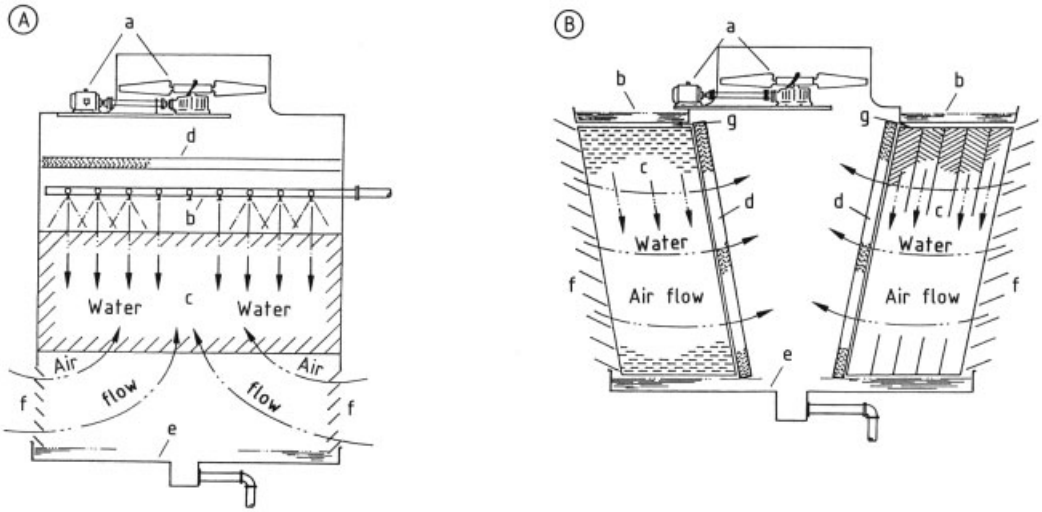

Figure 5. Cooling tower schematic configurations A) Counterflow; B) Crossflow
a) Mechanical equipment; b) Water distribution; c) Fill packing; d) Drift eliminators; e) Cold-water basin; f) Air inlet louvers; g) Redistribution area (only in B)

mechanical calculations should be performed to ascertain the maximum cold water available to optimize each individual design condition requirement [8, Chap. 9].

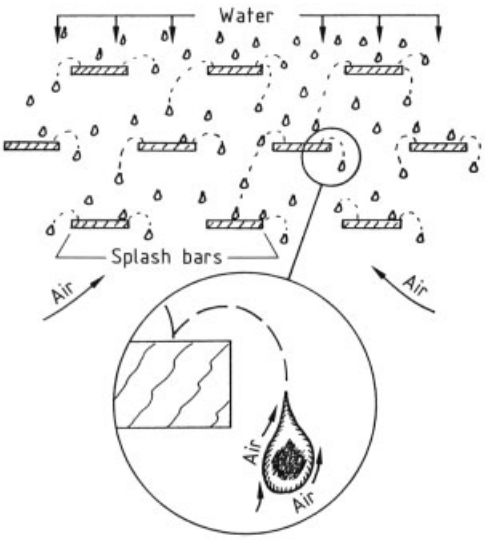

Figure 6. Splash bar cooling

Older, conventional fill consists of wooden slats. In wood slat splash bars (Fig. 6), droplets of water bounce from one layer of wood to the next, and rising air cools the outside of each sphere of the water droplets. Usually, however, the interior of the droplet retains some residual heat.

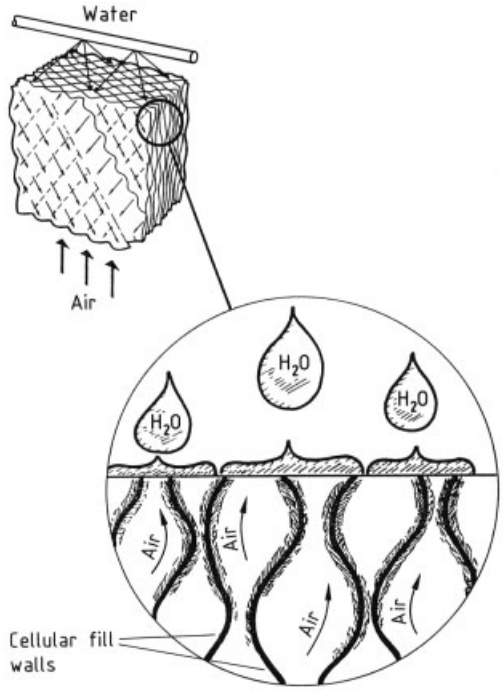

Figure 7. Cellular fill cooling

Modern state-of-the-art cellular fill (Fig. 7) takes the same droplet of water and spreads it

out in a very thin film, where the air can affect the entire surface of the film. Considerably more surface is available to the flowing air for evaporative cooling and sensible heat exchange to occur. The cellular fill pack contains more surface area than splash bars. Because the design of cellular fill permits air to flow through it with less total static pressure, it is extremely efficient, compared to the old-fashioned splash bar mixture system, for the same parameters of cooling.

For *crossflow towers* up to ca. 6 m (20 ft) in height, installation of cellular fill packing is economical. This provides the advantage of a high rate of cooling through the interaction of film fill and cooling air. In crossflow towers higher than 6 m (20 ft), the economic advantage and water loading per square foot favor the new shaped PVC [poly(vinyl chloride)] bars, which can be retrofitted to obtain higher levels of cooling. These bars are suspended on galvanized, bonderized, and PVC-coated wires, and several shapes are available from various manufacturers (Fig. 8). In changing any type of fill, whether cellular or shaped PVC splash bar, the fill manufacturer should be asked to provide thermal performance curves for their materials that have been tested and verified.

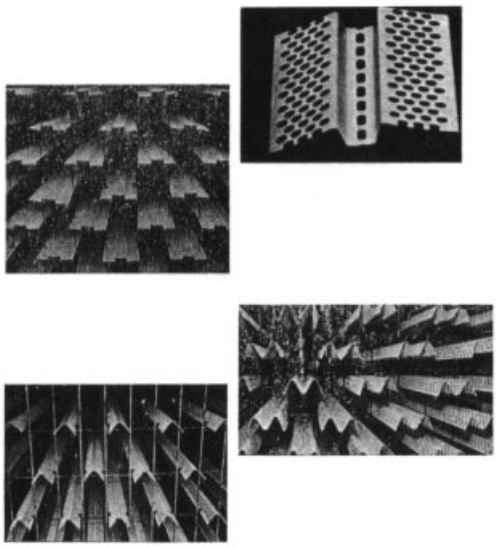

Figure 8. Shaped PVC fill bars designed for large crossflow cooling towers

Water Distribution System. For maximum performance, the water distribution system must provide a uniform pattern over the fill for optimum air – water interface. Crossflow hot water is distributed through holes (metered orifices) that drop the column of water on to splash grids. When a nozzle is clogged, it leaves a dry spot on the fill, and air follows the path of least resistance, rushing up this dry spot, thus wasting energy and cooling potential. This splash column distribution is typical of crossflow towers. A new modification to help develop a more uniform pattern involves a target nozzle with a built-in splash device (Fig. 9).

Figure 9. Target nozzles with splash devices to assist in developing a spray pattern

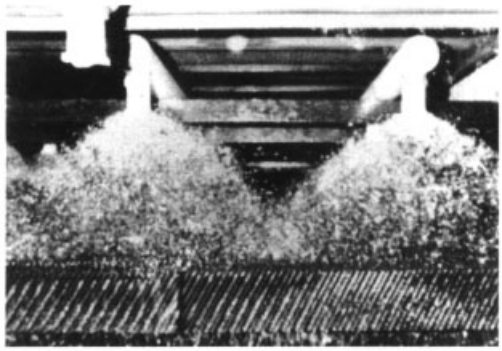

Figure 10. Modern PVC piping and nonclog square spray nozzle

In counterflow towers, rust and corrosion, typical of steel piping, cause breakage, with nozzles falling off. Wooden stave piping is also subject to breaking. A modern answer to the problem of deteriorated steel or wood piping is the use of PVC or fiberglass materials in conjunction with the ABS (acrylonitrile – butadiene – styrene) plastic square-pattern spray nozzle, which was de-

signed specifically for cooling towers and is practically clog-free (Fig. 10).

Drift Eliminators. A baffle, or drift eliminator, is placed between the water distribution system and the air discharge point to minimize dispersal of water droplets entrained in the airstream, into the surrounding atmosphere. Drift eliminator efficiency is an essential element of cooling tower design for many reasons: (1) conservation of water, (2) retention of chemical treatment in the cooling water, (3) prevention of staining by chemical additives such as chromates, and (4) compliance with local environmental-protection regulations.

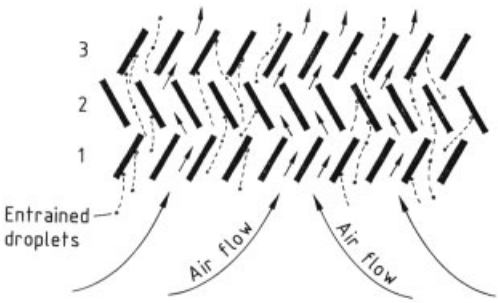

Figure 11. Schematic three-pass wooden drift eliminator slats

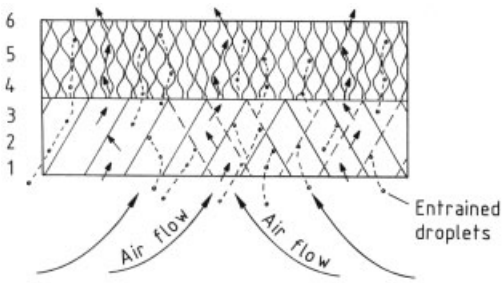

Figure 12. Schematic six-pass cellular PVC drift eliminator

Cooling tower performance, i.e., heat removal, is a balance between water flow and air volume; therefore, drift eliminators are normally designed to be efficient through a calculated range of air flow. Too great an air speed can result in excessive drifting of water from the tower, while poorly designed eliminators increase static pressure through the tower, thereby slowing the air flow and lowering the cooling capability and efficiency of the unit.

An efficient mechanical system for retaining water is necessary because the spray or water distribution system turns the heavy stream of condenser or process water into light droplets preparatory to cooling by the airstream. As the air moves counterflow or crossflow to the water, it picks up some mist or droplets and can carry them with the air flow out of the tower.

The schematic operation of a three-pass drift eliminator is shown in Figure 11. A newer, more efficient drift eliminator fabricated from self-extinguishing PVC has been developed. It causes the air to gently change direction approximately six times, thereby obtaining greater surface contact to release water droplets. Figure 12 illustrates the schematic mechanics of the six-pass cellular drift eliminator.

A common way of attempting to improve cooling tower performance is to increase air velocity by rotating the fan with more power and higher pitch angle to the blades. This often causes increased drift loss because the eliminators were not designed for this type of service. Cellular drift eliminators that do not increase the static pressure but do increase the drift-collecting capability can minimize such an eventuality. In fact, reduced static pressure resistance allows more air to flow through the tower at the same power expenditure and permits more cooling at the same cost.

Complete drift elimination is possible in theory, but impractical in application. A generally satisfactory level is less than 0.02 % of the water volume.

Crossflow towers are often susceptible to drift problems because of a tendency toward higher air flow velocities in this type of design at the top of the tower. Cellular drift eliminators are effective up to 220 m/min (720 ft/min), which is more than adequate for most commercial towers. Because the drift requirements are more stringent, a different type of eliminator system is used, with a very delicate balance design and performance. If one or two eliminator blades are out of position, drift occurs quite readily. Several manufacturers have developed cellular eliminators to overcome the problem. Here again, the lower resistance of these new eliminators, plus their drift elimination effectiveness, make them a desirable conversion from the less efficient wood.

Retrofitting of an Ammonia Plant. Output of anhydrous ammonia was to be increased from 120 000 to 225 000 t/a. The cooling tower structure, fabricated of old-growth clear-heart redwood, was about 25 years old. It contained eight cells, each with a 6.1-m (20-ft) diameter fan driven by a 56-kW (75-hp) motor. It was designed to cool 136 m^3 of water per minute, entering at 46 °C and leaving at 29.4 °C, with a wet-bulb temperature of 25.5 °C. This produced a 17 °C (30 °F) cooling range with a 4 °C (7 °F) approach.

Because the tower was overdesigned initially, doubling anhydrous ammonia output required that cooling-water circulation be increased only to 161.25 m^3/min.

Addition of a ninth cell was strongly recommended (crossflow configuration, 7.5-kW fan). However, the project engineer decided to investigate the matter more closely. An inspection revealed that the tower's structure, connectors, and mechanical equipment were in excellent condition. Thermal calculations indicated that, with the installation of 0.61 m of cellular fill at the bottom of the tower, the existing eight cells could, with no additional horsepower, easily cool 161.25 m^3 of water per minute.

Furthermore, the new fill could be installed so as to allow more fill to be added at a later date to again upgrade the tower's capacity. To improve water distribution, the 25-year-old spray-nozzle system was replaced by nonclogging ABS nozzles. Because the original polystyrene fill was in relatively good condition, it was not removed. Instead, the existing wet-decking fill was augmented by the addition of a 0.61-m layer of high-efficiency PVC film fill.

Savings from rebuilding the existing eight cells instead of adding a new cell amounted to ca. $ 100 000. However, of greater significance were the ongoing savings resulting from the fact that 75 kW (100 hp) was not required to operate a ninth cell. With electricity at 7¢/kW h, operating a fan motor for 8760 h would cost $ 55 869 annually. Over the projected ten-year life of the tower, if estimated escalations in the cost of utilities are considered, more than $ 1×10^6 in operating expense would be saved. Thus, upgrading the tower was clearly more economical than adding to it [9].

Maintenance. Like any other piece of mechanical equipment, a cooling tower requires maintenance to be productive. Over the years, the elements do deteriorate. They should be inspected periodically – at least every two years – by a qualified cooling tower specialist, who will analyze the thermal operation as well as the mechanical, structural, and physical condition of the tower. In a written report, recommendations will then be made for repair and rebuilding specifications, together with an estimate of their cost.

Economic Aspects. Cooling towers producing colder water earn money for a facility by keeping head pressures and temperatures of compressors low, yielding more condensate product by the use of colder condensing temperatures off the tower, and reducing the electric penalty of generating equipment. The cooling tower, while appearing to be a simple box, is an important mechanism in the refrigeration chain, chemical plant process operation, and electric generating facility. It demands professional maintenance and operational understanding, to maximize its usefulness and profit-making potential.

5. Other Evaporative Cooling Devices

Natural-draft cooling towers have been a familiar sight in Europe for many years since they were first developed in Germany. Modern natural-draft towers are generally constructed of reinforced concrete. The shape of the tower has changed over the years. Early towers were cylindrical; later, a design using two truncated cones was employed. The current shape is hyperbolic, giving the tower superior strength and an ability to match the natural flow of air through the tower shell.

Natural-draft towers do not use fans to move the air: the flow of air through the tower is created by the density difference between atmospheric air and air inside the tower, which is warmed as the water is cooled. Hence, the tower is little more than a chimney that draws cool air in at the base and exhausts warm air at the top. The natural-draft tower is employed when a large

quantity of water, from 380 to >1320 m^3/min (100 000 to >350 000 gpm), must be cooled in a small space.

The drawback of the natural-draft cooling tower is that the approach to the wet bulb is ca. 8 °C (13 – 15 °F). Mechanical-draft towers can reduce the approach to 4.5 °C (7 – 9 °F), which is more desirable from an economic view point.

A modern development of the large concrete shell tower is a truncated cylinder ca. 18.3 m (60 ft) high, usually with eight cells of 9.15-m (30-ft) diameter fans and 200-hp motors. In the United States, few plans exist for building these natural-draft giants (100 – 150 m (300 – 500 ft) diameter and height) because of the efficiency and low-temperature water provided by mechanical-draft concrete towers.

Cooling ponds are rarely used these days because they require a large expanse of water. The cost of real estate makes this prohibitive. Also, in the United States, environmentalists protest when water from cooling towers is discharged into lakes, ponds, and streams, because it raises the temperature of the water and is apparently detrimental to aquatic life.

Cooling Water → Water
Copolymers → Plastics, General Survey

6. References

General References

1. R. Burger: "Conservation Strategies for Cooling Towers," *CEP Chem. Eng. Prog.* **79** (1983) no. 12, 33 – 38.

Specific References

2. J. Willa: " Cooling Tower Operations," Cooling Tower Institute Proceedings, January, 1980.
3. B. D. Stanton: "Control Schemes for Distillation Towers," *Chem. Eng. N.Y.* **93** (1986) no. 22, 87.
4. C. Weiss: "Cooling Tower Inspection and Economics," *Nat. Eng.* **89** (1985) no. 3.
5. *ASHRAE Equipment Handbook,* American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, Ga., 1983.
6. *The Pressure Enthalpy Diagram, Its Construction, Use, and Value,* Allied-Signal Corp., Morristown, N.J., 1972.
7. *Cooling Tower Institute Performance Curves, Cooling Tower Theory,* Cooling Tower Institute, Spring, Texas, 1967.
8. R. Burger: *Cooling Tower Technology,* Burger Associates, Inc., Dallas 1979, revised 1987.
9. R. Burger: "Retrofitting Cuts Cooling-Tower Costs," *Chem. Eng. N.Y.* **93** (1986) no. 16, 117 – 119; "Koelwerken," *PT Procestechniek* **39** (1984) no. 9, 39 – 42.

Direct Heating with Circulating Heat Carriers

Hans-Jürgen Weiss, Lurgi GmbH, Frankfurt, Federal Republic of Germany

1. **Introduction** 1
2. **Direct Heating with Circulating Gas** . 1
3. **Direct Heating with Circulating Solid** 2
3.1. **Fundamentals** 2
3.2. **Applications** 4
4. **References** 5

1. Introduction

A gaseous, liquid, or solid substance may be heated by direct contact between the substance and a heat carrier. This process is called direct heating. The physical state of the heat carrier is of secondary importance; however, industrial processes generally use gaseous or granular solid heat carriers. Liquid heat carriers are limited by the narrow range between their melting and boiling points. For this reason they have found no major application to date.

Heat transfer is usually a cyclic process: after the heat has been transferred, the heat carrier is purified of any gases, vapors, or solids that may have been taken up in the course of the process, is then reheated, and finally recycled to the process.

2. Direct Heating with Circulating Gas

The circulating gas flows through the substance to be heated. Mass transfer may take place during this process. This form of direct heating is especially effective for lumpy or granular solids. Below the range 500 – 600 °C, heat transfer is achieved almost exclusively by convection; above the range 800 – 900 °C, radiation becomes significant, although this is affected by the contents of water vapor and carbon dioxide in the circulating gas. Above 900 °C, radiation becomes even more important as the temperature is increased.

The heat carrier flow may be cocurrent, countercurrent, or cross-current in respect to the substance to be heated. Cocurrent flow permits high initial heat carrier temperatures and avoids an excessive final temperature of the substance. For the same heat-flux density, countercurrent flow gives lower local temperature differences, and the product is heated to its final temperature more gently. Countercurrent heating is thermally more efficient than cocurrent heating because the temperature of the substance to be heated can be raised above the temperature of the exiting heat carrier. Cross-current heating is used less frequently than the other two possibilities; however, it is used occasionally, e.g., in the case of a travelling grate heater.

The term quasi-cocurrent flow is used to describe heat transfer in a fluidized bed. The good heat exchange results in an almost uniform bed temperature, the gas and the solid leaving nearly with the same temperature. The unfavorable course of the temperature difference between the circulating gas and the solid to be heated is a significant drawback to an otherwise advantageous process.

The composition of the circulating gas depends on the temperature range. At low temperature, air is generally used, while gases containing little or no oxygen–frequently combustion gases–are used at high temperatures. Often the gases or vapors released upon heating the substance, most often steam, are used as the heat carrier.

The simplest method of heating the circulating gas is to mix it with hot combustion gas. The furnace fuel may be coal, oil, or gas–depending on the material to be heated and on the possibility of corrosion, odor, etc., caused by the presence of oxygen, sulfur, tar vapors, and solids in the circulating gas. Oil and gas are ideal fuels in this respect as they are easy to control. If combustion gas cannot be allowed to come into contact with the substance to be heated or if the gases or vapors released from the heated substance should not be mixed with combustion gases, the circu-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_18

lating gas is heated indirectly in heat exchangers or heated directly by solids circulating in a second heat carrier loop.

A part of the circulating gas is continuously removed after each cycle to offset the amount of combustion gas added and the amount of gases and vapor released from the substance heated. Often, valuable byproducts can be separated by condensation from part or all of the circulating gas and withdrawn in liquid form. If dust is present, dust separators are used before the gas is cooled.

Applications. Circulating gases can be used in a number of ways to heat solids directly.

1) Gentle drying of substances to quickly remove vapors generated and to prevent sweating. Example: food drying with air recycle.
2) Cocurrent drying of substances in an entrained stream with hot flue gases containing little oxygen where the flue gas temperature is above the ignition temperature of the substance. Examples: flash dryer for coal (Babcock BSH), mill dryer for coal (Lurgi).
3) Steam drying of substances that are extremely sensitive to oxygen or that release hydrocarbons in addition to water vapor during drying. The hydrocarbons are recovered by partial condensation of the steam. Examples: rotary kiln steam drying of coal (Ahlsfelder Eisenwerke), steam flash drying of tar sands and oil-containing diatomaceous earth (Lurgi).
4) Carbonization/retorting of lumpy coal and oil shale. Example: Lurgi *Spülgas* retort, Unocal oil shale retort.

3. Direct Heating with Circulating Solid

3.1. Fundamentals

Frequently, a fine-granular solid is used as the heat carrier for thermal upgrading of solid, liquid, or gaseous fuels (retorting, coking, cracking) in which the emphasis is on the production of liquid and/or gaseous hydrocarbons. This process offers a number of advantages:

1) The vapors formed can be recovered undiluted by heat-carrier gases.
2) Cooling and condensation of the products is more economical because of the absence of heat-carrier gases.
3) The required heat-carrier volume is much smaller because the heat capacity per volume of solid is much greater than that per volume of gas.
4) Extremely short reaction times are achieved in the retorting of solid fuels such as coal, oil shale, or tar sands due to the high heat-transfer rates obtained by mixing the fine-grained heat carrier with the fine-grained feedstock. This normally leads to high yields of liquid hydrocarbons, and to relatively small reactor volumes.

A disadvantage of solid heat carriers is that their circulation system is generally more expensive than that for gases. In most cases, the solid reaction residue, the coke, spent shale, or spent sand is used as solid heat carrier. If the heat carrier is also to have a catalytic effect, then fine granular catalysts may be used as the heat carriers. Non-process-derived heat carriers without catalytic effect are of minor importance.

The coke contained in or deposited on the solid heat carrier is normally burnt with air at a controlled rate in a fluidized–or entrained–bed system. If the coke alone is not sufficient for reheating, a solid, liquid, or gaseous fuel must be added. Fluidized beds, mixers, and rotary kilns are used as reactors.

A temperature of ca. 1000 °C has been found to be the upper limit for the temperature of the solid heat carrier. The presence of small amounts of low-melting substances, e.g., salts, in the heat carrier, the substance to be heated, or the fuel may cause adhesion or sintering, which hinders recirculation.

In addition, processes recycling fine granular heat carriers have proved useful at lower temperatures because of their technological advantages. A disadvantage, however, is the necessity of equipment to remove the dust from the reaction and combustion gases. Other disadvantages include the abrasion of the internal lining of the circulation system and the loss of heat carrier due to abrasion. Because of the high investment costs such processes are generally economical only for large units. A schematic drawing of a

circulating solid heat carrier system is shown in Figure 1.

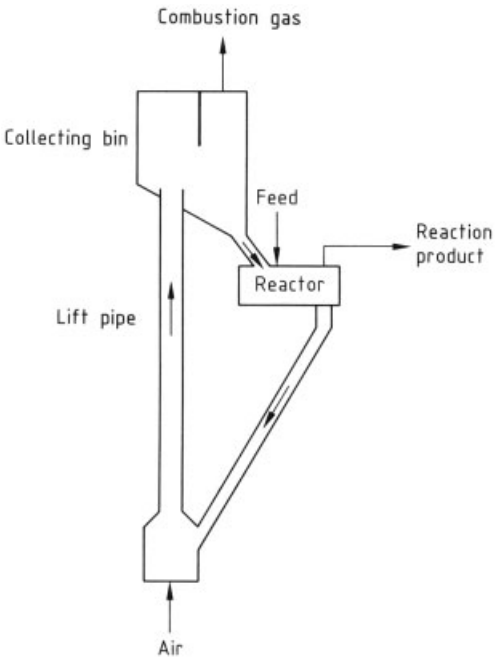

Figure 1. Lurgi – Ruhrgas process (LR process)

The fine-grained heat carrier is heated by burning its carbon content with air in a vertical lift pipe that simultaneously conveys it upward. The heat carrier is then separated from the combustion gases in a collecting bin and fed into the reactor at a controlled rate. After heat transfer, the cooled, coke-laden heat carrier is recycled to the lift pipe and reheated (regenerated), thus closing the loop. Heating and regeneration may also be carried out in a fluidized bed. In this case, however, an additional pneumatic transport device is required for the heat carrier.

Heat exchange between the hot gas and the solid heat carrier in the reactor may take place either in a countercurrent flow (when the solids travel through the reaction zone in a moving bed) or a quasi-cocurrent flow (when the solids are fluidized). For fine-granular materials, a fluidized bed or mixer is preferred as the reactor, primarily because of the favorable heat exchange between the reagent and the solid heat carrier. In addition, the throughput per unit cross section of the reactor reaches its optimum for a mixer. Another conceivable system is heat exchange between gas and a free-falling solid.

Numerous investigations [1–8] of heat and mass transfer between a gas and particle surfaces in a packed bed are available for various particle geometries. Empirical equations were derived for specific boundary conditions. The most comprehensive presentation was that of Jaeschke [9], who investigated packed beds with a broad range of void volumes and a variety of geometric forms. He presented his results in the form of a working diagram. His most important results are that transfer behavior depends strongly on the average specific void volume and on the effective flow velocity prevailing in the average void cross section.

For a bed made up of spheres with a specific void volume of ca. 0.4, the exchange is significantly greater than for a single sphere, under otherwise comparable conditions. The difference in transfer coefficients decreases with increasing void volume, becoming negligible for void volumes above 0.95.

For a fluidized bed, the average void volume depends primarily on the flow velocity. Jaeschke's results for particle beds of various degrees of loosening allow a comparison of convective transfer between nonfluidized and fluidized beds. According to Mosberger [10], along with the results of others [11, 12], there is no difference between a nonfluidized and a fluidized bed with respect to heat- and mass-transfer coefficients [13], for identical average void ratios under comparable conditions. Neither the intrinsic movement of the particles nor the frequently observed bumps and bubbles in a fluidized bed have a distinctly recognizable effect on heat and mass transfer. This is true for spherical and cubic particles with diameters between 0.7 and 9 mm. In fluidized beds of significantly smaller particles, different flow conditions, such as channelling, can be expected. In addition, an exact determination of particle surfaces becomes problematic. If heat- and mass-transfer values are known, then the residence time of the solid in the reactor, and thus the bed height, may be calculated. The reactor cross section is calculated from the gas flow rate and the permissible gas velocity.

If heating the reagent causes it to undergo a significant endothermal or exothermal reaction,

then bench-scale tests should be made to obtain information on the necessary solids flow rate.

The most favorable grain size, varying from fine dust to particles with 5-mm diameter, depends on the process conditions. Diameters between 0.1 and 3 mm are generally preferred. The void volume should be between 36 and 40 % for packed beds and between 50 and 80 % for fluidized beds. During operation, the particles assume a near spherical shape as a result of abrasion.

The number of spheres or exchange surface for arbitrary void ratios ε ($0 < \varepsilon < 1$) may be obtained by multiplication of the figures in Table 1 by the factor

$$\frac{1-\varepsilon}{0.6}$$

Owing to the large surface of granular solids (see Table 1) and to the favorable heat- and mass-transfer coefficients of packed and fluidized beds, the reaction volume required is relatively small. The effective gas veloctiy in moving bed reactors is smaller than the fluidizing threshold, whereas in fluidized beds, it is two- to sixfold greater.

Table 1. Number of spheres and heat-exchange surface per cubic meter of bed material as a function of sphere diameter for a void ratio $\varepsilon = 0.4$

	Spherical diameter, mm			
	1	2	5	10
Number of spheres, 10^6 m^{-3}	1147	143	9.17	1.15
Heat exchange surface, m^{-1}	3600	1800	720	360

Heat transfer between fine-grained solids in a mixed bed has been investigated relatively little. Available conclusions [14, 15] are that heat transfer from one fine-grained solid to another depends heavily on the avarage particle size and the mixing quality of the reactor. Nearly complete heat transfer in fluidized beds and mechanical agitators with average particle diameters of 0.5 – 1 mm is reached within seconds. In such systems heat is transferred partly by direct grain-to-grain contact, partly by convection of the gas in the void, and partly by radiation from solids and gases.

Theoretical evaluation is difficult because all these processes are mutually superposed. Most investigations involving fine-grained solids are empirical; however, they are generally adequate for practical purposes.

A solid heat-carrier should have the following properties:

1) High specific heat, to keep the circulating mass small
2) Good compressive strength at high temperatures
3) Low abrasive wear, even in the case of sharp temperature changes
4) Low abrasive action on the refractory linings
5) High thermal conductivity
6) Chemical resistance to the combustion gases and the substances to be heated
7) Catalytic action in special cases

The residence time of the solid in the regenerator and reactor may be only a few seconds when fine grained solids are used. Stability of the solids to thermal shock is of crucial importance in this case.

3.2. Applications

The fundamentals of heat transfer by means of circulating granular solids or "hot stones" was reported in German publications as early as 1927 [16–18]. The fluid coking process developed by Exxon for heavy oils or residual oils uses a fluidized bed of hot, fine-grained petroleum coke into which the oil is injected. In a second fluidized bed, the circulated coke is regenerated and reheated by partial combustion with air. The petroleum coke itself is formed in the first fluidized bed as a coking residue [19, 20]. Around 1940, heat transfer by circulating solids was introduced in the USA for the catalytic cracking of crude oils [21, 22]. In this case, the solid heat carrier is a catalyst.

The LR process for retoring fine granular coal was developed by Lurgi and Ruhrgas around 1950. It uses a mixer – reactor in which the feed material is intimately mixed with hot, circulated coke and retorted (devolatilized) within a few seconds [23–25]. The coke is then reheated in a pneumatic lift pipe by partial combustion with air as the coke is conveyed upwards into the collecting bin. In the collecting bin, it is separated from the combustion gases and returned to the mixer – reactor.

Since then, the LR process has also been used to produce olefins from various crude oil fractions with sand as the heat carrier (sand cracker) to retort oil shale and tar sand, and to coke residue oils (asphaltenes, hydrogenation residues, etc.)

4. References

1. J. De Acetis, G. Thodos, *Ind. Eng. Chem.* **52** (1961) 1003 – 1006.
2. C. B. Van der Decken et al., *Chem. Ing. Tech.* **32** (1960) 591 – 594.
3. R. Jeschar, *Arch. Eisenhüttenwes.* **35** (1964) 517 – 526.
4. H. Glaser, *Chem. Ing. Tech.* **34** (1962) 468 – 472.
5. T. Toei, M. Hiraoka, *Mem. Fac. Eng. Kyoto Univ.* **24** (1962) no. 1.
6. L. R. Galloway et al., *Can. J. Chem. Eng.* **35** (1957) no. 4, 139 – 150.
7. L. R. Pierson, *Trans. ASME* **59** (1937) .
8. J. T. McConnachie: *Transfer Process in the Flow of Gases Through Packed and Distended Beds of Spheres,* Dept. of Chem. Engng., Northwestern University, Evanston, Ill. 1961.
9. L. Jaeschke: *Über den Wärme- und Stoffaustausch und das Trocknungsverhalten ruhender luftdurchströmter Haufwerke aus Körpern verschiedener geometrischer Form in geordneter und ungeordneter Verteilung,* Dissertation, Technische Hochschule Darmstadt 1960.
10. E. Mosberger: *Über den Wärme- und Stoffaustausch zwischen Partikel und Luft in Wirbelschichten sowie über deren Ausdehnungsverhalten,* Dissertation, Technische Hochschule Darmstadt 1964.
11. J. Ch. Chu et al., *Chem. Eng. Prog.* **49** (1953) 141 – 149.
12. G. Thodos, R. E. Ricetti: "Mass Transfer in the Flow of Gases through Fluidized Beds," *AICHE J.* **7** (1961) 442 – 444.
13. O. Krischer, E. Mosberger, *Chem. Ing. Tech.* **37** (1965) 925 – 932 and 1253 – 1258.
14. W. Peters, *Chem. Ing. Tech.* **32** (1960) 178 – 184.
15. W. Peters: *Schnellentgasung von Steinkohlen,* Habilitation, Technische Hochschule Aachen 1963.
16. A. Thau: *Die Schwelung von Braun- und Steinkohle,* Knapp, Halle/Saale 1927, pp. 658 – 687.
17. P. Dolch: *Wassergas,* Barth, Leipzig 1936, pp. 236 – 238.
18. A. Thau: *Brennstoffschwelung,* **vol. 1,** Knapp, Halle/Saale 1949, pp. 150 – 152 and 373 – 377.
19. G. Webster: "Fluid Coking; New Tool for Old Job," *Oil Gas J.* **52** (1953) no. 15, 74 – 75.
20. H. Z. Martin et al., "More on the Fluid Coking Process," *Oil Gas J.* **53** (1954) no. 1, 166 – 171.
21. T. A. Burtis et al., *Chem. Eng. Prog.* **45** (1949) 97 –101.
22. S. C. Eastwood, A. E. Potas, *Pet. Process.* **3** (1948) 837 – 840.
23. R. W. Rammler, H. J. Weiss: "Lurgi Ruhrgas Process," *Handbook of Synfuels Technology,* McGraw-Hill, New York 1984.
24. H. J. Weiss: "The Lurgi Ruhrgas Process in Retorting of Oil Shale," *Energy,* 1983, 17, 18.
25. R. Rammler, H. J. Weiss et al.: "Oil Recovery by Distillation of Oil Shale and Tar Sand–Contribution to Energy Supply," *Ger. Chem. Eng. Engl. Transl.* **5** (1981) 241 – 250.

Direct Reduction → Iron
Disaccharide Alcohols → Sugar Alcohols
Dishwasher Soap → Cleansing Agents

Refrigeration Technology

Siegfried Haaf, Linde AG, Köln, Federal Republic of Germany

Helmut Henrici, Linde AG, Köln, Federal Republic of Germany

1. Refrigeration 2
1.1. Compressor Refrigerating Systems 3
1.1.1. Refrigerants for Compressor Refrigerating Systems 8
1.1.2. Compressors 10
1.1.2.1. Displacement Compressors 10
1.1.2.2. Dynamic Compressors 15
1.1.3. Condensers 16
1.1.4. Evaporators 19
1.1.5. Other Equipment 21
1.2. Absorption Refrigerating Systems 23
1.2.1. Ammonia – Water Absorption Refrigerating Systems 23
1.2.2. Equipment for Ammonia – Water Refrigerating Systems 25
1.2.3. Water – Lithium Bromide Absorption Refrigerating Systems 26
1.3. Steam-Jet Systems 27
1.4. Thermoelectric Refrigeration . . . 28
2. Industrial Applications of Refrigeration 28
2.1. Chilling of Liquids 28
2.2. Separation of Liquids by Crystallization 30
2.3. Cooling and Liquefaction of Gases 31
2.4. Remote Cooling 33
2.5. Air Conditioning 34
2.6. Refrigeration Plants for Cold Rooms and Freezers 34
2.7. Use of Cold in Construction 39
2.8. Air Refrigeration in Mining 39
2.9. Artificially Cooled Skating Rinks 39
3. Heat Pumps 39
4. References 40

Symbols

A heat-transfer area, m^2
b channel width of impeller, m
c correction factor
D outside diameter of impeller, m
e thermoelectric force, V/K
g acceleration due to gravity, m/s^2
h specific enthalpy, kJ/kg
$\dot{m}$ mass flow rate, kg/s
n rotational speed, s^{-1}
p pressure, Pa
P power consumption, kW
q capacity per unit mass of refrigerant, kJ/kg
Q useful heat generated, kJ
$\dot{Q}$ capacity, kW
R fouling resistance, m^2 K/W
s specific entropy, kJ kg^{-1} K^{-1}
u tip speed (centrifugal compressor), m/s
U overall heat-transfer coefficient, W m^{-2} K^{-1}
v specific volume, m^3/kg
$\dot{V}$ volume flow rate, m^3/s
w specific work, kJ/kg; velocity of cooling water, m/s
W fuel energy consumed, kJ
z figure of merit (Peltier effect), K^{-1}
β average coefficient of performance (electric heat pump)
γ average performance ratio
Δh discharge head (centrifugal compressor), kJ/kg
ζ performance ratio; heat ratio
ε coefficient of performance
ε_0 clearance volume (reciprocating compressor)
$\varkappa$ isentropic exponent
η efficiency
λ thermal conductivity, W m^{-1} K^{-1}; total volumetric efficiency
ξ concentration of refrigerant in solution, kg/kg
σ electrical conductivity, Ω^{-1} m^{-1}
ψ pressure coefficient

Subscripts

A absorber
c condenser; condensing
C Carnot

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_19

e	effective; external heat-transfer area
el	electrical
g	geometric
H	generator
HP	high-pressure; high-stage
i	indicated; internal heat-transfer area; intercooler
I	intermediate
IP	intermediate-pressure
LP	low-pressure; low-stage
p	pump
r	rich
R	reflux condenser; refrigerant
t	ideal (theoretical)
v	volumetric
w	wall; weak
0	evaporator; evaporating

1. Refrigeration [1–3], [28], [32], [34]

The technology of generating and using subambient temperatures is divided into three fields: refrigeration, low-temperature technology, and cryogenics. The boundaries between these fields are not clear cut. *Refrigeration* is generally understood to mean from ambient temperature down to the standard boiling point of ethylene (ca. 170 K); *low-temperature technology* is from 170 K down to the melting point of nitrogen (63 K); and *cryogenics* is below 120 K, as suggested by the Ausschuss für Tieftemperatur-Terminologie (Committee for Low-Temperature Terminology). This article deals with refrigeration; cryogenics is the subject of a separate article (→ Cryogenic Technology).

The purpose of refrigeration technology is to cool materials to subambient temperatures or hold them at such temperatures. The substance to be refrigerated is cooled by a working fluid, which moves in a closed cycle. Most of the cyclical processes employed for refrigeration are reversals of processes used in heat engines.

In *cold-gas processes*, the working fluid is always in the gaseous state; in *cold-vapor processes*, the working fluid changes state from liquid to gas and back to liquid. Cold-gas processes are energetically less favorable than cold-vapor processes at all temperatures of interest for refrigeration, i.e., down to ca. − 100 °C. In practice, cold-gas machines are important at lower temperatures, and these systems are therefore discussed under → Cryogenic Technology.

Cold-vapor systems use the latent heat of evaporation of a liquefied working fluid in a closed cycle to generate cold. The fluid (refrigerant) moves continuously through the cycle. The evaporating refrigerant remains at either constant pressure or constant temperature as heat is absorbed from the medium being cooled; it is then condensed at a higher pressure. The condensing pressure is dictated by the conditions of heat rejection to the surroundings and by the vapor pressure of the refrigerant. To keep the refrigerant circulating between the two pressures, energy must be supplied to the system in the form of mechanical work or heat.

Cold-vapor machines include the following types:

1) compressor refrigerating systems,
2) absorption refrigerating systems, and
3) steam-jet refrigerating systems.

Mechanical or electrical energy must be used to drive compressor refrigerating machines; absorption and steam-jet refrigerating systems are powered by thermal energy.

1.1. Compressor Refrigerating Systems [2], [3], [14], [17], [19], [20], [28], [35–43]

The most frequently used class of cold-vapor refrigeration machines is the compressor type. Figure 1 is a schematic diagram of a single-stage refrigeration system.

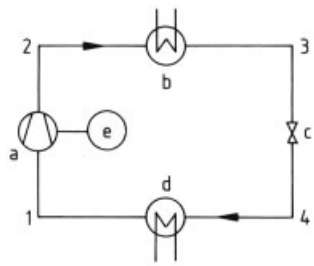

Figure 1. Single-stage compressor refrigerating system
a) Compressor; b) Condenser; c) Expansion valve; d) Evaporator; e) Compressor drive motor

A refrigerant enters the suction of a compressor (a) at a low pressure p_0, it is compressed to

a higher pressure p_c, and is then condensed in a condenser (b) by rejection of heat to the surroundings. The liquefied refrigerant is expanded to pressure p_0 in an expansion valve (c). The amount of refrigerant vapor (flash gas) produced depends on operating conditions; this part of the refrigerant cannot be used to generate cold. As it absorbs heat, the liquid refrigerant evaporates in the evaporator (d); it then returns to the suction of the compressor.

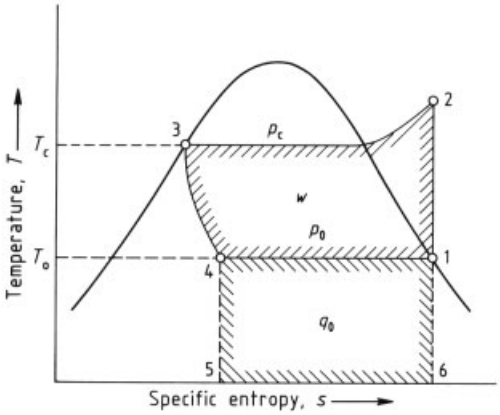

Figure 2. Diagram of the single-stage process in a compressor refrigerating system (temperature vs. specific entropy)
1 – 2 = Isentropic compression; 2 – 3 = Cooling and condensing at constant pressure; 3 – 4 = Expansion (isenthalpic); 4 – 1 = Evaporation at constant pressure; 1 – 2 – 3 – 4 – 1 = Specific work of compression; 1 – 4 – 5 – 6 – 1 = Specific refrigerating effect

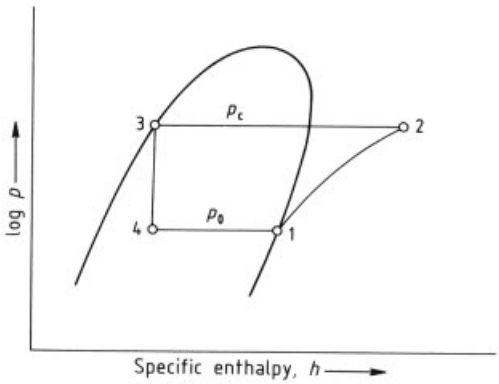

Figure 3. Diagram of the single-stage compressor refrigerating process (log pressure vs. specific enthalpy)
1 – 2 = Isentropic compression; 2 – 3 = Cooling and condensing at constant pressure; 3 – 4 = Expansion (isenthalpic); 4 – 1 = Evaporation at constant pressure

The thermal balance of the process is then

$$\dot{Q}_0 + P = \dot{Q}_c$$

where $\dot{Q}_0$ is the rate of cold production (refrigerating effect) in the evaporator, P is the compressor drive power, and $\dot{Q}_c$ is the rate of energy rejection in the condenser. When referred to the mass flow rate, this relation is

$$q_0 + w = q_c$$

where w is the specific work of compression. For energetic assessment of compressor refrigerating systems, the coefficient of performance (ε) is generally used; this is the ratio of refrigerating effect $\dot{Q}_0$ to the electrical or mechanical power supplied P:

$$\varepsilon = \frac{\dot{Q}_0}{P} = \frac{q_0}{w}$$

The coefficient of performance can be determined both for the real refrigeration process and for a suitable ideal (theoretical) reference process. The ratio of the coefficients of performance of these processes is called the efficiency factor (η):

$$\eta = \frac{\varepsilon}{\varepsilon_t}$$

For the fundamental ideal reference process in refrigeration, the Carnot process running between the absolute condensing temperature T_c and the absolute evaporating temperature T_0, the coefficient of performance is

$$\varepsilon_c = \frac{T_0}{T_c - T_0}$$

The smaller the temperature difference ($T_c - T_0$), the higher is the coefficient of performance.

The ideal process for a compressor refrigerating system is described by the temperature – entropy ($T - s$) diagram (Fig. 2), which shows the specific refrigerating effect q_0 and the specific compressor work w as separate areas. A log pressure vs. enthalpy diagram ($p - h$ or Mollier diagram, Fig. 3) is more useful because (1) both specific quantities are defined as enthalpy differences, (2) the process runs between two fixed pressures, and (3) expansion is a constant-enthalpy process. The quantities, now referred to 1 kg of circulating refrigerant, can be plotted as lengths. The specific compression work (w),

the specific refrigerating effect (q_0), and the specific thermal power (q_c) are defined as follows:

$$w = h_2 - h_1$$

$$q_0 = h_1 - h_4$$

$$q_c = h_2 - h_3$$

where the numerical subscripts refer to the points shown in Figure 3.

Performance Figures. The mass flow rate of refrigerant for a given refrigerating effect $\dot{Q}_0$ is

$$\dot{m}_R = \frac{\dot{Q}_0}{q_0} = \frac{\dot{Q}_0}{h_1 - h_4}$$

The specific refrigerating effect per unit specific volume q_{0v} is given by

$$q_{0v} = \frac{h_1 - h_4}{v_1}$$

where v_1 is the specific volume of refrigerant vapor at the suction. The volume flow rate at the suction of the compressor is thus

$$\dot{V}_1 = \frac{\dot{Q}_0}{q_{0v}} = \frac{\dot{Q}_0}{h_1 - h_4} v_1 = \dot{m}_R v_1$$

The coefficient of performance ε_t of the ideal process is

$$\varepsilon_t = \frac{q_0}{w} = \frac{h_1 - h_4}{h_2 - h_1}$$

The power consumption of the ideal process for a given refrigerating effect can be determined:

$$P_t = \frac{\dot{Q}_0}{\varepsilon_t} = \dot{Q}_0 \frac{h_2 - h_1}{h_1 - h_4}$$

The necessary rate of heat rejection needed in the condenser for the ideal process is

$$\dot{Q}_{ct} = \dot{m}_R q_{ct} = \dot{m}_R (h_2 - h_3)$$

In practice, compression is not an isentropic process, but a polytropic one in which the entropy increases. This nonideal behavior is described by the indicated efficiency η_i of the compressor used. The indicated power consumption P_i for a refrigerating effect $\dot{Q}_0$ is thus

$$P_i = \frac{P_t}{\eta_i} = \frac{\dot{Q}_0}{\varepsilon_t \eta_i}$$

When frictional losses in the compressor are taken into account, the effective power consumption P_e is

$$P_e = \frac{P_t}{\eta_e} = \frac{\dot{Q}_0}{\varepsilon_t \eta_e}$$

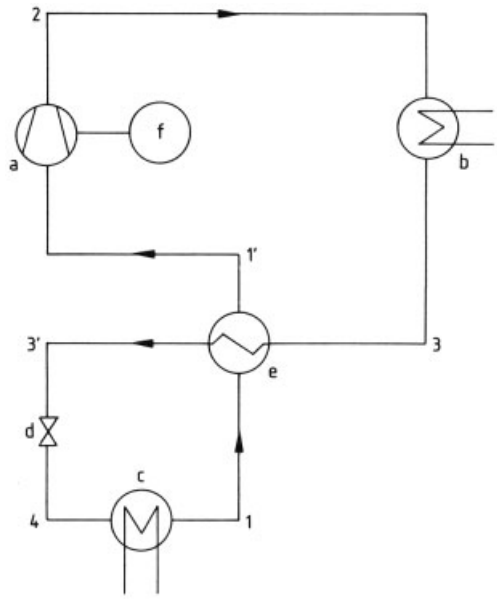

Figure 4. Single-stage compressor refrigerating system with internal heat exchange
a) Compressor; b) Condenser; c) Evaporator; d) Expansion valve; e) Heat exchanger; f) Compressor drive motor

Subcooling. The liquefied refrigerant can be subcooled by heat exchange with the cold vapor from the evaporator. A system with a heat exchanger provided for this purpose is depicted in Figure 4.

Subcooling the refrigerant results in a gain in specific refrigerating effect. The increase for some halocarbon refrigerants is greater than the increase in specific compressor work because compression is started in the superheating region instead of at saturation. The process is described by the $p-h$ diagram in Figure 5. This heat-exchange step can also be used to prevent liquid droplets from being entrained from the evaporator into the compressor. Valve damage in reciprocating compressors can thus be avoided.

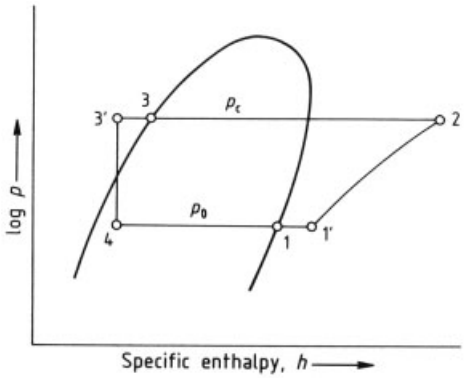

Figure 5. Diagram of the compressor refrigerating process with internal heat exchange (log pressure vs. specific enthalpy)
$h_{1'} - h_1 = h_3 - h_{3'}$

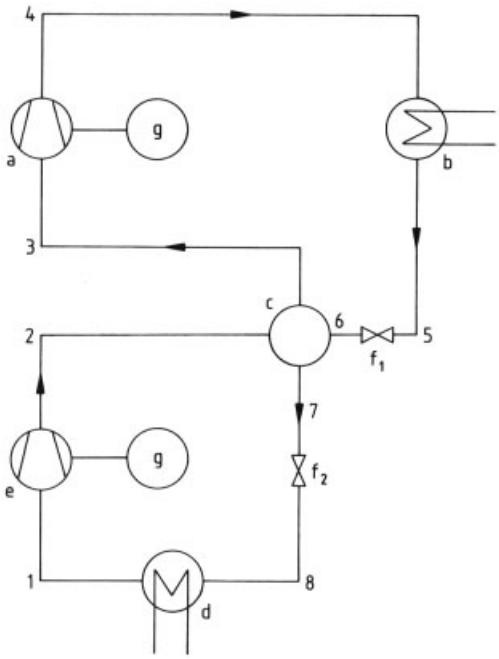

Figure 6. Two-stage compound compressor refrigerating process with two-stage expansion
a) High-stage compressor; b) Condenser; c) Intermediate-pressure vessel; d) Evaporator; e) Low-stage compressor; f) Expansion valve; g) Compressor drive motor

Compound Compression. In positive-displacement compressors, the volumetric and the indicated efficiencies of the compressor fall off sharply at higher compression ratios. In addition, compression temperatures may reach excessive values if the ratio is high. This happens at evaporating temperatures between ca. -20 and $-45\,^{\circ}$C, depending on the condensing temperature, the refrigerant, and the compressor design. In order to avoid these problems, compound (two-stage and multistage) compression systems are used. The optimum intermediate pressure for two-stage compression is

$$p_{\mathrm{I}} = (p_{\mathrm{c}} \cdot p_0)^{1/2}$$

Two-Stage Expansion. The refrigerating effect declines as the compression ratio increases, because of the high vapor content in single-stage expansion. Performance can be greatly improved by expansion in two stages. Figure 6 shows a process with two-stage compression and two-stage expansion; this arrangement is common in larger refrigerating systems that use ammonia as the refrigerant. Figure 7 is the corresponding $p-h$ diagram for this process.

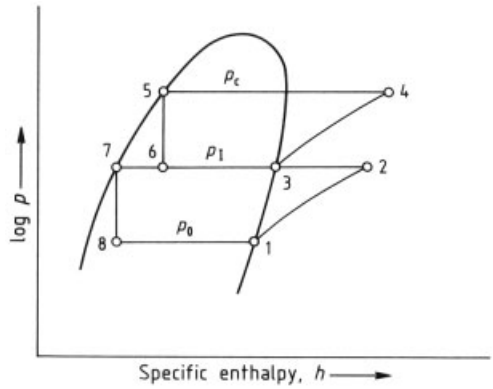

Figure 7. Diagram of compressor refrigerating process with two-stage compression and two-stage expansion (log pressure vs. specific enthalpy)

The low-stage compressor (Fig. 6, e) takes vapor from the evaporator (d) and compresses it to the intermediate pressure. This vapor goes to the intermediate-pressure tank (c), where it is cooled to saturation by refrigerant that has been expanded to the intermediate pressure. All the vapor passes to the suction of the high-stage compressor (a), where it is compressed and then condensed in the condenser (b). An expansion valve (f_1) allows the condensate to expand to the pressure prevailing in the intermediate-pressure tank; further expansion to the evaporating pressure takes place in a second expansion valve (f_2). The vapor produced by heat exchange in

the intermediate-pressure tank, along with the vapor from expansion, results in the mass flow rate of refrigerant being greater in the high-pressure part of the system than in the low-pressure part.

Two-Stage Compression with Two Evaporators. If additional refrigeration is necessary at the temperature corresponding to the intermediate pressure, another evaporator can be placed in parallel with the intermediate-pressure tank. Figure 8 shows such a system with one evaporator for low pressure (e) and one for intermediate pressure (d). The additional cooling of the vapor compressed by the low-stage compressor reduces energy consumption. The power balance for this two-stage process is as follows:

$$\dot{Q}_c = \dot{Q}_{0LP} + \dot{Q}_{0IP} + P_{LP} + P_{HP} + \dot{Q}_i$$

where $\dot{Q}_c$ is the rate of heat rejection in the condenser, $\dot{Q}_{0LP}$ is the refrigerating effect in the evaporator at p_0, $\dot{Q}_{0IP}$ is the refrigerating effect in the evaporator at p_I, P_{LP} is the drive power of the low-stage compressor, P_{HP} is the drive power of the high-stage compressor, and $\dot{Q}_i$ is the rate of heat rejection in the intercooler.

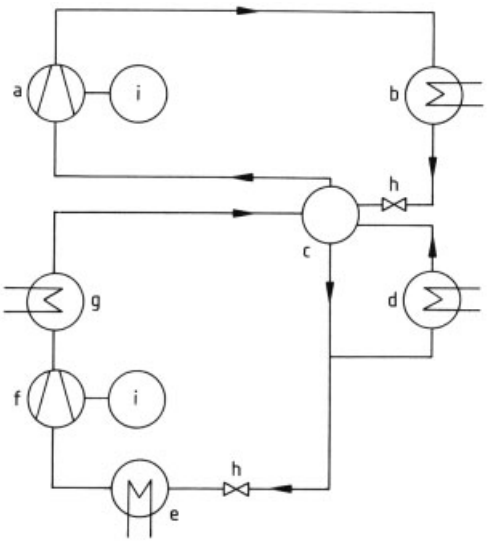

Figure 8. Two-stage compound compressor refrigerating system with two evaporators at different temperatures and with interstage cooling by cooling water and refrigerant
a) High-stage compressor; b) Condenser; c) Intermediate-pressure vessel; d) High-stage evaporator; e) Low-stage evaporator; f) Low-stage compressor; g) Interstage cooler; h) Expansion valves; i) Compressor drive motor

The refrigerant mass flow rates, the suction volume flow rates, and the effective power consumptions for the compressors can be determined from information provided on the single-stage process.

Two-Stage Compression with Single-Stage Expansion. On grounds of design simplicity, two-stage expansion is often not used in small systems with two-stage compression. This process is depicted in Figures 9 (schematic diagram) and 10 ($p-h$ diagram). In the low-pressure stage, the vapor is compressed from pressure p_0 to the intermediate pressure p_I. The vapor is then cooled by refrigerant from the condenser, which is injected through an expansion valve.

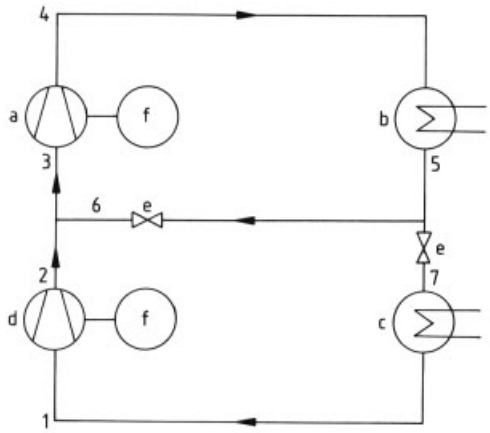

Figure 9. Two-stage compound compressor refrigerating system with vapor cooling by intermediate injection and single-stage expansion
a) High-stage compressor; b) Condenser; c) Evaporator; d) Low-stage compressor; e) Expansion valves; f) Compressor drive motor

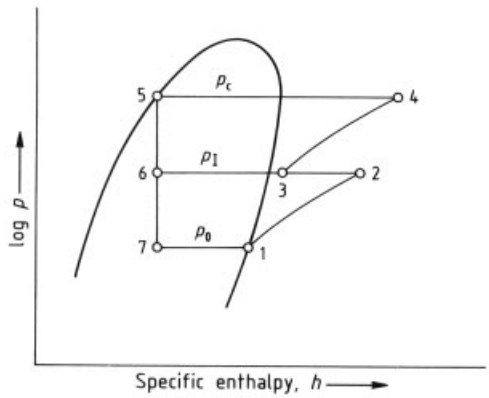

Figure 10. Diagram of the process illustrated in Figure 9 (log pressure vs. specific enthalpy.

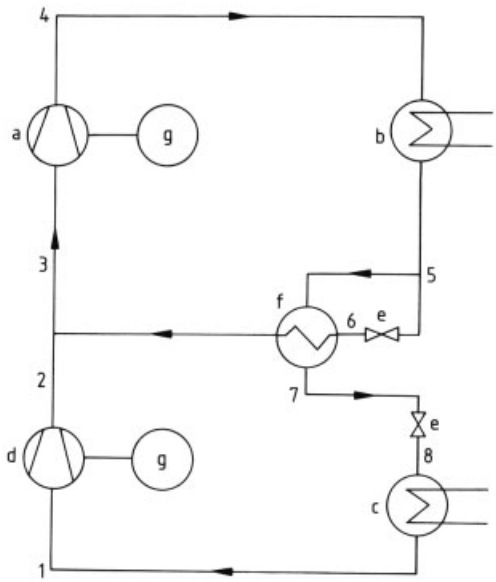

Figure 11. Two-stage compound compressor refrigerating system with single-stage expansion and refrigerant subcooling
a) High-stage compressor; b) Condenser; c) Evaporator; d) Low-stage compressor; e) Expansion valves; f) Subcooler; g) Compressor drive motor

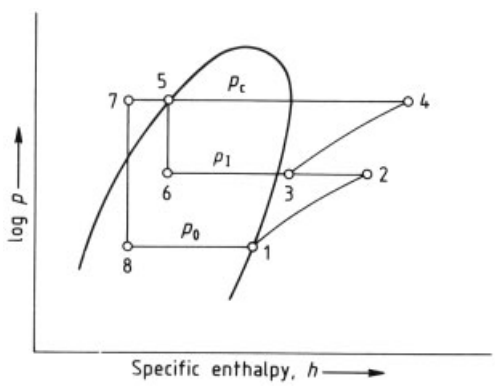

Figure 12. Diagram of the process illustrated in Figure 11 (log pressure vs. specific enthalpy)

When *halocarbon refrigerants* are employed, the system shown in Figures 11 (schematic) and 12 ($p-h$ diagram) is often used. Before expansion at low pressure, the refrigerant is subcooled almost to the saturation temperature corresponding to the intermediate pressure. Subcooling is accomplished by part of the refrigerant flow evaporating at the intermediate pressure. Energetically, the process is almost equivalent to two-stage expansion; it offers some advantages with regard to control of the refrigerant mass flow rate in the low-pressure section.

Cascade Compressor Refrigerating Systems. At evaporating temperatures between −70 and −110 °C, the large difference between the evaporating and compressing pressures has unfavorable effects. Depending on the refrigerant, the pressure on the condenser side may be excessively high or may exceed the critical pressure. On the low-pressure side, the pressure can be so low that entry of air into the system becomes a danger. The use of a given refrigerant may by ruled out by its freezing point and by the high specific volume of the suction vapor at low temperatures. The process is, therefore, divided into two separate loops and a suitable refrigerant is chosen in each loop for the pressures occurring there. Figure 13 shows a simple cascade system with two loops.

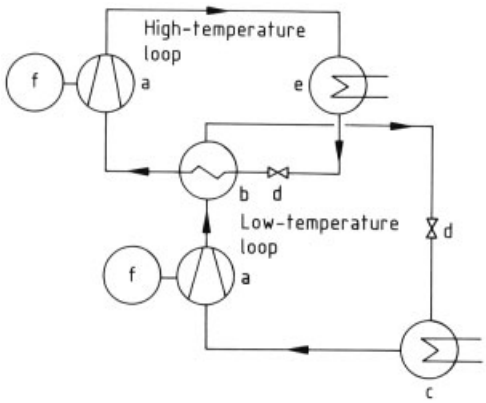

Figure 13. Cascade compressor refrigerating system
a) Compressor; b) Cascade cooler; c) Evaporator; d) Expansion valves; e) Condenser; f) Compressor drive motor

The refrigerant in the low-temperature circuit is condensed by evaporating high-temperature circuit refrigerant in the cascade cooler; i.e., the refrigerating effect of the high-temperature circuit is used to remove heat of condensation from the low-temperature circuit. In this way, only the evaporator with the lowest evaporating temperature generates the useful refrigerating effect. Depending on the compression ratios in the circuits of the cascade system, the refrigerant can be compressed in several stages.

1.1.1. Refrigerants for Compressor Refrigerating Systems [4], [14], [22], [28], [44–47]

The working fluids moving in a closed cycle in compressor refrigerating systems must satisfy many requirements. These requirements are primarily concerned with thermodynamic properties, heat-transfer characteristics, chemical and physiological behavior, and usability with various compressor designs.

A refrigerant must have good chemical stability under working conditions and in the presence of construction materials and lubricants. Its pressure must not be too high, but it should not be lower than atmospheric pressure at standard temperature and, if possible, in the evaporator. Other characteristics of a good refrigerant are low compression ratio (the ratio of the condenser pressure p_c to the evaporator pressure p_0, corresponding to the condensing temperature t_c and the evaporating temperature t_0), high specific enthalpy of evaporation, and small isentropic exponent (see Table 1). Furthermore, the volumetric refrigerating effect (the refrigerating effect per unit refrigerant volume flow rate at the suction) is a crucial factor; it depends on both the specific volume of the vapor and the specific enthalpy of evaporation.

The two main types of refrigerants are ammonia and chlorofluoro derivatives of methane and ethane (halocarbon refrigerants). Pure hydrocarbons are also used in petrochemical plants.

Ammonia has very good thermodynamic properties, but is toxic and flammable (→ Ammonia, Chap. 8.). Because it also attacks materials containing copper, it is unsuitable for sealed refrigeration compressors, where the winding of the drive motor comes in contact with the refrigerant. The use of ammonia is virtually restricted to large-capacity industrial refrigeration systems.

In contrast, *halocarbon refrigerants* are neither toxic nor flammable (→ Chlorinated Hydrocarbons, → Fluorine Compounds, Organic, Chap. 3.; → Bromine Compounds). They are employed in nearly all types of compressor refrigerating systems, from home refrigerators to large air-conditioning plants. However, some halocarbon refrigerants have harmful environmental effects, especially on the ozone layer.

Refrigerant nomenclature is standardized. Each name consists of the letter R (for refrigerant) and a multidigit number from which the chemical composition can be determined (→ Fluorine Compounds, Organic, Chap. 3.).

Table 1 gives some characteristic *thermodynamic data* for the most important refrigerants. For chlorotrifluoromethane, ethane, and ethylene, the last column gives the volumetric refrigerating effect at a condensing temperature of −30 °C and an evaporating temperature of −80 °C, because these substances are primarily low-temperature refrigerants for cascade systems. Table 2 gives the temperature ranges and applications for the refrigerants listed in Table 1.

Detailed vapor tables, $p-h$ diagrams, and equations of state for use on computers are available for designing refrigeration processes. Table 3 gives a selection of data from the vapor tables for the two most important refrigerants, ammonia and chlorodifluoromethane. Along with temperature and pressure, the table includes the specific volume (v) and specific enthalpy (h) of the liquids and saturated vapors.

Special requirements (purity, water content, and contents of gaseous impurities, acids, or residues after evaporation) for several halocarbon refrigerants are specified in DIN standards. With regard to the pressures occurring in the refrigerant loop, the operating and test pressures for system components are dictated by the refrigerant used. To ensure that no refrigerant escapes to the environment, leak tests are performed before a refrigeration system is placed in service. The most important safety principles for refrigerating systems are given in ISO 1662.

1.1.2. Compressors [5], [17], [30], [48–65]

Two kinds of machines are used to compress the refrigerant vapor:

1) Displacement machines (reciprocating, screw, multi-vane-rotary compressors)
2) Dynamic machines (centrifugal or turbine compressors)

The crucial difference in the operating properties of these two types is that the volume flow rate (gas flow) is highly dependent on the compression ratio in dynamic machines, whereas it

Table 1. Thermodynamic properties of refrigerants

Refrigerant	CAS registry number	Formula	Symbol	M_r	*mp*, °C	Standard *bp*, °C	Specific enthalpy of evaporation at standard *bp*, kJ/kg	Isentropic exponent at $p = 0$ and $t = 0$ °C	Volumetric refrigerating effect at −15 to +30 °C, kJ/m^3 *
Ammonia	[*7664-41-7*]	NH_3	R 717	17.03	− 78	− 33.4	1368	1.31	2175
Trichlorofluoromethane	[*75-69-4*]	$CFCl_3$	R 11	137.38	−111	23.8	182	1.12	204
Dichlorodifluoromethane	[*75-71-8*]	CF_2Cl_2	R 12	120.92	−158	− 29.8	166	1.14	1281
Chlorotrifluoromethane	[*75-72-9*]	CF_3Cl	R 13	104.47	−181	− 81.4	150	1.14	745 **
Bromotrifluoromethane	[*75-63-8*]	CF_3Br	R 13 B1	148.92	−168	− 57.8	118	1.12	2786
Chlorodifluoromethane	[*75-45-6*]	CHF_2Cl	R 22	86.48	−160	− 40.8	235	1.20	2091
Dichlorotetrafluoroethane	[*76-14-2*]	$C_2F_4Cl_2$	R 114	170.93	− 94	+ 3.6	137	1.09	369
Azeotrope of R 12 and R 152 a			R 500	99.31	−159	− 33.5	201		1507
Azeotrope of R 22 and R 115			R 502	111.60	ca. −160	− 45.6	170		2091
Ethane	[*74-84-0*]	C_2H_6	R 170	30.07	−183	− 88.6	491	1.20	1062 **
Ethylene	[*74-85-1*]	C_2H_4	R 1150	28.05	−170	−103.5	481	1.25	1995 **
Propane	[*74-98-6*]	C_3H_8	R 290	44.09	−188	− 41.6	433	1.15	1811
Propene	[*115-07-1*]	C_3H_6	R 1270	42.08	−185	− 47.7	439	1.16	2238

* Without superheating and subcooling.

** Volumetric refrigerating effect for $t_0 = -80$ °C and $t_c = -30$ °C.

Table 2. Application of refrigerants listed in Table (with temperature ranges)

Refrigerant	Temperature range, °C	Application
Ammonia	− 60 to +10	medium- and large-capacity systems with reciprocating, screw, and turbine compressors
Trichlorofluoromethane	− 10 to +20	water chillers with turbine compressors
Dichlorodifluoromethane	− 40 to +20	refrigeration systems and heat pumps with reciprocating, screw, and turbine compressors
Chlorotrifluoromethane	−100 to −60	low-temperature cascade systems, usually with reciprocating compressors
Bromotrifluoromethane	− 90 to −30	low-temperature systems with reciprocating compressors
Chlorodifluoromethane	− 70 to +20	all types of systems with reciprocating, screw, and turbine compressors
Dichlorotetrafluoroethane	− 20 to +20	heat pumps with reciprocating, screw, and turbine compressors
Azeotropic mixture R 500	− 40 to +20	refrigeration systems and heat pumps with reciprocating and screw compressors
Azeotropic mixture R 502	− 60 to 0	refrigeration systems with reciprocating compressors
Ethane	−110 to −70	low-temperature cascade systems, usually
Ethylene	−110 to −80	with turbine compressors
Propane	− 60 to −20	large-capacity refrigeration systems,
Propylene	− 60 to −20	usually with turbine compressors

Table 3. Vapor table values for ammonia and chlorodifluoromethane ∗

Temperature, °C	Ammonia					Chlorodifluoromethane				
	p, MPa	v', dm^3/kg	v'', dm^3/kg	h', kJ/kg	h'', kJ/kg	p, MPa	v', dm^3/kg	v'', dm^3/kg	h', kJ/kg	h'', kJ/kg
−70	0.0109	1.378	9006	−110.7	1356.2	0.0206	0.6706	940.1	123.02	372.97
−65	0.0156	1.389	6449	− 89.1	1365.2	0.0281	0.6765	705.3	127.87	375.52
−60	0.0219	1.401	4702	− 67.4	1374.0	0.0376	0.6825	537.2	132.84	378.07
−55	0.0302	1.412	3486	− 45.7	1382.6	0.0497	0.6888	415.0	137.92	380.59
−50	0.0409	1.424	2625	− 23.8	1391.1	0.0646	0.6954	324.8	143.11	383.09
−45	0.0545	1.436	2004	− 1.9	1399.3	0.0830	0.7021	257.2	148.40	385.55
−40	0.0717	1.449	1551	20.2	1407.3	0.1053	0.7092	205.9	153.81	387.97
−35	0.0931	1.462	1215	42.3	1415.1	0.1321	0.7165	166.5	159.31	390.34
−30	0.1195	1.475	962.6	64.6	1422.5	0.1640	0.7241	135.9	164.90	392.65
−25	0.1515	1.489	770.5	86.9	1429.7	0.2016	0.7320	112.0	170.57	394.89
−20	0.1901	1.504	622.8	109.3	1436.6	0.2455	0.7403	92.93	176.34	397.07
−15	0.2362	1.518	507.9	131.9	1443.2	0.2964	0.7490	77.69	182.16	399.16
−10	0.2908	1.534	417.7	154.5	1449.4	0.3550	0.7581	65.40	188.06	401.18
− 5	0.3548	1.549	346.2	177.2	1455.2	0.4219	0.7676	55.39	194.00	403.10
0	0.4294	1.566	289.0	200.0	1460.7	0.4980	0.7776	47.18	200.00	404.93
+ 5	0.5158	1.583	242.8	222.0	1465.9	0.5839	0.7882	40.40	206.03	406.65
+10	0.6150	1.601	205.3	245.9	1470.6	0.6803	0.7994	34.75	212.11	408.27
+15	0.7284	1.619	174.6	269.0	1474.9	0.7881	0.8112	30.03	218.21	409.77
+20	0.8573	1.639	149.4	291.4	1479.0	0.9081	0.8238	26.04	224.34	411.14
+25	1.003	1.659	128.4	314.9	1482.4	1.041	0.8373	22.66	230.50	412.38
+30	1.167	1.680	110.8	338.5	1485.3	1.188	0.8517	19.78	236.69	413.48
+35	1.350	1.702	95.96	362.3	1487.8	1.350	0.8673	17.31	242.93	414.42
+40	1.555	1.726	83.47	386.3	1489.7	1.527	0.8841	15.17	249.22	415.19
+45	1.782	1.750	72.85	410.5	1491.1	1.721	0.9024	13.32	255.57	415.76
+50	2.033	1.777	63.78	434.9	1491.9	1.933	0.9226	11.70	262.03	416.11
+55	2.310	1.805	56.07	458.6	1492.4	2.164	0.9449	10.29	268.62	416.20
+60	2.614	1.834	49.33	483.9	1491.8	2.415	0.9700	9.033	275.41	415.99

∗ p = pressure; v' = specific volume of the liquid; v'' = specific volume of the saturated vapor; h' = specific enthalpy of the liquid; h'' = specific enthalpy of the saturated vapor.

is virtually constant in positive-displacement devices. Figure 14 shows the characteristics of both types of machine.

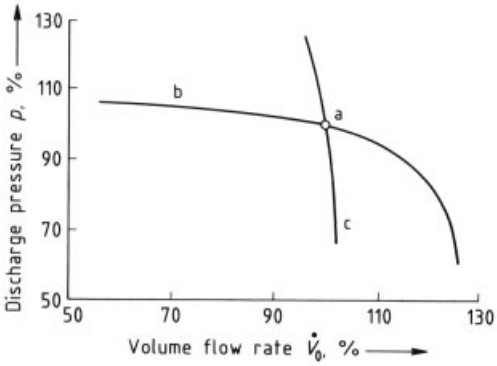

Figure 14. Operating curves for dynamic and positive-displacement compressors
a) Operating point; b) Dynamic compressor; c) Positive-displacement compressor

The gas flow and thus the refrigerating capacity are crucial in the selection of compression equipment. In Figure 15 the ranges of application of the most important compressor types are plotted as a function of the refrigerating capacity and evaporating temperature for a condensing temperature of 35 °C.

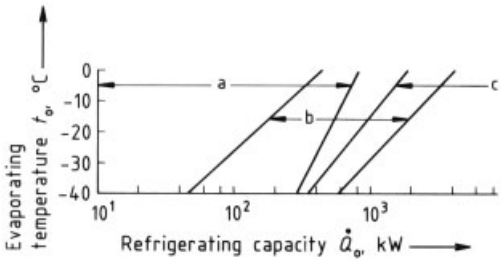

Figure 15. Range of application for refrigerant compressors
a) Centrifugal compressors; b) Screw compressors; c) Reciprocating compressors

1.1.2.1. Displacement Compressors

Reciprocating Compressors. *Piston-type reciprocating compressors* are used almost without exception for low and moderate capacities, conventional refrigerants, and evaporating temperatures of ca. − 60 °C.

These compressors are multicylinder machines (2 to 16 cylinders) with an inline-, V-, or W-configuration. The flow of refrigerant vapor alternates in direction, i.e., the cylinder head contains suction and discharge valves in an arrangement that causes the flow direction to reverse. These working valves are usually ring-plate or lamellar valves. Piston compressors operate at speeds between 700 and 1500 rpm, depending on size. The mean piston speed is ca. 4.0 m/s. In order to prevent wear of the cylinder contact surfaces and undue loading of the working valves, this figure should not be greatly exceeded. For compressors with gas flows $< 20\,m^3/h$, rotation speeds up to 3000 rpm are common. Piston compressors are usually driven by electric motors with either direct coupling or belt drives. For halocarbon refrigerants and capacities up to ca. 300 kW, the direct-coupled drive motor is often mounted in a housing with the compressor (hermetic refrigeration compressor) and cooled by the refrigerant vapor on the suction side. Advantages of this design include diminished leakage, space and weight savings, reduced noise, and lower maintenance costs.

For low evaporating temperatures (t_0 between − 20 and −45 °C), *multistage compressors* are employed. The stage compression ratios are dictated by the number of cylinders in the individual stages. Multistage compression can also be achieved by connecting two different-sized compressors in series.

Approaches used to match compressor output to varying cold requirements include not only speed control (change-pole motors) and throttling of the refrigerant vapor on the suction side, but also inclusion and exclusion (bypassing) of individual cylinders or groups of cylinders. In this type of control, compressed refrigerant vapor from the affected cylinders is returned to the suction side through pressure-reducing valves, or the suction-valve plates of individual cylinders are held open, usually by oil pressure. For simpler control, the capacity can also be divided among several smaller compressors, which are put on-line according to demand (interconnected refrigerating system).

The *volumetric losses* of a piston-type compressor are expressed in terms of the total volumetric efficiency λ, which is the ratio of the effective volumetric flow $\dot{V}_{v1}$ under suction conditions to the theoretical volumetric flow $\dot{V}_g$ calculated from the motor speed and the cylinder dis-

placement. The total volumetric efficiency declines as the compression ratio increases. It also depends on the refrigerant, the suction temperature, and the clearance volume:

$$\lambda = \frac{\dot{V}_{v1}}{\dot{V}_g} = \frac{\dot{m}_R v_{v1}}{\dot{V}_g}$$

where $\dot{m}_R$ is the mass flow rate of refrigerant and v_{v1} is the specific volume under suction conditions.

The nomograph of Figure 16 gives the total volumetric efficiency λ and the indicated efficiency η_i for compressors operating with ammonia as functions of the compression ratio p_c/p_0, evaporating temperature t_0, geometric stroke displacement per hour $\dot{V}_g$, and clearance volume ε_0. The nomograph is based on the relation

$$\lambda = \eta_v - (1 - \eta_w)$$

where η_v is the volumetric efficiency (i.e., the fraction of the reduction in delivery that is due to reexpansion of gas in the clearance volume and to valve resistance) and η_w is the wall efficiency. The wall efficiency allows for losses due to valve and piston leakage, as well as the decrease in suction-side gas density due to heating at the valves and the cylinder walls. The total volumetric efficiency found should be multiplied by a correction factor c for low suction temperatures when $t_0 < -25\,°C$. The correction factor is given by the following:

t_0, °C	−25	−30	−40	−50	−60	−70
c	1.0	0.95	0.90	0.85	0.80	0.75

The use of the nomograph is illustrated by a *sample calculation* (dashed lines in Figure 16):

Evaporating temperature $t_0 = -20\,°C$
Condensing temperature $t_c = +25\,°C$
Clearance volume $\varepsilon_0 = 0.04$; geometric displacement per hour $\dot{V}_g = 50\,m^3/h$
Compression ratio $p_c/p_0 = 5.3$
Volumetric efficiency $\eta_v = 0.82$
Wall losses $(1 - \eta_w) = 0.16$
Correction factor for all t_0 over $-25\,°C$, $c = 1$
Total volumetric efficiency $\lambda = [\eta_v - (1 - \eta_w)]$ $c = 0.66$; $\eta_i/\lambda = 1.18$
Indicated efficiency $\eta_i = 1.18 \times 0.66 = 0.78$

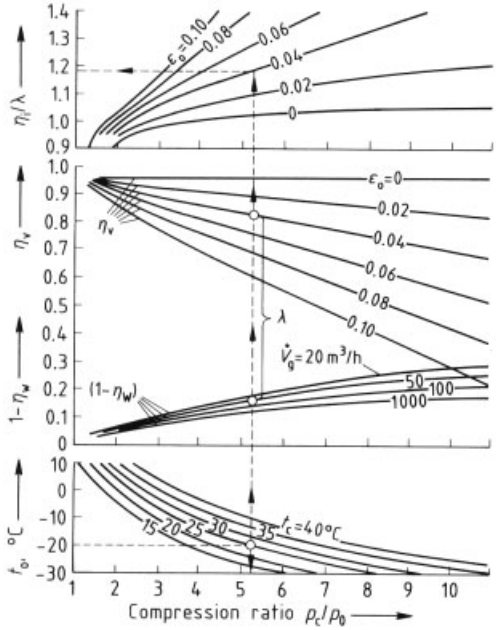

Figure 16. Total volumetric efficiency λ and indicated efficiency η_i of ammonia compressors (from DKV Specification 3-01)

The indicated power in terms of the indicated efficiency is given by

$$P_i = P_t/\eta_i$$

where P_t is the theoretical power consumption (see Section 1.1). The mechanical efficiency η_m must be included in the effective power consumption; this value is between 0.85 and 0.93 and depends on the displacement volume. For compressors operating on other refrigerants, this graphical representation gives usable values if the pressure drop in the working valves does not differ from that for ammonia.

Oil that enters the refrigerant loop in oil-lubricated piston compressors must be separated from the refrigerant stream before the inlet to the condenser, especially when no oil return from the evaporator is provided. This is necessary for plants using ammonia as refrigerant and for those using halocarbon refrigerants at lower evaporating temperatures. Gravity or inertial oil separators are installed for this purpose on the discharge side of the compressor. The separating efficiency can be enhanced by internal impingement collectors. Such oil separators ensure that an adequate quantity of oil is returned to the

compressor and that the lubricating oil will not degrade heat transfer in the heat exchangers.

In special cases, e.g., at very low evaporating temperatures, oil-less compression may be required. Nonlubricated compressors are used widely in process engineering and are suitable for this purpose. In these crosshead machines, either the piston does not make contact with the cylinder (labyrinth compressor) or the piston rings are made of a self-lubricating material, usually polytetrafluoroethylene or sintered carbon with appropriate additives (plastic-ring nonlubricated compressor).

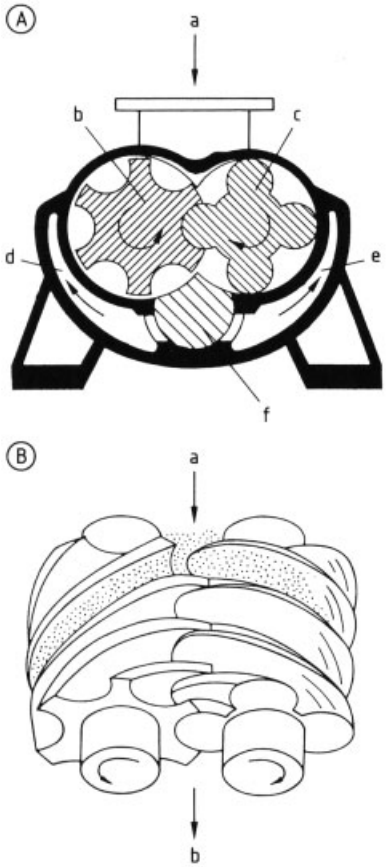

Figure 17. Principle of a rotating-screw compressor
A) Cross section: a) Inlet; b) Female rotor; c) Male rotor; d) Return channel; e) Return channel; f) Movable slidingB) Top view of the rotors: a) Inlet; b) Discharge

Screw Compressors. Screw compressors are being used increasingly in industrial refrigeration systems, especially those with moderate and high capacity.

Most of these devices are oil-flooded, two-shaft machines (Fig. 17). The refrigerant vapor is compressed by two screw-shaped rotors (b and c), which engage and make contact with each other. The rotors turn in opposite directions inside a closefitting case. As the rotors turn, the spaces between teeth continuously increase and decrease in size because of the screw shape. Thus, as in a reciprocating compressor, refrigerant vapor is pulled in, compressed, and discharged. The position and size of the discharge port in the housing determine the ratio of discharge to suction volume (installed volume ratio or, because for a given refrigerant conversion to pressure is possible, installed pressure ratio).

The screw compressor has neither clearance volume nor working valves. Precooled oil is usually injected into the compression space between rotors after completion of the suction process. This procedure allows direct torque transmission from the driven rotor (usually male) to the partner rotor, seals the clearances between rotors and housing, and removes the enthalpy of compression. Oil injection is driven by a pump or by the pressure of liquefaction. Independent of the compression ratio, pressure differences up to 2 MPa can be attained without the compression temperature exceeding 100 °C. Injected oil is separated from the refrigerant stream in one or two oil separators (depending on the degree of separation required), which are connected in series. The oil is cooled and returned to the compressor. The principal refrigerants used are ammonia (R 717) and chlorodifluoromethane (R 22); in heat-pump service, dichlorodifluoromethane (R 12) and dichlorotetrafluoroethane (R 114) are also employed.

Capacity is controlled by altering the mass flow rate with an axially adjustable movable sliding. When the sliding is moved toward the discharge port, usually by oil pressure, it creates an opening through which the suction-side vapor flows back into the inlet casing. Efficiency factors are favorable over a wide range of part-load conditions.

Small screw compressors, without a movable sliding and offering theoretical flow rates $\leq 100\ m^3/h$, are available. As with smaller reciprocating compressors, capacity can be divided

among several machines, which are turned on and off to meet demand.

Besides the tooth shape and number shown for the male and female rotors in Figure 17, *other rotor profiles* with different numbers of teeth are also available. Energetically optimal rotor tip speeds are generally 25 – 35 m/s; with smaller rotors, gearboxes are required between drive motor and compressor in order to attain these values. The drive motor speed is usually $3000\,\mathrm{min}^{-1}$.

Semihermetic screw compressors are used, especially in air conditioning. In these devices, the refrigerant vapor on the suction or discharge side is used to cool the rotor.

The most important qualities of the screw-type compressor are simple, compact design; low vibration and noise levels; high reliability; insensitivity to liquid refrigerant flow rates on the suction side; efficient, continuous capacity control; and efficiency factors similar to those of reciprocating compressors.

To reduce the specific power consumption, medium- and high-capacity screw compressors are fitted with an intermediate injection device (see Section 1.1). To allow further optimization of the efficiency factor under a variety of service conditions, the installed volume ratio of some types can be varied continuously during operation.

Multivane Rotary Compressors. The multi-vane rotary compressor (Fig. 18) consists of a cylindrical housing (a) in which an eccentrically mounted rotor (c) with several vanes (b) rotates.

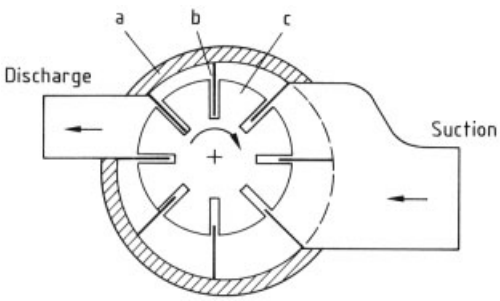

Figure 18. Principle of a multivane rotary compressor
a) Housing; b) Vane; c) Rotor

The vanes, guided in the rotor, are pressed by centrifugal force against the housing wall, forming a sequence of continuously varying compression spaces.

Overflow losses at the vanes limit compression ratios to only 1 : 5 for economic reasons. Therefore, multivane rotary compressors have restricted applications in refrigeration. They are used as boosters in reciprocating compressor plants with theoretical flow rates up to ca. $1000\,\mathrm{m}^3/\mathrm{h}$. Their advantages are low mass and small volume (due to the high rotation speeds) as well as vibration-free operation. Drawbacks include vane wear, especially when lubrication is inadequate, and the sharp drop in volumetrc efficiency at higher compression ratios.

1.1.2.2. Dynamic Compressors

Centrifugal Compressors. All turbine compressors (turbocompressors) for refrigeration service are radial-flow centrifugal designs; they are used for refrigerating capacities from ca. 300 to $>10^4$ kW. These machines have significant advantages over reciprocating compressors: low mass per unit capacity, decreased space requirement, vibration-free operation, oil-free compression, and high reliability due to the absence of reciprocating machine parts and valves.

For a given design, the discharge head Δh for one stage is set by the tip speed u of the gas:

$$\Delta h = \psi \frac{u^2}{2g}$$

where the pressure coefficient ψ depends on the vane shape (ψ = ca. 1.0 for vanes with a backward curve and 1.3 for vanes with radial ends). If the discharge head is known, then the compression ratio for isentropic compression is

$$\frac{p_c}{p_0} = \left(\frac{x-1}{x} \cdot \frac{\Delta h}{T_0} \cdot \frac{M_\mathrm{r}}{R} + 1\right)^{\frac{\varkappa}{\varkappa-1}}$$

where M_r is the relative molecular mass and R is the universal gas constant. Substances with high relative molecular masses thus favor the use of centrifugal compressors in terms of construction cost: achieving a given compression ratio then requires a lower discharge head, and often fewer stages, than with a low molecular mass fluid. Typical refrigerants for turbocompressor refrigeration plants are trichlorofluoromethane (R 11), dichlorodifluoromethane (R 12), and dichlorotetrafluoroethane (R 114); for very high

capacities, chlorodifluoromethane (R 22) and ammonia (R 717) are also candidates. Petrochemical refineries use ethylene (R 1150), propane (R 290), and propene (R 1270) as refrigerants.

Centrifugal compressors always offer lower compression ratios than positive-displacement machines. For an evaporating temperature of $-20\,°C$ and a condensing temperature of $+35\,°C$, the required compression ratio can be reached in a single stage with a reciprocating compressor; two stages of centrifugal compression are needed for R 22 and eight for R 717.

The main dimensions of centrifugal compressors are dictated not only by the refrigerant but also by the delivery volume and head. The delivery volume is proportional to nD^3b, and the head is proportional to n^2D^2 ; these relationships show how important the impeller outside diameter D is. Low delivery volumes allow small diameters, high rotation speeds n, and small impeller channel widths b. The minimum delivery volume is determined by frictional losses, which increase when the impeller channel width is small. In multistage compressors, often designed with equal-sized impellers, the suction-side volume of the last stage governs the minimum delivery volume.

The characteristic tip speed for the delivery head is limited by the strength of the impeller. For a medium with a relatively high molecular mass (low acoustic velocity), a further limit is that the mach number (ratio of flow velocity to acoustic velocity) at the inlet to the impeller must not exceed 0.8 – 0.9. Modern machines, with few exceptions, run at 4000 to 20 000 min^{-1}

Compressors are generally driven by electric motors provided with gearboxes; gas and steam turbines are also employed in larger units. For air conditioning, an enclosed construction is often used in which the electric motor is mounted in the housing and cooled by the refrigerant.

The operating characteristics of centrifugal compressors result from the heavy dependence of discharge volume on compression ratio and from the location of the pumping-limit curve. The pumping limit dictates the minimum quantities for various rotational speed curves; the speed must not go below this limit if unstable operation is to be avoided.

Capacity can be most economically controlled through the rotation speed, which is the usual arrangement with a turbine drive. If a constant-speed drive is employed, control by inlet guide vanes is also energetically favorable. A set of guide vanes ahead of the impeller imparts a twist to the inlet flow, thus altering the delivery head at constant rotation speed. A family of operating characteristics similar to those for speed control is obtained. This form of control is used in most motor-driven compressors.

Energetically unfavorable control techniques used to a limited extent include throttling the suction and varying the cooling-water rate. Pumping-limit control ensures that the flow under part-load conditions does not decrease below a level that makes operation unstable. If the flow corresponding to the desired refrigeration capacity is below the pumping limit, the difference flow rate must be additionally compressed. This quantity is usually returned by throttling (bypass control); expansion turbines are also used in large plants.

1.1.3. Condensers [3], [7], [15–17], [23], [27], [30], [33], [66–77]

Refrigerant exiting the compressor as a superheated vapor goes to the condenser, where it is cooled to the condensing temperature, liquefied, and subcooled to an extent that depends on equipment design and operating conditions. The heat rejected is usually absorbed by ambient air or cooling water. Accordingly, condensers are classified as air-cooled, water-cooled, or evaporative types.

In refrigeration systems with heat recovery and in heat pumps, the heat rejected in the condenser can also be used to heat water for industrial use or to generate space or process heat. A special design is the cascade cooler (see Section 1.1), in which the refrigerant in the low-temperature loop of a cascade compressor refrigerating system is liquefied by evaporating refrigerant in the high-temperature loop.

Air-Cooled Condensers. Air-cooled condensers (Fig. 19) are used in low- to medium-capacity refrigeration systems and, rarely, in large plants. This type of condenser is gaining

steadily in importance as freshwater becomes more scarce and more costly.

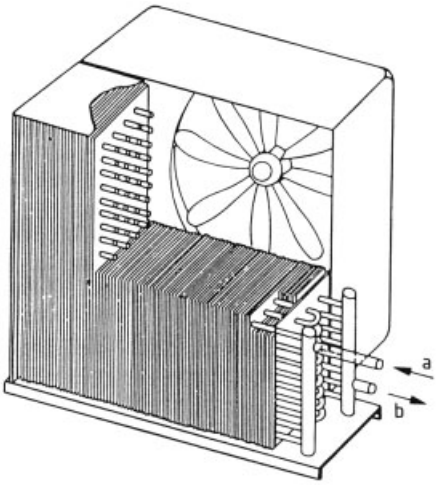

Figure 19. Air-cooled condenser
a) Refrigerant vapor inlet; b) Refrigerant condensate outlet

In very low-capacity equipment, such as home refrigerators, heat is rejected to the ambient air by free convection. In all other cases, cooling air is conveyed through the condenser, usually by several fans (forced-circulation condenser). Axial-flow fans are used when the condenser is installed outdoors; radial-flow fans are employed when additional resistance in air ducts must be overcome.

An air-cooled condenser consists of a system of tubes or hoses with fins on the air side to facilitate heat transfer. In this way, the external surface area is increased by a factor of 15 – 30. Steel tubes and fins used with ammonia as a refrigerant are galvanized to prevent corrosion and to provide a good heat-conducting surface. With halocarbon refrigerants, copper tubes and aluminum fins are generally employed. To improve heat transfer, the aluminum fins often have wavy edges and surface corrugation. The central tubes are commonly staggered in the direction of air flow to enhance turbulence. Fin spacing is 1.5 – 5 mm but should be at least 3 mm if dirty air is encountered. Heat-transfer coefficients referred to the total external surface are between 30 and 60 $\mathrm{W\,m^{-2}\,K^{-1}}$ for air face velocities of 2 – 4 m/s. The external surface area A required for heat rejection is determined by the relation

$$A = \frac{\dot{Q}_\mathrm{c}}{U \Delta t_\mathrm{m}}$$

where Δt_m is the logarithmic mean temperature difference between the condensing temperature t_c on the one hand, and the air inlet and outlet temperature on the other hand.

Air-cooled condensers are usually designed for a 10 – 15 K difference between the condensing temperature and the inlet air temperature. For economic reasons, small temperature differences are chosen for refrigeration systems with a high annual utilization, and larger differences are used when utilization is lower (e.g., in air conditioners).

If the condensing pressure becomes too low in the cold season, maintaining the needed flow of refrigerant through the expansion valve may not be possible. This is especially true of air-cooled condensers installed outdoors. To prevent this pressure loss, control of the condensing pressure is usually necessary. Control can be implemented by reducing the volume flow rate of air (blower speed control or shutoff of individual fans), altering the effective heat-transfer area (e.g., by impounding liquid refrigerant), raising the inlet air temperature by mixing it with recirculated hot air, or a combination of these methods.

A receiver is usually connected downstream of the air-cooled condenser. It prevents vapor from entering into the liquid line and, at the same time, compensates for dislocations of refrigerant caused by varying conditions in the system.

A venting device in the receiver removes foreign gases that have contaminated the refrigerant loop. (For water-cooled condensers, the vent can also be placed on the condenser.) The mixture of gas and refrigerant vapor in the top of the tank is cooled by evaporating refrigerant in the vent, so that the refrigerant condenses and flows back into the receiver. The foreign gases can then be released through a discharge valve without significant loss of refrigerant. A venting system is needed, in particular, when the evaporating pressure is < 0.1 MPa.

Water-Cooled Condensers. The most commonly used design is the horizontal shell-and-tube condenser (Fig. 20), which is used in systems with very small to very large capacities. Cooling water flows through the tubes in several

passes and the refrigerant condenses on the shell side. The end covers, containing the bends, are removable so that the tubes can be cleaned easily from the water side.

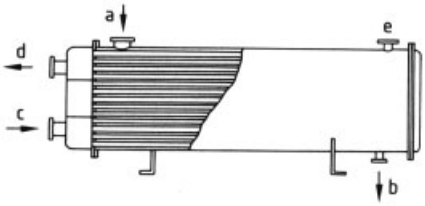

Figure 20. Water-cooled shell-and-tube condenser
a) Refrigerant vapor inlet; b) Refrigerant condensate outlet; c) Cooling water inlet; d) Cooling water outlet; e) Vent pipe

Table 4. Typical fouling resistances R_i for various cooling waters

Cooling water	Fouling resistance, $\times 10^{-4}$ m^2 K/W	
	$\leq$0.9 m/s *	>0.9 m/s *
Seawater	1	1
Brackish water	4	2
Water from cooling tower	2	2
Spring water	2	2
Stream water	4	2

* Water velocity

Sizing of this type of equipment is dictated not only by capacity, but also by the properties of the available cooling water.

Typical inlet water temperatures are

Spring water	10 – 15 °C
Stream water	15 – 25 °C
Seawater	20 – 32 °C
Water from cooling tower	25 – 35 °C

The water is commonly warmed 4 – 15 K (larger temperature differences for lower inlet water temperatures). The condensing temperature is generally 3 – 6 K above the outlet water temperature.

The relationship used to determine the heat-transfer area is the same as that for air-cooled condensers, but no allowance is made for superheating or subcooling temperatures.

The overall heat-transfer coefficient U depends on the coolant velocity w, the refrigerant, and the degree to which the inside walls of the tubes are contaminated. For shell-and-tube condensers in ammonia service, these effects can be seen in Figure 21. Typical fouling resistances R_i for various types of cooling water are given in Table 4.

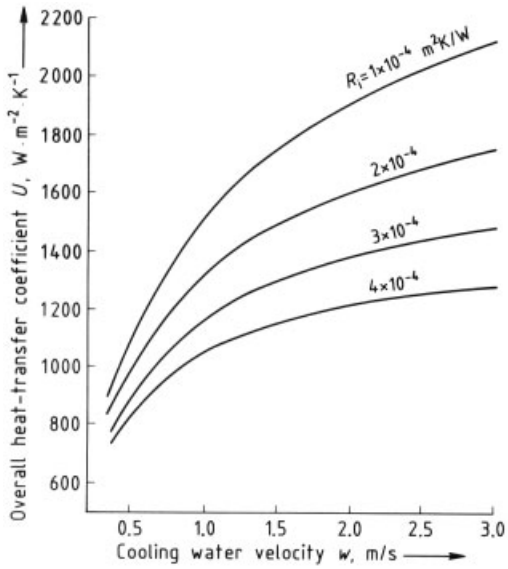

Figure 21. Heat-transfer coefficients for shell-and-tube ammonia condensers
R_i = internal fouling resistance

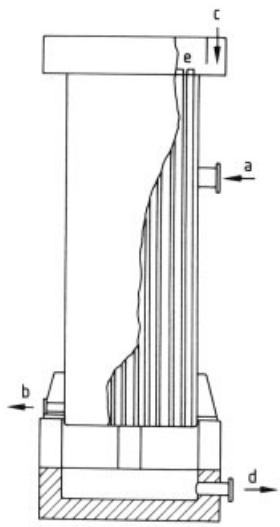

Figure 22. Vertical condenser
a) Refrigerant vapor inlet; b) Refrigerant condensate outlet; c) Cooling-water inlet; d) Cooling-water outlet; e) Spiral inserts to guide water to tube wall

Although bare steel tubes are used for ammonia, tubes finned on the refrigerant side (area ratio $A_a/A_i = 4 - 6$) are usually employed

with halocarbon refrigerants because of poorer heat transfer. The tubes are made of copper or brass; with seawater, they are made of copper – nickel alloys. The heat-transfer coefficients for halocarbon refrigerants, referred to the external tube area, are $600 - 1200\,\mathrm{W\,m^{-2}\,K^{-1}}$ at $w = 1.5 - 3$ m/s when the fouling resistance $R_i = 1 - 2 \times 10^{-4}\,\mathrm{m^2\,K/W}$.

If the water is very dirty or space is limited, vertical condensers are used (Fig. 22). These shell-and-tube devices are open on the water side and easy to clean; under certain conditions, cleaning does not require shutdown. Cooling water enters from a distribution tank at the top and is led down the inside of the tubes by spiral vanes. Thus, the water falls as a film, while the refrigerant condenses on the outside of the tubes.

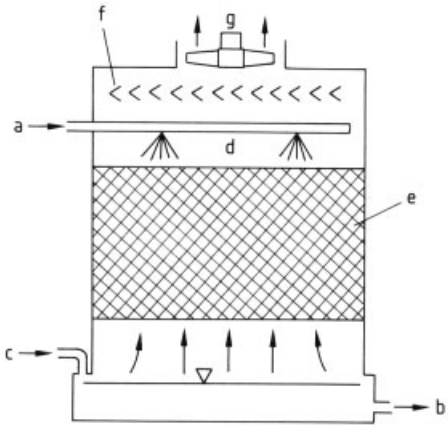

Figure 23. Cooling tower
a) Cooling-water inlet; b) Cooling-water outlet; c) Freshwater inlet; d) Spray nozzles; e) Cooling surface (fill); f) Water separator; g) Fan

For smaller capacities up to ca. 50 kW, double-pipe (coaxial) condensers are frequently used. An outer pipe, wound in a spiral encloses one or more central tubes, which carry water inside and are finned on the refrigerant side. This design is inexpensive, but difficulty in cleaning makes it unsuitable for use with dirty cooling water.

Because of the scarcity of spring water, water-cooled condensers are usually operated in conjunction with cooling towers; this allows water savings of ca. 95 % compared to freshwater operation. In the cooling tower (Fig. 23), warm cooling water is brought in contact with ambient air countercurrently. Because the partial pressure of water vapor in the air is lower than that of the cooling water, a small portion of the water evaporates and thus cools itself. The cooled water is collected in a catch basin and is then available to absorb heat again in the condenser.

Evaporative Condensers. An evaporative condenser combines the actions of a cooling tower and a water-cooled condenser. It is used for medium to large capacities when a cooling tower is not available and when cooling water is in short supply. Freshwater consumption is the same as that with a tower and water cooling.

The refrigerant condenses in a system of tubes (Fig. 24), which are sprayed with water from a tank. Air flows upward through the tube array either by free convection or, more often, by forced draft. The air absorbs the heat of condensation by evaporating water from the tube surfaces. The water level in the tank is held constant by freshwater supplied under float control. The mist collector above the array of pipes prevents water droplets entrained in the airstream from escaping.

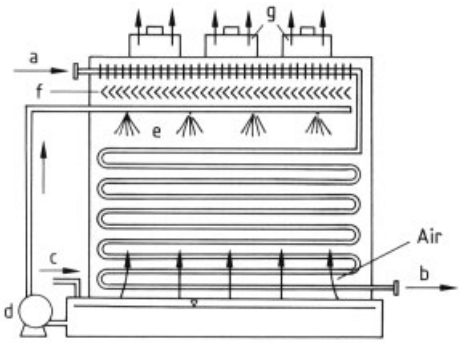

Figure 24. Evaporative condenser
a) Refrigerant vapor inlet; b) Refrigerant condensate outlet; c) Freshwater inlet; d) Water circulation pump; e) Spray apparatus; f) Water separator; g) Blower

Figure 25 illustrates the relationships of surface loading, condensing temperature, and inlet-air enthalpy for an ammonia condenser. To prevent severe scaling on the tubes, inlet water must have a low hardness. Many designs feature finned desuperheater tubes, which are not sprayed with water on the air outlet side, as well as special tubes for subcooling the refrigerant.

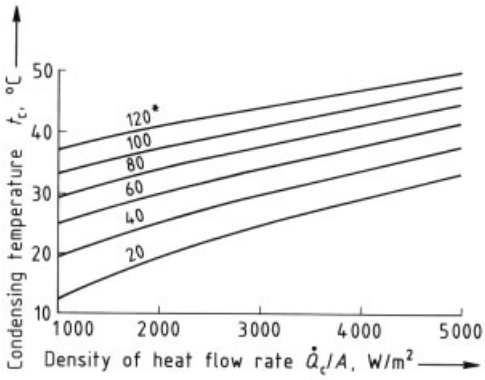

Figure 25. Relationships of density of heat flow rate, condensation temperature, and enthalpy of inlet air for an evaporative ammonia condenser as shown in Figure 24
∗ Enthalpy of suction air h, kJ/kg

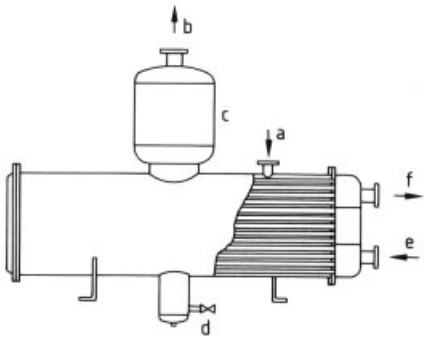

Figure 26. Horizontal shell-and tube evaporator for ammonia
a) Refrigerant inlet; b) Refrigerant vapor outlet; c) Separator; d) Oil drain; e) Brine inlet; f) Brine outlet

1.1.4. Evaporators [3], [7], [15–17], [23], [27], [30], [33], [78–90]

Evaporators for Liquid Cooling. For moderate to very high refrigerating capacities, flooded *horizontal shell-and-tube evaporators* are used (Fig. 26). These devices consist of a cylindrical shell with tube sheets at either end; the tubes are fastened in the sheets by welding, soldering, or rolling. Bare or finned tubes, usually of steel, copper, or brass, are used as appropriate for the refrigerant and the surface loading. The liquid to be cooled is pumped through the tubes and can be redirected one or more times at the ends of the evaporator. The refrigerant evaporates from the outside of the tubes. Heat-transfer coefficients are 400 – 1200 W m^{-2} K^{-1}, depending on the secondary (brine) and primary refrigerants.

Entrainment of liquid refrigerant from the evaporator, which can occur especially at higher surface loading, is prevented by an impingement separator located in the tube-free part of the evaporator or by a separator (usually using the effect of gravity) placed above the evaporator.

Special devices must be provided to recycle the oil carried into the apparatus with the refrigerant. The type of device is dictated by the refrigerant and the service conditions. Ammonia is insoluble in oil, and by virtue of its higher relative density, oil collects at the bottom of the evaporator, where it can be withdrawn. With oil-soluble halocarbon refrigerants, a side stream of refrigerant – oil mixture must be withdrawn from the evaporator. The refrigerant must then be removed from the side stream by steam or electrical heating, or by heat exchange with warm refrigerant at the condensing pressure.

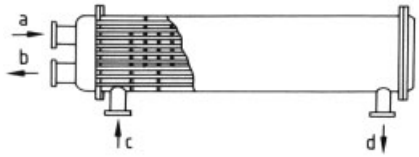

Figure 27. Shell-and-tube direct-expansion evaporator
a) Refrigerant inlet; b) Refrigerant vapor outlet; c) Brine inlet; d) Brine outlet

Because of the significant cost of recycling the oil, a different shell-and-tube evaporator design is often used with halocarbon refrigerants. In the *direct-expansion evaporator* (Fig. 27), refrigerant is completely evaporated in the tubes, while the liquid to be cooled is circulated around them by means of baffles. For water chillers especially, internally finned tubes are used instead of bare ones to enhance heat transfer on the evaporating side. If the tubes are connected in series so as to make several passes, improved heat transfer results from the increased mass flow density. However, the simultaneous pressure drop on the refrigerant side must be considered; this leads to a corresponding decline in the saturation temperature between the inlet and the outlet, and thus to a decrease in the mean temperature difference between the brine and the refrigerant. The heat-transfer coefficients re-

ferred to the external tube surface area may be 400 – 2000 W m^{-2} K^{-1}, depending on the brine being cooled and the tube design.

A *vertical-type evaporator* (Fig. 28) consists of a system of vertical, inclined, or hairpin tubes connected to one header at the top and another at the bottom, with a liquid separator in line downstream. The refrigerant stays at a constant level in the tubes as it evaporates, thus cooling the secondary refrigerant which is circulated around the tubes by an agitator. This kind of apparatus was formerly employed to cool brine in open vessels; today, it is used mainly as an ice-bank evaporator, where the ice frozen around the tubes is melted off periodically by the circulating water.

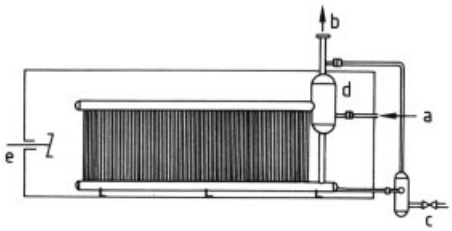

Figure 28. Vertical-type evaporator for ammonia
a) Refrigerant inlet; b) Refrigerant vapor outlet; c) Oil drain; d) Refrigerant separator; e) Agitator

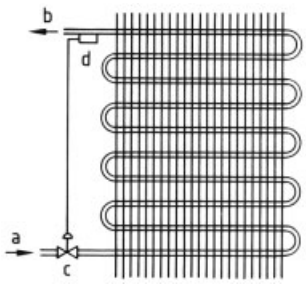

Figure 29. Finned air cooler as direct-expansion evaporator
a) Refrigerant inlet; b) Refrigerant vapor outlet; c) Thermostat-controlled expansion valve; d) Sensor

A *parallel-plate evaporator* consists of several plates, connected in parallel, in which the refrigerant evaporates. The corrugated plates are welded or soldered together to form narrow channels, which may be horizontal or vertical. This kind of device can thus be employed in direct expansion or flooded mode. The plates are made of galvanized or lacquered steel or, in special cases, stainless steel. These units are employed as drinking-water and brine chillers and for ice storage. The heat-transfer coefficients are similar to those for vertical-tube evaporators under the same conditions.

Evaporators for Gas Cooling. Gases, especially air for space cooling, are cooled on the outer surface of *finned evaporators* (Fig. 29).

The apparatus consists of a system of tubes with bonded fins that are connected in series and parallel on the refrigerant side. The gas is usually drawn or forced through the system of fins by a blower connected to the evaporator; natural convection cooling is employed less often. When the gas is cooled, water or other constituents often condense or freeze. If a frost coating forms on the tubes and fins, continuous operation is possible only if the system includes a second unit. Defrosting is accomplished from the pressure side of the compressor with warm compressed gas, which gives up its heat of condensation to the cold surfaces of the apparatus; otherwise, electric heating rods can be built into the evaporator system. To eliminate frequent defrosting, fin spacings of 8 – 20 mm are used in air coolers with a frost deposit. In air conditioners for operation above the freezing point of water, compact systems with close fin spacings (ca. 2 mm) are used.

System capacities, materials, and tube or fin sizes vary widely. As with air-cooled condensers, galvanized steel is used in ammonia systems, and copper tubes with aluminum fins are used for halocarbon refrigerants.

With halocarbon refrigerants, as in the case of liquid chillers, direct expansion is used. The refrigerant is expanded into the tubes as a liquid – vapor mixture; it is completely evaporated and superheated by several degrees. Where a network of evaporators with finned units exists (e.g., in ammonia refrigerant plants), expanded liquid refrigerant is transported through the system by pumps. Part of the refrigerant evaporates, and the resulting vapor – liquid mixture goes to a collecting tank that also serves as a refrigerant separator.

Instead of the evaporating refrigerant, a secondary refrigerant can be pumped through the chiller. At air temperature $<0\,°C$, heat-transfer coefficients are relatively low (ca. 7 – 17 W m^{-2} K^{-1}) because of frost formation. At temperatures $>0\,°C$, values of up to 30 W m^{-2} K^{-1} can be achieved with some designs. At gas pressure significantly greater than 0.1 MPa, shell-

and-tube devices are used; like similar liquid chillers, these can be designed for flooded or direct expansion operation.

1.1.5. Other Equipment [6], [15], [18], [23], [30], [91–96]

Safe and economical operation of vaporcompressor refrigerating systems involves not only the main components (compressor, condenser, and evaporator) but also a number of additional components.

In *expansion devices,* refrigerant coming from the condenser is allowed to expand to the evaporating pressure. These devices must supply the evaporator with as much refrigerant as it needs under the prevailing service conditions. Expansion valve design depends mainly on the selected evaporator design. The manual control valves formerly in use have now been replaced by automatic valves. High-pressure float controllers relieve the pressure on the liquid coming from the condenser. They are used in plants employing flooded evaporators, as well as in pumped-circulation systems where the refrigerant is expanded in a vessel at the evaporating pressure and the liquid refrigerant is pumped to the evaporator. In flooded operation with a highly variable cooling load, a low-pressure float controller is preferred to keep the liquid level in the evaporator constant.

Constant-pressure expansion valves and thermostatic expansion valves are employed with direct expansion evaporators. *Constant-pressure expansion valves* control the liquid flow to the evaporator in such a way that the evaporating pressure remains constant. Control is satisfactory only at constant load, and therefore this approach is now seldom used.

Constant refrigerant level in the evaporator and thus good heat transfer are achieved with the *thermostatic expansion valve* (Fig. 30). A temperature sensor (c) downstream of the evaporator controls the expansion valve so that the refrigerant leaves the evaporator at a temperature a few degrees higher (usually 5 – 7 K) than the evaporating temperature. In conventional thermostatic expansion valves, a vapor pressure that is dependent on the sensor temperature acts on the control diaphragm to alter the opening of the valve. More recent designs of thermostatic expansion valves use electronic actuation, which offers better dynamic properties; this can be an advantage especially when the cooling loads vary rapidly.

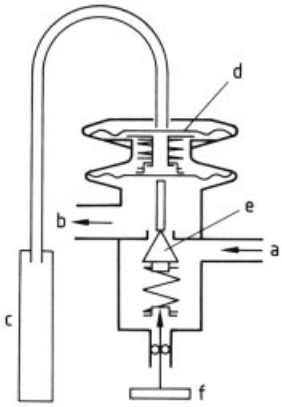

Figure 30. Thermostatic expansion valve
a) Refrigerant inlet; b) Refrigerant outlet, c) Temperature sensor; d) Control diaphragm; e) Valve cone; f) Adjusting spindle

Along with the expansion valves used as primary controllers in the refrigerant loop, *secondary controllers* can also be employed. These devices include:

1) thermostatic postinjection valves,
2) constant-pressure valves, and
3) capacity controllers.

Thermostatic postinjection valves are used when highly superheated suction-side refrigerants lead to excessive compression temperatures. To prevent this, the valves inject liquid refrigerant from the condenser into the suction line. This process is usually controlled by a temperature sensor on the compressor discharge side. Constant-pressure valves are employed when a refrigerant loop contains several evaporators at different temperatures and the evaporating temperature must not go below a certain limit in one of the units (e.g., to prevent icing in a water chiller). Capacity controllers match the compressor capacity to the refrigerating effect. If the available compressor controls are not adequate, the capacity controller can bypass part of the compressed refrigerant to the compressor suction, thus preventing a drop in suction pressure.

Thermostats are used to maintain constant temperature in cooled rooms or in the liquid outlet from evaporators. If the sensor in the con-

trol unit detects a change in temperature, a compressor capacity stage is turned on or off. If the evaporating pressure is to be held constant, a low-pressure *pressostat* is used instead of a thermostat. The high-pressure pressostat carries out a corresponding function on the condenser side. Oil-pressure differential cutouts are employed as oil delivery monitors in compressors with pressurized oil lubrication.

For freshwater-cooled refrigerating systems, *automatic water valves* reduce water consumption. They control the flow rate of cooling water so that a constant condensing pressure is maintained, independent of the condensing capacity required.

Automatic shutoff, especially of refrigerant piping, is accomplished with *solenoid valves*. These devices enable, for example, the supply of refrigerant to an evaporator to be discontinued. If intermediate positions between open and closed are needed, as is often the case in cooling-water and brine loops, *motor-actuated valves* are used.

Pressure safety relief devices are employed to prevent excessive pressure in system components. These devices include rupture disks, pressure safety relief valves, and safety pressure cutouts. Their design and use must comply with applicable national safety provisions and standards (DIN 8975 in the Federal Republic of Germany).

Systems using halocarbon refrigerants often have *heat exchangers* to subcool the liquid refrigerant flowing to the control device and superheat the refrigerant vapor flowing to the compressor. The exchangers improve the coefficient of performance and volumetric refrigerating effect with many refrigerants (R 12), (R 502) and the filling ratio in direct-expansion evaporators; they also protect the compressor against intake of liquid refrigerant. As a rule, refrigerant vapor is led through internally finned tubes, while liquid moves through the baffled shell-side space of the apparatus in crosscurrent – countercurrent flow.

Filters in the refrigerant line keep the compressor and control valve free of solid contaminants. Halocarbon refrigerant loops include *dryers,* which prevent refrigerant degradation and icing in expansion valves on expansion below 0 °C. Refrigerating units must have *pumps* to transport cooling water, brine, or other secondary refrigerant, as well as primary refrigerant. Virtually all the pumps used are centrifugal. Whereas open pumps are commonly employed for secondary refrigerants, hermetically sealed (canned-motor) pumps are usually found in primary-refrigerant service. Because of the low static pressure differences to be overcome in finned air chillers or air-cooled and evaporative condensers, the *fans* used with these units are generally axial-flow types. Some air-conditioning evaporators and evaporative condensers also have radial-flow fans. To avoid cold losses, all components that operate under evaporating pressure and are located in ambient-temperature areas are provided with *insulation*. Insulating materials include cork, cellular glass, foamed polystyrene, and polyurethane. The most economical insulation thickness depends strongly on refrigerant and ambient temperatures; it should always be sufficient to prevent sweating. *Vapor locks* must be included to prevent the diffusion of water vapor.

1.2. Absorption Refrigerating Systems

[3], [8], [14], [27], [30], [97–106]

Absorption refrigeration systems differ from compressor refrigerating systems in two main respects: (1) The refrigerant vapor drawn out of the evaporator is compressed not by a mechanical compressor but by a thermal one, which consists of an absorber, a solution pump, and a generator. (2) Operation of the thermal compressor requires a solvent (a second working fluid) as well as a refrigerant. The most important refrigerant – solvent pairs are:

1) *ammonia – water*, used for evaporating temperatures between +10 and −60 °C, and
2) *water – lithium bromide solution*, used for evaporating temperatures >0 °C (air-conditioning systems).

Although many other pairs have been suggested and studied, none is widely used.

1.2.1. Ammonia – Water Absorption Refrigerating Systems

Figure 31 is a schematic diagram of a single-stage ammonia – water absorption refrigerating system.

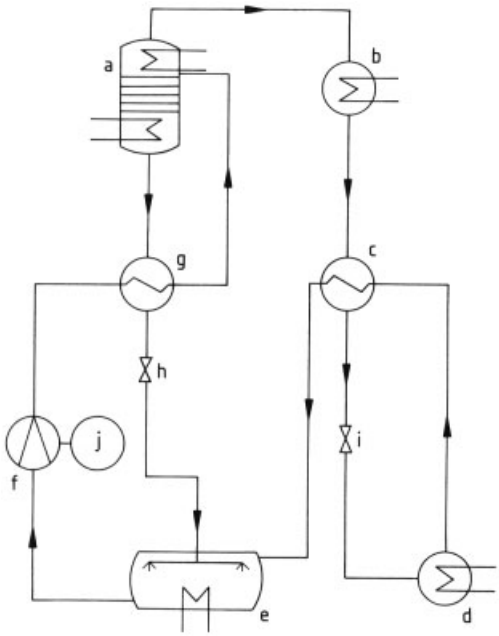

Figure 31. Single-stage absorption refrigerating system
a) Generator with distillation section and reflux condenser; b) Condenser; c) Heat exchanger; d) Evaporator; e) Absorber; f) Solution pump; g) Solution heat exchanger; h) Expansion valve for weak solution; i) Expansion valve for ammonia; j) Solution pump drive motor

In the generator (a), ammonia is driven out of an ammonia-water solution by heating under high pressure. The refrigerant is distilled in a column, usually located inside the generator, and condensed (b). The condensate is subcooled (c) by cold vapor from the evaporator, then expanded to the evaporating pressure and fed to the evaporator (d), where it absorbs heat and evaporates. From there, the refrigerant vapor passes through the heat exchanger (c) to the absorber (e). The weak (ammonia-poor) solution in (e) absorbs the refrigerant vapor, and the resulting heat is rejected to cooling water. The solution, now more concentrated, is pumped through the solution heat exchanger (g) to the generator. In the solution heat exchanger, the enriched cold solution from the absorber is brought into thermal contact with the warm weak solution, which flows from the generator to the absorber through an expansion valve.

A thermodynamic description of the process in an absorption system involves the relationship of pressure p, temperature T, and concentration ξ. Figure 32 shows curves of vapor pressure vs. temperature at varying concentrations ξ. The curves are straight lines in the coordinates chosen. The operating points of the most important devices can be plotted on such a diagram. Condensation and evaporation are each represented by a single point, because they proceed at a constant concentration $\xi \approx 1$, whereas generation and absorption take place in the concentration range bounded by the rich and weak solutions (ξ_r and ξ_w).

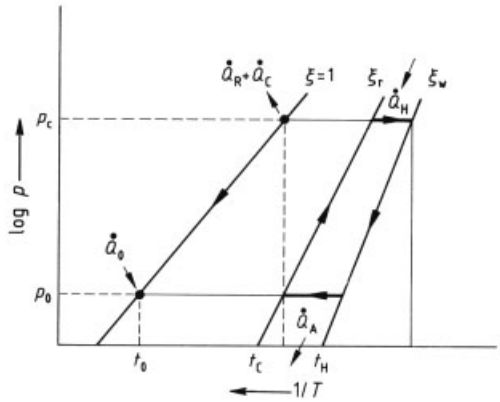

Figure 32. Diagram of absorption refrigerating process ∗ (log p vs. $1/T$)
t_0 = Evaporating temperature; t_c = Condensing temperature; t_H = Generator final temperature
∗ See the text for definitions of ξ_r, ξ_w, $\dot{Q}_0$, $\dot{Q}_c$, $\dot{Q}_H$, $\dot{Q}_A$, and $\dot{Q}_R$.

The $h-\xi$ diagram devised for ammonia – water by MERKEL and BOŠNJAKOVIĆ is useful for thermotechnical determinations, because in this plot quantities of heat appear as lengths.

The process is specified in terms of three temperatures:

1) the generator final temperature t_H, which is determined by the heating medium;
2) the condensing temperature t_c, which is determined by the cooling water; and
3) the evaporating temperature t_0, which is dictated by the application.

The remaining quantities p_c, p_0, ξ_r, and ξ_w are thus defined.

The heat balance of the absorption refrigeration process is as follows:

$$\dot{Q}_H + \dot{Q}_0 + P_P = \dot{Q}_c + \dot{Q}_A + \dot{Q}_R$$

where $\dot{Q}_H$ is the rate of heat supply to the generator; $\dot{Q}_0$ is the refrigerating effect supplied to

the evaporator; P_P is the drive power of the solution pump; and $\dot{Q}_A$, $\dot{Q}_c$, and $\dot{Q}_R$ are the rates of heat removal in the absorber, condenser, and reflux condenser, respectively. In many cases, P_P can be neglected in comparison with the other quantities.

Instead of the coefficient of performance for the compression system, the heat ratio ($\dot{Q}_0/\dot{Q}_H$) is used here for thermodynamic assessment:

$$\zeta = \frac{\dot{Q}_0}{\dot{Q}_H + P_P} \approx \frac{\dot{Q}_0}{\dot{Q}_H}$$

This ratio indicates how much refrigerating effect can be accomplished by a given thermal power supplied to the generator at the evaporating temperature t_0. The heat ratio for today's large-capacity ammonia – water absorption systems is ca. 0.7 when $t_0 \approx 0\,°C$; between 0.47 and 0.40 for t_0 from -20 to $-30\,°C$; and between 0.34 and 0.25 for t_0 from -40 to $-50\,°C$.

In the case of compressor refrigerating systems, economic considerations dictate a shift to two-stage or multistage operation above a certain compression ratio that depends on the compressor design. By contrast, two-stage absorption systems are employed when the temperatures of the heating medium, the cooling water, and the evaporating temperature (which determine the concentration difference between the weak and rich solutions) take on values so unfavorable that this concentration difference becomes $<5\,\%$. In two-stage absorption systems, the two solvent loops are connected in series; the pressure difference between generator and evaporator is divided into two steps.

Changes in the process (e.g., the use of multistage absorption, the injection of cold rich solution into the distillation column, or multistage desorption) lead to further improvement in the heat ratio and to a reduction in the necessary temperature of the heating agent.

Absorption refrigeration systems can be driven by relatively low-grade heat, which is often available from industrial processes. The process can then be economically optimized for given heating media and cooling-water temperatures.

Alteration of the solution flow rate and the heating capacity allows infinitely variable control down to the lowest heat-removal rate. Up to ca. 10 % of the refrigerating capacity, a proportional decrease of the consumption rates of heating agent and cooling water can be expected.

The most important advantage of absorption refrigerating systems is that they can utilize waste heat. However, the following points are also important for the use of these systems: safety in operation, good part-load performance, adaptability to changing service conditions, and low noise emission. Investment costs and energy consumption are higher than for comparable compressor refrigerating systems, but the low price of waste heat can often be crucial in the decision to use an absorption system.

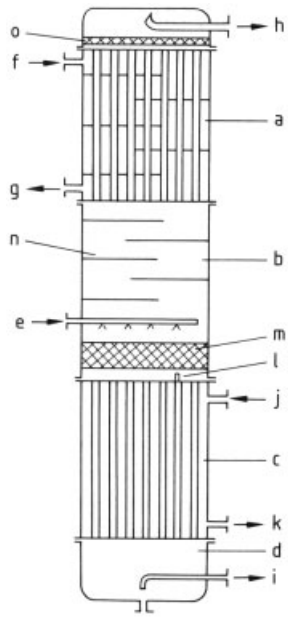

Figure 33. Steam-heated vertical sprayed generator
a) Reflux condenser; b) Distillation section; c) Heating system; d) Solution vessel; e) Rich solution inlet; f) Coolant inlet; g) Coolant outlet, h) Ammonia vapor outlet; i) Weak solution outlet; j) Inlet for heating steam; k) Condensate outlet; l) Guide inserts; m) Packed bed; n) Trays; o) Separator

1.2.2. Equipment for Ammonia – Water Refrigerating Systems

Absorption refrigerating systems employing water – lithium bromide solution are used primarily in air-conditioning systems, because this combination of substances is nontoxic. These systems, usually factory-assembled, with capacities of 200 – 6000 kW, are distinguished by

compact design, noise- and vibration-free operation, and simple, continuous capacity control. Their drawbacks are relatively high energy consumption, heavy demand for cooling water, and cost of anticorrosion measures in various parts of the system.

Generator. The generator is most commonly heated by steam, off-gas, or liquids. Either horizontal or vertical flooded or sprayed tube systems (usually in a shell) are used. The generator is ordinarily combined into a unit with a distillation column and a reflux condenser (dephlegmator). In rare cases, the generator is fired directly with natural gas, oil, or kerosene (multipass combined flue and smoke-tube principle).

Figure 33 shows a schematic diagram of a vertical sprayed generator. The device consists of a solution vessel (d), a heating system (c), a distillation section (b), and a reflux condenser (a). The countercurrent motion of steam and solution creates good conditions for heat and mass transfer, and heat-transfer coefficients of ca. 2800 $W\,m^{-2}\,K^{-1}$ are achieved.

Rich solution coming from the solution heat exchanger enters the packed bed (m), where it absorbs part of the water from the ascending steam. The descending solution then encounters the upper tube sheet of the heating system (c); it passes through guide inserts to the inside of the tubes, where it is degassed. The weak solution is caught in a collector and fed back to the absorber. The expelled steam is rectified by rich warm solution in a packed bed, then by reflux from the condenser in the distillation section (b). Heat removal in the reflux condenser can be accomplished with cooling water or with cold rich solution.

Absorber. The absorber is usually a horizontal shell-and-tube device (see Fig. 34). Several such units can be placed one above another depending on the capacity requirement. A spray device (e) over the tube bundle (f) supplies weak solution to the outside of the tubes while cooling water flows through them. The refrigerant vapor to be absorbed enters the shell-side space from several points over the length of the apparatus.

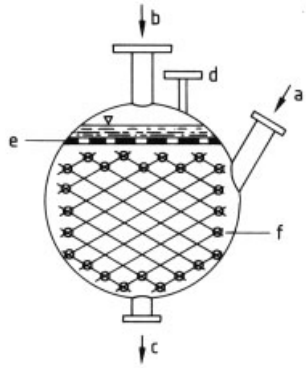

Figure 34. Cross section of a shell-and-tube absorber employing the spray principle
a) Ammonia vapor inlet; b) Weak solution inlet ; c) Rich solution outlet; d) Vent connection; e) Spray device; f) Tube bundle

Another shell-and-tube design, which has the advantage of requiring little floor space, is the *vertical sprayed absorber* (tower absorber). The weak solution is supplied through a distributing system to the vertical tubes, while cooling water flows through the tubes. The refrigerant vapor flows upward countercurrently to the solution.

Heat-transfer coefficients for these absorber designs are between 400 and 800 $W\,m^{-2}\,K^{-1}$, depending on the operating conditions.

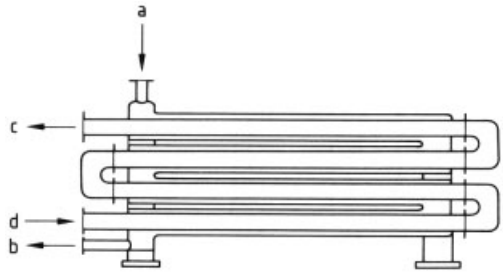

Figure 35. Solution heat exchanger
a) Rich solution inlet; b) Rich solution outlet; c) Weak solution outlet; d) Weak solution inlet

Solution Heat Exchanger. Weak solution is brought into thermal contact with rich solution in the countercurrent solution heat exchanger. Double-pipe heat exchangers are used for low transfer rates, and shell-and-tube devices for higher rates. The weak solution is usually inside the tubes, whereas the rich solution is transported through the shell-side space by a pump.

Figure 35 shows a high-capacity solution heat exchanger, consisting of several shell-and-tube units mounted one above another.

1.2.3. Water – Lithium Bromide Absorption Refrigerating Systems

The condenser and evaporator designs used in absorption refrigeration systems are largely the same as those employed in compression systems (Sections 1.1.3 and 1.1.4). Therefore, only the generator, absorber, and solution heat exchanger are discussed in this section. Other equipment worth mentioning includes extra heat exchangers, refrigerant and solution receivers, capacity controls, ventilation devices, and expansion valves.

Figure 36 is a schematic diagram of a lithium bromide absorption refrigerating system. The principal components are two vessels mounted one above the other. The lower vessel (a) contains the evaporator (c) and the absorber (d). The generator (e) and the condenser (f) are located inside the upper vessel (b). Because the system operates under a high vacuum, virtually no heat-transfer losses occur between the components in each vessel, despite the large temperature differences. A pump (g) delivers rich solution through the heat exchanger (h) to the generator, where part of the water is driven out of it.

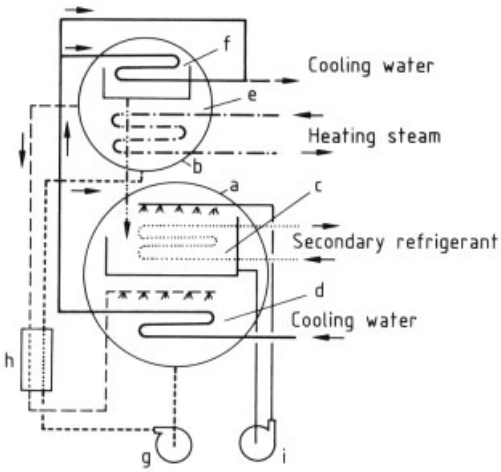

Figure 36. Water – lithium bromide absorption system
a) Lower vessel; b) Upper vessel; c) Evaporator; d) Absorber; e) Generator; f) Condenser; g) Pump; h) Heat exchanger; i) Pump

The weak, high-temperature solution is recycled through the solution heat exchanger to the absorber, where the vapor from the evaporator is absorbed by cooling water. Steam produced in the generator is condensed in the condenser; the condensate flows to the evaporator, where it absorbs heat from the secondary refrigerant. Another pump (i) circulates refrigerant to spray the evaporator tubes; the refrigerant vapor produced is absorbed by the weak solution in the bottom part of the vessel. An inhibitor (usually lithium nitrate or lithium chromate) is added to the loop to combat corrosion.

Lower capacity water – lithium bromide absorption systems have attracted interest in connection with solar energy utilization.

1.3. Steam-Jet Systems [3], [5], [30], [107], [108]

The steam-jet refrigeration system can be regarded as a thermally driven compression system in which vapor is compressed in a jet apparatus. Its operation is illustrated in Figure 37.

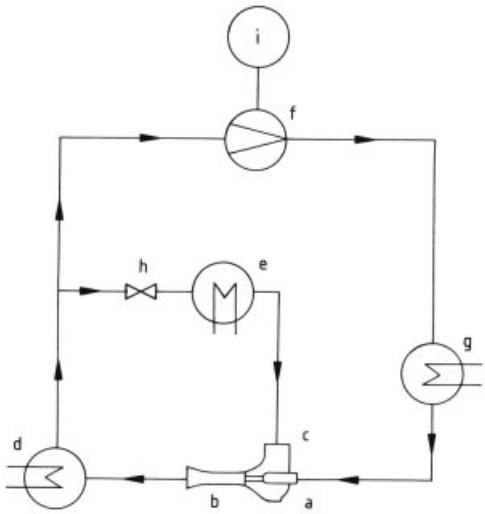

Figure 37. Steam-jet refrigeration system
a) Nozzle; b) Diffuser; c) Mixing chamber; d) Condenser; e) Evaporator; f) Pump; g) Steam boiler; h) Expansion valve; i) Pump drive motor

Steam is generated at pressure p_1 in boiler (g); the mass flow rate of steam is $\dot{m}_1$. In the nozzle (a), the steam is expanded to the pressure p_0 prevailing in the evaporator (e); then, in the

mixing chamber (c), it is mixed with steam (rate $\dot{m}_0$, pressure p_0) from the evaporator. Next, the pressure of the two streams $\dot{m}_1$ and $\dot{m}_0$ is raised to the condensation pressure in the diffuser (b). Virtually all the kinetic energy is transformed to pressure. When heat is withdrawn in the condenser (d), the steam condenses. The condensate is divided into two substreams $\dot{m}_1$ and $\dot{m}_0$; $\dot{m}_1$ is delivered to the steam boiler (g) by a pump (f), and $\dot{m}_0$ is led to the evaporator through an expansion valve (h). Here, substream $\dot{m}_0$ evaporates at pressure p_0, absorbing heat; the exiting expanded steam, moving at a higher velocity, sucks this stream into the mixing chamber.

Water, employed mainly as working steam and as refrigerant in the steam-jet system, is a low-cost, hazard-free working fluid with a high specific enthalpy of evaporation. The pressure p_0 is usually between 0.6 and 1.2 kPa; p_c depends on the cooling-water temperature and lies between 3 and 8 kPa.

This system's drawbacks are the large specific volume in the region of the evaporating temperatures, the high vacuum, and the limitation placed on the evaporating temperature by the freezing point. Lower temperatures can be reached if hydrocarbons and halocarbon refrigerants are used as working fluids.

As in absorption systems, the decisive quantity for assessing this process in energy terms is the *heat ratio* ζ, which indicates how much refrigerating effect $\dot{Q}_0$ is gained from the heating power $\dot{Q}_H$ and the pump power P_p supplied (see Section 1.2.1). The heat ratios attained in practice are much lower than those for an absorption system. With multistage jet units and multistage evaporation, the energy performance can be improved.

The advantages of the steam-jet system are low plant costs due to the simple design, utilization of waste heat, small space requirement, and low maintenance costs. These are partly offset by the drawback of high operating costs, especially because of the large consumption of cooling water. These features make the steam-jet process suitable for short-term refrigeration using cheap waste heat.

1.4. Thermoelectric Refrigeration [3], [28], [109–112]

Thermoelectric refrigeration is based on the *Peltier effect*, which is the reversal of the Seebeck principle. In an electrical circuit consisting of two different conductors and a d.c. voltage source, one of the junctions between the two conductors is heated while the other is cooled. If the direction of the current is changed, the warm and cold junctions change places. Energy is lost by Joule heating in the conductors, because of their resistance, and by conduction from the warm junction to the cold one; these losses greatly reduce the coefficient of performance. The maximum temperature difference that can be produced between the junctions and the coefficient of performance are primarily determined by the properties of the two conducting materials: thermoelectric force e, electrical conductivity σ, and thermal conductivity λ. These are grouped into a parameter called the *figure of merit* z:

$$z = \frac{e^2 \cdot \sigma}{\lambda}$$

The thermoelectric force and the electrical conductivity should be as high as possible, while the thermal conductivity should be as small as possible. This requirement is best met by alloys of semiconducting materials, such as lead – tellurium, antimony – bismuth, and germanium – silicon (→ Semiconductors).

Because the capacity of a single Peltier element is very low, a number of elements must be connected in a thermally parallel manner so that their refrigerating effects are additive. Figures of merit z of $(3.0-3.5)\times 10^{-3}\,\mathrm{K}^{-1}$ can be attained with current semiconductor technology, so the available coefficients of performance are comparable with those of other processes only at very small refrigerating capacities.

Thermoelectric refrigeration is used in special cooling devices for laboratory purposes and in low-capacity air-conditioning units.

2. Industrial Applications of Refrigeration [10–13], [21], [31]

Many industrial chemical processes require low temperature, often in conjunction with large re-

frigerating capacity. Examples of design of refrigerating plants for these applications are described in this chapter. Machines and equipment for ice making, food technology, climate control, and construction are also discussed.

Table 5. Important data for calcium chloride brines

Density at 15 °C, kg/m^3	Salt content, kg $CaCl_2$/ 100 kg solution	Start of freezing, °C	Specific heat at 0 °C, kJ kg^{-1} K^{-1}
1000	0.1	0.0	4.20
1010	1.3	− 0.6	4.13
1020	2.5	− 1.2	4.05
1030	3.6	− 1.8	3.98
1040	4.8	− 2.4	3.90
1050	5.9	− 3.0	3.83
1060	7.1	− 3.7	3.76
1070	8.3	− 4.4	3.69
1080	9.4	− 5.2	3.62
1090	10.5	− 6.1	3.56
1100	11.5	− 7.1	3.50
1110	12.6	− 8.1	3.44
1120	13.7	− 9.1	3.38
1130	14.7	−10.2	3.33
1140	15.8	−11.4	3.27
1150	16.8	−12.7	3.22
1160	17.8	−14.2	3.17
1170	18.9	−15.7	3.13
1180	19.9	−17.4	3.08
1190	20.9	−19.2	3.04
1200	21.9	−21.2	3.00
1210	22.8	−23.3	2.96
1220	23.8	−25.7	2.93
1230	24.7	−28.3	2.90
1240	25.7	−31.2	2.87
1250	26.6	−34.6	2.84
1260	27.5	−38.6	2.81
1270	28.4	−43.6	2.78
1280	29.4	−50.1	2.75
1286 *	29.9	−55.0	2.74
1290	30.3	−50.6	2.72
1300	31.2	−41.6	2.70
1310	32.1	−33.9	2.67
1320	33.0	−27.1	2.65
1330	33.9	−21.2	2.62
1340	34.7	−15.6	2.60
1350	35.6	−10.2	2.58
1360	36.4	− 5.1	2.55
1370	37.3	0.0	2.53

* Eutectic

2.1. Chilling of Liquids [18], [24], [26], [28], [30], [33], [113–117]

Liquid chilling usually refers to the cooling of secondary refrigerants in order to transfer cold to other products. The use of secondary refrigerants is suitable when cold is to be distributed to several consumer locations or when contact between primary refrigerant and product resulting from leaks must be avoided. The selection of a secondary refrigerant depends on temperature, heat-transfer properties, material compatibility, toxicity, and price. Figure 38 shows the temperature ranges for the most important secondary refrigerants; the shaded part of each range should be avoided whenever possible, because of poor heat-transfer properties. The favorable qualities of water mean that it is almost always preferred when the evaporating temperature of the refrigerant is > 0 °C. Down to -40 °C, calcium chloride brines are commonly used with steel; ethylene glycol – water mixtures are used more and more because of their good behavior with respect to ferrous and nonferrous metals.

The concentrations of aqueous mixtures are chosen so that the freezing point lies several degrees below the evaporating temperature. Tables 5 and 6 list important data for calcium chloride brines and ethylene glycol – water mixtures.

Table 6. Important data for ethylene glycol – water mixtures

Density at 15 °C, kg/m^3	Glycol content, kg glycol/100 kg solution	Start of freezing, °C	Specific heat at 0 °C, kJ kg^{-1} K^{-1}
1005	4.6	− 2	4.10
1010	8.4	− 4	4.06
1015	12.2	− 5	3.98
1020	16.0	− 7	3.89
1025	19.8	−10	3.85
1030	23.6	−13	3.77
1035	27.4	−15	3.73
1040	31.2	−17	3.64
1045	35.0	−21	3.56
1050	38.8	−26	3.52
1055	42.6	−29	3.43
1060	46.4	−33	3.35

Equipment. With few exceptions, chilling of the secondary refrigerant or other liquid takes place in shell-and-tube heat exchangers. Direct-expansion-type evaporators are employed with halocarbon refrigerants at capacities of up to ca. 1000 kW. At higher capacities and with ammonia as refrigerant, flooded shell-and-tube evaporators are used.

Liquid-chilling machines for temperatures > 0 °C, with treated water as secondary refrigerant, are widely used, especially for air condi-

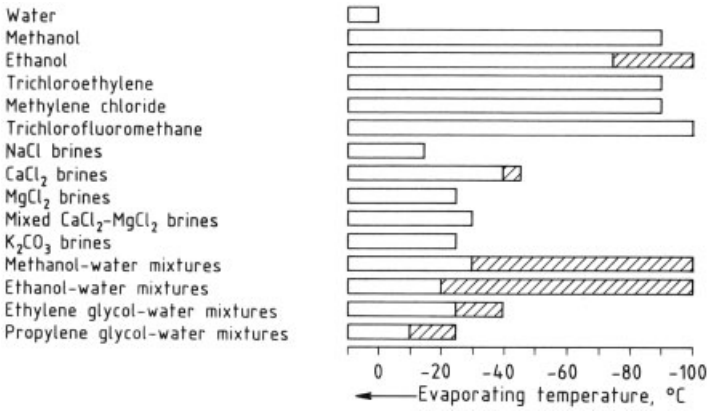

Figure 38. Application ranges of secondary refrigerants
Hatching indicates region of high viscosity

tioning, but also for manufacturing processes, food technology, cooling plastics machinery, etc. Nearly all of these units operate with halocarbon refrigerants.

Plants with reciprocating compressors are in service at refrigerating capacities up to ca. 1000 kW; above this range, centrifugal compressors are used most widely. Compact systems of this type are made for capacities up to ca. 5000 kW. These water chillers have compressor, evaporator, condenser, control, safety devices, and electrical switching; they are factory-assembled on a single frame, so that setup and commissioning are simple. Water- or air-cooled condensers are employed; in large air-cooled plants, the air-cooled condensers may be very large and, so separate placement of the water-chilling unit and the condenser (erected outdoors) is common. Virtually all water chillers with reciprocating compressors use R 22 as refrigerant.

Hermetic compressors are mainly employed at low refrigerating capacities up to ca. 50 kW, whereas semihermetic motor-driven compressors are used at higher capacities. Because air-conditioning service means frequent part-load operation, the refrigerating capacity is usually distributed between several smaller compressors, which can be turned on or off as needed. Larger units are also made with two separate refrigerant loops to enhance reliability.

Water chillers with screw compressors also use R 22 as their principal refrigerant. These units are distinguished by quiet operation and high reliability. Like reciprocating plants, they can be adapted to varying discharge pressures; this is important for air-cooled chillers. The compressors are driven by bipolar three-phase motors. In hermetic or semihermetic motor – compressors, the oil stream separation required on the discharge side of screw compressors can take place inside the motor space, so that compact construction is possible. In systems with a single compressor, capacity control is exerted with a slide valve; with several compressors, individual machines are turned on and off as in the case of reciprocating compressors.

Most centrifugal units use R 11 or R 12. The rotational speeds required are greater than bipolar three-phase motors can supply, so builtin gearing or frequency converters are employed. Even in large-capacity systems, semihermetic mounting of the drive motor and compressor is common. Both single-stage and two-stage units are built; the two-stage design with intermediate expansion offers somewhat better energy economy. Water chillers with centrifugal compressors usually feature water cooling, because of their high capacity and the inflexibility of centrifugal compressors with respect to varying discharge pressures. Whereas evaporators in systems with reciprocating and screw compressors operate in direct-expansion mode, centrifugal systems employ flooded evaporators. Usually a unit has just one compressor; part-load control is exerted by inlet guide vanes, the less economical bypass control being used only at cooling loads much lower than half the maximum capacity.

Units for chilling brine and other secondary refrigerant liquids are essentially similar in design to water chillers.

Very viscous or pasty liquids are chilled in screw or scraped-pipe coolers, whose walls are held at the required low temperature by evaporating primary refrigerant or by a secondary refrigerant. Such devices are frequently used in the food and chemical industries.

2.2. Separation of Liquids by Crystallization [13], [21], [118–123]

Recovering some products and separating materials from mixtures may require low process temperature, because of either the position of the thermodynamic equilibrium (solubility curves) or the temperature sensitivity of some substances.

For example, *p*-xylene is recovered from the mixture of ortho, meta, and para isomers by crystallization in a scraped-pipe heat exchanger (→ Xylenes); the process temperature is ca. − 70 °C. The xylene mixture flows inside a cylinder, while a stream of secondary refrigerant flows through the jacket. Scraper blades rotating inside the cylinder continuously remove the crystals precipitated on the wall and thus ensure constant heat-transfer conditions. Cold is produced in cascade-type refrigerating systems.

Figure 39 is a flow sheet of a cascade unit for chilling methane (secondary refrigerant) to − 77 °C. Both loops use three-stage centrifugal compressors; the high-temperature loop operates with R 12 and the low-temperature loop with R 13.

Another example is the dewaxing of crude oils. Alkanes left in the crude oil after distillation are crystallized by cooling to ca. −30 °C. The crystallizers, operating in several temperature stages, are similar in design to those used for xylene separation. The refrigerating systems are usually multistage ammonia compressor refrigerating systems or ammonia – water absorption refrigerating systems. Low-temperature fractional crystallization is also used to separate undesirable byproducts, especially when an impurity has a boiling point close to that of the desired product.

2.3. Cooling and Liquefaction of Gases [13], [21], [124–130]

Except in room air-conditioning systems (see Section 2.5), where air is used as secondary refrigerant, the purpose of cooling gases and gas mixtures is nearly always to bring about total or fractional condensation. The fractional condensation of gas mixtures effects separation; the desired product may be either the residual gas or the condensate.

A number of industrially important gases are produced by condensation from crude gases. The condensation temperature depends on the concentration of the gas in the mixture or on its partial pressure. The partial pressure can be raised by compressing the crude gas. The recovery (the ratio of gas condensed to gas in the feed stream) and the concentration of the gas in the feed are the chief factors dictating the temperature to which the crude gas must be cooled. Subsequent storage or transport of the condensed gas also determines the condensation temperature. Chlorine, ammonia, sulfur dioxide, and carbon dioxide are examples of important products recovered by condensation at subambient temperatures.

Chlorine is recovered between + 20 and − 50 °C, depending on the production process. Chlorine gas from electrolysis of sodium or potassium chloride contains, after preliminary treatment, 2 – 8 vol % air, hydrogen, carbon dioxide, and carbon monoxide (→ Chlorine, Chap. 11.). The choice for chlorine separation among the following three candidate processes depends on the chlorine content in the feed stream and the desired recovery:

1) The crude gas inlet at 0.1 MPa is condensed at this pressure. Multistage refrigerating systems may be needed (depending on the recovery) for evaporating temperatures at or below −45 °C.
2) The crude gas is compressed to ca. 0.3 MPa in nonlubricated reciprocating compressors (or centrifugal compressors when the quantity of gas is large). The increased partial pressure enables condensation to be carried out at evaporating temperatures of ca. − 15 to − 20 °C in the refrigerant loop.

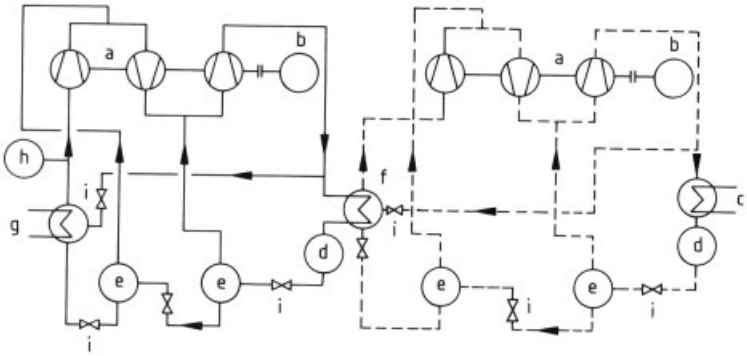

Figure 39. Flow sheet of cascade centrifugal refrigerating system for xylene separation
a) Centrifugal compressor; b) Drive motor; c) Water – cooled condenser; d) Receiving vessel; e) Expansion vessel; f) R 12 evaporator – R 13 condenser; g) Methanol cooler; h) Vessel for thermal expansion of R 13; i) Expansion valve
– – – R 12 Loop; ⟶ R 13 loop

3) The crude gas is compressed to 1.0 – 1.5 MPa in several stages and can then be condensed with water; no refrigerating machine is needed.

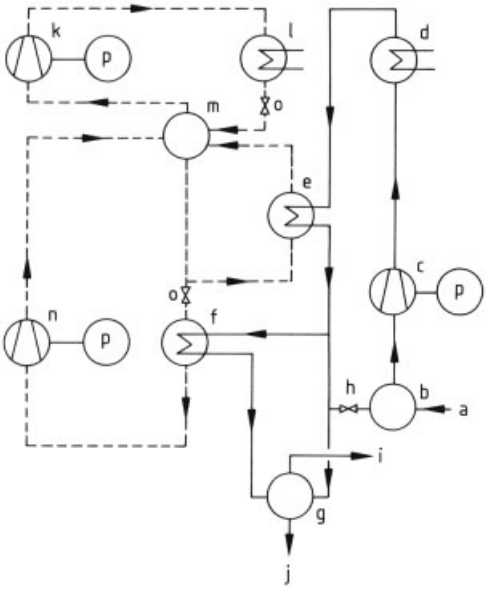

Figure 40. Flow sheet of a chlorine liquefaction system
a) Crude gas inlet; b) Precooler; c) Chlorine gas compressor; d) Water-cooled chlorine gas cooler; e) First-stage chlorine condenser; f) Second-stage chlorine condenser; g) Vessel for liquid chlorine; h) Expansion valve; i) Residual gas outlet; j) Liquid chlorine outlet; k) High-stage compressor; l) Condenser; m) Medium-pressure vessel; n) Low-stage compressor; o) Expansion valve; p) Drive motor
⟶ Chlorine loop; - - - - Refrigerant loop

These processes are often linked. For example, part of the chlorine gas from the compressed crude gas can be condensed with water; the remaining gas is then cooled to ca. −40 °C and another fraction condenses. The recovery, depending on feed-gas composition, is >99 vol %.

Figure 40 is a flow sheet of a *chlorine liquefaction system.* The crude gas is first treated to remove water vapor and other gases, then led through a precooler (b). Here, evaporating chlorine cools the gas to ca. − 30 °C so that, on subsequent compression to ca. 0.4 MPa (c), the final temperature does not exceed 80 °C, the allowed maximum on safety grounds. The compressed gas is first cooled against water in a heat exchanger (d). It then goes to the first chlorine condenser (e), which is cooled by evaporating primary refrigerant. The condensed fraction flows to a receiving vessel (g). The uncondensed fraction continues to a second chlorine condenser (f), where it is collected at a lower temperature and led to the vessel.

Ammonia can be condensed from converter gas in low-temperature refrigerating systems (→ Ammonia). Evaporating temperatures of the two-stage refrigeration process are between −50 and − 60 °C.

In the separation of *sulfur dioxide*, the feed gas is first compressed to ca. 0.5 MPa; up to one half of the initial sulfur dioxide content is recovered by condensation at − 40 °C (→ Sulfur Dioxide). The gas is then expanded to ca. 0.1 MPa, the gas temperature drops to − 50 °C, and the remaining sulfur dioxide condenses.

Carbon dioxide for the beverage industry and for dry-ice production is purified by scrubbing and adsorption, then condensed at low temperatures and stored (→ Carbon Dioxide). Evaporating temperatures are ca. − 30 to − 35 °C.

For *tank storage and transport*, gases are liquefied and recondensed with cold-vapor refrigeration processes, provided their standard boiling points are not much lower than − 100 °C. This group includes ethane, ethylene, propane, propene, butane, butadiene, ammonia, and others. Gases such as ethane and ethylene, which have critical pressures near or below ambient temperature, are liquefied in the evaporator of a closed refrigerating system resembling a cascade system. The process takes place at elevated pressure between − 50 and − 30 °C. Figure 41 is a flow sheet of a plant that recondenses ethylene for storage at atmospheric pressure. Higher boiling substances are condensed either by compression and cooling against water or air, or in a closed refrigerating system at storage pressure. Recondensing units with gas compression represent a possible application for nonlubricated reciprocating compressors, because the product must not be contaminated with lubricating oil.

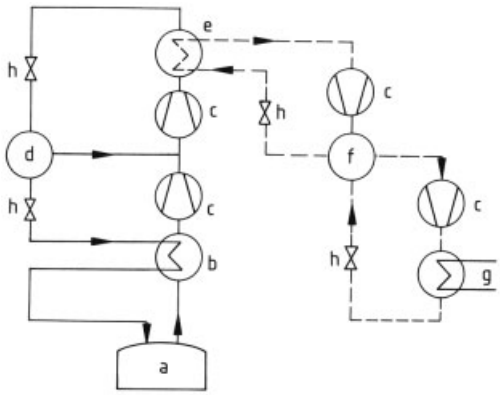

Figure 41. Flow sheet of an ethylene recondensing system
a) Ethylene tank; b) Heat exchanger; c) Compressor; d) Expansion vessel; e) Ethylene condenser – R 22 evaporator; f) Medium-pressure vessel; g) Water-cooled condenser; h) Expansion valve
– – – R 22 loop; ⟶ Ethylene loop

In the *drying of gases*, particularly of compressed air, the dew point must be lowered to such an extent that no water vapor condenses during later use of the gas. Refrigerating system are often employed from 25 to 5 °C when water cooling is inadequate and adsorptive drying is ruled out for economic reasons. To minimize energy consumption, the dry cold gas can be used to precool the moist gas stream. Smaller units generally employ R 12 and R 22 and thermostatically controlled expansion; for high capacities, ammonia and flooded operation are preferred. Finned cooling systems are used at standard pressure, whereas shell-and-tube exchangers with smooth or shallow finned tubes come into play at higher pressures. If dew points $< +5$ °C are required, the water vapor collects as frost on the heat-exchange surfaces. It must be melted periodically by using electrical heating or pressurized refrigerant vapor. This arrangement calls for two evaporators connected in parallel and operated alternately. To avoid short cycle times, the greatest possible areas are provided on the gas side, so that the frost layer grows slowly. Finned coolers best meet this requirement; in shell-and-tube devices, the air is on the shell side to prevent the tubes from becoming plugged with ice.

In refrigeration systems, the use of condensation to recover valuable feedstock and solvent and to return them to the production process contributes to saving raw materials. To reduce water and air pollution, increasingly lower temperatures (down to − 70 °C) are employed for off-gas treatment and other purposes, in order to remove even slight traces of foreign and hazardous substances by recondensation.

2.4. Remote Cooling [61], [131], [132]

Like steam power plants, central cooling systems in large chemical plants generate needed energy and distribute it through a network of pipes that is often highly branched. Such installations are a special application of centrifugal compressor ammonia refrigerating systems. Even though the properties of ammonia are not favorable for centrifugal compressors (low molecular mass requiring many stages), the low price of ammonia and the small pipe cross sections and heat exchangers make ammonia the most economical refrigerant. The cold consumers, often several hundred meters away, are supplied directly with liquid primary refrigerant; no intermediate secondary refrigerant system is necessary.

Figure 42 is a flow sheet of a remote cooling system with two stages. Before expanding in evaporator (g), the warm liquid refrigerant is subcooled (e), by the vapor coming from the evaporator, and in this way the vapor is heated close to ambient temperature. Consequently, the vapor piping can be left uninsulated with no significant energy loss. Before the refrigerant vapor enters compressor (i), it is saturated with expanded liquid in order to minimize superheating on compression. Vapor coming from the low-stage compressor is intercooled against water in the surface cooler (j), and expansion takes place in several stages. In this way, such installations can be operated economically.

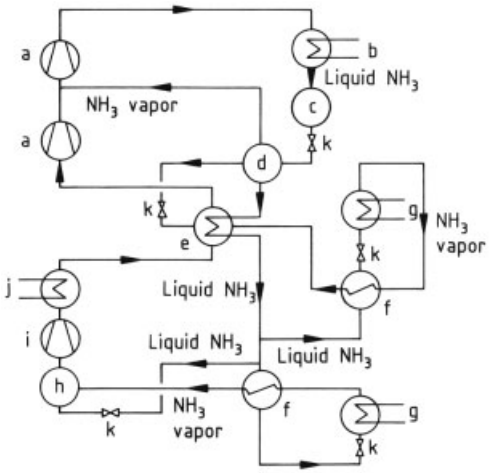

Figure 42. Flow sheet of an ammonia refrigerating system with centrifugal compression and two evaporation stages
a) High-stage compressors; b) Condenser; c) Receiver; d) Medium-pressure vessel; e) Medium-pressure vessel with subcooling coil; f) Heat exchanger; g) Evaporator; h) Saturating vessel; i) Low-stage compressor; j) Surface cooler; k) Expansion valve

2.5. Air Conditioning [13], [21], [24], [25], [29], [31], [133–137]

In many industrial plants, air must be held in some definite state for the manufacture, processing, and storage of products. The necessary conditions are dictated by the need to keep the temperature, humidity, air velocity, and air purity within narrow ranges. Such requirements apply, for example, in the manufacture of textiles, tobacco products, and photographic or electronic equipment; in printing, food, or pharmaceutical plants; in special fabrication processes in the machinery industry; at data-processing facilities; and elsewhere.

Conditioned air consists of fresh and recycled air in some proportion; the cooling surfaces are always $> 0\,°C$. Two main processes are used to cool air in an air-contitioning plant:

1) Cooling and simultaneous humidification or dehumidification in a wet air cooler (air washer)
2) Cooling and simultaneous dehumidification in a surface cooler

When the outside air is polluted, it is purified by filtering, scrubbing, or adsorption before being mixed with recycled air. Water, cooled in the evaporator of a water chiller, is usually employed as secondary refrigerant (see Section 2.1). Many air-conditioning systems incorporate heat recovery to warm the air after dehumidification.

2.6. Refrigeration Plants for Cold Rooms and Freezers [10–12], [15–18], [21], [31], [138–158]

Refrigeration plants for cold rooms are used mainly for the preservation of foodstuffs of plant or animal origin → Foods, 2. Food Technology, Chap. 3.; they can, however, be used to extend the shelf lives of other products, such as chemicals, pharmaceuticals, or flowers. The principal users are cold-storage plants, slaughterhouses, breweries and dairies, refrigerator trucks and railroad cars, and stockrooms in food stores.

Aside from a few special cases, such as the humid air cooling of vegetables or flowers, refrigeration plants for cold rooms use direct cooling by evaporating refrigerant, with no intermediate circulation of a secondary refrigerant.

Storage requirements vary, from 10 to 0 °C for coolers and from − 20 to − 30 °C for freezers; accordingly, refrigerating equipment operates at evaporating temperatures from 0 to − 10 °C (single-stage compression) or from − 30 to − 40 °C (two-stage compression). Two basic types of systems exist:

1) A *centralized unit*, consisting of a large refrigerating system from which the condensed primary refrigerant is distributed to evaporators in the individual chambers

2) *A decentralized system*, in which each chamber has its own refrigerating unit that may have one or several compressors

In large cold-storage plants, slaughterhouses, and breweries, the usual refrigerant is ammonia. In mobile refrigerating systems and in food stores, R 22 or R 502 is preferred. Plants using ammonia are centralized; the liquid refrigerant flows by gravity or is pumped from a receiving and separating tank on the low-pressure side to the individual chamber coolers. The air coolers must be defrosted from time to time; this is usually done by passing hot gas through the refrigerant piping (Fig. 43).

Refrigeration plants for cold rooms operating on R 22 or R 502 can be centralized or decentralized. The air cooler in such a unit is generally a thermostatically controlled direct-expansion evaporator. Decentralized machines are simpler in design, but require a much greater number of refrigerant compressors than centralized systems and also consume more energy in most cases. Whereas hot-gas defrosting is common in centralized plants, electrical defrosting is often used in decentralized systems with direct-expansion evaporators.

The products to be stored may be cooled or frozen in the chambers themselves, but special equipment is usually employed for this purpose. Such equipment lowers the product temperature much more rapidly (see following discussion of " Freezing Processes").

The coolers in a chamber are mounted either along the walls or below the ceiling. Air is often distributed by a system of ducts below the ceiling, whereas it is commonly exhausted from the chamber without any special ductwork. The rate of air circulation depends on the size of the room and the products stored there; an air speed in the chamber of ca. of 0.5 m/s is often selected. Maintenance of a given humidity level in the air, usually between 85 and 95 % R.H., is very important. If the humidity is too high, product shelf life is shortened; if it is too low, the product will dry out. Adequate sizing of the air-cooler surfaces keeps the humidity from dropping too low. The accuracy of temperature and humidity control can be improved by using additional heating and humidifying equipment. Other measures can extend the shelf life of foodstuffs. In fruit storage, for example, the oxygen content of the air is generally lowered (storage under controlled atmosphere).

In cold-storage plants, flat construction is favored because of better transport conditions, even though it involves greater cold consumption and higher land costs. Cold room chamber loading, depending on product and room height, is between 2 and 4 t/m^2 of floor area. Proper insulation of cold-storage rooms is important; the thickness of insulation is matched to the chamber temperature. The insulation must also incorporate a vapor barrier. The costs of insulating a cold-storage plant are often of the same order as those for the entire refrigerating system.

Freezing Processes. *Freezing in liquid brine* (immersion freezing) is appropriate only for foods in liquid-tight packages, such as canned frozen juices and shrink-wrapped poultry. Heat transfer is equally good everywhere on the surface, especially for products with uneven surfaces, so that freezing times are short. The product, such as packaged poultry, is moved by conveyors through a tank containing the brine, a mixture of propylene glycol and water. After leaving the tank, the product must be rinsed with water to remove residual brine.

In *contact freezing*, small unit packages of uniform thickness are frozen in plate-type equipment. The plates are sometimes vertical but most often horizontal. A unit has up to 20 aluminum plates with integrally cast channels for refrigerant or brine. The packages are laid on sheets between the plates, and the plates are pressed against the fixed lowermost plate to produce good contact between the cooled plates and the product. After freezing is completed, the plates must be warmed briefly with a defrosting circuit so that they can be separated.

Liquid, pulpy, or pasty material can be frozen at high rates in a *scraped-surface freezer.* The product is fed onto a slowly rotating horizontal cylindrical surface cooled by primary refrigerant. The devices used for application and removal depend on the product. The form of the frozen product (powdered or flaked) depends on the scraping apparatus and the starting material. The process is continuous and can be adapted to prevailing requirements by varying the roller speed, the evaporating temperature, and the coating thickness.

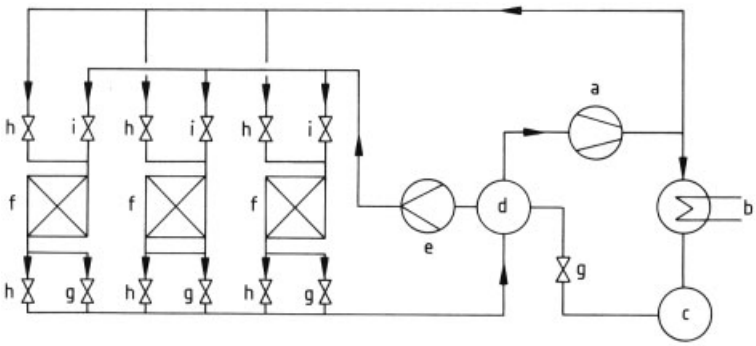

Figure 43. Flow sheet of a refrigerating plant for cold rooms with pumped circulation of refrigerant and hot-gas defrosting ∗ a) Compressor; b) Condenser; c) Receiver; d) Separator; e) Refrigerant circulation pump; f) Air coolers; g) Expansion valve; h) Switchover valve; i) Nonreturn valve
∗ The upper switchover valves are closed and the lower ones opened for cooling; the opposite is true for defrosting.

Cold-Blast Freezing. Freezing in air is suitable for products of any type, size, or form, whether in the loose or packaged state; all frozen products are processed in this way. Many unit designs and ratings are available. Achieving economical freezing rates with air requires greater temperature differences than with contact or immersion freezing. Air temperatures of −30 to − 50 °C are used.

The product is conveyed through the freezing tunnel on wheeled racks, belts, carousels, or slide ways. In this continuous process, the material enters the freezing unit through locks. Multiple fans blow cold air countercurrently or crosscurrently over the product at up to 7 m/s. The air is cooled against refrigerant evaporating at roughly −40 to − 55 °C in finned-pipe heat exchangers. Freezing rates of ca. 1 – 3 cm/h can be attained. The residence or freezing time can be controlled continuously by varying the belt or chain speed or the feed rate.

The refrigerating equipment, usually two-stage, employs ammonia or R 22 as refrigerant. Reciprocating and screw compressors are used.

A special form of cold-blast freezing is the *fluidized-bed process*; this is particularly suitable for free-flowing materials in the form of small lumps or fine grains, which are often encountered in food technology and the chemical industry.

Cold air, blown upward through a metal pan with a perforated bottom, suspends the product and flows around it. The good heat transfer (coefficients of 90 – 300 W m^{-2} K^{-1} for specific surfaces of 0.25 – 1 m^2/kg) results in short freezing times, so that equipment can be made compact and space requirements minimized. The air, which must be cooled to ca. −30 °C usually is generally blown by fans through several finned-tube evaporators. Pressure drops in the evaporators and the fluidized bed must be considered. The optimal bed depth is determined by the type of product being frozen and is controlled through air velocity and particle size. The evaporators are defrosted automatically at intervals, without interrupting the overall process; the refrigerant, usually R 22, is conveyed through the loop by pumping.

Freezing with Boiling Liquids. Two freezing processes involving boiling liquids have found limited application. One employs liquid nitrogen; in the other, boiling R 12 is placed in direct contact with the product.

Liquid nitrogen is sprayed onto the product in a tunnel (Fig. 44). The product is precooled by the cold gas produced by evaporation. Fans (h) circulate the gas around the product to prevent surface cracking, which would otherwise occur if the uncooled product suddenly came in contact with liquid nitrogen (d), which boils at ca. − 190 °C. To avoid excessive residence times and thus overlong tunnels, the large temperature difference between the surface and the interior of the product is not equalized in the tunnel. The product only remains in the tunnel as long as is necessary to remove the heat needed to bring its average temperature down to its desired storage temperature. Final temperature equalization between the core and the surface occurs later dur-

ing frozen storage. With 1 kg of liquid nitrogen (1.245 L), some 330 kJ of heat can be removed from the product. The large temperature difference and good heat transfer allow quick freezing and thus good product quality. Liquid carbon dioxide can be used instead of liquid nitrogen.

With R 12, the liquefied refrigerant is brought into contact with the product, again in a tunnel. The refrigerant evaporates at a temperature of ca. −30 °C. In contrast to nitrogen, R 12 can easily be recondensed with a condenser mounted in the tunnel. The condenser is held at about −45 °C by a two-stage R 22 refrigerating unit. Withdrawing a given amount of heat requires much less energy than with liquid nitrogen, because of the higher evaporating temperature. For economical operation, refrigerant losses resulting from leaks in the tunnel and entrainment with product must not exceed 1 – 2 wt % of product.

Freeze Drying (see also → Drying of Solid Materials, Chap. 2.5.). Frozen-dried products are of better quality than conventionally dried products because of reduced cell damage.

Freeze drying consists of the following steps: (1) freezing and cooling of the product to ca. −20 °C, (2) evacuation to 50 – 100 Pa, (3) addition of heat of sublimation through heated plates, and (4) condensation of water vapor on cold surfaces (−40 to −60 °C). The residual water content is <2 %. Finely divided products are more suitable than coarse ones, because both the freezing time and, more important, the drying time are shorter. The main difficulty is to remove water vapor quickly from the interior of the product without causing local overheating of the exterior. To prevent oxidation during storage, freeze-dried products are packaged under vacuum or in a nitrogen atmosphere.

Cold is produced with multistage compression and absorption machines. The energy consumption is high (10 to 15 times that required for frozen-food production) because the required refrigerating capacity is about eightfold and the evaporating temperatures are 10 – 20 K lower. However, energy costs for three months' storage of frozen food are roughly the same as for freeze drying. Nevertheless, the process is much more expensive than frozen storage, so freeze drying of foods is not very important; it is used increasingly for high-quality food and pharmaceutical products.

Ice Making. Ice, especially in finely divided form such as "tube" or "snow" ice, is used for cooling products during transport (factory ships, refrigerator railroad cars, etc.). It also helps avoid drying. In the chemical industry, finely divided ice is used to absorb the heat of reaction. Ice plays an important part in concrete construction work, where it removes the heat of setting.

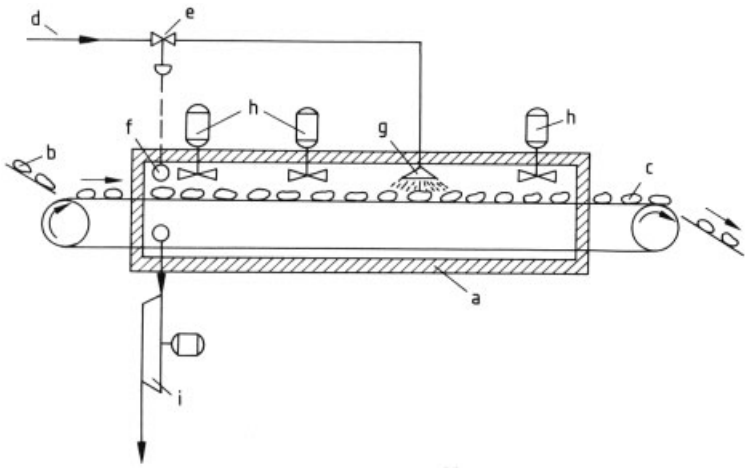

Figure 44. Quick-freezing apparatus with liquid nitrogen
a) Freezing tunnel with belt conveyor; b) Product feed; c) Product removal; d) Liquid nitrogen feed from supply tank; e) Control valve; f) Temperature sensor; g) Spray nozzle; h) Circulating fans; i) Suction blower

A further application is in cold storage. Large amounts of ice are often stored in nuclear reactor buildings, so that radioactive steam can be completely condensed on the stored ice if the steam pipes rupture. The storage of cold also allows better use of refrigerating capacity when cold demand is variable. Ice is frozen on a refrigerant evaporator (coiled-pipe or plate evaporator) located in an open, water-filled vessel. The water, which is close to 0 °C, serves mainly as a secondary refrigerant in food production, especially in dairies. When the demand for cold rises, the ice layer frozen on the evaporator coils is melted, providing additional cooling capacity. Ice-storage units are used in air conditioning, where ice produced with cheap nighttime power is used to meet peak daytime loads. In this way, refrigerating systems can be made smaller, power-connected load can be reduced, and energy costs can be lowered.

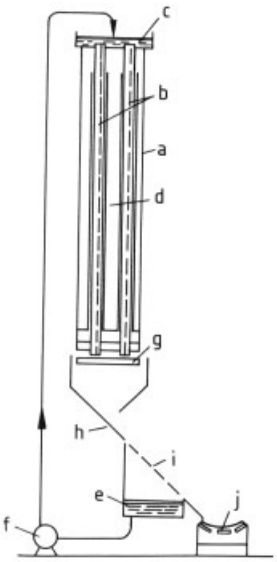

Figure 45. Schematic of a tube ice maker
a) Shell-and-tube heat exchanger; b) Freezing tubes; c) Upper water tank; d) Tube-side space of exchanger; e) Lower water tank; f) Water pump; g) Ice breaker; h) Ice chute; i) Draining screen; j) Belt conveyor

Tube Ice. Ice in short tubular pieces (diameter ca. 40 mm, wall thickness 8 mm) is produced in tube ice makers. Figure 45 shows such a machine. In a vertical shell-and-tube device, refrigerant evaporates in the shell-side space, while water from a tank flows through distributors and down the inside of the tubes. Some of the water freezes; the rest goes to a receiver, where a float controller admits fresh water to make up for the amount frozen. About twice an hour, a timer cuts off the water flow and admits pressurized gas from the condenser to start the defrosting process. The ice released from the tubes is broken, then passes over a draining screen and through a chute to a belt conveyor, which carries it to an ice hopper or directly to the consumer. For a tube ice maker operating with ammonia (evaporating temperature ca. − 12 °C) and having a capacity of ca. 400 kg/h, the specific cold consumption is ca. 780 kJ/kg of ice.

Flaked Ice. Ice in the form of thin flakes is made continuously in machines where freezing takes place on the inside or outside surface of a cylinder. Water is sprayed onto the inner surface of the cylinder from a revolving distributor; the cylinder is cooled from the outside by refrigerant evaporating in a jacket. The ice is removed by a revolving scraper. In another design, the cylinder is cooled from the inside by brine or evaporating refrigerant and rotates; the water feed and scraping devices are stationary. Capacities for this type of icemaking equipment are between 5 and 30 t/d.

Block Ice. Ice blocks are generally 1 m long and weigh 12.5 or 25 kg each. Water is charged into a large number of conical ice cells situated in an open brine tank. Because the brine used as secondary refrigerant is corrosive, the ice cells are made of lead-coated sheet steel or stainless steel. An agitator moves the brine to a vertical-tube evaporator in the brine tank, where it is chilled and then circulated around the ice cells. At a brine temperature of ca. − 7 °C, the freezing time is ca. 8 h for a 12.5-kg block or 19 h for a 25-kg block. After freezing, the ice cells are hoisted out of the brine tank and briefly dipped in warm water to release the ice blocks, which are then easily displaced by tipping.

To shorten the freezing time, double-walled ice cells are used in the "Rapid-Eis" freezing process. The refrigerant is evaporated in the jacket space and freezing is accelerated by means of an evaporator tube located inside the ice cell. To melt the ice and free it from the cell, the jacket and evaporator tube are connected to the condenser side of the refrigerating system. Drawbacks of this process are the high manufacturing costs and high mass of the ice cells, which must be designed for the refrigerant pres-

sure. The process can also operate with brine; cell costs are then lower. Refrigerating systems for this process are single-stage ammonia machines with reciprocating compressors.

Dry Ice. The method most commonly employed for producing dry ice (solid carbon dioxide) consists of expanding pressurized liquid carbon dioxide to ambient pressure; the temperature falls below the triple point and snowlike carbon dioxide is obtained (→ Carbon Dioxide, Chap. 9.). Roughly a third of the liquid is transformed to the solid state; the remaining gas is recycled. The carbon dioxide "snow" is molded into 12 – 100-kg blocks in hydraulic presses.

Dry ice sublimes at − 78.9 °C and has a heat of sublimation of 573 kJ/kg. It is used chiefly in the transport of deep-frozen products and, to a minor extent, in laboratory chilling apparatus.

2.7. Use of Cold in Construction [13], [21], [31], [159–162]

The heat of setting in concrete gravity dams ranges from 250 to 500 kJ/kg, depending on the aggregate used and the cement : sand ratio. This heat must be removed, especially in thick dams, so that it does not buildup and cause cracking. In *postcooling,* chilled water is pumped through pipes inside the structure; compact water-chilling plants are used for this purpose. About 75 000 kJ of cold is required per cubic meter of concrete. In *precooling*, excessive temperatures are prevented either by cooling the aggregate and water before mixing or by adding finely divided (flaked or tube-frozen) ice during mixing. Aggregate cooling is usually performed with chilled water, but other methods employing air chilled to − 20 °C also offer some operational advantages.

When tunneling or similar construction work requires shaft sinking in water-bearing ground, the ground around the planned excavation is stabilized by freezing. Double-walled pipes are driven into the soil. A secondary refrigerant (usually brine) at − 20 to − 40 °C flows down the inner pipe and returns through the outer space. The time required for freezing with these parallel pipes depends on the type and size of the excavation, the spacing of the pipes, and the temperature of the secondary refrigerant; months are often necessary. Transportable brine-cooling units with reciprocating or screw compressors, using ammonia as primary refrigerant, are employed.

2.8. Air Refrigeration in Mining [13], [163–165]

The temperature and humidity in mines often become so high that ordinary ventilation with outside air is inadequate. Refrigeration equipment can be set up below ground in the shaft, or central cooling plants can be erected at the surface to produce cold water, which is then led to air coolers near the working face.

Refrigerating systems for underground installation are compact, transportable, explosion-proof units that cool air to ca. 18 – 20 °C and dehumidify it. The air is introduced into the ventilation tubing ca. 50 – 100 m before the road ending. The plants usually have watercooled condensers and employ R 22 or R 12 as refrigerant. Such machines can be driven by a compressed-air motor instead of an electric motor.

2.9. Artificially Cooled Skating Rinks [13], [21], [31], [166]

An artificially cooled skating rink consists of tubes embedded in a concrete slab that rests on a frost-proof base. The pipes are parallel, spaced some distance apart, and ca. 25 – 30 mm below the surface. A secondary refrigerant (usually brine), evaporating ammonia, or R 22 is pumped through the pipes. The surface temperature of the ice over the entire rink surface must stay within narrow limits; stringent requirements thus apply to both the supply of secondary or primary refrigerant and the control system.

3. Heat Pumps [167–172]

Heat pumps operate on the refrigeration principles described in Sections 1.1 and 1.2. The process is shifted toward higher temperatures, and the temperature of the rejected heat allows a medium to be heated for a given application.

The economic significance of the heat pump lies in the saving of primary energy in thermotechnical processes. The useful energy consists partly of primary energy and partly of energy from the surroundings or another heat source.

In a heat pump driven by an *electric motor,* the heat rejected in the condenser is used for heating. The process is assessed from the energy standpoint in terms of the coefficient of performance ε, which is the ratio of thermal power to power consumption:

$$\varepsilon = \frac{\dot{Q}_c}{P_{el}}$$

It is determined essentially by the temperatures in the heat-source loop and the useful-heat loop.

An electrically driven heat pump is evaluated economically over a given span of time in terms of the average coefficient of performance β, which is the ratio of the heat generated to the electrical work consumed during this period:

$$\beta = \frac{Q_c}{W_{el}}$$

In many applications, the compressor is driven by an *internal-combustion (gas or diesel) engine.* Heat from the engine coolant and from the engine exhaust is used in addition to that produced in the condenser.

The heating medium absorbs heat in the condenser, and is then delivered by a pump to the motor coolant heat exchanger and, subsequently, to the exhaust heat exchanger. In this way, useful heat is obtained at higher temperatures than in heat pumps with electrical drive.

With regard to primary energy utilization, this heat-pump system is very economical as a heating method because a substantial part of the primary energy applied is available for heating purposes, and heat from the surroundings or some other heat source is also used.

Heat pumps driven with primary energy (including absorption heat pumps) are characterized by the performance ratio ζ. This quantity is the ratio of thermal power to power supplied by the fuel:

$$\zeta = \frac{\dot{Q}}{P_F}$$

The average performance ratio γ is, accordingly, defined as

$$\gamma = \frac{Q}{W_F}$$

where Q is the useful heat generated in the interval under consideration and W_F is the fuel energy consumed in the same interval.

The components used in heat-pump systems are largely the same as those in compression and absorption refrigerating systems. In heat pumps with internal-combustion drives, the engines and their control and monitoring equipment are supplemented by heat exchangers for recovering heat from the coolant and exhaust. Most such exchangers are of the shell-and-tube type.

With R 114, the thermal stability of the halocarbon refrigerant limits the temperature of the useful heat to between 100 and 130 °C. In general, however, these temperatures are between 50 and 80 °C when R 12 is used as refrigerant.

The most important application of heat pumps in industry is the generation of process heat, especially for drying, in the food industry, and for process technology. Heat generation with heat pumps is especially economical when the coupled production of cold and heat is possible, as in galvanizing, plastics fabrication, or food technology.

4. References

General References

1. R. Plank, (ed.): *Handbuch der Kältetechnik,* **vol. 1, 1954** Entwicklung, wirtschaftliche Bedeutung, Werkstoffe, Springer Verlag, Berlin.
2. *Thermodynamische Gundlagen,* **vol. 2,** 1953 in [1]
3. *Verfahren der Kälteerzeugung und Grundlagen der Wärmeübertragung,* **vol. 3,** 1959 in [1]
4. *Die Kältemittel,* **vol. 4,** 1956 in [1]
5. *Kältemaschinen, Kaltgasmaschinen, Kaltdampfmaschinen,* **vol. 5,** 1966 in [1]
6. *Automatik, Zubehör, Inbetriebnahme, Geräuschbekämpfung, Wärmepumpen,* **vol. 6 a,** 1969 in [1]
7. *Wärmeaustauscher,* **vol. 6 b,** 1988 in [1]
8. *Sorptionskältemaschinen,* **vol. 7,** 1959 in [1]
9. *Erzeugung sehr tiefer Temperaturen, Gasverflüssigung und Zerlegung von Gasgemischen,* **vol. 8,** 1957 in [1]
10. *Biochemische Grundlagen, Lebensmittelfrischhaltung,* **vol. 9,** 1952 in [1]

11. *Die Anwendung der Kälte in der Lebensmittelindustrie,* **vol. 10,** 1960 in [1]
12. *Der gekühlte Raum, der Transport gekühlter Lebensmittel und die Eiserzeugung,* **vol. 11,** 1962 in [1]
13. *Die Anwendung der Kälte in der Verfahrens- und Klimatechnik, Biologie und Medizin, Sicherheitsvorschriften,* **vol. 12,** 1967 in [1]
14. *Kältemaschinenregeln,* 7th ed., C. F. Müller, Karlsruhe 1981.
15. W. Pohlmann, W. Maake, H. J. Eckert: *Taschenbuch für Kältetechniker,* 17th ed., C. F. Müller, Karlsruhe 1988.
16. M. Bäckström, E. Emblik: *Kältetechnik,* 3rd ed., G. Braun, Karlsruhe 1965.
17. H. L. v. Cube (ed.): *Lehrbuch der Kältetechnik,* **vol. 1,** Physik, Chemie, Kälteerzeugung, Montage, C. F. Müller, Karlsruhe 1975.
18. *Regelung, Elektrotechnik, Kälteschutz, Kälteanwendung, Lebensmitteltechnik, Schallschutz, Neue Einheiten,* vol. 2 in [17]
19. H. Jungnickel, R. Agsten, W. E. Kraus: *Grundlagen der Kältetechnik,* 2nd ed., Technik, Berlin 1985.
20. International Institute of Refrigeration: *Saving of Energy in Refrigeration,* Paris 1980.
21. E. Emblik: *Kälteanwendung,* G. Braun, Karlsruhe 1971.
22. H. G. Hirschberg: *Kältemittel,* C. F. Müller, Karlsruhe 1966.
23. M. Mörsel: *Taschenbuch Kälteanlagen,* 3rd ed., VEB Technik, 1970.
24. H. Löwer: *Kältetechnik in Klimaanlagen,* C. F. Müller, Karlsruhe 1980.
25. H. Löwer: *Klimatechnik,* Springer, Berlin 1968.
26. L. Piatti: *Kühlflüssigkeiten und Kälteträger,* Sauerländer, Aarau, Frankfurt/M. 1968.
27. H. Titze: *Elemente des Apparatebaus, Grundlagen, Bauelemente, Apparate,* 2nd ed., Springer Berlin 1967.
28. *ASHRAE (Amer. Soc. of Heating, Refrig. and Air Cond. Eng.)* Handb., Fundamentals, New York 1981.
29. *ASHRAE Handb.,* Systems, New York, 1980.
30. *ASHRAE Handb.,* Equipment, New York, 1979.
31. *ASHRAE Handb.,* Applications, New York, 1982.
32. *Bull de l'Inst. Int. Froid,* Paris.
33. *VDI-Wärmeatlas,* 4th ed., Düsseldorf 1984.
34. K. Nesselmann: "Verfahren zur Kälteerzeugung unter −100 °C," *Chem. Ing. Tech.* **29** (1957) 198 – 200.
35. R. Plank: "Über den Idealprozeß von Kältemaschinen bei Verbund-Kompression," *Z. Gesamte Kälte Ind.* **35** (1928) 17 – 24.
36. K. Linge: "Der Einfluß des Ansaugzustands auf die volumetrische und die spezifische Kälteleistung," *Kältetechnik* **8** (1956) 75 – 79.
37. K. Linge: "Eine vereinfachte Berechnung zweistufiger Kältemaschinen," *Z. Gesamte Kälte Ind.* **44** (1937) 181 – 185.
38. A. Rasi: "Der günstigste Zwischendruck für zweistufige Kälteprozesse," *Ber. d. IX. Internat. Kältekongr.* **1** (Paris 1955) 3032 – 3039.
39. H. Lotz: "Fragen der Bemessung zweistufiger Kälteanlagen mit Kolbenkompressoren," *Kältetech. Klim.* **20** (1968) 38 – 46.
40. K. Baumann, E. Blass: "Beitrag zur Ermittlung des optimalen Mitteldruckes bei zweistufigen Kaltdampf-Verdichter-Kältemaschinen," *Kältetechnik* **13** (1961) 210 – 216.
41. H. Thumm: "Systemanalyse von Tieftemperatur-Kälteanlagen," *Kältetech. Klim.* **19** (1967) 36 – 43.
42. D. J. Missinur, W. L. Holladay: "Cascade Refrigeration Systems," *ASHRAE J.* **9** (1967) 70 – 73.
43. G. Ahnefeld, M. Giebe, Chr. Kluge: "Erzeugung tiefer Temperaturen mit Kaskadenkälteanlagen," *Luft Kältetech.* **12** (1976) 198 – 205.
44. W. Gosney: "A Survey of the newer Refrigerants," *Aust. Refrig. Air Cond. Heat* **5** (1962) 113 – 117.
45. J. H. Anderson: Refrigerants for Centrifugal Water Cooling Systems, *Aust. Refrig. Air Cond. Heat* **10** (1967) 100 – 104.
46. H. O. Spauchus: "Development in refrigeration: technical advances and opportunities for the 1990 s," *Int. J. Refrig.* **10** (1987) 263 – 270.
47. W. Frie: "Übersicht über die Berechnung der Zustandsgrößen von Kältemitteln," *Brennst. Wärme Kraft* **33** (1981) 320 – 324.
48. V. Chlumský: "Reciprocating and Rotary Compressors," *Public of Tech. Lit.,* SNTL, Prague 1965.
49. K. Linge: "Der Liefergrad und der indizierte Wirkungsgrad von Kältemaschinen," *Kälte* (Hamburg) 3 (1948) 116 – 123.
50. H. J. Müller: "Moderne Kälteverdichter der Hubkolbenbauart," *Linde Ber. Tech. Wiss.* **21** (1966) 25 – 37.
51. H. Najork: "Entwicklungstendenzen auf dem Gebiet der Kältemittelkompressoren," *Luft Kältetech.* **1** (1965) 2 – 9.

52. A. Bothe, J. Paul: "Kolben- und Schraubenverdichter in der Kälte-, Wärme- und Klimatechnik," *Ki Klima Kälte Heiz.* **11** (1983) 275 – 281.
53. U. Ritter: "Der Labyrinth-Kolbenkompressor für ölfreie Förderung von Gasen und Dämpfen," *Tech. Rundsch. Sulzer* **40** (1958) no. 2, 3.
54. H. J. Müller: "Weiterentwicklung der Trockenlaufkompressoren," *Linde Ber. Tech. Wissen.* **16** (1963) 11 – 18, **29** (1971) 12 – 17.
55. P. D. Laing, E. J. Perry: "The Development of oil-injected Screw Compressors for Refrigeration," *Prod. Instit. Refrig.* **60** (1963/64) 95.
56. A. Lundberg: "Die Eigenschaften des leistungsgeregelten Schraubenkompressors mit Öleinspritzung," *Kältetech. Klim.* **20** (1968) 102 – 107.
57. H. Soumerai: "Large Screw Compressors for Refrigeration," *ASHRAE J.* **9** (1967) 38 – 46.
58. B. Stenzel: "Schraubenverdichter für Kälte- und Wärmepumpenanlagen," *Ki Klima Kälte Heiz.* **11** (1983) 417 – 421.
59. I. W. Pillis: "Advancement in Refrigeration Screw Compressors," *ASHRAE Trans.* **92** (1986) no. 1.
60. W. Molls, P. Heimbach: "Ein Kältemittel-Drehkolbenverdichter mit saugseitiger Öleinspritzung," *Linde Ber. Tech. Wissen.* **9** (1960) 7 – 10.
61. K. H. Richter: "Moderne Ammoniak-Turbokälteanlagen," *Kälte (Hamburg)* **18** (1965) 2 – 20.
62. U. Senger: "Das äußere und innere Betriebsverhalten von Kälteanlagen mit Turbokompressoren," *Kältetechnik* **14** (1962) 238 – 245.
63. K. Jassniker: "Entwicklungstendenzen im Bau von Turbokompressoren für Kälteanlagen," *Kältetechnik* **14** (1962) 246 – 253.
64. M. Wemhöner, R. Göldner: "Turboverdichter in Kälteanlagen für Industrie-Kühlung und Klimatechnik," *Kältetechnik* **14** (1962) 326 – 332, 354 – 361.
65. D. Bachmann: "Letzte Entwicklungen in der Konstruktion von Turbokompressoren," *Kältetech. Klim.* **20** (1968) 150 – 155.
66. H. Dölz: "Anwendungsbereiche und Entwicklungstendenzen von Wärmeübertragern der Kältetechnik," *Luft Kältetech.* **4** (1968) 274 – 279.
67. F. Schulenberg: "Betriebsverfahren mit luftgekühlten Kondensatoren und Kühlern," *Chem. Ing. Tech.* **27** (1955) 262 – 268.
68. F. Schulenberg: "Wärmeübergang und Druckverlust bei der Kondensation von Kältemitteldämpfen in luftgekühlten Kondensatoren," *Kältetech. Klim.* **22** (1970) 75 – 81.
69. J. C. Chato: "Laminar condensation inside horizontal inclined tubes," *ASHRAE J.* **52** (1962) 52 – 60.
70. J. F. Pearson, J. G. Withers: "New Finned Tube Configuration Improves Refrigerant Condensing," *ASHRAE J.* **11** (1969) 77 – 82.
71. H. Göransson: "Wassergekühlte Kondensatoren," *Kältetechnik* **13** (1961) 398 – 403.
72. W. Niebergall: "Der Einfluß der Verschmutzung auf die Wärmeübertragung bei Apparaten der Kältetechnik und Verfahrenstechnik," *Allg. Wärmetech.* **8** (1957) 73 – 78.
73. K. H. Richter: "Entlüftungsprobleme bei Kälteanlagen," *Kälteprakt.* **2** (1962) 202 – 207.
74. P. Berliner: *Kühltürme – Grundlagen der Berechnung und Konstruktion,* Springer Verlag, Berlin, Heidelberg, New York 1975.
75. "Naßkühltürme – Wärmetechnische Grundlagen und Messungen," *VDI-Ber.* **298** (1977).
76. K. Nesselmann: "Die Berechnung von wassersparenden Berieselungsverflüssigern mit künstlicher Belüftung," *Z. Gesamte Kälte Ind.* **50** (1943) 97 – 102.
77. Th. E. Schmidt: "Bau und Betrieb von Verdunstungsverflüssigern," *Kältetechnik* **2** (1950) 221 – 225.
78. H. G. Hirschberg: "Zur Berechnung von Röhrenkesselverdampfern," *Kältetech. Klim.* **18** (1966) 155 –160.
79. K. Stephan: "Berechnung des Wärmeübergangs an siedende Kältemittel," *Kältetechnik* **15** (1963) 231 – 234.
80. E. Hofmann: "Wärmeübergangszahlen verdampfender Kältemittel," *Kältetechnik* **9** (1957) 7 – 12.
81. B. Slipcevic: "Wärmeübergang bei der Verdampfung von Frigenen," *Verfahrenstechnik* (Mainz) 5 (1971) 29 – 35.
82. A. Bürkholz, E. Muschelknauz: "Tropfenabscheider – Übersicht zum Stande des Wissens," *Chem. Ing. Tech.* 44 (1972) 503 – 509.
83. E. Hofmann: "Beitrag zur Berechnung von Flüssigkeitskühlern mit verdampfendem Kältemittel in Rohren," *Kältetech. Klim.* **23** (1971) 90 – 96.

84. E. U. Schlünder, J. M. Chawla, E. A. Thomé: *Wärmeübergangszahlen auf der Kältemittelseite in Verdampferrohren,* Inst. therm. Verfahrenstech. Univ. Karlsruhe 1968.
85. Th. E. Schmidt: "Die Wärmeleistung von berippten Oberflächen," *Abh. Dtsch. Kältetech. Ver.* **4** (1950), C. F. Müller, Karlsruhe.
86. E. Hofmann: "Wärmeübergang bei Einspritzverdampfern," *Linde Ber. Tech. Wissen.* **37** (1975) 24 – 55.
87. G. Mann: "Auslegung von Luftkühlern und Auswirkung der Konstruktionsdaten auf ihre Leistung," *Kälte* (Hamburg) 21 (1968) 161 – 175.
88. K. H. Rudolph: "Einige Gesichtspunkte bei der Gestaltung von Hochleistungsverdampfern," *Kälte Klim. Prakt.* **5** (1968) 155 – 164.
89. R. Polke: "Berippte Luftkühler mit direkter Verdampfung und Eisansatz," *Kälte (Hamburg)* **18** (1965) 661 – 669.
90. G. Lorentzen, R. Gronnerud: "Investigation of Liquid Hold-up, flow Resistance and Heat Transfer in an R12 Evaporator Coil with Recirculation," *Annexe 1970-1 de Bull. de l'Inst. Internat.* du Froid, Comiss. II, III. (London 1970), 179 – 203.
91. J. Lettner, R. Siegismund: "Thermostatische Expansionsventile," *Ki Klima Kälte Ing.* **5** (1977) 53 – 60.
92. K. Hartmann: "Kältemaschinen mit elektronischen Expansionsventilen," *Ki Klima Kälte Heiz.* **13** (1985) 443 – 447.
93. O. Möller-Olsen: "Automatischer Betrieb von Großkälteanlagen," *Kältetechnik* **7** (1955) 5 – 8.
94. P. Heimbach: "Automatische Saugdruck-Regelventile für Kälteanlagen," *Linde Ber. Techn. Wiss.* **17** (1964) 29 – 38.
95. L. Cojocaru: "Die Spaltrohrmotorpumpe und ihre Anwendung in der Kälteindustrie," *Kältetechnik* **14** (1962) 110 – 112.
96. G. Witt: "Kältemittelpumpen, die gestellten Aufgaben und Anforderungen sowie erprobte Bauformen", *Ki Klima Kälte Heiz.* **10** (1982) 431 – 434.
97. F. Merkel, F. Bosnjakovic: *Diagramme und Tabellen zur Berechnung der Absorptionskältemaschinen* Springer Verlag Berlin 1929.
98. W. Niebergall: "Ein- und mehrstufige Absorptions-Kälteanlagen zur Auswertung industrieller Abwärme," *Kältetechnik* **16** (1964) 199 – 210.
99. W. Malewski: "Auswahlkriterien für Absorptions-Kälteanlagen und deren Ausführungsformen," *Chem. Ztg.* **95** (1971) 186 – 192.
100. K. Brunnabend: "Kombination von Kompressions- und Absorptions-Kälteanlage zur wirtschaftlichen Ausnutzung von hochgespanntem Dampf," *Kältetechnik* **16** (1964) 211 – 212.
101. K. H. Richter: "Mit Heizöl oder Erdgas direkt befeuerte Absorptionskälteanlagen," *Kältetech. Klim.* **23** (1971) 142 – 150.
102. G. Holldorf: "Neuentwicklungen bei industriellen Absorptionskältemaschinen," *Kälte und Klimatech.* **12** (1977) 497 – 511.
103. G. Holldorf: "Methoden zur Senkung der Heizmittelqualität bei Absorptionskälteanlagen," *Ki Klima Kälte Heiz.* **12** (1984) 345 – 349.
104. M. J. P. Bogast: *Ammonia Absorption Refrigeration in Industrial Processes,* Gulf Publishing Company, 1981.
105. F. Ziegler, G. Alefeld: "Coefficient of performance of multistage absorption cycles," *Int. J. Refrig.* **10** (1987) 285 – 295.
106. G. Schuster: "Lithium-Absorptionskältemaschinen," *Kälte (Hamburg)* **19** (1966) 126 – 132.
107. T. Messing, W. Wöhlke: "Dampfstrahl-Kälteanlagen in der chemischen Industrie," *Chem. Ing. Tech.* **40** (1968) 219 – 223.
108. B. Bauer: "On the calculation of steamjet compressors with regard to the real state properties of the flow medium", *Klima + Kälte Ing.* **5** (1977) 389 – 393.
109. K. Sagel: "Grundlagen und Probleme der thermoelektrischen Kälteerzeugung," *VDI Z.* **104** (1962) 325 – 334.
110. H. Müller: "Bemessung und Aufbau von Peltieraggregaten," *Kältetechnik* **15** (1963) 137 – 143.
111. W. Hänlein: "Technologie und Anwendung des Peltiereffektes," *Kältetechnik* **16** (1964) 388 – 395.
112. A. W. Penn: "Cascaded Peltier Devices for low Temperature Operation", *Adv. Energy Convers.* **8** (1968) no. 2, 65 – 69.
113. F. Özvegyi: "Vergleich von Kälteübertragungsmitteln," *Kälte (Hamburg)* **23** (1970) 298 – 300.
114. R. Jünger: "Kühlsolen auf Chloridbasis," *Fette Seifen Anstrichm.* **61** (1959) 290 – 296.
115. H. Veith: "Flüssigkeitskühler in der Prozeß-kühlung," *Kältepraktiker* **4** (1964) 206 – 208.

116. H. Jacoby: "Kälteanlagen mit Turboverdichtern für industrielle Kühlung," *Linde Ber. Tech. Wiss. H.* **24** (1966) 36 – 42.
117. J. Rüedi: "Turbokompressoren in der Klimatechnik," *Kältetechnik* **14** (1962) 250 – 253.
118. G. Dinglinger: "Die Wärmeübertragung im Kratzkühler," *Kältetechnik* **16** (1964) 170 – 174.
119. A. J. Amstrong: "Kratzkühler als Kristaller," *Process Eng.* (London) 51 (1970) 59 – 65.
120. W. Niebergall: "Anwendung künstlicher Kälte in der Mineralöltechnik," *Kältetechnik* **15** (1963) 190 – 199.
121. H. Jacoby, K. H. Brendes: "Kaskaden-Turbo-Kälteanlage in Kompaktbauweise zur Tiefkühlung in einer Kristallisationsanlage," *Chem. Ing. Tech.* **40** (1968) 224 – 228.
122. E. Hofmann: "L'Exploitation industrielle du Paraxylène à l'aide d'un procédé d'élaboration à basses températures," *Proceedings Xth Internat. Congress of Refrigeration,* vol. 2, Kopenhagen 1959 pp. 425 – 430.
123. R. C. Burns: "Freezing Process for Conversion of Saline Water to Fresh Water," *ASHRAE J.* **5** (1963) 43 – 47.
124. W. Beutler: "Kältemaschinen in der chemischen Industrie," *Kälte (Hamburg)* **11** (1958) 177 – 180.
125. K. H. Richter: "Gaskühlverfahren," *Kälte* (Hamburg) 6 (1953) 74 – 80.
126. K. Heinze: "Befeuchtung und Kühlung von Preßluft," *Kältetechnik* **7** (1955) 150 – 155.
127. D. Dibbern: "Reif- und Schneebildung beim Abkühlen von Gas-Dampf-Gemischen in Gegenstrom-Wärmeaustauschern," *Kältetechnik* **16** (1964) 167 – 169.
128. A. Hauptmann: "Plan, Konstruktion and Arbeitsweise einer Lageranlage für verflüssigtes Erdgas," *Kälte* (Hamburg) 15 (1962) 422 – 425.
129. E. Emblick: "Über die Verflüssigung von Chlor durch Kälte," *Kältetechnik* **9** (1957) 72 – 75.
130. H. Abel, F. Özvegyi: "Chlorverdichtung und Chlorverflüssigung," *Tech. Rundsch. Sulzer* **50** (1968) no. 3, 123 – 132.
131. K. Dietrich: "Kältefernversorgung mit Ammoniak," *Kältetech. Klim.* **22** (1970) 184 – 188.
132. H. Schmitt: "Zentrale Kälteversorgung eines Chemiewerks," *Kälte (Hamburg)* **29** (1976) 422 – 433.
133. *Recknagel – Sprenger, Taschenbuch für Heizung und Klimatechnik,* R. Oldenburg-Verlag, München 1963.
134. T. C. Hines, B. R. Small: "Various Means of Supplying Make-up Air to Industrial Buildings," *ASHRAE-J* **5** (1963) 73 – 77.
135. H. Esdorn, A. Jahn: "Energetische Bewertung von Prozessen Raumlufttechn. Anlagen," *Heizung, Lüftung, Haustechnik* **31** (1980) 47 – 54.
136. H. Laakso, W. Mänz: "Kälteversorgung für die Klimaanlagen von Hochhäusern," *Heiz. Lüft. Haustech.* **14** (1963) 266 – 269.
137. K. Fitzner: "Sprühbefeuchter," *Ki Klima Kälte Heiz.* **12** (1984) 467 – 471.
138. W. Flechtenmacher: "Die vollautomatische, zentrale Kälteanlage einer neuen Fleischwarenfabrik," *Kältetechnik* **12** (1960) 190 – 196.
139. "Festschrift über Kühlhäuser und Kühlbahnhöfe, zum XI. Internationalen Kältekongreß in München 1963," *Kältetechnik* **15** (1963) 391 – 392.
140. I. A. Jorgensen: "Refrigeration in Modern Abattories," *Aust. J. Refrig. Air Cond. Heat.* **8** (1965) 195 – 200.
141. H. Ortner, A. Sauter: "Kälteversorgung und Wärmerückgewinnung eines neuen Fleischwarenbetriebes," *Klima Kälte Heiz.* **12** (1984) 17 – 22.
142. E. Lang: "Die Kühlung von Obst und Gemüse," *Kälte (Hamburg)* **27** (1972) 199 – 208.
143. F. W. Schröder: "Kältetechnische Einrichtungen von Kühlschiffen," *Kältetechnik* **17** (1965) 106 – 115.
144. J. T. Sills: "Recent Advances in the Liquid Nitrogen Freezing of Foods," *ASHRAE J.* **13** (1971), 36 – 38.
145. G. Schumacher: "Wirtschaftlicher Vergleich von Gefrierverfahren," *Linde Ber. Tech. Wiss. H.* **27** (1970) 10.
146. J. Remy: "Les moyens modernes de surgélation," *Rev Gen. Froid* **75** (1985) 21 – 49.
147. E. Hofmann: "Gefriereinrichtungen für Lebensmittelindustrie und Fischerei," *Linde Ber. Tech. Wiss., H.* **18** (1964) 58.
148. G. Dinglinger: "Tiefgefrieren von Lebensmitteln mit siedenden Flüssigkeiten," *Kältetech. Klim.* **22** (1970) 220 – 223.
149. H. G. Kessler: "Die Gefriertrocknung poriger Güter," *Kältetechnik* **14** (1962) 174 – 179.
150. U. Hackenberg: "Lebensmittelkonservierung durch Gefriertrocknung," *Umsch. Wiss. u. Tech.* **61** (1961) 428 – 431.

151. L. Rey: "Basic Principles of Lyophilization," *XIth Refr. Congress,* München 1963, Ref. P – G.
152. W. J. Davis: "Large Ice Storage Systems for Process water," *ASHRAE J.* **6** (1964) 66 – 69.
153. R. Plank: "Neue Entwicklungseinrichtungen des Rapid-Eis-Erzeugungssystems Wilbuchewich," *Kältetechnik* **15** (1963) 56 – 57.
154. E. Schmitt: "Erzeugung und Lagerung von Röhreneis," *Kältetechnik* **9** (1957) 108.
155. W. Maake: "Moderne Verfahren zur Blockeisherstellung," *Kältetechnik* **5** (1953) 37.
156. J. Kuprianoff: *Die feste Kohlensäure,* Enke Verlag Stuttgart 1953.
157. H. Dünkelmann: "Gewinnung und Vewendung von Kohlensäure und Trockeneis," *Linde Ber. Tech. Wiss.* H. 22 (1966) 36 – 45.
158. K. Nürnberger, H. Kubli: "Gewinnung von Kohlensäure und Trockeneis," *Escher Wyss.* **32** (1959) 16 – 27.
159. J. Pierre: "La Réfrigération du Barrage de Bort," *Rev. Gen. Froid* **30** (1953) 835 – 843.
160. W. Jurecka: "Betonkühlung bei der Staumauer Sariyar," *Linde Ber. Tech. Wiss.* H. 13 (1962) 32.
161. W. Ständer: "Die Verbesserung der Gefriertechnik auf wissenschaftlicher Grundlage," *VDI Z.* **108** (1966) 171 – 175.
162. R. Lauser: "Bodengefrieren," *BBC-Nachr.* **49** (1967) no. 12, 635 – 638.
163. G. Malter: "Der maschinentechnische Teil der Wetterkühlung," *Kältetechnik* **9** (1957) 21 – 24.
164. P. Weuthen: "Verfahren der Grubenwetterkühlung," *Glückauf* **98** (1962) 731 – 740.
165. W. Vogelwieche, E. Ganter: "Grubenbewetterung-Kühlsysteme in Bergwerken," *Energie* **40** (1988) 18 – 24.
166. W. Deyle: "Der Bau von Kunsteisbahnen," *Kältetechnik* **14** (1962) 10 – 17.
167. H. L. v. Cube, F. Steimle: *Wärmepumpen, Grundlagen und Praxis,* VDI-Verlag, Düsseldorf 1978.
168. G. Heinrich, H. Najork, W. Nestler: *Wärmepumpenanwendung in Industrie, Landwirtschaft, Gesellschafts- und Wohnungsbau,* VEB-Verlag Technik, Berlin 1982.
169. W. Eder, F. Moser: *Die Wärmepumpe in der Verfahrenstechnik,* Springer Verlag, Wien 1979.
170. A. Frank, H. Hohendorf, H. Moderegger, G. Trenkowitz: *Anwendung der Wärmepumpe in Industrie, Gewerbe und Landwirtschaft,* Verlag C. F. Müller, Karlsruhe 1980.
171. W. Malewski: "Industrielle Wärmepumpen für Nutztemperaturen bis 100 °C und Leistungseinheiten im MW-Bereiche," *VDI-Ber.* **289** (1977) 183 – 189.
172. K. Mötz: "Das Zusammenspiel von Verdichter und Gasmotor in Wärmepumpen bei Verwendung von Schrauben- und Hubkolbenverdichter," *Ki Klima Kälte Heiz.* **12** (1984) 167 – 170.

Cryogenic Technology

Randall F. Barron, Louisiana Tech University, Ruston, Louisiana 71 272, United States

1. Introduction 2
2. Uses 2
3. Liquefaction Systems 4
3.1. Joule – Thomson Effect 5
3.2. Basic Linde – Hampson System . . . 6
3.3. Precooled Linde – Hampson System 8
3.4. Linde Dual-Pressure System 9
3.5. Claude Liquefier 10
3.6. Liquefaction Systems for Natural Gas 11
3.7. Collins Helium-Liquefaction System 11
3.8. Miniature Liquefaction Systems . . 12
3.9. Comparison of Liquefaction Systems 14
4. Cryogenic Refrigeration Systems . . 15
4.1. Thermodynamically Ideal Refrigerator 15
4.2. Joule – Thomson Refrigerators . . . 15
4.3. CVI Cold-Gas Refrigerator 16
4.4. Stirling Refrigerator 17
4.5. Vuilleumier Refrigerator 18
4.6. Solvay Refrigerator 19
4.7. Gifford – McMahon Refrigerator . . 20
4.8. Magnetic Refrigerators 22
4.9. Dilution Refrigerator 24
5. Components of Liquefiers and Cryocoolers 26
5.1. Cryogenic Heat Exchangers 26
5.1.1. Types of Cryogenic Heat Exchangers 26
5.1.2. Heat Exchangers for Dilution Refrigerators 29
5.1.3. Heat Exchanger Design Problems . . 29
5.1.4. Regenerators 30
5.2. Compressors 31
5.3. Expanders 32
6. Low-Temperature Separation Systems 34
6.1. Minimum Work of Separation . . . 34
6.2. Properties of Mixtures 35
6.3. Flash Calculation 36
6.4. Principles of Cryogenic Rectification 37
6.5. Linde Single-Column System 39
6.6. Linde Double-Column System . . . 39
6.7. Heylandt System 40
6.8. Argon Separation 41
6.9. Neon Separation 42
6.10. Physical Adsorption 43
7. References 44

Symbols

a	constant in van der Waals equation, $\mathrm{Pa \cdot m^6/kg^2}$
b	constant in van der Waals equation, $\mathrm{m^3/kg}$
C	number of components; parameter in the BET equation
COP	coefficient of performance
e	heat exchanger effectiveness
F	number of independent thermostatic properties
FOM	figure of merit
h	specific enthalpy, J/kg
K	distribution coefficient
P	number of phases
p_i	partial pressure
p_m	mixture pressure
Q	heat flow rate, J/s
q_m	mass flow rate, kg/s
q_v	molar flow rate, mol/s
r	refrigerant mass flow rate ratio
s	specific entropy, $\mathrm{J\,kg^{-1}K^{-1}}$
v	specific volume, $\mathrm{m^3/kg}$
w	work flow rate, J/s
x	expander mass flow rate ratio;mole fraction in the liquid phase
y	liquid yield; mole fraction in the vapor phase
α	relative volatility
γ	specific heat ratio
η	efficiency
θ	parameter in the BET equation, K
μ_JT	Joule – Thomson coefficient, K/Pa
π	vapor pressure, Pa

Subscripts

A	absorbed or added, e.g., energy absorbed
act	actual, e.g., actual work
ad	adiabatic
c	cold; compressor
e	expander
f	liquid phase
g	vapor phase
h	hot or high, e.g., high temperature

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_20

i	intermediate pressure stream; inversion
id	ideal
l	low, e.g., low temperature
m	molar; mixture; number of plates in the lower section of the column
min	minimum
n	number of plates in the upper section of the column
o	outside; ambient
r	refrigerant
rev	reversible
R	rejected
sat	saturation

1. Introduction

Cryogenic technology deals with the development and utilization of low-temperature processes and equipment. The field of cryogenics generally involves systems that operate at temperatures below −150 °C (123 K) [1]. This dividing line between the cryogenic temperature range and the field of ordinary refrigeration is selected because the normal boiling points of the so-called permanent gases, such as helium, hydrogen, nitrogen, and oxygen, lie below −150 °C, whereas the freon refrigerants, hydrogen sulfide, ammonia, and other conventional refrigerants all boil above −150 °C.

2. Uses

Applications of cryogenic technology vary widely, both in scope and size. Some of the areas involving cryogenic technology are discussed briefly in this chapter.

Rocket Propulsion Systems. Many of the large launch vehicles used in the United States space program use liquid oxygen as the oxidizing agent [2]. The first flight test of a liquid-fuel rocket using liquid oxygen was made by Robert H. Goddard in 1926. The NASA Saturn V Apollo vehicle used liquid oxygen as the oxidizing agent for all three stages and liquid hydrogen as the fuel for the upper two stages. The Space Shuttle propulsion system uses liquid hydrogen as the fuel and liquid oxygen as the oxidizing agent.

Cryogenic fluids are used because of the high specific impulse (thrust per unit mass flow rate) that can be achieved with the hydrogen – oxygen combination. The fluids must be in the liquid phase so that a sufficient mass may be stored on board within a reasonable volume.

The use of liquid hydrogen as a fuel in space vehicles introduces several special problems in addition to the problem of the low density of the fluid. Because the normal boiling point of liquid hydrogen is only 20.3 K, ambient air liquefies on the outside of an exposed, uninsulated liquid hydrogen storage tank. The condensed air on the tank poses safety problems and results in excessive heat transfer to the cold liquid hydrogen. To avoid this problem, liquid hydrogen tanks must be insulated.

Liquid hydrogen has the second lowest viscosity of all liquids (liquid helium has the lowest viscosity), which makes ordinary sealing difficult. In many cases, the only solution is to use a welded construction, which makes component disassembly and inspection both complicated and expensive.

An advantage of liquid hydrogen is that it can be used as the coolant for rocket nozzles. Liquid hydrogen is usually at a pressure of ca. 7 MPa before it is injected into the combustion chamber. Because the critical pressure of hydrogen is ca. 1.3 MPa, no problems of two-phase flow occur in the nozzle coolant passages with regenerative cooling.

Manufacturing Processes. Oxygen is required in the *steel-making process*, and the economical production of oxygen involves separation of the components of air in a cryogenic system [3]. In the open-hearth process, the conventional use of air as the gas above the molten charge in the furnace requires additional heating by an auxiliary fuel. The use of oxygen allows the operation of open-hearth steel furnaces at ≥130 % of the capacity when air is used.

High-strength pressure vessels may be constructed by cryogenic forming, in which the cylinder is stretch-formed at liquid nitrogen temperatures. This process (called Ardeforming) involves placing the fabricated vessel in a bath of liquid nitrogen (77.4 K) and introducing high-pressure gaseous nitrogen into the vessel. The vessel is allowed to expand by ca. 15 % into a stainless steel die of the appropriate shape. After

this stretching process, the vessel is allowed to warm slowly to room temperature. An improvement in tensile strength as high as a factor of 3 has been observed.

Tool steels used in manufacturing processes have been cryogenically treated to improve wear resistance of the tool [4]. The tool steel is cooled slowly (cooling rate ca. 3 °C/min) to liquid nitrogen temperatures and is then allowed to soak at the cryogenic temperature for ca. 20 h. During this process, essentially all of the retained austenite in the tool steel is transformed to martensite, which is a more wear resistant phase (→ Steel). The useful lifetime of parts, such as slitter knives, high-speed drills, etc., has been increased by a factor of 3 – 6 by this technique.

Recycling of Rubber. One of the more difficult materials to recycle is automobile or truck tire. Each year, $(160-200) \times 10^6$ tires are scrapped in the United States alone [5]. By a conservative estimate, these tires contain ca. 1.5×10^6 t of rubber that could be reclaimed and recycled. The conventional reclamation method involves chemical devulcanization of the rubber and mechanically grinding the resulting material. The widespread use of steel-belted tires makes this process uneconomical today.

When rubber is cooled to a temperature below 210 – 175 K, it passes through a glass transition and becomes brittle. In this brittle condition, the rubber can be crushed easily into small particles. The cord and metal components in the original tire can be separated readily by magnetic and air blast techniques.

The *freezing process* is usually carried out in a freezing tunnel, in which the rubber is carried along by a conveyor belt and liquid nitrogen is sprayed onto the surface of the scrap tires. The vapor that is formed when the liquid nitrogen evaporates is recirculated through the tunnel by a series of fans which increase the velocity of the nitrogen gas. The high velocity of the gas enhances the heat transfer process and improves the effectiveness of the freezing process. After passing through the freezing tunnel, the material is transferred immediately to a size-reduction mill, where the cord and metal components in the original tire are separated from the rubber particles in a single operation.

The crumb rubber particles are virtually free of fiber and steel contaminants. This property allows the rubber to be used in a variety of applications, including highway repairs using tire – asphalt mixes, soil conditioning, and mixing with coal for use as a fuel extender.

Currently, the cryogenic process is the only economically effective method to recover the rubber from steel-belted radial tires.

Food Processing and Transportation. Freezing as a technique for preserving food was used as early as 1840.

Currently, frozen foods are prepared by moving the food packages on a conveyor belt through a liquid nitrogen spray in a specially insulated tunnel. Initial contact with the liquid nitrogen freezes all exposed surfaces and seals in both the flavor and aroma of the food product. The food packages are then passed through an "equilibrium zone" in the tunnel that is cooled by the cold nitrogen vapor.

The cryogenic food freezing process has several advantages over the conventional food freezing process that uses cold air. Cryogenic freezing requires ca. 7 min compared with ca. 30 min required by conventional freezing methods. In conventional freezing, ca. 3 – 15 wt % (primarily water) of the product is removed by the air blast. In contrast, cryogenic freezing results in only 0.2 – 0.3 wt % dehydration [6]. This reduction in moisture loss results in improved flavor, texture, and appearance of the frozen food when compared with the resulting product of conventional freezing. The cryogenic freezing system is fairly simple and requires ca. 10 % of the maintenance expenditures of a conventional system.

Liquid nitrogen has also been used as the refrigerant in frozen-food transport trucks and railway cars [7]. The system consists of an insulated container for the liquid nitrogen and a series of nozzles that are activated by a thermostatically controlled switch to spray liquid nitrogen into the refrigerated compartment. The system is quite versatile in that the foods may be maintained at any prescribed temperature. Temperatures lower than those economically achieved by mechanical refrigeration may be maintained with the cryogenic system. Moisture retention of the cargo is good because the inert atmosphere inside the storage space prevents oxidation and drying.

Electronics. Sensitive microwave amplifiers (masers) are cooled to liquid nitrogen or liquid helium temperatures to assure that thermal vibrations of the atoms of the amplifier element do not seriously degrade the performance of the system. Cryogenically cooled masers are used in radio astronomy applications and in communication systems such as the Telstar satellite [3, pp. 157 – 170]. These systems use a compact, continuously operating, closed-cycle helium refrigerator to maintain a temperature as low as 4.2 K.

Tiny superconducting elements, called SQUIDs (superconducting quantum interference devices), are used in extremely sensitive digital magnetometers and voltmeters [8]. In addition to SQUIDs, other electronic devices that use superconductivity in their operation include superconducting amplifiers, transformers, and magnets. Superconducting magnets (solenoids) are used as levitation systems for trains that operate at speeds up to 500 km/h [9]. These trains are used for commuter service between Tokyo and Osaka in Japan. The levitation system consists of a normal track conductor coil on the ground and a superconducting coil on board the train. The superconducting coils are maintained at liquid helium temperature (4.2 K) by means of a cryostat filled with liquid helium. The propulsion unit for the train is a linear induction motor, with the primary coil located on the ground and the secondary coil placed on the train.

Biological and Medical Applications. Containers cooled by liquid nitrogen are used to preserve whole blood, tissue, bone marrow, and animal semen for extended periods of time [10]. Glycerol is mixed with such material as bull semen to protect the cells during freezing. The semen is slowly cooled to ca. 5 °C and is maintained at this temperature for at least 5 h to allow the glycerol to penetrate the cells adequately. The cells are then cooled at 1 °C/min to −15 °C and then at 5 °C/min to −79 °C. The frozen material can then be stored indefinitely at −196 °C in liquid nitrogen.

Whole blood cells have been preserved by placing the blood in a flat metal container and immersing the container in liquid nitrogen. The blood is thawed for use by abrupt immersion in an agitated bath of warm water. Whole blood has also been frozen by spraying it directly into liquid nitrogen. Cryogenic surgery (cryosurgery) has been used in eye surgery and for the treatment of Parkinson's disease and various lesions. This surgical technique has several advantages over conventional surgery, including reduced problems with bleeding, greater control of the area of influence, and less trauma to the system [11].

Several special-purpose probes have been developed for cryosurgery. Generally, liquid nitrogen (or sometimes liquid argon) flows through a vacuum-insulated cannula to the exposed tip on the probe and freezes the tissue to be destroyed. The size of the frozen region is controlled by adjusting the mass flow rate of cryogen through the cannula.

3. Liquefaction Systems

Several different systems are used to liquefy cryogenic fluids. To compare the thermodynamics of these systems, the thermodynamically ideal liquefaction system must first be established [12]. The perfect refrigeration cycle is the Carnot cycle. However, liquefaction is essentially an open-system arrangement. Therefore, the ideal liquefaction system only requires the first two processes in the Carnot cycle, which are a reversible isothermal compression followed by a reversible adiabatic (isentropic) expansion process to the saturated liquid. The ideal process is shown in Figure 1.

The gas to be liquefied is compressed reversibly and isothermally from ambient conditions (point 1) to some high pressure (point 2). This high pressure is selected such that the gas will reach the saturated liquid-state point (point f) on expansion through a reversible isentropic expander.

By applying the first law of thermodynamics to the system consisting of the compressor, expander, and liquid receiver, the following equation is obtained:

$$Q_{\mathrm{rev}} - W_{\mathrm{id}} = q_{\mathrm{m}} (h_{\mathrm{f}} - h_1)$$

where Q_{rev} = heat transfer from the ideal compressor, W_{id} = total ideal work requirement, q_{m} = mass flow rate through the system, h_{f} = specific enthalpy of the liquid, and h_1 = specific enthalpy of the gas at ambient conditions. The heat transfer process for the ideal

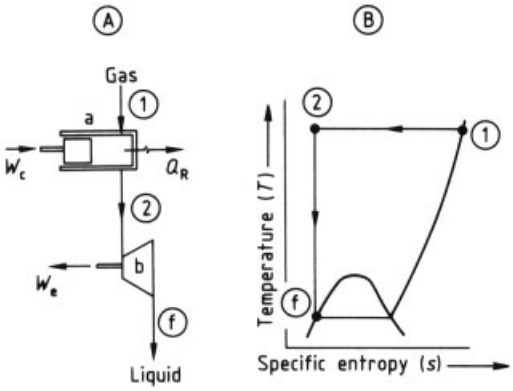

Figure 1. Thermodynamically ideal system for gas liquefaction ∗
A) Schematic: a) Isothermal compressor; b) Isentropic expander
B) Liquefaction thermodynamics
∗ Numbers on the temperature – entropy diagram correspond to similarly numbered points on the schematic diagram; f: fluid state

compressor is reversible and isothermal. Using the second law of thermodynamics gives

$$Q_{\mathrm{rev}} = q_{\mathrm{m}} T_1 (s_2 - s_1) = q_{\mathrm{m}} T_1 (s_{\mathrm{f}} - s_1)$$

where T_1 = ambient temperature, s_{f} = specific entropy of the liquid, s_1 = specific entropy of the gas at ambient conditions, and s_2 = specific entropy of the gas leaving the compressor ($s_2 = s_{\mathrm{f}}$). Substitution for the reversible heat transfer in the first expression gives the expression for the ideal work or the minimum work requirement for liquefaction of a gas:

$$-W_{\mathrm{id}}/q_{\mathrm{m}} = T_1 (s_1 - s_{\mathrm{f}}) - (h_1 - h_{\mathrm{f}}) \quad (1)$$

The negative sign appears in Equation (1) because of thermodynamic sign convention (i.e., work done by the system is positive; work done on the system is negative). The ideal work requirements for several common cryogenic fluids are given in Table 1.

3.1. Joule – Thomson Effect

Most liquefaction systems use an expansion valve or Joule – Thomson (J – T) valve to produce low temperature. Because heat transfer and kinetic energy changes are negligible and because there is no work transfer for the flow process through an expansion valve, the process is isenthalpic. The constant enthalpy lines on the temperature – pressure plane are shown in Figure 2. The effect of change in temperature for an isenthalpic change in pressure is represented by the *Joule – Thomson coefficient*, μ_{JT}, defined by

$$\mu_{\mathrm{JT}} = \left(\frac{\partial T}{\partial p}\right)_h$$

The Joule – Thomson (J – T) coefficient is the slope of the isenthalpic lines in Figure 2. In the region where $\mu_{\mathrm{JT}} < 0$, expansion through the valve (a decrease in pressure) results in an increase in temperature, whereas in the region where $\mu_{\mathrm{JT}} > 0$, expansion results in a decrease in temperature. The curve that separates the two regions is called the *inversion curve*. The J – T coefficient is zero along the inversion curve, because the slope of the isenthalpic line is zero.

Table 1. Ideal work requirements for liquefaction of gases beginning at 300 K and 101.3 kPa

Gas	Normal *bp*, K	Ideal work of liquefaction, kJ/kg
Helium [^{4}He]	4.21	6 819
Hydrogen	20.27	12 019
Neon	27.09	1 335
Nitrogen	77.36	768.1
Air	78.8	738.9
Carbon monoxide	81.6	768.6
Argon	87.3	478.6
Oxygen	90.2	635.6
Methane	111.7	1 091

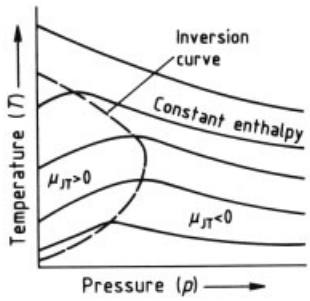

Figure 2. Temperature – pressure diagram for a real gas

Thermodynamics shows that the J – T coefficient may be expressed in terms of other thermodynamic properties as follows [13]:

$$\mu_{\mathrm{JT}} = \frac{1}{c_p}\left[T\left(\frac{\partial v}{\partial T}\right)_p - v\right] \quad (2)$$

where v is the specific volume of the material. For an ideal gas, $v = (RT)/p$ and

$$\left(\frac{\partial v}{\partial T}\right)_p = \frac{R}{p} = \frac{v}{T}$$

where R = gas constant and p = pressure. Substitution into Equation (2) gives $\mu_{\mathrm{JT}} = 0$ for an ideal gas. For a real gas, the J – T coefficient may be either negative or positive, depending on the temperature and pressure. One equation of state for gases, which illustrates the behavior of real gases, is the *van der Waals equation of state:*

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT$$

where the constant a gives a measure of the intermolecular forces between the gas molecules, and the constant b gives a measure of the finite size of the gas molecules. The expression for the J – T coefficient for a van der Waals gas is given by Equation (3):

$$\mu_{\mathrm{JT}} = \frac{(2a/RT)(1 - b/v)^2 - b}{c_p\left[1 - (2/vRT)(1 - b/v)^2\right]} \tag{3}$$

where c_p = specific heat at constant pressure. For large values of the specific volume, Equation (3) can be approximated by

$$\mu_{\mathrm{JT}} = \frac{1}{c_p}\left(\frac{2a}{RT} - b\right)$$

This expression shows that the J – T coefficient is positive when $T < 2a/bR$ and negative when $T > 2a/bR$. The inversion curve is represented by all points at which the J – T coefficient is zero. The inversion temperature from Equation (3) is given by

$$T_{\mathrm{i}} = \frac{2a}{bR}(1 - b/v)^2$$

The maximum inversion temperature for a van der Waals gas is the temperature on the inversion curve where $p = 0$ (or $b/v = 0$) or $T_{\mathrm{i,\,max}} = 2a/bR$. The maximum inversion temperature (in K) for several gases is as follows:

Helium [^{4}He]	45
Carbon monoxide	652
Hydrogen	205
Neon	250
Argon	794
Air	603
Oxygen	761
Nitrogen	621
Methane	939

The maximum inversion temperatures for helium, hydrogen, and neon are below ambient temperature. Therefore, liquefaction systems that use J – T expansion alone to produce low temperature cannot be used to liquefy these gases beginning at ambient temperature. Additional effects, such as precooling or expansion engines, must be used in liquefaction systems for these gases.

3.2. Basic Linde – Hampson System

Historically, the Linde – Hampson system was the second system used to liquefy gases at cryogenic temperatures (the cascade system was used first) [14]. The Linde – Hampson liquefaction system is shown in Figure 3 (see next page).

The gas is compressed (a) from ambient conditions to a pressure of ca. 20 MPa. The high-pressure gas then flows through a heat exchanger (b), in which the gas is cooled by heat exchange with the return stream of cold, low-pressure gas. At point 3, the cold high-pressure gas is expanded through a Joule – Thomson valve (c) to a pressure of the order of ambient pressure. At the exit of the valve (point 4), the stream is in the two-phase (liquid – vapor) region. The liquid phase is collected in the liquid receiver (d) and withdrawn. The saturated vapor is returned through the heat exchanger to furnish cooling for the incoming warm gas.

The liquid yield, or fraction of the warm gas stream that is liquefied, is given by

$$y = \frac{(q_{\mathrm{m}})_{\mathrm{f}}}{q_{\mathrm{m}}} = \frac{(h_1 - h_2) - (Q_{\mathrm{o}}/q_{\mathrm{m}})}{h_1 - h_{\mathrm{f}}}$$

where $(q_{\mathrm{m}})_{\mathrm{f}}$ = mass flow rate of liquid; q_{m} = total mass flow rate; Q_{o} = heat inleak from ambient; h_1 = enthalpy at the outlet at the warm end of the heat exchanger; h_2 = enthalpy at the inlet at the warm end of the heat exchanger; and h_{f} = enthalpy of the saturated liquid (fluid). (The numerical subscripts refer to the corresponding points in Figure 3.) The liquid yield may also be written in terms of the effectiveness of the heat exchanger, defined by

$$e = \frac{h_1 - h_{\mathrm{g}}}{h_1' - h_{\mathrm{g}}}$$

where h_1' = enthalpy of the low-pressure stream at pressure p_1 and at temperature $T_1{}' = T_2$;

h_g = enthalpy of the saturated vapor. If the heat exchanger were ideal, the cold gas would leave the warm end of the heat exchanger at the same temperature as that of the inlet temperature of the warm stream. The actual exit temperature of the cold stream is

$$T_1 = T_1' - (1-e)\left(T_1' - T_g\right)$$

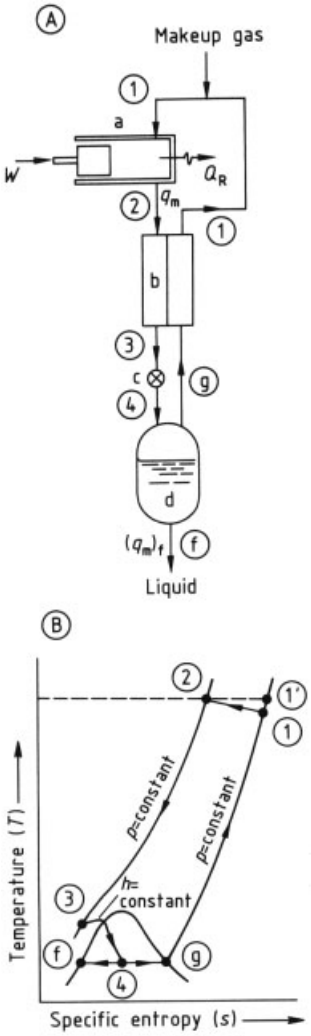

Figure 3. Linde – Hampson liquefaction system ∗
A) Schematic: a) Compressor; b) Heat exchanger; c) J – T valve; d) Liquid receive
B) Liquefaction thermodynamics
∗ Numbers on the temperature – entropy diagram correspond to similarly numbered points on the schematic; f: fluid state; g: vapor state

The term $(1 - e)$ is called the heat exchanger ineffectiveness [15]. The liquid yield for the Linde – Hampson system may be written in the alternative form:

$$y = \frac{\left(h_1' - h_2\right) - (1-e)\left(h_1' - h_g\right) - \left(Q_o/q_m\right)}{\left(h_1' - h_f\right) - (1-e)\left(h_1' - h_g\right)} \quad (4)$$

The heat exchanger effectiveness must be close to unity for the system to produce liquid. The minimum heat exchanger effectiveness may be found from Equation (4) by setting the liquid yield equal to zero. The result is

$$e_{min} = \frac{\left(h_1' - h_2\right) - \left(Q_o/q_m\right)}{\left(h_1' - h_g\right)} \quad (5)$$

A typical value of e_{min} for a nitrogen liquefier operating between ambient pressure and temperature (101.3 kPa and 300 K) and 20.3 MPa is ca. 0.86. The power requirement for the compressor is given by Equation (6):

$$-\frac{W}{q_m} = \frac{T_2\left(s_1' - s_2\right) - \left(h_1' - h_2\right)}{\eta_c} \quad (6)$$

where η_c = isothermal efficiency of the compressor. The negative sign is included in Equation (5) because work is done on the gas by the compressor. An important performance parameter for liquefaction systems is the figure of merit (FOM), defined by

$$\mathrm{FOM} = \frac{\left(-W_{id}/q_m\right)}{\left[-W_{act}/\left(q_m\right)_f\right]} = \frac{y\left(-W_{id}/q_m\right)}{\left(-W_{act}/q_m\right)}$$

where W_{act} = actual work requirement. Performance values for the Linde – Hampson system using various gases are given in Table 2 (see next page).

The basic Linde – Hampson system cannot be used to liquefy gases such as helium, hydrogen, and neon, which have a maximum inversion temperature lower than ambient temperature. For these gases at ambient temperature, the enthalpy difference $(h_1' - h_2)$ is negative; therefore, the calculated value of the liquid yield from Equation (4) is negative. The system can never get started from the warm condition, because the gas would experience an increase in temperature when it expands through the J – T valve at ambient temperature. Even if the system could be cooled somewhat initially, the gas would expand through the J – T valve into the vapor region

Table 2. Performance of the Linde – Hampson liquefaction system for different fluids [a]

Fluid	Normal *bp*, K	Liquid yield $y = (q_m)_f/q_m$		Work per unit mass liquefied, kJ/kg		FOM[b]	
		Ideal[c]	Actual[d]	Ideal[c]	Actual[d]	Ideal[c]	Actual[d]
N_2	77.4	0.0708	0.0581	6673	11 620	0.1151	0.0661
Air	78.8	0.0808	0.0672	5621	9 653	0.1313	0.0765
CO	81.6	0.0871	0.00749	5381	8 934	0.1428	0.0860
Ar	87.3	0.1183	0.1090	2750	4 265	0.1741	0.1123
O_2	90.2	0.1065	0.0957	3804	6 048	0.1671	0.1051
CH_4	111.7	0.1977	0.1888	3957	5 919	0.2758	0.1844

[a] State points are $p_1 = 101.3$ kPa, $p_2 = 20.265$ MPa, $T_2 = 300$ K. [b] Figure of merit. [c] Ideal system operates with $e = 1.00$ and $\eta_c = 1.00$, heat transfer from ambient is neglected. [d] Actual system operates, $e = 0.975$ and $\eta_c = 0.7$, heat transfer from ambient is neglected.

and completely miss the two-phase region [12, pp. 71 – 73].

The designer of a liquefaction system has little freedom in choosing the conditions at point 1, which is usually near ambient temperature and pressure. However, the designer can select the maximum pressure (at point 2) in the Linde – Hampson liquefaction system. To maximize the liquid yield for the system, the optimum pressure at point 2 must be the pressure that minimizes the enthalpy h_2. Mathematically, for minimum h_2, the following must apply:

$$\left(\frac{\partial h}{\partial p}\right)_{T=T_1} = 0$$

From the chain rule of calculus,

$$\left(\frac{\partial h}{\partial p}\right)_T = -\left(\frac{\partial h}{\partial T}\right)_p \left(\frac{\partial T}{\partial p}\right)_h = -c_p \mu_{JT}$$

Therefore, the pressure that minimizes h_2 is the pressure for which $\mu_{JT} = 0$; i.e., point 2 should lie on the inversion curve. For air at 300 K, the corresponding pressure is ca. 400 MPa. However, practical systems are generally designed to minimize the work per unit mass liquefied. The optimum pressure to achieve this condition can be found by numerical methods. For air, the optimum pressure at point 2 to minimize the power input per unit mass liquefied is ca. 20 MPa [14, p. 71].

3.3. Precooled Linde – Hampson System

A plot of the liquid yield y as a function of the temperature at the inlet of the heat exchanger (point 2) for the basic Linde – Hampson system is shown in Figure 4 for liquefaction of nitrogen. The performance of the liquefier can be improved if it is modified such that the gas enters the heat exchanger at a temperature lower than ambient temperature. This is the basic principle of the precooled system shown in Figure 5.

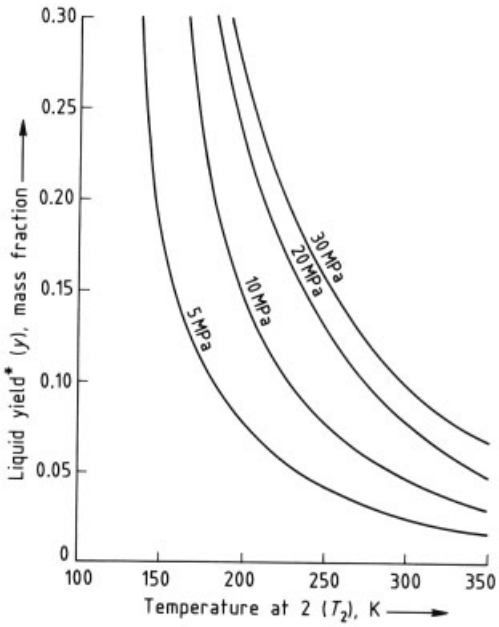

Figure 4. Yield of liquid nitrogen ∗ [$y = (q_m)_f/q_m$] for the Linde – Hampson liquefaction system as a function of the temperature of the gas at the warm end of the heat exchanger (see Fig. 3)
∗ Yield is mass fraction compared to ideal system.

A separate refrigeration system, using a fluid such as ammonia or a freon compound, is used to precool the main gas stream. The critical temperature of the auxiliary coolant must be greater than ambient temperature so that the precoolant can be condensed by using a condenser cooled with water or ambient air.

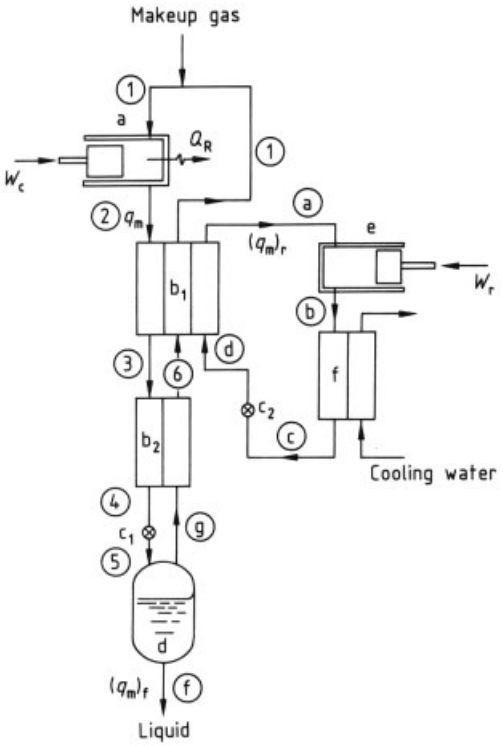

Figure 5. Precooled Linde – Hampson liquefaction system
a) Main compressor; b) Heat exchanger; c) J – T valve; d) Liquid receiver; e) Refrigerant compressor; f) Refrigerant condenser

The temperature at point 3 is limited by the minimum temperature of the precoolant at point d. This factor also places an upper limit on the liquid yield for this system.

The liquid yield for the precooled Linde – Hampson system is given by Equation (7):

$$y = \frac{(h_1 - h_2) + r\,(h_a - h_c) - (Q_o/q_m)}{(h_1 - h_f)} \tag{7}$$

where $r = (q_m)_r/q_m$ = refrigerant mass flow rate ratio; h_a = enthalpy of the auxiliary refrigerant at point a; and h_c = enthalpy at point c. The second term in Equation (7) represents the improvement in liquid yield over the basic Linde – Hampson system because of the use of precooling. The power requirement per unit mass flow rate through the main compressor is given by Equation (8):

$$-W/q_m = \frac{T_1\,(s_1 - s_2) - (h_1 - h_2)}{\eta_c} + r\,(h_b - h_a) \tag{8}$$

The last term in Equation (8) represents the additional work required by the auxiliary refrigeration system. Another limiting factor in the design of the precooled system is that the refrigerant mass flow rate $(q_m)_r$ must be small enough so that the liquid entering the heat exchanger at point d can be completely vaporized. If the auxiliary refrigerant is not completely vaporized at point a, the liquid enters the refrigerant compressor and damage to the compressor results.

3.4. Linde Dual-Pressure System

The basic Linde – Hampson system may be modified, as shown in Figure 6, by compressing only a portion of the main gas stream through the entire pressure range [16].

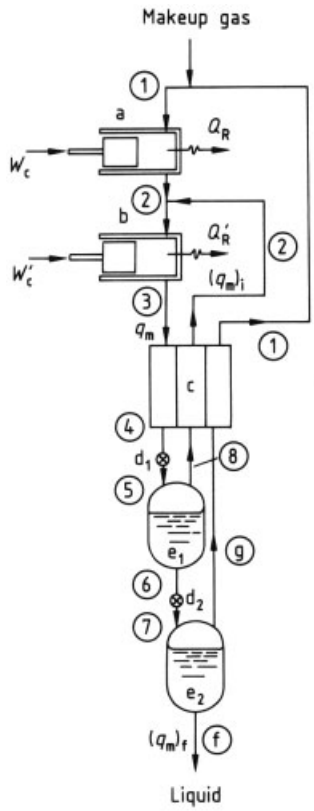

Figure 6. Linde dual-pressure liquefaction system
a) Low-pressure compressor; b) High-pressure compressor; c) Heat exchanger; d) J – T valve; e) Liquid receiver

In the Linde dual-pressure system, the gas to be liquefied is compressed from near ambient pressure to an intermediate pressure (a)

and then to a high pressure (b) after a return stream has been added. The high-pressure gas is passed through a three-channel heat exchanger (c) and expanded to an intermediate pressure, where some of the gas is liquefied. The vapor is returned through the three-channel heat exchanger to the high-pressure compressor. The liquid stream is finally expanded to the low pressure of the cycle.

The liquid yield for the dual-pressure system is given by

$$y = \frac{(h_1 - h_3) - \left[(q_\mathrm{m})_\mathrm{i}/q_\mathrm{m}\right](h_1 - h_2)}{(h_1 - h_\mathrm{f})}$$

where $(q_\mathrm{m})_\mathrm{i}$= mass flow rate of the intermediate pressure stream at point 8, and q_m= mass flow rate through the high-pressure compressor. The work requirement per unit mass compressed in the high-pressure compressor is given by

$$-W/q_\mathrm{m} = [T_1(s_1 - s_3) - (h_1 - h_3)]/\eta_\mathrm{c} - \left[(q_\mathrm{m})_\mathrm{i}/q_\mathrm{m}\right][T_1(s_1 - s_2) - (h_1 - h_2)]/\eta_\mathrm{c}$$

The work per unit mass liquefied may be determined from the following:

$$-W/(q_\mathrm{m})_\mathrm{f} = (-W/q_\mathrm{m})/y$$

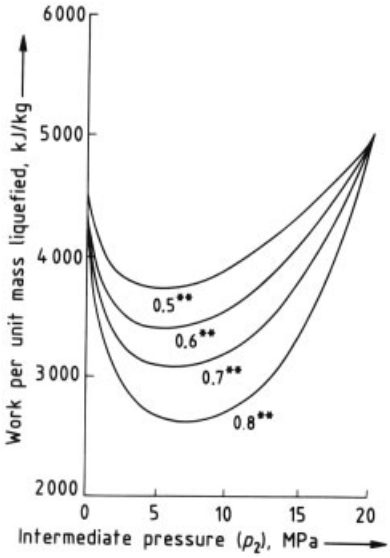

Figure 7. Work required to liquefy a unit mass of air in the Linde dual-pressure system ∗
∗ The intermediate pressure is the pressure of the gas between the low- and high-pressure compressors.
∗∗ Values for $(q_\mathrm{m})_\mathrm{i}/q_\mathrm{m}$

The work requirement for liquefaction of air in the Linde dual-pressure system is shown in Figure 7 (see below). At an optimum value of the intermediate pressure, the system work requirement is minimized for a given mass flow rate ratio, $(q_\mathrm{m})_\mathrm{i}/q_\mathrm{m}$. In addition, the work requirement for a given intermediate pressure is decreased by increasing the mass ratio. The mass ratio is thermodynamically limited, however, because $(q_\mathrm{m})_\mathrm{i} + (q_\mathrm{m})_\mathrm{f}$ must be less than the mass flow rate q_m through the high-pressure compressor.

3.5. Claude Liquefier

Expansion through a J – T valve is an irreversible process; therefore, the performance of a liquefaction system can be improved by removing energy by expansion through a turbine or expander [17]. The Claude system, shown in Figure 8, uses an expander (c) to provide the bulk of the cooling effect in the system.

The liquid yield for the Claude system is given by Equation (9)

$$y = \frac{(h_1 - h_2) + x\eta_\mathrm{ad}(h_3 - h'_\mathrm{e})}{(h_1 - h_\mathrm{f})} \tag{9}$$

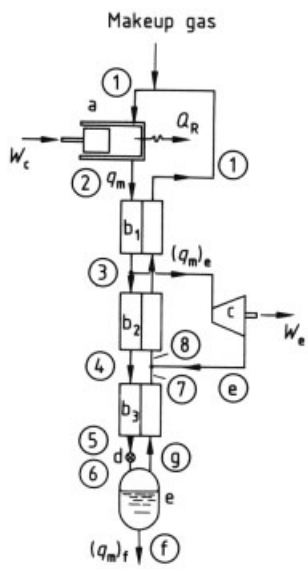

Figure 8. Claude liquefaction system
a) Compressor; b) Heat exchanger; c) Expander; d) J – T valve; e) Liquid receiver

where $x = (q_m)_e/q_m$ = expander mass flow ratio; $(q_m)_e$ = mass flow rate through the expander; η_{ad} = adiabatic efficiency of the expander; and h'_e = enthalpy at the expander outlet for an isentropic expansion. If the expander work is used to aid the compression of the gas, the work requirement per unit mass compressed is given by Equation (10):

$$-W/q_m = [T_1(s_1 - s_2) - (h_1 - h_2)]/\eta_c - x\eta_e(h_3 - h'_e) \quad (10)$$

where η_e = overall efficiency of the expander. In addition to the improved thermodynamic performance of the Claude system compared with that of the basic Linde – Hampson system, the Claude system has the advantage that it may be used without precooling for liquefaction of helium, hydrogen, and neon. If precooling is used for these gases, however, the system performance is improved significantly.

3.6. Liquefaction Systems for Natural Gas

Several plants for production of liquefied natural gas (LNG) operate on a modified cascade system, called the *mixed refrigerant cascade* (MRC) system [18], as shown in Figure 9. Operation of the system is possible because natural gas is composed of several components that condense at different temperatures. These component liquids are used to cool the feed stream, without requiring separate cooling circuits for the fluids, by carefully controlling the composition of the refrigeration cycle gas.

The natural gas feed stream generally enters the MRC system at ca. 4 – 5 MPa. The refrigerant cycle gas is compressed (a_1) and partially condensed in the compressor aftercooler. The propane-rich liquid phase is separated from the mixture (c_1), expanded through a J – T valve, and used as the coolant in the warm three-channel heat exchanger (b_1). The vapor from the phase separator is passed through the three-channel heat exchanger and partially liquefied; the ethane-rich liquid phase is separated in a second separator (c_2) to provide cooling in the intermediate three-channel heat exchanger (b_2). The vapor (which is mainly methane) from the second-phase-separator is cooled in the intermediate three-channel heat exchanger, expanded through a J – T valve, and recirculated to provide cooling in the cold three-channel heat exchanger (b_3).

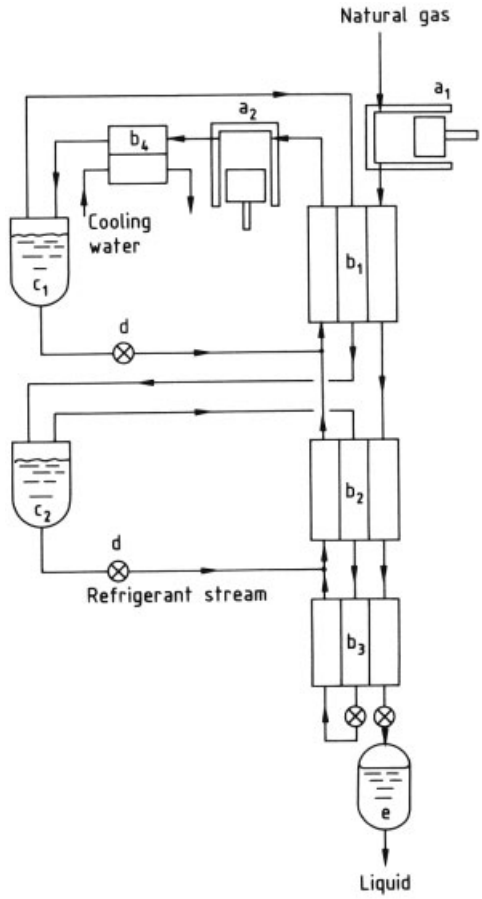

Figure 9. Mixed refrigerant cascade (MRC) system for liquefaction of natural gas
a) Compressor; b) Heat exchanger; c) Separator; d) J – T valve; e) Liquid receiver

3.7. Collins Helium-Liquefaction System

One of the most important developments in cryogenic liquefaction systems since the original construction of the Linde – Hampson and Claude liquefiers occurred in 1946 [19]. During that year, a helium liquefier was developed by Samuel Collins of the Massachusetts Institute of Technology. Commercial application was initially handled by the A. D. Little Co. of Cambridge, Mass.; later, commercial development

was assumed by an associated company, Cryogenic Technology, of Waltham, Mass. The introduction of the Collins helium liquefier transformed the existing situation, where research at liquid helium temperature was restricted to a few laboratories that were willing to endure the difficulties of constructing and operating helium liquefiers of their own design. The maintenance and operation of the special-purpose liquefiers were both difficult and expensive, whereas the Collins system was relatively simple to operate and required minimum maintenance because of its robust design (Fig. 10).

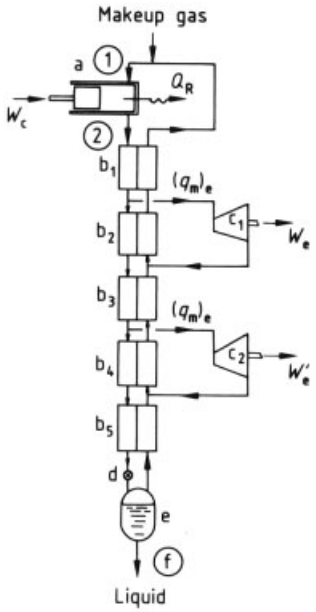

Figure 10. Collins helium liquefaction system
a) Compressor; b) Heat exchanger; c) Expander; d) J – T valve; e) Liquid receiver

Helium is compressed (a) to a pressure of ca. 1.27 MPa, and any entrained oil is removed in an oil trap. The compressed helium gas flows through the first heat exchanger (b_1), and a fraction (ca. 16 %) is bypassed through the first expander (c_1). The helium gas enters the expander at a temperature of 40 – 45 K. The remainder of the helium gas stream flows through the second heat exchanger (b_2), and leaves the heat exchanger at a temperature of ca. 15 K. A fraction of the stream (ca. 56 % of the total inlet flow) is then bypassed to the second expander (c_2). The remainder of the helium stream flows through the third heat exchanger (b_3) and expands through a J – T valve (d), in which a fraction of the stream is liquefied. The vapor formed during the expansion process is returned through the heat exchangers to provide cooling for the incoming helium gas stream.

The unit is enclosed in a Dewar vessel with a liquid-nitrogen-cooled shield to reduce ambient heat transfer. The heat exchangers and expanders operate in an atmosphere of low-pressure helium gas, so that minor leaks from the engines do not cause significant problems.

The liquid yield for the two-expander Collins liquefaction system is given by

$$y = \frac{(h_1 - h_2) + x\Delta h_e + x'\Delta h'_e}{h - h_f}$$

where $x = (q_m)_e/q_m$ = mass flow rate ratio for engine 1, $x' = (q_m)'_e/q_m$ = mass flow rate ratio for engine 2, Δh_e = enthalpy change through engine 1, and $\Delta h'_e$ = enthalpy change through engine 2. The expander work output in the Collins system is generally not used to aid in compression, but the energy is dissipated in a mechanical or hydraulic brake, which is external to the system. Liquid nitrogen is generally used to precool the system and to improve liquid yield. Typical operating characteristics of the Collins system are as follows [20]:

Liquid He production	25 – 32 L/h
Power requirement (He only)	45 kW
Helium compressed	5.2 kg/h
Liquid nitrogen consumption	25 L/h
Power requirement (including power for liquid N_2) per unit volume	3.1 kW · h/L

Collins liquefiers have also been constructed with as many as five expanders, depending on the inlet pressure of the helium gas.

3.8. Miniature Liquefaction Systems

Miniature cryogenic liquefaction systems have been used in two special applications: (1) cooling of electronic components in aerospace systems (particularly cooling of infrared detectors) and (2) providing cooling for cryosurgery applications. In both cases, liquefaction is based on the Linde – Hampson cycle.

Hymatic Cooler. A miniature liquefier used to cool electronic components is shown in Figure 11 [21]. Because of the small cross-sectional

area of the heat exchanger tubing, the system requires extremely pure air or nitrogen in its operation.

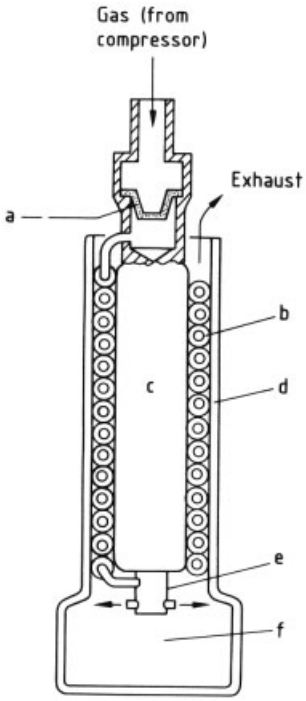

Figure 11. Miniature liquefier (Hymatic system)
a) Filter; b) Finned-tube heat exchanger; c) Center mandril; d) Outer jacket; e) Porous nozzle; f) Cold space

A sintered bronze filter (a) of 2 µm pore size at the inlet of the system excludes particles in the incoming stream. The heat exchanger (b) consists of miniature finned tubing wrapped around a center mandril (c). The outer jacket (d) around the tubing coil confines the flow of the return gas stream around the tubing. The high-pressure gas (6 – 40 MPa) flows inside the heat exchanger tube and is cooled by the return cold gas flowing over the fins.

The high-pressure gas expands through a porous nozzle into the cold space chamber, where the liquid formed during the expansion process is used to cool the electronic components.

Freezing of contaminant gases, such as CO_2 and water vapor, presents a serious operational problem with this type of liquefier. Typical maximum allowable concentrations of contaminants are as follows:

H_2O	1 mg/kg (dew point −77 °C)
CO_2	1.2 mg/kg
Hydrocarbons	0.045 mg/kg

CVI Liquefier. A similar miniature liquefier unit, using gaseous argon as the working fluid was developed by the CVI Corp. (Fig. 12), [22].

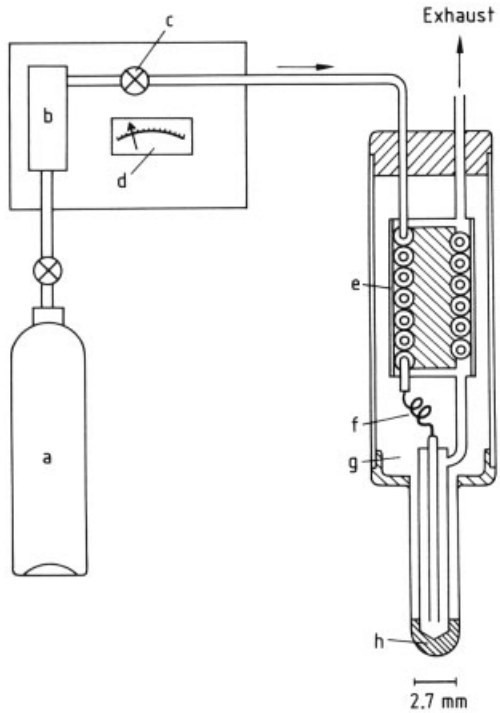

Figure 12. CVI liquefaction system for cryosurgery applications
a) Argon cylinder; b) Filter and dryer; c) Control valve; d) Temperature indicator; e) Heat exchanger; f) Capillary tube; g) Vacuum jacket; h) Probe tip

Compressed argon from a standard gas cylinder (a) passes through a filter and drier unit (b) to remove impurities. The heat exchanger (e) in the CVI system consists of miniature, spirally finned tubing wrapped around a formed mandril. The mandril grooves and spacers wrapped between the tube and the outer casing force the return gas stream to flow axially along the outside of the tube, thereby improving the effectiveness of the heat exchanger. The heat exchanger is 16 mm in diameter and 76 mm in length. The J – T valve is replaced with a fine capillary tube (f) having an inside diameter of a few micrometers. The liquid produced during the expansion process is passed through the cryosurgery probe

tip (h), where cooling freezes the tissue on the outside of the probe tip.

Precise temperature control is achieved by automatic or manual control (c) of the gas flow rate into the unit. A thermocouple in the tip senses the temperature of the probe tip. A small resistance heater within the probe tip is used when rapid thawing is required, as in eye surgery or retraction of the probe from a frozen tumor. The liquefier is enclosed in a vacuum jacket (g) having a 25-mm outside diameter to reduce heat transfer to the internal parts of the probe and to allow the probe to be hand-held.

3.9. Comparison of Liquefaction Systems

The thermodynamic performance of the liquefaction systems described in Sections 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, and 3.8, which use nitrogen as the working fluid, is given in Table 3 [23]. The first table entry is the thermodynamically ideal system, which has a liquid yield (y) of unity and a figure of merit (FOM) of unity. The remaining entries are presented for an ideal system (100 % effective heat exchangers and 100 % efficient compressors and expanders) and for a system with efficiency (η_c) and effectiveness (e) values commonly encountered in liquefier design.

The ideal precooled system and the ideal dual-pressure system have approximately the same FOM; however, the ideal Claude system has a significantly higher FOM. On the other hand, the Claude system is more sensitive to component inefficiencies. For example, the FOM of the basic Linde – Hampson system is decreased by a factor of 1.7 when component inefficiencies are included, whereas the FOM for the Claude system is decreased by a factor of 2.2 when inefficiencies are introduced. The performance of the Claude system is particularly sensitive to the efficiency of the expander, because a significant fraction of the cooling effect in the system is provided by expansion of the gas through the expander.

Table 3. Comparison of liquefaction systems ∗

Liquefaction system	Conditions	Liquid yield ∗∗ [$y=(q_m)_f/q_m$]	Work per unit mass liquefied ($-W/m_f$), kJ/kg	Figure of merit ∗∗ (FOM)
Ideal reversible system		1.000	768	1.000
Linde – Hampson				
Ideal system	$p_2=20.3$ MPa; $\eta_c=100$ %; $e=1.00$	0.071	6673	0.115
Practical system	$p_2=20.3$ MPa; $\eta_c=70$ %; $e=0.975$	0.058	11620	0.066
Precooled Linde – Hampson				
Ideal system	$r=0.15$; $p_2=20.3$ MPa; $T_3=240$ K; $\eta_c=100$ %; $e=1.00$	0.122	3945	0.195
Practical system	$r=0.15$; $p_2=20.3$ MPa; $T_3=240$ K; $n_c=70$ %; $e=0.975$	0.108	6317	0.122
Linde dual-pressure				
Ideal system	$p_2=6.1$ MPa; $p_3=20.3$ MPa; $\eta_c=100$ %; $(q_m)_i/(q_m)=0.80$; $e=1.00$	0.049	3679	0.209
Practical system	$p_2=6.1$ MPa; $p_3=20.3$ MPa; $\eta_c=70$ %; $(q_m)_i/(q_m)=0.80$; $e=0.975$	0.035	7346	0.105
Claude				
Ideal system	$p_2=6.1$ MPa; $x=0.60$; $\eta_c=\eta_{ad}=\eta_e=100$ %; $e=1.00$; $T_3=250$ K	0.258	1025	0.750
Practical system	$p_2=6.1$ MPa; $x=0.60$; $\eta_c=70$ %; $\eta_{ad}=80$ %; $\eta_e=72$ %; $e=0.975$; $T_3=250$ K	0.201	2232	0.344

∗ In all cases, the systems use nitrogen as the working fluid and are assumed to operate between 101.3 kPa at 300 K (26.8 °C) and the conditions stated; heat transfer from ambient is neglected.

∗∗ Compared to ideal reversible system; yield is the ratio of the mass of liquid produced to the mass compressed in the compressor.

4. Cryogenic Refrigeration Systems

Cryogenic refrigeration systems, called *cryocoolers*, are used in a variety of applications in which low temperature is required, including cooling of advanced electronic systems, superconducting magnets and motors, and space environment simulation chambers. In addition to cryocoolers that provide cooling by liquefying the working fluid, several others, including the Philips (based on the Stirling cycle), the Gifford–McMahon, and the Vuilleumier (VM), use entirely different concepts. Two special cryocoolers, i.e., the magnetic refrigerator and the dilution refrigerator, are used to achieve and maintain temperatures below 1 K.

4.1. Thermodynamically Ideal Refrigerator

The thermodynamically ideal refrigerator used for comparing the performance of real-life refrigerators is one of two types: (1) the *Carnot refrigerator*, for systems in which heat is absorbed by the refrigerator at constant temperature, and (2) the *isobaric-source refrigerator*, for systems in which the working fluid absorbs energy at constant pressure over a range of temperatures.

The figure of merit (FOM) for a refrigeration system is defined as the ratio of the ideal work requirement to the actual work requirement, for the same refrigeration effect. The FOM may be written in terms of the coefficient of performance (COP), defined by

$$\mathrm{COP} = \frac{Q_\mathrm{A}}{-W_\mathrm{net}} = \frac{Q_\mathrm{A}/q_\mathrm{m}}{-W_\mathrm{net}/q_\mathrm{m}}$$

where Q_A = energy absorbed from the low-temperature region, and W_net = net work input to the refrigerator; the work input is negative according to the thermodynamic sign convention. The FOM is then defined by Equation (11):

$$\mathrm{FOM} = \frac{(\mathrm{COP})_\mathrm{act}}{(\mathrm{COP})_\mathrm{id}} \tag{11}$$

where the subscripts act and id denote actual and ideal, respectively. The COP for the Carnot refrigerator depends only on the low and high temperatures of the cycle, and is given by Equation (12):

$$(\mathrm{COP})_\mathrm{id} = \frac{T_\mathrm{l}}{T_\mathrm{h} - T_\mathrm{l}} \tag{12}$$

where T_l = low temperature in the cycle, and T_h = high temperature in the cycle (usually ambient temperature). The COP for the isobaric-source ideal refrigerator is given by Equation (13):

$$(\mathrm{COP})_\mathrm{id} = \frac{h_2 - h_1}{T_\mathrm{h}\,(s_2 - s_1) - (h_2 - h_1)} \tag{13}$$

where points 1 and 2 are shown on Figure 13. If the specific heat of the material is approximately constant, Equation (13) may be written in terms of the temperatures, as follows:

$$(\mathrm{COP})_\mathrm{id} = \frac{T_2 - T_1}{T_\mathrm{h} \ln\,(T_2/T_1) - (T_2 - T_1)} \tag{14}$$

Equation (14) reduces to the expression for the Carnot COP (Equation 12) in the limit as T_2 approaches T_1.

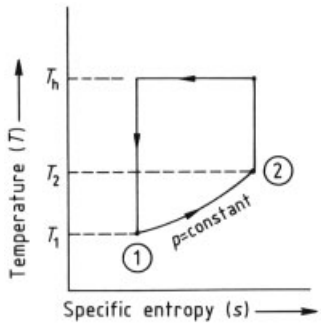

Figure 13. Thermodynamic cycle for the reversible isobaric-source refrigeration system

4.2. Joule–Thomson Refrigerators

Any of the liquefaction systems that do not use an expansion engine may be classed as Joule–Thomson (J–T) cryocoolers when operated as a refrigerator instead of as a direct liquefier. A schematic of a typical J–T cryocooler is shown in Figure 14. The refrigeration effect, $Q_\mathrm{A}/q_\mathrm{m}$, for this system is given by

$$Q_\mathrm{A}/q_\mathrm{m} = \left(h_1' - h_2\right) - (1-e)\left(h_1' - h_\mathrm{g}\right)$$

where e = heat exchanger effectiveness, h_1' = fluid enthalpy at pressure p_1 and temperature T_2, and h_g = saturated vapor enthalpy. The

work requirement per unit mass for the compressor is given by Equation (6); therefore, the COP for the J – T cryocooler may be written as follows:

$$\mathrm{COP} = \frac{\eta_c \left[(h_1' - h_2) - (1-e)\,(h_1' - h_g) \right]}{T_2\,(s_1 - s_2) - (h_1 - h_2)}$$

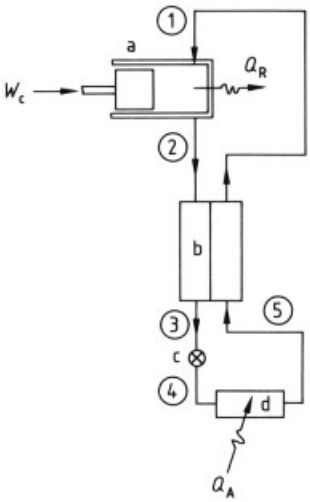

Figure 14. Joule – Thomson cryocooler (refrigerator)
a) Compressor; b) Heat exchanger; c) J – T valve; d) Evaporator

The liquid in the evaporator of the J – T cryocooler evaporates at a constant temperature; therefore, the FOM for the J – T system involves the COP for the Carnot refrigerator (Eq. 12). Liquid nitrogen may be used as the refrigerant for the temperature range of ca. 65 – 115 K. The low temperature limit is set by the triple-point temperature for the refrigerant and by problems in maintaining the low vacuum pressure with reasonable mass flow rates. The high temperature limit is set by the critical point for the refrigerant. If the evaporator pressure is lowered below the triple-point pressure, nitrogen "snow" forms in the evaporator and the expansion valve is likely to become clogged. In addition, the heat transfer in the evaporator is degraded. If the evaporator pressure is increased above the critical pressure, no liquid forms on expansion through the J – T valve. If maintenance of a constant temperature in the evaporator is not a concern, the temperature range for the J – T cryocooler using nitrogen gas can be extended up to ambient temperature; however, other working fluids, such as freon compounds, are more effective at temperatures above ca. 200 K. The temperature in the evaporator can be controlled by regulating the pressure in the evaporator.

Miniature J – T cryocoolers have been built by using an adsorption compressor [24].

Large quantities of gas are adsorbed by an adsorbent. The material is then heated in a closed container to produce a high-pressure gas. The gas flows through the J – T portion of the system, and the return gas is adsorbed in a second adsorbent chamber, which is cooled. The cold compressor temperature is 150 K, and the hot compressor temperature is 470 K. A COP of 0.22 was reported for nitrogen gas absorbing energy at 77 K, which corresponds to a FOM of 0.64.

4.3. CVI Cold-Gas Refrigerator

In many applications, a constant-temperature source is not required; helium or neon may be used as the working fluid, and the gas need not be condensed to provide refrigeration. The CVI cold-gas refrigerator (see Fig. 15) is an example of this type of cryocooler [25]. This unit has been used to provide cooling for the cryopanels in space environmental simulation chambers, for recondensation of boiled-off hydrogen gas, and for other applications involving cooling in the 20 K temperature range. The working fluid for the CVI refrigerator is helium.

The helium gas is compressed in a rotary compressor (a), and after passing through a filter system to remove traces of oil from the compressor, the gas is cooled in a three-channel heat exchanger (b). The high-pressure stream is further cooled in a liquid-nitrogen-cooled precooler (c) and a cold heat exchanger (d). Trace impurities in the helium, such as carbon dioxide, are removed in an adsorption purifier in the liquid nitrogen bath. The gas then expands through an expander (e) before passing through the load heat exchanger (which replaces the evaporator in J – T systems). The CVI refrigerator uses highly effective, brazed aluminum plate-fin heat exchangers.

The refrigeration effect for the CVI refrigerator is given by

$$Q_A/q_m = h_7 - h_6 = (h_7 - h_5) + \eta_{ad}\,(h_5 - h_6')$$

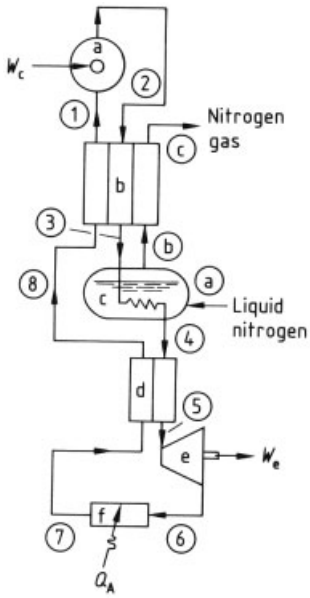

Figure 15. CVI cold-gas refrigerator using helium gas as the working fluid
a) Helium compressor; b) Heat exchanger; c) Precooler; d) Heat exchanger; e) Expander; f) Load heat exchanger

where η_{ad} = adiabatic efficiency of the expander, and h'_6 = enthalpy for an isentropic expansion from point 5 to 6. The system can operate without the liquid nitrogen precooling; however, the refrigeration effect is reduced by a factor of ca. 3. The mass of liquid nitrogen precoolant required per unit mass of helium gas is given by

$$\frac{(q_{\mathrm{m}})_{\mathrm{N}}}{(q_{\mathrm{m}})_{\mathrm{He}}} = \frac{(h_2 - h_1) + (h_8 - h_4)}{h_{\mathrm{c}} - h_{\mathrm{a}}}$$

The liquid nitrogen precoolant bath helps overcome the ineffectiveness of the warm heat exchanger in the system.

4.4. Stirling Refrigerator

The Stirling cycle was originally developed in 1816 by Robert Stirling for use in a hot-air engine [26]. The first Stirling cycle refrigerator was described by Alexander Kirk in 1864 [27]. Workers at the Philips Co. in Eindhoven were instrumental in developing the refrigerator into a commercially viable system in 1953 [28]. Since then, several other organizations have designed Stirling refrigerators for cooling infrared thermal imaging systems, spacecraft instrument systems, and superconducting elements.

A schematic of the operating cycle for the Stirling refrigerator is shown in Figure 16. The Stirling cryocooler consists of a cylinder enclosing a power piston (c) and a displacer piston (a). The two chambers in the engine are connected through a regenerator (b), which is a critical component of the cryocooler.

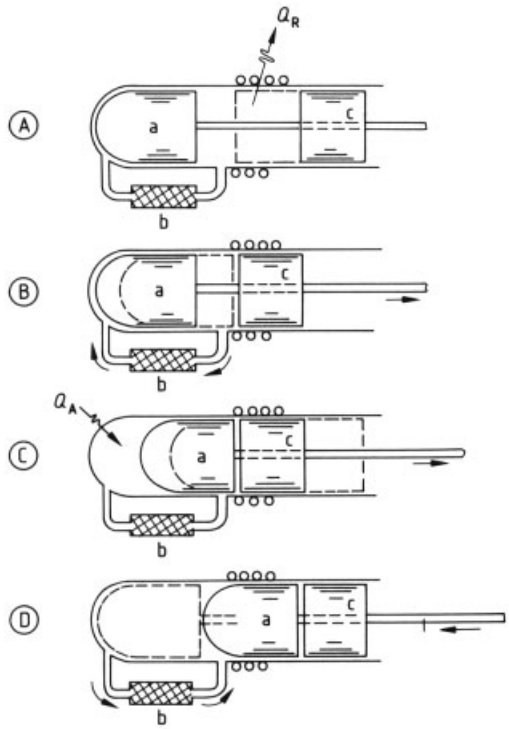

Figure 16. Sequence of operations for the Stirling cycle refrigerator
A) Process 1 → 2; B) Process 2 → 3; C) Process 3 → 4; D) Process 4 → 1
a) Displacer piston; b) Regenerator; c) Power piston

The sequence of processes for the Stirling cycle refrigerator is as follows:

Process 1 → 2 (Fig. 16A). The working fluid (usually helium gas, although hydrogen or neon can also be considered) is compressed by the power piston while heat is rejected to cooling coils at the warm end of the cylinder. In the ideal case, this compression process is isothermal.

Process 2 → 3 (Fig. 16B). With the power piston approximately stationary, the gas is forced through the regenerator by the motion of

the displacer piston. The gas is cooled by heat exchange with the cold regenerator matrix. In the ideal case, this is a constant-volume process.

Process 3 → 4 (Fig. 16C). The gas is expanded by motion of both the power and the displacer pistons, while heat is absorbed from the low-temperature region at the cold end of the piston. Ideally, this process is isothermal.

Process 4 → 1 (Fig. 16D). The cold gas is forced back through the regenerator by the motion of the displacer, while the power piston remains approximately stationary. The gas is warmed nearly to ambient temperature by heat exchange with the regenerator matrix. Ideally, this heating process is a constant-volume process.

For the ideal Stirling refrigerator, the COP is the same as that of the Carnot refrigerator, i.e., the FOM is unity.

$$\mathrm{COP} = \frac{T_3}{T_1 - T_3}$$

Frictional energy dissipation, pressure drops through the regenerator, and the ineffectiveness of the regenerator all interact to reduce the COP of the actual refrigerator to ca. 0.35, depending on the temperature at the cold end of the cylinder. An excellent analysis of the operation of the actual Stirling cryocooler has been given [16, pp. 131 – 142]; a more exact numerical analysis is presented in [29]. The success of the Stirling cryocooler depends largely on the effectiveness of the regenerator. The regenerator commonly used has a matrix of fine wires that results in a large surface area and high regenerator effectiveness. The fraction of the ideal refrigeration effect that is wasted because of the regenerator ineffectiveness is given by

$$\frac{\Delta Q}{Q'_{\mathrm{A}}} = \frac{(1-e)\,[(T_2/T_3)-1]}{(\gamma-1)\ln(v_4 v_3)}$$

where γ = specific heat ratio, v = specific volume, and Q'_{A} = ideal heat transfer rate from the low-temperature region; the numerical subscripts refer to the processes listed above. If the following numerical values are used, the importance of the regenerator may be demonstrated. For $\gamma = 1.67$ (helium gas), $v_4/v_3 = 1.50$, $T_2 = 300$ K, and $T_3 = 78$ K,

$$\Delta Q/Q'_{\mathrm{A}} = 10.48\,(1-e)$$

Thus, if the regenerator effectiveness is 0.99, the cooling effect loss is 0.108, i.e., >10 % of the ideal refrigeration effect is wasted because of a 1 % decrease in the regenerator effectiveness.

4.5. Vuilleumier Refrigerator

The Vuilleumier (VM) refrigerator, first patented by Rudolph Vuilleumier in 1918 in the United States [30], is similar to the Stirling cryocooler, except that the VM refrigerator uses a thermal compression process instead of the mechanical compression process used in the Stirling system. A schematic of the VM refrigerator is shown in Figure 17.

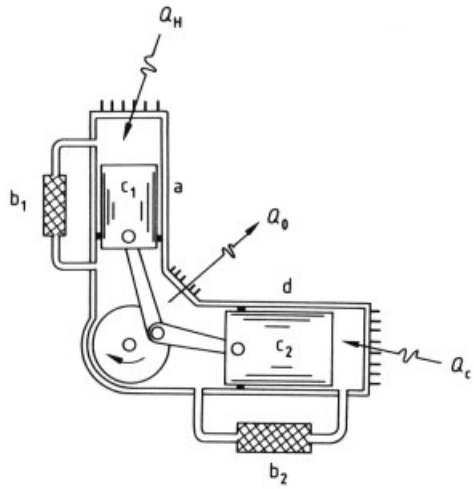

Figure 17. Vuilleumier refrigerator
a) Hot cylinder; b) Regenerator; c) Displacer piston; d) Cold cylinder

In the VM cryocooler, heat is added from a high-temperature source to the gas in the hot cylinder (a), and the displacer piston (c_1) moves downward to maintain the temperature of the gas constant at T_{h} [31]. At the same time, near-ambient-temperature gas flows from the intermediate volume through the regenerator (b_1) to the hot volume. The displacer then moves upward to displace the gas from the hot volume to the intermediate volume. Heat is rejected from the gas in the intermediate volume to maintain the temperature constant at T_0. The motion of the cold piston (c_2) is approximately 90° out of phase with the motion of the hot piston. As the cold piston moves to the left, heat is absorbed

by the gas in the cold volume from the low-temperature source to maintain the gas temperature constant at T_1. At the same time, gas from the intermediate volume flows through the cold regenerator (b_2) to the cold volume. The cold displacer moves back to the right, and gas is displaced from the cold volume back to the intermediate volume.

The ideal COP for the VM cryocooler is given by

$$\mathrm{COP} = \frac{Q_c}{Q_h} = \frac{T_c\,(T_h - T_o)}{T_h\,(T_o - T_c)}$$

where the subscript c denotes the cold cylinder, h the hot cylinder, and o ambient temperature. The factors that contribute to reduction of the COP for the actual system include the following: (1) shuttle loss, caused by the mismatch between the displacer and the cylinder wall, (2) pumping loss, caused by the gas volume between the displacer and the cylinder wall, (3) heat transfer through the displacers from the warmer end to the cooler end, (4) heat transfer through the cylinder wall from the warmer end to the cooler end, (5) frictional energy dissipation between the cylinder wall and the displacer seals, and (6) losses because of the ineffectiveness of the regenerators. An advantage of the VM cryocooler is that the thermal input may be provided by isotope energy or solar energy. This makes the unit attractive for cryogenic cooling in long-term space exploration and in applications where mechanical vibration of the drive engine must be avoided [16, pp. 186 – 187].

4.6. Solvay Refrigerator

The Solvay refrigerator was originally developed in Germany in 1887 [32]. It was one of the first refrigerators planned for use in an air liquefaction system that had an expansion engine. However, Solvay's prototype system was unable to achieve temperatures <178 K. The system was not considered for cryogenic refrigeration until the late 1950 s, when Gifford and McMahon of A. D. Little constructed a Solvay refrigerator for cooling a miniature infrared detector system [33]. A schematic of the Solvay cryocooler is shown in Figure 18.

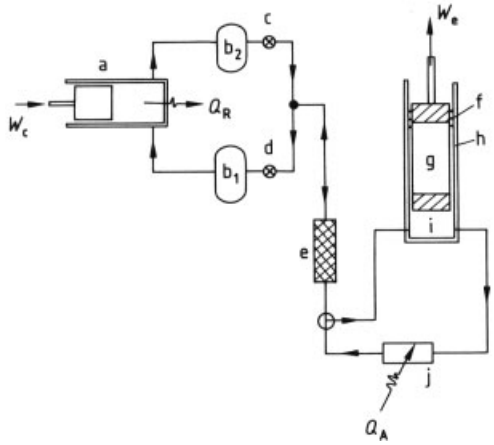

Figure 18. Solvay cryocooler
a) Compressor; b) Surge vessel (b_1 is at p_1, i.e., low pressure in the cycle, b_2 is at p_2, i.e., high pressure in the cycle); c) Inlet valve; d) Outlet valve; e) Regenerator; f) Seal; g) Displacer piston; h) Cylinder; i) Expansion space; j) Load heat exchanger

The system consists of a piston – cylinder arrangement (g) and (h) a regenerator (e), and a compressor (a) with inlet (c) and outlet (d) control valves. The cold part of the piston is at the lower end, whereas the upper portion of the cylinder operates at ambient temperature. The cylinder is constructed of thin-walled stainless steel tubing, and the piston is constructed of a low-conductivity plastic material so that heat conduction from the warm end to the cold is minimized. A sliding seal (f) is placed at the warm end of the piston to prevent leakage of the working gas from the system. Problems with operating a seal at cryogenic temperatures are avoided by placing the seal at the warm end. The piston generally fits quite loosely within the cylinder, with up to 0.5 mm clearance, except for the gas-tight seal at the warm end.

The regenerator uses punched copper disks or brass screens as the matrix. The screens or disks are separated by stainless steel wire spacers to prevent axial conduction along the regenerator.

The inlet and outlet control valves are coupled to the piston crankshaft so that the valves may be opened and closed at the appropriate time, depending on the position of the piston. These valves control the flow into and out of the expansion engine.

The sequence of operations for the Solvay cycle is as follows:

Process 1 → 2 (pressure buildup). With the piston at the bottom of the stroke, the inlet valve is opened, while the outlet valve remains closed. The high-pressure gas from the compressor flows through the regenerator, in which the gas is cooled; the system pressure increases.

Process 2 → 3 (intake stroke). With the inlet valve open and the outlet valve closed, the piston is raised to draw a volume of gas, which has been cooled by passage through the regenerator, into the cylinder.

Process 3 → 4 (expansion stroke). The inlet valve is closed and the piston continues to move toward the top dead center position. The gas within the cylinder expands (i) while doing work against the piston, and the gas temperature decreases to a low level.

Process 4 → 5 (exhaust process). When the piston reaches the top of its stroke, the outlet valve is opened and the gas pressure falls to the final minimum value. The exhaust value remains open as the piston moves back toward the bottom position to force the cold gas out of the cylinder through a load heat exchanger. Energy is absorbed from the external region, which is to be cooled during this process.

Process 5 → 1. The cold gas passes back through the regenerator, where it is warmed back to ambient temperature. The outlet valve is then closed, the inlet valve is opened, and the cycle is ready to be repeated.

The refrigeration effect for this cryocooler is given by

$$Q_{\mathrm{A}}/q_{\mathrm{m}} = \left(h_5 - h_4'\right) - (1-\eta_{\mathrm{ad}})\left(h_3 - h_4'\right)$$

where h_3 = enthalpy at the expander inlet, h_4' = enthalpy for an isentropic expansion from pressure p_3 to p_4, h_5 = enthalpy of the stream leaving the load heat exchanger, and η_{ad} = adiabatic efficiency for the expansion process between point 3 and point 4 (see process 3 → 4).

The *miniature Solvay cryocooler* described by GIFFORD and MCMAHON can maintain a temperature of 55 K with a single stage. The cylinder is a stainless steel tube having a 5-mm outside diameter, a 0.13-mm wall thickness, and a length of 51 mm. The piston is constructed of a linear base-laminated phenolic compound. The regenerator was actually contained within the piston in their design. The working fluid is helium gas, which varies in pressure between 0.345 and 1.725 MPa. The cool-down time for the unit is ca. 3 min.

4.7. Gifford – McMahon Refrigerator

A schematic of the Gifford – McMahon cryocooler is shown in Figure 19 [34]. This system differs from the Solvay refrigerator in that two spaces are used within the cylinder, and the piston does little work as it moves within the cylinder. The displacer piston simply moves the gas from one volume to the other during the cycle.

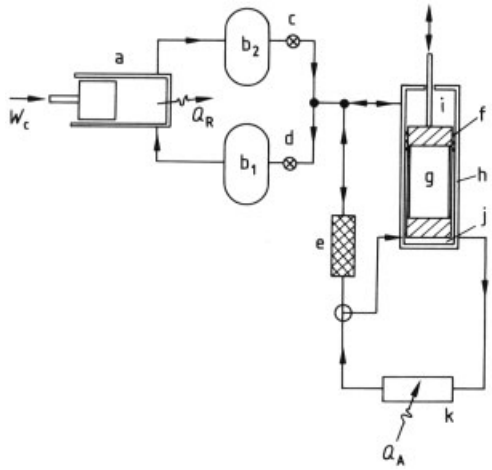

Figure 19. Gifford – McMahon refrigerator
a) Compressor; b) Surge vessel (b_1 is at p_1, b_2 is at p_2); c) Inlet valve; d) Outlet valve; e) Regenerator; f) Seal; g) Displacer piston; h) Cylinder; i) Upper expansion space; j) Lower expansion space; k) Load heat exchanger

The sequence of operations for the Gifford – McMahon cycle is as follows:

Process 1 → 2 (pressure buildup). With the displacer (g) at the bottom of the cylinder (h) and the outlet valve (d) closed, the inlet valve (c) is opened and the pressure within the cylinder is increased from the low pressure p_1 to the high pressure p_2. The volume of the lower expansion space (j) is practically zero during this process.

Process 2 → 3 (intake stroke). With the inlet valve open and the outlet valve closed, the displacer piston is moved from the bottom position to the top of the cylinder. This piston motion moves the gas that was originally in the upper (warm) expansion space (i) through the regenerator to the lower (cold) expansion space. Because the gas is cooled as it passes through the regenerator (e), its volume tends to decrease and

additional gas is drawn in through the inlet valve to maintain a constant pressure within the system.

Process 3 → 4 (pressure release and expansion). With the displacer at the top of the cylinder, the inlet valve is quickly closed and the outlet valve is slowly opened, thus allowing the gas within the lower expansion space to expand from pressure p_2 to the low pressure p_1. The gas that remains in the lower expansion space does work in forcing out the gas that leaves during this process; therefore, the temperature of the remaining gas drops. Ideally, the expansion takes place reversibly and adiabatically (isentropic expansion).

Process 4 → 5 (exhaust stroke). With the outlet valve open, the low-temperature gas is forced from the lower expansion space by moving the displacer down to the bottom of the cylinder. The cold gas flows through a heat exchanger (k) to absorb energy from the low-temperature source.

Process 5 → 1 The gas flows back through the regenerator, where it is warmed to ambient temperature, and the cycle is ready to be repeated.

The refrigeration effect for the Gifford–McMahon system is given by

$$Q_{\mathrm{A}}/q_{\mathrm{m}}=\eta_{\mathrm{ad}}\left[(q_{\mathrm{m}})_{\mathrm{e}}/q_{\mathrm{m}}\right]\left(h_5-h_4'\right)$$

where η_{ad}= adiabatic efficiency, for the expansion process from point 3 to point 4 (process 3 → 4), $(q_{\mathrm{m}})_{\mathrm{e}}$= mass of gas within the lower expansion space at the end of the expansion process, q_{m}= total mass compressed, h_4' = enthalpy for an isentropic expansion from point 3 to point 4, and h_5 = enthalpy of the stream leaving the heat exchanger. Because the volume of the cold expansion space remains essentially constant during the expansion process, the mass ratio may be calculated from the density ratio:

$$(q_{\mathrm{m}})_{\mathrm{e}}/q_{\mathrm{m}}=\varrho_4'/\varrho_3$$

The Gifford–McMahon refrigerator has several advantages over Claude-type refrigerators, which also use an expander. In the Gifford–McMahon cryocooler, all valves and seals operate at ambient temperature; therefore, problems with low-temperature seals and valves are eliminated. Problems with leakage around the piston seals are reduced because the viscosity of a gas near ambient temperature is higher than the gas viscosity at cryogenic temperatures. Whereas low-temperature operation requires special dry sealing rings, the Gifford–McMahon system allows the seals to be constructed of conventional materials.

The Gifford–McMahon refrigerator uses a regenerator instead of the heat exchangers employed in the Claude and J–T refrigerators. In general, regenerators are less expensive than heat exchangers, and the regenerator can be constructed with a larger heat transfer surface area per unit volume, which is important for miniature systems. Because of the back-and-forth motion of the gas through the regenerator, impurities deposited during the intake process are swept out of the system during the exhaust stroke; thus, problems with impurity buildup are minimized with the use of the regenerator.

The Gifford–McMahon system is also well-suited for multistaging [35]. In some cryogenic systems, refrigeration is required at intermediate temperatures, for example, to cool radiation shields around the unit. In this case the system can be operated with two stages, one for cooling the shields at 77 K, for example, and the other stage to provide the low-temperature refrigeration at 20 K. In addition, a multistage refrigerator generally has a higher thermodynamic performance than a single-stage unit. A typical two-stage Gifford–McMahon refrigerator is shown in Figure 20.

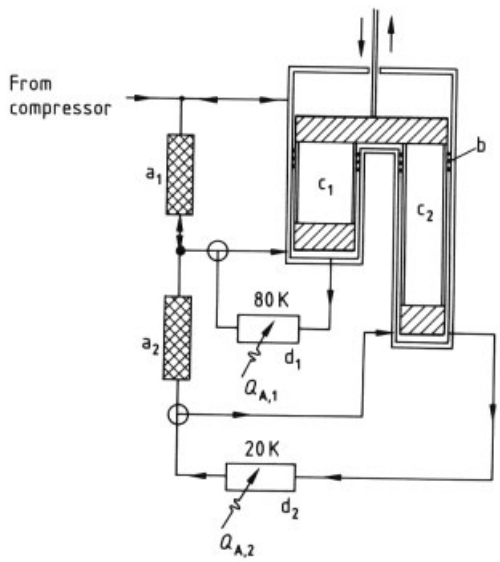

Figure 20. Two-stage Gifford–McMahon refrigerator
a) Regenerator; b) Seals; c) Displacer piston; d) Load heat exchanger

The principal disadvantage of the Gifford–McMahon cryocooler is that the COP is inherently smaller than that of the Claude cryocooler operating between the same pressure limits, because only a fraction of the total mass compressed in the Gifford–McMahon system is expanded during the expansion process.

4.8. Magnetic Refrigerators

When a liquid or gas is used as the working fluid in a refrigerator, the operating temperature is limited to values above ca. 1 K. All materials other than ordinary helium (^{4}He) and the helium isotope ^{3}He are in the solid phase at temperatures below 1 K. The low temperature that can be attained with boiling helium is determined by the vapor pressure of the helium. At 1 K, the vapor pressure of ^{4}He is 16 Pa, and the vapor pressure of ^{3}He is 1180 Pa. At 0.5 K, the vapor pressures are 1.9 mPa (^{4}He) and 22.2 Pa (^{3}He) [36]. Clearly, a refrigerator operating at these low pressures in the evaporator is impractical for relatively large capacities because of the size of the flow passages required for even moderate mass flow rates.

History. Giauque and Debye independently suggested a method to break this "temperature barrier" in the 1920 s [37, 38]: A paramagnetic material could be used instead of a liquid or gas as the working medium to reach temperatures <1 K, because the entropy associated with the electron-spin system becomes much larger than the entropy associated with the lattice vibrations. In 1933 Giauque and MacDougall (who won a Nobel prize for the feat) were able to verify the magnetic cooling concept experimentally [39]. They reached a temperature of 0.53 K with gadolinium sulfate after demagnetizing from a field of 0.8 T and an initial temperature of 3.5 K. DeHaas and coworkers attained a temperature of 0.27 K 2.5 weeks later by demagnetizing CeF_3 from 2.8 T at 1.3 K [40].

The principle of cooling by adiabatic demagnetization is similar to the cooling effect produced when a gas expands isentropically. When a gas is compressed isothermally, the entropy of the gas decreases because the order of the molecules increases (the atoms occupy a smaller volume with no increase in random molecular velocities). If the high-pressure gas is expanded reversibly and adiabatically, the entropy of the gas remains constant; therefore, the temperature of the gas decreases to offset the increase in disorder produced by moving the molecules farther apart.

A paramagnetic material produces order (or reduces entropy) differently. In the absence of a magnetic field, the dipoles of the magnetic material are more or less randomly arranged because of the disturbing effect of random thermal vibrations. If an external magnetic field is applied to the material while the material temperature is maintained constant, the entropy or disorder of the material decreases, because the dipoles tend to line up with the magnetic field. If the magnetic field is reversibly and adiabatically removed, the temperature of the paramagnetic material decreases to maintain constant entropy. This process, which is called *adiabatic demagnetization*, is the basis for the magnetic refrigerator used to reach temperatures down to 0.1 K.

A schematic of the basic magnetic refrigerator is shown in Figure 21. Lines of constant magnetic field strength replace the constant pressure lines on the temperature–entropy diagram for the cycle. The refrigerator consists of the following: a working salt (e), such as iron ammonium alum, which is the cooling medium for the refrigerator; a reservoir salt (f), which acts as a "thermal flywheel" to smooth out temperature fluctuations in the space to be cooled; thermal valves (b) to control heat transfer to and from the working salt; and solenoids (c and d), which are used to apply the required magnetic fields to the working salt and the thermal valves.

The sequence of operations for this refrigerator is as follows:

Process $1 \rightarrow 2$ (corresponds to the isothermal compression process in the Carnot cycle). With the upper thermal valve open (heat transfer is allowed) and the lower thermal valve closed (no heat transfer), the magnetic field around the working salt is increased from 0.14 to 0.72 T. Heat transfer occurs between the working salt and the liquid helium bath to maintain a constant temperature in the working salt. The working salt is thermally isolated from the reservoir salt during this process.

Process $2 \rightarrow 3$ (corresponds to the reversible adiabatic expansion process in the Carnot cycle).

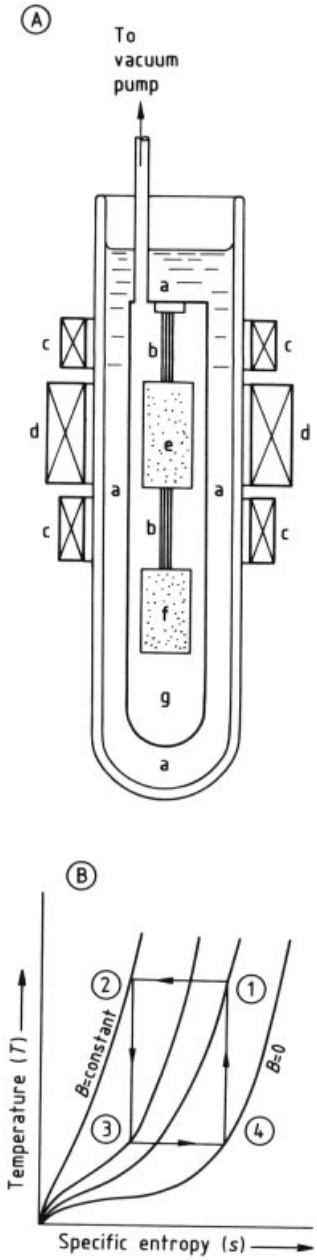

Figure 21. Magnetic refrigerator
A) Schematic diagram: a) Liquid helium; b) Thermal valve; c) Valve magnet; d) Main magnet; e) Working salt; f) Reservoir salt; g) Space to be cooled
B) Thermodynamic diagram for the cycle

Both thermal valves are closed so that the working salt is thermally isolated (adiabatic process) and the main magnetic field is reduced from 0.72 to 0.10 T. The temperature of the working salt decreases to ca. 0.16 K from an initial temperature of 1.15 K.

Process $3 \rightarrow 4$ (corresponds to the isothermal heat addition process in the Carnot cycle). The thermal valve between the working salt and the reservoir salt is opened to allow heat exchange between the two materials, and the magnetic field around the working salt is reduced to zero. Heat is transferred to the working salt from the reservoir salt to maintain constant temperature.

Process $4 \rightarrow 1$ (corresponds to the reversible adiabatic compression process in the Carnot cycle). The working salt is again thermally isolated by closing both thermal valves, and the magnetic field around the working salt is increased from 0 to 0.14 T. The temperature of the working salt increases to the initial value, and the cycle has been completed.

The energy absorbed as heat from the low temperature region (the reservoir salt) in the ideal case is given by Equation (15)

$$Q_{\mathrm{A}} = q_{\mathrm{m}} T_3 \left(s_4 - s_3\right) \tag{15}$$

where q_{m} = mass of the working salt and s = magnetic entropy of the working salt. Ideally, the heat rejected from the working salt to the helium bath is given by

$$Q_{\mathrm{R}} = q_{\mathrm{m}} T_1 \left(s_2 - s_1\right) = -q_{\mathrm{m}} T_1 \left(s_4 - s_3\right)$$

The net magnetic work done on the working salt is equal to the net heat transfer for a cycle; therefore,

$$W_{\mathrm{net}} = q_{\mathrm{m}} \left(T_1 - T_3\right) \left(s_4 - s_3\right) \tag{16}$$

Equations (15) and (16) show that the COP of the ideal cycle is the same as that of a Carnot cycle. Irreversibilities reduce the performance of the actual refrigerator below that of the ideal system. For example, at a source temperature of 0.4 K, the Daunt – Heer – Barnes refrigerator [41] has a refrigeration capacity of 27.5 μW, whereas the refrigeration capacity of the ideal system operating under the same conditions is 34.4 μW. At a source temperature of 0.3 K, the refrigeration capacity is 14.3 μW, compared with 25.8 μW for the ideal system.

One of the advantages of the magnetic refrigerator is that its operation is not dependent on the presence of a gravitational force; thus, the system can operate effectively in a zero-gravity environment. Infrared bolometers in space systems have been cooled by magnetic refrigerators.

One of the critical components of the magnetic refrigerator is the thermal valve or "switch." In the Daunt – Heer – Barnes refrigerator, a thin lead strip is used as the thermal valve. At a temperature of 1 K, the thermal conductivity of lead in the superconducting state is ca. 0.025 times the thermal conductivity of the material in a normal state; at 0.2 K, the ratio is ca. 0.010. Thus, the valve can be "closed" (little heat transfer through the valve) by decreasing the magnetic field around the valve to drive the lead into the superconducting state. The valve can be "opened" by increasing the magnetic field so that the lead strip returns to the normal state. Changes in magnetic field around the thermal valves are accomplished with small "valve magnets" around the lead strips. Other types of thermal switches have been tried, but they are generally not as effective as the superconducting switch. For example, mechanical switches allow zero thermal leakage when contact is broken; however, energy is dissipated when the contact is made or broken, and this energy dissipation cannot be tolerated in the region below 1 K.

Magnetic wheel refrigerators have been constructed for use in the temperature range between 2 and 10 K [42, 43]. A schematic of this type of refrigerator is shown in Figure 22.

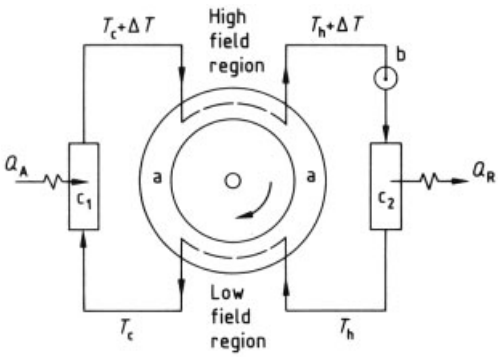

Figure 22. Schematic of the magnetic wheel refrigerator a) Wheel; b) Pump; c) Heat exchanger

Helium gas enters the wheel (a) in the low-field region at a temperature of T_h and is cooled to T_l by heat exchange with the magnetic material. The wheel is constructed of granules of gadolinium sulfate for improved heat exchange with the helium gas. The gas leaves the low-field region of the wheel and flows through a heat exchanger (c_1) in which the cooling load is absorbed by the helium gas. The gas enters the high-field region of the magnetic wheel at temperature ($T_c + \Delta T$), where ΔT is the intrinsic temperature change of the magnetic material as it moves from the low-field region to the high-field region. The helium gas is warmed in the high-field region to a temperature ($T_h + \Delta T$) by heat exchange with the magnetic material. The helium gas finally flows through a warm heat exchanger (c_2), in which heat is rejected to the heat sink, and the gas is cooled back to temperature T_h.

In a working model of the magnetic wheel refrigerator, a 3-T magnetic field was used in the high-field region, and the rotational speed of the 150-mm diameter wheel was 6 rpm. A temperature difference of 7 K was maintained with a 500-W thermal load on the cold side.

4.9. Dilution Refrigerator

One of the most widely used refrigerators for the temperature range between 0.005 and 1 K is the ^{3}He – ^{4}He dilution refrigerator. This is a relatively new development in cryogenic technology.

History. In 1951 H. London suggested that cooling could be achieved by dilution of liquid ^{3}He with superfluid liquid ^{4}He [44]. Little interest existed for developing a practical refrigerator at that time, because the phase separation of ^{3}He – ^{4}He mixtures had not been discovered. London [45] presented a practical design for the dilution refrigerator in 1962 after the phase separation phenomenon was discovered experimentally by Walters and Fairbank in 1956 [46]. The first dilution refrigerator was constructed by Das and coworkers [47] at the University of Leiden in 1965. In 1966 Hall and associates [48] and Neganov and others [49] in Russia constructed dilution refrigerators that achieved temperatures between 0.025 and 0.060 K. Commercial dilution refrigerators can continuously reach 0.005 K.

The entropy of pure liquid ^{4}He below 1 K is quite small and varies proportionally with T^3. On the other hand, the entropy of liquid ^{3}He below 1 K is somewhat larger and is proportional

to T. Thus, the ^{4}He in a ^{3}He – ^{4}He mixture contributes little to the total entropy of the mixture below 1 K.

Because ^{3}He has a higher entropy in the dilute phase mixture (dilute in ^{3}He), an amount of heat $Q = T \Delta S$ is absorbed for each mole of ^{3}He that passes from the pure phase to the dilute phase mixture in a reversible manner. The main problem is how to mix the two materials in a reversible manner. Fortuitously, nature provides a reversible mechanism for the dilution: the helium mixture separates into a dilute phase and a concentrated phase at temperatures below 0.86 K. The concentrated phase is less dense than the dilute phase and, therefore, floats on top of the dilute phase mixture. The concentrated phase is a normal fluid, whereas, the ^{4}He component in the dilute phase is a superfluid. Because the two phases are in thermodynamic equilibrium, ^{3}He can move reversibly from the low-entropy concentrated phase to the high-entropy dilute phase, analogous to the evaporation of liquid molecules into the vapor phase above the liquid. In addition, the mole fraction of ^{3}He in the dilute phase at absolute zero is not zero, but instead approaches a finite value of $x_D = 0.064$. This fact means that at low temperature the high-entropy phase has a particle density that is high enough to permit relatively high flow rates for the ^{3}He, which is important in the design of a practical refrigerator.

One other property that permits design of a practical dilution refrigerator is the fact that ^{3}He has a much higher vapor pressure than ^{4}He at the same temperature. This means that the vapor in equilibrium with the liquid phase is practically pure ^{3}He for temperatures below 1 K. A schematic of a ^{3}He – ^{4}He dilution refrigerator is shown in Figure 23.

The gas (which is practically pure ^{3}He) is compressed in a vacuum pump (a) from ca. 4 Pa to 4 kPa. The gas is cooled in a heat exchanger (b_1) and a liquid helium bath (c_1) at 4.2 K, and is then condensed in a liquid helium bath (c_2) at reduced pressure and a temperature of ≤ 1.2 K. The liquid ^{3}He is expanded through a capillary tube constriction (d) and is cooled further in the still (e), which operates at ca. 0.6 K. The liquid is further cooled in a heat exchanger (b_2) before it enters the mixing chamber (f), where ^{4}He is mixed with the ^{3}He at temperatures between 0.050 and 0.005 K.

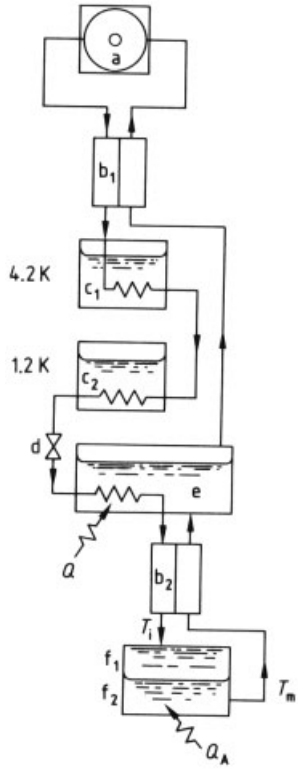

Figure 23. ^{3}He – ^{4}He dilution refrigerator
a) Vacuum pump; b) Heat exchanger; c) Helium bath; d) Constriction; e) Still; f) Mixing chamber: f_1 is concentrated mixture (85 % ^{3}He), f_2 is dilute mixture (6.4 % ^{3}He)

The liquid in the mixing chamber separates into two liquid phases, a less dense concentrated ^{3}He mixture (f_1) and a more dense diluted ^{3}He mixture (f_2). The ^{3}He molecules diffuse from the concentrated phase to the dilute phase, with heat being absorbed during this process. The dilute mixture returns through the heat exchanger to the still, where heat is added to evaporate the ^{3}He from the mixture. The mole fraction of ^{3}He in the liquid in the still is ca. 1 %, and the composition of the vapor is between 95 and 98 % ^{3}He.

The refrigeration effect of the dilution refrigerator is given by

$$Q_A = (q_v)_{He} (h_m - h_i)$$

where $(q_v)_{He}$ = molar flow rate of ^{3}He, h_m = molar enthalpy leaving the mixing chamber, and h_i = molar enthalpy of the ^{3}He entering the mixing chamber. For temperatures below ca.

0.04 K, the enthalpies may be approximated by Equations (17) and (18) [50]:

$$h_m = C_1 T_m{}^2 \text{ where } C_1 = 94\,\mathrm{J\,mol^{-1}\,K^{-2}} \quad (17)$$

$$h_i = C_2 T_i{}^2 \text{ where } C_2 = 12\,\mathrm{J\,mol^{-1}\,K^{-2}} \quad (18)$$

One of the more critical components of a dilution refrigerator is the heat exchanger between the mixing chamber and the still, which must have a high effectiveness; otherwise the system will not work. From Equations (17) and (18), the maximum temperature ratio (corresponding to zero heat transfer in the mixing chamber) is $T_1/T_m = 2.80$. For a value of $T_m = 0.040$ K, the largest value for T_1 for a positive refrigeration effect is 0.112 K, which yields a maximum temperature difference at the cold end of the exchanger of $(T_1 - T_m) = 0.072$ K. Heat exchangers with large surface areas per unit volume have been developed using sintered metal elements within the flow passages [51].

5. Components of Liquefiers and Cryocoolers

The three important subassemblies of liquefaction systems and cryogenic refrigerators are the heat exchanger or regenerator, the compressor, and the expander. Other components in these systems include the J – T valve, filters, and purifiers. However, this chapter addresses only the first three items.

5.1. Cryogenic Heat Exchangers

One of the most critical components in any liquefaction system or cryogenic refrigerator is the heat exchanger. In conventional systems using heat exchangers, such as regenerative gas turbine units, the system can operate satisfactorily with a heat exchanger having an effectiveness (e) as low as 0.50. However, a heat exchanger effectiveness of 0.85 or higher is required if the liquefier is to produce liquid at all. For example, nitrogen has the following enthalpy values [52]:

$h_1' = 462.15$ J/g at 300 K and 0.101 MPa
$h_2 = 431.50$ J/g at 300 K and 20.3 MPa
$h_g = 228.66$ J/g at 0.101 MPa and saturated vapor.

Using Equation 5 with zero heat inleak from ambient, the minimum heat exchanger effectiveness (e_{min}) for the basic Linde – Hampson liquefier of J – T cryocooler is found to be

$$e_{min} = 1 - \frac{(462.15 - 431.50)}{(462.15 - 228.66)} = 0.869$$

The effectiveness of the heat exchanger must be higher than 0.869 in this example for the liquefier to work. The importance of the heat exchanger may also be demonstrated by noting that the liquid yield for a Linde – Hampson system with a perfect heat exchanger is 0.0708, whereas for a heat exchanger with an effectiveness of 0.99, the liquid yield for the Linde – Hampson liquefier is decreased to 0.0658. A reduction of the effectiveness by 1 % results in a 7.1 % reduction in the liquid yield.

5.1.1. Types of Cryogenic Heat Exchangers

Many different types of heat exchangers (also called *recuperators*) are used in cryogenic liquefaction and refrigeration systems. The units may be classified as follows:

1) tubular exchangers
2) plate-fin exchangers
3) perforated-plate exchangers

Tubular Exchangers. Four basic tubular exchangers are shown in Figure 24. The simplest form of tubular exchanger is the *double-pipe or concentric-tube heat exchanger* (Fig. 24A, see next page). This was the unit used by Linde in his original liquefier. The fluids are in counterflow, with the high-pressure warm stream (a) in the inner tube and the low-pressure cold stream (b) in the annular region between the two tubes. To conserve space, the concentric tubes are wound in a large helix, and the unit is insulated to reduce heat transfer from the surroundings.

Concentric-Tube Exchanger with Wire Spacer. The performance of the concentric-tube exchanger may be improved by wrapping a wire or plastic spacer around the inner tube (Fig. 24B). This spacer tends to maintain a concentric relation of the inner tube with the outer tube and forces the low-density gas in the annular space (b) to follow a long helical path,

thereby increasing the gas velocity and improving the heat transfer coefficient.

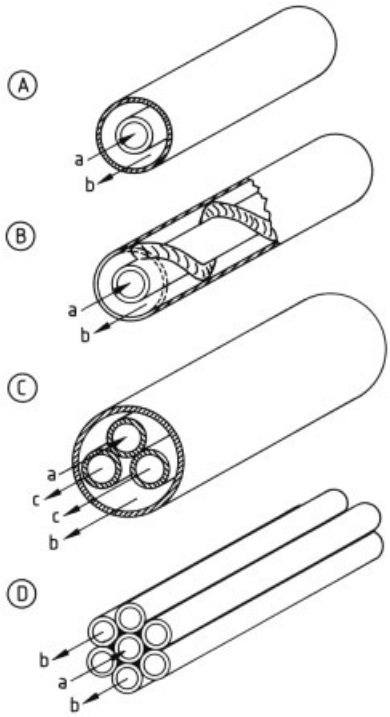

Figure 24. Cryogenic heat exchangers
A) Linde concentric-tube exchanger; B) Concentric-tube exchanger with wire spacer; C) Multiple-tube exchanger; D) Bundle-tube exchanger
a) High-pressure stream; b) Low-pressure stream; c) Intermediate-pressure stream

Several tubes may be contained inside a larger tube to form a *multiple-tube exchanger* (Fig. 24C). This unit can be used in the Linde dual-pressure system for the three-channel heat exchanger. In this case, the high-pressure stream (a) flows in one of the smaller tubes, the intermediate-pressure stream (c) flows in the other smaller tubes, and the low-pressure stream (b) flows in the space between the small tubes and the larger enclosing tube. Spiral windings, spacers, or longitudinal fins are used to locate the small tubes within the larger tube.

The *bundle- or cluster-tube heat exchanger* is formed by brazing or soldering several tubes around a central tube. The high-pressure stream generally flows in the central tube (a), while the low-pressure stream flows in the tubes (b) clustered around the center tube.

Giauque – Hampson Heat Exchanger. The classic heat exchanger used for medium- to large-scale cryogenic liquefiers is the Giauque – Hampson type shown in Figure 25.

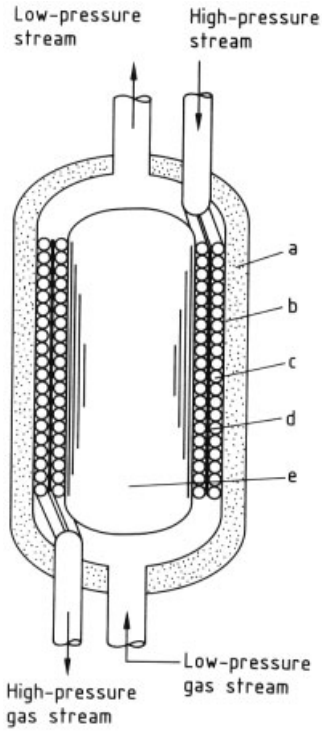

Figure 25. Giauque – Hampson heat exchanger
a) Insulation; b) Close-fitting sheath; c) Spiral-wound tubes; d) Punched spacer; e) Cylindrical mandrel

One or more coils of thin-walled stainless steel, aluminum, or copper tubes (c) are wound onto a low-conductivity mandrel (e), such as a stainless steel cylinder. A close-fitting sheath (b) of the same low-conductivity material is placed around the coil, and insulation (a) is added on the outside. The high-pressure stream flows inside the small tubes, while the low-pressure stream flows across the outside of the helical coil of small tubes between the inner mandrel and the outer sheath.

Uniform tube spacing is important in the construction of the Hampson heat exchanger; otherwise, the low-pressure gas tends to follow the path of least frictional resistance (through the widest spacing), and the flow is distributed unevenly. An important contribution to solve the spacing problem was made by Giauque at the

University of California in introducing *punched brass spacers* [53].

A thin strip of cellulose acetate was placed between the spacer and the tube as the coil was being wound. After winding was complete, the acetate was dissolved by acetone. Several layers of tubes were used to provide many parallel paths for the low-pressure fluid, thereby reducing the overall pressure drop on the low-pressure side. The length of the path of the high-pressure fluid inside the tubes was made constant by adjusting the pitch of the tube helix from layer to layer. Heat exchangers of this type are commonly called Giauque – Hampson exchangers.

Collins Heat Exchanger. A novel extended-surface heat exchanger design developed by Collins [54] and marketed by the Joy Manufacturing Co., Michigan City, Ind., is shown in Figure 26. This was the heat exchanger used by Collins in his helium liquefaction system (see Section 3.7).

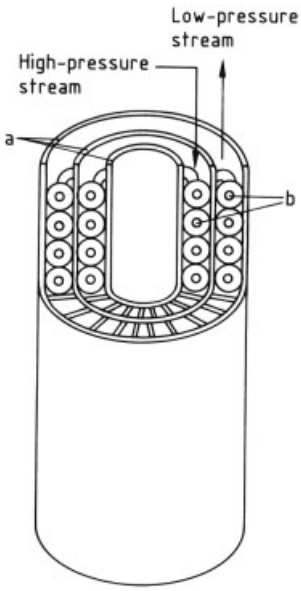

Figure 26. Collins heat exchanger
a) Concentric copper tubes; b) Edge-wound copper helix

The exchanger consists of several concentric copper tubes (a) with a close coiled helix of edge-wound copper ribbon (b) in the annular space between the tubes. The helix is soft-soldered to both sides of the annular space in which it is wound. The copper helix acts as a fin or extended surface to increase the heat transfer surface area of the annulus. These heat exchangers have been manufactured with as many as ten different flow passages.

Plate-Fin Exchangers. The plate-fin heat exchanger (an example is shown in Fig. 27) consists of stacks of alternating layers of corrugated die-formed metal sheets, i.e., fins (a), between thin metal sheets or plates (b). The edges of each layer are sealed with bars (c), and the flow passages formed within the corrugations are terminated with headers at each end of the stack. Crossflow, counterflow, or multipass arrangements can be constructed with the individual sandwiches. This type of heat exchanger has the advantage that a high surface area: volume ratio may be achieved (a compact arrangement). For example, as much as 1475 m^2 of heat transfer surface per cubic meter of exchanger volume may be attained with a plate-fin exchanger.

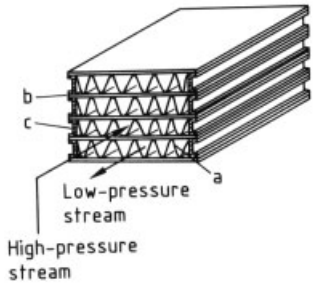

Figure 27. Plate-fin heat exchanger
a) Corrugated fin; b) Metal sheets; c) Bars

Perforated Plate Exchanger (Fig. 28) [55]. This exchanger consists of several parallel perforated plates (b) with spacers (a) between the plates. The plates are made of high-conductivity metal, such as copper or aluminum; the spacers are made of a low-conductivity plastic. The spacers are bonded to the plates to provide sealing between the high- and low-temperature streams and to minimize longitudinal conduction in the exchanger. Heat exchange between the two fluid streams occurs by lateral conduction through the copper or aluminum plates.

The diameter of the perforations can be made quite small so that a large heat transfer surface area per unit volume can be attained. The disruption of the flow between each plate results in high

heat transfer coefficients (and high friction factors); therefore, highly effective heat exchangers of this type can be constructed at relatively low cost.

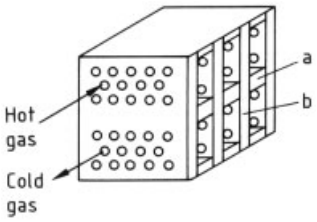

Figure 28. Perforated plate heat exchanger
a) Plastic spacer; b) Perforated metal plate

5.1.2. Heat Exchangers for Dilution Refrigerators

Heat exchangers for use in the dilution refrigerator system present special thermal problems because of the large thermal boundary resistance between helium and a solid surface at very low temperatures. This thermal resistance, called the *Kapitza resistance*, is inversely proportional to T^3, so the resistance can be quite large in the 0.01 K temperature range.

The *Frossati heat exchanger*, developed for a dilution refrigerator, is shown in Figure 29 [51].

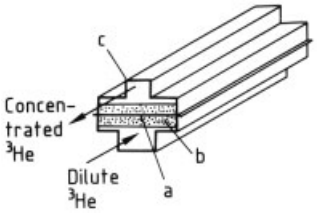

Figure 29. Frossati heat exchanger for a ^{3}He – ^{4}He dilution refrigerator
a) Separator; b) Sintered silver powder; c) Rectangular channel

The heat exchanger consists of flow passages separated by Cu – Ni foil (a), which is silver-plated and coated with a silver powder (particle diameter 1 μm). The coating is placed between two layers of sintered silver powder (b), which are covered by rectangular channels (c). The sintered metal coating on the separator provides a large surface area per unit volume, which reduces the Kapitza resistance. The liquid helium mixture flows through the rectangular channel, instead of through the sintered metal, so the pressure drop through the heat exchanger is not excessive.

The Frossati heat exchanger uses a sintered metal coating thickness of ca. 1 mm. The flow passages vary in size from 12.2 m^2 (dilute side) and 5.0 m^2 (concentrated side) at the cold end of the exchanger down to 4.0 m^2 (dilute side) and 1.6 m^2 (concentrated side) at the warm end. The heat exchanger is 1.040 m long and is wound in a spiral to conserve space. A low temperature of 0.002 K is achieved continuously with this refrigerator.

Sintered metal powder heat exchangers represent the present state of the art for dilution refrigerator heat exchangers and are used in commercial refrigerators that produce 0.005 K continuously.

5.1.3. Heat Exchanger Design Problems

Three problems may become acute in designing heat exchangers for cryogenic systems:

1) the effect of variable fluid properties, in particular variation of specific heat
2) the effect of longitudinal conduction
3) the effect of uneven distribution of flow

In many conventional heat exchanger design situations, the specific heats of the two fluids may be treated as essentially constant. For gases such as hydrogen in the cryogenic temperature range, the specific heat may vary significantly, and this variation can cause a *thermal pinch* effect within the heat exchanger.

Table 4 presents data on the temperature distribution in a gaseous helium – gaseous hydrogen counterflow heat exchanger operating between 50 and 250 K for the hydrogen stream and between 60 and 260 K for the helium stream. The mass flow rate of the hydrogen stream is 1 kg/s, and the mass flow rate of the helium stream is 2.425 kg/s. The specific heat of the hydrogen gas is 10.55 kJ $kg^{-1}K^{-1}$ at 50 K, and 14.04 kJ $kg^{-1}K^{-1}$ at 250 K, which corresponds to a 40 % increase in the specific heat through the exchanger. If the specific heats of the gases are constant, the log-mean-temperature-difference (LMTD) for the exchanger is 10 K However,

because of the specific heat variation, the temperature difference (ΔT) becomes as small as 3.1 K in the middle of the heat exchanger, and the actual LMTD is only 4.92 K. Thus, the actual heat transfer surface area that is required is about two times that calculated assuming constant specific heats.

Table 4. Temperature in a gaseous hydrogen – gaseous helium heat exchanger

Temperature of hydrogen (T_{H_2}), K	Temperature of helium (T_{He}), K	ΔT *
50	60.0	10.0
70	76.9	6.9
90	95.6	5.6
110	114.9	4.9
130	133.6	3.6
150	153.1	3.1
170	173.5	3.5
190	194.5	4.5
210	215.9	5.9
230	237.8	7.8
250	260.0	10.0

* $\Delta T = T_H^e - T_{H_2}$

The problem of "thermal pinch" may be alleviated by "unbalancing" the heat exchanger by increasing the hydrogen gas flow (if possible). Alternatively the helium gas flow can be decreased by bleeding off a portion of the flow at some intermediate point in the exchanger.

Longitudinal heat conduction along the walls of conventional heat exchangers is usually not a serious problem, because of the great length of the heat exchanger. Miniature cryogenic refrigerators can use heat exchangers that are as short as 50 mm. In addition, the effect of longitudinal conduction is most severe for a heat exchanger that is very highly effective.

One solution to the problem of longitudinal conduction is the use of thin-walled tubes of a relatively low-conductivity material (stainless steel). The tubes are coupled intermittently by a plating of a high thermal conductivity material (silver solder) between the tubes [56]. The high-conductivity material allows good heat transfer between the tubes, and longitudinal conduction is reduced by the thermal breaks between the sections of plating.

Many cryogenic heat exchangers have several short parallel flow passages for one or both fluid streams, instead of a single long flow passage for each stream. Ideally, the flow divides equally among the parallel flow passages; in practice, however, assurance of equal flow in each channel is not always possible because of differences in manufacturing tolerances. This results in a higher flow rate in the flow passages of smaller frictional resistance. The problems caused by uneven flow distribution are more severe for highly effective heat exchangers, such as the cryogenic heat exchangers.

The effect of uneven flow distribution in cryogenic heat exchangers has been analyzed [57]. For a heat exchanger with a design effectiveness of 0.995, uneven distribution of flow can reduce the effectiveness for the heat exchanger in operation to a value of only 0.95.

One technique for solving the problem of uneven flow distribution uses flow constrictions in the middle of the heat exchanger on the warm side [58]. If the temperature of the warm fluid increases because of a higher flow rate in the passage, the constriction presents a higher flow resistance, thereby reducing the warm fluid stream flow through that passage and balancing the flow.

5.1.4. Regenerators

In ordinary heat exchangers, the two fluids exchanging energy are separated by a solid surface (e.g., tube wall, plate). In a regenerator, the two fluids occupy the same space alternately, while the energy to be transferred is stored and released from the regenerator packing material or matrix. The temperature of the fluid within the regenerator and the temperature of the regenerator matrix are functions of both position within the regenerator and time of operation. As a result, the performance of the regenerator is a function of the mass of the matrix, the specific heat of the matrix material, and the frequency of switching of the flow through the regenerator.

The regenerator in a cryogenic system consists of a matrix material contained within a cylinder. For continuous operation, pairs of regenerators are usually used so that the matrix in one unit can be heated while the matrix in the other unit is being cooled. Various materials have been used for the matrix, including small metal spheres, metal wires or strips, punched copper disks, and layers of small-diameter, woven wires.

The main *advantages* of regenerators over ordinary heat exchangers (recuperators) are as follows:

1) A much larger heat transfer surface area per unit volume can be achieved for the regenerator. As much as 4000 m^2 of surface area per cubic meter of volume can be obtained in a regenerator, compared with 1475 m^2/m^3 for a compact plate-fin heat exchanger.
2) The regenerator tends to be self-cleaning because of the periodic flow reversal. Impurity buildup problems are not nearly as severe in a regenerator as those in ordinary heat exchangers.
3) Regenerators are generally simpler to fabricate, so the regenerator tends to cost less than a heat exchanger for the same heat transfer duty.

The main *disadvantage* of the regenerator is that some mixing of the warm and cold streams occurs during the switching process. In addition, some means must be provided for switching the flow from one regenerator to the other in the regenerator pair.

5.2. Compressors

The compressor is an important component of all liquefaction and refrigeration systems because (1) it represents a significant fraction of the total initial capital cost of the unit, (2) its power input accounts for a large portion of the system operating costs, and (3) it is usually the component that determines the reliability of the system. Compressors used in cryogenic systems are classed as either reciprocating compressors or rotary compressors.

Reciprocating Compressors. Reciprocating compressors are generally more massive and operate at slower speed (up to 3000 rpm) than rotary compressors. The reciprocating units are used in systems requiring high pressure (4 – 20 MPa) and relatively low flow rate (<300 t/d). They are easily adapted to situations in which the load changes suddenly.

The compression process for reciprocating compressors is polytropic because the gas cools in the cylinder during compression. Two stages of compression with intercooling and aftercooling can be used for pressure ratios ranging from up to 8 to 10. For higher pressure ratios, three or four stages can be used to limit the temperature of the gas at the end of the compression process.

Compressors used in cryogenic systems generally operate at ambient temperature; thus, they are similar in design to compressors used for any gas compression service, with one important modification. Cryogenic compressors must be designed to reduce leakage around seals and to reduce contamination of the gas being compressed. One of the primary contaminants is the oil used to lubricate seals in the compressor piston. In compressors lubricated with oil, some provision must be made to separate the entrained oil from the gas. Steel wool or glass wool filters are used to remove liquid oil droplets, and refrigerated adsorption units are used to remove oil vapor.

Some compressors are operated without oil lubrication to avoid the problem of contamination of the gas being compressed. Older units used carbon piston rings, sometimes filled with a metal powder reinforcement. However, when the dew point of the water vapor in the gas is less than −30 °C, the carbon piston ring is unsuitable because of changes in the surface chemistry of the carbon. Without the presence of a little moisture, the carbon behaves as an abrasive material.

In current dry-lubricated compressors, polytetrafluoroethylene (PTFE) piston rings are used. To improve the wear characteristics and to allow better removal of the frictional energy dissipation, the PTFE rings are usually reinforced with a hard filler of a metal powder, such as brass or copper. Glass-filled PTFE rings (known as Rulon-A) have been used with good results.

The following factors contribute to the compressor efficiency being less than 100 %:

1) *Nonisothermal Compression.* Because of the relatively small surface area available for heat transfer in the compressor piston, the compression process is polytropic, with a polytropic exponent of ca. 1.3 – 1.4 for compression of diatomic gases.
2) *Valve Losses.* Pressure drops occurring across the inlet and outlet valves reduce the actual pressure ratio of the gas within the compressor cylinder.
3) *Piston Friction.* Frictional energy dissipation in the piston seals, bearings, etc. increases the required power input to the compressor.

Rotary Compressors. Rotary compressors are generally smaller and operate at higher speed

than reciprocating units. The rotary units are more sensitive to sudden changes in load; they are either axial flow, centrifugal flow, or rotary screw compressors.

Axial flow compressors have several stages of compression because the practical pressure ratio for each stage is relatively small. Intercooling is required, along with aftercooling, to limit the temperature of the gas being compressed. Because of the small surface area within the compressor flow passages, the compression process is practically adiabatic, but not reversible.

Centrifugal flow compressors involve a simple construction with a single rotating disc or impeller for each stage of compression. The limiting pressure ratio across each stage is 4 : 1, so multistage units must be used to achieve pressures >400 kPa. Centrifugal flow compressors also require intercooling and aftercooling to limit the temperature of the gas being compressed. In air liquefaction plants, multistage centrifugal compressors have been used to compress air to pressures of 4.0 – 5.5 MPa at flow rates <500 t/d.

Rotary screw compressors are finding increased application in cryogenic systems. The rotary unit consists of two or more cylindrical rotors with multiple helical lobes that mesh as the rotors turn. A volume of gas is trapped between the lobes and is compressed as the volume of the space decreases. The rotors are lubricated with oil sprayed onto the rotors at the low-pressure end of the compressor.

Some of the factors that contribute to the efficiency of the rotary compressor being less than 100 % include

1) *Nonisothermal Compression.* The compression process is almost adiabatic, rather than isothermal. More work is required to compress a gas adiabatically than for isothermal compression.
2) *Disk Friction.* For centrifugal flow compressors, frictional energy is dissipated between the rotor and the gas in the space between the rotor and the housing.
3) *Impeller Losses.* For both axial flow and centrifugal flow compressors, frictional energy dissipation and flow separation occur within the blade flow passages and reduce compressor efficiency.
4) *Exit Losses.* The kinetic energy of the stream leaving the last stage is dissipated and results in additional energy loss.

5.3. Expanders

The expanders used by Claude (see Section 3.5) were basically steam engines modified to operate with air instead of steam. Initially, Claude attempted to liquefy the gas within the engine and to use the liquid air as a lubricant [59]. This proved unsuccessful because of the low viscosity of liquid air and excessive heat transfer; thus, he finally used a leather cup as the piston seal.

In 1912 Heylandt improved the efficiency of expanders by using the crowned piston concept (Fig. 30) [16, p. 325].

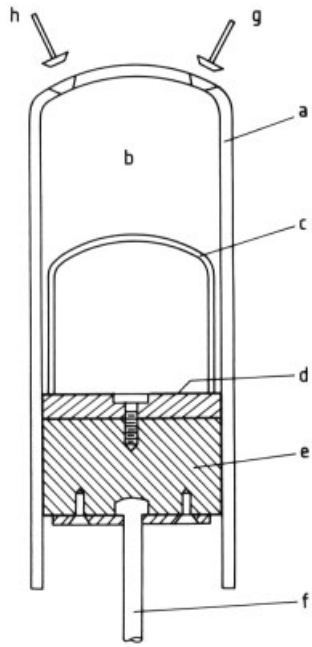

Figure 30. Piston used in the expander in the Heylandt liquefaction system
a) Cylinder; b) Expansion space; c) Low-conductivity cap; d) Packing ring; e) Sealing piston; f) Piston rod; g) Intake valve; h) Exhaust valve

The piston (e) is fitted with a long cap (c) of a low-conductivity material; the cylinder (a) is extended to accommodate the cap. The clearance between the cap and the cylinder is rela-

tively large so that little frictional energy is dissipated in this region. The expanded gas is in contact with the low-conductivity cap, and sealing is achieved in the piston near the ambient-temperature end (d).

Reciprocating expanders with oil-lubricated cylinders and exit temperatures as low as 90 K are used in air liquefaction plants. The piston rings are usually made of an aluminum – bronze alloy, and the cylinder has a high-strength, replaceable, cast iron liner. To reduce heat transfer down the piston, the stroke-to-diameter ratio is ca. 2 : 1, and the piston is constructed of stainless steel because of this material's low thermal conductivity. Hollow piston rods are used to further reduce conduction heat transfer. Expanders with oil-lubricated cylinders are equipped with oil filters in the discharge line to remove any oil particles.

For exit gas temperatures below ca. 90 K or for systems in which the presence of oil cannot be tolerated, nonlubricated PTFE pistons, similar to those in nonlubricated compressors, are used. The cylinder also contains a replaceable liner.

The operating characteristics of a reciprocating expander engine in an air separation plant have been studied [60]. The following components of the energy dissipation were recorded (adiabatic efficiency, η_{ad}= 82 %):

Nonideal expansion	6 %
Valve losses	4 %
Clearance losses	4 %
Incomplete expansion	2 %
Heat inleak	2 %
Total losses	18 %

Although several patents were obtained for rotary expanders from 1898 to 1922, the first successful commercial application of an expansion turbine was made by Linde in the 1930 s. His turbine was an axial flow, single-stage impulse turbine; inward radial flow units were constructed later.

A significant contribution to cryogenic expander technolgy was the development of high-speed, gas-lubricated bearings [61,62], which enabled the successful construction of miniature refrigerators using rotor diameters as small as 8 mm operating at a speed up to 500 000 rpm. The following loss components have been reported for a radial flow expansion turbine (adiabatic efficiency, η_{ad}= 84.0 %) [60]:

Inlet losses	5.5 %
Blade (wheel) losses	3.0 %
Leakage	3.0 %
Disk friction	2.5 %
Exit loss	1.0 %
Heat transfer	1.0 %
Total losses	16.0 %

The compressor efficiency does not influence the liquid yield in a liquefier or the refrigeration effect in a cryocooler; however, the expander efficiency has a significant effect on both quantities. For example, for a Claude liquefier operating between 101.3 kPa at 300 K and 6.08 MPa, with nitrogen as the working fluid, the enthalpy values are as follows [52]:

h_1= 462.15 kJ/kg at 300 K and 101.3 kPa
h_2= 450.52 kJ/kg at 300 K and 6.08 MPa
h_3= 392.21 kJ/kg at 250 K and 6.08 MPa
h'_{e} = 225.77 kJ/kg at 101.3 kPa and s_3
h_{f}= 29.40 kJ/kg at 101.3 kPa and saturated liquid.

For an expander mass flow rate ratio of x = 0.60, the liquid yield for a system with a 100 % efficient expander is given by Equation (9):

$$y = \frac{(462.15 - 450.52) + (0.60)(392.21 - 225.77)}{(462.15 - 29.40)} = 0.2576$$

For the same conditions but with an expander having an adiabatic efficiency of η_{ad}= 80 %, the liquid yield is

$$y = \frac{11.63 + (0.80)(0.60)(166.44)}{432.75} = 0.2115$$

The effect of the expander efficiency on the work per unit mass liquefied is even more dramatic. For an ideal compressor and ideal expander, the work requirement per unit mass compressed is found from Equation (10):

$$-W/q_{\mathrm{m}} = [(300)(4.421 - 3.169) - 11.63] - (0.60)(166.44) = 264.11\,\mathrm{kJ/kg}$$

The work per unit mass liquefied is

$$-W/(q_{\mathrm{m}})_{\mathrm{f}} = 264.11/0.2576 = 1025.3\,\mathrm{kJ/kg}$$

On the other hand, for a Claude liquefier with an ideal compressor and an expander with an overall efficiency of η_{e}= 72 %, the work per unit mass compressed is

$$-W/q_{\mathrm{m}} = 363.97 - (0.72)(0.60)(166.44) = 292.07\,\mathrm{kJ/kg}$$

The work per unit mass liquefied is

$$-W/(q_{\mathrm{m}})_{\mathrm{f}} = 292.07/0.2115 = 1380.9\,\mathrm{kJ/kg}$$

The expander inefficiencies increase the work required to produce 1 kg of liquid by ca. 35 %. If a compressor efficiency of 70 % is included in the calculations, the work per unit mass liquefied increases to 2118 kJ/kg.

6. Low-Temperature Separation Systems

The principles of separating gas and liquid mixtures at cryogenic temperatures are the same as those used in the design of separation systems for near-ambient temperatures and above; however, some special problems arise in low-temperature separation systems. For example, a cooling source for condensers in a cryogenic system is not readily available; whereas in petroleum separation systems, cooling water or ambient air may be used.

For several mixtures, such as air and coke-oven gas, cryogenic separation techniques are the most economical methods of separating the mixture components.

Cryogenic separation methods include rectification, partial condensation or evaporation, and adsorption.

6.1. Minimum Work of Separation

The minimum work of separation occurs when the separation process is carried out reversibly and isothermally at temperature T_0 of the surroundings. The work per mole of mixture is given by Equation (19):

$$-W_{\mathrm{id}}/(q_{\mathrm{v}})_{\mathrm{m}} = \sum y_{\mathrm{i}}\left[T_0(S_1-S_2)-(H_1-H_2)\right]_{\mathrm{i}} \tag{19}$$

where y_{i}= mole fraction of component i; the units for entropy (S) and enthalpy (H) are J mol^{-1} K^{-1} and J/mol, respectively. Subscript 1 denotes conditions before separation, and subscript 2 denotes conditions after separation. For conditions under which the materials may be treated as ideal gases, the entropy and enthalpy changes may be written as follows:

$$S_1-S_2 = c_p \ln(T_1/T_2) - R\ln(p_1/p_2)$$
$$S_1-S_2 = R\ln(p_{\mathrm{m}}/p_{\mathrm{i}})$$
$$H_1-H_2 = c_p(T_1-T_2) = 0$$

For isothermal separation, $T_1 = T_2 = T_0$. The initial pressure p_1 is equal to the partial pressure p_{i} of the component in the mixture, and the final pressure p_2 is equal to the mixture pressure p_{m}. The quantity R is the gas constant. Making these substitutions into Equation (19), the following expression is obtained for the minimum work of separation:

$$-W_{\mathrm{id}}/(q_{\mathrm{v}})_{\mathrm{m}} = RT_0 \sum y_{\mathrm{i}} \ln(1/y_{\mathrm{i}})$$

where $y_{\mathrm{i}} = p_{\mathrm{i}}/p_{\mathrm{m}}$ for an ideal gas.

Separation of Air. The ideal work of separation for air (a mixture of 78.1 mol % nitrogen, 21.0 mol % oxygen, and 0.9 mol % argon) at a temperature of 300 K is

$$-W_{\mathrm{id}}/(q_{\mathrm{v}})_{\mathrm{m}} = (8.3143)(300)\cdot$$
$$[0.781\ln(1/0.781) + 0.210\ln(1/0.210) + 0.009\ln(1/0.009)]$$
$$= 1405\,\mathrm{J/mol}$$

The ideal work per mole of oxygen is

$$-W_{\mathrm{id}}/q_{\mathrm{v}}(\mathrm{O_2}) = 1405/0.210 = 6690\,\mathrm{J/mol}$$

Similarly, the ideal work per unit mass of O_2 is

$$-W_{\mathrm{id}}/q_{\mathrm{m}}(\mathrm{O_2}) = 6690/32 = 209.1\,\mathrm{J/g}$$

If the oxygen is separated from the mixture and the nitrogen and argon are not separated, the ideal work per mole mixture is

$$W_{\mathrm{id}}/q_v = (8.3143)(300)\cdot$$
$$[0.790\ln(1/0.790) + 0.210\ln(1/0.210)]$$
$$= 1282\,\mathrm{J/mol}$$

The ideal work per mole and per unit mass of O_2 separated are

$$-W_{\mathrm{id}}/q_{\mathrm{v}}(\mathrm{O_2}) = 6105\,\mathrm{J/mol}$$
$$-W_{id}/q_{\mathrm{m}}(\mathrm{O_2}) = 190.8\,\mathrm{J/g}$$

Thus, the ideal work is smaller if only one of the components is separated from the mixture.

6.2. Properties of Mixtures

The number of independent thermostatic properties required to fix the state of a mixture is given by the *Gibbs phase rule:*

$$F = C - P + 2$$

where C = number of components and P = number of phases present. For a ternary (three-component) mixture in the vapor phase, four independent properties are required to determine the state of the mixture. For example, if the mixture pressure, mixture temperature, and mole fractions of two of the components are specified, all other thermostatic properties can then be determined.

Properties of mixtures may be presented in various forms. One of the common representations for binary (two-component) mixtures is the temperature – composition diagram, as shown for the nitrogen – oxygen system (Fig. 31).

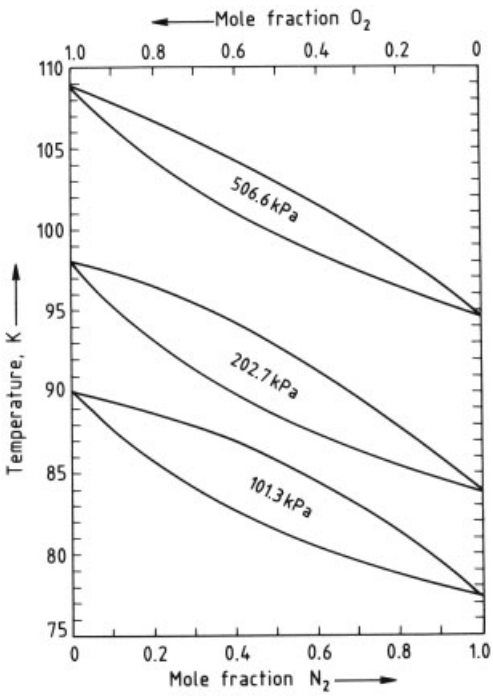

Figure 31. Temperature – composition diagram for nitrogen – oxygen mixtures

In this figure, mole fraction is plotted against mixture temperature for a fixed value of mixture pressure. The upper line in each pair of curves is called the *dew line* because this curve represents the temperature, for a given composition of the mixture, at which the vapor mixture begins to condense. The dew line corresponds to the condition of the saturated vapor in the two-phase mixture.

The lower line in each pair of curves is called the *bubble line* because this curve represents the temperature, for a given composition of the mixture, at which the liquid mixture begins to boil. The bubble line corresponds to the condition of the saturated liquid in the two-phase mixture.

The mixture properties may also be plotted as shown in Figure 32. This equilibrium diagram is important in the design of rectification columns to determine the number of "theoretical plates" required.

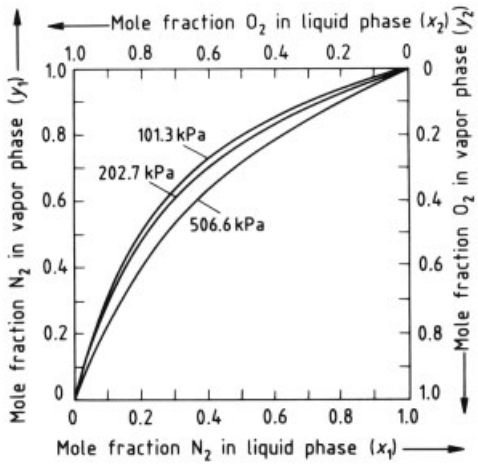

Figure 32. Equilibrium diagram for nitrogen – oxygen mixtures

The relationship between the mole fractions y and x in the vapor and liquid phases, respectively, may be expressed in terms of the distribution coefficient K (sometimes called the equilibrium ratio), defined for component j by [63]:

$$K_j = y_j / x_j$$

For an ideal-gas vapor phase and a perfect-solution liquid phase, the distribution coefficient is given by

$$K_j = \pi_j / p$$

where π_j = vapor pressure of component j at the mixture temperature and p = mixture pressure. For real gases and solutions, the distribution coefficients must be determined either experimentally or from thermodynamic relationships along with the equations of state for the materials. A plot of the distribution coefficients for nitrogen ($K_1 = x_1/y_1$) and for oxygen (K_2) is shown in Figure 33.

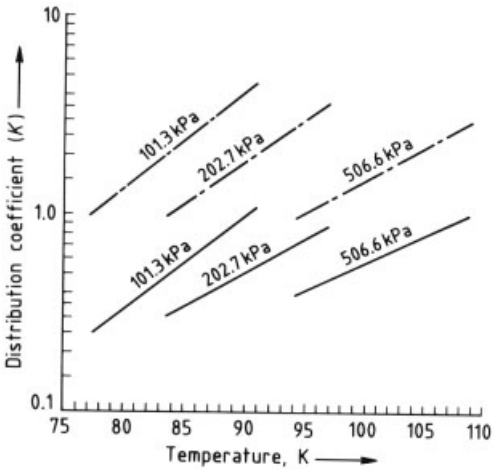

Figure 33. Distribution coefficients for nitrogen (– – –) and oxygen (–)

The behavior of mixtures can also be represented in terms of the relative volatility, α_{ij}, defined by

$$\alpha_{ij} = K_i / K_j$$

As shown in Figure 34, the relative volatility for cryogenic mixtures, such as nitrogen – oxygen, is a fairly weak function of the mixture temperature for a fixed value of the mixture pressure.

The composition of the vapor and liquid phases of binary mixtures may be determined from values of the distribution coefficients or the relative volatility. From the definition of the distribution coefficient,

$$y_1 = K_1 x_1$$
$$y_2 = K_2 x_2 = K_2 (1 - x_1)$$

For a binary mixture,

$$y_1 + y_2 = 1 = K_1 x_1 + K_2 (1 - x_1)$$

where subscript 1 refers to the lower boiling point component and subscript 2 refers to the

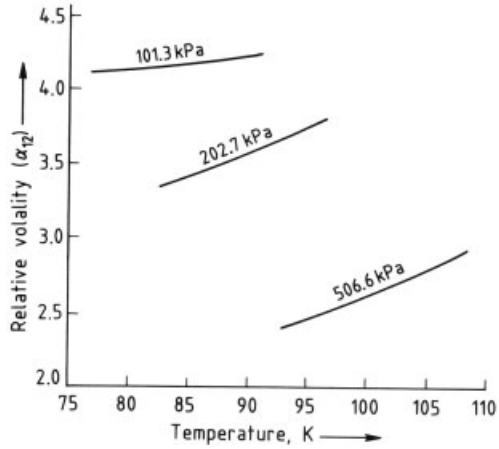

Figure 34. Relative volatility for nitrogen – oxygen mixtures

higher boiling point component in the mixture. Solving for the mole fraction of component 1 in the liquid phase, Equation (20) is obtained:

$$x_1 = \frac{1 - K_2}{K_1 - K_2} \tag{20}$$

The mole fraction of component 1 in the vapor phase may be determined from

$$y_1 = K_1 x_1 = \frac{K_1 (1 - K_2)}{K_1 - K_2} \tag{21}$$

Equations (20) and (21) can also be written in terms of the relative volatility α_{12}, as follows:

$$x_1 = \frac{1 - K_2}{K_2 (\alpha_{12} - 1)}$$
$$y_1 = \frac{\alpha_{12} (1 - K_2)}{\alpha_{12} - 1}$$

Another set of thermodynamic curves that are import in the design of separation systems is the enthalpy – composition diagram. Values for the enthalpy at various mixture pressures for the N_2–O_2 system are given in Table 5.

6.3. Flash Calculation

In the rectification columns used in cryogenic separation systems, the feed stream is introduced into the column by expansion through a J – T valve. Determining the condition of the feed stream after expansion through the valve is important. This solution for a two-phase mixture is called the flash calculation.

Table 5. Specific enthalpies of nitrogen – oxygen mixtures

Mole fraction	Specific enthalpies ∗, J/mol					
nitrogen	101.3 kPa		202.7 kPa		506.6 kPa	
	h_b	h_d	h_b	h_d	h_b	h_d
0.0	419	7252	764	7362	1315	7536
0.1	461	7231	807	7329	1403	7507
0.2	519	7210	860	7295	1499	7478
0.3	599	7185	923	7259	1591	7448
0.4	682	7151	996	7222	1675	7415
0.5	779	7118	1079	7184	1758	7377
0.6	879	7084	1172	7145	1851	7339
0.7	984	7042	1275	7105	1947	7298
0.8	1084	6992	1388	7063	2052	7252
0.9	1181	6933	1511	7020	2152	7201
1.0	1273	6871	1644	6976	2248	7147

$*h_b$ = specific enthalpy at the bubble point; h_d = specific enthalpy at the dew point for the mixture.

If the feed stream enters the J – T valve at a molar flow rate q_v with known properties (enthalpy h_{in} and mole fraction $x_{in,j}$), a mass balance across the J – T valve yields

$$x_{in,j} q_v = y_j q_v\,(g) + x_j q_v\,(l) = y_j\,(q_v - q_v\,(l)) + x_j q_v\,(l)$$

where $q_v(g)$ = molar flow rate of vapor from the J – T valve and $q_v(l)$ = molar flow rate of liquid from the J – T valve. Introducing the distribution coefficient, $x_j = y_j/K_j$, the mole fraction of the lower boiling point component in the vapor phase leaving the J – T valve may be determined:

$$y_j = \frac{x_{in,j}}{1 + \frac{q_v(l)}{q_v}\left(\frac{1}{K_j} - 1\right)}$$

An energy balance around the J – T valve yields the auxiliary relationship for the fraction of the outlet stream, which is in the liquid phase:

$$\frac{q_v\,(l)}{q_v} = \frac{L_g - L_{in}}{L_g - L_f} \tag{22}$$

The final governing equation for the flash calculation results from the fact that the sum of the mole fractions for the vapor phase is equal to unity:

$$\sum_j y_j = 1 = \sum_j \frac{x_{in,j}}{1 + \frac{q_v(l)}{q_v}\left(\frac{1}{K_j} - 1\right)} \tag{23}$$

The solution for the condition of the liquid and vapor phases after expansion through the J – T valve involves an *iterative technique.* First, a trial temperature for the mixture, which is between the dew point and bubble point temperatures at the exit pressure of the J – T valve, is chosen. The liquid fraction can be calculated from Equation (22) using this temperature, and the right side of Equation (23) can then be determined. If the right side is not equal to unity, as it should be, a new trial temperature is chosen; the process is repeated until Equation (23) is satisfied. When the correct temperature is found for the exit stream, then the composition of the stream may be determined from the distribution coefficients and from Equations (20) and (21). Computer-aided flash calculations have been described for multicomponent mixtures [64].

6.4. Principles of Cryogenic Rectification

The rectification process involves cascading several partial evaporations and condensations that are carried out in counterflow. A schematic of a rectification column is shown in Figure 35 (see next page).

The feed stream enters the column at a molar flow rate $(q_v)_f$, enthalpy h_{in}, and a mole fraction of the lower boiling point component x_{in}. The top product (the practically pure lower boiling point component, e.g., N_2 in the air separation column) leaves the column at a molar flow rate $(q_v)_D$, enthalpy h_D, and composition x_D. Similarly, the bottom product (the higher boiling point component, e.g., O_2 in the air separation column) leaves the column at a molar flow rate $(q_v)_B$, enthalpy h_B, and composition x_B. Heat Q_B is added to the column in the boiler (e), and heat Q_D is withdrawn from the column in the condenser (b).

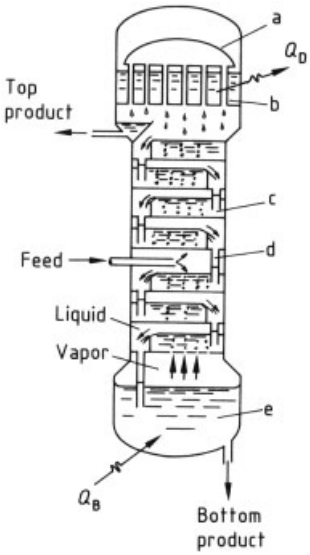

Figure 35. Schematic of a rectification column
a) Dome; b) Condenser; c) Liquid-on-plate; d) Downcomer (seal); e) Boiler

The vapor in the feed stream flows up the column through the layers of liquid (c) on plates or trays. As the vapor flows through the liquid, the vapor bubbles (which are at a higher temperature than the liquid) evaporate some of the lower boiling component from the liquid into the vapor bubble. The heat transfer from the bubble results in some condensation of the higher boiling component from the vapor bubble into the liquid phase. In this manner, the vapor phase becomes more enriched in the lower boiling component as the vapor flows up the column to the condenser. Similarly, the liquid phase in the feed stream flows down the column, and the liquid is enriched in the higher boiling component.

The composition of the top and the bottom product streams depends on the number of stages or plates in the upper and lower portions of the column. The number of *theoretical plates* for the column can be determined through one of two methods, the Ponchon – Savarit method [65] or the McCabe – Thiele method [66]. The Ponchon – Savarit method is an exact method, which can be applied for all materials; however, detailed enthalpy data are required. The McCabe – Thiele method is somewhat less accurate, in general, but it requires only equilibrium concentration data and is generally adequate for cryogenic systems.

The three governing relationships for the analysis of rectification columns are the conservation of mass, the conservation of mass for one component, and the conservation of energy. Applying these principles for the upper portion of the column gives

$$(q_v)_n\,(g) = (q_v)_{n+1}\,(l) + (q_v)_D \tag{24}$$

$$(y_n)\cdot(q_v)_n\,(g) = x_{n+1}\cdot(q_v)_{n+1}\,(l) + x_D\cdot(q_v)_D \tag{25}$$

$$h_n\,(g)\cdot(q_v)_n\,(g) = h_{n+1}\,(l)\cdot(q_v)_{n+1}\,(l) + h_D\cdot(q_v)_D + Q_D \tag{26}$$

where $(q_v)_n$ (g) = molar flow rate for the vapor leaving plate n, $(q_v)_{n+1}$ (l) = molar flow rate for the liquid leaving plate $(n+1)$ and entering plate n, y_n = mole fraction of the lower boiling point component leaving the nth plate in the vapor phase, x_{n+1} = mole fraction of the lower boiling point component leaving the $(n+1)$-plate in the liquid phase, h_n (g) = enthalpy of the vapor leaving plate n, and h_{n+1}(l) = enthalpy of the liquid leaving plate $n+1$. The plates are numbered from the feed plate ($n=1$) upward. The expression for the reflux ratio of the column may be determined by combining Equations (24) – (26):

$$\frac{(q_v)_{n+1}\,(l)}{(q_v)_n\,(g)} = 1 - \frac{h_n\,(g) - h_{n+1}\,(l)}{\frac{Q_D}{(q_v)_D} + h_D - h_{n+1}\,(l)} \tag{27}$$

$$\frac{(q_v)_D}{(q_v)_n\,(g)} = 1 - \frac{(q_v)_{n+1}\,(l)}{(q_v)_n\,(g)}$$

For nitrogen – oxygen mixtures and several other simple mixtures, the molar heat of vaporization is almost independent of the composition, the reflux ratio is practically constant, and the McCabe – Thiele method may be applied. The equation for the "operating line" is found by combining Equation 27 and the conservation of mass expression for the lower boiling component:

$$y_n = \left[\frac{(q_v)_{n+1}\,(l)}{(q_v)_n\,(g)}\right] x_{n+1} + \left[\frac{(q_v)_D}{(q_v)_n\,(g)}\right] x_D \tag{28}$$

The solution for the number of theoretical plates or "equilibrium stages" required for the upper portion of the column is found by successive application of Equation 28 and the equilibrium relationship $y_n = K_j x_{n+1}$, between the feed condition and the desired condition at the top of the column. Similar expressions can be written for the lower portion of the column:

$$\frac{(q_v)_B}{(q_v)_m\,(g)} = \frac{h_m\,(g) - h_{m+1}\,(l)}{\frac{Q_B}{(q_v)_B} - h_B + h_{m+1}\,(l)}$$

$$\frac{(q_v)_{m+1}\,(l)}{(q_v)_m\,(g)} = 1 + \frac{(q_v)_B}{(q_v)_m\,(g)}$$

The operating line for the lower section of the column is

$$y_m = \left[\frac{(q_v)_{m+1}\,(l)}{(q_v)_m\,(g)}\right] x_{m+1} - \left[\frac{(q_v)_B}{(q_v)_m\,(g)}\right] x_B$$

The plates in the lower section are numbered from the boiler section ($m = 1$) upward to the feed plate.

The solution for the number of theoretical plates can be determined graphically or by a direct numerical approach, using a computer.

6.5. Linde Single-Column System

One of the simplest systems for separation of air is the Linde single-column system, shown in Figure 36. The system is used to produce either liquid oxygen or liquid nitrogen, but not both.

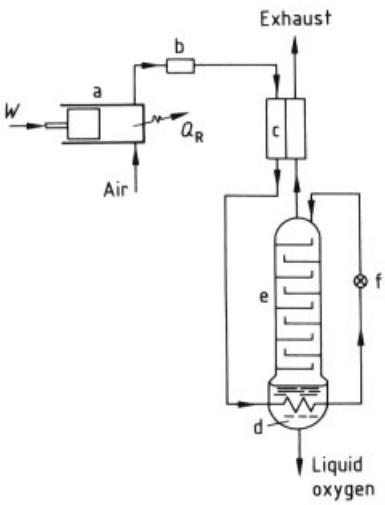

Figure 36. Linde single-column air-separation system
a) Compressor; b) Carbon dioxide and water removal; c) Heat exchanger; d) Boiler; e) Separation column; f) J – T valve

The air feed stream is compressed (a), water vapor and carbon dioxide are removed (b), and the air is passed through a heat exchanger (c). The cool air is further cooled in a heat exchanger (boiler, d) in the lower portion of the column (e) and then expanded through a J – T valve (f). The heat exchange in the boiler produces the upgoing vapor required in the column without using an external heat source.

The liquid formed during expansion through the J – T valve is introduced at the top of the column to provide the required liquid flow down the column. The vapor at the top of the column is passed back through the heat exchanger to provide cooling for the incoming feed stream. By use of the required number of plates in the column, nearly pure liquid or gaseous oxygen may be produced.

If the product is required as a gas, the oxygen gas is usually passed back through the heat exchanger (in this case, three-channel exchanger) to provide additional cooling for the incoming feed stream. Furthermore, the feed stream must be compressed to ca. 3 – 6 MPa, because of the additional cooling provided by the return cold oxygen stream through the heat exchanger. On the other hand, if the product is desired as a liquid, then the feed stream must be compressed to ca. 20 MPa for effective operation. The pressure within the column is usually 0.13 – 0.20 MPa.

6.6. Linde Double-Column System

If both liquid nitrogen and liquid oxygen products are required, the double-column system (Fig. 37) must be used.

The feed stream is compressed (a) and water and carbon dioxide are removed (b), as in the single-column system. The feed stream passes through a heat exchanger (c) and then through the boiler (d) of the lower column (e_1), where the feed stream is further cooled. Vapor is provided for the lower column by evaporating some of the liquid in the boiler. The feed stream is expanded through a J – T valve (f_1) and introduced in the middle of the lower column, which operates at a pressure of ca. 0.5 – 0.6 MPa [67].

A portion of the liquid in the boiler of the lower column is withdrawn, expanded through a J – T valve (f_2), and introduced into the middle of the upper column (e_2), which operates at a pressure of ca. 0.13 – 0.14 MPa. The composition of the liquid in the boiler is ca. 36 – 39 %

oxygen. At a pressure of 0.51 MPa, the boiling point of pure liquid nitrogen is 94.2 K, and the boiling point of pure liquid oxygen at 0.13 MPa is 92.7 K; therefore, the liquid oxygen in the upper column can be used to condense (g) the liquid nitrogen in the lower column without requiring external refrigeration.

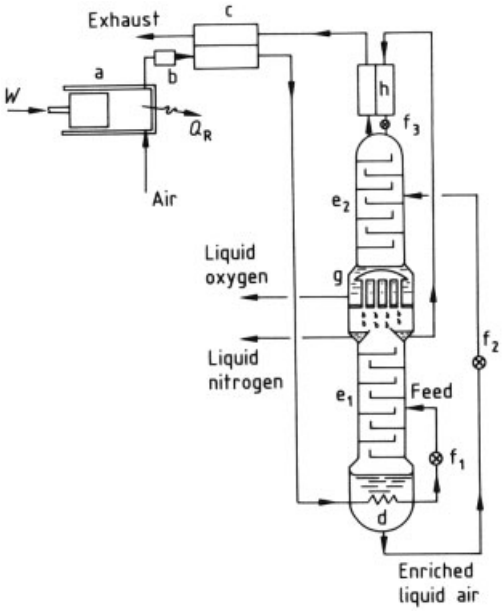

Figure 37. Linde double-column air-separation system
a) Compressor; b) Carbon dioxide and water removal; c) Heat exchanger; d) Boiler; e) Separation column; f) J – T valve; g) Condenser – reboiler; h) Subcooler

A portion of the liquid nitrogen from the lower column is passed through a subcooler (h), expanded through a J – T valve (f_3), and introduced at the top of the upper column.

6.7. Heylandt System

One of the commonly used air separation systems is the Heylandt system (Fig. 38). This system uses an expander to provide a significant portion of the refrigeration needed to liquefy the air.

The air feed stream is compressed (a) to a pressure of ca. 10 MPa. Water vapor is removed in the compressor intercoolers and the aftercooler. The air stream is then compressed further to ca. 14 MPa in a final stage of compression, which is driven by the expander.

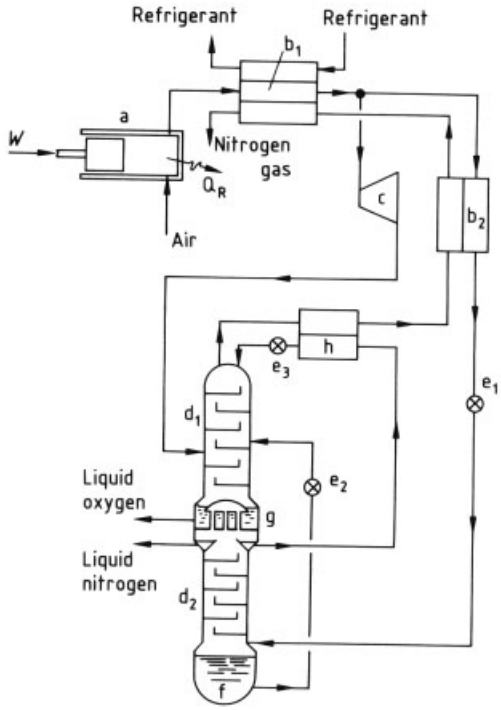

Figure 38. Heylandt air-separation system
a) Compressor; b) Heat exchanger; c) Expander; d) Separation column, e) J – T valve; f) Kettle; g) Condenser – reboiler; h) Subcooler

The air is cooled to ca. 230 K in an ammonia forecooler (precooler, b_1). About one-half of the feed stream is passed through the main heat exchanger (b_2) and the remainder of the feed stream is expanded through an expansion turbine (c) to a pressure of 0.2 MPa. The stream leaves the expander at ca. 110 K and is introduced into the middle portion of the upper column (d_1). The portion of the feed stream flowing through the main heat exchanger is cooled and expanded through a J – T valve (e_1) to a pressure of 0.7 MPa. The air is then introduced into the bottom of the lower column (d_2).

Enriched liquid air (oxygen content of 36 – 38 %) is withdrawn from the kettle (f) of the lower column, expanded through a J – T valve (e_2) to 0.2 MPa, and introduced into the middle of the upper column. To provide reflux liquid for the upper column, liquid nitrogen is withdrawn from the lower column below the condenser – reboiler (g), passed through a subcooler (h), and expanded through a J – T valve (e_3) before being introduced at the top of the upper col-

umn. Nitrogen gas from the upper column is returned through the subcooler (h), the main heat exchanger (b_2), and the precooler (b_1) to provide the necessary cooling in these units.

The product liquids are removed from the condenser – reboiler (oxygen from the upper column and nitrogen from the lower column).

6.8. Argon Separation

Atmospheric air is not simply a binary mixture of nitrogen and oxygen. Rather, the composition of dry, carbon dioxide free atmospheric air is as follows (mole fraction):

Nitrogen	0.78110
Oxygen	0.20953
Argon	0.00934
Other gases	0.00003
Neon	18.19 mg/kg
Helium	5.24 mg/kg
Krypton – xenon	1.23 mg/kg
Hydrogen	0.50 mg/kg

The predominant component, other than oxygen and nitrogen, is argon, which is present at a concentration of nearly 1 mol % [68]. This argon concentration in the air feed stream means that to obtain pure oxygen and pure nitrogen an additional column must be used to separate the argon from the other components.

Argon boils at 87.3 K and 101.3 kPa; thus, its boiling point at atmospheric pressure lies between that of nitrogen (77.4 K) and oxygen (90.2 K). If all of the argon were collected with the nitrogen product, the composition would be 98.8 mol % nitrogen and 1.2 mol % argon. On the other hand, if all of the argon were collected with the oxygen stream, the composition would be 95.7 mol % oxygen and 4.3 mol % argon. Because argon is soluble in both liquid nitrogen and liquid oxygen, the argon is generally present in both streams unless it is removed during the separation process. In addition, the equilibrium composition of the ternary mixture is significantly different from that of the binary nitrogen – oxygen mixture. This fact affects the calculation of the number of theoretical plates required for the column.

Argon may be separated in the double-column system by using a separate argon column, as shown in Figure 39.

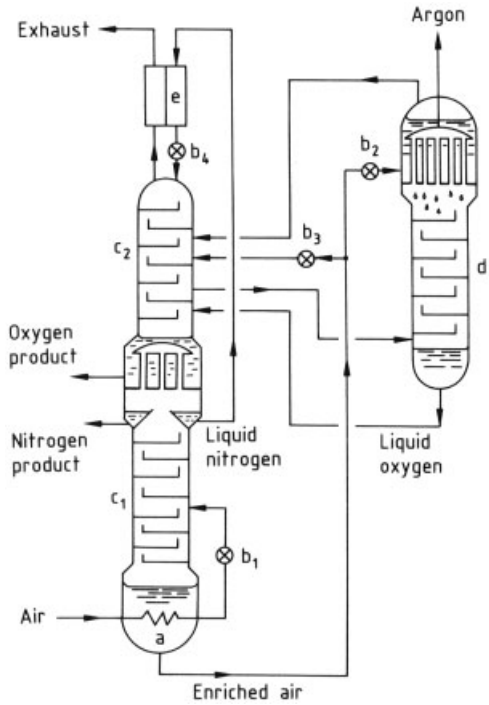

Figure 39. Argon recovery subsystem attached to the double-column air separation column
a) Kettle; b) J – T valve; c) Separation column; d) Argon column; e) Subcooler

Argon tends to collect in the part of the upper column (c_2) where the fluid has a high oxygen concentration and a relatively low nitrogen concentration. The mixture is removed from the main air-separation column at that point and is introduced into the lower portion of the argon side column (d). The vapor flows up the column and becomes enriched in argon, which has a lower boiling point than oxygen. At the top of the argon column, refrigeration is supplied by bypassing some of the liquid from the kettle (a) of the main column through the condenser of the argon column. The vapor from the condenser is returned to the upper portion of the main column.

The liquid that collects in the bottom of the argon column is primarily liquid oxygen, which is taken back to the main column. The vapor that collects in the top of the argon column is primarily argon and nitrogen, with a small amount of oxygen.

If the argon is desired as a product, the crude material from the argon column may be purified

by two techniques: (1) catalytic combustion to remove oxygen, followed by rectification to separate the argon and nitrogen, and (2) adsorption to remove oxygen, followed by rectification for nitrogen – argon separation.

In the *catalytic combustion system*, the crude argon is mixed with an appropriate amount of hydrogen gas and passed through a catalytic combustion unit to combine the oxygen impurity with hydrogen to form water vapor. The bulk of the water vapor is subsequently removed by passing the gas through a condenser; the last traces of water are removed by subsequently passing the gas through a set of alumina driers. The excess hydrogen gas is separated from the argon stream by partial condensation in the boiler of the rectification column for argon – nitrogen separation. The final gas mixture is rectified to produce essentially pure argon liquid or gas.

In the *adsorption system*, the crude argon is first passed through a rectification column to separate the nitrogen from the mixture. The oxygen – argon liquid is removed from the boiler of the column and evaporated in a heat exchanger. The gas is passed through a set of adsorbent traps filled with synthetic zeolite or other absorbent, where the oxygen is removed. If the oxygen is desired as a liquid, then the gaseous oxygen is condensed by using some of the liquid nitrogen product from the main column as the refrigerant in an argon condenser.

6.9. Neon Separation

The heat transfer rate in the condenser –reboiler in an air separation system begins to decrease after the double-column system has been in continuous operation for a long period of time (2 – 3 d). The reason for this drop in performance is the gradual collection of a noncondensable component (mainly neon and helium) on the nitrogen side of the heat exchanger. This accumulation lowers the partial pressure of the nitrogen in the mixture to the point where it no longer condenses at the condenser temperature.

This problem is solved by periodically venting a small portion of the gas from the dome of the condenser – reboiler, if the collection of neon is not required. The vent gas contains ca. 2.1 mol % neon and 0.8 mol % helium, with traces of hydrogen.

The commercial production and distribution of neon on a large scale began in the early 1960s in response to demand, e.g., as a refrigerant and for applications in decorative neon lighting [69]. Neon can be collected by using the air-separation column shown in Figure 40.

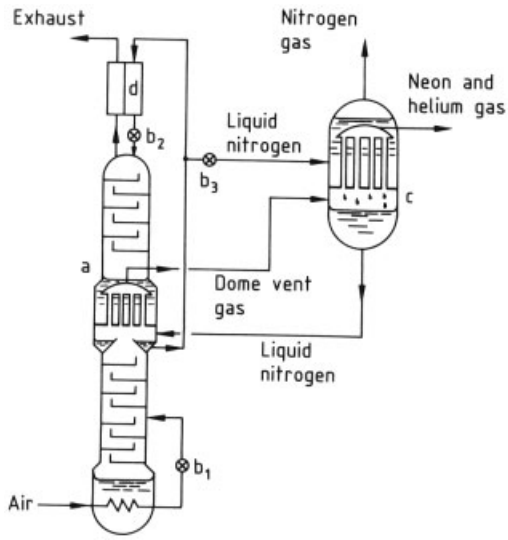

Figure 40. Neon recovery subsystem attached to the double-column air separation column
a) Main column; b) J – T valve; c) Neon condenser – rectifier; d) Subcooler

The neon – helium portion of the dome vent gas may be separated from the nitrogen by passing the gas mixture through a condenser – rectifier (c) that is cooled by liquid nitrogen expanded through a J – T valve (b_3) from the main column. The boiling points of neon (27.1 K) and helium (4.21 K) are so much lower than that of nitrogen (77.4 K) that the liquid formed in the condenser – rectifier is practically pure nitrogen. The typical composition of the vapor from the condenser – rectifier unit is 46 mol % neon, 19 mol % helium, 33 mol % nitrogen, and 2 mol % hydrogen.

The hydrogen can be removed from the mixture by using a catalytic combustion chamber followed by a drier to remove the water vapor formed during combustion. Excess oxygen and nitrogen are then removed in an adsorption trap. The helium and neon are separated in a con-

denser cooled by a closed-cycle hydrogen refrigeration system.

6.10. Physical Adsorption

When a gas contacts a solid surface, part of the gas is taken up by the solid. The general process is called sorption. If sorption is accompanied by a chemical reaction between the gas and the solid, the process is called chemical adsorption or chemisorption. If only the weak intermolecular forces between the gas and solid are involved, the process is called physical adsorption. Physical adsorption is used frequently in cryogenic separation systems because the adsorbent material can be regenerated and reused.

Several theories have been proposed to correlate the data on physical adsorption. One of the most successful theories was proposed by BRUNAUER, EMMETT, and TELLER (the BET theory), which yields the following expression for the volume of gas adsorbed per unit mass of adsorbent [70]:

$$\frac{V}{m_{\mathrm{a}}}=\frac{\mathrm{v_m}C\,(p/p_{\mathrm{sat}})}{(1-p/p_{\mathrm{sat}})\,[1+(C-1)\,(p/p_{\mathrm{sat}})]}$$

where V = volume of gas adsorbed at STP, m_{a} = mass of adsorbent, v_{m} = volume of gas required to form a monomolecular layer over the surface of the adsorbent per unit mass of adsorbent, p = partial pressure of the gas being adsorbed, and p_{sat} = saturation pressure of the gas at the adsorbent temperature. The parameter C is a function of the net energy of adsorption and the temperature of the adsorbent.

$$C=\exp\,(\theta/T)$$

where θ = difference between the heat of adsorption of the first layer and the latent heat of the gas. For adsorption of nitrogen gas on charcoal at 77.3 K, v_{m} = 0.1815 $\mathrm{m^3/kg}$, and θ = 300 K. For adsorption of argon gas on charcoal at 77.3 K, v_{m} = 0.2156 $\mathrm{m^3/kg}$ and θ = 341 K. Physical adsorption is widely used in cryogenic systems that separate and purify gas mixtures and in systems that produce and maintain a vacuum. One of the advantages of the use of an adsorbent as a vacuum pump is that the medium in the vacuum chamber is not contaminated by the oil used in mechanical and diffusion vacuum pumps. Cryogenic adsorption has substantially extended the range of separable gas mixtures. The gas mixtures that may be easily separated by using adsorption at cryogenic temperatures include methane – helium, neon – helium, hydrogen – helium, and nitrogen – carbon dioxide.

The basic arrangement for adsorption separation of a gas mixture is shown in Figure 41. The gas stream is first cooled by passing it through a heat exchanger (a) and then through one of the adsorbent beds (b_1). The component to be removed is adsorbed in the bed, which is usually refrigerated by liquid nitrogen. During this cycle, valves (c_1), (c_3), (c_6), and (c_8) are open, and valves (c_2), (c_4), (c_5), and (c_7) are closed. While adsorbent bed (b_1) is active in adsorbing the gas to be removed, adsorbent bed (b_2) is regenerated or reactivated by warming it to ambient temperature (or to a higher temperature) and purging it

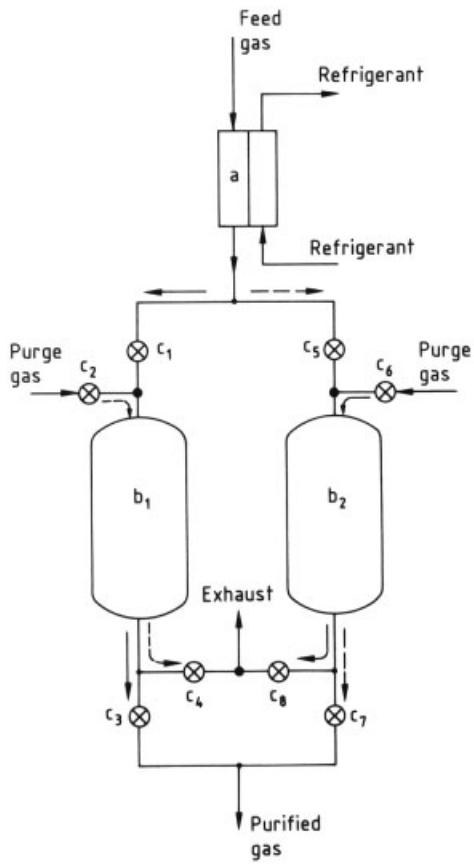

Figure 41. Adsorption separation system
a) Heat exchanger; b) Adsorbent bed; c) Control valve

with a gas. When the cold adsorbent bed (b_1) becomes saturated with the material to be removed, the valves are reversed, and the regenerated bed (b_2) begins to adsorb the gaseous impurity.

7. References

1. R. B. Scott: *Cryogenic Engineering,* D. van Nostrand Co., Princeton, N.J., 1959, p. 1.
2. C. A. Bailey: *Advanced Cryogenics,* Plenum Press, New York 1971, pp. 355 – 371.
3. M. Sittig, S. Kidd: *Cryogenics–Research and Applications,* D. van Nostrand Co., Princeton, N.J., 1963, p. 190.
4. R. F. Barron, *Cryogenics* **22** (1982) 409 – 413.
5. N. R. Brayton: *Cryogenic Recycling and Processing,* CRC Press, Boca Raton, Fla., 1980, pp. 211 – 219.
6. D. C. Brown: "The Application of Cryogenic Fluids to Freezing of Foods," *Advances in Cryogenic Engineering,* vol. 12, Plenum Press, New York 1967, pp. 11 – 22.
7. K. D. Timmerhaus, R. F. Kamm, J. D. Bays, "Whither Cryogenics," *Advances in Cryogenic Engineering,* **vol. 10 B,** Plenum Press, New York 1965, pp. 1 – 6.
8. J. E. Zimmerman, R. Radebaugh: "Operation of a SQUID in a Very Low-Power Cryocooler," NBS Special Publ. 508, 1978.
9. K. Oshima, Y. Kyotani: "High-Speed Transportation Levitated by Superconducting Magnet," *Advances in Cryogenic Engineering,* **vol. 19,** Plenum Press, New York 1974, pp. 127 – 136.
10. A. U. Smith: *Current Trends in Cryobiology,* Plenum Press, New York 1970, pp.153 – 180.
11. I. S. Cooper: "Cryogenic Surgery Introduction," in H. von Leden, W. G. Cahan (eds.): *Cryogenics in Surgery,* Medical Examination Publ. Co., Flushing, N.Y. 1971, pp. 31 – 38.
12. R. F. Barron: *Cryogenic System,* 2nd ed., Oxford Univ. Press, London 1985, pp. 61 – 64.
13. K. Wark: *Thermodynamics,* 3rd ed., McGraw-Hill, New York 1977, p. 571.
14. M. Davies: *The Physical Principles of Gas Liquefaction and Low Temperature Rectification,* Longmans, Green & Co., London 1949, p. 37.
15. F. Kreith, M. S. Bohn: *Principles of Heat Transfer,* 4th ed., Harper & Row Publ., New York 1986, pp. 407 – 408.
16. G. Walker: *Cryocoolers,* Plenum Press, New York 1983, Part 1, pp. 280 – 282.
17. J. G. Daunt: "The Production of Low Temperatures Down to Hydrogen Temperatures," *Handbuch der Physik,* vol. 4, Springer-Verlag, Berlin 1956.
18. D. T. Linnett, K. C. Smith, "Mixed Refrigerant Cascade Liquefiers for Natural Gas – Design and Optimization," Advances in Cryogenic Engineering, vol. 15, Plenum Press, New York 1970, pp. 8 – 17.
19. S. C. Collins, *Rev. Sci. Instrum.* **18** (1947) 157.
20. S. C. Collins, R. L. Cannaday: *Expansion Machines for Low Temperature Processes,* Oxford Univ. Press, London 1958.
21. J. McInroy, *Refrig. Sci. Technol.* **1** (1967) 59 – 68.
22. R. F. Barron in Reference 11, "Cryoinstrumentation," p. 115.
23. R. F. Barron: "Liquefaction Cycles for Cryogens," *Advances in Cryogenic Engineering,* **vol. 17,** Plenum Press, New York 1972, pp. 20 – 36.
24. W. H. Hartwig: "Cryogenic Refrigeration Concepts Utilizing Adsorption Pumping in Zeolites," *Advances in Cryogenic Engineering,* **vol. 23,** Plenum Press, New York 1978, pp. 438 – 447.
25. C. B. Hood, W. W. Vogelhuber, C. B. Barnes: "Helium Refrigerators for Operation in the 10 K to 30 K Range." *Advances in Cryogenic Engineering,* **vol. 9,** Plenum Press, New York 1964, pp. 406 – 506.
26. G. Walker: *Stirling Engines,* Oxford Univ. Press, London 1980.
27. A. Kirk, *Proc. Inst. Civ. Eng. (London)* **37** (1874) 244– 315.
28. J. W. L. Köhler: "The Gas Refrigerating Machine and Its Position in Cryogenic Technique," in K. Mendelssohn (ed.): *Progress in Cryogenics,* vol. 2, Academic Press, London 1960, pp. 41 – 68.
29. T. Finkelstein: "Computer Analysis of Stirling Engines," *Advances in Cryogenic Engineering,* **vol. 20,** Plenum Press, New York 1975, pp. 269 – 282.
30. R. Vuilleumier, US 1 275 507, 1918.
31. R. White: "Vuilleumier Cycle Cryogenic Refrigeration," AFFDL-TR-76-17, Wright-Patterson Air Force Base, Dayton, Ohio, AD/A-027-055, 1976.
32. E. Solvay: DE 39 280, 1887.
33. W. E. Gifford, H. O. McMahon: "A New Low Temperature Gas Expansion Cycle – Part II," *Advances in Cryogenic Engineering,* vol. 5, Plenum Press, New York 1960, pp. 368 – 372.

34. W. E. Gifford: "The Gifford-McMahon Cycle," *Advances in Cryogenic Engineering,* **vol. 11,** Plenum Press, New York 1961, pp. 152 – 159.
35. W. E. Gifford, T. E. Hoffman: "A New Refrigeration System for 4.2 K," *Advances in Cryogenic Engineering,* **vol. 6,** Plenum Press, New York 1961, pp. 82 – 94.
36. W. E. Keller: *Helium-3 and Helium-4,* Plenum Press, New York 1969, p. 112.
37. W. F. Giauque, *J. Am. Chem. Soc.* **49** (1927) 1870.
38. P. Debye, *Ann. Phys. (Leipzig)* **81** (1929) 1154.
39. W. F. Giauque, D. P. MacDougal, *Phys. Rev.* **43** (1933) 768.
40. W. J. de Haas, E. C. Wiersma, H. A. Kramers, *Nature (London)* **131** (1933) 719.
41. C. V. Heer, C. B. Barnes, J. G. Daunt, *Rev. Sci. Instrum.* **25** (1954) 1088.
42. J. A. Barclay, W. F. Stewart, W. C. Overton, R. J. Chandler, O. D. Harkleroad: "Experimental Results on a Low-Temperature Magnetic Refrigerator," *Advances in Cryogenic Engineering,* **vol. 31,** Plenum Press, New York 1985, pp. 743 – 752.
43. W. A. Steyert: "Applications of Closed-Cycle Cryocoolers to Small Superconducting Devices," NBS Special Publ. 508, 1978, p. 81. *J. Appl. Phys.* **49** (1978) 1227.
44. H. London: *Proc. Ist Int. Conf. Low Temperature Physics,* Oxford, Clarendon Lab 1951, p. 157.
45. H. London, G. R. Clark, E. Mendoza: *Phys. Rev.* **128** (1962) 1992.
46. K. G. Walters, W. M. Fairbank, *Phys. Rev.* **103** (1956) 262.
47. P. Das, R. de Bruyn Ouboter, K. W. Taconis: *Proc. 9th Int. Conf. on Low Temperature Physics,* Plenum Press, New York 1965, p. 1253.
48. H. E. Hall, P. J. Ford, K. Thompson, *Cryogenics* **6** (1966) 80.
49. B. Neganov, N. Borisov, M. Libury, *Zh. Eksp. Teor. Fiz.* **50** (1966) 1445. *Sov. Phys. JETP (Engl. Transl.)* **25** (1966) 959.
50. R. Radebaugh: "Thermodynamic Properties of H3 – He4 Solutions," NBS Technical Note 362, U.S. Govt. Printing Office, Washington, D.C., 1967.
51. G. Frossati: "Proc. LT15," *J. Phys.* (Paris) Collq. C6, 39 Suppl. no. 8 (1978) 1578.
52. T. E. Strobridge: "The Thermodynamic Properties of Nitrogen from 64 to 300 K between 0.1 and 200 Atmospheres," NBS Technical Note 129, 1962.
53. Reference 16, Part 2, pp. 6 – 8.
54. S. C. Collins, R. L. Canaday: *Expansion Machines for Low Temperature Processes,* Oxford Univ. Press, London 1958, p. 108.
55. R. B. Fleming: "A Compact Perforated-Plate Heat Exchanger", *Advances in Cryogenic Engineering,* vol. 14, Plenum Press, New York 1969, pp. 197 – 204.
56. P. G. Kroeger: "Plated Tube Heat Exchanger: Analytical Investigation of a New Surface Concept." *Advances in Cryogenic Engineering,* **vol. 12,** Plenum Press, New York 1967, pp. 240 – 351.
57. R. B. Fleming: "The Effect of Flow Distribution in Parallel Channels of Counterflow Heat Exchangers," *Advances in Cryogenic Engineering,* **vol. 12,** Plenum Press, New York 1967, pp. 352 – 362.
58. K. W. Cowans: "A Countercurrent Heat Exchanger That Compensates Automatically for Maldistribution of Flow in Parallel Channels," *Advances in Cryogenic Engineering,* **vol. 19,** Plenum Press, New York 1974, pp. 437 – 444.
59. G. Claude: *Liquid Air, Oxygen and Nitrogen,* P. Blakiston Son & Co., Philadelphia, Pa., 1913.
60. M. L. Land: "Expansion Turbines and Engines for Low Temperature Processing," *Advances in Cryogenic Engineering,* **vol. 5,** Plenum Press, New York 1960, pp. 250 – 260.
61. H. Sixsmith: "The Theory and Design of Gas-Lubricated Bearings of High Stability," *Proc. First Int. Symp. on Gas-Lubricated Bearings,* ACR-49, Office of Naval Research, Washington, D.C., 1959.
62. B. W. Birmingham, H. Sixsmith, W. A. Wilson: "The Application of Gas-Lubricated Bearings to a Miniature Helium Expansion Turbine." *Advances in Cryogenic Engineering,* **vol. 7,** Plenum Press, New York 1962, pp. 30 – 42.
63. W. L. McCabe, J. C. Smith: *Unit Operations of Chemical Engineering,* 3rd ed., McGraw-Hill, New York 1976, p. 478.
64. C. D. Holland: *Multicomponent Distillation,* Prentice-Hall, Englewood Cliffs, N.J., 1963, pp. 14 – 42.
65. M. Ponchon, *Tech. Mod* **13** (1921) 20 – 55.
66. W. L. McCabe, E. W. Thiele, *Ind. Eng. Chem.* **17** (1925) 605.
67. A. Arkharov, I. Marfenina, Ye. Mikulin: *Theory and Design of Cryogenic Systems,* Mir Publ., Moscow 1981, pp. 333 – 337.

68. G. A. Cook: *Argon, Helium, and the Rare Gases,* **vol. 1,** Interscience Publ., New York 1961, pp. 41 – 44.
69. G. F. Hagenbach, J. H. Schiffhauer: "Commercial Production of Neon," *Advances in Cryogenic Engineering,* **vol. 9,** Plenum Press, New York 1964, pp. 557 – 560.
70. S. Brunauer, P. H. Emmett, E. Teller: *J. Am. Chem. Soc.* **60** (1938) 309.

Vacuum Technology

Hans Peter Lorenz, Siemens AG, Erlangen, Federal Republic of Germany

1. Principles 2
1.1. Components 2
1.2. Physical Principles 2
1.2.1. Equations of State of Gases 2
1.2.2. Pressure and Temperature 4
1.2.3. Thermal Mean Velocity 4
1.2.4. Thermal Particle Interactions 4
1.2.5. Flow Regimes, Momentum, and Heat Transfer 5
1.2.6. Vacuum Regimes 6
1.3. Matter under Vacuum 7
1.3.1. Vapor Pressure 7
1.3.2. Sorption, Gettering, Permeation . . 7
2. Evacuation 9
2.1. Characteristic Quantities 9
2.2. Classification of Vacuum Pumps . 10
2.3. Technology of Vacuum Pumps . . 12
2.3.1. Displacement Pumps 12
2.3.1.1. Liquid and Gas Ring Vacuum Pumps 12
2.3.1.2. Sliding Vane Rotary, Rotary Plunger, Rotary Piston Pumps 14
2.3.1.3. Rotary Lobe Vacuum Pumps 15
2.3.1.4. Pump Oils for Rough and Medium Vacuum Pumps 16
2.3.2. Kinetic Vacuum Pumps 16
2.3.2.1. Working Fluid Jet Vacuum Pumps . 16
2.3.2.2. Diffusion Pumps 16
2.3.2.3. Jet Pumps–Working Fluids 17
2.3.2.4. Drag Vacuum Pumps–Turbomolecular Pump 18
2.3.3. Entrapment Vacuum Pumps 19
2.3.3.1. Sorption Vacuum Pumps 19
2.3.3.2. Getter Ion Pumps 19
2.3.3.3. Condensation Pumps 20
3. Vacuum Diagnostics 21
3.1. Total Pressure Measurement . . . 21
3.1.1. Mechanical Vacuum Instruments . . 21
3.1.2. Viscosity Vacuum Gauge 23
3.1.3. Thermal Conductivity Vacuum Gauge 23
3.1.4. Ionization Vacuum Gauge 24
3.1.5. Measuring Systems 25
3.2. Partial Pressure Measurement . . 25
3.3. Leak Surveys 27
4. Vacuum System Components . . . 28
4.1. Piping Components 28
4.2. Vacuum System Feedthroughs . . 29
4.3. Valves 29
5. Vacuum System Applications . . . 31
5.1. Perspectives on Design of Vacuum Systems 32
5.2. Examples of Vacuum Systems and Their Application 33
5.2.1. Decontamination of Groundwater . 33
5.2.2. Degassing Steel with a Steam Jet Vacuum System 33
5.2.3. High Vacuum Systems for Manufacturing Thin Films 33
5.2.4. Ultra High Vacuum System for the LEP Storage Ring at Geneva 35
6. References 37

1. Principles

Vacuum is the condition of a gas whose particle number density is less than that of the atmosphere at the surface of the earth. Since the particle number density is time- and location-dependent, within certain limits, an exact upper limit of vacuum cannot be specified [1]. Corresponding to this definition, the concept *vacuum* comprises an extraordinarily broad region of conditions of gases and gas mixtures. Equipment and instruments that serve to produce, apply, and diagnose these most varied vacuums are included within the concept *vacuum technology.*

Vacuum technology is an ancillary technique; it produces the vacuum needed to manufacture a product or to allow a technical component to function. Indeed, technologies that use vacuum may often be replaced by versions without vacuum. Applications of vacuums and vacuum technology are numerous and are to be found in all fields of technology. Examples are vacuum distillation in the chemical laboratory, freeze-drying in the food industry, vacuum facilities in large basic-research projects, such as fusion research and in everyday technology, such as control by vacuum in an automotive engine or the household vacuum cleaner. Figure 1 gives an

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b03_21

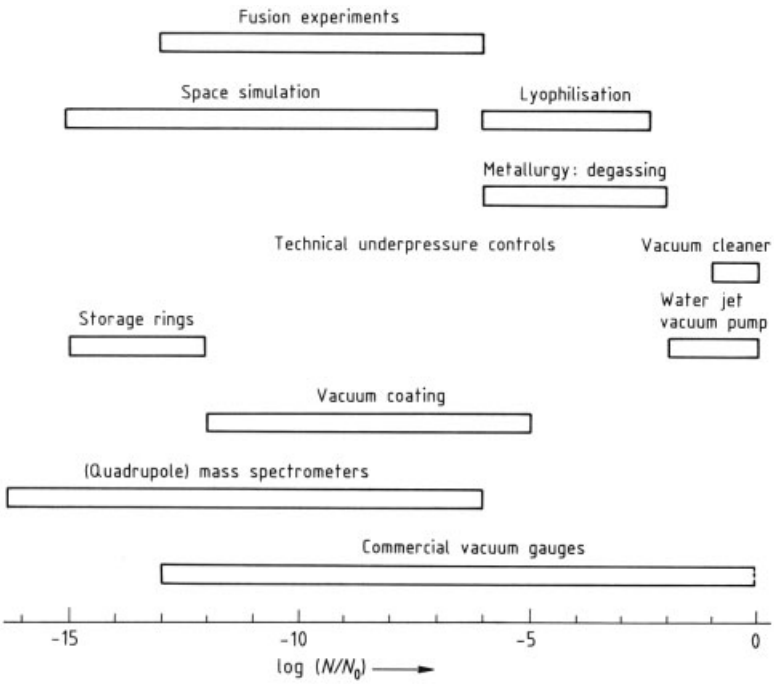

Figure 1. Vacuum technology: applications and relative number density of molecules, N/N_0.

overview of the application of vacuum technology. The abscissa scale, in powers of ten, serves to compare the particle density with that of the surrounding atmosphere.

1.1. Components

Industrial vacuum systems are built up of various components:

1) Vacuum chambers or receivers
2) Piping, pipe fittings, and flanges
3) A large variety of valves for shutoff and control
4) Pumps
5) Diagnostic instrumentation, (a) for measuring total pressure and (b) for analyzing the composition of gas mixtures
6) A multitude of ancillary equipment, (a) leak testers, (b) vacuum system leadthroughs, (c) cold traps, (d) electronic control systems

By means of symbolic figures, flow diagrams represent the interplay of all these components in a vacuum system. Figure 2 shows a small selection of these flow symbols [2].

1.2. Physical Principles

The physical principles of vacuum technology are derived from thermodynamics and the kinetic theory of gases.

1.2.1. Equations of State of Gases

According to the phenomenological theory of thermodynamics, the state variables temperature T, pressure p, volume V, and its component materials characterize a system. At sufficiently low pressure the interaction between the molecules of a gas may be neglected and the ideal gas equation describes the relationship between the state variables:

$$pV = nRT = NkT \tag{1}$$

where

n = number of mols per unit volume
R = universal gas constant, $8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}$
k = Boltzmann constant, $1.3805\times10^{-23}\,\mathrm{JK^{-1}}$

$N =$ number of particles per unit volume

With increasing pressure, the gas behavior deviates increasingly from that described by Equation (1). Corrections need to be made for the volume occupied by the molecules themselves and for the difference in the forces acting within and at the periphery of the volume, which lead to the van der Waals equation of state:

$$\left(p+a/V^2\right)(V-b)=RT \tag{2}$$

where *a, b* are constants specific to individual gases.

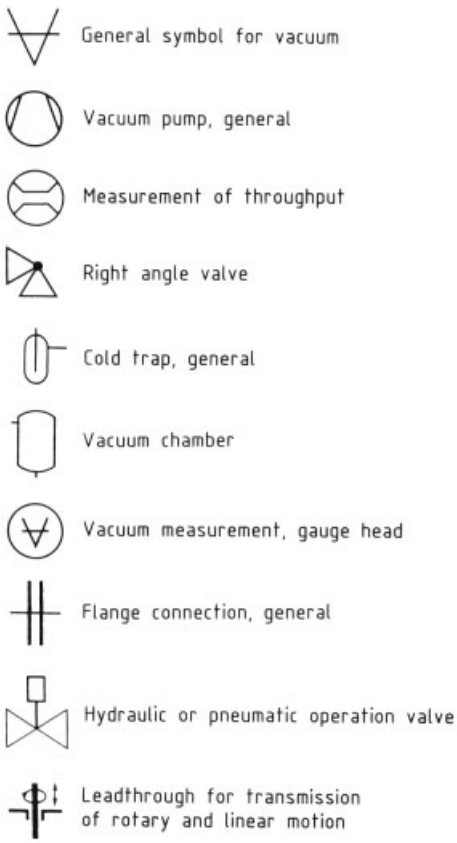

Figure 2. Some graphical symbols of vacuum technology [2]

1.2.2. Pressure and Temperature

The atomistic foundation of the kinetic theory of gases offers a satisfactory explanation of the phenomena associated with vacuum. It views an ideal gas as a collection of dimensionless mass points that move about in disarray at all possible velocities and in all possible directions without interaction. Elastic impacts of the gas particles with the container walls produce the macroscopically registered pressure:

$$p=nm\cdot\bar{v}^2/3=\varrho\cdot\bar{v}^2/3 \tag{3}$$

$\bar{v}^2$ = mean square particle velocity
n = number density of particles
m = mass of a particle
ϱ = macroscopic gas density

Pressure has the dimensions force per unit area. In pressure measurement, several units are still used.

The SI unit, the Pascal, abbreviated Pa [3] is

$$1\,\mathrm{Pa} = 1\,\mathrm{N/m^2} \tag{4}$$

In vacuum technology, the most frequently used unit is the millibar, mbar, 1 mbar = 100 Pa. Table 1 gives a compilation of the various pressure units often used.

Comparison of Equation (3) with the ideal gas equation (Eq. 1) yields Equation (5):

$$\bar{v}^2 = 3RT/M = 3kT/m \tag{5}$$

where M = molar mass of the gas

Equation (5) links the thermodynamic temperature T with the movement of the gas particles: the mean square velocity of the particles is proportional to the temperature T and inversely proportional to the mass m of the particles.

1.2.3. Thermal Mean Velocity

The $\bar{v}^2$ in Equations (3) and (5) is calculated using the *Maxwellian velocity distribution*, which can be derived from methods of statistical physics.

The number of particles dN with a velocity between v and dv is:

$$\mathrm{d}N=Nv\cdot(m/2kT)^{1.5}\cdot \exp\left(-mv/2kT2\right)\mathrm{d}v\cdot 1/4\cdot\sqrt{\pi}2$$

The resulting mean particle velocity is

$$\bar{v}=\sqrt{8RT/M\cdot\pi}=\sqrt{8kT/m\cdot\pi} \tag{6}$$

Table 1. Pressure conversion factors

	Pa	mbar	Torr	psi	μ
Pa	1	1×10^{-2}	7.50062×10^{-3}	1.45038×10^{-4}	7.50062
a) mbar	100	1	0.75	1.45×10^{-2}	750
b) Torr	133	1.33	1	1.93×10^{-2}	1000
c) psi	6.895×10^{3}	68.95	51.7	1	5.17×10^{4}
d) μm	0.133	1.33×10^{-3}	1×10^{-3}	1.93×10^{-5}	1

To convert from a unit in the first column to a unit in the top row multiply by the entry in the table.
a) 1 bar = 0.986923 atm
b) 1 Torr = 1 mm Hg, 760 Torr = 1 atm
c) pound force per square inch, lbf/in^2
d) micrometer

Table 2 lists the mean particle velocities at room temperature for some gases important in vacuum technology.

Table 2. Average velocities of molecules at 20 °C

Gas	M_r (mean value)	Velocity $\bar{v}$, m/s	Reference
Air	(29)	462	
H_2	2.016	1754	[4]
He	4.003	1245	[4]
N_2	28.013	471	[4]
O_2	31.999	440	[4]
H_2O	18.015	587	[4]
Balzers P-3 ∗	(190)	181	[5]
DC 705 ∗	546	107	[6]
Fomblin 06/6 ∗	(1800)	59	[5]
Apiezon C ∗	574	104	[5]

∗ Vacuum pump fluid

1.2.4. Thermal Particle Interactions

The interaction of gas particles with one another is determined in the kinetic theory of gases by the gas kinetic effective molecular diameter d. The collision cross section $\sigma = \pi\, d^2$. The mean free path is defined as the average distance a particle can travel between two collisions

$$\bar{l} = 1/\sigma n \sqrt{2} = kT/\sigma p \sqrt{2} \tag{7}$$

In a directed particle beam, N of N_0 participating particles have not sustained a collision after traversing a distance $x < \bar{l}$:

$$N = N_0 \exp\left(-x/\bar{l}\right) \tag{8}$$

The particle stream density through a cross section of the surface A, i.e., the number of particles per unit time per unit area impinging on a surface may be calculated from

$$j = \bar{v} n/4 = p/\sqrt{(2mkT\pi)} \tag{9}$$

where j is a measure of the maximum possible rate of vaporization in a high vacuum system. Or, in the regime of molecular flow, j can be considered as the value of the maximum particle pumping rate through an opening. For example, for nitrogen at a pressure of 10^{-6} mbar, j is $2.9\times10^{-18}\ \mathrm{s}^{-1}\ \mathrm{m}^{-2}$.

If n_{mono} designates the number of particles covering a surface in a monolayer, then, with Equation (9), the monolayer time t_{mono} for covering an initially gas-free surface may be calculated

$$t_{\mathrm{mono}} = n_{\mathrm{mono}}/\alpha j = n_{\mathrm{mono}} \cdot \sqrt{(2mkT\pi)}/p\alpha \tag{10}$$

where the sticking coefficient α is a measure of the retention probability of impinging particles.

Figure 3 shows $\bar{l}$ for helium, hydrogen, nitrogen – air, and water vapor as well as j and t_{mono} for nitrogen at room temperature.

1.2.5. Flow Regimes, Momentum, and Heat Transfer

Vacuum presumes the creation and maintenance of *pressure gradients,* which are generated by transport mechanisms caused by gradients of physical quantities. Gradients in the gas space cause transport. Thus, a concentration gradient leads to diffusion, combined with mass transfer; a velocity gradient gives rise to momentum transfer; heat transfer results from temperature differences. While diffusion plays no critical part in pumping processes, the pumping action

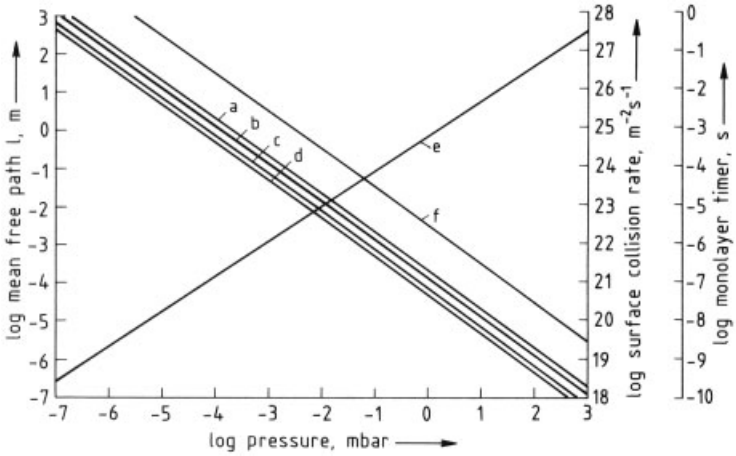

Figure 3. Mean free path, surface collision rate, monolayer time
a) Helium; b) Hydrogen; c) Nitrogen/Air; d) Water; e) j (nitrogen); f) t_{mono}(nitrogen)

of a substantial class of vacuum pumps depends on momentum transfer.

Thermal conductivity is the principle behind certain pressure gauges. The actual transport mechanisms depend on the pressure regime. If d is a typical length dimension for a vacuum system, then the cases $d \gg \bar{l}$, $d \sim \bar{l}$, and $d \ll \bar{l}$ must be distinguished. The Knudsen number K is introduced for classification:

$$K = \bar{l}/d \tag{11}$$

In the case of viscous flow, $K \ll 1$. Only collisions between the gas particles determine the behavior of the gas. Wall effects may be neglected. If a velocity gradient exists in a flowing stream, then momentum transfer occurs in that direction. The faster portions of the stream are decelerated, the slower accelerated, and internal friction arises. The coefficient of the internal friction, ε, is also known as the dynamic viscosity:

$$\varepsilon = n m v \bar{l}/3 \tag{12}$$

ε is independent of pressure as long as $d \gg \bar{l}$.

Heat transfer is described by the thermal conductivity λ:

$$\lambda = n m \bar{v} \bar{l} c_{\mathrm{v}}/3 = \varepsilon c_{\mathrm{v}} \tag{13}$$

where c_{v} is the specific heat capacity of the gas at constant volume.

In the region where $K \ll 1$, both ε and λ are independent of pressure. As long as viscous flow is governed only by internal friction, the flow regime is laminar. If the convection mechanism becomes important, then the flow regime becomes turbulent. The flow regimes may be classified according to their Reynolds number [4]. Normally, turbulent flows do not arise in vacuum technology.

Through pressure reduction, the mean free path length may become greater than a typical dimension of the vacuum container, and then the region of molecular flow is attained. Equation (12), for ε, and Equation (13), for λ, lose their meaning. Now wall – particle interactions dominate. The pressure-dependent coefficient of external friction characterizes the transfer of momentum. In this regime, wall collisions impart to the gas an impulse component parallel to the wall. Also, transfer of energy is pressure-dependent and influenced additionally by the specific properties of the surfaces involved. Accordingly, the thermal conductivity in the molecular flow regime is smaller than in the region of viscous flow.

The transition region, $K \sim 1$, is designated as the region of intermediate flow. Internal and external friction have equal influence on events. Equations that embody viscous and molecular flow as limiting cases can be derived.

1.2.6. Vacuum Regimes

The Knudsen number K may be used to effect the conventional, but relatively arbitrary, classification of the vacuum regimes. Table 3 shows the assignment of values of the vacuum parameters K, $\bar{l}$, and t_{mono} to rough vacuum (RV), medium vacuum (MV), high vacuum (HV), and ultra high vacuum (UHV). Since, in UHV, the mean free path length is always much larger than the container dimensions, both K and $\bar{l}$ are less meaningful quantities. The monolayer time, t_{mono}, however, gives better information about the conditions in the vacuum vessel.

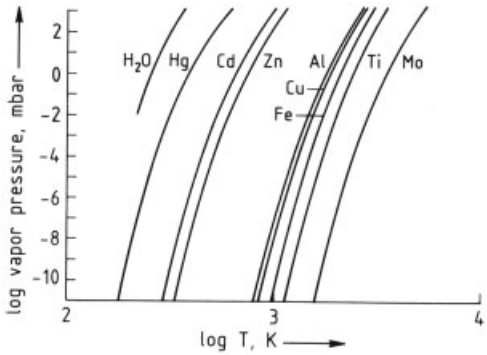

Figure 4. Vapor pressure of some materials used in vacuum technology (data from [8–10])

Rough and Medium Vacuum. The interactions within the gas characterize the regimes of rough and medium vacuum. In MV, the wall effects become stronger. Gas flows are laminar and seldom turbulent. Viscosity and thermal conductivity are independent of pressure. Sorption phenomena on the container walls may be ignored. Most vacuum processes in the chemical, food, and metallurgical industries take place in this regime.

High Vacuum. Gas – wall interactions are always a significant factor for operations in the high vacuum region. Outgassing the vessel walls contributes significantly to the gas particle balance sheet and to gas analysis, and may have major impacts on the process engineering of industrial processes. Important applications may be found in thin film technologies, e.g., manufacturing optical films, semiconductor coatings, and abrasion-resistant surfaces.

Ultra High Vacuum. In the UHV regime, the determining factors are the properties and condition of the surfaces of the vacuum equipment. Because of this, UHV equipment differs from that for the other vacuum regimes. Two basic requirements of UHV technology are accelerated degassing of surfaces by using temperatures up to 400 °C and using metal seals impervious to gases. Ultra high vacuum is required for particular coating manufacturing processes for semiconductor layers (molecular beam epitaxy [7]). Further examples of applications are analysis of the surfaces of solids, fusion research, and high-energy storage rings.

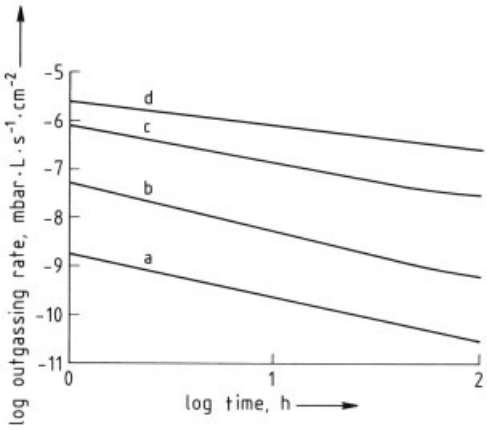

Figure 5. Time dependence of desorption rates (data from [15])
a) 304 stainless steel; b) 6016-T6 aluminum; c) Polyethylene; d) Chlorinated poly(vinyl chloride)

1.3. Matter under Vacuum

1.3.1. Vapor Pressure

The materials used for construction of vacuum units or the substances used within the unit, like internal parts and working fluids, may limit the attainable final pressure. The temperature-dependent pressure for equilibrium between gas and liquid or gas and solid phases can be derived by integration of the Clausius – Clapeyron equation. If a substance is present in more than one phase, then this equilibrium pressure represents the partial pressure of a substance in the system. It increases approximately exponentially with temperature (Fig. 4). For the most part, the vapor

Table 3. Vacuum ranges

Range	Pressure p, mbar	Knudsen number K	Mean free path $\bar{l}$, m	Monolayer time t_{mono}, s
Rough vacuum (RV)	$10^3 - 1$	$K \ll 1$	$10^{-7} - 10^{-4}$	$10^{-9} - 10^{-6}$
Medium vacuum (MV)	$1 - 10^{-3}$	$K \approx 1$	$10^{-4} - 10^{-1}$	$10^{-6} - 10^{-3}$
High vacuum (HV)	$10^{-3} - 10^{-7}$	$K \gg 1$	$\bar{l} > 10^{-1}$	$10^{-3} - 10$
Ultra high vacuum (UHV)	$<10^{-7}$			>10

pressures of the working fluids and sealing materials within the pumps limit the ultimate vacuum. Thus, water-filled liquid ring pumps achieve final pressures down to 15 mbar. Special pump oils with low vapor pressures make ultimate vacuums $<10^{-9}$ mbar possible with UHV pumps. Diligent attention must be directed at avoiding materials such as brass (zinc) and cadmium, which can contaminate high vacuum units at elevated temperatures.

1.3.2. Sorption, Gettering, Permeation

Van der Waals forces (physical adsorption) or chemical bonds (chemisorption) can bind gas particles to condensed matter. Deposition on the surface is called adsorption. If the gas penetrates into the body of the material, then the appropriate term is absorption. Desorption is the reversal of gas deposition. As the pressure is decreased desorption from the surfaces of vacuum equipment plays an increasing role in vacuum phenomena. If chemical processes (for example, reactive physical vapor deposition of thin coatings) take place in the vacuum gas space, these may be disturbed by desorbed gas, even at 10^{-3} mbar. The desorption also limits the attainable ultimate vacuum in the UHV region. The participating components distribute themselves according to the equilibrium between desorption and gas deposition resulting in a degree of surface coverage that depends on gas pressure. At constant temperature, the coverage can be characterized by adsorption isotherms. Through adsorption, substances with a vast internal surface, such as zeolites (molecular sieves [11]) or activated carbon can bond substantial amounts of gas and consequently pump gas. In addition to the surface adsorbed gases, the gases absorbed in the materials of the system, and the volatile constituents of the wall material itself, also contribute to gas emission into the vacuum. The gas emission is highly material-dependent and is affected by the pretreatment of the system, by the temperature, and by the duration of pumping. Table 4 lists typical desorption rates [12–14]. Figure 5 plots the desorption rate versus the pumping time. For metals, the outgassing rate is inversely proportional to the pumping time. As a rough classification, metals, glasses, and ceramics have gas emission rates $\leq 10^{-8}$ mbar m/s after a few hours. They are thus suitable for UHV. For polytetrafluoroethylene (PTFE), Viton, and a few other plastics, the gas emission is ca. 10^{-7} mbar m/s, and these materials can be used in high vaccum applications. Due to their high gas-emission rates, silicone rubber, butyl rubber, and epoxy resins have restricted use under high vacuum.

Table 4. Desorption rate of some materials at room temperature after 10 h pumping time

Material	Desorption rate, 10^{-10} mbar L s^{-1} cm^{-2}	Reference
PMMA (unfilled)	240 – 440	[12]
PMMA (filled)	770 – 4700	[12]
Epoxy (unfilled)	350 – 700	[12]
Epoxy (filled)	280 – 1000	[12]
Viton	2100	[14]
PVC	6700	[14]
Teflon	2	[14]
6063 aluminum	2	[13]
304 stainless steel	2	[15]
Alumina, dense	0.01	[14]

A further portion of the gas flow into the vacuum container comes about through permeation, i.e., by the flow of gas from the surrounding atmosphere through the seals and walls of the vacuum apparatus. The rate of diffusion in solids determines the rate of permeation. Unlike the rates of outgassing, permeation rates are independent of the duration of pumping. They depend ex-

ponentially on the temperature. Thus, they may limit the temperature at which a vacuum system can be degassed. Figure 6 shows the gas permeability for VITON, an elastomer in common use as a sealing material in vacuum applications. Table 5 lists some numerical values for permeability at room temperature for typical construction materials.

Table 5. Permeability of some materials at room temperature. Pressure difference 1 bar, Sample thickness 1 mm.

Material	Gas	Permeation flow density, mbar L s^{-1} cm^{-2}	Reference
Iron	H_2	8.1×10^{-9}	[16]
	N_2	6×10^{-19}	[16]
Quartz	H_2	7×10^{-13}	[16]
Copper	H_2	9.1×10^{-11}	[16]
Viton	H_2	1.5×10^{-7}	[17]
	N_2	1.9×10^{-8}	[17]
Palladium	H_2	9.1×10^{-7}	[16]

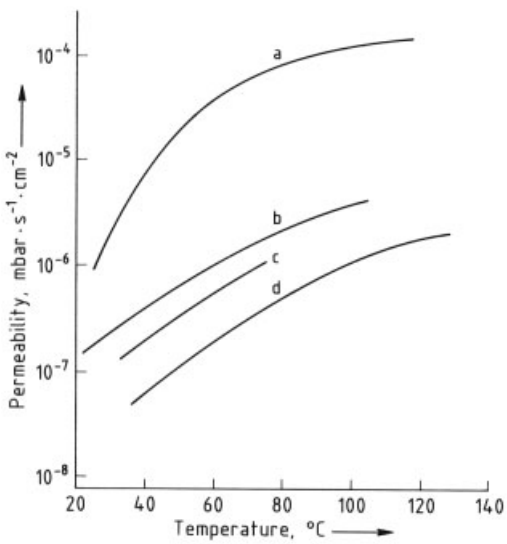

Figure 6. Permeation of various gases through VITON [17] a) Helium; b) Hydrogen; c) Argon; d) Nitrogen pressure difference 1 bar, sample thickness 1 mm

An adsorbent to which gas binds chemically can be termed a *getter*. In contrast to reversible physical adsorption (bond energies ≤ 10 kJ/mol), chemisorption is the result of bond energies > 100 kJ/mol and, in general, is irreversible. There is a distinction between sorption getters and vaporization getters. *Sorption getters* in the form of powders, sintered solids, or foil can bind gases within the body of the material. Getters in this group are titanium, thorium, zirconium and zirconium – aluminum alloys [18]. Subjecting a *vaporization getter* to evaporation produces a fresh, chemically active surface. The best vaporization getters are titanium, thorium, zirconium [19], barium, calcium, and strontium [18]. Getters can only bond chemically reactive gases, chiefly O_2, N_2, CO_2, and H_2, not the noble gases. Figure 7 shows examples of the effectiveness of various getters for hydrogen.

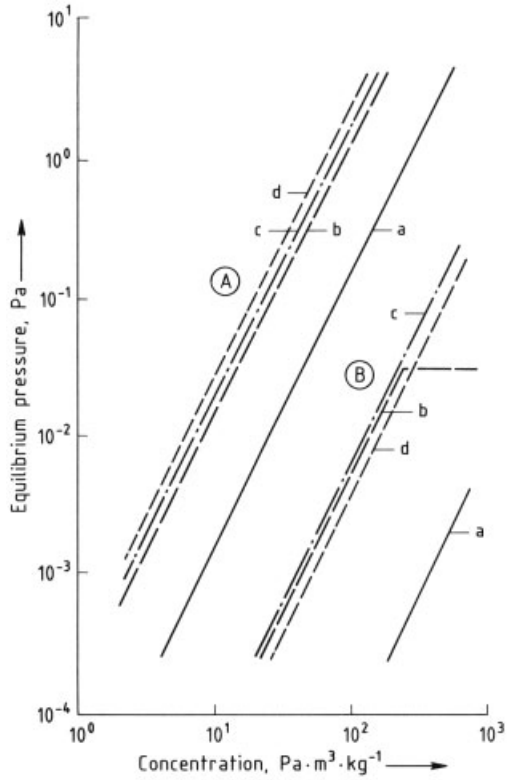

Figure 7. Hydrogen equilibrium pressure as a function of concentration in the getter material at different temperatures [18]
A) 700 °C; B) 400 °C
a) Zirconium – aluminum; b) Zirconium; c) Titanium; d) Zirconium – vanadium – iron

2. Evacuation

2.1. Characteristic Quantities

In order to produce a vacuum, gases must be transported through vessels, openings, piping, and pumping equipment. Some concepts, defined in detail, for example, in [16], serve to describe these operations. The throughput, $q_p\, v$, through the area A is a constant for the entire pump device:

$$q_{\mathrm{p}}\, v = p \cdot \Delta V / \Delta t = p\, A\, \bar{u} = \text{ constant} \tag{14}$$

where $\bar{u}$ = mean flow velocity over the cross section A

Suction capacity is the term used for the throughput for pumps. If p_1 is the pressure on the suction side of the pump, then one defines the pumping speed (ISO: volume flow rate) as:

$$S = q_{\mathrm{p}}\, \mathrm{v} / p_1 \tag{15}$$

Equations (9), (14), and (15) may be combined to give the theoretical pumping speed, the pumping rate required to remove all particles impinging on the surface A

$$S_{\mathrm{th}} = \mathrm{v} \cdot A / 4 \tag{16}$$

If p_2 designates the pump outlet pressure, then the pump compression ratio is:

$$\tau = p_2 / p_1 \tag{17}$$

The resistance to flow is:

$$R = (p_1 - p_2) / q_{\mathrm{p}}\, \mathrm{v} \tag{18}$$

And the flow conductance is:

$$C = 1/R = q_{\mathrm{p}}\, \mathrm{v} / (p_1 - p_2) \tag{19}$$

R and C are characteristic quantities for describing openings and piping.

If the pumping speed S is limited by a flow resistance with the conductance C caused by valves, baffles, or other component parts, as is always the case in a real layout, then one obtains an effective pumping speed by combining Equations (14), (15), and (19):

$$S_{\mathrm{eff}} = S \cdot C / (S + C) \tag{20}$$

Then the pump-down time t_{p} for a volume V necessary to reach a pressure p starting out from pressure p_0, may be determined using S_{eff}:

$$t_{\mathrm{p}} = V \cdot \ln (p_0 / p) / S_{\mathrm{eff}} \tag{21}$$

Assumptions for application of Equation (21) are negligible leak and outgassing rates and a constant S_{eff} in the pressure range under consideration. Therefore, in general, Equation (21) has absolute validity only in the medium vacuum region.

If several component elements are connected in series, their flow resistances are additive; if parallel, the conductances are additive.

Flow resistances and conductances depend on the geometry of the component parts and on the flow regime. For simple geometries, derived formulas can be produced, for example, see [16]. The ultimate vacuum is that produced after a long period of pumping by a continuously operating pump. The back pressure stability determines the maximum allowable pressure on the discharge side of the pump. For pumps with high compression ratio τ, the water vapor compatibility is also important. This decides the highest possible water vapor partial pressure in the intake gas.

2.2. Classification of Vacuum Pumps

To master the large range of particle densities embraced by the concept *vacuum* requires the application of extremely varied principles. The actual mechanism of the positive displacement pumps can be derived from the ideal gas law equation (Eq. (1)). The cavity in the pump whose volume varies periodically forces the gas from an inlet at pressure p_1 through the pump to the outlet at p_2. According to Equations (16) and (17), the pumping speed of a positive displacement pump operating periodically is independent of pressure: it is the product of the displacement volume V_1 and the frequency of filling the displacement volume ν_1:

S (for positive displacement pumps) = $V_1 \cdot \nu_1$

Yet, due to sealing losses (leaks) and the partial pressure of the pump working fluid, pressuredependence does arise at low pressures; S falls very quickly to the unusable low values shown by Curves a and b in Figure 8. Provided

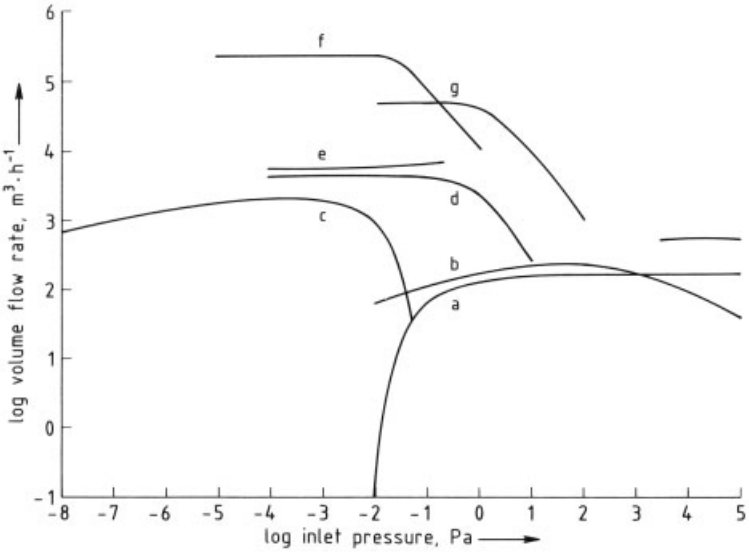

Figure 8. Volume flow rates of pumping units
a) Roots vacuum pump [23], b) Sliding vane rotary vacuum pump [21]; c) Diffusion pump [20]; d) Turbomolecular pump [25]; e) Getter ion pump [27]; f) Cryopump [24]; g) Diffusion ejector pump [22]; h) Liquid ring vacuum pump [26]

that isothermal conditions prevail and that Equation (1) is valid, the pump power requirement P may be calculated.

$$P = p_1 \cdot \ln\,(p_1/p_2)$$

Accordingly, for p_2 equal to atmospheric pressure, P goes through a maximum at p_1 equal to 390 mbar (Fig. 9). This power requirement curve must be considered in designing pumps for the anticipated inlet pressures and in providing for cooling. Positive displacement vacuum pumps are used almost exclusively in the rough and medium vacuum ranges, very often as the high pressure pumping stage, or forepump, in combination with high vacuum pumps.

For kinetic vacuum pumps, the fundamental basis for the pumping effect is momentum transfer. This is described by Equation (12) in the region of viscous flow and characterized in molecular flow by the pressure dependence of external friction. The class of working fluid vacuum pumps uses the gas – gas interchange effect; the molecular drag vacuum pumps use the gas – wall interchange effect. At low pressures, the pumping speed is limited by backflow of the working fluid or of the pumped gases. At high pressures, the pumping mechanism breaks down because the mean free path length becomes too small. Curves c and d in Figure 8 are examples of the trend of the pumping speed of kinetic vacuum pumps. Depending on the type of momentum transfer and on the nature of the working fluid, kinetic vacuum pumps cover all vacuum regimes from rough to ultra high vacuum.

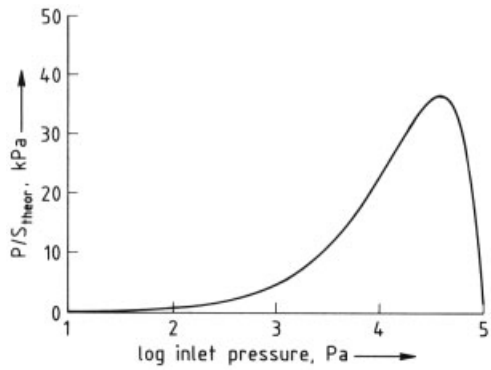

Figure 9. Pump power P/theoretical volume flow rate $S_{\mathrm{theor.}}$ vs. inlet pressure of positive displacement vacuum pumps

Positive displacement pumps, working fluid pumps, and molecular drag vacuum pumps move gases through the pumps by compression from the inlet to the outlet opening. Entrapment vacuum pumps accumulate the pumped gas in

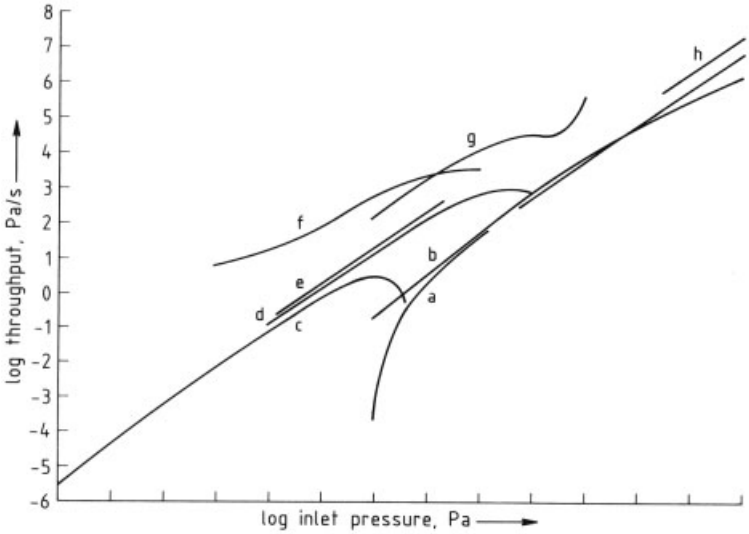

Figure 10. Throughput of pumping units
a) Sliding vane rotary vacuum pump [21]; b) Roots vacuum pump [23]; c) Getter ion pump [27]; d) Turbomolecular pump [25]; e) Cryopump [24]; f) Diffusion pump [20]; g) Diffusion ejector pump [22]; h) Liquid ring vacuum pump [26]

some form within the body of the pump itself. They use the gas bonding action of getter materials, including the use of electrical fields. Curve e of Figure 8 shows that pumping speed remains almost constant even at low pressures. Or, gas bonding takes place by cryocondensation and cryoadsorption at low temperatures. The pumping speed of these pumps achieves almost the theoretical value specified by Equation (16). The most prominent application for entrapment pumps is for ultra high vacuum, although modern cryopumps are increasingly applied in the medium vacuum range. While Figure 8 shows the pumping speed trend for a typical representative of the various pump types, Figure 10 shows the throughput of vacuum pumps, that is, the delivered gas volume per unit time referred to atmospheric pressure. Figure 11 gives an overview of the principal vacuum application ranges for various types of pumps.

2.3. Technology of Vacuum Pumps

2.3.1. Displacement Pumps

The reciprocating piston and membrane pumps are oscillating positive displacement pumps. These push gas through the pump by back and forth movement of the pistons or membranes. Membrane pumps are used for light services in the laboratory. Reciprocating piston units still find application in the chemical industry as pumps for rough vacuum. In comparison, rotary positive displacement pumps, which are treated in the following paragraphs, have great importance. In these, rotating internal parts periodically produce the displacement volume.

2.3.1.1. Liquid and Gas Ring Vacuum Pumps

The liquid ring vaccum pump is a simply constructed rotary positive displacement pump. A rotor provided with buckets is mounted eccentrically within a stationary housing filled with a working fluid. Its rotational motion produces a rotating liquid ring within the housing (Fig. 12 A). The liquid ring creates a sickle-shaped displacement volume, which is partitioned by the buckets into separate cells. The eccentricity of the rotor within the housing causes oscillation of the liquid ring within the cells. In turn, these oscillations result in compression of the gas to atmospheric pressure through correspondingly-shaped suction and discharge openings in the front and back walls of the housing. Adding working fluid through an additional

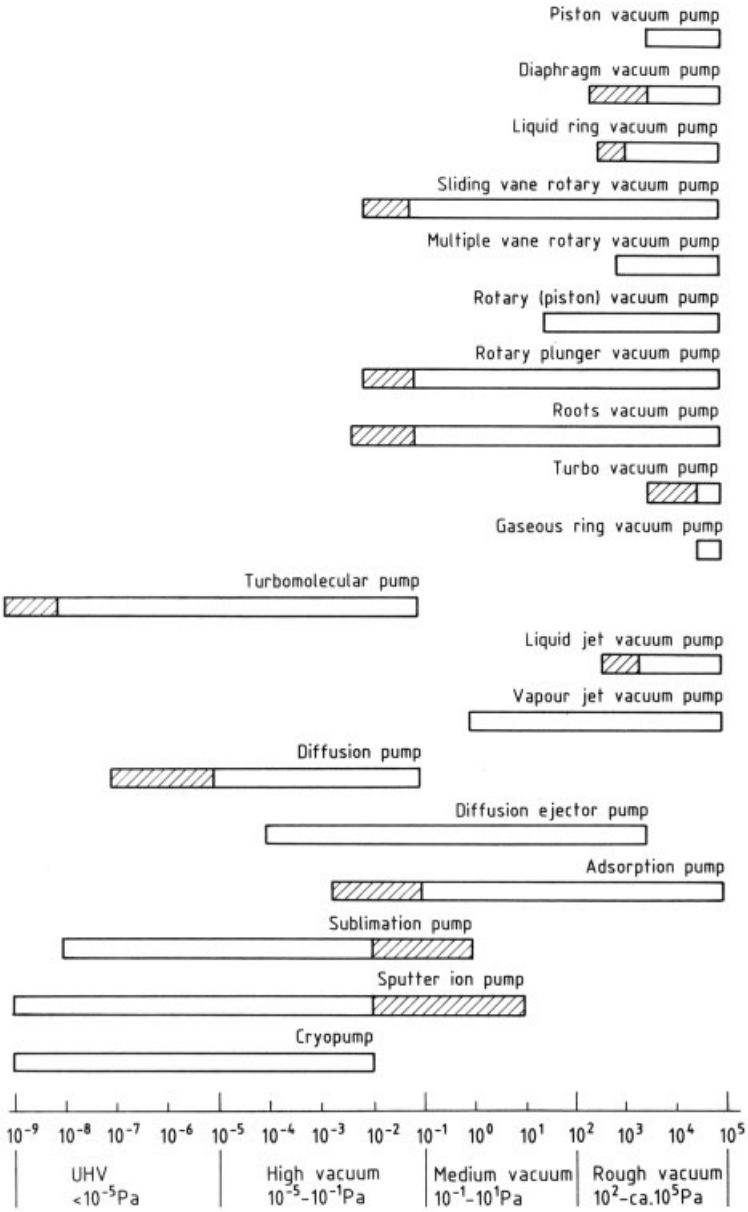

Figure 11. Vacuum ranges for pumping units [28]
Range of operation with special design or with particular operating data.

opening compensates for losses caused by combination with the gas exhaust.

With water as the working fluid (the usual case), these pumps can achieve an ultimate vacuum <30 mbar depending on the water temperature. If the gas being pumped contains water vapor, then the liquid ring pump acts simultaneously as a condenser, and its pumping speed rises above the value for positive displacement pumps. The pumps can handle up to 5 % of the volume resulting from the pumping speed in the form of dust or liquid droplets without difficulty [29]. The pumping operation effectively cools hot process gases. The simple construction permits the use of corrosion-resistant materials, for

example, ceramics. It also leads to high reliability and low maintenance cost.

A very similar pump in physical design and range of application is the gas ring vacuum pump. It consists of a turbine wheel provided with a multitude of buckets which accelerate the gas from the side, outwards into a toroidal-shaped channel leading around the pump. In this manner, a compression takes place to atmospheric pressure, or beyond. The gas ring pump, according to the physical principle involved, does not belong to the positive displacement pump group, but to the kinetic vacuum pumps treated later.

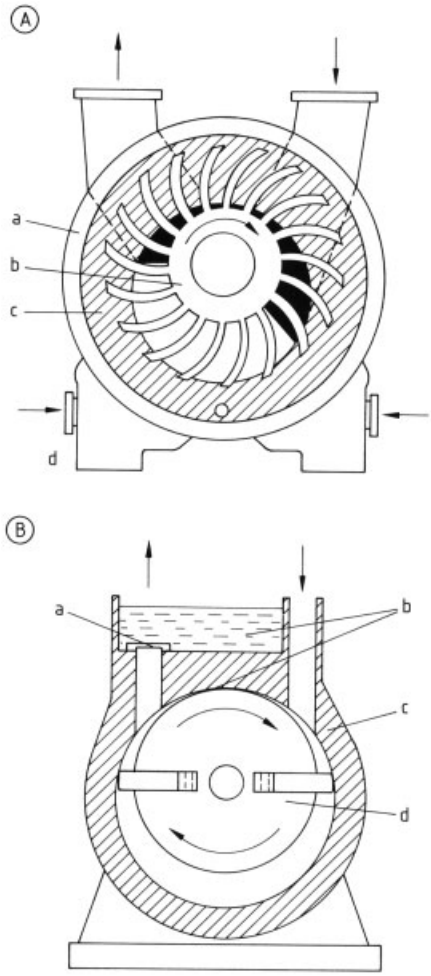

Figure 12. Rough vacuum and foreline pumps
A) Liquid ring vacuum pump: a) Stator; b) Rotor; c) Liquid ring; d) Liquid inlet
B) Sliding vane rotary vacuum pump: a) Outlet valve; b) Oil; c) Stator; d) Rotor

2.3.1.2. Sliding Vane Rotary, Rotary Plunger, Rotary Piston Pumps

Sliding vane rotary, rotary plunger, and rotary piston pumps are for the most part oil-sealed rotary positive displacement pumps that work with a rotating piston principle. The sliding vane vacuum pump (Fig. 12 B) has a rotor placed eccentrically within a stator. Movable sliding vanes (generally two) are fastened in a slot in the rotor. Forced outward by springs, these vanes slide on the stator inner wall, thereby producing reciprocating displacement volumes that act to compress the gas. The gas discharges to atmospheric pressure through an exhaust valve. Sliding vane vacuum pumps may be built with one pump stage or with two stages connected in series. The achievable ultimate pressure is ca. 0.01 mbar with a single stage and < 0.001 mbar with a two-stage series design. The achievable ultimate pressure depends on the partial pressure of the pump seal oil and the quantity of gas dissolved therein. The pumping speed of commercially available units ranges from 0.2 to $> 500\,m^3/h$.

The high compression ratio of the sliding vane pump may lead to condensation of the vapor contained in the inlet gas within the pump. As a result, pump capacity no longer suffices and the incompressible liquid may cause serious damage or corrosion. A remedy, up to a certain vapor concentration, is gas ballasting: supplementary air or inert gas is admitted into the pump. This reduces the partial pressure of the vapor in the gas at the discharge pressure to the point at which no condensation takes place. A measure of the allowable vapor in the inlet gas is the water vapor tolerance. For jacketed pumps having optimal construction and uniform heating of the pump internal volume, the allowable water vapor partial pressure lies at 70 mbar in the inlet gas.

Sliding vane rotary pumps are ideally suited for clean and dry vacuum operations and for small to medium inlet capacities in the range of atmospheric pressure down to medium vacuum. Contaminated oil fog in exhaust gas can be removed by separation. The amount of waste oil formed is usually small. If large amounts of condensable vapors or corrosive gases and dusts are pumped out with the gas, then a separator must

be installed. The pumping oil must then be purified in an external filtration loop. The residual gas from a sliding vane pump contains traces of oil vapors and cracked products from the pump seal oil.

The chief application for sliding vane rotary pumps in a vacuum system is as the high-pressure pumping stage for operation in the medium vacuum range. Various special designs are commercially available:

1) To overcome corrosion problems, continuous fresh oil feed may be used instead of a permanent oil charge
2) Tightly sealed pumps suitable for holding helium are available for the nuclear power industry
3) Dry-running sliding vane pumps avoid the possibility of oil vapors entering the vacuum vessel from the pump

Related to the sliding vane are the rotary plunger and the piston rotary vacuum pumps. In the former, a piston rotates eccentrically in a stator housing. A stationary sliding seal plate coupled with the piston serves to disengage the pump inlet from the discharge opening, allowing gas compression by the rotor. In the latter, the piston has a trochoidal cross section whose center of gravity produces a circular path as the pump moves. Both types of pump, being oil-sealed, have similar features to sliding vane pumps. They are applied as rugged rough-vacuum pumps for industrial processes, mostly with pumping speeds $>500\,m^3/h$.

2.3.1.3. Rotary Lobe Vacuum Pumps

In rotary lobe vacuum pumps, or *Roots pumps*, the pump housing contains two figure-eight-shaped rotors. These revolve counter to one another, turning with little clearance between them (Fig. 13 A). The compression ratio is always less than 100, therefore they can also pump gases with high concentrations of condensable vapors. Apart from rate of rotation and displacement volume, gas backflow through the clearance between rotors and walls, and gas left in the clearance volume also determine the pumping speed. Because of gas flow losses, the pumping speed depends on the type of gas (Eqs. (13) and (16)). The power requirement of a Roots pump is proportional to the pressure difference between discharge and suction pressure. Therefore, without incurring excessive temperatures, they can only be operated up to a maximum pressure difference of the order of 100 mbar. The necessary forepump must be chosen to assure the specified quantity of gas is pumped while maintaining the specified working vacuum. The pump operation against atmospheric pressure is also possible with a gas-cooled rotary lobe pump: cooled gas at atmospheric pressure, from an aftercooler, is mixed with the suction inlet gas between the suction and discharge pump stages.

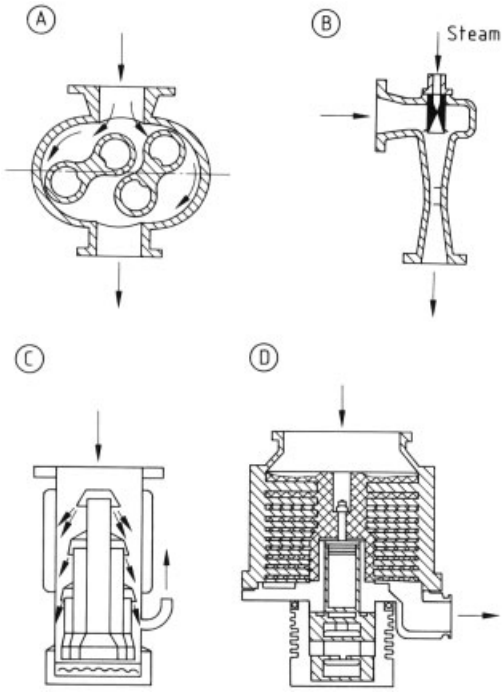

Figure 13. Medium and high vacuum pumps
A) Roots vacuum pump; B) Working fluid jet vacuum pump; C) Diffusion pump; D) Turbomolecular pump

Rotary lobe vacuum pumps in combination with suitable forepumps, correctly designed, achieve ultimate vacuums under 10^{-4} mbar. They are manufactured with pumping speeds from 150 to $>25\,000\,m^3/h$. They are applied universally in the rough and medium vacuum regimes, either as the low pressure stage in series with a pump discharging at atmospheric pressure or as an intermediate pump stage, for example, after a high vacuum diffusion pump. Since it

has no rubbing parts, the Roots pump is rugged and wear-resistant. Lubricating the bearings and drive requires oil, but, because of the low consumption, expensive oils with low vapor pressure can be used. A Roots pump can be driven by an electric motor enclosed within the pump housing, avoiding an opening for the shaft and the seal that goes with it. Thus it can be made completely vacuum-tight, minimizing the potential for escape into the surrounding air of, for example, radioactive or toxic gases.

2.3.1.4. Pump Oils for Rough and Medium Vacuum Pumps

Apart from some dry sealed special constructions, rotary vacuum pumps use oil for lubrication and sealing. Besides lubricity, the oils must exhibit the following characteristics:

1) Low vapor pressure, $\leq 10^{-3}$ mbar
2) Sufficient chemical inertness at the operating temperature, especially oxidation resistance
3) Low solubility for the gases being pumped

Apart from narrow mineral oil fractions, saturated paraffins, and mineral oils containing antioxidants, synthetic silicone oils, ethers, esters, and fluorocarbon compounds find use, especially under more severe chemical conditions [5]. The vapor pressure of commercial forepump oils at room temperature is less than 10^{-4} mbar, and the average molecular weight is mostly greater than 400. For fluorocarbon compounds, molecular weight is greater than 1000; the flash point is higher than 220 °C, or the compounds are not flammable. Synthetic pump oils cost up to one hundred times as much as simple mineral oils. Under certain circumstances, however, reprocessing expensive used pump oils can bring substantial cost savings [30].

2.3.2. Kinetic Vacuum Pumps

Kinetic vacuum pumps are gas-transfer pumps whose pumping action relies on the transfer of momentum between the gas and a liquid working fluid, or between gas streams having different velocities (Eq. (14)) or between the gas and a solid surface in some form (Eq. (16)). As a consequence, one differentiates between working fluid and mechanically acting kinetic vacuum pumps.

2.3.2.1. Working Fluid Jet Vacuum Pumps

Pumps operated with working fluids in the regions of viscous and intermediate flow are called working fluid jet vacuum pumps. The working fluid flows through a nozzle into a mixing chamber. The reduced static pressure (Bernoulli equation) forces gas from the suction into the mixing chamber. The working fluid entrains the suction gas into a diverging nozzle where, together, they are compressed and discharged (Fig. 13 B).

The best known example is the water jet pump. Its poor ultimate vacuum, ca. 20 mbar, its low pumping speed, mostly <1 m^3/h, and its high water consumption limits its use to simple laboratory applications. Gas jet or vapor jet vacuum pumps use gas or vapor working fluids, the latter usually steam. After passage through the diverging nozzle, the inlet gas mixed with the working fluid vapor passes into a condenser, which serves to separate working fluid and pumped gas. A forepump transfers the separated gas to the next, higher pressure stage. If desired, the condensed working fluid may be returned to the pumping operation through a closed recycle loop.

Vapor jet vacuum pumps have a maximum compression ratio of ca. 20. Therefore, to achieve an ultimate vacuum <1 mbar requires connecting several in series. To avoid pumping the working fluid vapor, and possibly suction inlet vapors, through all the pump stages, condensers are usually installed between the pump stages. The most economical operation requires careful consideration of the coordination of the condensers, the cooling temperature in the condensers, the pumping speeds, and the working fluid vapor pressures. Vapor jet vacuum pumps are commercially available in sizes from small units suitable for the laboratory to large installations with pumping speeds $>10^6$ m^3/h at 0.5 mbar inlet pressure.

Especially when equipped with intercondensers, operation is completely unaffected by high levels of dirt in the inlet suction gas. The low investment cost for the pump offsets the disadvantage of high working fluid consumption and cost. Due to the simple design, almost any material can be used to construct the pump. Therefore, vapor jet pumps can also be used in the chemical process industries to pump the most aggressive media.

2.3.2.2. Diffusion Pumps

Diffusion pumps [31] (Fig. 13 C) consist of three to four stages of Laval nozzles arranged around a central standpipe within an outer pipe (air-cooled in small, water-cooled in large units). The working fluid in a reservoir is heated, vaporized, and flows through the standpipe and issues from the nozzles at supersonic velocity. The gases being pumped diffuse into the vapor jets which are directed downward and against the outer pipe and thus, through momentum transfer, a downward velocity component is created. The working fluid condenses on the outer wall and flows back into its reservoir. Each stage develops a compression ratio between 100 and 1000. Compressed gases at pressure p_2 flow through the pump discharge nozzle into the suction nozzle of the forepump.

The working fluids are a wide variety of high vacuum oils, today only rarely mercury. Pumping speed, oil vapor backflow, and forepumping stability determine operational behavior. Velocity components of the working vapor molecules that are directed away from the jet, especially near the nozzles, result in oil vapor backflow into the chamber being pumped out. Also, the gas molecules being pumped may again separate from the working vapor jet in the direction of the receiver. That is why the actual pumping speed for air amounts to only 60 % and for hydrogen only 30 % [31] of the theoretical values given by Equation (16).

To achieve the ultimate vacuum of $<10^{-9}$ mbar, special condensation devices serve to avoid oil vapor backflow. Cold caps above the uppermost row of nozzles and baffles in the suction opening are appropriately-shaped water-cooled plates on which the oil vapor almost completely condenses. Further reduction of oil vapors results with liquid-nitrogen-filled cold traps. These devices, as well as the necessary high vacuum valve on the suction side, reduce the pumping speed of diffusion pumps by about 10 – 50 %.

Oil diffusion pumps are commercially available with pump speeds from 150 to 180 000 m^3/h at 10^{-5} mbar suction pressure. In the medium vacuum field, ejector diffusion pumps, which combine a diffusion pump with a vapor-jet pump stage, have even greater pump speeds. The forevacuum stability provides information about the maximum allowable pressure p_2 on the forepump suction side. For the usual oil diffusion pumps, p_2(max) is ca. 0.5 mbar. Diffusion pumps are typically applied in the high vacuum field. They are less suitable for the UHV region because of oil vapor backflow. In the medium vacuum regime, they are limited by low pumping speed. Since they have no moving parts, they are extremely reliable in operation. With suitable working fluids, they can also pump chemically aggressive gases without problems. By itself, dust, which would cause damage to mechanical pumps, does not adversely affect the functional effectiveness. Nevertheless, the importance of oil diffusion pumps has declined in favor of the much more expensive, but also much more effective turbomolecular and cryopumps. The latter may be applied over a broader range of vacuum and do not show the disadvantages of oil backflow and low back pressure stability.

2.3.2.3. Jet Pumps–Working Fluids [5]

The lubricating properties of the working fluid are unimportant for vapor jet and diffusion pumps. The following points require consideration:

1) The working fluids should have a low vapor pressure
2) The working fluids should not be subject to thermal decomposition at the operating temperatures, ca. 250 °C
3) The working fluids must vaporize easily
4) The gases being pumped should have low solubility in the fluid
5) The working fluids should be chemically inert to the gases being pumped

Since no liquid in itself exhibits all of these properties, the correct choice among the possible working fluids depends on the specific application. Table 6 lists some typical working fluids and their characteristics. Mineral oils are inexpensive; however, they do not have a low vapor pressure, are easily oxidized by oxygen, and are subjected to thermal decomposition. The silicone oils possess the lowest vapor pressure and are very resistant to oxidation. Various ethers have low vapor pressure, and they are resistant to corrosive gases. Fluorocarbon compounds are the most stable.

Table 6. Characteristics of some typical diffusion-pump fluids

Trade name	Compound	$\bar{M}_r$	p (25 °C), mbar	p_{end}, mbar	Reference
Apiezon C	mineral oil	479	1×10^{-8}	1.3×10^{-7}	[32]
DC 705	trisiloxane	546	4×10^{-10}	6×10^{-10}	[33]
Diffilen ultra	mineral oil	530	3.5×10^{-9}	1.3×10^{-9}	[33]
Fomblin 25/9	perfluorpolyether	3300	7×10^{-9}	$<2.5\times10^{-8}$	[35]
Kryotox 1625	perfluorpolyether	3700	9×10^{-10}		[36]
Santovac 5	pentaphenylether	447	1×10^{-10}	$<1\times10^{-8}$	[34]
Ultralen	polyphenylether	454	1.8×10^{-9}		[33]

2.3.2.4. Drag Vacuum Pumps–Turbomolecular Pump

Mechanical kinetic vacuum pumps include the gaseous ring vacuum pumps previously discussed, which are applied in the field of viscous flow, and liquid ring pumps and turbomolecular pumps (Fig. 13 D), which are used for pumping in the molecular flow region. Turbomolecular pumps are multistage units, each stage consisting of a set of rotor turbine blades and a set of stator blades stationed between each rotor wheel. Each blade represents an elemental unit that acts as a pump through transfer of momentum between the moving and the inwardly diffusing gas [37]. Hence, the turbomolecular pump represents an assembly of individual pump stages arranged in series and an assembly of many elemental pumps connected in parallel.

The pump speed is independent of the nature of the gas in the molecular flow regime. It does depend on the blade geometry and the circumferential velocity of the rotor. To achieve rotor velocities in the range of molecular kinetic velocities requires using rotational speeds up to 90 000 rpm. The efficiency decreases with increasing suction pressure and approach of the mean free path length to the pump dimensions. Because of the clearances between stator and rotor, limited by operational and manufacturing technical considerations to up to 1 mm, gas backflow in a pump stage is substantial and dependent on the nature of the gas.

The compression ratio, τ, is proportional to the square root of the molecular mass (Fig. 14). While τ is ca. 4×10^3 for hydrogen, the value for a gas with M_r 86 is 10^{15}. Therefore the residual gas above the pump consists essentially of hydrogen and oil vapors are not detected. With suitable forepumps, final pressures in the range of 10^{-11} mbar are possible.

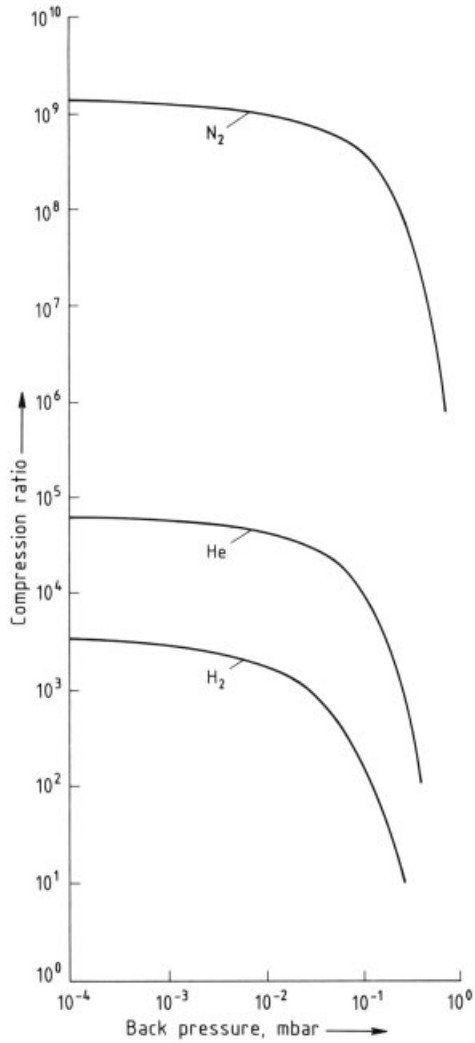

Figure 14. Compression curves of a turbomolecular pump [38]

Turbomolecular pumps can be started up at atmospheric pressure and also tolerate air break-

ins at full speed. Thus, they also do not require a high vacuum suction valve and can be flanged directly to the vacuum receiver.

Commercially available types of turbomolecular pumps have pump speeds between 100 and 32 000 m^3/h in the high vacuum region. They are all-purpose high vacuum and UHV pumps for producing hydrocarbon-free vacua. Their response to big upsets in gas flow is less sensitive than that for diffusion pumps. With special design, they are also suitable for pumping corrosive or radioactive gases. They are largely maintenance-free. Problems may arise in UHV applications due to the molecular-mass-dependent compression. Due to the high rotational speeds and the restricted mechanical tolerances, inflow of foreign particles may destroy the pump. There are two design configurations: double flow pumps with two pump rotor-stator sets horizontally supported symmetrical to the suction nozzles and single flow pumps with the pump shaft oriented in the direction of the suction nozzle. The single flow pump has been gaining increasing acceptance due to its lower space requirement and the possibility, at least for smaller pumps, of assembly at any site.

2.3.3. Entrapment Vacuum Pumps

In contrast to gas transfer pumps, the pumping action of the gas-bonding vacuum pumps is based on sorption and condensation of the gases being pumped within the body of the pumps. Thus, it is necessary to operate gas-bonding pumps cyclically. A regeneration operation, which restores the full gas-bonding capacity, follows each onstream pumping cycle.

2.3.3.1. Sorption Vacuum Pumps

Sorption vacuum pumps use the most varied potential physical and chemical sorption operations. Adsorption vacuum pumps consist of a closed, stainless steel vessel filled with a Zeolite. It is equipped not only for heating, but also for cooling with liquid nitrogen. By means of physical adsorption, the cooled sorbent easily condenses most gases, except neon and helium, and some hydrogen in the region from atmospheric pressure to ca. 0.01 mbar. Heating to 200 °C regenerates the pump. As forepumps, adsorption pumps produce completely hydrocarbon-free vacuums for UHV experiments.

Sorption getters are materials that bond gases into the solid state. They are used in sealed component parts like cathode ray tubes or vacuum switches to pump desorbed gases or gases flowing in through permeation. In addition, very special applications in mass spectrometers, space flight missions, in storage and accelerator rings, and for gas purification are well-known (NEG pumps, non-evaporable getters [16], [39]).

The use of the vaporizing getter, titanium, is more widespread. In the titanium sublimation pump, titanium from a hot coil or a sphere heated from within is vaporized, continuously or under pressure control, onto the surrounding surfaces. The freshly deposited film very effectively bonds all gases other than the noble gases. The pump speed depends on the nature of the gas, the pump configuration, the cooling of the pump surfaces, and the titanium evaporation rate. With liquid nitrogen cooling, it can approach the theoretical value for hydrogen and nitrogen (Eq. 16). Titanium sublimation pumps are supplemental auxiliary pumps of high pumping speed for the pressure range from 0.01 to $<10^{-8}$ mbar. They are used exclusively in UHV systems.

2.3.3.2. Getter Ion Pumps

Several getter ion pump designs are available. Units using the diode principle consist of flat titanium cathodes fastened on both sides of the opening of a tubular anode. Others using the triode principle combine flat anodes with cathodes in the form of grids behind which are collector electrodes at the anode potential. In both, a gas discharge is ignited between anodes and cathodes, intensified by magnetic fields from powerful permanent magnets and maintained even at the lowest pressures. The ions of the gas discharge atomize the titanium of the cathodes which, on the one hand, reacts with the gases forming compounds (getter effect), and on the other hand, also entrains gas particles mechanically (pumping the noble gases). Especially in the case of diode pumps, the overall pumping characteristic is dependent on the previous history; above all, entrained argon may again be-

come freed in a subsequent pumping operation (memory effect).

The getter ion pump is a UHV pump for ultimate vacuums in the range 10^{-11} mbar. Its pump speed depends on the pressure and on the nature of the gas. For air, single units are available with capacities of the order of 2000 m^3/h at 10^{-6} mbar. Frequently, titanium getter ion pumps and titanium sublimation pumps are combined in one unit, the so-called UHV booster.

2.3.3.3. Condensation Pumps

Freezing out vapors and gases being pumped onto sufficiently cooled surfaces is the basis for condensation pumps. The pump speed depends not only on the gas properties, the pump surfaces, and their temperature, but also on the degree of coverage of the surface by the gas and the effectiveness of screening against thermal radiation from the pump surroundings. When loaded with condensed gas, the pump is regenerated by warming.

Condensation pumps operating with cooling media temperatures above 120 K are called vapor condensers [27]. Piping, plates, and fins for the cooling medium are designed for a specific system to remove large quantities of water and solvent vapors from the process atmosphere. Vapor condensers are used in industrial processes in the range 0.2 – 0.01 mbar [40]. The pump speed for the case of water on a clean cooled surface at temperatures below −50 °C is close to theoretical. However, pump speed decreases with increasing loading of the cooling surface due to the poor thermal conductivity of the condensate. The value of the ice capacity characterizes the vapor condenser. The ice capacity designates the mass of ice which can be frozen out per day with a certain coating thickness on the whole vapor condenser. The pump operation is meant to be effective at the selected coating thickness.

Condensation pumps with temperatures <120 K are called cryopumps [41]. These are able to pump gases that are noncondensable at room temperature and if activated carbon is present even hydrogen, neon, and helium. The gas-bonding occurs by cryocondensation when the gas, at operating temperature and pressure, can be transferred into the solid or liquid phase; by cryotrapping when noncondensable gases are stored in the growing layer of condensate; and by adsorption on surfaces coated with adsorbent, usually activated carbon. In general, cryopumps can be operated so that the pump speed is equal to the maximum possible pump speed (11.6 L/s for air, 14.5 L/s for hydrogen, Eq. 16). For pump openings of equal size, cryopumps have the highest pumping speed of all vacuum pumps. They can achieve final pressures $<10^{-11}$ mbar.

The cooling medium in most cryopumps is helium. Cryopumps are differentiated according to the style of arrangement of cooling devices into bath cryopumps, vaporizer cryopumps, and refrigerator cryopumps. Bath cryopumps are vessels filled with boiling liquid helium. Therefore, the cooling medium temperature is 4.2 K at atmospheric pressure or even 2.5 K at reduced pressure. The cooling surfaces must be protected from thermal radiation by liquid-nitrogen-cooled shields. The arrangement of the pump is fixed by the need for charging with liquid helium.

In vaporizer cryopumps, vaporizing helium is pumped from cooling coils. Varying the flow rate of helium controls the cooling temperature. The pump manages without liquid nitrogen since the helium pumped out serves to cool the thermal-radiation shields. There is no size limitation for vaporizing cryopumps.

Refrigerator cryopumps are usually two-stage vaporizer cryopumps with a helium refrigerator creating a closed helium recycle loop. The second stage cools the "cold head" operating between 10 and 20 K. The cold head is partially covered with activated carbon. The first stage cools the radiation shield around the cold head and the shields often used around the vacuum chamber to 70 – 100 K. Easily condensable gases, like water vapor, deposit on these outer pump stages. Argon condenses in the second stage. Hydrogen and helium bond to the activated carbon of the cold head by cryotrapping and cryoadsorption.

Formerly, cryopumps had been used only as high vacuum and UHV pumps due to their limited capacity for gas uptake. However, newer designs also permit their application in reactive vacuum processes such as thin film coating by cathode sputtering. This is done by partially throttling the pumping speed for process gases while maintaining full speed for water. Here, cryopumps have entered suc-

cessfully into competition with turbomolecular pumps. Standard sizes have pumping speeds from 1200 – 180 000 m^3/h. Special designs for space simulation chambers or for fusion experiments reach 1.8×10^8 m^3/h [42]. Pump speeds of this large order of magnitude can only be realized economically with cryopumps. For lower pump speeds in the high vacuum region, cryopumps should be compared with others, for example, oil diffusion pumps. The expense of the consumption of working fluid and cooling media as well as the problems associated with waste disposal might soon exceed the higher investment cost of a system with cryopumps.

3. Vacuum Diagnostics

One measure of the state of the gas in a vacuum unit is the total pressure, the sum of all partial pressures of the individual gases. Others of great interest are the gas composition and the partial pressures. Just as in the case of producing a vacuum, in regard to vacuum diagnostics, no one single measurement principle suffices to describe the whole range of particle densities understood under the term vacuum.

3.1. Total Pressure Measurement

In principle, measurement of total pressure may be handled in accordance with the basic fundamentals expressed by Equation (3), directly measuring the quantity *force/area*. Beyond that, a multitude of pressure-dependent physical quantities are well suited to indirect measurement. However, many of these effects are dependent on the nature of the gas [43]. The instruments are then calibrated for air or nitrogen; for other gases, the pressure must be determined by use of correction factors. Important points to consider in applying instruments for total pressure measurement are the measurement range, sensitivity, signal accuracy, and reproducibility. The measurement range Δp and sensitivity ("pointer movement"/δp) may depend on the nature of the gas. Accuracy and reproducibility are inherent aspects of each instrument.

Vacuum meters may measure either the absolute pressure directly or a differential pressure referred to a preset vacuum. Certain measurement principles are customarily used in the case of each of the individual pressure regions (Fig. 15). Rough and medium vacuum, and the viscous flow regime, are the domain of pressure measurement based on mechanical phenomena. In the medium and high vacuum regimes, the transition region between viscous and molecular flow, pressure measurement may rely on transfer of heat and momentum. Gas ionization effects are the measure of total pressure in the HV and UHV regimes, the regions of molecular flow.

3.1.1. Mechanical Vacuum Instruments

Mechanical vacuum gauges directly record the force acting on a membrane or on the surface of a liquid or solid body, $p \cdot A$ (Eq. (3)). In principle, they can also be used as *manometers* in the positive pressure region. Fundamental examples are the U-tube and the compression vacuum gauges. The former measures the difference between the pressures on a liquid in both the legs of a tube bent into the shape of a U. In the latter, a well-known design being the McLeod gauge, mercury columns compress a volume V_1 at pressure p_1 to a known volume V_2. From the now measurable pressure p_2, p_1 may be calculated using Equation (1). The measurement range of accurate compression vacuum gauges extends to 10^{-6} mbar. The measurement error is highly dependent on pressure regime and is $\geq 5\,\%$ of the measured value. U-tube and compression vacuum gauge measurements do not depend on the nature of the gas. However, the compression vacuum gauge is unsuitable for measuring the pressure of gas mixtures containing portions of condensable vapors.

Bourdon tube vacuum gauges represent a group of mechanical instruments based on the principle of a capsule or a hollow spring. They are used in the rough vacuum regime. Pressure-dependent motion of the capsule or spring is transmitted by a mechanical linkage to move the indicator pointer. Units designed to the precision required for purposes of calibration reach an accuracy of 0.3 % of scale value in the vacuum range 0.1 – 1000 mbar.

The use of mechanical-electrical transducers permits digitizing of the signal output, indication remote from the point of measurement, and high accuracy. The physical phenomena include

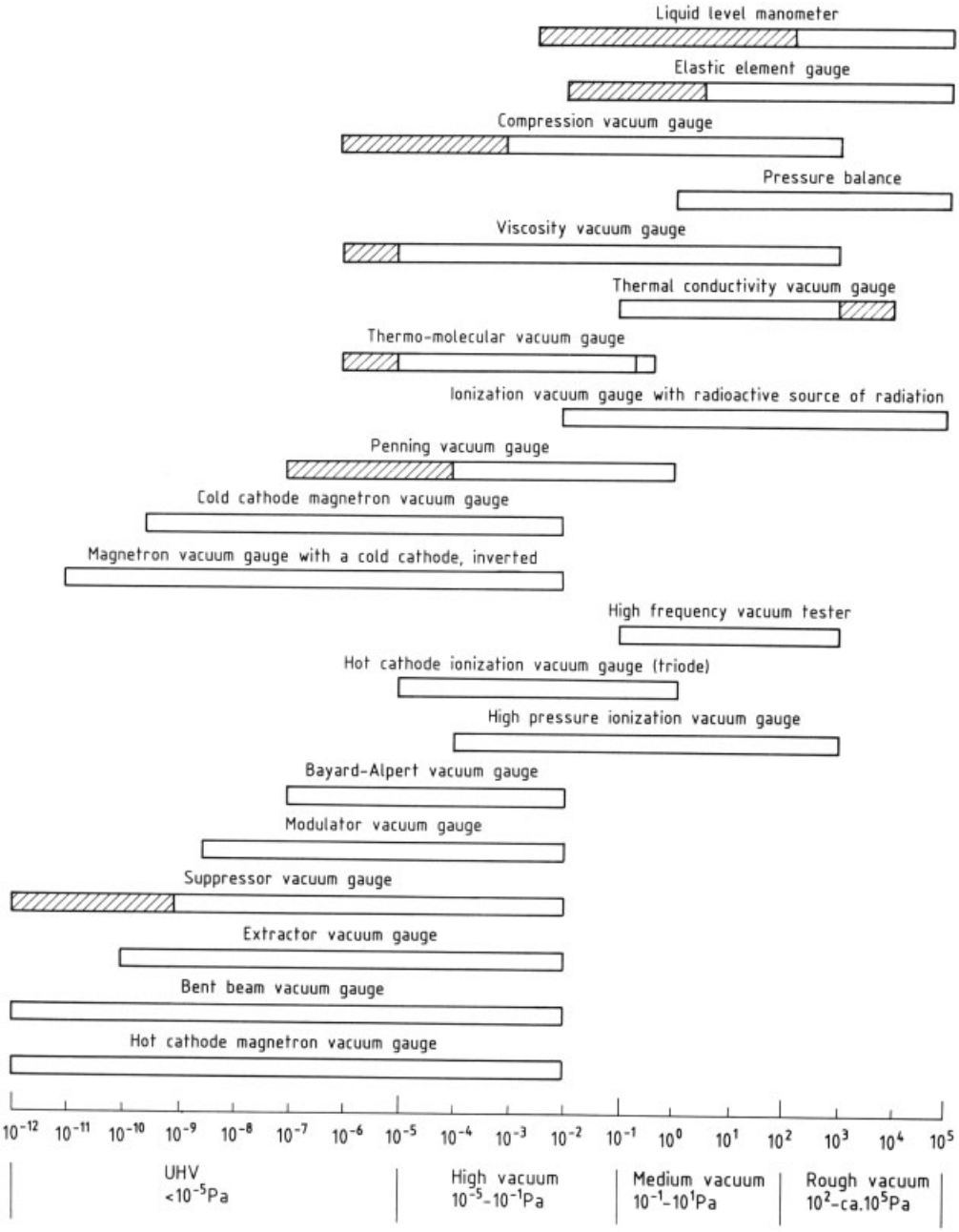

Figure 15. Pressure ranges of vacuum gauges [44]
Range of measurement with special design or with particular operating data

resistance changes in strain gauges, the piezoelectric effect, and measuring the electrical capacitance changes on deforming a membrane. Very sturdy and also corrosion-resistant designs are commercially available as pressure sensors in various sizes for pressures from one mbar to quite high positive pressures. The most accurate mechanical vacuum-measuring instruments are those based on measuring capacitance. They are available for measuring in the range 10^{-5} – 1000 mbar with a pressure-dependent minimum uncertainty of 0.05 % of the measured value.

Control of rough and medium vacuum pressures without high accuracy requirements employs the simple, problem-free, and inexpensive mechanical instruments. Capacitive pressure measurement suffices for high accuracy measurement, independent of the nature of the gas, e.g., for process gas control in the transition range between medium and high vacuum. In addition, capacitance vacuum gauges serve as transfer standards for calibration of measurement instruments in the pressure range 0.001 – 1000 mbar.

3.1.2. Viscosity Vacuum Gauge

The viscosity is pressure-dependent in the region where $d \leq \bar{l}$. The viscosity gauge uses this effect for pressure measurement. The momentum transfer described by the coefficient of external friction on a moving body retards it. Commercial devices consist of a freely-suspended sphere set into rotation in a magnetic field [45]. The instrument measures the pressuredependent decrease in the frequency of rotation due to friction. Figure 16 shows a pressure measuring head. The measurement range for this friction vacuum gauge extends from 10^{-7}–10^{0} mbar. The electromagnetic drive rotating the sphere produces a null effect, which determines the lower limit of the pressure range [46]. At high pressures, the restrictive factor is the change to a viscosity independent of pressure (Eq. (13)). Three factors combine to make this instrument an ideal standard for pressure measurement in the medium and high vacuum regimes (10^{-2}–10^{-6} mbar):

1) Accuracy of 1.5 % of the measured value
2) Long term stability, deviations <1 %/a
3) Calculation of the pressure measured requires only sphere diameter, sphere density, temperature, and frequency, all relatively easy to determine.

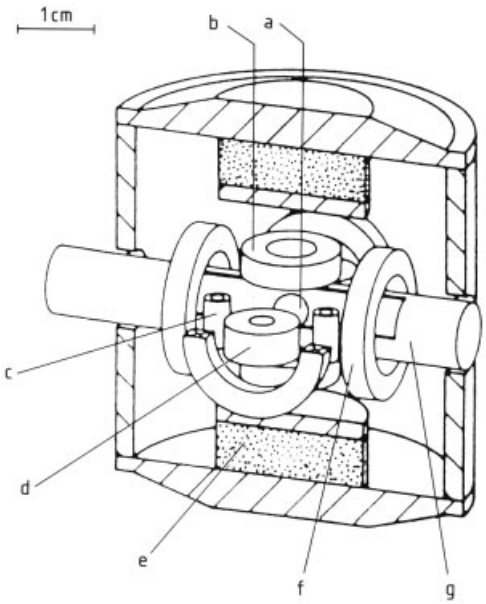

Figure 16. Sensor head of the spinning rotor gauge [45]
a) Rotor; b) Coils for pickup and control of axial rotor position; c) Coils of lateral damping system; d) Pickup coils; e) Permanent magnet; f) Drive coil; g) Vacuum enclosure

Further applications are direct measurement of the flow of gases from gas control instruments or desorbed from surfaces [47].

In the viscous flow region from 10^{0} – 10^{3} mbar, instruments also depending on gas friction are available for pressure measurement [48]. These utilize a second order influence on the resonance frequency of a quartz oscillator.

3.1.3. Thermal Conductivity Vacuum Gauge

A large variety of instruments utilize thermal conductivity to measure vacuum. Thermal conductivity depends on the pressure and nature of the gas. In principle, the rate of withdrawal of energy from an electrically-heated wire serves as the gauge of pressure. In practice, one measures the hot wire temperature thermoelectrically or by means of the resistance. Bridge circuits compensate for temperature deviations from a set value. The compensation serves as the measured variable and is related to pressure (Pirani principle). Utilizing convection allows extending the

range of reliable measurement into the viscous flow regime.

Heat losses from the lead wires and by radiation, which eventually outweigh heat transfer through the gas, result in a lower limit to the vacuum that can be measured. Therefore, thermal conductivity vacuum gauges have a measurement range of 10^{-4} – 10^{3} mbar. The calibration is done with air. The accuracy is ca. 20 % of the indicated value. Thermal conductivity vacuum gauges are simple, reliable, and inexpensive, and thus appropriate if extremely accurate measurements are not required.

3.1.4. Ionization Vacuum Gauge

The ionization vacuum gauges infer the pressure by measuring the stream of gas ions produced by electron collision ionization. Either of two methods are usual for producing the electrons necessary for gas ionization: an independent gas discharge (cold cathode ionization vacuum gauge) or glow emission of a heated wire (hot cathode ionization vacuum gauge). The measurement with either gauge depends on the nature of the gas. These gauges are usually calibrated with nitrogen.

The Penning vacuum gauge is the original, well-known form of the cold cathode vacuum gauge. It consists of two cathode sheets with a ring-shaped anode lying in between. A magnetic field in the direction of the electrode arrangement, produced by permanent magnets, amplifies the high voltage gas discharge by prolonging the electron paths. The discharge current is pressure-dependent over a broad range. Measuring tubes are available for the range 10^{-2} – 10^{-10} mbar. The measured value may be inaccurate by about a factor of two. However, this is scarcely a concern if the use of the Penning tube is limited to rough survey measurements. This kind of application is favored also by its extraordinary sturdiness and resistance to damage by sudden air break-in.

The underlying principle of the Penning gauge is the same as that for the ion getter pump. Just as the discharge current of the ion getter pump can serve as a pressure indicator, the Penning tube possesses a significant pump speed for gases, a possible cause for error in measurements.

The hot cathode vacuum gauge is built as a triode. A central glowing cathode emits electrons, which are accelerated to a concentrically-arranged anode grid. The positive ions produced by inelastic electron collisions give rise to a current to an ion collector located outside (Fig. 17 A). The ion current is directly proportional to the pressure and to the electron current:

$$i_{+} = C \cdot i_{-} \cdot p$$

The sensitivity C depends on the type of gas (tabulated for the common gases). At high pressures, C shows a maximum: measurement is no longer clear-cut. At low pressures, an ion current liberated by a photoeffect dominates (Roentgen limit). The Bayard – Alpert gauge (Fig. 17 B) shifts the Roentgen limit to lower pressures. This construction is a departure in which the ion collector, a fine wire, is in the center with a glowing cathode outside of the grid anode. Modulation of the ion current or extraction out of the ionization zone are further possibilities for broadening the useful range of measurement. For the conventional triode arrangement, the range is 10^{-1} – 10^{-6} mbar. The lower limit of the Bayard – Alpert gauge is 10^{-10} mbar. Commercial modulation tubes and extractor systems extend the range below 10^{-12} mbar. Special designs permit measurements below 10^{-13} mbar [49].

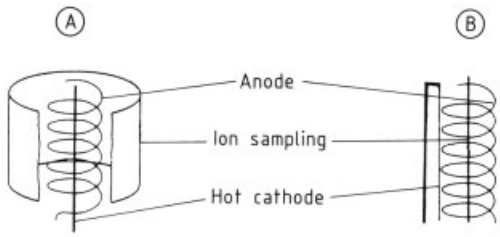

Figure 17. Principles of the A) triode gauge and B) Bayard – Alpert gauge

The potential measurement error of hot cathode vacuum gauges is about 10 % of the indicated value. In contrast, these gauges are sensitive to air break-in or to corrosion by gases. Especially in UHV applications, one must guard against a number of potential upsets [18]. Special ionization vacuum gauges equipped with gold electrodes are extremely stable and reproducible and are used as transfer standards for the range 10^{-4} – 10^{-9} mbar.

3.1.5. Measuring Systems

Frequently, vacuum gauges producing an electrical signal are combined with an electronic unit to create a measuring system. By automatic switching among the several devices, the entire pressure range can be measured. For process control, several pressure-dependent switch points are often available. In addition to the analog output always provided by pressure recording, more complex instruments provide computer-compatible outputs for remote control and readout.

3.2. Partial Pressure Measurement

Detection of the individual gas components and measurement of their partial pressures with a mass spectrometer offers substantially better insight into the nature of a vacuum than the total pressure.

Mass spectrometers consist of the following:

1) A gas ionization device
2) A separation system, which sorts the ions produced, mostly in accord with their charge/mass ratio
3) An ion detection apparatus
4) Control, evaluation, and indicator electronics

The gases are ionized, analogous to the operation of the glowing cathode vacuum gauge, by ionizing collisions with electrons from a glowing cathode. A focused beam of ions issues from the ionization zone and is directed into the analyzer section. For separating the ions, two principles are in widespread use: separation with a magnetic sector field and filtration by an electrical quadrupole field. In units with a magnetic sector, the ion beam is accelerated by the potential difference U, and then traverses a sector-shaped magnetic field perpendicular to the beam direction (Fig. 18 A). By this means, ions with the path radius r are diverted, where:

$$r = \vec{p} \,/\, (B\,q)$$

and

B = magnetic flux density
q = the ionic charge
$\vec{p} = m\ \vec{v}$ = the momentum of an ion

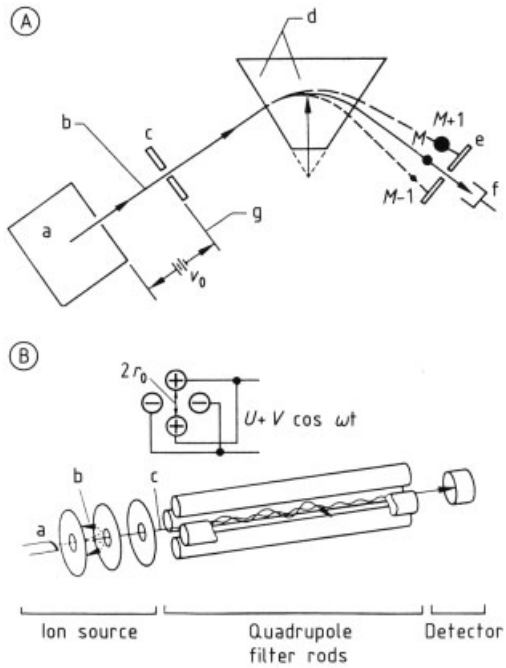

Figure 18. Partial pressure analyzers
A) Magnetic sector type [50]
a) Ion source; b) Beam of positive ions; c) Entrance slit; d) Magnetic field (points out of page); e) Exit slit; f) Detector; g) Ion acceleration region
B) Quadrupole type [51]
a) Cathode; b) Electrons; c) Ion beam

At constant U, the relation:

$$r \sim \sqrt{m/q}$$

is valid. If an ion detector is located at a fixed point at the outlet of the magnetic sector, then, by varying B or U, the gas ions can be sorted according to their ratio m/q and recorded.

In quadrupole mass spectrometers, the separation system consists of four, equally spaced rod-type electrodes (Fig. 18 B). In most cases, cylindrical electrodes approximate the ideal hyperbolic form. Between the respective opposite pairs of rod electrodes, a direct voltage is applied, with a high frequency alternating voltage superimposed. Ions shot axially into the electrode system experience transverse oscillations in the alternating voltage field. Thus, only ions with a certain ratio m/q can leave the filter system without touching the electrode surfaces and being neutralized. One may vary the ratio m/q to

obtain a linear mass scale by holding frequency constant and changing the direct and alternating voltage amplitudes.

The resolving power, $m/\Delta m$, measures the quality of the separation system. According to the convention, two equally high mass peaks are said to be separated if the height of the valley between both the peaks is lower than 10 % of the peak height. With both types of analyzer, this "unit resolution" can be met or exceeded for the range of mass between one and >500 amu (atomic mass unit).

The ions are detected either directly as an ion current by a Faraday collector or are converted into secondary electrons and amplified in a secondary electron multiplier. When high detector sensitivity is required, the arriving ions are counted as individual impulses with the secondary electron multiplier. The detector limits for partial pressure lie at 10^{-10} mbar for Faraday collectors, $<10^{-14}$ mbar with the secondary electron multiplier, and in the range of 10^{-16} mbar for individual counting.

Mass spectrometers with magnetic sectors require only simple electronic provisions. Thus, they achieve high detector stability. Disadvantages are the magnetic field, which can be the cause of breakdown, and the resolving power, which is controlled mechanically by means of the inlet and outlet slits.

More than 95 % of all installed mass spectrometers are now of the quadrupole type. This is a consequence of the broad development of control and readout devices made possible in combination with microelectronics. However, above all, the stability of units equipped with a secondary electron multiplier is only good enough for use as a process controller with great expense for calibration. Thus, the principal fields of application are the analysis of gas mixtures and the diagnosis of vacuum systems. Mass spectra show the ionic current as a function of the mass from the heights of the individual peaks. However, errors may arise: on the one hand through the probability of ionization of the gases in question, and on the other hand through the possible occurrence of several ionized species having equal q_1/m_1 ratios or of ionic fragments having the same charge q and mass m but a different composition. Ionization factors for ions and ionic fragments of gases must be taken from tables or catalogs of spectra [52–54].

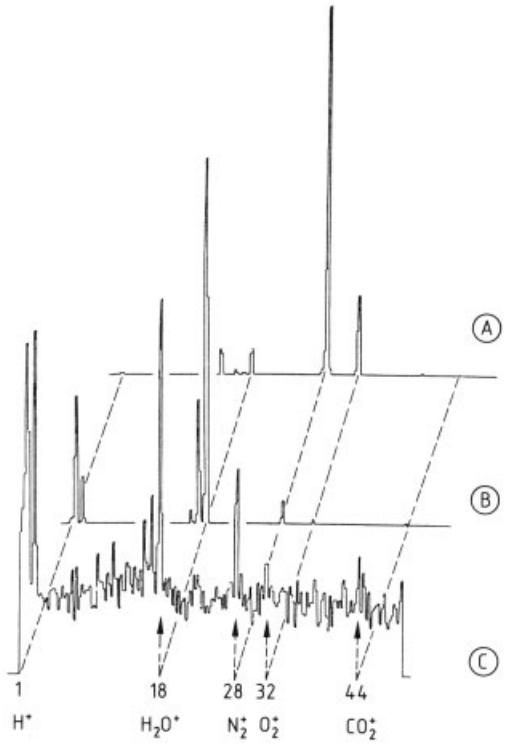

Figure 19. Mass spectra recorded by a simple quadrupole mass spectrometer (different pressure scaling!)
A) Mass spectrum of air at $p = 2\times10^{-4}$ mbar
B) HV-plant, strong water adsorption at the walls $p_{\mathrm{total}} = 3.8\times10^{-5}$ mbar
C) HV-plant, thouroughly degassed, $p_{\mathrm{total}} = 4.5 \times10^{-8}$ mbar

If all these influences are considered, a process of calculation can derive the accurate value of a partial pressure from the mass spectra. Figure 19 shows three typical quadrupole mass spectra:

1) The spectrum of air at 2×10^{-4} mbar, the ratio $p_{(\mathrm{N_2})} : p_{(\mathrm{O_2})} = 5$ is characteristic (Fig. 19 A)
2) The spectrum of an unheated high vacuum system, the water peak dominates (Fig. 19 B)
3) The spectrum of a well-heated high vacuum system, H_2 and CO_2 are represented with heavy lines, the background noise of the spectrometer is already relatively high (Fig. 19 C)

Figure 20 shows the spectrum of a system contaminated with forepump oil. The occurrence of a series of various ion fragments differing the mass of the group "$-CH_2-$" is typical.

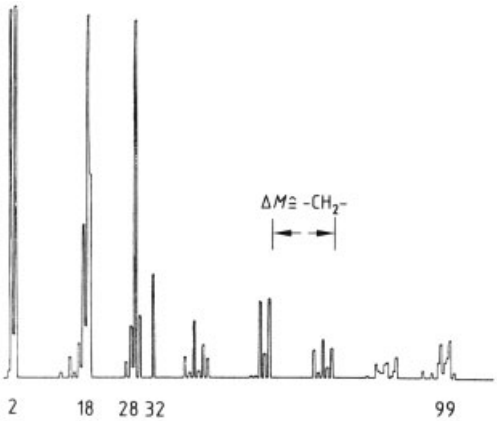

Figure 20. Mass spectrum of a HV-plant contaminated with rough pumping oil, $p_{\mathrm{total}} = 2.4 \times 10^{-5}$ mbar

Modern quadrupole mass spectrometers possess either electronically stored mass spectra of frequently occurring gases or determine the partial pressures using fixed, built-in evaluation programs. Connected to an external computer, mass spectrometers can mostly operate under program control. For evaluating the spectra, extensive programs are available. Simple vacuum diagnosis spectrometers record the masses 2 – 50 (100); the mass scales of routine equipment reach from mass one – 100 (200); research equipment measures out above mass 200.

Mass spectrometers can be applied directly only to pressures $\leq 10^{-4}$ mbar. Adding commercially available pump systems enables the region to and above atmospheric pressure to be attained. In these, restrictions independent of molecular mass produce the necessary pressure drops.

3.3. Leak Surveys

The term *leak* is understood to mean local breaches in the tightness of the vacuum container through which the gas can penetrate from outside into the vacuum. One must distinguish the gas streams caused by desorption or permeation from those resulting from the leak condition. It is not possible to avoid leaks completely. Leaks increase the oxygen partial pressure in the system. One must estimate how much the effect of leaks disturbs the system operation to determine the extent to which it is necessary to remove the leaks. The tolerable leak rates depend above all on the range of vacuum being used. Even when working with process gases in the medium vacuum area, a very low residual gas pressure, when important for the process, may also make necessary the removal of leaks to a degree that would impair the pump-out operation only in high vacuum.

For leak surveys including determining leak rates and locating leaks, various procedures have been developed which work with and without a test gas. In the rough vacuum regime, a direct determination of the leak rate is possible through measuring the gas flow discharged from the pump. One may fill the vessel to be tested with gas to positive pressure or evacuate it. Then, one can calculate the leak rate, q_{L}, through the drop or rise in pressure, Δp, with regard to the vessel volume V by:

$$q_{\mathrm{L}} = V \cdot \Delta p / \Delta t \tag{22}$$

The slope of the plot of pressure versus time permits distinction between desorption (initial high $\Delta p/\Delta t$) and leak rate (constant lower $\Delta p/\Delta t$ after a longer time).

Leak search methods involving test gases take substantially less time, and desorption is not a factor. The unit to be tested is filled with the test gas, traces of which outside the container are used to find the leak location. Or, by placing the unit full of test gas inside a vacuum container, measurement of the overall test gas flow into that container is possible. These methods are useful only with sealed vacuum containers, for example, with relay tubes. Filling can be done by storing the test unit in test gas under positive pressure.

In a second procedure, one can expose the test container, which is under vacuum, to the test gas on the outside either locally, at the place suspected of leaking, or globally, in a vessel enclosing the test container. Special leak-testing instruments detect the test gas flowing into the leak. Halogen leak detector instruments use a heated platinum-covered surface as a sensor emitting alkali-metal ions. The ion current increases greatly due to the effect of the halogen-containing gas; chlorofluorohydrocarbons are normally used. Leak rates $q_{\mathrm{L}} \geq 10^{-6}$ mbar L/s can be detected.

Helium leak detectors are more sensitive. With helium as the test gas, detection of leak

rates to 10^{-11} mbar L/s is possible. Helium leak detectors include a mass spectrometer with the magnetic sector adjusted to a fixed mass of four. Usually, they are also adjustable to the masses three and two. An integrated pumping system supplies the necessary vacuum, lower than 10^{-4} mbar, in the mass spectrometer. The helium leak detector can be flanged directly to the vacuum vessel to be checked. It is also possible to suck gas through a suction duct from the vessel surfaces to the mass spectrometer (sniffing method). Helium leak detectors are thoroughly automated. Self-acting range adjustment, background correction, and devices for programmed control of the components of a vacuum system are state of the art.

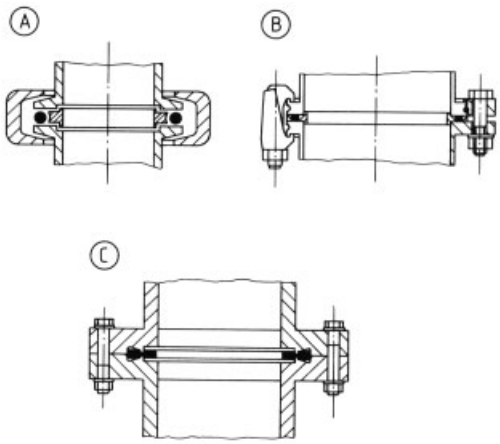

Figure 21. Vacuum flange connections
A) Small flange connection
B) Clamped flange connection
C) CF flange connection

4. Vacuum System Components

To a large extent, vacuum units can be assembled from standardized parts in accord with the principles of modularization. Only the vacuum chamber must be fabricated to suit the particular purpose. The necessary piping, flanges, erection materials, and shut-off devices are available in standard sizes [55]. Transition pieces between particular standard sizes, as well as between different flange systems, are offered in all variations. Therefore, the exchange of parts for the vacuum unit offers no special problems, either within a series from the same manufacturer or from other fabricators.

4.1. Piping Components

Piping components are pipes with both ends flanged, pipe stubs with flanges, crosses, tees, elbows, transition pieces between various standard sizes or between different flange systems and glass-to-metal connections. Spring supports and expansion joints of plastic and metal serve to isolate movement and vibration. Piping components of nominal sizes between 10 and 50 mm are available in glass-fiber-reinforced plastic, aluminum, or stainless steel. The engineering properties for vacuum service improve and the corrosion resistance, mechanical strength, as well as price increase from plastic to aluminum to stainless steel.

Flanges with gaskets between them connect the piping components. There are three common types of flanges, the small flange, the clamped flange and the CF flange (Fig. 21). Joints with small flanges in nominal sizes between 10 and 50 mm are held together by clamp rings or linked rings operable with one hand. They are sealed by centering rings containing round gaskets (“O” rings) made of butyl rubber or fluorocarbon ruber (“ Viton”). For applications requiring corrosion or heat resistance, or for other special purposes, many more gasket materials are available, including silicon rubber, Kalrez, PTFE-coated silicon rubber and nickel-powder-filled PTFE (TENIC). Sealing disks of pure aluminum are also used. With flawless seal surfaces and gasket rings, greasing with vacuum grease is not necessary. Greasing will not achieve greater tightness of the flanged joint.

The leak rate for small flange joints is $<10^{-8}$ mbar L/s, and yet an order of magnitude less with aluminum gaskets. With metal gaskets, small flanges can be heated to 200 °C. The field of application ranges from high vacuum to low positive pressure.

Clamped flange joints in nominal sizes from 65 to 500 mm show engineering properties for vacuum similar to those for small flanges. The joint and the necessary sealing pressure are produced by a number of clamp screws (PNEUROP clamp flanges) suitable for the particular size. Another type uses cap rings bolted together (DIN flanges and PNEUROP cap ring flanges).

CF flanges are used for UHV applications. Rings of pure copper (OHFC copper), or Viton for lesser requirements, serve as gaskets into which the stainless steel knife edges of the flange are pressed. These flanges, in nominal sizes from 16 to 250 mm, have leak rates $<10^{-10}$ mbar L/s, and can be heated to 500 °C. Larger UHV flanges, in nominal sizes to >1000 mm are built similarly; moreover, gasket wires of silver and gold are also employed.

Various clamp ring systems using screwed caps (Fig. 22) are preferred for tubing joints with sizes ≤10 mm. They are manufactured using stainless steel and various plastics. Stainless steel joints of this type are not only suitable for high vacuum but are also pressure-tight to several bar.

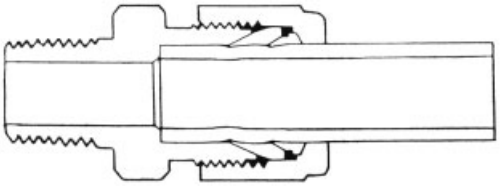

Figure 22. Threaded tube connection

4.2. Vacuum System Feedthroughs

A great variety of vacuum feedthroughs for all vacuum regimes permit control of experiments and manipulations in vacuum containers. These include:

1) Glass view ports, also of UV- and IR-transparent materials for optical measurements
2) Thermocouple feedthroughs
3) Leadthroughs for liquids
4) Electrical feedthroughs for measurement or for load ratings up to 5000 A or up to 20 kV.

Motion feedthroughs capable of turning or linear movement can serve as manipulators or permit built-in drives. They can be applied to the HV region with O-ring seals. They are suitable for the UHV region if the movement is transferred either by means of a spring bellows sealing mechanism or with magnetic liquids. Figure 23 shows an example of the abundance of possible components.

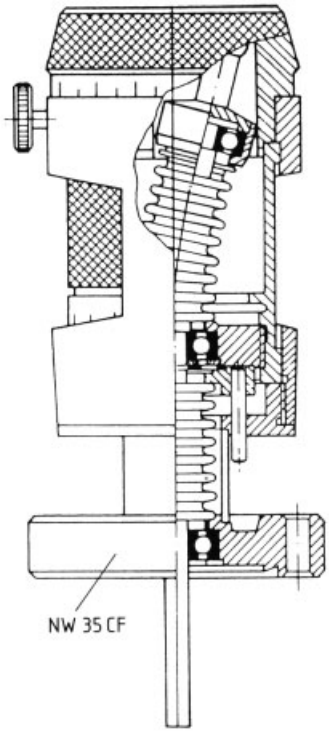

Figure 23. Leadthrough for transmission of rotary and linear motion [56]

4.3. Valves

Figure 24 shows the valves used as shutoff or lock devices between parts of vacuum units for gas inlet, gas control, throttling, and venting services. According to the design of the body, one makes a distinction between throughflow valves, with connecting flanges located on the valve axis, and angle valves, whose flanges form a right angle with one another. Valves of nominal sizes to 160 mm are operated manually; or, magnetic, electric motor, pneumatic, and hydraulic drives are available.

High vacuum valves are sealed with butyl rubber or Viton rings. Valve seats made of metal or ceramic must be used in UHV applications. Spring bellows usually separate the top works from the vacuum side. Springs, the drive mechanism, or the help of atmospheric pressure produce contact pressure on the seat. However, valves are preferred in which both sides of the seat close tight against the atmosphere. The valve function takes place by raising the valve

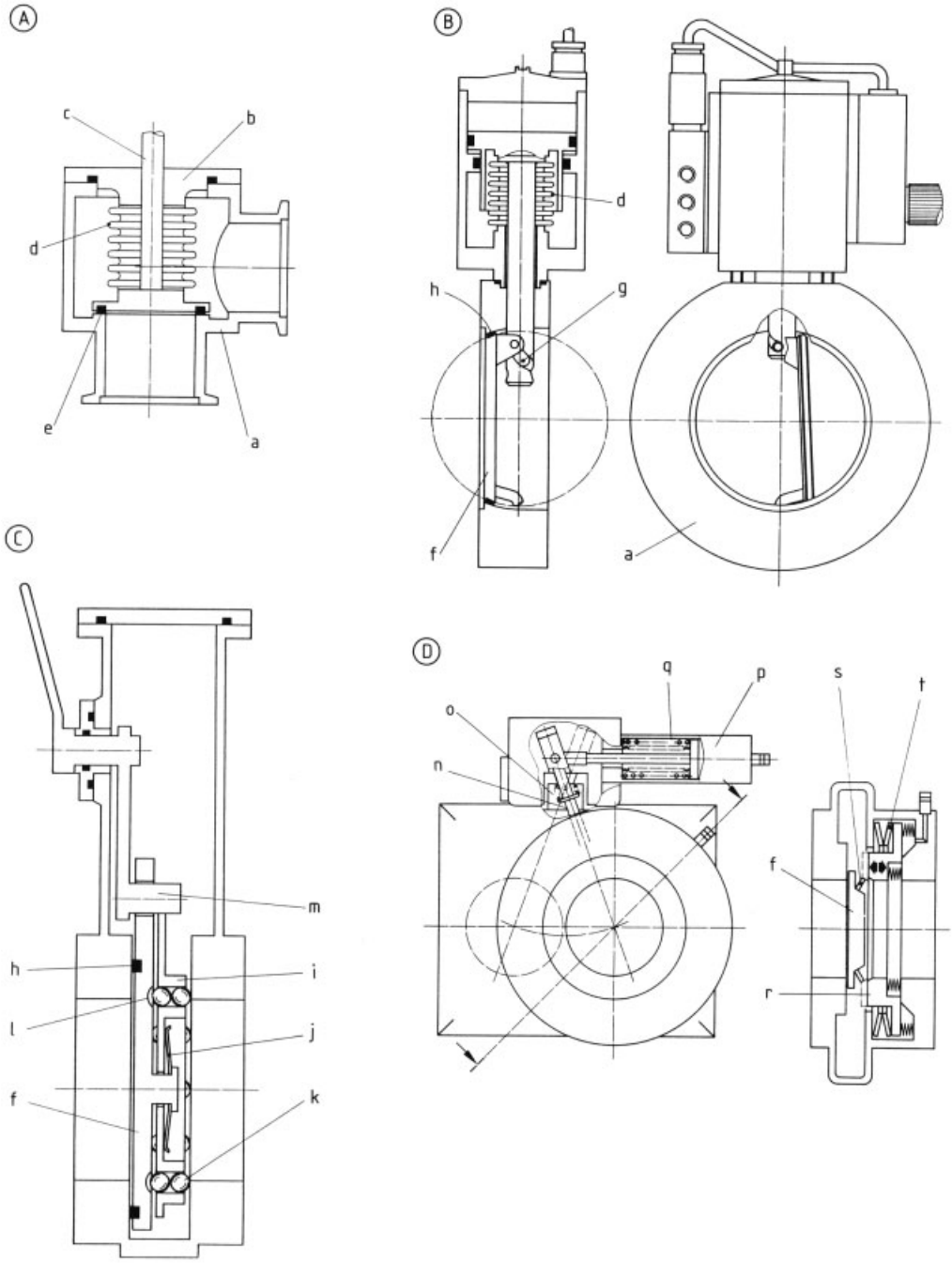

Figure 24. High vacuum valves [57]
A) Right angle valve with bellow seal
B) Butterfly valve
C) Gate valve, linear motion
D) Gate valve, sideward motion
a) Housing; b) Corer for housing; c) Control rod; d) Metal bellow; e) Disk seal; f) Valve disk; g) Flat coil; h) Seal; i) Sphere-control; j) Flat spring; k) Pair of spheres; e) Sphere sockets; m) Turning lever; n) Bearing rod; o) Membrane bellow; p) Pneumatic cylinder; q) Return spring (pendulum motion); r) Pressure cell; s) Sealing ring; t) Return spring (pressure cell)

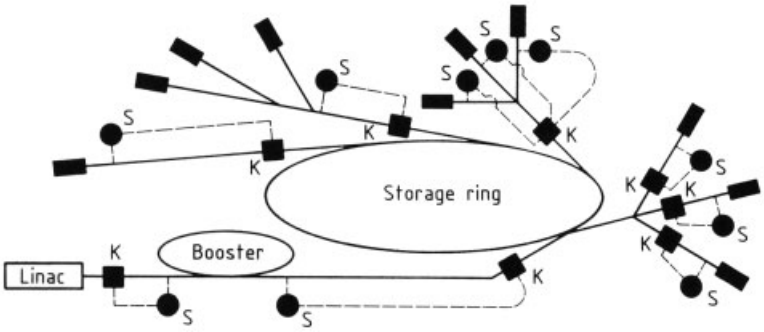

Figure 25. Example for safety shut-off devices in a storage ring plant [57]
S: pressure sensor
K: nonreturn valve

disk through a rotating gate (butterfly valve) or sideward motion or linear withdrawal (slide valves). The mechanisms are so designed that no sliding motion takes place on the seats. As with every additional component in a piping system, valves reduce the conductance, angle valves more intensely than throughflow valves, and also possess an unavoidable leak rate.

Isolating valves are available in nominal sizes from 10 to >1000 mm. In general, they are meant for operation in the fully open or fully closed positions. They are switching devices for individual pipes or shutoff devices for system components. In the form of special designs to suit a special purpose, they are also installed as locks for loading and unloading vacuum chambers. Simple isolating valves of small size serve as venting valves for admitting air into the system. So-called quick closing valves protect larger systems from air break-in by partitioning off the leaking portions of the system (Fig. 25). Gas admittance or regulating valves can be adjusted reproducibly to positions between on and off. In these, the position of a valve plug of varying diameter relative to a conical valve seat is changed to vary the valve port area, and thus conductance. Regulating valves are available with gas flow rates of 10^{-10} – 10^{3} cm^3/s, either hand-operated or electronically adjustable. At higher rates, the gas inflow can be handled with gas flow control instruments, largely independent of the pressure difference. A signal proportional to the gas flow, measured in a side stream conduit by a sensor system that depends on thermal conductivity, is the control variable for an electrically-operated valve. By this means, commercially available instruments control gas flows between $<10^{-3}$ to $>10^{5}$ cm^3/s with high accuracy and reproducibility. Either calibration or correction factors take into account the gas dependency. These are entered into the control instrument or are already stored there electronically and called up as needed. Conductance controllers using adjustable iris shutters or adjustable louvers regulate the vacuum process pressure and gas throughput. This allows adjustment of the vacuum system to the pumping speed of the pump system.

If a vacuum system is to be automated, then electrically-controlled valves can be used to do it. When these valves are equipped with valve position indicators, they provide the necessary signals indicating the switch position status needed for process system automation.

5. Vacuum System Applications

The applications of vacuum technology are numerous and a few fields of application are discussed here. Vacuum technology is used in preparation and processing of materials–for drying, freeze-drying, impregnating, distilling, modern welding procedures, remelting metals, production of metal powders, crystal pulling, and for degassing a wide range of materials. The foodstuff, pharmaceutical, metallurgical, and electronic industries have developed equivalent processes.

Manufacturing certain products from component parts can be done successfully only under vacuum. Examples of this are found in producing thin film surface coatings with thicknesses in the µm range for the electronic, optical, and tool industries. Vacuum switches, cathode ray tubes, transmitting tubes, and incandescent bulbs require vacuum for their operation. Modern physical surface analysis methods make use of vacuum technology, for example, in electron microscopy, and various other excitation and detection mechanisms. These permit conclusions to be drawn on the structure, composition, and state of the chemical bonds of solid surfaces. Particle accelerators, storage rings, nuclear fusion research plants, and space simulation belong to the most extensive scientific applications of vacuums.

Vacuum has an important industrial significance–here the term negative pressure is preferred–in conveying equipment for moving paper in printing machines, for moving plate-like loads, and for conveying granulated materials or liquids.

5.1. Perspectives on Design of Vacuum Systems

The working pressure p_W determines the design of a vacuum system. The following equation expresses the relationships between p_W, p_E, the ultimate pressure attainable by the pump system, and the gas flows shown in Figure 26:

$$p_W - p_E = (q_D + q_A + q_P + q_L + q_F)/S_{eff} \quad (23)$$

where

q_D = flow arising from desorption
q_A = flow arising from absorption
q_P = flow arising through permeation
q_L = flow arising through leakage into the system
q_F = flow originating in the process
S_{eff} = effective pumping speed at the vacuum receiver

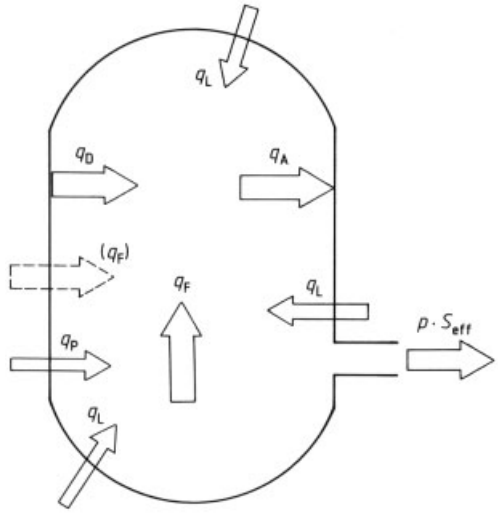

Figure 26. Gas throuphput of a vacuum chamber (see text for explanation of symbols)

Nearly all quantities on the right-hand side of Equation (23) are temperature- and pressure-dependent and thus, indirectly, a function of the sequence of the vacuum process. From this results the timedependence of the working pressure p_W. For estimates, the relationship shown by Equation (23) may be simplified. Thus, for working in the rough vacuum region, one may neglect desorption, absorption, and permeation and can concentrate, for example, on the leak rate for a given pump size. By comparison, for operations in the UHV region, q_L (and q_F) are normally negligibly small, and the importance of the quantities q_D,

q_A, and q_P is enhanced.

The economy of a vacuum system improves with careful adjustment of the effective pump speed S_{eff} to the system's requirements. The individual pump stages consisting of various pump types, which must be installed to work over wide ranges of vacuum, must also be suited to one another. The right choice of pump types and sizes can minimize investment costs and energy requirements. The choice of individual pump types and the components of a vacuum system must suit the anticipated working pressure and correspond to the nature of the vacuum process. In this, quantity and nature of the gases at hand, corrosion properties, presence of condensibles, entrainment in the form of liquid drops or solid particles are important. Factors assuming

even greater importance in conceptual design of vacuum systems are environmental impact and waste disposal problems. Above all, increased quantities of cooling water, or systems not having closed recycle loops for working fluid supply, cause additional burdens on the required environmental protection system. Necessary disposal of old oil and process-derived waste products is difficult and may act as a significant cost factor.

5.2. Examples of Vacuum Systems and Their Application

5.2.1. Decontamination of Groundwater

The careless handling of chlorinated hydrocarbons has in many cases led to contamination of the soil and the groundwater. Drawing air from wells through the ground represents one possible way to remove this pollutant. Blowing oil-free compressed air through the soil brings a further improvement in the groundwater ("in-situ stripping") [58].

Depending on the condition of the soil, the process requires pumping of 600 – 700 m^3/h at low pressure differences of 10 – 200 mbar. Beside the pollutant portion of up to 14 kg/h of chlorinated hydrocarbon [59], the pumped air may contain high amounts of water vapor and solids. Thus, this service requires a rugged vacuum pump having a low compression ratio and high pump speed in the rough vacuum region. In addition, these installations often include necessary separator equipment for sand on the suction side and filters with activated carbon on the discharge side for the hydrocarbons.

Problem-free, simple vacuum systems which fulfill these requirements may be assembled using, for example, gaseous ring pumps. Since they run dry, these offer no working fluid or lubricating problems. Technically, they are very simple and thus ideally suitable for mobile units to be set up at the polluted site.

Systems of this kind have been installed and operate successfully. They result in considerable savings compared to conventional methods of decontamination, as for example, excavation and removal of the soil.

5.2.2. Degassing Steel with a Steam Jet Vacuum System

Degassing molten steel takes place at a pressure of 0.5 mbar for which pump speeds to $>10^6$ m^3/h are necessary. The resulting dust is pyrophoric in air and must, therefore, be removed safely. Figure 27 shows a steel degassing unit using a three-stage steamjet pump installation to produce vacuum. Highest speed steamjet pumps are manufactured at competitive prices. Their operation is completely unaffected by dust in the suction gas. Pump stages a_1 and a_2 are directly connected one to another in series and compress to a pressure which permits condensation of the motive steam in the barometric condenser b_1. The pyrophoric dust is washed out by stream condensation and settles in a sedimentation basin. After pump stage a_3, the residual gas is purified in b_2 by the remaining dust. Lastly, the liquid ring pump d compresses the residual gas to atmospheric pressure. Residual operating liquid and condensate are led back, via the last separator b_3, into the cooling cycle. A cooling tower cools the condenser cooling water to close the circuit permitting water reuse with minimal environmental impact. The only condensate produced is condensed motive steam. The washing effect of the barometric condensers results in extensive removal of pollutants from the residual gas.

5.2.3. High Vacuum Systems for Manufacturing Thin Films

Vaporization and deposition by means of cathode sputtering are vacuum operations for producing thin films on substrates in the thickness range up to 10 µm. The palette of uses for these films includes the following: decorative films, optical films, mirror coatings on projector reflectors, heat-reflecting coatings on window glass, resistance films, electrode coatings for foil condensers, coatings for semiconductor components, tribologically-effective surface layers for dry lubricant coatings, and wear-protective coatings. It is possible to coat surfaces not only with metallic layers, but by carrying out reactive processes, with oxides, nitrides, and layers of other compounds.

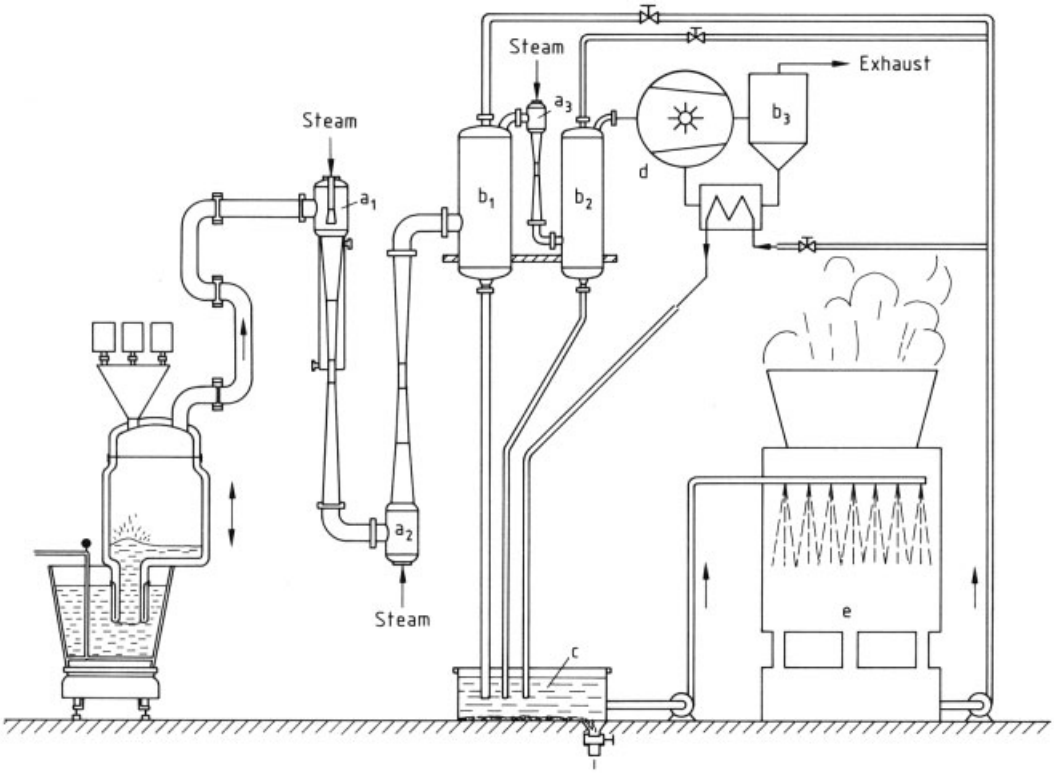

Figure 27. Degassing of steel with vapor jet vacuum pumps [60]
a_{1-3}) Vapor jet pumps; b_{1-3}) Condensing devices; c) Dust sedimentation; d) Liquid ring pump; e) Heat exchange

The basic principle of the vacuum system required for this is the same for all applications: a titanium nitride protective coating, for example, is deposited reactively by dispersing a titanium cathode in a nitrogen-containing gas discharge. The gas necessary for sputtering is a mixture of argon with a partial pressure of about 0.01 mbar and nitrogen at about 10^{-4} mbar partial pressure. Water vapor is decomposed in the gas atmosphere and the oxygen is incorporated into the resulting surface layer. To avoid producing this oxynitride film, it is necessary to reduce the water partial pressure to a value substantially below that of the nitrogen. The vacuum technology prerequisite, therefore, is a high vacuum system with an ultimate pressure in the 10^{-7} mbar range and with sufficient pumping speed, yet at 0.01 mbar suction pressure. These requirements are best resolved with a high vacuum pumping installation using a turbomolecular or a cryopump, as shown in Figure 28, as the high vacuum stage.

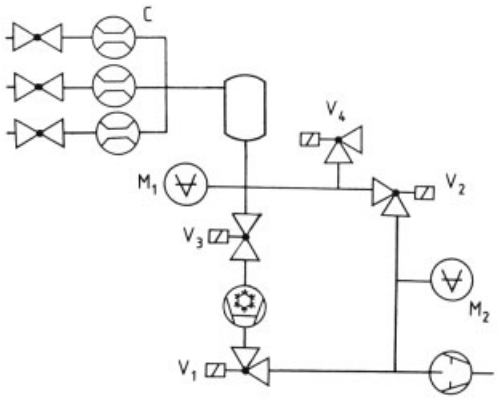

Figure 28. High vacuum plant–principle (see Fig. 2 for explanation of graphics)

The evacuation step proceeds in the customary way under automatic control. First, with valves V_1 and V_3 closed, the receiver is evacuated to a pressure that depends on the type of high vacuum pump. Then, with V_2 closed and V_1 and V_3 open, the receiver is evacuated to the ultimate pressure. V_4 can be used to vent the receiver. Gas flow controllers (C) regulate the inflow of process gas. Vacuum measuring points M_1 and M_2 indicate the pressure at the receiver and the vacuum produced by the forepump. While valve V_3 is necessary if one uses oil diffusion pumps, it may not be needed with turbomolecular or cryo-pumps.

At 0.01 mbar, tubomolecular and, particularly cryopumps possess sufficient pump speed compared to oil diffusion pumps, and permit pumping out to the desired high vacuum. At high gas pressure, the total pump speed of the turbomolecular pump must be throttled by a valve connected upstream. Whereas with the cryo-pump only the argon pumping speed for cryo-pump stage 1 need be reduced selectively via louvers, the water vapor pumping speed for pump stage 2 remains in full effect. Thus, low water vapor partial pressure is always assured, and the pump capacity for argon and nitrogen is preserved for hours or days.

The laboratory system corresponding to the flow scheme of Figure 28, when equipped with a refrigerator cryopump of 5400 m^3/h for air and a forepump of 60 m^3/h, has a potential operating cycle between regeneration steps of >15 h at 8×10^{-3} mbar argon pressure. The forepump is used only for pumping out the receiver after repressuring and for removing the melting gases during regeneration. To achieve short regeneration times, it is grossly oversized. During the making of the surface coating, the valve between the cryopump and forepump is closed and the forepump shut down.

The same scheme as that described for the laboratory system is used for the commercial system shown in Figure 29. It serves to deposit a layer of hard material on tools. With sputter cathodes of 1350×230 mm^2 and a substrate holder of 1050×600 mm^2 area, eight hours are needed to coat 6000 6-mm drills and 800 150-mm milling tools with a titanium nitride layer. An inlet lock chamber (a) and an outlet lock chamber (e) precede and follow the actual coating chamber (c), joined to them by lock valves (b). Movement of the substrate through the system is automatic; thus, the coating chamber is always held under vacuum. Rotary lobe, sliding vane pump combinations evacuate the repressured lock chambers at each loading-unloading cycle. These have high pump speeds up into the medium vacuum range. Eight turbomolecular pumps (only four are shown) with a total pump speed of 21 600 m^3/h maintain vacuum in the main chamber. The forepumps for these are again a rotary lobe/sliding vane installation.

Similar industrial vacuum systems, built mostly in modular fashion with chambers connectable to one another in series, are installed, for example, for coating plastic film with magnetic layers for storage media or for coating architectural glass of up to 3×6 m^2 with heat barrier layers.

5.2.4. Ultra High Vacuum System for the LEP Storage Ring at Geneva

At CERN in Geneva, the storage ring for the LEP machine (large electron positron collider) has been built for conducting electron-positron scattering experiments [54], [62]. In the final design, electron and positron energies exceeding 90 GeV should be possible. The storage ring, in the form of an octagon rounded at the corners, has a length of 27 km. An average pressure of 4×10^{-9} mbar is necessary to enable a beam lifetime, determined by the residual gas, of >20 h. This requires a base pressure $<10^{-10}$ mbar without the electron beam and also in some individual compartments with the beam.

In operation, the residual gas arises, above all, through photodesorption caused by the intense synchrotron radiation. This is incident on a strip of the beam tube wall only a few millimeters wide with a specific power of 1.4 kW per meter of beam tube length. A big part of this energy must be withdrawn as heat. For this, water-cooling ducts are required next to the beam tube. The beam tube has an elliptical cross section of 131×70 mm. This cross section has been greatly reduced from that possible by applying accelerating and focussing electromagnets.

Special concepts for materials and pumps are necessary to fulfill the catalog of these requirements and properties. The vacuum system

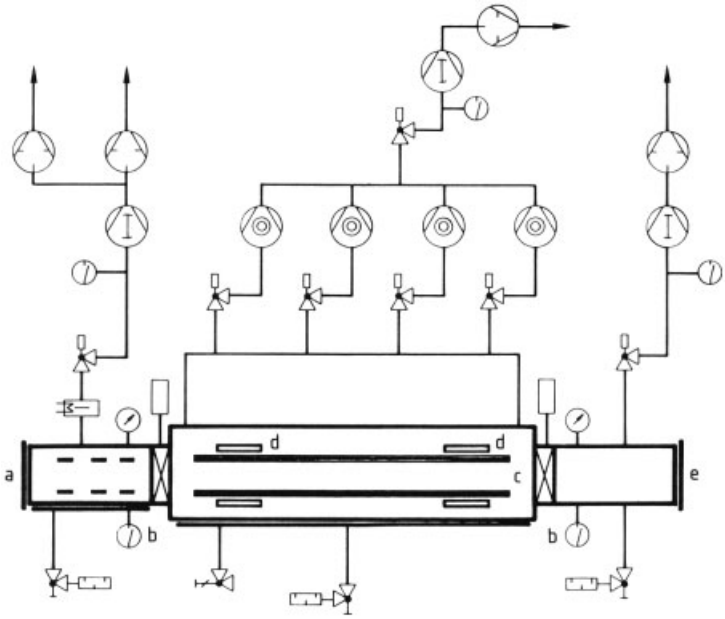

Figure 29. High vacuum thin film production plant [61]
a) Input vacuum interlock; b) Valves; c) Sputtering chamber; d) Cathodes; e) Exit vacuum interlock

must be built to meet the specifications for full UHV operation. This means materials suitable for high-temperature operation and metal seals. The beam tube is flanged together from separate pieces of length suited to the various magnets. Because of its good thermal conductivity and ease of fabrication, an aluminum alloy is used; it is extruded in the shape of the vacuum chamber cross section shown in Figure 30. Next to the elliptical cross section of the beam tube, one recognizes the three cooling ducts. In addition, there is the rectangular pump duct connected to the beam tube through slits.

A getter pump (NEG pump) with a total length of 20 km connects to the pump duct. This pump, distributed over the whole ring, removes the gases produced directly at the locations where they occur. The NEG pump is a two-phase mixture of zirconium and 16 % aluminum pressed on both sides of a 30-mm wide constantan heating strip. This can be heated by direct current. After a 10-min activation at 700 °C, in which the gases chemically bonded to the getter surface diffuse into the interior, a pump speed of ca. $1000\,\mathrm{s}^{-1}\,\mathrm{m}^{-1}$ is achieved, which depends on the type of gas.

In addition to the NEG main pump for the reactive gases, 1800 ion getter pumps, each having a pump speed of $108\,\mathrm{m}^3/\mathrm{h}$, are installed for pumping the noble gases and methane. Finally, simple turbomolecular pump systems, having $500\,\mathrm{m}^3/\mathrm{h}$ speed, are used for pumping from atmospheric pressure to the high vacuum region, for leak survey work, and while the vacuum system is being heated up.

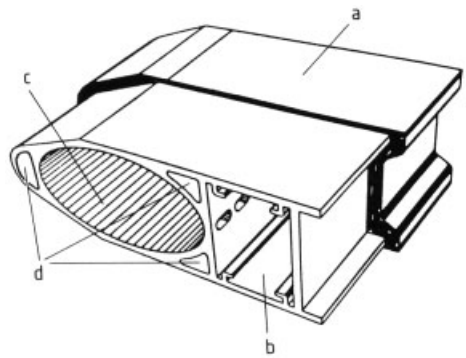

Figure 30. Cross section through the LEP vacuum chamber [63]
a) Lead shielding; b) Pump channel; c) Beam channel; d) Water cooling ducts

At least a twenty-year operating life is expected for the capacity of the NEG pump for the irreversibly bonded gases carbon monoxide and carbon dioxide at the planned operating times and parameters.

6. References

1. *Deutsche Normen DIN 28 400, Teil 1, Vakuumtechnik; Benennungen und Definitionen, Juli 1979.*
2. *Deutsche Normen DIN 28 401, Vakuumtechnik; Bildzeichen, Nov. 1976.*
3. *Deutsche Normen DIN 1314, Druck; Grundbegriffe, Einheiten, Feb. 1977.*
4. F. Kohlrausch, *Praktische Physik,* vol. 3, table 143, B. G. Teubner, Stuttgart 1968.
5. J. F. O'Manlon, "Vacuum pump fluids," *J. Vac. Sci. Technol.* **A2** (1984) 174 – 181.
6. Leybold-Heraeus GmbH, *Katalog HV 200 Vakuumtechnik, Teil 4, 11/85* [Catalog HV 200 Vacuum Technology, part 4, Nov., 1985] p. 4.5.
7. K. Ploog (ed.): *Crystals, Growth, Properties, and Applications,* Springer Verlag, Berlin – Heidelberg – New York 1980, pp. 73 – 162.
8. R. E. Honig: "Vapor pressure data for the more common elements," *RCA-Rev.* **18** (1975) 195 – 204.
9. d'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* **vol. I, 3rd ed.,** Springer Verlag, Berlin – Heidelberg – New York 1967.
10. R. C. Weast, S. M. Selby, C. D. Hodgman (ed.): *Handbook of Chemistry and Physics,* 46th ed., The Chemical Rubber Co., Cleveland 1965 – 1966.
11. Union Carbide, "Molekularsiebe."
12. S. S. Rosenblum: "Vacuum outgassing rates of plastics and composites for electrical insulators," *J. Vac. Sci. Technol.* **A4** (1986) 107 – 110.
13. J. R. Chen, K. Narushima, H. Ishimaru: "Thermal outgassing from aluminum alloy vacuum chambers," *J. Vac. Sci. Technol.* **A3** (1985) 2188 – 2191.
14. N. Yoshimura: "A differential pressure-rise method for measuring the net outgassing rates of a solid material and for estimating its characteristic values as a gas source," *J. Vac. Sci. Technol.* **A3** (1985) 2177 – 2183.
15. E. D. Erikson, T. G. Beat, D. D. Berger, B. A. Frazier: "Vacuum outgassing of various materials," *J. Vac. Sci. Technol.* **A2** (1984) 206 – 210.
16. C. Edelmann: *Wissensspeicher Vakuumtechnik,* VEB Fachbuchverlag, Leipzig 1985.
17. L. Laurenson, N. T. M. Dennis: "Permeability of common elastomers for gases over a range of temperatures," *J. Vac. Sci. Technol.* **A3** (1985) 1707 – 1710.
18. T. A. Giorgi, B. Ferrario, B. Storey: "An updated review of getters and gettering," *J. Vac. Sci. Technol.* **A 3** (1985) 417 – 423.
19. H. F. Dylla, P. H. LaMarche, W. R. Blanchard, R. V. Budny, et al.: "Chromium getter studies in the Tokamak Fusion Test Reactor," *J. Vac. Sci. Technol.* **A4** (1986) 1753 – 1757.
20. Balzers, *Katalog Komponenten für die Vakuumtechnik, Ausgabe 86/88, Öldiffusionspumpe DIF 1000* [Catalog Components for Vacuum Technology, 1986/88, Oil Diffusion Pump DIF 1000] p. E12.
21. Balzers, *Katalog Komponenten für die Vakuumtechnik, Ausgabe 86/88, Drehschiebervakuumpumpe DUO 016 B* [Catalog Components for Vacuum Technique, 1986/88, Rotary Vane Vacuum Pump DUO 016 B] p. B3.
22. Edwards Kniese, *Katalog Vakuumtechnik 1984, Booster-Pumpe 30B5A* [Catalog Vacuum Technology 1984, Booster Pump 30B5A] p. 47.
23. Edwards Kniese, *Betriebsanleitung 03-A363-01-880, 'Superpump' – Vakuumsysteme, Handbuch für Aufstellung und Betrieb* [Instructions 03-A363-01-880, "Super Pump" Vacuum Systems, Installation and Operation] April 1985.
24. Leybold-Heraeus, *Operating Instruction GA 771/3, RPK Refrigerator-Cooled Cryopumps, Technical Datas of Sputter Cryopumps, Kryopumpe RPK 1500 S3.*
25. Leybold-Heraeus, *Katalog HV 200, Vakuumtechnik, Teil 5, 12/85, Turbomolekularpumpe Turbovac 1000 C mit Flansch DN 250, D60A-Vorpumpe,*[Catalog HV 200, Vacuum Technology, part 5, Dec., 1985, Turbomolecular pumps Turbovac 1000 C with flange DN 250, D60 A Forepump] p. 5.15.
26. Siemens, *Katalogblatt PF1 19 Ausgabe A3, Flüssigkeitsringvakuumpumpe ELMO-F 2BE1 152-0, 1810 rpm* [Catalog PF1 19, A3, Liquid Ring Vacuum Pump ELMO-F 2BE1 152-0, 1810 rpm].
27. Varian, *Vacuum Catalogue, Diode and Triode VACTION® Pumps 110 to 1000 l/s, Diode pumps 500 l/s.*

28. *Deutsche Normen DIN 28 400, part 2, Vakuumtechnik; Benennungen und Definitionen; Vakuumpumpen, Oktober 1980.*
29. K. W. Mugele: "Neues über Vakuumpumpen," *Übersee-Post Pumpen Vakuumpumpen Kompressoren '85,* 1985, 50 – 54.
30. C. B. Whitman: "Reclamation of vacuum pump fluids," *J. Vac. Sci. Technol.* **A5** (1987) 255 – 261.
31. M. Wutz, "Gasdynamische Deutung der Wirkungsweise von Diffusionspumpen," part 1, *Vakuum-Technik* **28** (1979) 136 – 147, part 2, *Vakuum-Technik* **28** (1979) 172 – 179.
32. Edwards Kniese, *Katalog Vakuumtechnik 1984* [Catalog Vacuum Technology 1984] p. 136.
33. Leybold-Heraeus, *Katalog HV 200, Vakuumtechnik, Teil 4, 11/85* [Catalog HV 200, Vacuum Technology, part 4, Nov. 1985] p. 4.4.
34. Balzers, *Katalog Komponenten für die Vakuumtechnik, Ausgabe 86/88.* [Catalog components for vacuum technology, 1986/88] p. E2.
35. Montedison Group Montefluos, *Fomblin® Y-VAC.*
36. Du Pont, *KRYTOX® Vacuum Pump Fluids.*
37. W. Becker: "Die Turbomolekularpumpe," *Vakuum-Technik* **15** (1966) 211 – 218 and **15** (1966) 254 – 260.
38. Leybold-Heraeus, *Katalog HV 200, Vakuumtechnik, Teil 5, 12/85* [Catalog MV 200, Vacuum Technology, part 5, Dec., 1985] p. 5.4.
39. F. Doni, C. Boffito, B. Ferrario: "Hydrogen isotope sorption and recovery by a nonevaporable getter combined with a chemical compressor material," *J. Vac. Sci. Technol.* **A4** (1986) 2447 – 2451.
40. W. Schwarz: "Thermische Begrenzung des Wasserdampf-Saugvermögens maschinengekühlter Kaltflächen," *Vakuum-Technik,* **35** (1986) 66 – 71.
41. R. A. Haefer: *Kryo-Vakuumtechnik,* Springer Verlag, Berlin – Heidelberg – New York 1981.
42. C. B. Hood: "The development of large cryopumps from space chambers to the fusion program," *J. Vac. Sci. Technol.* **A3** (1985) 1684 – 1689.
43. K. F. Poulter: "Effect of gas composition on vacuum measurement," *J. Vac. Sci. Technol.* **A2** (1984) 150 – 158.
44. *Deutsche Normen DIN 28 400, part. 3, Vakuumtechnik; Benennungen und Definitionen, Vakuummeß-geräte, October, 1980.*
45. J. K. Fremerey: "The spinning rotor gauge," *J. Vac. Sci. Technol.* **A3** 1715 – 1720 (1985).
46. C. R. Tilford: "We do have vacuum gauge calibration standards!," *Res. Dev.* **28** (1986) 105 – 112.
47. G. Reich: "Die Bedeutung physikalischer Größen bei der Messung niedriger Drücke," *Vakuum-Technik* **35** (1986) 72 – 78.
48. K. Kokubun, M. Hirata, M. Ono, H. Murakami, Y. Toda: "Frequency dependence of a quartz oscillator on gas pressure," *J. Vac. Sci. Technol.* **A3** (1985) 2184 – 2187.
49. F. Watanabe: "Point collector ionization gauge with spherical grid for measuring pressures below 10^{-11} Pa," *J. Vac. Sci. Technol.* **A5** (1987) 242 – 248.
50. D. Lichtman: "Perspectives on residual gas analysis," *J. Vac. Sci. Technol.* **A2** (1984) 200 – 205.
51. Balzers, *Partialdruckmessung in der Vakuumtechnik Druckschrift Nr. B6 800 169 PD.*
52. A. Cornu, R. Massot, *Compilation of Mass Spectral Data,* Heyden & Son, London 1966.
53. H. Kienitz, *Massenspektrometrie,* Verlag Chemie, Weinheim 1968.
54. C. Biguenet: "Spectrométrie de masse, Index des Spectres," Le Vide, nos. 159 – 160 and 161 – 175 (1972).
55. *Deutsche Normen DIN 28 404, Vakuumtechnik; Flansche; Maße, October, 1986.*
56. Leybold-Heraeus, *Katalog HV 200, Vakuumtechnik, Teil 11, Dreh-Schiebedurchführung DND S 35,* [Catalog HV200 Vacuum Technology, part 11, Pulley operation DND S 35], p. 11.20.
57. VAT, *Katalog Vakuumventile 86.*
58. F. Bruckner, H. M. Hareß, D. Hiller: "Die Absaugung der Bodenluft – ein Verfahren zur Sanierung von Bodenkontaminationen mit leichtflüchtigen chlorierten Kohlenwasserstoffen," *BBR* **37** (1986) 1 – 8.
59. R. Kickuth: "CKW aus hundert Meter Tiefe," *Chem. Rundsch.* 13 Feb. 1987.
60. W. Hummel (GEA Wiegand): "Dampfstrahl-Vakuumpumpen mit geschlossenem Kreislauf des Kondensats, korrosionsbeständig und umweltfreundlich," *Chem. Anlagen + Verfahren,* 1978, 55 – 60.
61. W. D. Münz: "Reactive Sputtering of Nitrides and Carbides," 'Nitride and Carbide Coatings' Course, LSRH, Neufchâtel 1985; Leybold-Heraeus Special Publication 11-S07.2.

62. O. Gröbner: "27 km vacuum system for the Large Electron Positron Storage Ring," *Atomkernenergie – Kerntechnik* **46** (1985) 182 – 186.

63. H. P. Reinhard (LEP Vacuum Group, CERN): "The vacuum system of LEP," Proc. IX IVC – V ICSS, Madrid 1983, pp. 273 – 282.

Valeric Acid → Carboxylic Acids, Aliphatic
Valine → Amino Acids

Principles of Chemical Reaction Engineering

K. Roel Westerterp, Department of Chemical Engineering, Twente University, Enschede, The Netherlands

Ruud J. Wijngaarden, Koninklye/Shell-Laboratorium, Amsterdam, The Netherlands

1. Introduction 8
2. Fundamentals 9
2.1. Chemical Rates 9
2.2. Relative Degree of Conversion . . 9
2.3. Selectivity and Yield 10
3. Microkinetics 11
3.1. Elementary Reactions 11
3.1.1. Fundamentals 11
3.1.2. Influence of Temperature 11
3.1.3. Influence of Concentration 12
3.2. Chemical Equilibria 12
3.3. Complex Reaction Schemes 13
3.3.1. Fundamentals 13
3.3.2. Polymerization Reactions 15
3.3.2.1. Conversion – Time Curve 15
3.3.2.2. Radical Intermediates Concentration 17
3.3.3. Heterogeneous Catalytic Reactions 17
3.3.3.1. Monomolecular Reactions 18
3.3.3.2. Bimolecular Reactions 18
3.3.4. Biochemical Reactions 19
4. Macrokinetics: Single-Phase Systems 21
4.1. Introduction 21
4.2. Macromixing 22
4.2.1. Residence-Time Distribution (RTD) 22
4.2.2. Batch Reactor (BR) 25
4.2.3. Plug-Flow Reactor (PFR) 26
4.2.4. Continuous, Ideally Stirred Tank Reactor (CISTR) 27
4.3. Micromixing 28
4.4. Sequence of Mixing 31
5. Macrokinetics: Multiphase Systems – Part I: Mass Transfer without Reaction 32
5.1. Introduction 32
5.2. Penetration Models 33
5.3. Stagnant Film Model 34
5.4. Mass-Transfer Coefficients 35
6. Macrokinetics: Multi-phase Systems – Part II: Mass Transfer with Reaction 36
6.1. Mass Transfer with Reaction in Series 36
6.1.1. Introduction 36
6.1.2. Monomolecular First-Order Kinetics 36
6.1.3. Arbitrary Monomolecular Kinetics 37
6.1.4. Multiple Phases 38
6.1.5. Bimolecular Reactions 38
6.1.6. Case Study: Shrinking Core Model 39
6.1.7. Selectivities of Multiple Reactions . 41
6.1.7.1. Introduction 41
6.1.7.2. Parallel Reactions 42
6.1.7.3. Consecutive Reactions 43
6.1.7.4. Complex Reaction Schemes 44
6.2. Mass Transfer with Simultaneous Reaction 45
6.2.1. Introduction 45
6.2.2. Reaction of Gases in Porous Solids 45
6.2.2.1. Introduction 45
6.2.2.2. Monomolecular First-Order Kinetics 46
6.2.2.3. Monomolecular Arbitrary Kinetics 49
6.2.2.4. Bimolecular Reactions 50
6.2.2.5. Anisotropic Catalyst Pellets 51
6.2.3. Reactions of Gases in Liquids – Part I: Gaseous Reactants Only 53
6.2.3.1. Introduction 53
6.2.3.2. Higbie Penetration Model 53
6.2.3.3. Surface Renewal Model 55
6.2.3.4. Stagnant Film Model 55
6.2.3.5. Comparison of Models for First-Order Kinetics 56
6.2.3.6. Comparison of Hatta Numbers with Thiele Moduli 56
6.2.3.7. Hatta Numbers for Arbitrary Kinetics 57
6.2.3.8. Case Study: Ideally Mixed Gas – Liquid Reactors 58
6.2.3.9. Selectivities of Multiple Reactions . 61
6.2.4. Reactions of Gases in Liquids – Part II:Reactants from both the Gas and the Liquid 64
6.2.4.1. Introduction 64
6.2.4.2. Slow Reaction 65
6.2.4.3. Fast Reaction 65
6.2.4.4. Instantaneous Reaction 66
6.2.4.5. Van Krevelen – Hoftijzer Approximation 68
6.2.4.6. Arbitrary Kinetics 70
6.2.4.7. Selectivities of Multiple Reactions . 71
6.3. Combination of Simultaneous and Series Mass Transfer and Reaction 72

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_005

7. Macrokinetics: Multiphase Systems – Part III: Mass and Heat Transfer with Reaction 73
7.1. Introduction 73
7.2. Mass and Heat Transfer with Reaction in Series 73
7.2.1. Particle Mass and Heat Balances . . 74
7.2.2. Particle Multiplicity 75
7.3. Mass and Heat Transfer with Simultaneous Reaction 78
7.3.1. Frank – Kamenetzki Approximation 79
7.3.2. Thiele Moduli and Hatta Numbers . 80
8. References 81

Symbols:

a specific contact area between reaction and nonreaction phase, m^2/m^3 dispersion

a_A chemical activity of component A, mol/m^3

a_g $= \frac{\lambda_g}{\varrho_g C_{p,g}}$; gas thermal diffusivity, m^2/s

a_{gl} specific gas – liquid interface area per unit of dispersion volume, m^{-1}

a_{ls} specific catalyst particle external surface area per unit of dispersion volume, m^{-1}

A_p pellet external surface area, m^2

Al $= \frac{(1-\varepsilon)\, k_{1,A}}{a \mathbb{D}_A}$; Hinterland ratio, ratio between total liquid-phase volume and liquid-phase stagnant film volume

An $= \frac{E_a}{\Delta T_{ad} R} Le^{2/3}$; modified Arrhenius number

b_A adsorption constant of component A for slow surface reaction and relatively fast ad- and desorption, Pa^{-1}

$b_{A,\infty}$ pre-exponential factor of b_A, Pa^{-1}

Bi_A $= \frac{k_{g,A} r_p}{\mathbb{D}_{i,A}}$; Biot number of component A used in the shrinking core model

$\cosh(u)$ hyperbolic cosine function: $\cosh(u) = \frac{e^u + e^{-u}}{2}$

C total molar concentration of the reaction mixture, mol/m^3

C_A concentration of component A, mol/m^3

$\overline{C_A}$ concentration of component A in well-mixed fluid bulk, mol/m^3

$C_{A,c}$ concentration of component A at core boundary in shrinking core model, mol/m^3

$\overline{C_{A,g}}$ concentration of component A in well-mixed gas phase bulk, mol/m^3

$C_{A,i}$ concentration of component A at interface, mol/m^3

$C_{A,in}$ concentration of component A in reactor inlet stream, mol/m^3

$C_{A,i,g}$ concentration of component A at gas – liquid interface on gas-phase side, mol/m^3

$C_{A,i,l}$ concentration of component A at gas – liquid interface on liquid-phase side, mol/m^3

$C_{A,i,s}$ concentration of component A at catalyst pellet external surface, mol/m^3

$\overline{C_{A,l}}$ concentration of component A in well-mixed bulk of liquid phase, mol/m^3

$C_{A,p}$ concentration of component A inside catalyst pellet with negligible intra particle concentration gradient, mol/m^3

C_{A0} liquid inlet concentration of component A for bubble column, mol/m^3

$C_{A,0}$ concentration of component A for zero conversion, mol/m^3

$C_A —_t$ concentration of component A at time t, mol/m^3

$C_A —_{t=0}$ concentration of component A at $t = 0$, mol/m^3

$C_A —_z$ concentration of component A at location z in plug-flow reactor, mol/m^3

$C_A —_{z=0}$ concentration of component A at $z = 0$ or at inlet of plug-flow reactor, mol/m^3

$C_{A,BR}(t)$ concentration of component A in batch reactor after reaction time t, mol/m^3

C_I initiator concentration, mol/m^3

C_M monomer concentration, mol/m^3

C_{MO} concentration of metal oxide in virgin ore for shrinking core model, mol/m^3

$C_{p,g}$ specific heat of the gas, $J\,kg^{-1}\,K^{-1}$

$C_{\mathrm{P.}}$ concentration of growing, free-radical polymer chains, mol/m^3

$C_{\mathrm{R.}}$ free-radical concentration, mol/m^3

C_{T} concentration of tracer material, mol/m^3

d_{b} gas bubble diameter, m

d_{S} $= \frac{\sum d^3}{\sum d^2}$; Sauter mean bubble diameter, m

Da $= \frac{kt}{C_{\mathrm{A,0}}}$ (BR), $\frac{k\tau}{C_{\mathrm{A}}|_{z=0}}$ (PFR), $\frac{k\tau}{C_{\mathrm{A,in}}}$ (CISTR and CSSTR); Damköhler number for zeroth-order kinetics

$\mathbb{D}_{\mathrm{A}}$ molecular diffusion coefficient of component A, m^2/s

$\mathbb{D}_{\mathrm{A,g}}$ diffusion coefficient of component A in gas phase, m^2/s

$\mathbb{D}_{\mathrm{fl}}$ diffusion coefficient in bulk of fluid, m^2/s

$\mathbb{D}_{\mathrm{i}}$ effective diffusion coefficient inside porous pellet, m^2/s

$\mathbb{D}_{\mathrm{i,A}}$ effective diffusion coefficient inside porous pellet of component A, m^2/s

$\mathbb{D}^{h}_{\mathrm{i,A}}$ effective diffusion coefficient in ring-shaped or cylindrical porous catalyst pellet in longitudinal direction, m^2/s

$\mathbb{D}^{*}_{\mathrm{i,A}}$ effective diffusion coefficient, modified for pellet anisotropy, m^2/s

$\mathbb{D}^{r}_{\mathrm{i,A}}$ effective diffusion coefficient in ring-shaped or cylindrical porous catalyst pellet in radial direction, m^2/s

$\mathbb{D}_{\mathrm{Kn}}$ Knudsen diffusion coefficent, m^2/s

$\mathrm{erf}\,(u)$ error function:

$$\mathrm{erf}\,(u) = \frac{2}{\sqrt{\pi}} \int_0^u \exp\left(-x^2\right) \partial x$$

E_{a} activation energy, J/mol

E_{A} $= \frac{\overline{J_{\mathrm{A}}}\ \text{with reaction}}{\overline{J_{\mathrm{A}}}\ \text{without reaction}}$; enhancement factor of component A

E_{A0} $= \frac{\overline{J_{\mathrm{A}}}}{k_{1,\mathrm{A}} C_{\mathrm{A,i}}}$; modified enhancement factor used, e.g., in Van Krevelen – Hoftijzer approximation

$E_{\mathrm{A}\infty}$ $= \left(1 + \frac{1}{\nu} \frac{\mathbb{D}_{\mathrm{B}}}{\mathbb{D}_{\mathrm{A}}} \frac{\overline{C_{\mathrm{B}}}}{C_{\mathrm{A,i}}}\right) \left(\frac{\mathbb{D}_{\mathrm{B}}}{\mathbb{D}_{\mathrm{A}}}\right)^n$; enhancement factor for case of instantaneous reaction

E_{+} activation energy for reaction to right-hand side, J/mol

E_{-} activation energy for reaction to left-hand side, J/mol

$E\,(t)$ residence-time frequency distribution as function of t, s^{-1}

$E\,(\theta)$ residence-time frequency distribution as function of θ

f initiator efficiency; for $f = 1$, all initiator decomposes to active radicals; for $f = 0$, no active radicals are formed at all

$F\,(t)$ $= \int_0^t E\,(u)\ \partial u$; cumulative distribution of residence times

h longitudinal coordinate, m

H (1) parallelepiped height, m
(2) height of Raschig ring, m

Ha Hatta number, measure of square root of reaction rate divided by rate of mass transfer of component A; exact definition depends on kinetic expression

Ha_{i} modified Hatta number for reactions between reactants from a gas and a liquid, based on liquid-phase reactant interface concentration $C_{\mathrm{B,i}}$ (usual Hatta number is based on bulk concentration $\overline{C_{\mathrm{B}}}$)

$(-\Delta H_{\mathrm{ads}})$ adsorption enthalpy of component A on chemically active catalyst surface; since adsorption is an exothermic process, $(-\Delta H_{\mathrm{ads}}) > 0$, J/mol

$(-\Delta H_{\mathrm{r}})$ reaction enthalpy; for exothermic reactions, $(-\Delta H_{\mathrm{r}}) > 0$; for endothermic reactions, $(-\Delta H_{\mathrm{r}}) < 0$, J/mol

$\mathrm{I}_0\,(u)$ zeroth-order modified Bessel function of the first kind: $I_0\,(u) = \sum_{i=0}^{\infty} \left(\frac{u^i}{2^i\, i!}\right)^2$

$\mathrm{I}_1\,(u)$ first-order modified Bessel function of the first kind: $I_1\,(u) = \frac{u}{2} \sum_{i=0}^{\infty} \frac{(u/2)^{2i}}{i!\,(i+1)!}$

J_{A} molar flux of component A through interface, mol m^{-2} s^{-1}

$\overline{J_{\mathrm{A}}}$ average molar flux of component A for contact time τ of a stagnant fluid element in penetration theory, mol m^{-2} s^{-1}

$J_{\mathrm{A,p}}$ molar flux of component A at pellet external surface, mol m^{-2} s^{-1}

$(J_{\mathrm{A}})_{\mathrm{phys}}$ molar flux of component A through interface if no chemical reaction occurs, mol m^{-2} s^{-1}

$(J_{\mathrm{A}})_x$ molar flux of component A at location x, mol m^{-2} s^{-1}

$J_{\mathrm{A}}(r)$ molar flux of component A at location r, mol m^{-2} s^{-1}

k reaction rate constant; dimension depends on kinetic expression

k_{g} gas-phase mass-transfer coefficient for transfer from a gas to a liquid, m/s

$k_{\mathrm{g,A}}$ mass-transfer coefficient of component A in a gas, m/s

k_{l} liquid-phase mass-transfer coefficient for transfer from a gas to a liquid, m/s

k_{ls} liquid-phase mass-transfer coefficient for transfer from a liquid to a solid, m/s

$k_{\mathrm{l,A}}$ mass-transfer coefficient of component A in a liquid, m/s

$k_{n,m}$ reaction rate constant for bimolecular reactions with (n, m)-kinetics, m$^{3(n+m-1)}$ (mol A)$^{-(n-1)}$ (mol B)$^{-m}$, s^{-1}

k_{p} reaction rate constant based on pellet volume; dimension depends on kinetic expression

$k_{\mathrm{p,0}}$ pre-exponential factor for rate constant k_{p} (based on catalyst pellet volume)

$k_{\mathrm{p}}(T_{\mathrm{p}})$ reaction rate constant based on pellet volume for first-order kinetics at pellet temperature T_{p}, s^{-1}

k_{r} reaction rate constant; dimension depends on kinetic expression

k''_{r} surface reaction rate constant, m/s

k_{+} reaction rate constant of reaction to right-hand side; dimension depends on kinetic expression

k_{-} reaction rate constant of reaction to left-hand side; dimension depends on kinetic expression

k_{∞} reaction rate constant pre-exponential factor; reaction rate constant for infinite temperature; dimension depends on kinetic expression

K $= \frac{C_{\mathrm{P}}}{C_{\mathrm{A}}}, \frac{C_{\mathrm{P}}C_{\mathrm{Q}}}{C_{\mathrm{A}}C_{\mathrm{B}}}$; chemical equilibrium constant

K_{A} adsorption constant of component A, m^3/mol

K_{S} saturation constant in biochemical reactions, kg/m^3

$\mathrm{K}_0(u)$ zeroth-order modified Bessel function of the second kind:

$$K_0(u) = -\ln(u/2)\,\mathrm{I}_0(u) + \sum_{i=1}^{\infty}\left\{\mathrm{psi}(i+1)\left(\frac{u^i}{2^i i!}\right)^2\right\}$$

$\mathrm{K}_1(u)$ first-order modified Bessel function of the second kind:

$$\mathrm{K}_1(u) = \ln(u/2)\,\mathrm{I}_1(u) + -\frac{u}{2}\sum_{i=1}^{\infty} \times\left\{\frac{\mathrm{psi}(i+1) + \mathrm{psi}(i+2)}{2}\,\frac{(u/2)^{2i}}{i!\,(i+1)!}\right\}$$

L (1) length of plug-flow reactor, m
(2) parallelepiped length, m

Le $= \frac{a_{\mathrm{g}}}{\mathbb{D}_{\mathrm{A,g}}}$; gas Lewis number for component A

m (1) (apparent) reaction order of component B
(2) $= \frac{C_{\mathrm{A,i,l}}}{C_{\mathrm{A,i,g}}}$; dimensionless distribution coefficient or equilibrium constant of component A between gas and liquid phases

$m\,\mathrm{c}_k$ $\int_0^{\infty}(\theta-1)^k E(\theta)\,\partial\theta$; k-th central moment of dimensionless frequency distribution

$m\,\mathrm{o}_k$ $\int_0^{\infty}\theta^k E(\theta)\,\partial\theta$; k-th moment around origin of dimensionless frequency distribution

M_{A} molar mass of component A, kg/mol

M_{tot} total reaction mixture mass in closed system, kg

$M\,\mathrm{C}_k$ $\int_0^{\infty}(t-\bar{t})^k E(t)\,\partial t$; k-th central moment of frequency distribution, s^{k}

MO_k $\int_0^{\infty}t^k E(t)\,\partial t$; k-th moment around origin of frequency distribution, s^{k}

n (apparent) reaction order of component A

n_{A} number of moles of component A, mol

$n_{\mathrm{A},0}$ number of moles of component A for zero conversion, mol

$n_{\mathrm{T},0}$ number of moles of tracer injected in a pulse at $t = 0$, mol

psi (u) psi function; digamma function: $\mathrm{psi}\,(u) = -\gamma + \sum_{i=1}^{u-1} \frac{1}{i} \quad u \in \mathbb{Z}^+$

P_{A} partial pressure of component A, Pa

r radial coordinate, m

$r*$ radial coordinate in ring-shaped or cylindrical porous catalyst pellet, modified for pellet anisotropy, m

r_{c} core radius in shrinking core model, m

r_{p} (1) pellet radius in shrinking core model, m
(2) catalyst pellet radius, m

R ideal gas constant ($8.31434\,\mathrm{J\,mol^{-1}K^{-1}}$)

R_{A} chemical production rate of component A (if A is consumed, $R_{\mathrm{A}} < 0$; if A is produced, $R_{\mathrm{A}} > 0$), $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|R_{\mathrm{A}}|$ chemical conversion rate of component A (positive if reaction proceeds in direction of arrow), $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|R_{\mathrm{A}}|_{\mathrm{p}}$ conversion rate of component A based on catalyst pellet volume, $\mathrm{mol\,(m^3\,pellet)^{-1}\,s^{-1}}$

$|R_{\mathrm{A}}|_{\mathrm{P}}$ conversion rate of component A toward desired product P, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{X}}$ conversion rate of component A toward undesired byproduct X, $\mathrm{mol\,(m^{-3}\,s^{-1})}$

R_{i} inner radius of Raschig ring, m

R_{i}^* Raschig ring inner radius modified for pellet anisotropy, m

R_{u} outer radius of Raschig ring, m

R_{u}^* Raschig ring outer radius modified for pellet anisotropy, m

$|\mathbb{R}_{\mathrm{A}}|$ chemical reaction rate of component A; rate in absence of mass- or heat-transfer limitation, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\overline{\mathbb{R}}_{\mathrm{A}}|$ average reaction rate for two nonmixed fluid microelements, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\mathbb{R}''_{\mathrm{A}}|$ reaction rate based on pellet external surface area, $\mathrm{mol\,(m^2}$ external surface $\mathrm{area)^{-1}\,s^{-1}}$

$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{mixed}}$ reaction rate for two mixed fluid microelements, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}}$ reaction rate based on pellet volume, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\mathbb{R}''_{\mathrm{A}}|_{\mathrm{P}}$ rate of reaction yielding desired product P based on pellet external surface area, $\mathrm{mol\,(m^2}$ external surface $\mathrm{area)^{-1}\,s^{-1}}$

$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{r}}$ chemical reaction rate based on reaction mixture or reactor volume, $\mathrm{mol\,m^{-3}\,s^{-1}}$

$|\mathbb{R}''_{\mathrm{A}}|_{\mathrm{X}}$ reaction rate of reaction yielding undesired byproduct X based on the pellet external surface area, $\mathrm{mol\,(m^2}$ external surface $\mathrm{area)^{-1}\,s^{-1}}$

$\sinh(u)$ hyperbolic sine function: $\sinh\,(u) = \frac{\mathrm{e}^u - \mathrm{e}^{-u}}{2}$

S cross-sectional area of tube or plug-flow reactor, $\mathrm{m^2}$

s fraction of stagnant fluid elements replaced per unit of time in Danckwerts penetration or surface renewal model, $\mathrm{s^{-1}}$

$\tanh(u)$ hyperbolic tangent function: $\tanh\,(u) = \frac{\mathrm{e}^{2u} - 1}{\mathrm{e}^{2u} + 1}$

T (1) reaction mixture temperature, K
(2) temperature inside a catalyst pellet, K

$\bar{t}$ reactor average residence time, s

$t_1\%$ time lapse needed for polymer chain radical concentration to reach 99 % of steady-state value, s

ΔT_{ad} $= \frac{\overline{C_{\mathrm{A,g}}}(-\Delta H_{\mathrm{r}})}{\varrho_{\mathrm{g}} C_{\mathrm{p,g}}}$; adiabatic temperature rise for monomolecular gas-phase reaction, K

$\overline{T_{\mathrm{g}}}$ temperature of gas bulk surrounding catalyst pellet, K

T_{i} temperature at external surface of catalyst pellet, K

$T_{\max}$ temperature at which the maximum rate is found for exothermic equilibrium reaction with fixed concentration in the reaction mixture, K

T_{p} catalyst pellet temperature for negligible intra particle temperature gradients, K

$T_{\mathrm{p,max}}$ $= \overline{T_{\mathrm{g}}} + \Delta T_{\mathrm{ad}}\, Le^{2/3}$; maximum pellet temperature for heterogeneously catalyzed gas-phase reaction, K

u superficial velocity; velocity based on empty tube cross-sectional area, m/s

u_{b} gas bubble rise velocity in liquid, m/s

u_x fluid velocity in x-direction, m/s

u_y fluid velocity in y-direction, m/s

u_z fluid velocity in z-direction, m/s

$\mathrm{U}(u)$ unit step function or Heaviside step function: $U(u) = \begin{cases} 0 & u < 0 \\ 1 & u \geq 0 \end{cases}$

V total reaction mixture volume, m^3

V_{p} pellet volume, m^3

W parallelepiped width, m

x coordinate perpendicular to interface through which mass transfer occurs, m

z axial coordinate in plug-flow reactor, m

α_{g} heat-transfer coefficient between catalyst pellet and surrounding gas, $\mathrm{W\,m^{-2}\,K^{-1}}$

γ $= \lim_{n\to\infty}\left(\sum_{i=1}^{n}\frac{1}{i} - \ln(n)\right)$ $= 0.5772156649$; Eulers constant

γ_{A} activity or fugacity coefficient of component A

γ_{B} $= v_B \frac{\mathbb{D}_{i,A}}{\mathbb{D}_{i,B}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{B,i}}}$; excess amount of component B for bimolecular reactions

γ_{p} pore tortuosity factor for porous pellet

δ film thickness in stagnant film model, m

δ_{A} film thickness for component A in stagnant film model, m

δ_{r} thickness of reaction zone in stagnant film with thickness $\delta > \delta_{\mathrm{r}}$, m

$\delta(u)$ Dirac delta function: $\delta(u) = \lim_{\varepsilon\to 0}\begin{cases} 0 & u < -\varepsilon, u > \varepsilon \\ \frac{1}{2\varepsilon} & -\varepsilon \leq u \leq \varepsilon \end{cases}$

Δ_{FK} error introduced on using Frank-Kamenetzki approximation

$\Delta_{\mathrm{FK,max}}$ maximum value of Δ_{FK} (i.e., for $C_{\mathrm{A}} = 0$ or low effectiveness factors)

ε (1) porosity of packed bed of catalyst pellets; volume fraction void (2) volume fraction of gas in gas-in-liquid dispersion

ε_{p} porosity of porous particle; fraction void volume

ζ_{A} relative degree of conversion of component A

$\zeta_{\mathrm{A,BR}}$ relative degree of conversion of component A in batch reactor

$\zeta_{\mathrm{A,CISTR}}$ relative degree of conversion of component A in CISTR

$\zeta_{\mathrm{A,CSSTR}}$ relative degree of conversion of component A in CSSTR

$\zeta_{\mathrm{A,equi}}$ relative degree of conversion of component A at equilibrium conditions

$\zeta_{\mathrm{A,PFR}}$ relative degree of conversion of component A in plug-flow reactor

ζ_{MO} relative degree of conversion of metal oxide in shrinking core model

$\zeta_{\mathrm{M,de}}$ maximum relative degree of conversion of monomer obtained for dead-end polymerization

η effectiveness factor; amount of key reactant converted in porous catalyst pellet, divided by amount that could be converted if no retardation by diffusion were to occur

$\eta^{\ominus}$ $= \frac{\overline{J_{\mathrm{A}}}aV}{kC_{\mathrm{A,i}}(1-\varepsilon)V}$; degree of reactor utilization

η_{P} $= \sigma_{\mathrm{P}}\,\zeta_{\mathrm{A}}$; yield of product P; ratio of amount of P formed to amount that could have been formed if all of key reactant A were converted solely to P

θ $= \frac{t}{\tau}$; dimensionless residence time in reactor

θ_g $= \frac{\overline{T_{\mathrm{g}}}}{\Delta T_{\mathrm{ad}}} Le^{2/3}$; dimensionless gas temperature

θ_p $= \frac{T_{\mathrm{p}}}{\Delta T_{\mathrm{ad}}} Le^{2/3}$; dimensionless pellet temperature

ι $= \frac{R_{\mathrm{i}}}{R_{\mathrm{u}}}$; dimensionless inner radius of Raschig ring

κ_{A} mole fraction of component A

$\overline{\kappa}_{\mathrm{A}}$ mole fraction of component A in bulk of fluid

$\kappa_{\mathrm{A,i}}$ mole fraction of component A at interface

λ $= \frac{H}{R_{\mathrm{u}}}$; dimensionless height of Raschig ring

λ_{g} gas thermal conductivity, $\mathrm{W\,m^{-1}\,K^{-1}}$

λ_{p} catalyst pellet thermal conductivity, $\mathrm{W\,m^{-1}\,K^{-1}}$

μ $= \frac{(-\Delta H_r)\mathbb{D}_{i,A}C_{A,i}}{\lambda_p T_i}$; measure of maximum relative temperature increase inside pellet

ν stoichiometric coefficient

ν' kg enzymes or cells per kg substrate in biochemical reactions

ν_B stoichiometric coefficient of component B

ξ_0 $= \frac{k_g A_p}{k_{p,0} V_p}$; mass-transfer rate to catalyst pellet over reaction rate at infinitely high temperature

$\xi_{0,bo}$ value of ξ_0 for which pellet blowout occurs

$\xi_{0,ra}$ value of ξ_0 for which pellet runaway occurs

$\xi(T_p)$ $= \frac{k_g A_p}{k_p(T_p) V_p}$; measure of mass-transfer rate to catalyst pellet divided by reaction rate

ϱ (1) density of reaction mixture, kg/m^3
(2) $= \frac{r_c}{r_p}$; dimensionless core radius in shrinking core model

ϱ_g gas density, kg/m^3

ϱ_0 density of reaction mixture for zero conversion, kg/m^3

σ_P selectivity toward product P; ratio of amount of P formed over amount that could have been formed if all of key reactant A had been converted to P

$\sigma_{P,Ha<0.2}$ selectivity toward desired product P for parallel reactions in slow reaction regime $Ha < 0.2$

$\sigma_{P,Ha>2}$ selectivity toward desired product P for parallel reactions in fast reaction regime $Ha > 2$

σ_t $= \sqrt{\sigma_t^2}$; standard deviation of frequency distribution, s

$\sigma^2{}_t$ $= MC_2$; variance of frequency distribution, s^2

$\sigma^2{}_\theta$ $= mc_2$; variance of dimensionless frequency distribution

τ (1) reactor holding time, s
(2) contact time between small stagnant fluid element and interface in Higbie penetration model, s

τ_D characteristic time for shrinking core in regime of slow diffusion through converted shell, s

τ_L $= \frac{(1-\varepsilon)V}{\varphi_{V1}}$; average residence time in liquid phase in bubble column, s

τ_M characteristic time for shrinking core in regime of slow external mass transfer, s

τ_R characteristic time for shrinking core in regime of slow reaction kinetics, s

φ $\equiv \frac{1}{\eta}$ for low η values thus high φ values; generalized Thiele modulus

φ_A $= \frac{V_p}{A_p}\sqrt{\frac{k_p}{\mathbb{D}_{i,A}}}$; Aris' shape-generalized Thiele modulus for first-order kinetics

φ_m mass flow rate, kg/s

φ_T $= r_p\sqrt{\frac{k_p}{\mathbb{D}_{i,A}}}$; simple Thiele modulus for first-order kinetics; not generalized for arbitrary catalyst geometries

φ_v volumetric flow rate, m^3/s

$\varphi_{v,0}$ volumetric flow rate of liquid feed to bubble column, m^3/s

$\varphi_{v,1}$ volumetric flow rate of liquid stream out of bubble column, m^3/s

χ $= \frac{E_a}{RT_i} \cdot \frac{(-\Delta H_r)\mathbb{D}_{i,A}C_{A,i}}{\lambda_p T_i}$; measure of maximum relative reaction rate increase inside pellet due to intra particle temperature gradients

ψ'' heat flux from or to catalyst pellet, W/m^2

ω_A mass fraction of component A

$\omega_{A,0}$ mass fraction of component A for zero conversion

1. Introduction

A rational approach to all problems relating to a physical or chemical change of matter must be based on elementary physical conservation laws. They have been handled systematically in, for example, [1] and [2]. For the treatment of chemical engineering phenomena, the laws of conservation of matter and energy are of primary importance; they are the basis of the discussions below. Flow phenomena, governed by the principle of conservation of momentum, naturally are equally important for chemical engineering calculations (→ Fluid Mechanics).

The second kind of information needed for a quantitative treatment of chemical engineering is concerned with rates. The rate at which a reaction proceeds may be determined not only by the

chemical kinetics of the reaction proper, but also by physical transport phenomena (→ Transport Phenomena).

The interaction between chemical kinetics and physical transport rates has received considerable attention during the past decades. DAMKÖHLER [3], and particularly FRANK-KAMENETZKI [4] have systematically developed this field. The latter author distinguished between *microkinetics* (i.e., chemical kinetics) and *macrokinetics* (i.e., physical rates). Many scientists (e.g., VAN KREVELEN [5], DANCKWERTS [6], and ASTARITA [7]) have stressed the importance of the scale of scrutiny, at which the phenomena in a reacting system are considered. In principle, three different scales must be taken into account:

1) The scale of a molecule, where, for example, molecular collisions and the molecule mean free path (diffusion) are studied
2) The scale of an eddy or dispersion in a heterogeneous system, where, for example, a catalyst particle or a gas bubble is taken as the basis for deriving the appropriate equations
3) The scale of a reactor as a whole, where small-scale phenomena must be integrated over the entire reactor and macroscopic mixing becomes important.

The discussion in the following chapters focuses on the scale described in item 2. If appropriate, references giving more extensive treatments are cited.

2. Fundamentals

2.1. Chemical Rates

For the application of material balance to chemical engineering problems, an expression must be available for the *chemical production rate*. The production rate of component A, R_A, is defined as the number of moles of A produced per unit time and volume. If A is consumed in a chemical reaction, R_A is negative; if A is produced, R_A is positive. The production rate in mass units is found by multiplying R_A by the molar mass M_A.

For convenience, the concept of the *chemical conversion rate* —R— is often used; it is always positive when a reaction proceeds in the direction of the arrow. The conversion rate is expressed in moles of key reactant consumed or produced per unit time and volume. When a reaction proceeds according to

$$\nu_A\,A + \nu_B\,B \rightarrow \nu_P\,P + \nu_Q\,Q$$

and A is chosen as key reactant:

$$|R| = -R_A = -R_B\frac{\nu_A}{\nu_B} = R_P\frac{\nu_A}{\nu_P} = R_Q\frac{\nu_A}{\nu_Q} \tag{2.1}$$

The coefficients ν_A, ν_B, ν_P, and ν_Q are called the *stoichiometric coefficients.*

In this text, the conversion rate is used in relevant equations to avoid difficulties in applying the correct sign for the reaction rate in material balances.

Note that the chemical conversion rate is not identical to the *chemical reaction rate*, which is denoted —$\mathbb{R}$—. The chemical reaction rate only reflects the chemical kinetics of the system, that is, the conversion rate measured under such conditions that it is not influenced by physical transport (diffusion and mass transfer) of reactants toward the reaction site or of product away from it. The reaction rate generally depends only on the composition of the reaction mixture, its temperature and pressure, and the properties of the catalyst. The conversion rate, in addition, can be influenced by the conditions of flow, mixing, and mass and heat transfer in the reaction system. For homogeneous reactions in which the reactants are well mixed on a molecular scale, and for heterogeneous reactions that proceed slowly with respect to potential physical transport, the conversion rate approximates the reaction rate. In constrast, for homogeneous reactions in poorly mixed fluids and for relatively rapid heterogeneous reactions, physical transport phenomena may reduce the conversion rate. Hence, in general, the conversion rate is lower than the reaction rate.

2.2. Relative Degree of Conversion

For a single reaction as given above, the mass fractions of the reactants, ω_A and ω_B, decrease and those of the products, ω_P and ω_Q, increase as the reaction proceeds from left to right. For a closed system (no material added or withdrawn),

$$\omega_{\mathrm{A}}+\omega_{\mathrm{B}}+\omega_{\mathrm{P}}+\omega_{\mathrm{Q}}=\text{ constant} \tag{2.2}$$

For such a system the *relative degree of conversion* ζ can be introduced, which is a measure of the extent to which the reaction has proceeded. It can be defined quite generally as the fraction of the amount of a reactant, fed prior to and during reaction, that has been converted. For a closed system, ζ can be expressed in terms of mass fractions:

$$\zeta_J=\frac{\omega_{J,0}-\omega_J}{\omega_{J,0}} \tag{2.3}$$

with J = A, B, respectively. Since, by definition, the total mass within a closed system M_{tot} is constant, the relative degree of conversion can also be expressed in terms of the number of moles:

$$\zeta_J=\frac{\frac{\omega_{J,0}M_{\mathrm{tot}}}{M_J}-\frac{\omega_J M_{\mathrm{tot}}}{M_J}}{\frac{\omega_{J,0}M_{\mathrm{tot}}}{M_J}}=\frac{n_{J,0}-n_J}{n_{J,0}} \tag{2.4}$$

with M_J being the molar mass of either A or B. If the total volume of the reaction mixture is not a function of the relative degree of conversion, the latter can be expressed relatively easily in terms of concentrations:

$$\zeta_J=\frac{\frac{n_{J,0}}{V}-\frac{n_J}{V}}{\frac{n_{J,0}}{V}}=\frac{C_{J,0}-C_J}{C_{J,0}} \tag{2.5}$$

with V the volume of the reaction mixture. If this volume changes during the reaction, Equation (2.5) yields

$$\zeta_J=\frac{C_{J,0}\frac{M_{\mathrm{tot}}}{\varrho_0}-C_J\frac{M_{\mathrm{tot}}}{\varrho}}{C_{J,0}\frac{M_{\mathrm{tot}}}{\varrho}}=\frac{\mathrm{C}_{\mathrm{J},0}-\frac{\varrho_0}{\varrho}C_J}{C_{J,0}} \tag{2.6}$$

with ϱ_0 being the density of the reaction mixture for zero conversion and ϱ the density for a conversion ζ.

The chemical reaction rate is usually dependent on the molar concentrations of the reactants and not on their mass fractions, because it depends on the change of collision of molecules. However, here the definition of ζ in terms of mass fractions is preferred, because it can readily be incorporated into mass balances. A definition in terms of moles or molar concentrations might invite the use of mole balances instead of mass balances. Since, contrary to conservation of mass, there is no such thing as conservation of moles (e.g., one molecule might split up in several molecules, or several might condense into one), the use of mole balances is strongly dissuaded.

More information concerning the definition of conversion can be found in WESTERTERP et al. [8].

2.3. Selectivity and Yield

In industrial chemical operations a reactant A or reactants A and B very often react to form not only desired products, but also undesirable ones. Therefore in chemical reactor technology the concepts of selectivity and yield are often used.

The *selectivity* σ_{P} is the ratio between the amount of desired product P obtained and the amount of key reactant A converted; both quantities are usually expressed such that σ_{P} ranges between 0 (no P formed) and 1 (all A converted to P). This definition leads to:

$$\sigma_{\mathrm{P}}=\frac{\left(\omega_{\mathrm{P}}-\omega_{\mathrm{P},0}\right)M_{\mathrm{A}}\nu_{\mathrm{A}}}{\left(\omega_{\mathrm{A},0}-\omega_{\mathrm{A}}\right)M_{\mathrm{P}}\nu_{\mathrm{P}}} \tag{2.7}$$

For constant density of the reaction mixture, this can be written as

$$\sigma_{\mathrm{P}}=\frac{\left(C_{\mathrm{P}}-C_{\mathrm{P},0}\right)\nu_{\mathrm{A}}}{\left(C_{\mathrm{A},0}-C_{\mathrm{A}}\right)\nu_{\mathrm{P}}} \tag{2.8}$$

The *yield* η_{P} is the ratio between the amount of desired product P obtained and the amount that could be obtained if all of key reactant A were converted to P with 100 % selectivity. Therefore, η_{P} can be calculated from

$$\eta_{\mathrm{P}}=\sigma_{\mathrm{P}}\zeta_{\mathrm{A}} \tag{2.9}$$

Like σ_{P}, the yield η_{P} usually varies between 0 and 1. It is high when both the selectivity and the relative degree of conversion are high; it is low when either of them is low.

If in the product separation section of the reactor, the key reactant A can be removed and recycled to the reactor, selectivity is the key factor in economical operation of the plant. If, however, A cannot be recycled, then yield is the key factor.

3. Microkinetics

3.1. Elementary Reactions

3.1.1. Fundamentals

Elementary reactions are individual reaction steps that are caused by collisions of molecules. The collision can occur in a more or less homogeneous reaction medium or at the reaction sites on a catalyst surface. Only three elementary kinetic processes exist: mono-, bi-, and trimolecular processes. Of these, trimolecular processes are rarely found, because the chance of three molecules colliding at the same time is very small.

Each elementary reaction consists of an activation of the reactants, followed by a transition state and decomposition of the latter into reaction products:

$$\mathrm{A}(+\mathrm{B})\left[\underset{\text{Deactivation}}{\overset{\text{Activation}}{\rightleftharpoons}}\left\{\begin{matrix}\text{Energized}\\ \text{molecule}\end{matrix}\right\}\longrightarrow\left\{\begin{matrix}\text{Transition}\\ \text{state}\\ \text{molecule}\end{matrix}\right\}\right]\longrightarrow\text{Products}$$

Energized and transition state molecules are unstable and as such cannot be isolated.

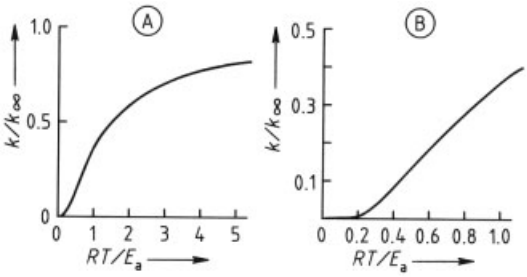

Figure 1. A) Temperature dependence of the reaction rate constant (note: as E_a/R ranges from 5000 to 35 000 K, in practice only the region $RT/E_a < 1$ is of interest); B) Enlargement of region $RT/E_a < 1$

3.1.2. Influence of Temperature

For elementary reactions, the influence of the temperature and composition of the reaction mixture can be represented separately. The influence of temperature is accounted for by the reaction rate constant k, referred to as a constant because it does not depend on the composition of the reaction mixture. The temperature dependence of k is as follows:

$$k = k_\infty \exp\left\{-\frac{E_a}{RT}\right\} \tag{3.1}$$

where E_a is the activation energy of the reaction.

Equation (3.1) is illustrated in Figure 1, where the reaction rate constant is plotted as a function of temperature (both quantities in dimensionless form). The term E_a/R is sometimes referred to as the *activation temperature*. For most reactions the activation energy lies in the range of 40 – 300 kJ/mol; hence the activation temperature E_a/R ranges from 5000 to 35 000 K. Therefore, industrially only the range $RT/E_a < 1$ is of interest: this region is shown enlarged in Figure 1 B. As can be seen from Figure 1 A, although at extreme temperatures k approaches k_∞, in practice it is several orders of magnitude lower.

The temperature increase required for doubling a reaction rate constant $\Delta T_{\mathrm{double}}$ as a function of the temperature of the reaction mixture and activation energy is given in Figure 2. The higher the activation energy, the more temperature sensitive are the reactions. To double the reaction rate at $T = 100\,°\mathrm{C}$, for example, a temperature increase of 3 °C ($E_a = 300$ kJ/mol) –20 °C ($E_a = 50$ kJ/mol) is required.

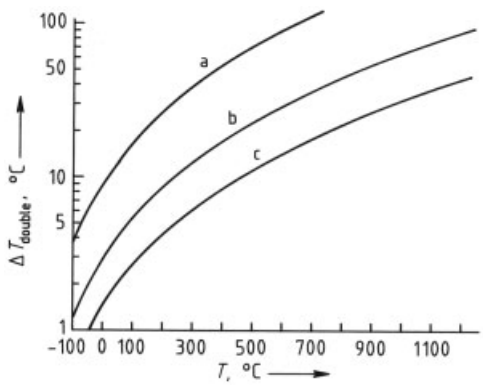

Figure 2. Temperature increase required to double the reaction rate constant ($\Delta T_{\mathrm{double}}$) for different values of the activation energy
a) $E_a = 50$ kJ/mol; b) $E_a = 150$ kJ/mol; c) $E_a = 300$ kJ/mol

3.1.3. Influence of Concentration

The influence of the composition of the reaction mixture on the chemical reaction rate is determined by the *chemical activity* (liquids) or *fugacity* (gases) of the reactants. These quantities can be expressed in relation to each other as

$$a_J = \frac{(\gamma_J P_J)}{RT} \tag{3.2}$$

where a_J is the chemical activity, γ_J the activity coefficient or fugacity coefficient, and P_J the partial pressure of component J. The product $\gamma_J P_J$ is called the fugacity. For ideal reaction mixtures, $\gamma_J = 1$ and the reaction rate is determined by the molar concentration or partial pressure of the reactants.

For elementary reactions and ideal reaction mixtures, the reaction rate is proportional to the concentration of each of the reactants, since the number of molecular collisions per unit time is proportional to the amount of molecules of each of the reactants. For example, for a bimolecular elementary reaction:

$$\mathrm{A} + \mathrm{B} \rightarrow \text{Products}$$

$$|\mathbb{R}| = k C_\mathrm{A} C_\mathrm{B} \tag{3.3}$$

If A is chosen as key reactant, $|\mathbb{R}|$ is expressed in moles of A converted per cubic meter of reaction mixture per second and k has the dimension of cubic meters of reaction mixture per mole of B per second. In this case, for a batch of V m^3 of reaction mixture,

$$|\mathbb{R}| = -\frac{1}{V}\frac{\partial n_\mathrm{A}}{\partial t} = -\frac{\nu_\mathrm{A}}{\nu_\mathrm{B}}\frac{1}{V}\frac{\partial n_\mathrm{B}}{\partial t} = k C_\mathrm{A} C_\mathrm{B} \tag{3.4}$$

or, if volume changes are negligible,

$$|\mathbb{R}| = -\frac{\partial C_\mathrm{A}}{\partial t} = -\frac{\nu_\mathrm{A}}{\nu_\mathrm{B}}\frac{\partial C_\mathrm{B}}{\partial t} = k C_\mathrm{A} C_\mathrm{B} \tag{3.5}$$

The notation given in Equation (3.5) is convenient for reactions carried out in liquids. For gases a more convenient notation is

$$|\mathbb{R}| = k' P_\mathrm{A} P_\mathrm{B} \tag{3.6}$$

with

$$k' = \frac{k}{R^2 T^2} \tag{3.7}$$

and

$$|\mathbb{R}| = -\frac{1}{RT}\frac{\partial P_\mathrm{A}}{\partial t} = -\frac{\nu_\mathrm{A}}{\nu_\mathrm{B}}\frac{1}{RT}\frac{\partial P_\mathrm{B}}{\partial t} = k' P_\mathrm{A} P_\mathrm{B} \tag{3.8}$$

For nonideal elementary reaction mixtures, the concentrations and partial pressures in the above equations must be replaced by the chemical activities and fugacities, respectively.

3.2. Chemical Equilibria

Section 3.1 assumes that the reactions described are irreversible. In practice, reactions are often reversible (i.e., the products can be converted into the reactants again):

$$\mathrm{A} + \mathrm{B} \underset{k_+}{\overset{k_-}{\rightleftharpoons}} \mathrm{P} + \mathrm{Q}$$

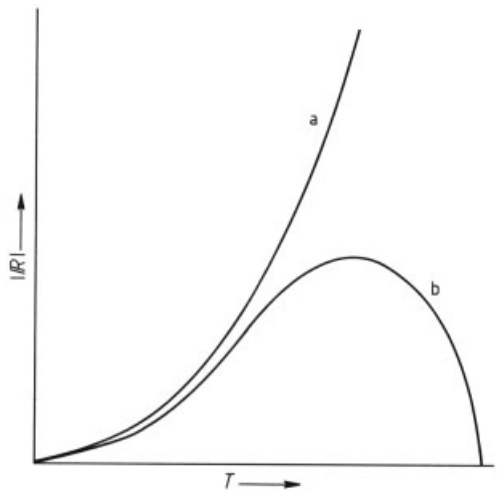

Figure 3. Dependence of reaction rate on temperature for equilibrium reactions
a) Endothermic reversible reaction; b) Exothermic reversible reaction

If both reactions are elementary and the reaction mixture is ideal, the chemical reaction rate is given by

$$|\mathbb{R}| = k_+ C_\mathrm{A} C_\mathrm{B} - k_- C_\mathrm{P} C_\mathrm{Q} \tag{3.9}$$

The activation energies of reaction E_+ and the reverse reaction E_- are related as

$$E_+ = E_- - (-\Delta H_\mathrm{r}) \tag{3.10}$$

where $(-\Delta H_\mathrm{r})$ is the reaction enthalpy [for exothermic reactions $(-\Delta H_\mathrm{r}) > 0$; for endothermic reactions $(-\Delta H_\mathrm{r}) < 0$].

Because of this, the temperature dependences of the reaction rate for exo- and endothermic equilibria are different (see Fig. 3). Initially, the rate increases with increasing temperature for both types of reaction. For *exothermic reactions* where $E_+ < E_-$, the value of k_- increases faster than the value of k_+. Above a certain temperature this can have the net result that —ℝ— in Equation (3.9) decreases with increasing temperature. Thus for exothermic reversible reactions, —ℝ— can, at constant reaction mixture composition, pass through a maximum as a function of temperature; it will become zero for the temperature at which the reaction mixture is at equilibrium (i.e., $k_+ C_A C_B = k_- C_P C_Q$) and negative beyond this value. For *endothermic reactions*, k_+ increases faster with increasing temperature than k_-; hence, —ℝ— continues to increase with increasing temperature, and the equilibrium shifts more and more to the right-hand side.

The temperature at which the maximum rate is found for exothermic reactions can be calculated from

$$T_{\max} = \frac{(-\Delta H_r)}{R}\left[\ln\left(\frac{C_P C_Q}{C_A C_B}\frac{k_- E_-}{k_+ E_+}\right)\right]^{-1} \tag{3.11}$$

Hence, this temperature depends on the composition of the reaction mixture. Therefore, for exothermic equilibrium reactions the temperature should be lowered as more reactant is converted, to keep the rate high. Thus this type of reaction is often carried out in multiple stages (e.g., multiple reactors or multiple zones within a reactor), each with its own temperature, adapted to the reaction mixture composition.

If the reaction rate becomes zero, chemical equilibrium is reached. For this situation the reaction mixture composition must fulfill the equilibrium condition

$$\frac{C_P C_Q}{C_A C_B} = K \tag{3.12}$$

where K is the equilibrium constant. Since —ℝ— = 0, from Equation (3.9), K can be calculated as

$$K = \frac{k_+}{k_-} \tag{3.13}$$

The equilibrium condition (Eq. 3.12) fixes the maximum obtainable relative degree of conversion, ζ_A, for a given initial reaction mixture composition and temperature. For example, if the reaction mixture exhibits ideal behavior and volume changes during the reaction are negligible,

$$C_A = C_{A,0} - C_{A,0}\zeta_A \tag{3.14}$$

$$C_B = C_{B,0} - C_{A,0}\zeta_A \tag{3.15}$$

$$C_P = C_{P,0} + C_{A,0}\zeta_A \tag{3.16}$$

$$C_Q = C_{Q,0} + C_{A,0}\zeta_A \tag{3.17}$$

Substitution in Equation (3.12) yields

$$\zeta_{A,\mathrm{equi}} = \frac{\alpha + \sqrt{\alpha^2 + 4\beta}}{2}$$
$$\alpha = \frac{1}{K-1}\left(\frac{C_{P,0}}{C_{A,0}} + \frac{C_{Q,0}}{C_{A,0}} + K\frac{C_{B,0}}{C_{A,0}} + K\right)$$
$$\beta = \frac{1}{K-1}\left(\frac{C_{P,0}C_{Q,0}}{C_{A,0}^2} - K\frac{C_{B,0}}{C_{A,0}}\right) \tag{3.18}$$

where $\zeta_{A,\mathrm{equi}}$ is the relative degree of conversion at equilibrium conditions. Similar equations can be derived for other equilibria.

3.3. Complex Reaction Schemes

3.3.1. Fundamentals

Only a few examples of chemical reaction rates determined by one elementary kinetic process are known. Conversion of reactants into products is mainly achieved via a scheme of elementary reactions. This scheme as a whole is generally referred to as the *mechanism* of a given overall conversion. A mechanism for the overall reaction A ⟶ P could be

$$\begin{array}{lll|c|c}
(1) & A \xrightarrow{k_1} 2\,X & & \times\frac{1}{2} & \\
(2) & A + X \xrightarrow{k_2} P + X & & \times\frac{1}{2} & \\
(3) & 2\,X \xrightarrow{k_3} P & & \times\frac{1}{2} & + \\
\hline
 & A \longrightarrow P & & &
\end{array}$$

The above mechanism is often found for the decomposition of polymerization initiators: A is the initiator (typically a peroxide or percarbonate); the intermediate X is a free radical that actually begins the growth of the polymer chain, and P is an inactive species. Reaction (1) is called

thermal decomposition, (2) induced decomposition, and (3) termination. The entire scheme is generally referred to as the induced initiator decomposition mechanism.

All of the above reactions are elementary; hence, if the reaction mixture is ideal and volume changes are negligible, the following rate equations are obtained:

$$\frac{\partial C_{\mathrm{A}}}{\partial t} = -k_1 C_{\mathrm{A}} - k_2 C_{\mathrm{A}} C_{\mathrm{X}} \tag{3.19}$$

$$\frac{\partial C_{\mathrm{X}}}{\partial t} = +2k_1 C_{\mathrm{A}} - k_3 C_{\mathrm{X}}^2 \tag{3.20}$$

To determine the concentration of A (and hence the relative degree of conversion) as a function of time, the above differential equations must be solved subject to the initial conditions:

$$t = 0 \rightarrow C_{\mathrm{A}} = C_{\mathrm{A},0} \quad \text{and} \quad C_{\mathrm{X}} = 0 \tag{3.21}$$

Thus, in general a simple kinetic expression is not obtained for complex reaction schemes, whereas for elementary reactions this is always the case. The above differential equations are not easily solved and do not yield a very clear picture.

To overcome these problems, the *pseudo-steady state assumption* can be made. This assumption postulates that the concentration of intermediates in a complex reaction scheme is pseudo-stationary compared to the transient behavior of the reactants and final products (i.e., the derivative of the concentration of intermediates with respect to time can be assumed to be zero). For the above reaction scheme, X is the intermediate and the pseudo-steady state assumption yields

$$\frac{\partial C_{\mathrm{X}}}{\partial t} = +2k_1 C_{\mathrm{A}} - k_3 C_{\mathrm{X}}^2 = 0$$

$$\Rightarrow C_{\mathrm{X}} = \sqrt{\frac{2k_1 C_{\mathrm{A}}}{k_3}} \tag{3.22}$$

Thus, a relatively simple expression is obtained for the intermediate concentration. Substituting the expression for C_{X} in the rate equation for A (Eq. 3.19), gives

$$\frac{\partial C_{\mathrm{A}}}{\partial t} = -k_1 C_{\mathrm{A}} - k_2 \sqrt{\frac{2k_1}{k_3}} C_{\mathrm{A}} \sqrt{C_{\mathrm{A}}} \tag{3.23}$$

This differential equation must be solved subject to

$$t = 0 \Rightarrow C_{\mathrm{A}} = C_{\mathrm{A},0} \tag{3.24}$$

which yields

$$C_{\mathrm{A}} = C_{\mathrm{A},0} \left[\left(1 + \sqrt{\frac{2k_2^2 C_{\mathrm{A},0}}{k_1 k_3}} \right) \exp\left(\frac{k_1 t}{2} \right) - \sqrt{\frac{2k_2^2 C_{\mathrm{A},0}}{k_1 k_3}} \right]^{-2} \tag{3.25}$$

For the pseudo-steady state assumption to be valid, we must have

$$\frac{\partial C_{\mathrm{X}}}{\partial t} << -2k_1 C_{\mathrm{A}} \tag{3.26}$$

or

$$\frac{-1}{2k_1 C_{\mathrm{A}}} \frac{\partial C_{\mathrm{X}}}{\partial t} << 1 \tag{3.27}$$

In analogy to Equation (3.22),

$$C_{\mathrm{X}} \approx \sqrt{\frac{2k_1 C_{\mathrm{A}}}{k_3}} \tag{3.28}$$

hence,

$$\frac{\partial C_{\mathrm{X}}}{\partial t} \approx \sqrt{\frac{k_1}{k_3 C_{\mathrm{A}}}} \frac{\partial C_{\mathrm{A}}}{\partial t} \tag{3.29}$$

Substituting Equation (3.23) gives

$$\frac{-1}{2k_1 C_{\mathrm{A}}} \frac{\partial C_{\mathrm{X}}}{\partial t} = \frac{1}{2k_3} \left(\sqrt{\frac{k_1 k_3}{C_{\mathrm{A}}}} + k_2 \sqrt{2} \right) << 1 \tag{3.30}$$

Thus, for this reaction scheme the pseudo-steady state assumption holds, e.g., if the termination rate is very fast (k_3 is very large). For an induced initiator decomposition (where X is a free radical), the termination reaction occurs between two free radicals. In general these reactions are very fast; hence the pseudo-steady state assumption is valid.

The rate equation for A (Eq. 3.23) was derived from the mechanism of the reaction. Most reactions are overall reactions; the exact mechanism behind them is often unknown, which means that kinetic data cannot be fitted to the exact rate equation. The chemical reaction rate can often be approximated by a so-called *power law equation*:

$$|\mathbb{R}| = k C_{\mathrm{A}}^n \tag{3.31}$$

If Equation (3.31) is used, the reaction is said to be of the order n with respect to A. The orders do not have to be whole numbers; thus, from Equation (3.23), for the discussed mechanism n can be 1 and 1.5, depending on the region covered by the kinetic data (temperature, concentration $C_{A,0}$). Also, if kinetic data are collected in a limited area where both right-hand terms in Equation (3.23) are important, n can take any value between 1 and 1.5.

For multimolecular overall reactions power laws can also be used; for example, for the bimolecular overall reaction

$$\nu_A\,A + \nu_B\,B \longrightarrow \text{Products}$$

the following equation can be written:

$$|\mathbb{R}| = k C_A^n C_B^m \tag{3.32}$$

The reaction is of order n with respect to A and of order m with respect to B; the total order of the reaction is $(n+m)$. Again, the orders n and m can be any numbers, not necessarily positive; negative values can also be found.

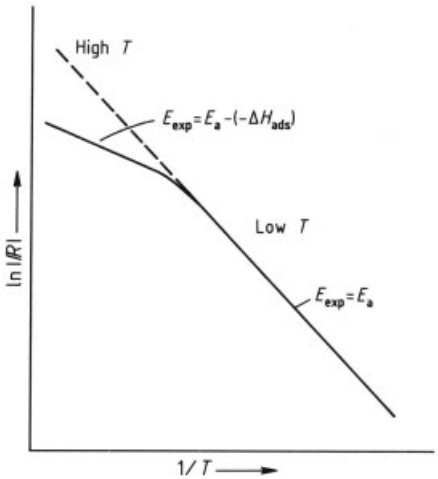

Figure 4. Arrhenius curve (i.e., ln — $\mathbb{R}$ — versus $1/T$) for monomolecular Langmuir kinetics

Hence, different sets of kinetic data, measured for different temperature or concentration ranges, yield different values for n and m. Therefore, if kinetic data or rate equations are obtained (e.g., from the literature), one should always ascertain that the region to which the kinetic data will be applied has indeed been explored experimentally. Extrapolation of kinetic data is generally impossible or very dangerous at least.

The principles discussed here can be used for any kinetic scheme in general. This will be illustrated further with some overall reactions often encountered in practice.

3.3.2. Polymerization Reactions

An ideal free-radical polymerization is studied as an example. The kinetic scheme of such a polymerization is given by

Initiation

$$\mathrm{I} \xrightarrow{k_{i,1}} 2\,\mathrm{R}^{\bullet}$$

$$\mathrm{R}^{\bullet} + \mathrm{M} \xrightarrow{k_{i,2}} \mathrm{P}_N^{\bullet}$$

Propagation

$$\mathrm{P}_N^{\bullet} + \mathrm{M} \xrightarrow{k_p} \mathrm{P}_{N+1}^{\bullet}$$

Termination

$$\mathrm{P}_M^{\bullet} + \mathrm{P}_N^{\bullet} \xrightarrow{k_t} \mathrm{P}_M + \mathrm{P}_N \quad \text{or} \quad \mathrm{P}_{M+N}$$

Transfer

$$\mathrm{P}_N^{\bullet} + \mathrm{T} \xrightarrow{k_s} \mathrm{P}_N + \mathrm{T}^{\bullet}$$

where I is the initiator, $\mathrm{R}^{\bullet}$ the free radical formed from the initiator; M a monomer unit; $\mathrm{P}_N^{\bullet}$ the growing, free-radical polymer chain; P_N the terminated, unreactive polymer chain; and T a chain-transfer agent (e.g., a polymer chain or monomer).

Induced initiator decomposition is not considered in the above scheme. To illustrate the pseudo-steady state assumption for the above scheme, the conversion – time curve is calculated and it will be investigated when the pseudo-steady state assumption holds.

3.3.2.1. Conversion – Time Curve

If the transfer reactions are neglected, the rate equations for the above scheme are

$$\frac{\partial C_I}{\partial t} = -k_{i,1} C_I \tag{3.33}$$

$$\frac{\partial C_M}{\partial t} = -k_{i,2} C_{R^{\cdot}} C_M - k_P C_{P^{\cdot}} C_M \tag{3.34}$$

$$\frac{\partial C_{R^{\cdot}}}{\partial t} = -k_{i,2} C_{R^{\cdot}} C_M + 2 f k_{i,1} C_I \tag{3.35}$$

$$\frac{\partial C_{\mathrm{P}^{\cdot}}}{\partial t} = -k_{\mathrm{t}} C_{\mathrm{P}^{\cdot}}^2 + k_{\mathrm{i},2} C_{\mathrm{R}^{\cdot}} C_{\mathrm{M}} \tag{3.36}$$

The parameter f is the initiator efficiency: if $f = 1$, all initiator decomposes to active radicals; for $f = 0$, no active radicals are formed from the initiator.

The aim is to arrive at a kinetic expression for the consumption of monomer, hence, the production of polymer. To achieve this, a pseudosteady state is assumed for the intermediate radicals $R^{\bullet}$ and $P^{\bullet}$. Pseudo-steady state for $R^{\bullet}$ yields

$$\frac{\partial C_{\mathrm{R}^{\cdot}}}{\partial t} = -k_{\mathrm{i},2} C_{\mathrm{R}^{\cdot}} C_{\mathrm{M}} + 2 f k_{\mathrm{i},1} C_{\mathrm{I}} = 0$$
$$\Rightarrow k_{\mathrm{i},2} C_{\mathrm{R}^{\cdot}} C_{\mathrm{M}} = 2 f k_{\mathrm{i},1} C_{\mathrm{I}} \tag{3.37}$$

Substitution of this equation into the rate equation for $P^{\bullet}$ (Eq. 3.36) gives

$$\frac{\partial C_{\mathrm{P}^{\cdot}}}{\partial t} = -k_{\mathrm{t}} C_{\mathrm{P}^{\cdot}}^2 + k_{\mathrm{i},2} C_{\mathrm{R}^{\cdot}} C_{\mathrm{M}}$$
$$= -k_{\mathrm{t}} C_{\mathrm{P}^{\cdot}}^2 + 2 f k_{\mathrm{i},1} C_{\mathrm{I}} \tag{3.38}$$

if a pseudo-steady state is assumed for $P^{\bullet}$,

$$\frac{\partial C_{\mathrm{P}^{\cdot}}}{\partial t} = -k_{\mathrm{t}} C_{\mathrm{P}^{\cdot}}^2 + 2 f k_{\mathrm{i},1} C_{\mathrm{I}} = 0$$
$$\Rightarrow C_{\mathrm{P}^{\cdot}} = \sqrt{\frac{2 f k_{\mathrm{i},1} C_{\mathrm{I}}}{k_{\mathrm{t}}}} \tag{3.39}$$

Substituting the terms $k_{\mathrm{i},2}\, C_{\mathrm{R}^{\bullet}}\, C_{\mathrm{M}}$ (Eq. 3.37) and $C_{\mathrm{P}^{\bullet}}$ (Eq. 3.39) in the rate equation for M (Eq. 3.34) yields

$$\frac{\partial C_{\mathrm{M}}}{\partial t} = -2 f k_{\mathrm{i},1} C_{\mathrm{I}} - k_{\mathrm{p}} \sqrt{\frac{2 f k_{\mathrm{i},1}}{k_{\mathrm{t}}}} \sqrt{C_{\mathrm{I}}} C_{\mathrm{M}} \tag{3.40}$$

If much less monomer is consumed in the initiation than in the propagation (as is the case, given the high molar masses of polymers) then

$$k_{i,2} C_{\mathrm{R}^{\cdot}} C_{\mathrm{M}} << k_{\mathrm{p}} C_{\mathrm{P}^{\cdot}} C_{\mathrm{M}}$$
$$\Rightarrow k_{\mathrm{i},1} C_{\mathrm{I}} << k_{\mathrm{p}} \sqrt{\frac{2 f k_{\mathrm{i},1}}{k_{\mathrm{t}}}} \sqrt{C_{\mathrm{I}}} C_{\mathrm{M}} \tag{3.41}$$

hence the rate equation for M becomes

$$\frac{\partial C_{\mathrm{M}}}{\partial t} = -k_{\mathrm{p}} \sqrt{\frac{2 f k_{\mathrm{i},1}}{k_{\mathrm{t}}}} \sqrt{C_{\mathrm{I}}} C_{\mathrm{M}} \tag{3.42}$$

In the literature, this equation is generally referred to as "ideal kinetics" for radical polymerization.

The initiator concentration as a function of time follows from Equation (3.33)

$$C_{\mathrm{I}} = C_{\mathrm{I},0} \exp\left(-k_{\mathrm{i},1} t\right) \tag{3.43}$$

Substitution in Equation (3.42) gives

$$\frac{\partial \zeta_{\mathrm{M}}}{\partial t} = \boldsymbol{k} \exp\left(-\frac{k_{\mathrm{i},1} t}{2}\right)(1 - \zeta_{\mathrm{M}}) \tag{3.44}$$

with ζ_{M} the relative degree of conversion of monomer:

$$\zeta_{\mathrm{M}} = \frac{C_{\mathrm{M},0} - C_{\mathrm{M}}}{C_{\mathrm{M},0}} \tag{3.45}$$

and $\boldsymbol{k}$ a constant defined according to

$$\boldsymbol{k} = k_{\mathrm{p}} \sqrt{\frac{2 f k_{\mathrm{i},1}}{k_{\mathrm{t}}}} \sqrt{C_{\mathrm{I},0}} \tag{3.46}$$

Equation (3.44) must be solved subject to the initial condition:

$$t = 0 \Rightarrow \zeta_{\mathrm{M}} = 0 \tag{3.47}$$

The solution to this equation is

$$\zeta_{\mathrm{M}} = 1 - \exp\left\{\frac{2\boldsymbol{k}}{k_{\mathrm{i},1}}\left[\exp\left(-\frac{k_{\mathrm{i},1} t}{2}\right) - 1\right]\right\} \tag{3.48}$$

This is a conversion time relation as it is typically found for an ideal free-radical polymerization (i.e., if macrokinetic phenomena are insignificant). In practice, in polymerizations these phenomena do become important for higher conversions (e.g., gel and glass effect) and therefore Equation (3.48) is of limited value only. More detailed discussions including macrokinetics specific for polymerizations are given in [9], [10]. From Equation (3.48), the maximum conversion is obtained for $t \to \infty$:

$$\zeta_{\mathrm{M,de}} = 1 - \exp\left[-2 k_{\mathrm{p}} \sqrt{\frac{2 f}{k_{\mathrm{t}} k_{\mathrm{i},1}}} \sqrt{C_{\mathrm{I},0}}\right] \tag{3.49}$$

This maximum conversion is less than unity because all the initiator is generally decomposed before all the monomer has had an opportunity to react. The fact that 100 % conversion cannot be reached due to initiator decomposition is commonly referred to as *dead-end polymerization*.

For $k_{\mathrm{i},1}\, t < 0.2$, Equation (3.48) can be approximated by

$$\zeta_{\mathrm{M}} = 1 - \exp\left(-\boldsymbol{k} t\right) \tag{3.50}$$

with $\boldsymbol{k}$ defined as in Equation (3.46). Hence, simple first-order kinetics are obtained; this is typical for the case in which thermal decomposition of the initiator can be neglected.

3.3.2.2. Radical Intermediates Concentration

To investigate the validity of the pseudo-steady state assumption the polymer chain radical concentration $C_{P\bullet}$ is calculated as a function of time. Initiator decomposition is assumed to be negligible, and for R· the pseudo-steady state assumption is taken as valid. Hence,

$$C_I = C_{I,0} \neq \text{function}\ (t) \tag{3.51}$$

and since Equation (3.37) holds,

$$k_{i,2} C_{R\cdot} C_M = 2 f k_{i,1} C_{I,0} \tag{3.52}$$

Substitution in the rate equation for P· (see Eq. 3.36) yields

$$\frac{6 C_{P\cdot}}{\partial t} = -k_t C_{P\cdot}^2 + 2 f k_{i,1} C_{I,0} \tag{3.53}$$

This differential equation must be solved subject to the initial condition

$$t = 0 \Rightarrow C_{P\cdot} = 0 \tag{3.54}$$

which gives the following solution:

$$C_{P\cdot} = \sqrt{\frac{2 f k_{i,1} C_{I,0}}{k_t}} \cdot \tanh\left(t\sqrt{2 f k_{i,1} k_t C_{I,0}}\right) \tag{3.55}$$

The steady-state concentration of polymer chain radicals, $C_{P\bullet,\text{steady}}$, is found for $t \to \infty$:

$$C_{P\cdot,\text{steady}} = \sqrt{\frac{2 f k_{i,1} C_{I,0}}{k_t}} \tag{3.56}$$

thus Equation (3.55) can be written as

$$\frac{C_{P\cdot}}{C_{P\cdot,\text{steady}}} = \tanh\left(t\sqrt{2 f k_{i,1} k_t C_{I,0}}\right) \tag{3.57}$$

Let $t_1\%$ be the time within which 99 % of the polymer chain radicals have reached the steady state:

$$\frac{C_{P\cdot}}{C_{P\cdot,\text{steady}}} = 0.99 \tag{3.58}$$

Then according to Equation (3.57),

$$t_1\% = \frac{\text{arctanh}\ (0.99)}{\sqrt{2 f k_{i,1} k_t C_{I,0}}} \frac{1.87}{\sqrt{f k_{i,1} k_t C_{I,0}}} \tag{3.59}$$

The values of $C_{P\bullet,\text{steady}}$ and $t_1\%$ are calculated below for the polymerization of methyl methacrylate with two specific initiators: bicyclohexylperoxybicarbonate (CPC) and 2,2′-azobisisobutyronitrile (ABI), respectively. The following data have been used:

$C_{i,0} = 5 \times 10^{-3}\text{mol/L}$
$k_{i,1}$ (CPC)
$= 1.37 \times 10^{16} \times \exp\left[-126 \times 10^3/(RT)\right] \text{s}^{-1}$
$k_{i,1}$ (ABI)
$= 1.58 \times 10^{15} \times \exp\left[-129 \times 10^3/(RT)\right] \text{s}^{-1}$
$k_t = 1.94 \times 10^{12} \times \exp\left[-29.3 \times 10^3/(RT)\right]$
$f = 0.5\,\text{L}\,\text{mol}^{-1}\text{s}^{-1}$

The values of $C_{P\bullet,\text{steady}}$ are

	40 °C	80 °C
CPC	5×10^{-8} mol/L	4×10^{-7} mol/L
ABI	1×10^{-8} mol/L	9×10^{-8} mol/L

and for $t_1\%$,

	40 °C	80 °C
CPC	2 s	0.07 s
ABI	11 s	0.4 s

The above time intervals show that the radical pseudo-steady state is reached in a matter of seconds or even fractions of a second. Thus, since residence times for polymerization are typically on the order of minutes to hours, pseudo-steady state can indeed be assumed.

3.3.3. Heterogeneous Catalytic Reactions

Heterogeneous catalytic reactions are reactions that take place on the surface of a solid catalyst. The following phenomena are significant in determining the microkinetics of these reactions:

1) Adsorption of the reactants
2) Reaction of the adsorbed reactants (e.g., with components in the gas phase or other adsorbed species)
3) Desorption of the products

The reaction itself can be monomolecular or bimolecular. Both situations are discussed below.

3.3.3.1. Monomolecular Reactions

A general mechanism for heterogeneously catalyzed monomolecular reactions is

$$A_{(gas)} \underset{k_a}{\overset{k_{-a}}{\rightleftharpoons}} A_{(ads)} \qquad \text{Adsorption}$$

$$\xrightarrow{k_r} P_{(ads)} + Q_{(gas)} \qquad \text{Reaction}$$

$$\xrightarrow{k_d} P_{(gas)} + Q_{(gas)} \qquad \text{Desorption}$$

A kinetic expression for the above reaction can be obtained by assuming a pseudo-steady state for the adsorbed species $A_{(ads)}$ and $P_{(ads)}$. A measure of the concentrations of these species is the *degree of occupation*, θ_A and θ_P, respectively: θ_A and θ_P are defined such that they range from zero (no active surface area covered at all) to unity (all active surface area covered). Applying the pseudo-steady state assumption to the rate equations of the adsorbed species gives

$$\frac{\partial \theta_A}{\partial t} \sim k_a P_A (1-\theta_A-\theta_P) - k_{-a}\theta_A - k_r\theta_A = 0 \quad (3.60)$$

$$\frac{\partial \theta_P}{\partial t} \sim k_r\theta_A - k_d\theta_P = 0 \quad (3.61)$$

The overall chemical reaction rate is obtained from

$$|\mathbb{R}| = -\frac{\partial C_A}{\partial t} = \frac{\partial C_P}{\partial t} = k_d\theta_P \quad (3.62)$$

Combining Equations (3.60), (3.61), and (3.62), and substituting expressions for θ_A and θ_P, gives

$$|\mathbb{R}| = \frac{k_r \frac{k_a}{k_{-a}} P_A}{\left(1+\frac{k_r}{k_{-a}}\right) + \left(1+\frac{k_r}{k_d}\right)\frac{k_a}{k_{-a}} P_A} \quad (3.63)$$

which is generally written as

$$|\mathbb{R}| = k\frac{K_A P_A}{1+K_A P_A} \quad (3.64)$$

k and K_A are lumped parameters, which can be obtained by comparing Equations (3.63) and (3.64). The above expression is referred to as *Langmuir kinetics*. For low partial pressures of A the rate is first order in A; for high partial pressures it becomes zero order because $K_A P_A >> 1$.

If the adsorption and desorption rates in the above scheme are much higher than the rate of the surface reaction (i.e., k_a, k_{-a}, $k_d >> k_r$), Equation (3.63) can be approximated by

$$|\mathbb{R}| \approx k_r \frac{\frac{k_a}{k_{-a}} P_A}{1+\frac{k_a}{k_{-a}} P_A} = k_r \frac{b_A P_A}{1+b_A P_A} \quad (3.65)$$

with

$$b_A = \frac{k_a}{k_{-a}} = b_{A,\infty} \exp\left[\frac{(-\Delta H_{ads})}{RT}\right] \quad (3.66)$$

and $(-\Delta H_{ads})$ the adsorption enthalpy of A on the active catalyst surface. Since the adsorption of a given component is an exothermic process, $(-\Delta H_{ads})$ is positive. An Arrhenius plot (i.e., the logarithm of the reaction rate versus the reciprocal absolute temperature) for this situation is given in Figure 4. For low temperatures, b_A is large (θ_A close to unity) and the rate is zero order:

$$|\mathbb{R}| \approx k_r = k_{r,\infty} \exp\left(-\frac{E_a}{RT}\right) \quad (3.67)$$

Hence, for low temperatures the energy of activation equals E_a and is relatively large. As the temperature increases, the value of b_A (or θ_A) decreases and the reaction becomes first order:

$$|\mathbb{R}| \approx k_r b_A P_A = k_{r,\infty} b_{A,\infty} \exp\left[-\frac{E_a - (-\Delta H_{ads})}{RT}\right] P_A \quad (3.68)$$

So, in this regime, a smaller, apparent activation energy is found that is equal to $E_a - (-\Delta H_{ads})$.

3.3.3.2. Bimolecular Reactions

For bimolecular, heterogeneously catalyzed reactions, two mechanisms can be distinguished:

1) *Eley–Rideal Mechanism*. One of the reactants is adsorbed on the catalyst surface, the other reacts from the gas phase :

$$A_{(gas)} \underset{k_a}{\overset{k_{-a}}{\rightleftharpoons}} A_{(ads)}$$

$$A_{(ads)} + B_{(gas)} \underset{k_r}{\overset{k_{-r}}{\rightleftharpoons}} P_{(ads)}$$

$$P_{(ads)} \xrightarrow{k_d} P_{(gas)}$$

2) *Langmuir – Hinshelwood Mechanism.* Both reactants are adsorbed and react as adsorbed species :

$$A_{(gas)} \underset{k_a}{\overset{k_{-a}}{\rightleftharpoons}} A_{(ads)}$$

$$B_{(gas)} \underset{k_b}{\overset{k_{-b}}{\rightleftharpoons}} B_{(ads)}$$

$$A_{(ads)} + B_{(ads)} \underset{k_r}{\overset{k_{-r}}{\rightleftharpoons}} P_{(ads)}$$

$$P_{(ads)} \xrightarrow{k_d} P_{(gas)}$$

Both mechanisms are illustrated in Figure 5. A description of these mechanisms is given below.

The pseudo-steady state assumption for the adsorbed species in the *Eley – Rideal mechanism* yields

$$\frac{\partial \theta_A}{\partial t} \sim k_a P_A (1-\theta_A-\theta_P) - k_{-a}\theta_A - k_r P_B \theta_A + k_{-r}\theta_P = 0 \tag{3.69}$$

$$\frac{\partial \theta_P}{\partial t} \sim k_r P_B \theta_A - k_{-r}\theta_P - k_d \theta_P = 0 \tag{3.70}$$

The overall chemical reaction rate is now obtained from

$$|\mathbb{R}| = k_r P_B \theta_A - k_{-r}\theta_P \tag{3.71}$$

Combining the above equations gives

$$|\mathbb{R}| = \frac{k_r \frac{k_a}{k_{-a}} P_A P_B}{1+\frac{k_{-r}}{k_d}+\frac{k_a}{k_{-a}}\left(1+\frac{k_{-r}}{k_d}\right)P_A+\frac{k_r}{k_{-a}}P_B+\frac{k_a k_r}{k_{-a}k_d}P_A P_B} \tag{3.72}$$

From this formula, several limiting regimes can again be obtained: for example, if the adsorption equilibrium rates are still much higher than the rate of the surface reaction (i.e., k_a, k_{-a}, $k_d >> k_r$, k_{-r}), Equation (3.71) can be approximated by

$$|\mathbb{R}| = k_r \frac{b_A P_A}{1+b_A P_A} P_B \tag{3.73}$$

with b_A again

$$b_A = \frac{k_a}{k_{-a}} \tag{3.74}$$

From the *Langmuir – Hinshelwood mechanism* a formula similar to Equation (3.72) can be derived. This formula, however, is rather complex and bulky. Therefore, here only the limit is discussed, where adsorption and desorption rates are again much higher than the rate of the surface reaction (i.e., k_a, k_{-a}, k_b, k_{-b}, $k_d >> k_r$, k_{-r}). For this situation,

$$|\mathbb{R}| = k_r \frac{b_A b_B P_A P_B}{(1+b_A P_A + b_B P_B)^2} \tag{3.75}$$

with b_A according to Equation (3.74) and

$$b_B = \frac{k_b}{k_{-b}} \tag{3.76}$$

If the partial pressure of one of the components (e.g., A) is increased, the reaction rate does not necessarily have to increase: for a certain value of P_A the maximum rate is obtained; above this pressure the rate decreases with increasing P_A. The explanation is that for high P_A values, adsorption of A is so strong that it hampers adsorption of reactant B and thus retards the surface reaction. Consequently, negative reaction orders can also be found: for both components the reaction order is between -1 and 1.

Examples of systems exhibiting Langmuir – Hinshelwood-type behavior are the oxidation of CO on platinum and palladium (order for CO is -1, and for O_2 between 0 and 1); the hydrogenation of C_2H_4 on nickel (order for C_2H_4 is between -1 and 0, and for H_2 between 0 and 1); and the reaction $N_2O + H_2 \longrightarrow N_2 + H_2O$ on platinum and gold (order for N_2O is between 0 and 1, and for H_2 -1).

3.3.4. Biochemical Reactions

Biochemical reactions typically occur according to the following scheme:

$$A \xrightarrow{k_P} P \begin{cases} \xrightarrow{k_X} X \\ \rightarrow Q \end{cases}$$

The stoichiometric coefficients are considered below.

A is called the substrate and P represents the enzymes or living cells. The substrate is consumed, the cells grow, and a desired product Q is formed. However, the enzymes or cells also decompose or die according to the reaction

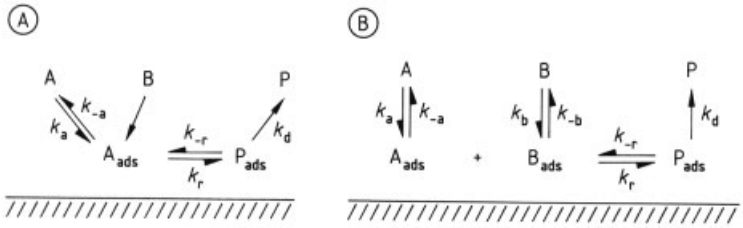

Figure 5. Elementary steps involved in bimolecular, heterogeneously catalyzed reactions
A) According to Eley – Rideal; B) According to Langmuir – Hinshelwood

P $\longrightarrow$ X. In wastewater treatment, A represents the water contaminants and Q represents CO_2 and H_2O. In ethanol production, A represents the sugar, P the yeast cells, X the dead cells, and Q the ethanol, CO_2, and H_2O.

In biochemical reactions where a living microorganism P (a type of bacteria or cell) is fed with a substrate A, the cell grows as a consequence, and the number of cells increases. A possible mechanism to describe this phenomenon was first postulated by MICHAELIS and MENTEN [11]. Based on their mechanism, a kinetic equation was proposed by MONOD [12]. Here, this kinetic equation is given without going into details of the exact mechanism. If the dying or decomposition rate $k_X\, C_P$ is not considered, the expression is

$$|\mathbb{R}_P| = \frac{\partial C_P}{\partial t} = k_P \frac{C_A}{K_S + C_A} C_P \tag{3.77}$$

Notice the resemblance to Eley – Rideal kinetics in the case where the adsorption and desorption rates are much higher than the rate of the surface reaction (see Eq. 3.73).

In Equation (3.77), K_S is often called the *saturation constant*: this is the concentration C_A at which — $\mathbb{R}_P$ — reaches half of its maximum value for a given value of C_P. The cell growth rate varies between zero and first order with respect to substrate concentration, so two extreme kinetic regimes can be distinguished:

1) The *uninhibited growth phase* ($C_A >> K_S$), where more than sufficient feed is available for the cells to reproduce in accordance with the number of cells available or — $\mathbb{R}_P$ — $= k_P\, C_P$
2) The *feed-limited growth phase* ($C_A << K_S$), with — $\mathbb{R}_P$ — $= k_P\, C_P\, (C_A/K_S)$, where insufficient feed is available and the cells will grow in accordance with the feed and the number of cells available

Concentrations and K_S are usually expressed in mass per unit volume (kg/m^3) because of the difficulty of determining molar concentrations in biotechnology. The dimension of k_P is per unit time (s^{-1}).

The substrate consumption rate equals

$$|\mathbb{R}| = -\frac{\partial C_A}{\partial t} = \frac{k_P}{\nu'} \frac{C_A}{K_A + C_A} C_P \tag{3.78}$$

In this equation, ν' is a kind of stoichiometric coefficient, which is often called the *yield factor*. It is expressed as mass of cells formed per mass of substrate consumed (kg/kg).

Monod kinetics are *autocatalytic*. The cells P have to be present to start reproduction of the cells and, as long as there is sufficient substrate available, the more cells have already been produced, the faster they will reproduce, i.e., the higher — $\mathbb{R}_P$ —.

Many extensions of Monod kinetics have been proposed in the past. A survey is given in [13]. One simple extension is, of course, inclusion of the fact that part of the microorganisms die with a rate of $k_X\, C_P$ as discussed above. Then,

$$|\mathbb{R}_P| = \frac{\partial C_P}{\partial t} = k_P \frac{C_A}{K_S + C_A} C_P - k_X C_P \tag{3.79}$$

The rate expressions for biochemical reactions are very sensitive to temperature and to the pH of the reaction mixture, as illustrated in Figure 6. An optimum temperature exists at which microorganisms grow most comfortably; below this temperature the microorganisms are relatively inactive, and above this temperature degradation sets in rapidly. Many microorganisms are optimal at $pH_{opt} = 6.5 \pm 1$, the exact value of the pH depending on the type of microorganism. Typically at $pH_{opt} \pm (1.0 - 1.5)$,

the activity has already been reduced to half of its value at pH_{opt}.

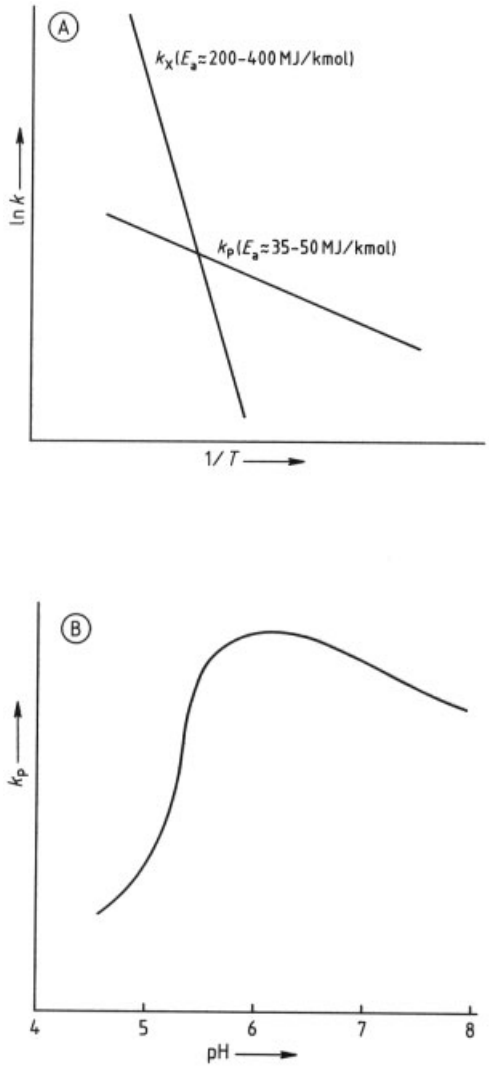

Figure 6. Effect of temperature (A) and pH (B) on cell growth in biochemical reactions

4. Macrokinetics: Single-Phase Systems

4.1. Introduction

A significant macrokinetic phenomenon for single-phase systems is the extent of *micromixing* (i.e., the extent to which the fluid travels as lumps of molecules). To visualize the difference between a micromixed and nonmicromixed fluid, one could think of a viscous fluid partially mixed with a colored fluid with the same properties. On a macroscale the fluid may seem of a uniform color, but viewed through a microscope, streaks and isolated domains of more or less intensive color might be seen (Fig. 7). If such nonhomogeneous domains exist, the fluid is called a (partially) *segregated fluid* or *macrofluid*. An emulsion of oil in water, where the oil is segregated in clearly distinguishable domains (oil droplets), is an extreme example of a macrofluid. Only if the mixture remains uniform on an ever-decreasing scale can it be considered a *microfluid*.

Based on the work of Kolmogoroff [14], [15], Batchelor [16], Corrsin [17], and Rosenweig [18], a characteristic dimension of the agglomerates in the fluid can be estimated as a function of hydrodynamic factors (e.g., stirrer power input) and fluid properties (e.g., viscosity).

The size of these agglomerates alone, however, is not sufficient to discriminate between a macro- or a microfluid. Generally, the effect of the state of segregation can be judged only in relation to conversion rate and reactor residence time: the same fluid at the same flow conditions may behave like both a macro- and a microfluid, depending on rate and residence time. If the reaction rate is low and thus reaction time is long compared to residence time or to the life span of an agglomerate in the reactor, the overall conversion can be calculated by assuming the fluid to behave like a microfluid. However, in the case of an extremely short reaction time or fast reaction and a short residence time, the life span of the agglomerate may be long in comparison, and the conversion can be found by assuming that the fluid behaves effectively as a macrofluid.

Therefore, to be able to assess the impact of micromixing, *macromixing* must be discussed. Macromixing means mixing on the scale of the whole reactor, as opposed to micromixing, which applies to the microscopic scale of the agglomerates (e.g., oil droplets). In Chapter 1, three scales are defined on which to consider the phenomena in a reacting system: (1) the scale of the molecules (microkinetics); (2) the scale of an eddy or dispersion, e.g., droplet, bubble, particle (macrokinetics); and (3) the scale of the

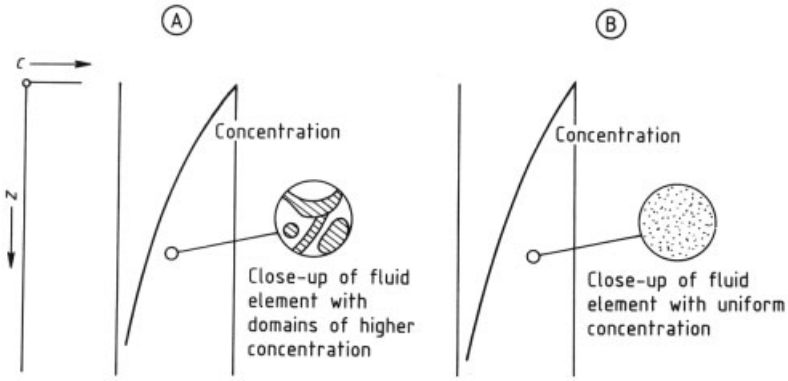

Figure 7. A) Illustration of a macrofluid (sometimes called partially segregated fluid); B) Illustration of a microfluid

reactor. Macromixing belongs to the third scale, whereas micromixing belongs to the second.

Since, in this chapter, the macrokinetic scale is dealt with, strictly speaking macromixing should not be considered here because it does not belong to this scale. However, as outlined above, to be able to understand the impact of micromixing one must have some knowledge of macromixing. Therefore, macromixing and the sequence of mixing are discussed below. For further references, see [2], [8], [19].

4.2. Macromixing

4.2.1. Residence-Time Distribution (RTD)

Because in many cases the hydrodynamic behavior of fluid flow through a real reactor cannot be predicted or even described, a need often exists for a simplified model with experimental data on fluid flow. A very powerful concept is the residence-time distribution, which is essentially a statistical approach for description of fluid flow. A typical element of the fluid is considered as it enters the reactor. The probability of leaving the reactor after a certain residence time there is expressed as a *residence-time distribution* (RTD). The idea was first discussed by DANCKWERTS [20].

The many elements of a fluid entering the reactor at a certain moment $t = 0$ generally do not reside in the reactor for exactly the same period; rather, there is a distribution of residence times. This can be represented in terms of the *frequency distribution*, usually called the E diagram in RTD descriptions (see Fig. 8). The shaded area $E(t)\,\partial t$ represents the fraction of fluid elements that enter the reactor at $t = 0$ and leave between t and $t + \partial t$. Stated otherwise, it represents the chance of a fluid element having a residence time between t and $t + \partial t$. The E diagram thus defined has an area of unity below the curve because all fluid elements have a residence time between 0 and ∞, and hence the fraction of fluid elements with a residence time between 0 and ∞ equals unity. Thus,

$$\int_0^\infty E(t)\,\partial t = 1 \tag{4.1}$$

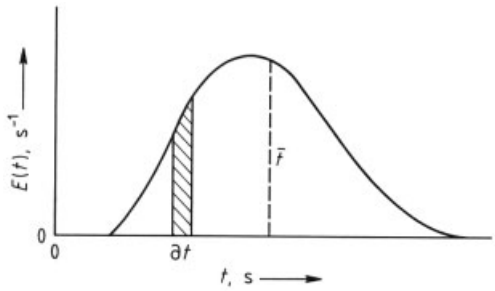

Figure 8. The residence-time distribution (RTD) given by an $E(t)$ diagram
The shaded area $E(t)\,\partial t$ represents the chance of a fluid element having a residence time between t and $t + \partial t$.

To characterize the residence-time distribution in terms of statistical parameters such as the average residence time, spread in residence time, or skewness of distribution, moments are used. The k th moment MO_k around the origin is defined as

$$\mathrm{MO}_k = \int_0^\infty t^k E(t)\,\partial t \tag{4.2}$$

Notice that the zeroth moment around the origin per definition must equal unity (see Eq. 4.1):

$$\mathrm{MO}_0 = \int_0^\infty E(t)\,\partial t = 1 \tag{4.3}$$

The first moment around the origin equals the average residence time $\bar{t}$:

$$\bar{t} = \mathrm{MO}_1 = \int_0^\infty tE(t)\,\partial t \tag{4.4}$$

The k-th central moment MC_k is defined as

$$\mathrm{MC}_k = \int_0^\infty (t-\bar{t})^k\, E(t)\,\partial t \tag{4.5}$$

The first central moment always equals zero, since by definition

$$\mathrm{MC}_1 = \int_0^\infty (t-\bar{t})\, E(t)\,\partial t$$
$$= \int_0^\infty tE(t)\,\partial t - \bar{t}\int_0^\infty E(t)\,\partial t = 0 \tag{4.6}$$

The second central moment is called the variance $\sigma^2{}_t$:

$$\sigma_t^2 = \mathrm{MC}_2 = \int_0^\infty (t-\bar{t})^2\, E(t)\,\partial t$$
$$= \mathrm{MO}_2 - \mathrm{MO}_1^2 \tag{4.7}$$

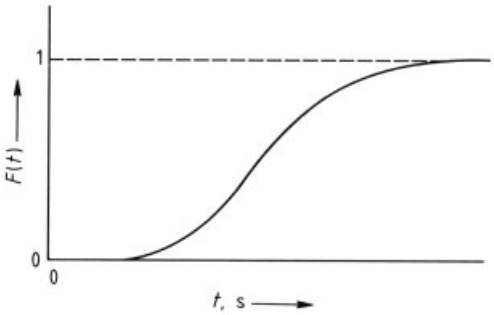

Figure 9. Cumulative RTD function or F diagram $F(t)$ represents the fraction of fluid elements in the vessel exit stream with a residence time shorter than $t \cdot$ s

The spread in the residence time is characterized by the standard deviation σ_t, defined as the square root of the variance:

$$\sigma_t = \sqrt{\sigma_t^2} \tag{4.8}$$

Higher moments, which are used in quantities such as skewness and peakedness of the RTD function, are employed less frequently.

The E diagram is often also presented, by using a reduced time, as $E(\theta)$ versus θ:

$$\theta = \frac{t}{\tau} \tag{4.9}$$

Here τ is the holding time, which for a fluid of constant density can be obtained from

$$\tau = \frac{V}{\varphi_v} \tag{4.10}$$

in which V is the volume accessible for the fluid (e.g., volume of the reaction mixture inside the reactor) and φ_v the fluid volumetric flow rate (m^3/s). The average residence time $\bar{t}$ can differ from the holding time τ, for example, if the fluid is adsorbed on a catalyst, reactor walls, or internals. In the case of adsorption, the average residence time is longer than the holding time. Below, adsorption phenomena are assumed to not play a significant role, in which case

$$\tau = \bar{t} \tag{4.11}$$

For the parameters t and θ, the respective E diagrams are related by

$$E(t) = \frac{1}{\tau}E(t/\tau) = \frac{1}{\tau}E(\theta) \tag{4.12}$$

As in Equation (4.2) for $E(\theta)$, moments around the origin are defined as

$$\mathrm{mo}_k = \int_0^\infty \theta^k E(\theta)\,\partial\theta \tag{4.13}$$

Since "the average dimensionless residence time" equals unity, $\mathrm{mo}_1 = 1$:

$$\mathrm{mo}_1 = \int_0^\infty \theta E(\theta)\,\partial\theta = \int_0^\infty tE\left(\frac{t}{\tau}\right)\partial\frac{t}{\tau}$$
$$= \frac{1}{\tau}\int_0^\infty tE(t)\,\partial t = 1 \tag{4.14}$$

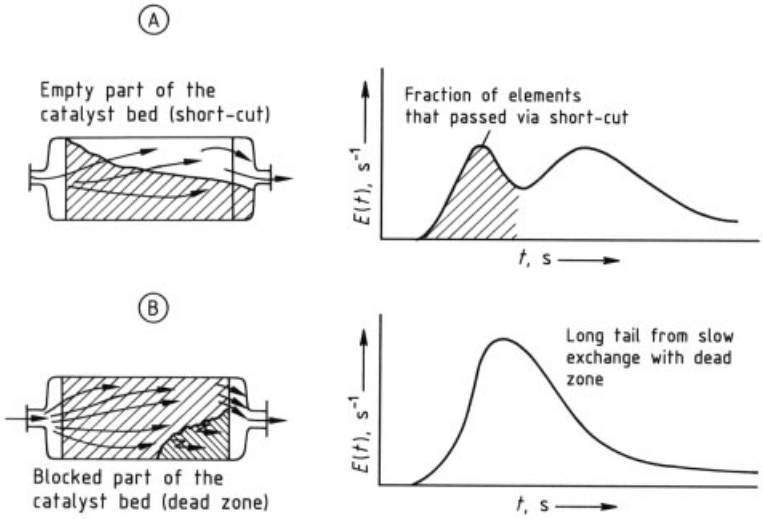

Figure 10. Examples of short-cutting (A) and dead zones (B) with their corresponding appearances in the $E(t)$ diagram

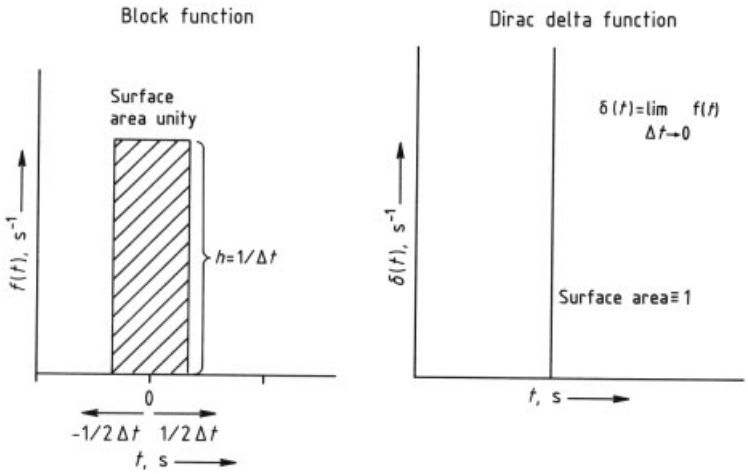

Figure 11. Idealized tracer injection pulse: the δ function or Dirac delta function (right) as an extreme case of a block function (left)

As above, the k-th central moment mc_k is:

$$\mathrm{mc}_k = \int_0^\infty (\theta - 1)^k \, E(\theta) \, \partial\theta \tag{4.15}$$

and the variance is

$$\sigma_\theta^2 = \mathrm{mc}_2 = \int_0^\infty (\theta - 1)^2 \, E(\theta) \, \partial\theta$$
$$= \mathrm{mo}_2 - \mathrm{mo}_1^2 \tag{4.16}$$

A complementary presentation of the residence-time distribution can be made by a cumulative function called the $F(t)$ diagram (see Fig. 9). Here the fraction $F(t)$ of the elements with a residence time shorter than t is plotted. The E and F diagrams are related by

$$F(t) = \int_0^t E(u) \, \partial u \tag{4.17}$$

The functions E and F are equivalent, but some properties are more easily demonstrated with an E diagram, so they are used more frequently.

Experimentally observed RTD functions can be used for identification of reactor malfunctioning. Two cases are illustrated in Fig-

ure 10. Short-cutting is visible in the front part of the E diagram, because a relatively large fraction of the fluid has a very short residence time. If, on the other hand, dead zones are present, a relatively large fraction of the fluid has a very long residence time, which is indicated by an excessively long tail on the E curve. In extreme cases the tail may not be detectable because of very slow renewal of the fluid in the dead zone by the active part of the fluid. In that case the average residence time observed from the E diagram is shorter than $\tau = V/\varphi_v$. If, in contrast, the average residence time observed from the E diagram is much longer than τ, then adsorption phenomena as discussed above must be taken into account.

The general procedure for an RTD determination is that fluid elements flowing into the reactor or the measuring section are marked and detected at the outlet. This marking is performed by changing some property with time. Mostly the concentration of a tracer substance is varied. Such a tracer substance must be easily detectable in low concentration (e.g., via changes in electrical or thermal conductivity, light adsorption, or radioactive radiation). Furthermore the tracer must not have a significant effect on bulk fluid properties (e.g., density, viscosity). Two types of incoming tracer signals are frequently used:

1) A sharp pulse, where at $t = 0$ a certain amount of tracer is injected into the inlet stream in the shortest possible time. Mathematically this is represented by the *Dirac delta function* $\delta(t)$, illustrated in Figure 11 where the delta function is depicted as an extreme case of a block function. An advantage of δ-pulse injection is that after normalization the $E(t)$ diagram is obtained directly as a response (albeit, this holds only if the reactor is a closed system [8]).
2) A step function where at $t = 0$ the incoming concentration is changed from one steady value (usually zero) to another. Mathematically this is represented by the *unit step function* $\mathrm{U}(t)$, also called the *Heaviside step function* (see Fig. 12). After normalization the direct response of the step function is the $F(t)$ diagram (again, for closed systems only).

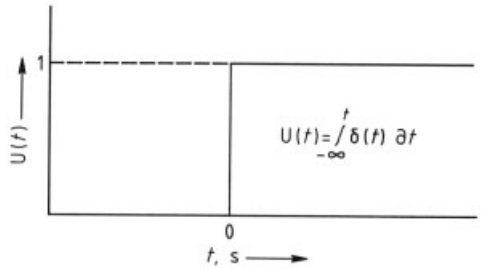

Figure 12. Idealized tracer injection function: unit step function $\mathrm{U}(t)$, also called the Heaviside step function

Other types of incoming tracer signals can also be used (a steady sinusoidal variation or even arbitrary signals) and are discussed in [8].

The RTDs of some model reactors are discussed in the following sections.

4.2.2. Batch Reactor (BR)

Basically a batch reactor is a vessel in which a reaction is carried out, with no material being supplied to or withdrawn from it during the reaction. Hence a batch reactor does not have a residence-time distribution: all elements in the reactor have the same residence times. If an ideal reaction mixture and negligible volume changes are assumed, for a momomolecular reaction the composition inside the batch reactor as a function of time follows from

$$\frac{\partial C_\mathrm{A}}{\partial t} = -|\mathbb{R}_\mathrm{A}| \tag{4.18}$$

where A is the reactant. By introducing the relative degree of conversion ζ_A as

$$\zeta_\mathrm{A} = \frac{C_\mathrm{A}|_{t=0} - C_\mathrm{A}|_t}{C_\mathrm{A}|_{t=0}} \tag{4.19}$$

Equation (4.18) can be written as

$$t = C_\mathrm{A}|_{t=0} \int_0^{\zeta_\mathrm{A}} \frac{\partial \zeta_\mathrm{A}}{|\mathbb{R}_\mathrm{A}|} \tag{4.20}$$

For example, for a first-order reaction where

$$|\mathbb{R}_\mathrm{A}| = kC_\mathrm{A}|_{t=0}(1-\zeta_\mathrm{A}) \tag{4.21}$$

Equation (4.20) yields

$$\zeta_\mathrm{A} = 1 - \mathrm{e}^{-kt} \tag{4.22}$$

For reactions of n th order, similar expressions can be found.

4.2.3. Plug-Flow Reactor (PFR)

In a plug-flow reactor the reaction mixture flows like a plug through a tube; the linear velocity u is the same at every point in a cross section perpendicular to the direction of flow and equals

$$u = \frac{\varphi_v}{S} \tag{4.23}$$

with S the cross-sectional area (m^2) and φ_v the volumetric flow rate of the reaction mixture (m^3/s).

The composition of the reaction mixture depends on the distance z from the inlet point (see Fig. 13). To calculate concentration profiles across the length, a material balance must be applied to a differential volumetric section $S\,\partial z$. applied to a differential volumetric section $S\,\partial z$. For a reactant A this yields (if volume changes during the reaction are negligible)

$$\varphi_v\,\partial C_A = -|\mathbb{R}_A| S\,\partial z \tag{4.24}$$

Introducing the relative degree of conversion ζ_A as

$$\zeta_A = \frac{C_A|_{z=0} - C_A|_z}{C_A|_{z=0}} \tag{4.25}$$

yields

$$\partial\zeta_A = \frac{|\mathbb{R}_A| S}{\varphi_v C_A|_{z=0}}\,\partial z \tag{4.26}$$

In this equation, only — $\mathbb{R}_A$ — depends on ζ_A. The reactor volume $V = SL$ needed to reach a relative degree of conversion equal to $\zeta_{A,L}$ can be calculated from

$$V = \varphi_v C_A|_{z=0} \int_0^{\zeta_{A,L}} \frac{\partial\zeta_A}{|\mathbb{R}_A|} \tag{4.27}$$

By introducing the reactor holding time (or since no adsorption is assumed, the reactor residence time) $\tau = V/\varphi_v$, Equation (4.27) can be written as

$$\tau = C_A|_{z=0} \int_0^{\zeta_{A,L}} \frac{\partial\zeta_A}{|\mathbb{R}_A|} \tag{4.28}$$

The integral on the right-hand side also occurs in calculation of the reaction time of a batch reactor (see Eq. 4.20). Apparently, for constant density, the reaction time for a batch reactor corresponds to the residence time in a plug-flow reactor. From a physical viewpoint this is because in a plug-flow reactor the reaction mixture flows like a plug through a tube. This plug does not mix with other plugs and therefore can be regarded as a "small batch reactor" by itself, with a reaction time equal to the residence time in the tube.

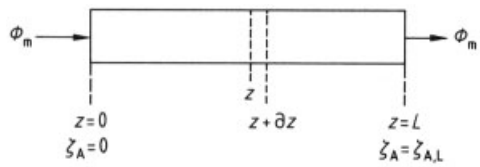

Figure 13. Microscale mass balance for tubular plug-flow reactor with total volume $V = SL$

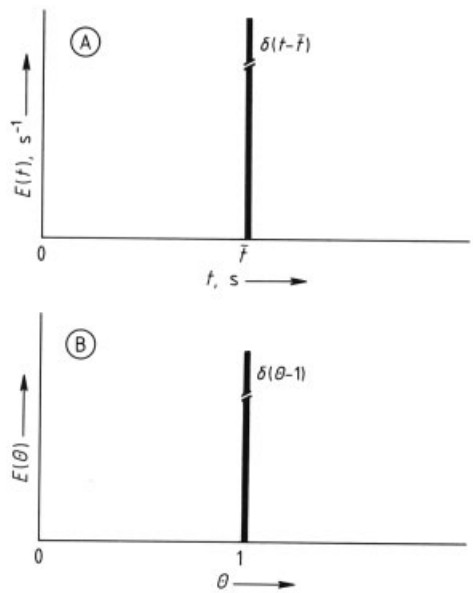

Figure 14. $E(t)$ (A) and $E(\theta)$ (B) diagrams for a plug-flow reactor

Conversion calculations for the plug-flow reactor can be done in a similar manner as for the batch reactor. Thus, for first-order kinetics,

$$|\mathbb{R}_A| = k\,C_A|_{z=0}\,(1-\zeta_A) \tag{4.29}$$

and Equation (4.28) yields

$$\zeta_A = 1 - e^{-k\tau} \tag{4.30}$$

A plug-flow reactor has no spread in residence time because the fluid is supposed to flow like a plug through the reactor. This means that the

input signal can be found back in the outlet with a time delay equal to the residence time. This is shown for the $E(t)$ and $E(\theta)$ diagrams in Figure 14. The RTD variance, therefore, must be zero in a plug-flow reactor:

$$\sigma_t^2 = \int_0^\infty (t-\bar{t})^2 \, E(t) \, \partial t = 0 \tag{4.31}$$

4.2.4. Continuous, Ideally Stirred Tank Reactor (CISTR)

A CISTR is a tank reactor whose contents are assumed to be "ideally mixed," so that conditions are the same throughout the tank and equal to conditions at the outlet (see Fig. 15). A steady-state material balance for the CISTR gives

$$\varphi_v \left(C_{\mathrm{A,in}} - C_{\mathrm{A}}\right) = |\mathbb{R}_{\mathrm{A}}| V \tag{4.32}$$

Introducing the relative degree of conversion ζ_{A}

$$\zeta_{\mathrm{A}} = \frac{C_{\mathrm{A,in}} - C_{\mathrm{A}}}{C_{\mathrm{A,in}}} \tag{4.33}$$

and assuming constant density of the reaction mixture yield

$$\tau = C_{\mathrm{A,in}} \frac{\zeta_{\mathrm{A}}}{|\mathbb{R}_{\mathrm{A}}|} \tag{4.34}$$

(cf. Eqs. 4.20 and 4.28). Again, τ is the holding time, or average residence time if adsorption effects are negligible.

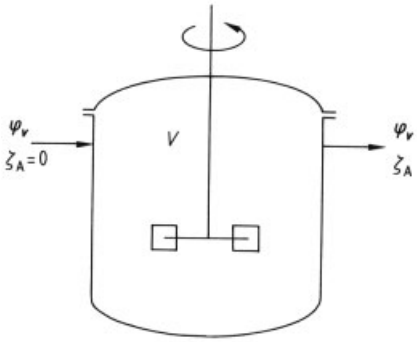

Figure 15. Schematic of a continuously operated, ideally stirred tank reactor (CISTR)

For example, in a first-order reaction where

$$|\mathbb{R}_{\mathrm{A}}| = k C_{\mathrm{A,in}} \left(1 - \zeta_{\mathrm{A}}\right) \tag{4.35}$$

Equation (4.34) yields

$$k\tau = \frac{\zeta_{\mathrm{A}}}{1-\zeta_{\mathrm{A}}} \Leftrightarrow \zeta_{\mathrm{A}} = \frac{k\tau}{1+k\tau} \tag{4.36}$$

The E diagram for a CISTR can be derived from a material balance after an ideal δ-pulse has been injected in the inlet stream. The pulse is immediately distributed over the entire reactor; furthermore, the concentration remains the same throughout the reactor and equals the outlet concentration. Assuming constant density of the reaction mixture gives

$$V \frac{\partial C_{\mathrm{T}}}{\partial t} = -\varphi_v C_{\mathrm{T}} \tag{4.37}$$

where C_{T} is the tracer concentration. Introducing the holding time $\tau = V/\varphi_v$ yields

$$\frac{\partial C_{\mathrm{T}}}{\partial t} = -\frac{C_{\mathrm{T}}}{\tau} \tag{4.38}$$

The initial condition is

$$t = 0 \Rightarrow C_{\mathrm{T}} = \frac{n_{\mathrm{T},0}}{V} \tag{4.39}$$

where $n_{\mathrm{T},0}$ equals the total number of moles of tracer injected in the pulse at $t = 0$. After integration, Equation (4.38) yields

$$C_{\mathrm{T}} = \frac{n_{\mathrm{T},0}}{V} \mathrm{e}^{-t/\tau} = \frac{n_{\mathrm{T},0}}{V} \mathrm{e}^{-\theta} \tag{4.40}$$

To find the E diagram, this equation must be normalized; i.e., the property $\int_0^\infty E(\theta) \, \partial\theta = 1$ must be adhered to. This gives

$$E(\theta) = \frac{\frac{n_{\mathrm{T},0}}{V} \mathrm{e}^{-\theta}}{\int_0^\infty \frac{n_{\mathrm{T},0}}{V} \mathrm{e}^{-\theta} \, \partial\theta} = e^{-\theta} \tag{4.41}$$

The $E(t)$ diagram reads, according to Equation (4.12),

$$E(t) = \frac{\mathrm{e}^{-t/\tau}}{\tau} \tag{4.42}$$

The result is plotted in Figure 16.

Now, because of the spread in the RTD, the variance does not equal zero. Equation (4.7) shows that for a CISTR,

$$\sigma_t^2 = \int_0^\infty (t-\tau)^2 \, E(t) \, \partial t = \tau^2 \tag{4.43}$$

For a detailed treatment of macromixing, see [8]

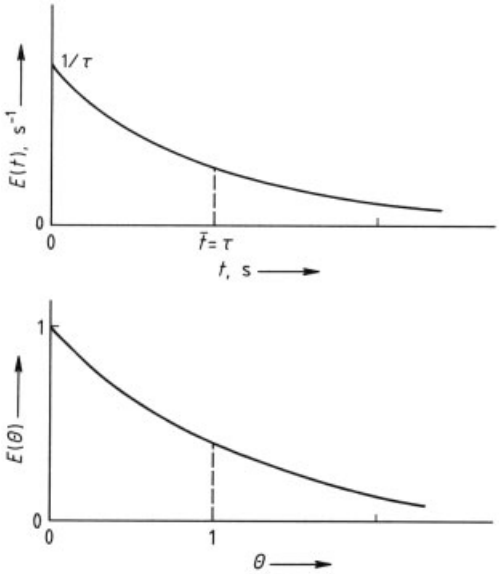

Figure 16. $E(t)$ and $E(\theta)$ diagrams for a continuously operated, ideally stirred tank reactor (CISTR)

4.3. Micromixing

When applying the above models to real reactor systems, the question remains whether these models reflect the real mixing behavior closely enough to predict reactor performance accurately. Especially in the case of non-first-order reactions or reactions with separate reactant inlets, the models should describe accurately enough not only the gross overall mixing behavior, but also the detailed intermixing of different fluid elements. To avoid problems related to micromixing, the assumption of an ideal microfluid has been made above. This is not necessarily true in practice.

For a continuously operated vessel, which shows the residence time of an ideal mixer, Equation (4.42) holds:

$$E(t) = \frac{e^{-t/\tau}}{\tau}$$

The different states of micromixing are illustrated in Figure 17. Three types can be distinguished:

1) The continuous, ideally stirred tank reactor (CISTR), a mixer that contains a microfluid
2) The continuous, segregated stirred tank reactor (CSSTR), a completely segregated mixer (here the fluid consists of lumps that travel through the reactor without exchanging matter with their neighbors)
3) The continuous, stirred-tank reactor (CSTR), which is a partially segregated mixer

Typical practical problems in which micromixing may have to be taken into account are

1) Reactors processing viscous fluids in which rapid reactions occur (e.g., polymerization reactors in which melt or solution polymerization is carried out)
2) Reactions in emulsions where the dispersed droplets have finite coalescence – redispersion rates
3) Reactors where rapid reactions occur between reactants that are fed through separate inlet points
4) Precipitation reactions
5) Highly nonisothermal reactors (e.g., flames)

For these systems an exact description is not possible and will probably not be practical for a long time to come. In many cases it is possible to calculate the extremes between which conversion and selectivity can be influenced by micromixing. In the following examples, only isothermal systems with constant density are considered.

In Figure 18 the reaction rate is depicted as a function of the concentration for an n-th-order reaction (i.e., $-\mathbb{R}_A- = k\,C_A^n$). Consider two elements 1 and 2 with equal volume, and assume that those elements leave the reactor at the same moment. The concentration C_A in the elements is not the same because the residence times of elements 1 and 2 are different. If the elements are not mixed (no micromixing), the average reaction rate for the two elements follows from

$$|\overline{\mathbb{R}_A}| = k\frac{C_{A,1}^n + C_{A,2}^n}{2} \tag{4.44}$$

If the two elements had in fact been mixed (complete micromixing), the reaction rate $-\mathbb{R}_A-_{\text{mixed}}$ would have been

$$|\mathbb{R}_A|_{\text{mixed}} = k\left(\frac{C_{A,1} + C_{A,2}}{2}\right)^n \tag{4.45}$$

This leads to the following conclusions:

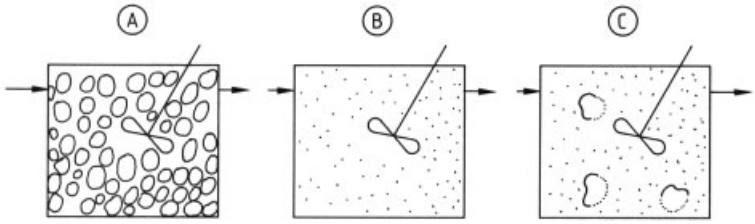

Figure 17. Possible states of micromixing in a stirred vessel
A) CSSTR, complete segregation; B) CISTR, ideal micromixing; C) CSTR, intermediate micromixing

1) $n < 1 \Rightarrow -\overline{\mathbb{R}_A} - < - \mathbb{R}_A -_{\text{mixed}}$ that is, for reaction orders below unity the absence of micromixing decreases the overall reaction rate
2) $n = 1 \Rightarrow -\overline{\mathbb{R}_A} - = - \mathbb{R}_A -_{\text{mixed}}$ that is, for a reaction order of unity, micromixing does not influence the rate
3) $n > 1 \Rightarrow -\overline{\mathbb{R}_A} - > - \mathbb{R}_A -_{\text{mixed}}$ that is, for reaction orders larger than unity the absence of micromixing increases the overall reaction rate

The intermixing of elements 1 and 2, which leave the reactor at the same moment, is not detectable by measuring the RTD function, so for non-first-order reactions the RTD does not provide sufficient data to calculate the conversion. For first-order reactions in which the degree of micromixing has no effect, the conversion can be calculated from the RTD.

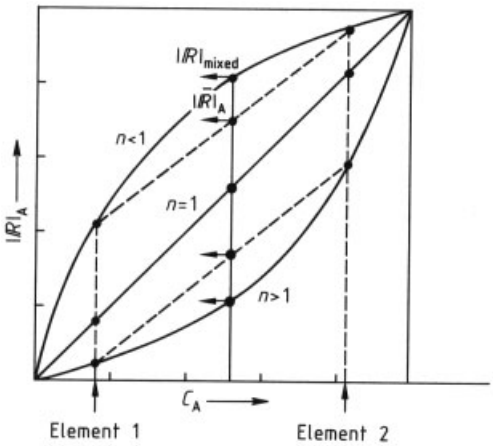

Figure 18. Influence of mixing on average reaction rate of two equal-volume fluid elements 1 and 2 that leave the reactor at the same time

For reactors with separate feed inlet points for different reactants, micromixing effects may be as important for plug-flow reactors as for CSTRs. In the extreme case of almost no micromixing, hardly any reaction occurs because the reactants remain in different aggregates and cannot come in contact with each other. These systems are discussed in [21].

Here the case is considered in which the reactants are premixed. For a plug-flow reactor with premixed reactants, the concept of micromixing is irrelevant because all elements that pass through the reactor can mix only with elements of exactly the same age. In a CSTR, two extreme states of micromixing exist: the microfluid (CISTR) and the completely segregated fluid (CSSTR). The first case has already been discussed; the latter is discussed here.

Completely Segregated Fluid. The fraction of aggregates having a residence time between t and $t + \partial t$ is $E(t)\,\partial t$, which follows directly from the definition of E given above. The average reactant concentration C_A in these aggregates equals the concentration $C_{A,\,BR}(t)$ obtained for a batch reactor in time t. This can be understood if one realizes that the aggregate has not mixed at all with other aggregates (complete segregation) and hence can be regarded a "small batch reactor" traveling through the CSSTR to the outlet. The average concentration at the CSSTR outlet is the sum of the products of concentration times fraction for all aggregates:

$$\overline{C_A} = \int_0^\infty C_{A,\,BR}(t)\,E(t)\,\partial t \tag{4.46}$$

In terms of the relative degree of conversion,

$$\zeta_{A,\,CSSTR} = \int_0^\infty \zeta_{A,\,BR}(t)\,E(t)\,\partial t \tag{4.47}$$

where $\zeta_{\mathrm{A,\,BR}}(t)$ is the conversion reached in a batch reactor after reaction time t. For a first-order reaction in a CSSTR, according to Equations (4.22), (4.42) and (4.47),

$$\zeta_{\mathrm{A,\,CSSTR}} = \int_0^\infty \left(1 - \mathrm{e}^{-kt}\right) \frac{\mathrm{e}^{-t/\tau}}{\tau}\, \partial t = \frac{k\tau}{1+k\tau} \tag{4.48}$$

If this is compared to Equation (4.36), for first-order reactions $\zeta_{\mathrm{A,\,CSSTR}} = \zeta_{\mathrm{A,\,CISTR}}$; this demonstrates again that for first-order reactions micromixing has no effect, as concluded from Figure 18.

For reaction orders not equal to unity, however, the situation is different. Below, the relative degree of conversion is calculated for a batch reactor, a plug-flow reactor (PFR), a CISTR, and a CSSTR for zero-order kinetics:

$$|\mathbb{R}_{\mathrm{A}}| = \begin{cases} k & C_{\mathrm{A}} \geq 0 \\ 0 & C_{\mathrm{A}} < 0 \end{cases} \tag{4.49}$$

At this point the Damköhler number *Da* should be defined:

$$Da = \begin{cases} \frac{k\,t}{C_{\mathrm{A,\,0}}} & \text{(BR)} \\ \frac{k\,\tau}{C_{\mathrm{A}}|_{z=0}} & \text{(PFR)} \\ \frac{k\,\tau}{C_{\mathrm{A,\,in}}} & \text{(CISTR/CSSTR)} \end{cases} \tag{4.50}$$

Then, for the different reactors, the following holds:

1) *Batch Reactor*. According to Equation (4.18),

$$\frac{\partial C_{\mathrm{A}}}{\partial t} = -\begin{cases} k & \text{for } C_{\mathrm{A}} \geq 0 \\ 0 & \text{for } C_{\mathrm{A}} < 0 \end{cases} \Rightarrow$$

$$\frac{\partial \zeta_{\mathrm{A}}}{\partial Da} = \begin{cases} 1 & \text{for } \zeta_{\mathrm{A}} \leq 1 \\ 0 & \text{for } \zeta_{\mathrm{A}} > 1 \end{cases} \tag{4.51}$$

This must be solved subject to

$$Da = 0 \Rightarrow \zeta_{\mathrm{A}} = 0 \tag{4.52}$$

which gives

$$\zeta_{\mathrm{A,\,BR}} = \begin{cases} Da & \text{for } Da \leq 1 \\ 1 & \text{for } Da \geq 1 \end{cases} \tag{4.53}$$

2) *Plug-flow reactor*. A micromaterial balance for a slice ∂z gives

$$u\frac{\partial C_{\mathrm{A}}}{\partial z} = -\begin{cases} k & \text{for } C_{\mathrm{A}} \geq 0 \\ 0 & \text{for } C_{\mathrm{A}} < 0 \end{cases} \Rightarrow$$

$$\frac{\partial \zeta_{\mathrm{A}}}{\partial\,(z/L)} = \begin{cases} Da & \text{for } \zeta_{\mathrm{A}} \leq 1 \\ 0 & \text{for } \zeta_{\mathrm{A}} > 1 \end{cases} \tag{4.54}$$

which is to be solved subject to

$$z = 0 \Rightarrow \zeta_{\mathrm{A}} = 0 \tag{4.55}$$

Integration to $z = L$ gives exactly the same result as found for the batch reactor (Eq. 4.53):

$$\zeta_{\mathrm{A,\,PFR}} = \begin{cases} Da & \text{for } Da \leq 1 \\ 1 & \text{for } Da > 1 \end{cases} \tag{4.56}$$

3) *CISTR*. The material balance (Eq. 4.32) now gives

$$\varphi_v\left(C_{\mathrm{A,\,in}} - C_{\mathrm{A}}\right) = \begin{cases} kV & \text{for } C_{\mathrm{A}} \geq 0 \\ 0 & \text{for } C_{\mathrm{A}} < 0 \end{cases} \tag{4.57}$$

This can be rewritten to again yield Equation (4.53):

$$\zeta_{\mathrm{A,\,CISTR}} = \begin{cases} Da & \text{for } Da \leq 1 \\ 1 & \text{for } Da > 1 \end{cases} \tag{4.58}$$

4) *CSSTR*. Now according to Equation (4.47),

$$\begin{aligned} \zeta_{\mathrm{A,\,CSSTR}} &= \int_0^\infty \zeta_{\mathrm{A,\,BR}}(t)\, E(t)\, \partial t \\ &= \int_0^{\frac{C_{\mathrm{A,in}}}{k}} \left(\frac{k\,t}{C_{\mathrm{A,in}}}\right)\left[\frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)\right] \partial t \\ &+ \int_{\frac{C_{\mathrm{A,in}}}{k}}^\infty \left[\frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right)\right] \partial t \\ &= \int_0^{\frac{1}{Da}} \theta \mathrm{e}^{-\theta}\, \partial\theta + \int_{\frac{1}{Da}}^\infty \mathrm{e}^{-\theta}\, \partial\theta \end{aligned} \tag{4.59}$$

which finally gives

$$\zeta_{\mathrm{A,\,CSSTR}} = Da\left[1 - \exp\left(-\frac{1}{Da}\right)\right] \tag{4.60}$$

Thus, for zeroth-order reactions $\zeta_{\mathrm{A,\,CSSTR}} \neq \zeta_{\mathrm{A,\,CISTR}}$ (see Fig. 19).

As expected from Figure 18, $\zeta_{A,\,CSSTR} < \zeta_{A,\,CISTR}$ because the reaction order is lower than unity. In a CISTR the conversion can reach 100 % for a finite residence time $\tau = C_{A,\,in}/k$, which follows from Equations (4.50) and (4.58). For a CSSTR, however, conversion in principle never reaches 100 %. Zeroth-order reactions are therefore very sensitive to micromixing at high conversions, so they are attractive as model reactions for assessing micromixing. Regretfully, many reactions that appear to be zeroth order at high reactant concentrations increase in apparent reaction order at higher conversions. This is the case, for example, if the reaction can be described by Langmuir or Michaelis – Menten kinetics as discussed in Sections 3.3.3 and 3.3.4.

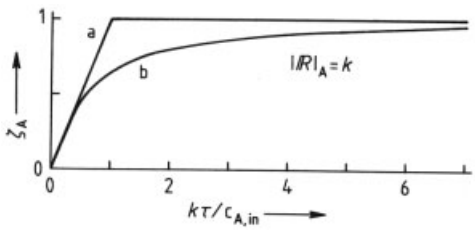

Figure 19. Influence of micromixing on the conversion of a zeroth-order chemical reaction in a stirred vessel (CSTR) a) Micromixed reactor; b) Completely segregated mixer

Intermediate Degree of Micromixing. The case of an intermediate degree of micromixing (e.g., between a microfluid and a completely segregated fluid) is very complex. Models describing intermediate micromixing usually do not give a clear physical picture of the actual phenomena but merely try to cover the span between the minimum and maximum degree of mixing by one or two adjustable parameters. These parameters must be determined experimentally (e.g., by measuring the conversion of a chemical reaction with non-first-order kinetics). A disadvantage is that the resulting micromixing parameter is merely a fitting parameter, which can have different values (e.g., for different conversions or reaction kinetics).

A more refined concept has been proposed by Weinstein and Adler [22] and improved by Villermaux and Zoulalian [23].

4.4. Sequence of Mixing

For the sake of completeness it should be mentioned here that macromixing and micromixing together are still not sufficient to give a unique prediction of the conversion. For non-first-order reactions the sequence of mixing is also relevant. Since, like macromixing, the sequence of mixing is not a macrokinetic, but a reactor-level phenomenon, the subject is touched on only briefly here by way of an example.

In Figure 20 the residence-time distribution of a plug-flow reactor and a CISTR in series is shown. Two cascades can be distinguished: the PFR before the CISTR or the PFR after the CISTR. The RTDs for both cascades are the same; hence, macromixing is the same for both cascades. Furthermore, a microfluid is assumed so the degree of micromixing is also the same. The only difference between the two cascades is the sequence of mixing.

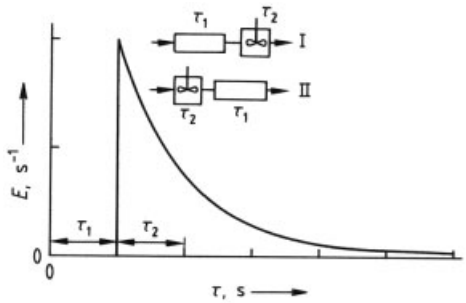

Figure 20. $E(t)$ diagram for cascades I and II
The order of the arrangement has no influence on the RTD.

For second-order reactions $-\mathbb{R}_A - = k\,C_A^2$ the conversion for both cascades can be calculated from

1) For cascade I:

$$\zeta_A = 1 + \frac{1}{2kC_{A,\,in}\tau_2}\left(1 - \sqrt{1 + 4\frac{kC_{A,\,in}\tau_2}{1 + kC_{A,\,in}\tau_1}}\right) \tag{4.61}$$

2) For cascade II:

$$\zeta_A = 1 + \frac{1 - \sqrt{1 + 4kC_{A,\,in}\tau_2}}{2kC_{A,\,in}\tau_2 + kC_{A,\,in}\tau_1\left(1 - \sqrt{1 + 4kC_{A,\,in}\tau_2}\right)} \tag{4.62}$$

This is illustrated in Figure 21. Obviously, the sequence of mixing can have a pronounced influence on the conversion of non-first-order reactions.

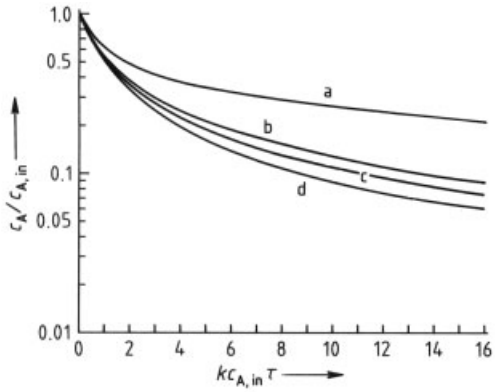

Figure 21. Influence of the sequence of mixing for a second-order reaction
a) CISTR; b) Cascade II (CISTR – plug-flow reactor); c) Cascade I (plug-flow reactor – CISTR); d) Plug-flow reactorCascades I and II are a PFR – CISTR and CISTR – PFR line up, respectively, with the CISTR and PFR having equal residence times $\tau/2$. In all cases a – d there are microfluids. Hence, only the sequence of mixing is changed.

5. Macrokinetics: Multiphase Systems – Part I: Mass Transfer without Reaction

5.1. Introduction

In this section, simple models describing the rate of mass transfer to or from an interface are summarized briefly. For a more detailed treatment, see → Transport Phenomena, Chap. 3.

If transport of component A is determined by *molecular diffusion* only (i.e., no flow of fluid), the net transport can be approximated by Fick's first law:

$$(J_A)_x = -\frac{C}{1-\kappa_A}\mathbb{D}_A\frac{\partial\kappa_A}{\partial x} \tag{5.1}$$

with $(J_A)_x$ the molar flux of A at location x (mol m^{-2} s^{-1}), C the total molar concentration (mol/m^3), κ_A the mole fraction of component A, $\mathbb{D}_A$ the coefficient of molecular diffusion of A (m^2/s), and x the coordinate perpendicular to the interface (m). In the majority of cases involving mass transfer with reaction, the total molar concentration does not change significantly near the interface. Furthermore, κ_A is often much less than 1. Then Fick's law reduces to

$$(J_A)_x = -\mathbb{D}_A\frac{\partial C_A}{\partial x} \tag{5.2}$$

In what follows, generally this simple expression (Eq. 5.2) is used. However, the validity of Equation (5.2) must always be checked. If large pressure, temperature, or mole fraction gradients occur in a system, a more rigorous approach must be chosen. For such systems, even the validity of Equation (5.1) is questionable, and use, for example, of the *dusty-gas model* is recommended (e.g., [24]).

If Equation (5.2) holds, a general material balance in Cartesian coordinates takes the following appearance:

$$\begin{aligned}
&\frac{\partial C_A}{\partial t} = && \text{Accumulation}\\
&-\frac{\partial(u_x C_A)}{\partial x} - \frac{\partial(u_y C_A)}{\partial y} - \frac{\partial(u_z C_A)}{\partial z} && \text{Convection}\\
&+\mathbb{D}_A\left(\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2}\right) && \text{Diffusion}\\
&-|\mathbb{R}_A| && \text{Reaction}
\end{aligned} \tag{5.3}$$

Note the minus sign in front of — $\mathbb{R}_A$ —: it is assumed that A is a reactant so that A is consumed.

For systems with constant density, Equation (5.3) can be simplified to

$$\frac{\partial C_A}{\partial t} = -u_x\frac{\partial C_A}{\partial x} - u_y\frac{\partial C_A}{\partial y} - u_z\frac{\partial C_A}{\partial z} + \mathbb{D}_A\left(\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2}\right) - |\mathbb{R}_A| \tag{5.4}$$

Often, further simplifications are possible. For example, for mass transfer in stagnant systems (no convection), $u_x = u_y = u_z = 0$; hence

$$\frac{\partial C_A}{\partial t} = \mathbb{D}_A\left(\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2}\right) - |\mathbb{R}_A| \tag{5.5}$$

which for steady-state conditions becomes

$$\mathbb{D}_A\left(\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2}\right) = |\mathbb{R}_A| \tag{5.6}$$

Alternatively, in the case of mass transfer without reaction (— $\mathbb{R}_A$ — = 0), Equation (5.5) yields

$$\frac{\partial \zeta C_A}{\partial t} = \mathbb{D}_A\left(\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2}\right) \tag{5.7}$$

which for steady-state conditions becomes

$$\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2} = 0 \tag{5.8}$$

Theoretically, the rate of mass transfer can be calculated from one of the Equations (5.3) to (5.8). The solution gives the concentration gradient at the interface, from which the molar flux of A through the interface can be calculated:

$$J_A = -\mathbb{D}_A \left(\frac{\partial C_A}{\partial x}\right)_{x=0} \tag{5.9}$$

In practice, however, this procedure is generally not applicable because of a lack of knowledge of the exact flow profile near the interface. Exceptions are mass transfer into or from quiescent media and in laminar flow systems. Therefore, more simple models are generally used, which lump the rate of mass transfer into one parameter. The three most widely used models are discussed below. Finally, the concept of the *mass-transfer coefficient* is introduced.

5.2. Penetration Models

The first penetration model originates from Higbie [25], where the mass-transfer process is described as follows: The interface is assumed to be covered with small stagnant fluid elements and mass transfer occurs into the elements. After a contact time τ with the interface, the elements return to the well-mixed bulk of the fluid from which they came, while new elements originating from the bulk take their place at the interface. Coming from the bulk, all stagnant elements have a uniform initial concentration $\overline{C_A}$. At the interface, mass transfer takes place by diffusion. This diffusion process is essentially transient. Figure 22 is a qualitative sketch of the concentration profiles versus the distance from the interface for several times. The depth of penetration of A into the element is assumed to be small with respect to the thickness of the element δ. Because this is a stagnant medium with no chemical reaction, Equation (5.7) applies. Since diffusion occurs perpendicular to the interface, only the x coordinate perpendicular to the interface is relevant and Equation (5.7) reduces to

$$\frac{\partial C_A}{\partial t} = \mathbb{D}_A \frac{\partial^2 C_A}{\partial x^2} \tag{5.10}$$

This partial differential equation must be solved subject to the initial condition

$$t = 0 \text{ and } x > 0 \Rightarrow C_A = \overline{C_A} \tag{5.11}$$

and the boundary conditions

$$t > 0 \text{ and } x = 0 \Rightarrow C_A = C_{A,i} \tag{5.12}$$

$$t > 0 \text{ and } x \to \infty \Rightarrow C_A \to \overline{C_A} \tag{5.13}$$

The last boundary condition can be used because the penetration depth is much smaller than the thickness of the element, so a semi-infinite medium can be assumed.

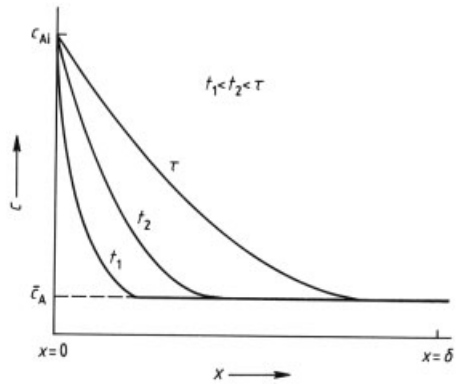

Figure 22. Concentration profiles in stagnant fluid element as a function of time (τ = maximum time: contact time in the Higbie penetration model)

The solution found for Equations (5.10)–(5.13) is

$$\frac{C_{A,i} - C_A}{C_{A,i} - \overline{C_A}} = \operatorname{erf}\left(\frac{x}{2\sqrt{\mathbb{D}_A t}}\right) \tag{5.14}$$

Differentiation of Equation (5.14) gives the molar flux of A through the interface:

$$J_A = -\mathbb{D}_A \left(\frac{\partial C_A}{\partial x}\right)_{x=0} = \sqrt{\frac{\mathbb{D}_A}{\pi t}}\left(C_{A,i} - \overline{C_A}\right) \tag{5.15}$$

This shows that the mass-transfer rate decreases with increasing time, which is caused by a decrease of $\partial C_A/\partial x$ at the interface as illustrated in Figure 22. The time-average molar flux of A for the entire contact time τ that the stagnant element remains at the interface equals

$$\overline{J_A} = \frac{1}{\tau}\int_0^\tau J_A\,\partial t = 2\sqrt{\frac{\mathbb{D}_A}{\pi\tau}}\left(C_{A,i} - \overline{C_A}\right) \tag{5.16}$$

The rate of mass transfer can be lumped in the parameter τ, which is called the *contact time* or sometimes the *penetration time*. Occasionally τ can be calculated from first principles (e.g., for model reactors such as the laminar jet or laminar film reactor [8]). For completely mobile bubbles it is suggested [25] that

$$\tau \approx \frac{d_b}{u_b} \tag{5.17}$$

with d_b the bubble diameter (m) and u_b the bubble rise velocity (m/s). However, for most applications, τ must be regarded as a fit parameter and must be obtained from experimental data.

A modification of the Higbie penetration model has been developed by DANCKWERTS [26]. Instead of a constant contact time τ, he assumes that the chance of a surface element being replaced within a given time is independent of its age. Thus, each element has an equal chance of being replaced during the next time unit. The fraction of elements that is replaced per unit time is s (s^{-1}). Thus, the following can be derived:

$$\overline{J_A} = \sqrt{\mathbb{D}_A s}\left(C_{A,i} - \overline{C_A}\right) \tag{5.18}$$

The penetration model of DANCKWERTS [26] is also known as the *surface renewal model*.

5.3. Stagnant Film Model

The stagnant film model is based on the following assumptions:

1) The fluid can be divided into two zones: a stagnant film near the interface with thickness δ and a well-mixed bulk behind it in which no concentration gradients occur
2) Mass transfer is a stationary process

Since a stagnant film, no reaction, and a steady state are assumed, Equation (5.8) applies in the film:

$$\frac{\partial^2 C_A}{\partial x^2} + \frac{\partial^2 C_A}{\partial y^2} + \frac{\partial^2 C_A}{\partial z^2} = 0 \quad \text{for} \quad 0 < x < \delta \tag{5.19}$$

If the radius of curvature of the interface is large compared to the film thickness and the x coordinate is perpendicular to the interface, Equation (5.19) reduces to

$$\frac{\partial^2 C_A}{\partial x^2} = 0 \quad \text{for} \quad 0 < x < \delta \tag{5.20}$$

This equation must be solved subject to the boundary conditions (see Fig. 23):

$$x = 0 \Rightarrow C_A = C_{A,i} \tag{5.21}$$

$$x = \delta \Rightarrow C_A = \overline{C_A} \tag{5.22}$$

Solving Equations (5.20) – (5.22) gives

$$\frac{C_{A,i} - C_A}{C_{A,i} - \overline{C_A}} = \frac{x}{\delta} \tag{5.23}$$

as illustrated in Figure 23. The molar flux of A then follows from

$$J_A = -\mathbb{D}_A \left(\frac{\partial C_A}{\partial x}\right)_{x=0} = \frac{\mathbb{D}_A}{\delta}\left(C_{A,i} - \overline{C_A}\right) \tag{5.24}$$

From a physical point of view, the film model is often not realistic because a stagnant film does not exist near an interface. The advantage of the model is its simplicity and mediative power.

All three models discussed above have in common the fact that the hydrodynamic conditions near the interface are lumped into one parameter: τ, s, or δ. For most applications this parameter must be regarded as a fit parameter to be obtained from experimental data. Because of the similarity of results, the application of more complicated (and possibly more realistic) mass-transfer models does not result in substantially better (i.e., more accurate) predictions (e.g., [27]). In particular, the knowledge of hydrodynamic, physical, and kinetic data is often not accurate enough. This explains why the simple film model is also applied in mass transfer to reaction problems in which the film theory is clearly an oversimplification with respect to the hydrodynamics of the system.

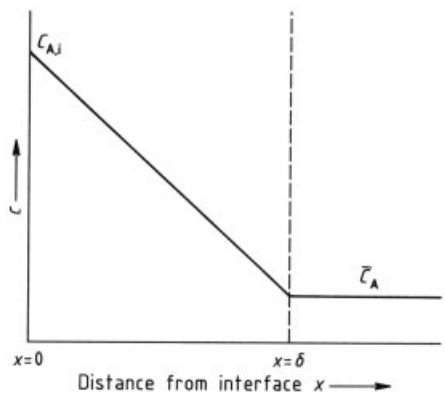

Figure 23. Concentration profile in the fluid film according to stagnant film theory

5.4. Mass-Transfer Coefficients

In practice the concept of the mass-transfer coefficient is often used. A mass-transfer coefficient is defined as

$$k_{\mathrm{l,A}} \quad \text{or} \quad k_{\mathrm{g,A}} \equiv \left| \frac{(J_{\mathrm{A}})_{\mathrm{phys}}}{C_{\mathrm{A,i}} - \overline{C}_{\mathrm{A}}} \right| \tag{5.25}$$

The subscript "phys" is added to J_A to stress the fact that the definition is restricted to the case in which A does not undergo a chemical reaction: only physical mass transfer of A is considered. The mass-transfer coefficients $k_{l,A}$ and $k_{g,A}$ have the dimension meters per second. If the molar fraction of A is low enough (< 0.1), the numerical values of $k_{l,A}$ and $k_{g,A}$ depend mainly on the geometry of the system, the physical constants (the viscosity, density, and diffusion coefficient), and the local hydrodynamics.

Substituting Equations (5.16), (5.18), and (5.24) in Equation (5.25) shows that each of the models discussed above gives its own relation for the mass-transfer coefficient:

$$k_{\mathrm{l,A}}/k_{\mathrm{g,A}} = \begin{cases} 2\sqrt{\frac{\mathbb{D}_{\mathrm{A}}}{\pi\,\tau}} & \textit{Higbie penetration model} \\ \sqrt{\mathbb{D}_{\mathrm{A}}\,s} & \textit{Surface renewal model} \\ \frac{\mathbb{D}_{\mathrm{A}}}{\delta} & \textit{Film theory} \end{cases} \tag{5.26}$$

The value of these relations is rather limited, because the parameters τ, s, and δ depend on the geometry of the system, the physical constants, and local hydrodynamics. In fact, τ, s, and δ are usually calculated from experimental mass-transfer coefficients according to the above equations. In calculating the rate of mass transfer of component A from one phase to another, the mass-transfer coefficients of both phases must be considered, as well as m, the distribution coefficient of A between the two phases. For example, for a gas liquid interface, a mass balance (see Fig. 24) yields

$$J_{\mathrm{A}} = k_{\mathrm{g,A}}\left(\overline{C_{\mathrm{A,g}}} - C_{\mathrm{A,i,g}}\right) = k_{\mathrm{l,A}}\left(\overline{C_{\mathrm{A,l}}} - C_{\mathrm{A,i,l}}\right) \tag{5.27}$$

For nearly all practical cases, the two phases at the interface can be assumed to be in equilibrium:

$$C_{\mathrm{A,i,l}} = m C_{\mathrm{A,i,g}} \tag{5.28}$$

The equilibrium constant or dimensionless distribution coefficient m generally depends on the physical properties of both phases (composition, pressure) and is usually strongly dependent on temperature. At a constant value of m, the proportionality given by Equation (5.28) is called Henry's law for gas – liquid equilibria and Nernst's law for liquid – liquid equilibria.

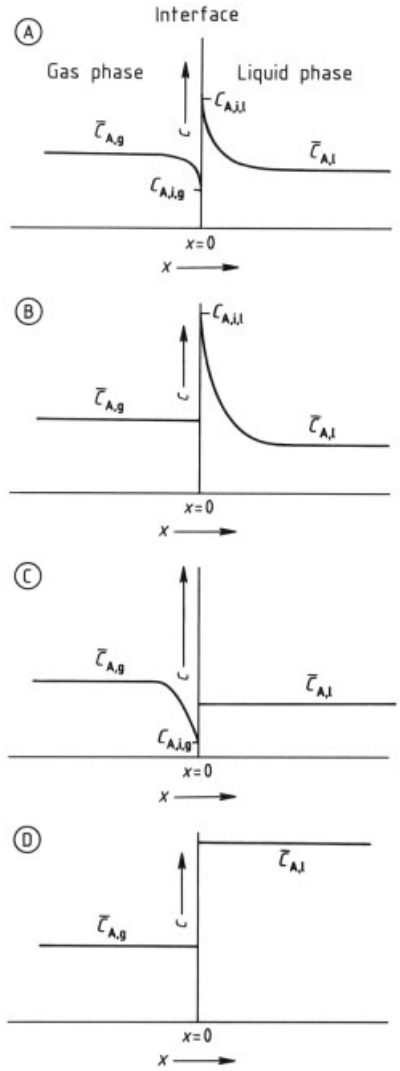

Figure 24. Concentration profiles near a gas – liquid interface in absorption
A) General case; B) Negligible gas-phase resistance; C) Negligible liquid-phase resistance; D) Equilibrium

Combining Equations (5.27) and (5.28) allows the interface concentrations to be eliminated and yields

$$J_A = \left(\frac{1}{mk_{l,A}} + \frac{1}{k_{g,A}}\right)^{-1} \left(\overline{C_{A,g}} - \frac{\overline{C_{A,l}}}{m}\right) \quad (5.29)$$

For $mk_{l,A}/k_{g,A} << 1$, the gas-phase resistance is negligible, and Equation (5.29) reduces to

$$J_A = mk_{l,A} \left(\overline{C_{A,g}} - \frac{\overline{C_{A,l}}}{m}\right) \quad (5.30)$$

For $mk_{l,A}/k_{g,A} >> 1$, the liquid-phase resistance is negligible, and Equation (5.29) reduces to

$$J_A = k_{g,A} \left(\overline{C_{A,g}} - \frac{\overline{C_{A,l}}}{m}\right) \quad (5.31)$$

Mass transfer stops for $\overline{C_{A,l}} = m\,\overline{C_{A,g}}$. Figure 24 qualitatively shows concentration profiles for the general case and for the asymptotic values.

6. Macrokinetics: Multi-phase Systems – Part II: Mass Transfer with Reaction

In the following sections, two cases of mass transfer with reaction are discussed: the case in which mass transfer and reaction occur in series, and the case of mass transfer with simultaneous reaction. The manner in which the two cases can be combined is also shown.

6.1. Mass Transfer with Reaction in Series

6.1.1. Introduction

The following steps occur in the general case of a reaction at the surface of a porous particle (e.g., catalyst):

1) Mass transfer of the reactants from the bulk of the fluid to the external surface of the particle
2) Transport of the reactants through the pores into the particle by diffusion
3) Adsorption of reactants on the solid surface
4) Chemical conversion in the adsorbed state
5) Desorption of products
6) Transport of products through the pores to the external surface of the particle
7) Transport of products from the external surface to the bulk

In the case of nonporous particles, steps 2 and 6 do not occur. In what follows, the assumption is made that the particles are not porous. Steps 2 and 6 come under simultaneous mass transfer and reaction, and are treated later. If steps 2 and 6 are absent, mass transfer and reaction occur in series. If only external mass transfer occurs, a rigorous analysis of this problem can be obtained by simply considering the values of the mass-transfer coefficients of each of the components. Whether the values of those coefficients are estimated with the Higbie penetration, surface renewal, or stagnant film model is not relevant: their value alone suffices. Consequently, for the analysis below, a distinction need not be made among the three models; therefore, they are not considered in this section.

6.1.2. Monomolecular First-Order Kinetics

Reactant A is assumed to be converted according to a first-order reaction. For convenience a chemical reaction rate — $\mathbb{R}''_A$ — is defined per unit of external surface:

$$|\mathbb{R}''_A| = \frac{V_p}{A_p} |\mathbb{R}_A|_p \quad (6.1)$$

where — $\mathbb{R}''_A$ — is the rate based on the external surface area of the pellet (moles of A per square meter of external pellet surface per second); V_p the pellet volume (m^3); A_p the external surface area of the pellet (m^2); and — $\mathbb{R}_A$ —$_p$ the reaction rate based on the pellet volume (moles of A per cubic meter of pellet volume per second). The latter parameter is discussed below in further detail. Corresponding to the reaction rate per unit surface is a surface kinetic constant k''_r for first-order kinetics:

$$|\mathbb{R}''_A| = k''_r C_{A,i} \quad (6.2)$$

where $C_{A,i}$ is the reactant concentration in the direct vicinity of the pellet surface (mol/m^3). Since first-order kinetics apply, the dimension of k''_r is meters per second.

If, for example, a gas – solid reaction is assumed, the stationary conversion rate can be calculated from Equation (6.2) together with

$$J_A = k_g\left(\overline{C_A} - C_{A,i}\right) \tag{6.3}$$

and under steady-state conditions

$$|\mathbb{R}''_A| = J_A \tag{6.4}$$

where k_g is the mass-transfer coefficient of A. Elimination of the unknown concentration $C_{A,i}$ gives the overall conversion rate

$$J_A = \left(\frac{1}{k''_r} + \frac{1}{k_g}\right)^{-1} \overline{C_A} \tag{6.5}$$

The two asymptotic solutions are (see Fig. 25)

1) The fast reaction regime for $k''_r/k_g >> 1$. The rate of mass transfer completely limits the conversion rate, which follows from

$$J_A = k_g \overline{C_A} \tag{6.6}$$

In this regime the conversion rate (hence, reactor capacity) can be enhanced by increasing the mass-transfer rate (e.g., by increasing the fluid velocity in the reactor or decreasing the average particle diameter).

2) The *slow reaction regime* for $k''_r/k_g << 1$. The chemical reaction rate completely limits the conversion rate, which now follows from

$$J_A = k''_r \overline{C_A} \tag{6.7}$$

In this regime, the conversion rate can be enhanced by increasing the chemical reaction rate, for example, by increasing the temperature of the reactor.

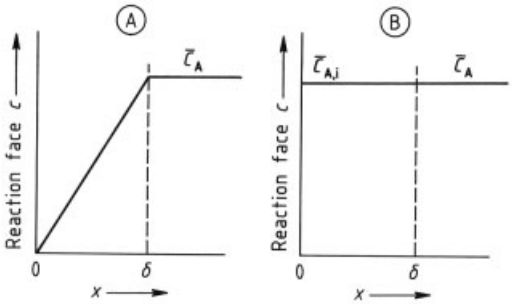

Figure 25. Two asymptotic concentration profiles for mass transfer and reaction in series (film theory)
A) Fast reaction regime; B) Slow reaction regime

6.1.3. Arbitrary Monomolecular Kinetics

The system analyzed above is easily extended to arbitrary kinetics. Langmuir kinetics are discussed here as an example. Then,

$$|\mathbb{R}''_A| = \frac{k''_r C_{A,i}}{1 + K C_{A,i}} \tag{6.8}$$

Combination with Equations (6.3) and (6.4) and elimination of $C_{A,i}$ give:

$$J_A = k_g \overline{C_A} \frac{2\beta}{1 + \alpha + \beta + \sqrt{(1 - \alpha + \beta)^2 + 4\alpha}}$$

$$\text{with} \quad \alpha = K\overline{C_A} \quad \text{and } \beta = \frac{k''_r}{k_g} \tag{6.9}$$

In the *slow reaction regime*, $k''_r/k_g << 1$ and thus $\beta \to 0$, Equation (6.9) becomes

$$J_A = k_g \overline{C_A} \frac{\beta}{1 + \alpha} = \frac{k''_r \overline{C_A}}{1 + K\overline{C_A}} \tag{6.10}$$

and mass-transfer limitation has no effect on the overall conversion rate.

In the *fast reaction regime* where $k''_r/k_g >> 1$ and $\beta \to \infty$, Equation (6.9) yields

$$J_A = k_g \overline{C_A} \tag{6.11}$$

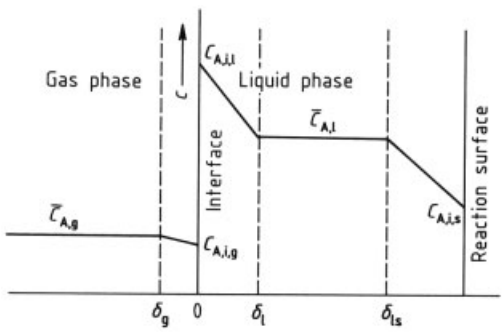

Figure 26. Concentration profiles for mass transfer and reaction in series in a gas – liquid – solid system (film theory)

Hence, due to mass-transfer limitation, an apparent first-order kinetics is observed in the fast reaction regime, irrespective of the numerical value of the term $K\overline{C_A}$. The physical explanation is that although the value of the term $K\overline{C_A}$

may be large, in this regime $C_{A,i}$ is much smaller than $\overline{C_A}$, so that $KC_{A,i}$ is much smaller than unity and the adsorption term in Langmuir kinetics becomes negligible. From this, the conclusion can be drawn that mass-transfer limitation may cause an apparent change in the reaction order. Below, more examples illustrating this statement are presented. For other forms of arbitrary kinetics, expressions for the conversion rate analogous to Equation (6.9) are obtained in a similar way.

6.1.4. Multiple Phases

The system discussed in Sections 6.1.2 and 6.1.3 is also easily extended to a three-phase reactor with mass transfer and reaction in series: for example, a gas is absorbed in a liquid in which nonporous particles are suspended. Reaction occurs at the surface of the particles. Examples are the hydrogenation of organic liquids with a solid catalyst and the alkylation of a liquid reactant with gaseous ethylene, for example, and solid $AlCl_3$ as catalyst.

The concentration profile could be that shown in Figure 26. At the gas – liquid interface the assumed equilibrium is given by

$$C_{A,i,l} = m\,C_{A,i,g} \tag{6.12}$$

where m is the distribution coefficient. The parameter a_{gl} is defined as the specific gas – liquid surface area (m^{-1}) and a_{ls} as the specific catalyst particle area (m^{-1}), both per unit volume of dispersion. k_g is the gas-phase mass-transfer coefficient for the transfer from gas to liquid, and k_l and k_{ls} are the liquid-phase mass-transfer coefficients for the transfers from gas to liquid and liquid to solid, respectively (all with the dimension m/s). If first-order kinetics are assumed, then the *overall conversion rate* — R_A — is given by

$$\begin{aligned} |R_A| &= a_{ls} k_r'' C_{A,i,s} \\ &= a_{gl} k_g \left(\overline{C_{A,g}} - C_{A,i,g}\right) \\ &= a_{gl} k_l \left(C_{A,i,l} - \overline{C_{A,l}}\right) \\ &= a_{ls} k_{ls} \left(\overline{C_{A,l}} - C_{A,i,s}\right) \end{aligned} \tag{6.13}$$

Combining Equations (5.27) and (6.13) and eliminating the unknown concentrations $C_{A,i,g}$, $C_{A,i,l}$, $C_{A,i,s}$, and $\overline{C_{A,l}}$ give

$$|R_A| = \overline{C_{A,g}} \cdot \left(\frac{1}{a_{gl} k_g} + \frac{1}{m a_{gl} k_l} + \frac{1}{m a_{ls} k_{ls}} + \frac{1}{m a_{ls} k_r''}\right)^{-1} \tag{6.14}$$

Reactor capacity per unit volume appears to depend on four resistances in series: the gas-phase transfer resistance, two liquid-phase transfer resistances, and the kinetic resistance. The highest resistance limits the capacity of the reactor.

Implicit in the derivation of Equation (6.14) is the assumption that all solid particles are in the bulk of the liquid phase. However, for

$$\frac{1}{a_{ls} k_{ls}} + \frac{1}{a_{ls} k_r''} << \frac{1}{a_{gl} k_l} \tag{6.15}$$

the concentration of A in the bulk of the liquid approaches zero. Then, only the particles in the liquid film at the gas – liquid interface may be active; hence mass transfer and reaction are parallel steps instead of being in series. This problem has been studied by SADA et al. [28], [29].

6.1.5. Bimolecular Reactions

Here again a two-phase situation is assumed – a gas – solid reaction on an external pellet surface.

If two components take part in a surface reaction, an extra balance for the second component must be considered as well. For

$$A + \nu, B \rightarrow \text{Products}$$

obtained

$$J_A = k_{g,A} \left(\overline{C_A} - C_{A,i}\right) \tag{6.16}$$

$$J_B = k_{g,B} \left(\overline{C_B} - C_{B,i}\right) \tag{6.17}$$

If, for example, the above reaction occurs according to (1,1)-kinetics, furthermore,

$$|\mathbb{R}_A''| = k_r'' C_{A,i} C_{B,i} \tag{6.18}$$

A material balance gives

$$J_A = |\mathbb{R}_A''| \tag{6.19}$$

$$J_B = \nu J_A \tag{6.20}$$

Combining Equations (6.16) – (6.20) and eliminating $C_{A,i}$, $C_{B,i}$, and J_B give

$$J_A = k_{g,A}\overline{C_A} \cdot \frac{2\beta_B}{1+\beta_A+\beta_B+\sqrt{(1-\beta_A+\beta_B)^2+4\beta_A}}$$

$$\text{with } \beta_A = \nu\frac{k_r''\overline{C_A}}{k_{g,B}} \text{ and } \beta_B = \frac{k_r''\overline{C_B}}{k_{g,A}} \tag{6.21}$$

For the *slow reaction regime* (i.e., $\beta_A \to 0$ and $\beta_B \to 0$), Equation (6.21) simplifies to

$$J_A = k_r''\overline{C_A}\,\overline{C_B} \tag{6.22}$$

thus, the mass-transfer resistance is negligible.

For the *fast reaction regime* (i.e., $\beta_A \to \infty$ and $\beta_B \to \infty$), Equation (6.21) simplifies to one of the following equations:

1) For $\nu\, k_{g,\,A}\,\overline{C_A} < k_{g,\,B}\,\overline{C_B}$,

$$J_A = k_{g,A}\overline{C_A} \tag{6.23}$$

Hence, the overall conversion rate is limited by mass transfer of the reactant with the lowest mass-transfer rate.

2) For $\nu\, k_{g,\,A}\,\overline{C_A} > k_{g,\,B}\,\overline{C_B}$,

$$J_B = \frac{1}{\nu}k_{g,B}\overline{C_B} \tag{6.24}$$

Now B limits the overall conversion rate because it has the lowest mass-transfer rate.

Thus, corrected for stoichiometry, the key component with the lowest rate of mass transfer determines the rate of conversion, and the component in excess at the catalyst interface has no influence at all on the rate in this case.

As illustrated for monomolecular reactions, the system analyzed here is easily extended to bimolecular reactions with arbitrary kinetics by modifying the rate equation of — $\mathbb{R}_A''$ — as given in Equation (6.18), albeit the algebra involved can become quite tedious.

6.1.6. Case Study: Shrinking Core Model

As an example, a solid metal oxide is reduced with a gaseous component (e.g., hydrogen, carbon monoxide) according to the equilibrium reaction

$$A\,(g) + \nu_{MO}\,MO_x\,(s) \rightleftharpoons \nu\,P\,(g) + \nu_{MO}\,M\,(s)$$

with

$$|\mathbb{R}_A''| = k_r''\left(C_A - \frac{C_P}{K}\right) \tag{6.25}$$

where A may be H_2 or CO and P may be H_2O or CO_2, respectively; MO_x is the metal oxide (ore) and M the pure metal.

The reaction is assumed to take place at a sharp boundary between the porous metal product layer and the shrinking unreacted nonporous metal oxide core. Then the following steps are involved:

1) Diffusion of reducing gas A through the external film surrounding the pellet
2) Diffusion of A through the porous product layer of pure metal
3) Chemical reaction at the sharp interface between A and MO_x
4) Diffusion of product gas P through the porous product layer
5) Diffusion of P via the external film into the bulk of the gas

An expression can be derived for the relative degree of conversion of the metal oxide ζ_{MO} as a function of time.

The mole fluxes of components A and P at the external surface $J_{A,\,p}$ and $J_{P,\,p}$ (see Fig. 27) are

$$J_{A,\,p} = k_{g,\,A}\left(\overline{C_A} - C_{A,\,i}\right) \tag{6.26}$$

$$J_{P,\,p} = k_{g,\,P}\left(C_{P,\,i} - \overline{C_P}\right) \tag{6.27}$$

The reaction rate at the core surface — $\mathbb{R}_{A,\,c}''$ — is given by

$$|\mathbb{R}_{A,\,c}''| = k_r''\left(C_{A,\,c} - \frac{C_{P,\,c}}{K}\right) \tag{6.28}$$

From the material balances,

$$4\pi r_p^2 J_{A,\,p} = \frac{1}{\nu}4\pi r_p^2 J_{P,p} = 4\pi r_c^2\,|\mathbb{R}_{A,\,c}''| \tag{6.29}$$

From the above set of equations the quantities $J_{P,\,p}$, — $\mathbb{R}_{A,\,c}''$ —, $C_{A,\,i}$, $C_{P,\,i}$, $C_{A,\,c}$, and $C_{P,\,c}$ must be eliminated. To achieve this, $C_{A,\,c}$ must be expressed in terms of $C_{A,\,i}$, and $C_{P,\,c}$ must be expressed in terms of $C_{P,\,i}$. These expressions are obtained from material balances for A and P across an infinitesimally thin spherical slice

within the porous product layer. This is illustrated for reactant A.

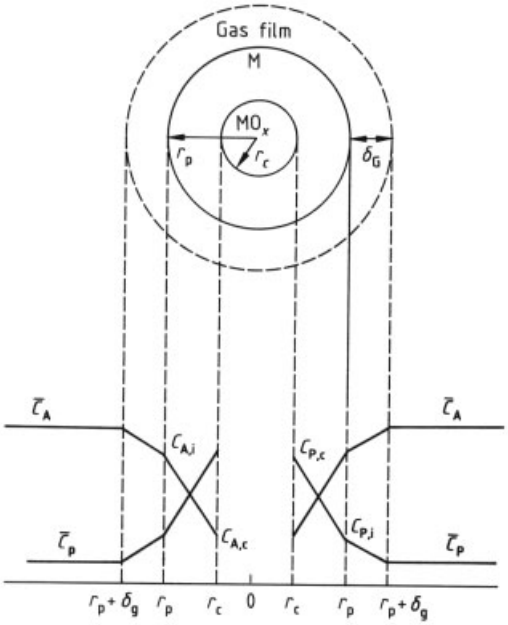

Figure 27. Concentration profiles for unreacted shrinking core model, for example, with the equilibrium reaction: $MO_x + A \rightleftharpoons M + AO_x$, where MO_x = metal oxide; M = metal; and A = reducing agent (for explanation of other symbols, see text)

Since the mole flow of A through each of the infinitesimally thin slices must be the same,

$$4\pi r^2 J_A(r) = 4\pi r_p^2 J_{A,p} \neq f(r) \tag{6.30}$$

If the diffusion coefficient of A in the porous metal layer is denoted by $\mathbb{D}_{i,A}$,

$$4\pi r^2 \mathbb{D}_{i,A} \frac{\partial C_A(r)}{\partial r} = 4\pi r_p^2 J_{A,p} \tag{6.31}$$

$\mathbb{D}_{i,A}$ is discussed in more detail in the next section. The differential equation (Eq. 6.31) must be solved subject to the boundary condition:

$$r = r_p \Rightarrow C_A(r) = C_{A,i} \tag{6.32}$$

The solution is

$$C_A(r) = C_{A,i} - \frac{J_{A,p}}{\mathbb{D}_{i,A}} \frac{r_p}{r}(r_p - r) \tag{6.33}$$

By substituting $r = r_c$, $C_{A,c}$ is obtained as a function of $C_{A,i}$:

$$C_{A,c} = C_{A,i} - \frac{J_{A,p}}{\mathbb{D}_{i,A}} \frac{r_p}{r_c}(r_p - r_c) \tag{6.34}$$

Similarly, for component P,

$$C_{P,c} = C_{P,i} + \frac{J_{P,p}}{\mathbb{D}_{i,P}} \frac{r_p}{r_c}(r_p - r_c) \tag{6.35}$$

Combining Equations (6.26)–(6.29), (6.34), and (6.35) gives the mole flux of A through the external surface as a function of core and particle diameter, concentrations, and physical constants:

$$J_{A,p} = \frac{1}{\xi(\varrho)} k_{g,A} \left(\overline{C_A} - \frac{\overline{C_P}}{K} \right) \tag{6.36}$$

where $\xi(\varrho)$ can be calculated from

$$\xi(\varrho) = 2 + Bi_A \frac{1-\varrho}{\varrho} + \frac{k_{g,A}}{k_r''} \varrho^2 + \frac{\nu}{K} \frac{k_{g,A}}{k_{g,P}} Bi_P \cdot \frac{1-\varrho}{\varrho} + 2\frac{\nu}{K} \frac{k_{g,A}}{k_{g,P}} \tag{6.37}$$

Here Bi_A and Bi_P are *Biot numbers*, which are defined as follows:

$$Bi_A = \frac{k_{g,A} r_p}{\mathbb{D}_{i,A}} \tag{6.38}$$

$$Bi_P = \frac{k_{g,P} r_p}{\mathbb{D}_{i,P}} \tag{6.39}$$

ϱ is a dimensionless core radius, which can vary from 1 to 0 during the reduction process:

$$\varrho = \frac{r_c}{r_p} \tag{6.40}$$

For $\varrho = 1$, no metal oxide has been reduced and an outer porous layer is still absent. For $\varrho = 0$, all metal oxide has been reduced and a porous metal particle without an oxide core remains.

If C_{MO} is the concentration of metal oxide in the virgin particle and core region, the rate of change of core diameter $-\frac{\partial r_c(t)}{\partial t}$ is obtained from

$$4\pi r_p^2 J_{A,p}(t) = -4\pi [r_c(t)]^2 \frac{\partial r_c(t)}{\partial t} \frac{C_{MO}}{\nu_{MO}} \tag{6.41}$$

where ν_{MO} is the stoichiometric coefficient of the metal oxide as given by the chemical equation. After substitution of Equation (6.36) in Equation (6.41),

$$\frac{\partial \varrho}{\partial t} = -\frac{\nu_{MO}}{C_{MO}} \left(\overline{C_A} - \frac{\overline{C_P}}{K} \right) \frac{k_{g,A}}{r_p} \frac{1}{\varrho^2 \xi(\varrho)} \tag{6.42}$$

For constant values of $\overline{C}_A$ and $\overline{C}_P$, Equation (6.42) can be integrated for the entire process subject to the initial condition $\varrho = 1$ for $t = 0$. With the relative degree of conversion of the metal oxide following from

$$\zeta_{MO} = 1 - \varrho^3 \tag{6.43}$$

this integration yields

$$\frac{2}{3}\left(1+\frac{\nu}{K}\frac{k_{g,A}}{k_{g,P}}\right)\zeta_{MO} + \frac{1}{6}\left(Bi_A + \frac{\nu}{K}\frac{k_{g,A}}{k_{g,P}}Bi_P\right) \cdot \left[1 - 3\,(1-\zeta_{MO})^{2/3} + 2\,(1-\zeta_{MO})\right] + \frac{k_{g,A}}{k_r''}\left[1-(1-\zeta_{MO})^{1/3}\right] = \frac{\nu_{MO}}{C_{MO}}\left(\overline{C_A} - \frac{C_P}{K}\right)\frac{k_{g,A}}{r_P}t \tag{6.44}$$

Equation (6.44) is very useful in conversion experiments with a thermal balance, where $\overline{C}_A$ and $\overline{C}_P$ are kept constant. In counter- or cocurrent-flow blast furnaces or in moving-bed gasifiers, $\overline{C}_A$ and $\overline{C}_P$ vary with ζ_{MO}, so Equation (6.44) must be applied anew for each step when integrating to the reactor level in stepwise calculations.

Depending on the controlling resistance, three asymptotic cases of Equation (6.44) can be distinguished:

1) *Slow External Mass Transfer* (i.e., $\mathbb{D}_{i,A}$, $\mathbb{D}_{i,P}$, $k_r'' \to \infty$). In this regime,

$$\zeta_{MO} = \frac{t}{\tau_M} \tag{6.45}$$

with

$$\tau_M = \frac{2}{3}\left(1+\frac{\nu}{K}\right)\frac{C_{MO}}{\nu_{MO}}\left(\overline{C_A} - \frac{\overline{C_P}}{K}\right)^{-1}\frac{r_P}{k_{g,A}} \tag{6.46}$$

2) *Slow Diffusion Through Converted Shell* (i.e., $k_{g,A}$, $k_{g,P}$, $k_r'' \to \infty$). In this regime,

$$1 - 3\,(1-\zeta_{MO})^{2/3} + 2\,(1-\zeta_{MO}) = \frac{t}{\tau_D} \tag{6.47}$$

with

$$\tau_D = \frac{1}{6}\left(1+\frac{\nu}{K}\frac{\mathbb{D}_{i,A}}{\mathbb{D}_{i,P}}\right)\frac{C_{MO}}{\nu_{MO}}\left(\overline{C_A} - \frac{\overline{C_P}}{K}\right)^{-1}\frac{r_P^2}{\mathbb{D}_{i,P}} \tag{6.48}$$

3) *Slow Chemical Reaction Kinetics* (i.e., $k_{g,A}$, $k_{g,P}$, $\mathbb{D}_{i,A}$, $\mathbb{D}_{i,P} \to \infty$). In this regime,

$$1 - (1-\zeta_{MO})^{1/3} = \frac{t}{\tau_R} \tag{6.49}$$

with

$$\tau_R = \frac{C_{MO}}{\nu_{MO}}\left(\overline{C_A} - \frac{\overline{C_P}}{K}\right)^{-1}\frac{r_P}{k_r''} \tag{6.50}$$

Figure 28 shows $\frac{\partial \zeta_{MO}}{\partial (t/\tau)}$ (with $\tau = \tau_M, \tau_D, \tau_R$) as a function of ζ_{MO} for the various asymptotic solutions. Notice that in practice the rate-controlling step may depend on the value of ζ_{MO}. For low conversions, diffusion through the porous metal layer is not necessary and external mass transfer or reaction will control the overall conversion rate. For large conversions and not too small particles, diffusion through the porous layer often controls the rate of conversion.

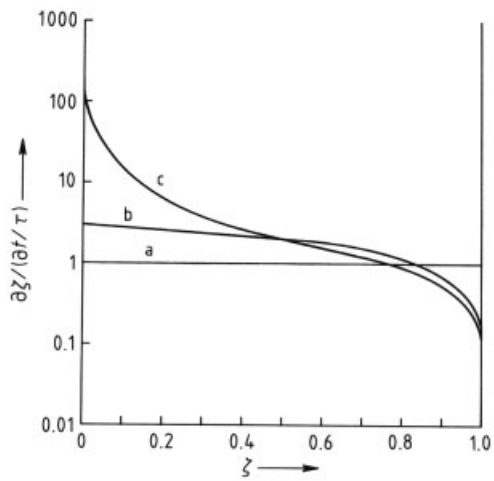

Figure 28. Conversion rate as a function of conversion according to the shrinking core model
a) External mass transfer in gas controls ($\tau = \tau_M$); b) Diffusion in converted shell controls ($\tau = \tau_D$); c) Chemical reaction rate controls ($\tau = \tau_R$)

6.1.7. Selectivities of Multiple Reactions

6.1.7.1. Introduction

Up to now, mass transfer and reaction in series have been described with one reaction only. More often than not, a reactant is involved in at least two reactions. Generally, only one of the

products is commercially desirable and the reactions must be carried out so that both an acceptable conversion and acceptable selectivity are obtained. Therefore, this section deals with the problem of selectivity for multiphase reactions and with mass transfer and reaction in series.

A distinction is made between *parallel* and *consecutive reactions* [30]:

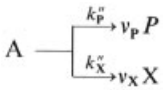

respectively,

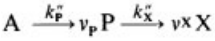

Combinations of both are also possible; in oxidation reactions the following scheme is often encountered:

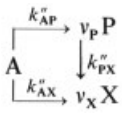

where A is the reactant (e.g., ethylene), P a desired product obtained after partial oxidation (ethylene oxide), and X the byproducts obtained after complete combustion (carbon dioxide, water).

6.1.7.2. Parallel Reactions

For the following reaction scheme

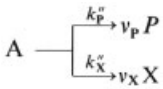

the desired product P is assumed to be formed with a reaction rate $-\mathbb{R}''_A(C_A)-_P$, whereas for the undesired product X the rate equals $-\mathbb{R}''_A(C_A)-_X$. The selectivity toward the desired product P then follows from (see Eq. 2.8)

$$\sigma_P = \frac{|\mathbb{R}''_A(C_{A,i})|_P}{|\mathbb{R}''_A(C_{A,i})|_P + |\mathbb{R}''_A(C_{A,i})|_X} \tag{6.51}$$

where σ_P can vary between 0 and 1.

For n-th and m-th-order kinetics

$$|\mathbb{R}''_A(C_A)|_P = k''_P C_A^n \tag{6.52}$$

$$|\mathbb{R}''_A(C_A)|_X = k''_X C_A^m \tag{6.53}$$

Equation (6.51) yields

$$\sigma_P = \left(1 + \frac{k''_X}{k''_P} C_{A,i}^{m-n}\right)^{-1} \tag{6.54}$$

If the desired reaction is of higher order ($n > m$), a high $C_{A,i}$ (thus high $\overline{C}_A$ and high k_g) will favor the selectivity. The high k_g requires small particles and a relatively high fluid velocity. For the special case in which the parallel reactions are of the same order (i.e., $n = m$), the selectivity is not influenced by mass-transfer limitation and

$$\sigma_P = \frac{k''_P}{k''_P + k''_X} \tag{6.55}$$

However, for this situation the relative degree of conversion ζ_A may still depend on the rate of mass transfer.

In general, selectivity can be influenced by mass-transfer limitation, as is the case for $n \neq m$. For such cases the concentration $C_{A,i}$ is needed to calculate the selectivity (e.g., see Eq. 6.54). The parameter $C_{A,i}$ follows from a mass balance for A:

$$J_A = k_g(\overline{C_A} - C_{A,i}) = |\mathbb{R}''_A(C_{A,i})|_P + |\mathbb{R}''_A(C_{A,i}|_X) \tag{6.56}$$

Solving the above equation, expressing $C_{A,i}$ in $\overline{C}_A$, and substituting this relation in Equation (6.51), gives a straightforward expression for σ_P in $\overline{C}_A$. Thus, for $n = 2$ and $m = 1$, Equation (6.56) yields

$$C_{A,i} = \overline{C_A} \frac{2}{1 + \beta_X + \sqrt{(1+\beta_X)^2 + 4\beta_P}}$$

$$\beta_P = \frac{k''_P}{k_g} \quad \text{and} \quad \beta_X = \frac{k''_X}{k_g} \tag{6.57}$$

which, upon substitution in Equation (6.54), gives

$$\sigma_P = \left\{1 + \frac{\beta_X}{2\beta_P}\left[1 + \beta_X + \sqrt{(1+\beta_X)^2 + 4\beta_P}\right]\right\}^{-1} \tag{6.58}$$

with β_P and β_X as in Equation (6.57). Since the desired reaction is of higher order ($n > m$), a high k_g yields the highest selectivity. This selectivity is found by substituting β_P and $\beta_X \to 0$ in Equation (6.58), which yields

$$\sigma_{\mathrm{P,max}} = \frac{k''_{\mathrm{P}}\,\overline{C_{\mathrm{A}}}}{k''_{\mathrm{P}}\,\overline{C_{\mathrm{A}}} + k''_{\mathrm{X}}} \tag{6.59}$$

This is the selectivity for slow reaction or no mass-transfer limitation at all.

For very low k_{g} values (i.e., in the fast reaction regime), the lowest selectivity is obtained. An expression is obtained by substituting β_{P} and $\beta_{\mathrm{X}} \to \infty$ in Equation (6.58), which yields

$$\sigma_{\mathrm{P,min}} = \frac{k_{\mathrm{g}} k''_{\mathrm{P}}\,\overline{C_{\mathrm{A}}}}{\left(k''_{\mathrm{X}}\right)^2} \tag{6.60}$$

Hence, in the fast reaction regime, selectivity is proportional to the mass-transfer coefficient k_{g}, which illustrates the importance of achieving as high a mass-transfer coefficient as possible. Expressions for σ_{P} for other forms of kinetics are obtained in a similar way as illustrated above.

6.1.7.3. Consecutive Reactions

The following reaction scheme is considered:

$$\mathrm{A} \xrightarrow{k''_{\mathrm{P}}} \nu_{\mathrm{P}}\mathrm{P} \xrightarrow{k''_{\mathrm{X}}} \nu_{\mathrm{X}}\mathrm{X}$$

The reaction of A to form the desired product P is assumed to have a rate — $\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A}}\right)$ — in moles of A per square meter per second; the rate of the reaction of P forming the undesired product X equals — $\mathbb{R}''_{\mathrm{P}}\left(C_{\mathrm{P}}\right)$ — in moles of P per square meter per second. The selectivity toward the desired product P then follows from (see Eq. 2.8)

$$\sigma_{\mathrm{P}} = \frac{\nu_{\mathrm{P}}\left|\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)\right| - \left|\mathbb{R}''_{\mathrm{P}}\left(C_{\mathrm{P,i}}\right)\right|}{\nu_{\mathrm{P}}\left|\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)\right|} \tag{6.61}$$

σ_{P} is always less than unity. It need not, however, be larger than 0: for — $\mathbb{R}''_{\mathrm{P}}\left(C_{\mathrm{P,i}}\right)$ — $> \nu_{\mathrm{P}}$ — $\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)$ —, more P is converted to X than formed from A. Hence, a net overall consumption of P occurs, which results in a negative value for σ_{P}.

The interface concentrations $C_{\mathrm{A,i}}$ and $C_{\mathrm{P,i}}$ are again obtained from mass balances:

$$J_{\mathrm{A}} = k_{\mathrm{g,A}}\left(\overline{C_{\mathrm{A}}} - C_{\mathrm{A,i}}\right) = \left|\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)\right| \tag{6.62}$$

$$\begin{aligned} J_{\mathrm{P}} &= k_{\mathrm{g,P}}\left(C_{\mathrm{P,i}} - \overline{C_{\mathrm{P}}}\right) \\ &= \nu_{\mathrm{P}}\left|\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)\right| - \left|\mathbb{R}''_{\mathrm{P}}\left(C_{\mathrm{P,i}}\right)\right| \end{aligned} \tag{6.63}$$

With the above set of equations, $C_{\mathrm{A,i}}$ and $C_{\mathrm{P,i}}$ can be expressed in the bulk concentrations $\overline{C_{\mathrm{A}}}$ and $\overline{C_{\mathrm{P}}}$, which upon substitution in Equation (6.61) give σ_{P} as a function of bulk concentration.

For cases in which both reactions are *first order*,

$$\left|\mathbb{R}''_{\mathrm{A}}\left(C_{\mathrm{A}}\right)\right| = k''_{\mathrm{P}} C_{\mathrm{A}} \tag{6.64}$$

$$\left|\mathbb{R}''_{P}\left(C_{\mathrm{P}}\right)\right| = k''_{\mathrm{X}} C_{\mathrm{P}} \tag{6.65}$$

Combining Equations (6.61) – (6.65) gives

$$\sigma_{\mathrm{P}} = \frac{1 - \Sigma\left(1 + \beta_{\mathrm{A}}\right)}{1 + \beta_{\mathrm{P}}}$$

$$\Sigma = \frac{1}{\nu_{\mathrm{P}}}\frac{k''_{\mathrm{X}}}{k''_{\mathrm{P}}}\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}}, \quad \beta_{\mathrm{A}} = \frac{k''_{\mathrm{P}}}{k_{\mathrm{g,A}}}, \quad \beta_{\mathrm{P}} = \frac{k''_{\mathrm{X}}}{k_{\mathrm{g,P}}} \tag{6.66}$$

Two asymptotic solutions can be distinguished:

1) *Slow Reaction Regime*. Here β_{A} and $\beta_{\mathrm{P}} \to 0$; hence,

$$\sigma_{\mathrm{P}} = 1 - \Sigma = 1 - \frac{1}{\nu_{\mathrm{P}}}\frac{k''_{\mathrm{X}}}{k''_{\mathrm{P}}}\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}} \tag{6.67}$$

which is the solution for no mass-transfer limitation at all.

2) *Fast Reaction Regime*. Here β_{A} and $\beta_{\mathrm{P}} \to \infty$, hence,

$$\sigma_{\mathrm{P}} = -\Sigma\frac{\beta_{\mathrm{A}}}{\beta_{\mathrm{P}}} = -\frac{1}{\nu_{\mathrm{P}}}\frac{k_{\mathrm{g,P}}}{k_{\mathrm{g,A}}}\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}} \tag{6.68}$$

Thus, in this regime the reaction rate constants do not influence selectivity. Furthermore, as soon as some P has been produced and is present in the bulk, the selectivity is always negative. This situation, of course, must always be prevented if P is the desired product.

The selectivity, in terms of "near surface" concentrations, is (see Eq. 6.61)

$$\sigma_{\mathrm{P}} = 1 - \frac{1}{\nu_{\mathrm{P}}}\frac{k''_{\mathrm{X}}}{k''_{\mathrm{P}}}\frac{C_{\mathrm{P,i}}}{C_{\mathrm{A,i}}} =$$

$$1 - \frac{1}{\nu_{\mathrm{P}}}\frac{k''_{\mathrm{X}}}{k''_{\mathrm{P}}}\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}} \times \frac{\overline{C_{\mathrm{A}}}}{C_{\mathrm{A,i}}}\frac{C_{\mathrm{P,i}}}{\overline{C_{\mathrm{P}}}} \tag{6.69}$$

Due to mass-transfer limitation, the term $\frac{\overline{C_{\mathrm{A}}}}{C_{\mathrm{A,i}}}\frac{C_{\mathrm{P,i}}}{\overline{C_{\mathrm{P}}}}$ may increase from unity (slow reaction regime) to infinity (fast reaction

regime). Hence, selectivity decreases tremendously with increasing mass-transfer limitation. Consequently, mass-transfer limitations must be avoided with heterogeneous consecutive reactions $A \longrightarrow P \longrightarrow X$. Therefore, a temperature increase sufficient to enhance the reaction rates always spoils the selectivity, even if both reactions have equal activation energies, because reaction rates usually increase much faster with temperature than mass-transfer rates.

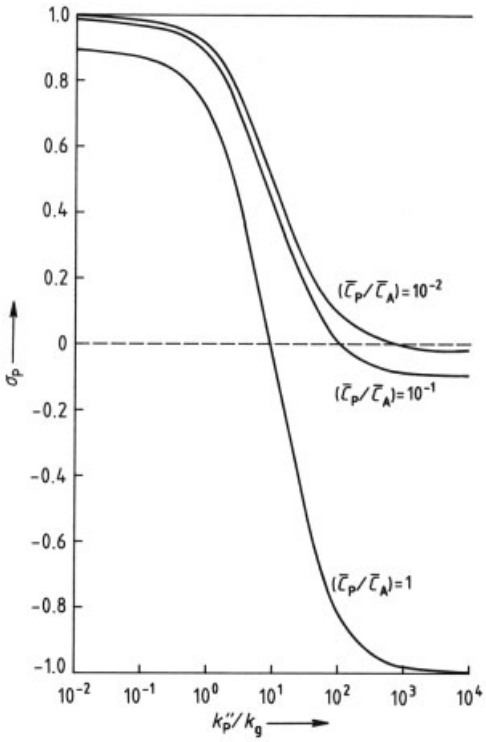

Figure 29. Selectivity σ_P as a function of k''_P/k_g for mass transfer and reaction in series with consecutive first-order reactions $A \longrightarrow P \longrightarrow X$ under the assumption that $k_{g,A} = k_{g,P} = k_g$ and $k''_X/k''_P = 1$
Several values of $\overline{C_P}/\overline{C_A}$ have been substituted.

Some other aspects are noteworthy. If no mass-transfer limitations occur, selectivity for the condition $\overline{C_P} = 0$ (e.g., initial selectivity at the reactor inlet) always equals 1, since $-\mathbb{R}''_P(\overline{C_P})- = 0$ because of the absence of P. When mass transfer is significant, this is no longer true. From Equation (6.66), for $\overline{C_P} = 0$ (i.e., $\Sigma = 0$),

$$\sigma_P = \left(1 + \frac{k''_X}{k_{g,P}}\right)^{-1} \tag{6.70}$$

which is only 1 for $k''_X/k_{g,P} << 1$ (i.e., in the slow reaction regime). Again, the reason for this decrease in selectivity is accumulation of P at the interface.

In general, the selectivity σ_P depends strongly on k''_X/k_g. Figure 29 shows how selectivity decreases dramatically with this ratio for a fixed value of k''_X/k''_P, several values of $\overline{C_P}/\overline{C_A}$, and $k_g = k_{g,A} = k_{g,P}$. For low ratios of k''_X/k_g, selectivities are very high due to the selected value of $k''_X/k''_P = 0.1$. With increasing k''_X/k_g, on entering the region of mass-transfer limitation, σ_P decreases rapidly to become constant again roughly for $k''_X/k_g \approx 10^3$. In the latter case the reactions are completely mass-transfer limited; hence σ_P becomes independent of the reaction rate constants and also of the ratio k''_X/k_g. Notice, the occurrence of negative values for σ_P, indicating a net consumption of P.

6.1.7.4. Complex Reaction Schemes

For more complex reaction schemes, selectivity calculations for mass transfer and reaction in series are always as straightforward as discussed above. An illustration for the following scheme is given:

where all three reactions are assumed to be first order. The rates $k''_{AP}\,C_A$, $k''_{AX}\,C_A$, and $k''_{PX}C_P$ have the dimensions mol m^{-2} s^{-1}.

For this scheme the selectivity is found as

$$\sigma_P = \frac{\nu_P k''_{AP} C_{A,i} - k''_{PX} C_{P,i}}{\nu_P k''_{AP} C_{A,i} + \nu_P k''_{AX} C_{A,i}} \tag{6.71}$$

σ_P is always smaller than unity and can again become negative, indicating a net consumption of P.

The interface concentrations follow from the mass balance for A and P:

$$J_A = k_{g,A}\left(\overline{C_A} - C_{A,i}\right) = k''_{AP} C_{A,i} + k''_{AX} C_{A,i} \tag{6.72}$$

$$J_P = k_{g,P}\left(C_{P,i} - \overline{C_P}\right) = \nu_P k''_{AP} C_{A,i} - k''_{PX} C_{P,i} \tag{6.73}$$

Combining Equations (6.71) – (6.73) gives the following expression for selectivity:

$$\sigma_{\mathrm{P}} = \frac{1}{1+\Sigma_{\mathrm{p}}} \cdot \frac{1-\Sigma_{\mathrm{c}}\,(1+\beta_{\mathrm{A}})}{1+\beta_{\mathrm{P}}}$$

$$\Sigma_{\mathrm{p}} = \frac{k''_{\mathrm{AX}}}{k''_{\mathrm{AP}}}\,,\ \Sigma_{\mathrm{c}} = \frac{1}{\nu_{\mathrm{P}}}\,\frac{k''_{\mathrm{PX}}}{k''_{\mathrm{AP}}}\,\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}}$$

$$\beta_{\mathrm{A}} = \frac{k''_{\mathrm{AP}}+k''_{\mathrm{AX}}}{k_{\mathrm{g,\,A}}}\,,\ \beta_{\mathrm{P}} = \frac{k''_{\mathrm{X}}}{k_{\mathrm{g,\,P}}} \tag{6.74}$$

Notice the resemblance between Equations (6.66) and (6.74). For the above reaction scheme, conclusions similar to those in the previous sections can be drawn. The only point highlighted here is that in the fast reaction regime (i.e., β_{A} and $\beta_{\mathrm{P}} \to \infty$),

$$\sigma_{\mathrm{P}} = -\frac{1}{\nu_{\mathrm{P}}}\,\frac{k_{\mathrm{g,\,P}}}{k_{\mathrm{g,\,A}}}\,\frac{\overline{C_{\mathrm{P}}}}{\overline{C_{\mathrm{A}}}}\,\frac{k''_{\mathrm{AP}}}{k''_{\mathrm{AP}}+k''_{\mathrm{AX}}} \tag{6.75}$$

hence, the reaction scheme discussed here differs from a purely consecutive reaction insofar as, in the fast reaction regime, selectivity is still a function of the kinetic constants.

6.2. Mass Transfer with Simultaneous Reaction

6.2.1. Introduction

The analysis in Section 6.1 did not require the Higbie penetration, surface renewal, or stagnant film models: values of the mass-transfer coefficients alone were sufficient. If mass transfer and reaction occur simultaneously, the distinction between the models can become significant. Simultaneous mass transfer and reaction mean:

1) For the Higbie penetration and surface renewal models: During the contact time of the element, mass transfer and reaction within the element interact.
2) For the stagnant film model: Mass transfer and reaction within the film exhibit interaction. Consequently, the concentration profile in the film need not be a straight line as was the case for mass transfer without reaction.

In this section, two general situations are discussed:

1) Reaction with diffusion of gases in a porous catalyst, where gaseous components react on the internal catalyst surface. This situation generally occurs in heterogeneous catalysis. For this application, either an interface covered with small stagnant fluid elements or a stagnant film with negligible curvature somewhere inside the pellet is difficult to imagine. Therefore the penetration and film models are less applicable here. Consequently, calculations for this type of problem are usually based on basic mass balances instead (e.g., Eq. 5.6).
2) Reaction with mass transfer in a liquid phase, where the reactant or at least one of the reactants is absorbed from a gas phase that is in contact with the liquid. Here the penetration and film models appear and are readily applicable, as can be illustrated.

In this section, mass transfer and reaction are assumed to occur simultaneously; the phenomenon of mass transfer and reaction in series is not considered:

1) For simultaneous reaction and diffusion of gases in a porous solid, mass-transfer resistance outside the solid is assumed to be infinitely small, so it need not be considered
2) For simultaneous reaction and mass transfer of gases in the liquid, mass transfer on the gas-phase side of the gas – liquid interface is assumed to be infinitely fast, so it need not be taken into account

The coupling of simultaneous and series mass transfer and reaction is discussed in Section 6.3.

6.2.2. Reaction of Gases in Porous Solids

6.2.2.1. Introduction

For the above situation the conversion is determined mainly by the following:

Microkinetic Scale. Conversion depends on

1) *Reaction properties*, such as the reaction kinetics, the number of reactants involved, and the complexity of the reaction scheme under consideration
2) *Catalyst pellet microproperties*, such as the pore-size distribution and tortuosity of the pores
3) *Physical properties*, such as the diffusion rates of the reacting components in the fluid phase and under Knudsen flow

Macrokinetic Scale. Conversion depends on the *catalyst pellet macroproperties*, such as the size and shape of the pellet and the possible occurrence of pellet anisotropy.

The microproperties of the catalyst pellet cannot be determined easily. Moreover, their interaction with the physical properties of the components is rather complex. Therefore, models that take microproperties into account tend to be very complex and inaccurate (e.g., [31]). Consequently, the microproperties of the catalyst and the physical properties of the components are usually accounted for by a lumped parameter, the so-called *effective diffusion coefficient* $\mathbb{D}_{\mathrm{i}}$. An often-used approximation for $\mathbb{D}_{\mathrm{i}}$ is

$$\mathbb{D}_{\mathrm{i}} = \frac{\varepsilon_{\mathrm{p}}}{\gamma_{\mathrm{p}}} \left(\frac{1}{\mathbb{D}_{\mathrm{fl}}} + \frac{1}{\mathbb{D}_{\mathrm{Kn}}} \right)^{-1} \tag{6.76}$$

with ε_{p} the particle porosity, γ_{p} the tortuosity of the particle pores, $\mathbb{D}_{\mathrm{fl}}$ the molecular diffusion coefficient of the component in the fluid (m^2/s), and $\mathbb{D}_{\mathrm{Kn}}$ its Knudsen diffusion coefficient (m^2/s), which is the diffusion coefficient found if the mean free path of the molecules is much larger than the average pore diameter of the particle. This is relevant only for a gas, since the mean free path of the molecules for liquids is on the order of the molecular dimensions and, hence, always much smaller than the pore diameter. Equation (6.76) is not applicable if the surface migration (sometimes called surface diffusion) of the component plays a role.

Using $\mathbb{D}_{\mathrm{i}}$ as lumping parameter for solid catalyst particles has proven useful, provided the particles can be regarded as being homogeneous on a microscale. However, if this is not the case (e.g., for zeolites in an amorphous matrix), care should be taken in using this approach [32], [33].

The concept of an effective diffusion coefficient is assumed not to introduce too large an error. Hence, the effect of catalyst microproperties and physical properties of the components are not examined here, and the value of $\mathbb{D}_{\mathrm{i}}$ is assumed to be known.

6.2.2.2. Monomolecular First-Order Kinetics

Assume that a simple reaction $\mathrm{A} \longrightarrow \nu_{\mathrm{P}}\ \mathrm{P}$ is carried out in a spherical, porous catalyst pellet with a radius r_{p}. A mass balance for an infinitesimally thin, spherical slice within the particle yields (compare with Eq. 5.6):

$$\mathbb{D}_{\mathrm{i,\,A}} \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial C_{\mathrm{A}}}{\partial r} \right) = |\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}} \tag{6.77}$$

where the coordinate r is zero in the sphere center and $r = r_{\mathrm{p}}$ at the outer surface; $-\!\!-\mathbb{R}_{\mathrm{A}}-\!\!-_{\mathrm{p}}$ is the chemical reaction rate in moles of A per cubic meter of catalyst particle volume per second, as used before. This rate is not the same as the rate $-\!\!-\mathbb{R}_{\mathrm{A}}-\!\!-_{\mathrm{r}}$ in moles of A per cubic meter of reactor volume per second. If $1-\varepsilon$ is the volume holdup of porous catalyst material in a reactor (i.e., ε is the porosity or void fraction), the two rates are related according to

$$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{r}} = (1-\varepsilon)\ |\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}} \tag{6.78}$$

To calculate the concentration profile inside the spherical particle, Equation (6.77) must be solved subject to the following boundary conditions:

$$r = 0 \Rightarrow C_{\mathrm{A}} \text{ is bounded} \tag{6.79}$$

$$r = r_{\mathrm{p}} \Rightarrow C_{\mathrm{A}} = C_{\mathrm{A,\,i}} \tag{6.80}$$

here $C_{\mathrm{A,\,i}}$ is the concentration of A in the gas at the interface with the pellet. For first-order kinetics, i.e.,

$$|\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}} = k_{\mathrm{p}} C_{\mathrm{A}} \tag{6.81}$$

solution of the differential equation gives

$$\frac{C_{\mathrm{A}}}{C_{\mathrm{A,i}}} = \frac{r_{\mathrm{p}}}{r} \frac{\sinh\left(\varphi_{\mathrm{T}} \frac{r}{r_{\mathrm{p}}}\right)}{\sinh\left(\varphi_{\mathrm{T}}\right)} \tag{6.82}$$

with φ_{T} the *Thiele modulus* defined as

$$\varphi_{\mathrm{T}} = r_{\mathrm{p}} \sqrt{\frac{k_{\mathrm{p}}}{\mathbb{D}_{\mathrm{i,\,A}}}} \tag{6.83}$$

The conversion rate of A for one pellet can be calculated from

$$\begin{aligned} 4\pi r_{\mathrm{p}}^2 J_{\mathrm{A,\,p}} &= 4\pi r_{\mathrm{p}}^2 \mathbb{D}_{\mathrm{i,\,A}} \frac{\partial C_{\mathrm{A}}(r)}{\partial r}\Big|_{r=r_{\mathrm{p}}} \\ &= 4\pi r_{\mathrm{p}} \mathbb{D}_{\mathrm{i,\,A}} C_{\mathrm{A,\,i}} \left(\frac{\varphi_{\mathrm{T}}}{\tanh\left(\varphi_{\mathrm{T}}\right)} - 1 \right) \end{aligned} \tag{6.84}$$

Use of the *effectiveness factor* is convenient. THIELE was one of the first to use this concept [34]. He defined the effectiveness factor as the

amount of a key reactant that is actually converted, divided by the amount that could have been converted if no retardation by diffusion had occurred. In this case the amount of A converted would equal $4/3\,\pi\, r_p^3\, k_p\, C_{A,i}$ (moles of A per m^3 pellet per unit time); hence, the effectiveness factor η is

$$\eta = \frac{3}{\varphi_T}\left(\frac{1}{\tanh(\varphi_T)} - \frac{1}{\varphi_T}\right) \tag{6.85}$$

In a similar way, equations are derived for other catalyst geometries. That is for an infinitely long cylinder with a radius r_p, Equation (6.77) must be replaced by

$$\mathbb{D}_{i,A}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C_A}{\partial r}\right) = |\mathbb{R}_A|_p \tag{6.86}$$

Again assuming first-order kinetics and going through the same routine as discussed above gives

$$\eta = 2\,\frac{I_1(\varphi_T)}{\varphi_T\, I_0(\varphi_T)} \tag{6.87}$$

with φ_T as given by Equation (6.83), and I_0 and I_1 being modified Bessel functions of the first kind of zeroth and first order, respectively.

For an infinitely large slab with r_p half the thickness of the slab, finally, Equation (6.77) must be replaced by

$$\mathbb{D}_{i,A}\frac{\partial^2 C_A}{\partial r^2} = |\mathbb{R}_A|_p \tag{6.88}$$

which, if first-order kinetics are assumed, can be shown to yield

$$\eta = \frac{\tanh(\varphi_T)}{\varphi_T} \tag{6.89}$$

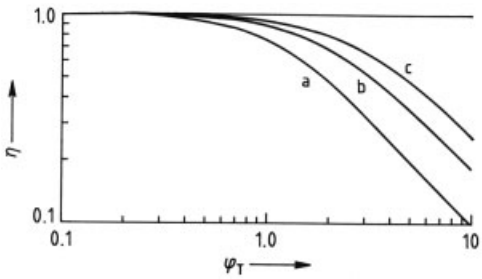

Figure 30. Effectiveness factor η versus "simple" Thiele modulus φ_T (not shape generalized) for first-order kinetics and the following catalyst particle geometries
a) Infinite slab; b) Infinite cylinder; c) Sphere

Equations (6.85), (6.87), and (6.89) are illustrated in Figure 30, where the effectiveness factor η is plotted versus the Thiele modulus φ_T for all three geometries discussed. As the Thiele modulus increases, the reaction becomes more limited by diffusion and thus the effectiveness factor decreases. For very low values of the Thiele modulus the effectiveness factor is unity; for very high values it is inversely proportional to the Thiele modulus.

The Thiele modulus may be regarded as a measure of the ratio of the reaction rate to the rate of diffusion. However, many different definitions are used in the literature. The best definition is a general one, which allows direct comparison of the Thiele moduli for arbitrary catalyst geometries. This definition is discussed below.

Aris was the first who noticed that all $\eta - \varphi_T$ curves can be brought together in the low η region if, instead of r_p, the characteristic dimension V_p/A_p is used, where V_p is the pellet volume and A_p the external surface area of the pellet [35]. Thus he defined the *shape-generalized Thiele modulus* for first-order kinetics as

$$\varphi_A = \frac{V_p}{A_p}\sqrt{\frac{k_p}{\mathbb{D}_{i,A}}} \tag{6.90}$$

In Figure 31, η is plotted versus φ_A for the three geometries discussed. The curves coincide in both the high and the low η region. In the intermediate region the spread among the curves is largest, but still limited.

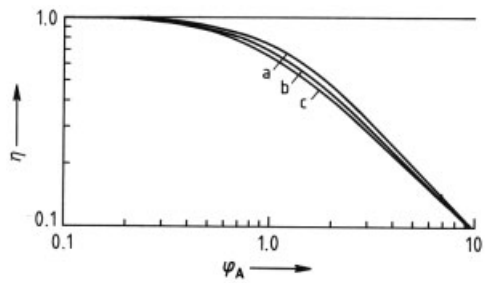

Figure 31. Effectiveness factor η versus the Aris shape-generalized Thiele modulus φ_A for first-order kinetics and the following catalyst particle geometries
a) Infinite slab; b) Infinite cylinder; c) Sphere

With the aid of the Aris' shape-generalized Thiele modulus, general statements on the possible regimes can be made that are met in reactions with internal diffusion:

Slow Reaction Regime ($\varphi_A < 0.2$). The effectiveness factor can be assumed to be unity, and mass transfer inside the pellet by diffusion is so fast that it can be neglected.

Fast Reaction Regime ($\varphi_A > 2$). Internal mass transfer by diffusion is significant. In this regime the conversion rate $-R_A-_p$ is linked to the reaction rate $-\mathbb{R}_A-_p$ as

$$|R_A|_p = \eta |\mathbb{R}_A|_p = \frac{|\mathbb{R}_A|_p}{\varphi_A} \tag{6.91}$$

Equation (6.91) implies for first-order kinetics that the apparent rate or conversion rate to be observed in the fast reaction regime equals

$$|R_A|_p = \frac{A_p}{V_p}\sqrt{k_p \mathbb{D}_{i,A}}\, C_{A,i} \tag{6.92}$$

which is different from the reaction rate that would be found if no diffusion limitation occurred:

$$|\mathbb{R}_A|_p = k_p C_{A,i} \tag{6.93}$$

Consequently, for $\varphi_A > 2$, the conversion is a strong function of catalyst particle diameter. Also the apparent activation energy of the reaction is only half of the true activation energy, since the internal diffusion coefficient is almost independent of temperature compared to the rate constant k_p.

The power of the Aris shape-generalized Thiele modulus can be illustrated further by using a ring-shaped catalyst or Raschig ring as an example (see Fig. 32). For this geometry the shape-generalized Thiele modulus is found to be

$$\varphi_A = \frac{1}{2}\left(\frac{1}{\lambda} + \frac{1}{1-\iota}\right)^{-1} R_u \sqrt{\frac{k_p}{\mathbb{D}_{i,A}}} \tag{6.94}$$

with

$$\lambda = \frac{H}{R_u} \tag{6.95}$$

and

$$\iota = \frac{R_i}{R_u} \tag{6.96}$$

Now the mass balance (Eq. 5.6) reduces to

$$\mathbb{D}_{i,A}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C_A}{\partial r}\right) + \frac{\partial^2 C_A}{\partial h^2}\right] = |\mathbb{R}_A|_p \tag{6.97}$$

which can be solved subject to the boundary conditions

$$r = R_i \Rightarrow C_A = C_{A,i} \tag{6.98}$$

$$r = R_u \Rightarrow C_A = C_{A,i} \tag{6.99}$$

$$h = 0 \Rightarrow C_A = C_{A,i} \tag{6.100}$$

$$h = H \Rightarrow C_A = C_{A,i} \tag{6.101}$$

Calculating the effectiveness factor as illustrated above gives

$$\eta = 1 - \frac{32}{\pi^2}\cdot\frac{\varphi_A^2}{1-\iota^2}\left(\frac{1}{\lambda} + \frac{1}{1-\iota}\right)^2 \cdot \sum_{m=1}^{\infty}\frac{1-\iota^2 - 2f(\iota,\alpha_m)}{(2m+1)^2\alpha_m^2} \tag{6.102}$$

with

$$\alpha_m = \sqrt{\left[\frac{(2m+1)\pi}{\lambda}\right]^2 + \left[2 + \left(\frac{1}{\lambda} + \frac{1}{1-\iota}\right)\varphi_A\right]^2} \tag{6.103}$$

$$f(x,y) = \frac{2 - y\,Q_0(xy,y) - xy\,Q_0(y,xy)}{y^2 P_0(xy,y)} \tag{6.104}$$

$$Q_0(x,y) = I_0(x)K_1(y) + I_1(y)K_0(x) \tag{6.105}$$

$$P_0(x,y) = I_0(x)K_0(y) - I_0(y)K_0(x) \tag{6.106}$$

where I_n and K_n are n-th-order modified Bessel functions of the first and second kind, respectively. Hence, for relatively simple catalyst geometries and reaction kinetics, the mathematics can be quite involved and the resulting formulas can become quite complex and tedious. For this reason the effectiveness factor is often approximated by the following relation:

$$\eta = \frac{\tanh(\varphi_A)}{\varphi_A} \tag{6.107}$$

Alternatively, the following approximation is sometimes used:

$$\eta = \frac{1}{\sqrt{1+\varphi_A^2}} \tag{6.108}$$

The above approximations can be used only if the shape-generalized Thiele modulus is employed. Both approximations are almost exact in the regimes $\varphi_A < 0.2$ and $\varphi_A > 2$. The largest error is found in the intermediate region and can be expected to be at most roughly 10 % for first-order reactions.

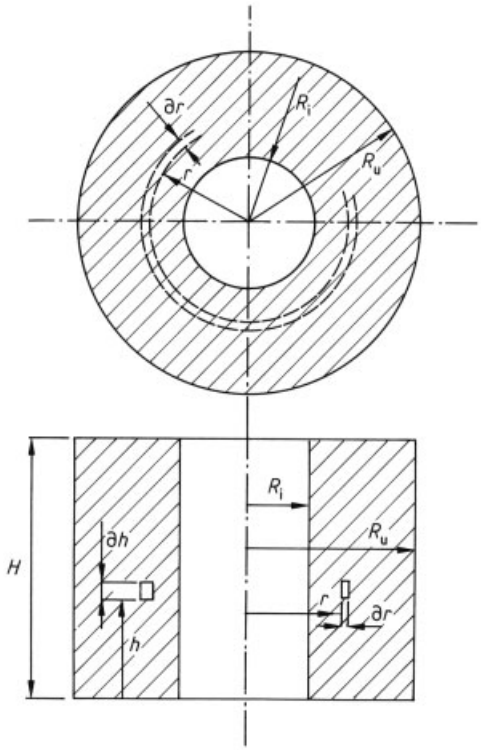

Figure 32. Microscale mass balance for catalyst pellet with the shape of a Raschig ring

6.2.2.3. Monomolecular Arbitrary Kinetics

Like the shape generalization of the Thiele modulus suggested by ARIS [35], a generalization for reaction kinetics can be made. This was introduced by BISCHOFF [36]. If the effective diffusion coefficient is independent of the concentration of A, for a Thiele modulus defined as

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \left| \mathbb{R}_\mathrm{A} \left(C_\mathrm{A,i} \right) \right|_\mathrm{p} \cdot \left(2\mathbb{D}_\mathrm{i,A} \int_0^{C_\mathrm{A,i}} \left| \mathbb{R}_\mathrm{A} \left(C_\mathrm{A} \right) \right|_\mathrm{p} \partial C_\mathrm{A} \right)^{-\frac{1}{2}} \tag{6.109}$$

in the low η region

$$\left| \mathbb{R}_\mathrm{A} \right|_\mathrm{p} = \frac{\left| \mathbb{R}_\mathrm{A} \right|_\mathrm{p}}{\varphi} \tag{6.110}$$

Hence, again in the low η region, all $\eta - \varphi$ curves are brought together, with this Thiele modulus for both arbitrary catalyst geometries and arbitrary kinetics. In what follows the Thiele modulus defined according to Equation (6.110) is used.

For *simple n-th-order kinetics*, i.e.,

$$\left| \mathbb{R}_\mathrm{A} \right|_\mathrm{p} = k_\mathrm{p} C_\mathrm{A}^n \tag{6.111}$$

Equation (6.109) yields

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{n+1}{2} \frac{k_\mathrm{p} C_\mathrm{A,i}^{n-1}}{\mathbb{D}_\mathrm{i,A}}} \tag{6.112}$$

which for $n = 1$ reduces to the shape-generalized Thiele modulus of Aris for first-order kinetics (Eq. 6.90):

$$\varphi = \varphi_\mathrm{A} = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{k_\mathrm{p}}{\mathbb{D}_\mathrm{i,A}}} \tag{6.113}$$

As a second illustration, for *Langmuir kinetics*,

$$\left| \mathbb{R}_\mathrm{A} \right|_\mathrm{p} = \frac{k_\mathrm{p} C_\mathrm{A}}{1 + K C_\mathrm{A}} \tag{6.114}$$

and it is found that

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \frac{K C_\mathrm{A,i}}{1 + K C_\mathrm{A,i}} \cdot \sqrt{\frac{k_\mathrm{p}}{2\mathbb{D}_\mathrm{i,A} \left[K C_\mathrm{A,i} - \ln \left(1 + K C_\mathrm{A,i} \right) \right]}} \tag{6.115}$$

For other kinetics, similar equations can readily be derived.

The mathematics necessary to derive the exact equations for η can be quite involved even for first-order kinetics. For arbitrary kinetics the problem is still worse: for most forms of kinetics, for example, Equations (6.77), (6.86), (6.88), and (6.97) cannot be solved analytically, so numerical techniques must be used. This explains why, for arbitrary kinetics, more often than not, approximations are used. The most popular approximations are similar to Equations (6.107) and (6.108):

$$\eta = \frac{\tanh(\varphi)}{\varphi} \tag{6.116}$$

or

$$\eta = \frac{1}{\sqrt{1 + \varphi^2}} \tag{6.117}$$

However, the error introduced by using one of the above approximations can be quite substantial. In general, it becomes larger for lower (apparent) reaction orders. Therefore, use of the

above approximations for (apparent) reaction orders below zero is not recommended without further studies. Only in the low η regime, where $\eta = 1\varphi$ is valid, are Equations (6.116) and (6.117) trustworthy. For (apparent) reaction orders above zero the equations give a fair indication, and maximum errors up to roughly 20 % must be expected in the intermediate region.

6.2.2.4. Bimolecular Reactions

For bimolecular reactions of the type

$$\mathrm{A} + \nu\,\mathrm{B} \longrightarrow \text{Products}$$

the concentrations C_A and C_B inside the catalyst pellet are interrelated via the microscale mass balance, which yields for arbitrary catalyst geometries

$$\mathbb{D}_{\mathrm{i,B}}\left(C_{\mathrm{B,i}} - C_\mathrm{B}\right) = \nu \mathbb{D}_{\mathrm{i,A}}\left(C_{\mathrm{A,i}} - C_\mathrm{A}\right) \tag{6.118}$$

For convenience, the *excess number for bimolecular reactions* is introduced [37]:

$$\gamma = \nu \frac{\mathbb{D}_{\mathrm{i,A}}}{\mathbb{D}_{\mathrm{i,B}}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{B,i}}} \tag{6.119}$$

Component A is chosen as the key reactant, so that B is present in excess; thus

$$0 < \gamma < 1 \tag{6.120}$$

$\gamma = 0$ represents a large excess of component B; if $\gamma = 1$ the molar quantities of A and B just match each other.

Equation (6.118) can now be written as

$$\frac{C_\mathrm{B}}{C_{\mathrm{B,i}}} = 1 - \gamma\left(1 - \frac{C_\mathrm{A}}{C_{\mathrm{A,i}}}\right) \tag{6.121}$$

By using this equation, the reaction rate, generally a function of both C_A and C_B, can be written as a function of C_A only:

$$\begin{aligned} &|\mathbb{R}\left(C_\mathrm{A}, C_\mathrm{B}\right)|_\mathrm{p} \\ &= |\mathbb{R}\left\{C_\mathrm{A}, C_{\mathrm{B,i}}\left[1 - \gamma\left(1 - C_\mathrm{A}/C_{\mathrm{A,i}}\right)\right]\right\}|_\mathrm{p} \end{aligned} \tag{6.122}$$

Substituting this in the definition of the Thiele modulus (Eq. 6.109) gives the generalized Thiele modulus for bimolecular reactions:

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} |\mathbb{R}_\mathrm{A}\left(C_{\mathrm{A,i}}, C_{\mathrm{B,i}}\right)|_\mathrm{p} \cdot \left(2\mathbb{D}_{\mathrm{i,A}} \int_0^{C_{\mathrm{A,i}}} |\mathbb{R}_\mathrm{A} \left\{C_\mathrm{A}, C_{\mathrm{B,i}}\left[1 - \gamma\left(1 - C_\mathrm{A}/C_{\mathrm{A,i}}\right)\right]\right\}|_\mathrm{p}\, \partial C_\mathrm{A}\right)^{-\frac{1}{2}} \tag{6.123}$$

For example, for (1,1)-kinetics

$$|\mathbb{R}\left(C_\mathrm{A}, C_\mathrm{B}\right)|_\mathrm{p} = k_\mathrm{p} C_\mathrm{A} C_\mathrm{B} \tag{6.124}$$

Equation (6.123) yields

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{k_\mathrm{p} C_{\mathrm{B,i}}}{\left(1 - \frac{\gamma}{3}\right)\mathbb{D}_{\mathrm{i,A}}}} \tag{6.125}$$

For $C_{\mathrm{B,i}}/C_{\mathrm{A,i}} \to \infty$ and hence $\gamma \to 0$, this yields

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{k_\mathrm{p} C_{\mathrm{B,i}}}{\mathbb{D}_{\mathrm{i,A}}}} \tag{6.126}$$

which is the generalized Thiele modulus for first-order kinetics with a rate constant $k_\mathrm{p}\, C_{\mathrm{B,i}}$. The physical explanation is that in view of the large excess of B, consumption of B in the pellet is negligible ($C_\mathrm{B} = C_{\mathrm{B,i}}$), so the rate equation reduces to

$$|\mathbb{R}\left(C_\mathrm{A}, C_\mathrm{B}\right)|_\mathrm{p} = \left(k_\mathrm{p} C_{\mathrm{B,i}}\right) C_\mathrm{A} \tag{6.127}$$

Similarly, for $\gamma = 1$, according to Equation (6.121), apparent second-order kinetics are obtained:

$$|\mathbb{R}\left(C_\mathrm{A}, C_\mathrm{B}\right)|_\mathrm{p} = \left(k_\mathrm{p} \frac{C_{\mathrm{B,i}}}{C_{\mathrm{A,i}}}\right) C_\mathrm{A}^2 \tag{6.128}$$

hence, according to Equation (6.112),

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{3}{2} \frac{k_\mathrm{p} C_{\mathrm{B,i}}}{\mathbb{D}_{\mathrm{i,A}}}} \tag{6.129}$$

This is consistent with Equation (6.125), as can be seen by substituting $\gamma = 1$.

The definition of the generalized Thiele modulus (Eq. 6.123) is readily applicable to bimolecular reactions with more complex kinetics. For example, for Langmuir – Hinshelwood kinetics, (see Eq. 3.75):

$$|\mathbb{R}\left(C_\mathrm{A}, C_\mathrm{B}\right)|_\mathrm{p} = \frac{k_\mathrm{p} C_\mathrm{A} C_\mathrm{B}}{\left(1 + K_\mathrm{A} C_\mathrm{A} + K_\mathrm{B} C_\mathrm{B}\right)^2} \tag{6.130}$$

Equation (6.123) yields

$$\varphi = \frac{V_\mathrm{p}}{A_\mathrm{p}} \sqrt{\frac{1}{2} \frac{k_\mathrm{p} C_{\mathrm{B,i}}}{\mathbb{D}_{\mathrm{i,A}}}} \cdot \frac{K_\mathrm{A} C_{\mathrm{A,i}} + \gamma K_\mathrm{B} C_{\mathrm{B,i}}}{\left(1 + K_\mathrm{A} C_{\mathrm{A,i}} + K_\mathrm{B} C_{\mathrm{B,i}}\right)^2} \cdot \sqrt{\left[2\gamma\left(1 + \beta\right) - 1\right]\left(\frac{1}{1 + \beta} - \ln\frac{1 + \beta}{\beta}\right)} \tag{6.131}$$

with

$$\beta = \frac{1+(1-\gamma)\,K_{\mathrm{B}} C_{\mathrm{B,i}}}{K_{\mathrm{A}}\,C_{\mathrm{A,i}} + \gamma K_{\mathrm{B}}\,C_{\mathrm{B,i}}} \tag{6.132}$$

In a similar way, expressions for other types of kinetics can be found.

If the reaction rate depends on the concentrations of more than two components, the Thiele moduli can be calculated by using the same approach outlined above. For example, for the following equilibrium:

$$>\mathrm{A} + \nu_{\mathrm{B}}\,\mathrm{B} \rightleftharpoons \nu_{\mathrm{P}}\,\mathrm{P} + \nu_{\mathrm{Q}}\,\mathrm{Q}$$

where A is the key reactant (thus B is present in excess and the overall conversion is to the right), the rate equation is given by

$$|\mathbb{R}\left(C_{\mathrm{A}}, C_{\mathrm{B}}, C_{\mathrm{P}}, C_{\mathrm{Q}}\right)|_{\mathrm{p}} = k_{\mathrm{p}}\left(C_{\mathrm{A}} C_{\mathrm{B}} - \frac{C_{\mathrm{P}}\,C_{\mathrm{Q}}}{K}\right) \tag{6.133}$$

As in the derivation for bimolecular reactions,

$$\gamma_{\mathrm{B}} = \nu_{\mathrm{B}} \frac{\mathbb{D}_{\mathrm{i,A}}}{\mathbb{D}_{\mathrm{i,B}}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{B,i}}} \tag{6.134}$$

$$\gamma_{\mathrm{P}} = \nu_{\mathrm{P}} \frac{\mathbb{D}_{\mathrm{i,A}}}{\mathbb{D}_{\mathrm{i,P}}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{P,i}}} \tag{6.135}$$

$$\gamma_{\mathrm{Q}} = \nu_{\mathrm{Q}} \frac{\mathbb{D}_{\mathrm{i,A}}}{\mathbb{D}_{\mathrm{i,Q}}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{Q,i}}} \tag{6.136}$$

Microscale mass balances now yield

$$\frac{C_{\mathrm{B}}}{C_{\mathrm{B,i}}} = 1 - \gamma_{\mathrm{B}}\left(1 - \frac{C_{\mathrm{A}}}{C_{\mathrm{A,i}}}\right) \tag{6.137}$$

$$\frac{C_{\mathrm{P}}}{C_{\mathrm{P,i}}} = 1 + \gamma_{\mathrm{P}}\left(1 - \frac{C_{\mathrm{A}}}{C_{\mathrm{A,i}}}\right) \tag{6.138}$$

$$\frac{C_{\mathrm{Q}}}{C_{\mathrm{Q,i}}} = 1 + \gamma_{\mathrm{Q}}\left(1 - \frac{C_{\mathrm{A}}}{C_{\mathrm{A,i}}}\right) \tag{6.139}$$

Hence the reaction rate can be expressed in concentration C_{A} only (cf. Eq. 6.133):

$$\begin{aligned} &|\mathbb{R}\left(C_{\mathrm{A}}, C_{\mathrm{B}}, C_{\mathrm{P}}, C_{\mathrm{Q}}\right)|_{\mathrm{p}} \\ &= k_{\mathrm{p}}\left\{C_{\mathrm{A}} C_{\mathrm{B,i}}\left[1 - \gamma_{\mathrm{B}}\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right]\right. \\ &- \frac{C_{\mathrm{P,i}} C_{\mathrm{Q,i}}}{K}\left[1 + \gamma_{\mathrm{P}}\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right] \\ &\left.\cdot\left[1 + \gamma_{\mathrm{Q}}\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right]\right\} \end{aligned} \tag{6.140}$$

Substitution in the definition of the Thiele modulus gives

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}}\,|\mathbb{R}_{\mathrm{A}}\left(C_{\mathrm{A,i}}, C_{\mathrm{B,i}}, C_{\mathrm{P,i}}, C_{\mathrm{Q,i}}\right)|_{\mathrm{p}} \cdot \left(2\,\mathbb{D}_{\mathrm{i,A}} \int_0^{C_{\mathrm{A,i}}} |\mathbb{R}_{\mathrm{A}}\left(C_{\mathrm{A}}, C_{\mathrm{B}}, C_{\mathrm{P}}, C_{\mathrm{Q}}\right)|_{\mathrm{p}}\,\partial C_{\mathrm{A}}\right)^{-\frac{1}{2}} \tag{6.141}$$

which yields the generalized Thiele modulus for equilibrium reactions:

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}}\left(1 - \frac{C_{\mathrm{P,i}} C_{\mathrm{Q,i}}}{K C_{\mathrm{A,i}} C_{\mathrm{B,i}}}\right) \cdot \sqrt{\frac{k_{\mathrm{p}} C_{\mathrm{B,i}}}{\mathbb{D}_{\mathrm{i,A}}\left[1 - \frac{\gamma_{\mathrm{B}}}{3} - \frac{C_{\mathrm{P,i}} C_{\mathrm{Q,i}}}{K C_{\mathrm{A,i}} C_{\mathrm{B,i}}}\left(2 + \gamma_{\mathrm{P}} + \gamma_{\mathrm{Q}} - \frac{\gamma_{\mathrm{P}}\gamma_{\mathrm{Q}}}{3}\right)\right]}} \tag{6.142}$$

For bimolecular reactions, the mathematics necessary to derive the exact equations for η are generally impossible to solve and numerical techniques must be used. Thus to estimate η, approximation equations (Eqs. 6.116 and 6.117) are generally used. Again, the error introduced when using one of the above approximations can be quite substantial, especially for low (apparent) reaction orders. Equations (6.116) and (6.117) are exact only in the low η regime, where $\eta = 1\varphi$.

6.2.2.5. Anisotropic Catalyst Pellets

Cylindrical catalyst pellets, with or without a hole in the center, are usually shaped by extrusion of a "wet" catalyst paste. During this process the diameter of the pores in the radial direction may become smaller, and the length of pores in the longitudinal direction may decrease. This can result in the radial effective diffusivity being lower than the longitudinal one. Thus cylindrical catalyst pellets can be anisotropic. This anisotropy can be accounted for in the Thiele modulus. Consider, for example, a ring-shaped catalyst pellet; here, for monomolecular reactions a microscale mass balance yields (see Eq. 6.97)

$$\mathbb{D}^{r}_{\mathrm{i,A}} \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C_{\mathrm{A}}}{\partial r}\right) + \mathbb{D}^{h}_{\mathrm{i,A}} \frac{\partial^2 C_{\mathrm{A}}}{\partial h^2} = |\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}} \tag{6.143}$$

If

$$r^* = r\sqrt{\frac{\mathbb{D}^{h}_{\mathrm{i,A}}}{\mathbb{D}^{r}_{\mathrm{i,A}}}} \tag{6.144}$$

is substituted, Equation (6.143) can be written as

$$\mathbb{D}^h_{\mathrm{i,A}}\left[\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\frac{\partial C_\mathrm{A}}{\partial r^*}\right)+\frac{\partial^2 C_\mathrm{A}}{\partial h^2}\right]=|\mathbb{R}_\mathrm{A}|_\mathrm{p} \quad (6.145)$$

The differential Equation (6.145) describes the concentration profile in an isotropic ring-shaped catalyst pellet with an effective diffusion coefficient $\mathbb{D}^h_{\mathrm{i,A}}$, an inner radius

$$R^*_\mathrm{i}=R_\mathrm{i}\sqrt{\frac{\mathbb{D}^h_{\mathrm{i,A}}}{\mathbb{D}^r_{\mathrm{i,A}}}} \quad (6.146)$$

and an outer radius

$$R^*_\mathrm{u}=R_\mathrm{u}\sqrt{\frac{\mathbb{D}^h_{\mathrm{i,A}}}{\mathbb{D}^r_{\mathrm{i,A}}}} \quad (6.147)$$

Hence, anisotropy can be accounted for in the Thiele modulus if the effective diffusion coefficient $\mathbb{D}_{\mathrm{i,A}}$ is replaced by $\mathbb{D}^h_{\mathrm{i,A}}$ and the characteristic dimension $V_\mathrm{p}/A_\mathrm{p}$ by

$$\left(\frac{V_\mathrm{p}}{A_\mathrm{p}}\right)^*=\frac{1}{2}\left(\frac{1}{H}+\frac{1}{R^*_\mathrm{u}-R^*_\mathrm{i}}\right)$$
$$=\frac{1}{2}\left(\frac{1}{H}\frac{1}{R_\mathrm{u}-R_\mathrm{i}}\sqrt{\frac{\mathbb{D}^r_{\mathrm{i,A}}}{\mathbb{D}^h_{\mathrm{i,A}}}}\right) \quad (6.148)$$

Thus, the following equation is obtained for the generalized Thiele modulus for monomolecular reactions (cf. Eq. 6.109):

$$\varphi=\left(\frac{V_\mathrm{p}}{A_\mathrm{p}}\right)^*|\mathbb{R}_\mathrm{A}\left(C_\mathrm{A,i}\right)|_\mathrm{p}\cdot\left(2\mathbb{D}^h_{\mathrm{i,A}}\int_0^{C_\mathrm{A,i}}|\mathbb{R}_\mathrm{A}\left(C_\mathrm{A}\right)|_\mathrm{p}\,\partial C_\mathrm{A}\right)^{-\frac{1}{2}} \quad (6.149)$$

Rather than accounting for anisotropy by modifying both the characteristic dimension and the effective diffusion coefficient, anisotropy is accounted for by modifying the effective diffusion coefficient only. Thus, for anisotropic catalyst pellets a modified effective diffusion coefficient $\mathbb{D}^*_{\mathrm{i,A}}$ is defined. Consequently, for anisotropic catalyst pellets and monomolecular reactions the generalized Thiele modulus follows from

$$\varphi=\frac{V_\mathrm{p}}{A_\mathrm{p}}|\mathbb{R}_\mathrm{A}\left(C_\mathrm{A,i}\right)|_\mathrm{p}\cdot\left(2\mathbb{D}^*_{\mathrm{i,A}}\int_0^{C_\mathrm{A,i}}|\mathbb{R}_\mathrm{A}\left(C_\mathrm{A}\right)|_\mathrm{p}\,\partial C_\mathrm{A}\right)^{-\frac{1}{2}} \quad (6.150)$$

Comparison of Equations (6.148)–(6.150) shows that for ring-shaped catalyst pellets the modified effective diffusion coefficient $\mathbb{D}^*_{\mathrm{i,A}}$ follows from

$$\mathbb{D}^*_{\mathrm{i,A}}=\left(\frac{\frac{\sqrt{\mathbb{D}^r_{\mathrm{i,A}}}}{R_\mathrm{u}-R_\mathrm{i}}+\frac{\sqrt{\mathbb{D}^h_{\mathrm{i,A}}}}{H}}{\frac{1}{R_\mathrm{u}-R_\mathrm{i}}+\frac{1}{H}}\right)^2 \quad (6.151)$$

Therefore, all conclusions drawn for isotropic catalyst pellets also hold for anisotropic catalyst pellets. For example for bimolecular reactions,

$$\varphi=\frac{V_\mathrm{p}}{A_\mathrm{p}}|\mathbb{R}_\mathrm{A}\left(C_\mathrm{A,i},C_\mathrm{B,i}\right)|_\mathrm{p}\cdot\left(2\mathbb{D}^*_{\mathrm{i,A}}\int_0^{C_\mathrm{A,i}}|\mathbb{R}_\mathrm{A}\left\{C_\mathrm{A},C_\mathrm{B,i}\right.\right.$$
$$\left.\left.\left[1-\gamma\left(1-C_\mathrm{A}/C_\mathrm{A,i}\right)\right]\right\}|_\mathrm{p}\,\partial C_\mathrm{A}\right)^{-\frac{1}{2}} \quad (6.152)$$

to be compared with Equation (6.123). As a matter of fact, with a single, constant catalyst geometry, no distinction is possible between isotropic and anisotropic pellets. The effect of anisotropy is lumped in the effective diffusion coefficient. If the catalyst pellet is *isotropic*, from Equation (6.151) the effective diffusion coefficient $\mathbb{D}_{\mathrm{i,A}}=\mathbb{D}^*_{\mathrm{i,A}}=\mathbb{D}^r_{\mathrm{i,A}}=\mathbb{D}^h_{\mathrm{i,A}}$ is measured. For *anisotropic pellets*, $\mathbb{D}^r_{\mathrm{i,A}}$ for pellets with a large height and $\mathbb{D}^h_{\mathrm{i,A}}$ for flat pellets are measured. For intermediate geometries the value of $\mathbb{D}^*_{\mathrm{i,A}}$ is between $\mathbb{D}^r_{\mathrm{i,A}}$ and $\mathbb{D}^h_{\mathrm{i,A}}$. The only situation in which anisotropy can give rise to errors is when the catalyst geometry–thus $(R_u-R_\mathrm{i})/H$–is changed, and the value of the effective diffusion coefficient is used, which was determined for another catalyst geometry. The procedure discussed above can readily be extended to other catalyst geometries. Consider the purely academic example of an anisotropic parallelepiped with the dimesions of width×length×height = $W\times L\times H$. Here the Thiele modulus can still be calculated from Equations (6.150) and (6.152). However, the modifed effective diffusion coefficient now becomes

$$\mathbb{D}^*_{\mathrm{i,A}}=\left(\frac{\frac{\sqrt{\mathbb{D}^w_{\mathrm{i,A}}}}{W}+\frac{\sqrt{\mathbb{D}^l_{\mathrm{i,A}}}}{L}+\frac{\sqrt{\mathbb{D}^h_{\mathrm{i,A}}}}{H}}{\frac{1}{W}+\frac{1}{L}+\frac{1}{H}}\right)^2 \quad (6.153)$$

6.2.3. Reactions of Gases in Liquids – Part I: Gaseous Reactants Only

6.2.3.1. Introduction

In this section, simultaneous reaction and mass transfer in a liquid phase are discussed, where the reactant or all of the reactants are absorbed from a gas phase in contact with this liquid. The assumption is that mass transfer on the gas-phase side of the gas – liquid interface is infinitely fast, so that it need not be considered.

For gas – liquid reactions in general, two cases can be distinguished:

1) A reactant or reactants are absorbed from the gas to react in the liquid. No components already present in the liquid are consumed in the reaction.
2) A reactant is absorbed from the gas to react in the liquid with an already dissolved component.

The former situation is discussed in Section 6.2.3, the latter in Section 6.2.4.

Although in the following a reactant is being absorbed from a gas phase, the theory can, of course, be used for a situation in which two immiscible liquids exist and a reactant is transported from one of the liquids to react in the other.

According to the models already presented, on the liquid side of the gas – liquid interface there is either an interface covered with small stagnant fluid elements or a stagnant film with a negligible curvature at the interface. Therefore the penetration and film models are the basis of the calculations for these cases.

6.2.3.2. Higbie Penetration Model

A small stagnant liquid element is attached to the interface for a contact time τ. In the period from 0 to τ the reactant is transferred into the liquid; simultaneously it reacts in that liquid. After τ s the element detaches to mix with the bulk of the liquid. The transient concentration profile in the element can be derived from Equation (5.10) with the extension of a reaction term:

$$\frac{\partial C_{\mathrm{A}}}{\partial t} = \mathbb{D}_{\mathrm{i,A}} \frac{\partial^2 C_{\mathrm{A}}}{\partial x^2} - |\mathbb{R}_{\mathrm{A}}| \tag{6.154}$$

to be solved subject to the boundary conditions

$$x = 0 \Rightarrow C_{\mathrm{A}} = C_{\mathrm{A,i}} \tag{6.155}$$

$$x \to \infty \Rightarrow C_{\mathrm{A}} = \overline{C_{\mathrm{A}}} \tag{6.156}$$

where $\overline{C_{\mathrm{A}}}$ is the concentration of A in the bulk of the liquid. The initial condition for Equation (6.154) is

$$t = 0 \Rightarrow C_{\mathrm{A}} = 0 \tag{6.157}$$

If A disappears by a first-order reaction, the above equation becomes

$$\frac{\partial C_{\mathrm{A}}}{\partial t} = \mathbb{D}_{\mathrm{A}} \frac{\partial^2 C_{\mathrm{A}}}{\partial x^2} - k C_{\mathrm{A}} \tag{6.158}$$

If the bulk concentration $\overline{C_{\mathrm{A}}} = 0$, Equation (6.158) gives [38]

$$\begin{aligned} \frac{C_{\mathrm{A}}}{C_{\mathrm{A,i}}} &= \frac{1}{2} \exp\left(-x\sqrt{\frac{k}{\mathbb{D}_{\mathrm{A}}}}\right) \\ &\quad \cdot \left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{\mathbb{D}_{\mathrm{A}} t}} - \sqrt{k t}\right)\right] \\ &+ \frac{1}{2} \exp\left(+x\sqrt{\frac{k}{\mathbb{D}_{\mathrm{A}}}}\right) \\ &\quad \cdot \left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{\mathbb{D}_{\mathrm{A}} t}} - \sqrt{k t}\right)\right] \end{aligned} \tag{6.159}$$

Differentiation and substitution of $x = 0$ give the flux through the interface:

$$\begin{aligned} J_{\mathrm{A}} &= -\mathbb{D}_{\mathrm{A}} \left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)\Big|_{x=0} \\ &= \sqrt{k \mathbb{D}_{\mathrm{A}}} \left[\operatorname{erf}\left(\sqrt{k t}\right) + \frac{\exp(-k t)}{\sqrt{\pi k t}}\right] C_{\mathrm{A,i}} \end{aligned} \tag{6.160}$$

The average mass-transfer rate between 0 and τ s follows from

$$\begin{aligned} \overline{J_{\mathrm{A}}} &= \frac{1}{\tau} \int_0^{\tau} J_{\mathrm{A}}\, \partial t = \sqrt{k \mathbb{D}_{\mathrm{A}}} \left[\left(1 + \frac{1}{2k\tau}\right) \operatorname{erf}\left(\sqrt{k\tau}\right)\right. \\ &\quad \left. + \frac{\exp(-k\tau)}{\sqrt{\pi k \tau}}\right] C_{\mathrm{A,i}} \end{aligned} \tag{6.161}$$

Equation (6.161) gives the mass-transfer rate according to the Higbie model for $\overline{C_{\mathrm{A}}} = 0$. In Figure 33, J_{A} and $\overline{J_{\mathrm{A}}}$ are plotted versus t and τ, respectively. Both fluxes decrease with increasing t as well as increasing τ.

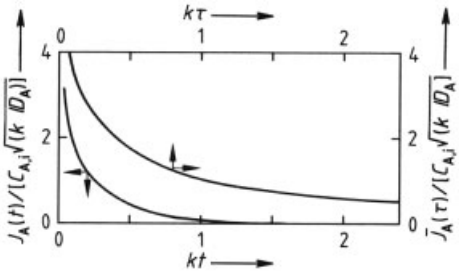

Figure 33. Molar flux $J_A(t)$ as a function of time (t) and average molar flux $\overline{J_A}(\tau)$ as a function of contact time (τ) for mass transfer with simultaneous first-order reaction, according to the Higbie penetration model

By using Equation (5.26), the Higbie contact time τ can be expressed in terms of k_l, the mass-transfer coefficient of A in the liquid phase as

$$\tau = \frac{4\mathbb{D}_A}{\pi k_l^2} \tag{6.162}$$

Substitution of this expression for τ in Equation (6.161) gives

$$\overline{J_A} = k_l C_{A,i} \frac{\sqrt{k\mathbb{D}_A}}{k_l} \left[\left(1 + \frac{\pi k_l^2}{8k\mathbb{D}_A}\right) \cdot \mathrm{erf}\left(\frac{2}{\sqrt{\pi}} \frac{\sqrt{k\mathbb{D}_A}}{k_l}\right) + \frac{1}{2} \frac{k_l}{\sqrt{k\mathbb{D}_A}} \cdot \exp\left(-\frac{4}{\pi} \frac{k\mathbb{D}_A}{k_l^2}\right)\right] \tag{6.163}$$

The dimensionless group $\sqrt{k\mathbb{D}_A/k_l}$ is known as the *Hatta number Ha*:

$$Ha = \frac{\sqrt{k\mathbb{D}_A}}{k_l} \tag{6.164}$$

The Hatta number is a measure of the ratio of the reaction rate to the rate of mass transfer of component A. Substituting Equation (6.164) in Equation (6.163) yields

$$\overline{J_A} = k_l C_{A,i} Ha \left[\left(1 + \frac{\pi}{8Ha^2}\right) \mathrm{erf}\left(\frac{2Ha}{\sqrt{\pi}}\right) + \frac{1}{2Ha} \exp\left(-\frac{4Ha^2}{\pi}\right)\right] \tag{6.165}$$

Asymptotic solutions for $\overline{J_A}$ are obtained for

1) $Ha < 0.2$

$$\overline{J_A} = k_l C_{A,i} \tag{6.166}$$

i.e., Ha is so small, or the reaction is so slow compared to mass transfer, that Equation (6.165) reduces to the equation for mass transfer without reaction

2) $Ha < 0.5$

$$\overline{J_A} = k_l C_{A,i} \left(1 + \frac{4Ha^2}{3\pi}\right) \tag{6.167}$$

3) $Ha > 1$

$$\overline{J_A} = k_l C_{A,i} Ha \left(1 + \frac{\pi}{8Ha^2}\right) \tag{6.168}$$

4) $Ha > 2$

$$\overline{J_A} = k_l C_{A,i} Ha \tag{6.169}$$

In the last three cases (i.e., roughly for $Ha > 0.2$), the transfer rate is enhanced compared to the rate obtained in the absence of reaction due to the reaction that occurs. This is a result of the increase of $\partial C_A/\partial x$ at the interface, because of a decrease in the concentration of A by reaction. A frequently used term is the so-called *enhancement factor* E_A defined as

$$E_A = \frac{\overline{J_A} \text{ with reaction}}{\overline{J_A} \text{ without reaction}} \tag{6.170}$$

where both fluxes $\overline{J_A}$ are based on the same $C_{A,i}$. Thus it follows from the above analysis that the enhancement factor E_A varies, according to the Higbie penetration model, from $E_A = 1$ (for $Ha < 0.2$) to $E_A = Ha$ (for $Ha > 2$), provided the reaction is first-order and provided the bulk concentration of A in the liquid is zero. In general, for this situation the Higbie penetration model yields

$$E_A = Ha \left[\left(1 + \frac{\pi}{8Ha^2}\right) \mathrm{erf}\left(\frac{2Ha}{\sqrt{\pi}}\right) + \frac{1}{2Ha} \exp\left(-\frac{4Ha^2}{\pi}\right)\right] \tag{6.171}$$

6.2.3.3. Surface Renewal Model

Starting from his modification of the penetration theory (i.e., again small stagnant liquid elements are attached to the interface, all elements have the same chance to be replaced by other elements and mix with the bulk of the liquid), DANCKWERTS derived the following expression for mass transfer with a first-order reaction in a liquid [6]

$$\overline{J_A} = k_l \left(C_{A,i} - \frac{\overline{C_A}}{1+Ha^2} \right) \sqrt{1+Ha^2} \tag{6.172}$$

This solution is relatively simple compared to the Higbie solution (Eq. 6.165), which is all the more remarkable because the Danckwerts solution also covers the case in which the concentration of A in the liquid bulk differs from zero. In the case of mass transfer of A to the liquid phase, $\overline{C_A} \leq C_{A,i}$ must hold, so that, roughly, for $Ha > 2$ the term $\overline{C_A}/(1+Ha^2)$ can be neglected in Equation (6.172).

The enhancement factor E_A defined as (see Eq. 6.170)

$$E_A = \frac{\overline{J_A}}{k_l \left(C_{A,i} - \overline{C_A} \right)} \tag{6.173}$$

follows from Equation (6.172)

$$E_A = \sqrt{1+Ha^2} \cdot \left(1 + \frac{\overline{C_A}}{C_{A,i} - \overline{C_A}} \frac{Ha^2}{1+Ha^2} \right) \tag{6.174}$$

For slow reactions ($Ha < 0.2$) the same asymptotic solution is found as with the Higbie model ($E_A = 1$). For fast reactions ($Ha > 2$), generally $\overline{C_A} \approx 0$; hence, the same asymptotic solution is also found as with the Higbie model ($E_A = Ha$). For the regime $0.2 < Ha < 2$ and $\overline{C_A} \neq 0$, E_A depends on the concentrations $C_{A,i}$ and $\overline{C_A}$. In this case, $\overline{J_A}$ is not proportional to the driving force $(C_{A,i} - \overline{C_A})$, as is the case for absorption without reaction. In practice, however, this situation does not occur frequently.

6.2.3.4. Stagnant Film Model

In the stagnant film model the following assumptions are made (see Section 5.3): the fluid is divided in a stagnant film near the interface with a thickness δ and a well-mixed bulk behind it; mass transfer is a stationary process; and the radius of curvature of the interface is large compared to the film thickness. Thus, in the case of homogeneous reaction, Equation (5.20) must be extended by a chemical reaction term

$$\frac{\partial^2 C_A}{\partial x^2} = |\mathbb{R}_A| \quad \text{for } 0 < x < \delta \tag{6.175}$$

The boundary conditions are the same as without reaction:

$$x = 0 \Rightarrow C_A = C_{A,i} \tag{5.21}$$

$$x = \delta \Rightarrow C_A = \overline{C_A} \tag{5.22}$$

The solution of Equations (6.175), (5.21), and (5.22) for first-order kinetics, i.e., $-\mathbb{R}_A- = k\,C_A$, gives, for $0 \leq x \leq \delta$:

$$\frac{C_A}{C_{A,i}} = \frac{\sinh\left[(\delta - x)\sqrt{\frac{k}{\mathbb{D}_A}}\right] + \frac{\overline{C_A}}{C_{A,i}} \sinh\left(x\sqrt{\frac{k}{\mathbb{D}_A}}\right)}{\sinh\left(\delta\sqrt{\frac{k}{\mathbb{D}_A}}\right)} \tag{6.176}$$

Differentiation and substitution of $x = 0$ give the flux through the interface:

$$\overline{J_A} = -\mathbb{D}_A \left(\frac{\partial C_A}{\partial x} \right)\Big|_{x=0} = \frac{\mathbb{D}_A}{\delta} C_{A,i} \cdot \delta \sqrt{\frac{k}{\mathbb{D}_A}} \frac{1}{\tanh\left(\delta\sqrt{\frac{k}{\mathbb{D}_A}}\right)} - \frac{\frac{\overline{C_A}}{C_{A,i}}}{\sinh\left(\delta\sqrt{\frac{k}{\mathbb{D}_A}}\right)} \tag{6.177}$$

From Equation (5.27) the mass-transfer coefficient according to the stagnant film model is found as

$$k_l = \frac{\mathbb{D}_A}{\delta} \tag{6.178}$$

The Hatta number thus becomes

$$Ha = \frac{\sqrt{k\mathbb{D}_A}}{k_l} = \delta\sqrt{\frac{k}{\mathbb{D}_A}} \tag{6.179}$$

Substitution of Equations (6.178) and (6.179) in Equation (6.177) gives

$$\overline{J_A} = k_l C_{A,i} \frac{Ha}{\tanh(Ha)} \left[1 - \frac{\overline{C_A}}{C_{A,i}} \frac{1}{\cosh(Ha)} \right] \tag{6.180}$$

Hence the enhancement factor E_A according to the stagnant film model becomes

$$E_A = \frac{Ha}{\tanh(Ha)} \left\{ 1 + \frac{\overline{C_A}}{C_{A,i} - \overline{C_A}} \left[1 - \frac{1}{\cosh(Ha)} \right] \right\} \tag{6.181}$$

Again, the same asymptotic solutions are found as with the penetration models ($E_A = 1$ and $E_A = Ha$ for $Ha < 0.2$ and $Ha > 2$, respectively). For the regime $0.2 < Ha < 2$ and $\overline{C_A} \neq 0$, E_A again depends on the concentration, in which case $\overline{J_A}$ is not proportional to the driving force $(C_{A,i} - \overline{C_A})$, as is the case for absorption without reaction. As already discussed, in practice this situation is seldom met.

6.2.3.5. Comparison of Models for First-Order Kinetics

In Sections 6.2.3.2, 6.2.3.3, and 6.2.3.4, analytical expressions have been presented for simultaneous mass transfer and irreversible first-order reactions according to the penetration models of Higbie and Danckwerts and the film model. Numerically, the differences between the three models are not very pronounced. All three models predict the same asymptotic solutions for $Ha < 0.2$ and $Ha > 2$, respectively. The differences are maximal for Hatta numbers roughly equal to unity. Even in this case the enhancement factors calculated from the three models agree within 10 % for $\overline{C_A} = 0$ and within 20 % for $\overline{C_A} \neq 0$. For example, for $Ha = 1$ and $\overline{C_A} = 0$ from Equations (6.171), (6.174), and (6.181),

Stagnant film	$E_A = 1.31$
Higbie penetration	$E_A = 1.38$
Danckwerts penetration	$E_A = 1.41$

For $Ha = 1$ and $\overline{C_A} = C_{A,i}/2$,

Stagnant film	$E_A = 1.78$
Danckwerts penetration	$E_A = 2.12$

However, as mentioned, for fast reactions the bulk concentration $\overline{C_A}$ can be assumed to be zero for almost all practical cases, so that the situation $Ha = 1$ and $\overline{C_A} = C_{A,i}/2$ is not often encountered.

The agreement between the enhancement factors is very satisfactory for most engineering purposes. Therefore, the distinction between the different models is generally not relevant. The Hatta number is usually calculated from the (measured) mass-transfer coefficient k_l (Eq. 6.164):

$$Ha = \frac{\sqrt{k\mathbb{D}_A}}{k_l}$$

Then the enhancement factor is estimated from one of the following equations:

$$E_A = \sqrt{1 + Ha^2} \tag{6.182}$$

or

$$E_A = \frac{Ha}{\tanh(Ha)} \tag{6.183}$$

These equations are generally selected because they are easy to use.

In the *fast reaction regime* ($Ha > 2$), all models predict

$$\overline{J_A} = k_l C_{A,i} Ha = \sqrt{k\mathbb{D}_A} C_{A,i} \tag{6.184}$$

Thus, for $Ha > 2$, the rate of mass transfer becomes independent of the mass-transfer coefficient. This result is far reaching; it implies that the overall conversion rate does not depend on local hydrodynamics near the gas – liquid interface. Furthermore, the bulk concentration $\overline{C_A}$ becomes zero, which implies that information on the exact flow pattern in the bulk of the liquid phase (e.g., macromixing) is not required for design purposes. Both facts may substantially simplify (experiments necessary for) the design of the reactor.

In the *slow reaction regime* ($Ha < 0.2$), all models predict

$$\overline{J_A} = k_l \left(C_{A,i} - \overline{C_A} \right) \tag{6.185}$$

Thus, the reaction does not enhance the rate of mass transfer in this case.

6.2.3.6. Comparison of Hatta Numbers with Thiele Moduli

In this section the film model is used for discussions of gas – liquid reactions because of its simplicity and mediative power. With the penetration models, however, the same conclusions would have been obtained.

The Hatta numbers and the Thiele moduli discussed in the previous section are very similar because, for both high Hatta numbers and

high Thiele moduli the penetration depth of A in the stagnant film or in the porous pellet is so small, that the thickness of the film with respect to the geometry of the pellet becomes irrelevant. For Hatta numbers this can be derived from Equation (6.184), where it is shown that for high Hatta numbers and first-order kinetics, the overall mole flux is not a function of k_1 (i.e., film thickness δ):

$$\overline{J_{\mathrm{A}}} = \sqrt{k\mathbb{D}_{\mathrm{A}}}\, C_{\mathrm{A,i}} \tag{6.186}$$

For porous pellets with high Thiele moduli and first-order kinetics, the effectiveness factor η becomes

$$\eta = \frac{1}{\varphi} = \frac{A_{\mathrm{p}}}{V_{\mathrm{p}}}\sqrt{\frac{\mathbb{D}_{\mathrm{i,A}}}{k}} \tag{6.187}$$

The number of moles of A converted within the porous pellet equals $V_{\mathrm{p}} — \mathbb{R}_{\mathrm{A}} —_{\mathrm{p}} \eta$; hence the overall flux through the pellet outer surface is

$$\overline{J_{\mathrm{A}}} = \frac{V_{\mathrm{p}}\,|\mathbb{R}_{\mathrm{A}}|_{\mathrm{p}}}{A_{\mathrm{p}}}\eta$$
$$= \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}}\, k_{\mathrm{p}} C_{\mathrm{A,i}}\, \frac{A_{\mathrm{p}}}{V_{\mathrm{p}}}\sqrt{\frac{\mathbb{D}_{\mathrm{i,A}}}{k}} = \sqrt{k_{\mathrm{p}}\mathbb{D}_{\mathrm{i,A}}}\, C_{\mathrm{A,i}} \tag{6.188}$$

which is similar to Equation (6.186).

In general, the Thiele modulus and Hatta number are identical: they differ only in the characteristic dismension. For the Thiele modulus this is the generalized pellet dimension $V_{\mathrm{p}}/A_{\mathrm{p}}$ introduced by ARIS, whereas for the Hatta number it is the characteristic dimension of the film $(V/A)_{\mathrm{film}}$ (i.e., the thickness δ, or alternatively written $\mathbb{D}_{\mathrm{A}}/k_1$). Thus,

$$Ha = \left(\frac{\mathbb{D}_{\mathrm{A}}}{k_1}\frac{A_{\mathrm{p}}}{V_{\mathrm{p}}}\right)\varphi \tag{6.189}$$

Sometimes a distinction between Hatta numbers and Thiele moduli is made, and sometimes reference is made to a single parameter (e.g., φ) that covers both numbers (e.g., see [8]).

For reactions with mass transfer in porous solids, the terms *effectiveness factor* and *retardation of reaction by mass transfer* are used, whereas for reactions with mass transfer of gases in liquids, *enhancement factor* and *enhancement of mass transfer by reaction* are generally employed. This classification is logical because for reactions of gases in porous solids, it is difficult to visualize mass transfer without chemical reaction: since there is no such thing as bulk behind a film, for porous pellets, mass transfer simply becomes nonexistent in the absence of chemical reaction. Consequently, an enhancement factor definition (Eq. 6.170) is not applicable, and the use of an effectiveness factor is more logical. In contrast, for gas – liquid reactions there is a large liquid bulk behind the film. Therefore the idea of mass transfer without reaction makes sense and so does the definition of enhancement factor.

6.2.3.7. Hatta Numbers for Arbitrary Kinetics

From the similarity between Hatta numbers and Thiele moduli, it is clear that generalization of Hatta numbers for arbitrary kinetics proceeds along the same as lines discussed in Section 6.2.2 for Thiele moduli. Thus, for monomolecular reactions and arbitrary kinetics,

$$Ha = \frac{\mathbb{D}_{\mathrm{A}}}{k_1}\,|\mathbb{R}_{\mathrm{A}}\left(C_{\mathrm{A,i}}\right)| \cdot \left(2\mathbb{D}_{\mathrm{A}}\int_0^{C_{\mathrm{A,i}}} |\mathbb{R}_{\mathrm{A}}\left(C_{\mathrm{A}}\right)|\,\partial C_{\mathrm{A}}\right)^{-\frac{1}{2}} \tag{6.190}$$

For simple *n* th-*order kinetics*,

$$|\mathbb{R}_{\mathrm{A}}| = k\,C_{\mathrm{A}}^{n} \tag{6.191}$$

Equation (6.190) yields

$$Ha = \frac{\mathbb{D}_{\mathrm{A}}}{k_1}\sqrt{\frac{n+1}{2}\,\frac{k\,C_{\mathrm{A,i}}^{n-1}}{\mathbb{D}_{\mathrm{A}}}} \tag{6.192}$$

For *bimolecular reactions*,

$$\mathrm{A} + \nu B \longrightarrow \text{Products}$$

where both A and B are transferred from the gas to react in the liquid; $— \mathbb{R}\left(C_{\mathrm{A}}, C_{\mathrm{B}}\right) —$ can again be written as a function only of C_{A}

$$|\mathbb{R}\left(C_{\mathrm{A}}, C_{\mathrm{B}}\right)| = |\mathbb{R}\left\{C_{\mathrm{A}}, C_{\mathrm{B,i}}\left[1-\gamma\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right]\right\}| \tag{6.193}$$

where γ is again the excess factor, defined as

$$\gamma = \nu\frac{\mathbb{D}_{\mathrm{A}}}{\mathbb{D}_{\mathrm{B}}}\frac{C_{\mathrm{A,i}}}{C_{\mathrm{B,i}}} \tag{6.194}$$

Thus the generalized Hatta number for bimolecular reactions is obtained:

$$Ha = \frac{\mathbb{D}_A}{k_l} \left| \mathbb{R}_A \left(C_{A,i}, C_{B,i} \right) \right| \cdot \left(2\,\mathbb{D}_A \int_0^{C_{A,i}} \left| \mathbb{R}_A \left\{ C_A, C_{B,i} \cdot \left[1 - \gamma \left(1 - C_A / C_{A,i} \right) \right] \right\} \right| \partial C_A \right)^{-\frac{1}{2}} \tag{6.195}$$

which, for example, for (1,1) kinetics

$$\left| \mathbb{R} \left(C_A, C_B \right) \right|_p = k_p C_A C_B \tag{6.196}$$

yields

$$Ha = \frac{\mathbb{D}_A}{k_l} \sqrt{\frac{k C_{B,i}}{\left(1 - \frac{\gamma}{3}\right) \mathbb{D}_A}} \tag{6.197}$$

Equations (6.195) and (6.197) can be used only if both A and B are transferred from the gas and not if one of the components (e.g., B) is already dissolved in the liquid. All the above formulas can be compared readily with similar equations for the Thiele modulus discussed in Section 6.2.2.

6.2.3.8. Case Study: Ideally Mixed Gas – Liquid Reactors

As a case study, the discussed theory is applied to ideally mixed gas liquid reactors [i.e., reactors for which the bulk concentration in the liquid (reaction) phase and the interface concentration are constant over the entire reactor]. Thus, a short excursion is made from macrokinetics to the reactor level.

The case study is restricted to first-order kinetics and only the film model is used (see Fig. 34). Two important parameters for the type of reactor shown in Figure 34 are the *gas – liquid dispersion volume* V (m^3) and the *gas – liquid specific contact area* a (square meters of interface per cubic meter of gas – liquid dispersion). The value of a is related to the gas volume fraction or void fraction ε and the *Sauter mean bubble diameter* d_S of the gas. If the gas phase consists of N spherical bubbles of different diameters d, then

$$a = \frac{N \pi \langle d^2 \rangle}{V} \tag{6.198}$$

where $\langle d^2 \rangle$ is the average of the square of the diameters. Similarly, the volume of the gas is

$$\varepsilon V = N \frac{1}{6} \pi \langle d^3 \rangle \tag{6.199}$$

Consequently, the specific contact area a equals

$$a = \frac{6\varepsilon \langle d^2 \rangle}{\langle d^3 \rangle} = \frac{6\varepsilon}{d_S} \tag{6.200}$$

Hence, the Sauter mean diameter d_S of the bubbles follows from

$$d_S = \frac{\sum d^3}{\sum d^2} \tag{6.201}$$

Equation (6.200) shows that a is inversely proportional to the average bubble size of the dispersed gas. This size depends not only on the properties of gas and liquid but also, to a large extent, on the interplay between the mechanisms responsible for the breaking up and mutual coalescence of the bubbles – hence, on reactor hydrodynamics.

Fast Reactions. In the case of a fast reaction in the above reactor ($Ha > 2$), the concentration of A in the liquid bulk $\overline{C_A}$ is zero. Thus, the overall conversion rate — R_A — can be calculated simply by multiplying $\overline{J_A}$ in Equation (6.186) by the total contact area aV between gas and liquid in the reactor:

$$|R_A| V = \sqrt{k \mathbb{D}_A}\, C_{A,i}\, a V \tag{6.202}$$

Slow Reactions. For $Ha < 2$, $\overline{C_A} \neq 0$, so at least part of the reaction takes place in the bulk of the liquid phase. For first-order kinetics a mass balance for A over the *bulk of the liquid phase* gives

In by diffusion through the plane ($x = \delta$) + in by flow – out by flow = consumed by reaction

$$-\mathbb{D}_A \left(\frac{\partial C_A}{\partial x} \right) \Big|_{x=\delta} a V + \varphi_{V0} C_{A0} - \varphi_{V1} \overline{C_A} - k \overline{C_A} (1 - \varepsilon) V = 0 \tag{6.203}$$

where φ_{V0} is the inlet liquid volumetric flow rate (m^3/s), φ_{V1} the outlet liquid volumetric flow rate (m^3/s), and C_{A0} the inlet concentration of

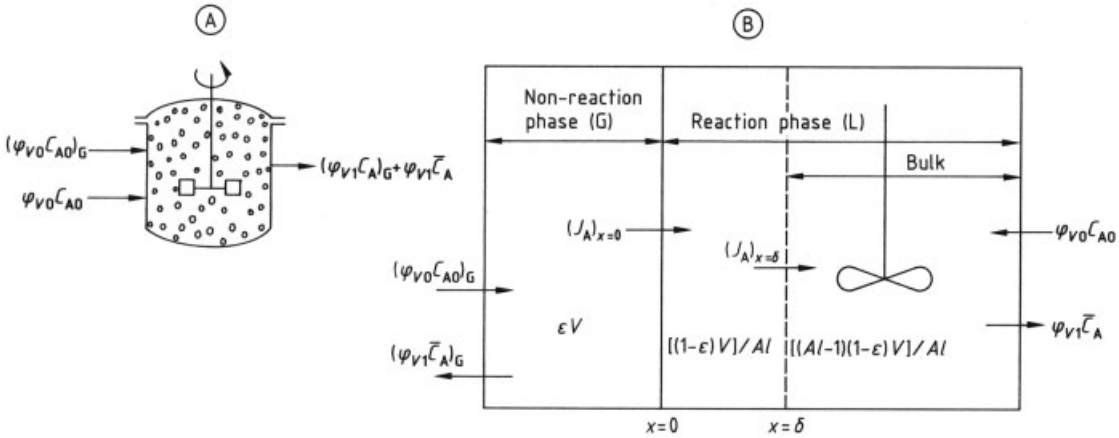

Figure 34. A) Continuous gas – liquid stirred-tank reactor or bubble column; B) Schematic of mass transfer and chemical reaction in such a reactor

reactant A in the liquid (mol/m^3). Because the kinetics are first order, the stagnant film model yields for the concentration gradient at the plane $x=\delta$ (see Eq. 6.176)

$$\left(\frac{\partial C_A}{\partial x}\right)\Big|_{x=\delta} = -C_{A,i}\sqrt{\frac{k}{\mathbb{D}_A}} \cdot \frac{1}{\sinh\left(\delta\sqrt{\frac{k}{\mathbb{D}_A}}\right)} - \frac{\frac{\overline{C_A}}{C_{A,i}}}{\tanh\left(\delta\sqrt{\frac{k}{\mathbb{D}_A}}\right)} \tag{6.204}$$

or by introducing the Hatta number from Equation (6.179)

$$Ha = \frac{\sqrt{k\mathbb{D}_A}}{k_1} = \delta\sqrt{\frac{k}{\mathbb{D}_A}}$$

Equation (6.204) becomes

$$\left(\frac{\partial C_A}{\partial x}\right)\Big|_{x=\delta} = -\sqrt{\frac{k}{\mathbb{D}_A}}\left[\frac{C_{A,i}}{\sinh(Ha)} - \frac{\overline{C_A}}{\tanh(Ha)}\right] \tag{6.205}$$

For convenience the so-called *Hinterland ratio* Al is defined as the ratio between the total liquid phase volume and the volume of the stagnant liquid film:

$$Al = \frac{1-\varepsilon}{a\delta} = \frac{(1-\varepsilon)k_1}{a\mathbb{D}_A} = \frac{1-\varepsilon}{6\varepsilon}\frac{d_S k_1}{\mathbb{D}_A} \tag{6.206}$$

Combination of Equations (6.203), (6.205), and (6.206) yields for $\overline{C_A}$

$$\frac{\overline{C_A}}{C_{A,i}} = \frac{1+\frac{\varphi_{V0}C_{A0}}{\varphi_{V1}C_{A,i}}\frac{1}{k\tau_L}Al\cdot Ha\sinh(Ha)}{\left(\frac{1}{k\tau_L}+\frac{Al-1}{Al}\right)Al\cdot Ha\sinh(Ha)+\cosh(Ha)} \tag{6.207}$$

where τ_L is the residence time of the liquid phase given by $\tau_L = (1-\varepsilon)V/\varphi_{V1}$. Substituting the expression obtained in Equation (6.207) for $\overline{C_A}$ in Equation (6.180) (which gives $\overline{J_A}$ as a function of $\overline{C_A}$) yields

$$\overline{J_A} = k_1 C_{A,i}\cdot Ha\left(\frac{1}{\tanh(Ha)} - \frac{\frac{1}{\sinh(Ha)}+\frac{\varphi_{V0}C_{A0}}{\varphi_{V1}C_{A,i}}\frac{1}{k\tau_L}Al\cdot Ha}{\left(\frac{1}{k\tau_L}+\frac{Al-1}{Al}\right)Al\cdot Ha\sinh(Ha)+\cosh(Ha)}\right) \tag{6.208}$$

Part of A that reaches the bulk of the liquid is converted there (part 1). The remainder leaves the reactor unconverted (part 2). The ratio of (1) over (2) is

$$\frac{k\overline{C_A}(1-\varepsilon)V}{\varphi_{V1}\overline{C_A}} = k\tau_L\left(\frac{Al-1}{Al}\right) \tag{6.209}$$

For an efficient reactor, this ratio must be high; otherwise the reactor woud act more like an absorber or extractor. Therefore, the following assumes that

$$\frac{Al-1}{Al} >> k\tau_L \tag{6.210}$$

Often the amount of A transferred from the gas into the liquid is much higher than the amount

of A introduced into the reactor via the feed of the liquid phase, which means that

$$k_1 C_{A,i} aV >> \varphi_{V0} C_{A0} \tag{6.211}$$

This condition is equivalent to

$$\frac{1}{\sinh(Ha)} >> \frac{\varphi_{V0} C_{A0}}{\varphi_{V1} C_{A,i}} \frac{1}{k\tau_L} Al \cdot Ha \tag{6.212}$$

Substitution of Equations (6.210) and (6.212) in Equations (6.207) and (6.208) gives

$$\frac{\overline{C_A}}{C_{A,i}} = \frac{1}{(Al-1)\, Ha \sinh(Ha) + \cosh(Ha)} \tag{6.213}$$

and

$$\overline{J_A} = \sqrt{k\mathbb{D}_A}\, C_{A,i} \frac{(Al-1)\, Ha + \tanh(Ha)}{(Al-1)\, Ha \tanh(Ha) + 1} \tag{6.214}$$

The above equations hold only if the liquid phase that is fed to or drawn from the reactor contains minor amounts of A compared to the amount of A transferred through the interface.

The enhancement factor E_A (see also Eq. 6.173) can be calculated by combining Equations (6.213) and (6.214):

$$\begin{aligned} E_A &= \frac{\overline{J_A}}{k_l \left(C_{A,i} - \overline{C_A}\right)} \\ &= Ha \frac{(Al-1)\, Ha \cosh(Ha) + \sinh(Ha)}{(Al-1)\, Ha \sinh(Ha) + \cosh(Ha) - 1} \end{aligned} \tag{6.215}$$

For convenience a *degree of reactor utilization* $\eta^{\ominus}$ is defined as the ratio of the conversion rate of A and the highest possible conversion rate that would occur if the entire liquid phase were in equilibrium with the interface (i.e., if $\overline{C_A} = C_{A,i}$). Thus,

$$\eta^{\ominus} = \frac{\overline{J_A}\, aV}{k C_{A,i} (1-\varepsilon) V} \tag{6.216}$$

Note that the degree of reactor utilization is not an effectiveness factor. Substituting Equations (6.213) and (6.214) in (6.216) gives

$$\eta^{\ominus} = \frac{1}{Al \cdot Ha} \frac{(Al-1)\, Ha + \tanh(Ha)}{(Al-1)\, Ha \tanh(Ha) + 1} \tag{6.217}$$

The asymptotic solutions are discussed below.

Slow Reactions. For slow reactions ($Ha < 0.2$) the derived equations simplify to

$$\begin{aligned} \frac{\overline{C_A}}{C_{A,i}} &= \frac{1}{1 + Al \cdot Ha^2} \\ \overline{J_A} &= k_l C_{A,i} \frac{Al \cdot Ha^2}{1 + Al \cdot Ha^2} \\ E_A &= 1 \\ \eta^{\ominus} &= \frac{1}{1 + Al \cdot Ha^2} \end{aligned} \tag{6.218}$$

Mass transfer is not enhanced. Depending on the value of $Al \cdot Ha^2$, the bulk either is saturated with A ($Al \cdot Ha^2 << 1$) or contains A only in relatively low concentration ($Al \cdot Ha^2 >> 1$). The reactor capacity, expressed as moles of A converted per cubic meter of reactor volume and unit time, is given by $J_A\, a$, which, in this case, is equal to $\eta^{\ominus} k C_{A,i} (1-\varepsilon)$ when $Al >> 1$ or the film volume is negligible compared to the bulk volume.

Fast Reactions. For $Ha > 2$, Equations (6.213), (6.215), and (6.217) simplify to

$$\begin{aligned} \frac{\overline{C_A}}{C_{A,i}} &= 0 \\ \overline{J_A} &= \sqrt{k\mathbb{D}_A}\, C_{A,i} \\ E_A &= Ha \\ \eta^{\ominus} &= \frac{1}{Al \cdot Ha} \end{aligned} \tag{6.219}$$

The reactor capacity now becomes independent of the mass-transfer coefficient k_l and of the bulk volume of the liquid phase $(1-\varepsilon)\, V$ because of the absence of reactant in the liquid bulk.

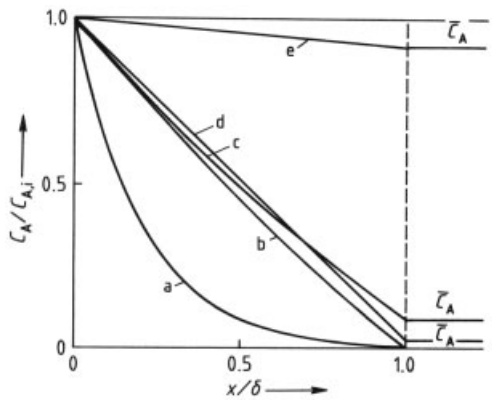

Figure 35. Concentration profiles in hypothetical mass-transfer film (film model) in a bubble column for several combinations of Ha and $Al \cdot Ha^2$
a) $Ha = 5$, any $Al \cdot Ha^2$; b) $Ha = 1$, $Al \cdot Ha^2 = 1000$; c) $Ha = 1$, $Al \cdot Ha^2 = 10$; d) $Ha = 0.1$, $Al \cdot Ha^2 = 1000$; e) $Ha = 0.1$, $Al \cdot Ha^2 = 0.1$

Figure 35 shows concentration profiles for several combinations of Ha and $Al \cdot Ha^2$ to further illustrate the different regimes.

In Figure 36 the degree of reactor utilization $\eta^{\ominus}$ is plotted versus the Hinterland ratio Al for some values of Ha. For a slow reaction ($Ha < 0.2$), the degree of utilization is approximately equal to $1/(1+Al \cdot Ha^2)$ or $\overline{C_A}/C_{A,i}$, as illustrated above. Hence, the degree of utilization approaches unity for $Al \cdot Ha^2 << 1$ and $Ha < 0.2$.

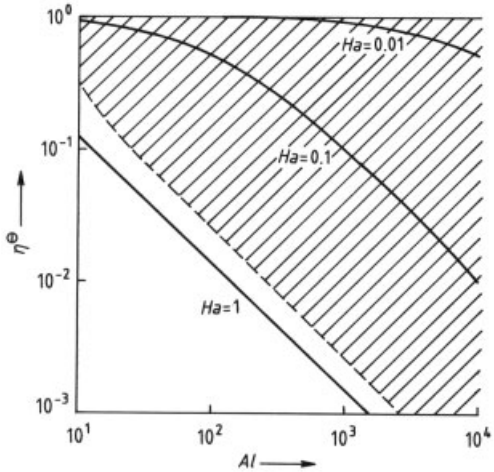

Figure 36. Degree of reactor utilization for simultaneous mass transfer and chemical reaction in a bubble column. In the shaded area more than 80 % of the conversion takes place in the liquid phase.

For a given Hatta number and below a certain value of the Hinterland ratio, say $Al^{\ominus}$, the reactor utilization is 100 % and above that value, $\eta^{\ominus}$ drops below unity. This value $Al^{\ominus}$ decreases with increasing Hatta number. In other words, the faster the reaction is, the larger must the value of the specific contact area be to ensure efficient use of the bulk of the liquid phase.

The Hinterland ratio strongly depends on the (*stirrer*) *specific power input* (i.e., power input per cubic meter of dispersion volume) [39]. The cost of intensive stirring usually exceeds the cost of a larger reactor by far. Therefore, a larger reactor with poor utilization of the liquid bulk is often a cheaper option than a smaller unit that consumes much more energy by intensive stirring. Often the reactor capacity can be increased much more efficiently by an increase in the reaction temperature or reactor pressure, which results in an increase of Ha and $C_{A,i}$ respectively. In that case the influence of temperature on the solubility of A must also be considered.

The above discussions can be summarized for a constant Hatta number as follows. If the degree of reactor utilization approaches 1 ($Ha < 0.2$ and $Al \cdot Ha^2 << 1$), the reactor capacity depends on the volume of the liquid phase and is independent of the value of the specific contact area. If the degree of utilization is much less than 1 ($Ha > 2$ or $Al \cdot Ha^2 >> 1$), the reactor capacity is proportional to the interfacial area and independent of the volume of the liquid phase.

6.2.3.9. Selectivities of Multiple Reactions

In this section the selectivity of *parallel reactions* and *consecutive reactions* is discussed for simultaneous mass transfer and reaction. Although not discussed here, the theory is readily applicable for reactions in porous solids, but the Hatta number must be replaced by the Thiele modulus. The treatment is limited to monomolecular reactions.

Parallel Reactions. Consider the following parallel reaction:

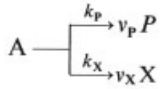

P is again the desired product; X an undesirable byproduct.

For *low Hatta numbers* ($Ha < 0.2$) and a sufficiently large mass-transfer coefficient, so that $\overline{C_A} = C_{A,i}$, selectivity follows from

$$\sigma_{P,Ha<0.2} = \frac{|\mathbb{R}_A\,(C_{A,i})\,|_P}{|\mathbb{R}_A\,(C_{A,i})\,|_P + |\mathbb{R}_A\,(C_{A,i})\,|_X} \tag{6.220}$$

For *high Hatta numbers* ($Ha > 2$) a simple general solution can also be found. In this regime the bulk concentration of A $\overline{C_A}$ equals zero, as discussed above. The penetration depth of A in the liquid δ_p is smaller than the thickness of the stagnant film δ (see Fig. 37). The concentration profile of A can be approximated by a straight line, which is given by

$$C_A = C_{A,i}\,\frac{\delta_p - x}{\delta_p} \tag{6.221}$$

The amount of A that is converted to P in the film, $-R_A-_P$, is determined from

$$|R_A|_P = a\int_0^{\delta_P} |\mathbb{R}_A(C_A)|_P\,\partial x \tag{6.222}$$

Substituting Equation (6.221) gives

$$|R_A|_P = a\frac{\delta_P}{C_{A,i}}\int_0^{C_{A,i}} |\mathbb{R}_A(C_A)|_P\,\partial C_A \tag{6.223}$$

Similarly the amount of A converted in X, $-R_A-_X$, is

$$|R_A|_X = a\frac{\delta_P}{C_{A,i}}\int_0^{C_{A,i}} |\mathbb{R}_A(C_A)|_X\,\partial C_A \tag{6.224}$$

The selectivity $\sigma_P = -R_A-_P/(-R_A-_P + -R_A-_X)$ for high Hatta numbers thus becomes

$$\sigma_{P,Ha>2} = \frac{\int_0^{C_{A,i}} |\mathbb{R}_A(C_A)|_P\,\partial C_A}{\int_0^{C_{A,i}} |\mathbb{R}_A(C_A)|_P\,\partial C_A + \int_0^{C_{A,i}} |\mathbb{R}_A(C_A)|_X\,\partial C_A} \tag{6.225}$$

This equation holds for any arbitrary kinetics and $Ha > 2$.

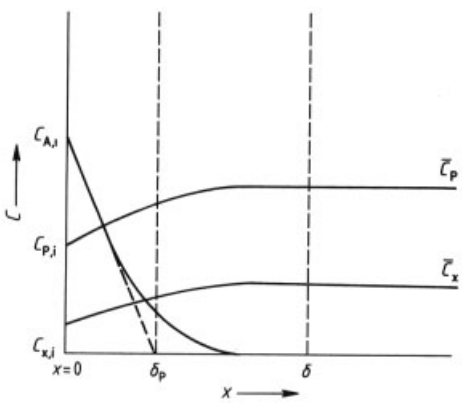

Figure 37. Concentration profiles (film model) for mass transfer with simultaneous parallel reactions

Consider, for example, the case of n-th- and m-th-order kinetics:

$$|\mathbb{R}_A(C_A)|_P = k_P C_A^n \tag{6.226}$$

$$|\mathbb{R}_A(C_A)|_X = k_X C_A^m \tag{6.227}$$

Equation (6.220) yields for $Ha < 0.2$ and $C_A = C_{A,i}$

$$\sigma_{P,Ha<0.2} = \left(1 + \frac{k_X}{k_P} C_{A,i}^{m-n}\right)^{-1} \tag{6.228}$$

whereas for $Ha > 2$, Equation (6.225) gives

$$\sigma_{P,Ha>2} = \left(1 + \frac{n+1}{m+1}\frac{k_X}{k_P} C_{A,i}^{m-n}\right)^{-1} \tag{6.229}$$

Hence, for $n > m$, $\sigma_{P,\,Ha<0.2} > \sigma_{P,\,Ha>2}$; for $n < m$, $\sigma_{P,\,Ha<0.2} < \sigma_{P,\,Ha>2}$. In other words, if the reaction toward the desired product is of a higher order, mass-transfer limitation decreases the selectivity. If, on the other hand, it is of a lower order, mass-transfer limitation increases the selectivity. Expressions can be derived similarly for other types of kinetics.

Equations (6.220) and (6.225) do not yield a value for selectivity in the intermediate region $0.2 < Ha < 2$. As an approximation for this regime, the authors of this article suggest

$$\sigma_P = \frac{\sigma_{P,\,Ha<0.2} + Ha\,\sigma_{P,\,Ha>2}}{Ha}\cdot\tanh(Ha) \tag{6.230}$$

This approximation can be used only if $\sigma_{P,\,Ha<0.2}$ and $\sigma_{P,\,Ha>2}$ do not differ too much, which usually holds for most practical cases.

Consecutive Reactions. Consider the following consecutive reaction:

$$A \xrightarrow{k_P} \nu_P P \xrightarrow{k_X} \nu_X X$$

Again, P is the desired and X the undesired product. If both reactions are assumed to be first order, the transport equations of A and P become

$$\mathbb{D}_A \frac{\partial^2 C_A}{\partial x^2} = k_P C_A \tag{6.231}$$

$$\mathbb{D}_P \frac{\partial^2 C_P}{\partial x^2} = k_X C_P - \nu_P k_P C_A \tag{6.232}$$

Here $k_P C_A$ is expressed in moles of A per cubic meter per second and $k_X C_P$ is expressed in moles of P per cubic meter per second. Both equations must be solved subject to the boundary conditions

$$\sigma_{\mathrm{P}}=\frac{1}{\beta_1}$$

$$\cdot\left\{1-\frac{\left(Al-1+\frac{\tanh(Ha_{\mathrm{P}})}{Ha_{\mathrm{P}}}\right)\left(\beta_2+\frac{\beta_3+\beta_4}{E_{\mathrm{A0}}}\right)}{\beta_2\left(Al-1+\frac{\tanh(Ha_{\mathrm{P}})}{Ha_{\mathrm{P}}}\right)+\beta_3\left((Al-1)\frac{\tanh(Ha_{\mathrm{P}})}{Ha_{\mathrm{P}}}+\frac{1}{Ha_{\mathrm{P}}^2}\right)}\right\}$$

$x=0\Rightarrow$

$$\mathbb{D}_{\mathrm{A}}\left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)=k_{\mathrm{g,A}}\left(\overline{C_{\mathrm{A,g}}}-\frac{C_{\mathrm{A,i}}}{m_{\mathrm{A}}}\right)$$

$$\mathbb{D}_{\mathrm{P}}\left(\frac{\partial C_{\mathrm{P}}}{\partial x}\right)=k_{\mathrm{g,P}}\left(\overline{C_{\mathrm{P,g}}}-\frac{C_{\mathrm{P,i}}}{m_{\mathrm{P}}}\right) \tag{6.233}$$

$x=\delta\Rightarrow$

$$a\mathbb{D}_{\mathrm{A}}\left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)=-(1-\varepsilon)\frac{Al-1}{Al}k_{\mathrm{P}}\overline{C_{\mathrm{A,l}}}$$

$$a\mathbb{D}_{\mathrm{P}}\left(\frac{\partial C_{\mathrm{P}}}{\partial x}\right)=-(1-\varepsilon)\frac{Al-1}{Al}$$

$$\cdot\left(k_{\mathrm{X}}\overline{C_{\mathrm{P,l}}}-\nu_{\mathrm{P}}k_{\mathrm{P}}\overline{C_{\mathrm{A,l}}}\right) \tag{6.234}$$

Furthermore, the mass-transfer coefficients in the liquid phase are related to the film thickness δ as follows

$$k_{\mathrm{l,A}}=\frac{\mathbb{D}_{\mathrm{A}}}{\delta} \tag{6.235}$$

$$k_{\mathrm{l,P}}=\frac{\mathbb{D}_{\mathrm{P}}}{\delta} \tag{6.236}$$

The molar flux of A is obtained by solving the above set of equations [40]:

$$\overline{J_{\mathrm{A}}}=\sqrt{\mathrm{k}\,\mathbb{D}_{\mathrm{A}}}\overline{C_{\mathrm{A,g}}}\left(\frac{1}{\frac{1}{m_{\mathrm{A}}}+E_{\mathrm{A0}}\frac{k_{\mathrm{l,A}}}{k_{\mathrm{g,A}}}}\right) \tag{6.237}$$

where E_{A0} is a modified enhancement factor, defined as

$$E_{\mathrm{A0}}=\frac{\overline{J_{\mathrm{A}}}}{k_{\mathrm{l,A}}C_{\mathrm{A,i}}}=Ha\frac{(Al-1)\,Ha+\tanh(Ha)}{(Al-1)\,Ha\tanh(Ha)+1} \tag{6.238}$$

and

$$Ha=\frac{\sqrt{k_{\mathrm{P}}\mathbb{D}_{\mathrm{A}}}}{k_{\mathrm{l,A}}} \tag{6.239}$$

Compare with Equations (6.164) and (6.214). The selectivity σ_{P} is

$$\beta_1=1-\frac{\mathbb{D}_{\mathrm{A}}}{\mathbb{D}_{\mathrm{P}}}\frac{k_{\mathrm{X}}}{k_{\mathrm{P}}}$$

$$\beta_2=\nu_{\mathrm{P}}m_{\mathrm{P}}\frac{k_{\mathrm{l,P}}}{k_{\mathrm{g,P}}}\frac{k_{\mathrm{P}}}{k_{\mathrm{X}}} \tag{6.240}$$

$$\beta_3=\nu_{\mathrm{P}}\frac{k_{\mathrm{P}}}{k_{\mathrm{X}}}$$

$$\beta_4=\frac{m_{\mathrm{P}}\overline{C_{\mathrm{P,g}}}}{m_{\mathrm{A}}\overline{C_{\mathrm{A,g}}}}\left(\frac{\mathbb{D}_{\mathrm{P}}}{\mathbb{D}_{\mathrm{A}}}\frac{k_{\mathrm{P}}}{k_{\mathrm{X}}}-1\right)$$

Ha_{P} is a modified Hatta number for the consecutive reaction

$$Ha=\frac{\sqrt{k_{\mathrm{X}}\mathbb{D}_{\mathrm{P}}}}{k_{\mathrm{l,P}}} \tag{6.241}$$

Equation (6.240) is simplified considerably if the mass-transfer resistance on the gas-phase side is negligible (i.e., $k_{\mathrm{g,P}}\to\infty$ and $\beta_2\to 0$) and if $\overline{C_{\mathrm{P,g}}}=0$ (thus $\beta_4\to 0$). Then,

$$\sigma_{\mathrm{P}}=\frac{Ha^2}{Ha^2-Ha_{\mathrm{P}}^2}\cdot\left(1-\frac{Ha_{\mathrm{P}}(Al-1)\,Ha\tanh(Ha)+1}{Ha\,(Al-1)\,Ha_{\mathrm{P}}\tanh(Ha_{\mathrm{P}})+1}\cdot\frac{(Al-1)\,Ha_{\mathrm{P}}+\tanh(Ha_{\mathrm{P}})}{(Al-1)\,Ha+\tanh(Ha)}\right) \tag{6.242}$$

Two limiting regimes are mentioned here:

1) In the *slow reaction regime* ($Ha<0.2$ and $Ha_{\mathrm{P}}<0.2$), Equation (6.242) yields

$$\sigma_{\mathrm{P}}=\frac{1}{1+(Al-1)\,Ha_{\mathrm{P}}^2} \tag{6.243}$$

2) In the *fast reaction regime* ($Ha>2$ and $Ha_{\mathrm{P}}>2$), Equation (6.242) yields

$$\sigma_{\mathrm{P}}=\frac{Ha}{Ha+Ha_{\mathrm{P}}} \tag{6.244}$$

In both cases the selectivity is between zero and unity. Of course it follows that $\sigma_{\mathrm{P}}>0$ because of the assumption $\overline{C_{\mathrm{P,g}}}=0$. In the slow reaction regime the interfacial area should be made as

large – and the liquid volume as small – as possible (decrease of Al), which is the case in a foam reactor.

For more complex reaction schemes, expressions for selectivity can be derived in a way similar to that illustrated above. However, the mathematics can become quite involved and tedious.

6.2.4. Reactions of Gases in Liquids – Part II:Reactants from both the Gas and the Liquid

6.2.4.1. Introduction

In this section the following system is discussed:

$$A\,(g) \longrightarrow A\,(l)$$

$$A\,(l) + \nu\,B\,(l) \longrightarrow \text{Products}\,(l)$$

where (g) and (l) refer to the gas and liquid phases, respectively (more generally they refer to phase 1 and phase 2, respectively). Hence, one reactant is transferred from the gas phase to react with a component already present in the liquid phase.

Three limiting cases are possible:

Slow Reaction. The reaction rate is low compared to the rate of mass transfer. Characteristic of this regime is that the rate of mass transfer is not enhanced by the reaction and the reaction proceeds mainly in the bulk of the liquid phase (see Fig. 38).

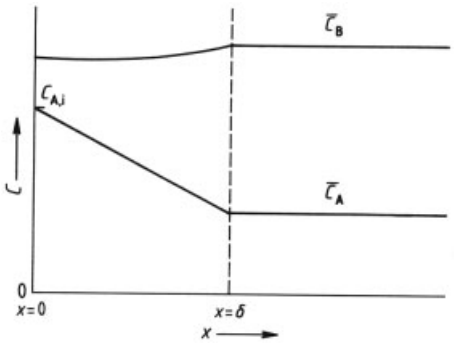

Figure 38. Concentration profiles (film theory) for reaction between reactants from the gas and liquid phases giving a slow reaction in the bulk of the liquid

Fast Reaction. The rate of reaction is so high compared to the mass-transfer rate that A is completely converted near the interf ace. For this situation the rate of mass transfer is enhanced by the chemical reaction and $\overline{C_A} = 0$ (see Fig. 39).

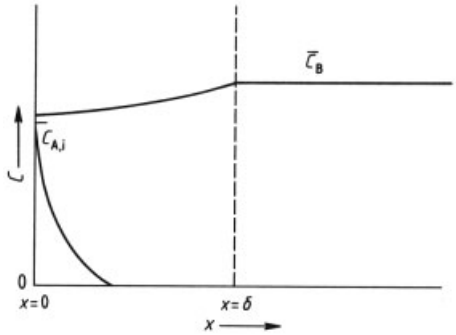

Figure 39. Concentration profiles (film theory) for reaction between reactants from the gas and liquid phases giving a fast reaction in the film on the liquid side

Instantaneous Reaction. An instantaneous reaction is, in fact, an asymptotic situation of the previous case. Now, the reaction rate is so high that the overall conversion rate is limited completely by diffusion of both A and B into the reaction zone. A and B do not occur at the same place. The place where A and B meet each other is known as the reaction plane in which both C_A and C_B are zero. In this case the enhancement factor E_A reaches its maximum value (see Fig. 40).

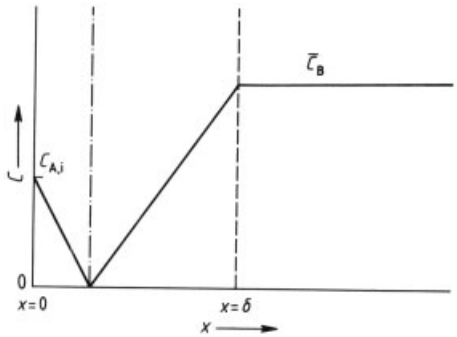

Figure 40. Concentration profiles (film theory) for a reaction between reactants from the gas and liquid bulk giving an instantaneous reaction

In the following sections, each case is analyzed in detail. Unless stated otherwise, (1,1)-kinetics are assumed:

$$|\mathbb{R}\,(C_A, C_B)\,| = k_{1,1} C_A\, C_B \tag{6.245}$$

6.2.4.2. Slow Reaction

Here, mass transfer through the interface is not enhanced by reaction. The reaction occurs mainly in the bulk of the liquid phase. This regime is identical to the case in which $Ha < 0.2$ for monomolecular first-order kinetics (see Eq. 6.218), albeit now the pseudo-first-order reaction rate constant k' is given by

$$k' = k_{1,1}\,\overline{C_{\mathrm{B}}} \tag{6.246}$$

so that the Hatta number is defined as

$$Ha = \frac{\sqrt{k'\mathbb{D}_{\mathrm{A}}}}{k_1} = \frac{\overline{k_{1,1}\,\overline{C_{\mathrm{B}}}\mathbb{D}_{\mathrm{A}}}}{k_1} \tag{6.247}$$

Notice that this equation is different from that for (1,1)-kinetics if both reactants are transferred from the gas phase (see Eq. 6.197). Equation (6.247) implies that in reactors where the bulk composition of B depends on the place inside the reactor, Ha is also place dependent. For such reactors the condition $Ha < 0.2$ must be verified to hold everywhere in the reactor.

In line with results obtained in the previous section, the degree of reactor utilization is again given by

$$\eta^{\ominus} = \frac{\overline{C_{\mathrm{A}}}}{C_{\mathrm{A,i}}} = \frac{1}{1 + Al \cdot Ha^2} \tag{6.248}$$

provided only negligible amounts of A enter or leave the reactor with the liquid phase (compared to the amount of A transferred through the interface).

Notice that the case of slow reaction does not exclude a concentration gradient of B in the film (in terms of the stagnant film theory). The above condition $Ha < 0.2$ guarantees a straight-line profile of A in the film (no chemical enhancement of mass transfer for A), but no statement is made about the gradient of B. However, as long as $Al >> 1$, possible concentration gradients of B in the film hardly influence the overall conversion rate because the reaction takes place mainly in the liquid bulk, even at low values of $\overline{C_{\mathrm{A}}}/C_{\mathrm{A,i}}$.

6.2.4.3. Fast Reaction

For $Ha > 2$ – with Ha defined by Equation (6.247) – the reaction takes place exclusively near the interface. In terms of stagnant film theory this means that all reaction occurs in the stagnant film. Thus far, the situation is identical with the previously treated case of first-order reactions with reactants from the gas only. The situation is complicated by the fact that the enhancement factor E_{A} depends on the concentration of B in the film (i.e., by whether diffusion limitation of B occurs in the reaction zone or not). If more than enough B can be supplied to the film, the situation $C_{\mathrm{B}} = \overline{C_{\mathrm{B}}}$ is approached everywhere in the film. Hence, the reaction can be treated as a pseudo-first-order monomolecular reaction with, again, $k' = k_{1,1}\,\overline{C_{\mathrm{B}}}$, so that the Hatta number equals

$$Ha = \frac{\sqrt{k_{1,1}\,\overline{C_{\mathrm{B}}}\mathbb{D}_{\mathrm{A}}}}{k_1} \tag{6.249}$$

and the enhancement factor $E_{\mathrm{A}} = Ha$ because $Ha > 2$.

The criterion determining whether or not this approach is allowed is derived below by using the stagnant film model. The assumption of a constant C_{B} $(= \overline{C_{\mathrm{B}}})$ all over the film is allowed if (see Fig. 41)

$$\frac{\overline{C_{\mathrm{B}}} - C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}} << 1 \tag{6.250}$$

The reaction takes place mainly near the interface in a smaller zone within the stagnant film of thickness δ. The thickness of this smaller reaction zone is denoted by δ_{r}. Its value follows from

$$E_{\mathrm{A}} = \frac{\overline{J_{\mathrm{A}}}}{k_1\left(C_{\mathrm{A,i}} - \overline{C_{\mathrm{A}}}\right)} = -\frac{\mathbb{D}_{\mathrm{A}}\left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)\Big|_{x=0}}{k_1 C_{\mathrm{A,i}}}$$

$$= \frac{\mathbb{D}_{\mathrm{A}}}{k_1 C_{\mathrm{A,i}}}\frac{C_{\mathrm{A,i}}}{\delta_{\mathrm{r}}} = \frac{\delta}{\delta_{\mathrm{r}}} \Rightarrow \delta_{\mathrm{r}} = \frac{\delta}{E_{\mathrm{A}}} = \frac{\delta}{Ha} \tag{6.251}$$

The conversion rate of B in the reaction zone must equal the transport rate of B from the bulk to the zone $0 < x < \delta_{\mathrm{r}}$. A conservative simplification gives

$$\overline{J_{\mathrm{B}}} \approx \frac{\mathbb{D}_{\mathrm{B}}}{\delta}\left(\overline{C_{\mathrm{B}}} - C_{\mathrm{B,i}}\right) \tag{6.252}$$

Actually $\overline{J_{\mathrm{B}}}$ will be somewhat higher since the slope at $x = \delta$ is higher than the above average slope. For the mole flux of A, Equation (6.251) gives

$$\overline{J_A} = \frac{\mathbb{D}_A}{\delta_r} C_{A,i} = \frac{\mathbb{D}_A}{\delta} C_{A,i}\, Ha \tag{6.253}$$

Combining Equations (6.252) and (6.253) together with $\overline{J_B} = \nu\, \overline{J_A}$ yields

$$\mathbb{D}_B \left(\overline{C_B} - C_{B,i}\right) \approx \nu \mathbb{D}_A C_{A,i}\, Ha \tag{6.254}$$

Thus, from this result, $(\overline{C_B} - C_{B,i})/\overline{C_B} << 1$ for

$$Ha << \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}} \tag{6.255}$$

Hence, in conclusion,

$$E_A = Ha \text{ for } 2 < Ha << \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}} \tag{6.256}$$

with Ha according to Equation (6.249).

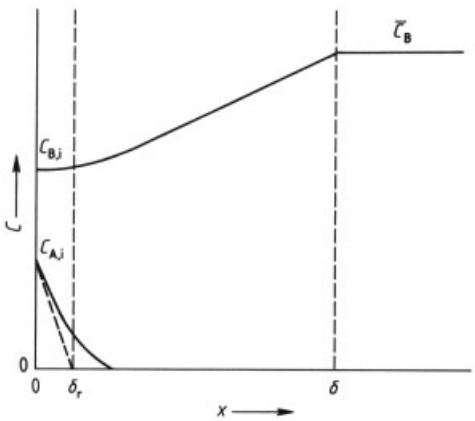

Figure 41. General concentration profile (film theory) for a reaction between reactants from the gas and liquid phases and $Ha > 2$

If Ha is not much lower than $(\mathbb{D}_B\, \overline{C_B})/(\nu\, \mathbb{D}_A\, C_{A,i})$, diffusion limitation of B occurs, the maximum transport rate of B becomes slow compared to the maximum possible conversion rate of B in the reaction zone, and $C_{B,i}$ becomes lower than $\overline{C_B}$. With increasing value of $Ha \left(\frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_{A,i} C_{A,i}}\right)^{-1}$, the concentration profile changes from that shown in Figure 42 A to that in Figure 42 D. The last reflects the asymptotic case of an instantaneous reaction. For the case of diffusion limitation of B, E_A will be lower than Ha. This is illustrated in the following section.

6.2.4.4. Instantaneous Reaction

When the reaction rate is very fast compared to the mass-transfer rates of both A and B, the reaction is called instantaneous. Section 6.2.4.4 shows that this is the case for both $Ha > 2$ and

$$Ha >> \frac{\mathbb{D}_B \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}} \tag{6.257}$$

The reaction zone reduces to a plane at $x = \delta_r$: A and B react so fast that they are not found at the same place, except in that plane. In terms of film theory, mass transfer of A and B to the reaction plane is given by (see Fig. 42 D)

$$\overline{J_A} = \frac{\mathbb{D}_A}{\delta_r} C_{A,i} \tag{6.258}$$

$$\overline{J_B} = \frac{\mathbb{D}_B}{\delta_B - \delta_r}\, \overline{C_B} \tag{6.259}$$

Combination with $\overline{J_B} = \nu\, \overline{J}_A$ and elimination of δ_r and $\overline{J_B}$ give

$$\overline{J_A} = \frac{\mathbb{D}_A}{\delta_B} C_{A,i} \left(1 + \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \tag{6.260}$$

The film thickness in this equation, δ_B, is that of component B. To express $\overline{J_A}$ in the mass-transfer coefficient of A, $k_{l,A}$, the film thickness for component A, δ_A, must be introduced:

$$\overline{J_A} = \frac{\mathbb{D}_A}{\delta_A} C_{A,i} \left(1 + \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \left(\frac{\delta_A}{\delta_B}\right) \tag{6.261}$$

or, because the film model is used,

$$\overline{J_A} = k_{l,A}\, C_{A,i} \left(1 + \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \left(\frac{k_{l,B}}{k_{l,A}} \frac{\mathbb{D}_A}{\mathbb{D}_B}\right) \tag{6.262}$$

Hence the enhancement factor E_A, which in this regime is generally referred to as $E_{A\infty}$, follows from

$$E_A \equiv E_{A\infty} = \left(1 + \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \left(\frac{k_{l,B}}{k_{l,A}} \frac{\mathbb{D}_A}{\mathbb{D}_B}\right) \tag{6.263}$$

The ratio $k_{l,A}/k_{l,B}$ generally equals $(\mathbb{D}_A/\mathbb{D}_B)^n$, where the number n depends on the model used or the local hydrodynamics. Thus, in general,

$$E_{A\infty} = \left(1 + \frac{\mathbb{D}_B\, \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \left(\frac{\mathbb{D}_A}{\mathbb{D}_B}\right)^{1-n} \tag{6.264}$$

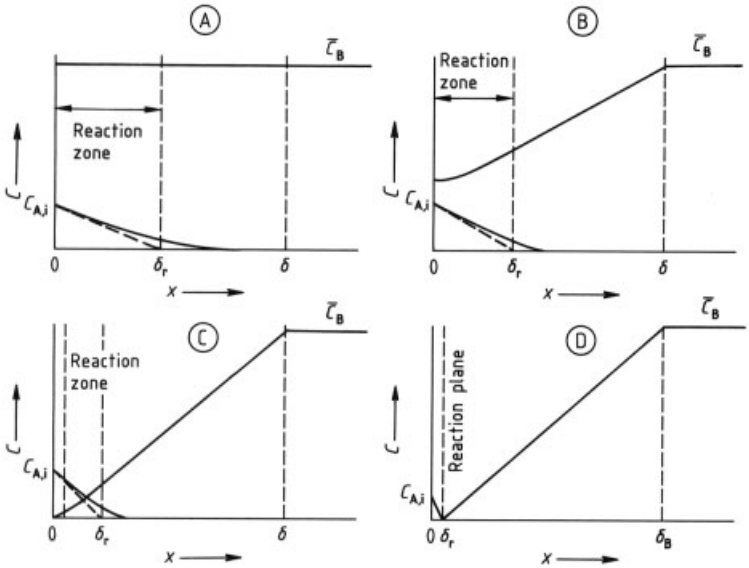

Figure 42. Concentration profiles (film model) for mass transfer with simultaneous irreversible reaction in the liquid phase for a reaction of the order (1,1) with $Ha > 2$ and the following:

A) $Ha << \frac{D_B \overline{C_B}}{v D_A C_{A,i}}$; B) $Ha \approx \frac{D_B \overline{C_B}}{v D_A C_{A,i}}$; C) $Ha \approx 5 \frac{D_B \overline{C_B}}{v D_A C_{A,i}}$; D) $Ha \approx 10 \frac{D_B \overline{C_B}}{v D_A C_{A,i}}$

Higbie's penetration model assumes that the contact time τ of the small stagnant fluid element attached to the interface is an only function of the hydrodynamics and does not depend on physical properties such as the diffusion coefficient $\mathbb{D}$. From this assumption, $k_1 = 2\sqrt{\mathbb{D}/(\pi\tau)} \sim \sqrt{\mathbb{D}}$ (see Eq. 5.26) and hence n in Equation (6.264) equals 1/2.

Generally, the enhancement factor for instantaneous reactions equals, at least to a good approximation [41], [42],

$$E_{A\infty} \approx \left(1 + \frac{\mathbb{D}_B \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}}\right) \left(\frac{\mathbb{D}_A}{\mathbb{D}_B}\right)^{1-n} \tag{6.265}$$

with, e. g.,

$n = 1/2$	Penetration model
$n = 2/3$	Laminar flow
$n = 2/3 - 3/4$	Fully developed turbulent flow

If $\overline{C_B}/(\nu\, C_{A,i})$ is not much larger than unity, the approximation

$$E_{A\infty} \approx 1 + \frac{\mathbb{D}_B \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}} \left(\frac{\mathbb{D}_A}{\mathbb{D}_B}\right)^{1-n} \tag{6.266}$$

is preferred [41].

Theoretically, Fick's law, which is the basis of all derivations discussed above, is applicable only for a plane through which no overall mass flux occurs (e.g., see [1]). Especially for instantaneous reactions, high mass-transfer rates can be obtained and the validity of this assumption becomes questionable. STEPANEK and SHILIMKAN derived a more general equation for the enhancement factor $E_{A\infty}$ for instantaneous reaction, on the basis of the film theory [43]. For the reaction

$$A + \nu_B B \rightarrow \nu_P P + \nu_Q Q$$

they obtained

$$E_{A\infty} = 1 + \frac{1}{\Delta\nu} \frac{\mathbb{D}_B}{\mathbb{D}_A} \frac{\ln\left(1 + \frac{\Delta\nu}{\nu_B} \overline{\kappa}_B\right)}{\ln\left(\frac{1}{1-\kappa_{A,i}}\right)} \tag{6.267}$$

with

$$\Delta\nu = \nu_P + \nu_Q - \nu_B \tag{6.268}$$

where $\kappa_{A,i}$ is the interface mole fraction of A and $\overline{\kappa}_B$ the mole fraction of B in the liquid bulk. For dilute systems (i.e., $\kappa_{A,i}$ and $\overline{\kappa}_B << 1$ with $\kappa_{A,i}$ and $\overline{\kappa}_B \approx C_{A,i}/C$ and $\overline{C_B}/C$, respectively,

since the molar density is constant), Equation (6.267) simplifies to

$$E_{A\infty} \approx 1 + \frac{\mathbb{D}_B \overline{C_B}}{\nu_B \mathbb{D}_A C_{A,i}} \tag{6.269}$$

which resembles Equation (6.266).

6.2.4.5. Van Krevelen – Hoftijzer Approximation

Here a more rigorous general solution is discussed for simultaneous mass transfer and reaction in a liquid with reactants from both the gas and the liquid phases, in which (1,1)-kinetics is assumed. The differential equations describing the reaction and mass transfer near the interface for the stagnant film model state

$$\mathbb{D}_A \frac{\partial^2 C_A}{\partial x^2} - k_{1,1} C_A C_B = 0 \tag{6.270}$$

$$\mathbb{D}_B \frac{\partial^2 C_B}{\partial x^2} - k_{1,1} C_A C_B = 0 \tag{6.271}$$

with the boundary conditions

$$x = 0 \Rightarrow C_A = C_{A,i} \quad \text{and} \quad \frac{\partial C_B}{\partial x} = 0 \tag{6.272}$$

$$x = \delta \Rightarrow C_A = \overline{C_A} \quad \text{and} \quad C_B = \overline{C_B} \tag{6.273}$$

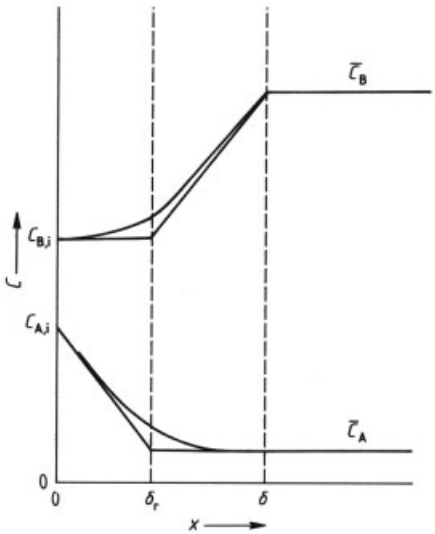

Figure 43. Concentration profile approximation (film model) as suggested by VAN KREVELEN and HOFTIJZER

The boundary condition $\partial C_B/\partial x = 0$ for $x = 0$ states that no B is transferred to the gas phase, meaning that B is not volatile (in gas – liquid systems) or not soluble in the phase where no reaction occurs (if liquid – liquid reactions are considered).

An approximate solution for the above set of equations has been presented by VAN KREVELEN and HOFTIJZER [44]. Their method of approximation has been applied successfully to solve many complex problems in mass transfer with simultaneous reaction. The essence of this method is the approximation of the concentration profile of component B by a constant $C_{B,i}$ over the entire reaction zone (see Fig. 43). The higher the value of δ/δ_r or, according to Equation (6.251), the higher the Hatta number, the better is the approximation. Then, left to the plane $x = \delta_r$, the (1,1)-kinetics reaction becomes pseudo-first order with a rate constant $k' = k_{1,1} C_{B,i}$:

$$|\mathbb{R}(C_A, C_B)| = (k_{1,1} C_{B,i}) C_A \tag{6.274}$$

Hence Equation (6.270) simplifies to

$$\mathbb{D}_A \frac{\partial^2 C_A}{\partial x^2} - (k_{1,1} C_{B,i}) C_A = 0 \tag{6.275}$$

subject to Equations (5.21) and (5.22):

$$x = 0 \Rightarrow C_A = C_{A,i}$$
$$x = \delta \Rightarrow C_A = \overline{C_A}$$

The solution is given by (see Eq. 6.180)

$$\overline{J_A} = k_{l,A} C_{A,i} \frac{Ha_i}{\tanh(Ha_i)} \left(1 - \frac{\overline{C_A}}{C_{A,i}} \frac{1}{\cosh(Ha_i)}\right) \tag{6.276}$$

with Ha_i a *modified Hatta number* based on the interface concentration $C_{B,i}$:

$$Ha_i = \frac{\sqrt{k' \mathbb{D}_A}}{k_l} = \frac{\sqrt{k_{1,1} C_{B,i} \mathbb{D}_A}}{k_l} = Ha \sqrt{\frac{C_{B,i}}{\overline{C_B}}}$$

where

$$Ha = \frac{\sqrt{k_{1,1} C_B \mathbb{D}_A}}{k_l} \tag{6.277}$$

The unknown concentration $C_{B,i}$ is obtained from mass balances. Since $\overline{J_B} = \nu \overline{J_A}$,

$$\frac{\mathbb{D}_B}{\delta - \delta_r} (\overline{C_B} - C_{B,i}) = \nu \frac{\mathbb{D}_A}{\delta_r} (C_{A,i} - \overline{C_A}) \tag{6.278}$$

or

$$\frac{C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}} = 1 - \left(\frac{\delta - \delta_{\mathrm{r}}}{\delta_{\mathrm{r}}}\right) \nu \frac{\mathbb{D}_{\mathrm{A}}}{\mathbb{D}_{\mathrm{B}}} \frac{C_{\mathrm{A,i}} - \overline{C_{\mathrm{A}}}}{\overline{C_{\mathrm{B}}}} \tag{6.279}$$

which for low bulk concentrations $\overline{C_{\mathrm{A}}}$ (high Hatta numbers) reduces to

$$\frac{C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}} = 1 - \left(\frac{\delta - \delta_{r}}{\delta_{\mathrm{r}}}\right) \nu \frac{\mathbb{D}_{\mathrm{A}}}{\mathbb{D}_{\mathrm{B}}} \frac{C_{\mathrm{A,i}}}{C_{\mathrm{B}}} \tag{6.280}$$

An expression for $\delta/\delta_{\mathrm{r}}$ is obtained by defining a *modified enhancement factor* E_{A0} according to

$$\overline{J_{\mathrm{A}}} = -\mathbb{D}_{\mathrm{A}} \left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)\Big|_{x=0} = \frac{\mathbb{D}_{\mathrm{A}}}{\delta} C_{\mathrm{A,i}} E_{\mathrm{A0}} \tag{6.281}$$

(see Eq. 6.238). For low bulk concentrations $\overline{C_{\mathrm{A}}}$ (thus high Hatta numbers), this yields

$$-\left(\frac{\partial C_{\mathrm{A}}}{\partial x}\right)\Big|_{x=0} = \frac{C_{\mathrm{A,i}}}{\delta_{\mathrm{r}}} = \frac{C_{\mathrm{A,i}}}{\delta} E_{\mathrm{A0}}$$
$$\rightarrow E_{\mathrm{A0}} = \frac{\delta}{\delta_{\mathrm{r}}} \tag{6.282}$$

This should be compared with Equation (6.251) where the same relation was found for E_{A} provided $\overline{C_{\mathrm{A}}} \approx 0$. However, E_{A0} is not identical to the enhancement factor E_{A} ; only for the case in which $\overline{C_{\mathrm{A}}} = 0$ does E_{A0} equal E_{A} . In general,

$$\overline{J_{\mathrm{A}}} = k_{\mathrm{l,A}} C_{\mathrm{A,i}} E_{\mathrm{A0}} = k_{\mathrm{l,A}} \left(C_{\mathrm{A,i}} - \overline{C_{\mathrm{A}}}\right) E_{\mathrm{A}} \tag{6.283}$$

and thus, $E_{\mathrm{A0}} \neq E_{\mathrm{A}}$.

Substituting the relation obtained for $\delta/\delta_{\mathrm{r}}$ by Equation (6.282) in Equation (6.280), gives

$$\frac{C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}} = 1 - (E_{\mathrm{A0}} - 1)\, \nu \frac{\mathbb{D}_{\mathrm{A}}}{\mathbb{D}_{\mathrm{B}}} \frac{C_{\mathrm{A,i}}}{\overline{C_{\mathrm{B}}}} \tag{6.284}$$

Furthermore, by approximating the enhancement factor for instantaneous reaction $E_{\mathrm{A}\infty}$ as in Equation (6.269)

$$E_{\mathrm{A}\infty} \approx 1 + \frac{\mathbb{D}_{\mathrm{B}} \overline{C_{\mathrm{B}}}}{\nu \mathbb{D}_{\mathrm{A}} C_{\mathrm{A,i}}} \tag{6.285}$$

Equation (6.284) becomes

$$\frac{C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}} = \frac{E_{\mathrm{A}\infty} - E_{\mathrm{A0}}}{E_{\mathrm{A}\infty} - 1} \tag{6.286}$$

Notice that this equation was obtained without using the rate equation of —$\mathbb{R}$—: thus Equation (6.286) also applies for reactions with arbitrary kinetics (this is used in Section 6.2.4.6) The parameter Ha_{i} can be expressed in terms of Ha by combining Equations (6.286) and (6.277),

$$Ha_{\mathrm{i}} = Ha \sqrt{\frac{C_{\mathrm{B,i}}}{\overline{C_{\mathrm{B}}}}} = Ha \sqrt{\frac{E_{\mathrm{A}\infty} - E_{\mathrm{A0}}}{E_{\mathrm{A}\infty} - 1}} \tag{6.287}$$

Combining Equations (6.276) and (6.283) yields

$$E_{\mathrm{A0}} = \frac{Ha_{\mathrm{i}}}{\tanh\,(Ha_{\mathrm{i}})} \left(1 - \frac{\overline{C_{\mathrm{A}}}}{C_{\mathrm{A,i}}} \frac{1}{\cosh\,(Ha_{\mathrm{i}})}\right) \tag{6.288}$$

hence,

$$E_{\mathrm{A0}} = \frac{Ha \sqrt{\frac{E_{\mathrm{A}\infty} - E_{\mathrm{A0}}}{E_{\mathrm{A}\infty} - 1}}}{\tanh\left(Ha \sqrt{\frac{E_{\mathrm{A}\infty} - E_{\mathrm{A0}}}{E_{\mathrm{A}\infty} - 1}}\right)} \cdot \left[\left(1 - \frac{\overline{C_{\mathrm{A}}}}{C_{\mathrm{A,i}}}\right) \frac{1}{\cosh\left(Ha \sqrt{\frac{E_{\mathrm{A}\infty} - E_{\mathrm{A0}}}{E_{\mathrm{A}\infty} - 1}}\right)}\right] \tag{6.289}$$

Equation (6.289) is an implicit relation from which the enhancement factor E_{A0} can be obtained by trial and error for any given values of $C_{\mathrm{A,i}}$, $\overline{C_{\mathrm{A}}}$, Ha, and $E_{\mathrm{A}\infty}$. Once E_{A0} is obtained, the molar flux of A through the interface follows from

$$\overline{J_{\mathrm{A}}} = k_{\mathrm{l,A}} C_{\mathrm{A,i}} E_{\mathrm{A0}} \tag{6.290}$$

The three limiting regimes of slow reaction, fast reaction, and instantaneous reaction are all included in Equation (6.289). The accuracy of the approximation is remarkably good: deviation from the exact solution is $< 3\,\%$ according to De Santiago and Farina [45]. Figure 44 presents the solution for $\overline{C_{\mathrm{A}}} = 0$, i.e., for $Al \cdot Ha^2 >> 1$.

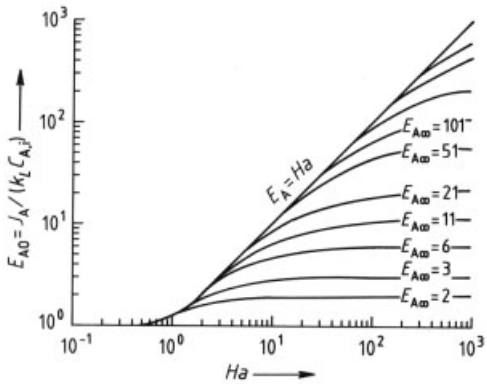

Figure 44. Enhancement factor for mass transfer with simultaneous (1,1)-kinetics and for a reaction between reactants from the gas and liquid phases; the figure holds only for $\overline{C_{\mathrm{A}}} = 0$

The above approximation was derived with the aid of the film model. A similar approximation can be derived, for example, from the Higbie penetration model. BRIAN et al. [46] showed that the solutions derived from the penetration and film models are nearly identical.

6.2.4.6. Arbitrary Kinetics

Also for arbitrary kinetics —$\mathbb{R}(C_A, C_B)$— an enhancement factor for mass transfer can be found by applying the Van Krevelen – Hoftijzer approximation. The general principle is always to assume that $C_B \approx C_{B,i}$ in the reaction zone, so the rate equation reduces to

$$|\mathbb{R}(C_A, C_B)| = |\mathbb{R}(C_A, C_{B,i})| \tag{6.291}$$

and is a function only of C_A. Hence, the problem has been reduced to the case of absorption of only one reactant from the gas and arbitrary kinetics, which is treated in Section 6.2.3.7. The Hatta number obtained is a modified one, Ha_i, and is based on $C_{B,i}$ and not on $\overline{C_B}$ like the usual Hatta number Ha. Since Equation (6.286) was derived without using the rate equation, both Hatta numbers can be related by using Equation (6.286)

$$\frac{C_{B,i}}{\overline{C_B}} = \frac{E_{A\infty} - E_{A0}}{E_{A\infty} - 1}$$

As an example, bimolecular reactions with (n,m)-kinetics are discussed:

$$|\mathbb{R}(C_A, C_B)| = k_{n,m} C_A^n C_B^m \tag{6.292}$$

With the Van Krevelen – Hoftijzer assumption, this equation reduces to

$$|\mathbb{R}(C_A, C_B)| = \left(k_{n,m} C_{B,i}^m\right) C_A^n \tag{6.293}$$

and becomes a function of C_A only. The modified Hatta number is derived from (see Eq. 6.190)

$$Ha_i = \frac{\mathbb{D}_A}{k_{l,A}} |\mathbb{R}_A(C_{A,i}, C_{B,i})| \cdot \left[2\mathbb{D}_A \int_0^{C_{A,i}} |\mathbb{R}_A(C_A, C_{B,i})| \partial C_A\right]^{-\frac{1}{2}} \tag{6.294}$$

For pseudo-n-th-order kinetics this becomes (see Eq. 6.192)

$$Ha_i = \frac{\mathbb{D}_A}{k_{l,A}} \sqrt{\frac{n+1}{2} \frac{k_{n,m} C_{A,i}^{n-1} C_{B,i}^m}{\mathbb{D}_A}} \tag{6.295}$$

The enhancement factor E_{A0}, defined according to Equation (6.290)

$$\overline{J_A} = k_{1,A} C_{A,i} E_{A0}$$

can be approximated with Equation (6.183):

$$E_{A0} = \frac{Ha_1}{\tanh(Ha_i)} \tag{6.296}$$

Analogous to Equation (6.277), the true Hatta number is defined as

$$Ha = \frac{\mathbb{D}_A}{k_{l,A}} \sqrt{\frac{n+1}{2} \frac{k_{n,m} C_{A,i}^{n-1} \left(\overline{C_B}\right)^m}{\mathbb{D}_A}} \tag{6.297}$$

so that

$$Ha_i = Ha \sqrt{\left(\frac{E_{A\infty} - E_{A0}}{E_{A\infty} - 1}\right)^m} \tag{6.298}$$

Combination of Equations (6.296) and (6.298) gives

$$E_{A0} = \frac{Ha\sqrt{\left(\frac{E_{A\infty} - E_{A0}}{E_{A\infty} - 1}\right)^m}}{\tanh\left[Ha\sqrt{\left(\frac{E_{A\infty} - E_{A0}}{E_{A\infty} - 1}\right)^m}\right]} \tag{6.299}$$

E_{A0} is obtained from this implicit relation; then the molar flux of A is calculated by using Equation (6.290). Equation (6.299) was first derived by HIKITA and ASAI [47]. The accuracy of this approximation is within 10 %.

For fast reactions and no diffusion limitation of B (see Eq. 6.256),

$$2 < Ha << E_{A\infty} = \frac{\mathbb{D}_B \overline{C_B}}{\nu \mathbb{D}_A C_{A,i}} \tag{6.300}$$

Equation (6.299) reduces to

$$E_{A0} = Ha \tag{6.301}$$

Hence, the conversion rate per cubic meter of dispersion volume follows from

$$|\mathbb{R}_A| = a\overline{J_A} = a k_{l,A} C_{A,i} Ha =$$

$$a\sqrt{\frac{2\mathbb{D}_A k_{n,m}}{n+1}} C_{A,i}^{[(n+1)/2]} \overline{C_B}^{(m/2)} \tag{6.302}$$

Equation (6.302) shows how the apparent reaction orders change, due to diffusion limitation of A. Hence, again mass-transfer limitation can obviously affect the apparent reactant order. Order (0,0) changes to ($\frac{1}{2}$,0); order (1,1), to (1,$\frac{1}{2}$); and (2,2), to ($\frac{3}{2}$,1).

Approximate solutions for (*n,m*)-kinetics by using the Van Krevelen – Hoftijzer assumption with the penetration models have been presented as well. HIKITA and ASAI [47] analyzed the Higbie penetration model; DE COURSEY [48] presented an approximation using the Danckwerts penetration model. The latter is rather practical because an explicit relation is obtained for E_{A0} :

$$E_{A0} = \sqrt{\frac{Ha^4}{4\,(E_{A\infty}-1)^2} + \frac{E_{A\infty}Ha^2}{E_{A\infty}-1} + 1} - \frac{Ha^2}{2\,(E_{A\infty}-1)} \tag{6.303}$$

Differences in accuracy between film and penetration model approximations are negligible. Use of the film model leads to easier derivation of the approximation.

General approximations for other types of kinetics can be derived with the Van Krevelen – Hoftijzer method in a way similar to that illustrated for (*n,m*)-kinetics.

6.2.4.7. Selectivities of Multiple Reactions

The mathematics for the treatment of multiple reactions become quite involved and tedious when reactant is absorbed from the gas whereas the other is already present in the liquid. An extensive overview is given in [8]; only some basic schemes are discussed here.

For mixed parallel – consecutive reactions of the type:

$$A\,(g) \longrightarrow A\,(l)$$

$$A\,(l) + B\,(l) \longrightarrow P\,(l)$$

$$A\,(l) + P\,(l) \longrightarrow X\,(l)$$

numerical and approximated enhancement factors according to stagnant film theory for reaction orders (*n,m*) and (*n′,m′*) respectively, have been presented by ONDA et al. [49]. Numerical and approximated Higbie penetration theory solutions are discussed in [50], [51].

For another type of mixed parallel – consecutive reaction network,

$$A\,(g) \longrightarrow A\,(l)$$

$$A\,(l) + B\,(l) \longrightarrow P\,(l)$$

$$P\,(l) + B\,(l) \longrightarrow X\,(l)$$

enhancement factor calculations have been carried out by BRIAN and BEAVERSTOCK [52] and by RAMACHANDRAN [53].

Simultaneous mass transfer with dependent parallel reactions

$$A_1\,(g) \longrightarrow A_1\,(l)$$

$$A_2\,(g) \longrightarrow A_2\,(l)$$

$$A_1\,(l) + B\,(l) \longrightarrow P\,(l)$$

$$A_2\,(l) + B\,(l) \longrightarrow X\,(l)$$

has been analyzed by GOETTLER and PIGFORD [54], starting from the Higbie penetration theory, and by SHKYLAR et al. [55], starting from the stagnant film model.

Simultaneous absorption followed by a complex consecutive reaction of the type

$$A_1\,(g) \longrightarrow A_1\,(l)$$

$$A_2\,(g) \longrightarrow A_2\,(l)$$

$$A_1\,(l) + B\,(l) \longrightarrow I\,(l) + P\,(l)$$

$$A_2\,(l) + I\,(l) \longrightarrow B\,(l)$$

has been analyzed by ONDA et al. [49].

Finally, the simultaneous absorption of a mixture of phosgene (A_1) and carbon dioxide (A_2) in water to remove phosgene, has the following scheme:

$$A_1\,(g) \longrightarrow A_1\,(l)$$

$$A_2\,(g) \longrightarrow A_2\,(l)$$

$$A_1\,(l) + B\,(l) \longrightarrow A_2\,(l)$$

$$A_2\,(l) + B\,(l) \longrightarrow X\,(l)$$

This system has been analyzed by RAMACHANDRAN and SHARMA [56]. Other, more complex reaction schemes are discussed by WESTERTERP et al. [8].

6.3. Combination of Simultaneous and Series Mass Transfer and Reaction

In all the above cases of mass transfer with simultaneous reaction, the effectiveness factor η and the enhancement factor E_A were written as a function of the interface concentration on the reaction-phase side of the reactants transferred to the reaction phase (e.g., $C_{A,i}$) and of the bulk concentration of the reactants already present in this phase (e.g., $\overline{C_B}$). If *no transfer resistance* of reactants from the nonreaction phase to the interface exists, the interface concentration on the reaction-phase side equals the bulk concentration in the nonreaction phase multiplied by the distribution coefficient m. For example, if the nonreaction phase is a gas and the reaction phase a liquid, for monomolecular reactions with reactant A the following holds:

$$C_{A,i,l} = m C_{A,i,g} = m \overline{C_{A,g}} \tag{6.304}$$

and all equations derived in Sections 6.1 and 6.2 can also be based on gas-phase conditions, provided $C_{A,i,l}$ in the equations is replaced by $m\,\overline{C_{A,g}}$.

If the *transfer resistance of reactants from the bulk of the nonreaction phase is not negligible*, a combination of simultaneous and series mass transfer and reaction can occur: the former in the reaction phase and the latter in the nonreaction phase. A calculation of the overall conversion rate for such cases is shown below. The nonreaction phase is assumed to be the gas phase; the reaction phase might be a porous catalyst pellet or a liquid in which reaction occurs.

For a monomolecular reaction with a reactant A, a mass balance gives

$$\overline{J_A} = k_g \left(\overline{C_{A,g}} - \frac{C_{A,i,l}}{m} \right) = \frac{\eta \left| \mathbb{R}\left(C_{A,i,l}\right) \right|}{a} \tag{6.305}$$

Here a is the specific contact area in square meters of interface area per cubic meter of dispersion volume (e.g., volume of porous pellets plus interstitial gas, or volume of gas bubbles plus liquid bulk). The reaction rate — $\mathbb{R}$ — has the dimension of moles of A per cubic meter of dispersion volume per unit time.

For reactions in porous pellets where $m = 1$, Equation (6.305) reduces to

$$\overline{J_A} = k_g \left(\overline{C_{A,g}} - C_{A,i,g} \right) = \eta \frac{V_p}{A_p} \left| \mathbb{R}\left(C_{A,i,g}\right) \right|_p \tag{6.306}$$

If the concept of an enhancement factor is used instead of the effectiveness factor, Equation (6.305) can be written as (e.g., for gas – liquid systems)

$$\overline{J_A} = k_g \left(\overline{C_{A,g}} - \frac{C_{A,i,l}}{m} \right) = E_A\, k_l\, C_{A,i,l} \tag{6.307}$$

where it is assumed that the reaction phase is a liquid, hence, the index $_l$ in k_l.

The value of η depends only on $C_{A,i,g}$, the reaction kinetics, the size of the catalyst pellet, and (to a lesser extent) the geometry; E_A depends only on $C_{A,i,l}$ and the reaction kinetics. Hence, $C_{A,i,g}$ and $C_{A,i,l}$ can be eliminated from Equations (6.306) and (6.307), respectively, to give the molar flux $\overline{J_A}$ – and thus the overall conversion rate – as a function of $\overline{C_{A,g}}$, the kinetic constants, and the mass-transfer parameters (a, k_g, and $\mathbb{D}_i$; k_g, k_l, and m).

Relatively simple solutions are found for first-order kinetics:

$$\left| \mathbb{R}\left(C_A\right) \right| = k C_A \tag{6.308}$$

In this case, neither the effectiveness factor η nor the enhancement factor E_A depends on the interface concentrations $C_{A,i,g}$ or $C_{A,i,l}$. Hence, Equation (6.306) yields

$$\overline{J_A} = \left(\frac{1}{k_g} + \frac{a}{\eta k} \right)^{-1} \overline{C_{A,g}} \tag{6.309}$$

and from Equation (6.307),

$$\overline{J_A} = \left(\frac{1}{k_g} + \frac{1}{E_A\, m\, k_l} \right)^{-1} \overline{C_{A,g}} \tag{6.310}$$

For arbitrary kinetics, calculations are similar to those above, albeit analytical solutions often cannot be obtained and numerical approximations must be used. Consider for example the case of Langmuir kinetics in a porous catalyst pellet:

$$\left| \mathbb{R}_A \right|_p = \frac{k C_A}{1 + K C_A} \tag{6.311}$$

(see Eq. 6.114). Then the Thiele modulus is found as (see Eq. 6.115):

$$\varphi = \frac{V_p}{A_p} \frac{K C_{A,i,g}}{1 + K C_{A,i,g}} \cdot \sqrt{\frac{k}{2\,\mathbb{D}_{i,A} \left[K C_{A,i,g} - \ln\left(1 + K C_{A,i,g}\right) \right]}} \tag{6.312}$$

The effectiveness factor can be approximated for example by Equation (6.116):

$$\eta = \frac{\tanh(\varphi)}{\varphi}$$

So, Equation (6.306) yields

$$k_g a \left(\overline{C_{A,g}} - C_{A,i,g}\right) = \frac{\tanh(\varphi)}{\varphi} \frac{k C_{A,i,g}}{1 + K C_{A,i,g}} \quad (6.313)$$

Since $C_{A,i,g}$ can be eliminated from Equations (6.312) and (6.313), the conversion rate can be obtained from

$$|\mathbb{R}_A| = \overline{J_A} a = k_g a \left(\overline{C_{A,g}} - C_{A,i,g}\right) \quad (6.314)$$

which gives the conversion rate as a function of $\overline{C_{A,g}}$, k, K, k_g, $\mathbb{D}_{i,A}$, a, and the characteristic pellet dimension V_p/A_p.

Numerical approximations must be used also for bimolecular reactions in general. Again, reaction in a porous catalyst pellet is considered for a reaction $A + \nu B \longrightarrow$ products. Both reactants A and B are supplied from the gas phase. Then, mass balances yield(see Eqs. 6.16, 6.18, and 6.19):

$$\overline{J_A} = k_{g,A} \left(\overline{C_{A,g}} - C_{A,i,g}\right) = \frac{\eta}{a} |\mathbb{R}\left(C_{A,i,g}, C_{B,i,g}\right)| \quad (6.315)$$

and since $\overline{J_B} = \nu \overline{J_A}$,

$$\overline{J_B} = k_{g,B} \left(\overline{C_{B,g}} - C_{B,i,g}\right) = \nu \frac{\eta}{a} |\mathbb{R}\left(C_{A,i,g}, C_{B,i,g}\right)| \quad (6.316)$$

If the excess factor γ is defined as

$$\gamma = \nu \frac{\mathbb{D}_{i,A}}{\mathbb{D}_{i,B}} \frac{C_{A,i,g}}{C_{B,i,g}} \quad (6.317)$$

the Thiele modulus is found from (see Eq. 6.123)

$$\varphi = \frac{V_p}{A_p} |\mathbb{R}\left(C_{A,i,g}, C_{B,i,g}\right)| \left(2\mathbb{D}_{i,A} \int_0^{C_{A,i,g}} |\mathbb{R} \cdot \left\{C_A, C_{B,i,g}\left[1 - \gamma\left(1 - C_A/C_{A,i,g}\right)\right]\right\}| \partial C_A\right)^{-\frac{1}{2}} \quad (6.318)$$

Finally, for the effectiveness factor an approximation can again be used (e.g., Eq. 6.116):

$$\eta = \frac{\tanh(\varphi)}{\varphi}$$

From Equations (6.116) and (6.315)–(6.318), the concentrations $C_{A,i,g}$ and $C_{B,i,g}$ can be eliminated to yield the conversion rate $\overline{J_A}\, a$ as a function of, for example, $\overline{C_{A,g}}$ and $\overline{C_{B,g}}$.

Generalizations like those illustrated above for the effectiveness factor η can be made in a similar way for the enhancement factor E_A.

7. Macrokinetics: Multiphase Systems – Part III: Mass and Heat Transfer with Reaction

7.1. Introduction

Heat effects occur in all chemical reactions. Either heat is produced and the reaction is said to be exothermic or heat is consumed and the reaction is endothermic. These heat effects can influence the temperature of the reaction mixture and, therefore, the chemical reaction rate and the overall conversion rate.

Consequently, for macrokinetic phenomena, apart from mass transfer, heat-transfer resistance can also play a significant role in the overall conversion rate. In this section, the effect of heat-transfer resistance in combination with mass-transfer resistance is discussed. The discussion is limited to the case in which the temperature dependence of the reaction rate is of the Arrhenius type and the physical properties (e.g., specific heat and density, mass and heat transport parameters) can be assumed to be independent of temperature.

Heat and mass transfer and chemical reaction can occur in series or simultaneously, as is the case for only mass transfer with chemical reaction. Both cases are discussed below.

7.2. Mass and Heat Transfer with Reaction in Series

Mass and heat transfer and chemical reaction in series are most relevant for chemical reactions in and on catalyst pellets. Here discussion is limited to heterogeneously catalyzed reactions in the gas phase. It is assumed that either a chemical reaction occurs at the outer surface of the pellet or that intra particle heat- and mass-transfer resistances are negligible. Furthermore, only monomolecular first-order kinetics are treated.

7.2.1. Particle Mass and Heat Balances

For first-order reactions A → products, carried out in or on a catalyst pellet, the overall mass and heat balances for the pellet yield a *mass balance* of:

$$k_g A_p \left(\overline{C_{A,g}} - C_{A,p}\right) = k_p (T_p) C_{A,p} V_p \tag{7.1}$$

and a *heat balance* of:

$$\alpha_g A_p \left(T_p - \overline{T_g}\right) = k_p (T_p) C_{A,p} (-\Delta H_r) V_p \tag{7.2}$$

Here $\overline{C_{A,g}}$ is the concentration of reactant A in the bulk of the gas phase and $C_{A,p}$ its concentration inside the pellet or at the pellet outer surface (since intra particle mass transfer is negligible, the two of them are equal); $\overline{T_g}$ is the temperature of the bulk of the gas and T_p the pellet temperature; $k_p(T_p)$ is the reaction rate constant at pellet temperature T_p; the reaction rate $k_p(T_p) C_{A,p}$ is defined as moles of A converted per cubic meter of catalyst volume per unit time; $(-\Delta H_r)$ is the reaction enthalpy (J/mol) and is positive for exothermic and negative for endothermic reactions; A_p and V_p are the outer surface area and the volume of the pellet, respectively; k_g is the *mass-transfer coefficient* of reactant A (m/s); and α_g is the *heat-transfer coefficient* of the pellet, (W m^{-2} K^{-1}). The heat-transfer coefficient α_g is defined in a manner similar to the mass-transfer coefficient k_g. The latter is the mole flux of A, J_A (moles of A per square meter of external surface area per unit time) divided by the driving force $\overline{C_{A,g}} - C_{A,p}$:

$$k_g = \frac{J_A}{\overline{C_{A,g}} - C_{A,p}} \tag{7.3}$$

Similarly, the heat-transfer coefficient α_g equals the heat flux from or to the pellet ψ'' (joules per square meter of external surface area per unit time) divided by the driving force $T_p - \overline{T_g}$:

$$\alpha_g = \frac{\psi''}{T_p - \overline{T_g}} \tag{7.4}$$

Both transfer coefficients are related by the *Chilton – Colburn analogy*:

$$\frac{\alpha_g}{k_g} = \varrho_g C_{p,g} Le^{2/3} \tag{7.5}$$

where ϱ_g (kg/m^3) is the density of the gas, $C_{p,g}$ (J kg^{-1} K^{-1}) its specific heat, and *Le* the *Lewis number* defined as

$$Le = \frac{a_g}{\mathbb{D}_{A,g}} \tag{7.6}$$

where a_g is the thermal diffusion coefficient of the gaseous bulk phase (m^2/s) and $\mathbb{D}_{A,g}$ (m^2/s) the molecular diffusion coefficient of A in the gas.

For convenience, the pellet concentration $C_{A,p}$ and the heat-transfer coefficient α_g are eliminated from Equations (7.1), (7.2), and (7.5). Then,

$$\frac{T_p - \overline{T_g}}{\Delta T_{ad}} = \left\{1 + \frac{k_g A_p}{k_p (T_p) V_p}\right\}^{-1} Le^{2/3} \tag{7.7}$$

with ΔT_{ad} the adiabatic temperature rise given by

$$\Delta T_{ad} = \frac{\overline{C_{A,g}} (\Delta H_r)}{\varrho_g C_{p,g}} \tag{7.8}$$

If the reaction rate is very fast compared to the rate of mass transfer, or $k_p(T_p) V_p >> k_g A_p$, the pellet operates in the *fast reaction regime* and its temperature reaches the maximum value:

$$T_{p,max} = \overline{T_g} + \Delta T_{ad} Le^{-2/3} \tag{7.9}$$

If the mass-transfer rate is much higher than the reaction rate, $k_p(T_p) V_p << k_g A_p$, the pellet operates in the *slow reaction regime* and its temperature equals the gas temperature. In that case the heat-transfer rate is also relatively fast according to the Chilton – Colburn analogy.

The following dimensionless numbers are also introduced for convencience:

$$\theta_p = \frac{T_p}{\Delta T_{ad}} Le^{2/3} \tag{7.10}$$

$$\theta_g = \frac{\overline{T_g}}{\Delta T_{ad}} Le^{2/3} \tag{7.11}$$

$$\xi(\theta_p) = \frac{k_g A_p}{k_p \left(\Delta T_{ad} Le^{-2/3} \theta_p\right) V_p} \tag{7.12}$$

Substitution of the above numbers in Equation (7.7) gives

$$\theta_p - \theta_g = \frac{1}{1 + \xi(\theta_p)} \tag{7.13}$$

Further, the assumption is made that the mass-transfer coefficient k_g is temperature independent and that the dependence of the reaction rate

constant $k_p(T_p)$ on temperature is of the Arrhenius type:

$$k_p(T_p) = k_{p,0} \exp\left(-\frac{E_a}{RT_p}\right) \tag{7.14}$$

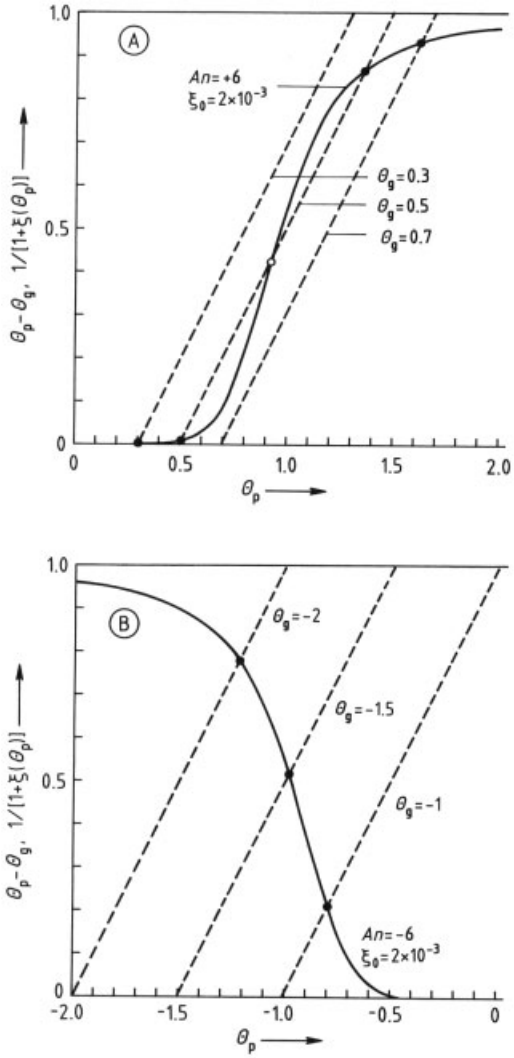

Figure 45. Relative heat production and heat withdrawal rates versus dimensionless pellet temperature ϑ_p
A) Exothermic reactions, θ_p, θ_g, $An > 0$; B) Endothermic reactions, θ_p, θ_g, $An < 0$
● = Stable operating point; ○ = Unstable operating point

Substitution in Equation (7.12) gives

$$\zeta(\vartheta_p) = \xi_0 \exp\left(\frac{An}{\vartheta_p}\right) \tag{7.15}$$

with

$$\xi_0 = \frac{k_g A_p}{k_{p,0} V_p} \tag{7.16}$$

An is a modified *Arrhenius number* defined as

$$An = \frac{E_a}{\Delta T_{ad} R} Le^{2/3} \tag{7.17}$$

An is the dimensionless ratio of the activation temperature E_a/R and the maximum temperature difference between the pellet and the gas (see Eq. 7.9). Substituting Equation (7.15) into Equation (7.13) yields

$$\theta_p - \theta_g = \left[1 + \xi_0 \exp\left(\frac{An}{\theta_p}\right)\right]^{-1} \tag{7.18}$$

From this equation, θ_p and thus T_p can be obtained for any given values of θ_g, ξ_0, and *An*. After calculation of $k_p(T_p)$, Equation (7.1) yields $C_{A,p}$, and hence the conversion rate $k_p(T_p)\, C_{A,p}$ is obtained.

7.2.2. Particle Multiplicity

The right-hand term of Equation (7.18) is a dimensionless pellet heat production rate (HPR) scaled between 0 and 1, being the ratio of the actual to the maximum possible heat production rate. The left-hand term can be regarded as a dimensionless pellet heat withdrawal rate (HWR).

If the left- and right-hand terms of Equation (7.18) are plotted separately versus θ_p, figures such as Figure 45 are obtained. The dimensionless HPR curve is sigmoidal because of the following: At low particle temperatures the system is in the *slow reaction regime;* hence the conversion rate equals the reaction rate. In this regime the conversion rate and thus the heat production rate are low. As pellet temperature increases, the conversion rate and thus HPR initially increase with temperature according to the assumed Arrhenius relation. If the pellet temperature becomes still higher, above a certain temperature the reaction rate is so high that mass transfer becomes rate limiting. The pellet then operates in the *fast reaction regime*, and the overall conversion rate equals the mass-transfer rate. Since mass transfer is virtually independent of temperature, in this regime the HPR does not increase further with increasing pellet temperature; it becomes constant.

The heat withdrawal rate equation is given by $\theta_p - \theta_g$, which is represented by a straight line with a slope of unity and an intercept with the θ_p-axis equal to θ_g.

For *exothermic reactions* ΔT_{ad} and thus also θ_p, θ_g, and *An* are positive. Depending on the parameter values, either one, two, or three operating points are obtained: that is, for the same values of θ_g, *An*, and ξ_0, one, two, or three values can be found for θ_p. This means that in Figure 45, one, two, or three intersection points between the HPR curve and the HWR line are obtained. The occurrence of more than one operating point for the same set of independent parameters is referred to as *multiplicity of steady states*.

For *endothermic reactions*, ΔT_{ad} is negative and hence θ_p, θ_g, and *An* are also. As a result, only a single operating point can be obtained (see Fig. 45), which means that for fixed values of θ_g, *An*, and ξ_0 the value of θ_p is also fixed. Thus for endothermic reactions, multiplicity can not occur.

Multiplicity of Steady States. If the heat withdrawal rate is represented by the line for $\theta_g = 0.3$ in Figure 45, the catalyst pellet is so effectively cooled by the surrounding gas that stable operation is possible only at low temperature and a low conversion rate. If the temperature of the gas is increased much further (e.g., as shown by the line for $\theta_g = 0.7$), cooling of the pellet is so ineffective that operation is possible only at high temperature and a relatively high conversion rate. For intermediate gas temperatures (e.g., the line given for $\theta_g = 0.5$) an HWR line is obtained that has three points of intersection with the heat production curve. For this situation, multiplicity occurs.

In case of multiplicity, of the three intersection points only those at the lowest and highest pellet temperature represent stable conditions: they are therefore called the *stable operating points*. The intermediate intersection point is referred to as an *unstable operating point*.

The stability of these points is determined by the slope of the HPR curve at the point of intersection with the HWR line. For the stable operating points its slope is less than 1, the slope of the HWR line. Therefore, if the temperature of the pellet increases somewhat, more heat is removed than is produced, which implies that the pellet temperature must decrease and thus return to the steady-state value. Similarly, a slightly too low pellet temperature causes more heat to be produced than removed, so the temperature increases. At the intermediate intersection point, on the other hand, the situation is reversed: the slope of the heat production curve is higher than unity. If the pellet temperature becomes slightly too high, more heat is produced than removed, and the pellet temperature increases further until the upper stable operating point is reached. Similarly, if the pellet temperature is a little too low, more heat is removed than produced and the temperature decreases further until the lower stable operating point is reached. So the intermediate operating point must be an unstable one.

In Figure 46, several heat production rate curves and one heat withdrawal rate line are plotted versus the dimensionless pellet temperature θ_p. The parameter values are $An = 3$ and $\theta_g = 0.3$, whereas ξ_0 is varied. Initially $\xi_0 = 0.1$, and the pellet conversion rate is low. Then the gas velocity is decreased. Subsequently the mass-transfer coefficient k_g decreases and thus ξ_0 (see Eq. 7.16). For $\xi_0 = 0.024$ a critical ξ_0 value is passed and the region of multiplicity begins. This multiplicity is illustrated in Figure 46 for $\xi_0 = 0.01$. The pellet remains at the lower stable operating point, until ξ_0 is decreased further to $\xi_0 = 0.0035$. Then the pellet is forced to leave the lower stable operating point and jumps to the higher stable operating point; this switch of operating points is sudden. The pellet is said to exhibit *runaway* or, alternatively, *ignition* of the pellet occurs. The pellet temperature in Figure 46 switches from $\theta_p = 0.34$ *to* $\theta_p = 1.26$. If ξ_0 is further decreased the pellet simply remains at the higher stable operating point, as illustrated for $\xi_0 = 0.001$.

If the velocity and thus ξ_0 are increased again, the pellet reenters the region of multiplicity for $\xi_0 = 0.0035$. In this case the pellet remains at the higher stable operating point until $\xi_0 = 0.024$ is reached. Then the pellet is forced to switch suddenly to the lower operating point. This phenomenon is referred to as *blowout*, or the reaction in the pellet is said to be *extinguished*. For the example given, the pellet temperature drops from $\theta_p = 0.86$ to $\theta_p = 0.30$.

For decreasing ξ_0 values a different line is found than for increasing ξ_0 values as shown in Figure 47. This is referred to as *hysteresis*. Hys-

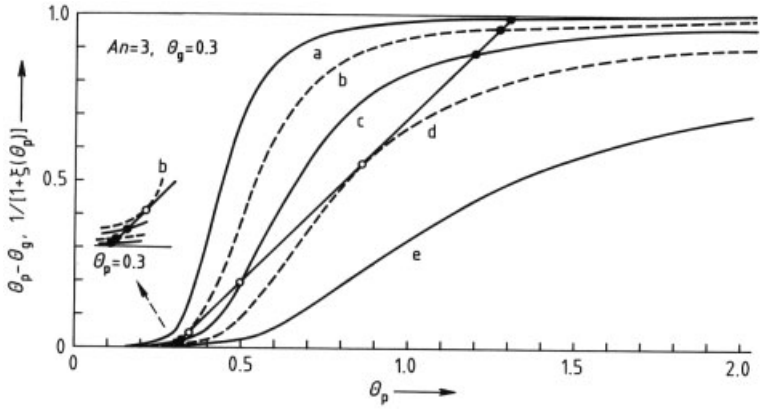

Figure 46. Relative heat production rate curves versus dimensionless pellet temperature θ_p for $An = 3$, $\theta_g = 0.3$, and several values of ξ_0
a) $\xi_0 = 0.001$; b) $\xi_0 = 0.0035$; c) $\xi_0 = 0.01$; d) $\xi_0 = 0.024$; e) $\xi_0 = 0.1$

teresis is commonly encountered when multiplicity of steady states is important.

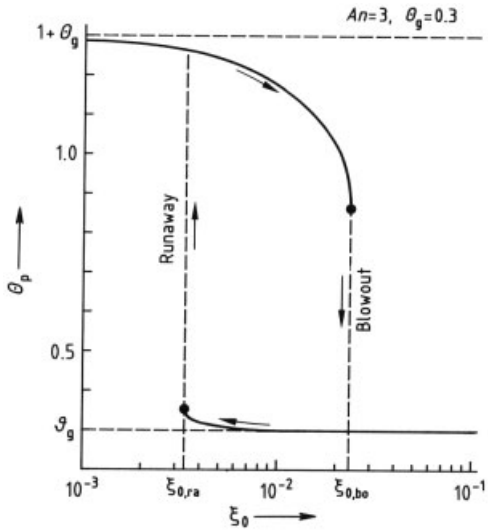

Figure 47. Dimensionless pellet temperature θ_p versus ξ_p for $An = 3$ and $\theta_g = 0.3$
Note the hysteresis, runaway, and blowout

The critical values of ξ_0 for runaway and blowout, $\xi_{0,\mathrm{ra}}$ and $\xi_{0,\mathrm{bo}}$, can be found by solving Equation (7.18) together with

$$\frac{\partial}{\partial \theta_p}\left[1+\xi_0 \exp\left(\frac{An}{\theta_p}\right)\right]^{-1} = 1 \tag{7.19}$$

for which case the HPR curve touches the HWR line. Solving Equations (7.18) and (7.19) and eliminating θ_p gives

$$\xi_{0,\mathrm{bo}} = \frac{2 + 2\theta_g + An - \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}}{An - 2\theta_g + \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}} \cdot \exp\left(-\frac{2An\,(1+An)}{An + 2\theta_g An + \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}}\right) \tag{7.20}$$

and

$$\xi_{0,\mathrm{ra}} = \frac{2 + 2\theta_g + An + \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}}{An - 2\theta_g - \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}} \cdot \exp\left(-\frac{2An\,(1+An)}{An + 2\theta_g An - \sqrt{An\left(An - 4\theta_g - 4\theta_g^2\right)}}\right) \tag{7.21}$$

The above equations are useful in the design of packed-bed reactors. For highly exothermic reactions, adiabatic temperature increases are on the order of 1000 – 2000 °C, which implies that at the higher operating point the catalyst temperature is so high that the catalyst would be destroyed. On the other hand, increasing the pellet size is economically attractive, so as to decrease the pressure drop over the packed bed and therefore the costs of compression. However, with increasing particle size the value of ξ_0 decreases and therefore the runaway limit given by Equation (7.21) is approached. Since runaway cannot be allowed, Equation (7.21) gives the maximum tolerable particle size for a given reaction, a given gas temperature, and a given gas velocity.

In Figure 48, $\xi_{0,\,\mathrm{ra}}$ is plotted versus θ_g for several values of *An*. A multiplicity and a uniqueness region can be distinguished. In this last region the slope in the point of inflection of the HPR curve is lower than that of the HWR line, which is 1, so multiplicity can never occur. To find the boundary line mentioned, Equation (7.18) must be solved together with Equations (7.19), and

$$\frac{\partial^2}{\partial\theta_\mathrm{p}^2}\left[1+\xi_0\exp\left(\frac{An}{\theta_\mathrm{p}}\right)\right]^{-1}=0 \tag{7.22}$$

which states that the derivative of the HPR curve is unity at the inflection point. Solving Equations (7.18), (7.19), and (7.22) and eliminating θ_p and *An* give

$$\xi_0=\frac{1+\theta_\mathrm{g}}{\theta_\mathrm{g}}\,\mathrm{e}^{-2(1+\theta_\mathrm{g})} \tag{7.23}$$

This is the boundary line equation giving the dashed line in Figure 48.

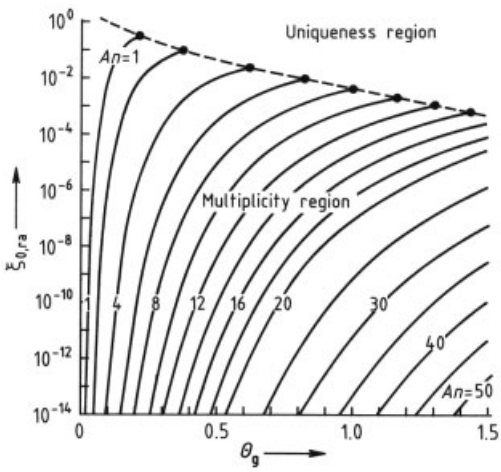

Figure 48. Value of $\xi_{0,\,\mathrm{ra}}$ for which runaway occurs versus dimensionless gas temperature θ_g

All pellets in a reactor cannot always be in the uniqueness region. Then, one must ensure that all pellets work at the lower operating point (see Fig. 47). As WESTERTERP et al. have shown, $k_\mathrm{g}\approx\sqrt{u/d_\mathrm{p}}$ with d_p, the particle diameter equivalent to a sphere, $d_\mathrm{p}=6\,V_\mathrm{p}/A_\mathrm{p}$. Thus, $\xi_{0,\,\mathrm{ra}}\sim u^{0.5}\,d_\mathrm{p}^{-1.5}$ [8]. Once the particle diameter has been fixed and safe operating conditions have been chosen for the design (e.g., in Fig. 48, the point given by $\theta_\mathrm{g}=0.3$ and $\xi_0=0.07$, see also Figs. 46 and 47), the diagram can be used further to determine operating adjustments. For example, lowering the throughput to 25 % of its design value would surely lead to (near) runaway, because now

$$\xi_{0,\,\mathrm{new}}=\xi_{0,\,\mathrm{old}}\cdot\sqrt{u_\mathrm{new}/u_\mathrm{old}}=0.5\xi_{0,\,\mathrm{old}}$$
$$=0.035.$$

In the first instance, θ_g should be lowered to prevent pellet runaway. This may lead to too low a temperature and hence too low a conversion. In this case the reactant concentration in the feed may be decreased to increase *An*, so that θ_g can be increased again to achieve sufficiently high reaction rates.

This reasoning also indicates the importance of gas flow rates. If, for example, in a tube of a multitubular reactor, a deposit is formed that decreases the gas velocity because of plugging, at some point pellet runaway will occur irreparably and, consequently, full runaway of that tube [57].

Multiplicity studies for other reaction kinetics proceed along the same lines, albeit simple analytical solutions often cannot be obtained, so numerical approximations must be used.

7.3. Mass and Heat Transfer with Simultaneous Reaction

Due to the heat effects of the reaction, the temperature in the reaction phase (e.g., a catalyst pellet) is not necessarily uniform. Mass and heat transfer with simultaneous reaction therefore must also be studied.

The effect of temperature gradients in the reaction phase can be incorporated readily into the Thiele modulus or Hatta number. In what follows, this will be derived for a porous pellet with intra particle temperature gradients. Extensions to gas – liquid reactions and Hatta numbers can, of course, be made; if, in the Thiele modulus, the characteristic dimension $V_\mathrm{p}/A_\mathrm{p}$ is replaced by $\mathbb{D}_\mathrm{A}/k_{\mathrm{l,\,A}}$ and the effective diffusion coefficient $\mathbb{D}_{\mathrm{i,\,A}}$ by $\mathbb{D}_\mathrm{A}$, the corresponding Hatta numbers are obtained. To simplify the Thiele modulus equations, the Frank-Kamenetzki approximation will be discussed first.

7.3.1. Frank – Kamenetzki Approximation

Consider a monomolecular reaction $A \longrightarrow products$ inside a porous catalyst pellet. By combining microscale mass and heat balances, the relation between the local temperature inside the pellet and the local reactant concentration can be found:

$$\frac{T-T_{\mathrm{i}}}{T_{\mathrm{i}}}=\mu\frac{C_{\mathrm{A,i}}-C_{\mathrm{A}}}{C_{\mathrm{A,i}}} \tag{7.24}$$

with

$$\mu=\frac{(-\Delta H_{\mathrm{r}})\,\mathbb{D}_{\mathrm{i,A}}\,C_{\mathrm{A,i}}}{\lambda_{\mathrm{p}}\,T_{\mathrm{i}}} \tag{7.25}$$

where T is the local temperature and C_{A} the local reactant concentration inside the pellet; T_{i} and $C_{\mathrm{A,i}}$ are the same quantities at the pellet external surface; $(-\Delta H_{\mathrm{r}})$ is the reaction enthalpy (J/mol); $\mathbb{D}_{\mathrm{i,A}}$ the effective diffusion coefficient (m^2/s); and λ_{p} the thermal conductivity of the pellet ($\mathrm{W\,m^{-1}\,K^{-1}}$). Notice that μ can be both positive [for $(-\Delta H_{\mathrm{r}})>0$, thus exothermic reactions] and negative [for $(-\Delta H_{\mathrm{r}})<0$, thus endothermic reactions].

The most extreme temperature T_{ext} is found for $C_{\mathrm{A}}=0$ (i.e., in the center of the pellet and for low effectiveness factors). This extreme temperature can be obtained from

$$\mu=\frac{T-T_{\mathrm{ext}}}{T_{\mathrm{ext}}} \tag{7.26}$$

If the temperature dependence of a reaction rate constant k or adsorption constant K is of the Arrhenius type, according to the above equation,

$$\frac{k\,(T_{\mathrm{i}})}{k\,(T)}=\frac{K\,(T_{\mathrm{i}})}{K\,(T)}=\exp\left[\chi\frac{1-C_{\mathrm{A}}/C_{\mathrm{A,i}}}{1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)}\right] \tag{7.27}$$

with

$$\chi=\frac{E_{\mathrm{a}}}{RT_{\mathrm{i}}}\,\frac{(-\Delta H_{\mathrm{r}})\,\mathbb{D}_{\mathrm{i,A}}C_{\mathrm{A,i}}}{\lambda_{\mathrm{p}}\,T_{\mathrm{i}}} \tag{7.28}$$

where $k\,(T_{\mathrm{i}})$ is the reaction rate constant and $K\,(T_{\mathrm{i}})$ the adsorption constant at the outer surface temperature. Like μ, χ can also have both positive and negative values.

Equation (7.27) can be simplified if the following assumption holds:

$$\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)<<1 \tag{7.29}$$

then,

$$\frac{k\,(T_{\mathrm{i}})}{k\,(T)}=\frac{K\,(T_{\mathrm{i}})}{K\,(T)}=\exp\left[\chi\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right] \tag{7.30}$$

Table 1. Comparison of χ, μ, and Δ_{FK} values for typical industrial processes [58]

Process	χ	μ	Δ_{FK}, %
Ethylene hydrogenation	0.038	0.0029	0.01
Benzene hydrogenation	0.48	0.030	1
Methane oxidation	0.12	0.0056	0.07
Ethylene oxidation	0.92	0.068	6
Naphthalene oxidation	0.33	0.015	0.5
Acrylonitrile synthesis	0.018	0.0019	0.003

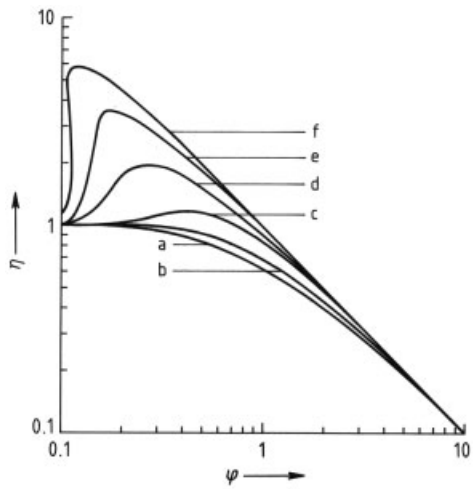

Figure 49. Effectiveness factor η versus Thiele modulus φ for first-order kinetics in an infinite slab and with significant intra particle temperature gradients (several values of χ have been assumed
a) $\chi=-\infty$; b) $\chi=0$; c) $\chi=+2$; d) $\chi=+4$; e) $\chi=+6$; f) $\chi=+8$

Equation (7.29) is generally referred to as the *Frank – Kamenetzki approximation*. Basically this assumption means that the Arrhenius-type dependence of reaction rates and adsorption constants can be approximated by an exponential relationship.

Comparison of Equations (7.27) and (7.30) shows that the error Δ_{FK} introduced by using the Frank – Kamenetzki approximation equals

$$\Delta_{\mathrm{FK}} = \left| \frac{\exp\left[\chi \frac{1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}}{1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)}\right] - \exp\left[\chi\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)\right]}{\exp\left[\chi \frac{1-C_{\mathrm{A}}/C_{\mathrm{A,i}}}{1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)}\right]} \right|$$

$$= \exp\left[\mu\chi \frac{\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)^2}{1+\mu\left(1 - C_{\mathrm{A}}/C_{\mathrm{A,i}}\right)}\right] - 1 \tag{7.31}$$

The largest value of Δ_{FK} is found for $C_{\mathrm{A}} = 0$; hence for low effectiveness factors and in the pellet center,

$$\Delta_{\mathrm{FK,\,max}} = \exp\left(\chi \frac{\mu}{1+\mu}\right) - 1 \tag{7.32}$$

In Table 1, values are given for χ, μ, and $\Delta_{\mathrm{FK,\,max}}$ for some typical industrial processes [58]. $\Delta_{\mathrm{FK,\,max}}$ is generally very small and $< 10\,\%$ for all cases. Hence, the Frank – Kamenetzki approximation can be used without introducing significant error.

7.3.2. Thiele Moduli and Hatta Numbers

The temperature inside the catalyst pellet can be related to the concentration of reactant A as (see Eq. 7.24)

$$T = T_{\mathrm{i}}\left[1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,\,i}}\right)\right] \tag{7.33}$$

Hence the reaction rate, which generally is a function of both the temperature T and the concentration C_{A}, can be written as a function of the concentration only:

$$|\mathbb{R}\left(C_{\mathrm{A}}, T\right)| = |\mathbb{R}\left\{C_{\mathrm{A}}, T_{\mathrm{i}}\left[1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,\,i}}\right)\right]\right\}| \tag{7.34}$$

Consequently, a general definition of the Thiele modulus for monomolecular reactions with arbitrary kinetics and significant intra particle temperature gradients follows from (see Eq. 6.123):

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}} |\mathbb{R}_{\mathrm{A}}\left(C_{\mathrm{A,\,i}}, T_{\mathrm{i}}\right)|_{\mathrm{p}} \left(2\mathbb{D}_{\mathrm{i,\,A}} \int_0^{C_{\mathrm{A,\,i}}} |\mathbb{R}_{\mathrm{A}} \cdot \left\{C_{\mathrm{A}}, T_{\mathrm{i}}\left[1+\mu\left(1-C_{\mathrm{A}}/C_{\mathrm{A,\,i}}\right)\right]\right\}|_{\mathrm{p}} \partial C_{\mathrm{A}}\right)^{-\frac{1}{2}} \tag{7.35}$$

In solving the above equations for any form of kinetics, use of the Frank – Kamenetzki approximation for the reaction rate and adsorption constants is advisable (i.e., assume an exponential instead of an Arrhenius-type temperature dependence). For example, for n th-order kinetics, use of Equation (6.191)

$$|\mathbb{R}\left(C_{\mathrm{A}}\right)| = kC_{\mathrm{A}}^n$$

with the Frank – Kamenetzki approximation Equation (7.35) yields

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}} \int \sqrt{\frac{k\left(T_{\mathrm{i}}\right) C_{\mathrm{A,\,i}}^{n-1}}{2\mathbb{D}_{\mathrm{i,\,A}}} \left(\mathrm{e}_0^{\chi 1} u^n \mathrm{e}^{-\chi u}\, \partial u\right)^{-1}} \tag{7.36}$$

where u is a dummy variable. For first-order ractions, this simplifies to

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}} \sqrt{\frac{k\left(T_{\mathrm{i}}\right)}{2\mathbb{D}_{\mathrm{i,\,A}}} \frac{\chi^2}{\mathrm{e}^{\chi} - 1 - \chi}} \tag{7.37}$$

In turn, for $\chi \to 0$, this simplifies to

$$\varphi = \frac{V_{\mathrm{p}}}{A_{\mathrm{p}}} \sqrt{\frac{k\left(T_{\mathrm{i}}\right)}{\mathbb{D}_{\mathrm{i,\,A}}}} \tag{7.38}$$

which is Aris' shape-generalized Thiele modulus for monomolecular first-order kinetics and insignificant intra particle temperature gradients.

In Figure 49 the effectiveness factor η is plotted against the Thiele modulus φ for first-order kinetics in an infinite slab and for varying values of χ. For $\chi = 0$ the curve discussed in Section 6.2.2.2 is obtained; for $\chi > 0$ (exothermic reactions) the effectiveness factor increases, and for $\chi < 0$ (endothermic reactions), η decreases. For high values of χ the effectiveness factor may even become larger than unity. In this case the temperature profiles inside the catalyst pellet are so pronounced that they can even compensate for the decreasing concentration inside the pellet. For very high values of χ, finally, even a region of multiplicity may be entered: for one value of the Thiele modulus, several values of the effectiveness factor can be obtained (in Fig. 49 for $\chi = 8$). This multiplicity results from the fact that

multiple solutions can be found if the mass and heat balances are solved simultaneously, as illustrated in Section 7.2 for mass and heat transfer and reaction in series.

To estimate the effectiveness factor from the Thiele modulus, one could again use the approximation (Eq. 6.116):

$$\eta = \frac{\tanh(\varphi)}{\varphi}$$

As shown in Figure 49 this approximation does not hold for high χ values ($\chi > 1$) and intermediate φ values. However, for mildly exothermic and all endothermic reactions the above approximation gives a fair estimate, usually within 10 % accuracy.

All the above formulas and conclusions also hold for systems in which Hatta numbers must be applied instead of Thiele moduli, as discussed above. For example, for monomolecular first-order reactions with mass and heat transfer with simultaneous reaction, the Hatta number becomes (see Eq. 7.37):

$$Ha = \frac{\mathbb{D}_A}{k_{l,A}} \sqrt{\frac{k(T_i)}{2\mathbb{D}_A} \frac{\chi^2}{e^{\chi} - 1 - \chi}} \tag{7.39}$$

It can be seen from the χ values in Table 1 that, in general, intrareaction-phase temperature gradients can be neglected, and the corrections for the Thiele moduli and Hatta numbers need not be applied.

8. References

1. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* Wiley, New York 1960.
2. W. J. Beek, M. K. Muttzall: *Transport Phenomena,* Wiley, New York 1975.
3. G. Damköhler: *Einfluss von Diffusion, Strömung und Wärmetransport auf die Ausbeute bei chemischtechnischen Reaktionen,* VDI Verlag, Leverkusen 1957.
4. D. A. Frank-Kamenetzki: *Diffusion and Heat Transfer in Chemical Kinetics,* 2nd ed., Plenum, New York 1969.
5. D. W. van Krevelen, *Chem. Eng. Sci.* **8** (1958) 5.
6. P. V. Danckwerts: *Gas – Liquid Reactions,* McGraw-Hill, London 1970.
7. G. Astarita: *Mass Transfer with Chemical Reaction,* Elsevier, Amsterdam 1967.
8. K. R. Westerterp, W. P. M. van Swaaij, A. A. C. M. Beenackers: *Chemical Reactor Design and Operation,* Wiley, Chichester 1987.
9. H. Gerrens, *Chem. React. Eng. Proc. Int. Symp.* **4th,** (1976) 421.
10. J. Villermaux, *Chem. React. Eng. Proc. Int. Symp.* **8th,** (1982) 545 – 554.
11. L. Michaelis, M. M. L. Menten, *Biochem. Z.* **49** (1913) 333.
12. J. Monod, *Annu. Rev. Microbiol.* **3** (1949) 371.
13. A. E. Humphrey, *ACS Symp. Ser.* **72** (1978) 262.
14. A. N. Kolmogoroff, *Dokl. Akad. Nauk SSSR* **30** (1941) 301.
15. A. N. Kolmogoroff, *Dokl. Akad. Nauk SSSR* **32** (1941) 16.
16. G. K. J. Batchelor, *J. Fluid Mech.* **5** (1959) 113.
17. S. Corrsin, *AIChE J.* **10** (1964) 87.
18. R. E. Rosenweig, *AIChE J.* **10** (1964) 92.
19. O. Levenspiel: *Chemical Reaction Engineering,* 2nd ed., Wiley, New York 1972.
20. P. V. Danckwerts, *Chem. Eng. Sci.* **2** (1953) 1.
21. T. N. Zwietering, *Chem. Eng. Sci.* **11** (1959) 1.
22. H. Weinstein, R. J. Adler, *Chem. Eng. Sci.* **22** (1967) 65.
23. J. Villermaux, A. Zoulalian, *Chem. Eng. Sci.* **24** (1669) 1513.
24. E. A. Mason, E. A. P. Malinauskas: *Gas Transport in Porous Media: the Dusty-Gas Model,* Elsevier, Amsterdam 1983.
25. R. Higbie, *Trans. Am. Inst. Chem. Eng.* **31** (1935) 365.
26. P. V. Danckwerts, *Ind. Eng. Chem.* **43** (1951) 1460.
27. H. J. Den Hartog, W. J. Beek, *Appl. Sci. Res.* **19** (1968) 311.
28. E. Sada, H. Kumazawa, M. A. Butt, T. Sumi, *Chem. Eng. Sci.* **32** (1977) 970.
29. E. Sada, H. Kumazawa, M. A. Butt, *Chem. Eng. Sci.* **32** (1977) 972.
30. A. Wheeler: *Advances in Catalysis,* **vol. III,** Academic Press, London 1951, p. 249.
31. D. K. Schmalzer, *Chem. Eng. Sci.* **13** (1969) 226.
32. N. Orrs, T. Dogu, *AIChE J.* **25** (1979) 723.
33. B. D. Kulkarni, V. K. Jayaraman, L. K. Doraiswamy, *Chem. Eng. Sci.* **36** (1981) 943.
34. E. W. Thiele, *Ind. Eng. Chem.* **31** (1939) 916.
35. R. Aris, *Chem. Eng. Sci.* **6** (1957) 262.
36. K. B. Bischoff, *Chem. Eng. Sci.* **29** (1974) 1348.

37. N. G. Karanth, H. P. Koh, R. Hughes, *Chem. Eng. Sci.* **29** (1974) 451.
38. P. V. Danckwerts, *Trans. Faraday Soc.* **46** (1950) 300.
39. K. R. Westerterp, *Chem. Eng. Sci.* **18** (1963) 495.
40. J. Bridgewater, *Chem. Eng. Sci.* **22** (1967) 185.
41. E. N. Lightfoot, *Chem. Eng. Sci.* **17** (1962) 1007.
42. G. Baldi, V. Specchia, *Chem. Eng. J. (Lausanne)* **11** (1976) 81.
43. J. B. Stepanek, R. V. Shilimkan, *Trans. Inst. Chem. Eng.* **52** (1974) 313.
44. D. W. van Krevelen, P. J. Hoftijzer, *Chim. Ind. XXI ème Congrès Int. Chim. Ind.,* (1948) p. 168.
45. M. De Santiago, I. H. Farina, *Chem. Eng. Sci.* **25** (1970) 744.
46. P. L. T. Brian, J. F. Hurley, E. A. Hasseltine, *AIChE J.* **7** (1961) 226.
47. H. Hikita, S. Asai, *Int. Chem. Eng.* **4** (1964) 332.
48. W. J. De Coursey, *Chem. Eng. Sci.* **29** (1974) 1867.
49. K. Onda, E. Sada, T. Kobayashi, M. Fujine, *Chem. Eng. Sci.* **25** (1970) 761.
50. K. Onda, E. Sada, T. Kobayashi, M. Fujine, *Chem. Eng. Sci.* **27** (1972) 247.
51. K. Onda, H. Takeuchi, M. Fujine, *Chem. Eng. Sci.* **28** (1973) 1637.
52. P. L. T. Brian, M. C. Beaverstock, *Chem. Eng. Sci.* **20** (1965) 47.
53. P. A. Ramachandran, *Chem. Eng. Sci.* **27** (1972) 1807.
54. L. A. Goettler, R. L. Pigford, *AIChE. J.* **17** (1971) 793.
55. S. L. Shkylar, Y. V. Aksel'rod, Y. A. Sokolinskii, *Theor. Found. Chem. Eng.* (*Engl. Transl.)* **10** (1976) 480.
56. P. A. Ramachandran, M. M. Sharma, *Chem. Eng. Sci.* **25** (1970) 1743.
57. R. J. Wijngaarden, K. R. Westerterp: "Incorporation of statistical distributions of particle properties in chemical reactor design and operation. The cooled tubular reactor," *Chem. Eng. Sci.* (1991), in press.
58. V. Hlaváček, M. Kubiček, *Chem. Eng. Sci.* **25** (1970) 1537.

Printed Circuits → Imaging Technology

Reactor Types and Their Industrial Applications

KLAUS-DIETER HENKEL, Buna AG, Schkopau, Federal Republic of Germany

1. Introduction 1
2. Basic Types of Reactors 2
3. Survey of Real Reactors and Their Uses 4
3.1. Reactors for Gas-Phase Reactions . 4
3.2. Reactors for Liquid-Phase Reactions 8
3.3. Reactors for Gas – Liquid Reactions 8
3.4. Reactors for Solid-Catalyzed Reactions 13
3.4.1. Reactors for Heterogeneous Gas Catalysis 13
3.4.2. Reactors for Liquid-Phase and Gas – Liquid Reactions over Solid Catalysts 13
3.5. Reactors for Noncatalytic Reactions Involving Solids 19
3.5.1. Reactors for Noncatalytic Gas – Solid Reactions 19
3.5.2. Reactors for Noncatalytic Liquid – Solid Reactions 21
3.5.3. Reactors for Noncatalytic Solid-Phase Reactions 21
3.6. Electrothermal Reactors 21
3.7. Reactors for Electrochemical Processes 24
3.8. Reactors for Biochemical Processes 27
3.9. Reactors for Photochemical and Radiochemical Processes 28
3.9.1. Photochemical Reactors 28
3.9.2. Radiochemical Reactors 32
4. References 33

1. Introduction

The reactor in which the chemical reaction takes place occupies a central position in the chemical process. Grouped around the reactor are the process steps involving physical treatment of the material streams, such as conveyance, heat transfer, and separation and mixing operations. The reactor provides the volume necessary for the reaction and holds the amount of catalyst required for the reaction. The energy required to overcome the activation threshold of each partial reaction is also supplied in the reactor, and the proper temperature and concentration are maintained.

The most important *reaction-related factors* for the design of a reactor are

1) The activation principle selected, together with the states of aggregation of the reactants and the resulting number and types of phases involved
2) The concentration and temperature dependence of the chemical reactions
3) The heat of the reactions taking place

The most important *activation principles* for a reaction mixture include

1) Activation by addition of heat
2) Catalytic activation
3) Activation by decomposition of an initiator
4) Electrochemical activation
5) Biochemical activation

Less important options for activation are visible or ultraviolet light and radioactive radiation.

With regard to *phase relationships* in the reaction space, a number of combinations are possible. The reactants and reaction products can be present, or be produced, in various states of aggregation. Furthermore, inert diluents or heat-transfer media can be present in different phases. Finally, the catalyst, which is generally in the solid or liquid phase, often has to be taken into consideration.

The (negative or positive) heat of the reactions taking place in a reactor influences the extent and nature of *provisions for heat transfer*. Exothermic or endothermic reactions frequently require supply or removal of large quantities of heat. Thermally neutral reactions involve considerably less technical sophistication.

The *concentration and temperature dependences* of a chemical reaction are described by the reaction rate. In practice most reaction systems are complex and include parallel, sequential, and equilibrium reactions. To obtain the highest possible yield of desired product under

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_087

these conditions, the temperature and pressure must be held within certain ranges, the temperature must be controlled along the reaction path, and a definite residence-time distribution in the reactor must be achieved. If, in addition, substances or energy have to be transferred from one phase to another, appropriate transport conditions have to be implemented. When catalysts are used, catalyst loss due to aging and poisoning must be considered. These factors impose a complex of requirements that must be kept in mind when designing a reactor.

Against the requirements established by the process, the designer must balance costs of fabrication, consumption of materials, and operational reliability. In practice, many possibilities are often available for realizing a chemical process, and in such cases the decision must depend on an assessment of the overall process as well as commercial constraints on the plant.

2. Basic Types of Reactors (→ Model Reactors and Their Design Equations)

A variety of reactor designs are used in industry, but all of them can be assigned to certain basic types or combinations of these. The basic types are as follows (see → Principles of Chemical Reaction Engineering, Chap. 4.2.):

1) Batch stirred-tank reactor
2) Continuous stirred-tank reactor
3) Tubular reactor

Given certain flow and thermal conditions, these types are also referred to as "ideal" reactors. With respect to *flow conditions* the ideal stirred-tank batch reactor is characterized by complete mixing on microscopic and macroscopic scales. In the ideal tubular reactor, plug flow is assumed, i. e., no mixing occurs in axial (flow) direction, but ideal mixing takes place in the ra-dial direction. Thus, as in the batch stirred-tank reactor, all particles experience a well-defined residence time. In contrast, the continuous stirred-tank reactor has a very broad residence-time distribution (→ Principles of Chemical Reaction Engineering, Chap. 4.2.1.). The ideal analysis is based on the assumption of a reaction system that is homogeneous as regards the phase. Thus transport resistance between phases does not occur.

The *thermally ideal operating states* are the isothermal and adiabatic states, i. e., either very intensive heat exchange with the surroundings or no exchange at all is assumed.

In practical operation, the ideal states are achieved only approximately. Examples of typical nonidealities include

1) The formation of real flow patterns, such as dead zones, short-circuit flows, and channeling
2) Transport processes in the individual phases, such as axial backmixing
3) The formation of concentration and temperature profiles as a result of transport resistances in and between phases
4) Segregation processes
5) Incomplete mixing of reactants

The essential advantages and disadvantages of the three basic reactor types are discussed in what follows.

Batch Stirred Tank (→ Stirred-Tank and Loop Reactors)

Principal Applications:

1) Liquid-phase reactions
2) Liquid – solid reactions

Advantages:

1) Quick production changeover possible; use for substances produced on a small scale
2) Process steps upstream or downstream of the reaction can also be performed in the reactor
3) Better process control than in continuous operation when solid or highly viscous phases form or are present
4) Well-defined residence time

Disadvantages :

1) Relatively high operating costs due to long downtimes and high manpower requirements
2) Quality differences between charges because reaction conditions are only partly reproducible
3) Limited temperature control capabilities, especially with highly endothermic or exothermic reactions

Continuous Stirred Tank
Principal Applications:

1) Liquid-phase reactions
2) Gas – liquid reactions
3) Gas – liquid reactions over suspended catalysts

Advantages:

1) Low operating costs, especially at high throughputs
2) Consistent product quality due to reproducible process control
3) Wide range of throughput

Disadvantages:

1) Final conversions lower than in other basic reactor types because of complete mixing (i.e., unreacted starting materials can get into the product stream)
2) High investment costs to implement continuous operation
3) Changeover to other products generally complex and time-consuming because of reaction-specific design

Tubular Reactor (→ Tubular Reactors)
Principal Applications:

1) Homogeneous gas-phase reactions
2) Liquid-phase reactions
3) Gas- and liquid-phase reactions over solid catalysts (→ Fixed-Bed Reactors)
4) Gas – liquid reactions

Advantages:

1) Favorable conditions for temperature control by heat supply or removal
2) No moving mechanical parts, hence especially suitable for high-pressure service

Disadvantages:

1) Very high degree of specialization, often with complicated design and high investment costs
2) Relatively large pressure drops

Reactors are interconnected to make up for the drawbacks of a single reactor, especially to adapt reaction conditions during scale-up capacity, as well as to optimize conversion and yield. Partial reactors can be combined in a single apparatus or connected in a system of reactors; these partial reactors may differ in shape and size.

Types of interconnections are series, parallel, and recycle.

Series Connection:

1) Multibed reactors
2) Tower reactors, reaction columns
3) Cascades of stirred tanks (→ Stirred-Tank and Loop Reactors)
4) Multiple-hearth reactors (→ Metallurgical Furnaces, Chap. 2.)
5) Different reactor types connected in series (e.g., stirred tank and tubular reactor)

Parallel Connection: Multitubular reactors
Recycle Connection: Loop reactors (→ Stirred-Tank and Loop Reactors)

Complicated reactor designs result, especially when different reactor types are combined in a single apparatus. At the same time, such a combination offers maximum adaptability to the requirements of a given reaction process. The designer must, of course, examine every case individually to ensure that the results justify the very high development and investment costs for such special reactors. The following survey of real reactors includes these special types of reactor designs only when their utility extends beyond a single case.

3. Survey of Real Reactors and Their Uses

The phase relationships in the reaction space are crucial in the design of reactors for *catalytic, thermal, and polymerization processes* and accordingly form the top-level classification feature for such reactors. Since many different combinations of phases are possible, the survey is based only on the state of the reactants at the inlet to the reactor or the beginning of the reaction and the phase of the reaction site (catalyst phase, liquid phase with dissolved reactant). Reaction products that form additional phases and inert substances of all types (except for solvents, as just noted) are ignored.

Reactors used in *electrothermal, electrochemical, biochemical, photochemical,* and *radiochemical* processes are treated separately. Reactor types for which no industrial application is currently known are not listed.

3.1. Reactors for Gas-Phase Reactions

Homogeneous gas-phase reactions utilized in industry are generally characterized by large positive or negative enthalpies of reaction and high reaction temperatures. To obtain the desired product spectrum, residence times must usually be very short. The high reaction temperature can be maintained or the requisite heat supplied by burning part of the feed.

Tables 1 and 2 and Figures 1 and 2 summarize the reactors used for such reactions as well as their applications.

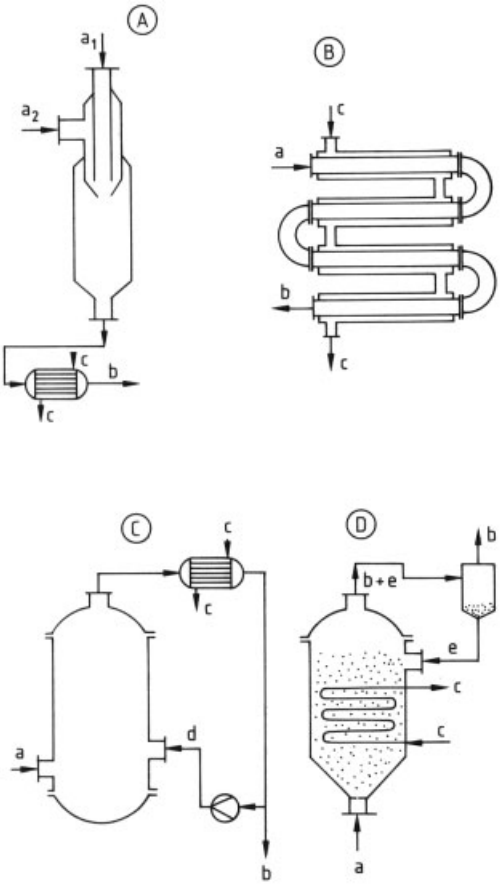

Figure 1. Reactors for exothermic gas-phase reactions
A) Burner; B) Tubular reactor; C) Reactor with recycle; D) Fluidized-bed reactor
a) Gaseous reaction mixture; a_1, a_2) Gaseous feed components; b) Gaseous product; c) Coolant; d) Partial stream of product; e) Catalyst

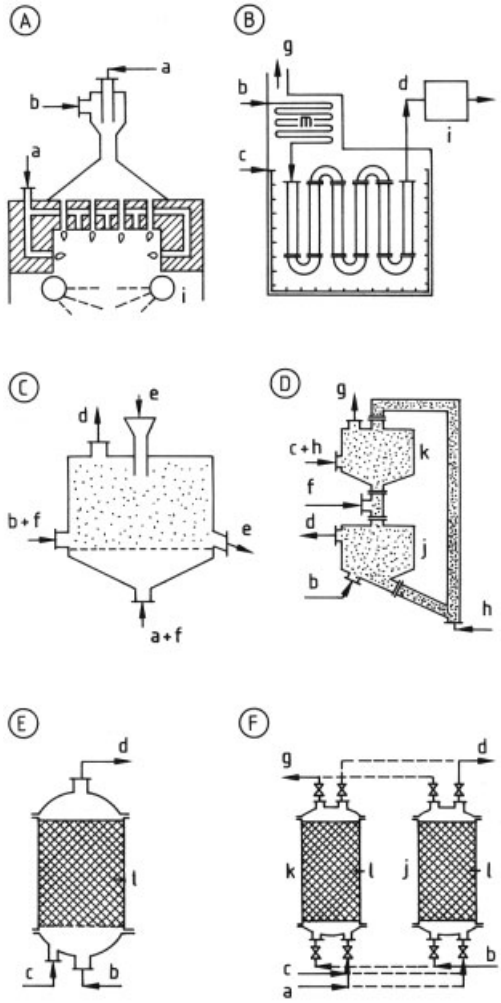

Figure 2. Reactors for endothermic gas-phase reactions
A) Burner; B) Reformer; C) Fluidized-bed reactor; D) Moving-bed reactor; E) Reactor with fixed bed of inerts; F) Regenerative furnaces
a) Oxygen or air; b) Hydrocarbon; c) Fuel gas; d) Product; e) Heat-transfer medium; f) Steam; g) Flue gas; h) Air; i) Quench; j) Reaction section; k) Regeneration section; l) Catalyst; m) Convection zone

3.2. Reactors for Liquid-Phase Reactions

In general, liquid-phase reactions are exothermic. In the case of multiphase systems, intensive mass and heat transfer must be provided for; this is possible only in reactors with compulsory mixing, such as stirred tanks. Along with a num-

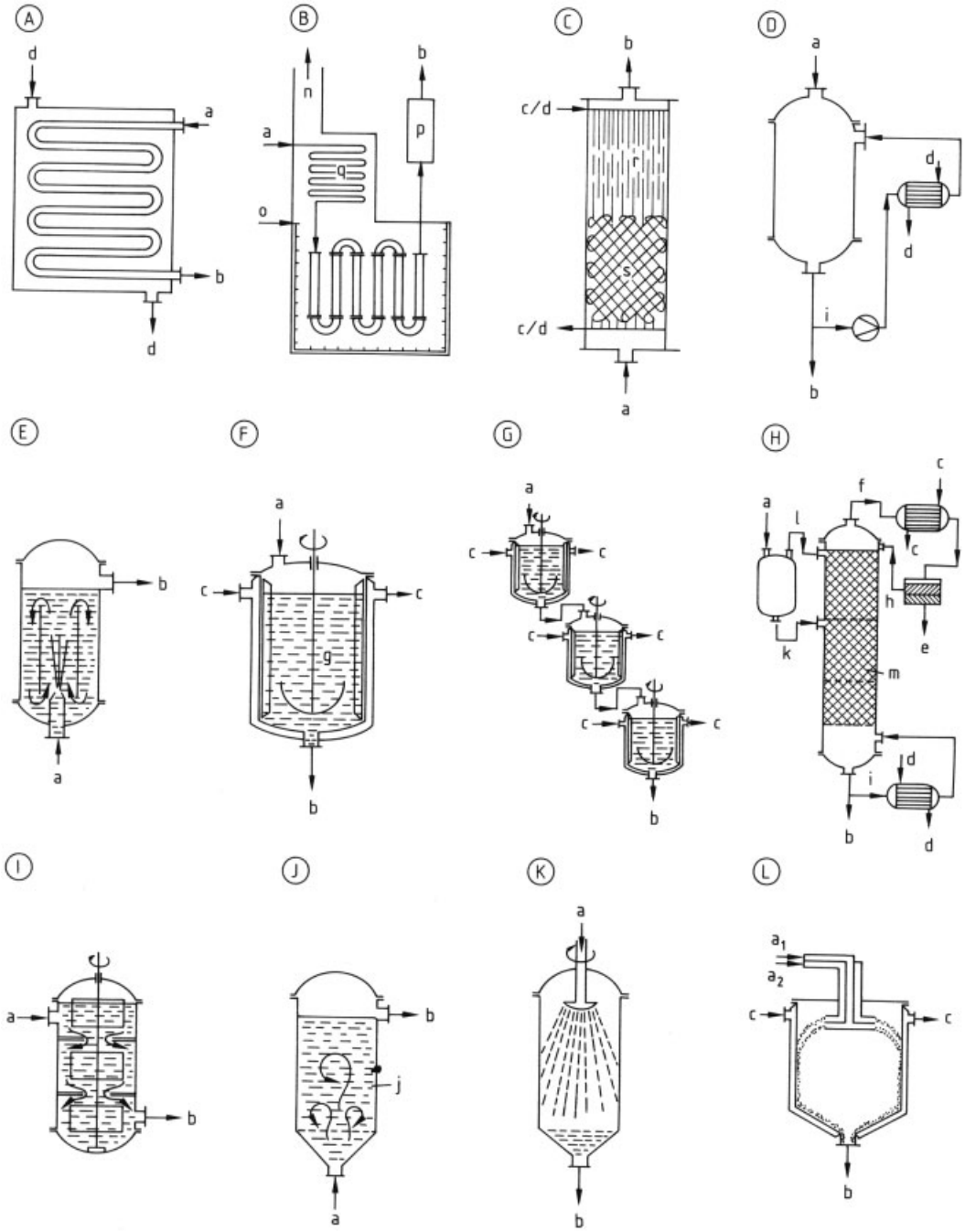

Figure 3. Reactors for liquid-phase reactions
A) Tubular reactor; B) Reformer; C) Sulzer mixer – reactor; D) Reactor with external recirculation; E) Reactor with internal recirculation (draft tube); F) Stirred tank; G) Cascade of stirred tanks; H) Column reactor; I) Multichamber tank; J) Fluidized-bed reactor; K) Spray reactor; L) Falling-film reactor
a) Liquid reaction mixture; a_1, a_2) Liquid feed components; b) Liquid product; c) Coolant; d) Heating agent; e) Water; f) Organic phase and water; g) Baffle; h) Organic phase; i) Partial stream of product; j) Catalyst; k) Reaction mixture from preceding reaction stage; l) Water from preceding stage; m) Packing; n) Off-gas; o) Fuel gas for burners; p) Quench; q) Convection zone; r) Mixing element consisting of tubes carrying heat-transfer medium; s) Mixing elements rotated 90°

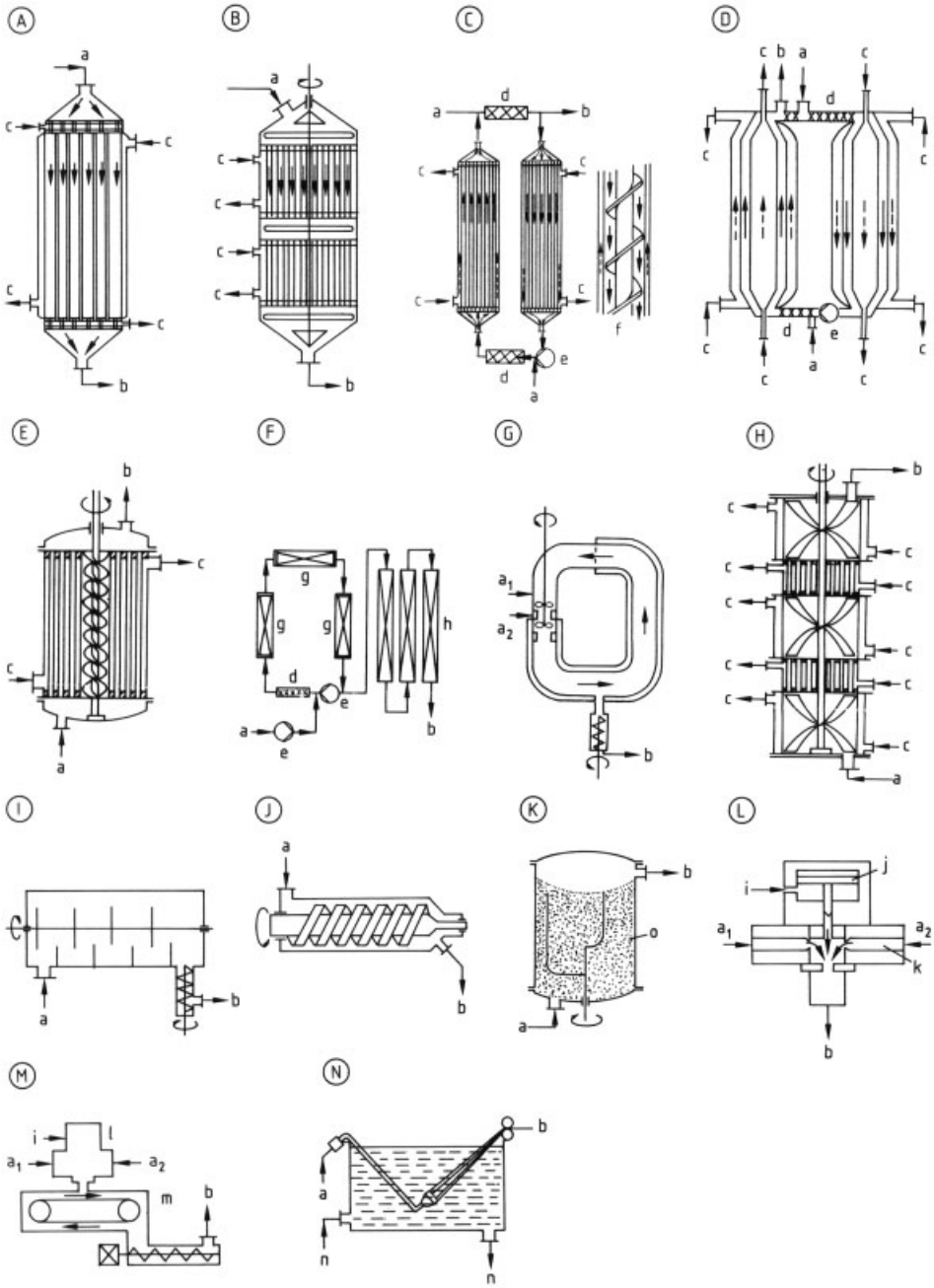

Figure 4. Special reactor designs for polymerization reactions
A) Multitubular reactor; B) Multistage multitubular reactor with interstage stirring; C) Reactor with external recycle (multitubular or screw-conveyor type); D) Reactor with external recycle (annular); E) Reactor with internal recirculation; F) Sulzer loop reactor (see Fig. 3C for detail of a single reactor); G) Loop reactor; H) Tower reactor; I) Ring-and-disk reactor; J) Extruder reactor; K) Powder-bed reactor; L) Mixing head; M) Belt reactor with mixing head; N) Spinning jet with coagulating bath
a) Polymerization mixture; a_1, a_2) Feed components; b) Polymerization product; c) Coolant; d) Static mixer; e) Pump; f) Screw-conveyor design for viscous media; g) Sulzer mixer – reactor; h) Sulzer mixer – reactors in plug-flow configuration; i) Air; j) Plunger; k) Nozzle; l) Mixing head; m) Belt reactor; n) Spinning bath; o) Packed bed of polymer granules

Table 1. Reactors for exothermic gas-phase reactions

Reactor type	Features	Examples of applications
Burner	for high reaction rates	combustion of H_2S to SO_2 (Claus vessel)
	very high reaction temperatures	carbon black production (furnace, gas, thermal carbon black processes)
	explosion limits must be taken into consideration	chlorine – hydrogen reaction
		chlorination of methane
		nitration of propane
Tubular reactor	well-defined residence time (tubes up to 1000 m long)	chlorination
		of methane
	intermediate injection possible	of propene to allyl chloride
	pressure drops	of butadiene to dichlorobutane
	good temperature control capability	chlorolysis of chlorinated hydrocarbons
Reactor with recycle	suitable for low reaction rates	chlorination of methane
	good mixing	
	cooling inside or outside reactor	
Fluidized-bed reactor	nearly isothermal conditions because heat transport is very efficient	chlorination
		of methane
	intensive mixing	of 1,2-dichloroethane to tri- and perchloroethylene
		chlorolysis of chlorinated hydrocarbons

Table 2. Reactors for endothermic gas-phase reactions

Reactor type	Features	Examples of applications
Burner	very high reaction temperatures attainable by partial combustion of reactants	Sachsse – Bartholomé process for acetylene production
	short residence times	high-pressure gasification for synthesis gas production (Texaco, Shell)
Reformer	high reaction temperatures attainable mainly by radiation	steam cracking of naphtha and other hydrocarbons to ethylene
	well-defined residence times	vinyl chloride production by cleavage of dichloroethane
		pyrolysis
		of acetic acid to ketene
		of 2-methyl-2-pentene
		to isoprene (in presence of HBr)
		of chlorodifluoromethane
		to tetrafluoroethylene
Fluidized-bed reactor	heat supplied along with solids	Lurgi Sandcracker
Moving-bed reactor	heat supplied along with solids	Langer – Mond process for production of ultrapure nickel
	continuous removal of solid products	
Reactor with fixed bed of inerts	fixed bed ensures heat storage and intensive mixing	Kureha process for acetylene and ethylene production
		production of CS_2 from CH_4 and sulfur vapor
Regenerative furnaces	battery operation	gas generation from heavy crudes
	no dilution by heat-transfer medium	

ber of other reaction types, nearly all industrially important polymerization reactions take place in the liquid phase. For the sake of completeness,a few important exceptions among polymerization reactions are included in this section, even though they do not fall under liquid-phase reactions according to the classification principle stated above. These are, in particular, "gas-phase polymerization" reactions, some of which take place over solid complex catalysts of the Ziegler – Natta type (high-density polyethylene, linear low-density polyethylene, and polypropylene).

The essential feature of polymerization reactions is that, in contrast to other liquid-phase reactions, the viscosity increases rapidly during the course of reaction and causes difficulties in heat and mass transport. In industry, this problem is countered by (1) the use of special stirring and kneading devices; (2) running the process in several stages; (3) raising the temperature as the conversion increases; and (4) carrying out polymerization in thin films.

Table 3. Reactors for liquid-phase reactions (one or more phases present)

Reactor type	Features	Examples of applications
Tubular reactor	well-defined residence time good temperature control capabilities	polymerization reactions bulk polymerization to LDPE ∗ polycondensation to PA 66 ∗ (2nd stage) hydrolysis reactions of ethylene oxide and propylene oxide to glycols of chlorobenzene to phenol and chlorotoluene to cresol of allyl chloride production of ethyl acetate from acetaldehyde production of isopropanolamine dehydrochlorination of 1,1,2-trichloroethane to vinylidene chloride
Reformer	high reaction temperature well-defined residence time	visbreaking delayed coking pyrolytic dehydrochlorination of tetrachloroethane to trichloroethylene high-pressure gasification of heavy crudes
Multitubular reactor	large heat-transfer area multistage design with stirring elements between stages is possible	bulk polymerization to PS ∗, HIPS ∗, and SAN ∗
Sulzer mixer – reactor (plug-flow configuration)	mixing elements consist of tubes carrying heat-transfer medium large heat-transfer area suitable for processes in which viscosity increases intensive radial mixing with little axial backmixing very narrow residence-time distribution	bulk polymerization to PS ∗ and polyacrylates temperature-controlled starch conversion
Reactor with external recirculation	good mixing and heat-removal conditions no moving parts suitable for low reaction rates heat exchanger can be placed outside reactor	cleavage of cumene hydroperoxide to phenol and acetone (2nd stage of Hock process) Beckmann rearrangement of cyclohexanone oxime to caprolactam production of hydroxylamine sulfate (Raschig process) production of phosphoric acid (wet process) saponification of allyl chloride bulk polymerization to PS ∗, HIPS ∗, SAN ∗, and PMMA ∗
Reactor with internal recirculation	very intensive mixing	production of melamine from molten urea (high-pressure process) production of aromatic nitro compounds production of adipic acid from cyclohexanol and nitric acid Bulk polymerization to PS ∗, HIPS ∗, and SAN ∗
Loop reactor	for slurry polymerization suspension is circulated at high velocity to prevent buildup	polymerization reactions slurry polymerization to PP ∗ production of HDPE ∗ and LLDPE ∗
Powder-bed reactor	liquid monomers supported on already polymerized granules for high conversion evaporating and condensing monomer acts as heat-transfer agent (boiling, cooling) vertical and horizontal designs	polymerization reactions polymerization to HDPE ∗ and PP ∗ block copolymerization to PE – PP ∗ precipitation polymerization to PAN ∗, IIR ∗, PE ∗, PP ∗

Table 3. Continued

Reactor type	Features	Examples of applications
Stirred tank, batch or semicontinuous	limited heat-transport capability mechanical stirring means suitable for slow reactions	polymerization reactions bulk polymerization to PS ∗, PMMA ∗, HIPS ∗, ABS ∗ (1st stage of each process) polycondensation to PA 66 ∗ solution polymerization to PVAC ∗, PAN ∗, PE ∗, PP ∗, EPM ∗, EPDM ∗, SB ∗, SB – S ∗, EO – PO ∗ polycondensation to UF ∗, MF ∗, PF ∗ resins precipitation polymerization to PVC ∗, PAN ∗, PE ∗, PP ∗, EPM ∗, EPDM ∗ suspension polymerization to PVC ∗, EPS ∗, PMMA ∗, PVAC ∗, and ion-exchange resins based on PS ∗, HIPS ∗, ABS ∗ (2nd stage) emulsion polymerization to numerous polymer dispersions production of aromatic nitro compounds sulfonation of benzene esterification of PA ∗ and alcohol to diphthalates many other syntheses of dyes and pharmaceuticals
Stirred tank, continuous	suitable for fast reactions with large negative or positive heat of reaction approximately complete mixing conversion generally not complete mechanical stirring means	polymerization reactions bulk and solution polymerization to PS ∗, PMMA ∗, HIPS ∗, and ABS ∗ (1st stage in each case); copolymers with nonazeotropic monomer ratios precipitation polymerization to PAN ∗, IIR ∗, PE ∗, PP ∗ emulsion polymerization to PVC ∗ and SAN ∗ esterification of acrylic acid with alcohol of acetic acid with ethanol dehydration of 1,4-butanediol to tetrahydrofuran of ethanol to diethyl ether saponification of benzyl chloride of fatty acids dehydrochlorination of 3,4-dichloro-1-butene to chloroprene of 1,1,2-trichloroethane to vinylidene chloride cyclization of glycols to 1,4-dioxane nitration of aliphatic hydrocarbons alkylation of isobutane with *n*-butenes production of melamine from molten urea (Montecatini) oxidation of cyclohexanone/ol with HNO_3 to adipic acid of mono- to dicarboxylic acids of allyl alcohol with H_2O_2 to glycerol
Cascade of stirred tanks	suitable for slow reactions adaptable to needed reaction conditions stage by stage residence-time distribution close to that of tubular reactor	polymerization reactions transesterification of DMT ∗ to DGT ∗ polycondensation to PETP ∗ and PBT ∗ solution polymerization to BR ∗, IR ∗, UP ∗, UF ∗, MF ∗, PF ∗ resins solution or precipitation polymerization to PE ∗, PP ∗, EPM ∗, EPDM ∗ emulsion polymerization to SBR ∗, CR ∗, NBR ∗ production of hydroxylamine sulfate (Raschig process) production of cyclohexanone oxime from cyclohexanol and hydroxylammonium sulfate nitration of aromatic hydrocarbons decomposition of ammonium carbamate to urea production of plasticizers from phthalic anhydride and alcohol production of MDA ∗ in conjunction with downstream tubular reactor production of methacrylamide from acetocyanohydrin production of MDI ∗ from MDA ∗ and TDI ∗ from TDA ∗

Table 3. (Continued)

Reactor type	Features	Examples of applications
Reaction column	reaction and separation in a single apparatus equilibrium can be modified by removing one or more components from reaction space	aldol condensation of *n*-butyraldehyde to 2-ethylhexenal saponification of chloropropanol with milk of lime of fatty acids esterification of acetic acid with butanol of phthalic anhydride with alcohols decomposition of amalgam of ammonium carbamate to urea and water
Multichamber tank	virtually identical to cascade of stirred tanks requires little space chamber-by-chamber feed injection possible	polymerization to LDPE ∗ (ICI) alkylation of isoparaffins with olefins (Kellogg)
Tower reactor	for continuous processes section-by-section temperature control possible little backmixing at high viscosity also in cascade or with upstream stirred tank	bulk and solution polymerization of PS ∗, HIPS ∗, ABS ∗, SAN ∗, PA 6 ∗
Ring-and-disk reactor	narrow residence-time distribution	final stage in production of PETP ∗ and PBT ∗
Extruder	for highly viscous media	polymerization reactions production of POM ∗ from trioxane final stage in production of PA 66 ∗
Fluidized-bed reactor	very good heat- and mass-transport conditions	polymerization to HDPE ∗, LLDPE ∗, PP ∗ fluid coking of heavy residual oils (Exxon) melamine production from molten urea
Mixing head with injection mold	special design for bringing several liquid reactants together	production of PUR ∗
Belt reactor with mixing head	for fabrication of sheets and films	production of PIB ∗, PMMA ∗, PUR ∗, PVAL ∗
Spinning jet (with coagulating bath)	for production of strands	viscose spinning
Spray reactor	direct heating in hot stream of gas	thermal H_2SO_4 cleavage production of MgO from $MgCl_2$ (spray calcination)
Falling-film reactor	gentle temperature control due to large heat-transfer area	sulfation of fatty alcohols diazotization of aromatic amines diazo coupling

∗ The following abbreviations are used: ABS = acrylonitrile – butadiene – styrene copolymer; BR = butadiene rubber; CR = chloroprene rubber; DGT = diglycyl terephthalate; DMT = dimethyl terephthalate; EO – PO = ethylene oxide –propylene oxide block copolymer; EPDM = ethylene – (propene – diene) copolymer; EPM = ethylene – propene copolymer; EPS = expandable polystyrene; HDPE = high-density polyethylene; HIPS = high-impact polystyrene; IIR = isobutylene – isoprene rubber (butyl rubber); IR = isoprene rubber (synthetic); LDPE = low-density polyethylene; LLDPE = linear low-density polyethylene; MA = maleic anhydride; MDA = 4,4′-diaminodiphenyl methane; MDI = methylene diphenylene isocyanate; MF = melamine – formaldehyde; NBR = butadiene – acrylonitrile copolymer (nitrile rubber); PA = polyamide; PAN = polyacrylonitrile; PBT = poly(butylene terephthalate); PE = polyethylene; PE – PP = polyethylene – polypropylene copolymer; PETP = poly(ethylene terephthalate); PF = phenol – formaldehyde; PIB = polyisobutylene; PMMA = poly(methyl methacrylate); PO = poly(propylene oxide); POM = polyoxymethylene; PP = polypropylene; PS = polystyrene; PUR = polyurethane; PVAC = poly(vinyl acetate); PVAL = poly(vinyl alcohol); PVC = poly(vinyl chloride); SAN = styrene – acrylonitrile copolymer; SBR = styrene – butadiene rubber; SB = styrene –butadiene block copolymer; SB – S = styrene – butadiene – styrene block copolymer; TDA = toluene diamine; TDI = toluene diisocyanate; UF = urea – formaldehyde; UP = unsaturated polyester.

Table 3 and Figures 3) and 4 summarize the types of reactors used in industry for liquid-phase reactions. Figure 4 shows special reactor designs for polymerization reactions.

3.3. Reactors for Gas – Liquid Reactions

Gas – liquid reactions include many industrially important processes, such as oxidation, alkylation, chlorination, and flue-gas scrubbing. The prerequisite for an efficient reaction is rapid mass transport between gas and liquid. Important criteria for assessment include

1) The interfacial area
2) The mass or volume ratio of gas to liquid
3) The energy required to mix the phases

Other important factors are temperature control, heat removal, and residence time (especially that of the liquid phase).

Reactor design is dictated largely by the way in which the interface is generated. The following methods are possible:

1) Reactors with continuous liquid-phase and fixed gas distribution devices [bubble columns (→ Bubble Columns), packed and tray reactors (→ Reaction Columns)]
2) Reactors with mechanical gas dispersion (sparged stirred tanks)
3) Reactors with continuous gas phase and liquid dispersing devices (spray reactors, liquid-ring pumps)
4) Thin-film reactors (→ Thin-Film Reactors)

Figure 5 illustrates reactor types for gas – liquid reactions. Important applications are listed in Table 4.

Table 4. Reactors for gas – liquid reactions

Reactor type	Features	Examples of applications
Tubular reactor (with injection)	large transfer areas can be provided temperature can be controlled by indirect transfer well-defined residence time little backmixing	production of adiponitrile from adipic acid and ammonia production of nitroaniline by amidation of chloronitrobenzene oxidation of ethylene to acetaldehyde (2nd stage)
Bubble column	can be equipped with internal draft tube or external recirculation simple design practically isothermal conditions can be achieved by evaporation of a liquid component limited temperature and pressure ranges suitable for continuous and semicontinuous operation high backmixing in liquid phase	oxidation of ethylene to acetaldehyde of toluene to benzoic acid of cyclohexane to cyclohexanone/ol of acetaldehyde to acetic acid and acetic anhydride of cumene to cumene hydroperoxide of *p*-xylene to dimethyl terephthalate (1st stage) of butane to acetic acid of butene to ketones of propane to acetone of copper chlorides to copper oxychlorides alkylation of benzene with ethylene and propene epoxidation of propene with organic hydroperoxides (Halcon) hydroformylation of propene (oxo synthesis) chlorination of benzene to chlorobenzene of ethylene to 1,2-dichloroethane of alkalies to hypochlorites of 1,2-dichloroethane or vinyl chloride to 1,1,2-trichloroethane production of acrylic acid or acrylates from acetylene, CO, and water or alcohol decarboxylation of benzoic acid to phenol production of vinyl ether from acetylene and alcohol carboxylation of methanol to acetic acid production of methylpyridine from ethylene and ammonia hydrolysis of trichloroethylene to monochloroacetic acid hydration of acetylene to acetaldehyde production of propionic acid from ethylene, CO, and water (Reppe process)
Reactor with external recirculation	heat exchangers in external loop mixing elements can be included suitable for continuous and batch processes	alkylation of *o*-toluidine with ethylene of phenol with isobutene air oxidation of liquid hydrocarbons
Liquid-ring pump	combines compression, absorption, and reaction	production of acetic anhydride from acetic acid and ketene

Table 4. Continued

Sparged stirred tank	continuous and semicontinuous operation possible stirrer shaft must be sealed much backmixing in both phases external liquid recirculation can be added	oxidation of *p*-xylene to terephthalic acid (Amoco process) of ethylene to acetaldehyde of propene to acetic acid of benzoic acid to phenol of cyclohexane to cyclohexanone/ol of toluene to benzaldehyde of butene to methyl ethyl ketone of cyclohexanol to adipic acid of acetaldehyde to acetic anhydride alkylation of benzene with ethylene chlorination of aromatics hydrogenation of natural fats to fatty acids or their esters acetylation of carboxylic acids to vinyl ethers production of pure SO_2 from liquid sulfur and SO_3
Cascade of stirred tanks	narrower residence-time distribution than single tank	epoxidation of propene with organic hydroperoxides (Halcon process) hydrolysis of fats to fatty acids and glycerol production of phthalic acid (Progil)
Reaction column	tray or packed column design combination of reaction and separation permits control of equilibrium cocurrent or countercurrent operation	oxidation of ethylbenzene, cumene, isobutane to the hydroperoxides (1st stage of Halcon process) production of chloropropanols from propene, chlorine, and water absorption of nitrogen oxides in nitric acid production of ammonia and CO_2 in soda production of gases in scrubbers (with amines) chlorination of ethylene to 1,2-dichloroethane of acetylene to tetrachloroethane of trichloroethylene to perchloroethane hydrochlorination of ethylene to chloroethane precipitation of $NaHCO_3$ in soda production production of sulfonic acids (benzene-, chlorosulfonic acids) high-temperature debromination of salt brines production of alkyl and dialkyl sulfates production of phthalimide from phthalic anhydride and NH_3 production of ammonium nitrate from HNO_3 and NH_3
Spray reactor (jet scrubber)	suitable for high gas throughputs and high gas-side solids transport resistance very large interfacial areas	gas treatment and off-gas scrubbing production of acetic anhydride from acetic acid and ketene production of H_3PO_4 (Knapsack process) acid refining of lubricants hydrolysis of dichlorodimethylsilane production of potassium hyperoxide by contacting of molten potassium with air
Falling-film reactor	gentle temperature control due to large heat-transfer area	Sulfonation of dodecylbenzene, fatty alcohols, and olefins
Rotary kiln	for reaction of molten solids with gases	oxidation of sodium to sodium peroxide (two-stage process) of molten sulfur to SO_2

3.4. Reactors for Solid-Catalyzed Reactions

Heterogeneous catalytic processes play a major role in chemical technology, because many key products and intermediates can be manufactured in this way. Fluid reactants react in the presence of a solid catalyst, the mechanism as a whole consisting of the reaction proper and a series of upstream and downstream transport steps.

3.4.1. Reactors for Heterogeneous Gas Catalysis

Reactors with a fixed catalyst bed are distinguished from those with moving catalyst.

Fixed-Bed Reactors (→ Fixed-Bed Reactors). The characteristic features of a reactor with fixed catalyst are the pressure drop of the flowing gas in the catalyst bed and the danger of unstable operation points, especially with strongly exothermic reactions, when flow through the catalyst bed becomes nonuniform. Fixed-bed reactors must be shut down after a certain time onstream to regenerate or replace the catalyst.

Fixed-bed reactors can be classified by the type of temperature control:

1) Reactors with no special temperature control features (adiabatic operation)
2) Reactor systems with stagewise temperature control (chiefly for equilibrium reactions)
3) Reactors with continuous heat exchange along the flow path (polytropic operation)

Fixed-bed reactors *without equipment for temperature control* are marked by a particularly simple construction and low flow resistance, which makes them suitable for high gas throughputs. A summary of these reactors appears in Table 5 and Figure 6.

Reactor systems with *stagewise temperature control* are used primarily for equilibrium reactions. Such a reactor consists of simple adiabatic reactor elements connected in series and takes the form of several units or a system housed in a common reactor shell. Temperature control is accomplished by heat transfer between reactor stages or by the injection of tempered gas or vapor streams at points along the flow path. Table 6 and Figure 7 present reactor systems of this type along with applications.

If the reaction process imposes special requirements on temperature control, heat-transfer surfaces must be located throughout the reactor volume. The best-known design for such a reactor is the *multitubular reactor*, which is frequently used in the chemical industry. The drawbacks relative to other fixed-bed reactors include the much more complicated design and the limitation on throughput due to the smaller cross-sectional area available for flow.

Temperature control is achieved by the use of gaseous and liquid heat-transfer media. One highly effective approach is the use of boiling liquids (e.g., pressurized-water and evaporatively cooled reactors). A special case is the autothermal process regime, in which the reaction mixture itself is used as a temperature control medium before it flows through the catalyst bed. Fixed-bed reactors with *continuous heat exchange* are described in Table 7 and Figure 8, along with applications.

Moving-Bed and Fluidized-Bed Reactors (→ Fluidized-Bed Reactors). In moving-bed reactors, transport of the catalyst is influenced by gravity and the drag force exerted by the flowing reaction fluid on the catalyst particles. The regime in the reactor can vary widely, depending on the ratio of these forces. The fol-lowing features must be taken into consideration when using reactors of this type:

1) The possibility of continuous catalyst regeneration
2) Increased mechanical loads on the catalyst and reactor materials
3) The favorable conditions for heat and mass transport, resulting from rapid movement of solids and small catalyst grain size

Table 8 and Figure 9 list reactor types and applications.

3.4.2. Reactors for Liquid-Phase and Gas – Liquid Reactions over Solid Catalysts

Fixed-bed reactors (trickle-flow reactors and packed bubble columns) are used for liquid-phase reactions, as well as gas – liquid reactions over solid catalysts. The presence of a liquid

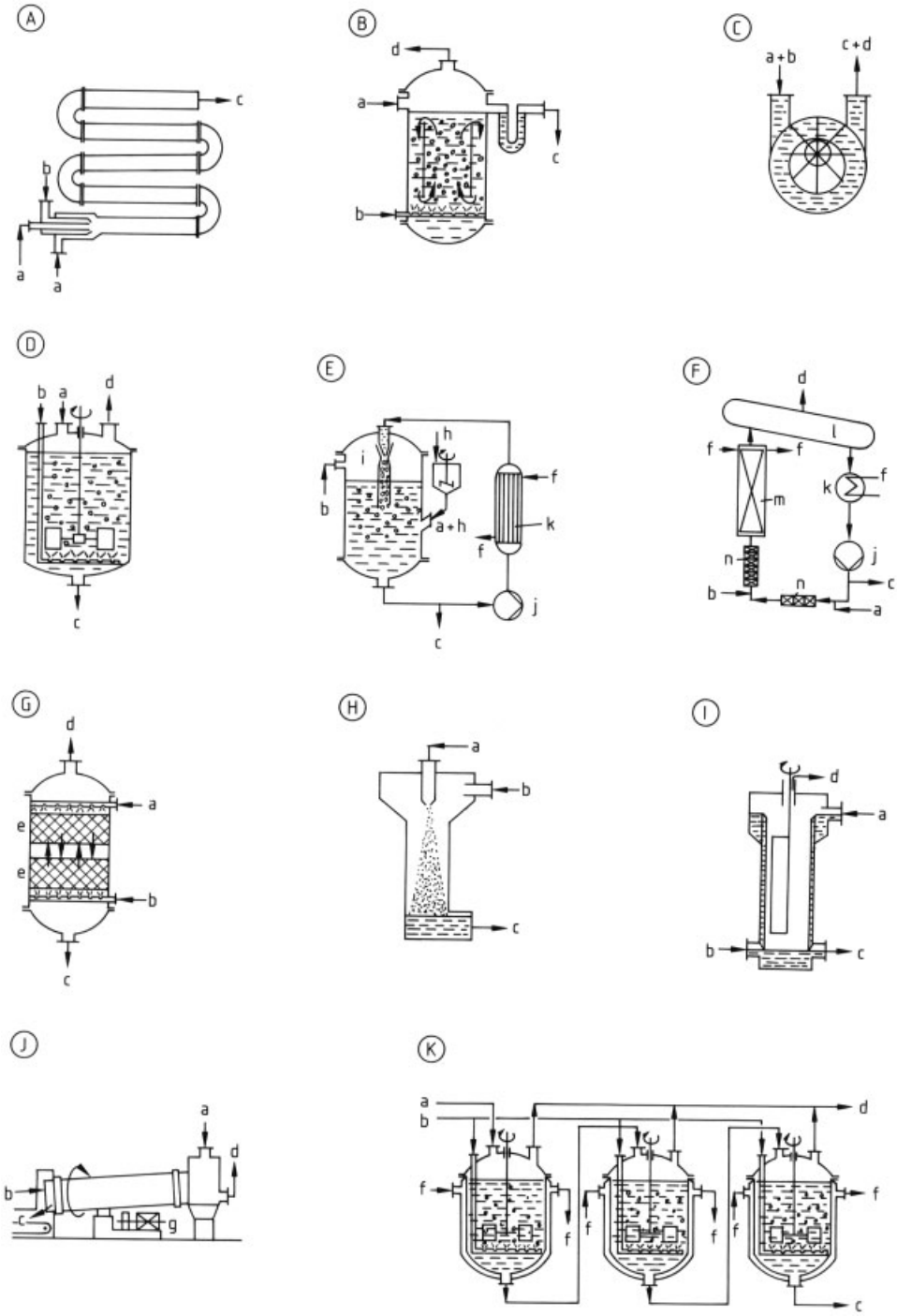

Figure 5. Reactors for gas – liquid reactions
A) Tubular reactor with injector; B) Bubble column; C) Liquid-ring pump; D) Sparged stirred tank; E) Buss loop reactor; F) Sulzer mixer – reactor in loop configuration; G) Reaction column; H) Spray reactor; I) Falling-film reactor; J) Rotary kiln; K) Cascade of stirred tanks
a) Liquid feed component; b) Gaseous feed component; c) Liquid product; d) Off-gas; e) Packing; f) Heating agent or coolant; g) Drive unit; h) Catalyst; i) Reaction mixer with mixing nozzle; j) Pump; k) Heat exchanger; l) Gas separator; m) Sulzer mixer – reactor (see Fig. 3C for detail of a single reactor); n) Static mixer

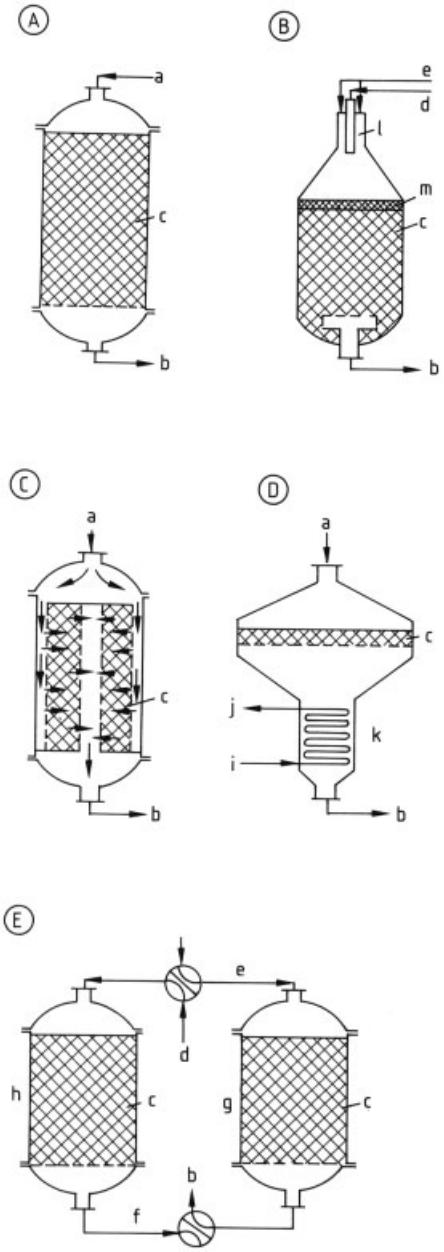

Figure 6. Fixed-bed catalytic reactors for gas-phase reactions with no special provisions for temperature control
A) Simple fixed-bed reactor; B) Fixed-bed reactor with combustion zone; C) Radial-flow reactor; D) Shallow-bed reactor; E) Regenerative furnace
a) Gaseous reaction mixture; b) Gaseous product; c) Catalyst; d) Air; e) Hydrocarbon; f) Flue gas; g) Reaction section; h) Regeneration section; i) Condensate; j) Steam; k) Steam generator; l) Burner; m) Inert guard bed

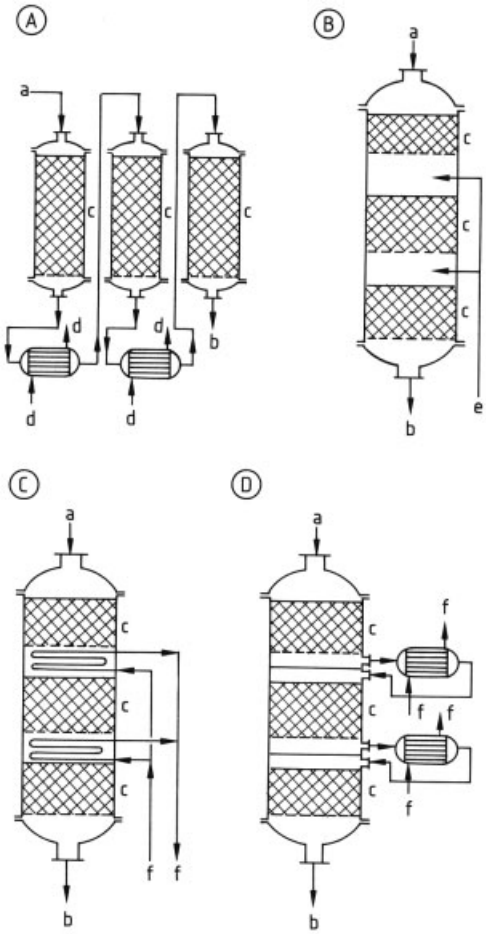

Figure 7. Fixed-bed catalytic reactors for gas-phase reactions with stagewise temperature control
A) Cascade of simple fixed-bed reactors; B) Multibed reactor with cold-gas or steam injection; C) Multibed reactor with intercooling (internal); D) Multibed reactor with intercooling (external)
a) Gaseous reaction mixture; b) Gaseous product; c) Catalyst; d) Heating agent; e) Cold gas; f) Coolant

Table 5. Fixed-bed catalytic reactors for gas-phase reactions with no special provisions for temperature control

Reactor type	Features	Examples of applications
Simple fixed-bed reactor (axial flow)	very simple design not suitable for reactions with large positive or negative heat of reaction and high temperature sensitivity	reforming (Platforming, Rheniforming, etc.) hydrotreating CO converting amination of methanol to methylamines desulfurization and methanation in synthesis-gas path upstream of primary reformer hydrogenation of nitrobenzene to aniline (Allied, Bayer) production of vinyl propionates from acetylene and propionic acid isomerization of *n*-alkanes dehydrogenation of ethylbenzene to styrene disproportionation of toluene to benzene and xylene
Fixed-bed reactor with combustion zone	direct heating by combustion of part of hydrocarbon feed	methane cleavage in secondary reformer
Radial-flow reactor	much lower pressure drop than axial-flow reactor multistage configuration possible enhanced backmixing due to small thickness of bed uniformity of flow requires exact sizing of distributing and collecting ducts	ammonia synthesis (Topsoe, Kellogg) dehydrogenation of ethylbenzene to styrene (Dow) reforming
Shallow-bed reactor	used for high reaction rates and unstable products very short residence time catalyst can also be in gauze form suitable for autothermal operation	oxidation of ammonia to NO_x oxidative dehydrogenation of methanol to formaldehyde production of hydrocyanic acid from ammonia, methane, and air (Andrussow process)
Regenerative furnace	suitable when catalyst ages rapidly and can be regenerated by burning off reaction heat can be supplied by catalyst regeneration battery operation	dehydrogenation of butane to butadiene (Houdry process) SO_2 reduction with methane (Andrussow process)

Table 6. Fixed-bed catalytic reactors for gas-phase reactions with stagewise temperature control

Reactor type	Features	Examples of applications
Cascade of fixed-bed reactors	large pressure and temperature differences are possible	reforming of heavy gasoline hydrocracking conversion of H_2S and SO_2 to elemental sulfur (Claus process) isomerization of five-to-six-ring naphthenes
Multibed reactor with cold-gas injection	used for exothermic equilibrium reactions injection of reaction mixture leads to lower conversion and thus increased number of stages injection of water lowers concentration at constant conversion adaptation of bed depth to progress of reaction	ammonia synthesis methanol synthesis hydrocracking hydrogenation of benzene desulfurization of vacuum gas oil
Multibed reactor with interstage cooling	used for exothermic equilibrium reactions internal or external heat exchangers no dilution effects adaptation of bed depth to progress of reaction	ammonia synthesis (ÖSW, Fauser, Montecatini) SO_2 oxidation (with interstage adsorption) hydrodealkylation of alkyl aromatics
Multibed reactor with heat supply	used for endothermic equilibrium reactions interstage heating or interstage injection of superheated steam	dehydrogenation of ethylbenzene to styrene (Dow)

Table 7. Fixed-bed catalytic reactors for gas-phase reactions with continuous temperature control

Reactor type	Features	Examples of applications
Multitubular reactor	temperature control with liquid, gaseous, or boiling heat-transfer agent in shell-side space relatively expensive design high cost and long downtime for catalyst replacement internals in shell-side space to improve heat transfer distribution devices needed for stream of reaction gas in large reactors relatively large pressure drop autothermal processing possible	oxidation of ethylene to ethylene oxide of *o*-xylene or naphthalene to phthalic anhydride of propene to acrolein or acrylic acid of acrolein to acrylic acid of methanol to formaldehyde (Formox process) of naphthalene to 1,4-naphthoquinone of ethanol to acetaldehyde alkylation of benzene with ethylene or propene to ethylbenzene, diethylbenzene, or cumene (Mobil, Badger processes) hydrogenation of CO to methanol (Lurgi low-pressure process) ammonia synthesis (TVA, NEC processes) of phenol to cyclohexanone/ol of natural fats to fatty acids dehydrogenation of ethylbenzene to styrene of diethylbenzene to divinylbenzene of cyclohexanol to cyclohexanone of 2-butanol to methyl ethyl ketone of ethanol to acetaldehyde hydration of propene to isopropanol dehydration of phenylmethylcarbinol to styrene of acetic acid to acetic anhydride of ethanol to diethyl ether oxychlorination of ethylene to ethylene dichloride (Stauffer process) production of vinyl acetate from acetic acid and ethylene or acetylene production of methyl *tert*-butyl ether from methanol and isobutene (1st stage) hydrochlorination of acetylene to vinyl chloride of methanol or ethanol to chloromethane or chloroethane
Tubular furnace	direct heating of catalyst tubes for endothermic reactions and high reaction temperatures	primary reformer
Fixed-bed reactor with heating or cooling elements	advantageous when only the heating or cooling system has to be designed for special pressure conditions classical examples are Fischer–Tropsch reactor and butadiene reactor of four-stage process	methanol synthesis (Linde isothermal reactor)

phase, however, leads to much greater drag and friction forces on the catalyst. If the reaction involves both gas and liquid phases, maintenance of uniform flow conditions through the catalyst bed and intensive mixing of the phases can be difficult. The crucial factor for the efficiency of catalytic processes is the wetting of the catalyst by the liquid. Since reactors of this type are usually operated adiabatically, local overheating is a danger, especially with exothermic reactions. Fixed-bed reactors are well suited to high-pressure processes by virtue of their simple design.

A second important group includes *suspension reactors*, in which very fine catalyst particles are distributed throughout the volume of the liquid (stirred tanks and bubble columns

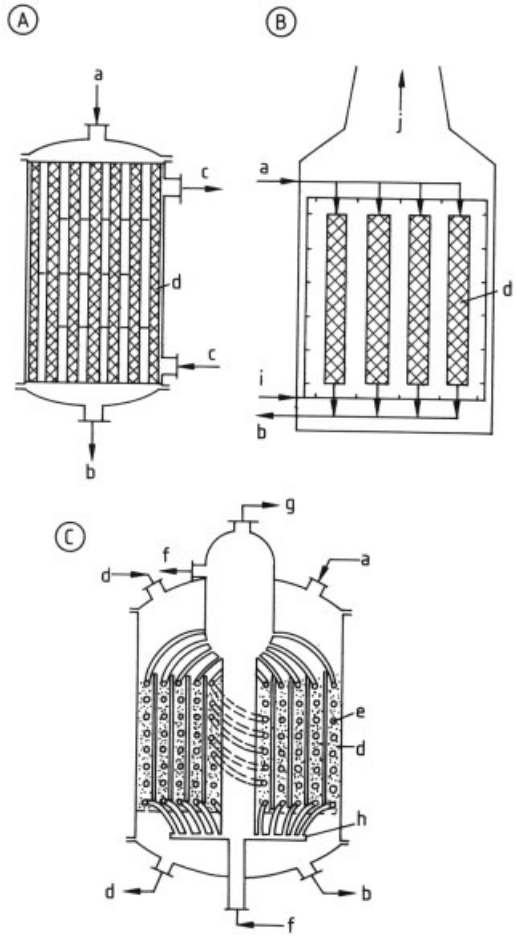

Figure 8. Fixed-bed catalytic reactors for gas-phase reactions with continuous temperature control
A) Multitubular reactor; B) Tubular reformer; C) Fixed-bed reactor with heating or cooling elements
a) Gaseous reaction mixture; b) Gaseous product; c) Heating agent or coolant; d) Catalyst; e) Cooling tubes; f) Circulating water; g) Steam; h) Tube sheet; i) Fuel gas for burners; j) Off-gas

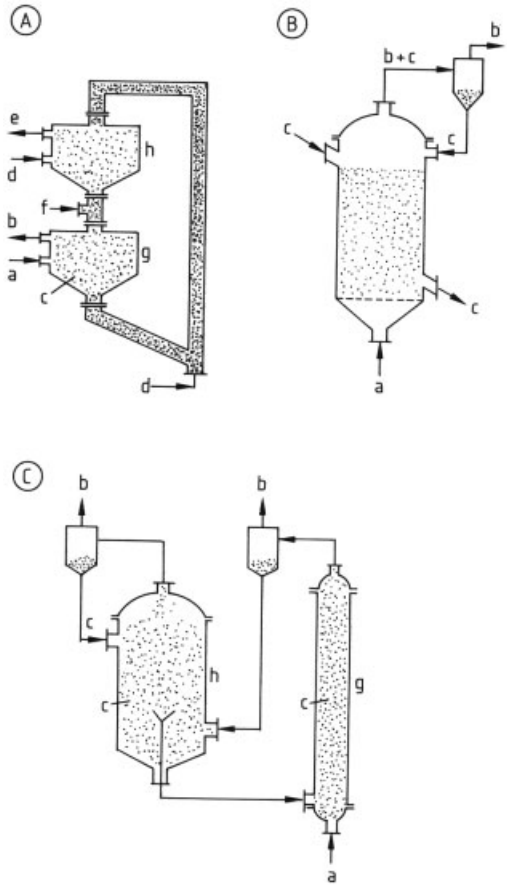

Figure 9. Moving-bed catalytic reactors for gas-phase reactions
A) Moving-bed reactor; B) Fluidized-bed reactor; C) Entrained-flow reactor
a) Reaction mixture; b) Gaseous product; c) Catalyst; d) Air; e) Flue gas; f) Blocking steam; g) Reaction section; h) Regeneration section

with suspended catalyst). Because transport resistances are reduced, these reactors offer a close approach to isothermal operating conditions and a favorable utilization of the catalyst volume.

Sophisticated techniques are required to separate the finely divided catalyst from the liquid. Equipment for this purpose can be installed inside or outside the reactor. At the same time, this arrangement permits continuous catalyst replacement. All suspension reactors have the disadvantage of increased backmixing, especially of the liquid phase, which affects product distribution.

The *fluidized-bed reactor* differs from the suspension reactor in the use of coarser catalyst particles and the formation of a well-defined agitated catalyst bed below the liquid level.

Industrially important reactors for liquid-phase and gas – liquid reactions over solid catalysts are listed, together with their applications, in Tables 9 and 10 and Figures 10 and 11.

Table 8. Moving-bed catalytic reactors for gas-phase reactions

Reactor type	Features	Examples of applications
Moving-bed reactor	gravity transport of catalyst	cracking (TCC, Houdry flow process)
	reaction conditions largely similar to those in fixed-bed reactor	dehydrogenation of butane
	advantageous when catalyst can be regenerated by burning off residues	
Fluidized-bed reactor	catalyst agitated by gravity and resistance force of gas flow	cracking (Kellogg, FFC, Flexicracking)
	almost isothermal conditions can be achieved in fluidized bed	hydrocracking
	pressure drop independent of gas throughput over a wide range	reforming
	form of fluidized bed can be varied as a function of geometric and hydraulic conditions	ammoxidation
	strong backmixing internals to improve mass transport and heat transfer are common	of propene to acrylonitrile (Sohio process)
	catalysts must have high abrasion resistance	of *o*-xylene to *o*-phthalodinitrile
		production of adiponitrile from adipic acid and ammonia
		oxychlorination of ethylene to 1,2-dichloroethane (Goodrich)
		production of melamine from urea (BASF)
		hydrogenation
		of nitrobenzene to aniline (BASF, Cyanamid)
		of ethylene
		oxidation
		of *o*-xylene or naphthalene to phthalic anhyride
		of butane to MA∗ (Du Pont)
		of SO_2 to SO_3
		of ethylene to ethylene oxide
		of NH_3 to NO
		of HCl to chlorine
		dehyrogenation
		of isopropanol
		of *n*-butane to *n*-butene
		production of chloromethylsilanes from chloromethane (catalytic gas – solid reaction)
		production of vinyl chloride (Cloe process)
		chlorination of methane and ethylene
		production of butadiene from ethanol
		isomerization of *n*-butane
		production of isoprene
		postchlorination of PVC∗
		combustion
Entrained-flow reactor	uses very fine-grained catalyst	Fischer – Tropsch process (Synthol process)
	whole quantity of catalyst circulates continuously between reaction section and tempering or regeneration unit	

∗ For abbreviations, see footnote to Table 3

3.5. Reactors for Noncatalytic Reactions Involving Solids

A variety of specialized reactors are available for noncatalytic reactions involving solids. The discussion that follows deals only with the industrially important types.

3.5.1. Reactors for Noncatalytic Gas – Solid Reactions

In general, noncatalytic gas – solid reactions are characterized by low overall reaction rates and

Table 9. Fixed-bed catalytic reactors for liquid-phase and gas – liquid reactions

Reactor type	Features	Examples of applications
Trickle-flow reactor	can operate in cocurrent or countercurrent	desulfurization and refining of petroleum products
	temperature control by intermediate injection or recirculation	hydrocracking
	danger of uneven liquid distribution and incomplete wetting of catalyst	production of butynediol from acetylene and formaldehyde
	narrow residence-time distribution	direct hydration of propene to 2-propanol (Texaco)
		hydrogenation
		of organic intermediates (butynediol, adiponitrile, ethylhexenal)
		of aldehydes, esters, and carboxylic acids to alcohols
		of natural fats to fatty acids
		of residues (low-temperature hydrogenation of tars)
		posthydrogenation
Packed bubble column	danger of flooding limit throughput capacity	amination of alcohols
	catalyst subject to greater mechanical stress (retention necessary)	cobaltizer and decobaltizer in oxo synthesis
	high liquid proportion promotes heat removal	disproportionation of toluene to benzene and xylene
	large amount of backmixing in liquid phase	

Table 10. Suspended-bed and fluidized-bed reactors for liquid-phase and gas – liquid reactions over solid catalysts

Reactor type	Features	Examples of applications
Bubble column with suspended catalyst	simple design	hydrogenation
	small pressure drop	of CO (Fischer – Tropsch synthesis)
	danger of undesired liquid-phase reactions	of tars and coals (bottom phase)
	inhomogeneous catalyst distribution must be prevented	of benzene to cyclohexane
	suitable if product drops out as solid	hydrodesulfurization
Reactor with external recirculation	heat-exchange and mixing devices in external loop	hydrogenation of organic intermediates (nitrobenzenes, nitriles, nitronaphthalenes, etc.)
	for continuous and batch operation	
	catalyst separation outside reactor	
Sparged stirred tank with suspended catalyst	can also be operated in semicontinuous and batch modes	hydrogenation of organic intermediates (nitro compounds, aromatics, butynediol)
	ensures intensive mixing of all phases	fat hydrogenation
	increased cost for sealing and maintaining stirrer drive	catalytic refining
Cascade of sparged stirred tanks with suspended catalyst	higher final conversions than in single stirred tank	hydrogenation of NO to hydroxylamine
	suitable for slow reaction rates	continuous hydrogenation of fats
	adaptable to intermediate injection and other interconnections	hydrolysis of fats to fatty acids and glycerol production of toluenediamine from dinitrotoluene
Fluidized-bed reactor	small pressure drop catalyst must have very high mechanical strength	hydrocracking and desulfurization of heavy petroleum fractions and still residues (H-Oil process; three-phase fluidized bed)

high process temperatures; in addition, the structure and geometry of the solid can change during the reaction.

Reactors for this service can essentially be grouped into those for semicontinuous operation, that is, *with no solids transport* (vertical shaft kilns and rotary drums), and those for continuous operation, that is, with *continuous solids transport*. The second type, in turn, can be divided into

1) Reactors with gravity transport of solids
2) Reactors with mechanical transport of solids
3) Reactors with pneumatic transport of solids

These three groups differ widely with respect to residence time, conditions of mass and heat transport between gas and solid phases, and heat-input capabilities. The first group includes *moving-bed reactors*. Since the gas has to flow through the bed of solids, mass and heat transport between the phases is relatively good. Temperature control can be effected by simultaneously carrying out exothermic and endothermic reactions in the same reactor.

Reactors with mechanical transport of solids include *rotary kilns* and *multiple-hearth furnaces* (→ Metallurgical Furnaces, Chap. 1., → Metallurgical Furnaces, Chap. 2.). Transport of gas and solid phases through the reactor largely occurs separately. Intensive heat and mass transfer occurs only at the surface of the bed of solids. Complete involvement of the solid phase in the reaction process depends on continuous, intensive mixing of the solids. Heat is often supplied directly by burners. More than one unit can be in operation in a single apparatus (e.g., drying, heating, cooling, and various reaction steps).

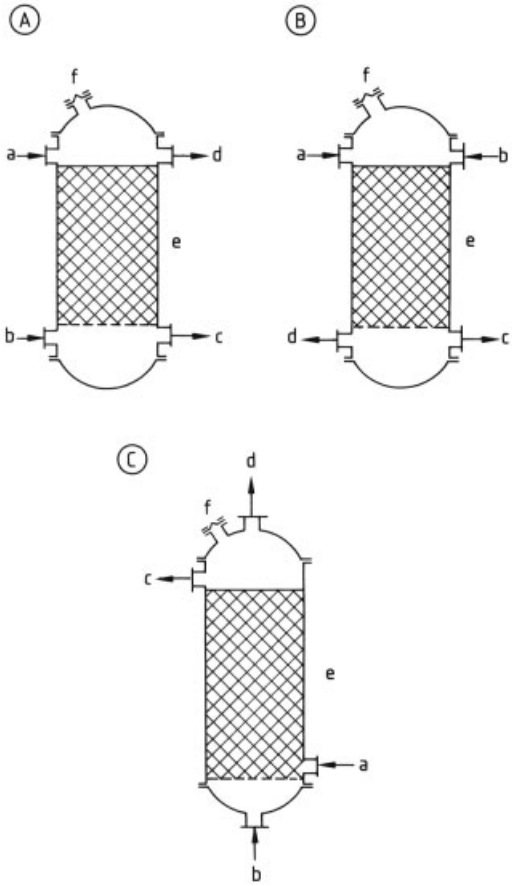

Figure 10. Fixed-bed catalytic reactors for liquid-phase and gas – liquid reactions
A) Trickle-flow reactor (countercurrent); B) Trickle-flow reactor (cocurrent); C) Packed bubble column
a) Liquid reactants; b) Gaseous reactants; c) Liquid product; d) Off-gas; e) Catalyst; f) Rupture disk

Solids transport by the gas stream is possible only with small particle sizes and the narrowest possible grain-size distribution. This group includes *fluidized-bed* and *entrained-flow reactors, dust roasters*, and *suspension furnaces*. Because of the favorable conditions for heat and mass transport, these reactors offer shorter residence times and thus higher throughputs than other types. The installation of heat-transfer surfaces, supplementary solid heat-transfer media, and direct heating is possible.

Industrially important reactor types for noncatalytic gas – solid reactions are listed in Table 11 and Figure 12 along with applications.
.

3.5.2. Reactors for Noncatalytic Liquid – Solid Reactions

Reactors used for noncatalytic liquid – solid reactions must be designed for the transport and mixing of phases, sometimes at high solids contents. Batch and semicontinuous designs are therefore dominant. Table 12 and Figure 13 present a survey of important reactor types for noncatalytic liquid – solid reactions and sample applications.

3.5.3. Reactors for Noncatalytic Solid-Phase Reactions

Reactors used for noncatalytic solid-phase reactions are similar to those used for noncatalytic gas – solid reactions. Long residence times and high reaction temperatures are necessary, especially for reactions between different solids, because of the low transport rates therein. Heat can be supplied by indirect or direct heating or by burning solid fuels.

Inert gases are employed for heat transport and agitation of the solids. Important applications are listed in Table 13.

3.6. Electrothermal Reactors

A variety of electrical heating schemes are used for some important noncatalytic reactions between gases and solids when very high reaction temperatures and large quantities of heat are required. In the simplest case, heating elements

Table 11. Reactors for noncatalytic gas – solid reactions

Reactor type	Features	Examples of applications
Shaft reactor, retort	simple design long residence times attainable semicontinuous operation	ore dressing autothermal coal gasification (Lurgi pressure gasification) thermal production of activated carbon chlorination of scrap iron to $FeCl_3$ reduction of copper compounds with HCl and H_2 to CuCl production of CS_2 from wood charcoal and sulfur vapor
Moving-bed reactor	long residence times attainable continuous operation heat supplied with gas stream	flue-gas desulfurization coal carbonization and coking (Lurgi purge-gas process) oil-shale retorting
Multiple-hearth reactor	for very long residence times individual hearths can be heated directly continuous operation gas can flow cocurrently or countercurrently small quantities of gas possible	roasting and reduction of ores production of activated carbon
Rotary kiln	relatively complicated design for long residence times direct heating possible continuous operation gas can flow cocurrently or countercurrently	roasting of ores production of calcium cyanamide from nitrogen and calcium carbide production of nickel carbonyl production of activated carbon production of ZnO from zinc burning of sulfur to SO_2
Fluidized-bed reactor	use of fine particles with narrow size distribution advantageous conditions can be made close to isothermal residence time controllable over a wide range entrainment of abraded material with gas stream necessitates downstream dust recovery step multistage operation and solids circulation possible use of supplemental solid heat-transfer agents possible high throughputs possible	burning of carbon for energy production autothermal coal gasification in Winkler generator in Ruhrgas process allothermal coal gasification (with solid heat-transfer agents) carbonization and coking fluidized-bed process Lurgi-Ruhrgas sulfur process roasting of ores chlorination of ores and lime production of activated carbon production of AlF_3 from Al_2O_3 and HF (multistage)
Entrained-flow reactor	solids in dust form with very narrow grain-size distribution very short residence time suspended dust can be conveyed with or without recycle for high throughputs internals for heat recovery by steam generation	carbonization and coking of coals (entrained-solids carbonization process) suspension smelting for beneficiation of copper and nickel ores (INCO process) coal gasification in Koppers–Totzek process in Prenflo process in Shell process
Dust roasters	very fine particles with narrow size distribution must be used solids drop downward velocity and residence time are governed by ratio of gravity to resistance force	ore roasting

(rods, strips, etc.) are used for this purpose. A much more efficient method, however, is direct , electric heating. Options here include arc, resistance, and induction heating.

The very high temperatures produced by the arc cause ionization in gases and thus activate the reactants; this feature is utilized in *plasma processes* for high-tempera-

Table 12. Reactors for noncatalytic liquid – solid reactions

Reactor type	Features	Examples of applications
Stirred tank	batch or semicontinuous operation predominant solids content limited by power of stirring apparatus	production of alkali cellulose and nitrocellulose reduction of nitrobenzene with metals to aniline or hydrazobenzene bauxite digestion production of salicylic acid from dry sodium phenolate (Kolbe – Schmitt process) hydrolysis of calcium cyanamide to cyanamide production of BF_3 from B_2O_3, CaF_2, and H_2SO_4 production of alkylaluminums from aluminum, olefin, and hydrogen production of tetraethyllead
Cascade of stirred tanks	for low reaction rates and high final conversions	apatite digestion
Tank with liquid recirculation	semicontinuous operation with solids fixed in tank and liquid recirculating	cellulose digestion production of ammonium sulfate from ammonium carbonate and gypsum
Rotary drum	for batch operation, high solids content	production of cellulose acetate and cellulose ethers production of AlF_3 by wet process
Fluidized-bed reactor	Semicontinuous operation intensive liquid circulation	water treatment
Steeping press	combination of reaction and liquid separation batch operation	production of cellulose ether
Kneader	used for highly viscous media	production of nitrocellulose, cellulose ether, and cellulose acetate
	for batch operation	production of celluloid from nitrocellulose production of superphosphate
Screw-conveyor reactor	used for highly viscous media batch operation	digestion of rutile or ilmenite with H_2SO_4
Multiple-hearth reactor	continuous operation long solids residence time	production of acetylene from carbide (dry gas generator)
Rotary kiln	direct heating for high reaction temperatures	digestion of fluorspar or phosphate with H_2SO_4reducing decomposition of H_2SO_4 in presence of carbon

Table 13. Reactors for noncatalytic solid-phase reactions

Reactor type	Features	Examples of applications
Shaft reactor	see Table 11	metallurgical processes, e.g., powder boriding of iron-based materials direct reduction of iron ores with carbon (Kinglor – Metor process)
Multiple-hearth reactor	see Table 11	calcination
Rotary kiln	see Table 11	cement production burning of lime, dolomite, gypsum, and magnesite calcination thermal decomposition of $FeSO_4$ and $BaCO_3$ reduction of barite with carbon to BaS reduction of ores with carbon (e.g., to ZnO)
Fluidized-bed reactor	see Table 11	burning of lime (multistage)

ture pyrolysis (→ Plasma Reactions, Chap. 2.1.; → Metallurgical Furnaces, Chap. 5.5.).

Equipment used for *solid reactions* includes arc and resistance-heated reduction furnaces and the Acheson furnace (→ Metallurgical Furnaces, Chap. 5.2., → Metallurgical Furnaces, Chap. 5.3.). The Acheson furnace is a resistance-heated device for pure solid – solid reactions; that is, in contrast to other processes, no melting of the solid charge occurs. All electrothermal processes are characterized by very high equipment cost and high electric power consumption. The prerequisite for their economical operation is a low unit price for energy.

This group of reactors and their applications are summarized in Table 14 and Figure 14.

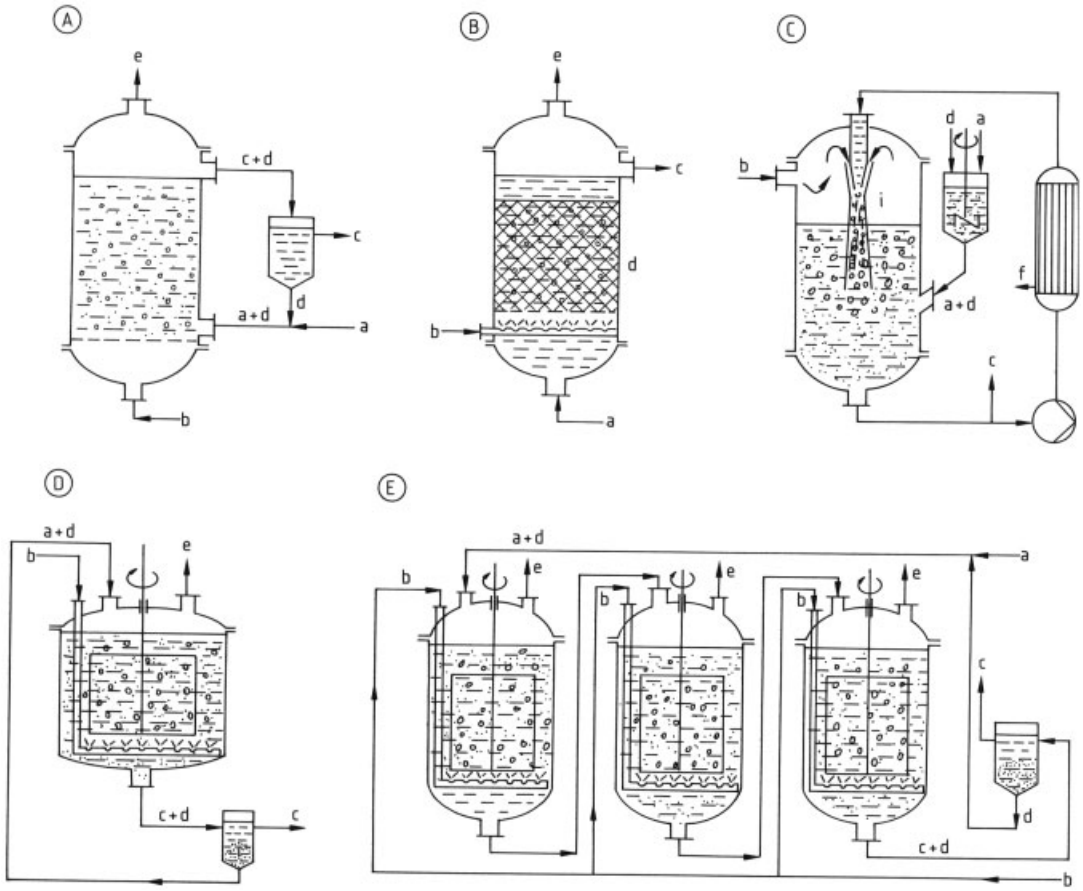

Figure 11. Suspended-bed and fluidized-bed reactors for liquid-phase and gas – liquid reactions over solid catalysts
A) Bubble column with suspended catalyst; B) Fluidized-bed reactor; C) Buss loop reactor; D) Sparged stirred tank with suspended catalyst; E) Cascade of sparged stirred tanks with suspended catalyst
a) Liquid feed components; b) Gaseous feed components; c) Liquid product; d) Catalyst; e) Off-gas; f) Heating agent or coolant; g) Heat exchanger; h) Pump; i) Reaction mixer with mixing nozzle

3.7. Reactors for Electrochemical Processes (→ Electrochemistry; → Metallurgical Furnaces, Chap. 5.7.)

In electrochemical reactions, electrons are supplied to a reactant in the electrolyte or re-moved from it with the aid of an electric current. A minimum voltage called the decomposition voltage must be applied to the electrodes for this purpose. In addition to the electrochemical reactions occurring on the electrode surface, transport processes and chemical reactions in the electrolyte bath are important.

Electrochemical processes have the following advantages:

1) High product purity (no secondary reactions)
2) Low reaction temperature (except for fused-salt electrolysis)
3) Easy control of reaction rate through variation of electrode voltage

They have the following disadvantages:

1) High energy losses in the system
2) Large space requirements
3) High investment costs

For these reasons, electrochemical processes are used only when no available thermal or catalytic process can accomplish the same purpose, which is especially true in the production of chlorine, aluminum, and copper. A survey of important applications for electrolytic processes is given in the following:

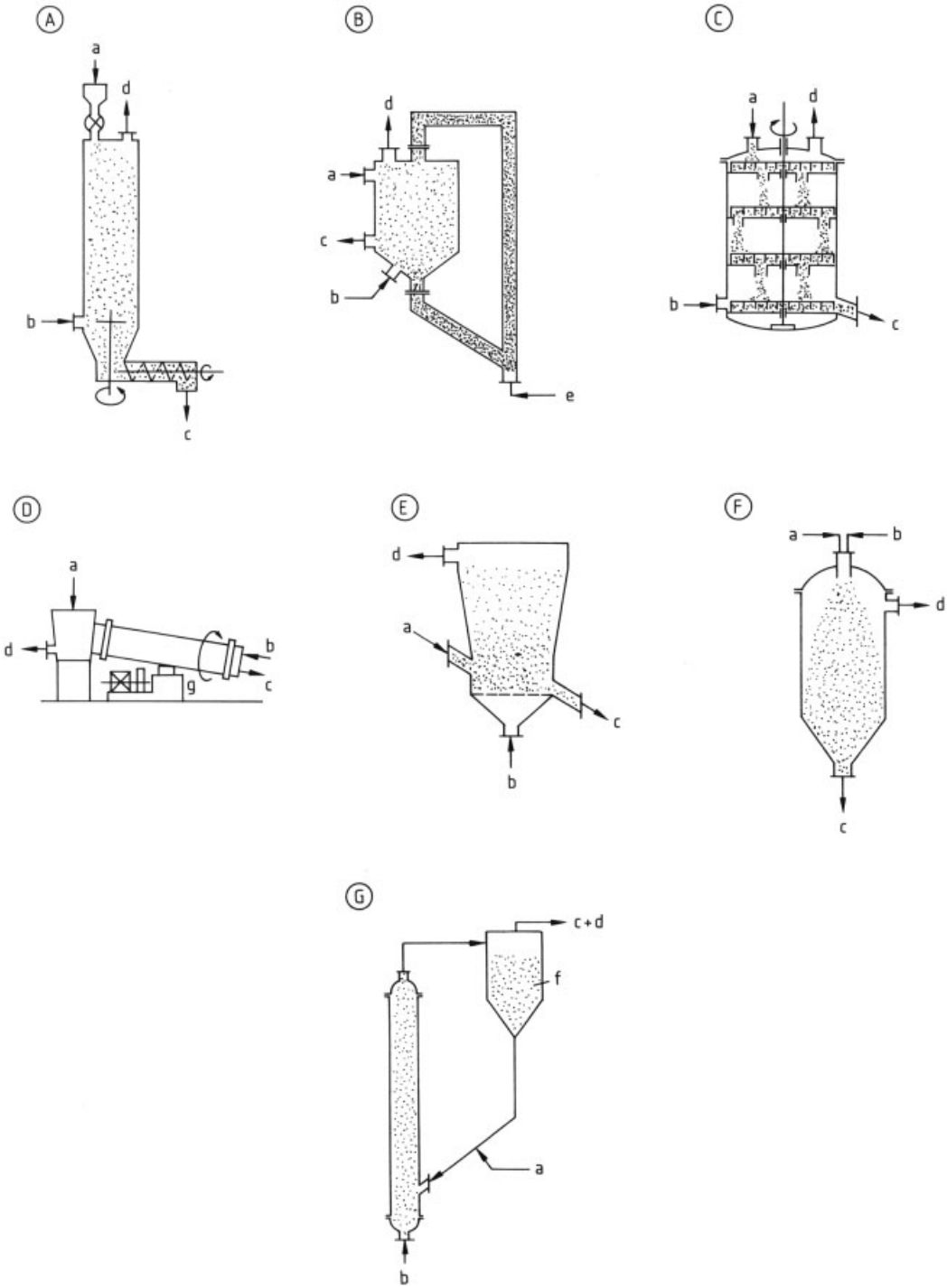

Figure 12. Reactors for noncatalytic gas – solid reactions
A) Shaft kiln; B) Moving-bed reactor; C) Multiple-hearth reactor; D) Rotary kiln; E) Fluidized-bed reactor; F) Spray reactor; G) Entrained-flow reactor
a) Solid feed components; b) Gaseous feed components; c) Solid product; d) Off-gas; e) Air; f) Cyclone; g) Drive unit

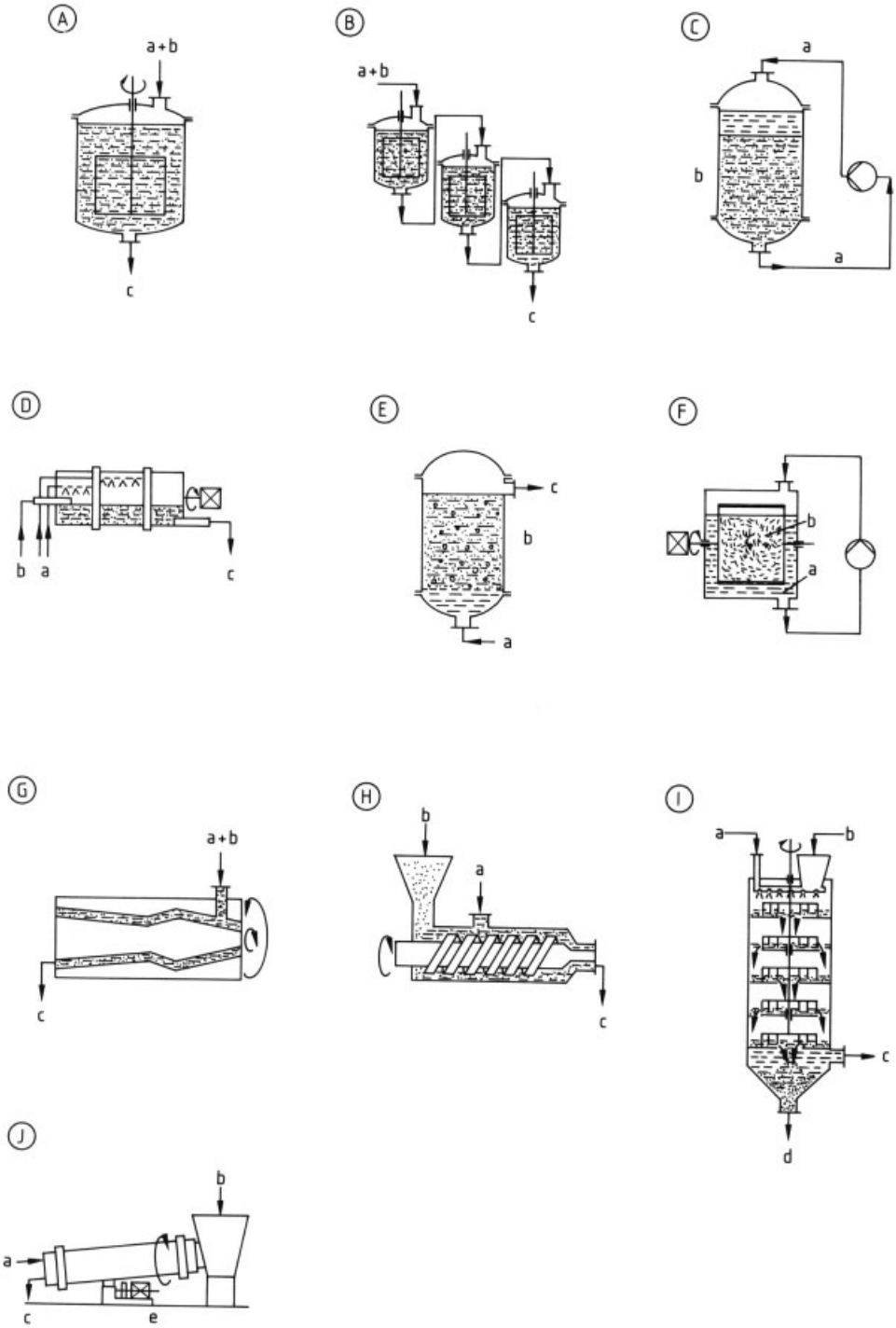

Figure 13. Reactors for noncatalytic liquid – solid reactions
A) Stirred tank; B) Cascade of stirred tanks; C) Tank with liquid recirculation; D) Rotary drum; E) Fluidized-bed reactor; F) Steeping press; G) Kneader; H) Screw-conveyor reactor; I) Multiple-hearth reactor; J) Rotary kiln
a) Liquid feed components; b) Solid feed components; c) Liquid product; d) Solid product; e) Drive unit

Table 14. Electrothermal reactors

Reactor type	Features	Examples of applications
Plasma torch	very high temperature for plasma generation very short residence times large product spectrum	high-temperature pyrolysis for the production of acetylene and ethylene are furnace, Hüls d.c. torch, Hüls a.c. plasma torch, Höchst
Fluohm reactor	resistance heating with graphite fluidized bed	production of hydrogen cyanide (Shawinigan process)
Reduction arc furnace	very high reaction temperatures continuous raw material feed and periodic discharge or batch process	production of calcium carbide production of phosphorus decomposition of $BaCO_3$ production of corundum metallurgical processes (e.g., production of ferroalloys)
Reduction resistance furnace	for low temperatures and lower specific electrical power than the arc furnace	production of glass melts metallurgical processes (e.g., reduction of metal oxides) chlorinating digestion of rare-earth ores
Acheson furnace	reaction takes place in the solid phase batch process very long residence times	production of graphite production of silicon carbide
Reactor with indirect electric heating		azotization of carbide in a furnace with heating elements production of $MgCl_2$ from MgO in a blast furnace with electric heating

Chlorine production by chlor–alkali electrolysis
- Mercury amalgam process
- Diaphragm-cell process
- Membrane process

Metal winning by fused-salt electrolysis
- Aluminum
- Magnesium
- Sodium

Metal refining
- Copper
- Nickel

Electrolysis of inorganic materials
- Electrolysis of water
- Fluorine production by electrolysis of hydrogen fluoride
- Production of sodium chlorate by electrolysis of sodium chloride
- Electrochemical oxidation of sodium chlorate to perchlorate
- Recovery of persulfuric acid
- Production of ozone

Electrolysis of organic materials
- Production of adiponitrile from acrylonitrile
- Production of dimethyl sebacate
- Reduction of nitrobenzene to aniline
- Production of perfluorocaprylic acid
- Production of dihydrostreptomycin

The design of the reaction system (i.e., cell geometry and flow configuration), the electrode arrangement and material, and control of phases and concentrations are highly process specific.

Typical designs are illustrated in Figure 15.

3.8. Reactors for Biochemical Processes

(→ Biochemical Engineering;
→ Biotechnology)

Some important biochemical processes, such as those used in making beer, wine, alcohol, and baker's yeast, have been known for centuries. Typical of these reactions is their use of enzymes as biocatalysts. The enzymes can be present as cell constituents of living microorganisms, or they can be isolated in dissolved form or bound to inert supports (→ Immobilized Biocatalysts). The prerequisite for the use of live microorganisms is the provision of favorable living conditions. Such conditions include the presence of optimal amounts of nutrients and oxygen (in aerobic processes); maintenance of the temperature, pressure, maintenance of pH in certain ranges, and sterile conditions.

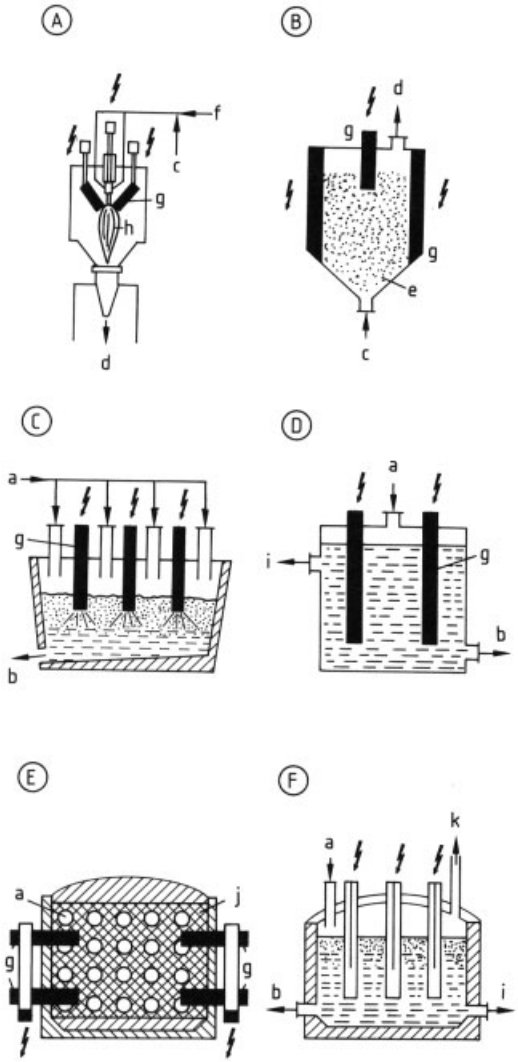

Figure 14. Reactors for electrothermal processes
A) Plasma torch; B) Fluohm reactor; C) Arc-heated reduction furnace; D) Resistance-heated reduction furnace; E) Acheson furnace; F) Reactor with indirect electric heating
a) Solids; b) Molten product; c) Gaseous reaction mixture; d) Gaseous product; e) Catalyst; f) Carrier gas; g) Electrodes; h) Plasma; i) Slag; j) Resistive charge; k) Off-gas

In addition to these factors, metabolism is important for reactor design. Aerobic processes require an adequate supply of oxygen. In anaerobic processes, the admission of gas from outside must be prevented; gases and solvent vapors resulting from the reaction must also be removed from the reactor.

Reactors for these processes can be classified as follows:

1) Reactors with dissolved or suspended biocatalysts (submerged processes) for aerobic or anaerobic conditions
2) Reactors with immobilized biocatalysts for aerobic or anaerobic conditions

Reactors for use in *submerged aerobic processes* have provisions for efficient aeration and intensive liquid circulation. Aeration is accomplished with fixed or moving distributors, nozzles, or submerged or rotating jets. Liquid circulation is ensured by various stirring systems or by forced or natural convection.

A summary of the most important reactor types and their applications is given in Table 15 and Figure 16.

Reactors for *anaerobic conditions* do not have aeration equipment. Usually, sealed vessels with or without stirrers are used (fermenters). Applications of these reactor types include fermentation processes (e.g., lactic acid fermentation, alcohol production, mash fermentation).

The *immobilization of enzymes* on suitable supports enables the use of reactor designs similar to those for heterogeneous catalytic processes. If the enzymes are supported on semipermeable membranes, separation and reaction can be combined in membrane reactors.

Reactors with immobilized biocatalysts, together with their applications, are listed in Table 16 and Figure 17.

3.9. Reactors for Photochemical and Radiochemical Processes

The photochemical and radiochemical principles are used to a very limited extent in industry because conditions for economical operation (e.g., high quantum efficiency) are seldom met.

3.9.1. Photochemical Reactors
(→ Photochemistry, Chap. 3.)

The rate of a photochemical reaction is determined by the concentration of reactants and by the intensity, quantity, and wavelength of light supplied. Light in the wavelength range that is absorbed by the reaction mixture can be formally

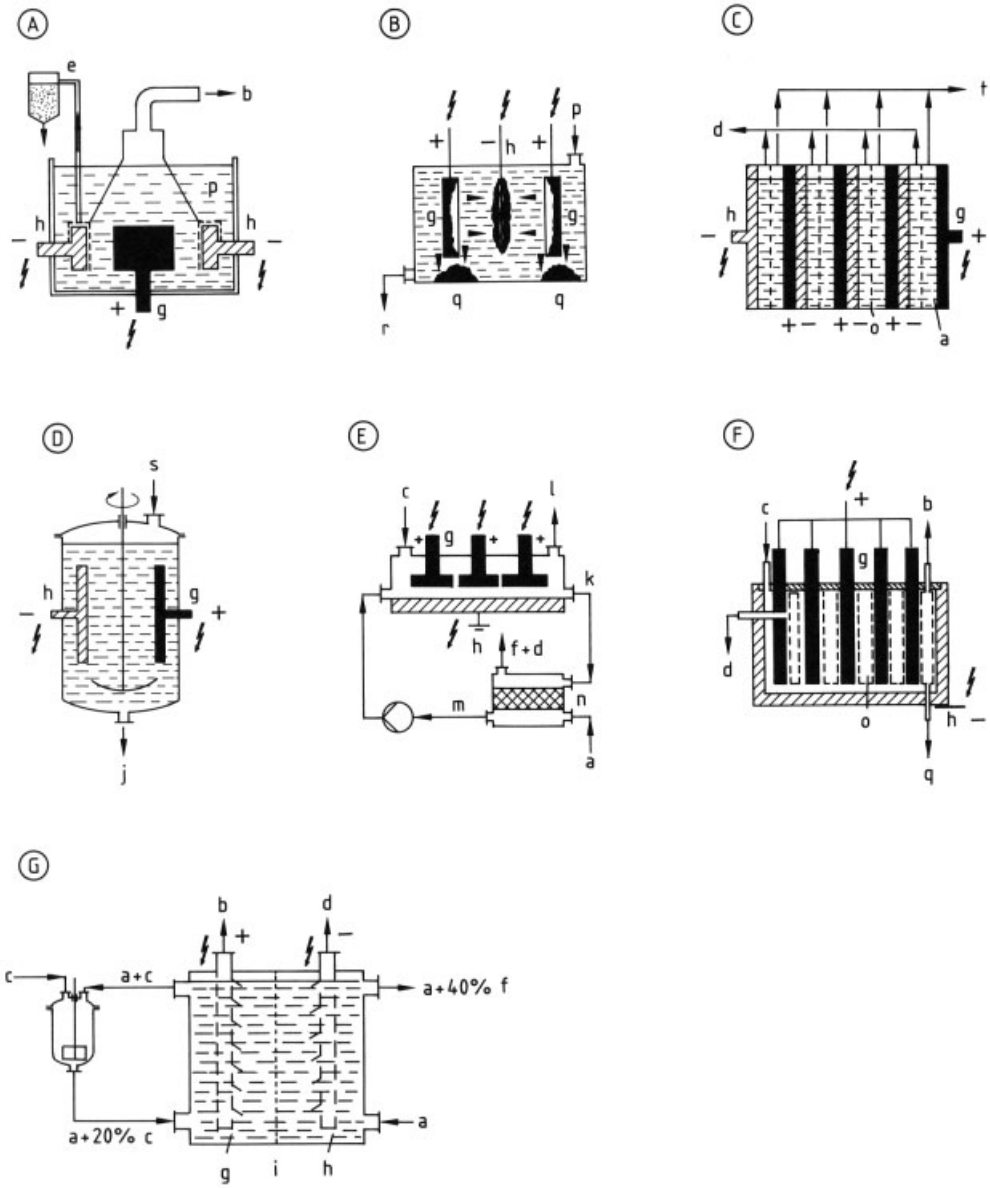

Figure 15. Reactors for electrochemical processes
A) Metal winning by fused-salt electrolysis; B) Electrolytic metal refining; C) Electrolysis of inorganic material; D) Electrolysis of organic material; E) Mercury amalgam process; F) Diaphragm-cell process; G) Membrane process
a) Water; b) Chlorine; c) Sodium chloride; d) Hydrogen; e) Sodium; f) Sodium hydroxide; g) Anode; h) Cathode; i) Membrane; j) Product; k) Amalgam; l) Recycle brine + chlorine; m) Mercury; n) Graphite; o) Diaphragm; p) Electrolytic salt solution of metal to be refined; q) Anode slime; r) Electrolyte removal; s) Organic feed solution; t) Oxygen

treated as a reactant. As a consequence, photochemical reactions exhibit a position dependence of the reaction rate, even with complete mixing, because the flux density of light quanta decreases with increasing distance from the light source. The feasible thickness of the reaction space, and thus the type and size of reactor that can be used, depend not only on the power of the emitter, but also on the optical properties of the reactor material and the reaction medium. In-

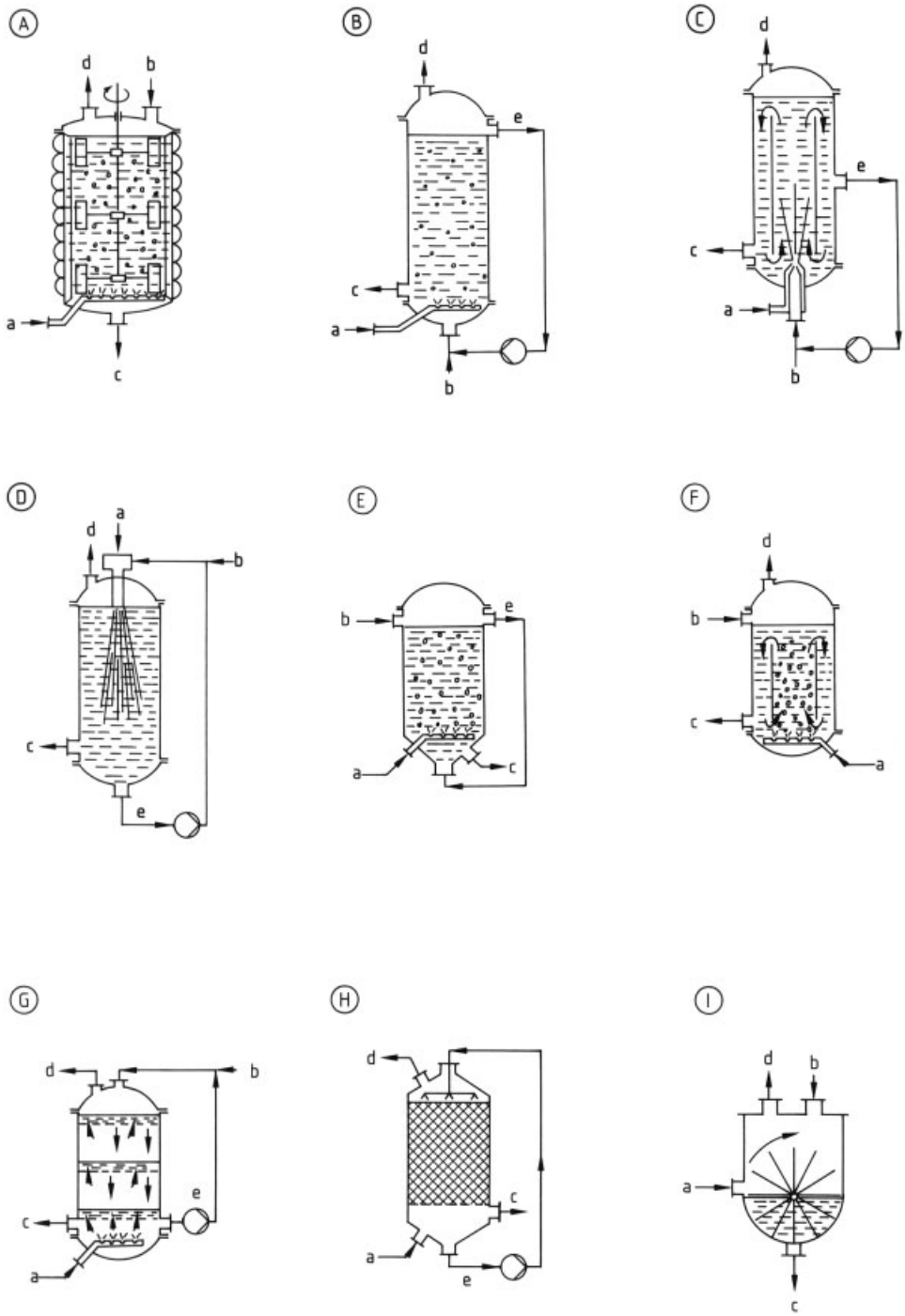

Figure 16. Reactors for submerged aerobic processes
A) Sparged stirred tank; B) Bubble column with forced circulation; C) Jet reactor with forced circulation; D) Submerged-jet reactor with forced circulation; E) Bubble column with natural circulation; F) Loop reactor; G) Sieve-tray tower; H) Trickle-bed reactor; I) Reactor with rotating internals
a) Gas; b) Fermentation medium; c) Product; d) Off-gas; e) Recycle stream

Table 15. Reactors for submerged aerobic processes

Reactor type	Features	Examples of applications
Sparged stirred tank	various stirring and circulation apparatus suitable for higher viscosities	production of antibiotics amino acids yeast
Reactors with forced circulation		
Bubble column	very broad residence-time distribution good dispersion properties	production of yeast aerobic wastewater treatment
Jet reactor	free jet, jet nozzle, or central tube designs possible for low viscosities high gas velocities, good mass transfer	
Submerged-jet reactor	very broad residence-time distribution good mass transfer danger of slime settling out	processing of spent sulfite liquor fermentation of waste substrates
Reactors with natural circulation		
Bubble column	much backmixing, broad residence-time distribution for low viscosities simple construction	production of biomass citric acid
Loop reactor	for low viscosities little dispersive action	
Sieve-tray tower	good mass transfer due to fine bubble structures	
Surface reactors		
Trickle-bed reactor	low mass-transfer coefficients and negligible dispersive action	production of acetic acid aerobic wastewater treatment
Reactor with rotating internals	use of paddles, cylinders, etc. suitable for viscous media	aerobic wastewater treatment

Table 16. Reactors for biochemical processes over immobilized biocatalysts (for aerobic and anaerobic conditions)

Reactor type	Features	Examples of applications
Fixed-bed reactor	high conversions possible high catalyst density danger of uneven flow through system combined with poor mass transfer high mechanical stress on catalyst large pressure drop simple design	production of amino acids fructose ethanol wastewater treatment
Stirred tank with suspended catalyst	continuous or batch operation good conditions for external mass transport only partial conversion achievable mechanical stress on catalyst grains due to shear forces possibility of continuous catalyst replacement	production of fructose
Fluidized-bed reactor	good conditions for external mass transport possibility of continuous catalyst replacement low catalyst concentration high flow velocities necessary	production of enzymes ethanol wastewater treatment
Membrane reactor	various membrane configurations possible catalyst separated from substrate capacity limited by transport resistance of membranes aging and fouling of membranes large pressure drop	production of enzymes pharmaceuticals amino acids

tensive mixing must be ensured, especially for thick beds. Light can be supplied from outside (through the reactor wall) or by submerged light sources. When high-power light sources are used a large amount of heat is evolved and supplemental cooling devices must be employed.

A survey of reactor types and their industrial applications appears in Table 17 and Figure 18.

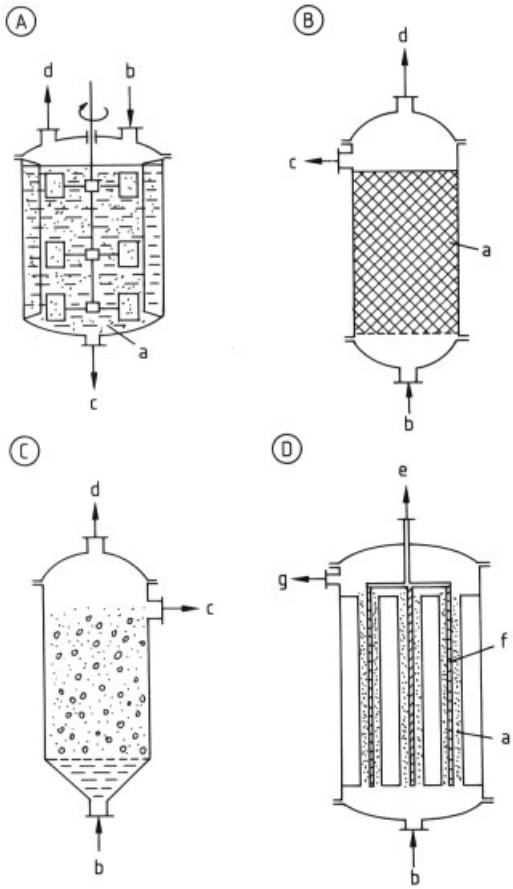

Figure 17. Reactors for biochemical processes over immobilized biocatalysts (for aerobic and anaerobic conditions)
A) Stirred tank with suspended catalyst; B) Fixed-bed reactor; C) Fluidized-bed reactor; D) Membrane reactor
a) Biocatalyst; b) Fermentation medium; c) Product; d) Off-gas; e) Permeate; f) Membrane tube; g) Retentate

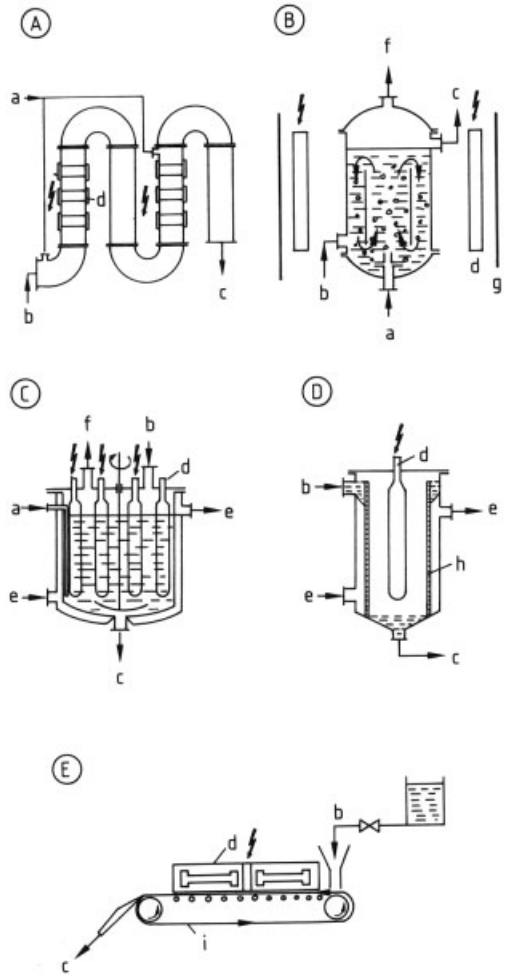

Figure 18. Reactors for photochemical processes
A) Tubular reactor; B) Bubble column; C) Stirred tank; D) Falling-film reactor; E) Belt reactor
a) Gaseous feed components; b) Liquid feed components; c) Product; d) Emitter; e) Coolant; f) Off-gas; g) External reflector; h) Falling film; i) Belt

3.9.2. Radiochemical Reactors (→ Radiation Chemistry)

Radiochemical reactions are induced by the action of ionizing radiation. In addition to high energy consumption, the extremely complex design of radiation sources and shielding works against the wider use of this reaction principle.

The following are known applications:

1) Production of ethyl bromide (Dow process, Fig. 19)
2) Radiative cross-linking of poly(vinyl chloride) and polyethylene
3) Production of alkyltin compounds
4) Degradation of polymers

Some reactions, such as chlorinations, can be implemented in either photochemical or radiochemical form.

Table 17. Reactors for photochemical processes

Reactor type	Features	Examples of applications
Tubular reactor	for homogeneous gas- and liquid-phase reactions	chlorination of benzene to hexachlorocyclohexane sulfochlorination chlorination of methane to dichloromethane
Bubble column	requires favorable optical conditions and low viscosity also used in cascades and with central tube	sulfochlorination of paraffins (cascade) side-chain chlorination of aromatics production of dodecanethiol from 1-dodecene and H_2S
Stirred tank	optically induced differences in reaction rate equalized by intensive stirring	oximation of cyclohexane with nitrosyl chloride production of provitamin D_3
Falling-film reactor	suitable for poor optical conditions because film is very thin	production of vitamin D_2
Belt reactor	especially for highly viscous media	polymerization to PAN, PAC, PVC, PVAC ∗

∗ PAN = polyacrylonitrile; PAC = polyacrylate; PVC = poly(vinyl chloride); PVAC = poly(vinyl acetate).

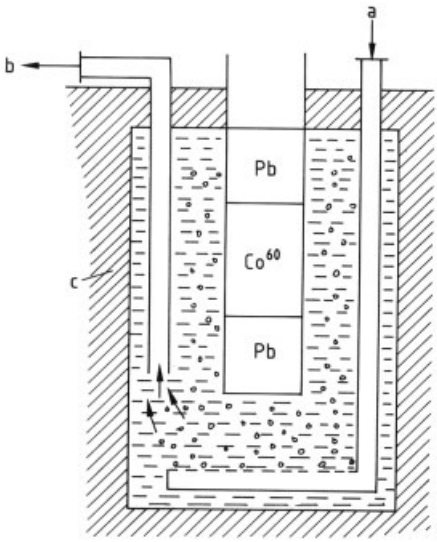

Figure 19. A reactor for a radiochemical process (production of ethyl bromide by the Dow process)
a) Gaseous reaction mixture; b) Liquid product; c) Shielding

4. References

1. "Chemische Reaktoren-Ausrüstungen und ihre Berechnung," *Verfahrenstechnische Berechnungsmethoden,* **part 5,** VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1981.
2. H. Gerrens: "Über die Auswahl von Polymerisationsreaktoren," *Chem. Ing. Tech.* **52** (1980) 477 – 488.
3. K. H. Reichert, W. Geiseler (eds.): *Polymer Reaction Engineering,* VCH Verlagsgesellschaft, Weinheim 1989.
4. W.-D. Deckwer: "Bioreaktoren," *Chem. Ing. Tech.* **60** (1988) 583 – 590.
5. K. Schügerl: "Characteristic Features of Bioreactors," *Bioreaction Engineering,* **vol. 2,** John Wiley and Sons, New York 1990.
6. A. Heger: *Technologie der Strahlenchemie von Polymeren,* Carl Hanser Verlag, München 1990.

Model Reactors and Their Design Equations

VLADIMIR HLAVACEK, Laboratory for Ceramic and Reaction Engineering, Department of Chemical Engineering, University of Buffalo, Buffalo, NY 14260, United States

JAN A. PUSZYNSKI, Laboratory for Ceramic and Reaction Engineering, Department of Chemical Engineering, University of Buffalo, Buffalo, NY 14260, United States

HENDRIK J. VILJOEN, Laboratory for Ceramic and Reaction Engineering, Department of Chemical Engineering, University of Buffalo, Buffalo, NY 14260, United States

JORGE E. GATICA, Laboratory for Ceramic and Reaction Engineering, Department of Chemical Engineering, University of Buffalo, Buffalo, NY 14260, United States

1. Introduction 3
2. Batch Reactors 4
2.1. Homogeneous Systems 4
2.1.1. Isothermal Reactors 4
2.1.2. Nonisothermal Reactors 4
2.2. Nonhomogeneous Systems 7
2.2.1. Gas – Liquid Reactions 7
2.2.2. Solid – Solid Reactions 8
2.2.3. Solid – Gas Reactions 10
3. Continuous Stirred-Tank Reactors (CSTR) 12
3.1. Isothermal Homogeneous System . 12
3.2. Isothermal Heterogeneous Reactors 16
3.3. Nonisothermal Continuous Stirred-Tank Reactors 16
3.4. Cascade of Tank Reactors 19
4. Packed-Bed Reactors 21
4.1. Mass and Energy Balances 21
4.1.1. The Single-Particle Model 22
4.1.2. The Two-Phase Model 23
4.1.3. The One-Phase Model 23
4.1.4. Boundary Conditions 24
4.2. Values of the Parameters 25
4.2.1. Average Bed Porosity 25
4.2.2. Effective Transport Coefficients 26
4.2.3. Wall Heat-Transfer Coefficient 27
4.3. Further Simplifications 27
4.4. Parametric Sensitivity 29
4.5. Flow Field Description 31
4.5.1. Boundary Conditions 32
4.5.2. Permeability and Inertia Coefficient . 32
4.5.3. Thermal-Expansion Coefficient 32
4.5.4. Mass-Expansion Coefficient 33
4.6. Thermomechanical Effects in the Reaction System 33
4.7. Numerical Simulation 34
4.7.1. Discretization of the Physical Domain 34
4.7.2. Discretization of the Governing Equations 35
5. Optimization of Chemical Reactors 36
5.1. The Objective Function for Chemical Reactors 36
5.2. Elementary Optimization Problems 37
5.2.1. Optimization of a Batch System 37
5.2.2. Optimization of a Continuous System (CSTR) 38
5.2.3. Optimization of Complex Systems: Illustrative Example 38
5.3. Optimization of Reactors by the Search Method 38
5.4. Optimum Temperature Profile and Optimization of Multistage Reactors 41
5.5. Optimization of a Multibed Adiabatic Reactor with Heat Exchange Between Catalytic Stages 42
6. References 43

Symbols

(see also → Principles of Chemical Reaction Engineering)

a	activity
a	thermal diffusivity, m^2/s
a	interfacial area per unit volume, m^2/m^3
A	heat transfer area, m^2
$B, B*$	dimensionless adiabatic temperature rise
c	concentration, $kmol/m^3$
c_J	molar concentration of species J, $kmol/m^3$
c_p, c_v	heat capacity per unit mass at constant pressure and volume, respectively, $J\,kg^{-1}\,{}^\circ C^{-1}$

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_121

c_{pm}	specific heat capacity of the cooling/heating medium
C	reduced concentration, c/c_0
C_p, C_v	molar heat capacity at constant pressure and volume, respectively, $\mathrm{J\,kmol^{-1}\,{}^\circ C^{-1}}$
d	characteristic dimension, diameter, m
d_I	diameter of impeller
D	diameter of reactor, m
D	coefficient of molecular diffusion, $\mathrm{m^2/s}$
e	particle emissivity
E_a	activation energy
$E(t)$	residence time distribution function or frequency function, $\mathrm{s^{-1}}$
F	molar flow rate
$\mathbf{g}$	gravitational acceleration, $\mathrm{m/s^2}$
h	overall heat-transfer coefficient, $\mathrm{W\,m^{-2}\,K^{-1}}$
H	enthalpy, J
H_J	molar enthalpy of species J, J/kmol
ΔH_r	heat of reaction at constant pressure as associated with the stoichiometric equation
$(\Delta H_\mathrm{r})_\mathrm{J}$	heat of reaction at constant pressure for the conversion of or to one molar unit of J, J/kmol
J	component
J	molar flux, $\mathrm{kmol\,m^{-2}\,s^{-1}}$
J_J	molar flux of species J, $\mathrm{kmol\,m^{-2}\,s^{-1}}$
k	homogeneous reaction velocity constant, dimension depends on the kinetics; for nth order: $\mathrm{m^{3(n-1)}/(kmol^{(n-1)}\,s)}$
K	chemical equilibrium constant
K_D	ratio between two concentrations
K_L	mass-transfer coefficient, m/s
L	length, m
m	total mass of a system, kg
m_J	mass of species J in a system, kg
M_J	molar mass of species J, kg/kmol
n	sequence number of a tank reactor in a cascade or of a bed in a multibed reactor
$\mathbf{n}$	outward unit vector on S
N	number of revolutions (impeller) per unit time
N_QI	dimensionless constant
Nu	$\alpha d/\lambda$; Nusselt number
p	total pressure, Pa
P	permeability
ΔP_r	pressure drop across the reactor
P_e	total pressure in the bulk stream
Pe	$\langle v\rangle L/D_1$; Péclet number for longitudinal dispersion
Pe_h	$\langle v\rangle L\,\varrho\,c_p/\lambda$; Péclet number for heat dispersion
Pe_t	$\langle v\rangle d_\mathrm{t}/D_\mathrm{t}$; Péclet number for transverse dispersion
Pr	$\mu\,c_p/\lambda$; Prandtl number
Q	amount of heat, J
$\dot{Q}$	heat flow, W
r	cylindrical or spherical coordinate, m
R	rate of reaction
R	gas constant, $8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}$
R_J	molar rate of production of J per unit volume of the reaction phase, $\mathrm{kmol\,m^{-2}\,s^{-1}}$
$\text{—}R\text{—}_\mathrm{J}$	molar rate of conversion per unit volume of the reaction phase, $\mathrm{kmol\,m^{-3}\,s^{-1}}$
Re	$v\,d/\nu$; Reynolds number
Re_p	$\langle v\rangle\,d_\mathrm{p}/\nu$; Reynolds number in packed bed related to particle diameter
s	ratio of reaction rate constant
S	surface, $\mathrm{m^2}$
Sc	ν/D; Schmidt number
St	$\alpha/\varrho\,c_p\langle v\rangle$; Stanton number for heat transfer
St	$K_\mathrm{G}/\langle v\rangle$ or $K_\mathrm{L}/\langle v\rangle$; Stanton number for mass transfer
St'	$UA/\varrho\,c_p\,\Phi_v$; modified Stanton number
t	time, s
t'	off-line time, s
$t*$	reaction time, s
T	temperature, K or °C
$T*$	optimum reaction temperature, K or °C
$\langle T\rangle$	average radial temperature for a two dimensional model
ΔT_ad	adiabatic temperature rise of a reaction mixture after complete conversion, °C
T^e	equilibrium temperature
u	internal energy per unit mass, J/kg
$\langle u\rangle$	average internal energy (over the reactor volume), J/kg
$\mathbf{u}$	linear velocity, m/s
$\mathbf{u_s}$	control volume velocity, m/s
U_J	molar internal energy of species J, J/kmol
ΔU_r	heat of reaction at constant volume as associated with a stoichiometric equation, J
$(\Delta U_\mathrm{r})_\mathrm{J}$	heat of reaction at constant reaction volume for the conversion of or to one molar unit of J, J/kmol

v	velocity, m/s
v_{F}	front velocity, m/s
v_{r}	relative velocity; velocity of one phase relative to another phase, m/s
V	volume, m^3
V_{r}	volume of reaction mixture, m^3
V_{R}	reactor volume, m^3
w_{J}	mass fraction of species J
W	mass of the catalyst
W	amount of work, J
$\dot{W}$	rate of work done on the surroundings, W
x	direction of propagation, m
x_{J}	mole fraction of species J in any phase
$\mathbf{x}$	coordinate, m
y_{J}	mole fraction of species J in any gas phase only used in connection with x_{J} for the liquid phase
z	flow direction, m
Z	dimensionless coordinate in the direction of flow, z/L
α	heat-transfer coefficient, $\mathrm{W\,m^{-2}\,K^{-1}}$
$\bar{\beta}$	dimensionless heat-transfer coefficient
γ	ratio between kinetic and mass-transfer rates
γ	temperature factor E/RT_{a}; index a varies, depending on situation
$\bar{\gamma}$	dimensionless activation energy
$\bar{\delta}$	dimensionless parameter
$\boldsymbol{\delta}$	displacement vector
ε	porosity
$\bar{\varepsilon}$	dimensionless parameter
ϑ	dimensionless temperature
Θ	dimensionless time
Θ_{c}	dimensionless cooling temperature
ζ	relative degree of conversion, fraction of reactant converted
$\bar{\zeta}$	average conversion
κ	ratio between reaction rate constant
$\boldsymbol{\lambda}_e$	effective thermal conductivity, $\mathrm{W\,m^{-1}\,K^{-1}}$
λ	thermal conductivity, $\mathrm{W\,m^{-1}\,K^{-1}}$
λ	Lamé constant
μ	Lamé constant
μ	dynamic viscosity, $\mathrm{Ns/m^2}$
ν	kinematic viscosity, $\mathrm{m^2/s}$
ν	stoichiometric coefficient
ξ_{J}	degree of conversion of species J
η	effectiveness factor
ϱ	density, specific mass, $\mathrm{kg/m^3}$
σ	Stefan – Boltzmann constant, $5.67\times10^{-8}\,\mathrm{J\,m^{-2}\,K^{-4}\,s^{-1}}$
τ	average residence time in reactor system, s
Φ	flow rate
Φ_{P}	production rate, kg/s; kmol/s
Φ_{mJ}	mass flow rate for species J, kg/s
Φ_{c}	circulation rate
Φ_v	volumetric flow rate, m^3/s
ψ	shape factor

Subcripts

0	initial
av	average
A, B	reactants
c	coolant, continuous
e	equilibrium
ext	extern
f	fluid
fus	fusion
F	fluid phase
g	gas
G	gas phase
i	interface
ign	ignition
J	arbitrary species
l	liquid, longitudinal
L	liquid phase
L	at the outlet of a tubular reactor of length L
m	medium, mass
n	output of the nth tank reactor in a cascade, or of a segment n
r	reaction, radial direction
s	solid
S	solid phase
t	transverse
v	volumetric
w	wall
z	axial direction

1. Introduction

Reactor modeling represents a simultaneous solution and analysis of mass, energy, and momentum transfer equations along with reaction rate and equilibrium data. Sometimes additional information (e.g., of economic nature in reactor optimization calculations) must be also included in the reactor model.

Important goals of reactor modeling are [1]:

1) Scale-up of experimental laboratory units to pilot-plant size
2) Scale-up of pilot-plant units to full scale size
3) Prediction of behavior of different reactor feedstocks and new catalysts
4) Prediction of transients in reactors for better control
5) Optimization of steady-state operating conditions
6) Better understanding of the system that may lead to process and design improvements

Reactor modeling in the hands of a practitioner is a powerful tool and can result in faster, more economical process design. It should always be coupled with pilot-plant experiments and process development.

The most important parts of reactor modeling are described in this article; namely, simulation of batch and continuous stirred-tank reactors, tubular systems, and packed beds. Some elements of optimization of chemical reactors are presented as well.

2. Batch Reactors

2.1. Homogeneous Systems

2.1.1. Isothermal Reactors

Batch reactors are usually operated as closed systems with no net inflow or outflow of mass. The systems are also well-mixed and no spatial gradients exist. The *material balance* for a component J in a batch reactor which meets these conditions, is:

$$\frac{dm_{\mathrm{J}}}{\mathrm{d}t} = m\frac{\mathrm{d}\xi_{\mathrm{J}}}{\mathrm{d}t} = M_{\mathrm{J}}\, R_{\mathrm{J}} V_{\mathrm{r}} \tag{2.1}$$

where V_{r} is the volume of the reaction mixture, m the total mass of the system (kg), m_{J} the mass of component J in the system (kg), ξ_{J} the degree of conversion of J, M_{J} the molar mass of component J (kg/kmol), and R_{J} the molar rate of production of J per unit volume of the reaction phase (kmol m^{-3} s^{-1}). In the isothermal case, R_{J} is only a function of the reactants for an irreversible reaction or a function of reactants and products for a reversible reaction. In terms of the density of the reaction mixture, $\varrho = m/V_{\mathrm{r}}$, the material balance can be written in integral form (Eq. 2.2) and the reaction time t that is required for a certain degree of conversion can be obtained by integrating Equation (2.1).

$$t = \int_0^{\xi_{\mathrm{J}}} \frac{\varrho\, \mathrm{d}\xi_{\mathrm{J}}}{M_{\mathrm{J}} R_{\mathrm{J}}} \tag{2.2}$$

If the density of the reaction mixture remains constant during the course of the reaction, the integration is simplified. In the case of *irreversible nth-order reactions* integration gives

$$t = \frac{(1-\xi_{\mathrm{J}})^{1-n} - 1}{k\,(n-1)\, c_{\mathrm{J}_0}^{n-1}}, n \neq 1. \tag{2.3}$$

$$t = \frac{\ln\,(1-\xi_{\mathrm{J}})}{-k}, n = 1 \tag{2.4}$$

The rate equations of *reversible first-order reactions* with initial concentrations c_{J_0} and c_{P_0} and rate constants k_1 and k_2 for the forward and backward reactions respectively, can also be integrated:

$$t = \frac{\ln\left[(1-\xi_{\mathrm{J}})\,(k_1+k_2) - k_2 \frac{c_{\mathrm{J}_0}+c_{\mathrm{P}0}}{c_{\mathrm{J}0}}\right]}{-\,(k_1+k_2)} \tag{2.5}$$

The general problem in reactor design is to calculate the volume of a reactor for a certain average rate of production. The volume can be obtained from the required degree of conversion and the corresponding reaction times. In batch operation, a certain time is used for filling, heating, cooling, and cleaning and during this time no production occurs. If t' denotes this off-line time and Φ_{P} denotes the desired production of product P, the reactor volume is

$$V_{\mathrm{R}} = \frac{\nu_{\mathrm{A}} \Phi_{\mathrm{P}}\,(t+t')}{\nu_{\mathrm{P}} M_{\mathrm{P}} c_{\mathrm{A}_0} \xi_{\mathrm{A}}} \tag{2.6}$$

where ν_{A} and ν_{P} are the stoichiometric coefficients of reactant A and product P, respectively.

2.1.2. Nonisothermal Reactors

All chemical reactions are accompanied by evolution or absorption of heat. The most important considerations which determine the choice of temperature in the reactor are:

1) Chemical equilibrium and the conversion rate of the desired reaction; the maximum conversion of a reversible exothermic reaction is reduced when the temperature is increased, but the kinetic rate increases, therefore an optimum temperature exists.
2) Undesirable side reactions (which also depend on temperature and their role in the overall equilibrium) can influence the maximum yield of product.
3) Phase transformations of reactants or products; there are examples where it is better to carry out the reaction in the liquid instead of the gaseous phase to achieve maximum conversion and mass transfer. An example is the formation of methyl *tert*-butyl ether from isobutene and methanol.
4) Cost implications of heating reactants to the reacting temperature; energy cost must be compared with the gain in yield at increased temperature as maximum yield and maximum profitability will not always be at the same temperature [2].

Constant Pressure. At constant pressure, the energy balance in a well-mixed batch reactor is:

$$\varrho c_p \frac{\mathrm{d}T}{\mathrm{d}t} + (\Delta H_\mathrm{r})_\mathrm{J}\, R_\mathrm{J} = \frac{\dot{Q}}{V_\mathrm{r}} \tag{2.7}$$

where c_p denotes the specific heat capacity, $(\Delta H_\mathrm{r})_\mathrm{J}$ the reaction enthalpy for the conversion of or to one molar unit of J, and $\dot{Q}$ denotes the rate of heat flow to or from the reactor. In the case of adiabatic operation, $\dot{Q} = 0$. Otherwise, $\dot{Q}$ can be determined as follows:

$$\dot{Q} = hA\left(T - T_{\mathrm{c,H}}\right) \tag{2.8}$$

where h is the overall heat-transfer coefficient and A the heat-transfer area. The external temperature can be cooler (T_c) or hotter (T_H) than T, depending on whether cooling or heating must be accomplished. The overall heat-transfer coefficient h is calculated from correlations [3] and it depends on the temperature of the cooling or heating medium. Thus, a heat-transfer coefficient $h_{\mathrm{a}v}$ is calculated at the average medium temperature. If Φ_m is the mass flow rate and $c_{p\mathrm{m}}$ is the specific heat capacity of the cooling or heating medium, then h is corrected according to:

$$h = \frac{h_{\mathrm{a}v}}{1 + \frac{h_{\mathrm{a}v} A}{2\,\Phi_\mathrm{m}\, c_{p\mathrm{m}}}} \tag{2.9}$$

Example. In a batch reactor with a volume of $5\,\mathrm{m}^3$, an exothermic reaction $\mathrm{A} \rightarrow \mathrm{P}$ is carried out in the liquid phase. The density of the mixture does not change during the reaction. The rate of reaction is given by

$$R_\mathrm{A} = -k c_\mathrm{A}$$

with $k = 4 \times 10^6 \exp(-7500/T)\,\mathrm{s}^{-1}$. The heat of reaction is $(-\Delta H_\mathrm{r})_\mathrm{A} = 1.67 \times 10^6$ J/kg, the initial concentration of A is $c_{\mathrm{A}_0} = 1$ kmol/m^3, $M_\mathrm{A} = 100$ kg/kmol and $\varrho C_p = 4.2 \times 10^6\,\mathrm{J\,m^{-3}\,K^{-1}}$. The initial temperature of the reaction mixture is 298 K. The reactor is fitted with a heat-exchange coil with a surface area of $3.3\,\mathrm{m}^2$. The heat exchanger is operated with steam ($T_\mathrm{H} = 393$ K, $h = 1000\,\mathrm{W\,m^{-2}\,K^{-1}}$) and with cooling water ($T_\mathrm{c} = 290$ K, $h = 1400\,\mathrm{W\,m^{-2}\,K^{-1}}$). Filling and emptying of the reactor take 600 s and 900 s, respectively.

For a required conversion $\xi_\mathrm{A} \geq 0.95$, the temperature profile over one reaction cycle and the duration of a reaction cycle shall be determined when the following policy of operation is followed: heat up to 350 K, the reaction then proceds under isothermal conditions until $\xi_\mathrm{A} = 0.95$, followed by cooling down to 310 K.

Solution. The following equations must be solved simultaneously for the first part of the reaction cycle:

$$\frac{\mathrm{d}\xi_\mathrm{A}}{\mathrm{d}t} = k\,(1 - \xi_\mathrm{A})$$

$$\varrho c_p \frac{\mathrm{d}T}{\mathrm{d}t} = (-\Delta H_\mathrm{r})_\mathrm{A}\, c_{\mathrm{A}_0} M_\mathrm{A}\, k\,(1 - \xi_\mathrm{A}) + \frac{hA_\mathrm{s}}{V_\mathrm{r}}\,[T_\mathrm{s} - T]$$

Starting from the initial concentration and temperature, a small temperature interval ΔT is selected. Average values for T and k over this interval are then calculated and used together with the initial value for ξ in the second equation to solve for Δt. This value is then used in the first equation to solve for ξ. An average value for ξ over this interval can then be calculated. With the average values of T, k, and ξ, an improved Δt value can be calculated from the second equation and the process is repeated until the required degree

of accuracy is reached. In Figure 1 the temperature profile is shown. Once the conversion is 0.95, cooling commences. The cooling period can be calculated by integrating the following energy equation from $T = 350$ K to $T = 310$ K:

$$\varrho c_p \frac{\mathrm{d}T}{\mathrm{d}t} = (-\Delta H_\mathrm{r})_\mathrm{A}\, c_{\mathrm{A}_0} M_\mathrm{A} k\,(1-\xi_\mathrm{A}) + \frac{hA_\mathrm{c}}{V_\mathrm{r}}\,[T_\mathrm{c} - T]$$

Integrating the first two equations from the initial conditions until T is 350 K gives a preheating period of 2384 s. The temperature is then kept constant and only the first equation is integrated until the conversion reaches 0.95; this period takes 1000 s. For the cooling-off period the third equation is integrated and this gives 5426 s. Adding 1500 s for filling and emptying gives a total period of one cycle of 10 310 s.

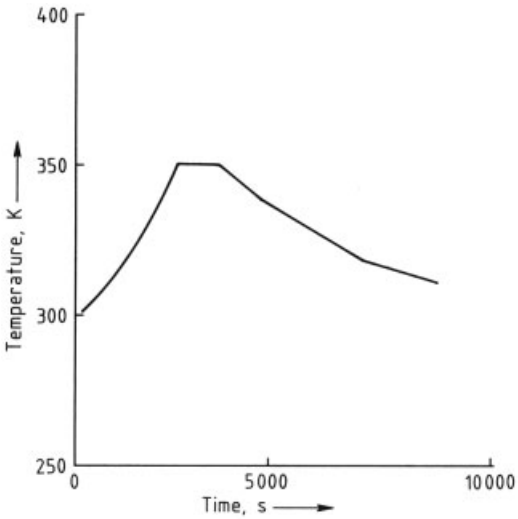

Figure 1. Temperature profile over a reaction cycle

Constant Volume. The energy balance for a constant volume process is

$$\varrho c_v \frac{\mathrm{d}T}{\mathrm{d}t} + (\Delta U_\mathrm{r})_\mathrm{J}\, R_\mathrm{J} = \frac{\dot{Q}}{V_\mathrm{r}} \tag{2.10}$$

Since ΔU_r is the change of internal energy at constant volume and temperature, for the case of an ideal gas ΔH_r and ΔU_r are related as follows:

$$\Delta H_\mathrm{r} = \Delta U_\mathrm{r} + RT\,\Delta n \tag{2.11}$$

This relation will be different for reaction mixtures with another equation of state.

Example. A closed tank with an internal volume of 250 m^3 contains a mixture of air with 4 % butane at 18 °C and a pressure of 100 kPa. The mixture is ignited and the reaction proceeds according to:

$$\mathrm{C_4H_{10}} + 6.5\,\mathrm{O_2} \rightarrow 4\,\mathrm{CO_2} + 5\,\mathrm{H_2O}$$

The heat of combustion at 18 °C and 100 kPa is $(-\Delta H_\mathrm{r}) = 2880$ kJ/mol. The pressure in the vessel shall be determined after the explosion.

Solution. Since the reaction proceeds adiabatically and at constant volume, the total internal energy of the reaction mixture does not change. The temperature rise is calculated from:

$$\int_{T_0}^{T} c_v\,\mathrm{d}T + \int_{\xi_{\mathrm{B}_0}}^{\xi_\mathrm{B}} \frac{(\Delta U_\mathrm{r})_\mathrm{B}}{M_\mathrm{B}}\,\mathrm{d}\xi_\mathrm{B} = 0$$

where the index B denotes butane.

To calculate $(\Delta U_\mathrm{r})_\mathrm{B}$, the relationship between U and H and the ideal gas law will be used:

$$\Delta H_\mathrm{r} = \Delta U_\mathrm{r} + RT\,\Delta n$$

The total change in number of moles is $(4 + 5 - 7.5 = 1.5)$, hence:

$$RT\,\Delta n = 1.5 \times 8.314\ \mathrm{J\,mol^{-1}\,K^{-1}} \times 291\ \mathrm{K} = 3.63 \times 10^3\ \mathrm{J/mol}$$

Since $\mathrm{O_2}$ is the limiting reactant (i.e., $\mathrm{O_2}$ is present in substoichiometric amounts), not all butane will react. The composition of the final mixture can then be calculated. In Table 1 the initial and final compositions are given. The specific heat capacity at constant volume (c_v) of the product mixture is a function of temperature and can be found in thermodynamic tables. In this example the average value of c_v is 1.06 kJ/kg. The mole fractions of butane before and after the reaction were 0.04 and 0.0089, respectively. These values can be converted to mass fractions by multiplying the mole fraction of each component in Table 1 by its molecular mass and normalizing. The mass fractions of butane before and after reaction are 0.077 and 0.017, respectively.

Table 1. Initial and final gas compositions

Compound	Mole fraction before reaction	Mole fraction after reaction
Butane	0.04	0.0089
Nitrogen	0.758	0.708
Oxygen	0.2028	
Carbon dioxide		0.182
Water		0.093

The temperature rise can now be determined:

$$c_v \times (T - T_0) + \left[\frac{(-\Delta H_\mathrm{r})_\mathrm{B} - RT\,\Delta n}{M_\mathrm{B}}\right] \times (\Delta\xi_\mathrm{B}) = 0$$

so:

$$\begin{aligned}&1.06\,\mathrm{kJ/kg} \times (T - 291\,\mathrm{K})\\&+\left(\frac{-2880\,\mathrm{kJ/mol} - 3.63\,\mathrm{kJ/mol}}{58\,\mathrm{g/mol}}\right)\\&\times (0.077 - 0.017) = 0\\&T = 3103\,\mathrm{K} = 2830\,^\circ\mathrm{C}\end{aligned}$$

Applying the ideal gas law, the pressure can be determined:

$$\begin{aligned}&P \times 250\,\mathrm{m}^3 = 1.046 \times n_0 \times 8.314\,\mathrm{J\,mol^{-1}\,K^{-1}}\\&\times (3103\,\mathrm{K}) = 1115.4\,\mathrm{kPa}\end{aligned}$$

where $1.046 \times n_0$ is the total number of moles present after the explosion. The last column in Table 1 is the mole fraction based on kmol/kmol original mixture and adding this column will give the relative change in the number of moles in the mixture.

2.2. Nonhomogeneous Systems

Batch systems are spatially homogeneous when the reactant mixture is well-mixed during the course of the reaction. Nonideal mixing can cause settling or separation of phases in liquid – gas mixtures, mixtures of nonmiscible liquids or liquid – solid mixtures. However, even reactants which are well-mixed, do not necessarily react in a homogeneous way. Even if a single phase is present, the system can still be spatially nonhomogeneous.

2.2.1. Gas – Liquid Reactions

In the following example, the design of a batch reactor, where both a liquid and a gas phase are present, will be demonstrated.

Example. A batch reactor must be designed for the production of 1×10^5 kg/day fats containing monounsaturated fatty acids by hydrogenation of cottonseed oil. The reaction is carried out on a nickel catalyst that is dispersed in hard stearin. Hydrogen is bubbled through the oil, thus two phases are present. The following equations describe the conversion and formation of acid moieties with diunsaturated (B), *cis*-monounsaturated (R_1), *trans*-monounsaturated (R_2), and saturated (S) aliphatic chains:

$$\begin{aligned}\frac{\mathrm{d}c_\mathrm{B}}{\mathrm{d}t} &= -(k_1 + k_2)\, c_{\mathrm{H_s}}^{0.5} c_\mathrm{B}\\ \frac{\mathrm{d}c_{\mathrm{R}_1}}{\mathrm{d}t} &= k_1 c_{\mathrm{H_s}}^{0.5} c_\mathrm{B} - k_3 c_{\mathrm{H_s}}^{0.5} c_{\mathrm{R}_1}\\ &\quad + k_4 c_{\mathrm{H_s}}^{0.5} c_{\mathrm{R}_2} - k_5 c_{\mathrm{H_s}} c_{\mathrm{R}_1}\\ \frac{\mathrm{d}c_{\mathrm{R}_2}}{\mathrm{d}t} &= k_2 c_{\mathrm{H_s}}^{0.5} c_\mathrm{B} + k_3 c_{\mathrm{H_s}}^{0.5} c_{\mathrm{R}_1}\\ &\quad - k_4 c_{\mathrm{H_s}}^{0.5} c_{\mathrm{R}_2} - k_6 c_{\mathrm{H_s}} c_{\mathrm{R}_2}\\ \frac{\mathrm{d}c_\mathrm{s}}{\mathrm{d}t} &= k_5 c_{\mathrm{H_s}} c_{\mathrm{R}_1} + k_6 c_{\mathrm{H_s}} c_{\mathrm{R}_2}\end{aligned}$$

The hydrogen concentration in the gas phase ($c_{\mathrm{H_i}}$) is in equilibrium with the hydrogen on the catalyst surface, $c_{\mathrm{H_s}}$. Since the gas is essentially pure hydrogen, $c_{\mathrm{H_i}}$ is constant. Furthermore it is assumed that the rate of consumption of adsorbed hydrogen on the catalyst surface is balanced by the rate of mass transfer from the gas phase. Let $K_\mathrm{D} = c_{\mathrm{H_s}}/c_{\mathrm{H_i}}$ denote the ratio between the two concentrations, then the following relation exists:

$$K_\mathrm{D}^{0.5} = \frac{-\left[1 + \frac{1}{s_2}\right][1 - \xi_\mathrm{B}]}{2\left(\frac{1}{\gamma} + \frac{Y_{\mathrm{R}_1}}{s_5} + \frac{Y_{\mathrm{R}_2}}{s_6}\right)} + \frac{\left[\left(1 + \frac{1}{s_2}\right)^2 (1 - \xi_\mathrm{B})^2 + \frac{4}{\gamma}\left(\frac{1}{\gamma} + \frac{Y_{\mathrm{R}_1}}{s_5} + \frac{Y_{\mathrm{R}_2}}{s_6}\right)\right]^{0.5}}{2\left(\frac{1}{\gamma} + \frac{Y_{\mathrm{R}_1}}{s_5} + \frac{Y_{\mathrm{R}_2}}{s_6}\right)}$$

where $Y_{\mathrm{R}_1} = c_{\mathrm{R}_1}/c_{\mathrm{B}_0}$, $Y_{\mathrm{R}_2} = c_{\mathrm{R}_2}/c_{\mathrm{B}_0}$, $\gamma = k_1 c_{\mathrm{B}_0}/K_\mathrm{L}\, a_v\, c_{\mathrm{H_i}}^{0.5}$, $s_2 = k_1/k_2$, $s_3 = k_1/k_3$, $s_4 = k_1/k_4$, $s_5 = k_1/k_5 c_{\mathrm{H_i}}^{0.5}$, $s_6 = k_1/k_6 c_{\mathrm{H_i}}^{0.5}$. The product of the mass-transfer coefficient and the bubble area a_v was determined experimentally as $K_\mathrm{L}\, a_v = 0.022\,\mathrm{s}^{-1}$. Whereas the s_i values denote the ratios of reaction rate constants, γ is a ratio between kinetic and mass transfer rates.

Experiments on the hydrogenation of cottonseed oil using a Rufert nickel catalyst gave the following values for the reaction rate constants [4]: $k_1 = 0.254\ \sqrt{\mathrm{L/mol}}/\mathrm{min}$, s_2= 2.89, s_3= 5.05, s_4= 15.1, s_5= 3.38, s_6= 3.38. The hydrogen concentration in the gas phase is $c_{\mathrm{H_i}}$ = 0.0129 mol/L. The feedstock composition (wt %) consists of 25.6 % S, 27 % R_1, 0.4 % R_2 and 47 % B. The molar concentration of B, c_{B} is 1.45 mol/L. The value for γ can now be calculated:

$$\gamma = \frac{0.254\ \sqrt{\mathrm{L/mol}}/\mathrm{min} \times 1.45\,\mathrm{mol/L}}{(60 \times 0.022)\ \mathrm{min}^{-1} \times 0.1136\ \sqrt{\mathrm{mol/L}}/\mathrm{min}} = 2.45$$

The reactions are irreversible, hence c_{s} does not appear in the rate expressions for components B, R_1, and R_2 and they can be integrated independently from the balancing equation for S. Time can be substituted in terms of the conversion of B:

$$\frac{\mathrm{d}Y_{\mathrm{R_1}}}{\mathrm{d}\xi_{\mathrm{B}}} = \frac{(1-\xi_{\mathrm{B}}) - \frac{Y_{\mathrm{R_1}}}{s_3} + \frac{Y_{\mathrm{R_2}}}{s_4} - \frac{Y_{\mathrm{R_1}}}{s_5} K_{\mathrm{D}}^{0.5}}{\left(1+\frac{1}{s_2}\right)(1-\xi_{\mathrm{B}})}$$

$$\frac{\mathrm{d}Y_{\mathrm{R_2}}}{\mathrm{d}\xi_{\mathrm{B}}} = \frac{\left(\frac{1-\xi_{\mathrm{B}}}{s_2}\right) - \frac{Y_{\mathrm{R_1}}}{s_3} + \frac{Y_{\mathrm{R_2}}}{s_4} - \frac{Y_{\mathrm{R_1}}}{s_5} K_{\mathrm{D}}^{0.5}}{\left(1+\frac{1}{s_2}\right)(1-\xi_{\mathrm{B}})}$$

These two equations, together with the algebraic equation for K_{D} can now be integrated, using the initial conditions $Y_{\mathrm{R_1}}$ = 27/47, $Y_{\mathrm{R_2}}$ = 0.4/47, and ξ_{B} = 0. In Figure 2 A, the total yield of monounsaturates is shown as a function of ξ_{B}. A maximum is obtained at ξ_{B} = 0.84 and the time it takes to obtain 84 % conversion of B is 115 min (see Fig. 2 B). The processing time, excluding the reaction time, is best ascerted from plant studies; for this example 2.5 h are taken. Hence the total time (conversion plus processing) is 4.42 h. Thus, five batches can be processed per day, each batch producing 20×10^3 kg. At 84 % conversion the product mixture consists of 67 kg monounsaturates per 100 kg of mixture, hence the reactor volume is:

$$V_{\mathrm{r}} = 1.05 \times \frac{100\,\mathrm{kg}}{67\,\mathrm{kg}} \times 20\,000\,\mathrm{kg} \times \frac{1}{800\,\mathrm{kg/m^3}} = 39.2\,\mathrm{m^3}$$

where the factor 1.05 is introduced because 5 % was added for head space and 800 kg/m^3 is the density of cottonseed oil. Using a height to diameter ratio of two, the specification of the vessel is 2.92×5.84 m.

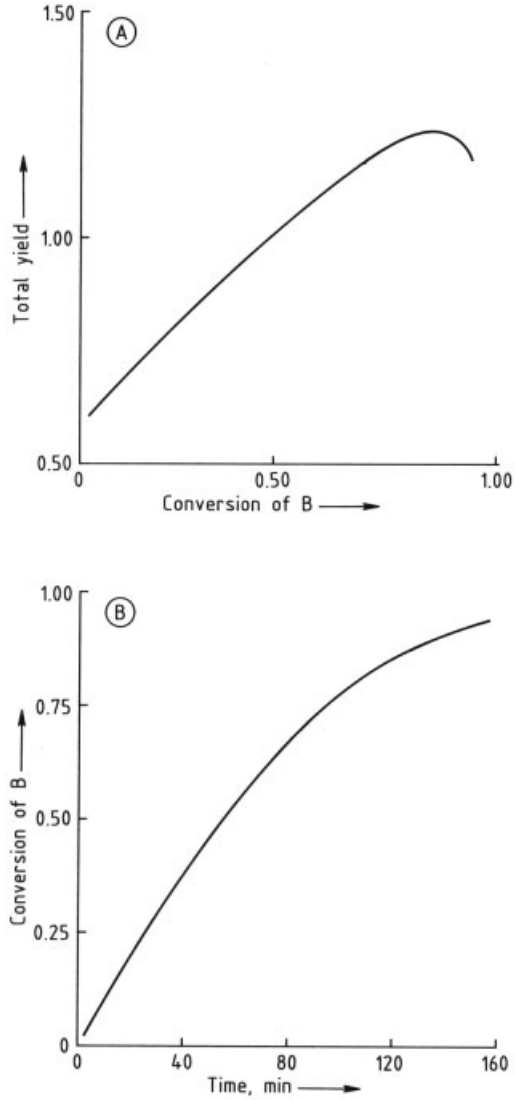

Figure 2. Total yield of monounsaturates in the hydrogenation of cottonseed oil (A) and conversion time of diunsaturates (B)

2.2.2. Solid – Solid Reactions

Even when only one phase is present, the reaction can progress in a nonhomogeneous way. Reactions with large activation energies require a certain amount of preheating before they be-

gin. If the temperature profile in a batch system is not uniform, the reaction rate will also be spatially distributed and if the heat of reaction is large, it will further contribute to the concentration of high temperatures in local areas inside the reactor.

Solid – solid reactions that are associated with large activation energies and large heats of reaction often take place in the combustion mode. The combustion mode is associated with strong spatial gradients and modeling these type of reactions requires special techniques. Examples of solid – solid reactions which occur in the combustion mode are those between Ti and C, Mo and Si, and Ta and C. The well-mixed reactants are usually prepared as a preform that is ignited at one end of the sample (see Fig. 3). Once the combustion wave has traversed the length of the sample, the reaction is complete. Conversion is high in these combustion reactions and under ideal conditions complete conversion can be accomplished. Two factors determine the degree of conversion. If the system loses a lot of heat during the reaction, the degree of conversion drops. Thus, the closer the system can be operated to adiabatic conditions, the higher will be the degree of conversion. The second factor is incomplete mixing of the reactants. In these type of reactions it is important to have intimate contact between the different species on the micro-level. Therefore, mixing of the reactants is one of the most time- consuming and expensive steps in the process.

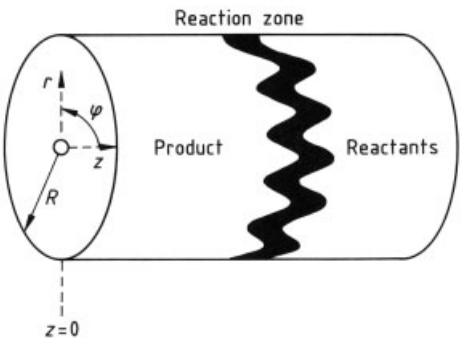

Figure 3. Typical configuration of solid – solid batch reactions

The mass and energy balances are

$$\frac{\partial c_{\mathrm{J}}}{\partial t} = -R_{\mathrm{J}} \tag{2.12}$$

$$\varrho c_p \frac{\partial T}{\partial t} = \nabla \cdot \lambda \nabla T + (-\Delta H_{\mathrm{r}})_{\mathrm{J}} R_{\mathrm{J}} \tag{2.13}$$

where λ denotes the thermal conductivity. Note that diffusion of the solid reactant is negligible. The geometry of the system is determined by the shape of the reactant mixture. A popular practice is to press the reactant mixture into a cylindrical preform. For this case, Equation (2.13) must be consistent with a cylindrical geometry, the dimensions of which are determined by the preform. The temperature at the reaction front can be very high and Equation (2.13) must be amended to take phase transformations into account. In the following section some aspects of the maximum temperature rise are discussed.

Maximum Temperature Rise. For $T_{\mathrm{ad}} < T_{\mathrm{fus}}$ the adiabatic temperature T_{ad} rise of a system where reactant A is transformed to product P, is given by:

$$T_{\mathrm{ad}} = T_0 + \frac{(-\Delta H_{\mathrm{r}})\, c_{\mathrm{A}}}{\varrho c_p} \tag{2.14}$$

When $T_{\mathrm{ad}} > T_{\mathrm{fus}}$ and

$(-\Delta H_{\mathrm{r}}) < \int_{T_0}^{T_{\mathrm{fus}}} c_p \mathrm{d}T + (-\Delta H_{\mathrm{fus}})$ the following equation holds:

$$T_{\mathrm{ad}} = T_0 + \frac{(-\Delta H_{\mathrm{r}})\, c_{\mathrm{A}}}{\varrho c_p} - \frac{\nu\,(-\Delta H_{\mathrm{fus}})}{c_p} \tag{2.15}$$

Otherwise:

$$T_{\mathrm{ad}} = T_{\mathrm{fus}} + \frac{(-\Delta H_{\mathrm{r}})\, c_{\mathrm{A}}}{\varrho c_{p\mathrm{l}}} - \frac{(-\Delta H_{\mathrm{fus}})}{c_{p\mathrm{l}}} - \frac{c_p}{c_{p\mathrm{l}}}\,[T_{\mathrm{fus}} - T_0] \tag{2.16}$$

where c_{p1} denotes the specific heat capacity of the liquid. Average properties over the temperature range must be used and the densities of the solid and liquid phases are assumed to be the same. The final temperature can be decreased by adding product to the reactant mixture to decrease c_{A}. There is a bound on the degree of dilution; if the system is too dilute, it will not ignite and self-propagation of the combustion front is not possible [5].

Equations (2.14) – (2.16) are derived on the basis of thermodynamic arguments, but the

batch process is transient and the maximum temperatures in the transient process can be different from those obtained from Equations (2.14) – (2.16). Unfortunately no rigorous bounds on the maximum temperature of a transient process are known and they must be determined by numerical methods. Furthermore, these transient maxima can exceed the adiabatic temperature and Equations (2.14) – (2.16) cannot be relied upon to give upper estimates of the reaction temperature.

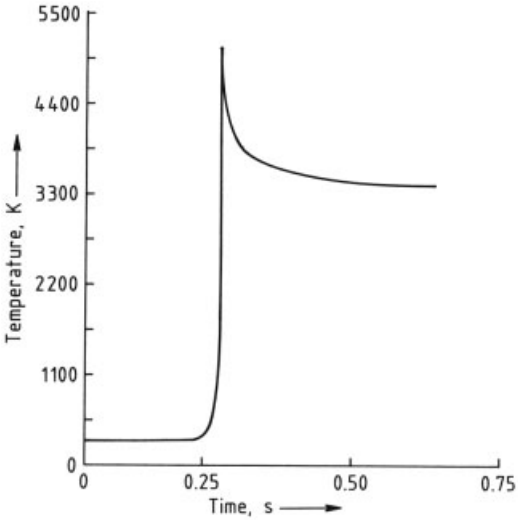

Figure 4. Temperature at the center of the cylindrical TiC sample

Example. Determine the maximum temperature of the following reaction:

$$\mathrm{Ti} + \mathrm{C} \rightarrow \mathrm{TiC}$$

A stoichiometric mixture of the reactants is pressed into a cylinder and simultaneously ignited at both ends. The cylinder is 2 cm long and perfectly insulated. The necessary properties of the system are listed below:

Density	$3240\,\mathrm{kg/m^3}$
Heat capacity	$1113.48\,\mathrm{J\,kg^{-1}\,K^{-1}}$
Thermal conductivity	$8.37\,\mathrm{W\,m^{-1}\,K^{-1}}$
Heat of formation at 25 °C	1.84×10^5 J/mol
Activation energy of formation reaction	8.29×10^5 J/mol
Frequency factor in expression for rate constant of formation reaction	$1.05\times10^5\,\mathrm{s^{-1}}$

Solution. Equations (2.12), (2.13) are integrated numerically and the maximum temperature at the center of the Ti – C cylinder is shown in Figure 4. The adiabatic temperature for this reaction is obtained from Equation (2.14):

$$T_{\mathrm{ad}} = 298\,\mathrm{K} + \frac{\left(1.8\times10^8\right)\ \mathrm{J/kmol}}{60\,\mathrm{kg/kmol}\times1113.48\,\mathrm{J\,kg^{-1}\,K^{-1}}} = 3052\,\mathrm{K}$$

The maximum transient temperature is 5100 K. The temperature overshoot is 2000 K. Since the melting point of TiC is 3420 K, the transient maximum temperature will exceed the melting point.

Reaction Wave Propagation. When the sample is ignited at one point by an external heat source, the reaction front can traverse the sample in a number of ways. The ideal propagation mode, from a processing point of view, is a stable front that propagates at a constant speed. If this front velocity is denoted by v_{F}, a similarity transformation can be made between time t and the direction of propagation x to render Equations (2.12), (2.13) time independent:

$$z = x + v_{\mathrm{F}}\,t \tag{2.17}$$

The integrity of this planar front is not preserved when external heat losses are present. Heat losses also reduce v_{F}. If heat losses are large enough, the system is quenched.

Other modes of front propagation also exist [6]. The planar front can, for example, become unstable and collapse into hot spots which are seated at different positions from the centerline. These hot spots propagate in a helical way around the centerline, leaving behind regions of incomplete conversion.

2.2.3. Solid – Gas Reactions

In solid – gas systems, combustion takes place between a solid and a gaseous component. A very important factor is the flow of gas through the porous solid by means of filtration or permeation. Hence this type of reaction is often termed *filtration combustion*. In Figure 5 various configurations for filtration combustion are shown.

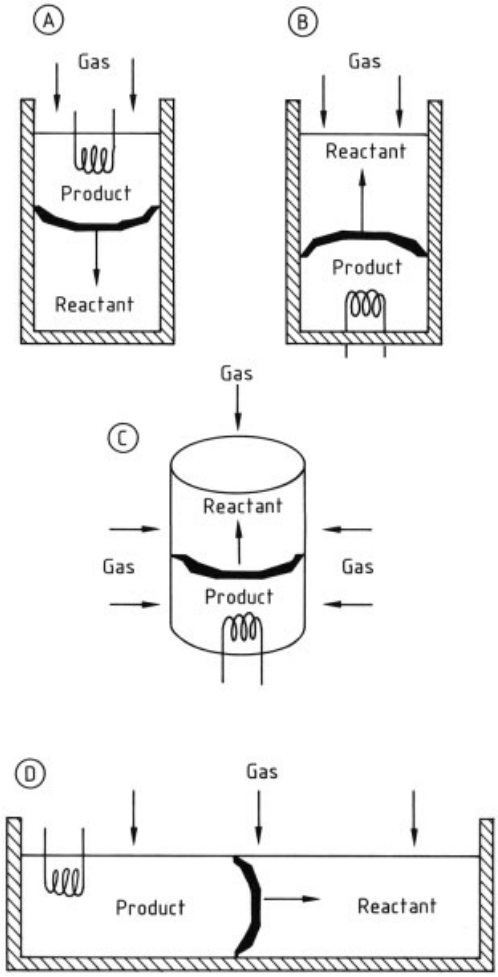

Figure 5. Different configurations of batch solid – gas reactions
A) Reaction initiated at top of container; B) Reaction initiated from bottom of container; C) Free-standing sample with gas penetrating from all sides; D) Design with both counter- and crossflow of reactant gases towards reaction front

The first two configurations are similar. In Figure 5 A, the reaction was initiated at the top of the container and the front propagates away from the open end. In Figure 5 B the reaction was initiated at the bottom of the container and the front propagates countercurrently to the flow of gas. An advantage of the second option is that the diffusion length for the gas decreases with increasing degree of conversion. Product materials also tend to sinter partially during the reaction, which makes them less permeable than the fresh material, which is another advantage of the second configuration. The sample shown in Figure 5 C is free standing and gas can penetrate from all sides into the sample. However, the reactant powder must be compressed to give the sample the necessary strength, therefore the powder's permeability will decrease. The configuration depicted in Figure 5 D is the most practical, because both counter- and crossflow of reactant gases towards the reaction front can be maintained without limiting the flow of the gaseous reactant through the solid medium. The filtration combustion model is described by the following set of equations:

$$\frac{\partial \varepsilon \varrho_{\mathrm{g}}}{\partial t} = -\nabla\left(\varepsilon \varrho_{\mathrm{g}} v\right) - \nu_{\mathrm{g}} R_{\mathrm{s}} \tag{2.18}$$

$$\frac{\partial\left[\varepsilon \varrho_{\mathrm{g}} c_{pg} + (1-\varepsilon)\, \varrho_{\mathrm{s}} c_{ps}\right]}{\partial t} = \nabla \cdot \nabla\left(\lambda T\right) - \nabla\left(\varepsilon \varrho_{\mathrm{g}} v T\right) + \left(-\Delta H\right)_{\mathrm{s}} R_{\mathrm{s}} \tag{2.19}$$

$$p = \frac{\varrho_{\mathrm{g}} R T}{M_{\mathrm{g}}} \tag{2.20}$$

$$v = -P \boldsymbol{\nabla} p \tag{2.21}$$

$$R_{\mathrm{g}} = -\varrho_{\mathrm{B}}^{0} p^{n} f\left(c_{\mathrm{s}}\right) k \tag{2.22}$$

where ε is the porosity of the bed, P the permeability in the bed, ϱ_{B}^{0} the initial bulk density of the solid, and $f\left(c_{\mathrm{s}}\right)$ the dependence of the reaction on the concentration of the solid phase. The porosity and permeability of the bed change as the reaction progresses, mainly because of sintering and melting; this affects the flow of gas to the front. The effects of sintering and melting can be incorporated into the model.

Example. A rectangular container is filled with titanium powder and placed in a chamber filled with nitrogen. The reaction is ignited at the top right of the container and the pressure in the chamber is kept constant. Low and high chamber pressures affect the conversion and in Figure 6 the results for both cases at different times (time increases from top to bottom) are shown.

The front propagates faster when the overhead pressure is higher and penetrates to the bottom of the container. When the overhead pressure is lower, conversion is incomplete and the

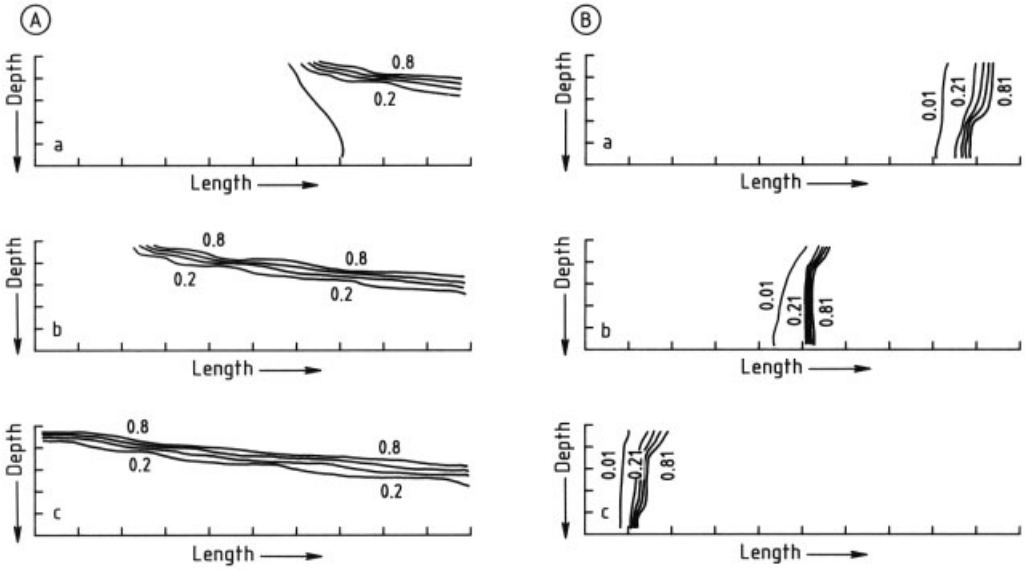

Figure 6. Conversion contours at different reaction times (time increases in the direction a – c) for A) low and B) high chamber pressures

reaction time is also much longer. When the depth of the container is increased, the chamber pressure must be further increased, but the exact values required can only be established through a numerical simulation.

3. Continuous Stirred-Tank Reactors (CSTR)

3.1. Isothermal Homogeneous System

The continuous, ideally mixed, stirred-tank reactor (CISTR) is a flow reactor in which all elements of the fluid have the same composition as the outlet stream of the reactor.

The continuous stirred-tank reactor is normally run at steady state, and is usually operated such that it is well-mixed. As a result, the CISTR is generally modeled as having no spatial variations in concentration, temperature, or reaction rate throughout the vessel. In systems where mixing is highly nonideal, the well-mixed model is inadequate and other techniques, such as residence-time distributions, have to be used to obtain meaningful results.

The CISTR has the following unique characteristics:

1) Temperature and heat transfer can be relatively easily controlled
2) Fine catalyst particles can be effectively suspended in a liquid reacting system
3) In a system with parallel reactions, a reaction with lower order is favored
4) In polymerization reactions, wider molecular mass distributions are expected
5) Very economical when long residence time is required
6) Recommended for liquid rather than gaseous reacting systems

Material Balance. The overall material balance for a CISTR which is schematically shown in Figure 7 reads:

$$\frac{d\,(m\xi_{\mathrm{J}})}{\mathrm{d}t} = -\Delta\Phi_{\mathrm{m}}\xi_{\mathrm{J}} + M_{\mathrm{J}}\,|R|_{\mathrm{J}}\,V_{\mathrm{r}} \tag{3.1}$$

where m is the total mass in the system; $-\Delta\Phi_{\mathrm{m}}$ the net inflow of mass by convection and diffusion; ξ_{J} the degree of conversion of component J; M_{J} — R —$_{\mathrm{J}}$ the mass rate of production of component J per unit volume; and V_{r} the volume of the reaction mixture.

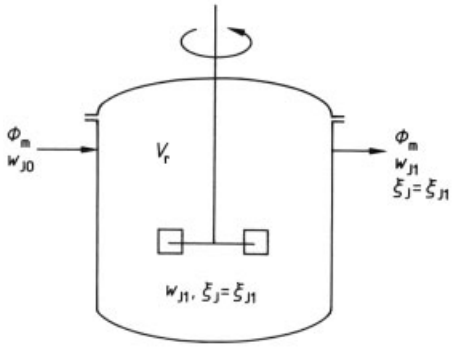

Figure 7. CISTR reactor
For explanation of symbols see text

When a CISTR is operated in a steady state, the required reactor volume is

$$V_r = \Phi_m \frac{\xi_J}{M_J\,|R|_J} \tag{3.2}$$

or in the case of constant density

$$V_r = \Phi_v \frac{c_{J_0}\zeta_J}{R_J} \tag{3.3}$$

where Φ_v is the volumetric flow rate; ζ_J the relative degree of conversion; c_{J_0} the molar concentration of species J; and R_J the molar rate of production of J per unit volume of the reaction phase. The mass balance equations for various reaction orders are listed in Table 2.

Table 2. Ratio of outlet to inlet concentrations in a CISTR for reactions of different order

Reaction order, n	Reaction rate equation, R_A	Concentration ratio, c_A/c_{A_0}
-1	$k c_A^{-1}$	$\frac{1}{2}\left[1 + \sqrt{\left(1 - \frac{4k\tau}{c_{A_0}^2}\right)}\right]$
0	k	$1 - \frac{k\tau}{c_{A_0}}$
$\frac{1}{2}$	$k\sqrt{c_A}$	$\sqrt{\left(1 + \frac{k^2\tau^2}{4c_{A_0}}\right)} - \frac{k\tau}{2\sqrt{c_{A_0}}}$
1	$k c_A$	$\frac{1}{1+k\tau}$
2	$k c_A^2$	$\frac{1}{2k c_{A_0}\tau}\left\lvert\sqrt{(1 + 4k c_{A_0})} - 1\right\rvert$

Most important in designing a homogeneous CISTR system is to select equipment and stirring procedures that will assure essentially perfect mixing. Choice of a suitable type of impeller and reactor arrangement, the required power, the mode of adding and removing reactants and products, and the cooling or heating rates are all essential concerns. For more details, see → Stirring, [7], [8].

Mixing. For a processing vessel to be effective, the volume of fluid or slurry that is circulated by the impeller must be sufficient to sweep out the entire vessel in a reasonable time. Also, the velocity of the stream leaving the impeller must be high enough to carry the currents to the remotest parts of the tank. In addition, an adequate level of turbulence is required in order to maintain the high effectiveness of the process. The proper selection of an impeller, its speed, and the geometrical dimensions of the reactor are very important design parameters (→ Stirring; → Continuous Mixing of Fluids).

In practice, for moderately fast reactions, a tank reactor may be approximately considered as ideally mixed if the discharge rate of the impeller Φ_c (the circulation rate) is at least five to ten times the feed rate Φ_m i.e.,

$$\frac{\Phi_c}{\Phi_m} > 5 - 10 \tag{3.4}$$

The impeller discharge rate depends on the type of impeller, the tank geometry, and operating conditions. This discharge rate can be estimated from:

$$\frac{\Phi_c}{N d_I^3} = N_{QI} \tag{3.5}$$

where d_I is the impeller diameter, N the number of impeller revolutions per unit time, and N_{QI} is a dimensionless constant. The value of this constant depends on the axial blade width of the impeller, the number of blades, curvature or pitch, and the ratio of impeller velocity to fluid rotational velocity [8]. The following values are recommended for agitators in baffled tanks which are commonly used in industry:

Marine-Type Impeller:

$$N_{QI} = 0.5$$

Turbine (with six blades and width-to-diameter ratio of 1 : 5):

$$N_{QI} = 0.93 D/d_I \quad \text{for} \quad Re_{imp} > 10^4$$

where $Re_{imp} = \varrho N d_I^2/\mu$ is the impeller Reynolds number, and D the diameter of the reactor.

Many reactions require rapid dispersal of reactants in order to avoid even momentary build-up of high concentrations in the system. In homogeneous reacting systems, the most rapid dispersal can be accomplished by discharging the

feed as close as possible to the center of the impeller. Discharge ports for product removal should be located at the opposite end of the flow pattern from the inlet nozzle. In general, the flow of a reacting fluid through a reactor is a very complex process. The interaction of flow pattern and chemical reaction cannot be analyzed rigorously even in the simple case of a single isothermal reaction. Therefore simplified and/or idealized models are necessary.

Residence-Time Distribution (RTD). (→ Principles of Chemical Reaction Engineering, Chap. 4.2.1.). The frequency function often used in statistics (called the E diagram in RTD description) can be derived for a continuous stirred-tank reactor from a mass balance of the reactor with an ideal pulse perturbation at the inlet stream. The pulse is immediately distributed over the reactor, moreover the product concentration remains the same everywhere and equals the outlet concentration. The $E(t)$ diagram reads:

$$E(t) = \frac{1}{\tau} \mathrm{e}^{-t/\tau} \tag{3.6}$$

For a continuously operated mixing vessel which shows a RTD of an ideal mixer (Eq. 3.6), the different states of micromixing are (→ Principles of Chemical Reaction Engineering, Chap. 4.3.):

1) CISTR: the ideal mixer (mixing on molecular level)
2) CSSTR: the completely segregated mixer where the fluid consists of lumps which move through the reactor without exchange of matter with their surroundings
3) CSTR: the partially segregated mixer

A detailed description of the amount of micromixing with one parameter is generally not possible. For non-first-order reactions, apart from the reaction rate not only the intensity of mass exchange between fluid elements is important, but also the residence time of a molecule in the reactor.

In the case of maximum micromixing and constant density of the reacting system, the conversion is identical to that calculated for the CISTR. When segregated flow takes place, the situation is different. Each aggregate in the outlet stream can be considered as a batch reactor with a reaction time equal to its residence time. The fraction of aggregates having a residence time between t and $t + \Delta t$ is $E\,\Delta t$. If the remaining reactant concentration after conversion in a batch reactor with residence time t is $c_{\mathrm{A}}^{\mathrm{Batch}}(t)$, the average concentration at the outlet of the segregated continuous flow reactor CSSTR can be calculated:

$$\frac{\bar{c}_{\mathrm{A}}}{c_{\mathrm{A}_0}} = \frac{1}{c_{\mathrm{A}_0}} \int_0^\infty c_{\mathrm{A}}^{\mathrm{Batch}}(t)\, E \mathrm{d}t \tag{3.7}$$

or

$$\frac{\bar{c}_{\mathrm{A}}}{c_{\mathrm{A}_0}} = \frac{1}{c_{\mathrm{A}_0}} \int_0^\infty c_{\mathrm{A}}^{\mathrm{Batch}}(t)\, \frac{1}{\tau} \mathrm{e}^{-t/\tau} \mathrm{d}t \tag{3.8}$$

The average outlet concentrations for different reaction kinetic expressions for both ideal micromixing and segregated flow are listed in Table 3 [9]. Outlet concentrations are identical for both cases if a first-order reaction takes place.

Example [9]. Reactant A ($M_{\mathrm{A}} = 104$ kg/kmol) is to be polymerized in the dispersed phase in a continuous stirred-tank reactor. Based on batch-scale experiments, the following reaction rate expression was found:

$$R_{\mathrm{A}} = -k\, c_{\mathrm{A}}^{1.5}$$
$$\text{with } k = 2.5 \times 10^{-4}\, \mathrm{s}^{-1}\, \mathrm{kmol}^{-0.5}\, \mathrm{m}^{1.5}$$

The required size of the tank reactor shall be calculated for a production of 10 t/d of polymer with a relative degree of conversion $\bar{\zeta}_{\mathrm{A}} = 0.9$, under the assumption that coalescence between the liquid drops does not occur. The feed of the dispersed phase consists of pure A, and the density of this phase remains constant at a value of 832 kg/m^3 during the reaction. The volume fraction of the dispersed phase in the reactor is 0.165.

Solution. The average conversion $\bar{\zeta}_{\mathrm{A}}$ at the outlet of the CSSTR is obtained by rearrangement of Equation (3.8):

$$\bar{\zeta}_{\mathrm{A}} = \int_0^\infty \zeta_{\mathrm{A}}^{\mathrm{Batch}}(t)\, \frac{1}{\tau} \mathrm{e}^{-t/\tau} \mathrm{d}t \tag{3.8a}$$

Table 3. Design equations for ideally mixed CISTR and for completely segregated fluids CSSTR

Kinetics	CISTR
$R_A = -k c_A$	$\frac{c_A}{c_{A_0}} = \frac{1}{1+N_r}$
$N_r = k\tau$	
$R_A = -k$	$\frac{c_A}{c_{A_0}} = 1 - N_r,\ N_r \le 1;\ \frac{c_A}{c_{A_0}} = 0,\ N_r \ge 1$
$N_r = \frac{k\tau}{c_{A_0}}$	
$R_A = -k c_A^{0.5}$	$\frac{c_A}{c_{A_0}} = 1 - \frac{N_r^2}{2}\left[-1 + \sqrt{\left(1 + \frac{4}{N_r^2}\right)}\right]$
$N_r = \frac{k\tau}{\sqrt{c_{A_0}}}$	
$R_A = -k c_A^2$	$\frac{c_A}{c_{A_0}} = \frac{-1 + \sqrt{(1+4N_r)}}{2N_r}$
$N_r = k\tau c_{A_0}$	
$R_A = -k c_A^n$	$\left(\frac{c_A}{c_{A_0}}\right)^n N_r + \frac{c_A}{c_{A_0}} - 1 = 0$
$N_r = c_{A_0}^{n-1} k\tau$	

Kinetics	CSSTR
$R_A = -k c_A$	$\frac{\bar{c}_A}{c_{A_0}} = \frac{1}{1+N_r}$
$N_r = k\tau$	
$R_A = -k$	$\frac{\bar{c}_A}{c_{A_0}} = 1 - N_r + N_r e^{-\frac{1}{N_r}}$
$N_r = \frac{k\tau}{c_{A_0}}$	
$R_A = -k c_A^{0.5}$	$\frac{\bar{c}_A}{c_{A_0}} = 1 - N_r\left[1 - \frac{N_r}{2}\left(1 - e^{-\frac{2}{N_r}}\right)\right]$
$N_r = \frac{k\tau}{\sqrt{c_{A_0}}}$	
$R_A = -k c_A^2$	$\frac{\bar{c}_A}{c_{A_0}} = -\frac{1}{N_r} e^{\frac{1}{N_r}} Ei\left(-\frac{1}{N_r}\right)^*$
$N_r = k\tau c_{A_0}$	
$R_A = -k c_A^n$	$\frac{\bar{c}_A}{c_{A_0}} = \frac{1}{\tau}\int_0^\infty [1 + (n-1) c_{A_0}^{n-1} k t]^{\frac{1}{1-n}} e^{-t/\tau} dt$
$N_r = c_{A_0}^{n-1} k\tau$	

$^*\ Ei(-x) = \int_x^\infty \frac{e^{-y}}{y} dy$

The material balance for a batch system together with the rate expression yields:

$$\bar{\zeta}_A^{\text{Batch}} = 1 - \left(\frac{2}{2 + k t \sqrt{c_{A_0}}}\right)^2$$

By integrating Equation (3.8 a) a value of $k\tau \sqrt{c_{A_0}} = 15.4$ is found for $\bar{\zeta}_A = 0.9$. The calculated residence time τ is 6 h and the required volume of the dispersed liquid $\Phi_m \tau / 0.9\, \varrho_A$ is equal to $3.34\,m^3$. The total required volume in the reactor is $3.34\,m^3/0.165 = 20.2\,m^3$.

Start-Up of a CISTR. For non-steady-state operations, the accumulation term in Equation (3.1) cannot be neglected. Two different periods of non-steady-state operations in a CISTR can be distinguished:

1) Filling a reactor with reactants (variable reactor volume)
2) Non-steady-state operation at a constant reactor volume

In the case of an irreversible liquid reaction ($A \rightarrow B$) which takes place in a CISTR under isothermal and constant density conditions, the instant concentration in the reactor can be calculated as follows: For $t \le \tau$, no liquid leaves the reactor and the material balance leads to:

$$\frac{-d(c_A V)}{dt} = -\Phi_v c_{A_0} + k c_A V \tag{3.9}$$

in which the reaction volume is $V = \Phi_v t$. Integration of Equation (3.9) gives:

$$c_A = \frac{c_{A_0}}{kt}\left(1 - e^{-kt}\right) \quad \text{for } t \le \frac{V}{\Phi_v} \tag{3.10}$$

For $t > \tau$ the reaction volume V is constant and equals V_r and the product flow out of the reactor equals the feed rate Φ_v. The material balance for reactant A now becomes:

$$\begin{aligned}\frac{-d(c_A V_r)}{dt} &= -V_r \frac{dc_A}{dt} \\ &= -\Phi_v\left(c_{A_0} - c_A\right) + k c_A V_r\end{aligned} \tag{3.11}$$

Integration yields:

$$\frac{c_{A_0} - (1+k\tau)\, c_A}{c_{A_0} - (1+k\tau)\, c_{A_\tau}} = e^{-(1+k\tau)(t-\tau)/\tau} \quad \text{for } t > \tau \tag{3.12}$$

where $c_A = \frac{c_{A_0}}{k\tau}\left(1 - e^{-k\tau}\right)$

After a long time ($t \rightarrow \infty$) the right-hand side of Equation (3.12) approaches zero, so that ultimately the concentration c_A in the reactor reaches the value $c_{A\infty}$. A conservative estimate for the time required to reach steady state can be made by substituting $c_{A_\tau} = c_{A_0}$ in Equation (3.12) for the required degree of conversion.

3.2. Isothermal Heterogeneous Reactors

Contrary to reactions in a single phase where mixing takes place on a molecular scale, in a multiphase system not all reactive molecules present may be available for chemical reaction due to mass transfer limitations. For a reacting system consisting of two phases L and G, with a reaction taking place in phase L only, the mass balance equations are:

$$\frac{dm_{\mathrm{JL}}}{\mathrm{d}t} = -\Delta\left(\Phi_{\mathrm{mL}} w_{\mathrm{JL}}\right) + J_{\mathrm{JL}} M_{\mathrm{J}} a V_{\mathrm{r}} + R_{\mathrm{J}} M_{\mathrm{J}} (1-\varepsilon) V_{\mathrm{r}} \tag{3.13}$$

and:

$$\frac{dm_{\mathrm{JG}}}{\mathrm{d}t} = -\Delta\left(\Phi_{\mathrm{mG}} w_{\mathrm{JG}}\right) + J_{\mathrm{JG}} M_{\mathrm{J}} a V_{\mathrm{r}} \tag{3.14}$$

The term $R_{\mathrm{J}} M_{\mathrm{J}} (1-\varepsilon) V_{\mathrm{r}}$ is the production rate of J by a homogeneous chemical reaction in phase L where $(1-\varepsilon)$ is the volume fraction of the reaction phase per unit volume of a reactor. The terms $(-\Delta\Phi_{\mathrm{m}} w_{\mathrm{J}})$ and $(J_{\mathrm{J}} M_{\mathrm{J}} a)$ represent the net supply of component J by convection and the mass flow of J from the interface to the bulk of the phase, respectively.

When a reaction takes place in a system of two immiscible or partially miscible liquid phases, the overall process can be controlled by chemical reaction and/or mass transfer between both phases. In the case of a *chemically controlled process*, the overall rate expression in Equation (3.1) is the sum of the reaction rates in the individual phases.

When *mass transfer* is the rate-determining step, phase equilibrium cannot be achieved and the reaction rate is fast compared to the rate of mass transfer. Thus, the reaction proceeds as the reactants diffuse through the interface into the bulk of the other phase. If reactant A is a major component of phase I while phase II consists of reactant B and if equilibrium is attained at the interface, the rate of mass transfer of A from phase I to phase II would be equal to the rate of disappearance of A by chemical reaction in phase II. For an irreversible, second-order reaction between A and B, mass transfer of A from phase I to phase II can be described as:

$$J_{\mathrm{AII}} = k_{\mathrm{II}} a \left(\frac{c_{\mathrm{AI}}}{K_{\mathrm{A}}} - c^{0}_{\mathrm{AII}}\right)(V_{\mathrm{I}} + V_{\mathrm{II}}) \tag{3.15}$$

where k_{II} is the mass-transfer coefficient in phase II per unit volume of both phases (m/s), a is the interfacial area per unit volume of both phases ($\mathrm{m}^2/\mathrm{m}^3$), c_{AI} is the concentration of reactant A in phase I, K_{A} is the distribution coefficient of A between both phases, and c^{0}_{AII} is the steady-state concentration of A in phase II.

The value of the product of mass-transfer coefficient and interfacial area strongly depends on the reactor arrangement and empirical correlations must be used to scale-up such reacting systems [10], [11].

3.3. Nonisothermal Continuous Stirred-Tank Reactors

All chemical reactions are in principle accompanied by evolution or absorption of heat and these effects often have to be taken into account. For flow processes, the energy balance equation is

$$\frac{d\left(\langle u\rangle m\right)}{\mathrm{d}t} = -\Delta\left(h\Phi_{\mathrm{m}}\right) + \dot{Q} - \dot{W} \tag{3.16}$$

where h is the enthalpy per unit mass of the reaction mixture and $\langle u\rangle$ is the average internal energy over the reactor volume. For the steady-state operation of a continuous tank reactor provided with a heat exchange area A, the energy balance becomes:

$$\Phi_{\mathrm{m}}\left(h_1 - h_0\right) = UA\left(T_{\mathrm{c}} - T_1\right) \tag{3.17}$$

where the subscripts 1 and 0 refer to the conditions at the outlet and inlet, respectively; T_{c} is the temperature of the cooling or heating medium. The difference in enthalpy $(h_1 - h_0)$ can be calculated from:

$$h_1 - h_0 = \int_{T_{0,\mathrm{O}}}^{T_1 \xi_{\mathrm{J},1}} \left[c_p\,\mathrm{d}T + (\Delta h_{\mathrm{r}})_{\mathrm{J}}\,\mathrm{d}\xi_{\mathrm{J}}\right] \tag{3.18}$$

where c_p is the specific heat capacity at constant pressure and (Δh_{r}) the enthalpy of reaction at constant pressure per unit mass of J converted. If the density and heat capacity of the system are constant, the energy balance for reactant A gives:

$$-(\Delta H_{\mathrm{r}})_{\mathrm{A}}\left(c_{\mathrm{A}_0} - c_{\mathrm{A}_1}\right) = c_p \varrho\left(T_1 - T_0\right) + \frac{UA}{\Phi_v}\left(T_1 - T_{\mathrm{c}}\right) \tag{3.19}$$

where $(\Delta H_r)_A$ is the heat of reaction per mole of A converted. In the case of a first-order irreversible reaction Equations (3.1) and (3.19) can be rewritten to:

$$-(\Delta H_r)_A\, c_{A_0} \frac{k\tau}{1+k\tau} = \varrho c_p (T_1 - T_0) + \frac{UA}{\Phi_v}(T_1 - T_c) \tag{3.20}$$

In order to find the maximum temperature T_1, Equation (3.20) must be solved by trial and error, either numerically or graphically, because k is an exponential function of temperature. The left-hand side of Equation (3.20) represents the heat produced per unit volume of a reaction mixture. The right-hand side represents the heat removed per unit volume of the reaction mixture as a result of the heat absorbed by the cold feed and the heat transferred to the cooling medium. The maximum temperature rise can be reached when the system is adiabatic (no heat transfer to the surroundings) and complete conversion is achieved:

$$\Delta T_{ad} = \frac{-(\Delta H_r)_A\, c_{A_0}}{\varrho c_p} \tag{3.21}$$

Autothermal Reactor Operation. The heat released during an exothermic reaction is often used for preheating inlet reactant streams. A reactor system, in which such a feedback of reaction heat to the incoming reactant stream takes place is said to operate under autothermal conditions. The theory of steady-state behavior of autothermal reactors is described in [12], [13].

Multiplicity (→ Principles of Chemical Reaction Engineering, Chap. 7.2.2.). In Figure 8 both right- (heat removal) and left-hand sides (heat production) are plotted for different inlet and operating conditions. At the intersection of the heat production (solid line) and heat removal curves (dashed lines), Equation (3.20) is obeyed. For a certain range of inlet parameters several solutions may exist (multiplicity of steady states). If the heat removal is represented by line 1 in Figure 8, the reactor is cooled so effectively that steady-state operation is possible only at a low temperature and a very low degree of conversion. If the cooling rate is lowered (line 2), three points of intersection occur with the heat generation curve. The two intersection points, the lowest and highest reaction temperature, represent stable conditions due to the fact that the rate of heat removal is higher than that generated in the system. The intermediate intersection point is unstable. Since the slope of the heat production curve at the intermediate point is greater than that of the heat removal line, any positive temperature perturbation will be amplified until the reactor reaches the upper stable operating point, and negative temperature deviation leads to extinction (lower stable operating point). If the feed temperature T_0 and the cooling capacity are such that the heat removal line 3 prevails (see Fig. 8), the reactor operates in an upper stable state and no special measures need be taken to start up a process.

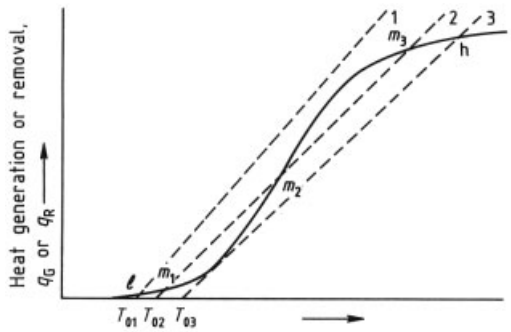

Figure 8. Heat production (—) and heat removal (– – –) in a nonisothermal CISTR for an irreversible exothermic reaction A → R

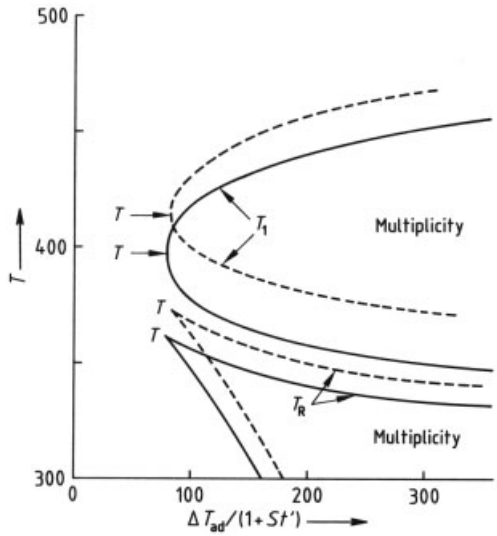

Figure 9. Regions of multiplicity in a CISTR [5]
T = Trifurcation point; —— $\tau = 1$ s; – – – $\tau = 0.5$ s

In Figure 9 a multiplicity region is presented as a function of the maximum temperature increase in the reactor, $\left(\frac{\Delta T_{\mathrm{ad}}}{1+St'}\right)$. In the area enveloped by the curves multiplicity will occur. Outside, only one (stable) operating point is possible. Operating points located above the multiplicity region are distinguished by a high degree of conversion, those below the multiplicity region by a low degree of conversion. The point with the lowest value of $\frac{\Delta T_{\mathrm{ad}}}{1+St'}$ where multiplicity occurs is often called the *trifurcation point*, and the curves departing from it, the *bifurcation lines*.

HLAVACEK et al. derived the criteria for existence of multiplicity of steady states for first-order reactions [14]. Multiplicity occurs when the dimensionless adiabatic temperature rise $B = \frac{(-\Delta H_{\mathrm{r}})\, c_{\mathrm{A}_0}}{\varrho c_p T_0} \frac{E}{RT_0}$ exceeds a critical value of$B*$:

1) Adiabatic case ($\bar{\beta} = 0$)

$$B^* = \frac{4\bar{\gamma}}{\bar{\gamma}-4}$$

2) Adiabatic case ($\bar{\beta} = 0$); $\bar{\gamma} \to \infty$

$$B^* = 4$$

3) Nonadiabatic case; $\bar{\gamma} \to \infty$

$$B^* = 4\bar{\delta}$$

4) Nonadiabatic case

$$B^* = \frac{4\left(\bar{\varepsilon}+\bar{\delta}\bar{\gamma}\right)^2}{\bar{\gamma}\bar{\delta}\left(\bar{\gamma}-4\right)-4\bar{\varepsilon}}$$

where $\bar{\gamma} = E/RT_0$ is a dimensionless activation energy, $\bar{\beta} = UA/\Phi_v\, \varrho\, c_p$ is a dimensionless heattransfer coefficient, $\bar{\delta} = 1 + \bar{\beta}$, $\bar{\varepsilon} = \bar{\beta}\,\bar{\Theta}_{\mathrm{c}}$, and $\Theta_{\mathrm{c}} = \frac{E}{RT_0^2}\left(T_{\mathrm{c}} - T_0\right)$ is a dimensionless cooling temperature.

Hysteresis in Autothermal Systems (→ Principles of Chemical Reaction Engineering).

If the feed temperature T_0 is increased at a constant feed rate, the reaction will be ignited at a certain feed temperature $(T_0)_{\mathrm{ign}}$ and the reactor will operate at the upper stable steady state. If T_0 is then lowered again, the reactor continues operating at a high conversion level until extinction takes place at a feed temperature $(T_0)_{\mathrm{ext}}$.

Stability of Autothermal CISTR. The transient behavior of strongly exothermic reacting systems may lead to unexpected, undesired changes in a reactor. This is a result of coupling between the material and energy balances (Eqs. 3.1 and 3.16). Detailed information on dynamic behavior of autothermal systems can be found in [15–17].

The mass and energy balance equations Equations (3.1) and (3.16) can be rewritten to:

$$\frac{d\zeta}{\mathrm{d}\Theta} = -\zeta + \frac{\tau R_{\mathrm{A}}}{c_{\mathrm{A}_0}} \tag{3.22}$$

and:

$$\frac{dT}{\mathrm{d}\Theta} = T_0 - T - \Delta T_{\mathrm{ad}} \frac{R_{\mathrm{A}}\tau}{c_{\mathrm{A}_0}} + St'\left(T_{\mathrm{c}} - T\right) \tag{3.23}$$

where $\tau = \varrho\, V_{\mathrm{r}}/\Phi_{\mathrm{m}}$ is the average residence time, $\Theta = t/\tau$ is a dimensionless time, and St' a modified Stanton number ($St' = UA/c_p\, \Phi_v$).

In the case of a first-order reaction Equations (3.22) and (3.23) can be rewritten to:

$$\frac{d\zeta}{\mathrm{d}\Theta} = -\zeta + k_{\mathrm{s}}\tau\left(1-\zeta\right)\mathrm{e}^{\Delta\vartheta} \tag{3.24}$$

$$\frac{d\vartheta}{\mathrm{d}\Theta} = -\left(1+St'\right)\Delta\vartheta + \Delta\vartheta_{\mathrm{ad}} k_{\mathrm{s}}\tau\left(1-\zeta\right)\mathrm{e}^{\Delta\vartheta} + \Delta\vartheta_0 + St'\,\Delta\vartheta_{\mathrm{c}} \tag{3.25}$$

where:

$$\Delta\vartheta_0 = \left(T_0 - T_{\mathrm{s}}\right)\frac{E}{RT_{\mathrm{s}}^2}$$

$$\Delta\vartheta_{\mathrm{c}} = \left(T_{\mathrm{c}} - T_{\mathrm{s}}\right)\frac{E}{RT_{\mathrm{s}}^2}$$

$$\Delta\vartheta_{\mathrm{ad}} = \Delta T_{\mathrm{ad}}\frac{E}{RT_{\mathrm{s}}^2}$$

and k_{s} is the reaction rate constant at temperature T_{s}.

When a small perturbation is imposed on a steady state, Equations (3.24) and (3.25) can be linearized as follows:

$$\frac{d\Delta\zeta}{\mathrm{d}\Theta} = -\frac{1}{1-\zeta_{\mathrm{s}}}\Delta\zeta + \zeta_{\mathrm{s}}\Delta\vartheta \equiv a_{11}\Delta\zeta + a_{12}\Delta\vartheta \tag{3.26}$$

and:

$$\frac{d\Delta\vartheta}{\mathrm{d}\Theta} = \frac{\Delta\vartheta_{\mathrm{ad}}\zeta_{\mathrm{s}}}{1-\zeta_{\mathrm{s}}}\Delta\zeta + \left[\Delta\vartheta_{\mathrm{ad}}\zeta_{\mathrm{s}} - \left(1+St'\right)\right]\Delta\vartheta \equiv a_{21}\Delta\zeta + a_{22}\Delta\vartheta \tag{3.27}$$

where $\Delta\zeta = \zeta - \zeta_s$ and $\Delta\vartheta = \vartheta - \vartheta_s$; ζ_s and ϑ_s are the conversion and the dimensionless temperature at steady state.

The solution of Equation (3.26) and (3.27) has the form:

$$\Delta\zeta = b_{11}e^{p_1\Theta} + b_{12}e^{p_2\Theta} \tag{3.28}$$

$$\Delta\vartheta = b_{21}e^{p_1\Theta} + b_{22}e^{p_2\Theta} \tag{3.29}$$

where:

$$p_{1,2} = \frac{(a_{11}+a_{22})}{2}\left[1 \pm \sqrt{\left(1 - \frac{4\,(a_{11}a_{22} - a_{12}a_{21})}{(a_{11}+a_{22})^2}\right)}\right]$$

The chemical process is stable when:

$$\frac{1 + St'}{\Delta\vartheta_{ad}} > \zeta_s\,(1 - \zeta_s) \tag{3.30}$$

and:

$$\frac{1}{1 - \zeta_s} - \Delta\vartheta_{ad}\,\zeta_s + 1 + St' > 0 \tag{3.31}$$

where $\zeta_s = k\,\tau/(1 + k\,\tau)$ in the case of first-order reaction.

If the second condition (Eq. 3.31) is met, the reactor is dynamically stable and operates at its steady state (ζ_s, T_s). However, if this condition is not satisfied the reactor is dynamically unstable and operates in a fixed *limit cycle* around the steady-state point (ζ_s, T_s) (Fig. 10).

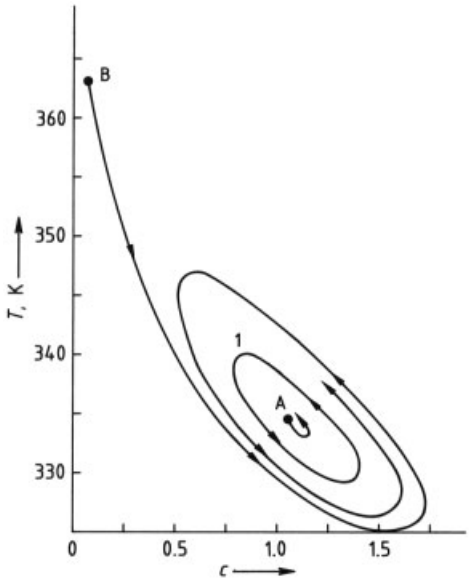

Figure 10. The limit cycle in CSTR [18]

3.4. Cascade of Tank Reactors [21]

Continuous stirred-tank reactors in series (Fig. 11) are used when the required residence time is long and more than one vessel is necessary for efficient, economical mixing. Conversions, yields, and residence times of the individual stirred-tank reactors are calculated analogously to those for a single CISTR. The steady-state *material balance* over the *n*th reactor for a component J reads:

$$\xi_{Jn} - \xi_{J\,(n-1)} = \frac{V_{rn}}{\Phi_m} M_J\, R_{Jn} \tag{3.32}$$

In most cases, the set of equations describing the degree of conversion in a cascade cannot be solved analytically. For a system consisting of *n* equal-sized CISTRs in series an analytical solution can be obtained for first-order irreversible reactions [19]:

$$c_{An} = c_{A_0}\,(1 + k\tau)^{-n} \tag{3.33}$$

and for second-order reaction: $c_{An} = 14\,k\,\tau_i$

$$c_{An} = \frac{1}{4k\tau_i}\left(-2 + 2\cdot\sqrt{-1\ldots + \sqrt{-1 + 2\sqrt{1 + 4k\,c_{A_0}\tau_i}}}\cdot s^n\right) \tag{3.34}$$

In other cases a solution has to be found by an algebraic stepwise method or by a graphical technique [9], [20].

Example. The second-order, reversible reaction $2\,A \rightleftharpoons C + D$ is carried out in a single CISTR with total volume of 2.665 m^3. The volumetric flow rate of reactant A is 7.866×10^{-4} m^3/s, and its inlet and outlet concentrations are 24.03 and 6.94 $kmol/m^3$, respectively. The forward rate constant at the reaction temperature is 1.156×10^{-4} $m^3\,kmol^{-1}\,s^{-1}$ and the equilibrium constant *K* is 16.0.

It now becomes necessary to replace the working reactor with a series of identical CISTRs. The volume of each reactor in a new reactor arrangement should be one-tenth of the current one. How many CISTRs are required?

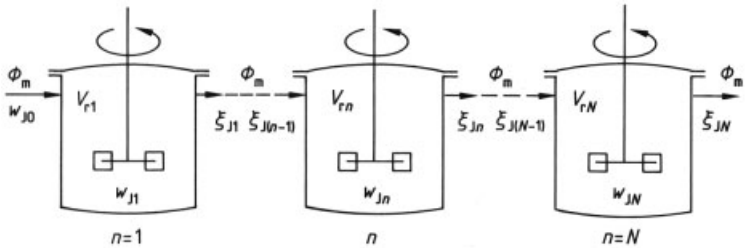

Figure 11. Continuous stirred-tank reactors in series (cascade)

Solution. The reaction rate can be expressed as follows:

$$R_{\mathrm{A}} = k_1 \left[c_{\mathrm{A}}^2 - \frac{\left(c_{\mathrm{A}_0} - c_{\mathrm{A}}\right)^2}{4K} \right]$$

The relation between R_{A} and c_{A} described by the above equation is shown in Figure 12. In addition, the balance equation

$$R_{\mathrm{A}} = \frac{\Phi_v}{V_{\mathrm{r}}} \left[c_{\mathrm{A}n} - c_{\mathrm{A}(n-1)} \right]$$

is also valid for each reactor in a series. The slope Φ_v / V_{r} in the case of a new arrangement is $2.95 \times 10^{-3}\,\mathrm{s}^{-1}$. A straight line through $c_{\mathrm{A}_0} = 24.03\,\mathrm{kmol/m^3}$ with this slope can be drawn to the intersection with the reaction rate curve. This process has to be repeated until $c_{\mathrm{A}n} = 6.94\,\mathrm{kmol/mol}$ is reached. From Figure 12, the number of stages required is 3.8. Hence four CSTRs should be used in this case.

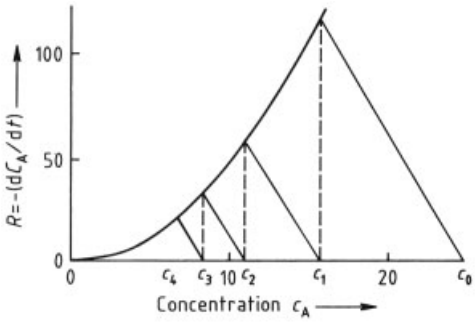

Figure 12. Graphical estimation of required number of CSTRs in series

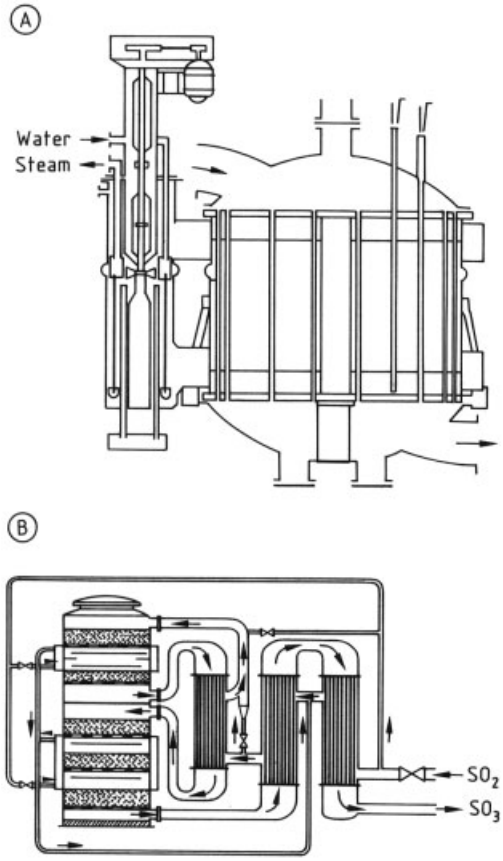

Figure 13. Different configurations for packed-bed reactors A) Multitube heat exchanger; B) Multistage adiabatic reactor
According to "Chemical Reactor Analysis and Design" G. F. Froment, K. B. Bischoff, John Wiley & Sons, New York 1990.

4. Packed-Bed Reactors

The packed-bed reactor constitutes the keystone in the production of several important chemicals. Several reactor configurations are encountered in practice (e.g., Fig. 13). The *heat-exchanger type* of reactor (Fig. 13 A) is used to carry out highly exothermic reactions. The *adiabatic multistage reactor* (Fig. 13 B) with cooling between stages is used to carry out exothermic equilibrium reactions. The *autothermal configuration*, an interesting alternative, is a heat-exchanger type of reactor in which the reactants circulate alternatively through shell and tube arrangements with the double purpose of being preheated and keeping the reactor operating within a prespecified thermal regime by cooling it down.

The analysis of transport phenomena for systems with chemical reaction is usually based on the transport equations resulting from the differential balance laws. To predict global effects, detailed information about the velocity profiles, and temperature and concentration fields is required. This information is extracted from the solution of the associated transport equations, subject to pertinent boundary conditions. When flow through a complex structure such as a porous medium is involved, these governing equations are valid even inside the pores. However, the geometric complexity of a randomly ordered particulate system prevents any general solution of detailed concentration, temperature, and flow fields. Instead, some form of macroscopic balances based on the average over a small volumetric element must be employed. A common practice is, with the help of some empirical relations, to replace the microscopic momentum, mass, and energy balances with corresponding macroscopic equations.

Due to the distributed nature of the system, a dynamic model contains two or more independent variables (time and spatial coordinates). Simultaneously, it is important in principle to include all relevant transport resistance (see Fig. 14) to avoid unrealistic results. Since a very complex model may be too difficult to solve, a number of simplifying assumptions are often adopted in practice. However, care must be taken when choosing the adequate degree of complexity for the reactor model, so as to guarantee sufficient accuracy for a given practical application.

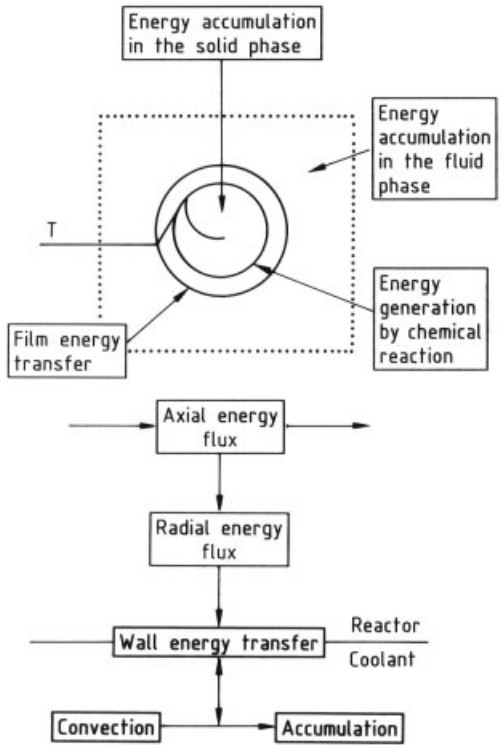

Figure 14. Energy transport mechanisms in packed beds

4.1. Mass and Energy Balances

Mathematical models for adiabatic and nonadiabatic packed-bed reactors have been systematically listed in reviews by Froment [22], Hlavacek [23], and Hofmann [24]. In these reviews an attempt is made to classify the models according to their complexity. Basically, the mathematical modeling of fixed-bed reactors can be performed either through the finite-stage or the continuum approach [25]. The continuum approach leads to a system of partial differential equations in which mixing effects are accounted for in the form of mass diffusion and heat conduction terms. This approach is being increasingly favored, mainly due to the continuous progress in computer software and numerical modeling.

In general, for the mathematical treatment of the heat and mass transfer processes in packed beds, with or without chemical reaction, three types of continuous models have been developed:

1) The single-particle model, where intra particle phenomena are accounted for separately
2) The two-phase (heterogeneous) model in which both phases, solid and fluid, exchange heat and/or mass, and such transport processes are accounted for
3) The one-phase (pseudohomogeneous) model in which the reactor is approximated as a quasi- continuum model

4.1.1. The Single-Particle Model

Basically the single-particle model is also a heterogeneous model. However, it differs from the two-phase model in that the solid phase is not regarded as a continuum since the concentration and temperature distribution inside the single pellet are accounted for.

Consider a catalyst pellet positioned in a flowing stream (Fig. 15). Three different regions can be distinguished: the bulk fluid phase, the boundary layer surrounding the particle, and the particle itself.

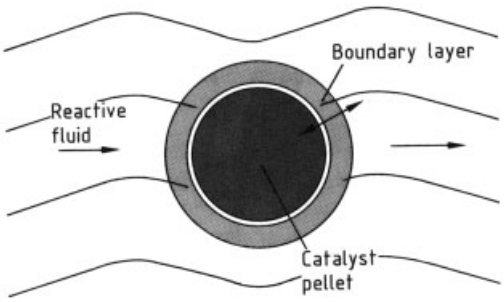

Figure 15. Schematic of a single particle in a flowing stream

Bulk Fluid Phase. The bulk fluid phase can be modeled as a constant temperature and constant concentration field. Perturbations at the inlet of the reactor propagate through the fluid and the conditions at the outer edge of the boundary layer change accordingly; this dynamic evolution of the boundary condition can be solved by using a heterogeneous model (see Section 4.1.2).

Boundary Layer. Temperature and concentration conditions in the boundary layer are frequently assumed to be identical to those of the bulk fluid phase. This is not always the case and, particularly if it is desired to trace the dynamic behavior of the concentration and temperature fields inside the catalyst pellets, the temperature and concentration in the boundary layer must be obtained by solving the corresponding mass and energy balances. The first step in solving these fields is to determine the extent of the boundary layer. This can be accomplished by performing overall energy and mass balances across the boundary layer. Once the thickness of the boundary layer has been found (e.g., Δr), the governing equations and their boundary conditions will be defined by:

Mass balance:

$$\frac{\partial C_{\mathrm{f}}}{\partial t} = D_{\mathrm{f}} \nabla^2 C_{\mathrm{f}} \tag{4.1}$$

Energy balance:

$$\varrho_{\mathrm{g}} C_{p,\mathrm{f}} \frac{\partial T_{\mathrm{f}}}{\partial t} = \lambda_{\mathrm{f}} \nabla^2 T_{\mathrm{f}} \tag{4.2}$$

subject to:

$$\left.\begin{aligned} C_{\mathrm{f}} &= C_{\mathrm{f}}(t) \\ T_{\mathrm{f}} &= T_{\mathrm{f}}(t) \end{aligned}\right\} \text{ at } r = R + \Delta r \tag{4.3}$$

$$\left.\begin{aligned} &\frac{\partial C_{\mathrm{f}}}{\partial r} = -\eta R(C_{\mathrm{s}}, T_{\mathrm{s}}) \\ &\lambda_{\mathrm{s}} \left.\frac{\partial T_{\mathrm{s}}}{\partial r}\right|_{r \to R^-} - \lambda_f \left.\frac{\partial T_f}{\partial r}\right|_{r \to R^+} \\ &= \eta(-\Delta H_r) R(C_{\mathrm{s}}, T_{\mathrm{s}}) \end{aligned}\right\} \text{ at } r = R \tag{4.4}$$

where η is the effectiveness factor, the subscript f denotes fluid, and $R(C_{\mathrm{s}}, T_{\mathrm{s}}) = k_0^* C_{\mathrm{s}} \exp(-E_{\mathrm{a}}/RT)$ stands for the equation for the reaction rate [26].

Catalyst Pellet. Similar to the analysis of the boundary layer, the pellet itself is modeled in terms of mass and energy transport. The pellet can be either porous or nonporous; in the latter case there is no mass diffusion of species into the pellet and only an energy balance is necessary. For certain classes of reactions, the assumption of nonporous catalyst is justifiable, e.g., for reactions whose high exothermicity prevents the reaction front from penetrating deep into the pellet even if it were porous. Under these assumptions the energy balance for the catalyst pellet gives:

$$\varrho_{\mathrm{s}} C_{p,\mathrm{s}} \frac{\partial T_{\mathrm{s}}}{\partial t} = \lambda_{\mathrm{s}} \nabla^2 T_{\mathrm{s}} \tag{4.5}$$

subject to:

$$\frac{\partial T_{\mathrm{s}}}{\partial \tau} = 0 \quad \text{at} \quad r = 0$$

$$k_{\mathrm{s}} \frac{\partial T_{\mathrm{s}}}{\partial r}\Big|_{\mathrm{r}\to R^-} - \lambda_f \frac{\partial T_f}{\partial r}\Big|_{r\to R^+} = \eta\left(-\Delta H_r\right) R\left(C_{\mathrm{s}}, T_{\mathrm{s}}\right) \text{ at } r = R \tag{4.6}$$

4.1.2. The Two-Phase Model

Of the two remaining models, the two-phase model is not only more realistic but it also constitutes the basis on which the derivation of the pseudohomogeneous approximation relies.

The most important feature of this model is that temperature and concentration differences are considered between solid and fluid phases. The mass and energy transport processes are governed by the following equations:

Fluid phase:

$$\varepsilon \frac{\partial\left(C_{\mathrm{i,f}}\right)}{\partial t} = -\varepsilon \nabla \cdot \left(\boldsymbol{u} C_{\mathrm{i,f}}\right) + \varepsilon \nabla \cdot \mathbf{D}_{\mathrm{f}} \cdot \nabla C_{\mathrm{i,f}} + a_v \lambda_{\mathrm{f,s}}\left(C_{\mathrm{i,s}} - C_{\mathrm{i,f}}\right) \tag{4.7}$$

$$\varepsilon \varrho_{\mathrm{f}} C_{\mathrm{p,f}} \frac{\partial T_{\mathrm{f}}}{\partial t} = -\varepsilon \varrho_{\mathrm{f}} C_{p,\mathrm{f}} \boldsymbol{u} \cdot \nabla T_{\mathrm{f}} + \varepsilon \nabla \cdot \lambda_{\mathrm{f}} \cdot \nabla T_{\mathrm{f}} + a_v h_{\mathrm{f,s}}\left(T_{\mathrm{s}} - T_{\mathrm{f}}\right) \tag{4.8}$$

Solid phase:

$$(1-\varepsilon)\, \varepsilon_{\mathrm{p}} \frac{\partial\left(C_{\mathrm{i,s}}\right)}{\partial t} = -\eta\,(1-\varepsilon)\, \varrho_{\mathrm{s}} R\left(C_{1,\mathrm{s}}, \ldots, C_{\mathrm{i,s}}, \ldots, C_{\mathrm{S,s}}, T_{\mathrm{s}}\right) + a_v\, \lambda_{\mathrm{f,s}}\left(C_{\mathrm{i,f}} - C_{\mathrm{i,s}}\right) \tag{4.9}$$

$$(1-\varepsilon)\, \varrho_{\mathrm{s}} C_{p,\mathrm{s}} \frac{\partial T_{\mathrm{s}}}{\partial t} = \eta\,(1-\varepsilon)\, \varrho_{\mathrm{s}}\left(-\Delta H_r\right) R\left(C_{1,\mathrm{s}}, \ldots, C_{\mathrm{i,s}}, \ldots, C_{\mathrm{S,s}}, T_{\mathrm{s}}\right) + (1-\varepsilon)\, \boldsymbol{\nabla} \cdot \lambda_{\mathrm{s}} \cdot \boldsymbol{\nabla} T_{\mathrm{s}} + a_v\, h_{\mathrm{f,s}}\left(T_{\mathrm{f}} - T_{\mathrm{s}}\right) \tag{4.10}$$

4.1.3. The One-Phase Model

By introduction of the effective transport concept heterogeneous fluid – solid systems can be treated as quasi- continuum media [27]. In the pseudohomogeneous model, additional transfer mechanisms are superimposed on the contributions that originate from the global movement of the fluid. These additional mechanisms are based upon the observation that the fluid traveling between two points of the packed bed undergoes a tortuous journey that is composed of a large number of random steps. It is assumed that the transfer mechanisms resulting from this process can be treated as diffusion- and conduction-like phenomena. The fact that the modeled system is actually heterogeneous is accounted for by defining transport coefficients where both contributions due to the solid and fluid phases are considered.

Although the two-phase models fulfill the obvious physical requirements of considering a system where two phases exchange mass and/or energy, they also contain a number of approximations. For instance, separation of thermal dispersion effects, which in fact are interconnected, is very difficult. An even more difficult problem arises when heat losses must be accounted for; in this case the wall heat transfer coefficient has to be resolved into its solid- and fluid-phase components. Ramkrishna and Arce discussed the validity of the pseudohomogeneous model to represent the heterogeneous nature of packed-beds [28], [29].

The simplest, most obvious pseudohomogeneous model was derived by Yagi et al. [30]. These authors made use of the fact that, at low Reynolds number, temperature differences between the solid and fluid phases are negligible and both phases can be assumed to have the same temperature. This approach has been generalized, and the one-phase model is adopted, even for high Reynolds number, based on the assumption of equal temperatures in both phases (e.g., [31]).

The idea of equivalence was extended by Vortmeyer et al. [32], [33] for arbitrary Reynolds numbers. These authors combined the energy balances for the fluid and solid phases (Eqs. 4.8 and 4.10) by assuming that the driving forces for heat conduction were similar in both phases. The main difference to the approach followed by Yagi and coworkers is that Vortmeyer and coworkers derived a dispersion term which reflects the exchange of energy between both phases; this term is added to the thermal

conductivity of the solid phase to form a quantity which is known as the effective thermal conductivity.

An additional assumption concerns the mass transport between the solid and fluid phases. The resistance to mass transfer of the film adjacent to the solid particles can often be considered negligible; thus, the surface reactant concentration can be considered similar to that in the bulk of the fluid phase.

Under these assumptions and by combining both energy balances (Eqs. 4.8 and 4.10), the transport processes in the packed-bed reactor can be considered as governed by the following set of equations:

Mass balance:

$$\varepsilon \frac{\partial (C_\mathrm{i})}{\partial t} = -\varepsilon \boldsymbol{\nabla} \cdot (\boldsymbol{u} C_\mathrm{i}) + \varepsilon \boldsymbol{\nabla} \cdot \boldsymbol{D}_\mathrm{e} \cdot \boldsymbol{\nabla} C_\mathrm{i} - \eta (1-\varepsilon)\, \varrho_\mathrm{s} R\, (C_1, \ldots, C_\mathrm{i}, \ldots, C_\mathrm{S}, T) \tag{4.11}$$

Energy balance:

$$\overline{\varrho C_p} \frac{\partial T}{\partial t} = -\varepsilon\, \varrho_\mathrm{f} C_{p,\mathrm{f}}\, \boldsymbol{u} \cdot \boldsymbol{\nabla} T + \boldsymbol{\nabla} \cdot \lambda_\mathrm{e} \cdot \boldsymbol{\nabla} T + \eta (1-e)\, \varrho_\mathrm{s} (-\Delta H_r)\, R\, (C_1, \ldots, C_\mathrm{i}, \ldots, C_\mathrm{S}, T) \tag{4.12}$$

where:

$$\overline{\varrho C_\mathrm{p}} = \varepsilon \varrho_\mathrm{f} C_{p,\mathrm{f}} + (1-\varepsilon)\, \varrho_\mathrm{s} C_{p,\mathrm{s}}$$

An alternative two-phase model was proposed by CARBERRY and WHITE [34]. This model combines the pseudohomogeneous and heterogeneous models; it is based on the catalyst particle but uses effective transport properties. As a hybrid between the two previous models it has been termed the "pseudoheterogeneous" model. By considering the solid phase to be discontinuous, this model appears to perform better when matching models. A thorough analysis of this model is given in [35].

4.1.4. Boundary Conditions

Since the classical works by HULBURT [36], DANCKWERTS [37], and WEHNER and WILHELM [38] appeared, boundary conditions in packed-bed reactors have received much attention. The formulation of the boundary conditions proposed by DANCKWERTS is the most often followed. This formulation meets the limiting mixing conditions of the continuous stirred-tank reactor (CSTR) and the plug-flow reactor (PFR). This was proven for isothermal [39] and nonisothermal [40] systems. These boundary conditions have also been shown to be adequate for systems with a multiplicity of steady states [41]. They also predict the behavior of nonadiabatic systems satisfactorily [42] even though the number of possible steady states in these systems could change if the WEHNER and WILHELM approach is followed [43].

Even though the formulation of the boundary conditions as suggested by WEHNER and WILHELM [38] appears more logical from a physical point of view, their correspondence with DANCKWERTS' suggestion for steady-state operation of adiabatic systems has been proven [44]. Modifications have been proposed when radial dispersion is also taken into account [45], and the boundary conditions have also been modified to treat the transient situation [46], [47].

For selection of the proper boundary and initial conditions for flow systems see [48][49][50]. A simple pseudohomogeneous model without radial dispersion is now considered. The reactor is visualized as a tube with three distinguishable sections (see Fig. 16): the reaction zone packed with the catalyst ($0 \le z \le L$), and a fore ($z < 0$) and aft ($z > L$) sections of inert packing. A general formulation of the inlet and exit boundary conditions can be written as:

Inlet ($z = 0$):

$$\left[D_{\mathrm{e},z} \frac{\partial C_\mathrm{i}}{\partial z} - u_z C_\mathrm{i} \right]_{0-} = \left[D_{\mathrm{e},z} \frac{\partial C_\mathrm{i}}{\partial z} - u_z C_\mathrm{i} \right]_{0+} \tag{4.13}$$

$$\left[\lambda_{\mathrm{e},z} \frac{\partial T}{\partial z} - u_z\, \varrho_\mathrm{f} C_{p,\mathrm{f}} T \right]_{0-} = \left[\lambda_{\mathrm{e},z} \frac{\partial T}{\partial z} - u_z\, \varrho_\mathrm{f} C_{p,\mathrm{f}} T \right]_{0+} \tag{4.14}$$

Exit ($z = L$):

$$\left[D_{\mathrm{e},z} \frac{\partial C_\mathrm{i}}{\partial z} \right]_{L-} = \left[D_{\mathrm{e},z} \frac{\partial C_\mathrm{i}}{\partial z} \right]_{L+} \tag{4.15}$$

$$\left[\lambda_{\mathrm{e},z} \frac{\partial T}{\partial z} \right]_{L-} = \left[\lambda_{\mathrm{e},z} \frac{\partial T}{\partial z} \right]_{L+} \tag{4.16}$$

where a given approach is followed by selecting the diffusivities/conductivities as indicated in Table 4 ("yes" means parameter is included in the model, "no" means not included).

Table 4. Dispersion coefficients for different boundary conditions

Parameter	Hulburt	Danckwerts	Wehner and Wilhelm
$D/\lambda_{e,z}$ — $_{0-}$	no	no	yes
$D/\lambda_{e,z}$ — $_{0+}$	no	yes	yes
$D/\lambda_{e,z}$ — $_{L-}$	yes	yes	yes
$D/\lambda_{e,z}$ — $_{L+}$	no	no	yes

The conditions for the mass balance at the lateral walls are such that it is assumed that no mass flux occurs because the lateral walls are often impermeable (radial reactors are an exception). The boundary conditions for the energy balance, however, vary depending on whether the reactor is operated adiabatically or nonadiabatically. For a pseudohomogeneous model with radial dispersion these boundary conditions are:

Reactor wall (r=R):

$$\frac{\partial C_i}{\partial r} = 0 \tag{4.17}$$

$$\lambda_{e,r}\frac{\partial T}{\partial r} = h_w\,(T_w - T) \tag{4.18}$$

Reactor centerline (r=0):

$$\frac{\partial C_i}{\partial r} = 0 \tag{4.19}$$

$$\frac{\partial T}{\partial r} = 0 \tag{4.20}$$

4.2. Values of the Parameters

Various steps have been taken towards the derivation of an integral model for the packed-bed reactor by considering the individual steps involved in the transport of a given property in a packed bed. In this section only few correlations for typical applications will be mentioned, for more details, see [25], [51][52–54].

4.2.1. Average Bed Porosity

The existence of oscillatory radial variations of the void fraction in packed beds has been recognized since the early work of Roblee et al. [55] and Benenati and Brosilow [56], [57]. This nonuniform distribution of the porosity causes the flow to be nonuniformly distributed, particularly near the reactor walls [58], [59]. The inclusion of radial porosity and velocity profiles improves the agreement between experimental and theoretical results [60], [61]. It has been found [62], however, that the predictions obtained by assuming constant average porosity and main-flow velocity for the entire packed bed will be qualitatively correct in most cases.

$$\left[D_{e,z}\frac{\partial C_i}{\partial z} - u_z C_i\right]_{0^-} = \left[D_{e,z}\frac{\partial C_i}{\partial z} - u_z C_i\right]_{0^+} \tag{4.13}$$

$$\left[\lambda_{e,z}\frac{\partial T}{\partial z} - u_z \varrho_f C_{p,f} T\right]_{0^-} = \left[\lambda_{e,z}\frac{\partial T}{\partial z} - u_z \varrho_f C_{p,f} T\right]_{0^+} \tag{4.14}$$

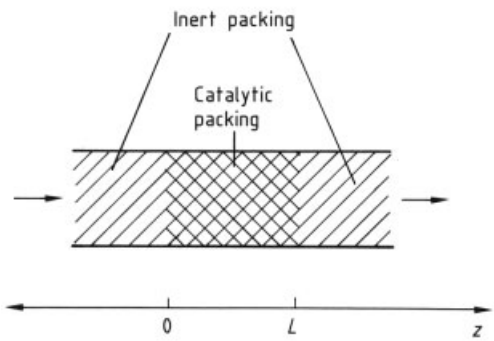

Figure 16. Schematic of the flow system in a packed-bed reactor

In order to use a constant average porosity, the following integral relation must be fulfilled:

$$\varepsilon = \frac{1}{\pi R^2}\int_0^R \varepsilon(r)\, r\,\mathrm{d}r$$

There is no unique functional relation between the radius and the porosity, but many functionalities have been proposed in the literature. To find a realistic value for the average porosity, the approach suggested by Martin can be followed [63]. The bed is considered to be formed by two concentric rings with their own properties; a central core with an average porosity ε_0 and a one-particle-diameter-thick concentric shell with porosity ε_w. The average porosity, as defined above, must satisfy the relation:

$$\varepsilon_0 A_0 + \varepsilon_w A_w = \varepsilon A$$

Then, from geometric considerations, the following equation results:

$$\varepsilon = \varepsilon_w - (\varepsilon_w - \varepsilon_0)(1 - d/D)^2$$

where the porosity for the outer shell (ε_w) is estimated as [63]:

$$\varepsilon_w = 1 - (1 - \varepsilon_{min}) \frac{2}{3} \left(\frac{D/d - 7/8}{D/d - 1/2} \right)$$

where ε_{min} is the minimum value observed for the void fraction ($\varepsilon_{min} = 0.23$ and $\varepsilon_0 = 0.39$ are taken from [56]).

4.2.2. Effective Transport Coefficients

Since the effective transport coefficients are mainly determined by the flow characteristics, in general, packed beds are anisotropic for effective transport, so that the radial transport component is different from its analogue in the axial direction. For the sake of simplicity, effective transport coefficients are often considered to be unaffected by secondary flow and dependent on the mainflow (forced) only.

Mass Transport. A correlation for the mass transport in both spatial directions is [64]:

$$D_{e(r,z)} = \langle u \rangle_z d \times \left[\frac{0.7}{ReSc} + \frac{C_1}{1 + \left(\frac{C_2}{ReSc} \right)} \right]$$

where the constants C_1 and C_2 depend on the geometry of the packing and the fluid flowing through the packed bed as listed in Table 5.

Table 5. Constants for the effective mass-transport coefficients [64]

Constant	Liquids		Gases	
	Spherical packing	Irregular packing	Spherical packing	Irregular packing
C_1 (longitudinal)	2.5	2.5	0.7	4
C_1 (radial)	0.08	0.08	0.12	0.12
C_2 (longitudinal)	8.8	7.7	5.8	5.1
C_2 (radial)			78 ± 20	

A more heuristic approach is to assume a constant value for the axial and radial Bodenstein (or mass Péclet) numbers, i.e.,

$$Pe_{m,z/r} = \frac{\langle u \rangle_z d}{D_{e,z/r}} = \text{constant}$$

where z/r denotes that the expression is valid for both axial and radial transport.

Experimental evidence has shown that *axial dispersion* in packed beds is typically characterized by $Pe_{m,z} \approx 2$. In other words, CSTR conditions (small $Pe_{m,z}$) are reached in the void within each void cell, and the packed-bed reactor can be viewed as (L/d) CSTRs in series [65]. Therefore, axial mass diffusion can be neglected for ratios $L/d > 50$. This criterion applies to systems with particle Reynolds number larger than unity (i.e., $Re_p = \varrho_f \langle u \rangle_z d/\mu_f > 1$).

Radial mass dispersion is characterized by a constant radial mass Péclet number:

$$Pe_{m,r} = \frac{\langle u \rangle_z d}{D_{e,r}} \approx 8 - 10$$

Energy Transport. Energy transport is supplemented by radiation and conduction in the solid phase. Therefore, it is more difficult to define constant values for the heat Péclet numbers. The problem of equivalence between the heterogeneous and pseudohomogeneous models has been addressed in [51], [66].

In general the effective thermal conductivity $\lambda_{e(r,z)}$ is formed by two contributions: the quiescent bed conductivity λ_s^0 and a contribution due to the flow, i.e.,

$$\lambda_{e(r,z)} = \lambda_s^0 + C \lambda_f RePr$$

where the conductivity for the quiescent bed λ_s^0 is obtained from the correlation proposed by Zehner and Schlünder [67], while the contribution due to the flow can be found from the analysis by Vortmeyer and Berninger [66]. The constant in the above equation is chosen as $C = 0.1$ for the radial (r) direction and $C = 0.1 - 0.2$ for the axial (z) direction.

The equivalence between the heterogeneous and pseudohomogeneous models has been proved for nonreactive conditions [51]. For further information on energy transport in packed beds see [51], [52], [68–71].

Effect of Radiation. Transport by radiation is frequently ignored when estimating heat transfer coefficients. This transport mechanism depends on the temperature range for the reactor operation; for a temperature of 600 K at the reactor wall, radiation has been found to contribute up to 20 % to the overall radial heat flux [72].

The contribution of radiation to effective energy transport in packed beds has been analyzed [72], [73]. In the model presented the heat transfer is considered to be controlled by two parallel mechanisms and the overall effective thermal conductivity is assumed as

$$\lambda_{\mathrm{e},r} = \lambda^0_{\mathrm{e},r} + \lambda^r a\, d_{\mathrm{e},r}$$

where the contribution due to radiation $\lambda^{\mathrm{r}}\, ad_{\mathrm{e},\,r}$ is

$$\lambda^r a d_{\mathrm{e},r} = 4\psi\sigma\, \mathrm{d}T^3 = \left(\frac{8}{2/e - 0.264}\right)\sigma\, \mathrm{d}T^3 = C_r T^3$$

where σ is the Stefan – Boltzmann constant ($5.67 \times 10^{-8}\,\mathrm{J\,m^{-2}\,K^{-4}\,s^{-1}}$), ψ is a shape factor, and e is the particle emissivity. The treatment of this contribution is carried out by adding an additional term to the conduction term in the energy balance, i.e.,

$$\nabla \cdot \lambda_{\mathrm{e}} \cdot \nabla T = \frac{1}{r}\frac{\partial}{\partial r}\left[\left(r\,\lambda^0_{\mathrm{e},r}\frac{\partial T}{\partial r}\right) + \left(r\,C_r\,T^3\frac{\partial T}{\partial r}\right)\right] + \ldots$$
$$= \lambda_{\mathrm{e},r}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r}\right) + 3C_r\,T^2\left(\frac{\partial T}{\partial r}\right)^2 + \ldots$$

It is apparent that this contribution can become important with increasing particle emissivity and its omission would lead to conservative temperature predictions.

4.2.3. Wall Heat-Transfer Coefficient

The wall heat-transfer coefficient can be estimated from many correlations. The correlations available, however, predict heat transfer coefficients which are largely scattered. For instance, for a particle Reynolds number $Re = 100$, the heat transfer coefficient can differ by almost one order of magnitude [53]. Odendaal et al. analyzed the available correlations and compared the radial temperature profiles predicted for the solid and fluid phases. Based on the physical restriction that both profiles can never cross, most of them were discarded and only a few could be considered reliable [53].

The most often accepted correlation has been presented by Dixon et al. [74]. A thorough analysis of the estimation of this coefficient has been presented [75]. The recommended correlation has the form:

$$Nu_{\mathrm{f,\,w}} = \frac{\alpha_{\mathrm{f,\,w}}\, d}{k_{\mathrm{f}}} = \left[1 - 1.5\,(D/d)^{-1.5}\right] Pr^{1/3} Re^{0.6} \tag{4.21}$$

where $\alpha_{\mathrm{f,\,w}}$ is the fluid-to-wall heat transfer coefficient.

4.3. Further Simplifications

Despite its simplicity, the pseudohomogeneous model has proven to be a very reliable representation for transport processes in packed beds [66]. If care is taken when evaluating the transport coefficients, predictions obtained from a one-phase model will not differ significantly from those yielded by a two-phase representation. Every model has its deficiencies, thus, the heterogeneous model cannot capture the radial temperature and concentration gradients present in the pellets under reactive conditions, and the pseudohomogeneous model lacks an adequate representation of the solid phase. Windes et al. [54] have suggested an approach to achieve equivalence between the pseudohomogeneous and heterogeneous models.

The questions concerning equivalence stem from the fact that packed beds behave differently under reactive and nonreactive conditions. This was initially interpreted as a change in the heat transport coefficients under reactive conditions [76]. Wijngaarden and Westerterp [77] have addressed the subject in detail and provided a comprehensive explanation for the difference between reactive and nonreactive situations. For the mixing coefficients independence of chemical reaction was demonstrated by Gunn and Vortmeyer [78].

One-Dimensional Models versus Two-Dimensional Models Although radial temperature profiles can be minimized in the design of the reactor, it is not necessary to suppress radial dependence of temperature to obtain reliable predictions with one-dimensional models. In general, the sophistication of the model should never exceed that of the experimental data. For this reason, one-dimensional models have become increasingly useful.

BEEK and SINGER [79] addressed this point by assuming a weak parabolic radial dependence of the temperature profiles on the radial dimension. Then it can be shown that predictions from a one-dimensional model equal the average radial temperature $\langle T \rangle$ for a two-dimensional model [25]. The radial temperature distribution is then obtained as [25]:

$$T(r/R) \approx T_w + \left[\frac{1-(r/R)^2 + (2/B_\mathrm{i})}{1+(4/B_\mathrm{i})}\right] 2\left(\langle T \rangle - T_w\right)$$

The equivalence between the one-dimensional model and the radial average temperature for a two-dimensional model is obtained by a proper computation of the overall heat-transfer coefficient. For a parabolic temperature profile:

$$\frac{1}{U} = \frac{1}{h_w} + \frac{D}{8k_{\mathrm{e},r}}$$

A more refined approach was followed by CRIDER and FOSS [80], who solved the heat transfer problem analytically under nonreactive conditions. The equivalence of one/two-dimensional models was achieved by combining the radial heat-transfer parameters as follows:

$$\frac{1}{U} = \frac{1}{h_w} + \frac{D}{6.133\lambda_{\mathrm{e},r}}$$

This was later confirmed by HLAVACEK et al. [81] who solved the linearized problem and found identical results. An important theoretical conclusion can be drawn from this last observation; as long as only small radial temperature gradients exist, the chemical reaction only seems to have a significant effect on the axial temperature dependence. Under these circumstances, the radial temperature dependence can be described via a radially lumped model.

Example. The above analysis is illustrated by solving for the steady-state temperature and concentration profiles in a typical industrial reactor with the partial oxidation of *o*-xylene to phthalic anhydride as an example. This highly exothermic process is carried out in heat-exchanger-type reactors. The reactor consists of a bundle of tubes with a cooling fluid circulating around the tube arrangement. The process (of German origin, but used worldwide) employs V_2O_5 on silica gel pellets as catalyst. The geometric description and kinetic information have been taken from [82].

If the pseudohomogeneous, two-dimensional model is lumped radially, i.e., if the equations are integrated across the cross-sectional area, the following steady-state, one-dimensional model results:

$$0 = -\frac{\partial}{\partial z}(u_z C_\mathrm{X}) + D_{\mathrm{e},z}\frac{\partial^2 C_\mathrm{X}}{\partial z^2} - (1-\varepsilon)\,\varrho_\mathrm{s} R\left(C_\mathrm{X}, \langle T \rangle\right) \tag{4.22}$$

$$0 = -\frac{\partial}{\partial z}\left(\varrho_\mathrm{f} C_{p,\mathrm{f}} u_z \langle T \rangle\right) + \lambda_{\mathrm{e},z}\frac{\partial^2 \langle T \rangle}{\partial z^2} - \frac{4U}{D}\left(\langle T \rangle - T_w\right) + (-\Delta H_r)(1-\varepsilon)\,\varrho_\mathrm{s} R\left(C_\mathrm{X}, \langle T \rangle\right) \tag{4.23}$$

where $R(C_\mathrm{X}, \langle T \rangle) = k_0\, p_\mathrm{O}\, p_\mathrm{X} \exp(-E_\mathrm{a}/R_\mathrm{u} \langle T \rangle)$ represents the rate of reaction, and p_X and p_O represent the partial pressures of *o*-xylene and oxygen, respectively. The reaction is performed in excess oxygen (i.e., $p_\mathrm{O}/p_\mathrm{tot} = 0.21 \approx$ constant). Although the reaction mechanism for this process is completed by parallel and consecutive reactions, only the main reaction is addressed here.

The results obtained with the one-dimensional (Eqs. 4.22 and 4.23) and two-dimensional models are presented in Figure 17 for two different operating conditions. The one-dimensional model predicts reactor behavior satisfactorily for the conditions of Figure 17 A ($T_w = 630$ K). For the conditions of Figure 17 B ($T_\mathrm{w} = 645$ K), however, the differences between this model and the two-dimensional models are significant. This can be better understood by analyzing the results presented in Figure 18 where the radial temperature profile is shown as a function of radial position r/R at different

axial locations. The high-temperature situation (Fig. 18 B) shows the nonlinear effect of the reaction rate and the temperature profile deviates markedly from a parabolic dependence(as shown Fig. 18 A). These results show that, for mild radial dependence of the temperature distribution, one-dimensional models can be used as a reliable design tool.

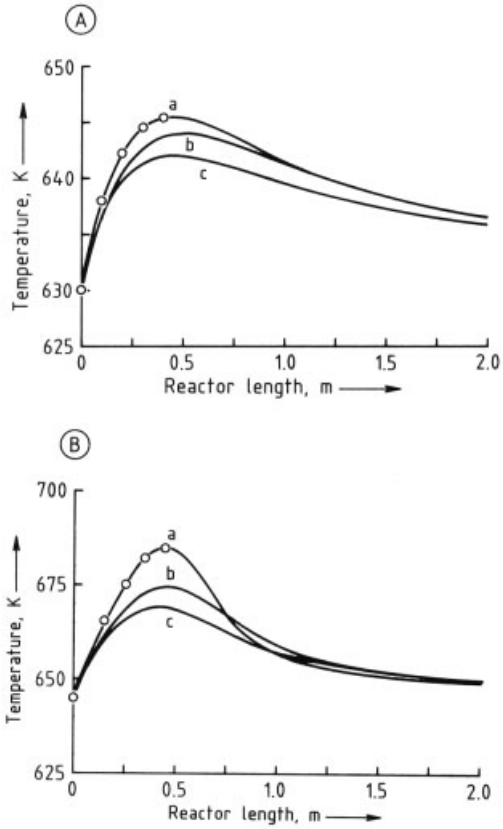

Figure 17. Comparison between one- and two-dimensional models
A) $T_w = 630$ K, $x_{x,0} = 0.00924$; B) $T_w = 645$ K, $x_{x,0} = 0.00924$
a) Two-dimensional model; b) Crider and Foss model (one-dimensional; c) Beek and Singer model (one-dimensional)
Open circles in Figures 17 A and 17 B denote axial locations for curves shown in Figures 18 A and 18 B, respectively.

4.4. Parametric Sensitivity

Since the 1960s, great improvements have been made in the design and operation of catalytic reactors. In industrial conditions, however, due to the high exothermicity of the chemical reaction, the temperature rises sharply in the catalytic bed towards a maximum or hot spot (usually located near the reactor inlet). These sharp axial temperature gradients can cause poor reaction selectivity and extreme temperatures can be responsible for rapid deactivation or even deterioration of the catalyst. Therefore, the hot spot must be kept within permissible bounds. Furthermore, the attainable conversion frequently has an upper safety limit at levels where the temperature profile becomes extremely sensitive to changes in operational and/or physicochemical parameters. Bilous and Amundson [83] termed such a reactor condition as "parametric sensitivity". Many studies have been reported on the prediction of this phenomenon, which in turn can lead the reactor to runaway operation [23], [84]. The danger inherent in running such systems is widely recognized, and the final reactor design must guarantee a safe mode of operation.

Most studies on parametric sensitivity have assumed constant temperature for the cooling medium [85–91]. This is a suitable approach for perfectly mixed coolant, or reactors employing a boiling liquid or a fluid of abnormally large heat capacity as cooling medium. However, in the more common case of molten salt circulation, the thermal gradients in the shell side cannot be neglected in the reactor model. Few papers have studied the nonisothermal situation. For more information; see [92–106].

Example. Consider again, the production of phthalic anhydride via partial oxidation of *o*-xylene in a multitube heat-exchanger-type reactor. The steady-state operation will be described by the one-dimensional pseudohomogeneous model (Eqs. 4.22 and 4.23). Parametric sensitivity can be estimated as originally proposed by Bilous and Amundson [83]. It is computed as the derivative of the variable to be analyzed with respect to one of the physicochemical parameters of the process. Thus, the parametric sensitivity (S_{ij}) of the variable χ_i with respect to the parameter π_j is defined as:

$$S_{ij} = \frac{\partial \chi_i}{\partial \pi_j}$$

In what follows the effect of the reactor wall temperature on the reactor thermal condition is analyzed, i.e., $\pi = T_w$ and $\chi = \langle T \rangle$. This is a typical problem found in industrial practice where

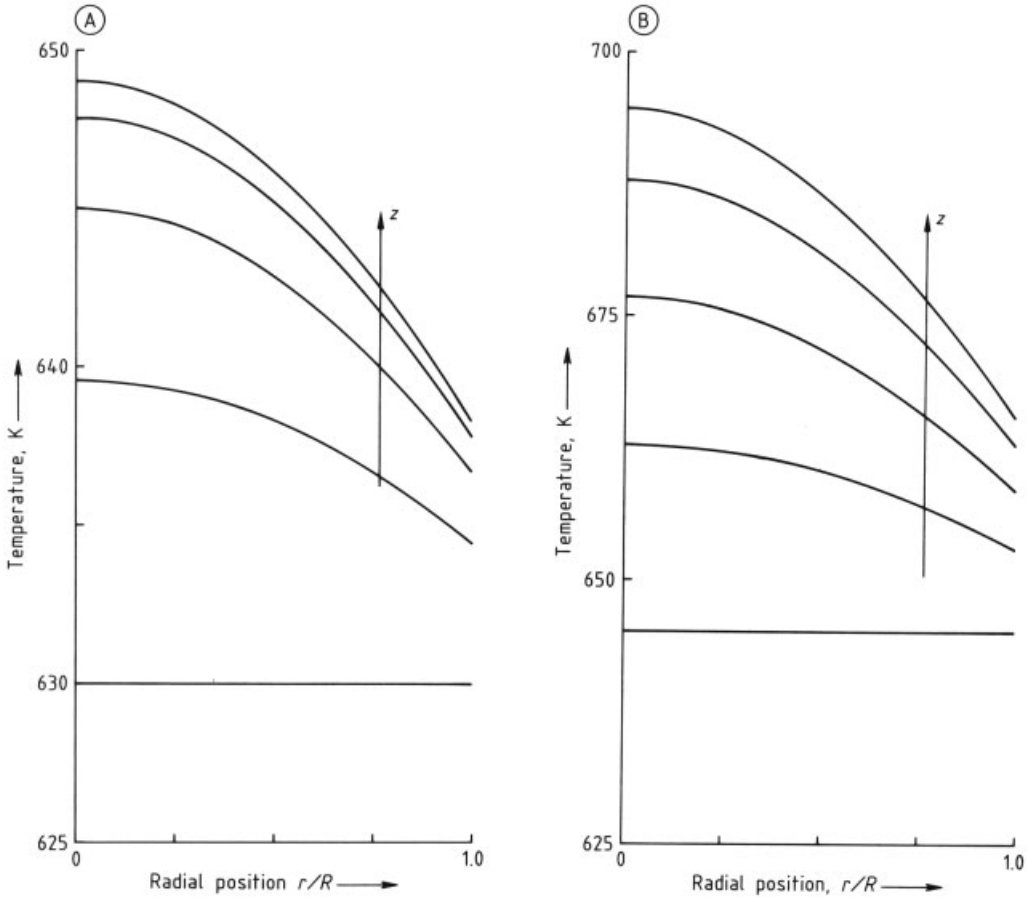

Figure 18. Radial temperature profiles
A) $T_w = 630$ K, $x_{\mathrm{x},\,0} = 0.00924$; B) $T_{\mathrm{w}} = 645$ K, $x_{\mathrm{x},\,0} = 0.00924$

the thermal state of the reactor is monitored (i.e., the observed variable would be the readouts of the reactor thermocouples) and the operator acts on the cooling fluid (i.e., the manipulated variable would be the temperature or coolant flow rate) to keep the reactor operating near the desired state.

In order to obtain numerical values for the parametric sensitivity of the reactant concentration (S_C) and the bed temperature (S_T), the governing equations are differentiated with respect to the wall temperature, i.e.,

$$0 = -\frac{\partial}{\partial z}\left(u_z S_C\right) + D_{\mathrm{e},z}\frac{\partial^2 S_C}{\partial z^2} - (1-\varepsilon)\,\varrho_{\mathrm{s}}\frac{\partial R}{\partial T_w} \tag{4.24}$$

$$0 = -\frac{\partial}{\partial z}\left(\varrho_{\mathrm{f}} C_{p,\mathrm{f}} u_z S_T\right) + k_{\mathrm{e},z}\frac{\partial^2 S_T}{\partial z^2} - \frac{4U}{D}\left(S_T - 1\right) + \left(-\Delta H_r\right)(1-\varepsilon)\,\varrho_{\mathrm{s}}\frac{\partial R}{\partial T_{\mathrm{w}}} \tag{4.25}$$

where:

$$\frac{\partial R}{\partial T_\omega} = k_0 p_{\mathrm{O}}\left(S_C + p_{\mathrm{X}}\frac{E_{\mathrm{a}}}{R_{\mathrm{u}}\langle T\rangle^2}S_T\right)\exp\left(-E_{\mathrm{a}}/R_{\mathrm{u}}\langle T\rangle\right)$$

If Equations (4.24) and (4.25) are integrated simultaneously with the reactor model (i.e., Eqs. 4.22 and 4.23), the sensitivity of the reactant conversion and bed temperature to changes in

the coolant temperature is obtained at any point of the reactor length. Of these two sensitivity profiles, only $S_T(z)$ guarantees a safe mode of operation.

The four additional boundary conditions needed are derived from the original boundary conditions, i.e.,

Inlet, $z = 0$:

$$0 = u_z S_C - D_{e,z} \frac{\partial S_C}{\partial z} \tag{4.26}$$

$$\varrho_f C_{p,f} u_z = \varrho_f C_{p,f} u_z S_T - k_{e,z} \frac{\partial S_T}{\partial z} \tag{4.27}$$

Exit, $z = L$:

$$0 = \frac{\partial S_C}{\partial z} = \frac{\partial S_T}{\partial z} \tag{4.28}$$

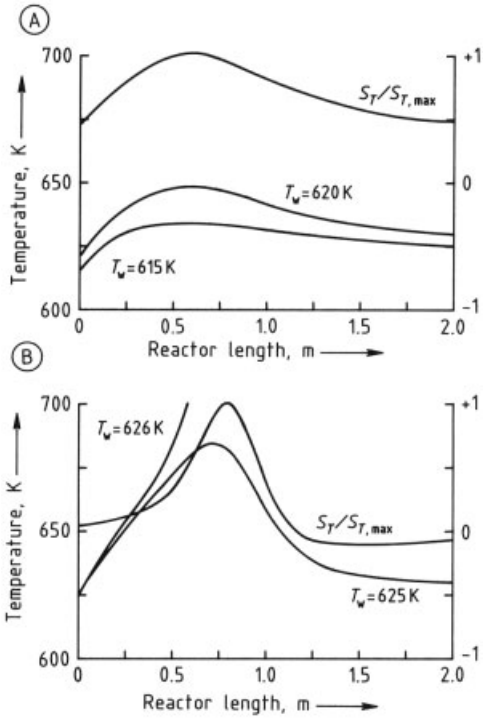

Figure 19. Parametric sensitivity in packed-bed reactors for different inlet temperatures and $x_{x,0} = 0.019$
A) $T_w = 620$ K and 615 K; B) $T_w = 626$ K and 625 K

In Figure 19 the axial temperature profiles for the previous example are shown for two different inlet temperatures and the normalized temperature sensitivity associated with the lower inlet temperature. Figure 19 B illustrates the concept of parametric sensitivity; a 1 K rise in the inlet temperature is sufficient to drive the reactor away from a safe operation regime. Figure 19 A, on the other hand, only shows a small difference in the reactor thermal state after a change of 5 K in the inlet temperature. Indeed the 5 K temperature increase is transmitted almost linearly along the reactor. These two situations can be better understood by inspecting the normalized temperature sensitivity associated with the lower inlet temperature operating condition. In Figure 19 A the temperature sensitivity remains almost constant along the reactor, i.e., the temperature profile should experience an almost constant increase for a temperature rise at the inlet of the reactor. In Figure 19 B, on the other hand, the temperature is predicted to grow dramatically in the vicinity of the hot spot. Furthermore, the sensitivity crosses to negative values after the hot spot axial location; this suggests a temperature overshoot promoting a high degree of conversion and mostly cooling afterwards. Indeed, the temperature profiles corresponding to the higher inlet temperature confirm the predictions extracted from analyzing the temperature sensitivity.

4.5. Flow Field Description

Usually the flow distribution in unbounded porous media is assumed to be well represented by a linear relation between the pressure drop and the fluid velocity (Darcy's law). When using such an approximation, however, two main points are disregarded:

1) Effect of the boundaries on the flow field
2) Increasing importance of the inertial effects as the flow speed increases

If the fluid obeys the Boussinesq approximation, the flow field will be governed by [110]:

$$\varrho_f \frac{\partial \mathbf{u}}{\partial t} = -\nabla p - \frac{\mu_f}{\varkappa} \mathbf{u} - \varrho_f \left[\beta_1 (T - T_0) + \beta_2 \frac{1}{C_0} (C_0 - C) \right] \boldsymbol{g} \tag{4.29}$$

$$\nabla \cdot \mathbf{u} = 0 \tag{4.30}$$

which is known as the Darcy – Oberbeck – Boussinesq model for flow through a porous medium [111].

As the flow speed increases, the *inertial terms* become important and the results of Darcy's model are inefficient in describing the momentum transfer in packed beds. The flow through porous media is now said to be described by a "*modified*" *Darcy's law*:

$$\varrho_{\mathrm{f}}\left(\frac{\partial \mathbf{u}}{\partial t}+\frac{1}{e}\mathbf{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right)=-\boldsymbol{\nabla}p-\frac{\mu_{\mathrm{f}}}{\varkappa}\boldsymbol{u}$$
$$-\varrho_{\mathrm{f}}\left[\beta_1\left(T-T_0\right)+\beta_2\frac{1}{C_0}\left(C_0-C\right)\right]\boldsymbol{g} \tag{4.31}$$

where the inertial forces are represented by the term $\mathbf{u}\cdot\boldsymbol{\nabla}\mathbf{u}$. Even though this term arises from a formal volume averaging in the point field equations [112], its inclusion may lead to inconsistencies between boundary conditions and governing equations [113]. FORCHHEIMER [114] first proposed addition of higher order inertial terms to the relation between pressure drop and fluid velocity and modified Darcy's equation to:

$$\frac{\mathrm{d}p}{\mathrm{d}z}+a_1u_z+a_2\left(u_z\right)^2=0$$

Later he added a third-order term to fit experimental data. FORCHHEIMER'S modification for the one-dimensional flow can be formalized for two and three dimensions as [115], [116]:

$$\boldsymbol{\nabla}p-\varrho_{\mathrm{f}}\boldsymbol{g}+\frac{\mu_{\mathrm{f}}}{\varkappa}\mathbf{u}+\frac{\varrho_{\mathrm{f}}}{\varkappa}b\mathbf{u}\,|\mathbf{u}|=0$$

where b can be considered a structure property associated with inertia effects in the porous matrix. Then the momentum equation becomes:

$$\varrho_{\mathrm{f}}\frac{\partial \mathbf{u}}{\partial t}=-\boldsymbol{\nabla}p-\frac{\mu_{\mathrm{f}}}{\varkappa}\mathbf{u}-\frac{\varrho_{\mathrm{f}}}{\varkappa}b\mathbf{u}\,|\mathbf{u}|$$
$$-\varrho_{\mathrm{f}}\left[\beta_1\left(T-T_0\right)+\beta_2\frac{1}{C_0}\left(C_0-C\right)\right]\boldsymbol{g} \tag{4.32}$$

4.5.1. Boundary Conditions

Only two boundary conditions are necessary to solve the momentum equations. The classical nonslip boundary conditions are imposed on the solid walls and a given velocity profile is imposed at the entrance and exit of the reactor, i.e.,

Reactor wall ($r=R$):

$$\mathbf{u}=[0,\,0] \tag{4.33}$$

Reactor centerline ($r=0$):

$$\mathbf{u}=[0,u_z] \tag{4.34}$$

Inlet ($z=0$):

$$\mathbf{u}=[0,u_z\,(r)] \tag{4.35}$$

Exit ($z=L$):

$$\mathbf{u}=[0,u_z\,(r)] \tag{4.36}$$

4.5.2. Permeability and Inertia Coefficient

The pressure drop in a porous medium can be approximated by the Ergun correlation, a linear combination of the Carman – Kozeny and Burke – Plummer equations [117]:

$$\frac{\Delta p}{H}=\frac{150\,(1-\varepsilon)^2}{\varepsilon^3d^2}\mu u_z+\frac{1.75\,(1-\varepsilon)}{\varepsilon^3d}\varrho u_z^2$$

However, for a certain range of fluid velocities, a linear dependence of the pressure drop on the fluid velocity (namely, the Carman – Kozeny term) provides a reliable representation description. Thus, as Darcy's law proposes a linear dependence between pressure drop and fluid velocity, the above equation can be used to estimate the packing permeability $\varkappa$ and the Forchheimer correction term b as:

$$\varkappa=\frac{d^2\varepsilon^3}{150\,(1-\varepsilon)^2}$$

$$b=\frac{1.75d}{150\,(1-\varepsilon)}$$

4.5.3. Thermal-Expansion Coefficient

The thermal-expansion coefficients are tabulated for most of the fluids of interest. When not available, a satisfactory estimation can be obtained for gaseous fluids by resorting to the ideal gas law. For instance, the expansion coefficient for temperature variations can be approximated as:

$$\varrho=\frac{pM}{RT}$$

$$\frac{\partial\varrho}{\partial T}=\frac{\varrho}{T}$$

then:

$$\beta_1\approx\frac{1}{T_0}\approx 10^{-3}$$

4.5.4. Mass-Expansion Coefficient

The expansion coefficient for concentration changes is not as readily obtained as its temperature analogue. A satisfactory estimation can still be obtained by assuming ideal gas behavior. For instance, for a binary mixture A + B where the reaction A → B takes place, the fluid density can be written as:

$$\varrho = \varrho_0 \left(\frac{1}{1 + cC_{\mathrm{A}} y} \right)$$

where c stands for a combination of the initial and final molar masses:

$$c = \frac{M_{\mathrm{i}} - M_{\mathrm{f}}}{M_{\mathrm{f}}}$$

If $c < 0$, then a lower molar mass product is formed in the chemical reaction (for instance, a decomposition reaction). In contrast $c > 0$ would indicate the formation of a product with a higher molar mass (for instance, a polymerization reaction).

If a perturbation is considered in a situation where no conversion has been achieved (i.e., $\xi \to 0$), the gas density can be represented in a linear approximation as:

$$\varrho = \varrho_0 \left(1 - \beta_2 C_{\mathrm{A}} \xi\right)$$

Taking into account that $c < 1$ and $\xi < 1$, the dependence of the density on the conversion can be written as:

$$\varrho \approx \varrho_0 \left(1 - cC_{\mathrm{A}} \xi\right)$$

then:

$$\beta_2 \approx c$$

In the case of a reaction in the liquid phase, the density must be determined from an appropriate equation of state, and the expansion coefficients should be found experimentally.

4.6. Thermomechanical Effects in the Reaction System

Several phenomena can be observed when analyzing the dynamic behavior of catalytic and noncatalytic packed-bed reactors, e.g., "parametric sensitivity", "wrong-way" behavior, travelling reaction fronts, and ignition and extinction phenomena. All these temperature excursions may disappear or be too small to drive the system out of control. However, the development of thermal stresses during these temperature excursions adds a new factor to the reactor design and operation. The mechanical behavior of the catalyst pellet can be analyzed following the basic theory of elastic behavior of solids (see [118]). For a general situation the equations can be written as:

Energy balance:

$$\varrho_{\mathrm{s}} C_{p,\mathrm{s}} \frac{\partial T}{\partial t} = k_{\mathrm{s}} \boldsymbol{\nabla}^2 T - (3\lambda + 2\mu)\, \alpha T_0 \frac{\partial}{\partial t} \left(\boldsymbol{\nabla} \cdot \mathbf{u}\right) \tag{4.37}$$

Thermomechanical equations:

$$\varrho_{\mathrm{s}} \frac{\partial^2 \delta}{\partial t^2} = \mu \boldsymbol{\nabla}^2 \delta + (\lambda + \mu)\, \boldsymbol{\nabla} \left(\boldsymbol{\nabla} \cdot \delta\right) - (3\lambda + 2\mu)\, \alpha \boldsymbol{\nabla} T \tag{4.38}$$

subject to:

$$\delta_r = 0 \quad \text{at } r = 0$$
$$\sigma_{rr} = 0 \quad \text{at } r = \mathrm{R} \tag{4.39}$$

where λ and μ are the Lamé constants, α is the coefficient of linear expansion, T_0 the temperature of stressless condition, $\boldsymbol{\delta}$ is the displacement vector, and σ_{rr} are the radial stresses.

Thus, the energy balance has been augmented from its previous form (i.e., Eq. 4.5 in Section 4.1.1), by an additional term. This term represents what is known as thermoelastic dissipation. Nevertheless, the temperature fluctuations due to deformations are usually small and, for most situations, they can be neglected. A more systematic criterion is to neglect the thermoelastic dissipation term whenever the inequality below holds:

$$\frac{(3\lambda + 2\mu)^2\, \alpha^2 T_0}{(\lambda + 2\mu)\, \varrho_{\mathrm{s}} C_{p,\mathrm{s}}} << 1$$

Another noticeable feature of the above equations is the hyperbolic nature of the thermoelastic equations. There is a close connection between the coupling discussed previously and neglection of the inertia term in the equation of motion. Indeed, the possibility of omitting the coupling terms does not depend only on satisfying the previous indicated inequality, but also on

the fact that the strain rates must be of the same order as the cooling or heating rates. This implies that the temporal behavior of the displacement vector closely follows the dynamic behavior of the temperature field. In other words, no pronounced lag or vibrations in the motion of the body must arise. Under these circumstances, the inertia term can be neglected and the equations become decoupled [119].

This model is completed by the compability and equilibrium equations (see Chap. 8 in [118]).

4.7. Numerical Simulation

(→ Mathematics in Chemical Engineering, Chap. 7.)

Due to the strongly nonlinear nature of the equations describing reactive flows, up to now analytical techniques have failed to produce solutions for the full form of the governing equations. Nevertheless, considerable progress in understanding reaction systems has been made possible through the application of analytical techniques to some limiting situations. For instance, since 1980 a wealth of information has been accumulated that shows results based on the use of large activation energy asymptotics. Although credit should be given to similar approaches for providing physical insight into many reaction problems, the models described are simplified and limitations on the techniques clearly exist. To overcome these limitations the systems of governing equations, together with the appropriate boundary conditions, must normally be solved by numerical techniques.

For the application of numerical methods, a number of grid points are distributed over the physical domain or the domain is cut up into small pieces (cells). The discretization of the governing equations yields a system of algebraic equations expressing relationships between values of the different variables at neighboring grid points.

Because of finite computer resources, it is not possible to monitor the solution at every point of the region in the space on which the problem has been defined. A finite number of grid points is therefore distributed over the field. Each of these points acts as an observer, monitoring the solution locally and relaying information about it to its neighbors. In most situations a regular, logical connectivity exists between the grid points; thus, in a two-dimensional situation each grid point is surrounded by eight neighbors. Because of this structure, a duality exists. Instead of looking at grid points, every four-point group could be seen as a quadrilateral cell. Discretization of the physical domain will thus be seen as constructing a collection of cells covering the entire domain.

However, efficiency in distributing these observers should be the guiding principle; a limited, fixed number of grid points or grid cells should be arranged in such a way that the largest possible amount of information can be extracted from their use. First, the geometry of the physical domain should be approximated in an accurate way and, second, the grid points should be able to monitor all essential features of the evolving solution without leaving gaps in any region of importance.

4.7.1. Discretization of the Physical Domain

In recent years the method of Orthogonal Collocation (OC) [120] (→ Mathematics in Chemical Engineering, Chap. 7.5., → Mathematics in Chemical Engineering, Chap. 7.6.) has been used successfully in cases of moderate values of the Damköhler number [31] even when the system is highly sensitive to changes in the operating conditions [62]. However, this approach gives inappropriate results at large values of the Péclet and/or Damköhler numbers when boundary layers associated with the flow and/or chemical reaction exist.

Certainly, in combustion-type problems most of the important physics is often restricted to small parts (or "boundary layers") inside the domain. For problems of this type Finite-Difference (FD) methods are superior to OC methods because a large number of grid points can be used and computations can still be effectively performed through tri-, penta-, and/or nano-diagonal matrix calculations.

By combining the low truncation error properties of OC methods and the ability of FD to locate grid points where needed, Carey and Finlayson [121] proposed a piecewise representation of the profiles. This method known as Orthogonal Collocation on Finite Elements

(OCFE) divides the domain into elements where the properties are expanded locally in terms of orthogonal polynomials. The OCFE method is very efficient when dealing with single boundary layers [121], [122] and one-dimensional propagation phenomena [123]. Although the OCFE methods possess several advantages, they have some undesirable features such as the drastic increase of the bandwidth with the consequence of high memory and computer time requirements for some two-dimensional problems. In particular, this method is clearly outclassed by FD methods when multiple two-dimensional boundary layers displace and deform in time. In such situations, a generalized mesh is usually the proper domain discretization. To discretize the governing equations on generalized meshes, FD schemes have proven simpler and exhibit several advantages for efficient programming.

4.7.2. Discretization of the Governing Equations

Once the physical domain has been discretized appropriately and a way has been chosen to change the grid points locations in response to the evolving solution, the governing equations can be approximated on the so constructed generalized mesh. The use of a generalized mesh makes approximation of the governing partial differential equations more complicated than it would be on an orthogonal cartesian or polar grid. Two different approaches have been followed to discretize the equations in a generalized coordinate system, each having its own merits [124].

The first approach is based on the transformation of the governing equations to the generalized curvilinear coordinates. Since the transformation involves additional terms ("cross derivatives") and variable coefficients ("metrics"), the resulting equations become more complicated. The additional terms appear because of the nonorthogonality of the curvilinear coordinates in the physical space. Depending on the rearrangement of the different terms after substitution of the transformation relations, the resulting transformed equations can be obtained in a conservative or nonconservative form. For more information, see [125].

The second approach is known as the Control Volume Formation (CVF). It takes into account the fact that the governing equations frequently represent the conservation of a transported quantity for an infinitesimally small control volume; thus, the conservation equations can easily be written for any arbitrary control volume of finite size. The approach consists of dividing the physical domain into smaller "cells" and rewritting the governing equations for these finite-size control volumes. In what follows, compressibility effects are neglected and the fluid and solid heat capacities are assumed to be constant.

The conservation law (the energy balance, for instance) applied to a control volume V is

$$\frac{\partial}{\partial t}\int_V \overline{\varrho C_p}\,T\mathrm{d}V = -\int_V \boldsymbol{\nabla}\cdot\left[\varrho_{\mathrm{f}} C_{p,\mathrm{f}}\mathbf{u}T - \boldsymbol{\lambda}_{\mathrm{e}}\cdot\boldsymbol{\nabla}T\right]\mathrm{d}V + \int_V (-\Delta H_r)\,R\,(C_1,\ldots,C_{\mathrm{i}},\ldots C_{\mathrm{s}},T)\,\mathrm{d}V \quad (4.40)$$

The convective and conductive transport terms can be rewritten by using the Gauss divergence theorem as:

$$-\int_V \boldsymbol{\nabla}\cdot\left[\varrho_{\mathrm{f}} C_p^{;\mathrm{f}}\mathbf{u}T - \lambda_{\mathrm{e}}\cdot\boldsymbol{\nabla}T\right]\mathrm{d}V = \oint_s \mathbf{n}\cdot\left[\varrho_{\mathrm{f}} C_{p,\mathrm{f}}\mathbf{u}T - \lambda_{\mathrm{e}}\cdot\boldsymbol{\nabla}T\right]\mathrm{d}S \quad (4.41)$$

where S is the surface enclosing the control volume V and $\mathbf{n}$ is the outward unit normal vector on S.

If the control volume moves and deforms in time, and a Lagrangian approach is followed, the integral of the time-dependent term becomes:

$$\frac{\partial}{\partial t}\int_V \overline{\varrho C_p}\,T\mathrm{d}V = \frac{\mathrm{d}}{\mathrm{d}t}\int_V \overline{\varrho C_p}\,T\mathrm{d}V - \oint_s \mathbf{n}\cdot\mathbf{u}_{\mathrm{s}}\left(\overline{\varrho C_p}T\right)\mathrm{d}S \quad (4.42)$$

where $\mathbf{u}_{\mathrm{s}}$ is the velocity at which the enclosing surface S moves and/or deforms.

Then, the integral form of the governing equations for the control volume V can be rewritten as:

Mass balance:

$$\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} \int_V C_\mathrm{i}\,\mathrm{d}V - \varepsilon \oint_\mathrm{s} \mathbf{n}\cdot\left[(\boldsymbol{u}-\mathbf{u}_\mathrm{s})\,C_\mathrm{i}\right]\mathrm{d}S$$
$$+\varepsilon \oint_\mathrm{s} \boldsymbol{n}\cdot\mathbf{D}_\mathrm{e}\cdot\boldsymbol{\nabla}C_\mathrm{i}\,\mathrm{d}S$$
$$-\int_V R\left(C_1,\ldots,C_\mathrm{i},\ldots,C_\mathrm{S},T\right)\mathrm{d}V \qquad (4.43)$$

Energy balance:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_V \overline{\varrho C_p}\,T\mathrm{d}V = -\oint_\mathrm{s} \mathbf{n}\cdot\left[(\mathbf{u}-\mathbf{u}_\mathrm{s})\,\varepsilon\,\varrho_\mathrm{f}C_p^{;\mathrm{f}}T\right]\mathrm{d}S$$
$$+\oint_\mathrm{s} \mathbf{n}\cdot\left[(\mathbf{u}_\mathrm{s}\,(1-\varepsilon)\,\varrho_\mathrm{s}C_{p,s}T)\,\mathrm{d}S - \oint_S \mathbf{n}\cdot\lambda_\mathrm{e}\cdot\boldsymbol{\nabla}T\mathrm{d}S\right]$$
$$+\int_V (-\Delta H_r)\,R\left(C_1,\ldots,C_\mathrm{i},\ldots,C_\mathrm{S},T\right)\mathrm{d}V \qquad (4.44)$$

The discretization of the governing equations now consists of approximating the integral form on the generalized mesh formed by quadrilateral cells. These last forms of the equations are extremely useful since continuity of fluxes through the boundaries of neighboring control volumes can be easily ensured. Conservation of the transported quantity, even for the discrete numerical approximation, will thus be maintained.

Further details of numerical approximation methods are discussed in detail in → Mathematics in Chemical Engineering, Chap. 2.

5. Optimization of Chemical Reactors

Optimization means calculation of the "best way" of operating a particular unit where a chemical reaction takes place. A number of parameters must be chosen during the design process of a chemical reactor; for instance, inlet temperature and concentration of reacting components, residence time, temperature of the cooling liquid, size of the catalyst particles, and dimension of the reactor tube. All these parameters may have a critical effect on a particular design. The optimization process usually includes economical factors in a direct or indirect way. In practical calculations there are many physical or technical constraints. For example, the composition of the inlet mixture of reacting gases must be outside of explosion limits; the maximum temperature in the reactor is given by catalyst deactivation (heterogeneous systems) or by mechanical properties of the reactor tube (homogeneous systems); maximum feed in the reactor may be limited by the capacity of downstream distillation columns, etc. The optimum design of the reactor must maximize the performance and consider the process constraints. The parameters entering into the optimization calculation can have different effects, some may have only a marginal effect while others can be critical for the optimization.

The selection of the criterion of optimality (also known as *objective function* or *performance index*) is crucial for design but somewhat arbitrary so the decision to select a certain criterion will always have a "fuzzy" character.

5.1. The Objective Function for Chemical Reactors

The objective function of a chemical reactor can have different forms. A list of functions as suggested by Edgar and Himmelblau [126] is given below:

1) Maximum conversion in a given equipment
2) Maximum production per batch
3) Minimum time of operation for a prescribed conversion
4) Maximum profit with respect to volume
5) Minimum consumption of energy
6) Minimum reactor volume with respect to certain concentrations
7) Maximum concentration by adjusting optimal temperature profile
8) Maximum selectivity for given feed
9) Minimum operating cost for a running chemical reactor

Typical independent variables are: pressure, inlet temperature and concentration, flow rate of the feed, temperature of the cooling system, and diameter of the catalyst pellets.

Frequently, the objective function has a complex character. It may be desirable, for example, to maximize profit and minimize pollution of the environment. Here two objective functions are introduced and the problem is called optimization of a vector objective function. To solve this

problem an efficient point is defined [127]. The system is said to be working at an efficient point when a particular scalar objective function cannot be optimized further without degrading the optima reached for the remaining functions.

Mathematically, the optimization problem can be formulated in the following way. A given objective function must be maximized:

$$F(x_1, x_2, \ldots, x_n) = 0$$

subject to the following relations:

$$g_i(x_1, x_2, \ldots, x_n) = 0 \quad \text{with} \quad i = 1, 2, \ldots, r < n$$

and constraints:

$$h_j(x_1, x_2, \ldots, x_n) \geq 0 \quad \text{with} \quad j = 1, 2, \ldots, m$$

Mathematically, the description of the problem is represented by algebraic equations (e.g., for the nonisothermal CSTR), ordinary differential equations of initial value type (for the nonisothermal packed-bed reactor), nonlinear differential equations of boundary value type (adiabatic packed-bed reactor with external heat exchanger), or parabolic partial differential equations (packed-bed tubular reactor). Frequently, in reaction engineering, these descriptions are highly nonlinear because of the complex kinetic expressions and exponential dependence on temperature of kinetic, adsorption, and equilibrium constants.

The mathematical techniques used to solve the optimization problems in reactor design most often reported in the literature are [128]:

1) Differential calculus
2) Nonlinear programming
3) Variational methods and maximum principle
4) Dynamic programming
5) Optimization via repeated simulations

An important aspect in optimum design of chemical reactors is the parametric sensitivity of the solution. The multiparametric problems are often so complex that direct optimization can be an unaffordable, time- consuming task. However, qualitative analysis of the problem could indicate that only few independent variables significantly affect the solution and that the optimum solution is rather insensitive towards some of the other variables. Very frequently, in the industrial environment, a quasi-optimization of the reactor problem is sought which only considers the important parameters. A quasi-optimal solution can be very close to the optimum one, and has the advantage that the operating parameters can be easily adjusted and safely controlled.

5.2. Elementary Optimization Problems

5.2.1. Optimization of a Batch System

A mass balance for the consecutive reaction:

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \xrightarrow{k_2} \mathrm{C}$$

can be easily integrated and the solution for the concentration of B, C_B is:

$$C_\mathrm{B} = C_\mathrm{A}^0 \frac{k_1}{k_2 - k_1} \left[\mathrm{e}^{-k_1 t} - \mathrm{e}^{-k_2 t} \right] \tag{5.1}$$

where C_A^0 is the initial concentration of A. The maximum concentration of B, C_B^*, can be calculated by differentiation and solving for:

$$\frac{\partial C_\mathrm{B}}{\partial t} = C_\mathrm{A}^0 \frac{k_1}{k_2 - k_1} \left[k_2 \mathrm{e}^{-k_2 t} - k_1 \mathrm{e}^{-k_1 t} \right] = 0 \tag{5.2}$$

i.e.,

$$t^* = \frac{1}{k_2 - k_1} \ln \left(\frac{k_2}{k_1} \right) \tag{5.3}$$

where $t*$ is the time for the batch operation resulting in maximum production of B. The maximum concentration of B achievable under $t*$ conditions can be calculated by substituting Equation (5.3) into Equation (5.1), i.e.,

$$C_\mathrm{B}^* = C_\mathrm{A}^0 \left(\frac{k_2}{k_1} \right)^{k_2/(k_1 - k_2)} \tag{5.4}$$

If the activation energies of both reactions are different, the maximum selectivity is also a function of the temperature T. The maximum concentration of B can then be calculated from the condition:

$$\frac{\partial C_\mathrm{B}}{\partial t} = \frac{\partial C_\mathrm{B}}{\partial T} = 0 \tag{5.5}$$

This problem cannot be solved analytically and Equation (5.5) must be solved numerically by a suitable method (e.g., Newton – Raphson) to obtain $t*$ and $T*$ (where $T*$ is the reaction temperature which results in a maximum production

of B). The maximum concentration C_B^* can then be calculated from Equation (5.1) after inserting $t*$ and $T*$. The procedure described above can be followed for any arbitrary sequence of first-order equations that can be integrated analytically. For a complex nonlinear reaction rate scheme, a simple search approach can be followed; for a given temperature the problem is integrated numerically as an initial value problem and $t*$ is determined for $\partial C_B/\partial t = 0$. Repeating the calculation for different values of T, a one-dimensional search for the optimum condition in T will result in the values of $t*$, $T*$, and C_B^*.

5.2.2. Optimization of a Continuous System (CSTR)

If the consecutive reactions described in Section 5.2.1 are carried out in a CSTR, in a similar way it can be found that:

$$t^* = \frac{1}{\sqrt{k_1 k_2}} \tag{5.6}$$

$$C_B^* = C_A^0 \frac{1}{\left(\sqrt{k_2/k_1}+1\right)^2} \tag{5.7}$$

5.2.3. Optimization of Complex Systems: Illustrative Example

Consider the following reaction scheme:

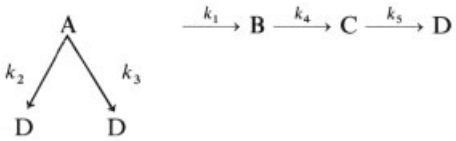

All reactions are first order with the rate constants k_i given in the form:

$$k_i = k_{0,i} \exp\left[-\frac{E_i}{R_u}\left(\frac{1}{T} - \frac{1}{658\,\mathrm{K}}\right)\right]$$

The preexponential factors and activation energies are given in Table 6.

Table 6. Values of preexponential factors and activation energies

Kinetic values	Reaction				
	1	2	3	4	5
$k_{0,i}$, s^{-1}	1.02	0.93	0.39	3.28	0.08
E_i, kJ/mol	67.00	58.70	62.80	41.90	62.80

The optimization problem consists of finding the optimal residence time and operation temperature to maximize the production of C, C_C^*. An analytical expression for C_C can be found as:

$$C_C = \frac{k_1 k_4}{k_A - k_4} \cdot \left[\frac{1 - e^{-(k_4 - k_5)t}}{k_4 - k_5} - \frac{1 - e^{-(k_A - k_5)t}}{k_A - k_5}\right] e^{-k_5 t} \tag{5.8}$$

where $k_A = k_1 + k_2 + k_3$. The problem can be solved by a search method via, for instance, the gradient method (→ Mathematics in Chemical Engineering, Chap. 10.). The results have been summarized in Table 7.

As can be seen from Table 7, the gradient method converges slowly and the maximum is rather flat in t and T. Evidently, $t = 0.40$ s, and $T = 773$ K can be chosen as reaction conditions. In this case only 0.014 is lost in selectivity; the reaction, however, can be carried out at a temperature that is 200 K lower than the optimum temperature. This represents a major energy saving. On the other hand, the residence time is five times higher than at optimum temperature which represents higher investment cost.

5.3. Optimization of Reactors by the Search Method

The optimization of a multiparametric reaction system can be accomplished by the search method. For more information, see [126].

The simplest gradient method is sufficient to illustrate the principle of the search technique. A new approximation is always selected in the gradient direction $\nabla F(\mathbf{x})$, i.e.,

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h^k \nabla F\left(\mathbf{x}^k\right) \tag{5.9}$$

where the value of h^k is selected in such a way that:

$$F\left(\mathbf{x}^{k+1}\right) < F\left(\mathbf{x}^k\right) \tag{5.10}$$

This method can be easily used for problems where the gradient of the function can be calculated analytically. Otherwise, the gradient must be calculated numerically by finite differences.

Table 7. Course of iterations using the gradient method

Iteration no.	T, K	t, s	C_C, kmol/m^3	T, K	t, s	C_C, kmol/m^3
0	1073	0.50	0.019	773	0.5	0.406
1	909	0.10	0.414	773	0.45	0.409
5	915	0.12	0.422	773	0.4	0.409
16	983	0.076	0.423	858	0.19	0.419
29	983	0.076	0.423	983	0.076	0.423
33	983	0.076	0.423	983	0.076	0.423

The rate of convergence of the gradient method is good if the initial point is far from the optimum; close to the optimum, however, the rate of convergence becomes very slow. The search by gradient method can be used easily for parameters with continuous values, complications arise for integer-value parameters (e.g., number of adiabatic reactor stages). The simplest way to overcome this problem is to perform a gradient search for a sequence of fixed integer value parameters and compare the results.

Illustrative Example [129]. Optimum operating conditions are to be found for a reactor in which nitrogen monoxide is oxidized to nitrogen dioxide. A dilute mixture of nitrogen monoxide (1.5 vol % NO) in air is fed into the reactor at a rate of 50 t NO/d. The following parameters should be optimized: catalyst type, feed composition, dew point of the inlet mixture, feed temperature, temperature control in the reactor, pressure, recirculation ratio, pressure drop, degree of conversion, mass velocity, diameter of the catalyst pellets, number of catalytic stages, and length and diameter of the catalyst bed.

A number of parameters listed here can be chosen a priori based on the information about the process.

Initial Concentration. The initial concentration is 1.5 vol % NO. This is an equilibrium concentration which results from synthesis of NO from N_2 and O_2 at ca. 2200 °C. The concentration of NO in the inlet gas can be increased by liquefaction and subsequent distillation. This modification would increase the processing costs.

Catalyst Type. Experimental results indicate that activated coke is a very good catalyst because it significantly increases the reaction rate. Silica gel is also a suitable catalyst, however, the reactor volume must then be 21 times larger. The homogeneous reactor (without catalyst) would require a volume 500 times larger. A disadvantage of activated coke is that there is the possibility of combustion in the presence of NO. This can be prevented by careful reactor temperature control, i.e., the adiabatic temperature increase in the reactor must be kept low.

Feed Temperature. The reaction rate is described by a Langmuir – Hinshelwood kinetic expression, thus the reaction rate is not very dependent on temperature. The rate of the surface reaction increases with the temperature; however, the surface concentration is then lowered because of desorption. The rate of the reverse reaction:

$$2\,NO \longrightarrow N_2 + O_2$$

can be neglected at temperatures below 200 °C. The exit temperature from the high-temperature reactor is 250 °C, the outlet gas is cooled by water to 20 – 30 °C. Cooling the inlet gas to the reactor below 20 °C would require new cooling equipment which would increase the investment costs.

Dew Point of Feed. Water is a catalyst poison for the oxidation of nitrogen monoxide both on silica gel and activated coke. For a reaction gas having a dew point 0 °C the amount of catalyst must be doubled compared with dry gas feed. For a dew point of − 30 °C the amount of catalyst required increases only by 10 %. Water can be removed from the inlet gas by drying it in an adsorption drying column packed with silica gel.

Temperature Control in the Reactor. The adiabatic temperature increase for 98 % conversion of 1.5 vol % NO in the feed is only 28 °C. Part of the reaction already occurs in the drying column (50 % conversion). The temperature increase involved (14 °C) does not promote undesirable catalyst oxidation. An adiabatic multistage reactor can thus be used for the reaction.

Recirculation. A simple calculation indicates that the reaction is completed after one pass so recirculation is not necessary.

Reactor Pressure. Doubling the pressure approximately doubles the reaction rate and halves the amount of catalyst required. However, since the mixture is strongly diluted, compression will be more expensive than the saving due to a smaller reaction volume. Therefore, the inlet pressure is chosen such that it only compensates for the pressure drop in the reactor.

Data for Calculation. The rate of oxidation of NO is described by Langmuir – Hinshelwood kinetics as:

$$R = \frac{p_{\mathrm{NO}}^2 p_{\mathrm{O_2}}}{a + b p_{\mathrm{NO}}^2 + c p_{\mathrm{NO_2}} + \omega p_{\mathrm{H_2O}}}$$

The governing equations are described by the one-phase model (see Section 4.1.3) and the pressure drop by the Ergun correlation (Section 4.5.2).

The *reactor costs* C_{R} are given by:

$$C_{\mathrm{R}} = 0.99\left[128.5 t_{\mathrm{sw}} D H_{\mathrm{t}} + 77.2 t_{\mathrm{dh}} D^2\right]^{0.648}$$

where t_{sw} is the thickness of the cylindrical side wall, D the internal diameter of the reactor, H_{t} the total length of the reactor, and t_{dh} the thickness of the dished reactor head. The reactor is designed to withstand 140 kPa overpressure.

The *ventilator costs* C_{V} are:

$$C_{\mathrm{V}} = 22\,900 \frac{\Delta P_r}{(\Delta P_r + P_{\mathrm{e}} - 0.579)^{0.915}}$$

where ΔP_r is the pressure drop across the reactor and P_{e} is the total pressure in the bulk stream at the reactor exit.

The *turbine costs* C_{T} are:

$$C_{\mathrm{T}} = 122 p_{\mathrm{H}}^{0.6}\left(\frac{\Delta P_r}{\Delta P_r + P_{\mathrm{e}} - 0.579}\right)$$

where:

$$p_{\mathrm{H}} = 22\,900\left[(\Delta P_r + P_{\mathrm{e}} + 0.421)^{0.283} - 1\right]$$

To estimate the *steam costs* (C_{S}) it is assumed that the turbine uses steam with 2.76 MPa pressure. The effectiveness of the turbine is 0.68; the turbine works 8000 h/a and steam is generated for \$ 1.10/1000 kg (1966). Then:

$$C_{\mathrm{S}} = 54 p_{\mathrm{H}}\left(\frac{\Delta P_r}{\Delta P_r + P_{\mathrm{e}} - 0.579}\right)$$

To estimate the *catalyst cost* (C_{C}), the life of the catalyst is assumed to be five years. Therefore, for \$ 0.2 per kilogram of catalyst the cost is

$$C_{\mathrm{C}} = 0.018 W$$

where W is the mass of the catalyst.

The loss C_{N} that is due to unconverted NO (\$ 2.40/100 kg NO) is

$$C_{\mathrm{N}} = 3.59 \times 10^5 (1 - \xi)$$

where ξ is the degree of NO conversion in the feed.

The *overall cost* is

$$Y = C_{\mathrm{R}} + C_{\mathrm{V}} + C_{\mathrm{T}} + C_{\mathrm{S}} + C_{\mathrm{C}} + C_{\mathrm{N}}$$

The optimization calculation performed here was based on repetitive calculations of the reactor using the Box – Wilson method [130] in which the base point of the first set of trials should be arbitrarily chosen. The range of the first set of trial points in the factorial design should be quite narrow. Preliminary calculations made by varying one factor at a time indicate certain ranges within which the optimum point probably lies. These ranges are used directly as the upper and lower levels in a two-level factorial experiment (see Table 8). The half-replicate experiments, the radial line trials, and the factor levels are given in Tables 9, 10, 11, respectively.

Table 8. Two-level factorial experimental design

Standarized variable	Natural variable	Factor level	
		− 1	+ 1
x_1	y	0.97	0.99
x_2	G [a]	200	400
x_3	D_{p} [b]	0.025	0.050
x_4	Z [c]	6	8

[a] G = mass velocity.
[b] D_{p} = diameter of catalyst particle.
[c] Z = number of catalyst layers.

The average value of trials 101 – 108 is 20 488. Since the value at the center of the response surface is much less than the average of the circumferential points, the existence of curvature in the explored region is obvious. The results of trials 109 – 112 along the radial line indicate the occurrence of a minimum in trial

110. The results of the optimum design are as follows:

Table 9. Half-replicate experiments

Trial	x_1	x_2	x_3	x_4	Y
101	−1	−1	−1	−1	20 692
102	+1	−1	−1	+1	17 081
103	−1	+1	−1	+1	21 019
104	+1	+1	−1	−1	22 862
105	−1	−1	+1	+1	21 147
106	+1	−1	+1	−1	18 453
107	−1	+1	+1	−1	20 636
108	+1	+1	+1	+1	22 018
Improved trial					
109	0	0	0	0	17 733

Table 10. Radial line trials

Trial	x_1	x_2	x_3	x_4	Y
110	+1/2	−1/2	−1/2	+1/2	16 509
111	+1/4	−1/4	−1/4	+1/4	17 077
112	+3/4	−3/4	−3/4	+3/4	17 173

Table 11. Factor levels

	−2	−1	0	+1	+2
x_1	0.98	0.9825	0.985	0.9875	0.99
x_2	200.0	225.0	250.0	275.0	300.0
x_3	0.025	0.028125	0.03125	0.034375	0.375

Catalyst	activated coke
Size of catalyst	≈ 0.01 m
Composition of gas entering bed	1.50 vol % NO
Number of layers	8
Thickness of each layer	≈ 0.80 m
Mass velocity	≈ 0.34 kg m^{-2} s^{-1}
Diameter of bed	≈ 4.0 m
Temperature of entering gas	293 K
Temperature of leaving gas	≈ 331.50 K
Pressure of entering gas	≈ 11.69 kPa
Pressure of leaving gas	≈ 11.0 kPa
Fraction conversion	0.986
Dew point, entering gas	243 K
Weight of catalyst	≈ 37 850 kg
Annual costs (1966)	
Reactor	$ 8230
Blower	$ 180
Turbine	$ 120
Steam	$ 1330
Catalyst	$ 1500
Nitrogen monoxide loss	$ 5210
Total	$ 16 570

5.4. Optimum Temperature Profile and Optimization of Multistage Reactors

For reversible exothermic catalytic reactions (e.g., water–gas shift synthesis, oxidation of SO_2 to SO_3, synthesis of methanol and ammonia), the optimum temperature profile plays an important role in optimum design.

Reaction rate isolines in the $\xi - T$ phase plane are presented in Figure 20. Three generic lines (a–c) representing three different reaction rate values (for an exothermic reversible reaction) in decreasing order are shown. The locus of optimum conditions (d) connects the points of maximum conversion for a given value of the reaction rate, the equilibrium line (e) represents the locus of reaction rate $R(C,T)=0$, and the adiabatic line (f) represents the history of the reaction rate for a reaction under adiabatic conditions.

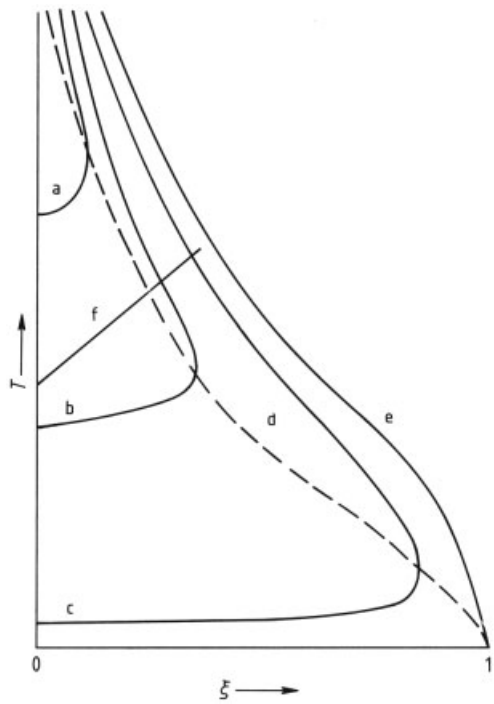

Figure 20. Reaction rate in exothermic reversible systems a)–c) Loci of constant reaction rate; d) Optimum temperature profile; e) Equilibrium curve; f) Adiabatic line

For an equilibrium reaction:

$$\nu_1 A + \nu_2 B \rightleftharpoons \nu_3 P + \nu_4 Q$$

the reaction rate is given by:

$$R = k_1 C_A^{\nu_1} C_B^{\nu_2} - k_2 C_P^{\nu_3} C_Q^{\nu_4} \tag{5.11}$$

and for the optimum temperature where the maximum rate can be achieved, the following relation holds:

$$\frac{\partial r}{\partial T} = C_A^{\nu_1} C_B^{\nu_2} \frac{\partial k_1}{\partial T} - C_P^{\nu_3} C_Q^{\nu_4} \frac{\partial k_2}{\partial T} = 0 \tag{5.12}$$

If k_1 and k_2 show an Arrhenius dependence on temperature then:

$$\frac{C_P^{\nu_3} C_Q^{\nu_4}}{C_A^{\nu_1} C_B^{\nu_2}} = \left(\frac{k_1 E_1}{k_2 E_2}\right)_{T^*} \tag{5.13}$$

where $T*$ is the optimum reaction temperature.

For equilibrium conditions the following expression holds:

$$\frac{C_P^{\nu_3} C_Q^{\nu_4}}{C_A^{\nu_1} C_B^{\nu_2}} = \left(\frac{k_1}{k_2}\right)_{T^e} \tag{5.14}$$

where T^e is the equilibrium temperature.

Combining Equations (5.13) and (5.14), gives:

$$\frac{T^*}{T^e} = \frac{1}{1 + \frac{R_n T^e}{E_2 - E_1} \ln\left(\frac{E_2}{E_1}\right)} \tag{5.15}$$

The mass of catalyst W required for processing a given molar flow rate F of reactants in an adiabatic packed bed can be calculated from:

$$\frac{W}{F} = Z = \int_0^y \frac{\mathrm{d}y}{R} \tag{5.16}$$

For the optimum temperature profile in the reactor the integral in Equation (5.16) is minimum, i.e.,

$$\frac{\partial Z}{\partial T} = \int_0^y \left(\frac{\partial r}{\partial T}\right) \frac{\mathrm{d}y}{r^2} = 0 \tag{5.17}$$

The optimum temperature profile is presented in Figure 20 (curve d).

For adiabatic tubular reactors the optimum temperature profile cannot be easily adjusted. Figure 20 reveals that optimum reaction rate is a decreasing function in the ξ vs. T diagram while the adiabatic line is an increasing function. To reconcile the different trends of optimum temperature profile and adiabatic line, heat must be withdrawn in the axial direction. The simplest configuration is cooling between adiabatic layers. This strategy of approximation the optimum temperature profile is shown in Figure 21.

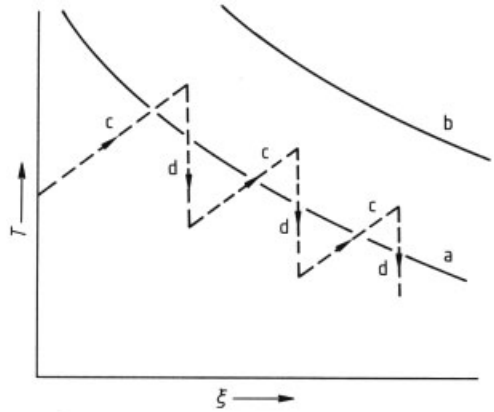

Figure 21. Approximation to the optimum temperature profile
a) Optimum profile; b) Equilibrium curve; c) Adiabatic operation; d) Intermediate cooling

5.5. Optimization of a Multibed Adiabatic Reactor with Heat Exchange Between Catalytic Stages

A schematic for optimization of a n stage adiabatic packed-bed reactor is shown in Figure 22. The mass of the catalyst W is a function of the inlet temperatures of the individual stages and the degrees of conversion obtained in the different stages:

$$W = W\left(T'_1, T'_2, \ldots, T'_n, \xi''_1, \xi''_2, \ldots, \xi''_n\right)$$

Assume that the degree of conversion in the nth stage is given. It is necessary to calculate $2n-1$ variables in such a way so that the mass of the catalyst is minimal. These parameters can be calculated by differential calculus, i.e.,

$$\left(\frac{\partial Z}{\partial T'_j}\right)_{\xi''_i, T'_i (i \neq j)} = 0 \quad \text{for} \quad i, j = 1, 2, \ldots, n \tag{5.18}$$

$$\left(\frac{\partial Z}{\partial \zeta_j}\right)_{T'_i, \xi''_i (i \neq j)} = 0 \quad \text{for} \quad i, j = 1, 2, \ldots, n \tag{5.19}$$

where ξ' and T' denote inlet conversion and temperature to a tray and ξ'' and T'' denote outlet conversion and temperature. After differentiation:

$$\frac{\partial Z}{\partial T'_j} = \int_{\zeta'_j}^{\zeta''_j} \frac{\partial}{\partial T}\left(\frac{1}{R}\right) \mathrm{d}\zeta = 0 \quad \text{for} \quad i,j = 1, 2, \ldots, n$$

$$\frac{\partial Z}{\partial \zeta'_j} = \frac{\partial}{\partial \zeta''_j} = \left[\int_{\zeta'_j}^{\zeta''_j} \frac{\mathrm{d}\zeta}{R\left(T'_j\right)} + \int_{\zeta'_{j+1}}^{\zeta''_{j+1}} \frac{\mathrm{d}\zeta}{R\left(T'_{j+1}\right)}\right] \tag{5.21}$$

$$= \frac{1}{R\left(T'_j, \zeta''_j\right)} - \frac{1}{R\left(T'_{j+1}, \zeta''_{j+1}\right)} = 0 \tag{5.22}$$

i.e.,

$$R\left(T'_j, \zeta''_j\right) = R\left(T'_{j+1}, \zeta''_{j+1}\right) \quad \text{for } i,j = 1, 2, \ldots, n \tag{5.23}$$

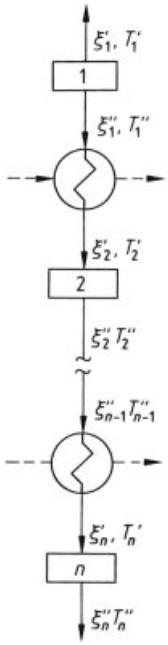

Figure 22. Schematic of a multibed adiabatic reactor with intermediate cooling
ξ' and T' denote inlet, ξ'' and T'' denote outlet conditions

There are $2n-1$ equations (Eqs. 5.20 and 5.23) to calculate the values of the optimum parameters. Numerically, Equation (5.20) can be solved easily by transforming it to a differential equation:

$$\frac{\partial Z}{\partial \zeta} = \frac{\partial}{\partial T}\left(\frac{1}{R}\right) = f(\zeta) \tag{5.24}$$

subject to the initial condition:

$$Z = 0 \quad \text{for} \quad \zeta = 0 \tag{5.25}$$

Equation (5.24) is integrated with Equation (5.25) until $Z = 0$. The calculated exit degree of conversion and exit temperature provide a value of reaction rate R. The inlet temperature in the second stage is calculated from Equation (5.23). The process is repeated for the remaining $n-1$ stages. In a similar way adiabatic reactors can be optimized with "cold shot", i.e., cooling between stages with fresh gas [131].

6. References

1. R. Shinnar, "Chemical Reaction Modeling – The Desirable and the Achievable," in: *Chemical Reaction Engineering – Houston,* ACS, Washington, D.C., 1978, p. 1.
2. K. R. Westerterp, W. P. M. van Swaaij, A. A. C. M. Beenackers: *Chemical Reactor Desgin and Operation,* J. Wiley & Sons, New York 1982.
3. R. B. Bird, W. E. Stewart, E. N. Lightfoot: *Transport Phenomena,* J. Wiley & Sons, New York 1960.
4. H. F. Rase: *Chemical Reactor Design for Process Plants,* J. Wiley & Sons, New York 1977.
5. V. V. Aleksandrov, A. A. Davydenko, Y. A. Kovalenko, N. P. Poddudnyi: "Influence of Two-Dimensionality of the Front with Heat Losses on the Limits of Stationary Gasless Combustion," *Fiz. Goreniya Vzryva* **23** (1987) 2 – 14.
6. J. Puszynski, J. Degreve, V. Hlavacek, "Modeling of Exothermic Non-Catalytic Reactions," *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 1424 – 1434.
7. F. A. Holland, F. S. Chapman: *Liquid Mixing and Processing in Stirred Tanks,* Reinhold, New York 1966.
8. V. W. Uhl, J. B. Gray (eds.): *Mixing Theory and Practice,* vols. I and II, Academic Press, New York 1966.
9. K. R. Westerterp, W. P. M. Van Swaaij, A. A. C. M. Beenackers: *Chemical Reactor Design and Operation,* J. Wiley & Sons, New York 1984.

10. G. Astarita: *Mass Transfer with Chemical Reaction,* Elsevier, Amsterdam 1967.
11. R. H. : *Perry's Engineering Handbook,* 6th ed. McGraw Hill, New York 1984.
12. C. van Heerden: "Autothermic Processes; Properties and Reactor Design", *Ind. Eng. Chem.* **45** (1953) 1242.
13. C. van Heerden: "The Character of the Stationary State of Exothermic Processes," *Chem. Eng. Sci.* **8** (1958) 133.
14. V. Hlavacek, M. Kubicek, J. Jelinek: "Modeling of Chemical Reactors – XVIII, Stability and Oscillatory Behavior of the CSTR," *Chem. Eng. Sci.* **25** (1970) 1441.
15. R. Aris, N. R. Amundson: "An Analysis of Chemical Reactor Stability and Control," *Chem. Eng. Sci.* **7** (1958) 121.
16. D. D. Perlmutter: *Stability of Chemical Reactors,* Prentice Hall, Englewood Cliffs, NJ, 1972.
17. R. A. Schmitz: "Multiplicity, Stability and Sensitivity of States in Chemically Reacting Systems," *Chem. React. Eng. Rev.* **148** (1975) 156.
18. A. H. Heemskerk, W. R. Dammers, J. M. H. Fortuin: "Limit Cycles Measured in a Liquid-Phase Reaction System," *Chem. Eng. Sci.* **35** (1980) 439.
19. O. Levenspiel: *Chemical Reaction Engineering,* 2nd ed., J. Wiley & Sons, New York 1972.
20. J. M. Eldridge, E. L. Piret: "Continuous Flow Stirred-Tank Reactor Systems," *Chem. Eng. Prog.* **46** (1950) 290.
21. N. H. Chen: *Process Reactor Design,* Allyn and Bacon, Newton, Mass. 1983.
22. G. F. Froment: "Packed Bed Reactors. Steady State Conditions," Review A 5 in *Chemical Reactor Engineering,* Proceedings of the Fifth European/Second International Symposium on Chemical Reaction Engineering, Amsterdam, Elsevier, Amsterdam 1972.
23. V. Hlavacek: "Aspects in Design of Packed Catalytic Reactors," *Ind. Eng. Chem.* **62** (1970) no. 7, 8 – 26.
24. H. Hofmann: "Probleme bei der mathematischen Modellierung von Schüttschicht-Reaktoren," *Chem.-Ing-Tech.* **46** (1974) no. 6, 236 – 242.
25. V. Hlavacek, J. Votruba: "Steady-State Operation of Fixed-Bed Reactors and Monolithic Structures," in L. Lapidus, N. R. Amundson (eds.): *Chemical Reactor Theory. A Review,* Chap. 6, Prentice Hall, Englewood Cliffs, N.J., 1977.
26. S. S. E. Elnashaie, D. L. Cresswell: "Dynamic Behaviour and Stability of Non-Porous Catalyst Particles," *Chem. Eng. Sci.* **28** (1973) 1387 – 1399.
27. G. F. Froment, K. B. Bischoff: *Chemical Reactor Analysis and Design,* J. Wiley & Sons, New York 1979.
28. D. Ramkrishna, P. Arce: "Can Pseudohomogeneous Reactor Models be Valid?," *Chem. Eng. Sci.* **44** (1989) 1949 – 1966.
29. P. Arce, D. Ramkrishna: "Pattern Formation in Catalytic Systems: The Role of Fluid Mixing," *AIChE J.* **37** (1991) no. 1, 98 – 110.
30. S. Yagi, D. Kunii, N. Wakao: "Studies on Axial Effective Thermal Conductivities in Packed Beds," *AIChE J.* **6** (1960) no. 4, 543 – 546.
31. A. Jutan et. al.: "Multivariable Computer Control of a Butane Hydrogenolysis Reactor: Part I. State Space Reactor Modeling," *AIChE J.* **23** (1977) no. 5, 732 – 742.
32. D. Vortmeyer, K. J. Dietrich, K. O. Ring: "Comparison of One- and Two-Phase Model Predictions for Adiabatic Packed-Bed Chemical Reactors," *Adv. Chem. Ser.* **133** (1974) 588 – 599.
33. D. Vortmeyer, R. J. Schaefer: "Equivalence of One- and Two-Phase Models for Heat Transfer Processes in Packed Beds: One Dimensional Theory," *Chem. Eng. Sci.* **29** (1974) 485 – 491.
34. J. J. Carberry, D. White: "On the Role of Transport Phenomena in Catalytic Reactor Behavior," *Ind. Eng. Chem.* **61** (1969) 27 – 31.
35. S. Feyo de Azevedo, M. A. Romero-Ogawa, A. P. Wardle: "Modelling of Tubular Fixed-Bed Catalytic Reactors: A Brief Review," *Chem. Eng. Res. Des.* **68 (A)** (1990) 483 – 502.
36. H. M. Hulburt: "Chemical Processes in Continuous-Flow Systems. Reaction Kinetics," *Ind. Eng. Chem.* **36** (1944) no. 11, 1012 – 1017.
37. P. V. Danckwerts: "Continuous Flow Systems. Distribution of Residence Times," *Chem. Eng. Sci.* **2** (1953) no. 1, 1 – 18.
38. J. F. Wehner, R. H. Wilhelm: "Boundary Conditions of Flow Reactor," *Chem. Eng. Sci.* **6** (1956) 89 – 93.
39. L. T. Fan, Y. K. Ahn: "Critical Evaluation of Boundary Conditions for Tubular Flow Reactors," *Ind. Eng. Chem. Process Des. Dev.* **3** (1962) no. 1, 190 – 195.

40. V. Hlavacek, M. Marek: "Axialer Stoff- und Wärmetransport im adiabatischen Rohrreaktor – II. Numerische Untersuchung – Ablauf einer einfachen Reaktion bzw. einer Folgereaktion," *Chem. Eng. Sci.* **21** (1966) 501 – 514.
41. V. Hlavacek, H. Hofmann: "Modeling of Chemical Reactors – XVII. Steady State Axial Heat and Mass Transfer in Tubular Reactors. Numerical Investigations of Multiplicity," *Chem. Eng. Sci.* **25** (1970) 187 – 199.
42. V. Hlavacek et. al.: "Boundary Conditions for a Tubular Nonisothermal Nonadiabatic Packed Bed Reactor with Wall Heat Transfer in the Fore and Aft Sections," *Chem. Eng. Commun.* **3** (1979) 451 –460.
43. J. Puszynski: "Experimental and Numerical Studies of Catalytic Nonadiabatic Reactors" Ph. D. Thesis, Institute for Chemical Technology, Prague, Czechoslovakia 1980.
44. K. B. Bischoff: "A Note on Boundary Conditions for Flow Reactors," *Chem. Eng. Sci.* **16** (1961) no. 1, 131 – 133.
45. L. C. Young, B. A. Finlayson: "Axial Dispersion in Nonisothermal Packed Bed Chemical Reactors," *Ind. Eng. Chem. Fundam.* **12** (1973) no. 4, 412 – 422.
46. A. Rasmuson: "Exact Solution of a Model for Diffusion in Particles and Longitudinal Dispersion in Packed Beds: Numerical Evaluation," *AIChE J.* **31** (1985) 518, 519.
47. A. Rasmuson: " Exact Solution of Some Models for the Dynamics of Fixed Beds using Danckwerts' Inlet Condition," *Chem. Eng. Sci.* **41** (1986) no. 3, 599, 600.
48. R. A. Novy, H. T. Davis, L. E. Scriven: "Upstream and Downstream Boundary Conditions for Continuous-Flow Systems," *Chem. Eng. Sci.* **45** (1990) no. 6, 1515 – 1524.
49. R. A. Novy, H. T. Davis, L. E. Scriven: "A Comparison of Synthetic Boundary Conditions for Continuous-Flow Systems," *Chem. Eng. Sci.* **46** (1991) 57 – 68.
50. R. A. Novy, H. T. Davis, L. E. Scriven: "A Comparison of Initial Conditions for Continuous-Flow Systems," *Chem. Eng. Sci.* **46** (1991) no. 7, 1725 –1737.
51. A. G. Dixon, D. L. Cresswell: "Theoretical Prediction of Effective Transfer Parameters in Packed Beds," *AIChE J.* **25** (1979) no. 4, 663 – 676.
52. D. Vortmeyer, R. Berninger in A. G. Dixon, D. L. Cresswell (eds.): "Comments on the Paper, Theoretical Prediction of Effective Heat Transfer Parameters in Packed Beds," *AIChE J.* **28** (1982) no. 3, 508 – 510.
53. W. Odendaal, W. Gobie, J. J. Carberry: "Thermal Parameter Sensitivity in the Simulation of the Nonisothermal, Nonadiabatic Fixed Bed Reactor – The Two-Dimensional Heterogeneous Model," *Chem. Eng. Commun.* **58** (1987) 37 – 62.
54. L. C. Windes, M. J. Schwedock, W. H. Ray: "Steady State and Dynamic Modeling of a Packed Bed Reactor for the Partial Oxidation of Methanol to Formaldehyde. I. Model Development," *Chem. Eng. Commun.* **78** (1989) 1 – 43.
55. L. H. S. Roblee, R. M. Baird, J. W. Tierney: "Radial Porosity Variations in Packed Beds," *AIChE J.* **4** (1958) 460 – 464.
56. R. F. Benenati, C. B. Brosilow: "Void Fraction Distribution in Beds of Spheres," *AIChE J.* **8** (1962) no. 3, 359 – 361.
57. V. M. H. Govindarao, G. F. Froment: "Voidage Profiles in Packed Beds of Spheres," *Chem. Eng. Sci.* **41** (1986) no. 3, 533 – 539.
58. B. C. Chandrasekhara, D. Vortmeyer: "Flow Model for Velocity Distribution in Fixed Beds Under Isothermal Conditions, " *Wärme Stoffübertrag.* **12** (1979) 105 – 111.
59. D. Vortmeyer, J. Schuster: "Evaluation of Steady Flow Profiles in Rectangular and Circular Packed Beds by a Variational Method," *Chem. Eng. Sci.* **38** (1983) no. 10, 1691 – 1699.
60. O. Kalthoff, D. Vortmeyer: "Ignition/Extinction Phenomena in a Wall Cooled Fixed Bed Reactor. Experiments and Model Calculations Including Radial Porosity and Velocity Distributions," *Chem. Eng. Sci.* **35** (1980) 1637 – 1643.
61. D. Vortmeyer, R. P. Winter: "Improvements in Reactor Analysis Incorporating Porosity and Velocity Profiles," *Ger. Chem. Eng. (Engl. Transl.)* **7** (1984) 19 – 25.
62. J. E. Gatica et. al.: "Steady and Non-Steady State Modeling of Tubular Fixed Bed Reactors," *Chem. Eng. Commun.* **78** (1989) 73 – 96.
63. H. Martin: "Low Péclet Number Particle-to-Fluid Heat and Mass Transfer in Packed Beds," *Chem. Eng. Sci.* **33** (1978) 913 – 919.
64. C. L. de Ligny: "Coupling between Diffusion and Convection in Radial Dispersion of Matter by Fluid Flow through Packed Beds," *Chem. Eng. Sci.* **25** (1970) 1177 – 1181.

65. J. J. Carberry: *Chemical and Catalytic Reaction Engineering,* McGraw-Hill, New York 1976.
66. D. Vortmeyer, R. Berninger: "Comparison of Dispersion Models for Calculation of Temperature Profiles in Packed Beds," *Ger. Chem. Eng. (Engl. Transl.)* **6** (1983) 9 – 14.
67. P. Zehner, E. P. Schlunder: "Wärmeleitfähigkeit von Schüttungen bei mäßigen Temperaturen," *Chem. Ing. Tech.* **42** (1970) 933 – 940.
68. A. G. Dixon: "Thermal Resistance Models of Packed-Bed Effective Heat Transfer Parameters," *AIChE J.* **31** (1985) no. 5, 826 – 834.
69. A. G. Dixon: "Wall and Particle-Shape Effects on Heat Transfer in Packed Beds," *Chem. Eng. Commun.* **71** (1988) 217 – 237.
70. D. L. Cresswell, A. G. Dixon: in [52]511 – 513.
71. A. G. Dixon, D. L. Cresswell: "Effective Heat Transfer Parameters for Transient PackedBed Models, " *AIChE J.* **32** (1986) no. 5, 809 –819.
72. D. Vortmeyer: "Radiation in Packed Solids," *Ger. Chem. Eng. (Engl. Transl.)* **3** (1980) 124 – 138.
73. C. L. Tien: "Thermal Radiation in Packed and Fluidized Beds," *J. Heat Transfer* **110** (1988) 1230 –1242.
74. A. G. Dixon, M. A. DiConstanzo, B. A. Soucy: "Fluid-Phase Radial Transport in Packed Beds of Low Tube-to-Particle Diameter Ratio," *Int. J. Heat Mass Transfer* **27** (1984) no. 10, 1701 – 1713.
75. E. Tsotsas, E.-U. Schlünder: "Heat Transfer in Packed Beds with Fluid Flow: Remarks on the Meaning and the Calculation of a Heat Transfer Coefficient at the Wall," *Chem. Eng. Sci.* **45** (1989) no. 4, 819 – 837.
76. R. E. Chao, R. A. Caban, M. M. Irizarry: "Wall Heat Transfer to Chemical Reactors," *Can. J. Chem. Eng.* **51** (1973) 67 – 70.
77. R. J. Wijngaarden, K. R. Westerterp: "Do the Effective Heat Conductivity and the Heat Transfer Coefficient at the Wall Inside a Packed Bed Depend on a Chemical Reaction? Weakness and Applicability of Current Models," *Chem. Eng. Sci.* **44** (1989) no. 8, 1653 – 1663.
78. D. J. Gunn, D. Vortmeyer: "The Reaction-Independence of Mixing Coefficients in Fixed-Bed Reactors," *AIChE J.* **36** (1990) no. 9, 1449 – 1451.
79. J. Beek, E. Singer: "A Procedure for Scaling-up a Catalytic Reactor," *Chem. Eng. Prog.* **47** (1951) 534 – 540.
80. J. E. Crider, A. S. Foss: "Effective Wall Heat Transfer Coefficients and Thermal Resistances in Mathematical Models of Packed Beds," *AIChE J.* **11** (1965) no. 6, 1012 – 1018.
81. V. Hlavacek, M. Marek, T. M. John: "Tubular Non-Isothermal Nonadiabatic Packed Bed Reactor. An Analysis of the One-Dimensional Case," *Collect. Czech. Chem. Commun.* **34** (1969) 3664 –3675.
82. G. F. Froment: "Fixed Bed Catalytic Reactors. Current Design Status," *Ind. Eng. Chem.* **59** (1967) no. 2, 18 – 27.
83. O. Bilous, N. R. Amundson: "Chemical Reactor Stability and Sensitivity: II. Effects of Parameters on Sensitivity of Empty Tubular Reactors," *AIChE J.* **2** (1956) no. 1, 117 – 126.
84. G. F. Froment: "Progress in the Fundamental Design of Fixed Bed Reactors," in L. K. Doraiswamy, R. A. Mashelkar (eds.): *Frontiers in Chemical Reaction Engineering 1,* Wiley Eastern, New Delhi 1984, p. 12.
85. P. L. Chambré: "On the Characteristics of a Nonisothermal Chemical Reactor," *Chem. Eng. Sci.* **5** (1956) 209 – 216.
86. C. H. Barkelew: "Stability of Chemical Reactors," *Chem. Eng. Prog. Symp. Ser.* **55** (1959) no. 25, 37 –46.
87. R. J. van Welsenaere, G. F. Froment: "Parametric Sensitivity and Runaway in Fixed Bed Catalytic Reactors," *Chem. Eng. Sci.* **25** (1970) 1503 – 1516.
88. C. McGreavy, C. I. Adderley: "Parametric Sensitivity and Temperature Runaway in Heterogeneous Fixed Bed Reactors," *Adv. Chem. Ser.* **133** (1974) 519 – 531.
89. R. A. Rajadhyaksha, K. Vasudeva, L. K. Doraiswamy: "Parametric Sensitivity in Fixed Bed Reactors," *Chem. Eng. Sci.* **30** (1975) 1399 – 1408.
90. A. Oroskar, S. A. Stern: " Stability in Chemical Reactors," *AIChE J.* **25** (1979) no. 5, 903 – 905.
91. M. Morbidelli, A. Varma: "Parametric Sensitivity and Runaway in Tubular Reactors," *AIChE J.* **28** (1982) no. 5, 705 – 713.
92. T. F. Degnan, J. Wei: "The Cocurrent Reactor Heat Exchanger: I. Theory," *AIChE J.* **25** (1979) no. 2, 338 – 344.
93. L. M. Akella, H. H. Lee: "A Design Approach Based on Phase Plane Analysis: Countercurrent Reactor/Heat Exchanger with Parametric Sensitivity," *AIChE J.* **29** (1983) no. 1, 87 – 94.

94. A. Soria Lopez, H. I. DeLasa, J. A. Porras: "Parametric Sensitivity of a Fixed Bed Catalytic Reactor: Cooling Fluid Flow Influence," *Chem. Eng. Sci.* **36** (1981) 285 – 291.
95. H. I. DeLasa: "Application of the Pseudoadiabatic Operation to Catalytic Fixed Bed Reactors," *Can. J. Chem. Eng.* **61** (1983) 710 – 718.
96. G. P. Henning, G. A. Pérez: "Parametric Sensitivity in Fixed-Bed Catalytic Reactors," *Chem. Eng. Sci.* **41** (1986) no. 1, 83 – 88.
97. G. P. Henning, G. A. Pérez: "A Generalization of Two Intrinsic Criteria for Runaway in Fixed Bed Catalytic Reactors: Part I. Development of the Criteria," *Chem. Eng. Commun.* **59** (1987) 107 – 125.
98. E. Bauman et al.: "Parametric Sensitivity in Tubular Reactors with Co-Current External Cooling," *Chem. Eng. Sci.* **45** (1990) no. 5, 1301 – 1307.
99. G. P. Henning, G. A. Pérez: "A Generalization of Two Intrinsic Criteria for Runaway in Fixed Bed Catalytic Reactors: Part II. Criteria Performance Analysis of Co-Current vs. Counter-Current Operation," *Chem. Eng. Commun.* **59** (1987) 127 – 136.
100. D. O. Borio, J. E. Gatica, J. A. Porras: "Wall-Cooled Fixed-Bed Reactors: Parametric Sensitivity as a Design Criterion," *AIChE J.* **35** (1989) no. 2, 287 – 292.
101. D. O. Borio et. al. "Cocurrently-Cooled Fixed-Bed Reactors: A Simple Approach to Optimal Cooling Design," *AIChE J.* **35** (1989) no. 11, 1899 – 1902.
102. M. Morbidelli, A. Varma: "Parametric Sensitivity and Runaway in Fixed-Bed Catalytic Reactors," *Chem. Eng. Sci.* **41** (1986) 1063 – 1071.
103. M. Morbidelli, A. Varma: "On Parametric Sensitivity and Runaway Criteria of Pseudohomogeneous Tubular Reactors," *Chem. Eng. Sci.* **40** (1985) no. 11, 2165 – 2168.
104. M. Morbidelli, A. Varma: "A Generalized Criterion for Parametric Sensitvity: Application to Thermal Explosion Theory," *Chem. Eng. Sci.* **43** (1988) 91 – 102.
105. M. Morbidelli, A. Varma: "A Generalized Criterion for Parametric Sensitivity: Application to a Pseudohomogeneous Tubular Reactor with Consecutive or Parallel Reactions," *Chem. Eng. Sci.* **44** (1989) no. 8, 1675 – 1696.
106. V. Balakotahia: "Simple Runaway Criteria for Cooled Reactors," *AIChE J.* **35** (1989) no. 6, 1039 –1043.
107. H. C. Brinkman: "A Calculation of the Viscous Force Exerted by a Flowing Fluid on a Dense Swarm of Particles," *Appl. Sci. Res. Lect. A:* **1** (1947) 27 – 34.
108. D. A. Nield: "The Boundary Correction for the Rayleigh-Darcy Problem: Limitations of the Brinkman Equation," *J. Fluid Mech.* **128** (1983) 37 – 46.
109. K. Vafai, C. L. Tien: "Boundary and Inertia Effects on Flow and Heat Transfer in Porous Media," *Int. J. Heat Mass Transfer* **24** (1981) 195 – 203.
110. D. D. Gray, A. Giorgini: "The Validity of the Boussinesq Approximation for Liquid and Gases," *Int. J. Heat Mass Transfer* **19** (1976) 545 – 551.
111. D. D. Joseph: *Stability of Fluid Motions,* vol. 2, chap. X and XII, Springer Verlag, Berlin 1976.
112. D. Drew, L. Segel: "Averaged Equations for Two-Phase Flow," *Stud. Appl. Math.* **50** (1971) no. 3, 205 – 231.
113. J. L. Beck: "Convection in a Box of Porous Material; Saturated with Fluid," *Phys. Fluids* **15** (1972) no. 8, 1377 – 1383.
114. Ph. Forchheimer: "Wasserbewegung durch Boden," *VDJZ* **45** (1901) no. 50, 1781 – 1788.
115. M. Choudhary, M. Propster, J. Szekely: "On the Importance of the Inertial Terms in the Modeling of Flow Maldistribution in Packed Beds," *AIChE J.* **22** (1976) no. 3, 600 – 603.
116. M. Choudhary, J. Szekely, S. W. Weller: "The Effect of Flow Maldistribution on Conversion in a Catalytic Packed-Bed Reactor: Part I. Analysis," *AIChE J.* **22** (1976) no. 6, 1021 – 1027.
117. S. Ergun: "Fluid Flow through Packed Columns," *Chem. Eng. Prog.* **48** (1952) 89 – 94.
118. B. A. Boley, J. H. Weiner: *Theory of Thermal Stresses,* J. Wiley & Sons, New York 1960.
119. B. A. Boley, I. S. Tollins: "Transient Coupled Thermoelastic Boundary Value Problems in the Half Space," *J. Appl. Mech.* **29** (1962) no. 4, 637 – 646.
120. J. V. Villadsen, W. E. Stewart: "Solution of Boundary-Value Problems by Orthogonal Collocation," *Chem. Eng. Sci.* **22** (1967) 1483 – 1495.
121. C. G. Carey, B. A. Finlayson: "Orthogonal Collocation on Finite Elements," *Chem. Eng. Sci.* **30** (1975) 587 – 596.

122. K. F. Jensen, W. H. Ray: "The Bifurcation Behavior of Tubular Reactors," *Chem. Eng. Sci.* **37** (1982) 199 – 222.
123. J. E. Gatica, J. Puszynski, V. Hlavacek: "Reaction Front Propagation in Nonadiabatic Exothermic Reaction Flow Systems," *AIChE J.* **33** (1987) no. 5, 819 – 833.
124. D. A. Anderson, J. C. Tannehill, R. H. Pletcher: *Computational Fluid Mechanics and Heat Transfer,* McGraw-Hill, New York 1984.
125. R. G. Hindman: "Generalized Coordinate Forms of Governing Fluid Equations and Associated Geometrically Induced Errors," *AIAA J.* **16** (1982) 1359 – 1367.
126. T. F. Edgar, D. M. Himmelblau: *Optimization of Chemical Processes,* McGraw-Hill, New York 1988.
127. J. H. Seinfeld, W. L. McBride: "Optimization with Multiple Performance Criteria," *Ind. Eng. Chem. Process Des. Dev.* **9** (1970) 53.
128. D. J. Wilde: *Optimum Seeking Methods,* Prentice Hall, Englewood Cliffs, NJ, 1964.
129. C. Chu, O. A. Hougen: "Optimum Design of a Catalytic Nitric Oxide Reactor," *Chem. Eng. Prog.* **57** (1961) no. 6, 51.
130. H. H. Rosenbrock, C. Storey: *Computational Techniques for Chemical Engineers,* Pergamon Press, New York 1966.
131. J. Caha, V. Hlavacek, M. Kubicek: "Study of the Optimization of Chemical Engineering Equipment. Numerical Solution of the Optimization of an Adiabatic Reactor," *Int. Chem. Eng.* **13** (1973) 466.

Stirred-Tank and Loop Reactors

For other industrial reactors and their applications, see → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

John C. Middleton, ICI Chemicals & Polymers Ltd., Runcorn, United Kingdom

Keith J. Carpenter, ICI Fine Chemicals, Manufacturing Organisation, Manchester, United Kingdom

1. Tanks 1
1.1. Description 1
1.2. Operation of Stirred-Tank Reactors 4
1.3. Examples 5
1.3.1. Multiproduct Batch Plant 5
1.3.2. Continuous-Campaign Stirred-Tank Reactor 6
1.3.3. Cascade of Continuous Stirred-Tank Reactors 6
2. Loop Reactors 7
2.1. Description 7
2.2. Fields of Application 8
2.2.1. Liquid-Phase Reactions 8
2.2.2. Two-Phase Reactions 9
3. Practical Limitations of Stirred-Tank and Loop Reactors 10
4. Design Methods 11
4.1. Experiments 11
4.2. Scale-up 11
4.3. Residence-Time Distribution 12
4.3.1. Relevance 12
4.3.2. Stirred Tanks 12
4.3.3. Loop Reactors 13
5. Reactor Modeling 13
5.1. Fitted Models 13
5.2. Computational Fluid Dynamics (CFD) 14
6. References 15

Symbols used in this article are defined in → Principles of Chemical Reaction Engineering and → Model Reactors and Their Design Equations.

1. Tanks

1.1. Description

Most stirred tanks (stirred vessels) used as chemical reactors are cylindrical with a vertical axis, and are equipped with a centrally positioned vertical shaft upon which one or more stirrers ("impellers") are mounted. This article will concentrate on these reactors, although variants such as horizontal-axis tanks (for use with high-viscosity fluids) or off-center shafts are also encountered. The ends are normally dished (as in Fig. 1), but flat ends can be used when the reaction is performed near atmospheric pressure. These tanks must be designed according to appropriate national standards.

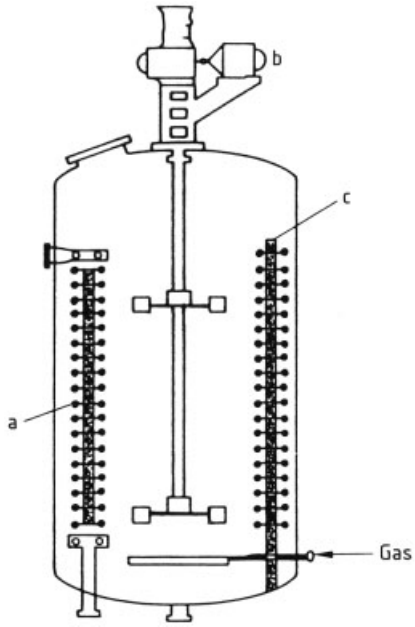

Figure 1. Baffled stirred tank (batch)
a) Heating or cooling coil; b) Motor drive; c) Baffle

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_167

The volume of the tank is normally determined by the residence time required, but the aspect ratio of the tank (height/diameter, *H/D*) can be selected. Bulk mixing is favored by $H/D = 1$; however, if heat must be transferred to a jacket, the *H/D* ratio must be large to maximize the exchange surface area per unit of reacting volume, so a compromise is often necessary.

Whenever possible, *wall baffles* are installed inside the vessel (Fig. 1). These baffles prevent the gross rotation ("swirl") of the contents with the stirrer, and produce axial bulk motion of the contents to achieve top-to-bottom mixing and good particle suspension. They also produce a high relative velocity between the stirrer blades and the fluid, as required for dispersion of an immiscible phase. Wall baffles are sometimes avoided in systems that are prone to fouling (polymerizers, for example) because of the risk of an uncontrolled buildup of solid, or in processes in which gas or a floating material is to be entrained in the liquid. In this case the swirling surface vortex that develops in unbaffled stirred vessels (Fig. 2) is advantageous (although it can be unstable).

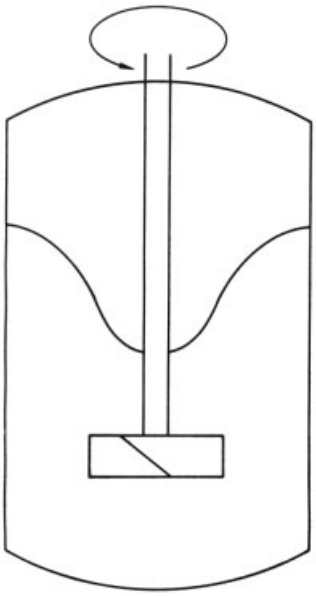

Figure 2. Unbaffled stirred tank (batch)

Jackets are often fitted around the vessel to provide heat transfer. Heat-transfer jackets are, however, insufficient for many exothermic reactions in all but the smallest vessels; in addition, a coolant flow through immersed tubes inside the tank, and even through hollow baffles, is commonly used to improve heat transfer.

Stirrers are used to enhance the contacting, mixing, mass, and heat-transfer processes relevant to the desired reactions. Especially in baffled tanks, the most appropriate type of stirrer must be selected, depending on the most important duty it has to perform; this task is not always obvious in complex multiphase reaction mixtures. The duties might be:

Bulk: Overall mixing (macromixing, i.e., "blending", "homogenization"); suspension of solid particles; distribution of bubbles or drops

Local: Heat transfer to heat-exchange surfaces; dispersion of immiscible liquid or gaseous reactants; control of size of product drops, bubbles, or particles; local mixing (micromixing) of liquid reactants

A wide variety of stirrers is available: some common examples are shown in Figures 3 and 4. Selection is based initially on the viscosity of the reaction mixture. The stirrers shown in Figure 3 are suited best to viscosities below ca. 0.5 Pa · s. Then, the type of duty and the number of phases present in the mixture are considered:

1) Propellers and their modern equivalent "hydrofoil" types (for low-viscosity liquid blending or particle suspension).
2) Intermigs and helical ribbons provide axial flow (efficient for bulk mixing and particle suspension).
3) Flat-blade turbines and disk turbines provide radial flow with high shear for dispersion of immiscible phases.
4) Pitched-blade turbines provide mixed axial and radial flow for multipurpose applications.
5) Anchors (which are used without baffles, at close clearance with the vessel wall) provide only enhancement of heat transfer between very viscous liquids and a jacket with very little mixing.

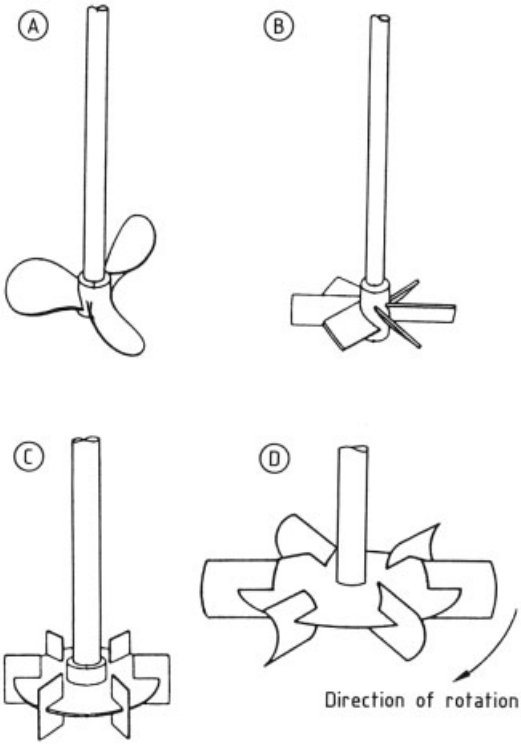

Figure 3. Agitators for low-viscosity liquids ($< 0.5\,\mathrm{Pa \cdot s}$)
A) Propeller; B) Pitched – blade turbine; C) Turbine; D) Curved – blade turbine

If the aspect ratio of the tank (*H/D*) is greater than 1.5, a second stirrer on the shaft is recommended; if the ratio is greater than 2.5, a third one; and so on. Spacing of the stirrers should approximately equal the tank diameter *D*. Closer spacing is advised for stirring non-Newtonian liquids; the Intermig is particularly suitable for these. Different types of stirrers are often combined on the shaft; for example, a disk turbine near the shaft base and downflow axial stirrers above are often used in exothermic gas – liquid processes (although the gas can produce flow instabilities on the axial stirrers). If $H/D = 2$ or 3, two or three radial-flow turbines on the shaft are also common, especially in fermentors. This arrangement can provide high power for good dispersion, but the flow pattern is compartmentalized so that overall top-to-bottom mixing is poor.

For more information on stirrer selection, including special types for high shear or for self-induction of gas (in this case the stirrer is designed to draw the gas into the liquid without a gas compressor), see → Stirring.

1.2. Operation of Stirred-Tank Reactors

Stirred-tank reactors can be operated either batchwise or continuously.

Batch Reactors. Stirred-tank batch reactors are very popular because of their flexibility and convenience. The large top surface per unit volume enables reactants to be added or samples to be removed easily by a variety of methods. The large volume available is sometimes used as a stock or buffer vessel as well as reaction space. Large tank reactors are available for reactions requiring long residence times (many hours) at commercial throughput rates. High-pressure tanks are also readily available although obtaining rotary seals for large shafts and high pressures can be a difficulty.

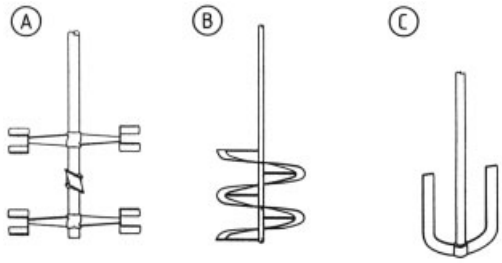

Figure 4. Agitators for high-viscosity liquids
A) Intermig (Ekato); B) Helical ribbon; C) Anchor

By adjustment of stirrer speed, and perhaps size and type, the mixing performance can be altered to suit a particular reaction step. These characteristics lead to the widespread use of batch stirred tanks in multiproduct plants.

Continuous Reactors. Stirred tanks are also widely used as continuous reactors. These reactors have one or more steady inflowing and outflowing streams, arranged so that vessel contents are at steady state. Ideally, the local concentrations of all components in the vessel are equal.

The flow pattern in a baffled stirred-tank approaches "ideal" back mixing, which means that

all fluid elements entering the vessel are "instantly" mixed with the vessel contents, and the concentrations of all components in the vessel are equal to those in the exit stream. Ideal back mixing in real stirred reactors is often assumed but not necessarily achieved in practice, especially on a large scale. [Misleadingly, the term continuous stirred-tank reactor (CSTR) is used in reaction engineering to mean an ideally back-mixed reactor; this can lead to confusion and unrealistic design assumptions. Ideal back-mixed reactor (IBMR) would be a better term]. By altering the rotational speed of the agitator the degree of back mixing can be adjusted within the limits of mechanical feasability.

The continuous stirred reactor has advantages in process control: its first-order time lag response is amenable to stable feedback control.

Cascade of Stirred-Tank Reactors. Other flow patterns can be achieved by linking several continuous stirred-tank reactors together. When linked in sequence (cascade), the more units in the cascade, the closer the overall residence-time distribution approaches plug flow. Stirred-tank stages can be assembled vertically, with a common shaft and interstage partition baffles, as in some contactors for immiscible liquids.

The significance of the residence-time distribution on reaction behavior is discussed further in Section 4.3.

Limitations. However, whether a stirred tank with batchwise or continuous configuration is selected, several important limitations should be borne in mind. If, after successful laboratory trials, inappropriate scale-up methods are applied, limitations may not be obvious until the production plant has been commissioned; the financial implications of the resultant disappointing yield or low capacity could then be serious. Small-scale work must be carried out under conditions that can be scaled up properly; preferably, a perceived process should be "scaled down" to operate under laboratory conditions. The temptation to run laboratory tanks at the highest agitation speeds or cooling rates should therefore be resisted unless these can be reproduced on a production scale (perhaps in some tubular reactor rather than a tank). Laboratory work should include a variation of stirrer speed to establish whether the rate-determining step is a physical one (interphase mass transfer or mixing; see Section 4.2).

Heat-Transfer Limitations. The achievable rate of heat transfer is the most common limitation. Although the heat-transfer intensity of stirred-tank reactors may be adequate for many cases, some reactions must be slowed down (to decrease the rate of heat evolution) when performed in a stirred-tank reactor. Often, the heating or cooling of a batch reactor takes as long as the reaction itself. Loop reactors should be considered in such cases (see Chap. 2).

Mixing Time Limitations. There are also limitations on mixing time (usually defined as the time required after injection of a pulse of soluble tracer liquid for its concentration to be within 1 % of the final equilibrium value at all locations in the tank contents). The mixing time is usually in the region of seconds to minutes. Thus, in the case of rapid reactions (99 % complete in ≤ 10 s with low-viscosity liquids), the concentrations, and hence reaction rates, will not be uniform throughout the tank contents. Localized overheating may result if the heat-transfer system was designed with the assumption of homogeneity. The selectivity of the reaction process may be adversely affected when competing reactions occur.

Both of the above limitations are exacerbated by higher viscosities of the reaction mixture, especially if the fluid is not fully turbulent.

Limitations of Multiphase Systems. Other limitations occur in multiphase systems. *Solid particles* generally must be kept in suspension rather than float on the surface or sit in heaps at the base if full contacting is to be ensured. Axial-flow agitators in downflow mode are usual for suspension of heavy particles. Floating solids may be drawn down into the liquid phase by a surface vortex that can be formed in vessels without baffles. The bottom bearings on the shaft should be designed to resist mechanical attack when solids are present.

Dispersion of an *immiscible liquid* in the main liquid is produced in high-shear and elongational-flow regions behind the agitator blade tips. To maximize dispersion, radial-flow agitators are used in baffled tanks. Except in those cases where the fraction of the dispersed phase is low (<10 %), it can be difficult to establish which will be the dispersed (droplet) phase. In continuous tanks the phase to be dispersed

should be fed into the stirrer blade-tip region if possible. In batch tanks, especially when starting with separated phases, the agitator is placed in the desired continuous phase so that it draws the other phase in. When fine dispersion of the dispersed phase is desired to achieve rapid mass transfer and reaction between the phases, but the phases must be separated subsequently by gravity (which requires larger droplets), a compromise in droplet size must be found.

The mechanism of *gas – liquid* dispersion in stirred tanks is similar to that described above for liquid – liquid systems. If flat-bladed turbines are used, an additional complication is that the low-pressure zone behind the blades draws gas in, forming a gas "cavity". This cavity "stream lines" the blades and partially blocks their liquid discharge path, so that local bubble breakup forces and bulk circulation are both impaired; the result can be up to 60 % reduction in power transmitted at typical gas flow rates. Agitators with curved blades (concave to the flow) alleviate this problem [1], [2].

To maximize the gas contact time, the gas is often introduced under the agitator blades, with a disk forming the center of the agitator to ensure that the gas does not bypass the dispersion zone. Alternatively, the gas can be drawn in through the top surface, either by a surface vortex (Fig. 2) or via a hollow agitator shaft with a vent hole above the liquid surface and a "self-inducing" type of agitator below.

The design of two-phase stirred tanks is described in more detail in → Stirring.

Shear Forces. The local shear forces in a stirred tank are not distributed uniformly, being an order of magnitude higher near the tips of the stirrer blade (as mentioned above) than in more remote parts of the tank. Some polymerizations and fermentations are sensitive to higher shear rates so a difficult compromise must be found between shear sensitivity and adequate mixing, phase dispersion, and heat transfer.

1.3. Examples

1.3.1. Multiproduct Batch Plant

Stirred-tank reactors are often used in multiproduct batch plants where a large number of similar processes are carried out, each at a relatively small annual production. Further complications in the manufacture of fine chemicals and specialty chemicals are the complex structures of the organic molecules undergoing reaction, the large number of possible side reactions, and the varied physical states (dry powder, wet solid, paste, liquid, emulsion, etc.) of the materials being used. Therefore, reactions must be delayed until analytical results are available; the actual concentrations of reactants in feeding vessels must be measured to adjust material charges depending on the analysis; and the reaction end points must be determined by sampling and testing.

In the example given here, several stirredtank reactors of 30- and 60-m^3 capacities are used to operate 150 separate but similar processes, with a total annual production of ca. 5000 t. The tanks are required to withstand a range of corrosive conditions and are made of rubber- or glass-lined steel. Mechanical design of the agitator baffles and cooling systems is constrained by the need to attach and maintain the rubber or glass lining. In a typical process, a wet solid reactant is charged to a concentrated acid solution to dissolve the reactant. The solution is then sampled to determine the reactant concentration. A second reactant is added, the exact quantity depending on the result of the analysis. The reaction is then performed under temperature control (in some cases by the addition of ice). The reaction product is moved to another stirredtank reactor, where it is reacted with another intermediate. This second intermediate is itself the product of one or more reaction stages. In this way, complex organic molecules (e.g., with molar masses of 500 – 1500 g/mol) are built up in multiple stages. The flexibility of the stirred-tank reactor, and the opportunity to intervene and adjust concentrations and reaction conditions, are ideally suited to this kind of manufacture.

1.3.2. Continuous-Campaign Stirred-Tank Reactor

The second example of a stirred-tank reactor is used to carry out a reaction between a gas and an organic liquid, catalyzed by a transition-metal catalyst. The reaction is mass-transfer controlled, and the reactor has baffles and an ag-

itator specially designed to promote good mass transfer between the gas, the liquid, and the solid catalyst. The reactor has a working volume of $30\,m^3$, but a significant part of this volume is left unused to allow for the occasional problem of disengaging the off-gas from the liquid, which may have a tendency to foam. The major difficulty is the removal of heat since the reaction is highly exothermic. The heat is removed partly by heat exchange in the vessel and partly by operating in an excess of product, which acts as a solvent and removes heat by boiling off and being condensed and returned to the reaction mixture.

Although this reaction might be carried out in a device where mass transfer is more intense, the advantage of the stirred tank is that production can be increased by simple modifications or enhancements, since its original installation and the use of a vaporizing and condensing solvent provide a safe, self-stabilizing system of heat removal.

1.3.3. Cascade of Continuous Stirred-Tank Reactors

The third example is a cascade of four continuous stirred-tank reactors. The reaction is the nitration of an organic compound. The heat of reaction for this is 100 J/h, and compared with the previous batch and campaign operations the reactors are relatively small, having working volumes of 3.8, 1.5, 1.5, and $1.5\,m^3$. The reaction is second order, and the system is designed to give a conversion $>99.9\,\%$.

In this example, heat is removed by cooling coils, and the heat load on the first reactor is 2.2×10^6 W. The reactor internals are therefore designed to operate as a forced-circulation heat exchanger, and the agitator is designed to give a high flow rate in a well-structured flow pattern over a tightly packed coil. The first reactor has a coil area of $90\,m^2$ in a total of $5\,m^3$ overall volume, and although theoretically a second-order reaction the overall reactor volume would be more productive if the largest reactor were the last in the cascade. The need to provide the required heat-transfer area determines the first reactor's overall size.

2. Loop Reactors (see also → Bubble Columns)

2.1. Description

A loop reactor consists of a pipe through which process fluids are recirculated. Often it also contains a heat exchanger, a tank, an injection device for reactants, and a separation device. Here, loop reactors are considered in which a pump is used for recirculation of the liquid (see Fig. 5). For loop reactors in which gas buoyancy is the driving force ("gas lifts"), see → Bubble Columns.

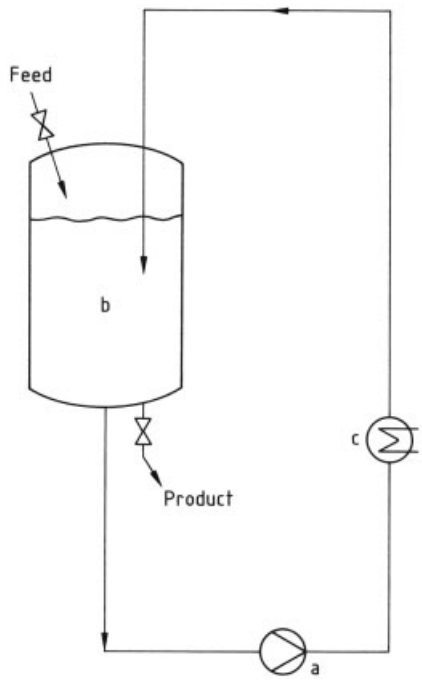

Figure 5. General batch loop reactor
a) Circulating pump; b) Tank; c) Heat exchanger

Loop reactors can be operated in either batch (closed-loop) or continuous configuration. A *batch loop reactor* generally contains a vessel (which may or may not be agitated) that is used to receive the initial charge of reactants and may also be employed for separation of a gaseous or solid product after the reaction. In some cases the energy of the recirculating stream can be used in the form of a jet to mix tank contents.

A *continuous loop reactor* has feedstreams (generally two) entering the loop (or tank, if present) and product streams leaving it (Fig. 6). The rate of recirculation around the loop can generally be set independently of throughput rate (an exception to this is the use of an ejector for gas addition, described below). This flexibility can be used to adapt the residence-time distribution of the overall reactor to requirements of the reaction. The higher the ratio of recirculating flow to throughflow, the nearer is the overall behavior of the loop reactor to ideal back mixing. Such adjustments are sometimes made between products in multiproduct reactors by altering pump speed or switching in more or less pipework.

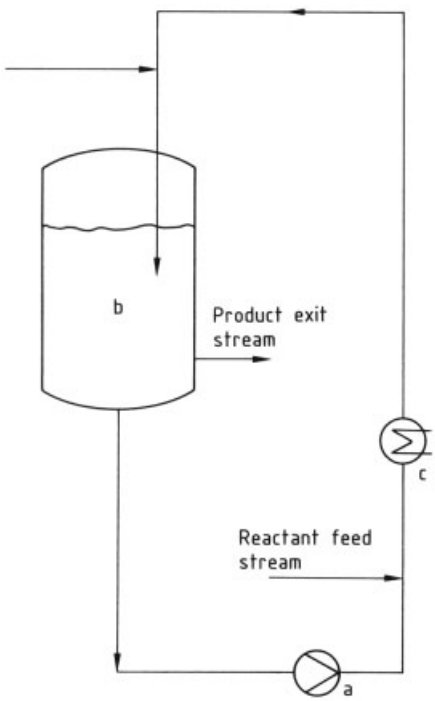

Figure 6. Continuous loop reactor
a) Circulating pump; b) Tank; c) Heat exchanger

Choice of the proper recirculation rate may also be influenced by the need to disperse a second phase (a gaseous or immiscible liquid reactant) in the reacting liquid. If a suitable in-line *disperser* (e.g., a jet, orifice, or "motionless mixer" in turbulent flow) is incorporated in the loop, specific energy dissipation rates for drop or bubble dispersion can be conveniently supplied, which are 10 – 100 times higher than those available in stirred tanks. Such dispersion intensities are useful to minimize the size and cost of the reactor and to maximize yields when the key reaction rates are intrinsically rapid but controlled by interphase mass-transfer rate.

The high dispersion intensity is especially valuable in gas – liquid systems, with their typically high interfacial tensions. Care must be taken if the gas is to be separated after the reaction because higher dispersion intensities produce smaller bubbles that make any subsequent separation more difficult. The energy in the recirculating liquid stream can be used to draw in a gas reactant by means of an ejector. The *Buss reactor* is based on this principle (Fig. 7) [3] and is often used for hydrogenation. It has the advantage of rapid gas – liquid mass transfer in the initial reaction zone, combined with high heat removal capability in the tubular heat exchanger.

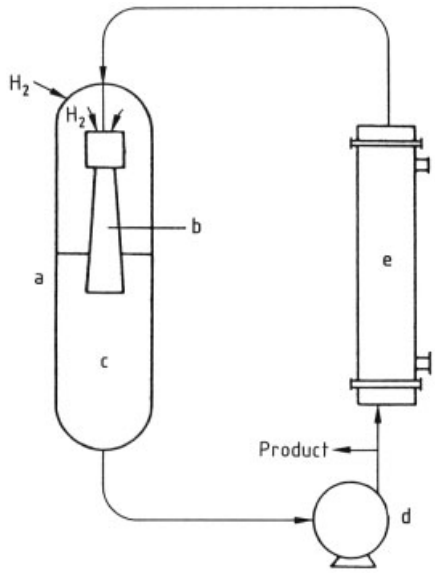

Figure 7. Buss loop reactor
a) Tank; b) Ejector nozzle; c) Reacting liquid; d) Pump; e) Heat exchanger

2.2. Fields of Application

2.2.1. Liquid-Phase Reactions

Loop reactors are preferred to stirred vessels when a large sensible heat-transfer area per unit volume is desired. A large heat-transfer area is particularly important for temperature control of rapid exothermic reactions. This is achieved by incorporating a tubular (often multitubular) heat exchanger in the loop and using a suitable fluid for sensible or evaporative cooling.

Coupling of Azo Dyes. A recirculating loop reactor is used for the coupling stage in the continuous manufacture of a dye. The reaction under consideration is

$$\text{Diazonium salt} + \text{Coupler} \rightarrow \text{Dye}$$

At a fixed pH the reaction rate can be represented simply by

$$\frac{\mathrm{d}}{\mathrm{d}t}\,[\text{Dye}] = K_2\,[\text{Diazonium salt}]\,[\text{Coupler}]$$

The rate constant is strongly dependent on pH and on the nature of the diazonium salt and coupler.

A competing decomposition of the diazonium salt can be represented by

$$\frac{\mathrm{d}}{\mathrm{d}t}\,[\text{Decomposition product}] = K_1\,[\text{Diazonium salt}]$$

Again K_1 is very dependent on the nature of the diazonium salt and on pH. Careful control of pH during the coupling reaction is therefore extremely important. To avoid problems caused by local rises in pH that occur when the pH control agent (e.g., sodium hydroxide) is added in a stirred vessel, a recirculating loop reactor may be used.

By recirculation of the buffer solution and addition of NaOH (Fig. 8), very tight control of pH and temperature conditions at the point of reaction can be obtained.

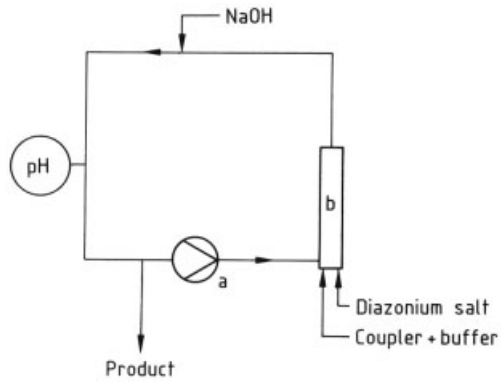

Figure 8. Loop reactor for azo coupling
a) Pump; b) Jet mixing section

The design of the mixer section is an important feature of the system. In some examples, a dye that precipitates may adhere to the surfaces. In recirculating loop reactors, a jet system can be used to keep the precipitate away from the tube walls until it has aged sufficiently to be nonsticky.

Production of Phenol. Another example is the cleavage of cumene hydroperoxide to produce phenol and acetone, a rapid liquid-phase reaction that is highly exothermic [4]. This reaction is autocatalytic, so back mixing is advantageous. By operating in a continuous loop reactor with a recirculation ratio of about 100, the temperature rise from inlet to outlet is maintained at only a few degrees.

Oxidation of organic compounds by nitric acid is carried out in tanks having such large recirculation streams through multitubular coolers that they could be called loop reactors. Other exothermic organic reactions (e.g., alkylation, hydration, oxidation) are also performed in loop reactors because of their high heat-transfer intensity and flexibility in residence-time distribution.

Evaporative crystallizers sometimes have an external heater and pumped circulation, which effectively makes them loop reactors. Such crystallizers have higher capacities than those using the traditional thermosiphon system.

2.2.2. Two-Phase Reactions

Immiscible Liquid Phases. A loop reactor is reportedly used in the manufacture of poly(vinyl alcohol) latex. The loop reactor is chosen for two-phase reactions to achieve tight control of the dispersion and a high degree of dispersion of the second, immiscible phase. This example is an attempt at changing a process that is traditionally performed batchwise to a continuous one. Care must always be taken in such exercises because 100 % conversion is not achievable in a continuous process.

Gas – Liquid Reactions. In the rapid gas – liquid reaction process shown in Figure 9 the gas is sparged into a tall narrow tank where it encounters the recirculating liquid stream in downflow. Downward drag from the liquid slows the rise of the gas bubbles and increases bubble contact time in the tank. Another gas – liquid reaction (alkane chlorination) in a loop reactor has been described in which the gas buoyancy is used to assist circulation of the liquid.

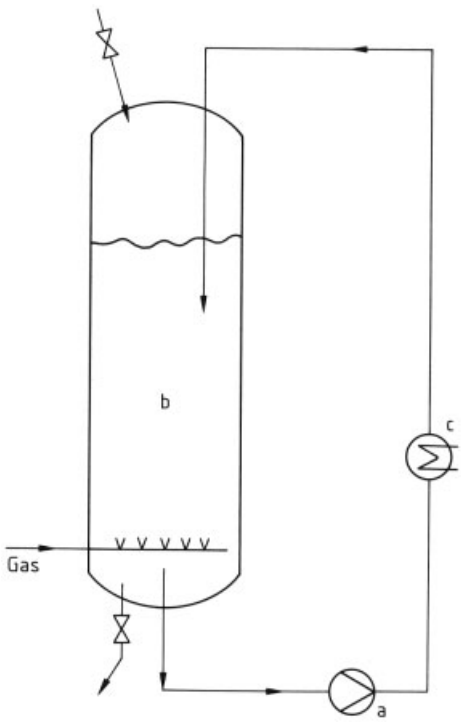

Figure 9. Gas – liquid loop reactor
a) Pump; b) Tank; c) Heat exchanger

If use of a pump for recirculation is undesirable and gas flow is sufficient, the buoyant energy of the gas stream can provide recirculation. This type of loop reactor, known as a gas-lift or air-lift reactor is employed for some fermentations, for more details, see → Bubble Columns.

3. Practical Limitations of Stirred-Tank and Loop Reactors

Reaction Volume.

Stirred reaction vessels for near-atmospheric pressure operation are available in very large sizes (up to ca. 5000 m^3). Drives, shafts, and seals suited for large reactor sizes are, however, very expensive. Tanks intended for operation at higher pressure are limited to smaller sizes due to limitations in maximum wall thickness, weight, and cost.

Loop reactors, which are assembled from pipe, can be virtually any size. However, volumes of more than ca. 100 m^3 in the pipe section would probably entail pumps of an impractical size especially if some energy is required to drive an intense mixer or contactor from the flow.

Heat Transfer. Consider, for example, an exothermic reaction in which temperature is controlled by heat transfer to walls only (e.g., a cooling jacket). Here, the heat removal capability is related to the wall area πHD, whereas the heat generation rate is proportional to the reaction volume $\pi HD^2/4$. The ratio of the two, $4/D$, therefore gives the maximum reaction heat removal capacity; thus, tubes are preferable to tanks because of their smaller diameter.

Mixing Time. Consider a fast reaction whose rate is determined by the speed of mixing of the miscible liquid reactants. To achieve a given degree of conversion in a continuous reactor, a given value of residence time/mixing time is required (this simple comparison ignores the effects of residence-time distribution). Typical values are given in Table 1 for systems with a low viscosity ($\approx$1 mPa · s) in various types of reactor. This table shows that a wide range of mixing times is available to suit various types of reaction and that the short residence time of in-line devices (coaxial jetmixer and static

mixer) can be matched by their rapid mixing performance.

Table 1. Comparison of mixing times ∗

Continuous reactors	Residence time, s	99 % Mixing time t_M, s
Laboratory scale		
Stirred tank (unbaffled)	1 → ∞	10 → 100
Stirred tank (baffled)	10 → ∞	1 → 10
Coaxial jet mixer	0.01 → 1	0.001 → 0.01
"Static mixer"	0.01 → 1	0.001 → 0.01
Plant scale		
Stirred tank (unbaffled)	10 → ∞	100 → 1000
Stirred tank (baffled)	100 → ∞	10 → 100
Coaxial jet mixer	0.1 → 10	0.01 → 0.1
"Static mixer"	0.1 → 10	0.01 → 0.1

∗ These very approximate values are for 'low' viscosity systems (∼1 mPa · s). Residence times do not include any loop pipework.

Two-Phase Dispersion. A similar comparison can be drawn between the capabilities of various reactors for mass transfer between a dispersed phase and a continuous phase. This is difficult to generalize for liquid – liquid systems because the phase ratio varies widely between systems. Therefore, performance is compared to the example of a typical gas – liquid system in which the reaction is so rapid that the overall rate is controlled by mass transfer.

The mass-transfer rate Φ_{MT} is usually described by

$$\Phi_{MT} = K_L a V (\Delta c)$$

where K_L is the mass-transfer coefficient, a is the interface area per unit volume, V is the corresponding volume, and Δc is the concentration gradient.

Again, if residence-time distribution differences are ignored (i.e., by assuming that Δc is the same function of inlet and outlet concentrations for all devices), a given degree of conversion (i.e., mass transferred) requires a given value of

$$K_L a V / \Phi_F = K_L a \tau_L$$

where Φ_F is the feed rate of the reactant and τ_L (= V/Φ_F) is the mean residence time of the liquid in the contact zone.

Various devices for gas – liquid mass transfer are compared in Table 2 [5]. Again the good performance of in-line devices is noteworthy. Turbulent flow conditions are assumed here. Higher viscosities of the liquid will result in longer mixing times and lower mass-transfer rates, especially if the flow is not turbulent. In the case of non-Newtonian fluids the performance is even lower.

4. Design Methods

4.1. Experiments

Where possible, reactor design is based on measurements of the rate constants and the orders of reaction of each important reaction step. If the reactions are other than first order, a proper selection of reactor type can be based on the optimum residence-time distribution (see Section 4.3) and the rate of reaction. Other considerations such as heat transfer, phase dispersion, and cost can also be introduced in the selection of reactor type. The size of the reactor can be established from the required degree of conversion and the rate equation of the slowest (rate-determining) step. The physical limitations of the design chosen are checked: if the reaction is very rapid the effective rate-determining step may be mixing or the interphase mass-transfer rate.

4.2. Scale-up

If detailed kinetic measurements are not possible, a convenient laboratory reactor (often a stirred vessel) is chosen and the overall performance of the reactor is measured under various reaction conditions. The effects of stirrer speed, pressure, and temperature are investigated to determine whether a physical rate limitation exists [5]. The conditions that enable optimum performance are then scaled up to the production design. Scale-up is reliable only if the laboratory work has been carried out carefully. With multiphase systems, scale-up is fraught with conflicts, and decisions about which design factors are most important must be made. Some examples are illustrated in [5].

Prediction of physical limitations such as reaction time and mass transfer rate is notoriously

Table 2. Comparison of different gas – liquid contacting devices (by permission of Butterworth & Co.) ∗

Device	$K_L a$, s^{-1}	V, m^3	$K_L a V$, m^3/s	a, m^{-1}	ε_L	Liquid mixing	Gas mixing	Power per unit volume, kW/m^3
Baffled agitated tank	0.02 – 0.2	0.002 – 100	10^{-4} – 20	∼ 200	0.9	back mixed	intermediate	0.5 – 10
Bubble column	0.05 – 0.01	0.002 – 300	10^{-5} – 3	∼ 20	0.95	approx. plug	plug	0.01 – 1
Packed tower	0.005 – 0.02	0.005 – 300	10^{-5} – 6	∼ 200	0.05	plug	approx. plug	0.01 – 0.2
Plate tower	0.01 – 0.05	0.005 – 300	10^{-5} – 15	∼ 150	0.15	intermediate	approx. plug	0.01 – 0.2
Static mixer (bubble flow)	0.1 – 2	up to 10	1 – 20	∼ 1000	0.5	approx. plug	plug	10 – 500

∗ These are orders of magnitude for comparison, not design figures. They are based on physical properties of air and water. Other considerations such as heat transfer, materials of construction, suspended solids, and hazardous liquids or gases (requiring low inventory) must obviously be taken into account.

unreliable and complex, and they should preferably be inferred from laboratory data and scaled up by using suitable correlations (→ Stirring, Chap. 4., → Stirring, Chap. 5.). Even the accuracy of various scale-up correlations is debatable. Figure 10 is an example of this for solid particle suspension; it compares the results of using various correlations to scale up the stirrer power per unit of liquid mass required to "just suspend" the particles [5]. If "scale factor = 1" is taken as the laboratory scale, movement to the right across the Figure represents scale-up to a plant-sized tank. The power input required (per unit mass) is obtained by following the curve for the chosen correlation and reading off the vertical scale. The disparity between the correlations is evident.

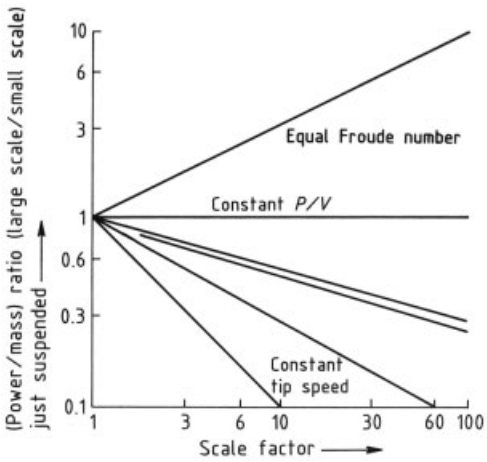

Figure 10. Conflicting scale-up relationships for particle suspension (by permission of Butterworth & Co.)

4.3. Residence-Time Distribution

4.3.1. Relevance

The ideal, "fully back-mixed" state has been defined in Section 1.1. The other extreme, for continuous reactors, is the "ideal plug-flow reactor" in which the residence time of all fluid elements is the same.

For a single reaction of order 1 or higher, the *required volume* for a given conversion will be greater for a fully back-mixed reactor than for a plug-flow or batch reactor (→ Principles of Chemical Reaction Engineering, Chap. 4.2.), for example, by a factor of 21.5 for 99 % conversion of a first-order reaction.

In more complicated reaction schemes, *selectivity* is also affected (→ Principles of Chemical Reaction Engineering, Chap. 2.3.). For example, for

$A \rightarrow B \rightarrow C$

plug flow will maximize the yield of B, whereas back mixing favors the formation of C. With parallel reactions,

$A \rightarrow B$ (order p in A)

$A \rightarrow C$ (order q in A)

if $p > q$, plug flow gives a high yield of B, whereas if $p < q$, maximum B is obtained from a back-mixed reactor.

4.3.2. Stirred Tanks

For the ideal, "fully back-mixed" state, the degree of conversion can be calculated from the following:

A mass balance over the reactor gives

$$\Phi_F c_i - \Phi_F c_o = V R_A$$

hence,

$$V/\Phi_F = \tau = (c_i - c_o)/R_A$$

where Φ_F is the feed rate, c_i the inlet concentration of reactant, c_o the outlet concentration of reactant, V the reacting volume, R_A the reaction rate, and τ the mean residence time.

For a first-order reaction

$$R_A = k c_v$$

where k is the reaction rate constant and c_v the concentration of reactant in the tank. With "full back mixing," $c_v = c_o$ hence,

$$\frac{c_o}{c_i} = \frac{1}{1+k\tau}$$

For a second-order reaction,

$$R = k c_v^2 = k c_o^2$$

$$\frac{c_o}{c_i} = \frac{(1+4k\tau c_i)^{1/2} - 1}{2k\tau c_i}$$

4.3.3. Loop Reactors

The *loop circulation* time t_L is given as

$$t_L = V/(\Phi_F + \Phi_R) = \tau/(1+R) \quad (1)$$

where Φ_R is the recirculation rate, $\tau\,(= V/\Phi_F)$ the mean residence time, and $R\,(= \Phi_R/\Phi_F)$ the recycle ratio. After mixing with the loop contents, the reactant concentration in the loop c_{iL} is

$$c_{iL} = (\Phi_F c_i + \Phi_R c_o)/(\Phi_F + \Phi_R) = (c_i + R c_o)/(1+R) \quad (2)$$

For second-order reaction

$$\frac{dc}{dt} = -k c_L^2$$

at

$$t = 0, c_L = c_{iL}$$

at

$$t = t_L, \frac{c_o}{c_{iL}} = \frac{1}{1+k c_{iL} t_L}$$

From Equations (1) and (2),

$$\frac{c_0}{c_i} = \frac{\left[\left(1+\frac{k\tau c_i}{1+R}\right)^2 + \left(\frac{4k\tau R c_i}{1+R}\right)\right]^{1/2} - \left(1+\frac{k\tau c}{1+R}\right)}{\left(\frac{2k\tau R c_i}{1+R}\right)}$$

Note that when $R \to \infty$

$$\frac{c_o}{c_i} = \frac{[1+4k\tau c_i]^{1/2} - 1}{2k\tau c_i}$$

corresponding to full back mixing, as above; when $R = 0$,

$$\frac{c_0}{c_i} = \frac{1}{1+k\tau c_i}$$

which is the equation for an ideal plug-flow reactor. More complex cases are dealt with in standard texts on reaction engineering [6], [7].

When a close approach to an ideal residence-time distribution cannot be assumed, design or scale-up requires the use of a mathematical model for the reactor that incorporates a description of the detailed flow patterns; this is dealt with in Chapter 5.

5. Reactor Modeling

5.1. Fitted Models

The earliest models for imperfectly mixed stirred vessels divided the reaction volume into *arbitrary mathematical zones of mixedness and segregation* [8], whose sizes and interlinking are described by parameters obtained by fitting to experimental data. The physical interpretation of the parameters is uncertain, so how they will change on scale-up is generally unpredictable (i.e., such methods are descriptive rather than predictive).

In the 1970s, "*semiphysical*" *models* of the reactions within the flow patterns were developed. One example is the "*zones-in-loops*" *model*, in which circulation patterns within the tank are described by loops of fully back-mixed zones in series [9]. Figure 11 shows this applied to a gas–liquid stirred reactor. The number of

zones to be used depends on the quality of mixing and can be obtained from measurements of the distribution of circulation times through the agitator. The usefulness of such models for scale-up therefore depends on the predictability of the circulation-time distribution, and data and correlations for this are limited.

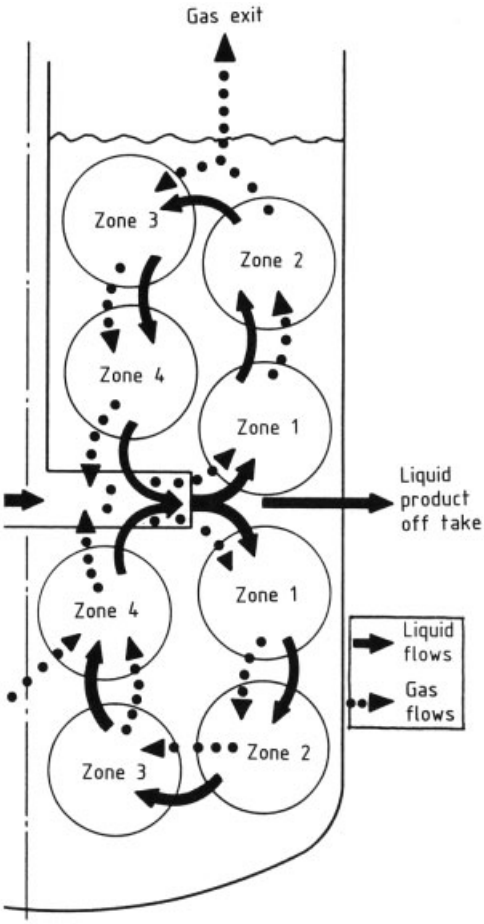

Figure 11. "Zones-in-loops" model for gas – liquid stirred reactor (by permission of Butterworth & Co.)

Exchange flows between circulation "loops" can be incorporated into the model in an attempt to represent the local turbulent mixing between circulating streams. Such models, with large numbers of mixed cells, have been used extensively by MANN and coworkers [10]. These models still require parameters (velocity profiles and exchange rates) fitted from experimental data, so they are somewhat limited in scope for general use, although the parameters seem to be well established for single six-blade disk turbine agitators in baffled tanks in the turbulent regime.

All the models described thus far are basically one (or at best two) dimensional so they cannot describe the mixing processes in baffled tanks in fine detail.

5.2. Computational Fluid Dynamics (CFD)

With modern computers and algorithms the basic equations of flow (Navier–Stokes equations plus turbulence models) can be solved together with concentration and reaction equations at many (typically 50 000) points throughout the tank. The solutions can be converged to give a detailed description of the three-dimensional mixing patterns and local concentrations. Examples of results in the form of a vertical-section velocity vector field and a tracer concentration map are shown in Figures 12 A and 12 B, respectively.

The only parameters that have to be determined experimentally are now in the turbulence model (these parameters are nearly universal) and at the agitator boundaries (universal to each type of agitator, with modifications for extreme wall proximity), so CFD offers a facility for much more predictive reactor modeling. Suitable computer programs for liquid flows are available as commercial packages; suitable boundary conditions and turbulence models have to be chosen with care. The effects of changing tank shape, internals, agitator type, inlet and outlet position, etc., can be readily predicted. Programs for handling two-phase reactions, incorporating breakup, coalescence, and interphase mass transfer are being developed.

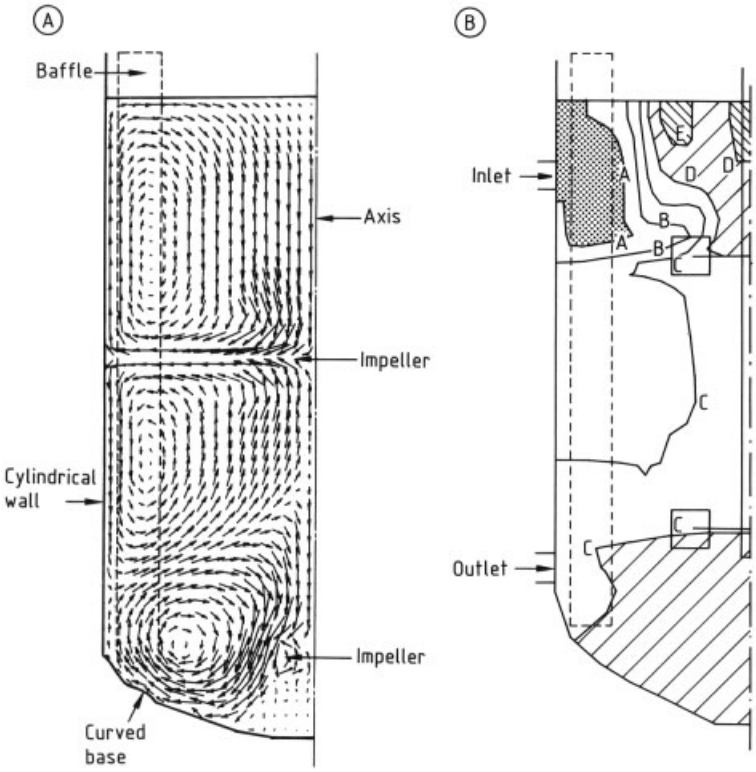

Figure 12. CFD analysis of stirred tanks
A) Flow in a batch stirred tank; test case example 5-m tank, velocity plot (XY plane) 15.0 → = 1.4 m/s; B) Concentration of tracer: contour values A = 0.55; B = 0.53; C = 0.50; D = 0.48; E = 0.45; relative to inlet value = 1.0

6. References

1. Scaba, US 477 9990, 1988.
2. Chemineer, Hollow-Blade Turbine Agitator, Bulletin no. 714.
3. R. J. Malone, *Chem. Eng. Progress* 1980 June, 53.
4. S. R. I. Reports 22 a (1972); 22 b (1977); Chem. Systems Report PERP 88/6.
5. N. Harnby (ed.): *Mixing in the Process Industries,* Butterworths, London 1985, chap. 17.
6. K. R. Westerterp, W. P. M. van Swaaij, A. A. C. M. Beenackers: *Chemical Reactor Design and Operation,* Wiley Interscience, New York 1984.
7. O. Levenspiel: *Chemical Reaction Engineering,* Wiley Interscience, New York 1976.
8. D. P. Rao, L. L. Edwards, *Chem. Eng. Sci.* **28** (1973) 1179.
9. R. Mann, J. C. Middleton, I. B. Parker, *Proc. Eur. Conf. Mixing 2nd,* Cranfield 1978.
10. R. Mann, P. P. Mavros, J. C. Middleton, *Inst. Chem. Eng. Symp. Ser.* (1978) no. 64, G 1.

Tubular Reactors

For other industrial reactors and their applications, see → Stirred-Tanc and Loop Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

David Cresswell, ICI Chemicals and Polymers Limited, Runcorn, United Kingdom (Chaps. 1, 5, and 6)

Arthur Gough, ICI Chemicals and Polymers Limited, Runcorn, United Kingdom (Chap. 2)

Graeme Milne, ICI Chemicals and Polymers Limited, Runcorn, United Kingdom (Chaps. 3 and 4)

1. Introduction 2
2. Olefin Production by Steam Cracking 2
2.1. Background 2
2.2. Chemistry of Olefin Production . . . 2
2.3. The Reactor 4
2.4. Reactor Design 4
2.5. Heat Recovery 7
2.6. Control of the Reactor 8
2.7. Coking 9
2.8. Alternatives to the Tubular Reactor 9
3. Gas-Phase Chlorocarbon Cracking Reactions 10
3.1. Reaction Description 10
3.2. Reaction Thermodynamics 10
3.3. Reaction Mechanism 10
3.4. Reactor Design and Operating Conditions 11
3.5. Reactor Modeling 11
3.6. Reactor Coking and Carbon Formation 12
4. Gas-Phase Halogenation Processes 12
4.1. Reaction Description 12
4.2. Substitutive Chlorination Reactors 12
4.2.1. Reactor Modeling 13
4.2.2. Reactor Coking and Soot Formation . 13
4.3. Substitutive Bromination Reactors 13
4.3.1. Reactor Design 14
4.3.2. Reactor Modeling 14
4.3.3. Reactor Coking and Carbon Formation 14
4.4. Perhalogenation Reactors 14
4.4.1. Reaction Description 14
4.4.2. Reactor Design and Operation 14
4.4.3. Reactor Modeling 14
4.4.4. Reactor Coking and Carbon Formation 15
5. Fundamental Modeling Approaches – An Evaluation 15
5.1. Thermal Decomposition of Ethane . 15
5.2. Improving the Model 17
6. Conclusions 18
7. References 18

Symbols used in this article are defined in → Principles of Chemical Reaction Engineering and → Model Reactors and Their Design Equations.

1. Introduction

Tubular reactors are in widespread industrial use for carrying out *homogeneous gas-phase reactions* requiring high temperature, plug-flow conditions, and high rate of heat transfer. Examples include (1) steam cracking of hydrocarbon feedstock to make olefins, (2) homogeneous dehydrochlorination of 1,2-dichloroethane to vinyl chloride, (3) chlorination of methane or chloromethane to higher chloromethanes, (4) pyrolysis of chlorodifluoromethane to tetrafluoroethylene, and (5) reaction of methane and ammonia to produce hydrogen cyanide.

Certain *liquid-phase reactions* requiring plug flow are also carried out in tubular reactors or, alternatively, in a cascade of continuously stirred tank reactors. The manufacture of ethylene glycol by hydrolysis of ethylene oxide is one such large-scale example.

Tubular reactors have found less application in *polymerization reactions*, where batch reactors often still predominate. The production of polyethylene is a notable exception.

This article deals mainly with the practical aspects of tubular reactors, with the intention of supplementing the more fundamental treatments given in → Principles of Chemical Reaction En-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_181

gineering and → Model Reactors and Their Design Equations. Factors such as tube metallurgy, coking, quenching, heat recovery, and reactor control are addressed. A historical perspective of the evolution of reactors and of possible future directions for development is given. To illustrate the practical aspects in some depth, the following case studies have been selected:

Olefin production by steam cracking
Dehydrochlorination of 1,2-dichloroethane
Chloromethane manufacture

The concluding section examines the efficacy of free-radical kinetic models to predict via first principles the product distribution from a model industrial reaction – the cracking of ethane – at ultrashort residence times and temperatures considerably higher than those achievable commercially at present.

2. Olefin Production by Steam Cracking (→ Ethylene)

2.1. Background

Production of olefins by steam cracking of hydrocarbons provides an excellent example of the use of tubular reactors for an important petrochemical process. The reaction is highly endothermic, and the highest selectivity toward desired olefinic products is achieved at short residence times with temperatures up to 900 °C at the reactor exit, so the reactor must be designed for optimum heat transfer.

Steam cracking is the major source of ethylene. Depending mainly on the feedstock choice, and to a lesser extent on reactor configuration and operating conditions, the ethylene yield can range from as high as 80 wt % from ethane (when unconverted ethane is recycled) to as low as 20 wt % for gas oil as feedstock. Thus, with higher molar mass (heavier) feedstocks, large amounts of byproducts are produced that have a deleterious effect on process economics. The major byproducts are propene (propylene), 1,3-butadiene, and benzene. Extraction from cracker C_4 streams is virtually the only source of butadiene (→ Butadiene, Chap. 4.2.). Typical product yield from commonly used feedstocks are given in Table 1.

2.2. Chemistry of Olefin Production

The cracking of hydrocarbons occurs almost entirely by homogeneous gas-phase free-radical chain reactions [1]. In the simplest case, ethane pyrolysis, the steps can be set out as follows:

$$C_2H_6 \longrightarrow CH_3^\bullet + CH_3^\bullet \quad (1)$$

$$C_2H_6 + CH_3^\bullet \longrightarrow C_2H_5^\bullet + CH_4 \quad (2)$$

$$C_2H_5^\bullet \longrightarrow C_2H_4 + H^\bullet \quad (3)$$

$$C_2H_6 + H^\bullet \longrightarrow C_2H_5^\bullet + H_2 \quad (4)$$

Step (1) is the initiation reaction, and steps (2) – (4) are propagation reactions. The chain is terminated when any two radicals combine to form one of the stable species, for example

$$CH_3^\bullet + H^\bullet \longrightarrow CH_4$$

or

$$C_2H_5^\bullet + C_2H_5^\bullet \longrightarrow C_4H_{10}$$

If these were the only reactions to occur, the products from ethane pyrolysis would consist of ethylene and hydrogen in equimolar quantities, with methane and *n*-butane as byproducts. The amount of methane is determined by the length of the radical chain; at low conversion this can be very large (several hundred) and < 1 % methane is formed, but at industrially important ethane conversions (55 – 75 %), chain lengths of 5 to 10 are the norm, several percent methane is produced. At higher conversions, secondary reactions become important. Examples of these are

$$C_2H_4 + R^\bullet \longrightarrow C_2H_3^\bullet + RH$$

$$C_2H_3^\bullet \longrightarrow C_2H_2 + H^\bullet$$

$$2\,C_2H_3 \longrightarrow C_4H_6$$

Butadiene can undergo Diels – Alder reactions with itself; the reaction of butadiene with acetylene or ethylene leads to the formation of aromatics.

In the cracking of higher hydrocarbons, rapid decomposition of larger radicals – formed by hydrogen abstraction or by carbon – carbon bond

Table 1. Typical yields of major products from primary olefin plant feedstocks (in wt % on a once-through basis) *

Product	Feedstock				
	Ethane	Propane	*n*-Butane	Naphtha	Gas oil
Hydrogen	3.7	1.4	1.1	0.9	0.8
Methane	4.4	23.0	20.0	19.0	11.0
Ethylene	49.0	35.0	38.0	29.0	20.0
Ethane	35.0	3.5	4.0	4.4	3.5
Propene	1.8	14.5	16.5	14.5	13.0
Propane	0.2	10.0	0.5	0.5	0.6
Butadiene	1.5	2.6	3.5	4.0	4.5
Butenes	0.6	1.2	2.2	4.0	4.5
Butanes	0.9	0.0	5.0	0.4	0.2
Pentanes	4.0	1.7	2.3	3.0	4.5
Benzene	0.6	3.0	2.5	7.0	6.0
Other gasoline	2.0	2.0	3.0	8.0	12.0
Fuel oil	0.2	0.5	0.5	4.0	19.0

∗ All yields are approximate. With LPG (i.e., propane – butane) as feedstock yields are affected only by tube design and operating conditions. Naphthas and particularly gas oils have variable composition, which affects yield as well as coil design and operating conditions. Unconverted ethane and propane are usually separated from the products and recycled to the furnace.

scission similar to that in Equation (1) – is an important contributor to olefin yields, for example,

$$C_7H_{15}^{\bullet} \longrightarrow C_5H_{11}^{\bullet} + C_2H_4$$

$$C_5H_{11}^{\bullet} \longrightarrow C_3H_7^{\bullet} + C_2H_4$$

A chemical model that can simulate product yields from complex feed mixtures with sufficient accuracy to be commercially useful for a full-scale furnace simulation requires more than a hundred different species and perhaps more than a thousand reactions. The wide range of concentrations of components gives rise to stiff differential equations, which are not easy to integrate along a tubular reactor, especially when combined with accurate representation of heat transfer into the reacting gas stream. Such models have been written but tend to be proprietary and expensive. Several examples covering the cracking of ethane and propane at high conversions have been published [2], [3]. One that covers the formation of heavier byproducts more adequately than most includes about 47 species and more than 400 reactions [3].

Surprisingly, despite the complexity of this chemistry, simple first-order kinetics can provide a quite accurate representation of conversion as a function of residence time and temperature up to 70 % conversion.

The chemical reactions outlined above do not involve steam, despite the name steam cracking. Steam has two purposes as an essentially inert diluent: First, it *lowers the partial pressure* of hydrocarbon while absolute pressure is maintained above atmospheric. The steam can then be separated easily from the products downstream of the reactor by condensation. Low partial pressure has the beneficial effect of increasing olefin yield at the expense of the formation of paraffins and heavy products with low commercial value. The highest molar mass products tend to condense in the quench system downstream of the reactor. Condensation is minimized by low partial pressure.

Second, steam aids *heat transfer* from the walls of the reactor to the process fluid and is important in effecting the vaporization of heavier feedstocks such as full-range naphtha and gas-oil immediately upstream of the reactor.

The steam is not entirely inert; it *reacts with coke* deposited on the walls of the reactor and removes some of it as CO and CO_2, which are formed by the water – gas and shift reactions:

$$C + H_2O \longrightarrow CO + H_2$$

$$CO + H_2O \rightleftharpoons CO_2 + H_2$$

Coke formation, which has profoundly adverse effects on reactor operation, is discussed in more detail in Section 2.7. Typical concentrations of CO in the reactor effluent are a few hundred ppm; the concentration of CO_2 is about onetenth that of CO. These proportions increase rapidly with

increasing severity of cracking. Provision must be made in the separation section of the plant for removal of these gases to avoid contamination of ethylene by CO_2 and of hydrogen by CO. Steam also reacts to a small extent with hydrocarbons to give a host of minor products. These include the lower carboxylic acids, acetaldehyde, acetone, and phenols – all present at only a few parts per million in reactor effluent. However, most of them can have some deleterious effect (e.g., corrosion, contamination of products, and polymer formation in the separation section).

Since the process water formed from condensation of dilution steam is contaminated by phenols and carboxylic acids, continuous addition of alkali is required to control pH and prevent corrosion. A fraction of the water must be blown down rather than completely recycled to the dilution steam generator to avoid buildup of contaminants. This accounts for a significant part of the effluent BOD from an olefin plant.

2.3. The Reactor

The chemistry outlined in Section 2.2 leads to the conclusion that the highest yields of valuable olefins will be obtained by cracking at short residence time, high temperature, and low partial pressure. Rapid quenching of effluent from the reactor is required to "freeze" the product spectrum and stop further polymerization of the olefins and dienes produced to give undesirable secondary products of high molar mass. The low partial pressure also helps to prevent these secondary reactions, which are usually higher than first order.

Short residence time and low partial pressure are particularly important with heavier feedstocks such as naphtha and gas oil. The yields of olefins from ethane and propane are affected less by increased residence times. These low molar mass feeds require more heat input per tonne than heavier feeds and are less easily cracked, requiring a higher reactor outlet temperature or residence time than naphtha or gas oil to achieve viable conversions. Heavier feeds also start cracking at a significant rate at lower temperature than low molar mass alkanes, so the rapid heat input must begin at a lower temperature (ca. 100 K lower for gas oil than for ethane). With heavier feedstocks, lower partial pressures are also more important to avoid secondary reactions. Typical reactor inlet and outlet conditions and steam – hydrocarbon ratios are given in Table 2.

These requirements were discovered empirically before the advent of detailed kinetic models. In fact, full kinetic models indicate that a small but valuable increase in ethylene yield can be gained by delaying the quench for a few milliseconds after exit from the reaction zone to allow the reaction to continue adiabatically until the temperature has dropped 10 – 20 K [4].

2.4. Reactor Design

Since rapid heating to reaction temperature is required and the reaction is highly endothermic, it must be carried out in tubes with a high surface-to-volume ratio. Over the years, reactor design has evolved steadily to achieve higher and higher heat fluxes in combination with low pressure drops. The rate of progress has been determined largely by progress in metallurgy.

When the first full-range naphtha crackers were designed, cracking the naphtha in directly fired tubes was considered undesirable because of the risk of high coking rates induced by the high metal temperature required. Instead, the hydrocarbon feed was heated until cracking began and then mixed with a large volume of high-temperature steam, and the bulk of the reaction took place adiabatically with a residence time typically of > 1 s and a reactor exit temperature of only ca. 700°C.

Ethylene yields achieved with this type of reactor were only ca. 21 % when naphthas were used as feedstocks, which in a modern high-severity furnace could be expected to yield more than 30 % ethylene. Because of this and the poor heat integration achieved on early crackers, this type of reactor was superseded in the early 1960s by tubes suspended in a directly fired furnace similar to that used for lighter feedstocks. Although many attempts have been made to replace the tube – firebox type of reactor (see Section 2.8), none has gained commercial significance, and all current olefin plants use tubular reactors in fireboxes.

In the *earliest tubular furnaces* the internal diameters of the reactors (usually referred to as the coil) were 100 – 150 mm, the total lengths

Table 2. Typical temperatures at inlet and exit of reactor and steam – hydrocarbon ratio for major olefin plant feedstocks

	Ethane	Propane	*n*-Butane	Naphtha	Gas oil
Inlet temperature ∗, °C	650	650	600	600	550
Exit temperature ∗∗, °C	850	850	850	850	770
Steam/hydrocarbon ratio, m/m	0.3	0.4	0.4	0.5	0.7

∗ Inlet temperature is the temperature at which the rate of cracking becomes significant and the hydrocarbon must then be heated rapidly.
∗∗ Exit temperatures vary up to ± 30 K depending on furnace design and are significantly higher in the Millisecond furnace.

of the reactor varied between 70 and 90 m [5]. Residence times were on the order of 1 s, and the outlet temperature was ca. 800 – 820 °C for naphtha. This gave ethylene yields of 25 – 26 % from naphthas. The tubes were made of a ductile alloy with a high nickel content (e.g., Incoloy 800), which allowed them to be formed by extrusion. Metal temperatures were limited to slightly more than 1000 °C. The high ductility of the metal meant that the tubes could not be suspended vertically in the firebox but had to be supported horizontally at 1- to 2-m intervals. Reactor wall thicknesses were between 6 and 10 mm.

At the start of the run, when the tubes were free of coke, the temperature drop between the inside of the tube wall and the process fluid was on the order of 100 K, and the temperature gradient across the wall ca. 20 K. Thus, when the coil exit temperature was 820 °C, the start-of-run temperature of the external surface (skin temperature) was ca. 940 °C. During the run, coke is deposited, especially toward the exit of the tube, and the skin temperature required to maintain a steady process temperature approaches the metallurgical maximum. Typical run lengths before the tubes required decoking were 30 to 40 days.

Introduction of less ductile alloys (which means that reactor tubes must be made by centrifugal casting) [6] allowed higher metal temperatures (1080 – 1100 °C), and the tubes could be suspended vertically in the firebox because they did not stretch under their own weight. This allowed a better mechanical design of the furnace. The tubes were usually serpentine, with six or eight vertical legs joined by return bends, and several such coils were usually found in a single firebox (Fig. 1). The higher metallurgical temperature meant that the process temperature could be increased; the heat flux could be increased to achieve shorter residence times; or the run length could be increased by allowing greater buildup of coke. Designers usually took advantage of a combination of these possibilities.

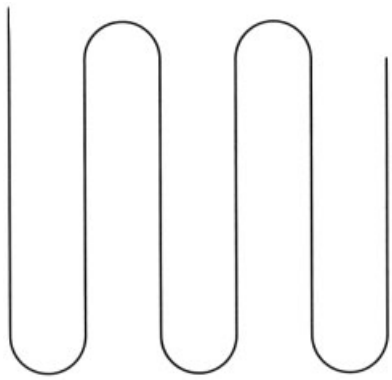

Figure 1. Typical early cracker tube design
Uniform internal diameter 100 – 150 mm; 70 – 80 m length; vertically mounted once suitable alloys were available

Typical of this development was the *Lummus SRT-I furnace* (SRT stands for short residence time) introduced in the mid-1960s (→ Ethylene) [7]. This furnace had four cracking coils of 100-mm internal diameter, each with eight vertical passes and a total length of ca. 75 m. The coils were heated by a combination of gas-fired wall burners and floor burners (gas or oil). Each coil was capable of cracking ca. 2.5 t/h of hydrocarbon; the residence time was approximately 0.5 s. At a coil exit temperature of 840 °C, the ethylene yield was typically 2 – 3 % better than in the older type of furnace. Heat fluxes in this type of furnace are on the order of 100 kW/m^2.

Each coil had independent flow controllers for the hydrocarbon mixture and the steam. The exit temperatures for each coil were controlled independently by controlling the firing in the zone opposite that coil. Some interaction occurred between adjacent zones.

Many of these furnaces are still in service, and although selectivity has been increased by subsequent developments, the increase in selectivity for ethylene is not necessarily economically attractive because many of the coproducts are valuable.

The search for ever shorter residence times and lower pressure drops to give higher selectivities led to various diversions from the simple serpentine coil.

One development was to make the first one or two passes of each coil from a smaller-diameter tube. This gave a larger surface-to-volume ratio, allowing higher heat inputs to be achieved and full cracking temperature to be reached more rapidly. Overall pressure drop was also reduced. In a typical arrangement (Fig. 2), four parallel 50-mm tubes were followed by two 75-mm tubes and then a single 100-mm tube. Later Lummus SRT designs have this general layout [8].

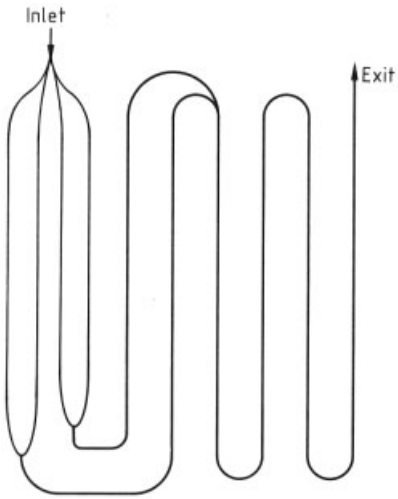

Figure 2. Example of reactor geometry giving improved heat transfer and lower pressure drop
First four parallel tubes have ca. 50-mm internal diameter; next two parallel tubes, ca. 75-mm internal diameter; later tubes, ca. 10-cm internal diameter

An alternative was to use a large number of short tubes with a relatively small diameter. A good example of this is the *Stone & Webster USC furnace* [9] (→ Ethylene). This furnace typically uses tubes of 50-mm internal diameter in the form of U tubes, which are less than 30 m long. Each tube has a capacity of less than 1 t/h of naphtha, but up to 48 tubes are accommodated in a single firebox. Individual flow controllers for each coil would be uneconomical, and flow is therefore distributed between the coils by sonic flow across orifice plates or venturis. Residence time is about 0.25 s, and coil exit temperature can reach 870 °C.

The *Pullman – Kellog Millisecond furnace* has tubes of only 25-mm internal diameter (→ Ethylene) [10]. These tubes have no return bends and are only about 20 m long. Residence time is approximately 0.1 s, and coil outlet temperature ca. 900 °C. Decoking of tubes must be performed relatively often, which is acceptable in return for the higher selectivities achieved toward olefins. As another aid to heat transfer, "rifled" tubes are employed to increase the internal surface area in contact with the process stream.

Most modern furnaces have a total hydrocarbon feed rate of 15 – 25 t/h, and a typical olefin plant will have between 8 and possibly 20 furnaces. This allows furnaces to be taken off-line for decoking and maintenance, without causing undue changes in the loading on the separation section of the plant, which is usually a single-train unit. Table 2 gives typical physical parameters for a modern cracking furnace tube at inlet and outlet conditions.

Virtually all heat transfer to the outside of the tubes occurs by radiation from hot combustion gases in the firebox and from the refractory of the walls. Therefore, the arrangement of any tube in the firebox must not cause shadowing of adjacent tubes. The thermal conductivity of the high-temperature alloys used is not high; thus heat conduction around the circumference of the tube from the hottest areas with a full "view" of the hot refractory to shaded areas is not very large, compared with the heat fluxes required. The uneven circumferential temperature distribution means that the hottest sections must be unnecessarily hot for a given average heat flux. As the rate of coke formation has a high activation energy, this can lead to excessive coking in the hottest areas and shorter run lengths between decokes.

The designer is left with very little option than to have a single row of tubes suspended down the center of the firebox. The box must be quite wide to give sufficient optical path length for radiative heat transfer from the hot gases and, in the case of floor or roof firing, to avoid flame impingement.

2.5. Heat Recovery

Heat recovery from the reactor system falls into two separate sections, recovery of the heat from firebox gases and recovery of the sensible heat from process gases leaving the reactor at up to 900 °C.

The very high heat fluxes required – well in excess of 100 kW/m^2 in modern furnaces – coupled with the high skin temperatures, especially as the end of the run is approached, require that fireboxes be at > 1200 °C. This means that the efficiency of the radiant section of the furnace (where the reactor tubes are) is only ca. 50 %, the rest of the heat generated by combustion of the fuel being carried out of the firebox as sensible heat in the flue gas. In later furnace designs the radiant efficiency has been raised a few percent by heating the air required for combustion before it enters the furnace, often by integration with a gas turbine drive from one of the main compressors in the separation section of the plant, but heat recovery is still very important to operation of the reactor.

Efficient heat recovery from *flue gases* forms an integral part of reactor design. Heat from the flue gases is used to preheat the feedstock, including vaporization in the case of liquid feedstocks, and to superheat the dilution steam. Optimization of the overall heat integration of the entire plant usually dictates that the hydrocarbon feed be preheated to 100 – 150 °C via the low-grade heat recovered in the fractionation section of the plant and that the heat required to vaporize the dilution steam (equal to the latent heat of vaporization) be provided from elsewhere in the plant. Full heat recovery from the flue gas is therefore achieved by preheating high-pressure boiler feedwater and superheating the steam produced in the exchangers that cool the process gas. Modern furnaces achieve overall thermal efficiencies in the mid-90 % region.

The *process gas* typically leaves the reactor at 800 – 900 °C and is still rapidly reacting. As explained above, a few extra milliseconds of adiabatic reaction can increase the yields of valuable products, but the reactants must then be quenched rapidly. Quench system forms an integral part of the reactor because the temperatures do not permit valves, etc., between the reactor and the quench section.

For *lighter feedstocks* (LPG and naphtha), quenching is usually achieved in tubular heat exchangers with ca. 100 tubes in parallel with high-pressure steam raised on the outside. These are known as transfer line exchangers (TLXs, sometimes TLEs). For *heavier feeds*, direct quenching with heavy oil (i.e., without exchanger) was considered necessary to avoid rapid fouling of the narrow exchanger tubes by the heavy tars produced from gas oil. This meant a considerable loss of thermal efficiency. Recent designs of exchangers have enabled everything except the effluents from cracking vacuum gas oil to be quenched indirectly.

The design of the TLX requires careful optimization. Variables at the disposal of the designer are the tube diameter and length, and the pressure of the steam coolant. The objective is to minimize the pressure drop while achieving high heat-transfer area and, at the same time, to avoid cooling the process stream to a temperature at which a large amount of heavy ends would condense at the exit end of the tubes. This condensation increases roughly exponentially with decreasing temperature [11]. Most TLXs now operate with a steam side pressure of 100 bar with steam – water in a narrow annulus sur-rounding each individual process tube to increase the flow rate and hence the waterside heat-transfer coefficient. The internal diameter of the process tubes ranges from 20 to 35 mm. Start-of-run exit temperature is ca. 330 °C, and end-of-run temperature up to 450 °C.

2.6. Control of the Reactor

Once the basic design parameters of the cracking furnace have been determined, the product spectrum from a given feedstock is fixed within relatively narrow limits. The operator, however, can control several parameters that may significantly affect the yields from the furnace and hence the profitability. On individual furnaces, the operator can usually control the hydrocarbon feed rate, dilution steam rate, and coil exit temperature. For the furnaces as a whole, absolute pressure can be controlled within a limited range by adjusting the suction pressure of the compressor. The coil exit temperature has by far the greatest effect on product spectrum, and its effect on product yields is shown qualitatively in

Figure 3. Note that the ethylene yield increases rapidly and monotonically with exit temperature, whereas propene goes through a maximum. Modern furnaces usually operate to the right of this maximum, so the ratio of these two important products can be influenced over a considerable range by temperature adjustment. Increasing temperature increases the rate of coking, leading to shorter run and an increased methane-to-olefin ratio, which can result in constraints and increased energy consumption in the separation end of the plant.

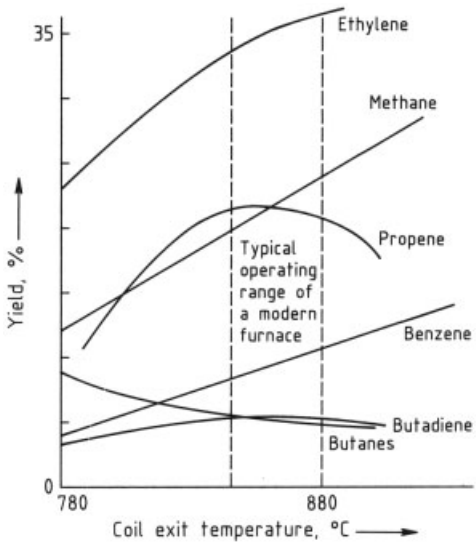

Figure 3. Schematic showing qualitative effect of temperature on yields of major products
The values given are approximate, exact values depend on the coil geometry and feedstock

Olefin production is favored and coke deposition reduced by low partial pressures, which can be achieved by either increasing the dilution steam-to-hydrocarbon ratio or reducing the compressor suction pressure. Both of these increase energy consumption, so a fine balance exists between them to obtain maximum benefit.

The total amount of heat that can be put into the reaction tubes is also limited. It varies from furnace to furnace, depending on the firing and the amount of coke deposited. This leads to complex choices between hydrocarbon feed rate and temperature.

For these reasons, most operators have computer control and optimization for their furnaces. This optimization is performed with a model relating the yields of at least the major products from the furnaces to the controllable parameters; it is coupled to a model of the separation section of the plant and to an economic model representing the value of the products, the cost of fuel, etc. The optimizer may be run off-line to set conditions for a few days in advance, but it is more often run on-line, generally with feedback from on-line analysis of the furnace effluent. Optimization of olefin plants was discussed soon after introduction of the computer; the first such system was installed in the late 1960s [12], but several are now available [13]. These systems are expensive to install and maintain, but cash flows from olefin plants are so large that these systems can increase profitability by several million dollars per year compared to manual control.

2.7. Coking

Coking is an extremely complex subject whose mechanism is not well understood. The field has been reviewed [14]; a brief summary is given here.

Carbonaceous deposits can form basically in four locations in the furnace; these deposits are collectively known as coke. The first location is the point at which the last part of the liquid feed is evaporated in the convection section of the furnace. This was a problem on some early furnaces, but improvements in design have virtually eliminated it. The second location is at the other end of the system, at the exit of the quench where the heaviest components of the cracked product condense on the cool surfaces and gradually dehydrogenate. This increases pressure drop and reduces heat recovery. For many furnaces that crack heavy feedstocks, this coke lay down often limits run length, rather than coke deposition in the reaction section. The TLX design does not usually allow on-line decoking; the furnace must be taken off-line and the exchanger water-washed. Mitsubishi has designed an exchanger that allows on-line decoking [15], and several patents claim on-line decoking of conventional TLXs [16].

Neither of these two deposits can be regarded as true coke. They are really heavy hydrocarbon tar, with a hydrogen content of 3 – 4 %. The other two deposits can be regarded as true cokes, being extremely hard and having very low hydrogen content. They are deposited toward the end of the reactor tube itself and in the inlet of the TLX. Deposition in the inlet of the TLX seems to occur only when ethane and LPGs are used as cracking feeds. Inlet cone geometry of the TLX appears to have a significant, but not well-understood, effect on this.

Considerable evidence has accumulated [14] that deposition in the tube itself is metal catalyzed, with nickel identified as the major culprit. Many patents claim to reduce this catalytic action, either by coating the inner surface of the tubes [17], by using a low nickel alloy lining [18], or by adding a coking inhibitor continuously to the feed.

Sulfur occurs naturally in many hydrocarbon feedstocks, and this can markedly reduce the catalytic effect of the tube metal, at least in the early part of the run. Sulfur is usually added to feeds that do not contain it at a level of 50 – 100 ppm, often in the form of dimethyl disulfide. Apart from sulfur, which has been used too long to be patent protected, patents cover phosphorus compounds [19], boron compounds [20], and various combinations of tin, antimony, bismuth, copper, and other compounds [21]. None of these seem to provide a complete cure, especially on heavier feeds, and all have potential problems such as contamination of the products, contamination of the environment, or corrosion. Coking remains a major unsolved problem.

2.8. Alternatives to the Tubular Reactor

Many attempts have been made to improve on the tubular reactor for olefin production. Two of them have come close to commercial success. One was the Carbide Kureha process [22], in which the hydrocarbon was mixed directly with hot combustion gases in a tubular reactor with very short residence time (a few milliseconds). This process was developed to an advanced stage for cracking crude oil or fuel oil, but the continued ready availability of lighter feeds made it unattractive.

The Stone & Webster/Gulf process [23] cracks the hydrocarbon feed by rapid mixing with very hot solids, which are recycled via a generator that burns off deposited coke. A unit with a throughput equivalent to a conventional cracking furnace has been built and operated for several years.

Thus far no one has moved away from the conventional tubular reactor for a full-scale plant.

3. Gas-Phase Chlorocarbon Cracking Reactions

3.1. Reaction Description

The term chlorocarbon cracking reactions means reactions of the type

$$\mathrm{RCl} \longrightarrow \mathrm{Olefin} + \mathrm{HCl}$$

where RCl is an chlorinated aliphatic hydrocarbon.

Two such processes that have been used on an industrial scale are the cracking of 1,2-dichloroethane [*107-06-2*] to produce vinyl chloride [*75-01-4*] (→ Vinyl Chloride) and the cracking of 1,1,2,2-tetrachloroethane [*79-34-5*] to produce trichloroethylene [*79-01-6*] (→ Chlorinated Hydrocarbons):

$$\mathrm{CH_2ClCH_2Cl} \longrightarrow \mathrm{CH_2{=}CHCl} + \mathrm{HCl}$$

and

$$\mathrm{CHCl_2CHCl_2} \longrightarrow \mathrm{CHCl{=}CCl_2} + \mathrm{HCl}$$

The installed reactor capacity for 1,2-dichloroethane cracking is of the order of 20×10^6 t/a [24]; that for tetrachloroethane cracking was once several hundred thousands tonnes per year, but is now much reduced (probably of the order of 10×10^4 t/a) [25].

3.2. Reaction Thermodynamics

Heats of reaction and equilibrium constants are listed in Table 3. These reactions are significantly endothermic. The cracking reactions require temperatures of up to 700 K to go to completion (Fig. 4).

Table 3. Reaction heats and equilibrium constants for cracking reactions of 1,2-dichloroethane and 1,1,2,2-tetrachloroethane

Reaction	ΔH^0_{298},	log K_p						
	kJ/mol	300 K	400 K	500 K	600 K	700 K	800 K	900 K
$CH_2ClCH_2Cl \longrightarrow CH_2{=}CHCl + HCl$	72.4	−5.17	−2.02	−0.12	1.15	2.06	2.74	3.26
$CHCl_2CHCl_2 \longrightarrow CHCl{=}CCl_2 + HCl$	54.4	−1.72	0.67	2.11	3.07	3.76	4.26	4.65

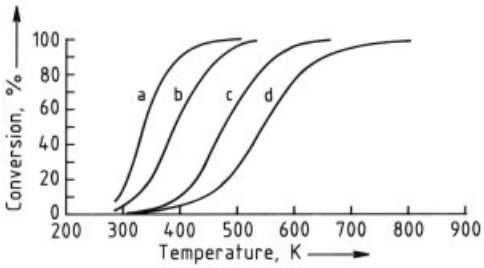

Figure 4. Equilibrium conversions for 1,2-dichloroethane and 1,1,2,2-tetrachloroethane cracking reactions
a) 1,1,2,2-Tetrachloroethane, 1 bar; b) 1,1,2,2-Tetrachloroethane, 10 bar; c) 1,2-Dichloroethane, 1 bar; d) 1,2-Dichloroethane, 10 bar

3.3. Reaction Mechanism

The cracking reactions of 1,2-dichloroethane and 1,1,2,2-tetrachloroethane are thermally initiated free-radical processes that require temperatures of 470 – 520 °C to proceed at industrially significant rates.

The chain length in these reactions is very long (of the order of 10^6); as a consequence the reaction rate is extremely sensitive to impurities and to the reactor surface. This sensitivity is reflected in the wide range of published values [26–29] for the rate of 1,2-dichloroethane cracking (Table 4).

Table 4. Published rate data for 1,2-dichloroethane cracking

log A	E, kJ/mol	$k \times 10^4$, s^{-1} (450 °C)	Reference
7.06	115.1	524	[26]
7.52	121.8	512	[27]
10.81	196.8	3.9	[28]
13.35	222.3	21.0	[29]

Byproducts arise from reactions of the olefin produced [30]: the amounts of such byproducts are therefore tied to the reaction conversion. The higher the conversion, the greater are the quantities of byproducts formed.

3.4. Reactor Design and Operating Conditions

Reactor design and operating conditions are usually a compromise between conflicting factors. Some of the factors that influence design are as follows.

The reactor is only one unit (albeit the most important one) in a plant. The conditions under which it operates are therefore influenced by the design philosophy adopted for the entire process. In this type of process the pressure at which the reactor operates is determined by the need to distill HCl from vinyl chloride. To achieve this without excessive refrigeration, large-scale reactors operate (typically) at ca. 13 bar.

The reaction temperature is then chosen to give the required conversion in a "sensible" residence time (typically of the order of 10 s).

The most important factor determining the shape of these reactors is the need to achieve high rates of heat transfer to the reactant gas, which is essential because these reactions are highly endothermic. In practice, this means a narrow tube that allows high gas velocities with high rates of heat transfer.

The reactor cross section cannot, however, be too narrow because of soot and coke formation, which covers the walls and causes significant pressure drops over a period of time. Reactor dimensions are therefore a compromise between conflicting factors.

3.5. Reactor Modeling

The overall reaction rate depends on the inner surface of the reactor (see Section 3.3) and on the purity of the feedstock. Therefore, reaction rates cannot be predicted from published rate data. Rates must be measured for the specific construction material of the reactor and the reactor feedstock. In any reactor model, the overall reaction rate must be treated empirically. In

practice this means that an initiation rate is incorporated into the model that fits the measured rate over as wide as possible a range of reactor conditions.

However, the amounts of byproduct formed are to a first approximation independent of the overall reaction rate: they depend only on the degree of conversion. The basic chemistry of these byproduct-forming reactions is now well understood [30]. Sufficient published data on radical reaction rates are available [31–33] to allow comprehensive reactor models to be constructed.

Such models should be designed with the aim clearly in view. Thus, constructing an elaborate chemical model that incorporates detailed byproduct patterns has little point when the aim of the work is the estimation of heat fluxes and reactor wall temperatures.

Detailed and accurate byproduct patterns can be predicted for a wide range of reactor and operating conditions. This allows an economic optimization of the reactor and reactor operating conditions to be made.

As an example of the type of information that a detailed reactor model can produce, a comparison of total byproduct formation from 1,2-dichloroethane cracking and from 1,1,2,2-tetrachloroethane cracking is given in Figure 5.

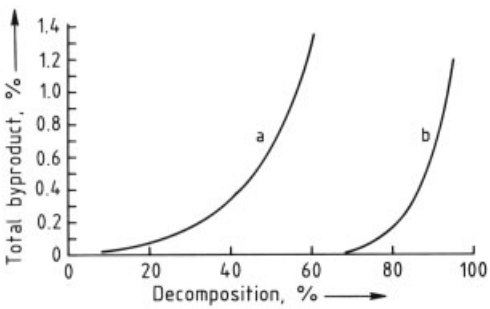

Figure 5. Byproduct formation in cracking of 1,2-dichloroethane and 1,1,2,2-tetrachloroethane
a) 1,2-Dichloroethane; b) 1,1,2,2-Tetrachloroethane

A large difference exists in the predicted total amount of byproduct formation as a function of the degree of conversion. This difference is found in practice and explains the fact that reactors in which 1,2-dichloroethane is cracked operate at around 50 % conversion with 1 % byproducts, whereas similar reactors for 1,1,2,2-tetrachloroethane crackers run at 95 % conversion with the same quantity of byproducts.

3.6. Reactor Coking and Carbon Formation

A major problem in the operation of 1,2-dichloroethane cracking reactors is the laydown of carbonaceous material on the reactor wall, which gradually blocks the reactor. At a certain point a pressure drop builds up across the reactor, leading to plant shutdown and reactor decoke.

The rate at which coke is layed down is tied closely to the degree of cracking. The higher the degree of cracking, the faster is the rate of coking. In practice, with 1,2-dichloroethane reactors, plant operation is dominated by the requirement to keep the degree of cracking to approximately 50 %. Operation at higher values leads to unacceptably high rates of coking.

Reactor coking thus has a major economic impact on the process. The limitation of only 50 % conversion per pass results in large process recycles, with concomitantly high energy requirements.

Various methods of reactor decoking have been employed, including steaming out, burning, and shot-blasting with steel pellets.

4. Gas-Phase Halogenation Processes

4.1. Reaction Description

Gas-phase halogenation reactions cover a wide range of processes, many of which employ tubular reactors. Three examples of such reactions that are used on an industrial scale are

1) Substitutive chlorination
   $Cl_2 + RH \longrightarrow RCl + HCl$
   where RH is an alkane or substituted alkane, and RCl an alkyl chloride
2) Substitutive bromination
   $Br_2 + RH \longrightarrow RBr + HBr$
   where RH is an alkane or substituted alkane, and RBr an alkyl bromide
3) Perhalogenation
   $7\ Cl_2 + C_3H_6 \longrightarrow CCl_4 + C_2Cl_4 + 6\ HCl$

On the face of it, these reactions are very similar, but their different characteristics result in significant differences in reactor design and operation.

4.2. Substitutive Chlorination Reactors

The classic example of substitutive chlorination reactions is the chlorination of methane [*74-82-8*] or monochloromethane [*74-87-3*] to produce dichloromethane [*75-09-2*], trichloromethane [*67-66-3*], and tetrachloromethane [*56-23-5*] [34–36] (→ Chlorinated Hydrocarbons, Chap. 1.3.3.). This is a stepwise radical chain reaction, with the following overall steps:

$$CH_4 + Cl_2 \longrightarrow CH_3Cl + HCl$$

$$CH_3Cl + Cl_2 \longrightarrow CH_2Cl_2 + HCl$$

$$CH_2Cl_2 + Cl_2 \longrightarrow CHCl_3 + HCl$$

$$CHCl_3 + Cl_2 \longrightarrow CCl_4 + HCl$$

These reactions are very exothermic (ca. 100 kJ of heat evolved per mole of Cl_2 reacted). The overall reaction rate is sufficiently rapid that once initiated, the reaction must be treated as adiabatic. The rate of heat generation is much too fast to be controlled by heat exchangers.

A variety of different reactor designs have been proposed to meet this requirement. These include injection of liquid, using evaporative cooling to control the temperature [37], and addition of chlorine along the length of a gas-permeable tubular reactor to spread the heat load down the length of the reactor [38].

However, most commercial reactors employ one of two different approaches. In the first, chlorine is fed to a train of reactors in series, with cooling of the reacted mixture between reactors before the addition of more chlorine. In the second, a single reactor is used with recycle of unreacted feedstock or HCl to maintain the heat balance.

Within these two approaches, significant variations can also occur in the detailed design of the reactor. Two commercial designs that give different spectra of products and have different operating stabilities are shown in Figure 6.

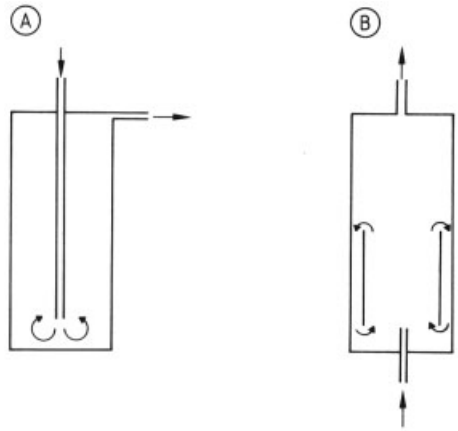

Figure 6. Examples of reactor design used in chloromethane processes
For further explanation, see text.

The crucial difference between these reactor designs is in the gas mixing in the reaction zone. Thus the design shown in Figure 6 A has a well-mixed zone at the bottom of the reactor that effectively acts as a continuously stirred-tank reactor (CSTR); the major section of the reactor (80 % of the total volume) acts as a plug-flow region in which the reaction goes to completion.

In the other reactor design (Fig. 6 B), gas fed into the reactor reacts in the expanding jet, with a well-controlled proportion of back-mixing. As a consequence, these reactors behave more like plug-flow tubular reactors than like CSTR's. This can be seen both in the product pattern that can be achieved and in byproduct formation.

These reactors can be considered hybrids between back-mixed and tubular – plug flow. An interesting characteristic is that the product pattern is dependent on the absolute size of the reactor and on the reaction rate. In small reactors the reaction of back-mixed gases is far from complete and the reactors behave as CSTR's (Fig. 6 A). In large commercial reactors (of the type shown in Fig. 6 B) the reaction in the back-mixed gases is complete before the gas is entrained into the reacting jet; in these circumstances, the reactor behaves as expected for a plug-flow system.

4.2.1. Reactor Modeling

Substitutive chlorination reactions are gas-phase free-radical processes. Apart from the initiating reaction(s), the detailed chemical reaction

mechanism is well known [34–36], [39]. Reactor models can therefore be constructed that accurately reflect plant performance over a wide range of potential operating conditions. Such models can be used to optimize plant operation and to explore problems with reactor stability.

4.2.2. Reactor Coking and Soot Formation

Two distinct carbon-forming processes are encountered with substitutive chlorination reactors. The first is a slow laydown of very hard sheets of carbonaceous material on all internal reactor surfaces. This problem is not significant, and the carbon can be removed during routine plant shutdowns.

The second type of carbon formation can, however, be more serious. Conditions can be reached at which thermal breakdown of the chlorinated product occurs, leading to the formation of large quantities of highly mobile, fluffy soot that can be deposited in downstream equipment. This problem is more likely to occur at high reactor temperatures, at high depths of chlorination, in reactors that operate as well-stirred systems (e.g., Fig. 6 A).

4.3. Substitutive Bromination Reactors

The only known commercial examples of substitutive bromination reactors are used in the manufacture of fire-extinguishing agents; for example, the reaction

$$CF_2HCl + Br_2 \longrightarrow CF_2BrCl + HBr$$

is carried out in tubular reactors at 480 – 550 °C and up to 3 bar [40], [41].

Such reactions are close to thermoneutral ($\Delta H = -10$ to $+15$ kJ/mol). They are equilibrium limited, with maximum conversions of the order of 50 %.

The reaction occurs via a thermally initiated free-radical mechanism and is very similar to the equivalent reaction with chlorine.

4.3.1. Reactor Design

Because substitutive bromination is almost thermoneutral, reactor design is simple. A tubular reactor with sufficient volume to provide the residence time at the reaction temperature is all that is required.

4.3.2. Reactor Modeling

The chemistry of substitutive bromination reactions is very well understood [42]. Reactor models for this type of process provide an accurate and reliable guide to plant performance; they can be used confidently for reactor design and process optimization.

4.3.3. Reactor Coking and Carbon Formation

Carbon formation is not a problem with substitutive bromination. The crucial difference between this class of reaction and the corresponding chlorination processes discussed in Section 4.2 is that bromination reactions are carried out in an excess of halogen, which inhibits coking and carbon formation. Reaction conversions are limited by thermodynamic equilibria, so that even where the feedstock alkane : Br_2 ratio is > 1, free bromine is present in significant concentrations throughout the reactor.

4.4. Perhalogenation Reactors

4.4.1. Reaction Description

Perhalogenation is widely used on a large scale for the production of tetrachloromethane and tetrachloroethylene. A typical reaction is as follows (→ Chlorinated Hydrocarbons):

$$C_3H_6 + 7\,Cl_2 \longrightarrow CCl_4 + C_2Cl_4 + 6\,HCl$$

The process is operated at 560 – 620 °C and ca. 3 bar. The reaction is very exothermic (for the case quoted above, $\Delta H = -695$ kJ per mole of propene reacted).

The process is not limited to propene as feed. Reactor temperatures are such that a wide range of C_1, C_2, and C_3 feedstocks can be used. In particular, chlorinated residues from vinyl chloride manufacture and similar waste streams, are suitable. The reaction can be used to produce exclusively CCl_4 or C_2Cl_4 or any mixture of the two. This is achieved by recycle of the undesirable product.

4.4.2. Reactor Design and Operation

The most important factor determining reactor design in perhalogenation is the high heat of reaction. The conventional method of heat dissipation is to inject liquid with the reactor feed, using the latent heat of vaporization to evaporate the liquid and maintain the required reactor temperature. This is very similar to the reactors described in Section 4.2. The liquid used is commonly a $CCl_4 - C_2Cl_4$ mixture, which forms as a primary condensate in the process. The reactor is designed around the need to interconvert C_1 and C_2 species via the following equilibria:

$$C_2Cl_4 + Cl_2 \rightleftharpoons C_2Cl_6$$

$$C_2Cl_6 + Cl_2 \rightleftharpoons 2\ CCl_4$$

The required product split is achieved by recycling the unwanted product and by varying the chlorine excess and the reactor pressure.

Hoechst has developed an interesting variant of the perchlorination process that uses a tubular reactor. It operates up to 620 °C and 150 bar [43]. The reaction is carried out in an excess of chlorine. The adiabatic temperature increase is controlled by the composition of the organic feed.

Under these conditions, even intractable chlorinated residues such as hexachlorobenzene react to form tetrachloromethane. A sophisticated reactor design has been developed to meet the stringent temperature and pressure requirements [44], [45].

4.4.3. Reactor Modeling

The perhalogenation reaction is a thermally initiated radical chain process. The first stages of this reaction, namely, the conversion of feedstock to a $CCl_4 - C_2Cl_4$ mixture, is extremely fast and is complete within a small fraction of the reactor residence time. The chemistry of these reactions is complex, and published rate data under reactor conditions are sparse. Development of a reactor model for this part of the reaction would be a major undertaking.

The second stage of the reaction, the interconversion of CCl_4 and C_2Cl_4, is straightforward. A model of the second stage can be used to optimize the reactor for different feedstock and product requirements, and to minimize formation of byproducts.

4.4.4. Reactor Coking and Carbon Formation

Reactor coking is not a problem with perhalogenation. The formation of "fluffy" carbon can occur under two sets of conditions. The most common problem occurs when C_2Cl_4 is being made. For production of C_2Cl_4, a low chlorine excess is required; carbon is formed if the chlorine excess drops too far. Carbon can also be formed during both CCl_4 and C_2Cl_4 production if the feeds are not mixed effectively before injection into the reactor.

5. Fundamental Modeling Approaches – An Evaluation

Tubular homogeneous reactors are designed, optimized, and operated according to proprietary empirical models that have been developed over many years of accumulated operating data. These data provide sufficiently accurate means of interpolating between operating conditions and feedstock changes that have been previously established.

So-called *fundamental models* attempt to provide a mechanistic description of the reaction chemistry, incorporating molecular, free-radical, and surface species. Such models often encounter many hundreds of reactions with very different magnitudes of rate constants. Integration is time consuming and has required the development of special numerical techniques capable of treating a large system of stiff differential equations. The development and testing of fundamental models are daunting tasks for all but the simplest feedstocks, yet the intellectual challenges and technological prizes remain undimmed by time. In particular, the development of new types of reactor, identification of new operating regions, and use of cheaper feedstock can all benefit from a fundamental modeling approach.

5.1. Thermal Decomposition of Ethane

Fundamental kinetic data for ethane decomposition are provided by Sundaram and Froment

[2]. A subset of 22 reactions is selected that includes the major molecular species isolated at high temperature and ultrashort reaction time (< 10 ms), namely, CH_4, C_2H_2, C_2H_4, C_2H_6, C_3H_6, C_3H_8, and H_2. The reactions and their kinetic constants are given in Table 5. Some of the reverse reaction rate parameters are not tabulated by SUNDARAM and FROMENT. They are estimated in the kinetics program from forward rates and equilibrium data.

The experimental reactor consisted of an inductively heated graphite cylinder containing several fine longitudinal flow channels down which ethane was passed. The cracked gas was rapidly quenched at the reactor exit by contact with a cold N_2 stream. Uniform skin temperatures could be obtained with this heating arrangement, although the flow channels were too narrow to permit measurement of gas temperature. Pure ethane at 1 bar was used as feed. Carbon and hydrogen atom balances were brought to within 2 and 3 %, respectively.

The mathematical model assumes plug flow, isothermal reaction, and instantaneous quench; it accounts automatically for volume expansion. Since gas temperature cannot be measured, it is adjusted within the model until a match is obtained between the predicted and the observed ethane conversion. Estimated gas temperatures and residence times are shown across the top of Figure 7, which displays the predicted yields of C_2H_4, C_2H_2, CH_4, and H_2 versus ethane conversion. Also shown are the observed values.

Clearly, the predicted results are in very poor agreement with observations. The model predicts a continuously increasing yield of C_2H_4, whereas the results observed pass through a maximum. The model predicts virtually no acetylene formation, whereas more than 12 wt % is observed at the highest ethane conversions. The yield of CH_4 is also vastly underpredicted. Only the H_2 yields are in reasonable agreement.

The kinetics of vinyl radical ($C_2H_3^\bullet$) formation (Table 5, reactions 9, 11, 13, 16) and destruction (reactions 10, 12, 14, 15) in this model appear to be seriously in error; furthermore, an insufficient number of methane-forming reactions may be included in the model.

Interestingly, the model's predictions are not too bad up to an ethane conversion of 60 %, which is within the commercial range. This leaves the impression that a certain degree of tuning has been built into this so-called fundamental model to render it consistent with plant data.

5.2. Improving the Model

OLSON and coworkers studied the high-temperature thermal decomposition of ethane [46]. Their work indicates that the kinetics of vinyl radical formation and destruction are much faster than those estimated from the Sundaram and Froment model. At 1000 °C, for example, the rate of formation of $C_2H_3^\bullet$ from reaction 9 of Table 5 is a factor of 14 faster, whereas the rate of disappearance of $C_2H_3^\bullet$ by reaction 15 is enhanced by 4×10^5.

The model was rerun with the improved parameters given in Table 6. The predicted and observed yields are compared in Figure 8.

A dramatic improvement in prediction of the C_2H_2 yield is clear, and the predicted H_2 yields agree closely with observed values. The predicted C_2H_4 yield now passes through a maximum, although the peak value is 15 % overestimated and occurs at too high an ethane conversion in comparison with the observed value. Methane yields are similar to those predicted previously, as expected, and remain much lower than observed values.

The model seems to require still further improvement, particularly with regard to the key methane-forming reaction

$$CH_3^\bullet + C_2H_4 \longrightarrow CH_4 + C_2H_3^\bullet$$

The additional reaction

$$CH_3^\bullet + H_2 \longrightarrow CH_4 + H^\bullet$$

might also be included. This reaction is known to possess a highly curved Arrhenius plot [47] and to become increasingly important at higher temperature.

However, wide tuning of the parameters in the fundamental model beyond their uncertainty limits, merely to obtain agreement with practice, would be inappropriate.

This study demonstrates the need for further research on the chemistry and mechanism of cracking, even for one of the simplest hydrocarbon molecules, ethane, before truly fundamental predictions can be claimed.

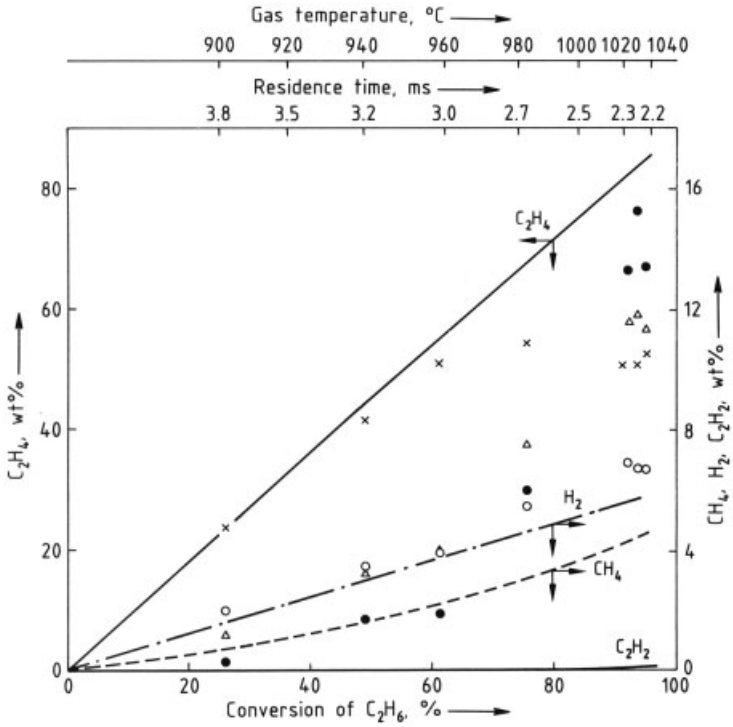

Figure 7. Comparison of product yields of C_2H_4, C_2H_2, CH_4, and H_2 as a function of C_2H_6 conversion between Sundaram and Froment kinetic model and experiment
$\times = C_2H_4$; $\triangle = CH_4$; $\bullet = C_2H_2$; $\circ = H_2$

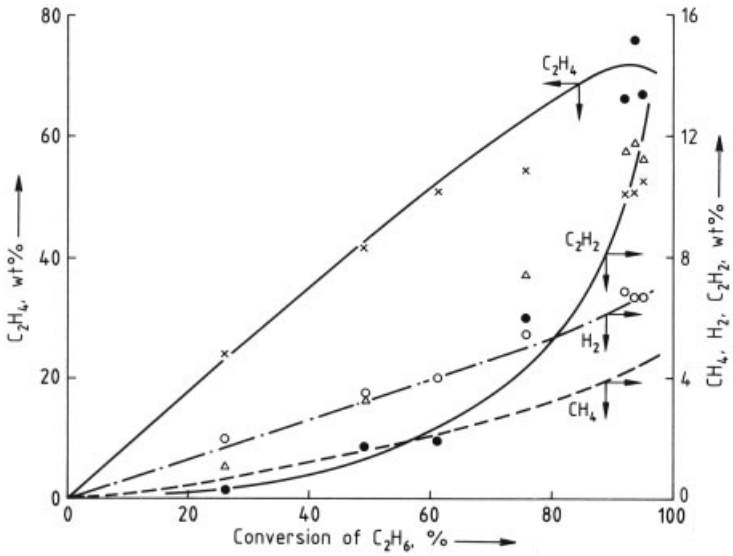

Figure 8. Comparison of product yields of C_2H_4, C_2H_2, CH_4, and H_2 as a function of C_2H_6 conversion between modified fundamental model and experiment
$\times = C_2H_4$; $\triangle = CH_4$; $\bullet = C_2H_2$; $\circ = H_2$

Table 5. Subset of reactions from SUNDARAM and FROMENT [2]

No.	Reaction	k_∞, s^{-1} or $m^3\ kmol^{-1}\ s^{-1}$	E, J/kmol
1.	$C_2H_6 \longrightarrow 2\,CH_3^\bullet$	3.98×10^{16}	3.664×10^8
2.	$2\,CH_3^\bullet \longrightarrow C_2H_6$	1.29×10^{10}	0
3.	$CH_3^\bullet + C_2H_6 \longrightarrow CH_4 + C_2H_5^\bullet$	3.8×10^{11}	6.908×10^7
4.	$C_2H_5^\bullet + CH_4 \longrightarrow C_2H_6 + CH_3^\bullet$	calculated within program	
5.	$C_2H_5^\bullet \longrightarrow C_2H_4 + H^\bullet$	3.24×10^{13}	1.675×10^8
6.	$H^\bullet + C_2H_4 \longrightarrow C_2H_5^\bullet$	1.0×10^{10}	6.280×10^6
7.	$H^\bullet + C_2H_6 \longrightarrow H_2 + C_2H_5^\bullet$	1.0×10^{11}	4.061×10^7
8.	$C_2H_5^\bullet + H_2 \longrightarrow C_2H_6 + H^\bullet$	calculated within program	
9.	$H^\bullet + C_2H_4 \longrightarrow H_2 + C_2H_3^\bullet$	7.94×10^8	1.675×10^7
10.	$C_2H_3^\bullet + H_2 \longrightarrow C_2H_4 + H^\bullet$	calculated within program	
11.	$CH_3^\bullet + C_2H_4 \longrightarrow CH_4 + C_2H_3^\bullet$	1.0×10^{10}	5.443×10^7
12.	$C_2H_3^\bullet + CH_4 \longrightarrow C_2H_4 + CH_3^\bullet$	calculated within program	
13.	$2\,C_2H_4 \longrightarrow C_2H_3^\bullet + C_2H_5^\bullet$	calculated within program	
14.	$C_2H_3^\bullet + C_2H_5^\bullet \longrightarrow 2\,C_2H_4$	3.16×10^9	0
15.	$C_2H_3^\bullet \longrightarrow C_2H_2 + H^\bullet$	2.0×10^9	1.318×10^8
16.	$H^\bullet + C_2H_2 \longrightarrow C_2H_3^\bullet$	3.98×10^{10}	5.443×10^6
17.	$2\,C_2H_5^\bullet \longrightarrow C_2H_6 + C_2H_4$	2.51×10^{10}	0
18.	$CH_3^\bullet + C_2H_5^\bullet \longrightarrow C_3H_8$	3.2×10^9	0
19.	$CH_3^\bullet + C_2H_4 \longrightarrow 1\text{-}C_3H_7^{\bullet\,a}$	1.0×10^9	8.374×10^6
20.	$1\text{-}C_3H_7^{\bullet\,a} \longrightarrow CH_3^\bullet + C_2H_4$	2.88×10^{14}	1.267×10^8
21.	$1\text{-}C_3H_7^{\bullet\,a} \longrightarrow C_3H_6 + H^\bullet$	6.31×10^{13}	1.591×10^8
22.	$C_3H_6 + H^\bullet \longrightarrow 1\text{-}C_3H_7^{\bullet\,a}$	3.63×10^{10}	1.0×10^7

[a] Primary propyl radical.

Table 6. Improved values for the Sundaram and Froment model according to [46]

No.	Reaction	k_∞, s^{-1} or $m^3\ kmol^{-1}\ s^{-1}$	E, J/kmol
9.	$H^\bullet + C_2H_4 \longrightarrow H_2 + C_2H_3^\bullet$	6.9×10^{11}	6.071×10^7
15.	$C_2H_3^\bullet \longrightarrow C_2H_2 + H^\bullet$	7.9×10^{14}	1.319×10^8

On a closing practical note, an increase in C_2H_4 yield from 47 % to a maximum of 55 % is observed when ethane conversion is increased from the current 60 % to 80 % by raising the gas temperature. The overall benefits of this require careful analysis and the fate of C_2H_2 would be a key component of this analysis.

6. Conclusions

In conclusion, some of the strengths and limitations of tubular reactors that have emerged in this article are listed below:

Strengths	Limitations
Plug flow	coking
High heat-transfer rates	temperature limits
Flexibility to different feedstocks	low capacity
	limited production flexibility for a given feedstock

Thus far, no one has abandoned the conventional tubular reactor in a full-scale plant. Its strengths, therefore, remain dominant. Progress in materials offers the prospect of reducing some of the more important disadvantages, such as tendency to coking and temperature limitations, which may make displacement of the tubular reactor even more difficult.

7. References

1. F. O. Rice, F. K. Hertzfeld, *J. Am. Chem. Soc.* **56** (1934) 284.
2. K. M. Sundaram, G. F. Froment, *Ind. Eng. Chem. Fundam.* **17** (1978) 174 – 182.
3. M. E. Dente, E. M. Ranzi: "Pyrolysis" in L. F. Albright, W. L. Crynes, W. H. Corcoran (eds): *Pyrolysis: Industrial Theory & Practice,* Academic Press, New York 1983, chap. 7.
4. ICI Ltd., GB 2 535 927, 1974 (P. Dyer, D. C. A. Waterman).
5. A. J. MacNab, *Hydrocarbon Process.* **66** (1987) no. 12, 43 – 45.
6. M. J. Maddock, R. K. Dorn, *Oil Gas J.* **69** (1971) Feb., 65 – 69.
7. J. M. Fernandez-Baujun, S. M. Solomon, *ACS Symp. Ser.* **32** (1976) 345 – 372.
8. Stone & Webster Engineering, FR 1 519 255, 1966 (H. P. Hallee, E. J. Green, H. N. Woebcke).
9. J. Dluzniewski, R. Orriss, J. Wallace, *Oil Gas J.* **79** (1981) April 6, 121.
10. J. Barton, *Oil Gas J.* **88** (1990) Jan. 29, 81 – 84.
11. S. M. Roberts, C. G. Lapse, *Ind. Eng. Chem.* **53** (1961) 343 – 348.
12. P. V. Youle, L. A. Duncanson, *Process. Eng. (London)* **51** (1970) 49 – 53.
13. M. L. Sourander et al., *Hydrocarbon Process.* **63** (1984) no. 6 63 – 69.
14. L. F. Albright, *Oil Gas J.* **86** (1988) Aug. 1, 35; *Oil Gas J.* **86** (1988) Aug. 15, 69; *Oil Gas J.* **86** (1988) Aug. 29, 44.
15. T. Sato et al., *CEER Chem. Econ. Eng. Rev.* **4** (1972) no. 7, 14 – 23.
16. BASF, DE 3 010 000, 1980 (A. Sliwka).
17. British Petroleum, US 4 099 990, 1975 (D. E. Brown, J. T. K. Clark, J. J. McCarroll, M. L. Sims). L. F. Albright, W. A. McGill, *Oil Gas J.* **85** (1987) Aug. 31, 46.
18. Toyo Engineering Kubota, EP 258 907, 1981 (J. Sugitani, K. Tsuchida, T. Shinohara, K. Shibata).
19. S. Vaish, D. Kunzru, *Ind. Eng. Chem. Res.* **28** (1989) 1293 – 1299. Nalco Chemical, US 4 542 253, 1984 (M. Kaplan, W. A. Linkley).
20. Betz Europe Inc., EP 0 207 745, 1985 (D. K. Reid, D. R. Forester). Betz Europe Inc., EP 0 219 960, 1985 (D. K. Reid, D. R. Forester).
21. IG Farben, US 1 847 095, 1926 (A. Mittasch, R. Wietzel, C. Pfaunler). Phillips Petroleum Co., US 4 551 227, 1984 (L. E. Reed, R. A. Porter).
22. R. L. Baldwin, G. R. Kamm, *Hydrocarbon Process.* **61** (1982) no. 11, 127 – 130.
23. R. J. Gartside, A. F. Ellis, *Chem. Eng. Prog.* **79** (1983) 82 – 85.
24. SRI International, Chem. Econ. Handbook, Marketing Report on Vinyl Chloride, Menlo Park, Calif., 1990.
25. SRI International, Chem. Econ. Handbook, Report on C2 Chlorinated Solvents, Menlo Park, Calif., 1988.
26. K. Kitabatake, T. Onouchi, *Kogyo Kagaku Zasshi* **65** (1962) 931.
27. T. Takahashi, *Kogyo Kagaku Zasshi* **71** (1968) 501.
28. D. H. R. Barton, K. E. Howlett, *J. Chem. Soc.* 1949, 155.
29. P. Goldfinger, G. Martens, M. Delaet : *Inhibition in Free Radical Reactions: The Pyrolysis of 1,2-Dichloroethane,* scientific note no. 2, US Army Dept. (Code no. DA-91-591-EUC-3200).
30. M. Schneider, J. Wolfrum, *Ber. Bunsenges. Phys. Chem.* **90** (1986) 1058 – 1062.
31. J. A. Kerr (ed.): *Handbook of Bimolecular and Termolecular Gas Reactions,* **vols. I–III,** CRC Press, Boca Raton., Fla., 1981 and 1987.
32. J. A. Kerr, M. J. Parsonage: *Evaluated Kinetic Data on Gas Phase Addition Reactions,* Butterworths, London 1972.
33. S. W. Benson, H. E. O'Neal: *Kinetic Data on Gas Phase Unimolecular Reactions,* National Standards Reference Data Series NSRDS-NBS 21, 1970.
34. H. B. Hass, E. T. McBee, *Ind. Eng. Chem.* **34** (1942) 296 – 300.
35. W. Hirschkind, *Ind. Eng. Chem.* **41** (1949) 2749 – 2752.
36. B. E. Kurtz, *Ind. Eng. Chem. Process Des. Dev.* **11** (1972) 332 – 338.
37. Dow Chemical, WO 90/12 774, 1989 (D. H. West, L. A. Herbert, R. L. Bowlin, M. Tholbrook).
38. Allied Corporation, US 4 513 154, 1980 (B. E. Kurtz).
39. M. A. A. Clyne, R. F. Walker, *J. Chem. Soc. Farad. Trans. 1* **69** (1973) 1547.
40. Kali-Chemie, DE 1 946 509, 1969 (J. Massone).
41. Firestone, US 2 731 505, 1956 (W. S. Barnhart).
42. J. C. Amphlett, E. Whittle, *Trans. Faraday Soc.* **64** (1968) 2130 – 2142.
43. H. Krekeler, W. Riemenschneider, *Chem. Ing. Tech.* **45** (1973) 1019 – 1021.

44. Hoechst, DE 1 918 270, 1969 (H. Krekeler, H. Meidert, W. Riemenschneider).
45. Hoechst, DE 2 218 414, 1972 (H. Krekeler, H. Osterbrink, W. Riemenschneider).
46. D. B. Olson, T. Tanzawa, W. C. Gardiner, Jr., *Int. J. Chem. Kinet.* **11** (1979) 23.
47. T. C. Clark, J. E. Dove, *Can. J. Chem.* **51** (1973) 2155.

Fixed-Bed Reactors

For other industrial reactors and their applications, see → Stirred-Tank and Loop Reactors, → Tubular Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

GERHART EIGENBERGER, Institut für Chemische Verfahrenstechnik, Universität Stuttgart, Stuttgart, Federal Republic of Germany

1. Introduction 2
2. Catalyst Forms for Fixed-Bed Reactors 4
2.1. Gas – Catalyst Mass and Heat Transfer 6
2.1.1. Random Packings 6
2.1.2. Monolith Structures 7
2.2. Flow and Pressure Loss in Fixed Beds 8
2.3. Heat Transport Transverse to the Flow Direction 9
2.4. Comparison and Evaluation of Different Catalyst Forms 9
2.4.1. Catalyst Forms for Adiabatic Operation 10
2.4.2. Catalyst Forms for Isothermal Operation 11
3. Adiabatic Reaction Control 13
4. Reaction Control with Supply or Removal of Heat in the Reactor 15
4.1. Introduction 15
4.2. Adiabatic Multistage Reactors with Interstage Heat Transfer 17
4.3. Reactors with Heat Exchange Integrated in the Fixed Bed 19
4.3.1. Heat Transfer Media for Fixed-Bed Reactors 19
4.3.2. Reactor Designs 22
4.3.3. Influencing the Course of the Reaction 24
5. Autothermal Reaction Control . . . 27
5.1. Introduction 27
5.2. Reaction Control with Periodic Flow Reversal 28
5.3. Reactor Behavior 29
6. Stability, Dynamics, and Control of Industrial Fixed-Bed Reactors . . . 32
6.1. Introduction 32
6.2. Parametric Sensitivity 32
6.3. Multiple Stationary States 34
6.4. Migrating Reaction Zones 36
6.5. Safety Aspects 38
6.6. Control of Fixed-Bed Reactors . . . 39
7. References 40

Symbols (see also → Principles of Chemical Reaction Engineering and → Model Reactors and Their Design Equations)

a_p specific outer surface area of catalyst, m^2/m^3 packing
A flow cross-sectional area, m^2
c molar concentration, mol/m^3
c_0 feed concentration, mol/m^3
c_{pG} specific heat of reaction gas, $kJ\,kg^{-1}\,K^{-1}$
c_s specific heat of catalyst, $kJ\,kg^{-1}\,K^{-1}$
D diffusion coefficient, m^2/s
d_p catalyst particle diameter, m
d_h hydraulic diameter ($d_h = 4A/U$), m
$f_{1,2}$ Ergun coefficients (Eqs. 2.4, 2.5)
G_z mass-flow velocity, $kg\,m^{-2}\,s^{-1}$
k_G gas – catalyst mass-transfer coefficient, m/s
L length of catalyst bed, m
$\dot{N}$ energy required for coolant circulation, kW
Nu Nusselt number
$\dot{Q}$ heat flow to be transferred, kW
Re Reynolds number
Sc Schmidt number
Sh Sherwood number (Eq. 2.10)
T temperature, K
T_0 feed gas temperature, K
U circumference, m
U_w gas – catalyst heat-transfer coefficient, $W\,m^{-2}\,K^{-1}$
v_G interparticle velocity, m/s
$\dot{V}$ volume flow rate, m^3/s
V_R reactor volume, m^3
w migration velocity of reaction zone, m/s
z space coordinate, m

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_199

Z_L reactor length, m
α_W wall heat-transfer coefficient, $W\,m^{-2}\,K^{-1}$
ε void fraction, m^3 free space/m^3 packing
η dynamic viscosity, $N\,s\,m^{-2}$
Δh_R reaction enthalpy, kJ/mol
Δp pressure drop, N/m^2
ΔT_{ad} adiabatic temperature change, K
ΔT_c cooling temperature change between inlet and outlet, K
Δx molar conversion, mol/mol
λ_{eff} effective axial thermal conductivity, $W\,m^{-1}\,K^{-1}$
λ_G molecular gas thermal conductivity, $W\,m^{-1}\,K^{-1}$
λ_r effective radial thermal conductivity, $W\,m^{-1}\,K^{-1}$
ϱ density, kg/m^3
ϱ_G gas density, kg/m^3
ϱ_S density of catalyst particle, kg/m^3
ξ pressure drop coefficient

1. Introduction

Catalytic fixed-bed reactors are the most important type of reactor for the synthesis of large-scale basic chemicals and intermediates. In these reactors, the reaction takes place in the form of a heterogeneously catalyzed gas reaction on the surface of catalysts that are arranged as a so-called fixed bed in the reactor. In addition to the synthesis of valuable chemicals, fixed-bed reactors have been increasingly used to treat harmful and toxic substances. For example, the reaction chambers used to remove nitrogen oxides from power station flue gases constitute the largest type of fixed-bed reactors as regards reactor volume and throughput, while automobile exhaust purification represents by far the most widely employed application of fixed-bed reactors.

With regard to application and construction, it is convenient to differentiate between fixed-bed reactors for adiabatic operation and those for nonadiabatic operation. Since temperature control is one of the most important methods of influencing a chemical reaction, adiabatic reactors are used only where the heat of reaction is small, or where there is only one major reaction pathway; in these cases no adverse effects on selectivity or yield due to the adiabatic temperature development are expected. The characteristic feature of adiabatic reaction control is that the catalyst is present in the form of a uniform fixed bed that is surrounded by an outer insulating jacket (Fig. 1 A). Adiabatic reactor designs are discussed in Chapter 3.

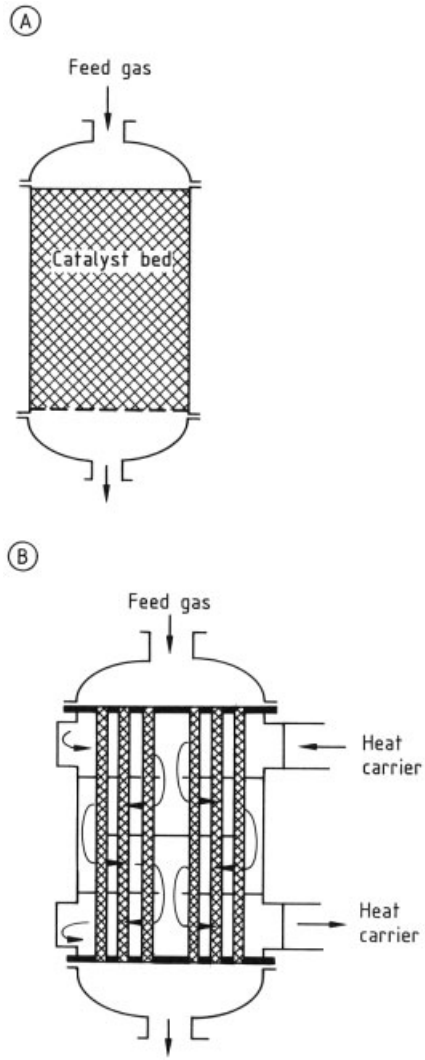

Figure 1. Basic types of catalytic fixed-bed reactors
A) Adiabatic fixed-bed reactor; B) Multitubular fixed-bed reactor

Since the incoming reaction gases in most cases must be heated to the ignition temperature of the catalytic reaction, adiabatic reaction control is often coupled with heat exchange between the incoming and exiting reaction gas resulting

in so-called autothermal reaction control. This type of reaction control offers certain specific features and development perspectives, which are discussed in Chapter 5.

Reactions with a large heat of reaction as well as reactions that are extremely temperature-sensitive are carried out in reactors in which indirect heat exchange occurs via a circulating heat transfer medium integrated in the fixed bed.

Since in most cases the task of the heat-transfer cycle is to maintain the temperature in the fixed bed within a specific range, this concept is frequently described as an "isothermal fixed-bed reactor". However, isothermal reaction control does not provide optimum selectivity or yield in all cases, and for this reason the concept of heat exchangers integrated in the fixed bed is also being increasingly used to achieve nonisothermal temperature profiles.

The most common arrangement is the multitubular fixed-bed reactor, in which the catalyst is arranged in the tubes, and the heat carrier circulates externally around the tubes (Fig. 1 B). Fixed-bed reactors with an integrated heat supply or removal are discussed in Chapter 4.

Fixed-bed reactors for industrial syntheses are generally operated in a stationary mode (i.e., under constant operating conditions) over prolonged production runs, and design therefore concentrates on achieving an optimum stationary operation. However, the nonstationary dynamic operating mode is also of great importance for industrial operation control. In particular, fixed-bed reactors with a strongly exothermic reaction exhibit an, at times, surprising operational behavior which is discussed in more detail in Chapter 6.

Within a production plant the reactor may justifiably be regarded as the central item of apparatus. However, compared to the remaining parts of the plant for preparing the feed and for separating and working-up the products, often it is by no means the largest and most cost-intensive component. In many cases the achievable conversion in the reactor is limited for thermodynamic (equilibrium) and kinetic reasons (selectivity). It is then usual to separate the material discharged from the reactor into products and unreacted feed components (see Fig. 2), which are recycled to the feedstock. This recycling procedure involves costs

1) For product separation
2) For recycle compression
3) For repeated heating and cooling of the circulating material to the reaction temperature and the temperature of the separating device
4) Due to loss of product resulting from the need to remove part of the circulating material to limit the amount of inert substances or byproducts in the recycle stream (bleed stream).

To minimize these costs it is therefore necessary to maximize the conversion in the reactor and avoid as far as possible inert accompanying substances in the reaction mixture. With irreversible reactions (e.g., partial oxidations) the trend is therefore towards a highly concentrated, approximately stoichiometric feed composition, which may occasionally be in the explosive range. The resulting problems are discussed in Chapters 4 and 6.

2. Catalyst Forms for Fixed-Bed Reactors

The heart of a fixed-bed reactor and the site of the chemical reaction is the catalyst. The processes taking place on the catalyst may formally be subdivided into the following separate steps:

1) Diffusion of the reactants from the gas space through the outer gas – particle boundary layer, macropores, and micropores
2) Chemisorption on active centers
3) Surface reactions
4) Desorption of the products
5) Back-diffusion of the products into the gas space

Since most reactions take place with a considerable heat of reaction, a corresponding heat transport is superimposed on the mass transport.

The control of the microkinetics, consisting of micropore diffusion, chemisorption, surface reaction, and desorption, is the task of the catalyst developer and is not discussed further here. If the catalyst is specified together with its microkinetic properties, then reaction conditions (feedstock concentrations, pressure, temperature profile, and residence time) can be found that lead to optimum yields. The reaction engineer must determine these conditions and ensure that they are maintained in an industrial reactor.

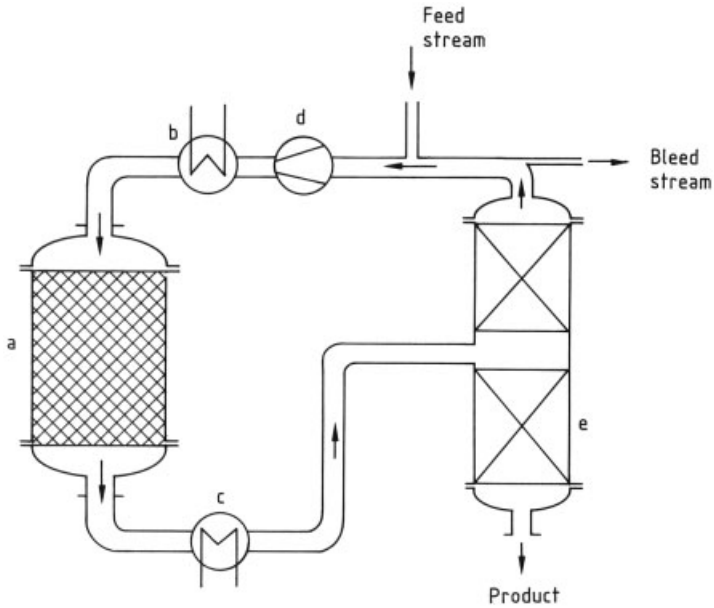

Figure 2. Reaction cycle for synthesis reactions with incomplete conversion
a) Fixed-bed reactor; b) Feed preheater; c) Exit cooler; d) Recirculation compressor; e) Separation column

In the case of selectivity-sensitive multistep reactions, any deviation from the optimum values inevitably leads to a decrease in yield. This applies to deviations from the uniform residence time distribution due to flow dispersion and flow bypass phenomena in the fixed bed, as well as to deviations from uniform reaction conditions in the catalyst as a result of mass-transport resistance in the particles and the outer boundary layer [1].

The influence of mass-transport resistance in the particles can only be excluded if the critical reaction rate is substantially lower than the mass transport velocity. This leads on the one hand to the need for good external mass transfer (i.e. to reasonable flow conditions in the packed bed), as well as to short diffusion paths in catalyst particles and sufficiently large pores. On the other hand (in the case of exothermic reactions) the local reaction rate must be controlled and limited by the packed-bed temperature.

Temperature control thus plays a predominant role in selective reaction control in general, and in particular in the case of exothermic multistep reactions. Under nonadiabatic conditions, catalysts must therefore be assembled and arranged in the fixed bed in such a way as to ensure good heat transport to the heat-transfer medium.

A further requirement placed on the catalyst is a low flow pressure loss. This applies particularly if the reaction conversion in a single throughput is low, so that the reaction has to be carried out with a large circulating gas ratio (Fig. 2), as well as to off-gas purification, in which large off-gas streams must be handled with minimal additional cost.

Finally, the catalyst should be available in a sufficiently high concentration in order to keep the construction volume of the reactor low. The decisive parameters here are the specific external catalyst surface a_p (= square meters of catalyst surface per cubic meter reactor volume) for reactions controlled by external mass transfer, the catalyst fraction $1 - \varepsilon$, where ε = cubic meters of free gas space per cubic meter of reactor volume, describing the void fraction of the fixed bed.

The above requirements are to some extent contradictory, which has led to the proposition of a large number of different catalyst forms and arrangements. However, only a few of these have proved really effective in practical operation. Suitable catalyst forms and arrangements include random packings of spheres,

solid cylinders, and hollow cylinders, as well as uniformly structured catalyst packings in the form of monoliths with parallel channels, parallel stacked plates, and crossed, corrugated-plate packets (Fig. 3).

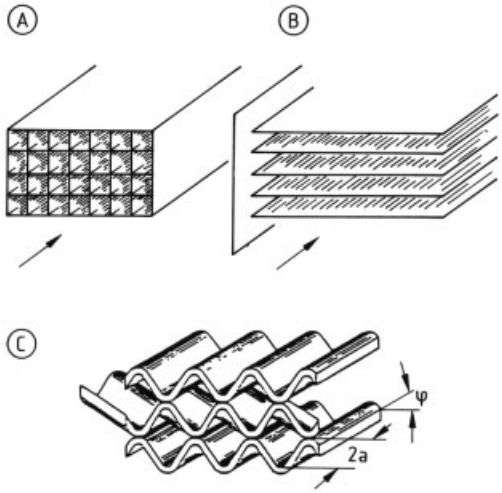

Figure 3. Usual shapes of monolith catalysts
A) Square-channel monolith; B) Parallel-plate monolith; C) Corrugated-plate monolith

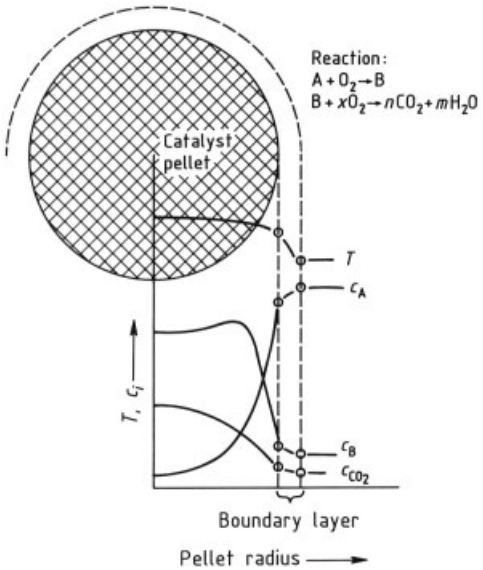

Figure 4. Temperature and concentration profiles for a partial oxidation reaction in a spherical catalyst pellet

2.1. Gas – Catalyst Mass and Heat Transfer

2.1.1. Random Packings

Industrial fixed-bed reactors are generally operated with a cross-sectional loading $G_z \geq 1$ kg gas per square meter of reactor cross section per second. This loading produces a sufficiently strong turbulence in random packing. As a result the external gas – catalyst mass-transfer resistance is small compared to the transport resistance in the catalyst pores.

However, generally the thermal conductivity of the catalyst matrix is larger than of the gas. This means that the external gas – catalyst heat transport resistance exceeds the thermal conduction resistance in the catalyst particles. The temperature and concentration conditions established in a spherical catalyst are illustrated for a partial oxidation reaction in Figure 4. The conditions can be calculated from the model equations given in → Model Reactors and Their Design Equations, Chap. 4.1.1. under the assumptions made there. An essential precondition is that the catalyst particle is uniformly exposed over its entire surface to a flow with uniform temperature and concentration. This is, of course, never the case in random packings. Figure 5 shows the local mass-transfer distribution, providing an insight into the overall highly inhomogeneous conditions. The visualization technique is based on the reaction of traces of ammonia in the gas stream with a catalyst surface impregnated with manganese chloride solution. The conversion to manganese dioxide is already so fast under ambient conditions that it practically depends only on the external mass-transfer resistance of the gas boundary layer. The intensity of the dark coloration is thus directly proportional to the local reaction rate of the surface reaction. At a constant ammonia concentration in the gas flow it is thus also proportional to the local mass transfer, and if the mass transfer and heat transfer are equivalent, also to the local heat transfer [2].

Figure 5 shows that the local conditions in random packings are much more complex than assumed in current conventional models. Nevertheless, the models for fixed beds containing a larger number of catalyst particles over the cross section can provide reliable information if it is

borne in mind that they only describe the mean value of a process that varies greatly as regards detail. For the same reason, it is inappropriate to compare model predictions with a few local temperature or concentration measurements. Indeed, a certain degree of averaging is also necessary in the measurement procedure (see Section 2.2).

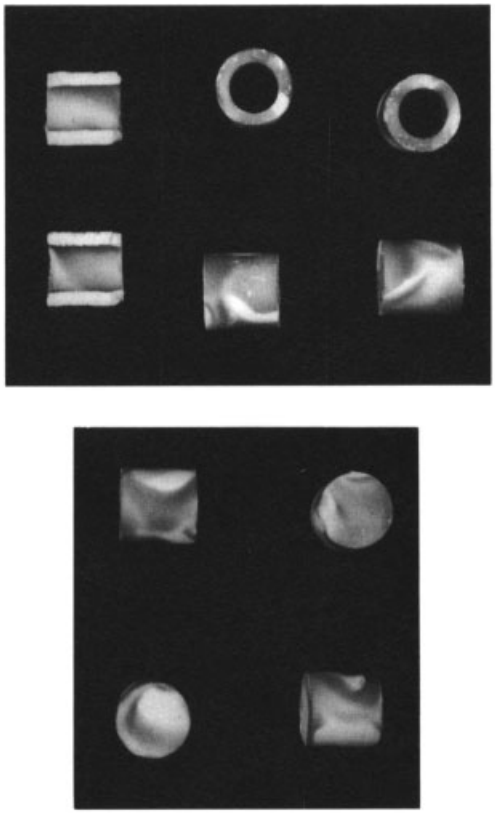

Figure 5. Local mass-transfer distribution at the surface of individual cylindrical or ring-shaped catalyst pellets in a fixed-bed packing

The literature contains a number of correlation equations for the mean gas – catalyst mass transfer and heat transfer as a function of gas properties, catalyst geometry, and flow conditions [3], [4]. However, in practice they play only a minor role for catalyst packings since design and simulation calculations are frequently performed with a model that is quasi-homogeneous, at least with regard to temperature (→ Model Reactors and Their Design Equations, Chap. 4.). The reason for this are the above-mentioned strong local fluctuations, which make differentiation between the gas temperature and catalyst temperature difficult.

2.1.2. Monolith Structures

In contrast to random packings, the external heat transfer and mass transfer in monolith catalysts is much more uniform, but they can also become limiting factors at high reaction rates. This applies in particular to channel-type monoliths with narrow parallel channels, where the flow is generally laminar under industrial operating conditions. Examples include monolith catalysts with a square-channel cross section for automobile exhaust purification and for the removal of nitrogen oxides from flue gases.

Figure 6 shows results of the visualization of the local mass transfer by using the ammonia – manganese chloride reaction (Section 2.1.1). The marked decrease in coloration in the flow direction is mainly the result of the increasing consumption of ammonia in the wall boundary layer, so that the reactants diffusing to the wall have to travel an ever increasing distance from the flow core (build up of the laminar boundary layer). The depletion is particularly pronounced in the corners, since two reaction surfaces meet here. The more acute their enclosed angle, the greater is the depletion in the corner region and the smaller is the contribution of the wall surface to further heat transfer and mass transfer. The efficiency of channel monoliths of equal cross-sectional area but different shape therefore decreases in the sequence: circle, hexagon, rectangle, triangle. This is illustrated in Table 1, which gives the asymptotic dimensionless mass-transfer and heat-transfer coefficients for tubes of the above cross sections. As Figure 7 shows, the asymptotic limiting value is established after a flow length roughly corresponding to 10 times the hydraulic channel diameter. The asymptotic limiting value from Table 1 can therefore be used to perform estimation calculations in conventional industrial monolith catalysts with a large length-to-diameter ratio. For more accurate calculations the radially and axially variable velocity, concentration, and temperature profiles must also be taken into account [7].

Table 1. Asymptotic dimensionless laminar flow heat or mass-transfer coefficients $Nu = \alpha_w \cdot d_h/\lambda_G$ (constant wall conditions) and fanning friction factor f for pressure drop $\Delta p = 2f(\eta Z_L/d_h^2) \cdot v_G$ for ducts of different cross section [6]

Geometry	*Nu*	*f*	d_h
	2.47	13.33	$2a/\sqrt{3}$
	2.98	14.23	a
	3.34	15.05	$2\sqrt{3}a$
	3.66	16.00	a
	7.54	24.00	$2a$

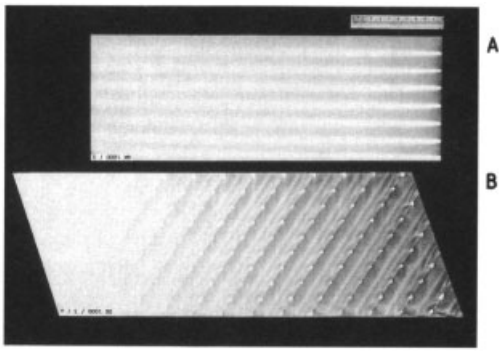

Figure 6. Visualization of mass transfer in monolith structures (flow from left to right)
A) Square duct; B) Corrugated monolith [5]

In contrast to the monolith types discussed so far, the flow conditions in corrugated-plate monoliths (Fig. 3 C) are turbulent under normal industrial velocities. The uniformity of the mass-transfer distribution [8] depends on geometrical parameters (wave form, amplitude, wavelength, angle of incidence). The transfer coefficients are considerably higher than those of laminarly traversed channel monoliths (Fig. 6 B), but the pressure loss is also high. These structures offer considerable advantages for convective heat transport transverse to the flow direction and for transverse mixing, which are discussed in more detail in the following sections.

2.2. Flow and Pressure Loss in Fixed Beds

Conventional industrial catalyst forms differ considerably as regards pressure loss. For example, for equal mean dimensions and the same proportion of empty space, random packings generally have a considerably higher pressure loss than monolith structures, and among these, corrugated structures have a higher pressure loss than monoliths with straight parallel channels. Mass transfer and heat transfer are strongly correlated with the pressure loss. For reasons of energy demand, the catalyst form for a given process is chosen to combine the required mass transport and heat transport with the lowest pressure loss. However, it should be borne in mind that the flow into the fixed-bed reactor (Fig. 1) is generally achieved by means of a feed pipe and a distribution hood. These must therefore be constructed so that the fixed bed or tube bundle is uniformly traversed, and the gas residence time in each tube or each flow filament of the fixed bed is the same. This requirement can always be met simply if the pressure loss in the fixed bed or tube bundle is substantially greater than in the entrance hood. In this case it need only be ensured that the fixed bed is packed so uniformly that no flow bypass is formed or, in the case of a multitube reactor, the pressure loss is equalized in all the individual tubes of the bundle. Because the pressure loss increases with the square of the flow velocity, a uniform distribution of the flow then occurs automatically.

Obtaining a uniform flow distribution is substantially more difficult if the pressure loss of the fixed bed is small. This holds in particular for monolith reactors with straight, laminarly traversed channels (see Section 2.3).

The Hagen – Poiseuille equation is used to calculate the pressure loss in laminarly traversed monolith channels

$$\Delta p = \frac{32\eta Z_L v_G}{d_h^2} \tag{2.1}$$

The pressure loss of packed tubes can be described either by means of a pressure loss coefficient and the pressure drop equation

$$\Delta p = \zeta \cdot \frac{\varrho}{2} \cdot v_G^2 \tag{2.2}$$

or by the Ergun equation:

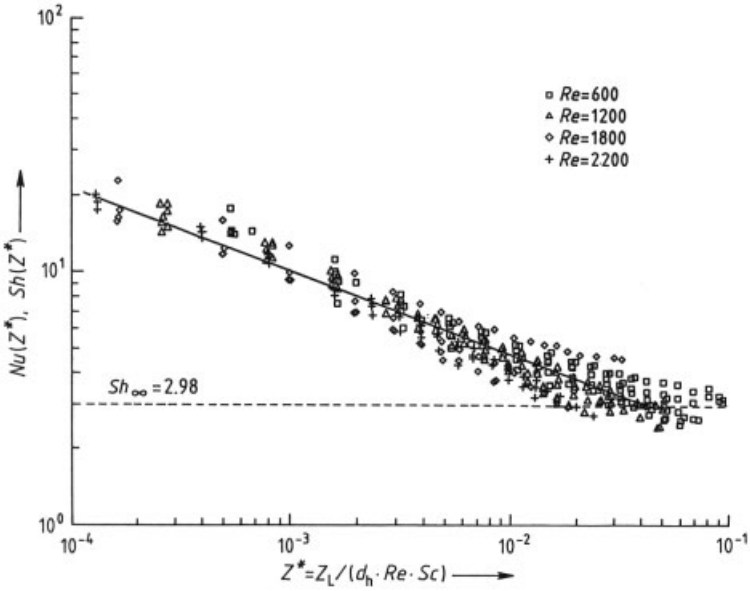

Figure 7. Behavior of the dimensionless heat-transfer and mass-transfer coefficients over the dimensionless tube length $Nu = \alpha_W \cdot d_h/\lambda_G$, $Sh = k_G \cdot d_h/D$

$$\Delta p = f_1 v_G + f_2 v_G^2 \tag{2.3}$$

where for spherical packings:

$$f_1 = 150 \cdot \eta \cdot \frac{(1-\varepsilon)^2}{\varepsilon^3 \cdot d_p^2} \tag{2.4}$$

$$f_2 = 1.75 \cdot \varrho \cdot \frac{(1-\varepsilon)}{\varepsilon^3 \cdot d_p} \tag{2.5}$$

Thus the pressure drop depends very strongly on the void fraction ε of the packing. Of the standard forms for packed catalysts, hollow cylinders of thin wall thickness ($\varepsilon \approx 0.6-0.8$) are therefore preferred to spheres ($\varepsilon \approx 0.37-0.4$) and solid cylinders ($\varepsilon \approx 0.35$).

The strong dependence of the pressure loss on the void fraction underlines the importance of packing catalyst beds carefully to avoid bypass flow due to local variations in the packing density. For the same reason the tubes of multitubular reactors for highly exothermic, selectivity-sensitive reactions are often filled uniformly by means of special devices and if necessary individually compensated for pressure loss.

The pressure loss coefficient ζ (Eq. 2.2) can be determined for typical packing forms, for example, according to [9].

2.3. Heat Transport Transverse to the Flow Direction

With nonadiabatic reaction control, heat must be transported perpendicular to the flow direction through the fixed bed to the heat exchange surfaces. At the usual mass flow rates of $G_z = 1\,\mathrm{kg\,m^{-2}\,s^{-1}}$, this heat transport takes place mainly by convection, i.e., the fixed bed must be constructed so that flow components transverse to the main flow direction occur locally. Monolith structures with straight parallel channels are thus unsuitable for nonadiabatic reaction control.

In catalyst packings transverse flow components are automatically established as a result of the nonuniform arrangement and the twisted flow around the pellets. Hollow and full cylinders with a length-to-diameter ratio of 1 to 3 are particularly effective in this respect.

Despite the fact that radial heat transport in the form of twisted flow takes place mainly by convection, it is formally described by means of a so-called effective thermal conductivity λ_r transverse to the flow direction.

In addition to heat transport through the fixed bed perpendicular to the main flow direction, the heat transport at the boundary between the fixed bed and heat exchange surface is also decisive for the heat exchange. The latter heat transport is generally described by a wall heat-transfer coefficient α_W, in which the complex interplay between convective flow at the tube wall and conduction transport by contact between the fixed bed and the heat exchange surface is described in overall terms. Heat transport in packed tubes has been investigated and discussed in detail [4], [10]. However, the correlations for α_W and λ_r given in the literature do not adequately take account of the actual velocity distribution in packed tubes. Plug flow was generally assumed, although the actual velocity profile is nonuniform with a pronounced slip at the wall. This is due to the fact that the pellets make only point contact with the wall, whereas they overlap and cross over one another inside the packing, thereby reducing the free volume and hence the velocity. The conditions for a spherical packing are illustrated schematically in Figure 8. The radially varying empty space distribution and velocity distribution must be taken into account in detailed reactor calculations, as well as in the determination of accurate heat-transport parameters [12], [13].

Existing correlation equations for calculating the heat-transport parameters were obtained from heating or cooling experiments without reactions and assuming plug flow; they therefore permit only a semiquantitative evaluation. However, this is adequate for qualitative comparison of catalyst structures.

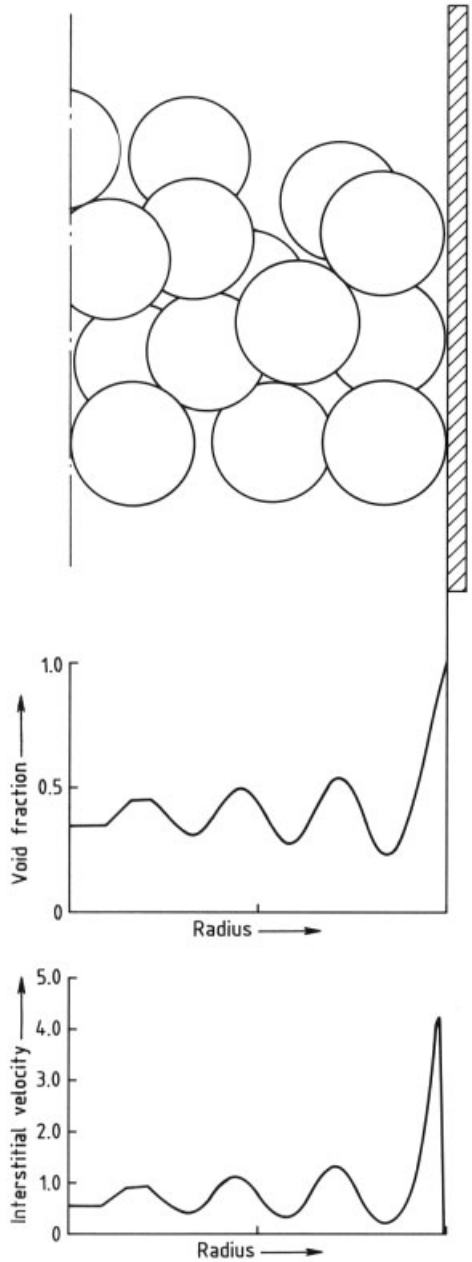

Figure 8. Radial distribution of the void fraction and axial flow velocity in a tube packed with spheres

2.4. Comparison and Evaluation of Different Catalyst Forms

The choice of a suitable catalyst form is always an optimization problem that can be completely specified only for a specific process. Even then, however, weighting the performance function is not easy since, for example, small pressure losses, uniform flow through the reactor, and good mass- and heat-transfer properties generally represent opposing requirements. The following assessment is therefore only a rough guide.

2.4.1. Catalyst Forms for Adiabatic Operation

Decisive parameters for adiabatic operation include:

1) The active catalyst surface available per unit reactor volume
2) The quality of the mass and heat transfer between the flowing gas and the active catalyst surface
3) The flow pressure loss
4) The uniformity of the flow through the reactor and thus the degree of utilization of the fixed bed

The major proportion of the active catalyst surface is located in the interior of the porous catalyst structure. However, if it is assumed that, in the case of sufficiently fast reactions, the reaction site is restricted to a thin layer underneath the external surface, then the active catalyst surface area can be taken as proportional to the specific external catalyst surface area a_{p}. If the uniform distribution of the flow entering the fixed bed is regarded as a separate problem, then for given kinetic conditions the evaluation may be restricted to three parameters: specific external surface area a_{p}, gas – solid mass-transfer coefficient k_{G}, and the flow pressure loss. The evaluation becomes particularly simple if, with a single dominant reaction, the external gas – solid mass transfer limits the reaction, since in this case the specific reaction kinetics have no influence. The relationships for a concentration c of a key component A and a simple reaction A + ... → products are outlined in [2.11]. Assuming mass-transfer limitation, the material balance for the key component along the flow path is

$$\varepsilon \cdot v_{\mathrm{G}} \cdot \frac{\partial c}{\partial z} = -k_{\mathrm{G}} \cdot a_{\mathrm{p}} \cdot c \tag{2.6}$$

If the reaction has a conversion of 99 %, i.e., an outflow concentration of 1 % of the inflow concentration, the length of the fixed bed Z_{L} is, by integration,

$$Z_{\mathrm{L}} = -\frac{\ln 0.01 \cdot \varepsilon \cdot v_{\mathrm{G}}}{k_{\mathrm{G}} \cdot a_{\mathrm{p}}} = \frac{4.6 \cdot \varepsilon \cdot v_{\mathrm{G}}}{k_{\mathrm{G}} \cdot a_{\mathrm{p}}} \tag{2.7}$$

For the required reactor volume V_{R} with the cross-sectional area A:

$$V_{\mathrm{R}} = Z_{\mathrm{L}} \cdot A = 4.6 \cdot \frac{\dot{V}}{k_{\mathrm{G}} \cdot a_{\mathrm{p}}} \tag{2.8}$$

Correlations for k_{G} can be found in [3], [4]. The computation is particularly simple for straight, laminarly traversed channels. The asymptotic limiting value given in Table 1 (Section 2.1) for the square channel

$$Sh = \frac{k_{\mathrm{G}} \cdot d_{\mathrm{h}}}{D} \approx 3 \tag{2.9}$$

gives, with $a_{\mathrm{P}} = \frac{4}{d_{\mathrm{h}} \cdot \varepsilon}$,

$$k_{\mathrm{G}} \cdot a_{\mathrm{p}} \approx \frac{12D}{d_{\mathrm{h}}^2 \cdot \varepsilon} \tag{2.10}$$

and

$$V_{\mathrm{R}} = 0.38 \frac{\dot{V} \cdot d_{\mathrm{h}}^2 \cdot \varepsilon}{D} \tag{2.11}$$

If Equation (2.1) is used for the pressure loss, then for the laminarly traversed monolith channel, from Equations (2.7) and (2.10)

$$\Delta p = \frac{12.28 \cdot \eta}{D} \cdot v_{\mathrm{G}}^2 \cdot \varepsilon^2 \tag{2.12}$$

Using the above equations, reactor variants for catalytic off-gas purification with a given volume flow $\dot{V}$ were calculated in [11]. The result is shown in Figure 9. The pressure drop of the fixed bed over the required reactor length is plotted; the curve parameter is the hydraulic diameter d_{h} of the monolith channel or packing body for different empty-tube gas velocities $v_{\mathrm{e}} = \varepsilon \cdot v_{\mathrm{G}}$. According to Equation (2.11) the required fixed bed volume V_{R} with square monolith catalysts and a given throughput $\dot{V}$ depends only on d_{h}, whereas the geometric arrangement (small bed cross section A and long bed length Z_{L}, or large bed cross section and short bed) has no influence on V_{R}. However, the bed cross-sectional area A for a given gas throughput $\dot{V}$ determines via

$$\dot{V} = A \cdot v_{\mathrm{G}} \cdot \varepsilon \tag{2.13}$$

the interparticle velocity v_{G}, and according to Equation (2.12) this then appears as a quadratic term in the pressure loss.

In conclusion, a minimum catalyst volume with minimum pressure loss is obtained by using a very flat bed with a large flow cross-sectional area A and small hydraulic channel diameter or particle diameter d_{h}. This result applies in general to all catalyst forms. The main difficulty with this arrangement is the uniform distribution of the flow velocity. Figure 9 shows the improvement in pressure loss that can be achieved by channel monoliths as compared to spherical packings. Other catalyst forms such as Raschig

rings or corrugated-plate packets lie between these boundary curves.

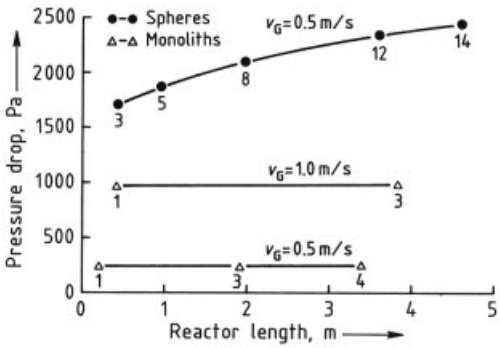

Figure 9. Pressure drop over the fixed-bed length for a catalytic combustion reactor with given throughput and 99 % conversion as a function of the hydraulic catalyst diameter d_h for different gas empty-tube velocities v_G [11]

2.4.2. Catalyst Forms for Isothermal Operation

In addition to the previously discussed quantities, the lateral heat-transport parameters of the fixed bed are further decisive parameters in isothermal reaction control. As shown in Section 2.3, heat transport can be characterized by the effective thermal conductivity perpendicular to the main flow direction λ_r and the wall heat transfer coefficient α_W. Both quantities are strongly dependent on the filling or packing form. Since transport in the industrially interesting region mainly occurs by convection, they are approximately proportional to the mass throughput.

For the purposes of overall comparison, the corresponding characteristic parameters are given in Figure 10 for some industrially important filling and packing forms in a tube of 50 mm internal diameter with a mass-flow velocity of $G_z = 1\ \mathrm{kg\,m^{-2}\,s^{-1}}$.

The dimensions of the packing bodies were chosen so that their specific external surface area a_p is ca. 500 $\mathrm{m^2/m^3}$. Under these conditions hollow, thin-walled cylinders have clear advantages over other packing forms, exhibiting the lowest pressure loss and the highest thermal conductivity. Only as regards wall heat transfer are they inferior to spheres or cylinders. However, good wall heat transfer is apparently less decisive from a reaction engineering viewpoint than good radial thermal conductivity, since the former can be compensated by an appropriate temperature profile of the heat-transfer medium, whereas the radial thermal conductivity directly influences the uniformity of the reaction conditions over the tube cross section. On the other hand, solid cylinders with a large length-to-diameter ratio have good heat transport values, but at the cost of a very high pressure loss.

Despite its poor heat transport properties, a monolith with straight, parallel channels, such as used for automobile exhaust control, is included in the comparison.

Monolith forms have very high specific surfaces combined with a very low pressure loss. Crossed corrugated structures are considerably more favorable for isothermal reaction control. They have a very high radial thermal conductivity that is almost independent of the specific surface area; the latter can be varied over a wide range by means of the channel dimensions. The catalyst of a reaction tube can thus be structured so that wide packings of small specific surface area can be used in the region of the main reaction zone, while packings of increasingly narrower structure, i.e., large specific surface area, are used downstream. In this way a more uniform reaction rate and temperature profile can be achieved over the tube length (see Section 4.3.3).

However, it must be remembered that with crossed corrugated structures, convective radial heat transport occurs only in one plane perpendicular to the main flow direction. In addition, the flow behavior in tubes of circular cross section is rather nonuniform over the circumference, which is why it is advantageous to arrange short packing sections in series, each section being displaced by 90 °. The heat transport parameters in Figure 10 were determined for structures arranged in this way.

A general problem in the use of monolith structures in reaction tubes is incomplete wall contact. Since individual tubes of multitubular reactors always have a diameter tolerance of ca. 1 mm and interlocking of the structure with the tube wall must be avoided, the bypass of flow at the wall is even greater than with random packings. Up to now there have been no specific investigations of the magnitude and effects of this phenomenon.

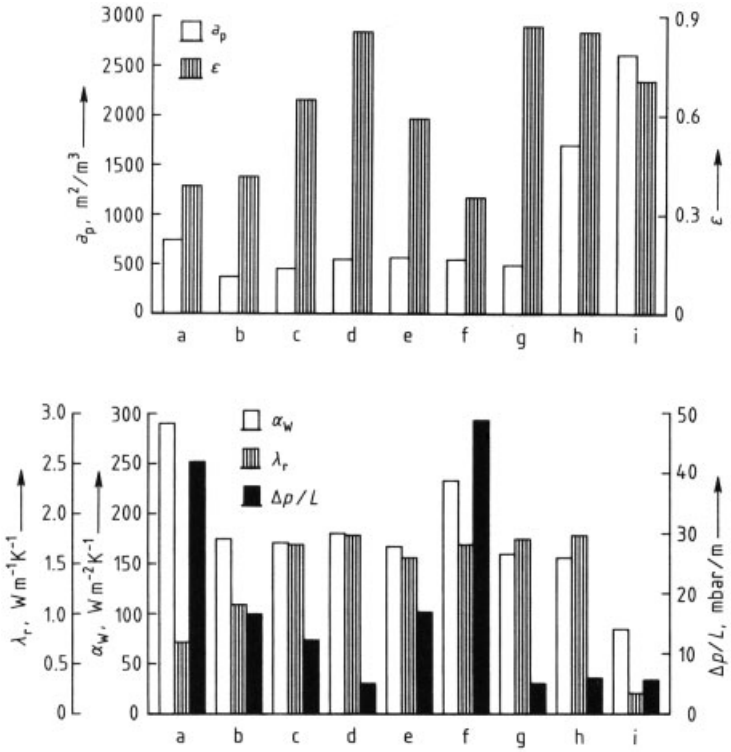

Figure 10. Specific outer surface area a_p, void fraction ε, effective radial heat conductivity λ_r, wall heat-transfer coefficient α_W, and pressure drop Δ_p for different packings in a tube of 50 mm internal diameter and an air mass-flow velocity $G_z = 1\ \mathrm{kg\,m^{-2}\,s^{-1}}$ at ambient conditions
a) Glass spheres, $d_p = 5$ mm; b) Glass spheres, $d_p = 10$ mm; c) Raschig rings, ceramic, $d_a = 10$ mm, $d_i = 6$ mm, $l = 11$ mm; d) Raschig rings, stainless steel, $d_a = 10$ mm, $d_i = 9$ mm, $l = 11$ mm; e) Hollow ceramic cylinders of irregular length, $d_a = 9$ mm, $d_i = 4.5$ mm, $l \approx 15$ mm; f) Full ceramic cylinders of irregular length, $d_a = 5$ mm, $l = 11$ mm; g) Crossed corrugated metal-plate packing (Sulzer Katapak) wide channels; h) Crossed corrugated metal-plate packing, narrow channels; i) Automotive exhaust monolith, Cordirith, square channels, $d = 0.9$ mm

3. Adiabatic Reaction Control

Adiabatic fixed-bed reactors are the oldest fixed-bed reactor configuration. In the simplest case they consist of a cylindrical jacket in which the catalyst is loosely packed on a screen and is traversed in the axial direction (Fig. 11 A). To avoid catalyst abrasion by partial fluidization, catalyst packings are always traversed from top to bottom. If fixed beds composed of monolith catalyst sections are used, the flow direction is arbitrary.

As discussed in Section 2.4.1, the requirement for a low pressure loss leads to a fixed bed of large diameter and low height (Fig. 11 B). Such an arrangement (disk concept) is used particularly when very short residence times, followed by direct quenching of the reaction, are required. Examples include ammonia oxidation in nitric acid production (→ Nitric Acid, Nitrous Acid, and Nitrogen Oxides, Chap. 1.3.) and oxidative dehydrogenation on silver catalysts (e.g., synthesis of formaldehyde by dehydrogenation of methanol, → Formaldehyde, Chap. 4.1., → Formaldehyde, Chap. 4.2.). In the first case the "fixed bed" consists of several layers of platinum wire gauzes, and in the second case, of a porous silver layer several centimeters in height. The bed diameters can be up to several meters.

On account of the difficulties involved with obtaining uniform flow as well as for structural reasons, the disk concept is limited to small catalyst volumes. The radial flow concept (Fig. 11 C) is used where large amounts of catalyst are involved. The catalyst packing is accommodated in the space between two concentric screen rings or perforated plate rings, and is traversed radi-

ally, either from the inside to the outside or from the outside to the inside. This design is particularly suitable for large catalyst volumes as well as for operation at elevated pressure, since at moderate reactor diameters the catalyst volume can be varied over a wide range by altering the reactor length, without affecting the flow-through length of the packing.

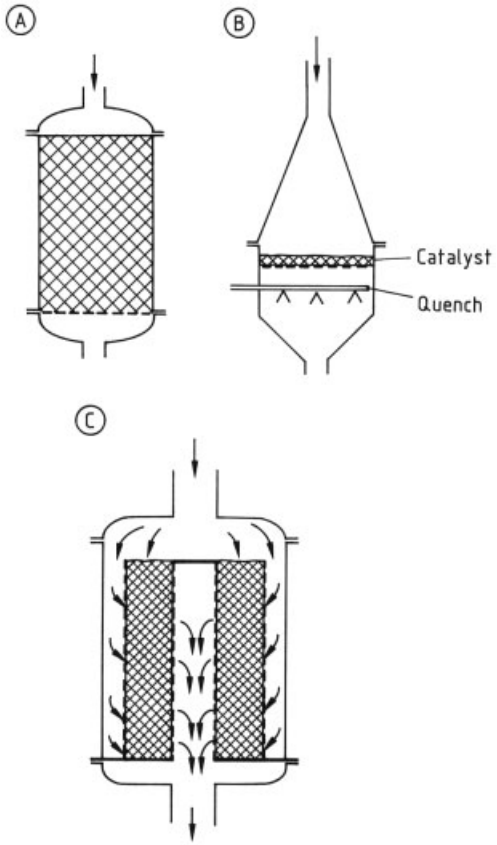

Figure 11. Main design concepts for adiabatic reactors
A) Adiabatic packed-bed reactor; B) Disk reactor;
C) Radial-flow reactor

A critical feature of packed radial-flow reactors is the shape of the upper bed closure. A simple horizontal covering is not practicable since a gap through which unreacted gas can pass is then formed due to the unavoidable settling of the packing. The arrangement shown in Figure 12 has proved effective since it produces mixed axial and radial flow through the bed in the upper bed closure. The required geometrical shape must be determined by simulation of the local two-dimensional flow through the packing [14].

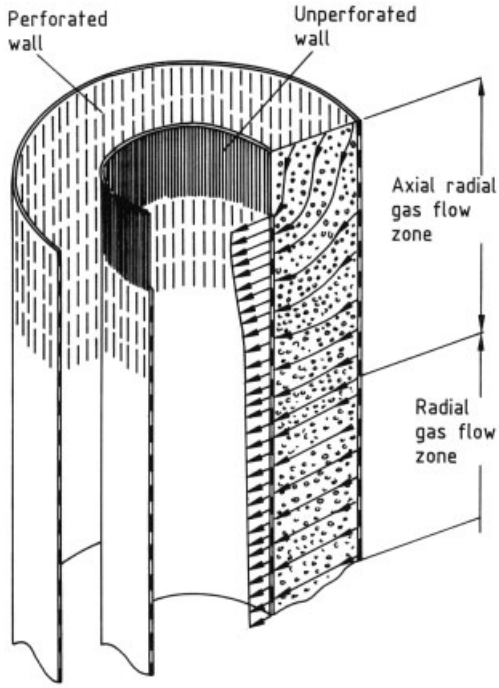

Figure 12. Upper bed closure in a radial-flow reactor [14]

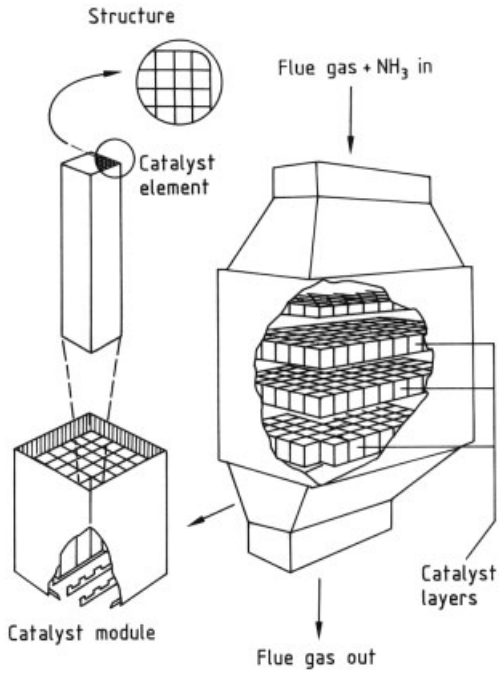

Figure 13. Reaction chamber for removal of nitrogen oxides from power station flue gas [15]

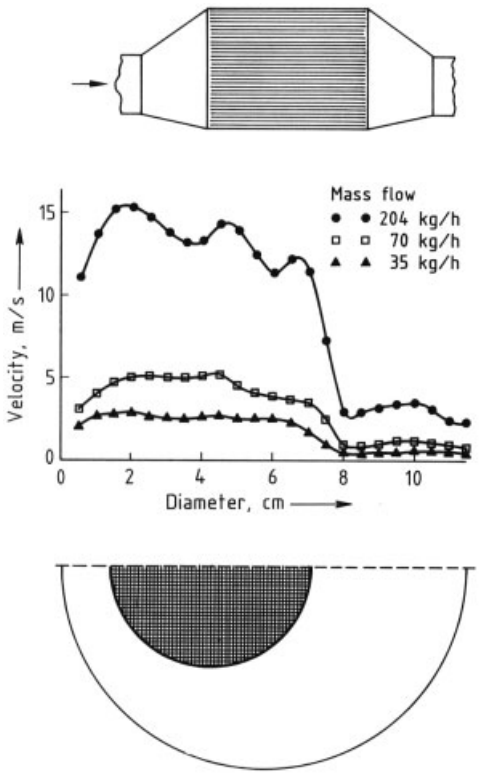

Figure 14. Velocity distribution in an industrially housed automobile catalyst [11]

The advantages of monolith catalysts with straight, parallel channels for adiabatic reactors have already been referred to in Section 2.4.1. Since monolith catalysts are usually produced with a rectangular cross section, the fixed bed is constructed by arranging these individual monoliths in rows in the form of a large "box". Conventional DENOX reactors for removing NO_x from power station flue gases are therefore designed as rectangular chambers (Fig. 13). The catalyst is often arranged in the form of several layers in series, the spaces between the individual layers permitting cross-mixing, so that the influence of nonuniform flow as well as any possible local blockage of the next layer can be compensated to some extent.

Reference is made in Section 2.2 to the importance of uniform flow into and through adiabatic fixed-bed reactors. This is not easy to achieve, particularly with low-pressure-loss monolith reactors, and requires a careful design of the inflow hood. On account of the low pressure loss, unfavorable flow conditions in the outflow hood may also affect the flow behavior through the catalyst bed.

Figure 14 shows the velocity distribution in front of the monolith inlet for an industrially housed automobile catalyst [11]. Since the flow cannot follow the sudden widening of the inlet funnel, one third of the total cross section is traversed at a velocity that is roughly three times the mean velocity. It can be estimated that, with uniform flow through the catalyst, half the catalyst volume would be sufficient for the same mean conversion. Also, bends in the feedpipe can lead to swirl-type components and thus contribute to nonuniform flow.

Purely adiabatic fixed-bed reactors are used mainly for reactions with a small heat of reaction. Such reactions are primarily involved in gas purification, in which small amounts of interfering components are converted to noninterfering compounds. The chambers used to remove NO_x from power station flue gases, with a catalyst volume of more than $1000\,m^3$, are the largest adiabatic reactors, and the exhaust catalysts for internal combustion engines, with a catalyst volume of ca. 1 L, the smallest. Typical chemical applications include the methanation of CO and CO_2 residues in NH_3 synthesis gas, as well as the hydrogenation of small amounts of unsaturated compounds in hydrocarbon streams. The latter case requires accurate monitoring and regulation when hydrogen is in excess, in order to prevent complete methanation due to an uncontrolled rise in temperature, a so-called runaway (see Chap. 6).

4. Reaction Control with Supply or Removal of Heat in the Reactor

4.1. Introduction

In the majority of fixed-bed reactors for industrial synthesis reactions, direct or indirect supply or removal of heat in the catalyst bed is utilized to adapt the temperature profile over the flow path as far as possible to the requirements of an optimal reaction pathway. Here a clear developmental trend can be observed, which is illustrated schematically in Figure 15.

Development started with the adiabatic reactor (Fig. 15 A), which on account of the adiabatic

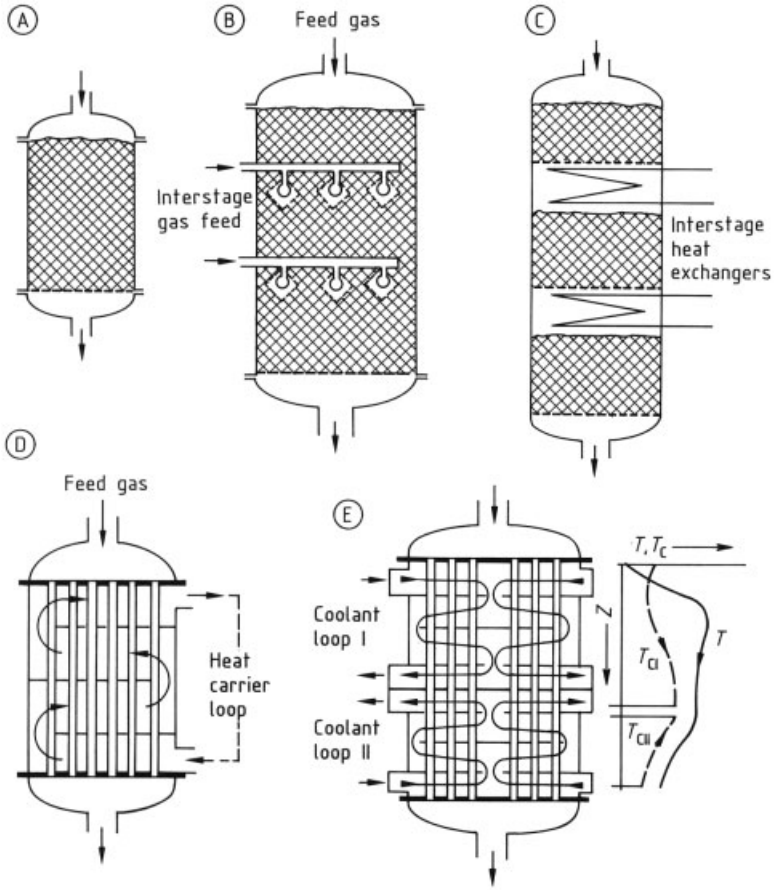

Figure 15. Development of fixed-bed reactors
A) Single-bed adiabatic packed-bed reactor; B) Adiabatic reactor with interstage gas feed (ICI concept); C) Multibed adiabatic fixed-bed reactor with interstage heat exchange; D) Multitubular fixed-bed reactor; E) Multitubular fixed-bed reactor with two cooling circuits and nonisothermal cooling temperature (the temperature profile shown is for a strongly exothermic reaction)

temperature change could only be operated to give a limited conversion. Higher conversions were achieved at the same mean temperature level when several adiabatic stages were introduced, with intermediate heating or cooling after each stage. The simplest form involves injecting hot or cold gas between the stages (Fig. 15 B). For a constant tube diameter, the main disadvantages of this temperature control strategy are cross-sectional loading, which increases from stage to stage, and the mixing of hot and cold streams, which is energetically unfavorable. The composition is changed by injection, which can have a positive or negative effect on the desired reaction.

The next development was the replacement of injection cooling by interstage heat exchangers, through which the required or released heat of reaction is supplied or removed (Fig. 15 C).

The development of reactors in which the heat-exchange surfaces are integrated in the fixed bed to supply or remove the heat of reaction as close as possible to the reaction site occurred in parallel with the development of multistage adiabatic reactors with intermediate heating or cooling. The multitubular fixed-bed reactor (Fig. 15 D) constitutes the oldest and still predominant representative of this class of fixed-bed reactors. Here the catalyst packing is located in the individual tubes of the tube bundle. The heat-transfer medium is circulated around the

tube bundle and through an external heat exchanger, in which the heat of reaction is supplied or removed. Whereas with endothermic reactions, circulating gas can be used as heat-transfer medium, for strongly exothermic reactions exclusively liquid or boiling heat-transfer media are used. Only in this way can the catalyst temperature (e.g., in the case of partial oxidations) be held in the narrow temperature range necessary for selective reaction control.

Initially, the integration of heat exchange in the fixed bed was utilized to ensure as isothermal a reaction control as possible, which is why reactors of the type shown in Figure 15 D are also commonly termed "isothermal reactors". They are characterized by reaction tubes of 20 – 80 mm internal diameter and a carefully designed flow control of the liquid heat-transfer medium, with largely constant heat-transfer conditions throughout the tube bundle and maximum temperature changes of the heat-transfer medium in the tube bundle of a few degrees.

The latest concepts are aimed at establishing a freely selectable (within limits) optimum temperature profile over the tube length. This requires complex heat-transfer medium control with several sections and temperature levels (Fig. 15 E).

The stimulus for the developments outlined in Figure 15 was the need for total raw material utilization as regards both mass and energy. This involves as main criteria the yield of primary end products (maximum), the yield of byproducts that must be removed and eliminated (minimum), the thermal energy consumption or recovery, and the mechanical energy requirements (gas compression, circulation of heat-transfer medium). In addition to the running costs, which are determined by the above criteria, the investment costs are decisive in an investment decision, and naturally rise sharply with increasing complexity of the reaction cycle.

On account of cost degression the projected plant size and the subsequent degree of utilization as well as the technological sophistication are decisive in calculating the product price. Several of the fixed-bed reactor variants illustrated in Figure 15 can, depending on the location of the production site and the estimated output, be used for the same process. For example, multistage adiabatic reaction systems with intermediate superheated steam feed, multitubular reactors with circulating gas heating, and strictly isothermal designs with multitubular reactors heated with molten salt are currently used for the endothermic synthesis of styrene from ethylbenzene. Also, exothermic, equilibrium-limited reactions such as methanol synthesis are carried out in multistage adiabatic reactors with interstage cooling as well as in multitubular reactors. Overall, however, there is a trend towards the more highly integrated designs, which will accelerate with rising energy and raw material prices. Adiabatic multistage designs and reactors with heat exchange integrated in the fixed bed are discussed in the following sections. Autothermal reaction control, in which the heat of reaction of moderately exothermic reactions is utilized to heat the incoming feedstock, is discussed separately in Chapter 5.

4.2. Adiabatic Multistage Reactors with Interstage Heat Transfer

Adiabatic multistage fixed-bed reactors with intermediate cooling or heating are nowadays used particularly where the reaction proceeds selectively to give a single product but is limited by the equilibrium conditions. Intermediate cooling or heating is used to displace the gas temperature in the direction of higher equilibrium conversion. Typical examples include the synthesis of ammonia, sulfur trioxide, and methanol. In these exothermic reactions the equilibrium conversion to the target product decreases with increasing temperature, as shown in Figure 16 A. For a given conversion x a temperature can therefore be found at which the reaction rate, with respect to the target product, becomes a maximum. This temperature must be below the equilibrium temperature but not so low that the reaction becomes too slow for kinetic reasons. The points obtained in this way can be joined to form a maximum reaction rate curve (Fig. 16 B).

Since, in the case of adiabatic reaction control the temperature increases linearly with the achieved conversion Δx according to the equation,

$$\Delta T = -\frac{\Delta h_{\mathrm{R}} \cdot c_0}{\varrho_{\mathrm{G}} \cdot c_{\mathrm{pG}}} \cdot \Delta x \tag{4.1}$$

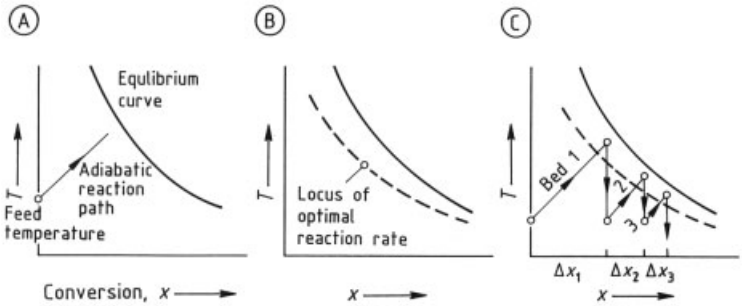

Figure 16. Equilibrium limitation of exothermic reactions
A) Equilibrium conversion as a function of temperature; B) Optimum reaction rate curve; C) Improvement of conversion by interstage cooling

each adiabatic reaction pathway of an exothermic reaction lies on a straight line of gradient $\Delta T/\Delta x$ (see Fig. 16 A).

A practicable reaction pathway for a multistage adiabatic reaction can thus be derived from Figure 16 by joining straight-line sections for the adiabatic reaction to vertical lines for the temperature reduction due to indirect intermediate cooling (Fig. 16 C).

The kinetically optimum reaction pathway with the smallest required catalyst volume results when the trajectory follows, in a large number of small steps, the line of maximum reaction rate. In practice, the apparatus and equipment expenditure involved in a large number of stages must be weighed up against the savings in catalyst. Conventional multistage reactors for this class of reaction therefore have three to five stages. Figure 17 shows the layout of an ammonia synthesis reactor designed on this basis. For structural reasons the heat exchanger is incorporated between the inflow and outflow in the lowest part of the pressure casing. The reaction gas then flows upward in the annular gap between the pressure casing and the fixed beds, whereby it is further heated and at the same time protects the pressure-bearing structural components against excessively high fixed-bed temperatures. The three adiabatic fixed-beds are traversed from top to bottom, part of the heat of reaction being utilized to generate steam in the two intermediate heat exchangers. To start up the cold reactor, hot gas must be added to the uppermost bed, for example through an external start-up preheater.

Industrial adiabatic multistage reactors often differ in many details from Figure 17, although they are of a comparable basic design. For example, radially instead of axially traversed beds can achieve a smaller pressure loss with a more favorable structural arrangement; heat exchange with the cold feedstock can be effected by heat exchange surfaces integrated in the first catalyst bed; or a cold gas quench can be used to achieve additional temperature regulation.

Modern, adiabatic multistage reactors may thus become so complex that the question arises whether a multitubular design according to Figure 15 D or E does not represent the more favorable alternative. The required heat-exchange surface area in the case where heat exchange is integrated in the fixed bed is smaller than in the case of free gas flow on account of the positive effect of the catalyst packing. Furthermore, it does not involve any additional pressure loss, and the optimum reaction rate curve (Fig. 16 B) can be better approximated by controlling the cooling temperature profile (see Section 4.3.2.) than by a stepwise temperature reduction.

On the other hand, a multistage arrangement may be considered for structural, operating technology, or kinetic reasons in the following cases:

1) If, in the case of large single-train plants, subdivision into several individual items of apparatus is necessary for reasons of transport or construction
2) If a catalyst must be replaced in individual stages at different times on account of different catalyst compositions and/or aging
3) If a gradual addition of a reactant has kinetic advantages compared to the total addition to the feed (here a suitably designed intermediate heat exchanger ensures a uniform dis-

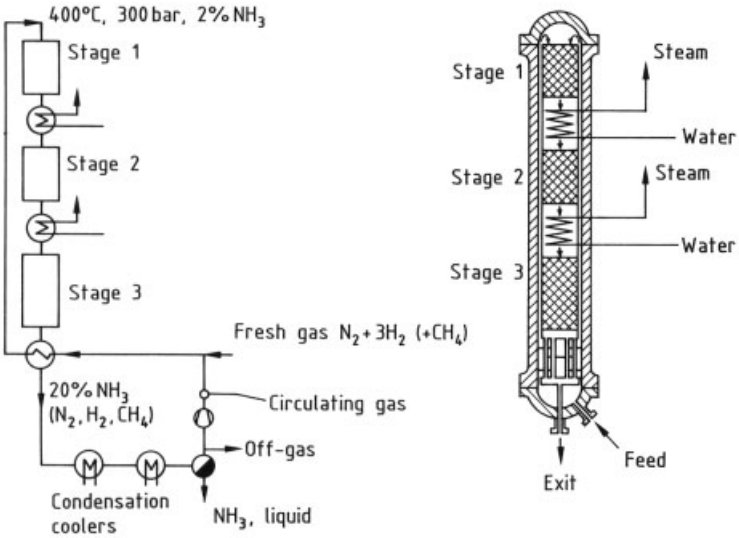

Figure 17. Schematic of a multistage reactor for ammonia synthesis

tribution and mixing with the reaction gas stream)

4) If the intermediate stages are used to extract a limiting product in the case of equilibrium-limited reactions; an example is the intermediate absorption of SO_3 before the last stage of the SO_3 synthesis
5) With reaction temperatures above 300 °C intermediate cooling can be performed directly with boiling water, whereas in a fixed bed a high-temperature heat-transfer medium must be used as coolant (see Section 4.3.1)

4.3. Reactors with Heat Exchange Integrated in the Fixed Bed

The aim of reaction control with heat exchange integrated in the fixed bed using a circulating heat-transfer medium is to maintain the catalyst temperature in a narrow optimum range under all operating conditions. With strongly exothermic successive reactions, such as partial oxidations and partial hydrogenations, on account of the danger of a runaway reaction (see Section 6.2) this requirement can be met only if

1) The temperature of the heat-transfer medium is close to the desired catalyst temperature
2) Large heat- exchange surfaces are available per unit catalyst volume
3) A sufficiently large mass flow velocity of the reaction gases ensures good heat transport from the packing to the heat-exchange surface

With exothermic equilibrium reactions and endothermic reactions these requirements are less stringent since these reactions cannot runaway, although here too it is beginning to be recognized that a tight and uniform temperature control over the reactor cross section is advantageous.

4.3.1. Heat Transfer Media for Fixed-Bed Reactors

The first of the above requirements presupposes an assortment of heat-transfer media that covers the whole temperature range of interest for gas-phase reactions in fixed-bed reactors. It is convenient to distinguish between gaseous, liquid, and vaporizing heat-transfer media. Gaseous heat-transfer media in the form of hot flue gases are used in the temperature range above 500 °C exclusively to supply heat for endothermic reactions.

Conversely, vaporizing heat-transfer media are used exclusively to remove heat from exothermic reactions. Whereas formerly petroleum fractions such as kerosene (e.g., in ethylene

oxide synthesis) were more widely used, they have now been largely replaced by boiling water on account of their flammability, lower heat of vaporization, and the need to produce steam in a downstream condenser/heat exchanger. Depending on the saturated vapor pressure, the temperature range from 100 to 310 °C (100 bar) can be covered with boiling water. In this range it is the preferred heat-transfer medium for evaporative cooling if an isothermal cooling temperature is required.

Locally variable cooling temperature profiles can be established most easily with liquid heat-transfer media that do not vaporize in the intended operating range. In order to avoid cavitation, pressurized water should be used only up to ca. 220 °C; heat-transfer oils cover the temperature range up to 300 °C, while above this temperature salt melts are now used exclusively in reaction technology [16]. Compared to heat-transfer oils they have the advantage that they are incombustible and stable, although they have the disadvantage that they solidify at about 200 °C (nitrate melts) or 400 °C (carbonate melts). The temperature ranges of possible heat-transfer media are compared in Figure 18. In addition to the thermal stability, the energy $\dot{N}$ per unit amount of transported heat $\dot{Q}$ required to circulate a heat-transfer medium is an important criterion of choice.

The following equation has been derived for the evaluation of heat-transfer media without phase change [26]:

$$\frac{\frac{\dot{Q}}{\Delta T_c}}{\dot{N}^{0.36}} \sim \frac{\varrho^{0.73} \cdot c_p}{\eta^{0.09}} \tag{4.2}$$

where $\dot{Q}$ is the heat flow transferred by the heat-transfer medium while its temperature increases by exactly ΔT_c, $\dot{N}$ is the required pump power, and ϱ, c_p, and η are the density, specific heat capacity, and viscosity of the heat-transfer medium. For liquid heat-transfer media, the equation is derived from the pressure drop in a turbulently traversed, hydraulically smooth pipe. It is practical to solve Equation (4.2) for the pump power:

$$\dot{N} \sim \left[\frac{\dot{Q}}{\Delta T_c}\right]^{2.75} \cdot \frac{\eta^{0.25}}{c_p^{2.75} \cdot \varrho^2} \tag{4.3}$$

Thus the required power varies as almost the third power of $\dot{Q}/\Delta T_c$. The component $\eta^{0.25}/c_p^{2.75} \cdot \varrho^2$, which depends only on the physical properties of the heat-transfer medium, is plotted as a function of temperature in Figure 19. This shows the exceptional heat-transfer properties of water that result from its density and, in particular, its high specific heat, as well as the properties of gases (air) and vapors (water vapor), which are about seven orders of magnitude worse. Common heat-transfer oils and salt melts (sodium nitrate) lie close together, but about an order of magnitude above water. The poor performance of molten sodium compared to heat-transfer oils and salt melts is noteworthy.

To test Equation 4.3 on a more realistic example of the cooling of a multitubular reactor, a quadratic tube bundle (cross section 1.5×1.5 m) with fourfold passage of the coolant was designed by using detailed equations for pressure drop and heat transfer [27] for various heat-transfer media. The heat release of 1650 kW in 1591 (37×43) tubes of 30 mm external diameter and 4 m length corresponds to the conditions in a partial oxidation reaction. A temperature rise between feed and outlet of $\Delta T_c = 5$ K was assumed for the heat-transfer medium (additionally 150 K for air and water vapor). Table 2 lists as results the required pump power $\dot{N}$, the pressure drop in the cooling circuit Δp (whereby $p = 1$ bar exit pressure was assumed for water vapor and air), and the excess temperature of the reactor wall above the heat-transfer medium temperature ΔT_W, assuming constant wall temperature along the length of the tube. It was shown that the results of these more precise design equations for the liquid heat-transfer media can be correlated with the equation

$$\dot{N} = 7800 \left[\frac{\dot{Q}}{\Delta T_c}\right]^{2.73} \frac{\eta^{0.27}}{c_p^{2.73} \cdot \varrho^2} \tag{4.4}$$

This represents a good confirmation of Equation (4.3).

Water is thus the ideal heat-transfer medium within its temperature range. For higher temperatures molten salts have increasingly replaced the previously more commonly used heat-transfer oils. Salt melts cover a larger temperature range and have the particular advantage over oils that they are incombustible. The potential danger of a relatively large amount of hot salt melt obviously exists, but is reliably dealt with by experienced reactor construction companies. Special

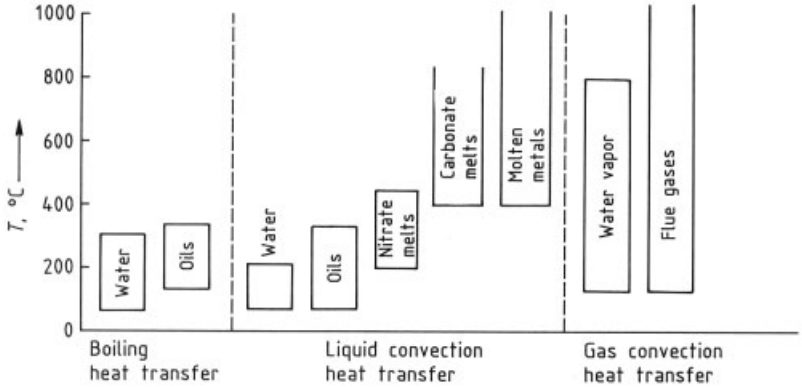

Figure 18. Application ranges for common heat-transfer media

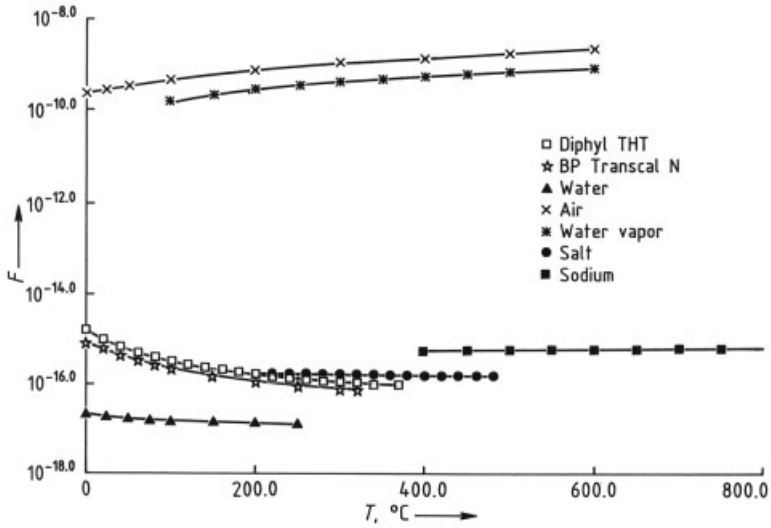

Figure 19. The material-specific factor $F = \eta^{0.25}/c_{\mathrm{p}}^{2.75} \cdot \varrho^2$ for various heat-transfer media as a function of temperature

Table 2. Results of the example calculation

Heat-transfer medium	Producer	Pump power $\dot{N}$, kW	Pressure drop Δp, bar	Mean temperature difference ΔT_{w}, K	Permitted cooling temperature increase ΔT_{c}, K
Water		0.060	0.008	0.75	5
Transcal N	BP	0.629	0.042	2.32	5
Diphyl THT	Bayer	0.862	0.054	2.35	5
Sodium		4.54	0.150	0.04	5
HT salt		1.31	0.105	1.45	5
Water vapor		0.119×10^6	8.759	1.30	5
Water vapor		0.254×10^3	0.175	14.22	150
Air		0.229×10^6	18.001	1.08	5
Air		0.456×10^3	0.341	12.29	150

nitrate melts (HITEC) can be used in the temperature range 200 – 500 °C. Gradual decomposition begins above this temperature, which can accelerate violently above 600 °C. Access of organic components to the melt (nitrate decomposition) and of water (steam explosion) must be excluded [16]. New salt melts, for example, based on carbonates, are being developed for the temperature range 400 – 800 °C. In this case it is not so much the thermal stability of the molten salt but rather the corrosion of the reactor materials that presents problems.

Gases are the only heat transfer media usable over the entire temperature range, but because of their low density they have an unfavorable heat transport behavior (Fig. 19). They are therefore used exclusively as flue gases to supply heat at high temperatures. However, large temperature differences between the heat-transfer medium and the reactor wall, with possible adverse effects on the uniformity of the heat supply, must then be accepted.

4.3.2. Reactor Designs

As with adiabatic reactors, a principal task of reactor development is to produce uniform reaction conditions over the whole reactor and maintain such conditions during the entire operating time. This involves the conditions in the catalyst packing (residence time, catalyst concentration and activity, heat transport) and in the heat-transfer medium circuit (throughput, temperature, heat transfer). When liquid or vaporizing heat-transfer media are used, the heat-transfer coefficients are usually one order of magnitude greater than those on the catalyst side, which facilitates the task as regards the heat-transfer medium. On the other hand, with strongly exothermic, selectivity-sensitive reactions (especially partial oxidations), a temperature constancy of the heat-transfer medium of ca. 1 °C is often required. This leads to a high energy requirement for circulation (according to Eq. 4.4) and necessitates extremely careful design and control of the heat-transfer medium circuit.

Tubular reactors of the type shown in Figure 15 D have been in use longest and have been furthest developed. In the case of a vaporizing heat-transfer medium an arrangement as shown in Figure 20 is generally chosen, in which the liquid heat-transfer medium surrounds the stationary tube bundle. The rising vapor bubbles escape through the ascending pipe into a vapor drum, where they are separated from the liquid. A circulation flow is established due to the difference in density in the downpipe (pure liquid) and in the reactor jacket, which means that circulating pumps are generally not required. The heat-transfer medium temperature is regulated via the saturated vapor valve. Specific details of this regulation are discussed in Section 6.6.

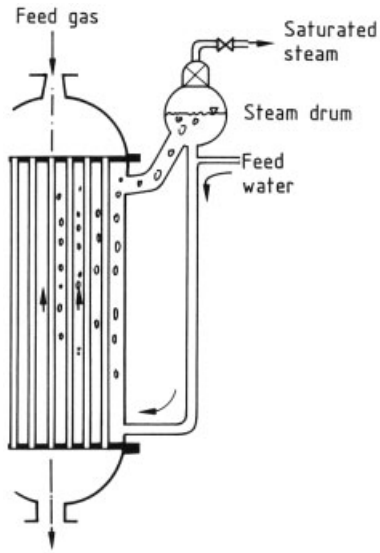

Figure 20. Multitubular reactor with boiling water cooling

If uniform supply and removal of heat-transfer medium via annular channels is ensured, this arrangement provides reliable isothermal reaction control. Since a steam cushion can form under the upper tube floor, the active catalyst layer should begin only at a deeper level.

Apart from the type of multitubular reactor shown in Figure 20, other multitubular reactors are sometimes used in which the catalyst bed is arranged around the tubes and the heat-transfer medium flows through the tubes (Fig. 21 B). An interesting new development has been introduced by Linde (Fig. 21 A): the tube bundle is composed of counterwound spirals in which upwardly flowing water is evaporated. The tubes run into a vertical vapor drum located at the reactor head. The tube bundle is connected to a central downpipe at the bottom so that, as in the arrangement in Figure 20, a natural circula-

tion of the evaporating water is established. Advantages in construction and in the heat transfer from the reaction gas to the tubes of the bundle are claimed for this design [28].

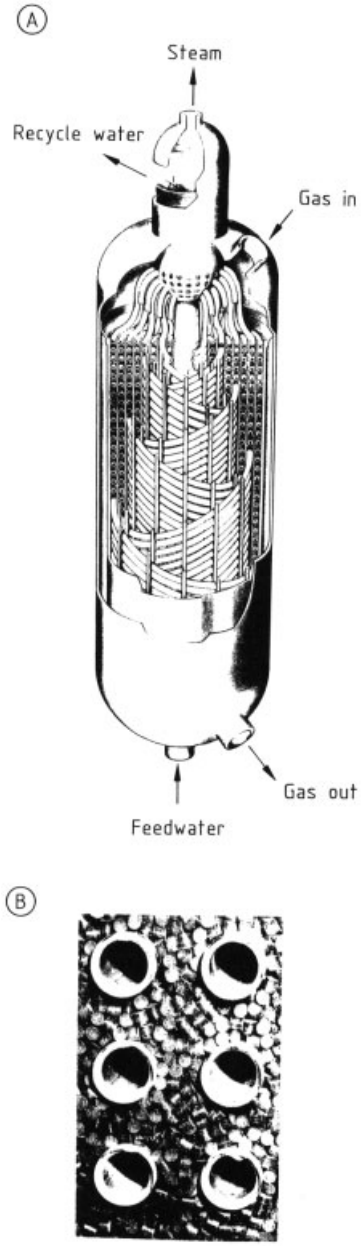

Figure 21. A) Design concept of the Linde isothermal reactor for methanol synthesis; B) Cut through the tube bundle surrounded by catalyst pellets (from [31])

Circulation systems with parallel and crossed cocurrent or countercurrent flow of the heat-transfer medium (Fig. 22) are commonly employed for liquid heat-transfer media. The main part of the heat-transfer medium is generally circulated with a high-capacity pump in order to achieve uniform heat-exchange conditions, while a partial stream is passed through a heat exchanger for supplying or removing the heat of reaction. The desired heat-transfer medium temperature is attained by regulating this partial stream. With exothermic reactions the heat exchanger is normally a steam generator which produces saturated steam at a pressure corresponding to a boiling point of 30 – 80 °C below the maximum cooling temperature.

A superheater fed with a further partial stream of the heat-transfer medium can if necessary be connected downstream of the steam generator. The arrangement with separate external units, shown schematically in Figure 22 has structural and maintenance advantages over a common arrangement in the tube-free interior of the reactor [29].

Apparatus construction companies specializing in these reactors have developed a detailed and comprehensive know-how as regards flow control of the heat-transfer medium [30]. This concerns the uniform supply and removal of the heat-transfer medium, which generally takes place via external annular channels, as well as the flow control within the reactor. Here a distinction is made between parallel flow control and crossed flow control. Parallel flow control (Fig. 22 B) is achieved by two rectifier plates with narrow bores. On account of the pressure loss through the bores suitably arranged over the reactor radius, a uniform flow profile is produced over the cross section with uniform heat removal conditions between the distributor plates. The advantage of this arrangement is that the whole reactor can be equipped with tubes.

However, due to the nonuniform flow conditions in the inflow and outflow region of the heat-transfer medium, only the region between the distributor plates should be filled with active catalyst. Heat transfer is low because of the parallel flow of the heat transfer medium, and the pressure loss in the distributor plates does not contribute to improving the heat transport. For this reason this arrangement is preferably used for reactions with a moderate heat of reaction.

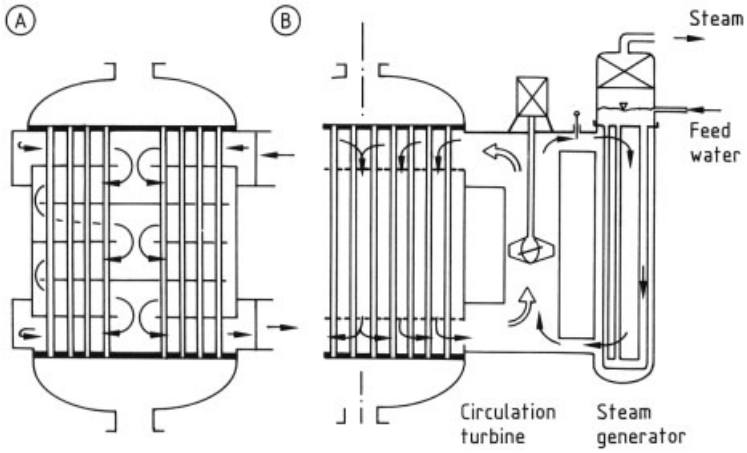

Figure 22. Heat-transfer medium control in tube-bundle fixed-bed reactors
A) Cross flow; B) Parallel flow

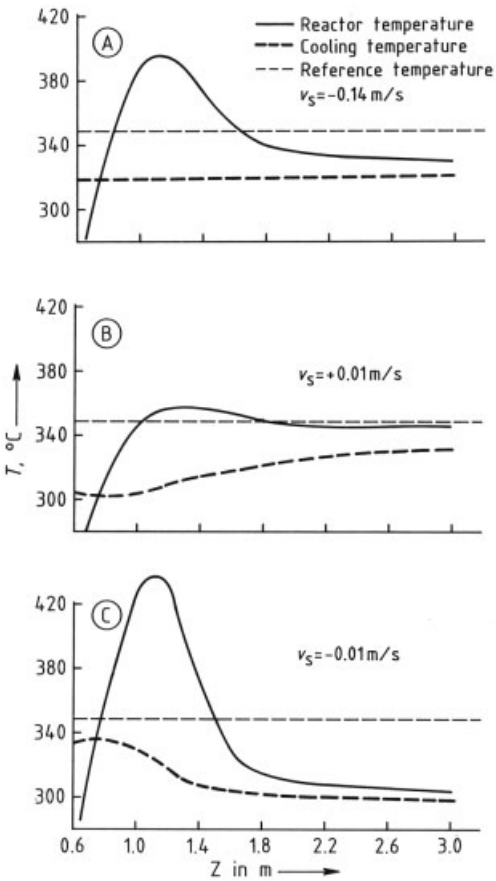

Figure 23. Influence of the coolant flow direction and flow velocity v_s on the reaction temperature profile
A) Isothermal; B) Cocurrent flow; C) Countercurrent flow

In reactions with a large heat of reaction, especially partial oxidations, transverse flow control is more widely used, generally in a radially symmetrical arrangement with baffle plates, as illustrated in Figure 22 A. Since the heat removal conditions are poorly defined in the region of the flow deflection, with the change from a transverse stream to a parallel stream and back to a transverse stream, this region should be free of tubes. A difficulty arises from the constant change of the flow cross section in the radial direction. However, this can be alleviated, for example, by providing pressure-relief bores in the baffle plates, increasing in number towards the reactor axis. In this way constant heat-transfer conditions should be achieved over the reactor radius. In summary, ensuring uniform heat transfer conditions as regards the heat-transfer medium thus requires a considerable flow technology know-how. Some publications illustrate the major differences in the behavior of different tube sections that can arise due to an inadequate design and layout of the heat-transfer medium circuit [17–19].

4.3.3. Influencing the Course of the Reaction

The course of the reaction (i.e., the conversion and selectivity or yield) can be decisively influenced by the arrangements made for controlling the heat-transfer medium. The most obvious, although technically most complex solution, is to

arrange different heat-transfer medium circuits so as to achieve a stepwise approximation of an optimum temperature profile. The purposeful utilization of the temperature change of the heat-transfer medium flowing through the reactor is, however, technically simpler, and is discussed here in connection with cocurrent or countercurrent cooling of a fixed-bed reactor involving an exothermic reaction.

Figure 23 shows temperature profiles for three different ways of controlling the cooling stream in a partial oxidation reaction. If the coolant is circulated so fast that its temperature in the reactor scarcely alters, then its flow direction is irrelevant and a temperature profile with a pronounced temperature maximum becomes established; this is typical of strongly exothermic reactions (Fig. 23 A). If the coolant is circulated in cocurrent and its velocity is chosen so that it becomes noticeably hotter over its path, an almost isothermal temperature behavior can be achieved (Fig. 23 B). This is because the reactive gas at the inlet is in contact with the coldest coolant and the cooling temperature rises in step with the consumption of the reactants, so that the reaction rate remains virtually constant over a fairly long section [20–22]. This stabilizing effect of cocurrent cooling has hardly been exploited up to now in industrial reactors. This may be due to the fear that, at the required low flow velocity (in the example of Fig. 23 B, $v_s = 0.01$ m/s), heat transfer will be inadequate and natural convection will occur in the cooling jacket. However, v_s describes the mean coolant velocity parallel to the tube axis. With a cross-cocurrent flow of the coolant, the actual flow velocity may in fact be substantially larger, depending on the number of deflections, with the result that the aforementioned problems do not arise.

Compared to cocurrent flow, countercurrent flow has a markedly destabilizing effect at low flow velocities (Fig. 23 C). Since the incoming reaction mixture in this case is in contact with the warm coolant outflow, the maximum temperature rises to much higher values. Countercurrent cooling can even lead to the occurrence of several stationary states, and in general favors the runaway of a strongly exothermic, irreversible reaction (see Section 6.3). In contrast to heat exchange without a reaction, countercurrent control of the heat-transfer medium in reactors involving exothermic reactions should therefore be chosen only in particular cases.

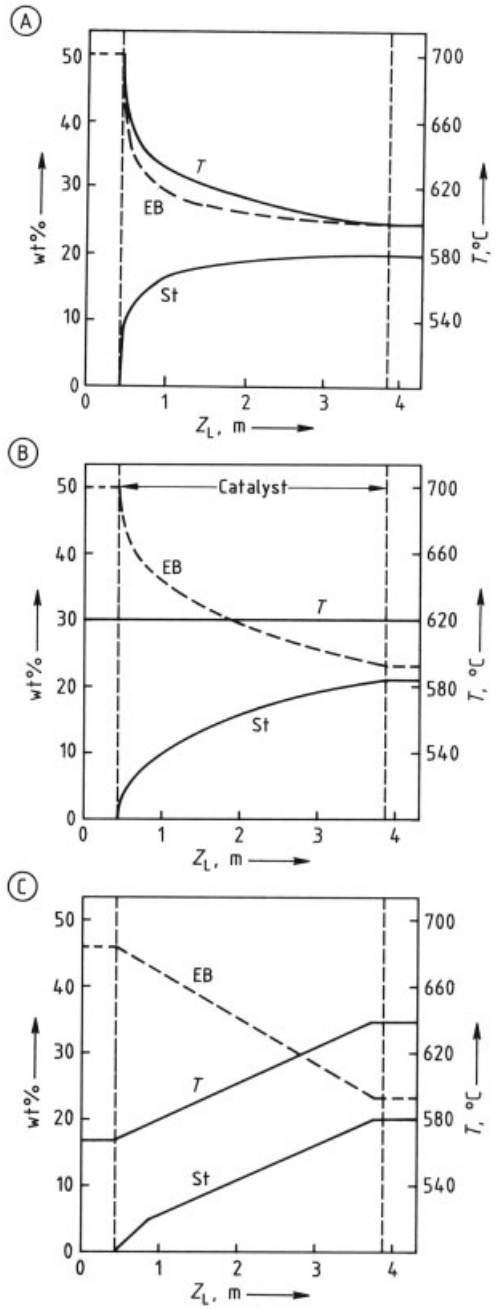

Figure 24. Influence of the heating agent flow direction on the temperature profile in styrene synthesis
A) Adiabatic; B) Isothermal; C) Countercurrent
EB = ethylbenzene; St = styrene

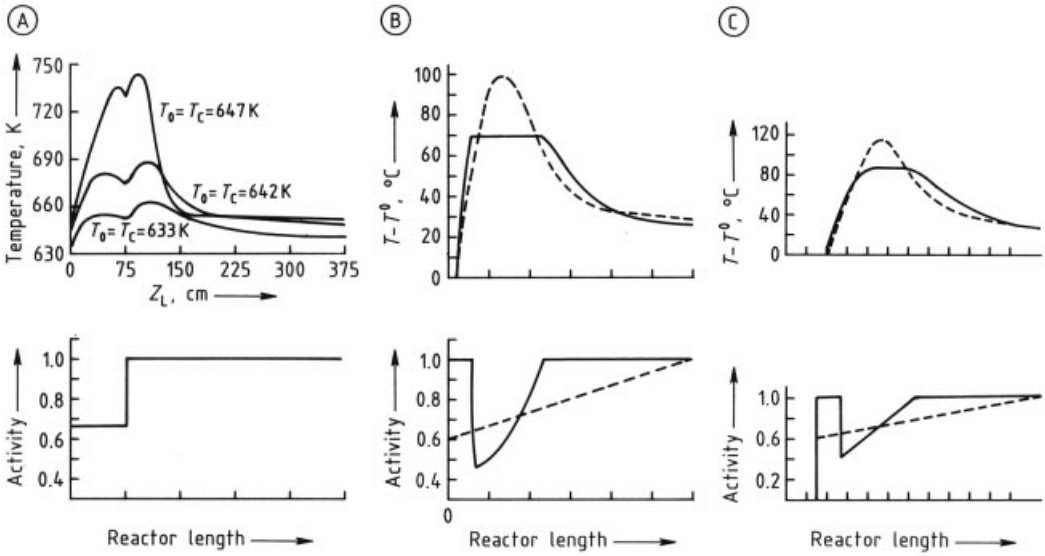

Figure 25. Influence of activity profiles on the temperature profile of a strongly exothermic reaction
A) Catalyst with 66 % activity in the front section of the tube and 100 % activity in the rear section of the tube [23]; B) Linear (broken line) and optimal catalyst activity distribution (full line) for limiting the maximum temperature to 370 °C (simulation result); C) Experimental verification of B [25]

The temperature control of an exothermic equilibrium reaction can constitute such a case. As illustrated in Figure 16 B, the optimum temperature profile should in this case decrease with increasing conversion, i.e., along the tube length. On account of the equilibrium inhibition of the reaction, it is not possible for the reaction to runaway in the front region. Countercurrent flow of the heat-transfer medium is also advantageous for endothermic equilibrium reactions. Figure 24 shows the calculated temperature and concentration profiles with different heating conditions in the synthesis of styrene (dehydrogenation of ethylbenzene). With adiabatic and isothermal reaction control, styrene formation decreases with increasing tube length, whereas it remains roughly constant with countercurrent control.

A significant advantage of nonisothermal control of the heat-transfer medium in cocurrent or countercurrent flow is the saving in circulation energy, since much smaller heat transfer-medium streams must be circulated. Overall, the combination of several heat-transfer medium circuits (Fig. 15 E) and the purposeful utilization of the temperature change of the heat-transfer medium in the reactor offer a wide range of possibilities of establishing optimum temperature profiles for a given reaction, and of counteracting any changes in activity that occur during the operating life of the catalyst by altering the temperature profiles.

A further possibility of influencing the course of the reaction is to use catalysts of different activities over the reactor length. Particularly with strongly exothermic reactions that are liable to runaway, such as partial oxidations, a less active catalyst is occasionally used in the front part of the reactor in order to avoid too high a maximum temperature. Figure 25 A illustrates the use of two catalysts of differing activity in series. The resulting temperature profiles have a typical double-hump shape [23], [24]. This can be avoided if an activity profile is established by using a continuously varying mixture of catalysts with different activities (Fig. 25 B, C) [25]. In these cases the fully active entry region (relative activity 1) is designed so that the temperature rises to a preselected maximum value. In order to maintain the temperature at this value, the activity in the following region is sharply decreased and is then raised to a relative activity of 1 as the reaction rate drops due to the depletion of the reactants. A smooth temperature profile

can be achieved even if the optimized activity profile in Figure 25 B is only approximately realized (Fig. 25 C).

The control of the maximum temperature by using locally differing catalyst activities presents problems if the main reaction zone moves into the region of high catalyst activity due to changes in the operating conditions. For example, in the case of Figures 25 B and C, a decrease in the throughput may already result in reaction runaway in the short, fully active front region. This can be counteracted by reducing the activity of this zone. Catalyst deactivation occurring during operation may have more serious effects. If, for example, the catalyst is poisoned in a front migrating from the entrance to the rear, the main reaction zone finally migrates to the highly active rear catalyst region, which may lead to excessively high temperatures that can no longer be controlled.

In general, influencing the reaction course via control of the cooling stream is more flexible than incorporating catalysts of different activities. Particularly with nonisothermal control of the heat-transfer media, the reaction course can be influenced over a wide range by means of the inflow temperature of the heat-transfer medium as well as by its volumetric flow rate.

5. Autothermal Reaction Control

5.1. Introduction

The expression autothermal reaction control is used with fixed-bed reactors if the heat of reaction is utilized to heat the reactor feed to the ignition temperature of the catalyst. Then, neither addition nor removal of heat occurs during stationary operation of the reactor. It thus follows that autothermal reaction control is restricted to exothermic reactions in which a moderately large amount of heat is produced. As a measure of the heat release, the adiabatic temperature increase ΔT_{ad} is usually used. This characterizes the temperature increase that the reaction mixture would experience at the maximum possible conversion Δx if no heat were released. As already given in Equation (4.1):

$$\Delta T = -\frac{\Delta h_{\mathrm{R}} \cdot c_0}{\varrho_{\mathrm{G}} \cdot c_{\mathrm{pG}}} \cdot \Delta x$$

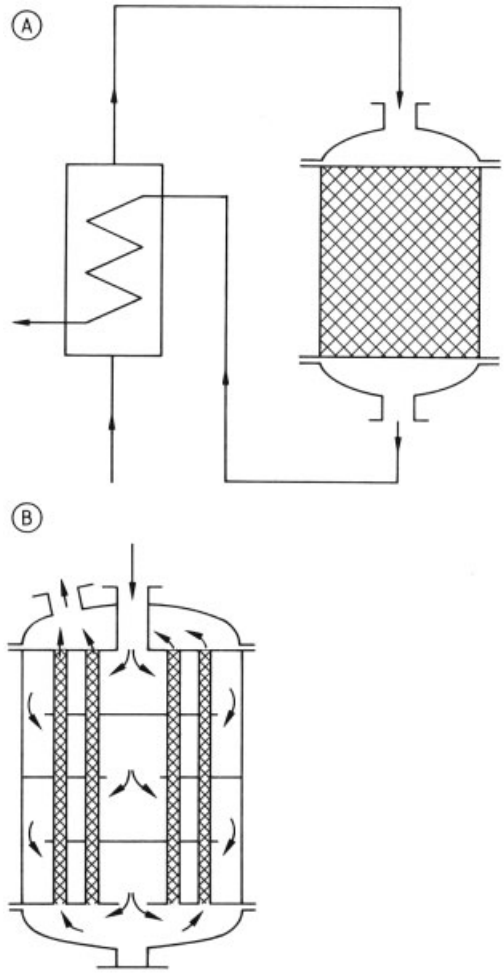

Figure 26. Reactor circuits for autothermal reaction control with indirect heat exchange
A) External heat exchange; B) Internal heat exchange

If the conversion is not limited by the reaction equilibrium, then $\Delta x = 1$ and ΔT_{ad} depends only on the inflow concentration c_0 of a key component. Autothermal reaction control is used in practice if the adiabatic temperature increase is in the range $20\,\mathrm{K} \leq \Delta T_{\mathrm{ad}} \leq 300\,\mathrm{K}$. Autothermal reaction control can be realized in several ways, the simplest being to couple an adiabatic reactor with a countercurrent heat exchanger in which the hot outflow heats the cold inflow (Fig. 26 A). If the heat exchanger is not to become impracticably large, this arrangement is suitable only for an adiabatic temperature increase $>$ ca. 200 K. If the heat exchange is in-

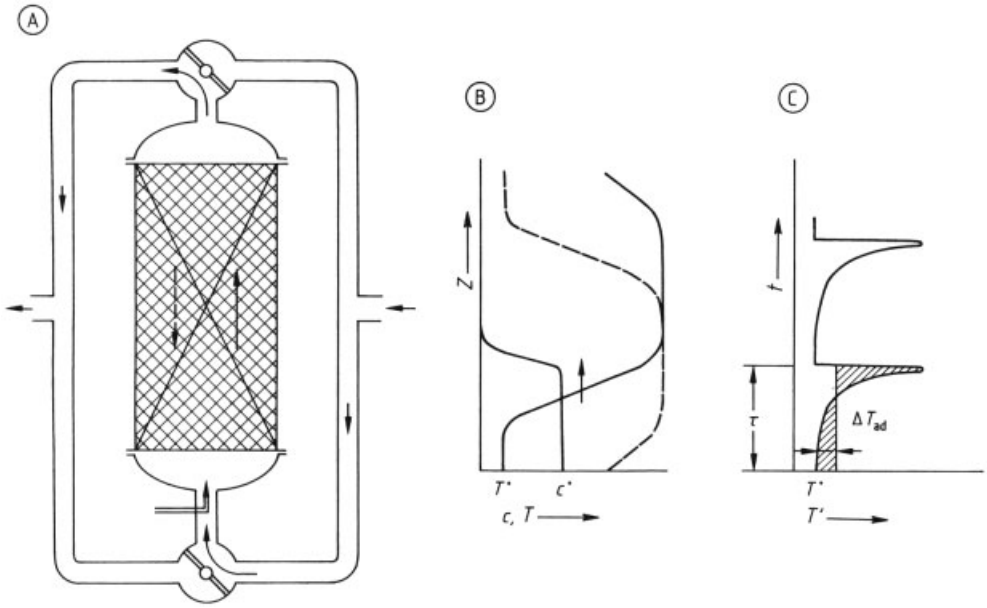

Figure 27. Autothermal reaction control with direct (regenerative) heat exchange for an irreversible reaction [11]
A) Basic arrangement; B) Local concentration and temperature profiles prior to flow reversal in steady state; C) Variation of outlet temperature with time in steady state

tegrated in the reactor (Fig. 26 B), then a better heat transfer can be achieved inside the catalyst-filled tubes as well as around the tubes as a result of the transverse flow, so that also gases with a somewhat lower adiabatic temperature rise can be reacted autothermally. Overall, however, indirect gas – gas heat exchange is a weakness of this reactor concept.

5.2. Reaction Control with Periodic Flow Reversal

A major improvement was the autothermal reaction control with direct, regenerative heat exchange, developed by Matros et al., in which the catalyst packing simultaneously acts as the regenerative heat exchanger [32–34]. Figure 27 A shows the basic arrangement. After the catalyst fixed bed has been heated to the reaction temperature, for example, with a pilot burner, the cold reaction gas flows into the packing, where it is heated by the hot catalyst packing and then reacts. At the same time the inflow part of the packing is cooled, so that the reaction front migrates into the packing (Fig. 27 B). Before the reaction front has reached the end of the bed, the flow direction is reversed by valves, so that the temperature front moves back again and heats the cooled part of the packing. In this way a periodic steady state is finally established, in which the temperature profile moves up a certain amount in one half cycle, and moves down by the same amount in the next half cycle. The upper and lower ends of the packing each serve as a regenerative heat exchanger, the hot reaction zone being maintained in each case in the interior of the packing.

If the fixed bed is adiabatically insulated, the heat of reaction can leave the reactor only via the outflow. This leads to the saw-toothed course of the outflow temperature T' with time shown in Figure 27 C. From the total energy balance it can easily be seen that the time-average value of the outflow temperature must lie exactly ΔT_{ad} above the inflow temperature. Nevertheless, the temperature peaks shown in Fig. 27 C may considerably exceed this mean value.

The main advantage of this form of reaction control is the very effective regenerative heat exchange in the catalyst packing, which does not require any additional built-in components and permits autothermal reaction of gas mixtures with an adiabatic temperature increase of less

than 20 K. The main application of autothermal reaction control with periodically varying flow direction is catalytic off-gas purification, in particular catalytic oxidation of small amounts of combustible substances in exhaust air [11]. In addition this form of reaction control has been tested by MATROS et al. for equilibriumlimited exothermic reactions such as SO_3, methanol, and ammonia synthesis, and has in some cases been employed on an industrial scale [32–34].

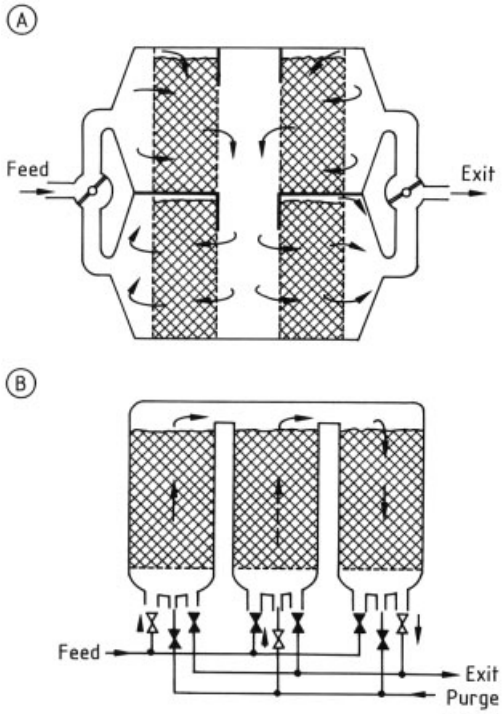

Figure 28. Alternative arrangements for autothermal reactor design with regenerative heat exchange
A) Radial flow concept; B) Multiple-bed arrangement

In addition to the standard variants shown in Figure 27, a number of modifications have been proposed and applied (Fig. 28). For example, the bed shown in Figure 28 A can be designed as a radial flow system, so that the hot region insulates itself, and a multibed arrangement may be used to avoid flushing losses when reversing the flow direction, one bed being backflushed each time (Fig. 28 B). The arrangements of Figure 29 offer an elegant possibility of a valveless, continuous operation with an integrated flushing stage [11]. Analogous to the Ljungström heat exchanger, the rotation of the segmented fixed bed with a stationary inlet, outlet, and flushing segment leads to a periodic change in the direction of flow through the bed, which can be either axial (Fig. 29 A) or radial (Fig. 29 B).

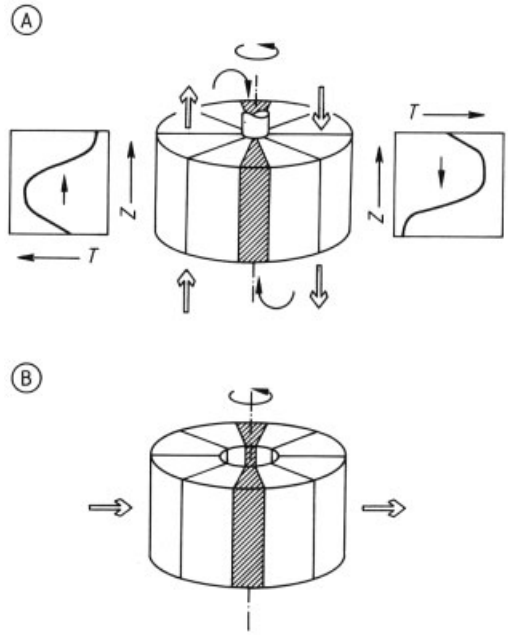

Figure 29. Schematic of flow control with a rotating catalyst bed [11]
A) Axial-flow reactor; B) Radial-flow reactor
The flushing segment is shaded.

5.3. Reactor Behavior

Due to the direct coupling of the reaction with the heat exchange, autothermal reaction control exhibits some specific features as regards reactor behavior that distinguish it from the behavior of conventionally operated fixed-bed reactors. Firstly, the autothermal procedure presupposes the existence of several stationary operating states (see Section 6.3). The "ignited" operating state is the desired state, for only then is so much heat released that the cold inflow can be heated. In the "extinguished" state the reactor cools to the inflow temperature, so that no further detectable conversion takes place. The object of reaction control of autothermal reactors is to prevent extinction of the reaction, but also to limit the rise in the maximum temperature in the reactor due to the positive thermal feedback (see Section 6.3).

With regard to the combined handling of autothermal reaction control with indirect and direct (regenerative) heat exchange, it is convenient to consider the limiting case of a very rapid reversal of the flow direction [11]. For the sake of simplicity the discussion is restricted to a simple, irreversible exothermic reaction. A typical example is the catalytic oxidation of a combustible component in a waste air stream.

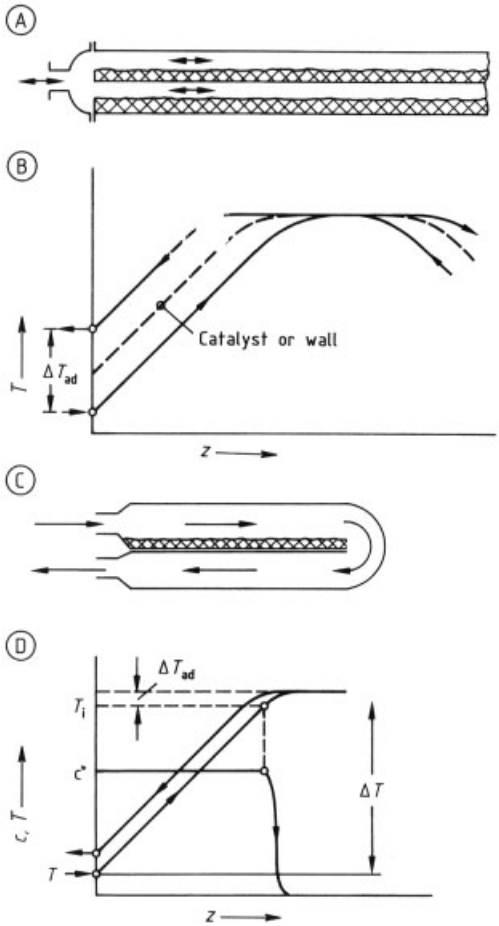

Figure 30. Equivalence of operating with periodical flow reversal and integrated countercurrent heat exchange
A) Matros reactor; B) Temperature profiles in the Matros reactor with rapid flow reversal; C) Countercurrent reactor with catalyst at the wall; D) Schematic concentration and temperature profiles in natros and countercurrent reactors [11]

The mean gas temperature and mean catalyst temperature in the packing are considered separately. With a very frequent reversal of the flow direction, the catalyst temperature remains practically constant because the heat capacity of the packing is ca. 1000 times that of the gas, whereas the gas temperature is quasi-stationary compared to the packing temperature and therefore, depending on the flow direction, is somewhat below or above the packing temperature. To illustrate this, the principal temperature profiles are shown disproportionately far apart in Figure 30 B. These profiles are completely equivalent to those of a countercurrent heat exchange reactor in which the heat-exchanging partition is coated with catalyst (Fig. 30 C).

Since the stationary behavior of the countercurrent reactor can be calculated much more simply than the steady state of the reactor with periodic flow reversal, the equivalence model is suitable for rapid parameter studies. In addition, it permits straightforward assessment and interpretation of the somewhat unusual operating behavior of autothermal reactors. Considering the concentration profile corresponding to the temperature behavior in Fig. 30 D it can be seen that the reaction starts at an "ignition temperature" T_i and goes to rapid completion over a short section.

Thus it can be assumed to a first approximation that the temperature difference ΔT in Figure 30 D can be calculated from:

$$\Delta T = T_i + \Delta T_{ad} - T_0 \tag{5.1}$$

A detailed analysis shows that for more accurate estimates, heat exchange in the region of the reaction zone must also be included. In the following discussion the ignition temperature is therefore designated somewhat pragmatically as the temperature that lies ΔT_{ad} below the maximum temperature.

The gradient of the temperature curves below the ignition temperature can be derived from the conditions of countercurrent heat exchange with axial thermal conduction [11]:

$$\frac{dT}{dz} = \frac{U_w \cdot a_p \cdot \varepsilon \cdot \varrho_G \cdot c_{pG} \cdot v}{2 \cdot U_w \cdot a_p \cdot \lambda_{eff} + 4\left(\varepsilon \cdot \varrho_G \cdot c_{pG} \cdot v\right)} \cdot \Delta T_{ad} \tag{5.2}$$

This allows a simple graphical interpretation of the influence of the operating parameters on the reactor behavior.

Figure 31 A shows the change in the temperature and concentration profiles in the case of sufficiently quick reversal or countercurrent operation for gases with different ignition temperatures. The feed concentrations in each case are

chosen such that the same adiabatic temperature increase occurs. The plateau region of the maximum temperature becomes increasingly narrow with rising ignition temperature. If the ignition temperature lies above the point of intersection T_s of the two lateral straight lines, then operation in the ignited state is not possible. Accordingly, the narrower the plateau region of the maximum temperature, the closer the operating point is to the stability boundary; the wider the maximum temperature plateau, the more stable is the operation.

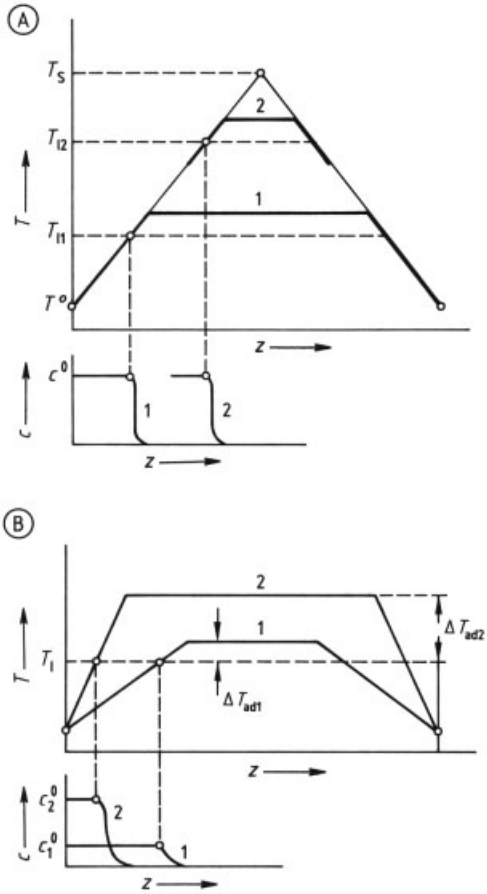

Figure 31. The influence of operating parameters on reactor behavior
A) Temperature and concentration profiles for gases with different ignition temperatures T_{I1}, T_{I2}; B) Influence of the feed concentration c_1^0, c_2^0, in the form of an adiabatic temperature rise ΔT_{ad} [11]

In this connection, the influence of catalyst deactivation can also be illustrated. Deactivation means a rise in the ignition temperature. In this case autothermal operation exhibits a fortunate tendency to self-stabilization: the maximum temperature adapts quasi-automatically to the rising ignition temperature, so that total conversion is ensured over a large activity range. If the catalyst activity is too low, then extinction from a state with highest maximum temperature occurs rather abruptly. In the case where the catalyst is deactivated by excess temperature, the self-stabilization of course leads to a vicious circle: the higher the maximum temperature, the greater the deactivation, and the higher the maximum temperature, etc.

The influence of inert front and end zones, which are frequently used with this type of reactor, can also easily be explained by means of Figure 31 A. If the ignition temperature is T_{I1} and the fixed bed of the reactor is inert up to the point z_2, then the catalytic reaction can be initiated at the earliest at z_2; the maximum temperature is thus raised from T_{I1} to T_{I2}, compared to the case of a continuously active catalyst.

Figure 31 B shows the influence of the feed concentration in the form of the adiabatic temperature rise on the temperature profile in the steady state. According to Equation (5.2), an increase in ΔT_{ad} broadens the temperature profile and raises the maximum temperature, whereby the increase in the maximum temperature above the ignition temperature is roughly proportional to ΔT_{ad}.

The above discussion refers to countercurrent operation and to the limiting case of a very rapid reversal of the flow direction. Normally during operation with periodic flow reversal, the reaction zone first migrates a certain extent into the reactor. As shown in Chapter 6, a relationship for the front migration velocity w can be derived from the total energy balance around the moving reaction front, and in the case of a large temperature rise ΔT and a small adiabatic temperature increase ΔT_{ad}, w can be approximated as follows:

$$w = \frac{\varepsilon \cdot \varrho_G \cdot c_{pG}}{\varepsilon \cdot \varrho_G \cdot c_{pG} + (1-\varepsilon) \cdot \varrho_s \cdot c_s} \cdot v_G \tag{5.3}$$

w from the above equation also corresponds to the migration velocity of a temperature front in a packing without reaction traversed with a velocity v_G. The influence of the period duration on the reactor behavior with periodic alternation of the flow direction can easily be estimated by

using Equation (5.3). For stable, ignited operation the maximum temperature and full conversion must be reached during the entire period. A semiperiod may only be so long that the temperature profile after reversal changes to the gradient specified in Equation (5.2), without the maximum temperature falling below the ignition temperature. Thus, a conservative estimate of the maximum semiperiod duration $\tau_{\max}$ is

$$\tau_{\max} = \frac{\Delta l_{\max}}{w}$$

with

$$\Delta l_{\max} = L - 2\,(T_{\mathrm{i}} - T_0) \cdot \frac{\mathrm{d}T}{\mathrm{d}z} \tag{5.4}$$

with catalyst bed length L as well as w, ΔT, and $\mathrm{d}T/\mathrm{d}z$ from Equations (5.3), (5.1), and (5.2), respectively. The behavior of autothermal combustion reactors with periodically alternating flow direction or integrated countercurrent heat exchange can be estimated by using the above equations [11].

6. Stability, Dynamics, and Control of Industrial Fixed-Bed Reactors

6.1. Introduction

Stability, dynamics, and control of fixed-bed reactors with strongly exothermic reactions has been studied in great detail since the early 1970s. The numerous publications could give the impression that this is a particularly critical reactor type with a large potential risk. In fact, the opposite is true. Compared to a liquid-phase reactor of the same size, a fixed-bed reactor with a gas-phase reaction contains a mass of reactants several orders of magnitude smaller. There is therefore no danger of a runaway exothermic reaction due to reactants accumulating in the reactor, especially as the heat capacity of the catalyst mass additionally damps the uncontrolled temperature rise.

Nevertheless, instabilities can arise in fixed-bed reactors, particularly with strongly exothermic reactions, and can lead to excess temperatures that can damage the catalyst and the reactor construction materials. Some causes of this are discussed below.

Fixed-bed reactors for industrial syntheses are generally operated over a long production period with almost constant operating parameters. The task of process control engineering is simply to keep these parameters optimal. In contrast, for supply or disposal plants that have several users or suppliers in a production network, there are frequent changes of feed material and throughput which require fast, automatic reaction control. Examples are fixed-bed reactors for synthesis gas production or off-gas treatment.

6.2. Parametric Sensitivity

Parametric sensitivity is the property typical of all highly exothermic reactions with high activation energy: small changes in the operating parameters can lead to large changes in the maximum temperature and yield. The reason is the exponential dependence of the reaction rate on temperature (Arrhenius law). Figure 32 shows calculated temperature profiles for partial oxidation in a wall-cooled, fixed-bed reactor tube of typical dimensions. In Figure 32 A only the main reaction is considered, while in Figure 32 B the total combustion to CO_2 and water is additionally taken into account. Since considerably more heat is liberated in the total combustion than in the desired main reaction, the sensitivity is increased considerably. As a measure of the parametric sensitivity, Figure 32 C shows the change in the maximum temperature via the cooling temperature for case B. The sensitivity is only moderate at low cooling temperatures, whereas above $T_c = 343\,°C$ small changes in T_c – and also in other parameters such as throughput, feed concentration, or pressure – lead to large changes in reactor behavior. Due to the unavoidable differences between individual tubes, multitubular reactors cannot be operated in the range of high parametric sensitivity. In this case the cooling temperature must be lowered to ca. 340 °C, and the tubes made longer to give a good conversion. This example emphasizes the requirement, discussed in Chapter 4, for making the conditions in the tubes of the tube bundle and in the cooling circuit as uniform as possible to avoid premature runaway reaction in individual tubes.

In the literature there are numerous runaway criteria with which operating ranges of high parametric sensitivity can be precalculated for

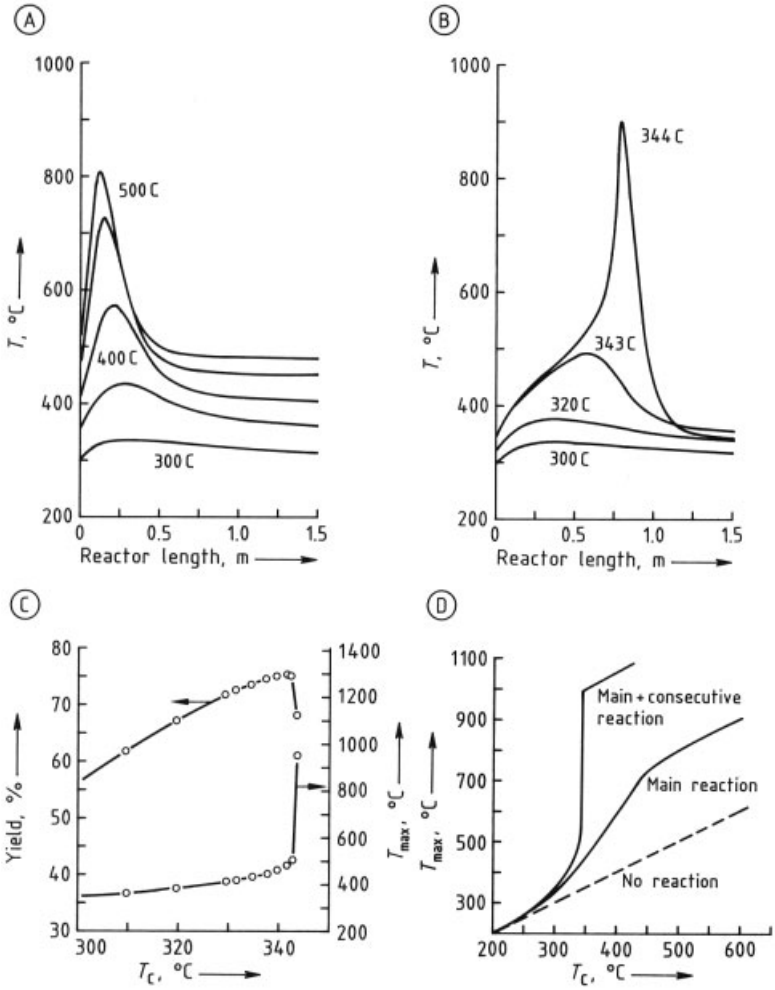

Figure 32. Parametric sensitivity of a partial oxidation reaction in a fixed-bed reactor of typical dimensions as a function of the coolant temperature T_C with $T(z=0)=T_C$
A) Temperature profile over reaction length (main reaction only); B) Temperature profile including total oxidation as side reaction; C) Maximum temperature T_{max} and yield as a function of coolant temperature T_C in case B; D) T_{max} as a function of T_C for both cases

known reaction kinetics [35–37]. In practice, however, these parameters are of only limited importance because they rarely take into account the peculiarities of individual cases. Sensitive reactions such as partial oxidation and partial hydrogenation are therefore generally tested in single-tube reactors of the same dimensions as those in the subsequent multitubular reactor. This allows the range of parametric sensitivity to be determined directly. Recalculation of the results for other tube diameters is only possible to a limited extent due to the uncertainties in the quantification of the heat-transfer parameters (see Section 2.3).

An operating fixed-bed reactor can enter the region of high parametric sensitivity through changes in the catalyst properties or operating conditions. Initially a few particularly sensitive tubes of the bundle will runaway, i.e., the reaction changes, for example, from a selective partial oxidation to a total combustion, and the temperature rises rapidly. In a multitubular reactor with thousands of tubes every tube cannot be equipped with temperature-profile measure-

ments; it is therefore likely that this runaway will remain undetected, especially if it involves only a few tubes. Although temperatures $> 1000\,°C$ can often be reached in the catalyst during such runaways, they are generally safe provided the tube is surrounded by heat-transfer medium. Because of the good heat transfer to the fluid the tube temperature remains close to that of the heat-transfer medium, and melting of the tube does not occur. The most certain method of detecting a runaway is on-line analysis of a product formed in the runaway reaction. For example, CO_2 can be monitored in the off-gas during the runaway-sensitive synthesis of ethylene oxide. If its concentration increases above a specified limit, the reactor must be shut down and for a certain period cooled to a lower cooling temperature before operation is recommenced. The reason why lowering the cooling temperature during operation does not extinguish the runaway reaction is discussed below.

6.3. Multiple Stationary States

The term parametric sensitivity applies when, according to Figure 32 D, a state quantity (T_{max}) is a single-valued function of a control quantity (T_C). However, in the runaway examples discussed in Section 6.1 this dependence can also be multiple valued. Then the runaway occurs not in a finite range of the control quantity, but on exceeding a fixed limit T_{ci}, often referred to as the ignition limit (Fig. 33). Between ignition and extinction of the runaway reaction (at $T_C = T_{cE}$), there is a hysteris region in which two stable stationary states are possible; at least one unstable intermediate state must lie between them. Various causes of multiple stationary states in fixed-bed reactors are known [38], [39]. Here, only the so-called thermal instabilities are discussed. They occur in exothermic reactions and arise because the evolution of heat increases exponentially with increasing temperature, while the heat removal at constant cooling temperature in-creases only linearly. Thus on exceeding a certain temperature limit, the heat release increases more rapidly than the heat removal, and the reaction ignites. On dropping below a second, lower temperature limit, the heat release becomes less than the heat removal, and the reaction is extinguished. In fixed-bed reactors, this process of ignition and extinction is associated with a local displacement of the main reaction zone. This can be most readily demonstrated by the example of an exothermic reaction in an adiabatic fixed-bed reactor.

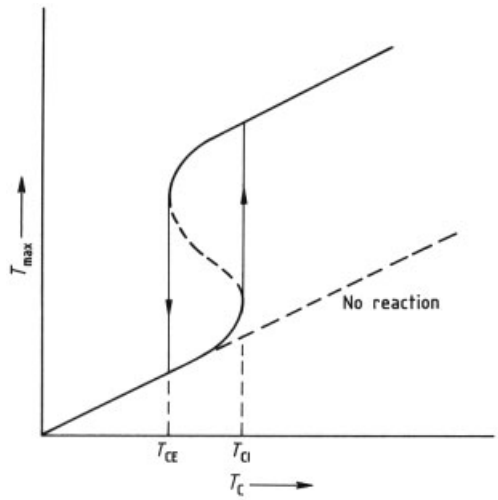

Figure 33. Ignition – extinction hysteresis in the region of multiple steady states
Ignition occurs at $T_c = T_{cI}$ and extinction at $T_c = T_{cE}$.

Figure 34 shows the temperature profile for the catalytic total oxidation of ethane as a function of feed temperature. If this temperature is raised only slightly above the value for profile I, the reaction ignites in the rear portion of the reactor. Due to the axial thermal conduction in the developing steep temperature front, the main reaction zone moves slowly towards the entrance and is stabilized in position II a. The feed temperature must be lowered considerably to the value of position II b before the reaction extinguishes completely to profile III upon a further slight lowering of the feed temperature. A similar ignition – extinction behavior can also arise on changing the throughput or the feed concentration. The multiple stationary states I–II a and II b – III are partly a result of the axial backward conduction of heat in the reaction front. This represents a positive feedback which stabilizes either the ignited or the extinguished state. Mechanisms that transport heat from the end of the bed to its beginning stabilize multiple stationary states in a similar manner. These include the axial thermal conduction in the tube wall [41] and especially the influence of countercurrent cooling when the temperature of the coolant rises sig-

nificantly [42]. In autothermal reaction control (see Chap. 5), countercurrent cooling is used to stabilize the ignited state, which is desired in this case. However, in the case of strongly exothermic reactions with a risk of runaway, countercurrent cooling should generally not be used. The destabilizing effect of countercurrent cooling is demonstrated by the examples discussed in Figure 23. Another effect that can result in multiple stationary states is single-grain instability (see → Principles of Chemical Engineering, Chap. 7.2., → Principles of Chemical Engineering, Chap. 7.3.), which in fixed-bed reactors is always coupled with the effect of the reverse axial backward thermal conduction of heat [43] and intensifies the above-mentioned hysteresis behavior.

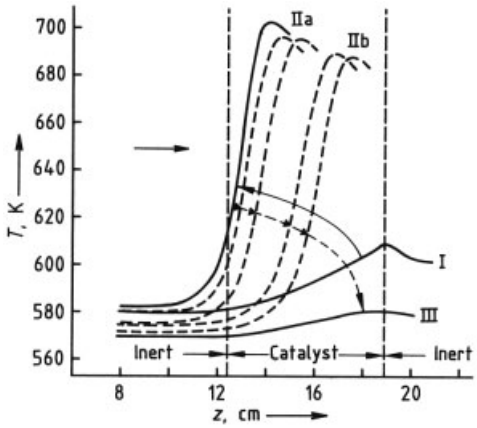

Figure 34. Measured axial temperature profile in the region of multiple steady states in an adiabatic fixed-bed reactor for the oxidation of ethane [40]
Ignition from profile I to II a; extinction from II b to III.

Whether a fixed-bed reactor goes runaway due to high parametric sensitivity or whether the reaction ignites on passing from a lower to a higher stationary state, the results in practice are the same: the maximum temperature increases rapidly to an unacceptably high value. However, in the case of parametric sensitivity, decreasing the cooling temperature or feed concentration during operation results in the maximum temperature returning to the safe range, whereas when multiple stationary states are present, this need not be the case. If a certain limit is exceeded on ignition, then the runaway reaction proceeds towards the ignited stationary state,

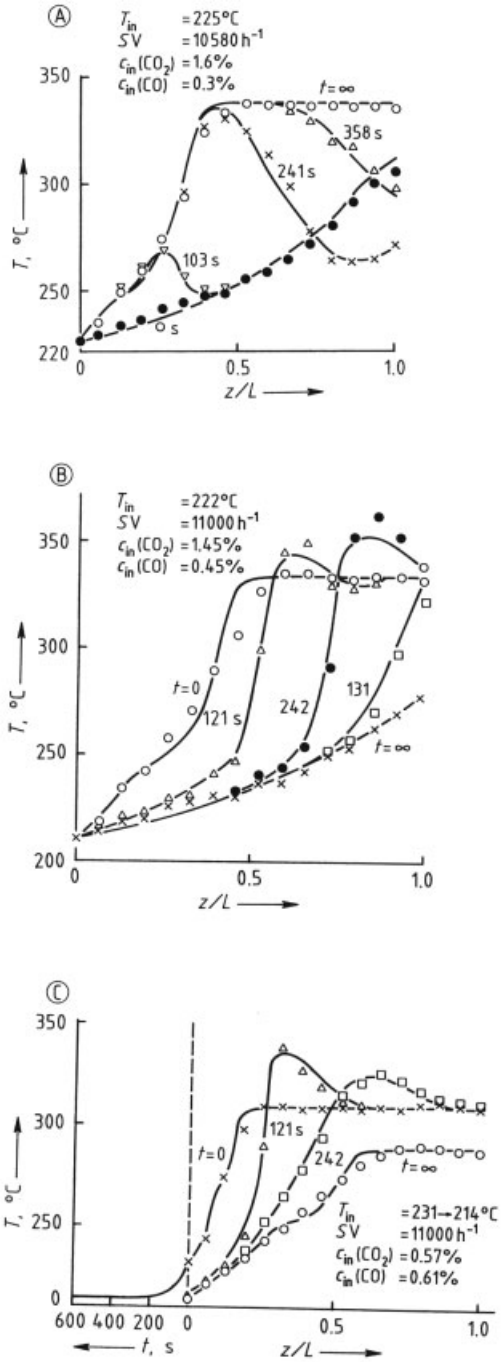

Figure 35. Measured (points) and calculated temperature profiles showing the transient of an adiabatic fixed-bed reactor for the methanation of CO and CO_2 [44]
A) Transition after increase in feed concentration; B) Transition after decrease in feed concentration; C) Transition after decrease in feed temperature
$SV = \dot{V}/V_R$ (space velocity)

even if the feed concentration or cooling temperature is lowered considerably. Extinction of the reaction occurs only on passing below the extinction limits, which particularly for strongly exothermic reactions lie far below the ignition conditions. In this case, the best strategy is to switch off the feed followed by inert gas purging, until all temperatures have fallen to uncritical values.

6.4. Migrating Reaction Zones

The discussion in the previous sections refers exclusively to the stationary state, i.e., the temperature and concentration profiles established after a sufficiently long operating time. However, in practice it is often more important in which way and with which velocity these profiles approach the next stationary state. This dynamic behavior of fixed-bed reactors with exothermic reactions exhibits several special features which are connected with the axial displacement of the main reaction zone in the form of a so-called migrating or moving reaction zone. The behavior can be most easily discussed by using the example of an adiabatic fixed-bed reactor.

Figure 35 shows measured temperature profiles for the methanation of traces of CO and CO_2. This reaction is carried out in an adiabatic fixed-bed reactor during ammonia synthesis gas production. If the feed concentration is increased, a new main reaction zone forms in the front part of the reactor, and the temperature rises here to a new maximum value. At the same time, the reactor outlet temperature initially drops below the original value before increasing to the new final value. This temporary reaction in the wrong direction is known as wrong-way behavior. More surprising is the wrong-way behavior in the examples of Figure 35 B (decrease in feed concentration) and Figure 35 C (decrease in feed temperature). Particularly in the latter case, the maximum temperature in the fixed bed initially increases rapidly while the main reaction zone moves slowly backwards out of the reactor. An explanation for this behavior in terms of the heat balance around a migrating combustion zone was first derived by Wicke and Vortmeyer [45]. In a migrating reaction zone that moves through an adiabatic fixed-bed reactor with velocity w due to lowering of the feed temperature to T^0 (Fig. 36), gas flows from the left into the reaction zone with velocity $v - w$ and temperature T^0 (v = gas velocity), while from the right catalyst enters the reaction zone with velocity w and temperature $T^0 + \Delta T$ and reacted gas leaves with velocity $v - w$ and the same temperature. The energy balance for the migrating reaction zone is thus:

$$\underbrace{\varepsilon \cdot A \cdot (v - w) \cdot \varrho_G \cdot c_{pG} \cdot \Delta T_{ad}}_{\text{Heat of reaction introduced and released with the gas}} = \underbrace{\varepsilon \cdot A \cdot (v - w) \cdot \varrho_G \cdot c_{pG} \cdot \Delta T}_{\text{Latent heat taken up by the gas}} + \underbrace{(1 - \varepsilon) \cdot A \cdot w \cdot \varrho_s \cdot c_s \cdot \Delta T}_{\text{Latent heat released by the packing}} \quad (6.1)$$

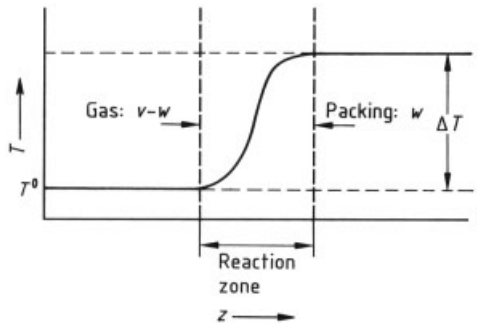

Figure 36. Derivation of the energy balance in a quasi-stationary migrating reaction zone in a fixed-bed reactor

It follows that for the temperature increase ΔT at the reaction front:

$$\Delta T = \frac{\Delta T_{ad}}{1 - \frac{(1-\varepsilon) \cdot w \cdot \varrho_s \cdot c_s}{\varepsilon \cdot (v - w) \cdot \varrho_G \cdot c_{pG}}} \quad (6.2)$$

Thus for $w > 0$ (i.e., reaction zone moving downstream), the temperature increase of the reaction front is always larger than the adiabatic temperature rise because the entering gas is heated by the hot catalyst (which is cooled) and because of the liberated heat of reaction.

The examples of autothermal reaction control with periodic reversal of the flow direction (Section 5.2) show that the temperature increase of the combustion zone can be conserably higher than the adiabatic temperature rise. According to Equation (6.2), the temperature increase in the combustion can even rise to arbitrarily high values when the denominator approaches zero; i.e.,

the migration velocity reaches the critical limit W_{crit}:

$$W_{\mathrm{crit}} = \frac{\varepsilon \cdot \varrho_{\mathrm{G}} \cdot c_{\mathrm{pG}}}{\varepsilon \cdot \varrho_{\mathrm{G}} \cdot c_{\mathrm{pG}} + (1-\varepsilon) \cdot \varrho_{\mathrm{s}} \cdot c_{\mathrm{s}}} \cdot v \tag{6.3}$$

In this case, the reaction front moves with exactly the velocity of a temperature front without reaction. If dispersion effects such as thermal conduction are neglected, the liberated heat of reaction can no longer be transported out of the reaction zone and thus accumulates there.

The migration velocity w of a reaction front can generally not be preset from outside, but is instead a result of the complex interaction of material and heat transport with the reaction, as is approximately described by fixed-bed reactor models. However, there is one noteworthy exception, which occurs when the catalyst in the main reaction zone is deactivated with exactly the critical velocity.

In the example of Figure 37, this deactivation was carried out deliberately by adding a catalyst poison (thiophene) to the feed at time $t = 0$. This results in a deactivation of the nickel catalyst by irreversible adsorption of thiophene, whereby the adsorption front moves through the bed with velocity w, pushing the reaction front before it. The accidental addition of a catalyst poison can thus lead to migrating combustion zones with very high maximum temperatures in both adiabatic and nonadiabatic reactors, if a migration velocity in the critical range is initiated.

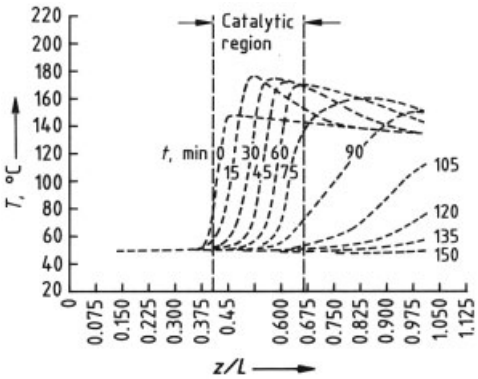

Figure 37. Migrating reaction front in an adiabatic fixed-bed reactor due to intentional catalyst poisoning (hydrogenation of benzene over nickel) [46]
At $t = 0$, a catalyst poison thiophene is added to the feed.

Particularly problematic is catalyst deactivation due to excess temperature deactivation of the catalyst because it results in a vicious circle: the catalyst is deactivated in the region of the main reaction zone due to an excessive maximum temperature. This leads to a migrating combustion zone in which the maximum temperature rises further, and so on. Impressive examples of this behavior in wall-cooled reactor tubes have been described by Blaum [47] and Emig [48].

Figure 38 shows measured profiles in a bed of thermally unstable nickel catalyst in the case of CO oxidation. Figure 39 shows the temperature profile for vinyl acetate synthesis over zinc acetate. Since the zinc acetate catalyst decomposes above 500 K, the result is a reaction front that moves downstream. The dip in the temperature profile (d) clearly indicates deactivation at the position of the original temperature maximum. In the past, this was sometimes cited as a good example of parametric sensitivity. The above discussion, however, shows that the cause of the "runaway" reaction should rather be attributed to the dynamic influence of the migrating combustion zone.

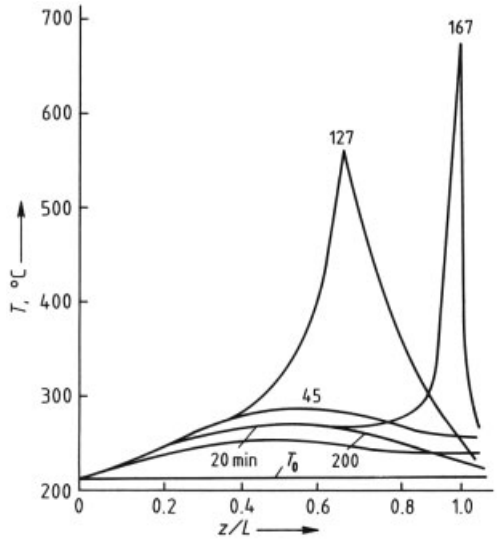

Figure 38. Migrating combustion zone in the case of CO oxidation over nickel, caused by thermal deactivation of the catalyst [47]

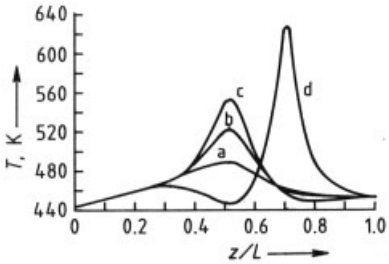

Figure 39. Migrating reaction front during the synthesis of vinyl acetate over zinc acetate, caused by thermal damage to the catalyst [48]
a) 60 min; b) 75 min; c) 100 min; d) 165 min

6.5. Safety Aspects

Because of the small mass storage capacity compared to liquid-phase reactors, the danger of sudden reaction of accumulated reactants in gas-phase fixed-bed reactors is low. Leaving out the peculiarities of individual cases, the following safety risks can be assumed for fixed-bed reactors:

1) Leaks which result in the release of large amounts of gas or vapor and the formation of explosive clouds
2) Leaks resulting in release of large amounts of liquid heat-transfer media (oils, salt melts)
3) Occurrence of ignitable or decomposable gas mixtures in the reactor
4) Melting of the reactor due to a runaway reaction

The safety aspects of liquid heat-transfer media are discussed in Section 4.3.1. Ignitable gas mixtures can arise particularly during partial oxidation reactions. They are especially critical where large packing-free volumes are present. This is the case in the inflow and outflow hoods of the reactor, while in the reactor tubes the catalyst packing dampens the propagation of a flame front due to its heat capacity. Complete avoidance of ignitable mixture is generally not possible in partial oxidations because at least during mixing of the gas streams prior to the reactor the ignition limit is exceeded locally. Nevertheless, in the past, operation of fixed-bed reactors with ignitable mixtures was avoided, either by dilution with inert gas or by operating in the nonstoichiometric range. The former requires additional efforts for heating, cooling, and separation of the inert gas, while the latter gives only low conversions of the reactants in a single pass. New developments in partial oxidation therefore aim for stoichiometric operation in the ignitable range [51]. A prerequisite for this is pressure-resistant construction with check valves and flame barriers so a possible that ignition is confined to the interior of the reactor.

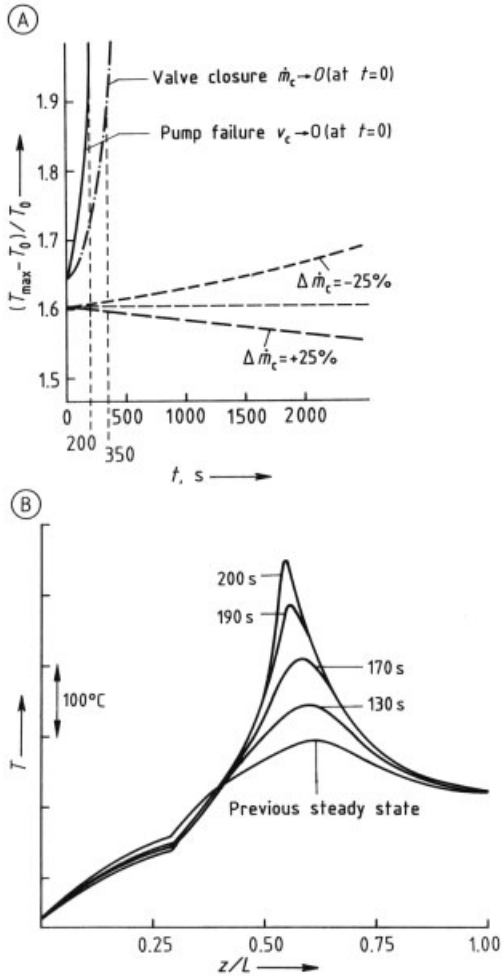

Figure 40. Effect of disturbances in the coolant circuit of a multitubular reactor (Fig. 41) with a partial oxidation reaction [49]
A) Profile of maximum temperature with time after change of the valve position ($\Delta\dot{m}_c$), after pump failure ($v_c \to 0$), and after valve closure ($\dot{m}_c \to 0$); B) Temperature profile after failure of the coolant circulation pump

Melting of reactor tubes during runaway reaction is only to be feared in multitubular reactors

if the respective tube is not surrounded by heat-transfer medium. Thus, appropriate design must ensure that running dry of reactor tubes cannot occur. In the case of corrosive reaction gases, apparatus for the detection of leaks caused by corrosion must be provided, particularly when pressurized or boiling water is used as coolant.

If the coolant circulation is interrupted during an exothermic reaction, the reactor must be shut down. Nevertheless, the heat storage capacity of the catalyst and the heat-transfer medium surrounding the tubes allows sufficient reaction time for controlled shutdown of the reactor. This is shown by simulation calculations for a multitubular reactor (Fig. 40) with a partial oxidation reaction. It was assumed that at $t = 0$ either the circulating pump fails or that the regulating valve in the pipe to the steam generator closes. As can be seen in Figure 40, there is a time span of 200 or 350 s, respectively before the maximum temperature in the catalyst reaches unpermissably high values.

In large-volume adiabatic fixed-bed reactors, local temperature maxima (hot spots) can form due to nonuniform distribution of the flow or inhomogeneities in the catalyst packing. They can be intensified by front migration phenomena (Section 6.4) and can lead to thermal damage of the reactor. This is particularly true of runaway-prone reactions such as partial hydrogenation with a large excess of hydrogen, which can change from selective hydrogenation to complete methanation. This can be detected by online analysis of methane in the product.

6.6. Control of Fixed-Bed Reactors

The control of fixed-bed reactors for chemical synthesis generally only involves maintaining the operating conditions at the optimum values. For runaway-prone reactions, this is supplemented by an extensive measurement and control system with built-in redundancy for giving alarm and shutting down the reactor. Slow changes in catalyst activity during multistage adiabatic reactors are corrected by "manual" adjustment of the set value for the feed temperature of each stage in multitubular reactors or by the adjustment of the heat-transfer medium temperature. The target is usually to keep the desired conversion constant. Mass flow control, feed stoichiometry control, control of the total pressure (generally in the reactor outlet), as well as feed temperature and heat-transfer medium temperature control are therefore the most important automatic control circuits in fixed-bed reactors.

In the following, some special points concerning the cooling temperature control of multitubular fixed-bed reactors are considered. In a multitubular reactor with molten-salt cooling (Fig. 41) the cooling temperature is controlled by a control unit in the connecting channel to the steam generator. In this way, varying amounts of colder salt melt can be introduced into the circulation system, with effective mixing taking place in the circulating pump. This cooling circuit can be unstable [37]. Since the reaction rate or liberated heat $\dot{Q}_{\mathrm{E}}$ increases exponentially with increasing cooling temperature, while the heat removed in the steam generator $\dot{Q}_{\mathrm{A}}$ increases only linearly, the conditions shown in Figure 42 may occur in the desired state. On a slight increase of the cooling temperature above $T_{\mathrm{C,\,set}}$, the cooling circuit warms up since $\dot{Q}_{\mathrm{E}} > \dot{Q}_{\mathrm{A}}$, and on dropping slightly below $T_{\mathrm{C,\,set}}$ cools down because $\dot{Q}_{\mathrm{A}} > \dot{Q}_{\mathrm{E}}$. Because of the large thermal inertia, however, the unstable operating point can easily be stabilized via the control TC (Fig. 41).

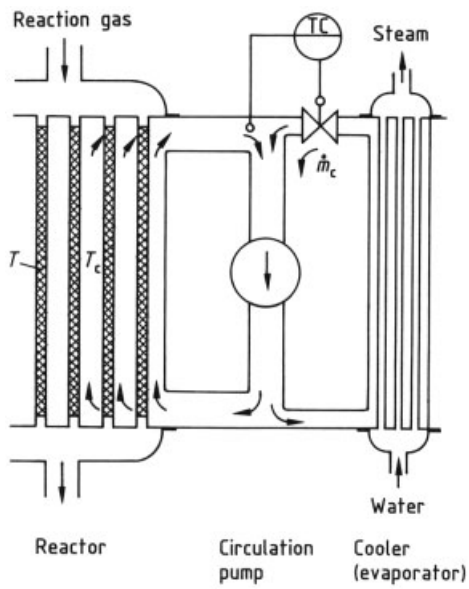

Figure 41. Temperature control in a fixed-bed reactor with molten-salt cooling

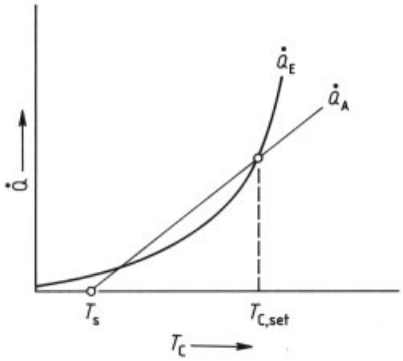

Figure 42. Heat generation $\dot{Q}_E$ and heat removal $\dot{Q}_A$ for the cooled multitubular reactor of Figure 41 as a function of coolant temperature [37]

To obtain optimum conversion, attempts are usually made to operate with a cooling temperature so high that a previously determined maximum temperature in the reactor is just reached but not exceeded. The maximum temperature can be measured by equipping several tubes of the bundle with multiple thermoelements which allow simultaneous measurement of, for example, 10 – 20 temperatures along the tube length. This allows regulation of the maximum temperature, which influences the set point of the cooling temperature in a cascade circuit. However, determining the maximum temperature from various discrete measurement points is not trivial [50]. Therefore, methods for the model-based measurement techniques should be used here.

In the case of cooling with a boiling medium, a multitubular reactor generally has the structure shown in Figure 43. The cooling temperature here corresponds to the boiling temperature in the vapor drum, which is controlled by a control unit in the saturated vapor line. The cooling temperature control is thus performed by controlling the saturated vapor pressure in the vapor drum. The dynamics of the natural convection cooling circuit has a particular effect on the behavior of this control loop. Thus rapid lowering of the cooling temperature via the drum pressure by opening the saturated vapor valve leads to increased boiling of the coolant in the circuit. If the liquid boils into the downpipe, the natural circulation is temporarily interrupted. Furthermore, the increase in volume associated with boiling pushes liquid into the vapor drum, which can result in liquid entering the vapor line. This is discussed in detail in [49].

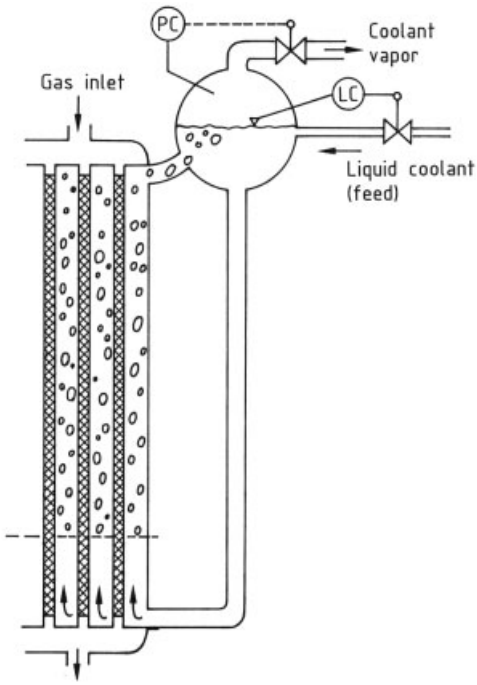

Figure 43. Multitubular reactor with evaporative cooling

If the fixed bed to be regulated is part of a supply or off-gas treatment plant that is subject to large variations in feed concentration and throughput, then a fixed control is difficult, mainly because of the complex wrong-way behavior (see Section 6.4). Hence, in practice a catalyst with a wide operating range is used and the reactor is made larger.

7. References

1. L. Riekert, *Appl. Catal.* **15** (1985) 89 – 102.
2. V. Kottke, H. Blenke: "Verfahren zur Bestimmung örtlicher Stoff- und Wärmeübertragung an beliebig geformten Oberflächen," *Verfahrenstechnik* (Mainz) 16 (1982) 504 – 509.
3. Verein Deutscher Ingenieure: *VDI-Wärmeatlas,* 5th ed., Section Gh, VDI-Verlag, Düsseldorf 1988.
4. E. Tsotsas: "Über die Wärme- und Stoffübertragung in durchströmten Festbetten," *VDI-Fortschr. Ber., Reihe 3,* 1990, no. 223.

5. G. Gaiser, V. Kottke: "Wärme- und Stoffübergang in Katalysatoren mit regelmäßiger Formgebung," *Chem. Ing. Tech.* **61** (1989) 729 – 731.
6. R. K. Shah, A. L. London: "Laminar Flow Forced Convection in Ducts," *Advances in Heat Transfer,* suppl. 1, Academic Press, New York 1978.
7. S. Quest, D. Mewes: "Der Einfluß des Temperaturfeldes auf den Stoffaustausch in laminar durchströmten Rohrreaktoren," *Wärme Stoffübertrag.* **32** (1988) 355 – 363.
8. G. Gaiser, V. Kottke: "Flow Phenomena and Local Heat and Mass Transfer in Corrugated Passages," *Chem. Eng. Technol.* **12** (1989) 400 – 405.
9. Verein Deutscher Ingenieure: *VDI-Wärmeatlas,* Section Le 1, VDI-Verlag, Düsseldorf 1988.
10. Verein Deutscher Ingenieure: *VDI-Wärmeatlas,* Section Mh 1, VDI-Verlag, Düsseldorf 1988.
11. G. Eigenberger, U. Nieken: "Katalytische Abluftreinigung: Verfahrenstechnische Aufgaben und neue Lösungen," *Chem. Ing. Tech.* **63** (1991) 781 – 791.
12. D. Vortmeyer, P. Winter: "Verbesserung der Analyse von Festbettreaktoren durch die Berücksichtigung von Porositäts- und Strömungsverteilung," *Chem. Ing. Tech.* **55** (1983) 312 – 313.
13. T. Daszkowski, G. Eigenberger: "A Reevaluation of Fluid Flow, Heat Transfer and Chemical Reaction in Catalyst Filled Tubes," *Chem. Eng. Sci.* **47** (1992) 2245 – 2250.
14. U. Zardi et al.: "A Novel Reactor Design for Ammonia and Methanol Synthesis," *IV. Intern. Conference Fertilizer Technology,* London 1981; Hydrocarbon Process (1982) 129 – 133.
15. "Umwelttechnik, Sekundärmaßnahmen zur Minderung der NO_x-Emission," *Steinmüller Schriftenreihe, L + C.* , Steinmüller, 1985.
16. C. B. Allen, G. J. Janz: "Molten Salts Safety and Hazards: an Annotated Bibliography," *J. Hazard. Mater.* **4** (1980) 145 – 175.
17. G. Eigenberger, W. Ruppel: "Probleme der Modellbildung bei technischen Festbettreaktoren," *Chem. Ing. Tech.* **57** (1985) 181 – 190.
18. A. Stankiewicz: "Advances in Modelling and Design of Multitubular Fixed-Bed Reactors," *Chem. Eng. Tech.* **12** (1989) 113 – 130, 170 – 175.
19. M. S. G. Cristina et al.: "Modelling Multitubular Catalytic Reactors: the Influence of Shell Side Flow," *Chem. Eng. Sci.* **47** (1992) 2565 – 2570.
20. T. F. Degnan, J. Wei: "The Co-Current Reactor Heat Exchanger," *AIChE J.* **25** (1979) 338 – 344.
21. A. S. Lopez, H. I. de Lasa, J. A. Porras: "Parametric Sensitivity of a Fixed-Bed Catalytic Reactor," *Chem. Eng. Sci.* **36** (1981) 285 – 291.
22. D. O. Borio, V. Bucala, J. A. Orejas, J. A. Porras: "Cocurrently Cooled Fixed Bed Reactors: A Simple Approach to Optimal Cooling Design," *AIChE J.* **35** (1989) 1899 – 1902.
23. J. C. Pirkle, I. E. Wachs: "Activity Profiling in Catalytic Reactors," *Chem. Eng. Progr.* (1987) 29 – 34.
24. P. H. Calderbank, A. D. Caldwell, G. Ross: "The Diluted Catalyst Fixed Bed Reactor for Exothermic Catalytic Reactions," *Chemie et Industrie-Gènie Chimique* **101** (1969) 215 – 230.
25. G. Eigenberger: "Practical Problems in the Modelling of Chemical Reactions in Fixed Bed Reactors," *Chem. Eng. Process.* **18** (1984) 55 – 65.
26. E. Klapp: *Apparate- und Anlagentechnik,* Springer-Verlag, Berlin 1980.
27. Verein Deutscher Ingenieure: *VDI-Wärmeatlas,* Section Gf, LD, VDI-Verlag, Düsseldorf 1988.
28. U. Lahne, R. Lohmüller: "Schüttschichtreaktoren mit gewickelten Kühlrohren, eine konstruktive Neuentwicklung zur Durchführung exothermer katalytischer Prozesse," *Chem. Ing. Tech.* **58** (1986) 212 – 215.
29. Deggendorfer Werft, DE 2 207 166, 1972 (O. Wanka et al.).
30. Deggendorfer Werft, DE 2 201 528, 1972 (O. Wanka et al.).
31. P. Grimm: "Six Years Successful Operation of Linde Isothermal Reactor," *Linde Reports on Science and Technology* **49** (1991) 57 – 59.
32. G. K. Boreskow, G. A. Bunimovich, Y. S. Matros, A. A. Ivanov: "Catalytic Processes Carried Out Under Non-Steady-State Conditions: Switching the Direction of the Feed," Kinetika i Kataliz **23** (1982) 402 – 406; *Int. Chem. Eng.* **22** (1982) 335 – 342.
33. Y. S. Matros: "Performance of Catalytic Processes under Unsteady Conditions," *Chem. Eng. Sci.* **45** (1990) 2097 – 2102.

34. Y. S. Matros: "Catalytic Processes under Unsteady-State Conditions," *Studies in Surface Science and Catalysis,* vol. 43, Elsevier, Amsterdam 1989.
35. G. F. Froment: *Front. Chem. Reakt. Eng.* **1** (1984) pp 12 – 38.
36. M. Morbidelli, A. Varma: "Parametric Sensitivity in Fixed-Bed Catalytic Reactor," *AIChE J.* **32** (1986) 297 – 306; *Chem. Eng. Sci.* **46** (1991) 3330 – 3332.
37. G. Eigenberger, H. Schuler: "Reactor Stability and Safe Reaction Engineering," *Chem. Ing. Tech.* **58** (1986) 655 – 665; *Int. Chem. Eng.* **29** (1989) 12 – 25.
38. G. Eigenberger: "Stabilität und Dynamik heterogenkatalytischer Reaktionssysteme," *Chem. Ing. Tech.* **50** (1978) 924 – 933.
39. K. F. Jensen, W. H. Ray: "The Bifurcation Behavior of Tubular Reactors," *Chem. Eng. Sci.* **37** (1982) 199-222.
40. O. Kalthoff, D. Vortmeyer: *Chem. Eng. Sci.* **35** (1980) 1637 – 1643.
41. G. Eigenberger: "Influence of the Wall on the Dynamic Behavior of Homogeneous Tubular Reactors with a Highly Exothermic Reaction," *Chem. Ser.* **133** (1974) 477 – 488.
42. G. Eigenberger: "Zur Modellbildung und Dynamik des homogenen Rohrreaktors," *Chem. Ing. Tech.* **46** (1974) 11 – 19.
43. G. Eigenberger: "On the Dynamic Behavior of the Catalytic Fixed-Bed Reactor in the Region of Multiple Steady States – II," *Chem. Eng. Sci.* **278** (1972) 1917 – 1924.
44. H. von Doesburg, W. A. de Jong: "Transient Behavior of an Adiabatic Fixed-Bed Methanator," *Chem. Eng. Sci.* **31** (1976) 45 – 58.
45. E. Wicke, D. Vortmeyer: "Zündzonen heterogener Reaktionen in gasdurchströmten Körnerschichten," *Ber. Bunsenges.* **63** (1959) 145 – 152.
46. T. H. Price, J. B. Butt: "Catalyst Poisoning and Fixed-Bed Reactor Dynamics II," *Chem. Eng. Sci.* **32** (1977) 393 – 412.
47. E. Blaum: "Zur Dynamik des katalytischen Festbettreaktors bei Katalysatordesaktivierung," *Chem. Eng. Sci.* **29** (1974) 2263 – 2277.
48. G. Emig, H. Hofmann, U. Hoffmann, U. Fiand: "Experimental Studies on Runaway of Catalytic Fixed-Bed Reactors," *Chem. Eng. Sci.* **85** (1980) 249 – 257.
49. G. Eigenberger: "Modelling and Computer Simulation as Aids for Safe Reactor Operation," *Chem. Ing. Tech.* **51** (1979) 1105 – 1110; *Int. Chem. Eng.* **3** (1980) 211 – 216.
50. G. Eigenberger: "Temperaturregelung von chemischen Rohrreaktoren mit stark exothermer Reaktion," *Regelungstechnik* (1975) 118 – 126.
51. Deggendorfer Werft, Reactor Systems and their Advantages, Technical information, May 1989.

Fluidized-Bed Reactors

JOACHIM WERTHER, Hamburg University of Technology, Hamburg, Germany

Related Articles For other industrial reactors and their applications, see → Stirred-Tank and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

1. Introduction . . . 3
1.1. The Fluidization Principle . . . 3
1.2. Forms of Fluidized Beds . . . 3
1.3. Advantages and Disadvantages of the Fluidized-Bed Reactor . . . 4
2. Fluid-Mechanical Principles . . . 5
2.1. Minimum Fluidization Velocity . . . 5
2.2. Expansion of Liquid–Solid Fluidized Beds . . . 6
2.3. Fluidization Properties of Typical Bed Solids . . . 6
2.4. State Diagram of Fluidized Bed . . . 7
2.5. Gas Distribution . . . 8
2.6. Gas Jets in Fluidized Beds . . . 10
2.7. Bubble Development . . . 10
2.8. Elutriation . . . 12
2.9. Circulating Fluidized Beds . . . 13
2.9.1. Hydrodynamic Principles . . . 13
2.9.2. Local Flow Structure in Circulating Fluidized Beds . . . 15
2.9.3. Design of Solids Recycle System . . . 16
2.10. Cocurrent Downflow Circulating Fluidized Beds (Downers) . . . 16
2.11. Attrition of Solids . . . 17
3. Solids Mixing in Fluidized-Bed Reactors . . . 20
3.1. Mechanisms of Solids Mixing . . . 20
3.2. Vertical Mixing of Solids . . . 21
3.3. Horizontal Mixing of Solids . . . 22
3.4. Solids Residence-Time Properties . . . 22
3.5. Solids Mixing in Circulating Fluidized Beds . . . 22
4. Gas Mixing in Fluidized-Bed Reactors . . . 23
4.1. Gas Mixing in Bubbling Fluidized Beds . . . 23
4.2. Gas Mixing in Circulating Fluidized Beds . . . 23
5. Heat and Mass Transfer in Fluidized-Bed Reactors . . . 24
6. Gas-Solid Separation . . . 25
7. Injection of Liquid Reactants into Fluidized Beds . . . 26
8. Industrial Applications . . . 26
8.1. Heterogeneous Catalytic Gas-Phase Reactions . . . 27
8.2. Polymerization of Olefins . . . 29
8.3. Homogeneous Gas-Phase Reactions 30
8.4. Gas–Solid Reactions . . . 30
8.5. Applications in Biotechnology . . . 36
9. Modeling of Fluidized-Bed Reactors 37
9.1. Modeling of Liquid–Solid Fluidized-Bed Reactors . . . 37
9.2. Modeling of Gas–Solid Fluidized-Bed Reactors . . . 37
9.2.1. Bubbling Fluidized-Bed Reactors . . . 37
9.2.2. Circulating Fluidized-Bed Reactors . 38
9.3. New Developments in Modeling Fluidized-Bed Reactors . . . 40
9.3.1. Computational Fluid Dynamics . . . 40
9.3.2. Modeling of Fluidized-Bed Systems . 41
10. Scale-up . . . 42
11. References . . . 43

The article gives an overview on fundamentals and applications of fluidized-bed technology. Some basic fluid-mechanical principles—minimum fluidization velocity, status diagram, gas distribution, bubble development, and solids entrainment and elutriation—are presented at the beginning, followed by a detailed discussion of circulating fluidized beds, downers, and attrition effects in fluidized-bed systems. Sections on solids mixing, gas mixing, and heat and mass transfer in fluidized beds complete the fundamentals section, which is followed by an overview on industrial applications. Fluidized-bed processes for heterogeneous catalytic gas-phase reactions, for the polymerization of olefins, for homogenous gas-phase reactions, gas–solid reactions and applications in biotechnology are described in detail. A final chapter is devoted to the modeling of fluidized-bed reactors.

© 2007 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_239.pub2

Symbols

a	volume-specific mass-transfer area between bubble and suspension phases, m^{-1}
A_0	cross-sectional area of orifice, m^2
Ar	Archimedes number, defined by Equation (5)
A_t	cross-sectional area of reactor, m^2
b	parameter def. by Equation (54)
c_v	solids volume concentration
c_b	bubble attrition rate constant, defined by Equation (50), s^2/m^4
c_c	cyclone attrition rate constant defined by Equation (51), s^2/m^3
c_j	jet attrition rate constant, defined by Equation (52),s^2/m^3
C_b	concentration in bubble phase, $kmol/m^3$
C_d	concentration in suspension phase, $kmol/m^3$
d_o	orifice diameter, m
d_p	Sauter diameter, defined by Equation (6), m
d_{pi}	diameter of particle size class i, m
d_t	bed diameter, m
d_v	local bubble volume equivalent sphere diameter, m
d_{v0}	initial bubble diameter, m
D	coefficient of molecular diffusion, m^2/s
D_{sh}	lateral solids dispersion coefficient, m^2/s
D_{sv}	vertical solids dispersion coefficient, m^2/s
Fr_p	Froude number, defined by Equation (29)
G_s	solids mass flow rate, based on reactor cross-sectional area, kg m^{-2} s^{-1}
h	height above distributor level, m
h_o	height above distributor where bubbles are forming, m
h_{gs}	gas-to-solid heat transfer coefficient, W $m^{-2}K^{-1}$
h_{wb}	wall-to-bed heat transfer coefficient, W $m^{-2}K^{-}$
H	expanded bed height, m
H_{mf}	bed height at minimum fluidization, m
k_G	mass-transfer coefficient, m/s
L	jet length, m
m_a	mass of elutriated solids, kg
$\dot{m}_{att}$	mass flow due to attrition, kg/s
m_b	bed mass, kg
$\dot{m}_s$	solids mass flow, g/s
n_p	number of passages through cyclone
p	pressure, Pa
$Pe_{r,c}$	Péclet number, defined by Equation (43)
Q_3	cumulative mass distribution
r_a	attrition rate, defined by Equation (33), s^{-1}
r_j	reaction rate, based on catalyst mass, kmol kg^{-1} s^{-1}
Re	Reynolds number
S_v	volume-specific surface area of particles, m^{-1}
t	time, s
TDH	transport disengaging height, m
u	superficial fluidizing velocity, m/s
u_b	local bubble rise velocity, m/s
u_c	velocity at cyclone inlet, m/s
u_{mf}	superficial minimum fluidizing velocity, m/s
u_o	jet velocity at orifice, m/s
u_{sl}	slip velocity, defined by Equation (27), m/s
u_t	single particle terminal velocity, m/s
$\dot{V}_b$	visible bubble flow, based on bed area, m^3 m^{-2} s^{-1}
$\dot{V}_{mf}$	minimum fluidizing flow rate, m^3/s
$\dot{V}_o$	flow rate of gas issuing from orifice, m^3/s
x_i	mass fraction of particle size fraction i in bed material
α	velocity ratio, defined by Equation (14)
Δp_d	pressure drop of the gas distributor, Pa
ε	bed porosity
ε_b	local bubble gas holdup
ε_i	porosity of catalyst particle
ε_{mf}	bed porosity at minimum fluidization
$\varkappa^*$	elutriation rate constant, kg m^{-2} s^{-1}
λ	average life time of a bubble, s
μ	solid-to-gas mass flow ratio
ν	kinematic viscosity, m^2/s
ν_{ij}	stoichiometric number of species i in reaction j
ϱ_f	fluid density, kg/m^3
ϱ_s	solids density, kg/m^3
ϑ	stress history parameter, defined by Equation (54)
ϑ_b	parameter, defined by Equation (23)
ψ	pressure ratio, defined by Equation (28)

1. Introduction

1.1. The Fluidization Principle

In fluidization an initially stationary bed of solid particles is brought to a "fluidized" state by an upward stream of gas or liquid as soon as the volume flow rate of the fluid exceeds a certain limiting value $\dot{V}_{\mathrm{mf}}$ (where mf denotes minimum fluidization). In the fluidized bed, the particles are held suspended by the fluid stream; the pressure drop Δp_{fb} of the fluid on passing through the fluidized bed is equal to the weight of the solids minus the buoyancy, divided by the cross-sectional area A_{t} of the fluidized-bed vessel (Fig. 1):

$$\Delta p_{\mathrm{fb}} = \frac{A_{\mathrm{t}} \cdot H \cdot (1-\varepsilon) \cdot (\varrho_{\mathrm{s}} - \varrho_{\mathrm{f}}) \cdot g}{A_{\mathrm{t}}} \tag{1}$$

In Equation (1), the porosity ε of the fluidized bed is the void volume of the fluidized bed (volume in interstices between grains, not including any pore volume in the interior of the particles) divided by the total bed volume; ϱ_{s} is the solids apparent density; and H is the height of the fluidized bed.

In many respects, the fluidized bed behaves like a liquid. The bed can be stirred like a liquid; objects of greater specific gravity sink, whereas those of lower specific gravity float; if the vessel is tilted, the bed surface resumes a horizontal position; if two adjacent fluidized beds with different bed heights are connected to each other, the heights become equal; and the fluidized bed flows out like a liquid through a lateral opening. Particularly advantageous features of the fluidized bed for use as a reactor are excellent gas–solid contact in the bed, good gas–particle heat and mass transfer, and high bed–wall and bed–internals heat-transfer coefficients.

The fluidization principle was first used on an industrial scale in 1922 for the gasification of fine-grained coal [1]. Since then, fluidized beds have been applied in many industrially important processes. The present spectrum of applications extends from a number of physical processes, such as cooling–heating, drying, sublimation–desublimation, adsorption–desorption, coating, and granulation, to many heterogeneous catalytic gas-phase reactions as well as noncatalytic reactions.

What follows is a survey of the fluid mechanical principles of fluidization technology, gas and solid mixing, gas–solid contact in the fluidized bed, typical industrial applications, and approaches to modeling fluidized-bed reactors. Further information is given in textbooks (e.g., [2]) and monographs (e.g., [3 – 8]). Summary treatments can also be found in [9 – 19]. Other useful literature includes reports of the Engineering Foundation Conferences on Fluidization [20 – 22], the Circulating Fluidized Bed Conferences (e.g., [23 – 25], and—for use of the fluidized bed in energy technology—the Fluidized Bed Combustion Conferences (e.g., [26 – 28]).

1.2. Forms of Fluidized Beds

As the volume flow rate $\dot{V}$ or the superficial velocity $u = \dot{V}/A_{\mathrm{t}}$ of the fluid increases beyond

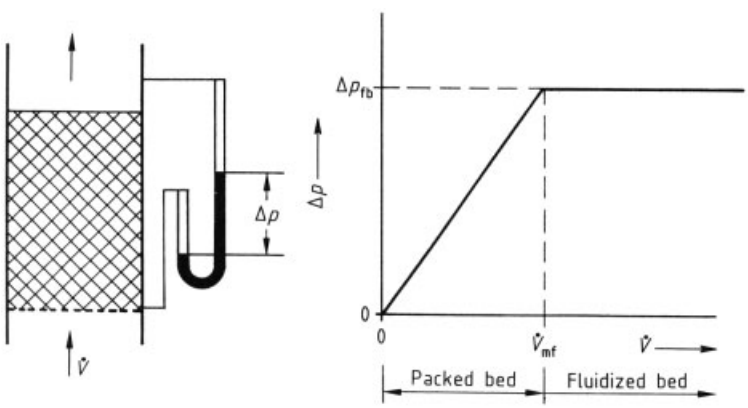

Figure 1. Pressure drop in flow through packed and fluidized beds

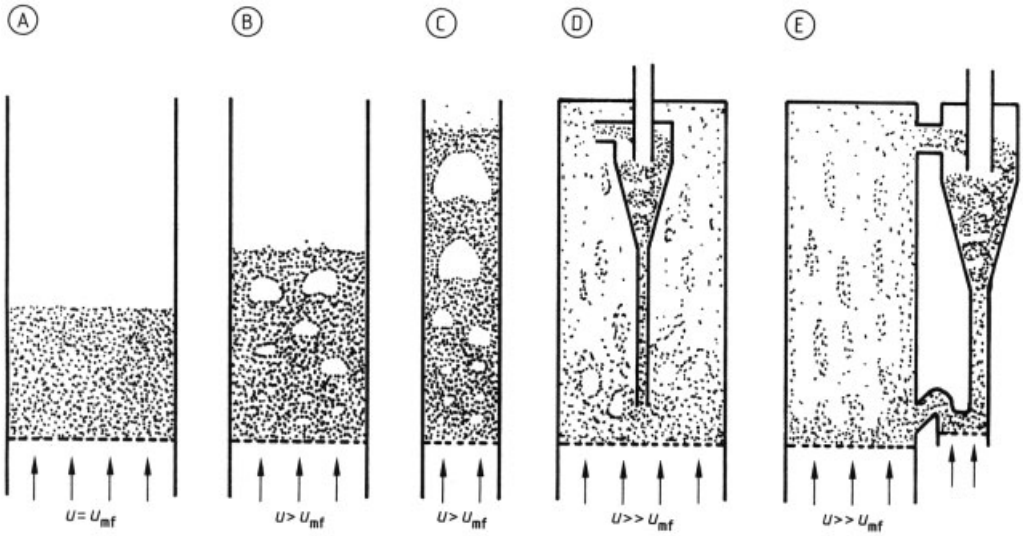

Figure 2. Forms of gas–solids fluidized beds

the value $\dot{V}_{\mathrm{mf}}$ or u_{mf} (Fig. 2 A) corresponding to the minimum fluidization point, one of two things happens: in *fluidization with a liquid*, the bed begins to expand uniformly; in *fluidization with a gas*—a process of greater industrial importance and the one discussed almost exclusively in the following material—virtually solids-free gas bubbles begin to form (Fig. 2 B). The local mean bubble size increases rapidly with increasing height above the grid because of coalescence of the bubbles. If the bed vessel is sufficiently narrow and high, the bubbles ultimately fill the entire cross section and pass through the bed as a series of gas slugs (Fig. 2 C). As the gas velocity increases further, more and more solids are carried out of the bed, the original, sharply defined surface of the bed disappears, and the solids concentration comes to decrease continuously with increasing height. To achieve steady-state operation of such a "turbulent" fluidized bed (Fig. 2 D), solids entrained in the fluidizing gas must be collected and returned to the bed. The simplest way to do this is with a cyclone integrated into the bed vessel and a standpipe dipping into the bed. A further increase in gas velocity finally leads to the circulating fluidized bed (Fig. 2 E), which is characterized by a much lower average solids concentration than the previous systems. The high solids entrainment requires an efficient external solids recycle system with a specially designed pressure seal (shown as a siphon in Fig. 2 E).

1.3. Advantages and Disadvantages of the Fluidized-Bed Reactor

The major advantages of the (gas–solid) fluidized bed as a reaction system include

1) Easy handling and transport of solids due to liquid-like behavior of the fluidized bed
2) Uniform temperature distribution due to intensive solids mixing (no hot spots even with strongly exothermic reactions)
3) Large solid–gas exchange area by virtue of small solids grain size
4) High heat-transfer coefficients between bed and immersed heating or cooling surfaces
5) Uniform (solid) product in batchwise process because of intensive solids mixing

Set against these advantages are the following disadvantages:

1) Expensive solids separation or gas purification equipment required because of solids entrainment by fluidizing gas
2) As a consequence of high solids mixing rate, nonuniform residence time of solids, backmixing of gas, and resulting lower conversion

3) In catalytic reactions, undesired bypass or broadening of residence-time distribution for reaction gas due to bubble development
4) Erosion of internals and attrition of solids (especially significant with catalysts), resulting from high solids velocities
5) Possibility of defluidization due to agglomeration of solids
6) Gas–solid countercurrent motion possible only in multistage equipment
7) Difficulty in scaling-up

Table 1 compares the fluidized-bed reactor with alternative gas–solid reaction systems: fixed-bed, moving-bed, and entrained-flow reactors.

2. Fluid-Mechanical Principles

2.1. Minimum Fluidization Velocity

The minimum fluidization point, which marks the boundary between the fixed- and the fluidized-bed conditions, can be determined by measuring the pressure drop Δp across the bed as a function of volume flow rate $\dot{V}$ (Fig. 1). Measurement should always be performed with decreasing gas velocity, by starting in the fluidized condition.

Only for very narrow particle-size distributions, however, does a sharply defined minimum fluidization point occur. The broad size distributions commonly encountered in practice exhibit

Table 1. Comparison of gas–solid reaction systems [2, 18]

Characteristics	Fixed bed	Moving bed	Fluidized bed	Entrained flow
Suitability for heterogeneous catalytic gas-phase reactions	only for catalyst that is deactivated very slowly	can also be used with catalyst that is rapidly deactivated		
	catalyst attrition negligible	catalyst attrition may be critical, depending on operating conditions		
	plug flow of gas ensures high gas conversion		backmixing of gas due to mixing motion of solids and bubble-gas bypass lead to lower conversion	gas in virtually plug flow; high conversion possible
Suitability for gas–solid reactions	unsuitable for continuous processes; batchwise operation yields nonuniform product	for uniform feed particle size with low fines content; large reactor capacities possible	no special requirements for feed particle-size distribution; high fines content also possible; continuous operation yields uniform product	possible for fast reactions; recycle of unreacted fines often difficult
Temperature distribution	danger of hot spots with exothermic reactions	temperature gradients can be held within limits by virtue of high solids circulation and high gas throughput	high solids mixing ensures uniform temperature distribution in bed; temperature control by heat exchangers immersed in bed or by admission and removal of solids	axial temperature gradients can be held within limits by high solids circulation
Heat supply and removal, heat exchange	poor heat exchange; heat transport limits scale-up	poor heat exchange; due to high heat capacity of solids transport of large quantities of heat by way of circulating solids	very efficient heat exchange, good heat transport by solids	properties intermediate between fluidized bed and moving bed
Particle size	large pellets (ca. 8–20 mm), as uniform as possible; no fines	medium size (ca. 2–6 mm) and uniform; no fines	broad particle-size distribution (ca. 0.02–6 mm); high fines content acceptable	fine (0.02–0.5 mm), with narrow particle-size distribution

a blurred range; conventionally, the minimum fluidization point is defined as the intersection of the extrapolated fixed-bed characteristic with the line of constant bed pressure drop typical of the fluidized bed (Fig. 1).

The measurement technique already contains the possibility of calculating the minimum fluidization velocity u_{mf}: The pressure drop in flow through the polydisperse fixed bed at the point $u = u_{\mathrm{mf}}$, given, for example, by the Ergun relation [29] (→ Fluid Mechanics), is set equal to the fluidized-bed pressure drop given by Equation (1). From the Ergun relation

$$\frac{\Delta p}{h} = 4.17 \cdot S_v^2 \cdot \frac{(1-\varepsilon)^2}{\varepsilon^3}\,\eta u + 0.29 S_v \cdot \frac{1-\varepsilon}{\varepsilon^3}\,\varrho_{\mathrm{f}}\,u^2$$

it follows

$$u_{\mathrm{mf}} = 7.14\,(1-\varepsilon_{\mathrm{mf}})\,\nu \cdot S_v \cdot \left[\sqrt{1+0.067\,\frac{\varepsilon_{\mathrm{mf}}^3}{(1-\varepsilon_{\mathrm{mf}})^2} \cdot \frac{(\varrho_{\mathrm{s}}-\varrho_{\mathrm{f}})\cdot g}{\varrho_{\mathrm{f}}\,\nu^2} \cdot \frac{1}{S_v^3}} - 1\right] \tag{2}$$

Accordingly, to calculate u_{mf}, the characteristics of the gas (ϱ_{f}, ν), the density ϱ_{s} of the particles, the porosity $\varepsilon_{\mathrm{mf}}$ of the bed at minimum fluidization, and the volume-specific surface area S_v of the solids must be known. The specific surface area defined by

$$S_v = \frac{\text{surface area of all particles in the bed}}{\text{volume of all particles in the bed}}$$

(this takes into account only the external area, which governs hydraulic resistance, not the pore surface area as in porous catalysts) cannot be determined very exactly in practice. Hence u_{mf} should not be calculated on the basis of the measured particle-size distribution of a representative sample of the bed solids; instead, it is better measured directly. Equation (2) can be employed advantageously to calculate u_{mf} in an industrial-scale process on the basis of minimum fluidization velocities measured in the laboratory under ambient conditions [30].

An equation from Wen and Yu [31] can be used for approximate calculations:

$$Re_{\mathrm{mf}} = 33.7\left(\sqrt{1+3.6\times10^{-5}\cdot Ar} - 1\right) \tag{3}$$

where

$$Re_{\mathrm{mf}} = \frac{u_{\mathrm{mf}} d_{\mathrm{p}}}{\nu} \tag{4}$$

$$Ar = \frac{g d_{\mathrm{p}}^3}{\nu^2} \cdot \frac{\varrho_{\mathrm{s}}-\varrho_{\mathrm{f}}}{\varrho_{\mathrm{f}}} \tag{5}$$

Here the surface mean or Sauter diameter calculated from the mass–density distribution $q_3\,(d)$ of the particle diameters

$$d_{\mathrm{p}} = \frac{1}{\int\limits_{d_{\min}}^{d_{\max}} d^{-1} \cdot q_3\,(d)\,\mathrm{d}\,(d)} \tag{6}$$

should be used for the characteristic particle diameter d_{p}.

Both the Ergun approach and the Wen and Yu simplification have been confirmed experimentally over a wide range of parameters. More recently, Vogt et al. [32] found that Equations (2) and (3) are also applicable to high-pressure fluidized beds in which the fluid is under supercritical conditions

2.2. Expansion of Liquid–Solid Fluidized Beds

The uniform expansion of a bed on fluidization with a liquid can be described by

$$\frac{u}{u_{\mathrm{t}}} = \varepsilon^n \tag{7}$$

according to Richardson and Zaki [33]. Here u_{t} is the terminal velocity of isolated single particles; the exponent n is given as follows, provided the particle diameter is much smaller than that of the vessel:

$$n = \begin{cases} 4.65 & 0 < Re_{\mathrm{t}} \leq 0.2 \\ 4.4 \cdot Re_{\mathrm{t}}^{-0.03} & 0.2 < Re_{\mathrm{t}} \leq 1 \\ 4.4 \cdot Re_{\mathrm{t}}^{-0.1} & 1 < Re_{\mathrm{t}} \leq 500 \\ 2.4 & 500 < Re_{\mathrm{t}} \end{cases} \tag{8}$$

The Reynolds number used above is calculated via the single-particle terminal velocity u_{t}:

$$Re_{\mathrm{t}} = \frac{u_{\mathrm{t}} \cdot d_{\mathrm{p}}}{\nu} \tag{9}$$

2.3. Fluidization Properties of Typical Bed Solids

In fluidization with gases, solids display characteristic differences in behavior that can also affect the operating characteristics of fluidized-bed reactors. Geldart has proposed an empirically based classification of solids into four groups (A to D) by fluidization behavior [34].

The parameters employed are those crucial for fluidization properties: the mean particle diameter (d_p) and the density difference ($\varrho_s - \varrho_f$) between solid and fluid. Figure 3 shows the Geldart diagram with the interclass boundaries theoretically established by MOLERUS [35].

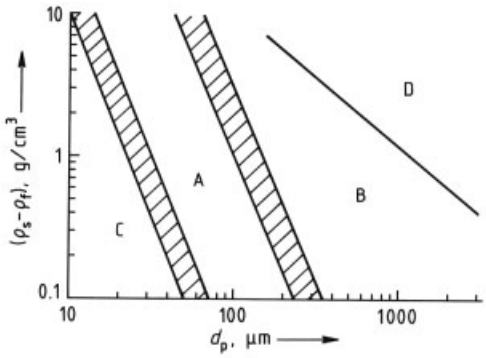

Figure 3. Geldart diagram (boundaries according to MOLERUS [35])
For explanation see text

Solids of group C are very fine-grained, cohesive powders (e.g., flour, fines from cyclones and electrostatic filters) that virtually cannot be fluidized without fluidization aids. The adhesion forces between particles are stronger than the forces that the fluid can exert on the particles. Gas flow through the bed forms channels extending from the grid to the top of the bed, and the pressure drop across the bed is lower than the value from Equation (1). Fluidization properties can be improved by the use of mechanical equipment (agitators, vibrators) or flowability additives, e.g., Aerosil.

Solids of group A have small particle diameters (ca. 0.1 mm) or low bulk densities; this class includes catalysts used e.g., in the fluidized-bed catalytic cracker. As the gas velocity u increases beyond the minimum fluidization point, the bed of such a solid first expands uniformly until bubble formation sets in at $u = u_{mb} > u_{mf}$. The bubbles grow by coalescence but break up again after passing a certain size. At a considerable height above the gas distributor grid, a dynamic equilibrium is formed between bubble growth and breakup. If the gas flow is cut off abruptly, the gas storage capacity of the fluidized suspension causes the bed to collapse rather slowly.

Group B solids have moderate particle sizes and densities. Typical representatives of this group are sands with mean particle diameters between ca. 0.06 and 0.5 mm. Bubble formation begins immediately above the minimum fluidization point. The bubbles grow by coalescence, and growth is not limited by bubble splitting. When the gas flow is cut off abruptly, the bed collapses quickly.

Group D includes solids with large particle diameters or high bulk densities; examples are sands with average particle diameters > 0.5 mm. Bubbles begin to form just above the minimum fluidization point, but the character of bubble flow is markedly different from that in group B solids: group D solids are characterized by the formation of "slow" bubbles (Section 2.7). On sudden stoppage of the gas flow, the bed also collapses suddenly.

2.4. State Diagram of Fluidized Bed

Whereas the onset of the fluidized state can be described by the minimum fluidization velocity, the bed operating range and the gas velocity needed to create a given fluidized state can be estimated with the help of the fluidized-bed state diagram (Fig. 4) devised by REH [36]. This plot shows the fluid mechanical resistance characteristics of the fixed bed, fluidized bed, and pneumatic transport. The ordinate is the quantity

$$\frac{3}{4}\frac{u^2}{g \cdot d_p} \cdot \frac{\varrho_f}{(\varrho_s - \varrho_f)}$$

and the abscissa is the Reynolds number Re_p formed with the fluidization velocity u and the particle diameter d_p. The state parameter in the fluidized-bed region is the mean bed porosity ε. The use of the diagram is facilitated by an auxiliary grid with lines of constant M and constant Archimedes number. While the dimensionless groups plotted as ordinate and abscissa each contain both the particle diameter and the fluidization velocity, this is not the case with the parameters Ar and M defined by

$$Ar = \frac{g \cdot d_p^3}{\nu^2} \cdot \frac{(\varrho_s - \varrho_f)}{\varrho_f} \tag{10}$$

$$M = \frac{u^3}{g \cdot \nu} \cdot \frac{\varrho_f}{(\varrho_s - \varrho_f)} \tag{11}$$

The Reh status diagram can answer a number of practical questions. If, for example, the properties of the gas (ρ_f, ν) and the solid (d_p, ρ_s) and the fluidization velocity u are given, the calculation of Ar and Re_p yields, via the status point S in the diagram (Fig. 4), the average voidage ε in the fluidized bed. Taking the line M = const. through S at the intersection with the line $\varepsilon \to 1$ at S_1 gives information on the particle size which is just elutriated when a particles with a size distribution are fluidized, and the intersection of the same line with the fixed-bed limit $\varepsilon = 0.4$ (S_2) indicates the particle size at which fluidization will break down if agglomeration occurs. The line Ar = const. through S can be used to find the minimum fluidization velocity at S_3 or—as a measure of the upper limit of fluidization—the maximum fluidizing velocity at S_4.

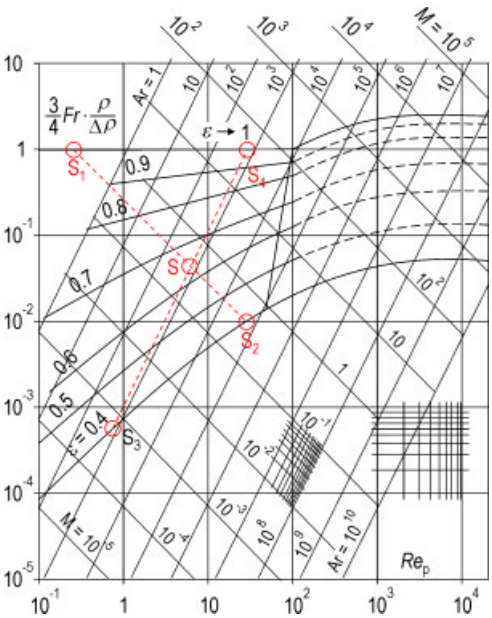

Figure 4. Reh status diagram with status points S and S_1–S_4 (for explanation, see text)

An important practical point is that the state diagram implies a classification scheme that relates various fluidized-bed systems to one another [37, 38] (Fig. 5). When a new fluidized-bed process is being designed, the position of the state point in the diagram will identify related fluidized-bed systems with potentially similar operating problems.

2.5. Gas Distribution

The gas distribution device must satisfy the following requirements:

1) Ensure uniform fluidization over the entire cross section of the bed (especially important for shallow beds)
2) Provide complete fluidization of the bed without dead spots where, for example, deposits can form
3) Maintain a constant pressure drop over long operation periods (outlet holes must not become clogged)

Often, the gas distributor design must also prevent solids from raining through the grid both during operation and after the bed has been shut off.

Porous plates of glass, ceramics, metal, or plastic are commonly used as gas distributors in *laboratory apparatus*; a variety of designs are used in *pilot-plant* and *full-scale fluidized-bed* reactors (see Fig. 6). Many more designs can be found, for example, in [2] and [39].

The principal requirement—uniform distribution of fluidizing gas over the bed cross section—can be met if the pressure drop Δp_d across the gas distribution grid is large enough. Suggested values for the ratio $\Delta p_d / \Delta p_{fb}$ are 0.1–0.3 (with a minimum Δp_d of 3.5 kPa) [40], 0.2–0.4 [41], and > 0.3 [42].

For a given pressure drop Δp_d the gas velocity in the nozzle u_o can be calculated from

$$\Delta p_d = \frac{\varrho_o}{2} \cdot C_D \cdot u_o^2$$

where ϱ_o is the gas density in the orifice and C_D is the drag coefficient. Applying the continuity equation

$$\dot{V} = N_o \cdot A_o \cdot u_o$$

either the number of nozzles N_o or the cross-sectional area of the individual nozzle A_o can be calculated for a given gas flow rate $\dot{V}$.

Problems related to the design of gas distributors are attrition of solids (see Section 2.11), erosion, and back-flow of solids. Erosion may occur at the distributor plate and at neighboring

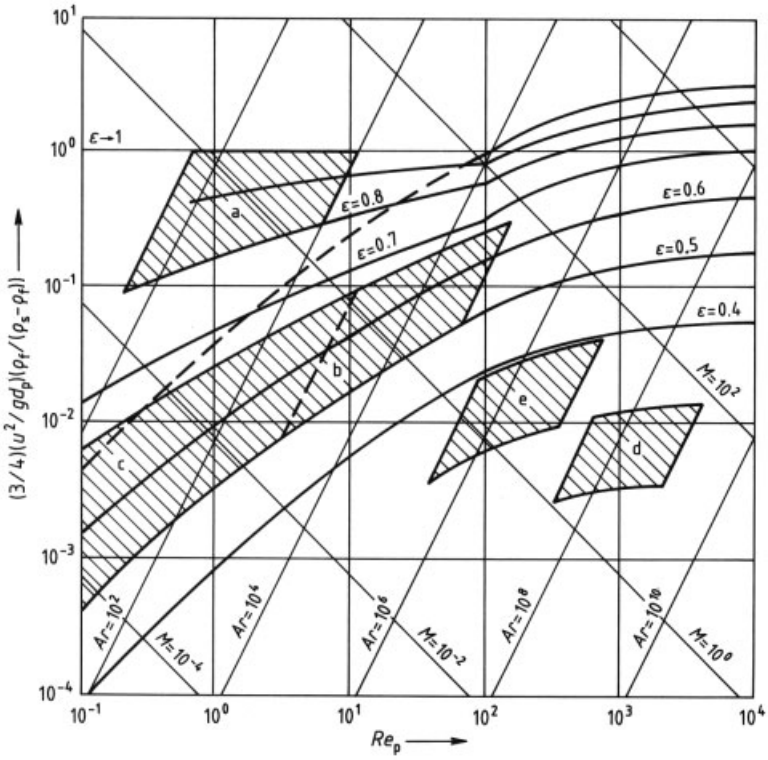

Figure 5. Reh's fluidized-bed state diagram with operating regions of different reaction systems
a) Circulating fluidized bed; b) Fluidized-bed roaster; c) Bubbling fluidized bed; d) Shaft furnace; e) Moving bed

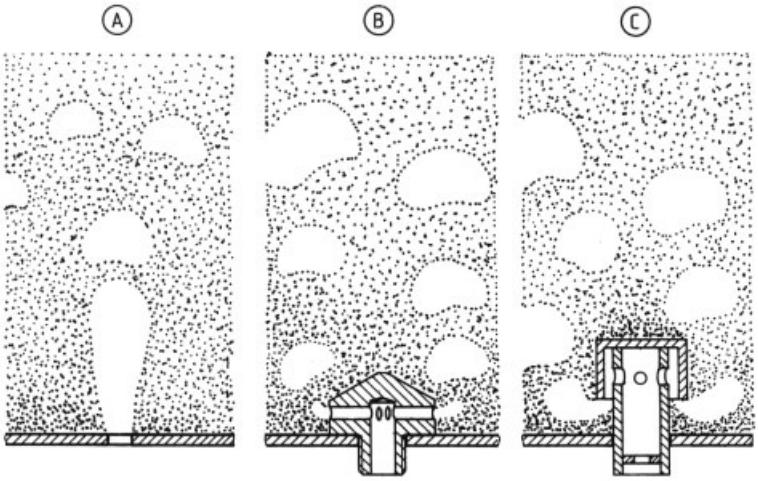

Figure 6. Industrial gas distributors
A) Perforated plate; B) Nozzle plate; C) Bubble-cap plate

nozzles or walls due to gas jets as well as at the nozzle itself. Back-flow of solids into the windbox is caused by pressure fluctuations. In order to prevent this either the design pressure drop has to be larger than the pressure fluctuations or—if this is not feasible for economic reasons—a design must be chosen which tolerates short periods of gas flow reversal without permitting the solids to penetrate into the windbox. For the latter case the bubble cap design has turned out to be advantageous [43].

In the operation of fluidized-bed reactors, the quadratic response ($\Delta p_{\mathrm{d}} \sim u^2$) of industrial gas-distributor designs must be kept in mind, because even if the fluidization velocity is lowered only slightly, an unacceptably low pressure drop across the gas distributor may occur. Industrial experience with different distributor designs, practical design rules, and a discussion of distributor-related problems, such as weepage into the windbox and erosion by grid jets and at grid nozzles, has been compiled in [44].

2.6. Gas Jets in Fluidized Beds

Gas jets can form at the outlet openings of industrial gas distributors and also where gaseous reactants are admitted directly into the fluidized bed. A knowledge of the geometry of such jets, in particular the depth of penetration, is important for the implementation of chemical operations in fluidized-bed reactors, and not just from the standpoint of reaction engineering. It is also vital for reasons of design: the strongly erosive action of these jets means that internals, such as heat-exchanger tubes, must not be located within their range.

The literature contains many empirical correlations for estimating the mean depth of jet penetration L (e.g., [2 – 4]); these must, however, be used with care and, whenever possible, only within the range of parameter values for which they were derived. By way of example, Merry gives the following correlations for *vertical gas* jets [45]:

$$\frac{L}{d_{\mathrm{o}}} = 5.2 \left(\frac{\varrho_{\mathrm{f}} d_{\mathrm{o}}}{\varrho_{\mathrm{s}} d_{\mathrm{p}}}\right)^{0.3} \left[1.3 \left(\frac{u_{\mathrm{o}}^2}{g d_{\mathrm{o}}}\right)^{0.2} - 1\right] \tag{12}$$

and for *horizontal jets* [46]:

$$\frac{L}{d_{\mathrm{o}}} = 5.25 \left[\frac{\varrho_{\mathrm{o}} u_{\mathrm{o}}^2}{(1-\varepsilon)\, \varrho_{\mathrm{s}} g\, d_{\mathrm{p}}}\right]^{0.4} \left(\frac{\varrho_{\mathrm{f}}}{\varrho_{\mathrm{s}}}\right)^{0.2} \cdot \left(\frac{d_{\mathrm{p}}}{d_{\mathrm{o}}}\right)^{0.2} - 4.5 \tag{13}$$

Here d_{o} is the diameter of the outlet opening, u_{o} is the outflow velocity, and ϱ_{o} is the density of the jet gas.

2.7. Bubble Development

For many applications, especially physical operations and noncatalytic reactions, the state of a fluidized bed can adequately be described in terms of a single quantity averaged over the entire bed, such as the mean bed porosity ε. In contrast, the design of fluidized-bed catalytic reactors requires that local fluid-flow conditions also be taken into account.

The local fluid mechanics of gas–solid fluidized beds are determined by the existence of bubbles, which influence the performance of fluidized-bed equipment in several ways: the stirring action and convective solids transport by the rising bubbles are helpful; the resulting intensive solids motion produces a uniform temperature throughout the fluidized bed and rapid heat transfer between the bed and the heating or cooling tubes submerged in it. The bubbles and the motion of solids that they cause, however, also have some drawbacks: attrition of solid particles, erosion of internals, and increased solids entrainment by bubbles bursting at the bed surface. The existence of bubbles is particularly detrimental in the case of a heterogeneous catalytic gas-phase reaction, because the bypass of reactant gas in the bubble phase limits the conversion achieved in the fluidized bed.

The ultimate cause of bubble formation is the universal tendency of gas–solid flows to segregate. Many studies on the theory of stability (e.g., [3, 4]) have shown that disturbances induced in an initially homogeneous gas–solid suspension do not decay but always lead to the formation of voids. The bubbles formed in this way exhibit a characteristic flow pattern whose basic properties can be calculated with the model of Davidson and Harrison [47]. Figure 7 shows the streamlines of the gas flow relative to a bubble rising in a fluidized bed at minimum fluidization conditions ($\varepsilon = \varepsilon_{\mathrm{mf}}$). The characteristic parameter is the ratio α of the bubble's upward

velocity u_b to the interstitial velocity of the gas in the suspension surrounding the bubble:

$$\alpha = \frac{u_b}{u_{mf}/\varepsilon_{mf}} \tag{14}$$

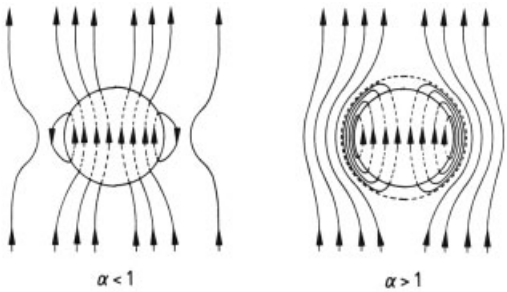

Figure 7. Gas flow for isolated rising bubbles in the Davidson model [47]

The case $\alpha > 1$ is typical for solids of Geldart groups A and B. The gas rising in the bubble flows downward again in a thin layer of suspension ("cloud") surrounding the bubble. An important point for heterogeneous catalytic gas-phase reactions is that the presence of a boundary between bubble gas and suspension gas leads to the existence of two distinct phases (bubble phase and suspension phase) with drastically different gas–solid contact.

If $\alpha < 1$, some of the gas in the suspension phase undergoes short-circuit flow through the bubble, while only part of the bubble gas recirculates through the suspension. This type of flow is typical for fluidized beds of coarse particles (Geldart group D).

Under the real operating conditions of a fluidized-bed reactor, a number of interacting bubbles occur in the interior of the fluidized bed. As a rule, the interaction leads to coalescence. As detailed studies have shown, this process is quite different from that between gas bubbles in liquids because of the absence of surface-tension effects in the fluidized bed [48, 49].

For predicting mean bubble sizes in freely bubbling fluidized beds, a differential equation for bubble growth should be used in the case of Geldart group A and B solids [50]:

$$\frac{\mathrm{d}}{\mathrm{d}h}\, d_v = \left(\frac{2\varepsilon_b}{9\pi}\right)^{\frac{1}{3}} - \frac{dv}{3\lambda\, u_b} \tag{15}$$

with the following boundary condition at $h = h_o$:

$$\frac{d_{v0}}{\mathrm{m}} = \begin{cases} 0.008\varepsilon_b^{1/3} & \text{porous plate} \\ 1.3\left(\frac{\dot{V}_o^2}{g}\right)^{0.2} & \text{industrial gas distributor} \end{cases} \tag{16}$$

where h_o is the height above the grid where the bubbles form (for a porous plate, $h_o \approx 0$; for a perforated plate, $h_o = L$; for a nozzle plate, h_o is the height of the outlet opening above the plate; and for a bubble-cap plate, h_o is the height of the lower edge of the cap above the plate). $\dot{V}_0$ is the volume flow rate of gas through the individual grid opening.

The local volume fraction of bubble gas ε_b is given by

$$\varepsilon_b = \dot{V}_b/u_b \tag{17}$$

and the visible bubble flow $\dot{V}_b$ is

$$\dot{V}_b \approx 0.8\ (u - u_{mf}) \tag{18}$$

The upward velocity u_b of bubbles depends not only on the bubble size but also on the diameter d_t of the fluidized bed:

where

$$u_b = \dot{V}_b + 0.71 \cdot \vartheta_b \cdot \sqrt{g d_v} \tag{19}$$

$$\vartheta_b = \begin{cases} 3.2\, d_t^{0.33} & 0.05 \le d_t \le 1\ \mathrm{m},\ \text{Geldart group A} \\ 3.2\, d_t^{0.5} & 0.1 \le d_t \le 1\ \mathrm{m},\ \text{Geldart group B} \end{cases} \tag{20}$$

Outside these limits, ϑ_b is taken as constant.

The differential equation (Eq. 15) describes not only bubble growth by coalescence but also the splitting of bubbles (second term on the right-hand side [51]). The crucial parameter here is the mean bubble lifetime λ:

$$\lambda \approx 280 \cdot \frac{u_{mf}}{g} \tag{21}$$

In practice, bubble growth is limited not only by the splitting mechanism based on the particle-size distribution of the bed solids, but also by internals (screens, tube bundles, and the like) that cause bubbles to break up. Computational techniques for estimating this process are given in [52, 53].

Hilligardt and Werther have derived a corresponding bubble-growth model for coarse-particle fluidized beds (Geldart group D) [50].

An example of a measured and calculated bubble-growth curve is presented in Figure 8.

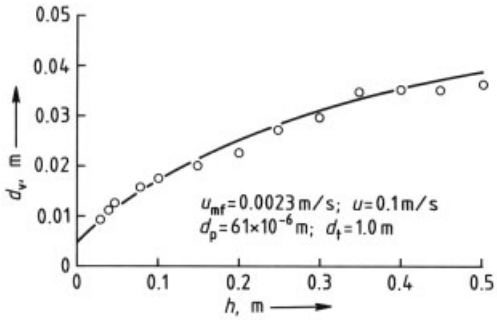

Figure 8. Bubble growth in a fluidized bed of fine particles (Geldart group A; data points from [54], calculation from [50])

2.8. Elutriation

When bubbles burst at the surface of the fluidized bed, solid material carried along in their wake is ejected into the freeboard space above the bed. The solids are classified in the freeboard; particles whose settling velocity u_t is greater than the gas velocity fall back into the bed, whereas particles with $u_t < u$ are elutriated by the gas stream. As a result, both the volume concentration of solids c_v and the mass flow rate of entrained solids in the freeboard show a characteristic exponential decay (Fig. 9). With increasing height above the bed surface, the "transport disengaging height" (*TDH*) is finally reached. Here the increased local gas velocities due to bubble eruptions have decayed, and the gas stream contains only particles with $u_t < u$. When the *TDH* can be reached in a fluidized-bed reactor, this is associated with minimum entrained mass flow rates and solids concentrations, and hence with minimum loading on downstream dust collection equipment. Design of the dust collection system requires knowledge of the entrained mass flow rate G_s and the particle-size distribution of the entrained solids. For the design of the fluidized-bed reactor, the distribution c_v (h) of the solids volume concentration and, for gas–solid reactions, the local particle-size distribution as a function of height in the freeboard must be known.

For solids of Geldart group A, the *TDH* can be estimated with the diagram shown in Figure 10 [55]. The following relation is given for the *TDH* of Geldart group B solids as a function of the size d_v of bubbles bursting at the bed surface [56]:

$$TDH = 18.2 \cdot d_v \tag{22}$$

Equation (25) was, however, derived for a bench-scale unit and may not scale to plant-size equipment.

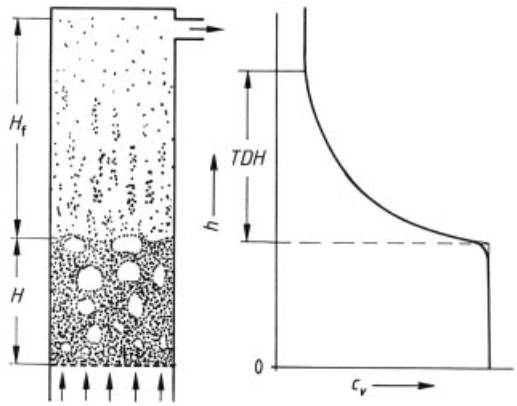

Figure 9. Schematic drawing of fluidized bed and freeboard

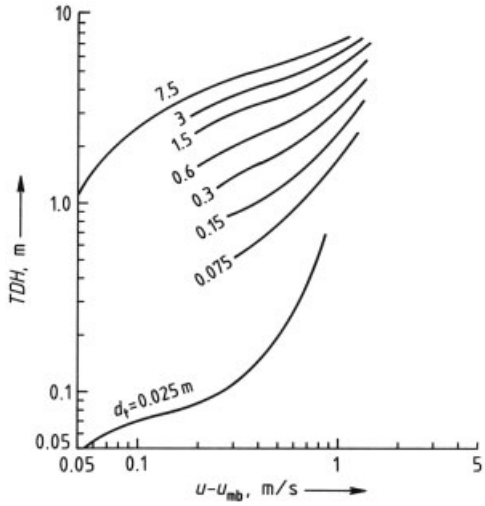

Figure 10. Estimation of transport disengaging height (*TDH*), according to [55]
u_{mb} = Fluidization velocity at which bubble development begins

The mass flow rate G_s of entrained solids per unit area leaving the fluidized-bed reactor is the sum of contributions from the entrainable particle size fractions ($u_t < u$):

$$G_s = \sum_i x_i \cdot \varkappa_i^* \tag{23}$$

Here x_i is the mass fraction of particle-size fraction i in the bed material and $\varkappa^*_i$ is the elutriation rate constant for this fraction. The literature contains a number of empirical correlations for estimating $\varkappa^*_i$ (e.g., [2 – 4]). More physical-based are the elutriation models of WEN and CHEN [57] and of KUNII and LEVENSPIEL [2, 58], which enable not only calculation of the exiting mass flow rate but also estimation of the concentration versus height c_v (h) in the freeboard. The model by SMOLDERS and BAEYENS additionally takes the effect of variable freeboard geometry into account [59].

A literature survey on the factors affecting elutriation and the available modeling tools is given in [60].

2.9. Circulating Fluidized Beds

2.9.1. Hydrodynamic Principles

In REH's state diagram of the fluidized bed [36], the circulating fluidized bed (CFB) is located above the single-particle suspension curve for $Re < 10^2$ and porosities ε greater than about 0.8 (dashed line in Fig. 5). The shortcoming of this diagram is that it does not show an important parameter in the operation of a circulating fluidized bed: the circulating solids mass flow rate per unit area G_s. The diagram of Figure 11 [61] attempts to remedy this by plotting the mean slip velocity u_{sl} between gas and solids

$$u_{sl} = \frac{u}{\varepsilon} - \frac{(G_s/\varrho_s)}{1-\varepsilon} \tag{24}$$

versus the mean solids concentration $c_v = 1 - \varepsilon$, with G_s as the parameter. The limiting conditions are high solids concentration (bed at minimum fluidization) and $c_v \to 0$ with $u_{sl} = u_t$ (isolated single particle). In the circulating fluidized-bed region, slip velocity increases with increasing G_s and can become much higher than the single-particle settling velocity (the physical justification for this statement comes from the formation of strands or clusters of particles). In the entrained-flow region the slip velocities again decrease with decreasing solids concentration.

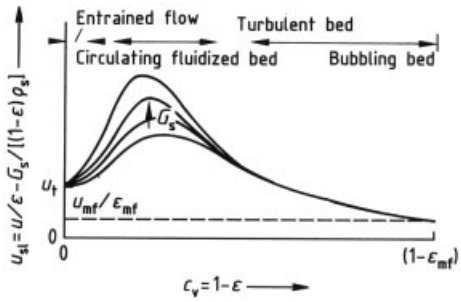

Figure 11. Fluidized-bed state diagram, according to [61]

The fluidized-bed state diagrams discussed thus far, as well as others (e.g., [62, 63]), are suitable mainly for the qualitative interpretation of flow phenomena. A diagram proposed by WIRTH (e.g., [11, 64, 65]) also provides quantitative assistance in the design of circulating fluidized beds. The schematic in Figure 12 applies to a given gas–solid system described by a constant value of the Archimedes number Ar. The ordinate is the dimensionless pressure drop of the fluidized bed

$$\psi = \frac{\Delta p}{(\varrho_s - \varrho_f)\ (1-\varepsilon_{mf})\ g\,\Delta h} \tag{25}$$

the abscissa is the particle Froude number

$$Fr_p = \frac{u}{\sqrt{\frac{(\varrho_s - \varrho_f)}{\varrho_f} g d_p}} \tag{26}$$

The dimensionless pressure drop ψ is the ratio of the pressure drop Δp along the flow path Δh to the maximum possible value for ascending flow (the value that would be attained if the pipe cross section were filled with solids corresponding to the concentration at the minimum fluidization point). The parameter of the family of curves is a volume flow rate ratio

$$\frac{\mu\ \varrho_f}{\varrho_s\ (1-\varepsilon_{mf})} \tag{27}$$

where μ is the ratio of solid-to-gas mass flow rates. The limiting curve bounds the region of stable, vertically upward gas–solid flow on the low gas velocity side.

Figure 13 shows how the state diagram of Figure 12 is constructed for a circulating fluidized bed with siphon recycle. If solids holdup in the recycle line and siphon is ignored, this case represents operation with a constant bed mass independent of velocity. At high gas velocities and if acceleration effects are neglected,

the bed material is distributed uniformly over the total height H_{cfb} of the fluidized bed (Fig. 13 C). The circulating fluidized bed then exhibits a single steady-state section with a constant pressure gradient ($\Delta p/\Delta h$). This pressure gradient can be calculated from the bed mass as

$$\psi_{\mathrm{hom}} = \frac{(\varrho_{\mathrm{s}}-\varrho_{\mathrm{f}})\,\mathrm{g}H_{\mathrm{mf}}\,(1-\varepsilon_{\mathrm{mf}})}{(\varrho_{\mathrm{s}}-\varrho_{\mathrm{f}})\,\mathrm{g}H_{\mathrm{cfb}}\,(1-\varepsilon_{\mathrm{mf}})} = \frac{H_{\mathrm{mf}}}{H_{\mathrm{cfb}}} \tag{28}$$

where H_{mf} is the bed height at minimum fluidization.

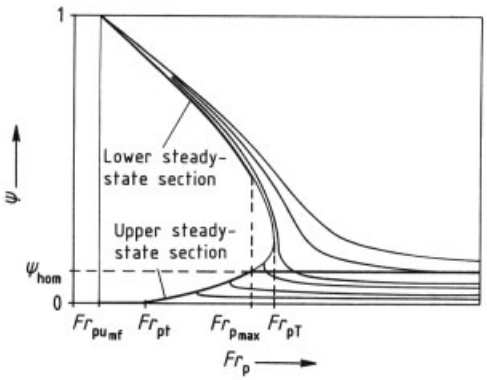

Figure 12. State diagram for the circulating fluidized bed with siphon, according to WIRTH [64]
Ar = const., parameter of family of curves is the volume flow rate ratio $\mu\ \varrho_{\mathrm{f}}/(\varrho_{\mathrm{s}}\ (1-\varepsilon_{\mathrm{mf}}))$; Fr_{p}= particle Froude number for superficial minimum fluidization velocity ($\mathrm{pu}_{\mathrm{mf}}$), single-particle terminal velocity (pt), and transport velocity (pT), respectively.

The states identified by ψ_{hom} to the right of the bounding curve in Figure 12 are accessible by increasing the gas velocity (corresponding to increasing Fr_{p}). With increasing Fr_{p} the volume flow ratio increases; that is, relatively more solids are elutriated (and thus circulated).

If Fr_{p} is allowed to drop below the limit $Fr_{\mathrm{p_{max}}}$ (Fig. 13 B, Fig. 12) two steady-state sections appear in the riser tube: the one in the lower part is marked by a high pressure gradient, that in the upper part by a lower gradient. Figure 13 illustrates the physical significance of these two pressure gradients. In practice, the transition between the two linear regions takes place gradually. The height of the transition zone corresponds to the transport disengaging height (TDH).

The picture changes further if the gas velocity declines to values lower than the settling velocity u_{t} of a single isolated particle. In this case (for $Fr_{\mathrm{p}} < Fr_{\mathrm{pt}}$, Fig. 13 A, Fig. 12), no more solids can be elutriated, and the pressure gradient in the upper linear region vanishes. All the solid material is now in the form of a bubbling or turbulent fluidized bed.

The *solids concentrations* averaged over the tube cross section $(1-\varepsilon)$ can be calculated from the dimensionless pressure drop:

$$1-\varepsilon = (1-\varepsilon_{\mathrm{mf}})\cdot\psi \tag{29}$$

Besides the pressure and solids concentration profile, the *circulating mass flow rate of solids* $G_{\mathrm{s}}\cdot A_{\mathrm{t}}$ is important for the design of the circulating fluidized bed. In particular, the design of the solids collection and recycle system depends very much on this quantity. The mass flow rate of solids depends on the flow regime. At gas velocities such that two steady-state sections are

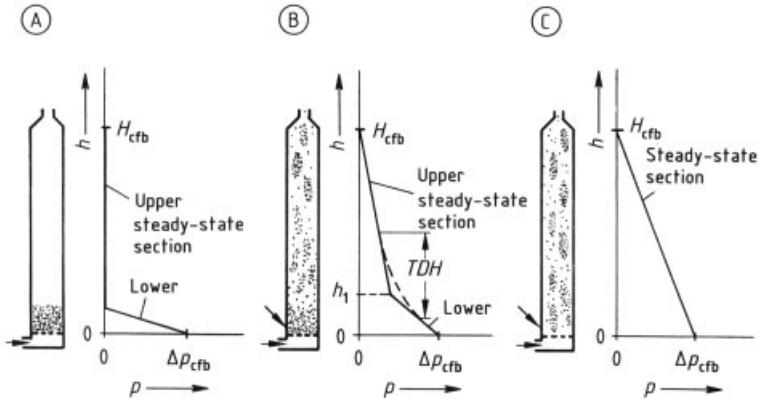

Figure 13. Pressure profile in the circulating fluidized bed with siphon, according to WIRTH [64]
A) $Fr_{\mathrm{pu_{mf}}}<Fr_{\mathrm{p}}<Fr_{\mathrm{pt}}$; B) $Fr_{\mathrm{pt}}<Fr_{\mathrm{p_{max}}}$; C) $Fr_{\mathrm{p}}>Fr_{\mathrm{p_{max}}}$

present in the bed vessel (i.e., $Fr_{\mathrm{p}u_{\mathrm{mf}}} < Fr_{\mathrm{p}} < Fr_{\mathrm{pT}}$), the mass flow rate of entrained solids depends on the physics of the gas–solid flow. Figure 14 plots the dimensionless solids mass flow rate versus Fr_{p}, with the Archimedes number as parameter. For a given Ar, the flow rate tends to zero as $Fr_{\mathrm{p}} \rightarrow Fr_{\mathrm{pt}}$ and reaches a maximum at $Fr_{\mathrm{p}} = Fr_{\mathrm{pT}}$. The slope of the elutriation curve becomes greater with increasing Ar; that is, the coarser the particles, the greater is the relative change in the circulating mass flow rate of solids with a change in gas velocity.

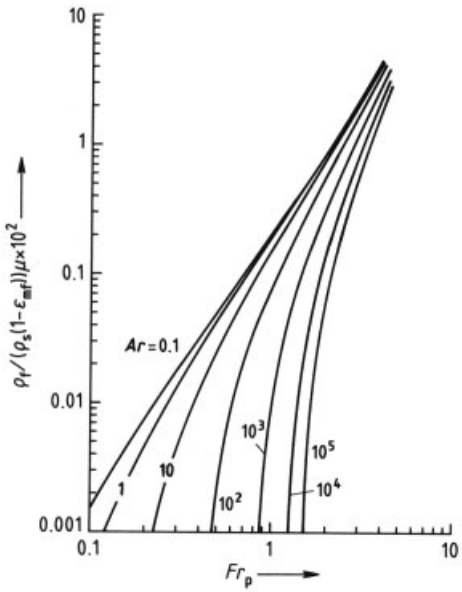

Figure 14. Elutriation diagram when the circulating fluidized bed contains two steady-state sections, according to WIRTH [64]

At high gas velocities in the circulating fluidized bed (i.e., when a single steady-state section exists), the entrained mass flow rate depends on the particle Froude number and the solids holdup. More detailed information about the application of Wirth's theory in practice may be found in [11].

Whereas WIRTH's analysis of the circulating fluidized bed starts from the pneumatic transport condition, the models of RHODES and GELDART [66], as well as KUNII and LEVENSPIEL [2, 58], are based on the bubbling fluidized bed and describe the circulating fluidized bed as a limiting case of a bubbling bed with a very high rate of solids entrainment.

2.9.2. Local Flow Structure in Circulating Fluidized Beds

The Wirth state diagram, as a first step toward the local characterization of flow regimes in a circulating fluidized bed, describes the vertical profile of the solids concentration. In the lower section of a circulating fluidized bed a dense region exits near the gas distributor. It has been observed that in this bottom zone bubble-like voids coexist with a surrounding dense suspension. The solids volume concentration is higher at the wall ($c_{\mathrm{v}} \approx 0.4$) then in the center ($c_{\mathrm{v}} \approx 0.15$) of the bottom zone [67]. The splash zone which links the bottom zone to the upper dilute zone is characterized by violent gas–solid mixing. Many recent experimental studies with various measurement techniques (e.g., X-ray tomography [68], capacitance tomography [69] and fiber-optical probes [70]) have shown that the upper section of the circulating fluidized bed exhibits characteristic horizontal profiles, with the concentration $c_{v,wall}$ near the vessel wall always significantly higher than the value $\bar{c}_v$ averaged over the vessel cross section; for example, $c_{v,\ \mathrm{wall}} = 2.3 \cdot \bar{c}_v$ [71].

Local measurements of the solids concentration and solids velocity show that upward-flowing regions of low solids concentration and downward-flowing aggregates of high solids concentration alternate in time at every point inside the fluidized bed, with downward-moving aggregates (strands, clusters) predominating near the wall and upward-moving regions of low suspension concentration predominating in the central zone. However, no significant downward flow of solids near the wall was observed in high-density circulating fluidized beds, e.g., [72]. The picture of the local flow structure in a circulating fluidized bed, as derived from these observations, is shown schematically in Figure 15.

A modeling approach which is based on the local flow structure of the CFB is the energy-minimization multiscale (EMMS) model [73]. It considers the tendency of a fluid in a gas–solid two-phase flow to pass through the particulate layer with least resistance and the tendency of the solids to maintain least gravitational po-

tential. Least resistance means that the volume-specific energy consumption for suspending and transporting solids is minimized, and minimization of the gravitational potential is equivalent to the requirement that the local mean voidage ε attains a minimum. The model has been applied as a description of fluid-mechanical phenomena in CFB risers of different sizes [74, 75] but also for the prediction of flow patterns of gas and solids in industrial-scale units, such as a CFB boiler [76] and a petrochemical processing unit [77]. Another promising line of development is the introduction of the EMMS concept into computational fluid dynamical calculations of multiphase flows; first results obtained with a drag model based on the EMMS model are encouraging [78].

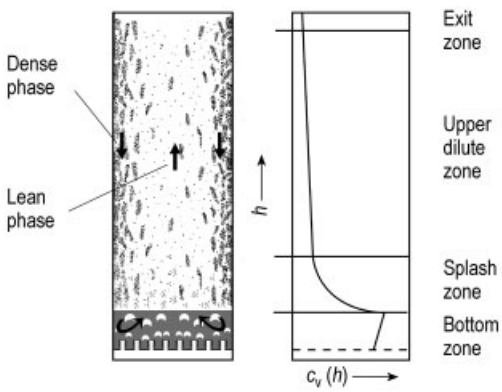

Figure 15. Schematic diagram of flow structure in a circulating fluidized bed

2.9.3. Design of Solids Recycle System

Solids carried over with the fluidized gas are generally collected in cyclones. In the case of *bubbling beds*, the solids can easily be returned to the bed through the standpipe of the cyclone, which dips directly into the bed.

Due to the large amounts of circulating solids, *circulating fluidized beds* require very large cyclones arranged beside and outside the bed, with special "valves" needed to connect the standpipe to the bed vessel. Figure 16 shows two design options, the siphon and the L-valve. With the siphon, the solids are fluidized (i.e., enabled to flow back into the reactor). In the L-valve design, the mass flow rate of the solids can be regulated by varying the gas supplied to the standpipe.

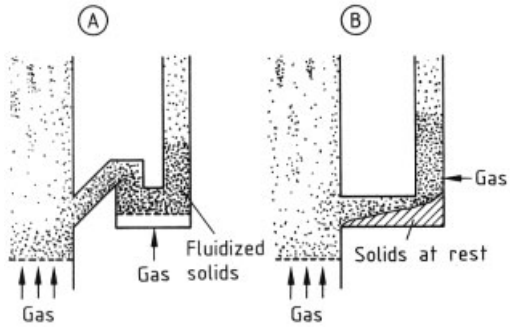

Figure 16. Design options for solids recycle
A) Siphon; B) L-valve

Because the solids path does not contain any sort of mechanical closure, the characteristic pressure distribution plotted in Figure 17 is obtained. The distribution of solids between the fluidized bed and the recycle line is directly related to this pressure distribution. Operating properties differ from one recycle design to another [79].

2.10. Cocurrent Downflow Circulating Fluidized Beds (Downers)

A certain drawback of circulating and bubbling fluidized beds when applied for gas-phase reactions is the backmixing which inevitably occurs in the gas phase. In bubbling fluidized beds it is the bubble-induced solids circulation, and in circulating fluidized beds the downflow of solids in the wall zone, which entrains gas in the upstream direction and thus lowers the yield of a catalytic reaction or gives rise to undesired consecutive or side reactions. These disadvantages caused by the hydrodynamic effects of both gas and solids flowing against gravity could be overcome in the so-called downer reactor, in which the flow directions of both gas and solids are downward, i.e., in the same direction as gravity [80]. Another incentive is the possibility of realizing short contact times between gas and solids of around or even below one second.

Downer systems have been intensely studied [80]. Hydrodynamics [81, 82], gas mixing

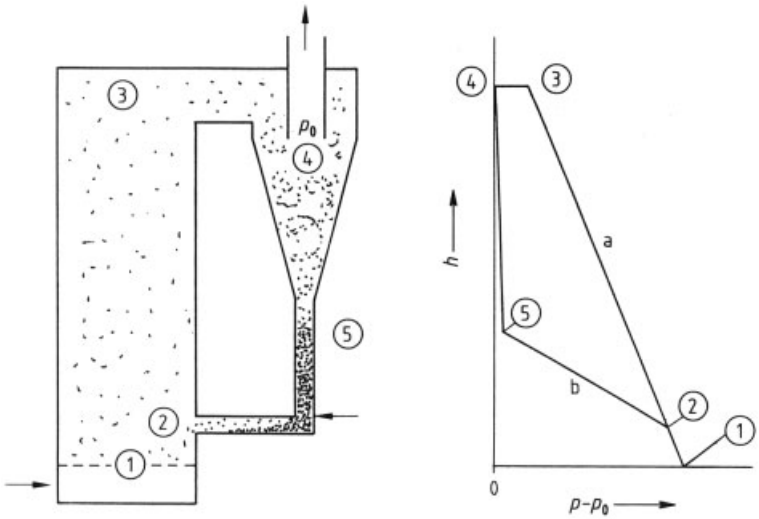

Figure 17. Pressure distribution in solids recycle system of a circulating fluidized bed
a) Fluidized bed; b) Return leg

[83], and solids mixing [84, 85] have been investigated both experimentally and by numerical simulation [86]. It has been found that the hydrodynamics of the downer are also characterized by a wall zone of increased solids concentration. However, axial and radial gas-solids flow structures are much more uniform than in conventional fluidized beds. Another result is that the length of the flow development zone is much shorter for the downer than for the riser, which means that reactions with very short contact times can be carried out under near-plug-flow conditions. However, the solids feeding process and the geometry of the entrance region are critical points that deserve special attention [87].

The patent and open literature suggest various applications for downer reactors, e.g., residual oil cracking [88], coal pyrolysis [89], and biomass pyrolysis [90]. The catalytic pyrolysis of heavy feeds for the production of light olefins has been investigated on the laboratory scale with promising results [88]. However, no large-scale industrial process has emerged yet.

2.11. Attrition of Solids

The attrition of solid particles is an unavoidable consequence of the intensive solids motion in the fluidized bed. The attrition problem is especially critical in processes where the bed material needs to remain unaltered for the longest possible time, as in fluidized-bed reactors for heterogeneous catalytic gas-phase reactions. Catalyst attrition is important in the economics of such processes and may even become the critical factor.

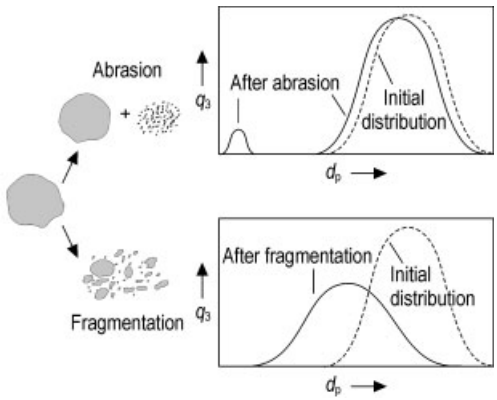

Figure 18. Attrition modes and their effects on the particle size distribution (q_3 = mass density distribution of particle sizes d_p)

Catalyst attrition in fluidized-bed reactors occurs normally as surface abrasion (Fig. 18) which means that surface asperities are abraded and edges of the catalyst particles are rounded off. Fragmentation may also play a role, especially for some fresh catalyst particles which on

entering the reactor may simply be crushed into pieces. If in an industrial process extraordinarily high catalyst losses are observed it is advisable to examine catalyst samples under the scanning electron microscope. If the sample contains many fragments this could be an indication of a wrong design (e.g., too high a velocity at the cyclone inlet or at the distributor).

When designing catalytic fluidized-bed processes, the attrition performance of candidate catalysts should be tested under standardized conditions in the process development stage. This test can be performed in a small laboratory apparatus; it consists essentially of an extended fluidization test in which the mass of solids carried out of the bed is recorded as a function of time. Figure 19 presents a typical test result: during the first hours of testing, both the attrited material and the fine fraction of the bed material are elutriated. Only after a relatively long operating period is a quasi-steady state attained. The attrition rate r_a in this steady state can be defined as

$$r_a = \frac{1}{m_b} \cdot \frac{\Delta m_a}{\Delta t} \tag{30}$$

where m_a is the elutriated mass and m_b the bed mass. Usually r_a is expressed as percentage per day; for relatively attrition-resistant, fluidized-bed catalysts, it is of the order of 0.2 % per day [9].

Many standard test apparatuses have been proposed for comparative attrition tests (e.g., [91, 92]), but all such equipment has been suitable only for comparative studies of different catalysts under consideration for the same process. The attrition measured in large-scale equipment can be far different from the values measured in a test apparatus.

A number of sources can be identified for catalyst attrition in industrial fluidized-bed reactors:

1) Jet attrition at gas distribution grid openings and nozzles where gaseous reactants are admitted to the bed
2) Bubble attrition in the bed due to solids motion caused by bubbles
3) Attrition in cyclones
4) Attrition in pneumatic conveying lines, such as those between reactor and regenerator beds

Empirical correlations are available for the attriting action of a gas jet in the fluidized bed [93] and for the size reduction effect of solids motion in the bed [94, 95].

WERTHER and coworkers [96] employ the laboratory apparatus shown schematically in Figure 20 which enables separate study of the attrition due to jets from nozzles of various diameters and that due to bubbles.

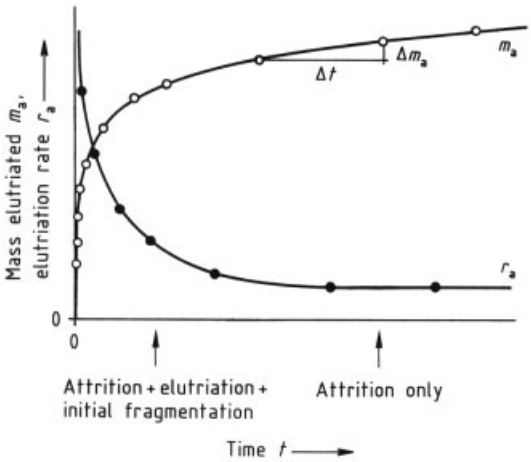

Figure 19. Result of an attrition measurement

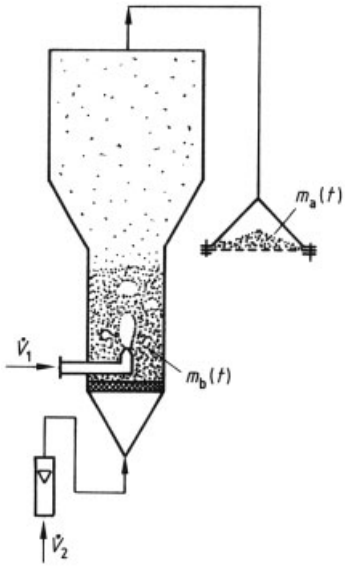

Figure 20. Experimental apparatus for attrition measurement

Under steady-state conditions the jet-attrition-related mass production of fines per unit time for a gas distributor with a number n_o of orifices from mother particles with diameter $d_{p,i}$ which are present in the catalyst inventory with a mass fraction ΔQ_{3i} is proportional to the particle size d_{pi}, the mass fraction ΔQ_{3i}, the density ρ_o of the gas issuing from the orifice, the square of the orifice diameter d_o, and to the cube of the jet exit velocity u_o [97, 98]:

$$\dot{m}_{\mathrm{att,jet},i}=c_j n_0 d_{pi} \Delta Q_{3i} \rho_0 d_0^2 u_0^3 \tag{31}$$

Attrition due to the bubble induced solids movement is given by [98]

$$\dot{m}_{\mathrm{att,bubble},i}=c_b d_{pi} \Delta Q_{3i} m_b (u-u_{mf})^3 \tag{32}$$

where m_b denotes the bed mass which contains bubbles (i.e., which is located outside the jet-dominated grid region). Equation (32) also denotes the mass production of attrited fines which is resulting from the size fraction d_{pi} in the bed.

The stress on the catalyst particles will be different in contact with a gas jet, in the bulk of the bubbling fluidized bed, and during its passage through a cyclone. Recent investigations of cyclone-induced catalyst attrition [99 – 101] have shown that the mass flow of attrited fines which is produced by attrition inside the cyclone when a solids mass flow $\dot{m}_c \cdot \Delta Q_{3ci}$ of particles of the size fraction d_{pi} enters the cyclone is given by

$$\dot{m}_{\mathrm{att,c},i}=c_c \cdot \dot{m}_c \cdot \Delta Q_{3ci} \cdot d_{pi} \cdot \frac{u_c^2}{\sqrt{\mu_c}} \tag{33}$$

where u_c is the gas velocity at the cyclone inlet, and μ_c the solids loading of the incoming gas flow

$$\mu_c=\frac{\dot{m}_c}{\rho_c \cdot u_c \cdot A_c} \tag{34}$$

where ρ_c is the density of the inflowing gas, and A_c the cross-sectional area of the cyclone inlet.

Equations (31)–(34) describe the catalyst attrition under conditions of steady state, i.e., when the particles are more or less rounded off (Fig. 18). To describe also the initial breakage and attrition of fresh catalyst particles, it is necessary to follow the fate of the particles on their introduction into the reactor, which is possible with population balance models (cf. Section 9.3.2). Klett et al. [102] and Hartge et al. [103] have defined a stress history parameter ϑ

$$\vartheta=\begin{cases} t/t_j^* \text{ for jet} - \text{induced attrition} \\ t/t_b^* \text{ for in} - \text{bed attrition} \\ n_p/n_p^* \text{ for attrition in cyclones} \end{cases} \tag{35}$$

where the definition of the characteristic parameters t_j^*, t_b^*, and n_p^* can be taken from Figure 21, n_p is the number of passages of a given particle through the cyclone, and t_b and t_j are the time periods during which the particle is subjected to bubble and jet stress, respectively. If it is assumed that the effects of the different stress mechanisms on the catalyst particles are additive, a uniform treatment of the overall stress history for all three attrition mechanisms is given by

$$\frac{\dot{m}_{\mathrm{att}}(\vartheta)}{\dot{m}_{\mathrm{att},\infty}}=\begin{cases} 1.1 \cdot \vartheta^b & \vartheta \leq 1.1^{-1/b} \\ 1 & \vartheta > 1.1^{-1/b} \end{cases} \tag{36}$$

The parameter b is characteristic of a given catalyst. Figure 22 shows measurements with FCC catalyst [103] which lead to b = −1.16. Equation (36) allows the description of the stress-history-dependent attrition rate and can be used for the simulation of fluidized bed reactors (see Section 9.3.2).

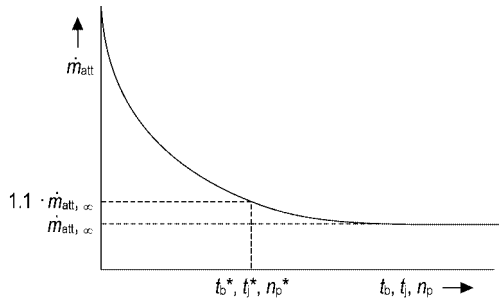

Figure 21. Dependence of attrition on time (bubble- and jet-induced attrition) and number of passages n_p through a cyclone.

A variety of approaches exist for reducing attrition in industrial fluidized-bed reactors. The jet attrition action can be controlled with special gas distributor designs ([9]; e.g., by the use of bubble caps, Fig. 6) such that gas jets do not issue directly into the bed at high velocity. Attrition due to bubbles can be lowered by limiting bubble growth (avoiding high gas velocities and large bed heights; use of fine catalysts with low

u_{mf}, as implied by Eqs. 18 and 24). Attrition in cyclones can be prevented, in the simplest case, by replacing the cyclones with devices such as filters. Attrition can also be minimized by cutting back the load on the cyclone, for example, by placing the cyclones above the TDH. Relatively high catalyst attrition also occurs in circulating fluidized beds where very large quantities of solids must be collected in the cyclones.

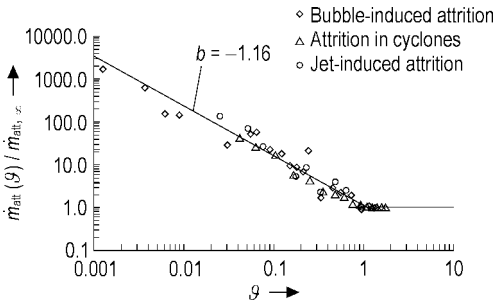

Figure 22. Dimensionless attrition rate of FCC catalyst as a function of stress history.

3. Solids Mixing in Fluidized-Bed Reactors

The intensive solids mixing typical of fluidized-bed reactors has several effects on performance. In *catalytic reactions*, the large-scale vertical solids mixing results in a transport of the gas components, adsorbed to the catalyst, so that the gas phase is backmixed and the conversion and selectivity are impaired. In *noncatalytic gas–solid reactions*, the mean solids residence time and residence-time distribution, as well as the propagation behavior of the solids from individual feed points, play a role. In general, fast and strongly exothermic reactions require fairly vigorous solids mixing to prevent temperature peaks near the reactant inlet.

3.1. Mechanisms of Solids Mixing

The wake of the rising bubbles produces a rather slight upward and lateral drift of the particles (Fig. 23 A) [104]. In addition, solid particles are drawn upward in the wake, portions of the wake are shed at irregular intervals during bubble motion, and new portions of solids are taken into the wake (Fig. 23 B). Solids transport in the wake is essentially the reason that vertical solids mixing is from one to two orders of magnitude better than horizontal mixing.

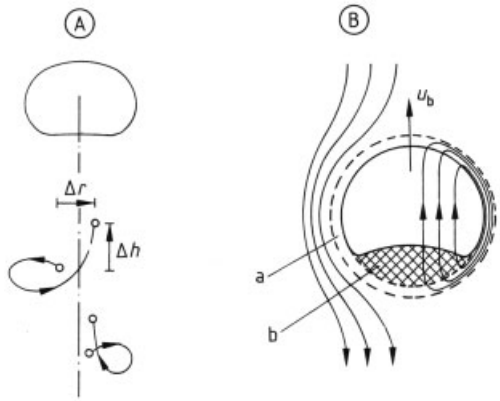

Figure 23. Solids mixing in bubbling fluidized beds due to particle drift (A) and wake transport (B)
a) Cloud; b) Wake

For reasons of continuity, the upward transport of particles by bubbles is coupled with a downward movement in the suspension phase that surrounds the bubbles. Measurements of the local bubble-gas flow have shown that the rising bubbles are not distributed evenly over the bed cross section. As a typical example, Figure 24 A gives a plot of the radial distribution of the bubble-gas flow at three heights above the grid in a fluidized bed 1 m in diameter. The profile is comparatively flat in the bottom zone but exhibits a steeper slope as the height increases, with an annular zone of preferentially rising bubbles. The resulting circulation of the solids also features an annular region of upward transport in the wakes with predominantly downward motion of the solids in the center and at the periphery of the bed (Fig. 24 B).

The large-scale solids circulation can be reinforced by uneven distribution of the fluidized gas over the distributor cross section [106]. Figure 25 presents examples of industrial fluidized-bed furnaces in which forced circulation of the solids is employed to improve coal burnup.

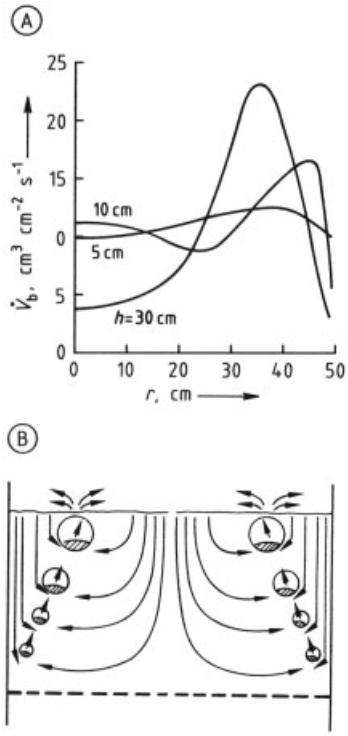

Figure 24. A) Radial distribution of bubble-gas flow; B) Relationship between bubble distribution and solids circulation [105]
d_t = 1 m, quartz sand, u_{mf} = 0.013 m/s, u = 0.2 m/s, H_{mf} = 0.5 m, $\dot{V}_b$ = visible bubble flow

3.2. Vertical Mixing of Solids

The propagation behavior of the solids in a fluidized bed can be described by a number of models (e.g., [2, 109]). Most commonly used is the *dispersion model*, in which solids transport is described by a diffusion law. The numerical value of the dispersion coefficient D_{sv} for solids mixing in the vertical direction increases with increasing gas velocity because of the growth in the number and size of bubbles. The following simple empirical correlation is given for fine particles (Geldart groups A and B) [2]:

$$\left(\frac{D_{sv}}{m^2/s}\right) = 0.06+0.1\left(\frac{u}{m/s}\right) \tag{37}$$

For a *plant-scale fluidized bed* (0.9 × 1.26 m^2 in plan, bed height 4 m) equipped with a bundle of horizontal tubes, a very similar relation was derived for a solid of Geldart group B [110]:

$$\left(\frac{D_{sv}}{m^2/s}\right) = 0.056\left(\frac{u-u_{mf}}{m/s}\right) \tag{38}$$

Because solids circulation becomes more marked in *larger-diameter fluidized beds*, the dispersion coefficient increases rapidly with increasing bed diameter d_t (Fig. 26). For this case the following expression is found [2]:

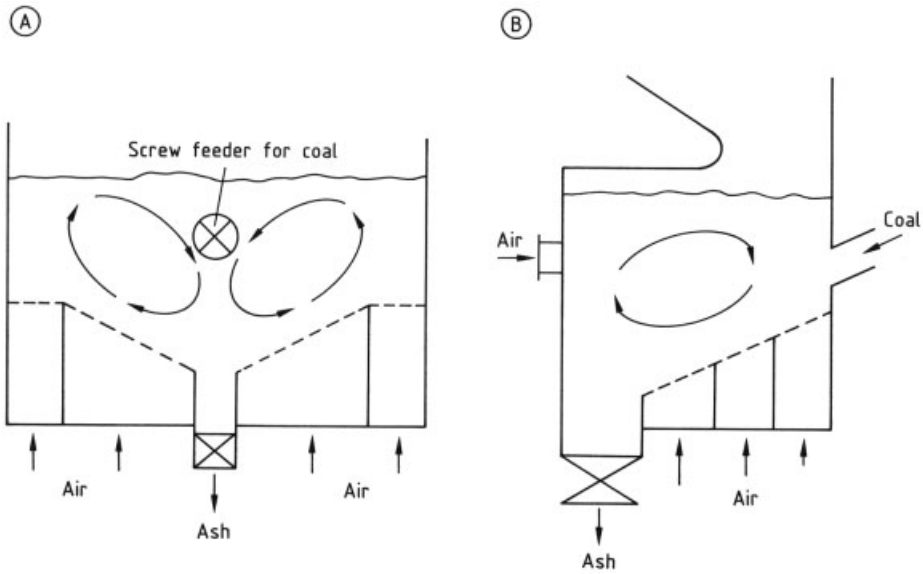

Figure 25. Fluidized-bed furnaces with forced circulation of solids
A) According to [107]; B) According to [108]

$$\left(\frac{D_{\mathrm{sv}}}{\mathrm{m}^2/\mathrm{s}}\right) = 0.030\left(\frac{d_{\mathrm{t}}}{\mathrm{m}}\right)^{0.65} \tag{39}$$

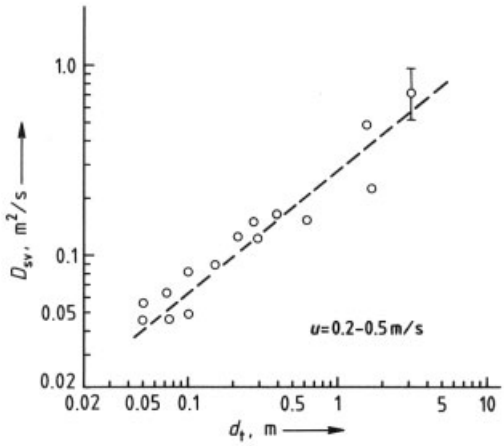

Figure 26. Vertical solids dispersion in fluidized beds of fine particles (Geldart groups A and B) [2]

The above correlations can provide only rough values. Other effects observed in practice include, in particular, a significant effect of particle-size range [111, 112].

3.3. Horizontal Mixing of Solids

In gas–solid reactions, the propagation behavior of the solids in the horizontal direction is important if, for example, the solid material is fed into the bed at isolated feed points. WERTHER and coworkers model the horizontal propagation of coal in a fluidized-bed furnace, describing the carbon conversion in terms of a simple first-order reaction (rate constant k with dimension s^{-1}) [113]. The crucial parameter is the ratio $\mathrm{k}d_{\mathrm{t}}^2/D_{\mathrm{sh}}$ between the rate of the chemical reaction and the rate of dispersive mass transport. For high values of k (fast reaction), large reactor diameters d_{t}, and low values of the dispersion coefficient D_{sh}, the local carbon concentration in the bed exhibits a rather steep horizontal profile, resulting in a significantly nonuniform distribution of gas emissions over the bed cross section.

On the basis of KUNII and LEVENSPIEL's model of bubble-induced solids mixing [114], an expression has been derived for calculating the horizontal dispersion coefficient D_{sh} averaged over the bed height H, given local bubble properties (bubble diameter d_v, bubble-gas holdup ε_{b}) [115]:

$$D_{\mathrm{sh}}=0.67\times10^{-3}+0.023\,\frac{1}{H}\int_0^H \frac{\varepsilon_{\mathrm{b}}}{1-\varepsilon_{\mathrm{b}}}\cdot\sqrt{g\cdot d_v^3}\,\mathrm{d}h \tag{40}$$

This correlation holds for solids of Geldart groups B and D with Archimedes numbers between 8 600 and 58 000.

3.4. Solids Residence-Time Properties

Many applications of fluidization technology involve continuous processing of solids. Important considerations in such cases are not only the mean solids residence time but also the residence-time distribution. Whereas all elements have the same residence time in a plug-flow system, a stirred tank exhibits a broad distribution of residence times. To a good approximation, the residence-time properties of the fluidized bed with respect to the solids are the same as those of a stirred tank. The mean residence time τ is the ratio of the solids mass m_{b} in the reactor to the solids throughput $\dot{m}_{\mathrm{s}}$:

$$\tau = \frac{m_{\mathrm{b}}}{\dot{m}_{\mathrm{s}}} \tag{41}$$

The mass fraction $\mathrm{d}m_{\mathrm{s}}/m_{\mathrm{b}}$ of solids having a residence time between t and $t + \mathrm{d}t$ is

$$\frac{\mathrm{d}m_{\mathrm{s}}}{m_{\mathrm{b}}} = \frac{1}{\tau}\,\mathrm{e}^{-\frac{t}{\tau}}\,\mathrm{d}t \tag{42}$$

Similarly, the fraction f of solids having a residence time less than t in the bed is calculated as

$$f = 1-\mathrm{e}^{-t/\tau} \tag{43}$$

The residence-time distribution can be narrowed by placing a number of fluidized beds in series. Multistage systems of this type are used, for example, in fluidized-bed drying [18].

3.5. Solids Mixing in Circulating Fluidized Beds

The circulating fluidized bed exhibits a complex gas–solid flow pattern as discussed in Section

2.9. Different regions can be discriminated with respect to the prevailing mechanisms of solids motion and mixing. An extensive survey on experimental findings in solids mixing is given in [116]. In the upper diluted zone of the circulating fluidized bed, clusters are formed with mainly upward flow in the core and predominantly downwards motion near the wall. While the wall region can be modeled by a plug-flow approach, the core region exhibits radial gradients. The Péclet number characterizing radial solids mixing in the core region

$$Pe_{\mathrm{r,\,s}} = \frac{u_{\mathrm{c}} \cdot 2R^*}{D_{\mathrm{r,\,s}}}$$

increases from 150 to 300 with increasing solids volume concentrations [117]. A recent investigation of solids mixing in the bottom zone with solid carbon dioxide as a tracer showed that in this zone solids are almost ideally mixed in the vertical direction but lateral mixing is limited with dispersion coefficients of about 0.1 m^2/s which corresponds to Péclet numbers of around 40, [118].

Counteracting to solids mixing, segregation occurs in applications using particles of a broad size distribution and/or different densities. Easily fluidized particles tend to be elutriated while others tend to sink. A dynamic equilibrium between solids mixing and segregation is established, causing a spatial distribution of particles with significantly different solids properties, as was shown in an experimental study with a mixture of iron powder and quartz sand with a broad particle size distribution [119].

4. Gas Mixing in Fluidized-Bed Reactors

The mixing and residence-time distribution of the gas are particularly important for catalytic reactions but are also significant for gas–solid reactions when gaseous reactants are to be converted to the greatest possible extent in fluidized beds (e.g., reduction of fine-grained iron ores to sponge iron with gaseous reductants [18]). Gas mixing is closely linked to the motion and mixing of the solids in the bed.

4.1. Gas Mixing in Bubbling Fluidized Beds

If the flow and mixing of gas in the bubbling fluidized bed are described by a simple one-phase dispersion model, the coefficients D_{gv} and D_{gh} of gas dispersion in the vertical and horizontal directions have similar numerical values and follow trends similar to those of the solids dispersion coefficients. By way of example, Figure 27 shows the effect of fluidized-bed diameter d_{t} on vertical gas dispersion. The increase in dispersion coefficient with vessel diameter might be attributable to the formation of large-scale solids circulation patterns, which becomes more marked in larger equipment. As in the solids case, the coefficients of horizontal gas dispersion are a factor of 10–100 lower than those of vertical gas dispersion.

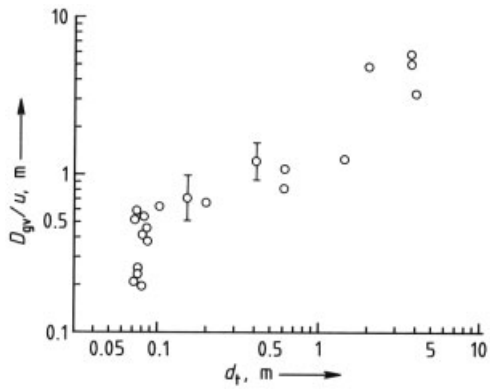

Figure 27. Vertical gas dispersion in a fluidized bed of solids of Geldart group A (measurements by various workers; [2])

A single-phase dispersion model gives only a rough description of gas mixing in bubbling fluidized beds. A more exact description comes from models that take account of local flow conditions in the bed, especially the presence of bubbles (see Chap. 9).

4.2. Gas Mixing in Circulating Fluidized Beds

Only a few detailed studies of gas mixing in circulating fluidized beds have been published, which are summarized in [120]. The bubbles

in a bubbling fluidized bed influence the gas residence-time distribution and mixing directly through the bypass action of the bubble-gas flow and gas exchange between the bubbles and the surrounding suspension phase, and also indirectly through the solids motion that they induce. In the circulating fluidized bed, on the other hand, the gas-mixing properties are controlled by segregation due to the formation of solid aggregates (jets, clusters) and the rapid downward movement of solids strands predominantly near the wall. Grace and coworkers, for example, show that a single-phase dispersion model cannot describe the tracer gas residence-time distributions that they measured [121]. They propose instead a two-phase model featuring exchange between a wall zone with stagnant gas and a core zone with plug flow.

For the case of *horizontal gas mixing*, Werther and coworkers [122, 123] have shown that, for the bed solids they used (quartz sand, d_p = 0.13 mm, Geldart group B), horizontal gas mixing in the top part of the circulating fluidized bed in the core zone can be described by the model of turbulent single-phase flow [124]. The Péclet number

$$Pe_{r,\,c} = \frac{u_c \cdot 2R^*}{D_{r,\,c}} \tag{44}$$

(defined in terms of u_c, the superficial velocity in the core zone; R^*, the radius of the core zone; and $D_{r,\,c}$, the horizontal dispersion coefficient in the core zone) has a value of 465, which is in fairly good agreement with values measured in single-phase flows [125]. This value is independent of the solids circulation rate G_s. The circulating fluidized bed thus exhibits no especially intensive horizontal gas mixing, at least in the upper section where solids concentrations are relatively low.

5. Heat and Mass Transfer in Fluidized-Bed Reactors

Fluidized-bed reactors exhibit a uniform temperature distribution even in case of highly exothermic or endothermic reactions. Approximations of the heat transfer rates are necessary for the design and control of fluidized-bed reactors in order to determine the appropriate design of internals for cooling or heating and to estimate the changes in the performance with changing operating conditions.

However, up to now there is no general theory on heat and mass transfer in fluidized beds. Numerous correlations for the calculation of heat and mass transfer coefficients are reported in the literature. Since these correlations are mainly based on experimental investigations performed under laboratory conditions, they may be different to the situation in large-scale reactors. Details on models of heat and mass transfer with their respective range of application are given in related surveys, e.g., [14 – 17, 126, 130].

Fluid-to-Particle Heat and Mass Transfer. Since the particle surface area is very large, fluid-to-particle heat and mass transfer is rarely a limiting factor in the design and operation of fluidized bed reactors. The heat-transfer coefficients of fluidized-beds range between characteristic values for flow through a fixed bed and flow around a single particle [127].

Fixed bed ($Re_p > 80$)

$$Nu = \frac{h_{gs} \cdot d_p}{\lambda_g} = 2 + 1.8\, Re_p^{0.5}\, Pr^{0.33}$$

Single particle

$$Nu = \frac{h_{gs} \cdot d_p}{\lambda_g} = 2 + 0.6\, Re_p^{0.5}\, Pr^{0.33}$$

where h_{gs} is the gas–solid heat-transfer coefficient, d_p the particle size, and λ_g the thermal conductivity of gas. The mass transfer coefficient can be determined applying the analogy of heat and mass transfer by replacing in the above formulas the Nusselt number *Nu* by the Sherwood number *Sh* and the Prandtl number *Pr* by the Schmidt number *Sc*. For particle Reynolds numbers below 100 and for fine particles, the transfer coefficients are significantly lower than estimated by the above formulas. If necessary, the effect of adsorption in mass transfer and of radiation in heat transfer needs to be taken into account additionally.

Heat Transfer to Submerged Surfaces. Heat-transfer coefficients between fluidized bed and submerged surfaces are one or two orders of magnitude larger than for gases alone [126]. For single phase flow a stagnant gas layer is established at the wall causing a hindered heat transfer. This layer is disrupted by solids transported

at the wall. The solids adsorb heat and are mixed into the fluidized bed [9].

An example of the time-averaged local heat transfer along the circumference of a tube immersed horizontally in a fluidized bed is given in Figure 28. It exhibits lower values of the heat-transfer coefficient below the tube due to a gas gap caused by bubbles and lower values on top of the tube because of solids being at rest. With intensified mixing this effect becomes less significant.

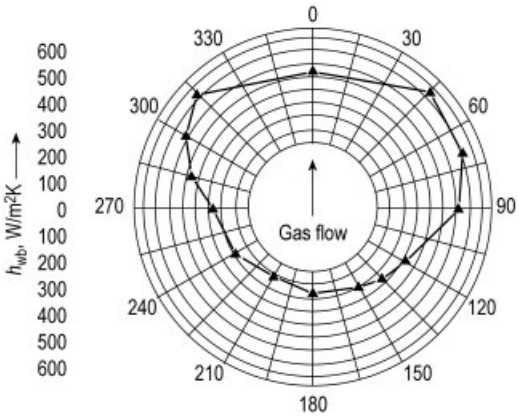

Figure 28. Local heat-transfer coefficient around a 35 mm diameter tube immersed horizontally in a fluidized bed of 0.37 mm alumina particles operated at a superficial gas velocity of 0.8 m/s and a temperature of 500 °C, adapted from [128]

The dependence of the heat-transfer coefficient on the superficial gas velocity is illustrated in Figure 29. Fluidized beds of fine particles yield a larger heat-transfer coefficient than coarse particles. According to Molerus and Wirth [126], different transfer mechanisms can be identified. In case of fine particles, solids act as agents transporting heat between walls and bed, whereas gas convective transport is the mechanism dominating the heat transfer of coarse particles. The heat-transfer coefficient of particles of intermediate sizes exhibits a maximum due to the superposition of these two transport mechanisms. Heat-transfer rates in circulating fluidized beds are lower than in bubbling fluidized beds due to reduced solids volume concentrations and are dominated by clusters and strands [130].

The heat-transfer coefficient increases with increasing pressure [131] and temperature. The effect of radiation has to be considered for temperatures above 500°C, but opaque particles can form an effective radiation shield [132].

6. Gas-Solid Separation

The fluidizing gas inevitably carries fine catalyst particles by entrainment to the reactor exit. Not only for environmental reasons (i.e., to minimize emissions) is it necessary to separate the solids from the gas. It may also be necessary to stop the main reaction and to avoid unwanted side or consecutive reactions or to protect following process steps or machines from particle-laden streams. In fluidized-bed technology cyclones are mostly used for this purpose. Knowlton [133] has given a survey on the state of the art of cyclone design and application in fluidized-bed reactors.

The cyclone should not be considered as a separate apparatus following the fluidized bed but should be seen as an integral part of the fluidized-bed process. The reason is that, not only in circulating fluidized beds but also in bubbling or turbulent fluidized beds, the catalyst particles which are recovered in the cyclone are recycled to the fluidized bed. The collection efficiency of the cyclone is thus responsible for maintaining the particle size distribution in the bed inventory, which in turn determines the fluidized-bed fluid mechanics and the chemical performance of the bed as a reactor. The interrelation between fluidized bed and cyclone is discussed in Section 9.3.2.

The influence of cyclone performance on the overall process performance is increasingly considered. For example, Pulupula et al. [134] investigated the role of cyclones in the regenerator system of a commercial FCC unit. Arnold et al. [135] were able to trace the deterioration of plant performance in the ALMA maleic process back to problems with cyclone efficiency. A change of the cyclone design improved the particle size distribution of the bed inventory and consequently bed hydrodynamics and chemical conversion. Smit et al. [136] report on cyclone performance in turbulent fluidized-bed Synthol reactors for Fischer–Tropsch synthesis. Carbon deposition on the catalyst particles influences the bed hydrodynamics, which in turn, via the

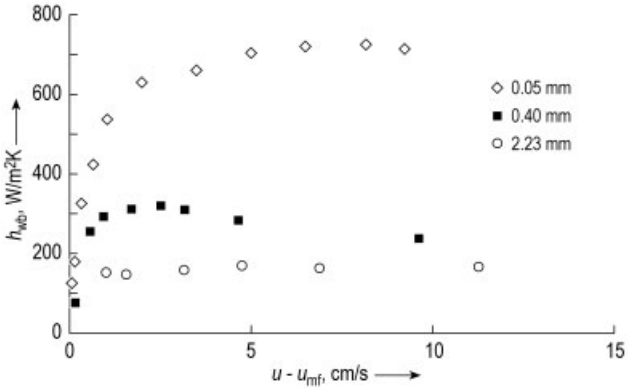

Figure 29. Heat-transfer coefficients determined with a tube immersed vertically in a fluidized bed of glass beads of different size operated at ambient conditions, adapted from [126, 129]

elutriation mechanism, influence cyclone performance.

7. Injection of Liquid Reactants into Fluidized Beds

The injection of reactants in liquid form into the bed was already an essential part of the first fluidized-bed catalytic process. In the FCC process (Section 8.1) crude oil is injected at the base of the reactor and evaporated in contact with the hot catalyst particles. The direct heat transfer is very efficient and avoids a separate evaporator for the feed. The cooling action of the evaporating reactant is a further advantage in the case of an exothermal reaction. Liquid-feed injection is therefore practiced not only in the FCC process but also, for example, in the syntheses of aniline (→ Fluid Mechanics), caprolactam (→ Fluid Mechanics), and melamine (→ Fluid Mechanics) and in BP Chemicals' Inovene process [137] for the gas-phase production of low-density polyethylene.

Despite its industrial significance knowledge about the mechanisms of liquid mixing and evaporation in the fluidized bed is relatively scarce. Investigations with nonvaporizing horizontal gas–liquid spray jets have shown that with proper design of the injection nozzle it is possible to penetrate over several decimeters into the bed before the jet breaks up [138, 139]. On the other hand, it was found that under vaporizing conditions for atomizer nozzles with spray angles between 20 and 120° the injected liquid wetted the bed particles and subsequently evaporated from their surface while the particles were mixing in the bulk of the bed [140, 141]. This latter mechanism helps to transport the reactant away from the location of the nozzle and thus contributes to equalization of the feed distribution inside the reactor. The special case that a large oil droplet impinges on a smaller hot catalyst particle was recently investigated in a 3D direct numerical simulation to analyze droplet–particle collisions in the Leidenfrost regime [142]. The calculations were carried out for conditions prevailing near the feed nozzle in an FCC riser. Vapor layer pressure induced by evaporation and the droplet surface tension are the driving forces for droplet recoiling and rebounding. The contact time for a FCC particle and an oil droplet turned out to be about 140 μs.

8. Industrial Applications

In this chapter the industrial uses of fluidized-bed reactors are classified as follows:

1) Heterogeneous catalytic gas-phase reactions
2) Polymerization of olefins
3) Homogeneous gas-phase reactions
4) Gas–solid reactions
5) Biotechnology applications

In each of these areas, the most important applications are listed and a few typical examples are analyzed in more detail. For further descriptions of processes, the reader is referred to relevant articles in the A series. Complete descriptions of industrial uses of the fluidized-bed reactor can also be found in [2, 10, 18, 19].

8.1. Heterogeneous Catalytic Gas-Phase Reactions

The fluidized-bed reactor offers the following principal advantages over the fixed-bed reactor for heterogeneous catalytic gas-phase reactions:

1) High temperature homogeneity, even with strongly exothermic reactions.
2) Easy solids handling, permitting continuous withdrawal of spent catalyst and addition of fresh if the catalyst rapidly loses its activity.
3) Ability to operate in the explosion range, provided the reactants are not mixed until they are inlet to the fluidized bed. This is because the high heat capacity of the bed solids, together with intensive solids mixing, prevents the propagation of explosions.

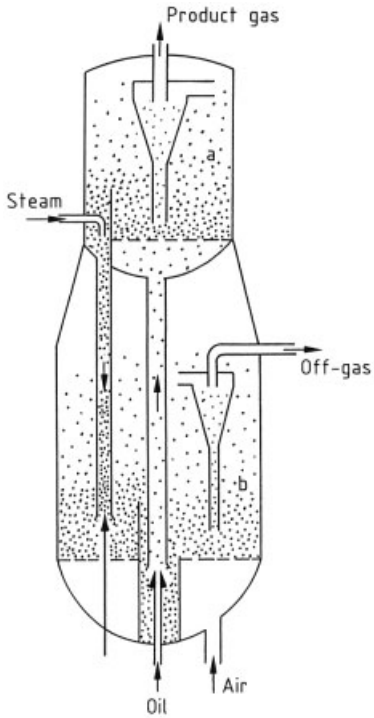

Figure 30. Fluid catalytic cracking process (Kellogg-Orthoflow system; according to [143, 144])
a) Reactor; b) Regenerator

Catalytic Cracking (→ Oil Refining, Chap. 3.2). The ease of solids handling was the basic reason for the success of catalytic cracking of long-chain hydrocarbons in the fluidized bed (Fig. 30). The cracking reaction is endothermic and involves the deposition of carbon on the catalyst surface, which quickly renders the catalyst inactive. Accordingly, the catalyst must be continuously discharged from the reactor and regenerated in an air-fluidized regenerator bed (b), where its carbon loading is lowered from 1–2 to 0.4–0.8 wt %. The combustion in this bed simultaneously furnishes the heat required for the cracking reactor; the catalyst acts as a heat carrier. The temperature in the regenerator is 570–590 °C and in the reactor, 480–540 °C [2]. In a stripper, steam is admitted to remove hydrocarbons adhering to the catalyst before it is forwarded to the regenerator.

With the advent of high-activity zeolite catalysts in the 1960s, the bubbling fluidized bed, operated at gas velocities between 0.31 and 0.76 m/s [2], was replaced by the *riser cracker* (Fig. 31), in which the oil fed in at the bottom of the riser (c) is vaporized in contact with the hot catalyst and the mixture of oil vapors and cracking gas transports the catalyst up through the riser. In the reactor bed (a), solids are collected before passing through the stripper (b) to the regenerator (f). By virtue of the short contact time of the order of a few seconds and the narrow gas residence-time distribution, the high activity of the zeolite catalyst is optimally utilized and a higher gasoline yield is achieved [2, 10].

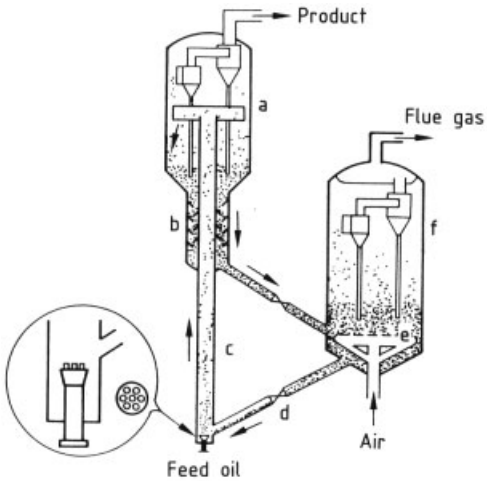

Figure 31. Riser cracking process (UOP system), [2]
a) Reactor; b) Stripper; c) Riser; d) Slide valve; e) Air grid; f) Regenerator

Synthesis of Acrylonitrile. The crucial factor in the successful use of the fluidized-bed reactor for the synthesis of acrylonitrile by the ammonoxidation of propene (Sohio process) (→ Acrylonitrile) was reliable control of this strongly exothermic reaction:

$$C_3H_6 + NH_3 + 3/2\,O_2 \rightarrow C_3H_3N + 3\,H_2O \quad \Delta H_r = -515\,\mathrm{kJ/mol\ of\ acrylonitrile.}$$

The reaction is carried out at a bed temperature of 400–500 °C and gas contact time of 1–15 s [145] or 5–20 s [2]. Figure 32 is a schematic of the reactor. Air is fed to the bottom of the fluidized-bed vessel. The reactants ammonia and propene are fed in through a separate distributor (b). Catalyst regeneration by carbon burnoff occurs in the space between the air distributor and the feed-gas distributor. The heat of reaction is removed by bundles of vertical tubes (a) inside the bed (horizontal tubes are used in other designs [146]).

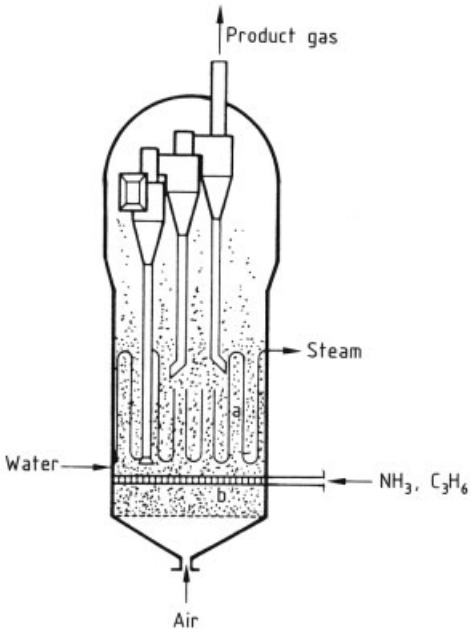

Figure 32. Synthesis of acrylonitrile (Sohio process) [2]
a) Cooler with internals; b) Distributor

Fischer–Tropsch Synthesis. The Fischer–Tropsch synthesis of hydrocarbons is used on a large scale for fuel production in the Republic of South Africa [149]. Synthesis gas generated from coal in Lurgi fixed-bed gasifiers enters the Synthol reactor (Fig. 33), where it is reacted over an iron catalyst at ca. 340 °C. The reactor works on the principle of the circulating fluidized bed. The mean porosity in the riser is 85 %, and the gas velocity varies between 3 and 12 m/s [2]. Reaction heat is removed by way of heat-exchanger tube bundles placed inside the riser.

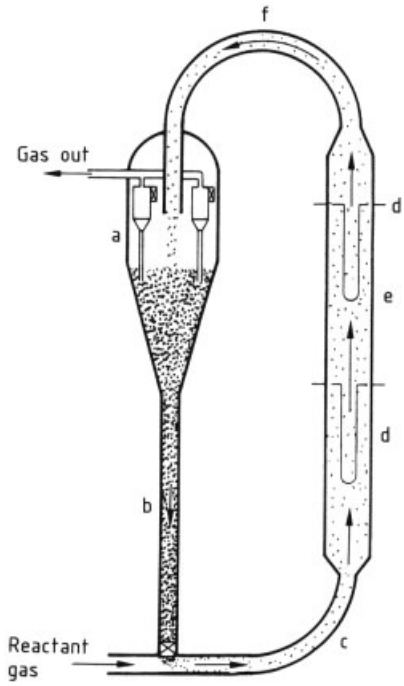

Figure 33. Fischer–Tropsch synthesis in the Synthol reactor [2, 147]
a) Hopper; b) Standpipe; c) Riser; d) Cooler (coil); e) Reactor; f) Gooseneck

However, experience has shown that this reactor is costly, relatively expensive to operate and maintain, and scale-up to the size of the reactors in operation is probably close to the maximum achievable for operation at 350 °C and 2.5 MPa. Therefore, in the 1990s the 16 circulating fluidized-bed reactors operating at Sasol's Secunda site were replaced by eight turbulent fluidized-bed reactors each of 10.7 m diameter, which achieve a higher per-pass syngas conversion [150].

Different process routes have been developed for the synthesis of maleic anhydride. The Mitsubishi process [152, 153] used the naphtha

cracker C_4 fraction. The ALMA process uses *n*-butane as feedstock [154, 155]. A more recent development is the Du Pont process, which is also based on *n*-butane but uses a circulating fluidized bed as reactor (Fig. 34) [156]. It is based on a vanadium phosphorus oxide (VPO) catalyst which oxidizes *n*-butane to maleic anhydride by a redox mechanism on its surface layers [157]. In the riser *n*-butane is selectively oxidized by the oxidized catalyst. In the fluidized-bed regenerator the spent catalyst is reoxidized. Since 1996 a commercial plant has been operating in Asturias, Spain [158].

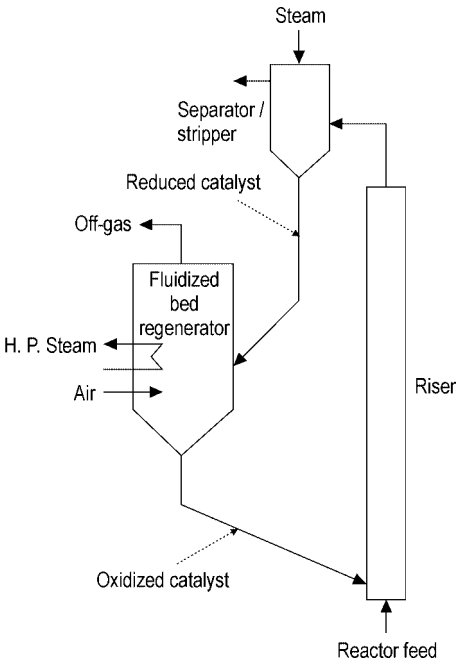

Figure 34. The Du Pont maleic anhydride process [158].

Other Processes. Other catalytic reactions carried out in fluidized-bed reactors are the oxidation of naphthalene to phthalic anhydride (→ Phthalic Acid and Derivatives) [2, 10, 151]; the ammoxidation of isobutane to methacrylonitrile [2]; the reaction of acetylene with acetic acid to vinyl acetate [2]; the oxychlorination of ethylene to 1,2-dichloroethane (→ Chlorinated Hydrocarbons) [2, 10, 159, 160]; the chlorination of methane [2]; the reaction of phenol with methanol to cresol and 2,6-xylenol [2, 161]; the reaction of methanol to gasoline [162, 163]; the synthesis of phthalonitrile by ammoxidation of *o*-xylene (→ Phthalic Acid and Derivatives) [164]; the synthesis of aniline by gas-phase hydrogenation of nitrobenzene (→ Aniline, Chap. 3.2) [165]; and the low-pressure synthesis of melamine from urea (→ Melamine and Guanamines) [166].

An overview on the various fluidized-bed catalytic processes has been given [167].

8.2. Polymerization of Olefins

The gas-phase polymerization of *ethylene* in the fluidized bed was developed by Union Carbide (Unipol process [168]; see Fig. 35) (→ Polyolefins). The reaction gas (ethylene and its comonomers butene or hexene) fluidizes the bed at 75–115 °C and 20–30 bar. Extremely fine-grained catalyst is metered into the bed. Polymerization occurs on the catalyst surface and yields a granular product with diameter ranging from 0.25 to 1 mm. Ethylene conversion is comparatively low, 2 % per pass; so the reaction gas is recycled. The heat of reaction is removed by cooling the recirculating gas. The catalysts used have such a high activity that more than 10^5 parts by volume of polymer can be produced per unit weight of active substance in the catalyst [2]. Because of the high degree of catalyst dilution in the granular polymer, the catalyst need not be removed from the product. In the process developed by BP Chemicals, prepolymers with a diameter from 0.2 to 0.25 mm rather than catalyst particles are fed into the fluidized bed [169].

Mitsui Petrochemical Industries has developed a process for the gas-phase fluidized-bed polymerization of propene (→ Polyolefins); a plant using this process came on stream in 1984 [170]. The Unipol–Shell process was jointly developed by Union Carbide and Shell and commissioned in 1986.

Burdett et al. [171] have given a broad overview on this still-developing technology, which presents many challenges for the engineer. One of the biggest problems is the stickiness of the particles under the operating conditions of the process, which has often led to particle sintering with subsequent defluidization of the bed. Seville et al. [172] monitored the motion of particles in a scaled polymer reactor and studied the sintering kinetics in order to determine

a safe operating window. Cai and Burdett [173] developed a model of single-particle polymerization in the fluidized bed to simulate particle growth and particle-temperature evolution with the residence time of a catalyst particle in the reactor.

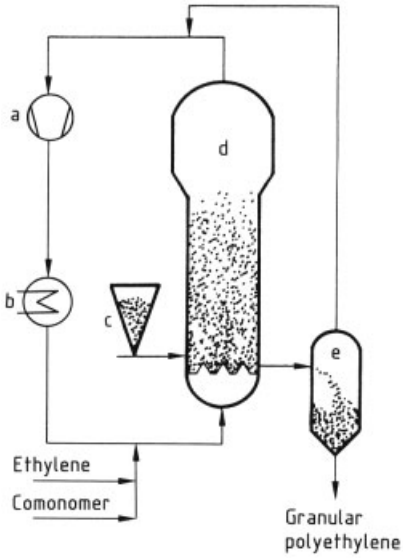

Figure 35. Gas-phase polymerization of ethylene (Unipol process) [2]
a) Compressor; b) Cooler; c) Catalyst feed hopper; d) Reactor; e) Separator

8.3. Homogeneous Gas-Phase Reactions

The decisive advantage of the fluidized bed for homogeneous gas-phase reactions is the ability to carry large quantities of heat into or out of the reactor by using direct heat exchange via the bed solids. An example is the *Exxon fluid coking process* (Fig. 36; [2, 18, 174, 175]), which converts heavy residual oils to petroleum coke and gas-oil. The reactor (d) and heater (e) beds are connected in a single solids loop. The bed material is coke generated in coking at 480–570 °C, which grows to spherical particles 0.1–1 mm in diameter in the reactor. The coke is discharged continuously from the reactor and heated to 500–690 °C by partial combustion in the heater. The hot coke stream then transports the heat needed for the endothermic coking reaction into the reactor. Excess coke is removed as a coarse fraction in a classifier connected to the heater. Fluid coking is used, e.g., for refining bitumen from the Athabasca tar sands in Canada. To make efficient use of the product coke, Exxon combined the fluid coking process with a fluidized-bed gasification reactor [2, 175]. This Flexi-Coking process was first implemented in 1976 in Japan; the daily capacity of one plant is ca. 3400 t of vacuum residue.

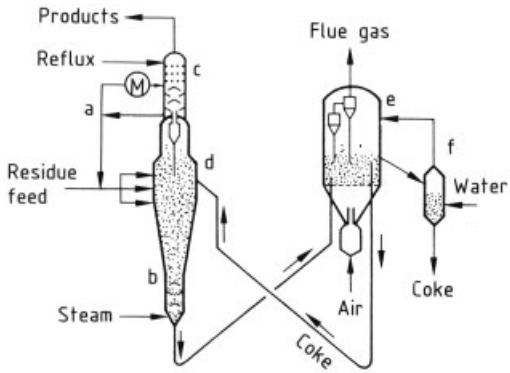

Figure 36. Fluid coking process [2, 18]
a) Slurry recycle; b) Stripper; c) Scrubber; d) Reactor; e) Heater; f) Quench elutriator

The bed solids also find use as heat-transfer agents in the *thermal cracking of naphtha*, a process carried out in the Lurgi sand cracker [2, 18, 176]. The solids circulating between the reactor and the heater consist of coarse sand particles (ca. 1 mm in diameter). When the coke deposit produced in cracking is burned off the particle surface with air, the solids are heated to 800–850 °C and can thus deliver the heat required for endothermic cracking. The temperature in the reactor is ca. 700–750 °C.

Other thermal cracking processes include the BASF Wirbelfliess process [2, 18, 177], and the Kunii–Kunugi process [2].

8.4. Gas–Solid Reactions

Coal Combustion. The high heat capacity of the fluidized bed permits stable combustion at low temperature (ca. 850 °C), so that the formation of thermal and prompt nitrogen oxides [178] can be suppressed and total nitrogen oxide emissions can be reduced. If limestone is added to the bed, the calcination reaction

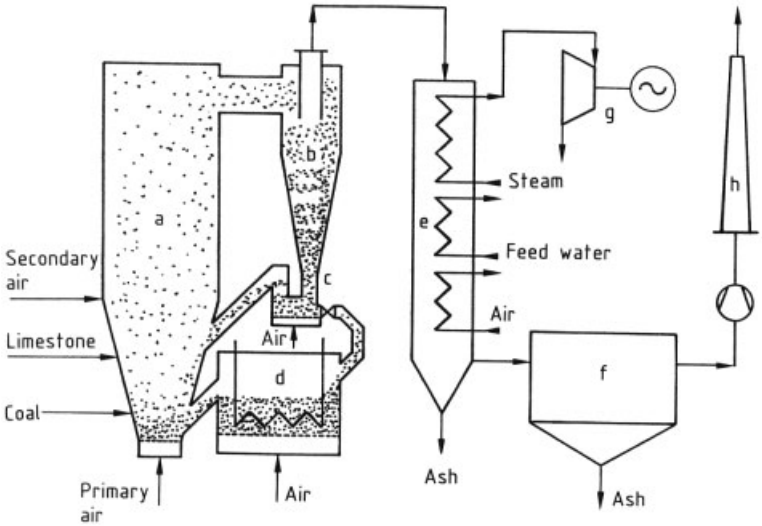

Figure 37. Power plant with circulating fluidized-bed furnace (Lurgi process) [180]
a) Circulating fluidized-bed reactor; b) Recycling cyclone; c) Siphon; d) Fluidized-bed heat exchanger; e) Convective pass; f) Dust filter; g) Turbine; h) Stack

$$CaCO_3 \rightarrow CaO + CO_2$$

yields CaO, which can bind in situ the SO_2 produced in combustion:

$$SO_2 + CaO + 1/2\,O_2 \rightarrow CaSO_4$$

During the 1980s the fluidized bed was established in power-plant engineering. The unit size rapidly increased from 5 MW_e in 1970 to about 350 MW_e during this time [179]. Meanwhile (ca. 2006) some 500 power plants are in operation worldwide. By far the majority of these plants operate with circulating fluidized beds. As an example, Figure 37 shows a Lurgi design.

The staged admission of the combustion air minimizes NO production from nitrogen in the fuel in the lower part of the combustion chamber. The admission of secondary air completes the combustion in the upper part of the chamber by oxidizing most of the CO. Some of the circulating solids are led through the external fluidized-bed cooler, which enhances the flexibility of control and permits load variation over a wide range.

More recent developments aim at even larger capacities with a further enlargement of the combustion chamber and making use of supercritical steam conditions and once-through boiler design. One problem associated with the size enlargement is the distribution of both the coal and the secondary air from the sidewalls over the cross section of the combustion chamber. Since the lateral mixing of gas and solids in a circulating fluidized bed is quite slow, sufficient numbers of feed ports for the coal and air injection nozzles have to be arranged on the sidewalls. The "pants-leg" design shown in Figure 38 is one possibility to provide sufficient lateral mixing at the bottom of the combustion chamber. A first 450 MW_e unit is being built in Lagisza/Poland [181] and 600 MW_e CFB combustors are in the design phase [182].

If a fluidized-bed furnace running under a pressure of 12–16 bar is linked to a gas turbine, the efficiency of the power plant can be markedly enhanced [183]. At the same time, however, this concept imposes severe requirements on gas cleaning [184]. Pressurized fluidized-bed combustion has been tested in several large experimental plants (e.g., [186]). In the meantime several plants of the 80 MW range are in commercial operation. Recently, a pressurized fluidized-bed combustor with an electrical power of 360 MW was erected by ABB Carbon [185] (Fig. 39); these units employ bubbling fluidized beds [187, 188]. Pressurized fluidized-bed boilers employing circulating fluidized beds are still under development (e.g., [189]).

For further details on fluidized-bed combustion systems see the proceedings of the Fluidized

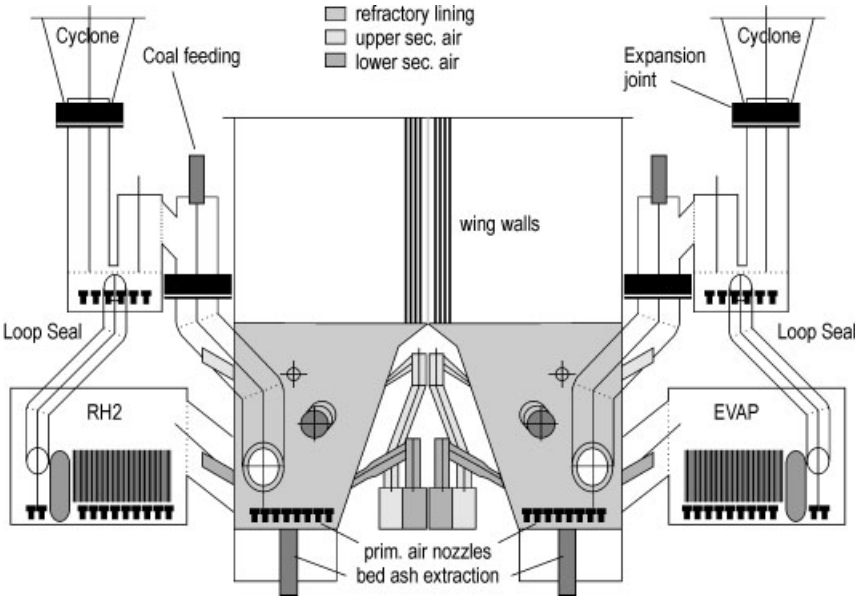

Figure 38. Furnace cross section of a large CFB combustor (after [179])

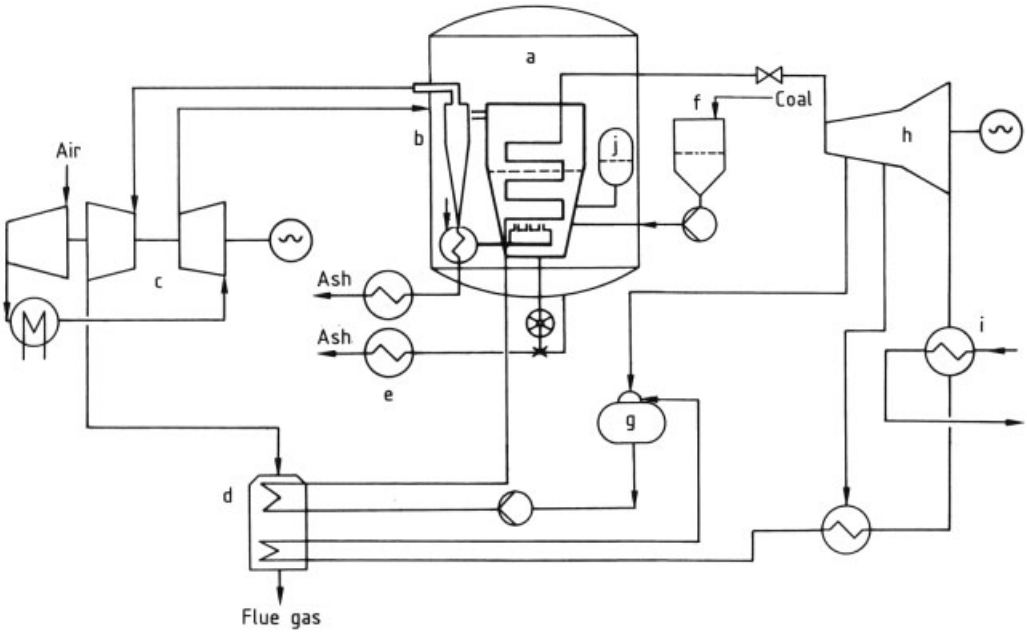

Figure 39. Power plant with pressurized fluidized-bed combustor (ABB design) [187]
a) Pressurized fluidized-bed boiler; b) Cyclone; c) Gas turbine; d) Economizer; e) Ash removal; f) Fuel feed; g) Feed-water tank; h) Steam turbine; i) Condenser; j) Bed material hopper

Bed Combustion Conferences [25] and a monograph [190].

Waste Incineration. The incineration of *municipal sewage sludge* in fluidized-bed furnaces is now practiced in many facilities [191]. The total amount of sludge thermally treated in Germany in 1996 was 513 000 t (dry basis) [192]. Figure 40 is a diagram of an incinerator. Moist, centrifuged sludge is fed from above to

the fluidized bed by means of piston pumps. The fluidizing air is preheated in an oil-fired muffle before reaching the furnace. Fuel oil, used as a supplemental fuel, is metered directly into the bed. Developments in sludge incineration have achieved energy autarky by recovering waste heat and utilizing it to predry the sludge so that self-sustaining combustion is possible [194 – 196]. Later, the more stringent emission limits set forth in the 17th Bundesimmissionschutzverordnung (BImSchV, regulation in the Federal Republic of Germany concerning the limitation of immissions) may necessitate staged combustion (as in power generation), particularly to control NO_x emissions [197, 198].

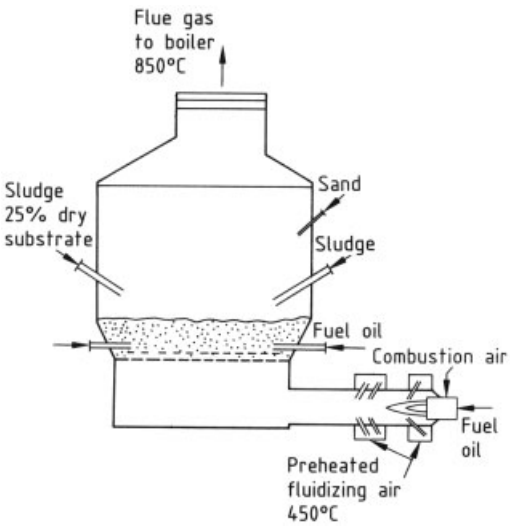

Figure 40. Fluidized-bed furnace for municipal sludge incineration (Uhde system) [193]

For the *incineration of municipal waste*, a furnace with a "rotating" fluidized bed has been developed. The inclined distribution grid in this design generates two rollerlike flows of circulating bed solids, leading to rapid and efficient mixing of the waste in the bed [199].

Coal Gasification. A number of fluidized-bed processes have been developed for gasifying coal (→ Gas Production; e.g., [2]). Interest in these processes for cogeneration power plants has recently become more intense. In the cogeneration system shown in Figure 41, high-pressure gasification is combined with combustion in a circulating fluidized bed; efficiencies of more than 40 % are expected, depending on the available gas turbine technology [180]. The Rheinische Braunkohlenwerke company has developed a high-temperature Winkler (HTW) process based on Winkler gasification (Fig. 42) [200]. The pressure (ca. 10 bar) and temperature (ca. 1100 °C) are higher than in the Winkler process; coal is gasified with oxygen and steam. Recycling of solids from the cyclone to the fluidized bed results in a much higher carbon conversion than in the Winkler process.

Gasification of Solid Waste. In comparison with incineration, the gasification of solid waste offers the advantage of a smaller volume of off-gas, so the cleaning system can be made smaller. In the Japanese Pyrox process, heat required by the gasification reactor is supplied by sand heated in a fluidized-bed furnace, so that a high-Btu-gas can be generated [2].

A cement kiln plant at Ruedersdorf in Germany is operated with a biomass gasification reactor. This circulating fluidized-bed reactor designed by Lurgi supplies fuel gas for the calciners [201].

Calcination. The calcination of aluminum hydroxide in the Vereinigte Aluminiumwerke (VAW)/Lurgi circulating fluidized-bed process features an overall thermal efficiency of more than 70 %, which is achieved through downstream heat recovery from the calcined alumina and the off-gas [18, 202 – 204]. The circulating fluidized bed proper (c) is coupled to two venturi fluidized beds (a), in which the moist hydroxide is first dried and heated by direct contact with the off-gas before it is forwarded to the calcination furnace (see Fig. 43). The five-stage fluidized-bed cooler (d) downstream of the furnace serves to preheat the combustion air. A furnace 3.8 m in diameter and 20 m tall, with a fluidization velocity of 3 m/s and mean particle diameter of 0.04–0.05 mm, produces more than 500 t/d of Al_2O_3 [18].

Other applications include the calcination of limestone (in multistage fluidized-bed furnaces), lime muds, and crude phosphates [2, 18].

Roasting Processes. Fluidized-bed roasting follows the general reaction equation

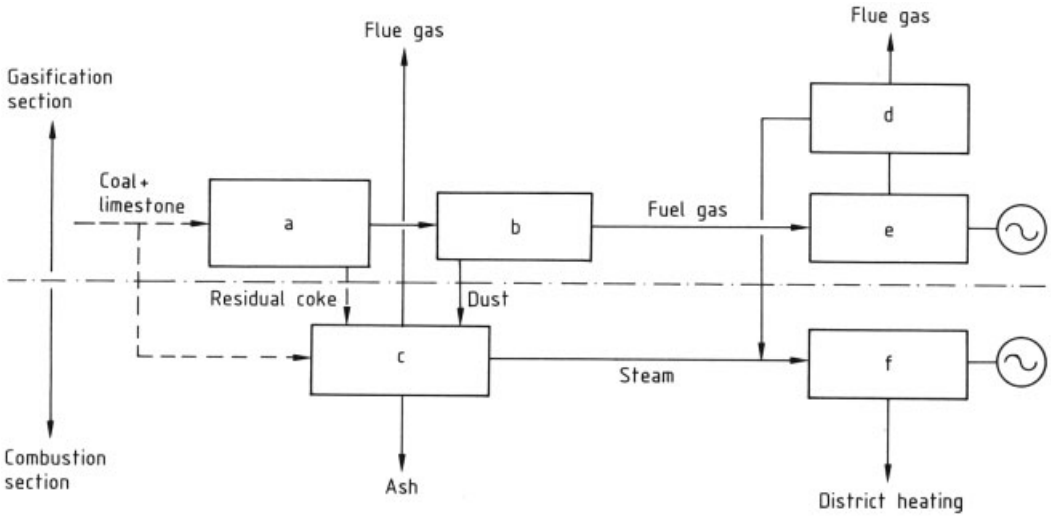

Figure 41. Concept for cogeneration power plant based on high-pressure gasification in circulating fluidized bed [185]
a) High-pressure CFB gasification; b) Gas cleaning; c) CFB combustion; d) Waste-heat boiler; e) Gas turbine; f) Steam turbine

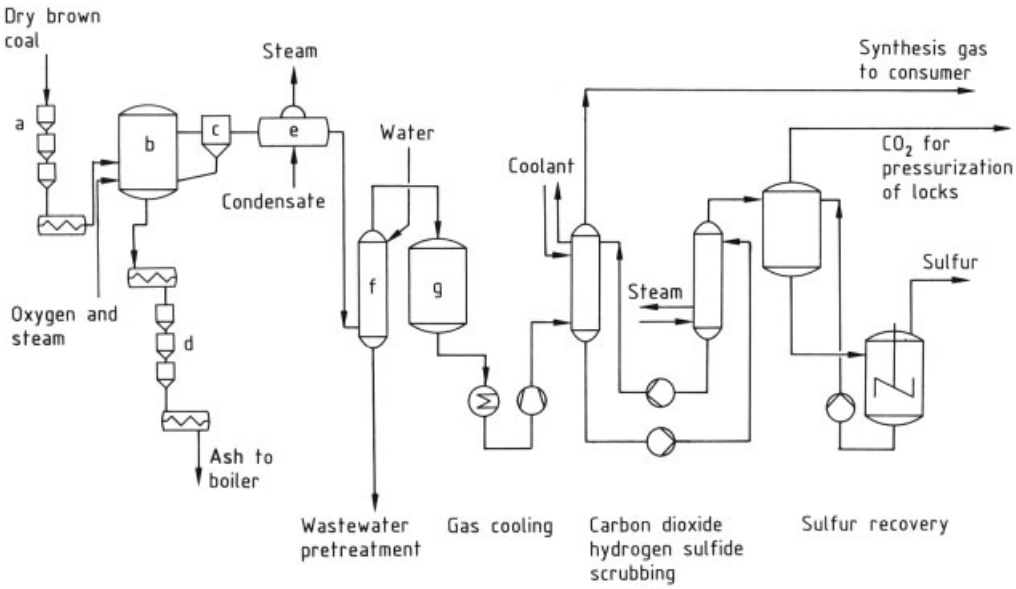

Figure 42. Flow sheet of HTW demonstration plant [200]
a) Coal lock hopper system; b) Gasifier; c) Cyclone; d) Ash lock hopper system; e) Raw gas cooler; f) Wet dust separator; g) Carbon monoxide conversion

Metal sulfide + Atmospheric oxygen

→ Metal oxide + Sulfur dioxide

This is one of the earliest industrial uses of fluidization. Many such processes are used in the roasting of pyrite, zinc blende, and other sulfide ores. Bubbling fluidized beds with gas velocities between 0.5 and 2.3 m/s [2] are employed; heat generated by the exothermic roasting reaction is removed by tube banks immersed in the bed, via a solid heat-transfer agent, or by simple water injection. Roasting furnaces are available in very large sizes (bed diameters up to 11 m) with capacities of several hundred tonnes of ore per day [2, 18].

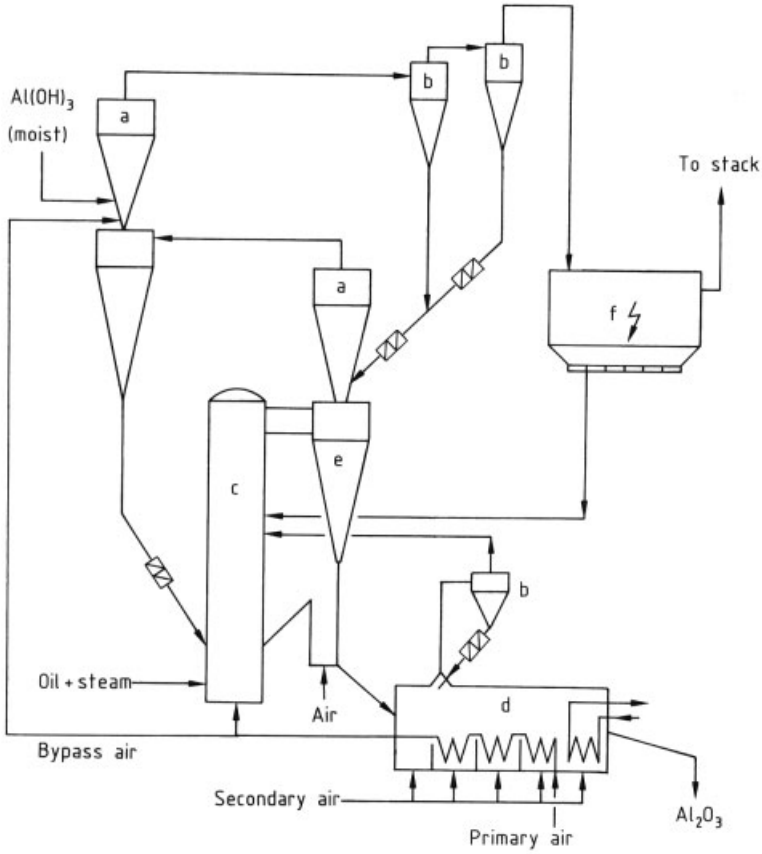

Figure 43. Fluidized-bed calcination of aluminum hydroxide (VAW/Lurgi system) [18]
a) Venturi fluidized bed; b) Cyclones; c) Fluidized-bed furnace; d) Fluidized-bed cooler; e) Recycle cylone; f) Electrostatic precipitator

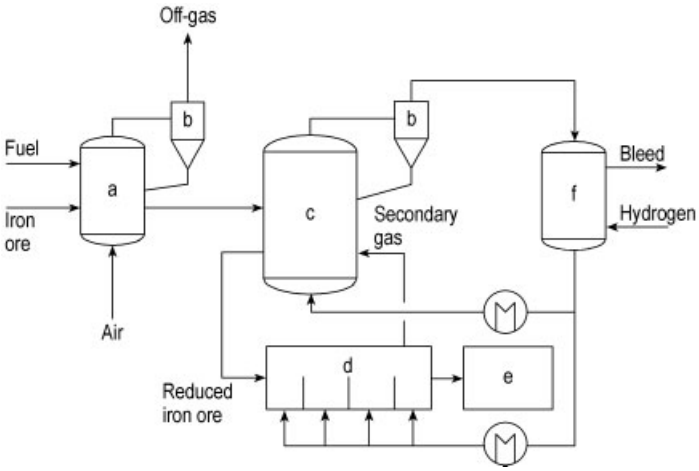

Figure 44. Flow sheet of Lurgi Circored process [206]
a) Preheater; b) Cyclone; c) First stage reactor; d) Second stage reactor; e) Briquetting unit; f) Gas cleaning unit

Iron Ore Direct Reduction. (→ Iron). For the direct reduction of iron ore, Lurgi has developed two processes [205]. Hydrogen is applied as reductant in the Circored process, coal gas is used in the Circofer process. A plant for the Circored process has been built in Trinidad with a capacity of 500 000 t iron briquette per year [206]. Applying two stages, a circulating fluidized-bed reactor reduces the preheated iron ore (800 °C) at 630–650 °C to a degree of metallization of 65–85 % and a bubbling fluidized-bed reactor proceeds at temperatures up to 680 °C to achieve a degree of metallization of 93–95 % (see Fig. 44). The metallized product is then transported to the hot briquetting unit. At a pressure of 4 bar, the process gas is recycled to a gas cleaning unit and made up with hydrogen.

Other Processes. Fluidized-bed processes for the production of high-purity silicon and activated carbon and the chlorination and fluorination of metal oxides are described in [2]. A detailed description of TiO_2 synthesis in a fluidized-bed reactor and a survey of the use of fluidized-bed processes in the production of nuclear fuels are given in [10].

8.5. Applications in Biotechnology

A comprehensive survey of the use of fluidized-bed reactors in biotechnology is given in [10, 207].

Liquid–solid and liquid–gas–solid systems are used in aerobic and anaerobic wastewater treatment (nitrification and denitrification); the microorganisms are grown as a biofilm on particulate supports to prevent their entrainment from the reactor with the fluidizing medium. The advantages of the fluidized-bed reactor over the fixed-bed reactor include higher capacity per unit volume and less susceptibility to plugging [208]. A study showing the potential of liquid–solid circulating fluidized beds in biotechnological processes such as fermentation has been published recently [209].

Full-scale fluidized-bed *biogas production reactors* have come on stream since 1984 [210, 211]. The process consists of two stages, acidification and methanation. Sand particles 0.1–0.3 mm in diameter serve as support for the microorganisms; at fluidization velocities of 8–20 m/h, the biofilm grows to a thickness of 0.06–0.2 mm on these particles. The reactors are large devices with diameters of 4.6 m and bed heights of 21 m.

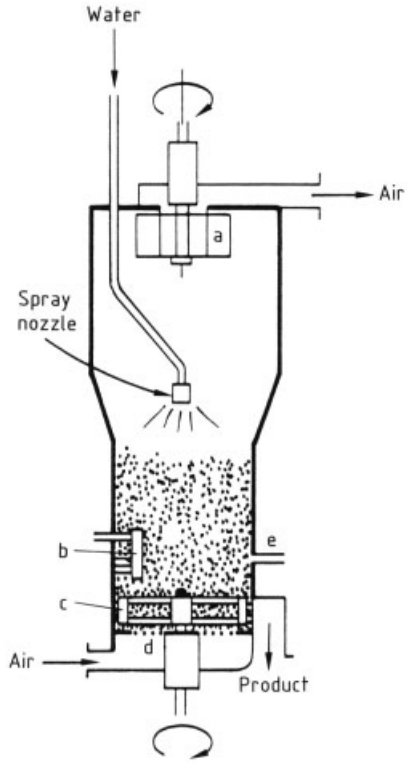

Figure 45. Solid-state fermentation of *Aspergillus sojae* in the fluidized bed (adapted from [216])
a) Separator; b) Electrode; c) Agitator; d) Distributor; e) Ejector

Gas–Solid Systems. Gas–solid fluidized-bed fermenters have been investigated on a pilot scale for the growth of *Saccharomyces cerevisiae* [212], the production of ethanol with *S. cerevisiae* [213, 214], and the enrichment of glutathione in yeast by *S. cerevisiae* [214, 215]. In these applications, the substrate is metered into the bed in liquid form. A process used in Japan for the culture of *Aspergillus sojae* on wheat groats employs a solid substrate [216, 217]. The latter process is in service on a plant scale (bed mass 500 kg, bed diameter 1.5 m). The reactor (Fig. 45) contains an agitator (c) just above the distributor (d), as well as a rotating separator (a) in the top of the vessel. Water is sprayed onto

the bed from above to maintain the proper moisture level; electrodes (b) dipping into the bed measure this parameter. The moisture content of the solids is generally a critical parameter for the fluidized-bed fermenter; the bioreactions extinguish if it becomes too low, whereas the particles agglomerate and fluidization is disrupted if it becomes too high. Sterilized air is used for fluidization. Seed spores of the microorganisms are fed into the bed via the ejector (e). This system achieves a considerable gain in cell yield and an enrichment of certain enzymes by a factor of 5–15 over conventional fixed-bed cultures. The generated biomass forms the basis for soy sauce production.

9. Modeling of Fluidized-Bed Reactors

9.1. Modeling of Liquid–Solid Fluidized-Bed Reactors

An expansion formula of the Richardson–Zaki type, Equation (7), describes the hydrodynamics of liquid–solid fluidized beds fairly well [218]. The difficulty in modeling this kind of reactor for bioreactions thus lies not so much in determining the flow and mixing conditions in the fluid as in describing the diffusion processes in the biofilm and the kinetics of the biological reactions [219]. In view of the small number of experimental studies reported thus far, no final judgment can be made on the suitability of various models [208].

9.2. Modeling of Gas–Solid Fluidized-Bed Reactors

Exhaustive literature surveys can be found in [2, 9, 10, 220]. [221]. Many models exist in the literature, which are classified in the cited references under various schemes. The available information can be summed up as follows: No generally accepted model of the fluidized-bed reactor exists; instead, many models have been proposed on the basis of more-or-less extensive experimental findings for various applications.

Any fluidized-bed reactor model can be broken down into separate components that describe, with varying degrees of accuracy, the hydrodynamics (depending on solid properties, operating conditions, and geometry), gas–solid contact, and reaction kinetics. The essential point is that the reactor geometry effect, which is important for scale-up (Chap. 10), manifests itself in the flow conditions and must therefore be included in the hydrodynamic part of the model.

Before a reactor model found in the literature can be applied to a given problem, the designer must determine whether numerical values are available for all model parameters, that is, whether the model is appropriate for design calculations or is a "learning model" [222] in which the numerical values of important parameters can be determined only after the model is adapted to actual test results.

Reaction kinetics may be determined in a fixed-bed reactor, provided measurements are performed under conditions comparable to those that prevail in the fluidized-bed reactor (e.g., the same solids composition and particle-size distribution, the same activity state in the case of catalysts) [223]. However, the kinetic parameters can also be determined directly by measurements in a bench-scale fluidized-bed apparatus [224].

9.2.1. Bubbling Fluidized-Bed Reactors

By far the majority of fluidized-bed reactor models described in the literature deal with reactions in bubbling fluidized beds [2, 9, 10, 225, 226]. For a specific application, modeling depends on the bubble flow regime. For *slow-bubble systems* (Fig. 7, left), the short-circuit flow of gas through the bubbles must be taken into account [227]. For *fast-bubble systems* (Fig. 7, right), the species have to be balanced separately in the bubble and suspension phases.

If models from the literature are employed, it should be taken into account that those devised in the past, when adequate computing hardware was not available, often sought to obtain analytical expressions for the degree of conversion of a single reaction (usually taken as first-order). The simplifying assumption of a single "effective" bubble size for the entire fluidized bed was therefore made [2], or the mass-transfer area between the bubble and suspension phases was taken as uniformly distributed over the height of the bed (HTU or NTU concept, where HTU

denotes height of transfer unit and NTU denotes number of transfer units [228]. Today, in view of the computing power available at the PC level, the recommended procedure is to start from local mass-transfer relations, write balance equations for the differential volume element of the reactor, and then numerically integrate these equations.

Figure 46 presents a model used by WERTHER for a constant-volume reaction [224, 229]. Here the simplifying assumption is that flow through the suspension phase is at the minimum fluidization velocity u_{mf}. For a heterogeneous catalytic gas-phase reaction, the material balances for species i in the unsteady-state cases are as follows:

Bubble phase

$$\varepsilon_{\mathrm{b}} \frac{\delta C_{\mathrm{b}i}}{\delta t} = -\left[u - u_{\mathrm{mf}}\left(1-\varepsilon_{\mathrm{b}}\right)\right] \cdot \frac{\delta C_{\mathrm{b}i}}{\delta h} - k_{\mathrm{G},i} a \left(C_{\mathrm{b}i} - C_{\mathrm{d}i}\right) \tag{45}$$

Suspension phase

$$\begin{aligned}&\left(1-\varepsilon_{\mathrm{b}}\right)\left[\varepsilon_{\mathrm{mf}} + \left(1-\varepsilon_{\mathrm{mf}}\right)\varepsilon_i\right] \frac{\delta C_{\mathrm{d}i}}{\delta t} = \\ &-u_{\mathrm{mf}}\left(1-\varepsilon_{\mathrm{b}}\right) \cdot \frac{\delta C_{\mathrm{d}i}}{\delta h} + k_{\mathrm{G},i} \cdot a \left(C_{\mathrm{b}i} - C_{\mathrm{d}i}\right) \\ &+ \left(1-\varepsilon_{\mathrm{b}}\right) \cdot \left(1-\varepsilon_{\mathrm{mf}}\right) \varrho_{\mathrm{s}} \sum_{j=1}^{M} \nu_{i,j}\, r_j\end{aligned} \tag{46}$$

In Equations (45) and (46) the following simplifying assumptions have been made:

1) Plug flow through the suspension phase at an interstitial velocity ($u_{\mathrm{mf}}/\varepsilon_{\mathrm{mf}}$)
2) Bubble phase in plug flow, bubbles are solids free
3) Reaction in suspension phase only
4) Constant-volume reaction (see [224] for handling a change in number of moles)
5) Sorption effects neglected (see [229] for handling sorption)

Here ε_i is the porosity of the catalyst particles; a is the local mass-transfer area per unit of fluidized-bed volume, which can be calculated as

$$a = \frac{6\varepsilon_{\mathrm{b}}}{d_v} \tag{47}$$

for spherical bubbles; r_j is the rate of partial reaction j per unit mass of catalyst; and ν_{ij} is the stoichiometric number of species i in reaction j. The relation

$$k_{G,i} = \frac{u_{\mathrm{mf}}}{3} + \sqrt{\frac{4 D_i \varepsilon_{\mathrm{mf}} u_b}{\pi d_v}} \tag{48}$$

proposed by SIT and GRACE [230] has proved useful for describing the mass-transfer coefficient $k_{\mathrm{G}i}$ associated with component i in mass transfer between the bubble and suspension phases; D_i is the molecular diffusion coefficient of species i.

The *freeboard space* above the bubbling fluidized bed must be considered in the reactor model if the entrainment rate is high and the reactions in the freeboard are not quenched, for example, by cooling.

Most fluidized-bed models include concentration profiles only for the vertical direction. This one-dimensional modeling is acceptable when the reactants are admitted uniformly over the bed cross section. If, however, reactants are metered into the bed at individual feed points, three-dimensional modeling may become necessary. Such models have been devised for the combustion of coal in bubbling fluidized beds [232 – 234].

As a rule, the *modeling of solids* behavior in bubbling fluidized-bed reactors is based on that in stirred tanks. Fluidized-bed combustion is one of the few exceptions; here the model must take account of the propagation of coal from the feed point if the furnace emission behavior is to be described correctly [232, 234].

Temperature homogeneity is a virtually fundamental property of the fluidized-bed reactor. Even so, one exception is industrially important: in high-pressure fluidized-bed furnaces, the high energy density can cause local hot spots near the fuel injection points [235]. Reactor models that take care of this have been described, e.g., in [236].

9.2.2. Circulating Fluidized-Bed Reactors

In the early days of circulating fluidized-bed reactor modeling, negligible axial dispersion and laterally uniform flow structure were believed to characterize these systems. Thus, simple plug-flow models were used [237]. This approach was found to oversimplify the behavior of circulating fluidized-bed reactors, because a significant amount of axial dispersion was observed. As a result, the plug-flow model has often been modified by adding a dispersion term to the balance equations. Axial dispersion coefficients have

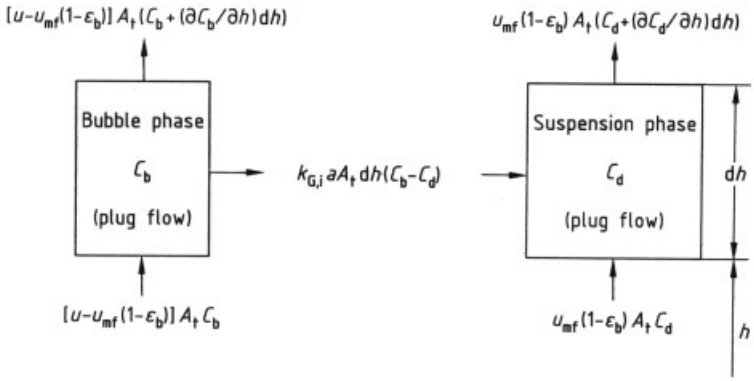

Figure 46. Two-phase model of the fluidized-bed reactor

been determined by many authors who measured the residence time distribution of tracer gases [238, 239]. Typical values of Péclet numbers they found are on the order of 10.

By means of a model reaction it has been proved that in many cases circulating fluidized-bed reactors cannot be characterized by solely considering mixing phenomena [240]. Instead, the presence of mass-transfer limitations and bypassing was found to have a significant influence. In analogy to low-velocity fluidized beds a detailed description of the local flow structure within the reaction volume must serve as a basis for appropriate reactor modeling.

The highly nonuniform flow structure of circulating fluidized beds described in 2.9.2 has led to reactor models which separately deal with different axial zones. The bottom zone–if it exists under the given operating conditions–can be described by models whose basic approaches were originally developed for modeling of bubbling fluidized beds [241]. Modeling of the upper section of the circulating fluidized bed is in most cases based on a proper description of the heterogeneous core–annulus flow structure [242 – 244]. These state-of-the-art models are one-dimensional and define two phases or zones which are present at every axial location:

1) A dense phase or annulus zone: high solids concentration, gas stagnant or moving downwards
2) A dilute phase or core zone: low solids concentration, gas flowing quickly upward.

Similar to the situation in bubbling fluidized beds, the two phases exchange gas with each other and are modeled by separate equations which are obtained from mass balances for each component in each phase.

Today's models still suffer from the problem that not all fluid-mechanical variables can be predicted on the basis of the operating conditions. Instead, reasonable estimations or measurements in cold-flow models are used to obtain numerical values for many variables.

A common feature of all models for the upper part of circulating fluidized beds is the description of the mass exchange between dense phase and dilute phase. In analogy to low-velocity fluidized beds the product of the local specific mass transfer area a and the mass-transfer coefficient k may be used for this purpose. Many different methods for determination of values for these important variables have been reported, such as tracer-gas backmixing experiments [241], non-steady-state tracer-gas experiments [245], model reactions [244] and theoretical calculations [243].

Similar to the bubbling fluidized-bed reactor, the solids behavior of the circulating fluidized-bed reactor can usually be described as completely mixed. This does not hold for riser reactors with very high gas velocities, such as those used in FCC risers ($u > 10\ ms^{-1}$). Here, better modeling results will be obtained by assuming dispersed plug flow of solids [239].

Like for bubbling fluidized beds, it can be assumed that circulating fluidized beds exhibit a high degree of temperature homogeneity even in the case of highly exothermic reactions. However, in the case of very large circulating fluidized beds for coal combustion, significant hor-

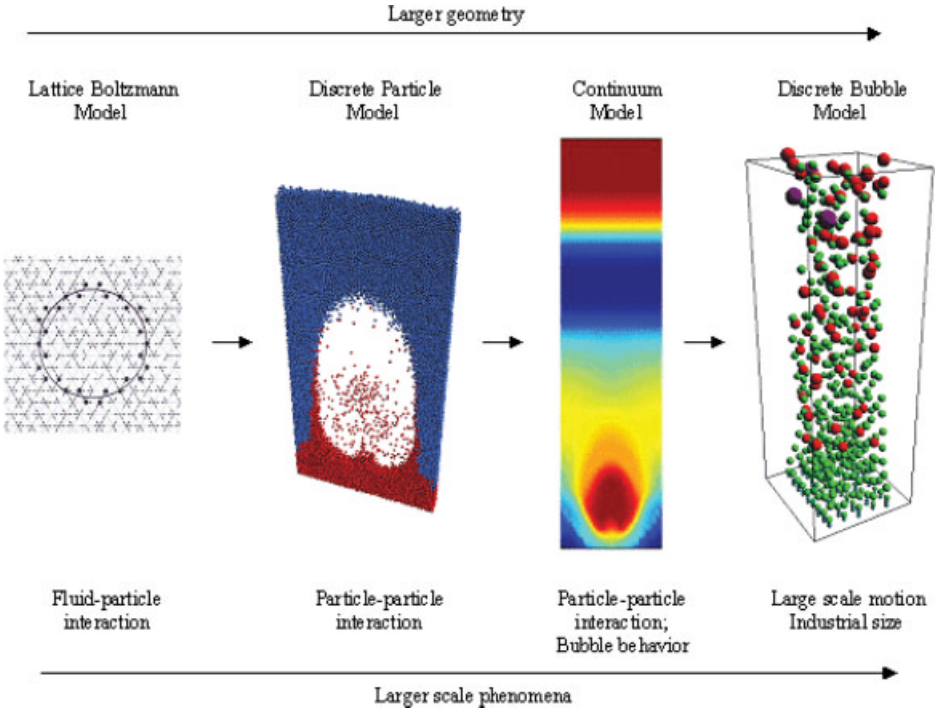

Figure 47. The multiscale approach for CFD modeling of fluidized-bed reactors [254].

izontal and vertical temperature profiles have been observed inside the combustion chambers [246].

Despite the many uncertainties, circulating fluidized bed reactors have been modeled successfully. For example, three-dimensional gas and solids concentration profiles were calculated in circulating fluidized-bed boilers with local injection of reactants [247] and coal feeding via discrete feeding points [248].

9.3. New Developments in Modeling Fluidized-Bed Reactors

9.3.1. Computational Fluid Dynamics

The models described above follow the "classical" chemical engineering approach which replaces the complex particle–fluid interaction in the fluidized bed by idealized configurations (plug flow, stirred tank, either valid overall or in regions) with mixing and mass-transfer coefficients describing the transport of matter. However, more recently, there has been a strong tendency to model the fluid mechanics of fluidized-bed reactors from first principles. The problem of computational fluid dynamics (CFD) modeling in this area is that the particle–particle and particle–fluid interaction must be considered on the particle scale, while the reactor performance must be described on a much larger scale, typically on the order of several meters. This leads to computational difficulties and problems with available computing capacities. At present (ca. 2006) there is no generally accepted CFD model of the fluidized-bed reactor available, but rapid progress can be seen in this area [249 – 253]. A promising approach appears to be multiscale modeling strategy [254].

The idea essentially is that fundamental models which take into account the relevant details of fluid–particle (lattice Boltzmann model) and particle–particle (discrete-particle model) interactions are used to develop closure laws to feed continuum models which can then be used to simulate the flow structures on a larger scale. Figure 47 illustrates this approach, which finally leads to the discrete-bubble model and should be applicable to the large industrial scale of the

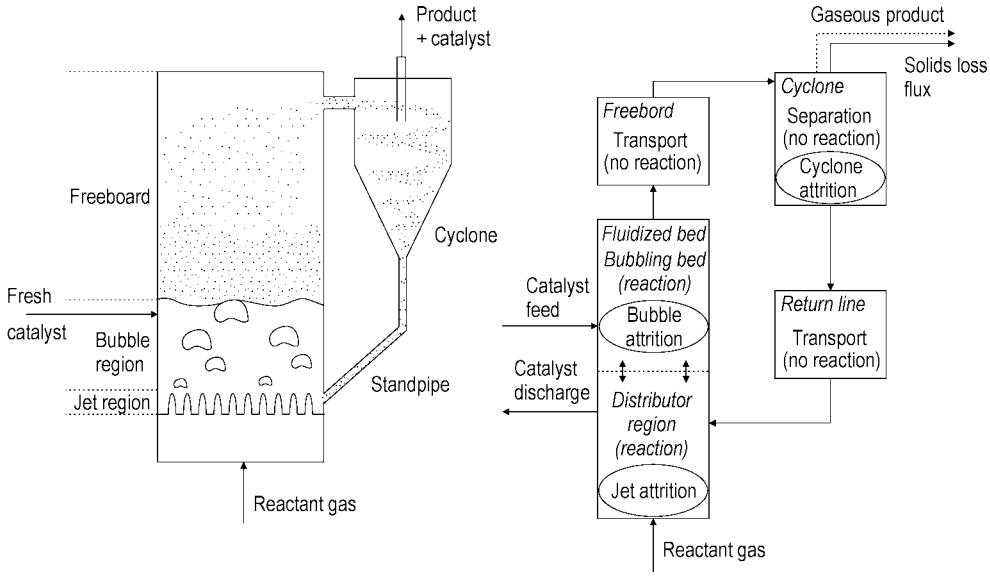

Figure 48. Fluidized-bed reactor model system [256].

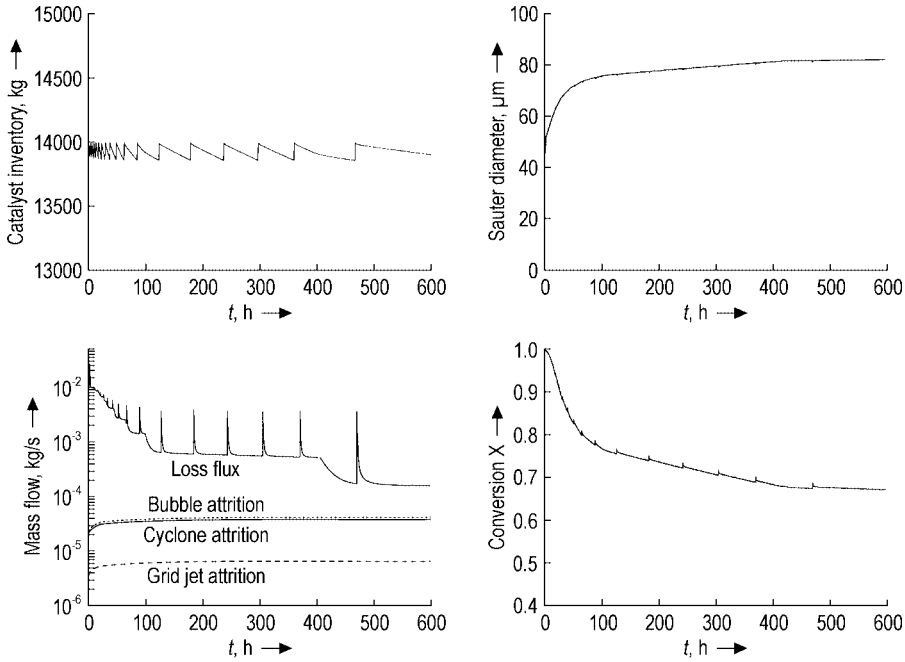

Figure 49. Reactor behavior as a function of operating time [256].

bubbling fluidized-bed reactor. The multiscale methodology [255] still requires development work but provides a good chance to arrive at more realistic fluidized-bed reactor models in the not too far future.

9.3.2. Modeling of Fluidized-Bed Systems

Another line of development is the modeling of fluidized-bed reactor systems. Whereas previously the isolated fluidized bed was modeled,

the focus now is on the coupling between the fluidized bed and the cyclone for catalyst recovery and recycle [256] or even on the coupling between two fluidized-bed reactors [257], e.g., reactor–regenerator systems as are used in the FCC and maleic anhydride processes.

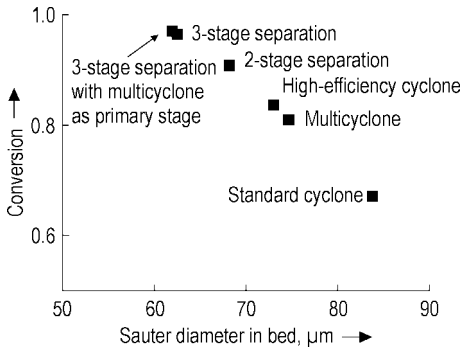

Figure 50. Influence of the Sauter diameter on the chemical conversion of a simple first-order reaction [256].

As an example, Figure 48 shows a fluidized-bed coupled with a cyclone and its translation into the model system. Attrition leads to a loss of material from the system, which requires the addition of fresh catalyst after some time (Fig. 49). A population balance model which considers the changes in the catalyst particle size intervals allows the change in the catalyst inventory with time to be followed. We see that it takes several weeks for the system to reach a quasisteady state. As a consequence of attrition and incomplete separation in the cyclone, the mean particle diameter in the bed increases with time, and this leads to larger bubbles and a reduced area of mass transfer between bubbles and the surrounding suspension in the bed. As a further consequence the conversion rate of a simple first-order reaction falls off with time. Finally, Figure 50 shows that improvements in the efficiency of the solids-recovery system are able to increase the conversion rate again, which is in agreement with large-scale industrial experience [258, 259].

10. Scale-up

Typical diameters of bench-scale fluidized-bed reactors are roughly 30–60 mm, and of pilot-scale units 450–600 mm, which should allow a reliable scale-up [9]. Full-scale fluidized-bed reactors used in the chemical industry have diameters up to ca. 10 m. Circulating fluidized bed combustors are even bigger with bed cross-sectional areas reaching 200 m^2 [261]. As equipment size increases, characteristic changes take place in the gas–solid flow that can decisively affect reactor performance. Such changes result either directly from the geometry or indirectly from design changes made as the unit is enlarged. In particular, experience has shown that the following factors affect the performance of bubbling fluidized beds during scale-up [262]:

Bed Diameter. According to Equation (22), the mean upward bubble velocity increases as the bed diameter d_t increases. As a result, the bubbles have a shorter residence time in the bed; hence the exchange area between the bubble and suspension phases is smaller, so conversion is reduced [263]. In case of circulating fluidized-bed combustors, measurements have shown that the downwards velocity of solids in the wall zone increases drastically with increasing size of the combustor [260].

Grid Design. In the laboratory, porous plates are the preferred type of gas distributor because of the ease of working with them. Gas distribution becomes worse when these are replaced by industrial distributor designs; thus the exchange area between the bubble and suspension phases is reduced, again with consequently lower conversion [43].

Internals. Whereas the laboratory fluidized bed is generally operated with no internals, plant equipment often must contain bundles of heat-exchanger tubes. Screens, baffles, or similar internals are frequently used to redisperse the bubble gas in industrial reactors. The mass-transfer area is thus increased relative to the fluidized bed without internals; the extra area can be utilized to partially offset the conversion-reducing effects of bed diameter and gas distributor [263].

Catalyst Particle-Size Distribution. Bubble growth is influenced by the proportion of fines in the particle-size distribution of the bed (usually measured as the weight fraction $<$ 0.044 mm) or by the mean grain size d_p (via

u_{mf}, Eq. 18). If the content of fines increases, bubbles collapse sooner and the equilibrium bubble size becomes smaller, with a resultant greater bubble–suspension mass-transfer area. This effect generally is fully developed only in the plant-scale reactor, where bubbles can grow without the hindrance of vessel walls. Thus, in principle, the performance of catalytic fluidized-bed reactors can be controlled by modifying the catalyst particle-size distribution [112, 264]. The recommended content by weight of fines ($<$ 0.044 mm) for "good fluidization" is 30–40 % [265], but maintaining this high a fines content in the system over a long span of time requires a very efficient solids recovery system.

Lateral Mixing of Reactants. On a laboratory scale, reactants experience compulsory uniform distribution over the bed cross section. In plant equipment, on the other hand, reactants often arrive in the reactor via individual feed points. The resulting uneven distribution of reactants can have a marked effect on reactor performance, which has been shown for the effect of coal feeding on the emission properties of fluidized-bed furnaces [248].

Secondary Reactions in the Freeboard. In a bench-scale apparatus, the fluidized gas is rapidly cooled by the vessel wall in the freeboard space after leaving the bed, so secondary reactions in the freeboard are often negligible. Such is not the case in the plant-scale reactor. The action of wall cooling is not significant here, and the entrainment rate is high because of the higher fluidization velocities common in full-scale equipment. Both effects—lack of cooling and high solids concentration in the freeboard—may lead to marked secondary reactions in the freeboard of industrial fluidized-bed reactors. In the case of a system of consecutive reactions where the desired product is formed as an intermediate, the freeboard reactions will generally lower the selectivity. The effect of freeboard reactions has been demonstrated for the example of NO and CO emissions from a fluidized-bed furnace [232].

Catalyst Attrition. Catalyst attrition is minimal in laboratory apparatus, because of the use of porous plates as gas distributors, as well as the low gas velocities and bed depths. Attrition is necessarily greater in industrial reactors. To reduce this risk in scale-up, the attrition tests described in Section 2.11 should be carried out and the results converted to the full-scale conditions with the aid of Equations (34), (35) and (36).

Other Factors. In addition to the factors just listed, many other effects become apparent when a fluidized-bed reactor is scaled up that are difficult to calculate. Examples are the risk of nonuniform gas distribution over very large cross sections in shallow fluidized beds; the formation of deposits in the bed; the fouling of heat-exchange surfaces; and catalyst aging and poisoning. On the whole, accordingly, the scale-up of fluidized-bed reactors is a complex process, commonly requiring a large amount of pilot-scale experimentation. Current knowledge about the fluid mechanics in the fluidized bed, however, enables simulation calculations of many of the scale-up effects, so the amount of testing during process development may be decreased and the risk can be at least limited.

11. References

1. I. G. Farben, DE 437 970, 1922 (F. Winkler).
2. D. Kunii, O. Levenspiel: *Fluidization Engineering*, Butterworth–Heinemann, Boston 1991.
3. D. Geldart (ed.): *Gas Fluidization Technology*, John Wiley & Sons, Chichester 1986.
4. J. F. Davidson, R. Clift, D. Harrison: *Fluidization*, Academic Press, London 1985.
5. W.-C. Yang (ed.): *Fluidization*, Solids Handling and Processing, Noyes Publications, Westwood 1999.
6. D. Gidaspow: *Multiphase Flow and Fluidization*, Academic Press, London 1994.
7. J. R. Grace, A. A. Avidan, T. M. Knowlton (eds.): *Circulating Fluidized Beds*, Chapman & Hall, London 1997.
8. W. -C Yang (eds.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003.
9. M. Pell: *Gas Fluidization*, Elsevier, Amsterdam 1990.
10. J. G. Yates: *Fundamentals of Fluidized-Bed Chemical Processes*, Butterworths, London 1983.

11. K.-E. Wirth: *Zirkulierende Wirbelschichten* – Strömungsmechanische Grundlagen, Anwendungen in der Feuerungstechnik, Springer Verlag, Heidelberg 1990.
12. P. Basu: *Combustion and Gasification in Fluidized Bed*, Taylor & Francis, Boca Raton 2006.
13. K. B. Mathur, N. Epstein: *Spouted Beds*, Academic Press, New York 1974.
14. K.-E. Wirth in E.-U. Schlünder (ed.): *VDI Wärmeatlas*, 8th ed., VDI Verlag, Düsseldorf 1997, pp. Lf 1–9.
15. K.-E. Wirth in E.-U. Schlünder (ed.): *VDI Heat Atlas*, 1st English ed., VDI Verlag, Düsseldorf 1993, pp. Lf 1–9.
16. H. Martin in E.-U. Schlünder (ed.): *VDI Wärmeatlas*, 8th ed., VDI Verlag, Düsseldorf 1997, pp. Mf 1–9.
17. H. Martin in E.-U. Schlünder (ed.): *VDI Heat Atlas*, 1st English ed., VDI Verlag, Düsseldorf 1993, pp. Mf 1–9.
18. *Ullmann*, 4th ed. **3**, 433–460.
19. *Ullmann*, 4th ed. **3**, 480–493.
20. U. Arena, R. Chirone, M. Micchio, P. Salatino (eds.): *Fluidization XI*, Engineering Conferences International, Brooklyn 2004.
21. M. Kwauk, J. Li, W.-C. Yang (eds.): *Fluidization X*, Engineering Foundation, New York 2001.
22. L.-S. Fan, T. M. Knowlton (eds.): *Fluidization IX*, Engineering Foundation, New York 1998.
23. K. Cen (eds.): *Circulating Fluidized Bed Technology VIII*, International Academic Publishers, Beijing 2005.
24. J. R Grace, J.-X. Zhu H. de Lasa (eds.): *Circulating Fluidized Bed Technology VII*, Can. Soc. Chem. Engng., Ottawa 2002.
25. J. Werther (ed.): *Circulating Fluidized Bed Technology VI*, DECHEMA, Frankfurt 1999.
26. L. Jia (ed.): *Proc. 18th Int. Conf. Fluidized Bed Combustion*, ASME, New York 2005. CD-ROM, ISBN 0-7918-3755-6.
27. S. Pisupati (ed.): *Proc. 17th Int. Conf. Fluidized Bed Combustion*, ASME, New York 2003. CD-ROM, ISBN 0-7918-3755-6.
28. F. Preto (ed.): *Proc. 16th Int. Conf. Fluidized Bed Combustion.*, ASME, New York 2001, CD-ROM, ISBN 07918-3523-5.
29. S. Ergun, *Chem. Eng. Prog.* **48** (1952) 89–97.
30. J. Werther, *Chem.-Ing.-Tech.* **54** (1982) no. 10, 876–883.
31. C. Y. Wen, Y. H. Yu, *AIChE J.* **12** (1966) 610–612.
32. C. Vogt, R. Schreiber, G. Brunner, J. Werther, *Powder Technol.* **158** (2005) 102–114.
33. J. F. Richardson, W. N. Zaki, *Trans. Inst. Chem. Eng.* **32** (1954) 35–54.
34. D. Geldart, *Powder Technol.* **7** (1973) 285–292.
35. O. Molerus, *Powder Technol.* **33** (1982) 81–87.
36. L. Reh: "Das Wirbeln von körnigem Gut in schlanken Diffusoren als Grenzzustand zwischen Wirbelschicht und pneumatischer Förderung", Dissertation, TH Karlsruhe 1961.
37. L. Reh, *Chem.-Ing.-Tech.* **46** (1974) 180–189.
38. L. Reh, *Chem.-Ing.-Tech.* **49** (1977) 786–794.
39. VGB-Merkblatt M218H, Gasverteilerböden von Wirbelschichtsystemen, VGB-Kraftwerkstechnik GmbH, Essen (1994)
40. J. C. Agarwal, W. L. Davis, D. T. King, *Chem. Eng. Prog.* **58** (1962) 85–91.
41. F. J. Zuiderweg, *Proc. Int. Symp. Fluid.* 1967, 739–750.
42. F. Zenz, D. Othmer: *Fluidization and Fluid/Particle Systems*, Reinhold Publ. Corp., New York 1960.
43. E.-U. Hartge, J. Werther in L.-S. Fan, T. M. Knowlton (eds.): *Fluidization IX*, Engineering Foundation, New York 1998, pp. 213–220.
44. S. B. Reddy Karri, J. Werther in W.-C. Yang (ed.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp. 155–170.
45. J. M. D. Merry, *AIChE J.* **21** (1975) 507–510.
46. J. M. D. Merry, *Trans. Inst. Chem. Eng.* **49** (1971) 189–195.
47. J. F. Davidson, D. Harrison: *Fluidised Particles*, Cambridge University Press, Cambridge 1963.
48. R. Clift, J. R. Grace, *Chem. Eng. Prog. Symp. Ser.* **66** (1970) no. 105, 14–27.
49. R. Clift, J. R. Grace, *AIChE Symp. Ser.* **67** (1971) no. 116, 23–33.
50. K. Hilligardt, J. Werther, *Chem. Eng. Technol.* **10** (1987) 272–280.
51. R. Toei, R. Matsuno, M. Oichi, K. Yamamoto, *J. Chem. Eng. Jpn.* **7** (1974) no. 6, 447–450.
52. J. Werther, D. Bellgardt, H. Groenewald, K. Hilligardt, *Proc. Int. Conf. Fluid. Bed. Combust.* **9th** (1987) 515–522.
53. K. Hilligardt, J. Werther, *Chem.-Ing.-Tech.* **59** (1987) no. 7, 596–597, MS 1613/87.
54. J. Werther in D. Kunii, R. Toei (eds.): *Fluidization*, Engineering Foundation, New York 1983, pp. 93–102.
55. F. A. Zenz, *Hydrocarbon Process.* **53** (1974) no. 4, 119–124.
56. S. E. George, J. R. Grace, *AIChE Symp. Ser.* **74** (1978) no. 176, pp. 67–74.

57. C. Y. Wen, L. H. Chen, *AIChE J.* **28** (1982) 117–128.
58. D. Kunii, O. Levenspiel, *Powder Technol.* **61** (1990) 193–206.
59. K. Smolders, J. Baeyens, *Powder Technol.* **92** (1997) 35–46.
60. J. Werther, E.-U. Hartge, in W. C. Yang (ed.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp. 113–128.
61. J. Yerushalmi, N. T. Cankurt, D. Geldart, B. Liss, *AIChE Symp. Ser.* **74** (1978) no. 176, 1–12.
62. Y. Li, M. Kwauk in J. R. Grace, J. Matsen (eds.): *Fluidization*, Plenum Press, New York 1980, pp. 537–544.
63. J. M. Matsen, *Powder Technol.* **32** (1982) 21–33.
64. K.-E. Wirth, *Chem. Eng. Technol.* **11** (1988) no. 1, 11–17.
65. K.-E. Wirth, *Chem. Eng. Technol.* **14** (1991) 29–38.
66. M. Rhodes, D. Geldart in P. Basu (ed.): *Circulating Fluidized Bed Technology*, Pergamon Press, Toronto 1986, pp. 193–200.
67. P. Schlichthaerle, J. Werther, *Chem. Eng. Sci.* **54** (1999) 5485–5494.
68. T. Grassler, K.-E. Wirth in J. Werther (ed.): *Circulating Fluidized Bed Technology VI*, DECHEMA, Frankfurt 1999, pp. 65–70.
69. M. J. Rhodes, M. Sollaart, X. S. Wang in L.-S. Fan, T. M. Knowlton (eds.): *Fluidization IX*, Engineering Foundation, New York 1998, pp. 141–148.
70. C. Rudnick, J. Werther in L.-S. Fan, T. M. Knowlton (eds.): *Fluidization IX*, Engineering Foundation, New York 1998, pp. 573–580.
71. K.-E. Wirth, M. Seiter in R. J. Anthony (ed.): *Fluidized Bed Combustion*, ASME, New York 1991, pp. 311–315.
72. J. R. Grace, A. S. Issangya, D. Bai, H. Bi, J. Zhu, *AIChE J.* **45** (1999) 2108–2116.
73. J. Li, M. Kwauk : *Particle-Fluid Two-Phase Flow–The Energy-Minimization Multi-Scale Method.*, Metallurgical Industry Press, Beijing 1994.
74. W. Ge, J. Li, *Chem. Eng. Sci.* **57** (2002) 3993–4004.
75. J. Li., W. Ge, J. Zhang, M. Kwauk, *Trans. ChemE, Part A Chem. Eng. Res. Des.* **83** (2005) 574–582.
76. J. Li., C. Cheng, Z. Zhang, J. Yuan, A. Nemet, F.N. Fett, *Chem. Eng. Sci.* **54** (1999) 574–582.
77. J. Li, J. Ouyang, S. Gao, W. Ge, N. Yang, W. Song: *Multi-Scale Simulation of Particle-Fluid Complex Systems*, Science Press, Beijing 2005,
78. N. Yang, W. Wang, W. Ge, J. Li, *Chem. Eng. J.* **96** (2003) 71–80.
79. T. M. Knowlton in P. Basu, J. F. Large (eds.): *Circulating Fluidized Bed Technology II*, Pergamon Press, Oxford 1988, pp. 31–41.
80. Y. Jin, Y. Zheng, F. Wei in J. R. Grace, J. Zhu, H. de Lasa (eds.): *Circulating Fluidized Bed Technology VII, Can. Soc. Chem. Eng.*, Ottawa 2002, pp. 40–60.
81. P. Lehner, K.-E. Wirth, *Chem. Eng. Sci.* **54** (1999) 5471–5483.
82. M. Zhang, Z. Quian, H. Yu, F. Wei, *Chem. Eng. J.* **92** (2003) 161–167.
83. F. Wei, J. Z. Liu, Y. Jin, Z. Q. Yu, *Chem. Eng. Tech.* **18** (1995) 59–62.
84. H. E. Liu, F. Wei, Y. H. Yang, Y. Jin in J. R. Grace, J. Zhu, H. de Lasa (eds.): *Circulating Fluidized Bed Technology VII, Can. Soc. Chem. Eng.*, Ottawa 2002, pp. 83–89.
85. C. Huang, Z. Quian, M. Zhang, F. Wei, *Powder Technol.* **161** (2006) 48–52.
86. M. Zhang, K. Chu, A. Yu, F. Wei in K. Cen (ed.): *Circulating Fluidized Bed Technology VIII, Int. Academic Publ.*, Beijing 2005, pp. 299–305.
87. H. Brust, K.-E. Wirth, *Ind. Eng. Chem. Res.* **43** (2004) pp. 5796–5801.
88. R. Deng, F. Wei, Y. Jin, Q. Zhang, Y. Jin, *Chem. Eng. Tech.* **25** (2002) 711–716.
89. H. Brachold, C. Peuckert, H. Regner, *Chem. Ing. Tech.* **65** (1993) 293–297.
90. A. S. Bassi, C. L. Briens, M. A. Bergougnou in A. A. Avidan (ed.): *Circulating Fluidized Bed Technology IV*, AIChE, New York 1994, pp. 25–32.
91. W. L. Forsythe, W. R. Hertwig, *Ind. Eng. Chem.* (1949) 1200–1206.
92. J. E. Gwyn, *AIChE J.* **15** (1969) no. 1, 35–39.
93. F. A. Zenz, E. G. Kelleher, *J. Powder Bulk Solids Technol.* **4** (1980) nos. 2/3, 13–20.
94. W. G. Vaux, J. S. Schruben, *AIChE Symp. Ser.* **79** (1983) no. 222, 222–228.
95. W. G. Vaux, J. S. Schruben, *Chem. Eng. Commun.* **33** (1985) 333–347.
96. W. Xi, J. Werther, "Jet Attrition of Catalyst Particles in Gas Fluidized Beds", *Int. Symp. on Attrition and Wear in Powder Technology,* Utrecht 1992.
97. J. Werther, W. Xi, *Powder Technol.* **33** (1993) 39–46.
98. J. Werther, J. Reppenhagen, in W.-C. Yang (eds.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp. 201–237.

99. J. Reppenhagen, J. Werther, *AIChE J.* **45** (1999) 2001–2010.
100. J. Reppenhagen, J. Werther, *Powder Technol.* **113** (2000) 55–69.
101. J. Reppenhagen, A. Schetzschen, J. Werther, *Powder Technol.* **112** (2000) 251–255.
102. C. Klett, E.-U. Hartge, J. Werther, *Proc. Combustion Inst.* **30** (2005) 2947–2954.
103. E.-U. Hartge, C. Klett, J. Werther, AIChE J. (2007), in press.
104. P. N. Rowe in J. F. Davidson, E. Harrison (eds.): *Fluidization*, Academic Press, London 1971, pp. 121 – 192.
105. J. Werther, *Chem.-Ing.-Tech.* **49** (1977) no. 3, 193–202.
106. J. M. D. Merry, J. F. Davidson, *Trans. Inst. Chem. Eng.* **51** (1973) 361–368.
107. M. Virr, *Inst. Energy Symp. Ser. (London)* **4** (1980) IA–1–1-11.
108. D. Ormston, E. Robinson, D. Buckle, *Inst. Energy Symp. Ser. (London)* **4** (1980) IIA–1–1-10.
109. J. J. van Deemter in J. F. Davidson, R. Clift, D. Harrison (eds.): *Fluidization*, Academic Press, London 1985, pp. 331–355.
110. J. Werther, H. Groenewald, A.-K. Schroeder-Lanz, *Energietechnik (Leipzig)* **40** (1990) no. 5, 164–167.
111. J. H. De Groot, *Proc. Int. Symp. Fluid.* 1967, 348–358.
112. R. J. de Vries, W. P. M. van Swaaij, C. Mantovani, A. Heijkoop, *Chem. React. Eng. Proc. Eur. Symp.* **5th** (1972) session B 9, pp. 56–69.
113. D. Bellgardt, F. W. Hembach, M. Schößler, J. Werther, *Proc. Int. Conf. Fluid. Bed Combust.* **9th** (1987) 713–722.
114. D. Kunii, O. Levenspiel, *J. Chem. Eng. Jpn.* **2** (1969) 122–129.
115. D. Bellgardt, J. Werther, *Powder Technol.* **48** (1986) 173–180.
116. B. Hirschberg, J. Werther in J. R. Grace, A. A. Avidan, T. M. Knowlton (eds.): *Circulating Fluidized Beds*, Chapman & Hall, London 1997, pp. 119–148
117. R. Koenigsdorff, J. Werther, *Powder Technol.* **84** (1995) 317–329.
118. P. Schlichthärle, J. Werther, *Powder Technol.* **120** (2001) 21–33.
119. B. Hirschberg, J. Werther, *AIChE J.* **44** (1998) 25–34.
120. U. Arena in J. R. Grace, A. A. Avidan, T. M. Knowlton (eds.): *Circulating Fluidized Beds*, Chapman & Hall, London 1997, pp. 86–118.
121. C. M. H. Brereton, J. R. Grace, J. Yu in P. Basu, J. F. Large (eds.): *Circulating Fluidized Bed Technology II*, Pergamon Press, Oxford 1988, pp. 307–314.
122. J. Werther, E.-U. Hartge, M. Kruse, W. Nowak in P. Basu, M. Horio, M. Hasatani (eds.): *Circulating Fluidized Bed Technology III*, Pergamon Press, Oxford 1990, pp. 593–598.
123. J. Werther, E.-U. Hartge, M. Kruse, *Powder Technol.* **70** (1992) 293–301.
124. A. Klinkenberg, H. J. Krajenbrink, H. A. Lauwerier, *Ind. Eng. Chem.* **45** (1953) no. 6, 1202–1208.
125. T. K. Sherwood, R. L. Pigford, C. R. Wilke: *Mass Transfer*, McGraw Hill, New York 1975.
126. O. Molerus, K.-E. Wirth, *Heat Transfer in Fluidized Beds*, Chapman & Hall, London 1997.
127. R. Turton, T. J. Fitzgerald, O. Levenspiel, *Int. J. Heat Mass Transfer* **32** (1989) 289–296.
128. J. S. M. Botterill, Y. Teoman, K. R. Yuregir, *Powder Technol.* **39** (1984) 177–189.
129. R. Wunder, A. Mersmann, *Chem.-Ing.-Tech.* **51** (1979) no. 3, 241.
130. L. R. Glicksman in J. R. Grace, A. A. Avidan, T. M. Knowlton (eds.): *Circulating Fluidized Beds*, Chapman & Hall, London 1997, pp. 261–311.
131. J. Schweinzer, O. Molerus, *Chem. Eng. Technol.* **12** (1989) 38–44.
132. A. P. Baskakov, B. Leckner, *Powder Technol.* **90** (1997) 213–218.
133. T. M. Knowlton, in W.-C. Yang, (ed.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp. 599–617.
134. S. Pulupula, H. Singh, S. J. Chopra, in L. S. Fan, T. M. Knowlton (eds.): *Fluidization X*, Engineering Foundation, New York 1998, pp. 733–739.
135. S. C. Arnold, M. Cecchini, G. Fenati, J. W. Stanecki, in M. Kwauk, J. Li, W. C. Yang (eds.): *Fluidization X*, Engineering Foundation, New York 2001, pp. 707–714.
136. P.W. Smit, S. Baloyi, T. Shingles, in U. Arena, R. Chirone, M. Micchio, P. Salatino (eds.): *Fluidization XI*, ECI Engineering Conferences International, Brooklyn 2004, pp. 587–593.
137. Y. Jiang, K. M. Mc Auley, J. C. C. Hsu, *Ind. Eng. Chem. Res.* **37** (1997) 1176–1180.
138. S. Ariyapadi, D. W. Holdsworth, C. J. D. Norley, F. Berruti, C. Briens, *Int. J. Chem. Reactor Eng.* **1** (2003) A 60. http://www.bepress.com/ijcre/Vol1/A60

139. S. Ariyapadi, F. Berruti, C. Briens, P. Griffith, C. Hulet, *Can. J. Chem. Eng.* **81** (2003) 891–891.
140. S. Bruhns, J. Werther, *AIChE J.* **51** (2005) pp. 766–775
141. J. Werther, S. Bruhns *Int. J. Chem. Reactor Eng.* **2** (2004) A31. http://www.bepress.com/ijcre/Vol 2/A31
142. Y. Ge, L.S. Fan in U. Arena, R. Chirone, M. Micchio, P. Salatino (eds.): *Fluidization XI*, ECI Engineering Conferences International, Brooklyn 2004, pp. 259–266.
143. B. Riediger: *Die Verarbeitung des Erdöls*, Springer Verlag Berlin 1971.
144. D. Kunii, O. Levenspiel: *Fluidization Engineering*, John Wiley, New York 1969.
145. J. L. Callahan, R. K. Grasselli, E. C. Milberger, H. A. Strecker, *Ind. Eng. Chem. Process Des. Dev.* **6** (1970) 134–140.
146. H. S. van der Baan in R. Prins, G. C. A. Schuit (eds.): *Chemistry and Chemical Engineering of Catalytic Processes*, Nato Advanced Study Institute Series E: Applied Sciences no. 39, Academic Press, London 1980.
147. M. E. Dry: *Applied Industrial Catalysis*, Academic Press, New York 1983.
148. R. W. Silverman et al. in K. Ostergaard, A. Sorensen (eds.): *Fluidization*, Engineering Foundation, New York 1986, pp. 441–448.
149. A. P. Steynberg, R. L. Espinoza, B. Jager, A. C. Vosloo, *Appl. Catal. A* **188** (1999) 41–54.
150. S. Sookai, P.L. Langenhoven, T. Shingles,in M. Kwauk, J. Li and W.C. Yang (eds.): *Fluidization X*, Engineering Foundation, New York 2001, pp. 620–628.
151. J. J. Graham, *Chem. Eng. Prog.* **66** (1970) 54–60.
152. Mitsubishi Kasei Kogyo, DE 2 165 323 C 2, 1971 (K. Terahata et al.).
153. Mitsubishi Chemical Industries, DE 2 658 861, 1976 (S. Kamimura, T. Otaki).
154. S. C. Arnold, G. D. Suciu, L. Verde, A. Neri, *Hydrocarbon Process.* **64** (1985) no. 9, 123–126.
155. G. Stefani, F. Budi, C. Fumagalli, G. D. Suciu in G. Centi, F. Trifiro (eds.): *New Developments in Selective Oxidation*, Elsevier Science Publ., Amsterdam 1990, pp. 537–552.
156. R. M. Contractor, J. Ebner, M. J. Mummey, in C. Centi, F. Trifiro (eds.): *New Developments in Selective Oxidation*, Elsevier, Amsterdam 1990, pp. 553–562.
157. K. G. Golbig, J. Werther, *Chem. Eng. Sci.*, **52** (1997) pp. 583–595.
158. R.M. Contractor, *Chem. Eng. Sci.* **54** (1999) pp. 5627–5632.
159. T. Miyauchi et al., *Adv. Chem. Eng.* **11** (1981) 275–448.
160. R. Pearce, W. R. Patterson: *Catalysis and Chemical Processes*, Leonard Hill, Glasgow 1981.
161. T. Katsumata, T. Dozono, *AIChE Symp. Ser.* **83** (1987) no. 255, 86–96.
162. C. D. Chang: *Hydrocarbons from Methanol*, Marcel Dekker, New York 1983.
163. H. H. Gierlich, W. Dolkemeyer, A. Avidan, N. Thiagarajan, *Chem.-Ing.-Tech.* **58** (1986) no. 3, 238–239, MS 1462/86.
164. *Ullmann*, 4th ed., **18**, 521–544.
165. *Ullmann*, 4th ed., **7**, 566–576.
166. M. Schwarzmann, *Hydrocarbon Process.* **48** (1969) no. 9, 184–186.
167. B. Jazayeri in W.-C. Yang (ed.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp.421–444.
168. F. J. Karol, *Catal. Rev. Sci. Eng.* **26** (1984) 557–565.
169. T. Xie, K. B. McAuley, J. C. C. Hsu, D. W. Bacon, *Ind. Eng. Chem. Res.* **33** (1994) 449–479.
170. H. Koda, T. Kurisaka in M. Kwauk, D. Kunii, Z. Jiansheng, M. Hasatani (eds.): *Fluidization '85*, Science and Technology, Science Press, Beijing 1985, 402–413.
171. J. D. Burdett, R. S. Eisinger, P. Cai, K. H. Lee, in M. Kwauk, J. Li, W. C. Yang (eds.): *Fluidization X*, Engineering Foundation, New York 2001, pp. 39–52.
172. A. M. Salleh, A. Ingram, A. McCormack, R. Coreenwood, V. Reiling, in U. Arena, R. Chirone, M. Micchio, P. Salatino (eds.): *Fluidization XI*, Engineering Conferences, New York 2004, pp. 211–218.
173. P. Cai, I. D. Burdett in K. Cen (ed.): *Circulating Fluidized Bed Technology VIII*, Int. Academic Publ., World Publ. Corp., Beijing 2005, pp. 410–417.
174. F. T. Barr, C. E. Jahnig, *Chem. Eng. Prog.* **51** (1955) 167–177.
175. J. M. Matsen in G. Hetsroni (ed.): *Handbook of Multiphase Systems*, McGraw Hill, New York 1982, pp. 8/178–183.
176. P. Schmalfeld, *Hydrocarbon Process Pet. Refin.* **42** (1963) no. 6, 145–155.
177. A. Steinhofer, *Chem.-Ing.-Tech.* **36** (1964) no. 9, 889–898.
178. J. Zelkowski: *Kohleverbrennung*, VGB-Kraftwerkstechnik GmbH, Essen 1986.

179. G. Stamatelopoulos, J. Seeber, R.S. Skowyra in L. Jia (ed.): Proc. 18th Int. Conf. Fluidized Bed Combustion, ASME, New York 2005, CD ROM, ISBN 0-7918-3755-6.
180. Lurgi AG, Jahresbericht 1989, Frankfurt/Main 1989.
181. F. van Dijen, P. Savat, J. Vanormelingen, H. Sablon, *VGB PowerTech* (2005) no. 11, pp. 64–66.
182. G. Stamatelopulos, J. Seeber, *VGB PowerTech* 2005, no. **11**, 38–43
183. L. Carpenter et al.in R. J. Anthony (ed.): *Proc. Int. Conf. Fluid. Bed Combust*. ASME, New York, 1991, p. 467–474.
184. H. Martin, *VGB Kraftwerkstech.* **71** (1991) no. 4, 347–354.
185. S. A. Jansson, J. Anderson in R. B. Reuther (ed.): *Proc. Int. Conf. Fluid. Bed Combust.*, ASME, New York 1999, FBC99–0191.
186. G. Dehn, H. Meier, H. Möllenhoff, H. Rehwinkel, G. von Wedel, *Proc. Int. Conf. Fluid. Bed Combust*. (R. J. Anthony, ed.), ASME New York 1991, p. 345–362.
187. K. K. Pillai, S. A. Jansson, *Proc. Int. Conf. Fluid. Bed Combust*. (R. J.Anthony, ed.) ASME New York 1991, p. 493–502.
188. J. Anderson, S. A. Jansson, *Proc. Int. Conf. Fluid. Bed Combust*. (R. J.Anthony, ed.) ASME New York 1991, p. 787–793.
189. D. Adams, A. Dodd, D. Geiling, A. Robertson in R. B. Reuther (ed.): *Proc. Int. Conf. Fluid. Bed Combust.*, ASME, New York 1999, FBC99–0076.
190. M. A. Cuenca, E. J. Anthony (eds.): *Pressurized Fluidized Bed Combustion*, Chapman & Hall, London 1995.
191. J. Werther, T. Ogada, *Progress Energy Combust. Sci.* **25** (1999) 55–116.
192. J. Werther, M. Saenger, *J. Chem. Eng. Japan* **33** (2000) 1–11.
193. Uhde: Verbrennung pastöser Rückstände, company brochure, Dortmund 1988.
194. J. Werther, L. Mühlhaus, W. Liebig, *Maschinenmarkt* **94** (1988) no. 22, 50–54.
195. U. Urban, *WLB Wasser Luft Betr*. (1987) no. 9, 66–68.
196. Sulzer-Escher-Wyss: Verfahren für die umweltfreundliche und energiesparende Behandlung von Klärschlamm, company brochure, Ravensburg 1989.
197. C. Berse, H. Teufel, G. Müller, *Chem. Eng. Process.* **24** (1990) no. 5, 36–39.
198. H. v. Raczeck, J. Werther, M. Wischniewski, *VDI-Ber.* **522** (1991) 585–595.
199. H. Hannemann et al.: *Wirbelschichtsysteme 1990*, VGB-TB 212, VGB Techn. Vereinigung der Großkraftwerksbetreiber e.V., Essen 1990, V 29/1–23.
200. W. Adlhoch, J. Keller, P. K. Herbert: "Kohlevergasung 1991" *VGB Conference*, Dortmund, May 16th/17th, 1991.
201. R. Wirthwein, P. Scur, H. Hirschfelder, H. Vierrath in J. Werther (ed.): *Circulating Fluidized Bed Technology VI*, DECHEMA, Frankfurt 1999, pp. 737–746.
202. L. Reh, *Chem. Eng. Prog.* **67** (1971) no. 2, 58–68.
203. L. Reh, *Chem.-Ing.-Tech.* **42** (1970) no. 7, 447–451.
204. L. Reh, J. Ernst, H. W. Schmidt, K. H. Rosenthal, *Aluminium (Düsseldorf)* **47** (1971) no. 11, 681–685.
205. W. Bresser, P. Weber, *Iron Steel Eng.* (1995) 81–85.
206. R. Husain, P. Weber, A. Orth, H. Eichberger in J. Werther (ed.): *Circulating Fluidized Bed Technology VI*, DECHEMA, Frankfurt 1999, pp. 593–598.
207. L.-S. Fan, *Gas-Liquid-Solid Fluidization Engineering*, Butterworth, Boston 1989.
208. K. Schügerl, *Can. J. Chem. Eng.* **67** (1989) 178–184.
209. J.-X. Zhu, Y. Zheng, D. G. Karamanev, A. S. Bassi, *Cand. J. Chem. Eng.* **78** (2000) 82–94.
210. J. J. Heijnen, *Chem.-Ing.-Tech.* **56** (1984) 526–532.
211. J. J. Heijnen, A. Mulder, W. Enger, F. Hoeks: *Anaerobic Treatment, a Grown-up Technology*, NVA Water Treatment Conference Aquatech '86, Amsterdam Sept. 15–19, 1986.
212. O. Moebus, M. Teuber, *Kiel. Milchwirtsch. Forschungsber.* **33** (1981) 3–23.
213. L. Röttenbacher: "Entwicklung und Modellierung eines Gas/Feststoff Wirbelschichtfermenters für die Erzeugung von Ethanol mit S. Cerevisiae", Dissertation, Techn. Universität Hamburg-Harburg 1985.
214. W. Bauer in K. Ostergaard, A. Sorensen (eds.): *Fluidization V*, Engineering Foundation, New York 1986, pp. 619–626.
215. W. Bauer, *Can. J. Chem. Eng.* **64** (1986) 561–566.
216. T. Akao, Y. Okamoto in D. Kunii, R. Toei (eds.): *Fluidization*, Engineering Foundation, New York 1983, pp. 631–637.
217. T. Akao, Y. Okamoto, *Kagaku Kogaku* **49** (1985) 349–355.
218. M. Asif, *Chem. Eng. Technol.* **21** (1998) 77–82.

219. A. Tanyolak, H. Beyenal, *Chem. Eng. J. & Biochem. Eng. J.* **2** (1998) 207–216.
220. J. R. Grace, K. S. Lim in J. R. Grace, A. A. Avidan, T. M. Knowlton (eds.): *Circulating Fluidized Beds*, Chapman & Hall, London 1997, pp. 504–524.
221. P. Jiang, F. Wei, L.S. Fan in W. C. Yang (ed.): *Handbook of Fluidization and Fluid-Particle Systems*, Marcel Dekker, New York 2003, pp. 309–342.
222. R. Shinnar, *ACS Symp. Ser.* **72** (1978) 1–13.
223. W. Böck, G. Emig, W. Sit, J. Werther, *Chem. React. Eng. Proc. Int. Symp.* **8th** (1984) 479–486.
224. W. Sitzmann, M. Schößler, J. Werther, *Chem.-Ing.-Tech.* **59** (1987) no. 1, 68/69 and MS 1553/87.
225. W. P. M. van Swaaij in J. F. Davidson, R. Clift, D. Harrison (eds.): *Fluidization*, Academic Press, London 1985, pp. 595–629.
226. J. R. Grace in D. Geldart (ed.): *Gas Fluidization Technology*, John Wiley & Sons, Chichester 1986, pp. 285–340.
227. D. Kunii, O. Levenspiel: *Fluidization Engineering*, Butterworth-Heinemann, Boston 1991, pp. 303–305.
228. W. P. M. van Swaaij, F.J. Zuiderweg, Proc. 5th Eur. Symp. Reaction Eng., Amsterdam 1972, pp. B9–25.
229. D. Rokita: "Dynamische Wirbelschichtmodellierung einer komplexen Reaktion", Dissertation, Technische Universität Hamburg-Harburg 1991.
230. S. P. Sit, J. R. Grace, *Chem. Eng. Sci.* **36** (1981) 327–335.
231. L. H. Chen, C. Y. Wen, *AIChE J.* **28** (1982) no. 6, 1019–1029.
232. M. Schößler, H. v. Raczeck, J. Werther, *Chem.-Ing.-Tech.* **61** (1989) no. 8, 652–653 and MS 1771/89.
233. T. Minoura, Y. Sakamoto, S. Toyama in A. M. Manaker (ed.): *Proc. Int. Conf. Fluid. Bed Combust.*, ASME, New York 1989, pp. 1155–1161.
234. D. Bellgardt, F. Hembach, M. Schössler, J. Werther in J. P. Mustonen (ed.): *Proc. 9th Int. Conf. Fluid. Bed Combust.*, ASME, New York 1987, pp. 713–722.
235. P. Raven, G. A. Sparham, *Proc. Int. Conf. Fluid. Bed Combust.* **7th** (1982) 275–283.
236. S. Artlich, E.-U. Hartge, J. Werther, *Ind. Eng. Chem. Res.* **37** (1998) 782–792.
237. W. P. M. vanSwaaij in D. Luss, V. W. Weekman (eds.): *Chemical Reaction Engineering Review*, ACS, Washington 1978, pp. 329–338.
238. M. Edwards, A. A. Avidan, *Chem. Eng. Sci.* **41** (1986) pp. 829–836.
239. P. I. Viitanen, *Ind. Eng. Chem. Res.* **32** (1993) pp. 577–583.
240. S. Ouyang, J. Lin, O. E. Potter, *Powder Technol.* **74** (1993) pp. 73–79.
241. H. Schoenfelder, J. Werther, J. Hinderer, J. F. Keil, *AIChE Symposium Series* **90** (1994) 301, 92–104.
242. H. Kagawa, H. Mineo, R. Yamazaki, K. Yoshida in P. Basu, M. Horio, M. Hasatani (eds.): *Circulating Fluidized-Bed Technology III*, Pergamon Press, Oxford 1991, pp. 551–559.
243. T. S. Pugsley, G. S. Patience, F. Berruti, J. Chaouki, *Ind. Eng. Chem. Res.* **31** (1992) pp. 2652–2660.
244. S. Ouyang, O.E. Potter in A. Avidan (ed.): *Circulating Fluidized-Bed Technology IV*, AIChE, New York 1993, pp. 515–521.
245. G.S. Patience, P.L. Mills, AIChE Annual Meeting, St. Louis 1993.
246. J. Werther in K. Cen (ed.): *Circulating Fluidized-Bed Technology VIII*, International Academic Publishers, Beijing 2005, pp. 1–25.
247. T. Knoebig, J. Werther, *Chem. Eng. Technol.* **21** (1999) 656–659.
248. E.-U. Hartge, K. Luecke, J. Werther, *Chem. Eng. Sci.* **54** (1999) 5393–5407.
249. J.A.M. Kuipers, B.P.B. Hoomans, W.P.M. vanSwaaij in L.S. Fan, T.M. Knowlton (eds.): *Fluidization IX*, Engineering Foundation, New York 1998, pp. 15–30.
250. B.A. Kashiwa, W.C. Yang in J.R. Grace, J. Zhu, H. de Lasa (eds.): *Circulating Fluidized Bed Technology VII*, Can. Soc. Chem. Eng., Ottawa 2002, pp. 27–39.
251. N. Yang, W. Ge, G. Niu, C. Yang, J. Li in K. Cen (ed.): *Circulating Fluidized-Bed Technology VIII*, International Academic Publishers, Beijing 2005, pp. 291–298.
252. M. Zhang, K. Chu, A. Yu, F. Wei in K. Cen (ed.): *Circulating Fluidized-Bed Technology VIII*, International Academic Publishers, Beijing 2005, pp. 299–305.
253. X. Xiao, W. Wang, H. Yang, J. Zhang, G. Yue in K. Cen (ed.): *Circulating Fluidized-Bed Technology VIII*, International Academic Publishers, Beijing 2005, pp. 394–401.
254. M. A. van der Hoef, M. vanSint Annaland, J. A. M. Kuipers, *Chem. Eng. Sci.* **59** (2004) pp. 5157–5165.

255. J. Li, W. Ge, J. Zhang, M. Kwauk, *Trans. I ChemE, Part A Chem. Eng. Des.*, **83** (2005) pp. 574–582.
256. J. Werther, E. -U. Hartge, *Ind. Eng. Chem. Res.* **43** (2004) pp. 5593–5604.
257. J. Werther, E. -U. Hartge, *Powder Technol.* **148** (2004) pp. 113–122.
258. M. Pell, S.P. Jordan, *AIChE Symp. Ser.* **84** (1988) no. 262, pp. 68–73.
259. R. J. de Vries, W. P. M. van Swaaij, C. Mantovani, A. Heijkoop, Proc. 5th Eur. Symp. Chem. Reaction Eng., Session B9, Amsterdam 1972, pp. 56–69.
260. E.-U. Hartge, J. Werther, V. Wiesendorf in J.R. Grace, J. X. Zhu,H. de Lasa (eds.): *Circulating Fluidized Bed Technology VII*, Can. Soc. Chem. Eng., Ottawa 2002, pp. 325–332.
261. W. Nowak, Z. Bis, J. Laskawiec, W. Krzywoszynski, R. Walkowiak in J. Werther (ed.): *Circulating Fluidized Bed Technology VI*, DECHEMA, Frankfurt 1999, pp. 1003–1008.
262. J. Werther, *Chem. Eng. Sci.* **47** (1992) 9–11, 2457–2462.
263. W. Bauer, J. Werther, *Chem. React. Eng. Proc. Int. Symp.* **7th** (1982); *ACS. Symp. Ser.* **196** (1982) 121–132.
264. M. Pell, S. P. Jordan, paper presented at AIChE Meeting, New York, November 1987.
265. Y. Ikeda, in M. Kwauk, D. Kunii, J. Zheng, M. Hasatani (eds.): *Fluidization '85–Science and Technology*, Elsevier, Amsterdam 1985, pp. 1–10.

Bubble Columns

For other industrial reactors and their applications, see → Stirred-Tanc and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

PETER ZEHNER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

MATTHIAS KRAUME, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 2
2. Bubble Columns and Modifications 4
2.1. Design and Applications 4
2.2. Gas Distribution 5
2.3. Flow Regimes 6
2.4. Fluid Dynamics 7
2.5. Bubble Size 8
2.6. Bubble Rise Velocity 9
2.7. Dispersion of the Liquid Phase 9
2.8. Dispersion of the Gas Phase 10
2.9. Gas Holdup 10
2.10. Specific Interfacial Area 12
2.11. Volumetric Mass-Transfer Coefficient 13
2.12. Heat Transfer 14
2.13. Slurry Bubble Columns 14
2.14. Airlift Loop Reactors 16
3. Downflow Bubble Columns 18
3.1. Design and Applications 19
3.2. Operating Conditions and Gas Holdup 20
3.3. Mass Transfer 21
4. Jet Loop Reactors 22
4.1. Design and Applications 23
4.2. Typical Dimensions 26
4.3. Energy Balance 27
4.4. Mixing Behavior and Fluid Dynamics 27
4.5. Gas Holdup 28
4.6. Mass Transfer 30
4.7. Three-Phase Loop Reactor 31
5. References 31

Symbols (see also → Principles of Chemical Reaction Engineering and → Model Reactors and Their Design Equations)

Variables

a specific interfacial area, m^{-1}
A interfacial area, m^2
d diameter, m
d_h diameter of holes, m
d_i inner diameter of draft tube, m
d_n nozzle diameter, m
D diffusion or dispersion coefficient, m^2/s
$D_{G,L}$ diffusion coefficient of dissolved gas in liquid, m^2/s
e_M energy dissipation rate per unit mass, W/kg
e_n jet power per unit volume, W/m^3
e_V energy dissipation rate per unit volume, W/m^3
f fraction of cross-sectional area
f_i fraction of cross-sectional area of draft tube
F cross-sectional area, m^2
F_i cross-sectional area of draft tube, m^2
F_R cross-sectional area of reactor, m^2
h height, m
h_R height of gas – liquid mixture, m
h_t height of reactor, m
J_D dispersion flow
k_L liquid-phase mass-transfer coefficient, m/s
P power, W
r radial distance from column axis, m
t time, s
u superficial velocity, m/s
v velocity, m/s
v_{rG} relative velocity of bubble swarm in liquid, m/s
v_{rS} relative velocity of particle swarm in liquid, m/s
V volume, m^3
$\dot{V}$ volumetric flow rate, m^3/s
z axial coordinate, m

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_275

Greek symbols

α	heat-transfer coefficient, $W\,K^{-1}\,m^{-2}$
ε	volume fraction
ε_G	gas holdup
ζ	drag coefficient of circulation flow
η	dynamic viscosity, $kg\,m^{-1}\,s^{-1}$
ν	kinematic viscosity, m^2/s
ϱ	density, kg/m^3
$\Delta\varrho$	density difference between liquid and gas, kg/m^3
$\Delta\varrho_S$	density difference between liquid and solids, kg/m^3
σ	surface tension, N/m
φ	mass concentration, kg/m^3

Subscripts

a	annular space
b	bubble
bS	Sauter diameter
c, circ	circulation
D	downflow
G	gas phase
h	hole
i	inside draft tube
L	liquid
max	maximum value
min	minimum value
M	per unit mass
n	nozzle
p	particle
r	relative
R	upflow, reaction mixture
slip	slip
S	solids
t	reactor
V	per unit volume

1. Introduction

Bubble columns are devices in which gas, in the form of bubbles, comes in contact with liquid. The purpose may be simply to mix the liquid phase. Far more often, however, substances are transferred from one phase to the other, for example, when gaseous reactants are dissolved in a liquid or when liquid reaction products are stripped. Both processes can take place simultaneously. A chemical or biological reaction nearly always proceeds in the liquid phase. Depending on the application, special measures to intensify mass transfer between the two phases may be useful, or the residence-time distribution of one or both phases may be modified.

The liquid may also contain inert, catalytically active, or reactive particles in suspension. Oxidation, hydrogenation, chlorination, phosgenation, alkylation, and other processes have long been performed in bubble-column reactors in the chemical industry. In 1978, more than 10^7 t/a of chemical products were made in bubble columns [1]. Since then, marked growth has occurred. Industrial reactors for high-tonnage products have capacities of 100 – 300 m^3. Larger bubble columns, with capacities up to 3000 m^3, are employed as fermenters for protein production from methanol. The largest units (20 000 m^3) are those for wastewater treatment.

Scientific interest in bubble columns has increased considerably in the past 10 – 15 years. Up to the mid-1970s, only 10 to 20 publications appeared annually; by the mid- to late 1980s, the number had increased to 80 per year. This led to the development of many empirical correlations and theoretical models enabling the mathematical simulation of bubble- column reactors. Some academic research groups and commercial software developers have offered simulation programs.

The mixing of a liquid and a gas having only partial mutual solubility is one of the unit operations in chemical technology. As Figure 1 shows, this operation takes one of three principal forms. The simplest design is the *bubble column* (Fig. 1 A) in which gas is fed into the column at the bottom and rises in the liquid, escaping from it at the upper surface; the gas is consumed to a greater or lesser extent (depending on the intensity of mass transfer and chemical reaction). When the off-gas contains high concentrations of valuable reactants, part of it is recycled to the reactor. This recycle design, however, lowers the concentration profile in the bubble column and must be optimized from an economic standpoint. In a simple bubble column the liquid is led in either cocurrently or countercurrently to the upward gas stream and has a long residence time. The flow direction of the liquid phase has little effect on the gas-phase residence time, which is comparatively short. Thus, in the

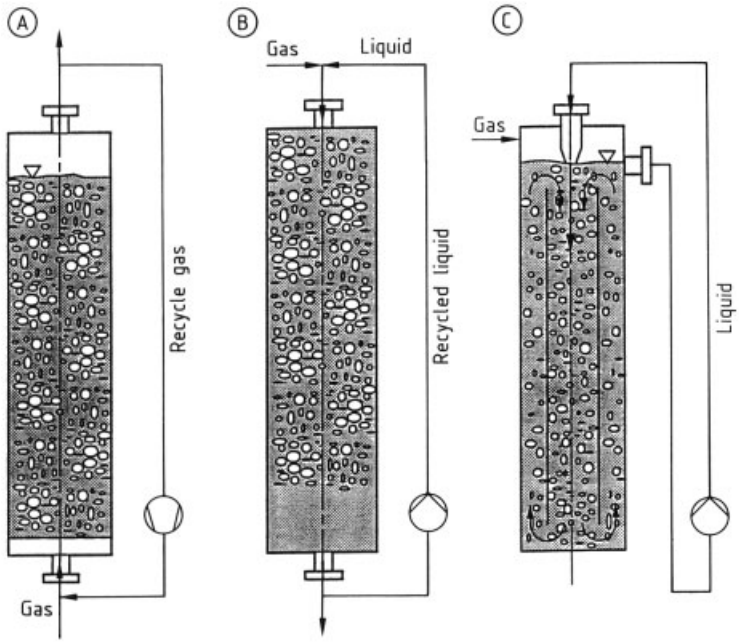

Figure 1. Principal methods of gas – liquid mixing
A) Bubble column; B) Downflow bubble column; C) Jet loop reactor

simple column, the flow of gas is always from bottom to top, and the stream can be made up of both fresh and recycle gas.

Longer gas-phase residence times can be achieved with the *downflow bubble column* shown in Figure 1 B. The liquid is pumped down through the column at a velocity of more than 20 cm/s, so that gas let in at the top is entrained in the flow and can even be held in a suspension-like state until it has reacted completely. Usually, however, unconsumed gas is removed with the liquid and separated. Special designs permit phase separation inside the apparatus. The downflow bubble column is used mainly when large liquid streams are to be contacted with small gas streams and a short liquid residence time is required. The necessary velocity cannot always be obtained with the liquid inlet to the reactor. Thus, like the gas in an ordinary bubble column, the liquid in the downflow bubble column can be recycled. Typical applications for downflow bubble columns are the ozonation of drinking water and the treatment of water in swimming pools. A special use of such devices in the evacuation and compression of gases has also been reported [2].

In both types of column energy must be supplied continuously to the two-phase system to keep the liquid and gas mixed. Only in this way can separation of the phases be counteracted or reversed. In the first case, the simple bubble column, this energy is supplied by the gas. In the downflow bubble column the energy is supplied by the downflowing liquid.

A different mechanism comes into play in the *jet loop reactor* (Fig. 1 C). Here no net flow of gas or liquid occurs along the column; instead, an internal circulating flow is produced. One way to achieve this is with a propeller, but other approaches exist. In the most commonly used type of loop reactor, the *jet loop reactor*, the flow is driven by a high-velocity liquid jet. As in the downflow bubble column, gas is let in at the top and dispersed by the jet energy. Bubbles can be distributed throughout the reactor volume only if the downward liquid flow velocity in the internal tube is greater than the slip velocity of the bubbles. Accordingly, a minimum power input is required.

These three basic methods of dispersing gas in liquid are generally not used in their pure forms. The variety of problems in chemical and biotechnical processes has led to many different contacting devices that combine these basic techniques.

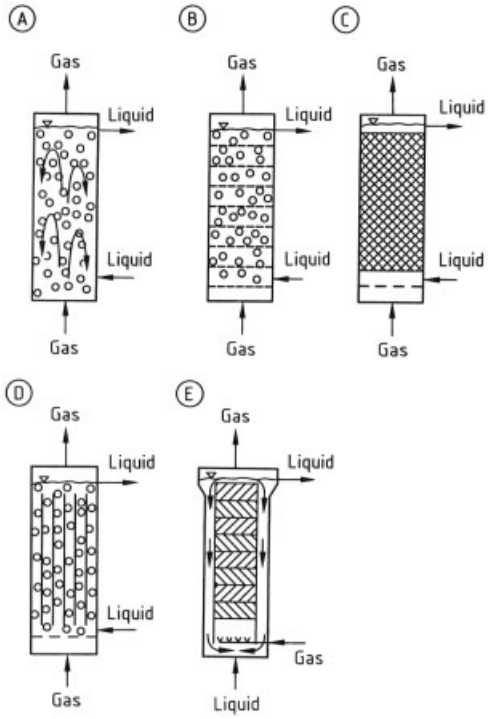

Figure 2. Types of bubble-column reactors
A) Simple bubble column; B) Cascade bubble column with sieve trays; C) Packed bubble column; D) Multishaft bubble column; E) Bubble column with static mixers

2. Bubble Columns and Modifications

2.1. Design and Applications

Bubble columns are very adaptable gas – liquid contacting devices; possible designs are shown in Figure 2. The simplest form of bubble column (Fig. 2 A) consists of a vertical tube with no internals. Gas is fed in at the bottom while liquid is led through the apparatus cocurrently or countercurrently. This simple form is seldom used in practice; instead, a number of modifications are employed. The back-mixing of gas and liquid phases in the simple bubble column and the nonuniform distribution of gas bubbles over the cross section can be reduced by the installation of trays (Fig. 2 B), packings (Fig. 2 C), or shafts (Fig. 2 D). All these devices can operate either cocurrently or countercurrently. To set up the most homogeneous possible bubble flow, static mixer elements can also be placed in the ascending flow section (Fig. 2 E).

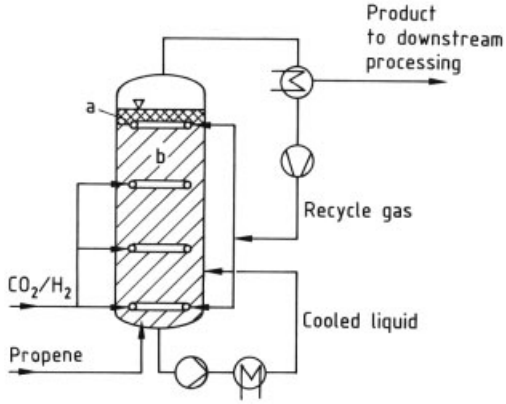

Figure 3. Hydroformylation of propene
a) Stripping zone; b) Reaction zone

Hydroformylation. The hydroformylation of propene is carried out in simple bubble columns. The reaction is homogeneously catalyzed by rhodium complexes. Usually the propene and the CO/H_2 gas mixture are let in at the bottom of the reactor. Incompletely reacted gas, saturated with the reaction product, exits the reactor. The hydroformylation product is separated from the gas stream by condensation and forwarded to downstream processing, while the gas is recycled to the reactor. Because the heat of reaction cannot be completely removed by evaporative cooling using the enthalpy of vaporization of the product, the bubble column is also equipped with an external cooling loop.

One great advantage of the process is that the product is recovered from the reaction mixture without additional separation operations which would damage the expensive catalyst system. The close coupling between the product and the recycle gas necessary to discharge it (i.e., a certain quantity of gas is required for product discharge for thermodynamic reasons), however, presents some problems. First, the gas flow rate causes a high gas holdup, which reduces the reaction volume and thus decreases the productiv-

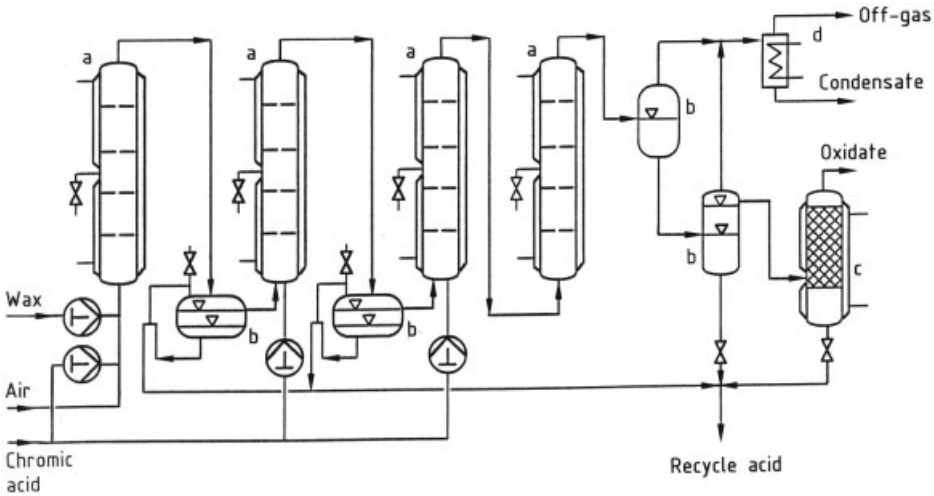

Figure 4. Oxidation of montan waxes in cascade bubble columns
a) Cascade bubble-column reactors; b) Separators; c) Final purification of wax oxidate; d) Off-gas treatment

ity of the reactor. Second, large bubbles occur, which limit the delivery of gaseous reactants to the liquid phase in the reactor. For these reasons, recycle gas is admitted to the bubble column at two levels (Fig. 3) [3]. About half of the recycle gas is fed via the bottom sparger to disperse reactants into the overlying reaction zone. The remaining recycle gas is let in via the top sparger, which lies slightly below the liquid surface, to facilitate separation of the reaction product. Finally, the CO/H_2 reactant stream is fed at various levels to supply CO that has been consumed by the reaction in the liquid phase.

Oxidation of Montan Waxes. Bubble columns are used in a cascade when a narrow residence-time distribution is required, for example, to prevent or limit undesired consecutive reactions. Reducing back-mixing (i.e., a narrow residence-time distribution) is also useful when reaction-engineering considerations dictate that the gas must be fed to various points in the reactor or when a liquid reactant must be degraded to the greatest extent possible.

Montan waxes from brown coal must be deresinified, oxidatively bleached, and esterified (optional) [4], [5]. Oxidation of the waxes consists of several consecutive reactions; the first three steps (oxidation of resins and dark- colored substances, saponification of montan waxes, oxidation of wax alcohols) are desirable, whereas the fourth (oxidative degradation of wax acids) is not. The residence-time distribution in the reactor must be controlled so that the desired reactions go as far as possible without the undesirable reaction occurring to any marked extent. Oxidation is performed in four cascaded bubble columns connected in series (Fig. 4). In the first bubble column, the crude wax for bleaching is metered in along with half of the required amount of chromic acid. Air is supplied to enhance mixing of the reactants. The spent chromic acid is separated from the wax downstream of both the first and the second bubble columns. Another 25 % of the total acid required is added to the second and third columns. The reaction preferably takes place at 100 – 125 °C and 1 – 5 bar, with a residence time of 1 – 3 h for the entire cascade. The enthalpy of reaction is removed by partial evaporation of the water contained in the chromic acid. After exiting the fourth bubble column, the oxidized product, spent acid, and off-gas are separated in two separators.

2.2. Gas Distribution

Usually, the gas is dispersed to create small bubbles and distribute them uniformly over the cross section of the equipment to maximize the intensity of mass transfer. The formation of fine bubbles is especially desirable in coalescence-hindered systems and in the homogeneous flow

regime (Section 2.3). In principle, however, significant mass transfer can be obtained at the gas distributor through a high local energy-dissipation density [6], [7].

In most cases, gas bubbles are generated by pores or holes or in the shear zone of a liquid jet. Figure 5 shows typical forms of *"static" gas spargers*, in which bubble formation occurs without any additional energy supplied from outside. The simplest of these devices, the *dip tube* (Fig. 5 A), only gives an acceptably uniform gas distribution over the cross section at some distance above the sparger. *Perforated plates* (Fig. 5 B) and *perforated ring spargers* (Fig. 5 C) are more effective. Both of these require a certain minimum gas flow rate to achieve uniform distribution and prevent the liquid from getting into the sparger [8–10]. Very fine bubbles can be generated by the use of porous plates (Fig. 5 D), but their pores are susceptible to fouling, and this type of sparger is seldom used in full-scale equipment.

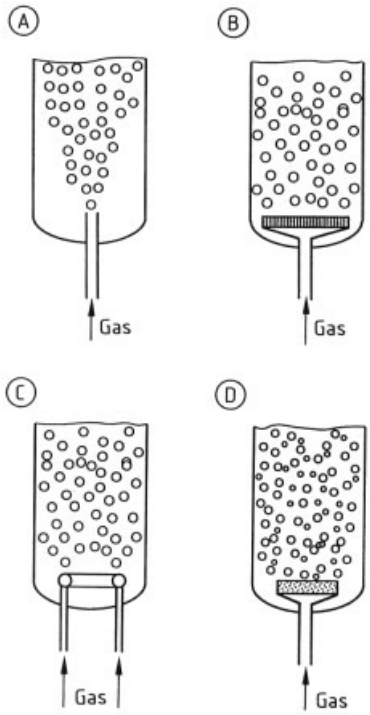

Figure 5. Static gas spargers
A) Dip tube; B) Perforated plate; C) Perforated ring sparger; D) Porous plate

Dynamic spargers offer an alternative to the static types. They use the power of a liquid jet to disperse gas in a zone of high energy-dissipation rate [11–13]. Figure 6 illustrates several frequently used dynamic gas spargers. The simple two-phase jet nozzle alone (Fig. 6 A) or with momentum-transfer tube (Fig. 6 B) is not able to simultaneously disperse gas and suck in the gas stream. This can be achieved, however, with the ejector jet nozzle (Fig. 6 C), the ejector (Fig. 6 D), and the Venturi tube (Fig. 6 E). In nozzle selection the ratio of the gas – liquid volumetric flow rates must always be considered. Common values lie between 0.5 and 2. However, much higher values can be achieved in special cases with momentum-transfer tubes [12].

2.3. Flow Regimes

The upward motion of bubbles gives rise to three distinct flow regimes. The crucial quantity for a flow regime is the superficial gas velocity. The *homogeneous flow regime* is marked by a narrow bubble-size distribution, and bubbles are distributed relatively uniformly over the cross section of the apparatus. This regime extends to superficial gas velocities of 0.03 – 0.08 m/s, depending on the gas – liquid system and gas sparger type.

The uniform distribution of gas bubbles vanishes at higher gas rates, and a highly turbulent flow structure appears. In this *heterogeneous* or *churn-turbulent flow regime*, large bubbles or agglomerates of bubbles form and travel upward at high velocity (see Section 2.6), mainly in the axis of the column. The circulating flow that results may be so vigorous that bubbles of a size corresponding to that in the homogeneous regime are actually transported downward in the zone near the column wall (see Section 2.4).

In the small-diameter columns often used as laboratory equipment, *slug flow* occurs at high gas flow rates. Large bubbles are stabilized by the column wall and take on the characteristic slug shape.

The relationship between superficial gas velocity and reactor diameter is illustrated by the flow map of Figure 7 [14]. The broad transition regions are due to the effects of the gas distributor, the gas – liquid system, and the liquid rate. A knowledge of the flow regime is particularly important because it strongly affects the productivity of bubble- column reactors.

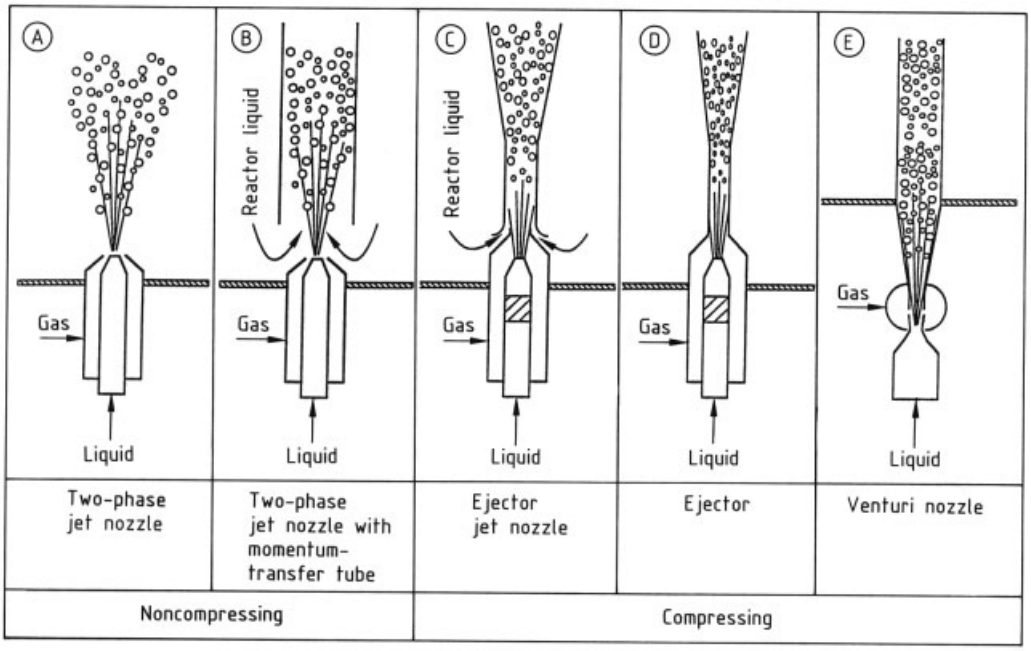

Figure 6. Dynamic gas spargers

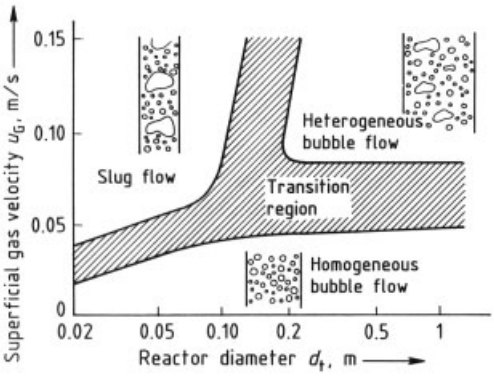

Figure 7. Flow regimes in bubble columns

2.4. Fluid Dynamics

Rising gas bubbles entrain liquid in their wakes. As a rule, this upward flow of liquid is much greater than the net liquid flow rate. Because of continuity, regions therefore exist in which the liquid is predominantly moving downward.

Many theoretical and experimental studies have described the flow behavior of the liquid phase [15]. The circulation velocity is given as a function of superficial gas velocity, column diameter, gas holdup, bubble diameter and rise velocity, viscosity of the liquid, and dispersion height. Published analyses deal with both laminar liquid circulation, which is only of theoretical interest [16–18], and turbulent flow, to which the following discussion is devoted. For example, MIYAUCHI and coworkers use a *force balance* over an annular, axially symmetrical volume element to obtain the velocity profile shown in Figure 8, [19]. Calculation of the velocities, however, requires knowledge of the gas holdup as a function of radial position.

Models of circulation velocity based on *energy balances*, in contrast, assume a cell structure in the bubble column similar to that shown in Figure 9 [20], [21]. In slender bubble columns, both calculations and experimental results show that the height of the circulation cells h_c is equal to the apparatus diameter d_t [20], [22]. JOSHI and SHARMA take into account the energy input due to gas compression and energy losses by dissipation in the wakes of the rising bubbles, as well as liquid transport across the liquid surface (hydraulic pump), thus obtaining a velocity profile over the cross section. HILLS [23] and KOJIMA and coworkers [24] have determined velocity profiles experimentally in bubble columns

having diameters of 0.14 and 5.5 m, respectively (Fig. 10). For the mean circulation velocity $\bar{v}_{\mathrm{L,c}}$ in bubble columns with additional liquid rate u_{L}, JOSHI and SHARMA [20] give the expression

$$\bar{v}_{\mathrm{L,c}} = 1.4\sqrt[3]{g d_{\mathrm{t}}\left(u_{\mathrm{G}} \pm \frac{\varepsilon_{\mathrm{G}} u_{\mathrm{L}}}{1-\varepsilon_{\mathrm{G}}} - \varepsilon_{\mathrm{G}} v_{\mathrm{rG}}\right)} \tag{2.1}$$

where ε_{G} is the gas holdup (+ for countercurrent, − for cocurrent). ZEHNER, using a force balance, arrives at a similar relation for the mean circulation velocity [21]:

$$\bar{v}_{\mathrm{L,c}} = \sqrt[3]{\frac{1}{2.5} \cdot \frac{\Delta\varrho}{\varrho_{\mathrm{L}}} g d_{\mathrm{t}} u_{\mathrm{G}}} \tag{2.2}$$

The velocity profiles derived from the models and, in particular, the mean velocities enable calculation of the essential fluid-dynamic parameters in bubble columns [20], [21], [25].

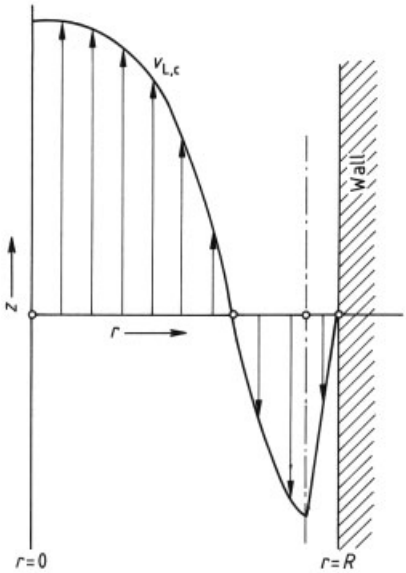

Figure 8. Radial distribution of liquid velocity in a bubble column

2.5. Bubble Size

Analysis of bubble size in bubble columns must distinguish between bubble-size distribution just after bubble formation at the sparger and size distribution further away from the distributor. Because of breakup and coalescence of the rising bubbles, the two distributions can differ significantly. Since the efficiency of bubble columns depends chiefly on bubbles far from the gas distributor,the following discussion only concerns these.

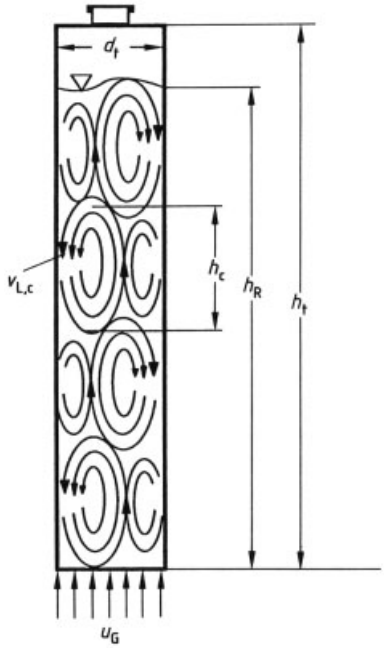

Figure 9. Cell structure in bubble columns

Two basic methods – photography and probe techniques – exist for determining bubble size; however, they do not lead to identical results. Both methods are subject to certain limitations in view of the marked bubble selection that may occur (i.e., not all bubble sizes can be detected) [26], [27]. In particular, any measurement method only leads to realistic results if the flow is homogeneous (i.e., a narrow bubble-size distribution is found). As yet, no method can be recommended for the measurement of large bubbles in the heterogeneous flow regime.

If bubbles are generated in a region of high turbulence (as with *dynamic gas spargers*), the following formula [28] can be used to describe the Sauter diameter d_{bS} (mean bubble diameter, calculated from the volume to surface ratio) [29], [30].

$$d_{\mathrm{bS}} = \frac{2}{e_{\mathrm{M}}^{0.4}} \left(\frac{\sigma}{\varrho_{\mathrm{L}}}\right)^{0.6} \varepsilon_{\mathrm{G}}^{0.5} \left(\frac{\eta_{\mathrm{G}}}{\eta_{\mathrm{L}}}\right)^{0.25} \tag{2.3}$$

This formula is based on Kolmogorov's theory of isotropic turbulence.

When *static gas spargers* are used, the bubble diameter is only weakly dependent on gas velocity. Descriptive correlations [31–34] are applicable only to the systems and sparger geometries for which they were obtained; a generally valid description of bubble size does not yet exist. The maximum bubble diameter $d_{\mathrm{b,\,max}}$ can be used for purposes of estimation [27], [35]. For low-viscosity liquids, the maximum bubble diameter is given by

$$d_{\mathrm{b,max}} = \sqrt[3]{\frac{\sigma}{g\varrho_{\mathrm{L}}}} \tag{2.4}$$

where σ is the surface tension. For the water – air system, $d_{\mathrm{b,\,max}} = 8$ mm. Larger bubbles have a high probability of being unstable and thus breaking up. The Sauter diameter for real distributions is between 40 and 60 % of the largest stable bubble diameter. This estimate is not, however, applicable to the heterogeneous flow regime due to the binodal bubble-size distribution in this regime.

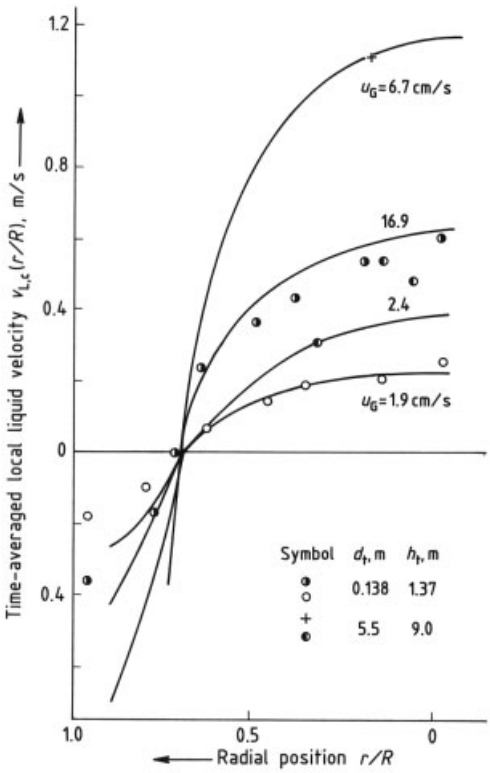

Figure 10. Calculated radial profiles of liquid velocity in bubble columns [20]

2.6. Bubble Rise Velocity

In the homogeneous flow regime, bubbles of almost uniform size and shape rise in the form of a swarm distributed uniformly over the column cross section. When the regime changes, larger bubbles or agglomerates of bubbles form in addition to the bubbles already present [36], [37]. These aggregates rise at a markedly higher velocity than the small bubbles. Figure 11 shows measured velocities for large and small bubbles [36]. Large bubbles first appear at a superficial gas velocity of ca. 0.03 m/s. The formation of large bubbles, however, depends strongly on the type of sparger used. With sintered plates, for example, larger bubbles do not appear at gas rates lower than ca. 0.1 m/s. As shown in Figure 11, large bubbles have a rise velocity that is four or more times larger than small ones. Thus, most of the gas transport in the heterogeneous flow regime is accomplished by large bubbles. In this regime, the quantity of gas transported by small bubbles remains constant, whereas the quantity transported by large bubbles increases linearly with gas velocity. This relationship applies to coalescing and coalescence-hindered gas – liquid systems.

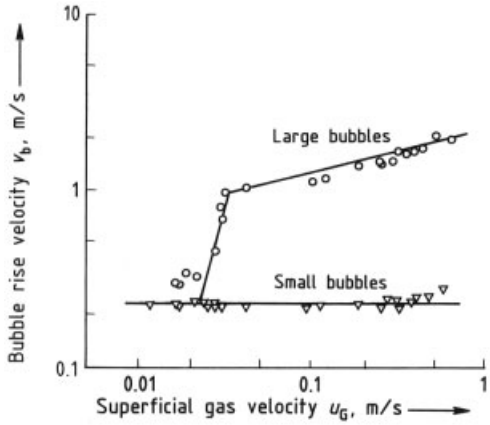

Figure 11. Velocities of rising bubbles for the system water – air
Reactor: $d_t = 0.44$ m, $h_t = 5$ m; Gas distributor: perforated plate ($d_h = 3$ mm)

2.7. Dispersion of the Liquid Phase

Because of the large-scale circulation flows, back-mixing occurs in both phases. The resulting dispersion flow J_D is usually governed by an

equation analogous to Fick's first law for molecular diffusion. For the one-dimensional case of axial dispersion, which is generally sufficient for a description, follows

$$J_{\mathrm{D}} = -D_{\mathrm{L}} \frac{\mathrm{d}c}{\mathrm{d}z} \tag{2.5}$$

The dispersion coefficient D_{L} is essentially a function of the superficial gas velocity and the column diameter (e.g., see [38]). Flow direction or liquid velocity does not show any effect, provided the superficial liquid velocity remains within the range common in industry ($u_{\mathrm{L}} < 0.03$ m/s). The dispersion coefficient can be estimated fairly accurately on the basis of fluid-dynamic models. For example, Joshi and Sharma [20] and Zehner [21] give dispersion coefficients derived from the mean circulation velocity. Each of these formulas gives a good description of the experimentally determined dispersion coefficients known from the literature. By way of example, Figure 12 compares experimental results reported by various workers with the theoretical relation derived by Zehner:

$$D_{\mathrm{L}} = \frac{d_{\mathrm{t}} \bar{v}_{\mathrm{L,c}}}{2} = \frac{d_{\mathrm{t}}}{2} \sqrt[3]{\frac{1}{2.5} \cdot \frac{\Delta\varrho}{\varrho_{\mathrm{L}}} g d_{\mathrm{t}} u_{\mathrm{G}}} \tag{2.6}$$

The equation emphasizes that D_{L} strongly depends on column diameter.

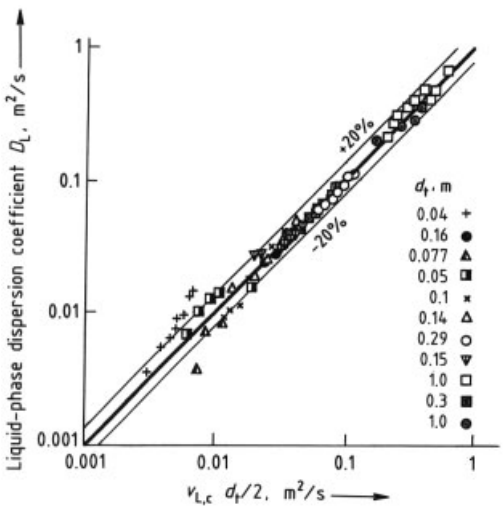

Figure 12. Liquid-phase dispersion coefficient measured by various authors [39]

2.8. Dispersion of the Gas Phase

Due to the large-scale circulation flow both the liquid and gas phases are dispersed. Furthermore, the formation of large and small bubbles, coalescence, and breakup result in additional dispersion in the gas phase. Whereas the gas phase in a bubble column with a smaller diameter flows with virtually no back-mixing, large units behave more like stirred tanks. The gas-phase dispersion coefficient depends more strongly on gas velocity and column diameter than does that of the liquid phase. For this reason, the degree of axial gas mixing is especially relevant for scale-up when the gas phase is expected to show strong concentration variations.

Many formulas in the literature describe the dispersion coefficient as a function of different independent variables. A particularly suitable formula is [40]:

$$D_{\mathrm{G}} = 5 \times 10^{-4} \left(\frac{u_{\mathrm{G}}}{\varepsilon_{\mathrm{G}}} \right)^3 d_{\mathrm{t}}^{1.5} \tag{2.7}$$

This formula is not, however, dimensionally homogeneous (D_{G} in cm^2/s, u_{G} in cm/s, d_{t} in cm), and the gas holdup must be known. By contrast, the equation

$$\varepsilon_{\mathrm{G}} D_{\mathrm{G}} = 0.2 d_{\mathrm{t}} u_{\mathrm{G}} \frac{\frac{\Delta\varrho}{\varrho_{\mathrm{L}}} g d_{\mathrm{t}} u_{\mathrm{G}}}{v_{\mathrm{rG}}^3} \tag{2.8}$$

derived by Zehner and Schuch is dimensionally correct [41]. However, more recent measurements [42] have shown that this correlation must be modified in the heterogeneous flow regime ($u_{\mathrm{G}} \geq 5$ cm/s) because the proportionality is somewhat different:

$$\varepsilon_{\mathrm{G}} D_{\mathrm{G}} \sim (u_{\mathrm{G}} d_{\mathrm{t}})^{1.65}$$

2.9. Gas Holdup

Gas holdup is one of the most important operating parameters because it not only governs phase fraction and gas-phase residence time but is also crucial for mass transfer between liquid and gas. Gas holdup depends chiefly on gas flow rate, but also to a great extent on the gas – liquid system involved. Accordingly, many correlations that

have been published only apply to the systems investigated.

Gas holdup is defined as the volume of the gas phase divided by the total volume of the dispersion:

$$\varepsilon_G = \frac{V_G}{V_G + V_L} \tag{2.9}$$

The relationship between gas holdup and gas velocity is generally described by the proportionality

$$\varepsilon_G \sim u_G^n$$

In the homogeneous flow regime, n is close to unity. When large bubbles are present, the exponent decreases, i.e., the gas holdup increases less than proportionally to the gas flow rate (Fig. 13). The higher the contribution of large bubbles to the total gas holdup, the smaller is the exponent n. In the fully developed heterogeneous flow regime, n finally takes on values between 0.4 and 0.7, depending on the gas – liquid system.

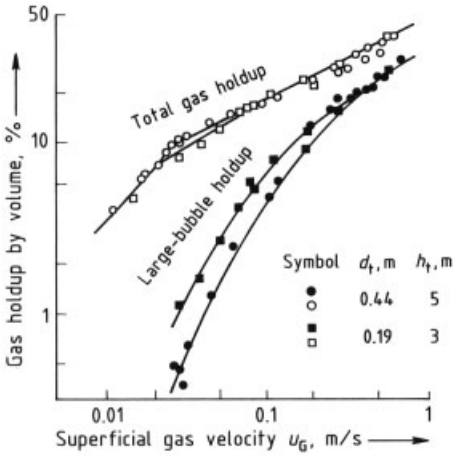

Figure 13. Gas holdup and fraction of large bubbles (system: water – air; gas distributor: perforated plate $d_h = 3$ mm)

The effect of low *liquid velocities* u_L on gas content is generally negligible. At high flow rates the gas holdup decreases in cocurrent systems because gas bubbles pass through the column more quickly. In contrast, the gas holdup rises in countercurrent systems; this can lead to extremely high gas holdup, especially in downflow bubble columns [43].

Above 0.1 m, the reactor diameter is of secondary importance for gas holdup, as measurements on units having diameters between 0.1 and 5.50 m show [34], [36], [44], [45].

The effects of *physical properties* on gas holdup are exceedingly complex. Increasing the viscosity of the liquid phase leads to increased bubble coalescence and thus a decrease in gas holdup. Above ca. 50 mPa · s, however, the gas holdup remains constant [46]. Although surface tension is not very important for the gas holdup, a change in coalescence behavior may have lasting effects. When gas-phase residence times are long and gas distribution is obtained with perforated or sintered plates, the presence of salts or alcohols that counteract coalescence has little effect [44]. In contrast, gas holdup increases markedly in systems sparged by two-phase nozzles when coalescence is hindered [47]. Such behavior can be attributed to a small-bubble gas holdup higher than that in coalescing systems, whereas the content of large bubbles is identical [36]. Small bubbles formed under high shear stresses in the region near the two-phase nozzle cannot recombine so the gas holdup increases significantly with this type of gas distributor.

The relation of Akita and Yoshida [48] is suitable for estimating the gas holdup and is based on the investigation of numerous systems:

$$\frac{\varepsilon_G}{(1-\varepsilon_G)^4} c_1 = \left(\frac{g d_t^2 \varrho_L}{\sigma}\right)^{\frac{1}{8}} \left(\frac{g d_t^3}{\nu_L}\right)^{\frac{1}{12}} \left(\frac{u_G}{\sqrt{g d_t}}\right) \tag{2.10}$$

For pure liquids and nonpolar solutions the constant c_1 is 0.2, for electrolyte solutions it is 0.25. However, reliable results cannot be expected for systems that have not been investigated in this study. The effects of *reactor pressure* on gas holdup have not been fully explained. Although some authors find no effect between 1 and 16 bar [49], others find that gas holdup increases with pressure in systems with small sparging holes ($d_h \leq 1$ mm) or with sintered plates [44], [50–52]. Transition from the homogeneous to the heterogeneous regime occurs at higher gas flow rates as pressure increases.

Gas holdup is generally a function of position in the bubble column. Axial profiles of gas holdup show a zone near the gas distributor in which the holdup increases to the value that characterizes the following equilibrium zone. The gas holdup at the top of the column, in the zone

of bubble breakup, is markedly higher than the equilibrium value [45].

Gas holdup also depends on radial position. The profile shows gradients only near the wall in the homogeneous flow regime [23], [45], [53]. In contrast, a parabolic radial gas holdup distribution appears in the heterogeneous regime [23], [37], [45], as a consequence of the preferential rising of large bubbles or agglomerates of bubbles in the axis of the column. Figure 14 shows radial profiles of gas holdup at various gas flow rates in the water – air system.

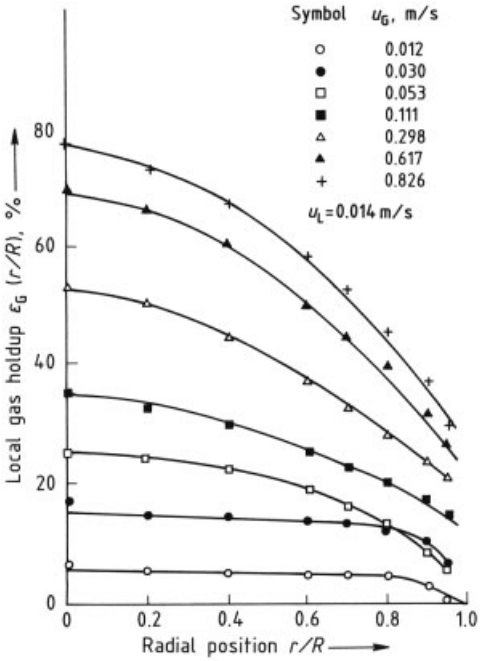

Figure 14. Radial profiles of local gas holdup
d_t = 0.45 m; h_t = 6.2 m; h = 3.03 m (at measurement point); perforated plate d_h = 1 mm

2.10. Specific Interfacial Area

The area of the gas – liquid interface is one of the most important process parameters. Especially at high reaction rates (e.g., when a bubble column is employed as an absorber), the interfacial area becomes a crucial factor in equipment sizing. Like gas holdup, interfacial area depends on the geometry, operating conditions, and gas – liquid system. Gas holdup and interfacial area per unit volume are related as

$$a = \frac{A}{V_R} = \frac{6\varepsilon_G}{d_{bS}} \tag{2.11}$$

where V_R is the volume of the reaction mixture and d_{bS} is the mean bubble diameter (Sauter diameter, Section 2.5). As Figure 15 shows, the interfacial area increases with increasing gas flow rate. An exception occurs when a porous-plate sparger is used; like gas holdup, interfacial area decreases on transition to the heterogeneous flow regime and then approaches the same values observed with perforated plates. The growth in interfacial area with increasing gas velocity is always greater in the homogeneous than in the heterogeneous flow regime. The reason lies in the formation of large bubbles in the heterogeneous regime: the interfacial area of large bubbles per unit volume is markedly lower than that of smaller ones.

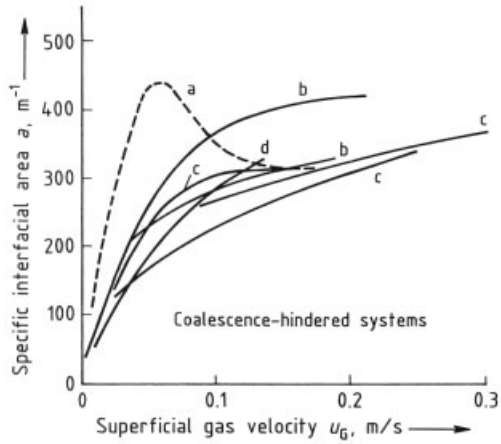

Figure 15. Specific interfacial area as a function of superficial gas velocity
a) d_t = 0.102 m; b) d_t = 0.29 m; c) d_t = 0.14 m; d) d_t = 0.1 m; – – – Porous plate; —- Perforated plate

The specific interfacial areas attainable in various gas – liquid reactors can be compared on the basis of power input P per unit volume [29]. Experimental values can be described by the relation

$$a = k\left(\frac{P}{V_R}\right)^m \varepsilon_G^n \tag{2.12}$$

The exponent m is between 0.4 and 1 [54]. The plot in Figure 16 enables a direct comparison to be made between reactors with respect to the energy required to produce a given interfacial area.

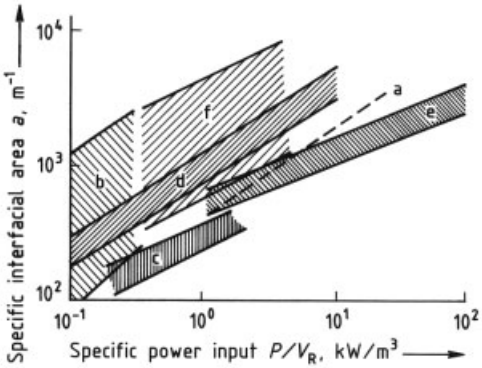

Figure 16. Specific interfacial area as a function of specific power input [55]
a) Stirred tank; b) Bubble column with porous plate; c) Bubble column; d) Bubble column with two-phase jet nozzle (jet loop reactor); e) Packed column; f) Bubble column with injector nozzle

2.11. Volumetric Mass-Transfer Coefficient

The mass transfer between the gas and the liquid phase in a bubble column can be described in most cases by the volumetric mass-transfer coefficient $k_L\,a$, which is the liquid-phase masstransfer coefficient k_L multiplied by the specific interfacial area. Gas-phase resistance can usually be neglected, so $k_L\,a$ gives an adequate description. To determine the mass-transfer rate, however, the driving concentration difference must be known which in turn requires a knowledge of mixing behavior in the gas and the liquid phase. In industrial units ($d_t > 1$ m), estimates can be based on the assumption of complete mixing in both liquid and gas phases.

Like gas holdup and interfacial area, $k_L\,a$ also depends on the gas flow rate, type of sparger, and gas – liquid system. The mass-transfer coefficient and the gas rate are again proportional to one another:

$$k_L a \sim u_G^n$$

where n can be between 0.7 and 0.92 [31], [56–59]. Mass-transfer coefficients two- to threefold higher can be achieved in the homogeneous flow regime if a porous plate is used as sparger instead of a perforated plate (Fig. 17). In the heterogeneous regime, however, the effect of the sparger is negligible.

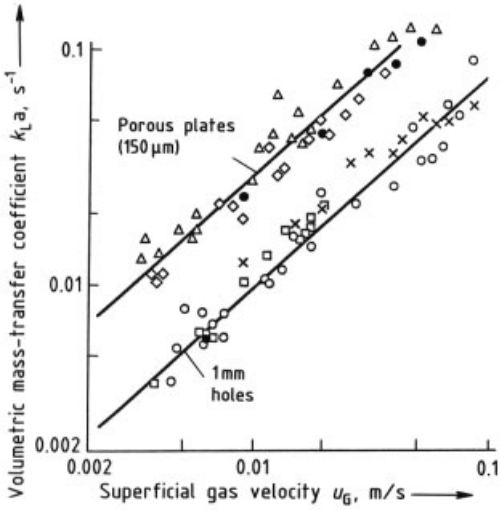

Gas distributor	d_t, m	h_t, m	Flow direction	Symbol
Cross sparger	0.20	0.250	↑↑	×
		0.723	↑↑	○
		0.723	↑↓	□
Porous plate	0.10	0.250	↑↑	●
	0.15	0.440	↑↑	△
	0.15	0.440	↑↓	◇

Figure 17. Mass-transfer coefficients in bubble columns

According to experimental results, the column diameter above about 15 cm has no effect on mass-transfer coefficient. Some correlations nonetheless include reactor diameter [31], [57], [60]. Akita and Joshida [31] state that the value of the column diameter used for calculation should not be increased beyond 0.6 m. Based on this premise, their correlation for $k_L\,a$ is

$$\frac{k_L a d_t^2}{D_{G,L}} = 0.6\left(\frac{\nu_L}{D_{G,L}}\right)^{0.5}\left(\frac{g d_t^2 \varrho_L}{\sigma}\right)^{0.62}\left(\frac{g d_t^3}{\nu_L^2}\right)^{0.31}\varepsilon_G^{1.1} \tag{2.13}$$

and has the best experimental support. The mass-transfer coefficient increases in coalescence-hindered systems [54], [61]. This increase depends on the system and the concentration of coalescence-hindering substance. The maximum gain in mass-transfer coefficient due to the presence of electrolytes, however, is only 30 %.

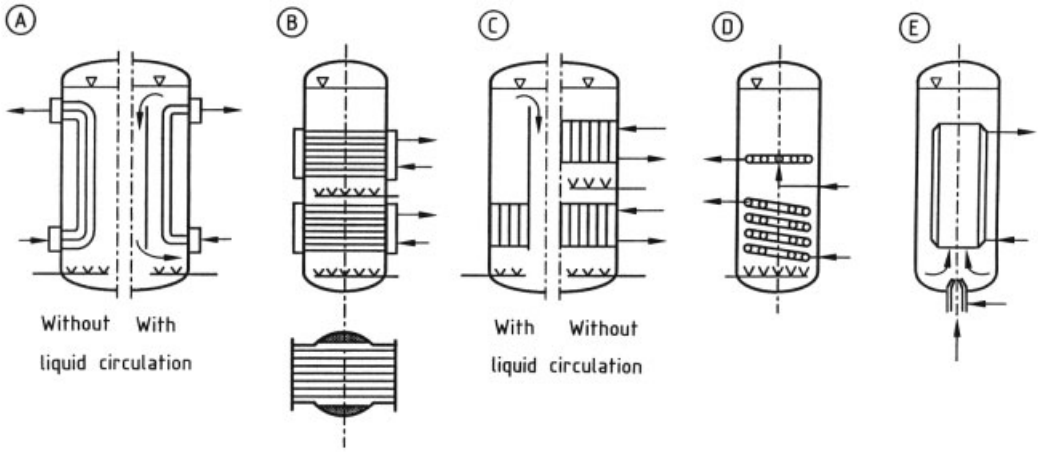

Figure 18. Examples of heat exchanger inserts
A) Parallel single tubes; B) Tube bundle in cross flow; C) Longitudinal tube bundle; D) Tube spiral or helix; E) Draft tube with jacket

2.12. Heat Transfer

In many cases, heat must be removed when operating bubble columns. A particularly simple solution is to utilize the latent heat of vaporization of the liquid phase for heat removal, although this is not always feasible. In addition, many possibilities exist for heat transfer through heated or cooled surfaces, as shown in Figure 18. In this way, up to ca. 30 m^2/m^3 of heat-transfer area can be installed in a bubble column.

The turbulent flow generated by rising gas bubbles increases heat transfer even at low gas rates (Fig. 19). The increase in heat-transfer coefficient α, with gas throughput is markedly greater in the homogeneous than in the heterogeneous regime.

The heat-transfer coefficient does not depend on the column diameter, type of sparger, or coalescence behavior of the gas – liquid system.

Two distinct concepts are used to describe the heat-transfer coefficient at the wall. Whereas Kast [62] and Deckwer [63] consider radial flow and heat transported by it, Joshi and coworkers [64] and Zehner [21] use circulation velocities derived by them for physical modeling. Here with the following relation for the heat-transfer coefficient α can be derived [21]:

$$\alpha = 0.18\,(1-\varepsilon_G)\left(\lambda_L^2 \varrho_L c_{pL} \cdot \frac{\bar{v}_{L,c}^2}{l \cdot \nu_L}\right)^{1/2} \tag{2.14}$$

where

$$l = d_b \left(\frac{\pi}{6\varepsilon_G}\right)^{1/3}$$

and $\bar{v}_{L,c}$ is calculated by Equation (2.2). On the whole, these two approaches correlate well with literature data. Heat transfer in bubble columns with heat-exchange internals has not been intensively studied [65–69]. For tube bundles arranged in an axial direction (Fig. 18 C), the heat-transfer coefficient increases with increasing tube pitch and decreases when the free cross-sectional area increases [68], [69]. A similar relationship is found for a tube bundle arranged in cross flow (Fig. 18 B), but here a marked effect of liquid throughput occurs [67].

The installation of tube bundles leads to an overall change in fluid dynamics and thus in mixing behavior. For example, tubes installed in cross flow hinder flow in the longitudinal direction and thus reduce dispersion in the liquid phase [70]. In contrast, the arrangement of heat-transfer surfaces in the flow direction leads to more intense mixing of the liquid phase by intensifying circulation [71], [72].

2.13. Slurry Bubble Columns

Solid particles are present in bubble columns in a wide variety of processes; they must be held

in suspension by the rapid liquid circulation already discussed. The presence of the solid phase in a slurry bubble column means that all process parameters behave differently, and in a more complicated way, than in a two-phase bubble column.

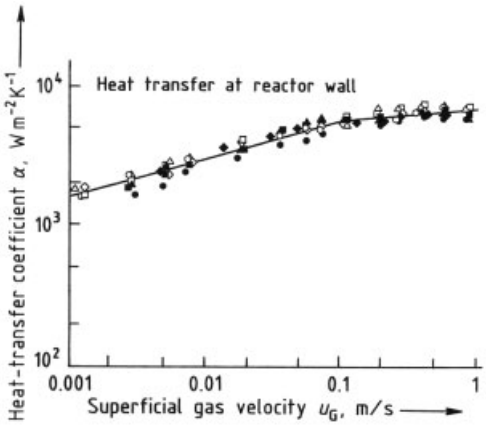

Symbol	Distributor	d_h, mm	System
○	porous plate	0.02	water – air
△	porous plate	0.15	
◇	perforated plate	1.00	
□	perforated plate	2.00	
●	porous plate	0.02	NaCl solution – air
▲	porous plate	0.15	
◆	perforated plate	1.00	
■	perforated plate	2.00	

Figure 19. Heat-transfer coefficient at reactor wall $d_t = 0.196$ m; $h_t = 6.20$ m; liquid velocity $u_L = 1.2$ cm/s

The *minimum gas velocity* necessary to hold the solids in suspension increases as the concentration and density of the particles increase. The increment depends, however, on the physical properties of the solid and liquid phases. Many empirical equations for the critical gas velocity for complete suspension show a marked increase with increasing single-particle settling velocity. Equation (2.15) [73] can be used for design purposes:

$$\frac{u_{G,\min}}{v_{rS}} = 0.801 \left(\frac{\varrho_S - \varrho_L}{\varrho_L}\right)^{0.6} \left(\frac{\bar{\varphi}_S}{\varrho_S}\right)^{0.146} \left(\frac{\sqrt{g d_t}}{v_{rS}}\right)^{0.24} \left[1 + 807 \left(\frac{g \eta_L^4}{\varrho_L \sigma^3}\right)^{0.578}\right] \tag{2.15}$$

where v_{rS} is the relative settling velocity of the particle swarm in the liquid. At low solids concentration (< 10 wt %) and low settling velocity of the particles the *gas holdup* is nearly unchanged [74]. In contrast, the gas holdup decreases at higher settling velocities with increasing solids concentration [75–77]. The strength of this effect differs from one flow regime to another. The decrease is particularly marked when an increase in solids content leads to a change from the homogeneous to the heterogeneous regime. On the other hand, in the heterogeneous regime the reduction in gas holdup is only slight with increasing solids content. YASUNISHI and coworkers [78] verify and recommend the gas holdup relation of KOIDE and coworkers [79]

$$\frac{\varepsilon_G}{(1-\varepsilon_G)^4} = \frac{0.277 \left(\frac{u_G \eta_L}{\sigma}\right)^{0.918} \left(\frac{g \eta_L^4}{\varrho_L \sigma^3}\right)^{-0.252}}{1 + 4.35 \left(\frac{\bar{\varphi}_S}{\varrho_S}\right)^{0.748} \left(\frac{\varrho_S - \varrho_L}{\varrho_L}\right)^{0.881} \left(\frac{d_t u_G \varrho_L}{\eta_L}\right)^{-0.168}} \tag{2.16}$$

for a wide range of solids concentrations and liquid properties. For coalescence-hindered aqueous electrolyte solutions a coefficient 0.364 must be used instead of 0.277 in Equation (2.16). The *mixing behavior* of the liquid and the gas phase is very similar to that in the two-phase bubble column. The axial solids distribution can be described by a one-dimensional sedimentation – dispersion model. The *solids dispersion coefficient* is generally lower than the corresponding liquid dispersion coefficient [80]. The difference between the two values increases rapidly with increasing settling velocity of particles. For small solid particles ($v_{rS} < 0.01$ m/s), the effect of superficial gas velocity on the axial solids concentration profile is negligible above the minimum gas velocity for suspension [81].

The *specific interfacial area* declines continuously with increasing solids content [76], [82]. This phenomenon can be explained by the formation of larger gas bubbles, due to the presence of solid particles that lead to the observed decrease in gas holdup. If, however, very fine particles are used in aqueous electrolyte solutions (i.e., systems with hindered coalescence), the interfacial areas produced do not differ from those in the two-phase system [83].

The effect of solids on the *volumetric mass-transfer coefficient* depends largely on particle properties, solids content, and physical properties of the liquid. At low solids concentration (up

to ca. 3 – 5 wt %), the mass-transfer coefficient matches that of the two-phase bubble column. A higher solids content, like increasing particle size, leads to a drop in $k_L\,a$ relative to the solid-free condition [78], [79], [84], [85]. For very fine particles ($d_P < 36\,\mu m$), however, $k_L\,a$ also decreases with decreasing particle diameter [86]. Overall, the behavior of $k_L\,a$ is governed by the interfacial area per unit volume because the change in k_L is generally small.

The mass-transfer coefficient between *liquid and solid* increases roughly as the fourth root of the gas flow rate, decreases with increasing liquid viscosity and particle diameter, and becomes partly independent of these factors at high gas rates. Two distinct models can be used for the mathematical description, but they lead to similar values if the solids are completely suspended. Values reported by various authors are compared in Figure 20 [87].

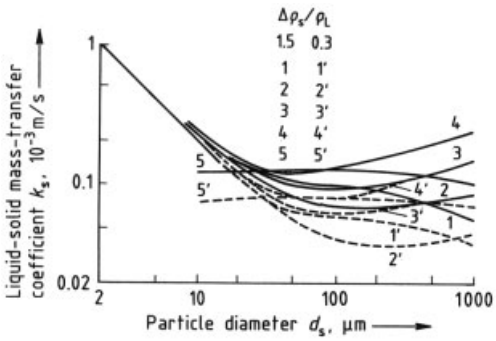

Figure 20. Liquid – solid mass-transfer coefficient estimated with five different correlations [87]
$\nu_L = 10^{-6}\,m^2/s$; $\varrho_L = 1000\,kg/m^3$; $D_{G,L} = 10^{-9}\,m^2/s$;
—· $\Delta\varrho_S/\varrho_L = 1.5$; – – – $\Delta\varrho_S/\varrho_L = 0.3$

Hydrogenation of Benzene. In the IFP process, benzene is hydrogenated to cyclohexane in a slurry bubble column [88] (→ Cyclohexane, Chap. 4.1.). This process was used to produce 1.8×10^6 t of cyclohexane in 23 plants worldwide in 1991.

Figure 21 shows the slurry bubble column (a) in which benzene is hydrogenated on suspended Raney nickel. The hydrogen-rich gas that is let in at the bottom of the main reactor provides hydrogen for the reaction and also strips the product cyclohexane out of the reactor. Thus, the process can be operated without the need for expensive equipment to separate the product from the catalyst. In the external cooling loop, the high heat of reaction is removed and suspension of the solid catalyst is assisted by the circulating liquid stream. Complete conversion of benzene is accomplished in a fixed-bed reactor (b) installed downstream on the gas side.

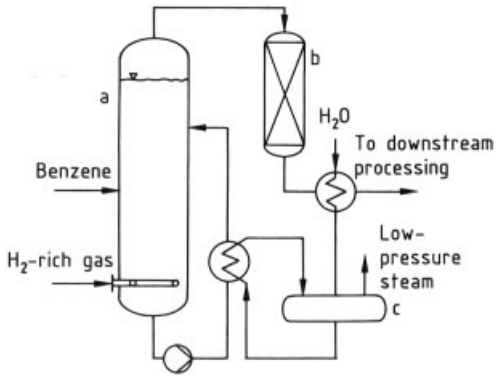

Figure 21. Hydrogenation of benzene to cyclohexane
a) Main reactor; b) Secondary reactor; c) Steam drum $h_t = 10$ m; $d_t = 2.5$ m; $T = 195\,°C$; $p = 22$ bar; Gas velocity: 7.5 cm/s; Liquid residence time ~ 3 h

The usual reaction conditions in the bubble-column reactor are 200 °C and 22 bar. Typical gas velocities are ca. 0.08 m/s; liquid-phase residence time is ca. 3 h.

2.14. Airlift Loop Reactors

In contrast to bubble columns, airlift loop reactors are characterized by a well-defined liquid circulation, which is achieved by dividing the reactor into sections with and without gas sparging. The difference in gas holdup between these two zones drives the liquid circulation. In principle, two types of airlift loops can be identified (Fig. 22). In the first (airlift reactors with internal loop), either a concentric tube (Fig. 22 A, B) or a plane partition (Fig. 22 C) divides the column into riser and downcomer sections. In the second (airlift reactors with external loop, Fig. 22 D), two separate tubes form the upflow and downflow zones; the tubes are joined by two horizontal sections at top and bottom.

The dependence of liquid circulation velocity on superficial gas velocity is described by the purely empirical relation

$$\bar{v}_{L,c} = C_1 u_G^{C_2} \tag{2.17}$$

The value of the constant C_1 is determined by reactor geometry and the physical properties of the system; C_2 depends on both flow regime and reactor geometry [89–91]. A physical model based on an energy balance [39], [92] leads to the relation

$$\bar{v}_{\mathrm{L,c}} = c \cdot \sqrt[3]{g h_{\mathrm{R}} u_{\mathrm{G}}} \tag{2.18}$$

An exponent of 0.33 for gas flow rate is of the order of usual experimental results.

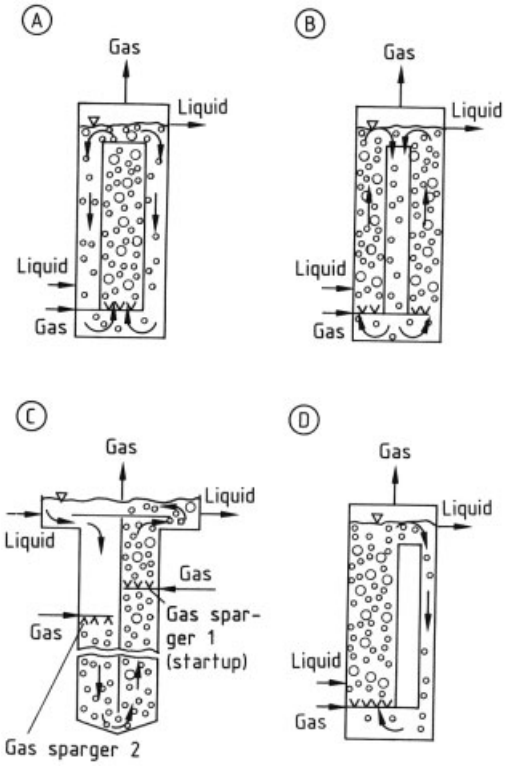

Figure 22. Types of airlift loop reactors
A) Concentric draft tube with external recycle; B) Concentric draft tube with internal recycle; C) Deep shaft reactor (ICI); D) External loop

Airlift reactors with external loop (Fig. 22 D) are usually run at much higher gas and liquid flow rates than conventional bubble columns. The high circulation velocities significantly change the nature of the two-phase flow, namely, the gas holdup declines with increasing circulation velocity (see Fig. 23). The highest gas holdup occurs in the bubble column ($u_{\mathrm{L}} = 0$), where the absolute velocity of the rising bubbles is lowest because of the zero liquid velocity. Parameters such as surface tension, coalescence, and viscosity have much less effect on airlift loop reactors than on bubble columns because the interactions between bubbles are far weaker as a result of the high circulation velocity. For the same reason, the homogeneous flow regime in airlift loop devices extends to much higher gas rates than in bubble columns [91], [93].

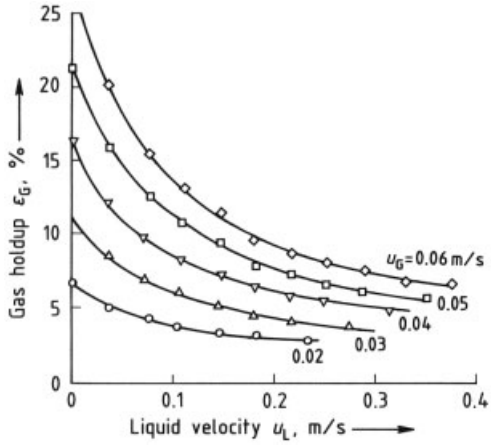

Figure 23. Gas holdup in airlift reactors with external loop (system: 0.1 mol/L NaCl solution – air) $d_t = 0.1$ m; $h_t = 8.5$ m; Porous plate = 150 µm

The gas holdups that occur in *airlift reactors with internal loop* are only slightly lower than those in bubble columns. The decrease in gas holdup in the riser is partly offset by an increase in the downcomer [94], [95].

As in the bubble column, the *volumetric mass-transfer coefficient* increases with increasing gas flow rate. Because the liquid-phase mass-transfer coefficients k_{L} are the same in bubble columns and airlift reactors [96], [97], the difference in $k_{\mathrm{L}}\,a$ results from differences in interfacial area. Airlift reactors with external loop always have lower mass-transfer coefficients than bubble columns because the lower gas holdup implies a smaller area for mass transfer. The $k_{\mathrm{L}}\,a$ values for airlift reactors with internal loop, on the other hand, are similar to the values for bubble columns (Fig. 24) because here the gas holdups differ only slightly [93], [98–100].

In contrast to bubble columns, the residence-time distribution of airlift reactors is influenced not just by longitudinal mixing but also by back-mixing due to the circulation flow. Reactors with an external loop exhibit axial dispersion coefficients whose values are up to 20 % lower than

those of bubble columns, depending on the circulation velocity [93]. The consequences of this back-mixing for the reactor yield of airlift loops, however, are far less serious because the high liquid velocities lead to far larger Bodenstein numbers (for a definition of the Bodenstein number, see → Mathematical Modeling). The axial dispersion coefficients of airlift reactors with internal loop are much lower. Measurements of axial mixing of the gas and liquid phases show a decrease in the dispersion coefficients by roughly a factor of three [101].

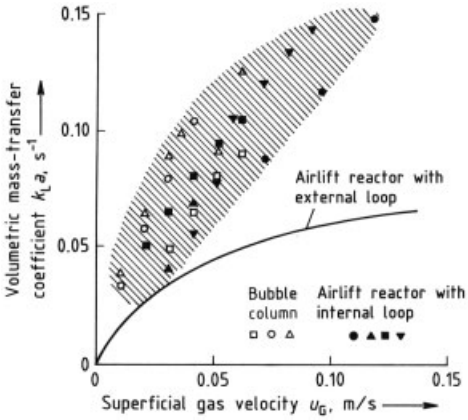

Figure 24. Comparison of mass-transfer coefficients for airlift reactors and bubble columns
(system: salt solutions; gas distribution: small bubbles)

As in bubble columns, mixing times decrease with increasing superficial gas velocity because the circulation velocity becomes greater. Since the circulation time t_c and the mixing time t_M are directly proportional [92]

$$\frac{t_M}{t_c} = \gamma \left(\frac{F_D}{F_R}\right)^{0.5} \tag{2.19}$$

(where $\gamma = 3.5$ for internal circulation and 5.2 for external circulation), the relationship between mixing time and circulation velocity can be expressed as

$$t_M \sim \bar{v}_{L,c}^{-1}$$

The mixing time is also directly proportional to the distance traveled. Five to six passes are required for a degree of mixing of 90 %. The liquid circulation in airlift reactors, with their high circulation velocities, leads to higher heat-transfer coefficients than in bubble columns [102]. As in bubble columns, the heat-transfer coefficient in airlift loop reactors increases with gas flow rate.

Biological Wastewater Treatment. Airlift loops are employed to provide well-defined back-mixing of the liquid phase. This is desirable, for ex-ample, when uniform temperature and concentration distributions must be maintained in the reaction medium to equalize feed variations as quickly as possible or to prevent settling of solids from the mixture.

Airlift reactors are used in biological wastewater treatment [5], [103], [104]. These units are closed vessels ca. 15 – 25 m tall and 10 – 45 m in diameter; they have small space requirements, very good oxygen utilization, and greatly reduced off-gas and noise emissions. The contents of the reactor circulate through one or more draft tubes (Fig. 25); sparging occurs outside the draft tubes. During operation, the growing microorganisms must constantly be provided with sufficient oxygen and substrate, and adequate mixing of the wastewater – activated-sludge mixture must be insured. If these conditions are not satisfied, solids will settle and anaerobic fermentation processes may occur. The usual conditions are as follows: superficial gas velocity 1 – 3 mm/s, gas holdup between 1 and 3 %, and roughly 25 circulations of reactor contents per hour. The utilization of atmospheric oxygen is more than 50 %. The wastewater has a residence time between 6 and 15 h.

3. Downflow Bubble Columns

Chapter 2 described ordinary bubble columns in which gas flows from bottom to top and has a short residence time (gas sparging method A in Fig. 1). In downflow bubble columns, by contrast, the gas and liquid phases are transported together from top to bottom (Fig. 1 B). This regime demands liquid velocities v_L greater than the relative velocity v_{rG} between the two phases. Depending on the liquid velocity chosen, very low gas velocities can be achieved

$$v_G = v_L - v_{rG} \tag{3.1}$$

or long residence times. This is an advantage especially when a large-volume liquid stream must come in contact with a small-volume gas stream. In the extreme case, a virtually suspended state

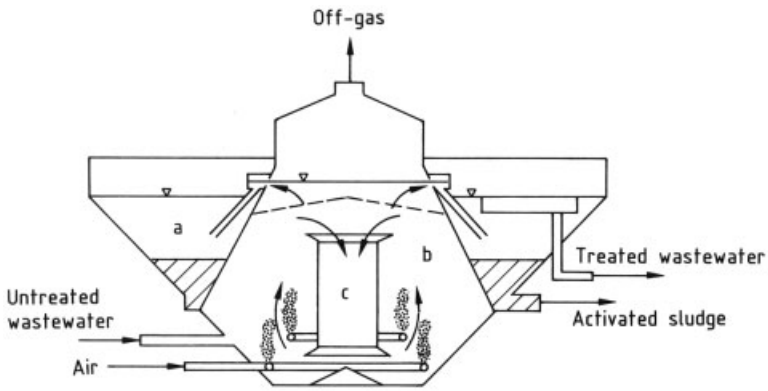

Figure 25. Biohoch reactor (Hoechst)
a) Settling zone; b) Aeration chamber; c) Draft tube

of the bubbles can be realized, with an arbitrarily long residence time. Under certain conditions, this permits complete conversion of the gas. Normally, however, part of the gas must be assumed to exit the reactor without reacting, mainly when the gas contains components that do not react. In such cases, the cocurrent motion of the phases is a disadvantage because only one theoretical transfer unit can be realized.

3.1. Design and Applications

As with bubble columns, a variety of designs exist for downflow bubble columns. These differ mainly in the way the gas is let in, the bubbles are generated, and the unreacted gas is removed. Figure 26 shows some examples.

The *simple downflow bubble column* (Fig. 26 A) is particularly suitable for gases that are soluble in the liquid phase and/or fast reactions. Unreacted gases cannot be separated in the column, so an extra separator may be required. The simple downflow bubble column is often employed at high pressure (> 100 bar). A slender geometry makes it possible to reduce the wall thickness of the cylindrical reactor. To improve mass transfer between gas and liquid phases, the vessel can be packed with particles, which also reduce both the required liquid rate and the axial mixing of liquid and gas. Usually, however, packings are used as catalyst supports. The classical application of this type of device is the hydrogenation of a wide range of substances.

Adding a liquid recycle creates diverse process design options. The back-mixing involved, which is usually undesirable, can often be acceptable. With a liquid recycle the downflow bubble column can be operated on small feedstreams. The recycle loop also provides a simple way of adding or removing heat, so that the temperature profile in the reactor becomes more uniform.

From the standpoint of process engineering, the *downflow bubble column with integrated separator* (Fig. 26 B) differs little from the simple downflow bubble column. The integrated separator is well suited when larger quantities of off-gas must be removed. A typical application is in the ozone treatment of water air or oxygen with a low ozone content is fed to the reactor and the quantity of exit gas is almost the same as the quantity of inlet gas. The design of the reactor with integrated separator is simple. The shouldered form is not suitable for high pressure.

The *downflow – upflowbubble column* (Fig. 26 C) combines a downflow bubble column and an ordinary bubble column. Particularly long gas residence times are possible. The liquid routing shown in Figure 26 C gives a frequently desirable residence-time distribution: the downflow section features mixing similar to a stirred tank by virtue of the pump stream. This is advantageous with a high heat of reaction, which can be removed with the pump stream. The bubble column in the outer annular space merely carries the outflow, and back-mixing in this zone can be suppressed by internals (pack-

ing, static mixers, sieve trays). High conversion are achieved in higher-order reactions.

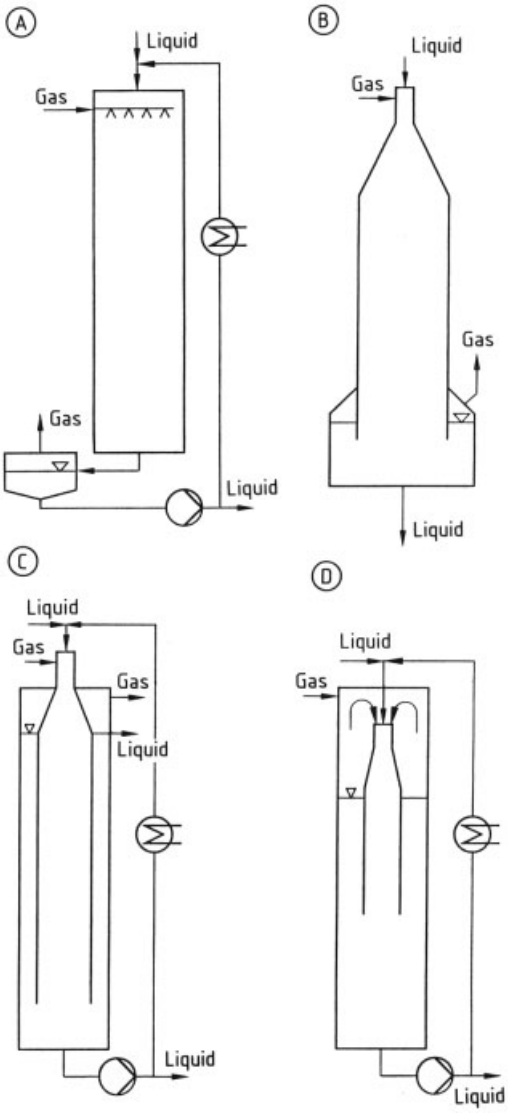

Figure 26. Designs of downflow bubble columns
A) With external gas separator; B) With integrated gas separator; C) Combined downflow – upflow with bubble column in annulus; D) Dip-tube gas sparging with internal gas recycle

Downflow – Upflow Bubble Column (Fig. 26 D). Another design combines downflow and ordinary bubble columns. However, the top of the downflow section is in the gas space of the reactor. Fresh gas together with recycle gas (that has escaped from the liquid surface) is drawn in here and dispersed in the liquid. Pure gases or gases with low inerts content can be completely converted under pressure with this method, also called *dip-tube sparging*.

The lower part of the downflow bubble column serves as the separator. Only small gas bubbles are carried out of the reactor, which still have some reactivity. The pump is therefore protected against excessive contents of gas in the liquid, even in coalescence-hindered systems.

3.2. Operating Conditions and Gas Holdup

Gas is fed in at the top of the column and distributed as uniformly as possible over the cross section. Large gas holdups ($\varepsilon_G \approx 0.3-0.35$) can be obtained even in coalescing systems such as water – air [105–110]. When the system is coalescence-hindered, for example, if carboxymethyl cellulose or ethanol has been added to the water – air system, values up to $\varepsilon_G \approx 0.45$ can be achieved [108].

The gas holdup can be estimated as

$$\varepsilon_G = \frac{u_G}{v_L - v_{rG}} \tag{3.2}$$

The relative velocity v_{rG} for coalescing aqueous systems takes on values in the range $0.2\,\text{m/s} \leq v_{rG} \leq 0.3\,\text{m/s}$. The liquid velocity v_L can be calculated from the liquid rate u_L and the gas holdup ε_G:

$$v_L = \frac{u_L}{1-\varepsilon_G} \tag{3.3}$$

Hence the gas holdup can be expressed as

$$\varepsilon_G = \frac{u_G}{\frac{u_L}{1-\varepsilon_G} - v_{rG}} \tag{3.4}$$

If gas and liquid rates are given, Equation (3.4) can be solved for the gas holdup:

$$\varepsilon_G = \frac{B}{2}\left[+\sqrt{\left(\frac{2}{B}\right)^2 \frac{u_G}{v_{rG}} + 1} - 1\right]$$

where

$$B = \frac{u_L + u_G}{v_{rG}} - 1 \tag{3.5}$$

Finally, Equation (3.4) can be written so as to yield the liquid rate:

$$u_L = (1-\varepsilon_G)\left(\frac{u_G}{\varepsilon_G} + v_{rG}\right) \tag{3.6}$$

A flow chart can then be derived for the water – air system ($v_{rG} = 0.23$ m/s). In Figure 27, the gas velocity is plotted versus the superficial liquid velocity, with the gas holdup ε_G as a parameter. For $\varepsilon_G > 0.35$, flow is in the heterogeneous regime. In this regime a highly turbulent two-phase flow develops, resulting in strong mixing of liquid and gas. In the extreme case, gas accumulates at the top and can propagate throughout the reactor from there. KULKARNI reports somewhat different results [111], possibly because of less uniform gas feeding.

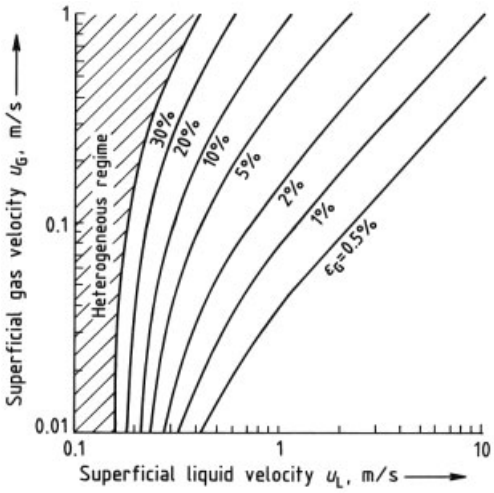

Figure 27. Flow chart for downflow bubble columns
Gas velocity is plotted as a function of liquid velocity, with gas holdup as parameter, for the water – air system ($v_{rG} = 0.23$ m/s) calculated with Equation (3.6).

For systems with hindered coalescence, determination of these values and other process parameters normally requires experimentation [105].

3.3. Mass Transfer

The following statements are applicable only to coalescing systems. In the homogeneous flow regime, all bubbles in the downflow bubble column are almost equal in size. Diameters of $3\text{ mm} \le d_b \le 4\text{ mm}$ are observed [112]. A slight decrease in bubble size with increasing liquid flow rate and rising pressure has been reported [108].

If the gas holdup is known, the specific interfacial area can be estimated as

$$a = \frac{6\varepsilon_G}{d_{bS}}$$

which gives maximum values up to $a = 450 - 700\text{ m}^{-1}$. Figure 28 compares these values to those measured in simple bubble columns and packed columns [108] (referred not to the total volume but to the liquid volume). This figure illustrates the advantages of downflow bubble columns, which have higher a_L values at low gas rates.

A *liquid-phase mass-transfer coefficient* $k_L \approx 3.7 \times 10^{-4}$ m/s has been calculated by plotting the volumetric mass-transfer coefficient $k_L\,a$ versus the volumetric gas holdup [110] and the bubble size stated above. Somewhat lower estimates, up to 3×10^{-4} m/s, have been reported in [109], [112]. However, the mass-transfer coefficient k_L cannot be measured directly, and large uncertainties are to be expected. The volumetric mass-transfer coefficients $k_L\,a$ measured in bubble columns at equal gas holdups are roughly the same. Because of the very different volumetric gas flow rates in upflow and downflow bubble columns, different gas conversions are achieved; Figure 29 compares these figures for sulfite oxidation with air [110].

Data on the axial back-mixing of the gas phase have been reported [111]. At gas rates of $u_G = 0.001 - 0.01$ m/s, the dispersion coefficient takes on constant high values: $D_G \approx 0.2\text{ m}^2/\text{s}$ ($d_t = 0.025$ m, $u_L = 0.334$ m/s, $\varepsilon_G = 0.025 - 0.09$).

Back-mixing of the liquid phase under comparable conditions is also markedly less than in normal bubble columns [113].

Examples. STEINER and HERBRECHTSMEIER studied the *oxidation of sulfite solutions* with air in the downflow bubble column, finding a twelvefold higher depletion than with a simple bubble column (Fig. 29) [110].

A process for the *absorption of gases containing nitrogen oxides* has been reported [114]. Virtually complete conversion can be obtained through the use of a downflow bubble column 3 m high.

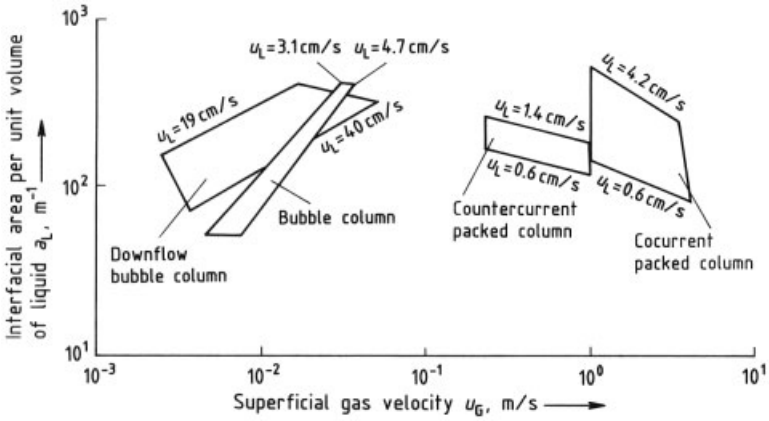

Figure 28. Comparison of mass-transfer area per unit liquid volume for various sparged devices as a function of gas velocity, with liquid velocity as parameter, according to [108]

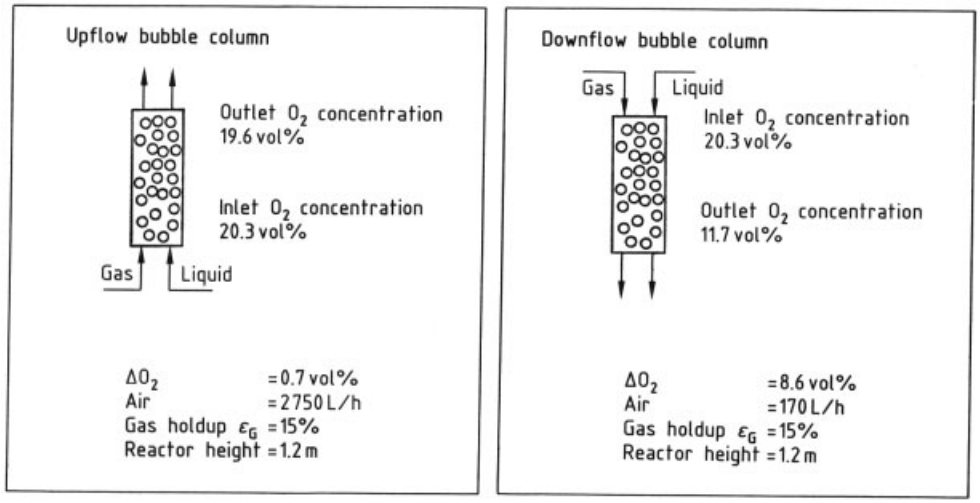

Figure 29. Comparison of oxygen depletion in upflow and downflow bubble columns for sulfite oxidation with air [110]
$T = 22\,°C$; $[SO_3^{2-}] = 0.4 - 0.8$ mol/L; $[Co^{3+}] = 7 \times 10^{-6}$ mol/L; pH = 8.0

The *degradation of organic contaminants by ozonolysis* is a well-known method of water treatment. For economic and safety reasons, virtually complete depletion of the ozone is desirable, which can be achieved in downflow bubble columns as reactors [115]. No danger of fouling exists in these devices, and high liquid throughputs can be handled. Figure 30 is a flow sheet of the entire process. Only a single theoretical mass-transfer unit can be realized, but this drawback can be overcome by the proposed use of a reactor cascade [115]. The phases are led through the cascade countercurrently.

4. Jet Loop Reactors

Jet loop reactors are among the most versatile gas – liquid contactors. The momentum of the liquid jet issuing from the nozzle enhances internal circulation and opposes demixing of the phases (distribution method C in Fig. 1). The liquid jet can be utilized to suck in, compress, and disperse fresh or recycle gas (Figs. 31, 32, 33, 34). The liquid-jet data are important process parameters. The liquid volumetric flow rate, velocity, and power can be varied over wide ranges. Finally, the size of the draft tube and the upper

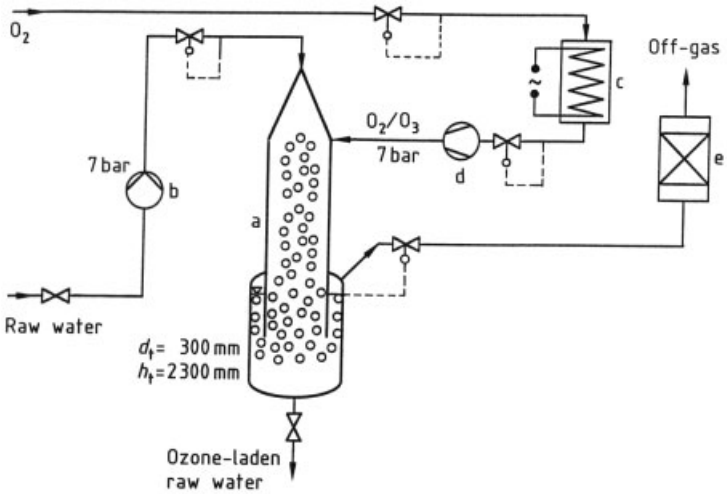

Figure 30. Plant with downflow bubble column for ozone treatment of water
a) Downflow reactor; b) Pressurizing pump; c) Ozone generator; d) Compressor; e) Deozonizer

flow-reversal zone strongly affect fluid dynamics and gas separation. Other possible variations are offered by the nozzle configuration (Figs. 32, 33, 34).

The four examples in Figure 31 illustrate options for the direction flow pattern phases. In all cases the gas is incorporated into the liquid via the nozzle located in the gas space. The liquid jet entrains gas bubbles until the nozzle orifice is closed by the rising liquid surface. The incorporation of more gas submerges the nozzle and blocks the gas inlet, the surface level then drops again; this self-regulating mechanism enables the gas holdup to be controlled.

A jet loop reactor (also called a gas-circulation reactor) [116] without net gas or liquid throughput is shown in Figure 31 A. Both phases are let in at the top and discharged at the top (the gas phase is consumed to a higher or lower degree). This corresponds to dispersion method C of Figure 1.

If the gas is under pressure, it can also be let in at the bottom to intensify circulation. The result is a net gas rate, as in the bubble column (Fig. 31 B); here, distribution methods A and C of Figure 1 are combined.

Figure 31 C shows a combination of sparging methods B and C in Figure 1. The liquid is fed at the top and discharged at the bottom. This jet loop reactor has an additional net flow of liquid, as in a downflow bubble column.

A combination of all three distribution methods (bubble column, downflow bubble column, and jet loop) is shown in Figure 31 D. The process characteristics of one sparging type will predominate, depending on the selected gas and liquid flow rates. At high gas flow rates, for example, the liquid surface level rises above the nozzle orifice. The liquid jet then no longer entrains gas bubbles, serving only to drive the circulation and disperse the bubbles. This versatile type of distribution can be further refined through variation of the nozzle position and the use of self-priming ejectors (Figs. 32, 33, 34). To evaluate a design, the essential process parameters must be estimated, which is not always feasible because of the many possible variations. For the basic forms shown in Figure 31, however, some information can be derived from a power balance (Section 4.2).

4.1. Design and Applications

Loop Reactors with Downflow Liquid Jets. Figure 32 illustrates several loop reactors with downward-pointing nozzles. These reactors are even suitable for suspension catalysis. If the pump or feedstream is cut off or fails, the nozzle drains clear, and the danger of the nozzle being plugged by the suspended catalyst is thus reduced. Another advantage is the long gas res-

idence time. From the inlet at the top, the gas circulates through the loop at least once. Moreover, the devices are designed so that the gas is internally recycled. This is important for the complete conversion of gases containing little or no inerts.

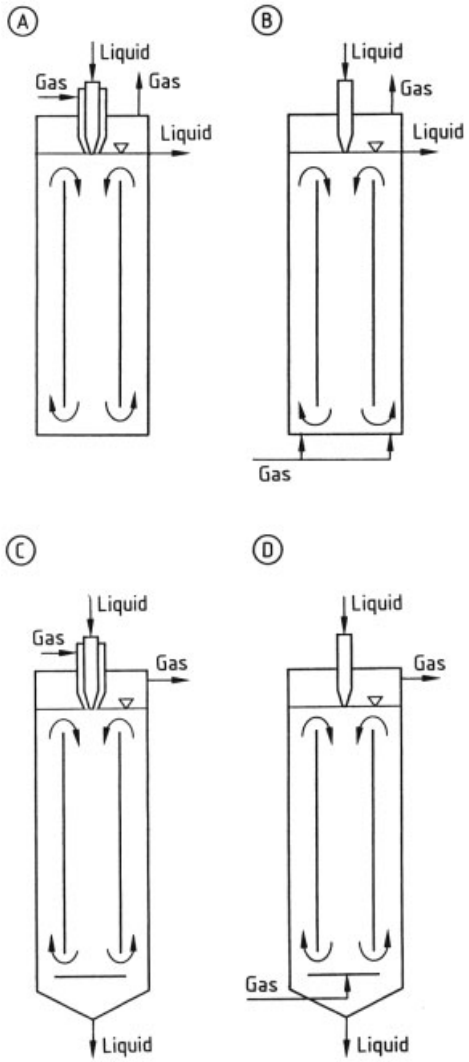

Figure 31. Options for phase routing in the "gas-circulation" type of jet loop reactor
A, C) With surface gas sparging; B, D) With pressure sparging (with gas throughput); C, D) With bottom outlet for liquid, inlet at top (with liquid throughput)

In the gas- circulation reactor of Figure 32 A (see also Fig. 31) [116], the gas can also be let in at the bottom of the reactor, independently of the nozzle, if the gas is available at reactor pressure. For a given jet power, this design offers much higher gas holdups and better mass-transfer performance (Sections 4.5 and 4.6). The jet only has to supply the recycle gas. Another marked increase (up to a factor of two) in the gas holdup is achieved by installing a *momentum-transfer tube* in the reactor. This can be submerged (Fig. 33 B) or can extend above the liquid surface (Fig. 33 C).

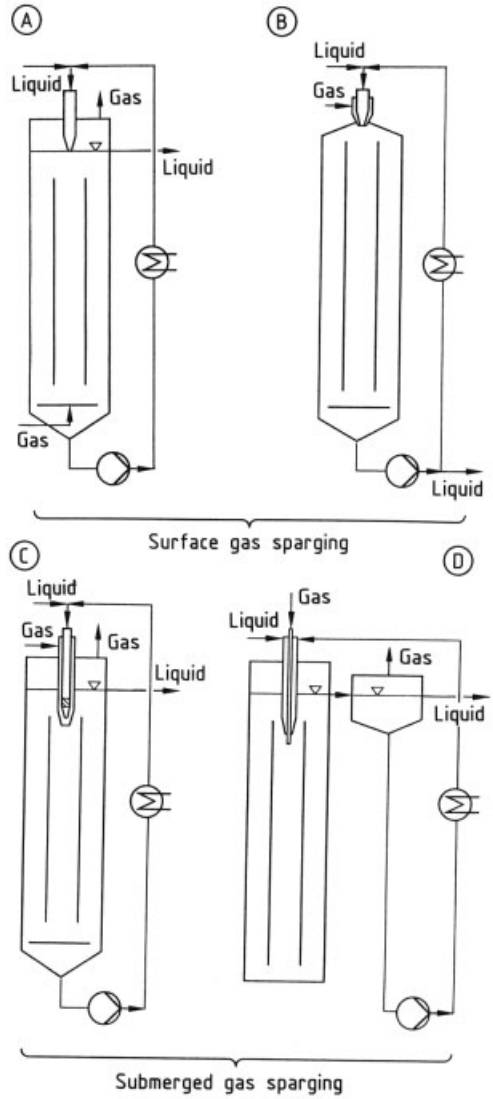

Figure 32. Types of jet loop reactors with downward liquid jet
A) Gas-circulation reactor with pressure sparging; B) Gas-circulation reactor without connected gas space; C) Gas-circulation reactor with submerged nozzle; D) "Compact" reactor with annular space for liquid injection

To prevent separation of the gas phase, the nozzle can be built directly into the reactor top (Fig. 32 B) [117]. Gas bubbles separating in this zone are immediately entrained by the liquid jet and redispersed into the circulating flow.

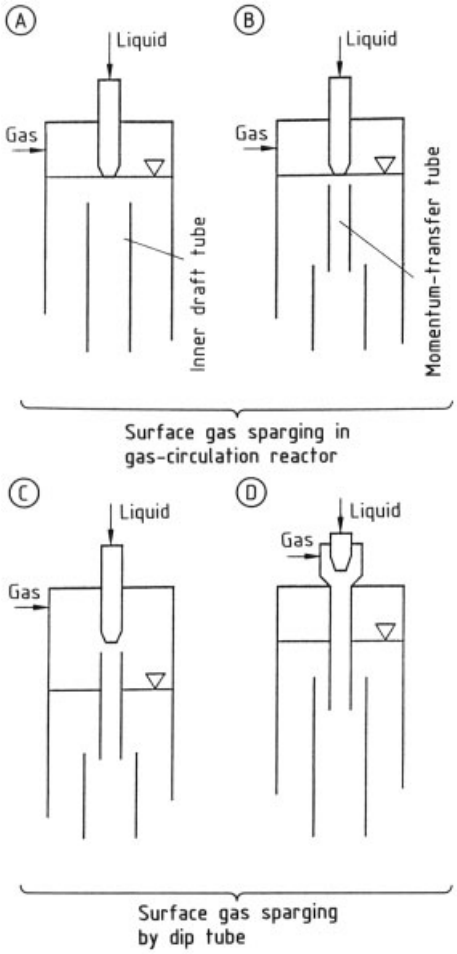

Figure 33. Options for spontaneous gas sparging from top Surface sparging in gas-circulation reactor without (A) and with (B) momentum-transfer tube; dip-tube sparging without (C) and with (D) self-priming ejector

The submerged nozzle in Figure 32 C can either accept pressurized gas from outside the system [118] or suck the gas in. Internal gas recycle is also possible. The suction of the nozzle can be enhanced by applying a swirl to break up the liquid jet, provided the liquid nozzle orifice is set back somewhat to the rim of the nozzle. The ejectors and ejector nozzles discussed in Chapter 2 are suitable for deeper submergence (see also Fig. 34 B and C).

In the jet loop reactor proposed by RÄBIGER and coworkers, the liquid is fed in via an annular nozzle (Fig. 32 D) [119], [120]. Gas can be drawn in via the center tube (ejector fashion) or supplied under pressure (injector fashion). For application of this reactor type in wastewater treatment see [121], [122].

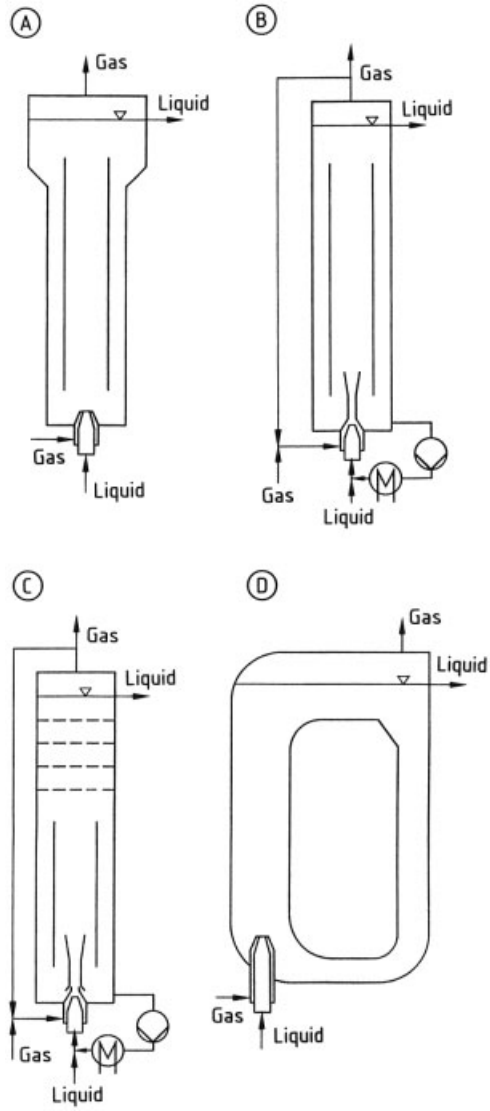

Figure 34. Design options for loop reactors with upward liquid jet
A) With expanded head for gas separation; B) With gas recycle via a self-priming ejector; C) With gas recycle via an ejector jet nozzle and perforated plates; D) With external recycle tube

Submerged nozzles supply the gas in the loop flow. The penetration depth of the jet differs. The reactor is thus easier to start up and the circulation flow can be build up in a simpler way. This is particularly important in batch processes. In principle, submerged nozzles represent an intermediate stage between surface sparging (Fig. 32 A and B) and pressurized gas sparging through nozzles at the bottom of the reactor (Fig. 34).

In the gas-circulation reactor of Figure 33, gas sparging takes place through the free surface. If a momentum-transfer tube is used (Fig. 33 B), gas enters the circulating flow at a greater depth. This arrangement functions even when the liquid level drops and the momentum-transfer tube extends into the gas space (Fig. 33 C). In Figure 33 D, the tube is led outside the reactor, so that gas can be delivered directly from outside without any mixing with recycle gas. The last two types of gas sparging (Fig. 33 C and D) are also referred to as dip-tube sparging designs.

Jet Loop Reactors with Upflow Liquid Jet (Fig. 34). In the second major variant of the jet loop reactor, the nozzle points upward. This design is closely related to the airlift loop reactor (Section 2.14). The liquid jet mainly produces smaller gas bubbles so that conversion of the gas phase can be improved. On the other hand, the circulation velocity is increased. At least a portion of the gas is thereby driven through the reactor faster than in airlift reactors, with the possible result of lower conversion; this danger exists particularly in the reactor of Figure 34 A, which uses a jet nozzle. The widened separation zone at the top can reduce gas recirculation or cut it off altogether; the result is improved mixing of the liquid phase because the circulation velocity increases with increasing gas holdup in the downcomer. Gas recycle is recommended whenever part of the gas leaves the column unconverted; it can be accomplished without a mechanical recycle-gas compressor if an ejector or an ejector jet nozzle is employed (Fig. 34 B and C). The ejector jet nozzle offers a higher compression efficiency than the ejector and produces larger interfacial areas [12]. From the chemical reaction engineering standpoint the combination of a loop reactor with a series of perforated plates is very interesting (Fig. 34 C). The lower part (loop reactor) exhibits the residence-time distribution of a stirred tank. Back-mixing is suppressed in the second section (perforated plates). For higher-order reactions, the conversion of both liquid and gaseous components can be increased in this way. With pure gases, recycle ensures adequate distribution because the gas flow rate is sufficient for an even gas load of the perforated plate.

Figure 34 D shows a reactor (analogous to the airlift loop of Fig. 22 D) with external recirculation. This device has an operational behavior comparable to that of reactors with internal circulation. Its advantages include better gas separation and simpler heating or cooling facilities. The accessible heat-transfer area is larger, and a conventional heat exchanger can be integrated directly into the loop.

The most important *characteristics* of jet loop reactors with upward-pointing nozzles can be summed up as follows:

1) They are particularly suitable for higher gas throughputs or when the gas has a high content of inerts
2) They have larger interfacial areas than bubble columns, especially with system with hindered coalescence
3) They offer more intensive back-mixing than bubble columns
4) They are less suitable for suspension catalysis, because the nozzle can become plugged

4.2. Typical Dimensions

In jet loop reactors, *height-to-diameter ratios* h_t/d_t of 5 to 20 are common. When several nozzles are used along with internal tubes, arbitrarily small height-to-diameter ratios can be achieved. Values larger than 20 are also seen in high-pressure operation and in pilot plant reactors.

The optimal *diameter* d_i *of the internal tube* is dictated by the direction of flow. If flow in the tube is *downward*, tubes with a diameter ratio d_i/d_t of 0.2 – 0.5 are suggested [116], [119]. The narrower internal tubes have the principal advantages of higher gas holdup and better mass-transfer at low energy dissipation rate (up to 1 kW/m^3). As jet power increases, the wider tubes become better in these respects [119], [123].

For *upward flow*, BLENKE and coworkers determined the optimal internal tube diameter as $d_i/d_t = 0.59$ [124]; this result applies to single- and two-phase reactors.

A crucial parameter of the jet loop reactor is the *nozzle diameter* d_n. Common values are in the range of $d_n/d_t = 0.02-0.1$. For a given jet power, large nozzles are more efficient than smaller ones for coalescing systems. In systems with hindered coalescence, in contrast, this relationship can be reversed.

4.3. Energy Balance

In the general case of Figure 31 D (combination of bubble column, downflow bubble column, and jet loop reactor), a total of five mechanical power terms can be identified [125]. These can be referred to the reaction volume

$$V_R = \frac{\pi}{4} d_t^2 h_R \tag{4.1}$$

Three types of power are *delivered to the reactor*:

1) Jet thrust power e_{thrust} per unit volume:

$$e_{thrust} = f_i \varrho_L \frac{v_n^2 v_{L,i}}{h_R} \left(\frac{d_n}{d_t}\right)^2 \tag{4.2}$$

where f_i is the fraction of cross-sectional area of the inner tube and v_n is the nozzle velocity

2) Power input per volume due to liquid flow e_L (as in a downflow bubble column) :

$$e_L = \Delta\varrho g \varepsilon_G u_L \tag{4.3}$$

3) Power input per unit volume due to gas flow e_G (as in a bubble column; holds approximately only for small pressure changes):

$$e_G \approx \Delta\varrho g \left(1-\varepsilon_G\right) u_G \tag{4.4}$$

This power is *transformed to heat* by two mechanisms:

1) Power dissipation per unit volume due to circulation flow:

$$e_{circ} = -\zeta f_i \bar{\varrho} \frac{v_{L,i}^3}{2h_R} \tag{4.5}$$

2) Power dissipation per unit volume due to friction between phases (slip power):

$$e_{slip} = -\Delta\varrho g \varepsilon_G \left(1-\varepsilon_G\right) v_{rG} \tag{4.6}$$

A combination of these terms gives the balance equation:

$$e_{thrust} + e_G + e_L + e_{slip} + e_{circ} = 0 \tag{4.7}$$

In general, the quantities $v_{L,i}$ (i.e., e_{circ}) and ε_G (i.e., e_{slip}) are unknown, so direct evaluation is impossible.

4.4. Mixing Behavior and Fluid Dynamics

The flow processes in jet loop reactors are particularly crucial for the mixing and residence-time distribution of both phases. The investigation of the relationship between mixing time and liquid circulation time showed that complete homogenization of the liquid phase requires ca. ten passes [126].

The *residence-time distribution of the liquid* in the jet loop reactor has been investigated both experimentally and theoretically [127], [128]. Values of dispersion coefficients in the liquid are presented in [118], [128]. The crucial parameter for mixing is the internal circulation flow generated by the liquid let in through the nozzle. In certain cases, this quantity can be estimated from the energy balance (Eq. 4.7).

The *residence-time distribution of the dispersed gas phase* can be found in [129]. The dispersion coefficients are not substantially different from those of the liquid phase.

Fluid Dynamics of Single-Phase Flow. The simplest formulation is that of single-phase flow. For $e_G = 0$, $e_{slip} = 0$, $e_L = 0$, the following holds for liquid circulation flow

$$v_{L,i} d_t = \sqrt{\frac{2}{\zeta f_i}} v_n d_n \tag{4.8}$$

The drag coefficient of circulation flow ζ can be found in [118], [130–132]. Typical values for bottom gas feed are $\zeta = 4$; for top gas feed, $\zeta \approx 0.25-2$.

Fluid Dynamics of Two-Phase Flow. In two-phase flow the momentum balance can be solved for the circulation flow $v_{\mathrm{L,i}}$ only if the gas holdup ε_{G} is known, which is not normally the case. Only in the gas-circulation reactor (Fig. 32 A, B) can the gas holdup be definitely specified within a certain range. Under the assumption of low gas holdup ε_{G}, for example, the following implicit relation is reported [123]:

$$(v_{\mathrm{n}} d_{\mathrm{n}})^2 = \left\{ \frac{\zeta f_{\mathrm{i}}}{2} v_{\mathrm{L,i}}^2 + \frac{\Delta\varrho}{\varrho_{\mathrm{L}}} g h_{\mathrm{R}} \frac{\varepsilon_{\mathrm{G}} (v_{\mathrm{rG}} - u_{\mathrm{L}}) - u_{\mathrm{G}}}{v_{\mathrm{L,i}}} \right\} d_{\mathrm{t}}^2 \quad (4.9)$$

The resistance coefficients ζ obtained for single-phase flow can be used to a good approximation. Figure 35 shows a logarithmic plot of Equation (4.9) for $u_{\mathrm{L}} = 0$ and $u_{\mathrm{G}} = 0$. The liquid velocities in the two-phase regime are always lower than in single-phase flow. If the velocity goes below a minimum value, the flow becomes unstable and stops, as shown by the nonlinear behavior of the curves at low velocities. Gas sparging then takes place only in the upper part of the internal tube. These features are illustrated in Figure 35. The minimum flow velocities $v_{\mathrm{L,\,i,\,min}}$ are a factor of $\sqrt{3}$ smaller than the single-phase velocities. This means that all possible velocities in two-phase flow are in the range

$$\frac{1}{\sqrt{3}} \leq \frac{v_{\mathrm{L,i}} d_{\mathrm{t}}}{\sqrt{\frac{2}{\zeta f_{\mathrm{i}}}} v_{\mathrm{n}} d_{\mathrm{n}}} \leq 1$$

The minimum flow velocity can also be obtained from Equation (4.9):

$$v_{\mathrm{L,i,min}} = \sqrt[3]{\frac{\Delta\varrho/\varrho_{\mathrm{L}}}{\zeta f_{\mathrm{i}}} g h_{\mathrm{R}} \left[\varepsilon_{\mathrm{G}} (v_{\mathrm{rG}} - u_{\mathrm{L}}) - u_{\mathrm{G}}\right]} \quad (4.10)$$

At high gas velocities u_{G} or when the nozzle is at the bottom of the reactor (Fig. 34 A), gas holdup cannot be freely selected. Instead, it adjusts itself as a function of fluid-dynamic conditions. At present, flow velocities cannot be calculated in advance.

BOHNER's measurements for the jet loop reactor are plotted in Figure 36 in the form

$$\bar{v}_{\mathrm{L,c}} d_{\mathrm{t}} = f\,(v_{\mathrm{n}} d_{\mathrm{n}}, u_{\mathrm{G}})$$

[133], [134]. Two regions can be identified. At low jet velocities v_{n}, the jet loop behaves like an airlift loop reactor (Section 2.14). The liquid circulation velocity is almost independent of jet velocity. Only at larger values does $v_{\mathrm{L,\,c}}$ increase linearly with v_{n} as in single-phase flow [133–135]. Investigations on radial velocity profiles in loop reactors are found in [135].

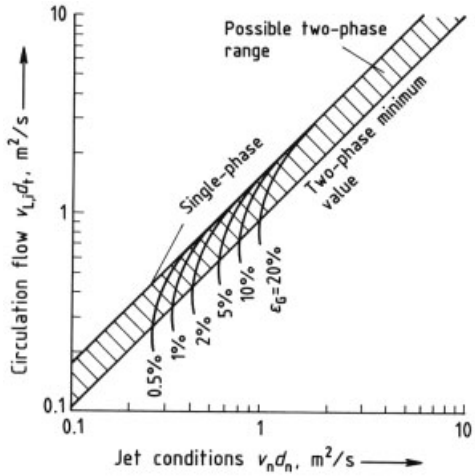

Figure 35. Circulation flow rate in gas-circulation reactor as a function of jet conditions and gas holdup, calculated with Equation (4.9)
Assumptions: $\zeta = 2.5$; $v_{\mathrm{rG}} = 0.23$ m/s; Geometric dimensions: $d_{\mathrm{t}} = 0.3$ m; $d_{\mathrm{h}} = 0.015$ m; $h_{\mathrm{t}} = 2.0$ m; $f_{\mathrm{i}} = 0.25$

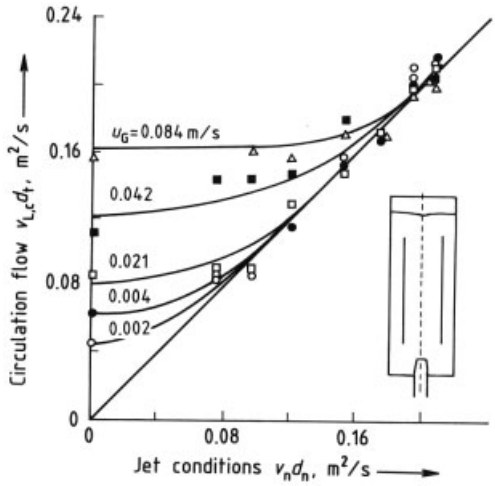

Figure 36. Measured circulation flow in pressure-sparged jet loop reactor as a function of jet conditions, with gas velocity as parameter [133]
$d_{\mathrm{t}} = 0.29$ m; $h_{\mathrm{t}} = 2.0$ m; $f_{\mathrm{i}} = 0.35$

4.5. Gas Holdup

In the *gas-circulation reactor* (Figs. 31 and 32 A and B), the gas holdup can be set arbitrarily

within certain limits. For a given jet velocity v_n, the maximum values $\varepsilon_{G,\max}$ can be calculated from the power input by the liquid jet

$$P_n = \frac{\varrho_L}{2} v_n^3 F_n \text{ where } F_n = \frac{\pi}{4} d_n^2 \tag{4.11a}$$

or the jet power per unit volume

$$e_n = \frac{P_n}{V_R} = \frac{\varrho_L}{2} \cdot \frac{v_n^3}{h_R} \left(\frac{d_n}{d_t}\right)^2 \tag{4.11b}$$

The equations reported by Tebel and Zehner are presented here in simplified form [123]. The maximum specific energy dissipation rate caused by the slip between the two phases (Eq. 4.6) is given by

$$e_{\text{slip},\max} = \frac{4}{3^{1.5}} \sqrt{\frac{2}{\zeta f_i}}\, e_n \frac{d_n}{d_t} \tag{4.12}$$

which can be solved as follows for the maximum gas holdup:

$$\varepsilon_{G,\max} = \frac{2}{3^{1.5}} \sqrt{\frac{2}{\zeta f_i}} \cdot \frac{2 e_n \frac{d_n}{d_t}}{\Delta\varrho g v_{rG}} = \frac{2}{3^{1.5}} \sqrt{\frac{2}{\zeta f_i}} \cdot \frac{2 P_n \frac{d_n}{d_t}}{\Delta\varrho g v_{rG} V_R} \tag{4.13}$$

Equation (4.13) is compared with measured values in Figure 37. The lowest gas holdups are obtained in the coalescing system water – air. Even tiny amounts of methanol reduce the tendency to coalesce. The primary gas bubbles generated by the jet retain approximately the same size over the whole apparatus because they do not coalesce. This reduces their slip velocity v_{rG} so that much higher gas holdups and interfacial areas can be obtained.

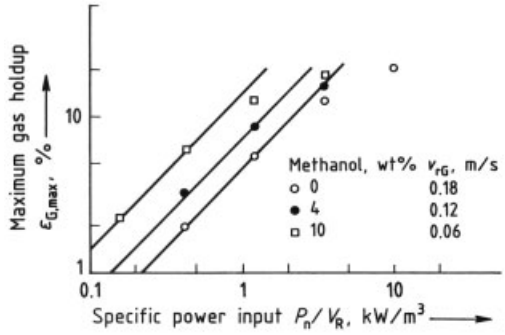

Figure 37. Comparison of measured and calculated maximum gas holdup as a function of specific jet power in gas-circulation reactor
System: water – air with added methanol; measured values after [123]; calculations with Equation (4.13)
$d_t = 0.14$ m; $h_t = 1.32$ m; $d_i = 0.055$ m; $d_n = 4.9$ mm

More complex relationships apply if pressurized gas is let in at the bottom of the gas-circulation reactor. The gas holdup then sets itself, analogously to behavior in a bubble column, an airlift loop reactor (Sections 2.9 and 2.14),

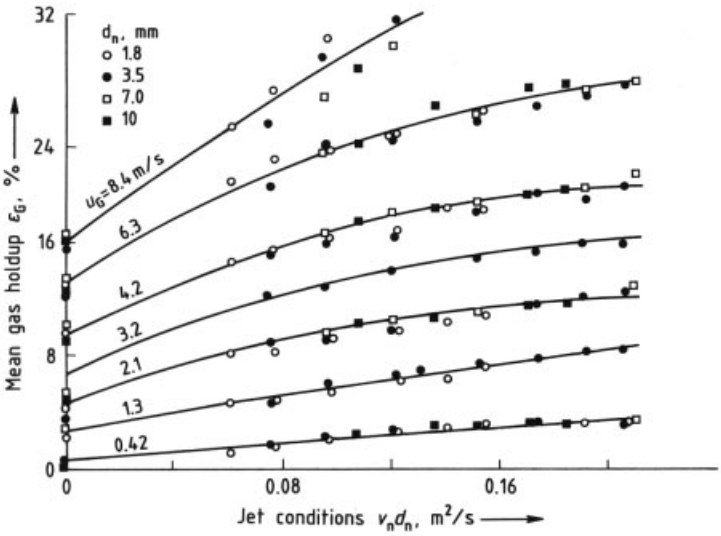

Figure 38. Gas holdup measured in the pressure-sparged jet loop reactor [133] versus jet conditions at various gas velocities
$h_t = 2$ m; $d_t = 0.29$ m; $f_i = 0.33$

or the ordinary loop reactor (Fig. 34 A). Typical measured values for this type of gas sparging are presented in Figure 38. As the gas rate increases, gas holdup increases rapidly. The jet velocity, by contrast, has relatively little effect.

ZEHNER and THELEN obtained the expression

$$\frac{e_{\mathrm{slip}}}{2e_{\mathrm{n}}\frac{d_{\mathrm{n}}}{d_{\mathrm{t}}}} = f\left(\frac{e_{\mathrm{G}}}{2e_{\mathrm{n}}\frac{d_{\mathrm{n}}}{d_{\mathrm{t}}}}\right) \tag{4.14}$$

for jet loop reactors with pressurized gas sparging [125]. Figure 39 shows the gas holdup for different reactor types and sparging types, based on this relation. To within measurement error, the same values are obtained for the gas-circulation reactor with pressurized gas sparging (Fig. 31 B) and for the jet loop reactor (Fig. 34). These values are fitted well by the correlation

$$\frac{e_{\mathrm{slip}}}{2e_{\mathrm{n}}\frac{d_{\mathrm{n}}}{d_{\mathrm{t}}}} = 1.5\left(\frac{e_{\mathrm{G}}}{2e_{\mathrm{n}}\frac{d_{\mathrm{n}}}{d_{\mathrm{t}}}}\right)^{0.8} \tag{4.13a}$$

which can be solved directly for the gas holdup:

$$\varepsilon_{\mathrm{G}} = 1.5\left(\frac{2P_{\mathrm{n}}\frac{d_{\mathrm{n}}}{d_{\mathrm{t}}}}{\Delta\varrho g v_{\mathrm{rG}} V_{\mathrm{R}}}\right)^{0.2}\left(\frac{u_{\mathrm{G}}}{v_{\mathrm{rG}}}\right)^{0.8} \tag{4.13b}$$

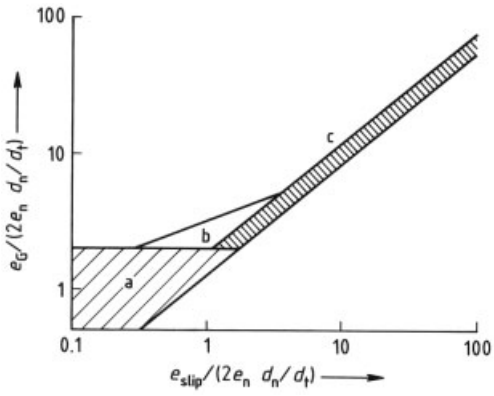

Figure 39. Relationship between slip power ϱ_{slip}, jet power ϱ_{n}, and gas compression power ϱ_{G} for determination of gas holdup
a) Surface sparging in gas-circulation reactor; b) Transition between surface and pressure sparging; c) Pressure sparging in gas-circulation reactor and normal jet loop reactor

4.6. Mass Transfer

The mean bubble diameter in a sparged reactor is always the result of distribution and coalescence processes. In the *easily coalescing water – air system*, the bubble diameter corresponds essentially to the local energy dissipation rate, a quantity that is distributed very unevenly over the volume in jet loop reactors. Zones of particularly high energy-dissipation rate include the immediate action region of the liquid jet and, to a lesser degree, the regions of loop flow reversal. This is the reason why the smallest bubbles are observed near the jet in the compact reactor (see Fig. 32 D) [120]. As Figure 40 shows, bubble size decreases from ca. 3 mm to almost 2 mm with increasing energy dissipation rate $P_{\mathrm{n}}/V_{\mathrm{R}}$. In the other much larger regions, the bubbles quickly coalesce to bigger (3 – 4 mm) units. Jet loop reactors accordingly do not feature smaller air-in-water bubbles, on average, than bubble columns or downflow bubble columns operated in the homogeneous flow regime. For a given gas holdup, comparable mass-transfer should therefore be expected. The relationships among the mean bubble diameter (Sauter diameter) d_{bS}, the gas holdup ε_{G}, the specific interfacial area a, and the volumetric mass-transfer coefficient $k_{\mathrm{L}}\,a$ have been pointed out in Section 3.3. These considerations lead to the following guideline values for the water – air system:

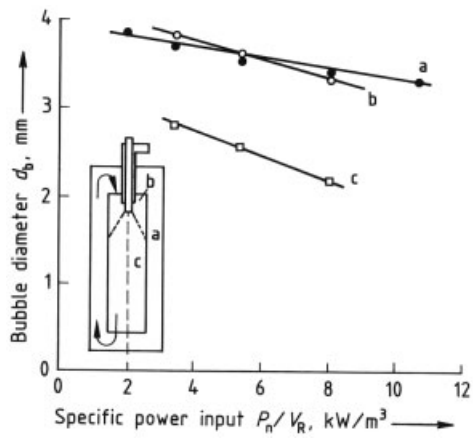

Figure 40. Bubble sizes in the compact reactor (system: water – air)
$d_{\mathrm{t}} = 0.1$ m; $h_{\mathrm{t}}/d_{\mathrm{t}} = 5.9$; $d_{\mathrm{n}} = 5$ mm; $d_{\mathrm{i}}/d_{\mathrm{t}} = 0.6$

Usual specific energy dissipation rates:
$P_{\mathrm{n}}/V_{\mathrm{R}} \approx 1 - 10\,\mathrm{kW/m^3}$
Mean bubble diameter (Sauter diameter):
$d_{\mathrm{bS}} \approx 3.5$ mm
Maximum gas holdup: $\varepsilon_{\mathrm{G}} \approx 6 - 30$ %

Special interfacial area: $a \approx 100-600\,\mathrm{m}^{-1}$
Volumetric mass-transfer coefficient: $k_L\,a \approx 0.04-0.2\,\mathrm{s}^{-1}$

If liquid mixtures and ionic or detergent solutions exhibit a *noncoalescing behavior*, much higher gas holdups and smaller bubble diameters (significantly less than 1 mm) can be achieved. The volumetric mass-transfer coefficients and specific interfacial areas behave in a similar way; in systems with hindered coalescence they may be a factor of 5 to 10 higher than the water – air values. A compilation of volumetric mass-transfer coefficients in various types of sparged apparatus can be found in [136].

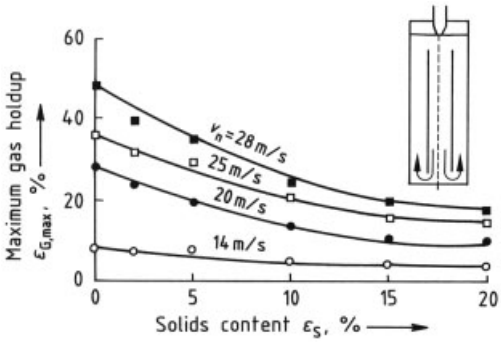

Figure 41. Effect of solids on gas holdup in the gas-circulation reactor
System: 1 wt % NaCl – air – glass spheres; $d_P = 0.075-0.15$ mm

4.7. Three-Phase Loop Reactor

When solids are suspended in sparged loop reactors the same engineering considerations are necessary as in slurry bubble columns (Section 2.13). Many workers have studied the effect of solid particles on fluid dynamics and mass-transfer performance [76], [137–139].

The fluid-dynamic principles of solid – liquid systems are comparable to those of gas – liquid systems (Sections 4.3 and 4.4) [132], [140]. These considerations have been extended to threephase systems so that theoretical models are available for complex multiphase flow [120], [123], [141]. In principle, the solid phase is accounted for by a supplemental energy term for the slip power dissipation of the particle swarm (Section 4.3). How interactions between phases influence individual slip velocities is not clear. For example, KÜRTEN and coworkers examined the maximum possible gas holdup in a gas-circulation reactor with gas sparging at the surface [76]. As Figure 41 shows, the solids content ε_S has a strong effect that cannot be accounted for merely by the additional slip power dissipation of fine particles. Instead, the slurry must be assumed to yield larger gas bubbles because of its higher apparent viscosity. The higher slip velocity of these larger bubbles might then explain the marked dependence on solids concentration.

RÄBIGER [138] and WACHSMANN [139] also studied the effect of solids. At constant energy dissipation rate and constant volumetric gas flow rate, the dependences are roughly similar to those found in [76]. Only for low solids and gas holdups does a slight increase in gas holdup occur relative to two-phase systems.

This has been explained by bubble breakup by large solid particles [142]. The effect should not occur below the critical Weber number

$$We = \frac{\varrho_S d_S v_b^2}{\sigma} = 3$$

However, this statement partially contradicts RÄBIGER's results [138]. Technically, these discrepancies are insignificant. Up to a solids concentration of 10 vol %, no major differences exist in gas holdup and mass-transfer between normal and three-phase jet loop reactors. Higher volumetric particle concentrations do not, however, normally occur in jet loop reactors.

5. References

1. H. Gerstenberg, *Chem.-Ing.-Tech.* **47** (1975) no. 5, 209.
2. M. Bobik, *Chem.-Ing.-Tech.* **53** (1981) MS 922.
3. BASF, EP 254 180, 1991 (P. Zehner et al.).
4. Hoechst, DE-OS 2 855 263, 1978 (A. Riedel et al.).
5. R. Steiner, *Chem. Eng. Process.* **21** (1987) 1 – 8.
6. W.-D. Deckwer, J. Hallensleben, M. Popovic, *Can. J. Chem. Eng.* **58** (1980) 190.
7. H. Buchholz, Dissertation, Universität Hannover 1979.
8. K. Ruff, T. Pilhofer, A. Mersmann, *Chem.-Ing.-Tech.* **48** (1976) no. 9, 759 – 764.
9. G. Neubauer, T. Pilhofer, *Chem.-Ing.-Tech.* **50** (1978) no. 2, 115 – 116.

10. Mersmann, A., *Ger. Chem. Eng. (Engl. Transl.)* **1** (1978) 1.
11. P. Krötsch, *Fortschr. Verfahrenstech. Abt. D* **16** (1978) 389.
12. P. Zehner, *Chem.-Ing.-Tech.* **47** (1975) no. 5, 209.
13. H. Gerstenberg, *Chem.-Ing.-Tech.* **51** (1979) no. 3, 208 – 216.
14. Y. T. Shah, B. G. Kelkar, S. P. Godbole, W. D. Deckwer, *AIChE J.* **28** (1982) 353 – 379.
15. J. B. Joshi, Y. T. Shah, *Chem. Eng. Commun.* **11** (1981) 165 – 199.
16. K. Rietema, S. P. P. Ottengraph, *Trans. Inst. Chem. Eng.* **48** (1970) T 54.
17. J. R. Crabtree, J. Bridgwater, *Chem. Eng. Sci.* **24** (1969) 1755.
18. S. M. Bhavraju, T. W. F. Russel, H. W. Blanch, *AIChE J.* **24** (1978) 454.
19. K. Ueyama, T. Miyauchi, *AIChE J.* **25** (1979) 258 – 266.
20. J. B. Joshi, M. M. Sharma, *Trans. Inst. Chem. Eng.* **57** (1979) 244.
21. P. Zehner, *Ist. Chem. Eng.* **26** (1986) no. 1, 22 – 35.
22. A. Lübbert: "Mass Transfer with Chemical Reactions in Multiphase Systems," NATO Advanced Study Institute, Cesme/Izmir, Aug. 10 – 11, 1981.
23. J. H. Hills, *Trans. Inst. Chem. Eng.* **52** (1974) 1 – 9.
24. E. Kojima et al., *J. Chem. Eng. Jpn.* **13** (1980) 16.
25. M. Kraume, P. Zehner, *German/Japanese Symposium Bubble Columns,* Schwerte 1988, preprints.
26. R. Buchholz, K. Schügerl, *Eur. J. Appl. Microbiol. Biotech.* **6** (1979) 301 – 315.
27. E. Blaß, *Chem.-Ing.-Tech.* **60** (1988) no. 12, 935 – 947.
28. P. H. Calderbank, *Chem. Eng.* **45** (1976) CE 209.
29. O. Nagel, B. Hegner, H. Kürten, *Chem.-Ing.-Tech.* **50** (1978) no. 12, 934 – 944.
30. H. Unno, I. Inoue, *Chem. Eng. Sci.* **35** (1980) 1571.
31. K. Akita, F. Yoshida, *Ind. Eng. Chem. Proc. Des. Dev.* **13** (1974) 84 – 91.
32. T. Miyahara, Y. Matsuba, T. Takahashi, *Int. Chem. Eng.* **23** (1983) 517 – 523.
33. T. Miyahara, N. Naga, T. Takahashi, *Int. Chem. Eng.* **23** (1983) 524 – 531.
34. K. Koide et al., *Chem. Eng. Jpn.* **12** (1979) 98 – 104.
35. A. Mersmann, *Chem.-Ing.-Tech.* **61** (1989), no. 2, 97 – 104.
36. H. Wezorke, Dissertation, Universität Dortmund 1986.
37. R. Beinhauer, Dissertation, TU Berlin 1971.
38. Y. T. Shah, G. J. Stiegel, M. M. Sharma, *AIChE J.* **24** (1978) 369 – 400.
39. M. Kraume, P. Zehner, *Chem.-Ing.-Tech.* **61** (1989) no. 4, 332 – 333.
40. K.-H. Mangartz, T. Pilhofer, *VT, Verfahrenstech.* **14** (1980) no. 1, 40 – 44.
41. P. Zehner, G. Schuch, *Ger. Chem. Eng. (Engl. Transl.)* **8** (1985) 282 – 289.
42. G. Blasey, Dissertation, Universität Dortmund 1987.
43. P. Herbrechtsmeier, R. Steiner, *Chem.-Ing.-Tech.* **52** (1980) 468.
44. W. Hikita et al., *Chem. Eng. J.* **20** (1980) 59 – 67.
45. R. Nottenkämper, A. Steiff, P. M. Weinspach, *Ger. Chem. Eng.* **6** (1983) 147 – 155.
46. H. F. Bach, T. Pilhofer, *Ger. Chem. Eng. (Engl. Transl.)* **1** (1978) 270. H. Buchholz, R. Buchholz, H. Niebeschütz, K. Schügert, *Eur. J. App. Microbiol. Biotechnol.* **6** (1978) 115.
47. K. Schügerl, J. Lücke, U. Oels, *Adv. Biochem. Eng.* **7** (1977) 1. G. Keitel, U. Onken, *Chem.-Ing.-Tech.* **54** (1982) no. 3, 262 – 263.
48. K. Akida, F. Yoshida, *Ind. Eng. Chem. Proc. Des. Dev.* **13** (1974) 84 – 91.
49. H. Kölbel, E. Borchers, H. Langemann, *Chem.-Ing.-Tech.* **33** (1961) no. 10, 668 – 675. W.-D. Deckwer, Y. Luisi, A. Zaidi, M. Ralek, *Ind. Eng. Chem. Proc. Des. Dev.* **10** (1980) 699.
50. M. H. Oyevaar, K. R. Westerterp, *Chem. Eng. Process.* **25** (1989) 85 – 98.
51. M. H. Oyevaar, R. Bos, K. R. Westerterp, *Chem. Eng. Sci.* **46** (1991) 1217 – 1231.
52. P. M. Wilkinson, L. v. Dierendonck, *Chem. Eng. Sci.* **45** (1990) 2309 – 2315.
53. A. Serizawa, I. Kataoka, I. Michiyoshi, *Int. J. Multiphase Flow* **2** (1975) 235 – 246.
54. W.-D. Deckwer, *Reaktionstechnik in Blasensäulen,* Salle und Sauerländer, Aarau 1985.
55. U. Oels, J. Lücke, R. Buchholz, K. Schügerl, *Ger. Chem. Eng. (Engl. Transl.)* **1** (1978) 115.
56. W.-D. Deckwer, R. Burchhart, G. Zoll, *Chem. Eng. Sci.* **29** (1974) 2177.
57. Y. Kawase, B. Halard, M. Moo-Young, *Chem. Eng. Sci.* **42** (1987) 1609 – 1617.
58. H. Hikita et al., *Chem. Eng. J.* **22** (1981) 61 – 69.
59. J. S. Cho, N. Wakao, *J. Chem. Eng. Jpn.* **21** (1980) 576 – 581.

60. S. P. Godbole, A. Schumpe, Y. T. Shah, N. L. Carr, *AIChE J.* **30** (1984) 213 – 220.
61. K. Akita, *Int. Chem. Eng.* **29** (1989) 127 – 135.
62. W. Kast, *Int. J. Heat Mass Transfer* **5** (1962) 329 –336.
63. W.-D. Deckwer, *Chem. Eng. Sci.* **35** (1980) 1341 –1346.
64. J. B. Joshi et al., *Chem. Eng. Commun.* **6** (1980) 257 –271.
65. A. N. Khoze et al., *J. Eng. Phys.* **20** (1971) 759 –763.
66. L. S. Aksel'rod, N. I. Vorotnikova, A. A. Kozlov, *Heat Transfer-Sov. Res.* **8** (1976) 25 – 33.
67. R. Wendt, Dissertation, Universität Dortmund 1983.
68. H.-J. Korte, Dissertation, Universität Dortmund 1987.
69. H.-J. Korte, A. Steiff, P.-M. Weinspach, *German/Japanese Symposium Bubble Columns,* Schwerte 1988, preprints.
70. R. Wendt, A. Steiff, P.-M. Weinspach, *Chem.-Ing.-Tech.* **55** (1983) no. 10, 796 – 797.
71. Y. T. Shah, C. A. Ratway, H. G. McIlvried, *Trans. Inst. Chem. Eng.* **56** (1978) 107 – 112.
72. K. Bernemann, Dissertation, Universität Dortmund 1989.
73. K. Koide, T. Yasuda, S. Iwamoto, E. Fukuda, *J. Chem. Eng. Jpn.* **16** (1983) 7 – 12.
74. S. Kara et al., *Ind. Eng. Chem. Proc. Des. Dev.* **21** (1982) 584 – 594.
75. Y. Kato et al., *Ist. Chem. Eng.* **13** (1979) 562 – 567.
76. H. Kürten, P. Zehner, *Ger. Chem. Eng. (Engl. Transl.)* **2** (1979) 220 – 227.
77. T. Sauer, Dissertation, Universität Paderborn 1986.
78. A. Yasunishi, M. Fukuma, K. Muroyama, *J. Chem. Eng. Jpn.* **19** (1986) 444 – 449.
79. K. Koide, A. Takazawa, M. Komura, H. Matsunaga, *J. Chem. Eng. Jpn.* **17** (1984) 459 – 466.
80. T. Matsumoto, N. Hidaka, S. Morooka, *AIChE J.* **35** (1989) 1701 – 1709.
81. H. Mürtz, Dissertation, Universität Dortmund 1990.
82. M. Fukuma, K. Muroyama, A. Yasunishi, *J. Chem. Eng. Jpn.* **20** (1987) 28.
83. E. Sada, H. Kumazawa, C.-H. Lee, H. Nakamura, *Ind. Eng. Chem. Proc. Des. Dev.* **26** (1987) 112.
84. K. Nguen-Tien, H. Patwari, A. Schumpe, W.-D. Deckwer, *AIChE J.* **31** (1985) 194.
85. H.-G. Springmann, A. Steiff, P.-M. Weinspach, *Chem.-Ing.-Tech.* **63** (1991) 162 – 163.
86. J. Kratochvil, K. Winkler, J. Zahradnik: *Coll. Czech. Chem. Commun.* **50** (1985) 48 – 60.
87. K. Muroyama, *German/Japanese Symposium Bubble Columns,* Schwerte 1988, preprints.
88. "Cyclohexane," *Hydrocarbon Processing* 1977, no. 11, 144.
89. U. Onken, P. Weiland, *Adv. Biotechnol. Processes* **1** (1983) 67 – 95.
90. M. Y. Chisti, *Airlift Bioreactors,* Elsevier Applied Science, London 1989.
91. P. Verlaan, J. C. Vos, K. van't Riet: *J. Chem. Technl. Biotechnol.* **45** (1989) 109 – 121.
92. R. A. Bello, C. W. Robinson, M. Moo-Young, *Can. J. Chem. Eng.* **40** (1984) 573 – 577.
93. P. Weiland, U. Onken, *Chem.-Ing.-Tech.* **52** (1980) 986 – 987.
94. P. Weiland, *Ger. Chem. Eng. (Engl. Transl.)* **7** (1984) 374 – 385.
95. Y. Kawase, M. Moo-Young, *Chem. Eng. Commun.* **40** (1986) 67 – 83.
96. T. Reith, Ph. D. Thesis, Universität Delft 1968.
97. W. Hirner, H. Blenke, *VT, Verfahrenstech.* **11** (1977) no. 5, 297 – 303.
98. K. Koide, H. Sato, S. Iwamoto, *J. Chem. Eng. Jpn.* **16** (1983) 407 – 413.
99. K. Koide et al., *J. Chem. Eng. Jpn.* **16** (1983) 413 – 419.
100. K. Nakao et al., *German/Japanese Symposium Bubble Columns,* Schwerte 1988, preprints.
101. K. Schügerl, *Bioreaktionstechnik,* vol. 2, Salle und Sauerländer, Aarau 1991.
102. P. K. Ouyoung, M. Y. Chisti, M. Moo-Young, *Chem. Eng. Res. Des.* **67** (1989) 451 – 456.
103. G. Müller, G. Sell, *DECHEMA-Monogr.* **86** (1980) 589.
104. S. J. Heijnen et al., *Chem. Eng. Technol.* **13** (1990) 202 – 208.
105. Y. Bando et al., *J. Chem. Eng. Jpn.* **21** (1988) no. 6, 607 – 612.
106. A. Ohkawa, Y. Shiokawa, N. Sakai, *J. Chem. Eng. Jpn.* **18** (1985) no. 6, 172 – 174.
107. A. Ohkawa, Y. Shiokawa, N. Sakai, N. Imai, *J. Chem. Eng. Jpn.* **18** (1985) no. 6, 466 – 469.
108. P. Herbrechtsmeier, H. Schäfer, R. Steiner, *Chem.-Ing.-Tech.* **56** (1984) MS 1221.
109. P. Herbrechtsmeier, H. Schäfer, R. Steiner, *Chem.-Ing.-Tech.* **53** (1981) MS 884.
110. P. Herbrechtsmeier, R. Steiner, *Chem.-Ing.-Tech.* **50** (1978) no. 12, 944 – 947.
111. A. Kulkarni, Y. T. Shah, *Chem. Eng. Commun.* **28** (1984) 311 – 326.
112. A. Ohkawa et al., *J. Chem. Eng. Jpn.* **20** (1987) no. 1, 99 – 101.

113. W. J. Albrecht, G. Schulze-Pillot, *Chem.-Ing.-Tech.* **50** (1987) no. 22.
114. R. Steiner, *Chem. Eng. Process.* **21** (1987) 1 – 8.
115. R. Steiner, *DECHEMA-Monogr.* **94** (1983) 164 – 197.
116. H. Kürten, P. Zehner, *Annual Conference of Chemical Engineering,* Straßburg, Oct. 1980.
117. BASF, EP 0 087 670, 1983 (H. H. Schuster, H. Dreher, J. Hambrecht).
118. H. Blenke in: Loop Reactors, Advances in Biochemical Engineering 13, Springer-Verlag, Berlin 1979.
119. N. Räbiger, A. Vogelpohl, *Chem.-Ing.-Tech.* **53** (1983) MS 1111.
120. N. Räbiger, Habilitationsschrift, Technische Universität, Clausthal 1988; *Praxiswissen thermische Verfahrenstechnik,* Verlag TÜV Rheinland, Köln 1988.
121. H. Loosen, *ZFL* **40** (1989) no. 9, 494 – 496.
122. B. Lohrengel, A. Vogelpohl, *Chem.-Ing.-Tech.* **62** (1990) MS 1856.
123. K. H. Tebel, P. Zehner, *Chem. Eng. Technol.* **12** (1989) 274 – 280.
124. H. Blenke, K. Bohner, W. Hirner, *Verfahrenstechnik (Mainz)* **3** (1969) no. 10, 444 – 452.
125. P. Zehner, H.-G. Thelen: "Mehrphasenströmung," Vortrag auf dem GVC-Fachausschuß, Bremen 1991.
126. A. Jaramillo, I. Adler, A. Fiechter, *Third European Congress on Biotechnology,* Munich, Sept. 10 – 14, 1984.
127. J. Lehnert, *Verfahrenstechnik (Mainz)* **6** (1972) 58.
128. H.-J. Warnecke, M. Weidenbach, J. Prüss, H. Langemann, *Chem.-Ing.-Tech.* **59** (1987) no. 6, 496 –499.
129. H.-J. Warnecke, D. Vaupel, I. Prüss, H. Langemann, *Chem.-Ing.-Tech.* **61** (1989 no. 10, 815 – 817.
130. H. Blenke, K. Bohner, S. Schuster, *Chem.-Ing.-Tech.* **37** (1965) no. 3, 289 – 294.
131. H. Blenke, K. Bohner, W. Pfeifer, *Chem.-Ing.-Tech.* **43** (1971) no. 1 – 2, 10 – 17.
132. P. Zehner, Habilitationsschrift, Universität Kaiserslautern 1988; *Fortschr.-Ber. VDI Z. Reihe 3,* 1988, no. 160.
133. K. Bohner, Dissertation, Universität Stuttgart 1971.
134. K. Bohner, H. Blenke, *Verfahrenstechnik (Mainz)* **6** (1972) no. 2, 50 – 57.
135. D. K. Jain et al., *Can. J. Chem. Eng.* **68** (1990) 1047 –1051.
136. H.-J. Warnecke, M. Geisendörfer, D. Ch. Hempel, *Chem. Eng. Technol.* **11** (1988) 306 – 311.
137. L.-S. Fan, S.-J. Hwang, A. Masuura, *Chem. Eng. Sci.* **39** (1984) no. 12, 1677 – 1688.
138. N. Räbiger, *Chem.-Ing.-Tech.* **57** (1985) no. 3, 248 –249.
139. U. Wachsmann, N. Räbiger, A. Vogelpohl, *VDI-Ber.* **607** (1986) 1537 – 1566. U. Wachsmann, N. Räbiger, A. Vogelpohl, *Ger. Chem. Eng. (Engl. Transl.)* **8** (1985) 411 – 418.
140. P. Zehner, *Chem.-Ing.-Tech.* **62** (1980) MS 850.
141. K. H. Tebel, P. Zehner, *Annual Meeting of the Process Engineers,* Straßburg, Sept 17 – 19, 1986.
142. J. C. Lee, *4th Proc. European Symp., Chem. React. Eng.,* 1971, p. 27.

Bubble Memories → Information Storage Materials
Buffing → Abrasives
Building Bricks → Construction Ceramics

Three-Phase Trickle-Bed Reactors

For other industrial reactors and their applications, see → Stirred-Tanc and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Reaction Columns, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

K. Roel Westerterp, Chemical Reaction Engineering Laboratories, Twente University of Technology, Enschede, Netherlands

Wino J. A. Wammes, Chemical Reaction Engineering Laboratories, Twente University of Technology, Enschede, Netherlands

1. Introduction 2
2. Design 3
2.1. Flow Regimes 3
2.2. Liquid Holdup 5
2.3. Pressure Drop 6
2.4. Mass Transfer 7
2.4.1. Gas-Side Mass Transfer 7
2.4.2. Liquid-Side Mass Transfer 8
2.4.3. Liquid – Solid Mass Transfer 8
2.5. Heat Transfer 9
2.6. Wetting Efficiency 9
3. Scale-Up 10
4. Industrial Applications 12
5. References 12

Symbols (see also → Principles of Chemical Reaction Engineeringand → Model Reactors and Their Design Equations)

A_e	effective area
A_{gl}	specific gas – liquid interfacial area
A_s	specific external packing area
C	concentration, kmol/m^3
d_p	particle diameter, m
D	reactor diameter, m; molecular diffusivity coefficient
$Eö$	Eötvös number
f	friction factor
g	gravitational acceleration, m/s^2
G	mass flux of gas, kg m^{-2} s^{-1}
Ga	Galileo number
$Ga*$	modified Galileo number
H	height, m
L	mass flux of liquid, kg m^{-2} s^{-1} ; length, m
ΔP	pressure difference
R	reaction rate, mol s^{-1} m^{-3}
Re	Reynolds number
Sc	Schmidt number
Sh	Sherwood number
U	real velocity, m/s
V	reactor volume, m^3 ; superficial velocity; m/s
X	Lockhart – Martinelli parameter, $(\Delta P_l/\Delta P_g)^{1/2}$
β	liquid holdup
ε	porosity
λ	empirical parameter
η	viscosity
ϱ	density, kg/m^3
σ	surface tension, N/m
τ	mean residence time, s
φ	fraction of effective area; dimensionless two-phase pressure drop, $(\Delta P_{lg}/\Delta P_l)^{1/2}$
ψ	empirical parameter

Subscripts

dyn	dynamic
g	gas
l	liquid
res	residual
tot	total

1. Introduction

In a gas – liquid trickle-bed reactor the liquid phase always flows downward through a packed bed, whereas the gas phase can flow cocurrently or countercurrently to the liquid. Countercurrent operation is often preferred in cases where a high driving force is required for the gas – liquid mass transfer, as in distillation or physical absorption processes. From a hydrodynamic

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_309

point of view, this way of operating is disadvantageous because the capacity of the reactor (column) is limited by flooding. For processes in which the transferred component is removed by irreversible reactions, as in chemical absorption and heterogeneously catalyzed gas – liquid reactions, an operation with cocurrent downflow of both phases is more advantageous. In cocurrent downflow operation no limitations exist in the gas and liquid throughputs. In addition, at any given gas and liquid flow rates the pressure drop over the packed bed is lower for cocurrent than for countercurrent flow. Therefore, if a trickle-bed reactor is used to carry out reactions between gas and liquid components, cocurrent flow is usually employed. Some examples of large industrial applications are hydrodesulfurization and catalytic hydrodenitrification of crude oil fractions (→ Oil Refining, Chap. 3.3.), oxidative treatment of wastewater, and the historically important synthesis of 2-butyne-1,4-diol from acetylene and formaldehyde (see Chap. 4; → Butanediols, Butenediol, and Butynediol, Chap. 1.1.).

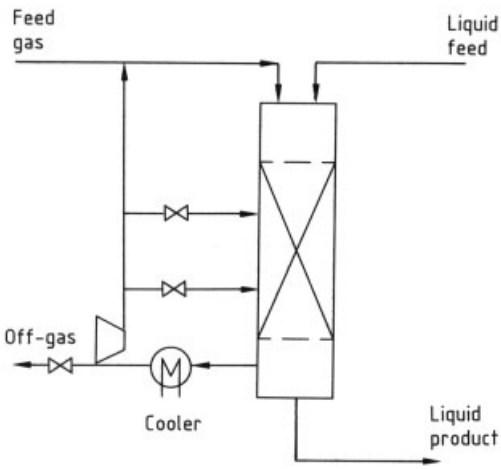

Figure 1. Cocurrent trickle-flow operation with gas recycling and interstage cooling

The operation of commercial trickle-bed reactors is essentially adiabatic. In many cases the amount of gaseous reactant fed to the reactor far exceeds the stoichiometric amount. Subsequent to gas – liquid separation in the bottom of the reactor, the major part of the gas phase is recycled to the entrance after cooling. This excess of gas is applied for example, for kinetic reasons in hydrocracking processes; for hydrodynamic reasons, since a better liquid distribution or higher gas – liquid interfacial area may be achieved; or for thermal reasons, especially to control the temperature rise over the reactor.

Figure 1 shows a typical cocurrent trickle-bed reactor. The superficial gas and liquid velocities, based on an empty reactor cross section, are typically up to 0.4 m/s and 8 mm/s, respectively. To increase the concentration of gaseous reactant in the liquid phase, nearly all trickle-flow processes are performed at elevated pressures up to ca. 20 – 30 MPa.

The height of packing ranges typically from 5 to 30 m. Packings are arranged in one or multistage beds. The height-to-diameter ratio of the reactor H/D, is generally between 5 and 25. The maximal reactor diameter is limited by the ability to distribute the liquid phase at the top of the bed evenly over the cross-sectional area; the cross-sectional area does not normally exceed 4 – 5 m. The catalyst packing consists of cylinders, spheres, extrudates, needles, or beads. This results in an overall bed porosity in the range 0.35 –0.40. The nominal size of the particles is of the order of 1 – 5 mm.

2. Design

In the trickle-bed reactor the conversion of reactants, the yield, and the selectivity toward desired products depend not only on reaction kinetics, operating pressure, and temperature, but also on the hydrodynamics of the packed bed. The gas and liquid flow on a pellet scale is illustrated in Figure 2. The flow of the gas and liquid phases through the voids between the catalyst particles is very complex and difficult to describe on a microscale. Usually the hydrodynamics are described by means of global parameters. The most important parameter for characterizing flow behavior is the flow regime in the packed bed. Other parameters that characterize the hydrodynamic situation are the liquid holdup, the pressure drop in the bed, the dispersion of gaseous and liquid phases, the catalyst wetting, and the mass- and heat-transfer coefficients. These parameters depend on the gas and liquid throughputs, the type of packing, and the physical properties of the three phases.

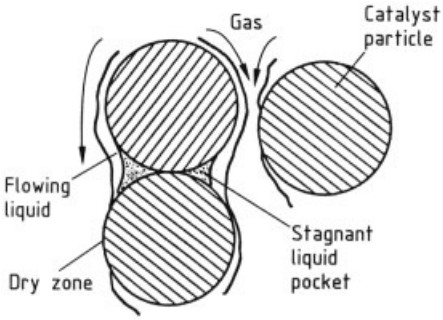

Figure 2. Flow pattern of gas and liquid phases

Much research on a laboratory scale has been published concerning these factors. Reviews on the hydrodynamics of trickle-bed reactors are in [1–3]. Laboratory reactors usually have a diameter between 0.03 and 0.2 m, whereas their length can be up to several meters. Near the reactor wall, the structure of the packing differs from the core of the packed bed; the porosity is higher at the wall. Hence, the gas – liquid flow behavior in the wall region is different. In industrial reactors where the ratio of reactor-to-particle diameter D/d_p is large–typically between 100 and 1000 – the influences of wall flow deviations on overall flow behavior can be neglected. A rule of thumb is that for ratios higher than 15, the wall effect is negligible. The D/d_p ratio has a significant influence at values lower than 10 [4]. Most laboratory experiments on hydrodynamics of trickle-bed reactors have been performed under isothermal conditions at room temperature. This is in contrast to industrial conditions where reactions are performed adiabatically and at higher temperature. Nearly all laboratory studies have been conducted under atmospheric pressure with gas density $< 2\,\mathrm{kg/m^3}$, whereas most industrial processes are operated at much higher density. Studies at higher pressure and higher gas density have been performed by only a few groups [5–7]. Therefore, care must be taken in applying design equations derived from laboratory-scale trickle-bed reactors under atmospheric pressure to typical operating conditions in industrial columns.

2.1. Flow Regimes

In a trickle-bed reactor, four different flow regimes can be distinguished. At low liquid and gas flow rates, the liquid trickles over the packing in the form of a thin film with a thickness of the order of one-tenth of a millimeter, and the gas phase is the continuous phase filling the voids in the packing. This regime is generally called *trickle flow or gas-continuous flow*. The flow behavior is very similar to that encountered in countercurrent trickle-bed operation at low gas flow rates below the loading point (i.e., the operating point at which the pressure drop becomes proportional to a gas flow rate power higher than two). At low liquid and high gas flow rates, the *spray- or mistflow-regime* appears. The liquid film is broken into fine droplets entrained by the gas phase. The *pulse-flow regime* can be obtained at higher liquid and high gas loads. The liquid periodically blocks the small channels between catalyst particles and forms liquid plugs that prevent the flow of gas. Above the liquid plugs, the local gas pressure increases rapidly until it becomes so high that the liquid plugs break and the liquid is pushed downward through the column. The velocity of the pulses can reach 1 m/s. In laboratory-scale columns, the pulses span the entire cross-sectional area of the packed bed.

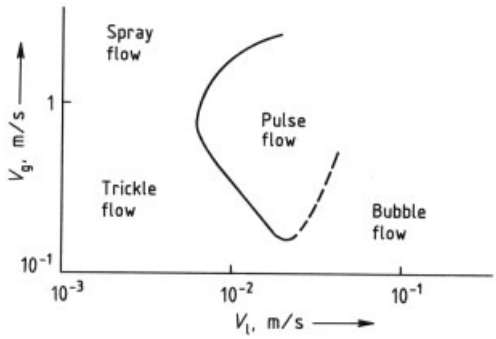

Figure 3. Different flow regimes as a function of superficial velocities of the gas and liquid phases for a given three-phase system

Further increase of the liquid flow rate results in a higher frequency of the pulses. Above a frequency of 8 – 10 Hz, individual pulses cannot be distinguished any more; the liquid holdup is relatively large; and the liquid then becomes the continuous phase, whereas the gas flows as a *uniform bubble dispersion* downward through the column. A more detailed description of the flow regimes is given in [8]. Extensive experimental study of the pulse-flow regime has been

carried out [9–11]. Figure 3 shows the different flow regimes plotted as a function of the superficial gas and liquid velocities V_g and V_l, respectively. The curve indicating the transition region between spray/trickle flow and pulse flow is typically C-shaped. On an industrial scale, trickle-bed reactors normally operate in the trickle-flow regime. Processes operating at higher flow rates, such as commericial hydrodesulfurization of light petroleum feed stocks (boiling range ca. 80 – 350 °C), may also operate in the pulse-flow regime [1].

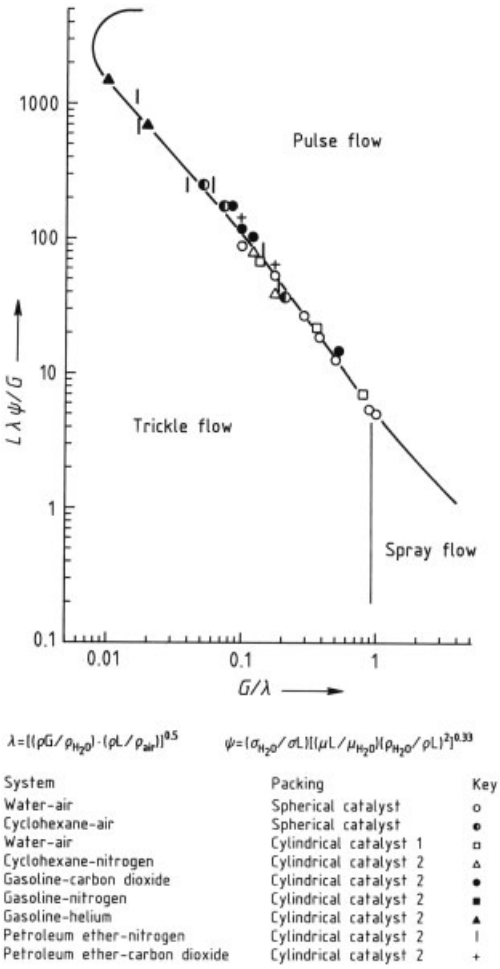

Figure 4. Flow pattern diagram for nonfoaming liquids [18]

In reactor design, knowledge of the flow regime in which the reactor will operate is very important because other hydrodynamic parameters, especially the mass-transfer rates, are affected differently in each regime. For example, the conversion is sensitive to the type of the flow regime in processes that are carried out under mass-transfer rate-limited conditions. The gas and liquid velocities at which transitions between different flow regimes occur depend on the physical properties of the liquid and the geometry and wettability of the solid phase [12], [13]. Wammes et al. found that for $\varrho_g > 2.5\,\mathrm{kg/m^3}$, the gas density also influences the flow regime transition [14]. The mechanisms governing the transitions are not well understood, so theoretical prediction of transition conditions is difficult. The first steps toward a more fundamental approach to the transition between trickle and pulse flow have been taken by Grosser et al. [15] and Danckworth et al. [16]. At present, a convenient method for predicting the flow regime is to make use of generalized flow maps. Several flow maps, all based on (semi-)empirical considerations, have been presented in the literature [4], [17], [18]. The flow diagram of Char-pentier and Favier [18], shown in Figure 4, seems to be the most accepted one: it is based on various three-phase systems and has been verified by other investigators (e.g., [19]).

The variables G and L are the mass fluxes in $\mathrm{kg\,m^{-2}\,s^{-1}}$ of the gas and the liquid, respectively. The variables λ and ψ are empirical parameters, originally proposed by Baker [20], which take the physical properties surface tension σ, liquid-phase viscosity η, and densities of the gas ϱ_g and liquid phase ϱ_l into account with respect to the properties of air and water. The type of packing is not explicitly accounted for.Gianetto et al. modified this flow map by using $G/(\lambda\,\varepsilon)$ as ordinate with the variable ε as the overall porosity of the bed [2]. Wammes et al. have found good agreement with Figure 4 for gas densities $< 2.5\,\mathrm{kg/m^3}$ [14]. However, the diagram did not predict that the liquid flow rate has to be increased for higher gas densities in order to obtain pulses. This means that high-pressure trickle-bed reactors, which operate in the pulse-flow regime according to Figure 4, may operate in reality in trickle flow. Christensen et al. executed a study in a packed bed of a rectangular cross section with a width of 0.46 m [11]. They concluded that operating conditions at the transition between trickle flow and pulse flow are well described by the flow diagram shown in

Figure 4. However, the pulses occurred locally and could only span the entire cross section at much higher gas and liquid flow rates.

2.2. Liquid Holdup

In design, the liquid holdup inside the packed bed is important because the mean residence time τ of the liquid in the reactor, and hence the conversion achieved, depends on it. If β_{tot} is the total liquid holdup in the void volume $\varepsilon_{\mathrm{bed}} V_{\mathrm{reactor}}$, the residence time is given by $\tau = \varepsilon_{\mathrm{bed}} \beta_{\mathrm{tot}} H/V_{\mathrm{l}}$. The total liquid holdup can be divided into an *internal holdup* inside the porous catalyst particles and an *external holdup* β_{ext} occupying part of the void volume between the particles. For hydrodynamics, the external part is important because the wetting of the outer surface of the catalyst and the pressure drop are related to it. As the external holdup increases, the real gas velocity in the interstices between the catalyst particles increases and with that the pressure gradient. Under stationary operating conditions, the external holdup is determined by the balance of the resist forces on the liquid (friction with the packing and capillary forces) and the driving forces (gravity and pressure gradient). Because of the complex hydrodynamics, these different forces are difficult to describe on the basis of first principles. Therefore the holdup is generally modeled by means of empirical or semiempirical equations.

The external liquid holdup can be divided into a static or *residual part* and a *dynamic part*. The latter, β_{dyn}, is the amount of liquid that leaks out of the column after the gas and liquid flows have stopped. The residual part, β_{res}, remains around the catalyst contacting points due to capillary forces and is independent of previous operating conditions.

During operation, each contact point is also surrounded by a more or less stagnant, nonactive part of liquid. However, the amount depends on operating conditions, especially liquid flow rate; see [21] and Section 2.4.3. Nevertheless, β_{res} can be taken as a first estimation for the nonactive part during operation.

For the *residual holdup*, a successful graphical relation has been developed by Charpentier et al. by correlating $\varepsilon_{\mathrm{bed}} \cdot \beta_{\mathrm{res}}$ as a function of the Eötvös number $Eö = \varrho_{\mathrm{l}}\, g\, d_{\mathrm{p}}^2/\sigma_{\mathrm{l}}$, which represents the ratio of the gravitational to the capillary forces. Up to Eötvös numbers of ca. 10, which applies to the usual catalyst packings, the residual holdup $\varepsilon_{\mathrm{bed}} \cdot \beta_{\mathrm{res}}$ amounts to 0.05.

Two approaches can be found to describe *dynamic liquid holdup :*

1) The dynamic holdup is based on the gas-phase pressure drop $(\Delta P/L)_{\mathrm{g}}$, which must be determined when only gas is flowing through the bed, and on the liquid-phase pressure drop $(\Delta P/L)_{\mathrm{l}}$, which must be determined with only liquid flowing through a bed that is completely filled with liquid. In this case the holdup is described as a function of the so-called Lockhart – Martinelli parameter $X = (\Delta P_{\mathrm{l}}/\Delta P_{\mathrm{g}})^{1/2}$: $\beta_{\mathrm{dyn}} = f\,(X)$.
   The most widely used correlation to estimate the pressure drop for one-phase flow in packed beds, is the Ergun equation [23] (→ Fixed-Bed Reactors, Chap. 2.2.). For an accurate prediction of X, the constants in the Ergun equation need to be determined experimentally for each specific packed bed. Several different functions between β_{dyn} and X have been proposed. A survey is given in [12].
2) The dynamic holdup is described on the basis of dimensionless numbers. In the case of trickle flow without gas flow (i.e., a freely trickling liquid), the holdup has often been described successfully as proportional to the product $Re_{\mathrm{l}}^{\mathrm{a}}\, Ga_{\mathrm{l}}^{\mathrm{b}}\, P^{\mathrm{c}}$, where Re_{l} and Ga_{l} are the Reynolds and Galileo numbers of the liquid $Re = \varrho_{\mathrm{l}}\, V_{\mathrm{l}}\, d_{\mathrm{p}}/\eta_{\mathrm{l}}$, and $Ga_{\mathrm{l}} = d_{\mathrm{p}}^3\, \varrho_{\mathrm{l}}^2\, g/\eta_{\mathrm{l}}^2$ (e.g., see [24]). The parameter P describes packed-bed properties.

In the case of both gas and liquid flow, liquid holdup decreases. A pressure gradient $\Delta P/L$ over the packing arises as a consequence of gas flow. The gradient is an additional driving force acting on the liquid phase in addition to the constant gravitational force; therefore, the holdup decreases, especially at higher gas velocity and density. To take the influence of gas flow into account, the pressure drop can be added to the gravitational force in the form of a modified Galileo number:

$$Ga^* = d_{\mathrm{p}}^3\, \varrho_{\mathrm{l}}^2\, g/\eta_{\mathrm{l}}^2\, [1 + \Delta P/(\varrho_{\mathrm{l}}\, g L)]$$

It has been verified experimentally that for $\Delta P/(\varrho_{\mathrm{l}}\, g L) \leq 0.1$, (i.e., for low gas density and

velocity), the influence of the gas flow on the holdup is negligible. For values higher than ca. 0.25 (e.g., at high pressure), the holdup decreases significantly [25]. The following dimensionless power product seems to be well suited to determine the dynamic holdup in the trickle-flow regime:

$$\beta_{\mathrm{dyn}} = 3.8 \left(\frac{\varrho_{\mathrm{l}} V_{\mathrm{l}} d_{\mathrm{p}}}{\eta_{\mathrm{l}}}\right)^{0.55} \cdot \left[\frac{d_{\mathrm{p}}^3 \varrho_{\mathrm{l}}^2 g}{\eta_{\mathrm{l}}^2}\left(1+\frac{\Delta P}{\varrho_{\mathrm{l}} g L}\right)\right]^{-0.42} \left(\frac{A_{\mathrm{s}} d_{\mathrm{p}}}{e_{\mathrm{bed}}}\right)^{0.65}$$

where A_{s} is the specific external packing area. The validity of this power product has been verified over a wide range of operating conditions: $0.3 \leq Re_{\mathrm{l}} \leq 3000$, $3.2 \times 10^3 \leq Ga_{\mathrm{l}} \leq 0.32 \times 10^6$, $0 \leq \Delta P/(\varrho_{\mathrm{l}}\, g\, L) \leq 16$, $0.3 \leq \varrho_{\mathrm{g}} \leq 81\,\mathrm{kg/m^3}$, and different packings [26–28].

The influence of the gas – liquid surface tension is negligible. However, for small catalyst particle diameters (e.g., < 1 mm), capillary forces may increase β_{dyn}. The above equation is valid only for nonfoaming liquids; for foaming systems, the hydrodynamic behavior is quite different (e.g., the holdup is much lower and is also largely affected by the gas – liquid surface tension). Experimental work on foaming systems can be found in [12], [29].

To calculate β_{dyn} from the above-mentioned equation, the pressure gradient must be known. This gradient can be obtained by using a pressure drop correlation (see Section 2.3), but it can also be measured relatively easily under operating conditions in pilot- as well as large-scale reactors.

For the *pulse-flow regime*, empirical correlations for the holdup have been proposed [27], [29]. The influence of pulses on the holdup is probably less significant in large-diameter columns. As mentioned before, pulses span the entire reactor diameter in small-scale columns, but in large-scale columns, they probably appear only locally and other parts of the column remain in trickle flow.

2.3. Pressure Drop

As a consequence of gas flow downward through the packed bed, a pressure drop arises over the trickle-bed reactor due to friction at the gas – liquid interface. Knowledge of this pressure drop is essential for sizing the compression equipment. The pressure gradient depends not only on the gas throughput but also on the liquid throughput, because the external liquid holdup directly influences the interstitial velocity of the gas. At the transition from trickle to pulse flow, no large changes in pressure drop have been observed [9], [19]. In describing the pressure gradient, three different approaches have been followed:

1) A modified Ergun equation is used. Here the packed-bed porosity term $\varepsilon_{\mathrm{bed}}$ has been corrected for the volume occupied by the liquid phase; see [27], [30]. This type of equation is restricted to the trickle-flow regime.
2) The dimensionless two-phase pressure drop $\varphi = (\Delta P_{\mathrm{lg}}/\Delta P_{\mathrm{l}})^{1/2}$ is correlated to the Lockhart – Martinelli parameter $X = (\Delta P_{\mathrm{l}}/\Delta P_{\mathrm{g}})^{1/2}$; see also Section 2.2. For a number of three-phase systems, Midoux et al. correlated the parameter φ to

$$\varphi = 1 + \frac{1}{x} + \frac{1.14}{x^{0.54}}$$

This equation is valid in the range $0.1 < X < 80$, $0.15 < \varrho_{\mathrm{g}} < 1.9\,\mathrm{kg/m^3}$, and for different flow regimes [29].
3) The pressure drop is decribed by means of the friction factor $f = [\Delta P/(0.5\, \varrho_{\mathrm{g}} V_{\mathrm{g}}^2)] \cdot (d_{\mathrm{p}}/L)$, which is correlated to dimensionless numbers, such as the Reynolds numbers of the gas and liquid phases. For the *trickle-flow regime*, Rao et al. proposed the following equation for the friction factor, based on several three-phase systems [31]:

$$f = 4.6 \times 10^3 \frac{Re_{\mathrm{l}}^{0.4}}{Re_{\mathrm{g}}^{1.05}} \left[\frac{(1-\varepsilon_{\mathrm{bed}})^2}{\varepsilon_{\mathrm{bed}}^3}\right]^{0.75}$$

For *other flow regimes*, the same types of equations have been proposed by Rao, with different values for the constant and exponents. The advantage of this approach is that, once the flow regime is known, the pressure gradient can be determined directly from the operating variables.

Most of the proposed equations have been derived for two-phase flow with water as the liquid phase and nitrogen or air as gas phase under

atmospheric conditions. Overviews of the available correlations, valid for this gas density range, are given in [31], [32].

At present, only WAMMES et al. have determined pressure drops for gas densities up to 81 kg/m^3 [7]. For the *trickle-flow regime* they proposed the following equation

$$\frac{\Delta P}{0.5\varrho_g U_g^2}\cdot\frac{d_p}{L} = 155\left[\frac{\varrho_g U_g \varepsilon_{bed} d_p}{\eta_g\,(1-\varepsilon_{bed})}\right]^{-0.37} \cdot\left[\frac{1-\varepsilon_{bed}}{\varepsilon_{bed}\,(1-\beta_{tot})}\right]$$

for the range of Reynolds numbers of the gas phase $200 < [\varrho_g U_g \varepsilon_{bed} d_p]/[\eta_g(1-\varepsilon_{bed})] < 5000$ and for a total external liquid holdup $\beta_{tot} > 25\,\%$ of the void volume between particles. The Re_g and the kinetic energy term ($0.5\,\varrho_g U_g^2$) have been based on the average real interstitial gas velocity U_g, given by $U_g = V_g/[\varepsilon_{bed}(1-\beta_{tot})]$. In this range of Re_g numbers, the pressure drop is proportional to $\varrho_g^{0.63} U_g^{1.63}$. For $Re_g < 200$ the pressure drop is overestimated when using this relation. The equation of RAO et al. demonstrates that for lower Re_g, the pressure drop is approximately proportional to the superficial gas velocity V_g and independent of the gas density ϱ_g.

For foaming systems, pressure drop values are much higher compared with nonfoaming flow under the same operating conditions; see [29].

2.4. Mass Transfer

Several transport steps are involved in the overall conversion rate: (1) transport of a gas-phase reactant from the gas bulk through the gas – liquid interface into the liquid bulk; and (2) transport of reactants from the liquid bulk to the external surface of a porous catalyst and then into the catalyst pores. Determination of the rate-limiting step(s) plays an important role in design. If the gas – liquid or liquid – solid mass-transfer rates are low compared to the intrinsic reaction rate, the hydrodynamic behavior largely dominates the conversion achieved in a trickle-bed reactor because it affects these mass-transfer rates.

2.4.1. Gas-Side Mass Transfer

Experimental studies on the volumetric gas-side mass-transfer coefficient $k_g A_{gl}$ under operating conditions as encountered in trickle beds are scarce. However, the resistance in the gas film is in most cases negligible compared to that in the liquid film, especially when the solubility of the gaseous reactant in the liquid is low. According to the penetration model, k_g is proportional to the square root of the diffusion coefficient of the gas, which is inversely proportional to the gas density. Therefore, at higher reactor pressures the gas-side mass-transfer resistance may become important in the overall conversion rate.

2.4.2. Liquid-Side Mass Transfer

Many studies have been made on determination of the volumetric liquid-phase mass-transfer coefficient $k_l A_{gl}$ and on the specific gas – liquid interfacial area A_{gl}. In general, the liquid-side mass-transfer rate increases at higher gas and liquid throughputs. Several differing correlations have been proposed, and an overview of the available relations is given in [33].

The gas – liquid mass-transfer rate depends strongly on the hydrodynamic flow pattern in the trickle-bed reactor. The mass-transfer resistance is highest in the trickle-flow regime, in comparison to pulse, spray-, and bubble-flow conditions. For flow rates and packings typical in catalytic trickle beds, the $k_l A_{gl}$ values are in the range of 0.01 – 0.3 s^{-1} for *trickle flow*.

In the *pulse-flow regime*, values > 1 s^{-1} have been reported. Under nonfoaming trickle-flow conditions, the specific interfacial area A_{gl} is in the range of 20 – 80 % of the specific external packing area A_s. This has also been verified for high pressure (see [7]). At the transition to *pulse flow*, A_{gl} increases sharply, probably due to the contribution of small gas bubbles inside the liquid pulses, and exceeds the geometric packing area.

Apart from flow regime, the *foaming nature* of the liquid phase has a great influence on A_{gl}. MORSI et al. showed that the interfacial area of a foaming system can be more than 10 times higher than that in a nonfoaming system for both systems operating in the trickle-flow regime [34].

Clearly if literature correlations are used to estimate the liquid-phase mass-transfer rate, a careful selection must be made on the basis of comparable operating conditions and liquid-phase properties.

2.4.3. Liquid – Solid Mass Transfer

The liquid – solid mass-transfer resistance becomes significant when the following inequality holds [1]:

$$\left(\frac{10\,d_{\mathrm{p}}}{C}\right) R\,(1-\varepsilon_{\mathrm{bed}}) > k_{\mathrm{ls}}$$

in which C is the concentration of the gaseous component in the liquid phase at saturation and R is the chemical reaction rate in moles per second per cubic meter of reactor volume. Hence, the liquid – solid mass-transfer resistance may become limiting in the case of highly active catalysts or large catalyst particles.

The hydrodynamic flow regime strongly affects the volumetric liquid – solid mass-transfer coefficient $k_{\mathrm{ls}}\,A_{\mathrm{e}}$, where A_{e} is the effective area for liquid solid mass transfer. In the pulse-flow regime this coefficient is usually several times higher than in the trickle-flow regime [35–37]. The effective area for liquid – solid mass-transfer A_{e} is not necessarily equal to the external catalyst area A_{s}. As shown in Figure 2, a more or less stagnant liquid zone is present around the contacting points of the catalyst particles. The volume of this zone is not necessarily equal to the residual holdup β_{res}, which remains in the packed bed after the column has been drained. The mass-transfer rate through these zones is relatively low and makes a minimal contribution to the overall volumetric liquid – solid mass-transfer rate. The effective area A_{e} increases slightly with gas flow rate and particle size, and more strongly with liquid flow rate (see [38]). In the pulse-flow regime, with pulses over the entire bed diameter, $A_{\mathrm{e}}/A_{\mathrm{s}} = 1$ is likely because the pulses continuously refresh the complete external catalyst surface.

Several differing correlations have been presented, which are mostly of the type $Sh \cdot \varphi = a \cdot Re_{\mathrm{l}}^{\mathrm{b}}\,Sc^{1/3}$, in which φ is the fraction of effective area for mass transfer $A_{\mathrm{e}}/A_{\mathrm{s}}$, Sh is the Sherwood, and Sc the Schmidt number. An overview is given in [35]. In the trickle-flow regime, the influence of gas flow rate on the liquid – solid mass-transfer rate is relatively small, whereas the effect of liquid flow rate is more pronounced because it affects the real liquid velocity $U_{\mathrm{l}} = U_{\mathrm{l}}/[\varepsilon_{\mathrm{bed}} \cdot \beta_{\mathrm{tot}}]$ and, by that, the turbulence around the particle, as well as the effective mass-transfer area.

On the basis of several three-phase systems, Rao and Drinkenburg proposed for the Reynolds range $70 < \varrho_{\mathrm{l}}\,U_{\mathrm{l}}\,d_{\mathrm{p}}/\eta_{\mathrm{l}} < 300$ [35]:

$$\left(\frac{K_{\mathrm{s}}\,d_{\mathrm{p}}}{D}\right) \cdot \frac{A_{\mathrm{e}}}{A_{\mathrm{s}}} = 0.24\left(\frac{\varrho_{\mathrm{l}}\,U_{\mathrm{l}}\,d_{\mathrm{p}}}{\eta_{\mathrm{l}}}\right)^{0.75}\left(\frac{\eta_{\mathrm{l}}}{D}\right)^{1/3}$$

in which D is the molecular diffusivity coefficient and d_{p} the equivalent diameter of a sphere having the same external surface area as the catalyst particles.

2.5. Heat Transfer

The consideration of thermal phenomena is important because the conversion and selectivities are strongly related to them. The design and operating conditions must be developed in such a way that large temperature increases in the trickle bed are avoided because they may lead to excessive evaporation of the liquid, dry catalyst zones, hot spots, sintering, and runaway.

Three heat-transfer steps must be distinguished: (1) heat transport inside the catalyst particle, (2) heat transport between the external catalyst surface and the contacting phases, and (3) heat transport in axial and radial directions through the bed. Criteria to assess whether or not intra particle thermal gradients can be neglected have been developed [39], [40].

The heat of reaction generated in catalyst particles must be transported to the phases surrounding it. Unfortunately, no experimental data on interphase heat-transfer coefficients are available. To estimate the order of magnitude of the fluid – solid temperature difference, the analogy between mass and heat transport is often applied [41]. Little information is available on axial and radial heat transport in the bed by the gas and liquid phases, in contrast to one-phase flow in packed beds of inert solids. Only a few experimental studies on temperature profiles in trickle beds have been published [42–45]. Hence, the dependence of reactor temperature profiles on

operating conditions is difficult to predict; this results in a large uncertainty in design and scale-up.

To control the adiabatic temperature rise, a quench fluid is often applied: in hydrogenation reactions, fresh cold hydrogen gas, cooled hydrogen-rich outlet gas, or cooled liquid products can be used. The recycle of products reduces the plug-flow characteristics of the reactor. The amount of quench fluid and its injection positions must be determined carefully to maintain the temperature sufficiently high throughout the catalyst bed and to avoid large hot spots. Several studies dealing with aspects of quenching are available [46–48]. BARKELEW and GAMBHIR developed empirical criteria to prevent runaway in a trickle-bed reactor and checked these with unpublished incidents in plants [49].

Another possibility for heat removal is to partly vaporize the solvent [50], [51]. The amount of evaporating solvent must be controlled carefully to keep the liquid throughput sufficiently high to maintain fully wetted catalyst particles.

2.6. Wetting Efficiency

The flow around catalyst particles is not uniform (see Fig. 2). As mentioned previously, part of the liquid is more or less stagnant and exhibits a high resistance to mass transfer which leads to nonuniform reactant concentration around the catalyst, in cases of high intrinsic reaction rates. Another hydrodynamic phenomenon reducing reactor efficiency is incomplete wetting of the catalyst particles. In principle, incomplete wetting can take place on a reactor scale as well as on the scale of a particle. Normally, *incomplete wetting of large catalyst zones* can be prevented by proper design of the distributor and correct choice of the H/D and D/d_p ratios.

Incomplete wetting of a particle may occur at low liquid flow rates. Here the holdup is relatively small and the liquid film breaks into rivulets leading to incomplete liquid coverage of the external catalyst surface. If the key reactant in the liquid phase has a low vapor pressure, as in hydrodesulfurization processes, the reactant reaches the catalyst only through the liquid-covered surface. When the catalyst is incompletely wetted, a decrease in the effective area for the liquid-to-solid mass-transfer rate will occur, resulting in a lower overall conversion rate. On the other hand, when the limiting reactant is also present in the gas phase, as in the hydrogenation of unsaturated hydrocarbons of low molar mass, the direct mass-transfer from gas to solid may increase the overall reaction rate; see [52], [53]. In this case the locally enhanced reaction rates, together with the poor heat-removal ability of the gas phase, may lead to hot-spot formation or even reactor runaways.

As a consequence of nonuniform concentration profiles around the catalyst, conventional expressions for the catalyst effectiveness factor are no longer applicable. Several theoretical models have been derived dealing with the dependence of the effectiveness factor on partial wetting of the catalyst [54], [55]. These models divide the external catalyst surface into a dry and an actively wetted part. A model that incorporates the presence of stagnant liquid zones has also been proposed [56], [57].

Above all, to predict the influence of partial wetting on reaction rates, the wetting efficiency f must be known a priori. A survey of experimental studies on wetting efficiencies has been given [58]. At present, no valid correlations are available. The broad hatched region in Figure 5, originally proposed in [1], indicates the possible range of values for the wetting efficiency. In this diagram the wetting efficiency is taken as a function only of the superficial liquid velocity V_l or the liquid-phase Reynolds number Re_l. The broad hatched region encompasses most of the experimental studies; except for the data of SCHWARTZ, which do not show an influence of the liquid flow rate on the wetting efficiency. At $Re_l > 10$, the entire external catalyst surface appears to be covered by flowing liquid.

3. Scale-Up

Due to the complicated hydrodynamics, no straightforward scale-up strategy for trickle-bed reactors can be given [59]. For scale-up, a combination of procedures must be used.

Use of Dimensionless Groups of Parameters. The dimensionless groups are kept constant during scale-up. In trickle-bed reactors,

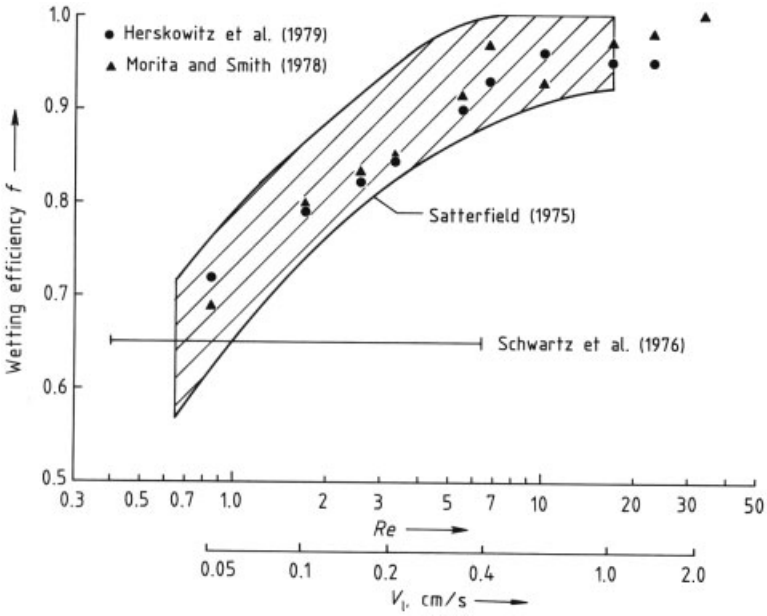

Figure 5. Wetting efficiency in a trickle-flow reactor as a function of liquid load

many dimensionless groups are involved because many independent parameters play a role in process description. Unfortunately all dimensionless groups cannot generally be kept constant during the scale-up.

Use of Mathematical Models. In the derivation of a mathematical process model, the first step is to analyze what simplifications can be made. One of the simple models, which is applied in various hydrotreating processes, is the so-called pseudohomogeneous model. The following assumptions have been made: plug flow; absence of external mass transfer limitations, of thermal gradients, and of stagnant zones; complete particle wetting and pseudo-first-order kinetics. If this simple model does not hold it has to be analyzed so as to find which simplifying assumption induces the largest error. Several other models have been developed to deal with the shortcomings of the pseudohomogeneous model [1], [59].

Use of Small-Scale Reactors. The use of small-scale reactors seems inevitable in the development of new chemical processes for trickle beds. Small, flexible, laboratory recycle or differential reactors are used mainly for catalyst development and obtaining kinetic data. Pilot plants are useful in studying temperature profiles, catalyst deactivation, selectivity, and yield.

Normally, *pilot reactors* have a diameter of 5 – 20 cm and a height of up to 8 m. A criterion for reactor diameter is that $D/d_p > 15$ must be satisfied. In the choice of *catalyst size*, the catalyst efficiency and the pressure drop must be balanced: small catalyst particles give a high catalyst effectiveness but also a large pressure drop over the packed bed. The *reactor length* must be large enough to obtain plug-flow characteristics. A criterion has been given by MEARS [60]. Since industrial reactors usually operate under *adiabatic conditions*, measures must also be taken to operate pilot reactors adiabatically because temperature influences the conversion and selectivity. Sufficient insulation or heat compensation methods can be applied to avoid large temperature profiles in the radial direction. An essential part of the analysis of small-scale experiments is determination of the *rate-controlling step*(*s*). If *external mass transfer* influences the overall conversion rate, the hydrodynamics (i.e., the flow rates) should not be changed too much in scale-up. Therefore, the ratio between the cross-sectional area of the reactor and the liquid and gas throughput must be held constant. A diam-

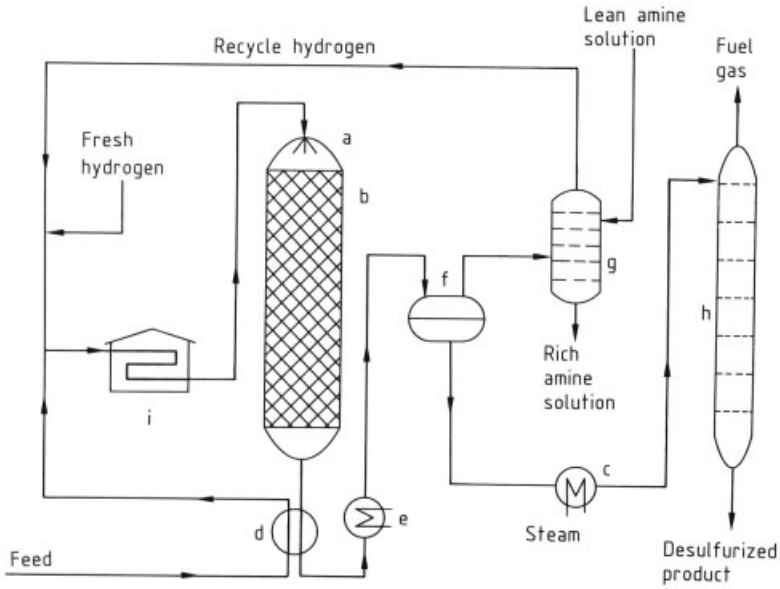

Figure 6. Schematic flow sheet for the Shell HDS process for light feedstocks
a) Trickle-bed reactor; b) Catalyst bed; c) Heater; d) Heat exchanger; e) Cooler; f) Gas – liquid separator; g) Scrubber; h) Stripper

eter increase is limited by the requirement of $H/D > 5$ and by the possibilities of even liquid distribution. In practical cases, this means that if the catalyst volume is enlarged, the height of the reactor must also be increased because an even distribution of liquid is one of the key aspects of successful operation. Poor liquid – solid contacting in catalyst zones may lead to temperature increase, deactivation, coking, and eventually to blocking or plugging of part of the bed. Satisfactory liquid – solid contacting can generally be achieved by a proper initial distribution and relatively high liquid throughputs ($> 12\,m^3\,m^{-2}\,h^{-1}$), as well as the use of liquid redistribution devices.

If the *intrinsic reaction rate* is limiting, the conversion should be kept constant (i.e., via the residence time), particularly when selectivities are important. The catalyst volume can be increased by changing reactor height or diameter while keeping the ratio of liquid holdup (see Section 2.2) to liquid throughput constant.

The largest uncertainties in scale-up are caused mainly by the difficulty in predicting the temperature gradients on a particle scale, as well as on a reactor scale, and the rate of catalyst deactivation.

4. Industrial Applications

The largest class of chemical reactions carried out in cocurrent trickle-bed reactors are hydrogenations of various crude petroleum fractions and other refinery streams and hydrogenations of single chemical compounds. With respect to the applications in the refining industries, the *hydrodesulfurization* (HDS) has gained high importance since the 1970s. The major reactions occurring are the cleavage of the C – S and S – S bonds of compounds which act as catalyst poisons in the downstream process units. In addition, the sulfur contents in the fuel products must be lowered to the allowable limits. The porous catalyst particles have diameters of 1.5 – 3 mm. Particle size and geometry of the pores play an important role because – due to the high gas and liquid flow rates in the trickle-bed reactor–the transport processes within the catalyst pores are probably rate limiting [1], [61]. A schematic flow sheet of the Shell HDS process is given in Figure 6. The petroleum feedstock is mixed with fresh hydrogen and the hydrogen-rich recycle gas.

Under the operating conditions of 300 – 380 °C and 3.0 – 6.0 MPa, the feedstock remains

partly in the liquid phase. After the gas – liquid separator (f), the H_2S is washed out in a scrubber (g). Catalyst deactivation mainly occurs through coke deposition. The reactor temperature is raised gradually to compensate for it. The accumulated coke is periodically burnt off by passing it over a mixture of steam and air. The service life of the catalyst can be as long as 10 years. Other examples in *refining industries* are:

1) Hydrocracking of heavier petroleum fractions into more valuable lighter fractions, e.g., production of LPG from naphtha. Reactor temperatures and pressures are typically in the range 350 – 450 °C and 3.5 – 35 MPa, respectively. The hourly liquid space velocity is normally $1 - 2\,h^{-1}$.
2) Hydrodenitrification and hydrodemetallization reactions.

In the chemical industry trickle-flow reactors are used, for example, in the production of 2-propanol from propene by direct hydration (Deutsche Texaco). Water (liquid phase) and gaseous propene (molar ratio 15 : 1) trickle down over a sulfonic acid ion-exchange resin. The exothermic hydration occurs at 130 – 160 °C and 8.0 – 10 MPa. Under these conditions an aqueous solution of 2-propanol is formed. The water is injected at several positions along the height of the reactor. The life of the cation exchange resin is at least 8 months.

5. References

1. C. N. Scatterfiled, *AIChE J.* **21** (1975) 209 – 228.
2. A. Gianetto, G. Baldi, V. Specchia, S. Sicardi, *AIChE J.* **24** (1978) 1087 – 1104.
3. M. Herskowitz, J. M. Smith, *AIChE J.* **29** (1983) 1 – 18.
4. S. Fukushima, K. Kusaka, *J. Chem. Eng. Jpn.* **10** (1977) 461 – 474.
5. F. Larachi, N. Midoux, A. Laurent, *10th International Congress of Chemical Engineering, Chemical Equipment Design and Automation,* Prague 1990.
6. F. Larachi, N. Midoux, A. Laurent, *2nd Int. Symp. High Press. Chem. Eng.* 1990, Erlangen, pp. 441 – 446.
7. W. J. A. Wammes et al., *AIChE J.* **37** (1991) 1849 –1862 .
8. Y. Sato et al., *J. Chem. Eng. Jpn.* **6** (1973) 315 – 319.
9. J. R. Blok, J. Varkevisser, A. A. H. Drinkenburg, *Chem. Eng. Sci.* **38** (1983) 687 – 699.
10. J. R. Blok, A. A. H. Drinkenburg, *Chem. Eng. J. (Lausanne)* **25** (1982) 89 – 99.
11. G. Christensen, S. J. McGovern, S. Sundaresan, *AIChE J.* **32** (1986) 1677 – 1689.
12. P. S. T. Sai, Y. B. G. Varma, *Can. J. Chem. Eng.* **66** (1988) 353 – 360.
13. T. S. Chou, F. L. Worley, D. Luss, *Ind. Eng. Chem. Process Des. Dev.* **16** (1977) 424 – 427.
14. W. J. A. Wammes, S. J. Mechielsen, K. R. Westerterp, *Chem. Eng. Sci.* **45** (1990) 3149 – 3158.
15. K. Grosser, R. G. Carbonell, S. Sundaresan, *AIChE J.* **34** (1988) 1850 – 1860.
16. D. C. Dankworth, I. G. Kevrekidis, *Chem. Eng. Sci.* **45** (1990) 2239 – 2246.
17. E. Talmor, *AIChE J.* **23** (1977) 868 – 878.
18. J. C. Charpentier, M. Favier, *AIChE J.* **21** (1975) 1213 – 1218.
19. G. Tosun, *Ind. Eng. Chem. Process Des. Dev.* **23** (1984) 29 – 35.
20. O. Baker, *Oil Gas J.* **53** (1954) 185 – 191.
21. S. Sicardi, G. Baldi, V. Specchia, *Chem. Eng. Sci.* **35** (1980) 1775 – 1782.
22. J. C. Charpentier, C. Prost, W. van Swaaij, P. Le Goff, *Chem. Ind. Gènie Chim.* **99** (1968) 803 – 826.
23. S. Ergun, *Chem. Eng. Prog.* **48** (1952) 89 – 94.
24. S. Whitaker, A. E. Cassano: *Concepts and Design of Chemical Reactors,* Gordon and Breach, New York 1986, p. 307.
25. W. J. A. Wammes, K. R. Westerterp, *Chem. Eng. Sci.* **45** (1990) 2246 – 2254.
26. W. J. A. Wammes, S. J. Michielsen, K. R. Westerterp, *Chem. Eng. Sci.* **46** (1991) 409 – 417.
27. V. Specchia, G. Baldi, *Chem. Eng. Sci.* **32** (1977) 515 – 523.
28. M. Kohler, G. Richarz, *Chem.-Ing.-Tech.* **56** (1984) H 10.
29. N. Midoux, M. Favier, J. C. Charpentier, *J. Chem. Eng. Jpn.* **9** (1976) 350 – 356.
30. B. E. T. Hutton, L. S. Leung, *Chem. Eng. Sci.* **29** (1974) 1681 – 1685.
31. V. G. Rao, S. Ananth, Y. B. G. Varma, *AIChE J.* **29** (1983) 473 – 483.
32. P. S. T. Sai, Y. B. G. Varma, *AIChE J.* **33** (1987) 2027 –2036.

33. N. Midoux et al., *Chem. Eng. Sci.* **39** (1984) 781 – 794.
34. B. I. Morsi, N. Midoux, A. Laurent, J. C. Charpentier, *Int. Chem. Eng.* **22** (1982) 142 – 151.
35. V. G. Rao, A. A. H. Drinkenburg, *AIChE J.* **31** (1985) 1059 – 1068.
36. J. A. Ruether, C. Yang, W. Hayduk, *Ind. Eng. Chem. Process Des. Dev.* **19** (1980) 103 – 107.
37. T. S. Chou, F. L. Worley, D. Luss, *Ind. Eng. Chem. Fundam.* **18** (1979) 279 – 283.
38. V. Specchia, G. Baldi, A. Gianetto, *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) 362 – 367.
39. C. D. Prater, *Chem. Eng. Sci.* **8** (1958) 284 – 290.
40. D. E. Mears, *Ind. Eng. Chem. Process Des. Dev.* **10** (1971) 541 – 549.
41. A. H. Calderbank: *Mass Transfer in Mixing,* Academic Press, New York 1967.
42. S. B. Jaffe, *Ind. Eng. Chem. Process Des. Dev.* **15** (1976) 410 – 418.
43. J. Hanika, K. Sporka, V. Ruzicka, J. Hrstka, *Chem. Eng. J. (Lausanne)* **12** (1976) 193 – 200.
44. J. Hanika, V. Vosecky, V. Ruzicka, *Chem. Eng. J. (Lausanne)* **21** (1981) 108 – 115.
45. A. H. Germain, A. G. Lefebvre, G. A. l'Homme, *Adv. Chem. Ser.* **133** (1974) 164 – 170.
46. T. Y. Yan, *Can. J. Chem. Eng.* **58** (1980) 259 – 266.
47. R. D. Mhaskar, Y. T. Shah, J. A. Paraskos, *Ind. Eng. Chem. Process Des. Dev.* **17** (1978) 27 – 35.
48. Y. T. Shah, R. D. Mhaskar, J. A. Paraskos, *Ind. Eng. Chem. Process Des. Dev.* **15** (1976) 400 – 410.
49. C. H. Barkelew, B. S. Gambhir, *ACS Symp. Ser.* **237** (1984) 62 – 81.
50. K. B. van Gelder, J. K. Damhof, P. J. Kroijenga, K. R. Westerterp, *Chem. Eng. Sci.* **45** (1990) 3159 –3170.
51. K. B. van Gelder, P. C. Borman, R. E. Weenik, K. R. Westerterp, *Chem. Eng. Sci.* **45** (1990) 3171 – 3192.
52. A. Mata, J. M. Smith, *AIChE Symp. Ser.* **77** (1981) 29 – 35.
53. M. Herskowitz, R. G. Carbonell, J. H. Smith, *AIChE J.* **25** (1979) 272 – 283.
54. P. L. Mills, M. P. Dudukovic, *Chem. Eng. Sci.* **35** (1980) 2267 – 2279.
55. C. Tan, *Chem. Eng. Sci.* **43** (1988) 1281 – 1286.
56. S. Sicardi, G. Baldi, A. Giantto, V. Specchia, *Chem. Eng. Sci.* **35** (1980) 67 – 73.
57. S. Sicardi et al., *Chem. Eng. Sci.* **36** (1981) 228 – 229.
58. C. M. Ruecker, R. K. Hess, A. Akgerman, *Chem. Eng. Commun.* **49** (1987) 301 – 315.
59. H. Hofmann, *Conf. Chem. Reaction Eng.* **6,** Ghent, 1979.
60. D. E. Mears, *Chem. Eng. Sci.* **26** (1971) 361 – 369.
61. G. H. Henry, J. B. Gilbert, *Ind. Eng. Chem. Proc. Des. Dev.* **12** (1973) 328.

Threonine → Amino Acids
Thulium → Rare Earth Elements

Reaction Columns

For other industrial reactors and their applications, see → Stirred-Tank and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Thin-Film Reactors, → Metallurgical Furnaces, and → Biochemical Engineering.

OTTO WÖRZ, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

HANS-HORST MAYER, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 2
2. Mathematical Model 4
3. Case Studies 5
3.1. Effect of Feed Tray and Number of Plates 5
3.2. Effect of Equilibrium Constant 5
3.3. Different Relative Volatilities of A and B 6
3.4. Increasing Relative Volatility of D . . 7
3.5. Declining Relative Volatility of A . . . 7
4. Important Applications 7
5. Summary 8
6. References 8

Symbols (see also → Principles of Chemical Reaction Engineering and → Model Reactors and Their Design Equations)

a_i, b_i, c_i coefficients in Antoine equation for species i
A, B, C, D species
E_{ij} symbol in phase-equilibrium Equation (5) for species i on tray j
F_j feed rate to tray j, kmol/h
G_j vapor rate from tray j, kmol/h
h_j specific enthalpy of liquid stream from tray j, kJ/kmol
H_j specific enthalpy of vapor stream from tray j, kJ/kmol
$H_{F,j}$ specific enthalpy of feed to tray j, kJ/kmol
K equilibrium constant of reaction, kmol/kmol
L_j liquid stream from tray j, kmol/h
M_{ij} symbol in mass balance of species i on tray j
n total number of theoretical plates
P_j pressure on tray j, Pa
q_j reaction heat on tray j, kJ/h
Q_j heat flow to tray j, kJ/h
R_j symbol in reaction equation on tray j
S_{1j}, S_{2j} symbols in equations for sums of mole fractions on tray j, index 1 denotes liquid phase, index 2 denotes gas phase
$S_{G,j}$ withdrawal rate of vapor side stream from jth tray, kmol/h
$S_{L,j}$ withdrawal rate of liquid side stream from jth tray, kmol/h
T_j temperature on tray j, °C
U_j conversion on tray j, kmol/h
W_j symbol in heat balance on tray j
x_{ij} mole fraction of species i in liquid phase on tray j, kmol/kmol
y_{ij} mole fraction of species i in gas phase on tray j, kmol/kmol
z_{ij} mole fraction of species i in feed to tray j, kmol/kmol

Subscripts
i component number ($1 \leq i \leq 4$)
j tray number ($1 \leq j \leq n$)
1 liquid phase
2 gas phase

1. Introduction

Reaction and separation of products generally take place in separate devices. The most common type of reactor is the stirred tank; the most common type of separation apparatus, the distillation (rectification) column. The term reaction column is used when the reaction can be carried out in a column together with distillation. Such an arrangement leads to savings in investment costs.

If a reaction proceeds to chemical equilibrium, depending on the equilibrium constant, the

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_321

reaction may have to be combined with distillation in order to shift the equilibrium to the product side [1–3]. One possibility for such coupled operation is a continuous countercurrent reaction column implementing a large number of theoretical plates (Fig. 1A). Alternatives are a batch reactor with an attached column (Fig. 1B) and one ($n = 1$) or a cascade ($n > 1$) of continuous stirred-tank reactors with attached columns (Fig. 1C). In the last case, the design may include a column on one reactor receiving volatile components from all the reactors.

An increasingly important question is which of these three configurations is energetically most favorable for a given reaction system. The comparison between *reaction column* and *continuous stirred-tank reactor plus column* has been studied in detail [4,5]. With regard to energy consumption, the reaction column is fundamentally preferable to the CSTR plus column. The energy consumption of the reaction column is the more favorable, the more difficult it is to separate the more volatile product from the reaction mixture. The two designs are energetically equivalent only in the limiting case where the separation of the volatile product is very easy. *Cascades* can be considered as a continuous transition between the reaction column and CSTR plus column. Therefore, the batch reactor – column combination is the only configuration that has to be considered as an alternative to the reaction column. Both designs are compared in this article.

Three prerequisites for the use of a reaction column follow:

1) The reaction must proceed so fast that equilibrium is reached, or is at least approached, on each tray. The residence times are ca. 1 min for packed columns, ordered packings, or sieve trays and 5 – 10 min for bubble-cap trays. Longer residence times can be achieved only by expensive, special designs, e.g., separate residence time containers.
2) The reaction must have a low activation energy; the activation energy must be lower when the temperature profile of the column becomes steeper. Otherwise, the reaction would take place on just a few trays, and the energetic advantages of the reaction column would disappear; the steeper the temperature profile, the more important is the use of a correctly designed control system [6].
3) The reaction must not be markedly exothermic or endothermic. The more *exothermic* a reaction is, the greater is the danger of uncontrolled vaporization in the apparatus and the more difficult is the problem of control. The more *endothermic* the reaction, the further is the equilibrium at the very top of the column shifted toward the starting material (given a strong temperature gradient in the column).

All three of these constraints vanish in the case of the batch reactor – column combination (and also the cascade design). For purposes of comparison, the frequent case $A + B \rightleftharpoons C + D$ is considered, where B and C are the more volatile components on either side of the equilibrium and C has the highest relative volatility of all reaction partners. (Otherwise the equilibrium could not be shifted by withdrawing the C stream at the top of the column.)

As the design of an apparatus becomes more specific, the number of reaction systems for which it is suitable decreases, but at the same time the advantages to be gained with these reaction systems increase. By comparison with the batch reactor plus column, the reaction column is quite a specific design. The objective must be to investigate the suitability of a reaction system for the reaction column by using those properties of the system that are easiest to determine.

First, a system is examined in which *only the highly volatile product C is difficult to separate* from the other three reaction partners. (If C is very easy to separate, e.g., because of its high relative volatility, reactor design has little effect on the energy requirement.) If reactants A and B are assumed to have roughly equal volatilities, while product D has a low volatility and is thus easy to separate, the isolation of C in the rectification section is comparable in the two designs; that is, the reaction column does not require any additional property of the reaction system for the rectification section. The reaction column, therefore, will show its greatest energetic advantages in this case. If, on the other hand, *A and B have very different relative volatilities*, the system is not well suited to the reaction column because the reactants quickly separate in the column, producing a narrow reaction zone. This trend can be countered by feeding the two

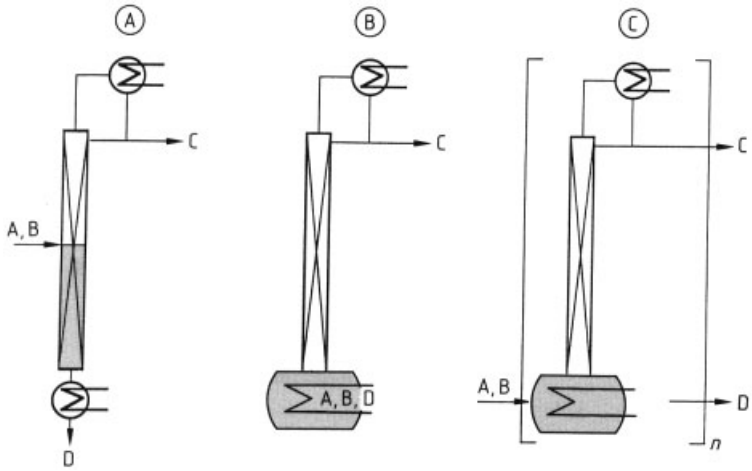

Figure 1. Alternative reaction modes (shading indicates reaction zones)
A) Continuous mode: reaction column; B) Batch mode: reactor with column; C) Continuous mode: CSTR or cascade of stirred-tank reactors

reactants to different trays. In the batch design, however, a low volatility for one of the reactants is actually an advantage, since the difficulty of separation is lower. Thus, the energetic advantages of the reaction column can be expected to be less pronounced in this situation. In addition, the advantages of the reaction column are assumed to decline *as the relative volatility of product D increases,* since this product is to be recovered in pure form at the bottom of the column. In the batch design, on the other hand, an increase in the relative volatility of D is unimportant. Furthermore, the reaction column is expected to have an advantage, particularly *when the equilibrium constant is low*. The higher the equilibrium constant, the further would the reaction proceed to completion anyway, so that the specific ability of the reaction column to shift the equilibrium becomes less important. The conclusions drawn here are verified and quantified in what follows.

To reduce the number of cases considered, several restrictions are applied: (1) The reaction requires a poorly volatile catalyst; that is, the reaction proceeds only in the stripping section in the reaction column and only in the tank of the batch reactor with column. (2) Species A and B are inlet to the reaction apparatus in a 1: 1 molar ratio; in the reaction column, they are usually fed to the same tray.

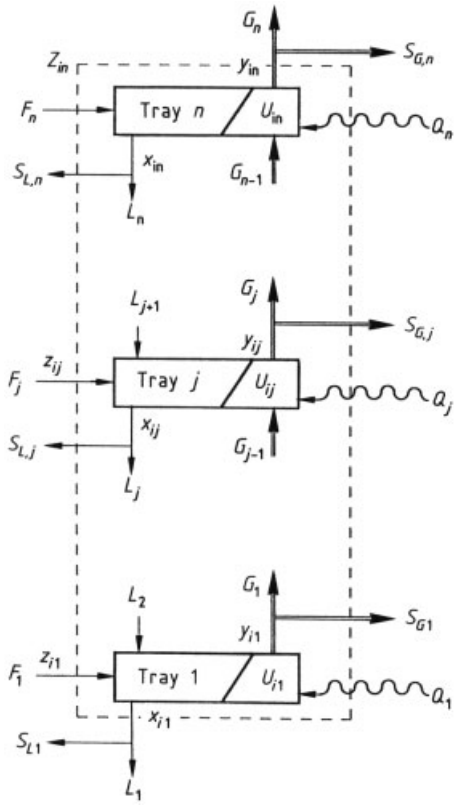

Figure 2. Model for simulation of reaction columns

2. Mathematical Model

Because of the complexity of the system, a mathematical model should be used to determine how the different parameters affect the conversion [5–9]. The model used for simulating continuous reaction columns was developed from the theoretical-plate model for pure distillation, with the reaction incorporated into the mass and thermal balances. The column model consists of trays coupled to one another in countercurrent (Fig. 2).

The following conditions are imposed:

1) Phase equilibrium between the streams leaving a tray
2) Complete mixing of the reaction phase on each tray

Ideal vapor – liquid equilibrium is assumed; the vapor pressures are described by the Antoine equation, with constant relative volatilities assumed between individual species. The specific heat and heat of vaporization are taken to be equal for all four species. Under these conditions, the model equations for each tray ($1 \leq i \leq 4$, $1 \leq j \leq n$) are as follows [1]:

Mass balance

$$M_{ij} \equiv L_{j+1} x_{ij+1} - \left(L_j + S_{L,j}\right) x_{ij} - \\ - \left(G_j + S_{G,j}\right) y_{ij} + G_{j-1} y_{ij-1} + \\ + F_j z_{ij} + U_j = 0 \quad (1)$$

where M_{ij} denotes the mass balance of component i in tray j, L_j denotes the liquid stream from tray j, S_L the withdrawal rate of liquid side stream, S_G the withdrawal rate of gaseous side stream, G the vapor rate, F the feed rate, and U the conversion.

Heat balance

$$W_j \equiv L_{j+1} h_{j+1} - \left(L_j + S_{L,j}\right) h_j - \\ - \left(G_j + S_{G,j}\right) H_j + G_{j-1} H_{j-1} + \\ + F_j H_{F,j} + q_j + Q_j = 0 \quad (2)$$

where W is the heat balance, h the specific enthalpy of liquid stream from a tray, H the specific enthalpy of vapor stream from a tray, H_F the specific enthalpy of feed to a tray, q the reaction heat on the tray, and Q the heat flow to a tray.

Sums of mole fractions

$$S_{1j} \equiv \sum_{i=1}^{4} x_{ij} - 1 = 0 \quad (3)$$

$$S_{2j} \equiv \sum_{i=1}^{4} y_{ij} - 1 = 0 \quad (4)$$

Phase equilibrium

$$E_{ij} \equiv y_{ij} - \exp\left[a_i - b_i / \left(c_i + T_j\right)\right] / P_j x_{ij} = 0 \quad (5)$$

Reaction equilibrium

$$R_j \equiv x_{3j} x_{4j} - K x_{1j} x_{2j} = 0 \quad (6)$$

To determine the two degrees of freedom remaining when the pressure, feed rate, number of plates, and feed tray have been specified, the required purity of the overhead product and the heat input are generally stated.

Solution of the model equations by a global Newton – Raphson method and the extension of the model equations to the batch case are described in [5].

3. Case Studies

3.1. Effect of Feed Tray and Number of Plates

A reaction system is assumed in which the four species have the following relative volatilities: A = 1.0, B = 1.1, C = 4.0, D = 0.1. Component C is withdrawn at the top of the column with a purity of 99.9 %. Such a system meets the requirements for advantageous use of the reaction column, since D is very poorly volatile whereas A and B have similar volatilities. Figure 3 shows the relative heating power needed to achieve a given conversion of reactant A (or B) if the equilibrium constant $K = 1$. The reaction column and the column atop the batch reactor have ten theoretical plates (n). At a given conversion, the location of the feed tray in the reaction column is an important factor for the energy requirement of the reaction column. Figure 3 compares the batch case with calculations for reaction columns with the feed at various trays. Whether the reaction column or the batch reactor is preferable depends on the feed location. As the required conversion increases, the optimal location of the feed tray moves upward in the column.

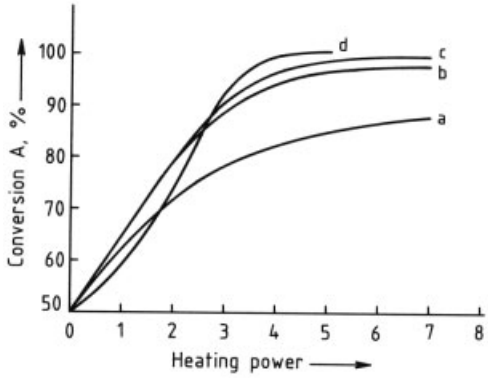

Figure 3. Effect of feed location
$K = 1$; $n = 10$; A = 1.0; B = 1.1; C = 4.0; D = 0.1
a) Reaction column: feed location is tray 1; b) Reaction column: feed location is tray 2; c) Batch reactor with column; d) Reaction column: feed location is tray 4

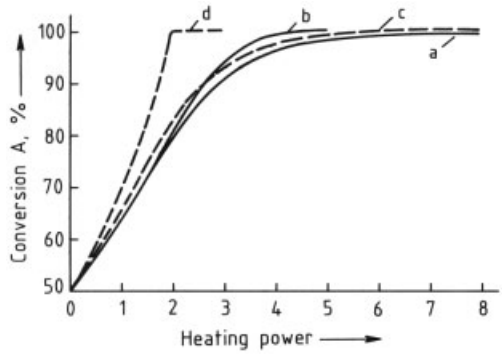

Figure 4. Effect of number of plates
$K = 1$; A = 1.0; B = 1.1; C = 4.0; D = 0.1
—· $n = 10$; – – – $n = 20$
a); c) Batch reactor with column; b); d) Reaction column

In the following cases the common feed location for A and B is varied until a minimum heat requirement is reached. Figure 4 presents optimal curves for reaction columns with 10 and 20 theoretical plates, compared with the curves for the batch reactor – column combination. For $n = 10$ and a conversion below ca. 80 %, no difference exists between the two designs in terms of energy consumption. The advantages of the reaction column become apparent only at higher conversions. For $n = 20$, the reaction column is better at all degrees of conversion; that is, the batch reactor with column cannot utilize the additional theoretical plates as well. The reaction column offers great advantages, especially at very high conversion. The energy requirements of the two designs are nearly unchanged on going to 40 theoretical plates.

3.2. Effect of Equilibrium Constant

In Figure 5, reactions with different equilibrium constants are investigated. The optimal region for the reaction column corresponds to $K = 0.01$ (curve b). At 95 % conversion, the reaction column needs ca. 5 % of the heating power required by the batch reactor with column (curve a). If the equilibrium constant is increased to $K = 10$ (curves e and f), on the other hand, most of the advantages of the reaction column are lost.

Thus, if the relative volatilities of the four components are matched to the specific properties of the reaction column, for an equilibrium constant $K \leq 1$ the reaction column yields significant energetic advantages over the batch design when many theoretical plates are present and a high conversion is sought. As the equilibrium constant decreases, this energetic advantage increases rapidly.

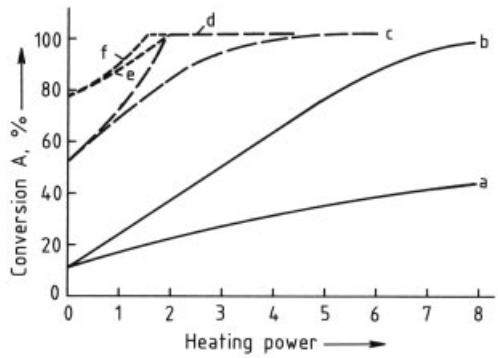

Figure 5. Effect of equilibrium constant
$n = 20$; A = 1.0; B = 1.1; C = 4.0; D = 0.1
—· $K = 0.01$; – – $K = 1$; – – – – $K = 10$
a); c); e) Batch reactor with column; b); d); f) Reaction column

3.3. Different Relative Volatilities of A and B

Figure 6 shows an example in which the relative volatilities of species A and B differ markedly

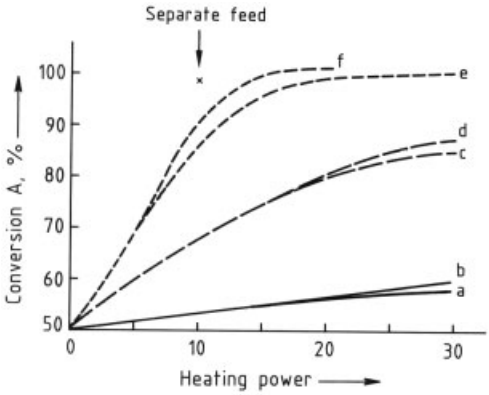

Figure 6. Effect of different reactant volatilities
$K = 1$; A = 1.0; B = 3.0; C = 4.0; D = 0.1
——— $n = 25$; – – $n = 30$; – – – – $n = 40$
a); c); e) Batch reactor with column; b); d); f) Reaction column

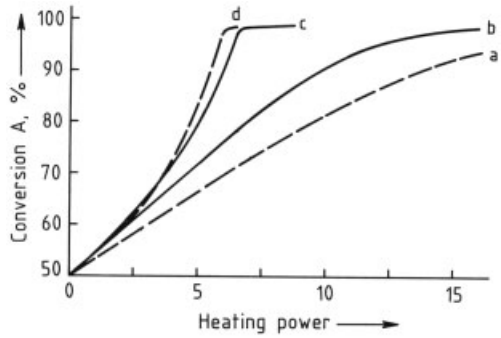

Figure 7. Effect of reducing overhead product purity from 99.9 to 98 %
$K = 1$; $n = 40$; B = 3.0; C = 4.0; D = 0.1
——— A = 1.0; – – A = 2.9
a); b) Batch reactor with column; c); d) Reaction column

(1: 3), that is, for a system not ideally matched to the reaction column. In contrast to Figure 4 (volatilities 1: 1.1), the reaction column has virtually no advantage, even for a large number of plates. However, increasing the relative volatility of B simultaneously changes the volatility ratio of B and C. The B – C separation problem may become so dominant that other differences are concealed. In Figure 7 (curves —-) and all subsequent examples, the required purity of overhead product C is therefore decreased from 99.9 to 98 %. The work of separation is thus reduced, and the difference between the two reactor designs does indeed become larger again. For comparison, Figure 7 (curves – –) shows another case in which the relative volatility of A has been raised to 2.9. Curves (—-) and (– –) in Figure 7 do not differ in the B – C separation problem but only in the A to B ratio. Under these conditions, the effect of the A to B ratio is not so significant.

The higher the difference in relative volatilities of A and B, the lower are the energetic advantages of the reaction column, as expected. The flattening out of the differences is greater as B comes closer to C in relative volatility and the required purity of overhead product C increases. This effect can be countered by separating the feed locations of the reactants. The point "separate feed" in Figure 6 represents separate feed for 99.9 % conversion with both feed locations optimized.

3.4. Increasing Relative Volatility of D

The case of Figure 8 (curves – – –) again roughly corresponds to the ideal case for the reaction column. The relative volatilities of A and B differ (1: 2), but they are so far from C that the reaction column offers a substantial advantage at high conversions. In Figure 8 (curves – –), the relative volatility of D has been increased so that it lies between those of A and B. Now the curves for both types of reactor design intersect. In this case the reaction column is better at low conversion and worse at high conversion than the batch reactor with column. If the relative volatility of D is increased beyond that of B, batch operation is preferable at all conversions; the differences become very large at high conversions (Fig. 8 curves —-).

The strong influence of the relative volatility of D is similar to the effect of the equilibrium constant. Just as a decrease in equilibrium constant rapidly boosts the energetic advantages of the reaction column, so does an increase in the relative volatility of D bring out the advantages of batch operation.

3.5. Declining Relative Volatility of A

Figure 9 (curves —-) repeats Figure 8 (curves – – –). In Figure 9 (curves – –), the relative volatil-

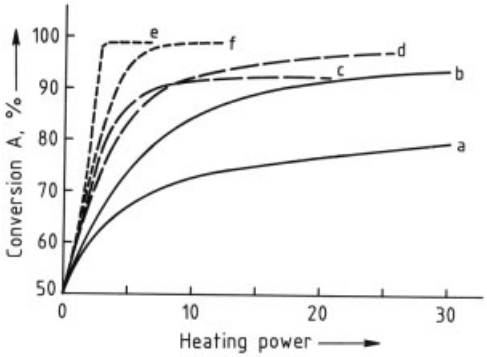

Figure 8. Effect of volatility of D
$K = 1$; $n = 20$; A = 1.0; B = 2.0; C = 4.0
- - - - D = 0.1; – – D = 1.5; ——— D = 2.5
a); c); e) Reaction column; b); d); f) Batch reactor with column

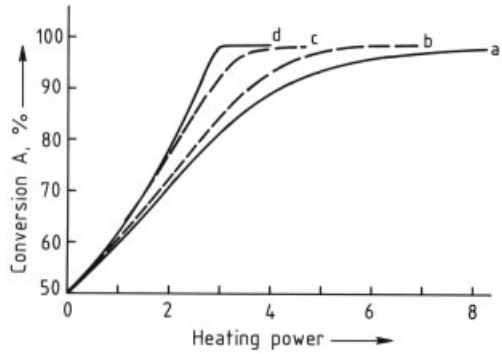

Figure 9. Effect of volatility of A
$K = 1$; $n = 20$; B = 2.0; C = 4.0; D = 0.1
——— A = 1.0; – – A = 0.1
a); b) Batch reactor with column; c); d) Reaction column

Table 1. Estimation of advantages of reaction columns in comparison to batch reactors with column reaction system

K	Relative volatilities [a]				Preferred design [b]		
	A	B	C	D	RC	RC ≈ BR	BR
0.01	1.0	1.1	4.0	0.1	+++		
1	1.0	1.1	4.0	0.1	++		
10	1.0	1.1	4.0	0.1	+		
1	1.0	3.0	4.0	0.1	+	+	
1	2.9	3.0	4.0	0.1	++		
1	1.0	2.0	4.0	0.1	++		
1	1.0	2.0	4.0	1.5		+	+
1	1.0	2.0	4.0	2.5			+++
1	0.1	2.0	4.0	0.1	+		
1	0.1	2.0	4.0	1.0		+	

[a] A ratio of relative volatilities of 1.1: 1.5: 10 corresponds to a difference in boiling points of ca. 3 °C: 10 °C: 100 °C.
[b] RC = reaction column; BR = batch reactor with column.

ity of A has been lowered to the same value as that of D (0.1). The advantages of the reaction column show little diminution. If, however, the relative volatility of D is increased at the same time, a picture similar to Figure 8 (curves – –) is obtained, as expected. Again, increasing the number of plates restores a small advantage to the reaction column.

Lowering the relative volatility of A thus decreases the energetic advantages of the reaction column, as does increasing the relative volatility of D. This effect, however, is not as strong and only leads to an advantage for batch operation at high conversions when the relative volatility of A falls below that of D.

4. Important Applications

Reaction columns are now widely used in industry, mainly for esterifications. These include the production of simple esters such as ethyl formate, methyl acetate, and butyl acetate as well as the esterification of long- chain acids (e.g., oleic acid, tallow fatty acids, and coconut oil acids) with methanol. Reactions with polyols are also carried out, examples include the esterification of acetic acid with glycerol, and of terephthalic acid with ethylene glycol. Reaction columns are also used for the saponification of esters, for transesterifications (e.g., the reaction of dimethyl terephthalate with ethylene glycol), and

for the cleavage of acetals. Another application is the reaction of methacrylic acid with acetic anhydride to give methacrylic anhydride.

5. Summary

The equilibrium constant of a reaction and the relative volatilities (boiling points) of the reaction partners are easily determined properties of a reaction system. With the cases analyzed in the figures, one can decide quickly whether the reaction column or the batch reactor with column offers significant energetic advantages or whether no marked differences can be expected. Table 1 summarizes the results.

Laboratory tests, as a rule, are first performed in a batchwise mode. If a reaction system is ideally suited to a reaction column, then laboratory trials in which the reaction partners are in reaction vessels with an attached column will result in much greater thermal stresses on the reactants and products than would be the case in a reaction column. Large selectivity losses may occur, leading to the abandonment of the process at the bench-testing stage. At an early point, therefore, it is important to know that higher selectivities may be achieved in a reaction column.

6. References

1. U. Block, B. Hegner: "Verfahrensvarianten eines Veresterungsverfahrens," *VT Verfahrenstech.* **11** (1977) no. 3, 157 – 160.
2. H. Hartig, H. Regner: "Verfahrenstechnische Auslegung einer Veresterungskolonne," *Chem. Ing. Tech.* **43** (1971) no. 18, 1001 – 1007.
3. J. Suzuki, H. Yagi, H. Komatsu, M. Hirata: "Calculation of Multicomponent Distillation Accompanied by a Chemical Reaction," *J. Chem. Eng. Jpn.* **4** (1971) no. 1, 26 – 33.
4. U. Block: "Verfahrensführung kontinuierlicher Reaktionen mit überlagerter Destillation," *Chem. Ing. Tech.* **49** (1977) no. 2, 151;MS 444/77.
5. G. Kaibel, H. H. Mayer, B. Seid: "Reaktionen in Destillationskolonnen," *Chem. Ing. Tech.* **50** (1978) no. 8, 586 – 592.
6. U. Trilling, G. Kaibel, *RTP Regelungstech. Prax.* **21** (1979) no. 5, 118 – 124.
7. S. Nagel, A. Lieberam: "Die Berücksichtigung chemischer Reaktionen bei der mathematischen Simulation von Stufentrennprozessen," *Chem. Ing. Tech.* **46** (1974) no. 5, 197, MS 026/74.
8. J. Jelinek, V. Hlavacek: "Steady State Countercurrent Equilibrium Stage Separation with Chemical Reaction by Relaxation Method," *Chem. Eng. Commun.* **2** (1975) 79 – 85.
9. U. Block, B. Hegner: "Simulationsmethode für homogene Flüssigphasenreaktionen mit überlagerter Destillation," *VT Verfahrenstech.* **11** (1977) no. 2, 101 – 104.

Thin-Film Reactors

For other industrial reactors and their applications, see → Stirred-Tanc and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Metallurgical Furnaces, and → Biochemical Engineering.

BERNHARD GUTSCHE, Henkel KGaA, Düsseldorf, Federal Republic of Germany

CHRISTOPH BREUCKER, Henkel KGaA, Düsseldorf, Federal Republic of Germany

GÜNTER PANTHEL, Henkel KGaA, Düsseldorf, Federal Republic of Germany

1. Design of Thin-Film Reactors 2
1.1. Gas – Liquid Interface 2
1.2. Fluid Dynamics 3
1.3. Heat Transfer 4
1.4. Mass Transfer 5
1.5. Models 5
2. Applications of Thin-Film Reactors . 6
3. Designs and Operating Modes of Thin-Film Reactors 6
4. Thin-Film Reactors for Industrial Sulfonation 7
5. References 9

Symbols (see also → Principles of Chemical Design Engineering and → Model Reactors and Their Design Equations)

A	transfer area
a	interfacial area per unit reaction volume
b	constant
c_p	isobaric heat capacity
d	diameter
$E(\tau)$	distribution function of residence time
L	length
$\dot{m}$	mass flow
n	number of tubes
q	heat
T	temperature
t	time
v	volume
$\dot{V}$	throughput
Ha	Hatta number
Nu	Nusselt number
Pr	Prandtl number
Re	Reynolds number
Sc	Schmidt number
Sh	Sherwood number
Ta	Taylor number
α	heat-transfer coefficient
β	mass-transfer coefficient
δ	film thickness
θ	dimensionless residence time
λ	coefficient of thermal conduction
ν	kinematic viscosity
τ	residence time
ω	angular velocity

Subscripts

c	coolant
G	gas
h	hydraulic (diameter, see Eq. 17)
i	internal (diameter of the inner cylinder of the reactor, see Eq. 17)
if	interface
l	laminar
L	liquid
t	tube, turbulent
W	wall

A thin-film apparatus, with or without rotating internals, is typically used for mild, nondestructive distillation, absorption, or desorption (stripping) in process engineering. Starting from this field of application, it has also acquired an important role in gas – liquid reactions.

1. Design of Thin-Film Reactors

Thin-film reactors (*falling-film reactors and film reactors with mechanical wiping*) are contacting devices for gas – liquid systems. The liquid phase is in the form of a thin film that runs down a vertical wall while gas flows over the liquid surface cocurrently or countercurrently. In the *rising-film reactor*, gas flowing upward at a high

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_329

rate causes a thin liquid film to ascend (gas and liquid cocurrent).

Thin-film reactors are used chiefly when strongly endothermic or exothermic reactions require a very high rate of heat transfer. These reactors are operated continuously. Thin-film reactors consist of one tube, or a number of tubes bundled as in a shell-and-tube heat exchanger, cooled or heated from the outside (Fig. 1). In another design [2], two large concentric pipes form an annular clearance. The liquid is distributed over both surfaces of the annular space and reacts with the gas flowing in the free volume between the pipes.

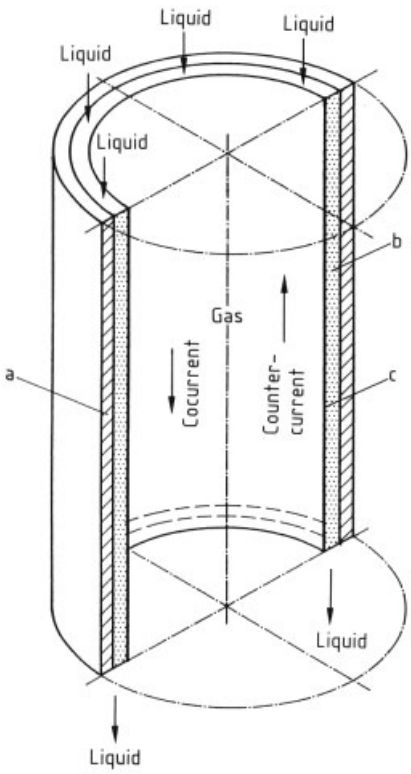

Figure 1. Schematic of a thin-film reactor [1]
a) Tube wall (heat transfer); b) Liquid film; c) Gas – liquid interface (reaction)

1.1. Gas – Liquid Interface

The thin-film reactor is the only type of gas – liquid contact apparatus in which both gas and liquid are present as continuous phases and, in fact, the dispersion of either phase must be prevented. Entrainment or breakup of the film cannot be permitted since heat transfer would be impaired and local overheating might degrade the product in the case of exothermic reactions. Options for calculating entrainment are summarized in [3]. To prevent entrainment, tubes must be installed precisely vertically, and the tube surface must be adequately finished. For highly viscous liquids, a rotating wiper system can be used to maintain a uniform film, afford better mixing of the liquid, and promote heat transfer (*wiped-film reactor*).

Thin-film reactors have a very small specific surface area for mass transfer between gas and liquid compared to other reactors. The mass transfer area A is governed only by reactor geometry (e.g., reactor length L, tube diameter d_t, number of tubes n). For a laminar film of thickness δ

$$A = n\pi L(d_t - \delta) \tag{1}$$

The increase in film surface area due to wave formation in the turbulent-flow regime is negligibly small. Since $\delta << d_t$ in most cases, the geometric area of the tube can be used for practical calculations:

$$A = n\pi L d_t \tag{2}$$

Hence the interfacial area per unit reaction volume is

$$a = \frac{n\pi L d_t}{n\frac{\pi}{4}Ld_t^2} = \frac{4}{d_t} \tag{3}$$

If the geometric interfacial area is divided by the volume of liquid in the film, the ratio is

$$a_L = \frac{n\pi L d_t}{n\pi L d_t \delta} = \frac{1}{\delta} \tag{4}$$

1.2. Fluid Dynamics

With regard to back mixing, the thin-film reactor is similar to the tubular reactor. Neither the gas nor the liquid phase shows much back mixing, and both can be treated as plug flows. Thin-film reactors are therefore also used when a narrow residence-time distribution enhances selectivity and yield. Figure 2 shows typical residence-time distributions for a number of reactors [1].

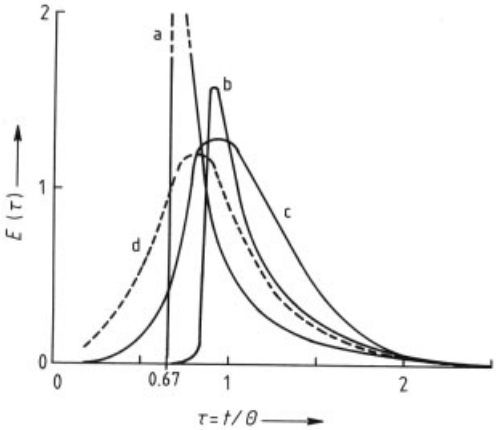

Figure 2. Residence-time distribution for different reactors [1]
a) Laminar falling film (ideal); b) Turbulent falling film; c) Turbulent tubular reactor; d) Film reactor with mechanically produced film

The residence times of the two phases are very short, and the liquid holdup is very small. Both quantities can be calculated easily from the film thickness and the reactor geometry. The ratio of gas and liquid throughputs in this type of reactor is usually in the range

$$10\,\frac{\mathrm{m}^3}{\mathrm{m}^3} < \dot{V}_\mathrm{G}/\dot{V}_\mathrm{L} < 2000\,\frac{\mathrm{m}^3}{\mathrm{m}^3} \tag{5}$$

If constraints, such as residence time or optimal reaction conditions, make it necessary to alter this ratio, an inert gas component can be added to dilute the reactant gas.

Falling films can be analyzed in three characteristic flow regimes. If the Reynolds number for the film is defined as

$$Re_\mathrm{L} = \frac{\dot{V}_\mathrm{L}}{d_\mathrm{t}\pi\nu_\mathrm{L}} \tag{6}$$

the boundaries of the regimes are as follows:

$Re_\mathrm{L} \leq 25$ laminar, smooth film
$25 < Re_\mathrm{L} < 400$ transition regime, laminar film
with waves
$Re_\mathrm{L} \geq 400$ turbulent film

For the laminar regime, the velocity distribution in the falling film can be used in calculating film thickness [4]. The calculated thickness is in satisfactory agreement with experimental data [5].

The following equation generalizes the results:

$$\delta = b_1 \left(\frac{3\nu_\mathrm{L}^2}{g}\right)^{1/3} Re^{b_2} \tag{7}$$

Table 1 gives the parameters for different flow regimes.

Table 1. Parameters from Equation (7) for different flow regimes

Re_L	b_1	b_2	Author	References
<80	1	1/3	Nusselt	[4]
100 – 420	0.93	1/3	Kapitza	[6]
>400	0.435	7/12	Zhivaikin	[7]
>420	0.369	1/2	Feind	[8]

1.3. Heat Transfer

The great advantage of thin-film reactors is their large heat-transfer area per unit liquid volume. The heat-transfer area is roughly the same as the mass-transfer area (see Eqs. 3 and 4). Such reactors are employed, accordingly, for strongly exothermic and endothermic reactions. Because these reactions are often accompanied by the absorption of a gaseous reactant, the evolved heat of an exothermic reaction and heat of absorption raise the temperature of the liquid, especially at the gas – liquid interface, thereby decreasing the solubility of the gas in the liquid. The absorption and the reaction rate therefore, strongly depend on the rate of heat removal.

A compilation and comparison of literature data on heat transfer for liquid films are given by Yüksel [9], who determined the heat-transfer and mass-transfer coefficients independently of each other and measured the film surface temperature by a contactless method. He confirmed that the turbulence in the film at the surface is damped, resulting in an additional thermal resistance near the film surface besides the resistance in the boundary layer of liquid next to the wall (Fig. 3).

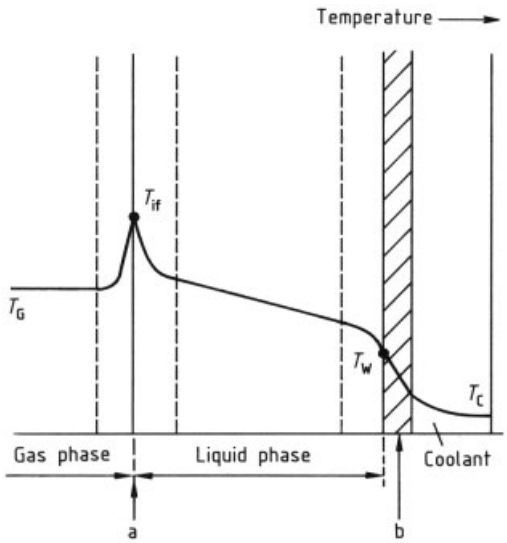

Figure 3. Temperature profile of a thin-film reactor for an exothermic reaction
a) Interface; b) Wall

On the basis of heat-transfer measurements in condensation on a falling film with a variety of liquids ($2.6 \leq Pr < 52$), plus literature data, MÜLLER [10] reports the following practical formulas for the mean heat transfer over the film length:
Laminar film:

$$Nu_{\mathrm{l}} = 0.606 Re^{-2/9} \tag{8}$$

Turbulent film:

$$Nu_{\mathrm{t}} = 0.0105 Re^{0.35} Pr^{0.43} \tag{9}$$

where

$$Nu = \frac{\alpha \sqrt{\nu_{\mathrm{if}}/g}3}{\lambda_{\mathrm{if}}} \tag{10}$$

$$\alpha = \frac{\dot{q}}{T_{\mathrm{if}} - T_{\mathrm{W}}} \tag{11}$$

$$Pr = \frac{\nu_{\mathrm{if}}\, c_p\, \varrho}{\lambda_{\mathrm{if}}} \tag{12}$$

The fourth powers of the laminar and turbulent contributions can be added. Because poorer heat transfer in higher-viscosity liquids gives rise to marked viscosity differences in the film, the temperature dependence of the material parameters is also introduced.

$$Nu = \left(Nu_{\mathrm{l}}^4 + Nu_{\mathrm{t}}^4\right)^{0.25} \left(\frac{Pr}{Pr_{\mathrm{W}}}\right)^{0.25} \tag{13}$$

Computational equations for heat transfer in mechanically produced films can be found [1].

1.4. Mass Transfer

Because of the limited transfer area and the short gas and liquid residence times, thin-film reactors are best employed only for fast ($0.3 < Ha < 3$) or instantaneous ($Ha > 3$) reactions. Most of the reaction occurs in the boundary layer, while the contribution of the bulk liquid phase is negligible. The rate-determining step is accordingly gas-side mass transfer. For the gas-side local mass-transfer coefficient β_{G}, which is a function of the Sherwood number, BRAUN and HIBY give the following dimensionless formulas [11]:
Cocurrent flow:

$$Sh = 0.18\, Re_{\mathrm{G}}^{0.4}\, Re_{\mathrm{L}}^{0.16}\, Sc_{\mathrm{G}}^{0.44} \left(1 + 6.4 \frac{L}{d_{\mathrm{t}}}\right)^{-0.75} \tag{14}$$

Countercurrent flow:

$$Sh = 0.015 \cdot Re_{\mathrm{G}}^{0.75}\, Re_{\mathrm{L}}^{0.16}\, Sc_{\mathrm{G}}^{0.44} \left(1 + 5.2 \frac{L}{d_{\mathrm{t}}}\right)^{-0.75} \tag{15}$$

The last term in Equations (14) and (15) takes care of the fact that the mass-transfer coefficient declines to a constant final value as reactor length increases.

For the wiped-film reactor with rotating annular clearance that is operated with gas and liquid ascending cocurrently [12], BROSTRÖM gives the following empirical equation for gas-side mass transfer [13]:

$$Sh = 0.079 Re^{0.67}\, Ta^{0.16}\, Sc^{0.52} \tag{16}$$

where the Taylor number Ta is defined as

$$Ta = \left(\frac{d_{\mathrm{h}}}{d}\right)^{0.5} \frac{\omega \pi d_{\mathrm{i}}}{2} \cdot \frac{d_{\mathrm{h}}}{2} \cdot \nu \tag{17}$$

d_{h} is the hydraulic diameter of the annulus and d_{i} is the internal diameter of the inner cylinder.

Calculations for mass transfer in the liquid phase can be found, e.g., in [1] and [9]. The discussion above leads to the concentration curve of the gaseous component shown in Figure 4. The entire reaction takes place in the film. The concentration of dissolved gas in the bulk liquid approaches zero. Absorption of the gaseous components should therefore be enhanced by the reaction. JANA and coworkers [14] find correspondingly high enhancement factors, up to 12. MANN [15], on the other hand, measures enhancement factors < 1. This disagreement can be explained by strong thermal effects, leading to a temperature rise at the gas – liquid interface (Fig. 3) and a correspondingly low solubility of the reactant gas in the reacting liquid.

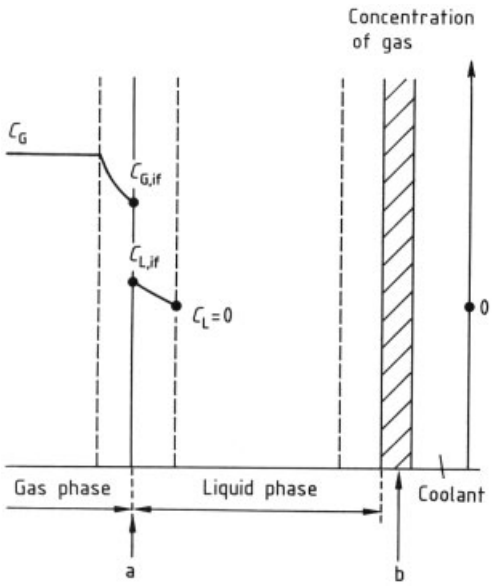

Figure 4. Concentration of gas in a thin-film reactor for a fast reaction
a) Interface; b) Wall

1.5. Models

Many publications have dealt with models of absorption and reaction in films, especially for the thin-film reactor. VILLADSEN and NIELSEN [16], [17] give a survey of existing models (examples are the chlorination of decane and the sulfonation of alkylbenzenes) [18], [15]. JANA and coworkers present a general model for the countercurrent absorption of gas in a laminar film with first-order reaction [14]. For the absorption of carbon dioxide in dilute sodium hydroxide solution, HAIMOUR and SANDALL show that the chemosorption process can be precalculated [19].

GUTIÉRREZ-GONZÁLEZ and coworkers propose a special model for the reactions of sulfur trioxide with alkylbenzenes in a cocurrent setup [20]. The model is compared with the authors' experiments. A simple model for cocurrent ascending sulfonation in a special wiped-film reactor is described by BROSTRÖM [21].

2. Applications of Thin-Film Reactors

In industry, film reactors are employed for a variety of processes involving either strongly exothermic reactions, where quasi-isothermal operation is achieved by cooling the reaction tube, or endothermic reactions, where evaporation or desorption can be achieved only by supplying heat. In the second case, thermodynamic constraints (vapor – liquid or sorptive equilibrium) must be accounted for. For both applications a high selectivity can be attained only with precise temperature control avoiding side reactions and sequential reactions.

The most important industrial use of the *falling-film reactor* without rotating internals is the sulfonation of organic products with sulfur trioxide, but other applications are also mentioned frequently in publications and patents. Table 2 summarizes the fields in which these reactors are used. The reactions in Table 2 are divided up with respect to the direction of mass transfer: absorption with reaction (industrially the most important form), reaction with superimposed absorption/desorption (where a tower-type reactor is more often employed [22]), and reaction with desorption of a byproduct, e.g., water formed in esterifications. In reaction-with-desorption processes used in polymer chemistry, film reactors with rotating internals (i.e., vertical or horizontal thin-layer evaporators) are frequently used. A good survey of fast gas – liquid reactions is found in [23].

Table 2. Practical applications of gas – liquid reactions in thin films without sulfonation/sulfation (see Chap. 4 for data on sulfonation with SO_3)

Absorption	Absorption and desorption	Desorption
Chlorination of hydrocarbons [15]	countercurrent esterification of fatty acids with short-chain alcohols [26]	esterification, transesterification [26]
Oxidation [23]		polycondensation (polyesters) [1]
Nitration [24]		
Neutralization, (e.g., of CO_2) [19]		Vilsmeier reaction [27]
Alkylation of isobutane with butene in solvent containing catalyst [25]		polymerization (polyamide 66) [28]

3. Designs and Operating Modes of Thin-Film Reactors

Many possible designs and modes of operation are described in the literature. Most of them find use only in the laboratory.

Figures 5,6,7,8 show examples of designs. The *falling-film reactor* illustrated in Figure 5 as a cocurrent multitube apparatus [29], requires good liquid distribution to form the liquid film. In addition, the viscosity must not increase much during the reaction.

The surface area of the single-tube falling-film apparatus is optimally utilized in the form of an *annulus reactor* [30], [31], since both the inner wall of the outer pipe and the outer wall of the inner pipe serve as reaction surfaces. Internals that swirl the gas stream give improved mass transfer, at least in the laboratory [32]. A corresponding thin-film reactor design, operated with gas and liquid ascending cocurrently, is the *rising-film reactor* [33].

For high-viscosity products, *film reactors with rotating internals* are employed. The internals can be installed either vertically or horizontally. Figure 6 shows an annulus reactor (Votator) [12] in which the gas and the liquid are fed in from below cocurrently. To improve mass transfer inside the liquid film, the rotating inner pipe is fitted with many small pins.

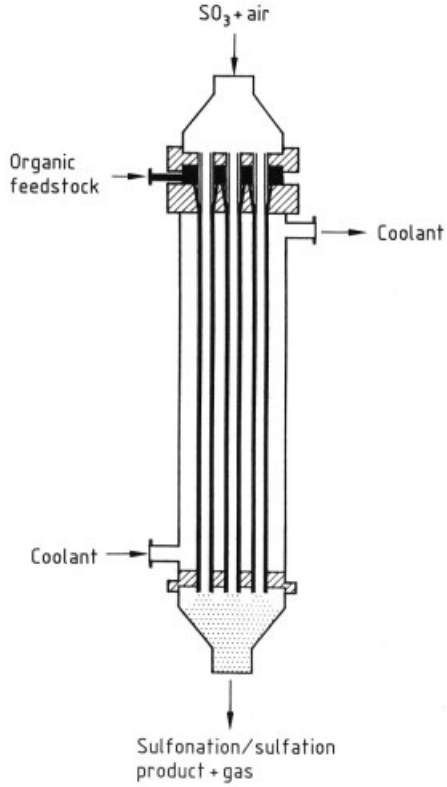

Figure 5. Ballestra falling-film sulfonation reactor [29]

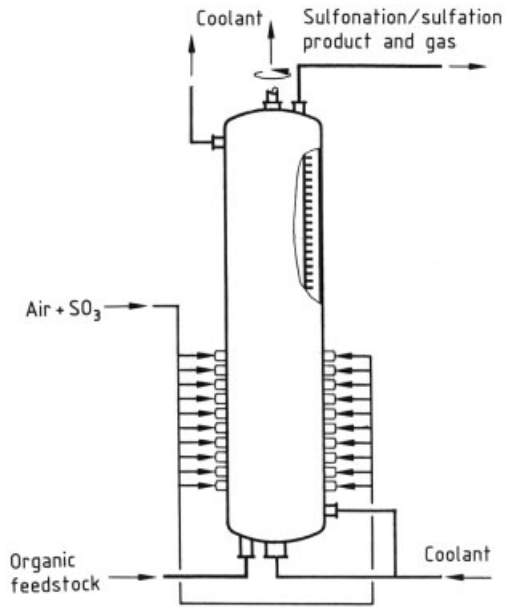

Figure 6. Berol Kemi annulus reactor with rotating inner pipe (Votator) [12]

A wiped-film *reactor for countercurrent operation* is shown in Figure 7. This device is used for reactions with superimposed absorption and possibly desorption, e.g., esterification of viscous fatty acids with short-chain alcohols.

Figure 8 illustrates a *wiped-film reactor* with specially designed wiper elements, intended for use with high-viscosity liquids (up to 80 Pa · s) [34]. This reactor is employed for reactions with desorption in polymer production.

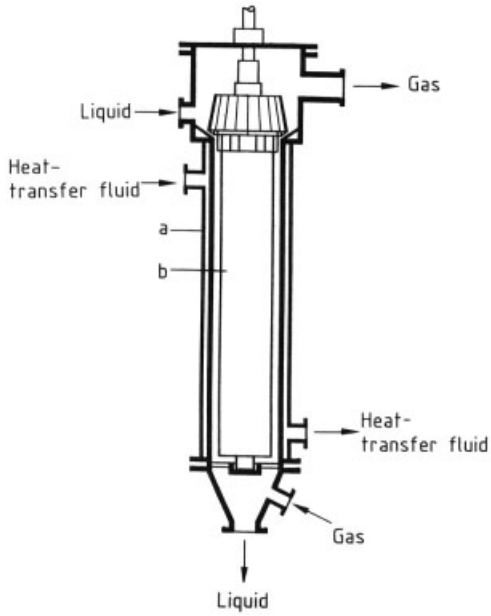

Figure 7. Thin-film reactor for countercurrent operation (schematic)
a) Heating jacket; b) Rotor

4. Thin-Film Reactors for Industrial Sulfonation

The literature contains many publications on the preparation of surfactants by sulfonating and sulfating organic feedstocks with sulfur trioxide. Along with the chlorination of hydrocarbons, the sulfonation of alkylbenzenes and the sulfation of fatty alcohols have served as test systems for the development of film reactors. Selected publications on sulfonation of different organic feedstocks with sulfur trioxide in film reactors of various types are listed below:

Petroleum distillates, crude oils	[35]
α-Olefins	[30], [36], [37]
Isoolefins	[38]
Alkylbenzene	[12], [36], [37], [39]
Ethoxylated alkylphenol	[36]
Fatty alcohol	[32], [33], [36], [37]
Ethoxylated fatty alcohol	[30], [33], [36], [37]
Methyl esters	[37], [40], [41]

The most important sulfonation feedstocks today are alkylbenzenes, α-olefins, and fatty alcohols and their derivatives:

Alkylbenzenes

α-Olefins

$$R\text{–}CH{=}CH_2 + SO_3 \rightarrow R\text{–}CH{=}CH\text{–}(CH_2)_n\text{–}SO_3H$$

Fatty alcohols

$$R\text{–}OH + SO_3 \rightarrow R\text{–}OSO_3H$$

Fatty alcohol ethoxylates

$$R\text{–}(OC_2H_4)_n\text{–}OH + SO_3 \rightarrow R\text{–}(OC_2H_4)_n\text{–}OSO_3H$$

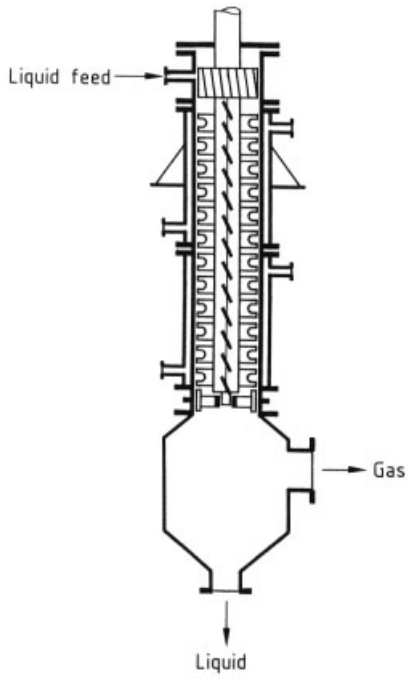

Figure 8. Buss-SMS wiped-film reactor for high-viscosity liquids (polymer production) [34]

The sulfonic acids or alkylsulfuric acids obtained are then neutralized to the desired alkali salts (mainly sodium salts).

Organic feedstocks are very sensitive to formation of degradation products (oversulfonation); these byproducts lead to discoloration or impurities in the products. The feedstocks are therefore reacted with gaseous sulfur trioxide cocurrently, in proportions as close as possible to equimolar. Sulfonation reactions are strongly exothermic ($\Delta H \approx -150$ kJ/mol), and the heat of reaction must be removed quickly from the liquid phase. Not all products of the reaction with sulfur trioxide are stable so they must be neutralized immediately after the sulfonation/sulfation reaction. Examples of such products are alkylsulfuric acids formed from fatty alcohols and sulfur trioxide.

Thin-film reactors ideally satisfy the two principal requirements for sulfonation reactions: short residence time of reactants in the reactor and good removal of heat of reaction from the reactor. Both single-tube and multitube reactors are used for this process.

In *single-tube reactors*, diluted sulfur trioxide gas is led into the annular clearance formed by two concentric pipes. The liquid film flows down the respective inner and outer surfaces of the pipes. Each pipe has intensive cooling on the side away from the liquid film. A typical representative of such single-tube reactors is the Chemithon reactor [42].

Multitube reactors have frequently been used since the 1980s and have up to 120 individual reaction tubes, the organic feedstock being fed to the inner surface of each. After the liquid film has formed, diluted sulfur trioxide gas is fed to each reaction tube. The heat of reaction is removed just as if the reactor were a vertical shell-and-tube heat exchanger. Figure 5 shows the Ballestra multitube reactor [29]. The *Mazzoni multitube reactor* has a special feature [43]: After the feedstock is admitted, air is blown into each reaction tube to spread out the liquid, thus promoting the formation of a uniform film of organic liquid. Only when this step is complete is sulfur trioxide gas admitted. As shown in Figure 9, each reaction tube has a double wall so that it can be cooled separately. It is sometimes argued that the equalizing air would retard the transport of sulfur trioxide to the reaction site and thus reduce local overheating due to the fast reaction. This effect, however, certainly results from further dilution of the sulfur trioxide gas by air, and the reduction in concentration could just as well be performed before the gas inlet to the reactor.

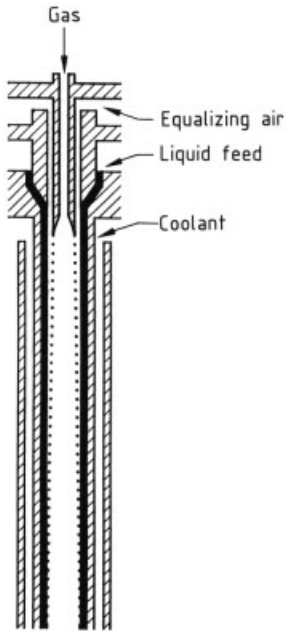

Figure 9. Schematic diagram of one tube in a Mazzoni multitube reactor [43]

Along with the development of hardware, more stringent product quality standards have also led to changes in reaction conditions. For example, the sulfur trioxide concentration in the process gas has been decreased to allow better handling of the heat of reaction. The dewpoint of the process air has also been lowered so that the formation of oleum from residual moisture can be minimized. Table 3 lists the ranges of the most important reaction parameters.

Table 3. Operating parameters of sulfonation reactors

Parameter	Single-tube reactor	Multitube reactor
Number of tubes	1	4 – 200
Tube diameter, mm	150 – 1000	25.4
Tube length, mm	ca. 3000	ca. 6000
$\dot{V}_G$, m^3/h (STP)	2000 – 7000	20 – 70 per tube
$\dot{m}_L$, kg/h	150 – 3500	18 – 36 per tube
Dewpoint of air, °C	− 60 to − 80	− 60 to − 80
SO_3, vol %	2 – 7	2 – 7

5. References

1. P. Trambouze, H. van Landeghem, J. P. Wauquier: *Les réacteurs chimiques,* Editions Technip, Paris 1984.
2. Lion Corp., DE 2 629 009, 1976 (T. Ogoshi, Y. Miyawaki, F. Kondo, S. Sakurai).
3. R. G. Krebs, E. U. Schlünder, *Chem. Eng. Process.* **18** (1984) 341 – 356.
4. W. Nusselt, *VDI Z.* **60** (1916) 541 – 546.
5. H. Brauer: *Grundlagen der Einphasen- und Mehrphasenströmungen,* Sauerländer, Aarau-Frankfurt/M. 1971.
6. P. L. Kapitza, *Z. Exp. Theor. Phys.* **18** (1948) 3 – 18.
7. L. Y. Zhivaikin, *Int. Chem. Eng.* **2** (1962) 337 – 341.
8. K. Feind: "Strömungsuntersuchungen bei Gegenstrom von Rieselfilmen in lotrechten Rohren," *VDI-Forschungsh.* **481** (1960).
9. M. L. Yüksel: "Wärme- und Stoffübergang bei der nicht isothermen Absorption am Rieselfilm," *Fortschr. – Ber., VDI* Reihe 3, no. 133, VDI-Verlag, Düsseldorf 1987.
10. J. Müller: "Einfluß der Stoffwerte auf den Wärme-übergang bei der Kondensation am Rieselfilm," lecture held at GVC Fachausschuß *Wärme- und Stoff-übertragung,* April 25, 1991, Freudenstadt, to be published in *Chem. Eng. Process.*
11. D. Braun, J. W. Hiby, *Chem. Ing. Tech.* **42** (1970) 345 –349.
12. Berol Kemi, DE-OS 2 523 875, 1975 (A. Broström).
13. A. Broström, *Trans. Inst. Chem. Eng.* **53** (1975) 26 – 28.
14. S. C. Jana et al., *Chem. Eng. Commun.* **79** (1989) 27 –37.
15. R. Mann, H. Moyes, *AIChE J.* **23** (1977) 17 – 23.
16. P. H. Nielsen, J. Villadsen, *Chem. Eng. Sci.* **38** (1983) 1439 – 1453.
17. J. Villadsen, P. H. Nielsen, *Chem. Eng. Sci.* **41** (1986) 1655 – 1671.
18. G. R. Johnson, B. L. Crynes, *Ind. Eng. Chem. Process Des. Dev.* **13** (1974) 6 – 14.
19. N. Haimour, O. C. Sandall, *AIChE J.* **29** (1983) 277 –281.
20. J. Gutièrrez-Gonzàlez, C. Mans-Teixidó, J. Costa Lòpez, *J. Dispersion Sci. Technol.* **6** (1985) 303 – 315.
21. A. Broström, *Trans. Inst. Chem. Eng.* **53** (1975) 29 – 33.
22. B. Schleper, B. Gutsche, J. Wnuck, L. Jeromin, *Chem. Ing. Tech.* **62** (1990) 226 – 227.
23. Y. T. Shah: *Gas-Liquid-Solid Reactor Design,* McGraw-Hill, New York 1979.
24. A. Beenackers, W. van Swaaij, *Chem. Eng. J. (Lausanne)* **15** (1978) 25 – 38.
25. I. Pervez, C. A. Karagiozov, C. B. Boyadjev, *Chem. Eng. Sci.* **46** (1991) 1589 – 1594.
26. VEB Deutsches Hydrierwerk Rodleben, FR 1 534 622, 1967.
27. B. Covelli, U. Lattmann, *Wärme-Stoffübertrag.* **12** (1979) 233 – 241.
28. D. D. Steppan, M. F. Doherty, M. F. Malone, *Ind. Eng. Chem. Res.* **29** (1990) 2012 – 2020.
29. Ballestra Chimica, DE 3 006 791, 1980 (G. Moretti, S. Noe).
30. I. Yamane, *J. Am. Oil Chem. Soc.* **55** (1978) 81 – 86.
31. Lion Corp., US 3 839 391, 1971 (R. Suzuki, S. Tanimori, S. Toyoda, T. Ogoshi); US 3 925 441, 1973 (S. Toyoda, T. Ogoshi, M. Maruyama, Y. Miyawaki).
32. Yu. D. Panaev, D. I. Zemenkov, V. G. Pravdin, V. N. Sokolov, *Int. Chem. Eng.* **17** (1977) 337 – 338.
33. K. Takei, K. Tsuto, S. Miyamoto, J. Wakatsuki, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 341 – 347.
34. Buss-SMS, High Viscosity Technology, company brochure, Butzbach 1976.
35. T. L. Ashcraft, R. K. Saunders, *JCPT J. Can. Pet. Technol.* **19** (1980) 47 – 53.
36. W. Fricke, *Tenside* **4** (1967) 317 – 320.
37. M. Ballestra, G. Moretti, *Chim. Oggi* **7 – 8** (1984) 41 – 43.
38. D. W. Roberts, D. L. Williams, *Tenside Deterg.* **22** (1985) 193 – 195.
39. A. Ujhidi, B. Babos, L. Farády, *Chem. Tech. (Leipzig)* **18** (1966) 652 – 654.
40. K. H. Schmid, H. Baumann, W. Stein, H. Dolhaine, *Henkel Ref.* **21** (1985) 11 – 19.
41. T. Ogoshi, Y. Miyawaki, *JAOCS J. Am. Oil Chem. Soc.* **62** (1985) 331 – 335.
42. The Chemithon Corp., Sulfonation Process Equipment, company brochure, Seattle.
43. Mazzoni, DE 2 107 968, 1971 (A. Lanteri).

Metallurgical Furnaces

For other industrial reactors and their applications, see → Stirred-Tanc and Loop Reactors, → Tubular Reactors, → Fixed-Bed Reactors, → Fluidized-Bed Reactors, → Bubble Columns, → Three-Phase Trickle-Bed Reactors, → Reaction Columns, → Thin-Film Reactors, and → Biochemical Engineering.

Jörg Grzella, Lurgi AG, Frankfurt, Federal Republic of Germany (Chaps. 1, 3, 4)

Peter Sturm, Lurgi AG, Frankfurt, Federal Republic of Germany (Chap. 2)

Joachim Krüger, Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany (Chap. 5)

Markus A. Reuter, Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany (Chap. 5)

Carina Kögler, Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany (Chap. 5)

Thomas Probst, Institut für Metallhüttenwesen und Elektrometallurgie, Rheinisch-Westfälische Technische Hochschule, Aachen, Federal Republic of Germany (Chap. 5)

1. Rotary Kilns 2
1.1. Design 2
1.1.1. General 2
1.1.2. Structural Elements 2
1.2. Process Engineering 3
1.3. Applications 4
1.3.1. Roasting and Calcining 4
1.3.2. Calcination of Limestone, Dolomite, and Magnesite 4
1.3.3. Production of Cement Clinker . . . 4
1.3.4. Iron Ore Reduction 5
1.3.5. Volatilization Processes 6
1.3.6. Other Applications 6
2. Multiple-Hearth Furnaces 6
2.1. Description 8
2.2. Applications 9
2.2.1. Roasting of Sulfide Ores 9
2.2.2. Use for Endothermic Reactions . . . 10
2.2.3. Clarifier Sludge Treatment 10
3. Shaft Furnaces 11
4. Smelting, Melting, and Refining in Bath and Flash Smelting Reactors 13
4.1. Bath Smelting Furnaces 14
4.2. Flash Smelting Furnaces 14
4.3. Bath Melting and Refining 14
4.4. Converters 15
5. Electrothermal Reactors 16
5.1. Introduction 16
5.1.1. History 16
5.1.2. Electrothermal Furnaces 17
5.1.3. Energy Consumption and Production Capacities 19
5.1.4. Economic Aspects 20
5.2. Resistance Furnaces 20
5.2.1. Basic Principles 20
5.2.2. Reduction Resistance Furnaces . . . 22
5.2.2.1. Furnace Construction and Operating Parameters 22
5.2.2.2. Metallurgical Significance 24
5.2.3. Refining Resistance Furnaces 26
5.2.4. Solid-State Resistance Furnaces . . 28
5.3. Arc Furnaces 29
5.3.1. Basic Principles 29
5.3.2. Electric Arc Furnaces 29
5.3.3. Vacuum Arc Refining (VAR) Furnaces 30
5.4. Induction Furnaces 31
5.4.1. Basic Principles 32
5.4.2. Furnace Construction 33
5.4.2.1. Crucible Induction Furnace 33
5.4.2.2. Channel Induction Furnace 34
5.4.2.3. Special Induction Furnaces 37
5.5. Plasma Furnaces 37
5.6. Electron-Beam Furnaces 38
5.7. Fused-Salt Electrolysis Cells . . . 40
5.8. Modeling 42
6. References 42

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_339

1. Rotary Kilns

A rotary kiln is an inclined, rotating cylindrical reactor through which a charge moves continuously. The rotary kiln is used when thermal processing of solids that is more severe than drying is required [1]. The furnace walls (normally lined) make intermittent contact with the flue gas and the charge. Heat required for the various physical and chemical processes is delivered to the charge by lifting and overturning the charge as it moves through the interior of the rotary kiln [2].

The first proposal for using a furnace defined in this way is probably that found in an English patent from 1865 [1]. The most widespread application of this principle today, the production of cement clinker, began in 1885 [3], [4]. Nowadays, limestone calcining [138], iron ore reduction, and PbZn volatilization, mainly from electric are furnace dust [134], [135] are important too.

1.1. Design

1.1.1. General

The essential design elements of a rotary kiln are depicted in Figures 1 and 2. The rotary kiln consists of a lined hollow cylinder, mounted in an inclined position on rolls and rotated slowly by a drive. The charge material moves from the feed end to the discharge as a result of the rotary motion and gravity. The inclination is between 1.5 and 5 % and varies only in experimental kilns. Speed is between 0.2 and 2 rpm; variable-speed drives used to control the residence time are common.

Kiln diameter is usually constant over the full length. Diameters have increased to more than 7 m, especially in the cement industry; kilns for wet cement processing can be more than 200 m long.

Some rotary kilns have internals such as conveying or lifting flights, built in crossed-hanging link chains, or ring dams. In some processes, air-feed pipes or burner tubes for gas or oil are installed on the furnace shell. Air or other gases can also be introduced through ports in the lining.

1.1.2. Structural Elements

The *steel shell* of the kiln (Fig. 2, b) is conically tapered at the ends and may have other tapered sections. It experiences torsion due to the drive, and flexural stress due to its own weight and the weight of the lining and charge. Because of partial filling and pointwise support, the ideally circular shell cross section is deformed into an oval shape. The shell is designed in accordance with the laws of thin-shell statics or by approximation methods [5–12].

The *riding rings* (Fig. 2, a) that help support the furnace shell are one-piece steel castings. Field-welded riding rings have also been used [13]. Riding rings up to a diameter of 5 m are welded to the shell in some designs. Toothed riding rings are also common. Usually, however, especially in large-diameter units, riding rings are slipped loosely onto the thickened shell ring. The riding ring moves relative to the shell when in rotation. Correct sizing of the riding ring, play, and shell ring thickness are crucial for lining life in a kiln section; this represents a difficult design problem [14], [15]. Measurements of oval deformation with the "shell tester" aid the designer [16–18]. Careful installation and maintenance of kilns, most of which are supported at several points, are also important. The alignment of long rotary kilns is critical for the load distribution on the riding rings and the shell [19–21].

Two smooth *trunnion rolls* (Fig. 2, d) per riding ring, shrunk onto journals, generally turn in plain bearings with immersion or pressure-flow lubrication. Antifriction bearings have also been used.

One or more *thrust rolls* (Fig. 2, f) arranged parallel to the kiln axis, bear the downslope forces exerted by the kiln. These have hydraulic position adjustments so that the kiln can be moved in the longitudinal direction.

The kiln is rotated by a *gear ring* (Fig. 1, j) that is elastically attached to the shell and is driven by one or two pinions (depending on the drive torque, (Fig. 1, k). Frequency- controlled multiphase motors allow continuous control of rotational speed with very little loss of energy. In case of malfunction, mechanically or electrically actuated couplings can be used to engage an auxiliary drive (electric or internal combustion motor) to keep the kiln turning slowly and

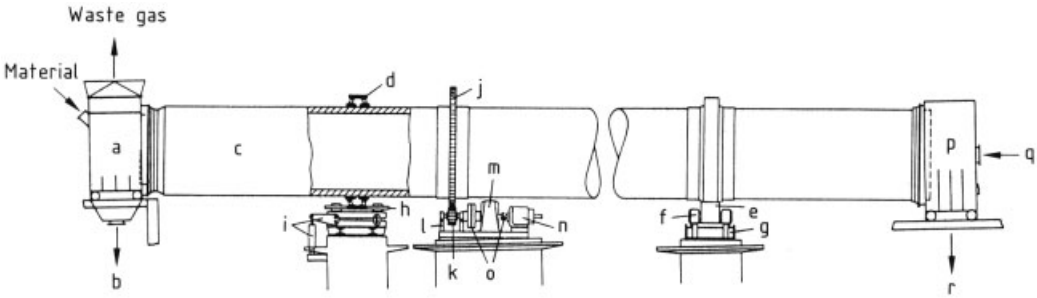

Figure 1. Schematic of a rotary kiln [50]
a) Kiln inlet head; b) Dust discharge; c) Rotary kiln; d) Riding ring; e) Roller; f) Roller bearing; g) Base plate; h) Thrust roller; i) Hydraulic system; j) Gear ring; k) Pinion; l) Pinion bearing; m) Gear box; n) Drive; o) Coupling; p) Kiln discharge head; q) Central burner; r) Discharge

prevent damage due to overheating on one side [21], [22].

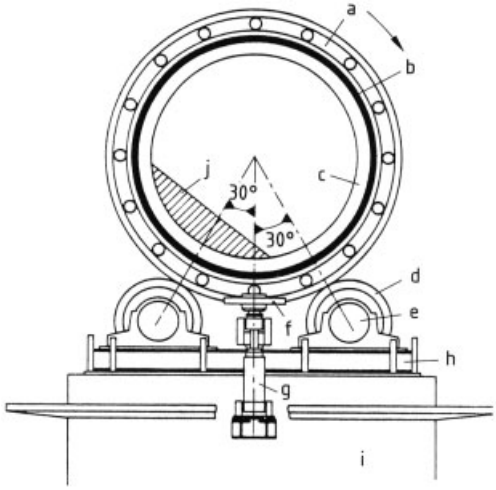

Figure 2. Section through kiln support [50]
a) Riding ring; b) Kiln shell; c) Lining; d) Trunnion roller; e) Roller bearing; f) Thrust roller; g) Hydraulic system; h) Base plate; i) Plinth; j) Charge

An important component of the rotary kiln is the *lining* (Fig. 2, c). Its thickness, physical properties, and chemical composition are dictated by the process to be carried out. For example, a cement kiln usually has a magnesite brick lining in the clinker zone and acid-insulating brick made from silicate compounds in the preheating zone. In iron ore reduction, low-iron alumina or magnesium – spinel bricks are required [23], [24]. An important factor determining the service life of the bricks is the mechanical stability of the shell [25]. Interlocking bricks (with tapered groove and tongue) are used in large-diameter rotary kilns [26], [27]. Qualities and lining techniques have improved greatly and the service life of linings has been extended significantly [28], [29]. Insulating liners would be very desirable for heat retention but have not proved useful in kilns with more than 3.6 m diameter.

1.2. Process Engineering

Movement of Material. The rotary kiln carries out several functions simultaneously: it is a device for conveying, mixing, heat transfer, and reaction. These functions must be in harmony. The charge in the kiln moves both radially and axially. Radial motion is determined by the degree of filling (percentage of cross-sectional area occupied by the charge) and the rotational speed. The angle of repose and the kiln inclination govern axial motion.

The interior of the charge tends to have a higher bulk density than the exterior, and grain size increases toward the outside [30]. This tendency can be counteracted by the internals, which also improve heat transfer into the charge. Dust production can be limited by pelletizing the feed.

Heat transfer occurs principally from the combustion gas (generated by a burner usually installed at the discharge end of the kiln) to the charge. The driving force is generally the temperature difference. The gas can move co- or countercurrent to the longitudinal motion of the charge. Cocurrent gas flow is advantageous only when the charge temperature does not have to

exceed a certain value. The countercurrent arrangement is preferred because it involves an increased total energy consumption. The different types of heat transfer in a rotary kiln are shown in Figure 3. Fundamental work is reported in [136], [137].

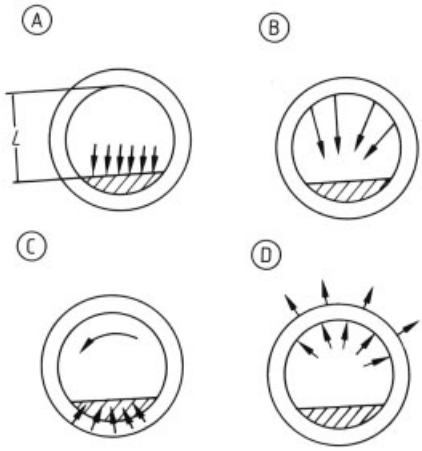

Figure 3. Types of heat transfer in a rotary kiln
A) Heat transfer to material by gas radiation and convection; B) Heat transfer to material by brick radiation; C) Conductive heat transfer to material from brick; D) Heat transfer to brick by radiation, convection, and heat loss by shell radiation and convection

1.3. Applications

1.3.1. Roasting and Calcining

Rotary kilns were employed for the exothermic roasting of sulfidic ores and for the endothermic removal of water of hydration and carbon dioxide from fine-grained materials such as ores, phosphates, alumina, ilmenite, and titanium dioxide. Today these processes are performed almost exclusively in fluidized-bed reactors (→ Fluidized-Bed Reactors), which offer better heat- and mass-transport conditions.

However, rotary kilns have advantages where softening, sticking, or even partial melting of the material cannot be avoided.

1.3.2. Calcination of Limestone, Dolomite, and Magnesite

Whereas coarsely crushed limestone can be converted to quicklime by heating in the shaft furnace, the rotary kiln is preferred for particle sizes ≤ 60 mm because it delivers a more uniform product at a higher rate (→ Lime and Limestone, Chap. 3.3.2.). This quicklime is particularly suitable for steelmaking and has been used in flue-gas desulfurization since the 1980s [31], [138].

Typical dimensions of rotary kilns with a quicklime capacity of 560 t/d are

Kiln without preheater	145 m × 3.45 m diameter
Kiln with vertical preheater	56 m × 4.0 m diameter

The rotary kiln also competes with the shaft furnace in the production of sintered magnesite and dolomite (→ Magnesium Compounds, Chap. 4.2.5.). As in limestone calcination, the rotary kiln offers the advantage of handling charge material of finer size with a broad size range (e.g., 2 – 60 mm) and can also handle throughputs at high calcination temperature (1600 – 1900 °C). A rotary kiln 170 m long with a shell diameter of 4.5 m, used for the production of sintered magnesite, is reported to have a capacity of 530 – 600 t/d. The same author indicates that the same mathematical relationships cannot be used for the design of cement, dolomite, and magnesite kilns [32].

1.3.3. Production of Cement Clinker

Cement is produced almost exclusively in rotary kilns (→ Cement and Concrete, Chap. 1.4.3.). World production was ca. 730×10^6 t in 1976 [33] and 1363×10^6 t/a in 1994 (USA 188×10^6 t/a, Europe 251×10^6 t/a). Developments in clinker production have, therefore, always spurred rotary kiln technology as a whole. Energy accounts for as much as 50 % of the total production costs of cement clinker [34]. Energy-saving approaches have led to thermally sophisticated, large- capacity units. The trend is also toward using coal as the energy source. Long rotary kilns with slurry feed (Fig. 4 A) have increasingly been replaced by "heat-exchanger" kilns (Figs. 4 B, C) in which the raw meal is heated in upstream suspension-type heat exchangers [36–38]. Development work with improved dust collection systems has led to energy consumption in the order of 3.35 MJ per kilogram of clinker.

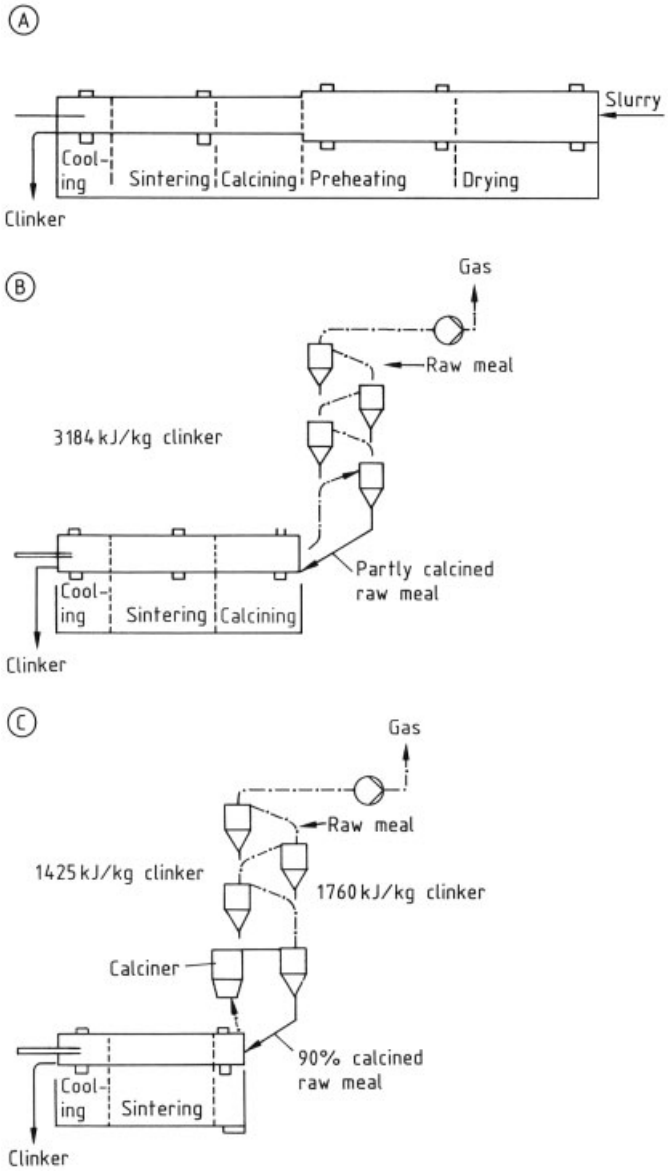

Figure 4. Kiln systems for cement clinker production of 2000 t/d [35]
A) Wet process; B) Dry process; C) Process with precalcining

1.3.4. Iron Ore Reduction

The rotary kiln should continue to grow in importance as a reduction apparatus. It can employ a wide range of carbon carriers not suitable as reductants for shaft furnaces (e.g., the blast furnace): from anthracite and coke breeze to charcoal fines, lignite, and brown coal.

The charge (ore and reductant) usually moves through the rotary kiln countercurrent to the hot gases. Coupled reactions – ore reduction by carbon monoxide and reaction of carbon with carbon dioxide with regeneration of carbon monoxide – occur in the charge. The reactivity of carbon is critical for the process as a whole [39]. Some of the carbon monoxide formed es-

capes from the charge so that oxidizing gases from the free kiln volume cannot permeate into it.

Reducing conditions depend on the temperature, reactivity, and quantity of the reductant; the residence time; and the charge holdup at the discharge end of the kiln. The rate of reduction can be controlled to meet a variety of objectives, from the formation of magnetite to the production of carburized molten pig iron. Details are sketched in Figure 5 and described briefly below.

In *magnetizing roasting*, the iron content of lean ore is transformed to magnetite (Fe_3O_4) at 700 – 1000 °C. The product is concentrated by grinding and magnetic separation. The temperature profile and reducing conditions can be controlled with gas- or oil-fired shell heaters located along the kiln; fuel is supplied to the rotating kiln through special seals. The reducing action can also be aided by the addition of small amounts of coal.

The most frequently used process for the production of sponge iron is the SL/RN process [40], [139], illustrated in Figure 6. Lump ore, pellets, titanomagnetite-containing sea-sand concentrates, zinc- and lead-containing residues from iron and steel production, and pellets from leach residues are reduced in existing plants (→ Iron, Chap. 2.6.4.).

1.3.5. Volatilization Processes

Metals and metal compounds with a high vapor pressure can be recovered by volatilization from their raw materials in a rotary kiln. The volatilization may be enhanced by selective reduction. Examples include arsenic, antimony, lead, cadmium, mercury, silver, tin, and zinc in elemental form or as compounds, for example, zinc as the metal and lead as PbS or $PbCl_2$.

The way in which the charge moves through the rotary kiln (a rolling motion called "wälzen" in German) has given its name to the *Waelz process* used for decades in the volatilization of zinc (→ Zinc) and lead (→ Lead, Chap. 4.2.2.3.). Low-zinc ores and other raw materials are fed together with solid reductants to the kiln, where the metal is volatilized under reducing conditions. The vaporized metal is oxidized in the free volume above the charge and recovered as dust (the so-called Waelz oxide) from the off-gas [41]. Figure 7 is a schematic of a modern Waelz plant. The Waelz process has also proved suitable for zinc- and lead-containing residues from iron and steel production [42]. A flow sheet of this process is shown in Figure 8. Higher productivity with less energy consumption and less mechanical dust formation can be achieved by pelletizing the feed. The Waelz kiln oxide can be purified by washing with Na_2CO_3 solution.

1.3.6. Other Applications

Other applications of the rotary kiln include:

1) Oxidation of ilmenite [43]
2) Calcination of pellets after preliminary heat treatment on a traveling grate (e.g., see → Iron, Chap. 2.1.1.5.)
3) Calcination of petroleum coke (→ Petroleum Coke, Chap. 3.2.) [44]
4) Reduction of heavy spar
5) Processing of gypsum to sulfuric acid and cement (gypsum – sulfuric acid process)
6) Production of TiO_2 pigment (→ Pigments, Inorganic, Chap. 2.1.3.1.)
7) Production of mercury (→ Mercury, Mercury Alloys, and Mercury Compounds, Chap. 4.1.3.)
8) Volatilization of zinc, lead, and copper with $CaCl_2$ [45]

Waste incineration and coal gasification may be important future applications of the rotary kiln.

2. Multiple-Hearth Furnaces

From a dominant position as a roasting furnace for sulfide ores (chiefly pyrites in sulfuric acid production), the multiple-hearth furnace has been almost completely displaced by fluidized-bed roasting equipment since the 1960s. Fluidized-bed devices (→ Fluidized-Bed Reactors) permit much higher throughputs than multiple-hearth furnaces, with substantially better control of reaction temperature and oxygen partial pressure in the roasting gas. Nonetheless, the multiple-hearth furnace will continue to find use in some special areas of process engineering.

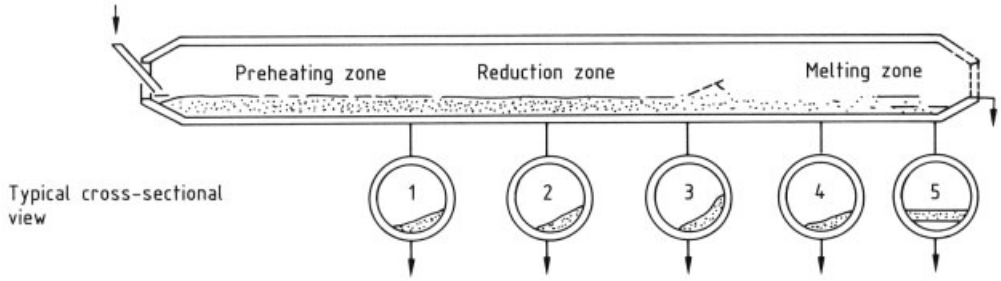

No.	1	2	3[a]	3[b]	4	5
Consistency of kiln discharge	solid	solid	solid	solid	semiliquid	solid (clinker) liquid (pig iron)
Preferred iron-content of ore, %	30-60		30-60	55-63	25-45	50-67
Size of ore feed, mm	<20	<20	<10	5-25[b]	<5	<0.2
Influence of basicity of charge	no influence	no influence	no influence	no influence	0.3	2.8-3.0
Maximal temperature of charge, °C	600-900	900-1100	900-1100	900-1100	1200-1300	1400-1500
Oxygen removal	12 (Fe_2O_3)	20-70	>90	>90	>90	100
Typical processes	Lurgi	Highveld Udy Larco Elkem	RN	SL/RN Krupp	Krupp-Renn	Basset

[a] % CaO/% SiO_2.
[b] Ilmenites and iron sand: 0.05-0.5 mm.

Figure 5. Characteristics and possibilities for the reduction of iron ores in rotary kilns

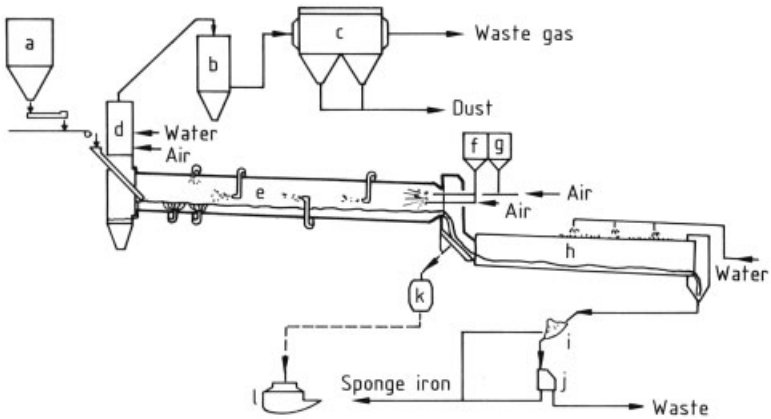

Figure 6. SL/RN process for sponge iron production
a) Bins for raw material; b) Waste-heat boiler; c) Aftercombustion chamber; d) Electrostatic precipitator; e) Rotary kiln; f) Coal bin; g) Fine-ore bin; h) Rotary cooler; i) Screening; j) Magnetic separation; k) Hot charging; l) Electric-arc furnace

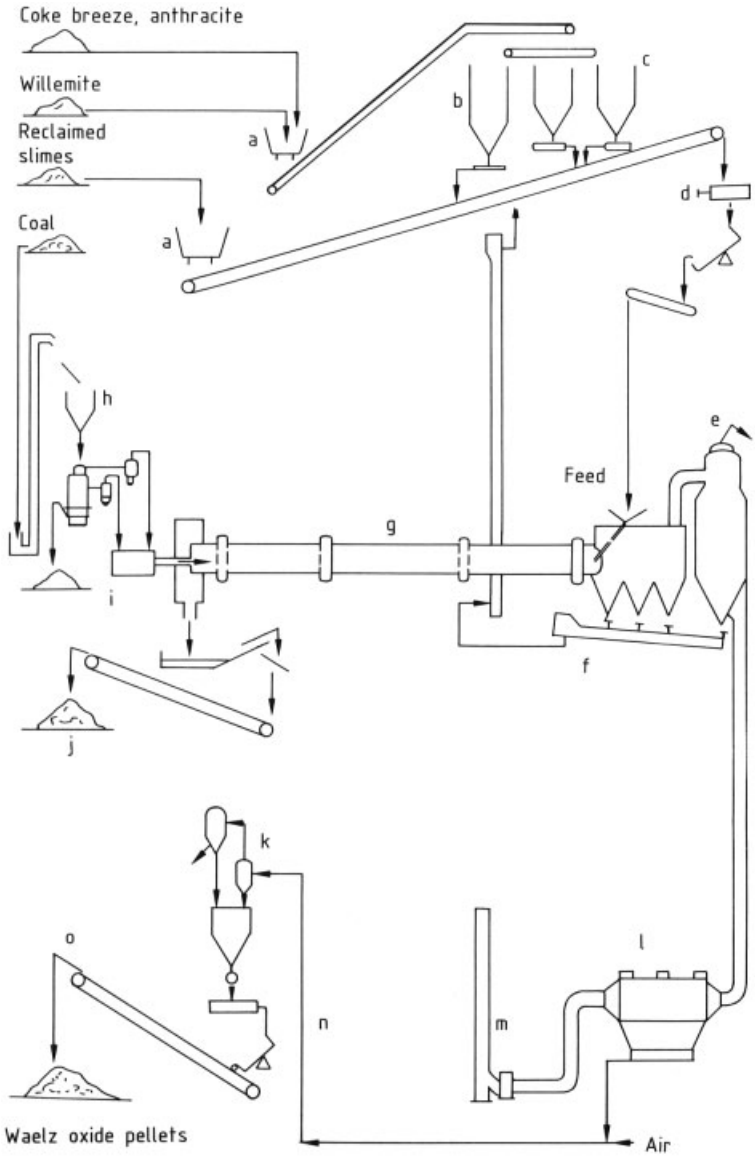

Figure 7. Flow sheet of a modern Waelz plant
a) Plate feeder; b) Willemite bin; c) Coke and anthracite bins; d) Mill; e) Cooling tower; f) Settling chamber; g) Waelz kiln; h) Producer gas plant; i) Burner; j) Waelz slag pit and scraper; k) Cyclones and bag filters; l) Two electrostatic precipitators; m) Waste-gas fan and stack; n) Pelletizing disk; o) Waelz oxide bins

2.1. Description

A multiple-hearth furnace (Fig. 9) consists of an internally lined steel cylinder with a number of horizontally mounted, lined platforms called hearths. The circular hearths are thinner near the center, which has an opening for a vertical shaft. An adjustable-speed drive with overload protection turns the shaft at 0.2 – 5 rpm. From one to four rabble arms per hearth are latched to the shaft in a gastight manner. These arms bear oblique stirring teeth to move the solids over the hearth; on one hearth the motion is from center to edge, on the next from edge to center depending on the inclination of the stirring teeth.

The openings in the hearths, through which the charge travels from the top of the furnace to the bottom, therefore alternate from central to peripheral.

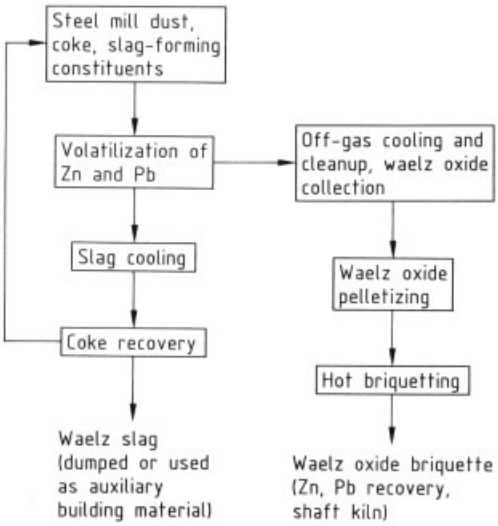

Figure 8. Flow sheet of Waelz process with steel mill dust as feed and briquetting of product

Because of the high temperature in the furnace, the shaft and rabble arms are air cooled. The shaft has double walls; cool air supplied by a fan enters the outside space, passes through the shaft and arms, and exits the furnace at 200–300 °C by way of the center space. Each hearth has several doors, which allow monitoring of the reaction and replacement of the rabble arms. The doors can be sealed tightly or can have adjustable air slots to admit cooling or combustion air if a slight subatmospheric pressure is maintained in the furnace.

Multiple-hearth furnaces are built in various sizing, ranging from 2 to 8 m in diameter and having 3–16 hearths. Residence time in the furnace is easily controlled by varying the shaft rotation speed or the number of rabble arms and teeth on each arm. Suitable reactions for the multiple-hearth furnace are

1) Slow reactions (because long residence times can be achieved in the multiple-hearth furnace)
2) Reactions between solids and quantities of gas that are too small to maintain a fluidized bed
3) Processes in which the solid is inlet in slurry form and the slowest and most gentle drying possible is desired
4) Processes in which solids must be exposed to a stepwise varying reaction temperature during thermal processing (within certain limits, the temperature and the gas atmosphere can be varied from hearth to hearth)
5) Reactions in which the solid undergoes slight softening, agglomeration, or sintering so that fluidized-bed processes cannot be employed

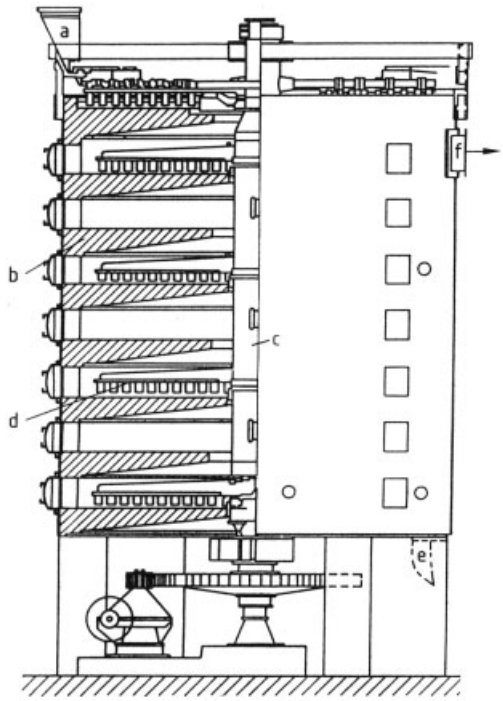

Figure 9. Lurgi multiple-hearth furnace [50]
a) Hopper; b) Circular hearth; c) Hollow shaft; d) Rabble arm with teeth; e) Discharge outlet; f) Gas outlet

2.2. Applications

2.2.1. Roasting of Sulfide Ores

Since the development of fluidized-bed techniques, sulfide ores such as pyrite, pyrrhotite, zinc blende, and chalcopyrite have not been roasted in multiple-hearth furnaces. Ores that can be roasted only with great difficulty, such as molybdenum disulfide, are still treated in multiple-hearth furnaces.

Because roasting reactions are exothermic, the furnace usually has to be heated only at the start of the process. The material fed to the topmost hearth is distributed by the teeth on the rabble arms, slowly transported to the center of the hearth, and dried. Then the ore falls into the first roasting zone, where it is heated in contact with hot roasting gas until it ignites. The reaction goes to completion as the charge is transported further over the hearths. On the last hearths, roasting air drawn or blown into the furnace from the bottom is preheated by cooling the residue. The progress of the reaction is monitored by measuring the temperature on the individual hearths.

The roasting of molybdenum disulfide (Eqs. 1 – 3) is carried out at ca. 630 °C (→ Molybdenum and Molybdenum Compounds, Chap. 4.2.). The hearths on which the exothermic roasting reactions (1) and (2) occur are cooled by the admission of air or water.

$$2\,MoS_2 + 7\,O_2 \longrightarrow 2\,MoO_3 + 4\,SO_2 \quad (1)$$

$$MoS_2 + 6\,MoO_3 \longrightarrow 7\,MoO_2 + 2\,SO_2 \quad (2)$$

$$2\,MoO_2 + O_2 \longrightarrow 2\,MoO_3 \quad (3)$$

After the roasting reaction has gone largely to completion, the molybdenum dioxide product is oxidized to the trioxide (Eq. 3). This reaction, however, is not sufficiently exothermic to keep the temperature high enough to burn off the residual sulfur. In a furnace with a total of nine hearths, the seventh and eighth are therefore heated with oil or gas.

2.2.2. Use for Endothermic Reactions

For endothermic reactions such as the calcination of a magnesite, dolomite, bauxite, clay, or zinc carbonate, the heat required is delivered from outside into the furnace. If the feed materials can tolerate some overheating by flame radiation, one or more burners can be built into the furnace shell at the hearths where heating is required (see Fig. 10). Otherwise, hot flue gas is admitted at these locations after generation in one or more combustion chambers. Heat losses are higher and throughput capacities lower than those of fluidized-bed furnaces and rotary kilns, so the calcination reactions mentioned above are rarely carried out in multiple-hearth furnaces. The use of this type of equipment is restricted to residues containing vanadium and tungsten.

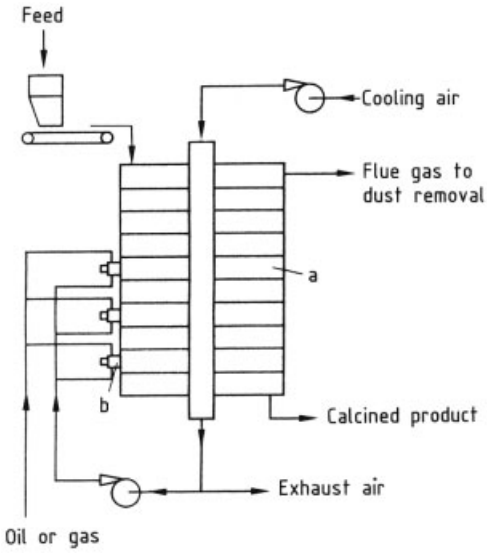

Figure 10. Heating of individual hearths with jacket burners [50]
a) Multiple-hearth furnace; b) Burner

Most *vanadium* is produced from vanadium-containing slags such as those from steelmaking. Because most of the vanadium in slag is in the tri- and tetravalent forms, which are insoluble in water, it must be converted to water-soluble pentavalent sodium vanadate by thermal oxidation and the addition of Na_2CO_3 and NaCl. This process is carried out in the multiple-hearth furnace (→ Vanadium and Vanadium Compounds).

The multiple-hearth furnace can also be used to oxidize *tungsten- containing slags* or residues from the metal fabrication industry (turning chips, grinding waste) in order to recycle the material in the form of WO_3 to tungsten production.

2.2.3. Clarifier Sludge Treatment

The flexibility of the multiple-hearth furnace led, in the 1980s, to its increasing use for the incineration of waste from a variety of sources. The main products treated by incineration are sludges from wastewater treatment and from cellulose and paper mills.

Today, most sludge incinerators built around multiple-hearth furnaces are designed so that combustion is self-sustaining, i.e., no fuel needs to be added. The sludge is therefore prepared by mechanical dewatering in centrifuges or filter presses.

If the degree of dewatering is insufficient for self-sustaining combustion, a thermal drying step is interposed. For sludge or waste incineration, the multiple-hearth furnace is equipped with an afterburning chamber (f) and gas vapor recycle (g) (see Fig. 11). The afterburning zone is needed for the residue-free combustion of volatile components at 800 – 900 °C. Recycling of vapors and flue gases by means of a fan thus helps by minimizing pollution and controlling temperature.

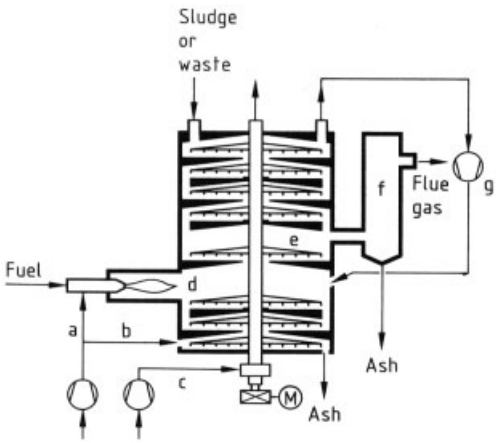

Figure 11. Multiple hearth for sludge incineration
a) Combustion air; b) Cooling air for ash; c) Cooling air; d) Startup burner; e) Multiple hearth; f) Afterburning chamber; g) Recycle fan

The use of multiple-hearth furnaces for sludge incineration led to the development of combined multiple-hearth – fluidized-bed (MHFB) furnaces in 1980. Furnaces of this type (Fig. 12) have the advantages of both the multiple-hearth furnace (slurry feed) and the fluidized-bed furnace for combustion processes.

In principle, such a device is a fluidized-bed furnace with a few hearths added on top as needed to predry the sludge. Flue gases ascending from the fluidized bed are drawn through the hearths in countercurrent to the slurry feed. About 50 % of the water is vaporized outside the fluidized bed, and the sludge is thus predried. Vapor produced by predrying can be recycled to the burnup zone of the fluidized-bed. One advantage of this arrangement, compared to the conventional fluidized-bed furnace with slurry feed, is that the air excess can be reduced to 30 – 40 %, so neither supplemental fuel nor external predrying is usually necessary (given an appropriate degree of dewatering).

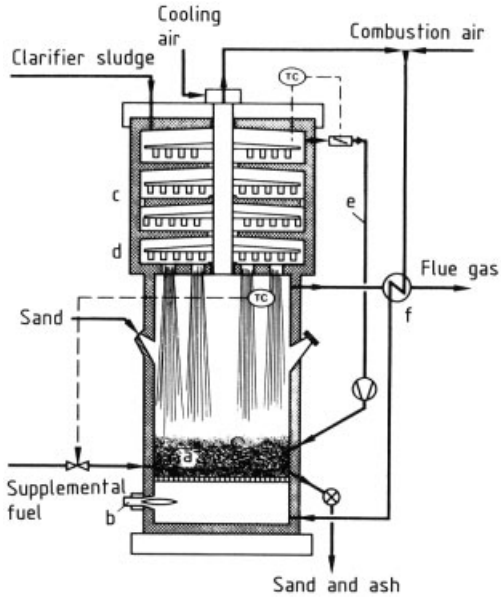

Figure 12. Combined multiple-hearth – fluidized-bed furnace (MHFB)
a) Fluidized bed; b) Startup burner; c) Drying hearths; d) Distribution hearths; e) Gas – vapor recycle; f) Preheater

3. Shaft Furnaces

The shaft furnaces have a variety of uses in metallurgy. The *Rachette furnace* is employed in lead production (→ Lead, Chap. 4.2.2.2.). Figure 13 shows a chair jacket profile used today. The distinguishing feature of the lead blast furnace is that the throat widens upward. The products, lead bullion and slag, are obtained by reduction; the slag contains any zinc present as ZnO. The *Imperial Smelting furnace* (Fig. 14) was developed for the simultaneous processing of lead and zinc concentrates (→ Zinc). It differs from ordinary lead blast furnaces by having a hot stack and a zinc splash condenser.

In the following, the *iron blast furnace* (→ Iron, Chap. 2.5.) is described in more detail as it is the most important type of shaft fur-

nace. The blast furnace has the form of two truncated cones set with their large bases together. The widening of the stack from top to bottom reduces the frictional resistance as the burden moves downward.

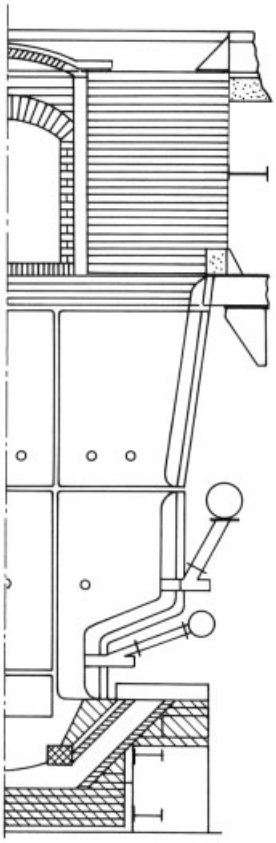

Figure 13. Profile of a lead blast furnace

A typical iron blast furnace is shown in Figure 15. Modern blast furnaces have a hearth diameter of 10 – 14 m, a volume of 2000 – 4000 m^3, reach throat gas pressures of up to 0.35 MPa, and have production capacities of $(1.5 - 3.5) \times 10^6$ t/a of pig iron. The blast furnace is divided into the following sections: throat, stack, bosh, and hearth.

The blast furnace is charged with burden, coke, and additives via the charging platform (throat). The blast-furnace gas or top gas is also withdrawn from the furnace through the throat. For design reasons, the throat diameter should be kept as small as possible. The *stack* is conical and terminates in the throat at the top. It expands downward to the *belly*. The next section down is the *bosh*, which narrows toward the bottom and forms the transition to the *blast-furnace hearth*. The narrowing of the bosh correspondsto the decrease in burden volume as the coke is consumed. This cylindrical section has water-cooled *tuyeres* through which the blast is introduced into the furnace. The hearth should have a relatively small diameter because the hot blast does not have a long penetration depth and otherwise would not reach the interior of the burden. The two main products of the furnace, slag and iron, collect in the hearth. Excess slag is withdrawn through slag holes, while the hot metal which has a higher density, is withdrawn through tap holes.

The main function of the iron blast furnace is to reduce iron oxides to metallic iron. The reducing melting of iron- containing substances takes place in a countercurrent arrangement; the ascending hot gas from coke combustion heats and reduces the descending burden.

The dimensions of a modern "high-top-pressure" blast furnace are as follows

Hearth diameter	11.5 m
Belly diameter	12.6 m
Throat diameter	8.4 m
Total hearth height	6.23 m
Distance from stool to tap hole	2.38 m
Distance from tap hole to slag hole	2.05 m
Distance from tap hole to tuyere level	3.25 m
Bosh height	3.30 m
Belly height	2.30 m
Shaft height	18.35 m
Throat height	2.75 m
Usable capacity (from tuyere level to 1 m below large bell in let-down position)	2.353 m^3
Total capacity	2.820 m^3
Diameter of large bell	6.10 m
Diameter of small bell	2.20 m
Number of tuyeres	28
Number of tap holes	3
Number of cooling plates in tuyere level and bosh	476
Number of stack cooling boxes	800

The throat of this modern blast furnace is closed by a dual- chamber structure consisting of a lower bell, a small bell, and four clack valves. The furnace consists of a fully armored steel structure: a multistory frame built on a square-plan portal frame. The furnace frame accepts the loads from the furnace top, the four gas offtake pipes, the over-throat construction, the bustle

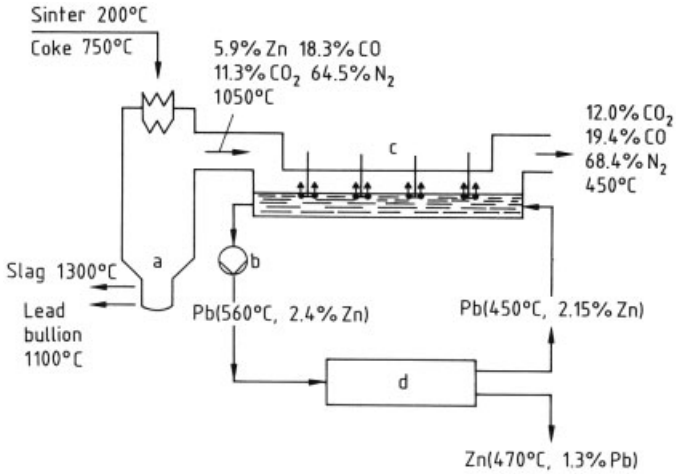

Figure 14. Imperial Smelting blast furnace with zinc condensation by molten lead [50]
a) Blast furnace; b) Pump; c) Spray condenser; d) Cooling system

pipe, and the working platforms. Compensators take care of length changes: one compensator is located between the upper furnace and the furnace end; three articulated compensators are located in each of the four gas offtake pipes. This blast furnace is designed for the production of 6500 t/d of pig iron [46–48].

Furthermore, metallurgy shaft furnaces are used for

1) Anode copper refining (→ Copper, Chap. 6.1.)
2) Pure copper melting in cast shops in ASARCO shaft furnaces (→ Copper, Chap. 6.3.)
3) Recovery of copper from secondary materials (→ Copper, Chap. 5.9.).
4) Cast iron melting in Cupola furnaces (→ Steel, Chap. 3.5.5.2.).

For shaft furnaces, ores, concentrates, or metals must be lumpy or be agglomerated by sintering, pelletizing, or briquetting. To some extent, fine ores or secondaries can be blown through the tuyeres directly into the raceway for very fast reduction.

4. Smelting, Melting, and Refining in Bath and Flash Smelting Reactors

Fine concentrates or secondaries can be smelted without agglomeration in flash (after drying) or bath smelting furnaces. The heat of exothermic chemical reactions provides the energy for autogenous smelting. All these furnaces are operated continuously.

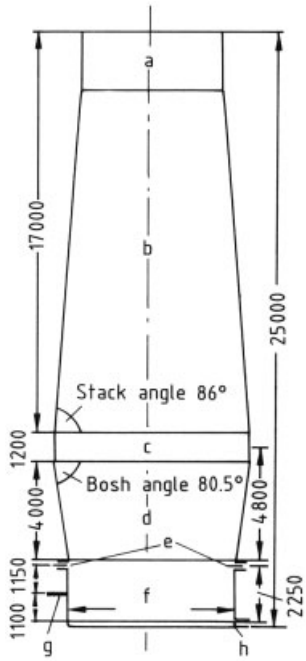

Figure 15. Schematic of a typical iron blast furnace with common heights (in mm) and angles
a) Charging platform (throat), b) Stack; c) Belly; d) Bosh; e) Tuyeres; f) Hearth; g) Slag hole; h) Tap hole

A wide variety of such smelting furnaces and converters are used in metallurgy. These smelting furnaces represent a modern approach of intensive metal smelting under environmentally "clean" conditions.

4.1. Bath Smelting Furnaces

In bath smelting furnaces and converters oxygen or oxygen-enriched air is blown into molten metal or matte baths via tuyeres (Fe, Cu), lances (Fe, Ni, Pb) or injectors (Fe) to oxidize elements that are to be removed as impurities. Coal or reducing gases may also be blown into molten slags via tuyeres, lances, and injectors for slag reduction (Pb). In the case of aluminum, chlorine is blown through pipes and stirrers into the molten metal bath to remove alkaline and alkaline-earth elements.

Bath smelting furnaces have been described in detail in the Encyclopedia and elsewhere [107], [140–148]. They are already in use for the production of copper by the Mitsubishi (→ Copper, Chap. 5.7.2.) and Isasmelt/Ausmelt (→ Copper, Chap. 5.4.4.) processes; the production of copper matte from concentrates and secondary materials in the Noranda (→ Copper, Chap. 5.7.1.), CMT/Teniente (→ Copper, Chap. 5.4.6.), and Vanyukov (→ Copper) processes; and for the production of lead bullion following the QSL and Isasmelt technologies from concentrates and secondary materials (→ Lead, Chap. 4.4.3.). Bath smelting furnaces are in development for the direct steel production from iron ores (→ Iron, Chap. 2.7.2., → Iron, Chap. 2.7.3.).

4.2. Flash Smelting Furnaces

In 1949 the first Outokumpu flash smelter started operation. By now 36 Outokumpu flash smelters treat copper concentrates, and six smelt nickel concentrates; 50 % of the global primary copper and 30 % of the nickel are flash smelted in Outokumpu-licensed plants. In addition, four Inco flash smelters and three Kivcet furnaces are in operation.

The Outokumpu furnaces consist of

- A circular reaction shaft (7 – 9 m high, 6 – 7 m in diameter) for roasting and smelting of dry concentrates in suspension with highly enriched air
- A settling hearth with an area of 120 – 160 m^2 for collection of the droplets and separation of matte (metal) and slag
- An off-take shaft for waste gas and flue dust.

More details are given in the Encyclopedia (→ Copper, Chap. 5.5.1., → Nickel, Chap. 4.2.1., → Lead) and elsewhere [107], [108], [141–144], [149].

4.3. Bath Melting and Refining

Bath melting and refining furnaces are often used for melting, refining, and alloying metals. The furnace size ranges from 1 t to several hundred tons. These furnaces are operated batchwise and are fed with solid and liquid metal. They are of the stationary, tilting, or rotary type.

Their applications include:

1) Electric arc furnaces and induction furnaces for stell and cast iron melting and refining (→ Steel, Chap. 3.5.5.2.)
2) Stationary and tilting hearth furnaces and rotary for aluminum melting, refining, and alloying
3) Stationary hearth furnaces for copper matte and ferronickel smelting
4) Rotary furnaces for anode copper melting and refining
5) Stationary and tilting hearth furnaces for copper scrap melting and refining
6) Rotary furnaces for secondary lead smelting

An example of a gas-heated aluminum rotary furnace for melting and refining of Al scrap is shown in Figure 16 [150].

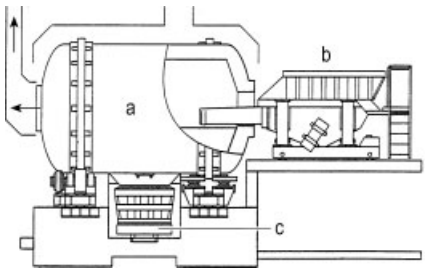

Figure 16. Conventional rotary furnace for melting Al scrap a) Rotary furnace; b) Loading machine; c) Salt slag bucket

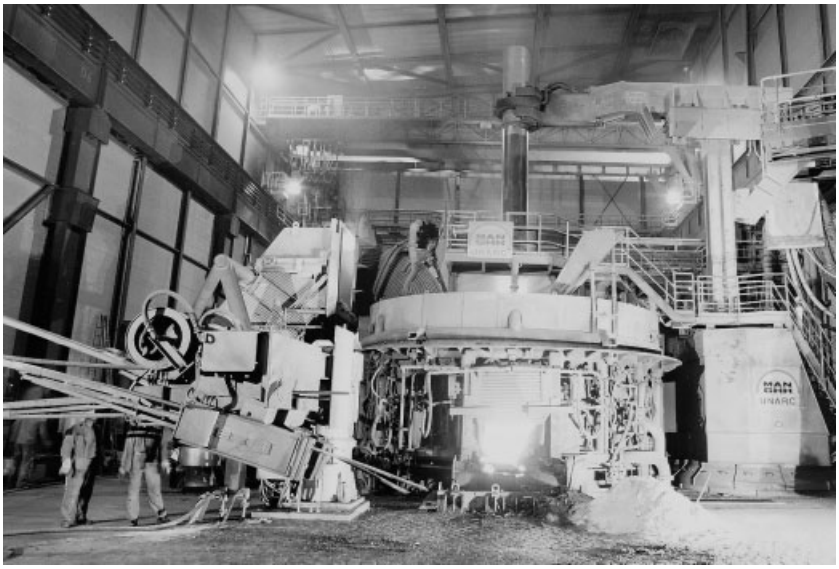

Figure 17. 100 t-steel melting direct-current arc furnaces, type UNARC (courtesy of MAN GHH).

Figure 17 shows two modern gas-heated tiltable aluminum melting and casting furnaces. For mixing the melt, such furnaces are stirred mechanically, inductively, or by gas.

4.4. Converters

Converters are mainly used for

1) Conversion of pig iron together with scrap into steel (→ Steel)
2) Conversion of copper matte into blister copper (→ Copper, Chap. 5.6.) and the refining of secondary black copper (→ Copper, Chap. 5.9.)
3) Refining of secondary aluminum
4) Conversion of nickel matte into nickel (→ Nickel, Chap. 4.2.2.) and lead matte (→ Lead, Chap. 4.4.2.2.)

Copper is up to now the only metal continuously converted from matte to blister copper on an industrial scale in several plants [140–144], [149], [151] (→ Copper, Chap. 5.7.).

The Mitsubishi converter has been operated for more than 30 years on liquid matte feed with lance blowing [140–144], [151]; the Noranda converter is a newer development [141], [151] and is operated with tuyeres.

Another development is flash converting of ground copper matte to liquid blister copper with oxygen. The first industrial plant started operation a few years ago in the United States. A second plant, in Peru, will follow soon [132], [140], [141], [151].

The main converting operation is steelmaking with ca. 450×10^6 t/a in 1996. Converters reach sizes of up to 400 t per batch, with annual capacities of up to 4×10^6 t. Molten iron together with steel scrap (for cooling) is transformed to steel in the converter. Accompanying elements, particularly carbon, are oxidized to a desired level by oxygen.

In the past, air was used for converting in the Bessemer (since 1855) and Thomas (since 1877) converters. Nowadays only oxygen converters are in use (Fig. 18) which employ lances for top blowing (e.g., LD, LD – AC, Kaldo) and injectors for bottom blowing [e.g., OBM, Q-BOP] or side blowing (AOD). See also → Steel).

The *LD process* (see Fig. 18 A) was developed in 1952 in Linz and Donawitz, Austria. Oxygen is blown onto the bath of molten pig iron from above through a water- cooled oxygen lance. The LD process can handle pig iron containing up to 0.5 % phosphorus. The *LD – AC variant* (A for Arbed, Luxembourg; C for CNRM, Liège, Belgium, Fig. 18 B) has been developed for high-phosphorus iron. Multiple

slagging makes it possible to dephosphorize the melt.

The Bottom-Blown Oxygen Process (BOP) is called the OBM (Oxygen-Bottom-Maxhütte) process in Europe (Fig. 18 C), and is generally used with high-phosphorus pig iron. In the United States and Japan it is called the Q-BOP process and is used with low-phosphorus pig iron. In the bottom-blown converter developed in 1968, oxygen is admitted from below to the reaction space through a number of tuyeres (5 to 15). The oxygen jet is surrounded by a protective medium containing hydrocarbons, which serves to cool the tuyeres. The bottom-blown process features short blowing times and low iron losses.

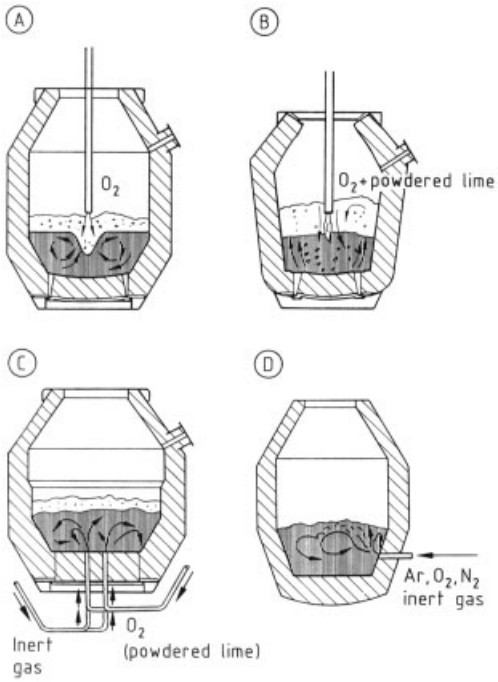

Figure 18. Different types of converters
A) LD converter (BOF); B) LD – AC converter; C) OBM converter; D) AOD converter [131].

The *AOD (Argon-Oxygen-Decarburization) converter* (Fig. 18 D) permits oxidation and degassing of steel before casting. Figure 19 shows working positions of a converter. Loading always takes place through the mouth of the converter. Steel is tapped via a taphole situated in the upper side by tilting the converter.

A modern development of melting and converting of scrap is the *energy-optimizing furnace (EOF)* which combines preheating, melting, side blowing, and direct heat recovery from the reactor gases [52]. Plants with batch weights of 30 – 100 t have been built. For the 30 t reactor in Pains, Brazil some consumption and performance figures (per tonne of molten steel from 100 % scrap) can be given [154]:

Oil	3 – 5 kg
Coke	85 – 100 kg
O_2	110 – 135 m^3 (STP)
Tap-to-tap time	65 min
Blowing time	40 – 50 min
Tap temperature	1680 °C
Refractory consumption	6 kg

5. Electrothermal Reactors

5.1. Introduction

5.1.1. History [54–62]

The use of electricity for metallurgical applications started in at the beginning of the nineteenth century. As early as 1810, Davy performed experiments in which he produced alkali metals via *fused salt electrolysis*. In 1888 Héroult patented a small electric furnace for the production of aluminum, which formed the basis of the present-day electrolytic recovery of aluminum from alumina [54]. The use of *induction* as a heating method was patented by de Ferranti in 1887 [54]. Although the experimental use of an electric arc can be traced back to 1810, the development of the first *electric arc furnace* is attributed to Siemens, who developed a small arc furnace in 1878 – 1879 [54]. From this humble beginning, the use of an arc for melting and smelting in furnaces progressed from an arc furnace (developed by Héroult in 1900) to the ultrahigh power (UHP) technology (developed in 1960 – 1962) [55], [56], with which very high melting rates can be attained.

The earliest recorded application of *electroslag* refining appears in a book by Slavyanov published in 1892 [57], but reported work performed in the United States could be dated 20 years earlier. By the time of Hopkins' development of the electroslag process in 1935, consumable vacuum arc refining was making great headway in the United States [57].

Vacuum melting

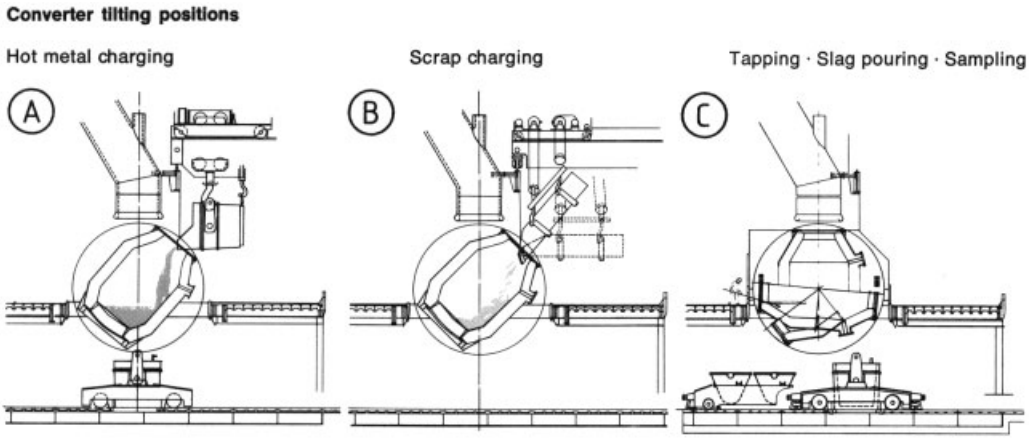

Figure 19. Working positions of a converter [131]

was introduced as a plant-scale operation in 1917 when ROHN melted nickel-base alloys by resistance heating. In 1923, Heraeus Vacuumschmelze (Hanau) was founded that operated vacuum induction furnaces; by about 1956 the potential of vacuum arc remelting for steels, as well as nickel- and titanium-base alloys had been exploited for the production of improved gas turbine disks, shafts, and casings [58].

The use of electron-beam technology for smelting and melting is as young as plasma metallurgy, although experiments with electron beams (at the time known as cathode rays) commenced as early as 1852 [59]. The conceptual birth of this technology is attributed to VON PIRANI, who patented his ideas in 1907. No further significant developments occurred until semicontinuous electron-beam melting was first performed in a cold mold crucible by SMITH and coworkers in 1954, who attempted to consolidate titanium powder [59]. By 1957 facilities were available for processing titanium ingot 80 mm in diameter and 1.5 m in length.

Plasma metallurgy was merely a topic of science fiction up to the 1950s [60]. The first investigations into carbothermic reduction of oxides were carried out in the late 1950s [61]. In a 1962 publication the use of plasma torches for melting scrap was traced back to the Linde Company, a division of Union Carbide [62].

5.1.2. Electrothermal Furnaces [55–89]

The short historical overview shows that most of the basic principles of electrothermal furnaces have been known for a considerable time. Electrothermal furnaces are classified on the basis of these principles as follows:

Resistance Furnaces. In these furnaces heat transfer occurs either directly or indirectly according to Ohm's and Joule's laws. They can be operated in an a.c. or d.c. mode.

Applications: Furnaces which implement *direct resistance* include:

1) Resistance furnaces for solid-state reactions, e.g., production of graphite and carbides [63].
2) Resistance furnaces for the production of copper, nickel, iron, tin, and zinc or their intermediate products from oxidic and sulfidic materials and slags [64–73].
3) Electroslag refining (ESR) for the production of clean ferrous or nonferrous metals and alloys such as titanium, steels, and superalloys, [57], [73], [154].
4) Furnaces that employ a combination of arc and resistance heating are used for the production of calcium carbide, ferroalloys, pig iron, phosphorus, and silicon compounds by reduction. Electrodes are situated within the charge; reduction and melting occur within the high-temperature zone around the electrodes [74], [75].

5) Fused-salt electrolysis cells for the production of aluminum and alkali metals may also be considered to be electrochemical thermal reactors [76], [77].
6) Salt-bath furnaces for the heat treatment of metals.

Indirect heating is mostly performed in a closed chamber in which heating elements, situated in the walls or suspended in the chamber, transfer their heat via radiation, conduction, and convection to the object to be heated. This type of furnace is applied for heat treatment of metals, ceramics, and for the production of carbides [55], [56], [78]. These types of furnaces are not discussed in this article.

Arc Furnaces. In arc furnaces the heating occurs because of the high temperature (> 6000 °C) of a gas plasma created by the arc. These furnaces may be operated either in an a.c. or d.c. mode.

Applications:

1) Steel scrap and cast iron scrap (in the past also copper scrap) are melted in arc furnaces in which the arc is situated above the molten slag. The slag acts as a heat distributor and a refining agent for the molten metal [154].
2) In vacuum arc remelting (VAR) a consumable electrode, made from the metal to be melted/refined, is melted in a vacuum to produce high-purity metals. An example is the production of high-purity titanium from titanium sponge [58], [79], [154].

Induction Furnaces. In induction furnaces the heat is generated according to Lenz's, Ohm's, and Joule's laws. Two types of furnaces are generally used for melting operations the *crucible furnace* and the *channel furnace* [80], [81], [94], [95], [154].

Applications:

1) *Crucible furnaces* are used mainly to melt particulate cast iron, (stainless) steel, base metals, copper, and aluminum.
2) *Channel furnaces* are applied in the ferrous industry as a holding furnace. In the nonferrous industry they are also used as a melting furnace for aluminum, zinc, and copper.
3) In *vacuum induction furnaces* volatile components are removed from less volatile metals and alloys. Examples are the refining of special steels, superalloys, and copper. Silver crusts from lead refining are also enriched in silver by volatilization of zinc.

Induction heating is also used during annealing, hardening (→ Steel), brazing, soldering (→ Soldering and Brazing), and longitudinal welding (→ Welding and Cutting).

Electron-Beam Furnaces. In electron-beam furnaces electron guns produce high-energy electrons, which impart their energy to the furnace charge to affect its melting [58], [59], [155], [156].

Applications. Electron-beam furnaces are used to melt and/or refine refractory metals such as vanadium, niobium, and tantalum; metals such as molybdenum and tungsten; reactive metals such as zirconium and hafnium; and ceramics such as zirconia, alumina, and uranium carbide.

Plasma Furnaces. In these furnaces heating is performed by either transferred or nontransferred arc plasma torches (→ Plasma Reactions). In contrast to arc furnaces, where arcs are produced between a graphite electrode and the charge, plasma furnaces use gas-stabilized plasma arcs produced by water-cooled plasma torches with nonconsumable electrodes. Consumable hollow carbon electrodes have also been applied [60–62], [82], [83].

Applications. Plasma furnaces have been employed commercially, e.g., to produce ferrochromium, to melt steel scrap, and to recover valuable metals from steel flue dust [125].

Laser Heating. Although "laser furnaces" are nonexistent, many applications of laser heating have been reported in metallurgy, e.g., in cutting, welding, machining, and surface treatment of metals [55], [56].

Miscellaneous Types of Heating. In addition to the heating methods discussed above, infrared, dielectric, and microwave heating are also available. As these are not presently used to effect a chemical reaction at high temperature on an industrial scale, they fall outside the scope of this article.

Table 1. Recovery of metals in selected electrothermal furnaces as a function of the type of reduction, reductant, and heating method

Metal	Type of reduction	Reductant	Type of heating
Aluminum/magnesium	fused-salt electrolysis	electrochemical	resistance
Copper/nickel	matte smelting		resistance
Ferroalloys (Cr, Mn, Si, Ni)	carbo-/silicothermal	C/SiMe	resistance
Flue dusts (Zn, Pb)	carbothermal	C	plasma, resistance
Magnesium	silicothermal	FeSi	resistance
Scrap/reduced iron pellets (Fe)		C	arc
Slags (Cu, PbZn)	slag smelting	C	resistance
Phosphorus, $CaSi_2$, Ca_2C	carbothermal	C/Ca_2C	resistance arc

Table 2. Metal refining operations

Metal	Type of furnace	Type of heating	Removed impurities
Steel	arc furnace	arc	C, O, and diverse impurities
Steel	electroslag	resistance	inclusions
Ti, Zr	vacuum arc	arc	H, Mg, Cl
V, Nb, Ta, Hf, Zr	electron beam	electrons	C, O, N, H, and diverse impurities
Superalloys, special steels	vacuum induction	induction	C, O, N, H, and diverse impurities

Table 3. Advantages and disadvantages of electrothermal furnaces

Specific property	AF [a]	RF [b]	IF [c]	ESR [d]	VAF [e]	VIF [f]	EBF [g]
High power density	+	+	+			+	+
Very high temperature	+	+					+
Very good heat transfer		+	+	+		+	
Good mixing		+	+			+	
Water-cooled inert crucible				+	+		+
Controlled atmosphere	+	+	+	+	+	+	+
Very low p_{O_2}			+		+	+	+
Low off-gas volume	+	+	+	+	+	+	+
High specific energy consumption				+			+
Low smelting rate				+			+

[a] AF = arc furnace
[b] RF = reduction furnace
[c] IF = induction furnace
[d] ESR = electroslag refining
[e] VAF = vacuum arc furnace
[f] VIF = vacuum induction furnace
[g] EBF = electron-beam furnace

Fields of Applications of Electrothermal Furnaces. A number of applications have already been given [78], [84–89] and further general information on electrothermal furnaces may be obtained from [93–96]. Electrothermal furnaces may be used for the *production of metals from raw materials by reduction*. Table 1 gives a summary of the applications of the above-mentioned techniques according to the metals produced.

Electrothermal furnaces may also be applied during metal refining. A summary of selected furnace types and their application during the *refining of specific metals* is presented in Table 2. The removed impurities are also given.

If cheap electrical energy is available heating in electrothermal furnaces has many advantages over heating with fossil fuels (Table 3).

Specific advantages associated with particular furnaces used to remelt certain metals are summarized in Table 4.

5.1.3. Energy Consumption and Production Capacities

Typical energy consumption for the production of various metals and alloys are given in Table 5 as a function of the type of furnace. The total energy consumption is a very important parameter for optimization for economic reasons. Elec-

trothermal reactors are also applied to produce nickel and ferronickel. Table 6 gives the energy consumptions for various ore compositions. The large content of platinum group metals (PGMs) in the ore of the Rustenburg operation in South Africa makes this process economically viable despite its high energy consumption.

Table 4. Remelting of metals and the advantages of specific furnaces

Metal	Type of furnace	Reason for application
Steel	arc	high melting capacity
Steel, cast iron, Al alloys, Cu alloys, Zn	induction	control of the atmosphere and the alloy composition
Ti, Zr	vacuum arc	no crucible and gas reactions

5.1.4. Economic Aspects

World production and energy consumption data for the production of metals, alloys, and special metals in electrothermal furnaces are summarized in Tables 7 and 8.

5.2. Resistance Furnaces

5.2.1. Basic Principles [55], [56], [67]

All resistance furnaces operate on the basis of Joule's law, i.e., an electrical conductor emits heat when a current flows through it. The quantity of electric power that is converted into heat may be expressed by:

$$P = UI \quad (\text{Joule's law}) \tag{1}$$

where P is the electric power (W), U the voltage (V), and I the current (A).

Ohm's law relates the resistance of the conductor (R in Ω) and the current through it to the applied voltage:

$$U = RI \quad (\text{Ohm's law}) \tag{2}$$

Substituting Equation (2) into Equation (1) gives

Table 5. Process data for smelting and melting processes

Type of furnace	Metal	Crucible capacity, t	Furnace power, MW	Throughput, t/h	Energy consumption, kW · h/t
Induction	Al	1 – 40	0.3 – 4	1 – 8	500
	Cu	5 – 25	1 – 3	3 – 7	400
	cast iron	0.08 – 60	up to 17	0.5 – 32	520
	steel	0.08 – 60		0.2 – 30	650
			up to 17		
Vacuum induction	NiCo	1 – 10	1 – 2	0.4	1200
	Cu	3 – 10		1.5	400
Electroslag refining	steel	2 – 160	1.5	0.5 – 1.5	1300
	NiCo	< 5		0.3	900
Electron beam	Nb, Ta	0.5 – 2	< 1.2	0.02 – 0.4	6000 – 15 000
	Mo, W			0.02 – 0.1	< 5000
	Ti, NiCo		≤ 5	0.2 – 1.0	1000
Vacuum arc	steel, Ni	50			
	Ti	10			
	Mo, Nb, Ta		0.3 – 3		

Table 6. Process data for the production of nickel/ferronickel from sulfidic and oxidic ores in resistance furnaces

Nickel concentrate	Ore body	Ore composition, %			Energy consumption	
		Ni	Cu	Co	kW · h/t ore	MW · h/t Ni + Cu
Sulfide ore	Manitoba	10.5	0.4	0.2	430	5.5
	Rustenburg (≤ 30 ppm PGM) ∗	3.5	2.0	0.1	500	9.5
Oxide ore	Cerro Matoso	3.0			450	16.0
	Larymna	1.2			420	35.0

∗ PGM = platinum group metals

Table 7. World production and energy consumption data for the production of metals and alloys in electrothermal furnaces

Metal	Production		Size of furnace	Energy consumption, MW · h/t
	10^3 t/a	Year		
Aluminum	22 600	1998	$\leq$ 300 kA	14
Calcium carbide	6400	1982	60 MW	3
Calcium silicide	100	1980	30 MW	12
Chromium	10	1980		
Electrosteel	244 000	1996	100 MW	0.5 – 0.6
Ferrochromium	2900	1980	40 – 150 MW	
Carburé				4
Affiné				8
Ferromanganese	4600	1980	30 MW	
Carburé				3
Affiné				1.5 – 2.0 + SiMn
Suraffiné				1.8 – 2.5 + SiMn
Ferroniobium	30	1998		
Ferrosilicon	3300	1978		$\leq$ 50 MW
FeSi 45				5
FeSi 75				9
FeSi 90				13
FeSi 98				14.5
Magnesium				
Fused-salt electrolysis	290	1998	100 kA	14
Silicothermal	84	1998	6 MW	11 + FeSi 75
Manganese	100	1980	20 – 60 MW	
Nickel from sulfide	800	1998	20 – 30 MVA	see Table 6
Nickel from oxide	205	1998	45 – 60 MVA	see Table 6
Phosphorus	1300	1977	$\leq$ 70 MW	13
Silicon	500	1980		
Silicochromium	365	1980		7
Silicomanganese	1700	1980	20 – 60 MW	3.7 – 6

$$P = RI^2$$

The amount of energy (heat produced in Joules after t seconds) may be represented by:

$$Q = Pt = RI^2 t$$

The above equations show that the higher the resistance of a conductor or the larger the current, the more electric energy is converted into heat. The resistance of a conductor depends on the resistivity of the material and on its shape (length and cross-section):

$$R = \sigma l / S$$

where σ is the resistivity ($\Omega \cdot$ cm), l the length, and S the cross-section.

The resistivity is also a function of the temperature:

$$\sigma_T = \sigma_{273\,\mathrm{K}} (1 + \alpha T)$$

where σ_T is the resistivity at T Kelvin, $\sigma_{273\,\mathrm{K}}$ the resistivity at 273 K, T the temperature in K,

and α the resistivity temperature coefficient in K^{-1}.

In most cases slag acts as the main resistance in furnaces. The resistivities for typical metallurgical slag materials are presented in Table 9.

Table 8. Annual world production of special metals

Metal	Production, 10^3 t/a	Year
Vanadium	0.1	1992
Niobium	1.4	1985
Tantalum	1.2	1984
	0.65	1985
Titanium	90	1984
Zirconium	6	1982
	8	1983
Hafnium	0.1	
	0.25	1983

Table 9. Resistivities of slag materials [65]

Compound	Temperature, °C	Resistivity, $\Omega \cdot$ cm
Ferrous/nonferrous slags		
$CaSiO_3$	1550	3.85
Ca_2SiO_4	1550	3.13
$FeSiO_3$	1400	0.83
Fe_2SiO_4	1400	0.20
FeO	1400	0.004
Al electrolyte		
Na_3AlF_6	1000	0.45
Na_3AlF_6	1100	0.42
Ti slags		
70 % TiO_2	1800	0.033
80 % TiO_2	1800	0.020
90 % TiO_2	1800	0.009

Direct resistance furnaces exploit the above principles for metallurgical purposes (see page 17). Indirect resistance furnaces fall outside the scope of this article and are not discussed here.

In direct resistance furnaces the feed or charge is the conductor and its resistance to the current flowing through it creates the heat for melting. The current in these reactors flows between at least two or more electrodes, one of which may be submerged in the material to be heated/melted. Direct resistance furnaces may be divided into three categories on the basis of the reactions taking place in them:

1) Reduction resistance furnaces
2) Refining resistance furnaces
3) Solid state resistance furnaces

5.2.2. Reduction Resistance Furnaces [97–108]

Electrothermal reduction furnaces use electric energy to heat the feed to the required operating temperature. In conventional reduction furnaces, heat liberated during the exothermic reaction between, for example, coke and oxygen is utilized to heat the feed. In electrothermal furnaces, coke is only needed for reduction.

Reduction resistance furnaces have a number of advantages over furnaces using fossil fuels:

1) Electrothermal processes permit higher productivities due to higher temperatures and energy densities
2) Higher grade products can be produced because ash and impurities (e.g., phosphorus and sulfur) derived from the fossil fuel are absent
3) Since the reactors are usually covered, dross formation can be minimized because atmospheric oxygen may be excluded
4) A smaller volume of off-gas volume is produced by resistance reactors with a higher carbon monoxide concentration because only reduction reactions take place
5) Electric furnaces can easily be enclosed, permitting tight control of the off-gas composition
6) Metallurgical coke can be replaced by, for example, charcoal which contains much less impurities
7) If the furnace is operated under vacuum conditions, reduction and degassing can take place in a single step

An obvious disadvantage of electrothermal furnaces are high energy costs that result from their energy-intensive operation. However, in countries where hydropower or inexpensive or excess electric power is available, this is an economically viable method for producing metals.

5.2.2.1. Furnace Construction and Operating Parameters [55], [56], [65], [84]

Normally two molten phases are present in a reduction furnace: a salt or oxide slag floats on the second phase (the produced metal or matte) that collects at the bottom of the furnace. Small

amounts of solid feed are added continuously. This open-bath operation can be used for the production of FeSi, Si, CaSi, and Ti slags, for example. The electric power is usually transferred to the furnace by three or six electrodes that are submerged in the molten slag. For the production of ferroalloys electrode diameters are 0.1 – 1.0 m and current densities are 5 – 15 A/cm^2, whereas for matte smelting the electrode diameters lie between 0.6 and 1.4 m and current densities between 2 and 3 A/cm^2. Typically, three-phase current circuits are used. The electric resistance of the slag produces heat which keeps the molten phases in a liquid state and supports the reduction reactions.

In some cases large amount of solids (e.g., the feed) can float on top of the slag (covered-bath operation for the production of FeCr, FeMn, SiMn, FeNi, CaC_2). In addition to resistance heating, these solids are also heated by electric arcs (see Section 5.3) which are created between the feed and the electrodes, and between the individual feed grains. Hence, these furnaces may also be called submerged-arc or arc-resistance furnaces. However, since no clear classification between this type of furnace operation and a pure resistance furnace can be made, they are discussed in this section.

The *electrodes* used in reduction resistance furnaces are either graphite or the cheaper Söderberg type (diameter of Söderberg eletrode 0.5 – 2 m and current densities 2 – 3 A/cm^2) type. The carbon paste for the Söderberg electrodes is filled into molds above the furnace. The electrodes are then baked by the heat produced by the current flow and the furnace, consequently they need not be pretreated in a separate furnace as is the case for graphite electrodes. Electrode consumption depends on the application, e.g., consumption of prebaked electrodes in steel production amounts to 5 – 15 kg per tonne steel, for ferrochromium (68 % Cr) production 45 – 50 kg per tonne ferrochromium and for ferromolybdenum production 100 – 120 kg per tonne ferromolybdenum.

The electrodes can be arranged on the vertices of an equilateral triangle or in a row (three or six). The best arrangement for a particular application depends on parameters such as furnace efficiency, type of feed, and investment costs.

The type of *refractory* used in the walls and the roof of reduction furnaces depends on the type of application and exposure it must withstand. They include magnesite, chrome magnesite, alumina, and graphite.

Figure 20 shows that the resistance heat is produced directly between the electrodes and themetal. The metal itself has a very low resistance and therefore produces very little heat. The resistivity of the slag, the slag depth, and the distance between the electrodes are the critical parameters controlling the amount of heat produced in the slag. The distance between the electrodes is chosen large enough to ensure that the current passes through the slag and metal and does not short- circuit between the electrodes (for six electrodes with diameter D the distance must be larger than 2.75 D; for three electrodes $\geq 2.6\,D$). Laboratory investigations have shown that pure resistance heating may be expected with electrode current densities up to 12 A/cm^2. Higher electrode current densities also show unstable arcing [103].

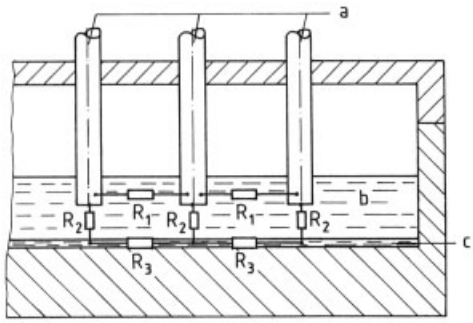

Figure 20. Flow of current through slag and metal in resistance furnaces, shown as a circuit containing resistances (R) [97]
a) Electrodes; b) Slag; c) Matte – metal

Figure 21 depicts the circulation of the slag, which can reach velocities between 20 and 30 cm/s close to the electrode, but $<$ 1 cm/s in colder sections of the furnace. This circulation is created by:

1) Hot slag being replaced by colder slag from the surface due to density differences
2) Carbon monoxide bubbles at the electrode surface, which drag the slag upwards
3) Electromagnetic forces created by the current flowing between the electrodes (three-phase furnaces create a downward movement of the slag, whereas d.c. furnaces create an upward movement if the bottom electrode

is a cathode and an upward movement if the bottom electrode is an anode)

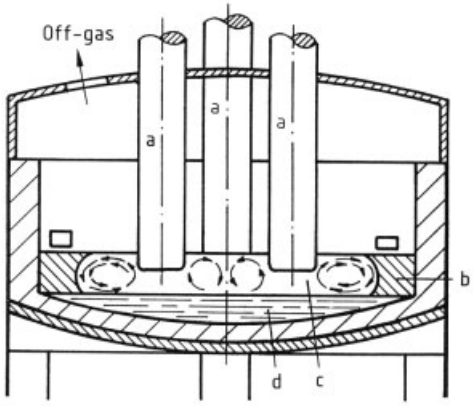

Figure 21. Flow distribution in resistance furnaces [82]
a) Electrodes; b) Dead volume; c) Slag; d) Matte – metal

In slag – matte furnaces the slag (40 – 150 cm deep) shows an upward circulation around the electrode. The movement is in the opposite direction in the matte (40 – 50 cm) due to electromagnetic forces (electrodynamic effects) and because the matte and slag flow in the same direction at the interface. In addition to the above-mentioned parameters affecting furnace operation, a number of operating parameters pertaining to the slag, can be modified to achieve optimal recovery of the valuable metals and compounds from the feed. These parameters are a function of the composition of the slag or material to be melted or smelted and temperature reached. They include:

1) Viscosity (Fig. 22), which affects the mixing and settling of the reduced and melted metal drops
2) Density differences between slag (iron slag: $\varrho = 2.2 - 2.6$ g/cm^3; copper slag: $\varrho = 3 - 3.6$ g/cm^3); matte ($\varrho = 3.8 - 5.2$ g/cm^3), speisses ($\varrho = 7.2 - 7.5$ g/cm^3) and metals effect the settling behavior of the phases or metals to be recovered
3) Resistivity, which has an effect on the amount of heat created and hence on the temperature via Joule's law
4) Oxygen permeability of the slag, a low oxygen permeability allows the required reducing conditions to be maintained within the metal – matte phase as well as in the slag

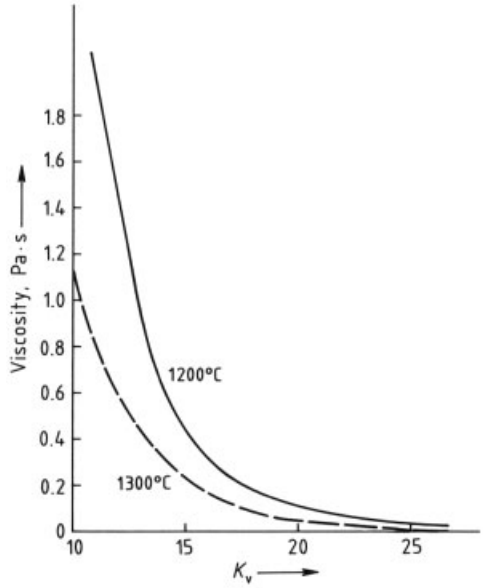

Figure 22. The viscosities of slags as a function of $K_V = (\%FeO + \%Fe_3O_4 + \%CaO + \%MgO)/(\%SiO_2 + \%Al_2O_3)$ and temperature [65], [102]

An additional factor which affects the performance of these furnaces is the residence time τ of the material to be treated in the reactor:

$$\tau = \frac{\text{active reactor volume}}{\text{volumetric flow of slag through the furnace}}$$

Tracer tests in a typical six-electrode resistance furnace for the production of Cu – Ni matte revealed that 65 % of the volume exhibits mixed flow characteristics, 1 % plug-flow characteristics, and 34 % is attributed to the dead volume [69, p. 274].

5.2.2.2. Metallurgical Significance

Reduction furnaces are used in many different processes for primary and secondary production:

1) Reduction of lead and zinc from oxidic materials (e.g., ores, roasted sulfidic ores, secondary oxidic materials)
2) Production of ferroalloys from oxidic ores (e.g., Ni, Cr, Mn, Si)
3) Ilmenite reduction to iron and a titanium slag
4) Iron and steel production from ores and prereduced pelletized iron ores

5) Matte production from sulfidic concentrates of copper and nickel
6) Production of calcium carbide and corundum
7) Cleaning of slag (i.e., recovery of valuable elements) produced by primary and secondary smelting operations (e.g., Pb, Cu)
8) Melting and cleaning of secondary metals (e.g., scrap)

Products include heavy metals and ferroalloys, mattes, special carbides, slags, and ceramics. The production of selected compounds will be treated in the following (→ Chromium and Chromium Compounds; → Manganese and Manganese Alloys; → Nickel; → Titanium, Titanium Alloys, and Titanium Compounds).

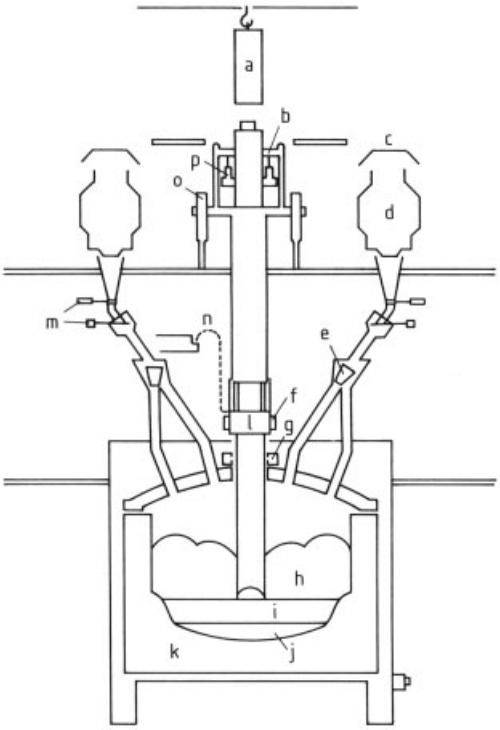

Figure 23. Schematic drawing of an electric FeNi reduction furnace [106]
a) Electrode addition; b) Holding ring; c) Ventilation hood; d) Calcine car; e) Charge diverter bucket; f) Pressure ring; g) Ventilation ring; h) Calcine; i) Slag; j) Metal; k) Refractory; l) Contact clamp; m) Change valves; n) Bus; o) Electrode hoist; p) Slipping ring

Production of Carbides. Calcium carbide furnaces (→ Calcium Carbide, Chap. 4.) belonged to the first reduction furnaces that were used on an industrial scale. The furnace feed consists of calcined lime and coke that react to form calcium carbide and carbon monoxide when electric energy is transferred to the charge via electrodes. The original furnaces (ca. 1895) operated with direct current. One electrode was submerged in the feed, and the bottom of the furnace formed the other electrode. Alternating current replaced the direct current installations due to their low efficiency and the high investment costs for rectifiers. Presently the energy consumption of this process amounts to 3 MW · h per tonne calcium carbide.

Production of Ferroalloys. At several plants (Larymna in Greece, Cerro Matosa in Colombia, Doniambo in New Caledonia, Falconbridge in Dominicana, Inco in Guatemala) laterite ores are smelted to *ferronickel* in electric reduction furnaces (see Fig. 23). The ore is prereduced in rotary kilns at about 800 – 950 °C and continuously fed to the reduction furnaces opposite the slag tapping hole. Ferronickel is produced at 1500 – 1650 °C in the electric furnaces (three or six prebaked or Söderberg electrodes, diameter 1 – 1.3 m, electrode consumption several kilogram per tonne ferronickel) [70], [74], [75], [104–106] (→ Nickel, Chap. 8.2., → Nickel, Chap. 8.3.).

To produce *ferromanganese alloys*, manganese ores (≥ 42 – 54 wt % Mn) are mixed with flux (lime) and coke and either charged to a shaft furnace or an electric reduction furnace (→ Manganese and Manganese Alloys, Chap. 6.1.2.). For the production of low-carbon ferromanganese only the electric furnace is suitable. Energy consumption is 2100 – 8000 kW · h/t depending on the manganese, silicon, and carbon content (carburé, affiné, suraffiné) of the product [74], [75]. *Ferrochromium alloys* are produced in electric furnaces only. The Perrin process is applied for the production of low-carbon ferrochromium. Most of the high-carbon ferrochromium production is performed directly in electric furnaces. For the processing of chromite fines for high-carbon ferrochromium production, the route pelletizing, prereduction, and submerged arc furnace smelting may be followed [74], [75], [157] (→ Chromium and Chromium Alloys, Chap. 3.3.1., → Chromium and Chromium Alloys, Chap. 3.3.2., → Chromium and Chromium Alloys, Chap. 3.3.3.).

Ferromolybdenum (→ Molybdenum and Molybdenum Compounds, Chap. 6.3.), ferrotungsten (→ Tungsten and Tungsten Compounds), ferrovanadium (→ Vanadium and Vanadium Compounds), ferrotitanium (Titanium and Titanium Alloys), phosphorus, (→ Phosphorus, Chap. 4.6.), ferrophosphorus (→ Phosphorus, Chap. 4.1., → Phosphorus, Chap. 4.2.), and ferrosilicon (→ Silicon) are also products of submerged arc furnaces.

Production of Matte. The numerous examples of the production of Cu – Ni matte include (→ Nickel, Chap. 4.3.)

1) Inco at Thompson with two 30 MVA rectangular furnaces containing six Söderberg electrodes (diameter 2 m) to produce a Cu – Ni matte with 26 % Cu + Ni at 400 kW · h per tonne charge [105]
2) Falconbridge with two 33 MVA rectangular furnace each having six electrodes producing Cu – Ni matte with 24 % Cu + Ni [104]
3) Rustenburg with a 19.5 MVA rectangular furnace with six electrodes (diameter 1.24 m) producing a matte with 24 % Cu + Ni [69]
4) Mufulira with a 36 MVA rectangular furnace with six electrodes (diameter 1.6 m) to produce a Cu matte using less than 400 kW · h per tonne [69].

Production of Ti Slags. Some large plants treat ilmenite for the production of a slag containing 75 – 85 % TiO_2 and iron (→ Pigments, Inorganic, Chap. 2.1.2.2.).

Among them are the plants of Quebec Iron and Titanium Corporation in Canada, Richard's Bay Iron and Titanium in South Africa, and Tinfos Titan & Iron in Norway.

Slag Cleaning [84], [108], [157]. In order to produce an environmentally clean slag and optimize recovery, slag cleaning operations often follow the primary and secondary production of metals such as lead and copper. This implies that heavy metal, matte, and oxide are recovered as far as possible.

Slag cleaning is mostly carried out in separate resistance furnaces, for example at several *copper smelters* employing the Outokumpu process (e.g., Onsan, KGHM-Glogów, Hidalgo, Hindustan, Norddeutsche Affinerie, Botswana Concession Limited, Rio Tinto) [84]; at Kamioka [99] where lead and zinc are recovered from slags; and in the Mitsubishi process for the production of copper at Naoshima and Timmins [84].

A slag-cleaning or reduction furnace is, however, sometimes directly integrated in the primary smelter, as for example in three Outokumpu furnaces and the Kivcet furnaces (Fig. 24). The Kivcet furnace is used for the *production of lead from sulfide ores* (→ Lead, Chap. 4.4.2.1.). In this operation the dry feed is oxidized and smelted in suspension in the reaction shaft (c) of the furnace during which most of the lead enters the slag phase as lead oxide. Most of this slag is reduced in the lower part of the reaction shaft by coke breeze floating on top of the liquid phases (metal and slag). Metal and slag continuously underflow a barrier to the slag-cleaning or reduction section of the furnace, in which three electrodes (a), placed in a row, transfer electric power to the melt. Additional coke is added to produce an atmosphere that reduces the remaining lead oxide to lead, which then collects at the bottom of the furnace. Some zinc oxide is also reduced to zinc, which evaporates and reoxidizes in the gas phase and is recovered from the filters in the off-gas cleaning units. Lead leaves the furnace by a syphon [101], [107].

Other slag-cleaning furnaces have been positioned downstream of copper or lead secondary smelting furnaces, and converters (e.g., at Metallurgie Hoboken, Belgium and USMR smelter of AMAX, New Jersey). Typical process data for slag reduction and slag cleaning processes are given in Table 10.

5.2.3. Refining Resistance Furnaces [57], [129], [130], [154]

The *electroslag refining process* (ESR) is the classical example for refining operations in a resistance furnace. The high resistance of the slags (CaF_2– CaO – Al_2O_3) is used to produce enough heat to melt and refine high-grade steels (→ Steel) and titanium (→ Titanium, Titanium Alloys, and Titanium Compounds). The main components of the furnace are its consumable electrode (raw material), a water-cooled mold, and a bottom plate, which are connected to an alternating current source (Fig. 25). Initially the

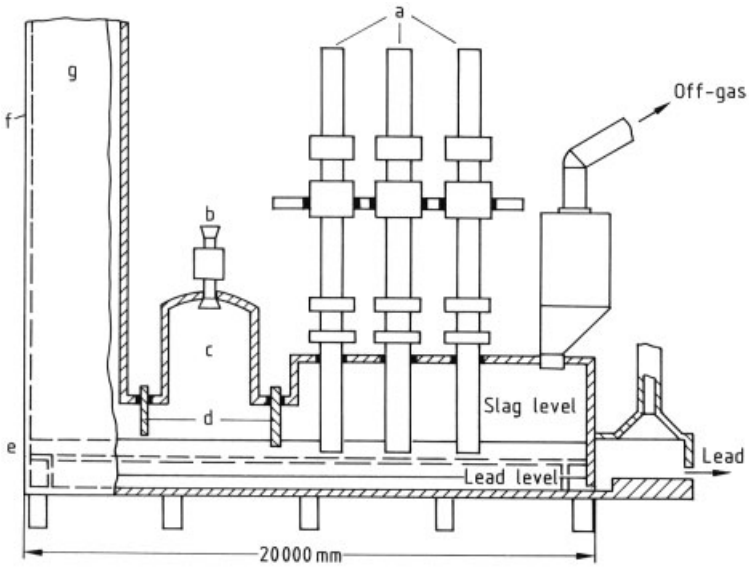

Figure 24. The Kivcet reactor for the production of lead [101], [107]
a) Electrodes; b) Concentrate burner; c) Smelting shaft; d) Partition walls; e) Belt of water-cooled elements; f) Water-cooled refractory lining; g) Gas removal shaft

Table 10. Production data for electrothermal slag reduction in various processes

Application	Estimated number of plants	Production capacity, 10^3 t/a	Furnace size, MVA	Energy consumption, kW · h/t
Cu flash smelting	30 – 35	10 000	6 – 10	50 – 100
Pb flash smelting	2			
Conventional Pb reduction	5	500	3 – 6	150 – 700
Secondary Cu smelting	2	200	6	150 – 200

electrode (a) is submerged in a high-melting liquid slag (b), the mold (e) and bottom plate (f) contain the slag. At sufficiently high voltages the liquid slag starts to melt the consumable electrode at its surface. The molten metal collects at the tip of the electrode and drops through the slag. During this process it is refined due to its interaction with the slag. The molten metal collects in the mold, and the bottom plate moves downwards at the same rate as that of metal production. The refined metal is cooled by the water-cooled mold and solidifies. The mold and bottom plate are made of copper and water cooled to prevent reactions and to ensure high heat transfer (Fig. 26). This cooling system is responsible for high energy losses (see Tables 3 and 7).

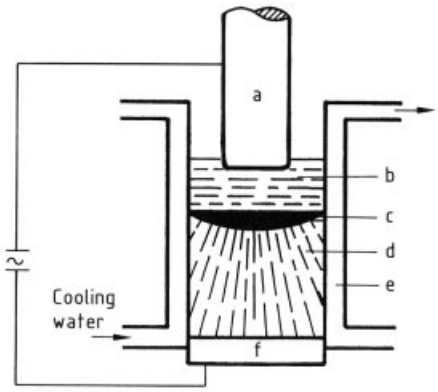

Figure 25. Sketch of an electroslag refining furnace [73]
a) Consumable electrode; b) Slag; c) Molten metal; d) Solid metal; e) Water-cooled mold; f) Movable cooled copper plate

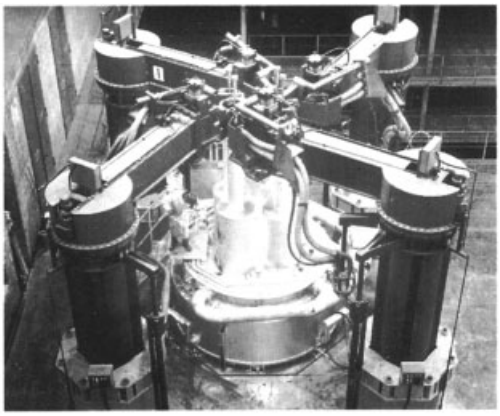

Figure 26. An industrial ESR plant at Saarstahl GmbH [65], [130]
Maximum block diameter 2.3 m, maximum block mass 160 t, with a remelting rate of 2.3 t/h

ESR processes can also be used to *remelt and refine secondary materials* in a single step. Reactor construction is similar to that described above for the standard ESR, the only difference being that a nonconsumable electrode is used and material is fed in particulate form.

5.2.4. Solid-State Resistance Furnaces [63], [71]

The feed material for this type of furnace has a high electric resistance. A current is applied to the feed via two electrodes and the heat generated permits the desired reactions to take place in the solid state. Two typical applications are the graphite furnace and the silicon carbide furnace (see below).

Graphite Furnace (→ Carbon). In this furnace petroleum coke is converted into graphite which is used for electrodes in other electrothermal furnaces. The feed (petroleum coke with tar as a binder) is mixed and usually extruded into the desired shape of the electrodes. The electrodes are heated up to 800 – 1300 °C in fuel-heated furnaces to carbonize the binder. Subsequently the electrodes are charged to a resistance furnace (20 m × 4 m × 4 m), where they are stacked parallel to the face sides (Fig. 27). The void areas are filled with crushed graphite to ensure that the electrodes are in good electric contact with one another. An electric current flows through the furnace and produces the operating temperatures (2400 – 2800 °C). The amorphous coke is transformed to graphite (hexagonal lattice) within several days. Since the conductivity of the amorphous coke is much lower than that of graphite, the furnace voltage is changed in order to maintain the required temperature. The energy consumption for this process is 4 – 6 MW · h per tonne graphite for a connecting power of 5 – 16 MW. Such a graphite furnace produces ca. 50 t graphite per batch.

Silicon Carbide Furnace. (See also → Silicon Carbide). In this case two solid materials (sand and coke) are mixed and filled into a furnace which is similar to that used for graphite production. The reaction

$$SiO_2 + 3\,C \rightleftharpoons SiC + 2\,CO$$

takes place in the solid state. The sand – coke mixture (with additions of sawdust and sodium chloride) is charged into the furnace around a central core of graphitized coke (Fig. 28). This

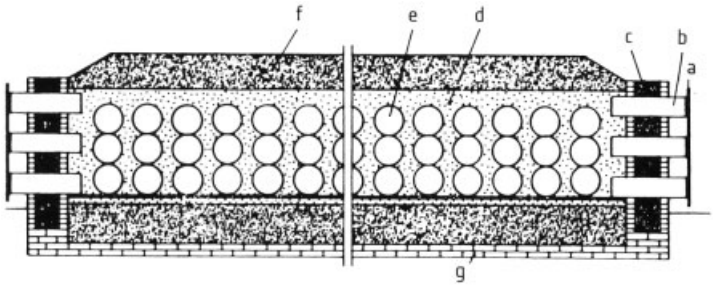

Figure 27. Sketch of a graphitizing furnace [63]
a) Current supply; b) Electrodes; c) Furnace head; d) Resistance body; e) Electrodes to be graphitized; f) Carborundum layer; g) Furnace bed

central core is connected to electric power and the coke is used as a conductor, which reacts in a similar way as described in the graphitization furnace above. Silicon carbide formation starts at 1700 °C, at 2400 – 2500 °C a coarse crystalline product is formed. Sodium chloride is added to evaporate undesired elements such as aluminum, iron, and titanium as chlorides. Only ca. 20 –30 % of the reaction product is high-grade material. Energy consumption is between 8.5 and 12.5 MW · h per tonne silicon carbide.

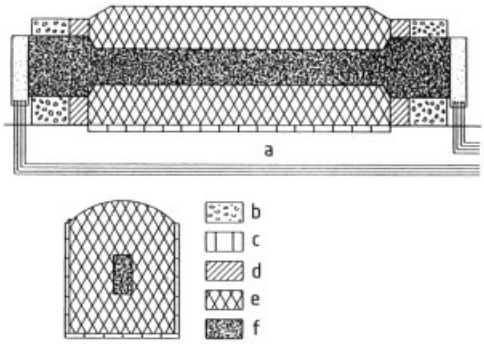

Figure 28. Sketch of a silicon carbide furnace without degassing from the furnace bottom [63]
a) Current supply; b) Fire clay concrete; c) Fire clay bricks; d) SiC-protection layer; e) Reaction mixture; f) Electrode and resistance core

5.3. Arc Furnaces

5.3.1. Basic Principles [55], [56], [67]

An electric arc can be produced and sustained between two electrodes or an electrode and a liquid melt if the voltage is high enough. The transformation of electric energy to heat takes place through the current in the ionized plasma of the arc, in which the temperatures may reach 6000 °C.

Electric arc furnaces may be classified according to whether transfer of heat from the arc to the furnace feed is indirect or direct. In indirectly heated furnaces (Fig. 29 A), the arc burns between two electrodes without contact with the feed; heat is transferred solely by radiation and convection. This technique was employed in single-phase rotary furnaces, which are, however, no longer in use because they are not economical. In directly heated furnaces (Fig. 29 B) an arc burns between the electrode and the melt. Heat is transferred via the fire point produced by the arc by conduction, radiation, and convection to the melt. Directly heated arc furnaces can be used for melting and refining.

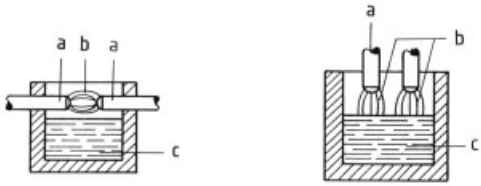

Figure 29. Indirectly (A) and directly (B) heated arc furnaces [67]
a) Electrodes; b) Arc; c) Liquid melt

The arc-resistance and submerged-arc furnaces mentioned in Section 5.2 are mainly used for reduction. Since their construction is similar to that of resistance furnaces the discussion is not repeated here.

5.3.2. Electric Arc Furnaces [109–112], [154]

Today, 32 % (240×10^6 t/a) of the total world steel is produced in electric arc furnaces, this is their principal application (→ Steel).

Furnace Construction. One-electrode furnaces usually operate in the d.c. mode with the electrode forming the cathode and the bottom of the furnace forming the anode (e.g., 11 MW furnace at Florida Steel Corporation for melting scrap, electrode diameter 0.45 m, electrode consumption 2 kg per tonne steel). This kind of furnace may, however, also be operated in an a.c. mode [111].

Three-electrode furnaces (Fig. 30) are mostly a.c. furnaces. The electrodes (a) are switched into a three-phase circuit and are usually placed on the vertices of an equilateral triangle in a round furnace.

Most furnaces are equipped with removable roofs in order to charge via baskets from the top. The charging of sponge iron, for example, takes place batchwise in baskets or continuously through tube feeding systems. The required atmosphere can be provided inside the furnace which can be tilted during tapping.

Dolomite and magnesite are the refractories mostly used for lining the lower walls and the bottom of the steel shell of the arc furnaces.

The upper walls and the roof of the furnaces are water-cooled to provide good resistance against high temperature and temperature changes. The high temperature obtained in arc furnaces allows high production capacities. When scrap is melted down, crater-type holes are burnt into the scrap which protect the furnace roof and walls from overheating.

Modern furnaces may be tilted only 12° during slagging off operations, steel tapping (Fig. 30) is done by eccentric bottom tapping. Sometimes the mixing in the furnace is improved by an induction stirring coil underneath the furnace or by bubbling gas through a porous plug.

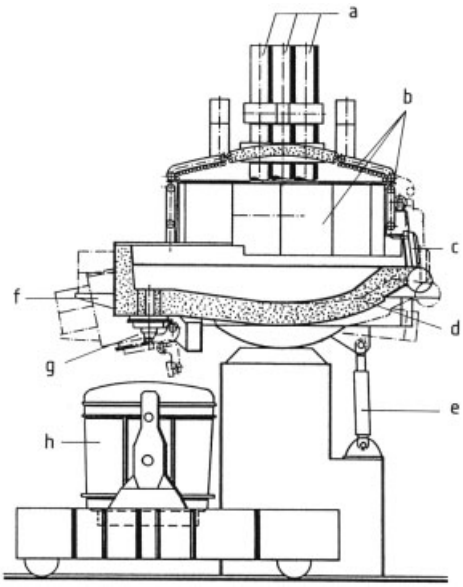

Figure 30. Typical three-electode arc furnace with water-cooled walls and roof, and eccentric bottom tapping for steel production [65], [78]
a) Electrodes; b) Water-cooled walls and roof; c) Tiltable furnace; d) Refractory lining; e) Tilting mechanism; f) Spout; g) Eccentric tapping; h) Ladle

Metallurgical Significance. The advantages of electric arc furnaces are listed below:

1) High temperatures can be reached within a short time
2) Continuous temperature adjustment is possible via computer control systems which control the level of the arc
3) The atmosphere in the furnace can be tightly controlled
4) In contrast to conventionally heated furnaces, impurities are not brought into the steel by the energy input (e.g., sulfur by coke)
5) The furnace can easily be brought on- or off-line with relatively little energy and time, furnace operation may be controlled to meet the production requirements of the cast shops and mills.
6) Oxidation, reduction, and alloying can be carried out in one charge with little loss
7) Desulfurization is possible

Development of arc furnace technology has led to the construction of ultrahigh power (UHP) furnaces [63], [109] with a connected power (MVA) to feed (tonnes) ratio $\leq 1:2$. Examples of UHP furnaces are given in Table 11. The excellent electrical efficiency of these furnaces can be attributed to low Joule-effect losses and conductors with small cross sectional areas. The short, high-current arcs improve heat transfer. These furnaces are computer controlled (energy input, phase balancing, and regulation of cooling water flow rates in the electrodes, walls, and roof). For a 24 MVA furnace (diameter 4.6 m, weight 40 t, electrode diameter 0.45 m) power consumption amounts to 560 kW · h/t, which decreases to < 500 kW · h/t for a 90 MVA furnace (diameter 7 m, weight 150 t, electrode diameter 0.6 m).

5.3.3. Vacuum Arc Refining (VAR) Furnaces [129], [130], [154]

Furnace Construction. Vacuum arc furnaces are normally used for consolidation and refining reactions. The raw material (sponge or scrap) is pressed into an electrode shape, which is used as a consumable electrode in the arc furnace. Normally d.c. current is used to provide a stable arc between the consumable electrode and the counter electrode at the bottom of the furnace (Fig. 31). After the arc is ignited the electrode starts to melt, volatile elements and entrapped gases are removed through the high temperature and the vacuum conditions. The refined metal drops to the water or liquid Na/K- cooled copper bottom where it solidifies. Copper is used as the mold material to prevent reactions with refractories.

Table 11. Process data for examples of arc furnaces (> 100 t) [112]

Company, location	Furnace capacity, t	Capacity, 10^3 t/a	Power, MVA ∗	Specific power, kVA/t	Electrode diameter, mm	Furnace diameter, mm
Thyssen, Oberhausen, Germany	130	600	96 AC	740	610	7050
Krupp, Bochum, Germany	140	550	85 AC	710	610	7000
Toshin – Azuma Steel, Himeji, Japan	150	600	66 AC	440	610	7000
Chaparral Steel, Midlothian, United States	150	800	138 AC	920	610	6710
Lukens Steel, Cloatesville, United States	160	800	67 AC	420	610	6710
Terni, Italy	180	500	110 AC	611	600	7300
Nakorntai strip, Chonburi, Siam	180	1500	130 AC	722		8700
North Star, Delta, USA	163	1360	140 AC	848	600	7300
Nucor Steel, Nicoman USA	155	1000	164 DC		711	7315
Nylsa, Monterrey, Mexico	135	900	120 DC		711	7315
Megasteel, Klang, Malaysia	160	1200	160 DC		800	7100

∗ Power for transformer

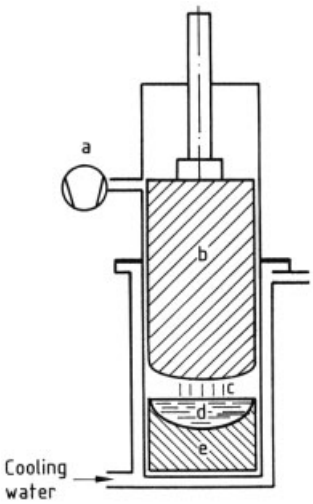

Figure 31. Typical vacuum arc refining furnace with consumable electrode [110]
a) Vacuum pump; b) Consumable electrode; c) Arc; d) Molten metal; e) Solid metal

VAR furnaces may also be used with a permanent electrode if the material is fine scrap [110]. This is fed in separately. These furnaces are also used for alloying high-melting metals such as niobium, tantalum, or titanium.

Metallurgical Significance. Molten titanium is very reactive, as are molybdenum and zirconium. These metals must therefore be melted in a controlled atmosphere or a vacuum. Vacuum arc refining furnaces are thus suitable for meltingand/or recycling these highly reactive metals (also possible in electron-beam furnaces). As with ESR furnaces, metal sponge and scrap may be mixed and pressed into consumable electrodes and used as such in the VAR furnace. Often refining involves several steps to achieve a pure product, this explains the relatively high energy consumption for titanium (3.8 MW · h per tonne titanium produced).

Vacuum arc furnaces are also used to degas certain steel grades during remelting, energy consumption is ca. 1.0 MW · h per tonne produced metal.

5.4. Induction Furnaces [80], [89], [113–118], [129], [130], [154]

The increasing demand for high-quality metals and alloys under ecologically clean conditions has increased the application of induction furnaces in foundries for cast iron, steel, and nonferrous metals (mainly Al, Cu, Mg, Ni, precious metals and their alloys). At least 20 % of metals produced by electrothermal reactors are melted in induction furnaces.

There are two basic types of induction furnace construction: the crucible induction furnace and the channel induction furnace (Figs. 32 and 33). In this chapter discussion of the application of induction furnaces is mainly limited to melting and refining of metals and alloys [56], [65], [81], [94], [95], [119]. Special applications are zone refining and levitation melting [58], [81]. Inductive heating is also used for brazing and welding, different types of heat treatment

(e.g., surface treatment, heating and annealing), metal transport and dosing, and electromagnetic casting (EMC) [56], [65], [81], [94], [95], [119]. Since these processes do not occur within reactor vessels, they will not be discussed here.

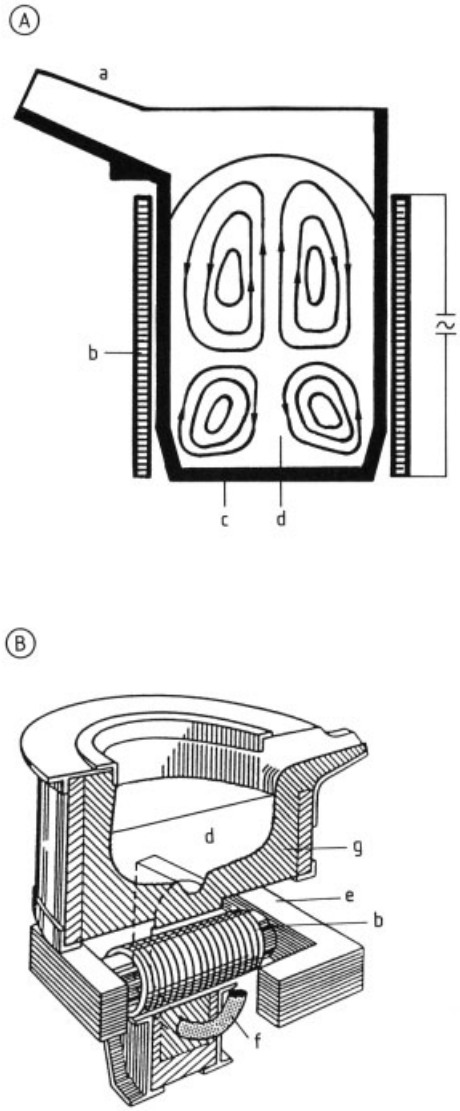

Figure 32. Crucible induction (A) and channel induction (B) furnaces [65]
a) Spout; b) Induction coil; c) Crucible; d) Molten metal; e) Iron core; f) Channel; g) Rammed lining

5.4.1. Basic Principles [55], [56], [67], [78], [81], [154]

Electromagnetic induction heating is a direct method for contactless heating with a high power density which implies that the material to be heated (melted) is not contaminated by heating gases or electrode materials. However, contact and reactions with the crucible (graphite or refractories) may take place. Advantages of these types of furnaces are excellent alloying and mixing conditions, good temperature control of the melt, a low dross formation, and low off-gas volume.

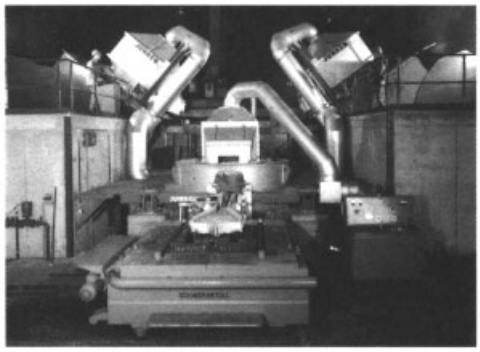

Figure 33. A 30-t channel induction furnace between two 24-t crucible induction furnaces [65]

The heating principle is as follows: an a.c. coil induces a potential in an electrical conductor (in induction furnaces solids and/or melts) situated inside the coil due to the changing magnetic field, which creates eddy currents (Lenz's law). The eddy currents or induced current produce heat according to Joule's law (see Section 5.2.1)

$$U_{\mathrm{ind}} = -\frac{w\,\mathrm{d}\Phi}{\mathrm{d}t}\ (\mathrm{Lenz's\,law})$$

where U_{ind} is the induced potential (V), Φ the magnetic field strength (Tesla), and w the number of coil windings.

The resistivity and relative magnetic permeability of the material to be heated and the frequency of the primary current circuit are the most important parameters of induction heating. The induced current within the conductor is not evenly distributed, but mostly located near the surface (skin effect, Fig. 34).

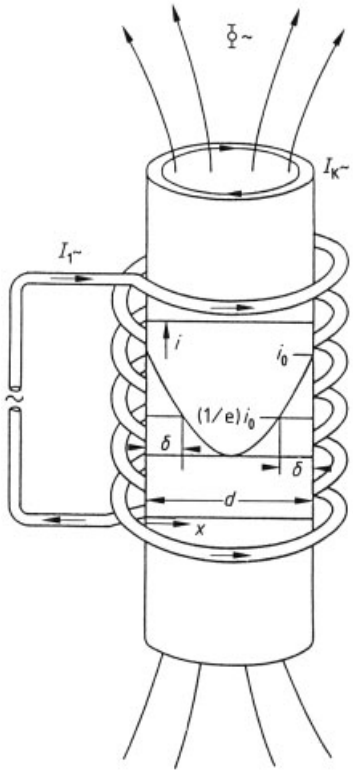

Figure 34. Inductive heating of a metal cylinder [81] δ = current penetration depth = (1/e) i_0; d = diameter; I_1 = inducing current; I_K = induced current; i_0 = surface current

5.4.2. Furnace Construction [55], [56], [67], [78], [81], [154]

5.4.2.1. Crucible Induction Furnace

Crucible induction furnaces have a cylindrical induction coil consisting of water-cooled copper tubes. The crucible is located inside the coil and usually consists of rammed or brick refractories, graphite, or clay – graphite. In rare cases a more expensive prefabricated ceramic or steel crucible may be used (Fig. 32 A and 33). An interesting development is the inert cold-wall induction crucible.

The crucible can and should be covered to reduce heat losses. For special applications (metal distillation, refining, melting of reactive or high-quality metals and alloys, etc.) the furnace is operated as a vacuum or controlled atmosphere induction furnace [58], [93].

There are two basic types of vacuum induction furnace, the compact vacuum crucible induction furnace and the vacuum chamber induction furnace (Figs. 35 and 36). The *compact vacuum crucible induction furnace* (Fig. 35 A) is operated at a low vacuum (ca. 10^{-3} bar). It has a simple design and lower investment costs than the vacuum chamber induction furnace, but the load capacities can be higher. The coil, vacuum chamber, and crucible form an integrated unit. In contrast, the crucible and induction coil of a *vacuum chamber induction furnace* (Fig. 35 B and 36) are located within a sealed, water-cooled chamber. It allows a vacuum of up to 10^{-9} bar and permits the metal to be poured under vacuum within the sealed chamber [56], [94], [95], [154].

Industrial induction crucible furnaces have different sizes (up to 60 t) and are operated at different frequencies (typically up to 500 Hz), under air or vacuum. Laboratory scale furnaces have capacities down to a few grams and frequencies reaching up to MHz range. Electrical efficiency is > 60 %.

One of the biggest advantages of a crucible induction furnace is the excellent bath mixing behavior. Interaction of the magnetic field and electrical currents create electromagnetic forces, which leads to the elevation of molten metal in the center of the bath and mixing (Fig. 32 A).

As a result of stirring, crucible induction furnaces can melt loads of small pieces of material (e.g., swarfs, chips), which are rapidly drawn into the bath. This facilitates operations such as alloying and refining, and ensures a uniform bath composition. The scale of mixing is inversely proportional to the applied frequency.

The diameter of the charged material d has to be at least 3.5 times larger than the penetration depth δ of the induced current to ensure maximum transference of power (Fig. 37). The minimum charge to start melting should be 20 % of the crucible capacity. In most cases the starting amount is two-thirds of the crucible capacity.

Metallurgical Significance. Crucible induction furnaces are used for melting and refining numerous metals and alloys. They are usually applied in cast iron, steel, and nonferrous metal foundries (aluminum, magnesium, copper, brass, bronze, zinc, nickel, precious metals,

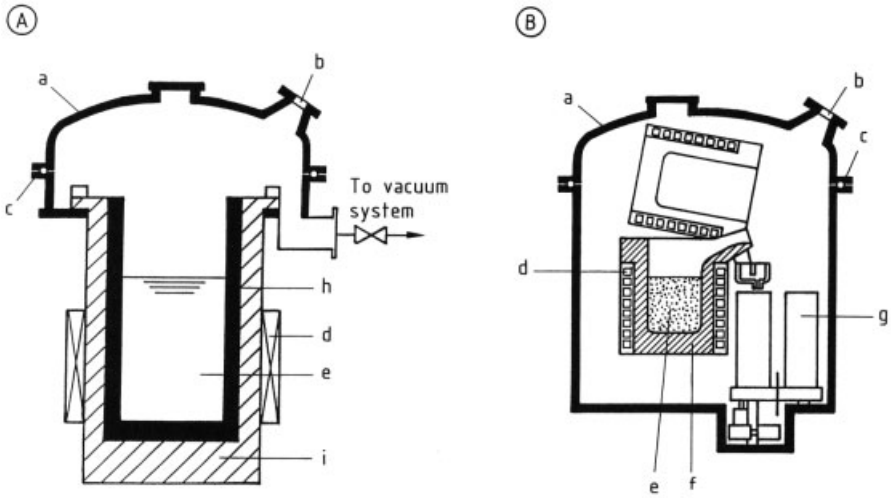

Figure 35. Compact vacuum crucible and (A) and vacuum chamber (B) induction furnace [56], [65]
a) Water-cooled vacuum hood; b) Viewing glass; c) Vacuum seal; d) Induction coil; e) Molten metal; f) Tiltable crucible; g) Centrifugal casting setup; h) Crucible; i) Rammed lining

superalloys). They are unsuitable for slag metallurgical operations since only the metal is heated by induction and not the slag. In addition the advantages of numerous crucible designs (e.g., rammed or brick lining; ceramic, graphite, steel, or copper crucible) render them very attractive. They also permit various operations, such as al-ternative or continuous melting, continuous melting and pouring, quick alloy changes, and application of vacuum or controlled atmosphere. They are sometimes used as buffer furnaces or grading furnaces.

Line-frequency crucible induction furnaces (Table 12) are mainly employed in cast iron, heavy metal, and aluminum foundries. They are sometimes used in (high-grade) steel foundries. High furnace capacities are possible.

Medium-frequency crucible induction furnaces (Table 13) are mainly used in operations that require frequent alloy changes for melting high-grade steel, alloyed cast iron, copper, aluminum, and precious metals in medium-size furnaces. The low power input required for holding furnaces means that short induction coil furnaces can be used for these applications.

If possible, the crucibles have a rammed lining (acidic monolithic linings) due to their lower costs, which produces significant savings in larger furnaces. They are widely used in cast iron, steel, and nonferrous metal foundries, especially in furnaces with high capacities and little or no metal and alloy changes. The linings are partially infiltrated by the metals/alloys and hence cause contamination during an alloy change. Nonferrous metal foundries, which frequently change alloys often employ more expensive prefabricated crucibles. Precious metal foundries need prefabricated crucibles to minimize metal losses in the linings. Magnesium and its alloys are melted solely in steel crucibles due to their chemical behavior. Alloys that attack acidic linings require basic linings (Al_2O_3, MgO) or graphite crucibles. Nonconducting materials (e.g., quartz and glass) are melted in graphite crucibles which heat the charge at low temperature.

5.4.2.2. Channel Induction Furnace

In the channel induction furnace the induction coil is located under or beside the crucible and has an iron core. The coil is enclosed in a channel (Fig. 32 B). This construction acts as a transformer with a short-circuit secondary coil producing the best power transfer efficiency (0.95 to 0.98; furnace efficiency amounts to 0.9). The metal in the channel is heated and pumped back to the metal reservoir in the crucible by the repulsive force of the coil and the induced currents.

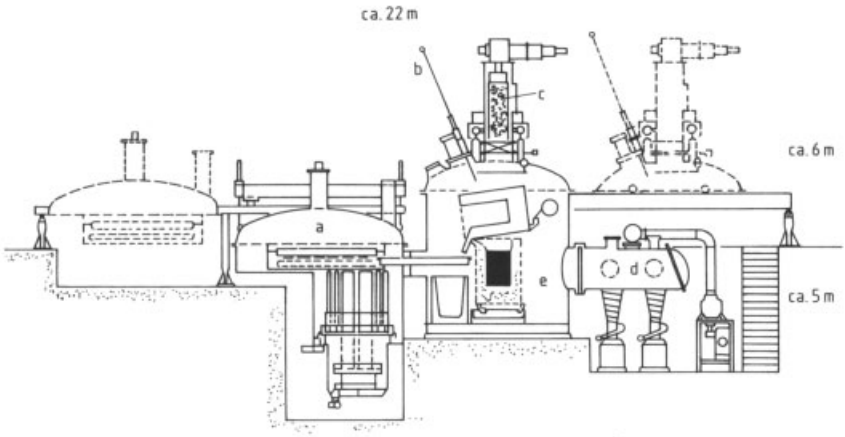

Figure 36. Multichamber vacuum induction furnace [65]
a) Mold chamber; b) Bridge breaker; c) Charging chamber; d) Vacuum pumping system; e) Melt chamber

Table 12. Process data for examples of line-frequency crucible induction furnaces [133]

Product	Pouring temperature, °C	Range of			Energy consumption,
		Capacity, t	Power, MW	Melting rate, t/h	kW · h/t
Cast iron	1500	0.75	0.3	0.47	610
		60.0	17.0	32.5	545
Steel	1650	0.75	0.25	0.32	740
		60.0	17.0	29.5	570
Copper	1200	0.6	0.15	0.32	465
		36.4	7.0	20.0	385
Brass	1000	0.6	0.15	0.51	320
		35.0	7.0	26.0	270
Aluminum	700	0.8	0.24	0.42	570
		12.0	3.8	7.3	530

Table 13. Examples of medium-frequency crucible induction furnaces [133]

Product	Pouring temperature, °C	Range of			Energy consumption,
		Capacity, t	Power, MW	Melting rate, t/h	kW · h/t
Cast iron	1500	0.08	0.16	0.23	600
		17.0	8.0	16.0	520
Steel	1650	0.08	0.16	0.21	650
		17.0	8.0	14.5	545
Copper	1200	0.15	0.16	0.32	500
		19.5	4.0	10.0	390
Brass	1000	0.15	0.16	0.41	350
		18.6	4.0	14.0	265
Aluminum	700	0.25	0.25	0.45	580
		6.2	3.0	5.7	530

This means that the furnace must always be charged with a certain amount of molten metal.

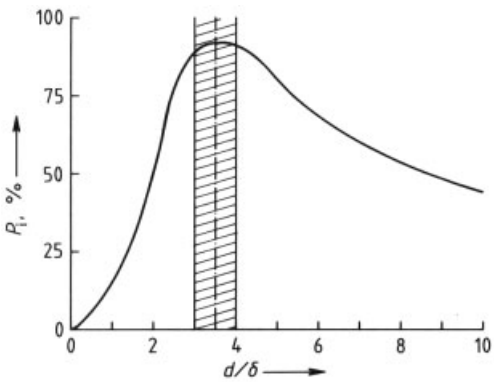

Figure 37. Power absorption P_i as a function of the standardized particle diameter d/δ

Various designs of channel induction furnaces exist: inclined and vertical channel furnaces, horizontal channel furnaces, and open channel furnaces. They are normally used as holding or casting furnaces and less frequently as melting furnaces. Due to large flow rates in the channel they are more susceptible to erosion of the refractory within the channel, however, they permit high capacities and power efficiency is greater than 80 % which is larger than that for crucible induction furnaces. Large holding furnaces may have more than one channel. The quantity of refractory materials used for channel furnaces is high. Careful maintenance of the channel is also necessary.

Metallurgical Significance. Due to the relatively low specific power requirements, channel induction furnaces are mainly used for holding and superheating cast iron and nonferrous metals, but less often for steel. Channel induction furnaces may have high capacities (up to ca. 250 t), especially in the iron industry (holding and superheating iron from blast furnaces prior to refining in the oxygen converter: up to 1500 t, 15 MW [56]). They are often used in combination with other furnaces and act as buffer furnaces. They are sometimes used as melting or

casting furnaces, these include automatic air-pressure pouring systems (Fig. 38).

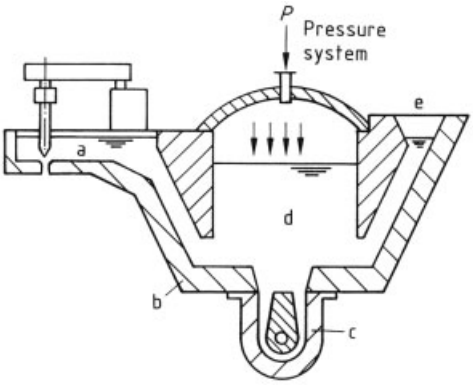

Figure 38. Automatic air-pressure pouring system [56]
a) Casting system; b) Rammed lining; c) Inductor; d) Molten metal; e) Charge system

Channel induction furnaces are not as widely applied as crucible induction furnaces in nonferrous industries. Clogging of the channel is a handicap, especially in the melting of aluminum. However, the different designs of channel inductors permit various applications, such as coating pots (zinc) or pouring furnaces.

Various data for channel induction furnaces are given in Tables 14 and 15.

5.4.2.3. Special Induction Furnaces

Reactions with the crucible or refractory material are undesirable for special high-grade metals and alloys. Use of special induction coil design and suitable frequencies and induction currents, allow metals or alloys to be melted and centered inside the coil by levitation melting. This melting method is only used on a laboratory scale for small amounts of pure metals and alloys.

Induction zone melting is used to refine very pure metals or alloys, especially for the production of semiconductors (→ Crystal Growth, Chap. 6.4.). Germanium and silicon mono crystals are produced by this technique.

5.5. Plasma Furnaces

A plasma is a partially ionized (up to 50 %) gas that contains electrons, ions, energized molecules, dissociated molecules, neutral molecules, and atoms (→ Plasma Reactions). The plasma operates at atmospheric pressure and is sufficiently conducting to permit stable transfer of electric power between two or more electrodes. The plasma attains temperatures > 2000 K (e.g., 10 000 – 20 000 K for an argon plasma operated at an arc voltage of 25 V and current of 1.5 kA).

Basic Principles and Construction [56], [61], [62]. The principal difference between plasma furnaces and arc furnaces is the use of a plasma torch instead of electrodes.

In a plasma torch thermionic electrons are emitted from a cathode and accelerated towards the anode. They collide with gas molecules and ionize them. The positively charged gas ions are accelerated in the opposite direction towards the cathode with which they collide, releasing their energy and hence sustaining the thermionic emission. Depending on the type of torch and its construction materials, water cooling may be applied. Various a.c. or d.c. torches are available; they include transferred arc (Fig. 39 A), nontransferred arc (Fig. 39 B), and superimposed arc.

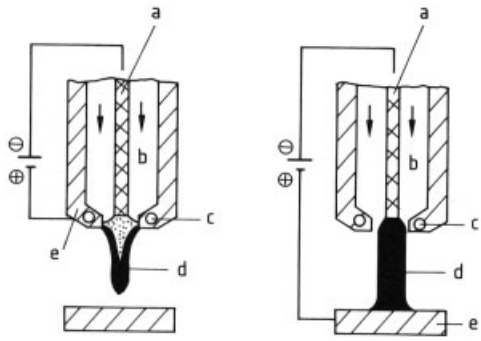

Figure 39. Transferred (A) and nontransferred (B) arc torches [83]
a) Cathode; b) Plasma gas; c) Cooling water; d) Plasma jet; e) Anode

Furnace construction is similar to that of arc resistance and arc furnaces. Refractories and other components are selected to suite the specific application.

Metallurgical Significance [62], [120], [130], [155]. Plasma furnaces have only recently been introduced into primary and secondary metallurgy. Examples include:

Table 14. Examples of channel induction furnaces for melting [133]

Product	Pouring temperature, °C	Range of			Energy consumption, kW · h/t
		Capacity, t	Power, MW	Melting rate, t/h	
Copper	1200	1.0	0.2	0.60	270
		26.0	3.3	11.2	280
Brass	1000	1.0	0.2	0.93	200
		26.0	3.3	18.5	175

Table 15. Examples of channel induction furnaces for holding and pouring cast iron (pouring temperature 1500 °C) [65].

Capacity, t	Power, MW	Throughput, t/h	Holding power, kW
10.0	0.3	7.2	145
10.0	0.5	12.0	145
115.0	1.2	27.0	390
115.0	3.0	87.0	390

1) Remelting of steel scrap (e.g., former VEB Edelstahlwerke, Freital, Germany: 35-t furnace with three plasma torches, total power input 19.8 MW, melting capacity of 20 t/h [121]; Voest – Alpine AG, Linz, Austria: 45 t-furnace (Fig. 40) with four plasma torches, total power input 24 MW, melting capacity $\leq$ 25 t/h [122]
2) Production of ferrochromium from chromite fines and the refining the high-carbon ferrochromium (e.g., 40 MVA d.c. one electrode furnace, diameter 9 m, height 3.8 m, Middelburg Steel and Alloys, South Africa) [62], [64], [123]
3) Recovery of aluminum from aluminum dross (e.g., ALCAN, Canada using one 1.5-MW torch for each of two 9-t rotating furnaces) [124]
4) Recovery of FeNiCr alloy, slag, and lead and zinc from flue dusts in the stainless steel industry (e.g., three × 6-MW torches, SCANDUST in Landskrona, Sweden producing 70 kt/a) [62], [125]
5) Plasma cold hearth melting is used for melting Ti alloys. Furnaces up to 3.5 MW are in operation [155]
6) Several units use plasma torches in research furnaces (e.g., SKF, Tetronics, Krupp) [62]

5.6. Electron-Beam Furnaces

Basic Principles [56], [58], [59], [156]. In electron-beam furnaces the kinetic energy of highly accelerated electrons in a vacuum is transferred to the surface of a material on impact. This leads to local temperature increases and melting of the material.

Furnace Construction [56]. The principle component of an electron-beam furnace is the electron gun which may be work-accelerated or self-accelerated. The self-accelerating gun is most commonly used. Here the voltage is applied between the cathode and an auxiliary hollow anode situated just below the cathode. In work-accelerated guns the anode is the material to be melted, this techniques is used to good advantage in zone melting.

The self-accelerating electron gun (Fig. 41) consists of an electron emitting cathode (a), an accelerating hollow anode (c), and electromagnetic focusing and deflection coils (d). In addition to the gun the furnace comprises of a work chamber in which melting and refining is performed in water-cooled crucibles, vacuum equipment for the electron gun and work chamber, and a computerized control system which controls and monitors the operating parameters.

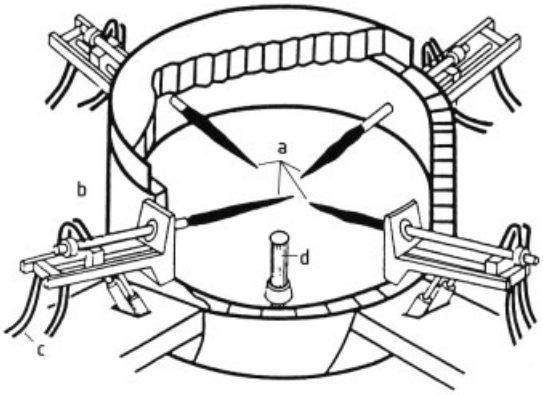

Figure 40. Schematic drawing of the Voest – Alpine furnace for remelting steel scrap [62]
a) Plasma zones; b) Plasma torch; c) Cathode; d) Anode

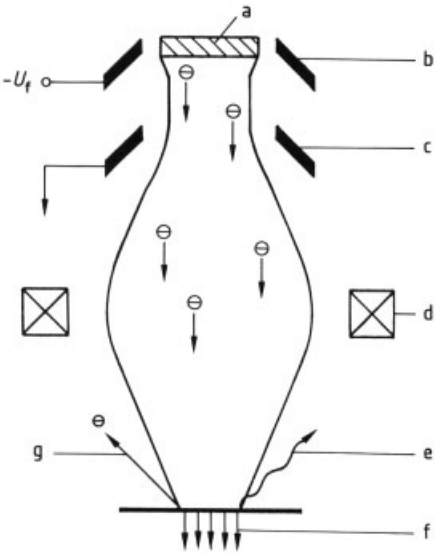

Figure 41. Self-accelerating electron gun [56]
a) Cathode; b) Control electrode; c) Anode; d) Electromagnetic lens; e) Secondary electromagnetic radiation; f) Electron beam penetration; g) Reflected and secondary electrons

Various types of self-accelerating guns and gun configurations have been developed; they include annular, axial (or Pierce), axial differential pumped, transverse and radial multifilament gun arrangements.

Feeding of material to the furnace is facilitated by drip melting of vertically or horizontally fed consumable electrodes, or bath melting of granulated material and scrap. The melted product may be continuously cast into bars or molds.

Metallurgical Significance [58], [59], [155]. Although electron-beam melting has higher investment costs per installed kilowatt hour than other melting techniques (see Table 5), it has distinct advantages. Success may be ascribed to its abilities to fuse any low-volatility material that undergoes a solid – liquid transition, and to maintain the fused condition for as long as is required to obtain the desired refining action. Production of high-quality refractory metals is possible, which is not the case with the other methods. Although investment and maintenance costs are high, they are compensated by increased quality.

The refining action in the electron-beam furnace may be classified into:

1) Removal of suboxides with high vapor pressures from their respective metals (i.e., NbO can be separated from Nb since their vapor pressure ratio at 2000 K is 10^3, whereas that for TiO/Ti is 1)
2) Removal of interstitials such as nitrogen and hydrogen
3) Removal of oxygen by carbon as carbon monoxide
4) Removal of highly volatile metals from the metal to be melted and refined (e.g., removal of Cr and Fe from Zr)
5) Removal of insolubles from the melt material and of inclusions which float to the surface of the melt; since most materials to be processed are usually clean this contribution to refining is small

In addition to the usual refractory metals (V, Nb, Ta, Mo, W, Zr, Hf) which are normally melted and/or refined in electron-beam furnaces, beryllium, cobalt, uranium, nickel, and alloys such as special steels have also been melted. For example, the refining action during the melting of hafnium in an electron-beam furnace is far superior to that in vacuum arc melting. For technical data see Table 5.

For economic reasons titanium is mostly processed via vacuum arc melting, although electron-beam furnaces are also applied. The impurities oxygen, nitrogen, carbon, and iron remaining after refining lie in similar ranges for both processes, hydrogen values are, however, far lower for electron-beam melting. Ceramics (e.g., ZrO_2 and Al_2O_3), high-quality steels, and carbides of uranium, thorium, and zirconium are also melted in electron-beam furnaces.

Electron beams are also used for metallization, vacuum deposition, centrifugal casting, machining, and welding.

5.7. Fused-Salt Electrolysis Cells

Fused electrolysis can be classified both as a resistance electrothermal reactor and as an electrochemical reactor. It is not, however, a purely electrothermal reactor, and is therefore discussed separately from resistance furnaces.

Due to their low reduction potentials elements such as aluminum, magnesium, and sodium cannot be recovered from aqueous solutions – hydrogen ions would be reduced in preference to these elements to produce hydrogen gas. Recovery of these elements by electrolysis can only be carried out in stable, nonaqueous electrolytes. In order to illustrate the technology of fused-salt electrolysis, two examples are described: the recovery of aluminum from alumina and sodium from sodium chloride. For magnesium electrolysis, see → Magnesium, Chap. 4.1.

Aluminum Recovery [76], [77], [87], [90–92]. The importance of aluminum electrolysis (→ Aluminum, Chap. 4.) is indicated by the 22.6×10^6 t of Al produced in 1998, with a power consumption of 14 MW · h/t (see Table 7).

The feed material for the production of aluminum is alumina (Al_2O_3). Due to alumina's high melting point (2050 °C), fused-salt electrolysis of pure alumina is not economically viable. The Hall – Héroult process, developed at the end of the nineteenth century, permits electrolysis at 940 – 980 °C, this temperature is produced due to the resistance of the stable salt electrolyte based on the $NaF - AlF_3 - Al_2O_3$ system to the electric current. This electrolyte system has a sufficiently low liquid temperature and a high alumina solubility. Usually about 2 – 8 % Al_2O_3 is dissolved in the cryolite-based (Na_3AlF_6) electrolyte. The density difference between the molten aluminum (ca. 2.3 g/cm^3) and the molten electrolyte (ca. 2.1 g/cm^3) is sufficiently large to ensure good separation of the molten electrolyte and molten aluminum.

The overall cell reaction for the electrolysis may be given as:

$$Al_2O_3 + xC \rightleftharpoons 2\,Al + (3-x)\,CO_2 + (2x-3)\,CO \quad 1.5 \le x \le 3$$

Usually the anode gas is a mixture of carbon dioxide and monoxide. Use of carbon anodes lowers the required electrical energy, at the expense of carbon anode consumption (ca. 0.4 kg C/kg Al).

At the normal current density (0.6 – 1.0 A/cm^2) an additional anodic overpotential of ca. 0.5 V must be added to the theoretical potential of the above reaction (1.06 – 1.18 V at 977 °C) as well as a cathodic overpotential of about 0.5 V. This implies that the effective electrolysis potential or polarization potential is ca. 1.7 V. If the voltage drops across the electrodes, then electrolyte, conductors, and anode effects are included, the potential required for aluminum electrolysis lies in the range 3.9 – 4.6 V.

A typical electrochemical cell for the production of aluminum consists of:

1) A carbon-lined steel vessel (carbon blocks at the bottom of the furnace 40 cm deep)
2) Two rows containing a total of 16 – 24 consumable prebaked carbon anodes
3) Molten salt electrolyte (cryolite-based)
4) A liquid aluminum cathode
5) A layer (20 – 40 cm) of insulation (fire clay refractory) at the bottom of the furnace
6) Anode and cathode bus bars

In modern cells (Fig. 42) small amounts of alumina are fed between the two rows of sus-

pended anodes at frequent intervals. Crust breakers between the two rows of anodes break the frozen bath crust, a ventilation shaft removes the 5000 – 8000 m^3/h of anode gas evolved per cell.

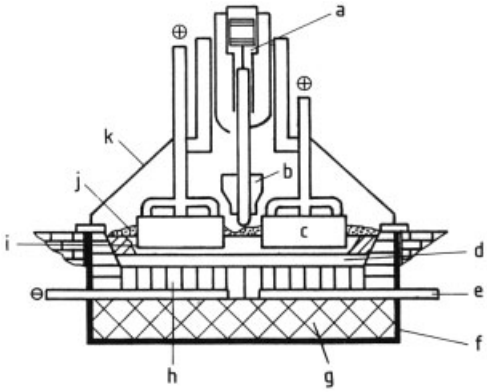

Figure 42. Typical Hall – Héroult aluminum production cell [76]
a) Charging system; b) Crust breaker; c) Graphite anodes; d) Molten aluminum; e) Bus bar; f) Steel container; g) Refractory; h) Graphite bottom (cathode); i) Molten electrolyte; j) Frozen electrolyte alumina crust; k) Cover hood

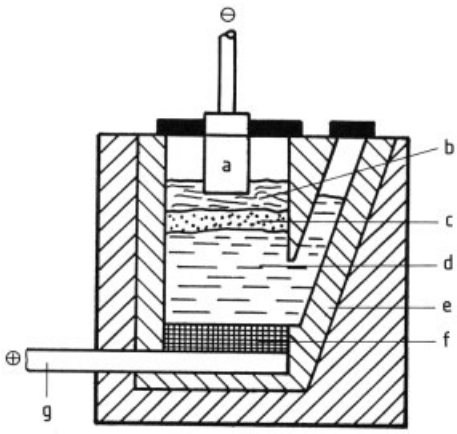

Figure 43. Typical aluminum refining cell [76]
a) Graphite cathode; b) High purity aluminum; c) Electrolyte; d) Anode alloy; e) Refactory; f) Carbon base; g) Anode bus bar

High-purity aluminum is produced in a refining cell [76]. This cell (Fig. 43) comprises the following elements:

1) A carbon base (f), which acts as the anode
2) An anode alloy (d) containing 30 % copper (density at 750 °C 3.4 – 3.7 g/cm^3) above the anode
3) An AlF_3–NaF – BaF_2– CaF_2-based electrolyte (c) (density 2.7 – 2.8 g/cm^3) in the middle
4) High-purity molten aluminum (b) (density 2.3 g/cm^3)

It is operated at a current density of 0.2 – 0.5 A/cm^2 and at a temperature of 750 °C, maintained by heat liberated due to the resistance of the electrolyte to the electric current.

Sodium Recovery [76]. Sodium is recovered from molten sodium chloride by fused-salt electrolysis in a Downs cell (Fig. 44); see also → Sodium. The salt is maintained in the molten state due to the resistance of the electrolyte to the electric current.

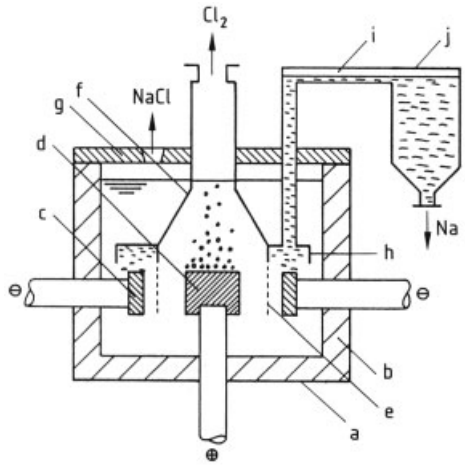

Figure 44. Downs cell for the production of sodium and chlorine from fused sodium chloride [76]
a) Steel container; b) Ceramic refractory; c) Iron cathode; d) Graphite anode; e) Diaphragm; f) Chlorine collection hood; g) Isolation lid; h) Collection bell for sodium; i) Sodium overflow; j) Sodium collector

The electrolyte is a molten mixture of $BaCl_2$, $CaCl_2$, and NaCl (*mp* ca. 600 °C). A cathode cylinder (c) is concentric with a graphite anode (d), the interelectrode gap being ca. 50 mm. A diaphragm (e) in the gap between the anode and cathode is attached to collection bells, allowing separate recovery of sodium and chlorine. Since sodium has a lower density (0.8 g/cm^3) than the molten electrolyte (2.7 g/cm^3), it drifts to the surface of the molten electrolyte where it is recovered in the collection bell.

These cells are typically run at current of up to 45 kA and potentials of 6.5 – 7 V, producing

about 800 kg sodium per day. Typical energy requirements lie between 9.8 and 10 kW · h/kg Na.

5.8. Modeling

The complex nature of the mass transfer processes taking place in electrothermal reactors makes it rather difficult to model them. However, fluid flow has been modeled using fundamental equations (e.g., Navier – Stokes), for example in Hall – Héroult aluminum recovery cells and in induction furnaces [96], [127], [128].

6. References

1. G. Johannsen, *Metall und Erz* **36** (1939) 325 – 333.
2. W. Heiligenstaedt: *Wärmetechnische Rechnungen für Industrieöfen,* Verlag Stahleisen, Düsseldorf 1951.
3. W. H. Duda: *Cement Data Book,* Bauverlag, Wiesbaden 1977, pp. 319 – 378.
4. K. E. Peray, J. J. Waddel: *The Rotary Cement Kiln,* Chemical Publishing Co., New York 1972, p. 1.
5. W. Flügge: *Statik und Dynamik der Schalen,* 3rd. ed., Springer Verlag, Heidelberg 1962.
6. R. Schardt, J. Steingaß, *Stahlbau* 1970, no. 3, 65 – 73; 1970, no. 5, 146 – 150.
7. J. Steingaß, Dissertation, TH Darmstadt, 1972.
8. H. Mettler, *Zem.-Kalk-Gips* 1969, no. 7, 306 – 311.
9. H. Mettler, Dissertation, RWTH Aachen, 1968.
10. V. Ramamurti, K. Raganatha Sai, *Zem.-Kalk-Gips* 1978, no. 9, 433 – 435.
11. K. Inove, T. Matsumoto, T. Ozaki, Technical Revue, no. 73, Mitsubishi Heavy Ind., 1972.
12. H. Meedom, *Zem.-Kalk-Gips* 1976, no. 12, 568 – 571. I. Gulida, *Izv. Vyssh. Uchebn. Zaved Mashinostr.* 1971, no. 8, 47 – 49.
13. C. Grinfeder, *Miner. Process.* (1968) no. 1, 14 – 19.
14. V. Ramamurti, L. S. Gupta, *Zem.-Kalk-Gips* 1978, no. 12, 614.
15. M. Garyk, M. Genda, *Zem.-Kalk-Gips* 1978, no. 9, 436 – 438.
16. E. Steinbiß, *Zem.-Kalk-Gips* 1976, no. 7, 321 – 328.
17. W. Bonn, B. Saxer, *Zem.-Kalk-Gips* 1978, no. 7, 329 –332.
18. K. W. Liebler, *Zem.-Kalk-Gips* 1978, no. 7, 565 – 567.
19. H. Paulsen, *Zem.-Kalk-Gips* 1971, no. 5, 204 – 207.
20. H. Xeller, H. Jöhnk, *Zem.-Kalk-Gips* 1976, no. 12, 557 – 564.
21. G. Weislehner, *Zem.-Kalk-Gips* 1976, no. 10, 443 –446.
22. W. Schroebler, *Zem.-Kalk-Gips* 1974, no. 2, 41 – 46.
23. H. Erni, *Zem.-Kalk-Gips* 1974, no. 10, 486 – 498.
24. G. Routschka, A. Majdic, *Zem.-Kalk-Gips* 1974, no. 10, 469 – 485.
25. G. Hotz, P. Bartha, *Zem.-Kalk-Gips* 1975, no. 6, 236 – 240.
26. M. Künnecke, H. Naefe, *Zem.-Kalk-Gips* 1969, no. 6, 271 – 275.
27. A. Künnecke, A. Kleinevoss, K. Wieland, *Zem.-Kalk-Gips* 1976, no. 7, 298 – 304.
28. G. Hotz, H. J. Schmidt, *Zem.-Kalk-Gips* 1976, no. 7, 305 – 307.
29. H. Barthel, *Zem.-Kalk-Gips* 1976, no. 7, 308 – 312.
30. R. Zengler, Dissertation, TU Clausthal, 1974.
31. F. W. Plank, *Zem.-Kalk-Gips* 1969, no. 75 – 81. T. Gori, F. Schwarzkopf, *Zem.-Kalk-Gips* 1978, no. 8, 392. S. Levine, *Pit & Quarry* 1977.
32. L. M. Ludera, *Zem.-Kalk-Gips* 1978, no. 12, 606 – 612.
33. *Handelsblatt, Westdeutsche Wirtschaftszeitung,* Düsseldorf, July 29, 1977.
34. G. A. Schroth, T. K. Diener, *Rock Prod.* (1976) 60 –62, 64.
35. L. Kwech, *Zem.-Kalk-Gips* 1977, no. 12, 597 – 607. *Cement, Lime, Gravel* **49** (1974) no. 1, 3 – 6.
36. E. Bomke, *Zem.-Kalk-Gips* 1978, no. 12, 589 – 594.
37. K. Wilck, *Refra Colloquium* 1976.
38. G. K. Kapoor, *Zem.-Kalk-Gips* 1978, no. 12, 602 –605. H. Ritzmann, *Zem.-Kalk-Gips* 1976, no. 5, 207 –212. G. Hotz, *Refra Colloquium,* 1976.
39. H. Serbent, H. Krainer, *Techn. Mitt. Krupp, Forschungsber.* **23** (1965) 49 – 68.
40. Krakatau Steel: "Current Status of Coal-Based Direct Reduction for Iron Making," *Iron Making Technology Seminar,* Cilegon, Indonesia, May 7 – 9, 1990. G. Reuter, W. Schnabel, H. Serbent, *Stahl Eisen* **98** (1978) 1232 – 1237. *Ullmann,* 4th ed., **3,** 415 – 432.
41. G. Kossek, H. Lommert, H. Serbent, *Erzmetall* **32** (1979) 135 – 139.

42. H. Maczek et al., *Stahl Eisen* **96** (1976) 1233 – 1238.
43. K. Meyer, H. Pietsch, *Erzmetall* **27** (1974) 345 – 353.
44. A. Vaillant, H. W. Bodlaender, *Aufbereit. Tech.* (1972) no. 10, 631 – 638. G. Dumont, P. R. Belanger, *JEC Process Design & Development,* **17** (1978) no. 2, 107 – 114.
45. Kowa Seiko Pyrite Cinders Treatment Process, *Sulphur,* no. 83, July/Aug. 1969, 34 – 35, 38.
46. H. W. Gudenau, Eisenhüttenmännische Verfahrenstechnik Hochofen, TH Aachen.
47. L. V. Bogdandy, H. J. Engell: *Die Reduktion von Eisenerzen,* Verlag Stahl-Eisen, Düsseldorf 1967.
48. *Ullmann,* 4th ed., **3,** 395.
49. W. Weißbach: *Werkstoffkunde und Werkstoffprüfung* 1988, 9th ed., Vieweg, Wiesbaden.
50. *Ullmann,* 4th ed., **3,** pp. 395 – 432.
51. M. Hirsch, Verfahrenstechnische Merkmale des EOF-Stahlerzeugungsverfahrens, Lecture held at Lurgi-Maitagung, Frankfurt, May 21, 1986.
52. Korf Lurgi Steel Engineering, *Stahl Eisen* **110** (1990) no. 7.
53. Bernt Rollinger, *Handelsblatt,* **no. 219,** Nov. 13, 1987
54. W. Borchers: *Die elektrischen Öfen,* Verlag von Wilhelm Knapp, Halle a.S. 1907.
55. K. Kegel: *Electroofentechnik in der Metallurgie,* VCH Verlagsgesellschaft, Weinheim 1988.
56. M. Orfeuil: *Electric Process Heating–Technologies/Equipment/Applications,* Batelle Press, Columbus Richland 1987.
57. G. Hoyle: *Electroslag Processes – Principles and Practice,* Applied Science Publishers, London 1983.
58. O. Winkler, R. Bakish (eds.) : *Vacuum Metallurgy,* Elsevier, Amsterdam 1971.
59. R. Bakish: *Introduction to Electron Beam Technology,* John Wiley and Sons, New York 1962.
60. V. Dembovsky: Plasma Metallurgy – The Principles, *Material Science Monogr.* **23,** Elsevier, Amsterdam 1985.
61. S. M. L. Hamblyn: "Plasma Technology and its Application to Extractive Metallurgy," *Min. Sci. Engng.* **9** (1977) no. 8, 151 – 176.
62. N. A. Barcza: "The Development of Large-scale Thermal Plasma Systems," *J. S. Afr. Inst. Min. Metall.* **86** (1986) no. 8, 317 – 333.
63. *Ullmann,* 4th ed., **3,** pp. 519 – 542.
64. H. Walde: *Elektrische Stoffumsetzungen in Chemie und Metallurgie in energiewirtschaftlicher Sicht,* Klepzig Verlag, Düsseldorf 1968.
65. "Elektroofentechnik in der Metallurgie," 19. *Metallurgische Seminar der Gesellschaft der Metallhütten- und Bergleute,* p. 130.
66. C. A. Otto: *Electric Furnaces,* George Newnes, Strand 1958.
67. J. H. Brunklaus, F. J. Stepanek: *Industrieöfen – Bau und Betrieb,* Vulkan-Verlag, Essen 1986.
68. A. K. Biswas, W. G. Davenport: *Extractive Metallurgy of Copper,* Pergamon International Library, Oxford 1976.
69. J. C. Yannopoulos, J. C. Agarwai (eds.): *Extractive Metallurgy of Copper – Pyrometallurgy and Electrolytic Refining. An International Symposium,* The Metallurgical Society of AIME, New York 1976.
70. P. Queneau (ed.): *Extractive Metallurgy of Copper, Nickel and Cobalt,* Interscience Publisher, New York 1961.
71. Rheinisch-Westfälisches Elektrizitätswerk AG, Elektrothermische Reduktion, company brochure, Essen 1985.
72. Rheinisch-Westfälisches Elektrizitätswerk AG, Elektroprozeßwärmeverfahren in der Industrie, company brochure, Essen 1991.
73. F. Denkhaus: "Herstellung hochreiner Stähle durch den Einsatz von Schrott beim Elektroschlacke-Umschmelzen," Dissertation, RWTH Aachen 1987.
74. Durrer, G. Volkert, K.-D. Frank: *Metallurgie der Ferrolegierungen,* 2nd ed., Springer-Verlag, Berlin 1972.
75. V. P. Elyutin, Yu. A. Pavlov, B. E. Levin, E. M. Alekseev: *Production of Ferroalloys/Electrometallurgy,* 2nd ed., The Israel Program for Scientific Translations, Jerusalem 1961.
76. *Winnacker Küchler,* 4th ed., **4,**Hansa, München 1986 p. 724.
77. F. Pawlek: *Metallhüttenkunde,* De Gruyter, Berlin 1983.
78. M. Rudolph, H. Schaefer: *Elektrothermische Verfahren–Grundlagen/Technologien/Anwendungen,* Springer-Verlag, Berlin 1989.
79. H. Bollinger, W. Teubner: *Industrielle Vakuumtechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leibzig 1980.

80. H. Wenzel: *Induktives Schmelzen und Warmhalten von Gußeisen,* RWE AG, Essen 1987.
81. J. Fasholz, G. Orth: *Induktive Erwärmung – Physikalische Grundlagen und technische Anwendungen,* RWE, AG, Essen 1991.
82. R. F. Baddour, R. S. Timmins: *The Application of Plasmas to Chemical Processing,* MIT Press, Massachusetts 1967.
83. D. Neuschütz: *Development of 3-phase a.c. Plasma Furnaces at Krupp, Iron Steel Eng.* **61** (1985), no. 5, 27 – 33.
84. M. Xu: "Massnahmen zur Erhöhung der Reduktionsgeschwindingkeit NE-metallhaltiger Schlacken," Dissertation, RWTH Aachen 1991.
85. J. Höfler: "Reduktion NE-metallhaltiger Schlacken im Eletroofen unter besonderer Berücksichtigung der Verwendung gleichgerichteten Schmelzstroms," Dissertation, RWTH Aachen 1988.
86. C. B. Gill: *Nonferrous Extraction Metallurgy,* Wiley & Sons, New York 1980.
87. P. C. Hayes: *Process Selection in Extractive Metallurgy,* Hayes Publishing Co., Brisbane 1985.
88. H. W. Gudenau: *Eisenhüttenmännische Verfahrenstechnik,* Institut für Eisenhüttenkunde, Aachen 1985.
89. E. Decker: "Rationeller Einsatz von Elektrowärme in der Metallindustrie," *VDI-Fachtagung: Rationelle Energieanwendung in der Industrie,* Essen 1981.
90. K. Grjotheim et al.: *Aluminium Electrolysis – The Chemistry of the Hall-Heroult Process,* Aluminium-Verlag GmbH, Düsseldorf 1977.
91. K. Grjotheim, H. Kvande: *Understanding the Hall-Heroult Process for Production of Aluminium,* Aluminium-Verlag GmbH, Düsseldorf 1986.
92. K. Grojtheim, B. J. Welch: *Aluminium Smelter Technology-A Pure and Applied Approach,* Aluminium-Verlag GmbH, Düsseldorf 1980.
93. E. Höffken, F.-J. Strunck, J. Wendorff, J. Wolf: "Das Edelstahlwerk der Thyssen Stahl AG in Oberhausen und seine Entwicklung von 1979 – 1987," *Stahl Eisen,* **108** (1988) no. 6, 281 – 288.
94. K.-H. Brokmeier: *Induktives Schmelzen,* Verlag W. Girardet, Essen 1966.
95. *Proceedings of "Junker-Ofentagungen,"* Otto Junker GmbH, Lammersdorf.
96. *Proceedings of "Light Metals,"* The Minerals and Metals Society, Warrendale, Pa.
97. G. Rath: "Der Elektroreduktionsofen – Ein Aggregat für die Nichteisenmetallurgie," *Erzmetall* **43** (1990) 81 – 85.
98. G. N. Angelopoulus, S. H. Pavlou, D. C. Papamantellos: "Ferronickel," *Erzmetall* **42** (1989) 107 – 113.
99. H. Nakamura, S. Hirakawa, E. Nomura: "Electrothermal Slag-fuming Process and Zinc Leach Residue Treatment by Lead Blast Furnace at Kamioka Smelter," *TMS Paper Selection,* A 77 – 9, The Metallurgical Society of AIME, New York 1977.
100. H. P. Rajcevic, W. R. Opie: "Development of Electric Furnace Slag Cleaning at a Secondary Copper Smelter," *J. Met.* 34 (1982) March, 54 – 56.
101. P. F. Feri, A. Perillo: "The New Lead Smelter of Portovesme," *Extraction Metallurgy '85,* The Institute of Mining and Metallurgy, London 1985.
102. T. P. Battle, J. P. Hager: "Viscosities and Activities in Lead-Smelting Slags," *Metallurgical Transactions B,* **21 B** (1990) 501 – 510.
103. W. P. Channon, R. C. Urquhart, D. D. Howat: "The Mode of Current Transfer Between Electrode and Slag in the Submerged-Arc Furnace, *J. S. Afr. Inst. Min. Metall.* **74** (1974) no. 8, 4 – 7.
104. G. P. Tyroler, C. A. Landolt (eds.): *Proceedings of Extractive Metallurgy of Nickel and Cobalt,* Warrendale, Pa., 1988.
105. *Proceedings of Symposium Nickel,* GDMB, Clausthal-Zellerfeld, Sept. 24 – 25, 1970.
106. D. J. I. Evans, R. S. Shoemaker, H. Veltman (eds.): *Proceedings of Laterite Symposium,* AIME, New York 1979.
107. T. S. Mackey, D. Prengaman (eds.): *Proceedings of Lead-Zinc '90,* TMS, Warrendale 1990.
108. W. G. Davenport, E. H. Partelpoeg: *Flash Smelting-Analysis, Control and Optimization,* Pergamon Press, New York 1987.
109. H. Brod, F. Kempkens, H. Strohschein: "Energierückgewinnung aus einem UHP-Elektrolichtbogenofen," *Stahl Eisen* **109** (1989) 229 – 238.
110. K. Rüdinger: "Erschmelzen von Titan im Lichtbogenvakuumofen," *Metall* **42** (1988) 774 – 779.
111. D. Meredith et al., "Der neue Gleichstrom-Lichtbogenofen bei der Florida

Steel Corporation," *Stahl Eisen* **108** (1988) 796 – 800.
112. "Stahl & Eisen," *Jahrbuch Stahl 1991,* Verlag Stahleisen mbH, Düsseldorf 1990.
113. E. Decker: "Energiebedarf und Wirtschaftlichkeit beim Schmelzen von Gußeisen in Kupol- und Induktionsöfen," *Elektrowärme International* **38** (1980) B 2, 92 – 97.
114. A. Dickopp: "Überblick über elektrisch beheizte Anlagen zum Feuerverzinken von Stahlteilen," *Elektrowärme International* **32** (1974) B 3.
115. E. Decker, A. Dickopp: "Energieeinsatz und Wärmerückgewinnung beim Schmelzen von Aluminium," *Aluminium* **57** (1981) 3.
116. E. Decker, A. Dickopp: "Induktionstiegelöfen für sinnvolle Energieausnutzung in der Eisengießerei," *Maschinenmarkt* **87** (1981) 30.
117. H. Ronig: "Induktionsöfen für das Warmhalten und Vergießen von Leichtmetall," *Metall* **37** (1983) no. 5, 459 – 465.
118. D. Hartmann, H. Rohn: "Der Netzfrequenz-Induktions-Tiegelofen zum Schmelzen von Nichteisen-Schwermetall-Legierungen," *Zeitschr. Metallkunde* **48** (1957) no. 3, 85 – 90.
119. *Conference Proceedings of the "International Conference on Vacuum Metallurgy,"* Tokyo 1973, München 1982, Linz 1985.
120. C. W. Chang, J. Szekely: "Plasma Applications in Metals Processing," *J. Met.* **34** (1982) Feb. 57 – 64.
121. F. Esser, H. Fiedler, W. Lachner: "Zur Theorie des Ein- und Niederschmelzen festen, metallischen Einsatzguts in Plasmaöfen – ein Beitrag zur Verfahrensoptimierung des Plasmaprimärschmelzens," *Neue Hütte* **19** (1974) no. 10, 577 – 586.
122. *Proceedings 8th International Conference on Vacuum Metallurgy,* Linz, Austria, 1985.
123. "A world first – The DC plasma arc furnance," *The S. Afr. Mechanical Engineer* **39** (1989) no. 5, 223 – 228.
124. S. Lavoir, C. Dubé, G. Dubé: "The ALCAN Plasma Dross Treatment Process," Light Metals 1991, E. L. Rooy (ed.): The Minerals and Metals Society (1990) 981 – 985.
125. SKF, Publications and brochures on PLASMARED, Landskrona, Sweden.
126. H. H. Kellog in N. J. Themelis, P. F. Duby (eds.): *International Symposium on Quantitative Description of Metal Extraction Processes,* The Minerals, Metals & Materials Society, Warrendale 1991.
127. J. Szekely: "The Role of Mathematical Models in New Process Development," in M. Koch, J. C. Taylor (eds.): *Productivity and Technology in the Metallurgical Industries,* The Minerals, Metals & Materials Society, Warrendale 1989, pp. 105 – 143.
128. J. Szekely: "The 1987 Extractive Metallurgy Lecture," *Met. Trans.* **19 B** (1988) 525 – 540.
129. G. K. Bhat (ed.): *Proceedings of the Conference "Special Melting and Processing Technologies",* San Diego, CA, Noyes Publications, New Jersey 1989.
130. A. Choudhury: *Vacuum Metallurgy,* ASM International, 1990.
131. Mannesmann Demag Hüttentechnik, Brochure, Duisburg, Germany.
132. P. Hanniala, T. Mäkinen, M. Kytö: "Flash Technology for Converting," in J. Vereecken (ed.): *EMC '91: Non-Ferrous Metallurgy – Present and Future,* Elsevier Science Publishers Ltd., New York 1991, pp. 191 – 203.
133. Rheinisches-Westfälisches Elektrizitätswerk AG, Inductive Erwärmung, company brochure Essen 1991.
134. T. S. Mackey, R. D. Prengaman (eds.): *Lead – Zinc '90,* TMS, Warrendale 1990, pp. 453 – 528.
135. J. Dutrizac, V. Ramachandran, A. Gonzales: *Lead – Zinc 2000,* TMS, Warrendale 2000.
136. J. K. Brimacombe, A. P. Watkinson, *Metall. Trans. B* **9 B** (1978) 201 – 219.
137. P. V. Barr, J. K. Brimacombe, A. P. Watkinson, *Metall. Trans. B* **20 B** (1989) 391 – 419.
138. A. P. Watkinson, J. K. Brimacombe, *Metall. Trans. B* **13 B** (1982) 369 – 378.
139. V. Venkateswaran, J. K. Brimacombe, *Metall. Trans. B* **8 B** (1977) 387 – 398.
140. J. D. Mc Cain, J. M. Floyd (eds.): *Converting, Fire Refining and Casting,* TMS, Warrendale 1993.
141. W. J. Chen, C. Diaz, A. Lurashi, P. J. Mackey (eds.): *Pyrometallurgy of Copper,* **Vol. IV, Copper 95,** CIM 1995.
142. G. A. Eltringham, N. L. Piret, M. Sahoo (eds.): *Plenary Lectures, Copper '99,* TMS, Warrendale 1999.
143. D. B. George, W. J. Chen, P. J. Mackey, A. J. Weddick (eds.): *Smelting Operations and Advances, Copper '99,* TMS, Warrendale 1999.
144. D. B. George, J. C. Taylor (eds.): *Copper Smelting – An Update,* TMS, Warrendale 1981.

145. J. K. Brimacombe, P. J. Mackey, G. J. W. Kor, C. Bickert, M. G. Ranade (eds.): *Savard/Lee International Symposium on Bath Smelting,* TMS, Warrendale 1992.
146. K. Tozawa (ed.): *Zinc "85",* MMIJ, Sendai, Japan 1985.
147. J. Dutrizac, J. A. Gonzales, G. L. Bolton, P. Hancock (eds.): *Zinc and Lead Processing,* CIM, Montreal 1998.
148. J. Dutrizac, V. Ramachandran, A. Gonzales (eds.): *Lead – Zinc 2000,* TMS, Warrendale 2000.
149. A. K. Biswas, W. G. Davenport: *Extractive Metallurgy of Copper,* 3rd ed., Pergamon, 1994.
150. K. Krone: *Aluminium-recycling,* Vereinigung Deutscher Schmelzhütten e.V., Düsseldorf 2000.
151. Intensivierung metallurgischer Prozesse, 37. Metallurg. Seminar, Heft 87, Schriftenreihe der GDMB, Clausthal-Zellerfeld 2000.
152. J. A. Asteljoki, R. L. Stephens (eds.): *Sulfide Smelting 98/Proc. Symp.,* TMS, Warrendale 1998.
153. R. G. Reddy, R. N. Weizenbach, C. A. Landolt (eds.): *Extractive Metallurgy of Copper, Nickel and Cobalt,* TMS, Warrendale 1993.
154. K.-H. Heinen: *Elektrostahlerzeugung,* 4. ed., Stahl & Eisen, Düsseldorf 1997.
155. R. Bakish (ed.): *Proc. Conf. Electron. Beam Melting and Refining,* Bakish Materials Corp., Englewood, NJ 1997.
156. S. Schiller, U. Heisig, S. Panzer: *Electron Beam Technology,* J. Wiley & Sons, New York 1982.
157. Schlacken in der Metallurgie, Heft 83, Schriftenreihe der GDMB, Clausthal-Zellerfeld 1999.

Biochemical Engineering

Martin Krahe, Bioengineering AG, 8636 Wald, Switzerland

1. Introduction 4
2. Requirements 5
2.1. Microorganisms: Growth and Bioreaction 5
2.1.1. Biomass Concentration – Cell Density 5
2.1.2. Unstructured Model of Growth . . . 5
2.1.2.1. Monod Model 6
2.1.2.2. Inhibition Kinetics 6
2.1.2.3. Nutrient Uptake Rate 6
2.1.3. Aerobic Bioreactions 6
2.1.3.1. Stoichiometry of an Aerobic Growth without Product Formation 7
2.1.3.2. Oxygen Uptake Rate (*OUR*) 8
2.1.3.3. Generated Heat 8
2.2. Sterility 9
2.2.1. Thermal Microbial Death Rates . . 9
2.2.2. Steam Sterilization Processes 10
2.2.2.1. Sterilization in the Autoclave 10
2.2.2.2. "Full" Sterilization In Place (SIP) . 11
2.2.2.3. "Empty" Sterilization In Place (SIP) 11
2.2.2.4. Continuous Water/Steam Sterilization 11
2.2.3. Sterile Filtration 11
2.2.4. Incinerators 11
2.3. Cleaning 12
2.3.1. Manual Cleaning 12
2.3.2. Cleaning in Place (CIP) 12
2.4. Homogenous Conditions due to Mixing 14
2.4.1. Mechanical Mixing 15
2.4.1.1. Stirred Tank Reactors – Energy Dissipation 15
2.4.1.2. Vibromixer 16
2.4.1.3. Hydraulic Mixing 16
2.4.2. Pneumatic Mixing 16
2.5. Gas – Liquid Mass Transfer 17
2.5.1. Solubility of Oxygen and Other Gases 17
2.5.2. Oxygen Transfer Rate for Stirred Tank Reactors 17
2.5.2.1. Liquid Mass-Transfer Coefficient k_L 17
2.5.2.2. Specific Interfacial Area a 18
2.5.3. Aeration of Stirred Tank Reactors . 18
2.5.3.1. Gas Hold-up and Foaming in Stirred Reactors 18
2.5.3.2. Volumetric Mass-Transfer Coefficient in Stirred Reactors 18
2.5.4. Volumetric Mass-Transfer Coefficient in Airlift Reactors 19
2.6. Measurement and Control Loops 19
2.6.1. Temperature Measurement and Control 19
2.6.2. Agitation Control 20
2.6.3. pH Measurement and Control 21
2.6.4. Measurement and Control of Dissolved Oxygen 21
2.6.5. Measurement and Control of the Volume or the Level 22
2.6.6. Measurement and Control of Gas-Flow Rate 23
2.6.7. Measurement and Control of Liquid Flows 23
2.6.8. Other Measurement Systems 24
2.7. Containment 24
3. Processes and Related Equipment 24
3.1. Media Preparation and Sterilization 24
3.1.1. Media Preparation Techniques . . . 24
3.1.2. General Procedure of a "Full" SIP . 25
3.1.3. General Procedure of an "Empty" SIP 27
3.2. Inoculation 27
3.3. Batch Cultivation 27
3.4. Fed-Batch Cultivation 28
3.5. Continuous Cultivation 29
3.6. Continuous Cultivation with Cell Retention 30
3.6.1. Filtration Systems for Suspended Organisms 30
3.6.1.1. Cross-Flow Filtration 30
3.6.1.2. Rotor Filter 30
3.6.2. Immobilized Organisms 31
3.6.2.1. Pellets, Porous Matrices, or Beads . 31
3.6.2.2. Fluidized-Bed Reactors 32
3.6.2.3. Fixed-Bed Reactors 32
3.7. Dialysis Cultivation 33
3.8. Selection of Equipment Related to Specific Processes or Products . . 33
4. Materials 34
4.1. Austenitic Steel 34
4.1.1. Different Alloys of Austenitic Steels 34
4.1.2. Welding 36
4.1.3. Surfaces of Austenitic Steels 38
4.1.3.1. Definition and Measurement of Surface Roughness 38

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_381

4.1.3.2. Mechanical Surface Treatment . . . 41
4.1.3.3. Chemical Surface Treatment 41
4.1.3.4. Electrochemical Surface Treatment 42
4.1.4. Storage and General Working Rules 42
4.2. Polymers 42
4.3. Other Materials 43
4.3.1. Glass 43
4.3.2. Grease and Lubricants 43
5. System Components and Detailed Engineering 43
5.1. The Vessel 44
5.1.1. Components Mounted or Welded on the Vessel 44
5.1.1.1. Seals 44
5.1.1.2. Ports 46
5.1.1.3. Viewing Glasses 46
5.1.1.4. Manways 46
5.1.1.5. Heating/Cooling Jackets 46
5.1.1.6. Heating/Cooling Coils 47
5.1.1.7. Pressure Relief and Safety Devices 47
5.1.2. Standard Stirred Tank Reactor . . . 47
5.1.3. Standard Airlift Reactor 48
5.1.4. Glass Vessels 49
5.1.5. Special Polyamide Film Sleeve Bioreactors 49
5.2. Agitation System Design 50
5.2.1. Top or Bottom Drive? 50
5.2.2. Mechanical Seals versus Magnetic Coupling 51
5.2.2.1. Shafts with a Mechanical Seal . . . 51
5.2.2.2. Magnetic Coupled Shafts 53
5.3. Transfer Lines 53
5.3.1. Needle Connection 53
5.3.2. Steam-Sterilizable Transfer Line . . 53
5.3.3. Steam-Sterilizable Transfer Line with CIP 53
5.3.4. Manually Operated Transfer Panel . 54
5.3.5. Ring Transfer Panel 55
5.4. Valves 55
5.4.1. Diaphragm Valves 55
5.4.1.1. The Valve Body 57
5.4.1.2. Diaphragms 57
5.4.2. Bellows-Type Sealed Valves 57
5.4.3. Piston Valves 58
5.4.4. Ball and Butterfly Valves 58
5.4.5. The Actuators 59
5.4.6. Relieve Valves, Safety Valves 60
5.4.7. Check Valves (Nonreturn Valves) . 60
5.5. Pipes 60
5.5.1. Pipe and Tube Sizes 60
5.5.2. Pipe Applications and Sloping of Lines 60
5.5.3. Bends, Reducing-Fittings, T-Piece Connectors 61
5.5.4. Sterile Coupling of Pipes 62
5.6. Gassing Devices for Bioreactors . 62
5.6.1. Tubes and Ring Spargers 62
5.6.2. Sinter Metal and Porous Membrane Spargers 62
5.6.3. Surface Gassing 63
5.6.4. Bubble-Free Membrane Gassing . . 64
5.7. Mechanical Foam Separators . . . 64
5.8. Sterile Filtration 65
5.8.1. Depth Filters 65
5.8.2. Membrane Filters 65
5.9. Hydraulic Pumps 66
6. References 68

Abbreviations

AF	Anti-foam
AISI	American Iron and Steel Institute
ASME	American Society of Mechanical Engineers
BgVV	Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin
BPVC	Boiler and Pressure Vessel Code
cdw	Cell dry weight
CFR	Code of Federal Regulations (USA)
CHO	Chinese hamster kidney
CIP	Cleaning in place
cww	Cell wet weight
ELISA	Enzyme linked immunosorbent assay
EPDM	Ethylene-propylene-diene-rubber
FIC	Flow indicator and controller
GILSP	Good Industrial Large Scale Practice
GMP	Good Manufacturing Practice
GTAW	Gas tungsten arc-welding
HE	Heat-exchanger
HEPA	High efficiency particulate air (filter)
IIW	International Institute of Welding
MDBK	Madin darby bovine kidney
MV	Measured value
NC	Normally closed
NO	Normally open
OD	Optical density
P	Controller with a proportional algorithm

PED	Pressure Equipment Directive
PI	Controller with a proportional and an integration algorithm
PI	Pressure indicator (pressure gauge)
PIC	Pressure indicator and controller
PID	Controller with a proportional an integration and a derivative algorithm
PLC	Programmable logic controller
PTFE	Polytetrafluoroethylene
pO_2	Partial pressure of dissolved oxygen
Pt-100	Temperature measurement sensor
SCBS	Swiss Interdisciplinary Commitee for Biosafety in Research and Technology. Zürich / Switzerland
SIC	Stirrer speed indicator and control
SIP	Sterilisation in place
SP	Set point
STR	Stirred tank reactor
TIC	Temperature indicator and controller
TIS	Temperature indicator and switch
TOC	Total organic carbon
TS	Temperature switch
VERO	African green monkey kidney
vvm	Gas volume per liquid reactor volume per minute
WFI	Water for injection

Symbols

A	area, m^2
a	interfacial area per unit volume, m^{-1}, number of moles H atoms in C mole biomass, mol/C-mol
b	number of moles O atoms in C mole biomass, mol/C-mol
c	number of moles N atoms in C mole biomass, mol/C-mol
$c^*_{o_2}$	O_2 concentration in equilibrium with the gas phase, mol/m^3
$\bar{c}_{o_2}$	mean O_2 concentration in the broth, mol/m^3
d_e	stable bubble diameter in bubble column, mm
d_i	diameter of impeller, m
D	dilution rate, h^{-1}
D_d	diameter of draft tube, m
D_i	diffusivity of compound i in the liquid, cm^2/s
D_j	diffusivity of compound j in the liquid, cm^2/s
D_s	diameter of the settler, m
D_t	diameter of the tank, m
D^z_{temp}	decimal reduction time at "temp" for z, min
F	volumetric liquid flow rate, m^3/h
F_b	bleed, volumetric liquid flow rate, m^3/h
F_g	volumetric gas flow rate, m^3/s
F_p	volumetric product flow rate, m^3/h
F^z_{temp}	lethal time at given temperature, min
H	total height of the tank, m
H_b	length of the baffle, m
H_l	liquid height below lower impeller, m
H_L	total liquid height in tank, m
H_i	liquid height in between two impellers, m
H_t	height of liquid above top impeller, m
I	volume of inoculum per working volume, %
k	specific death rate of microorganisms, min^{-1}
k_L	liquid side mass transfer coefficient, m/h
k_La	volumetric liquid side mass transfer coefficient, h^{-1}
K_s	Michaelis-Menten constant of substrate, g/L
K_p	inhibition constant of a product, g/L
L_i	blade length of the turbine impeller, m
l	number of moles H atoms in C mole substrate, mol/C-mol
m	number of moles O atoms in C mole substrate, mol/C-mol, multiplying factor for k_La in pure water
m_s	maintenance factor for substrate, g substrate $\cdot$ (g biomass)$^{-1}$ $\cdot$ h^{-1}
m_{O_2}	maintenance factor for oxygen, g oxygen $\cdot$ (g biomass)$^{-1}$ $\cdot$ h^{-1}
M_{bio}	molar mass of biomass, g/C-mol
n	agitation speed, s^{-1}
N	number of viable microorganisms, CFU
N_o	initial number of viable microorganisms, CFU
N_s	number of viable microorganism after sterilization, CFU
OTR	oxygen transfer rate, mmol $\cdot$ L^{-1} $\cdot$ h^{-1}
OUR	oxygen uptake rate, mmol $\cdot$ L^{-1} $\cdot$ h^{-1}
OUR_{max}	maximal oxygen uptake rate, mmol $\cdot$ L^{-1} $\cdot$ h^{-1}

p	product concentration, g/L
P	biomass production, $g \cdot L^{-1} \cdot h^{-1}$
P_o	ungassed power input of one impeller, W
P_g	gassed power input of one impeller, W
Q_g	generated heat due to bioreaction, $kJ \cdot L^{-1} \cdot h^{-1}$
$q_{\underline{sx}}$	rate of organic substrate consumption, g substrate $\cdot$ (g biomass)$^{-1} \cdot h^{-1}$
q_{o_2x}	rate of oxygen consumption, g oxygen $\cdot$ (g biomass)$^{-1} \cdot h^{-1}$
R_a	arithmetical average roughness, µm
R_q	root mean square deviation of profile, µm
R_z	average peak to valley height, µm
R_{max}	maximum peak to valley height, µm
s	substrate concentration, g/L
s_o	initial substrate concentration, g/L
t_D	doubling time, generation time, h
u_s	linear gas velocity, m/s
V_L	liquid reactor volume, m^3
W_b	width of the baffle, m
W_i	blade width of the turbine impeller, m
x	biomass concentration, g/L
Y	sorption number
Y_{os}	oxygen per substrate requirement, g oxygen/g substrate or mL oxygen/g substrate
Y'_{xo_2}	biomass yield coefficient per oxygen, C-mol biomass/mol O_2
$Y_{xo_2,obs}$	observed biomass yield coefficient per oxygen, g biomass/g O_2
$Y_{xo_2,max}$	max. biomass yield coefficient per oxygen, g biomass/g O_2
Y''_{xs}	biomass yield coefficient, C-mol biomass/C-mol substrate
$Y_{xs,obs}$	observed biomass yield coefficient per substrate, g biomass/g substrate
$Y_{xs,max}$	maximal biomass yield coefficient per substrate, g biomass/g substrate
x	biomass concentration, g/L cdw
z	z-value, °C
α	molar flow of substrate, $mol \cdot L^{-1} \cdot h^{-1}$
β	molar flow of ammonia, $mol \cdot L^{-1} \cdot h^{-1}$
γ	molar flow of oxygen, $mol \cdot L^{-1} \cdot h^{-1}$
γ'_s	degree of reduction of substrate (NH_3), –
γ'_x	degree of reduction of biomass (NH_3), –
δ	molar flow of biomass, C-mol $\cdot L^{-1} \cdot h^{-1}$
ε	molar flow of water, $mol \cdot L^{-1} \cdot h^{-1}$
κ	molar flow of carbon-dioxide, $mol \cdot L^{-1} \cdot h^{-1}$
μ	growth rate, h^{-1}
η_{o_2}	oxygen efficiency factor of growth
ρ	density of fluid, kg/m^3
ν	kinematic viscosity, m^2/s

Indices

max	maximum value of the parameter

Dimensionless numbers

Ne	Power or Newton number	$Ne \equiv \frac{P_o}{r \cdot n^3 \cdot d_i^5}$
Re	Reynold Number for stirrer	$Re \equiv \frac{n \cdot d_i^2}{u}$
Q	Flow-rate or aeration number	$Q = \frac{F_g}{n \cdot d_i^3}$
Y	Sorption number	$Y \equiv \frac{k_L a}{u_s} \cdot \left(\frac{u^2}{g}\right)^{1/3}$

1. Introduction

Biochemical reactions involve biocatalysts, such as microorganisms, plant cells, animal cells or enzymes, and result in the transformation and production of biochemical substances. Biochemical engineering covers the design of vessels and apparatus suitable for performing biochemical reactions.

This article will treat the design of apparatus for the cultivation of microorganisms, plant and animal cells. Apparatus for enzymatic reactions, the waste-water treatment and for solid-state fermentation are not discussed.

The article focuses on sterilizable reactors for suspended cultivation. In American English a differentiation is made between bioreactors and fermentors, the first characterizing reactors used to cultivate animal cells and the latter to cultivate bacteria or fungi [1].

Generally when running a bioprocess in a reactor, the periphery and the reactor are first sterilized. Organisms are then inoculated to a liquid medium and start growing and metabolizing substrates. Such processes are widely used in life science industry to produce vaccines, vitamins, starter cultures, monoclonal antibodies,

aroma, enzymes, antibiotics and many others (→ Biotechnology). Reactor volumes from 1 to 500 000 L are treated. Reactors of smaller scale such as petri dishes, roller flasks, and shakers are not discussed.

2. Requirements

Fermentors and bioreactors should provide optimal *growth conditions* for microorganisms to achieve conversion and/or production of biological products. Generally: only one specific form of life has to grow and produce the desired product (axenic culture). To prevent contamination with other microorganisms the *sterility* of the reactor is an indispensable requirement. *Cleaning* is important to make sure that impurities will not affect the bioconversion or spoil the final product. *Homogeneous conditions* with respect to temperature, pH, dissolved oxygen, substrate, and product concentration have to be maintained in the reactor. The bioprocess needs to be *controlled* and all available measurements should be logged to enable quality assurance. *Safety regulations* have to be adhered in order to prevent accidents and the release of toxic products. Documents describing the parts, instruments, control loops, the operation and the maintenance of the equipment, are required.

2.1. Microorganisms: Growth and Bioreaction (→ Biotechnology)

Microorganisms, plant and animal cells are described in [2], [3]. Generally, organisms with larger cell size, such as animal cells compared to bacteria, have a complex demand of nutrients and a lower growth rate. On the other hand their ability to produce complicated proteins is increased [4].

2.1.1. Biomass Concentration – Cell Density

For a better understanding of the bioprocess it is important to measure or calculate the biomass concentration x. Several units and measurement methods are in use.

The recommended unit to employ if a balance of substrates and products (usually in g/L or mol/L) has to be calculated is the cell dry weight (cdw, in g/L). This is the most widely used unit for bacteria and fungi density analysis. The cell wet weight (cww, in g/L) can be measured far more quickly but not as precisely. The cdw corresponds to 20 to 25 % of the cww [5]. The measurement of the optical density (OD) of the culture medium is a fast and relatively easy process used to determine the cell density (the wavelength of the measurement should always be indicated, e.g., 548, 550, 556, 600 nm). Nevertheless, the density values depend very much on the procedure and on the photometer. For a medium containing *E. coli* an OD of 1 (at 548 nm) was found to correspond to 0.25 g/L cdw [5].

The density of cells of larger organisms, such as insect or animal cells, or very low densities of bacteria can be given in number of cells per milliliter. The cells are counted in counting chambers (Neubauer or Thoma chamber) under the microscope. When staining methods are applied, viable cells can be differentiated from dead ones.

Most of these determinations are performed off-line. The optical density can be measured in situ.

2.1.2. Unstructured Model of Growth

Growth is defined as the increase of viable biomass. This increase can be due to an increase in size of the microorganisms or due to proliferation. During the exponential growth phase, cells divide according to a defined doubling time (generation time) t_{D} [6]. The growth during this phase follows the equation:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mu \cdot x \tag{2.1}$$

with μ, the growth rate, being

$$\mu = \frac{\ln 2}{t_{\mathrm{D}}} \tag{2.2}$$

The equation of biomass growth and the following equations in this chapter show that high bioreaction rates will be generated at high cell density with high growth rate. The consideration of high bioreaction rates allows the calculation of the maximal required mass flow rates of substrates, products and the generated heat of exothermic bioreactions. The considered "high" bioreaction rate has a direct impact on the design of the fermentor.

2.1.2.1. Monod Model

The Monod model allows the mathematical correlation of substrate concentration with growth kinetics [7]. The model is based on enzymatic Michaelis–Menten kinetics:

$$\mu(s) = \mu_{\max} \cdot \frac{s}{K_{\mathrm{s}} + s} \tag{2.3}$$

where $\mu_{\max}$ is the maximum specific growth rate of the organisms, s the substrate concentration, and K_{s} the Michaelis-Menten constant. K_{s} varies with the type of substrate and the diameter of the organism. Examples for $\mu_{\max}$ and K_{s} are given in Table 1. The values correspond to the efficiency of microorganisms in substrate uptake. *E. coli*, for example, reaches 95 % of its maximal growth rate with a glucose concentration of only 0.076 g/L. With a dissolved oxygen concentration of only 0.22 % of air saturation (air saturation corresponds to a dissolved oxygen concentration of ≈ 6 mg O_2/L) 95 % of maximal growth is reached. In practice when cultivating freely suspended organisms the dissolved oxygen concentration is kept above 5 to 20 % air saturation.

The examples demonstrate that when homogeneous conditions are achieved in a reactor and the organisms are freely suspended in the broth, only relatively low substrate concentrations are required to maintain maximal growth.

2.1.2.2. Inhibition Kinetics

At high concentrations, substrates may inhibit growth. The extended Monod model includes the growth inhibition by substrates or products:

$$\mu = \mu_{\max} \cdot \frac{s}{K_{\mathrm{s}} + s} \cdot \frac{K_{\mathrm{p}}}{K_{\mathrm{p}} + s} \tag{2.4}$$

where K_{p} is the inhibition constant. For *E. coli* growing on glycerol the constants were found to be $K_s = 0.68$ g/L, and $K_p = 87.4$ g/L [10]. In the case of acids usually only the undissociated part will inhibit growth, so the pH of the medium has to be taken into account.

2.1.2.3. Nutrient Uptake Rate

Biomass increase requires a certain amount of substrate, expressed in the observed biomass yield coefficient per substrate $Y_{\mathrm{xs,obs}}$:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = \frac{1}{Y_{\mathrm{xs,obs}}} \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1}{Y_{\mathrm{xs,obs}}} \cdot \mu \cdot x \tag{2.5}$$

Besides substrate uptake for growth, organisms require a certain amount of energy for their maintenance, m_s. This energy is needed to maintain the correct ionic composition and turnover of cellular constituents and for several other vital processes. An equation relates growth yields, maintenance factor, and specific growth rate to the rate of substrate or oxygen consumption per biomass ($q_{\mathrm{s}x}$, $q_{\mathrm{o}_2 x}$) [11]:

$$\frac{\mathrm{d}s}{\mathrm{d}t} = q_{\mathrm{s}x} \cdot x = \left(m_{\mathrm{s}} + \frac{\mu}{Y_{\mathrm{xs,max}}} \right) \cdot x = \frac{\mu}{Y_{\mathrm{xs,obs}}} \cdot x \tag{2.6}$$

$$\frac{\mathrm{d}o_2}{\mathrm{d}t} = q_{\mathrm{o}_2 x} \cdot x = \left(m_{o_2} + \frac{\mu}{Y_{xo_2,\max}} \right) \cdot x = \frac{\mu}{Y_{xo_2,\mathrm{obs}}} \cdot x \tag{2.7}$$

The equations show that cell maintenance requires a constant amount of substrate and oxygen, independent of the growth rate. Therefore at low growth rates the observed biomass yield coefficients on substrate will be much smaller than at a high growth rate [11].

$$Y_{\mathrm{xs,obs}} = \frac{\mu}{q_{\mathrm{sx}}} \tag{2.8}$$

$$\frac{1}{Y_{\mathrm{xs,obs}}} = \frac{1}{Y_{\mathrm{xs,max}}} + \frac{m_{\mathrm{s}}}{\mu} \tag{2.9}$$

Several examples of maintenance and yield factors are given in Table 2.

2.1.3. Aerobic Bioreactions

The majority of organisms are chemoheterotrophs, requiring an organic carbon source and a chemical energy source. The stoichiometry of a bioreaction can be used to evaluate the mass-flow rates of nutrients, the demand of oxygen, the accumulation of products and the heat generated during growth. The model also indicates how many mass flows need to be measured to estimate the remaining mass flows [12].

Table 1. Maximum growth rates $\mu_{\max}$ and Michaelis–Menten constant values K_s for selected substrates and microorganisms (data from [8])

Microorganism	Substrate	$\mu_{\max}$, h^{-1}	K_s, g/L
Escherichia coli	glucose	0.8 – 1.4	$2-4 \times 10^{-3}$
Escherichia coli	glycerol	0.87	2×10^{-3}
Escherichia coli	lactose	0.8	20×10^{-3}
Escherichia coli	oxygen		7.1×10^{-7} *
Saccharomyces cerevisiae	glucose	0.5 – 0.6	25×10^{-3}
Baker's yeast	oxygen		2.06×10^{-5} *
Candida tropicalis	glucose	0.5	$25-75 \times 10^{-3}$
Klebsiella aerogans	glycerol	0.85	9×10^{-3}
Aerobacter aerogenes	glucose	1.22	$1-10 \times 10^{-3}$

* Data from [9].

Table 2. Maintenance m and true yield coefficients Y on the basis of glucose and oxygen for selected microorganisms (from [11])

Microorganism	m_s, $g \cdot g^{-1} \cdot h^{-1}$	$Y_{xs,\max}$, g/g	m_o, $g \cdot g^{-1} \cdot h^{-1}$	$Y_{xo,\max}$, g/g
Escherichia coli	0.072	0.35	0.6	24.7
Escherichia coli	0.090	0.53	3.0	42
Saccharomyces cerevisiae	0.018	0.51	0.6	34.5
Aerobacter aerogenes	0.054	0.380	1.4	44

2.1.3.1. Stoichiometry of an Aerobic Growth without Product Formation

Based on the elemental composition of the cell and assuming constant cell composition and growth only on ammonia, oxygen, and a carbon source to form biomass, water, and carbon dioxide, the equation of the bioreaction can be written:

$$\alpha \cdot CH_lO_m + \beta \cdot NH_3 + \gamma \cdot O_2 \longrightarrow \delta \cdot CH_aO_bN_c + \varepsilon \cdot H_2O + \kappa \cdot CO_2 \tag{2.10}$$

with: $CH_aO_bN_c$ elemental composition of the cell (C mole of biomass) CH_lO_m elemental composition of the carbon source (C mole of substrate) [13], [8] The elemental composition of bacteria and yeast cells does not vary strongly. An average composition was estimated to be $CH_{1.64}O_{0.52}N_{0.16}$ [8]. Based on a general ash content of biomass of 5 %, this will result in a molar mass on C basis of $M_{\text{bio}} \approx 25.5$ g/C mol. The elemental composition of several carbon sources is given in Table 3.

Six unknown flows of molecules α, β, γ, δ, ε, κ (mol L^{-1} h^{-1}) have to be determined. According to the atomic balance four equations can be formulated. Following rules of linear algebra two flows are therefore free. When assuming or measuring two flows the other can be estimated [12].

In the following, the flows of biomass (δ) and substrate (α) are considered to be known. A macroscopic yield factor Y''_{xs} for the number of C moles of biomass produced per C mole of substrate consumed is considered:

$$Y''_{xs} = \left|\frac{\delta}{\alpha}\right| \tag{2.11}$$

The other mass flow rates are calculated accordingly: Substrate flow

$$\alpha = \frac{\delta}{Y''_{xs}} \tag{2.12}$$

Ammonia flow β

$$\beta = c \cdot \delta \tag{2.13}$$

Oxygen flow γ

$$\gamma = \frac{1}{4} \cdot \left(\frac{\gamma'_s}{Y''_{xs}} - \gamma'_x\right) \cdot \delta \tag{2.14}$$

Biomass flow

$$\delta = \delta \tag{2.15}$$

Water flow ε

$$\varepsilon = \frac{1}{2} \cdot \left(\frac{l}{Y''_{xs}} - a + 3c\right) \cdot \delta \tag{2.16}$$

CO_2 flow κ

$$\kappa = \left(\frac{1}{Y''_{xs}} - 1\right) \cdot \delta \tag{2.17}$$

where c is the number of moles of N atoms in C mole biomass, and a the number of moles of

Table 3. Elemental composition of averaged biomass and selected substrates (from [8])

Microorganism substrate	Molecular formula	Molar mass, g/mol	Molar mass on C mole basis, g/mol	Degree of reduction γ' (NH_3)
Biomass	$CH_{1.64}O_{0.52}N_{0.16}$		25.5**	4.17 ($s = \pm 3$ %)*
Methane	CH_4	16	16	8
Pentadecane	$C_{15}H_{32}$	212	14.1	6.13
Methanol	CH_4O	32	32	6
Ethanol	C_2H_6O	46	23	6
Glycerol	$C_3H_8O_3$	92	30.7	4.67
Mannitol	$C_6H_{14}O_6$	182	30.33	4.33
Acetic acid	$C_2H_4O_2$	60	30	4
Lactic acid	$C_3H_6O_3$	90	30	4
Glucose	$C_6H_{12}O_6$	180	30	4
Formaldehyde	CH_2O	30	30	4
Succinic acid	$C_4H_6O_4$	118	29.5	3.50
Citric acid	$C_6H_8O_7$	192	32	3.0
Formic acid	CH_2O_2	46	46.0	2.0
Oxalic acid	$C_2H_2O_4$	90	45.0	1.0

* s = standard error expressed in %.
** 5 % ash is considered in the calculation of the molar mass.

H atoms in C mole biomass. To simplify Equation (2.14), the so-called generalized degree of reduction of biomass γ'_x and of substrate γ'_s are introduced [14]:

$$\gamma'_x = 4 + a - 2b - 3c \tag{2.18}$$

$$\gamma'_s = 4 + l - 2m \tag{2.19}$$

where b is the number of moles of O atoms in C mole biomass, l the number of moles of H atoms in C mole substrate, and m the number of moles of O atoms in C mole substrate. The rationale behind the degree of reduction is the concept of the number of electrons available for transfer to oxygen during combustion of a compound. However, this concept must be used carefully, since the degree of reduction depends on the nitrogen source introduced in the bioreaction. The equations for the degree of reduction of biomass and substrate are in fact linear combinations of the equations of the atomic balances. Therefore these equations can be seen as a simplification but do not give any further information about the bioreaction. An excellent review of the calculation, considering also products or other nitrogen sources in the bioreaction, is given in [13], [15].

2.1.3.2. Oxygen Uptake Rate (*OUR*)

The theoretical oxygen uptake rate *OUR* can be calculated using Equation (2.14) [13], [15]. The yield of C mole biomass per consumed mole oxygen, Y'_{xo_2}, can be calculated by introducing the oxygen efficiency factor of growth, η_{o_2}:

$$\eta_{o_2} = \frac{Y''_{xs} \cdot \gamma'_x}{\gamma'_s} \tag{2.20}$$

$$Y'_{xo_2} = \frac{\delta}{\gamma} = \frac{4}{\gamma'_x} \cdot \left(\frac{\eta_{o_2}}{1 - \eta_{o_2}} \right) \tag{2.21}$$

Since the degree of reduction of biomass γ'_x can be assumed to be constant, the yield of biomass growth per mole oxygen will only depend on η_{o_2}. Table 4 shows the yield and efficiency factor of several microorganisms growing on minimal media. With the indication given in Tables 1, 2, 3, and 4 the maximal oxygen requirement OUR_{max} at maximal growth rate and cell density can be calculated:

$$OUR_{max} = \frac{\mu_{\max} \cdot x_{\max}}{Y'_{xo_2} \cdot m_{\text{bio}}} \tag{2.22}$$

The *OUR* of mammalian cells can be estimated by indicating the oxygen demand of a single cell per hour. Respiration parameter in the range of 2 to 8×10^{-9} mg $O_2 \cdot \text{cell}^{-1} \cdot \text{h}^{-1}$ are reported [16]. Typical cell densities are in the range of $1-3 \times 10^6$ cells/mL at the end of a batch cultivation and $0.8-2 \times 10^7$ cells/mL for continuous cultivation with cell recycle.

2.1.3.3. Generated Heat

Cooney et al. measured that during aerobic metabolism the rate of heat production Q_g correlates with the rate of oxygen consumption

Table 4. Selected microorganisms with indication of the substrate (minimal medium) and yield coefficients [11], [8]

Microorganism	Carbon source	Y_{xs}, g/g	Calculated* Y''_{xs}, C-mol/C-mol	Calculated* η_{o_2}	Calculated* Y'_{xo_2}, C-mol/mol
Candida utilis	glucose	0.504	0.593	0.618	1.553
Candida utilis	glucose	0.515	0.606	0.632	1.645
Candida utilis	ethanol	0.687	0.620	0.431	0.726
Saccharomyces cerevisiae	glucose	0.556	0.654	0.682	2.056
Penicillium chrysogenum	glucose	0.483	0.568	0.592	1.394
Escherichia coli	glucose	0.377	0.444	0.462	0.825
Escherichia coli	glucose	0.522	0.614	0.640	1.707
Aerobacter aerogenes	glucose	0.404	0.475	0.495	0.942
Aerobacter aerogenes	mannitol	0.524	0.623	0.600	1.437
Bacillus subtilis	glucose	0.4517	0.531	0.554	1.191

* For calculation the values of Table 3 were used.

(*OUR*) [17]. LUONG and VOLESKY found the correlation factor to be [18]:

$$Q_{\mathrm{g}} = (0.465 \pm 0.084) \cdot OUR \tag{2.23}$$

This correlation is predicted by the theory of the degree of reduction and the thermodynamic treatment of the energetics of growth (see Section 2.1.3.1). The correlation is independent of growth rate, substrate, or microorganism. The theoretical correlation factor was calculated to be 0.460 kJ/mmol [13], [14], [19]. The theoretical value is very consistent with the measurement performed by LUONG and VOLESKY. The measurement from COONEY et al. is slightly higher (0.519 ± 0.013 kJ/mmol).

2.2. Sterility

Sterility is defined as a state being free from microorganisms or spores which could reproduce themselves (CEN-prEN 13092, 1998). This absolute statement regarding the absence of viable microorganisms cannot be proven. It can only be said that no organisms were present that were capable of growing under the specific test conditions [20]. In practice processes will be called sterile, when established and recognized methods of sterilization are used.

The process of inactivation of viable microorganisms during a sterilization procedure is usually described by an exponential function. Mathematically the results of such functions can be reduced to very low fractions, but never to zero. A degree of inactivation or a certain probability of contamination can be calculated. Different administrations [21] accept a probability of contamination for samples (infusion flasks, ampoules) of 10^{-6}. It means that one out of 1 million samples might be contaminated.

The sterilization procedure used for the bioprocess should guarantee a contamination probability of $< 10^{-3}$. A sterile system inoculated with only one desired form of life, or a pure culture, is called "axenic".

2.2.1. Thermal Microbial Death Rates

When organisms (microorganisms, viruses, plant or animal cells) are exposed to dry heat or saturated pressurized steam their death rate is proportional to their number:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = -k \cdot N \tag{2.24}$$

$$\ln\left(\frac{N(t)}{N_{\mathrm{o}}}\right) = -k \cdot t \tag{2.25}$$

The proportionality factor k depends on the temperature and the type of microorganisms.

For more convenience the decimal reduction time D_{temp} was defined. It is the time required to reduce the amount of microorganisms by 90 % of the initial value (for a given temperature expressed in °C).

$$\ln\left(\frac{1}{10}\right) = -k \cdot D_{\mathrm{temp}} \tag{2.26}$$

$$D_{\mathrm{temp}} = \frac{2.303}{k} \tag{2.27}$$

The relation between the decimal reduction time and the temperature is represented in the thermal killing time curve. The microorganism specific z

value indicates the temperature increase (in °C) required to decrease the D value by a factor of 10.

$$\log\left(\frac{D^z_{T1}}{D^z_{T2}}\right) = \frac{T_2 - T_1}{z} \tag{2.28}$$

The F^z_{temp} value is the lethal time in minutes of a sterilization process. This calculation assumes a specific organism (therefore indicating z), an initial number of these organisms, N_o, and a final number of organisms after sterilization, N_s.

$$F^z_{\text{temp}} = \log\left(\frac{N_\text{o}}{N_\text{s}}\right) \cdot D^z_{\text{temp}} \tag{2.29}$$

The advantage of the calculation is the quantification of the lethal effect in terms of the F^z_{temp}. Based on this theory, equivalent processes at other temperatures and sterilization times can be calculated and compared (Table 5):

$$\frac{F^z_{T_1}}{F^z_{T_2}} = 10^{\left(\frac{T_2 - T_1}{z}\right)} \tag{2.30}$$

Several examples are given in Table 5.

Table 5. Lethal time and decimal reduction time at different temperatures of a steam sterilisation. A sterilisation at 121 °C for 20 min is regarded as safe. A microorganism with $D^6_{121} = 1.5$ min is considered. Values from [22]

Temperature T, °C	Lethal time* for a constant lethal effect F^6_T, min	D^6_T, min
100	63 246	4743
105	9283	696
110	1363	102
115	200	15.0
120	29	2.20
121	20	1.50
125	4.309	0.323
130	0.632	0.047
135	0.093	0.007
140	0.014	0.001

* Required lethal time to achieve the same lethal effect as with $F^6_{121} = 20$ min.

The decimal reduction time depends strongly on the organism and on the sterilization procedure. Dry heat (e.g., *Bacillus subtilis*: $D^{23}_{121} = 30$ min) is much less effective than saturated steam (e.g., Bacillus stearothermophilis: $D^6_{121} = 1.5$ min) [22].

When considering an initial concentration of microorganisms (specified with $D^6_{121} = 1.5$ min) of $N_0/V_\text{L} = 10^3$ CFU/mL, a working volume $V_\text{L} = 100\,\text{m}^3$ and a final number of organisms of $N_\text{s} = 10^{-3}$ the required lethal time F^6_{121} can be calculated to be 21 min. The probability of contamination is 1 to 1000.

However the difficulty of sterilizing a bioreactor is not the determination of the required lethal time for sterilization or the determination of the initial cell concentration, but to make sure that saturated steam and temperature prevail in all dead-ends, crevices and lowest points of the reactor.

2.2.2. Steam Sterilization Processes

When performing steam sterilization, it is most important to evacuate all air and other gases from the system to be sterilized. First because the lethal effect is dramatically reduced by air in the system and second because when controlling the temperature of a system with an air-steam mixture a higher pressure will be obtained compared to a system filled with saturated steam only. A system with 50 % air will have an absolute pressure of 3 bar at 121 °C while a system with saturated steam has 2 bar at 121 °C.

2.2.2.1. Sterilization in the Autoclave

Small bioreactors made of glass (3 to 15 L) are often sterilized in an autoclave. Stainless steel vessels of several hundred liters on wheels are seldom autoclaved.

The advantage of an autoclavable system is that the reactor and all connections to transfer lines are sterilized together. The complete system is therefore relatively simple with a small number of valves and clamps.

The disadvantages are multiple:

- The transfer in between the production place and the autoclave is always accompanied with the risk to damage the fragile glass vessel.
- It is not always possible to guarantee a complete removal of air out of the tubes, lines, and headspace of the reactor. To overcome this disadvantage some autoclaves have the possibility to apply a vacuum to the system prior to steam sterilize it.
- Due to lack of mixing of the media, the heat-up times in the autoclave are long and true temperatures difficult to check. The inhomogeneous condition lead to long sterilization times, which might harm the media.

2.2.2.2. "Full" Sterilization In Place (SIP)

A "full" SIP procedure heats up the liquid in the reactor (media, parts of the media, or water) to generate the steam necessary for steam sterilization. Heating jackets, coils, electrical heat fingers or direct steam injection can be used to heat up the liquid. The temperature of the liquid is controlled and homogeneous conditions achieved by mixing. Air or other gases must be evacuated from the reactor. When leaving the upper point of a reactor open, the generated steam will expel the air at around 100 °C. After this step the reactor is closed to pressurize it.

"Full" SIP procedures are fast, do not require an autoclave, and enable good quality assurance compared to sterilization performed in an autoclave. These advantages are paid with increased complexity. Transfer lines and filters require several valves and need to be sterilized separately in most cases (see Section 5.3).

2.2.2.3. "Empty" Sterilization In Place (SIP)

Empty SIP is performed by injecting clean steam in the empty vessel. Heat-up time might be supported by heating the liquid in the temperature jacket (or coils) or by draining the liquid in the jacket prior to starting the sterilization. Air in the vessel has to be evacuated through the top and/or bottom, while condensate is drained at the lowest point of the vessel. The main steps are the same as for a "full" SIP. When cooling down, high flow rates of sterile air are required to break the vacuum at around 100 °C.

Empty SIP is preferred when heat sensitive media are used (cell culture) or to reduce the heat and cooling times (especially large vessels 10 to 500 m^3). Media require a separate sterilization procedure before being added to the sterile reactor. A disadvantage is that pO_2 and pH probes suffer more during an empty SIP compared to a full SIP.

2.2.2.4. Continuous Water/Steam Sterilization [23], [24]

Instead of being sterilized batchwise in the reactor or the media tank, medium can be sterilized continuously. Generally shorter holding times (≈ 90 s) at higher temperatures (> 130 °C) are used (Table 5).

The process requires a pump to pressurize the medium; a heat exchanger to heat up the medium to the sterilization temperature; a holding section; a heat exchanger to cool the medium; a pressure relief valve (Fig. 9). Before starting the sterilization procedure, the equipment itself must be sterilized.

The liquid flow of an ideal continuous sterilization process has a plug-flow characteristic. Plug-flow or low axial dispersion is achieved with a turbulent liquid flow ($Re > 2 \times 10^4$, high Péclet number: $Pe > 1000$) [23], [24].

2.2.3. Sterile Filtration (→ Filtration)

During cultivation several flows enter and escape the reactor. The entering flows must be sterilized to avoid contamination of the reactor, while the escaping flows should be sterilized in order not to contaminate the environment. This is done by filtration. Liquids and gases can be filter sterilized to withdraw microorganisms and viruses (see Section 5.8).

A filter is characterized by the removal of particles of a stated size at a stated efficiency. Such ratings can be absolute removal (100 % for membrane filter), or percent removal of a specific particle size fraction (depth filter). The absolute filter rating is the diameter of the largest hard spherical particle that passes through the filter under specified test conditions (e.g., flow rate, temperature, pressure drop).

A filter or filter cartridge is usually located in a filter housing which requires to be sterilized with all involved tubing before operation. This is usually performed by steam sterilization (Fig. 11).

Absolute filters (membrane filters) have in comparison to depth filters a low capacity of retaining particles before clogging. To take advantage of both types of filters first a depth filter immediately followed by an absolute filter can be installed. While the depth filter retains most of the particles without clogging, the absolute filter ensures that all particles are retained (Fig. 8).

2.2.4. Incinerators

In bioreactor systems incinerators are applied for sterilization of exhaust gases. Processes involving pathogenic microorganisms might use

such a procedure as a second barrier in the exhaust-gas line (Fig. 1). An incinerator operates continuously between 350 to 450 °C (CEN, prEN13095, 1998). The exhaust gas flows either directly through a flame or is heated by electrical heaters.

2.3. Cleaning

After running a process, the reactor must be cleaned. Deposits of sugar, fats, proteins and salts must be removed to avoid contamination and pollution of the following process. Residues of fungi, baking media in the upper region of the heating jacket (see Section 5.1.1.5), and the headspace of a tank or the exhaust gas line (foam) are among the most difficult parts to clean. The cleaning process is influenced by four parameters:

- Physical shear stress (e.g., mixing, fluid velocity)
- Temperature
- Chemical influence, efficiency of the chemical agent
- Time required for cleaning

Generally smooth surfaces without crevices, edges, and fissures are easier to clean. Validated cleaning processes require the measurement of selected "key" components occurring in the biological and in the cleaning process. The measured values must be below a critical value, which is considered tolerable. Extremely accurate measurements and detectors are therefore required (e.g., determination of total organic carbon (TOC), DNA test, enzyme-lined immunosorbent assay ELISA test, Lowry's nonspecific protein detection). However the definition of "clean", the required tolerances and the procedures often lead to controversial discussions.

2.3.1. Manual Cleaning

Manual cleaning is performed in small and in very large reactors.

Often the dirty reactor is first rinsed with normal tap water. The vessel is afterwards filled with cleaning solution, the solution in the reactor powerfully mixed and temperature regulated. Different cleaning steps can follow each other (DIN 11483):

- 0.1 – 1 M NaOH 60 to 140 °C for several hours
- 0.1 – 0.5 M H_3PO_4 or HNO_3 up to 90 °C for less than 1 h
- Protease solutions for several hours

At the end and in between the steps the vessel is rinsed with tap water, demineralized water, or water for injection (WFI), depending on the required quality of cleaning.

Laboratory- and pilot-scale reactors are partly dismounted for in-depth cleaning. Sensors, baffles, shaft plugs, or other parts in the vessel should be removed to allow separate cleaning. Soft sponges, brushes, or tissues are used to physically remove the dirt. It must be avoided to scratch or damage the surface of stainless steel parts. Thorough procedures imply the removal and greasing of O-rings and O-ring grooves.

Manual cleaning in larger reactors requires a crane to remove the lid or a manway to get into the vessel. The reactor can be cleaned with a high-pressure water injector, either from the top or by a person inside the vessel.

2.3.2. Cleaning in Place (CIP)

Dismantling a reactor for manual cleaning is always time consuming. Cleaning of reactors in place (CIP), especially when being partly or completely automated is fast and more convenient.

In general for CIP the reactor needs to be equipped with CIP supply and return lines and spraying nozzles. Cleaning solutions are prepared in a "CIP kitchen".

A CIP-kitchen allows the preparation (concentration and temperature adjustment) of cleaning solutions (Fig. 2). To adjust the concentration conductivity probes or scales might be used. For temperature adjustment usually an external heat exchanger with a recirculation pump is installed. CIP kitchen with multiple tanks have higher investment costs but require less time for cleaning, since cleaning solutions are stored ready to use in holding tanks.

The reactor to be cleaned in place requires additional valves to supply the cleaning solutions through transfer lines into the vessel. Spraying nozzles must be placed in the top area of the

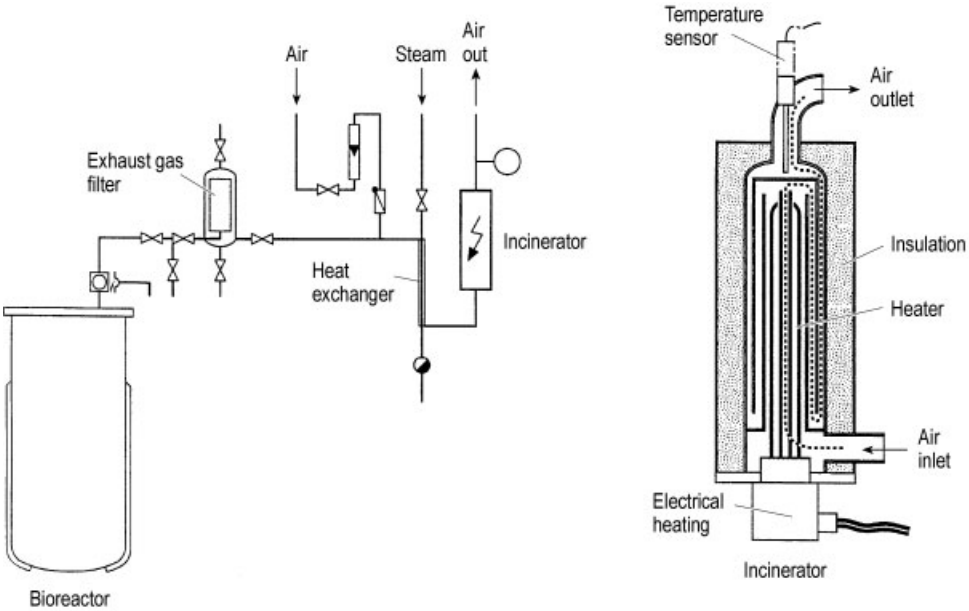

Figure 1. Incinerator: continuous sterilization in the exhaust gas-line

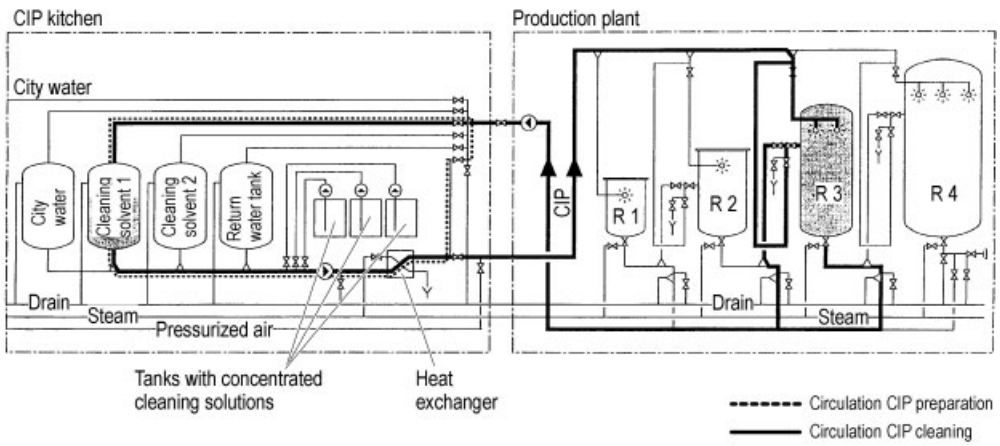

Figure 2. Example of a CIP kitchen preparing and supplying the cleaning solutions to a plant consisting of four vessels (R1 to R4). A preparation and a cleaning step are indicated.

vessel in order to cover all surfaces with cleaning solution. To avoid shadows on parts installed in the vessel, several spray nozzles or special spray valves need to be installed (Fig. 3). To clean pipes, flow velocities of 1.5 to 2.5 m/s are required. Vessels should be covered with at least $30-40\,\mathrm{L}\cdot\mathrm{min}^{-1}\cdot$ (m circumference of vessel)$^{-1}$. To avoid aerosols the pressure drop on the nozzles should be below 2.5 bar.

Typical cleaning blocks consist in: (a) flushing with tap water, (b) flushing or recirculation with cleaning solution 1 (usually an alkali solution), (c) flushing or recirculation with cleaning solution 2 (usually an acid solution), and (d) final

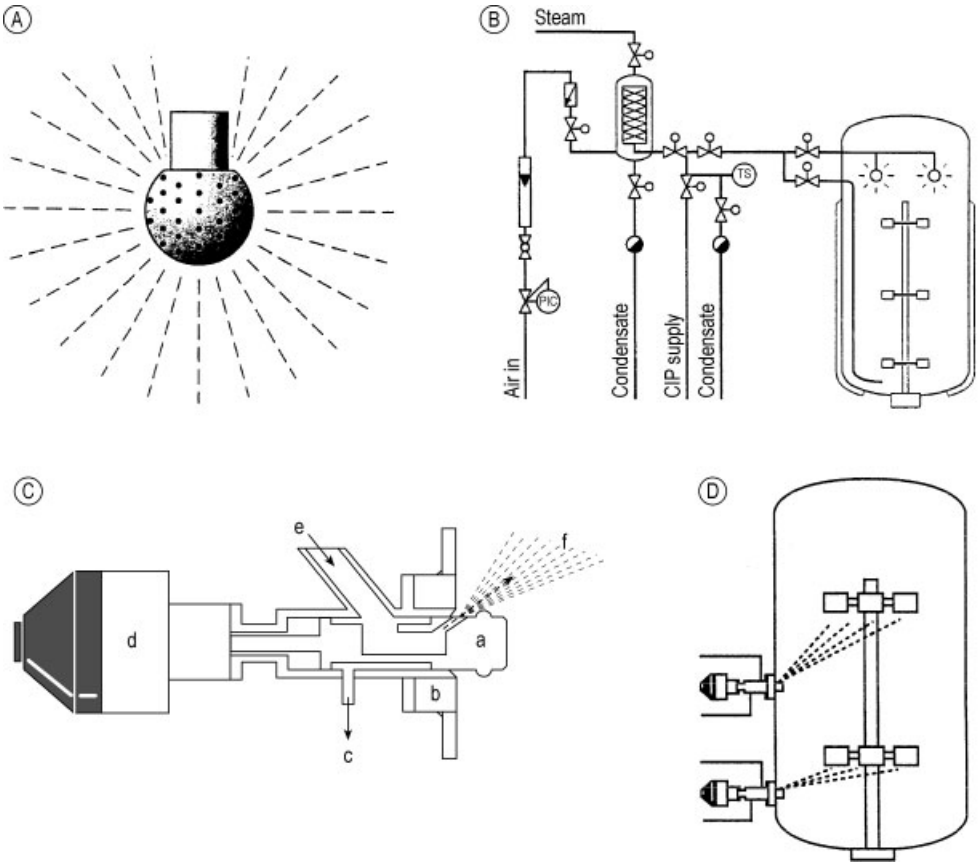

Figure 3. Spray nozzles and application examples to clean a stirred tank [25]
A) spray ball with action radius of 360 ° and flow rates from 1 to 15 m^3/h; B) Stirred vessel requiring a minimum of two spray balls; C) Spray valve; D) Application to sparge directly on critical parts
a) Spray nozzle part; b) Flange on vessel; c) Condensate out; d) Pneumatic actuator; e) Cleaning detergent or steam; f) Cleaning detergent outlet (spray direction)

rinse with WFI or deionized water. In between and at the end of those blocks the equipment is blown-out with air. The quality of the cleaning procedure can be estimated by measuring the conductivity of the rinsing liquid in the return line. When WFI solution is used to rinse the equipment the criteria for efficient rinsing is below 1 to 5 µS/cm. Conductivity probes without temperature compensation should be used for this purpose [26].

Good CIP systems are characterized by efficient cleaning, short cleaning times and low water consumption.

2.4. Homogenous Conditions due to Mixing

Efficient mixing (→ Mixing, Introduction; → Stirring) in fermentors and bioreactors is one of the main objectives to reach:

- Homogenization: equalization of differences in concentration, temperature, and pH
- Intensification of heat transfer between the heating surface and the liquid
- Dispersion of gas in liquid and thereby intensification of the gas-liquid mass transfer
- Suspension of solids such as microcarriers or pellets which must be kept in suspension

Mixing in reactors is limited by:

- The difficulty in removing the dissipated energy generated by mixing

- The intensive foaming of the medium due to mixing and gassing.
- The gas hold-up requiring additional volume in the bioreactor
- Shear sensitivity of the microorganisms

Two main categories of mixing are usually distinguished, the mechanical and the pneumatic mixing. When considering sterile application, about 95 % of the reactors are mechanically and 5 % are pneumatically mixed. Mixing in almost all mechanical mixed reactors is achieved by stirring. The turbulence in the media is generated by rotating impellers and in most cases promoted by static baffles.

2.4.1. Mechanical Mixing

Three categories of mechanical mixing can be distinguished:

- Stirring by a rotating shaft with impellers, (largest category, about 98 % of the cases).
- Vibromixing, based on a translatory vibrating instead of a rotating motion of a shaft.
- Hydraulic mixing, usually using a rotating impeller inducing a liquid flow, which is afterwards dissipated to create turbulence.

2.4.1.1. Stirred Tank Reactors – Energy Dissipation

A broad range of impellers were investigated for stirring (→ Stirring, Chap. 1.3.). Mainly two types are used for reactors: The six blade turbine impeller (also called turbo or Rushton impeller) for aerobic growth of microorganisms requiring high *OTR* and the marine impeller (propeller) for gentle mixing at low shear rate (e.g., animal cell cultures). The Rushton impeller is chosen in the following to calculate the dissipated energy in a stirred tank reactor (STR).

When mixing is investigated the rheology of the fluid has to be considered. Broth of most microorganisms and animal cells have a Newtonian fluid characteristic and keep it over the cultivation time [24].

The general procedure to calculate the power input of stirrers is widely described in literature [27], [24], [8] (→ Stirring, Chap. 2.3.) and consists in calculating the power input of a nongassed system P_o and adjusting the calculation to the gassed system P_g.

To determine power input P_o the dimensionless power or Newton number *Ne* was defined by RUSHTON et al. (1950) [28]:

$$Ne \equiv \frac{P_o}{r \cdot n^3 \cdot d_i^5} \tag{2.32}$$

The "power characteristic of a stirrer" (→ Stirring, Fig. 10, Chap. 2.3.) is given in the graph correlating the dimensionless Reynold number for stirrers to the power number:

$$Re \equiv \frac{n \cdot d_i^2}{v} \tag{2.33}$$

with n (s^{-1}) the rotational speed, d_i (m) the diameter of the impeller, ν the kinematic viscosity (m^2/s) of the broth and ρ (kg/m^3) the density of the broth. Since reliable kinematic viscosities and densities of the cultivation media are lacking, the viscosity and density of water at 30 °C is used for calculation ($\nu = 0.8 \times 10^{-6}$ m^2/s, $\rho \approx 1000$ kg/m^3). The "power characteristic of a stirrer" can be separated into three types of flow regimes: laminar, transient, and turbulent. A Rushton impeller with baffles in the vessel reaches turbulent mixing when $Re > 5 \times 10^3$. Under these conditions the power number is constant ($Ne = 5$, for other flow regimes or stirrers, see → Stirring, Chap. 2.3.2.). Thus the ungassed power input of one impeller P_o under turbulent flow regime is:

$$P_o = 5 \cdot r \cdot n^3 \cdot d_i^5 \tag{2.34}$$

The power input of a gassed impeller P_g is reduced compared to a nongassed system. Consequences can be quite dramatic and even result in a flooding of the impeller. At that stage the impeller is completely surrounded with gas bubbles and its mixing efficiency dramatically reduced.

To estimate the gassed power input again a dimensionless factor, the aeration- or flow-rate number Q is introduced.

$$Q = \frac{F_g}{n \cdot d_i^3} \tag{2.35}$$

with F_g (m^3/s) the volumetric aeration rate. A graph showing the gassed power number *Ne* as a function of the aeration-rate number Q is used for the estimation (→ Stirring, Fig. 17) of P_o and

P_g of single impellers. With aeration rate numbers $Q > 0.05$ the gassed power P_g is reduced by 50 % compared to the ungassed power P_o. The impeller can be considered to be flooded.

If multiple impellers are used and if these are properly spaced (distance in between the impellers 1 to 2 times the diameter of the impeller), P_o can be multiplied by the number of impellers to obtain the total ungassed power absorbed in the broth.

2.4.1.2. Vibromixer

A vibromixer consists of a disk with conical perforation, perpendicularly attached to a shaft moving up and down with a controlled frequency and amplitude. By increasing the amplitude and the frequency of the translatory movement the turbulence in the reactor is increased. The sealing of the shaft is achieved with a membrane or bellows. Vibromixer do not require a dynamic sealing, are a closed system and are therefore considered as very safe (containment). The mixing is efficient and gentle (low shear forces). Vibromixers find application in production of tetanus toxin and animal cell culture.

2.4.1.3. Hydraulic Mixing

In hydraulic mixing turbulence is created with a hydraulic pump. Two-phase nozzle injectors generate small gas bubbles by dissipating the kinetic energy of the liquid [29]. Liquids with low coalescence can efficiently be mixed and aerated with injectors. However aseptic design of pumps in a reactor system is always a problem and probably the reason why until now injectors are mainly found in wastewater treatment.

In fluidized-bed (Figs. 15 and 16) or fixed-bed reactors (Fig. 17), hydraulic pumps might be used to circulate the media. These two processes require an additional gassing system to allow gas-liquid mass transfer.

2.4.2. Pneumatic Mixing

Pneumatic mixing is achieved by sparging gas into liquid. Bubbles rising in the liquid will locally vary the density of the broth and thereby generate turbulences and liquid flow. Bubble sizes will depend on the type of the sparger and the coalescence of the media. No mechanical impeller or other device is used to crush the bubbles.

When bubbles freely rise, the bioreactor is called *bubble column* (→ Bubble Columns). *Airlift reactors* might be built with a concentric draft tube (→ Bubble Columns, Chap. 2.14.). In such units bubbles usually rise outside the draft tube and create thereby a liquid upflow (riser), on the inner side of the tube the liquid flows downwards (downcomer). Downcomer and riser can also be built in separate tubes (*loop reactors*). Draft tubes, riser, and downcomer are guiding the bubbles and thereby the liquid flow. A better control of the flow is thereby achieved, whether this results in improved mass transfer or homogeneity is questionable [16]. Large bubble column bioreactors ($> 100\,m^3$) produce excellent results, when using spargers that efficiently disperse the air bubbles.

Characteristics and main advantages of pneumatically mixed reactors include:

- Simplified design and low maintenance costs compared to mechanically agitated systems (mechanically moving parts have not to be introduced and sealed in the aseptic area)
- Mixing and gassing is performed in a single process
- Gentle mixing at low shear forces When related to the necessary power input, heat and mass transfer are excellent; absolute values will be lower than in the STR (Rushton impeller)
- Airlift or bubble columns are usually long and narrow and therefore require only a small base area

This type of reactors is used for large bulk productions and animal cell cultures. The main parameters influencing the mixing and the gas – liquid mass transfer are the geometry of the gas sparger, the gas flow rate, the length of the reactor, the coalescence of the bubbles, and the rheology of the broth. Higher gas flow rates will increase the turbulence and the mixing. Small bubble size will improve the gas – liquid mass transfer.

2.5. Gas – Liquid Mass Transfer

The oxygen demand of the cells (*OUR*, see Section 2.1.3.2) and the low solubility of oxygen in water, require a continuous transfer of oxygen from the gas phase into the liquid phase. This mass transfer is called oxygen transfer rate (*OTR*).

On the other hand product gases, such as CO_2 or H_2S from aerobic and anaerobic cultivations, might accumulate in the broth and inhibit growth. To prevent inhibition, these gases have to be stripped from the medium. With a high gas – liquid mass transfer, the stripping effect is more effective.

2.5.1. Solubility of Oxygen and Other Gases

The solubility of gases in liquid media is described by Henry's law. These solubilities vary as a function of the temperature and the concentration of salts in the solution (Table 6). The solubility of CO_2 in water is about 30 times higher than that of oxygen.

Table 6. Oxygen solubility $c^*_{o_2}$ ($\mathrm{mol \cdot m^{-3} \cdot bar^{-1}}$) for various temperatures in demineralized water (unless otherwise described)

Temperature T, °C	$c^*_{o_2}$ [$\mathrm{mol \cdot m^{-3} \cdot bar^{-1}}$]
15	1.54
20	1.38
25	1.26
30	1.16
30	0.75 to 0.875 *
35	1.09
37	0.96 **
40	1.03

* In a broth used to cultivate microorganisms. ** In a cell culture medium.

2.5.2. Oxygen Transfer Rate for Stirred Tank Reactors

The *OTR* ($\mathrm{mol \cdot m^{-3} \cdot h^{-1}}$) per unit of reactor volume V_L in an ideally STR is given by:

$$OTR = k_L a \cdot \left(c^*_{o_2} - \bar{c}_{o_2}\right) \tag{2.36}$$

where k_L (m/h) is the liquid mass transfer coefficient, a (m^2/m^3) the specific interfacial area (gas to liquid phase) of mass transfer per liquid reactor volume, $c^*_{o_2}$ (mol/m^3) the hypothetical oxygen concentration in the broth in equilibrium to the partial pressure of oxygen in the gas phase (or saturation concentration) and $\bar{c}_{o_2}$ (mol/m^3) the mean oxygen concentration in the broth. From this equation, reliable guidelines regarding the design of a STR for aerobic processes are extracted. In order to increase the oxygen transfer rate, either the mass transfer coefficient k_L or the specific interfacial area a have to be increased. The interfacial area a increases, when the intensity of the dissipated energy (power input per reactor volume) or when the gas flow rate increases. Another possibility consists in increasing the saturation concentration c^*, by either increasing the percentage of oxygen in the inlet gas or the overall pressure applied on the broth. The relation further indicates, what will happen during an aerobic cultivation. At constant $k_L a$ and $c^*_{o_2}$ and with increasing oxygen demand of the cultivation, the *OTR* is increased by a decrease of the oxygen concentration in the broth. If $\bar{c}_{o_2}$ drops below a critical concentration (Table 1), growth will be oxygen limited.

When considering a biological process with maximal *OUR* (from Eq. 2.22) the maximal required $k_L a$ can be calculated for a maximal allowed driving force ($c^*_{o_2} - \bar{c}_{o_2}$) using Equation (2.36), assuming that the *OUR* equals the *OTR*:

$$k_L a_{max} = \frac{\mu_{max} \cdot x_{max}}{Y'_{xo_2} \cdot m_{bio} \cdot \left(c^*_{o_2} - \bar{c}_{o_2}\right)_{max}} \tag{2.37}$$

When designing a STR the geometry and the rotation speed must meet the maximal required $k_L a$ value. Several correlations exist between the gassed power input, the aeration rate and the mass-transfer coefficient (see Section 2.5.3.2).

2.5.2.1. Liquid Mass-Transfer Coefficient k_L

The reciprocal of k_L is the resistance to gas – liquid mass transfer in bubble aeration (two-film theory, → Transport Phenomena, Chap. 3.3.2.). The value depends on the diffusivity D_i of the gas i in the liquid. For stagnant media the correlation $k_L \sim D_i$ exists. For moving bubbles with mobile surfaces the following correlation was demonstrated: $k_L \sim \sqrt{D_i}$ (penetration theory).

These correlations are useful, when the mass-transfer coefficient of another gas or at another

temperature has to be estimated. The mass-transfer coefficient $k_{L,i}$ of a gas with D_i in a given system can be estimated if the $k_{L,j}$ for a gas with the diffusivity D_j is known:

$$k_{Li} = k_{Lj} \cdot \sqrt{\frac{D_i}{D_j}} \tag{2.38}$$

2.5.2.2. Specific Interfacial Area *a*

The interfacial area a is the surface area A of the gas – liquid interface divided by the reactor working volume V_L. To increase the interfacial area the gas-flow rate or the gas hold-up in the gas-liquid mixture can be increased. Another possibility, for a given gas hold-up, consists in decreasing the bubble size. Many small bubbles will have a larger interfacial area than some few large bubbles. Parameters such as the dissipated energy in the gassed liquid, the geometry of the gas sparger, the coalescence of the bubbles and surfactants in the media therefore influence the interfacial area.

The equations correlating parameters with the mass-transfer coefficients usually are indicated with tolerances of 20 to 40 %. The reason for this large tolerance is found in the difficulty to estimate the coalescence and thus the a of the k_La coefficient. ZLOKARNIK formulated the "multiplying factor" m, relating the k_La of deionized water to solutions of different salt concentrations [29].

$$m = \frac{(k_La)_{\text{solution}}}{(k_La)_{\text{pure water}}} \tag{2.39}$$

Multiplying factors of up to 6 were found for solutions with high salt concentrations. Surfactants used as antifoam agents considerably reduce m. Low viscous Newtonian-type media used to cultivate bacteria or mammalian cell cultures often have multiplying factors in the range of 1.5 to 2.

2.5.3. Aeration of Stirred Tank Reactors

2.5.3.1. Gas Hold-up and Foaming in Stirred Reactors

When aerating a stirred tank reactor the gassed volume will be higher compared to the ungassed working volume V_L. The volume increase per total working volume V_L is the gas hold-up.

The gas hold-up is influenced by the foaming properties or the coalescence of a medium. The gas hold-up of a nonfoaming, bubble coalescing medium will be in the range of 3 to 12 % compared to 5 to 25 % for a foaming medium [30]. Calculations of the gas hold-up are based on the power dissipated per reactor volume P_g/V_L and the superficial gas velocity u_s [8]. Since the gas hold-up is related to the interfacial area a, foaming media have higher *OTR* values than nonfoaming media.

When an antifoam agent is added to a foaming medium, foam will immediately disappear and the gassed liquid level be lowered. At the same time the measurement of oxygen partial pressure in the broth drops. This demonstrates that the driving force $(c^*_{O_2} - \bar{c}_{O_2})$ is increased to compensate the effect of a lower k_La value. To achieve high *OTR* values in foaming media, mechanical foam breaker instead of chemical antifoam agents should be used. Fermentors should be designed to allow high gas holdups (25 %).

2.5.3.2. Volumetric Mass-Transfer Coefficient in Stirred Reactors

VAN'T RIET compared data on mass transfer measurements in stirred tanks with volumes in the range of 0.0025 to 5.1 m^3 and power per reactor volumes P_g/V_L in the range of 0.5 to 10 kW/m^3 [31]. According to these measurements he proposed a correlation for "totally" coalescing media (air sparged in pure water):

$$k_La = 93.6 \cdot \left(\frac{P_g}{V_L}\right)^{0.4} \cdot u_s^{0.5} \pm 20 - 40\% \tag{2.40}$$

where u_s (m/s) corresponds to the linear gas velocity ($u_s = F_g/(\pi \cdot D_t^2)$). For media with salts repressing the coalescence:

$$k_La = 7.2 \cdot \left(\frac{P_g}{V_L}\right)^{0.7} \cdot u_s^{0.2} \pm 20 - 40\% \tag{2.41}$$

These high tolerances and differences in the exponent indicate the difficulty to predict the mass-transfer coefficient.

Generally, when designing a STR the P_g/V_L should not exceed 5 kW/m^3. When using higher

P_g/V_L the mass transfer efficiency (mass-transfer coefficient per power input) is considerably decreased. To further increase the *OTR* not an increase in the k_La value but alteration of other parameters such as higher oxygen partial pressures or reduced usage of antifoam agent (using, e.g., mechanical foam separator instead) should be considered.

The air-flow rate indicated in the correlation as u_s is another important design criterion. For bioprocesses requiring high *OTR*s an often found rule of thumb is the aeration with 1 to 2 vvm [(gas volume) · (liquid reactor volume)$^{-1}$ · (minute)$^{-1}$]. When using this rule of thumb, one should be aware that with higher reactor volumes, u_s will considerably increase [32]. By increasing u_s, the gas hold-up is increased; the gassed power input P_g and the heat transfer are reduced. High u_s might also create excessive foaming.

2.5.4. Volumetric Mass-Transfer Coefficient in Airlift Reactors

In bubble columns due to oxygen transfer, bubbles reduce their oxygen partial pressure when rising to the surface. Smaller bubbles might flow downwards and decrease even more their oxygen partial pressure. This situation complicates the modeling of the system because it is questionable whether really an ideally mixed STR can here be considered. Other models for long and narrow reactors [33] or loop reactors [34] might be more appropriate. However for a first approach the model of the STR (Eq. 2.36) is used here.

For bubble columns the mass-transfer coefficient k_La is proportional to the linear gas velocity (→ Stirring) [33]:

$$k_La \propto u_s \tag{2.42}$$

Combining this proportionality with Equation (2.36) shows that the *OTR* is directly proportional to the:

- Concentration difference ($c^*_{O_2} - \bar{c}_{O_2}$)
- Liquid height in the reactor H_L
- Gas flow rate F_g

This correlation is valid for noncoalescing media and for u_s values below 200 m/h [33]. Maximal gassing rates for standard airlift reactors are indicated with 720 m/h [35].

To calculate the mass transfer coefficient the dimensionless sorption number Y has been introduced:

$$Y \equiv \frac{k_La}{u_s} \cdot \left(\frac{v^2}{g}\right)^{1/3} \tag{2.43}$$

The sorption number is a function of the final stable bubble diameter d_e in noncoalescing systems (Table 7), it decreases with increasing bubble diameter. The kinematic viscosity of water is $\nu = 0.8 \times 10^{-6}$ m^2/s; the acceleration due to gravity $g = 9.81$ m/s^2.

Table 7. Sorption number Y for bubble columns as a function of the final stable gas bubble diameter d_e or for gassing devices. Only noncoalescing media are considered [33]

Stable bubble diameter d_e or gassing device	Sorption number, $10^5 \cdot Y$
0.5 mm	129
0.6 mm	107
0.7 mm	92
1.0 mm	64
2.0 mm	32
Sparger with frit and ($u_s \leq 180$ m/s)	14
Bubble column, sparger with 1 mm Ø holes	9

2.6. Measurement and Control Loops

To maintain optimal growth of biomass, several parameters such as the temperature, pH, pO_2, and a sufficient substrate concentration must be maintained in the reactor. Sterilizable sensors in the reactor can measure on-line.

By taking samples of the broth, off-line measurements can be performed.

2.6.1. Temperature Measurement and Control

The temperature of the liquid is usually measured with a Pt-100 sensor ("Pt 100" because it provides a resistance of 100 Ω at 0 °C). Usually a PID (Proportional Integral Differential) controller activates control elements (valves) for heating or cooling (Fig. 4) An accuracy of ± 0.5 °C should be achieved.

In small reactors heating is achieved by activating an electrical heater immersed in the broth. To cool the reactor coolant is circulated through a coil in contact with the broth.

Larger reactors (10 to 3500 L) usually have a jacket with a temperature circuit. A pump circulates a liquid (usually tap water) through the jacket. A multitude of possibilities exist to heat-up the fluid. For relatively small units (10 to 150 L), and when steam is not available, an electrical heater can be mounted in the loop. Steam can also directly be injected in the loop through a steam injector. A popular solution consists in mounting a heat exchanger in the loop, heated-up with industrial pressurized steam (Fig. 4).

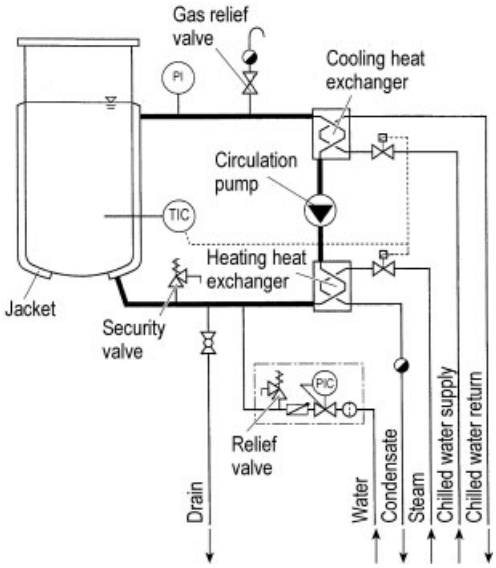

Figure 4. Temperature control of a jacketed bioreactor. The temperating fluid is heated and cooled by two heat exchanger and circulates through the jacket.

To cool down the reactor, tap water is either directly injected into the loop or supplied into a heat exchanger. Heat exchanger for cooling allow the use of closed chilled water systems such as glycerol – water (Fig. 4).

Very large vessels (> 1500 L) might have several parallel temperating loops either flowing through different sections of the jacket or into separate coils.

Temperature measurement might also be used to trace the sterilization procedure at several "critical spots" (Section 3.1.2). Either special fittings, allowing temperature probes to be mounted, are welded into the pipework or the measurement is performed on the outer side of the pipe. With the first method the exact temperature is measured inside the pipe but the added complexity (O-rings, crevices, edges) to the system is questionable. The second method does not change the internal pipework, is a flexible system, but measures the temperature on the outer side of the pipe. Temperature differences between the outer measurement and the inner temperature of less than 3 °C can be reached with a device shown in Figure 5.

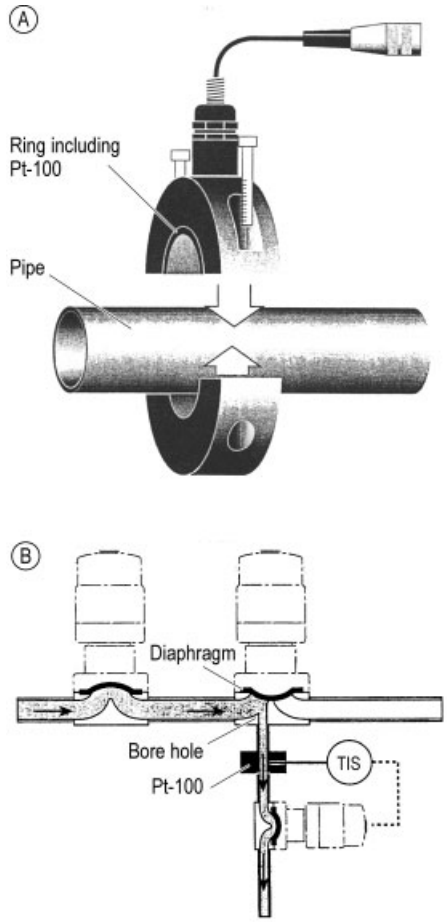

Figure 5. Measurement of the temperature on the outer side of the pipe to control the sterilization process. The temperature difference between the inner and outer side is less than 3 °C (at 121 °C internal temperature). The example shows a measurement on a sterile cross.

2.6.2. Agitation Control

To guarantee homogeneity of the parameters the agitation speed of the stirrer should be continuously controlled. A minimal speed is normally set at about 10 % of the maximal rotation speed of the motor. Gears or pulleys with belts are used

to reduce the rotating speed of the electrical motor. Compared to d.c. motors a.c. motors do not require any maintenance. Usually an accuracy of ± 3 % of the maximum rotation speed of the motor is acceptable.

2.6.3. pH Measurement and Control

For pH measurement and control in-situ sterilizable pH sensors measuring a potential difference (+ 500 to − 500 mV) between the broth and a reference electrolyte of pH 7 are in use. Before installation the sensor has to be calibrated in buffer solutions of known pH. The sensor is in-situ sterilized. Probes with liquid electrolyte (3 M aqueous KCl) require to be pressurized during sterilization to prevent penetration of broth or steam into the electrolyte (through the diaphragm or the membrane). pH electrodes with gel electrolyte need not be pressurized, their electrolyte however can not be exchanged for maintenance purposes. Membrane and diaphragm of the sensor must be cleaned on a regular basis (pepsin solution to clean the diaphragm, the glass membrane must be etched).

A proportional band controller with deadband should be used for pH control.

The output of the controller can for example activate a liquid pump or valve in the transfer line supplying acid (e.g., 1 – 5 M H_2SO_4) or alkali solution (e.g., 1 – 5 M NaOH, 20 % NH_3) to the media. The supply of gaseous ammonia raises the pH and balances the nitrogen demand of a culture (e.g., *E. coli*) without diluting the media.

To reduce the pH of an animal cell culture, CO_2 can be supplied to the inlet gas. Due to its high solubility it is often sufficient to supply the CO_2 to the headspace of the reactor thereby avoiding to produce bubbles (shear stress of "exploding" gas bubbles at the liquid surface might harm the cells [36]). To increase pH, CO_2 has to be stripped out of the media by sparging with air or N_2.

pH should be controlled with an accuracy of ± 0.1. A severe problem might result from drifting of the pH electrode. While media like those used for animal cell culture usually do not tend to induce drifting of probes, others such as those containing sulfur compounds affect the measurement considerably. Depending on the bioprocess it might be useful to recalibrate the pH probe during cultivation. The online measurement can for example be adjusted to the measurement obtained from a sample by a single point calibration. This method is often criticized, since due to gassing out of CO_2 from the sample and changed conditions (pressure, temperature) the measured value might be severely different. When accurate recalibration of the pH sensor is required during cultivation, interchangeable probe devices should be used (Fig. 47).

2.6.4. Measurement and Control of Dissolved Oxygen

Dissolved oxygen is usually measured with an in-situ sterilizable Clark-type electrode immersed in the broth (→ Chemical and Biochemical Sensors, Chap. 2.3.1.). A current proportional to the oxygen partial pressure (pO_2) or the fugacity of oxygen in the media is measured.

The probe should be calibrated in-situ after sterilization and before inoculation. Probes used to monitor aerobic growth are usually one-point calibrated at air saturation. When no current is measured it is assumed that the probe is not exposed to oxygen.

To control the concentration of dissolved oxygen in the broth, parameters influencing the *OTR* must be activated (see Section 2.5):

- Agitation speed of the impeller
- Oxygen or air flow-rate
- Oxygen partial pressure of the inlet gas
- Overall pressure of the tank

One or several of these parameters can be used to control the pO_2. A PI type controller should be used for control. The output signal of this controller can be used to change the set points of one or several of the above parameters. Such a system of two controllers in series is called cascade controller (Fig. 6).

The "pO_2 controlled substrate limited fed-batch strategy", is another method to control the pO_2, which consists in feeding a controlled amount of substrate to the cultivation (see Section 3.4).

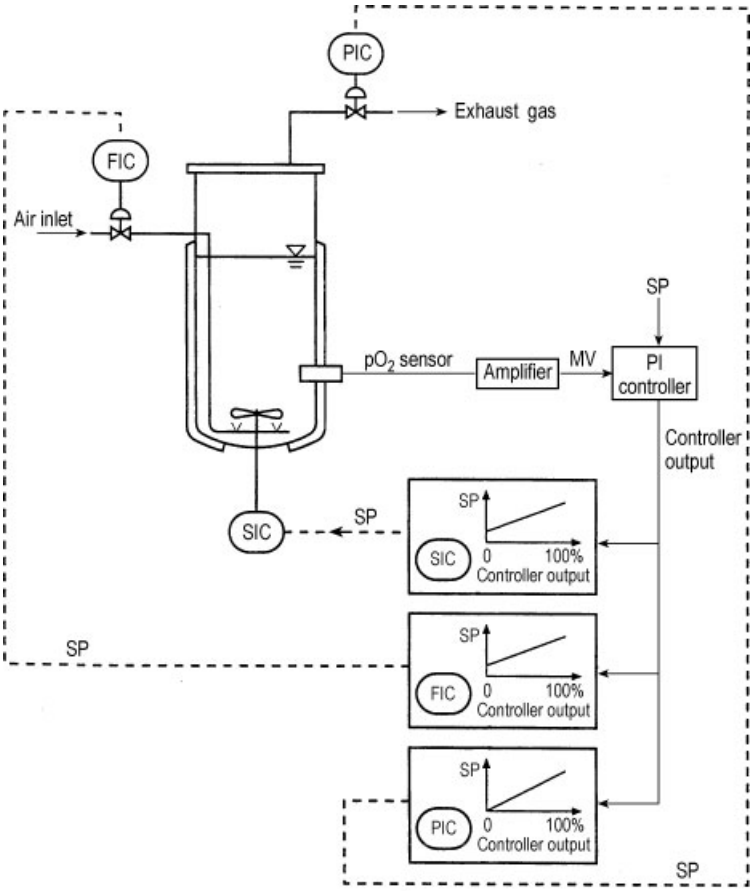

Figure 6. O_2 cascade control loops
SIC = Stirrer indicator and controller; PIC = Pressure indicator and controller; FIC = (Air) flow rate indicator and controller; SP = Set point; MV = Measured value

2.6.5. Measurement and Control of the Volume or the Level

In fed-batch or continuous cultivation the broth volume is an important parameter for the bioprocess. A popular but expensive method to evaluate the volume of the broth is to weigh the entire reactor. Platform scales or load scales have a precision of 0.05 to 0.1 % of the maximal load. The load of the vessel itself must therefore be considered. A vessel of a total volume of 150 L will for example weigh around 500 kg. The maximum range of the scale must therefore be above 650 kg. A platform scale for 1000 kg is for example chosen. A theoretical accuracy of around $\pm$ 1 kg results. In fact the main difficulty is to withdraw all cross influences of transfer lines connected to the vessel. These lines must be flexible and/or bent so that they do not exert a force (or only a stable force) on the vessel to be scaled. Lateral forces on the reactor must also be avoided. The temperature dependent dilatation of the vessel during sterilization must be considered. Because of these influences the load measured at varying temperatures is often affected with an accuracy far above the $\pm$ 1 kg for 150 kg liquid. During operation it must further be considered that any manipulation on the vessel will influence the measurement.

Far less expensive than weighing, are measurement principles, which determine the liquid level in the reactor. To convert the level height into a reactor volume, the cross-sectional area as a function of height should be known or better be constant. If the level is measured with the capacitive measurement principle the gas hold-

up and the foam are parameters influencing the measurement.

Another possibility consists in measuring the pressure difference in between the headspace of the reactor and the bottom, the height of the water column is thereby measured (Fig. 7). An accuracy of ± 6 mm water column can be achieved. For a 150-L tank with standard dimensions ($H : D_t = 1 : 3$) this corresponds to an accuracy of ± 0.8 kg. A possible cross influence is the pressure fluctuation due to turbulences (agitation, aeration) in the broth. The main advantage of this measurement principle is that it is independent of the gas hold-up of the liquid and external forces on the vessel.

To control the liquid volumes, pumps or valves are activated to supply or withdraw the liquids. Media tanks and transfer-lines are required.

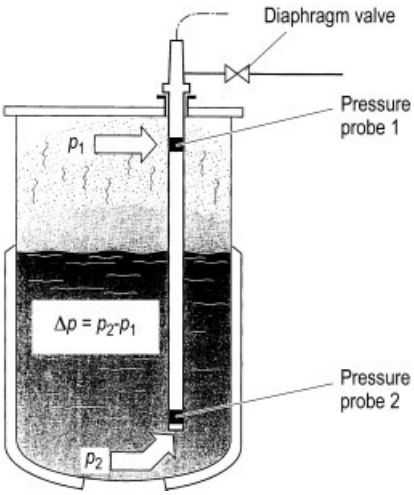

Figure 7. Weighing probe, the measurement of the pressure difference in between the headspace and the reactor bottom allows to quantify the water column and thereby the volume of the bioreactor

2.6.6. Measurement and Control of Gas-Flow Rate

The measurement of gas-flow rates is a nonsterile operation. Sterilization occurs usually downstream of the flow meter.

A cheap and reliable method to measure gas-flow rates consists in using float meters (rotameters). Here, a float hovers in the upward stream of gas passing through a vertical tube that is slightly tapered, the top of the tube being wider than the bottom. The reading is affected by the gas pressure and temperature but can be corrected by calculation. Accuracy up to 2 % can be achieved.

Thermal mass flow meters measure the heat transferred to the gas flow. The heat transfer is directly proportional to the mass flow. The measurement is not affected by the pressure. The measurement is very reliable for clean gases but difficult to service and recalibrate.

To control the gas flow, often a thermal mass flow meter is used, followed by a continuous valve. The combination is a stand-alone unit supplied with a controller. The units often have pressure drops of 1 to 2 bar.

2.6.7. Measurement and Control of Liquid Flows

Liquid flow meters are used in transfer lines and must therefore be cleanable and sterilizable. If a good accuracy is achieved the integration of the measurement is used to determine the transferred volume. To control the flow a throttle valve is activated by a flow controller.

Magnetic or electromagnetic flow meters measure the voltage created by a conducting liquid flowing through a magnetic field. The voltage is directly proportional to the flow rate. Temperature, solids in the liquid, and pressure changes do not affect the measurement. Gas bubbles must be avoided. The measurement requires a conductivity of the liquid above 10 µS/cm, WFI can therefore not be measured.

The Coriolis flow meter consists of a bent tube where the medium flows through. The tube is activated to vibrate at its normal frequency. When measuring the amplitude at two different points of the bent tube, a phase shift can be detected, which is due to Coriolis forces. This phase shift is proportional to the mass flow. External forces or mechanical stress on the bent tube must be avoided. Vibration of the plant might affect the measurement. If these parameters are well controlled, an accuracy of 0.5 % can be expected.

Liquid flow might also be controlled by calibrated metering pumps. Piston or diaphragm pumps generating flows independently of the back pressure are suitable. The system is efficient but has the disadvantage of not measuring the flow (no feed back control).

2.6.8. Other Measurement Systems

The inlet and exhaust gas may be analyzed to calculate the *OTR*. Carbon dioxide is measured using infrared spectrometry. The gas must be dried prior to measurement. Oxygen is measured with a paramagnetic analyzer.

In-situ sterilizable probes measuring on-line parameters such as the concentration of alcohol, the optical density, or the redox potential exist.

The torque of the agitator is measured to indicate the gassed power input to the bioreactor.

The pressure is measured and controlled to increase the *OTR* (see Section 2.5.4).

2.7. Containment

The release of biological material from the bioreactor has to be limited for the following reasons:

- Larger amounts of proteins, which are foreign to the human body, might produce an allergic reaction in humans.
- Microorganisms might be pathogenic to humans, animals or plants due to their toxicity or the toxicity of their metabolic products.

Bacteria, viruses and fungi are classified in four risk classes: harmless, low risk, medium risk, and high risk (e.g., [37]). To each risk class corresponds a safety level with defined precautions [38]. As an example the minimal precautions for the exhaust gas line are listed in Table 8 (see also CEN, prEN 13095, 1998). Internal standards of large companies are often much higher than those required by the official regulations.

The official or the company internal regulations have a strong impact on the design of the bioreactor (e.g., sampling system, transfer line, mechanical seal of the agitator [39]). Generally the regulations differentiate between, laboratory/pilot, and production scale processes [40].

3. Processes and Related Equipment

3.1. Media Preparation and Sterilization

3.1.1. Media Preparation Techniques

The simplest way to prepare media for a cultivation, is to use the reactor directly. It is filled with water (WFI, demineralized, or tap water) and the media components are added in form of powders or concentrated buffer solutions. The components are dissolved or homogeneously dispersed by mixing and heating the media in the reactor. Media sterilization is combined with the "full" SIP of the reactor (see Section 2.2.2.2).

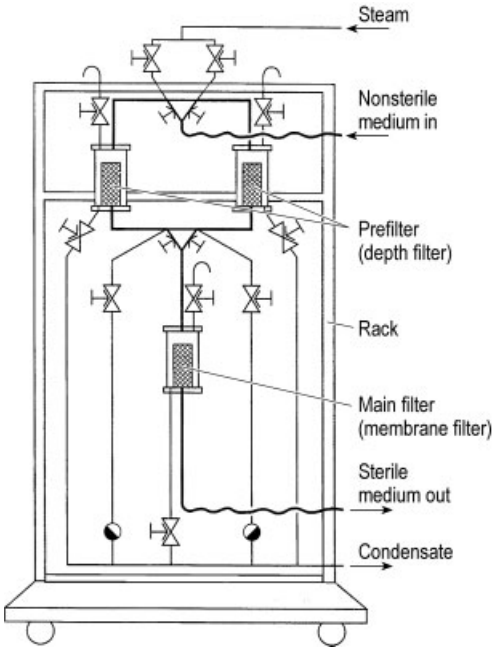

Figure 8. Combination of prefilter (depth filter) and an absolute filter (membrane filter) for the continuous filtration of liquids. Two prefilter are mounted in parallel to allow the exchange, cleaning and sterilization of one of them, while the filtration process is on-going with the other one.

Media components such as vitamins, certain amino acids, or proteins are heat-labile and might be destroyed by steam sterilization (121 °C, 20 min). In order to prevent this, only heat-resistant compounds will be sterilized in the reactor and heat-labile compounds are added afterwards by sterile filtration. It is also reported that a sterilization at higher temperatures for shorter times is less harmful to certain heat-labile compounds, such as thiamin [27].

Media for animal cell cultures are usually prepared separately and added by sterile filtration into the sterile reactor (Fig. 8). Nowadays, completely sterile media are readily available for delivery in bags. The filtration, as well as the bag

Table 8. Risk classes with corresponding safety precautions. Example of minimal required precautions to be taken in the exhaust gas line (pilot and production scale)

Risk class	Safety precautions	Minimal required precautions for the exhaust gas line (pilot and production scale [37])
Harmless microorganisms	GILSP (Good Industrial Large Scale Practices)	no treatment of exhaust gas is required
Low risk microorganisms	containment category 1	the exhaust gas is to be treated to limit the release of microorganisms to a minimum (e.g., mechanical foam separator)
Medium risk microorganisms	containment category 2	the exhaust gas must be sterilized by either passing it through a HEPA filter system or by heat treatment (incinerator)
High risk microorganisms	containment category 3	the exhaust gas must be passed through two HEPA filter systems. The filter must be exchanged under aseptic conditions

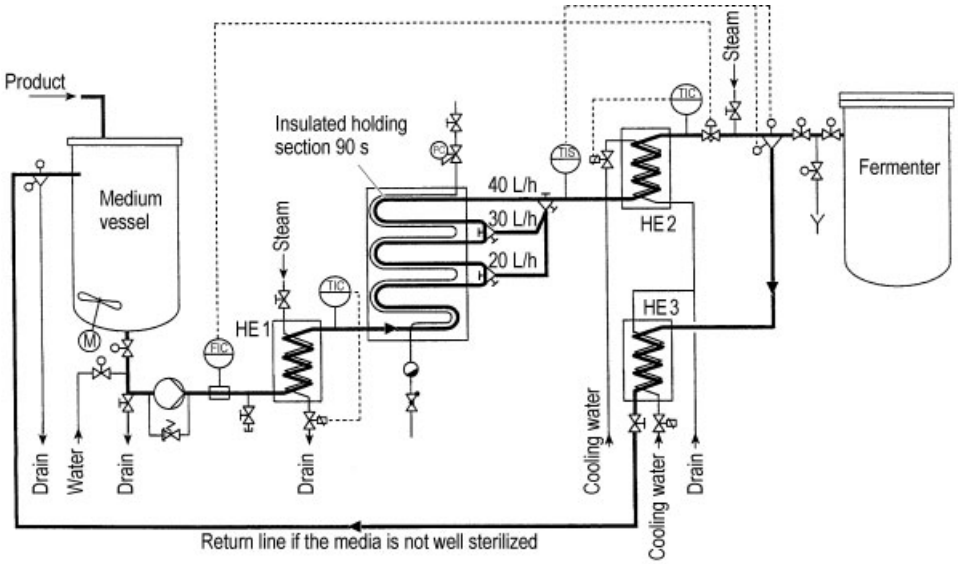

Figure 9. Continuous sterilization with constant holding time and variable length for different flow rates
FIC = Flow indicator and controller; HE1 = Heat exchanger to heat up the medium; TIC = Temperature indicator and controller; TIS = Temperature indicator and switch; HE2 = Heat exchanger to cool down the medium

method require a sterile transfer into the reactor. Large reactors are often sterilized empty to reduce the heat-up time. Media may be sterilized continuously (Fig. 9) to fill the sterile reactor.

Apart from the degradation of heat-labile components, certain reactions between components may contribute to the loss of nutrient quality during steam sterilization. A common phenomenon is the occurrence of Maillard-type browning reactions of carbonyl groups (usually from reducing sugars) with amino acids and proteins [41], [42]. To prevent the degradation a separate sterilization of the carbohydrate components (e.g., glucose, glycerol etc.) from the rest of the media is necessary.

A sterilizable media preparation tank with agitation and temperature control and a sterilizable transfer line is often used to prepare the carbohydrate part of the media (Fig. 10). Such a preparation tank prepares, sterilizes, and stores media used in fed-batch and continuous cultivation.

3.1.2. General Procedure of a "Full" SIP

Before applying a SIP procedure to a reactor often the gas supply and the exhaust gas filter must be sterilized first. When two sections are separately steam sterilized, these are usually connected by a sterile cross (Fig. 11).

The general SIP procedure of a "full" reactor can be described as follows:

1) Sensors requiring off-line calibration (e.g., pH sensors) are calibrated and mounted; the reactor is equipped with all fittings, connections, and plugs.

2) The reactor is filled with media or another liquid.
3) The medium is heated up while being homogeneously mixed.
4) At around 100 °C the liquid boils, the generated steam expels the air from the reactor.
5) Above 100 °C, the reactor must be heated and pressurized to reach the sterilization temperature. Therefore the vents must be closed. Out- and incoming pipes require a flow of steam to heat them up during the sterilization process. Steam-traps or pulsed valves can be used for this purpose (Fig. 11). The temperature of such "critical spots" might be monitored to control the lethal effect (e.g., calculation of F^{z}_{temp}).
6) Depending on the sterilization strategy the temperature measured in the reactor and on all "critical spots" has to be above a critical sterilization temperature (e.g., 121 °C) for a defined time (e.g., 20 min).
7) When cooling down, the pressure of the reactor drops. All valves or steam traps must be closed. To avoid an underpressure in the reactor, sterile gas should be injected to keep the pressure above ambient pressure. The exhaust gas line should be closed to avoid accumulation of condensate.
8) The reactor cools down and once the cultivation temperature is reached the air flow through the reactor should be operated to dry the inlet and outlet gas filter.

Tubes dipping into medium, such as spargers, require special care during full sterilization. When simply connecting the dipped tube to a pulsing valve, steam trap, or other vent, media would flow out of the reactor, as soon it is steam pressurized (above 100 °C). This must be prevented by opening a bypass in between the head space of the reactor and the line connected to the dipped tube (Fig. 11) or by directly supplying clean steam through the dipped tube.

The filter is separated from the tank by a sterile cross (valves 4,5 and 6). In a first step the filter is steam sterilized, steam is supplied via valve 1 and condensate drained through valve 3 and 6. The filter cartridge, the pipe in the direction of valve 4 and the diaphragm seat of valve 4 are sterilized and the procedure is monitored by measuring the sterilization temperature (TIR). In a second step the filled tank is sterilized. The medium is heated by heating the jacket of the tank. To prevent draining of the medium through the dip tube, the bypass valve 7 needs to be opened. Valve 4 is closed, valve 5 and 6 are open to drain the condensate. Again the sterilization is monitored with the TIR.

Figure 10. Media preparation tank for separate sterilization of carbohydrates or concentrated feed, steam-sterilizable, temperature-controlled, and agitated.

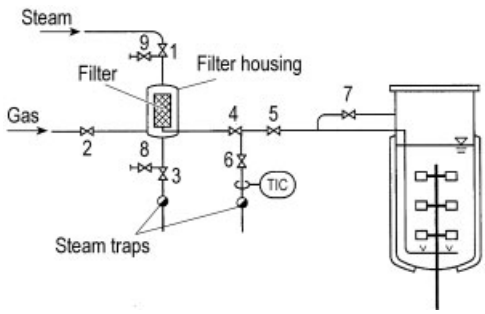

Figure 11. Example of a design for a gas supply filter. Separate sterilization of the filter, temperature monitoring and integrity test are the main characteristics of the design.

Valves 8 and 9 (always manual valves) are required for the integrity testing of the filter cartridge (absolute filter). Since the valves are mounted on the non-sterile side of the filter, the

integrity test can be performed before or after sterilization of the filter. Usually the pressure decay test is performed [39].

3.1.3. General Procedure of an "Empty" SIP

The general "empty" sterilization procedure is very similar to the "full" SIP. Steam injected to the reactor directly drives the air off. The tube of the sparger or of a CIP line can for example be used to supply the steam to the reactor. Condensate accumulating on the bottom must be drained. To check the sterilization process, the temperature on the lowest point of the reactor must be traced. The design of the reactor must consider that larger quantities of air need to be removed in the heat-up phase and a higher flow rate of sterile air is required to break the vacuum during cooling (compared to the full SIP).

3.2. Inoculation

The inoculation volume required for the total working volume is an important parameter when designing a plant of several reactors. Depending on the bioprocess the inoculation volume is 2 to 20 % of the working volume. Larger inoculation volumes will shorten the cultivation time, but require additional or larger reactors to produce the inoculum. If the initial amount of biomass is supposed to be proportional to the inoculation percentage (I = inoculation volume per working volume), the time savings t when starting the cultivation with a higher inoculation percentage (I_h compared to the lower I_l) can be calculated as:

$$t = \frac{1}{\mu_{\max}} \cdot \ln\left(\frac{I_h}{I_l}\right) \tag{3.1}$$

For a given microorganism growing at $\mu_{\max} = 0.5\,\mathrm{h}^{-1}$ the time saving when inoculating at 10 instead of 2 % will be 3.2 h. When considering a 1500 L production reactor (1000 L working volume) the total volume of the seed reactor used to produce the inoculum will be 150 (10 %) instead of 30 L (2 %). Each project requires an analysis of the resulting process times and investment costs for the vessels. The timing must be integrated in the upstream and downstream process. Small reactors (about 2 to 30 L total volume) are inoculated with cultures grown in Erlenmeyer flasks (→ Biotechnology (Inoculation), Chap. 3.5.2.).

A fermentation plant at production scale often consists of one large production reactor, which may be inoculated via two seed reactors. This results in a higher security level of the production process, if one seed reactor is contaminated or organisms are not well growing, the other remains as inoculum.

3.3. Batch Cultivation

A batch cultivation is characterized by the growth of microorganisms without supply of additional substrate after inoculation. While the substrate is metabolized, biomass and products are formed during cultivation (Fig. 12 A). The batch is stopped and harvested when the desired product concentration is maximal. Thereafter the reactor is cleaned and sterilized for the next batch.

During batch cultivation several components other than substrate might be added. In aerobic cultivation air or O_2 is continuously added through a sparger. To maintain constant pH, alkali or acid components are added to the broth. Antifoam agents are required for foam control. Agents might be added to the broth to induce the production of a desired protein. These examples show that batch processes require several sterilizable transfer lines between the holding tanks and the reactor.

The "repeated batch" operation mode consists first of a batch cultivation. Instead of harvesting the total volume some broth remains in the reactor as an inoculum. Fresh sterile media are added to immediately restart a batch. This procedure can be continued until the production rate declines. The time-consuming cleaning and sterilization in between two batches is avoided. The problem here consists in the validation of the process. The starting conditions are not clearly defined, because residues from former batches might influence the cultivation.

Batch processes are simple, reliable and therefore widely used in biotechnology.

The limit of a batch process is set by the maximal initial substrate concentration. As described in Section 2.1.2.2, high concentrations of substrates might inhibit growth.

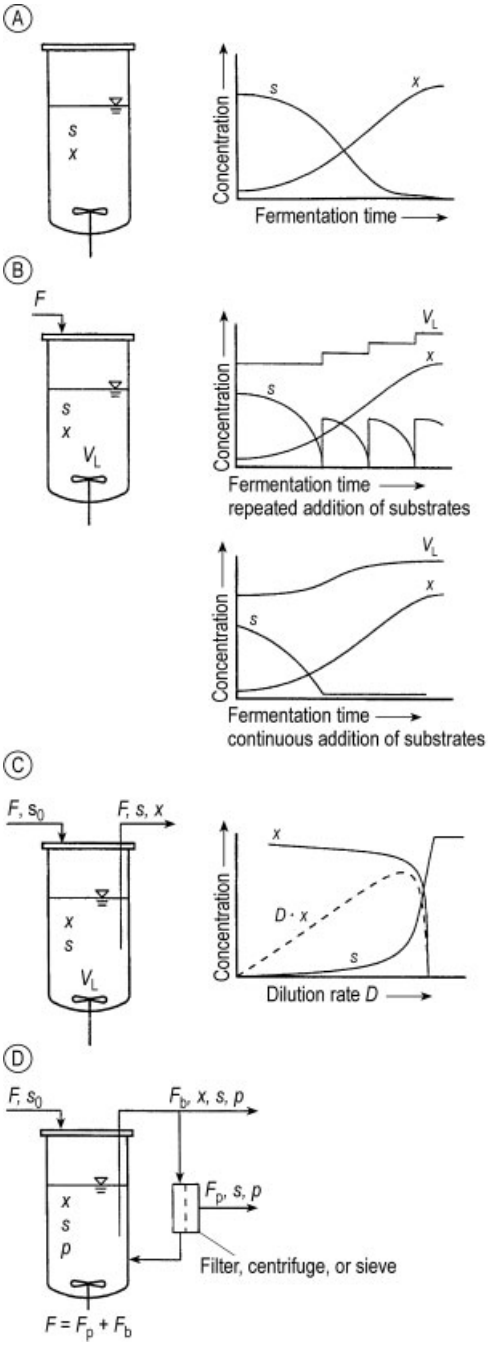

Figure 12. Several standard bioprocesses
A) Batch cultivation; B) Fed-batch cultivation; C) Continuous cultivation; D) Continuous cultivation with cell recycle
V_L = Working volume; F = Flow of fresh substrate; F_P = Permeate flow, cell free flow of product; F_b = Bleed, flow of broth with cells out of the bioreactor; s = Concentration of substrate; x = Concentration of biomass; p = Concentration of product; s_o: Concentration of substrate in the flow F; $D = F/V_L$ Dilution rate

3.4. Fed-Batch Cultivation

To overcome the growth limitation due to the initial substrate concentration (see Section 3.3), substrates can be added to the broth during cultivation. This mode of operation is called fed-batch cultivation. Substrates are either supplied stepwise during the cultivation or continuously (Fig. 12 B). Depending on the bioprocess different fed-batch techniques are in use:

1) *Constant or Predefined Growth Rate.* To maintain a steady growth rate or exponential growth, the flow of substrate must, to a certain extent, increase exponentially. Linear growth of biomass is achieved by a constant flow of substrate.
2) *Bioprocess Requiring a Constant Substrate Concentration.* To maintain a constant substrate concentration in the media, the substrate concentration must be measured or estimated. The substrate feed-rate is adjusted according to the measurement. The feeding strategy can be automated if an on-line estimation or measurement of the substrate concentration and a controller for the feed are installed. However, measurement (estimation) of the substrate concentration is complex and difficult. The controller requires measurement kinetics, which are about 10 times faster than the growth kinetics.
3) *Substrate-Limited Fed-Batch. "Manual adjustment".* Several bioprocesses require low substrate concentration for optimal production. For this method the substrate must be supplied continuously with a pump. Defined cultivation media (without complex nutrients such as yeast extract) are required. A minimal substrate concentration can be maintained, if the substrate feed rate is kept below the maximum amount, which could be metabolized by the organisms. To check the substrate limitation the supply is stopped from time to time. When in the following seconds the dissolved oxygen concentration is sharply rising, this indicates, that the process is substrate limited. The observations are used to adjust the feed rate, which is immediately started after the check.

4) *pO_2 Controlled feeding* is a variation of the substrate-limited feeding strategy mentioned above. pO_2 Controlled feeding is used when low substrate concentration are promoting the production. The control strategy requires a constant $k_L a$ and a constant hypothetical saturation concentration $c^*_{O_2}$ (stirrer speed, aeration flow rate, O_2 partial pressure of the inflowing gas and pressure in the vessel must be kept constant). Under these conditions and if the fed-batch is substrate limited, the dissolved pO_2 will rise if the feed rate is reduced and vice versa (see Section 2.6.4). This behavior is used to automate the controlled feeding.

The fed-batch technique is the state of the art for the efficient production of bacteria and fungi. Modern biotechnology processes with genetically engineered microorganisms, often use fed-batch techniques to reach high cell densities before adding the inducing agent to start the production of the desired product.

Two reasons may limit the growth of fed-batch cultivation: The accumulation of inhibiting products in the broth or lack of reactor volume for substrate addition. Optimizing a fed-batch consists in concentrating the added substrates and starting the batch with a low working volume. The minimal working volume of a reactor is therefore an important criterion. It depends on the location of the sensors which must all dip into the broth, the jacket or heat exchanger which must ensure a safe temperature control and the position of the lowest impeller (STR) to ensure good mixing.

This mode of operation requires a separate tank where the concentrated feed is prepared and sterilized before being transferred.

3.5. Continuous Cultivation

Continuous cultivation operation is characterized by the continuous addition of fresh media and the withdrawal of broth at constant volume [24]. Fresh nutrients promoting the growth are added, while growth inhibiting metabolic products are withdrawn from the system. The parameter characterizing the continuous cultivation is the dilution rate D:

$$D = \frac{F}{V_L} \tag{3.2}$$

with F representing the flow rate of fresh medium and V_L the working volume of the cultivation. When formulating a mass balance for the substrate and biomass, without considering the product formation or the substrate requirement for maintenance, the following equations can be written (see Section 2.1.2 and Fig. 12 C):

$$\frac{dx}{dt} = (\mu(s) - D) \cdot x \tag{3.3}$$

$$\frac{ds}{dt} = D \cdot (s_o - s) - \frac{\mu(s) \cdot x}{Y_{xs,obs}} \tag{3.4}$$

Of special interest are the steady states of a continuous cultivation. It shows that the specific growth rate $\mu(s)$ equals the dilution rate D. From Equations (3.3) and (3.4), and when considering the Monod kinetics (Eq. 2.3) the substrate and biomass concentration can be calculated as:

$$s = \frac{D \cdot K_s}{\mu_{max} - D} \tag{3.5}$$

$$x = Y_{xs,obs} \cdot \left(s_o - \frac{D \cdot K_s}{\mu_{max} - D} \right) \tag{3.6}$$

The maximal dilution rate D_{max} before washing microorganisms out of an ideally mixed reactor is:

$$D_{max} = \frac{\mu_{max} \cdot s_o}{s_o + K_s} \tag{3.7}$$

When the biomass is the desired product the productivity P is:

$$P = D \cdot x \tag{3.8}$$

The main advantage of continuous processes is that productivity is theoretically maintained for indefinitely. The turn-down time of a plant is kept low. For slowly growing organisms such as animal cells with product formation proportional to the biomass growth continuous processes are often chosen.

There are two standard types of continuous cultivation. The *chemostat cultivation* is characterized by a continuous substrate feed, where at least one nutrient is limiting. The steady states of the cultivation are of interest and are used to characterize the growth of the organism. In a *turbidostat cultivation*, the biomass concentration is measured and kept constant by varying the dilution rate. In both cases a sterile media supply tank and a harvesting tank are required. Level controller, scales or metering pumps are used to control the feed rate (Section 2.6.7).

The productivity of a continuous bioprocess is limited by the washing out of the cells (cell density x) and the limited dilution rate D.

3.6. Continuous Cultivation with Cell Retention

To overcome the production limitation of a continuous cultivation, organisms must somehow be retained in the reactor. Such a process would allow the increase of the cell density and of the dilution rate, both factors increasing the productivity (see Section 3.5). The process would uncouple the supply of nutrients and the withdrawal of inhibiting products from the cell density in the reactor.

Most of the applications of continuous cultivation with cell retention are found in animal cell cultivations. Several techniques retain suspended cells in the broth while other immobilize the cells on matrices.

3.6.1. Filtration Systems for Suspended Organisms

To retain suspended cells in reactors a filtration process can be used. The biomass is concentrated by withdrawing cell-free culture broth. If the desired product is dissolved in the broth (extracellular production), the procedure enables the continuous harvest in the cell-free permeate (Fig. 12 D). The production rate compared to a classical continuous cultivation is increased, because of higher cell densities and dilution rates.

To continuously withdraw dead cells from the cultivation and control the cell density a small flow rate, called bleed (F_b) is taken from the reactor.

Filtration units are either directly incorporated in the reactor or placed in a bypass. The first design has the advantage of being a compact system. The second is more flexible, because the unit can be dismounted, cleaned, sterilized and reconnected during cultivation.

Filtration or membrane systems are characterized by the membrane material, flow/pressure characteristics across the membrane and the required membrane surface-to-volume ratio. When scaling-up a membrane system, the surface-to-volume ratio should be constant. It is the factor, that limits the possibilities of scale-up.

Another difficulty is the fouling or clogging of the membrane. The transmembrane flux has the tendency to settle particles or organisms on or within the membrane. Thus the pressure drop across the membrane is increased or the permeate flow decreased [43]. The filtration methods described in the following try to overcome this fouling using special techniques.

3.6.1.1. Cross-Flow Filtration

(→ Biochemical Separations, Chap. 4.1.2., → Biochemical Separations, Chap. 4.1.3.; → Filtration, Chap. 2.4.)

To keep the membrane surface free of organisms, cross-flow filtration units maintain a tangential flow of the feed stream parallel to the separation surface. The aim is to provide sufficient shear at the membrane surface to keep solids and other particles from settling [44]. At a laboratory scale, filtration units built into the reactor [45] or in a bypass [46] were very successful. However the difficulty to maintain a constant ratio of membrane surface per reactor volume of about $20\,m^2/m^3$ limits the possibilities of scale-up.

3.6.1.2. Rotor Filter

The rotor filter (also called rotating sieve or spin filter) represents another method to avoid the fouling of the membrane (Fig. 13). When built in a narrow cylinder, Taylor vortices create the necessary turbulence in the vicinity of the sieve preventing solids from settling. Centrifugal forces might also have a positive influence [47]. The rotor filter is very popular for animal cell cultivation and is found in many production processes and design variations [48]. The membrane consists of a cylindrical stainless steel sieve available with various mesh sizes (2 to 60 µm). It is interesting that cells of about 5 to10 µm in diameter are retained by up to 90 % by a sieve with mesh size of 20 µm, rotating at 1000 rpm (cylinder diameter 100 mm). The permeate is not completely cell-free, but dead cells are thereby directly withdrawn from the system (bleed flow). Stable permeate flow rates of up to 40 to $150\,L \cdot m^2 \cdot h^{-1}$ can be achieved over several weeks with animal cell culture densities of up to 10^7 cells per milliliter. Such cell densities are reached by exchanging the working volume one to two times per day ($D = 0.041$ to

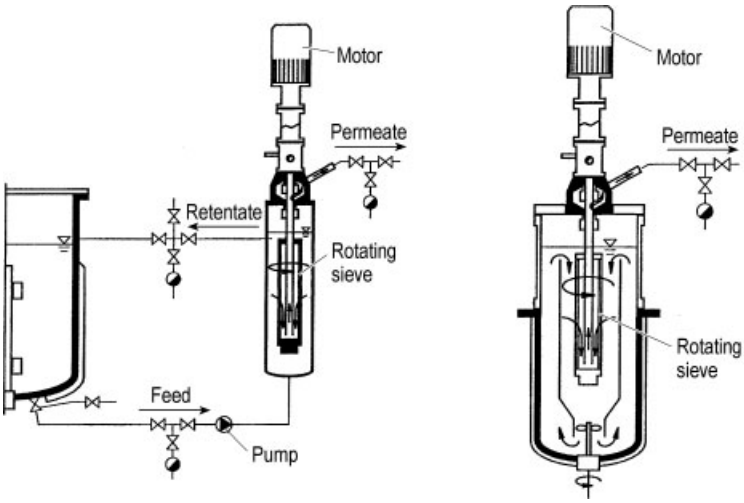

Figure 13. Rotor-filter build into the bioreactor or in a bypass circuit. The rotating speed can be adjusted separately from the stirrer speed.

$0.083\,h^{-1}$). This results in reasonable sieve areas per reactor volume of about 0.3 to $2\,m^2 \cdot m^{-3}$.

When placing the rotor filter within the reactor, the system does not require a pump for recirculation of broth. The bypass system can be separated, cleaned, sterilized, and reconnected to the reactor (Fig. 13). Twin filter systems enable to switch over to a second rotor filter and clean the first in the meantime (Fig. 14). Magnetic drive gear pumps, diaphragm pumps, or peristaltic pumps are suitable to circulate the broth without cell damage. The circulation flow should be about 10 to 20 times higher than the permeate flow.

3.6.2. Immobilized Organisms

Organisms are immobilized when they form cell aggregates (pellets), adhere on porous matrices, or are entrapped in beads. The possible technique depends on the considered organism and its characteristics. Immobilization is mainly applied to animal cells and mycelia type fungi.

Immobilized organisms are cultivated in fixed-bed or fluidized-bed reactors.

3.6.2.1. Pellets, Porous Matrices, or Beads

Mycelia type fungi can spontaneously form pellets if the media or the environment are manipulated to a certain extent (pH, addition of components). An immobilization on porous microcarrier such as celite beads (porous diatomaceous earth) is also possible. A positive side effect is a higher *OTR* compared to the viscous broth of mycelia [49].

Although relevant industrial mammalian cell lines, such as Chinese hamster ovary (CHO) or Baby hamster kidney (BHK) have been successfully adapted to grow in suspension, similar to "classical" suspendable cell lines like hybridoma, most primary cells (i.e., hepatocytes, epidermal cells) and most cell lines derived from tissues still need a surface for adherence in order to proliferate. Among these anchorage-dependent cells are cell lines (i.e., VERO, MDBK) for the production of viral particles used as veterinary and human vaccines.

Immobilized animal cells are protected from shear forces. This allows higher mixer speed and aeration rates. Because of this advantage even non-adherent cell lines are often immobilized for cultivation with cell retention.

Anchorage-dependent cells can be immobilized on porous carriers such as glass, cellulose, dextran, collagen, or gelatin. Macrocarriers have diameters of 0.4 to 5 mm and allow the cells to grow inside the porous structure. Microcarriers have diameters of only 0.1 to 0.2 mm. Here, cells grow mainly on the surface in monolayers and therefore are not so well protected. Cell densi-

ties on or in the carriers of 0.5 to 5×10^8 cells per milliliter carrier were reported. The homogeneity of parameters and the mass transfer of nutrients or metabolic products in or out of carriers might become critical for larger carriers.

To immobilize nonadherent cells, carriers with modified surfaces enabling an adherence can be used. Another possibility consists in encapsulating the cells into beads or perform an entrapment in a matrix (e.g., calcium alginate, agarose, carrageenan [50], [51]).

Figure 14. Twin rotor-filter for cell retention in continuous cultivation. Flexible separate unit connected by steam sterilizable transfer-lines to the bioreactor.

3.6.2.2. Fluidized-Bed Reactors

Reactors keeping the solids (carriers, pellets, or beads) in suspension are called fluidized-bed reactors. Usually three phases (solid, liquid, and gaseous) are involved. To keep the solids in suspension all types of mixing procedures can be used provided the solids are not damaged.

When a settler is used the density of the solids should be slightly higher than that of the media. A good settling is achieved, using beads with densities of 1.03 to 1.045 kg/L. Reactors with settling zones allow direct withdrawal of the cell- and particle-free broth (Figs. 15 and 16). A settler in an external loop (e.g., lamella clarifier) can also be used.

Stirred tank reactors, which usually do not have any settling zone require a separation technique to withdraw the permeate. A simple stainless steel mesh might already be sufficient for small-scale reactors (< 30 L). At larger scale a rotor filter (Section 3.6.1.2) can be used for separation.

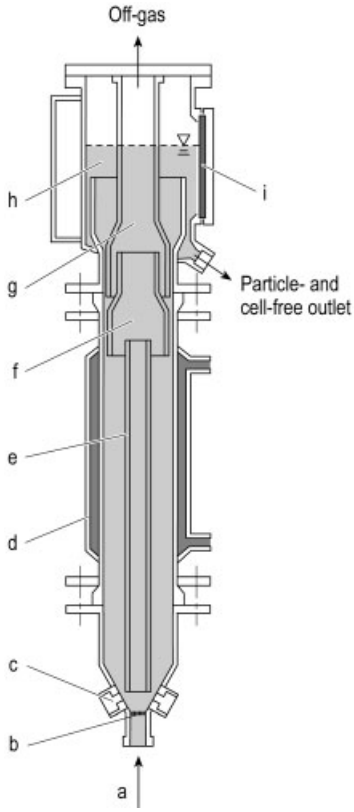

Figure 15. Three phase fluidized-bed reactor for immobilized cultures. The separation system in the headspace of the reactor allows the particle-free withdrawal of media.
a) Air inlet; b) Gas disperser, e.g., frit or jet; c) Sampling port; d) Jacket for temperature control; e) Draft tube; f) Circulation chamber; g) Degassing chamber; h) Separation chamber; i) Viewing glass

3.6.2.3. Fixed-Bed Reactors

Instead of keeping the immobilized cells in suspension (fluidized-bed reactor) they can be trapped in a fixed bed. The system therefore does not require a separation technique to withdraw

permeate. The medium is circulated through the fixed bed to supply nutrients and withdraw metabolites or products from the cells. The medium has to be oxygen enriched before being circulated through the fixed bed. The length of the fixed bed is the critical parameter, because the dissolved oxygen concentration in the medium decreases when flowing through the fixed bed with metabolizing cells [52]. A radial flow through a cylindrical fixed bed is proposed to allow a scale-up into length (Fig. 17 and 18) [53].

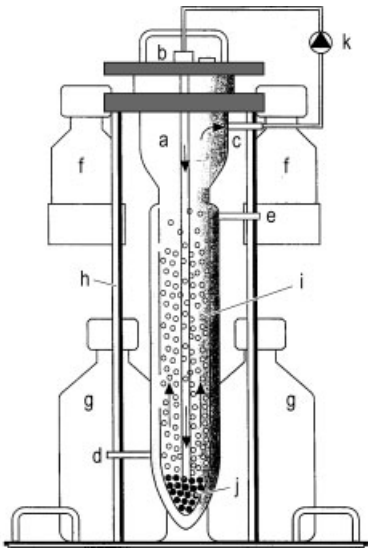

Figure 16. Fluidized-bed reactor. The medium is circulated with a peristaltic pump. Upflowing media keep carrier in suspension. In the upper part, due to the larger diameter, the media flow is reduced to allow sedimentation of carrier.
a) Glass vessel with double jacket; b) Medium circulation inlet; c) Medium circulation outlet; d) Temperature control fluid inlet; e) Temperature control fluid outlet; f) Storage vessels base/acid; g) Storage vessels harvest substrate; h) Frame; i) Carrier material, pellet; j) Steel ball; k) Hydraulic pump

3.7. Dialysis Cultivation

High cell densities or high production rates can be achieved by performing a dialysis process during cultivation. Extremely high cell densities of *E. coli* up to 174 g/L cdw were for example reached [54].

The driving force of the diffusive mass transfer across a membrane for a given compound is its concentration difference between both sides of the membrane. Mass transfer in both directions across the membrane is therefore possible [55], [57]. Nutrients can be supplied across the membrane to the cultivation while inhibiting products are withdrawn. Depending on the cut-off of the membrane compounds of higher molecular mass (e.g., > 10 000) will be retained by the membrane. That way a concentration of products can be achieved.

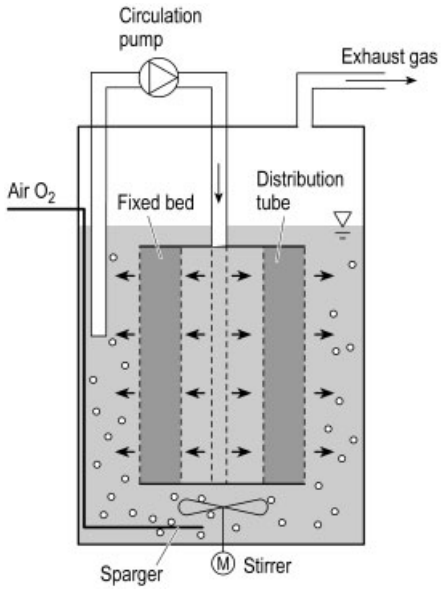

Figure 17. Radial flow fixed-bed bioreactor; media circulation is performed with an external hydraulic pump

Figures 19 and 20 show a reactor with an integrated dialysis membrane separating two reaction chambers. A coculture of two different strains separated by the membrane can be studied in such a laboratory system. Generally fast growing bacteria require a dialysis membrane surface ratio per reactor volume of at least 10 to 20 m^2/m^3. Reactors with a volume of more than 30 L therefore require the dialysis membrane to be mounted in an external loop [56].

Batch, fed-batch, continuous, or continuous cultivation with cell recycle can be performed in dialysis membrane reactors.

3.8. Selection of Equipment Related to Specific Processes or Products

Tables 9, 10, 11 are intended to give an overview on suitable equipment for different types of or-

ganisms, products, and the type of production. Such tables can only give a rough idea of what kind of equipment might be selected for a specific process, because each project has its special features. The huge variety in biotechnology makes it very difficult to standardize equipment for bioreactors and fermentors.

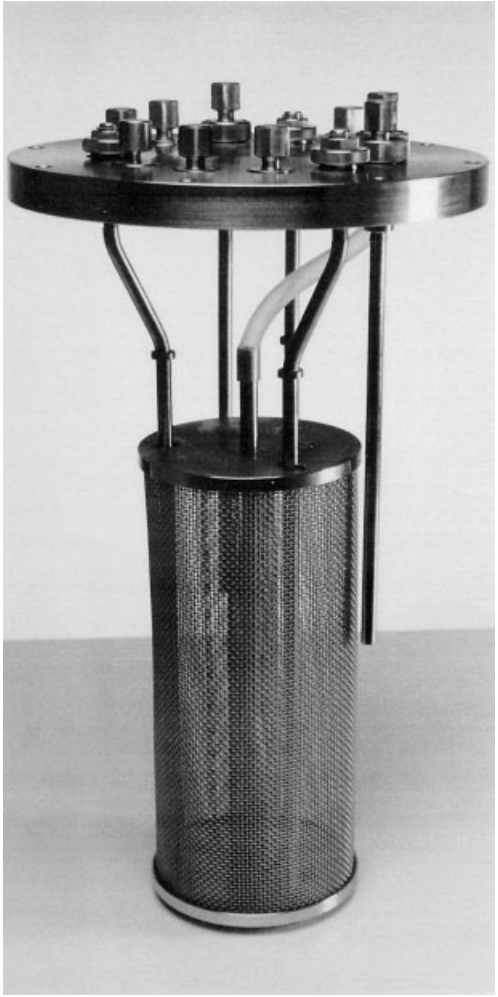

Figure 18. Photograph of a radial flow fixed-bed bioreactor

4. Materials

Materials used in the construction of fermentors and bioreactors (vessel, pipes, fittings, instruments) are usually long-term exposed to moderate conditions, i.e., temperatures between 28 and 40 °C, pH values between 6 and 8 and a salt concentration below 5 g/L. Exceptions are reactors used for cultivation of extremophiles or anaerobic microorganisms. These reactors contain corrosive media (pH 1.5 – 10, NaCl concentration up to 100 g/L, sulfur compounds). Parts in the periphery of the main reactor such as storage tanks for acid (hydrochloric, nitric, phosphoric, or citric acids) or alkali (ammonia, sodium hydroxide) and transfer pipes are also exposed to highly corrosive media. Sterilizable reactors must resist temperatures up to 125 °C and pressures of 1.5 bar for a short time. Temperature circuits are exposed to 140 °C and 2.5 bar. Aggressive agents are used to clean the inside of vessels and pipes. In some cases ethanol, formaldehyde, or other aggressive agents are also applied to the outer side of the equipment. In that case painted ferritic steels and thermoplastics should not be used, preference should be given to stainless steel.

The surfaces of materials must be smooth and easy to clean. Generally it is differentiated between materials in contact and those not in contact with the product (culture broth).

4.1. Austenitic Steel (→ Steel)

With a few exceptions, austenitic steels (with a maximum of 0.08 % carbon) are the material of choice in biotechnology. Compared to the nonaustenites, these steels show improved corrosion and heat resistance, and they are nonmagnetic. Other advantages are good tenacity and workability [58], [59].

4.1.1. Different Alloys of Austenitic Steels

The characteristics, composition and name of the different austenitic steels used in biotechnology are summarized in Table 12.

In general, vessels used in biological processes are manufactured from 316 or 316L steel. The expensive 316L has both excellent electropolishing and welding characteristics. The cheaper and less corrosion-resistant steel grades (304, 304L) are mostly used in food technology or for certain noncritical harvest or storage tanks such as waste tanks. Parts of the vessel not in contact with the product, such as the heating circuit the heat-transfer jacket or the platform,

Table 9. Suitable equipment for specific processes

	Type of organism									
	Bacteria aerobic	*E.coli*	Bacteria anaerobic	Yeast	*Pichia pastoris*	Animal cells	Hybridoma cells	Insect cells	Adherent cells on carriers	Algae (phototroph)
Equipment, mixing										
STR	**	**	**	**	**	**	**	**	*	**
Marine impeller			*			**	**	**	*	**
Rushton impeller	**	**	*	**	**	*	*			
Baffles	**	**	*	**	**	*	*			*
Airlift reactor	*			*		**	**		**	**
Vibromixer			*			*	*	*	**	**
Hydraulic mixing	*			*					**	
Mechanical seal	**	**	**	**	**	*	*	*		**
Magnetic coupling	*					**	**	**	**	**
Top drive	*	*	*	*	*	**	**	**	**	*
Bottom drive	**	**	**	**	**	**	**	**	*	**
Jacket for heat transfer	*	*	*	*	*	**	**	**	**	
Coils for heat transfer	*	*	*	*	*					*
Equipment, other										
Stainless steel vessel	**	**	**	**	**	**	**	**	**	*
Glass vessel	*	*	*	*	*	*	*	*	*	**
Plastic film sleeve reactor	*	*	*	*	*	*	*	*	*	**
Depth filter (gas)	*	*	*	*	*					
Membrane filter (gas)	*	*	*	**	**	**	**	**	**	**
Ring sparger, fish tail for gassing	**	**	*	**	**	*	*			**
Sinter metal sparger			*			**	**	*	*	**
Surface gassing			**			**	**	**	**	*
Bubble free membrane gassing						*	*	**	**	
Process										
Batch	**	**	**	**	**	**	**	**	**	**
Fed-batch	**	**	**	**	**	**	**	**	**	**
Continuous cultivation	*	*	*	*		**	**	**	**	**
Rotor-filter						**	**	**	**	*
Fluidized bed						**	**	**	**	*
Fixed-bed						**	**	**	**	*
Dialysis	*	*	*	*	*	**	**	**	**	*
CIP	*	*	*	*	*	*	*	*	*	*
SIP	*	*	*	*	*	*	*	*	*	*
Sterilization in the autoclave	*	*	*	*	*	*	*	*	*	*
Containment	*	*	*	*	*	*	*	*	*	

** = Very suitable. * = Suitable, possible, but depending on other parameters. = usually not suitable, / = not applicable.

Table 9. (Continued)

	Type of organism									
	Bacteria aerobic	*E.coli*	Bacteria anaerobic	Yeast	*Pichia pastoris*	Animal cells	Hybridoma cells	Insect cells	Adherent cells on carriers	Algae (phototroph)
Type of production										
GMP production	*	*	*	*	*	*	*	*	*	*
Bulk product	*	*	*	*	*					*
Product for food industry	*	*	*	*	*					*
Large scale $> 3.5\,m^3$	**	**	**	**	**	**	**	**	**	**
$0.3 <$ pilot scale $< 3.5\,m^3$	**	**	**	**	**	**	**	**	**	**
Laboratory scale $< 0.3\,m^3$	**	**	**	**	**	**	**	**	**	**

mm

** = Very suitable. * = Suitable, possible, but depending on other parameters. = usually not suitable, / = not applicable.

are usually made of 304, 304 L or non-stainless steel materials.

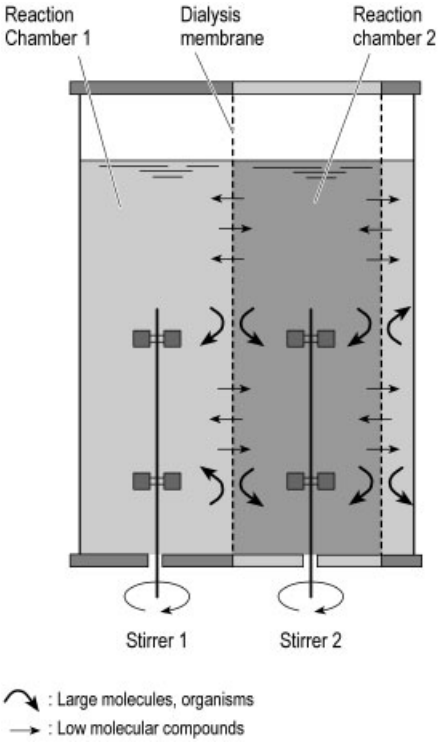

Figure 19. Schematic of a dialysis membrane reactor. A dialysis membrane separates two reaction chambers from each other. Low molecular compounds, driven by the concentration difference in between the two chambers, permeate through the membrane. One or both chambers can be used for the cultivation of microorganisms.

4.1.2. Welding

Gas tungsten arc welding (GTAW) is the most common technique to weld vessels and pipes in biotechnology. Fillet weld joints mostly applied to vessels are polished on the inside of the vessels. Polishing the partly overlapping fusion points of such fillet welds on the outside of a vessel gives a smooth surface but a visual control of the welds is no longer possible. Automatic welding such as orbital welding (*butt welds*) is used to connect pipes wherever the geometry of the pipework allows this technique.

A considerable effort is put into inspection and labeling of weld seams. To control the welding each weld seam needs to be labeled on labeled pipes and components. Drawings or isometrics of pipes may be provided to indicate the location of the weld seam. The date, the name of the welder or parameters of the welding machine, and the label of the pipe are reported for each weld seam in a log file.

Classes of acceptance and acceptance criteria for weld seams are defined by the International Institute of Welding (IIW, classes 1 – 5; 5: highest quality) or by CEN 25817 (class B, C and D; B: highest quality).

Examples of inspections include:

1) Spot check of 10 – 100 % of the welds with a boroscope to detect cracks, incomplete fusions or alignment deviations. Photographs, or video of this inspection might be stored.
2) Surface and color of the seam have to satisfy acceptance criteria.

Table 10. Suitable equipment for specific products

	Products											
	Amino acids, vitamins	Organic acids, ethanol	Technical enzymes	Single cell protein	Baker yeast	Antibiotics	Vaccines	Tetanus vaccine	Protein (pharma industry)	Monoclonal antibodies	Virus (VERO cells)	Viruses from bacteria
Equipment, mixing												
STR	**	**	**	*	**	**	**	*	*	**	*	**
Marine impeller			*				*	*	*	**	*	*
Rushton impeller	**	**	**	*	**	**	**		*	*		**
Baffles	**	**	**	*	**	**	**		*	*		**
Airlift reactor	*	*	**	**	**	*	*		*	*	*	
Vibromixer							*	**	*	*	*	
Hydraulic mixing	*	*		*	*							
Mechanical seal	**	**	**	*	**	**	*		*	*	*	**
Magnetic coupling							**	**	*	**	**	*
Top drive	**	**	**	*	**	**	**	**	*	*	**	*
Bottom drive	*	*	*	*	**	**	**		*	**	*	**
Jacket for heat transfer	*	*	*	*	*	*	**	**	*	**	**	**
Coils for heat transfer	**	**	**	**	**	**	*		*			*
Equipment, other												
Stainless steel vessel	**	**	**	**	**	**	**	**	**	**	**	**
Glass vessel												
Plastic film sleeve reactor												
Depth filter (gas)	*	**	**	**	**	**						
Membrane filter (gas)	**	*	*	*	*	**	**	**	**	**	**	**
Ring sparger, fish tail for gassing	**	**	**	**	**	**	*		*	*	*	**
Sinter metal sparger							*	*	*	**	**	
Surface gassing							*	*	*	**	**	
Bubble free membrane gassing							*		*	*	*	
Process												
Batch	**	**	**	**	**	**	**	**	*	*	**	**
Fed-batch	**	**	**	**	**	**	**	*	*	*	**	**
Continuous cultivation		**	*	*	*				*	**		
Rotor-filter									*	**		
Fluidized bed									*	**	**	
Fixed-bed									*	**	**	
Dialysis	*	*	*	*	*	*	*	*	*	*	*	*
CIP	*	*	*	*	*	*	*	*	*	*	*	*
SIP	**	**	**	**	**	**	*	**	*	*	*	*
Sterilization in the autoclave							*	*	*	*	*	*
Containment							**	**	*	*	**	**

** = Very suitable. * = Suitable, possible, but depending on other parameters. = usually not suitable, / = not applicable.

Table 10. (Continued)

	Products											
	Amino acids, vitamins	Organic acids, ethanol	Technical enzymes	Single cell protein	Baker yeast	Antibiotics	Vaccines	Tetanus vaccine	Protein (pharma industry)	Monoclonal antibodies	Virus (VERO cells)	Viruses from bacteria
Type of production												
GMP production						*	**	**	**	**	**	**
Bulk product	**	**	**	**	**	**						
Product for food industry	**	**	**	**	**	**						
Large scale $> 3.5\,m^3$	**	**	**	**	**	**	*	*	*	*	*	*
$0.3 <$ pilot scale $< 3.5\,m^3$	*	*	*	*	*	**	**	**	*	*	**	*
Laboratory scale $< 0.3\,m^3$	*	*	*	*	*	*	**	**	*	*	**	*

** = Very suitable. * = Suitable, possible, but depending on other parameters. = usually not suitable, / = not applicable.

3) Spot check (e.g., 10 %) with X-ray inspection to detect pores or slag intrusions.
4) Welders certificates need to be provided.
5) Parameters of automatic welding machines must be logged. Tests of weld seams may be performed and kept for later inspection.

The time required for those inspections and the data logging can be estimated with $\frac{1}{2} - 1$ h per weld seam.

Generally the surface criteria (corrosion, hygienic constraint) is more important than the detection of pores/inclusions (affecting the stability) in the inside of a weld seam. Therefore boroscope inspection makes more sense than X-ray inspection. In addition, X-ray inspection is by far more expensive than the boroscope method.

4.1.3. Surfaces of Austenitic Steels

Surface treatments of austenitic steels in biotechnology are important for two reasons. First, the cleaning of the equipment largely depends on the surface roughness. The smoother the surface is, the easier the cleaning will be, and less product is retained in the pores. However, there is a lower limit of surface roughness of $R_a = 0.4\,\mu m$, under which cleaning time will not further be decreased [35], [60]. Second, on polished surfaces a homogeneous microscopically thin chromium oxide layer will be formed. This layer increases the corrosion resistance of stainless steels.

In general, both the inside and the outside of a reactor have a defined surface finish. Different methods such as cauterization, ball blasting (mainly for outside vessel), polishing, electropolishing, and passivation are used for surface treatment, the final result of the treatment being directly related to the steel quality and the previous working steps [58]. For the inside of vessels a standard surface roughness of $R_a \leq 0.8\,\mu m$ and for the inside of pipes of $R_a \leq 5\,\mu m$ is suggested [35]. Straight pipes are available with surface roughness of $R_a \leq 0.5 - 0.8\,\mu m$. However when bending such pipes the surface roughness will become 2.5 to 4 times higher. For the outside specification of $R_a \leq 2\,\mu m$ are common. High ($R_a \leq 0.5\,\mu m$) and very high qualities of ($R_a \leq 0.25\,\mu m$) surface finishes are sometimes required for the inside of vessels. The increasing demands for surface quality are often difficult to achieve, expensive, and extremely time-consuming. It is questionable whether this expense is made up with improved cleaning or biological results.

4.1.3.1. Definition and Measurement of Surface Roughness

In the past the roughness was defined by the grit size (P60, P100 . . . P320, P400) of the abrasive

Table 11. Suitable equipment for specific types of organisms

	Type of production					
	GMP production	Bulk product	Product for food industry	Large scale $> 3.5\,m^3$	$0.3 <$ pilot scale $< 0.3\,m^3$	Laboratory scale $< 0.3\,m^3$
Equipment, mixing						
STR	*	*	*	**	**	**
Marine impeller	*			**	**	**
Rushton impeller	*	*	*	**	**	**
Baffles	*	*	*	**	**	**
Airlift reactor	*	*	*	**	**	**
Vibromixer	*				**	**
Hydraulic mixing		*	*	**	**	**
Mechanical seal	*	*	*	**	**	**
Magnetic coupling	*			*	**	**
Top drive	*	*	*	**	**	*
Bottom drive	*	*	*	*	**	**
Jacket for heat transfer	*	*	*	**	**	**
Coils for heat transfer	*	*	*	**	*	
Equipment, other						
Stainless steel vessel	*	*	*	**	**	**
Glass vessel	*					*
Plastic film sleeve reactor						*
Depth filter (gas)		*	*	**	*	*
Membrane filter (gas)	**	*	*	**	*	*
Ring sparger, fish tail for gassing	*	*	*	*	*	*
Sinter metal sparger	*				*	*
Surface gassing	*				*	*
Bubble free membrane gassing	*				*	*
Process						
Batch	*	*	*	**	**	**
Fed-batch	*	*	*	**	**	**
Continuous cultivation	*	*	*	**	**	**
Rotor-filter	*			*	**	**
Fluidized bed	*				**	**
Fixed-bed	*				*	**
Dialysis	*				*	**
CIP	**	*	*	*	*	*
SIP	**	**	**	**	**	*
Sterilization in the autoclave	*					*
Containment	*			*	*	*
Type of production						
GMP production	/			**	**	**
Bulk product		/		**		
Product for food industry			/	**		
Large scale $> 3.5\,m^3$	*	*	*	/	/	/
$0.3 <$ pilot scale $< 3.5\,m^3$	*	*	*	/	/	/
Laboratory scale $< 0.3\,m^3$	*			/	/	/

** = Very suitable. * = Suitable, possible, but depending on other parameters. = usually not suitable, / = not applicable.

Table 12. Austenitic steels with application in biotechnology (based on DIN 17440/41)

Name/indication			Alloy component according to DIN W N°, wt %				
DIN, W N°	AISI	Other	C	Cr	Mo	Ni	Other
1.4301	304	V2A	≤ 0.07	17.0 – 19.0		8.5 – 10.5	
1.4306	304L	V2A	≤ 0.030	18.0 – 20.0		19.0 – 12.5	
1.4541	321	V2A	≤ 0.08	17.0 – 19.0		9.0 – 12.0	Ti = (5 × %C) – 0.80
1.4401	316	V4A	≤ 0.07	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	
1.4404	316L	V4A	≤ 0.030	16.5 – 18.5	2.0 – 2.5	11.0 – 14.0	
1.4571	316Ti	V4A	≤ 0.08	16.5 – 18.5	2.0 – 2.5	10.5 – 13.5	Ti = (5 × %C) – 0.80
1.4439	316L	V4A	≤ 0.030	16.5 – 18.5	4.0 – 5.0	12.5 – 14.5	N = 0.12 – 0.22, S $\leq$ 0.025
1.4435	316L	V4A	≤ 0.030	17.0 – 18.5	2.5 – 3.0	12.5 – 15.0	S $\leq$ 0.025

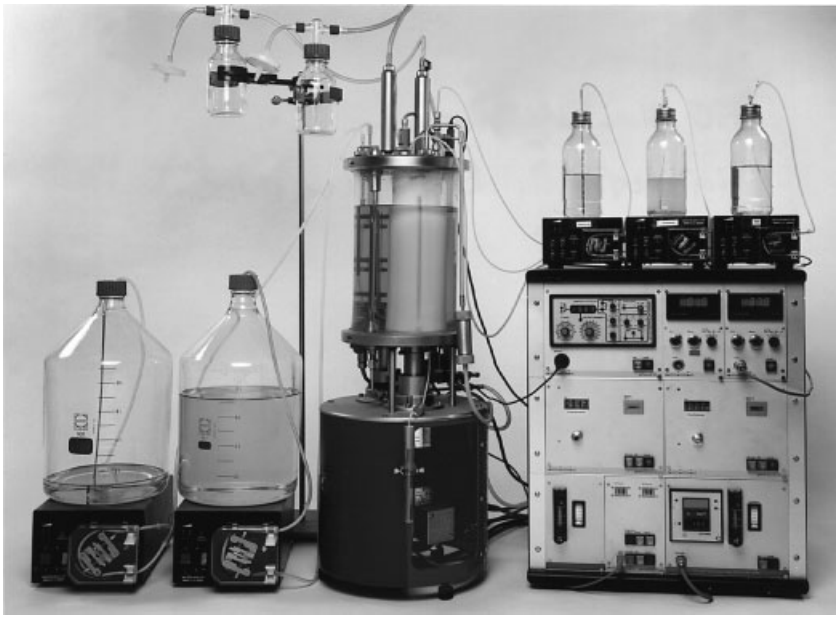

Figure 20. Dialysis membrane reactor

belts. The roughness achieved by polishing with a defined grit size can widely vary since it is related to the belt quality.

Therefore not the grit size but an unambiguous measurement as, e.g., with the electric stylus instrument should be used to evaluate the roughness of a surface (Fig. 21, [61], [62]). Several parameters characterizing the roughness profile are calculated.

1) The arithmetical average of the absolute values of the distances y from mean line to roughness profile within the sampling length (DIN 4768):

$$R_a = \frac{1}{l_m} \int_{x=0}^{x=l_m} y \mathrm{d}x \tag{4.1}$$

2) The average peak to valley height R_z of five adjoining single sampling lengths (Fig 22).

$$R_Z = \frac{Z_1 + Z_2 + Z_3 + Z_4 + Z_5}{5} \tag{4.3}$$

3) The maximum peak to valley height $R_{\max}$ is the highest of the 5 peak to valley heights Z_i (Fig. 22).

In practice surface definitions based on the mechanical treatment with different grain or grit sizes of polish belts are still in use. Manufacturer and customer should define a smoothness such as R_a, which is easily measured and unambiguous. Table 13 gives a rough correlation between the grit and the average roughness R_a.

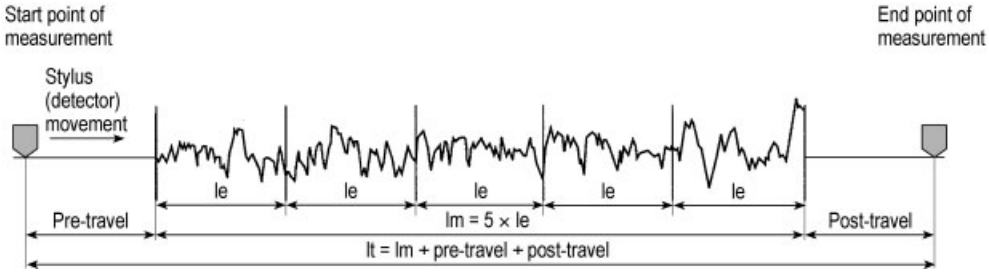

Figure 21. The roughness profile measured with the electrical stylus instrument

Table 13. Approximative correspondence of the arithmetic average roughness R_a, the average peak to valley height R_z and the maximum peak to valley height R_{max} to the grit size

Grit size	60	100	180	220	280	320	400
$\approx R_a$, µm	2.2	1.8	1.3	1.1	0.5	0.45	0.4
$\approx R_z$, µm	13.4	12.8	9.7	7.3	4.4	3.2	2.8
$\approx R_{max}$, µm	18.2	17.1	11.6	8.6	6.5	4.7	3.5

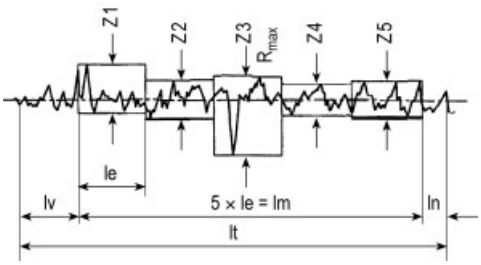

Figure 22. Determination of the average peak to valley height R_z and R_{max} (DIN) from the roughness profile

4.1.3.2. Mechanical Surface Treatment

The surface quality of steels can be greatly improved by mechanical treatment.

Polishing. Polish belts of increasing grit sizes are used (Table 13) to obtain specified smoothness values. The surface quality obtained with mechanical treatment depends on and is limited by the steel quality. Some examples are:

- Titanium-stabilized chrome-nickel or molybdenum steels, 1.4541 (AISI 321) and 1.4571 (AISI 316Ti), are not suited for high surface qualities
- Best possible smoothness values ($R_a = 0.25$ to 0.5 µm) are reached with steel grades 1.4306 (AISI 304L) and 1.4406 (316L)
- For later mirror polishing, a preceding polish to $R_a \approx 0.6$ µm (P220) is completely satisfactory. Often, finer polishing leads to poorer smoothness values after mirror polishing
- Steel used for vessels should be cold-rolled sheet material. It is extremely difficult and time-consuming to polish warm rolled sheets

Ball blasting is used for special surface finishing and gives various possibilities for optical surface effects with a dull-finished or dull-polished appearance. The ball blasting decreases the risk of tension cracks and corrosion. It is used mostly for the vessel exterior.

4.1.3.3. Chemical Surface Treatment

Cauterization. Impurities and surface imperfections prevent the formation of a perfect passive layer. The purpose of cauterization is to dissolve these flaws with suitable acid mixtures, such as 15 – 25 vol % aqueous HNO_3 and 1 – 8 vol % aqueous HF for steel with chromium content above 15.5 %. The impurities that are cauterized may be oxidation tint, welding slag residues, surface impurities containing unalloyed steel, and fine scales or overlaps generated during the working process.

Passivation. The natural passive layer is formed when stainless steel is exposed to air. An artificial passivation with dilute nitric acid is frequently used for corrosion resistance at critical conditions. The oxidizing effect of the nitric

acid accelerates the formation of a dense passive layer. Passivation is either used after cauterization or as a final treatment of ground, brushed, or polished steels with special surface structures that would be destroyed when cauterized. New studies on passivation show that alternatives to nitric acids, such as organic acid chelating agents can produce a long-lasting passive layer [63].

4.1.3.4. Electrochemical Surface Treatment

(→ Electrochemistry, Chap. 4.8.)

Mirror polishing (electropolishing) is used to smoothen and polish rough and dull surfaces. The quality of the mirror polishing is only as good as the source material and its mechanical pretreatment.

Advantages of mirror polishing are:

- Despite good R_a values, the surface may have roughness in the microscopic range, which is smoothed by the electrolytic process
- Corrosion resistance is increased due to exposure to high oxygen concentration, which enables and increases the formation of a dense chromium oxide layer

Disadvantages of electrolytic polishing are:

- Formerly existing imperfections on the material surface are more accentuated after mirror polishing (R_a increasing after electropolishing)
- The difficulty to mirror polish rolled plate steel
- It is impossible to work titanium-stabilized steels (1.4541-AISI 321; 1.4571-AISI 316Ti)

After mirror polishing, the parts have to be thoroughly cleaned to remove remaining electrolyte.

4.1.4. Storage and General Working Rules

Tools such as saws and files must only be in contact with one type of austenitic steel. Never use these tools with ferritic steels.

Workpieces have to be stored on a clean and non-absorbing base support, protected from dust and water spray. Workshops handling ferritic material are not suited for storage. Also, locations near railways and tramways (which may produce ferritic brake dust) and pavements in winter (dry sand and salt) should be avoided. For long storage periods, it is best to cover and wrap up delicate parts, as well as workpieces, in polyethylene film.

4.2. Polymers

Polymers are used for O-rings, gaskets, diaphragms, or filters. Those materials have to withstand temperatures between 0 and 125 °C, in some cases up to 135 °C. An optimal grade of hardness for sealing material is "medium" to "firm" between 65 and 75 Shore A. Polymers must be resistant to acid and alkali solutions (see Chap. 4, Introduction).

Polymers, which are in contact with the product (culture broth), must be recognized as safe for use in food or food packaging. The German BgVV issues recommendations of types of materials and the allowed migration, five categories are declared. Only quite safe materials are admitted.

The FDA standard relating to elastomers is the FDA Code of Federal Regulations (CFR), Chapter 1, Title 21, part 177 (ref. 21 CFR 177.2600 in section 2600 "Rubber articles intended for repeated use" [64]).

An alternative is the USP 23 approval for elastomers. The applicable standard in this case is laid down in section <88> Biological Reactivity Tests in Vivo Plastic Classes I to VI. This standard relates to results of animal testing by exposure to elastomers or plastics by injections of extracts or implants from test material. For sanitary process application elastomers must meet the specification of class V or VI [65].

Several synthetic polymers are recommended [35], [66] and widespread in biotechnology because of their excellent temperature, acid and alkali resistance. Their main advantages and disadvantages are shown in Table 14.

EPDM has versatile qualities and is therefore widely used for O-rings, diaphragms of valves and gaskets. The material is appropriate to resist to flows of hot water and steam. When PFTE is used an increased resistance to chemicals is required. The disadvantages of PFTE are that

Table 14. Characteristics of synthetic polymers used in reactors

Synthetic polymers	Ethylene-propylene-diene rubber	Fluoroelastomer	Silicone or fluorosilicone elastomer	Fluoropolymer
ASTM code	EPDM	FPM, FKM	Q, MQ, VMQ, FVMQ	PTFE
Registered trade name, e.g.	Nordel, Buna AP, Keltan	Viton, Tecnoflon, Fluorel	Silopren, SE, Blensil, Silastic	Teflon Fluon, Hostaflon
Temp. min	− 40 °C	− 20 °C	− 55 °C	− 25 °C
Max. dry temp.	135 °C	210 °C	200 °C	260 °C
Max. steam temp.	140 °C	120 °C	125 °C	150 °C
Resistance to conc. acids	reasonable	very good	good	very good
Resistance to dilute acids	good	excellent	very good	very good
Resistance to alkali	very good	not recommended	reasonable	very good
Resistance to steam	good	poor	poor	very good

it cannot be stretched, is hard and has cold-flow properties under pressure. Higher temperatures dramatically increase this cold flow behavior. PFTE is used to coat EPDM diaphragms of valves, when these have to resist specified chemicals.

4.3. Other Materials

4.3.1. Glass (→ Glass)

Because of its excellent temperature and chemical resistance and because it enables transparent sight into the reaction chamber, glass is a frequently used material for reactor vessels of volumes below 15 L.

A disadvantage of glass is its brittleness. When bouncing glass, this may break or even worse have micro-fissures, which are not directly discovered. This may lead to accidents during sterilization.

To obtain smooth glass surface (sealing purposes) glass must be polished. The design parameters for pressure glass vessels are summarized in [75].

Glass vessels are used for viewing glasses built into stainless steel vessels.

4.3.2. Grease and Lubricants

Lubricants are used for O-rings and gaskets, to prevent their damage during installation and prolong service life. Lubricants have incidental contact with the product and are subject to regulations such as USDA H1. A positive list of allowed substances and their limitations in the product is given under "Indirect Food Additives: Adjuvant, production aids, and sanitizers" of the CFR (21CFR178.3570) [64].

5. System Components and Detailed Engineering (Fig. 23)

Cleaning and asepsis are the two most important points of consideration when designing components for bioreactors and fermentors. Good design practice follows several cardinal rules:

1) All parts must be self-draining, puddles must be avoided or minimized.
2) Dead-ends, resulting in poor heat transfer and air pockets must be avoided or minimized.
3) Corners, crevices, and edges (screws) must be avoided. The design principle of maximizing the crevice should be applied to enhance cleaning (see Fig. 24 A).
4) "Noncontrolled" closed spaces (e.g., in between two seals, see Fig. 24 B) must be avoided.

The design must consider that organisms have to be killed during sterilization and kept outside the cultivation system to avoid contamination.

Higher design criteria are necessary, when the release of organisms or products to the environment must be avoided. In the case of pathogenic organisms (see Section 2.7) the contact to operators and the direct surroundings of the reactor must be avoided. These additional requirements have a considerable impact on the design of components.

Figure 23. Components for a bioreactor

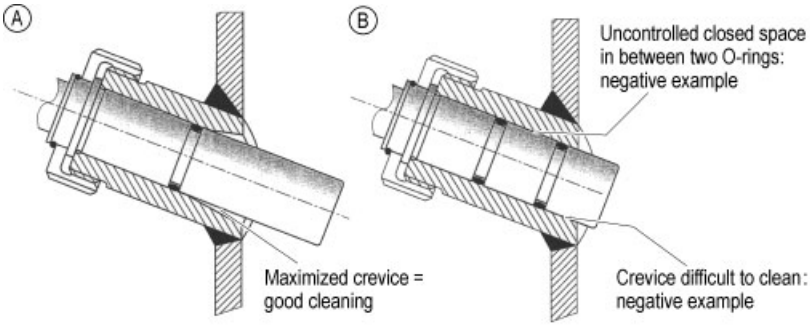

Figure 24. Inclined 25 mm port (Ingold type)
A) Principle of maximizing unavoidable crevices. B) Negative example of a port with crevice difficult to clean, double O-ring creating an uncontrolled space

5.1. The Vessel

A reactor system is often composed of several vessels. The core is the cultivation vessel, which might be surrounded by acid, alkali, antifoam agent, fresh media, storage, or harvest tanks.

Up to a size of 1500 L, stainless steel vessels are usually built with a removable lid sealed by an axial positioned O-Ring. For vessels above this volume the costs to manufacture. Such large sealing surfaces dramatically increase. Therefore non-removable dished tops with a manway are designed [24].

A vessel and its jacket (if any) are designed to withstand a certain pressure. On an inscription plate the maximal allowed pressure and the volume of the tank respectively the jacket must be indicated. Further the manufacturer, the year of construction and a serial number must be indicated. Steam sterilized vessels must be inspected according to the national pressure vessel regulation (e.g., ASME, PED 97/23, BPVC).

5.1.1. Components Mounted or Welded on the Vessel

5.1.1.1. Seals

Sealing a reactor correctly according to the hygienic and sterile design rules is one of the main challenges. Seals can be divided in two groups: *static seals* which are stressed statically and *dynamic seals* (see Section 5.2.2.1), which have a sliding contact to another surface.

Wherever removable components are mounted static seals are required. Static seals

are divided in two subgroups: *round shaped* seals like O-rings, which seal along a line and *flat seals* (gaskets, tri-clamp type seals), which seal a surface. Common materials are EPDM or fluoroelastomers such as Viton.

O-Rings. Using O-rings is the state of the art. The design of grooves and its tolerances for the O-rings are described in DIN 3771 or ISO 3601/I.

O-rings must be elastic and incompressible. The dimension of the O-ring groove should be designed to enable the O-ring contact with all four surfaces during sealing to minimize the space where dirt or broth residues could accumulate or remain. On the other hand, to prevent damage of the incompressible O-ring, the groove should leave enough space for the O-ring to react to compression. The swallow-tail groove, often used for O-rings sealing the lid minimizes the "dirty" space (Fig. 25). Large diameters (> 200 mm) such as those for manways and lids are sealed with an axial O-ring. Radial O-rings are preferred for smaller diameters (< 200 mm) to avoid dead-ends (Fig. 26). The O-ring should always be positioned closed to the end to minimize the space where "dirt" could accumulate.

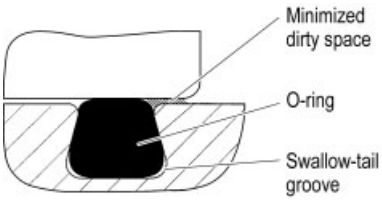

Figure 25. Swallow-tail groove with compressed O-ring, the large groove allows the expansion of the compressed O-ring

High containment design is sometimes performed by designing seals with two O-rings being steam sterilized in between. However in practice the lifetime of O-rings often exposed to steam is dramatically reduced. When considering the additional pipework and the complexity of the system it is questionable whether the system will really be safer [66].

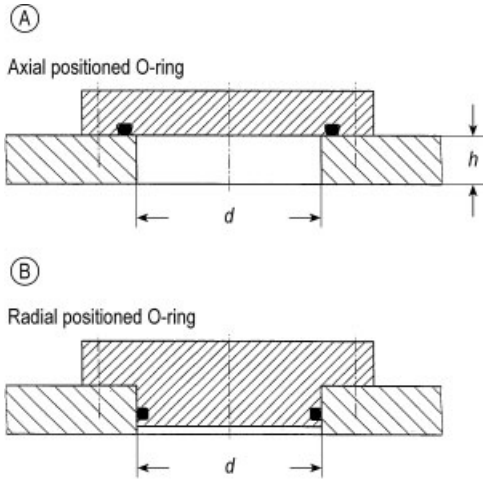

Figure 26. Axial and radial position of an O-ring
A) If $h > d$ there is an increased risk of dead-end. Axial positioned O-rings will therefore be used for large diameter ($d > 200$ mm) B) Radial positioned O-rings require a high precision of the sealing parts, which can only be achieved at reasonable costs for small diameter ($d < 200$ mm)

Gaskets and Tri-Clamps. The sealing surface of a gasket is large compared to an O-Ring. This surface must be clean and even to properly seal. To reach a comparable pressure on the sealing surface, the force applied on a gasket, compared to an O-ring, must be higher. Despite these disadvantages several application require a flat sealing:

1) Noncircular openings such as longitudinal viewing glasses.
2) The sealing of stainless steel with another brittle material such as bursting disks or glass. Gaskets avoid strain peaks compared to O-rings.
3) Depth filters, which must be sealed over a surface.
4) Tri-clamp connections for pipes are widespread in North America and accepted by the FDA. While adepts of tri-clamps emphasize the easy handling (no tools required) others argue that when the pipe is pressurized the operator has no chance to feel the danger while opening the connection, further three hands are required to open and close a tri-clamp.

5.1.1.2. Ports

Ports on vessels are required for probes and transfer lines. In Europe more or less standardized ports with inner diameter of 12, 19, 25 (Ingold type) and 32 mm are in use. Stanardization encompasses dimension and tolerance of the inner diameter. Probe manufacturer can therefore supply probes with an adaptor sealing with a radial O-ring (Fig. 24). However different lengths of the Ingold type port are in use and make it therefore difficult to place the O-ring on the optimal location [67]. Lateral ports on the vessel can be mounted horizontally since an O-ring is placed next to the vessel wall (Figs. 44 and 45). Probes containing electrolyte (e.g., pH, pO_2) need to be installed in inclined ports (usually 15 °, Fig. 24).

In North America a different standardization is employed. Not the diameter of the port is standardized but the end, which should always be a tri-clamp. The seal is located on the tri-clamp and not next to the vessel wall, this situation always creates a small dead-leg, which is difficult to clean. Lateral ports need to be slightly inclined to allow draining of the small dead-leg.

5.1.1.3. Viewing Glasses

As a matter of fact almost all stainless steel reactors are equipped with lamps and viewing glasses, although the opinions about viewing glasses on the lid or the vessel wall are controversial. A first group estimating that every gasket/O-ring is a source of risk will not agree to install a viewing glass. Further it is argued that longitudinal viewing glasses on the vessel wall interfere with the liquid circulation in the jacket, reduce the surface of heat transfer and the strength of the pressure vessel. A second group would not like to miss the quick look into the vessel to control the broth/liquid level and the actual condition of foam.

The viewing glass flange should be flush with the internal wall of the vessel and have rounded edges to promote drainage. Glass flanges are usually sealed with gaskets. Construction details such as frames, size and geometry are defined in the national pressure vessel guidelines (e.g., DIN 7080, 7081, 28210, 28121).

During cultivation foam residues and condensate accumulates on lid viewing glasses. For a transparent sight into the reactor the glass must be cleaned. There are two cleaning systems available: a manually operated wiper blade or a steam jet (Fig. 27). The disadvantage of the wiper blade is that it is designed with a dynamic O-ring.

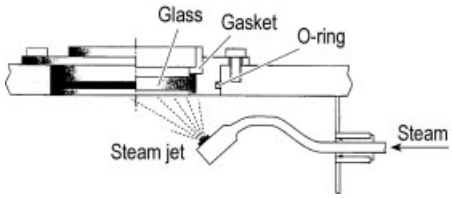

Figure 27. Round viewing glass with steam jet cleaning unit

5.1.1.4. Manways

Vessels with a volume above 1500 L seldom have removable lids and require manways for cleaning and servicing the inside of the vessel. Baffles, impellers, and the shaft are also introduced or removed through the manway.

The manway is closed with a lid and sealed with an O-ring on a flange. Nuts or kind of butterfly nuts are used to lock the manway. Usually a hinge is designed to open the lid of the manway. The design of the manway must avoid dead spaces and consider that all surfaces are accessible for cleaning fluids (CIP) but drain readily.

5.1.1.5. Heating/Cooling Jackets

Jackets are responsible for heat transfer in and out of the vessel.

Three general concepts predominate in the layout: the double jacket, the full pipe, and the half pipe (see Table 15). Double jackets are the most common solutions for vessels up to about 3 m^3. A tangential inlet and outlet on the jacket should be designed to create a fluid circulation around the vessel. This allows the total surface of the jacket to be used for heat transfer. However, ports and longitudinal viewing glasses interfere with this circulation (inhomogeneous conditions and pressure drop). Jacketing the bottom area of the tank is difficult to manufacture, but advantageous when small working volumes are used or the tank is poorly mixed. It is recommended

to jacket up to the unaerated maximal working volume (about 2/3 of the total volume). This choice is a tightrope walk between lower jackets, which prevent baking of media during sterilization of low working volumes, and higher jackets, which increase the surface area of the jacket and thereby improve the heat transfer of aerated full working volumes.

5.1.1.6. Heating/Cooling Coils

Large fermentors or those with extremely high power input per volume, high *OTR*, and thereby strong exothermic reactions must remove large amounts of heat and are therefore equipped with internal cooling coils. A *spiral coil* is more effective if an axial flow pattern is created. Tube bundles arranged *vertically* along the vessel wall are preferred for stirred reactors with Rushton impellers and replace baffles. Vertical coils are also preferred to spiral coils since they have less horizontal surfaces, on which residues might remain, when the vessel is drained.

5.1.1.7. Pressure Relief and Safety Devices

Vessels and jackets resist a certain pressure. Pressure may rise above the limit, if the temperature control of the vessel or the pressure reducing valve of the inlet gas is defect. For safety reasons vessels are therefore equipped with a pressure relief valve or a bursting disk. The advantage of a pressure relief valve is that it closes after a release as soon the pressure drops below the maximal allowed pressure. Such an event would not contaminate an aseptic vessel. Usually such valves are used for small vessels of up to 100 L.

Vessels above 100 L are usually equipped with bursting disks, made of carbon or synthetic polymers. The disadvantage of a bursting disk is the fact that it is broken after an overpressure. Furthermore it should be exchanged on a regular base.

Vent lines from a pressure relief valve or a bursting disk are directed to the floor or connected to a kill tank (contained drain). Safety devices are not required on the vessel if all lines providing pressure to the vessel are equipped with safety devices rated below or at the maximal pressure of vessel.

5.1.2. Standard Stirred Tank Reactor

About 95 % of the reactors are stirred tank reactors, which are available from laboratory scale (2 L) up to 250 m^3. Often encountered volume sizes due to standardized components are: 30, 42, 300, 500, 750, 1500, 3000, 5000, 7500, 15 000, and 30 000 L [35], [66]. As a rule of thumb about 2/3 of the total vessel volume is the maximal unaerated working volume of a stirred tank. The liquid broth volume depends mainly on the aeration rate (gas hold-up) and the foam development during the process. The minimal working volume is defined by the location of the necessary sensors and the lowest impeller. It requires a sufficient heating/cooling surface area of the jacket to ensure a proper temperature control.

Baffles, absolutely necessary in reactors stirred with Rushton type impellers, improve mixing by eliminating the vortex and promoting turbulence. Usually four baffles are either permanently welded to the vessel or designed as removable units [28] (→ Stirring, Chap. 1.2.). Storage tanks or bioreactors for animal cells requiring low mixing can do without baffles. Impellers generating an axial flow pattern (e.g., marine impeller) situated off center and/or at an angle to the axis are used here. Top and bottom drives are available.

Figure 28 illustrates the most important vessel dimensions and components. Several dimensions are related to each other and rules of thumb for ratios can be formulated for the most important dimensions.

Especially the H/D ratio is an often discussed parameter. Reactors with a high H/D ratio (≈ 3) generally have the following advantages over lower ratios:

- Higher surface/volume ratio resulting in better heat transfer via jackets.
- Higher hydrostatic pressure at the sparger, longer bubble residence time in the reactor resulting in higher *OTR*.

On the other hand reactors with a low H/D ratio (≈ 2) feature also some advantages:

- Lower overall height (fit better into rooms with low ceiling).

Table 15. Design of Jackets

Type	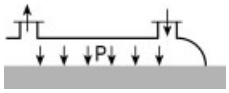 Double jacket	Full pipe	Half pipe
Advantages	– Max. contact surface – Min. pressure drop – Few welding seams	– Totally independent from the vessel, no pressure on the vessel wall	– Good contact surface
Disadvantages	– High pressure on the vessel, vessel wall has to be reinforced-reduced energy transfer	– High pressure drop – Min. contact surface – Many welding seams	– High pressure drop – Many welding seams

- Higher reactor cross-section-to-volume ratio enables better surface gas-liquid mass transfers when gassing the headspace (e.g., bubble-free gassing of O_2, CO_2 for animal cell culture).
- Better axial mixing (only of importance at large scale $> 3\,m^3$).
- Shorter shafts, with usually only two impellers, having better mechanical stability (only of importance at large scale $> 3\,m^3$).

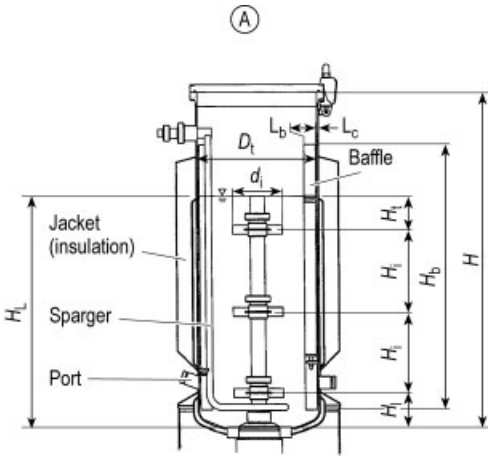

Figure 28. Standard stirred tank reactor, schematic drawing with ratios of main dimensions
$H/D_t = 2-3$; $d_i/D_t = 0.3-0.5$; $H_l/d_i = 0.3-0.5$; $H_i/d_i = 1-2$; 4 baffles $L_b/D_t = 0.08-0.1$; $H_b \simeq \times H$; $L_c/D_t = 0.02$

Reactors requiring a high *OTR* and an efficient heat transfer (e.g., aerobic cultivation of prokaryotic cells and fungi) will often have a H/D ratio of around 3, on the other hand reactors requiring homogeneous mixing at low shear forces (e.g., animal cell culture) will have ratios of about 2. However, the importance of the dimensions must not be overestimated, most processes will run equally with both extremities of dimensions. Therefore often the available space in the building will define the overall dimensions of the reactor.

5.1.3. Standard Airlift Reactor

Good mixing, an efficient *OTR*, low shear forces and a relative simple design (no agitation shaft to be inserted in the vessel) are the main advantages of an airlift reactor (→ Bubble Columns, Chap. 2.14.). However, the high *OTR* values reached in a STR are not achieved.

The height to diameter ratio of an airlift reactor varies between 5 and 10 [34], [35]. Compared to STR (H/D ratio 2 to 3), jackets along the vessel cylinder of airlift reactors (Fig. 29) have therefore much better heat transfer rates.

Airlift reactors are available from laboratory to very large production scale. Due to their relatively simple design these reactors are well suited for very large applications (100 to $500\,m^3$). Standard total volumes of airlift reactors are 16.8, 26, 50 and 1500 L, resulting in a working volume of 12, 21.5, 30 and 1200 L, respectively [35].

Gas distribution is achieved by perforated pipes, sintered bodies, hydrophobic tubes or static mixers (→ Stirring). The *OTR* depends on the aeration flow-rate, the reactor height, and

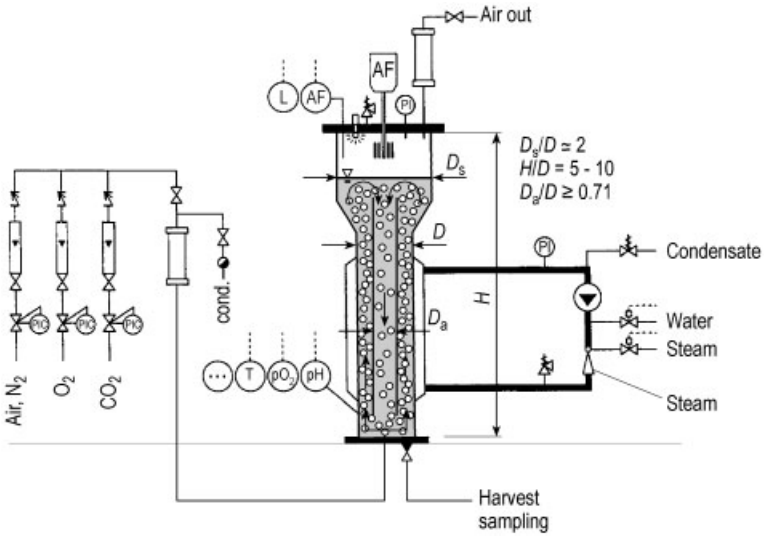

Figure 29. Standard airlift reactor with draft tube

the oxygen concentration difference between the gas and liquid phase (see Section 2.5.4). As shown in Figure 29 airlift reactors often have larger diameters on the top end to act as gas separators.

5.1.4. Glass Vessels

Glass vessels are used up to total volumes of around 15 L. Glass is appreciated in R&D for an overall view inside the reactor. When cultivating phototrophic organisms light is transmitted through the glass.

Glass vessels are often *sterilized in an autoclave* and designed with a jacket for heating and cooling (Fig. 30). All transfer lines with bottles (antifoam agent, alkali solution, etc.) can be sterilized together with the glass reactor. The components of such systems are therefore quite simple and are easy to maintain. On the other hand transport and handling of the fragile glass vessel into the autoclave and back to the laboratory bench might be delicate. Long heat-up times during sterilization in the autoclave, due to inhomogeneous conditions, must be accepted. Heat-up times are improved when the jacket of the vessel is filled with water.

In-situ sterilizable (SIP) glass vessels consist usually of a glass cylinder, which is closed at the bottom and the top by stainless steel plates. The construction must absolutely avoid strain peaks on the glass. Heat transfer is achieved by jacketing the bottom of the reactor or by introducing heat exchangers (electrical heater, cooling coils) into the reactor. The cylinder must be handled with care and inspected before each sterilization. The explosion of a damaged glass vessel filled with liquid at 121 °C at 1 bar overpressure represents a major risk in the laboratory. To protect the surroundings it is necessary to position a cylindrical metal shield around the glass vessel during sterilization. This shield must withstand a possible explosion of the glass reactor. Tests have shown that slightly conical cylinders allowing the emission of vapor and glass splitter into a vertical direction are best suited.

5.1.5. Special Polyamide Film Sleeve Bioreactors

A safe alternative to the in-situ sterilizable glass vessel is the in-situ sterilizable STR made of a special polyamide film sleeve. The film sleeve is supported by a stainless steel base and lid (Fig. 31). Metal shells are used to provide support for the film during sterilization (overpressure!) and are removed afterwards. This construction principle is the only feasible and safe solution for larger in-situ sterilizable reactors (15 to 100 L), which need external illumination

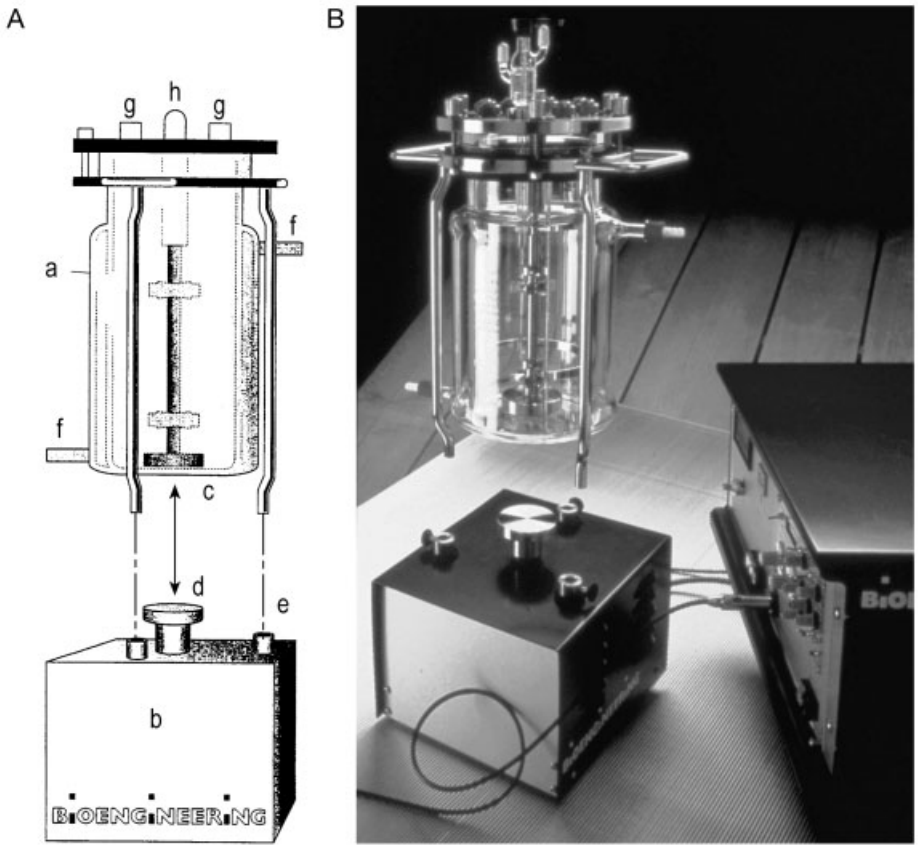

Figure 30. Autoclavable stirred tank reactor with glass jacket (5 L); stirrer with axial magnetic coupling, bottom driven
a) Glass vessel with a glass jacket; b) Bottom drive; c) Upper magnetic coupling; d) Lower magnetic coupling; e) Plug-in support; f) Temperature control fluid; g) Inlets/outlets and probes; h) Stirrer shaft with magnetic coupling

(phototrophic culture). Long airlift reactors were designed with a polyamide film sleeve (Fig. 32). The construction is limited by the strength of the plastic film sleeve exposed to the hydrostatic pressure of the liquid. Main design parameters are reported in [55].

5.2. Agitation System Design

5.2.1. Top or Bottom Drive?

The advantages of a *bottom driven* shaft are:

- A shorter length of the shaft (less vibration, reduced design dimensions, more stability)
- Space on the top head of the reactor for transfer connections and sensors (especially laboratory reactors)
- Lubrication of the primary mechanical seal or of the bushing type bearing of magnetic coupled shafts by the medium/broth

The advantages of a *top driven* shaft are:

- Abrasive media can be cultivated without damage of the bearing seal
- No media leakage in case of a failure of the mechanical seal

Reactors equipped with a bottom drive are therefore suitable for most applications.

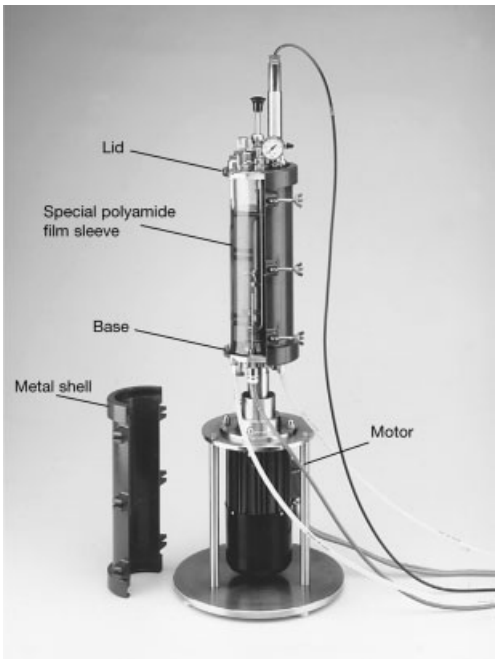

Figure 31. Visual safety bioreactor, 2.4 L total volume. In-situ sterilizable stirred tank reactor designed with a special polyamide film sleeve supported by the stainless steel base and lid, metal shells are used to provide support for the film during sterilization and are removed afterwards.

pathogenic organisms), the magnetic coupling should be chosen.

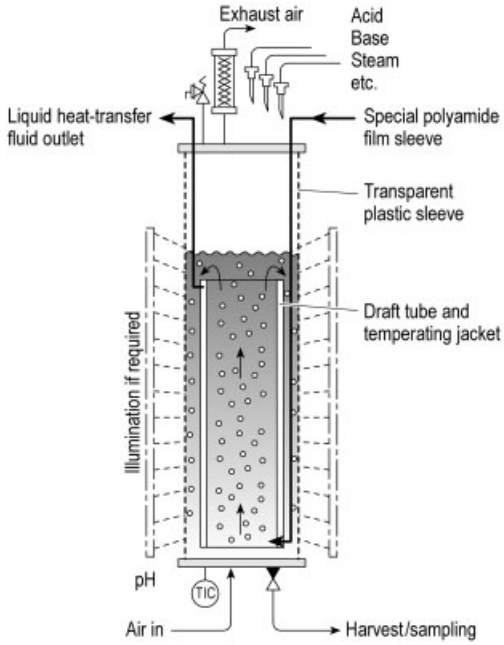

Figure 32. Visual safety fermentor; the outer cylinder of the airlift reactor is a transparent plastic-film sleeve; the heat exchanger is integrated in the draft tube

5.2.2. Mechanical Seals versus Magnetic Coupling

The difficulty consists in introducing a rotating shaft into a "closed system".

When using a mechanical seal, the rotating shaft is held by bearings located outside the reactor and the mechanical seal seals the rotating shaft against the static vessel. The design allows good cleaning, it must, however tolerate minor leakage of broth in case of failure of the mechanical seal.

When using a magnetic coupling the vessel can be considered closed. Torque is transmitted through the reactor wall with magnets. The disadvantage is that the shaft is held inside the vessel by bushing type bearings, which are difficult to clean.

If the most important requirement to the reactor is very high torque transmission and/or efficient cleaning (CIP) the mechanical seal should be used. If containment is most important (use of

5.2.2.1. Shafts with a Mechanical Seal

The mechanical seal consists of the static seat ring on the vessel and the rotating ring on the shaft. The polished surfaces of both rings must be perfectly parallel to slide one on each other. This is achieved by positioning one of the rings on springs to form a floating bearing (Fig. 33). The springs are compressed to produce a defined force necessary to seal both parts.

Reactors often have a double mechanical seal. The primary mechanical seal is in contact with the medium, the secondary with the environment (Fig. 33) [32]. The materials mostly used in biotechnology are carbon for the rotating ring and ceramics for the seat ring. More resistant but also more expensive materials are silicon or tungsten carbide (static and rotating ring same material).

Lubrication of the seal is very important. The primary mechanical seal of a bottom drive is automatically lubricated by the medium/broth. A top drive or a secondary mechanical seal requires

an external lubricant. Either steam condensate or glycerol is used.

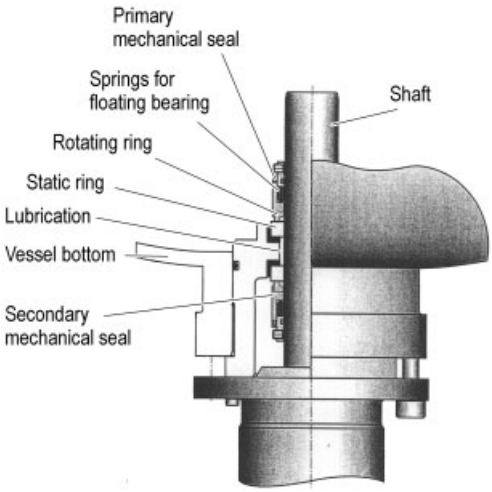

Figure 33. Double mechanical seal of a bottom drive

When using *steam condensate* as a lubricant, the space in between the two mechanical seals can be sterilized. For sterilization, the condensate is evacuated and the space in between the two mechanical seals steamed (Fig. 34). On the end of the sterilization, to again create condensate, cold water is circulated through a heat exchanger in the condensate reservoir. In cultivation mode, the condensate is pressurized by sterile air at a pressure above the pressure in the tank. A leakage from a defect primary or secondary seal would therefore result in a level decrease in the condensate reservoir. By mounting a float switch in the reservoir an alarm can be send out. Broth will not leak out of the reactor.

When using *glycerol* the space in between the mechanical seals is not sterilized. The design thus requires far less valves and less maintenance. The system is reliable since glycerol is a better lubricant than steam condensate. Glycerol is therefore an excellent alternative for processes where a small leakage of broth is tolerated [68].

Steam condensate as well as glycerol are used for top or bottom drives.

Shafts mounted with external bearings are called overhung shafts. When reaching a critical length shafts need not be supported by a steady bearing inside the reactor. Bottom drives have shorter shafts than top drives, and do not need a steady bearing up to volumes of about 3500 L. Larger reactors will require a steady bearing, usually a bushing type bearing, to hold the end of the shaft. Materials such as synthetic coals, met-

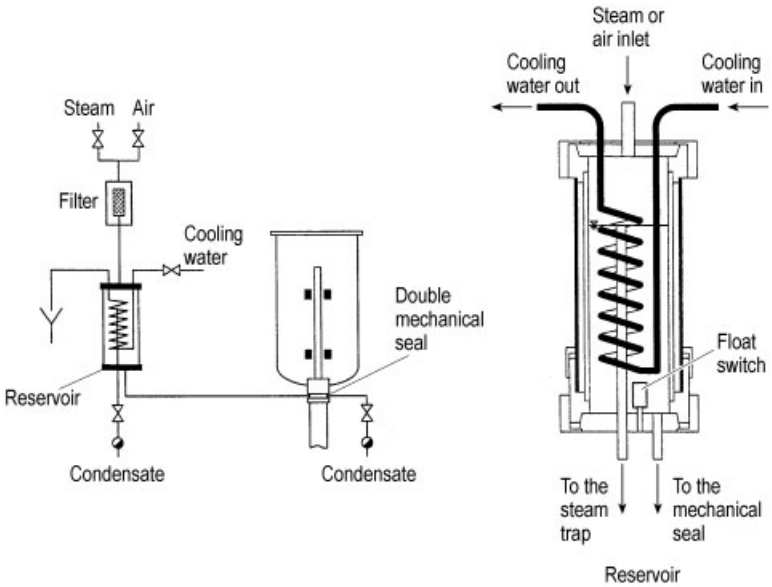

Figure 34. Lubrication of a double mechanical seal with steam condensate. Steam condensate is created in a reservoir connected to the mechanical seal. During operation the reservoir and therefore the lubricating condensate is pressurized with sterile air. A float or level switch in the reservoir detects a leak of the lubricant.

als, or metal carbides such as silicon and tungsten carbide are used. To ensure lubrication and allow cleaning, the bushing often has a groove through which the lubricant (medium/broth) can flow.

5.2.2.2. Magnetic Coupled Shafts

A magnetic coupling provides torque transmission from the motor shaft to the agitator shaft. The magnetic torque is transmitted through the reactor wall. It is therefore the ideal technique when complete containment is required. Generally it is differentiated between axial (Fig. 30) and radial torque transmission (Fig. 35). Radial couplings can transmit a higher torque than axial couplings. The torque transmission is not the limiting factor up to vessel volumes of 750 L. The stability of the shaft is, however, not so easy to achieve, because the shaft is hollow and mounted on an arbor by bushind type bearings. The stability is given by the arbor, which has a smaller diameter. The stability compared to a shaft directly introduced into a reactor (mechanical seal) is therefore reduced.

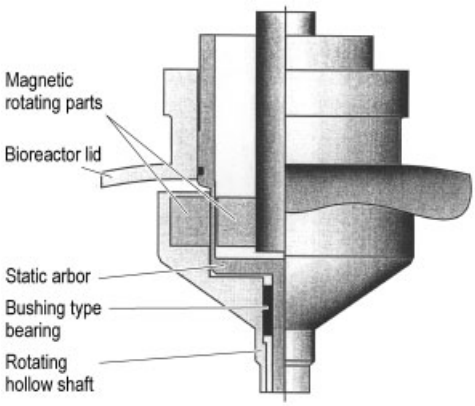

Figure 35. Radial magnetic torque transmission, bushing type bearing of the shaft in contact with the product

The bushing type bearings require lubrication and cleaning. To meet this requirement holes are drilled on the top and bottom of the shaft. Broth and cleaning solution can thereby penetrate into the narrow gaps.

5.3. Transfer Lines

To inoculate, feed, supply or withdraw a liquid from or to a reactor, sterile transfer lines are required. These lines must be cleanable and sterilizable. Depending on the containment class, the size, the length and the slope of the line, the design can considerably differ.

5.3.1. Needle Connection

A bottle can be connected via a flexible hose to a sterile reactor with a needle connection (Fig. 36). The bottle, the hose, and the needle are protected from the nonsterile environment by a tube with filter and are sterilized in an autoclave. By a burning few drops of ethanol on the septum the surface of the septum is sterilized. The reactor is sterile and needs to be pressureless when connecting the needle to it. The needle is taken off from its protecting tube and pierced through the flame and septum. These types of transfer lines are widely used on a laboratory scale.

5.3.2. Steam-Sterilizable Transfer Line

Figure 37 shows a transfer line from the harvest valve of vessel 1 to a sterile cross connection on the top of vessel 2. The harvest valve is mounted flush to the bottom of an in-situ sterilizable tank; the steam line is mounted next to the sealing. The sterile cross on the end of the transfer line, next to vessel 2, is connected to a steam trap. The transfer-line can be sterilized by supplying steam to the harvest valve and draining condensate through the sterile cross, while the two vessels remain in operation. The configuration shown in Figure 37 is only valid for transfer lines of small nominal diameters ($<$ DN 10). In pipes of low diameters condensate accumulating at the lowest point of the line is chased by the steam flow. Larger lines require a steam trap at the lowest point of the transfer-line in order to drain condensate.

5.3.3. Steam-Sterilizable Transfer Line with CIP

Figure 38 shows a steam-sterilizable transfer line, which can be cleaned in place. The CIP supply line, connected to the sterile cross of vessel 2, provides cleaning solution or WFI. The connection is used to either clean the transfer line or the

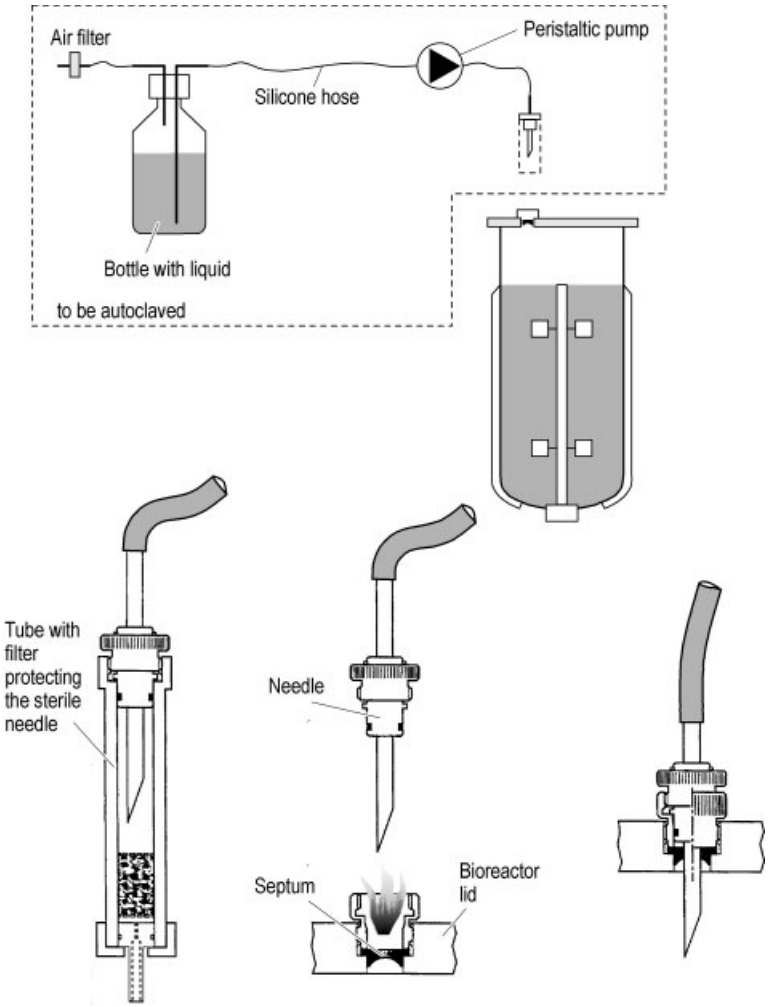

Figure 36. Transfer line with needle connection. Separate sterilization of the needle, bottle, and the flexible hose in the autoclave.

nozzle of vessel 2. On the other hand the CIP return line is used to drain the cleaning solutions, when the transfer line or vessel 1 is cleaned. Instead of using steam traps as shown in Figure 37, pneumatic actuated pulsed valves with nozzles are used to drain the condensate. When briefly opening, those valves drain the accumulating condensate. The pulsation is controlled by the programmable logic controller (PLC).

5.3.4. Manually Operated Transfer Panel

The previous example show that a large number of valves are required to design a sterilizable and cleanable transfer line. When multiple transfer-lines in between several tanks are needed, the number of valves increases tremendously. In order to reduce the number of valves so-called transfer panels manually operated with pivoting elbows, bends, or connecting pipes can be designed (Fig. 39). Thus the number of valves is strongly reduced, but skilled operators, are required to connect the elbows. Manually operated transfer panels are open systems, since operators might be in contact with the product.

5.3.5. Ring Transfer Panel

Ring transfer panels have similar functions as manual transfer panels but have the additional advantage of being a closed system. Special T- and corner-type diaphragm valves installed in a ring configuration allow the transfer of any inlet to any outlet valve (Fig. 40).

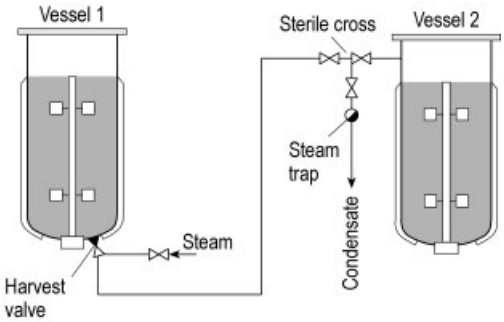

Figure 37. Steam sterilizable transfer line (SIP) for tubes or hoses withstanding 2 bar pressure and diameters below DN 10

5.4. Valves

Valves used in equipment for sterile and hygiene application, must constitute a sterile barrier when closed. Valves should have a minimal number of joints and connections to reduce the maintenance costs and downtime of a plant. Parts requiring maintenance (e.g., actuators, diaphragms) must be removable.

Ball, gate, butterfly, and piston valves have cavities, which restrict their use for hygienic/sterile applications. Thus, diaphragm valves and bellows type valves are widely used valves in biotechnology [35], [39], [69], [70].

5.4.1. Diaphragm Valves

Diaphragm valves are best suited for sterile/hygienic applications in stainless steel pipework. The main advantages of a diaphragm valve are:

1) Pocketless and cavity-free design
2) Good draining in horizontal pipes
3) In- and outlet of the diaphragm valve is in-line with the pipe

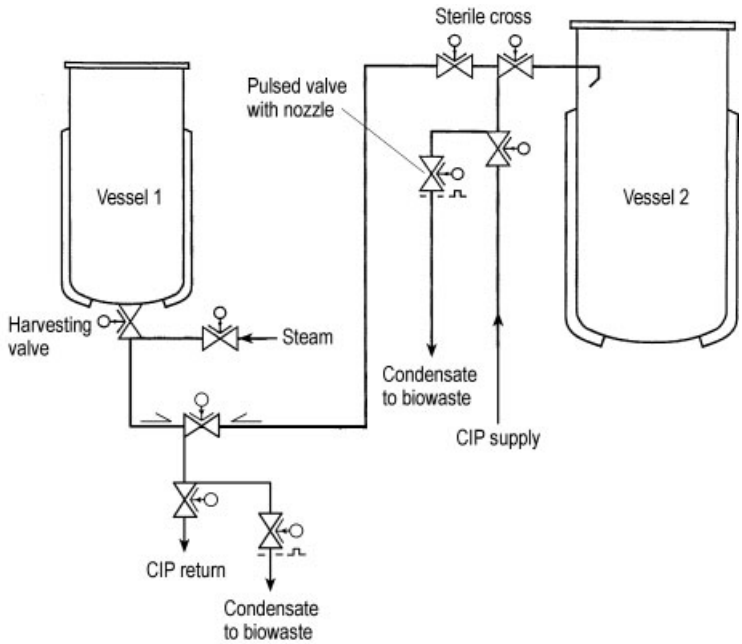

Figure 38. Steam sterilizable transfer line with CIP supply and return line to clean the transfer. Pulsed valves with nozzles are used instead of steam traps to drain the condensate.

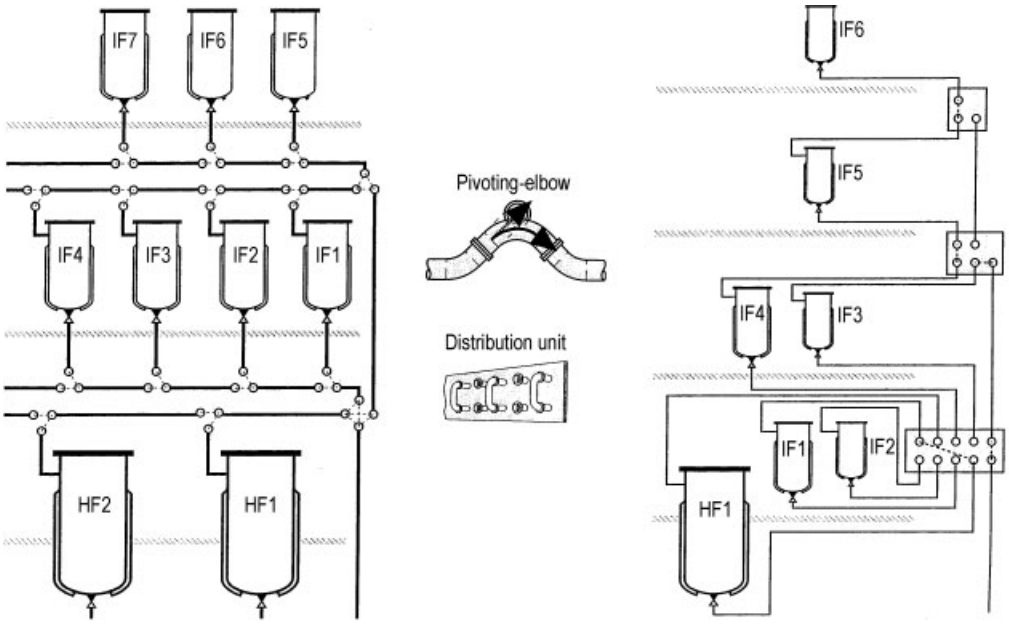

Figure 39. Transfer panel to connect several tanks to each other via pivoting elbows or bent pipes.

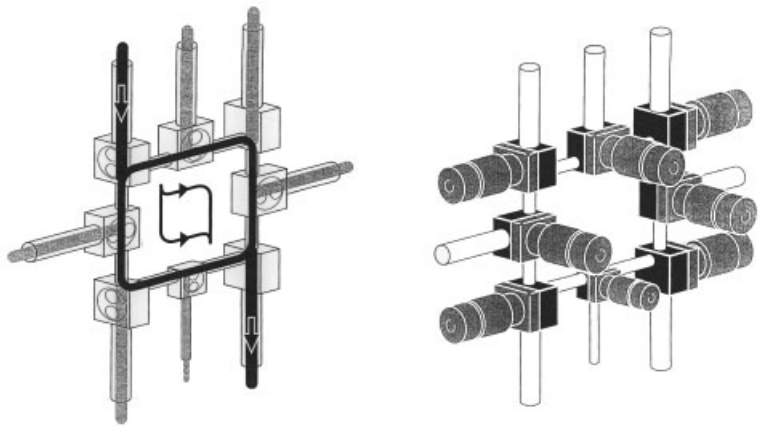

Figure 40. Ring transfer panel with SIP and CIP. Special T- and corner-type diaphragm valves allow a ring design without dead-spaces. Any inlet can be connected to any outlet.

4) Maintenance can be performed on build-in valve bodies, therefore valve bodies can be welded in the pipework
5) Manual and powered actuators (electrical or pneumatic) are available

A diaphragm valve consists of a valve body, the diaphragm, and the actuator (Fig. 41). Complete closure of the valve is achieved by contact of the elastic diaphragm with the metal weir of the valve body.

Beside this classical diaphragm valve there are models with a built-in drain line connection adjacent to the weir (Fig. 42). Combinations to form sterile crosses, and rings are also available. To ensure free draining of horizontal pipes, diaphragm valves need to be positioned in a lateral or inclined position.

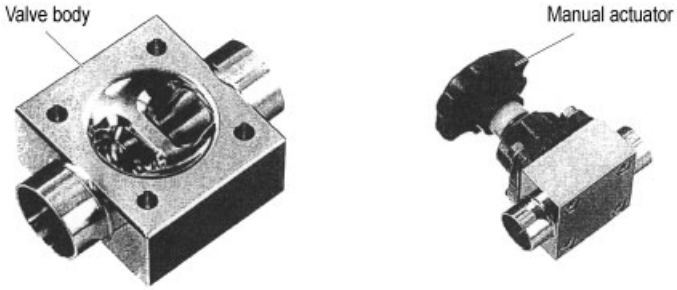

Figure 41. Diaphragm valve, valve body and body with manual actuator

5.4.1.1. The Valve Body

Valve bodies can be cast, forged, welded, or be machined out of bar material. In the past most valve bodies were cast. The surface of such bodies is of poor quality and a big effort is required to reach an acceptable roughness (by welding, polishing). Further the ferrite content is high and cavities are possible. Forged bodies, where cavities can be excluded and the amount of ferrite be reduced below 1 %, are therefore preferred [35]. Excellent quality and flexibility is reached when bar material is machined. The quality of stainless steel (e.g., 316L) is chosen and the pipe connections are milled according to the specification of each customer. Surface roughness of $R_a \leq 0.8\,\mu m$ is standard, but even higher demands down to $R_a \leq 0.25\,\mu m$ can be reached. Bodies of bar material and those, which are forged, are usually available for pipe sizes in between DN 08 to DN 32. Tube sizes with higher nominal diameter, up to DN 150, require cast or welded (Fig. 43) valve bodies, which are remachined.

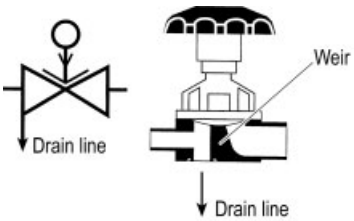

Figure 42. Diaphragm valve with drain line adjacent to the weir

Valve bodies of diaphragm valves offer the choice of all type of end connections (tri-clamp, butt-welded, flare-type fitting, etc.). Whenever possible, valves should be welded into the pipework to eliminate unnecessary joints. For automated orbital welding the butt tube of the body requires a minimal length, depending of the tube size (e.g., 11 mm for DN 10).

5.4.1.2. Diaphragms

The weakest point of a diaphragm valve is the diaphragm itself. It is a wear and tear part which has to be checked on a regular basis. Usually those diaphragms that are in direct contact with steam or abrasive media first begin to show deficiencies. In a plant diaphragms are therefore differently stressed and have different lifetimes. When specially stressed diaphragms are identified in a plant, these can be used as an indicator to define the time to change all the diaphragms of a plant.

Usually EPDM or EPDM covered with PTFE is used.

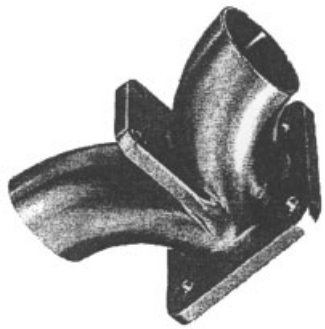

Figure 43. Picture and drawing of a welded diaphragm valve body (DN 100)

5.4.2. Bellows-Type Sealed Valves

Bellows-type sealed valves can be designed in order to seal flush with a tank wall. The closed valve can be sterilized separately from the tank.

Therefore this type of valves finds applications as harvesting (Fig. 44) and sampling valves (Fig. 45). Bellows-type valves find also application as control and process valves in pipework.

Bellows materials are stainless steel or PTFE. No dynamic O-rings are used in this type of design.

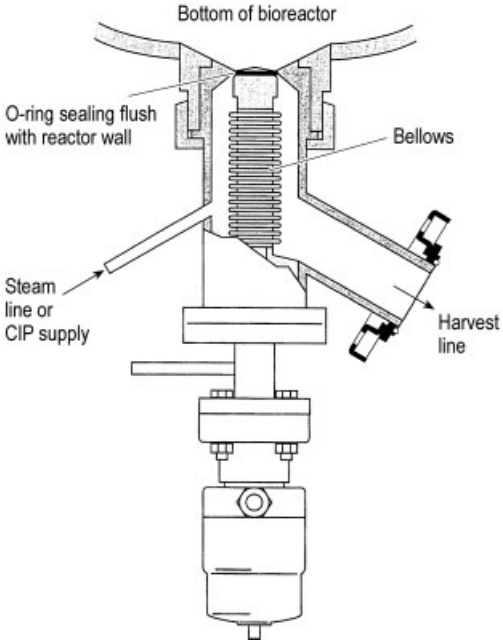

Figure 44. Pneumatic actuated harvest valve with bellows

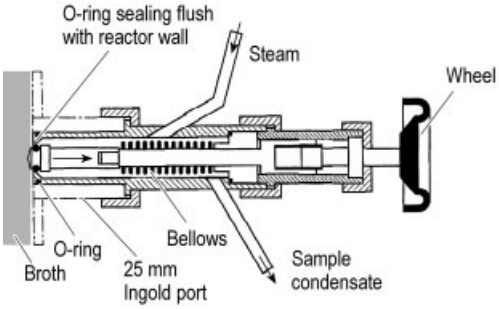

Figure 45. Flush mounted sampling valve with static O-rings and bellows in closed position. When turning the wheel the bellows contracts and the valve opens.

The weak point of this valve are the bellows. To control the integrity of the bellows a leakage pipe can be used as an indicator. Double bellows with internal pressure indicator are also available. Cleaning of the bellows might be difficult in special applications involving sticky media or mycelia-forming fungi.

Similar valve systems, for the same applications, replacing the bellows by a rubber joint are available.

5.4.3. Piston Valves

Piston valves always have a dynamic sealing, which requires special maintenance. Such seals (often O-rings or PTFE packing) should be inspected before usage. The piston surface sliding on the seal should be smooth, clean, and lubricated. The advantages of such a piston valve are the simple design and that the sealing can be designed flush with the tank wall. Piston valves are therefore used as sampling (Fig. 46), harvesting and injection valves. Piston type valves are the simpler and therefore cheaper alternative design to the bellows-type valves.

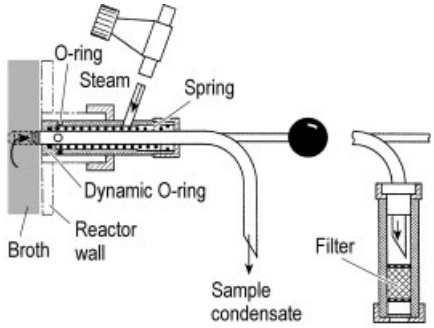

Figure 46. Flush mounted sampling valve with dynamic O-ring

A special application of a piston valve is found in an interchangeable probe armature (Fig. 47). The armature enables external probe maintenance during cultivation. The probe is sterilized before being reinserted.

5.4.4. Ball and Butterfly Valves

Ball and butterfly valves can be mounted in any position, when open they are always self-draining and have almost no pressure drop. To open and close the valves a swivel drive needs to be turned by 90 °. To a limited extent these valves can be used as control valves. Problematic is the packing of the ball or the sealing of the shaft. Cavities at these locations cannot be excluded, this limits their use for sterile and hygienic applications [35]. Since they are cheaper than diaphragm valves they are usually found in the nonsterile pipework of the plant (cooling water circuit, air line before sterile filter, etc.).

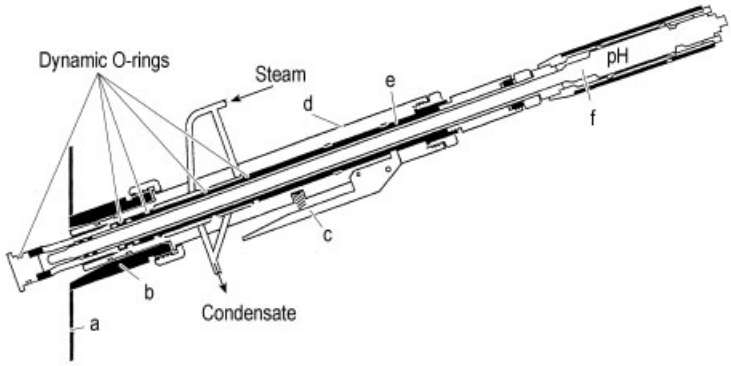

Figure 47. Interchangeable probe in contact with the medium (measurement position). When being retracted the probe is separated from the medium, the probe can be steam sterilized or dismounted to be recalibrated and inspected. Before being reinserted to the media the sensor is steam sterilized.
a) Bioreactor wall; b) Port 25 mm; c) Handle; d) Housing; e) Sliding part; f) Sensor

An adapted butterfly valve for sterile conditions has a steam-sterilizable dual shaft sealing. Such rather complicated units are gaining interest for lines larger than 100 mm in diameter [70].

5.4.5. The Actuators

Actuators of valves can be divided into two groups, those describing an axial and others describing a rotating movement. An axial movement is used to actuate diaphragm, bellows-type, and piston valves. The rotating movement is used to actuate ball, gate, and butterfly valves. Actuators are manually, electrically, or pneumatically powered. Generally the pneumatic actuators for axial movement are bigger than the valve body. The actuator of a butterfly valve has the advantage to be smaller than that of the diaphragm valve. This is a further reason why butterfly valves become interesting for lines larger than 100 mm in diameter.

Process, harvesting, and sampling valves have an open and a closed position. Control valves require a continuous movement of the valve to control a flow rate or a pressure.

Axial Actuators. Manual axial actuators for diaphragm and bellows-type valves are usually designed with an insert screw. Several turns of a hand wheel are necessary to open and close the valve. To not over-tighten the valve and thereby damage the seal the actuator should be mechanically blocked in the closed and open position. A visual indication showing the position of the actuator is advantageous. The insert screw of the actuator makes it possible to use the valve as a control valve.

Manual axial actuators for piston valves can be designed with a hand lever, which directly acts on the piston. A spring ensures that the valve is automatically in the closed position when not activated (Fig. 46).

Automated actuators for process valves are often pneumatically and seldom electrically powered. So called NO (normally open) or NC (normally closed) actuators are available. This indication defines the position of the actuator during pressure or power loss (Fig. 48). Used materials for pneumatic actuators are aluminum, synthetic polymers, or stainless steel. Pneumatic actuators for control valves also exist.

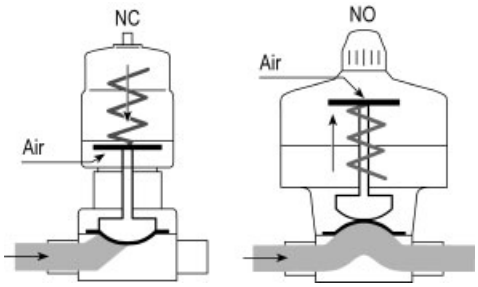

Figure 48. Normally closed (NC) and a normally open (NO) pneumatic actuators

Quarter Turn Actuators. The swivel drive of butterfly and ball valves requires to be rotated by 90°. Manual actuators are open when the hand lever is directed into pipe direction. The pneumatic actuator is available as single-action or double-action. The single-action actuator uses a spring to close and air pressure to open the valve. The double-action actuator is opened and closed with air pressure.

Position Feedback. To provide a position feedback of actuators and valves, limit switches are used. By use of micro-switches or by the principle of induction the closed and open position of a valve can be detected and transmitted to a PLC.

5.4.6. Relieve Valves, Safety Valves

To prevent pipes and vessels from bursting due to overpressure, relieve or safety valves are installed (see Section 5.1.1.7).

An often found design principle consists of an O-ring sealed piston valve held by a spring. The preloading of the spring defines the critical pressure.

The released fluid needs to be drained to the floor or into a kill-tank.

5.4.7. Check Valves (Nonreturn Valves)

Check valves allow the gas or fluid to flow only into one direction. They find many applications in the piping of utilities. Check valves are designed with seals, balls, and springs, have a pressure drop and are seldom self-draining. This limits their use in sanitary piping or pumps. Cleaning of a sanitary pipe with a built-in check valve in countercurrent is for example not possible. Check valves in sanitary piping are therefore being replaced by process-controlled diaphragm valves.

5.5. Pipes

Generally the following criteria can be postulated for pipework [70]:

- The number of lines and their length has to be kept to a minimum.
- All valves and instruments have to be located so as to be both easily operable and maintainable.
- All pipes should be drainable, pockets need to be minimized if any.
- The amount of reducers, elbows, adapters and connections have to be minimized, this criteria requires a correspondence of the armatures to the tubing.
- Welded connections will always be preferred to dismountable connections with joints.

5.5.1. Pipe and Tube Sizes

Pipes or tubes in contact with the product are made of stainless steel (standard 1.4435/316L) as discussed in Section 4.1. Many standards, defining the outer diameter and the wall thickness of a pipe, are in use in biotechnology [67]. To avoid tremendous welding (to connect pieces of one standard to the other) it is important to have a consistent standard within one plant. Three types of standardization are the most used. ISO tubes (in fact ISO 1127, corresponding to DIN 2462/3) are frequently used in Europe. This standardization does not define very small tube sizes and therefore the metrical standardization DIN 11850 is used for inner diameters of 6, 8, 10, and 15 mm. In the US mainly "imperial" pipes (also called O.D. tubing), based on an inch measurement of the outer diameter, are used.

It is differentiated between seamless pipes and welded tubing. The inner surface of a seamless pipe is not as good as that of a welded pipe made out of a cold rolled plate. Seamless tubes therefore might require polishing to achieve higher inner surface qualities. Cold rolled welded pipes can have excellent inner surface finish without extra polishing and are therefore often preferred. Standard inner surface roughness of straight tubes are available in the range of $R_a \leq 0.2$ to 0.8 µm.

5.5.2. Pipe Applications and Sloping of Lines

Pipes must be routed in order to be self-draining. A slope of 0.3 to 0.5 % is generally sufficient to allow low viscous media to be drained.

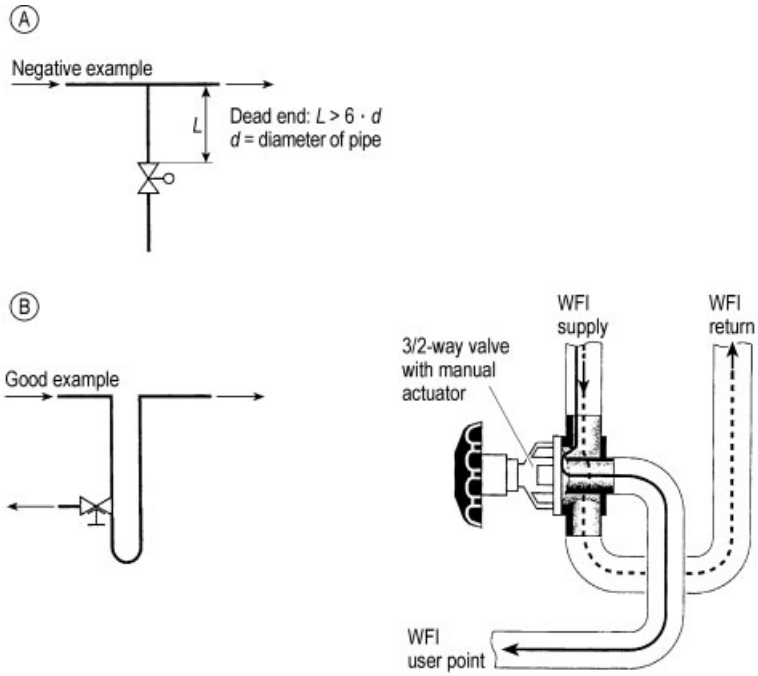

Figure 49. Distribution piping for WFI or deionized water
A) Negative example: piping with dead-end to the point of use; B) Good example: piping with a loop down to the point of use. By using a 3/2 way (T) valve dead-ends can be avoided.

Fluid distribution piping with T-piece connectors should avoid dead-ends longer than 6 times the diameter of the pipe (6 *d* rule) to the point of use. In fact such dead-ends can normally be avoided by using special T-valves and clever routing of pipes (Fig. 49).

Distribution piping for steam should be taken from the top of the horizontal main line, to avoid taking condensate to the point of use (Fig. 50).

5.5.3. Bends, Reducing-Fittings, T-Piece Connectors

Pipes can be bent but the surface quality of bent pipe is reduced compared to the original surface. Generally the surface roughness R_a of a bent pipe is about 2.5 to 4 times higher than the one of the original straight tube.

Prefabricated 90° bends achieve lower surface roughness by polishing the internal surface after bending the pipe. However, it must be considered that with each prefabricated bend two welds are added to the plant. Bends of 60° or lower are not considered to be very critical. When bending welded tubes, the weld seam should be on the lateral side of the bend, because this line is neither condensed nor stretched.

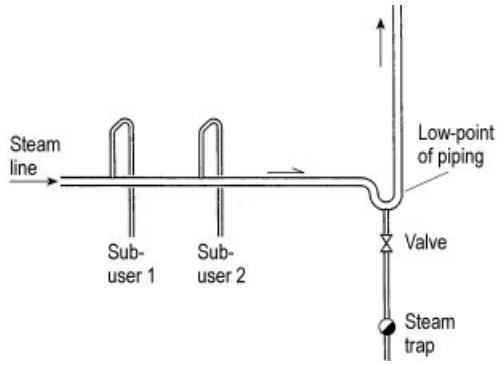

Figure 50. Pipe routing for steam lines. Lowest points and end of pipes require a valve – steam-trap combination. T-connections from horizontal pipes to sub-users must start from the top of the pipe.

Reducing fittings exist in eccentric and concentric design. In vertical lines concentric reducing fittings will be used while in horizontal lines

eccentric fittings are build in to allow the fluid to be drained (Fig. 51). T-piece connectors exist in a broad range of variations adapting all kind of tube sizes.

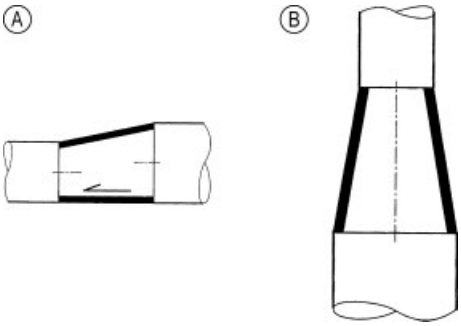

Figure 51. A) Eccentric reducing fittings can be used in horizontal, slightly inclined pipes; B) Concentric reducing fittings are only used in vertical pipes

5.5.4. Sterile Coupling of Pipes

Wherever possible, pipes should be welded for a maintenance-free coupling. On the other hand the possibility to dismount parts of piping to perform maintenance or exchange parts of a plant is often required. Several criteria can be formulated to pipe couplings:

- A static seal is required. For hygienic design reasons, seals with an O-ring are preferred compared to a gasket.
- The compression and position of the seal should be defined by the design of the coupling
- The seal should be clamped on one pipe end to avoid failure when being assembled.
- Pipe-ends should be identical. Male/female ends should be avoided.
- The in-line position of the two pipe ends must be defined by the shape of the coupling.
- When using a screwed pipe connection threads with a broad, steep and round pitch should be used. Such threads, originally used in milk industry, are very robust.
- Easy and safe handling.

Figure 52 shows different types of pipe couplings with their advantages and drawbacks.

5.6. Gassing Devices for Bioreactors

(→ Bubble Columns)

The formation and size of bubbles depends on the gas sparging device, the coalescence of the media and the turbulence in the vessel. Small gas bubbles have a higher interfacial surface area a than large bubbles. Therefore the OTR can be increased when smaller bubbles are created (see Section 2.5).

Shear stress generated by "exploding" gas bubbles at the liquid surface may damage suspended animal cells. The number of gas-bubbles must therefore be minimized or completely be avoided (bubble-free aeration).

The bubble size also has an influence on the structure of the foam. Foam created by small bubbles in a protein rich-medium (e.g., serum) is more stable than the one created by bigger bubbles.

5.6.1. Tubes and Ring Spargers

Tubes, fish tail and ring spargers are made of simple stainless steel pipes. The sparger can end with a tube opening, a fish tail (squashed pipe end) or with a perforated ring (ring sparger). In all cases large bubbles are created, which are usually smashed and dispersed by a Rushton impeller rotating at high velocity just above the sparger (Fig. 53). The sparger pipe must be self-draining to avoid a puddle when emptying the reactor. The ring sparger requires holes on the bottom of the ring to be drained.

5.6.2. Sinter Metal and Porous Membrane Spargers

When small bubbles are required and high stirrer speed is not applicable, sinter metal or porous hydrophobic membranes are mounted on the end of the aeration tube. Such spargers are often used for animal cell cultivation or for airlift reactors. The final bubble diameter size in the upper region of the reactor will much depend on the coalescence capability of the medium (see Section 2.5.2.2). Such sparger need to be dismounted for separate cleaning.

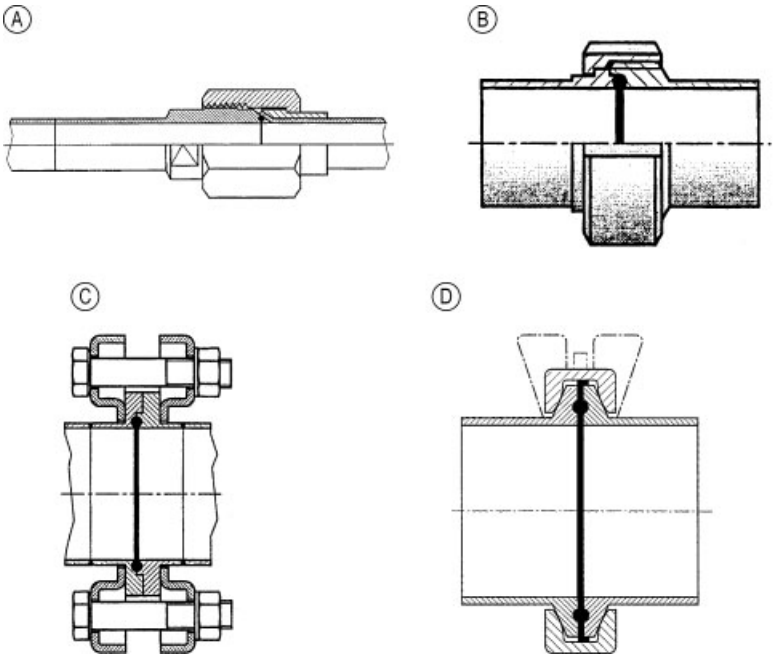

Figure 52. Sterile coupling of pipes
A) Flared tube connection with O-ring for small pipes (DN 6-10); B) Sterile coupling designed with a clamped O-ring, robust round thread and a shape aligning the pipes, for DN 15 to DN 50 pipes; C) Sterile flange connection with clamped O-ring, shape aligning the pipes, for large pipes > DN 50; D) Tri-clamp connection (gasket, clamped on pipe-end, identical tube ends)

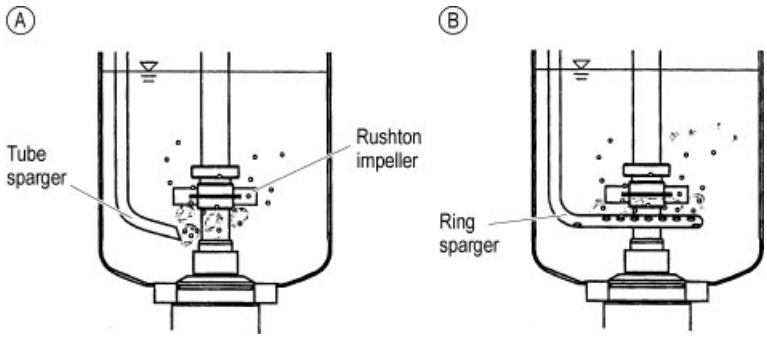

Figure 53. Tube type sparger below Rushton impeller
A) Tube sparger; B) Ring-sparger with sparging holes on the top and few holes on the bottom to drain the sparger when emptying the reactor

5.6.3. Surface Gassing

To avoid bubbles in the broth the reactor headspace can be gassed. Of course the supply of oxygen is limited by the liquid surface. Aeration through the surface is therefore reserved to small reactor volumes or to cell cultures with low densities. The gas – liquid mass transfer is increased by mixing the media (the liquid boundary layer at the phase surface is reduced).

Surface gassing is performed in animal cell cultivation to control the pH. CO_2 is stripped from the media by supplying air to the headspace, or solved in the media by supplying CO_2. The first will increase and the second reduce the pH.

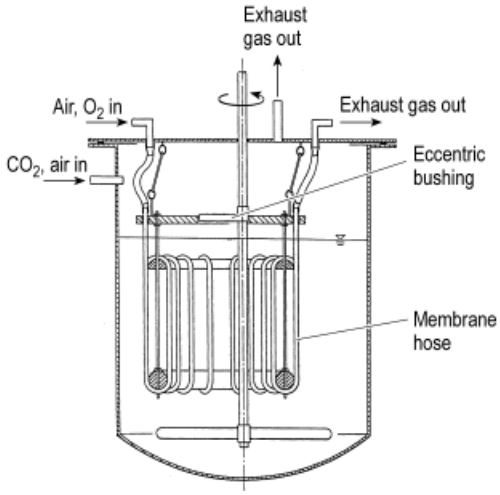

Figure 54. Support with winded membrane hose for bubble free aeration. The membrane is kept in a translatory movement [71]

5.6.4. Bubble-Free Membrane Gassing

Bubble-free membrane gassing is used for shear-sensitive animal cell cultivation (e.g., insect cells) or in few applications where highly volatile products should not be stripped out of the media. The main advantage of the method is that foam formation is completely avoided, thus foaming media such as those containing serum can be used for cultivation.

The technique uses a membrane separating the liquid from the gas phase. Mass transfer of gases from one phase to the other is achieved through the membrane. Usually thin-walled membrane hoses of silicone or porous hydrophobic materials such as polypropylene are used. The gas phase is pressurized, but the pressure drop across the membrane must be controlled to prevent the formation of gas bubbles on the liquid side of the membrane. By creating a turbulence in the liquid phase the mass transfer can be increased and gas bubbles be avoided. One possible method consists in winding the membrane hose on a support, which is kept in a translatory movement (Fig. 54) [71].

Another method consists in using reinforced silicone hoses to be able to increase the pressure of the gas phase and thus increase the gas transfer [72].

The possibilities to scale-up membrane gassed reactors is limited by the scale-up criterion membrane area to reactor volume ratio. Realistic are ratios of 10 to 25 m^2/m^3 in tanks up to 150 L. The *OTR* will usually be high enough for animal cell cultures cultivated without cell retention (see Section 2.1.3.2).

5.7. Mechanical Foam Separators

Foaming can be one of the major problems during aerobic growth of bacteria. Antifoam agents might severely reduce the *OTR* (see Section 2.5.3.2) or have a detrimental impact on the product recovery later in the downstream processing. Their use can be avoided when a mechanical foam separator working on the principle of a cyclone is used. Due to centrifugal forces the foam is separated into liquid and gas phases. The exhaust gas is withdrawn through the center of the separator. Mechanical foam separators are mounted in the headspace of the reactor or in the exhaust gas line (Fig. 55). The liquid phase is usually recirculated to the broth. The shaft of the rotor must be sealed. Usually single or double mechanical seals, which require lubrication, are used.

5.8. Sterile Filtration [73]

In biotechnology filters are used for sterile filtration of gases and liquids. Typical particles sizes to be separated by filters are bacteria (0.3 – 1 µm), mycoplasma (0.1 – 0.2 µm) or even viruses (0.04 µm). Filters are often characterized by their absolute filter rating expressed in diameter length of the largest hard spherical particle that passes through the filter under specified test conditions.

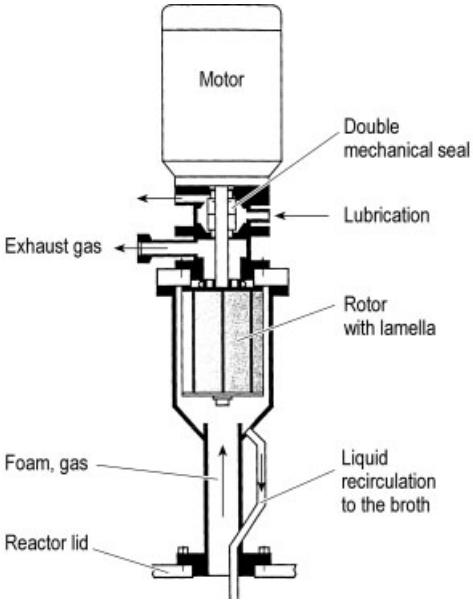

Figure 55. Mechanical foam separator mounted in the the exhaust gas line

5.8.1. Depth Filters

Depth filters used in biotechnology consist of a porous structure or fibers, which form an irregular three-dimensional net. Particles are not only retained on the surface but also inside the filter. Electrostatic forces, adsorbing forces, diffusion and the impact of the particles on the filter material are responsible for the retention. Depth filters can retain a high load of particles but the gas velocity across the filter must be adapted (≈ 0.3 m/s). Depth filters do not have an absolute rating since their retention rate will be proportional to their depth length, they therefore have to be sealed with gaskets over their depth.

Typical filter materials are ceramics, sinter metal, polypropylene, cellulose acetate and microglass. The advantages of materials such as ceramics and sinter metal lie in their resistance to acids and superheated steam. Depth filter are therefore robust and can easily be cleaned or regenerated. Gases in the laboratory as well as in large scale production are often sterilized with depth filter. For liquid filtration they are often used as a prefilter because of their ability to retain large amounts of particles without clogging (Fig. 8).

Depth filter cannot be integrity tested without destroying them, this limits their use for GMP production.

5.8.2. Membrane Filters

Membrane filters, also called absolute filters, retain particles on their surface similar to a sieve. Often filter cartridges are designed with a pleated membrane in a polypropylene support cage. The cartridge has O-rings, which fit into a filter housing (Fig. 56).

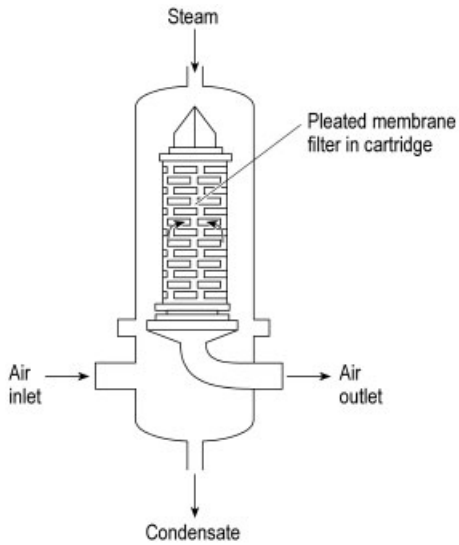

Figure 56. Filter housing for a pleated membrane filter in cartridge

Membrane filters can be used to sterilize liquids (hydrophilic membranes) as well as gases (hydrophobic membranes). Typical wet filter ratings are:

- 0.45 µm: in most cases sufficient to retain bacteria

- 0.2 µm: absolutely secure against bacterial contamination
- 0.1 µm: secure against mycoplasma contamination
- 0.04 µm: secure against virus contamination

Dry removal filter rates are smaller than the indicated wet ones.

The main advantage of membrane filters is their ability to be integrity tested after sterilization without contaminating the process. Standard test methods consist in wetting the membrane, applying a certain pressure below the bubble point and measuring the diffusion flow (forward flow test). By measuring the pressure decay the diffusion flow can be calculated (pressure hold test). Those nondestructive test methods have to be validated by the filter vendor by comparing them to destructive test methods. Automated integrity test instruments are available to perform, monitor, and log the test. A GMP production usually require the test to be performed after the sterilization, before-, and immediately after the production process.

Since the sterilization of membrane filters requires clean steam, the filter housing is often designed to be sterilized separately from the reactor. To perform a filter integrity test two supplementary connections on the non-sterile side of the filter cartridge are required (Fig. 11).

5.9. Hydraulic Pumps (→ High-Pressure Technology)

Hydraulic pumps can be grouped in centrifugal and positive displacement pumps. In the following a selection of pumps are presented and their advantages, disadvantages, and application fields for equipment used in biotechnology are discussed [74]. The following characteristics should be considered when choosing a pump:

- Total dynamic pressure of the pump (usually expressed in height of water column)
- Flow rate
- Ability to be self-draining
- Hygiene, sterilizable design
- Materials in contact with the product
- Ability to be self-priming
- Steady continuous flow versus pulsating flow
- In-line design of connections
- Ability to handle abrasive media
- Ability to handle viscous media
- Ability to handle shear-sensitive cells
- Ability to run dry
- Noise
- Costs (price per flow rate)

Centrifugal pumps Centrifugal pumps are often found in CIP units. To meet the hygienic design requirements centrifugal pumps should be equipped with a mechanical seal. The pump is not self priming and difficult to drain. When taking into account these disadvantages in the process design, the pump can, however, be integrated in the equipment. An advantage is that a broad range of centrifugal pumps with different flow-rates and pressures are available on the market.

Peristaltic pumps depend on progressive squeezing of flexible tubing by rollers mounted on a rotating plate. The liquid is pushed towards the discharge as the roller travels its circular path. The major advantage of such principle is that nothing but the tubing itself is ever in contact with the media. Typical tubing selections are silicone and Viton.

The pump can run dry, is self-priming up to a certain degree, can handle abrasive media and provides constant flow as long as the tubing is well adapted to the rollers and not too stressed. Usually the tube is sterilized in the autoclave and afterwards inserted in the pump-head. Large reinforced tubes can also be sterilized in-situ. Those advantages make it to be the pump par excellence in small to middle scale. Peristaltic pumps are operating at low shear forces and are therefore suitable to pump media with animal cells. Their usage is just limited by the pressure. Depending on the spring forces of the rollers, the wall thickness, the tubing diameter and material, pressures of 0.5 to 2 bar can be achieved.

Piston pumps (→ High-Pressure Technology, Chap. 3.3.1.) are self-priming, metering pump that provide accurate low flow rates at high or even very high pressure (up to 500 bar). For safety reasons a pressure relief valve should therefore always be installed behind the pump. Often frequency and stroke can be regulated to control the flow rate. Pulsation can be reduced

when using two or three pistons or when building in dampeners.

The sealing of the piston and the check valves are the weak points. Usually only clean nonviscous media can be delivered, solids would score the piston. A piston pump therefore requires further design tricks, such as a steam barrier sealing of the piston or the usage of a membrane (Fig. 57) to meet the standards of sanitary design.

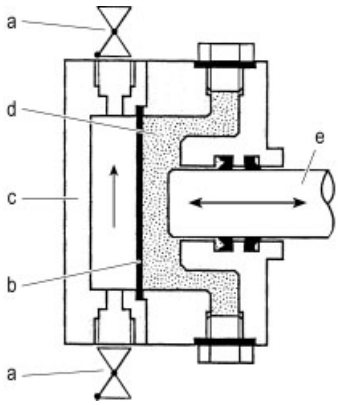

Figure 57. Piston-type pump with diaphragm. The axial movement of the piston lubricated by oil, is separated from the product by the diaphragm. Cleaning in counter-current is not possible due to the check valves.
a) Check valve; b) Diaphragm; c) Pump head; d) Oil chamber; e) Piston

Diaphragm pumps (→ High-Pressure Technology, Chap. 3.3.1.) are very reliable. A common model used for sanitary transfers in between vessels is the pneumatically driven double diaphragm pump. Two diaphragms, each separating a wetted chamber, are connected by a shaft inside the air motor, so that when one is on the pressure stroke pushing forward the other is correspondingly pulling back to create vacuum.

The advantages of double diaphragm pumps are their ability to stall if run against too great a head pressure, their ability to prime abrasive or shear sensitive media, their self-priming and dry running capability and the inherent variable flow rate capability (regulated by the air flow). Disadvantages are the need for clean, dry, compressed air, the air turbulence created around the pump (usually to be avoided in clean rooms), the fundamental inefficiency of a compressor/pump system, the noise of the check valves and the high repair and maintenance costs those pumps in practice usually demand.

The check valves of a double diaphragm pump can be replaced by diaphragm valves (Fig. 58). Cleaning (CIP) in counter-current to the transfer can thereby be achieved.

Other Pumps with Rotating Shaft. Gear, lobe, piston, circumferential piston, flexible impeller, sliding vane, and progressing cavity pumps are all positive displacement pumps with a rotating shaft. The difficulty is to seal the inser-

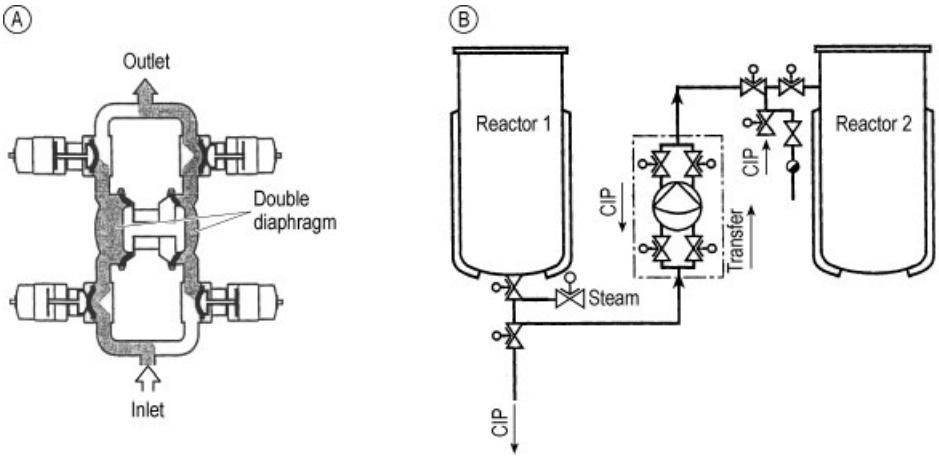

Figure 58. Air-driven double diaphragm pump with diaphragm valves instead of check valves
A) The opening and closing sequence of all diaphragm valves is computer controlled; B) Example of usage as a transfer-pump from reactor 1 to 2. CIP usually performed in counter-current is possible with such pump.

tion of the shaft, which might be accomplished by using mechanical seals or magnetic couplings (Fig. 59). The pumps are usually self-priming.

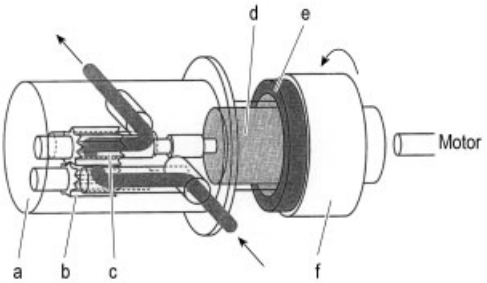

Figure 59. Gear-pump with magnetic coupling to the motor-drive. The pump-head is either sterilizable in place or separately in the autoclave.
a) Pump housing; b) Gear driving part; c) Gear receiving part; d) Inside magnet; e) Outside magnet; f) Magnet cap

6. References

1. E. L. Stadler (1998) Dual purpose fermentor and bioreactor? A capital quandary! Pharmaceutical Engineering 18. 3.
2. H.-J. Rehm (1980) Industrielle Mikrobiologie. Springer Verlag
3. H. G. Schlegel: *Allgemeine Mikrobiologie,* Thieme Verlag, Stuttgart 1992.
4. H. Dellweg: *Biotechnologie. Grundlagen und Verfahren.* VCH, Weinheim 1987.
5. H. Märkl et al.: "Cultivation of Escherichia coli to high cell densities in a dialysis reactor". *Appl. Microbiol Biotechnol.* **39** (1993) 48 – 52.
6. W. Crueger, A. Crueger (1989) Biotechnologie – Lehrbuch der angewandten Mikrobiologie. München; Wien; R. Oldenburg Verlag.
7. J. Monod (1942) Recherches sur la Croissance des Cultures Bactériennes. Hermann & Cie. Paris
8. B. Atkinson, F. Mavituna: *Biochemical Engineering and Biotechnology Handbook,* Nature Press, 1991.
9. I. S. Longmuir (1954) Respiration rate of bacteria as a function of oxygen concentration. Biochem 57. 81 – 87.
10. A. Ch. Dubach, H. Märkl (1992) Aplication of an extended Kalman filter method for monitoring high density cultivation of *Escherichia coli*, J. of Ferment Bioeng. 73, 5, 396 – 402.
11. B. O. Solomon, L. E. Erickson (1981) Biomass yields and maintenance requirements for growth on carbohydrates, Process Biochemistry Feb/March 44 – 49.
12. Stephanopoulos et al. Studies on On-Line Bioreactor Identification I to IV. Biotechnol. Bioeng. 26, 1176 – 1218.
13. J. A. Roels (1983) Energetics and kinetics in biotechnology, Elsevier biomedical press.
14. I. G. Minkevich, V. K. Eroshin (1973) Productivity and Heat Generation of Fermentation Under Oxygen Limitation. Folia Microbiol. 18, 376 – 385.
15. H. W. Blanch, D. S. Clark (1996) Biochemical engineering, ed. M. Dekker cop. New York, Basel.
16. J. G. Aunins, H.-J. Henzler: Aeration in cell culture bioreactors. In Biotechnology Vol.3 ed. Stephanopoulos. VCH Weinheim, 1993.
17. C. L. Cooney, D. I. C. Wang, R. I. Mateles (1969) Measurement of heat Evolution and correlation with oxygen consumption during microbial growth. Biotechnol. Bioeng. 11, 3, 269 – 281.
18. J. H. T. Luong, B. Volesky "Determination of Heat of some Aerobic Fermentations".*The Canadian Journal of Chemical Engineering* **58** (1980) 497 – 504.
19. M. S. Kharasch, "Heats of cumbustion of organic compounds". *Bur. Stand. J. Res.* **2** (1929) 359.
20. F. Bader in A. L. Demain, N. A. Salomon (eds.): "Sterilization: Prevention of Contamination", *Manual of Industrial Microbiology and Biotechnology,* Chap. 25, American Society for Microbiology, Washigton DC 1986.
21. K. H. Wallhäuser (1985) Sterilization in: Biotechnology Vol. 2 Fundamentals of Biochemical Engineering. Ed. H.-J. Rehm and G. Reed. VCH, Weinheim.
22. K. H. Wallhäuser (1995) Praxis der Sterilisation - Desinfektion – Konservierung. Georg Thieme Verlag Stuttgart – New York.
23. G. K. Raju, C. L. Cooney (1993) Media and Air Sterilization. In: Biotechnology Vol.3 ed. G. Stephanopoulos. VCH Weinheim.
24. Sh. Aiba, A. E. Humphrey, N. F. Millis: *Biochemical Engineering,* Academic Press Inc., 1973.
25. F. Wilde (1997) Sterilization and cleaning. Chemical Plants and Processing, 3.
26. H. Schunn (1998) Hohe Massstäbe, Leitfähigkeitsmessung nach USP 23. Process 11.
27. D. I. C. Wang et al. (1979) Fermentation and Enzyme Technology J. Wiley & Sons.

28. Rushton et al. (1950) Power characteristics of mixing impellers II. Chem. Eng. Prog. 46: 467 – 476.
29. M. Zlokarnik (1985) Tower-shaped reactors for aerobic biological waste-water treatment in Biotechnology Vol. 2 ed. H. Brauer VCH Weinheim.
30. M. Yasukawa et al. (1991) Gas Holdup, Power Consumption, and Oxygen Absorption Coefficient in a Stirred-Tank Fermentor under Foam Control. Biotechnol and Bioengineering, Vol. 38, 629 – 636.
31. K. Van't Riet (1979) Review of Measuring Methods and Results in Nonviscous Gas-Liquid Mass Transfer in Stirred Vessels. Ind. Eng. Chem. Process Des. Dev. 18. 357 – 364.
32. M. Charles, J. Wilson (1994) fermentor Design; In: Bioprocess Engineering Edited by B. K. Lydersen et al.; J. Wiley & Sons, Inc. New York, 3 – 67.
33. M. Zlokarnik (1999) Rührtechnik Theorie und Praxis. Springer Verlag.
34. H. Blenke (1987) Process Engineering Contributions to bioreactor design and operation. In: Biochemical engineering. Eds: Chmiel et al. G. Fischer Verlag. 69 – 91.
35. Dechema (1991) Standardisierungs- und Ausrüstungsempfehlungen für Bioreaktoren und periphere Einrichtungen. Dechema Frankfurt a. M.
36. I. Jöbses et al. (1991) Lethal events during gas sparging in animal cell culture. Biotechnol. Bioeng. 37. 484 – 490.
37. SCBS (1995) Guidelines for work with genetically modified organisms. Zürich Switzerland.
38. Belgian Biosafety Server http://biosafety.ihe.be.
39. ISSA (2000) Control of Risks in Work with Biological Agents. Part 3: Production. Prevention Series N° 2039.
40. M. M. Meagher (1992) Design of a University-Based. Biosafety Level 2, large-Scale Fermentation facility. BioPharm 10. 30 – 35.
41. J. Mauron (1981) The Maillard reaction in food. Progress In Food and Nutrition Science. 5, 5 – 35.
42. M. Krahe, G. Antranikian, H. Märkl (1996) Fermentation of extremophilic microorganisms. FEMS Microbiol. Rev. 18, 271 – 285. C. Kroll (1992) Neuere apparative Konzepte in der Bioreaktortechnik. Chemie-Technik 8 p 14 – 18.
43. R. Rautenbach, R. Albrecht (1981) Membrantrennverfahren Ultrafiltration und Umkehrosmose. Ed. Otto Salle Verlag Frankfurt a. M.
44. R. J. Datar, C-G. Rosén (1993) Cell and cell debris removal: Centrifugation and Crosslow filtration. In Biotechnology Vol. 3 ed. Stephanopoulos. VCH Weinheim.
45. T. Suzuki (1996) A dense culture system for microorganisms using a stirred ceramic membrane reactor incorporating asymmetric porous ceramic filters. J. ferment. Bioeng. 82/ 3/ 264 – 271.
46. Hayakawa et al. (1990) High density culture of Lactobacillus casei by cross-flow culture method based on kinetic properties of the microorganism. J. ferment. Bioeng. 70/ 6/ 404 – 408.
47. A. Sambanis, W.-S. Hu (1993) Cell culture bioreactors. In Biotechnology Vol.3 ed. Stephanopoulos. VCH Weinheim.
48. J. D. Macmillan et al. (1987) Monoclonal antibody production in stirred reactors. In: Large scale cell culture technology. Ed. B. K. Lydersen. Hanser Publishers Munich.
49. L. A. Behie et al. (1987) The application of continuous three phase fluidized bed bioreactors to the production of pharmaceuticals. In: Biochemical Engineering. Ed. Chmiel et. al. G. Fischer Verlag Stuttgart.
50. K. Nilsson (1987) Entrapment of cultured cells in agarose beads. In: Large scale cell culture technology. Ed. B. K. Lydersen Hanser Publishers Munich.
51. R. Rupp et al. (1987) Cellular microencapsulation for large-scale production of monoclonal antibodies. In: Large scale cell culture technology. Ed. B. K. Lydersen Hanser Publishers Munich.
52. D. Fassnacht, R. Pörtner, (1999) Experimental and theoretical considerations on the oxygen supply of animal fixed-bed cultures. J. Biotechnol. 72. 169 – 184.
53. R. Pörtner et al.: "Immobilization of Mammalian Cells in Fixed Bed Reactors".*Bioforum International,* G.I.T. Verlag Publishing Ltd. 1999.
54. R. Pörtner, H. Märkl (1998) Dialysis cultures. Appl. Microbiol. Biotechnol. 50. 403 – 414.
55. H. Märkl (1989) Folien und Membranen als neue Elemente im fermentorbau. forum mikrobiologie. GIT Verlag GmbH **12** 234 – 237.

56. J. Ogbonna, H. Märkl (1993) Nutrient-Split Feeding Strategy for Dialysis Cultivation of *Escherichia coli*. 41. 1092 – 1100.
57. H. Märkl et al. (1990) A new dialysis fermentor for the production of high concentrations of extracellular enzymes. J. Ferment. Bioeng. 69. 244 – 249.
58. C. Dillon, D. Rahoi, A. Tuthill (1992) Stainless steel for Bioprocessing, in BioPharm 5: Part 1: Materials selection, (April), 38 – 42; Part 2: Classes of Alloys, (May), 32 – 35; Part 3: Corrosion phenomena, (June), 40 – 44.
59. P. Meyer in B. K. Lydersen et al. (eds.): "Vessels for Biotechnology: Design and Materials".*Bioprocess Engineering,* J. wiley & Sons Inc., New York 1994, pp. 189 – 214.
60. R. Schmidt (1999) Nicht nur sauber sondern steril, PROCESS 11.1999 p. 46 – 49.
61. W. Küppers (1976) Die Oberfläche von nichtrostenden Feinblechen – Anlieferung und Weiterverarbeitung. Thyssen Edelstahlwerke AG, Technische Veröffentlichung 6327/6 Krefeld, Germany.
62. G. Graf (1987) Die Oberflächenbehandlung von Nichtrostenden Chromnickelstählen, Metall. 9.
63. D. Coleman, R. Evans (1990) Fundamentals of Passivation and Passivity in the Pharmaceutical Industry, Pharm. Eng. 10, 43 – 49.
64. http://www.acessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/showcfr.cfm
65. Ph. Bellow, J. Larsen: (1999) Elastomers in aseptic processing. In: Pharmaceutical processing. (Sept.), 106 – 107.
66. F. Menkel (1992) Einführung in die Technik von Bioreaktoren, R. Oldenbourg Verlag München.
67. P. Meyer (2000) Handlungsbedarf: Standards international vereinheitlichen. Pharma + Food 5. 66 – 69.
68. E. Breuker, U. Drees 1984) Pilotfermentor in der Industrie: Kriterien und konzipierung eines praxisnahen Systems. In: forum mikrobiologie 7 p. 12 – 18.
69. W. Storhas (1994) Bioreaktoren und periphere Einrichtungen. Vieweg-Verlag Braunschweig.
70. H. Adey, M. S. Polan (1994) Piping and Valves for Biotechnology.; In: Bioprocess Engineering Edited by B. K. Lydersen et al.; J. Wiley & Sons, Inc. New York, 215 – 252.
71. Büntemeyer, Lehmann (1987) Device and process for bubble-free gasification of liquids, in particular culture media for tissue culture reproduction. PCT/EP86/00566.
72. I. Sekoulov, H.-J. Bräutigam: (1987) Process for supplying oxygen from fermentation plant and device for implementation of the process. Patent PCT/EP86/00744; WO87/03615.
73. J. M. Martin et al. (1994) Cartridge filtration for Biotechnology. In: Bioprocess Engineering. Edited by B. K. Lydersen et al.; J. Wiley & Sons, Inc. New York, 189 – 214.
74. R. Stover in B. L. Lydersen et al. (eds.): "Pumps". *Bioprocess Engineering,* J. wiley & Sons Inc., New York 1994, 189 – 214.
75. AD-Merkblatt N4. (1983) Druckbehälter aus Glas. Beuth Verlag GmbH, Berlin (Sept).

Process Development

HERBERT VOGEL, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany

1. Introduction 1
1.1. The Objective of Industrial Research and Development 1
1.2. The Production Structure of the Chemical Industry 2
1.3. The Task of Process Development 4
2. Process Data 4
2.1. Chemical Mechanism 4
2.2. Physicochemical Data 5
2.3. Processing 6
2.4. Patenting and Licensing Situation 6
2.5. Development Costs 7
2.6. Location 7
2.7. Market Situation 7
2.8. Raw Materials 8
2.9. Plant Capacity 8
2.10. Waste-Disposal Situation 10
2.11. End Product 10
3. The Standard Course of Process Development 11
3.1. The Iterative Nature of Process Development 11
3.2. Drawing up an Initial Version of the Process 13
3.2.1. Tools used in Drawing up the Initial Version of the Process 15
3.2.1.1. Databases 15
3.2.1.2. Simulation Programs 15
3.2.1.3. Expert Systems 16
3.3. Checking the Individual Steps . . 17
3.4. Testing the Entire Process on a Small Scale 18
3.4.1. The Miniplant Technique 18
3.4.1.1. Introduction 18
3.4.1.2. Construction 19
3.4.1.3. The Limits of Miniaturization 21
3.4.1.4. The Limitations of the Miniplant Technique 22
3.4.2. Pilot Plant 22
4. Process Evaluation 23
4.1. Preparation of Study Reports . . 23
4.1.1. The Summary 23
4.1.2. Basic Flow Diagram 23
4.1.3. Process Description and Flow Diagram 23
4.1.4. Waste-Disposal Flow Diagram . . . 24
4.1.5. Estimation of Capital Expenditure . 24
4.1.5.1. Introduction 24
4.1.5.2. ISBL Investment Costs 24
4.1.5.3. OSBL Investment Costs 26
4.1.5.4. Infrastructure Costs 26
4.1.6. Calculation of Production Costs . . 26
4.1.6.1. Feedstock Costs 27
4.1.6.2. Utilities 27
4.1.6.3. Waste-Disposal Costs 30
4.1.6.4. Staff Costs 30
4.1.6.5. Maintenance Costs 31
4.1.6.6. Overheads 31
4.1.6.7. Capital-Dependent Costs (Depreciation) 31
4.1.7. Technology Evaluation 31
4.1.8. Assessment of the Experimental Work 35
4.2. Return on Investment 35
4.2.1. Static Return on Investment 36
4.2.2. Dynamic Return on Investment . . . 36
4.3. Economic Risk 37
4.3.1. Sensitivity Analysis 37
4.3.2. Amortization Time 37
5. Trends in Process Development . 37
6. References 38

1. Introduction

1.1. The Objective of Industrial Research and Development

In the chemical industry, ca. 6 % of turnover is spent on research and development [1, pp. 282, 302], an amount which is often of the same order of magnitude as the profit of the undertaking.

It is the job of research management to use these resources to achieve competitive advantages. After all, for the chemical industry the halcyon days of the sellers' market (demand > supply) have receded into the distant

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_437

past; it is now the customer who calls the tune (buyers' market).

Unlike consumer goods such as cars and clothes, most commercial chemical products are faceless (e.g., hydrochloric acid), and as a rule the customer is therefore only interested in sales incentives such as price, quality, and availability.

All the research activities of an industrial enterprise must therefore ultimately boil down to two basic competitive advantages, namely those of being "cheaper" and "better" than the competitor.

At the same time, "better" now refers not only to availability and product quality, but also environmental compatibility of the process, and the quality assurance approach of the supplier, etc.

1.2. The Production Structure of the Chemical Industry

If the production structure of the chemical industry is examined, it is seen that there are only a few hundred major basic products and intermediates that are produced on a scale of at least a few thousand to several million tonnes per annum worldwide.

This relatively small group of key products, which are in turn produced from only about ten raw materials, are the stable foundation on which the many branches of refining chemistry (dyes, pharmaceuticals, etc.), with their many thousands of often only short-lived end products, are based [2, p. 57 ff]. This has resulted in the well-known family tree (see Fig. 1), which can also be regarded as being synonymous with an integrated production system, with synergies that are often of critical importance for success.

A special characteristic of the major basic products and intermediates is their longevity. They are statistically so well protected by their large number of secondary products and their wide range of possible uses that they are hardly affected by the continuous changes in the range of products on sale. Unlike many end products, which are replaced by better ones in the course of time, they do not themselves have a so-called life cycle. However, the processes for producing them are subject to change. This is initiated by new technical possibilities and advances opened up by research but is also dictated by the current raw material situation (see Fig. 2).

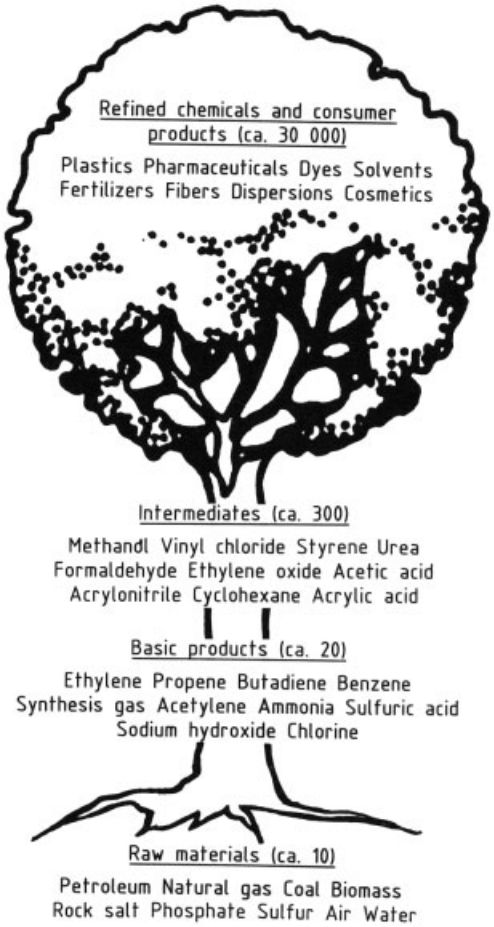

Figure 1. Product family tree of the chemical industry: starting from raw materials and progressing through the basic products and intermediates, to the refined chemicals and final consumer products [3], [4]

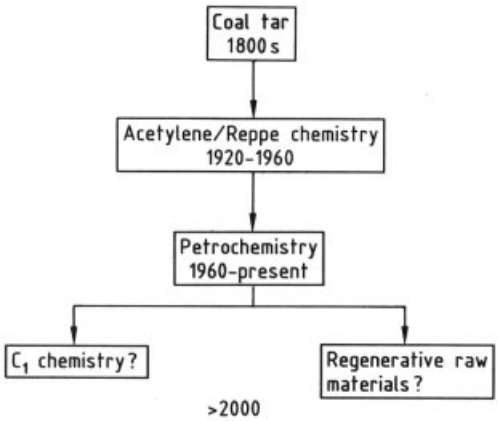

Figure 2. How the raw material base of the chemical industry has changed with time

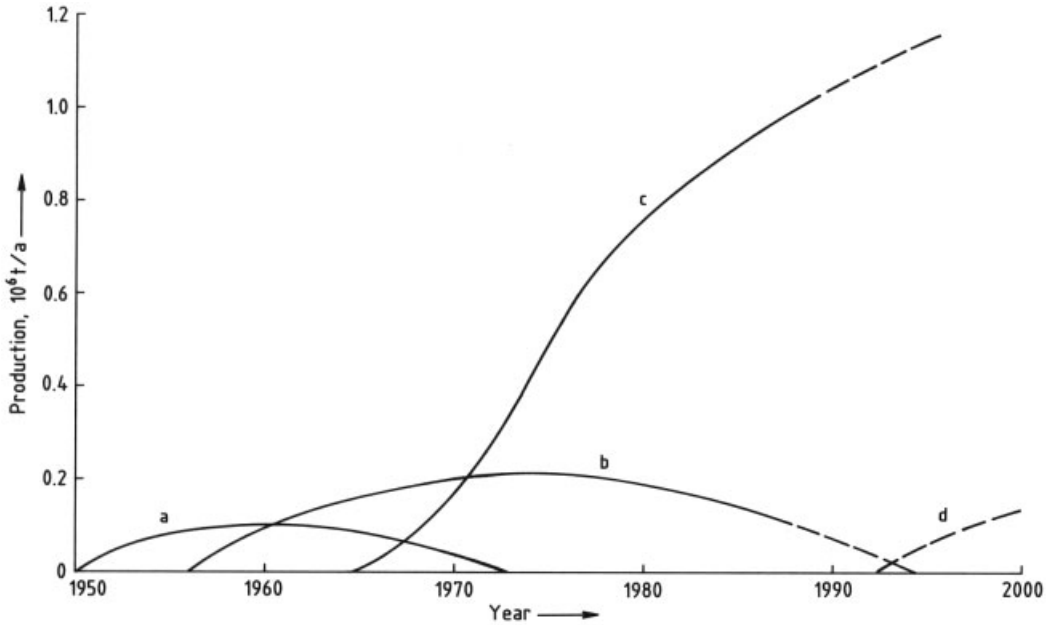

Figure 3. Life cycles of the acrylic acid processes
a) Cyanohydrin,acrylonitrile,and propiolactone processes; b) Reppe process; c) Propylene oxidation; d) New process?[4]

Here it is not the individual chemical product, but the production process or technology which has a life cycle. Figure 3 shows the life cycles of the acrylic acid processes [4].

Enterprises which already have competitive advantages must take account of this technology S curve [5] in their research and development strategy (see Fig. 4). The curve shows that as the research and development expenditure on a particular technology increases, the productivity of such expenditure decreases with time [6]. If enterprises are approaching the limit of a particular technology, they must accept disproportionately high research and development expenditure, with the result that the contribution made by these efforts to the research objectives of "cheaper" and/or "better" becomes increasingly small, thereby always giving the competitor the opportunity of catching up on the technical advantage.

Once an enterprise has reached the upper region of the product or technology S curve, the question arises whether it is necessary to switch from the standard technology to a new trend-setting technology in order to gain a new and sufficient competitive advantage [7]. Figure 4 depicts this switch to a new technology schematically and shows that on switching from a basic technology to a new trend-setting technology, the productivity in the research and development sector increases appreciably, thus enabling substantial competitive advantages to be achieved.

It is precisely on this innovative activity that the prosperity of highly developed countries with limited raw material sources such as Germany and Japan is based.

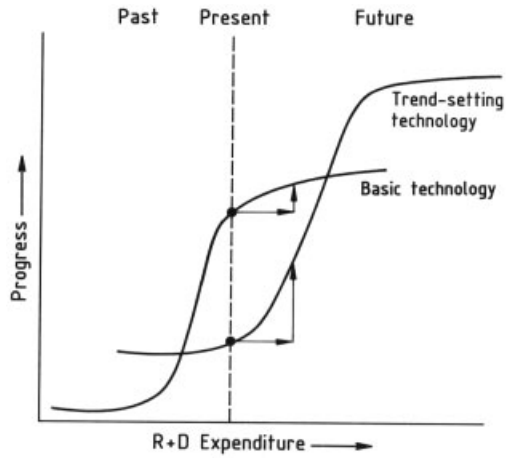

Figure 4. The technology S curve [8]: The productivity of the research and development expenditure increases considerably by switching from the basic technology to a new trend-setting technology

To assess whether a "better" and/or "cheaper" research and development strategy is still acceptable in the long term for a given product or production process, the R & D management must develop an "early warning system" that determines the optimum time for switching to a new product or a new technology [9].

1.3. The Task of Process Development

The task of process development is to extrapolate a chemical reaction discovered and researched in the laboratory to an industrial scale, taking into consideration the economic, safety, and ecological boundary conditions.

The starting point is the laboratory equipment and the outcome of development is the production plant; in between, process development is required. The following account shows how this task is generally handled. Although the sequence of steps in the development process described are typical, they are by no means obligatory, and it is only possible to outline the basic framework.

2. Process Data

When the laboratory phase has been completed and before the actual process development is started, further information must be obtained since the latter stage is normally associated with high costs.

2.1. Chemical Mechanism

The core of a chemical plant is the reactor. Its input and output decide the structure of the entire plant built up around it. Therefore, detailed knowledge of the chemical reaction must be available at the earliest possible stage. In particular, the following questions must be clarified:

1) Thermodynamic equilibria
2) Kinetics of the main, secondary, and side reactions
3) Dependence of selectivity and conversion on the process parameters
4) Heats of reaction

Thermodynamic Equilibrium. Many reactions, such as esterifications and dimerizations, proceed to chemical equilibrium. The thermodynamic equilibrium provides information about the maximum possible conversion, and to assess the potential conversion it is essential to know how far the intended reactions are from chemical equilibrium.

Example 2.1. The formation of trioxane from formaldehyde

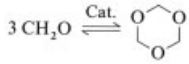

(2.1)

in the gas phase is a typical equilibrium reaction.

The equilibrium constant, which is given by

$$K_{\mathrm{P}} = \frac{p_{\mathrm{T}}}{p_{\mathrm{F}}^{3}} = \frac{4\left(p_{\mathrm{F}}^{0} - p\right)}{\left(3p - p_{\mathrm{F}}^{0}\right)^{3}} \tag{2.2}$$

where

p = total pressure at equilibrium
p_{F}^{0} = initial partial pressure of formaldehyde
$p_{\mathrm{T}}, p_{\mathrm{F}}$ = equilibrium partial pressure of trioxane or formaldehyde, respectively,

can be expressed according to [10] as:

$$\log\left(K_{\mathrm{P}}/\mathrm{bar}^{-2}\right) = \frac{7350}{T/\mathrm{K}} - 19.8 \tag{2.3}$$

It follows that if the process conditions assumed are $T = 353$ K and $p^{0}{}_{\mathrm{F}} = 0.2$ bar, a maximum formaldehyde conversion of:

$$U = \frac{3 \cdot p_{\mathrm{T}}}{p_{\mathrm{F}}^{0}} \cdot 100 = 35\% \tag{2.4}$$

can be achieved in one throughput. The conversion actually achieved will be $< 35\,\%$ and depends on the catalyst and reactor systems.

The thermodynamic data (standard enthalpy and entropy of formation, specific heat of the reactants) required to calculate the equilibrium constant can be found in published tables or can be estimated by approximation methods (see → Estimation of Physical Properties) [11].

Kinetics. In order to specify the type of reactor to be used later, information must be available on the potential reaction routes to the main and secondary products and the byproducts. The rate of formation and its dependence on process parameters such as temperature, pressure, and catalyst concentration should, if possible, be known quantitatively.

Example 2.2 [12, pp. 173 – 175]. The starting material A reacts not only to form product P, but also (1) to form an undesirable byproduct Y in a parallel reaction or (2) to form an undesirable secondary product X in a sequential reaction:

$$\begin{array}{l} A \xrightarrow{k_1} P \xrightarrow{k_2} X \\ \quad\;\; \vert \xrightarrow{k_3} Y \end{array} \tag{2.5}$$

1) First-order parallel reaction
   For the starting material A:

$$\frac{dc_A}{dt} = -(k_1+k_3)\cdot c_A \tag{2.6}$$

   with $c_A = c_A^0$ at time $t = 0$, integration results in:

$$c_A = c_A^0 \cdot \exp\{-(k_1+k_3)\cdot t\} \tag{2.7}$$

   For the useful product P:

$$\frac{dc_P}{dt} = k_1 \cdot c_A \tag{2.8}$$

   with $c_P = c_P^0$ for $t = 0$, integration gives

$$c_P = c_P^0 + \frac{k_1 \cdot c_P^0}{(k_1+k_3)} \cdot \{1 - \exp[-(k_1+k_3)\cdot t]\} \tag{2.9}$$

   The same applies to c_Y.
2) First-order sequential reaction
   This is covered by three simultaneous differential equations:

$$\frac{dc_A}{dt} = -k_1 \cdot c_A \tag{2.10}$$

$$\frac{dc_P}{dt} = k_1 \cdot c_A - k_2 \cdot c_P \tag{2.11}$$

$$\frac{dc_X}{dt} = k_2 \cdot c_P \tag{2.12}$$

   If $c_P^0 = c_X^0 = 0$, integration of these equations yields:

$$c_A = c_A^0 \cdot \exp(-k_1 \cdot t) \tag{2.13}$$

$$c_P = \frac{k_1 \cdot c_A^0}{k_1 - k_2} \cdot \{\exp(-k_2 \cdot t) - \exp(-k_1 \cdot t)\} \tag{2.14}$$

$$c_X = c_A^0 \cdot \left\{1 + \frac{k_1}{k_2 - k_1} \cdot \exp(-k_2 \cdot t) - \frac{k_2}{k_2 - k_1} \cdot \exp(-k_1 \cdot t)\right\} \tag{2.15}$$

2.2. Physicochemical Data

A knowledge of exact physicochemical data has become more important for a number of reasons:

1) The increasing use of simulation programs has made the requirements for exact physicochemical data increasingly important. The result of a simulation calculation can only be as good as the quality of the physicochemical data.
2) To approve chemical plants, the authorities demand information on the toxicity, degradability, and safety of the materials involved [13], [14].
3) The public are demanding more information on the effect of the materials being handled on the environment.

At the start of process development, material data files for pure substances as well as binary and ternary mixtures are started. As development advances, these grow and must be continually updated. They subsequently form a document which is passed to the planning and plant construction departments.

The layout shown in Table 1 has proved useful as a model for the collection of pure-substance data.

Only data which have been reliably evaluated are kept in the files and these are the only values fed into the simulation programs.

At the start of process development, the first step is to collect all available literature data on a material (see→ Estimation of Physical Properties).

Many companies now maintain their own material databases. The data have been evaluated and should have been found satisfactory when used in practice. The times when a single company had 10 different sets of Antoine parameters for water between 0 and 100 °C ought now to be over.

If no material values can be found in the literature, which is often the case for binary and ternary data, they should initially be estimated by using empirical formulae or realistic values that have proved themselves for many substances (e.g., heat capacity≈2 kJ kg^{-1} K^{-1} for most organic liquids). However, experimental determination or confirmation of the most important values on which the plant design is based is unavoidable.

It is sheer negligence to rely on physicochemical data originating from a single literature reference.

2.3. Processing

Processing is intimately linked to the chemical reaction. While the chemical aspects are still subject to extensive modification (for instance, in the choice of catalyst, solvents, etc.), there is little point in paying a great deal of attention to processing. It is only when the reaction mixture is being produced in a representative manner that an initial processing procedure can be devised (see Section 3.2). There is no generally accepted way of going about this. It is more an art than craft, and reliance must be placed on inspiration and experience. It has, however, always been helpful to gather as many experienced specialists as possible and to discuss the problems regularly. Although tentative efforts are now being made to compile collective know-how in expert systems for processing strategies, these are still far from being universally applicable.

The above deliberations culminate in a separation concept which can be broken down into individual unit operations. Initially, these operations can be examined in the laboratory either on a batch basis, or in a continuous process to determine whether they are feasible in principle (e.g., are azeotropes formed?, how difficult is separation?, what is the dissolution rate?, can the phases be separated?, etc.). This gives the initial process concept, which can be used as a basis for starting the actual process development.

Table 1. Important pure-substance data frequently required in process development using acrylic acid as an example [15]

Product name	acrylic acid
CAS no.	[*79-10-7*]
Formula	$C_3H_4O_2$
Molar mass, kg/kmol	72.06
Melting point, °C	13.5
Boiling point, °C	141.0
Antoine parameters *	
A	9.135
B	−3 245
C	216.4
Vapor pressure at 20°C, mbar	10
Heat of evaporation at *bp*, kJ/kg	633
Heat capacity, kJ kg^{-1} K^{-1}	1.93 (liquid)
Density, kg/m^3	1 040 at 30°C
Viscosity, mPa · s	1.149 at 25°C
Heat of formation, kJ/mol	
Gross calorific value, kJ/kg	19 095
Heat of transformation, kJ/kg	1 075 (polymerization)
Heat of fusion, kJ/kg	154 at 13°C
Solubility in H_2O	miscible
Solubility of H_2O in acrylic acid	miscible
MAK	
Toxicity (LD_{50}, rat, oral)	340 mg/kg
Water hazard class	1 ("low water hazard")
Odor threshold	pungent
Flash point, °C	54
Ignition temperature, °C	390
Lower explosion limit, vol %	2.4 (47.5°C)
Upper explosion limit, vol %	16 (88.5°C)

* ln (p/bar) = $A + B/(C + T/°C)$.

2.4. Patenting and Licensing Situation

(See also → Patents)

If a new idea is to be used in developing a process, a careful check should be made to determine whether any third-party proprietary rights would be infringed.

The best situation is one in which the ideas are completely new and the company can file its own patents to ensure it can use the ideas without restriction. However, patent applications not only provide protection but also represent a source of information for competitors [16]. Careful consideration must therefore be given to the question of whether it is desirable to keep an idea secret or whether it should be patented.

The risk involved in deciding against patenting in-house knowledge can be reduced by eliminating the patentable novelty of that information on a worldwide basis through publication in a suitable form, e.g., in an in-house journal or a technical bulletin, etc.

If the idea is restricted by third-party proprietary rights, the following questions should be answered:

1) When do those rights expire?
2) How can they be rendered void?
3) How can they be circumvented by modifications?

If the idea is completely covered by third-party proprietary rights, the only way out is to enter into licensing negotiations.

Normally, a process should only be developed after the legal situation has been clarified.

2.5. Development Costs

Development costs often constitute a considerable proportion of the total cost of a project. If the process is completely new, these costs are often around 50 % of the investment costs for the industrial plant, and perhaps more if a pilot plant must be built.

Development costs can be quoted with sufficient accuracy if the following items are assessed:

Man years: This involves determining how many chemists and engineers will be employed on the project and for how long. Since the costs associated with each of the specialists working on the individual investigations are known, the man-year costs can then be calculated.

Setting up costs: A rough estimate of the costs of setting up a small-scale plant is sufficient since this item is normally the smallest in the development costs. Once some notion of the process is available (a rough process flow diagram), these costs can be determined with reasonable accuracy.

Operating costs: The operating costs can be estimated from the assumed development time and the supervisory staff (shift personnel) required. A rule of thumb is that the annual operating costs of an experimental plant will be at least equal to those required to set it up.

More difficult to estimate are the amounts required annually for modifications and repairs to the experimental plant. These depend very much on whether new or already proven technologies are being used.

The development costs determined in this way do not contain any contributions arising from personnel not directly involved in development, such as management or patent specialists. Nor is account taken of unsuccessful developments. It nevertheless seems reasonable to apportion the direct development costs to the product and to allow for the other research and development costs in some other way, for example by means of a general cost provision (see Section 4.1.6.6). Although this method only accounts for the costs of successful development projects, the financial burdens on the product will be relatively high if it must bear all the development costs of the first plant.

2.6. Location

The main competitive advantages such as price and availability are nowadays often dependent on the location of the production plant. Thus, there is now an increasing tendency to move the production of major basic chemicals such as methanol to the oil-producing countries. However, products whose production requires considerable know-how are less location-dependent and will continue to be produced in the industrialized countries, even in the future. Extreme cases are those in which legal requirements decide the location, as is the case in the field of gene technology.

In deciding on location, consideration should be given to the following aspects:

1) Proximity to the sales market
2) Proximity to the sources of raw materials or precursors
3) Transport facilities (ship, rail, road)
4) Quality of the available labor
5) Availability of cooling water
6) Waste-disposal management (sewage treatment plant, incineration plant)
7) Political considerations (tax, investment aid, approval situation)

2.7. Market Situation

The competitive advantages generated by process development should ultimately benefit the

market and its customers. It is therefore impossible to target process development correctly without a precise knowledge of that market. The following information is required if the basic data needed to plan production (plant size, maximum production costs) are to be obtained

1) Market price (how it will develop and vary with time)
2) Demand, broken down into in-house and third-party demand
3) Market growth
4) Utilization of capacity in existing plants, with a breakdown in terms of location (United States, Western Europe, Japan) and production process
5) Competitive situation (who is the largest competitor?)
6) Customer situation (are there many small customers or a few large ones?)

The question of whether a process development or improvement will be worthwhile can be analyzed by using a portfolio representation (Fig. 5).

Portfolio analysis is a strategic planning tool, used to concentrate research investment on products for which market prospects appear favorable and competitive advantages can be exploited [17].

2.8. Raw Materials

Raw materials, their availability, and their price structure have always been crucial factors responsible for shaping the technological base of the chemical industry and, consequently, its growth and expansion (see Fig. 2).

The chemical industry produces a wide variety of products from only a few inorganic and organic raw materials.

Thus, raw materials such as natural gas, petroleum, air, and water are used to produce basic chemicals such as synthesis gas ($CO-H_2$ mixture), acetylene, ethylene, propene, benzene, ammonia, etc. These are used in turn to manufacture intermediates such as methanol, styrene, urea, ethylene oxide, acetic acid, acrylic acid, cyclohexane, etc. (see Fig. 1).

Any change in the price and availability of raw materials will therefore have a substantial effect on the production processes for secondary products.

Since the exact purchase prices (transfer prices) are an important source of information in assessing manufacturing costs, they must always be available and kept up to date in a chemical company. Figure 6 shows the variation in the price of some basic chemicals.

It is not only the price of potential raw materials which must be determined but also their availability and quality (purity, state as delivered).

It is never possible to reach a generally applicable decision on a particular raw material since much depends on its availability at a particular location and variation with time.

2.9. Plant Capacity

Optimum plant capacity, i.e., the size of plant which yields the maximum return (see Section 4.2), will be achieved if a degree of capacity utilization of 100 % is reached 3 – 5 years after start-up. However, since the plant capacity must be decided about 4 – 5 years before start-up for planning and legal approval reasons, and plants which are not fully utilized over a period of time increase the specific production costs, it is essential to be able to look 7 – 10 years into the future in determining the optimum capacity. Therefore, precise knowledge of the market potential and of market growth trends is required.

The relationship between production price and degree of capacity utilization is as follows [19]

$$\frac{P}{P_0} = \frac{1+F/V}{1+F/V \cdot K_{\text{eff}}/K_0} \tag{2.16}$$

where

P = production price in monetary units per kilogram,
P_0 = production price at rated capacity,
F = fixed costs,
V = variable costs,
K_{eff} = effective plant capacity, in tonnes per annum,
K_0 = rated capacity, in tonnes per annum.

Thus, as the degree of capacity utilization drops, the production price increases hyperbolically and this dependence is the greater, the

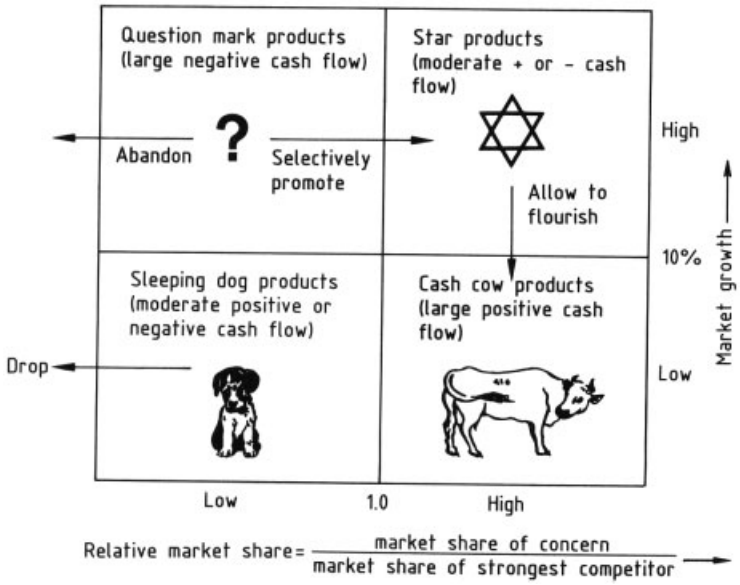

Figure 5. Portfolio matrix [17]
A) Sleeping dog products These are problem products whose competitive situation is weak because of their low market share. In such cases process development would not be worthwhile.
B) Cash cow products These are products which, although their market growth is low, have a high market share. To exploit the cost advantages to the full, there is little point in heavily investing in process improvements. Instead, these products should be "milked" and the resulting cash flow used to support the star products.
C) Star products All the efforts should be concentrated on these products since they will safeguard the survival of the company in the long term.
D) Question mark products In these cases the company must decide whether to increase the market share (turn them into star products) or to abandon them if the prospects are slim.

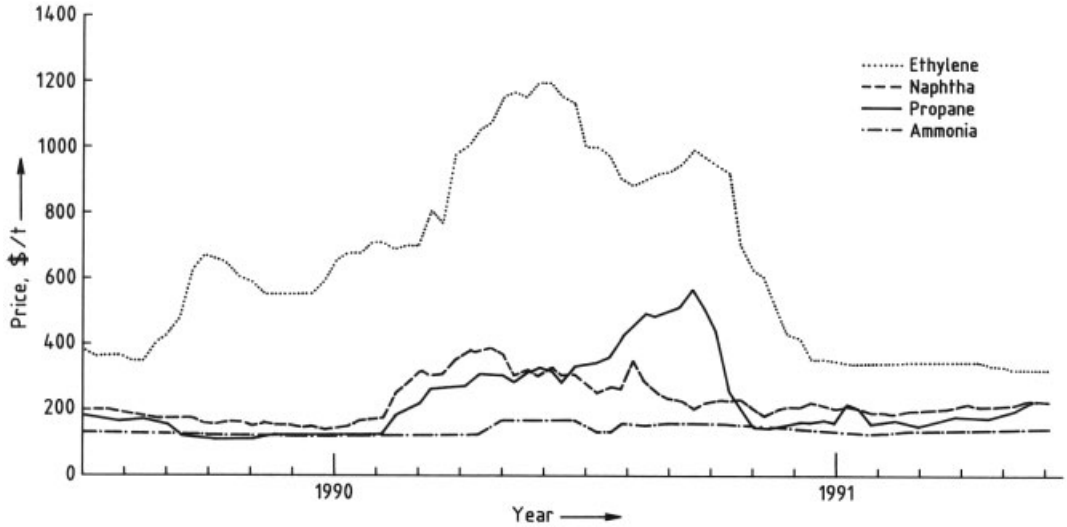

Figure 6. Variation in the price of some basic chemicals with time [18]

higher the contribution made by the fixed costs (see Fig. 7).

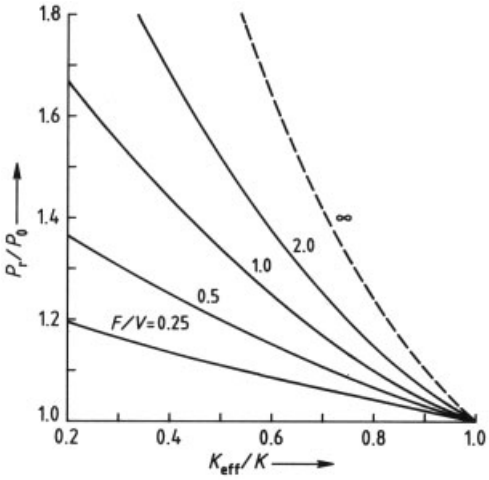

Figure 7. Dependence of the relative production price on degree of capacity utilization for different fixed cost/variable cost ratios F/V

The greater the uncertainty in the market studies and the newer the technology used, the greater will be the preference for choosing a lower capacity. Many Japanese chemical companies adopt this approach and prefer to set up several plants with smaller capacities. However, lower capacities result in higher specific plant costs and therefore in higher production costs. This is represented by the following empirical relationship:

$$I_2 = I_1 \cdot (K_2/K_1)^{\chi} \tag{2.17}$$

where

I = capital expenditure (investment),
K = capacity,
χ = degression exponent.

For individual items of equipment, $\chi = 0.6$, and for plants containing the same number of items of equipment $\chi = 2/3$ [20].

2.10. Waste-Disposal Situation (see Section 4.1.6.3)

Waste-disposal facilities (incineration, sewage treatment plant, waste-disposal site) depend on the location. However, at the beginning of process development the location will generally still be undecided and the process should therefore be designed to produce only waste products which can be readily incinerated. Processes which produce salts or large amounts of dilute aqueous solutions containing materials that are difficult to degrade biologically are inherently location-dependent.

2.11. End Product

It is essential to determine the demand for the end product as well as its specification and achievable price. A strongly competitive situation will mean particularly severe changes with time, and in this case it is necessary to adopt a particularly critical approach to forecasting. The product specification is of particular significance for process development. In the simplest case, it consists of a minimum purity, for example, that the end product should contain at least 99 % of a particular chemical compound. The permissible impurity content may vary considerably from one product to another. Thus while it may be a few percent in the case of dyes, the limit is often only a few parts per million for monomers.

However, in most cases the end product cannot be specified so simply since its subsequent use may be affected in different ways by individual impurities. In such cases, it is not sufficient to specify an upper limit for the sum of impurities; instead, limits must be specified for important individual components. Often, the chemical analysis is not sufficient or is not sufficient on its own for specifying the end product, a case in point being, for example, polymers. In such cases chemical analysis is supplemented by process engineering properties which, in the simplest case, can be reduced to simple physical properties.

The process development engineer endeavours to obtain as detailed a specification as possible for the end product before he starts work, but usually the specification will only be finalized during process development.

Statements such as "as pure as possible" are inadequate since the removal of the final traces of impurities is time consuming and expensive. Individual specifications which are of no significance in characterizing the end product should be avoided. If possible, a composition which can be determined analytically is preferable to a

purely physical characterization since it is more closely related to the process sequence.

3. The Standard Course of Process Development

3.1. The Iterative Nature of Process Development

The development of chemical processes is a complex procedure.

The first hurdle in establishing a new process is overcome when a promising synthetic route, usually with associated catalysts, is discovered, but many issues must be clarified and many problems solved before industrial implementation is feasible and all the documents necessary to design and operate a chemical plant have been assembled [21], [22]. How are these planning documents arrived at?

The conventional procedure is carried out in three stages (see Fig. 8):

An initial process concept is developed on the basis of an optimized laboratory synthesis. This is not usually continuous (see Section 3.2), and the individual process steps are examined independently of each other in the laboratory (see Section 3.3). A continuous laboratory plant (the so-called miniplant) is then designed, set up, and operated (see Section 3.4.1). This is a small but complete plant handling production quantities of ca. 100 g/h and consisting of a synthesis section, working-up, and all recycling streams. Once the process concept has been confirmed in the miniplant, the next step is to design and set up a trial plant with a much higher capacity, the scale of this pilot plant being between that of the miniplant and that of the industrial plant. The quantities produced are a few kilograms per hour or tonnes per annum and enable, for example, application tests to be carried out on the product or large-scale deliveries to be made to customers. Operation of the pilot plant makes it possible to complete and verify data and documentary information obtained at an earlier stage of process development (see Section 3.4.2).

The scale-up factor from one stage to the next is always limited by the "minimum principle", i.e., the process stage or piece of equipment with the lowest scale-up feasibility determines the maximum capacity of the next larger plant whose operating performance can be calculated. It is here that the process development engineer has an opportunity to save time and money. If reliable documents can be drawn up which enable the maximum scale-up of four powers of ten ($100\,g/h \times 10^4 = 1\,t/h$) to be spanned in designing a production plant, omitting the intermediate stage also eliminates the costs of the pilot plant and 3–4 years' development time [23], providing an important advantage in terms of cost effectiveness and marketing. Consequently, efforts are now made to extrapolate directly from the miniplant to the production scale.

The following list summarizes some of the important process steps and typical maximum scale-up values above which reliable scale-up is no longer possible [24]:

Reactors	
Multitubular reactor, homogeneous tube, homogeneous stirred tank	>10 000
Bubble column	< 1000
Fluidized-bed reactor	50–100
Separation processes	
Distillation and rectification	1000–50 000
Absorption	1000–50 000
Extraction	500–1 000
Drying	20–50
Crystallization	20–30

Larger scale-up factors are possible for gas-phase reactors such as multitubular reactors.

An overview of the main types of reactor and a discussion of the criteria for the choice of a particular reactor and reactor scale-up during process development can be found in Parts A and B of this volume. The use of dimensional analysis in scale-up is discussed in→ Scale-Up in the Chemical Industry.

The unit operations of distillation, rectification, and absorption can be scaled up without an intermediate stage. This explains, for example, why much effort is expended on obtaining operating conditions that allow the reaction and processing of gases and liquids. The behavior of gases and gas–liquid equilibria can be explained well in physical terms and calculation is therefore straightforward.

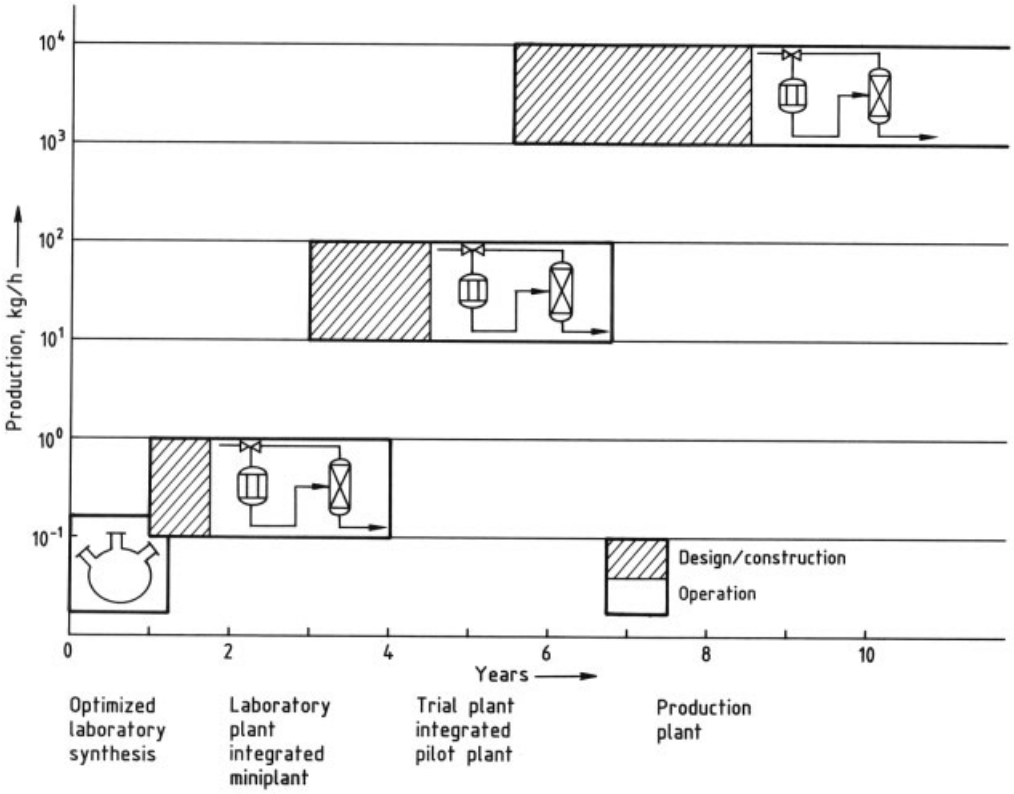

Figure 8. The stages involved in developing a process
The time intervals specified for design, installation, and operation are only tentative and, in specific cases, the real values may differ greatly from those given

Example 3.1. Vapor – liquid equilibria can readily be measured experimentally and can be described well in mathematical form. The general relationship for the equilibrium between a liquid phase and an ideal gas phase takes the form:

$$y_{\mathrm{i}}\,(x_{\mathrm{i}},T) = \frac{\gamma_{\mathrm{i}}\,(x_{\mathrm{i}}) \cdot P_{\mathrm{i}}^{0}\,(T)}{P} \cdot x_{\mathrm{i}} \tag{3.1}$$

where

y_i = mole fraction of component i in the gas phase,
x_i = mole fraction of component i in the liquid phase,
$P_{\mathrm{i}}^{0}(T)$ = saturation pressure of pure component i at the system temperature T,
P = vapor pressure of the mixture,
γ_i = activity coefficient of component i.

The vapor pressure of the pure component i, $P_{\mathrm{i}}^{0}(T)$, is often represented by a three-parameter Antoine expression:

$$P_{\mathrm{i}}^{0}\,(T) = A_{\mathrm{i}} + \frac{B_{\mathrm{i}}}{C_{\mathrm{i}} + T} \tag{3.2}$$

Several models are available for describing the activity coefficient γ_{i}. Examples are the Wilson model, which can be applied to completely miscible mixtures, and the NRTL model, which is suitable for systems with a miscibility gap [25], [26].

For solids, scale-up factors are several orders of magnitude lower because of complications such as deposits, encrustations, and abrasion. This is true not only for synthesis (e.g., in fluidized-bed reactors), but also for processing (e.g., drying and crystallization).

Process development does not, however, take place as a one-way street. Assumptions are made for the individual development stages which are

only confirmed or refuted when the next stage is being worked on. It may be necessary therefore to go through the individual stages several times with modified assumptions, resulting in a cyclic pattern (see Fig. 9).

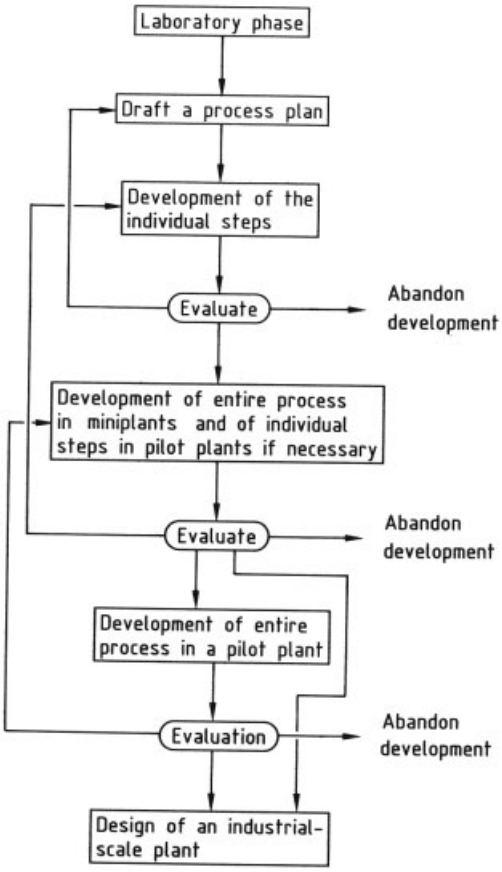

Figure 9. The cyclic pattern of process development

The most important task is to find the weak points and subject them to particularly close scrutiny. The entire process will then be examined again with the improved data obtained in this way, and so on. The fact that many decisions must be taken with incomplete knowledge is a fundamental but inevitable difficulty. To delay development until all uncertainties have been eliminated would be just as wrong as starting industrial development on the basis of laboratory discoveries alone.

A start should be made on choosing between various processes or process variants as early as possible so that consideration of a larger number of possibilities is restricted to the laboratory. Prolonged investigation of two variants on a trial-plant scale should be avoided.

Most mistakes are made at the beginning of the activity, but it is still relatively easy and cheap to eliminate them at the miniplant stage. However, the further process development advances, the more expensive it becomes to eliminate mistakes (see Fig. 10). In the final production plant, corrections can only be made with an enormous expenditure of time and money.

As shown in Figure 9, each development stage is followed by an evaluation to decide whether development should be continued, stopped, or started again at an earlier development level (see Section 4).

3.2. Drawing up an Initial Version of the Process

Once all the information has been assembled, an initial version of the process is drawn up [27]. In general, it has been found useful to adopt the following procedure:

After the initial versions of the process, which should be the most reasonable ones possible at that stage, have been drawn up, they should be discussed by the project team. This discussion should take account of all the information gathered hitherto and, after all the advantages and disadvantages have been reviewed, only one version of the process should remain.

There is at present no method which will result in the "best" version of the process being invariably chosen at the outset and there will presumably not be one in the future either [28], [29]. Finding the optimum version of a process requires not only knowledge and experience but also a considerable creativity, since the number of possible ways of carrying out a given task is almost infinite and many of the rules can still not be quantified [30].

Example 3.2 : The esterification of an impure organic acid with methanol yields a reaction mixture containing at least 5 components ($N = 5$): unreacted acid, unreacted alcohol, ester, volatiles such as H_2O, lower esters, dimethyl ether, and high-boiling components such as higher esters.

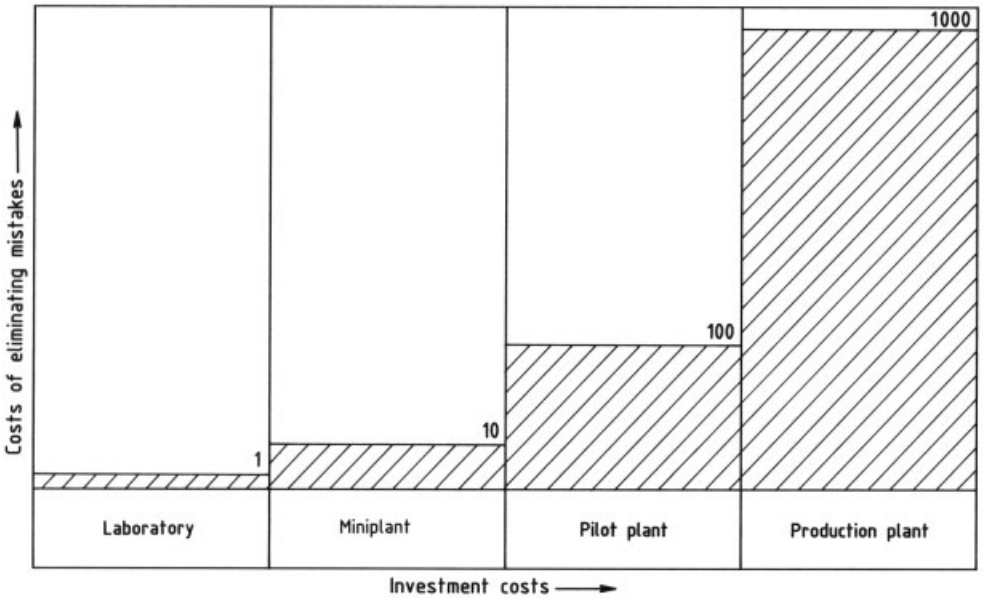

Figure 10. The costs of eliminating mistakes and the investment costs increase by a factor of 10 from one development stage to the next

According to STEPHANOPOLOUS [31] the possible number of arrangements for the separation operations can be calculated from the relationship:

$$Z = \frac{[2 \cdot (N-1)]!}{N! \cdot (N-1)!} \cdot S^{(N-1)} \tag{3.3}$$

where

Z = number of alternative arrangements,
N = number of components,
S = number of separation processes.

If all components can be separated by distillation, $(N-1) = 4$ separation columns are necessary, which can be arranged in 14 different ways. If a second separation process – for example, an extraction, if the ester, alcohol, and water form azeotropes – must be added, this already results in 224 possible arrangements. If the number of possible reaction procedures (e.g., batch/continuous; gas/liquid; ion exchanger/mineral acid) is also taken into consideration, there will be an almost infinite number of ways of carrying out the process (a so-called combinatorial explosion).

To find the best process among all these variants is quite impossible simply because the ideal solution is unknown. To calculate all possible variants, to compare them, and then select one is not feasible because of the large number. Many of the variants can of course be eliminated at the start for trivial reasons (industrial, economic, legal reasons, etc.). However, even the number of potential variants left is still so large that it is impossible to calculate investment and production costs for all of them. Furthermore, at this point the information available for doing this is often inadequate and there is consequently a risk of eliminating promising variants.

Perhaps this example reveals why inventing processes is still an art rather than a craft. However, the tools available for performing this art are being continuously improved [30], [32], [33] (see Section 3.2.1).

On the basis of this initial version, an industrial plant is designed. The individual unit operations (reactor, absorber, distillation, etc.) are designed on the basis of the existing information (approximate size and diameter of columns, etc.).

By scaling down this hypothetical large-scale plant a trial plant is designed, which is nowadays generally a miniplant (see Section 3.4.1). At the

same time, the industrial plant is simulated by using a computer program and an initial complete mass and energy balance sheet is drawn up.

3.2.1. Tools used in Drawing up the Initial Version of the Process

Requirements which must be met by modern EDP tools are as follows:

1) Fast access to the information
2) Information which can be conveniently handled at the workplace
3) The information must be continuously updated

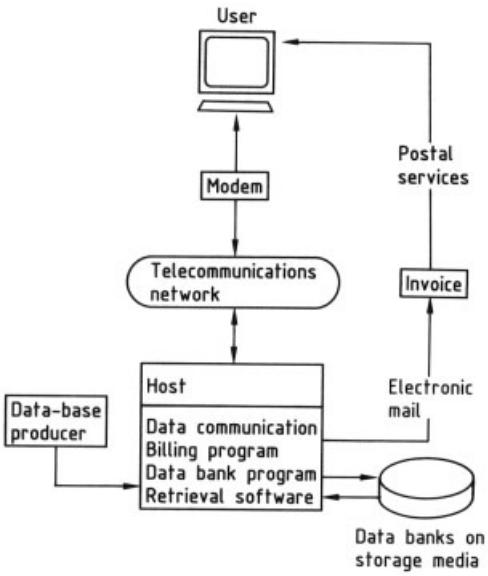

Figure 11. On-line service components [35]

3.2.1.1. Databases [34]

To carry out process development as efficiently as possible, it is important to avoid duplicate activities. This includes taking account of existing knowledge or exploiting it. For this reason, chemical companies not only build up internal databases but also utilize external databases. Figure 11 provides an overview of the structure of an on-line data bank.

Databases differ as to whether they contain the information itself (primary information or factual data) or whether they refer to other information sources (secondary information or literature data).

Some types of data which are important for process development are [36]:

1) Material properties [37]
2) Physicochemical data
3) Ecological and toxicological data
4) Costs of raw materials, intermediates, and end products
5) Energy and equipment costs
6) On-line literature searches

Table 2. A selection of simulation programs for steady-state processes

Program	Supplier
ASPEN PLUS	Aspen Technology Inc., 251 Vassar Street Cambridge, MA 02139 United States
CHEMASIM	BASF AG, Ludwigshafen, Germany
CHEMCAD	Chemstations Engineering Software COADE 10375 Richmond Ave., Suite 1225 Houston, TX 77042 United States
DESIGN II	Chemshare
FLOWPACK	ICI
HYSIM	Hydrotech. Ltd. 400, 119 14th Street N.W. Calgary, Alberta T2N1Z6 Canada
PROCESS	Simulation Science Inc.,
PRO/II	1 051 W Bastanchury Road, Fullerton, CA 92633 United States
VTPLAN (CONTI)	Bayer AG, Leverkusen, Germany

3.2.1.2. Simulation Programs [38]

Simulation using computer programs furnishes the process development engineer with a very effective tool. The development of very fast, large computers and effective physical models has now made it possible to simulate individual

units (e.g., distillation), networks of units of the same type (e.g., a stirred tank cascade, a heat exchanger network) and networks of dissimilar units (e.g., chemical plants).

In practice, the simulation program used depends on the objective; there is, after all, little point in trying to solve detailed problems with a simulation program which, although it is all-embracing, is slow if there are very much more efficient programs available which are designed for the specific requirement. Programs for solving specific problems are usually developed by the user himself and are not generally available or are available only in rare cases.

Numerous programs for dealing with flow sheet problems are available from universities, industry, or the commercial market (see Tables 2 and 3) [33].

Table 3. A selection of simulation programs for dynamic systems [39–42]

Program	Supplier
CHEMADYN	BASF AG, Ludwigshafen, Germany
DIVA	Prof. Dr. Gilles, University of Stuttgart, Germany
SATU	Höchst AG, Frankfurt, Germany
SIMUSOLV	The Dow Company Midland, Mi 48674, United States
SPEEDUP	Prosys Technology Ltd., Cambridge, England, (since 15.10.1991 Aspen Technology Inc.)

3.2.1.3. Expert Systems

The difference between simulation programs and expert systems is that in the former the modeling is rigidly embodied by an algorithm, while in the latter, knowledge of the model is stored in the knowledge base independent of the deductive mechanism [43–45].

An expert system is a computer program equipped with knowledge and ability which can carry out complex tasks by imitating human intelligence. Such systems are therefore particularly suitable for use in process development where complex problems are solved by drawing conclusions and using experimental knowledge. In this case, however, the knowledge does not only consist of facts which can be stored in databases but is acquired by carrying out process development, experiencing failures, being successful, repeating the same process, and learning at the same time; in short, by acquiring a feel for a problem. Knowledge is gained as to when it is necessary to stick to the rules and when they can be broken; a stock of practically proven, heuristic knowledge is built up [30].

Some examples of heuristic rules in process development are:

1) Carry out each reaction or separation only once (e.g., volatiles should only be removed once)
2) Remove substances which can be readily isolated immediately after they are produced and do not spread them over the entire working-up process
3) Design recycling paths so that, if possible, they only result in small feedbacks
4) If possible, use substances which are in any case present in the reaction output as auxiliary solvents since this may reduce the working-up costs considerably

Combinatorial problems such as drawing up an initial version of a process or developing a complex catalyst system result in a combinatorial explosion (see Section 3.2). Time does not allow all the possible solutions to be tested. The human expert copes with such a combinatorial explosion by eliminating those possibilities which seem to him to be unfruitful, concentrating on those which seem to be feasible and using heuristic knowledge which steers him towards the optimum solution but does not guarantee it.

The main components of an expert system are as follows (see also Fig. 12):

1) Knowledge base: heuristic knowledge is represented in this case by a system of symbols
2) Knowledge editor: a module which assists in changing, adding, or removing rules and formalizes their evaluation
3) The deductive mechanism, which combines knowledge and problem data by deriving further data from the rules stored in the knowledge base
4) input – output systems

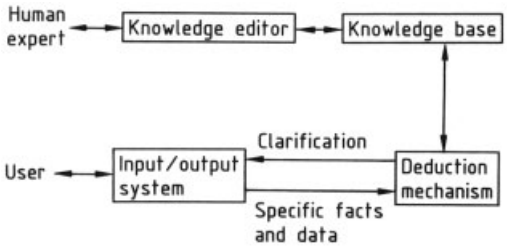

Figure 12. Structure of an expert system [46]

This technique was first applied in 1969 for describing and designing composite heat exchanger systems [47]. It was rule-oriented, i.e., the expert system involved a series of rules whose combination led to different results. The system was controlled by weighting factors, which were in turn dictated by success. The system "learned" through this method of changing weights and was therefore capable of yielding increasingly better results.

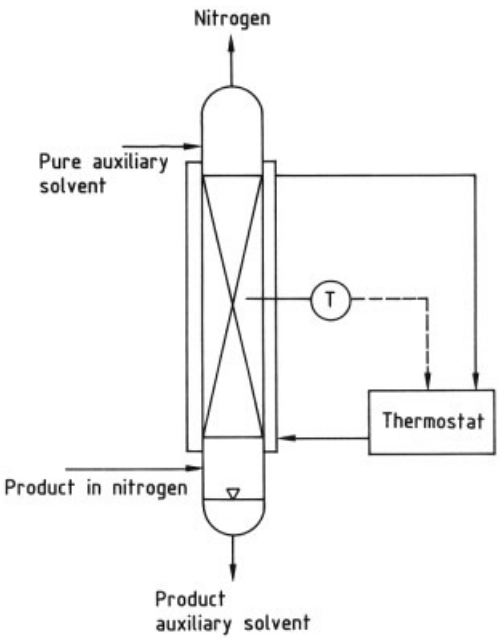

Figure 13. Checking the individual steps in the laboratory using pure feedstock with the absorber stage as an example (see Fig. 14)

The first expert systems were laboriously written in FORTRAN, but the technique acquired a new impetus in the 1980s when programming languages such as LISP (List Processor) were introduced. These are able to process symbols and symbol structures using a computer [48]. In addition a wide range of computer tools became available. Although these have made the development and handling of expert systems easier, they have contributed almost nothing to an understanding of the phenomena [49].

It is now beyond dispute that expert systems have justified themselves, although the euphoria of the 1970s has given way to a more sober assessment. The belief that a lack of knowledge and understanding can be overcome by an expert system has been found to be completely erroneous. On the contrary, an expert system can only be used efficiently by experts.

3.3. Checking the Individual Steps

Like the reaction step or steps, the chosen preparatory and separation steps are first checked individually and independently of one another on a laboratory scale. Even at this stage, new requirements may be imposed on the reaction stage. For example, difficulty in removing a byproduct may make it necessary to alter the reaction conditions or modify the catalyst.

Along with the experimental work, e.g., testing an absorber (see Fig. 13), thermodynamic simulation calculations are carried out for the unit concerned.

Laboratory experiments are generally carried out with pure, well-defined materials. Thus, the solvent selected for the absorption is used as received from the chemical stores. This differs, of course, from the subsequent situation, where the solvent must be recycled for cost reasons and inevitably becomes enriched with byproducts, some of which are unknown, however, at this point in time.

The feedstock used is a synthetic mixture of gases prepared from pure materials which also does not contain byproducts. The experiment shows whether it is physically possible to absorb the useful product from the gas stream with reasonable amounts of solvent. If the absorber can be simulated with the available material data in a calculation run in parallel, work on this step will be complete. The unit is then another piece which can be fitted into the jigsaw puzzle of the

entire process, which is the next item which must be considered in an integrated trial plant.

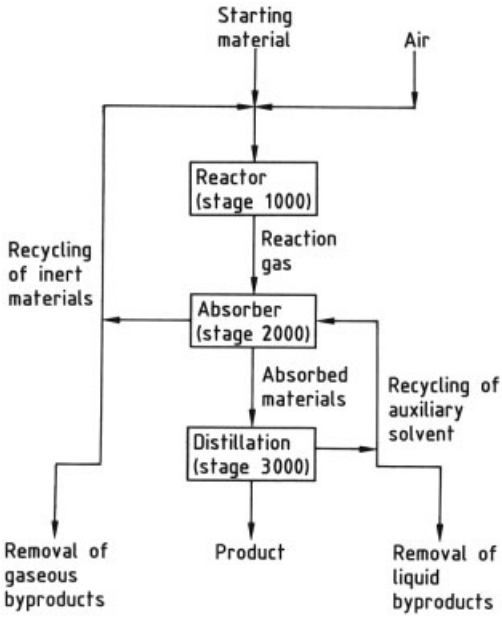

Figure 14. An initial version of a process for oxidizing a starting material with air and then isolating the product by absorption and distillation

Once all the individual steps have been successfully tested, it is possible to draw up a reliable flow sheet of the entire process (see Fig. 15). However, if any subsidiary step is found not to be feasible at this stage, the process concept must be changed.

This preliminary flow sheet, which has not as yet been fully tested as a whole, can be used to draw up an initial rough cost estimate (see Section 4).

3.4. Testing the Entire Process on a Small Scale

As Figure 15 shows, when the individual steps are put together, recycling streams are created (in this case one gas and one solvent recycling path). These recycling streams or feedbacks are an economic necessity, but they raise new process engineering problems. Thus, material recycled to the reactor may drastically affect the activity and service life of the catalyst. In addition, the solvent circuit may become enriched in byproducts and this may lower product purity.

In principle, these problems can only be solved by mathematical simulation, but since many of the quantities that are required for a mathematical description are unknown, an experimental approach must be adopted.

Example 3.3. When a plant which includes recycling streams (see Fig. 15) is operated, the solvent circuit becomes enriched in a high-boiling substance whose existence was not previously known because it was below the analytical limit of detection in the reaction product. This high-boiling substance also has the troublesome property of depositing when it reaches a certain concentration in the solvent and forms a coating in the column.

Since the high-boiling substance was not known before, it was not possible to include it in the mathematical simulation. Now that it has been discovered as a result of operating an integrated plant, it can be characterized and incorporated in the simulation.

3.4.1. The Miniplant Technique

3.4.1.1. Introduction

Improvement in computer programs for modeling processes has resulted in the integrated miniplant technology acquiring ever increasing importance because the synergism between the miniplant technique and mathematical simulation means that the margin of safety in scaling up is as good as that obtained by setting up a pilot plant [50–52], [53, pp. 179–182].

The miniplant technique has the following characteristics:

1) The trial plant includes all the recycling paths (it is a so-called integrated miniplant) and it can consequently be extrapolated with a high degree of reliability.
2) The components used (columns, pumps, condensers, pipelines) are often the same as those used in the laboratory. In addition, they are largely reusable standardized components which have been tried and tested in continuous operation. For these reasons, investment costs are low and flexibility is high.

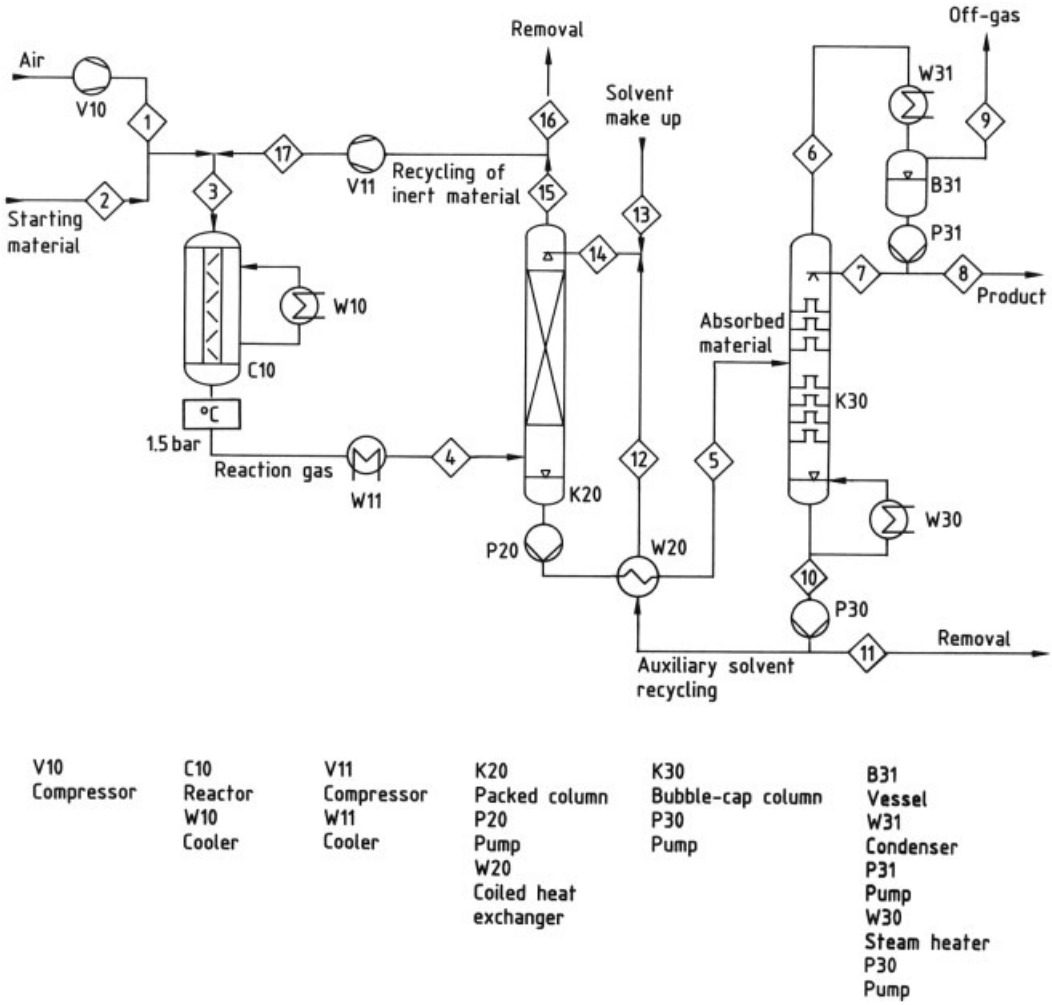

Figure 15. Example of a simplified process flow chart produced after testing the individual steps

3) The plant is operated round the clock on a shift basis for weeks and it therefore must be as fully automated as possible to keep operating costs low, but not, however, at the expense of flexibility.
   The measuring and control instruments used are standard components which are nowadays generally connected to a small process control system [54, pp. 183 – 186], thus making it possible to carry out modifications in the measurement and control rapidly.
4) The entire plant is normally set up in an extracted chamber so that explosion-proof operation is possible and safety-at-work requirements can be met more easily.

3.4.1.2. Construction

The following design documents are required as a minimum for constructing a miniplant:

1) A process flow sheet showing all the quantity flows in the miniplant and the associated temperature and pressure conditions
2) Engineering flow diagram
   Required information:
   a) All equipment and machines
   b) Internal diameters and pressure ratings of pipelines and construction material

c) The objectives of the instrumentation
d) Information on the insulation of the equipment and pipelines

3) List of measuring points (measuring point number, specification, parts list, point of installation)
4) Safety concept, i.e., how the sensors (e.g., for temperature, pressure, flow) are linked to the actuators (e.g., valves, pumps) by the safety logic

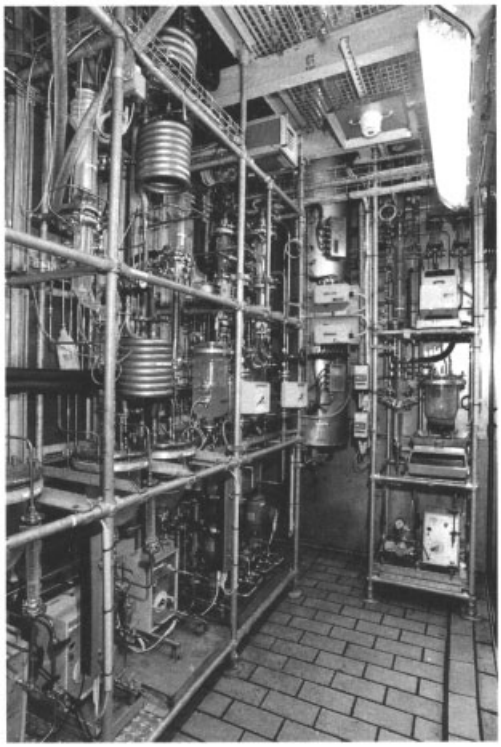

Figure 16. Example of a miniplant

Specification of Equipment Size. In a miniplant, the hold-up for a given throughput is almost always greater than in an industrial-scale plant. The size of the vessels and column bottoms should therefore be minimized, otherwise it will take too long for the miniplant to reach a steady state because of long residence times. In the column bottoms, which are particularly critical, a hold up of 100 – 200 mL is normally the lower limit. Below this level, a laboratory evaporator can usually no longer be operated efficiently (because of excessive temperature differences, foaming, etc.).

Example 3.4. A miniplant produces a high-boiling residue at a rate of 1 g/h. This residue is removed from the bottom of a distillation column with a distillation boiler capacity of 200 mL, so that the mean residence time is 200 h. The time required to reach the steady state is consequently about 25 d (about 3 times the mean residence time). Only then will it be possible to assess this bottom residue and draw conclusions about its fouling behavior, composition, properties, etc.

Specification of Material Balance. In Section 3.2 it was shown that a miniplant should be designed by scaling down a hypothetical large-scale plant. The scale-down factor is determined by finding the points in the process where the smallest and largest quantity flows occur. The levels at which these flows can be handled determine the scale-down factor.

Normally only flows of $\geqq$ 1 g/h and $\leqq$ 10 kg/h can be handled in a miniplant. At the same time, the minimum flow is primarily determined by the time required to reach a steady state (see Example 3.4), while the upper limit of the flow is set by the maximum hydrodynamic load of the laboratory column used.

Example 3.5. In a hypothetical large-scale plant producing 5 t/h, the minimum flow is 50 kg/h of a high-boiling substance which must be removed and incinerated, while the maximum flow is 100 t/h of the reaction throughput, which contains the useful product as a 5 % aqueous solution.

Scale-down factor:
for the minimum flow: 50 kg/h ÷ 1 g/h = 50 000,
for the maximum flow: 100 t/h ÷ 10 kg/h = 10 000.

Thus the miniplant should be designed to produce not less than 100 g or not more than 500 g of useful product per hour.

Specification of Type of Equipment. Unlike normal laboratory experiments in which the equipment operates only during the daytime, miniplants must run continuously over a period of weeks to be able to fulfil their task. This is due to the fact that the time required to reach

a steady state is often long and because it is only possible to draw reliable conclusions about long-term effects such as corrosion, fouling, catalyst deactivation after a long time (1 week to 6 months).

Unlike laboratory plants, the requirement imposed on the reliability of miniplant components are therefore very high and almost the same as for a large-scale plant. Occasionally this sets very tight restraints on choosing equipment and in some cases necessitates in-house development if nothing suitable is available on the market.

3.4.1.3. The Limits of Miniaturization

Table 4 lists the minimum dimensions for which it is possible to scale-up a process.

Table 4. Examples of tried and tested miniaturization limits [51]

Example	Tried and tested limit
Columns	30 mm diameter packed columns
	35 mm diameter columns with structured packing, dual-flow plate columns
	50 mm diameter bubble-cap columns
Pumps	1 mL/h injection pumps
	10 mL/h piston pumps
	100 L/h rotary pumps
Belt filter	50 mm belt width
Centrifuges	delivery 5 L/h
Pipelines	1.5 mm diameter, metal or Teflon

Columns can still be operated effectively down to 35 mm diameter when filled with an ordered packing. For plate columns having one bubble cap per plate, the limit is ca. 50 mm. At still smaller dimensions, wall effects are difficult to suppress.

Experimental equipment for handling solids or dealing with extreme operating conditions is often not available. This often necessitates in-house developments and cooperation with specialist firms, university institutes, and research companies.

Expenditure on instrumentation is often very much lower for a miniplant than for the corresponding large-scale plant. For this to be the case, however, most of the miniplant components must be made of glass and therefore operate at low-pressure, thus making it much easier to monitor liquid levels, deposits, foam formation, etc., than in the large-scale plant. SCHULZ-WALZ has given examples of measuring and metering techniques for miniplants [55]. In the last two decades, important advances have been made in sensor technology and, along with electronics and processor technology, this has considerably simplified the miniaturization of sensors for miniplants (see Table 5).

Table 5. Established miniaturization limits for instrumentation components [51]

Measured quantity	Principle of measurement	Established minimum measurement range
Gas flow	thermal	0.02 – 0.6 L/h
	volumetric	2 – 200 L/h
Liquid flow	thermal	2 – 30 g/h
	Coriolis force	0.07 – 1.5 kg/h
	magnetic inductance	0.6 – 6 L/h
	volumetric	0.1 – 1 L/h for piston meters
		2 – 40 L/h for oval disk meters
		1 mL/h upwards for metering pumps
	gravimetric	2 – 50 g/h (balance range 0 – 5000 g)
Temperature	resistance change	$T_{\min}$ range 30°C (maximum range −200 to 600°C)
	thermoelectric voltage	$T_{\min}$ range 50°C (maximum range 0 – 1000°C)
Level	conductive	>50 mm
	capacitive	>50 mm
	hydrostatic	>4 mbar
Differential pressure	strain gauge	0 – 1.25 mbar
	piezoresistive	
	capacitive	

There are four stages in the automation of a miniplant:

Stage 1. At this stage, the electrical instrumentation of the miniplant consists of compact, self-contained individual components, e.g., balances, metering controls, metering pumps, thermostats, temperature controllers, pH sensors, individual controllers, control valves, pressure sensors, and plotters. Monitoring and operation is the responsibility of the shift personnel.

Stage 2. At this stage the structure of the sensors, actuators, and individual components is the same as in Stage 1, but the individual components may be connected to a data processing

computer via a data interface. Monitoring and operation is again the responsibility of the shift personnel.

Stage 3. Here the system for logging the measured data is the same as in Stage 2, but the lower-ranking self-contained individual components are controlled and provided with set points by a higher-ranking automation software, so that shift personnel are able to concentrate on monitoring tasks.

Stage 4. At this stage the miniplant is equipped with a compact process control system. While the actuator and sensor system is the same as at Stage 1, all the other functions such as controlling, measuring, regulating, calculating, processing data, optimizing, keeping records, alerting, logging, operating, and observing are performed by the process control system.

The choice of automation stage is decided during the miniplant design phase on the basis of the specific plant requirements and boundary conditions as well as on the results of trials. The higher automation Stages 3 and 4 are only justified if a miniplant is to be operated over a long period of time.

3.4.1.4. The Limitations of the Miniplant Technique

A disadvantage of the miniplant approach emerges if individual steps in the process are subject to an unduly large risk in extrapolating from the miniplant to the industrial-scale plant. For example, scale-up factors of ca. 10^4 are still not feasible for extraction and crystallization steps even today. This disadvantage can often be circumvented by isolating critical sections of the process and working on them in an intermediate stage (pilot stage). If enough of the unmodified product can then be collected from an integrated miniplant at an acceptable cost for it to be possible to operate, e.g., an extraction column or a crystallizing plant on a pilot scale with a representative feed for a sufficiently long time, the entire plant can again be extrapolated upwards with a calculable risk. This may obviate the need for constructing a cost-intensive and time-consuming pilot plant for the entire process [56].

3.4.2. Pilot Plant

It may be necessary to construct a pilot plant if one of the following conditions applies:

1) The scaling-up risk is too large to proceed directly from the miniplant to the industrial-scale plant. The reasons for this may be:
   a) The process involves several critical stages (e.g., handling of solids) which cannot be described by physical models
   b) a difficult or completely novel technology is being used
2) It is necessary to provide representative product quantities, e.g., for the market launch, and these cannot be produced by the miniplant in a reasonable time

The operation of the pilot plant should clarify all the issues which have not been fully dealt with in the miniplant.

These may include, for example [56], [57]:

1) Checking design calculations
2) Solving scale-up problems
3) Checking experimental results obtained with the miniplant
4) Measuring the true temperature profiles in the reactors and columns under adiabatic conditions
5) Gaining process know-how (dynamic behavior of the plant, start-up and shut-down procedures)
6) Producing representative sample quantities in fairly large amounts
7) Precisely assessing fairly small flows (e.g., residues)
8) Training the personnel who are to run the plant
9) Improving the estimate of the expected service life
10) Carrying out material tests under realistic conditions

A pilot plant should be designed as a scaled-down version of the industrial-scale plant and not as a larger copy of the existing miniplant [58]. Nowadays it costs almost as much to design, construct, and operate a pilot plant as an industrial-scale plant.

The decision to build a pilot plant results in considerable costs. As a rule of thumb, the initial investment is at least 10 % of the investment costs of the subsequent industrial-scale

plant. Moreover, if they use particularly toxic substances, pilot plants now often require approval, and this may considerably lengthen the time required to put them into operation.

Once process development on a pilot scale has been successfully concluded, the pilot plant must be kept on stand-by until the industrial-scale plant is running satisfactorily. Normally, when the larger plant is started up, the pilot plant is operated simultaneously so that any problems which occur in the former can be dealt with rapidly.

4. Process Evaluation

After each development stage (see Section 3.1) the existing knowledge should be documented and the status of the process evaluated. This involves preparing study reports and is now standard in many enterprises.

At this stage it is necessary to answer the question posed in Section 1 as to whether the process is "better" and "cheaper" than that of the competitors.

4.1. Preparation of Study Reports

The study report documents knowledge about the status of the development of a process after specified time intervals or specified development stages to provide the basis for making a decision. The objective of the study report is to answer three important questions:

1) Can the production process be implemented in this way in principle?
2) What is the return on investment?
3) How big is the risk in economic and technological terms?

At least three project study reports must be completed in the course of the development of a process:

1) One at the beginning of process development
2) One in the course of working out the details of the process
3) One at the end of process development

A study report should include the following items:

1) A summary
2) A basic flow diagram
3) A process flow diagram and a description of the process
4) A waste-disposal flow diagram
5) An estimate of the capital expenditure
6) A calculation of the production costs
7) A technology evaluation
8) The level to which the experimental details have been worked out

4.1.1. The Summary

The summary must start with a precise definition of the objective. This is followed by a short description of the process and of the most important process stages.

The chemistry of the process should be represented by overall reaction equations, with the assumed conversion and selectivity being specified.

Details of the result of the study, the production costs, the capital expenditure for a specified annual production output with an indication of the accuracy of the estimate and the planned location of the plant, if known, are other important points.

The summary concludes with information on the state of knowledge and on the major risks.

4.1.2. Basic Flow Diagram

The basic flow diagram provides a quick overview of the total process. It shows the input, output, and recycling streams and identifies the individual process stages (see Fig. 14).

4.1.3. Process Description and Flow Diagram

All the individual steps of the process should be described as accurately as possible on the basis of all the information available at the time, reference being made to the process flow diagram (see Fig. 15).

The latter should clearly show the process in detail and the composition of the flows.

The following information should be included in the process flow diagram:

1) All equipment and machines (columns, vessels, heat exchangers, reactors, etc.)

Table 6. Example of a material balance table

Material number	Name of material	Molecular mass	Flow no. 1 kg/h	%	Flow no. 2 m^3/h	vol %	...
1	A		1000	66.7			
2	B		500	33.3			
3	C				2000		
4	D				200		
$\sum^{kg/h}_{m^3/h}$			1500		2200		
State			liquid		gaseous		
Pressure, bar			4		1.2		
Temperature, °C			102		220		

2) All product streams (with numbering)

Beneath the process flow diagram the material balance should be presented in a clearly laid out table (see Table 6).

When development is at an advanced stage or is complete, the process flow diagram should show the following additional information:

1) Identification of the energy carriers (type of steam, cooling water, compressed air, electric power, nitrogen, deionized water, heating gas, refrigerants, etc.)
2) Details on the construction materials, size and power of the equipment and machines
3) Typical operating conditions (e.g., pressure and temperature in columns and pipelines)
4) Important fittings
5) Purpose of instrumentation

4.1.4. Waste-Disposal Flow Diagram

The waste-disposal flow diagram illustrates the waste streams for which there is no further economic use in the plant (see Section 1.3) and indicate their flow numbers from the process flow diagram. Precise information should be given on the toxicity, degradability, water danger, flash point, ignition temperature, MAK, odor threshold, etc. of the individual components, their mass flows, and the state of aggregation of the flow (viscosity, sediment content, etc.). This information makes it possible to specify the type of waste disposal required, for example:

1) Sewage treatment plant
2) Incineration in a power station or an incineration plant
3) Central hazardous waste incineration
4) Waste-disposal site

4.1.5. Estimation of Capital Expenditure

4.1.5.1. Introduction

An important reason for preparing study reports is to determine the capital expenditure (investment) of a project [59], [60]. This is made up of the independent location ISBL (inside battery limits) process plant costs such as investment in the production plant and the associated control rooms, laboratories, employee facilities, tank farms, and loading and unloading stations, as well as the location-dependent OSBL (outside battery limits) costs of all the service facilities, such as investment in storage buildings, cooling towers, and waste-disposal facilities. Infrastructure costs are also to be included (see Fig. 17).

The accuracy of the investment cost estimate depends critically on the level of maturity the process has reached [62].

4.1.5.2. ISBL Investment Costs

Of the many methods for determining the capital required to set up a plant, a method of estimation which derives the costs of a new plant from the known costs of existing plants with an accuracy of $\pm 30\,\%$ is now firmly established (the additional-cost calculation method). Such a method is only possible because the structure of chemical plants is always very similar, i.e.,

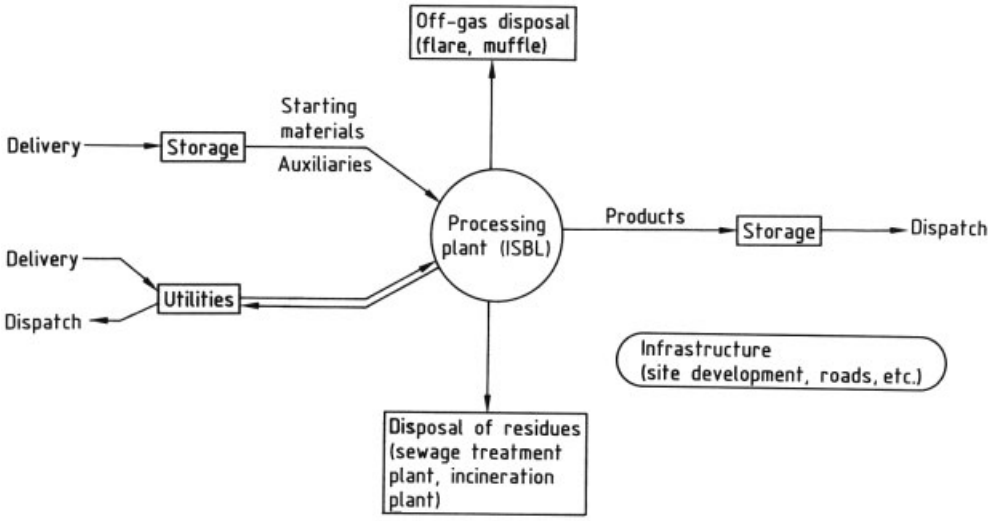

Figure 17. Subdivision of a project into ISBL, OSBL, and infrastructure activities [61]

they are made up of interlinked pieces of equipment of similar construction whose number does not vary widely. For this reason, the overall investment cost of the entire plant when installed exceeds the average value of the machines and equipment (= total machine and equipment costs divided by the number of machines and pieces of equipment) by a factor which depends on the capacity.

The total machine and equipment costs are determined from the parts list, while the costs of individual units can be obtained from in-house databases or directly from the manufacturers.

Example 4.1 [59]. The overall factor for the chemical plant construction projects handled by BASF in 1984 was on average 3.9; i.e., the direct plant costs of an average chemical plant is 3.9 times the machine and equipment costs.

The overall factor is essentially determined by the following items:

1) Operating conditions (pressure, temperature)
2) Type of construction materials used (steel, stainless steel, special materials)
3) Size of machines and pieces of equipment

According to Miller [63], the effect these items have on the overall factor can be determined from the average value of the machines and equipment (see Fig. 18).

This results in the following simple relationship:

$$\text{direct plant costs} = \text{overall factor} \times \Sigma\,(\text{values of machines and items of equipment})\,. \quad (4.1)$$

Since direct plant costs are always estimated for some date in the future, it is necessary to determine the effect of increases in price. Use is therefore made of price indexes which are published for chemical plants in the individual countries (see Table 7). The price index i determined by extrapolation is divided by the index i_0 at the date of the cost determination. The ratio i/i_0 is then a measure of the increase in investment costs to be expected.

The direct plant costs make no allowance for the cost of engineers (usually 10 – 20 %) or for contingencies (ca. 5 – 10 %).

This method of estimation has the disadvantage that while the machine and equipment costs can be determined fairly precisely, the average value of these items is multiplied by a comparatively inexact overall factor. The accuracy can be considerably increased by breaking the overall factor down into individual factors.

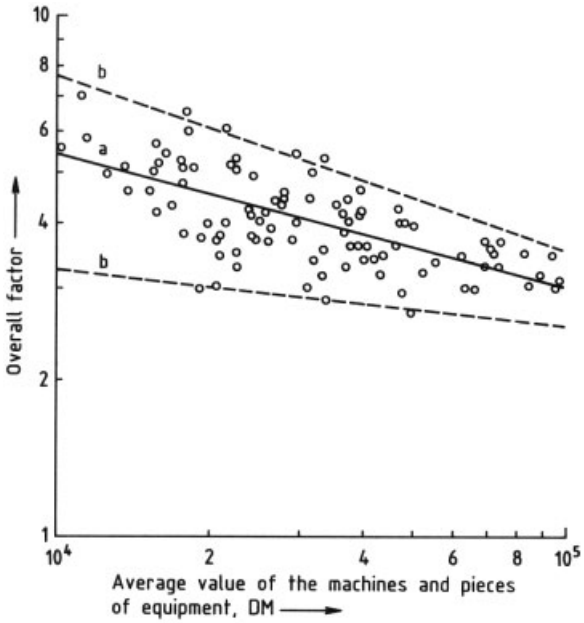

Figure 18. Overall factor for the direct plant costs as a function of the average value of the machines and pieces of equipment in 1984 [59]:
a) Best-fit curve; b) Confidence limits for a statistical confidence interval of 95 %

Table 7. Price indexes i for chemical plants in Germany , [64, p. 280]

Year	*i*
1976	100
1980	125.3
1981	132.9
1982	140.5
1983	144.4
1984	147.2
1985	150.4
1986	154.0
1987	159.6
1988	164.7
1989	170.9

4.1.5.3. OSBL Investment Costs

The OSBL investment depends on the location of the plant. It includes, for example, the cost of storage buildings, cooling towers, waste-disposal facilities, linking the OSBL facilities to the location in question (e.g., water and electricity supply), etc. Since the question of location is generally still unresolved at an early stage in development, it is not possible to specify the OSBL costs explicitly but they can be tentatively taken as 20 – 30 % of the ISBL costs. For small plants the percentage tends to lie near the upper limit, while for larger plants it is likely to be close to the lower limit.

If the location is known, the OSBL costs can be estimated very accurately.

4.1.5.4. Infrastructure Costs

The infrastructure costs depend on the location of the plant. They are not, however, related exclusively to the project since the infrastructure, such as workshops, the central store, and the canal system, is also used by other production facilities. Obviously, this type of cost can only be specified if the precise position of the plant at a location is known.

4.1.6. Calculation of Production Costs

The cost effectiveness of a new process depends on the production costs of a product (unit of currency divided by unit of quantity) (see Section 4.2). The following information is required to determine these costs:

1) Material balance

2) Waste-disposal flow diagram
3) Utilities
4) Investment

The production costs can be calculated by adding up the following items:

1) Feedstock costs (raw materials and auxiliaries)
2) Manufacturing costs, which can be subdivided into:
   a) Utility costs
   b) Waste-disposal costs
   c) Staff costs (wages and salaries)
   d) Maintainance
   e) Miscellaneous costs (concern overheads)
3) Capital-dependent costs (depreciation, interest, insurance taxes)

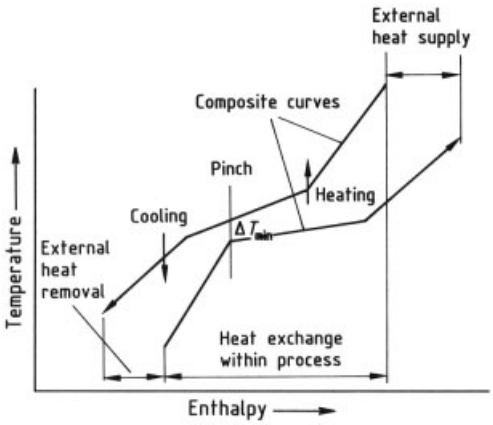

Figure 19. Temperature – enthalpy diagram

4.1.6.1. Feedstock Costs

The raw material requirements are given by the material balance for the chosen capacity, but to determine the feedstock costs the prices of the raw materials are required. The change in raw material prices while the process is being developed can be estimated, but it is more difficult to allow for the fact that the erection of the plant will have an impact on the raw material market and consequently alter the price structure. Therefore, in some cases it is necessary to enter into negotiations with the raw material suppliers at an early stage. If the raw materials originate from another plant in the same concern, market prices are replaced by transfer prices [65], but this does not overcome the problem of the impact of the proposed plant on the market – it only shifts it. The significance of raw material costs in cost estimation varies widely. Thus, in processes involving a high level of material refinement (as is the case for some pharmaceuticals) or in those with high specific energy costs (e.g., chlorine production), precise determination is not as important as in the case of cheap, mass-produced products, such as petrochemicals, manufactured in large plants.

Credits for byproducts can be set against the raw material costs from the outset.

4.1.6.2. Utilities

For large-scale processes, the cost effectiveness depends appreciably on the cost of the utilities. The combination of utilities required in the plant therefore acquires considerable importance and consideration must be given to the ideal thermal coupling even at an early stage in the process development since it can have a considerable impact on the configuration of the process [66]. For example, whether the energy derived from condensing vapors can be utilized depends on the behavior of the materials. Clean and stable condensates usually present no problem in using the heat of condensation to raise the temperature of cold streams, but in the case of products which tend to form deposits (by decomposition, polymerization, etc.) optimum thermal utilization of this type of energy is not possible since the vapors must be quenched to prevent deposits and this results in loss of energy. This type of question should be resolved in the miniplant phase.

A suitable aid for systematizing such considerations is Linhoff analysis [67], [68] (see → Pinch Technology). The Linhoff analysis determines the process configuration for which the energy supplied from or lost to the surroundings is a minimum. The problem which must be solved is how to couple the heat sinks and heat sources within the plant such that minimum energy consumption is achieved. The initial step in the calculation is to determine the so-called composite curves and pinch temperature, which is the temperature above which heat can be lost and below which heat can be absorbed (see Fig. 19).

For a given interconnection system, the Linhoff analysis shows how far the heat utilization

is from being ideal. However, there are often reasons for not choosing an ideal interconnection:

1) Start-up operations often require start-up heat exchangers, and this means higher investment costs
2) Energy utilization only becomes possible as a result of increasing the column pressure, and this means higher investment costs and may result in material problems (decomposition, side reactions, etc.)
3) Problems associated with the formation of deposits during heat transfer, which necessitates direct heat removal (quenching)

Generally, as the plant capacity increases, reducing the operating utilities (i.e., optimizing the thermal combination) rather than increasing the investment will have a more beneficial effect on production costs.

The energy analysis is used to find the amounts of energy required or released and to calculate the energy per unit quantity. The determination of energy costs requires information on the prices of the different types of energy, which depend on the quantities required and the location.

Steam is the energy carrier most widely used for heat columns because it offers the following advantages:

1) It is inexpensive since it can often be generated directly in the synthesis section of the plant
2) There are no explosion hazards provided the hot pipe surfaces are properly insulated
3) The temperature range covered is large (see Table 8)

Table 8. Enthalpy and temperature of some frequently used steam pressures [69]

Gauge pressure, p, bar	Temperature, °C	Heat of vaporization ∗, kWh/t
0	100	627
1.5	127	605
4	153	585
16	204	534
25	226	508
40	252	477
100	312	361

∗ With indirect heating, only the difference between the incoming stream and the discharged condensate can be utilized.

An initial estimate of the steam requirement of a column is given by:

$$\mathrm{t/h\,steam} = 1/5\,(\mathrm{reflux} + \mathrm{head\,production})\ \mathrm{t/h} \quad (4.2)$$

This is based on the assumption that:

1) The thermal combination is optimum (e.g., the hot discharge from the bottom is used to heat the feed)
2) Normal organic liquids are involved (heats of evaporation ca. 110 kWh/t)

Electrical energy is more than twice as expensive as steam, and because of the explosion hazard it is normally not suitable for heating columns. Its most important applications are in feeding material (pumps, impellers, compressors) and in carrying out operations such as size reduction, mixing, and separation.

The pump delivery N required can be estimated using the following formula [70]:

$$N/\mathrm{kW} = \frac{\dot{V} \cdot H \cdot \varrho \cdot g}{3600 \cdot 10^3 \cdot \eta_{\mathrm{tot}}} \quad (4.3)$$

where

$\dot{V}$ = the useful feed flow in m^3/h,
H = delivery head in m,
ϱ = density in kg/m^3,
g = acceleration due to gravity (10 m/s^2),
η_{tot} = overall efficiency.

The overall efficiency is made up of the mechanical and the electrical efficiencies:

$$\eta_{\mathrm{tot}} = \eta_{\mathrm{mech}} \cdot \eta_{\mathrm{electr.}} \quad (4.4)$$

For rotary pumps, the overall efficiency is ca. 0.7 – 0.8 while for piston pumps it is 0.8 – 0.9. For small pumps ($< 5\ m^3$/h), the overall efficiency may, however, be much lower.

Another application for electrical energy in chemical plants is in auxiliary electrical heating systems (e.g., frost protection).

Fuels are only used in chemical plants if streams must be heated to such high temperatures that steam is no longer economical (the limit is ca. 100 bar of steam, i.e., 310°C). However, the use of fuels in chemical plants always entails a risk from the ignition source.

Only the net calorific value of the fuel can be used since the steam formed can generally not be condensed.

A few typical values of the standard enthalpy of combustion are given in Table 9.

Table 9. Standard enthalpy of combustion at 25°C [71]

Substance	kWh/t
Acetaldehyde	7 360
Acetic acid	4 040
Carbon	9 110
Carbon monoxide	2 810
Ethane	14 440
Ethanol	8 250
Formaldehyde	5 190
Formic acid	1 590
Hydrogen	39 740
Methane	15 460
Methanol	6 310

Cooling water can be drawn from a body of surface water or from the groundwater [72]. Compared with surface water, brackish water or seawater has the advantage of a more constant temperature and a lower infestation with algae, but requires special measures to be taken against corrosion. A recirculating procedure (Fig. 20) is preferred to a once-through system, with only sufficient cold river water being used for topping up as is required to maintain the specified temperature (the advantages of this are that economies in river water are possible in the cold season, the minimum temperature can be kept constant, and high flow rates in the circuit prevent corrosion and fouling).

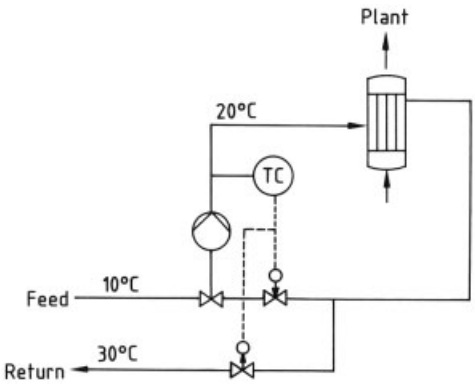

Figure 20. Example of a recirculation system for cooling a chemical plant with river water

The cooling water required to remove a given quantity of heat depends on the temperature of the water and its quality. If more detailed information is not available, it is sufficient as a first approximation to determine the amount of cooling water required by assuming its temperature rises by 20°C.

For high product temperatures, an alternative to water cooling is to use air coolers. Although these require a higher investment than water-cooled condensers, they are often cheaper to operate and are less prone to fouling.

However, unless there is a shortage of cooling water at a known location, it is generally sufficient to assume water cooling in an initial cost effectiveness calculation.

Refrigeration Energy. If temperatures below approximately 25 – 30°C must be achieved in a process, energy must be expended on refrigeration. Refrigeration is generally achieved by compressing, cooling, and adiabatically expanding coolants in the plant itself; i.e., the refrigeration requirement can be expressed as a requirement for electrical energy and cooling water. To remove a given quantity of heat, about 20 – 50 % of that quantity, depending on the temperature level required, must additionally be expended in the form of electrical energy, thereby increasing the amount of heat which must be removed to about 120 – 150 %.

Compressed Air. Small quantities of compressed air are drawn from pipelines. An example is the control air required by closed-loop control instruments (the rule of thumb is 1 m^3/h (STP) of control air per instrument).

Larger amounts are produced by an air compressor in the plant and the requirement can then be expressed in electrical energy [73]. The power N expended on compressing gases is given by:

$$N_{\text{theo}} = P_1 \cdot \dot{V}_1 \cdot \frac{n}{n-1} \cdot \left\{ \left(\frac{P_2}{P_1} \right)^{\frac{n-1}{n}} - 1 \right\} \tag{4.5}$$

where

n $= c_p/c_v$,

P_1 or P_2 = initial or final pressure, respectively,

c_p or c_v = heat capacity of the gas at constant pressure or constant volume,

$\dot{V}_1$ = initial volumetric flow.

For air compression ($n = 1.4$, overall efficiency $\eta \approx 0.7$):

$$N/\mathrm{kW} = 0.14 \cdot \dot{V}_1 / \left(m^3/\mathrm{h}\right) \cdot \left\{ (P_2/\mathrm{bar})^{0.286} - 1 \right\} \quad (4.6)$$

4.1.6.3. Waste-Disposal Costs

The waste-disposal situation has become much more difficult for chemical plants, a trend which is continuing. The delivery of effluent and wastes to central sewage plants, waste incineration plants, or waste-disposal sites is restricted by official regulations, and as a result waste-disposal costs have risen dramatically in Germany in recent years (see Fig. 21).

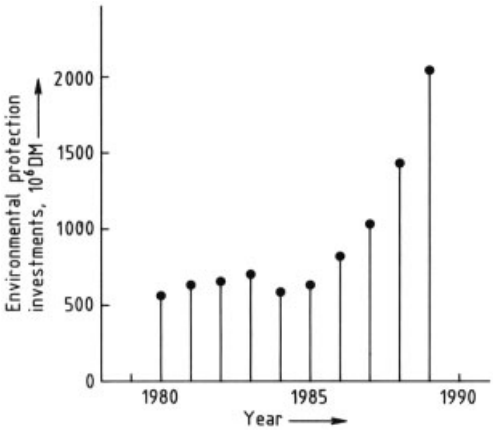

Figure 21. Expenditure on environmental protection investment in the chemical industry in Germany [74, p. 303]

This poses a challenge to process development. The slogan "built-in rather than add-on environmental protection" expresses the current demand for environmental pollution to be reduced in the manufacturing process itself and for efforts to be made to put even the byproducts of a chemical process to good use if possible [75–78].

Sewage Treatment Plant. Effluents discharged into a sewage treatment plant must be readily capable of biological degradation and must not be toxic to the bacteria in the plant. In addition, each substance in an effluent stream should have a sufficiently high biological degradation rate, regardless of its total content in the stream. This requirement prevents individual undegradable substances being lost in an otherwise readily degradable mixture. To assess whether disposal of a stream via a sewage treatment plant is possible, the following information must be collected for each substances it contains:

1) BOD_5 and COD value [79]
2) Zahn – Wellens test [80]
3) Bacteria and fish toxicity [81–84]
4) Water hazard class [85]

Incineration Plants [86, pp. 213 – 217]. Waste streams which consist only of compounds containing carbon, hydrogen, and oxygen as aqueous solutions with a concentration >10 % are ideally suitable for disposal in an incineration plant or a power station (if the content is >10 %, the heat of incineration is roughly equal to the heat of evaporation of the water).

Disposing of waste by incineration is more difficult if other elements such as nitrogen, chlorine, sulfur, and metals are present. Such wastes can only be disposed of economically in a central waste incineration plant equipped with suitable absorption systems for NO_x, HCl, SO_2, etc. [87], [88].

Waste-Disposal Site. The restrictions imposed on wastes which can be dumped are continuously increasing, and new dumping space is no longer available. Consequently, this waste-disposal option is virtually no longer available for new processes.

4.1.6.4. Staff Costs

Operating staff costs in chemical plants can vary widely. For large automated plants, they may be less than 3 % but for small batch processes they may amount to 25 % of the manufacturing costs. Often they are 5 – 10 % of the manufacturing costs.

In new plants, whether batch-operated or of the single-train type, the trend is towards the fullest possible automation. The introduction of process control engineering has made it possible to operate and monitor virtually every aspect of a plant using pictorial displays [89, pp. 158 –162].

Almost regardless of the size of the plant, a well-planned and automated single-train installation requires the following minimum staff:

1) One shift foreman or his deputy
2) Two shift employees for the control room
3) One shift employee for outside operations
4) One shift employee as a reserve

This minimum staff of five should be multiplied by a factor, specific to the company, which reflects the type of shift model, the sickness statistics, and vacations to give the total number of employees required. Special operating requirements such as frequent changes in production schedule, periodic start-up and shut-down of the plant, loading and unloading of tankers may make additional staff necessary.

4.1.6.5. Maintenance Costs

Annual repair and maintenance costs generally amount to 3 – 6 % of the invested capital [90]. About half of this amount is material costs and the rest wages. Their long-term increase can therefore be assumed to be between that of the investment costs (see Table 7) and that of the wage rates (see Table 10). Experience shows that high repair costs are incurred shortly after a plant is put into operation, in eliminating start-up difficulties, and after a prolonged operation because of the need to carry out major repairs. Depending on how new a process is, annual costs of 5 – 10 % of the estimated total investment should be allowed for in an initial estimate.

Table 10. Gross hourly earnings of an average worker in the chemical industry in Germany , p. 286[91]

Year	Earnings, €/h
1980	7.22
1981	7.68
1982	8.10
1983	8.47
1984	8.72
1985	8.99
1986	9.30
1987	9.69
1988	10.06
1989	10.42
1990	11.11

4.1.6.6. Overheads

The overheads of the company consist of the total costs of the central facilities (research, personnel, energy-supply department), works management, staff facilities, internal plant traffic, etc. These costs are very dependent on the structure of the company. As a rule they are ca. 5 – 20 % of the combined energy, staff, and depreciation costs.

4.1.6.7. Capital-Dependent Costs (Depreciation)

The plant capital can be depreciated by various methods [92]. For a cost efficiency calculation it is generally sufficient to assume linear depreciation over a ten-year utilization period. The annual depreciation is then 10 % of the plant capital:

$$\text{Depreciation} = \frac{I}{10 \cdot N} \tag{4.7}$$

where

I = total investment in currency units
N = nominal capacity in kg/a

4.1.7. Technology Evaluation

Technology evaluation should provide information on the technical risk associated with a process. This risk is equivalent to the economic damage which would occur in the extreme case if the process did not work at all or could only be rendered operational by carrying out subsequent improvements.

Technology evaluation should also reveal whether the individual unit operations or the equipment and machines used are of technically established types and whether particular risks are incurred by breaking new ground (size, construction, material, scale-up factor, etc.). It is also necessary to include a statement as to whether the process is based on a technology with which the company is familiar (e.g., high pressure, gas-phase oxidation, phosgenation technology, etc.) or whether fundamental innovations are involved. If limits are exceeded (e.g., if the dimensions of the largest extraction

column previously operated by the company are considerably exceeded in the new project), the limits which have been tested should be specified.

This information should be elaborated in detail as early as possible in order to recognize, consider, and evaluate potential risks and causes of failure. During operation of the miniplant, close attention should be paid to deposits, frequently occurring malfunctions, etc., since such problems often subsequently lead to insuperable difficulties in the industrial-scale plant (e.g., foaming, emulsion formation, aerosol formation, etc.).

The technology evaluation gives a list of weak points arranged in order of technical risk (e.g., catalyst service life, corrosion, etc.). This list can then be used to take steps which reduce such risks. In principle, two steps are possible:

1) Expenditure on R & D can be increased at the weak point (e.g., the unit concerned may be checked on a pilot scale or a well-established solution may be sought). This option should always be chosen if, in an extreme case, failure of the unit concerned would be accompanied by a total investment loss.
2) Failure scenarios can be developed, i.e., what can be done later if problems occur or if a unit fails completely?

Normally, these two options have to be weighed against each other, i.e., it is necessary to clarify whether it is more economic to minimize the risk by increasing R & D expenditure (e.g., by building a pilot plant), or to reduce or eliminate it at a later date by technical measures (e.g., additional instrumentation or backup units).

Some general aspects, which may be useful in assessing the technical reliability, are summarized below.

From the estimate of the risk and the measures it suggests, the economic effect can be quantified very exactly (surcharges on the development costs or on the investment and repair costs). This transforms technical reliability into cost effectiveness considerations and results in the possibility of expressing the risk of a development project in monetary units.

Improving Technical Reliability

Decoupling Process Steps. The susceptibility to failure of a process carried out in a single-train plant increases as the number of process steps and the probability of failure of individual steps increase: a single-train plant can only function if each step is operational.

If partial streams are fed back to earlier stages, the coupling between them is particularly strong. This is also true of the energy cross-links within a process (see Section 4.1.6.2). The increase in risk as the number of process steps increases is independent of whether the process is continuous or batch.

In practice, various options are now adopted in resolving this dilemma (see Figs. 22 and 23):

1) Decoupling by inserting tanks for intermediates (the greater the buffer tank, the greater the independence). For larger quantities of product or gaseous intermediates, this approach is virtually impracticable.
2) Decoupling by installing parallel arrangements of equipment which is particularly susceptible. For example, filters and fittings that are susceptible to failure should be provided with a by-pass if safety considerations permit (e.g., heat exchangers in the main flow in which severe fouling is to be expected).

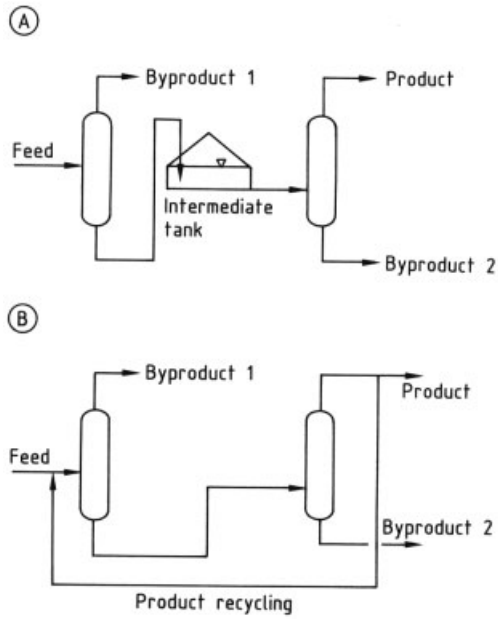

Figure 22. Example of decoupling a single-train plant: a) Intermediate tank to smooth out variations in feed flow; b) Recycling to smooth out variations in feed flow

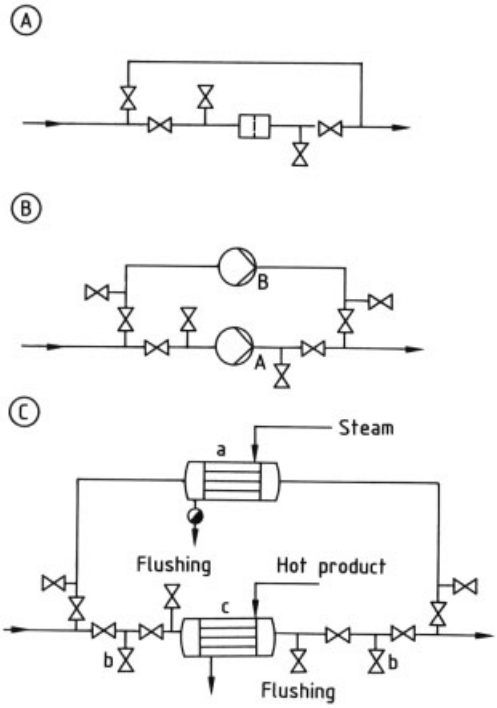

Figure 23. Decoupling of a single-train plant by including redundant equipment
A) Filter station with single by-pass; B) Pumping station with A and B pump; C) Preheater with flushing facilitya) Auxiliary preheater; b) Double block and bleed; c) Preheater

Each measure which is aimed at increasing the uptime of a single-train plant results in an increase in investment costs. In practice it is often the case that such steps are only taken after the plant is put into operation for the first time.

Choice of Process Steps which can be Safely Scaled-Up. An example of the dependence of technical reliability on the operational method employed is mass transfer. While simple standard equipment can be used for rectification, a wide variety of equipment in which phases are mixed in several stages and then separated has been developed for extractions. Such equipment generally requires more expenditure for safe operation than distillation columns.

The complexity of an operation increases as the number of substances and phases it involves increases and as the operating conditions become more extreme. Thus, the presence of two liquid phases requires special precautions to be taken. The same applies if particularly high or low temperatures or pressures are involved.

If solids are processed, the process must be assessed with particular care. The scaling-up of equipment for separating solids should be subjected to particularly careful scrutiny. Even the bunkering and transportation of solids present difficulties which are often specific to the materials.

If new technology is employed, the operational safety can be expected to be lower than that of well-established processes. It can be increased if similar types of equipment are used in the experimental plant as in the industrial-scale plant. This allows scaling-up to be possible but in the case of reaction equipment, it is not always the case.

Materials to be Processed. If toxic, inflammable, caustic, or otherwise hazardous substances are to be handled, a higher degree of operational safety is required from the outset than in the case of harmless substances.

Corrosion is an ever-present risk and necessitates the use of special materials. In making an initial selection, use may be made of published tables (see → Corrosion), but in most cases it is necessary to carry out corrosion tests. Trace impurities can lead to considerable difficulties if they accumulate at particular points in the equipment. The most reliable information in this connection is provided by operating an integrated miniplant.

If the flow rate of mixed phases is high, erosion of the wall material can be expected. The extent of erosion can often be lowered by the choice of material and by effective design. Increases in wall thickness are generally of no help because the erosion usually takes the form of pits. Miniplant experiments are of no assistance in this connection since they are usually not representative in hydrodynamic terms.

In the case of moving solids, not only the walls of the equipment but also the solids themselves are subject to abrasion. If a process is intended to give a particular size, the abrasion loss may have a considerable effect and, moreover, an increase in this abrasion is always expected on scaling up a process.

Solids often deposit inside equipment and result in fouling. Valuable information about this is provided by miniplant results. In heat exchangers, allowance is made for fouling by increasing the heat transfer resistance by empirically determined amounts (fouling factors) (see Fig. 24).

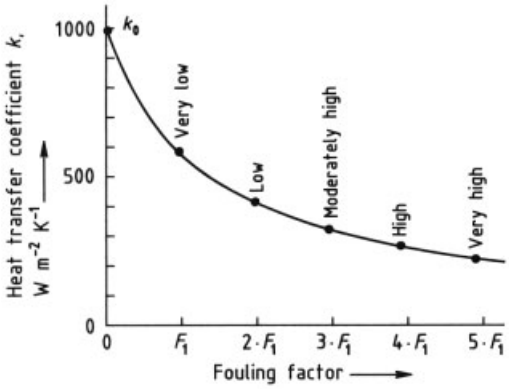

Figure 24. Effect of the fouling factor F_i on the heat transfer coefficient k
$F_1 = 7 \times 10^{-4}$ m^2 K W^{-1} (typical value for river water);
$k_0 = 1000$ W m^{-2} K;
$\dot{Q} = k\,A\,\Delta T$;
$1/k = 1/k_0 + F_i$;
$F_i = S_i/\lambda_s$
where S_i = thickness of fouling layer; λ_s = thermal conductivity of fouling layer; $\dot{Q}$ = heat flux; A = surface area of heat exchanger

If the expected operating time is inadequate, a back-up heat exchanger must be installed in parallel. In larger plants, heat exchangers are usually subdivided into several units for design reasons. Thus, if suitable fittings (double block and bleed) are provided to ensure that each heat exchanger can be cleaned while the plant is operating, the effect on the availability of the entire installation is small (see Fig. 23). However, this is only true provided the cleaning frequency is technically feasible. The design of heat exchangers and the cleaning operations should be carefully matched to the type of fouling, and extensive automation is always worthwhile in the case of large pieces of equipment.

If fouling is expected in columns, larger columns should be designed for whole-body access to each plate, while in small columns a manual-access hole should be provided for each plate to allow cleaning. Furthermore, designs should be used whose operation is relatively insensitive to fouling and which can readily be cleaned.

The choice of reaction equipment can be considerably influenced by fouling. As fouling increases, the pressure drop in fixed-bed reactors increases, whereas fluidized-bed reactors are virtually unaffected. Moreover, fixed-bed reactors must be opened and emptied in order to regenerate the contaminated catalyst, but fluidized-bed catalysts can usually be regenerated in the reactor. Often it is also possible to remove, regenerate, and replace the catalyst in a sidestream while the plant is operating. Fouling that arises both from the material being processed and from the material of which the equipment is made is more troublesome. For example, in many polymerization reactors, wall deposits can be suppressed by suitable choice of the wall material and processing method. In such cases, steps should be taken to ensure that the same wall material is used in the trial plant and in the industrial-scale plant.

If wall erosion or deposit growth occurs at a constant rate, the difficulties can be expected to decrease in bulk-storage equipment and pipelines with increasing scale-up but to remain unchanged in columns and heat exchangers (equipment with large surface areas).

Machines. Since machines have moving parts, they are always a potential source of malfunction and the reliability of a process decreases as the proportion of machines it employs increases. In general, machines having reciprocating parts are regarded as less reliable than those employing only rotating parts. For this reason it is particularly important to provide a reasonable level of back-up. Because of their relatively low cost, duplicate full-capacity pumps are usually installed, but occasionally two pumps having different functions can be provided with a common backup. In the case of reciprocating machines, in addition to a complete backup (2×100 % capacity), 2×66 % capacity, or in exceptional cases 3×50 % capacity, may also be considered.

Turbo machines are usually installed at 1×100 % capacity, but provision is made for rapid replaceability of parts which are subject to wear. Alternatively, two machines with 50 % capacity are often installed.

Instrumentation. The narrower the limits in which quantity flows and operating conditions have to be kept for reliable operation, the higher will be the requirements imposed on the instrumentation.

To increase plant availability, provision is made for n measurements instead of one measurement in the case of critical circuits and control loops. The complete unit will operate as specified provided m of the n measurements

($m<n$) are in the "good" range (m of n circuits). The greater m is, the higher will be the safety margin and the lower the plant availability. In the case of chemical plants, "two out of three" circuits have been found to be satisfactory, thus providing a compromise between safety margin and availability.

Multiple-Train Design. According to Equation (2.17), the increase in investment costs on scaling up the plant is approximately proportional to the capacity to the power 2/3. Consequently, the specific investment costs for a single-train plant are approximately 20 % lower than for a twin-train plant of the same total capacity. In addition, the personnel costs are lower for a single-train plant than for several individual plants of the same total capacity. However, certain plant components can often not be scaled up to the required total capacity with an adequate safety margin. Thus, the maximum capacity of polymerization vessels may be limited because a breakdown of the cooling system must be insufficient to bring about dangerous operating conditions, or it may be impossible to increase the size of an oxidation reactor for cost reasons. In such cases, the total capacity is provided by using the requisite number of individual pieces of equipment.

A significant advantage of multiple-train design is that it is easier to cope with a temporary reduction in load, which would normally mean shutting down and then starting up again. If the plant is underloaded, the specific consumption figures of multiple-train plants are lower. Furthermore, the capacity of such plants can more easily be increased at a later date, particularly if this has been planned from the outset. The advantages of multiple-train design can, of course, only be exploited if it is possible to link any line in a process stage with any of the lines of the subsequent stage. For this purpose manifolds will be needed.

4.1.8. Assessment of the Experimental Work

This part of the project study should contain information on whether the existing experimental results have been obtained in-house or are derived from third parties. If in-house knowledge is available, the test units in which the experiments were carried out (e.g., laboratory, integrated miniplant) should be specified and the duration and conditions of the experiments given.

The aim is to arrive at an as objective assessment as possible of the quality of the data obtained since this is of major importance for the accuracy of the project cost estimate.

4.2. Return on Investment

After the individual cost factors have been estimated, a rough cost effectiveness calculation is possible. This should provide information on the return on investment (= ratio of profit to capital employed) for the planned project. Investment in the process under consideration will be profitable if the sum of the revenues exceeds the total outlay and the profit (revenue – outlay) makes it possible to amortize and pay reasonable interest on the capital invested.

The return on investment provided by a process can be increased by minimizing the production costs. Exposing the main cost factors (raw material, energy, waste disposal, personnel costs, depreciation) will therefore indicate the direction the development should take in order to improve the process (see Table 11).

Table 11. Examples of measures which may result in the reduction of the main cost factors

Main cost factor	Measures
Raw materials	a) Increase selectivity of catalyst b) Minimize processing losses c) Use raw materials of lower quality
Energy	a) Optimize the thermal integration b) Use multistage evaporation c) Extraction instead of distillation d) Air condenser instead of water condenser
Waste disposal	a) Integrated waste utilization b) Choice of location
Depreciation	a) Simplify the reaction control and work-up b) Reexamine choice of materials c) Reexamine dimensioning of pipeline system d) Exploit synergies of the location

Minimizing the production costs does not, however, maximize the return on investment if the full capacity of the plant is not reached. In such cases, a plant with lower capacity and higher production costs might provide a larger return on investment despite being less economical.

Some of the indicators frequently used in practice to assess return on investment are described below.

4.2.1. Static Return on Investment

In its simplest form, the return on investment calculation widely adopted in practice relates the expected annual profit to the capital invested:

$$R = \frac{(E - PC)}{I} \cdot 100 \tag{4.8}$$

where

R	is the percentage return in %/a
E	is the annual sales proceeds (= amount sold multiplied by production or transfer price) in monetary units per year
PC	is the annual production costs in monetary units per year
I	is the capital expenditure in monetary units
$E - PC$	is the annual profit in monetary units per year

The method can be made more informative by taking the sales into consideration separately:

$$R = \frac{(E - PC)}{E} \cdot \frac{E}{I} \cdot 100 \tag{4.9}$$

The first factor indicates the sales success and the second the capital turnover.

Instead of R, use may also be made of the reciprocal value, i.e., the payback time 100/R. This indicates the time required to recover the invested capital so that the plant begins to make a profit.

The return on capital from the process being studied should be compared with that of known processes. Above a threshold value, which varies from case to case, the process will provide a return on investment. A minimum return R of 20 % is often employed, but because of the considerable vulnerability to technical risk and market certainty, this is only a guideline. While the value may be lower if the risk is low, it is prudent to impose a minimum value up to a factor of two higher in the case of new high-risk products.

Generally the process development engineer is not faced with the question of how profitable a project is but of whether a particular process engineering approach is more profitable than an other, alternative solution within the framework of the process being studied or not [61]. In this case, the sales are, of course, unchanged but the production and investment costs are:

$$PC_2 = PC_1 - \Delta PC$$
$$I_2 = I_1 - \Delta I \tag{4.10}$$

where

PC_1, I_1	are the production and investment costs before the change,
PC_2, I_2	are the production and investment costs after the change.

The return offered by a particular approach is then given by the ratio of the changes:

$$R_{12} = \frac{\Delta PC}{\Delta I} \tag{4.11}$$

The change in the overall return resulting from adopting this particular approach is then:

$$R_2 = \frac{R_1 \cdot I_1 + \Delta PC}{I_1 + \Delta I} \tag{4.12}$$

The weakness of static return-on-investment indicators is that they take a short-term view and, since they do not allow for future changes, they favor short-lived investments. Their advantage is that they can easily be determined.

4.2.2. Dynamic Return on Investment

The dynamic return on investment method is superior to the static method because allowance is made for the timing of the cash movements. Since only contemporaneous monies are comparable, the investment outlay and the earnings from the profit must be discounted to a fixed point in time, usually the start of production. If a certain rate of interest p is assumed, all the capital movements can be referred to that point to obtain the so-called net value of a capital movement:

$$(E - A)_{t=0} = (E - A)_t \cdot (1 + p)^{-t} \tag{4.13}$$

where

E	= expected earnings, e.g., from product sales in monetary units per year
A	= expected outlay, e.g., investment costs, production costs, in monetary units per year
p	= interest rate

t = number of years (negative before start-up and positive after start-up)

For example, when started up, a plant has to bear double the cost of an outlay incurred five years before start-up if the rate of interest is 15 %. Conversely, a revenue five years after start-up will only carry half the weight and one ten years after start-up will be virtually insignificant. The capital value K of a project is therefore the sum of the net values:

$$K = \sum_{t=-m}^{t=+n} (E - A)_t \cdot (1+p)^{-t} \tag{4.14}$$

where

m = duration of process development in years

n = useful life of project in years (usually 10 years)

To compare different projects, a useful life of ten years is usually assumed for chemical plants.

An investment is only beneficial if its capital value is equal to zero or is positive.

The internal rate of return method (discounted cash flow, DCF) tries to find the rate of interest which will result in a capital value of zero, i.e., the rate at which the net values of the incoming and outgoing payments are equally large (the internal rate of return). The internal rate of return p_0 is determined by setting the above capital value function equal to zero:

$$\sum_{t=-m}^{t=+n} (E - A)_t \cdot (1+p_0)^{-t} = 0 \tag{4.15}$$

Projects are compared by comparing their calculated internal rates of return, the project with the highest internal rate of return being the most profitable.

4.3. Economic Risk

The return on capital calculation indicates whether or not a project is economically worthwhile. To calculate the return, assumptions must be made about the most probable product prices, the production outlay, the plant loading (actual production/nominal capacity), investment costs, etc. Time will see whether these assumptions are correct. The impact of these uncertainties on the cost effectiveness of the project can be quantified by using the following variables.

4.3.1. Sensitivity Analysis

Sensitivity analysis yields information about the sensitivity of the process to variations in individual influencing variables and therefore about the reliability of the result of the return on capital analysis. It also provides pointers to the direction in which the process should be developed further. The development is particularly worthwhile if it alters influencing variables to which the return on capital is highly sensitive. The procedure is to vary the different influencing variables in Equation (4.15) one after the other in the unfavorable direction (while keeping all the others constant) by a certain fraction (for example, 10 %) and then determine how much the internal rate of return p_0 differs from the original value (see Table 12).

Table 12. Relative change in the internal rate of return on varying the influencing variable x ∗

Parameter x	$\Delta x/x$
Development costs	+10 %
Investment costs	+10 %
Product price	−10 %
Capacity utilization	−10 %
Net raw material costs	+10 %
Operating resources costs	+10 %
Staffing costs	+10 %

∗ By successively changing the parameters x in the unfavorable direction, the change in the internal rate of return can be calculated from Equation (4.15).

4.3.2. Amortization Time

The risk as a function of time can be quantified by means of the payback time t_w. This indicates the time which will be necessary for the cost of the capital employed to be recovered. As an initial estimate, t_w is the ratio of the investment costs I to the annual surplus income $(E - A)$:

$$t_w \approx \frac{I}{(E-A)} \tag{4.16}$$

5. Trends in Process Development

The pattern of process development should be such that it produces competitive advantages, i.e., a process should be developed rapidly and cheaply. Some trends which promote this objective are highlighted below [93]:

1) Nowadays only a team made up of members from all the scientific disciplines is capable of coping with the complex task of developing a process. For this reason, a project team, usually consisting of a chemical process engineer, a chemist specializing in preparation, an engineer, and possibly the future works manager should be brought together at the beginning [94].
   If necessary, this group may recruit or consult other specialists. Team work and team spirit are essential to the success of the activity, the synergies resulting from this type of activity being exploited to the full.
   When process development is complete, it is advantageous for at least one member of the project team to continue working on the project during planning, construction, commissioning, and start-up, since otherwise a great deal of detailed knowledge would be lost.
2) In the case of short-lived consumer-related products (dyes, active substances, dispersions, etc.), which are preferably produced in batch processes, the trend is towards equipment which is as flexible as possible (multipurpose production plants, multipurpose small-scale plants). Innovations in such products are usually due to changes in formulation, and much less a result of modifications to the equipment. The ideal is to produce a wide variety of products in a batch plant over a long period of time.
3) In the case of basic products and intermediates produced in continuous plants, process development can make a contribution to competitiveness by reducing the time and the cost required to develop improved manufacturing processes. This objective can be achieved by increasing the scale-up factors (see Section 3.1). In order to limit the resulting increase in technical risk, use is now made of the synergy between an integrated miniplant and a mathematical simulation. This makes it possible to dispense with the pilot plant stage, i.e., the synergism between an integrated miniplant technique and a mathematical simulation results in scale-up safety margins of similar quality to those obtained by using a pilot plant. As a result it is possible to shorten the time required for process development by ca. 30 %, which can have a noticeable effect on the cost effectiveness of a new process.
4) Another trend is the development and use of expert systems. These are already available for simple, limited systems such as the optimization of individual unit operations. For much more complex systems such as catalyst development or the optimization of entire plants, expert systems are at present still a long way from general practical application.

Future tasks in process development will be in the following fields [95]:

1) Improvement of plants and processes with a view to an integrated environmental protection [96–98]
   a) *Residue reduction:* Reduction in the amounts of byproducts by improving the catalysts and the reaction control
   b) *Integrated residue utilization:* Integrated work-up of unavoidable waste streams within the main plant, i.e., exploitation of on-site synergies (thermal and material coupling) and effective reuse of worked-up wastes
2) Development of new processes for the basic products and intermediates on the basis of:
   a) natural gas (C_1 chemistry)
   b) renewable raw materials
   c) waste streams as raw material
3) Development of processes that result in smaller, more economical, and cheaper plants.

6. References

1. *Jahrbuch der Chemiewirtschaft,* VCH Verlagsgesellschaft, Weinheim 1991.
2. H.-B. Amecke: *Chemiewirtschaft im Überblick,* VCH Verlagsgesellschaft, Weinheim 1987.
3. H.-J. Quadbeck-Seeger: "Chemie für die Zukunft," *Angew. Chem.* **102** (1990) 1213 – 1224.
4. W. Jentzsch: "Was erwartet die Chemische Industrie von der Physikalischen und Technischen Chemie?" *Angew. Chem.* **102** (1990) 1267 – 1273.

5. C. Marchetti: "Die magische Entwicklungskurve," *Bild Wiss.* **10** (1982) 115 – 128.
6. E. G. Krubasik: "Angreifer im Vorteil," *Wirtschaftswoche* (1984) no. 23, 48 – 56.
7. M. Perlitz, H. Löbler: "Brauchen Unternehmen zum Innovieren Krisen?", *Zeitschrift für Betriebswirtschaft* **55** (1985) 424 – 450.
8. G. Specht, K. Michel: "Integrierte Technologie- und Marktplanung im Innovationsportfolio," *ZfB* **58** (1988) 502 – 520.
9. M. E. Porter: "Technology and Competitive Advantage," *Journal of Business Strategy* (1985) 60 – 78.
10. W. Busfield, D. Merigold, *J. Chem. Soc. (A)* 1969 2975 – 2977.
11. H. Knapp: "Methoden zur Beschaffung von Information über Stoffeigenschaften," *Chem.-Ing.-Tech.* **59** (1987) 367 – 376.
12. E. Fitzer, W. Fritz: "Technische Chemie," Springer-Verlag, Berlin 1975.
13. B. Streit: *Lexikon Ökotoxikologie,* VCH, Weinheim 1991.
14. L. Roth, G. Rupp: *Sicherheitsdatenblätter,* VCH, Weinheim 1991.
15. BASF, Acrylsäure rein, technical information, TI/ED 1330 d, Ludwigshafen 1987.
16. F. Dolder: "Geheimhalten oder patentieren?" *Management Zeitschrift* **60** (1991) 64 – 68.
17. E. D. Dichtl, O. Issing: *Vahlens Großes Wirtschaftslexikon,* Verlag C. H. Beck, Verlag Franz Vahlen, München 1987.
18. Chemical Week, International ICIS/Chemical Week Price Report, Europe spot prices, $/m.t.
19. *Ullmann,* 4th ed., **4,** 24.
20. K. W. Schneider: "Großanlagen in der Mineralölindustrie und Petrochemie," *Chem.-Ing.-Tech.* **37** (1965) 875 – 877.
21. G. Lück: "Verfahrensentwicklung," *Fortschr. Verfahrenstech.* **21** (1983) 473 – 486.
22. H. Harnich, R. Steiner, K. Winnacker: *Chemische Technologie,* Carl Hauser Verlag, München 1984.
23. W. Hofen, M. Körfer, K. Zetzmann: "Scale-up-Probleme bei der experimentellen Verfahrensentwicklung," *Chem.-Ing.-Tech.* **62** (1990) 805 – 812.
24. J. Krekel, G. Siekmann: "Die Rolle des Experiments in der Verfahrensentwicklung," *Chem.-Ing.-Tech.* **57** (1985) 511 – 519.
25. J. Gmehling, U. Onken, W. Aret: *Vapor – Liquid Equilibrium Data Collection,* DECHEMA, Frankfurt 1977.
26. J. M. Praussnitz: *Molecular Thermodynamic of Fluid-Phase-Equilibria,* Prentice Hall, Englewood Cliffs, N.Y. 1969.
27. I. Hacker, K. Hartmann: "Heuristische Regeln zum Entwurf verfahrenstechnischer Systeme," in *Probleme der modernen chemischen Technologie,* Akademie-Verlag, Berlin 1980, pp. 194 – 244.
28. E. Blaß: "Methodische Entwicklung verfahrenstechnischer Prozesse," *Chem.-Ing.-Tech.* **57** (1985) 201 –210.
29. G. Kaibel, E. Blaß, J. Köhler: "Gestaltung destillativer Trennungen unter Einbeziehung thermodynamischer Gesichtspunkte," *Chem.-Ing.-Tech.* **61** (1987) 16 – 25.
30. H. H. Erdmann, J. Kussi, K. H. Simmrock: "Möglichkeiten und Probleme der Prozeß-Synthese," *Chem.-Ing.-Tech.* **56** (1984) 32 – 41.
31. G. Stephanopoulos, A. W. Westerberg: "Studies in Process Synthesis-II," *Chem. Eng. Sci.* **31** (1976) 195 –204.
32. E. Blaß: *Entwicklung verfahrenstechnischer Prozesse – Methode, Zielsuche, Lösungssuche, Lösungsauswahl,* 1st ed., Verlag Salle und Sauerländer, Frankfurt 1989.
33. K. Hartmann, L. Dietzsch: "Modellierung und Simulation verfahrenstechnischer Systeme," in *Probleme der modernen Technologie,* Akademie-Verlag, Berlin 1980, pp. 43 – 77.
34. E. Blaß: "Aufgabengerechte Informationssammlung für die Verfahrensentwicklung," *Chem.-Ing.-Tech.* **56** (1984) 272 – 278.
35. M. Buder, W. Rehfeld, T. Seeger: Grundlagen der praktischen Information und Dokumentation, vol. 1, K. G. Saur, München 1990, pp. 366 ff.
36. "Dechema-Informationssysteme für die Chemische Technik und Biotechnologie," *Achema-Jahrbuch 1991,* **vol. 1,** Frankfurt 1990, p. A60.
37. R. Berenz: "Das Rad nicht zweimal erfinden," *Jahrbuch der Chemiewirtschaft,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 172 – 174.
38. M. Zeitz: "Simulationstechnik," *Chem.-Ing.-Tech.* **59** (1987) 464 – 469.
39. G. Wozny: "Software macht den Fortschritt," *Jahrbuch der Chemiewirtschaft,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 153 – 158.
40. D. Gilles, P. Holl, W. Marquardt: "Dynamische Simulation komplexer chemischer Prozesse," *Chem.-Ing.-Tech.* **58** (1986) 268 – 278.

41. W. Scheiding, H. Hartmann, P. Tönnishoff: "SATU 86 – Simulationsrahmen zur dynamischen Prozeßsimulation für den industriellen Einsatz," *Chem.-Ing.-Tech.* **61** (1989) 292 – 299.
42. W. Marquardt: "Rechnergestützte Erstellung verfahrenstechnischer Prozeßmodelle," *Chem.-Ing.-Tech.* **64** (1992) 25 – 40.
43. W. Klar: "Expertensysteme," *Elektronik* **8** (1991) 108 – 123.
44. J. J. Ferrada, J. M. Holmes: "Developing Expert Systems," *Chem. Eng. Prog.* **86** (1990) 34 – 41.
45. A. Lieberam: "Expertensysteme für die Verfahrenstechnik," *Chem.-Ing.-Tech.* **58** (1986) 9 – 14.
46. P. Trum, N. R. Judica: "Expertensysteme – Aufbau und Anwendungsgebiete," *Chem.-Ing.-Tech.* **58** (1986) 6 – 9.
47. A. H. Masso, D. F. Rudd: "The Synthesis of System Designs," *AICHE* **15** (1969) no. 1, 10 – 17.
48. G. L. Steele: *Common LISP,* Digital Press, 1984.
49. L. Puigjaner, A. Espuña: *Computer-Oriented Process Engineering,* Elsevier Science Publishers, Amsterdam 1991.
50. D. Greß, H. Hartmann, G. Kaibel, B. Seid: "Einsatz von mathematischer Simulation und Miniplant-Technik in der Verfahrensentwicklung," *Chem.-Ing.-Tech.* **51** (1979) 601 – 611.
51. S. Maier, G. Kaibel: "Verkleinerung verfahrenstechnischer Versuchsanlagen – was ist erreichbar?" *Chem.-Ing.-Tech.* **62** (1990) 169 – 174.
52. L. A. Robbins: "The Miniplant Concept," *Chem. Eng. Prog.* **75** (1979) no. 9, 45 – 48.
53. W. Brust, F. Heimann, G. Kaibel, S. Maier: "Klein, aber fein," *Jahrbuch der Chemiewirtschaft 1991,* VCH, Weinheim 1991, pp. 179 – 182.
54. U. Wörsdörfer, U. Müller-Nehler: "Ein Gewinn für die Forschung," *Jahrbuch der Chemiewirtschaft 91,* VCH, Weinheim 1991, pp. 182 – 186.
55. A. Schulz-Walz: "Meß- und Dosiertechnik beim Betrieb von Miniplants," *Chem.-Ing.-Tech.* **62** (1990) 453 – 457.
56. J. M. Berty: "The Changing Role of the Pilot Plant," *Chem. Eng. Prog.* **75** (1979) no. 9, 48 – 51.
57. J. G. Lowenstein: "The pilot plant," *Chem. Eng.* **9** (1985) no. 23, 69 – 76.
58. R. P. Palluzi: "Understand Your Pilot-Plant Options," *Chem. Eng. Prog.* **87** (1991) no. 1, 21 – 26.
59. P. Prinzing, R. Rödl, D. Aichert: "Investitionskosten-Schätzung für Chemieanlagen," *Chem.-Ing.-Tech.* **57** (1985) 8 – 14.
60. E. Muthmann: "Ermittlung der Investitionskosten von Chemieanlagen in verschiedenen Projektphasen," *Chem.-Ing.-Tech.* **56** (1984) 940 – 941.
61. W. Frey, F. Heimann, S. Maier: "Wirtschaftliche und technologische Bewertung von Verfahren," *Chem.-Ing.-Tech.* **62** (1990) 1 – 8.
62. J. Jung: "Aspekte der Vorausberechnung von Anlagenkosten bei der Projektierung verfahrenstechnischer Anlagen," *Chem. Tech. (Heidelberg)* **12** (1983) 9 – 18.
63. C. A. Miller: "New Cost Factors Give Quick, Accurate Estimates," *Chem. Eng.* **13** (1965) 226.
64. *Jahrbuch der Chemiewirtschaft,* VCH Verlagsgesellschaft, Weinheim 1991, p. 64.
65. K. E. Lunde: "Transfer Pricing," *Chem. Eng.* **4** (1985) 85 – 87.
66. H. Körner: "Optimaler Energieeinsatz in der Chemischen Industrie," *Chem.-Ing.-Tech.* **60** (1988) 511 –518.
67. B. Linnhoff, J. A. Turner: "Heat Recovery Networks: New Insights Yield Big Savings," *Chem. Eng.* **88** (1981) no. 22, 56 – 70.
68. B. Linnhoff: "New Concepts in Thermodynamics for Better Chemical Process Design," *Chem. Eng. Res. Des.* **61** (1983) 207 – 223.
69. D'Ans Lax: "Taschenbuch für Chemiker und Physiker," 3rd ed., vol. 1, Springer Verlag, Berlin 1967.
70. F. A. Henglein: *Grundriss der chemischen Technik,* Verlag Chemie GmbH, Weinheim 1963.
71. *Chemiker Kalender,* 3rd ed., Springer Verlag, Berlin 1984.
72. W. Hellmund, R. Auelmann, J. Wasel-Nielen, A. Rothert: *Chem. Ind. (Düsseldorf)* **114** (1991) no. 9, 12 – 13; **114** (1991) no. 10,10 – 12; **114** (1991) no. 11, 10 – 12.
73. J. M. Coulson, J. F. Richardson: *Chemical Engineering,* 4th ed., Pergamon Press, Oxford 1990.
74. *Jahrbuch der Chemiewirtschaft 1991,* VCH Verlagsgesellschaft, Weinheim 1991, p. 303.

75. W. Swodenk: "Umweltfreundlichere Produktionsverfahren in der chemischen Industrie," *Chem.-Ing.-Tech.* **56** (1984) 1 – 8.
76. G. Lipphardt: "Produktionsintegrierter Umweltschutz, Verpflichtung der Chemischen Industrie," *Chem.-Ing.-Tech.* **61** (1989) 855 – 860.
77. T. Bakay, K. Domnick: "Produktionsintegrierter Umweltschutz – Verminderung von Reststoffen, gezeigt anhand ausgeführter Beispiele," *Chem.-Ing.-Tech.* **61** (1989) 867 – 870.
78. H. Lenz, M. Molzahn, D. W. Schmitt: "Produktionsintegrierter Umweltschutz – Verwertung von Reststoffen," *Chem.-Ing.-Tech.* **61** (1989) 860 – 866.
79. DIN 38 409,Bestimmung des Biochemischen Sauerstoffbedarfs (BSB), DIN 38 409 part 51; DIN 38 409, Bestimmung des Chemischen Sauerstoffbedarfs (CSB); DIN 38 409 part 41.
80. "Modifizierter Zahn-Wellens-Test," *OECD Guidelines for Testing of Chemicals,* Paris 1981, p. 302 B.
81. I. Juhnke, D. Lüdemann: "Ergebnisse der Untersuchung von 200 chemischen Verbindungen auf akute Fischtoxizität mit dem Goldorfentest," *Z. Wasser Abwasser Forsch.* **11** (1978) 161 – 164.
82. G. Bringmann, R. Kühn: "Grenzwerte der Schadwirkung wassergefährdender Stoffe gegen Bakterien (Pseudomonas putida) und Grünalgen (Scenedesmus quadricauda) im Zellvermehrungshemmtest," *Z. Wasser Abwasser Forsch.* **10** (1977) 87 – 98.
83. DIN 38 412,Testverfahren mit Wasserorganismen, 1991, part 8.
84. DFG Deutsche Forschungsgemeinschaft: *Toxikologische Untersuchungen an Fischen,* Verlag Chemie, Weinheim 1983.
85. *Kühn Birett Merkblätter Gefährliche Arbeitsstoffe,* **25th suppl.,** Nov. 1984, "VI-4 Wassergefährdende Stoffe."
86. P. Reher: "Produktion mit umweltgerechter Entsorgung," *Jahrbuch der Chemiewirtschaft,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 213 – 217.
87. W. Hüning, H. Nentwig, C. Gockel: "Die verbrennungstechnische Entsorgung eines Chemiebetriebes," *Chem.-Ing.-Tech.* **56** (1984) 521 – 526.
88. W. Hüning, P. Reher: "Verbrennungstechnik für den Umweltschutz," *Chem.-Ing.-Tech.* **61** (1989) 26 –36.
89. G. Färber: "Standardtechnik löst Probleme," *Jahrbuch der Chemiewirtschaft 1991,* VCH Verlagsgesellschaft, Weinheim 1991, pp. 158 – 162.
90. D. Wagener: "Kostenermittlung am Anlagenbau für die chemische Industrie," *Tech. Mitt.* **65** (1972) 542 –547.
91. "Bruttostundenverdienst," *Jahrbuch der Chemiewirtschaft 1991,* VCH Verlagsgesellschaft, Weinheim 1991, p. 286.
92. H. Kölbel, J. Schulze: "Neuentwicklungen zur Berechnung von Preisindices für chem. Anlagen," *Chem. Ind. (Düsseldorf)* **19** (1967) 340 – 343, 701 – 706.
93. G. Bogenstätter, H. Gerrens: "Die chemische Fabrik der Zukunft," *Chem. Ind.* **37** (1985) 275 – 277.
94. J. Platz, H. J. Schmelzer: *Projektmanagement in der industriellen Forschung und Entwicklung,* Springer, Berlin 1986.
95. R. Shinnar: "The Future of the Chemical Industries," *CHEMTECH* 1991, 58 – 64.
96. M. Zlokarnik: "Umweltschutz – eine ständige Herausforderung," *Chem. Ing. Tech.* **61** (1989) 378 –385.
97. G. Scharfe, B. Sevekow: "Das Beispiel Bayer," *Chem. Ind.* (1991) 17 – 20.
98. J. Wiesner: *Produktionsintegrierter Umweltschutz in der chemischen Industrie,* DECHEMA, Frankfurt, 1990.

Chemical Plant Design and Construction

ERICH MOSBERGER, Lurgi AG, Frankfurt, Federal Republic of Germanyand coauthors

1. Introduction . . . 3
2. Feasibility Study . . . 4
2.1. Initial Work . . . 4
2.2. Cost Estimation . . . 5
2.2.1. Investment Costs . . . 5
2.2.1.1. Global Methods of Investment Cost Estimation . . . 6
2.2.1.2. Detailed Methods of Investment Cost Estimation . . . 7
2.2.1.3. Item-by-Item Calculation . . . 9
2.2.1.4. Cost Indexes . . . 9
2.2.2. Operating Costs . . . 9
2.2.3. EDP Support . . . 10
2.3. Profitability Analysis . . . 11
2.3.1. Profitability Analysis as an Engineering Task . . . 11
2.3.2. Methods of Profitability Analysis . 11
2.4. Site Selection . . . 12
2.5. Decision between Alternative Investments . . . 13
3. Preliminary Design . . . 13
3.1. Preliminary Design Costs . . . 14
3.2. Final Selection of Site Locations . 14
3.3. Process Design . . . 16
3.3.1. Optimization . . . 20
3.3.2. Safety Aspects and Environmental Control . . . 23
3.3.2.1. Protection Against Emissions . . . 24
3.3.2.2. Noise Control . . . 24
3.3.2.3. Occupational Safety and Health . . 26
3.3.2.4. Plant Availability . . . 27
3.3.2.5. Authority Engineering . . . 27
3.4. Basic Engineering . . . 29
3.4.1. Equipment Specification from the Process Engineering Standpoint . . 29
3.4.2. Materials of Construction . . . 30
3.4.3. Plant Layout . . . 33
3.4.4. Preliminary Piping and Instrumentation Diagram . . . 35
3.5. Calculation of Plant Costs . . . 35
3.5.1. Equipment . . . 36
3.5.2. Bulk Materials . . . 37
3.5.3. Other Costs . . . 38
3.6. Conclusion of Preliminary Design Phase . . . 40
4. Contract Writing and Forms of Contracts . . . 40
4.1. Licensing Agreements . . . 40
4.1.1. Patent Licenses . . . 40
4.1.2. Process Licenses . . . 40
4.1.3. Process Licenses via Engineering Contractors . . . 41
4.1.4. Know-How Contracts via Engineering Contractors . . . 42
4.2. Design and Supply Contracts with Engineering Contractors . . . 42
4.2.1. Selection of Engineering Contractors . . . 42
4.2.1.1. Importance of Risk in the Plant Business . . . 42
4.2.1.2. Selection and Award Criteria . . . 43
4.2.2. Form and Content of Contracts . . . 44
4.2.2.1. Basic Concerns in Contract Writing 44
4.2.2.2. Contract Types and Provisions . . . 45
4.2.2.3. Essential Elements of a Contract . . 46
5. Execution of the Project . . . 47
5.1. Scope of Work . . . 47
5.2. Project Organization and Management . . . 48
5.2.1. Matrix Project Management . . . 49
5.2.2. The Project Manager . . . 49
5.2.3. The Project Team . . . 50
5.2.4. The Start Phase of a Project . . . 50
5.3. Project Control (Schedules, Progress, Costs) . . . 51
5.3.1. Time Scheduling . . . 51
5.3.2. Progress Planning and Control . . . 54
5.3.3. Cost Planning and Control . . . 54
5.3.4. Project Report . . . 57
5.4. Detail Engineering . . . 57
5.4.1. Process Engineering . . . 57
5.4.2. Plant Layout . . . 57
5.4.3. Apparatus and Machinery . . . 59
5.4.4. Piping . . . 61
5.4.5. Control Systems . . . 67
5.4.6. Electrical Design . . . 69
5.5. Procurement . . . 70
5.5.1. Purchase of Equipment and Services 71
5.5.2. Expediting . . . 72
5.5.3. Shipping . . . 72
5.6. Planning and Execution of Civil Work and Erection . . . 72
5.6.1. Planning of Civil Work and Erection 72
5.6.1.1. Planning of Civil Work (Including Structural Steel Work) . . . 72
5.6.1.2. Erection Planning . . . 73

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_477

5.6.2. Execution of Construction 74
5.6.2.1. Construction-Site Organization and Management 74
5.6.2.2. Time Scheduling and Progress Control 75
5.6.2.3. Construction Work 76
5.7. Commissioning 77
5.7.1. Plant Design and Commissioning . 77
5.7.2. Operating Manual 77
5.7.3. Responsibility and Organization . . 79
5.7.4. Preparation for Commissioning . . . 80
5.7.5. Plant Startup 80
6. Computer Support 80
6.1. Role of Computers in Project Execution 81
6.2. EDP Infrastructure and Systems 82
6.3. Coordination and Interfaces . . . 82
7. Quality Assurance 83
8. Training of Plant Personnel 84
9. References 85

1. Introduction

Since the 1930s, the design and construction of chemical plants have become increasingly specialized. Chemists and engineers collaborate to develop process and engineering concepts, which engineers and designers then transform into detailed plans and specifications for all the components of a chemical plant. Purchasing agents procure equipment from specialist manufacturers. Construction and installation firms are put under contract to build the plant.

Plant design and construction starts as an idea of the potential owner. Increasingly complex markets and the interrelations of the world economic and political systems require critical examination of every project for feasibility, economic relevance, and environmental impact. As a rule, this is done with the aid of a feasibility study that includes preliminary design work. Market analyses are carried out to determine potential sales, future demand dynamics, availability of raw materials, and the competitive situation. The plant capacity and location are specified. Particular attention must be paid to environmental protection. Studies are supplemented by sufficiently accurate estimates of capital requirements and profitability.

Once the decision has been made to go ahead with the construction project, whereby the owner may have opted not to carry it out himself, the owner will prepare an accurate, comprehensive definition of the plant which is used as the basis for inviting bids from competent engineering contractors. This approach is increasingly employed, especially for large projects. (If the owner has an adequate pool of experienced design engineers, construction specialists, and procurement staff, he may plan and construct the plant himself.) The conceptual phase of the project ends when an appropriate engineering firm is chosen and the contract signed.

The implementation phase consists of the engineering of the chemical plant, procurement of plant equipment and material, construction, and commissioning. The engineering contractor either performs all of these tasks or brings in subcontractors or personnel employed by the owner to carry out a portion of the work.

Plant equipment is fabricated by specialized manufacturers. Only in rare cases does the engineering contractor or the plant owner have production facilities.

After a successful test run, the plant is handed over to the owner. (Fig. 1 shows a schematic diagram of the development of a chemical plant project.)

This article does not discuss the design and construction of small, simple plants that specialized firms can supply "off the rack." It deals with larger, more complicated projects in the field of chemical plant construction. The term chemical plant design and construction is used in a very broad sense. It also relates to allied technologies, such as metallurgy, environmental protection, fiber and food production, and petroleum and natural gas processing. Some of these basic project development and execution principles can also be applied to general industrial plant construction.

2. Feasibility Study

At the start of a project, the bases used for planning are still very inexact. No major costs should be incurred until it is known whether the project is feasible or not. Nevertheless, all alternatives must be considered. A great deal of experience is needed if uneconomic variants are to be discarded without generating high design

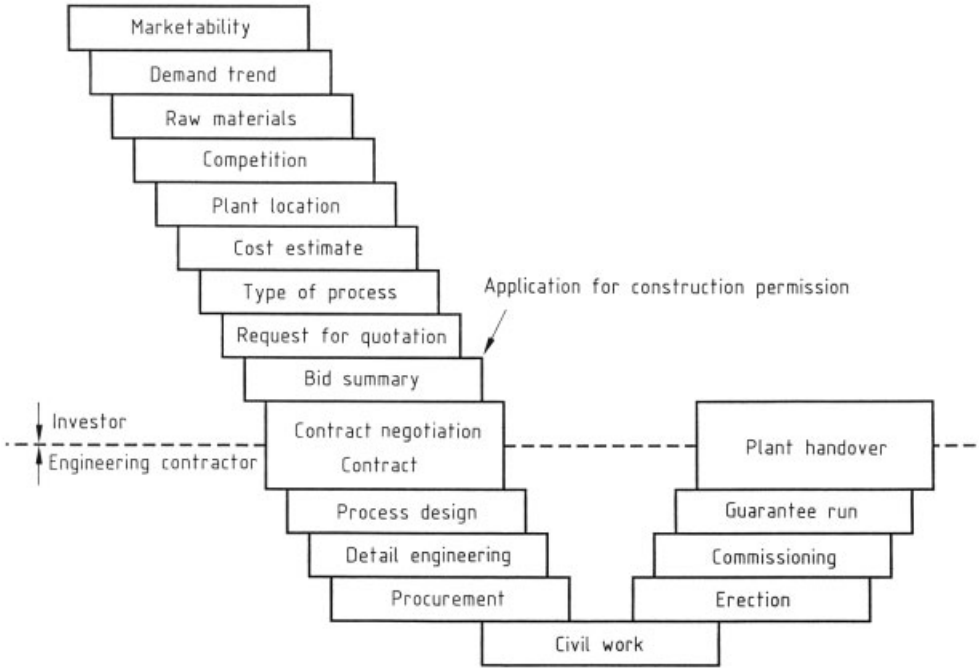

Figure 1. Development of a project

costs. The first steps toward accurate definition of the project are carried out by the company that wishes to erect the plant. A small working group is formed, consisting of process engineers, sales engineers, and other engineering personnel. It is useful to appoint a project leader to coordinate the team. Normally, the same person will later become responsible for implementing the project if the decision is taken to proceed.

Preliminary studies aimed at defining the objective partly depend on initial economic estimates and generally include:

1) Market analysis and trend analysis
2) Fixing of production capacity
3) Examination of competing processes and of the patent and license situation
4) Legal aspects
5) Estimation of investment costs
6) Estimation of production costs
7) Estimation of profitability
8) Selection of an appropriate plant location

2.1. Initial Work

Market Analysis and Production Capacity. Before the economically optimum plant capacity is set, a careful market analysis must be carried out. This analysis must be performed by experienced market analysts, even if the product has a comprehensive international sales history. These specialists evaluate literature on the development of similar products, determine the capacity of existing production facilities, carry out representative surveys, obtain suitable conditions from downstream processors in the case of intermediate products, and forecast the future market for the product. They must also provide a realistic evaluation of the competition and the world economic situation. Forecasts of costs for raw materials and working capital play an important role in the economic analysis.

A new chemical product usually experiences slow early growth with a relatively high price and low output. The market then expands; production climbs faster and prices drop. Finally,

prices stabilize at a low level, and older, smaller plants are shut down (see Fig. 2). If market analysis shows that world output of the product is still increasing rapidly, expansion of capacity through erection of a new plant may have a good outlook if other conditions are favorable. The future price decline must, however, be allowed for in the calculations.

In the upper part of the trend curve, investment only makes sense if a clear demand is perceptible in the market. The production curve for the old process may fall off if a new, more economical production process appears. (Example: low-pressure process for methanol production forces shutdown of older high-pressure process.) The new process may open up new markets and lend new impetus to development [1–4].

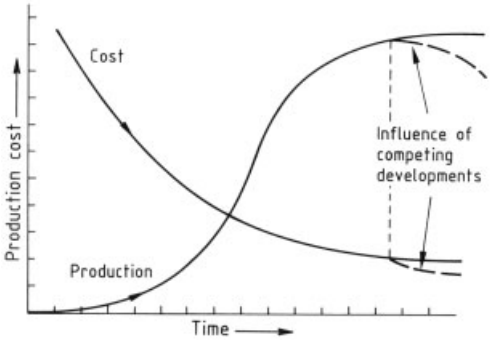

Figure 2. Development of a new product

Competing Processes, Patent and License Situation. The patent and license situation must be investigated at the beginning of the study. Foreign patents may block the construction project if the owner of the patent is not prepared to grant a license (see also Chap. 4).

Legal Aspects. Extensive, far-reaching environmental regulations mean that it is essential to make an early approach to authorities that will later have to approve the operation of the plant (see Section 3.3.2.5).

2.2. Cost Estimation

If initial planning work has shown that investment is desirable and the market analysis has led to a tentative capacity figure, a first estimate of investment costs and subsequent operating costs is performed.

Simply calculating and analyzing the investment costs is not sufficient because, over the service life of a plant, operating costs make up a much greater proportion of life-cycle costs than investment costs do. The goal of cost estimation in the conceptual phase is to optimize the life-cycle costs of a chemical plant. This often means increasing investment costs so as to lower operating costs. Operating costs are favorably influenced by long component and equipment lifetimes, improvements in maintenance and consumption, and other factors. Cost prediction as the basis for profitability analysis should therefore include the determination of investment costs as well as subsequent production and operating costs.

The following sections discuss methods of determining investment and operating costs.

2.2.1. Investment Costs

A proven technique in investment cost estimation is to subdivide the project into onsite and offsite items.

Onsite Items. The term onsites denotes all facilities required to make the desired product.

Offsite Items. Offsites are all facilities that are normally not located within the process plant. They include facilities for the delivery of steam, electric power, gas, solid or fluid fuels, water, compressed air, and instrument air. Furthermore, this group includes stockpiles and warehouses for raw materials and semifinished and finished products; service facilities (administrative buildings, canteens, workshops, stores, laboratories, parking, fire protection, roads, tracks, and harbor facilities); and, finally, power plants; loading docks; facilities for treating raw materials, off-gas, wastewater; and waste disposal facilities.

In the United States, offsites are usually divided into storage and handling (stocks of raw materials and finished products), utilities (generation or delivery of energy as steam, electricity, and water), and service facilities (e.g., offices, recreation rooms, laboratories, workshops, warehouses).

On grounds of cost, an attempt must be made to carry out an "order-of-magnitude" estimate

at minimal cost. This estimate is of course inexact, but later it makes it easier to decide whether to bear the costs for an accurate analysis. The approximate determination of investment costs is subdivided into simple "global" methods and "detailed" methods. Methods for estimating investment costs are now well-established; those now in use are described in publications mainly dating from 1960 – 1984.

2.2.1.1. Global Methods of Investment Cost Estimation

Global methods permit investment costs to be estimated relatively easily and with an accuracy between ± 30 % and ± 50 %. Several methods are outlined below [5].

Single Complexity Factor. In the single-complexity-factor method [6] processes are classified as having a low, medium, or high complexity factor.

Low Complexity Factor.

This class includes all batch processing plants and all processes involving simple syntheses (e.g., production of sulfuric acid).

Medium Complexity Factor.

This class comprises processes with gas and fluid phases that run at ordinary pressures and temperatures.

High Complexity Factor.

This class covers processes with high pressures and/or temperatures, as well as polymerization processes.

Investment costs depend on the complexity factor and the required plant capacity, and are determined from empirical data obtained in other processing plants. Auxiliary and utility units are taken into account by adding 45 %.

This method has the advantage that it permits estimation of investment costs in the orientation phase when little information about the process is available.

Turnover Ratios. The turnover-ratio method allows costs to be estimated without process information by using market information such as product sale price and sales volume [7].

On the basis of plants already on-stream, a capital turnover ratio is obtained by dividing the annual return on sales by the investment costs. A statistically determined turnover ratio and an expected annual return on sales are then used to estimate the investment costs of new plants. Turnover ratios in the chemical industry lie between 1.2 and 1.5.

Statistical turnover ratios can be found for the analysis of individual plants, companies (based on balance sheets and profit-and-loss statements of typical firms), or a whole industrial sector.

Degression Exponents. The use of degression exponents (cost-versus-capacity exponents) per-mits relatively accurate cost estimation. The technique is based on costs for plants already on-stream. The exponents are used to estimate investment costs for the planned facility as a function of plant capacity.

Investment costs for proces plants are published from time to time and can be used for initial cost estimation. The costs of identical or similar plants within the same company can be used in a similar way. In 1967, J. E. Hasselbarth [8] published the costs of process plants for 60 chemical products, including investment costs per tonne of annual capacity. His figures referred to costs within battery limits (i.e., within the plant boundary), exclusive of land and offsite facilities. In 1970, K. M. Guthrie [9] compiled the investment and operating costs for 54 chemical and refinery processes covering a wider capacity range. The degression exponents cited in both publications allow the calculation to be applied to other capacities. (Example: Given a degression coefficient of 0.7, doubling the capacity leads to an increase in investment costs by a factor of $2^{0.7} = 1.65$.)

When specific degression exponents are used it should be noted that the error range grows with the capacity scaling factor. This type of calculation generally gives acceptable results only for scaling factors of up to 1:3. Furthermore, the use of the method depends on the state of the art because changes in processes, apparatus, and mechanical technique can change the exponents.

Comparative Methods. When adequate data are available from an existing plant similar to that being planned, the investment costs of the old plant can obviously be used to calculate the costs of the new one.

The following information is needed [5]:

1) Production capacity
2) Construction time
3) Investment costs (inside and outside battery limits)
4) Location

Costs inside and outside battery limits are both determined with degression exponents. Cost estimation outside battery limits must, however, be preceded by a critical analysis of the auxiliary and utility units needed. The figures are adapted to the location by the use of indexes to adjust for the following [5]:

1) Construction location
2) Economic situation of the industry
3) Taxes
4) Labor market
5) Qualifications of available labor

2.2.1.2. Detailed Methods of Investment Cost Estimation

If the preliminary planning as embodied in technical documents has reached an advanced stage, it can form the basis for investment cost estimation that takes into account specific details of the project. Methods used for this are mainly based on analogies with plants that are already on-stream. Detailed methods of plant matching [10] and multiplication-factor techniques of cost determination are employed.

Lang Factor Method. If the process is well characterized, the required capacities of furnaces, apparatus, and machinery can be specified in preliminary flow sheets [11]. These specifications cover main plant items such as furnaces, columns, filters, reactors, heat exchangers, vessels, and machinery. The estimator can determine the costs "free-on-site" for such items or obtain the costs from suppliers.

Lang used cost analyses of existing plants to derive multiplication factors that allow determination of the investment costs for process units within battery limits if the costs of the main equipment items are known [12]. The factors depend on the type of plant. Lang distinguishes three types according to the state of aggregation of the raw material and product: "solid" (e.g., ore sintering), "solid – fluid" (e.g., oil – shale retorting with shale – tar recovery), and "fluid" (e.g., petroleum refineries, petrochemical plants).

If the costs for the main plant items are taken as 100, total processing-plant costs are found by multiplying by 3.10 (solid-processing plant), 3.63 (solid – fluid-processing plant), or 4.74 (fluid-processing plant).

Chilton [13] and Hand [14] have improved these approximate Lang factors by introducing supplements to the costs of main equipment items. Estimates of the total costs of a "grass-roots" plant can then be made. Figure 3 outlines the procedure for preliminary calculations by the Lang – Chilton method.

Other authors have extended and refined the Lang method. For example, Burgert in 1979 published an analysis of investment- cost structures for more than 100 projects [15].

In 1965, Miller [16] devised another system based on modified Lang factors. Miller assumed that the factors are influenced by three other parameters besides those used by Lang (solid, solid – fluid, fluid):

1) Size of main equipment items
2) Material from which the plant is constructed
3) Pressure for which the plant is built

Increasing size, more refined materials, and higher operating pressure increase the relative costs of the main equipment items in relation to storage, utilities, and service facilities, thus diminishing the factors. According to Miller, all factors can be referred to the mean per-piece costs of plant parts and depend on these.

Guthrie's Modular Technique. The modular technique published by Guthrie [17] in 1968 is also based on Lang's method and is the best of the multiplication-factor approaches. The project is first broken down into six modules [10]:

Five direct modules
- Chemical processes
- Solids handling
- Site development
- Industrial structures (civil work)
- Auxiliary and service facilities outside battery limits (offsites)

One indirect module
- Indirect project costs

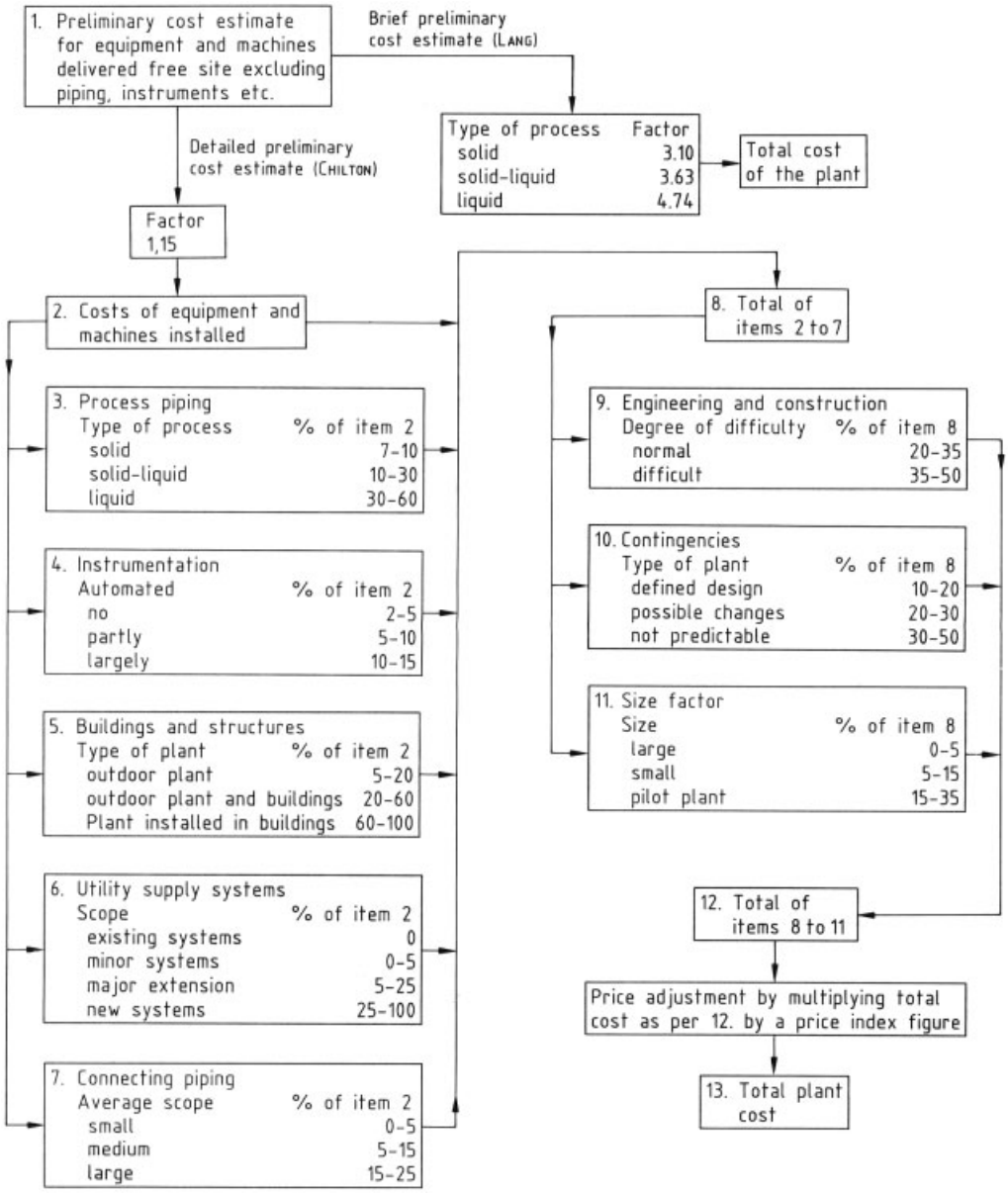

Figure 3. Preliminary cost estimate using factors (Lang – Chilton method)

The key costs of the direct modules are determined first. For the "chemical process" module, these might be costs for machinery and equipment. As shown in Figure 4, the key costs of the direct modules are multiplied by the gross module factors. The sum of the individual "gross module costs" gives the investment costs for on- and off-battery facilities.

GUTHRIE proposes a variety of methods for calculating the equipment costs of a module. The multiplication factors required for this include not only size of the component (magnitude fac-

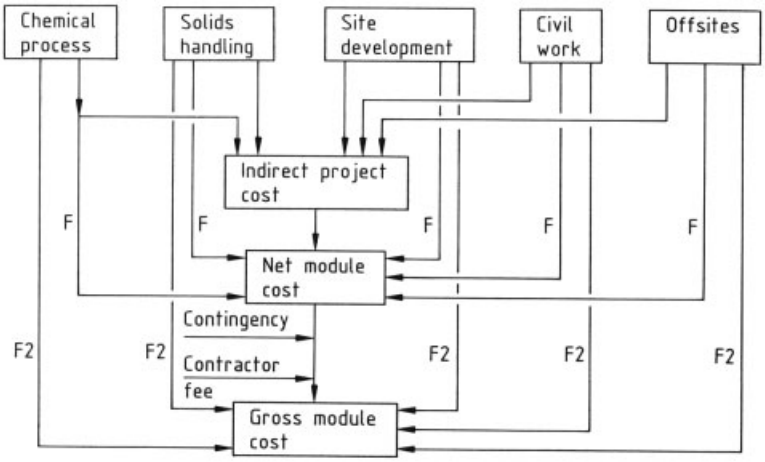

Figure 4. Simplified modular concept for estimating investment costs
F = net module factor; F2 = gross module factor

tor) and the alloy factor, but also indirect effects. This relatively accurate modular technique has not found wide acceptance, however, because it is relatively difficult to perform the calculation and maintain the statistical data base.

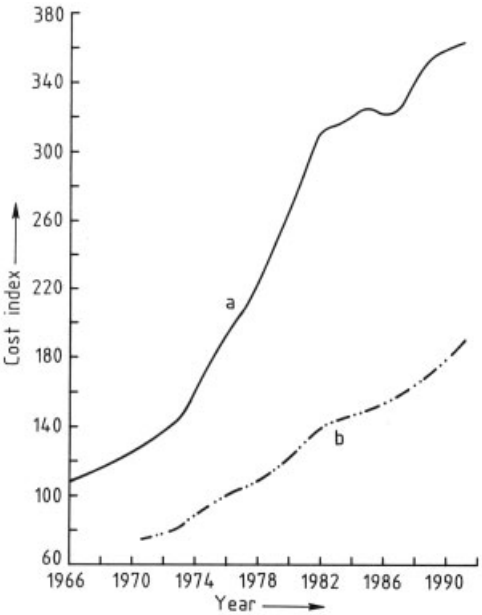

Figure 5. Development of cost indexes
a) CE Plant cost index, 1959 = 100, successively published in [18] (1982 revision of productivity factor from 2.50 to 1.75); b) Kölbel – Schulze index, 1976 = 100, successively published in [19]

2.2.1.3. Item-by-Item Calculation

When detailed methods of investment- cost estimation do not give sufficient accuracy, the only alternative is to calculate investment costs item by item. The plant equipment and the engineering work must be specified. The procedure and amount of work required for such a cost estimation are the same as those described for calculating plant costs (see Section 3.5). Close cooperation between the subsequent operator of the plant and an engineering firm has proved advantageous for this approach.

2.2.1.4. Cost Indexes

The methods of investment-cost estimation discussed above are generally based on historical statistics derived from existing plants. The cost figures obtained by these methods are therefore referred to given periods, such as 1985. If investment costs are to be estimated for the year 1990, the estimates must be adjusted to current prices. Every industrial country publishes one or more indexes for this purpose. Some widely used indexes for the United States and Germany follow:

1) Bureau of Labor Statistics cost index for equipment, machinery, and materials in the U.S. market
2) Chemical Engineering Plant cost index [18–20]

3) Kölbel – Schulze index for chemical plants (Kölbel – Schulze Index für Chemieanlagen) [21]
4) Producer price index for commercial products (Index der Erzeugerpreise gewerblicher Produkte) compiled by the German Federal Statistical Service (Statistische Bundesamt)

These indexes are based both on chemical-plant cost structures and on national primary price indexes. Figure 5 compares important cost indexes over time.

2.2.2. Operating Costs

Along with the investment costs, the operating costs incurred in the production of a given product also play an important part in deciding whether to erect a plant. The technical and economic literature, however, contains little information on the preliminary calculation of operating costs. Possible reasons are the complexity of the problem and the company's possible loss of maneuvering room if internal operating data were published [22].

The methods of estimating operating costs discussed below are based on data from comparable plants or empirical data from plants belonging to the same company [22].

They are related, but differ as regards starting information: published data, empirical data, business information, physical data, correlations, and information from comparable plants. In order to check the reliability of the results, operating costs should be estimated by several methods so that the calculations can be verified and error ranges given.

Graphical Method. The graphical method is based on statistical evaluation of operating costs in existing plants. Operating costs per unit of product are plotted versus plant capacity. It is important to be aware of the scope of the plotted costs. The graph usually includes only the manufacturing costs of a product: raw materials, power, catalysts, chemicals, wages, depreciation, and maintenance. Plant overheads, fixed costs, and indirect production costs should also be taken care of by multiplication factors.

Business Analysis. The analysis of balance sheets and profit-and-loss statements from companies that manufacture the product in question as their main commodity may also be helpful.

Energy-Based Methods. Chemical production processes involve large amounts of energy. The chemical reactions themselves often contribute very little to energy requirements but upstream and downstream operations do. This fact provides the basis for several methods used to estimate operating costs [13], [23–25].

Key Cost Categories. If it is assumed that chemical plants show relatively constant operating- cost structures for a given product, operating costs can be calculated with multiplication factors if a single cost category is known accurately [22].

Scaleup Methods. When operating costs for similar plants are known,specific data can be used to derive scaling coefficients for proportional, personnel-dependent, and investment-dependent costs. The operating costs can then be estimated.

2.2.3. EDP Support

A number of manufacturers and operators have established electronic data processing (EDP) programs for estimating investment and operating costs. Examples are Factest (ICI) [26] and the ASPEN package [27]. The ASPEN PLUS software, a flow sheet simulation program, is supplemented with a costing module.

The program sizes the most important equipment and machinery from the process simulation. Investment costs are estimated by the use of multiplication factors and cost indexes to adjust to current price levels. ASPEN PLUS allows the calculation not only of fixed costs for an investment, but also operating costs.

To determine operating costs, the program calculates fixed and variable components separately. Variable costs include raw materials, fuels, catalysts, disposal, and "running royalties". Fixed costs comprise personnel costs for maintenance and operation, overheads, insurance, and taxes. The program generates summaries and details of annual operating costs. Finally, the software can evaluate a variety of profitability measures (Section 2.3).

2.3. Profitability Analysis

2.3.1. Profitability Analysis as an Engineering Task

In profitability calculations, it is necessary to keep in mind that a plant erected without reserves (standby units) for unavoidable shutdowns and repairs will produce for only 330 days (or 8000 h) a year; that is, it will attain only ca. 90 % of rated capacity on a long-term basis. Manufacturers generally rate their equipment for operation at 10 % over capacity. However, this figure only applies to intermittent overloads and is not guaranteed.

If full-load operation is required all year, the plant must not be designed as a single-train (single processing line) facility, unless it is set up for 110 % capacity and adequate storage is provided for the finished product. Large storage areas are needed if sales are seasonal (e.g., fertilizers).

Because the feasibility study includes comparisons between alternative processes, two processes with equal profitability need not be ranked equally. For example, both may have equal production costs but different fixed costs.

Fixed costs usually comprise interest payments and operator wages. Nearly all other costs depend on output and are therefore variable (e. g., costs of raw materials, power, and fuel). If market conditions make it necessary to reduce the output to, say, 80 % of rated capacity, the plant with high fixed costs will become unprofitable more quickly. Thus, in the case of two equally profitable plants, the plant with the lower fixed-cost contribution will be preferred. High fixed costs often have to be accepted, if the need for reliable operation dictates that critical parts of the plant must be designed with 100 % standby capacity or the plant must be subdivided into parallel trains.

The profitability calculations must take into account that 2 – 2.5 years usually elapse between the start of planning and the commissioning of the plant. Interest therefore has to be paid on design and construction costs. Working capital is also needed for storage of raw materials and product. Finally, for technical or market-related reasons, a period of 6 – 12 months generally elapses after commissioning before the utilization of the plant is sufficient to cover the costs (break-even point). Only then does the return flow of capital begin.

The history of an investment is illustrated schematically in Figure 6.

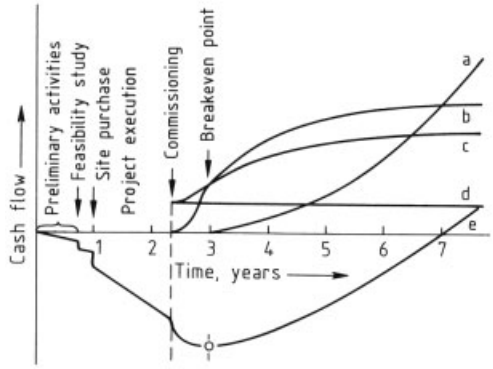

Figure 6. Schematic showing the course of capital investment
a) Accumulated profits; b) Return on investment; c) Total production costs (fixed and variable); d) Fixed costs; e) Accumulated cash flow

2.3.2. Methods of Profitability Analysis

Methods for assessing the profitability of a project [1, pp. 285 – 401], [28–33] differ in the way investment, revenue, and risk are associated. The three most important techniques are described briefly below.

These profitability calculations are combined and linked with operations research techniques [34], [35], so that alternative proposals can be obtained with a justifiable amount of effort.

Payout (Payback) Period. If alternatives are only to be compared, it may be adequate to divide invested capital by gross excess revenues (proceeds from sales minus operating costs). This method gives a quick indication as to whether the investment is attractive. If, for example, a value of 3 (payout period 3 years) or less is obtained, the investment should be profitable. This does not mean, however, that the plant will be completely written off after three years on stream.

A more realistic figure can be obtained by including taxes and interest under expenses and listing revenues by year after commissioning (higher operating costs in the first year when full capacity has not been reached, future decline in earnings).

Return on Investment (ROI). In large chemical companies, investments depend not so much on the payout period but on whether the investment will increase total profits, i.e., dividends to the stockholders. Thus a large company will only invest if a certain return on invested capital is ensured. The ROI method is employed in such cases. The annual profit before taxes and interest, but after depreciation, is divided by the invested capital. In contrast to the preceding method, startup losses are neglected and data for design- capacity operation are used in the calculation. The result is a pretax rate of return on invested capital.

Dynamic calculations are based on the discounted cash flow (DCF) method. The DCF method is an advance over ROI. It allows for the fact that investment costs precede revenues. Since only funds existing at the same time can be compared, all revenues and expenses that are directly or indirectly related to the project from the start of planning onwards are discounted to a fixed time, usually the start of production. A rate of return is then sought that makes the sum of the discounted annual excess revenues equal to the cash value of the total investment at the start of production. Since excess revenues are spread throughout the entire year, they are all recalculated to the middle of the year and discounted from then on. The calculation runs over the economic life of the project, but usually only 10 years, since the longer-term market position can scarcely be foreseen. Furthermore, the equivalent value of revenues after more than 10 years is so small that it has little effect on the result (internal rate of return).

2.4. Site Selection

The selection of an optimal site is a high-priority item in the feasibility study. Every economic region has its advantages and disadvantages. The present and future importance of correct site selection can scarcely be overstated. Wrong site choices cannot be corrected later and have led to the downfall of many companies.

The development of world trade and the decline of tariffs have changed the environment of many existing plants for the worse. Formerly, for example, the steel mill was located near the coke plant and iron ore transported to it. Today, the most important Japanese steel mills are located on the coast and import coal and ore from overseas. Petroleum refineries used to be sited at petroleum sources. Now that giant tankers and pipelines have come into wide use, refineries are often located in consumption centers.

Many large chemical companies are now investing in coastal and foreign property. Good surveys of publications important in site selection can be found in [1, p. 439], [36–38]. Stobough has devised a selection system, based on a point ranking, which allows unsuitable sites to be eliminated quickly. The short list of remaining sites should then be examined as described below.

Site Quality, Topography, Soil Conditions, Climate, Flood Risk. The short list should include only locations that appear suitable for the plant and possible subsequent expansions from the standpoint of size, price, transportation facilities, and buildability.

If the terrain is not flat, it should be established whether grading or filling is necessary and whether piles must be driven for foundations. If transportation facilities (roads, railroad tracks, water routes) are lacking, the expense of developing them must be ascertained. Only costs for developed sites can be compared.

Raw-Material Availability Including Power, Fuels, and Water. Raw materials of good quality must be available at a favorable price throughout the service life of the plant. Other prerequisites are availability of sufficient fresh water, electric power, and fuels. Obtaining power and steam on a grass-roots site can be expensive.

Environmental Conditions, Environmental Legislation, and Infrastructure of the Construction Site. Improved roads, rail connections, and location on harbors or year-round navigable rivers are important criteria.

Long-Term Availability of Qualified Labor. In developing countries, leading personnel – from foremen to management – must meet very stringent requirements. Only after years of schooling will local personnel have the required standard of education.

Raw-material costs, wages, and maintenance costs are also crucial in site selection. Finally, it must be taken into account that the phase of startup losses will be longer if an industrial site has to be developed from scratch.

2.5. Decision between Alternative Investments

Preparation for Decisionmaking. Large investment projects involve exploration of widely varying options. In addition to straight profitability analysis, qualitative factors should also be considered (e.g., special site problems, political environment, market development). These qualitative factors involve risks, which must also be assessed. The preparation phase for decisionmaking therefore includes not only calculation of investment and operating costs, but also risk and sensitivity analysis [39].

Risk Analysis. In risk analysis, all constraints that can have an adverse effect on life-cycle costs and earnings must be identified, documented, and assessed. These include:

1) Estimation of sales market
2) Energy cost development
3) Availability of raw materials
4) Plant construction risks
5) Management-related risks (e.g., site-dependent problems, reliability of vendors and subcontractors)
6) Organizational risks

The financial effects of the identified risks are quantified by experts in risk-assessment procedures and the results then evaluated in a Monte Carlo simulation [39].

Sensitivity Analysis. Sensitivity analysis should be carried out to find out how the profitability and risk situation changes when certain assumptions and constraints are varied. For example, it might be asked how much the investment costs would have to be reduced to allow the desired profit and an acceptable amortization.

A calculation of annual revenues based on the quantity of product that can be sold in the market and its price allows a variational calculation of this type to be made. The expected profit is deducted from the sum over the project life. The remainder represents the maximum available investment and operating costs that have to be optimized in several steps. This approach is called "design to cost." Investment costs are often minimized in this approach by designing low-cost plants (usually open-air plants with simple equipment and no storage capacity for intermediate products).

Several iterations are usually required to reach an optimum between investment costs and operating costs.

Decisionmaking. Profitability, sensitivity, and risk analyses may lead to the conclusion that execution of the project is not desirable. Recent publications may have already shown that similar projects undertaken by third parties will oversaturate the market. Other companies may have access to such good raw-material sources that the company considering the project cannot compete. It may be that environmental regulations have a prohibitive effect on costs. The planning work should then be stopped until new information or analyses suggest a different conclusion. The avoidance of huge losses on a badly planned facility far outweighs the costs incurred up to this point. If, however, the feasibility study reveals a positive situation, the next step is to incorporate the capital requirement into the company's investment program and give the go-ahead to start preliminary design.

3. Preliminary Design

The following results from the feasibility study provide the basis for deciding to proceed to the preliminary design phase:

1) Plant capacity has been set on the basis of market research
2) The choice between expansion of an existing facility and construction of a new one has been made
3) The list of potential sites has been shortened to two or three alternatives
4) The projected capital outlay has been determined
5) The projected production costs for the product have been determined
6) The payout time and profitability have been estimated

The main task in preliminary design is to obtain a more exact calculation that takes into consideration all costs up until commissioning. The

first step toward this objective is to work out the engineering details.

A qualified project leader directs the preliminary design. Specialist engineering teams (e.g., for process calculations, equipment design, plant layout, and estimations) provide advisory support [40].

3.1. Preliminary Design Costs

The funds and time spent on preliminary design can be considerable, depending on how precise planning and budget are to be. Although highly accurate results are always sought, technical documentation and calculations should be refined only to the degree necessary for subsequently deciding whether to implement the project. Events during the preliminary design period may force the abandonment of the project.

If the feasibility study predicts very favorable profitability, less accurate calculations (say ± 20 %) may be acceptable; if the project is expected to be marginally profitable, better accuracy (e.g., ± 10 %) is needed. Figure 7 [41] shows the basic information required for given accuracy levels. Even if the upper and lower limits of the percentage ranges in Figure 7 are made equal, the probability of exceeding the projected costs is greater than that of falling short of them. This is primarily due to subsequent additions that are not known at the time when the calculation is performed. A "contingencies" item is therefore commonly included in the calculation.

Design costs for preliminary calculations at various accuracy levels, according to the American Association of Cost Engineers [42], are presented in Table 1. Specific empirical figures are given in [43].

Preliminary design requires a special project team consisting of persons with the necessary expertise. A chemical company whose engineering staff is oriented mainly toward maintenance and the occasional addition of new pieces of equipment should not attempt to perform the preliminary design of a large plant or a branch plant in-house. Such a department lacks appropriate experience and is also short of capacity. The need to hold down design costs and maintain the performance level of in-house engineering staff on their specific tasks, forces even large chemical firms to collaborate closely with external engineering contractors who will later be in charge of executing the project. External engineering firms usually offer cost advantages because they have so much experience in their routine fields that they can quickly estimate reliable cost figures for projects based on well-known processes.

The client's role is limited – at least as far as established processes are concerned – to making process knowhow available, purchasing licenses, and cooperating with the engineering contractor in customizing the plant to relevant requirements.

Contract forms have been developed for the collaboration between the customer and the engineering firm in the basic-design and/or detailengineering phases, as well as for procurement, supply, construction, construction supervision, and commissioning (Section 4.2.2).

Another option for cutting design costs is for the chemical firm to obtain a process license, with the licenser providing the complete "basic design."

3.2. Final Selection of Site Locations

After the feasibility study two or three suitable sites are often available for the new plant. A final decision can be made only after detailed study.

If the potential sites are close to the client's parent plant, similar fringe conditions can be assumed to apply for the purpose of site comparison. If, however, a branch plant is to be erected in a foreign country, conditions are usually different which means that each of the alternatives has to be carefully analyzed.

Important constraints besides those already listed in Section 2.4 include:

1) Medium- and long-term capacity of the local (national) market to absorb the product
2) Export to nearby countries
3) Availability, quality, and price of raw materials
4) Political situation and risks (e.g., risk of nationalization)

	Required information	Detailed estimate ±5%	Definitive estimate ±10%	Preliminary estimate ±20%	Study estimate ±30%	Order of magnitude estimate >±30%
Location	Location factors	●	●	●	●	○
	General description	●	●	●	○	○
	Soil conditions	●	●	●	○	○
	Space required and dimensions for tracks, embankments, roads, fences	●	●	●	○	○
	General plan for development, topographical maps	●	●	○	○	○
	Site and partial site plan	●	○	○	○	○
Process flow diagrams	Block flow diagrams	○	○	○	●	○
	Preliminary technological schemes	○	○	●	○	○
	Process flow diagrams	●	●	○	○	○
Apparatus and machinery	Preliminary equipment list (size and material)	○	○	●	●	○
	Detailed design and equipment specifications	●	●	○	○	○
	Equipment data sheets	●	●	○	○	○
	General arrangement drawings					
	-preliminary	○	●	●	○	○
	-detailed	●	○	○	○	○
Buildings	Estimated size and type of buildings	○	○	●	●	○
	Foundation outlines	○	●	●	○	○
	Architectural and construction layout drawings	●	●	●	○	○
	Preliminary building drawings	○	○	●	○	○
	General arrangement drawings incl. elevations	●	●	○	○	○
	Detailed drawings	●	○	○	○	○
Energy requirements	Estimated energy requirements (water, steam, electric power, etc)	○	○	○	●	○
	Preliminary heat balance	○	○	●	○	○
	Preliminary energy flow diagrams	○	○	●	○	○
	Final heat balance	●	●	○	○	○
	Final energy flow diagrams	●	●	○	○	○
	Detailed drawings and specifications	●	○	○	○	○
Pipes and fittings	Preliminary piping and instrumentation diagrams	○	○	●	●	○
	Final piping and instrumentation diagrams	●	●	○	○	○
	Detailed drawings and specifications	●	○	○	○	○
Insulation	Preliminary specifications	○	○	●	○	○
	Preliminary lists of piping and equipment to be insulated	○	●	○	○	○
	Detailed specifications for insulation (wallthickness and materials)	●	●	○	○	○
	Detailed design and specifications	●	○	○	○	○
Control systems	Preliminary instrumentation list	○	○	●	○	○
	Final piping and instrumentation diagrams	●	●	○	○	○
	Detailed drawings and specifications	●	○	○	○	○
Electrical engineering	Preliminary list of motors including design data	○	○	●	●	○
	Final list of motors including design data	●	●	○	○	○
	Size and number of transformers	●	●	●	○	○
	Specification of power distribution	●	○	○	○	○
	Preliminary lighting plan	○	○	●	○	○
	Preliminary plans for locking and control instruments	○	●	○	○	○
	One-line diagrams for power and light	●	●	○	○	○
	Detailed drawings and specifications	●	○	○	○	○
Man hours	Engineers and draftsmen	●	●	●	●	○
	Specialists (craftsmen, e.g., for piping models)	●	○	○	○	○
	Site supervision	●	○	○	○	○
Scope of project (standard process)	Characteristic data of the plant including capacity, location, space requirements, requirements of service and auxiliary plants, storage	○	○	○	○	●

Figure 7. Accuracy of cost assessment based on available information [41]

Table 1. Design costs for cost estimates [43]

Type of estimate	Synonymous terms	Accuracy, %	Costs as percentage of project value
Order of magnitude estimate	ratio estimate seat of the pants estimate ballpark estimate guesstimate	± 30 – ± 50	0 – 0.1
Study estimate	evaluation estimate predesign estimate factored estimate	± 20 – ± 30	0.1 – 0.2
Preliminary estimate	sanction estimate funding estimate authorization estimate budget estimate	± 10 – ± 25	0.4 – 0.8
Definitive estimate	project control estimate	± 5 – ± 15	1 – 3
Detailed estimate	tender estimate contractor's final cost estimate	± 2 – ± 5	5 – 10

5) Tax laws, tariffs, possibility of repatriating profits

If, for example, one country has cheap raw materials but a limited capacity to absorb the product, partial upgrading of the raw material and fabrication of intermediate products should be considered. Manufacture of the end product would then take place where there is adequate long-term demand and a suitable distribution network.

A team of the client's experts should perform a thorough on-site examination of each alternative before the final decision is made; close contact with national and local government agencies is important. The recommendations of this team should weigh heavily in the choice of site.

Partnership with qualified domestic enterprises is becoming increasingly popular (e.g., joint ventures).

3.3. Process Design

See also → Process Development.

The feasibility study defines the process objective, i.e., it specifies products (type and quantity), feedstocks and auxiliaries (type and quantity), and local conditions (environmental situation, elevation, climate, energy situation). This objective, together with the overall state of the art and the experience of the operater, licenser, or plant design and construction contractor, provide the basis for process selection.

Process selection can be done most simply in the form of a block flow diagram (Fig. 8), in which each block represents a unit operation or, in complex plants, a plant section containing several unit operations. The blocks are joined by lines representing the principal material and energy streams.

The first step in process design is to establish the operating parameters for the major stages in the process:

1) In the case of *chemical reactions*, the pressures, temperatures, concentrations, reactor type, and reactor size are defined or estimated on the basis of the reaction kinetics and experience gathered in existing plants
2) In the case of *mixing* (stirring, gas dispersion, suspension) or *separation* (distillation, drying, precipitation, filtration) of substances, the pressures, temperatures, concentrations, and type and size of apparatus are defined or estimated on the basis of established rules.

If information needed for setting the operating parameters or designing reactors/apparatus is not known, it must be obtained in bench or pilot-plant tests or calculated approximately on the basis of similar reactions or unit operations.

The unit operations used in process engineering are described in [44–49].

Process Flow Diagram. The next step is to prepare a process flow diagram from the block flow diagram. Standard symbols (e.g., defined in DIN 28 004) are used to represent reactors and other apparatus, including equipment for conveyance and control of important streams.

In complex plants, it may be necessary first to prepare a synoptic flow diagram (Fig. 9) and

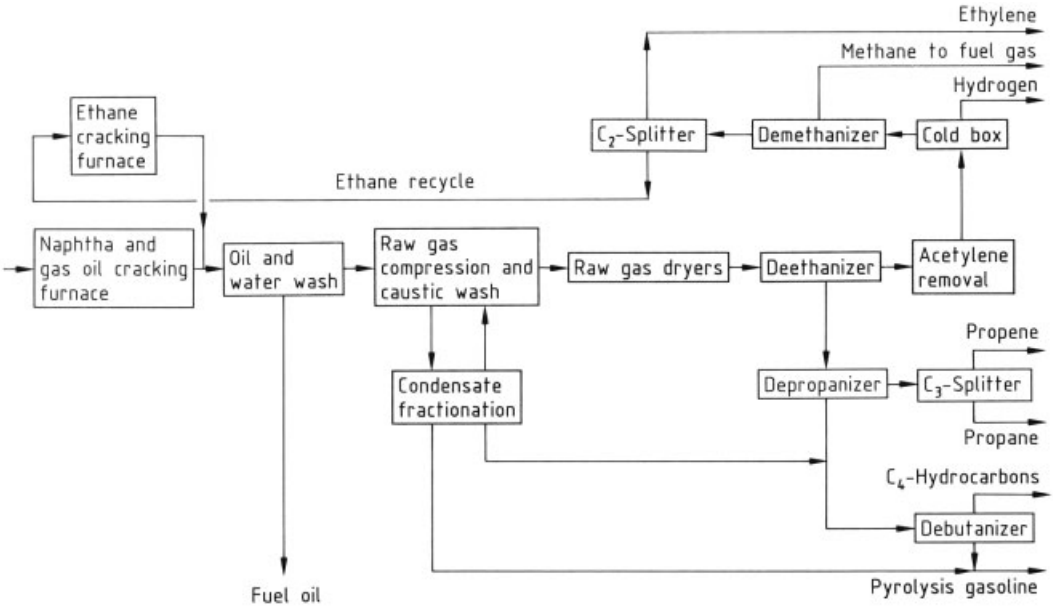

Figure 8. Block flow diagram (olefin plant)

then to draw up individual flow diagrams, giving the needed detail for plant sections and auxiliary operations.

Determination of Final Process Data. When the desired effective operating time per year has been set (e.g., 8000 h, corresponding to 91 % availability), the design capacity of the plant (i.e., the mass throughput per unit time) is defined. The next step is to compile the specifications for all feedstocks, auxiliaries, catalysts, utilities, and end products, and to estimate what intermediate products are to be expected. These specifications include relevant physical and chemical properties as well as the battery-limit states of all substances received and shipped.

On the basis of the process flow diagram, the preliminary process parameters, and the above-mentioned specifications are used to prepare material and energy balances for the process steps and finally for the entire process. This objective is not generally achieved in a single set of computations. The experienced process engineer must use an iterative procedure to modify the process flow diagram and/or the process parameters so that a closed material and energy balance is attained as simply as possible. The following must always be ensured:

1) Compliance with emission limits.
2) Plant safety (i.e., adequate margins of safety relative to critical operating conditions).
3) Maintenance of product quality.
4) Control of startup operations and of planned and unplanned shutdowns. This includes specification of components and media required (e.g., heatup burners, cooling and purge gases, and pressure-reducing valves).

These calculations must also take account of long-term effects, such as increasing contamination (resulting in a higher pressure drop and less efficient heat transfer) or aging of catalysts (lower conversion, changes in temperature and concentration profiles).

The calculations give mass and energy values for all important points of the plant, preferably downstream of each unit operation. The mass and energy flow rates and the final process parameters are tabulated at the foot of each process flow diagram and keyed to points in the plant (Fig. 10).

The preliminary sizes of the reactors and apparatus are now checked and, if necessary, modified in the light of the final process parameters.

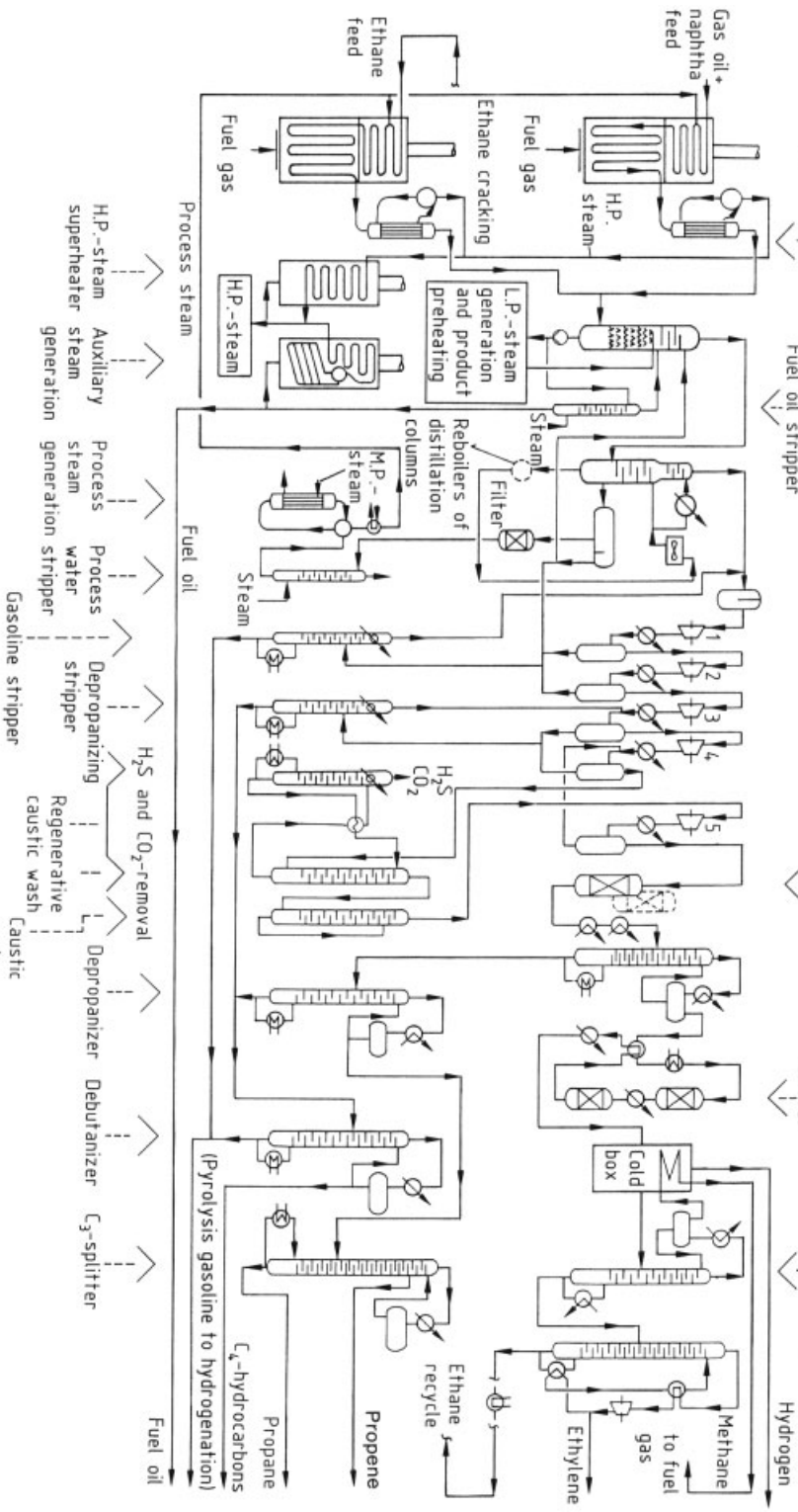

Figure 9. Overall process flow diagram (olefin plant)

If a plant is to be operated at reduced capacity for short or long intervals, appropriate calculations must be performed. If unacceptable operating conditions or bottlenecks are found in certain plant sections, additional measures must be taken (e.g., supplementary heating, gas recycle, or the partial shutdown of some unit operations). A similar treatment applies when occasional overloading of the plant is expected.

Other factors are then determined, namely the quantities of substances that have to be available for the initial charges and the storage capacities (including those outside the battery limits) needed for feedstocks, auxiliaries, intermediates, and end products.

Finally, a detailed process description is written. Assessments of known processes are available on a subscription basis from Chem Systems Inc. and SRI International. A survey of production processes for initial and intermediate organic products is given in [50].

Process Simulation. Process design calculations for multistage, interconnected processes with material and energy recycle soon become very complex. Computer tools enable designers to develop key concepts in a reasonable time and thus to optimize the process according to certain criteria (investment cost, yield, energy economy, production costs). "Flow sheeting" programs are complicated computer programs that model unit operations mathematically and allow them to be interconnected. A plant can thus be represented as a network of unit operations with material and energy streams and thus simulated [51], [52]. The material and energy balances can be calculated as functions of the process parameters. An iterative procedure can be carried out to bring the balances to equilibrium for individual and interconnected unit operations.

Simulation programs are available for steady-state operation and dynamic conditions (e.g., ASPEN PLUS, DESIGN II, PROCESS, HYSIM, and SPEED UP).

Design calculations for reactors and other equipment (e.g., heat exchangers) can be performed with special-purpose design programs and process simulation packages. The program for a fired tubular furnace (e.g., steam reformer) allows calculation of process conditions as a function of the configuration and geometry of the tubes, burners, and combustion chamber. An important function of such programs is to simulate the equipment when connected into systems and to identify and remedy bottlenecks (debottlenecking).

Process simulators have integrated substance data bases that meet the needs of most applications. Further data can be obtained from data bases [53] and compilations of physical and chemical data [54–56]. Proprietary data and empirical factors (interactions, long-term effects) can also be input.

3.3.1. Optimization

The object of plant optimization is to obtain an optimal economic result. This is a continual problem during the operation of a plant, especially if its capacity is diminished by aging or increased by expansion, or if product earnings, costs, or expected profits change during its lifetime. The parameters necessary for subsequent optimization must, if possible, be established in the design stage. They relate to process design, the plant concept, the selection of equipment, and the process control system.

The operating result is primarily determined by product earnings, expected profit, and costs. Obviously, costs decrease with longer plant service life and higher availability. The feasibility study should provide data about both of these factors.

Most chemical production facilities are operated around the clock. Since fixed costs represent a significant proportion of the operating costs (interest payments, personnel, energy supply, overheads), high availability is essential for an optimal result and is often more important than maximal process optimization.

Single-train plants with many unit operations in series are more susceptible to breakdowns than plants in which the critical sections are multitrained or have standby equipment. The drawback of a larger capital investment must be weighed against the advantage of a higher expected availability.

In established methods of risk analysis, the plant concept is systematically analyzed on the basis of process flow diagrams, process description, equipment lists, and operating experience (e.g., reliability and maintainability, RAM). An initially higher expenditure on equipment not

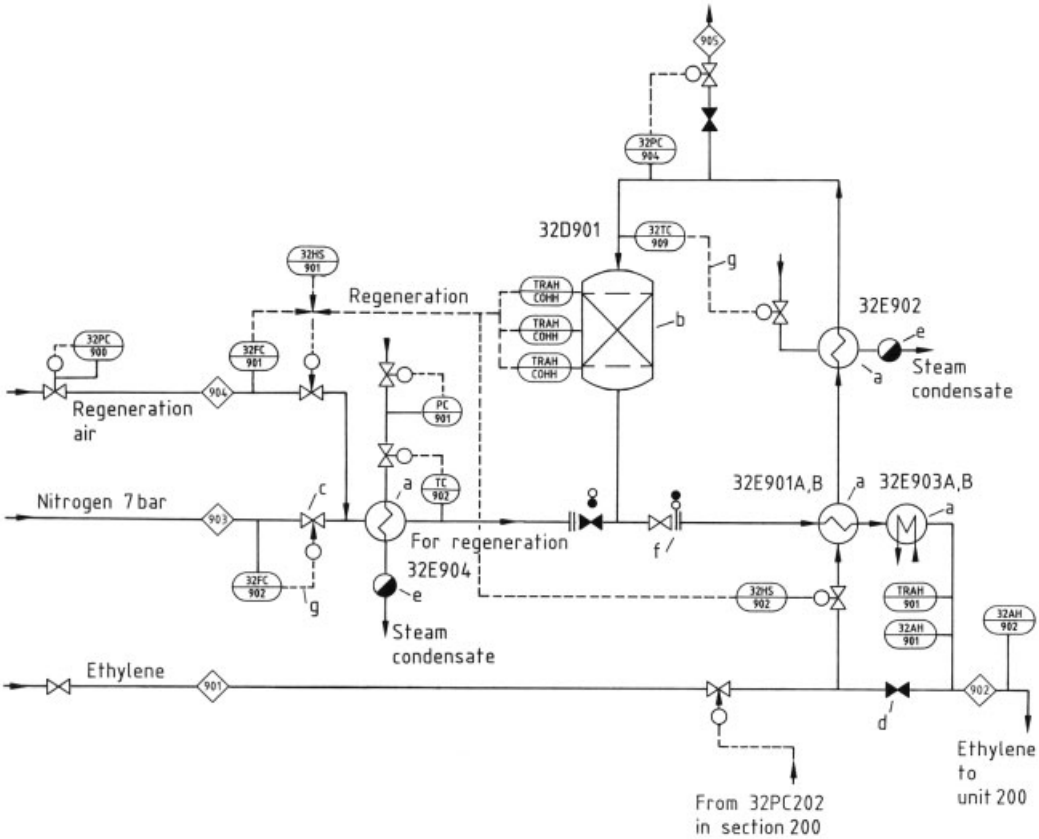

Stream no.		901	902	903	904	905
Designation		Ethylene from unit 100	Ethylene to unit 200	Nitrogen to 32D901	Regeneration air to 32D901	Regeneration off-gas
Mass flow	kg/h	10277.50	10277.51	3125	161	3286
Mass flow	kmol/h	366.33	366.33	111.55	5.58	117.13
Volume	m^3/h	8205.69	8205.69	Max.2500	Max.125	Max.2625
Temperature	°C	15	33	15	1520	Max.200
Pressure	bar	22	16	7	3.7	1.1
M_r	kg/kmol	28.06	28.06	28.01	28.9	28.05
Density	kg/m^3	26.11	17.87	8.29	4.45	0.80
Design mass flow	kg/h	13000	13000	3125	178	3303

Figure 10. Individual process flow diagram for a plant section
a) Heat exchanger; b) Purification tower; c) Control valve; d) Block valve; e) Steam trap; f) Blind; g) Control loop (temperature, pressure, flow)

only results in better availability but also lowers repair and maintenance costs.

Optimization of process engineering must then be investigated. The process design objective of a unit operation or a plant can be achieved with various combinations of capital costs and variable production costs (feedstocks, auxiliaries, utilities, energy consumption, disposal) as the following simplified examples show:

1) In plants with high gas throughputs, the cross-sectional areas of piping, fittings, and

reactors/apparatus determine the pressure drop in the plant and thus the energy that must be expended to transport the gaseous media. Large cross sections, with correspondingly higher capital costs, lead to lower energy costs and vice versa.

2) In heat exchangers, such as those used for waste-heat recovery, the quantity of transferred heat increases with increasing exchange area (= higher capital costs) and increasing flow velocities (= higher pressure drop = higher energy consumption) and vice versa.

In each case, there is an optimum for given economic parameters. The second example is more complicated, because not only must the cost optimum for a given quantity of heat be determined but also the optimal quantity of heat to be exchanged. It must further be taken into consideration that the cost of waste-heat transfer to the environment decreases when more heat is recovered.

Similar examples could be cited for the optimization of conversions, the optimization of the product mix when products are coupled, and the simultaneous minimization of byproducts and residues and of their reprocessing or disposal.

The examples lead to the following conclusions: identical plants do not have a unique optimum; instead, they have a variety of optima that depend on the crucial technical and economic constraints of a given site location. Optimal design of the plant is based on economic parameters employed in the design phase. Subsequent deviations may partly offset one another or may accumulate in the result.

The objective is thus to assign cost factors to the functional dependences of the process design or, in mathematical terms, to express the dependence of the costs on the engineering variables in the form of a cost function. Limiting parameters always have to be introduced, e. g., emissions, plant reliability, product quality, and safety factors for equipment and material (to ensure maximum lifetime and on-line time). Optimization can be solved by a suitable method (linear and nonlinear optimization, mixed-integer programming, gradient and search procedures).

The large number of variables in a plant design and the high degree of interaction in plants with many process loops mean that a facility cannot be completely optimized during the design phase. The dimensions of such a program exceed any reasonable and justifiable computer effort.

It is useful to begin by performing an economic – technical analysis of the process based on the "standard design" (not yet optimized). This analysis reveals the process stages in which significant fractions of the costs are incurred, consumed, or transformed, and expresses them as a proportion of the total costs. As a rule, a few key points dominate the economics of the entire plant. Optimization work can thus be facilitated right from the start.

If individual cost items in subsequent operating periods are expected to deviate significantly from the values chosen initially, new calculations can be made for the few key points, taking maximum and minimum values (sensitivity analysis).

A practical application of this method is the comparison of designs "on an evaluated basis." The costs for the necessary capital (often with allowance for the expected return on investment and tax considerations) and for capital goods are represented in a formula. The actual requirements for capital and capital goods are substituted into the formula. The lowest resultant value identifies the best concept for the selected premises. Details of cost structure and financing need not be known.

Optimization of Individual Tasks. For the calculation of various reactor types, see → Mathematical Modeling and → Model Reactors and their Design Equations.

Enthalpy – temperature (H/T) diagrams and exergy analysis (formerly also availability analysis) are being increasingly used in conjunction with simulation programs to optimize the energy economy of the plant [57]. This is especially worthwhile for processes with high heat turnover or high compression ratios. Apparatus costs must be determined separately.

The design of heat-exchanger networks is based on analysis of the heat fluxes in the network as a whole. One representative of such methods is the Linnhoff "pinch" method, see → Pinch Technology. This technique uses the H/T diagram with cumulative curves for the quantities of heat dissipated and absorbed in various sections of the plant at defined tempera-

tures. The method can be applied to utility systems and to the integration of thermal engines [58]. Software is available for the design of heat-exchanger networks by this method (e.g., HEXTRAN and ADVENT). An alternative method for heat-exchanger/utility networks is based on the mixed-integer method [59].

Synthesis-gas processes offer an example of integrated loops. These reactions are mostly exothermic, and an attempt is made to transfer the excess heat of reaction to steam instead of cooling water. Use of waste heat from gas production to supply the heat required for gas purification and converting processes has led to a variety of integrated loops with significant reductions in operating costs, for example in ammonia and methanol synthesis and in the production of nitric, sulfuric, phthalic, and maleic acids.

Another optimization possibility is the use of heat pumps. The mechanical compression of vapors and their condensation at higher pressure offers interesting solutions with relatively low investment costs, for example in the distillative separation of components with similar boiling points (e.g., such as ethylene – ethane and propene – propane).

The operability and economics of integrated loops must be checked during process design. This can be done by using LINNHOFF's method of heat integration analysis [60], [61].

Process Synthesis. The development of optimal combinations of unit operations is an iterative process that is now carried out with the aid of simulation programs. An advanced program performs computer-aided process synthesis by automatically generating design variants and selecting the best ones under consideration of uncertain data, i.e., when material properties, thermodynamic data, and kinetics of partial processes are incomplete. Results can be applied to some practical problems [62–64] especially to heat-exchanger networks (HENs) and utility systems.

The design of separation processes is much more difficult because of the great number of species present. Process synthesis has therefore not reached a comparable level. A survey of the design of column cascades appears in [65].

Optimization has to be customized for every application, key points can be identified only by economic and technical analysis of a process. The boundaries between "routine" optimization during process design and a special, supplemental optimization study are not clearcut. The practical operability of a plant should never be ignored. Overly complicated circuits, highly complex control systems, dewpoints and corrosion limits that are too close together, and increased maintenance costs can all wipe out calculated cost savings.

Continuous optimization is needed throughout the service life of a plant. Digital process monitoring and control systems allow appropriate data acquisition, storage, and archiving. Process data processing systems print data in the form of diagrams, graphs, and tables. They are an indispensable aid in the commissioning and optimal operation of plants. Process-based simulation programs allow on-line and off-line balancing and prompt detection of abnormal occurrences.

3.3.2. Safety Aspects and Environmental Control

In the following section the essential elements of safety and environmental control are treated – relevant to the design and construction of a chemical plant, using Germany as an example. The laws and decrees cited are only valid in Germany, however, there are similar laws in other countries, e. g., those issued by the EPA and OSHA in the United States and the "Stoomwezen" in The Netherlands.

The materials present in a chemical plant, their processing, and processing equipment constitute a source of risk to persons and property in the plant area and its surroundings. The level of risk depends on the nature, extent, and probability of occurrence of injury or damage [66]. Safety engineering institutes measures that reduce (limit) the risk to a degree acceptable to the public require that:

1) Potential hazards must be identified
2) Effective safety standards against these hazards must be established
3) The standards must be transformed to engineering and/or management safety practices
4) It must be proved that the safety level meets requirements
5) The effectiveness of the adopted measures must be evaluated and improved [67]

Safety-related functions are the responsibility of governmental supervisors, the plant operator, and the engineering contractor. In approving the plant, the regulatory authorities not only ratify the safety objectives but also evaluate the scope and quality of the safety measures. The operator is responsible for correct execution of these measures. The design and construction firms must assist the operator in complying with the standards [68–70].

3.3.2.1. Protection Against Emissions

The following important environmental protection goals are necessary for the approval of a new plant and are therefore the concern of authority engineering:

Landscape and Surroundings. The erection of a plant on a site not expressly intended for industrial use can run into difficulties [71]. It is also necessary to clarify in advance what effects the plant will have on nearby residential areas and what levels of emissions and noise are acceptable in industrial areas [72–75].

The appearance of the plant must be appropriate for the surroundings. This factor is governed by restrictions on building height, design, materials, color, the construction of visual barriers, and landscape plantings.

Air Pollution. Pollution control begins with the classification of hazardous substances and prevention or minimization of emissions to the atmosphere. Release under normal operating conditions is prevented by using appropriate design and process conditions (e.g., gas collection and recovery systems, use of low-polluting burners, catalytic gas purification, use of high-quality flange connections and seals) [76–78].

Safe operation of a chemical plant must be ensured to prevent or minimize the hazard to the surroundings. Possible sources of danger (accidents) include release of large amounts of hazardous substances (gases, liquids, solids), fire in the plants, and explosions [72].

Safety aspects must be taken into consideration in process design, plant layout, and equipment selection; in the construction, operation, and maintenance of the plant; and in the training of plant personnel [69].

The "hazard and operability study" (HAZOP) is often used and can reveal weak points in plant equipment and operation while the facility is still under design. This method of risk analysis has been proven in years of use [79–81].

Wastewater. The wastewater generated in a plant is collected and treated in systems classified according to water quality (e.g., severely contaminated, moderately contaminated, uncontaminated, rain water). If the systems are properly designed, the treated wastewater can be released into a system approved by the local authority [82], [83].

Wastewater produced in case of fire (fire fighting water) must be collected in a retention basin whose capacity is usually sufficient to hold the water used in 1 h against the design fire- case [84].

Soil. Soil protection is afforded by sealing all plant areas that might be contaminated by hazardous liquids or solids under normal operating conditions or in an accident [85].

Waste. "Waste" denotes all those substances and parts that cannot be recycled to the production process or otherwise reused.

As early as the process selection step, special attention must be paid to waste prevention because waste is difficult to dispose of. In large chemical complexes, wastes from one process can often be used as feedstocks in other processes. This alternative should be investigated closely in feasibility studies performed at a very early stage of the project [86–88].

3.3.2.2. Noise Control

The primary objectives of noise- control design are in compliance with contractual and legal provisions while ensuring that the plant is easy to maintain and runs economically. Noise can be controlled at the source by selecting low-noise products and processes. Low-noise equipment often has the welcome side effect that it offers low energy consumption and longer service life.

Basic concepts and research in acoustics are discussed as examples in [89–92].

Regulations. Noise- control regulations are intended to protect the surroundings and the plant personnel.

Protection of the Surroundings . The basis for immission noise limits in Germany is the Bundes-Immissionsschutzgesetz (Federal Immission Control Act) and the TA Lärm (Engineering Directive on Noise Control) contained in it [93]. The directive lays down permissible immission noise levels as a function of land use (Table 2). The site of measurement varies; as a rule, it is 0.5 m in front of an open window of the residence. Regulatory practice has been to treat the permissible levels cumulatively: if several plants emit to one receiver, the maximum level must be distributed among all the emitters. Furthermore, if a plant is expanded, its permissible noise-level contribution must not be exceeded.

Table 2. Permissible noise immission levels

Type of builtup area	Standard value, dB(A)	
	Day	Night
Exclusively industrial utilization	70	70
Mainly industrial utilization	65	50
Equal housing and industrial utilization	60	45
Mainly housing areas	55	40
Exclusively housing, health resorts, and hospitals	50	35

This approach ensures that the permissible total noise level in the residential area is not exceeded. On the other hand, new plant sections are only approved if they meet strict noise standards.

Protection of Plant Personnel. The basis for noise regulations at the workplace in Germany comprises the Arbeitsstätten Verordnung (Workplace Regulation) [94] and the relevant sections of the Unfallverhütungsgesetz Lärm (Accident Prevention Code) [95] which contain references to DIN standards and VDI guidelines. These regulations set a maximum personal noise level of 85 dB(A) for an 8-h shift and recommend the use of personal hearing protection above 85 dB(A). Personal hearing protection must be worn at levels over 90 dB(A).

The most important regulations for plant design follow:

Measurement of noise at machines	DIN 45 635 part 1, and following parts
Average level and assessed level temporal fluctuating sound processes	DIN 45 641
Sound propagation in the open	VDI 2714
Assessment of working noise in the neighborhood	VDI 2058 sheet 1
at workplace regarding danger to hearing	VDI 2058 sheet 2
with regard to various occupations	VDI 2058 sheet 3
Sound projection from industrial construction	VDI 2571
Noise from piping	VDI 3733
Noise abatement for ventilation and air conditioning plants	VDI 2081
Personal soundproofing	VDI 2560
Soundproofing by mufflers	VDI 2567
Soundproofing by shielding	VDI 2720 sheets 1 + 2
Soundproofing by metal cladding	VDI 2711

Design. The engineering firm must guarantee compliance with noise regulations at the workplace and in residential areas. The contract should also stipulate the type of noise measurement and the operating conditions at the time of measurement.

Planning for noise control begins in the preliminary design phase and continues through basic and detail engineering. The engineer must carry out the following activities:

1) Calculate the permissible sound level for the entire plant on the basis of the maximum level allowed in the residential area and available studies on the workplace noise-level limits.
2) List noise-producing components and establish their permissible sound levels. Noise-protection practices are dictated by practical and economic considerations.
3) Prepare noise specifications (including permissible sound levels) as part of the bid specifications for all noise-emitting equipment.
4) Check the bids (e.g., for machinery, control valves) to ensure that noise requirements are satisfied.
5) Compile specifications for silencers, hoods, and insulation; check bids and orders for these items.
6) Compile noise specifications for the building.
7) After the design is complete, write a design report on residential and workplace noise levels.
8) Inspect installation of equipment at the construction site.

9) Measure noise levels of noise-generating equipment at the workshop where it is manufactured, and also later when it is installed in the plant and the plant is on stream.
10) Prepare a final report.

Manufacturers of noise-generating equipment have many ways of reducing noise. Examples follow:

Electric motors: Reduction of cooling air requirement, use of low-noise cooling fans, use of improved insulation.
Control valves: Division of large-expansion cross sections into smaller areas (perforated cage), division of pressure release into multiple stages, optimized flow control.
Air coolers and small cooling towers: Reduction of rotor peripheral velocity, modification of blade profile, increase in number of blades, use of low-noise gears.
Pumps: Optimization of impeller design, avoidance of cavitating conditions, reduction of impeller peripheral velocity (below 45 m/s).
Steam generators and process furnaces: Low-noise, forced-air burners; use of ceramic-fiber linings on interior walls.
Compressors: Frequency- controlled drives, low-noise oil systems, direct drives (no gears), low-noise surge-limit control.

Important secondary practices of noise abatement include:

1) Soundproof enclosures
2) Soundproof hoods and barriers
3) Soundproofing insulation on piping, ducts, and machine housings
4) Silencers in the form of absorbers, resonators, or combinations of both types

In most chemical plants, on-battery sound pressure levels can be held to 85 dB(A) by means of careful design. Large compressor sections are an exception; levels up to 100 dB(A) can be expected and such areas must be designated as high-noise areas.

Investment costs for noise control normally range from 0.5 to 3 % of total plant material and installation costs. These percentages are exceeded if residential restrictions make it necessary to enclose the entire plant.

Plant safety and reliability must not be impaired by noise- control measures. Low-noise machines must therefore be preferred over loud machines with soundproof enclosures. If hoods or enclosures cannot be avoided, accessible areas inside noise enclosures on gas-handling devices must be provided with adequate ventilation, gas alarms, and possibly fire fighting systems. Noise- control barriers and shielding must not block escape routes.

3.3.2.3. Occupational Safety and Health

The principal requirement in occupational safety and health is that physiological and psychological burdens on the employees caused by working and other conditions must be limited to a generally accepted level. Measures must also be taken to prevent or control risks in case of an accident [70].

The occupational safety and health authorities have defined general conditions for ventilation, lighting, and ambient temperatures at workplaces, protection against weather and noise, and the safe use of traffic routes inside the plant. Accident insurance regulations are especially important for plant layout and process design of chemical plants. These rules concern: harzardous operations, work involving hazardous substances [77], [78], use of special auxiliary equipment, and ergonomic design of the workplace [96]. Suitably equipped "social" facilities such as rest and changing rooms, washrooms, toilets, and medical stations must also be included in the design [94].

Warning of unavoidable dangers must be given, and appropriate protective measures must be instituted (e.g., signs marking fire or explosion hazard zones and appropriate safeguards against ignition). Escape routes and protected areas with emergency lighting, fixed personal protection facilities (emergency showers), and alarm systems must be provided [94], [97].

Appropriate layout of buildings and apparatus or enclosures around particularly dangerous equipment help to minimize injuries and damage [96].

Effective fire fighting measures include short access routes as well as fixed or mobile fire-extinguishing equipment with an assured supply

of extinguishing media and an adequate action radius.

Occupational safety and health measures also apply during construction of the plant. They relate to the structural design and size of the plant and the construction methods. As much as possible of the equipment used in subsequent normal operation of the plant must be available during commissioning for protection of the workers [96].

3.3.2.4. Plant Availability

Capacity and profitability calculations for a chemical plant are normally based on 8000 h/a on-stream (corresponding to roughly 330 d/a).

Shutdowns are generally planned at intervals of one or two years to allow cleaning of heat exchangers, apparatus, and piping; charging of new catalysts and chemicals; replacement of worn mechanical parts; and prescribed inspections by regulatory authorities.

Unplanned shutdowns are usually caused by mechanical defects or automatic emergency shutdowns when operating parameters are outside the limits for normal plant operation.

Important items of equipment are duplicated (standby equipment) to ensure continuous operation between scheduled shutdowns. This is standard practice for continuously operated pumps and reciprocating compressors. Standby turbine compressors are not necessary because of their longer maintenance intervals. In auxiliary systems (e.g., the lubrication system), however, the necessary reliability is provided by installing standby pumps and filters.

If operating conditions make regeneration or cleaning necessary during production, standby equipment is again used (e.g., fixed-bed reactors, molecular-sieve adsorbers, dryers, and filters).

Piping (as used to convey suspended solids or powders) may be susceptible to plugging. Installation of standby piping or flushing connections may be desirable.

Measures to ensure reliable plant operation during power outages must be considered at an early stage (emergency planning).

The drives of important process equipment (e.g., cooling-water pumps, instrument-air compressors, boiler feedwater pumps) are usually dual (electric motor plus steam turbine).

Emergency power-generating units are needed to maintain safety-relevant drive and control functions (including emergency lighting, communication facilities, and computer-aided process control systems) during power outages. In the case of particularly critical equipment, a separate power source (e.g., battery, instrument-air reservoir, nitrogen reservoir) must be provided to bridge the startup time of the emergency generator.

Control functions and process control systems play a vital role in reliability and safety. Relevant process parameters and the points where they are to be measured must be defined during process design so that indicators and alarms will warn the operating personnel promptly whenever operating conditions become abnormal.

Allowance must also be made for operator error so that consequential damage (e.g., overheating or pressure buildup in uncontrolled exothermic reactions) can be avoided. Automatic emergency-shutdown devices are provided for such cases. They are often redundantly included in both the measurement and shutdown systems. They must be regularly inspected by the plant operator or the regulatory authorities.

3.3.2.5. Authority Engineering

The construction and operation of chemical plants are affected by many legislative provisions and regulations that are concerned with environmental protection and plant safety [68]. These requirements have major consequences for plant design and construction. It is no longer sufficient to comply with all laws, standards, and specifications applicable to a given plant at a given site. Instead, a formal approval process, usually with public participation, must be gone through in the design phase. Only then can construction of the plant begin [98], [99].

The length of the approval process depends on the type of plant, the environmental sensitivity of the plant site, and the nature of the approval procedure. Another important point is whether the project involves expansion of an existing facility or an entirely new "grass-roots" plant on a previously nonindustrial site.

The time taken for an application for a construction permit to be approved is usually six

months to a year, but sometimes longer. This is preceded by a lead time of about a year, during which the conditions for the plant are discussed and preliminary talks are held with the regulatory authorities. The total span from the investment decision to groundbreaking is therefore about 1.5 years. This is a substantial fraction of a total project duration of 2 – 3 years.

The granting of construction and operational permits requires close collaboration between the plant owner, the engineering firm, and the authorities (Fig. 11). All activities aimed at devising an approved plant concept are grouped under the heading "authority engineering" and are usually the responsibility of the engineering firm. Typical activities carried out during the design phase are listed below.

Feasibility Study/Preliminary Design . The feasibility study and preliminary design include technical and economic optimization of the process with minimization of environmentally harmful factors. The following parameters are established

1) Nature and quantities of substances present
2) Waste and residues
3) Wastewater
4) Emission of air pollutants
5) Type and design of plant equipment
6) Nature (open or closed) of processing systems
7) Safety and reliability standards [77], [78]

As soon as site selection is complete, the following questions must be answered:

1) Which authorities are responsible for approving the construction and operation of the plant?
2) Which laws must be observed?
3) Which regulations, especially local ones, must be complied with?
4) What is the public opinion at the intended location?
5) Has any nearby project gone through approval proceedings recently? If so, with what result? How much time did the proceedings take?
6) Will special restrictions over and above normal legislation and regulations apply to the site? [75], [100], [101].

In order to clarify these points, informal contact with the regulatory authorities should be initiated as soon as the initial concept of the plant is set.

Process Design/Basic Engineering. Results obtained during the preliminary phase are used for process design/basic engineering (e.g., for planning off-gas collection and combustion systems and wastewater systems). A preliminary safety analysis may be useful to identify potential risks (e.g., threat to the environment due to release of substances, or risk to the plant due to hazards in the vicinity). This kind of analysis is called an *environmental impact study* [73]. The results of such an analysis may influence the plant layout [72], [75].

Detail Engineering, Construction, and Commissioning. Many authority engineering activities take place in the detail engineering phase. They include:

1) Preparation of project documentation
2) Preparation of a detailed safety analysis
3) Engineering escort duty during approval proceedings
4) Implementation of design changes required during approval proceedings
5) Management of partial approval procedures that take place in parallel with the main approval procedure
6) Selection of experts and technical cooperation with them to clarify detail questions raised during approval proceedings

The activities of engineering and authority engineering must be well coordinated if the progress of work at the construction site is not to be held back.

An important point for the success of a project is that the engineering activities concerned with early civil work activities should be taken care of early. This is especially important for the plot plan and buildings; escape, emergency and access routes; and fire fighting concept.

The authorities grant partial approvals (e.g., for civil work, erection, and commissioning of plant sections) so that progress is not unnecessarily delayed. They also check that relevant regulations and provisions are observed during construction and installation. When construction is

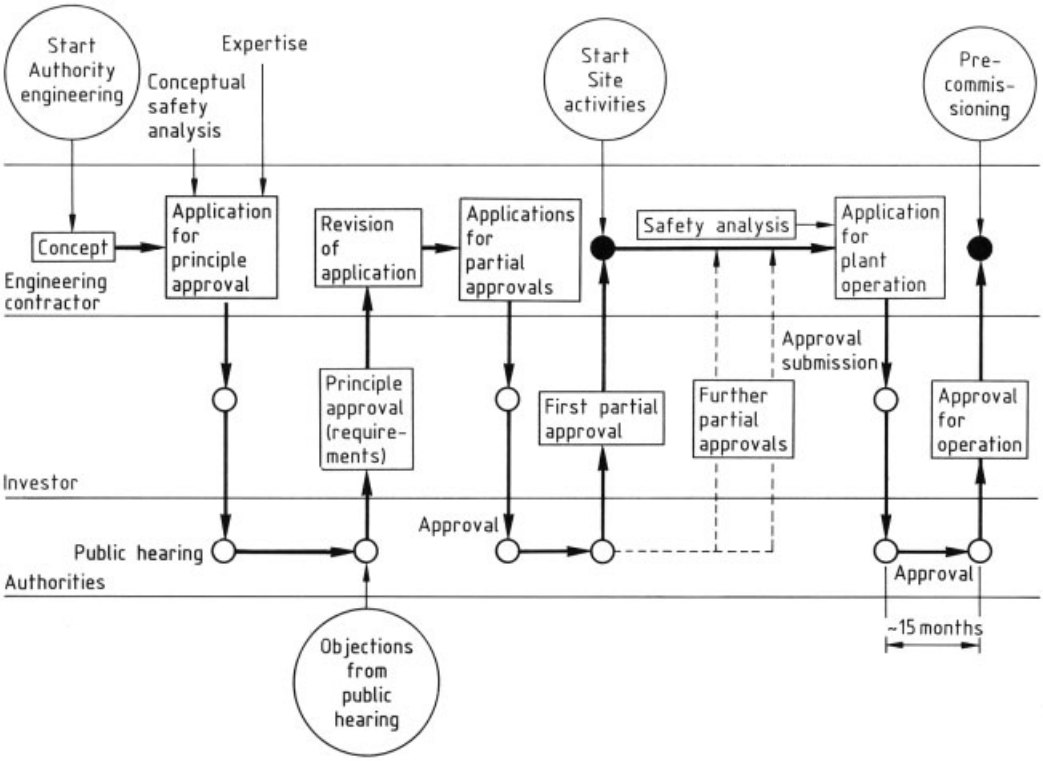

Figure 11. Principle flow scheme of German authority approvals procedure

complete the entire plant is examined by the authorities. Deficiencies must usually be remedied immediately. The authorities only grant permission for commissioning when the acceptance report has been made.

During commissioning proof of compliance with approved levels of emissions, wastewater values, etc. must be submitted. Measurements are difficult to perform and time consuming. If improvements to the plant are required at this stage, they may be very expensive.

3.4. Basic Engineering

The focus has so far been on process design (process flow diagrams and parameters such as operating temperatures and pressures and flow rates). Now the geometric dimensions of individual equipment items, the design temperatures and pressures, the materials of construction, and the layout of the entire plant must be established. (The main elements of basic engineering documentation are described in Section 5.1).

3.4.1. Equipment Specification from the Process Engineering Standpoint

Equipment dimensions and capacities are dictated by the process. Data from the process flow sheet can be used for the sizing of process equipment, machinery, piping, etc. Examples of data that can be derived in this way are the diameter, number of trays, and tray spacing for distillation towers. These design data are entered in process engineering data sheets (Fig. 12) that contain all relevant specifications for the specialist engineer. For example, the data sheet for process equipment includes a schematic drawing with overall dimensions, operating and design pressures and temperatures, number and nominal diameters of nozzles and manholes, material of construction, corrosion allowance, insulation thickness, etc.: in short, the information that

the equipment design engineer needs in order to perform strength calculations and prepare a more detailed drawing. For pumps, the data sheet must show normal, maximum, and minimum flow rates, inlet and outlet pressures, operating temperature, material of construction, type of medium, and physical properties of the medium. On the basis of this information the mechanical engineer can select the optimal pump with the best efficiency.

Specifications are prepared similarly for control systems, safety valves, and all other items of plant equipment. The nominal diameters of piping are calculated for the specified flow rates, physical properties of the streams, and economically acceptable pressure drops.

3.4.2. Materials of Construction

See also → Construction Materials in Chemical Industry.

Materials of construction in chemical plants must be able to withstand mechanical, chemical, and thermal stresses and must not be attacked by the medium with which they come in contact. Two criteria must be met:

1) Materials must be approved for use in pressure-bearing parts under the pertinent regulations, i.e., their guaranteed values, processing, and permissible service conditions are defined and can be reproduced at any time.
2) Materials must be suitable from the corrosion standpoint. Their behavior and properties should show little or no change under the action of the media with which they come in contact.

Material selection should be solved by close collaboration between the materials specialist, the designer, and the process engineer or chemist [102–109]. The materials most commonly employed in process engineering are unalloyed, low-alloy, and high-alloy steels. Both solid steel and cladded steel fabricated by rolling, weld overlaying, or explosion bonding are used in vessels, towers, heat exchangers, storage tanks, piping, and other equipment.

Stresses. Materials of construction must withstand temperatures from ca. −180 to 1100 °C and pressures from vacuum to several hundred bar.

The selection of a material is dictated by three criteria: mechanical stress, thermal stress, and chemical attack. Seldom do these three types of stresses occur singly; the usual case, in which two or all three are present at once, governs material selection.

The most important ferrous materials used in plant construction, are the following steels:

Predominantly mechanical stress:
- Structural steels
- Fine-grained structural steels
- Quenched and tempered steels
- Steels for low-temperature service

Coexisting mechanical and thermal stresses:
- Steels for high-temperature service
- High-strength alloy steels for high-temperature service
- Heat-resisting steels
- Steels for hydrogen service at elevated temperature and pressure

Chemical attack together with thermal and mechanical stresses:
- Ferritic chromium alloy steels
- Austenitic chromium – nickel steels
- Ferritic/austenitic steels (duplex steels)

The selection of steels for predominantly mechanical stress depends on strength, toughness, and weldability. States of mechanical stress in individual parts are often incompletely known. Design is therefore based on approximate rules derived from simple loading modes (tension, crushing, bending). The time dependence of load (e.g., static or cyclic) must also be considered.

The operating pressure, operating temperature, and number of load cycles are crucial for strength calculations of apparatus and piping. The temperatures determine the strength (yield point). The use of fine-grained structural steels, with higher yield points than normal carbon steels allows design of equipment with thinner walls and thus results in significant savings in weight and welding work.

Equipment must be sized so that it does not fail by ductile fracture, brittle fracture, fatigue, or creep (Table 3).

The choice of steels for chemical service, or where chemical attack occurs together with thermal and mechanical loads, is difficult because

LURGI

Lurgi GmbH

Vessel Engineering
Pressure Vessel
Process Data Sheet

P-CUYO-63333
Docum. 1200
Unit 500
1.9.2 /15

Sketch

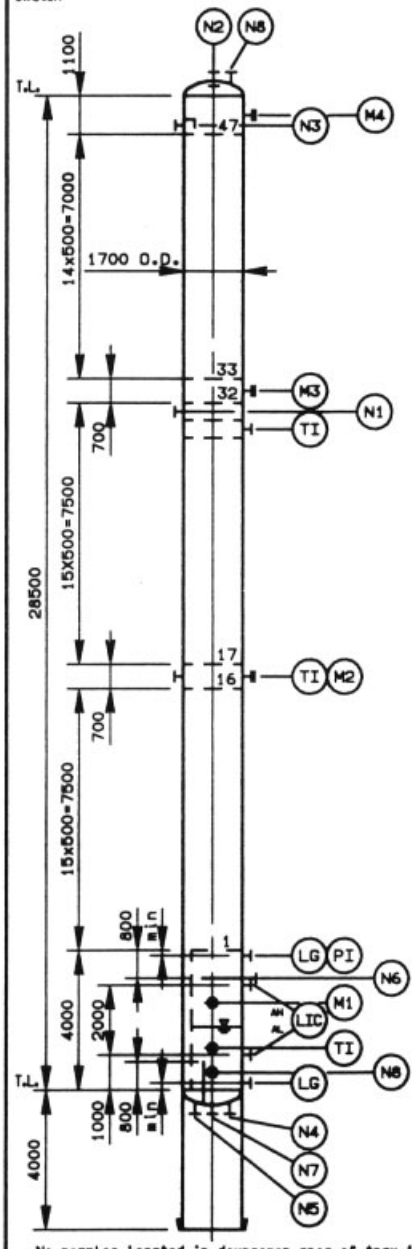

No nozzles located in downcomer area of tray 1

☒ Mark with a cross where applicable

01	Item No.	No. req.	K 501	1 Piece
02	Designation		Deethanizer	
03	Position		☐ horizontal	☒ vertical
04				
05				
06	Medium		Light hydrocarbons	
07	Corrosive components		—	
08	Density			kg/m³
09	Operating temperature			°C T:52,7/B:75,9
10	Operating pressure			bar T:30,8/B:31,0
11	Design temperature			°C 100
12	Design pressure			bar 35
13	Cycle loading			☐ yes2) ☒ no
14	H_2 partial press.(abs.)	H_2S service 3)	bar -	
15	Material / -group	Corrosion allowance	- mm	CS. 1
16	Heat treatment, for process reasons			☐ yes4) ☒ no
17	Lining / cladding 1)			kind —
18	~~inside~~, outside insulation 1)			kind yes
19				
20				
21	Internals due to process reason:			
22	Characteristic data:			Material:
23	Type of trays			Sieve trays
24				
25				
26				
27	Tray spacing			mm 500
28	Attachments mounted to vessel directly:			
29	☐ no ☐ yes	☐ Evaporator	☐ Siphon	☐ Stack
30		☐ Submerged pump	☐ Agitator	
31		☐		
32	Table of nozzles			

33 Item	No. req.	Nom. Size	Designation
34 N1	1	3"	Feed
35 N2	1	4"	Overhead
36 N3	1	4"	Reflux
37 N4	1	3"	Bottoms
38 N5	1	10"	To Reboiler
39 N6	1	14"	From Reboiler
40 N7	1	2"	Drain
41 N8	2	2"	Steam out + Vent
42 M	4	24"	Manhole
43 LIC	2	1½"	Level
44 LG	4	¾"	Level
45 PI	1	¾"	Pressure
46 TI	3	1½"	Temperature
47			
48			
49			
50			

51 Remarks:
Pressure values are in gauge, if not stated otherwise
1) Delete if not applicable
2) Expected load cycles incl. start-up: .....
.....
3) Classification in to service groups acc. to Lurgi NB 0099
4) Kind of heat treatment:

Completed: Wh 7.11.88	Process: PVP	Job or Project No	Job:	Revised Line Modified:		↑
Checked:	Document-Type: 2 5 4 No.: D 6 3 3 3 3 0 0 4 L 0 1 5 8 9			Sheet:	Rev.:	Date:	Checked:

Figure 12. Data sheet with process information for a deethanizer

several types of corrosion are possible, each resulting in a different type of failure. The problem of material selection in this case is often very complicated because corrosion is due to multicomponent systems.

Table 3. Nature of failure of materials

Nature of fault	Cause	Measure
Ductile fracture	unacceptably high loads	calculation with elasticity limit or safety correction values
Brittle fracture	stress also below the permissible elasticity limit by influences from	calculation of the brittle fracture safety with fracture mechanical values on the basis of
	1) multiaxial state of stresses 2) thermal stress 3) state of material 4) state of defect 5) geometry of component	the combined effect of material faults and stresses highly dependent on 1) material production 2) quality of manufacture 3) fault finding and 4) fault description with nondestructive testing
Fatigue	introduction of crack and crack propagation by alternating stress	calculation with conventional strength values for alternating stress (e.g., fatigue limit) taking into account all stresses as with the danger of brittle fracture
	fatigue crack propagation at existing fault sites	consideration of mechanical fracture principles and their mathematical utilization
Creep, time interval	time-dependent deformations and fractures at high temperature and stresses below the elasticity limit	1) calculation with creep strength 2) calculation of time yield limit 3) crack propagation by existing defects

Alloys. Corrosion-resistant alloy systems have been developed which form a passive layer that affords protection against corrosion. Such alloys must remain stable during subsequent treatment (e.g., welding) so that phase precipitation does not affect the passivity of the material.

The most commonly used alloy systems for chemical plant construction are those based on iron, nickel, and copper. Figure 13 shows these three groups schematically, with the maximum concentrations of the alloying elements singly or in combination. A great number of alloys with the desired properties can be produced with these systems.

Other metals and alloys used in plant construction are aluminum, titanium, zirconium, and tantalum. See also → Construction Materials in Chemical Industry.

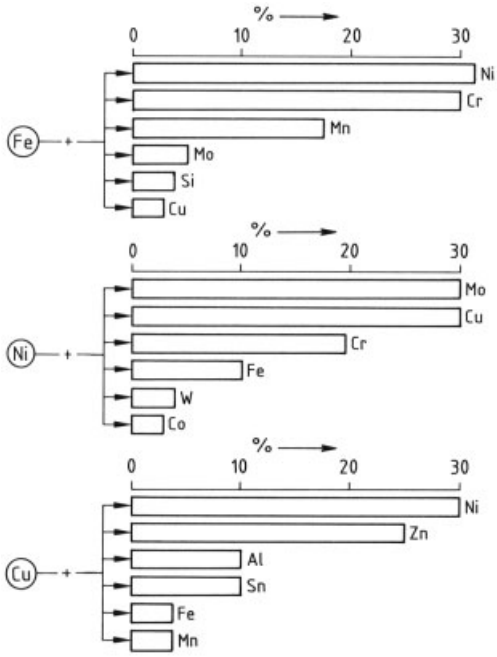

Figure 13. Alloy systems most commonly used in chemical plant construction

Corrosion. For a detailed description of corrosion, see → Corrosion. Corrosion phenomena (e.g., selective corrosion, diffusion, crevice corrosion, and pitting) must be given special attention but are not amenable to design calculations.

The most frequent form, general corrosion, can be dealt with by appropriate corrosion allowances of ca. 1 – 5 mm. The corroded material must not, however, interfere with the process or affect the product (discoloration or flavor changes in the pharmaceutical, beverage, and food industries).

Charts, tables, empirical values, and processing guidelines are available for material selection according to corrosion criteria. Carbon steel, for example, is attacked only slightly by water, but severely by water in the presence of air. Pitting occurs if air-containing water is heated, or if oxygen is present in steam and condensate piping; this often rules out the use of car-

bon steel. Although stainless steels resist many media, they also suffer pitting or stress corrosion cracking in the presence of halogen compounds.

Material selection has to take into account both the chemical nature and the flow of the medium. Rapid motion of a liquid along a pipe wall can accelerate corrosion, whereas a protective film often forms on the surface when the velocity is low.

Intergranular stress corrosion cracking may be caused, for example, by alkaline solutions above 100 °C under pressure in welded vessels of unalloyed steels.

Transgranular stress corrosion cracking can occur in equipment made of austenitic Cr – Ni steels, while crevice corrosion results from improper fastening of tubes in tube sheets; these types of corrosion are both accelerated by vibrations (corrosion fatigue).

Electrolytic corrosion can take place when metals widely separated in the electrochemical series (e.g., copper, iron, and aluminum) are not insulated from one another in the same piece of equipment apparatus and come in contact with conductive liquids.

Other corrosion-related problems are hydrogen embrittlement, embrittlement at low temperature, and high-temperature corrosion.

The same degree of corrosion protection can often be obtained with several materials. In such cases, price, service life, and profitability must be balanced [110].

Costs can be saved by applying more noble materials as a coating or cladding. Electroplated coatings cannot be employed in chemical plant apparatus because they do not offer long-term protection. Claddings of nonferrous metals, their alloys, or austenitic steels can be used if the contact between the corrosive stream and the load-bearing steel shell of an apparatus can be prevented by appropriate design.

In simpler cases, plastic coatings can be employed in place of cladding. Stoving finishes (duroplastic) have proved suitable for contaminated cooling water and weak acids in heat exchangers (service up to 240 °C). Polytetrafluoroethylene coatings are highly resistant to acidic and alkaline substances (up to 300 °C). Rubber and hard rubber, respectively, are effective against weak acids and salt solutions up to ca. + 100 °C. It should be kept in mind that plastic coatings tend to swell and unbond in the presence of organic substances, especially at elevated temperature.

3.4.3. Plant Layout

The layout becomes a high-priority item during basic engineering [111].

For large plants, the planner begins with 1: 500 layout sketches that merely show the space required for individual units. Such preliminary layouts are used to determine the most expedient arrangement with possible transfer points at battery limits for material and energy streams. Conventional layout techniques employ adhesive films and overlaying of transparent sketches; PC and CAD programs are also finding use.

The layout should include approximate data on the positions and sizes of storage areas for raw materials, intermediates, and end products, as well as pipe bridges, roads, underground piping, and sewerage. Control room and electrical substations, compressor buildings and service facilities, and road and rail connections are shown with their overall dimensions. The accessibility of plant equipment for repair and maintenance, construction aspects, safety, and inspection must be considered from the very start. The consequences of later expansion must also be taken into account.

If the existing information is adequate, the layout should be drawn to scale. The layout and the process flow diagram then permit assessment of the production sequence, mass transport, and storage (Fig. 14).

On the basis of the layout and specifications, buildings should be inserted next as needed (ground plan, number of floors, height between floors).

Layout Model. The layout can also be made into a block model (Fig. 15). Layout models have proved especially useful for complicated installations such as refineries and petrochemical plants. They convey a general impression of the final appearance of the plant, even at an early stage. Approximate 1 : 100 or 1 : 50 models of equipment items are made of foamed polystyrene or a similar material and can then be moved around to establish suitable positions and spacings.

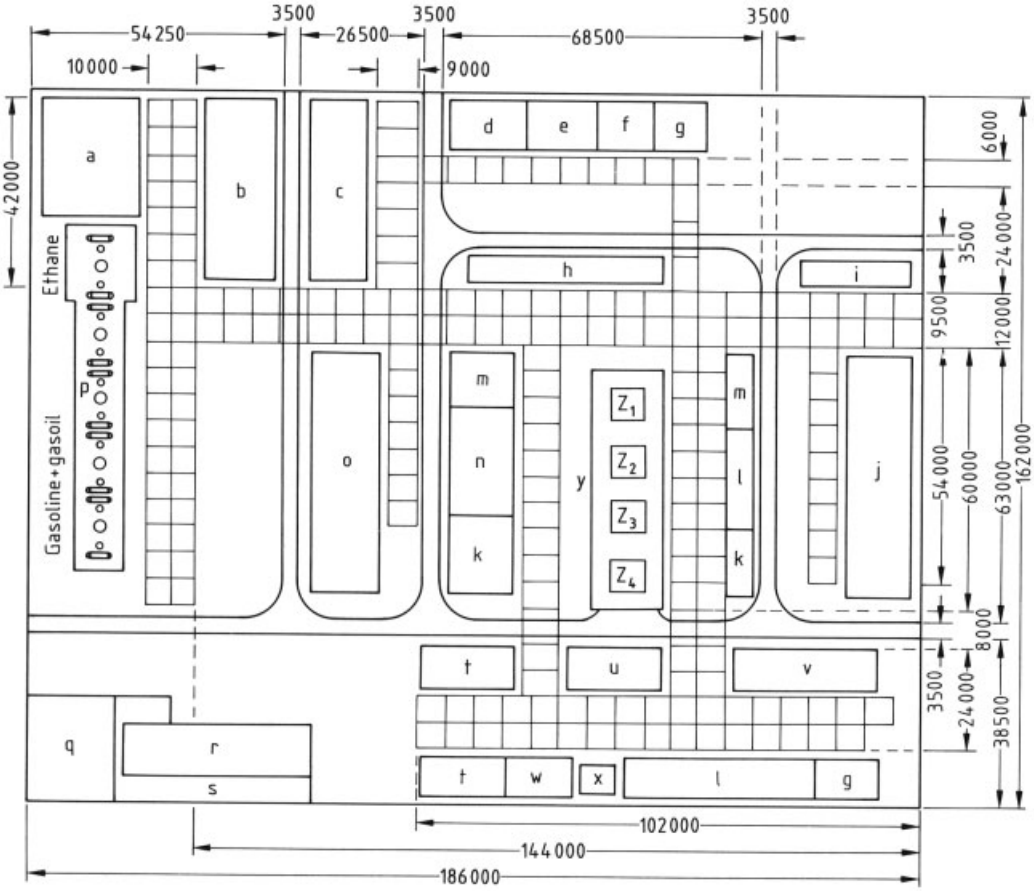

Figure 14. Plant layout plan
a) Steam generation; b) Process steam generation and fuel oil system; c) Oil wash with dispersed oil separator; d) Propene – propane separation; e) Debutanizer; f) Depropanizer; g) Deethanizer; h) Condensate and slop system; i) Hydrogen methanization; j) Hydrogenation; k) Propene refrigeration; l) Ethylene refrigeration; m) Charge gas compressor; n) Caustic wash; o) Water wash with gasoline stripper; p) Cracking furnaces; q) Substation, transformer station; r) Control room; s) Social facilities; t) Ethylene – ethane separation; u) Charge gas dryer; v) Acetylene conversion; w) Demethanizer; x) Cold box; y) Compressor house; z) Pumps, compressors

The layout model includes steel structures but does not cover the details of pipe runs. It mainly serves as a basis for discussion, permitting a number of fundamental questions to be clarified.

At this stage of basic design (at the latest), the design engineer is thus able to decide which parts of the plant will be outdoor or enclosed facilities, and whether it will be on one or more levels. A process plant should be outdoors unless there are pressing reasons to enclose it. This allows substantial savings in construction costs; machinery and equipment remain more accessible; and the danger of fire and explosion is reduced for processes involving flammable materials. The outdoor setting, however, entails more expense for shelter against weather, heat, and cold.

For some types of plant equipment, the layout criteria are nearly always the same. For example, turbocompressors, reciprocating compressors and their drives; coolers and oil circulation lines should be located in sheds or compressor buildings. Pumps, in contrast, are usually placed outdoors; they are installed in pump houses only if this is necessary on environmental grounds (e.g., for noise protection).

Production facilities that are sensitive to ambient conditions (e.g., film and fiber produc-

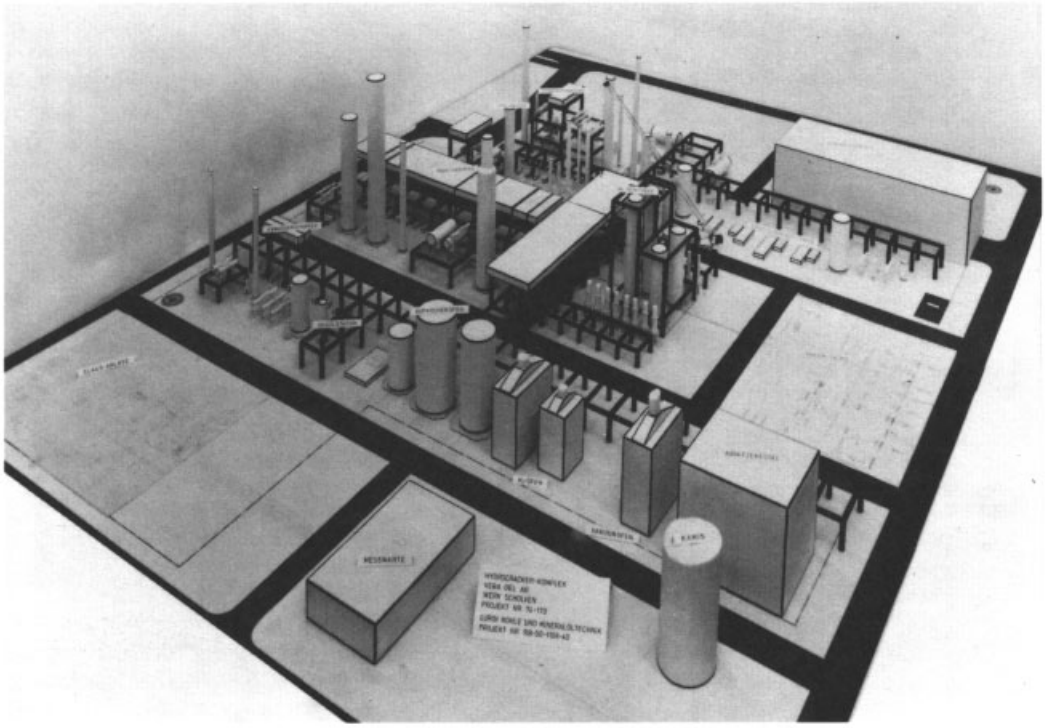

Figure 15. Layout model (hydrocracker complex), courtesy of Lurgi AG

tion, paper mills, pharmaceutical plants, catalyst plants, packaging facilities) must be set up inside buildings. Raw materials and products that are sensitive to moisture and sunlight must be stored in covered areas or enclosed buildings.

In contrast to the solidly constructed buildings required in Europe, buildings in tropical regions can be lightly constructed. Sheds open on two sides, affording protection against wind and rain only, may be adequate if the process does not call for climate control.

3.4.4. Preliminary Piping and Instrumentation Diagram

A piping and instrumentation diagram, based on the process flow diagram, is needed for more accurate calculations at the preliminary design stage. For the basic engineering package it should contain the following information:

1) All equipment and machinery, drives, piping or transport routes, and fittings (including installed spares)
2) Nominal diameter, pressure, material of construction, and design information of piping
3) Field instruments, control devices, and connections between them
4) Special fittings required for process control and safety, e.g., check valves, safety valves, level gauges, condensate drain lines
5) Significant dimensions of equipment and machinery
6) Essential data on materials of construction

3.5. Calculation of Plant Costs

Once the engineering documents for the calculations at the requisite accuracy are available, plant costs can be determined. It is useful to employ a scheme as shown in Table 4 for this. A "standard" scheme has been proposed by Aries and Newton [112] and other authors. The calculation is subdivided into three groups of items: equipment, bulk materials, and indirect costs; each has to be calculated differently.

The *equipment group* includes all itemized plant equipment such as towers, reactors, and heat exchangers, vessels and machinery that are

Table 4. Example of a calculation scheme for plant costs

Designation	% of 1	% of 1 + 2	% of 1 to 3	% of 1 to 4
Equipment				
Columns	12.5	7.4	4.7	3.4
Reactors	6	3.5	2.2	1.6
Heat exchangers	22.5	13.3	8.4	6.1
Tanks	18	10.6	6.7	4.8
Furnaces	10	5.9	3.7	2.7
Machines	26	15.3	9.6	7
Other equipment	5	3	1.9	1.4
Total 1	*100*	*59*	*37.2*	*27*
Bulk material				
Steel structures	8.3	4.3	3.1	2.2
Piping	25	14.7	9.3	6.7
Control systems	15	8.8	5.5	4
Electricals	9.5	5.6	3.5	2.5
Noise abatement	0.7	0.4	0.3	0.2
Catalysts	2.5	1.5	0.9	0.6
First charges	1.2	0.7	0.4	0.3
Spare parts	7.5	4.4	2.8	2
Total 2	*69.7*	*41*	*25.8*	*18.5*
Total 1 + 2		*100*		
Package units				
Erection (including material for insulation and painting)		43	27	19.5
Civil work		16	10	7.2
Total 3		*59*	*37*	*26.7*
Total 1 + 2 + 3			*100*	
Other costs				
Packing			3	2.2
Transport			2	1.4
Insurances			1.5	1.1
Planning costs			14	10.1
Supervision of construction			5	3.6
Commissioning			4	2.9
Allowance for inflation			4	2.9
Unforeseen			5	3.6
Total 4			*38.5*	*27.8*
Total 1 to 4 = project costs				*100*

calculated ex factory without packaging. These items are normally specified individually in engineering design allowing a rapid, detailed, preliminary calculation to be made.

The *bulk materials group* covers items such as pipes, control systems, electrical equipment and materials, insulation, and paint. Exact determination of these costs is expensive and time consuming and can be done only on the basis of detailed material takeoffs (MTOs). This group also takes in civil work, structural steel and erection work.

Other costs include engineering, procurement, construction supervision, commissioning, travel costs; price inflation during the construction period; insurance, duties, and contingencies. These items can be only calculated for the project as a whole and cannot be determined until the equipment and bulk materials costs are known.

3.5.1. Equipment

Because most pieces of equipment are "tailor-made," manufacturers' price lists are not often available for cost calculations. The best way to get accurate prices is to submit enquiries to the manufacturers. This can be done on the basis of the design specifications, but applicable standards (e.g., DIN or ASA), design specifications, and acceptance conditions (e.g., AD-Merkblatt documents, ASME Code, TÜV, Lloyd's) must be indicated. Such inquiries are, however, time-consuming and make a great deal of work for the vendors; they are accordingly unwelcome. Therefore inquiries are only made with complicated equipment and machinery. Simpler items are estimated on the basis of in-house documents.

An engineering firm normally has a data base containing important data (including weights

and prices) on all equipment and machinery covered by orders and inquiries in recent years. The data base must be kept up to date, this requires close cooperation with the purchasing department. Armed with such a system, the firm can make fairly accurate estimates, for example, of the cost of heat exchangers per square meter or per kilogram, given the size, type, pressure, and material of construction. The same holds for towers, reactors, and vessels; here also the size, pressure, temperature, and material must be known. The costs of simple machinery can be determined similarly. The use of computer programs makes it an easy matter to calculate the weights of equipment items.

Before such prices are incorporated into the estimate, correction factors must be determined so that current prices can be obtained. There are two reasons for such a correction:

1) The figures stored in the data base cannot be completely up to date, and there may be a lag of half a year or more before the order is issued
2) Purchase prices depend not only on market trends but also on the economic situation of the manufacturer

The use of such a data base system necessitates the availability of appropriate cost-index figures for price adjustment. More information on cost-index figures is given in Section 2.2.1.4.

3.5.2. Bulk Materials

Provided the engineering documentation (layout, piping and instrumentation diagrams etc.) is exact enough (see Fig. 7, Section 3.1), bulk material costs can be calculated fairly accurately by applying unit prices (per piece, m, m^2, m^3, kg, etc.) to material takeoffs. Installation costs are also derived from takeoffs classified according to functional disciplines (unit prices for construction activities plus costs of site preparation, erection equipment, materials, etc.). If unit prices from an in-house data base system are used for bulk materials, they must be adjusted in the same way as equipment prices (see Section 3.5.1).

Another option is to obtain detailed current unit prices from the manufacturers or construction contractors. This method yields accurate results, but is time consuming and involves high costs. It should therefore be employed only when a highly accurate estimate is needed.

The engineering documents available in the design stage are usually incomplete. The trend in bulk materials estimation is therefore toward applying multiplication factors to the estimated total equipment costs. The latter can be obtained only, however, by analyzing projects already realized. In order to achieve satisfactory accuracy, separate multiplication factors must be derived for each plant type and each process. The larger the number of facilities evaluated, the more accurate are the multiplication factors. The percentages given in Table 4 represent multiplication factors for the individual cost items of a typical chemical plant; normally, material and construction costs for a bulk item are determined separately. Other multiplication factors are published in [7].

If the required accuracy is $\leq \pm 10$ %, however, these two estimation approaches are combined.

Civil Work and Structural Steel. For structures such as compressor houses, control rooms, laboratories, and workshops the enclosed volume is usually determined and the costs estimated by using unit costs. In the case of adverse soil conditions or heavy structures, it may be necessary to perform a preliminary static calculation so that the foundation size and/or the number of piles needed can be determined more accurately. The other civil-work items are often handled by means of multiplication factors.

The same holds for structural steel, although here the costs are found in terms of the weight and the price per tonne for heavy, medium, and light steel construction.

Control Systems. The trend toward plant automation and increasingly strict safety standards have increased the contribution of control systems to plant costs. The degree of sophistication of the control system of a chemical plant depends strongly on the future operator. Accordingly, it is often impossible to work with general multiplication factors. Costs are therefore calculated from the number of control circuits, required computer capacity, number of displays and control panels, etc. Multiplication factors are used only for installation materials.

Electricals. Since the number and sizes of electric motors are normally known from the detailed equipment estimate, it is relatively simple to determine cost-intensive electrical items such as motors, transformers, switchgear, and cables. Costs for other items, such as grounding, lightning protection, lighting, and installation materials are handled with multiplication factors.

Erection. The determination of equipment installation costs on the basis of unit-weight formulas gives adequate accuracy; it is also simple because the cost of major equipment (other than machinery) is usually based on weight. A similar method can be used for machinery, but the calculation process is often divided into equipment setting (unit-weight formulas) and equipment alignment (hourly basis).

Installation costs for bulk material items (e.g., control systems, electrical, insulation, and paint) and for catalysts and first charges are usually determined by applying multiplication factors to the material costs.

The civil work discussed above normally includes installation costs. Erection costs for structural steel are established on the basis of unit-weight installation formulas and subdivided into heavy, medium, and light steelwork.

Serious difficulties arise in the accurate estimation of piping installation because the exact material takeoff and the exact routes of pipe runs are seldom known. The labor cost of installation depends on the number of welds and flange connections, pipe diameters, wall thicknesses, materials of construction, and installation heights. Estimation is therefore done with unit-weight prices tabulated for each nominal diameter or nominal-diameter group. A separate calculation is needed for each material of construction.

Installation costs can also be estimated from the number of welds per meter of straight pipe and the nominal diameter but this requires detailed knowledge of the piping design.

Vendors normally offer package units, such as cooling towers and refrigeration units, including installation; the entire scope of supply and installation costs given in the bid specifications can therefore be incorporated in the calculation as a lump sum.

The calculation of the construction of a plant must also cover the costs of site preparation, utilities, construction equipment, transportation equipment, scaffolds, site roads, fencing, security, first aid, and other items.

3.5.3. Other Costs

Bulk material costs are commonly determined for each section of the plant, whereas "other" costs are calculated for the project as a whole. Multiplication factors are employed for items such as packaging, shipping, insurance, price escalation, and contingencies.

Before the contingency factor is established, it is necessary to make certain that the calculations for the various functional disciplines do not contain any multipliers of this type, since otherwise contingency amounts grow out of control. Engineering, procurement, construction supervision, and commissioning costs are, however, usually determined individually.

Engineering and Procurement Costs. In addition to the approximate method of obtaining engineering and procurement costs by applying multiplication factors to plant costs, two more accurate methods are available: calculation of engineering hours and procurement hours on the basis of equipment items or the documents to be prepared for each functional discipline. The latter is more accurate.

Itemized Equipment Approach. The number of hours per equipment item can be referred to the work done on each item; all equipment items are then added together with multiplication factors for other functional disciplines. Alternatively, a total number of hours per equipment item can be determined for all functional disciplines. This total number of hours varies widely, depending on the processes employed, the scope of authority engineering, and existing design documents. It generally lies between 600 and 1200 h per equipment item. This method can be used only if experience has been gathered with existing plants of the same type.

Calculation Based on Documents to be Prepared. The estimation of man-hours based on the number and size of drawings and documents presupposes a great deal of experience in project execution and the availability of document breakdowns from earlier projects (including empirical figures for the hours spent on all

other activities, such as procurement and expediting, dealing with vendors, and inspection of equipment). This procedure also calls for an exact definition of the individual services (service catalog) so that the various functional disciplines can be clearly isolated from one another. In practice, a blend of the two methods of calculation is often used.

After the man-hour requirements for engineering and procurement of a project have been determined, they are subdivided into categories and multiplied by appropriate hourly rates to determine the costs. Travel, communications, reproduction, computer costs, and the cost of the model must be added.

Construction supervision costs are often obtained by applying multiplication factors to the equipment and bulk materials groups or to the engineering hours. A better, slightly more complex method is to determine the costs on the basis of construction and manpower schedules.

The manpower schedules include supervisory personnel for the functional disciplines used by the operator and the engineering contractor; also the installation specialists provided by the manufacturers of complex apparatus, machinery, and package units; and supervisory personnel, if any, assigned by the licenser.

In contrast to engineering costs, estimates are based not on hours but on person-days and daily rates. These rates normally include accomodation and food costs for personnel. The calculation must also cover overtime, travel expenses, and costs for the construction office and its operation.

Commissioning Costs. The costs of commissioning relate to the time from the end of erection to the start of production. They are also classified as investment and must therefore be incorporated into the estimate. Commissioning usually lasts one to three months, but up to six months for very large multistage plants. This cost group includes the following items:

Personnel Costs. Personnel costs are generally determined in the same way as installation supervision costs.

1) Operating personnel on a rotating shift basis
2) Laboratory personnel on a rotating shift basis
3) An on-call installation crew who remedy mechanical problems and/or reinforce the shift crew in the workshop
4) Specialists assigned by vendors for commissioning special equipment and machinery
5) Consulting or supervision by licenser's personnel
6) Possible increased deployment of process and specialist engineers from the operator's engineering department
7) Consultation or supervision by personnel belonging to the engineering contractor

Other Commissioning Costs. These costs have to be estimated case by case. For routine processes where most products are directly marketable, raw material, auxiliary, and fuel costs can be virtually neglected. In larger, more innovative plants running difficult processes, these costs may be very high. Further difficulties may arise if the market for a new product has to be created. Other commissioning costs are divided into

1) Training of company personnel (possibly in the licenser's facility)
2) Raw materials and auxiliaries as well as utilities needed to replace defective charges or product not up to specification
3) Travel expenses for personnel not belonging to the company

Indirect Costs. Other indirect costs can account for a large portion of investment costs. They are:

1) Licensing fees
2) Land costs
3) Land development costs
4) Fees for government inspections and approvals
5) Financing costs
6) Administrative costs and costs for possible expansion of sales organization to market the product
7) Public relations work to inform the public about the new plant
8) Working capital, such as raw materials stocks and finished product storage

3.6. Conclusion of Preliminary Design Phase

The last step in preliminary design is the preparation of a report containing the following information:

1) The updated feasibility study.
2) Profitability calculation with graphs of accumulated revenues and expenses, the break-even point, and profits (similar to Fig. 6, Section 2.3.1). If a clear decision in favor of one of the alternative investments is possible, the profitability calculation is limited to a single proposal. If, however, entrepreneurial questions figure in the analysis, the report should include the two best alternatives.
   Inaccuracies in the primary documents and the amounts added to allow for these should be established and specified in all calculations.
3) Schedule for financing requirements.
4) Schedule for personnel requirements.
5) Time schedule for the phases of execution up to the start of production.
6) The project manual containing all the studies and results from preliminary design (basic engineering documentation), including the estimate.

4. Contract Writing and Forms of Contracts

4.1. Licensing Agreements

Process development has become so expensive that development costs can seldom be covered by royalties. Aside from government-funded research, new processes are developed chiefly by production-oriented chemical companies or petrochemical concerns, which regard license fees as incidental income. In cases where engineering firms hold licensing rights, they either arise from license agreements with production companies or represent improvements to established processes.

Licensing agreements may relate to patent licenses (rights to use granted patents, Section 4.1.1), process licenses (which include patent rights and all know-how, Sections 4.1.2 and 4.1.3), and straight transfers of know-how for processes not protected by patent (Section 4.1.4).

4.1.1. Patent Licenses

Suppose chemical company A has developed a process that is under patent protection but is wholly or partly covered by the earlier protection of chemical company B. Company A will seek to join B by exchanging patent rights, a mutual grant of rights to use the patents, or the purchase of a patent license from B. Usually, company B asks for compensation in an amount that depends on the age, the importance of its rights, and the degree of overlap. This fee only covers the particular use in question. The licenser does not disclose experience above and beyond the published contents of the patents. If the fee asked is too high for A or there are doubts about the validity of B's patents, then in the absence of an agreement A can either seek to have the courts declare B's patents invalid or can disregard them and risk legal action by B.

4.1.2. Process Licenses

As a further example, B has plant-scale experience while A only has experience with a pilot plant for the same or a similar process. Even if the new process is clearly an improvement, it is expedient for A to purchase B's experience because inevitable setbacks and lost time on going from the pilot plant to a first production-scale facility ("scaleup risk") usually cost more than a process license.

In acquiring a process license, company A will commonly be confronted with one or more of the following possiblities:

Case 1. To date, company B has only built its own full-scale plant. It has not yet granted any licenses to third parties and therefore has no design documentation on plants with different feedstocks and utilities, as needed by A for its plant. B is not prepared to guarantee third-party plants, but only to offer a patent and know-how contract. Once such a contract has been signed, it permits A to examine the operating records and assemble the documentation needed for the projected plant.

Case 2. Company B is interested in worldwide licensing of the process, maintains a large in-house engineering department, and has worked out the complete process design. In addition to the license and know-how contract, B will offer guarantees, and may even be ready to undertake engineering, procurement, construction supervision, and commissioning with separate billing for these services.

A variant of this case is when B has developed the process in collaboration with an engineering contractor and granted exclusive use or construction rights to that company. Here again, licensee A cannot seek competing bids because the process is a "monopoly."

Case 3. As in case 2, B owns the process design and is prepared to give guarantees in the context of a license and know-how contract. B leaves it to licensee A, however, to select an engineering contractor to construct the plant. Company A requests competitive bids from several engineering firms, each of which must enter into a confidentiality agreement regarding the working documents delivered to it. Licenser B will opt for this procedure when its patents have given it the lead on the market for a long enough time and disclosure to several engineering contractors will lead to the wider use of the process and generate royalties.

Case 4. Since patent protection for processes declines in value with age, it is in B's interest to secure its know-how to the maximum extent. Accordingly, B will hand over the acquisition and licensing rights to one or two trusted engineering firms in the form of a license agreement (which may be exclusive or nonexclusive, for stated countries or for the world). Licensee A can then acquire process rights, engineering, and plants only by contracting with one of these engineering companies.

Process licensers usually give limited guarantees on the functioning of the process and product quality, but their liability commonly extends only to half of the license fee. Commissioning assistance is generally given – in special cases, after training of the operating personnel in B's own plant or another licensee's plant. In developing countries, licenses are only granted with long-term management contracts, which many banks require as a condition of financing.

For profitable processes, royalties amount to 1 – 4 % of the product value for the life of the patent, but for at least 10 years. A *running royalty* is a license fee levied on a specific measurable quantity of production from a specified technology. A *paid-up royalty* generally represents the current value of 10 years' royalties and is thus equal to 5 – 6 times an annual royalty. If supply contracts are in effect, the royalty may also be calculated in terms of the plant value (e.g., 5 % of the value of the entire facility or 10 % of the value of the process facility proper) or as a fixed amount per unit product, often with price escalator clauses.

4.1.3. Process Licenses via Engineering Contractors

Cooperation with engineering contractors often increases the licenser's chances of obtaining additional revenue by licensing its entire operating know-how. The licenser simultaneously reduces the load on in-house engineering capacity because the external engineering firm bears the main load of establishing engineering documentation, interpretation, and training. Licensing is therefore being increasingly handled by engineering firms who check and extend the process documentation and may also optimize some equipment items or process steps. Improvements are possible when new operational know-how has not yet been implemented as redesigns in the licenser's first full-scale plant; this is difficult with good processes because of market requirements (supply contracts for products). Engineering firms are able to undertake all design, supply, and plant construction tasks, including commissioning. Competitors offering the same process will design different plants because of their different design and construction experience. In comparing bids, the customer must therefore consider not only the price, but also the references, guarantees, and quality of the bidders.

Production companies granting licenses only tend to guarantee the typical basic features of the process (yield and product quality). They never take responsibility for plants erected by third parties. Engineering firms thus see an opportunity to guarantee the whole plant. This is, of course, only possible when the company has built many plants and has wide experience and

knowledge. The licenser generally remunerates the engineering firm for this aid by paying it part of the royalty in return for the additional responsibility.

If the engineering firm has no special experience, it will supplement the licenser's process guarantee with a guarantee of correct engineering design; in the case of supply contracts, it will relieve the customer of some risk by giving a material and price guarantee. Depending on experience, the firm may also guarantee the capacity of the plant and the utilities consumption within stated tolerances. When processes are licensed from the licenser either directly or via an engineering firm, it is important for the licenser to ensure that "secret know-how" remains secret and that the exchange of experience, including supplementary discoveries of the licensee (feedback), remains confidential. Naturally, the licenser will compensate the licensee for economically valuable improvements.

4.1.4. Know-How Contracts via Engineering Contractors

The know-how contract is a modification of the process license agreement. When patent protection ceases, the process licenser attaches great importance to keeping its undisclosed know-how secret. This know-how is given to a few engineering firms under strict confidentiality agreements, which are binding on the firms' employees. Companies interested in the construction of plants may be told solely what has already been published. They can only obtain more precise information after having signed a secrecy agreement.

The texts of know-how contracts are comparable to those of license agreements, but the fee is lower. The activity of engineering firms and any guarantees they might give are similar to those discussed under license agreements.

Some engineering companies have often managed to improve current processes that are no longer protected so that new applications result. In most cases, however, the activity of such firms is limited to simpler processes (e.g., gasification and gas purification), unit operations, or improvements in thermal economy, automation, environmental protection, and disposal technology.

4.2. Design and Supply Contracts with Engineering Contractors

Should an investor bring in an engineering firm to construct its plants? This question arises whenever the investor does not have adequate in-house engineering capacity. In the following sections it is assumed that the investor needs an outside engineering contractor for a construction project.

4.2.1. Selection of Engineering Contractors

4.2.1.1. Importance of Risk in the Plant Business

The intention to build a chemical plant presents the investor and the candidate engineering firm not only opportunities for success but also risks. The erection of a large plant entails above-average, qualitative and quantitative risks. The following features of the large-plant business are crucial in risk assessment:

1) Complexity and duration of the project
2) Responsibility for long-term operation of the plant
3) Transfer of investor's entrepreneurial risks to the engineering firm and vice versa
4) Priority of the project in investor's and engineering firm's business results
5) High financing budget
6) Heavy dependence on market development of the product

A unique feature of the plant design and construction business is that these risks are generally cumulative, so that both the investor and the engineering contractor must practice risk management, i.e., measures aimed at identifying, assessing, and limiting risks.

Since the investor and the engineering contractor deal with potential risks in different ways, the formulation of a plant construction contract between these parties is extremely important: both parties must be satisfied that their interests are protected.

Generally, there is no benefit to the investor in looking too hard for every competitive advantage. This is especially so when the engineering

contractor may fall into the danger of trading off plant quality for savings as a result of accepting below-cost prices and unsatisfactory conditions; such shortcuts often become apparent only after the plant is on-stream. As a rule, sharp competition forces engineering firms to seek an optimal outcome at minimum cost. These considerations have promoted cooperation between investor and engineering contractor and have become crucial points in contract drafting.

4.2.1.2. Selection and Award Criteria

The investor develops his own criteria for selecting an engineering firm to carry out a construction project. Bid invitations and bid comparisons are used to award a contract that is acceptable to both parties. The most important criteria follow:

1) Does the engineering firm have experience or references relevant to the tasks it may have to perform?
2) Does the firm have experience in the country and locality where the plant is to be built (logistics)?
3) Is there a danger of communications problems during execution of the project (language barrier)?
4) What kind of staff situation exists within the engineering firm? Will it be able to organize good project management?
5) What standards and specifications will govern the work? Can the engineering firm ensure compliance with standards that may be new and unfamiliar to it?
6) What references does the firm offer with regard to the quality of supply and services and compliance with time schedules?
7) Is the firm's credit rating adequate for the requirements of a large contract?

In the bid invitation, the investor states what is wanted and required with regard to the technical concept (process technology, scope of delivery and services, quality requirements, etc.) and the commercial concept (type and contents of contract, financing, marketing, etc.).

A thorough, detailed bid represents a significant cost to the bidder. A company will only make this investment if it has a good chance of winning the competition. The bid invitation must therefore be written in a clear, understandable manner so that the potential bidder can precisely identify the task to be performed.

In order to make bids comparable, all bidders must be given the same documents and information. The bid invitation should always include the following:

1) General description of project with site layout plan.
2) Services and deliveries desired (e.g., basic engineering and know-how, complete engineering, complete engineering plus procurement assistance, delivery with or without construction and commissioning).
3) Precise data on the process or process stages to be covered by the bid. If the bidder does not have rights to the process, the invitation must state whether the process license is also to be covered or whether this will be handled by the investor directly with the process licenser.
4) Production capacity and product quality.
5) Available feedstocks, auxiliaries, and utilities (composition and quantity available).
6) Availability of electric power and water for construction purposes.
7) Available infrastructure (e.g., workshops, laboratories, warehouses, and social facilities).
8) Overall time schedule.
9) Guarantees desired for production, quality, consumption rates, materials, date of completion.
10) Pricing terms of payment desired.
11) Standard codes and guidelines to be applied.
12) Deadline for bid submission.
13) Period for which bid must remain binding.
14) General conditions.

A qualified engineering firm may still decide, not to submit a bid – possibly because it sees too little chance of realization of the project or a chance of success in the face of stiff competition.

The simplest approach to bid comparison is to prepare schedules of prices, delivery times, terms of payment, terms of delivery, exclusions, and miscellaneous conditions. Large investor companies send preformulated tables to bidders, especially for standard processes; the bidder merely has to enter the services, exclusions, and miscellaneous conditions.

The differences found in the bid comparison should be analyzed and then discussed with the bidder. Discussions are often time consuming and should be carefully planned. An efficient negotiating program should cover the following topics, in order:

1) Fundamentals, infrastructure, overall concept, process engineering, economics
2) Equipment, material, and services to be provided; exclusions
3) Special and general conditions; guarantees
4) Price
5) Time schedule, including intermediate targets

4.2.2. Form and Content of Contracts

4.2.2.1. Basic Concerns in Contract Writing

In writing the contract, both the party inviting bids and the bidder can propose, negotiate, and conclude agreements. The elements of the contract must therefore cover all those aspects that the potential contracting parties deem not satisfactorily governed by other provisions (usually by relevant legislation).

Efforts to reach contractual agreement aside from relevant legislation are primarily devoted to making the potential risks manageable. Details on contract drafting can be found in [37], [38], [113–115].

Since the profit margins of the engineering contracting business are small, it is important to identify, assess, and allot all risks. For a multimillion-dollar project differences of a few tenths of a percent can decide whether a bid is accepted or rejected.

Often the only other possibility for absorbing risks is the contractor's profit, which averages ca. 2 – 4 % of contract value in the plant design and construction business. Such a small margin can only cover limited cost overruns on a single project.

Methods of dealing with potential risks at the contract drafting stage can be classified into two groups. First, each of the parties must consider the well-established policies of its company. The extent of potential risk for the company resulting from deviation from such a policy governs which person in the company hierarchy has the power of decision. The person in charge of the negotiations must therefore know when the limits of his jurisdiction have been reached. The decision whether to accept or reject conditions proposed by the other party must then be passed on to an appropriately higher level in the company hierarchy.

Second, there are modifiable (dispositive) goals for the negotiating parties. The limits of negotiability (breaking points) are reached when one or the other party's willingness to make compromises has been exhausted or when acceptance of the other party's proposals would result in incalculable risks.

As a rule, the engineering contractor has to decline to accept liability for indirect and consequential damages. Furthermore, the contractor cannot be held responsible if postponements of deliveries and services occur as a result of circumstances beyond his control (e.g., war, revolution, strikes). Other possible breaking points include excessive guarantee requirements.

The engineering contractor will reject a proposal for unlimited liability. Limitations on liability include:

1) Limitation of total liability for delays and/or nonattainment of process guarantees and/or for costs incurred for the correction of design errors in equipment or during construction and installation.
2) Time limits on the guarantee for materials, typically x months after commissioning of the plant; if commissioning is delayed and this is not the fault of the engineering contractor, a "not later than" date must be agreed on.
3) The liability provisions set forth in the contract should bar recourse to liability laws because these do not make appropriate allowance for the special features of the plant construction business.
4) No liability for indirect or consequential damages, such as lost profits, loss of use, and loss of production. From the engineering contractor's standpoint, this is a nonnegotiable clause. If such liability were accepted, the existence of even financially sound contracting companies might be endangered.

4.2.2.2. Contract Types and Provisions

Although there are many types of contracts, the following are particularly important:

Engineering Contract. An engineering contract normally governs compensation for engineering and procurement services, and the supervision of construction, and commissioning. It usually provides for reimbursement of incidental costs such as travel, communications, computer support, and accomodation.

The contract can be for complete engineering, including procurement and supervision at the construction site up to the successful completion of the guarantee run.

The contract may, however, be limited to partial services, such as basic engineering, detailed engineering, procurement services, supervision of construction, supervision of commissioning, and supervision of third-party engineering services.

An engineering contract does not include delivery of plant equipment or the performance of plant construction. The procurement services that may be included are generally carried out "in the name and on behalf of" the investor.

The contractor's liability in an engineering contract extends to correcting design errors at no cost to the investor, and often includes a percentage of costs incurred as a result of correcting problems in equipment or during construction.

Reimbursable Contract. In the reimbursable contract, every hour of work performed by the engineers and procurement staff is paid for, plus all incidental costs. Proof of working hours is required.

A reimbursable contract offers an advantage when the scope of work is not well-defined (so that it is impossible to determine a fixed price) or when abandonment of the project is anticipated.

The hourly or daily rates specified in the contract are often classified by category. The contract must state what is included in the hourly rates (e.g., salary, other payroll costs, workplace costs, profit).

An important drawback of this type of contract is that the investor can strongly influence the contractor's execution of the project. Because the scope of engineering services is generally not adequately defined, control of the service budget is difficult.

Reimbursable Contract with Target Price. In this type of contract, the engineering contractor accepts a share of the risk for staying within the calculated service costs. A target price is set on the basis of the calculated costs; the engineering contractor generally takes a certain percent participation in any overrun, but receives a bonus if the work comes under the target price.

The target-price provision can also extend to equipment and construction costs. A sufficiently exact definition of the services is required so that a fairly reliable calculation is possible. The investor gains the advantage of better budget control.

Lump-Sum Engineering Contract. If the bid invitation for a chemical plant defines the scope of work well enough, a lump-sum bid can be prepared. This type of contract has the advantage for the investor that the budget is set in advance (provided no additional work is needed in the course of the project). The engineering contractor has much better control over the project with regard to costs (man hours) and schedules. In contrast to the reimbursable contract, the investor has a limited say in decisions taken during the project.

Supply Contract, Turnkey Contract. As a rule, a *supply contract* commits the engineering contractor to provide engineering, and supply of all plant equipment and materials needed for the construction of a complete chemical plant. The contract must define the limit of responsibility of the engineering contractor for plant equipment; the boundary may be the ship, railroad car, or truck used to transport the equipment to the construction site. The engineering contractor's responsibility can, however, extend to the receipt of equipment on site.

A *turnkey contract* includes the provisions of the supply contract, plus construction services. If the engineering contractor does not have the personnel resources for construction work, it generally subcontracts this work to specialist firms but assigns its own management and supervisory personnel to maintain single-line responsibility to the investor.

In supply and turnkey contracts, all goods and services are procured in the name and on behalf of the engineering contractor. The contractor thus assumes liability to the investor for the mechanical functioning of equipment as provided by the contract. In turn, the liability

risks associated with this material guarantee are shifted to the relevant suppliers and subcontractors. Similarly, the engineering contractor will impose penalties on the suppliers and subcontractors for late delivery or installation to reduce its own time risk.

Supply and turnkey contracts are based on a lump-sum price, sometimes with a provision for price escalation. In order to keep the price risk quantifiable, the bid must be based on a detailed, accurate definition of the scope of deliveries and services for the plant to be erected.

4.2.2.3. Essential Elements of a Contract

The structure and delimitation of scopes of delivery and work are very important, especially because these factors affect liability questions, contractual provisions, and the negotiation process. They are expressed in the form of

1) Specifications of services
2) Description of the scope of delivery with delivery boundaries
3) Guarantees and liabilities
4) Criteria for handing the plant over to the investor

Specification of Services. Although the specification of technical and commercial services is not usually problematic from the standpoint of contract law, it is often highly relevant as a basic fact and for later reference. Since this specification normally comprises a very large set of documents, it should be formulated as one or more appendices to the contract. Much unnecessary discussion between the contracting parties can be avoided if the specification of services is carefully prepared in detail and states clearly what services must be performed by whom.

It is useful to prepare a list of services for the project, which can be based on the specification services contained in the bid invitation, as well as in the bid. Any modifications made in the course of contract negotiations must be incorporated.

The specification of services must also contain the detailed conditions and constraints needed to carry out the assignment, a procedural description, and a list of relevant regulations and standards.

The list of services should give a detailed description of technical and commercial services to be performed, indicating what design documentation is to be prepared (e.g., drawings, data sheets, specifications, flow sheets and schematics, lists, approval documents).

Moreover, it must be stated which of the services will be performed by the engineering contractor, the investor and third parties.

Scope of Delivery/Delivery Boundaries. A supply contract must describe the equipment and materials to be supplied; a turnkey contract must also define the construction services. In addition, the scopes of delivery and work must be delimited with respect to third parties.

The description of the scopes of delivery and work should contain the most complete list possible of apparatus and machinery; a description of electrical equipment and materials and control systems; approximate quantity requirements for piping, concrete, cables, and structural steel; and a description of construction services. It is not, however, expedient to set down exact numbers because final data on equipment size, pipe routing, etc. cannot be made until the design work is under way. Establishing the details too early could hinder subsequent optimization of the design, this would not be in the investor's interest.

Guarantees. Forms of guarantee for chemical plants have been devised that are technically meaningful and can be verified at a reasonable cost.

Mechanical (material) guarantees relate solely to individual pieces of equipment, but not parts subject to wear. A period (e.g., 12 or 18 months) is defined in which repair or replacement is undertaken by the contracting party or suppliers at no cost to the investor.

Process guarantees generally relate to the capacity of the plant, product quality, and the consumption figures for utilities such as steam, cooling water, and electric power.

The contract should state the conditions used to verify the process guarantee (e.g., analysis of feedstock, quality of utilities). Furthermore, it should define the measurement and analysis methods used in the guarantee run and its duration.

The forms of guarantees are different for utilities (steam, electric power, fuel gas), where a

certain amount of over- and underconsumption can offset each other, and for plant capacity and product quality, where shortfalls result in penalties.

If the plant does not meet the agreed figures in the guarantee run, the engineering contractor is given an opportunity to make improvements. If these are not successful either, the relevant contractual consequences take effect.

Handing Over the Plant. The criteria for handing the plant over to the investor are usually stated in the contract. The plant is often handed over section by section and/or in phases. Several forms of handover are employed:

1) *Transfer to the operational custody of the investor :* This, as a rule, takes place at the beginning of test operation. The investor's personnel run the plant under the direction of the engineering contractor.
2) *Provisional handover*: Provisional handover takes place after a report on the successful guarantee run has been signed, or after a "not later than" date has passed (if the guarantee run has not been done for reasons that the investor has to justify).
3) *Final handover*: Final handover comes after all defects listed in the report made after the provisional handover have been corrected.

Final inspection of the plant, performed by both parties, completes the handover procedure.

5. Execution of the Project

Design and construction of chemical plants are complex operations comprising many interrelated activities (Fig. 16). Plant design and the procurement of equipment occupy workers from nearly a dozen special engineering and procurement disciplines.

Medium-sized to large projects [contract value $(50-250)\times10^6$ DM; $(30-170)\times10^6$ \$] have a design period of approximately 15 – 18 months and a total project duration (up to commissioning) of ca. 18 – 24 months. Some 100 – 150 engineers and procurement staff are involved at the peak of such a project in the design offices alone.

5.1. Scope of Work

Work on total chemical plant project comprises basic engineering, detail engineering, procurement, civil work and erection, and commissioning. (The "authority engineering" activity, in which documents are prepared for submission to the authorities having jurisdiction, is generally carried out during both basic and detail engineering.)

Depending on its engineering capabilities and resources the investor may execute part of the project (basic engineering and/or procurement, and/or construction). The remaining activities (or the whole project) are executed by an engineering firm under contract.

Basic engineering is based on process design (Section 3.3) which is performed during the preliminary design phase by the investor, by the licenser, or by an engineering firm in close cooperation with the investor. Process design is a component of basic engineering.

Basic engineering documentation includes the process flow diagrams, piping and instrumentation diagrams, plant layout, equipment list (broken down according to plant sections), utilities distribution scheme, process engineering data sheets, noise protection concept, electrical equipment list, summary of electrical consumers (one-line diagram), data sheets for control equipment and instrumentation, functional process control and instrumentation plan, description of the process control system, and soil report.

The content and level of detail in basic engineering documentation is such that detail engineering can be done (possibly by another engineering firm) without significant difficulties.

Detail Engineering. In detail engineering the engineering and procurement teams prepare detailed plans, drawings, specifications, calculations, and descriptions so that the following steps can be carried out:

1) Preparation of bid invitations for all plant equipment, material, and services such as civil work and erection of plant equipment
2) Selection of manufacturers, vendors, subcontractors; placing of orders
3) Execution of quality assurance operations at manufacturers' and vendors' workshops

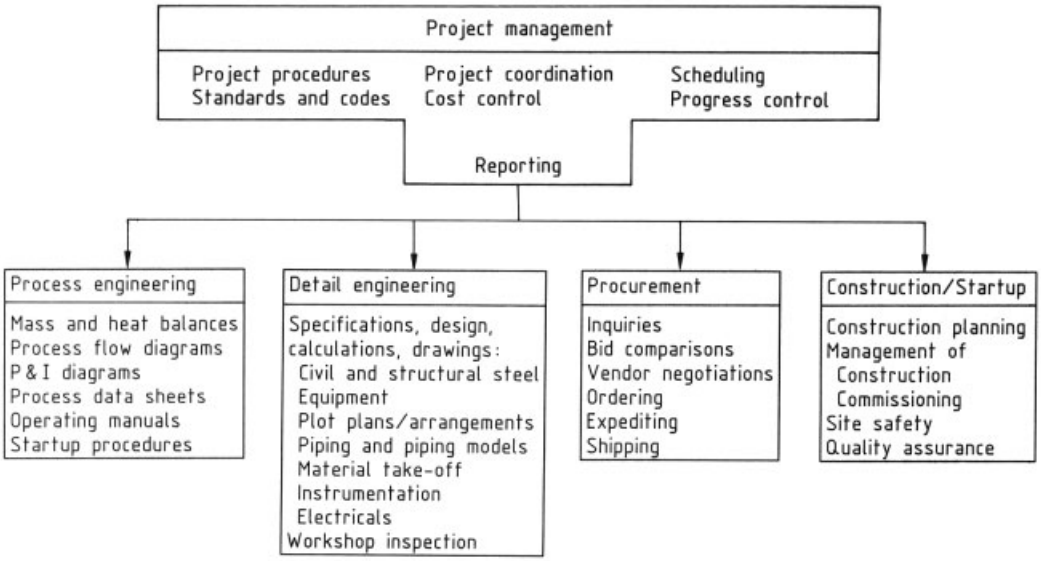

Figure 16. Main tasks of project execution

4) Shipping of plant equipment to the plant site
5) Execution of civil work and erection of plant equipment
6) Commissioning

Procurement services include all activities in connection with preparation of bid invitations, evaluation of bids and bid comparisons, and placing of orders. They also involve ensuring on-time performance of manufacturers and vendors (expediting), as well as the planning and supervision of the shipping of plant equipment to the plant site.

Civil work and erection of a facility are generally subcontracted to specialist firms by the engineering contractor or by the investor. The work is often supervised by the same organization that does the detail engineering. It must not be overlooked that the construction documentation prepared by the engineering contractor will not be free of errors, so that corrections have to be made on-site. Such corrections should be carried out by the engineering contractor's specialists.

The contract should state which partner is responsible for construction execution time and costs: the investor, the engineering contractor, or (as at large, complex plants) a general contractor accepting overall responsibility.

Commissioning of a plant comprises all the work done after "mechanical completion" up to certification of the guarantees embodied in the contract. The term "mechanical completion" should be defined unambiguously and in detail. Once "mechanical completion" has been certified, the plant passes into the custody of the builder. Preparatory work for commissioning and commissioning itself are usually performed by employees of the future operator under the direction of the engineering contractor or licenser. When the guarantee tests defined in the contract have been passed, the plant is handed over to the owner. Minor activities remain that are carried out by the engineering and/or construction contractor. This work is specified in a *punch list* and must be completed within a stated time.

5.2. Project Organization and Management

The design and construction of a chemical plant within a predefined time and budget calls for careful organization of the people working on the project, a clear definition of their responsibilities and competences, and an appropriate management concept.

A chemical-plant engineering firm usually executes many such projects, differing in size,

duration, and complexity, at any one time. The contracting firm is geared to this kind of assignment; it often develops its own set of "tools" and is organized exclusively for these activities.

The following description of the relatively complex steps in the execution of a chemical plant project is given from the viewpoint of an engineering contractor. It should not, however, be forgotten that investors and/or operators of chemical plants also follow a similar procedure to design and construct plants in their own areas of expertise.

5.2.1. Matrix Project Management [116], [117]

The first powerful management concepts for the execution of complex, single projects were devised in the United States during the last years of World War II. It was recognized that existing forms of organization were not suited to the solution of complicated armaments assignments on a crash basis. This management philosophy was subsequently employed in the space program, and later found use in the industries of Western Europe. It can be defined as follows: "Project management is a management concept for the solution of a well-defined problem within a predefined time subject to a cost framework specified for the project" [117].

Personnel from a wide range of disciplines within a company work on a defined project in an interdisciplinary setting and must be coordinated and directed. This means:

1) Assignment of appropriate personnel to the project from functional organizations in the company
2) Organization of the personnel into a project team for a defined period of time
3) Designation of a project manager
4) Establishment of responsibilities and competences for the duration of the project

The organization of the project team and, above all, the responsibility and competences of the project manager and his colleagues, have three basic forms: functional, matrix, and autonomous project management.

The matrix form of project management, either in pure or modified form, is the most suitable concept for engineering contractors because they perform "multiproject" management. Such an organization represents the best solution for integration of temporary project teams into the company organization. It also ensures that the projects are optimally executed.

Matrix project management is characterized by division of competences (Fig. 17). The project manager is responsible to company management for the realization of project targets. The managers of the functional departments (engineering and procurement) involved in the project are responsible for delegating appropriate persons to the project team. The managers also carry the technical responsibility for the performance of these team members. The project manager determines the "what" and "when" of a task in the project, while the department managers determine the "who" and "how."

For smaller, simpler projects and projects of a repetitive character, which do not keep a project manager busy full-time with management tasks, the matrix system can easily be modified so that project management is taken over by a qualified person from a functional department, who also carries out the required specialist work.

5.2.2. The Project Manager

The execution of a project in the matrix organization concept places heavy demands on the project manager [117], [118]. He assumes "project responsibility" to company management for the assigned project: compliance with budgeted project costs, agreed time schedules, and all contractual conditions, particularly the attainment of the agreed quality level.

For the customer, the project manager is the first representative of the engineering firm for the project, and also his first partner in discussions. The management tasks of the project manager are to set targets, establish guidelines, promote information flow, supervise, analyze and correct, motivate, and report. He must accord the same weight to the interests of the customer as he does to his own company.

The technical capabilities of the project manager should include not only specialized knowledge of the project and the technology, but also knowledge of how design and construction work is executed and ability in engineering, procurement, contract drafting, and project control.

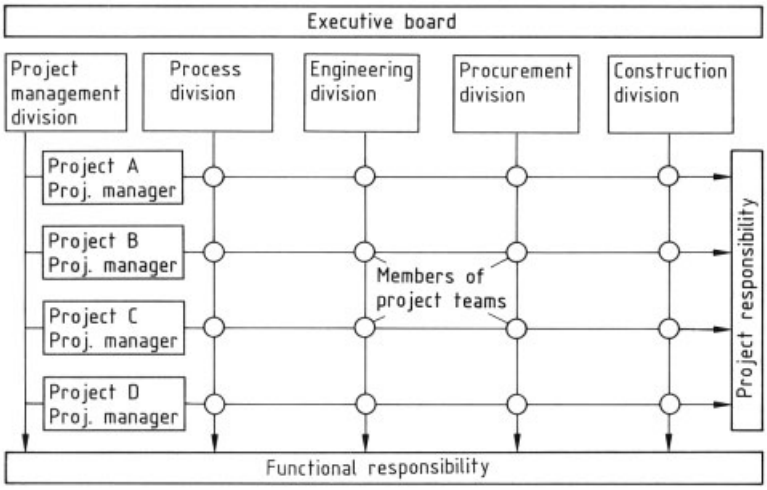

Figure 17. Matrix project management

These capabilities should be complemented by familiarity with the client's language, country, and the client, as well as skill in giving presentations and in moderating discussions. Other qualities include analytical ability, leadership and negotiation skills, ability to work under physical and psychological stress, ability to make decisions, and at the same time ability to integrate and to make compromises.

In larger projects, project engineers are put under an experienced project manager; they take responsibility for the execution of portions of the project.

5.2.3. The Project Team

A project team comprises employees from technical and procurement departments. If they are not fully occupied by a project (e.g., as it winds down), they may also belong to a team working on another project at the same time. The organization of a project team (Fig. 18) must be adapted to the specific needs of the project. Tasks, responsibilities, and competences are unambiguously defined. The project manager bears responsibility for meeting targets until the project is handed over to the owner, he is therefore placed above the construction site manager, who in turn heads the construction site organization.

Efficient coordination, with short lines of communication within the project team, is crucial. An effort is therefore made to separate the project team members physically from their functional departments within the company and locate them close together.

The team members remain subordinate to the project manager in regard to project requirements and for the duration of their participation in project execution. They continue to make their technical and functional "residence" in their home departments.

5.2.4. The Start Phase of a Project

The most important phase in the execution of a project is the start phase [119], [120]. The first 1 – 3 months are crucial to the success of the project. In this period all basic decisions must be made; all project-specific documents needed by the team members and the project manager must be generated.

The essential decisions and documentation include:

1) Selection of project manager and key personnel
2) Creation of project team and its organization
3) Working out project structure, including definition of subprojects and work packages
4) Planning of project execution and preparation of project execution guidelines

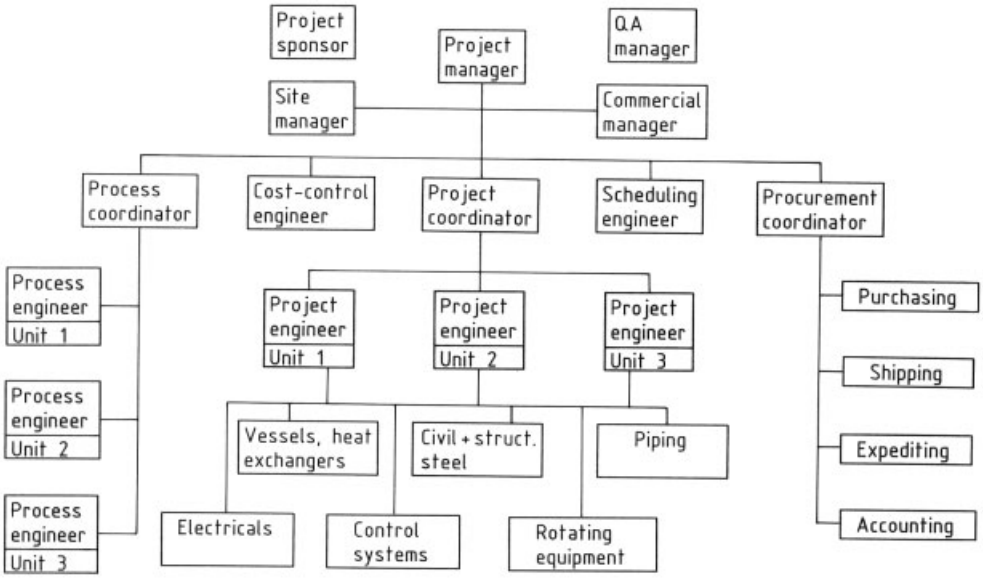

Figure 18. Organization of a project team

5) Identification of codes and standards governing the project (usually included in the contract)
6) Preparation of detailed time schedules (bar charts and network diagrams) as a basis for time scheduling and control
7) Drawing up a detailed budget as a basis for cost control
8) Preparation of a detailed man-hour budget for each of the technical and procurement disciplines
9) Determination of nominal progress curves for all disciplines involved, as a basis for monitoring compliance work progress
10) Manpower planning (office and construction site)

5.3. Project Control (Schedules, Progress, Costs)

(Project control during the construction phase is described in Section 5.6.2.2). The control of project execution is one of the most important tasks of the project manager. As early as possible, the project control team (schedule and cost control engineers) provides information about the time situation, the progress of work, and the consumption of man-hours in relation to due dates and determines the cost situation for deliveries and services [121]. Regular analysis of outstanding activities, expected time consumption, and cost and man-hours up to the end of the project is used to update these values. A comparison of the actual/expected/target values shows the reliable status of the project, enabling the project manager to identify problems at an early stage. After appropriate analysis, he and his coworkers devise solution options, which then lead to decisions about further steps. This procedure generally allows corrective measures to be implemented so that the target status can be regained.

5.3.1. Time Scheduling

Time scheduling and control involve the use of bar charts or network diagrams. The choice is made at the start of execution and depends on the size, complexity, and criticality of the project, and the customer's requirements. Schedules are usually prepared and updated by computer (mainframe and/or PC).

Bar Charts. Bar charts are used for planning and control of small, simple projects employing established processes and involving little time risk. These charts show the planned elapsed time for each discipline and their important activities. In general, the following areas are included: process engineering, plant layout, civil

work and structural steel, mechanical equipment (furnaces, process equipment, machinery), piping, electrical, and control systems.

The schedules cover the full span of the contract period and are broken down into the following activities: engineering (specifications, drawings); procurement (bid invitations, bid comparisons, orders, delivery time, shipping); construction and commissioning.

Figure 19 shows an example. A "dual-bar" system is employed to check the status of individual activities with respect to the due dates at any time during execution. The upper bar shows the planned duration of the activity. The lower bar indicates, as of the cutoff date, the actual or expected start and end of the activity. The progress of the activity is indicated by filling in the lower bar.

If needed, bar charts for individual disciplines can be expanded to show more detail. This is done, for example, when an early activity plan is prepared for the first three months of project execution, or when a detailed chart is drawn up if the status of work in one area is critical.

Network Diagrams. The network technique is widely used in industrial operations research [116], [121–124]. According to DIN 69 900, the term "network diagram technique" includes all methods used to analyze, describe, plan, oversee, and control processes on the basis of graph theory; considered parameters include time, costs, and materials. The main application of network diagrams is in project planning and supervision. The project is broken down into steps (activities), which are placed in a network diagram according to their logical dependences (precedences).

Formal representation employs arrows and nodes (a node may be represented by a rectangle or a circle). Network diagram techniques differ in how the logical conditions "event" and "activity" are assigned to these fundamental elements.

Event-on-node plan: The network is formulated mainly in terms of events, which are represented by nodes. An event is the start or finish of an activity and has no duration.

Activity-on-node plan: The network is formulated mainly in terms of activities, which are represented by nodes.

Activity-on-arrow plan: The network is formulated mainly in terms of activities, which are represented by arrows.

The following network methods based on the above approaches were devised independently of one another.

The *Program Evaluation and Review Technique (PERT)* is an event-on-node method and was devised in 1958 by the U.S. Navy in collaboration with the Boots & Hamilton consulting firm and Lockheed. The impetus for its development was the program to design and construct nuclear submarines. PERT uses three time estimates for each activity: optimistic, probable, and pessimistic. The analysis is run with each estimate and the results are averaged.

The *Critical Path Method (CPM)* was also developed in the United States in 1957. It is an activity-on-arrow network method that was devised by Du Pont and Remington Rand. It was first used for planning maintenance and conversion work in the chemical industry.

The *Metra Potential Method (MPM)* was first used in the design of nuclear power plants and was developed in 1958 by the French consulting firm SEMA. It is an activity-on-node technique.

Several other network methods have since been worked out. The *Precedence Diagram Method (PDM)* was developed from MPM and allows greater flexibility in the way precedences are represented.

The advantages of network diagram planning can best be exploited with computers. By the end of the 1970s, these techniques had been optimized for mainframe systems. Since the beginning of the 1980s, a variety of personal computer programs have been written, and network diagram techniques have come into wide use. Mainframe and PC software are capable of outputting the results of network calculations in the form of due date lists, bar charts, and network plots.

To control progress effectively, the actual status of project execution must be determined at regular intervals, the results compared with the target data from the network chart, and the consequences ascertained by recalculating the network to the end of the project.

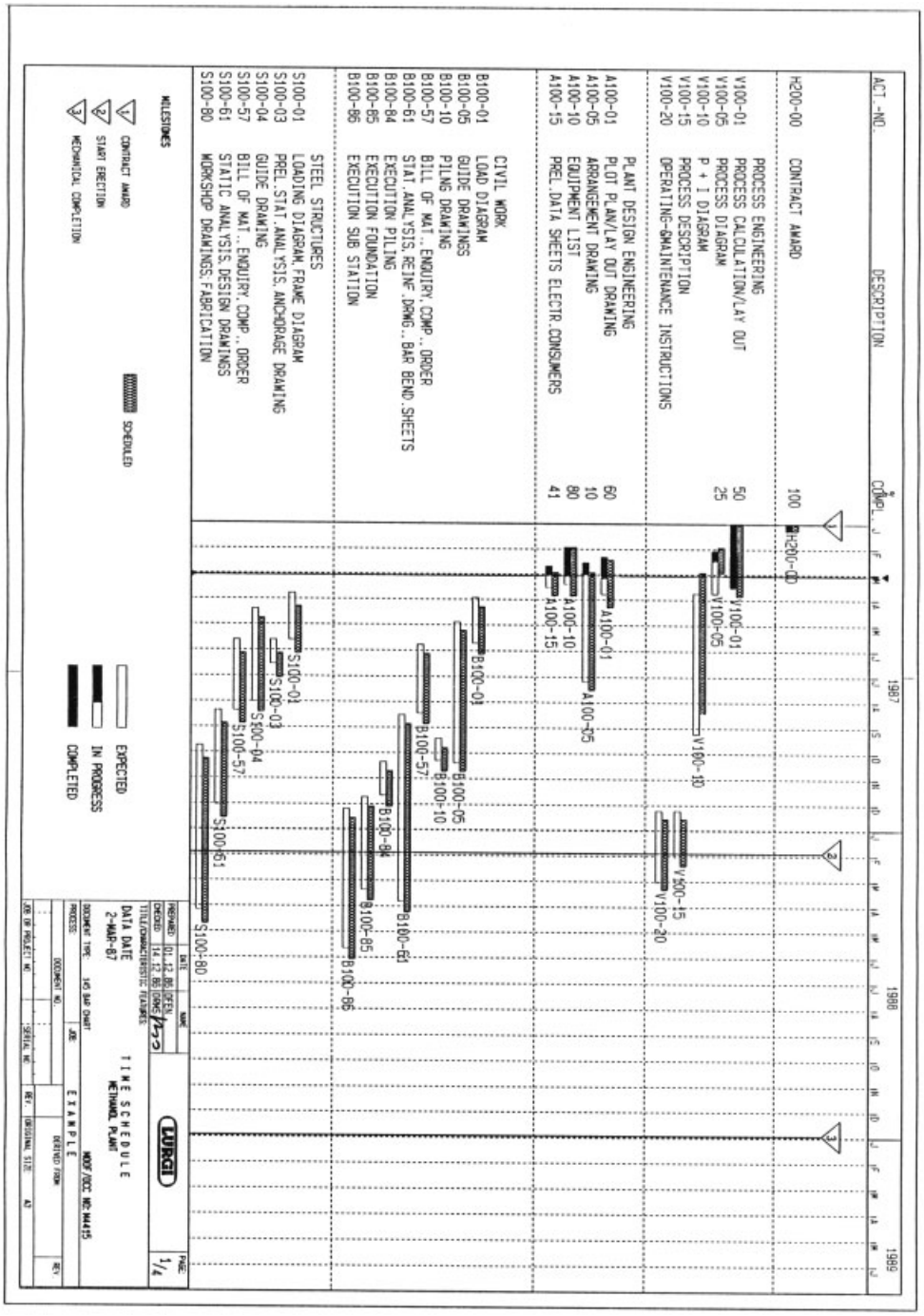

Figure 19. Bar chart

Event-oriented network systems play a secondary role in project planning and control. Activity-on-node and activity-on-arrow charts are about equally important in the planning and control of all kinds of projects, including research and development.

Some engineering firms have their own software for calculating network plans. Lurgi, for example, uses the PDM method to prepare network charts (unless the customer requires otherwise). The calculation is done with the Lurgi Netzplan System, which was developed in-house and is specially adapted to the conditions of chemical plant construction. Network diagrams are used for large, complex projects, chiefly when there is a great time risk. The project is broken down into individual activities of the various divisions involved, similar to the procedure for bar charts (see above). The level of detail is predefined by the project manager and scheduling engineer. The stored data can be quickly updated on screen (interactive operation) and the network diagram immediately recalculated.

When the network calculations are complete, the results can be output as lists of due dates sorted under all kinds of criteria, or bar charts with or without interdependence. Figure 20 shows the steps involved in project control with network planning.

5.3.2. Progress Planning and Control

Progress planning and control is performed for each discipline involved. For management needs, these individual plans are assembled into an overall plan.

The progress plan is generally plotted graphically. A "target" curve for the planned progress of work, for example of engineering, is drawn up in the following way. The work of each engineering discipline is broken down into individual activities. The duration and the expected monthly consumption of work units (e.g., man-hours) available from experience is specified for each activity. The sums of the monthly work units planned for individual activities are expressed as a fraction of the total number of work units required to give a target progress curve. These planning curves are prepared at the start of the project. During execution, they form the basis for control of actual work progress and actual consumption of work units (e.g., man-hours).

Actual work progress(physical progress)is determined by scheduling engineers who query the discipline engineers and evaluate the work performed. An example of such a work progress chart is shown in Figure 21. The positions of the curves of physical progress and work units consumed, relative to the target curve, give the project manager a clear indication of the time compliance situation and the efficiency of the personnel involved. If there are significant deviations from the target curve, the causes should be analyzed and appropriate corrective measures taken.

5.3.3. Cost Planning and Control

Costs. The objective of cost control is to provide correct information, as early as possible, on the cost status for equipment, materials, services, and indirect costs, and to identify reliable trends in cost development.

The project manager instructs the cost control engineers to analyze planned orders for equipment and services so that cost budgets are met. This "cost engineering" already involves the cost engineer at the stage of bid invitation, price comparison, and assignment of order stages.

Cost control is used throughout the entire project for engineering and procurement, as well as for construction. All costs connected with the project are considered, such as costs of equipment and bulk materials (mechanical, electrical, control systems, structural steel); spare parts; civil work and erection; third-party services; indirect costs (e.g., travel, computer services, communications); and costs for engineering, procurement, and construction supervision.

Figure 22 illustrates the operation of a computer-supported project cost control system.

Cost control is based on the calculation used to obtain the contract price. A detailed budget is prepared at the start of project execution. In the course of project execution, this budget will be modified if change orders are received from the client. The project team members responsible for each of the disciplines involved, in collaboration with the cost engineer, keep a continual watch on the technical and cost developments of the project in their respective areas of responsibility.

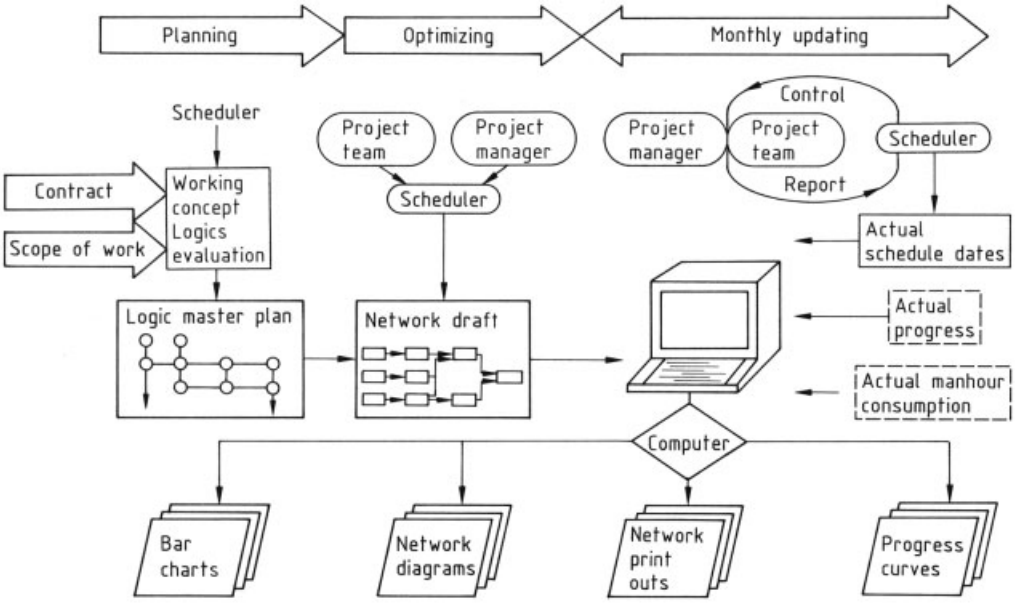

Figure 20. Project control with network system

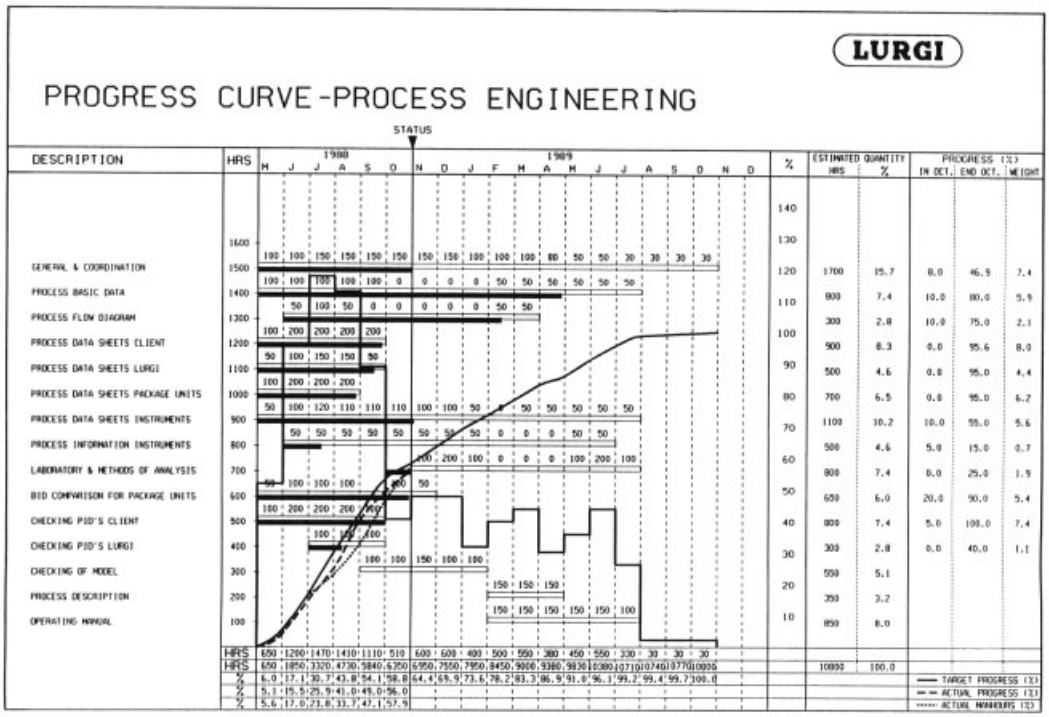

Figure 21. Engineering work progress

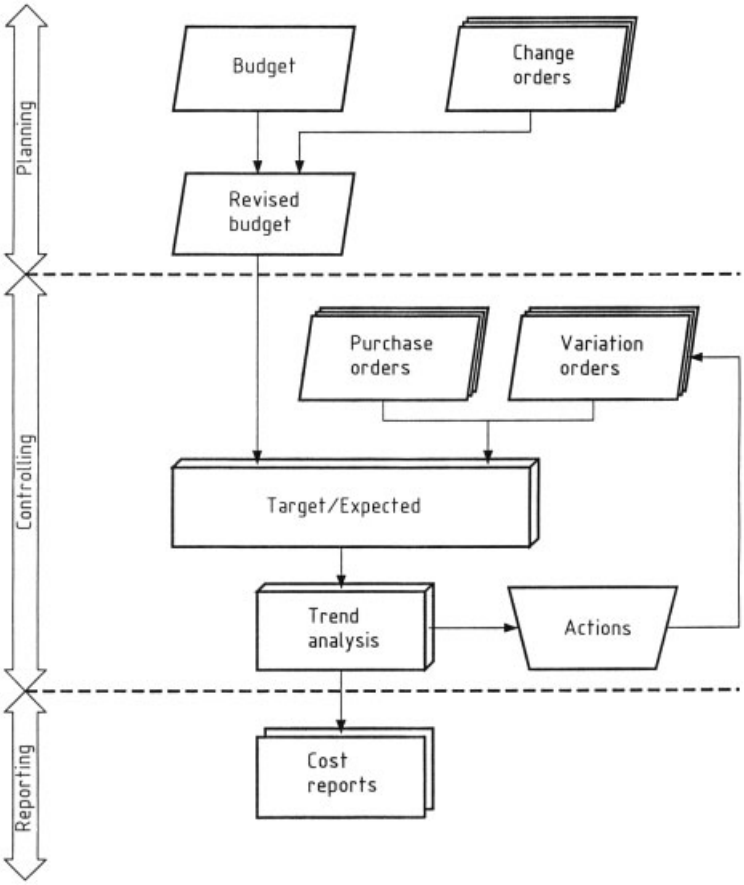

Figure 22. Principles of cost controlling

A data link between purchasing and cost control allows updating of actual costs. Cost control, however, is based on comparison of calculated and expected values, i.e., the calculated values are always compared with the costs forecasted up to the end of the project. This forecast is repeatedly updated for each discipline by the specialist engineer and the cost engineer. Forecasts are based on the order values at the cutoff date, the calculated values of equipment not yet ordered, and expected cost increases or decreases. The cost control cycle closes when data for the forecast and costs for approved changes are input to the cost control system.

Risk Analysis. The analysis of orders still to be carried out and services still to be performed is fed into the latest forecast. This analysis often reveals potential risks that may have a decisive influence on the project. The effects of these risks on costs and their probability of occurrence cannot generally be clearly defined. One approach to risk quantification is the Lurgi "project risk analyis" concept [125], [126]. The concept is based on the risk profile developed during expert conferences.

Cashflow Schedule. Knowledge of how to procure funds for equipment and services at the appropriate time is important for the investor, but also for the engineering firm working on supply or turnkey contract. A project cashflow schedule is a combination of cost planning, time scheduling, and progress planning. The following procedure is used to prepare such a schedule:

1) Planning of deliveries and services, including payment modalities agreed on with the suppliers of equipment and subcontractors

2) Determination of relevant actual figures
3) Deviation analysis (e.g., postponements)
4) Revision of the plan

Figure 23 is a graphical representation of a cashflow schedule. Payments can be read off in detail from the associated lists.

5.3.4. Project Report

The preparation of regular reports for the management of the engineering firm and the investor is the task of the project manager. These project status reports should provide comprehensive information; information on cost and man-hours should not, however, be included in the report to the investor in the case of a lump-sum contract.

The report should contain the following items:

1) Text
   a) Summary (highlights): project status, trends, problems
   b) Detailed information
      – Progress of individual disciplines (process engineering;civil engineering,structural steel;apparatus; piping, machinery, electrical, control systems, procurement civil work, and construction)
      – Problems in execution and recommended remedies
      – Forecast for the following month and key points
2) Graphs and lists
   – Time schedules
   – Progress curves (for individual disciplines and the project as a whole)
   – Network diagrams and lists
   – Procurement lists
3) Cost report
   – Analysis and assessment
   – Status of costs and man-hours
   – Cost report (computer printout), summary
   – Cost report (computer printout), detailed

5.4. Detail Engineering

5.4.1. Process Engineering

Most of the process engineering of a project is performed at the basic engineering stage (Sections 3.3 and 3.4). In addition, the process engineers devise the plant control concepts in collaboration with the control system engineers. The contribution of process engineers to the completion of piping and instrumentation diagrams extends far into the detail engineering phase. The process engineers also prepare the process description, which states how the process operates and identifies important control functions and special features.

In cooperation with the appropriate specialty engineers, the process engineers write detailed startup instructions which are needed for commissioning. The process engineers responsible for process design should also be involved in commissioning because they are familiar with the process and control details of the plant. They can also gain considerable expertise in operation of a plant, know-how that they can use in future design.

If the engineering firm is responsible for "authority engineering," the process engineer also makes a significant contribution to the approval documentation (see also Section 3.3.2.5), which includes specifications for products, byproducts, residues, and wastes; emissions; safety analyses; and HAZOP studies.

5.4.2. Plant Layout

Plant layout is an interdisciplinary activity. Its most important components are the layout (plot plan) and the piping and instrumentation diagram. Both documents are created, in preliminary form, during basic engineering (see Sections 3.4.3 and 3.4.4). They are continually updated in the course of detail engineering.

Plot Plan [38], [111]. The layout of, for example, a petrochemical plant is usually drawn to 1: 100 scale. It shows the outlines of equipment items, pipe bridges, and buildings, all dimensioned. In the case of buildings and platforms on tanks and towers, horizontal projections in appropriate planes and vertical sections are also included (see Fig. 24).

An experienced erection engineer should be involved in the layout planning so that allowance can be made for erection operations, in particular the space required for the installation of heavy

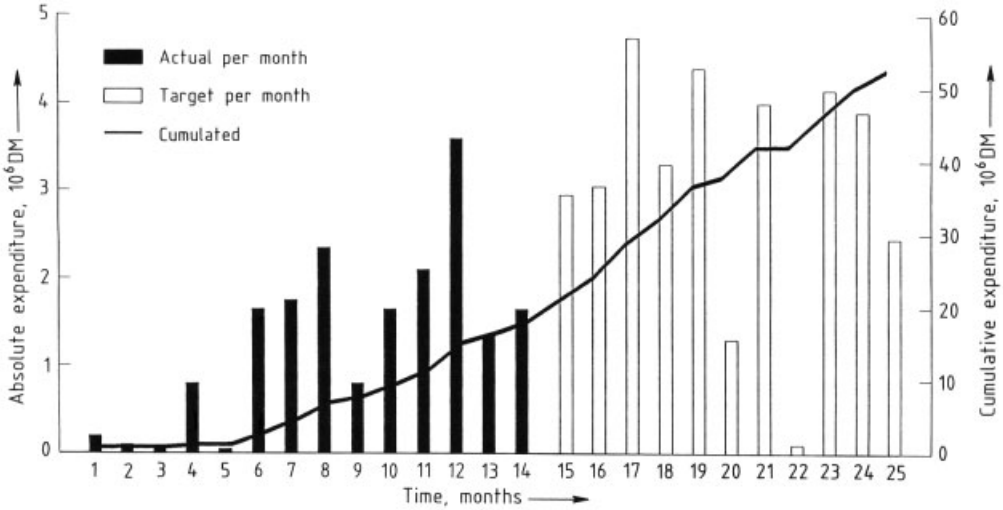

Figure 23. Cashflow schedule

apparatus. Other factors that must be considered include maximal exploitation of available space, placement of heavy apparatus on the ground floor, operability of important block valves, easy maintenance of machinery, and short runs of pipes with large diameters and/or made from expensive materials. There must be adequate safety distances between pieces of equipment, escape routes for operating personnel, and access for fire fighting vehicles.

It is also advisable to discuss the basic plot plan with the client's personnel (plant operators, maintenance personnel) at an early stage. The progress of construction and new detailed information provided by equipment vendors during detail engineering necessitate continual changes in the layout. In order to avoid delays in the start of construction, an attempt must be made to "freeze" the building dimensions and the positions of heavy apparatus as early as possible.

The final layout is the result of design work performed by all the disciplines involved in the execution of the project.

Piping and Instrumentation Diagram. The piping and instrumentation (P & I) diagram or mechanical flow diagram is based on the preliminary P & I diagram from basic engineering. The latter is developed and elaborated in detail engineering, as new information is acquired from the engineering disciplines and equipment vendors.

The final P & I diagram describes the whole plant in detail (by use of coding). The contents of the P & I diagram from the basic engineering stage (Section 3.4.4.) are supplemented during detailed engineering by:

1) Operating data and dimensions of equipment
2) Design data for machinery
3) Data on insulation and heating of equipment, machinery, and piping
4) Elevation of machinery and equipment
5) Information on noise abatement
6) Delivery boundaries
7) Codes for equipment nozzles
8) Codes for fittings
9) Detailed information on electricals and control systems

The utility systems (e.g., steam, instrument air, and condensate systems) are diagrammed separately from the process systems. In larger, more complicated chemical plants, P & I diagrams may comprise > 100 DIN A2 sheets.

The P & I diagram contains all essential information developed by the individual disciplines: process engineering; equipment, machinery, piping engineering; engineering for electricals and control systems. It must also include data provided by the manufacturers of equipment and machinery (e.g., the control system of a compressor). This information becomes available over a prolonged span of time.

The P & I diagram is therefore revised several times in the engineering stage (Fig. 25 is a

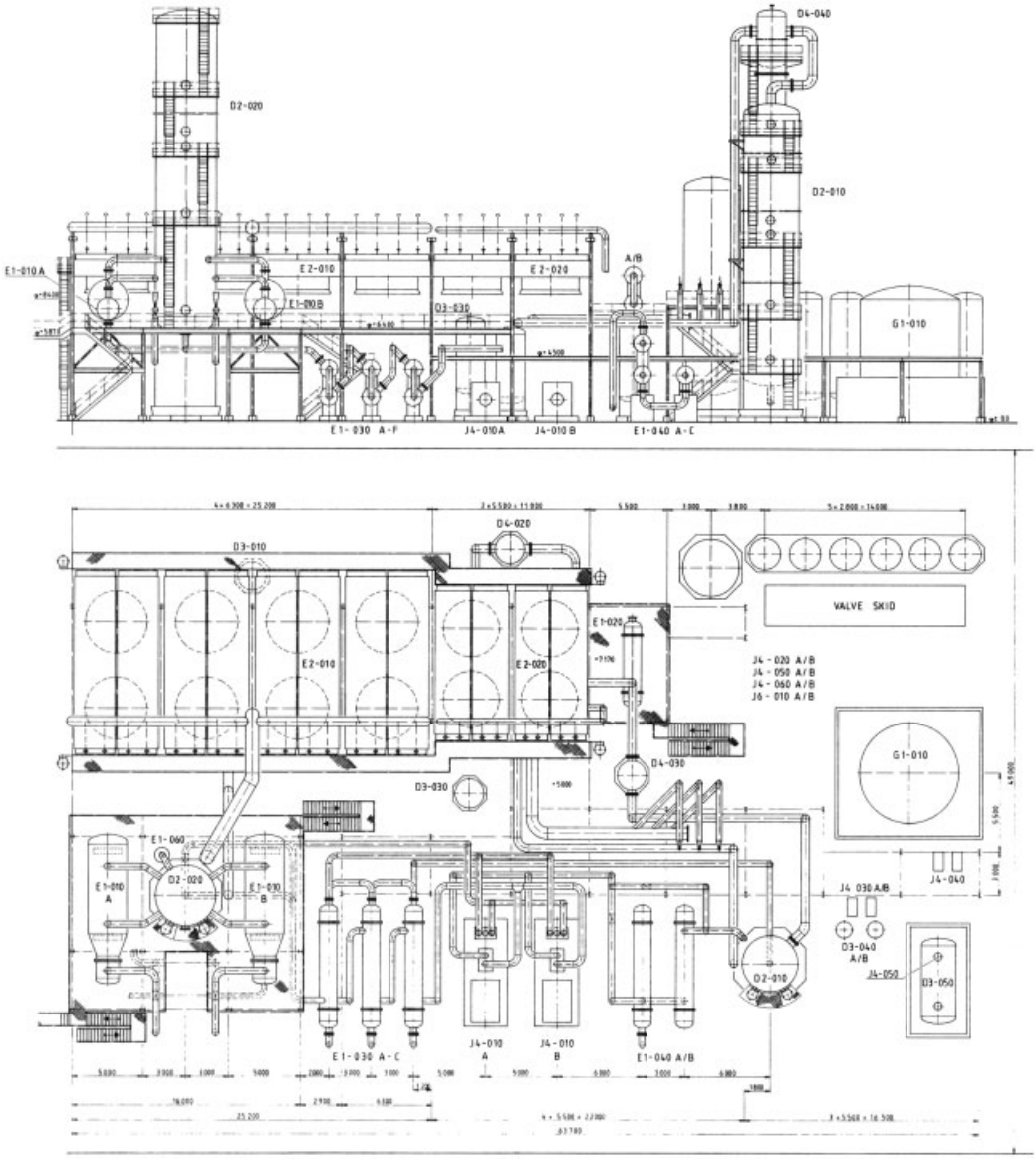

Figure 24. General arrangement drawing

portion of a P & I diagram). An attempt should be made to review the P & I diagram with all responsible persons (and, if possible, representatives of the client). After this review, the P & I diagram is "frozen" and only essential changes should be subsequently allowed (e.g., changes concerned with plant safety). Minor changes are often made during commissioning; these should be incorporated into the P & I diagram so that the document reflects the "as built" state of the plant.

5.4.3. Apparatus and Machinery

All important process engineering data for apparatus (e.g., heat exchangers, reactors, towers, vessels, tanks) and machinery (e.g., pumps, blowers, compressors, turbines) are specified

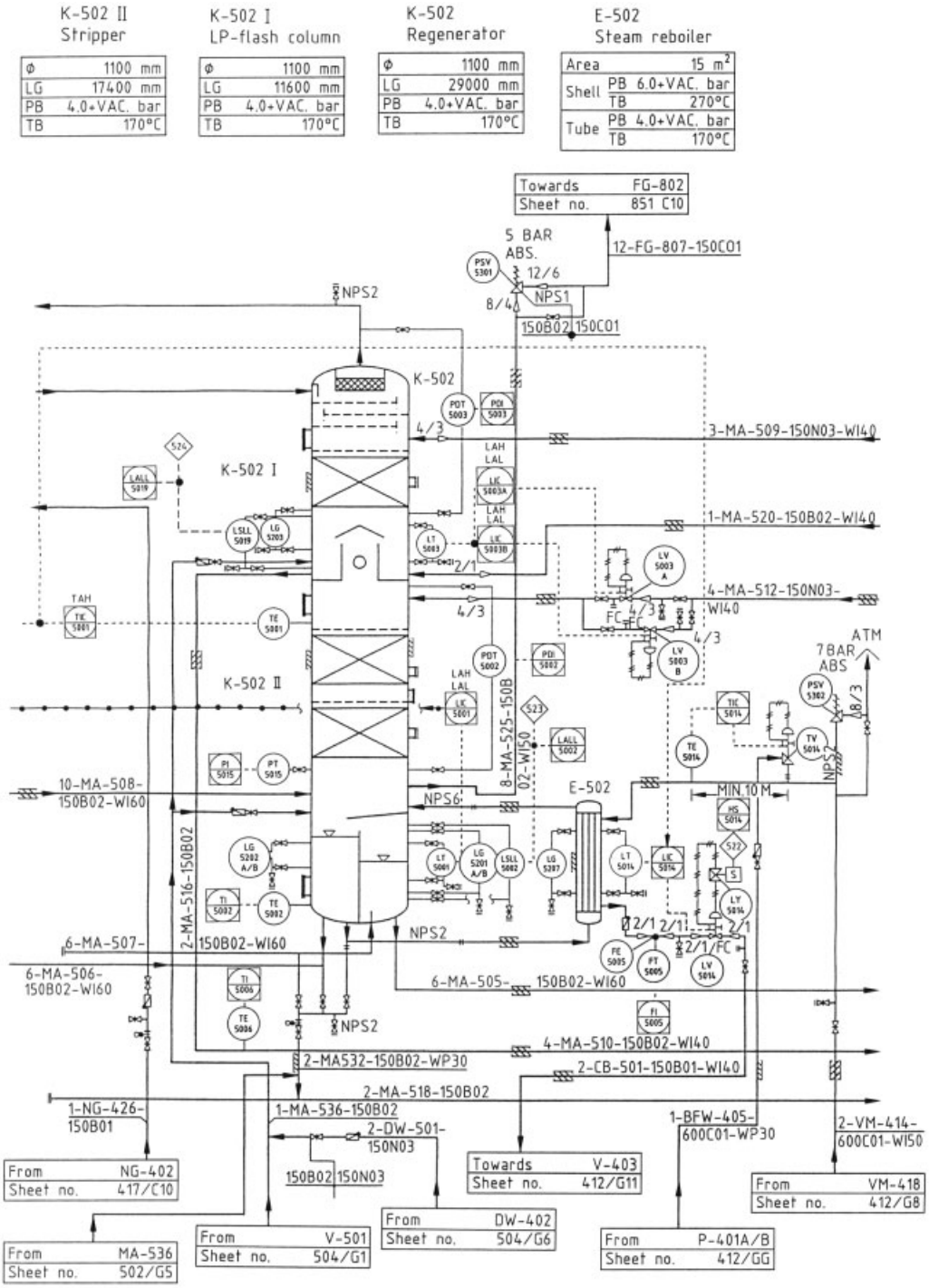

Figure 25. Section of a piping and instrumentation (P & I) diagram
(e.g., LIC = level control; TE = local temperature indicator; TI = temperature indicator in the control room; TIC = temperature control)

during basic engineering. Data sheets prepared in this stage contain essential information on overall dimensions, pressures, temperatures, quantities, materials of construction, etc., of each piece of apparatus and machinery. In detail engineering the apparatus and machinery engineers complete this information. The result of this work is a set of specifications in the form of drawings and descriptions, which enable qualified manufacturers to submit bids for apparatus and machinery.

The equipment engineers prepare so-called *guide drawings*, which are scaled drawings indicating all dimensions dictated by process engineering (e.g., number and diameter of trays in a tower, spacing of trays, and tower height). All dimensions of importance for shipping are also shown. Relevant legal provisions must be taken into account. Nozzle tabulation and other important design data are attached to the guide drawing (see Fig. 26). The wall thickness is estimated so that the weight of the apparatus can be calculated The number and dimensions of the nozzles, and frequently their elevations, are stated. The horizontal orientation of the nozzles is determined later when the exact position can be ascertained from the piping design.

The guide drawings and supplementary information form the technical portion of the bid invitation, which is sent to selected manufacturers. The information in the bid invitation must be presented in such a way that the bidders can submit comparable bids (see also Section 5.5.1).

The design office of the manufacturer prepares detailed workshop drawings and calculates the final wall thicknesses. The workshop drawings are checked by the engineering firm who also fixes the position of the nozzles and informs the manufacturer of any changes.

Once the drawings have been approved, production can begin. The manufacturer is responsible for compliance with legal provisions.

Specialist engineers periodically inspect complicated equipment, even during its production. Such inspections are independent of those per-formed by a third party (e.g., one of the German TÜV organizations or Lloyds) when mandated by law (e.g., for pressure vessels). The equipment is cleared for shipping only after final acceptance by the same specialist engineers.

The specification and procurement of machinery are similar to that for apparatus. In contrast to apparatus, which is usually custom-built and thus individually designed and drawn, an attempt is made to use off-the-shelf machinery. There are two reasons for doing so: to minimize engineering costs and to hold down the purchase price.

The machinery engineers prepare specifications for every machine to be procured. The specifications are based on the data sheets compiled by the process engineers, which contain all information relevant to the process (operating conditions, materials of construction). As an example, Figure 27 shows a data sheet for centrifugal pumps taken from the bid specification. The machinery manufacturer supplements the data sheet with further information on the model he has selected for the bid.

An important element of order handling is the time schedule according to which the machinery manufacturer is to submit information about the machine (e.g., dimensions, weight, vibratory behavior). It is important for the engineering firm to obtain this information as early as possible to avoid delays in the design of footings and foundations, buildings, and piping. Such information should also be finalized as soon as possible to avoid duplication of work.

For noise-abatement design see Section 3.3.2.2.

Spare parts required for plant startup and the first two years of on-stream operation are commonly ordered at the same time as the machinery. The manufacturer recommends the type and quantity of spares. The subsequent operator of the plant makes the final decision once the specialist engineer has checked the bid.

5.4.4. Piping

The objective of piping design is to prepare all drawings and specifications needed for procurement and installation of the piping components. The engineering of piping systems is closely linked with the engineering of all other disciplines. In the initial phase of engineering, information is incomplete and often preliminary. The data become more complete and exact as work progresses. Because the need for on-time availability of piping data requires early information from the other engineering disciplines, a step-by-step procedure is employed. Often the first

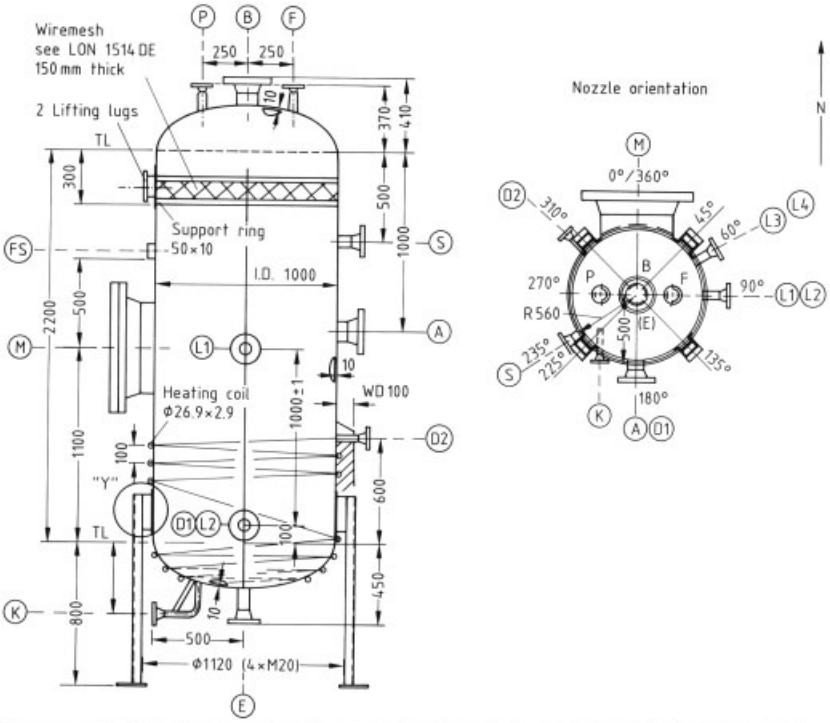

(B) Design						
10 Capacity	2.0 m³	21	Weight	Empty without int. and ext.	~1200 kg
9 Corr. components	-	20		Empty with int. and ext.	~1300 kg
8 Fluid γ=600–1000 kg/m³	Fuelgas	19		Operating	- kg
7 Corr. allowance	3 mm	18		with int. ext. and water	~3300 kg
6 Joint efficiency (shell/heads)	85%	17 Manufacturer			
5 Operating temp. (1A)	20°C	16 Manufacturer job no.			
4 Max. allow. working temp. (1B)	120°C	15 Inspection by			
3 Test press. (PP)	Dim. at manufacturer	14 Official inspection by			
2 Operating press. (PA)	5.5	13 Dimensional check			
1 Max. allow. working press. (PB)	10.0	12 Heat treatment			-
(A) General data		11 Partial press.			- bar

Item	No. reqd.	NPS	Class.	Facing	Length	OD. × wall thick.	Designation
S	1	50	40	C	S. DRWG.	60.3 × 5.0	Safety valve
P	1	25	40	C	S. DRWG.	60.3 × 5.0	Press. indication
M	1	500	25	C	250	508.0 × 14.2	Manhole (blind)
L1,2	2	50	40	C	190	60.3 × 5.0	Level indication
K	1	25	40	C	S. DRWG.	26.9 × 2.9	Condensate outlet
F	1	25	40	C	S. DRWG.	60.3 × 5.0	Vent (blind)
E	1	50	40	C	S. DRWG.	60.3 × 5.0	Drain (blind)
D2	1	25	40	C	190	26.9 × 2.9	Steam inlet
D1	1	50	40	C	190	60.3 × 5.0	Steam outlet
B	1	100	40	C	S. DRWG.	114.3 × 6.3	Fuelgas outlet
A	1	100	40	C	150	114.3 × 6.3	Fuelgas inlet
FS	1	–	–	–	–	–	Nameplate

Table of nozzles

Figure 26. Guide drawing for a vessel

LURGI

Lurgi GmbH

Machinery
Centrifugal Pump
Engineering Data Sheet (ITD)

01	●	Plant: HYDROCHLORIC ACID PLANT							
02	●	Service: FEED PUMPS			●	Specification class:			
03			No. req.	Pump type and size horizontal / ~~vertical~~ 1)	Manufacturer serial No.		Driver Kind	Type, size	Item No.
04	●	Operation	1	RN 32/250 C2	132 579	●	AC-MOTOR	180 M	P-1701 A
05	●	Standby	1	RN 32/250 C2	132 580	●	"	180 M	P-1701 S
06		Purchaser:		Enq.-No.: A 5184	Date: 11.1.90		Supplier:	Prop.-No.: 910.1.0323	Date: 13.2.90
07				Order-No.:	Date:		Rheinhütte	Contr.-No.:	Date:
08		Site conditions							
09		**Operating conditions**							
10	●	Liquid 5)		CHLORINATED HYDRO-		●	Inlet gauge	rated	0,4 bar
11	●	Hazard		CARBONS		●	pressure 2)	maximal	bar
12	●	Corrosion by		TRACES OF HCL		●	Outled gauge	rated	9,4 bar
13	●	Solids	Type	FINE COKE PARTICLES		●	pressure 2)	maximal	bar
14	●		Mass percentage	-	%	●	Diff. pressure rated		9 bar
15	●	Operating temp. t_A		20	°C		Head curve	rated / max.	69,5 / 72 m
16	●	Density at t_A		1,321	kg/dm³	●	NPSH at rated and	$NPSH_A$	7,1 / 3,7 m
17	●	Kinematic viscosity at t_A		0,35 - 1,3	cSt ~~mm²/s~~		max. flow	$NPSH_R$	3,2 m
18	●	Vapour pressure abs. at t_A		0,11 - 0,55	bar		Pump speed rated		2900 min⁻¹
19	●	Flow	rated	7	m³/h		Pump power input	rated	9,3 kW
20	●		normal / max.	6,5 / 7	m³/h		P_a	normal	- kW
21	●		min. req. / perm.	3 / 2	m³/h		P_a max.	at rated impeller-⌀	14 kW
22								at max. impeller-⌀	20,5 kW
23							Rated driver output		15 kW
24						●	Self priming		~~yes~~ / no 1)
25		**Construction features**							
26	●	Nominal pressure		16			Casing support		Feet below
27		Allowable	pump	16 bar	at 120 °C		Casing split		radial / ~~axial~~ 1)
28		work. pressure	auxiliaries	bar	at °C		Total clearance	Wear ring / plates 1)	- mm
29		Test pressure 2)		20	bar			Shaft bushes	- mm
30		Number of stages		1				Balance drum	- mm
31		Impeller-⌀	rated / installed	235 / 232	mm		Thrust reduction by		IMP. HOLES
32			maximal / minimal	265 / 200	mm		Radial bearing	Type, Size	-
33			Type	CLOSED			Axial bearing		3308 / C3
34		Rotation facing	pump 3)	clockwise / ~~anticlockwise~~ 1)			Connection bearing		-
35		pump driven end	driver 4)	~~clockwise~~ / anticlockwise 1)			Lubrication	supply	GREASE
36		Inlet	Size / Position	50	AXIAL		Shaft seal	Arrangement	MECH. SEAL
37	●	flange	Rating and facing	DIN 2501	PN 16			Manufacturer	Burgmann 4)
38	●	Outlet		DIN 2501	PN 16			Type, Size	M7N / M7N
39		flange	Size / Position	32	TOP			Allowable work. pressure	16 bar at 120 °C
40		Vent		-			Aux. piping	Arrangement	see sheet 9
41		Drain		YES			Shaft seal	Supplied by	Rheinhütte
42		Bearing housing size		II			Volute / Diffuser 1)		single / ~~double / multi~~ 1)

43 ↑ ● The marked lines are to be completed by the purchaser. 5) COMPOSITION SEE SHEET 8

Remarks: 1) Cross out where not applicable. 3) Viewed from the drive to the pump.
2) Pressure are tobe stated in gauge pressure. 4) Viewed from the pump to the drive. 6) WITH PACIFIC THERMOSYPH.-VESSEL

Completed: FHU	Process: TTL	Job or Project No.	Job:		Line revised in Rev. Modified: ↑	
Checked: Mo	Document-Type 1,5,6 No. D,6,6,3,5,7,0,0,4,L,0,0,0,53			Sheet: 6	Rev.: 3	Date: 9.1.90	Checked: Mo

Figure 27. Data sheet for a centrifugal pump

steps are based on assumptions, so that frequent corrections are required later.

Piping accounts for a relatively high proportion of chemical plant costs and piping engineering may represent as much as 20 – 40 % of total engineering. Refineries and petrochemical plants lie at the upper end of this range.

Piping engineering can begin once the following information (at least in preliminary form) is available:

1) Standards and codes (of the investor or the engineering firm)
2) P & I diagram
3) Layout model
4) Plot plan
5) Guide drawings for apparatus
6) System drawings for machinery
7) Preliminary civil and structural steel drawings
8) Data on electricals and control systems

Piping Specification. For the sake of efficiency in engineering, procurement, and piping installation, and in view of the wide variety of piping components and design, a piping specification is first prepared. This document is based on the standards and codes applicable to the plant and relevant engineering regulations. The piping specification also contains special piping design guidelines for the project.

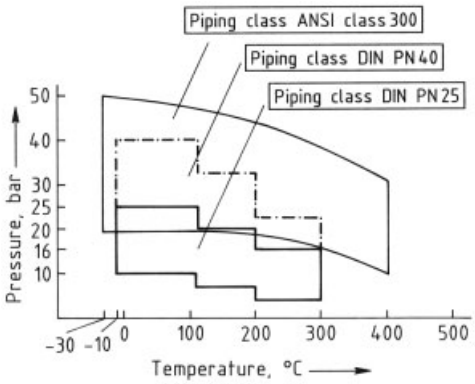

Figure 28. Piping classes (pressure – temperature diagram)

An important element in the piping specification is the piping classification which minimizes the different types of piping components required. All piping components (pipe, fittings, flanges, bolts, seals, etc.) are classified on the basis of flowing media, pressures, and temperatures that occur in the plant.

A piping class comprises the expected dimensions of components and their materials of construction for a given set of media, pressures, and temperatures. The classification is based on pressure – temperature diagrams from DIN 2401 or ANSI B 16.5 (see Fig. 28). Each of the areas represents one piping class. All components within one such area are uniformly sized. Fittings and flanges are standardized according to pressure level. Wall thicknesses are calculated from the pressure and temperature. Once the piping classes have been worked out, the figures are stored in a data base and can be retrieved as needed. The use of such a data base greatly reduces the amount of work to be done in specifying the piping system for a plant.

Piping List. All pipe runs are identified by a code number and a piping class, and are compiled in a piping list. The associated data are stored in a data file. The piping list is prepared at the same time as the P & I diagram.

Isometric Piping Drawings and Piping Model. Drawings that show both the geometry of the run and its location in the plant are needed for the prefabrication and installation of piping. It used to be common to plot every pipe run and every fitting in a piping diagram with plan, elevation, and section views. This method has been largely replaced by a diagram of a single pipe run and its components, along with measurement and control devices and piping supports.

The initial piping studies and the final isometric drawing of a pipe run are done in parallel with the construction of a piping model. A bill of materials containing all piping components is drawn up for every pipe run. Figure 29 gives an example of such an isometric drawing. Computer-aided design (CAD) techniques are used increasingly in preparing isometrics.

The isometric drawings are supplemented by plans for pipe bridges and underground pipe runs.

The model shop uses the plot plan to make a basic model at 1 : 33 1/3 or 1: 25 scale that includes all apparatus, buildings, frameworks for equipment, stairs, ladders, platforms, and pipe bridges (Fig. 30). The piping model forms

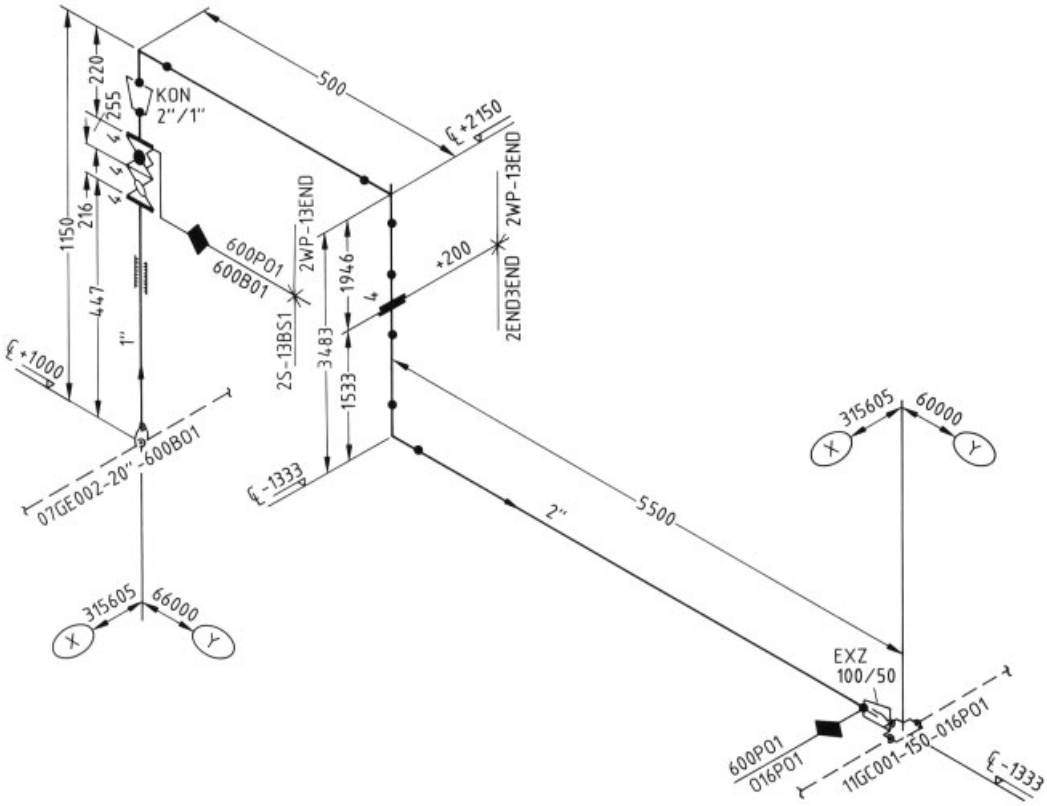

Figure 29. Isometric piping drawing
Symbols and designations mark piping components (e. g., valves, flanges, reducers), dimensions, position of the pipe run

the center for coordination of all detail designs [127].

Installation work carried out on the model in parallel with piping includes air conditioning and cable ducts, control panels, hoists, and cranes.

A piping model has the following advantages:

1) Piping routes are easy to check for collisions with other equipment
2) The plant operator can check ease of operation and maintenance
3) The model can serve as a training facility for operating personnel and as a form of instruction during installation

The completed model should be thoroughly assessed by all discipline engineers, the later maintenance engineer, and the operator. Any necessary changes made at this stage are much less costly than if they are made later on the construction site.

Piping Calculations. Piping calculations cover strength calculations for individual elements (wall thicknesses, flanged joints) and stress analysis of the piping system.

Wall thicknesses are calculated from the pressure and temperature ratings of the piping class.

The safe functioning of a piping system depends on correct sizing and proper layout. Special attention should be given to the elasticity of the piping and the use of supports and anchors. Temperature changes give rise to stresses in piping systems, which in turn generate forces and moments at connection and support points. Computer-aided elasticity calculations are performed to make certain that the strains resulting from stresses in a given piping layout are within the elastic range. If the stresses are too high, a different configuration must be selected or compensators must be inserted. For small-diameter pipes and moderate temperatures, these expen-

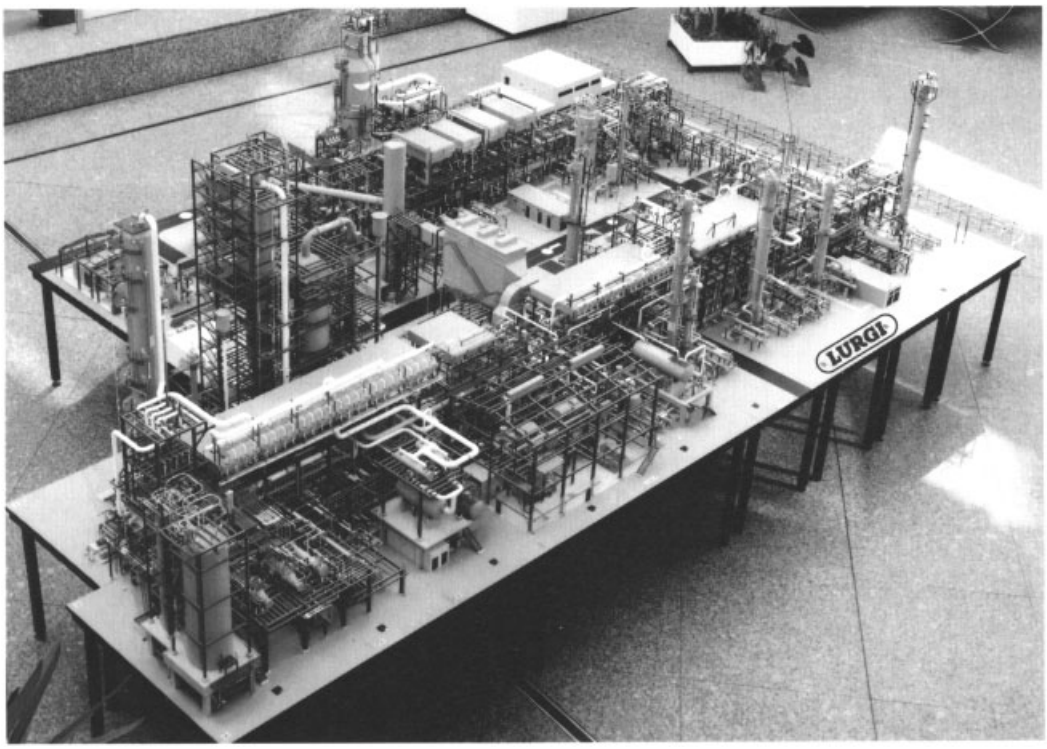

Figure 30. Piping model (fluid catalytic cracker complex, courtesy of Lurgi AG)

sive calculations are often superfluous, since experienced piping engineers lay out such piping with adequate elasticity.

Material Takeoff and Procurement. A first rough material takeoff can be performed when the P & I diagram has reached a certain level of completeness and the plot plan has been drawn up so that fittings can be counted and the lengths of the main pipe runs can be estimated. The objective of this preliminary takeoff step is to invite bids and place orders for piping components with long delivery times. Sufficient material can thus be made available on site when piping installation begins.

As piping design advances, isometrics and piping plans with bills of material are generated. Another (not yet final) material takeoff is then prepared and has a higher degree of accuracy than the preliminary takeoff. Further orders are then placed. At this stage, bid invitations can be sent for the installation work, and the piping installation contractor can be selected. The final material takeoff is worked out after all isometrics and piping plans are complete.

The preparation of the material takeoffs is computer-aided. Sorting and condensing programs calculate the quantities that have to be ordered [128]. An integrated materials management system allows the print out of lists and calculations for every step (e.g., bid invitation, ordering, expediting, and material handling on site). The efficient use of such a system requires a consistent high-order data structure as well as the unambiguous definition of piping components in terms of piping classes.

Insulation and Coatings. The thickness of insulation needed on equipment and piping must be established at a fairly early stage since this value may influence other parameters such as the length of nozzles and the width of pipe bridges. Insulation thicknesses are entered on the P & I diagram and in the piping list. Thermal insulation for pipes carrying hot media generally consists of mineral wool enclosed in galvanized or aluminum jackets. Polyurethane foam enclosed

in sheet metal is widely used for pipes carrying cold media to prevent icing and cold bridges.

Uninsulated surfaces of tanks, piping, and steel structures must be painted to protect them against corrosion. As a rule, machinery is delivered with the specified prime and topcoats. Rust must be removed from the surface before application of the prime coat. In many cases, a second prime coat is needed before the first and second topcoats are applied.

Underground pipe is either coated with asphalt or jacketed in plastic.

5.4.5. Control Systems [129–133]

Automation of chemical plants is increasing. Rapid progress in microprocessor technology has led to the development of distributed control systems (DCS) that can meet the increasingly stringent requirements of modern process operation. The objectives are to improve the availability of operating process plants, enhance their reliability, and optimize their operation.

The distributed functions facilitate the engineering, operation, and maintenance when broken down into levels (Fig. 31).

Centralization of process control systems means that plants are chiefly fitted with electronic devices, since this kind of equipment with its reliable signals is suitable, even for explosion hazard areas. Pneumatic instrumentation is limited to pneumatically actuated controllers and low-order local control loops.

With the help of process monitoring and control systems, advanced control strategies can be built up in modular form. The modules perform both computing and dynamic functions, so that a variety of signal processing algorithms can be selected for optimal control strategy.

The control of material streams, plant optimization, and balancing is implemented by process control computers at a level above the process control systems. These computers may have an on-line function or a data management function. Software vendors offer appropriate modular software. Such programs are linked together to permit control of products by centralized programs.

Optimization programs did not gain wide use until stable on-line analytical instruments with short response times were developed. The additional investment is amortized after as little as one to two years. Complicated analytical systems, however, have higher costs for maintenance, which is performed by specially trained personnel.

Analytical instruments are built in prefabricated enclosures and tested at the vendor's workshop. Installation simply involves connecting the process loop lines, utilities, and data cables.

Plant Safety and Availability. Regulations on plant safety and environmental protection have rapidly become more stringent, influencing the choice of automation hardware and system structure. Safety control requires the use of redundant systems approved by the regulatory authorities. Interfaces connect these systems to the process control system, special attention should be paid to transmission time between the different systems.

Process Monitoring and Control System. In petroleum refineries and petrochemical plants, it is often necessary to operate 3000 – 4000 loops and give the plant operator access to these in a meaningful order.

Sensor signals relating to the process and control functions, along with signals to motor control centers (MCC) and valve actuators, are handled in the processing stations, which perform configured tasks such as signal conditioning, control, and signal processing. The processing stations are assembled from modules and tailored to individual functions. Process information is transmitted to the operating and monitoring system via serial busses.

The chemical plant is controlled with extended software functions for graphical display, along with process graphics overlays. The hierarchical information structure of data representation leads the operator to the proper level in the information structure. The alarm functions notify the operator directly of the inititating measuring point in the process. The automation structure is governed by the following important criteria:

1) Size of process plant
2) Continuous or batch process
3) Behavior of process over time
   - Steady and stable
   - Product output or quality strongly affected by load variations and/or perturbations

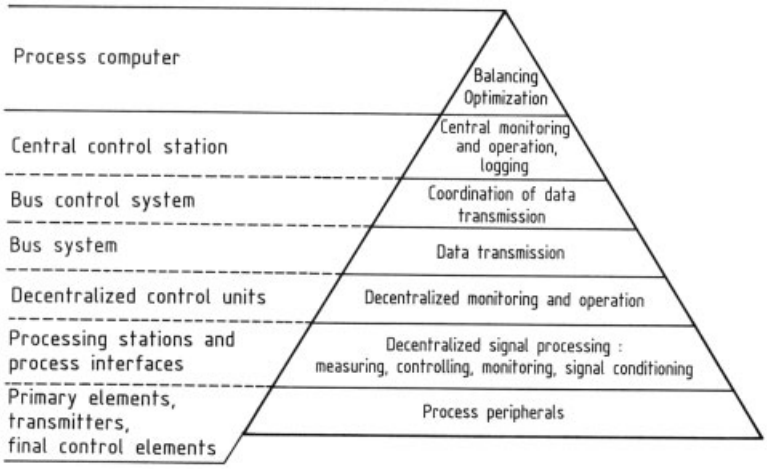

Figure 31. Levels of distributed control systems (DCS)

4) Complexity of process
   – Simple control strategies
   – Complex interdependences
   – Frequently changing recipes
5) Local or central process control
6) Startup and shutdown strategies
7) Upgradability
8) Amenability to changes
9) Capability of linking with automation systems at other plants, other monitoring/control levels, or information systems
10) Type of reporting
11) Safety, availability
12) Standardization
13) Environmental restrictions
14) Maker and service capabilities
15) Personnel considerations (crew size, qualifications)
16) Economic and management aspects

Only close collaboration between process engineering and process control specialists belonging to the staff of the engineering contractor and the owner can ensure proper decisions.

In the *central control room*, the operating and monitoring devices collect all needed process information. All signals from the plant (flow, pressure, temperature, etc.) are dynamically displayed on a screen. Two to three screens per workstation, with the necessary operating features (touch screen, light pen, keyboard), have proved optimal with regard to cost and volume of information. The process control system includes reporting features that maintain a continuously updated record of alarm and condition reports, series of measurements, balances, and operator actions. Trend displays of the process variables are replacing conventional chart recorders. The latter are needed only as required by the regulatory authorities (e.g., emission measurements) or to record guaranteed values (e.g., temperatures in a catalyst bed).

The workstations and peripherals should be arranged so that the operator sees the whole working field as a closed area and operators can exchange information in order to coordinate their actions in case of abnormal occurrences (Fig. 32).

Engineering of Control Systems. It is useful to break the engineering of modern measurement and control equipment into field devices and central control rooms (process control systems). Different levels of detailed knowledge are needed for these two areas. The high rate of innovation in process monitoring and control systems demands continuous retraining of the design engineers.

The main activities involved in engineering follow and make use of computer-aided engineering (CAE) systems with different software requirements:

1) Preparation of basic documents such as coding system, power supply and distribution, materials of construction, P & I diagrams, control strategy, functional diagrams for process and device control systems, instrument

Figure 32. Control room, courtesy of Lurgi AG

list, quantity structure list, requirements for process measurement and control system

2) Preparation of instrument specifications such as data sheets for all instruments to be installed, with information on designation of measurement location, process data, manufacturer's data, materials of construction
3) Engineering of process measurement and control system: configuration documentation, process graphics, loop sheets, description of process measurement and control system
4) Planning of central facilities such as power distribution, instrument cabinets, monitoring rooms
5) Preparation of installation documents such as cable run plans, cable lists, hookups, and list of installation materials

The pareparation of as-built documents once the plant has been commissioned and the maintenance of important documents thereafter facilitate plant maintenance and the remedying of malfunctions.

Information on reliability, maintenance cost, spare parts management, experience, availability over an extended time, and service are important factors in the selection of instruments and systems. Ease of access to instruments and systems greatly reduces the number of plant malfunctions and thus increases profits.

5.4.6. Electrical Design

The objective of electrical design is to supply electric power reliably and economically to all consumers. The designer does not create isolated solutions component by component but must find the optimal solution for the system as a whole.

Design begins where high-voltage power enters the plant, and may include medium- and low-voltage switchgear, transformers, generators,emergency backup sys-

tems,lighting,grounding, and communications. It covers three areas:

1) Planning of power generation and distribution
2) Planning of electric utilities
3) Installation planning

Planning of Power Generation and Distribution. The operator wants to optimize use of the electric power and to insure that the system can handle short- circuit loads, that the power grid can handle short- circuit loads, that the power grid offers the necessary reliability, that investment costs are minimized, and that operating costs are held down.

The following points therefore have to be examined carefully in the design of the distribution grid:

1) Selection of voltage levels
2) Determination of transformer ratings
3) Location of load centers
4) Location of distribution stations (with allowance for danger zones)
5) Reliability of supply from electric utilities and/or in-plant generating capacity
6) Materials of construction

The elements of the electrical grid must be selected and sized. The use of powerful computer programs is indispensable (to keep the consumer list, perform short- circuit/load – flow calculations, determine the run up behavior of motors, and carry out the sizing of cables). The results of these steps are entered in the block diagram.

The list of electrical consumers gives a detailed description of energy consumption and provides the basis for the energy balance, in which the installed power and net power demand are calculated.

Planning of Electric Utilities. Once the results of the above activities are available, documents needed for specifying the electrical utilities are prepared. These include technical specifications, engineering data sheets, circuit diagrams, cable lists, terminal diagrams, and mimic diagrams. These technical procurement documents form the basis for bid comparisons and order specifications. The plans are prepared by CAD methods.

Installation Planning. Electrical installation accounts for a significant fraction of investment costs, so detailed planning is a prerequisite for economic execution. At this stage access to a plant model is extremely helpful.

The installation plans are drawn up by CAD methods on the basis of layout plans and comprise:

1) Position plan for electrical consumers
2) Cable run plan
3) Cable run sections
4) Grounding position plan
5) Lighting layout plan
6) Hazardous area classification
7) Layout plan for communication systems

On the basis of the designs and documents prepared, the quantities needed are determined and bid invitations are prepared for bulk materials and installation. The ordered equipment is inspected at the workshops to check for compliance with the specifications.

The cost of electrical equipment and materials makes up 6 – 10 % of the total chemical plant costs. The high end of the range applies to grass-roots plants where a new infrastructure must be created. For further information, see [134], [135].

5.5. Procurement

Procurement activities consist of three main tasks:

1) The purchase of plant components and services
2) Expediting during the fabrication of plant components (i.e., supervising the fabrication sequence of plant components according to an agreed time schedule)
3) The shipping of plant components to the construction site

In U.S. oriented regions, workshop inspection of plant components also comes under this heading.

Depending on the terms of the contract, the engineering firm procures plant equipment in its own name or in the name and on behalf of the investor. The procurement department of an engineering firm is acquainted with the world market

and carefully observes trends. The procurement and engineering activities are closely linked together. The purchasing and shipping agents as well as the expediters are also members of the project team, and thus subordinate to the project manager. Procurement man-hours make up 8 – 12 % of total engineering hours spent on project execution.

5.5.1. Purchase of Equipment and Services

The main steps in purchasing are:

1) Preparation of a vendor list, possibly in collaboration with the investor
2) Preparation and dispatch of bid invitations based on requisitions written by the specialist engineers
3) Handling queries from bidders, checking on-time receipt of bids, checking received bids for completeness
4) Checking bids for comparability by the engineer
5) Preparation of bid comparison and order recommendation by the purchasing department
6) Checking the order recommendation (by cost engineers, specialist engineers, and possibly the project manager)
7) Negotiations with the assistance of the specialist engineer and possibly the project manager
8) Preparation of order documents
9) Checking order confirmation
10) Approval of bill payments after confirmation from the specialist engineer
11) Compilation of lists of bid invitations, bid comparisons, and orders

Bid Invitations and Comparisons. The engineer responsible for a particular discipline prepares specifications for the components he needs in the form of data sheets, guide drawings, descriptions, and information on when each item will be needed. These documents are sent to the purchasing department, which adds relevant business conditions and sends the packages to selected bidders. The potential bidders are chosen by the responsible engineers and purchasing agents when the vendor list is drawn up. If goods are purchased in the name and on behalf of the investor, the investor commonly has a say in the process. Good definition of equipment items covered by the bid invitations is important so that bids from competing vendors are comparable as to content and thus price.

The above procedure is also followed in the procurement of services. The technical documentation (e.g., for installation of piping) is prepared by the specialist engineers responsible for piping, in close cooperation with the erection planning department. This documentation includes the piping material takeoff, plant layout, specifications for piping installation, information on material storage capabilities, and schedules. The received bids for plant components and services are examined by the responsible engineers to insure that they are comparable and conform to the requirements stated in the bid invitations. Bidders may often be asked to correct their scope of delivery and services. The bid price may be adjusted as a result.

Technical bid comparison is followed by commercial bid comparison and an order recommendation. Often, the final decision is only made after verbal negotiations with two or three bidders. Decision criteria include not only lowest price but also the technical reliability of the plant component covered by the bid, the experience and reliability of the manufacturer, and the vendor's solvency and workshop capacity utilization.

Orders. Especially in the case of complicated plant components and large service packages, the order often goes out in abbreviated form by Telex, to make the best possible use of the agreed-on lead time. The detailed order document follows immediately. The order is prepared by computer-aided techniques so that, for example, the value of the order is transferred directly into the computer-aided cost control system.

If the plant components are clearly defined, the specification used for the bid invitation can also be employed for ordering. For larger and more complicated items, the scope of delivery or service must be unambiguously described with a statement of exclusions. Imprecision at this point can result in unpleasant confrontations with the vendor. Receipt of the order is confirmed in writing by the vendor.

5.5.2. Expediting

The engineering firm has an obligation to the investor to erect the plant within a certain time. The time schedule agreed between the engineering firm and the vendor must therefore also be complied with. The engineering firm establish a production schedule monitoring system for this purpose. Supervision begins as soon as stocks of material are ordered, it covers the manufacturer's design work (workshop drawings) and other production steps. Expediters from the engineering firm carry out their checks by telephone calls and regular visits to material vendors and the manufacturers of plant equipment. They report regularly to the expediting engineer and the project manager.

When delays are expected, corrective measures must be instituted in collaboration with the manufacturer (e.g., changing material vendors, weekend work, night shifts).

5.5.3. Shipping

The packaging and shipping of plant components to the construction site may be included in the order given to the vendor, or may be the responsibility of the engineering firm. In larger projects, the engineering firm should take responsibility for coordinating and supervising packaging and shipping. The principal activities of the shipping department are:

1) Checking the specifications for packaging, shipping, payments, and duties (these are usually part of the contract with the investor)
2) Drawing up plans for the delivery of equipment
3) Issuing standard invoices for import licenses
4) Obtaining packaging bids and issuing orders
5) Checking container lists and obtaining transport approval for large and heavy containers

Further activities include booking freight space, procuring insurance, and supervising loading and transport. Shipping, customs, and bank documents (including invoices) must also be prepared. Finally, damage and faults must be taken care of.

5.6. Planning and Execution of Civil Work and Erection

The main execution phases of a project up to mechanical completion (engineering, procurement, civil work, erection) overlap one another in time. The sequence of engineering work should guarantee that

1) Plant equipment with long delivery times (e.g., compressors, complicated apparatus) can be ordered as early as possible
2) Civil work (e.g., foundations, cable ducts, buildings) is begun early so that equipment erection is not delayed

Civil work should be begun as soon as the engineering work is 25 – 30 % complete.

By way of example, Figure 33 shows the project master schedule for the design and construction of the expansion of a refinery complex including the progress curves for engineering and construction. The time to mechanical completion is 30 months if basic engineering (which must be performed by the licenser) is complete at the outset. Usually the time required for basic engineering is four to six months. Figure 33 also shows that engineering office work is only complete by the time of "mechanical completion." The remaining work includes the preparation of final documentation.

5.6.1. Planning of Civil Work and Erection

The planning of civil work and erection is part of detail engineering.

5.6.1.1. Planning of Civil Work (Including Structural Steel Work)

As a rule, engineering for civil work and structural steel is done by engineers in the civil engineering department assigned to the project team. Often, however, their activity is limited to basic civil design, while detailed civil design is assigned to engineering firms in the country where the plant will be built. These firms are familiar with local conditions, know the local regulations, and have short lines of communication to the construction site and the firm performing the civil work.

Important information required at the start of basic civil design includes:

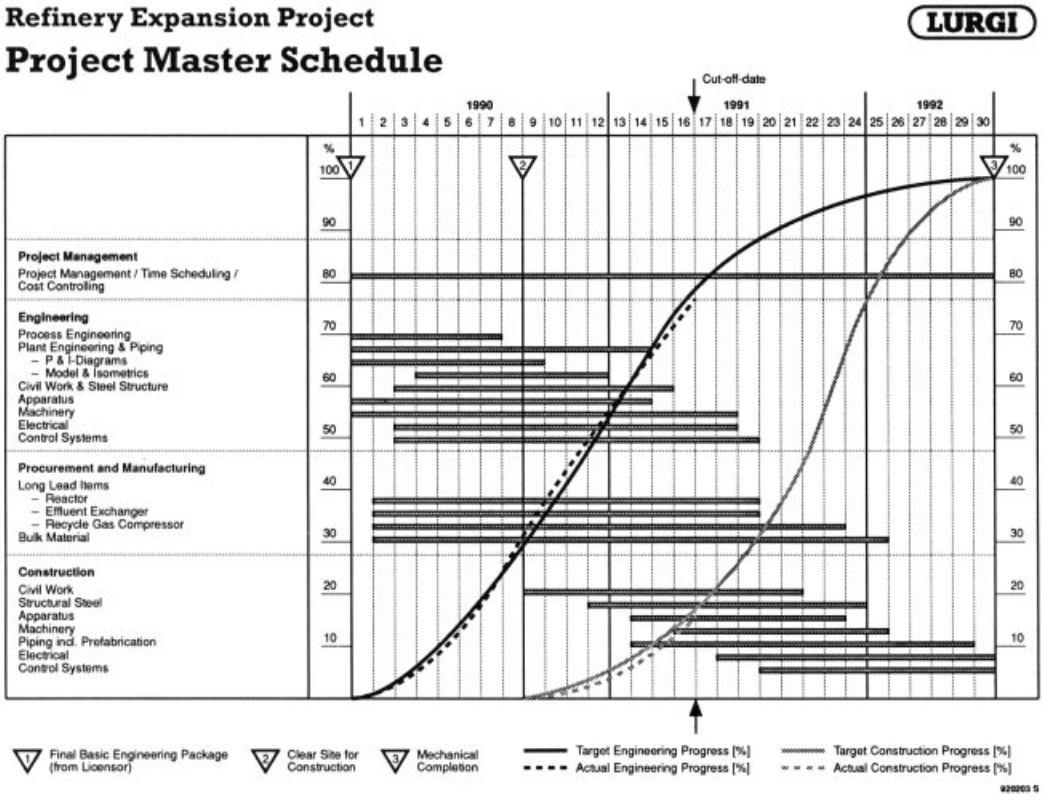

Figure 33. Project master schedule (refinery expansion project)

1) Soil evaluation report that contains, data on the subsurface soil conditions at the plant site, water table, water analysis, and soil bearing capacity. It should also give settlement calculations so that plant components subjected to severe dynamic loads can be calculated and designed.
2) Data on the proposed wastewater system.
3) An approved layout.
4) Geological and climatic figures such as earthquake factor, prevailing wind direction, and severe snow conditions.
5) Static and dynamic load data for the foundations of machinery, equipment, furnaces, and steel structures.
6) Footprint dimensions of machinery and equipment: piping and cable cutouts in floors, platforms, and walls.

On the basis of this information, preliminary plans are drawn for foundations, buildings, steel frames, traffic routes, underground piping, sewers, wastewater systems, and heating/ventilation/air conditioning systems. They are the basis for detailed planning by the civil engineering subcontractor or in the office. Together with the estimated quantities determined for steel and concrete, these documents are used in inviting bids. The detail engineering for structural steel is usually carried out by the structural steel supplier.

After the submitted bids have been evaluated, the contractors for civil work and structural steel are selected on the basis of qualitative and price aspects.

5.6.1.2. Erection Planning

Planning the installation of plant equipment starts at a relatively early stage in the engineering process. The sequence of installation activities can strongly influence the detailed scheduling of engineering and procurement. The sched-

ule for overall project execution should be developed backward from the agreed mechanical completion date and specifically for equipment with long delivery times. In large plants, separate schedules are worked out for each plant section. Erection schedules are revised at intervals throughout detail engineering and procurement as the agreed equipment delivery times are incorporated into the schedules.

Drawing up the plot plan calls for the cooperation of an experienced erection engineer, who must consider in particular the erection requirements for heavy components (space requirement, accessibility).

The land on which the plant will later be built must be prepared prior to construction work. The planning of temporary facilities is generally the responsibility of the engineering firm. Besides surveying and leveling, it is necessary to plan for the delivery of utilities and the removal of runoff and sewage. Other facilities include construction offices, stores, open-air storage areas, site roads, site fencing, piping prefabrication shops, communication facilities, toilets, first-aid station, guardroom, changing rooms, and accommodation for subcontractor personnel.

The engineering firm subcontracts installation work to qualified firms specializing in the erection of structural steel, apparatus, machinery, and installation of piping, electricals, and control systems. Subcontracting takes place in the detail engineering stage as soon as sufficiently exact information is available on the plant components and bulk materials. This is especially important when installation is covered by a fixed-price contract.

5.6.2. Execution of Construction

The construction work performed by specialist subcontractors is generally directed by the engineering firm. If the contract provides for the investor or a third party to do the construction, the engineering firm only supplies technical advisory services.

5.6.2.1. Construction-Site Organization and Management

The construction manager and his team supervise, coordinate, and direct construction. Responsibility for all site activities belongs to the construction manager, as specified in the contract, relevant codes, and regulations. The construction manager is answerable to the project manager and is the engineering firm's principal representative to the investor on the site. The organization of the construction team must take account of the size and complexity of the project, time schedule, local conditions, and contractual obligations.

Figure 34 shows a typical site organization for a large project. Commonly the investor maintains a similar, but smaller organization so that discussions can be carried on at all technical levels.

The main tasks of the construction site team are:

1) To plan, coordinate, and manage all site activities
2) To arrange for site offices and establish site security systems
3) To organize and oversee materials management
4) To perform scheduling and progress control
5) To define working methods
6) To prepare and carry out quality control
7) To establish and supervise work safety procedures
8) To coordinate and oversee the work of construction subcontractors
9) To manage and clarify the construction documentation
10) To implement a cost control system and arrange payments
11) To prepare construction-site orders
12) To prepare deficiency reports and control insurance cases
13) To prepare as-built drawings
14) To implement a change order management system
15) To submit reports to the investor and the project manager
16) To direct and oversee functional tests
17) To initiate and direct the final plant inspection and, when the plant is mechanically complete, to pass it on to the commissioning manager or turn it over to the custody of the future plant operator

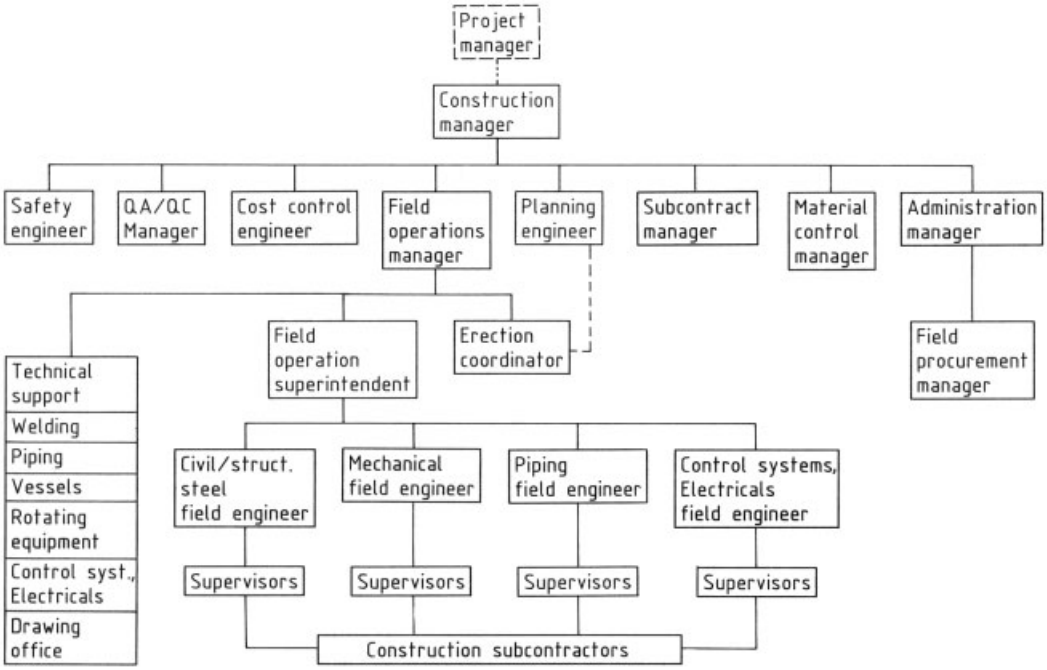

Figure 34. Organization of a construction team

5.6.2.2. Time Scheduling and Progress Control

Time Schedules. An overall construction schedule is created during detail engineering. This plan is continually refined as new information becomes available.

Construction work calls for more detailed, individual schedules for on-site activities, which allow better monitoring of individual jobs and fast response to schedule changes.

Experience shows that detailed construction network diagrams are unwieldy because of the large quantity of data. Individual schedules for use at the construction site are therefore usually prepared as bar charts that also show the interdependencies between the various activities.

Two types of bar charts are usually used:

1) Detail schedules for plant sections, subdivided according to functional disciplines. These charts form the basis for assessing progress of work.
2) Detailed schedules for each functional discipline that include all plant sections. These charts are used for capacity planning for construction personnel and their tools, equipment, materials, and consumables.

Progress of Work. Regular evaluation of progress in construction provides reliable information about the current status of the project. Schedule changes show up and measures can be planned and carried out early enough to insure on-time completion of the plant.

Progress planning is based on a detailed schedule, list of plant components, material takeoffs, and specific rating factors. A specific rating factor is an empirical number of hours required for a specified activity (e.g., hours per tonne, m^3, or piece). If there are no rating factors, the hours required for stated activities are estimated in advance.

The total number of hours thus found for each activity is allotted to the planned execution time for individual activities. Expressing such allotments in percent allows a target progress curve to be determined for each functional discipline in each plant section; this curve serves as a reference for monitoring construction progress.

Progress in each special discipline is evaluated every two to four weeks. Activities not completed at the time of progress assessment must be included. The activities of the special disciplines

are therefore divided into steps and evaluated. A breakdown for above-ground piping installation serves as an example:

Prefabrication:	
Pick up material	2 %
Prepare and tack parts	16 %
Weld	16 %
Fit up and weld small parts	6 %
Installation:	
Transport	5 %
Install, tack	20 %
Weld	11 %
Attach clamps, supports	9 %
Inspection:	
Do preliminary tests, remedy deficiencies	6 %
Pressure test, flush	7 %
Do final test and prepare report	2 %
Total	100 %

Assessments of individual operations result from years of experience. The progress values found for each discipline are then summarized for each plant section, yielding the progress for each plant section or the plant as a whole. These values are compared with the planned target progress values. If there are deviations, the causes are analyzed and appropriate measures taken.

The progress report contains the above information and is an essential part of the regularly updated construction-site report.

5.6.2.3. Construction Work [123]

Work at the construction site begins with preparation of the terrain. The plot must be surveyed, graded, and terraced if necessary. Access roads must be laid out and old structures demolished. Utilities for construction must be brought in. Construction-site offices, storage areas, and workshops must be built. Accomodation (containers) for construction personnel must be provided.

The construction work begins with excavation and foundation work. If the soil quality necessitates driving piles, this must be done first. The sequence of pouring foundations depends on the order in which equipment is to be delivered and installed.

The first step in erection is the erection of heavy equipment (reactors, towers) and steel structures (pipe bridges and equipment supporting frames). Very large process equipment often cannot be shipped in one piece. Tall towers are divided into sections, while large tanks are delivered in the form of prefabricated pieces. The tools needed for assembly (e.g., welding and cutting machines) and facilities for stress-relief annealing of the welds must be provided at the site. If possible, the delivery of heavy units should be scheduled so that they can be placed on their foundations or supported in their frames immediately. Specialists provided by the manufacturer usually assist in the installation of pumps, compressors, and turbines, as well as "package units" such as refrigeration systems and complicated conveyors; these experts later commission the components installed.

Piping installation at a chemical plant is often the most labor intensive and longest phase of installation. It starts with the placement of underground pipelines,the mounting of straight piping on bridges, and the prefabrication of piping. When the number of connection points to apparatus and machinery is great enough (i.e., when the devices have been delivered and put in place), the prefabricated piping sections are connected. Satisfactory progress in prefabrication and smooth installation of piping depends on skillful scheduling of the preparation of isometric drawings and the observance of this schedule. Furthermore, the procurement of piping material should be scheduled so that the material for prefabrication arrives at the site on time. Computer-aided integrated material management systems are a great help in the handling of bulk material.

Weld inspection is carried out by X-ray methods. Which piping is to be inspected depends on the quality assurance specifications.

After a pipe run has been installed, it is pressurized with water to reveal any leaks. The pressure test is documented in a report and the piping is approved for painting or insulation.

Insulation work starts at vessels, towers, and reactors and often requires the construction of complex scaffolds. Pipes should not be insulated until a given plant section has a sufficiently large number of pipe runs that have been approved for insulation.

The installation of electronic devices and control systems takes place after a section of piping has been completed. Devices in control rooms and substations can, however, be installed independent of other work as soon as the buildings have been completed. Underground electri-

cal cables are layed after piping. Electrical and pneumatic cables for measurement and control are installed in cable ducts after the completion of underground work. The laying of cables on cable trays is put off until as late as possible to prevent damage during simultaneous piping installation. The same applies to the installation and junctions of field instruments.

Furnaces are lined with refractories before shipping to the site or on site. Trays are installed in towers after access to the towers has been provided by platforms and ladders. Lightning protection and grounding wires are installed at an early stage, during the fill work of foundation excavations.

Functional tests of the installed equipment mark the end of erection work. These tests are done with the plant in the cold condition and with no product.

The contract must precisely define "mechanical completion," since the contractual obligations of the engineering contractor to the investor often end at this point. Responsibility for the commissioning of the plant may lie with the investor, the licenser, or the engineering contractor. Generally speaking, a plant is mechanically completed when subsequent commissioning will not be delayed or disrupted by installation work and that the safety of the plant is fully guaranteed.

The certificate of mechanical completion is usually accompanied by a "punch list" that defines all installation work that is still outstanding and to be done in the commissioning phase. The certificate is generally granted when the following activities have been performed:

1) Pressure testing of equipment and vessels with air, water, or nitrogen
2) Purging and, if necessary, chemical cleaning and pressure testing of piping
3) Testing of stress-free piping connections to machinery and checking of rotation direction and coupling seating
4) Brief trial run of pumps (with water) and of machinery and motors (as possible without product)
5) Calibration of measuring instruments, alarms, interlocks, and cutoff points
6) Functional checking of electrical equipment and control systems

Figure 35 shows a three-train reformer plant under construction.

5.7. Commissioning

5.7.1. Plant Design and Commissioning

Commissioning must be considered even during basic and detail engineering in the design of equipment, piping, and control systems. Faulty process design can have serious effects on the time required for commissioning and the amount of corrective work needed. The start of production may be significantly delayed and the owner may suffer a substantial loss of production and revenue.

Difficulties in commissioning and causes of delays have been identified as [136]:

26 – 29 %	faulty design
56 – 61 %	failure of plant components
13 – 15 %	errors by operating personnel

Commissioning costs as a percentage of total plant investment are:

5 – 10 %	for established processes
10 – 15 %	for relatively new processes
15 – 20 %	for novel processes

Commissioning should be a special consideration when the piping and instrumentation diagram is designed. A commissioning engineer with relevant experience should be brought in during the planning work. This engineer should prepare the complete operating manual which should be available before the final version of the piping and instrumentation diagram.

The experienced commissioning engineer along with specialists (and maintenance engineers) working for the future plant operator should also be involved in checking the piping model. Errors in pipe routing and poor access for the servicing, installation, and removal of equipment can thus be remedied at an early stage.

5.7.2. Operating Manual

The operating manual is a condensed "reference book" for the entire plant. It should contain all important details about the design and operation. The typical contents of an operating manual for a chemical plant follow [123], [136], [137]:

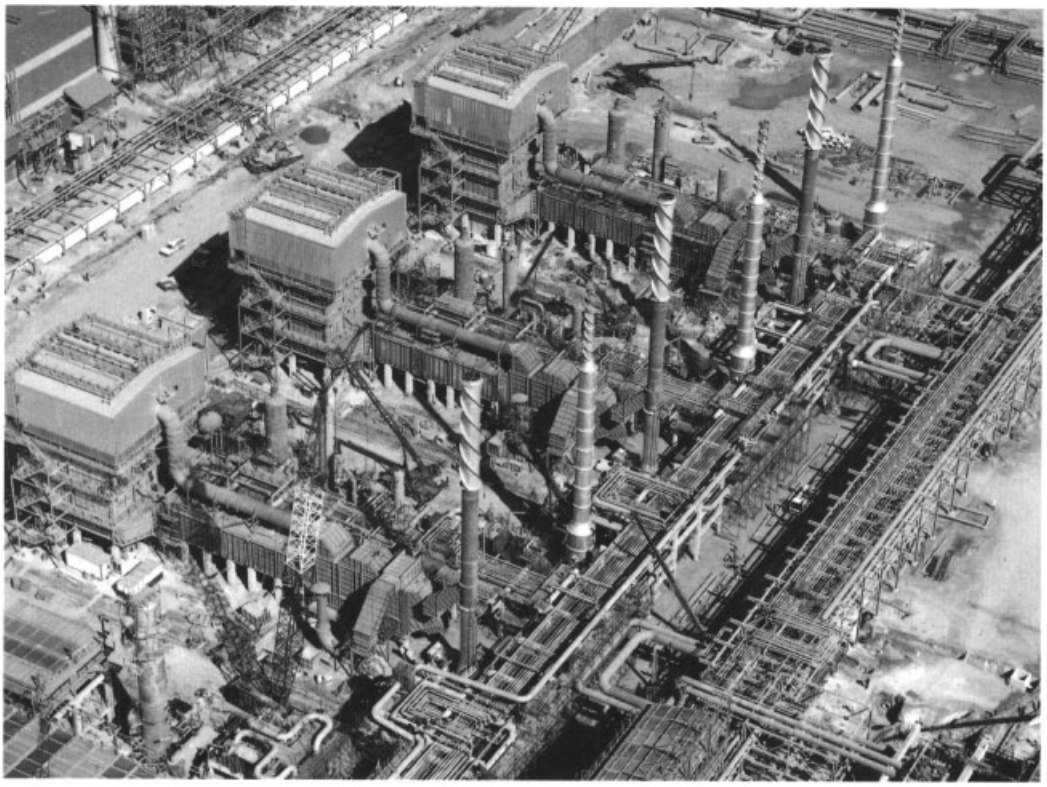

Figure 35. Reformer plant during construction, courtesy of Lurgi AG

Part I. Operating Instructions

1. Design Principles

Statement of the type, purpose, and capacity of the plant; specification of quality and quantities of feedstocks and products (including waste streams); utilities and consumables.

2. Description of the Process and the Plant

2.1. Description of the process with its principles (e.g., chemical and physical principles of the process stages). The process itself is demonstrated by process flow diagrams showing equipment, machinery, and instruments and important process conditions.

Auxiliary systems such as refrigeration, steam, or slop (wastewater) systems are similarly shown. The delivery and disposal of utilities and consumables is also discussed.

2.2. The plant description covers the designs of the individual plant sections, the functions of their components, and the control of the installed unit operations.

2.3. Material balances.

2.4. Process principles and guidelines for plant operation explain the theoretical basis of the process, as well as the process variables and their effects on product quality or composition. Diagrams, formulas, nomographs, and tables allow estimation of these variables (e.g., raw-materials composition, cooling water temperatures).

3. Special Equipment

This chapter contains an in-depth description of special-purpose or critical equipment (e.g., reactors, compressors, turbines). A subchapter deals with particularly important or complex control loops or interlockings and emergency shutdown systems.

4. Preparation of the Plant for Commissioning

This chapter lists the preparatory steps required for commissioning (e.g., flushing, cleaning, and neutralizing of piping, equipment, and plant sections; inspections; pressure and leak tests; inspection of safety devices; mechanical tests of machinery; drying

of furnace refractory or reactor linings; and specifications for charging catalysts and consumables).

5. Plant Startup

An outline of the overall plan for starting up the plant is first given. All startup operations are then described in detail step-by-step. Special precautions and unusual design conditions are highlighted.

The startup instructions are broken down as follows:

a) Initial startup after installation is complete.
b) Startup after a prolonged shutdown.
c) Restart after a brief shutdown when the plant is still warm.
d) Procedures for catalyst regeneration or replacement.
e) Measures to be taken after abnormal occurrences. Possible disturbances are listed together with their effects and countermeasures. This section discusses how to remedy problems during operation; how to keep plant sections in operation while problems are being remedied; and how to perform restart afterward.

6. Plant Shutdown

This chapter describes procedures for planned and unplanned shutdown of the plant. It is subdivided as follows:

a) Partial shutdown for periodic catalyst regeneration or removal of cracking deposits from furnace tubes
b) Procedure for brief shutdown
c) Procedure for extended shutdown
d) Shutdown on utilities outage
e) Emergency shutdown and special precautions

7. Analysis Specifications

Analysis specifications: required or recommended number of analyses during commissioning under steady-state operating conditions and exceptional operating conditions.

8. Operating Report

This chapter describes which data are to be recorded during plant operation. A standard form for reporting during steady-state operation or commissioning is recommended.

9. Safety Practices

The safety regulations are summarized. Potential hazards are discussed, and the behavior of the operating personnel is recommended or prescribed. Information about safety facilities and the locations of first-aid stations is given.

10. Miscellaneous

List of blinds, setpoints for alarm and switching functions.

Part II. Drawings and Equipment Specifications

1. Drawings

This chapter contains drawings that relate to the process or to the plant as a whole:

a) Process flow diagrams and P & I diagram
b) Plant layout plan
c) Underground plan
d) Selected overall drawings of important equipment (e.g., reactors)

2. Specifications

Equipment design specifications are compiled. Relevant drawing numbers, technical procurement specifications, and other documents containing supplementary information are also noted.

3. Equipment Manufacturers Operating Instructions

Part III. Technical Documentation and Drawings

Manuals are prepared for each engineering discipline. These manuals show all specifications and drawings relating to plant components and their operation and maintenance.

5.7.3. Responsibility and Organization

Responsibility for commissioning generally lies with the party granting the process license: the investor/owner, licenser, or the engineering firm. The commissioning team is led by the commissioning manager. The key positions are occupied by experienced startup engineers. Startup operation goes on around the clock, so that an adequate number of startup engineers must be available for shift work.

In large plants involving two or more process steps, it is desirable to break the plant down into sections and assign responsibility for each to a smaller startup team.

On mechanical completion of the plant most of the installation personnel leave. Some specialist engineers remain on site, however, especially

those involved with piping, electricals, control systems, and machinery, who solve problems that arise during commissioning.

The actual manual commissioning activities are generally performed by personnel of the operator who have already gone through classroom training. The operating personnel can often be trained in similar plants operated by associated firms.

5.7.4. Preparation for Commissioning

The commissioning manager must ensure that the plant is supplied with the necessary quantities and qualities of feedstocks, utilities, consumables, and energy in time for the planned start of commissioning. Spare parts and a fully equipped repair shop must also be available. The cooperation of specialists provided by the manufacturers is essential for the commissioning of complicated equipment (e.g., compressors, refrigeration plants).

The plant laboratory has a vital function during commissioning. Analytical data are an important input to process control. Sampling and analysis programs must be prepared and discussed with the laboratory.

The commissioning team should be present on site during the final installation phase. While the installation team performs function tests, the commissioning engineers perform a detailed check of the plant, focusing on process design and operation (e.g., inspection of towers, internals, and control systems). Changes requested by the commissioning team can then be carried out by the installation team.

When the commissioning manager is convinced that the plant is ready for operation, he takes over responsibility for further activities from the erection management.

5.7.5. Plant Startup

The measures described in Sections 5.7.3 and 5.7.4 apply to chemical plants in general, whereas activities during the initial startup of a plant depend on the type of process.

Commissioning takes place step by step as specified in the operating manual. The first units to be started are utilities and off-sites (e.g., cooling water loop, steam generation). At the same time, lined furnaces are dried and heated in accordance with the vendor's specifications. The catalyst is charged, reduced when necessary, and brought up to reaction temperature. Plants in which combustible media circulate must be purged with inert gas so that they are oxygen-free before charging. Steam lines must be carefully dewatered.

All measurements are recorded and balances are run so that incorrect behavior of the plant can be quickly detected and corrective measures instituted.

Initial disorders are almost always encountered: these result, for example, from utilities outages, mechanical damage, and hot running of bearings or stuffing boxes. An attempt should be made, however, to get the plant running first and start up all systems, provided the safety of personnel and equipment is not endangered. The defects can then be remedied during the first scheduled shutdown of the plant.

After operation has stabilized, conditions are optimized. When the planned values of product quantity and quality, utilities consumption etc., have been attained, the guarantee test is carried out. Guarantee values and conditions for performing the guarantee test are stipulated in the contract.

If the test results are satisfactory, a report for handover of the plant to the owner is signed. Responsibility for the plant and its operation now shifts to the owner. Any defects still to be remedied are entered in a punch list and a deadline for corrective action is established. Generally the service life of plant equipment is guaranteed for a further, contractually agreed period (parts subject to wear are usually exempt from this guarantee). Whether this guarantee is the responsibility of the manufacturers or the engineering firm depends on the contract. Figure 36 shows part of a plant complex for olefin production.

6. Computer Support

Most engineering contractors have invested heavily in computerization, with the emphasis on computer-aided design (CAD), computer-aided engineering (CAE), design calculations, data-base management,and office communication sys-tems [138]. Decisions on the use of such systems, in particular CAD and CAE, are

Figure 36. Section of an olefin complex, courtesy of Lurgi AG

driven by benefits, chiefly reductions in costs and turnaround times, gains in the transparency of methods used, and systematic support for project procedures.

6.1. Role of Computers in Project Execution

A variety of systems based on discrete and closed mathematical models are employed in process engineering for the simulation and design of processes. They can be accessed from mainframe computers, workstations, or personal computers (PCs).

For special processes, firms use internally developed programs and modules based on standard PC software. Internationally recognized simulation programs are employed to prepare mass and energy balances and to optimize heat-exchanger systems. Powerful systems are also available for computer-aided process analysis in large plants, dynamic simulation, and plant optimization.

Project management uses computer software with a high degree of integration for scheduling, cost planning, and project control. These systems are generally accessible on both mainframes and PCs.

The specific applications of computer systems vary. Graphical documents (e.g., process flow diagrams, P & I diagrams, loop diagrams) are prepared with CAD systems which are being increasingly linked to engineering data bases. Two-dimensional design instruments are in widespread use for site layout planning and plant de-sign; three-dimensional design systems are occasionally used in special piping-intensive projects [139].

The advantages of graphics in plant design are the consistent and systematic use of models and overlays and the reuse and evaluation of graphical elements with variable intelligence [140].

CAD use improves collaboration between individual disciplines. For example, if plant and civil engineering both use the same CAD system, plant design can be optimized "at the source."

Engineering calculations can be performed with PC programs developed in-house or with internationally recognized standard software products.

Standard systems are predominant in direct daily use by engineers. In subfields, such as finite-elements design, tasks are delegated to specialists who can perform optimizations with special computer tools.

Every step in the procurement of equipment is computer-aided. All data and functions are integrated into a system that performs bid invitation, bid comparison, ordering, expediting, and shipping. Relevant data are available not only to the cost control system but also to accounting. As a rule, the entire accounting process is also computer-aided.

Specialized relational data-base systems are widely used in the procurement of plant components and bulk materials. These systems are linked by interfaces.

Administrative functions are supported by integrated office communications systems. Clerical staff work with PC support connected to laser printers.

6.2. EDP Infrastructure and Systems

Progressive engineering companies strive for complete system integration so that engineering, commercial, and management data can be utilized in interlinked modules. In addition to mainframe computers, PCs are increasingly used for individual support at engineering workstations.

The growing demand for "distributed intelligence" and advances in computer capabilities are now leading to the use of interconnected, decentralized workstations with alphanumeric applications (data bases, calculations) and graphical ones (CAD). Extensive standardization of the data-processing infrastructure is desirable.

The variety of software systems used in project execution are illustrated by the following examples:

Design Calculations. Many software systems are developed in-house for the process design and engineering of plant equipment. The following general-purpose programs are also commercially available:

PROCESS, ASPEN PLUS	simulation programs, flow sheeting
SDC	material data compiler
HTRI	design of heat exchangers
ROHR2	strength calculations for piping
PROBAD/FEZEN	strength calculations for apparatus
ANSYS, STRUDL	finite-element method (FEM) programs for structural analysis and material and heat flow

Data-Base Management Systems. The following list includes systems developed by Lurgi from relational data-base systems:

ANSY	mechanical equipment
LUPREA	electrical equipment
MASY	control systems
MVS/LUROMAK	piping
DISPO	disposition of bulk materials
MOSY	management of materials on-site
BISAM	bid invitation and ordering
VERONA	shipping
ATERM	expediting
KAPAZ	capacity planning
KOKO	cost control
DOSY	documentation and archiving
REPRÜ	accounting

Graphics. Standard systems for graphics applications include:

CADAM/AEC	two- and three-dimensional plant layout and design
INTERGRAPH/IGDS, CADEX	preparation of P & I diagrams
AUTOCAD	preparation of P & I diagrams in special fields

6.3. Coordination and Interfaces

Growing importance is now attached to the fast, error-free transfer of information generated in computer-aided operations of the engineering firm during project execution, to the owner and to engineering partners (and vice versa). Accordingly, the partners must arrive at a good understanding as to the content of documents and data files to be transmitted; standards and codes (data structures, nomenclature, symbols); and data formats and EDP procedures.

The definition of interfaces between two different EDP systems is an important factor in data transfer. Standardization of interfaces to a reliable extent does not yet exist for plant design and construction. Various engineering companies have devised interfaces for two-dimensional plant layout and design so that drawing data can be flexibly transmitted to any computer system.

7. Quality Assurance

The primary goal in the design and construction of chemical plants is to satisfy quality requirements. These are defined by agreement between the investor and the engineering contractor, by legal regulations, and by objectives set by the engineering firm. The quality requirements are generally specified in the contract between the investor and the engineering firm.

Quality assurance is ensured by installation of relevant systems. These quality assurance systems cover all technical and organizational practices needed to achieve the desired quality.

Requirements for quality assurance systems are defined in standards. The international standards ISO 9000 – 9004 have already been incorporated into most national standards systems [141]. Increasing numbers of quality assurance systems have been developed, introduced, and documented in production and service companies. Certification of these quality assurance systems by neutral organizations is in development.

The idea behind the creation of quality assurance systems is that the quality of a product should not only be established after its production, but rather that the entire production process should be subjected to appropriate quality assurance practices on a phase-by-phase basis. Quality assurance practices must be defined and implemented for all services performed by the engineering firm itself (e.g., project management, engineering, procurement, supervision of construction, commissioning).

The company quality assurance system is usually documented in a "quality assurance manual" which contains information on organizational structure and processes, as well as the procedures, means, and methods used to assure quality. It may also include references to internal procedures and work instructions which are not part of the manual.

The manual gives the owner and third parties a summary of the company's quality policy and quality system. It is also an instrument for communicating the company's quality policy to management and employees.

A typical table of contents follows:

1) Quality policy of the company
2) Brief description of the company
3) Elements of quality assurance:
   - Management tasks
   - Company quality assurance system
   - Marketing
   - Research and development, engineering, project management
   - Documentation
   - Procurement
   - Fabrication, civil work, erection, commissioning
   - Measuring and test equipment, inspection status, corrective action
   - Quality records
   - Internal quality audits
   - Education and training

The quality assurance system must not be regarded as a fixture that schematizes all procedures – it needs to be continuously improved in the light of practical knowledge.

An engineering firm does not usually have its own fabrication capacity or perform construction services. Equipment as well as construction and installation services must therefore be procured. Suppliers of equipment and firms performing construction work must demonstrate their own quality assurance systems to the engineering firm and allow them to be verified. In addition to this, employees of the engineering firm monitor the fabrication of equipment and the construction work in accordance with established rues to ensure that delivery and performance are in accordance with quality requirements and planned schedules. Such supervision does not, however, release a manufacturer or a construction contractor from its contract obligations.

In the normal case, a project can be executed and meet the quality requirements if the provisions of the quality system are satisfied. In complicated projects or those involving a high degree of risk or innovation, a quality assurance plan must be drawn up [142]. This sets forth in detail the quality practices to be followed during execution of the project. A quality assurance manager is designated for the project who, after consultations with the project manager and the heads of the functional divisions, directs the quality assurance activities. He is independent of the project team, reports directly to company management, and confers with the investor on all questions of quality assurance.

Many quality assurance practices are involved in the execution of a project, a few examples follow:

1) Checking contracts against checklists
2) Defining the degree of checking of technical specifications and drawings for equipment
3) Implementing design change control
4) Instituting design reviews to check, for example, process flow diagrams, layout plans, piping and instrumentation diagrams, and piping models
5) Selecting competent manufacturers for critical equipment
6) Identifying "hold points" and intensity of inspection for equipment during fabrication

Quality obtained on the basis of an assurance system "tailored" to the company results in several important benefits:

1) After consultation of quality assurance documents, the customer of a production or service company can convince himself of the company's ability to achieve the agreed-on quality of the product or service
2) A company's quality assurance systems, examined and certified by a competent neutral organization, can mean a competitive advantage
3) If a quality assurance system is organized in a meaningful and expedient manner, agreed-on quality of a product can be obtained at low cost and with little expenditure of time

Quality assurance practices must already be applied, for example, during the design stage for a piece of plant equipment. Expensive post-fabrication corrections on a wrongly specified device are then avoided.

8. Training of Plant Personnel

Preliminary Planning. The people who are to operate and maintain a chemical plant must have the necessary theoretical background, practical training, and know-how.

This applies in particular to personnel in developing countries, who should participate in specially developed know-how transfer programs.

Many owners and operators of chemical plants write into their contracts with engineering firms the transfer of operating and maintenance knowledge to their specialists, supervisors, engineers, and technicians. Training covers technical and commercial jobs as well as middle and upper management.

The following questions should be explored at the feasibility study stage:

1) What level of education exists in the region?
2) Will skilled labor be available in the region?
3) Is the project a newly-built plant or an expansion of an existing facility?
4) If it is an expansion, can skilled operating personnel be found?
5) Is the plant a labor-intensive production facility or one that can run automatically?
6) Can the plant operate autonomously (e.g., in a virgin forest area or on an island) or does it require an industrial infrastructure?

Training Plan. The training plan must be adapted to the needs of the plant and its surroundings and should be workable regardless of the initial qualifications of the workers being trained. It should also convey state-of-the-art knowledge.

The training plan comprises organizational charts, job descriptions, definition of minimum qualifications of future jobholders, training schedules, and identification of facilities for practical training. It may also include administrative topics (e.g., how to obtain a visa and residence and work permits, assistance in finding accomo-dation, where to get work and safety clothing, personal insurance and medical care during the training period).

The planning documents provide a basis for hiring tests and results are compared with the profile of requirements in the job descriptions. The tests should pose and evaluate technical and management skills questions. If necessary, an institute or an industrial psychologist can be brought in.

Execution of Training. Every participant should receive a training schedule in which the subject, day/time, and discussion partner or instructor are listed.

Training should take place as late as possible so that there is no lag between training and job assignment (fluctuation danger). However, training activities should be started early enough so

that the future operator's personnel can see their own plant demonstrated in the final installation phase and can perform some functions themselves. This improves their sense of responsibility.

The training program is usually broken down as follows:

Phase I.	Presentation of basic information about the plant.
Phase II.	Presentation of generally important instructions on plant operation, maintenance of machinery, safety practices, and organization.
Phase III.	The trainees are divided into operating personnel, maintenance personnel, and administrative and management personnel. When possible, these groups are trained on the same or similar facilities for the jobs they will later perform.
Phase IV.	During the final phase of installation, the plant personnel familiarize themselves in depth with their own plant. The training period ends with active participation in plant commissioning.

Training Costs. Training costs comprise personnel, nonpersonnel, and incidental costs.

Personnel costs are incurred for the people who prepare, execute, and coordinate training, as well as the trainees salaries. *Nonpersonnel costs* include payments to the operators of facilities where training takes place and the outfitting of training rooms on the construction site. *Incidental costs* comprise costs for accomodation, travel, work clothing, insurance, and utilities. These may make up a significant fraction of total costs if the trainees are sent abroad.

9. References

1. H. Popper: *Modern Cost Engineering Techniques,* McGraw-Hill, New York 1970.
2. H. Kölbel, J. Schulze: *Der Absatz in der chemischen Industrie,* Springer-Verlag, Berlin 1982.
3. J. T. Thorngren: "Probability Technique Improves Investment Analysis," *Chem. Eng. Cost File,* vol. 9, McGraw-Hill, New York 1967.
4. D. J. Massey, I. H. Black: "Predicting Chemical Prices," *Chem. Eng. Cost File,* vol. 11, McGraw-Hill, New York 1969.
5. J. Jung: "Globale Vorausberechnung von Investitionskosten," in *GVT-Hochschulkurs: Angewandte Kosten- und Wirtschaftlichkeitsberechnungen bei der Projektierung verfahrenstechnischer Anlagen,* University of Dortmund 1984, 90 – 114.
6. V. D. Herbert, A. Bisio: "The Risk and the Benefit," part 2, *CHEMTECH* **6** (1976) no. 7, 422 – 429.
7. H. Kölbel, J. Schulze: *Projektierung und Vorkalkulation in der Chemischen Industrie,* Springer-Verlag, Berlin 1982.
8. J. E. Haselbarth: "Updated Investment Cost for 60 Types of Chemical Plants," *Chem. Eng. (N. Y.)* **74** (1967) Dec. 4, 214.
9. K. M. Guthrie: "Capital and Operating Cost for 54 Chemical Processes," *Chem. Eng. (N. Y.)* **77** (1970) June 15, 140 – 156.
10. J. Jung: "Detaillierte Vorausberechnung von Investitionskosten," in *GVT-Hochschulkurs: Angewandte Kosten- und Wirtschaftlichkeitsberechnungen bei der Projektierung verfahrenstechnischer Anlagen,* University of Dortmund 1984, pp. 115 – 137.
11. F. Strailk: "Das Verfahrensschema als Grundlage bei der Planung in der chemischen Technik," *Chem. Appar.* **92** (1968) 419 – 434.
12. H. C. Lang, *Cost Engineering in the Process Industries,* McGraw-Hill, New York, 1960, pp. 7 – 14.
13. C. H. Chilton, *Chem. Eng. (N. Y.)* **73** (1966) 184 – 190; C. H. Chilton: *Cost Engineering in the Process Industries,* McGraw-Hill, New York 1960.
14. W. E. Hand, "From Flowsheet to Cost Estimate," *Pet. Refin.* **37** (1958) 331.
15. W. Burgert "Kostenschätzung mit Hilfe von Kostenstrukturanalysen," *Chem-Ing.-Techn.* **51** (1979) 484 –487.
16. C. A. Miller: "Factor Estimating Refined for the Appropriation of Funds," *Chem. Eng. Cost File,* vol. 7, McGraw-Hill, New York 1965.
17. K. M. Guthrie: "Date Techniques for Preliminary Capital Cost Estimating," *Chem. Eng. (N. Y.)* **76** (1969) Jan. 13, 138;**76** (1969) March 24, 114 –142;**76** (1969) April 14, 201 – 216. *Chem. Eng. Cost File,* vol. 11, McGraw-Hill, New York 1969.
18. *Chemical Engineering,* McGraw-Hill, New York –Düsseldorf, published monthly.
19. *Chemische Industrie,* Verlag Handelsblatt, Frankfurt am Main (published monthly).
20. P. M. Kohn: "CE Cost Indexes Maintain 13-Year Ascent," in: *Modern Cost Engineering: Methods and Data,* McGraw-Hill, New York 1979.
21. J. Schulze: "Modernisierter Preisindex für Chemieanlagen," *Chem. Ind. (Düsseldorf)* **32** (1980) no. 10, 657 – 663.

22. J. Jung: "Globale Vorausberechnung von Betriebskosten," in *GVT-Hochschulkurs: Angewandte Kosten- und Wirtschaftlichkeitsberechnungen bei der Projektierung verfahrenstechnischer Anlagen,* University of Dortmund 1984, pp. 90 – 114.
23. J. T. Sommerfeld, C. T. Lenk, *Chem. Eng.* **77** (1970) May 4, 136 – 138.
24. H. Gaensslen, *Chem. Ing. Tech.* **48** (1976) no. 12, 1193 – 1195.
25. F. C. Vilbrandt, C. E. Dryden: *Chemical Engineering Plant Design,* McGraw-Hill, New York 1959.
26. ICI "Factest," *Eur. Chem. News* **21** (1972) March 17, 32.
27. ASPEN: TECHNOLOGY ASPEN plus Costing Manual, Cambridge 1984.
28. M. S. Peters, K. D. Timmenhaus: *Plant Design and Economics for Chemical Engineers,* McGraw-Hill, New York 1980.
29. Verband der chemischen Industrie: *Kostenrechnung in der Chemischen Industrie,* Gabler, Wiesbaden 1962.
30. R. I. Reul: "Which Investment Appraisal Technique Should You Use?" *Chem. Eng. Cost File,* **vol. 10,** McGraw-Hill, New York 1968. W. R. Park: "The Investment Profit-Prophet," *Chem. Eng. Cost File,* **vol. 10,** McGraw-Hill, New York 1968.
31. D. Spethmann: "Flexibilitätsüberlegungen bei Investitionsprogrammen," *Vortrag auf dem 22. Deutschen Betriebswirtschaftler Tag,* Deutsche Gesellschaft für Betriebswirtschaft, Berlin 1971.
32. H. Jonas: *Investitionsrechnung,* De Gruyter, Berlin 1964.
33. J. Leibson, C. A. Trischmann: "When and How to Apply Discounted Cash Flow and Present Worth," *Chem. Eng. (N. Y.)* **78** (1971) Dec. 13, 97 – 106.
34. A. Fletscher, G. Clarke: *Mathematische Hilfsmittel der Unternehmensführung,* Verlag W. Dummer & Co., München 1966.
35. H. Müller-Merbach: "Operations Research/Rev. 1," *Methoden und Modelle der Optimalplanung,* 3rd ed., München 1973.
36. H. F. Rase, M. H. Barrow: *Project Engineering of Process Plants,* J. Wiley, New York 1957.
37. B. Aggteleky: *Fabrikplanung,* Hanser Verlag, München 1987, 1990.
38. E. Mach: *Planung und Errichtung chemischer Fabriken,* Sauerländer, Frankfurt 1971.
39. A. Franke: "Risk Analysis in Project Management," *Int. J. Project Management* **5** (1987) no. 1, 29 – 34.
40. A. Jaafari: "Management Know-how for Project Feasibility Studies," *Int. J. Project Management* **8** (1990) no. 3, 167 – 172.
41. W. T. Nichols, *Ind. Eng. Chem.* **43** (1951) no. 19, 2295.
42. American Association of Cost Engineers: *A Guide to Capital Estimating* (1988) 14.
43. R. J. Loring: "Cost of Preparing Proposals," *Chem. Eng. (N. Y.)* **77** (1970) Nov. 16, 126.
44. P. Grassmann: *Physikalische Grundlagen der Chemie-Ingenieur-Technik,* Sauerländer, Frankfurt 1983.
45. E. U. Schlünder: *Heat Exchanger Design Handbook,* Hemisphere Publ. Corp., Washington 1983.
46. J. H. Perry: *Chemical Engineers Handbook,* McGraw-Hill, New York 1984.
47. McGraw-Hill, New York: L. Clarke, R. L. Davidson: *Manual for Process Engineering Calculations,* 1962; E. W. Comings: *High Pressure Technology,* 1956; D. L. Katz: *Handbook of Natural Gas Engineering,* 1959; A. L. Kohl, F. C. Riesenfeld: *Gas Purification,* 1960; W. L. Nelson: *Petroleum Refinery Engineering,* 1958; M. van Winkle: *Distillation,* 1967. H. Harnisch, R. Steiner, K. Winnacker: *Chemische Technologie,* C. Hanser Verlag, München 1984.
48. Gulf Publ. Co., Houston: R. N. Watkins: *Petroleum Refinery Destillation,* 1979; E. E. Ludwig: *Applied Process Design for Chemical and Petrochemical Plants,* 1964.
49. R. H. Perry: *Perry's Chemical Engineers' Handbook,* McGraw-Hill, New York 1984.
50. K. Weissermel, H.-J. Arpe: *Industrial Organic Chemistry,* VCH Verlagsgesellschaft, Weinheim 1993.
51. L. B. Evans: *Foundations of Computer-Aided Chemical Process Design,* **vol. 1,** Engineering Foundation, New York 1981, 425 – 469.
52. R. Raman: *Chemical Process Computations,* Elsevier Appl. Sci. Publ., London 1985.
53. DECHEMA, DETHERM-SDC (Stoffdaten-Compiler), Frankfurt, updated regularly. Design Institute of Physical Property Data (DIPR), American Institute of Chemical Engineers (AIChE), DIPPR Data Compilation, uptated regularly. Thermodynamics Research Center (TRC), Vapor Pressure Datafile, updated regularly.
54. Landolt-Börnstein: *Zahlenwerte und Funktionen aus Naturwissenschaft und Technik,* Springer-Verlag, Berlin 1985.
55. D'Ans-Lax: *Taschenbuch für Chemiker und Physiker,* Springer-Verlag, Berlin 1970.

56. D. Behrens, R. Eckermann: *Chemistry Data Series,* DECHEMA, Frankfurt 1977 ff.
57. M. Munsch, T. Mohr, E. Futterer, *Chem. Ing. Tech.* **62** (1990) 995 – 1002.
58. D. W. Townsend, B. Linnhoff, *AIChE J.* **29** (1983) 742 – 771.
59. M. A. Duran, I. E. Grossmann, *AIChE J.* **32** (1986) 123 – 138.
60. B. Linnhoff: "New Concepts in Thermodynamics for Better Chemical Process Design," *Chem. Eng. Res. Des.* **61** (1983) 207 – 223.
61. B. Linnhoff: "Wärme-Integration und Prozeßoptimierung," *Chem. Ing. Tech.* **59** (1987) no. 11, 851 –857.
62. A. W. Westerberg, *Comput. Chem. Eng.* **13** (1989) 365 – 376.
63. V. Hlavacek, *Comput. Chem. Eng.* **2** (1978) 67 – 75.
64. N. Nishida, G. Stephanopoulos, A. W. Westerberg, *ALChE J.* **27** (1981) 321 – 351.
65. A. W. Westerberg, *Comput. Chem. Eng.* **9** (1985) 421.
66. G. Hosemann: "Methodisches Grundkonzept technischer Sicherheit," DIN-Mitteilung 70, no. 3, Beuth Verlag, Berlin 1991.
67. E. Tröster: "Sicherheitsbetrachtungen bei der Planung von Chemieanlagen," *Chem. Ing. Tech.* **57** (1985) no. 1, 15 – 19.
68. Bundes-Immissionsschutzgesetz vom May 14, 1990 (BGBl I, 880)
69. Zwölfte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Störfallverordnung) vom September 20, 1991 (BGBl I 1991, 1891)
70. Gewerbeordnung, Jan. 1, 1987 (BGBl. I, 1987, 925). Amended by Gesetz zur Änderung der Gewerbeordnung, April 5, 1990 (BGBl. I, 1990, 706) and Gesetz zur Änderung der Gewerbeordnung Nov. 9, 1990 (BGBl. I, 1990, 2442)
71. Bundesnaturschutzgesetz (Mar. 12, 1987) Bundesgesetzblatt (BGBl) I (1987), 889.
72. G. Hohe: "Sicherheitsabstände – Definition und Bedeutung in Rechtsvorschriften und Technischen Regelwerken," *Berichtsband zum BPU-Seminar: "Sicherheitsabstände bei Planung und Betrieb industrieller Anlagen,"* BPU GmbH, Berlin 1990.
73. Gesetz zur Umsetzung der Richtlinie des Rates, June 27, 1985 Concerning environmental compatibility testing in public and private projects (85/337/EWG). (BGBl. I, 1990, 205).
74. Gesetz über die Umwelthaftung, Bundesgesetzblatt (BGBl) F (1990), 2634.
75. Abstände zwischen Industrie- bzw. Gewerbegebieten und Wohngebieten im Rahmen der Bauleitplanung (Abstandserlaß), Minister für Umwelt, Raumordnung und Landwirtschaft, March 21 1980 (MBl.NW, 504 / SM Bl. NW, 283).
76. Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft – TA Luft), Feb. 27, 1986.
77. DIN 52 900,Feb. 1983, DIN Safety Data Sheet for Chemical Substances and Preparations, Beuth Verlag, Berlin 1983.
78. T. Redeker, W. Möller: "Chemsafe-Datenbank für sicherheitstechnische Kenngrößen," *TU* **29** (1988) no. 5, 174, VDI Verlag, Düsseldorf.
79. V. Pilz: "Sicherheitsanalysen zur systematischen Überprüfung von Verfahren und Anlagen – Methoden, Nutzen und Grenzen," *Chem. Ing. Tech.* **57** (1985) no. 4, 289 – 307.
80. *A Guide to Hazard and Operability Studies,* Chemical Industry Safety and Health Council of the Chemical Industries Association, Alensbie House, 93 Albert Embankment, London SE 1 7TU. Reprinted 1985.
81. Risikobegrenzung in der Chemie, PAAG-Verfahren (HAZOP), ISSA Prevention Series No. 2002, May 1990. Published by: IVSS Internationale Vereinigung für soziale Sicherheit, Berufsgenossenschaft der chemischen Industrie, Geisbergstraße 11, D-6900 Heidelberg.
82. Wasserhaushaltsgesetz Sept. 23, 1986 (BGBl. I, 1986, 1529)
83. Allgemeine Rahmen-Verwaltungsvorschrift über Mindestanforderungen an das Einleiten von Abwasser in Gewässer (Rahmen Abwasser VwV) vom Sept. 8, 1989 (GMBI, 1989, 518).
84. H.-O. Braubach, E. Sommer, K. Zellmann: "Anforderungen an verfahrenstechnische Anlagen durch die Weiterentwicklung der Gewässerschutz-Vorschriften," *Chem. Ing. Tech.* **57** (1985) no. 3, 113 –118.
85. Allgemeine Verwaltungsvorschrift über die nähere Bestimmung wassergefährdender Stoffe und ihre Einstufung entsprechend ihrer Gefährlichkeit (VwV wassergefährdende Stoffe), Mar. 9, 1990 (GMBI, 1990, 114).
86. Gesetz über die Vermeidung und Entsorgung von Abfällen (Abfallgesetz – AbfG), Aug. 27, 1986 (BGBl I, 1986, 1410).

87. Verordnung zur Bestimmung von Abfällen nach § 2 Abs. 2 des Abfallgesetzes (Abfallsbestimmung-Verordnung – AbfBestV) April 3, 1990 (BGBl. I, 1990, 614).
88. Allgemeine Abfallverwaltungsvorschrift über Anforderungen zum Schutz des Grundwassers bei der Lagerung und Ablagerung von Abfällen, Jan. 31, 1990 (GMBI, 1990, 74).
89. MAGS-Lärmschutz beim Raffineriebau, Minister für Arbeit, Gesundheit und Soziales des Landes Nordrhein-Westfalen, 1977.
90. Forschungsbericht 79-105-03-302, Stand der Technik bei der Lärmminderung in der Petrochemie, commissioned by the Bundesumweltamtes, Dec. 1979.
91. H. M. Bohuy: *Lärmschutz in der Praxis,* R. Oldenbourg Verlag, München 1986.
92. M. Heckl, H. A. Müller: *Taschenbuch der Technischen Akustik,* Springer-Verlag, Berlin 1975.
93. Allgemeine Verwaltungsvorschrift über genehmigungsbedürftige Anlagen nach § 16 der Gewerbeordnung GewO, Technische Anleitung zum Schutz gegen Lärm (TA Lärm), July 16, 1968 (Beilage zum BAnz Nr. 137, July 26, 1968)
94. Verordnung über Arbeitsstätten, March 20, 1975 (BGBl. I, 729). Modified by decrees passed on Jan. 2, 1982 (BGBl. I, 1) and Aug. 1, 1983 (BGBl. I, 1057)
95. Unfallverhütungsvorschrift Lärm (UVV-Lärm) Oct. 10, 1990, Hauptverband der gewerblichen Berufsgenossenschaft.
96. Gesetz über Betriebsärzte, Sicherheitsingenieure und andere Fachkräfte für Arbeitssicherheit, Dec. 12, 1973 (BGBl. I, 965). Modified by law passed on April 12, 1976 (BGBl. I, 965)
97. Verordnung über Druckbehälter, Druckgasbehälter und Füllanlagen (Druckbehälterverordnung), Feb. 27, 1980 (BGBl. I, 173). Modified by decree passed on April 21, 1989 (BGBl. I, 830)
98. M. Pütz: *Die Genehmigungsverfahren nach dem Bundes-Immissionsschutzgesetz. Handbuch für Antragsteller und Genehmigungsbehörden,* Erich Schmidt, Berlin 1986.
99. Neunte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Grundsätze des Genehmigungsverfahrens – 9. BImSchV), Feb. 18, 1977 (BGBl. I, 274) and May 19, 1988 (BGBl. I, 608, 623)
100. Berücksichtigung von Emissionen und Immissionen bei der Bauleitplanung sowie bei der Genehmigung von Vorhaben (Planungserlaß), Minister für Landes- und Stadtentwicklung, Minister für Arbeit, Gesundheit und Soziales und Minister für Wirtschaft, Mittelstand und Verkehr, July 8, 1982 (MBl. NW, 1366 /SM Bl. NW, 2311).
101. Die Beteiligung an der Bauleitplanung (Beteiligungserlaß), Minister für Landes- und Stadtentwicklung, July 16, 1982 (MBl. NW, 1375 / SM Bl. NW, 2311).
102. A. Rahmel, W. Schenk: *Korrosion und Korrosionsschutz von Stählen,* Verlag Chemie, Weinheim 1977.
103. U. Heubner et al.: *Nickellegierungen und hochlegierte Sonderstähle; Kontakt und Studium,* vol. 153, Expert-Verlag, Sindelfingen 1985.
104. H. Gräfen et al.: *Die Praxis des Korrosionsschutzes; Kontakt und Studium,* vol. 64, Expert-Verlag, Grafenau 1981.
105. U. R. Evans: *Einführung in die Korrosion der Metalle,* Verlag Chemie, Weinheim 1965.
106. *DIN Taschenbuch 219: Korrosion und Korrosionsschutz,* Beuth-Verlag, Berlin 1987.
107. H. Gräfen et al.: *Kleine Stahlkunde für den Chemieapparatebau,* VDI Verlag, Düsseldorf 1978.
108. VDI-Berichte 385: Schadensverhütung durch Qualitätssicherung von Werkstoffen und Bauteile-Optimierung des Bauteilhaltbarkeit. Tagung, München 1980, VDI-Verlag, Düsseldorf.
109. D. Behrens: *DECHEMA Corrosion Handbook,* VCH Verlagsgesellschaft mbH, Weinheim 1988 ff.
110. J. R. Brauweiler: "Economics of Longlife – Shortlife Material," *Chem. Eng. (N. Y.)* **70** (1963) no. 2, 128.
111. R. E. Meissner, D. C. Shelton, *Chem. Eng. (N.Y.),* April 1992, 81 – 85, 89 – 94.
112. R. S. Aries, R. D. Newton: *Chem. Eng. Cost Estimation,* New York (1955) 2, 77.
113. J. T. Gallagher: "A Fresh Look at Engineering Construction Contracts," *Chem. Eng. Cost File,* vol. 9, McGraw-Hill, New York 1967, p. 23.
114. J. in 't Veld, W. A. Peters: "Keeping large Projects under Control: The Importance of Control Type Selections," *Int. J. Project Management* **7** (1989) no. 3.
115. A. Morna, N. J. Smith: "Project Managers and the Use of Turnkey Contracts," *Int. J. Project Management* **8** (1990) no. 3, 183 – 189.
116. J. Madauss: *Projektmanagement,* 3rd ed., C. E. Poeschel Verlag, Stuttgart 1989.

117. E. Mosberger: "Projektmanagement im Anlagenbau," *Chem. Ing. Tech.* **63** (1991) no. 9, 921 – 925.
118. D. Arnold: "Projektmanagment im internationalen Großanlagenbau und Konsequenzen für die Aus- und Weiterbildung von Führungskräften," *Zfbf Sonderheft* 24 (1989) 55 – 77 (Verlagsgruppe Handelsblatt, Düsseldorf-Frankfurt).
119. D. Kumar: "Developing Strategies and Philosophies Early for Successful Project Implementations," *Int. J. Projekt Management* **7** (1989) no. 3.
120. L. F. Rolstad: "Project Start-Up in Tough Practice," *Int. J. Project Management* **9** (1991) no. 1, 10 – 14.
121. H. Reschke, H. Schelle, R. Schnoop: *Handbuch Projektmanagement,* vol. 1, Verlag TÜV Rheinland, Köln 1989.
122. N. Thumb: *Grundlagen und Praxis der Netzplantechnik,* Verlag Moderne Industrie, München 1969.
123. G. Bernecker: *Planung und Bau verfahrenstechnischer Anlagen,* 3rd ed., VDI Verlag, Düsseldorf 1984.
124. H. Groh, R. Gutsch: *Netzplantechnik,* 3rd ed., VDI Verlag, Düsseldorf 1982.
125. A. Franke: "Risikobewußtes Projektcontrolling, Risikomanagement als Element eines integrierten Projectcontrollings," Dissertation, Universität Bremen 1991.
126. S. C. Ward, C. B. Chapman, B. Curtis: "On the Allocation of Risk in Construction Projects," *Int. J. Project Management* **9** (1991) no. 3, 140 – 147.
127. E. Diegelmann, H. Benesch: "Industriemodelle – Methodik, Abwicklung," 3 R International (1983) no. 11, 535 – 541.
128. E. Diegelmann: "Erfassung, Beschaffung und Montage von Rohrleitungsteilen," *Chem. Ing. Tech.* **54** (1982) no. 4, 303 – 313.
129. J. Hengstenberg, B. Sturm, O. Winkler: *Messen, Steuern, Regeln in der Chemischen Technik,* 3rd ed., vol. V, Springer-Verlag, Berlin 1985.
130. *VDI 3546 – Konstruktive Gestaltung von Prozeßleitwarten,* Beuth-Verlag, Berlin 1987 – 90.
131. *VDI 3683 – Beschreibung von Steuerungsaufgaben, Pflichtenheft,* Beuth-Verlag, Berlin 1986.
132. *VDI 3693 – Verteilte Prozeßleitsysteme, Prüfliste für den Einsatz,* Beuth-Verlag, Berlin 1985.
133. *VDI 2180 – Sicherung von Anlagen in der Verfahrenstechnik,* Bl. 1 bis 5, Beuth-Verlag, Berlin 1984 – 88.
134. ABB (ed.): *Schaltanlagen,* 8th ed., Cornelsen-Verlag, Düsseldorf 1988.
135. Siemens (ed.): *Handbuch der Elektrotechnik,* 3rd ed., Verlag Girardet, Essen 1971.
136. J. Matley: "Keys to Successful Plant Start-Ups," *Chem. Eng. (N.Y.)* **76** (1969) Sept. 8, 110 – 130.
137. W. Langhoff: "Inbetriebnahme," Beitrag im Kursushandbuch "Planung und Bau von Großanlagen der Chemischen Industrie," DECHEMA, Erfahrungsaustausch, D. Behrens, K. Fischbeck, Frankfurt 1969.
138. VDI-Richtlinien, Datenverarbeitung in der Konstruktion, Einführungsstrategien und Wirtschaftlichkeit von CAD-Systemen, VDI 2216 (October 1990, draft).
139. W. Hilkert, R. Pfohl: Anlagenkonstruktion mit CAD/CAE-Werkzeugen – ein Erfahrungsbericht – VDI-Berichte Nr. 861.5, 1990.
140. H. Schmidt-Traub: "Integrierte Informationsverarbeitung im Anlagenbau," *Chem. Ing. Tech.* **62** (1990) no. 5, 373 – 380.
141. R. Gareis: *Projektmanagement im Maschinen- und Anlagenbau,* Manz-Verlag, Wien 1991.
142. S. J. Heisler: "Project Quality and the Project Manager," *Int. J. Project Management* **8** (1990) no. 3, 133 – 137.

Continuous Mixing of Fluids

HANS-JÜRGEN HENZLER, Bayer AG, Wuppertal, Federal Republic of Germany

1. Introduction 2
2. Principles 3
3. Mixing in Pipes 5
3.1. Mixing Properties 5
3.2. Pressure Drop 6
4. Jet Mixers 6
4.1. Mixing Properties 6
4.2. Pressure Drop 8
5. Static Mixers 9
5.1. Mixing Properties 10
5.2. Pressure Drop 12
6. Dynamic Mixers 13
6.1. Stirred Tanks and Stirred Columns 13
6.2. In-Line Mixers 15
6.3. Pressure Drop 16
6.4. Stirrer Power 16
7. Residence-Time Distribution 17
8. Mixing of Non-Newtonian Fluids . 18
8.1. Mixing Behavior 19
8.1.1. Pipes and Jet Mixers 19
8.1.2. Static Mixers 19
8.1.3. Dynamic Mixers 20
8.2. Pressure Drop for Non-Newtonians Fluids 21
8.3. Power Requirements of Dynamic Mixers 21
9. Selection and Comparison of Continuous Mixers 22
9.1. Parameters for Energetic Comparison 22
9.2. Energetic Comparison 23
9.2.1. Mixers with Fixed Internals 24
9.2.2. Dynamic Mixers 25
9.2.3. Comparison between Mixing Pipes and Dynamic Mixers 25
10. References 26

Symbols

c	concentration, kg/m^3
$\bar{c}$	mean concentration after complete mixing, kg/m^3
Δc	maximum concentration difference, kg/m^3
d	stirrer diameter, m
d_{B}	diameter of orifice, m
d_{E}	diameter of discharge opening, m
d_{i}	diameter of inlet lines to mixer, m
D	diameter of mixer, m
D_{h}	hydraulic diameter, m
h	stirrer blade height, m
Δh	spacing between stirrers, m
ID	diffusion coefficient, m^2/s
K	consistency index, Pa $\cdot$ s^{m}
k	empirical constant in Metzner – Otto law, Equation (36)
L	length of mixer, m
L_{e}	length of a diverter element, m
m	flow exponent
n	stirrer rotation speed, rpm
n	number of samples
N	number of stirred tanks
Δp	total pressure drop in mixer, N/m^2
P	power of net flow through device, W
P_{n}	agitator power, W
t	time, s
$\bar{t}$	mean residence time, s
t_{H}	homogenization time in batch operation, s
t_{s}	torsion factor (static mixers)
u	mean velocity of net flow, m/s
u_{B}	mean flow velocity in orifice, m/s
u_{i}	flow velocities at inlet to mixer, at nozzle discharge, m/s
V	volume, m^3
V_{tot}	total volume of a cascade of tanks, m^3
$\dot{V}$	volume flow rate in mixer, m^3/s
$\dot{V}_i$	volume flow rates of components, m^3/s
$\dot{V}_{\mathrm{B}}$	volume flow rate of backflow eddy (jet mixer), m^3/s
$\dot{\gamma}$	shear rate, 1/s
η	dynamic viscosity, kg m^{-1} s^{-1}
ν	kinematic viscosity, m^2/s
$\bar{\nu}$	representative kinematic viscosity, m^2/s
$\bar{\nu}_{\mathrm{P}}$	representative kinematic viscosity for shear gradient of Equation (38), m^2/s
$\bar{\nu}_{\mathrm{s}}$	representative kinematic viscosity for shear gradient of Equation (35), m^2/s
ϱ	mean density, kg/m^3
ϱ_i	densities of components, kg/m^3

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_561

τ shear stress, Pa

Dimensionless groups

Ar Archimedes number, $gD^3 \Delta \varrho / \nu^2 \varrho$
c_{H} mixing number for continuous operation, $n\bar{t}$
c_{PHD} mixing power number for given mixer diameter, $P D^4 / \varrho \dot{V}^3$
c_{PHL} mixing power number for given mixer length, $P L^4 / \varrho \dot{V}^3$
c_{PHV} mixing power number for given mixer volume, $PV^{4/3} / \varrho \dot{V}^3$
Eu_{i} Euler number of inlet to mixer, $2 \Delta p_{\mathrm{i}} / \varrho u_{\mathrm{i}}^2$
Eu_{B} Euler number of internals in mixer, $2 \Delta p_{\mathrm{B}} / \varrho u_{\mathrm{B}}^2$
Eu_{E} Euler number of mixer discharge opening, $2 \Delta p_{\mathrm{E}} / \varrho u_{\mathrm{E}}^2$
Fr Froude number, $u^2 \varrho / D g \Delta \varrho$
I_{S} degree of segragation, $(c-\bar{c})^2 / \bar{c}(1-\bar{c})$
I_1 variation coefficient $\sqrt{(c-\bar{c})^2} / \bar{c}$
I_2 degree of inhomogeneity, $\Delta c / \bar{c}(1-\bar{c})$
I_3 degree of inhomogeneity, $\Delta c / \bar{c}$
$(L/D)_{\mathrm{I}}$ mixing length for degree of inhomogeneity, I
Ne Newton number of net flow, $PD^4 / \varrho \dot{V}^3$
Ne_{n} Newton number of stirrer, $P_{\mathrm{n}} / \varrho n^3 d^5$
Re Reynolds number of pipe flow, $u D / \nu$
Re_{n} Reynolds number of stirrer, nd^2 / ν
$Re'_{\mathrm{V}*}$ modified Reynolds number of pipe flow, $\dot{V} / V^{1/3} \nu$
Sc Schmidt number, $\nu / I D$
Wi Weissenberg number
ε porosity
λ pipe friction coefficient
ψ pressure drop coefficient (static mixer)

Subscripts

DC decolorization method
P photoelectric method
I degree of inhomogeneity
WB baffles at the vessel wall

1. Introduction

Mutually soluble liquids or gases must be mixed together in many industries. Continuous processes, moreover, often require the use of continuous mixers.

Continuous mixers can be classified by residence time, residence-time behavior, and the way in which mixing energy is introduced. The required residence time and residence-time behavior are governed by the process conditions, among other factors; in chemical and biological reactions, they also depend on reaction kinetics. The method of supplying energy, in contrast, can be freely chosen. A distinction is made between mixers with moving internal parts and those without. The latter include mixing ducts, jet mixers, and mixers with stationary internals (also called static mixers); here the energy for mixing is withdrawn from the flowing medium. Mixers with moving parts, also known as dynamic mixers, include stirred tanks, stirred columns, and in-line mixers. In these devices, the energy required for mixing is introduced by stirrers. Figure 1 shows a classification of mixer types, initially according to residence time.

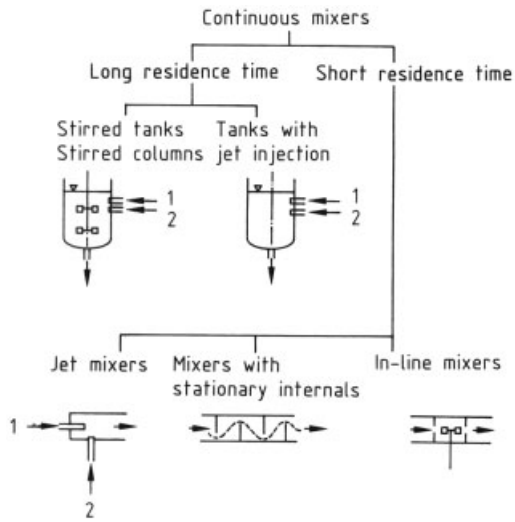

Figure 1. Types of mixers

2. Principles

When mutually soluble liquids or gases are mixed together, the purpose is to reduce concentration or temperature differences until a specified degree of homogeneity is reached. This is caused by the macroscopic and microscopic

mixing resulting from flow in the mixer; molecular diffusion also contributes to micromixing.

Macromixing, which is caused by large-scale motion in the mixer, is the rate-determining step in most mixing processes taking place in the turbulent to laminar regimes. As a rule, therefore, it also determines whether a mixer is suitable for a given mixing process. Exceptions are mixing processes used to carry out fast, competitive reactions, in which micromixing is also important and backmixing must be prevented as much as possible so that no undesirable byproducts are formed [1], [2].

Species entering a continuous mixer at L=0, with volume flow rates $\dot{V}_1$ and $\dot{V}_2$ and concentrations c_1 and c_2 are mixed together as the material moves along the mixer length L. The state of mixing can be described by the degree of inhomogeneity, which is defined in various ways. The degree of inhomogeneity represents the departure of the concentration c_i at a given place and time in the mixer from the arithmetic mean concentration $\bar{c}$ reached after complete mixing.

The Danckwerts degree of separation is often used [3]. This figure is the ratio of the local variance

$$\overline{(c-\bar{c})}^2=\sum_{i=1}^{n}(c_i-\bar{c})^2/n-1 \text{ where } \bar{c}=(\Sigma c_i)/n$$

to the variance at the mixer inlet $(\bar{c}-c_1)(c_2-\bar{c})$. If the concentration is normalized with c_1=0 and c_2=1, the degree of segregation is

$$I_S=\frac{\overline{(c-\bar{c})}^2}{\bar{c}\,(1-\bar{c})} \tag{1}$$

where $\bar{c}=\dot{V}_2/(\dot{V}_1+\dot{V}_2)$.

The degree of segregation takes on the values 1 for complete separation and 0 for complete mixing.

Another widely used measure is the variation coefficient I_1, which is defined as the ratio of the standard deviation $\sigma=\sqrt{\overline{(c-\bar{c})}^2}$ to the mean concentration $\bar{c}$:

$$I_1=\frac{\sigma}{\bar{c}}=\frac{\sqrt{\overline{(c-\bar{c})^2}}}{\bar{c}} \tag{2}$$

Degrees of inhomogeneity derived from the maximum concentration difference $\Delta c = (c-\bar{c})_{\max}$ are also employed [4–6]:

$$I_2=\frac{\Delta c}{\sqrt{\bar{c}\,(1-\bar{c})}} \tag{3a}$$

$$I_3=\frac{\Delta c}{\bar{c}} \tag{3b}$$

In practice, homogeneities of $I_S=10^{-4}$, $I_1=\sigma/\bar{c}=0.01$, and $I_2=0.05$ are adequate.

The quality of mixing can be calculated in advance by solving the appropriate differential equations, only for a few simple cases. Mixing processes are so complex that reliable information can be obtained only by experiments. Experimental results must be formulated as dimensionless quantities by means of similarity theory, so that the design knowledge gained can be used for scaling-up.

In general, dimensionless groups applicable to mixing are derived by considering a mixer in which volume flows differing in density and viscosity are mixed by both the flow through the mixer and a stirrer (Fig. 2). For a given mixer geometry, the desired homogeneity depends on the length L of the mixer, the rotational speed n of the stirrer, the diameter D of the mixer, the volume flow rates $\dot{V}_1$ and $\dot{V}_2$, the densities ϱ_1 and ϱ_2, the viscosities ν_1 and ν_2, the diffusion coefficient ID, and the acceleration due to gravity g.

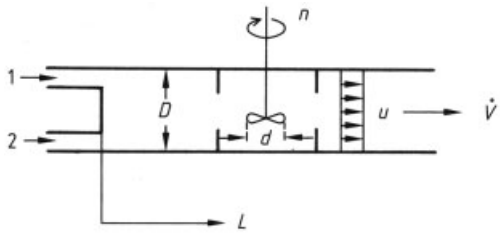

Figure 2. Schematic of a continuous-flow mixer

If one of the degrees of inhomogeneity I defined by Equations (1) – (3) is considered as the target figure, the following relevance list is obtained

$$I=f_1\left(L,n,D,\dot{V}_1,\dot{V}_2,\varrho_1,\varrho_2,\nu_1,\nu_2,ID,g\right) \tag{4}$$

If geometrical similarity is preserved, dimensional analysis gives a set of eight independent dimensionless groups:

$$I=f_2\left(\frac{L}{D},nt,\frac{uD}{\nu},\frac{gD^3}{\nu^2},\frac{\Delta\varrho}{\varrho},\frac{\nu}{ID},\frac{\dot{V}_1}{\dot{V}_2},\frac{\varrho_1}{\varrho_2},\frac{\nu_1}{\nu_2}\right) \tag{5}$$

Here, $Re=uD/\nu$ is the Reynolds number of the pipe flow, $Ar=g\,D^3\,\Delta\varrho/\nu^2\varrho$ is the Archimedes number, and $Sc=\nu/I\,D$ is the Schmidt number. For convenience, these are formulated by using the mean velocity u, an average viscosity, and an average density (the data of the homogeneous mixture).

The Schmidt number usually plays a secondary role because of the dominance of macromixing. The only exception would be processes involving competitive reactions in which the reaction rate is greater than the velocity of micromixing. Because the time required for micromixing depends on the size of the smallest turbulent eddies, the progress of the reaction in such processes depends not only on the viscosity and the diffusion coefficient but also on the energy dissipation.

If the degree of inhomogeneity I is given, two choices of target figure are possible, depending on the type of mixer and the mode of analysis. Either the mixing length $(L/D)_{\mathrm{I}}$ or, when stirrers are employed, the mixing number for continuous mixing $c_{\mathrm{HI}}=(nt)_{\mathrm{I}}$ [4], [5] can be used. These become functions of the remaining groups in Equation (5).

Because comprehensive experimental data are lacking, the results presented in the following sections constitute only a part of the dependences formulated in Equation (5).

The decrease in concentration differences when uniform mixing occurs is described by an exponential function, so that if, for example, the mixer is made longer,

$$I=I_0\exp\left(-BL/D\right) \tag{6}$$

where B is an empirical constant.

In these cases, the mixing length L_{I} determined experimentally for a degree of inhomogeneity I_{I} can be related to the mixing length L_{II} corresponding to another degree of inhomogeneity I_{II}:

$$\frac{L_{\mathrm{II}}}{L_{\mathrm{I}}}=\frac{\ln\left(I_0/I_{\mathrm{II}}\right)}{\ln\left(I_0/I_{\mathrm{I}}\right)} \tag{7}$$

If the degree of segregation is used, $I_0=1$.

The total power P_{tot} that must be dissipated to mix several components i in the generalized mixer of Figure 2 consists of the power P required to sustain the flow and the power P_{n} needed to drive the stirrer:

$$P_{\mathrm{tot}}=P+P_{\mathrm{n}} \tag{8}$$

where

$$P=\sum_i \Delta p_i\dot{V}_i=Ne\,\frac{\varrho\dot{V}^3}{D^4} \tag{8a}$$

$$P_{\mathrm{n}}=Ne_{\mathrm{n}}\varrho n^3 d^5 \tag{8b}$$

The Newton numbers Ne and Ne_{n} are empirical quantities that must be determined by experiment. They depend chiefly on the mixer geometry and the Reynolds number $Re=uD/\nu$ of the net flow or the Reynolds number $Re_{\mathrm{n}}=nd^2/\nu$ of the stirrer. If a marked interaction occurs between pipe flow and stirrer flow, an additional dependence on the velocity ratio dn/u must be incorporated (see Section 6.3).

The Newton number Ne of the pipe flow can also be estimated by adding the individual flow losses in the mixer. As a rule, the interference of various flow fields results in values somewhat smaller than the calculated values.

Large pressure drops occur in the mixer because of the abrupt widening of flow channels downstream of the nozzles ($\Delta p_{\mathrm{i}}=Eu_{\mathrm{i}}\,\varrho\,u_{\mathrm{i}}^2/2$) and downstream of flow-guiding internals such as orifices or static mixers ($\Delta p_{\mathrm{B}}=Eu_{\mathrm{B}}\,\varrho\,u^2/2$), and also because of possible constrictions of flow at the mixer outlet ($\Delta p_{\mathrm{E}}=Eu_{\mathrm{E}}\,\varrho\,u^2/2$). If these pressure drops are considered, the Newton number for net flow is given by

$$Ne=\frac{8}{\pi^2}\left[\sum_i Eu_i\frac{\varrho_i}{\varrho}\left(\frac{\dot{V}_1}{\dot{V}}\right)^2\left(\frac{D}{d_i}\right)^4+Eu_{\mathrm{B}}+Eu_{\mathrm{E}}\right] \tag{9}$$

As a rule, the Euler numbers Eu_{i}, Eu_{B}, and Eu_{E} must be determined experimentally.

3. Mixing in Pipes

3.1. Mixing Properties

In turbulent flow ($Re>2300$), mixing takes place even in hydraulically smooth and straight pipes and even if the components to be mixed are inlet axially and at the same speed. Concentration differences in the pipe decrease exponentially. For

equal volume flow rates ($\dot{V}_1/\dot{V}_2$=1) and densities (ϱ_1/ϱ_2=1), the following equation applies [7]:

$$I_S = 10^{-m\frac{L}{D}} \tag{10a}$$

where $m = \frac{1.89(Re-2300)^{0.61}}{Re}$

From Equation (10 a), the mixing length *L/D* is given by the general equation

$$\frac{L}{D} = -\frac{\log I_S}{m} = -0.53\,\frac{Re\cdot\log I_S}{(Re-2300)^{0.61}} \tag{10b}$$

The mixing length depends on the Reynolds number. It has a weak minimum at *Re*=6000, then increases with increasing Reynolds number because the flow velocity increases more rapidly than the turbulence. This effect was also observed by EVANS, who reports $L/D \sim Re^{0.17}$ [8]. The observed dependence is similar to that for the pipe friction coefficients λ for smooth pipes ($\lambda \sim Re^{-0.25}$ according to BLASIUS [9]); this result indicates the analogy between mass and momentum transfer. Accordingly, the turbulent regime in rough pipes should be characterized by a diminishing Reynolds number dependence and also by shorter mixing lengths.

A markedly greater length of pipe is required for mixing if the components differ in density (Fig. 3).

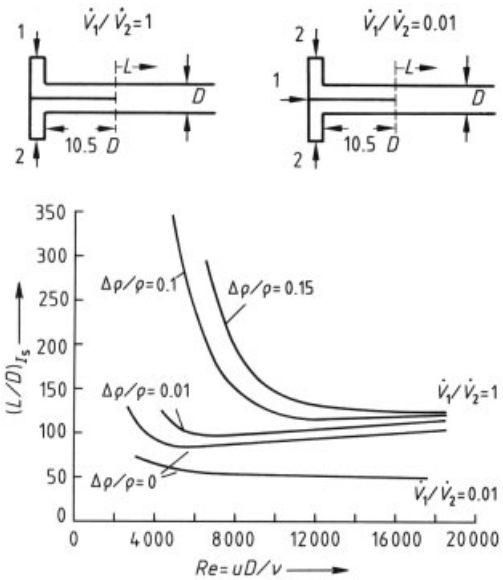

Figure 3. Effect of density differences on mixing in hydraulically smooth pipes, I_S=10^{-4} [7]

If the ratio of volume flow rates is small and both species are inlet axially at the same velocity, the mixing lengths are somewhat shorter according to HARTUNG [10]; see Figure 3.

Mixing can be accelerated by predistribution of the components; by velocity differences between species; and by curves, bends, and internals [4], [7], [10], [11]; see Chapters 4, 5, 6.

3.2. Pressure Drop

The pressure drop in pipes with no internals is due to frictional losses in the boundary layer next to the wall. It can be calculated from the pipe friction coefficients λ:

$$\Delta p = \lambda\cdot\frac{L}{D}\cdot\frac{\varrho}{2}\cdot u^2 \tag{11}$$

For fully developed flow in a pipe, these depend on the Reynolds number and the wall roughness. For smooth pipes,

$$\lambda = 64/Re;\ Re<2300 \tag{12}$$

for laminar flow, and

$$\lambda = 0.3164/Re^{0.25};\ 2300<Re<10^5 \tag{13}$$

for turbulent flow [9].

4. Jet Mixers

A jet mixer is a device in which one or more components are inlet through nozzles with a velocity greater than the mean velocity in the mixing tube.

4.1. Mixing Properties

Jet mixers exhibit much shorter mixing lengths than ducts, even if the components to be mixed are admitted axially and in the same direction [11]. The reason is that the jets bring about large-scale motion (see Fig. 4) and the flow is much more turbulent than mixing motion in simple pipe flow. Above certain Reynolds numbers, the mixing length *L/D* required to attain a given homogeneity therefore remains constant. HENZLER gives the following relations for the limiting Reynolds numbers Re_i for jet flow [4], [5], [12]:

$$Re_i = \frac{u_i d_i}{\nu} = A\frac{d_i}{D} \text{ for } 0.04 < \frac{d_i}{D} < 0.4 \tag{14}$$

where A has the following empirical values: $A \approx 5 \times 10^4$ for a jet mixer; $A \approx 10^4$ for a jet mixer with a downstream orifice having a diameter ratio $d_B/D \approx 0.35$.

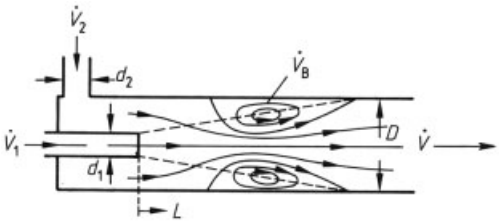

Figure 4. Jet mixer with backflow eddy

Concentration differences also decay exponentially in jet mixers, as is shown by the results of HARTUNG and HIBY [11] and of HENZLER [4], [12]. In very efficient mixers with highly nonuniform inlet flow, however, exponential behavior does not set in until most of the inhomogeneities have been smoothed out.

Mixing in jet mixers is largely complete after short flow distances ($L/D = 1.5 - 10$; see Figs. 5, 6, 7).

System	No.	d_1/D	d_2/D	ρ_1/ρ_2
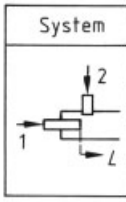	1	0.175	0.175	0.65
	2		0.25	
	3	0.250	0.175	
	4		0.25	

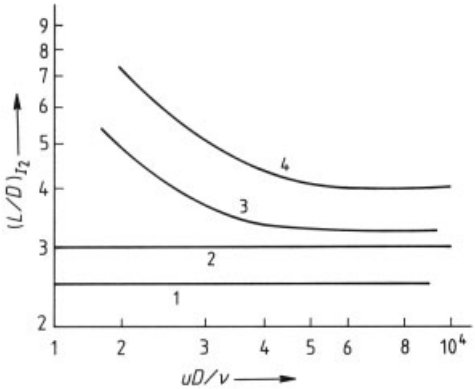

Figure 5. Mixing lengths (I_2=0.05) for jet mixers: $\dot{V}_1/\dot{V}_2$=1, ϱ_1/ϱ_2=0.65 [4]

For mixers with at least one component inlet axially, the mixing effect is governed by the expansion of the dominant axial jet. If the free jet formed in the upstream part of the mixer is able to draw in a flow of material greater than the total flow rate of the other components $\dot{V}_2$, a backflow eddy forms in the mixing tube (Fig. 4). Such an eddy leads to a distinct improvement in mixing. Accordingly, the influence of the volume flow rate ratio and the density ratio on mixing length can be described in terms of the volume backflow rate $\dot{V}_B$ [4], [12]

$$\frac{\dot{V}_B}{\dot{V}_1} = \frac{D}{d_1}\sqrt{\frac{\varrho_1}{\varrho_2}} - \frac{\varrho_1}{\varrho_2} - \frac{\dot{V}_2}{\dot{V}_1} \tag{15}$$

Figure 6 shows such a correlation of mixing lengths of various jet mixers. In the mixer with two tangential inlets to the outer pipe (type 2) and in the cross-flow configuration (type 3), the effect of the geometry of inlet 2 can be described simply in terms of the product of $\dot{V}_B/\dot{V}_1$ and the diameter ratio D/d_2. The design of inlets to the mixer has a crucial influence on mixing action. The shortest mixing lengths are achieved in mixers where the jets come into the most direct contact, as in the cross-flow type (type 3) and, even more so, the countercurrent arrangement (type 4) (see Figs. 6 and 7). Downstream orifices or restrictions in the mixing tube also have a beneficial effect. Division of a jet into several small jets does not promote mass transfer because of the tendency of parallel jets to coalesce (hydrodynamic paradox) [4], [12].

Mixers with swirling flow display a poorer mixing performance than those with radial inflow [4], [12], [13].

In the efficient countercurrent mixers (type 4 in Fig. 7), the volume flow rate has been found not to affect mixing action in the range $0.25 < \dot{V}_1/\dot{V}_2 < 4$ [4], [5], [12]. If the differences between the two volume flow rates are larger, the components should be combined upstream of the jet mixer, as shown in Figure 8.

The good performance of jet mixers has also been observed by HARTUNG [10] and HARTUNG, HIBY [11], who cite the following relation for a mixer as shown in Figure 4:

$$\frac{L}{D} = -1.7\sqrt{\frac{J}{J_1}\ln I_S} \tag{16}$$

in which the momentum – flow ratio

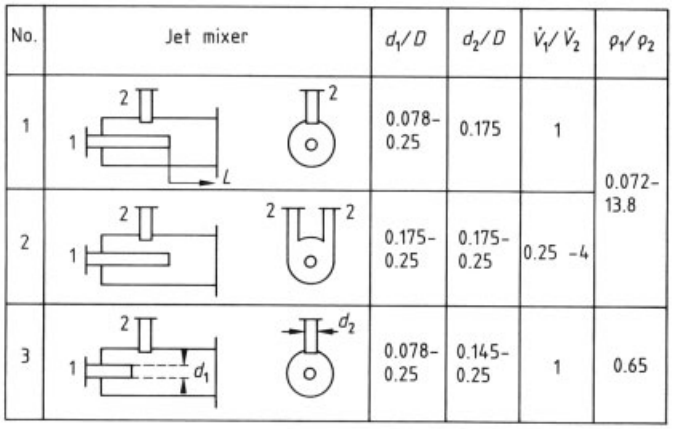

No.	Jet mixer	d_1/D	d_2/D	$\dot{V}_1/\dot{V}_2$	ρ_1/ρ_2
1		0.078–0.25	0.175	1	0.072–13.8
2		0.175–0.25	0.175–0.25	0.25 –4	
3		0.078–0.25	0.145–0.25	1	0.65

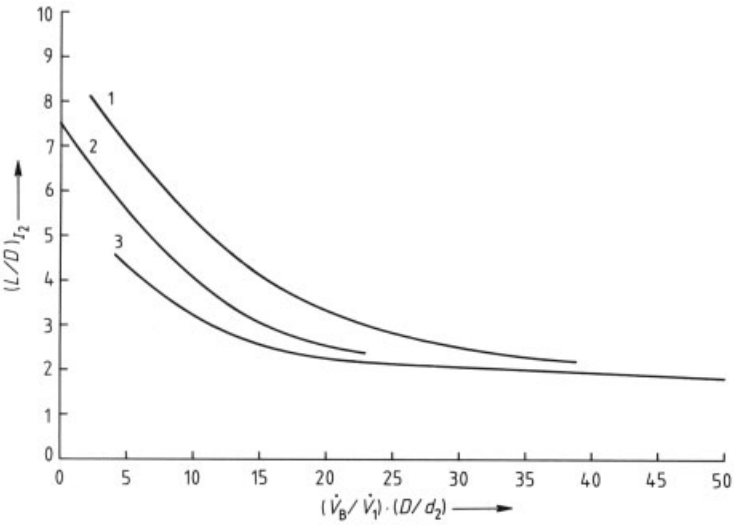

Figure 6. Correlation of mixing lengths ($I_2 = 0.05$) of jet mixers in turbulent flow regime [4]

$$\frac{J}{J_1} = \frac{\dot{m}u}{\dot{m}_1 u_1} = \left[\frac{d_1}{D}\left(1+\frac{\dot{V}_2}{\dot{V}_1}\right)\left(1+\frac{\varrho_2}{\varrho_1}\frac{\dot{V}_2}{\dot{V}_1}\right)\right]^2 \quad (17)$$

takes account of the effects of the volume flow rate ratio and the density ratio (analogous to Eq. 15 for volume backflow rate).

In this type of mixer, the radical inlet to the outer pipe lies about 5.3 D upstream of the mouth of the axial nozzle. A relatively uniform flow can accordingly form in the annular space before the actual mixing section is reached. The mixing lengths under given operating conditions are therefore somewhat longer than those in other mixer types [4], [12]. (For comparison with the results cited above, which apply to a degree of inhomogeneity I_2, the relation $I_2 = 6\, I_S^{1/2}$ can be applied [7].)

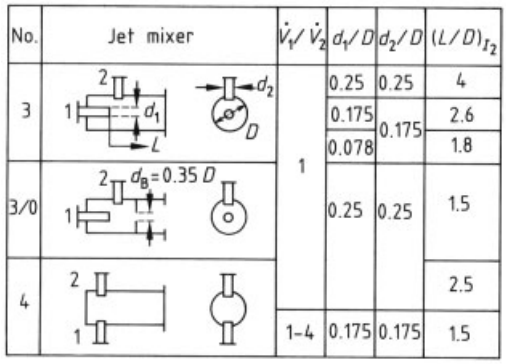

No.	Jet mixer	$\dot{V}_1/\dot{V}_2$	d_1/D	d_2/D	$(L/D)_{I_2}$
3		1	0.25	0.25	4
			0.175	0.175	2.6
			0.078		1.8
3/0	$d_B = 0.35\,D$		0.25	0.25	1.5
4					2.5
		1–4	0.175	0.175	1.5

Figure 7. Mixing lengths ($I_2 = 0.05$) of jet mixers in the turbulent flow regime: ϱ_1/ϱ_2=0.65 [4], [12]

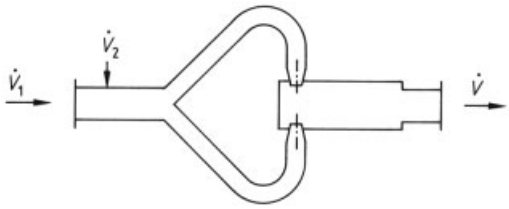

Figure 8. Countercurrent jet mixer for $\dot{V}_1 >> \dot{V}_2$ and $\varrho_1 \approx \varrho_2$

4.2. Pressure Drop

Significant pressure drops result from abrupt expansion of the flow when various components are inlet to the upstream part of the mixer (Δp_i) and from the internals in the mixing pipe (Δp_B). The total pressure drop Δp is calculated from the Euler number of the mixer, which must generally be determined by experiment.

Except for countercurrent devices (mixer type 4 in Fig. 7), Euler numbers for jet mxers with nozzle inlets and no internals can be calculated from the simplified equation:

$$Eu_1 = \frac{\Delta p_1}{\frac{\varrho_1}{2} \cdot u_1^2} \approx 1 - 2\left(\frac{d_1}{D}\right)^2 \left[1 + \frac{J_1}{J_2} - \frac{J}{J_1}\right] \quad (18)$$

which is derived from the conservation of momentum. This equation holds for $J_1/J_2 > 0.25$ [4], [12]. If the nozzle cross section is small compared to the mixing tube ($d_1^2/D^2 < 1$), the Euler number approaches unity. For countercurrent mixers, experimental results lead to the following approximation [4], [12]:

$$Eu_1 \approx 0.9 - 0.45\left(\frac{d_1}{d_2}\right)^2 \frac{J_2}{J_1} \quad (19)$$

Equations (18) and (19) also hold for the second component if all subscripts 1 and 2 are interchanged.

If the mixer also has internals, the total pressure drop for stream i can be estimated by adding the individual drops:

$$\Delta p_{\mathrm{tot},i} \approx \left[Eu_i \frac{\varrho_i}{\varrho}\left(\frac{\dot{V}_i}{\dot{V}}\right)^2 \left(\frac{D}{d_i}\right)^4 + Eu_\mathrm{B} + Eu_\mathrm{E}\right] \frac{\varrho}{2} u^2 \quad (20)$$

If an orifice (diameter d_B) is present to improve mixing performance, the turbulent Euler number Eu_B is given by

$$Eu_\mathrm{B} = \left[\frac{1}{\alpha}\left(\frac{D}{d_\mathrm{B}}\right)^2 - 1\right]^2 \quad (21a)$$

If there is a restriction (diameter d_E) at the end of the mixer,

$$Eu_\mathrm{E} = \left(\frac{D}{d_\mathrm{E}}\right)^4 \left(\frac{1}{\alpha} - 1\right)^2 \quad (21b)$$

For sharp-edged restrictions with turbulent flow, $Re_\mathrm{B} = d_\mathrm{B}\, u_\mathrm{B}/\nu$ and $Re_\mathrm{E} = d_\mathrm{E}\, u_\mathrm{E}/\nu > 10^4$ for a contraction coefficient $\alpha = 0.586 + 0.12(d_\mathrm{B/E}/D)^2$ [14]. Values for lower Reynolds numbers are given in [15].

5. Static Mixers

Static mixers are mixing pipes with fixed internals arranged so as to create a layering of fluid elements and induce eddy motion in the flow. In addition to very simple internals such as orifices and baffles (see Figs. 1 and 9 A), a number of special diverting structures are employed (e.g., see [16–18]).

Lamellar or web-type internals (Fig. 9 B) are distinguished from helical internals (Fig. 9 C). A mixer consists of a number of such elements, each with length L_e, arranged in series. Each element makes a 90° angle with the preceding one. In the case of helical internals, right-handed and left-handed elements alternate.

The diverting structures bring about layering of the liquid elements, which leads to very efficient mixing even in laminar and creeping flow regime.

Modified versions of the SMX mixer shown in Figure 9 B are the SMXL and SMF types. The SMXL is a stretched version of the SMX with low pressure drop [19]. The SMF mixer consists of inclined webs attached to the pipe wall, which do not touch one another. This design is especially well suited to media that tend to cause blockages. In the SMR-Sulzer type the mixing elements are constructed of heat-exchanger tubes (Fig. 9 B).

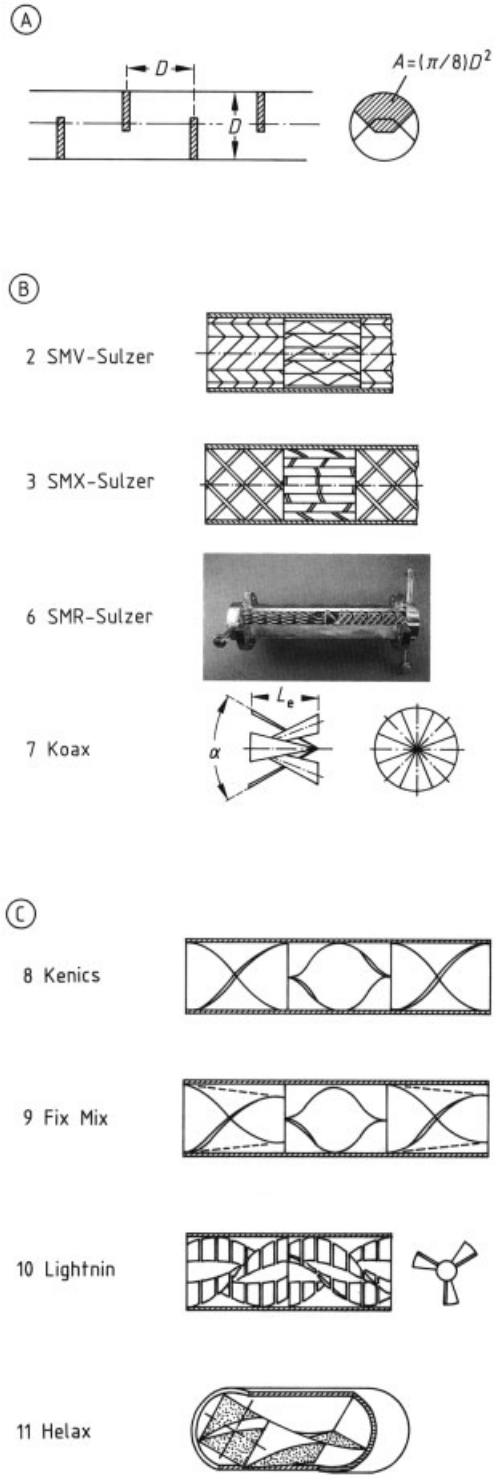

Figure 9. A) Mixer with baffles [11]; B) Static mixers with web-type internals; C) Static mixers with helical internals

5.1. Mixing Properties

As in smooth pipes and jet mixers, the concentration changes exponentially when the inlet flow is undisturbed or after the disturbance has decayed (Figs. 10 and 11). The various designs give very different mixing actions, whose intensities depend on the Reynolds number. Constant mixing lengths are found for both laminar and turbulent flow (Fig. 12). In the case of turbulent flow, homogeneity ($I_S = 10^{-4}$) is achieved at very short mixing lengths ($L/D = 2-8$). In the laminar regime the mixing lengths for a given mixer type are a factor of 4 – 8 longer.

The Kenics mixer shows a decline in the mixing action at $2<Re<300$ because longitudinal mixing is enhanced preferentially over transverse mixing. Similar but less marked behavior has been observed in other static mixers [23]. This behavior is typical of the transition regime and can be suppressed by installing additional structures, such as orifices, to prevent longitudinal mixing.

Differences in mixing lengths reported by various authors (e.g., compare the values for Kenics and Helax elements in Figs. 10, 11, 12) may result from the use of mixing elements that are not geometrically identical, as well as the use of different measurement techniques.

According to Tauscher and Streiff [22], with various gases in the SMV-Sulzer mixer, the ratio of volume flow rates under turbulent flow conditions has virtually no effect on mixing quality, if the degree of segregation in Equation (1) is used as degree of inhomogeneity; see Figure 10. (In the literature [24] and in publications by the manufacturers of static mixers [19], the same is claimed to hold true for laminar flow. Experimental proof of this claim, however, is not presented.) In practice, mixing is usually considered adequate when a certain coefficient of variation $\sigma/\bar{c}$ (Eq. 2) is attained; then, however, since $\sigma/\bar{c}=I_S(\dot{V}_1/\dot{V}_2)^{1/2}$, the quality of mixing depends on the ratio of volume flow rates.

For the SMV-Sulzer mixer in the turbulent flow regime, no effect of density differences has been found at Froude numbers $Fr=u^2\,\varrho/(Dg\,\Delta\varrho)>2.5$; see Figure 10 [22]. This should also be true of other static mixers with intensive transverse mixing.

No.	Internals		Media	$\Delta p/p$	$\dot{V}_1/\dot{V}_2$	Ref.
	Smooth pipe		liquids	0	1	[7] [10] [20]
1	Baffles					
8	Kenics					
4	Sulzer	SMXL	CO_2/air	ca. 0.5	0.11	[21]
5	Sulzer	SMF				
2	Sulzer	SMV				
	Sulzer	SMV/9 plates	misc. gases	0.53–1.71	0.005–0.43	[22]

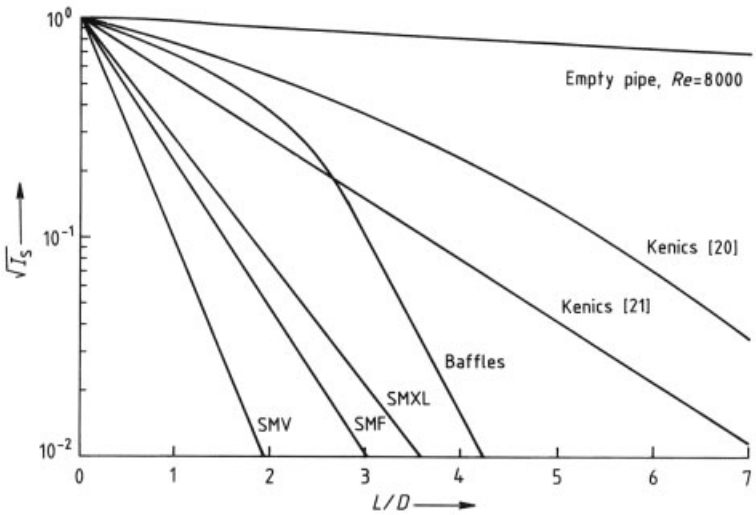

Figure 10. Smoothing out of inhomogeneities in an empty pipe and in static mixers in the turbulent flow regime: $Re > 5 \times 10^3$ (see Table 1 for data) [7], [10], [20], [21], [22]

Under turbulent flow conditions, a mixing effect is also observed downstream of flow-guiding internals. In the SMV-Sulzer mixer this effect continues for two pipe diameters beyond the mixing element [22].

Viscosity differences between the components make them more difficult to mix. The SMX-Sulzer mixer is suitable in such cases [24], [25]; if, for example, a low-viscosity component (viscosity η_2) is to be mixed into a high-viscosity component (η_1), the mixing length L/D is only slightly greater than the value for laminar flow $(L/D)_1$ with no viscosity difference:

$$\frac{L}{D} = \left(\frac{L}{D}\right)_1 + a \log\left(\frac{\eta_1}{\eta_2}\right) \tag{22}$$

The coefficient a=1 for SMX mixers and 2 for SMXL mixers.

If density or viscosity differences exist between the media present in helical mixers, care must be taken that there is no gap between the mixing elements and the pipe wall because such clearances would lead to short-circuit flows and thus much poorer mixing.

When the ratio of volume flow rates differs from unity, the components with low throughput should be distributed as uniformly as possible. An especially good way to do this is to divide the stream into smaller jets, evenly spaced over the mixing tube. Axial or concentric feed is preferable to feed from just one side [26].

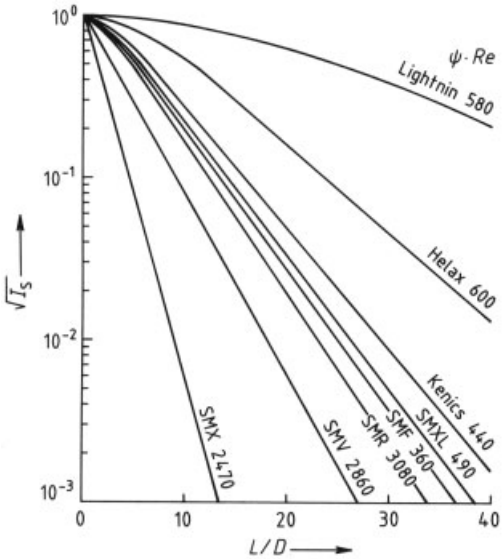

Figure 11. Smoothing out of inhomogeneities in static mixers in the laminar flow regime: $\dot{V}_1/\dot{V}_2$=0.1 [24]

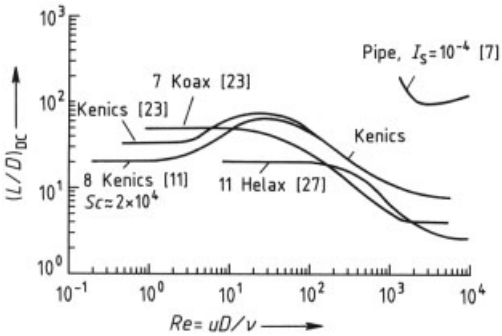

Figure 12. Mixing lengths for static mixers (see Table 1 for data) [7], [11], [23], [27]

5.2. Pressure Drop

As in pipe flow, the pressure drop in static mixers is proportional to pipe length. A coefficient ψ

$$\psi = \frac{Eu_{\mathrm{B}}}{L/D} \tag{23}$$

independent of the length is therefore used to describe the pressure drop:

$$\Delta p = \psi \cdot \frac{L}{D} \cdot \frac{\varrho}{2} \cdot u^2 \tag{24}$$

where ψ depends on the Reynolds number and mixer geometry; see Figure 13 and Table 1. For laminar flow, ψ is proportional to $1/Re$, that is, $\psi\, Re$=constant. As the Reynolds number increases, this dependence becomes less pronounced, so that constant ψ values are expected for fully turbulent flow. This flow regime is not, however, reached until $Re = 10^4$ [23], [27], [28], [31], [32].

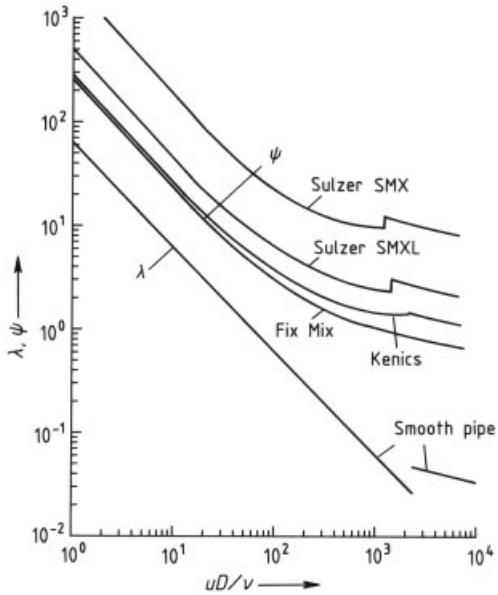

Figure 13. Euler number versus Reynolds number for static mixers [28]

For the Sulzer mixer KALBITZ finds an abrupt increase in ψ at the onset of turbulent flow ($Re \approx 1.3 \times 10^3$) in static mixers, analogous to the behavior of the pipe friction coefficient λ [28]; see Figure 13. Since the curve of ψ resembles that of λ in the range of Reynolds numbers studied ($1.3 \times 10^3 < Re < 5 \times 10^3$), the ratio ψ/λ for turbulent flow is listed in Table 1, along with other literature data. The ratio represents the pressure drop as a multiple of the value for a hydraulically smooth pipe.

The highly varied geometries of static mixers have hindered development of a more generally valid representation of pressure-drop characteristics based on geometrical quantities such as hydraulic diameter, which takes account of the specific geometry of the flow channels [31–33].

Table 1. Pressure-drop coefficients of static mixers

No.	Type		$\frac{L_e}{D}$	Laminar flow			Turbulent flow			Ref.
				Re	ψ *Re*	ψ / λ	*Re*	ψ	ψ / λ	
1	Baffles		1				> 4500	7		[11]
2	Sulzer	SMV	0.5		2860	45		6.8		[22], [29]
3	Sulzer	SMX	1		1893	31	> 1350		230	[28]
					2470		10^3	10		[24]
4	Sulzer	SMXL	4.4		502	8	> 1530		58	[28]
					590		10^3	2		
5	Sulzer	SMF		< 50	360	5.6	10^3	2		[24]
6	Sulzer	SMR	0.5		3080					
7	Koax		0.87		830	13	> 5000	3.5		[23]
8	Kenics				281	4.4	> 2145		31	[28]
			1.5		420	6.5			14	[30], [11]
9	Fix Mix				271	4.2	> 850		19	[28]
10	Lightnin				580	9				[24]
11	Helax		1.3	< 100	805	12.5	$> 10^4$	1		[27]
					600	9.4				[24]

6. Dynamic Mixers

In dynamic mixers, the energy for mixing is supplied by rotating stirrers. This class includes large-capacity devices (single-stage stirred tanks and stirred columns) with multiple stirrers, as well as continuous-flow mixers with smaller capacities designed for in-line installation. The target quantity for homogenization in continuous-flow stirred mixers with fixed length (*L/D*=const.) is the mixing number c_{H}, defined [4], [5] analogously to the target quantity for batch operation [4], [6] (→ Stirring, Chap. 2.2. The value c_{H} is the product of agitator speed n and mean residence time $\bar{t}=V/\dot{V}$; it is directly proportional to the ratio of the stirrer tangential velocity v necessary for mixing to the mean pipe flow velocity u:

$$c_{\mathrm{H}}=n\bar{t}=\frac{1}{\pi}\cdot\frac{L}{D}\cdot\frac{v}{u} \tag{25}$$

The set of dimensionless groups given in Equation (5) applies to c_{H}.

6.1. Stirred Tanks and Stirred Columns

The positions of the inlet and outlet are particularly important for the quality of mixing in a stirred tank with a single stirrer. They should be arranged so that short-circuit flows cannot form. Figure 14 illustrates systems that, according to LIEPE, are well suited to continuous homogenization with turbulent flow ($Re_{\mathrm{n}}=nd^2/\nu>10^4$) over a wide range of flow rate ratios ($\dot{V}_1/\dot{V}_2=2-10^4$) [6]. The figure lists not only the mixing number c_{H} but also the ratio $\bar{t}/t_{\mathrm{H}}$ needed to reduce the inhomogeneities to $I_3=0.05$ (Eq. 3 b). This ratio represents the mean residence time $\bar{t}$ needed in the mixer as a multiple of the homogenization time t_{H} in batch operation.

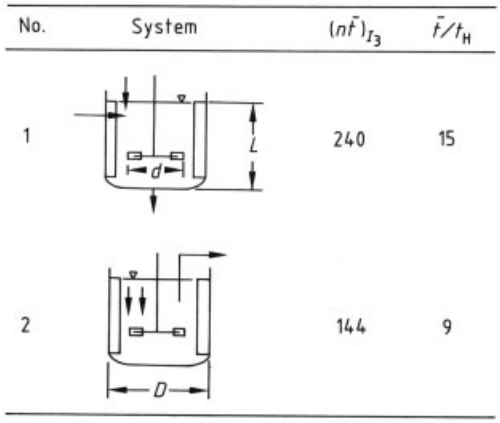

Figure 14. Mixing parameters of continuous stirred tanks ($d/D=0.4$, $L/D=1$, $\varrho_1/\varrho_2\approx 1$, $\nu_1/\nu_2\approx 1$, $Re>10^4$, $I_3=0.05$)

As in other types of mixers, inhomogeneities decay exponentially [4], [6]. According to LIEPE,

$$I_3 = 0.22 \exp\left(-0.16\bar{t}/t_{\mathrm{H}}\right) \tag{26}$$

for stirred tank 1 in Figure 14 [6]. The results can be converted to different mixing conditions analogous to Equation (7):

$$\frac{c_{\mathrm{HII}}}{c_{\mathrm{HI}}} = \frac{\ln\left(I_0/I_{\mathrm{II}}\right)}{\ln\left(I_0/I_{\mathrm{I}}\right)} \tag{27}$$

The ratio $\bar{t}/t_{\mathrm{H}}$ needed to attain practical homogeneity in a single-stage stirred tank is very large for the turbulent and transition regimes because of marked longitudinal mixing. The value therefore cannot be extended to other flow regimes. At lower Reynolds numbers, the only results available apply to multistage stirred equipment, which performs better than single-stage stirred tanks. Figures 15 and 16 present some data in plots of c_{H} versus *Re*. Figure 15 shows results for a two-stage stirred tank with a jet system for the injection of components 1 and 2. In the transition regime ($6 < Re < 60$), the use of jet systems requires the installation of orifices between stirrers because longitudinal mixing is faster than transverse mixing in this flow regime (see Section 5.1), so mixing performance is degraded (Fig. 15).

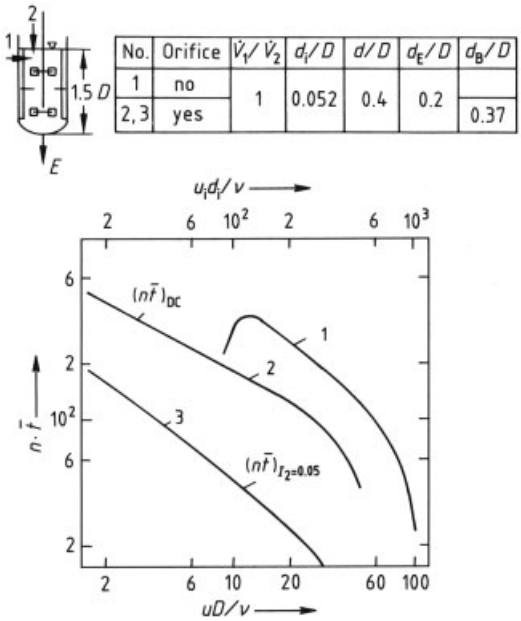

Figure 15. Mixing numbers for two-stage stirred tanks with jet system ($\dot{V}_1/\dot{V}_2 \approx 1$, $\varrho_1/\varrho_2 \approx 1$, $\nu_1/\nu_2 \approx 1$, I_2=0.05) [4], [5]

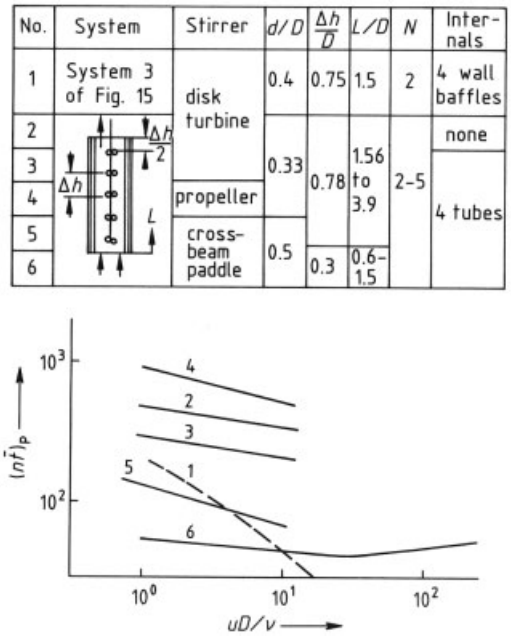

Figure 16. Mixing numbers for multistage stirred columns ($\dot{V}_1/\dot{V}_2 \approx 1$, $\varrho_1/\varrho_2 \approx 1$, $\nu_1/\nu_2 \approx 1$)
Curves 3 – 6 are derived from [34]

Similar effects are observed in stirred columns without jet systems (Fig. 16). With the very small stage-to-stage spacing $\Delta h/D$=0.3 used here, this effect was observed only at higher Reynolds numbers (Re>30). As results show, agitators such as the paddle stirrer and the cross-beam stirrer, which generate marked flow transverse to the flow direction, give lower mixing numbers than the axial-flow propeller and MIG stirrers.

The mixing performance of stirred columns improves as the stage-to-stage spacing $\Delta h/D$ decreases and the diameter ratio d/D increases [34]. For cross-beam paddle stirrers with $d/D = 0.5 - 0.75$, $h/d = 0.2$, $\Delta h/d = 0.3 - 1.3$, $Re = 0.32 - 14$, and $Re_{\mathrm{n}} = 10 - 300$, the mixing number is given by [34]

$$c_{\mathrm{H}} = 150 \left(\frac{\pi}{4}\right)^a \left(\frac{D}{d}\right)^{0.7} \left(\frac{\Delta h}{D} - 0.1\right)^{0.75} \cdot Re^{a-1} \tag{28}$$

where $a = 0.74\left(\frac{\Delta h}{D} + 0.1\right)^{-0.43}$. As in most mixing problems, the use of baffles has a beneficial effect [34].

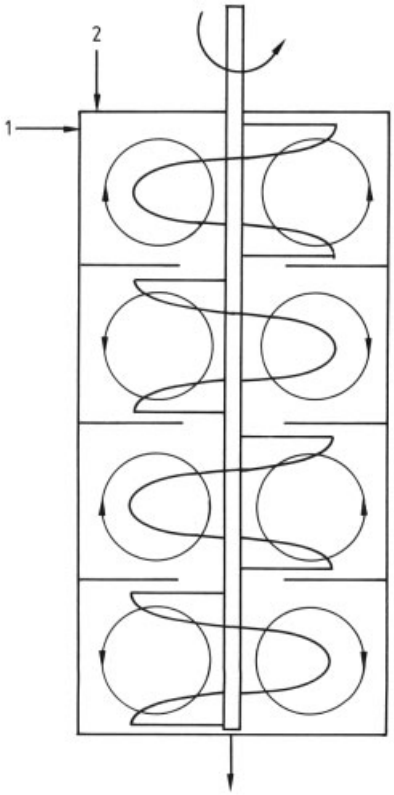

Figure 17. Dynamic mixer for laminar flow regime

For small density differences between the components being mixed, horizontal positioning of the stirrers is recommended; for larger differences ($\Delta\varrho/\varrho \geq 0.29$), a vertical arrangement with flow from top to bottom is preferred [34].

The sharp drop in mixing number at higher Reynolds numbers ($Re > 20$) in Figure 15 is due to jet mixing. Because of the small nozzle diameters in this system, a relatively large amount of energy is dissipated at the nozzles. As a result, premixing improves as the turbulence of the pipe flow increases. Curves 1 and 3 in Figure 16, which are for similar mixing systems, give an indication of the effect of jet mixing. The limiting Reynolds numbers, above which jet mixing alone yields adequate homogenization, are in the same range as the result of Equation (14) for jet mixers. In stirred equipment with injection of components through a jet system as shown in Figure 15, the mixing action becomes poorer as the ratio of volume flow rates increases [4].

The literature reports only a few studies for laminar flow ($Re < 2$) [4]. The equipment recommended for this flow regime is similar to that for batch homogenization: mixers with forced-agitation stirrers such as screw or helical ribbon stirrers (Fig. 17).

6.2. In-Line Mixers

An in-line mixer is a device having a high-speed agitator installed in-line to provide additional mixing in pipelines. In this type of mixer, restrictors can be used to create a forced flow of the components through the stirring zone and thus prevent short-circuit flows (see Fig. 18).

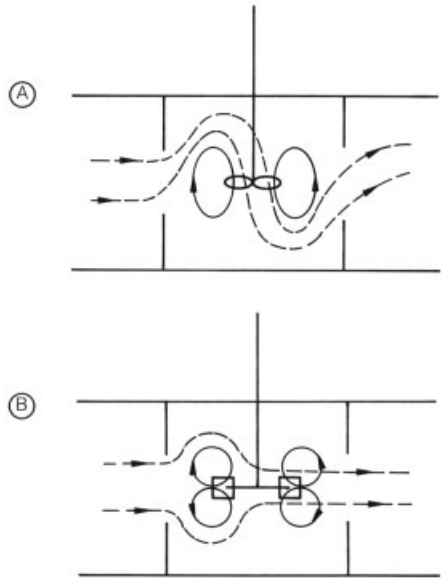

Figure 18. Flow in continuous-flow mixers

The homogenizing curves of these systems resemble those of mixers already discussed (Fig. 19). Again, devices that generate intensive motion transverse to the flow direction have lower mixing numbers (Fig. 18 A). Given the geometry of the device shown, this requires not radial-flow but axial-flow impellers such as propellers and pitched-tube stirrers (Fig. 20). For $Re > 20$, the effect of jet mixing can again be seen as a result of the increasing turbulence of flow through the device.

6.3. Pressure Drop

The total pressure drop $\Delta p_{\mathrm{tot,i}}$ between the inlet i and the outlet of a mixer can be calculated from the sum of the individual pressure drops. Significant contributions are due to the abrupt widening downstream of each inlet i and the restriction of flow in orifices and at the outlet of the mixer, so that Equations (18) – (21) derived for jet mixers can be employed.

In dynamic mixers, the pressure drop increases further due to impeller flow, particularly when the pipe flow and the impeller flow are in opposite directions (see pressure-drop coefficients in Fig. 21). According to WEISSGERBER, all types of impellers show a greater pressure drop if the ratio of impeller tangential velocity to pipe-flow velocity $\pi\, dn/u > 1$ [35].

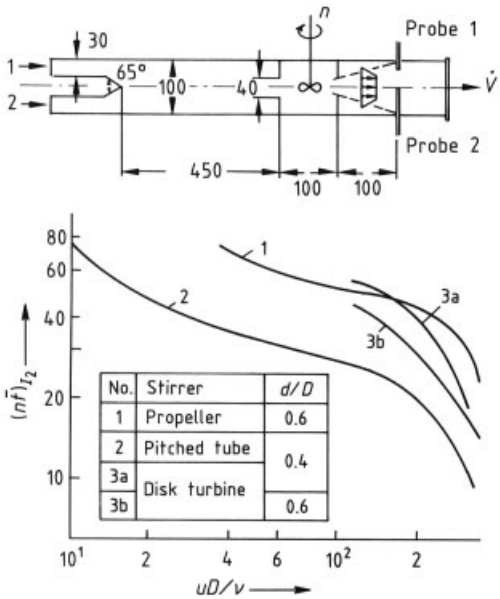

Figure 19. Mixing numbers for continuous-flow mixers ($\dot{V}_1/\dot{V}_2 \approx 1$, $\varrho_1/\varrho_2 \approx 1$, $\nu_1/\nu_2 \approx 1$), derived from [35]

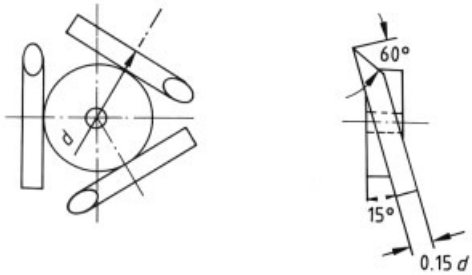

Figure 20. Pitched-tube stirrer from Figures 19 and 21

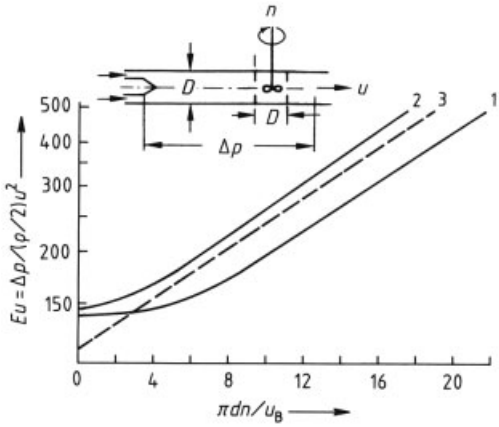

Figure 21. Euler numbers for the continuous-flow mixers of Figure 19 ($Re=uD/\nu>30$) [35]

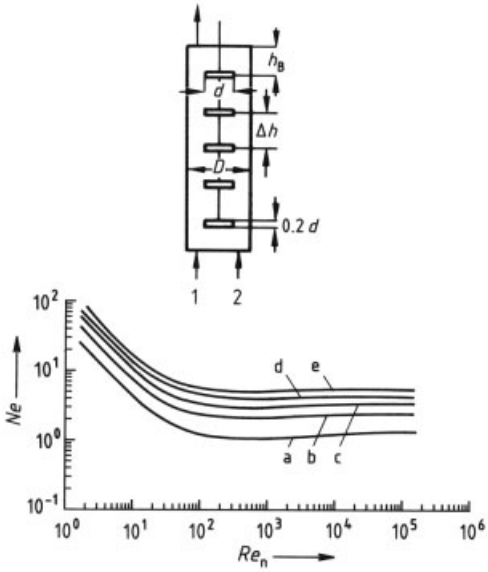

Figure 22. Power characteristics for stirred columns with cross-beam stirrers ($d/D = 0.33$, $\Delta h/D = 0.78$, $h_B/D = 0.5$, 4 round baffles) [34]
a) One stirrer stage; b) Two stages; c) Three stages; d) Four stages; e) Five stages

6.4. Stirrer Power

The power P_n delivered by the stirrer can be calculated from Equation (8 b). The Newton number Ne needed for this calculation depends on the type and geometry of the stirrer and is a function of the impeller Reynolds number $Re_n = nd^2/\nu$

(Figs. 22 and 23). If the homogeneity at the outlet of a dynamic mixer is adequate for the intended purpose, the ratio $u/\pi\, dn$ of the pipe-flow velocity to the required impeller tangential velocity is generally less than unity and has no effect on *Ne* [35], so that the Newton numbers for batch operation can be used [4], [6] (see also → Stirring, Chap. 2.3.).

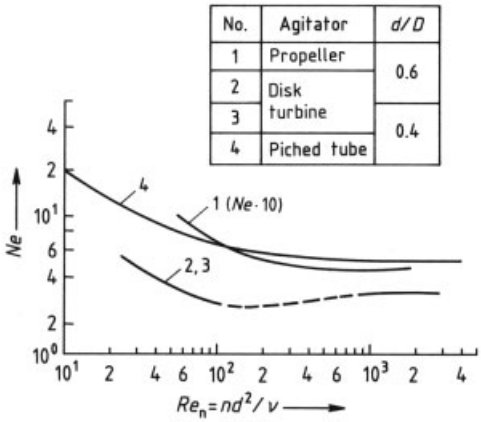

Figure 23. Newton numbers for the continuous-flow mixers of Figure 19 [35]

7. Residence-Time Distribution

For any chemical or biological reaction of order greater than zero (i.e., where the reaction rate depends on the concentration of reactants), the residence – -time distribution affects the reaction rate and thus the size of reactor needed. Accordingly, the residence-time distribution is an important criterion for selection of continuous-flow reaction equipment. The residence-time distribution depends on macroscopic mixing motion in the apparatus. It can be characterized by the residence-time spectrum, that is, the time dependence of concentration at the mixer outlet after a defined change in feed concentration. Two types of residence-time spectrum are distinguished: (1) that obtained after an impulse mark (addition of a quantity $m_z = c_0 V_{\mathrm{tot}}$ all at once – a Dirac impulse function) and (2) that obtained after a displacement mark (abrupt change in feed concentration from $c/c_0 = 0$ to $c/c_0 = 1$). The residence-time spectrum $(c/c_0)_V$ for the second case is found by integrating the impulse mark spectrum and is also referred to as the cumulative residence-time distribution:

$$S = \left(\frac{c}{c_0}\right)_V = \int_0^{t/\bar{t}} \frac{c}{c_0} \cdot d\left(\frac{t}{\bar{t}}\right) \tag{29}$$

This cumulative curve gives the fraction of the liquid that has a residence time in the reactor shorter than t.

The residence-time spectra of N equal-sized tanks with volume $V = V_{\mathrm{tot}}/N$ connected in series, calculated under the assumption of ideal mixing, are given in Equation (30) and Figure 24.

$$\frac{c}{c_0} = \frac{N^N}{(N-1)!} \left(\frac{t}{\bar{t}}\right)^{N-1} \exp\left(-N \cdot \frac{t}{\bar{t}}\right) \tag{30a}$$

$$\left(\frac{c}{c_0}\right)_V = 1 - \sum_{i=1}^{N} \frac{N t/\bar{t}}{(i-1)!} \exp\left(-N \cdot \frac{t}{\bar{t}}\right) \tag{30b}$$

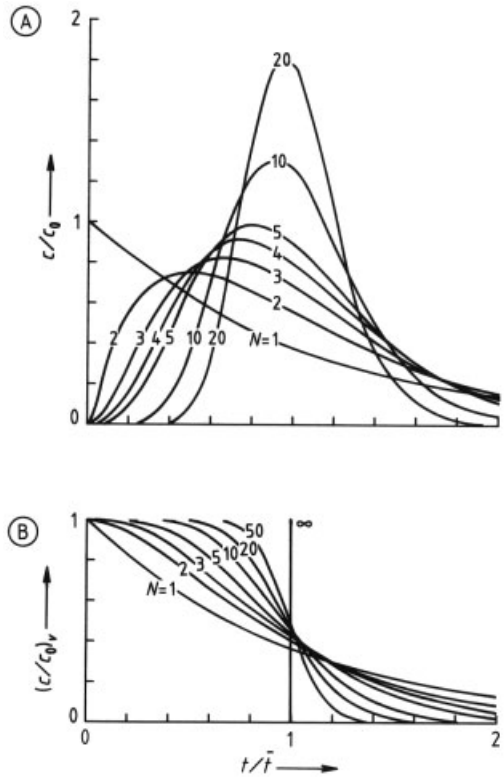

Figure 24. Residence-time distributions for N equal-sized ideal mixers in series ($\bar{t} = V_{\mathrm{tot}}/V$)
A) Impulse mark; B) Displacement mark

These calculated values can be used in practice when the mixing time in the reactor is small compared to the mean residence time in the individual stages or when the actual value of the product $n\bar{t}$ is larger than the mixing number c_{H}. An ideally mixed tank has a very broad residence-time spectrum (Fig. 24 A). As more reactors are connected in series (increasing N), the spectra become narrower and narrower until, at $N=\infty$, plug-flow distribution is obtained.

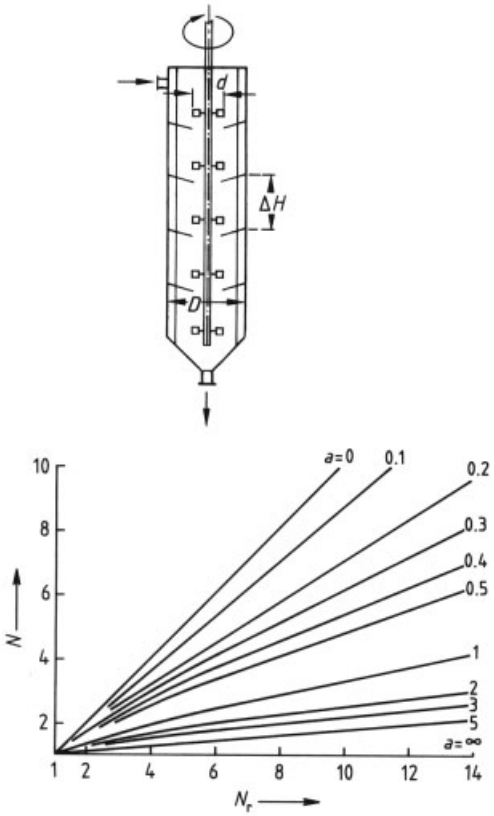

Figure 25. Theoretical number of stages N for stirred columns associated with spectra given in Figure 24 versus actual number of stages N_{r} for various backmixing factors a as defined in Equation (31) [6]

Because of backmixing between chambers, stirred columns have a broader residence-time spectrum than stirred tanks in series. The relationship between the actual and theoretical numbers of stages associated with a given residence-time spectrum in Figure 24 is illustrated in Figure 25 [6]. Each curve corresponds to a different backmixing factor a, which is calculated from the following empirical relation [6]:

$$a=BNe_{\mathrm{n}}^{1/3}\left(\frac{nd}{u}\right)\left(\frac{D}{\Delta H}\right)^{1/2}\left(\frac{d}{D}\right)^{2/3} \tag{31}$$

The proportionality constant B also depends on other geometric dimensions of the column that do not appear in Equation (31). For columns with horizontal trays, B=0.035; for trays inclined at 15°, B=0.052. In each case, the stirrers are centrally located in the chambers.

Jet mixers, which provide very intensive mixing in a short length of pipe, offer virtually ideal mixing.

Internals in mixing pipes cause the residence-time distribution to become narrower; as in stirred columns, the theoretical number of stages increases with increasing number of internals. The same holds for static mixers. These have a much narrower residence-time spectrum, especially in the laminar flow regime, than stirred tanks and stirred columns with few stages. The residence-time distributions of the SMX-Sulzer and Kenics mixers correspond to the distribution for a cascade of ca. 20 stirred tanks; that of the SMV-Sulzer mixer, to a cascade with about five stages.

For further comparison, Figure 26 includes plots for the mixer types just discussed and also for a smooth pipe, a fixed bed of Rasching rings, and plug flow.

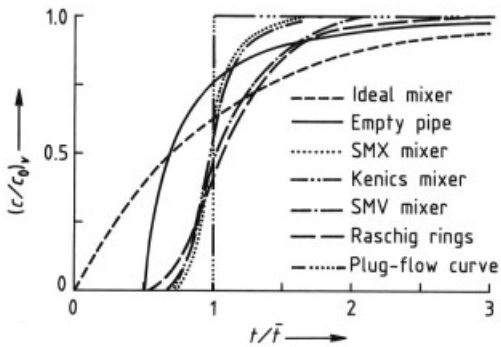

Figure 26. Residence-time spectra for the laminar flow regime [26]

8. Mixing of Non-Newtonian Fluids

The flow properties of many liquids and gases used in industry cannot be described by Newton's law $\tau=\eta\,\dot{\gamma}$, since the viscosity $\eta=\tau/\dot{\gamma}$ of

such media depends on the shear rate $\dot{\gamma}$. These non-Newtonian fluids often exhibit elastic behavior in addition to viscous behavior [36], [37]. Non-Newtonian flow behavior can alter the hydrodynamic conditions so much that the parameters given above for the laminar and transition flow regimes cannot be used in calculating mixing performance, pressure drop, and power dissipation. Only in strongly turbulent flow are conditions roughly the same as for Newtonian fluids.

To describe the behavior of non-Newtonian fluids, all material parameters needed to determine flow properties must be included, instead of the constant viscosity that is sufficient for Newtonian fluids. These parameters are obtained from the material functions in dimensionless form [36–38]. Because the rheological equations of state are multiparametric, the set of independent variables becomes larger. When elastic properties are negligible, at least one additional process parameter must be included (a parameter formed from the process variables and the geometrical and rheological parameters), and in the simplest case where the power law

$$\tau = K\gamma^m \tag{32}$$

applies, the additional material parameter m is necessary to describe viscous flow behavior. If the material functions are complicated, other material parameters enter into the calculation [36], [37]. The Weissenberg number Wi is useful in describing elastic flow properties. It represents the ratio of the first normal stress difference to the shear stress and is a measure of the ratio of elastic to viscous properties [36].

The calculation of engineering parameters for non-Newtonian fluids is made more difficult by the additional parameters. For this reason, many attempts have been made to use a representative viscosity and the Weissenberg number, both determined from the material functions for a shear rate representative of the process under consideration. If this can be done, the set of parameters for the case of negligible elastic properties will be same as for Newtonian fluids. When the elastic properties are significant, the Weissenberg number must also be included in the calculation.

8.1. Mixing Behavior

8.1.1. Pipes and Jet Mixers

For fully developed turbulent flow ($Re > 10^4$), the mixing behavior should be similar to that for Newtonian fluids. The Reynolds number is calculated with a representative viscosity read off the flow curve at the following representative shear rate:

$$\bar{\gamma} = \sqrt{\frac{2}{3}} \cdot \frac{8u}{D} \tag{33}$$

Markedly viscoelastic liquids, such as polyacrylamide (PAA) solutions, are an exception. In these liquids, a reduction in the pipe friction coefficients is observed even in the turbulent flow regime [33], so that a worsening of pipe mixing is also expected.

8.1.2. Static Mixers

For static mixers, the representative shear rate is derived from the model of Langer et al. [33]:

$$\bar{\gamma} = \sqrt{\frac{2}{3}} \cdot \frac{t_s}{\varepsilon} \cdot \frac{8u}{D_h} \tag{34}$$

Here ε is the ratio of the free volume V_f in the mixer to the total volume V; t_s is a torsion factor determined empirically by fitting to the pressure-drop curves; and $D_h = 4\,V_f/A_0$ is the hydraulic diameter obtained from the wetted surface area A_0 of the static mixer, consisting of the surface area of the pipe plus internals.

For fluids exhibiting pseudoplastic viscosity, such as carboxymethyl cellulose (CMC) solutions, Langer's results indicate that the mixing characteristics in Kenics, SMX-Sulzer, and SMV-Sulzer mixers are the same as in Newtonian systems (Fig. 27) if the shear rate defined by Equation (34) and the following modified Reynolds number Re' are used [33]:

$$Re' = Re\frac{D_h}{4\varepsilon D} = \frac{uD_h}{4\varepsilon\bar{\nu}_s} \tag{35}$$

For elastic fluids, mixing is much slower (Fig. 27). Correlation of all measured results, including the data for elastic fluids, requires the additional incorporation of at least the Weissenberg number.

Table 2. Geometric dimensions of static mixers characterized in Figures and (see Eqs. 34 and 35)

Mixer	D, mm	L_e/D	D_h/D	ε	t_s
SMV-Sulzer-2 Y	9.6	1.06	0.041	0.697	1.4
SMV-Sulzer-8 Y	42	0.95	0.042	0.812	1.4
SMX-Sulzer	38.6	1.75	0.0738	0.913	1.79

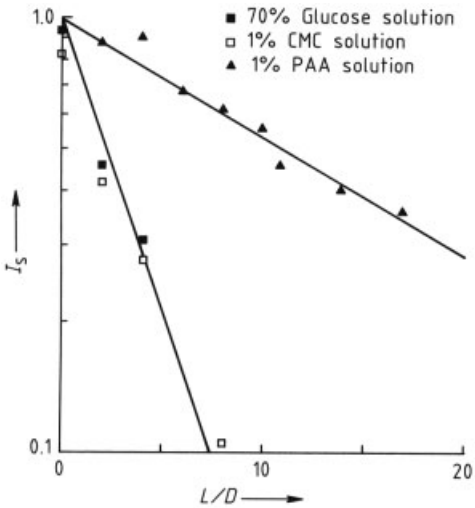

Figure 27. Mixing performance curves for an SMX-Sulzer mixer at $Re' = 1$ [33]
See Table 2 for geometric dimensions.

8.1.3. Dynamic Mixers

The only studies on the effect of non-Newtonian fluid properties have been carried out in batch mixers. Several deductions applicable to continuous mixers can be derived: the representative shear rate based on the METZNER and OTTO rotational velocity concept [39]

$$\bar{\gamma}_n = kn \tag{36}$$

the wall shear gradient between two concentric cylinders, which takes the following form if the power law of Equation (32) applies,

$$\bar{\gamma}_W = \frac{4\pi n}{m}\left[1-(d/D)^{2/m}\right]^{-1} \tag{37}$$

and the power concept [40], [41],

$$\bar{\gamma}_P = \sqrt{\frac{P_{tot}}{\varrho V}\cdot\frac{1}{\bar{\nu}_P}} \tag{38}$$

where P_{tot} is the total power dissipated in the mixing volume V.

In Equation (36), k is an empirical constant obtained by fitting a curve to the Newton numbers in the laminar flow regime. Its value depends, to some extent, on the type and geometry of the stirrer and on flow behavior (see Section 8.3).

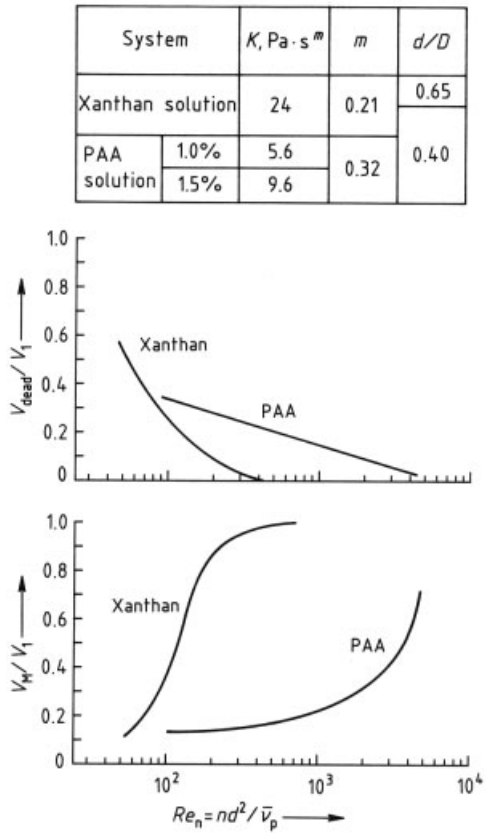

Figure 28. Unstirred volume V_{dead} and well-mixed volume V_M for stirring of high-viscosity non-Newtonian fluids with a rusthon turbine (V_1 = capacity of stirred reactor) [41]

For laminar and transition flow, dead spots (volume V_{dead}) and poorly mixed regions can form, especially when the stirrers are small and do not cause forced circulation of the reactor contents [41]. This behavior has been observed for liquids with pseudoplastic behavior at Reynolds numbers $Re_n = nd^2/\nu < 500$ and in more elastic liquids at $Re_n < 5\times10^3$. As Figure 28 shows, the effective volume of the continuous mixer – and thus the residence time – may

be decreased, so that adequate mixing of the total content requires higher stirrer speeds than are needed with Newtonian liquids.

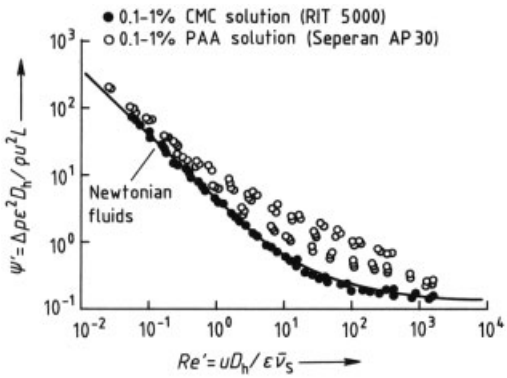

Figure 29. Pressure-drop characteristic for an SMV-Sulzer mixer [33]
See Table 2 for geometric dimensions.

Changes in the mixing behavior of CMC and PAA solutions have also been found by MÜLLER [42] and PÖRTNER [43]. An attempt to obtain improved correlation of the mixing numbers by using the Weissenberg number has not, however, led to satisfactory results. LIEPE has summarized other findings on the mixing behavior of non-Newtonian fluids [6].

8.2. Pressure Drop for Non-Newtonians Fluids

The pressure-drop curves for Newtonian fluids can be used in approximate calculations for non-Newtonian fluids, provided no unmixed regions occur in the mixer and boundary-layer effects can be neglected (as in jet mixers and dynamic mixers). This applies to turbulent flow, which exists above a certain Reynolds number that depends on flow properties (Fig. 28). Practically turbulent mixing conditions exist when $V_{\mathrm{dead}} \to 0$ and $\mathrm{V_M/V_1} \to 1$. The shear rate as defined in Equations (33), (34), and (36) – (38) can be used to determine the representative viscosity.

In pipes and static mixers for liquids with pseudoplastic behavior, the same curves can be used as for non-Newtonian liquids in all flow regimes (e.g., see the values for CMC solutions in SMV-Sulzer mixers, Fig. 29) if the shear rate is calculated with Equation (33) (pipes) or (34) (static mixers). In Figure 29, the modified Reynolds number Re' from Equation (35), which was introduced by LANGER on the basis of a model [33], is used along with the pressure-drop number ψ':

$$\psi' = \psi \frac{\varepsilon^2 D_{\mathrm{h}}}{8D} \tag{39}$$

Accordingly, viscoelastic substances (which generally exhibit a decrease in resistance in turbulent pipe flow [33], [44] and an increase in resistance in static mixers) have no correlation with values for Newtonian media. See, for example, the values in Figure 29 for PAA (Separan AP 30) solutions.

LANGER first obtained a more general correlation of the pressure-drop characteristics for Newtonian, pseudoplastic, and viscoelastic fluids by using the Weissenberg number [33]. The following expression gives a satisfactory fit to the data of Figure 29:

$$\psi' Re' = 4 + 0.14 Re' + 0.6 \sqrt{Re' Wi} \tag{40}$$

The curve for the SMX-Sulzer mixer is described by [33]

$$\psi' Re' = 6.4 + 0.3 Re' + 1.7 \sqrt{Re' Wi} \tag{41}$$

Even these representations still include a perceptible dependence on the elastic properties of the solution in the transition regime. (See fit of experimental data in [31]).

8.3. Power Requirements of Dynamic Mixers

The Newton numbers for the laminar regime (e.g., see Figs. 22 and 23) can also be used for some non-Newtonian fluids if the representative viscosity ν_{n} used for calculating the stirrer Reynolds number $Re_{\mathrm{n}} = nd^2/\nu_{\mathrm{n}}$ is obtained by applying the Metzner – Otto concept $\gamma_{\mathrm{n}} = kn$ to the flow curve [6], [39], [45], [46]. If the power law $\tau = K\gamma^{\mathrm{m}}$ holds, the Reynolds number is given by

$$Re_{\mathrm{n}} = \frac{nd^2}{\bar{\nu}_{\mathrm{n}}} = \frac{k^{1-m} n^{2-m} d^2 \varrho}{K} \tag{42}$$

The empirical factor k depends on the type and geometry of the stirrer and on flow properties.

For agitators with $d/D < 0.5$, the geometry has relatively little effect. Values between 10 and 13 have been found for k in pseudoplastic ($m<1$) and plastic fluids [39], [45]. Larger k values are obtained at larger diameter ratios and, especially, for agitators moving near the vessel walls [45], [47], [48]. For example, HIMMELSBACH finds k=15 for the INTERMIG stirrer with $d/D = 0.65$ [48]. In dilatant liquids ($m>1$), CALDERBANK reports $k=22(D/d)^2/[(D/d)^2-1]$ [45].

In the transition regime ($Re_n>20-50$ up to very large values, 10^4-10^5), the Ne values for non-Newtonian media differ from those for Newtonian systems [6], [46]. The differences are especially large for liquids with marked elastic properties ($Wi>2$).

9. Selection and Comparison of Continuous Mixers

Equipment selection is determined by the required residence time, the residence-time spectrum, the Reynolds number range, and the energy consumption. Suitable equipment for short residence times includes jet mixers, mixers with stationary internals, and in-line mixers, and for long residence times, tanks with jet or agitator systems.

The residence-time distribution can be adjusted by varying the number of tanks or stirrers and by using special internals in the mixers.

In contrast to static mixers and mixers with moving internals, jet mixers and tanks with jet systems are appropriate for mixing only in the turbulent flow regime ($Re>10^3$).

In small-capacity mixers, constant metering of the components to be mixed is necessary to avoid concentration fluctuations over time. The ratio of mean flow velocity u to the frequency f_P of feed concentration fluctuations must satisfy $u/f_P<D$ [6]. If this condition is not met, a larger-capacity mixer should be selected or a well-mixed buffer tank must be added.

Table 3 summarizes important advantages and disadvantages of various continuous mixers. Jet mixers and static mixers should be used chiefly for pure mixing tasks, where a longer residence time is not required to smooth out variations in concentration over time. Dynamic mixers are more flexible because the throughput and the mixing performance are decoupled; they are preferable when additional unit operations, such as suspension and dispersion, are to be performed.

9.1. Parameters for Energetic Comparison

The user's interest is in mixers that offer the lowest possible investment and operating costs. These costs depend on, among other factors, the mixing performance required and the size of the equipment (volume, diameter, length).

Mixers that can perform a given task under given constraints with low power consumption are therefore favored. One approach to mixer selection is the energetic comparison of systems. Such a comparison is easily understood if it is done with dimensionless groups derived from the total power consumption per unit mass of the components processed $P_{tot}/\varrho\dot{V}$. If only mixers with no moving internals are considered first, the following relation can be derived from Equation (8 a):

$$\frac{P}{\varrho\dot{V}}=[Ne]\frac{\dot{V}^2}{D^4}=\left[Ne\left(\frac{L}{D}\right)^4\right]\frac{\dot{V}^2}{L^4}$$
$$=\left[Ne\frac{V^{4/3}}{D^4}\right]\frac{\dot{V}^2}{V^{4/3}} \tag{43}$$

The dimensionless groups in square brackets were first introduced by HENZLER [4] and later used by other workers [24–26], [29], [32]. Which of them is used in the energetic comparison depends on the question to be answered. In the design of a continuous mixer, the volume flow rate $\dot{V}$ is always given. The groups in Equation (43) provide a measure of power consumption per unit mass if one of the following is also given

Diameter D:

$$c_{PHD}=Ne=\frac{PD^4}{\varrho\dot{V}^3} \tag{44}$$

Mixer length L:

$$c_{PHL}=Ne\left(\frac{L}{D}\right)^4=\frac{PL^4}{\varrho\dot{V}^3} \tag{45}$$

Volume V, or mean residence time $\bar{t}$:

$$c_{PHV}=Ne\frac{V^{4/3}}{D^4}=\frac{PV^{4/3}}{\varrho\dot{V}^3}=\frac{P\bar{t}^{4/3}}{\varrho\dot{V}^{5/3}} \tag{46}$$

Table 3. Properties, advantages, and disadvantages of continuous mixers

Type	Advantages	Disadvantages
Jet mixer	simple design	suitable only for turbulent regime
Tank with jet system	no moving parts low maintenance costs	only partly suitable for large density differences and for production of suspensions and dispersions
Static mixer	simple design no moving parts low maintenance costs narrow residence-time spectrum low energy consumption in turbulent regime	only partly suitable for large density differences and for production of suspension and dispersions
Dynamic mixer	decoupling of mixing power and throughput versatility: large density differences, production of suspensions and dispersions residence-time spectrum controllable through number of stages	complex design moving agitators higher maintenance costs
Stirred tank/stirred column Tank with jet system	smoothening of concentration fluctuations over time	

Because the Newton number and the mixing length may depend on the Reynolds number in various flow regimes, these dimensionless groups should be expressed as functions of the modified Reynolds numbers:

$$c_{\mathrm{PHD}}=f\left(Re'_{\mathrm{D}*}\right) \text{ with } Re'_{\mathrm{D}*}=\frac{\dot{V}}{D\nu} \tag{47}$$

$$c_{\mathrm{PHL}}=f\left(Re'_{\mathrm{L}*}\right) \text{ with } Re'_{\mathrm{L}*}=\frac{\dot{V}}{L\nu} \tag{48}$$

$$c_{\mathrm{PHV}}=f\left(Re'_{\mathrm{V}*}\right) \text{ with } Re'_{\mathrm{V}*}=\frac{\dot{V}}{V^{1/3}\nu}=\frac{\dot{V}^{2/3}}{\bar{t}^{1/3}\nu} \tag{49}$$

For static mixers, which exhibit a constant mixing length in the laminar flow regime, the product of any mixing power number c_{PH} and the corresponding Reynolds number is a constant because $Ne \cdot Re$ is constant. Therefore, characterizing the mixing performance in these terms is meaningful. If the pressure-drop coefficient ψ as defined by Equations (23) and (24) is introduced, the following values can be obtained from Equations (44) – (46) for static mixers in the laminar flow regime:

$$c_{\mathrm{PHD}}\, Re'_{\mathrm{D}*}=\frac{2}{\pi}\cdot\psi Re\cdot\frac{L}{D}=\frac{PD^3}{\eta\dot{V}^2} \tag{50}$$

$$c_{\mathrm{PHL}}\, Re'_{\mathrm{L}*}=\frac{2}{\pi}\cdot\psi Re\left(\frac{L}{D}\right)^4=\frac{PL^3}{\eta\dot{V}^2} \tag{51}$$

$$c_{\mathrm{PHV}}\, Re'_{\mathrm{V}*}=\frac{1}{2}\psi Re\left(\frac{L}{D}\right)^2=\frac{PV}{\eta\dot{V}^2}=\frac{P\bar{t}}{\eta\dot{V}} \tag{52}$$

If the energy for mixing is supplied by both the pipe flow and moving stirrers, the energy used in the comparison will be the sum of these two contributions. Consider, for example, the case of a given mixer volume. As with Equation (43), the following can be obtained from Equation (8) after introduction of the mixing number c_{H}:

$$\frac{P_{\mathrm{tot}}}{\varrho\dot{V}}=\left(Ne\cdot\frac{V^{4/3}}{D^4}+Ne_{\mathrm{n}}c_{\mathrm{H}}^3\cdot\frac{d^5}{\dot{V}^{5/3}}\right)\frac{\dot{V}^2}{V^{3/4}} \tag{53}$$

where

$$c_{\mathrm{PHV}}=\left(Ne\cdot\frac{\dot{V}^{4/3}}{D^4}+Ne_{\mathrm{n}}c_{\mathrm{H}}^3\cdot\frac{d^5}{V^{5/3}}\right)$$
$$=\frac{P_{\mathrm{tot}}V^{4/3}}{\varrho\dot{V}^3}=\frac{P_{\mathrm{tot}}\bar{t}^{4/3}}{\varrho\dot{V}^{5/3}}$$

In the mixing power number c_{PHV}, the right-hand term accounts for the transformed flow energy, while the left-hand term represents the stirrer energy. These numbers and others similarly formed are thus suitable for comparisons among mixing pipes, static mixers, and dynamic mixers.

9.2. Energetic Comparison

An exact energetic comparison between systems is possible only if all the systems have been

investigated with the same measurement techniques, comparable or interconvertible definitions of the degree of inhomogeneity, and (in the case of the method of final decolorization) by the same experimenter [4]. Because only a few measurements are available, and fewer still have been performed under these conditions, only a limited comparison between systems is possible.

Tables 4 and 5 and Figures 30, 31, 32 present a comparison between various systems based on the dimensionless groups discussed above.

	Mixer	I
Curves 1 and 2	Jet mixers	$I_2 = 0.05$
S	Sulzer	$I_S = 10^{-4}$
K	Kenics	
P	Baffles	
R	Smooth pipe	

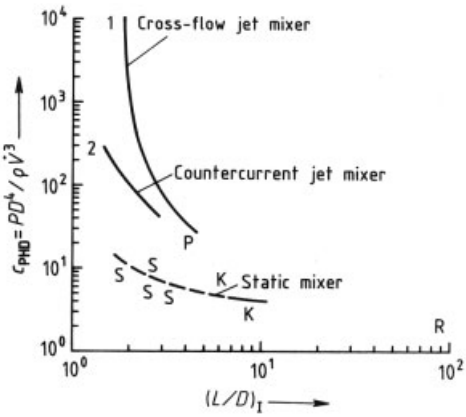

Figure 30. Energetic comparison of mixers in the turbulent flow regime for a given mixer diameter
Curves 1 and 2: Data from Figures 6 and 7
S – R: Data from Table 1 and Figure 10

9.2.1. Mixers with Fixed Internals

In the turbulent flow regime, the energy requirement for mixing decreases with increasing mixing length (Fig. 30). Because turbulence plays the decisive role, similar curves are obtained for comparable mixer types. In the turbulent case, static mixers with comparable mixing lengths exhibit a lower energy requirement than mixers with baffles or jet mixers. The lowest energy requirement is that of the hydraulically smooth pipe, although the mixing length is extremely long.

No.	Stirrer	d/D	z	$\Delta h/D$	L/D
1a	Propeller	0.33	2	0.78	1.56
1b			5		3.9
2	Disk turbine		2		1.56
3	MIG	0.5			
4	Cross-beam paddle	0.75			
5		0.5			
6a				0.3	0.6
6b			5		1.5

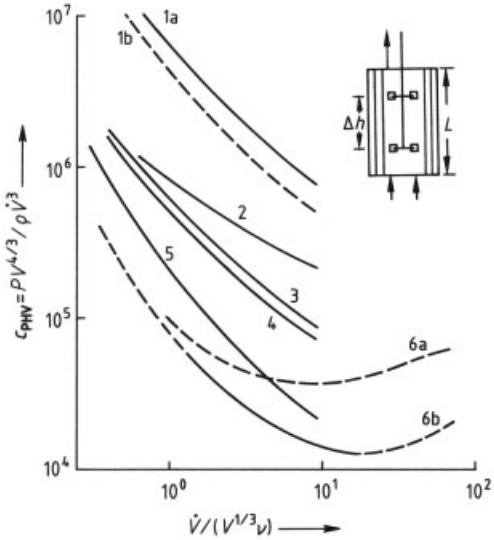

Figure 31. Energetic comparison of stirred columns for a given mixer volume, after Figures 16 and 22

If the length or volume is given, static mixers again offer the lowest energy consumption; for a given energy consumption, conversely, they have the shortest length or the smallest volume (Table 4).

At lower Reynolds numbers ($Re < 2300$), pipe flow does not afford sufficient mixing, nor does jet action (see Eq. 14 for the limiting Reynolds numbers). In these flow regimes, however, static mixers can be used. Lack of turbulence means that mixing is based on efficient splitting and layering of fluid elements, and the efficiency of these processes varies widely from one mixer type to another. Accordingly, the range of energy consumptions in laminar flow is particularly broad (Table 5). The SMF-Sulzer, SMXL-Sulzer, and Kenics mixers have the lowest en-

Table 4. Energetic comparison of mixers in the turbulent flow regime: $I = I_2 = 0.05$ for jet mixers, $I = I_S = 10^{-4}$ for other mixers

Energetic comparison			Diameter	Length	Volume		Data from	
Mixer		$(L/D)_I$	c_{PHD}	$c_{PHL} \times 10^2$	c_{PHV}	Re		
Cross-flow jet mixer	$d_i/D = 0.3$	4.5	25	102	135	$> 5 \times 10^3$	Figure 7 and Table 1	
Countercurrent jet mixer	$d_i/D = 0.175$	1.5	216	11	269			
	$d_i/D = 0.25$	2.5	52	20	127			
SMV-Sulzer		1.9	10.5	1.4	18	$> 5 \times 10^3$	Table 1 and Figure 10	
SMF-Sulzer		3	4.9	3.9	15	10^3		[21]
SMXL-Sulzer		3.6	5.8	9.7	23			[28]
Kenics		7.20	6.5	175	65	6000		[24]
Baffles (see Fig. 9 A)		4.25	24.1	79	120	$> 5 \times 10^3$		
Kenics		8.5	3.5	181	44	6000		[11]
Smooth pipe		85	2.5	1 300 000	670	6000		

Table 5. Energetic comparison of static mixers in the laminar flow regime: $Re < 2$, $I_S = 10^{-4}$ (data from Fig.)

Energetic comparison			Diameter	Length	Volume
Mixer	$(L/D)_{I_S}$	$\psi\, Re$	$c_{PHD}\, Re'_{D*} \times 10^{-3}$	$c_{PHL}\, Re'_{L*} \times 10^{-6}$	$c_{PHV}\, Re'_{V*} \times 10^{-4}$
SMX-Sulzer	9	2470	14	10	10
SMV-Sulzer	18	2860	33	190	46
SMR-Sulzer	23	3080	45	550	81
SMF-Sulzer	25	360	6	90	11
SMXL-Sulzer	26	490	8	140	17
Kenics	29	440	8	200	19
Helax	42	600	16	1190	53

ergy requirement. The SMX-Sulzer and SMF-Sulzer mixers have the smallest length and volume for a given energy input.

9.2.2. Dynamic Mixers

The comparison in Figures 31 and 32 and the discussion in Chapter 6 lead to the following recommendations for the design of dynamic mixers:

1) The stirrers should be arranged so that the dominant mixing motion of the agitator is transverse to the net flow.
2) As in batch homogenizing, the use of baffles is advantageous.
3) In single-stage stirred tanks, short-circuit flows must be prevented by appropriate placement of the inlet and outlet.
4) In multistage stirred tanks, a short stage-to-stage spacing and a large number of stages are recommended.
5) In the transition and turbulent flow regimes ($Re = u\,D/\nu > 10$), short-circuit flows can be prevented by the use of restrictors and baffles between the agitators. This applies especially to the use of jet systems for injection of individual components.
6) As in batch homogenizing, screw or helical stirrers (Fig. 17) should be used in the laminar and creeping flow range ($Re < 5$).

9.2.3. Comparison between Mixing Pipes and Dynamic Mixers

Results for dynamic mixers are few, so only the comparison performed in [5] can be discussed. The use of data for batch mixers [24], [26], [32] is not appropriate because optimal continuous mixers differ in design, and the comparison involves an assumption about the residence time required for sufficient mixing (needed for calculation of mixing number c_H). The assumption $t_H = \bar{t}$ made in the literature may not be realistic.

When the mixing power numbers are plotted versus the Reynolds number as in Figure 32, all mixers give similar curves. Although the mixing power required is constant in the turbulent flow regime ($Re > 5 \times 10^3$), a very sharp increase

occurs in the transition regime ($2<Re<500$) because of decreasing turbulence. In this flow regime, because of inertial effects, unsuitable mixer types (such as the Kenics mixer or the stirred tank without internals) exhibit much worse mixing performance than other dynamic or static mixers.

No.	System	Stirrer	d/D	d_i/D	d_B/D	L/D
1		none		0.25		
2					0.35	1.8
3	Kenics pipe					
4		pitched tube	0.4	0.30	0.40	5.5
5		propeller	0.6			
6		disk turbine				
7			0.4			
8		disk turbine	0.4	0.052	0.37	1.5

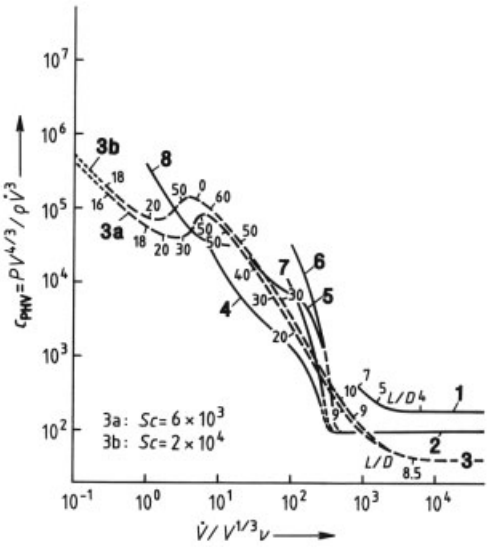

Figure 32. Energetic comparison of mixers with and without moving internals for a given mixer volume
Data for Kenics mixer (I_2=0.06) from [11]; other values from Figures 19 and 21 and from [5] (I_2=0.05, $\dot{V}_1/\dot{V}_2=\varrho_1/\varrho_2=\nu_1/\nu_2\approx1$)

In the laminar regime ($Re<2$), mixer systems that cause forced flow of the components are characterized by mixing power c_{PH} proportional to $1/Re$. This behavior is found for static mixers, as shown by the dependence of mixing length on Re (Fig. 12). Similar to the results for batch homogenization [4], [6], the same proportionality is expected for dynamic mixers with forced-flow stirrers, such as screw and helical stirrers (Fig. 17). Dynamic mixers of these types can deliver similar mixing performance to static mixers at equal power consumption.

As Figure 32 shows, static mixers are markedly superior in the turbulent flow regime ($Re>10^3$). The forced motion of fluid elements obviously leads to large-scale motions that favor homogenization.

10. References

1. J. Bourne, P. Rys, K. Suler: *Chem. Eng. Sci.* **32** (1977) 711 – 716.
2. K. H. Tebel, H.-O. May: "Der Freistrahlreaktor – ein effektives Reaktordesign zur Unterdrückung von Selektivitätsverlusten durch schnelle, unerwünschte Folgereaktionen," *Chem. Ing. Tech.* **60** (1988) 11.
3. P. V. Danckwerts: "The Definition and Measurement of some Characteristics of Mixtures," *Appl. Sci. Res.* **A 3** (1953) 279 – 296.
4. H.-J. Henzler: "Untersuchungen zum Homogenisieren von Flüssigkeiten und Gasen," *VDI-Forschungsh.* **587** (1978) 1 – 60.
5. H.-J. Henzler: "Eignung von kontinuierlich durchströmten Mischern zum Homogenisieren," *Chem. Ing. Tech.* **51** (1979) no. 1, 1 – 8.
6. F. Liepe: "Verfahrenstechnische Berechnungsmethoden," *Stoffvereinigen in fluiden Phasen,* VCH Verlagsgesellschaft, Weinheim 1988; *Stoffvereinigen in fluiden Phasen,* **part 4/1,** VEB Verlag für Grundstoffindustrie, Leipzig 1978.
7. W. Hiby: "Mischung in turbulenter Rohrströmung," *Verfahrenstechnik* (Mainz) 4 (1970) no. 12, 538 – 543.
8. G. V. Evans: "A Study of Diffusion in Turbulent Pipe Flow," *Trans. ASME* (1967) 624 – 632.
9. H. Blasius: "Das Ähnlichkeitsgesetz bei Reibungsvorgängen in Flüssigkeiten," *Forsch. Ingenieurwes.* 1913, H 131.
10. K. H. Hartung: "Untersuchungen technischer Misch-vorgänge mittels pH-Indikatoren," Dissertation, TH Aachen, 1975.

11. K. H. Hartung, J. W. Hiby: "Beschleunigung der turbulenten Mischung in Rohren," *Chem. Ing. Tech.* **44** (1972) no. 18, 1051 – 1056.
12. H.-J. Henzler: "Homogenisieren von Flüssigkeiten und Gasen in Strahlmischern," *Chem. Tech.* (Leipzig) 27 (1957) no. 7, 399 – 401.
13. T. Maruyama et al.: "Optimum Conditions for Jet Mixing in Turbulent Pipe Flow," *Int. Chem. Eng.* **23** (1983) no. 4, 707 – 716.
14. F. Liepe: *Taschenbuch Maschinenbau,* "Strömungslehre", vol. 1/I, VEB Verlag der Technik, Leipzig 1985.
15. J. Idelcik: *Handbuch der hydraulischen Widerstände,* Staatlicher Verlag für Energetik, Moscow 1960.
16. M. H. Pahl, E. Muschelknautz: "Einsatz und Auslegung statischer Mischer," *Chem. Ing. Tech.* **51** (1979) no. 5, 347 – 364.
17. M. H. Pahl, E. Muschelknautz: "Statische Mischer und ihre Auslegung," *Chem. Ing. Tech.* **52** (1980) no. 4, 285 – 291.
18. *25 Jahre Statisches Mischen,* preprints, Haus der Technik, Essen Oct. 16 – 17,1991.
19. Sulzer, company brochure, Winterthur, Schweiz 1991.
20. K.-H. Hartung, J. W. Hiby: "Kontinuierliche Laminarmischung im Kenics-Rohr," *Chem. Ing. Tech.* **47** (1957) 7.
21. H. Falk: "Vermischen von niedrigviskosen Medien," in [18]
22. W. A. Tauscher, F. A. Streiff: CEP 4 (1979), 61 – 65. W. Tauscher, F. Streiff, R. Bürgi: "Statisches Mischen von Gasströmen in großen Leitungen," *VGB Kraftwerkstechnik* **60** (1980) no. 4, 290 – 296.
23. V. Novak, V. Jandourek, F. Rieger: "Homogenization Efficiency of Motionless Mixers," *5th European Conference on Mixing in Würzburg,* June 10 – 12, BHRA, Cranfield 1985.
24. F. Streiff in M. Pahl, K. Sommer, F. Streiff, A. Limper (eds.): *Mischen in Kunststoffen und Kunststoffprodukten,* vol. 2, VDI Gesellschaft Kunststofftechnik, Düsseldorf 1991.
25. F. A. Streiff, G. Schneider: *Mischen von Kunststoffen,* VDI-Verlag, Düsseldorf 1983.
26. M. Pahl: *Mischen beim Herstellen und Verarbeiten von Kunststoffen,* VDI-Verlag, Düsseldorf 1986.
27. J. Kabatek, P. Ditl, V. Novak: "Helax – a New Type of Static Mixer," *Chem. Eng. Process.* **25** (1989) 59 – 64.
28. H. Kalbitz: "Einfluß statischer Mischer auf den Wärmeübergang und den Druckverlust in Rohrwärmetauschern," *Chem. Ing. Tech.* **63** (1991) no. 3, 270 – 271; "Wirkung statischer Mischer auf die Strömung, den Wärmeübergang und den Druckverlust in Rohrwärmeaustauschern," Dissertation, TU Braunschweig, 1990.
29. F. Streiff: "Statische Mischer mit großer Anpassungsfähigkeit," *Chem. Ing. Tech.* **52** (1980) no. 6, 520 – 522.
30. J. H. Sununu: "Heat Transfer with Static Mixer Systems," Technical Report 1002, Kenics-Corporation, Danvers, Mass., 1970.
31. G. Langer, U. Werner, M. S. Ajam, *Fortschr. Ber. VDI Z., Reihe 3,* 1987 no. 145.
32. F. A. Streiff: "Adapted Motionless Mixer Design," *3rd Conference on Mixing in York,* April 4 – 6, 1979.
33. G. Langer: *25 Jahre Statisches Mischen,* preprints, Haus der Technik, Essen, Oct. 16 – 17, 1991.
34. P. Todtenhaupt: "Homogenisieren von Flüssigkeiten mit kontinuierlich betriebenen Rührkolonnen," Dissertation, TU München.
35. H. Weissengerber: "Kontinuierliches Mischen in Rohrleitungen mit bewegten Einbauten," *Mitteilungen des Forschungszentrums für Chemieanlagen Dresden* **9** (1969) no. 4, 115 – 121.
36. H.-J. Henzler: "Rheologische Stoffeigenschaften – Erklärung, Messung und Bedeutung," *Chem. Ing. Tech.* **60** (1988) no. 1, 1 – 8.
37. H.-J. Henzler, E. E. Schäfer: "Viskose und elastische Eigenschaften von Fermentationslösungen," *Chem. Ing. Tech.* **59** (1987) 12.
38. J. Pawlowski: *Ähnlichkeitstheorie in der physikalisch-technischen Forschung,* Springer-Verlag, Heidelberg 1980.
39. A. B. Metzner, R. E. Otto, "Agitation of non-Newtonian Fluids," *AIChE J.* **3** (1957) no. 1, 3 – 10.
40. H.-J. Henzler, J. Kauling: "Scale-up of Mass Transfer in Highly Viscous Liquids," *5th European Conference on Mixing,* paper 30, BHRA, Cranfield, 1985, pp. 303 – 312.
41. H.-J. Henzler, G. Obernosterer: "Effect of Mixing Behaviour on Gas-Liquid Mass Transfer in Highly Viscous Stirred non-Newtonian Liquids," *Chem. Eng. Technol.* **14** (1991) no. 1, 1 – 10.
42. W. Müller: "Untersuchungen von Homogenisiervorgängen in nicht-Newtonschen Flüssigkeiten," *Fortschr. Ber. VDI Z., Reihe 3* 1985 no. 103, 1 – 110.

43. R. Pörtner: "Zur Homogenisierung Newtonscher, strukturviskoser und viskoelastischer Flüssigkeiten in Rührbehältern," Dissertation, Universität Dortmund, 1989.
44. G. Wellinghoff: "Zum Einfluß von Reibungsminderern auf den Druckverlust und den Wärmeübergang in Fernwärmeleitungen," Dissertation, Universität Dortmund, 1985.
45. P. H. Calderbank, M. B. Moo-Young: "The Prediction of Power Consumption in the Agitation of non-Newtonian Fluids," *Trans. Inst. Chem. Eng.* **39** (1961) 338 – 347.
46. H. Höcker: "Untersuchungen zum Leistungsbedarf und Stoffübergang in Rührreaktoren," Dissertation, Dortmund, 1979.
47. S. Yüce, D. Schlegel: "Wärmeübergang an nicht-Newtonsche Substanzen in Statischen Mischern," *Chem. Ing. Tech.* **62** (1990) no. 5, 418 – 419.
48. W. Himmelsbach: "Untersuchungen zum Stoff- und Wärmeübergang sowie zur Homogenisierung in geührten Fermentern," *Chem. Ing. Tech.* **57** (1985) 6.

High-Pressure Technology

Gerhard Vetter, Universität Erlangen-Nürnberg, Federal Republic of Germany (Chaps. 1 and 3)

Erich Karl, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Chaps. 2 and 4)

1. Applications of High-Pressures . . 2
2. Pressure Vessels 3
2.1. Solid-Wall Vessels 3
2.2. Multiwall Vessels 3
2.2.1. Designs with a Purely Mechanical Joint 3
2.2.2. Welded Designs (Layered-Wall Vessels) 4
2.3. Strength Calculations 5
2.3.1. Cylindrical Wall 5
2.3.1.1. Stresses Due to Internal Pressure . . 5
2.3.1.2. Initial Stresses in Shrink Joints . . . 6
2.3.1.3. Residual Stresses Due to Autofrettage 6
2.3.1.4. Thermal Stresses 7
2.3.1.5. Avoiding Brittle Fracture 7
2.3.2. Cylindrical Wall with Radial Bore . 7
2.3.3. End Pieces 8
2.4. Corrosion in High-Pressure Plant 8
2.5. Material Selection 9
2.6. Design Details 9
2.6.1. Corrosion Protection 9
2.6.2. Covers and Their Attachment 10
2.6.3. Cover Seals 11
2.6.4. Seals for Piping 12
2.7. Example of the Design of Pressure Vessels 12
3. High-Pressure Machinery for Chemical Plants 13
3.1. Special Features of High-Pressure Machines 13
3.2. Creation of Pressure by Pumps and Compressors 14
3.3. Pumps 15
3.3.1. Reciprocating Displacement Pumps 17
3.3.1.1. Metering Pumps 17
3.3.1.2. Transfer Pumps 19
3.3.2. Rotary Displacement Pumps 20
3.3.3. Centrifugal Pumps 21
3.3.3.1. Multistage Centrifugal Pumps . . . 23
3.3.3.2. High-Speed Centrifugal Pumps . . . 23
3.3.3.3. Hermetic Centrifugal Pumps 24
3.4. Compressors 26
3.4.1. Piston Compressors 28
3.4.2. Diaphragm Compressors, Laboratory High-Pressure Compressors . . 28
3.4.3. Turbo Compressors 30
3.5. Other High-Pressure Machines . 30
3.6. Special Problems Involving High-Pressure Machines 31
3.6.1. Strength of the Components 31
3.6.2. Seals 33
3.6.3. Wear and Vibration 36
4. Piping and Fittings 36
5. References 38

1. Applications of High-Pressures

Internal combustion engines, thermal power plants, hydraulic power transmission, the arms industry, the process industries, and the manufacturing industries commonly use high pressures. Process and manufacturing plants use pressures in the range 100 – 4000 bar. In research pressures up to 10 000 bar may be employed. High pressures are used to improve the efficiency of an operation in some way [1].

In chemical reaction engineering, high pressure may create a favorable reaction environment or determine the reaction phases. Pressure may decisively affect the reaction rate, state of equilibrium, yield, and products. High pressure is particularly advantageous for reactions in which a reduction in volume takes place [2].

Typical industrial high-pressure chemical reactions are hydrogenation, amination, polymerization, carbonylation, oxo synthesis, oligomerization, oxidation, and cracking (Table 1). Some well-known products of high-pressure processes are methanol; ammonia; urea; propionic, acetic, and terephthalic acids; butyraldehyde (oxo synthesis); butanediol; and polyethylene. In many processes, e.g., extraction of ores, high pressure ensures that reaction takes place in the liquid phase.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b04_587

Table 1. High-pressure processes

Polyethylene	< 3000 bar
Acetic acid	< 700 bar
Methanol, ammonia, urea, oxo synthesis, butanol, coal and heavy oil hydrogenation, hydrocracking, steam cracking, heavy water	< 300 bar
Fischer – Tropsch synthesis, aluminum alkyls (Ziegler), alcohols (Alfol), gas treating	< 200 bar
Ore digestion (bauxite), coal pressure gasification	< 150 bar
Supercritical extraction	< 300 bar
Reverse osmosis	< 100 bar
Spray drying	< 300 bar

Some separation processes, e.g., supercritical gas extraction [3], become more effective at high pressure because of favorable changes in the solubilities. Other separation processes – reverse osmosis, ultrafiltration, and spray drying – require high pressures for hydromechanical reasons.

The number of manufacturing processes that use high pressure as a tool or as a reservoir of potential energy (Table 2) appears to be expanding [4]. The pressure is used not only statically, e.g., in isostatic molding, but also dynamically, as stream-flow kinetic energy, e.g., in water-jet cutting and homogenizing. In many cases, high pressure merely serves to overcome hydraulic resistances.

Table 2. Pressure as a form of energy

Autofrettage	< 12 000 bar
Isostatic molding	< 5000 bar
Water-jet cutting	< 4000 bar
High-pressure cleaning	< 1500 bar
Homogenizing, mixing, extrusion, cell cracking, tobacco impregnation	< 500 bar
Petroleum and natural gas recovery	< 400 bar
Spinning pumps, polymer melt pumping and filtration	< 500 bar
HPLC	< 1000 bar
Water and oil hydraulic systems	< 500 bar
Slurry conveying	< 200 bar

2. Pressure Vessels

Pressure vessels are chiefly used as reactors, scrubbers, heat exchangers, buffer vessels, or pulsation dampers. The cylindrical wall of thick-walled pressures vessels (i.e., those with an outer to inner diameter ratio $d_o/d_i > 1.2$) can either be produced as one solid piece (solid-wall vessels) or be constructed from several layers (multiwall vessels).

2.1. Solid-Wall Vessels

The term solid-wall vessel is applied to all designs in which the cylindrical wall consists of a single layer. Solid-wall vessels are suitable for all types of pressure vessels, in particular for high temperatures.

Thermal stresses arising during heating or cooling are smaller than with multilayer vessels because of the good thermal conduction through the wall. Solid-wall vessels are therefore good for batch processing.

Production. Solid-wall vessels can be produced by the following methods:

1) A one-piece hollow body is forged without welding seams.
2) Forged rings are joined to form the vessel by circular seams.
3) Thick sheets are bent around and welded into sections by one or more longitudinal seams. The sections are then joined by circular seams to form the vessel.
4) Pressure vessels are also produced by submerged-arc welding. The vessel wall is constructed together with flange and bottom by deposit welding onto a thinner tube. Vessels of any shape can be produced in this way.

2.2. Multiwall Vessels

Multiwall vessels are suitable for all types of pressure vessels, in particular those with large diameters, lengths, and wall thicknesses. Thermal conduction through the wall is lower than that for solid-wall vessels, so that heating and cooling rates must be taken into account. The type of joint between the wall elements is used as the distinguishing feature.

2.2.1. Designs with a Purely Mechanical Joint

Multiwall vessels can be produced by shrinking one or more seamlessly forged cylinders onto a core shell. The difference required between the outer diameter of the smaller cylinder and the inner diameter of the larger cylinder

must be precalculated. Before being shrunk, the larger cylinder is heated and the smaller may be cooled. During the shrinking-on process, initial compressive stresses are produced in the smaller cylinder, and initial tensile stresses in the larger cylinder, both in the circumferential and longitudinal directions.

Nowadays, this process is applied only in special cases, for example, for compressor cylinders or for vessels used in batch processing at pressures exceeding 2000 bar. Figure 1 shows an example.

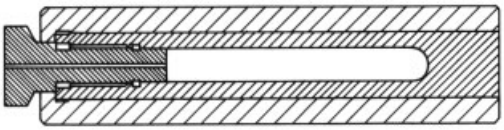

Figure 1. Two-layer shrunk autoclave for 10 000 bar

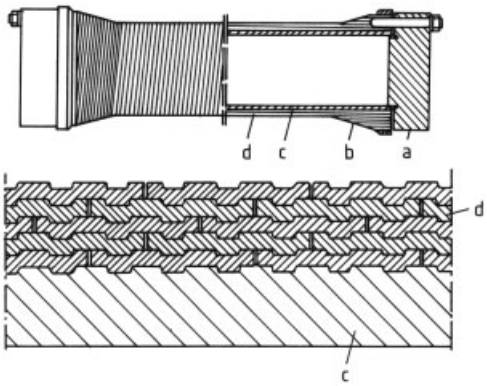

Figure 2. Pressure vessel produced by the BASF strip-winding process showing a sectional view of the wound wall a) Cover; b) Strip-wound flange; c) Core tube; d) Strip-wound vessel wall

Figure 2 shows a pressure vessel produced by the BASF Schierenbeck strip-winding process. In this process, a profiled steel strip is heated to 650 – 950 °C and helically wound onto a core tube. The strip is cooled and annealed by blowing on air and shrinks at high tension onto the underlying strip layers. The strip layers are joined to one another and to the core tube by profiled grooves, so that apart from the circumferential stresses they are also able to absorb longitudinal stresses produced by the internal pressure. The end pieces are produced from forging ingots and are joined to the cylindrical shell by overwinding. The flanges are wound from the same profiled strip, and the threaded bores for the cover bolts can then be drilled directly into the wound flange.

2.2.2. Welded Designs (Layered-Wall Vessels)

Welded designs consist of sheets which are welded together to form a shell. Figure 3 shows widely used manufacturing processes for layered-wall vessels.

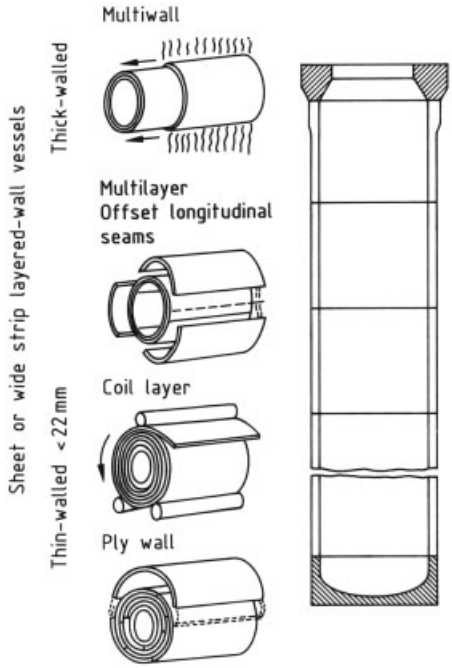

Figure 3. Various designs of layered-wall vessels [10]

Multiwall Design. Tubes having different precalculated diameters are produced from 25 – 50-mm thick sheets welded with longitudinal seams. These tubes are shrunk on top of one another without prior machining to produce individual shell sections. These are joined to one another and to the forged end pieces and flanges by circular seams.

Multilayer Design. Half-shells, formed from 6 – 22-mm thick sheets, are applied to the core tube, which is seamless or welded by a longitudinal seam. The shells are joined to one another and to the underlying layer by two longitudinal seams (three-sheet seams). The half-shells are

tangentially prestressed by the welding contraction. Uniform distribution of the welds around the circumference is achieved by offsetting the longitudinal seams from layer to layer. When larger diameters are required, three or four sheet segments are arranged on the circumference instead of half-shells. The individual sections are joined to one another by circular seams.

Coil Layer Design. The shell is produced by spirally winding a 3 – 5-mm thick sheet strip onto the core tube. The outer shell is applied after completion of the winding. The beginning and end of the winding require wedges. Longitudinal seam welds are made at the beginning and end, or whenever a sufficiently long strip is not available.

Ply Wall Design. This design is a combination of the coil layer and multilayer designs: the spiral construction of the shell section is combined with three-sheet longitudinal weld seams.

2.3. Strength Calculations

2.3.1. Cylindrical Wall

2.3.1.1. Stresses Due to Internal Pressure

The stresses generated by internal pressure in a cylindrical body with the outer to inner diameter ratio $u = d_o/d_i$, are listed in Table 3.

Table 3. Stresses and loads due to internal pressure p (N/mm^2) in a closed hollow cylinder ∗

Property	Inside	Outside
Tangential stress	$\sigma_t = p\frac{u^2+1}{u^2-1}$	$p\frac{2}{u^2-1}$
Longitudinal stress	$\sigma_l = p\frac{1}{u^2-1}$	$p\frac{1}{u^2-1}$
Radial stress	$\sigma_r = -p$	0
Stress according to the GE hypothesis	$\sigma_v = p\frac{u^2\sqrt{3}}{u^2-1}$	$p\frac{\sqrt{3}}{u^2-1}$
Stress according to the shear stress hypothesis	$\sigma_{red} = p\frac{2u^2}{u^2-1}$	$p\frac{2}{u^2-1}$

∗ p = internal pressure in N/mm^2, u = diameter ratio d_o/d_i.

Figure 4 shows the stresses produced by internal pressure in a cylinder with $u = 2$. The stress is substantially higher on the inner than on the outer wall. The differences between the inside and outside increase with the square of the diameter ratio. This means that with thick-walled vessels the material is only partly exploited if a safety margin of, for example, 1.5 with respect to the tensile yield strength is maintained on the inside of the wall. The unfavorable stress distribution over the wall thickness can be improved by appropriate design methods (shrink joints); it can be altered by using an increased test pressure (autofrettage), or it can be ignored by selecting a different computational method (complete plastic yielding).

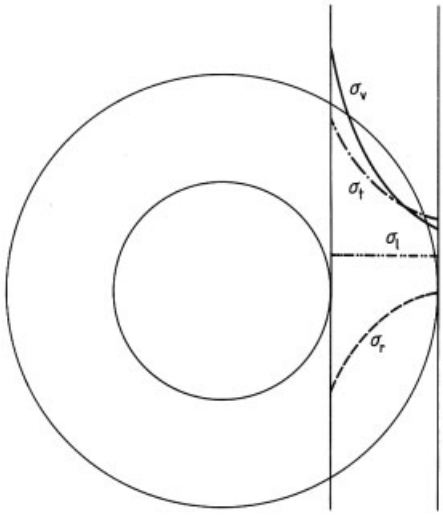

Figure 4. Stress components in a thick-walled hollow cylinder under internal pressure
σ_t = tangential stress; σ_r = radial stress; σ_l = longitudinal stress; σ_v = stress intensity

Selection of the operating pressure depends on the safety margin with respect to the pressure at which the vessel is deformed to an impermissible degree or bursts. When pressure is increased in a thick-walled vessel, the wall expands elastically in a uniform manner until the stress at the inside of the wall is equal to the tensile yield strength of the material. With further pressure increase, the outside of the wall continues to expand elastically, whereas the inside the wall is increasingly deformed plastically. Once the plastic deformation has reached the outside of the wall as well, so-called complete plastic yielding occurs.

In the computation according to the GE hypothesis (v. Mises), plastic deformation begins on the inside of the wall if:

$$p_{el} = \sigma_{0.2}\frac{u^2-1}{u^2\sqrt{3}} \quad (1)$$

and complete plastic yielding is reached when:

$$p_{\text{complete pl.}} = d_{0.2} \frac{2}{\sqrt{3}} \ln u \tag{2}$$

where $\sigma_{0.2}$ is the 0.2 % proof strength.

Figure 5 shows the ratio $p_{\text{complete pl.}}/p_{\text{el}}$ as a function of the diameter ratio u. Thus the material of thick-walled vessels is more fully exploited if complete plastic yielding is used as the design limit.

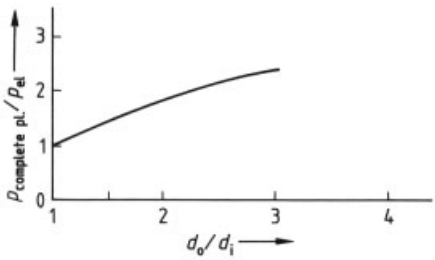

Figure 5. Ratio of complete plastic yielding to plastic deformation as a function of the diameter ratio

An ASME (American Society of Mechanical Engineers) committee is preparing a new code for calculating high-pressure vessels (Sect. VIII, Div. 3). It uses the type of calculation presented above, and selects a safety factor of $S = 2$ relative to complete plastic yielding according to the GE hypothesis, or $S = 1.732$ with respect to complete plastic yielding according to the shear stress hypothesis. The permissible internal pressure is thus:

$$p = \sigma_{0.2} \frac{1}{1.732} \ln u \tag{3}$$

The following points must also be considered:

1) The stress must not exceed the tensile yield strength at any point
2) The fatigue strength must be taken into account
3) Brittle fracture must not occur

Initial stresses, for example caused by shrinking or autofrettage, may also be considered in these calculations.

2.3.1.2. Initial Stresses in Shrink Joints

The unfavorably high tangential stresses on the inside of the wall can be substantially reduced if two or more cylinders are shrunk onto one another. Shrink joints are generally produced by heating the outer cylinder prior to joining. However, it is also possible to press conical cylinders onto one another. In order to calculate the initial stresses generated by shrinking, it is important to make an exact measurement of the diameter at the joint surface before shrinking takes place.

The shrinkage should be selected such that at operating pressure the inner and outer cylinders have the same safety margin with respect to the tensile yield strength or the fatigue strength.

Figure 6 shows the tangential stresses after shrinking and due to the internal pressure.

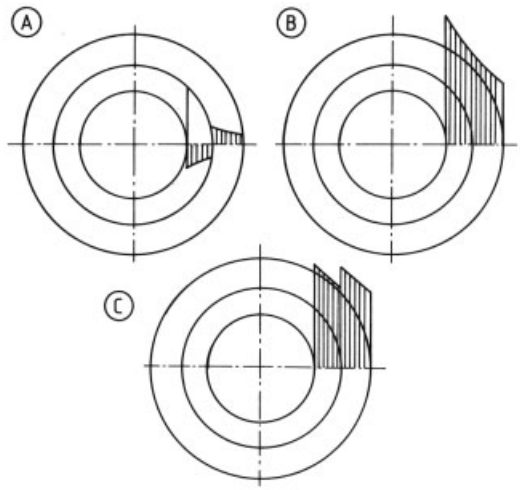

Figure 6. Tangential stress in two-layer shrink vessels
A) After shrinking; B) Due to internal pressure alone; C) Tangential stress resulting from shrinkage and internal pressure

2.3.1.3. Residual Stresses Due to Autofrettage

Autofrettage can be used to generate a favorable stress distribution in the vessel wall. A higher internal pressure causes the material on the inside of a thick-walled vessel to flow, while the outside is only elastically expanded. The pressure is released after part of the wall, or even the entire wall, has been plastically deformed. The material in the vicinity of the inside of the wall is strongly plastically deformed in the autofrettage process and tends to remain deformed. The material near the outside of the wall is least

deformed, or may not be deformed at all, and therefore tends to return to its original shape. As a result, tangential residual stresses are produced as compressive stress on the inside of the wall and as tensile stress on the outside; they vary nonuniformly over the wall thickness. The residual compressive stresses on the inside of the wall later counteract the tangential stresses produced by the internal pressure. The result is that with pulsating pressure there is a reduction in the mean stress, which increases service life.

The risk of brittle fracture is also reduced by autofrettage since the stress intensity factor at a crack tip or a cracklike inclusion is lowered. This slows down crack growth and increases the critical crack depth.

The residual stresses obtained by autofrettage can be computed according to ASME Section VIII/3; three methods are used:

1) Measurement of the mean tangential expansion at the outside of the wall at the maximum autofrettage pressure
2) Measurement of the mean residual expansion at the inside of the wall

When calculating the load, the initial stresses can be superimposed on the other stresses. The first step is thus to add together the individual stresses in each of the three main axes, from which the load can then be calculated.

Autofrettage pressure is often selected so that ca. 30 % of the wall thickness is plastically deformed. Figure 7 shows the tangential stress occurring in a pipe after autofrettage, and the stress resulting from autofrettage and internal pressure.

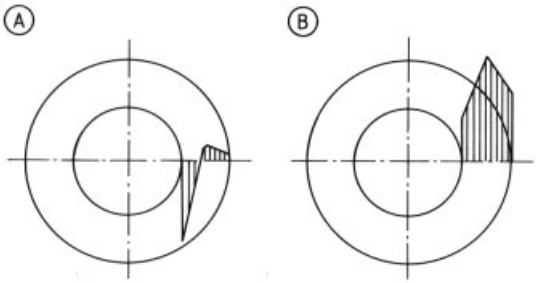

Figure 7. Tangential stress in vessels treated by autofrettage A) After autofrettage; B) Tangential stress resulting from autofrettage and internal pressure

2.3.1.4. Thermal Stresses

When pressure vessels are operated at constant temperature, thermal stresses do not generally need to be taken into account. If, however, the vessel is heated or cooled under internal pressure, the stress on the vessel wall can increase considerably. Thermal stress must be taken into account when heating occurs from outside through the wall or when the vessel wall is cooled by the contents.

2.3.1.5. Avoiding Brittle Fracture

Brittle fracture is a type of failure in which a crack present in a part subjected to tensile stresses spreads suddenly and without plastic deformation and thus destroys the component.

The tendency to brittle fracture is increased by:

1) High tensile strength of the material and thus low toughness
2) Low temperature and thus low toughness
3) Large tensile stresses
4) Large wall thickness
5) Unfavorable design (peak stresses)
6) Corrosion

The calculation of the critical crack length in vessels under internal pressure is difficult, since steep stress gradients are present at the risk sites, and high internal pressure at the crack tip has a substantial influence on brittle fracture behavior. In the case of the steels that are usually used, this calculation is generally only necessary for large vessels in which the operating pressure exceeds 1000 bar.

2.3.2. Cylindrical Wall with Radial Bore

Radial bores are used for thermal measurements and pipe connections.

Stresses Due to Internal Pressure. In the region where the bore penetrates the inside of the cylinder, the internal pressure produces a stress which is approximately three times that on the intact cylinder wall. This site is therefore often the starting point for fatigue cracks. The risk of cracks can be reduced by rounding off and polishing the edges.

Residual Stresses as a Result of Test Pressure or Autofrettage. Even at relatively low pressure, the tensile yield strength is exceeded at the point of penetration. As a result, an autofrettage effect can often be achieved in this region during pressure testing. As an example, a residual compressive stress of 0.5 times the tensile yield strength is generated for a diameter ratio u of 2.25 at the point of penetration if the inside of the wall of the intact cylinder has been loaded up to the tensile yield strength during pressure testing.

2.3.3. End Pieces

As a rule, end pieces that are forged or welded on are constructed either as hemispherical or flat plates. *Hemispherical end plates* require a wall thickness of approximately 60 % of the cylinder wall. They ensure good distribution of stress at the transition from the hemisphere to the cylinder. A *flat end plate* is frequently chosen for production engineering reasons. High stresses, which are strongly influenced by the form of the transition from cylinder to end plate, are produced at the transition from the flat plate to the cylinder wall. In the case of a favorably shaped conical transition (Fig. 8) the load is only slightly higher than in the cylindrical wall.

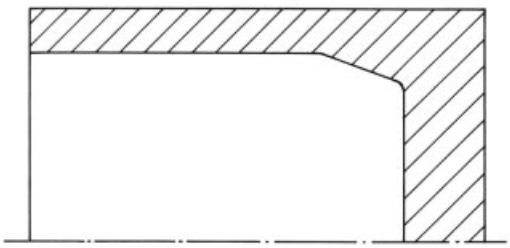

Figure 8. Favorably shaped conical transition from cylinder to end plate

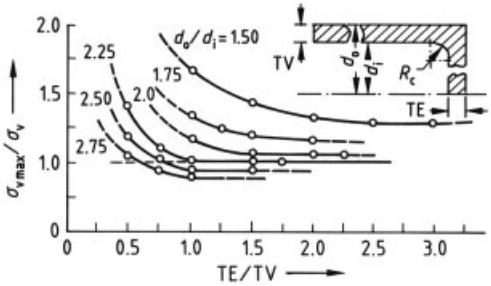

Figure 9. Ratio $\sigma_{v\,max}/\sigma_v$ at the radius of transition from cylinder to end plate according to [6] (only if R_c/TV = 0.25)

The shape factor $\sigma_{v\,max}/\sigma_v$ is shown in Figure 9 for a wall thickness ratio $d_o/d_i > 1.5$ and applies only if the transition radius is one quarter of the wall thickness (R_c/TV = 0.25). The stress can be reduced only slightly by using a larger radius.

2.4. Corrosion in High-Pressure Plant

Only those types of corrosion which are influenced by high pressure are dealt with in this chapter.

Hydrogen Attack at Elevated Temperatures (→ Corrosion, Chap. 3.3.1.6.). The action of molecular hydrogen at high pressure on steels must be considered in a number of high-pressure processes, e.g., hydrogenation and ammonia production. At elevated temperatures (> 200 °C) and under pressure, hydrogen tends to decarburize carbon steels, which causes embrittlement and deterioration of the mechanical properties. Hydrogen attack can be prevented by using carbon-binding alloying elements such as chromium, molybdenum, vanadium, and tungsten. The hydrogen-resistant steels used nowadays contain 2 – 6 % Cr, 0.2 – 0.6 % Mo, and a few types are additionally alloyed with up to 0.86 % V and 0.45 % W. Figure 10 shows the resistance limits for different steels.

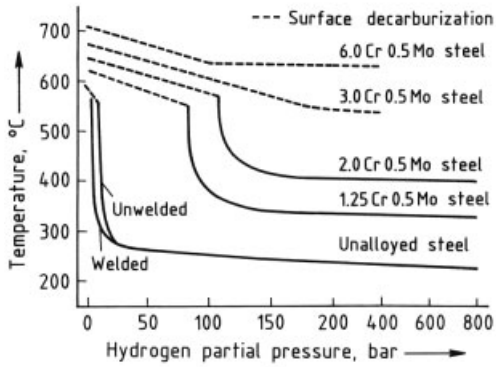

Figure 10. Resistance limits for the attack of pressurized hydrogen on steels (Nelson diagram)

Corrosion by Carbon Monoxide (→ Corrosion, Chap. 4.6.). Carbon monoxide under pressure attacks unalloyed and low-alloy steels at 130 – 140 °C with formation of iron pentacarbonyl. Attack virtually ceases above 350 °C. With high-alloy chromium and chromium–nickel steels the damage is significantly less. Chromium steels with 30 % Cr and austenitic steels with 25 % Cr and 20 % Ni are completely resistant. If the iron pentacarbonyl is entrained in a synthesis cycle, it may decompose exothermically at a higher temperature with the formation of active, pyrophoric iron.

Nitriding by Ammonia (→ Corrosion, Chap. 4.6.). At temperatures above 250 °C, ammonia forms a nitride layer on steels and embrittles the underlying material. Welding is then difficult owing to the nitrogen enrichment, and these parts are often irreparable after lengthy operating times. Apart from Armco iron, austenitic steels are the least subject to attack. Although a nitride layer does form with austenitic steels above 350 °C, no loss of material occurs because this layer does not flake off. This layer must be ground off before repair welding, but the welding properties of the base material are not affected.

2.5. Material Selection

Low-alloy steels (as specified in, e.g., DIN 17 200) are generally used for the load-bearing wall of pressure vessels. In the case of relatively thick walls, good annealing characteristics and good toughness are important. Wall temperatures exceeding 350 °C may produce temperature embrittlement, which can be reduced by limiting the content of impurities (S, P, Zn, As). Hydrogen-resistant materials are necessary at temperatures above 200 °C with hydrogen-containing media. It is generally necessary to use high-strength steels at operating pressures above 2000 bar. Since elongation at rupture and toughness decrease with increasing strength, good toughness is particularly important with these steels. Fracture-mechanical investigations are advisable if these steels are used for wall thicknesses exceeding 150 mm.

Some typical steels for high-pressure vessels and piping are listed in Table 4.

Table 4. Steels used for high-pressure applications

Type of steel	Material no.	Uses
10 CrMo 910	1.7380	for forgings and pipes to prevent hydrogen attack above 200 °C
12 CrMo 910		for forgings and pipes to prevent hydrogen attack above 200 °C
24 CrMo 5	1.7258	for covers, flanges, bolts
21 CrMoV 5 7	1.7709	for forgings at high temperatures, not hydrogen-resistant
30 CrNiMo 8	1.6580	pipes, forgings at pressures above 1000 bar
24 CrMo 10	1.7273	forgings to prevent hydrogen attack above 200 °C
20 CrMoV 13 5	1.7779	pipes, forgings to prevent hydrogen attack above 200 °C
X 20 CrMoV 12 1	1.4922	pipes and forgings to prevent hydrogen attack above 450 °C

2.6. Design Details

2.6.1. Corrosion Protection

The types of corrosion that are specifically intensified by high pressure are described in Section 2.4. Apart from these, there are other types of corrosive attack that are independent of pressure.

Vessel Walls Made from Corrosion-Resistant Material. This is the simplest corrosion protection method, but itis only possible in a few cases because of cost reasons. In the case of hydrogen attack, the wall can be protected by using the alloying elements Mo, Cr, Ni, and V. Multilayer vessels are advantageous since they require hydrogen-resistant steel only for the inner layer. The other layers need not be hydrogen-resistant if design (e.g., venting channels) ensures that hydrogen diffusing through the inner layer is drawn off.

Cladding. Abrasion can be prevented by roll-bonding cladding or weld overlay cladding. However, this type of protection is often not possible with the steels and wall thicknesses used for high-pressure vessels. An advantage of this method is that the cladding materials (e.g., austenitic steel) often have a different thermal expansion than the vessel wall. This leads to thermal stresses in the cladding during heating and cooling. If the cladding is joined solidly to the vessel wall, the expansion of the base material is imposed on every part of the cladding.

The thermal stresses are therefore equal in all parts of the cladding.

Disadvantages of this method include:

1) The base material is attacked if the cladding is damaged at any point and the damage often cannot be detected in time.
2) In weld overlay cladding, the solidifying melt forms columnar crystals perpendicular to the wall surface. Along the grain boundary of these crystals, the corrosion resistance is lower than in the rolled material. In the case of heavy corrosion, at least a two-layer weld overlay cladding should therefore be selected.
3) Cladding does not protect against hydrogen attack, since hydrogen diffuses through the cladding to the base material.

Lining. With loose linings, the tightness can be monitored during operation. For this purpose, the space between the lining and the pressure-bearing wall is monitored by means of venting holes. The lining must be attached to the pressure-bearing wall in such a way that different thermal expansion cannot produce a displacement between the lining and the vessel wall.

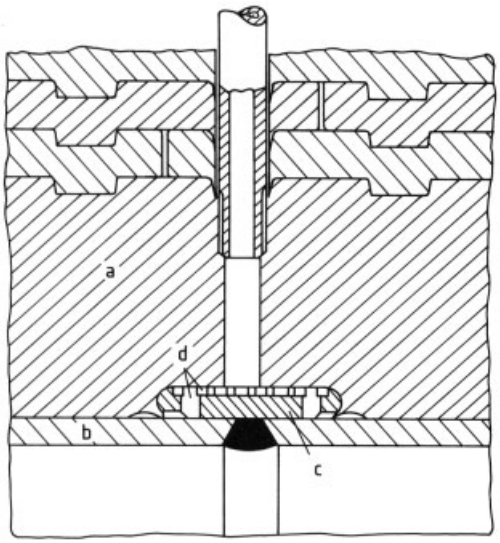

Figure 11. Titanium lining in a pressure vessel attached in the axial direction by mortised titanium strips
a) Core tube of a multilayer vessel or vessel wall; b) Titanium sheet; c) Titanium strip with grooves; d) Bores for blanketing gas and leakage monitoring

An advantage over cladding is that the lining can be checked for leaks at any time. Disadvantages of this method compared to cladding include:

1) Given large differences in thermal expansion or an unsuitable attachment to the pressure-bearing wall, the expansions can be concentrated on the weakest point and cause damage there
2) The joining of a nozzle lining to the lining of the vessel can be a weak spot

Figure 11 shows the attachment of a loose titanium lining by means of mortised titanium strips. The venting hole through the vessel wall is used to introduce the protective gas.

Floating Bladders. The "floating bladder" principle is used under extreme corrosion conditions. The balloon is made from an alloy that is resistant to the extremely aggressive medium and is surrounded by pressurized water so that the internal pressure of the balloon and the pressure of the water surrounding it are virtually identical (see Fig. 12, next page).

An advantage of this method is that the high-pressure shell and the high-pressure seals are not directly exposed to the aggressive medium. Leaks in the bladder can be detected by chemical monitoring of the pressurized water. A disadvantage of this method is that the required pressure equalization between the balloon and the pressurized water space necessitates additional outlay on piping, measuring instruments, and balancing tanks.

2.6.2. Covers and Their Attachment

The most commonly used types of closure are shown in Figure 13 (see next page).

The *cover seals* (Fig. 13 A–E) are attached by means of threaded bolts and are used for all diameters and pressure ranges. The bolts are generally tightened hydraulically or pneumatically. It is, however, also possible to prestress the bolts longitudinally and then to tighten the nuts.

In the *Bredtschneider seal* (Fig. 13 F), the force acting on the cover due to the internal pressure is led via the sealing ring (a) onto the split ring (b), which is supported in the groove of the shell. This design requires only small bolts

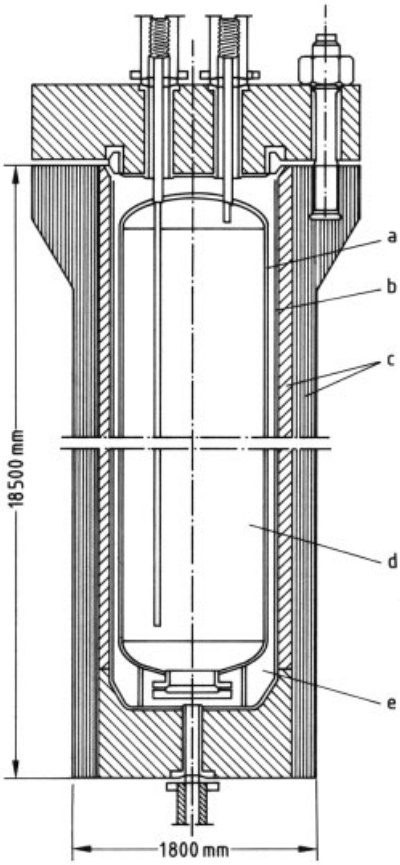

Figure 12. Pressure vessel with internal bladder of highly corrosion-resistant material
a) Internal bladder of highly corrosion-resistant material; b) Austenitic steel lining; c) BASF-Schierenbeck multilayer vessel (low-alloy steel); d) Highly corrosive medium (pressure p_a); e) Water (pressure, p_w; monitored by chemical analysis); $p_w = p_a + 2$ bar

which bring the cover and sealing ring together in the pressureless state.

The *clamp seal* (Fig. 13 G) permits rapid opening and closing of the pressure vessel. The two- or three-part clamp is pushed over the collar on the cover and vessel, and fixed. The elastic seal is pressed onto the sealing surface by the internal pressure. It is easy to automate the closing process, which is why this joint is frequently used for batch processing, e.g., in extraction with supercritical gases.

Screwed-in covers (Fig. 13 H) are used for smaller inner diameters. The Hahn and Clay *finger-pin seal* (Fig. 13 I) is suitable for large diameters. An elastomeric seal is necessary. Opening and closing can easily be automated.

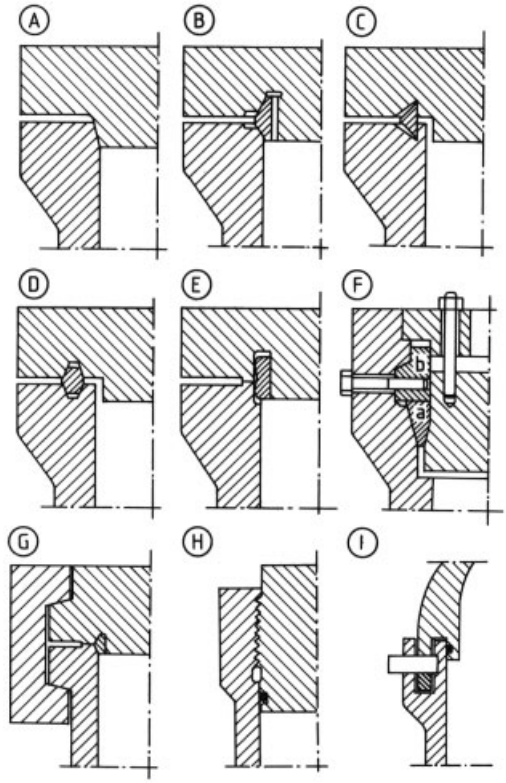

Figure 13. Forms of cover and forms of seals (see text for explanation)

2.6.3. Cover Seals

Metal seals are usually necessary as cover seals in chemical high-pressure plants. The single cone (Fig. 13 A) with a conical slope of 10° is chiefly used for cover openings of up to 800 mm diameter and pressures up to 400 bar. The double cone ring (Fig. 13 B) with a conical slope of 30° is more complicated than the single cone, and is used for cover openings larger than 500 mm and pressures up to 700 bar. The delta ring (Fig. 13 C) and the RTJ ring (Fig. 13 D) can be regarded as smaller versions of the double cone ring. The wave-ring seal (Fig. 13 E) is pressed by the internal pressure against the cylinder sealing surfaces and is plastically deformed in the process. This is a special seal for stirred-tank polyethylene reactors where they are used up to 300 °C and 2500 bar. Flat sealing rings and metallic O-rings are seldom used in chemical plants.

2.6.4. Seals for Piping

These seals are used for substantially smaller diameters than the cover seals. For cost reasons, the aim is to have simple shapes which can also be replaced easily. Larger tightening forces during assembly are then acceptable.

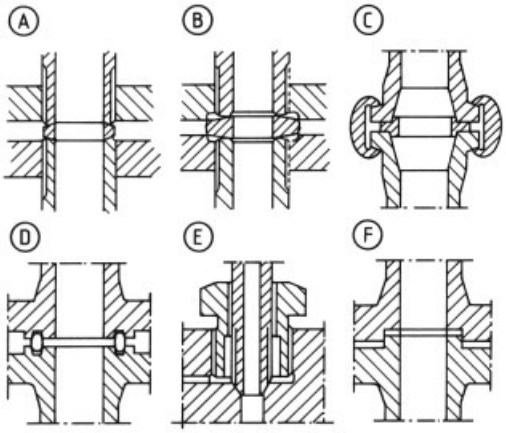

Figure 14. Seals for high-pressure piping (see text for explanation)

Figure 14 shows the most common designs for piping seals. The lens ring seal (Fig. 14 A) is used in Europe for pressures up to 4000 bar. The conical ring seal (Fig. 14 B) is the most widely used in the United States for pressures from 1000 to 2000 bar. The Grey lock seal (Fig. 14 C) is used in the United States predominantly for pressures up to 2000 bar. The ring joint seal (Fig. 14 D) can be used for pressures up to 200 bar and temperatures up to 700 °C. The conical seal with cap nut (Fig. 14 E) can be used for small nominal pipe sizes and pressures of up to 6000 bar. Flat seals (Fig. 14 F) are rarely used in high-pressure plants.

2.7. Example of the Design of Pressure Vessels

Figure 15 shows the upper part of a reactor having internal cooling tubes, a central high-pressure connection, and an inserted temperature-measuring tube. The reactor is lined inside with stainless steel, and the cover is sealed to the vessel wall by means of a double cone seal. The cooling liquid is led through the lateral branches (nominal diameter 125 mm) into the distribution chamber above the cover. The liquid flows through the dip pipes to the lower end of the cooling tubes, which are closed at the bottom, and returns to the top around the dip pipes.

Figure 16 (see next page) shows the dimensions of the reactors used in ammonia plants since 1910. Even larger pressure vessels are built as hydrocrackers and desulfurization reactors which have inner diameters of 4.1 m, a

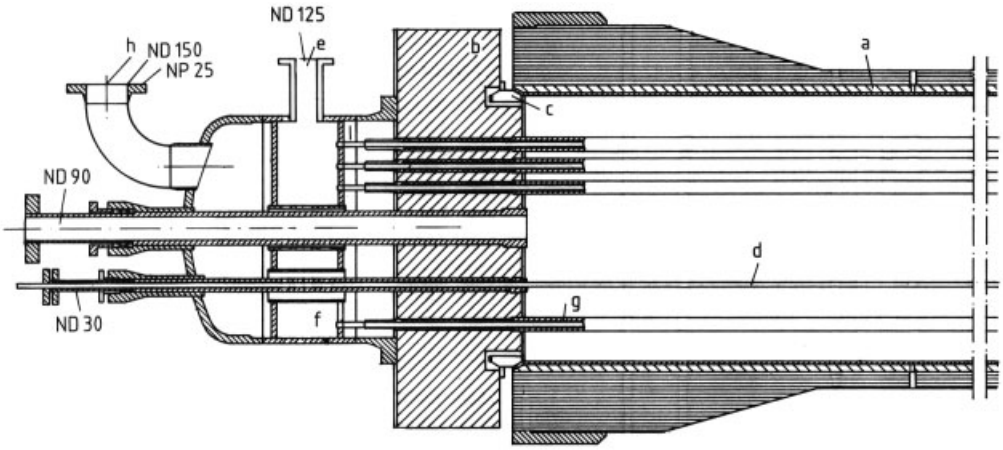

Figure 15. Reactor with internal cooling tubes
a) High-pressure vessel; b) Flat cover; c) Double cone sealing ring; d) Thermowell; e) Inlet nozzle for cooling fluid; f) Distribution chamber g) Field tubing for cooling; h) Outlet nozzle for cooling fluidND = nominal diameter in millimeters; NP = nominal pressure in bar

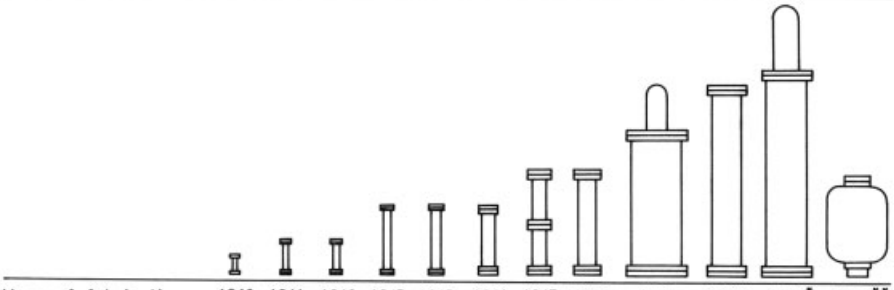

Year of fabrication	1910	1911	1912	1913	1913	1914	1915	1956	1963	1969	1972*	1982**
Inner diameter, mm	90	90	160	200	300	500	800	1200	2900	2000	2400	3350
Height, m	1.8	4	4	8	8	8	12	12	22	22	34	9
Weight, t					3.5		65	105	270	386		190
Volume of catalyst, m^3					0.09		1.1	4.85	64	36	56	53
Operating pressure, bar					200		300	300	150	350	210	210
Capacity, t/d					3.6–4.8		85	195	910	1200	1580	1800

*Horizontal **2 Reactors

Figure 16. Development of the dimensions of ammonia reactors since 1910 [20]

wall thickness of 280 mm, and a total weight of 1100 t. The material employed has largely been ASTM A 336-F 22 with special purity requirements. These reactors generally have a stainless steel weld cladding.

3. High-Pressure Machinery for Chemical Plants

Because of severe stresses, it is necessary to design and fabricate high-pressure machines very carefully. Experience has shown that, in high-pressure processes, the high-pressure pumps and compressors present the greatest problems as regards lifetime and maintenance. In process and production engineering, high pressure is generally understood as 100–4000 bar. Most important applications also fall within this range [25] (see Tables 1 and 2).

3.1. Special Features of High-Pressure Machines

High-pressure machines are expensive to develop, produce, supply with energy, and maintain. Since the components are subject to high stress, careful stress analysis is necessary. For dynamic stresses, a design with few notches and with excellent surface quality is required. At points that are especially prone to risk, local material properties must be improved, and stress peaks must be reduced.

The solution of sealing problems, especially under dynamic stress, requires a great deal of experience. The piston seals of reciprocating displacement pumps present special problems since these pumps overcome very high pressure differences (up to several thousand bar) in a single stage. The sealing of problem fluids, such as suspensions and materials with low viscosity and poor lubricating action, requires special measures.

In selecting materials for high-pressure components, special attention should be paid to high strength, good isotropy, homgeneity, toughness, and quality. Apart from high-strength ferritic tempering steels fabricated with special purity, soft martensitic, semiaustenitic, and precipitation-hardened chromium – nickel steels are increasingly being used to meet corrosion resistance requirements. Homogeneity is achieved by the smelting method and isotropy is achieved by forging the slugs as much as possible from all sides.

For dynamic seals, all capabilities of modern materials development for plastics, bronzes, sintered metal carbides, and layers of hard material are utilized. The special material properties of fluids at high pressures, e.g., compressibility, viscosity changes, density, and solidification temperature, must be taken into account.

3.2. Creation of Pressure by Pumps and Compressors

The approximately isentropic change of state during a pressure increase (Fig. 17) causes a temperature rise which is small in liquids but large in gases. This temperature rise limits the pressure ratio which can be produced by one stage in compressors. The required work is generally determined by the difference of the total enthalpies. For gas compressors, efficiency data refer either to isothermal or isentropic changes of state, which leads to considerable differences (17).

For *reciprocating piston machines*, the clearance volume, fluid compressibility, and working-space elasticity determine the volumetric efficiency (also known as the transfer or discharge level), which is a measure for the utilization of the structurally determined swept volume. High-pressure piston compressors require a multistage arrangement with intermediate cooling to approximate the energetically favorable isothermal change of state for economical operation. The pressure ratios produced by one stage amount to 2 – 5 in the case of process gas compressors (Fig. 18, left). By contrast, piston pumps for liquids overcome very high pressure differences in a single stage because liquids are much less compressible (Fig. 18, right). The fatigue of the pump component with a cyclical interior pressure [26, 27] and the lifetime of the piston seals are limiting factors.

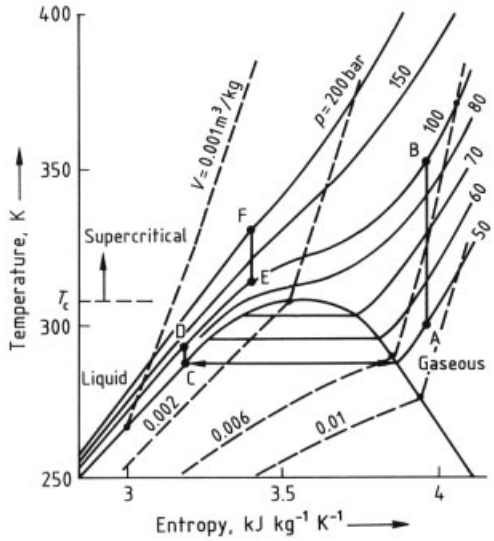

Figure 17. Isentropic change of state of carbon dioxide A – B gases; C – D liquids; E – F supercritical fluids

With *rotary displacement machines*, e.g., gear pumps, interior leakage losses determine the volumetric efficiency. For high-pressure applications, such pumps are suitable only for highly viscous fluids.

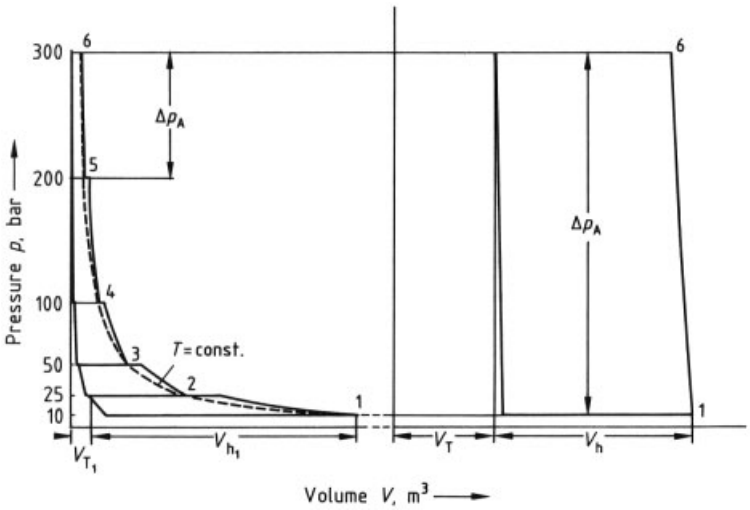

Figure 18. Gas and liquid compressor (pump). The $p - V$ diagram for a
Left: five-stage gas compression; Right: single-stage liquid compression
V_h = stroke volume; V_T = clearance volume; Δp_A = pressure increase per stage

When pressure is generated hydrodynamically, as in *turbomachines* (Fig. 19), the circumferential velocity u of the commonly used radial impellers is very important for the pressure increase per stage $\Delta p_i = \psi \varrho u^2/2$ (where ψ is the coefficient of pressure, ϱ is the fluid density). The rotor stress limits the circumferential velocity for pumps to about 100 – 200 m/s. For turbo compressors, the speed of sound (ca. 300 – 400 m/s) is the limit; generally Mach numbers < 0.7 are used.

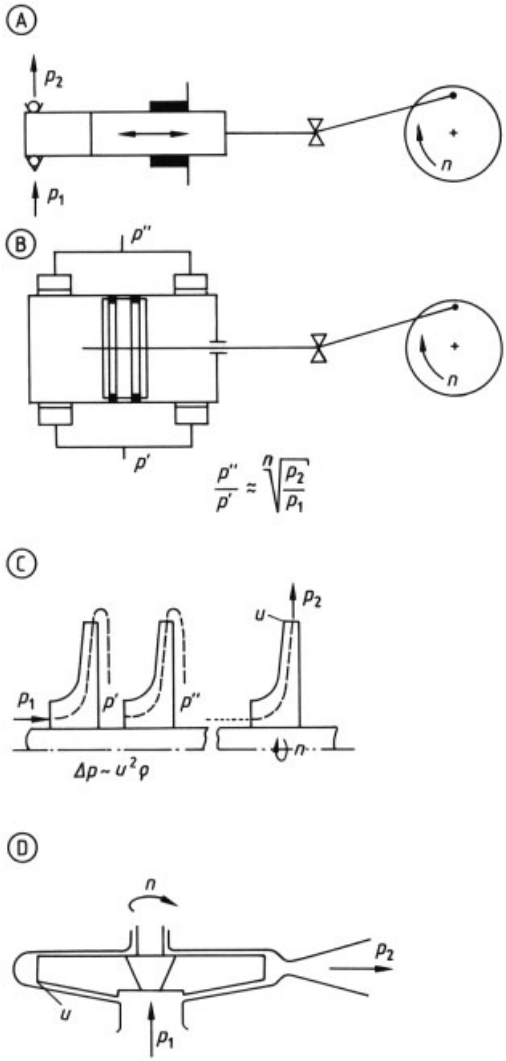

Figure 19. Principles of pressure generation
A) Single-stage piston pump; B) Piston compressor stage (double-action); C) Multistage centrifugal pump (or turbo compressor); D) High-speed (partial emission) single-stage centrifugal pump

Nevertheless, pressure increases of 50 – 250 bar per stage can be achieved for liquids, depending on the power. For gas compression, more stages are needed than for liquid pumps because of the lower fluid density, and intermediate cooling is required.

Compared to hydrodynamic turbomachines, hydrostatic displacement machines exhibit pulsation of the flow rate (Fig. 20). Generally displacement machines are more efficient for comparable power, but turbomachines have the advantage of being smaller in size, thus requiring less space and running quieter.

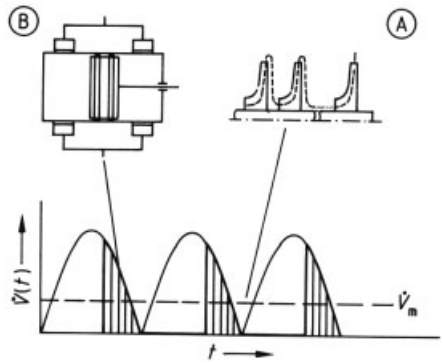

Figure 20. Time behavior of the fluid flow
a) Hydrodynamic pump or compressor; b) Hydrostatic (double-action) piston machine
$\dot{V}_m$ = average-volume flow; $\dot{V}$ (t) = time volume flow

3.3. Pumps

In high-pressure systems, pumps are used for volumetric metering, transfer, circulation, and pressure maintenance. Volume flow rates range from ca. 1 cm^3/h to $> 10^3$ m^3/h. The pressure range generally extends up to 300 – 1000 bar, more rarely into the range of 1000 – 4000 (10 000) bar (Fig. 21).

The characteristic curves of pump construction types differ mainly in their pressure stiffness (the influence of pressure on flow rate) (Fig. 22 A). Because of their pressure-elastic curve, centrifugal pumps can also be regulated by throttling (Fig. 22 A: the operating point shifts from A to B). Regulation of the rotational speed is more favorable energetically (Fig. 22 A: the operating point shifts from A to B′). Bypass regulation, in which part of the pump's flow

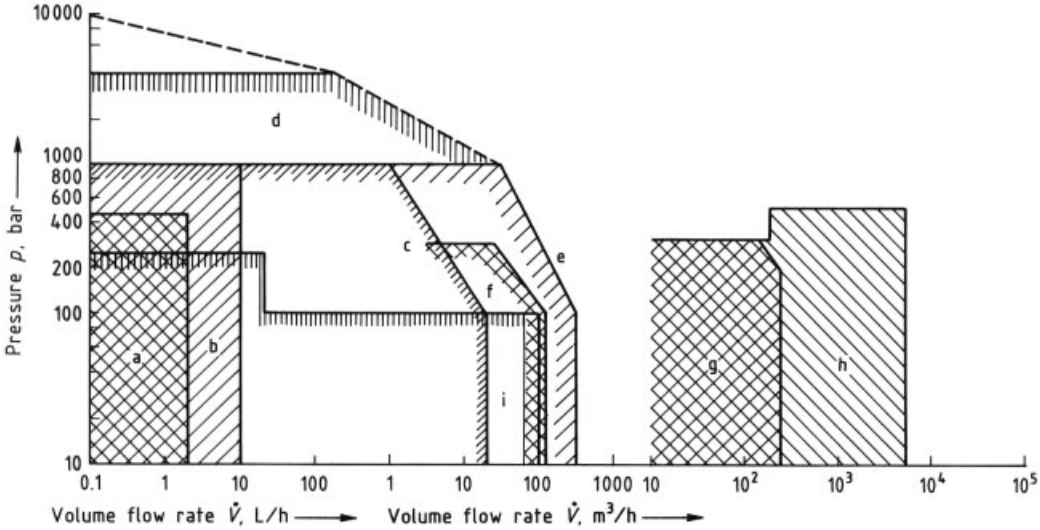

Figure 21. Power range of various types of high-pressure pumps
a) Micrometering pumps (piston, diaphragm, gear, spinning pumps); b) Metering pumps for laboratory and pilot systems (piston, diaphragm/hydraulic); c) Metering pumps (piston, diaphragm/hydraulic); d) Ultrahigh-pressure piston pump; e) Multicylinder piston pump; f) Multicylinder diaphragm pump; g) Centrifugal pumps (high-speed, single-/two-stage); h) Centrifugal pumps (multistage); i) Rotating displacement pumps (gear, spindle)

rate is recycled to suction through a regulation valve, is also customary for large machines and low energy costs. Displacement machines have pressure-stiff characteristic curves; they thus impress their flow rate on the pumping system (Fig. 22 A: lines d and e). As a rule, the flow rate depends approximately linearly on the regulated variables (rotational speed n, stroke length h) (Fig. 22 B). With reciprocating displacement pumps, the high differential pressure shifts the characteristic curve from the origin due to elasticity effects of the working space and the fluid [28]. The same applies to rotary displacement pumps at higher pressure differences due to interior clearance leakages. With rotational speed n_0 (e.g., gear pump) and stroke length h_0 (e.g.,

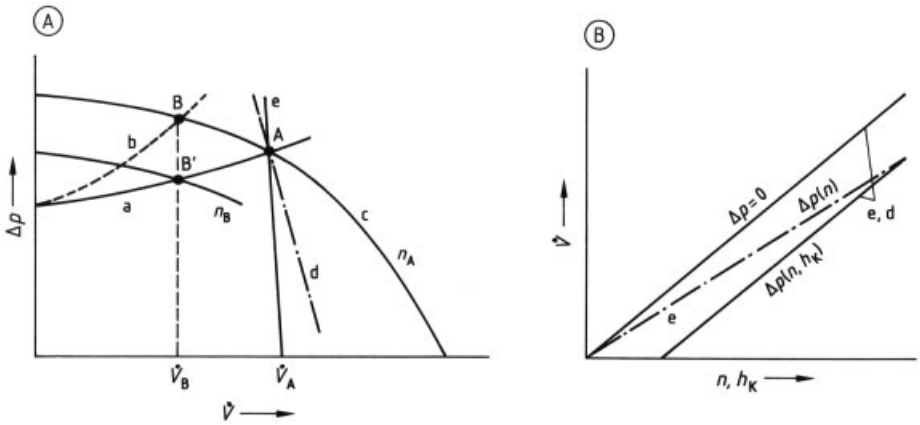

Figure 22. Characteristic curves for various types of pumps
A) Pump in the system, throttle/rpm regulation of centrifugal pumps; B) Regulation characteristics of displacement pumps
a), b) Characteristic curve for the system; c) centrifugal pumps; d) rotating displacement pumps; e) reciprocating displacement pumps
$\dot{V}$ = fluid flow; Δp = pressure difference; n = rpm; h_K = stroke length; A, B, B′ = operating points

reciprocating metering pump), the flow rate is 0. Only in the case of reciprocating displacement pumps is there proportionality between the flow rate $\dot{V}$ and the regulated variable n, because the interior leakage currents tend to vanish (Fig. 26 B).

3.3.1. Reciprocating Displacement Pumps

In high-pressure engineering, reciprocating piston and diaphragm pumps are used in many variants up to pressures of 10 000 bar. They may have electrical, hydraulic or pneumatic, linear or mechanical crank, or eccentric drives, and their stroke or speed may be adjustable. The following sections refer to metering and transfer pumps, and are oriented in terms of important applications.

3.3.1.1. Metering Pumps

For very small, pulsation-free metering flows (e.g., $10-10^3$ cm^3/h), hydraulic and mechanical (spindle) linear drives are used for pressures up to several thousand bar. By operating two systems in parallel, these can also be operated without interruption (Fig. 23 a and b) [28]. By superposing two displacement pistons or by controlling the angular velocity ω with a special cam drive, small metering flows with low pulsation are achieved up to a pressure range of about 700 bar. This can even be done with leak-free hydraulic diaphragm pump heads (Fig. 23 c and d). Such pumps are also used in HPLC analysis [29, 30]. If high-frequency metering flow pulsations do not interfere with the process, as is the case in most applications, reciprocating piston or diaphragm metering pumps with adjustable stroke and speed are used.

Magnetic linear drives (Fig. 23 e) are distinguished by their ease of operation and their small construction size. They are suitable for flow rates <1 L/h and pressures up to 500 bar with diaphragm pump heads. For larger metering flows, reciprocating metering pumps with an adjustable stroke and with mechanical transmissions have proved suitable. These pumps can be designed with several cylinders for metering several components or to smooth out pulsations (Figs. 23 f and g) [31]. The automated systems are activated by stroke actuators or by drives with variable speed. Diaphragm metering pumps with a hydraulic diaphragm drive are suitable for very high pressures. Up to about 500 bar, their diaphragm consists of polytetrafluoroethylene (PTFE); beyond this, up to about 3000 bar, it consists of metal (e.g., chromium – nickel steel) [31, 32]. The leak-free design operates reliably, is easy to maintain, presents no problems with hazardous fluids, and circumvents piston sealing problems.

With smaller metering flows, spring-loaded cam drive systems predominate because they make it possible to adjust the stroke without backlash (Fig. 23 h). For larger metering flows (hydraulic power up to 50 kW per metering element), all types of drive designs with adjustable stroke are in use. All of these are based on the familiar conventional crank mechanism (Fig. 24).

The lifetime of PTFE diaphragms, which are primarily used up to 500 bar, exceeds 10 000 h by far. Diaphragm rupture can be signalled by sandwich diaphragms [33, 34], which consist of two diaphragms, the space in between being connected to a pressure sensor via a check valve. If one of the diaphragms ruptures, the discharge pressure operates a pressure switch. In normal operation, the two diaphragms are connected by an adhesive intermediate fluid during suction and by direct mechanical support on one another during discharge.

For very high pressures (>1500 bar), diaphragm pumps with metal diaphragms are used only for laboratory and pilot systems because they are quite sensitive to surface damage by dust particles, which reduces the fatigue limit in this pressure range.

The trend in the development of reciprocating high-pressure metering pumps is therefore towards diaphragm technology. Exceptions are the above-mentioned piston designs for HPLC analysis, where extremely favorable operating conditions prevail.

For process and production engineering, piston metering pumps with a low-frequency, long-stroke linear drive (Fig. 25), also known as pressure converters are suitable for very high pressures (1500 – 4000 bar) (Fig. 23 i). The uniform, slow piston motion leads to an operationally reliable piston seal and, when two pump heads are superposed, to low pulsation of the flow rate [35].

Designation		Functional principle	Displacer kinematics	Adjustment	Displacer system
Spindle drive one- and two-cylinder electro-mechanical hydraulic	a		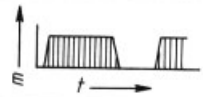	$n(h)$ $n < 30\ \mathrm{min}^{-1}$	P DH
	b	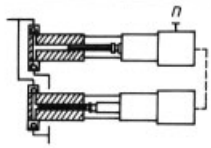	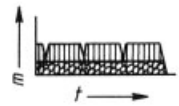		
Cam with pulsating angular velocity	c	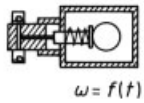	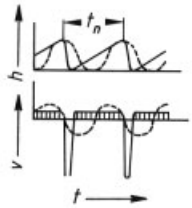	n $n < 60\ \mathrm{min}^{-1}$	P DH
Superposition drive two cylinder with cams	d	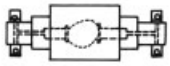	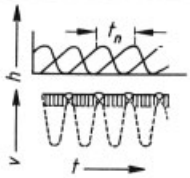	n $n < 100\ \mathrm{min}^{-1}$	P DH
Linear drive single-cylinder magnetic	e	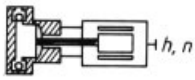	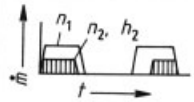	h, n $n < 100\ \mathrm{min}^{-1}$	P DH
Stroke-adjustable quasi-harmonic drive single and multi-cylinder	f	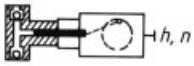	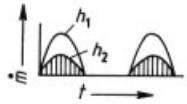	$h(n)$ $n < 350\ \mathrm{min}^{-1}$	P DH
	g	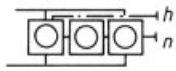	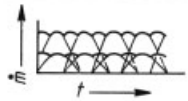		
Stroke-adjustable drive with lost motion drive	h	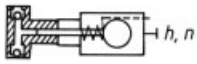	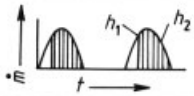	$h(n)$ $n < 150\ \mathrm{min}^{-1}$	P DH
Linear drive super- position, two-cylinder hydraulic	i	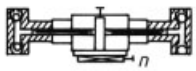	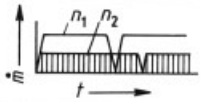	n $n < 100\ \mathrm{min}^{-1}$	P DH

Figure 23. Types of reciprocating metering pumps
P = piston pump head; DH = diaphragm pump head (hydraulic)

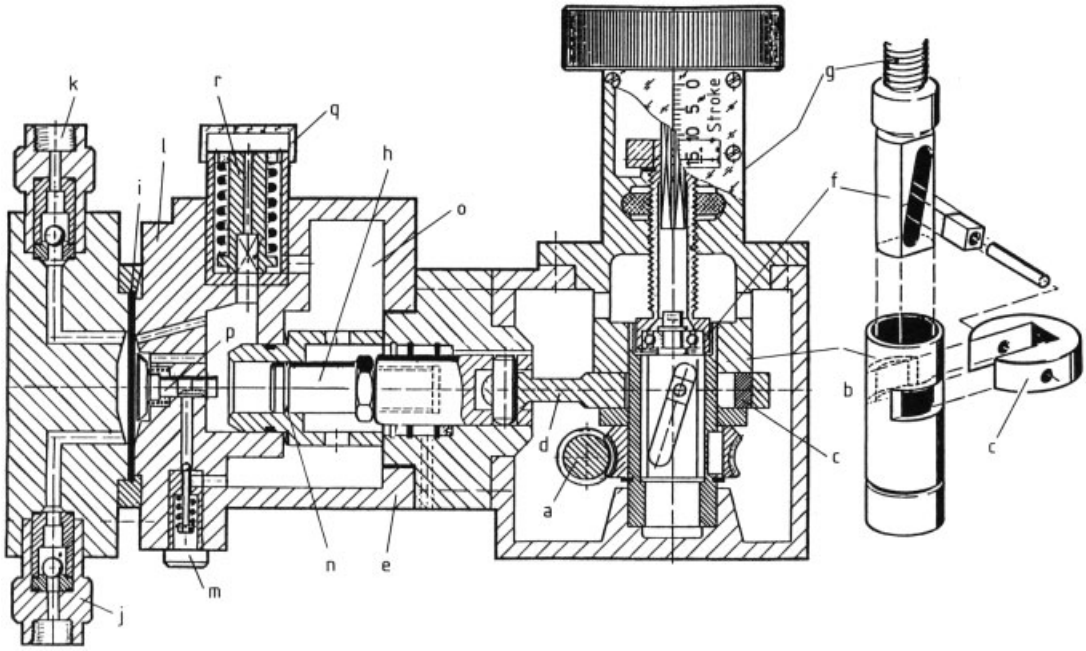

Figure 24. Diaphragm metering pump with stroke adjustment through an eccentric adjustment system (Lewa)
a) – g) Eccentric adjustment drive; h) Piston; i) Diaphragm; j), k) Pump valves; l) Pump cylinder; m) Replenishing valve; n) Piston seal; o) Hydraulic supply vessel; p) Sliding gate; q) Overflow valve; r) Venting valve

For pressures of several thousand bar, such as those required for metering in polyolefin production or for water-jet cutting, the piston heads are designed with coaxial pump valves to avoid the notch effect (see Fig. 28 and Chap. 3.6.1). Hydraulic pressure converter pumps can also be easily controlled at high powers for automation purposes.

3.3.1.2. Transfer Pumps

In contrast to metering pumps, which generally have a single cylinder and an adjustable stroke, transfer pumps are always designed with several cylinders (e.g., triplex) in a multihead structure, in order to smooth out pulsations in the flow rate (Fig. 23 g). Powers over 1 MW are achieved in a vertical or horizontal mode of construction. Piston pump heads currently still predominate. Drives for multihead piston and diaphragm pumps are very similar (Fig. 26 a). At high pressure, and when using corrosion-resistant high-strength steels, the mono-block mode of construction (Fig. 27 A, B) is replaced by a component mode of construction (Fig. 27 C). The advantage is that the parts which are most subject to fatigue can be replaced more easily. Above 100 – 300 bar, a coaxial valve arrangement (Fig. 28) is suitable because, in this way, the notch stresses can be reduced by at least a factor of two compared to the customary T-intersection of the pump head bores (see Chap. 3.6.1). Such designs can deal with differential pressures up to ca. 1500 bar with long-term operational reliability. Such pressures are presently required, for example, in high-pressure cleaning technology.

Diaphragm transfer pumps in a multihead structure with powers up to about 1 MW have also been built for the pressure range up to 1000 bar (Fig. 26 B) [33, 36]. The avoidance of piston seals with fluid contact and leakages leads to remarkable operational reliability. For this reason, multihead diaphragm pumps are economical for hazardous fluids at high pressures, as compared to multihead piston pumps, despite the higher investment costs.

With the development of pipeline transport of coal or ores, the chemical refining of ore, and coal hydrogenation, high-pressure pumps

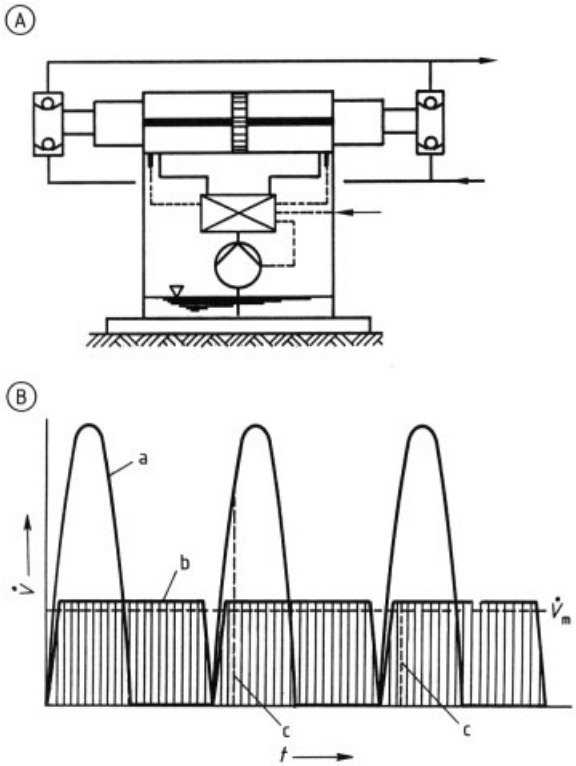

Figure 25. Superposition system with hydraulic linear drive (Uhde)
A) Principle arrangement; B) Pulsation of the fluid flow
a) Single-cylinder piston pump; b) Two-cylinder linear drive (as in Fig. 25 A); c) Basic effect of reduced volumetric efficiency due to elasticity
$\dot{V}_m$ = average flow rate of fluid; t = time

are currently being used for suspensions that are loaded with solids. There are various methods to keep abrasive particles away from the piston seal, e.g., hydraulic isolating lines, sedimentation spaces, and flushing systems [37]. For predominantly aqueous, abrasive slurries with a high concentration of solids, multicylinder diaphragm pumps are suitable up to a pressure of 200 bar and for powers of ca. 1.5 MW. They generally have a four-cylinder arrangement and elastomer diaphragms [38].

A new branch of high-pressure engineering is the conveyance of materials with high consistency, e.g., mortar, concrete, sludge, slurries, over long distances for which pressures of up to 200 bar may be necessary. This can be carried out with slow hydraulic long-stroke piston displacement elements if the inlet opening of the two-cylinder arrangement is controlled with sliding switches which operate as valves [39].

3.3.2. Rotary Displacement Pumps

At high differential pressures, rotary displacement pumps are suitable only for viscous fluids. As the viscosity increases, the internal sealing clearances seal better, and the pump efficiency and the attainable pressure difference reach satisfactory values. Gear and screw pumps are suitable for process engineering applications up to 300 bar, with fluid viscosities > ca. 100 mPa · s. Typical applications for gear pumps are the injection, conveyance, filtration, and spinning of polymer melts (viscosities $>10^6$ mPa · s). Delivery pressures up to 450 bar occur in these applications. In the case of hot and direct spinning processes, the heated high-pressure gear pump with a large inlet cross-section takes the polymer melt from a reactor, which may be under vacuum, and presses it through filters towards the gear spinning pumps, whose flow rate is es-

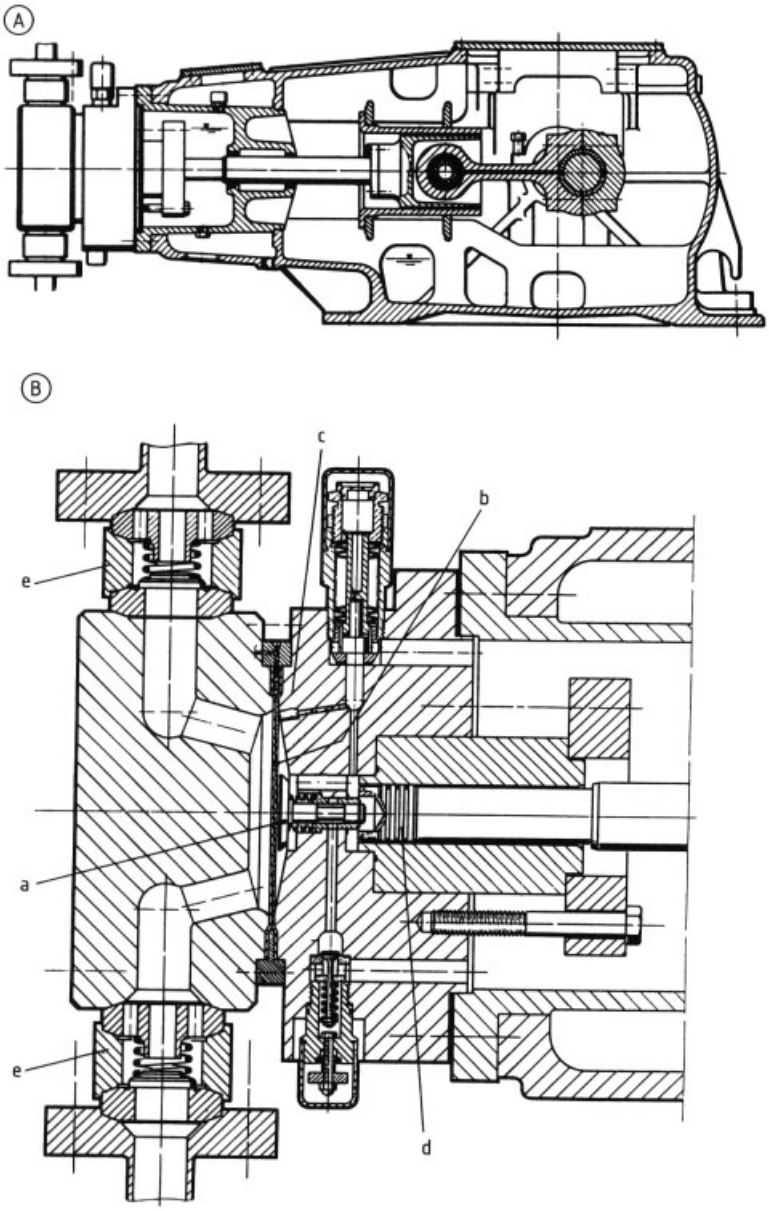

Figure 26. Process diaphragm-pump for high pressure (Lewa)
A) Propulsion system with straight thrust crank drive for multihead arrangement; B) Diaphragm pump head for the pressure range up to 500 bar
a) Sliding gate for controlling the position of the diaphragm; b) Diaphragm (PTFE, sandwich form); c) Hydraulic cylinder; d) Piston with seal; e) Pump valves

pecially uniform and does not depend greatly on the delivery pressure. Gear pumps (Fig. 29) are used behind compounding extruders to increase the pressure (e.g., from 100 to 300 bar) and to smooth out fluctuations in the extruder output [40–44].

3.3.3. Centrifugal Pumps

Centrifugal pumps require high circumferential speeds and several or many stages at high pressure differences. Economic application of centrifugal pumps depends on the specific speed of

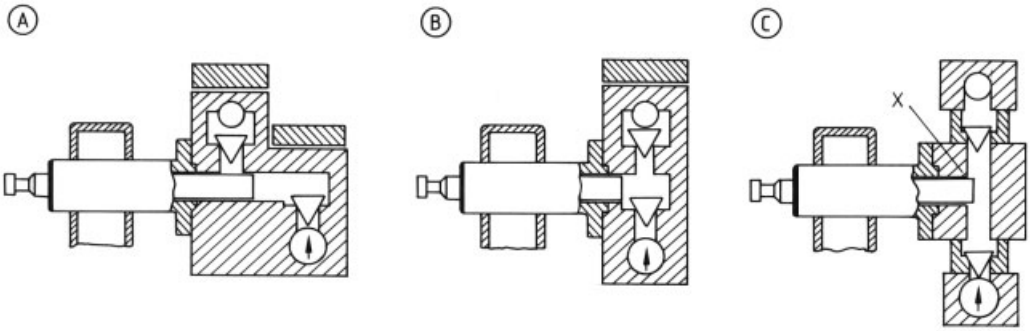

Figure 27. Monoblock and component mode of construction (URACA)
A) Monoblock, valves displaced; B) Monoblock, valves flush; C) Component mode of construction

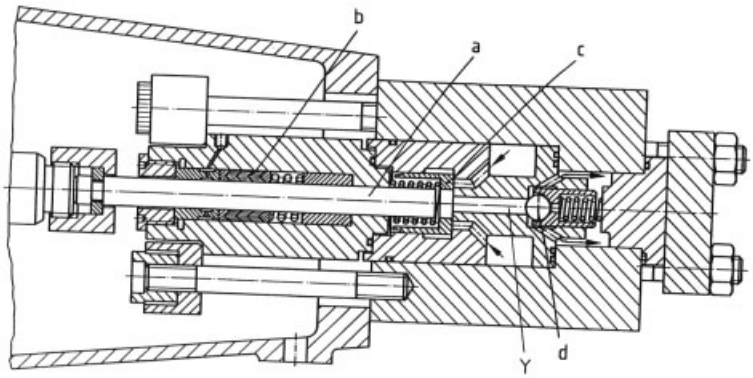

Figure 28. Coaxial arrangement of the pump valves
a) Piston; b) Piston seal; c) Suction valve; d) Pressure valve

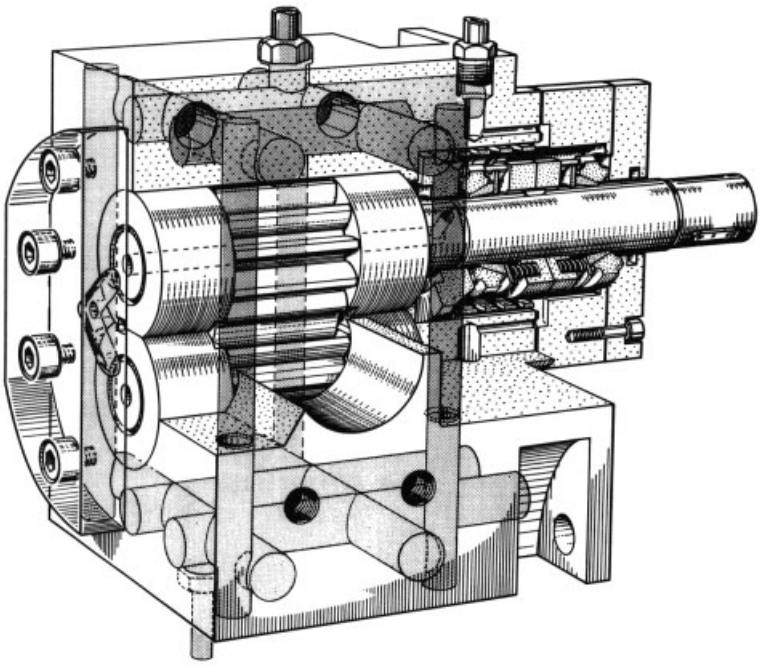

Figure 29. High-pressure gear pump (heated) for polymer melts (Maag) [44]

the single stage (at least 10 – 20 rpm) and thus on the magnitude of the flow rate and the hydraulic power.

3.3.3.1. Multistage Centrifugal Pumps

High-pressure centrifugal pumps have long been used as feed pumps for steam boilers [45,46]; power extends to beyond 30 MW, and pressures up to 400 bar. The design of high-pressure centrifugal pumps for process engineering varies depending on the pressure level, the temperature conditions (including those during starting), and the operational requirements, as well as the directives of the international regulations.

For high-pressure process engineering, centrifugal pumps with vertically-split barrel construction have become established for powers above ca. 1 MW in petrochemistry as well as in petroleum and natural gas exploration, especially as injection pumps [47,48]. Injection pumps for surface water and seawater impose much more stringent requirements as regards corrosion and erosion resistance of their components (power up to ca. 20 MW, pressure up to 300 bar). The entire running gear can be withdrawn without disassembling the housing, which facilitates repairs (Fig. 30). For pressures below 250 bar and lower powers, multistage centrifugal pumps are also used with staged housings and modular construction. They generally require lower investment costs. For the pressure range up to 150 bar, a longitudinally divided (horizontally split) housing design (pipeline pump) is suitable for many applications, including boiler feeding and transferring liquids into chemical processes.

3.3.3.2. High-Speed Centrifugal Pumps

High-speed single-stage and two-stage centrifugal pumps have been introduced (Fig. 31). They now represent a significant alternative to multistage centrifugal pumps in the range up to 100 m^3/h. The required high circumferential speed (max. 200 m/s) is effected by high rotor speed by using gears (up to ca. 30 000 rpm), so that pressure increases per stage up to about 200 bar are possible. Although the efficiency is then generally poorer than in the case of multistage pressure increase, high-speed centrifugal pumps have a compact construction [49].

Two-stage gear-centrifugal pumps (Fig. 31 B) achieve higher pressures than single-stage pumps (Fig. 31 A). However, the shaft seal of the second stage is exposed to higher pressure; the slow-running booster stage that is shown in Figure 31 B is used to increase the suction pressure or the head above vapor pressure to avoid cavitation.

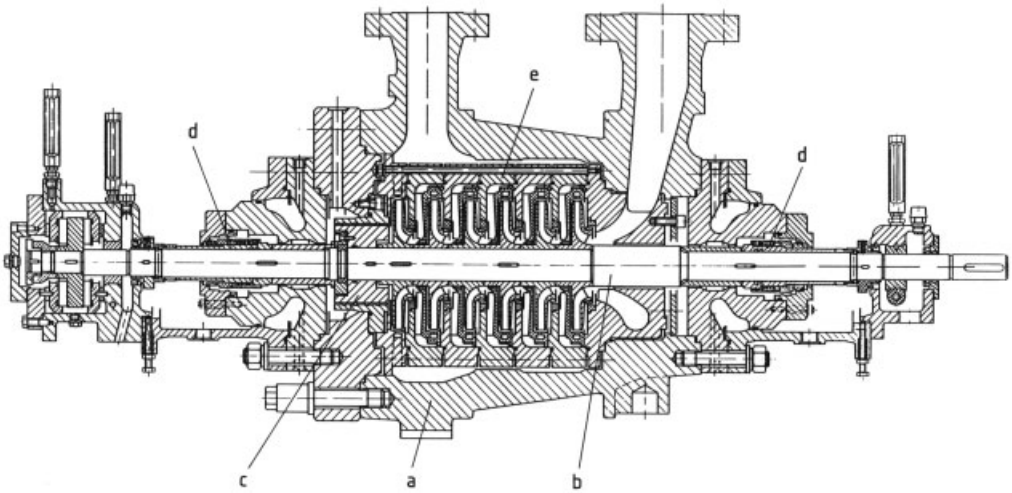

Figure 30. Six-stage vertically split barrel housing centrifugal pump (Sulzer)
a) Barrel housing; b) Shaft; c) Axial thrust compensation; d) Shaft seal; e) Six internal pump stages

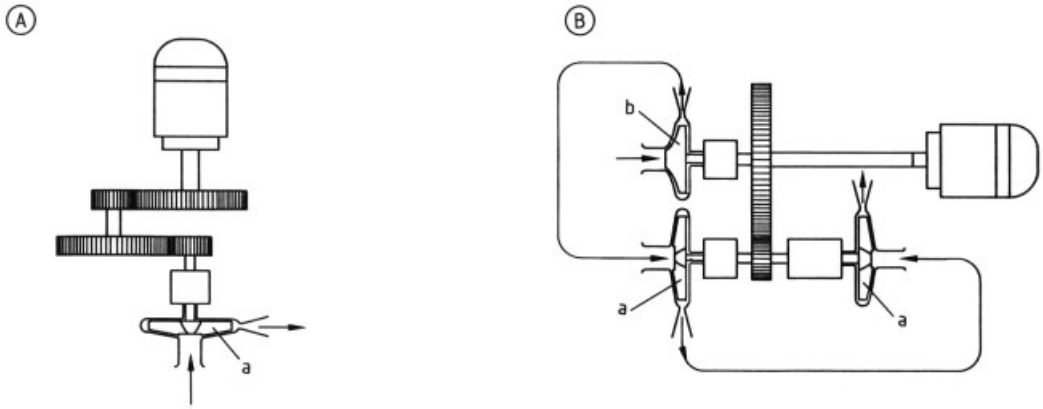

Figure 31. High-speed gear-centrifugal pumps (Sunstrand)
A) Single stage; B) Two stage with booster stage for improving the NPSH; a) Stages; b) Booster stage

For hydraulic powers below 1 MW (pressures from 100 to 300 bar), only a detailed cost-benefit calculation will generally indicate the most favorable choice between multistage cen-trifugal pumps, high-speed one- or two-stage centrifugal pumps, or reciprocating multihead piston or diaphragm pumps. The energy consumption and expenses for investment and maintenance determine the optimal solution.

3.3.3.3. Hermetic Centrifugal Pumps

With centrifugal pumps, there is also a trend towards avoiding leakages and problems with shaft seals by means of a leak-free design. Normally, the shaft seals of centrifugal pumps are exposed only to intake pressure, and thus can be dealt with relatively easily. With higher intake pressures or system pressures, it is recom-

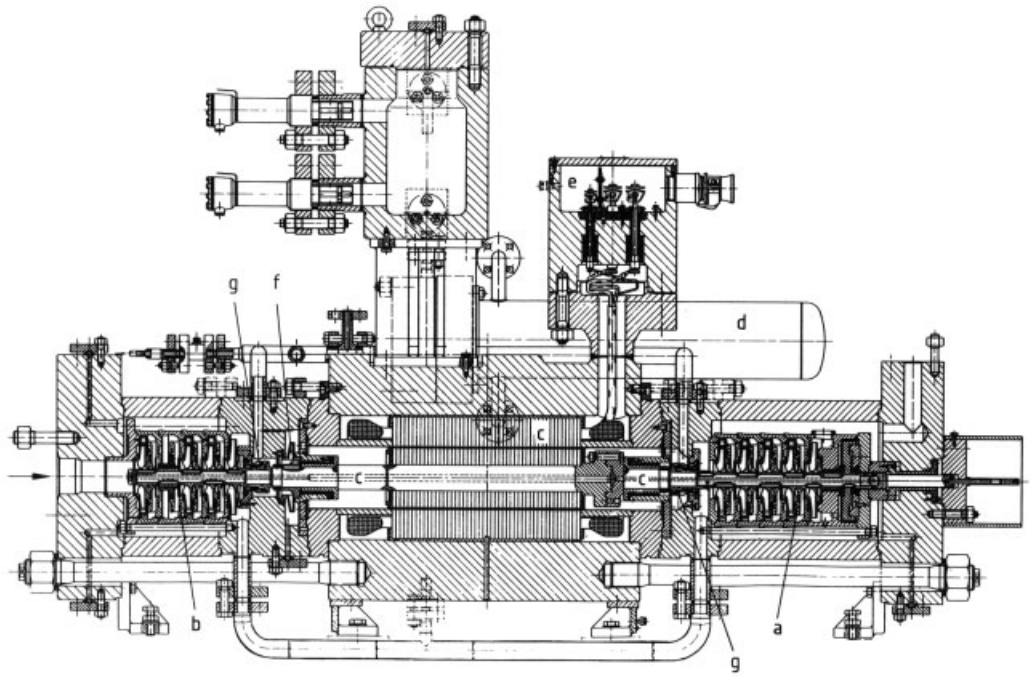

Figure 32. Ten-stage canned-motor centrifugal pump in tandem arrangement for carbon dioxide circulation (HERMETIC)
a), b) Tandem rotor; c) Canned motor; d) Cooler; e) Terminal box; f) Cooling circuit impeller; g) Mechanical shaft seals

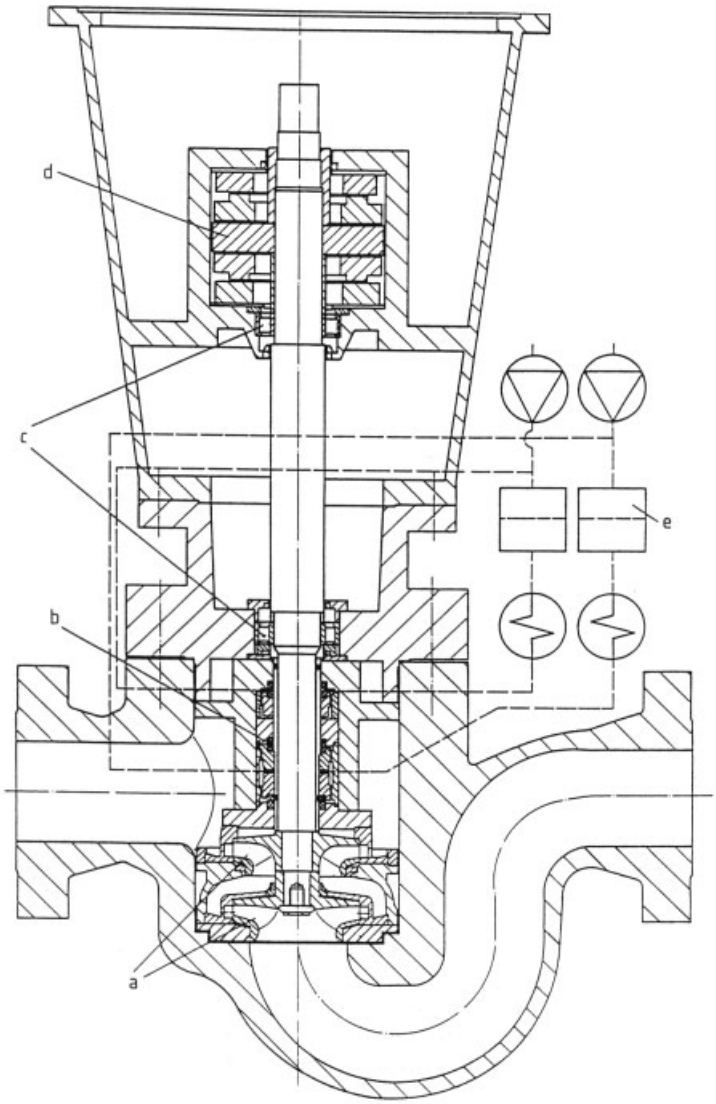

Figure 33. Two-stage high-pressure centrifugal pump for circulating carbon dioxide at a system pressure of 300 – 400 bar (Grassel)
a) Pump stages; b) Sealed high-pressure mechanical seal (two-stage); c) Radial bearing; d) Axial bearing; e) High-pressure blocking system

mended that centrifugal pumps with a canned-motor drive be used, especially in the power range below 500 kW and if the fluids are toxic or hazardous [50, 51]. Experience is available up to system pressures of 1000 bar; the support of the thin-walled containment shells of austenitic steel in the grooved stator against the high pressure is especially important. The rotor bearings, lubricated by the delivery fluid, are now frequently made of silicon carbide.

Figure 32 (see previous page) shows a multistage canned-motor centrifugal pump with a separate water-cooling circuit for the motor, such as is used, for example, to circulate carbon dioxide at 300 bar system pressure. The pump rotor is mounted in the motor area; the tandem pump

stages and the cooling-circuit impeller operate over-mounted on both sides of the bearings. The entire internal cooling flow passes the canned motor. The cooling circuit is separated from the liquid circuit by mechanical shaft seals (semi-hermetic design).

Permanent magnet drives for pumps and stirrers are increasingly being used in high-pressure technology [50, 52, 53]. The containment shell of the magnet drive, in contrast to canned motors, must take up the full pressure difference and must have appropriately thick walls.

If larger flows of nonproblematic fluids must be circulated at high pressure (e.g., carbon dioxide at 400 bar), two-stage mechanical seals with pressure blockage are sometimes used. The axial thrust is taken up, for example, by a segment bearing (Fig. 33). Hydrodynamic mechanical seals make it possible to control system pressures up to 300 bar even in a single stage [54].

3.4. Compressors

Optimal compressor types for the various power ranges have been developed. Figure 34, which shows power ranges for various compressors, was calculated on the basis of 1 bar intake pressure and an isothermal efficiency of 64 % [55]. It can give only approximate reference points for the most favorable area of application, which varies depending on the manufacturer. If the intake pressure is greater than 1 bar, as is the rule, it is necessary to recalculate. (For example: Compression of ethylene at a flow rate of 6400 kg/h $\approx$ 51 130 m^3/h (STP) from 231 to 2151 bar, corresponds to a real intake volume flow rate at 231 bar of 161 m^3/h and a power consumption of 8200 kW).

The application boundary between piston and radial turbo compressors lies approximately at 10^4 m^3/h intake volume flow rate (power > 1 – 5 MW). In individual cases, the most economical solution must be determined in the boundary region. Development of piston compressors has partially stagnated in recent years. Rotating screw-type compressors and radial turbo machines have taken over certain areas. For large, multistage horizontal process-gas piston compressors, built up to powers in excess of 10 MW, development has been largely concluded. The trend toward radial turbo compressors dominates for economic reasons. Their advantages are that they operate with little vibration and pulsation, are more easily sealed, and require less space.

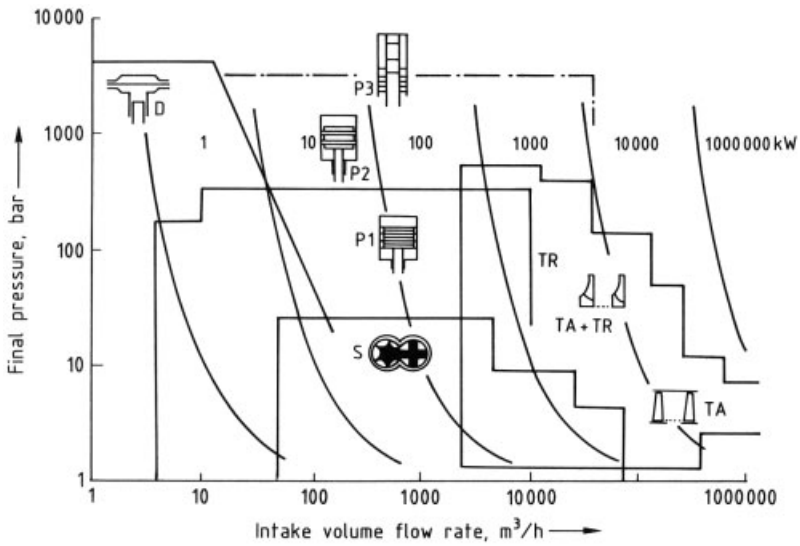

Figure 34. Power range of various compressor types for high pressures [55]
D = diaphragm compressor; P 1 = piston compressor, dry running (piston ring, labyrinth); P 2 = piston compressor, lubricated (piston ring); P 3 = ultrahigh-pressure compressor, lubricated (plunger piston); TR = radial turbo compressor; TA = axial turbo compressor; S = screw compressor

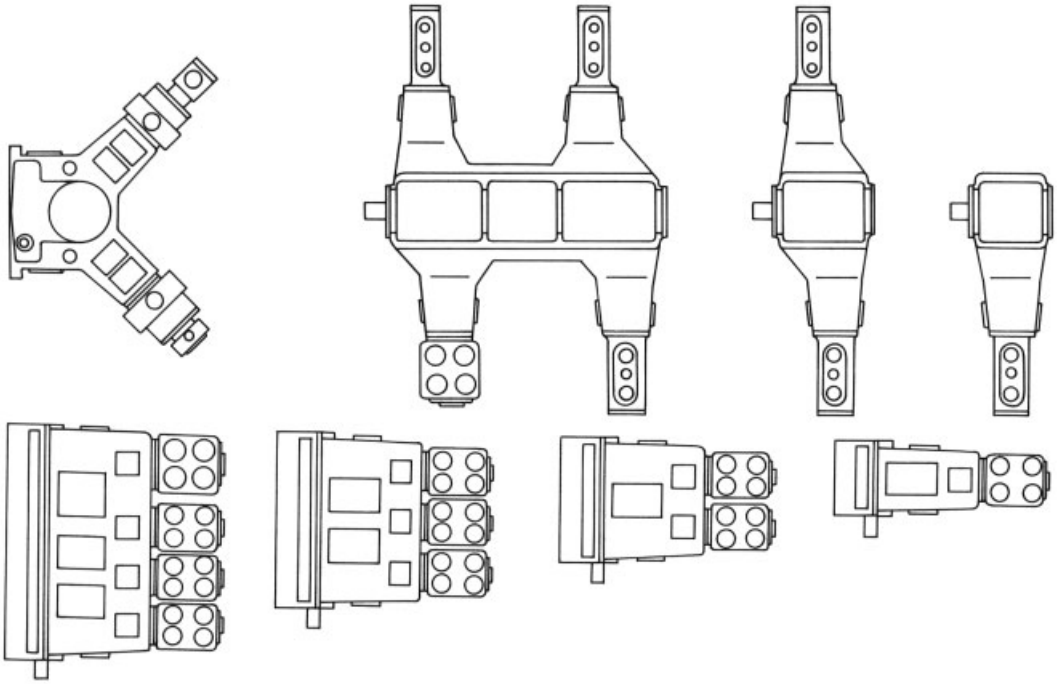

Figure 35. Piston compressor in modular construction (Neuman & Esser)

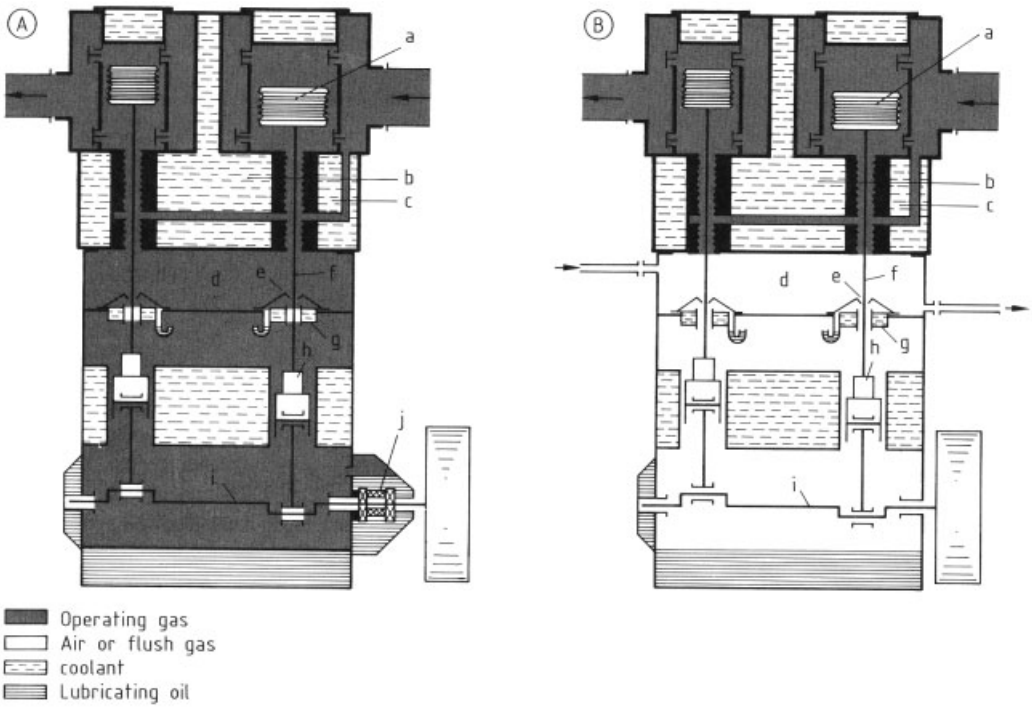

Figure 36. Sealing concepts for labyrinth compressor [58]
A) Pressureproof housing; B) Spacer
a) Labyrinth piston; b) Cylinder; c) Labyrinth piston rod seal; d) Spacer; e) Oil wiper; f) Piston rod; g) Guide bearing; h) Cross head; i) Crank shaft; j) Mechanical seal

3.4.1. Piston Compressors

At lower powers (<2 MW), there is much development of piston compressors in the direction of a higher stroke frequency, a more compact modular construction with a horizontal, vertical, or V-shaped arrangement (Fig. 35, see previous page), dry-running compression, sparing lubrication, and simpler maintenance [56–59]. The dry-running range for continuous operation extends up to about 300 bar, depending on the type of gas. Labyrinth and piston-ring compressors have characteristic fields of application. The completely encapsulated labyrinth compressor with a pressureproof housing is suitable for gases of higher density (Fig. 36, see previous page). Dry-running piston-ring compressors can be used universally as far as gas density is concerned; to achieve satisfactory running times (>10 000 h), modern design and material concepts must be implemented for the piston rings. The so-called sparing lubrication also plays a role here in reducing the lubricant contamination of the gas. However, most high-pressure piston compressors are designed to be lubricated so that trouble-free running times above 10 000 h are the rule.

Ethylene secondary compressors are a special case. These compress typically from ca. 200–300 bar to 2000–3500 bar in two stages, and are built for powers of up to ca. 10 MW [60–62]. In contrast to crank compressors, their drives are designed with an interior, robust cross-head guide (Fig. 37 D) and pairwise opposite cylinders. The most important principle is to keep any transverse forces away from the piston seal. For components subject to cyclical pressure, notches are avoided (Fig. 38, see next page). Valve bodies with cone valves are used here; at large capacities several valves are arranged in parallel. Superpressure compressors mainly work with stationary piston sealing elements; this is a fundamental difference from standard piston compressors with mobile piston rings. Only with a stationary sealing arrangement does one achieve the required effective high-pressure lubrication, which furthermore requires extremely hard, smooth, and precisely machined plunger pistons.

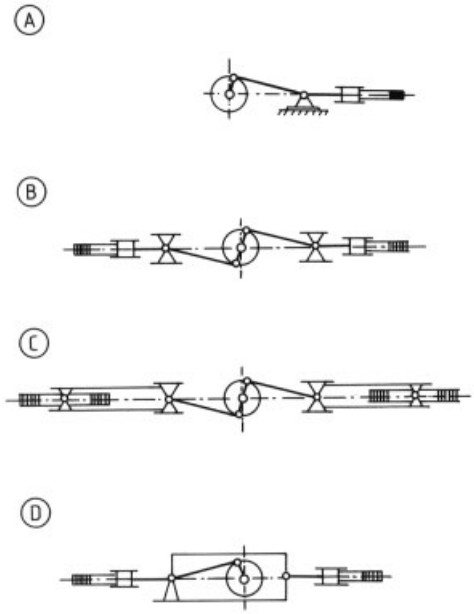

Figure 37. Designs of propulsion systems for high-pressure piston compressors
A) Vertical crank; B) Horizontal crank; C) Horizontal crank (double action); D) Horizontal sliding block

3.4.2. Diaphragm Compressors, Laboratory High-Pressure Compressors

The diaphragm compressor (Fig. 39, see next page), one of the oldest leak-free machines, is limited to lower powers (<100 kW) because of its sensitive metal diaphragms which must be used because of the high compression temperatures. Its applications include laboratories, pilot installations, and special production facilities. An attractive feature is their very high pressure ratio (up to 20) produced by a single stage as a result of the small clearance volume and good cooling. At every stroke, the diaphragm is pressed against the upper contact surface by a hydraulic fluid (smallest clearance volume) [63]. However, for this reason it is sensitive to particles, coatings, liquid droplets, and surface defects, so that its lifetime is generally less than 5000 h. It is possible to signal diaphragm rupture by means of sandwich diaphragms [64]. Diaphragm characteristics and rupture control are similar to diaphragm pumps (see Section 3.3.1.1).

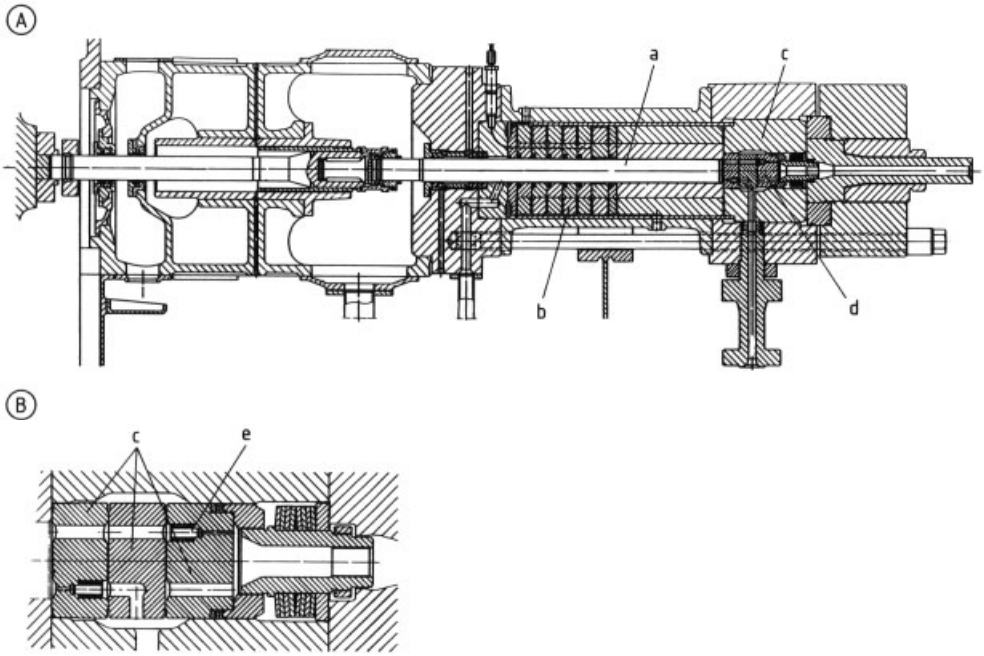

Figure 38. Final stage of an ultrahigh-pressure secondary compressor (Sulzer Burkardt) [60]
A) Overall arrangement; B) Valve block
a) Plunger piston; b) Stationary sealing elements; c) Valve housing; d) Compressor valve arrangement; e) Plug valve

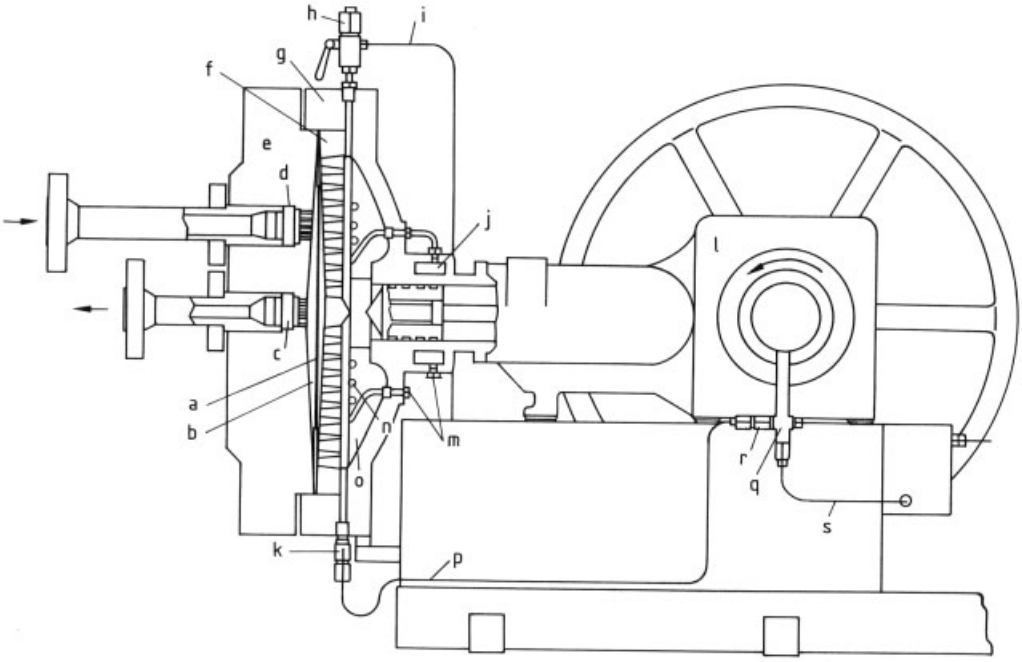

Figure 39. Diaphragm compressor (Hofer) [63]
a) Diaphragm; b) Gas space; c) Pressure valve; d) Suction valve; e) Diaphragm cover; f) Perforated plate; g) Hydraulic cylinder; h) Oil overflow valve; i) Oil return; j) Cylinder cooling; k) Check valve; l) Crank drive; m) Cooling water in/out; n) Oil cooling coil; o) Oil chamber; p) Oil injection (leakage compensation); q) Compensation pump; r) Check valve; s) Oil supply

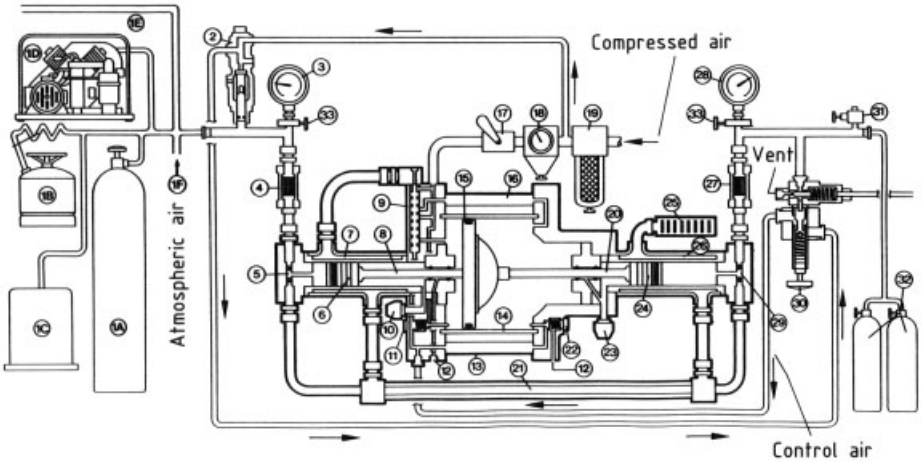

Figure 40. Two-stage filling compressor with pneumatic linear drive (Haskel)

Diaphragm compressors are built for pressures up to 4000 bar, at least for short-term operation. They are economical when leak-free operation with no gas contamination is required at low volume flow rates and at high pressure.

For very small delivery quantities (0.1 – 1 m^3/h) and pressures up to 2000 bar, air-driven piston and diaphragm compressors (Fig. 40) are suitable for transferrring and compressing gases. Because of their compact construction, their explosion protection, and their ease of regulation, such machines are attractive for short-term operation in research facilities, especially for booster and filling applications.

3.4.3. Turbo Compressors

For large volume flow rates (2×10^3–10^5 m^3/h), radial turbo compressors have now largely displaced process-gas piston compressors even in the range above 150 bar. Typical applications in urea, ammonia, and methanol plants, for compressing synthesis, natural, and refinery gas, as well as for transporting and injecting gas, extend up to the pressure range of 400 bar [65].

For high-pressure technology, radial compressors are designed similar to high-pressure centrifugal pumps, in vertically split barrel construction. For hazardous gases, this design is also used for pressures beginning at about 80 bar on account of the more favorable seal. There are good possibilities for compensating the axial thrust by an appropriate arrangement of impeller stages with intermediate cooling.

Multistage gear – radial turbo compressors are suitable as preliminary stages for high-pressure stages for pressure ratios up to about 50 bar [66, 67]. For pipeline gas transport, single-stage and two-stage vertically split barrel designs have been produced for final pressures up to 80 bar and for powers above 20 MW. These designs are easy to install and maintain.

A series of prototypes were developed to expand the range of application of turbo compressors. By way of example, Figure 41 (see next page) shows a six-stage high-pressure stage (350 – 800 bar). The impellers are machined smooth and are made up of two disks (e.g., by soldering); an exit gap of a few millimeters can be achieved here. Nevertheless, the shaft seals here must seal against the suction-side pressure level (350 bar); they work with pressurized oil blockage [68].

3.5. Other High-Pressure Machines

The problem of sealing rotating shafts is also encountered with high-pressure stirrers and high-pressure extruders, which greatly resemble rotating pumps [40–44, 52, 53]. High-pressure sluices for bulk goods present special sealing problems [69]. A new type of high-pressure machine has arisen in the area of high-pressure extractors with automatically opening locks [70].

3.6. Special Problems Involving High-Pressure Machines

3.6.1. Strength of the Components

Since high-pressure components are generally subject to high loads, a precise stress analysis is necessary, possibly with the finite element method or with photoelastic testing. This is especially true for impellers that have a high circumferential velocity, and also for the vibrational analysis of rotor shafts.

In general there is no special problem about dealing with stresses in statically loaded pressure housings especially since a suitable stress flux can achieve a favorable strain (e.g., vertically split barrel design). By contrast, it is much more difficult to dimension components that are stressed in cyclical fashion by internal pressure. The largest cyclical internal pressure stresses occur with reciprocating liquid pumps. The permissible dynamic material stress lies below the yield point of the material. The maximum stresses occur at notch points.

Compared to the "smooth" thick-walled pipe, the usual notch points (Fig. 42) have notch factors from 1.5 to 3.5 [27,71]. The notch factor characterizes the local stress sizing effect. Thus in the T-intersection crossbore (Fig. 27), at point X the tangential and radial stresses superimpose, whereas the smooth pipe section (Fig. 28, point Y), is loaded only by the stress of the corresponding single bore. Studies on ferritic tempering, soft martensitic, semiaustenitic, and precipitation-hardened high alloy chromium – nickel steels, make it possible to design dynamically loaded thick-walled components [26,27,72–80]. With tough, high-alloy chromium – nickel steels, an increase of their dynamic loadability by autogeneous dynamic autofrettage (see Section 2.3.1.3) has been observed which explains the favorable behavior of these steels in practice. Studies show that thick-walled forgings can exhibit anisotropy, and that the surface quality has a perceptible influence. Furthermore, the mechanical properties of the material depend on its purity. With components subject to very high stress, the use of high-purity steel grades, smelted under inert gas or in vacuum, are therefore always recommended. All-round forging promotes isotropy.

At very high pressures, for parts made of brittle steels that are subject to cyclical stresses, a clear discrepancy exists between the mechanical properties of the material for uniaxial and thick-walled samples [76]. Various studies have shown that the penetration of pressures into the microcracks (e.g., slag inclusions) which are al-

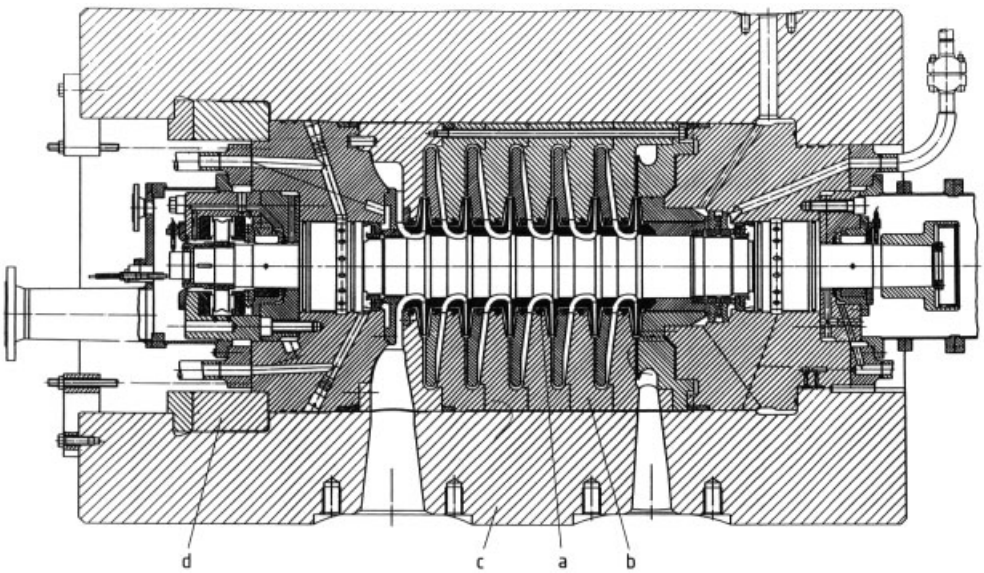

Figure 41. Six-stage radial compressor with final stage 350 – 800 bar (Mannesmann-Demag)
a) Rotor with compressor stages; b) Stage housing; c) Vertically split barrel housing; d) Seal for vertically split barrel housing

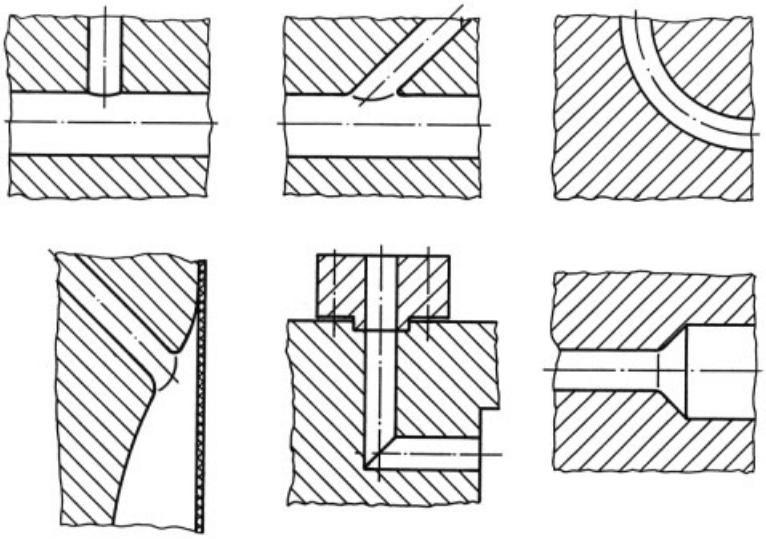

Figure 42. Typical notch points in thick-walled high-pressure components

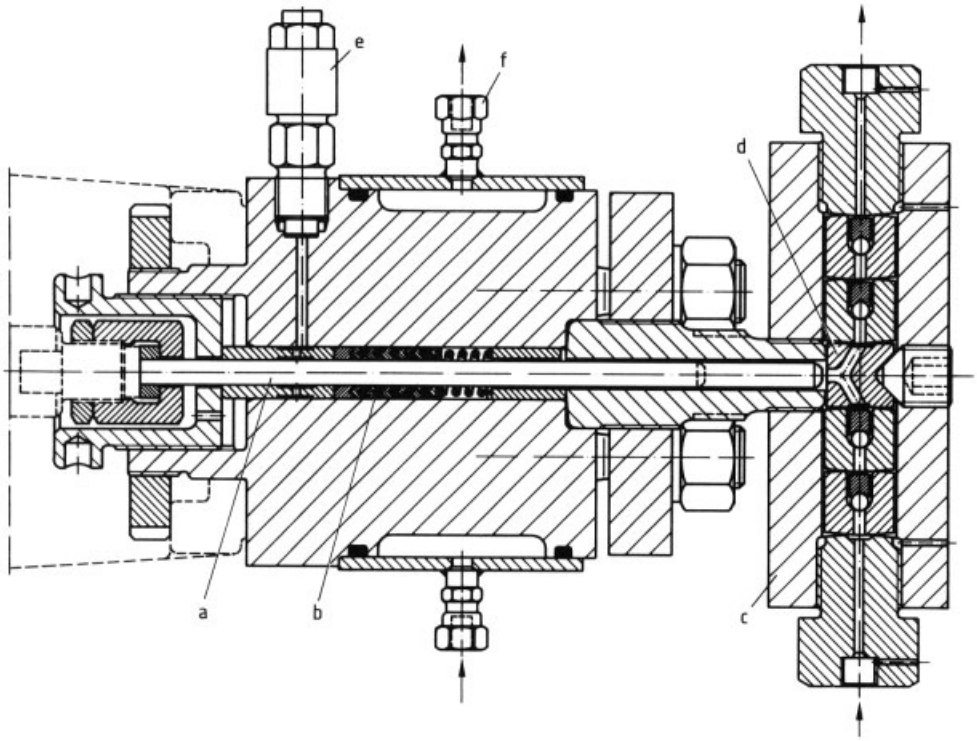

Figure 43. Piston pump head for the pressure range up to 3000 bar (Uhde)
a) Piston; b) Piston sealing elements; c) Valve head; d) Y-intersection; e) Lubrication; f) Cooling

ways present in the material creates a noticeable additional stress. However, with high-alloy chromium – nickel steels, these stresses appear to be reduced by local plastification.

For thick-walled components, the following design measures help to deal with cyclical pressures:

1) Avoiding notches, for example, by a coaxial arrangement of the pump valves (see Section 3.3.1.2, Fig. 28). If this cannot be achieved, then the point that is subject to the greatest stress must be rendered safe by giving it a Y-shape (rounded) instead of a T-shape (Fig. 43, see previous page). A tangential layout of the intersecting bores can also yield advantages [80].
2) Application of compressive prestresses, in the limiting case up to pure cyclic compressive stress by autofrettage, cold working, shrinkage, or local shot-peening.
3) The use of pure, isotropic materials that are as tough as possible, and the avoidance of corrosion, which lowers fatigue strength [81]. Reliable data on the fatigue strength of thick-walled components have recently become available for modern, high-strength soft-martensitic, semiaustenitic, and precipitation-hardened steels [82].

In general, the design of components that are subject to cyclic pressure requires great experience. In compressor construction, the problems are generally easier to solve because of the smaller pressure amplitudes.

3.6.2. Seals

In high-pressure technology, metallic seals – lenses, ring joints, cuts – are used as *static seals*. Self-sealing elastomer sealing rings, e.g., O-rings, have also proved suitable for easily loosened connections as long as the pressure, temperature, and fluid permit their use.

Where such seals are unfavorable, for example, with horizontally split compressor housing, structural measures, e.g., vertically split barrel housings, are used to guarantee a reliable seal. The clamping and sealing of diaphragms for diaphragm pumps and compressors [33] require special static seals. The problem points in high-pressure machines are the *dynamic shaft* and *piston seals*. Seals create special problems when they slide rapidly under high pressure and if the fluid being sealed does not have any lubricating action. Cyclic stress additionally creates fatigue problems in the seal. *Noncontact seals* with defined narrow clearances, such as labyrinth seals (Fig. 44 A), are less problematic due to the generally low differential pressures (<50 bar); however, the fluid should not have an erosive effect. Differential pressures up to several hundred bar occur at the axial thrust compensation pistons of centrifugal pumps and compressors, but these are distributed over the many labyrinth stages. Labyrinth seals on reciprocating pistons require very precise alignment, and are limited to gases that do not have too low a molecular mass, as well as to pressure differences $\Delta p < 150$ bar (Fig. 44 B).

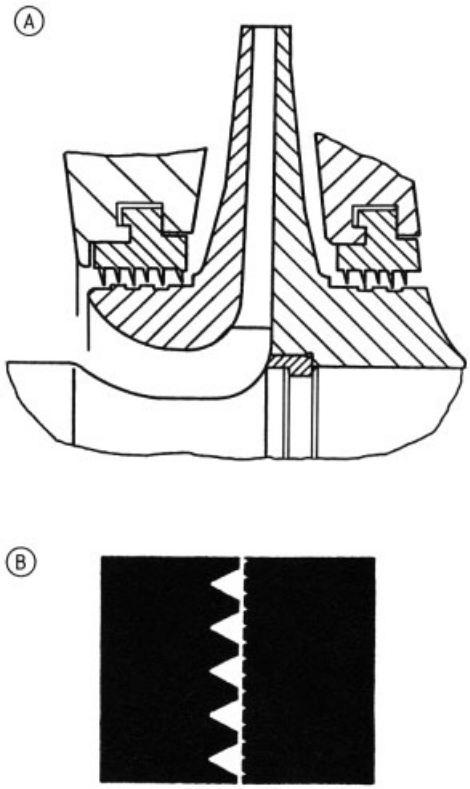

Figure 44. Labyrinth seals
A) Radial impeller; B) Labyrinth piston

The leakage flow of noncontact gap and labyrinth seals is considerable and must be re-

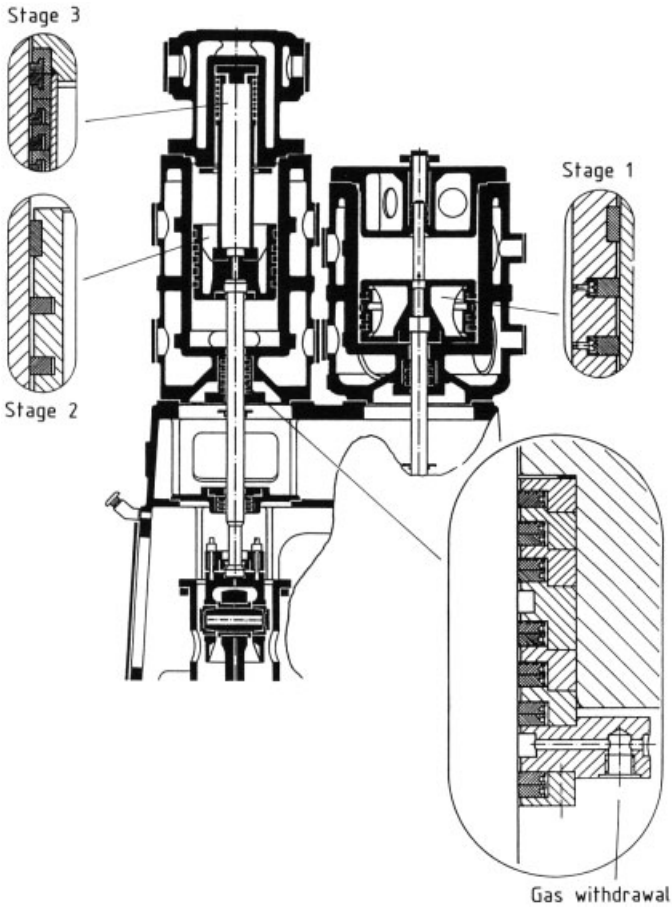

Figure 45. Various designs of piston and rod seals for piston compressors (Neuman & Esser)

cycled internally. With the *floating ring seal*, a typical shaft seal for high-pressure turbo compressors, the high-pressure sealing oil system maintains a slight overpressure compared to the gas pressure at the floating ring, so that the sealing problem is reduced to the gap seal for the oil [68].

Conventional mechanical seals for liquids are limited to pressures below 100 bar [83]. The limits are determined by the product of the sliding speed v and the differential pressure Δp, with the rule of thumb $\Delta p \cdot v < 2000 - 5000\,\mathrm{bar} \cdot \mathrm{m\,s^{-1}}$. This value is based on the friction power arising in the gap. With pressures above 100 bar, sealing pressure blocking systems are used to reduce the differential pressure, to avoid evaporation, to improve lubrication, and to provide cooling [54]. High-pressure mechanical seals are needed for centrifugal pumps only if the system pressure or vapor pressure is high. For lower powers, leak-free canned-motor pumps are generally used.

It is especially difficult to seal pistons with reciprocating motion. Piston rings are suitable as piston seals in piston compressors up to ca. 1500 bar; above 200 bar, lubrication is generally required. With dry-running or sparsely lubricated piston compressors, the use of piston rings made of PTFE composites (Fig. 45) has proven effective as an alternative to noncontact labyrinth seals. At high pressure, the "captive piston ring" is especially interesting (Fig. 45, stage 3). It is supported after a run-in process, and thus makes contact at low compression [84]. For lubricated high-pressure stages up to 700 bar, metallic piston rings are used. Generally these are made of perlitic gray cast iron. At still higher pressures (up to 1500 bar), piston

rings made of forged bronze have also proven suitable.

Piston seals of ethylene secondary compressors for polethylene production are a special case (Fig. 46). Supercritical ethylene is compressed from 300 – 350 bar, in two stages, to 3000 – 3500 bar (specific volume at 4500 bar: 1.5 dm^3/kg). Stationary bronze seals which run against hard metal pistons have proven suitable. The sealing elements consist of a sequence of radially and tangentially slotted rings, generally preceded by a throttle ring. Lubricating fluid is injected between the individual sealing elements against the high pressure produced by the compressor [85]. Similar sealing elements are also successfully used for sealing piston rods at pressures up to 500 bar [86].

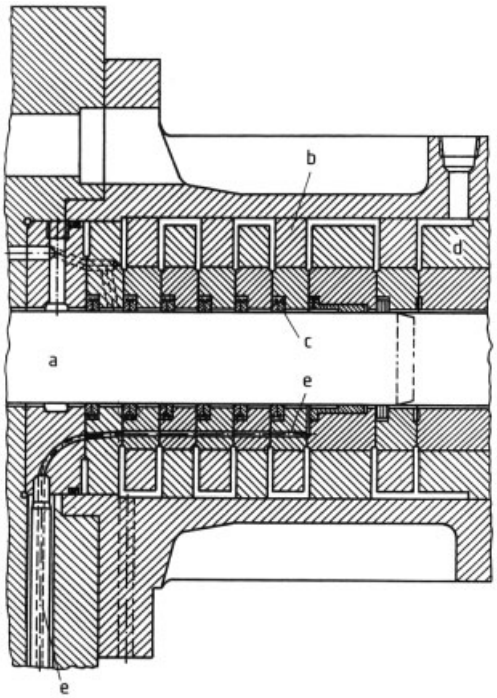

Figure 46. Plunger piston seal of an ethylene ultrahigh-pressure compressor for 2000 – 3000 bar (Nuovo Pignone)
a) Plunger piston; b) Seal housing; c) Sealing rings; d) Compressor cylinder; e) Oil supply

Piston seals for reciprocating displacement pumps must cope with the highest dynamic pressure differences, because compression is always performed in a single stage, as well as with a wide spectrum of fluid properties (e.g., toxicity, corrosivity, abrasiveness, low vapor pressure). Lubrication is often ineffective because the pumped fluid acts as a solvent. The seals vary widely in terms of shape and material. Solid or braided PTFE composites are often used. The piston should be hard, smooth, geometrically precise, and guided centrally so that it runs in the seal "without force". The sealing elements should press just hard enough as is required for sealing. Dry running must be avoided, possibly by lubrication or flushing (for design forms, see Figs. 28 and 43). Spring-tensioned chevrons with lubrication are used successfully up to very high pressures (up to 3000 bar). As the pressure increases, the lifetime decreases quite drastically (e.g., 3000 h at 500 bar, 1000 h at 3000 bar). Packing rings with a separately clamped lubricating or flushing chamber have proven to be a robust design for the pressure range from 200 to 300 bar. With heterogeneous fluids, or with fluids of low lubricity, it is advantageous to inject a flushing fluid into the working space (Fig. 47, see next page).

The maintenance and installation of piston seals requires a great deal of experience; the necessary flushing and lubricating systems are sometimes quite complicated. Frequently, the optimal design (3000 – 8000 h lifetime) must be determined empirically by systematic troubleshooting. At very high pressures (>1000 bar), special measures are needed: pistons and guide bushings of sintered hard metals; extremely precise, smooth machining; and seals made of PTFE compounded with carbon, graphite, ceramic, or metal powder. To this is added an external lubrication, which supports the discharge fluid lubrication, as well as sealing elements which operate as automatically as possible under pressure. A low piston speed reduces the friction power and increases the lifetime of piston sealing elements. In this respect, hydraulic linear drives (Fig. 25) are more favorable. With high-pressure piston seals for abrasive suspensions, the particles must be kept away from the seals by injection flushing or by using sedimentation spaces, as experience with coal slurry pumps (3000 bar) has shown [37].

Problems with piston or shaft seals can be avoided totally by using hermetic diaphragm or canned-motor pumps.

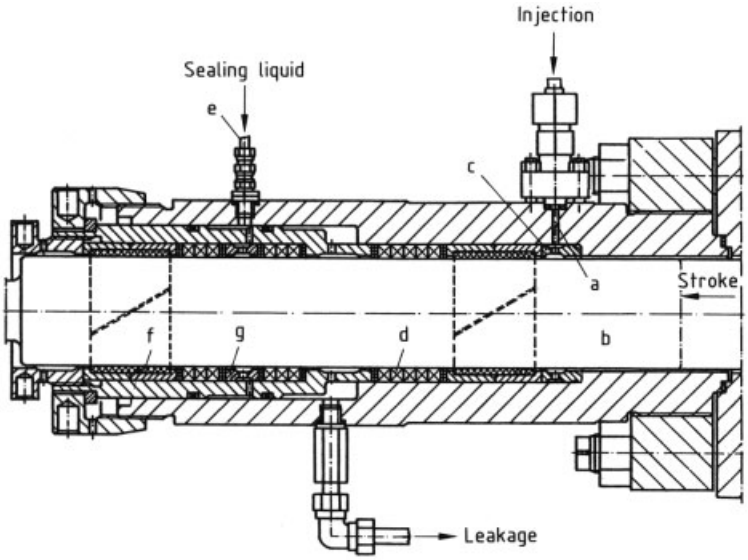

Figure 47. Piston seal with packing rings, lubrication, and injection flushing (URACA)
a) Injection; b) Piston; c) Injection bushing; d) High-pressure seal; e) Flushing, lubricating; f) Guide bushings; g) Lowpressure seal

3.6.3. Wear and Vibration

The wear of automatic pump valves and the determination of pressure oscillations in reciprocating pumps and compressors at high pressure are dealt with in [87–92].

4. Piping and Fittings

Standard DIN parts are not generally used in the chemical industry for high-pressure piping. Since the large-scale chemical industry used high pressures and temperatures earlier than, for example, the power engineering sector, the development of piping elements took place independently and was concluded as early as 1930 for pressures up to 700 bar. The so-called IG Standard developed at that time was used in all German high-pressure plants, and was also common outside Germany; it is still used today. From 1950, further pressure stages up to 4000 bar were added for polyethylene plants.

The following pressure ratings are used for new plants:

Nominal pressure	325 bar for nominal bores up to	200
Nominal pressure	700 bar for nominal bores up to	58
Nominal pressure	3600 bar for nominal bores up to	65

The following steels are predominately used:

CK 15 and CK 35	for pipes, flanges, bolts	up to 700 bar
20 CrMoV 135	for pipes above 200 °C	up to 700 bar
30 NiCrMo 8	for pipes, flanges, bolts	up to 3600 bar
18/8 CrNi steels	for corrosion-resistant piping	up to 700 bar

Pipes are used only in seamless designs. Pipe bends are produced from pipes which are bent when cold or hot, or from hollow cylinders which are bent in a forging die. Tighter bending radii can be achieved with bending in a die, but the reduction in wall thickness arising during bending must be taken into account.

Only forged steel is used for valves and fittings (e.g., T-pieces). Pipes and valves are connected by flanged joints with threaded flange and lenticular sealing rings. The piping itself can be flanged or welded for use at pressures up to 700 bar. Welding is not possible at higher pressures because the steels used here are, with very few exceptions, not weldable.

For ease of production, shut-off valves are designed predominantly as angle valves.

Figures 48 and 49 (see next page) show angle valves for 325 and 3600 bar. A coupling between the upper and lower valve spindles of the 325-bar valve eliminates the need for the spindle to rotate in the seal during opening and closing. V-shaped rings of reinforced polytetrafluoroethyl-

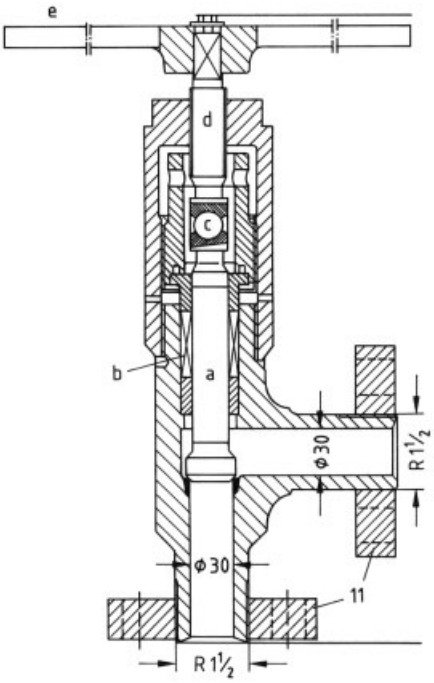

Figure 48. Manually operated angle valve for 325 to 700 bar
a) Lower stem; b) Packaging rings; c) Coupling between upper and lower stem to prevent rotation of lower stem; d) Upper stem; e) Handle

ene have proved to be good packing materials for the gland at temperatures of up to 250 °C. This valve can be operated manually. In contrast, the large valves for 3600 bar can be operated only hydraulically or by an electric drive.

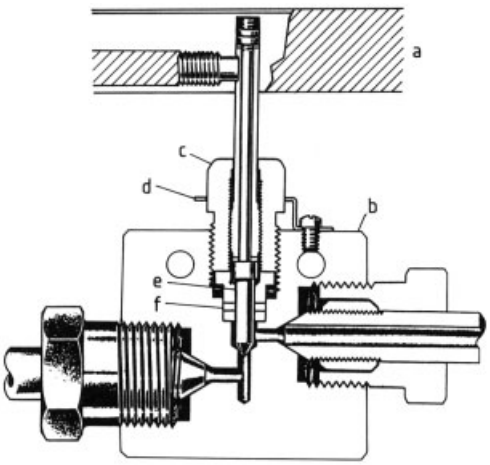

Figure 50. Straight-way valve for small nominal sizes
a) Designer handle; b) Stainless steel body; c) Packing gland; d) Loading device; e) Nonrotating stem; f) Adjustable packing below threads

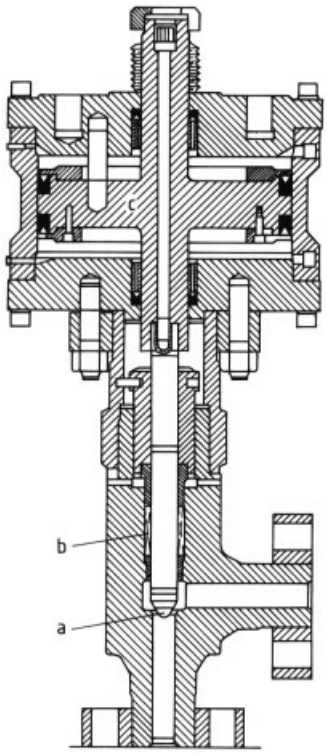

Figure 49. Angle valve with operation by oil plunger for 3600 bar
a) Stem; b) Packaging rings; c) Hydraulic piston

For instrument lines and high-pressure laboratories, frequent use is made of valves of the form shown in Figure 50. The screwed fitting can be used for pipes with an inner diameter of up to ca. 8 mm.

Relief valves are constructed as spring-loaded proportional valves or full-stroke values. Figure 51 shows a full-stroke relief valve for 3600 bar. A stainless steel ball takes over the function of sealing and assisting the stroke, since it is lifted by the back pressure generated during blow-off. The disk spring chamber is designed for an internal pressure of 325 bar, so that the gas to be discharged can be collected in a pressure vessel. The valve can be opened by the lifting magnet even if the internal pressure has not yet risen to the set blow-off pressure. Rupture disks are also used up to 3000 bar instead of the relief valves.

Nonreturn ball valves are manufactured up to nominal width 30 mm; spring-loaded plate valves are used for larger nominal sizes. The check valves with pear-shaped cones that were

developed for coal hydrogenation plants are now rarely used. Check valves should be installed vertically with the flow direction upwards, so that the dead weight of the sealing element supports the closing force of the spring. The lengths of the valves and fittings are coordinated with one another so that they can be interchanged, e.g., a pipe bend for a corner valve or T-piece.

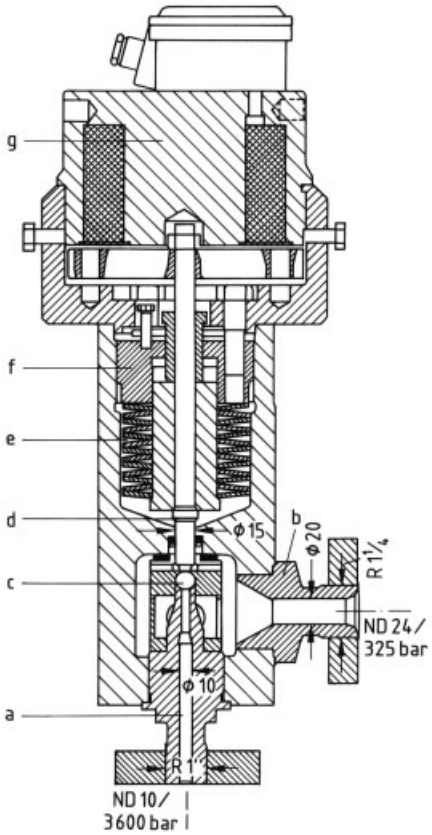

Figure 51. Full-stroke relief valve for 3600 bar
a) Inlet nozzle; b) Outlet nozzle; c) Stainless steel ball; d) Stem; e) Disk springs; f) Adjusting screw for spring load; g) Magnetic lifting device

5. References

1. S. Maier, F. J. Müller: "Reaktionstechnik bei industriellen Hochdruckverfahren," *CIT* **58** (1986) no. 4, 287 – 296.
2. K. H. Schmidt: "Neuentwicklungen in der homogenen Katalyse," *Chem. Ind.* **37** (1985) 762 – 765.
3. R. Eggers: "Energetische Optimierung der Hochdruckextraction," *CIT* **53** (1981) 551 – 554.
4. G. Vetter: "Pumpen und Verdichter für hohe Drücke – Festigkeit, Dichtungen, Verschleiß," *CIT* **58** (1986) 183 – 185.

General References

5. I. Class, A. F. Mayer: "Bauarten von Hochdruck-hohlkörpern in Mehrteil-, insbesondere Mehrlagenkonstruktion," *Chem.-Ing.-Techn.* **24** (1952) 4 – 9.
6. I. Class: "Entwicklung und besondere Vorteile der zylindrischen Mehrlagenhohlkörper für hohe Innendrücke," *Werkst. Korros.* **7** (1956) 8 – 11.
7. J. Schierenbeck: "Wickelverfahren zur Herstellung von Synthese-Hochdruckhohlkörpern," *Brennst.-Chem.* 1959, 375 – 381.
8. W. Witschakowski: "Large Pressure Vessel Manufacture," *Capital and Process Engineering* **1** (1968) 63 – 66.
9. H. C. Rauschenplat, A. F. Fino: "Design, Fabrication, Inspection and Testing of Multiwall Pressure Vessels," *Am. Soc. Mech. Eng. Pressure Vessels Piping Div.* [Publ.] 71-PVP-57.
10. R. Tschirsch: "Der Mehrlagenbehälter, Eigenschaften, Versuche und Stand der Anwendung," *Stahlbau* **4** (1976) 108 – 119.
11. Nooter Technical Service, Plywall-Containing High-Pressures Economically, bulletin no. 102, Nooter Corp., St. Louis, Mo., 1973.
12. J. Ehle, A. Pollanz: "Grundlagen des Hochdruck-Apparatebaues," *Chem.-Ztg.* **95** (1971) no. 1, 28 – 33.
13. E. Karl: "Dichtungen für Hochdruckbehälter," *Chem.-Ing.-Techn.* **43** (1971) 698 – 704.
14. E. Karl: "Strip-Wound Pressure Vessels," *Chem. Eng. Prog.* **68** (1972) no. 11, 56 – 60.
15. E. Karl: "Wickel-Druckbehälter für Hochdruck-Polyäthylen-Anlagen," *CZ-Chemie-Techn.* **2** (1973) no. 1, 13 – 17.
16. J. P. König, H. Spähn: "Korrosion und Korrosionsschutz in der Hochdrucktechnik," *Werkst. Korros.* **24** (1973) 468 – 477.
17. H. H. Buchter: *Apparate und Armaturen der Chemischen Hochdrucktechnik,* Springer Verlag, Berlin 1967.
18. E. Karl: "Seal Design for Pipes and Vessels used in Chemical High-Pressure Plants," ASME-PVP-Conference 1981, vol. 48, 37 – 51.

19. W. R. D. Manning, S. Labrow: *High-Pressure Engineering,* Leonard Hill, London 1971.
20. H. Spähn: "Druckgefäße," *VGB-Kraftwerkstech.* **58** (1978), no. 7, 511 – 528.
21. H. Ford, E. H. Watson, B. Crossland: "Thoughts on a Code of Practice for Forged High-Pressure Vessels of Monobloc Design," *Am. Soc. Mech. Eng. Pressure Vessels Piping Div.* [Publ.], 78-PVP-62.
22. D. E. Witkin, G. J. Mraz: "Design Philosophy of Pressure Vessels for Service above 10 KSI (70 MPa)," *Am. Soc. Mech. Eng. Pressure Vessels Piping Div.* [Publ.], 76-PVP-62.
23. J. Watanabe et al.: "Temper Embrittlement of 2 Cr – 1 Mo Pressure Vessel Steel," *ASME 29th Petroleum Mechanical Engineering Conference,* Dallas, Texas, Sept. 15 – 18, 1974.
24. A. Chaaban: "Static and Fatigue Design of High Pressure Vessels with Blind Ends and Cross-Bores," Ph. D. Thesis, University of Waterloo, Canada, 1985. References for Chapter 3
25. H. Wendt: "Apparate und Maschinen der Hochdrucktechnik, Achema, Berichte," *Chem. Ing. Tech.* **57** (1985) no. 12, 1062 – 1067.
26. G. Vetter, D. Lambrecht, G. Mischorr: "The Fatigue of Thickwalled Components with Softmartensitic and Semiaustenitic Chrome-Nickel-Steels under Pulsating Pressure," *2nd Int. Symp. High-Pressure Chemical Eng.,* Erlangen 1990, Abstrakt Handbook, DECHEMA, Frankfurt 1990, pp. 499 ff.
27. G. Mischorr: "Zur Ermüdung dickwandiger Rohre aus weichmartensitischen und halbaustenitischen Chrom-Nickel-Stählen durch schwellenden Innendruck," Dissertation Universität Erlangen-Nürnberg 1990.
28. G. Vetter: "Ausführungskriterien und Störeinflüsse bei oszillierenden Dosierpumpen," in G. Vetter (ed.): *Jahrbuch Pumpen,* Vulkan-Verlag, Essen 1987, pp. 521 – 529.
29. H. Funke: "Eluentenfördersysteme in der HPLC," *Labor Praxis* 1984, 18 – 29.
30. H. Fritsch: "Leckfreie Mikro-Dosierpumpen," in G. Vetter (ed.): *Leckfreie Pumpen,* Vulkan-Verlag, Essen 1990, pp. 97 – 118.
31. G. Vetter: "Oszillierende Dosierpumpen in leckfreier Ausführung," in G. Vetter (ed.): *Leckfreie Pumpen,* Vulkan-Verlag, Essen 1990.
32. J. Jarosch: "High-Pressure Diaphragm Pumps for High Flow Rates," *2nd Int. Symp. High-Pressures Chemical Eng.,* Erlangen 1990, Abstract Handbook DECHEMA, Frankfurt 1990, pp. 517 ff.
33. G. Vetter: "Reliability and Future Development of High-Pressure Diaphragm Pumps for Process Service," *Proc. of the 5th International Pump Users Symposium,* Houston, Texas 1988, pp. 49 ff.
34. R. Bräuer: "Leckfreie Dosierpumpen für schwierige Anwendungen, *3* [*Drei*] *R Rohre Rohrleitungsbau Rohrleitungstransp.* **27** (1988) no. 7, pp. 494 – 497.
35. J. P. Körner, P. S. I. Crofton: "Entwicklung und Konstruktion eines Drucküberset zers für den Dauerbetrieb bei Druck über 3000 bar," *Chem. Tech. (Heidelberg)* **8** (1980) no. 11, 561 – 565.
36. H. Fritsch: "Prozeßmembranpumpen," in G. Vetter (ed.): *Leckfreie Pumpen,* Vulkan-Verlag, Essen 1990, pp. 118 – 138.
37. W. Dettinger: "Eigenschaften von Kohlemaischepumpen für Hydrieranlagen," *Chem. Ing. Tech.* **54** (1982) no. 5, 500 – 501.
38. G. H. Holthuis, P. W. H. Simons: "The Economics of Positive Displacement Slurry Pumps," *5th Int. Techn. Conf. on Slurry Transportation,* Las Vegas 1981.
39. U. Schuster: "Hydraulically Driven Twocylinder Pistonpumps for Solid Transport with Viscous Materials in Pipelines," *1st Int. Congr. Fluid Handling Systems, Abstract Handbook,* Essen 1990, p. 507 ff.
40. W. T. Rice, S. Lavin: *Thermorex-Zahnradpumpen – Anwendung in Verbindung mit Extrudern,* Druckschrift Maag-Zahnräder, Zürich 1981.
41. W. Branscheid: "Die Zahnradpumpe als volumetrisches Dosieraggregat für Kunststoffe," *VDI-Fachtagung: Speichern, Fördern und Dosieren von Kunststoffen,* VDI, Düsseldorf 1976.
42. J. M. McKelly, V. Maire, F. Haupt: "How Gear Pumps and Screw Pumps Perform in Polymer Processing Applications," in: *Fluid Movers,* McGraw-Hill, New York 1979, pp. 290 – 298.
43. G. Balder, H. Langhorst: "Zahnrad-Schmelze-Pumpen in der Extrusion," *Kunststoffe* 1988, no. 1, 27 – 33.
44. T. Bartilla: "Zahnradpumpen im Extrusionsprozeß," *Der Extruder im Extrusionsprozeß – Grundlagen für Qualität und Wirtschaftlichkeit,* VDI-Verlag, Düsseldorf 1989, pp. 201 – 240.
45. G. Feldle: "Neue Baureihe Hochdruck-Kreiselpumpen – Einfluß der Gesamtkostenbetrachtung auf die

Entwicklung," *KSB Technische Berichte* (1986) no. 20, 50 – 55.
46. H. B. Matthias: "Speisepumpen für Dampfkraftwerke und Nuklear-Kraftwerke bis 1300 MW Blockleistung," *Pumpen und Pumpenanlagen,* Lexika-Verlag Grafenau 1979.
47. G. Füssle: "Betriebserfahrungen mit Injektionspumpen," *Techn. Rundsch. Sulzer* (1985) no. 2, 12 – 17.
48. P. E. Huber: "Seawater Injection Pumps for Secondary Oil Recovery," World Pumps 1983, pp. 125 – 127.
49. V. S. Lobanoff, R. R. Ross: *Centrifugal Pumps – Design and Application,* Gulf Publ. Co., Houston 1985.
50. R. Krämer, R. Neumaier: *Centrifugal Pumps and Rotary Positive-Displacement Pumps of Hermetic Design,* Hermetic-Pumpen GmbH, Gundelfingen 1988.
51. R. Krämer, R. Neumaier: "Hermetic Drive Systems in High-Pressure Circuits," *2nd Int. Symp. High-Pressure Chemical Eng.,* Erlangen 1990, Abstract Handbook pp. 507 – 516.
52. M. Knorr: "Stopfbuchslose Kreisel- und Verdrängerpumpen mit permanentmagnetischer Synchronkupplung," in G. Vetter (ed.): *Leckfreie Pumpen,* Vulkan-Verlag, Essen 1990.
53. J. P. Körner: "Developmental Tendencies in the Design and Constructions of High-Pressure Vessels, especially of Large Scale Stirrer Vessels with Permanent-magnet Rotary Stirrer Drives," *2nd Int. Symp. High-Pressure Chem. Eng.,* Erlangen 1990, Abstract Handbook, pp. 489 – 496.
54. K. H. Victor: "Kontakt und verschleißfrei arbeitende Elasto-Hydrodynamik-Gleitringdichtungen für den Hochdruckeinsatz," *Pumpentagung,* Karlsruhe 1988.
55. H. R. Kläy: "Gasverdichter (Übersicht)," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990, pp. 1 – 5.
56. H. Gernandt: "Kolbenkompression für die Verfahrenstechnik," in G. Vetter (ed.): *Verdichter Handbuch,* 1st. ed., Vulkan-Verlag, Essen 1990, pp. 200 – 206.
57. H. Gernandt: "Some Problem Solutions with High-Pressure Reciprocating Compresors," *2nd Int. Symp. High-Pressure Chem. Eng.,* Erlangen 1990, Abstract Handbook, pp. 523 – 529.
58. H. R. Kläy: "Labyrinth-Kolben-Kompressoren," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990, pp. 207 – 212.
59. Industrial Gases Committee (IGC): Kolben-Verdichter für Sauerstoff-Betrieb, IGC-Dok. 10/81/D.
60. C. Matile: "Industrie-Höchstdruck-Verdichter," *Techn. Rundsch. Sulzer* (1971) no. 1, 1 – 8.
61. E. Giacomelli, P. Pinzanti, S. Corsi: Autofrettage in Secondary Ethylene Compressors: Some Practical Aspects, *Quaderni Pignone,* **44** (1988) 19 – 26.
62. A. Traversari, P. Beni: "Approaches to Design of a Safe Secondary Compressor for High-Pressure Polyethylene Plants," *Saf. High-Pressure Polyethylene Plants* AICHE, 1974.
63. I. Meyn: "Membrankompressoren für hohe Drücke," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990, pp. 107 – 111.
64. W. Rieß: "Einsatzgebiete und konstruktive Ausführung von Turbokompressoren – Stand der Technik und Zukunftsperspektiven," in VDMA (ed.): *Pumpen, Vakuumpumpen, Kompressoren* '89, Harnisch Verlag, Nürnberg 1989, pp. 44 – 50.
65. E. Rothstein: "Neue Lösungen in der Prozeß-industrie – der Einsatz von sechsstufigen Getriebe-Turboverdichtern," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990, pp. 213 – 222.
66. I. Ispas, W. Bosen: "Mehrstufige Getriebe-Turboverdichter," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990, pp. 128 – 138.
67. J. Hütgens, F. J. Mangelmann, H. O. Jeske: "Gastransport mit einstufigen axial ansaugenden Pipelinekompressoren," in G. Vetter (ed.): *Verdichter Handbuch,* 1st ed., Vulkan-Verlag, Essen 1990.
68. *Turboverdichter für sehr hohes Druckniveau,* Mannesmann-Demag Druckschrift MA 25.69 dt/11.81, 1981.
69. R. Reimert: "Schleusen für Druckreaktoren – Konzepte und Ausführungen," *Chem. Ing. Techn.* **53** (1981) 335 – 344.
70. J. P. Körner: "New Development in the Design and Construction of Industrial-size SCGE Plants," *Proced. of the Int. Symp. on Supercritical Fluids,* Nice 1988, p. 633 ff.
71. A. Chaaban: *Static and Fatigue Design of High-Pressure Vessels and Blind-Ends and Cross-bores,* Thesis, University Waterloo 1985.

72. J. L. M. Morrison, B. Crossland, J. S. C. Parry: "Strength of Thick Cylinders Subjected to Repeated Internal Pressure," *Proc. Inst. Mech. Eng.* **14** (1960) 95 ff.
73. J. S. C. Parry, R. W. E. Shannon: "Fatigue and Fracture of Thickwalled Pressure Vessels," *Proc. Int. Conf. High-Pressure Engineering,* Brighton 1975, I. Mech. E. 1977, p. 267 ff.
74. G. H. Haslam: "The Fatigue Limit of Cylinders Subjected to Repeated Internal Pressure," *High-Pressures – High Temperatures* **1** (1969) 705 ff.
75. P. S. J. Crofton: "Optimization of Duplex Cylinders for Fatigue Resistance," *ASME Pressure Vessels and Piping Conf.,* Pittsburgh USA 1988.
76. P. S. J. Crofton: *The Role of Inclusions in the Initiation and Propagation of Fatigue Cracks in Low Alloy Steels,* Doctoral Thesis, Imp. Coll. of Science and Techn., London 1980.
77. E. Karl: "Fatigue Resistance of Thick-Walled Pipes under Pulsating Internal Pressure," *Int. Symp. High-Pressure Chem. Eng.,* Erlangen 1984, Abstract Handbook, p. 41.
78. M. Nishihara, Y. Yamaguchi, S. Hattori: "Fatigue Design of Thick-Walled Cylinders for Very High-Pressure," *R & D Res. Div. (Kobe Steel Ltd.)* **26** (1978) no. 2.
79. G. Vetter, G. Mischorr: "Fatigue of Thick-walled Cylinders from High-alloyed Corrosion Resistant CrNi-Steels under Pulsating Pressure," *Proc. Int. Conf. on Fract. and Fract. Mech.,* Shanghai 1987, p. 721 ff.
80. B. N. Cole, G. Graggs, I. Ficinec: "Strength of Cylinders Containing Radial or Offset Cross-Bores," *J. Mech. Eng. Sci.* **18** (1976) no. 6, 279 – 285.
81. U. Gramberg: "Beitrag zur Problematik der Auslegung schwingend beanspruchter Bauteile," *Chem. Ing. Tech.* **56** (1984) 661 – 666.
82. G. Vetter, D. Lambrecht, G. Mischorr: "Fatigue of Thick-Walled Pipes from Soft Martensitic and Semiaustenitic Chrome-Nickel Steels under Pulsating Internal Pressure," *Chem. Eng. Technol.* **15** (1992) 1 – 13.
83. E. Mayer: *Axiale Gleitringdichtungen,* VDI-Verlag, Düsseldorf 1982.
84. Neuman & Esser, *Process Gas Compressors,* Druckschrift 1989.
85. K. Scheuber: "Dynamische Druckverteilung in der Zylinderdichtung von Höchstdruckkompressoren," *Ölhydraulik und Pneumatik* **25** (1981) 583 – 586.
86. P. de Haas: BORSIG-Druckschrift, *Besondere Merkmale von Borsig-Kolbenkompressoren,* 1989.
87. G. Vetter, U. Störk: "Zum Verschleiß selbsttätiger Pumpenventile oszillierender Verdrängerpumpen durch abrasive Suspensionen," *Konstruktion* **41** (1989) 67 ff.
88. G. Vetter, F. Schweinfurter: "Vermeidung störender und gefährlicher Druckschwingungen durch Hochdruckpumpen," *VDI-Ber.* **748** (1989) 21 ff.
89. H. Fritsch: "Hydraulische Schwingungen und ihre Dämpfung in Rohrleitungen oszillierender Verdrängerpumpen," *Chem.-Tech. (Heidelberg)* **13** (1984) no. 10, 61 – 74.
90. W. Nimitz: "Reciprocating Compressor Performance," *Pipeline and Gas Journal* **212** (1985) 12 – 20.
91. H. J. Brokowski, H. Kayss, T. Wierschem: "Sichere Gestaltung von Arbeitsmaschinen und deren Umfeld," *Chem. Tech.* **15** (1986) 23 – 26.
92. H. Lankenau: "Entstehung von Schwingungen in Kolbenkompressorenanlagen unter Berücksichtigung der Wirkung von Pulsationsdämpfern," in G. Vetter (ed.): *Verdichter Handbuch* 1st ed., Vulkan-Verlag, Essen 1990, pp. 165 – 173.

High-Tech Ceramics → Ceramics, Advanced Structural Products
High-Temperature Composites → Composite Materials
High-Temperature Composites → Metal-Matrix Composites
High-Temperature Fibers → High-Performance Fibers

Analytical Chemistry: Purpose and Procedures

GÜNTHER TÖLG, Institut für Spektrochemie und Angewandte Spektroskopie, Dortmund, Federal Republic of Germany

HELMUT GÜNZLER, Weinheim, Federal Republic of Germany

ALEX WILLIAMS, Mytchett, Camberley, UK

1. The Evolution of Analytical Chemistry . . . 1
2. The Functional Organization of Analytical Chemistry . . . 4
3. Analysis Today . . . 5
4. Computers . . . 7
5. Analytical Tasks and Structures . . . 8
5.1. Formulating the Analytical Problem 8
5.2. Research and Application . . . 8
5.3. An Organogram . . . 9
5.4. Physical Organization of the Analytical Laboratory . . . 11
5.5. The Target of Analysis . . . 11
6. Definitions and Important Concepts 13
6.1. Sensitivity, Limit of Detection, and Detection Power . . . 13
6.2. Reliability – Measurement Uncertainty . . . 14
6.3. Elemental Analysis . . . 15
6.4. Elementary Analysis . . . 16
6.5. Microanalysis and Micro Procedures 16
6.6. Stereochemical and Topochemical Analysis . . . 16
6.7. Microdistribution Analysis . . . 17
6.8. Surface Analysis . . . 17
6.9. Trace Analysis . . . 17
6.10. Trace Elements . . . 18
6.11. Multistep Procedures . . . 19
6.12. Hyphenated Methods . . . 19
6.13. Radioanalytical Methods and Activation Analysis . . . 19
6.14. Species Analysis (Speciation) . . . 20
6.15. Chemometrics . . . 20
6.16. DNA Analysis . . . 21
7. References . . . 21

1. The Evolution of Analytical Chemistry

"Analytical chemistry" (more simply: *analysis*) is understood as encompassing any examination of chemical material with the goal of eliciting information regarding its constituents: their *character* (form, quality, or pattern of chemical bonding), *quantity* (concentration, content), *distribution* (homogeneity, but also distribution with respect to internal and external boundary surfaces), and *structure* (spatial arrangement of atoms or molecules). This goal is pursued using an appropriate combination of chemical, physical, and biological methods [1–5]. From a strategic standpoint the challenge is to solve the analytical problem in question as completely and reliably as possible with the available methods, and then to interpret the results correctly. Sometimes it becomes apparent that none of the methods at hand are in fact suitable, in which case it is the methods themselves that must be improved, perhaps the most important rationale for intensive basic research directed toward the increased effectiveness of problem-oriented analysis in the future.

More comprehensive contemporary definitions of analytical chemistry have been proposed [6], [7], underscoring above all the complexity of the discipline – which the authors of this introduction were also forced to confront.

Consistent with its close historical ties to chemical synthesis, modern analysis is still firmly embedded within the broader framework of chemistry in general. This is inevitably the case, because systematic analysis depends absolutely upon a solid, factual knowledge of matter. This point is as valid now as it was in 1862 when C. R. FRESENIUS stated in his classic *Introduction to Qualitative Chemical Analysis* [8]: "Chemical analysis is based directly on general chemistry, and it cannot be practiced without a knowledge thereof. At the same time it must be regarded as one of the fundamental pillars upon which the entire scientific edifice rests; for analysis is of almost equal importance with respect

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_001

to all the branches of chemistry, the theoretical as well as the applied, and its usefulness to doctors, pharmacists, mineralogists, enlightened farmers, technologists, and others requires no discussion."

The *tools* of modern analysis are nevertheless based largely on physical principles. Mathematical techniques related to information theory, systems theory, and chemometrics are also making increasingly important inroads. It would in fact no longer be presumptuous to go so far as to describe "analytical science" as an independent discipline in its own right.

The pathway leading to the present exalted place of analysis within the hierarchy of chemistry specifically and the natural sciences generally has not always been a straight one, however. Indeed, from earliest times until well into the eighteenth century the very concept of "analysis" was purely implicit, representing only one aspect of the work of the alchemists and various practitioners of the healing arts (iatrochemists). Some more tangible objective always served as the driving force in an investigation, and "to analyze" was almost synonymous with the broader aim: a quest for precious metals, a desire to establish the content of something in a particular matrix, or a demonstration of pharmacological activity. Only after the time of Lavoisier and with the emergence of a separate chemical science – a science largely divorced from external goals – is one able to discern what would today be regarded as typical "analytical" activity. The term "analysis" appears explicitly for the first time around the turn of the nineteenth century in the title of the book, *Handbuch zur chemischen Analyse der Mineralkörper* ("Handbook for the Chemical Analysis of Minerals") by W. A. Lampadius. Further information regarding the history of analysis is available from the monograph by Szabadvary [9].

Many of the greatest discoveries in chemistry could fairly be described as classic examples of successful analyses, including the discovery of oxygen, the halogens, and several other elements. Well into the nineteenth century, discovering a new chemical element was regarded as the highest and most prestigious achievement possible for an academic chemist, as documented, for example, by desperate attempts to gain further insight into the "rare earths", or to detect the elusive (but accurately predicted) homologues of lanthanum and cerium. Mosander in fact devoted his entire life to the latter search.

Fresenius once again deserves credit for noting, toward the middle of the nineteenth century, that new analytical techniques invariably lead to fresh sets of discoveries. Whereas the element germanium was found on the basis of "classical" methods (ClemensWinkler, 1886), Fresenius' observation clearly applies to the discovery of the alkali metals rubidium and cesium (by Robert W. Bunsen after he and G. R. Kirchhoff first developed emission spectroscopy in 1861). Other relevant examples include the discoveries of radium and polonium (by Marie Curie), hafnium (Hevesy and Coster, 1922), and rhenium (I. Tacke and W. Noddack, 1925), all with the aid of newly introduced X-ray spectrometric techniques. This is also an appropriate point to mention the discovery of nuclear fission by Otto Hahn and Fritz Strassmann (1938), another accomplishment with strongly analytical characteristics [9].

Robert Bunsen is rightfully acknowledged as the harbinger of modern analysis, but much of the discipline's distinctive scientific character was provided by Wilhelm Ostwald [10] building on the activities of Jacobus H. van't Hoff and Walther Nernst.

Analytical chemistry in these early decades was often accorded the secondary status of a faithful servant, but even the few examples cited here demonstrate quite convincingly that it also pursued its own unique set of principles – and for its own sake, with a strictly scientific orientation. The principles themselves were shaped by Berzelius and Wöhler; experiment rather than theoretical speculation was the starting point and source of inspiration in this era characterized largely by chemical reactions. Readers of the present essay should in fact take the time to examine the third edition of *Ullmann's* [11] and discover there what the expression "analytical procedure" actually meant even as late as the end of World War II. There can be no mistaking the fact that "purely chemical" methods were still dominant, and that the synthetic process constituted the model, particularly in the field of organic analysis.

Analytical chemistry has been responsible for many important contributions to our basic understanding of matter (e.g., the existence of the various elements, gas theory, stoichiometry,

atomic theory, the law of mass action, nuclear fission, etc.), but the growth and development of a separate chemical industry ushered in a phase during which the scientific aspects of analysis suffered serious decline. The demand for analytical services shifted markedly in the direction of routine quality control, particularly with respect to synthetic organic products; indeed, significant resources were invested in the effort to dismember, resolve, and decompose synthetic substances into their simpler constituents (e.g., the chemical elements) – in strict conformity with the original meaning of the word "analysis" ($\alpha\nu\alpha\lambda\upsilon\sigma\iota\sigma$, *resolution*). For many years organic elementary analysis was virtually the only analytical approach available for characterizing synthetic organic reaction products. The denigration suffered by analysis at that time relative to synthesis (and production) continues to exert a negative influence even today on the university training of analytical chemists.

Elemental analysis in certain other quarters enjoyed a climate much more congenial to further development, especially in the metalworking industry and geochemistry. The indispensable contributions of analysis were recognized here much earlier, particularly with respect to optimizing product characteristics (e.g., of steels and other alloys), and to providing detailed insight into the composition of the Earth's crust to facilitate the extraction of valuable raw materials. Geochemistry and the steel industry were particularly receptive to BUNSEN's new methods of spectral analysis, for example, which in turn provided a powerful stimulus for the development of other modern instrumental techniques. These techniques encouraged the exploitation of new and innovative technologies, first in the fields of semiconductors and ultrapure metals, then optical fibers and superconductors, and, most recently, in high-temperature and functional ceramics. Extraordinarily stringent demands were imposed upon the various analytical methods with respect to detection limits, extending to the outermost limits of the possible, especially in the attempt to characterize impurities responsible for altering the properties of particular materials. At the same time, the information acquired was expected to reflect the highest possible standards of reliability – and to be available at an affordable price. These three fundamental quality criteria are in fact closely interrelated, as indicated in Figure 1.

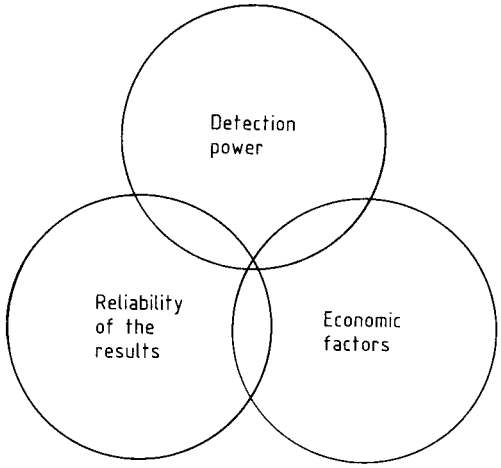

Figure 1. Qualitative criteria for use in evaluating analytical procedures

The increasing effectiveness of analytical techniques in general led ultimately to progress in the area of organic materials as well, especially with the rapid development of chromatographic and molecular spectroscopic methods. At the same time it also became necessary to acknowledge that technological advances inevitably bring with them new safety and health risks. For this reason analysis today plays an essential role not only in supporting technological progress but also in detection and minimization of the associated risks.

Just as FRESENIUS predicted, analysis has advanced rapidly toward becoming a science in its own right, with interdisciplinary appeal and subject to intense interest extending far beyond the bounds of chemistry itself: to the geological and materials sciences, the biosciences, medicine, environmental research, criminology – even research into the history of civilization, to mention only a few of the most important areas of application. The chemical industry today is the source of only a relatively small fraction of the samples subject to analysis. Rocks, soils, water, air, and biological matrices, not to mention mankind itself and a wide array of consumer goods, together with raw materials and sources of energy constitute the broad spectrum of analytical samples in the modern era (Fig. 2).

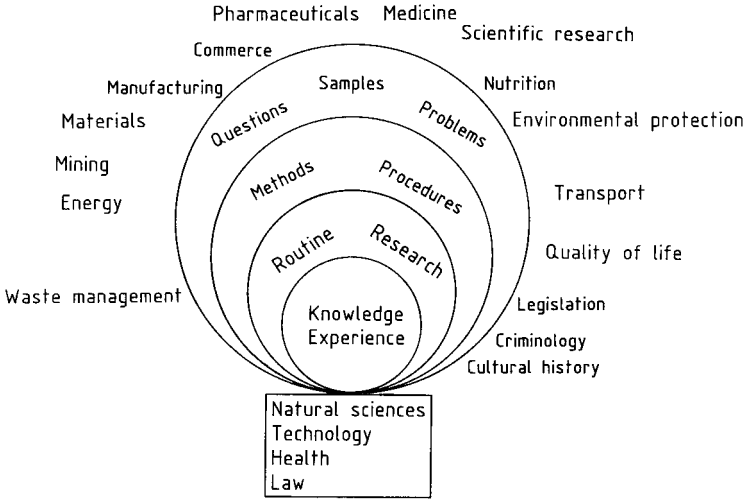

Figure 2. Overall task structure associated with analytical chemistry

Given this diversity of appeal the question has frequently been raised as to whether analysis really is an independent discipline, or if it should not instead be regarded simply as a service activity. The question is of course unrealistic, because analysis by its very nature is clearly both. Equally clear is the crucial importance of analysis to modern society. While the service function is undoubtedly more widely appreciated than other activities characterized by a "strictly scientific" focus, the latter also have an indispensable part to play in future progress.

The diversity characterizing the beneficiaries of analysis has actually remained fairly constant in recent decades, though immediate priorities have undergone a steady shift, particularly during the last 20 years with respect to ecology. Such a "paradigm shift" (THOMAS KUHN), marked by profound changes over time in both motivation and methodology, can occasionally assume revolutionary proportions. It remains an open question whether external change induces analysts to adapt and further develop their methodologies, or if the methodology itself provides the driving force. Here as elsewhere, however, there can be little doubt that "necessity is the mother of invention", capable of mobilizing forces and resources to an extent unimaginable in the absence of pressing problems.

Change also provides an incentive for deeper reflection: should we perhaps reformulate our understanding of the overall significance of analysis, lift it out of its customary chemico-physical framework and broaden its scope to include, for example, KANT'S "analytical judgments", or even psychoanalysis? Some would undoubtedly dismiss the questions as pointless or exaggerated, but from the perspective of the theory of learning they nevertheless provoke a considerable amount of interest and fascination [12], [13].

2. The Functional Organization of Analytical Chemistry

Attempting to summarize analytical chemistry in a single comprehensive schematic diagram is a major challenge, one that can only be addressed in an approximate way, and only after considerable simplification (Fig. 2). The fundamentals supporting the analysis must ultimately be the individual analyst's own store of knowledge, including the basic principles and laws of science and mathematics, together with the scope – and limitations – of existing analytical methods and procedures. Indispensable prerequisites to the successful resolution of an analytical problem include experience, a certain amount of intuition, and thorough acquaintance with a wide variety of modern analytical techniques. Familiarity with the extensive technical literature is also important (including the

sources ones cited at the end of this article), an area in which modern systems of documentation can be of considerable assistance. For example, an astonishing level of perfection can almost be taken for granted with respect to computer-based systems for locating spectra. Another essential component of the analyst's information base is knowledge regarding the source of each analytical sample – whether it comes from industry, the environment, or from medicine. After all, only the analyst is in a position to provide an overview of the analytical data themselves when the time comes for critical interpretation of the experimental results.

Immediately adjacent to "knowledge" in the functional diagram characterizing analytical chemistry (Fig. 2) is a region occupied by two parallel lines of endeavor: routine analysis on one hand, and research and development on the other, with the latter directed toward new methods and procedures. Both are subject to initiatives and incentives from outside, including other branches of science, medicine, regulatory agencies, commerce, and industry, all of which encourage and foster innovative developments within analysis itself.

Figure 2 also underscores the fact that an analyst's primary activities are of a problem-oriented nature, determined largely by the needs of others. The problems themselves, represented here by the outermost circle, might originate almost anywhere within the material world. Analysis can even play a significant role in the very definition of a scientific investigation. Consider the case of archaeology, for example, a considerable part of which is now "archaeometry", simply a specialized type of analysis.

With respect to the development of new products – such as materials, semiconductors, pharmaceuticals, crop protection agents, or surfactants – analysis plays a companion role at every stage in the progression from research laboratory to marketplace. Studies related to physiological and ecological behavior demand comprehensive analytical efforts as well as intimate knowledge of the materials in question.

3. Analysis Today

Figure 3 provides a representative sample of methods to be found in the arsenal of the modern analyst. The figure also highlights the rapid pace of developments in analytical chemistry during the twentieth century [14]. Continued success in meeting present and future analytical challenges involves more than simply the tools, however, most of which have already been perfected to the point of commercialization. Appropriate strategies are required as well, just as a hammer, a chisel, and a block of marble will not suffice to produce a sculpture. Analytical strategies are at least as important as the methods, and the strategies must themselves be devised by qualified analysts, because every complex analytical problem demands its own unique strategic approach.

It is this context that establishes the urgent need for reactivating as quickly as possible the long-neglected training of qualified analysts. New analytical curricula must also be devised in which special emphasis is placed on the close symbiotic relationship in modern analysis between chemistry and physics.

Figure 4 depicts in a generalized way the multileveled complex of pathways constituting a typical analytical process and linking a particular problem with its solution. From the diagram it becomes immediately apparent that the "analytical measurement", which is the focal point of most modern physical methods, in fact represents only a very small part of the whole, despite the fact that the treatise to which this essay serves as a preface focuses almost exclusively on the principles of instrumental methods and their limitations.

Physical methods clearly occupy the spotlight at the moment, but *chemical* methods of analysis are just as indispensable today as in the past. Especially when combined with physical methods, chemical techniques frequently represent the only means to achieving a desired end. This is generally the case in extreme trace analysis [15], for example, where attaining maximum sensitivity and reliability usually requires that the element or compound of interest first be isolated from an accompanying matrix and then concentrated within the smallest possible target area or solution volume prior to the physical excitation that leads ultimately to an analytical signal. Combination approaches involving both chemical and physical methods are today commonly referred to as *multistep procedures* (see Section 6.11), where some chemical step (e.g., digestion, or enrichment) often precedes an in-

pre-1850	1850-1900	1900-1925	1925-1950	1950-1975	1975-1992 (selection)
Titrimetry	Electrogravimetry	Raman spectroscopy	γ-Spectroscopy	Neutron spectroscopy	Atomic force microscopy
Gravimetry	Coulometry	X-Ray emission spectroscopy	Electron spin resonance	Laser spectroscopy	Scanning tunnel microscopy
Gas-volumetry	Optical emission spectroscopy (OES)	X-Ray diffraction	NMR	Electron tunneling spectroscopy	Atom probe
		Mass spectroscopy	Gas chromatography	HPLC	Analytical electron microscopy
		IR spectroscopy	Paper chromatography	Candoluminescence	FT spectroscopy
		Isotope-dilution analysis	Distribution chromatography	MECA spectroscopy	ICR mass spectroscopy
		Radiochemical analysis	High-frequency titration	RHEED	TXRFA
		Adsorption chromatography	Electron diffraction	LEED	IR microscopy
		Conductometry	Activation analysis	Desorptiometry	Scanning auger electron spectroscopy
			Electrophoresis	SIMS	Dynamic SIMS
			Polarography	Mößbauer spectroscopy	CARS
			RBS	ESMA	Photoacoustic spectroscopy
				NQR	SERS
				EELS	FANES
				Auger electron spectroscopy	MONES
				Photoelectron spectroscopy	Appearance-potential spectroscopy
				Ionography	IBLE
				FIM	FAB-MS
				FEM	GD-AES
				EDX	GD-MS
				Neutron diffraction	Ion chromatography
				β-Backscattering	Ion neutralization spectroscopy
				Atomic absorption spectroscopy	EXAFS
				Reflection IRS	SNMS
				Microwave spectroscopy	UPS, XPS
				PIXE	Capillary gas chromatography
					Gel chromatography
					Supercritical-fluid chromatography
					HPTLC
					Scanning DC
					Plasma spectroscopy
					ICP-OES, MIP-OES
					ETA-AAS
					2D-NMR
					Multinuclear NMR
					Time-resolution spectroscopy
					GC-MS
					MS-MS-(MS)
					AAS-GC
					ICP-MS
					ICP-GC
					Laser ICP-MS
					Thermospray MS, Electrospray MS
					HPLC-MS
					C-13 NMR
					Charged-particle activation analysis
					Raman microprobe
					RP chromatography
					Solid phase AAS
					Headspace GC
					PIGE
					Static TOFSIMS
					LAMMA

Figure 3. Chronological summary of the arsenal of experimental methods available to analytical chemists; based on [14]

strumental measurement, or an analysis is facilitated by preliminary chromatographic separation. Chromatographic separation in turn sometimes requires some type of prior chemical transformation [16], as in the gas-chromatographic separation of organic acids, which is usually preceded by esterification.

The terms "preanalysis" and "postanalysis" have been coined for characterizing steps that precede or follow a "true" analytical operation. Unfortunately, classification in this way tends to denigrate the importance of an operation like sampling or the evaluation of a set of final results, suggesting that these are secondary and relatively peripheral activities – reason enough for exercising considerable caution in use of the terms.

There can be no justification whatsoever for dismissing the importance of chemical reactions in analysis, as "superprogressive" instrumental analysts occasionally tend to do, treating chemical methods as relics of an outmoded past.

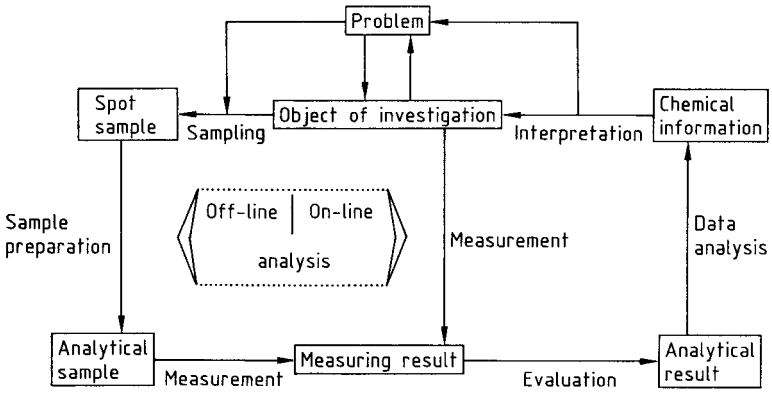

Figure 4. Schematic diagram of the analytical process; based on [14]

Chemical reactions still have a crucial part to play in many operations: sometimes as useful adjuncts, but often enough at the very heart of the determination. It is worth recalling in this context that gravimetry – together with the volumetric methods to which it gave birth – remains virtually the only viable approach to direct and reliable absolute determination (i.e., to *calibration-free analysis*). Such analyses rely on "stoichiometric factors", which were painstakingly compiled over the course of decades in conjunction with the equally arduous and prolonged quest for an exact set of atomic masses.

Most physical methods, especially those associated with spectroscopy, lead only to *relative* information acquired through a comparison of two signals. This in turn presupposes a procedure involving a *calibration standard*, or reference sample of known composition. The only exceptions to this generalization – at least theoretically – are instrumental activation analysis (which involves the counting of activated atomic nuclei), isotope dilution, and coulometry (assuming the strict validity of Faraday's Law).

In view of quality assurance, the named methods, jointly with gravimetry, volumetric analysis, and thermoanalysis, were recently designated as *primary methods of measurement* [17–20]. They play an important role in achieving traceable results in chemical measurements.

Some may feel that the foregoing observations direct excessive attention to the virtues of classical analytical chemistry. If so, the justification is a continuing need to emphasize the fact that optimal results are achieved when there is a close coupling between chemical and physical methods, and this despite antagonisms that persist between champions devoted to one approach or the other. Even today, classical principles – appropriately adapted – often constitute the most reliable guide.

4. Computers

A few remarks are necessary at this point on the subject of *electronic data processing* and the vital supportive role computers now play in analysis.

Developments in this area began with the central mainframe computer, to which a wide variety of isolated analytical devices could be connected. The trend has shifted strongly toward preliminary data processing via a minicomputer located directly at the site of data collection, followed in some cases by network transfer of the resulting information to a central computing facility. Often, however, the central computer is dispensed with entirely, with all data evaluation occurring on the spot. The powerful impact of electronic data processing on modern analysis dictates that it be addressed elsewhere in greater detail [→ Chemometrics ; → Laboratory Information Management Systems (LIMS)].

The benefit of computers in modern analysis has been clearly established for some time. Computers now provide routine management and control support in a wide variety of analytical operations and procedures, and they are an almost indispensable element in data interpretation, processing, and documentation. Indeed,

the lofty goals of "good laboratory practice" (GLP) would probably be beyond reach were it not for the assistance of computers. Computers also have a key role in such wide-ranging activities as automated sample introduction and the control of calibration steps (*robotics*). Process-independent tasks closely related to the ongoing work of a laboratory have long been delegated to computers, including the storage, retrieval, and management of data.

Nevertheless, the claim that we have entered an age of "computer-based analytical chemistry" (COBAC) is inappropriate and overly optimistic; "computer-aided" analysis would be a more satisfactory description, and one more consistent with terminology adopted in other disciplines. "Artificial intelligence", so- called expert systems [21], neural networks, and genetic algorithms will undoubtedly be increasingly important in the analytical chemistry of the future, but in most cases probably in the context of relatively complex routine investigations supported by extensive previous experience. It is unlikely that such methods will prove optimal even in the long term with respect to analytical research in uncharted waters, especially if results are required near the limit of detectability.

5. Analytical Tasks and Structures

5.1. Formulating the Analytical Problem

Generally speaking, problem-oriented analytical tasks can best be defined with reference to criteria most easily expressed as questions:

1) How has the problem at hand already been stated? Is the problem as stated truly relevant? If so, what is the maximum expenditure that can be justified for its solution, considering both material and economic resources? (Note that not every problem warrants the pursuit of an optimal analytical solution!)
2) What type and size of sample is available? What content range is predicted with respect to the analyte, and what mass of sample would be required to produce an answer?
3) What analytical strategy (including choice of a particular method) is most appropriate within the context set by considerations (1) and (2) ?
4) Will critical assessment of the analytical results be possible, with evaluation of an uncertainty budget aiming to determine an *expanded uncertainty* of the analytical result [24], [25]? (see Section 6.2)

Ensuring the correctness of a set of results is extremely important, because nothing is more wasteful than acquiring a wrong answer, especially when account is taken of the subsequent interpretation and application of analytical data with respect to matters of safety, health, and the environment. The ultimate validity of an analytical result can be placed in serious jeopardy as early as the sampling stage, since inappropriate sampling can be a very significant source of error.

Such mathematical tools as statistical tests and error analysis are prerequisite to the practical application of an analytical result. In any situation involving verification of compliance with conventions, agreements, regulations, or laws, analysis is expected to provide the meaningful and objective criteria required for assessing the material facts. This means that observed analytical values must be supplemented with quality criteria applicable to the analytical procedure itself, such as the limit of detection, limit of determination, standard deviation, and measurement uncertainty.

5.2. Research and Application

Two major branches of analytical chemistry can be distinguished by the types of challenges they address. The first is the problem-oriented service sector, or *routine analysis*. Here one is usually in a position to rely on existing and proven methods and procedures, though some adaptation may be required to accommodate a method to the particular task at hand.

The second area, *basic analytical research*, is the key to resolving an increasingly complex set of problems today and in the future – problems not subject to attack with tools that are currently available, or amenable only to unsatisfactory solutions (with appropriate regard for economic factors). This underscores the high degree of innovative scientific character associated with analysis as a discipline, innovation that often approaches revolutionary proportions. It is

perfectly possible for epochal developments to emerge from basic principles that are themselves already well established. A striking example is provided by the path leading from organic elementary analysis as first introduced by JUSTUS LIEBIG, starting with rather large samples, via the work of F. EMICH and F. PREGL, and culminating in today's highly perfected micro techniques, a path that runs parallel to the development of the analytical balance.

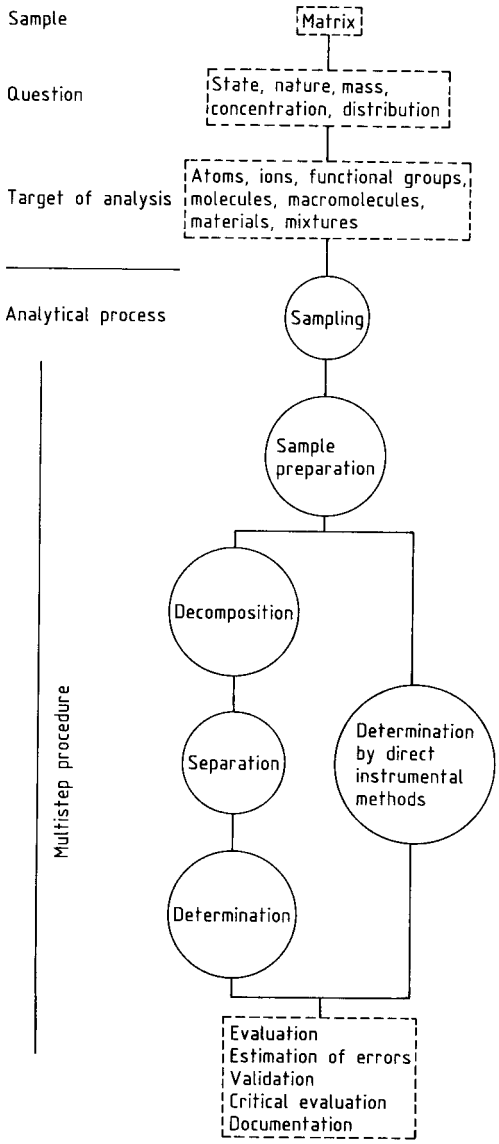

Figure 5. Strategic organization of the analytical process

It is also interesting to consider in this context the source of some of today's most important innovations, which increasingly result from a close symbiotic relationship between university research centers on one hand, and commercial instrument manufacturers on the other (where the latter often have access to extensive in-house research facilities of their own, and may be in a position to introduce important independent initiatives). The reason for the collaborative trend is obvious: continued progress has been accompanied by a disproportionate increase in costs, and the resulting burden can no longer be borne by universities alone. Collaboration between industry and higher education is certainly to be welcomed, but not to the point that technical shortcomings still evident at the conclusion of a joint commercialization venture are suppressed or trivialized in the interest of profit, as has unfortunately occurred on more than one occasion.

5.3. An Organogram

The two complementary branches of analytical chemistry rely on a common foundation of structure and content, illustrated in the "organogram" of Figure 5.

Starting with an analytical sample (the matrix), and proceeding via the formulation of a specific question regarding the state, nature, mass, concentration, or distribution of that sample, as well as a definition (or at least partial definition) of the true target of the analysis (atoms, ions, molecules, etc.), two different paths might in principle be followed in pursuit of the desired objective. Both commence with the extremely critical steps of sampling (→ Sampling) and sample preparation (→ Sample Preparation for Trace Analysis), which must again be recognized as potential sources of significant error. Under certain conditions it may then be possible to embark immediately on qualitative and/or quantitative analysis of the relevant target(s) through direct application of a physical method in the form of an "instrumental" analysis (e.g., a spectroscopic determination following excitation of the sample with photons, electrons, other charged particles, or neutrons). Such instrumental methods can be subdivided into *simultaneous* and *sequential* methods, according to whether several analytes would be determined

at the same time (as in the case of multichannel optical emission spectrometry) or one after another (with the help of a monochromator).

Immediate application of a direct instrumental method (e.g., atomic spectroscopy in one of its many variants) usually represents the most economical approach to elemental analysis provided the procedure in question is essentially unaffected by the sample matrix, or if one has access to appropriate standard reference materials similar in composition to the substance under investigation [22]. The alternative is an analytical method consisting of multiple operations separated by either space or time, often referred to as a *multistep procedure*, as indicated at the left in Figure 5. The possibility of combining two or more discrete techniques adds a whole new dimension to chemical analysis, although there is a long tradition of observing a formal distinction between "separation" and true "determination". Separation in this sense has often been understood to include chemical reactions undertaken for the purpose of preparing a new, more readily separable compound – as a solid phase, for example – together with the actual separation step itself (e.g., filtration or extraction), although the term is sometimes interpreted more literally and limited to the latter activity alone. Cases also come to mind in which individual "separation" and "determination" steps cannot be clearly differentiated (e.g., in chromatography).

A separation step might be preceded by some preliminary treatment of the sample, such as a prechromatographic operation [16], and this might also warrant special attention. *Trace enrichment* is typical of the fields in which prechromatographic techniques have much to offer.

Particularly in trace analysis, and in the absence of standard samples for calibration purposes, there still is no satisfactory alternative to relying at least initially on "wet- chemical" multistep procedures. This entails a detour consisting of sample decomposition with subsequent separation and enrichment of the analyte(s) of interest relative to interfering matrix constituents. A suitable form of the analyte(s) is then subjected to the actual determination step, which may ultimately involve one or more of the direct instrumental methods of analysis.

Multistep procedures are even more indispensable in the analysis of organic substances, where a chromatographic separation is often closely coupled with the actual method of determination, such as IR or mass spectrometry. Separations based on chemical reactions designed to generate new phases for subsequent mechanical isolation (e.g., precipitation, liquid – liquid partition) have also not been completely supplanted in elemental and molecular analysis.

Progress in analytical chemistry is marked by dramatic developments in two areas: (1) an enormous increase in the number of available analytical methods and opportunities for applying them in combination, and (2) new approaches to mathematical evaluation (chemometrics). As a result, most matrices are now subject to characterization with respect to their components both in terms of the bulk sample and at such internal and external phase interfaces as grain boundaries and surfaces – extending in some cases even into the extreme trace range. As in the past, the safest course of action entails separating the component(s) of interest in weighable form, or taking an indirect route via gravimetry or tritrimetry as a way of establishing a state indicative of complete reaction.

Many modern methods of separation and determination result in the generation of some type of "signal", whereby an appropriate sensor or detector is expected to react in response to concentration or mass flow – perhaps as a function of time, and at least ideally in a linear fashion throughout the range of practical interest. Devices such as photocells, secondary electron multipliers, Golay cells, thermal conductivity cells, thermocouples, and flame ionization detectors convey information related to concentration changes. This information takes the form of an electrical signal (either a voltage or a current), which is fed to some type of measuring system, preferably at a level such that it requires no amplification. Sensor development is an especially timely subject, warranting extensive discussion elsewhere (→ Chemical and Biochemical Sensors).

Further processing of an analytical signal may have any of several objectives, including:

1) Incorporation of a "calibration function" that permits direct output of a concentration value
2) Establishing feedback control as one way of managing the data-acquisition process (e.g., in a *process computer*)

3) Recasting the primary signal to reflect more clearly the true analytical objective (e.g., "on-line" Fourier transformation, a common practice now in both IR and NMR spectroscopy)

5.4. Physical Organization of the Analytical Laboratory

Depending on the situation, assignments with which a particular analytical team is confronted might be linked organizationally and physically with the source of the samples in various ways. The following can be regarded as limiting cases:

1) Direct physical integration of the analytical function into the production or organization process, where "on-line" analysis represents the extreme
2) Strict physical separation of the sample source from subsequent analytical activities

It would be pointless to express a general preference for one arrangement or the other, but a few relevant considerations are worth examining.

Analysis "on the spot" eliminates the complications of sample transport, and it offers the potential for saving a considerable amount of time. This in turn facilitates rapid processing, an especially important factor when process control is dependent upon analytical data (e.g., in blast furnace operation). Analysis of this type is always associated with a very specific objective, usually involving a single analytical method and a single specialized type of instrumentation, and its economic viability must be critically evaluated on a case-by-case basis. Costs related to acquisition, amortization, and the repair of expensive equipment must all be considered, as must demands for personnel – who are likely to require special skills and training.

The obvious alternative to integrated analysis is a physically separate, central analytical facility like that traditionally maintained by a large chemical corporation. A laboratory of this sort typically reflects an interest in analysis in the most general sense, with provisions for the utilization of as many as possible – preferably all – of the conventional and fashionable analytical methods in anticipation of a very broad spectrum of assignments. Routine analysis in such a setting can conveniently be combined with the innovative development of new methods and procedures, thereby assuring optimal utilization of equipment that is becoming increasingly sophisticated and expensive. Considering the rapid pace of developments in major instrumentation, and the risks entailed in implementing modern approaches to automation, data processing, and laboratory operations generally, it often becomes apparent that centralization is the only economically justifiable course of action.

Similar considerations underscore the critical importance of continuing education for laboratory personnel, who must of necessity adapt to any changes in hardware. This perspective also sheds additional light on the independent scientific character of analysis, both in the industrial sphere and in academia. The problems encountered are essentially scientific in nature, the questions are fundamental, and the tools engaged in their solution reflect a complex development process that is technically demanding in the extreme.

5.5. The Target of Analysis

One of the fields in Figure 5 (the diagram singling out various stages in an analytical procedure) bears the label "Target of Analysis", and its structure deserves closer scrutiny. Until relatively recently the "target of an analysis" was always a list of constituent elements, together with the corresponding overall composition. An arduous trail of analytical research leads from the dualistic theory of matter (Berzelius and his contemporaries) to an understanding of the fine structure and conformation of molecules in the solid and liquid (dissolved) states, culminating in direct proof of the existence of atoms. In planning an analysis today it is almost self-evident that the first question to be addressed concerns the particular level in the hierarchically ordered concept "target" at which the investigation is to be conducted.

One important property of this hierarchy is that every higher level of order implies a specific set of properties at each of the lower levels. The reverse is not true, however, since the lower stages are independent and do not presuppose any higher degree of structure. Thus, in order to conduct a molecular structure determination on

an organic substrate it is first necessary to ascertain the corresponding elemental composition. Needless to say, analysis at any level in the object hierarchy depends upon the availability of suitable procedures.

Atoms. As shown schematically in Figure 6, the hierarchy of targets begins with *atoms* (and the various *isotopes* of the elements) as the smallest fundamental units with analytical chemical relevance. This is already a rather profound observation in the case of certain geochemical questions, for example, since it is well known that the isotope ratios for such isotopically mixed elements as sulfur or uranium are by no means constant, and an isotope ratio (of chlorine, perhaps) can also be a useful or even indispensable parameter in the practice of mass spectrometry (→ Mass Spectrometry).

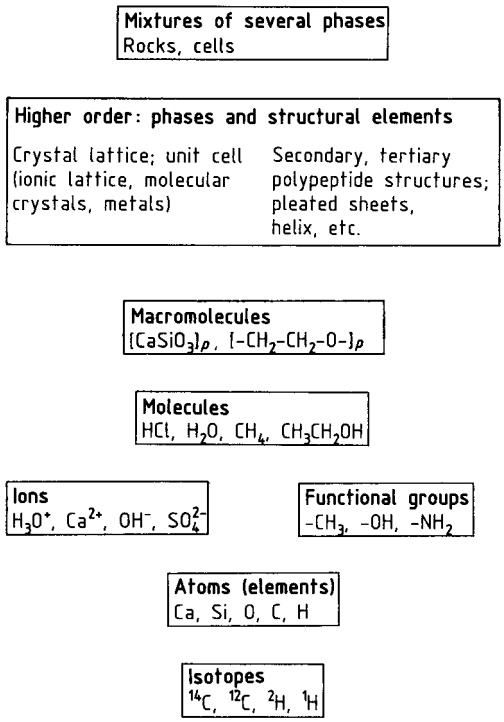

Figure 6. Hierarchical ordering of the various possible targets of analysis

Molecules. *Ions* and *functional groups* have been assigned to a level of their own, located between that of the atoms and that of the *molecules*, which represents the next formal stage in our hierarchical scheme. Chemical reactions long constituted the sole basis for analysis at the molecular level, and together with the methods of atomic and molecular spectroscopy they continue to serve as the foundation of modern analysis.

Macromolecular Species. The transition from high molecular mass substances (*macromolecules*) to the highly ordered macroscopic crystalline range is somewhat indistinct, and it is also here that the concept of the molecule loses its validity (ignoring for the moment the notion of molecular crystals). Qualitatively distinct properties emerge, and these can in turn be exploited for purposes of analytical characterization: external form (crystalline habit), symmetry, ionic lattice structure (as well as molecular lattices and the structures of mixed forms), mechanical and electrical characteristics, specific gravity, phenomena associated with phase formation and conversion, melting point, etc. In the course of analysis, a solid material is degraded to a *matrix*, which often turns out to be a troublesome source of interference. At this stage the solid itself is regarded as homogeneous with respect to its composition, although the structure, surface properties, and microdistribution in terms of atoms or other components might also be subjected to analysis.

Heterogeneous mutliphase systems (*mixtures*) lay claim to a special status, constituting the highest level in the hierarchical classification system. The prototype for this category is a rock: granite, for example, which is composed of three homogeneous substances. Precise identification at the mixture level typically requires the use of special techniques drawn from other disciplines, such as mineralogy or petrography.

Analysis of the Various Targets. A formal distinction can be made today between the detection and determination of free atoms, molecules, or ions on the one hand (via mass or optical spectroscopy in the gaseous state), and the analysis of molecules or ions in solution on the other (using such techniques as UV – VIS spectrophotometry, electrochemistry, and NMR or ESR spectroscopy). Mention should also be made of such "colligative" methods as ebullioscopy, cryoscopy, and gas-density determination. The

overall importance of colligative methods has declined somewhat, but they continue to provide useful information regarding molecular mass and various dissociation phenomena. Their principal field of application today is the analysis of high polymers (macromolecules), the level immediately above molecules in our hierarchy of targets.

Polymeric molecules and ions have been assigned to a level of their own, primarily because of a unique set of methods that has been developed specifically for dealing with analytical targets in this size range. The most important are ultracentrifugation, certain types of liquid chromatography, light scattering, and – not to be overlooked – chemical approaches to the determination of end groups. Certain colligative methods are also very important here because of their extremely high sensitivity.

With respect to the optical analytical methods it could be claimed (cum grano salis) that *dispersive methods* – which today extend into the short-X-ray region – are increasingly being supplemented by *image-forming methods* (e.g., microscopy, polarizing microscopy, stereoscanning, electron microscopy, ultrasound microscopy). Image-forming methods appear to have reached a plateau (at least for the time being) with the quasimechanical/optical principle as manifested in scanning tunneling microscopy, which has succeeded for the first time in making atomic structures "visible". Methods for the study of surfaces [e.g., Auger spectrometry, X-ray photoelectron spectroscopy (XPS), and secondary ion mass spectroscopy (SIMS)] must also now be incorporated under the heading of general structural analysis (→ Surface and Thin-Film Analysis).

Major advances have occurred in direct instrumental approaches in the *bulk analysis* of condensed matter, providing integrated insight into the various components comprising a complex sample; examples of such methods include X-ray fluorescence, atomic emission, atomic absorption, and atomic fluorescence spectrometry. Especially in optical atomic emission spectrometry and mass spectrometry, traditional sources of excitation such as arcs and sparks are increasingly giving way to alternative techniques: direct-current, high-frequency, and microwave plasmas; glow discharges; and lasers (→ Laser Analytical Spectroscopy).

Information regarding the qualitative and quantitative distribution of elements within individual phases and at phase interfaces (grain boundaries) is the primary goal of *microdistribution analysis*, in which special probe techniques involving electrons, ions, photons, and neutrons are used to excite the sample under investigation. In the organic realm, biological cells might be singled out as the prototypical analytical substrate. A fundamental distinction must of course be made here between "nondestructive" methods and methods that concentrate on the sample as a whole.

6. Definitions and Important Concepts

Terminology plays an important role in analysis, and several technical terms have in fact already appeared in preceding sections of this essay. Here we consider explicitly a select subset of these terms, mainly ones whose widespread usage is relatively recent, as well as a few that are often utilized incorrectly.

6.1. Sensitivity, Limit of Detection, and Detection Power

These three expressions tend to be used very loosely, even among analysts – despite the fact that each is subject to very precise definition [1].

The *sensitivity* E of an analytical method expresses the dependency of a measured response on the analytical value of primary interest. It is defined as the first derivative of the measurement function:

$$F(y)\ E = F'(y) = \mathrm{d}x/\mathrm{d}y$$

In the case of a linear calibration function of the type $x = a_x \cdot y + b_x$, the sensitivity is equal to the slope of the calibration line; i.e.,

$$E = \Delta_x/\Delta_y = a_x$$

The *limit of detection* of an individual analytical procedure is the lowest amount of an analyte in a sample which can be detected but not necessarily quantified as an exact value [29]. Expressed as a concentration c_{L} or a quantity q_{L} the *limit of*

detection is derived from the smallest signal x_L which can be detected with reasonable certainly for a given analytical procedure. The value of x_L is given by the equation

$$x_L = x_{bl} + k s_{bl}$$

where x_{bl} is the mean of the blank measurements. s_{bl} is the standard deviation of the blank measurements, and k is a numerical factor chosen according to the level of confidence required. For many purposes, the limit of detection is taken to be $3\,s_{bl}$ or $3 \times$ signal to noise ratio, but other values of k may be used if appropriate. The concept "limit of detection" is applicable only to a particular analytical procedure, one that can be precisely defined with respect to all its parameters, whereas the *detection power* is a crude estimate associated with an idealized analysis, in which external interfering factors are largely ignored. This term is therefore reserved for characterizing an analytical *principle*. The demand for analytical procedures with ever greater detection power is especially acute in the context of biologically relevant trace elements because of the ubiquitous concentrations of these materials in all natural matrices. It is the environmental concentrations that effectively establish lowest levels of the corresponding elements that are subject to determination in any biotic matrix. In most cases these levels are in the range > 0.1 ng/g, and thus within a region that could be regarded as practically accessible – at least in principle. Exceptions include the concentrations of certain elements in Antarctic or Arctic ice samples, for example, or samples from research involving ultrapure substances, where the goal is to prepare materials (e.g., metals) of the highest possible purity. Relevant impurity concentrations in cases such as these may fall in the pg/g range, often far below the environmental background concentrations of the elements in question.

Establishing the trace-element content for a bulk material typically requires a sample at the upper end of the milligram or lower end of the gram range, but increasing importance is being attached to the acquisition of detailed information regarding the distribution and bonding states of elements within specific microregions (see Sections 6.5, 6.7, 6.14).

Studies related to construction materials and other solids frequently rely on information regarding the distribution of elements at external and internal boundary surfaces, including grain boundaries. With biotic matrices the attention may be focused on tissue compartments, individual cells, or even cell membranes, whereas an environmental analysis might be concerned with individual aerosol particles. Investigations in areas such as these – especially projects involving in situ microdistribution analysis – serve to further promote the ongoing quest for ever lower absolute limits of detection. Assume, for example, one wished to determine the elemental distribution in various protein fractions from blood serum. Blood-serum background levels are in the low ng/g range, so an analytical method would be required with a detection power 10 – 100 times greater. Attempting to determine quantitatively an elemental concentration on the order of 1 ng/g in a sample weighing only 1 µg presupposes an absolute detection power in the femtogram range (1 fg = 10^{-15} g).

An interesting question in this context is the minimum mass of an element that would theoretically be required for a successful determination with a given statistical degree of certainty. If the statistical error is not to exceed 1 %, for example, then the root-N law of error analysis specifies that at least 10 000 atoms must be present for the determination to succeed. For the element zinc (atomic mass 60) this would correspond to a mass of only one attogram (10^{-18} g) [23]. Given the potential of laser spectroscopy (→ Laser Analytical Spectroscopy), attaining such a goal is no longer considered utopian.

The question "Is it really necessary that analysis continue to strive for greater sensitivity?" must therefore be answered with an unequivocal "Yes". It is naive to suggest (as some have) that the blame for the many problems with which we are today confronted lies exclusively with the availability of increasingly powerful analytical tools. Such a biased perspective reflects at best uncritical – indeed, irresponsible – misuse of the powers of analysis, although misuse of this type can never be ruled out completely.

6.2. Reliability – Measurement Uncertainty

One of the first problems faced by analysts is whether a method will provide a result that is fit

for its intended purpose. i.e., whether it will produce a result of the required "accuracy". A quantitative indication of the accuracy is required if the user of this result is to make any judgement on the confidence to be placed in it, or to compare it in a rational way with the results of other analyses. The statement of a result is not complete without information about the "accuracy" or the "uncertainty".

There will always be an uncertainty about the correctness of a stated result, even when all the known or suspected components of error have been evaluated and the appropriate correction factors applied, since there is an uncertainty in the value of these correction factors. In addition, there will the uncertainty arising from random effects.

Recent developments have led to the formulations of consistent and quantitative procedures for evaluating and reporting the uncertainty of a result, which are applicable in all areas of measurement. These procedures have been set out in the ISO Guide to the Expression of Uncertainty in Measurement [24] and their application to analytical chemistry is described in the EURACHEM (Cooperation for Analytical Chemistry in Europe) Guide Quantifying Uncertainty in Analytical Measurement [25].

The approach set out in the ISO Guide treats all sources of uncertainty in a consistent manner and thus avoids the difficulties encountered in some previous approaches to the evaluation of uncertainty, which treated the uncertainty arising from systematic effects in a different manner to that arising from random effects.

In essence, for chemical analysis the ISO definition of uncertainty is:

> A parameter associated with the result of an analysis that characterizes the dispersion of the values that could reasonably be attributed to the concentration of the analyte.

Thus the analyst when reporting the result of an analysis is also being asked to provide a parameter that gives a quantitative indication of the range of the values that could reasonably be attributed to the concentration of the analyte. The ISO Guide recommends that this parameter should be reported as either:

A standard uncertainty defined as:

> uncertainty of the result of a measurement expressed as standard deviation (→ Chemometrics)

or as *expanded uncertainty* defined as:

> a quantity defining an interval about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could be attributed to the measurand (concentration of the analyte) and which is obtained by multiplying the *standard uncertainty* by a *coverage factor*, which in practice is typically in the range 2 – 3.

It is common practice to report the standard uncertainty using a value of 2 since this gives an interval with confidence level of appoximately 95 %.

The evaluation of uncertainty requires a detailed examination of the measurement procedure. The first step is to identify the possible sources of uncertainty. The next step is to evaluate the size of the contribution from each source, or the combined contribution from a number of sources, expressed as a standard deviation. These contributions are then combined to give the standard uncertainty. Detailed examples are given in the EURACHEM Guide [25].

6.3. Elemental Analysis

A formal distinction between "elemental analysis" and "elementary analysis" (Section 6.4) is seldom carefully observed in English. *Elemental analysis* in the present context is understood to mean a determination of essentially all the elements present in a sample, irrespective of the type of bonding involved or the constitution of the matrix. Means toward that end include not only the classical methods (gravimetric analysis, titrimetry, spectrophotometry, electrochemical and kinetic methods, etc.) but also atomic spectrometric and radioanalytical methods, some of which are essentially nondestructive. From the standpoint of reliability, classical chemical methods are rarely surpassed by instrumental methods, though the latter typically do provide lower limits of detectability, and they are faster and more economical, generally offering the added potential for simultaneous multielement determination.

6.4. Elementary Analysis

Elementary analysis in the classical sense refers to quantitative determination of the constituent elements in an organic compound, especially carbon, hydrogen, oxygen, nitrogen, sulfur, the halogens, and phosphorus, although the definition would today be expanded to cover determination of any element present in an organic structure.

Precise characterization of organic substances became possible for the first time as a result of elementary analysis techniques developed by BERZELIUS, LIEBIG, DUMAS, KJELDAHL, and many others. The fundamental principle, which relies essentially on combustion of the material under investigation with subsequent determination of the combustion products, was adapted to microscale analysis by F. PREGL and later developed by other analysts to such a point that "microelementary analysis" now requires only a few milligrams of sample. This has led not only to shorter analysis times, but also the possibility of analyzing extremely small amounts of valuable natural substances.

Microelementary analysis has now been carried to a high degree of perfection as well as almost complete automation. Despite the advent of modern physical methods (MS, NMR, etc.) the classical techniques have lost little of their significance for synthetic chemists and biochemists, thanks mainly to their considerable advantage of providing reliable *absolute* values of actual mass proportions [26], [27].

There has been no lack of attempts to reduce requisite sample sizes even farther into the lower microgram range [28], [29], but the associated techniques are often quite time-comsuming, and efforts in this direction have been largely superseded by mass spectrometry. Nevertheless, work in the field of "ultramicroelementary analysis" is still of interest, especially since it has produced impressive evidence of the limits of classical microanalysis.

6.5. Microanalysis and Micro Procedures

"Microanalysis" is a term originally associated with classical analytical techniques capable of providing very accurate results from as little as ca. 1 mg of substance (relative error ≤ 1 %; see the discussion of organic microelementary analysis in Section 6.4).

The term "micro procedure" is one of several now defined with respect to mass range by a DIN standard (DIN 32 630):

	Sample size
macro procedure	> 100 mg
semimacro procedure (also known as a semimicro procedure)	1 – 100 mg
micro procedure	1 µg – 1 mg
submicro procedure	< 1 µg

Progress in analytical chemistry in recent decades has been so extensive that most procedures could today be included in the "micro" or even "ultramicro" category, especially ones involving chromatography or spectroscopy. Under these circumstances the term "microanalysis" should now be restricted exclusively to the field of elementary analysis. "Chemically" oriented techniques for ultramicro elemental analysis in the microgram range [23] are based on extremely small reaction volumes (e.g., as little as ca. 1 µL) and correspondingly designed facilities for sample manipulation (ultramicro balances, microscopes, capillaries, and extremely tiny tools). Such methods have now been largely superseded by total reflection X-ray fluorescence (TRXRF), which permits rapid, simultaneous, quantitative determination of most elements with atomic numbers > 11 at levels extending into the lower picogram range [30], a striking example of the impressive advance in instrumental microbulk analysis (→ X-Ray Fluorescence Spectrometry).

6.6. Stereochemical and Topochemical Analysis

In this case the goal is to describe the target object (an atom, a molecule, or some other component of a solid phase or a solution) with respect to its spatial orientation in its surroundings. Sometimes the frame of reference is an external surface or an internal boundary surface (grain boundary). *Structural analysis* has as its ultimate objective describing all aspects of the overall structure of a particular phase, including the conformations of individual structural elements, perhaps also as a function of time. *Consti-*

tutional analysis produces information regarding relative and absolute arrangements of atoms or atomic groupings (functional groups) within a molecule.

6.7. Microdistribution Analysis

Microdistribution analysis is a special type of topological analysis directed toward establishing lateral and depth distributions of the various elements making up a solid – preferably with explicit reference to the ways in which these elements are bonded, all expressed with the highest possible degree of positional resolution. Distribution with respect to phase boundaries may be important here as well. The designated goal is approached with the aid of techniques that permit beams of rays (e.g., laser photons, electrons, ions, or neutrons) to be focused extremely sharply, with a typical cross section of 1 μm or less. Alternative techniques maintain such spatial relationships as may exist within a series of signals, employing a multidimensional detector to transform the crude data into images.

Every effort is made here to achieve the highest possible absolute power of detection. Microdistribution analysis represents the primary field of application for *microprobe* techniques based on beams of laser photons, electrons, or ions, including electron microprobe analysis (EPMA), electron energy-loss spectrometry (EELS), particle-induced X-ray spectrometry (PIXE), secondary ion mass spectrometry (SIMS), and laser vaporization (laser ablation). These are exploited in conjunction with optical atomic emission spectrometry and mass spectrometry, as well as various forms of laser spectrometry that are still under development, such as laser atomic absorption spectrometry (LAAS), resonance ionization spectrometry (RIS), resonance ionization mass spectrometry (RIMS), laser-enhanced ionization (LEI) spectrometry, and laser-induced fluorescence (LIF) spectrometry [31–40].

6.8. Surface Analysis

Surface analysis is in turn a specialized form of microdistribution analysis, one that provides information on the coverage, distribution, and content of components either at the surface of a solid or in discrete layers located near the surface. Elemental analysis in this context is conducted in a single plane with no attempt at lateral resolution, utilizing, for example, total reflection X-ray fluorescence (TRXRF), glow-discharge mass spectrometry (GDMS), or secondary neutron mass spectrometry (SNMS). Positional resolution can also be achieved with probe techniques such as Auger electron spectrometry (AES) or secondary ion mass spectrometry (SIMS), and to a limited extent with Rutherford back-scattering (RBS) and X-ray photoelectron spectroscopy (XPS) [40–43] (→ Surface and Thin-Film Analysis).

6.9. Trace Analysis

Here we encounter another term associated with a range whose definition has changed considerably with time. A "trace" was once understood to be a no longer determinable but nonetheless observable concentration of some undesired companion substance (impurity) within a matrix. In the meantime, trace analysis has become an important and very precise field of inquiry – subject to certain restrictions with respect to the achievable reliability, but indispensable in a number of disciplines (→ Trace Analysis). There is little point in attempting to express a "trace" in terms of absolute mass units; data should instead be reported on the basis of content (concentration) in a form such as "μg/kg" (mass proportion) or "μg/L" (mass concentration). These units are to be used in preference to the very popular abbreviations "ppm", "ppb", and "ppt", which need an additional indication to the respective unit (mass, volume, amount of substance).

A warning is in order against the practice of emphasizing wide disparities in content through inappropriate comparisons that make a sensational impression at the expense of reality. Phrases like "a needle in a haystack", or "a grain of wheat in a hundredweight of rye" used as metaphors for "ppm" are inconsistent with the fact that what is really at issue is a trace *proportion*; that is, a homogeneous distribution or quasicontinuum within which single individuals can be isolated only hypothetically, and only at the molecular level.

Trace constituents at levels as low as a few micrograms per ton (i.e., in a mass ratio of 1 to 10^{12}) are today subject to meaningful analytical determination thanks to highly developed methods of separation and enrichment (multistep procedures, see Chap. 3 and Section 6.11) [44–47]. The principal challenge facing the trace analyst is that diminishing concentration leads to a rapid increase in systematic error [15], [47]. Extreme trace analysis with respect to the elements is therefore subject to large systematic deviations from "true" content, even though results obtained with a particular method may be quite reproducible.

Trace determinations based on atomic spectroscopy are usually matrix-dependent, *relative* methods, requiring the availability of standard reference samples for calibration. Unfortunately, no such standards yet exist for extreme ranges, so one is forced to rely instead on multistep procedures whereby trace amounts of elements of interest are excited in isolated form and within the smallest possible volume of analyte. "Limits of detectability" reported in the literature for methods of trace elemental analysis are almost invariably extrapolations based on determinations actually carried out at fairly high concentration. In no sense can these limits be regarded as reflecting real conditions owing to the problem of systematic error. Systematic errors are extremely difficult to detect, so it is advisable that one verify the validity of data acquired at each stage of the work using an alternative analytical approach. Only when data from different procedures agree within the appropriate statistical limits of error should one speak in terms of "reliable" results. Success in identifying and eliminating all the sources of error in an extreme trace analysis therefore presupposes a considerable amount of experience and a well-developed capacity for self- criticism.

Preferred methods in trace determination of the elements include atomic absorption spectrometry (AAS), optical emission spectrometry (OES) with any of a wide variety of excitation sources [e.g., sparks, arcs, high-frequency or microwave plasmas (inductively coupled plasma, ICP; microwave induction plasma, MIP; capacitively coupled microwave plasma, CMP), glow discharges (GD), hollow cathodes, or laser vaporization (laser ablation)], as well as mass spectrometry (again in combination with the various excitation sources listed), together with several types of X-ray fluorescence (XRF) analysis [46].

A special place is reserved for methods of *activation analysis*, involving slow and fast neutrons, charged particles, or photons, applied either directly or in combination with some type of radiochemical separation (Section 6.13). These methods quickly became almost indispensable, especially in extreme trace analysis of the elements, owing to a low risk of contamination and detection levels in at least some cases that are exceptionally favorable (→ Activation Analysis).

Electrochemical methods continue to be important as well, including inverse voltammetry, coulometry, amperometry, and potentiometry (→ Analytical Voltammetry and Polarography); indeed, their overall role has actually been expanded with the development of such chemical techniques as ion chromatography and chelate HPLC.

Problems associated with extreme trace analysis of the elements also affect extreme trace analyses of organic compounds, although background levels tend to be less relevant in this case [48]. All the separation methods most commonly applied to organic substances are chromatographic in nature, including thin layer chromatography (TLC; → Thin Layer Chromatography), high-performance (or high-pressure) liquid chromatography (HPLC; → Liquid Chromatography), gas chromatography (GC; → Gas Chromatography), and electrophoresis (more recently: capillary electrophoresis; → Electrophoresis), preferably combined with on-line mass or infrared spectrometry.

6.10. Trace Elements

The term "trace element" was first introduced in biochemistry after it became apparent during the 1920s that very low levels of certain elements in food can be important to life. Nine such elements had been identified by 1959, whereas today more than twenty different elements are regarded as essential, including several previously recognized only as toxic (e.g., arsenic, lead, and cadmium). Ambivalent physiological characteristics have now been ascribed to many elements, where toxicity may be manifested at high concentrations, but a low concentration is

an absolute requirement, since a concentration even lower – or complete absence of such an element – leads directly to symptoms of illness [49]. A more appropriate descriptive term applicable in a nonbiological context might be "elemental traces".

6.11. Multistep Procedures

Situations frequently arise in which direct instrumental methods of analysis are inapplicable, perhaps because the corresponding detection power is insufficient, or in the case of a matrix-dependent method because no suitable calibration standards are available to correct for systematic errors. The best recourse is then a *multistep procedure*, in which actual determination is preceded by sample preparation, digestion, separation, or preconcentration steps. Individual operations within such a procedure must be linked as closely as possible, as in a "one-pot method" or one of the flow-injection or continuous-flow techniques that lend themselves so readily to automation. The goal is to concentrate an analyte from a rather large volume of solution (on the order of milliliters) for subsequent analysis on the microliter scale. Systematic errors can be minimized with on-line procedures, permitting highly reliable analysis at the picogram-per-milliliter level.

A search will also continue for elemental analysis techniques based on direct instrumental methods with enhanced powers of detection and more or less complete matrix independence. The motivation for this search goes beyond mere economic factors: direct methods are also less likely to be held hostage to blank readings, because physical sources of excitation (e.g., photons, electrons, charged particles, neutrons), which H. Malissa and M. Grasserbauer [50] characterize as "physical reagents", are essentially free of material contamination.

The trace analysis of organic substances is especially dependent on multistep procedures. In this case losses due to adsorption and vaporization are more worrisome potential sources of systematic error than elevated blank values.

6.12. Hyphenated Methods

This unfortunate piece of terminology is intended to emphasize the fact that multiple techniques, usually of an instrumental nature, often lend themselves to direct physical coupling, resulting in combinations whose formal designations contain hyphens (e.g., GC – MS). In contrast to the multistep procedures discussed previously, this type of combination involves a "real time" connection and true physical integration. The greatest challenge is to develop satisfactory interfaces for joining the various separation and detection systems. This particular problem is one that has long plagued the otherwise promising HPLC – MS combination.

6.13. Radioanalytical Methods and Activation Analysis

Methods based on nuclear reactions are restricted to laboratories specially equipped for handling radioactive substances (radionuclides) under the close supervision of trained personnel. The first important breakthroughs in trace analysis of the elements (e.g., in semiconductor applications) accompanied the development of activation analysis, which was originally based on excitation with slow or fast neutrons but later broadened to encompass the use of charged particles and photons as well. Activation methods were long held in exceptionally high esteem in the field of trace analysis, although competition eventually surfaced in the form of atomic spectroscopy. The drawbacks of activation methods (long analysis times, high cost, and fear of radioactivity) are now perceived by many to outweigh the advantages (high detection power for many elements and relatively high reliability due to minimal complications from matrix effects or contamination), and the activation technique has been demoted to the status of one approach among many. This actually increases the need for stressing that activation analysis must still be regarded as an indispensable technique. Sometimes it in fact represents the only viable solution to a problem in extreme trace elemental analysis (e.g., in high-purity substance studies), whether applied directly (*instrumental* activation analysis) or – more often – in combination with radiochemical methods of separation (*radiochem-*

ical activation analysis). Activation analysis also plays an essential role in the preparation of standard reference samples because of the fact that it is so reliable.

The same considerations apply with even greater force to the use of radioactive tracers in elemental and especially molecular analysis (→ Radionuclides in Analytical Chemistry).

6.14. Species Analysis (Speciation)

In biology the term *species* is used to describe a population of organisms with hereditary features that survive even after cross-breeding. The related chemical term *species analysis* was first introduced by biochemists, where a "chemical species" is understood to be a particular molecular form (configuration) of atoms of a single element or a cluster of atoms of different elements. Biologists thus define the term "species" very clearly, whereas chemists apply it in various ways:

1) For the *analysis* of a species, leading to its identification and quantification within some defined region of a sample (*speciation*)
2) For describing the *abundance* or *distribution* of various species of an element within a particular volume
3) In conjunction with the *reactivity* of a given species
4) With respect to the *transformation* of one species into another

Accordingly, various *categories* of chemical species can also be distinguished, including "original" or "conceptional" species, "matrix" species, and "analyte" species.

A species is said to be *original* or *conceptional* if it is resistant to change in contact with other matrices. If a chemical change does occur as a result of such contact, the material is called a *matrix* species. The third category refers to a species that undergoes a change during the course of an analytical procedure, in which case it is some new species (the *analyte* species) that becomes the subject of analytical detection. Consider, for example, the original (conceptional) species CH_3Hg^+, which in soil forms the matrix species CH_3Hg – humic acid, but is subsequently determined as the analyte species CH_3HgCl. Transformations of this type are a major source of the considerable challenge posed by species analysis relative to determining the total content of an element. Transformation may well begin as early as the sampling process, continuing throughout the period of sample storage and at the time of the analysis itself, so that original species present in situ and in vivo are never actually detected [51], [52]. Above all, in both environmental analysis and toxicology, detection of the original species is becoming more and more important (→ Sample Preparation for Trace Analysis)

6.15. Chemometrics

Chemometrics is a relatively new field encompassing those aspects of chemical analysis associated directly with measurement techniques, especially principles underlying the various types of detection. Opinions differ with respect to the meaning, purpose, and limitations of this new discipline, but a relatively clear set of ideas is beginning to prevail [53]. According to K. Doerffel et al., chemometrics is concerned with evaluation of observed analytical data with the aid of mathematics, especially statistical methods. Chemometric methods facilitate the extraction of useful information even when the noise level of a signal is high, as well as the establishment of relationships linking multiple observations even when the results themselves seem widely divergent. A good example is provided by the mathematical resolution of partially overlapping signals ("peaks") in a chromatogram.

The tools of chemometrics encompass not only the familiar (univariant) methods of statistics, but especially the various multivariant methods, together with a package of "pattern-recognition" methods for time-series analyses and all the known models for signal detection and signal processing. Chemometric methods of evaluation have now become an essential part of environmental analysis, medicine, process analysis, criminology, and a host of other fields.

Chemometric methods have also been adapted to the development of labor-saving analytical strategies – the establishment, for example, of the ideal sampling frequency in a process analysis, or simplification of a multicomponent

analysis so that it reflects only the truly relevant features. In addition, chemometrics plays an important part in quality-assurance programs directed toward analytical investigations. In the future, chemometrics should make a valuable contribution to the design of "legally binding" analyses with statistically assured results [54–58] (→ Chemometrics ; see also Chap. 10).

6.16. DNA Analysis

DNA techniques are already being applied in a number of ares of analysis such as human health, idnetification of sex in certain species, personal identification, environmental and food analysis. The Polymerase Chain Reaction (PCR) is one of the most powerful methods of producing material for analysis from very small samples and can achieve up to a 10^6-fold increase in the target DNA. For example, with this technique it is possible to detect a range of pathogenic microorganisms with a sensitivity which is orders of magnitude greater than previously achievable, and that is beyond the limits required for public health. It is also a very powerful technique for checking for food adulteration and food specification. Applications of this technique are growing very rapidly and it could be one of the most important of the recent developments in analytical science.

7. References

1. K. Danzer, E. Than, D. Molch, L. Küchler: *Analytik-Systematischer Überblick,* 2nd ed., Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig 1987.
2. R. Bock: *Methoden der Analytischen Chemie,* **vol. 1: "Trennungsmethoden, " vol. 2: parts 1, 2, and 3: "Nachweis- und Bestimmungsmethoden",** VCH Verlagsgesellschaft, Weinheim 1974 – 1987.
3. H. Kienitz et al. (eds.): *Analytiker-Taschenbuch* **vols. 1 – 11,** Springer-Verlag, Berlin 1980 – 1993.
4. G. Svehla (ed.): *Wilson and Wilson's Comprehensive Analytical Chemistry,* **vols. 1 – 28,** Elsevier, Amsterdam 1959 – 1991.
5. R. Kellner, J.-M. Mermet, M. Otto, H. M. Widmer (eds.): *Analytical Chemistry,* Wiley-VCH, Weinheim 1998.
6. K. Cammann, *Fresenius J. Anal. Chem.* **343** (1992) 812 – 813.
7. M. Valcarcel, *Fresenius J. Anal. Chem.* **343** (1992) 814 –816.
8. C. R. Fresenius: *Anleitung zur qualitativen chemischen Analyse,* 12th ed., Vieweg u. Sohn, Braunschweig 1866, p. 4.
9. F. Szabadvary: *Geschichte der analytischen Chemie,* Vieweg u. Sohn, Braunschweig 1966.
10. W. Ostwald: *Die wissenschaftlichen Grundlagen der analytischen Chemie,* Leipzig 1894.
11. Ullmann, 3rd ed., vol. 2/I.
12. H. Malissa, *Fresenius J. Anal. Chem.* **337** (1991) 159.
13. H. Malissa, *Fresenius J. Anal. Chem.* **343** (1992) 836.
14. K. Danzer, *Mitteilungsblatt der Fachgruppe Analytische Chemie der GDCh* **4/1992** M 104 – M 110.
15. G. Tölg, *Naturwissenschaften* **63** (1976) 99.
16. W. Dünges: *Prächromatographische Mikromethoden,* Hüthig-Verlag, Heidelberg 1979.
17. W. Richter in *Report on the Comité Consultarif pour la Quantité de Matière,*1st meeting, Paris 1995.
18. X. R. Pan, *Accred. Qual. Assur.* **1** (1996) 181 – 185.
19. P. De Bièvre, *Accred. Qual. Assur.* **3** (1998) 481.
20. W. Wegscheider, *Accred. Qual. Assur.* **4** (1999) 478 – 479.
21. J. W. A. Klaessens, G. Kateman, B. G. M. Vanderginste, *TrAC Trends Anal. Chem.* **4** (1985) 114.
22. B. Griepink, *Fresenius J. Anal. Chem.* **337** (1990) 812.
23. G. Tölg in G. Svehla (ed.): *Wilson and Wilson's Comprehensive Analytical Chemistry,* **vol. III,** Elsevier, Amsterdam 1975, pp. 1 – 184.
24. *ISO Guide to the Expression of Uncertainty in Measurement,* ISO, Geneva, Switzerland, 1993.
25. The Quantification of Uncertainty in Chemical Analysis. Available for download from the EURACHEM web site (www.vtt.fi/ket/eurachem).
26. F. Ehrenberger, S. Gorbach: *Quantitative organische Elementaranalyse,* VCH Verlagsgesellschaft, Weinheim 1991.
27. T. S. Ma, R. C. Rittner: *Modern Organic Elemental Analysis,* Marcel Dekker, New York 1979.

28. G. Tölg: *Ultramicro Elemental Analysis,* Wiley-Interscience, New York 1970.
29. W. J. Kirsten: *Organic Elemental Analysis – Ultramicro, Micro, and Trace Methods,* Academic Press, New York 1983.
30. G. Tölg, R. Klockenkämper, *Spectrochim. Acta Part B* **48 B** (1993) 111 – 127.
31. K. Kiss: *Problem Solving with Microbeam Analysis,* Elsevier, Amsterdam 1988.
32. E. Fuchs, H. Oppolzer, H. Rehme: *Particle Beam Microanalysis, Fundamentals, Methods and Applications,* VCH Verlagsgesellschaft, Weinheim 1990.
33. J. C. Vickerman, A. E. Brown, N. M. Reed: *Secondary Ion Mass Spectrometry: Principles & Applications,* Oxford University Press, Oxford 1990.
34. A. Benninghoven et al. (eds.): "Secondary Ion Mass Spectrometry SIMS VII", *Proceedings of the 7th International Conference on Secondary Ion Mass Spectrometry,* J. Wiley & Sons, Chichester 1990.
35. M. Grasserbauer, H. W. Werner (eds.): *Analysis of Microelectronic Materials and Devices,* J. Wiley & Sons, Chichester 1991.
36. L. Moenke-Blankenburg: "Laser Micro Analysis", *Chemical Analysis,* vol. 105, J. Wiley & Sons, New York 1989.
37. K. Niemax in: *Analytiker-Taschenbuch,* vol. 10, Springer-Verlag, Heidelberg 1991, pp. 1 – 28.
38. J. Uebbing, A. Ciocan, K. Niemax, *Spectrochim. Acta Part B* **47 B** (1992) 601.
39. C. M. Miller, J. E. Parks (eds.): "Resonance Ionization Spectroscopy 1992", *Inst. Phys. Conf. Ser.* **128,** Institute of Physics Publishing, Bristol 1992.
40. J. M. Watts: *Methods of Surface Analysis,* C.K.P., Cambridge 1989.
41. M. Grasserbauer, *Philos. Trans. R. Soc. London* **A 333** (1990) 113.
42. J. C. Riviere: *Surface Analytical Techniques,* Oxford University Press, Oxford 1990.
43. D. Briggs, M. P. Seah: *Practical Surface Analysis,* 2nd ed., J. Wiley & Sons, Chichester, "Auger and X-Ray Photoelectron Spectroscopy", vol. 1, 1990; "Ion and Neutral Spectroscopy", vol. 2, 1992.
44. A. Mizuike: *Enrichment Techniques for Inorganic Trace Analysis,* Springer-Verlag, Heidelberg 1983.
45. J. Minczewski, J. Chwasstowska, R. Dybczinski: "Separation and Preconcentration Methods", in *Inorganic Trace Analysis,* Ellis Horwood Ltd., Chichester 1982.
46. G. Tölg, *Anal. Chem. Acta* **238** (1993) 3 – 18.
47. G. Tölg, P. Tschöpel: "Systematic Errors in Trace Analysis", in Z. B. Alfassi (ed.): *Determination of Trace Elements,* Balaban Pub., Rehovot, in press.
48. K. Beyermann: "Organische Spurenanalyse", in H. Hulpkje, H. Hartkamp, G. Tölg (eds.): *Analytische Chemie für die Praxis,* Thieme-Verlag, Stuttgart 1982.
49. G. Tölg, in H. Malissa, M. Grasserbauer, R. Belcher (eds.): *Nature, Aim and Methods of Microchemistry,* Springer-Verlag, Wien 1981, p. 203.
50. M. Grasserbauer, *Angew. Chem.* **93** (1981) 1059.
51. M. Bernhard, F. F. Brinckman, K. J. Irgolic: "The importance of Chemical 'Speciation' in Environmental Processes", in M. Bernhard, F. E. Brinckman, P. J. Sadler (eds.): *Dahlemkonferenzen 1984,* Springer-Verlag, Heidelberg 1986.
52. Group Report, Importance and Determination of Chemical Species in Biological Systems in : "The Importance of Chemical 'Speciation' in Environmental Processes, " in M. Bernhard, F. E. Brinckman, P. J. Sadler (eds.): *Dahlemkonferenzen 1984,* Springer-Verlag, Heidelberg 1986, pp. 17 – 38.
53. K. Doerffel, K. Danzer, G. Ehrlich, M. Otto, *Mitteilungsbl. Chem. Ges. DDR* **31** (1984) 3.
54. D. L. Massart et al.: *Chemometrics: A Textbook,* Elsevier, Amsterdam 1988.
55. M. A. Sharaf, D. L. Illman, B. R. Kowalski: *Chemometrics,* J. Wiley, New York 1986.
56. R. G. Brereton: *Chemometrics, Applications of Mathematics and Statistics to Laboratory Systems,* Ellis Horwood, Chichester 1990.
57. B. G. M. Vandeginste, *Fresenius J. Anal. Chem.* **337** (1990) 786.
58. St. J. Haswell: *Practical Guide to Chemometrics,* Marcel Dekker, New York 1992.

Weighing

WALTER E. KUPPER, Madison, New Jersey 07940, United States

1. **Introduction** 1
2. **The Principle of Magnetic Force Compensation** 3
3. **Automatic and Semiautomatic Calibration** 3
4. **Processing and Computing Functions** 3
5. **Balance Performance** 4
6. **Fitness of a Balance for Its Application** 5
7. **Gravity and Air Buoyancy** 6
8. **The Distinction Between Mass and Weight** 6
9. **Qualitative Factors in Weighing** 7
10. **Governmental Regulations and Standardization** 7
11. **References** 8

1. Introduction

The field of mass measurement, commonly called weighing, is conveniently structured according to the degree of numerical precision involved. The classification and nomenclature of modern laboratory balances follows a decimal pattern based on the step size d of the associated digital display:

1) Precision balances: $d = 1$, 0.1, 0.01, or 0.001 g
2) (Macro) analytical balances: $d = 0.1$ mg
3) Semimicro balances: $d = 0.01$ mg
4) Microbalances: $d = 0.001$ mg
5) Ultramicrobalances: $d = 0.0001$ mg

Weighing capacities range from 5 g for an ultramicrobalance to >60 kg for the largest precision balances (Figs. 1, 2, 3, 4, 5). Precision weighing instruments of greater capacity, known as industrial precision scales and not generally used in the laboratory, are available in a weighing range up to 6 t and $d = 0.1$ kg.

The words "balance" and "scale" are often used interchangeably. "Balance" is derived from the Latin *bilanx*, having two pans [1]. "Scale" and "scales" are short forms of "a pair of scales" (old English, meaning dishes or plates). Modern weighing instruments no longer have two weighing pans, but the words "balance" and "scale" survive. The term "balance" is preferred for the more precise weighing instruments found in a laboratory, while "scale" is appropriate for all other weighing equipment.

Figure 1. Microbalance Mettler MX5, 5100 mg × 1 µg
Automatic self-calibration with built-in reference masses.

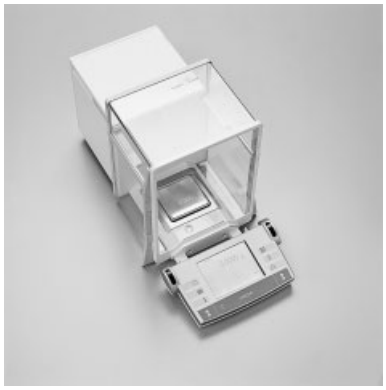

Figure 2. Analytical balance Mettler AX201, 220 g × 0.01 mg
Automatic self-calibration with built-in reference masses.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_027

Figure 3. Analytical balance Mettler AB204-S, 210 g × 0.1 mg
Self-calibration with built-in reference mass.

Figure 4. Precision balance Mettler PR8002, DeltaRange, 8100 g × 0.01 g
Self-calibration with built-in reference mass.

Figure 5. Precision balance Mettler SR32001, 32 100 × 0.1 g
Self-calibration with built-in reference mass.

2. The Principle of Magnetic Force Compensation

Analytical and precision balances are based on the principle of electromagnetic force compensation, where the weight of a sample is counterbalanced by a corresponding electromagnetic force. The measuring sensor or transducer in a balance is best described as a linear motor; i.e., it is an electromechanical aggregate that generates a straight-line force and motion on the basis of an electric current.

The operating principle of a loudspeaker (Fig. 6) provides a useful analogy. In a loudspeaker, an oscillatory, voice-modulated current is fed through a coil (c) to create a small magnetic field that interacts with the larger field from a surrounding permanent magnet (d) to generate an oscillatory force that in turn moves a diaphragm or voice cone (a), thus producing sound. In a balance sensor, direct current through a coil (c) similarly produces a static force of precisely the magnitude required to counterbalance the weight of a sample load (b). A position-maintaining feedback circuit (e) controls this current in such a way as to keep the overall magnetic force in equilibrium with the weight. The strength of the required current in this case provides a measure of the weight; i.e., the balance establishes a mass in grams on the basis of a current measured in milliamperes.

A typical electronic balance incorporates a system of links and levers like that shown schematically in Figure 7. The weighing pan (a) is guided by two parallel horizontal links and counterbalanced by the force coil (g) acting through a lever (c). The pivots (b) in this linkage are metallic leaf springs, called *flexures*. In the weighing process, the pan is initially depressed by the sample, after which it is restored to its null level by the servoaction of the force motor. This balancing process is controlled with sub-micron sensitivity by a photoelectric position sensor (f).

3. Automatic and Semiautomatic Calibration

The electronic signal-processing part of a balance is comparable to a digital voltmeter, except that the display is calibrated in units of mass.

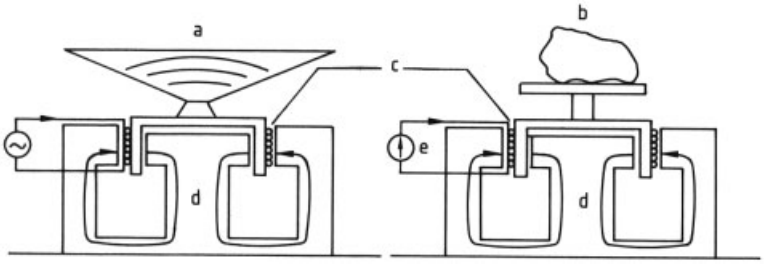

Figure 6. Analogy between a loudspeaker and a laboratory balance, as explained in the text
a) Diaphragm; b) Load; c) Coils; d) Magnets; e) Feedback circuit

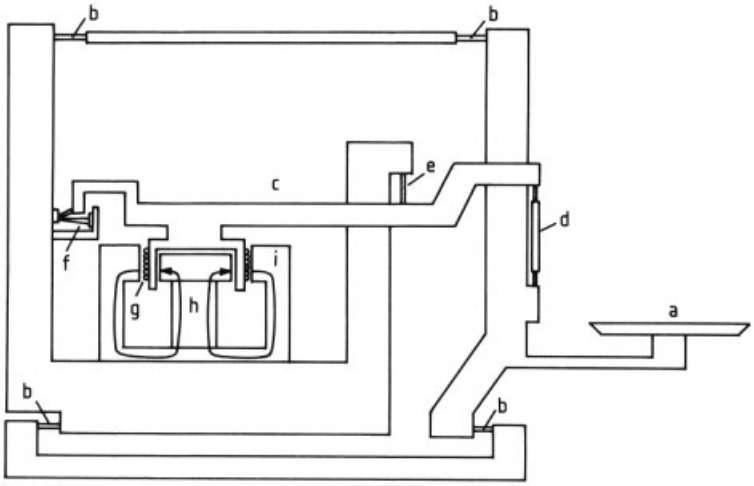

Figure 7. Schematic cross-section through an analytical balance
a) Weighing pan, cantilevered from a parallel-motion linkage; b) Flexure pivots of the parallel-motion linkage; c) Main lever or balance beam; d) Coupling between linkage and beam; e) Main pivot or fulcrum; f) Photoelectric position sensor; g) Force coil; h) Cylindrical permanent magnet; i) Cylindrical soft-iron shell, channeling the magnetic flux

The *accuracy* (correspondence between a displayed value and the true mass on the balance) of a modern professional-grade balance is ensured by automatic or semiautomatic self-calibration. In a self-calibration cycle, a reference mass is either deposited on the balance internally or set on the pan manually by the operator, after which the balance's microprocessor updates the value of a calibration factor stored in a nonvolatile (power-independent) memory. This calibration function represents a direct relationship between the weighing object and the displayed value, in that the weighing process is referenced directly to a mass. A more sophisticated form of automatic calibration employs two built-in reference masses to assure both calibration at full capacity and linearity at midrange of the balance (see Figs. 1 and 2).

4. Processing and Computing Functions

Thanks to the capabilities of microprocessors, a number of special computation routines are common to most balances:

1) Pushbutton tare for subtracting container weights
2) Automatic zero tracking to keep the display zeroed when no weight is on the pan
3) Vibration filtering to make displayed values more immune to disturbances
4) Stability detection to prevent premature reading or transfer of transient or fluctuating results

In addition to these signal processing features, some balances also provide useful conve-

nience options such as conversion to nonmetric units, parts counting, weight statistics, percentage calculation, and animal weighing.

5. Balance Performance

A weighing instrument is commonly tested by loading it with calibrated mass standards and then ascertaining whether the displayed weight values are accurate, meaning that each result corresponds to the value of the test standard within some predetermined tolerance limits. In the most elementary check of balance accuracy, a single test mass is placed once on the balance. This is a test that is often performed for extra assurance in critical applications.

A complete balance test consists of a systematic set of weighings designed to evaluate all aspects of balance performance, including repeatability, eccentric loading errors, linearity, and span calibration. For a balance to be declared in proper working condition, all test results must fall within the tolerance limits specified by the manufacturer. These tolerances normally apply only to new or newly serviced equipment. In other cases a so-called in-service tolerance – equal to twice the original tolerance – is customarily allowed.

Repeatability for successive weighings of the same load is calculated as standard deviation from ten weighings of the same mass, where the balance is reset to zero as necessary before each weighing. The procedure is usually conducted with a mass of approximately one-half the weighing capacity of the balance. This test measures short-term fluctuations and noise of the type present to some degree in every electronic instrument. Excessive standard deviation may be a result of adverse environmental conditions, such as building vibrations or air drafts. A factor equal to three times the standard deviation is customarily called the *uncertainty* of the balance, meaning that virtually all observed values (theoretically, 99.7 %) would fall within this limit on either side of the average.

In the *eccentric load test* a weight equal to one-half the capacity of the balance is placed at the center of the weighing pan and the balance is reset to zero. The weight is then moved, in turn, half-way to the left, right, front, and rear of the pan. Weight readings from the four eccentric positions, if different from zero, are called eccentric loading errors. If not specified separately, the tolerance limit for such errors is of the same order of magnitude as that for nonlinearity. Larger errors indicate that the mechanical geometry of the balance is out of adjustment, usually as a result of abusive treatment. A correctly adjusted balance is practically insensitive to eccentric placement of the load, but any residual errors will be minimized if samples and containers are approximately centered on the balance pan.

Nonlinearity refers to deviations from a mathematically straight line in a graph of display readings versus actual weights, where the line is drawn from the zero point to the full-capacity endpoint. Linearity errors in a balance are caused by practical imperfections in the theoretically linear scheme of electromagnetic force compensation. Nonlinearity manifests itself as a variation in results when the same object is weighed together with differing amounts of tare, and this is, in fact, how the nonlinearity of a balance is tested. A weight equal to one-half the capacity is weighed once without tare, then with a tare of the same magnitude as the weight itself. The mid-range linearity error of the balance is equal to one-half the difference between the two results. This test detects the most common type of balance nonlinearity, an error curve of parabolic shape with maximum positive or negative error at mid range. Higher-order components in the error curve can be detected with four weighings of a test weight equivalent to one-fourth the capacity, using tares equal to 0, 25, 50, and 75 % of the balance capacity. The advantage of such a procedure is that it does not require calibrated weights; the test weight as well as the tares represent nominal values only.

The *calibration error*, also known as sensitivity error or span error, is a small relative factor (a constant percentage of the applied load) by which all weight readings may be biased. For a balance with a self-calibration device this error is limited to the error tolerance of the reference weight, which is of the order of 0.0001 % (1 ppm) in the case of an analytical balance. If the balance lacks self-calibration, calibration errors may develop:

1) With large changes in room temperature
2) When the balance is moved to a different location

3) As a long-term aging effect in the electromagnetic force sensor

Calibration errors should be ascertained with a certified test weight of substantially lower uncertainty than the calibration tolerance of the balance (insofar as this is possible or necessary in practice). In the finest analytical balances, the achievable calibration-error tolerance is dictated by the accuracy of the available mass standards. However, this limitation is not relevant at the level of 0.1 % to 0.01 % accuracy applicable to most laboratory work.

The temperature dependence of the calibration error (temperature drift) is tested in new balances at the factory. Test weighings are made in an environmental test chamber at various temperatures (e.g., 10, 20, and 30 °C). A typical tolerance for the span drift (sensitivity drift) as a function of temperature for an analytical balance is 1 ppm/ °C, also expressed as 10^{-6}/ °C. With a test weight of 100 g, this means that a correct reading of 100.0000 g at 20 °C could change to 99.9999 g or 100.0001 g (or anywhere in between) if the temperature were to change to either 19 or 21 °C.

6. Fitness of a Balance for Its Application

To select an appropriate balance for a given application, one needs to answer two basic questions:

1) How heavy are the samples to be investigated?
2) How accurately must they be weighed?

For example, if the accuracy objective is 0.1 %, then a balance should be selected for which the sample weights would represent at least a thousand times the value of the uncertainty, or three thousand times the standard deviation. Adherence to this rule may be encouraged by placing an appropriate label on the balance, such as "Minimum net sample weight 60 milligrams" for a balance with a standard deviation of 0.02 mg. To weigh smaller samples, one would select a balance with a smaller standard deviation. This requirement of 0.1 % accuracy and its interpretation in terms of the standard deviation are mandated for pharmaceutical weighing according to the *United States Pharmacopeia* [2].

7. Gravity and Air Buoyancy

All weighing relies ultimately on the concept of equilibrium with a symmetrical lever (Fig. 8). A *weight*, in the common usage of the term, is represented by the total value of calibrated weighmasses required to put the lever in balance. According to international convention, the reference masses are to be fabricated from stainless steel with a density of 8 g/cm^3 [3]. The forces in equilibrium arise primarily from the downward pull of gravity on the two sides of the balance, but small buoyant forces from the surrounding air must also be considered. In Figure 8, the gravity and buoyancy forces are symbolized by arrows. The equation under the diagram expresses the equilibrium in mathematical terms. Solving the equilibrium equation for *m* leads to the conversion from weight to mass.

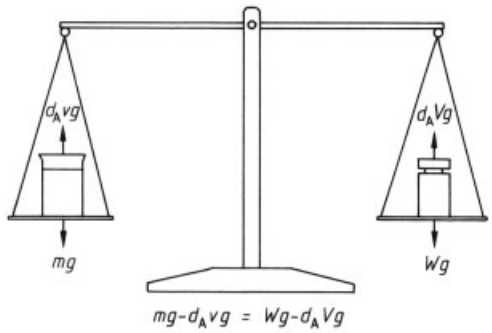

Figure 8. Gravity and buoyancy forces acting on a two-pan balance
v = sample volume. V = volume of the mass standard. g = acceleration due to gravity.

$$m = W \cdot (1 - d_A/D) / (1 - d_A/d) = W \cdot k \quad (1)$$

m	sample mass, expressed in the same metric unit as that applicable to W
W	sample weight, defined in terms of the counterbalancing steel mass
d_A	density of ambient air, g/cm^3
D	density of the mass standard, normally 8 g/cm^3
d	density of the sample, g/cm^3
k	weight-to-mass conversion ratio for the sample

According to this expression, there are two cases in which sample mass equals sample weight ($k = 1$): at zero air density (vacuum), or if the sample density d equals the density D of the mass standard.

Typical values for air density are approximately 1.2 mg/cm^3 at sea level and 1.0 mg/cm^3 in Denver, Colorado (altitude 1600 m). The air density may also vary by as much as 6 % with changes in the weather. Given that the ambient air density is variable, it follows from Equation (1) that the same mass will produce slightly different weights at different times and places, as illustrated in Table 1. This variation is too small to be relevant in commercial applications, however, and the same is true for most laboratory weighings as well.

The conclusions derived above for a two-pan balance also apply to modern electronic balances. Weighing still entails comparison of the test object or sample with a calibrated steel mass. However, in the case of a single-pan electronic balance, the standard is placed on the pan at a particular point in time in order to calibrate the balance, and the weighing process occurs on the same pan at a subsequent time. Thus, weighing on an electronic balance involves a sequential rather than simultaneous comparison, which has one important implication: an electronic balance is the functional equivalent of a classical two-pan balance only if the surrounding gravity field and atmosphere are the same during weighing as when the balance was calibrated. Therefore, electronic balances must always be calibrated at the place where they are to be used, particularly since the force of gravity varies by more than 0.1 % over a territory as limited as the continental United States.

8. The Distinction Between Mass and Weight

Weight, in the primary definition of the term, means the direct result of a weighing, without any correction for air buoyancy. Mass, in the context of weighing, is understood as an absolute, buoyancy-corrected quantity. A problem arises if a weighing result is reported as a mass, when in fact no buoyancy correction has been made. The discrepancy amounts to about 0.1 % for a material with a density of 1 g/cm^3. In science there is a general desire to avoid the term "weight" in the sense of a weighing result, because scientific nomenclature treats weight as a *force*, properly expressed in dynes or newtons, whereas the gram is a unit of mass. To prevent misunderstandings, if weighing results are reported as "mass" then the context should make it clear whether or not stated values have been subjected to an air-buoyancy correction.

9. Qualitative Factors in Weighing

A proper environment and sufficient technical skill on the part of the operator are critical for achieving satisfactory weighing results. Precision balances weighing to 0.1 g perform well in almost any situation, but with analytical and microbalances a number of factors can be detrimental to accurate weighing, including air drafts, direct sunlight and other forms of radiant heat, building vibrations, and electrostatic and magnetic forces. A stable indoor climate is generally sufficient, but balances should be kept away from doors, windows, ovens, and heat and air conditioning outlets, and they should never be placed under ceiling outlets. A mid-range level of relative humidity is best, since this will prevent the buildup of electrostatic charge from excessively dry air as well as moisture absorption by samples and containers from air that is too humid. The balance table should rest on a solid foundation so the balance will not be affected if the operator moves or leans on the table. There should be no elevators or vibrating machinery nearby. An operator may find it necessary to conduct several experiments in a search for appropriate procedures when weighing hygroscopic, volatile, or ferromagnetic samples. Samples and containers should always be acclimated to room temperature to prevent air convection around the sample that might interfere with weighing.

10. Governmental Regulations and Standardization

Weighing is subject to governmental regulation in situations where errors or fraud might result in material or health hazards. Weighing in trade and

Table 1. Weights at sea level (W_S) and in Denver, Colorado (W_D), for samples with a mass of one gram and various densities d

Sample density d, g/cm^3	Weight at sea level W_S, g	Weight in Denver W_D, g	Difference $(W_D - W_S)/W_S$, %
0.5	0.997749	0.998124	0.0376
1.0	0.998949	0.999124	0.0175
2.0	0.999549	0.999624	0.0075
4.0	0.999849	0.999874	0.0025
8.0	1.000000	1.000000	0.0000
20.0	1.000090	1.000074	−0.0017

commerce is commonly supervised by a governmental "Weights and Measures" agency through some system of certification and inspection of scales and weights [4], [5]. In areas other than commerce, the nature and scope of government control over weighing varies depending upon the field, as in the case of pharmaceuticals [2], nuclear materials [6], or military defense contracts [7]. In the United States, critical weighing applications are subject to a system of government-supervised self-control. Any organization conducting weighing processes subject to regulatory control must have adopted formal mass assurance procedures as part of an overall quality assurance system. The government exercises its oversight function through periodic audits. The main elements of a typical mass assurance program are as follows:

1) Balances available to the staff must conform to the accuracy required for the intended use
2) Balances must be maintained and calibrated on a regular cycle, with records kept on file
3) All balance operators must be instructed in correct weighing technique, and training records must be kept on file
4) Formal and documented weighing procedures must be in place
5) Orderly records must be maintained for all weighings that fall within the scope of the regulated activity

The most widely accepted guideline for establishing a system of mass assurance is found in International Standard ISO 9001, Section 4.11., "Inspection, Measuring and Test Equipment" [8]. Under ISO 9001, the quality system of an organization is to be audited and certified by independent quality assessors, analogous to the audits performed by the government on regulated organizations.

11. References

1. B. Kisch: *Scales and Weights: A Historical Outline,* Yale University Press, New Haven 1965.
2. *The United States Pharmacopeia,* 22nd ed., United States Pharmacopeia Convention, Inc., Rockville, Md., 1990.
3. OIML Recommendation R-33, Conventional Value of the Result of Weighing in Air, Bureau International de Métrologie Légale, Paris, France, 1973.
4. OIML International Recommendation R 76-1, Non-automatic Weighing Instruments. Part 1: Metrological and Technical Requirements – Tests, Bureau International de Métrologie Légale, Paris, France, 1993.
5. NIST Handbook 44, U.S. Dept. of Commerce, National Institute of Standards and Technology, Gaithersburg, Md., 1993.
6. MIL-STD-45662A, Calibration Systems Requirements, Department of Defense, Washington, DC 20360, 1988.
7. Nuclear Regulatory Commission, Codes 10 CFR 21 and 10 CFR 50, 1989.
8. ISO 9001,Quality Systems – Model for Quality Assurance in Design/Development, Production, Installation and Servicing, 1994.

Weld Cladding → Corrosion

Chemometrics

René Henrion, Weierstrass Institute of Applied Analysis and Stochastics, Berlin, Germany
Günther Henrion, Humboldt University Berlin, Berlin, Germany

1.	**Introduction**	2
2.	**Measurements and Statistical Distributions**	2
2.1.	**Measurements**	2
2.2.	**Statistical Distributions**	2
2.3.	**Estimates**	3
2.4.	**Accuracy and Precision**	4
3.	**Statistical Tests**	4
3.1.	**General Procedure**	5
3.2.	**Tests on Parameters of One or Two Measurement Series**	6
3.3.	**Outliers, Trend and Nonparametric Tests**	8
4.	**Comparison of Several Measurement Series**	8
4.1.	**Homogeneity of Variances**	9
4.2.	**Equality of Expected Values**	9
5.	**Regression and Calibration**	10
5.1.	**Regression Analysis**	10
5.2.	**Calibration**	11
6.	**Characterization of Analytical Procedures**	12
7.	**Signal Processing**	14
7.1.	**Fourier Transform**	14
7.2.	**Data Smoothing**	15
7.3.	**Signal Resolution**	15
8.	**Basic Concepts of Multivariate Methods**	16
8.1.	**Objects, Variables, and Data Sets**	16
8.2.	**Correlation and Distance Matrices**	17
8.3.	**Data Scaling**	18
9.	**Factorial Methods**	18
9.1.	**Principal Components Analysis**	18
9.2.	**Factor Analysis**	20
10.	**Classification Methods**	21
10.1.	**Cluster Analysis**	21
10.2.	**Supervised Classification**	22
11.	**Multivariate Regression**	24
11.1.	**Multiple Linear Regression**	24
11.2.	**Latent Variable Regression**	24
12.	**Multidimensional Arrays**	25
13.	**References**	26

Abstract

The article provides an introduction to some of the basic concepts and methods in chemometrics. The first part is devoted to classical univariate statistics with emphasis on estimation and comparison of statistical parameters from one or several measurement series and its application to measurements from chemistry. The second part presents special techniques of data evaluation in analytical chemistry such as calibration, signal processing, and characterizations of analytical procedures. The third and last part introduces modern concepts of multivariate statistics (including multiway data analysis) with applications to chemometrical data sets. Numerical examples serve as illustrations for the presented methods.

Symbols

b	sensitivity ($=dy/dx$)
d	distance
$\boldsymbol{D}$	distance matrix
n	number of measurements (or objects, Chap. 8)
n_{j}	number of parallel measurements
P	probability
r	correlation coefficient
$\boldsymbol{R}$	correlation matrix
s	estimate of standard deviation (sdv)
$t\,(P, f\,)$	integration limit for Student's distribution
$\boldsymbol{v}_j$	eigenvectors
x	independent variable (mostly: concentration)
x_{d}	decision limit

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_035.pub2

x_D	detection limit
x_i	i-th value (predicting variable, Chap. 11)
$\bar{x}$	estimate of mean value
$\boldsymbol{X}$	matrix
$\boldsymbol{X}^T$	transpose of the matrix
y	dependent variable (mostly: signal)
λ_j	eigenvalues
μ	mean value
σ	standard deviation

1. Introduction

Using a widely accepted definition, *chemometrics* can be understood as "the chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analyzing chemical data" [1]. Defined in this way, chemometrics shares the impact of mathematical modeling with many other disciplines carrying the suffix "metrics", such as biometrics, psychometrics, and econometrics. Nevertheless, the application to problems of chemistry puts emphasis on particular issues not present or less important in these other sciences. For example, keywords like calibration and signal resolution can be mentioned. However, chemometrics is not just the recipient of mathematical progress, it also stimulates the development and the foundation of new mathematical methods. For instance, the use of modern hyphenated analytical methods producing huge amounts of data with increasing complexity of structure has become a driving force for algorithms in multimodal statistics. Over time, chemometric methods have become indispensable tools, e.g., for quality control, environmental and forensic analysis, medicine, and process control. Chemometrics, however, does not only interpret data or find optimal strategies for analytical work. It also provides a basic for theoretical treatment of analytical chemistry and establishes the field of analytical science as an independent discipline of chemistry. For comprehensive introductions to the field of chemometrics, the reader is referred to the basic monographs [1] and [2].

2. Measurements and Statistical Distributions

2.1. Measurements

One of the final aims of chemical analysis is the quantitative determination of species concentrations based on appropriate measurements. Usually these measurements exploit a functional relationship between concentrations and a physical (optical, electrochemical, etc.) signal. For a given sample under investigation, the obtained signal will never be constant when repeating the measurement. The reason is that a variety of conditions, such as temperature, conductivity, pH, change slightly over time and their superimposed impact on the measurement is out of control in the sense of a functional model. Therefore, it is useful to understand a measurement as a random variable x which can have many different outcomes or realizations x_i. In a fixed experiment, the set of realizations $x_1,\ldots,x_N$ is called a sample of x with sample size N. Usually, a sample is characterized by its mean value $\bar{x}$ and its empirical standard deviation s:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i;\ s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Example 2.1. The analysis of an element concentration by repeated (N = 16) measurements of one laboratory gave the following results:

30.3, 29.9, 32.0, 32.0, 30.0, 31.0, 30.9, 30.3, 30.9, 30.1, 29.4, 29.6, 30.4, 30.9, 28.0, 29.6.

$N = 16$, $\bar{x} = 30.33$, $s = 0.99$.

2.2. Statistical Distributions

Typically, the outcomes of random variables are not arranged uniformly but show regions of differing density. The points corresponding to the measurements in Example 2.1 are arranged on a straight line in Fig. 1 a. Their density can be illustrated by means of so-called frequency histograms, where the range of measurements is subdivided into an appropriate number of subintervals of equal length. The number of measurements x_i falling inside a given interval is represented by a rectangle of corresponding height on

top of the interval. With increasing sample size N, the histograms take on a more regular shape, which is often well approximated by a smooth curve $f(t)$, as in Figure 1 b, where f is the density function of x. Clearly, the realizations of x are most likely to fall inside regions with high density values of f. More precisely, the probability of an outcome of x falling into a set A is given by the integral

$$P(A) = \int_A f(t)\,\mathrm{d}t \tag{2.1}$$

Of particular interest are sets given by left open intervals: $A = (-\infty, \tau)$. In this case, τ is called a P(A) quantile of the distribution of x, which means that the probability for an outcome of x not to exceed τ equals P(A). The density function is always normalized so as to render the integral over the whole real line equal to one. In this way, a density function completely determines the distribution of a random variable. The most prominent representative of probability distributions is the normal or Gaussian distribution with density function

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$$

The bell-shaped profile of this function is shown in Figure 1 c. The parameters μ and σ determine the position of the maximum of f and the distance of the inflection points of f from this maximum. Actually, μ corresponds to the expected value E(x), and σ to the standard deviation σ(x) of the normal distribution. For general distributions with densities, these quantities are defined as

$$E(x) = \int f(t)\,t\mathrm{d}t$$
$$\sigma(x) = \sqrt{\int f(t)\,[t - E(x)]^2 \mathrm{d}t}$$

The square of the standard deviation $\sigma^2(x)$ is referred to as the variance of x. The expected value, in some sense, represents the most typical outcome of a random variable, whereas the variance measures the average squared deviation of outcomes from the expected value. For the normal distribution, the probability of a realization falling inside the interval $\mu \pm \sigma$ — that is, closer to the expected value than one standard deviation — is 68.3 %. According to Equation (2.1), this value corresponds to the area portion under the density function for this interval (Fig. 1 c). Taking the larger interval $\mu \pm 3\,\sigma$ gives a probability of 99.7 %. This justifies the 3 σ rule, which states that realizations of normally distributed random variables almost surely fall inside this interval.

The importance of the normal distribution relies on the so-called central limit theorem, which roughly states that the distribution of the sum of a large number of independent random variables tends to be a normal one. In particular, measurements which are influenced by a large number of small independent errors are well approximated by a normal distribution with appropriate parameters μ and σ. Nevertheless, other random variables may follow distributions different from the normal one. Counting methods (e.g., X-ray fluorescence) must be described by a Poisson distribution, whereas concentrations in trace analysis are better modeled by a logarithmically normal distribution, that is, the logarithm of the measurements is normally distributed. The latter fact is intuitively clear, since concentrations cannot have negative values, so the density function should be zero on the negative axis (Fig. 1 d). The log-normal distribution also shows that the expected value does not coincide, in general, with the maximum of the density. Other important statistical distributions arise from transformations of normally distributed random variables, e.g., the t, F and χ^2 distributions.

2.3. Estimates

It is important to distinguish between parameters of statistical distributions such as density function, expected value, and standard deviation on the one hand, and corresponding sample characteristics such as histograms, mean value and empirical standard deviation on the other. The latter are calculated from concrete samples and will converge — in a stochastic sense — with increasing sample size towards the distribution parameters which are the usually unknown characteristics of the abstract probability distribution behind one and the same experiment. The sample-based quantities are also called estimates of the corresponding statistical parameters.

Estimates are random variables themselves. For instance, repeating the same experiment as in Example 2.1., one would arrive at a different

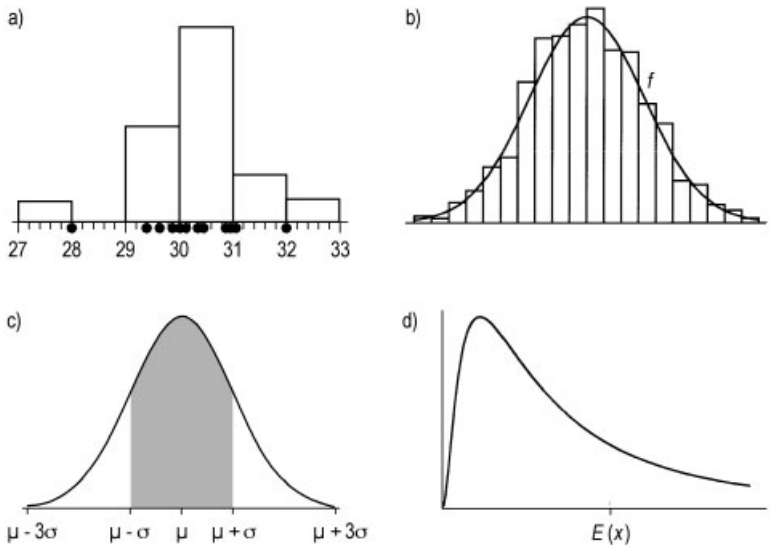

Figure 1. Illustration of densities for probability distributions.
a) Histogram for a small concrete measurement series; b) Histogram for a large series and limiting density function; c) Normal density function with symmetric interval of 68.3 % probability around the mean; d) Density and expected value for the logarithmic normal distribution

measurement series with slightly changed values of $\bar{x}$ and s. Repeating this experiment many times, one arrives at a distribution for $\bar{x}$ and s similar to that obtained before for x. Not surprisingly, the expected value of $\bar{x}$ stays the same as that of x: $E(\bar{x}) = E(x)$. The standard deviation, however, decreases by a factor of $\sqrt{N}$, where N is the sample size: $\sigma(\bar{x}) = \sigma(x)/\sqrt{N}$. Hence, the distribution of $\bar{x}$ is much narrower than that of x, (Fig. 2). Increasing the sample size of some measurement series by a factor of nine will therefore decrease the standard deviation of the associated mean value by a factor of three.

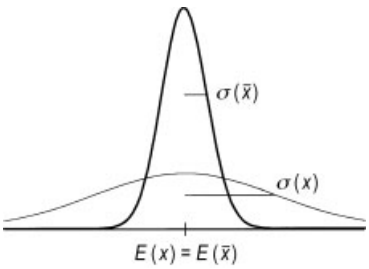

Figure 2. Densities for the normal distribution of a measurements x and of associated mean value $\bar{x}$

2.4. Accuracy and Precision

Assume that c is the true value of some analytical quantity to be determined. For a given measurement series as in Example 2.1, the deviation $|\bar{x} - c|$ characterizes its accuracy. We already know that $\bar{x} \to E(x)$ for $N \to \infty$. In the ideal case, $E(x) = c$; hence, the accuracy will become perfect with increasing sample size. Otherwise, the determination has some bias $|E(x) - c| > 0$, and the accuracy can only improve up to the value of this bias. A statistical test for presence of bias is discussed in Section 3.2. The concept of precision refers to the scattering of measurements around the true value. In the unbiased case, precision is simply given by the (empirical) standard deviation of a measurement series. As long as precision refers to the results of a single laboratory, it is identified with the concept of *repeatability*. Precision of results provided by different laboratories is identified with the concept of *reproducibility*. For a quantification of both concepts, see Chapter 4.

3. Statistical Tests

Statistical tests serve the purpose of verifying hypotheses on parameters of distributions. As

these parameters are usually unknown, the tests have to rely on estimates which are random variables, hence their outcome has to be interpreted in a probabilistic way.

3.1. General Procedure

The general procedure of a test may roughly be sketched as follows:

1) Formulate a "null hypothesis" H_0 usually involving an equality statment about a parameter of interest.
2) Fix a significance level P (usually $P = 95\,\%$ or $P = 99\,\%$).
3) Choose an appropriate "test statistic" T which is a random variable depending on the measurement sample and has a well-known distribution under validity of H_0.
4) Select some set A such that the probability of T having values in A equals P.
5) For a given sample, check if the value of T falls outside A. If yes, then reject H_0; otherwise, accept H_0.

Note that, according to this scheme, the probability of rejecting H_0 although H_0 holds true (error of the first kind) is less than $1 - P$ (say 5 % or 1 %), which in this case gives the probability of the observed fact that T falls outside A. However, nothing is said about the probability of accepting H_0 although H_0 is false (error of the second kind) if we observe that T takes a value in A. The realization of step 4 requires knowledge of quantiles for the distribution of T. Because of limited space, the necessary values for the different tests are not given here; reference is made to standard monographs on applied statistics or handbooks containing the corresponding data tables and also giving a more detailed introduction to statistical theory (e.g., [3, 4]). All tests presented in the following are based on normally distributed data. If this assumption, which should be satisfied in most cases, fails to hold, then one can use nonparametric tests as an alternative (see Section 3.3).

Test for Mean with Known Standard Deviation. As an illustration, consider a test for bias in laboratory results. Assume that a measurement series of size N is given. As null hypothesis, we formulate coincidence of the expected value behind the measurements with some known true value c:

$$H_0{:}E\,(x) = c$$

A significance level of $P = 95\,\%$ is fixed and the test statistic $T = \bar{x}$ is chosen. From Section 2.3, we know that, $E(T) = E(x)$ and $\sigma\,(T) = \sigma\,(x)\,/\sqrt{N}$. Furthermore, T is normally distributed whenever x is (which one has good reason to assume). It is convenient to select the set A, as required in step 4, in a symmetric way around the expected value (see shaded area in Fig. 1 c). For normal distributions, we already know that the intervals $[\mu - \sigma,\ \mu + \sigma]$ and $[\mu - 3\,\sigma,\ \mu + 3\,\sigma]$ have probabilities of 68.3 and 99.7 %, respectively. To realize the chosen probability of $P = 95\,\%$, one would have to consider the interval $[\mu - 1.96\,\sigma,\ \mu + 1.96\,\sigma]$, where the value of 1.96 can be read off from appropriate data tables. Since we have to relate the parameters μ and σ here to the expected value and standard deviation of the distribution of $\bar{x}$ rather than that of x, the appropriate choice is (taking into account that $E(T) = E(x) = c$ for validity of H_0)

$$A = \left[c{-}1.96\sigma\,(x)\,/\sqrt{N}, c{+}1.96\sigma\,(x)\,/\sqrt{N}\right] \qquad (3.1)$$

Now a decision according to step 5 can be made on the basis of a concrete sample and a concrete resulting value for T. Note, however, that the described procedure requires knowledge of the distribution parameter $\sigma(x)$, which is rarely available. A related test avoiding this assumption is described in Section 3.2.

Example 3.1. For the data of Example 2.1., it shall be checked if there is some bias to an assumed true value of $c = 30$. We also assume that the standard deviation of the (abstract) measurement distribution is known: $\sigma(x) = 0.7$ (e.g., $\sigma(x)$ may have been approximated by empirical standard deviations s on the basis of a large sample size). The test interval calculates as $A = 30 \pm 1.96 \times 0.7/4 = 30 \pm 0.343$. Since the mean value $T = \bar{x} = 30.33$ is contained in this interval, we accept H_0 which means that — up to unavoidable random errors — the laboratory meets the true value (no bias). Recall, that nothing is said about the risk of a false statement here. If, in contrast, the true value were $c = 29.5$, then T is not in A, hence we would reject H_0 and deduce the presence of some bias with a probability of error below $1 - P = 5\,\%$.

Confidence Intervals. The preceding derivations have shown that the mean value $\bar{x}$ belongs to the interval $E(x) \pm 1.96\sigma(x)/\sqrt{N}$ with probability 95 %. But, equivalently, $E(x)$ belongs to the interval $\bar{x} \pm 1.96\sigma(x)/\sqrt{N}$ with the same probability. Hence, given a mean value from a concrete sample, the unknown distribution parameter $E(x)$ can be included in a symmetric interval around $\bar{x}$ with a given probability. This is called a P % confidence interval for the statistical parameter. For the data of Example 2.2. one obtains the 95 % confidence interval 30.33 ± 0.343. This is the usual way of indicating the result for an analytical concentration etc. For higher probabilities, the interval enlarges accordingly. Choosing, for instance the 99 % level, one would have to replace the above factor of 1.96 by 2.58 yielding the interval 30.33 ± 0.452.

One-Sided Tests. Frequently, some additional prior information on the considered statistical parameter is available. There may be, for instance, some danger of bias to the true concentration value as a consequence of sample preparation. From the logical background it might be clear, for instance, that the concentration of the analyte is increased if at all. Then, in the above terminology, we know that $E(x) \geq c$. Endowed with such one-sided information, it is reasonable to test the null hypothesis formulated above with a set A chosen as a one-sided open interval rather than using the symmetric interval as before. Indeed, here we do not expect our mean value to be much below c, and it is the large values of the mean which are of potential interest now. In this way, we gain information on the crucial side of the inequality to be tested (here: $E(x) > c$) since the test set A is shifted to the left, and this makes it easier to detect excessive deviations on the right. Figure 3 shows that, although both intervals represent 95 % probability, a specific mean value $\bar{x}$ may fall outside the one-sided interval whereas it is inside the two-sided interval. This gives a gain in decision power when exploiting one-sided information. Therefore, we fix A in the form $(-\infty, \tau]$, where τ is chosen such that $P(A) = 95$ %. Hence, τ has to be selected as the 95 % quantile (see Section 2.2) of the normal distribution with parameters $\mu = E(x)$ and $\sigma = \sigma(x)$. This can be read off from data tables as the value $\tau = 1.65$. Accordingly, for the data of Example 3.1, the mean value now falls outside the interval $A = 30 \pm 1.65 \times 0.7/4 = 30 \pm 0.289$. That is why, in contrast to Example 3.1, the null hypothesis now has to be rejected, and it can be stated that there is a 95 % significant overestimation in the determinations as compared to the true value. The opposite case of knowing that $E(x) \leq c$ is treated analogously by exploiting the symmetry of the normal distribution.

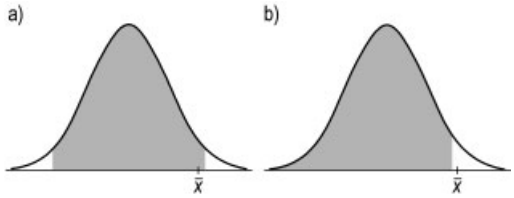

Figure 3. Symmetric (a) and one-sided (b) intervals around the expected value, both having equal probability

3.2. Tests on Parameters of One or Two Measurement Series

Table 1 lists tests on statistical parameters of one or two measurement series. The test intervals are formulated for two-sided tests with probability $P = 95$ %. The occuring quantities refer to quantiles of corresponding distributions (F, t, χ^2) which can be found in data tables (e.g., [4]). Indices involving sample sizes refer to so-called degrees of freedom of the corresponding distribution. For instance, $t_{N-1;0.975}$ refers to the 97.5 % quantile of the t- distribution with N − 1 degrees of freedom (note that in the literature quantiles are frequently indicated in the complementary way as critical values, so that 97.5 % correspond to 2.5 %, and vice versa). A two-sided symmetric interval of probability 95 % would then be given by $[t_{N-1;0.025}, t_{N-1;0.975}]$ (since 97.5 % – 2.5 % = 95 %). Owing to the symmetry of the t distribution, $t_{N-1;0.025} = -t_{N-1;0.975}$, whence the interval as indicated in Table 1. There is no problem in changing to other probability levels or one-sided tests (see examples below).

Example 3.2. Using the data of Example 2.1, we want to test the hypotheses $E(x) = 30.8$ and $\sigma(x) = 2$. The 97.5 % quantile of the t-distribution with 15 degrees of freedom equals 2.13, hence $A = [-2.13, 2.13]$. The test statis-

Table 1. Sample information, null hypothesis, test statistic, and test interval for common tests on parameters of one or two measurement series

Test	Sample data	H_0	T	A (two-sided, 95 %)
Comparison with fixed standard deviation	s, N	$\sigma(x) = \sigma_0$	$(N-1)s^2/\sigma_0^2$	$\left[\chi^2_{N-1;0.025}, \chi^2_{N-1;0.975}\right]$
Comparison with fixed expected value	$\bar{x}, s, N$	$E(x) = \mu_0$	$(\overline{x} - \mu_0)\sqrt{N}/s$	$[-t_{N-1;0.975}, t_{N-1;0.975}]$
Comparison of two standard deviations	s_1, N_1, s_2, N_2 $(s_1 \geq s_2)$	$\sigma(x_1) = \sigma(x_2)$	$\frac{s_1^2}{s_2^2}$	$[0, F_{N_1-1,N_2-1;0.975}]$
Comparison of two expected values $\sigma(x_1) = \sigma(x_2)$	$\bar{x}_1, s_1, N_1$ $\bar{x}_2, s_2, N_2$	$E(x_1) = E(x_2)$	$\frac{\overline{x}_1 - \overline{x}_2}{s\prime}\sqrt{\frac{N_1 N_2}{N_1+N_2}}$ $s\prime = \sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}$	$[-t_{N-1;0.975}, t_{N-1;0.975}]$

tic calculates for the test on the expected value to $T = (30.33 - 30.8) \times 4/0.99 = -1.89$. Since this value belongs to A, H_0 is accepted and there is no reason to assume a bias in the determinations with respect to the (hypothetical) true result of 30.8. If, however, one knows in advance that the expected value cannot exceed the true one, i.e., $E(x) \leq 30.8$, then A can be chosen as a right-open interval of the same probability $P = 95\,\%$: $A = [t_{N-1;0.05}, \infty) = [-t_{N-1;0.95}, \infty)$ (owing to the symmetry of the t-distribution). For the concrete data with $N - 1 = 15$, one gets $A = [-1.75, \infty)$. Now, the value calculated above falls outside the test interval, hence, with the one-sided extra information one can deduce that there is a negative bias to 30.8 in the determinations of the laboratory at the probability level of 95 %. Note that we have just dealt with a one-sided test in the opposite direction to that considered in Section 3.1. Another important difference to that simpler test is that now the (theoretical) standard deviation need not be known. Instead, its empirical estimate can be used. As for the test on standard deviation, one obtains $A = [6.26, 27.5]$ and $T = 3.68$; hence, the standard deviation behind the given measurement series must be considered different at the 95 % level from the value of 2 which was to be tested. For one-sided tests with additional a priori information that $\sigma(x) \leq 2$ or $\sigma(x) \geq 2$, one would fix the test intervals as $\left[\chi^2_{N-1;0.05}, \infty\right)$ and $(-\infty, \chi^2_{N-1;0.95}]$, respectively. Finally, with the given data, confidence intervals for the statistical parameters are easily constructed along the same line as described in Sect ion 3.1. For instance, at the 95 % level, one obtains the interval $\bar{x} \pm 2.13$ for covering $E(x)$.

Frequently, it is not the statistical parameters of a single measurement series which have to be tested against specific values, but rather different series are to be tested against each other. For instance, one may ask if two laboratories participating in the same experiment provide equal accuracy and/or precision. This can be answered by the last two tests recorded in Table 1.

Example 3.3. Consider the measurement data of Example 4.1 below. It shall be tested whether the expected values and standard deviations of the five measurements provided by laboratories A and B are different. Starting with the test on standard deviations, the measurement series with the higher empirical standard deviation is set to be the first one (with respect to index) by definition. So, the symbols s_1,N_1 in Table 1 refer to lab B and s_2,N_2 to lab A. As there is no a-priori information on how the two standard deviations relate to each other, a two-sided test is performed. Accordingly, using the 97.5 % quantile of the F-distribution with $N_1 - 1 = N_2 - 1 = 4$ degrees of freedom which is 9.60, one calculates $T = 0.09^2/0.08^2 = 1.27$ which is clearly inside the test interval. Consequently, there is no reason to assume differences between the underlying theoretical standard deviations of the two labs. For a one-sided test at the same 95 % probability level, the 97.5 % quantile of the F-distribution would have to be replaced by the 95 % quantile.

As for the comparison of expected values, one has to take into account first that the application of this test requires the theoretical

standard deviations of the two labs to coincide (more sophisticated tests exist in case this assumption is violated). From the foregoing test, we have no reason to doubt about this coincidence for the data of labs A and B. Following the recipe in Table 1, we calculate $s\prime = \sqrt{(4\times0.09^2+4\times0.08^2/8} = 0.085$ and $T = (45.32-45.21)\sqrt{5\times5/(5+5)}/0.085 = 2.04$.

With $t_{4;0.975} = 2.78$, we see that T remains inside the test interval. Hence, we cannot deduce 95 % significant differences in the accuracies of labs A and B.

3.3. Outliers, Trend and Nonparametric Tests

Before extracting statistical characteristics from a measurement series, such as mean value, standard deviation, confidence interval, etc., it has to be checked whether the data are proper with respect to certain criteria. For instance, the sample may contain extremely deviating measurements owing to a gross error (e.g., simply a typing error or improper measurement conditions). Such values are called *outliers* and must be removed form further consideration as they strongly falsify the characteristics of the sample. The same argumentation holds true for the presence of a *trend* in measurements. This might be caused by directed changes of experimental conditions (e.g., temperature). This leads to a continuous shift of the mean value and increase of the empirical standard deviation with growing sample size, so there is no chance of approaching the theoretical parameters. The presence of outliers and trends can be tested as follows: the test statistic $T = d/s$, where s is the empirical standard deviation and d refers to the maximum absolute deviation of a single value x_i in the measurement series from its mean value, is checked against the values recorded in Table 2 (Grubbs test). If T exceeds the tabulated value, then the measurement x_i is considered as an outlier at the 99 % significance level (smaller levels should not be used, in general, for outlier testing). Similarly, the Trend test according to Neumann and Moore calculates the test statistic

$$T = \sum_{i=2}^{N}(x_i - x_{i-1})^2 / \left((N-1)\, s^2\right)$$

If it is smaller than the corresponding value in Table 2, a trend is evident at the 95 % level of probability.

Table 2. Critical values for the Grubbs test and Neumann/Moore test on the presence of outliers and trend, respectively

N	Outlier (99 %)	Trend (95 %)
3	1.16	
4	1.49	0.78
5	1.75	0.82
6	1.94	0.89
7	2.10	0.94
8	2.22	0.98
9	2.32	1.02
10	2.41	1.06
11	2.48	1.10
12	2.55	1.13
13	2.61	1.16
14	2.66	1.18
15	2.71	1.21
16	2.75	1.23

Example 3.4. For the data of Example 2.1., the most outlying measurement with respect to the mean value is $x_{15} = 28.0$, hence $d = 2.33$ and $T = 2.33/0.99 = 2.35$, which is less than the tabulated value 2.75 for $N = 16$. Hence, the series can be considered outlier-free. The trend test statistic for the same data becomes $T = 1.59$ which is smaller than the tabulated value of 1.23. Hence, the series is also free of trends at the 95 % level.

All tests presented so far relied on the assumption of normally distributed measurements. Sometimes, strong deviations from this assumption can lead to incorrect conclusions from the outcome of these tests. If there is doubt about normality of the data (which itself can be tested as well) then it is recommended to apply so-called nonparametric counterparts of these tests. Nonparametric means that these tests are not based on a concrete distribution. In this way, they are robust with respect to deviations from normality, although of course , less efficient in the presence of normality. Typically, nonparametric tests evaluate some ranking of specifically arranged measurements.

4. Comparison of Several Measurement Series

Up to now, we have dealt with statistical characterizations of single measurement series or pairwise comparisons. Frequently, a larger group of

measurement series must be analyzed with respect to variance and mean. In the case of commercial products, for example, quality is tested in different laboratories. The different working conditions in individual laboratories provide an additional source of random error. In such situations, it is of interest to analyze the homogeneity of data with respect to standard deviations and mean values. In case that homogeneity can be assumed, much sharper confidence intervals for the precision and accuracy can be obtained from the pooled data rather than from a single lab's results. In the opposite case, the standard deviations relating to repeatability and reproducibility (see Section 2.4) can be estimated from *analysis of variance*. Testing for homogeneous data is not only useful, however, in the context of interlaboratory comparisons but also for the important question of representative sampling of materials to be analyzed.

4.1. Homogeneity of Variances

Assume that we are given a group of p measurement series with $N_j, \overline{x}_j, s_j$ denoting sample size, mean value and empirical standard deviation of series j. Further we set $N = \sum_{j=1}^{p} N_j$. We want to check whether all series have equal precision, i.e., H_0:

$$\sigma(x_1) = \ldots = \sigma(x_p)$$

This may be realized by the Bartlett test, in which, the following averaged standard deviation within the series is first determined:

$$s_g = \sqrt{\frac{\sum_{j=1}^{p} (N_j - 1)\, s_j^2}{N - p}} \tag{4.1}$$

Now, a test statistic is calculated from the data according to

$$T = \frac{2.306}{c}\left((N-p)\log s_g^2 - \sum_{j=1}^{p}(N_j - 1)\log s_j^2\right)$$

(where "log" = decadic logarithm) and

$$c = 1 + \frac{\left(\sum_{j=1}^{p} \frac{1}{N_j - 1} - \frac{1}{N-p}\right)}{3(p-1)}$$

H_0 will then be rejected at probability level 95 % if $T > \chi^2_{p-1;0.95}$.

Example 4.1. A sample of FeSi was analyzed in $p = 7$ laboratories with the following results (% Si)

Laboratory	A	B	C	D	E	F	G
	45.09	45.20	45.37	45.23	45.40	45.63	44.93
	45.19	45.27	45.45	45.26	45.40	45.65	44.95
	45.22	45.30	45.48	45.31	45.45	45.73	44.95
	45.25	45.40	45.60	45.39	45.60	45.85	45.14
	45.31	45.43	45.62	45.44	45.60	45.85	45.17
$\overline{x}_j$	45.21	45.32	45.50	45.33	45.49	45.74	45.03
s_j	0.08	0.09	0.11	0.09	0.10	0.11	0.12

For the data from this table, one obtains the results

$$p = 7, N_j = 5\,(j = 1, \ldots, p), N = 35, s_g = 0.10,$$
$$c = 1.10, T = 0.64, \chi^2_{6;0.95} = 12.6$$

Consequently, the null hypothesis cannot be rejected at the 95 % level, and one can assume homogeneous variances among the seven different measurement series.

4.2. Equality of Expected Values

The hypothesis for homogeneity of expected values is H_0:

$$E(x_1) = \ldots = E(x_p)$$

This hypothesis is tested by analysis of variance in which — apart from the already determined standard deviation within laboratories (Equation 3) — a so-called standard deviation of laboratory means (standard deviation between laboratories) figures as an important ingredient:

$$s_z = \frac{1}{p-1}\sum_{j=1}^{p} N_j \left(\overline{x}_j - \overline{\overline{x}}\right)^2 \tag{4.2}$$

where

$$\overline{\overline{x}} = \frac{\sum_{j=1}^{p} N_j \overline{x}_j}{N} \tag{4.3}$$

is the weighted mean of all single mean values or, equivalently, the overall mean of all measurements in the data table. H_0 will be rejected at probability level 95 % if $T = s_z^2 / s_g^2 > F_{p-1, N-p; 0.95}$. The reason for using the one-sided test interval here in contrast to the

two-sided one in Section 3.2 is that the theoretical standard deviations σ_g, σ_z behind their empirical estimates s_g, s_z can be shown always to satisfy the one-sided relation $\sigma_g \leq \sigma_z$, where equality holds exactly in the case that H_0 is true. If H_0 can be rejected on the basis of this test, then the standard deviations in Equations (4.1) and (4.2) represent the repeatability and reproducibility, respectively, of the given laboratory data. Otherwise, the expected values can be considered as being equal, and there is no reason to assume laboratory bias in the given determinations. In this case the data material is homogeneous and the overall mean value in Equation (4.3) is taken as an estimate of the true value to be determined by the laboratories. As a mean over means, this overall mean has a much smaller variance than the single means (which in turn have smaller variances than the single measurements). This gives rise to very narrow confidence intervals for the true value and explains why interlaboratory homogeneity comparisons are so valuable in producing reference materials.

Similar to the test on equality of two expected values described in Section 3.2, its generalization to the multiple case requires homogeneity of variances, this means a Bartlett test must be performed first.

Example 4.2. For Example 4.1, the Bartlett test gave no reason to doubt the homogeneity of variances, so the test on multiple equality of expected values can be carried out. For the data of the example, one calculates

$$s_z = 0.51,\ \bar{\bar{x}} = 45.37, T = 26.58, F_{6.28;0.95} = 2.44$$

Consequently, the expected values behind the single measurement series must be considered different at the 95 % level. In other words, there exists a strong laboratory bias which can be characterized by the strongly different values for s_z and s_g.

5. Regression and Calibration

5.1. Regression Analysis

Regression is a tool for modeling a set of N observed data pairs (x_i,y_i) by means of a functional relationship $y = f(x)$, where f belongs to a specified family of functions (e.g., linear functions, polynomials, bi-exponential functions, etc.). Usually, it is assumed that x is an independent variable which can be fixed without errors, whereas the variable y is dependent on x and is subject to random errors. In analytical chemistry, the role of y is frequently played by measured signals responding to some concentration x or recorded at some wavelength x. The main applications of regression in analytical chemistry are calibration (see Section 5.2) and signal resolution. Figure 4 a illustrates a calibration problem, while Figure 4 c shows the decomposition of a signal profile into a sum of two Gaussian profiles.

The family to which f is supposed to belong to is defined via some set of parameters which are coupled in a particular way with x. Therefore, it is reasonable to extend the above mentioned functional relationship to $y = f(x,p)$, where the parameter p is variable. Some prominent families of regression models are:

1) Linear model $f(x,p) = p_1 + p_2 x$
2) Polynominal model $f(x,p) = p_1 + p_2 x + p_3 x^2 + \ldots + p_n x^n$
3) Multi-exponential model $p_1 e^{p_2 x} + \ldots + p_{2n-1} e^{p_{2n} x}$

Of course, the linear model is a special case of the polynomial model. Generally, a model is called *quasilinear* when f is a linear function of p. This does not exclude the case that f is nonlinear in x. In particular, the polynomial model is quasilinear although the functional dependence on x may be quadratic, as in Fig. 4 b. Given the data pairs (x_i,y_i), the parameter p yielding the best approximation of f to all these data pairs is found by minimizing the sum of the squares of the deviations between the measured values y_i and their modeled counterparts $f(x_i)$:

$$p \rightarrow \text{minimize} \sum_{i=1}^{N} (y_i - f(x,p))^2$$

The advantage of quasilinear models is that the exact solution for p is easily obtained by solving a system of linear equations (see Section 11.1). In the simple case of linear models, one can even directly indicate the explicit solution:

$$p_1 = \frac{s_{xx} s_y - s_x s_{xy}}{d}; p_2 = \frac{N s_{xy} - s_x s_y}{d} \tag{5.1}$$

where

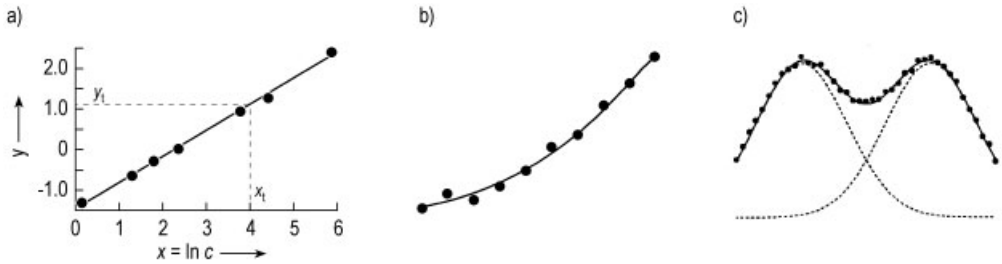

Figure 4. Illustration of different regression models. Linear model a), polynomial model b) and nonlinear model

$$s_x = \sum_{i=1}^{N} x_i; s_y = \sum_{i=1}^{N} y_i; s_{xy} = \sum_{i=1}^{N} x_i y_i;$$
$$s_{xx} = \sum_{i=1}^{N} x_i^2; d = N s_{xx} - (s_x)^2$$

The coefficients p_1 and p_2 refer to the intercept and slope of the straight line fitting the data points (see Fig. 4 a). Being based on the random variables x and y, the coefficients p_1 and p_2 are outcomes of random variables themselves. The true coefficients can be covered by the following 95 % confidence intervals:

$$p_1 \pm t_{N-2;0.975} s_R \sqrt{\frac{1}{N} + \frac{\bar{x}^2}{s_Q}}; \quad p_2 \pm \frac{t_{N-2;0.975} s_R}{\sqrt{s_Q}} \tag{5.2}$$

where

$$s_R = \sqrt{\frac{\sum_{i=1}^{N} (y_i - p_1 - p_2 x_i)^2}{N-2}}$$

and

$$s_Q = s_{xx} - \frac{(S_x)^2}{N}.$$

In contrast to the polynomial model, the multi-exponential model is not quasilinear. This makes the determination of the parameters p more complicated. Then iterative methods have to be employed in order to approach a solution. This issue falls into the framework of nonlinear regression analysis. For more details, see [5].

5.2. Calibration

The purpose of quantitative chemical analysis is to determine concentrations of certain analytes on the basis of concentration-dependent instrumental responses or signals. The relation between signals (or functions thereof) and concentrations (or functions thereof) is established by calibration. Calibration consists of two steps: in the first step a functional dependence $y = f(x)$ is modeled on the basis of so-called calibration samples with known concentrations (e.g. reference materials) and corresponding signals, altogether giving a set of data pairs (x_i, y_i). To these data points, an appropriate function is fitted as described in Section 5.1. In the second step, the obtained model is applied to a test sample with unknown concentration on the basis of a measured signal. Here, the inverse function is used: $x = f^{-1}(y)$.

Example 5.1. The following calibration results were obtained for the determination of Ni in WO_3 by optical emission spectroscopy (OES):

Sample	1	2	3	4	5	6	7
c/ppm	1	3.5	6.0	11.0	51.0	101	501
x = ln *c*	0.00	1.25	1.79	2.40	3.93	4.62	6.22
y	−1.31	−0.66	−0.33	−0.04	0.84	1.15	2.26

The signal is in good linear relation with the logarithm of concentrations. The data are plotted in Fig. 4 a. From the data table one calculates the values required in Equation (5.1) as

$$s_x = 20.21; s_y = 1.91; s_{xy} = 21.16; s_{xx} = 86.0;$$
$$d = 193.6; p_1 = -1.36; p_2 = 0.57$$

For a test sample, a signal with value y_t = 1.0 was measured. Starting from the obtained model $y = -1.36 + 0.57x$, one calculates the inverse function as $x = (y + 1.36)/0.57$; hence, $x_t = 2.36/0.57 = 4.14$. Finally, this last result is transformed back to concentrations via $x = \ln c$. This gives $c_t = e^{x_t} = 62.8$ as the concentration of Ni in the test sample.

In linear calibration, the intercept p_1 corresponds to the signal of the blank, i.e., the signal

occurring in absence of the analyte. The slope p_2 is called the *sensitivity* of the signal. For nonlinear calibration curves, this sensitivity changes and has to be calculated as the first derivative dy/dx at a given concentration value x.

The methodology presented so far corresponds to the *classical calibration*. A different approach is *inverse calibration,* where the functional relationship is directly modeled in the form required for prediction of concentrations in test samples: $x = f(y)$. This approach has proven useful in the context of multivariate calibration (see Chap 11).

6. Characterization of Analytical Procedures

An analytical procedure is characterized by its range of concentration (including calibration), precision (random error), trueness (systematic error), selectivity, and principal limitations.

Trueness [6, 7]. A systematic error can occur as an *additive error* (e.g., an undetected blank) or a *multiplicative error* (e.g., an incorrect titer). Systematic errors are detected by analyzing a short series of m samples with "known" contents x_i and "found" contents y_i as the results. Evaluation by linear regression (Chap. 5) yields $y = a + b\,x$. An intercept $a \neq 0$ is due to an additive systematic error, whereas a slope $b \neq 1$ indicates a multiplicative error. The significance of a and b is tested by verifying that 0 and 1 do not belong to the 95 % confidence intervals around a and b, see Equation (5.2).

If samples with known concentration are not available (e.g., very often in the case of bioproducts), the multiplicative error can be detected in the following way. The sample (solution) is divided into two halves. One half is analyzed directly (result x_1). To the other half an exactly defined amount of the analyte x_+ is added (result x_2). Then the "recovery rate" b is given by

$$b = (x_2 - x_1)/x_+$$

This is performed for m samples, resulting in

$$\bar{b} = \frac{1}{m}\sum_{i=1}^{m} b_i; s_b = \sqrt{\sum_{i=1}^{m}\left(b_i - \bar{b}\right)^2 / (m-1)}$$

Testing for the absence of a multiplicative systematic error amounts to testing the statistical hypothesis $E(b) = 1$, where $E(b)$ denotes the expected value of the slope b. According to Section 3.2 (see Example 3.2), the hypothesis above can be rejected at the 95 % level if the test statistic $\left(1 - \bar{b}\right)\sqrt{m}/s_b$ falls outside the interval $\left[-t_{m-1;0.975}, t_{m-1;0.975}\right]$. Equivalently, a multiplicative systematic error is proven with 95 % probability if

$$t_b := \left|1 - \bar{b}\right|\sqrt{m}/s_b > t_{m-1;0.975}.$$

Example 6.1. Trueness of determination of arsenic levels in yeast by hydride atomic absorption spectroscopy (AAS) should be tested. However, samples with known contents were not available. Therefore, the recovery rate (with x_+ = 30 µg As) was determined from m = 5 unknown samples.

Sample	1	2	3	4	5
x_1, µg As	5.8	13.8	30.0	43.1	66.8
x_2, µg As	35.2	43.3	59.6	72.4	96.5
b_i	0.9800	0.9833	0.9867	0.9767	0.9900

$\bar{b} = 0.9833$, $s_b = 0.005266$
$t_b = 7.09$; $t_{4;0.975} = 2.78$

Since $t_b > t_{4;0.975}$ the recovery rate differs undoubtedly from $b = 1$. The results gained by hydride-AAS are on average about 2 % too low.

Selectivity [6]. In a multicomponent system consisting of an analyte and other components, each component contributes to the signal of the analyte. Selectivity is given if the signal of the analyte is only randomly influenced by the other components. This can be tested by a model

$$y = c_1x_1 + c_2x_2 + \cdots + c_px_p$$

of multiple linear regression (Section 11.1), where the x_j refer to the concentrations of the analyte ($j = 1$) and the remaining components ($j = 2, \ldots, p$), y is the measured signal and the c_j denote the sensitivities of the respective components. In order to estimate these sensitivities, an experimental design [8] is set up which allows the variation of the signal to be characterized in dependence on variations of the component concentrations. The calculation of sensitivities and their confidence intervals (indicating statistical significance) is carried out on the basis of Equations (11.2) and (11.3).

Table 3. Experimental design for variation of component concentrations (g/L) and measured signals

Experiment	[Na]	$[Ca^{2+}]$	$[K^+]$	$[Mg^{2+}]$	$[SO_4{}^{2-}]$	$[Cl^-]$	$[PO_4{}^{3-}]$	Signal
1	0.1	0.1	0.1	0.1	0.0	1.0	0.0	132
2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	137
3	0.1	0.1	0.0	0.1	0.0	0.0	0.1	119
4	0.1	0.0	0.1	0.0	0.0	1.0	0.1	111
5	0.1	0.1	0.0	0.0	0.1	1.0	0.1	125
6	0.1	0.0	0.0	0.1	0.1	1.0	0.0	114
7	0.1	0.0	0.1	0.1	0.1	0.0	0.1	93
8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	129

Example 6.3. The selectivity for the determination of sodium (xNa = 100 mg/L) by flame photometry in the presence of Ca^{2+}, K^+, Mg^{2+}, $SO_4{}^{2-}$, Cl^-, and $PO_4{}^{3-}$ had to be tested. The experimental design for the variation of components as well as the corresponding measured signals (scale reading) are recorded in Table 3.

The calculations (according to Section 11.1) provided the following result:

$$y = (1292.5\pm84.0)\,[\mathrm{Na}] + [165\pm63.5]\,\left[Ca^{2+}\right] - (35\pm63.5)\,\left[K^+\right] - (110\pm63.5)\,\left[Mg^{2+}\right] - (55\pm63.5)\,\left[SO_4^{2-}\right] + (1\pm6.4)\,\left[Cl^-\right] - (160\pm63.5)\,\left[PO_4^{3-}\right]$$

The significance of a component can be read off from whether or not the value 0 belongs to the respective confidence interval. Accordingly, apart from the analyte's contribution (which is evident), the presence of Ca^{2+}, Mg^{2+}, and $PO_4{}^{3-}$ must be considered as significantly influencing the signal, whereas the perturbations due to K^+, $SO_4{}^{2-}$, and Cl^- are negligible from the statistical point of view. Summarizing, selectivity is not given for the considered system.

Principle Limitations. The analysis of traces is often influenced by the impurities of the reagents. Then a blank $\bar{y}_{\mathrm{bl}}\neq0$ (standard deviation σ_{bl}) is measured at the concentration $x = 0$ of the analyte. From Equation (3.1) we know that an analytical signal $\bar{y}_{\mathrm{a}}$ (n_{j} parallels) is different from the blank with 95 % probability if

$$\bar{y}_{\mathrm{a}}-\bar{y}_{\mathrm{bl}}>1.96\frac{\sigma_{\mathrm{bl}}}{\sqrt{n_j}}$$

Extending the factor 1.96 to the value 3 would then even yield a difference with 99.7 % probability according to the 3σ-rule (see Section 2.2). Now, taking into account that the expected value of the signal of some analyte is never smaller than the blank, one can sharpen the statement above: the analytical signal is different from the blank with 99.85 % probability if

$$\bar{y}_{\mathrm{a}}>\bar{y}_{\mathrm{bl}}+3\frac{\sigma_{\mathrm{bl}}}{\sqrt{n_j}} = y_{\mathrm{c}}$$

where y_{c} is the critical value. The improvement in probability here is due to using one-sided additional information.

In the case of a blank the calibration function takes the form (see Fig. 5 A)

$$y = a+bx = \bar{y}_{\mathrm{bl}}+bx$$

For $y = y_{\mathrm{c}}$ the evaluation function will be

$$x = (y_{\mathrm{c}}-a)\,/b = x_{\mathrm{d}}$$

x_{d}—*the limit of decision* [9]—is due to the minimum concentration of the analyte that gives a significant analytical signal.

A signal $y = y_{\mathrm{c}}$ results from a concentration $x = x_{\mathrm{d}}$. The frequently repeated analysis of such a sample yields a signal $y < y_{\mathrm{c}}$ as often as $y > y_{\mathrm{c}}$ caused by the random error of the blank (see Fig. 5 B). Expressed as probabilities gives

$$P\,(y<y_{\mathrm{c}}) = P\,(y>y_{\mathrm{c}}) = 0.5$$

Therefore the value x will be interpreted as "analyte present" $\rightarrow$ (+) as often as "only blank $\rightarrow$ analyte absent" $\rightarrow$ (−), or again expressed as probabilities

$$P_x^+ = P_x^- = 0.5$$

Therefore the limit of decision x_{d} never can be used as guarantee of purity. A sufficiently high

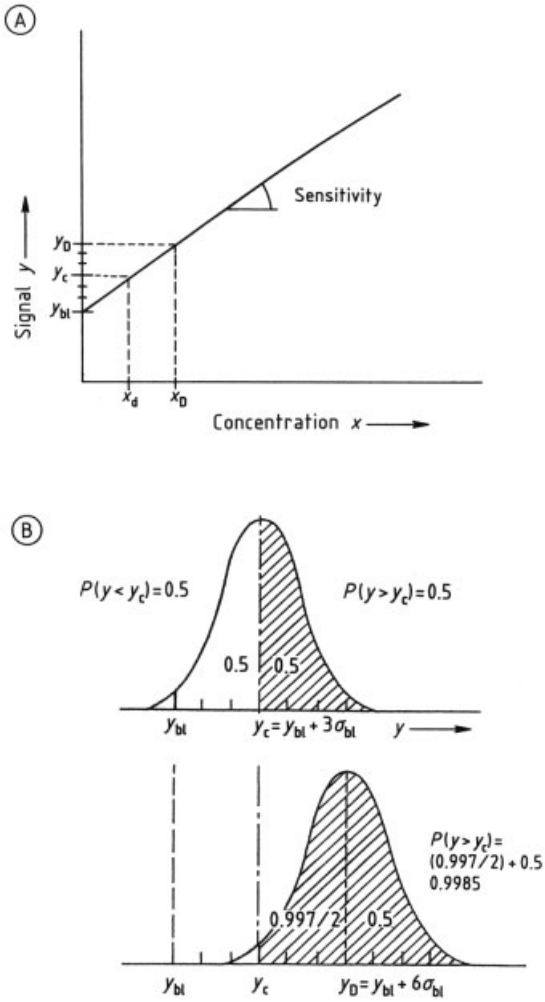

Figure 5. Limit of decision x_d and of detection x_D
A) Calibration function in presence of a blank;
B) Probabilities for the detection of an impurity at limit of decision (upper curve) and at limit of detection (lower curve)

reliability for such a guarantee is given for the signal value

$$y_D = \bar{y}_{bl} + \frac{6\sigma_{bl}}{\sqrt{n_j}}$$

or the associated concentration

$$x_D = \frac{6\sigma_{bl}}{b\sqrt{n_j}}$$

x_D is termed the *detection limit* [9]. Then,

$$P_x^+ = P(y_c) = 0.0085$$
$$P_x^- = P(y<y_c) = 0.0015 \qquad (6.5)$$

An impurity can be detected now at a very high probability level. Therefore, only this limit of detection allows the characterization of high-purity material.

Example 6.4. The photometric determination of iron with triazine (absorptivity $A = 2.25 \times 10^3$ m^2/mol) yielded $\bar{y}_{bl}$ (= E_{bl}) = 0.08 and σ_{bl} = 0.02. For the analysis with duplicates, the critical value is

$$y_c\ (= E_c) = 0.08 + 3\times 0.02/\sqrt{2} = 0.122$$

The evaluation function gives

$$x_d = (0.122 - 0.080)\,/2.25\times 10^3 = 1.87\times 10^{-5}\ \text{mol/L}$$

and the limit of detection x_D follows from

$$y_D = 0.08 + 2\times 3\times 0.02/\sqrt{2} = 0.165$$

$$x_D = (0.165 - 0.080)\,/2.25\times 10^3 = 3.78\times 10^{-5}\ \text{mol/L}$$

7. Signal Processing

Recording instrumental responses to the presence of some analyte is at the heart of qualitative and quantitative chemical analysis. The physical nature of these responses or signals may be different according to appropriateness with respect to the given analyte (e.g., optical, electrochemical etc.). In the simplest case, just one signal is recorded, e.g., absorption at some wavelength. Modern instrumentation, however, allows signals to be monitored over varying physical quantities such as wavelength or time or both of them coupled. Signals are always subjected to perturbations of different kind and origin. Such perturbations can be induced by noise related to the analyzing technique and its apparatus. Furthermore, signals of a certain analyte under investigation are frequently perturbed by interference with signals of different species, which leads to falsifications both in qualitative and quantitative analysis.

7.1. Fourier Transform

Contamination of analytical signals by high-frequency noise (e.g., 50 Hz modulated noise caused by the power line) can be detected and corrected by Fourier transformation. Figure 6

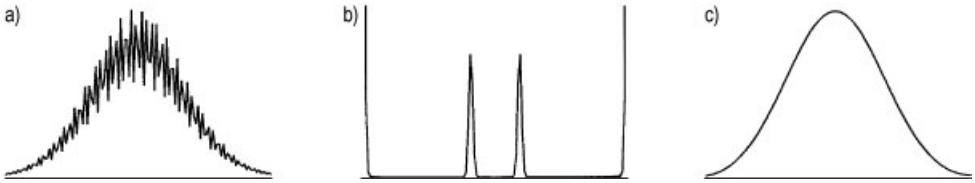

Figure 6. Illustration of Fourier transform
a) Noisy Gaussian signal peak; b) Resulting Fourier transform; c) Filtered signal

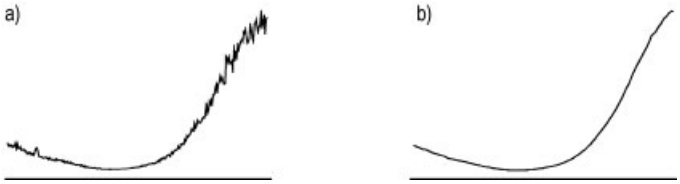

Figure 7. Signal smoothing
a) Noisy fluorescence spectrum; b) Spectrum smoothed by using moving average

a shows a Gaussian signal (e.g., spectroscopic peak over the wavelength domain) containing sinoidal noise. For a discretized signal f with n values $f(0), \ldots, f(n-1)$ the (discretized) Fourier transformed signal is defined as

$$F(u) = \frac{1}{n}\sum_{t=0}^{n-1} f(t)\, e^{-2\pi uti/n} \; (u = 0, \ldots, n-1)$$

where i is the imaginary unit and n is the discretization number. As the Fourier transform is complex, one usually plots its absolute value as shown in Figure 6 b. The symmetry of the Fourier transform around the center point means that only the first half of the spectrum contains substantial information. Peaks in the Fourier transform indicate the presence of periodic contributions in the original signal. The higher the frequency of this contribution, the more this peak shifts to the center. Hence, each of the symmetrically arranged two peaks in the center of Figure 6 b corresponds to the presence of a high-frequency contribution in the original signal, whereas the two peaks on the boundary represent the (low-frequency) signal itself. In this way, noise and signal are clearly separated in the Fourier transform. To arrive at such separation in the original representation of the signal, one may exploit so-called filtering: in a first step, the Fourier transformed signal is convoluted with a so-called kernel which typically is some simple function having a peak in the center of the frequency domain; the convoluted signal is transformed back then to the original domain of the signal by Inverse Fourier transformation. The result is shown in Figure 6 c. A clear improvement of the signal-to-noise ratio becomes evident when compared with Figure 6 a.

7.2. Data Smoothing

Not all kinds of noise are efficiently removed by Fourier transformation. This is true, in particular, for nonperiodic perturbations such as single spikes. Then, direct data smoothing may be helpful. The simplest method is the moving average, in which each data point of the original signal is replaced by the average of its k nearest neighbors. The appropriate choice of the parameter k is crucial. Enhancing k improves the smoothing effect, which is desirable as far as the reduction of noise is concerned. If k is chosen too large, then information of the signal itself will be smoothed away. Figure 7 a shows an originally measured fluorescence spectrum (267 data points) with clear contamination by noise, whereas Figure 7 b plots a version smoothed by moving averages with $k = 20$. More advanced techniques are based on polynomial smoothing.

7.3. Signal Resolution

Frequently, the analyte cannot be completely separated by chemical means from other species

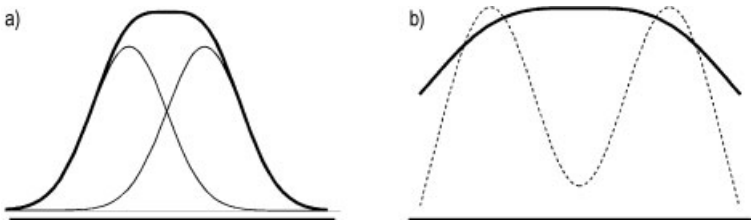

Figure 8. Signal sharpening by differentiation
a) Single peak as a superposition of two close Gaussian peaks; b) Second differential of the superposed signal, yielding two resolved peaks again

whose signals interfere with the signal of interest. Then, the resulting superposed signal may be distorted so much that quantitative analysis becomes impossible without additional mathematical tools. In the worst case, even the presence of an interfering component is hidden. Figure 8 a shows two idealized Gaussian peaks the superposition of which results in a single peak only. Such junction of peaks occurs when the distance between them is below a critical value. In this case, signal sharpening by differentiation is a possible remedy. Fig. 8 b shows the superposed single peak of Fig. 8 a in a shrinked domain along with its second derivative. Now, the presence of two underlying components is clearly evident again. In principle one could enhance the order of differentiation, but the gain in signal resolution quickly becomes negligible in comparison with the increase in noise.

8. Basic Concepts of Multivariate Methods

8.1. Objects, Variables, and Data Sets

The increasing performance of modern instruments allows the characterization of certain objects under investigation by a whole set of properties, which yield a typical fingerprint for each object. These properties are usually statistical variables, whose realization is measured experimentally.

Example 8.1. Several water samples (objects) are analyzed with respect to concentrations of a set of relevant trace elements (variables). Collecting the resulting data gives a constellation as in Table 4. In such data tables or data sets the rows correspond to objects and the columns to variables.

Table 4. Small example of a fictive data table (arbitrary units)

Sample no.	[Cd]	[Pb]	[Cu]
1	3	3	1
2	1	2	3
3	4	3	3
4	2	2	0

Real life data sets may be generated by very large numbers of objects or variables, depending on the problem. From a more technical point of view a distinction can be made between experimental and instrumental data sets. *Instrumental data sets* are the direct output of instruments, such as digitized spectra, chromatograms, or images in image processing. Here, the number of objects or variables may reach some ten or hundred thousands; therefore, data processing and evaluation generally are automated. *Experimental data* result from single observations (e.g., physicochemical parameters of natural water at different locations) or from condensation of instrumental data (e.g., one concentration value for a sample instead of a digitized spectrum). As a consequence, acquisition of experimental data is much more expansive, leading to smaller data sets (in the range of several hundreds of objects or variables). As a rule, experimental data tables can be neither processed nor evaluated automatically but require consideration of specific circumstances in the treated problem.

Compared to classical measurement series restricted to a single variable, the multivariate approach yields a considerably greater amount of information for object characterization. This advantage, however, is connected with a loss of direct interpretability. If only two variables were measured, the whole data structure might be still visualized in a diagram, but this approach fails for larger problems. That is why the increase in information must be accompanied by some

efficient tool of *data reduction* that enables interpretation or visualization of the essential part of the data structure, thereby separating it from unessential noise. In this sense, data reduction can be understood as the main goal of multivariate methods in chemometrics. Whereas theoretical foundations were developed at the beginning of this century, practical applications of these procedures have revived considerably only during the last few decades (from the 1970s onward) in connection with modern computational equipment (for introductory surveys of multivariate statistics see [10, 11]).

8.2. Correlation and Distance Matrices

An important task of multivariate data analysis is classification of objects and variables (i.e., subdivision of the whole data set into homogeneous groups of similar objects or variables, respectively). Similarity of variables is usually measured by their correlation coefficient, whereas similarity of objects is expressed in terms of the geometric distance. The *correlation coefficient* r of two variables x_1, x_2 is computed according to

$$r\,(x_1,x_2) = \frac{\sum_{i=1}^{n} (x_{i1}-\bar{x}_1)\,(x_{i2}-\bar{x}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1}-\bar{x}_1)^2}\cdot\sqrt{\sum_{i=1}^{n} (x_{i2}-\bar{x}_2)^2}} \qquad (8.1)$$

where n is the number of realizations (i.e., number of objects or rows of the data table); x_{ij} is the ith realization of the jth variable (i.e., entry of the data table that is located in the ith row and jth column); and $\bar{x}_j$ is the mean of the jth variable.

For the fictive example of Table 4, for instance, the following coefficients r are obtained r (Cd, Pb) = 0.89, r (Cd, Cu) = 0.09, and r (Pb, Cu) = 0.19. All information on correlation can conveniently be collected in the correlation matrix $\boldsymbol{R}$, which is the basis of many latent variable methods such as principal components, factor, or discriminant analysis:

$$\begin{array}{c} \\ \text{Cd} \\ \text{Pb} \\ \text{Cu} \end{array} \begin{array}{c} \begin{array}{ccc} \text{Cd} & \text{Pb} & \text{Cu} \end{array} \\ \begin{pmatrix} 1.00 & 0.89 & 0.09 \\ 0.89 & 1.00 & 0.19 \\ 0.09 & 0.19 & 1.00 \end{pmatrix} \end{array}$$

Sometimes, instead of correlation the covariance of two variables is used, which is just the numerator of Equation (8.1).

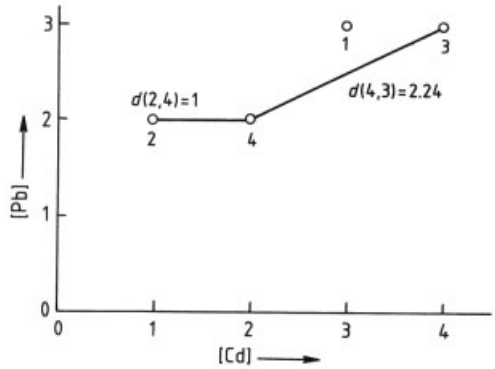

Figure 9. Plot of four samples from Table 4 in an elementary diagram

To recognize similarities among objects, they can best be considered as points in p-dimensional Euclidean space, where p is the number of variables (columns) of the data set. If in Table 4 only the concentrations of Cd and Pb had been measured, then all samples could be represented in a diagram as in Figure 9 (i.e., in a two-dimensional space or plane). Although for technical reasons, such representations cannot be realized if the number of variables exceeds 3, the mathematical concept of p-dimensional space is nevertheless useful, for example, when defining the distance of two objects. In a planar display, distances are easily computed by the Pythagorean theorem by taking the square root of the sum of squared coordinate differences between two points (e.g., the distance between samples 4 and 3 equals $\sqrt{5}$). Obviously, the more similar two objects are in all their coordinates the smaller their computed distance is. In this way, homogeneous classes of similar objects may be detected as those that have only small distances between them. Of course, computation of distances is not restricted to the plane but is extended to p-dimensional constellations in a straightforward way:

$$d\,(x_a,x_b) = \sqrt{\sum_{j=1}^{p} (x_{aj}-x_{bj})^2} \qquad (8.2)$$

where x_a, x_b are the objects a and b; and x_{aj} is the entry of the data table in row a and column j.

Taking all three columns of Table 4 into account gives, for instance:

$$d(4,3) = \sqrt{(2-4)^2+(2-3)^2+(0-3)^2} = 3.74$$

All pairwise distances are again collected to give the distance matrix $\boldsymbol{D}$, which forms the basis of many computational approaches including cluster analysis or multidimensional scaling:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0.00 & 3.00 & 2.24 & 1.73 \\ 3.00 & 0.00 & 3.16 & 3.16 \\ 2.24 & 3.16 & 0.00 & 3.74 \\ 1.73 & 3.16 & 3.74 & 0.00 \end{pmatrix} \end{array}$$

In this matrix the most similar pair of (different) objects is (1, 4), while the most dissimilar pair is (3, 4). Apart from the classical Euclidean distance defined by Equation 8.2, some further relevant measures exist such as Mahalanobis or Manhattan distance. The Mahalanobis distance, for instance, which is important in classification (see Chap. 10), is computed according to

$$d(x_a,x_b) = \sqrt{(x_a-x_b)^{\mathrm{T}} S^{-1} (x_a-x_b)} \tag{8.3}$$

where $\boldsymbol{S}$ is the matrix of all pairwise covariances between variables.

8.3. Data Scaling

Data pretreatment is a necessary condition of multivariate analysis. Besides appropriate transformation of data according to their type of distribution (e.g., taking logarithms for concentration values in the range of the detection limit is recommended), data scaling is necessary as a first step of analysis for almost all methods. Scaling means that each variable is linearly transformed to have zero mean and unit standard deviation. In this way the fact of different variables being measured in different units (e.g., concentrations in parts per million or percent is circumvented, to avoid an arbitrary influence in the total variance or the pairwise object distances. After scaling, all variables have equal a priori influence.

9. Factorial Methods

9.1. Principal Components Analysis

Principal components analysis (PCA) as a statistical method was introduced by Hotelling in 1933. Principal components are linear combinations of the original variables with optimal features: the first PC defines maximum variance among all possible linear combinations; the second PC defines maximum variance among all linear combinations uncorrelated with the first one, etc. In this way a small set of a few PCs (generally much fewer than the number of original variables) will suffice to represent the greatest part of the total data variation. Furthermore, in contrast to the original variables, different PCs are always uncorrelated, which makes them useful for many statistical applications. Geometrically, PCs may be identified as new coordinate axes that point to directions of large variance. Using the coordinate system defined by the first two PCs is a very popular approach for visualizing data structure in a diagram. From the computational viewpoint, PCs are obtained by solving the eigenvalue problem

$$R \cdot v_j = \lambda_j \cdot v_j \; (j = 1,\ldots p) \tag{9.1}$$

where $\boldsymbol{R}$ is the correlation matrix (see Chap. 8); $\boldsymbol{v}_j$ are the eigenvectors; and λ_j are the eigenvalues of $\boldsymbol{R}$.

The number of linearly independent eigenvectors is at most p. Each $\boldsymbol{v}_j$ consists of p components v_{ij}, which are the coefficients (weights) of the original variables. The ith realization of the jth PC is (with x_{ik} being an entry of the scaled data table)

$$\mathrm{PC}_{ij} = \sum_{k=1}^{p} x_{ik} \cdot v_{kj} \tag{9.2}$$

By using just the first two PCs, a diagram plot for the *objects* is obtained with PC_{i1}, PC_{i2} as coordinates of the ith object. These coordinates are frequently called scores. Similarly, in the same plot a representation of *variables* may be given by using v_{i1}, v_{i2} as coordinates (which are called loadings) for the ith variable. Finally, the eigenvalues contain information on the amount of data structure covered by a PC: λ_j is the variance of the jth PC. Frequently, indication of the variance percentage is preferred by relating one eigenvalue to the sum of all eigenvalues, which in the

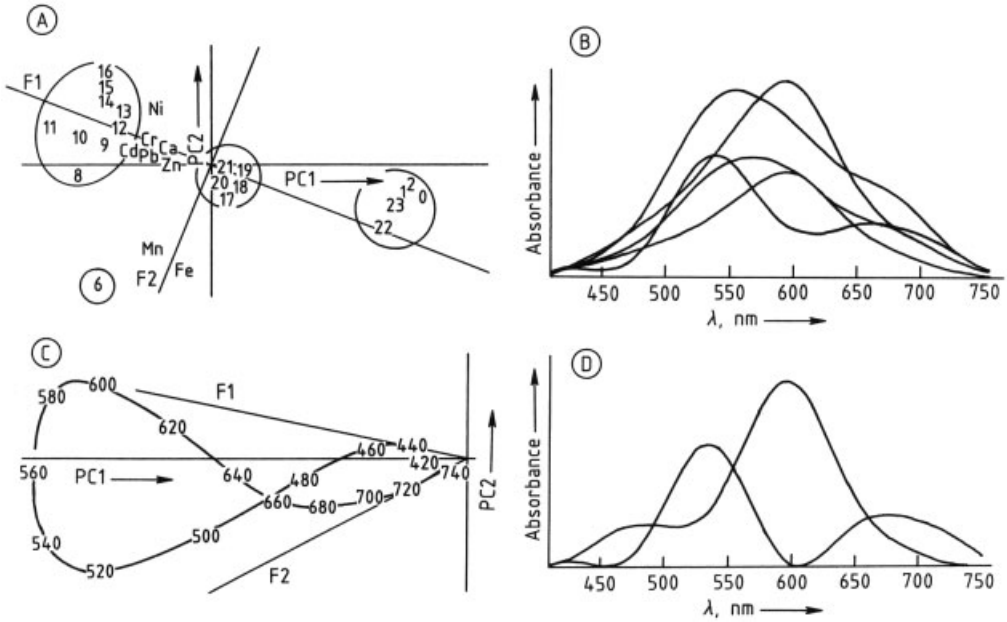

Figure 10. Examples of PCA and FA
A) PC diagram for trace element characterization of wastewater samples; B) Simulated (noise-added) absorbance spectra of five mixtures of two pure components; C) PC wavelength plot for B) including the axes F1 and F2 of pure components; D) Pure component spectra related to B) using the self-modeling method

case of scaled data simply equals p (the number of original variables).

Example 9.1. Figure 10 A provides the PC plot (PC1 versus PC2) for a data set containing analytical results of eight trace elements (variables) in 20 samples (objects) of urban wastewater. The numbers in the plot refer to hours of sampling time in the course of a day. The element patterns of samples do not change continuously, but rather abruptly; four clusters of sampling time are visible: (6 h), (8 – 16 h), (17 – 21 h), and (22 – 2 h). The plot of variables reveals two major groups of correlating elements: (Fe, Mn), and (Ni, Cr, Cd, Cu, Pb, Zn). Even a joint interpretation of scores and loadings with respect to influence of variable groups in object groups is possible. This aspect is reconsidered below. In this example, the first two of eight eigenvalues were $\lambda_1 = 4.78$, $\lambda_2 = 1.49$; hence, ca. 78% of total data variation is exhausted by the first two PCs, which is much more than by using any two of the original variables (recall that due to equal weighting by scaling, each pair of original variables covers a percentage of 2/8 = 25% of the total variance).

An important kind of chemometric application of PCA is to find out the number of (unknown) components determining the spectra of a set of mixtures of these components.

Example 9.2. Figure 10 B shows simulated and noise-added spectra of five mixtures as an example. Each of these spectra may be digitized (in the example, 680 equidistant wavelengths in the range between 410 and 750 nm were used) to yield a data table with wavelengths as objects and absorbances of the mixtures at these wavelengths as variables. The total variance within such a data set is composed by a systematic part due to changing proportions of the underlying components within the mixtures and by a generally much smaller part due to noise. Consequently, an eigenvalue analysis should provide a group of large eigenvalues corresponding to the principal components, which reflect chemical composition, and a group of very small eigenvalues relating to PCs, which reflect only noise. Various methods exist for objective separation of both groups: graphical (e.g., scree-test), statistical (e.g., Bartlett's test on sphericity or cross-validation) or empirical procedures (e.g., Malinowski's indicator function). In the given artifi-

cial example the eigenvalue sequence is 4.453, 0.540, 0.004, 0.002, 0.000, making clear, even without any objective evaluation, that the number of underlying chemical components is probably two.

A review on PCA from the chemometric viewpoint may be found in [12].

9.2. Factor Analysis

Factor analysis (FA), developed in 1947 by Thurstone, is a statistical method with different model assumptions compared to PCA. In chemometric applications, however, FA is frequently considered as an additional step after PCA, with the aim of making the results of the latter easier to interpret (a chemometric approach to FA is found in [13]). One important feature of FA is rotation of PCA solutions. In general, PCs are abstract variables without physical meaning. Therefore, an attempt is made to rotate the axes corresponding to the first few significant PCs so as to make the new axes better fit certain groups of variables and facilitate their interpretation as physically meaningful latent factors. The variety of rotations is subdivided into orthogonal (rotations in the true sense) and oblique transformations. *Orthogonal transformations* leave unchanged the variance percentage covered by the rotated axes. *Oblique transformations*, in contrast, do not provide uncorrelated factors but have more degrees of freedom for fitting groups of variables. From the computational viewpoint a successful transformation is achieved by optimizing an appropriate goal function (e.g., Kaisers varimax criterion, compare [13]). In Example (9.1) (Fig. 10 A) the optimal pair of orthogonal factors is denoted by F1, F2, respectively. It obviously fits the variable (and object) structure better than the PCs. Now, the above-mentioned joint interpretation of objects and variables may be given. Obviously, F1 is determined by (Ni, Cr, Cd, Pd, Cu, Zn) and it discriminates three major groups of sampling times with decreasing concentrations in the mentioned six elements from the beginning to the end of a day. This makes it reasonable to identify F1 as a factor in industrial pollution of wastewater. F2 consists of the (Fe, Mn) group, which is related to the 6-h sample. High concentrations of Fe and Mn in wastewater in the early morning might result from water pipes in private households. With some caution, this could represent a second, independent factor in wastewater exposure.

Another, frequently employed aspect of FA is target testing. This procedure is very useful for deriving new hypotheses from the given data set (for details of the method and applications, particularly in chromatography, see [13]). In Example (9.2) a specific chemical component might be suspected to be present in the mixtures. To verify this, the vector of digitized spectrum of this component could be considered a target to be tested. If the presence of the fixed component is probable, its digitized spectrum vector should be a linear combination of the score vectors of significant PCs since these are, in turn, linear combinations of the spectra of the unknown pure chemical components. The test is performed by multiple linear regresssion (see Chap. 11.1), indicating whether the target spectrum may be successfully predicted (within the range of experimental error) by a suitable linear combination of the PC spectra.

A powerful method of FA is *self-modeling of mixture spectra* [14], which allows, under a given assumption, the spectra of underlying pure components to be isolated even if these cannot be modeled by a certain band shape such as Gaussian or Lorentzian. This is a good alternative to conventional least-squares fit. Figure 10 C shows the PC score plot (680 points for different wavelengths in the range 410 – 750 nm), for Example (9.2). According to Beer's law, absorbances of mixtures are positive combinations of the absorbances of pure components. Therefore, the wavelength points of mixture spectra in a PC plot have positive coordinates with respect to factor axes representing the pure components. Consequently, these pure component axes must define some cone that contains all the wavelength points. This is easily illustrated if, as in the example, the number of pure components equals two. Of course, some degree of freedom exists for choosing the axes F1 and F2. Under the specific assumption, however, that each pure component absorbs radiation uniquely at least at one wavelength (within the considered range), the cone defined by F1 and F2 must be a minimal one (i.e., both axes must pass the wavelength loop tangentially; at 440 and 730 nm in the example). Having found F1 and F2 in this

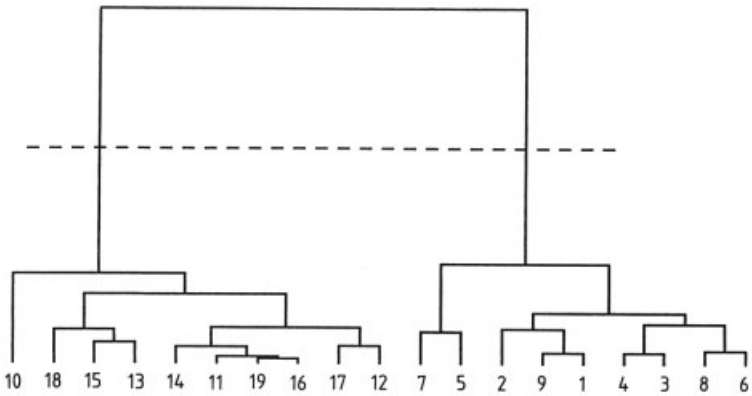

Figure 11. Dendrogram for hierarchic clustering of 19 tungsten powder samples

way, the pure component spectra are easily established: The absorbances of the first and second pure components, respectively, are given by projecting corresponding wavelength points of the PC plot onto F1 and F2, respectively. Projection must be carried out parallel to both axes. The result is illustrated in Figure 10 D.

10. Classification Methods

10.1. Cluster Analysis

The classification of objects by their patterns in the data set plays a central role in chemometrics. Depending on the prior knowledge concerning the given problem such classification may be performed in a supervised or nonsupervised way. The main tool of nonsupervised or automatic classification is cluster analysis, which produces a partition of the set of objects into several homogeneous subgroups (for an introduction to this field, compare [15]). From a rough viewpoint, cluster analysis may be subdivided into hierarchic and nonhierarchic clustering. *Hierarchic clustering* generates a sequence of partitions that satisfy the definition of a hierarchy; i.e., if any two classes of possibly different partitions are compared, either these classes are disjoint or at least one class is contained in the other (overlapping is excluded). This sequence may be visualized by a so-called dendrogram similar to genealogical trees in taxonomic classification.

Example 10.1. Consider Figure 11, which resulted from hierarchic clustering of 19 tungsten powders being characterized by 11 trace elements. Partitions with different degrees of refinement are obtained by horizontal cuts of the dendrogram. At the lower end, all objects occur as separated classes; at the upper end, all objects are fused in a single class. The level at which certain subclasses are fused (vertical direction) may be interpreted as their dissimilarity. As a consequence, homogeneous groups are found below long branches of the tree. In the example, a bipartition (1 – 9 and 10 – 19) of the object set is suggested. In fact, the two groups of trace element patterns are in coincidence with two different chemical treatments of tungsten raw materials.

Depending on whether the sequence of partitions is generated by starting at the lower end (successive fusion) or at the uper end (successive splitting), agglomerative and divisive procedures can be distinguished. Most of the well-known methods are *agglomerative* and obey the following type of general algorithm:

1) Given a data set X compute the distance matrix $\boldsymbol{D}$ (see Chap. 8) and define each object to form a single class.
2) Find out the most similar pair of classes [e.g., C_1, C_2 ($C_1 \neq C_2$)], realizing the smallest distance $d_{1,2}$ in $\boldsymbol{D}$. Fuse all elements of C_1 and C_2 to yield a new, larger class $C_{1,2}$. Fix $d_{1,2}$ as the index of fusion (height in the dendrogram). Update $\boldsymbol{D}$ (with a reduced number of classes) by recomputing all distances between the new class $C_{1,2}$ and all old classes that differ from C_1 and C_2. Repeat step 2 until all objects are fused in a single class.

In fact, agglomerative procedures differ only by the way of updating the new distance matrix in each step. As an essential drawback of hierarchic clustering, "mistakes" of fusion at a low level of aggregation are retained, due to hierarchy, in the whole sequence of partitions.

If different partitions are obtained in an independent way by varying some parameter, *nonhierarchic clustering* is used. In so-called partition-making algorithms, such as k-means or Forgy's method, the number of classes expected must be predefined. Different numbers of classes yield different partitions that might have some overlapping; this would indicate an uncertainty in the class structure. Various criteria are used to decide which number of classes or which partition is appropriate. Other important representatives of nonhierarchic clustering are fuzzy clustering and potential clustering (the nonsupervised version of the ALLOC method discussed below).

10.2. Supervised Classification

The general principle of supervised classification may be sketched as follows: For specific predefined classes of interest (e.g., different labels of brandies in food chemistry) a number of samples can be characterized by certain properties (e.g., areas of relevant gas-chromatographic peaks). These samples are called learning objects because their origin is known. Based on the data set of learning objects, rules are derived for subdividing the space of all possible patterns into regions corresponding to the classes. Given any additional objects of unknown origin, they can then—after the same set of properties are determined for test as for learning objects—be assigned to the most probable class (according to the region into which the test object's pattern falls).

An important feature of supervised learning is estimation of the probability of misclassification which has to be kept as small as possible. Since a decision on correctness of classification cannot be made with real test objects having unknown origin, the risk of misclassification must be estimated from the learning objects. The simplest way of doing so would be to plug the data on learning objects into the classification rules that were derived before by the same objects. This method, which is called *resubstitution*, is easily and quickly applicable but it suffers from an underestimation of the real risk of errors. A useful alternative to resubstitution is the *leave-one-out method*, which deletes each learning object one at a time, derives the classification rule by means of the remaining objects, and applies this rule to the variable pattern of the left-out object. The number of misclassifications is much more realistic in this case. Formally applied, the leave-one-out method is very time consuming, but for most classification methods, efficient updating formulas exist to give reasonable computing times. The leave-one-out method is a special type of cross-validation, where not one but several objects are left out simultaneously.

Some of the most frequently used classification methods are illustrated in Figure 12. For the purpose of visualization the objects are arranged in a plane (e.g., samples that are characterized by just two trace elements or peak areas, etc.). For simplicity, the discussion is restricted to the two-class problem (filled and unfilled circles represent different classes of learning objects). Figure 12 A relates to the *k-nearest neighbor* (*KNN*) *method*. Given the pattern (coordinates) of any test object (asterisk) the Euclidean distances can be computed between this test object and all the learning objects by using Equation (8.2). For a certain number k (e.g., $k = 3$) the test object is assigned to the class holding the majority under the k smallest distances. In Figure 12 A the test object would be assigned to the filled-circle class. Given a data set, the appropriate number for k is determined to yield minimum misclassification rate by the leave-one-out procedure. The main disadvantage of the KNN method is that all distances must be recomputed for each new test object, which makes this very time consuming for larger data sets.

Figure 12 B illustrates the soft independent modeling of class analogy (*SIMCA*) *method* [16], where each class of learning objects is modeled separately by a principal components approximation (see Section 9.1). In the figure, each of the two learning classes is sufficiently described—up to noise— by a straight line or one principal component, respectively. Now, a test object is classified according to its distance (orthogonal projection) to the approximating linear subspaces. In the figure the asterisk is assigned to the filled-circle group. The optimal

number of principal components for each class is determined by cross-validation.

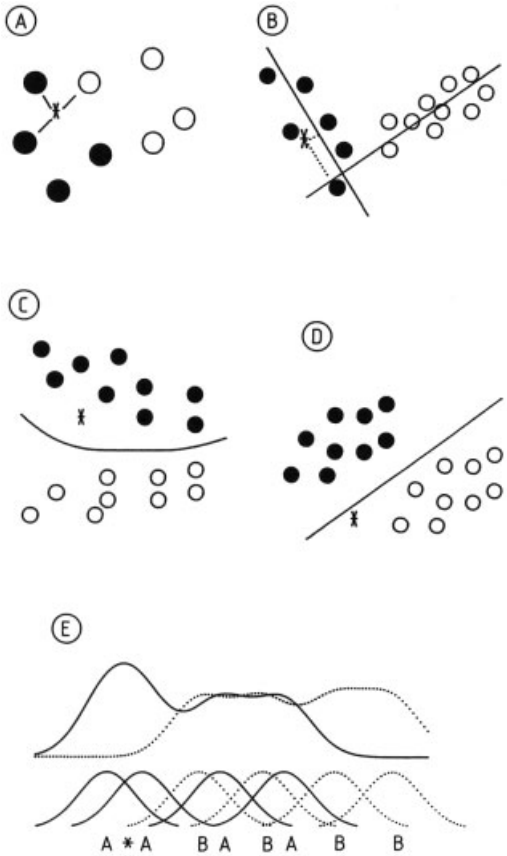

Figure 12. Illustration of supervised classification methods (filled and unfilled circles represent learning objects of different classes; asterisk indicates a test object)
A) The *k*-nearest neighbor (KNN) method; B) SIMCA method; C) Quadratic Bayesian classification; D) Linear Bayesian classification; E) ALLOC method

A classical method based on the assumption of normal distributions is *Bayesian classification* (Figs. 12 C and D). Roughly speaking (with some simplification), a test object is assigned to the class the centroid of which is nearest in the sense of Mahalanobis distance (see Eq. 8.3). Here, centroid refers to an average object having mean values of a certain class in all variables. The Mahalanobis distance is essentially based on the covariance matrix or, in illustrative terms, the shape of the class distribution. In the case of spherical distribution the Mahalanobis distance coincides with the conventional Euclidean distance (Eq. 8.2). In this special case the test object would be assigned to the class with nearest (in the sense of conventional distance) centroid. This trivial classification rule, however, is not appropriate for nonspherical distributions occurring due to correlations among variables. In contrast to this, the Mahalanobis distance yields an optimal classification rule, provided multivariate normality can be assumed. If the Mahalonobis distance is computed for the covariance matrices of each class separately (by assuming different shapes of distribution), a nonlinear decision curve (surface for more than two variables) is generally obtained (see Fig. 12 C, where the test object is assigned to the filled-circle class). This situation is referred to as quadratic classification. Under the assumption of equal shapes of distribution as in Figure 12 D, decision lines are obtained (hyperplanes for greater dimensions). Therefore this method is called linear classification. Here the test object would be assigned to the unfilled-circle class. A drawback of quadratic classification is the need to estimate many parameters (different class covariance matrices). If this estimation is not supplied with enough learning objects, the decision surface or classification rule, respectively, will become rather uncertain. This is why linear classification is often used even if the shapes of the distribution are unequal. Then an equalized covariance matrix is generated by averaging.

Another type of classification is the allocation of test objects *(ALLOC) method* [17], where the densities of class distributions are estimated by potential functions resulting from so-called density kernels. These kernels, usually Gaussian shaped, are constructed around each of the learning objects. Summing them up pointwise for each class separately yields the desired density estimations. Now a test object is assigned to the class having maximum "influence" in it; this means attaining the highest value of density function at the position of the object. In Figure 12 E, where objects are arranged in one dimension for simplicity, the test object would be assigned to class A. Here, solid curves (Gaussian kernels and summed-up potential function) refer to class A and dashed curves to class B. As for KNN (optimal *k*) and SIMCA (optimal number of PCs), some parameter must be adjusted in ALLOC, namely, the smoothing parameter, which determines the flatness of the kernels. Ob-

viously, neither too sharp nor too flat kernels are desirable for optimal class separation. The best decision can be evaluated once again by some cross-validation technique.

11. Multivariate Regression

11.1. Multiple Linear Regression

Multiple linear regression (MLR) is a method that estimates the coefficients $c_1,\ldots,c_p$ in a linear dependence

$$y = c_1 \cdot x_1 + \cdots + c_p \cdot x_p \tag{11.1}$$

of some response variable $\boldsymbol{y}$ on several predicting variables $\boldsymbol{x}_i$. If both predicting variables and response are observed n times, the resulting data can be collected in a data set $\boldsymbol{X}$ (n rows and p columns) and a vector $\boldsymbol{y}$ (n components). As an example, one may consider Table 3 consisting of 8 (data) rows and 8 (data) columns. The first 7 columns correspond to the X-block of values for predicting variables, while the last column contains the observations of the response variable y. Hence, in this table, one has $n = 8$ and $p = 7$. Note that an intercept c_0 is easily included into Equation (11.1) by adding an additional formal predicting variable x_0 which realizes the constant value 1. MLR may be used, for instance, to fit the polynomial model mentioned in Section 5.1 upon putting $x_i := x^i$ for $i = 1,\ldots,n$. The coefficient vector c in Equation 11.1 (p components) is chosen as to minimize the sum of squared deviations (ssd)

$$\text{ssd} = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} c_j x_{ij} \right)^2 \tag{11.2}$$

between observed and predicted responses. It can be obtained by solving the system

$$X^{\mathrm{T}} X c = X^{\mathrm{T}} y \tag{11.3}$$

of linear equations. The purpose of MLR may be twofold, namely, prediction and modelling. Prediction means that, once the coefficient vector c has been determined from n experiments with measured observations of y, Equation (11.1) may then be used to forecast y from new constellations of the predicting variables. In the modelling context, one is more interested in the interpretation of the coefficient vector c itself. For instance, statistical significance of these coefficients may be tested and nonsignificant contributions deleted from the model [5]. Such model reduction may lead not only to less expensive future experiments but also to more reliable predictions. To be more precise, the 95 % confidence interval for the coefficient c_j is calculated as

$$c_j \pm \Delta_{0.95} c_j \quad \text{where}$$
$$\Delta_{0.95} c_j = t_{n-p;0.975} \sqrt{\gamma_i \cdot \text{ssd}/(n-p)} \tag{11.4}$$

where γ_i is the i-th diagonal element of the matrix $(X^T X)^{-1}$ and the t-statistic is used.

A typical application of MLR arises in multicomponent calibration: Given the spectrum of a sample that is a mixture of p components, the concentrations of these components in the mixture sample need to be determined. Digitizing the spectra of the mixture and of the pure components, respectively, by n distinct wavelengths yields vectors $\boldsymbol{y}$ and $\boldsymbol{x}_i$, respectively, which are related according to Equation (11.1) by Beer's law, where the coefficients c_i represent the desired concentrations. The pure components need not even be known exactly. It would be sufficient to start computations with a large set of pure components that are possibly included. In an ideal situation, the correct subset could then be detected by some variable selection technique as mentioned above. A serious disadvantage of MLR is the multicollinearity problem. This problem occurs when all or some of the predicting variables are highly correlated. As a consequence, the estimation of regression coefficients in Equation (11.1) becomes more and more unreliable. The fact may be illustrated by the extreme situation in which all predicting variables are identical and, hence, the response variable may be described by arbitrary coefficients. In the example of spectral data, high correlations are generated by components having similar spectra (strong overlapping). Such situations would lead to unreliable computed concentrations.

11.2. Latent Variable Regression

The multicollinearity problem can be circumvented by application of *principal components regression* (PCR), which, roughly speaking, is MLR applied not to the original set of predicting

variables but to a smaller set of principal components of these variables. Since PCs are always uncorrelated (see Chap. 9), multicollinearities may be avoided. The deleted components, however, although describing only a smaller amount of variance in the set of predictors, might account for a great part of prediction of the response variable that will be lost. A procedure for selecting latent variables as predictors by taking both multicollinearity and predictive ability into consideration is called latent root regression [18].

For multivariate calibration in analytical chemistry, the *partial least squares* (PLS) method [19], is very efficient. Here, the relations between a set of predictors and a set (not just one) of response variables are modeled. In multicomponent calibration the known concentrations of l components in n calibration samples are collected to constitute the response matrix Y (n rows, l columns). Digitization of the spectra of calibration samples using p wavelengths yields the predictor matrix $\boldsymbol{X}$ (n rows, p columns). The relations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are modeled by latent variables for both data sets. These latent variables (PLS components) are constructed to exhaust maximal variance (information) within both data sets on the one hand and to be maximally correlated for the purpose of good prediction on the other hand. From the computational viewpoint, solutions are obtained by a simple iterative procedure. Having established the model for calibration samples, component concentrations for future mixtures can be predicted from their spectra. A survey of multicomponent regression is contained in [20].

12. Multidimensional Arrays

The increasing complexity of experimental designs and hyphenated methods requires a generalization of data tables to higher dimensional arrays. Frequently, a fixed classical objects/variables setting is observed several times under different conditions, yielding a separate data table for each condition. Piecing all these together, one arrives at a cubic or three-dimensional data array, as sketched in Figure 13. In this data example, a set of six trace element concentrations was determined in five different fodder plants for 32 samples (eight parallel determinations in each of four different sampling regions labeled a to d. While the general entry of a data table may be referred to as x_{ij} (see Chap. 8), it must be denoted by x_{ijk} for a three-dimensional array. As before, the indices relate to the corresponding ordered item in each of the three dimensions. In the array of Figure 13, for example, x_{123} represents the concentration of Cd determined in grass for the first sample of region a, where the order of indices is: objects (samples), variables (elements), and "conditions" (fodder plants).

Such arrays raise the question of generalizations of the table-oriented techniques presented in Chapters 9 to 11. The most prominent representatives of higher dimensional factorial methods are the so-called Tucker3 [21] and PARAFAC (parallel factor analysis) [22] models. For three-way arrays, the Tucker3 model is expressed as

$$x_{ijk} \approx \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} c_{pqr} e_{ip} g_{jq} h_{kr} \tag{12.1}$$

where, P, Q, R denote relatively small numbers of components for each of the three dimensions, the e_{ip}, g_{jq}, h_{kr} are entries of so-called component matrices (frequently required to be orthonormal), and the c_{pqr} refer to elements of a so-called core matrix. The approximation sign in Equation (12.1) is to be understood in a least-squares sense. The component matrices contain the scores or loadings of the factors for objects, variables, and conditions, and as such are generalizations of the principal components for data tables [indeed, Eq. (9.2) relates to a special case of Eq.(12.1) when reducing the dimension of the array from three to two]. In particular, choosing $P = Q = R = 2$, the component matrices contain the coordinates for diagrams of objects, variables, and conditions. In this way, generalizations of plots like Figure 10 a can be obtained for three dimensions. The core elements indicate how single factors of different dimensions are linked with each other. Their squared values c^2_{pqr} provide the amount of variation in the data explained by jointly interpreting factors p, q, and r, from objects, variables, and conditions, respectively. As an illustration, one may consult the diagram plots for samples, elements and fodder plants collected in Figure 13. For a brief interpretation of a first factor, one recognizes that all elements and fodder plants are given

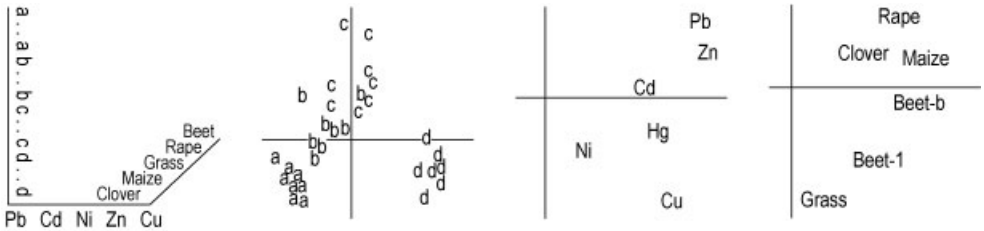

Figure 13. Example of a three-dimensional data array from environmental chemistry (left) and diagram plots for the first two factors of each of the three dimensions (to the right)

some signed (positive) scores on the first axis, which at the same time clearly differentiates between sampling regions in the order d > c, b > a. Hence, this first axis can be interpreted as one of general pollution exposure (not differentiating between elements and plants) with regions affected by pollution according to the just-stated order. A second factor then would differentiate between plants (e.g., grass as opposed to rape) or elements (Pb, Zn as opposed to Cu).

The PARAFAC model is slightly more restrictive, hence simpler, than Tucker 3:

$$x_{ijk} \approx \sum_{p=1}^{P} e_{ip} g_{jp} h_{kp} \tag{12.2}$$

This model is particularly important for analysis of data from hyphenated methods. If, for instance, mixed samples of P chemical components are characterized by spectroscopic-chromatographic measurements, then the signal intensity x_{ijk} may be approximated as in Equation (12.2) by means of the spectral and chromatographic profiles g_{jp}, h_{kp} at wavelength j and retention time k of component p at unit concentration and the concentration e_{ip} of component p in mixture i. A peculiarity of the PARAFAC model is that its decomposition according to Equation (12.2) is unique, in contrast to the accordingly reduced decomposition of two-dimensional tables: whereas for data tables the factors can be rotated without changing the result of approximation, this is not possible for three-dimensional arrays. As a consequence, it is not necessary, for the decomposition to be followed by an appropriate rotation to arrive at chemically meaningful factors (e.g., spectra with nonnegative intensities or nonnegative concentration values). The PARAFAC model based on three-dimensional arrays is in principle able to find the correct and unique decomposition of the given data array in a single step. This fact highlights the importance of the model for qualitative (e.g., identification of spectral and chromatographic profiles of unknown pure components in the mixtures) and quantitative (e.g., concentrations of pure components in the mixture) analysis at the same time.

The Tucker 3 and the PARAFAC models are easily generalized to arrrays of arbitrarily high dimension by using tensorial notation (for a review, see [23]). An introduction to the analysis of multidimensional arrays in the context of chemometrics can be found in [11].

13. References

1. D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte, L. Kaufman: *Chemometrics: a textbook*, Elsevier, Amsterdam 1988.
2. M. A. Sharaf, D. L. Illmann, B. R. Kowalski: *Chemometrics*, John Wiley & Sons, New York 1986.
3. D. C. Montgomery, G. C. Runger: *Applied Statistics and Probability for Engineers*, John Wiley & Sons, New York 1999.
4. L. Sachs: *Angewandte Statistik*, Springer, Berlin 1999.
5. S. Chatterjee, A. S. Hadi, B. Price: *Regression Analysis by Example*, Wiley-Interscience, New York 2000.
6. K. Doerffel: *Statistische Methoden in der analytischen Chemie*, 5th ed., Deutscher Verlag für Grundstoffindustrie, Leipzig 1990.
7. W. J. Youden: "Technique for Testing Accuracy of Analytical Data," *Anal. Chem.* **19** (1947) 946 – 950.
8. S. N. Deming, S. L. Morgan: *Experimental Design: A Chemometric Approach*, Elsevier, Amsterdam 1993.
9. J. C. Miller, J. N. Miller: *Statistics for Analytical Chemistry*, 2nd ed., Ellis Horwood Ltd., Chichester 1989.

10. R. A. Johnson, D. W. Wichern: *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey 1982.
11. B. S. Everitt, G. Dunn: *Applied Multivariate Data Analysis*, Arnold, London 2001.
12. S. Wold, K. Esbensen, P. Geladi, *Chemom. Intell. Lab. Syst.* **2** (1987) 37 – 52.
13. E. R. Malinowski, D. G. Howery: *Factor Analysis in Chemistry*, J. Wiley and Sons, New York 1980.
14. W. H. Lawton, E. A. Sylvestre, *Technometrics* **13** (1971) 617.
15. M. R. Anderberg: *Cluster Analysis for Applications*, Academic Press, New York 1973.
16. S. Wold, *Pattern Recogn.* **8** (1975) 127 – 139.
17. J. Hermans, J. D. F. Habbema: *Manual for the ALLOC Discriminant Analysis Programs*, University of Leiden, Leiden 1976.
18. R. F. Gunst, R. L. Mason: *Regression Analysis and its Applications*, Marcel Dekker, New York 1980.
19. H. Wold in K. G. Jöreskog, H. Wold (eds.): *Systems under Indirect Observation*, North Holland Publ., Amsterdam 1981.
20. H. Martens, T. Næs: *Multivariate Calibration*, J. Wiley and Sons, Chichester 1989.
21. L. R. Tucker, *Psychometrika* **31** (1966) 279 – 311.
22. R. A. Haeshman, *UCLA Working Papers in Phonetics* **16** (1970) 1 – 84.
23. R. Henrion, *Chemom. Intell. Lab. Syst.* **25** (1994) 1 – 23.
24. A. Smilde, R. Bro, P. Geladi: *Multi-way Analysis*, Wiley, New York 2005.

Sampling

James N. Miller, Department of Chemistry, Loughborough University, Loughborough, United Kingdom

1. **Introduction and Terminology** 1
2. **Probability Sampling** 2
3. **Basic Sampling Statistics** 3
4. **Acceptance Sampling** 4
5. **Conclusions** 6
6. **References** 6

1. Introduction and Terminology

The determination of the concentrations of one or more components of a chemical material is a multistage process. Research in analytical chemistry understandably gives emphasis to methods of pretreating materials prior to analysis (e.g., dissolution, concentration, removal of interferences); to the production of standard materials and methods, such as those used to check for bias in measurement; to the development of analytical techniques demonstrating ever-improving selectivity, sensitivity, speed, cost per sample, etc.; and to new methods of optimization and data handling which enhance the information generated by these techniques. Many of these improvements are in vain, however, if the very first stage of the analytical process is not planned and managed properly: this is the *sampling* stage, which should ensure that specimens subjected to analysis are fully representative of the materials of interest. Repeated investigations show that much of the variation encountered when (for example) similar analyses are performed by different laboratories, arises at the sampling stage (see below). This article outlines the general and statistical aspects of the sampling process. Detailed coverage of the chemical aspects of sampling in particular applications is to be found in reviews and monographs [1, 2].

The word *sample* is used in two rather different, but related senses in analytical science. In each case, the general implication is clear—a sample is a small and hopefully representative portion of a larger object. But some differences are also clear. Statisticians use the term to mean a small number of measurements which are assumed to be representative of a theoretically infinite *population* of measurements. Chemists and others use the term to mean a small number of specimens, taken in cases where the materials of interest are too large or too numerous to be examined in their entirety. Thus, a water analyst might take ten specimens of water from a particular stretch of river, knowing or assuming that they form a representative sample of the whole of that part of the river at a particular time. Laboratory staff might then make five replicate pH measurements on each specimen: in each case, these five pH values are a sample representing a potentially infinite number of pH values—the population. (It is conventional to use Greek letters for population statistics, and the Roman alphabet for sample statistics.) In this example, the chemical and statistical uses of the word sample are to some degree distinct. In a different application area, an analyst might remove every hundredth capsule from a pharmaceutical production line, and analyze each capsule once to determine the level of the active component. The concentrations obtained would form a sample (in both chemical and statistical senses of the word) representing the behavior of the production line. The latter example also reminds us that sampling theory is closely related to *time series* statistics, and to *quality control* methods.

Once the overall objectives of an analytical process have been set, the sampling steps must be carefully planned: how many specimens (or *sample increments;* see Chap. 2) should be taken? How should they be selected? How big should they be (in cases of bulk samples)? One difficulty is that in some analyses the nature of the sample and the precision of the sampling and measurement steps may not initially be known. In such cases, simple preliminary sampling plans may be used to provide such infor-

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_059.pub2

mation, and the resulting data can be used iteratively to improve the sampling procedures for long-term use. Good sampling plans should be available in written form as protocols which describe experimental procedures in detail, set criteria on the circumstances in which a sample can be rejected, and list the additional information that should be gathered when a sample is taken (temperature, weather, etc.). In principle, such protocols allow samples to be taken by untrained staff, but it is desirable (if not always practicable) for a sample to be taken by the analytical scientist who is to perform the subsequent measurements.

The terminology of sampling and sample preparation has recently been conveniently summarized in a technical brief published by the Sampling Uncertainty and Quality Sub-Committee of the Analytical Methods Committee of the Royal Society of Chemistry [3].

2. Probability Sampling

Most sampling procedures combine a suitable method of selecting sample items or increments (individual portions) with a knowledge of probability theory. From the analytical data conclusions are drawn on, for example, the risks involved in accepting a particular product. Such methods are known collectively as probability sampling methods. The first step in such a procedure is the selection of the sample items or increments: it is critically important that these are not biased in any conscious or unconscious way, otherwise they will not be representative of the corresponding population. Usually, each part or element of the population must have an equal chance of selection. (This is not absolutely necessary, but if the probabilities of selecting different parts of the population are not equal, those probabilities should be known.) The best method of ensuring such equal probabilities is to use *random sampling.* This method (definitely not the same as haphazard sampling!) involves using a table of random numbers to indicate the items or increments to be selected. The approach can be applied to bulk samples as well as itemized/numbered materials. For example, a truckload of coal could be regarded as being divided into numbered cells by dividing the truck's internal dimensions both horizontally and vertically: the cells to be sampled are then obtained from the random number table. Sampling normally proceeds without replacement, i.e., an item or cell once sampled is not available for further sampling.

An important component of the sampling process—and an example of how it can be developed and improved with time—is to analyze for any systematic relationships within these random samples. For example, samples taken late in the working day may give results that differ significantly from those taken early in the morning. Such a trend, which if unnoticed might bias the results, can be taken into account once identified. Alternative sampling approaches are less satisfactory than a properly controlled random sample. Simplicity suggests that it might be easier, for example, to sample every hundredth item from a production line: but if the manufacturing process has an undesirable periodicity such a regular sampling plan might conceal the problem.

In some applications it is possible to use a *composite sample,* produced by taking several separate increments from a bulk material, and blending them. This process requires care (especially when solids of different particle size, etc., are involved) to produce a single sample that is analyzed several times and assumes that it is only the average composition of the bulk material which is of interest. When replicate measurements are made on this composite sample the sampling variance is greater (i.e., worse) than would have been the case without blending, but there is a benefit in that fewer measurements are necessary—considerations of time and cost often arise in the development of sampling plans.

Some materials to be sampled (e.g., ores or minerals) are often heterogeneous, i.e., the components of interest are distributed through the bulk in a nonrandom way. In such cases it is clearly inappropriate to use simple random sampling, and a more complex approach is necessary. This involves dividing the bulk material into a sensible number of segments or strata, and then (using random numbers, as before) taking samples from each stratum. It is usual to use strata of equal size, in which case the sample size for each stratum is also the same, but if the strata are by nature unequal in size, the number of samples taken should be proportional to these sizes. Such *stratified random sampling* clearly

involves a compromise between the number of strata chosen and the labor involved in the analyses. If too few strata are chosen, it is possible to take acceptable numbers of samples from each, but some inhomogeneities in the bulk material may remain concealed; if too many strata are chosen, the number of measurements necessary is large, if the sampling variance is not to be too high. This is another instance where prior knowledge and/or iteratively gained experience helps to formulate a sensible sampling plan.

A useful qualitative guide to the principles underlying probability sampling is provided by an ASTM standard [4].

In many cases sampling schemes are regulated by statutes, such as those published by individual governments, the European Community, the Codex Alimentarius Commission, and so on. Such schemes are particularly common in the analysis of foodstuffs, and may specify the mass and number of the sample increments to be taken as well as the statutory limits for the analyte(s) under study.

3. Basic Sampling Statistics

The overall random error of an analytical process, expressed as the variance s^2, can be regarded as the sum of two other variances, that due to sampling s_1^2, and that due to the remaining measurement components of the process s_0^2. These variances are *estimates* of the corresponding population variances σ^2 and its components σ_1^2 and σ_0^2. The overall standard deviation s is calculated as usual from single measurements on each of h sample increments, and the confidence limits of the true mean value of the population μ are obtained from:

$$\mu = \bar{x} \pm ts/\sqrt{h} \tag{1}$$

where $\bar{x}$ is the mean of the h measurements, and the t-value is at the appropriate confidence level. This first calculation of the mean and standard deviation allows a sampling plan to be initiated; further measurements then give refined values.

From Equation (1), the variance of $\bar{x}$ is s^2/h, used as an estimate of σ^2/h. To reduce, i.e., improve this variance, it is necessary to take n measurements of each of the h increments. This allows the separation of the sampling and measurement variances by one-way analysis of variance. The replication of the measurements is expected to reduce the measurement variance σ_0^2, and the overall variance of the mean becomes $\sigma_0^2/nh + \sigma_1^2/h$. This relationship provides general guidance on the best practical way of reducing the overall variance. The measurement variance can be minimized by using a high-precision analytical method or by using higher values of n, but in practice the sampling variance is often much larger than the measurement variance, so the choice of h is much more important (see below). If a composite sample is used, i.e., h sample increments blended to form one sample which is then measured n times (see above), the variance of the mean of such replicates is $\sigma_0^2/n + \sigma_1^2/h$. This variance is higher (worse) than when h increments are measured n times each, but the number of measurements is reduced (n as opposed to nh).

When sampling a bulk material, it is clearly important to come to decisions about: (1) how many increments should be taken; and (2) how large they should be. The minimum number of increments h necessary to obtain a given level of confidence is:

$$h = t^2 s_1^2 / E^2 \tag{2}$$

where E is the largest permissible difference between the sample estimate $\bar{x}$ and the corresponding population value μ for the determinand (the equation assumes that each increment is measured once only). The t-statistic is used at the desired confidence level, e.g., 95 % or 99.7 %. Since the value of t depends on h, it is necessary to make an initial estimate of h, using the $h = \infty$ value of t (e.g., $t = 1.96$ for 95 % confidence), thus obtaining from Equation (2) a preliminary value of h: a new value of t can then be taken, and a reliable final value of h approached iteratively. The ASTM standard [5] recommends the use of 99.7 % confidence levels. It should be noted that Equation (2) is only applicable if the determinand is distributed according to the Gaussian, or "normal", distribution: different distributions require separate expressions for h.

When a bulk material is examined, the size of each increment is also of importance. Clearly, if each increment is too large, it may conceal the extent of variation within the bulk material: if it is too small, many increments are necessary to reveal the extent of the sampling variance.

INGAMELLS [6] utilized the fact that s_1^2 decreased as the increment size increased to develop the equation:

$$W R^2 = K_S \tag{3}$$

where W is the increment weight, R is the relative standard deviation of the sample composition, and K_S is a constant. The definition of K_S depends on the user (e.g., it might be the weight necessary to limit the uncertainty in the composition of the material to 2 % with 95 % confidence); it is estimated by carrying out preliminary experiments to determine how s_1 and hence R vary with W. Once K_S is known, the minimum weight W required to produce a particular value of R is readily calculated.

Equations analogous to (2) and (3) can be derived for the more complex situations that arise in stratified random sampling [1, 7]. A computer program has been developed [8] to assist in the solution of sampling problems: it is especially directed at geochemical and other areas where the sizes and shapes of particulate solids may affect R^2.

It is now widely accepted that sampling steps make major contributions to the overall uncertainty of many if not most analytical processes. RAMSEY has provided a major tutorial review of this topic showing how robust analysis of variance methods can be used to distinguish the sampling and analytical components of a typical process. Examples from the area of geochemistry are provided [9].

4. Acceptance Sampling

This chapter deals with the problems associated with the acceptance or rejection of a product by a purchaser, on the basis of an examination of a *sample* taken from the whole *lot* of the product. The general aim of the methods summarized here is to protect both the manufacturer (e.g., of a pharmaceutical product which must contain a specified amount or range of the active ingredient), and also the consumer. The manufacturer is protected if good or valid lots of the product are (wrongly) rejected only infrequently; and the consumer is protected if bad or invalid lots are (wrongly) accepted only infrequently. Since the contents of the components of each lot will show variation (generally assumed to be Gaussian), the term "infrequently" has to be defined numerically (e.g., as 5 % of all occasions), and it is also necessary to give quantitative definition to "good/valid" and "bad/invalid". In some cases, lots are tested by means of a qualitative criterion (sampling by attributes), but more usually a quantitative measure is used (sampling by variables—this approach requires smaller samples than sampling by attributes to achieve a given degree of discrimination), and good and bad are thus defined numerically. Once these definitions are known, it is possible to develop a sampling plan, and describe it by an *operating characteristic* (OC) curve, sometimes known as a power curve, because of its relationship to the concept of the power of a statistical significance test.

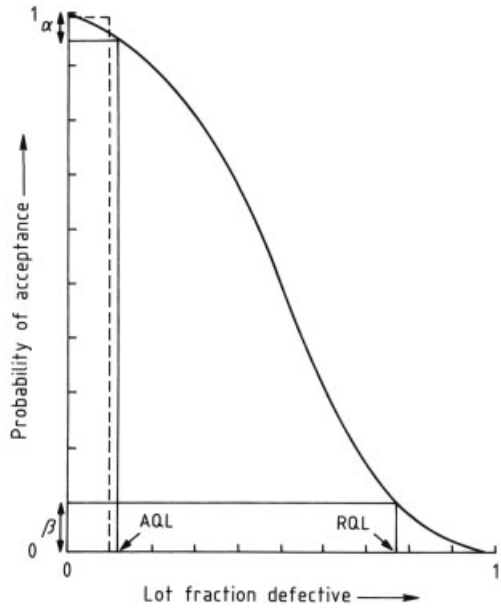

Figure 1. A typical OC curve
The broken line shows the result for an idealized sampling plan

An OC curve for sampling by attributes plots the *probability of acceptance* of a lot (y-axis) against the *lot fraction defective* (x-axis), i.e., the proportion of items in a lot which fail to meet the predetermined "good" criterion. A typical curve is shown in Figure 1. The success of a sampling plan in discriminating between good and bad lots depends on the shape of its curve. An ideal plan

would give the shape shown in the figure as a broken line: the probability of acceptance would be 1.00 until the lot fraction defective reached a predetermined value, and would then fall abruptly to zero. In practice a curve is always obtained, so two points on it are particularly important. The first is called the *acceptance quality level* (AQL): this is the percentage of defective items that is (just) tolerable with a predetermined frequency, e.g., 5 %. A lot with this proportion of defectives will be rejected with a probability α. This is known as the *producer's risk.* It is often set at 5 %, but this is not universal; in some cases, plans are developed so that α = 10 % for small lots, but only 1 % for large lots. A second point of importance is the *rejectable quality level* (RQL), also known as the *lot tolerance percent defective* (LTPD), or the *limiting quality level* (LQL). If β is the probability of acceptance at the RQL, this is clearly the *consumer's risk,* i.e., the risk that the consumer will accept the lot, even though it has a more substantial portion of defective items. Quite commonly, β = 10 %.

To define a sampling plan when the sampling is by attributes, three numbers must be fixed. These are the sample size *h*, the *acceptance number,* and the *rejection number.* The acceptance number, sometimes given the symbol *c*, is the number which must not be exceeded by the number of defective items in the sample if the lot is to be accepted. The rejection number is the number at or above which the number of defectives in the sample leads to rejection of the lot. It might seem that, by definition, the rejection number is *c* + 1, but this is not so, if the sampling plan is a *multiple* one, i.e., a plan in which a second or subsequent sample might be taken, depending on the results obtained for the first sample. In such cases, if the number of defectives in the first sample did not exceed *c*, the lot would be accepted; if the number of defectives exceeded the rejection number (to be defined, but > c), the lot would be rejected; and if the number of defectives fell between these two extremes, one or more further samples might be taken. It is easy to see that multiple sampling plans might be highly discriminating, but the price paid is the much larger number of measurements needed, and such plans are not very widely used.

In the case of sampling by variables, the criteria needed to define the OC curve are inevitably different, and can best be illustrated by a simple example. Suppose that a manufactured fluorescent dye is of use to a laboratory analyst only if it contains at least 99 % of the active fluorophore. Any lot containing $\geq$ 99 % is acceptable, but the analyst is reluctant to buy the dye if the fluorophore content falls below 96 %. It is necessary to quantify these requirements: let us say that the manufacturer requires that a dye lot containing 99 % fluorophore must have a 95 % probability of acceptance, and the analyst requires that the probability of accepting a lot containing only 96 % fluorophore is only 1 %. It is also postulated that the sample size is *h*, and that the standard deviation σ for the fluorophore analysis, which reflects both sample inhomogeneities and measurement variation, is known exactly. (This is the case because of long experience of the production and testing of the dye.) It is necessary to estimate *h* and it is also stipulated that a dye lot will be accepted whenever the average fluorophore content over these *h* measurements equals or exceeds a value *k*, which is also to be estimated.

Suppose first that the true fluorophore content of a lot is 99 %. The sample average $\bar{x}$ is then normally distributed, with mean 99 and standard deviation σ/h. The probability of acceptance of the lot is the probability that $\bar{x} > k$; this is equal to the probability that the standard normal deviate:

$$z = (\bar{x} - 99)/\left(\sigma/\sqrt{h}\right) > (k-99)/\left(\sigma/\sqrt{h}\right)$$

This probability has been defined as 0.95, so from the standard statistical tables for *z*, the result is:

$$(k-99)/\left(\sigma/\sqrt{h}\right) = -1.64$$

In the alternative situation, suppose that the true fluorophore content of the dye lot is only 96 %. By exactly analogous arguments:

$$(k-96)/\left(\sigma/\sqrt{h}\right) = +2.33$$

the latter number again being obtained from the tables for *z* at a probability of 0.99. Combining these results:

$$(k-99)/(k-96) = -(1.64/2.33) = -0.704$$

from which *k* = 97.76. If it is assumed that σ = 2.0, *h* can be obtained from either of the two above relationships. Thus:

$$(97.76-99)/\left(2.00/\sqrt{h}\right) = -1.64$$

from which h is 7, almost exactly. The sampling plan thus consists of taking seven measurements, and rejecting a lot whenever the average fluorophore content does not reach 97.76 %. A noteworthy feature of these calculations is that σ does not have to be known for k to be determined, so, as in other forms of sampling, a reasonable plan can be obtained, even when σ is initially unknown. It is possible to make a conservative estimate of σ, use it along with k to estimate h, and refine the estimate with increasing experience of the variance.

Sampling plans of this kind generate OC curves of the same form as those described above, although in cases of sampling by variables the x-axis is normally plotted as the mean determinant concentration (the fluorophore level in the example above). Other aspects of OC curves, of multiple sampling plans, and of *sequential* sampling plans, in which samples are taken one at a time, are dealt with in more advanced texts [10 – 12].

5. Conclusions

This brief survey of sampling and its application in various circumstances should indicate that the subject is of critical importance to the performance of the overall analytical process, whether the latter is laboratory based or process based. It is also a very good example of the principle that methods based on statistics and probability can, and should, be applied in the planning stage of an analysis—to use such methods only when data have been obtained is almost invariably too late!

6. References

1. B. Kratochvil, J. K. Taylor, *Anal. Chem.* **53** (1981) 924 A – 938 A.
2. D. T. E. Hunt, A. L. Wilson: *The Chemical Analysis of Water*, 2nd ed., Royal Society of Chemistry, London 1986.
3. AMC Technical Brief 19, Terminology – the Key to Understanding Analytical Science, Part 2: Sampling and Sample Preparation, Royal Society of Chemistry, London 2005; http://www.rsc.org.
4. ASTM Standard E 105–04, Standard Recommended Practice for Probability Sampling of Materials, ASTM, Philadelphia 2004.
5. ASTM Standard E 122–72, Standard Recommended Practice for Choice of Sample Size to Estimate with a Specified Tolerable Error the Average quality of a Lot or Process, ASTM, Philadelphia 2000.
6. C. O. Ingamells, *Talanta* **21** (1974) 141 – 146.
7. V. Barnett: *Elements of Sampling Theory*, Hodder and Stoughton, London 1974.
8. P. Minkkinnen, *Anal. Chim. Acta* **196** (1987) 237 – 245.
9. M. H. Ramsey, *J. Anal. At. Spectrom.* **13** (1998) 97–104.
10. D. C. Montgomery: *Introduction to Statistical Quality Control*, Wiley, New York 1985.
11. E. L. Crow, F. A. Davis, M. W. Maxfield: *Statistics Manual,* Dover, New York 1960.
12. D. L. Massart, B. G. M. Vandeginste, L. M. Buydens, S. De Jong, P. J. Lewi, J. Smeyers-Verbeke, *Handbook of Chemometrics & Qualimetrics*, **Part 1**, Elsevier, Amsterdam 1997, Chapter 20.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Sample Preparation for Trace Analysis

Standard Article

Jutta Begerow[1] and Lothar Dunemann[2]

[1]Hygiene-Institut des Ruhrgebiets, Gelsenkirchen, Germany

[2]Hygiene-Institut des Ruhrgebiets, Gelsenkirchen, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.b05_065.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. A Strategy Appropriate to Trace Analysis
1.2. Avoidance of Systematic Errors
1.2.1. Trace Losses and Contamination
1.2.2. Uncertainty
2. Sample Preparation and Digestion in Inorganic Analysis
2.1. Sample Treatment after the Sampling Process
2.1.1. Stabilization, Drying, and Storage
2.1.2. Homogenization and Aliquoting
2.1.3. Requirements with Respect to Materials and Chemicals
2.2. Sample-Preparation Techniques; General Considerations
2.2.1. Special Factors Associated with Microwave-Assisted Digestion
2.2.2. Safety Considerations
2.3. Wet Digestion Techniques
2.3.1. Wet Digestion at Atmospheric Pressure
2.3.2. Pressure Digestion
2.3.2.1. Thermally Convective Pressure Digestion
2.3.2.2. Microwave-Assisted Pressure Digestion
2.4. “Dry” Digestion Techniques
2.4.1. Combustion in Air
2.4.2. Combustion in Oxygen
2.4.3. Cold-Plasma Ashing
2.4.4. Fusion
2.5. Illustrative Examples
2.5.1. Sample Preparation as a Function of Analytical Method
2.5.2. Combined Use of Multiple Decomposition Techniques
2.5.3. Comparative Merits of the Various Sample-Preparation Techniques
2.5.4. Decomposition Procedures for Determining Nonmetals
2.6. Evaluation Criteria
2.6.1. Completeness
2.6.2. Uncertainty
2.6.3. Time Factors
2.6.4. The Final Result
2.7. Concentration and Separation of Inorganic Trace Materials
2.8. Automation and Direct Analysis
2.8.1. Automation
2.8.2. Direct Analysis
2.9. Analysis of Element Species

3. Sample Preparation in Organic Analysis
3.1. Sample Treatment after the Sampling Process
3.1.1. Stabilization, Drying, and Storage
3.1.2. Homogenization and Aliquoting
3.1.3. Requirements with Respect to Materials and Chemicals
3.2. Separation of the Analyte
3.2.1. Hydrolysis
3.2.2. Liquid - Liquid Extraction
3.2.3. Soxhlet Extraction
3.2.4. Microwave-Assisted Solvent Extraction
3.2.5. Supercritical Fluid Extraction (SFE)
3.2.6. Solid-Phase Extraction (SPE)
3.2.7. Solid-Phase Microextraction (SPME)
3.2.8. Stir-Bar Adsorptive Extraction (SBSE)
3.2.9. Miscellaneous Techniques
3.3. Headspace Techniques
3.3.1. Static Headspace Technique
3.3.2. Dynamic Headspace Technique (Purge and Trap)
3.4. Determination of Trace Organic Materials in Air Samples
3.5. Analyte Concentration
3.6. Derivatization
3.7. Coupled Techniques

Trace analysis is a very relevant and applications-oriented branch of analytical chemistry. The sample preparation for trace analysis must be custom-tailored to the problem at hand. Systematic errors can arise by contact with vessel materials, reagents, or the ambient atmosphere, as well as any change in chemical or physical state.

In inorganic analysis, sample preparation has to meet the requirements for a substantially trouble-free determination of the analyte. Digestion of the matrix (microwave digestion, wet digestion, dry digestion techniques) and subsequent careful comparison of several decomposition techniques is therefore an essentially important step. Some separation and concentration techniques of the analytes are liquid –liquid extraction, solid-phase extraction, special precipitation reactions, and electrolytic deposition. The introduction of laboratory robots should make it possible to incorporate a significant degree of automation into the time-consuming, labor-intensive area of sample preparation as well, leading to more efficient, reliable, and reproducible sample work-up.

The goal of sample preparation in organic trace analysis is to isolate the analyte from the sample matrix (e.g., liquid-liquid extraction, Soxhlet extraction, microwave-assisted solvent extraction, steam distillation) and then concentrate it and convert it into a form suitable for analysis by the selected method. Separation and concentration of an analyte must often be followed by some type of derivatization. Various coupled sample preparation and determination processes are increasingly utilized in trace organic analysis.

[Top of Page]

1. Introduction

Modern analysis begins with a definition and outline of the problem and ends only after a detailed critical evaluation of the relevant analytical data is complete, permitting the presentation of a “result” (⟶ Analytical Chemistry: Purpose and Procedures). The analyst must therefore retain the ability to monitor a sample conscientiously and knowledgably throughout the analytical process. Only the analyst is in a position to assess the quality of a set of results and the validity of subsequent conclusions, although defining the problem and presenting the conclusions is almost always a cooperative multidisciplinary effort.

Analysis, and particularly trace analysis, thus entails more than the mere qualitative or quantitative detection of a particular element or chemical compound. For example, it presupposes knowledge on the part of the analyst with respect to the origin and structure of the sample matrix. The analyst must also possess a specialist's insight into analogous problems from other disciplines in order to assure the plausibility of the questions raised and critically evaluate and interpret the results, at least in a provisional way.

From these preliminary observations it will already have become clear that trace analysis, the focus of much concern in sample preparation, cannot be regarded as an end in itself. Rather, it is a very relevant and applications-oriented branch of analytical chemistry generally, one that, from a historical perspective, developed from within but became independent of chemical analysis as a whole, which was long regarded simply as a servant to the traditional subspecialities of chemistry and other disciplines (⟶ Analytical Chemistry: Purpose and Procedures). Nevertheless, the same tools, equipment, and methodological principles remain common to both general chemical analysis and modern trace analysis.

1.1. A Strategy Appropriate to Trace Analysis

Trace analytical efforts directed toward the determination of particular elements or compounds require the analyst

steadfastly to pursue a single six-step strategy consisting of

1. Rigorous definition of the problem
2. Problem-oriented selection of a sample
3. Appropriate sample preparation
4. Quantitative determination of the analyte(s)
5. Plausibility tests with respect to the results, together with further evaluation
6. Professional interpretation and presentation of the findings

This strategy must be regarded as an indissoluble whole, with all the constituent parts remaining insofar as possible and practicable in the hands of a responsible trace analyst. Accordingly, it is absolutely essential that the analyst begins with a clear understanding of the problem — if necessary even formulating it. Only after the problem has been defined is it possible to establish the most appropriate approach to sampling and sample preparation. Sampling and sample preparation are followed by the determination itself in the strictest sense of the term, the outcome of which — an analytical result — must be checked for accuracy and plausibility before the findings can be presented. In many cases, the findings will require interpretation developed with the help of specialists from other disciplines, who in turn must familiarize themselves with the analyst's procedures.

Sample preparation plays a central role in the process, but it too often leads a “wall-flower” existence, with primary attention being directed to the determination step. This sense of priorities is reflected all too conspicuously in the equipment and investment planning of many analytical laboratories. However, a welcome trend in recent years points toward fuller recognition of the true importance of sample preparation in the quest for high-quality analytical results and valid conclusions [1], [2].

1.2. Avoidance of Systematic Errors

Trace analysis can satisfy the stringent requirements outlined above only by recognizing and minimizing possible sources of error. Unfortunately, the trace analyst almost always suffers from a lack of knowledge of the correct (or “true”) analytical value. Nevertheless, it is precisely this fact that is the source of the analyst's peculiar responsibility with respect to the results in question. A purely theoretical examination of the dilemma provides little guidance for routine work, since it leads simply to a recognition that the “true” value represents an unachievable ideal.

At this point only a general consideration is feasible regarding possible sources of systematic errors that might influence the correctness of an analysis. Every sample preparation technique is associated with its own particular advantages as well as specific sources of error. In practice, the complete exclusion of systematic errors is possible only in rare cases, because optimization at one stage tends to cause problems elsewhere.

Generally speaking, any contact with vessel materials, reagents, or the ambient atmosphere, as well as any change in chemical or physical state, might result in systematic errors. It is therefore crucial that sample preparation be custom-tailored to the problem at hand with particular regard for the nature of the determination step.

Systematic errors arising from contact between a sample and various materials and chemicals are discussed in the context of the analyte (Sections Sample Treatment after the Sampling Process and Sample Treatment after the Sampling Process), because different analytes require different types of treatment.

1.2.1. *Trace Losses and Contamination*

Trace *losses* are particularly common in the case of highly volatile analytes and as a consequence of adsorption effects, whereas trace *contamination* with otherwise prevalent elements and compounds may arise from laboratory air, vessels, chemicals, and various desorption effects.

The loss of trace analytes can be minimized by conducting all operations in a system that is almost completely sealed hermetically from the ambient atmosphere, and by using vessel materials characterized by a small effective surface area [3]. Surfaces should be free of fissures and preconditioned as necessary to minimize physical adsorption of the analyte (e.g., through ion exchange or hydrophobic interactions). It is also important to consider the possibility of diffusion into or through the vessel walls.

Addition of a radioactive tracer may make it possible to monitor losses that occur during the course of an analysis [4] (⟶ Radionuclides in Analytical Chemistry). Isotope dilution analysis is also useful for establishing the correctness of a result [5], [6].

Attention must be directed toward contamination in the early stages of laboratory planning and design to ensure that areas and rooms of equivalent sensitivity are situated adjacent to each other and shielded from less sensitive areas by appropriate structural and ventilation provisions. Analyses involving concentrations below the mg/kg or mg/L range, and especially in the ng or pg range, should of course be conducted in a clean-room environment. Precise requirements with respect to ventilation vary from case to case.

1.2.2. *Uncertainty*

The uncertainty on the result arises from both random and systematic effects but in trace analysis systematic effects largely determine the uncertainty of an analytical result. The search for and correction of systematic errors is therefore an important responsibility of every trace analyst. Even after correction for systematic errors the uncertainties on there corrections need to be evaluated and included in the overall uncertainty. Failure to correct for systematic errors leads to the considerable scattering frequently observed with collaborative analyses, and ultimately to inaccurate results. The uncertainty on the result increases disproportionately with decreasing amounts of analyte in the sample.

In trace analysis it is not sufficient simply to report a level of reproducibility for the actual determination of an analyte. Evaluating the quality of an analysis requires a knowledge of the reproducibility and the uncertainty arising from systematic effects (⟶ Chemometrics). Errors in sampling and/or sample preparation may be orders of magnitude greater than the standard deviation observed in several repetitions of a determination.

Whenever possible, reference materials should be included in the experimental plan as a way of checking for bias [7]. However, it is also important that the chemical state and environment of the analyte be as nearly identical as possible in both reference material and sample. The method of *standard addition* (where the spike should again be in the same chemical state as the native analyte) is particularly useful in a search for systematic errors, as is a comparison involving different analytical methods.

Internal and external quality control measures are a “must” in all trace analytical procedures, and an analytical result can only be accepted if all necessary actions in this respect have been carefully considered and the uncertainty on the result evaluated (Chapter Introduction). Quality control may therefore not only be a part of the determination step itself, but must also include the full analytical procedure including sample preparation. Internal quality control means the intra-laboratory procedure for quality control purposes including the control of uncertainty, whereas external quality control means the inter-laboratory comparison of the results of analyses performed on the same specimen by several independent laboratories (according to the definition of [8]).

[Top of Page]

2. Sample Preparation and Digestion in Inorganic Analysis

The object of sample preparation in inorganic analysis is to meet the requirements for a substantially trouble-free determination of the analyte. The most important of these with respect to trace analysis in any matrix include:

1. Conversion of the sample into a form consistent with the determination (dissolution)
2. Destruction of the matrix (digestion)
3. Isolation of the analyte from interfering substances that may be present (separation)
4. Enrichment of the sample with respect to trace analytes (concentration)

The dissolution step is designed to compensate for inhomogeneities in the sample. Dissolution and digestion also simplify the subsequent calibration step, ensuring that both the sample and the calibration solutions are in essentially the same chemical and physical state. The extent to which matrix constituents interfere in the determination process is significantly reduced by digestion, leading to a lower limit of detection for the determination. Digestion also facilitates concentration and separation steps.

2.1. Sample Treatment after the Sampling Process

2.1.1. *Stabilization, Drying, and Storage*

In trace analysis, sampling must always be followed by an appropriate stabilization step, with due regard for the nature of the matrix and the analyte.

Aqueous samples such as drinking water, surface water, and waste water but also beverages and urine samples, should always be acidified with mineral acids for stabilization purposes immediately after collection. This is especially true for the prevention of desorption processes during sampling and storage of samples in the course of trace metal analysis. Acidification reduces the tendency for ions to be adsorbed onto active sites at the surface of the containment vessel, and it also inhibits bacterial growth [9]). Glacial acetic acid and 65 % nitric acid have been shown to be suitable for this purpose. Approximately 1 mL of acid should be added per 100 mL of sample. Acidified urine in appropriate containers can be safely kept in a deep-frozen state for several months [10], [11]. Stored urine samples often deposit a sediment that may include considerable amounts of analyte (e.g., arsenic, antimony, copper, chromium, mercury, selenium, zinc) [10].

Loss of water from aqueous matrices (e.g., tissue, fruit, vegetables, and soil samples) may occur during storage. For this reason analytical results should always be reported in terms of dry mass to avoid false interpretations. Drying is best conducted immediately after sampling. Water removal can be accomplished by oven drying at elevated temperature, use of desiccating materials, or freeze drying (lyophilization). Freeze drying has been shown to be the most satisfactory procedure since it minimizes the loss of highly volatile elements and compounds. Drying at a temperature as low as 120 °C can result in the loss of up to 10 % of most elements, and losses for mercury, lead, and selenium may be considerably higher (e.g., 20 – 65 %) with certain matrices [12], [13].

Blood samples must be rendered incoagulable by the addition of an anticoagulant immediately after sampling. Suitable substances for this purpose include salts of ethylenediaminetetraacetic acid (EDTA), citric acid, oxalic acid, and heparin. A particular anticoagulant should be selected with reference to the analyte in question as well as the proposed analytical procedure. Particular attention should be directed to ensuring that the anticoagulant is not itself contaminated with the analyte. For recommendations concerning appropriate levels of anticoagulants see, for example, [14]. Blood sampling kits that already contain the aforementioned anticoagulants are now available commercially. Blood samples in suitable containers can be either refrigerated or deep-frozen prior to analysis. Such samples can normally be kept for several months at – 18 °C without measurable change in the concentration of an analyte. One exception is mercury; in this case reduction to elemental mercury leads to significant volatilization losses within as little as a few days [15]. If storage over several years is envisaged, the sample should either be dried at – 18 °C or quick-frozen in liquid nitrogen at – 196 °C and then stored at or below – 70 °C.

Determination of an analyte in serum or plasma requires that the analyte be isolated subsequent to sampling. This is no longer possible after deep-freezing because of the hemolytic nature of blood samples. In the case of elements at very low

concentration it is preferable that an investigation be conducted on serum rather than plasma. This reduces the risk of contamination, because unlike blood plasma, serum is recovered without addition of an anticoagulant.

2.1.2. *Homogenization and Aliquoting*

In solid samples, elements are normally distributed in an inhomogeneous way. Trace-element determinations are usually restricted to relatively small samples, which requires that a fairly large sample be comminuted and homogenized prior to removal of an aliquot for analysis.

Comminution of the sample presents significant opportunity for contamination. Contamination of the sample from abrasion of the comminution equipment is fundamentally unavoidable, so efforts must be made to select the best possible equipment for each particular analytical task. Equipment is preferred in which the sample comes into contact only with surfaces fabricated from such high-purity plastics as polytetrafluoroethylene (PTFE), since this permits the sample to be used without restriction for the determination of a large number of elements. Friability can be increased by deep-freezing or drying the sample prior to comminution.

2.1.3. *Requirements with Respect to Materials and Chemicals*

Whenever samples, standard solutions, or reagents come into contact with containers, apparatus, or other materials, two opposing physicochemical processes occur at the interface: *adsorption* and *desorption* of ions and molecules.

Ion adsorption at active sites on a contact surface may lead to losses on the order of several nanograms per square centimeter. The importance of this factor therefore increases as the concentration of the analyte decreases. Adsorptive losses are a function of the nature and valence state of the analyte, the nature and concentration of accompanying ions in the solution, pH, temperature, and the duration of contact. Factors related to the containment vessel include the material from which it is constructed, its surface area, and the nature of any pretreatment.

Adsorption can be minimized through the use of quartz, PTFE, polypropylene (PP), perfluoroalkoxy resin (PFA), or glassy carbon vessels. There is no single ideal material, but quartz is preferred in the case of biological samples provided the use of hydrofluoric acid can be avoided. The surface area of the container and the contact time should be minimized, whereas the analyte concentration should be as high as possible. Cations are adsorbed less strongly from acidic solutions than from solutions that are neutral or basic because such ions are displaced from active sites on the container surface by the much more abundant protons . Adsorptive losses are also lower from solutions with high salt concentrations (e.g., urine) relative to those with low salt concentrations, such as standard solutions. Adsorption can be further reduced by adding a suitable complexing agent so long as this has no adverse affect on the subsequent analysis. For example, mercury can be stabilized as $[HgI_4]^{2-}$, and silver as $[Ag(CN)_2]^-$.

Containers, pipettes, and other laboratory implements contain traces of various elements both in their structure and adsorbed on the surface, and these are subject to desorption on contact with a sample, calibration solutions, or reagents, thereby contributing to a nonreproducible increase in blank results. Table 1 summarizes the inorganic impurities likely to be encountered with various vessel materials.

Table 1. Typical inorganic impurities in selected vessel materials [16], [17]

Element	Content, mg/kg				
	Quartz	GC [a]	PTFE [b]	PE [c]	BS [d]
V		4			< 2
Cr		0.5 – 5	10 – 30	15 – 300	< 3
Mn	0.01	0.1	60	10	< 6
Fe	0.2 – 0.8	1 – 15	10 – 35	600 – 2100	200
Co	< 0.001	< 0.01	0.3 – 1.7	0.1 – 5	< 0.1
Ni		0.5			< 2
Cu	< 0.01 – 0.07	0.2	20	4 – 15	1
Zn	< 0.01	< 0.3	8 – 10	25 – 90	2 – 4
As	0.0001 – 0.08	0.06			0.5 – 22
Cd		< 0.01			1
Sn		25 – 50			< 4
Sb	< 0.001	< 0.01	0.4	0.2 – 5	7 – 9
Hg	0.001	0.001			
Pb		0.4			

[a] Glassy carbon.
[b] Polytetrafluoroethylene.
[c] Polyethylene.
[d] Borosilicate glass.

Careful and thorough cleaning of all vessels is extremely important in trace analysis both for inorganic and organic

analytes. Materials must be specially selected with a specific view to the requirements of each task during sample handling and storage, with explicit attention being paid to trace purity. Thorough pretreatment (rinsing, boiling, steaming) with pure nitric, sulfuric, or hydrochloric acid in the case of metal analysis is strongly recommended [18-21]. Figure 1 illustrates a steaming apparatus useful for cleaning analytical vessels which are to be prepared for use in inorganic analyses, namely metals but also other inorganic analytes [22]. For organic analyses repeated cleaning of vessel surfaces with suitable solvents is essential.

One-way vessels and pipettes can be an appropriate means to reduce the risk of contamination during sampling and sample handling. Nevertheless, the blank values of one-way materials have also to be tested on a regular basis.

Figure 1. Steaming apparatus for cleaning analytical vessels with nitric acid, hydrochloric acid, or water (with permission from Kürner, Rosenheim/Oberbayern, Germany)

a) Reflux condenser; b) Steam chamber; c) Rack of quartz tubes; d) Overflow; e) Round-bottom flask (1 – 2 L capacity); f) Heating mantle

Reagents for trace analytical applications must also satisfy stringent purity requirements, often exceeding those of commercially available reagents. Acids such as hydrochloric, nitric, sulfuric, hydrofluoric, and perchloric can be prepared relatively easily in the laboratory to a high standard of purity by distillation below the boiling point ("subboiling process"). An apparatus for this purpose is illustrated in Figure 2. Table 2 compares the elemental contents of commercially available acids with those of acids prepared by the subboiling process.

Table 2. Elemental content of selected acids with varying degrees of purity [16]

Acid	Elemental content, µg/L				
	Cd	Pb	Cu	Fe	Zn
10 N HCl, analytically pure	0.1	0.5	1.0	100	8.0
10 N HCl, extra pure	0.03	0.13	0.2	11.0	0.3
2 N HCl, subboiled	0.01	< 0.05	0.07	0.6	0.2
15 N HNO_3, analytically pure	0.1	0.5	2.0	25.0	3.0
15 N, HNO_3, extra pure	0.06	0.7	3.0	14.0	5.0
15 N, HNO_3, subboiled	0.001	< 0.02	0.25	0.2	0.04

Figure 2. Quartz apparatus for purifying acids or water by subboiling distillation (with permission from Kürner, Rosenheim/Oberbayern, Germany); this device is also available in PFA for the subboiling distillation of hydrofluoric acid (with permission from Berghof Maassen, Eningen u. A., Germany)

a) Heater; b) Cold-finger condenser

All chemicals and reagents should be checked regularly for purity, and replaced as necessary. It is also important to remember the ever-present danger of contamination during reagent use as a result of interactions with container materials and the entrainment of impurities. For this reason reagents and standard solutions should always be prepared immediately before their use. As a general rule, the addition of reagents should be kept strictly to a minimum.

Possible contamination due to atmospheric dust and particle emission attributable to laboratory staff activities (smoking, use of cosmetics, etc.) can be avoided in trace analytical determinations at the µg/kg, µg/L, and µg/m^3 ranges by working under clean-room conditions. A laminar-flow "clean bench" ensures that particles with a diameter > 0.1 µm will be removed with an efficiency of 99.9 % through high-efficiency,submicron-particulate air filtration. As indicated in Table 3, this can result in purification factors on the order of 100 with respect to such elements as manganese and vanadium [23].

Table 3. Typical levels of contamination by selected elements as a result of dust in the laboratory [23]

Element	Contamination level, ng cm^{-2} d^{-1}

	Conventional laboratory	Clean workstation
Al	0.82	0.09
V	0.037	0.0003
Mn	0.063	0.002
Fe	1.5	0.2
Co	0.0036	0.0003
Cu	0.032	0.012
As	0.07	0.03
Sb	0.013	0.001

2.2. Sample-Preparation Techniques; General Considerations

Sample-preparation techniques are classified, somewhat arbitrarily, according to external conditions. Categories include wet digestion, dry ashing, pressure digestion, microwave digestion, and cold-plasma ashing. However, these conventional designations are often imprecise or even misleading with respect to the actual mechanism of the process. Several very different names are sometimes applied to a single technique, which presents a considerable obstacle for anyone (particularly a nonspecialist) interested in acquiring a quick overview of systems applicable to a specific task. For example, the expression “pressure digestion” is inappropriate, because it is a relatively high boiling temperature that ensures more effective digestion, not the associated high pressure. The conventional designation should nevertheless be retained in this case, if only because of its wide acceptance; attempting to rename the procedure now would introduce more confusion than clarity.

Another — very timely — example of incorrect terminology involves uncritical use of the expression “microwave digestion” for both acid digestion with microwave excitation and cold-plasma ashing. Although both techniques make use of microwave radiation, the direct effects of this radiation are of minor importance at most. Expressions like “microwave excitation,” “microwave induction,” or “microwave assistance” would be preferable.

2.2.1. *Special Factors Associated with Microwave-Assisted Digestion*

Rapid, straightforward digestion processes have long been in demand in a wide variety of disciplines. The disparity between the several hours required for a conventional digestion process and the few seconds or minutes involved in elemental determination has become increasingly obvious in recent years with the increasing automation of analytical procedures. Recently introduced rapid digestion processes, which have profited greatly from modern microwave technology, can contribute to reducing this disparity.

The first published report on microwave ashing was that of KOIRTYOHANN et al., which appeared in 1975 [24]. Initially it attracted little attention because of the lack of appropriate expertise and apparatus.

Modern commercial microwave ovens operate at a frequency of 2.45 GHz. The corresponding energy is so low that such microwaves are incapable of rupturing molecular bonds directly, leading only to rotational excitation of dipoles and molecular motion associated with the migration of ions. No vibrational or electronic excitation occurs [25], [26]. It is for this reason that the designation “microwave-assisted digestion” has been suggested. Microwave energy introduced as radiation is dissipated (converted into heat) by ionic conduction and dipole rotation. Oscillating alignment and subsequent relaxation of dipoles occurs about 5×10^9 times per second at a frequency of 2.45 GHz, causing rapid heating of the sample.

Microwave radiation is generated by means of a magnetron (microwave diode). The radiation is then directed through a waveguide into the interior of the oven. Even distribution of the radiation is facilitated by a “mode stirrer” that prevents the development of standing waves, as well as a turntable and perhaps a rotating antenna located beneath the oven floor.

Starting in the late 1980s, microwave-based techniques became more and more available for analytical purposes [27-30]. In trace metal and non-metal analyses, microwave-assisted digestion has been successfully applied to the decomposition of broad range of organic matrix components in geological and biological samples as well as analysis and residue monitoring [22].

2.2.2. *Safety Considerations*

Necessary protective measures with respect to the laboratory workforce must be implemented for the handling of acids and other strongly oxidizing or poisonous chemicals. This includes protective clothing and an efficient fume extraction system. Perchloric acid as a digestion reagent offers advantages due to its powerful oxidizing action, but it should be used only when absolutely necessary since it entails extensive safety precautions (e.g., working in a special hood), and a definitive set of regulations must be scrupulously observed.

Special statutory regulations apply to operations involving pressure vessels, and pressure digestion techniques must be carried out with exceptional care and attention. Ashing an organic matrix can cause the pressure to rise to extremely high levels within a short period of time, and the reaction may become uncontrolled. The corresponding risk of a ruptured pressure vessel cannot be overemphasized. When developing a pressure digestion method it is important to begin with a small sample ($\leqq$100 mg dry mass; even less for pure carbon).

Additional safety considerations apply to the use of microwave technology. First, the equipment should be operated in strict accordance with instructions to ensure that no detectable amounts of microwave radiation will escape. Leak tests should be conducted on a regular basis by a trained technician. Vessels or other objects made of metal should never be placed in the oven chamber. Microwave equipment must also never be operated empty (i.e., in the absence of a sample solution, water, acid, etc.), since the magnetron might be destroyed by reflected radiation. The behavior of unknown samples should be tested on a small scale (see above).

2.3. Wet Digestion Techniques

Wet digestion with oxidizing acids is the most common sample-preparation procedure [31], [32]. This category can be extended to include processes involving bases or nonoxidizing acids as ashing reagents. Concentrated acids with the requisite high degree of purity are available commercially, but they can be purified further by subboiling distillation (see Section Requirements with Respect to Materials and Chemicals).

The most suitable acid for a digestion is a function of the sample matrix, the analyte, and the proposed determination method. Nitric acid is an almost universal digestion reagent, since it does not interfere with most determinations and is available commercially in sufficient purity. However, nitric acid has a rather low boiling point (122 °C), and its oxidizing power is often insufficient under atmospheric-pressure conditions. Hydrogen peroxide and hydrochloric acid can usefully be employed in conjunction with nitric acid as a way of improving the quality of a digestion. Hydrochloric acid and sulfuric acid may interfere with the determination of certain metals through the formation of stable compounds. As noted previously (Section Safety Considerations), safety considerations are particularly important when using perchloric acid. Silicate samples require the further addition of hydrofluoric acid.

2.3.1. *Wet Digestion at Atmospheric Pressure*

This category includes all the various ambient-pressure wet digestion techniques involving concentrated acids (or bases). This very inexpensive technique is of inestimable value for routine analysis because it can easily be automated; all the relevant parameters (time, temperature, introduction of digestion reagents) lend themselves to straightforward control [33].

Systems of this type are limited by a low maximum digestion temperature, which cannot exceed the ambient-pressure boiling point of the corresponding acid or acid mixture. As noted previously, the oxidizing power of nitric acid with respect to many matrices is insufficient at such low temperatures. One possible remedy is the addition of sulfuric acid, which significantly increases the temperature of a digestion solution. Whether or not this expedient is practical depends on the matrix and the determination method. High-fat and high-protein samples are generally not subject to complete ashing at atmospheric pressure. Other disadvantages relate to the risk of contamination through laboratory air, the necessarily rather large amounts of required reagents, and the danger of trace losses. Nevertheless, systems operated at atmospheric pressure are preferred from the standpoint of workplace safety.

This category also includes the *solubilization* of simple matrices with low carbon content using special detergents or other solubilization reagents, as well as the enzymatic digestion techniques that are often used in food control and monitoring. Enzymatic digestion in trace analytical chemistry has been described in [34] for the example of the non-specific protease enzyme (pronase), which was used as an alternative sample preparation technique for the determination of trace elements in blood serum by ICP-MS. Using enzymatic digestion, a higher degree of instrument stability was achieved over 3 h period. The reported results are in good agreement with the certified values of the reference material [34]. Also “extraction – digestion” techniques belong to this category, e.g., using aqua regia at relatively low temperatures (e.g., 70 °C) over a long period of time (up to 5 d) as tool for the pseudo-total analysis of metals in soil samples [35]. These techniques are some importance as sample preparation techniques, but one should be aware that they do not lead to complete digestion of the organic matrix. Extraction – digestion techniques have become standard methods for the examination of water, waste water, sludge, and sediments using aqua regia for the subsequent determination of the acid-soluble portion of metals [36].

Thermally Convective Wet Digestion. The conventional approach to wet digestion, which has proven its worth over many years, entails a system equipped with heating units operating either at a fixed temperature or in response to a temperature program. Commonly employed digestion agents include nitric acid, sulfuric acid, hydrogen peroxide, hydrofluoric acid, and perchloric acid, as well as various combinations of these [33]. Most applications of wet digestion involve aqueous or organic matrices, such as surface waters, waste water, biological samples (tissue, body fluids, etc.), food samples, as well as soil and sewage sludge, coal, high-purity materials, and various technical materials [35].

Microwave-Assisted Wet Digestion. Microwave-assisted digestion in open vessels (atmospheric digestion) [37] is generally applicable only with simple matrices or strictly defined objectives, and the results are reproducible only if the specified ashing parameters are strictly observed. Losses may be encountered with mercury and possibly also with organometallic compounds (e.g., those containing arsenic, antimony, or tin). Addition of sulfuric acid is essential in order to achieve a sufficiently high digestion temperature in atmospheric-pressure equipment, where the boiling point of the acid establishes the maximum ashing temperature, although it is important to remember that the presence of sulfate interferes with many procedures for metal determination (e.g., graphite-furnace atomic absorption spectrometry, GF-AAS). Specific experimental conditions for microwave-assisted digestions have been summarized in [9], [22] as a function of the matrix to be ashed.

Ultraviolet Digestion (Photolysis). Ultraviolet (UV) digestion is utilized mainly in conjunction with uncontaminated or slightly contaminated waters, such as sea water and surface water. Liquids or slurries of solids are decomposed by UV radiation in the presence of small amounts of hydrogen peroxide and acids (e.g., beverages, wastewater, soil extracts). The corresponding digestion vessel should be placed in the closest possible proximity to the UV lamp to ensure a high light yield. In photolysis (see Fig. 3) the digestion mechanism can be characterized by the formation of OH radicals from both water and hydrogen peroxide that is initialized by the aid of the UV radiation [9], [38]. These reactive radicals are able to oxidize the organic matrix of simple matrices containing up to about 100 mg/L of carbon to carbon dioxide and water. Effective cooling of the sample is essential, since losses might otherwise be incurred with highly volatile elements.

Contamination can be minimized by the use of a nongas-tight stopper. Hydrogen peroxide addition may need to be repeated several times to produce a clear sample solution [39], [40]. Complete elimination of the matrix is of course possible only with very simple matrices (e.g., slightly contaminated water) or by combining photolysis with other digestion techniques [41], [42]. The advantage of UV photolysis is to reduce spectral and nonspectral interferences and to improve the long-term stability of instruments such as ICP-MS. At the same time, analytical blank values can be kept very low because of the only minimal reagent addition. The use of ultrapure nitric acid and hydrogen peroxide for UV photolysis in combination with high-resolution ICP-MS allows the determination of platinum-group elements, thorium, and uranium down to the upper pg/L level [43], [44].

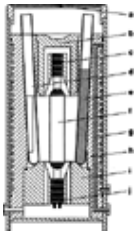

Figure 3. Apparatus for the UV digestion (photolysis) of liquid samples with low carbon content (with permission from Kürner, Rosenheim/Oberbayern, Germany)

a) UV protection glass; b) Lamp cooler; c) Lamp socket; d) Sample vessel; e) Sample solution; f) UV lamp; g) Coolant (water); h) Sample cooler; i) Lamp cooler; j) Lamp socket

A new promising development is the microwave-supported UV digestion, which also allows the digestion of liquid organic samples with higher carbon contents. The technique is based on a closed, pressurized microwave digestion device which allows maximum reaction temperature up to about 250 °C in combination with UV irradiation. This results in an accelerated degradation and in a higher sample throughput [45]. The advantage over classical pressure digestion techniques (see Pressure Digestion) is the low acid demand leading to very low reagent blanks.

2.3.2. *Pressure Digestion*

Pressure digestion offers the advantage that the operation is essentially isolated from the laboratory atmosphere, thereby minimizing contamination, and digestion occurs at relatively high temperature due to boiling-point elevation effects. The pressure itself is in fact nothing more than an undesirable — but unavoidable — side effect. The principal argument in favor of this form of digestion is the vast amount of relevant experience acquired in recent decades. The literature is a treasure trove of practical information with respect to virtually every important matrix and a great number of elements (see, for example [46]). Pressure digestion is particularly suitable for trace and ultratrace analysis, especially when the supply of sample is limited.

Since the oxidizing power of a digestion reagent shows a marked dependence on temperature, a distinction must be made between “simple” pressure digestion and high-pressure digestion. Simple pressure digestions (< 20 bar) are limited to a temperature of ca. 180 °C, whereas with high-pressure apparatus (> 70 bar) the digestion temperature may exceed 300 °C.

2.3.2.1. *Thermally Convective Pressure Digestion*

Most sample containers for use in *thermally convective pressure digestion* are constructed from PTFE or PFA, although special quartz containers with PTFE holders are available for trace-analysis purposes. The sample container is mounted in a stainless-steel pressure vessel (autoclave) and then heated in a furnace to the desired temperature. Pressure digestions based on apparatus of the Tölg, Bernas, and Parr type are all feasible below ca. 180 °C, but above this temperature PTFE begins to “flow”, rendering it unsuitable for use in high-pressure applications. Investigations with organic matrices have shown that at 180 °C as much as 10 % of the carbon present in high-fat and high-protein samples remains unashed [47]. Such samples must be processed further prior to analysis through evaporation of the digestion solution with perchloric acid.

All thermally initiated digestions have the disadvantage that a considerable amount of time is consumed in preheating the digestion solutions due to the need for heating the autoclave and sample vessels.

High-Pressure Digestion. The introduction of a high-pressure ashing (HPA) technique by Knapp [48] has not only reduced the effective ashing time to ca. two hours but also opened the way to digestion of extremely resistant materials [49]. High-pressure ashing is conducted in quartz vessels, with a maximum ashing temperature as high as 320 °C at a pressure of ca. 100 bar. The quartz vessel is stabilized during the digestion process by subjecting it to an external pressure roughly equivalent to that developed within. Even thermally stable organic compounds undergo acid decomposition at such high temperatures. Essentially complete ashing can be accomplished with the vast majority of samples so far investigated. Nitric acid alone is a sufficiently powerful reagent in many cases.

2.3.2.2. *Microwave-Assisted Pressure Digestion*

The digestion of lightly contaminated water, many geological samples, and low-fat plant matrices can be accomplished successfully with a domestic microwave oven and relatively primitive sample holders. The introduction of complete commercial microwave systems specifically designed for sample ashing has led to important further developments. Such systems also offer special safety features and improved facilities for controlling and regulating the pressure and/or temperature [50]. Moreover, the distribution of microwave radiation inside the oven cavity tends to be significantly more homogeneous than with domestic devices, in part because the magnetron control is subject to shorter switching cycles.

Nitric acid alone does not always lead to satisfactory microwave-assisted ashing. Addition of hydrochloric acid or hydrogen peroxide is often helpful, and hydrofluoric acid must of course be present in the case of silicate- containing materials [39], [51]. The maximum attainable pressure seldom exceeds 10 bar.

A serious disadvantage of this simple approach is the frequent need for multistep ashing, in which a preliminary digestion

is followed by a second treatment, perhaps with added hydrogen peroxide [40].

High-Pressure Microwave Digestion. High-pressure (or high-temperature) microwave-induced processes are designed to support digestion up to a pressure of perhaps 70 bar, opening the way to substantially higher temperatures (e.g., ca. 250 °C) without a corresponding increase in overall processing time. Temperature and digestion time ultimately determine the effectiveness of a digestion, with residual carbon content serving as a useful measure for quantitative assessment. The striking advantage of high-pressure, i.e., high-temperature (!) microwave digestion (see Fig. 4) is that almost complete dissolution can be accomplished with a single-stage process, even in the case of complex matrices. Examples found in the recent literature consist of new successful developments in the technique and apparatus [52] and applications to biological tissues [53] or technical products [54].

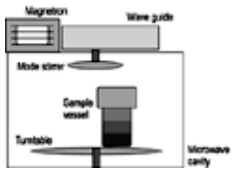

Figure 4. Apparatus for microwave-assisted pressure digestion

Examples of almost completely mineralizable matrices include resistant inorganic and high-fat organic samples. Nevertheless, limitations with respect to subsequent use of electrochemical methods of determination (Section Sample Preparation as a Function of Analytical Method) still represents a critical problem [55], and the results are often inferior to those from a conventional high-pressure asher (HPA).

2.4. “Dry” Digestion Techniques

The term “dry” digestion is intended to encompass all processes based on gaseous or solid digestion reagents. Such a distinction relative to wet digestion processes is not absolutely essential, but it does offer certain practical advantages [46].

2.4.1. *Combustion in Air*

Strictly speaking, “dry ashing” refers to the oxidation (combustion) of a substance in air at a temperature of several hundred degrees Celsius, often in a muffle furnace or similar apparatus.

Dry Ashing in a Muffle Furnace. Carbon- containing substances generally decompose satisfactorily under these conditions without the use of auxiliary agents. It is advantageous that reasonably large samples can be digested. In addition to routine biological samples (plants, foods), coal has also been subjected to dry ashing [56]. The importance of this technique has declined, primarily because it is so time-consuming; it also entails a high risk of contamination since the samples are kept in open vessels [57]. Losses must be anticipated in the case of volatile substances [58]. However, the literature provides numerous examples of decompositions that have been optimized through the addition of “ashing auxiliaries” (salt solutions or acids). Such additives often reduce the ashing time, especially for organic substances, and they may prevent the volatilization of element traces.

Dry Ashing in a Microwave Oven. The time required for dry ashing in air can be reduced by microwave heating of a special block containing the sample [25]. This block is fabricated from a material that is highly absorbent with respect to microwaves (e.g., silicon carbide), and is is surrounded by quartz insulation. At maximum power such a block can be heated to about 1000 °C within ca. 2 min. In addition to time saved during the heat-up phase this system offers the further advantages of a low energy requirement and a procedure that spares the technicians from exposure to intense heat when inserting and removing samples, in sharp contrast to work with a conventional muffle furnace [25].

2.4.2. *Combustion in Oxygen*

Combustion in oxygen is a decomposition procedure that eliminates many of the potential problems of sample contamination associated with digestion reagents [59]. The sole reagent here is gaseous oxygen, which is commercially available in very high purity. Related combustion techniques are based on the use of other oxidizing agents (e.g., sodium peroxide) from which free oxygen is released during the course of the operation. Analogous methods rely on oxidizing agents other than oxygen (e.g., chlorine [60], fluorine), but these have not been widely accepted, in part because they require working with very aggresive materials, often not available with the required level of purity and themselves capable of interfering with the determination process.

Atmospheric-Pressure Combustion. Although combustion in an open vessel is relatively easy to accomplish, it has the disadvantage of possible volatilization of element traces. Significant advantages are associated with the compromise of a quasi- closed quartz glass system consisting of a central combustion chamber, an ignition device (an infrared source), and a reflux condenser charged with liquid nitrogen. The analyte, together with inorganic combustion residues, is redissolved by refluxing with a small amount of acid. Such a system is applicable to almost all the elements, and it leads to nearly complete decomposition [59], [61]. Applications have been described with respect to a large number of organic matrices, including petroleum products, coal, and soils. Combustion systems also exist for analysis of nonmetals. SCHÖNIGER developed a particularly simple combustion process (now bearing his name) for general elemental analysis [62]. A special variant — combustion in an oxyhydrogen flame — is carried out in a Wickbold apparatus (Fig. 5) [63].

Figure 5. Wickbold combustion apparatus

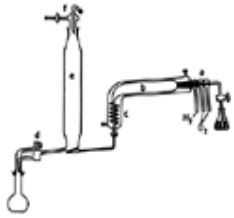

a) Injection burner; b) Combustion chamber; c) Condenser; d) Multiport stopcock; e) Absorption chamber; f) Flushing port

Combustion in a Closed Vessel. The risk of volatilization of trace analytes (e.g., mercury) can be reduced further by conducting the combustion in a special combustion bomb (e.g., a Parr bomb) [64]. Oxygen gas is introduced up to some predetermined pressure, after which the sample is subjected to explosive combustion with the aid of a detonation device (ignition wire). The analyte, together with inorganic residues, is again collected in a small volume of acid [65]. This technique is also applicable to the determination of nonmetals [66].

Microwave-induced oxygen combustion is a new combustion technique for the determination of halogens, which has been first presented at Analytica 2004. It consists in an accessory which is used in combination with a commercial microwave sample preparation system [67].

2.4.3. *Cold-Plasma Ashing*

Decomposition by cold-plasma ashing is accomplished with excited-state oxygen under a pressure of a few millibars. The plasma is formed at reduced pressure (0.01 – 1 mbar) in a high-frequency field established either with a magnetron (2.45 GHz) or with a semiconductor high-frequency generator. In the simplest application samples are ashed in flat dishes (e.g., Petri dishes) arranged in trays [68]. A quasi-closed system with a cold-finger condenser should of course be employed for trace analysis to minimize potential losses [69]. Oxidation of organic matrices occurs by way of reactive, short-lived radicals, leading to more complete matrix decomposition relative to most other approaches. This ashing technique is therefore particularly recommended as a precursor to electrochemical determinations (e.g., polarography). The oxygen plasma is a “cold plasma,” which means that the sample temperature never exceeds ca. 150 °C. Risk with respect to the loss of volatile elements is therefore restricted to mercury. A reaction period of ca. 2 – 4 h is usually sufficient to disrupt the organic matrix to such an extent that the residue can be removed from the sample vessel and cold finger by refluxing with a small volume of concentrated acid (2 mL). The most important application of cold-plasma ashing is the mineralization of biological samples [70], although coal, graphite, and plastics (including PTFE) can also be ashed by this technique [69]. Samples as large as one gram or more can be accommodated.

2.4.4. *Fusion*

Fusion is the traditional approach to sample preparation for industrial and geological analyses [71], especially for the analysis of mineral samples (e.g., lithium metaborate melts [72]). The major problems associated with fusion include restricted means for purifying the required (solid) reagents and the presence of high salt concentrations in the resulting analyte solution. Especially for trace analysis, overloading of the matrix in this way is inadvisable.

2.5. Illustrative Examples

The sections that follow highlight a few particularly interesting examples of sample preparation, taking into account both the analytical method of choice and effective mineralization of specific matrices.

2.5.1. *Sample Preparation as a Function of Analytical Method*

In the case of *atomic spectrometry* [e.g., atomic absorption (AA), inductively coupled plasma – optical emission spectrometry (ICP –OES), inductively coupled plasma – mass spectrometry (ICP – MS)], matrix interference due to incomplete digestion may manifest itself as differences in the suction rates and aerosol yields for samples relative to calibrating solutions. Such differences may also reflect differences in the bonding states of the elements, which in turn leads to systematic errors and problems with calibration.

Electrochemical determination methods are particularly sensitive to incomplete sample mineralization. The result is matrix interference manifested in the case of inverse voltammetry, for example, as ghost peaks, signal suppression, and an enhanced hydrogen signal. With simple, aqueous samples such as seawater and river water, voltammetric determination of zinc, cadmium, lead, and copper can be effectively coupled with UV decomposition [73], [74]. The same determination method is applicable to these heavy metals in simple plant products such as sugars provided decomposition (ashing) is effected by combustion in a stream of oxygen (typical sample mass: 1.5 g) [75]. Combustion is also appropriate prior to the potentiometric determination of aluminum in organic samples [76]. Biological samples can be prepared for electrochemical analysis by cold-plasma or high-pressure ashing [41], [48], [59], [77].

An improved procedure has been described for atmospheric-pressure combustion of samples from high-purity metals [78]. High-pressure digestion with nitric and hydrofluoric acids [79], [80] and combustion techniques have both been shown to be suitable for the decomposition of samples of silicate- containing materials and fuels. Combustion in this case is carried out either in a stream of oxygen [81] or, to provide a closed system, in an oxygen bomb [82]. Mercury can be satisfactorily determined after prior Wickbold ashing [83].

Halogens such as iodine and bromine cannot successfully be analyzed after digestion with oxidizing acids such as nitric acids, because this would result in the oxidation to elemental halogens and losses due to volatilization. Schöniger combustion and the subsequent uptake into a basic medium such as sodium hydroxide or tetramethylammonium hydroxide (TMAH) prevent such losses. As a rule, TMAH is preferable, because it results in a lower salt burden to instruments such as AAS or ICP-MS. Schramel et al. [84] applied Schöniger combustion in combination with ICP-MS for

the determination of iodine in biological matrices such as urine and serum.

The decomposition of inorganic substances can be accomplished with a mixture of hydrochloric and nitric acids in a closed vessel (pressure digestion). Silicate-containing minerals and glasses require the addition of hydrofluoric acid, and the presence of perchloric acid, phosphoric acid, or sulfuric acid is sometimes useful as well [85]. The resulting metal fluorides can be redissolved by adding a solution of boric acid. The same procedure is effective for ores and slags as well as quartz [86].

Cold-plasma ashing is useful as a way of excluding matrix interferences in the detection of elements with secondary-ion mass spectrometry (SIMS) [87].

2.5.2. *Combined Use of Multiple Decomposition Techniques*

A single ashing procedure is often insufficient for the complete decomposition of a complex matrix, leading some authors to recommend a combination of two or more techniques. One example will suffice to illustrate the principle: pressure ashing followed by UV photolysis. Thus, it has been shown that analysis of olive leaves for heavy metals by voltametric methods leads to distorted results after pressure digestion alone. Reliable data can be obtained only by supple-menting the digestion with UV irradiation to ensure adequate decomposition of the matrix [41].

2.5.3. *Comparative Merits of the Various Sample-Preparation Techniques*

High-fat samples present special problems with respect to decomposition due to the presence of highly resistant components. The result is often incomplete ashing. Moreover, there is always a risk of explosive decomposition caused by the formation of reactive radicals in the digestion solution. Unlike electrochemical methods of determination, which require the complete degradation of organic matrices, atomic absorption (particularly the flame technique) is compatible with a certain amount of residual carbon. Any digestion solution that is colorless and free of particles is a reasonable candidate for AAS analysis.

A systematic comparison of microwave-assisted decomposition techniques with cold-plasma ashing, conventional thermally convective pressure digestion, and high-pressure digestion is presented in [40]. Generally speaking, comparable results can be obtained with high-quality commercial ashing systems of all the common types, with the potential for nearly complete decomposition [e.g., the Paar microwave-assisted pressurized microwave decomposition (PMD) system, the Büchi high-pressure adapter, and the Knapp HPA system]. Microwave ashing of high-fat samples in an open vessel presents a serious risk of significant analyte loss despite the use of a reflux condenser, but losses with a microwave pressure-digestion system have been found to be no greater than with a corresponding convective heating system. Results also suggest that “memory effects” are less noticeable with a microwave system than in the case of a conventional heat source [40].

Table 4 provides a comparison of the time required for various steps in microwave versus convective thermal ashing, taking into account all the associated preparations, the cool-down phase, and cleaning of the apparatus, which has the effect of reducing somewhat the marked time advantage of the microwave-assisted ashing period. The three-fold time advantage of the complete microwave-assisted procedure is reduced even more if, for example, six samples are digested, because the PMD cannot process more than two samples at once. Table 5 shows that microwave-assisted pressure ashing is the most rapid of the techniques investigated, faster than the frequently used Tölg technique (Section Thermally Convective Pressure Digestion) by a factor of six. Cold-plasma ashing takes even longer. It should be noted, however, that this comparison ignores the advantages that might be derived from parallel ashing of several samples, and it does not reflect special requirements associated with particular matrices or analytes.

Table 4. A comparison of the time required for microwave pressure digestion versus thermally convective digestion (for one sample and for six samples)

Operational step	Time, min			
	Microwave digestion (PMD)		Knapp digestion (HPA)	
	One sample	Six samples	One sample	Six samples
Preparation for digestion	10	30	10	30
Ashing, including warm-up	10	30 *	90 – 150	90 – 150
Cool-down	10	30	20	20
Preparation for analysis	5 – 10	30	5 – 10	30
Cleaning the equipment	10 – 20	20	10 – 20	20
Complete procedure	45 – 60	140	135 – 210	190 – 250

* Limited to two parallel samples.

Table 5. Overall ashing times associated with various procedures (for one sample)

Sample type	Ashing time, h

	Pressure microwave *	Open microwave	Tölg **	Knapp **	Cold plasma
Plant matrix	0.75	0.92	4.5	3	10
Sunflower oil	0.75	0.92	4.5	3	20

* Single-stage digestion.
** Thermal convection heating.

Earlier investigations confirm these results. According to [41], however, complete degradation of many samples is achieved only through cold-plasma ashing or by a combination of wet ashing and UV irradiation. In a determination of aluminum in dialysis liquids it was observed that microwave-assisted ashing produced results comparable to those from conventional pressure ashing [55].

Five different digestion techniques are compared in [88] in an attempt to optimize the determination of mercury in soil via cold-vapor AAS. Only with closed systems was loss-free digestion assured. Open systems resulted either in mercury losses or incomplete recovery due to incomplete ashing. On the other hand, wet ashing in an open system with nitric and sulfuric acids was found to be preferable for the mineralization of biological samples, since it was easier to accomplish and led to lower blank readings [89].

Fusion with lithium metaborate is feasible for geological samples provided the melt is subsequently dissolved in acid in the course of a pressure- or microwave-assisted ashing process [90].

2.5.4. *Decomposition Procedures for Determining Nonmetals*

A pertinent example of an analysis for nonmetallic constituents is the digestion of water or food samples with peroxodisulfate, according to the Koroleff method [91], which was developed for the determination of all bound nitrogen in seawater, inland waters, rain water, groundwater, wastewater, or effluents. Measurable concentrations are in the range 0.02 – 4 mg of nitrogen per liter. The nitrogen-containing sample is first subjected to a pressure digestion in the course of which organic nitrogen compounds, nitrite ion, and ammonium ion are all oxidized to nitrate. This oxidation proceeds quantitatively only under alkaline conditions, achieved with the aid of a buffer system consisting of boric acid and sodium hydroxide. Reaction commences at pH 9.7 and ends at pH 5 – 6, at which point the nitrate exists in stable form in the acidic digestion solution. Nitrogen is not released in this way from certain five-membered heterocyclic systems [92], but the resulting error is probably negligible with most of the samples subject to investigation. The total nitrogen content in the digestion solutions can be established by photometric or ion-chromatographic methods. A Kjeldahl analysis can be used as a standard for comparison, using either the conventional technique or microwave-assisted heating. The accuracy of the results has been verified on the basis of standard reference material (IAEA A-11, full-cream milkpowder, in the case of foods) and by interlaboratory comparison (river water, wastewater) [93].

The fluoride content of silicates can be established after fusion with lithium metaborate [94], where fluoride is subsequently separated with the aid of superheated steam [95].

2.6. Evaluation Criteria

The maxim “one method is no method” is just as applicable to sample preparation as it is to quantitative determinations. A careful comparison of several decomposition techniques is the only way of assuring accurate results, particularly when little experience is available with respect to the decomposition of a specific matrix, or existing reports are contradictory.

2.6.1. *Completeness*

Decomposition in the strictest sense has as its goal a complete elimination of the matrix. Nevertheless, many decomposition procedures fall short of this ideal, permitting traces of the matrix to remain. Residual carbon content can be established with a total organic carbon (TOC) analyzer. This decreases with increasing oxidizing power of the decomposition reagent as well as with an increase in the duration and temperature of the digestion [46].

2.6.2. *Uncertainty*

As with all aspects of trace analysis, strict attention to possible sources of error is important in the successful application of a decomposition technique. The best determination is rendered completely worthless if the sample preparation step leads to deviations from the true value that exceed the reproducibility of the determination itself by several orders of magnitude. The risks are particularly great when one is working with very low concentrations. Even very small amounts of contamination introduced with the reagents, or “memory effects” from previous experiments, can generate utterly incorrect results. The laboratory atmosphere and materials present in the sample containers can also contribute to contamination and loss, and the potential for trace volatilization (of mercury, for example) at elevated temperature must always be taken into account. Sample preparation is an ideal place to apply the principles of “Good Laboratory Practice” (GLP), ⟶ Analytical Chemistry: Purpose and Procedures; even better, “Good *Analytical* Practice” (GAP). Construction materials for the equipment used in sample preparation must be selected very carefully, with appropriate consideration of possible adsorption – desorption effects. Steaming with concentrated acid is the most effective method for cleaning the equipment (Section Requirements with Respect to Materials and Chemicals).

The nature of any safety devices employed can also play a role with respect to uncertainty in the case of a pressure ashing system. For example, serious losses may be incurred if sample constituents are entrained during a pressure blow-off. With an irreversible safety device like a rupture disc such a problem may be immediately apparent, but the activation of a reversible device like a valve generally cannot be documented even though major losses may be excluded by automatic reclosure of the device. The optimal solution is almost always a function of the particular application in question.

2.6.3. *Time Factors*

Sample preparation is increasingly coming to be regarded as the “bottleneck” in analysis, because the actual determination methods are becoming increasingly rapid and ever more stringent demands are imposed with respect to precision and accuracy. The only remedies available are acceleration of the decomposition process and simultaneous preparation of multiple samples. However, the latter expedient inevitably leads to higher investment costs and increased operating expenses. Decomposition can be accelerated by working at higher temperature. In the case of wet and dry decompositions, heating can be accelerated by the use of microwave radiation. The subsequent cool-down phase also consumes less time, because the vessel walls are heated only indirectly via the hot sample, and there is no metal autoclave present, with its correspondingly high heat capacity.

Finally, an important parameter to consider in evaluating a routine sample preparation method is the number of staff involved. Several manufacturers have for some time been making a deliberate attempt to reduce staff requirements by introducing automation (or partial automation) into the regimen of sample preparation. The results have been quite satisfactory in the case of apparatus operating at atmospheric pressure, but the technical challenge is greater with systems operating under pressure, especially high pressure (see Section Automation and Direct Analysis).

2.6.4. *The Final Result*

It is impossible to separate completely the consideration of a specific decomposition process from the subsequent analytical determination. Every decomposition must be adapted to suit a particular determination, which means there can be no such thing as an ideal decomposition method appropriate to all situations and applications. The preferred method must be established with reference to the following factors:

1. Nature of the analyte(s)
2. Concentration of the analyte
3. Matrix characteristics
4. Sample size
5. Required completeness of matrix degradation
6. Amount of time involved
7. Proposed method of determination

If the greatest possible degree of attention must be directed toward completeness in the trace range, decomposition in a closed system is preferred. Ensuring essentially complete decomposition requires the application of high temperature over a period of at least two hours. If sample preparation must be accomplished more rapidly (as in the case of food analysis, environmental studies, or quality control) tests should be conducted on the feasibility of a microwave-assisted process or direct analysis (see Section Direct Analysis).

An investigation into the accuracy of an analysis should never be restricted to optimization of the determination itself: equal concern is warranted with respect to sample preparation. The key considerations listed above should be regarded only as rough starting points.

2.7. Concentration and Separation of Inorganic Trace Materials

Subsequent to any decomposition, but also in the case of liquid samples such as water and urine, the analytes of interest are generally present in dilute solution together with a large excess of foreign ions (e.g., alkali-metal and alkaline-earth cations). Separation and concentration of the analytes may be necessary to improve the limit of detection and exclude interference. Useful techniques in this regard include liquid –liquid extraction, solid-phase extraction, special precipitation reactions, and electrolytic deposition.

Liquid – liquid extraction (see also ⟶ Liquid–Liquid Extraction*)* is a widely used technique for separating and concentrating traces of various elements. For this purpose the analyte is first complexed with a suitable chelating agent, and the complex is then extracted into a water-immiscible solvent. Generally speaking, any complexing agent can be used provided it leads to stable, extractable complexes of the analyte in question. The complexing agent should be capable of separating the analyte as selectively and quantitatively as possible from accompanying ions. Such an extract can usually be used directly for a determination based on AAS. For example, nickel [96], cobalt [97], and cadmium [98] can all be separated and concentrated directly from urine after chelation with hexamethylene ammonium – hexamethylene dithiocarbamidate (HMA – HMDC), using the solvent mixture methyl isopropyl ketone – xylene (70:30). The separation and enrichment of nickel and antimony in the form of HMA – HMDC complexes from digestion solutions of livers and kidneys has also been described, where copper and iron are removed in a preliminary step as the corresponding cupferron complexes [99]. Thallium can be extracted from a digestion solution after complexing with HMA – HMDC, diethyl dithiocarbamate, dithizone, and 8-mercaptoquinoline [100], or cupferron [101]. Antimony in urine has been extracted after nitric acid digestion as an ammonium pyrrolidine dithiocarbamate complex [102].

The advantage of this approach to concentration is that it permits very selective separation provided a complexing agent is selected with sufficient regard for accompanying ions. Its principal disadvantage is that the potential concentration factors tend to be rather small.

Metal complexes separable by liquid – liquid extraction can also be concentrated on and subsequently eluted from such adsorber resins as XAD, a reverse phase (e.g., C_{18}) [103], activated charcoal [104], or cellulose [105], leading to the technique of *solid-phase extraction*. Numerous reports describe the concentration of analytes on ion-exchange resins [106], [118] and such chelate-forming ion exchangers as Chelex [106], [119], [120] or Hyphan [121], [122]. Solid-phase extraction may result in concentration factors as high as 1000, particularly with dilute samples like drinking water or rain water.

Separation of element traces is also possible with *precipitation* reactions. This technique permits rapid and extensive

concentration in a relatively uniform matrix, but it is not very selective. Aluminum hydroxide, magnesium hydroxide, iron hydroxide, and hydrogen sulfide have all been used for trapping trace amounts of various elements [106], [123-125], as have such organic precipitating agents as thionalide, cupferron, and dithiocarbamate [106], [126], [127]. A subsequent determination is carried out either directly on the separated precipitate or after restoring it to solution [123].

The *electrolytic deposition* of ions or complexes has also been described with such elec-trode materials as carbon, copper, or platinum. For example, this technique is utilized in the method of inverse voltammetric determination, where analyte is concentrated at the cathode [106].

New techniques and strategies have been introduced into sample preparation in the area of preconcentration of inorganic analytes and the separation of such analytes from the matrix [107-109]. MANSFELDT and BIERNATH [110] described a micro-distillation apparatus (Fig. 6) for the determination of total cyanide in soils and sludge.

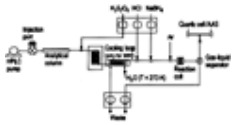

Figure 6. Schematic flow diagram of an on-line coupling of HPLC via UV or microwave-assisted reactors to hydride generation AAS for the determination of As species

Another paper has suggested solid-phase extraction (SPE) for the preconcentration and speciation of Cr in waste water samples prior to AAS determination [111]. Interferences of di- and trivalent cations and diverse anions were studied. The detection limits for these species are about 1 µg/L. Other examples for SPE procedures are given in [112] for seawater analysis using on-line ICP-MS and in [113] for the separation and determination of precious metals.

New aspects of liquid – liquid extraction after a complexation step have been discussed for noble metals [114] and for rare earth elements [115].

Also, supercritical-fluid extraction (SFE) has been used for analyte enrichment and matrix separation in metal determinations [116]. The solubilities of metal dithiocarbamates in supercritical carbon dioxide have been characterized [117].

As an additional step in trace analysis, preconcentration techniques are resorted to only when either the sensitivity of the analytical determination is inadequate or severe matrix effects worsen the detection of the analytes [108].

2.8. Automation and Direct Analysis

2.8.1. *Automation*

Efforts have long been directed toward automation as a means of minimizing the labor and time involved in an analysis. The introduction of laboratory robots should make it possible to incorporate a significant degree of automation into the time- consuming, labor-intensive area of sample preparation as well, leading to more efficient, reliable, and reproducible sample work-up. An example of an automated system for microwave-assisted pressure ashing is presented in [128].

On-line procedures based on UV-photolysis or microwave-assisted dissolution have been established for the rapid digestion of relatively simple matrices such as water and waste water samples [135], [136]. In [129] on-line UV digestion with a segmented-flow device was applied for the determination of total cyanide prior to amperometric detection. The detection limit was found to be 0.2 µg/L. The throughput is about 30 analyses per hour.

Another example [130] is the on-line microwave and UV digestion of urine samples prior to hydride generation AAS (Fig. 7). The authors managed to determine as many as six arsenic species quantitatively [As(III), As(V), MMA, DMA, arsenobetaine, arsenocholine)]. For more information on species analysis see Section Analysis of Element Species).

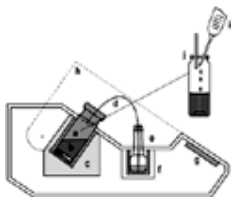

Figure 7. Apparatus for micro-distillation

a) Sample vessel; b) Frit; c) Heater; d) Capillary; e) Absorption vessel; f) Cooler; g) Keyboard and display; h) Protection hood; i) Syringe for reagent feeding; j) Septum (with permission from Dr. Tim Mansfeldt, Ruhr-Universität Bochum, Germany)

Ozone has also been used as a reagent with high quality criterions. For the determination of Hg in a batch cold-vapor system, ozone was used instead of permanganate or peroxidisulfate. The oxidation efficiency of ozone was found to be very high and the digestion was complete in less than 2 min [131].

Microwave decomposition at atmospheric pressure can also be used as an on-line technique, e.g., in a way that the sample passes through a PTFE coil with an inner diameter of about 1 – 2 mm which is situated in a focused microwave-heated oven. A fully automated on-line sample pretreatment system which combines microwave digestion with sample preconcentration and matrix separation for heavy metals in blood samples is described in [132]. The digested sample solution was transferred on-line to a column packed with iminodiacetate resin for separation of matrix elements (alkaline, alkaline earth, non-metals) that might interfere with the analyte mass signals in the ICP-MS detection method used. For Ni and Pb the authors reached detection limits in the upper ng/L range, for Fe, Cu, and Zn in the lower µg/L range. The sample throughput is given to be 6 samples/h.

PICHLER et al. [133] have shown that such an on-line microwave procedure can also be adapted to pressurized ashing. In this special case, the sample is injected into a constant stream of nitric acid which passes the microwave heating zone. Expansion of the coil due to the high vapor pressure is prevented by pressurizing its surroundings with nitrogen at up to 35 bar. The temperatures that can be reached in such a system may be as high as 230 °C or even higher.

An on-line SFE/HPLC procedure for the determination of Rh and Pd via chelates has been described in [134].

Flow-injection systems and continuous-flow systems can also make a contribution at the concentration and separation stages as a way of enhancing the potential advantages of automated atomic spectrometric determinations by the flame and graphite-furnace techniques [137], [138]. Another possible application of flow injection is in the conversion of an analyte into the particular oxidation state called for in some sample preparation or determination procedure, as in hydride generation or cold-vapor AAS [139], [140].

2.8.2. *Direct Analysis*

There are certain situations in which sample preparation can be completely or largely omitted, as in the case of graphite-furnace AAS analysis, which permits the direct introduction of solid samples. Here the “ashing step” has been incorporated into the determination itself, which is then referred to as a “direct process.” This integration permits an analysis to be carried out more quickly, and the chance of errors is reduced by limiting the number of individual operations [141]. On the other hand, it is essential that the accuracy of such an analysis be confirmed by comparisons with standard reference materials.

Examples of the AAS analysis of powdered samples, as well as such solid foods as chocolate and flour, are presented in [142], [143]. The direct determination of mercury in soils, coal, and ash is discussed in [144]. ICP – OES (and ICP – MS) can also be transformed into a direct process through solid sample introduction and electrothermal pretreatment [145].

2.9. Analysis of Element Species

For the determination of the Hg species methylmercury, phenylmercury and Hg(II), ozone was used successfully in a batch cold-vapor system [131]. The preconcentration and speciation of Cr(III) and Cr(VI) in water sample can be performed using solid-phase extraction (SPE) [110]. An SPME (solid-phase microextraction) technique has been used as the sample preparation system for the innovative simultaneous multielement/multi-species determination of six different mercury, tin, and lead species in waters and urine with GC/MS-MS [146], [147]. The determination of arsenic species [As(III), As(V), MMA, DMA, arsenocholine, arsenobetaine)] has also been shown to work with an on-line digestion step prior to hydride generation AAS [130] (see Fig. 7).

[Top of Page]

3. Sample Preparation in Organic Analysis

Recent decades have witnessed significant advances in the efficiency and productivity of instrumental methods in the field of organic trace analysis. Chromatographic and spectroscopic methods in particular have improved greatly with respect to sensitivity. There has also been constant improvement in selectivity, to the point where some samples can now be subjected to analysis without prior preparation, although this is certainly not true in the majority of cases. The goal of sample preparation in organic trace analysis is to isolate the analyte from the sample matrix and then concentrate it and convert it into a form suitable for analysis by the selected method.

Microwave-based techniques have become versatile tools not only for total digestion of organic and inorganic matrices for use in inorganic analysis, but also in drying processes, solvent extraction, clean-up steps and specific reactions such as sample preparation steps in organic analysis [148].

3.1. Sample Treatment after the Sampling Process

3.1.1. *Stabilization, Drying, and Storage*

In the context of organic trace analysis, appropriate stabilization and storage precautions are a function of the nature of the analyte and its concentration. To avoid contamination by ambient air and dust, all operations should be conducted under clean-room conditions (clean-bench environment). It is of paramount importance to assure that contamination from vessels, covers, septa, and stabilizers is rigorously excluded.

Volatilization losses can pose a problem in the case of analytes with high vapor pressure. Evaporative loss can be largely avoided by completely filling each sample container and fitting it with an air-tight seal (water samples, biological fluids). This procedure ensures that there will be no oxidative loss due to oxygen present in the vapor space above the sample. For example, blood samples to be analyzed for volatile organic compounds (VOCs) are first treated with an anticoagulant, after which they can be stored for several months at +4 °C in Sovirel or Pyrex test tubes protected only by screw- cap closures and PTFE seals [149]. Similarly, VOCs adsorbed from air samples onto Tenax or activated charcoal can be sealed and stored for several months prior to analysis. Losses have been reported only with cyclic and aliphatic ketones (catalytic oxidation) and esters (hydrolysis) [150]. For these groups of substances, losses approach a maximum of 20 % after a four-week storage period. No losses were observed with benzene, toluene, and xylenes over a 24-month storage period [151]. Samples containing analytes of lower volatility can be stored in suitable containers for a few days at +4 °C, and for several months deep-frozen at – 18 °C or lower.

If an analyte is to be separated by Soxhlet extraction or supercritical fluid extraction, it is essential that the sample ist first dried. Standard practice is to add a drying agent like anhydrous sodium sulfate. Freeze-drying has also been recommended, but care must be exercised to prevent large nonreproducible losses of the analyte. For example, freeze-drying of milk leads to some loss of most PCB congeners. Losses may be as great as 50 – 74 % for monochlorobiphenyl, dichlorobiphenyl, and trichlorobiphenyl [152], [153].

Microwave drying permits a relatively soft drying of samples without hot spots and charring of the sample. As each molecule generates heat and the microwaves penetrate deeper into the sample, the whole volume is heated simultaneously. This deep penetration offers a substantial advantage over conventional heating, where heat can only penetrate into the material via the surface. The major benefit of microwave drying is the reduction of the drying time, which results in a higher sample throughput. Typical applications are for example the drying of wood, soil, herbs, and food [154-156].

3.1.2. *Homogenization and Aliquoting*

It is best to avoid homogenization prior to the determination of highly volatile analytes, since this often leads to nonreproducible losses. Liquid samples — preferably solids as well — should be transferred immediately after sampling to gas-tight sample vessels, followed by a direct determination on the basis of headspace gas chromatography (Section Headspace Techniques). If comminution is absolutely necessary, this should be accomplished in the deep-frozen state [157].

For relatively nonvolatile analytes, in such solid matrices as tissues, foods, or soil, the samples should be homogenized prior to aliquoting. This is often accomplished by trituration with sea sand, although use of a drying agent instead results in simultaneous removal of water, producing a homogenate suitable for immediate liquid extraction in a Soxhlet extractor (Section Soxhlet Extraction) or for supercritical fluid extraction (SFE; Section Supercritical Fluid Extraction (SFE)). Small amounts of sample can be homogenized in an ultrasound bath after the addition of water [158].

3.1.3. *Requirements with Respect to Materials and Chemicals*

Plastic containers are generally inappropriate for storing organic samples, since losses may be incurred through adsorption or migration through the vessel wall, and interfering substances may migrate into the sample from the container itself (plasticizers, antioxidants, monomers, etc.). In the case of analytes with high vapor pressures, volatilization losses must also be anticipated with plastic containers. Samples for organic trace analysis should always be stored in glass vessels with appropriate PTFE-sealed screw- cap closures or glass stoppers.

Materials and chemicals must be carefully checked for contamination before use, and if necessary cleaned and purified. Heating for several hours [159] or several days [160] is absolutely essential and very effective in the trace analysis of volatile analytes. Precleaning with boiling nitric acid may be necessary, as may a rinse with an organic solvent like acetone. Equipment for use in the determination of nonvolatile analytes is most effectively cleaned by rinsing with solvents that will be used subsequently in the course of the preparation process [161].

Glass vessels may also contribute to loss of an analyte, especially from aqueous solution, through adsorption on the vessel surface. This can be largely prevented by silanization of active sites on the glass. Alternatively, the vessel should be rinsed with solvent after removal of the sample, with the rinse solution then added to the sample itself [152].

Liquid – solid extraction procedures are accompanied by a serious risk of contamination. For example, blank tests with commercial extraction columns have revealed traces of plasticizers and antioxidants as well as various alkanes and alkenes [162]. For this reason certain manufacturers now offer solid-phase extraction columns made from glass. The extent of the po-tential contamination depends on the manufac-turer, the adsorbent load, conditioning steps, and the solvent used for the extraction [162], [163]. If no satisfactory commercial column is available, one can easily be prepared in the laboratory from a suitable adsorbent and an appropriate glass tube. For example, XAD can be purified to an extent sufficient for this purpose by Soxhlet extraction [164], [165], which reduces the level of impurities by a factor of 30. Silica gel and aluminum oxide can be purified by washing with solvents, followed by activation in a drying oven [160]. For example, unpurified silica gel has been shown to release as much as 160 ng/g of PCBs [163].

The risk of contamination must also be considered when a diffusive sampler is used to obtain air samples for determining volatile organic compounds. Thus, benzene, *n*-tetradecane, *n*-pentadecane, *n*-hexadecane, *n*-heptadecane, and dioctyl phthalate [166], along with trichloromethane, 1,1,1-trichloroethane, trichloroethene, and tetrachloroethene [167] have all been detected in extracts from unexposed diffusive samplers. The measured concentrations varied from batch to batch, ranging from 0.005 µg to 0.14 µg per diffusive sampler in the case of the chlorinated hydrocarbons [167], and from 0.1 – 2.6 µg per diffusion collector for the aliphatic hydrocarbons and dioctyl phthalate [166].

Ultrapure water can be prepared in the laboratory by filtration through an ion exchanger and activated charcoal. According to the manufacturer, one commercial water treatment system of this type reduces the residual hydrocarbon content of water to < 20 µg/L. Further reduction to < 5 µg/L is possible through UV irradiation (as with the “zero-water unit” manufactured by Gräntzel, Karlsruhe, Germany). If necessary, additional purification can be achieved with a stream of nitrogen at 80 °C [159], or by extraction with the solvent that will be used for extracting the analyte.

Most solvents are available commercially in sufficient purity, but blank test values should be established on a regular basis to ensure the absence of contaminants.

3.2. Separation of the Analyte

3.2.1. *Hydrolysis*

Many of the harmful substances and pollutants in biological samples, including their metabolites, are present wholly or in part in a conjugated (bound) form. This is the case, for example, with pentachlorophenol and phenol, which are excreted in the urine as glucuronide or glucuronide and sulfate, as well as with aromatic amines, present to some extent in the blood as hemoglobin conjugates. In such a situation the first step must accomplish release of the primary analytes. Conjugates can be cleaved by acid hydrolysis [168-171], basic hydrolysis [170], or enzymatic hydrolysis [170], [172]. Hydrolysis with the aid of enzymes, such as sulfatase and glucuronidase, has the significant advantage that bonds are cleaved extremely selectively, and information can also be obtained regarding the binding partners of the analyte. A disadvantage is the considerable amount of time required, typically several hours.

In many cases a portion of the matrix can be removed or converted into some more easily removable form before actual separation of the analytes is attempted. Relatively involatile, chemically stable analytes [e.g., organochlorine pesticides, PCBs, or polychlorinated dibenzo-*p*-dioxines (PCDDs)] can frequently be released from a matrix by acidic or alkaline hydrolysis with concentrated acid or base. Proteins can often be precipitated from biological samples by addition of acid, salt, or an organic solvent, or they can be broken down by enzymatic hydrolysis [173-175].

3.2.2. *Liquid – Liquid Extraction*

Extraction with organic solvents is the most common extraction technique for isolating an analyte from a liquid sample. The technique is based on distribution of the analyte between two immiscible liquid phases.

The decisive parameter with respect to extraction yield is the *distribution coefficient* for the analyte between the particular phases involved. A distribution coefficient can often be influenced advantageously by establishing a specific pH, thereby dividing the sample into strongly or weakly acidic, neutral, or basic fractions [176]. An example is provided by the extraction of aromatic amines from blood and urine [177]. Further possibilities include formation of ion pairs [178], complexation with metal salts, or salting-out of an aqueous phase.

If the distribution coefficient is sufficiently large, the simplest approach to liquid – liquid extraction is shaking the sample with an appropriate amount of an organic solvent. With smaller distribution coefficients or large sample volumes, continuous extraction or countercurrent extraction is required to achieve a complete separation. The apparatus for continuous extraction causes a liquid immiscible with the sample solution to circulate continuously and in finely divided form through the sample [179-181]. Extracted analytes are concentrated by distillation at appropriate times between individual extraction cycles.

Disadvantages of liquid – liquid extraction include the high dilution of the extract, which must subsequently be concentrated, and a high rate of solvent consumption. Both problems can be minimized by employing microextraction methods. Devices for extracting aqueous samples as large as 1 L with 200 µL of organic solvent have been described [182], [183].

Another variant of liquid – liquid extraction takes advantage of a liquid phase immobilized on a solid sorbent such as kieselguhr, Celite, Chromosorb W, or Chromosorb P [184]. The immobilized phase may be either aqueous or nonaqueous.

The formation of emulsions frequently presents problems in liquid – liquid extraction. Often such emulsions can be broken by centrifugation, freezing, or the addition of salt.

3.2.3. *Soxhlet Extraction*

Soxhlet extraction is chiefly applied to the separation of relatively nonvolatile analytes from solid samples. For example, this technique is useful for extracting pesticides, PCBs, and PCDD/PCDFs from various matrices, including fatty tissue, soil, and paper [185], [186]. The advantage of the method is high yield, achieved by continuous extraction of the sample with fresh extractant. Disadvantages include the considerable amount of time required (typically several hours to several days), the thermal stress to which analytes are subjected, and the fact that the analytes are obtained in very dilute form in a solvent, and must therefore be concentrated in a subsequent step. Any fat present in the sample is co-extracted, and this must also be removed separately.

3.2.4. *Microwave-Assisted Solvent Extraction*

Microwave assisted solvent extraction (MASE) is a method which uses microwave radiation to heat the extracting solvent quickly without primarily heating the vessel. Therefore, the rate of heating is faster than in conventional methods. By using a closed system, extraction can be performed at elevated temperatures and pressures, which enhances the extraction rates and minimizes extraction time and solvent consumption. Closed-vessel systems also allow solvents to be heated to extraction temperatures distinctly higher than their atmospheric boiling points. In commercial MASE systems, a number of extraction vessels can be located in a carousel which can be extracted simultaneously. Target analytes in MASE are organic pollutants such as semivolatiles, PAH, PCB, dioxins, organochlorine pesticides and herbicides, organophosphorous pesticides, and phenols from environmental samples such as soils, plants, and waters [187-190] or the extraction of fat from food [191]. MASE can in many cases be used in place of Soxhlet extraction.

However, MASE has some drawbacks, which have to be kept in mind. Because solvent extractions typically involve nonpolar solvents with very small dielectric constants, which do not absorb microwave energy and were thus not heated. Thus a cosolvent must often be added to the solvent to aid heating of the solution. The use of a cosolvent leads to the extraction of a broader spectrum of compounds which can cause interference problems during the analyses of the analytes of interest [187]. Solvents which strongly absorb microwaves are for example alcohols such as methanol, ethanol, propanol, and butanol. Solvents such as toluene, hexane, tetrahydrofurane, and tetrachloromethane absorb microwaves only poorly. Water can also be used as solvent. Typical applications using water as extractant are the extraction of pesticides and acid herbicides from solid samples such as soil [189], [192].

3.2.5. *Supercritical Fluid Extraction (SFE)*

An extraction method long practiced in industry — supercritical fluid extraction (SFE) — has also been introduced into the analytical sector. As the name suggests, SFE employs supercritical fluids for extraction purposes in place of the organic solvents of conventional extraction.

A compound is said to be in a supercritical state when the critical pressure and critical temperature for that particular material are exceeded. Critical parameters for several relevant compounds are listed in Table 6. Supercritical fluids exhibit simultaneously properties associated with both gases and liquids (see Table 7). Thus, like gases, they are compressible, but they also display solvencies similar to those of liquids. Any increase in temperature at constant pressure reduces the solvent power of a supercritical fluid, but it also leads to an increase in diffusion rate, which tends to lower the minimum

required extraction time. Compared to conventional extractants, supercritical fluids have low viscosities and diffusion rates that are higher by a factor of 10 – 100, both of which contribute to reduced extraction times. With supercritical CO_2 and N_2O, which are gases under normal conditions, the extractant is separated by reducing the pressure to atmospheric levels, leading to simultaneous concentration of the extract.

Table 6. Critical parameters for selected substances [193]

Substance	Critical temperature, °C	Critical pressure, MPa
Xe	16.6	5.83
CHF_3	25.9	4.83
$CClF_3$	28.8	3.92
CO_2	31.1	7.37
N_2O	36.4	7.24
NH_3	132.2	11.27
CH_3OH	239.4	8.09
H_2O	374.1	22.04

Table 7. Typical characteristics of gases, liquids, and supercritical fluids

Property	Gas	Liquid	Supercritical fluid
Density, g/mL	$10^{-4} – 10^{-3}$	0.6 – 1.4	0.1 – 1
Diffusion coefficient, cm^2/s	10^{-1}	10^{-5}	$10^{-3} – 10^{-4}$
Solvency	no	yes	yes
Compressibility	yes	no	yes

Supercritical CO_2 is the most frequently used extractant for SFE. It has the advantage of being chemically rather inert, and its critical temperature is low, so it is valuable for the extraction of such thermolabile analytes as steroids and fragrances. The low critical pressure for CO_2 opens the way to a relatively broad range over which the solvency can be varied through adjustment of the pressure. Other advantages of CO_2 as an extractant include low toxicity, high purity, and low cost. The principal disadvantage of CO_2 is a relatively low polarity. Its solvent power with respect to polar analytes can be improved, however, by adding such polar *modifiers* as methanol, acetone, hexane, or dichloromethane.

Extraction with supercritical CO_2 has been recommended for separating a wide variety of analytes, including pesticides, PCBs, vitamins, and fragrances from meat, fish, baby food, and animal feed [194-199]. PAHs, PCBs, PCDO/PCDFs, and other substances have been extracted from soil, fly ash, sediment, air particles, polymers, and plants using supercritical CO_2 together with a modifier [200-210].

3.2.6. *Solid-Phase Extraction (SPE)*

Solid-phase extraction (SPE) is used for the selective separation and concentration of analytes from liquid samples. Extraction of the analytes is based in this case on the distribution of dissolved substances between a solid-phase surface and the sample liquid. Separation of various sample constituents may be a result of differing polarities, differences in molecular size, or differences with respect to ion-exchange capacity.

Many of the comments above regarding liquid – liquid extraction also apply to solid-phase extraction. However, the latter has the advantages that it can be accomplished more rapidly, requires less solvent, provides more highly concentrated extracts, and is relatively easy to automate. Laboratory robots and suitable column arrangement have rendered large parts of the sample preparation process automatic, thereby facilitating reproducibility and efficient operation [211], [212].

Many different adsorbents are applicable in this context (see Table 8). The most common adsorbents for solid-phase extraction are based on silica gel, the surface of which has been modified in some way.

Table 8. Important adsorption agents classified according to type

Nonpolar	Polar	Anion exchangers	Cation exchangers
Octadecyl	Cyano	Primary amine	Carboxylic acid
Octyl	Kieselguhr	Secondary amine	Sulfonic acid

Butyl	Silica gel	Quaternary ammonium salt
Cyclohexyl	Florisil	
Phenyl	Aluminum oxide	
Amino		
Diol		

Octadecyl surface phases (C_{18}) are used for the reversed-phase extraction of nonpolar substances from aqueous solutions. Typical applications include the extraction of organochlorine pesticides [213-215], organophosphorus pesticides [216], chlorinated and unchlorinated hydrocarbons, triazines [214], [215], PAHs and nitro-PAHs [214], carbamates [214], chlorophenols [214], [217], aflatoxins [218], plasticizers [219], vitamins [220], and medicaments such as barbiturates and antibiotics [221], [222].

Shorter octyl phases (C_8) are used for extracting substances of medium polarity. Substances that bind irreversibly to C_{18} phases can often be concentrated and re-eluted successfully with C_8 phases.

Normal phase materials, including unmodified silica gel, aluminum oxide, and Florisil, separate sample constituents into fractions of comparable polarity. They are often utilized to separate and concentrate pesticides [223-225], PCBs [152], [223], and PCDD/PCDFs [226-230] from such biological samples as blood, breast milk, fatty tissue, and foods. Normal phases are also used to extract polar sample constituents, such as amines, alcohols, phenols, dyes, medicaments, or vitamins [231].

Anion exchangers are mainly employed in the extraction of carbohydrates, peptides, nucleosides, and amino acids. Cation exchangers are useful for extracting amino acids and nucleosides [231].

Apart from modified silica gel, the most frequently used solid-phase adsorbents are activated charcoal, macroreticular resins (XAD), ordinary silica gel, aluminum oxide, and Florisil. Activated charcoal is a universal adsorbent for concentrating trace organic materials in aqueous solutions and air. XAD resins are also commonly employed for extracting organic trace constituents, such as alkoxyacetic acids [165], organochlorine pesticides [214], [232], carbamates [212], [216], and triazines [232] from aqueous matrices.

The activity of these solids must be accurately adjusted to ensure reproducible results and well-defined fractions, and sample solutions must always be carefully dried prior to extraction.

Exceptionally low limits of detectability can be achieved with automated on-line systems. Here the complete eluate is analyzed rather than simply an aliquot, which lowers the risk of evaporative losses and sample contamination. Automated sample preparation through solid-phase extraction is becoming increasingly important in extreme trace analysis, as in the determination of PCDDs in tissue or in the blood of uncontaminated persons [233].

A variant of solid-phase extraction is extraction with the aid of *extraction disks* [234], [235], membrane filter plates on which an appropriate solid-phase material has been immobilized. Advantages of extraction disks include higher flow rates and more rapid material exchange. Another variant involves the use of SPE for simultaneous extraction and derivatization. For this purpose the analyte is adsorbed onto a solid phase that has been impregnated with a derivatizing agent, permitting the separation and concentration of analytes that would otherwise not be adsorbed at all. Solid-supported reagents are useful in such applications as the determination of organic acids or cannabinoids in plasma [236], [237].

3.2.7. *Solid-Phase Microextraction (SPME)*

A new technique, which is applicable for sampling in air and liquids or in the headspace above a liquid or a solid sample, is solid-phase microextraction (SPME). The mechanism of SPME, which has been developed by PAWLISZYN et al. [239], [240], is based on the partition equilibrium of the analytes between the sample or the headspace above the sample, respectively, and a fused silica fiber coated with a suitable stationary phase. The amount of analyte extracted by the fiber is proportional to the initial analyte concentration in the sample and depends on the type of fiber. After sampling, the fiber can be thermally desorbed directly into the injector of a gas chromatograph. SPME combines sampling, analyte enrichment, matrix separation, and sample introduction within one step [240]. Since its development, this innovative technique has found widespread use in environmental analysis. It has, for example, been applied in the determination of volatile organic compounds [241], [242], phenols [243], pesticides [244], polyaromatic hydrocarbons, and polychlorinated biphenyls [245] in water.

SPME fibers have been used as air sampling devices for volatile organic compounds in ambient and workplace air and give results that are in good agreement with traditional sampling methods [246]. Furthermore, grab sampling is used with stainless steel canisters or glass bulbs in combination with SPME [240], [247]. The dependence of the sampling rate on humidity and air temperature can be eliminated by correction factors [247]. The main drawback is the low storage stability of the samples, due to uncontrolled losses of analytes by adsorption on the walls of the canisters or by evaporation from the loaded fiber.

3.2.8. *Stir-Bar Adsorptive Extraction (SBSE)*

Another new approach for adsorptive sample enrichment is stir-bar adsorptive extraction (SBSE). Sorption is based on dissolution of the analyte in a liquid polymeric material (e.g., polydimethylsiloxane) which is covering a stir bar. This stir bar can be placed into liquid samples such as waters, waste water, beverages, and body fluids. A multitude of analytes such as pesticides, herbicides, fungicides, flavours, and polycyclic aromatic hydrocarbons [248-253] is dissoluted in this polydimethylsiloxane film and in this way extracted and enriched without using organic solvents. These analytes are

subsequently determined by thermal desorption in combination with capillary gas chromatography. SBSE can be up to a factor of 1000 more sensitive than SPME. A review comparing SBSE and SPME for sorptive sample preparation is given in [254].

3.2.9. *Miscellaneous Techniques*

An alternative approach to separating volatile compounds from liquid and solid samples is distillation. Simple distillation can accomplish the isolation of a volatile analyte from a nonvolatile residue, or separate multiple sample constituents with widely differing boiling points. A further application is separation of a sample into fractions with different boiling ranges.

Steam distillation is an effective way to separate from a matrix such steam-volatile analytes as phenol [238], alkylphenols [168], or formic acid [255]. Steam distillation offers the advantage that the analyte is recovered in a nearly matrix-free condition, although it is diluted with a large amount of water from which it must subsequently be separated and concentrated. Solid-phase extraction is particularly useful for this purpose (Section Solid-Phase Extraction (SPE)).

Another technique involving distillative concentration of sample constituents is *sweep co-distillation*. Here the sample is treated with a highly volatile solvent introduced with the aid of a stream of carrier gas. The solvent in turn transports soluble components of the sample to a cooled distillation receiver. The method is useful for such applications as the isolation of volatile pesticides from animal and plant fats [256], [257].

Membrane techniques, including dialysis, ultrafiltration, and reverse osmosis, are also applicable to sample preparation problems [258-260]. *Dialysis* separates analytes on the basis of their ability to diffuse through a membrane as a result of a concentration gradient. It is most frequently used for concentrating analytes in biological liquids. Automation and on-line coupling with liquid chromatography have both been reported [261], [262]. *Reverse osmosis* is very similar to dialysis, but in this case a pressure difference is used to cause the solvent to migrate through a membrane from a region of high analyte concentration to one of low analyte concentration. Reverse osmosis is frequently called upon for concentrating large volumes of very dilute solutions. *Ultrafiltration* is a method for concentrating large sample molecules present in dilute solution. Separation in this case depends directly on molecular size, and occurs when a sample solution is filtered through a membrane with an appropriate pore size. One possible application is in the separation of free molecules present in a biological fluid from similar molecules that are bound to receptors [263]; another is the desalination of a liquid sample.

3.3. Headspace Techniques

3.3.1. *Static Headspace Technique*

Highly volatile analytes can be separated for subsequent quantitative determination by one of several *headspace techniques*. All headspace techniques are based on the Henry – Dalton law, which states that, in a closed system, the vapor-space concentration of an analyte depends exclusively on the temperature and the corresponding analyte concentration in solution or on a solid surface. Consequently, raising the temperature causes volatile analytes to separate from their matrix and become concentrated in the surrounding vapor space. In the *static* headspace technique, a gas chromatographic determination is made of the analyte once equilibrium has been established. The separation efficiency characteristic of gas chromatography and the virtual absence of matrix-derived detector noise means that chlorinated, aromatic, and aliphatic hydrocarbons, along with alcohols, ketones, and esters, are all subject to determination essentially free of interference regardless of the matrix. Limits of detectability in the µg/L range permit volatile organic compounds to be determined in body fluids from persons occupationally exposed to contaminants [264], [265], as well as in foods [266] and water [267]. Nevertheless, the distribution coefficient does depend on the matrix, so a relatively complicated and time-consuming calibration is essential. Depending on the matrix, calibration can be based on matrix standards, standard addition, or internal standardization.

3.3.2. *Dynamic Headspace Technique (Purge and Trap)*

In the dynamic headspace technique, a carrier gas is passed continuously through the sample under investigation. Constant contact with fresh carrier gas causes volatile analytes to be removed gradually and almost completely from the sample. These are then concentrated on an adsorbent such as Tenax or Chromosorb. Quantitative transfer of the analytes to a gas chromatographic column is accomplished by thermodesorption, using either on-line or off-line techniques.

The concentration that results (typically by a factor of ca. 100) leads to limits of detectability in the ng/L range, permitting the simultaneous analysis of environmentally conditioned concentrations of as many as 48 volatile organic compounds in water and such body fluids as blood and urine [149], [268-270].

3.4. Determination of Trace Organic Materials in Air Samples

Air sampling for analysis of gases and vapors is accomplished either with the aid of a pump and subsequent concentration on a solid adsorbent (*active sampling*), or with a collector, in which case analytes reach the collecting phase via diffusion or permeation (*passive sampling*). Gaseous trace organic materials in air are often collected by adsorption on activated charcoal, Tenax, or XAD resin, although other adsorbents such as aluminum oxide or Florisil can also be used.

Analytes are usually desorbed from the adsorbent by washing with appropriate solvents, after which the eluates can be analyzed by gas chromatography. Carbon disulfide is the most commonly employed desorption agent [271]. For the desorption of polar analytes, small amounts of such polar solvents as methanol, 2-propanol, or 2-butanol can be added to the carbon disulfide as a way of increasing the desorption yield [272].

Analytes can also be desorbed thermally and analyzed on-line with a gas chromatograph. This expedient results in a significant increase in sensitivity, since the entire sample is utilized for the analysis. One disadvantage is that the binding power of activated charcoal (the most suitable adsorption material for most analytes) is too great to be overcome with conventional thermodesorbers, eliminating the possibility of using this combination for quantitative analysis. Quantitative desorption from charcoal can be accomplished with a microwave thermodesorption device, however, and this microwave thermodesorption is a mild process, suitable even for such labile analytes as fragrances [273].

Particle-bound substances like polycyclic aromatic hydrocarbons (PAH) are determined by collecting particulate air constituents on glass-fiber or PTFE filters, or in impactors, followed by extraction.

3.5. Analyte Concentration

Subsequent to one of the aforementioned extraction procedures (e.g., liquid – liquid extraction, steam distillation, Soxhlet extraction) the analytes are obtained in extremely dilute form in a large volume of solvent. Various approaches to concentration are possible depending on the nature of the analyte, the solvent, the initial volume, and the target volume.

Aqueous analytes, such as those from a steam distillation, can be concentrated by either solid-phase or liquid – liquid extraction (Sections Liquid – Liquid Extraction and Solid-Phase Extraction (SPE)). An analyte dissolved in a highly volatile solvent (as with liquid – liquid or Soxhlet extraction) is concentrated most effectively by evaporation of the solvent in a gas stream (e.g., nitrogen, helium), or with a rotary evaporator or a Kuderna – Danish concentrator. Losses must be anticipated with a rotary evaporator, however, especially of volatile analytes, though even less volatile constituents may also be lost through codistillation. Recoveries of volatile analytes are higher with a Kuderna – Danish concentrator. Concentrating an extract by solvent evaporation in a gas stream is appropriate only with relatively small volumes on account of the low vaporization rate. Analyte loss is possible through aerosol formation or evaporation.

3.6. Derivatization

Separation and concentration of an analyte must often be followed by some type of *derivatization*. Derivatization is conducted with one or more of the following objectives:

1. To facilitate chromatographic separation
2. To increase selectivity
3. To improve the limit of detection

The first objective applies mainly to gas chromatography, and involves the preparation of a more volatile or less polar form of the analyte. Improved selectivity is generally less important in this case due to the large number of theoretical plates available with capillary gas chromatography columns, in contrast to HPLC. However, the limit of detection plays an important role in both gas and liquid chromatography.

Most derivatizations for gas chromatography are esterifications or etherifications. For example, analytes containing carboxyl groups are often converted into methyl [164], [169], pentafluorobenzyl [274-277], or triethyl esters [278-280]. Analytes containing acidic hydroxyl groups (phenols, chlorophenols, glycol ethers) or amino groups (e.g., aniline) can be derivatized with such perfluorinated compounds as heptafluorobutyric anhydride, pentafluoropropionic anhydride [177], or pentafluorobenzoyl chloride [281].

Derivatization with halogenated compounds offers the advantage that the derivatives are subject to extremely sensitive detection with an electron capture detector (ECD). The disadvantage is that excess derivatization agent remains in the sample solution after the reaction, necessitating its removal prior to gas chromatographic determination because of the potential for interference during ECD detection.

Derivatizations for HPLC are designed mainly to improve the limit of detection, permitting the use of highly sensitive or selective detectors inapplicable to the analytes themselves. Enhanced absorption of UV/visible light is achieved by the introduction of chromophoric groups. Analytes can also be rendered fluorescent by the introduction of fluorophoric groups.

Carboxylic acids, including formic, acetic, lactic, propionic, malic, tartaric, and citric acids, can be transformed with benzyl, naphthacyl, phenacyl, or bromophenacyl bromides [282], [283]; *p*-nitrobenzyl bromide [284]; *p*-nitrophenacyl bromide; methoxyaniline; or *p*-nitrobenzyl-*N*,*N*′-diisopropylurea [285], [286] into esters that absorb UV or visible light. α-Keto acids (e.g., glycolic, glyoxylic acids) are detectable with UV light after derivatization with phenylhydrazine [287], [288].

Fluorescent compounds are obtained by reacting carboxylic acids with 4-bromomethyl-7-methoxycoumarin [289-292] or 4-bromomethyl-7-acetoxycoumarin [293], [294].

Analytes containing hydroxyl groups, such as phenols, glycols, and alcohols, can be converted with 3,5-dinitrobenzoyl chloride [295] or dabsyl chloride [296-298] into compounds that absorb UV or visible light. Fluorescent derivatives can be obtained with 7-(chlorocarbonyl) methoxy-4-methylcoumarin [299].

Derivatizations for HPLC purposes are accomplished either off-line or on-line. An on-line process may involve either pre-column or post- column reaction depending on the objective.

3.7. Coupled Techniques

Various coupled sample preparation and determination processes are increasingly utilized in trace organic analysis, whereby separated analytes pass directly in an on-line way into a chromatographic analysis system. The main advantages of such on-line procedures reflect the quantitative nature of analyte transfer from the sample preparation stage to the analytical determination, leading to an optimal limit of detection. This approach also minimizes the risk of sample loss and contamination. However, coupled techniques are relatively difficult to imple-ment, since parameters must be optimized not only with respect to sample preparation but also for the subsequent chromatographic separation. One of the most important coupled techniques at the present time is LC – GC. Automatic sample preparation is in this case accomplished via HPLC, permitting the separation of a complex matrix, or even a concentration of trace material. A subsequent capillary gas chromatographic analysis results in a high degree of resolution (with isomeric materials, for example) and the high signal-to-noise ratio essential for trace analysis thanks to the availability of such extremely sensitive and selective devices as electron capture detectors (ECD), flame photometry detectors (FPD), phosphorus – nitrogen-selective detectors (PND), and mass spectrometric detectors (MSD) [300-302]. Coupling is achieved with a sample loop (loop technique), which is

used to transfer a small portion of the LC eluate to an uncoated GC precolumn (retention gap) [303]. Determinations of pesticide metabolites from maize [304] and organochlorine pesticides from fat [305] constitute two examples of the widespread application of this technique.

LC – MS coupling is also becoming an increasingly common analytical method now that effective interfaces have been developed (particle beam, thermospray, electrospray). Determinations of pesticides in sea water [234], carbamates and phenylurea derivatives in water samples [306], and organophosphorus pesticides and chlorophenols [307] are just a few examples of the use of this method.

LC – TLC coupling constitutes one possibility for carrying out multidimensional LC. Thus HPLC (usually reversed-phase chromatography) is coupled with thin layer (adsorption) chromatography in such a way that the eluate from the HPLC column is transferred via a capillary column to a mechanically transported TLC plate [308], [309].

The coupling of SFE with gas chromatography has also been described, as has chromatography with supercritical fluids (SFC). Examples include the separation and determination of PCBs, PAHs, and pesticides in such environmental samples as soil and sediments [310-313]. A frequently employed technique for determining traces of pesticide in aqueous samples is the on-line coupling of solid-phase extraction or dialysis with HPLC [261], [262], [314].

[Top of Page]

References

1. P. Tschöpel: “Sample Treatment” in M. Stoeppler (ed.): *Hazardous Metals in the Environment,* vol. 12, Elsevier, Amsterdam 1992.
2. O. Behne, *J. Clin. Chem. Clin. Biochem.* **19** (1981) 115. Links
3. G. Knapp, *Int. J. Environ. Anal. Chem.* **22** (1985) 71. Links
4. D. v. Renterghem, R. Cornelis, R. Vanholder, *Anal. Chim. Acta* **257** (1992) 1.
5. J. Trettenbach, K. G. Heumann, *Fresenius Z. Anal. Chem.* **322** (1985) 306.
6. L. B. Fischer, *Anal. Chem.* **58** (1986) 261. Links
7. B. Grieping, H. Marchandise in: *Analytiker Taschenbuch,* vol. 6, Springer Verlag, Berlin 1986, p. 3.
8. DFG Deutsche Forschungsgemeinschaft in J. Angerer, K. H. Schaller (eds.): *Analyses of Hazardous Substances in Biological Materials*. vol. **5**, VCH Verlagsgesellschaft, Weinheim, Germany 1997, pp. XXI – XXII.
9. J. Begerow, L. Dunemann, in L. Matter (ed.): *Elementspurenanalytik in biologischen Matrices*, Spektrum Akademischer Verlag, Heidelberg – Berlin – Oxford 1997, p. 27.
10. D. Behne, G. V. Iyengar: “Spurenelementanalyse in biologischen Proben”, *Analytiker-Taschenbuch,* vol. 6, Springer Verlag, Berlin 1986.
11. M. Kiilunen, J. Järvisalo, O. Mäkitie, A. Aitio, *Int. Arch. Occup. Environ. Health* **59** (1987) 43. Links
12. G. V. Iyengar, B. Sansoni: “Elemental Analysis of Biological Materials: Current Problems and Techniques with Special Reference to Trace Elements,” *IAEA Technical Report 197*, Wien 1980, p. 73.
13. G. V. Iyengar, K. Kasperek, L. E. Feindegen, *Sci. Total Environ.* **10** (1978) 1. Links
14. J. Angerer: “Spezielle Vorbemerkungen,” in D. Henschler (ed.): *Analysen in biologischem Material,* 10th suppl., VCH Verlagsgesellschaft, Weinheim 1991.
15. J. C. Meranger, B. L. Hollebone, G. A. Blanchette, *J. Anal. Toxicol.* **33** (1981) 5.
16. P. Tschöpel *et al.*, *Fresenius Z. Anal. Chem.* **302** (1980) 1.
17. K. Heydorn: *Neutron Activation Analysis for Clinical Trace Element Research*, vol. **1**, CRC Press, Boca Raton Fla. 1984, p. 42.
18. J. R. Moody, P. Lindström, *Anal. Chem.* **47** (1977) 2264.
19. S. P. Ericson *et al.*, *Clin. Chem. (Winston-Salem N.C.)* **32** (1986) 1350.
20. S. B. Adeloju, A. M. Bond, *Anal. Chem.* **57** (1985) 1728. Links
21. R. Cornelis et al. in: *Nuclear Activation Techniques in the Life Sciences,* IAEA, Wien 1979, p. 165.
22. J. Begerow, L. Dunemann in M. Stoeppler (ed.): *Sampling and Sample Preparation*, Springer-Verlag, Berlin – Heidelberg – New York 1997, p. 155.
23. P. Lievens, J. Versieck, R. Cornelis, J. Hoste, *J. Radioanal. Chem.* **37** (1977) 483.
24. A. Abu-Samra, J. S. Morris, S. R. Koirtyohann, *Anal. Chem.* **47** (1975) 1475. Links
25. E. D. Neas, M. J. Collins in H. M. Kingston, L. B. Jassie (eds.): *Introduction to Microwave Sample Preparation*, Am. Chem. Soc., Washington, D.C. 1988, p. 7.
26. R. A. Nadkarni, *Anal. Chem.* **56** (1984) 2233. Links
27. Z. Sulcek, P. Povondra: *Methods of Decomposition in Inorganic Analysis*, CRC Press, Boca Raton Fla. 1989.
28. P. Aysola, P. Anderson, C. H. Langford, *Anal. Chem.* **59** (1987) 1582. Links
29. R. T. White Jr., G. E. Douthit, *J. Assoc. Off. Anal. Chem.* **68** (1985) 766.
30. L.-Q. Xu, W.-X. Shen, *Fresenius Z. Anal. Chem.* **332** (1988) 45.
31. G. Knapp, *Int. J. Environ. Anal. Chem.* **22** (1984) 71.
32. G. Knapp, *Fresenius Z. Anal. Chem.* **317** (1984) 213.
33. K. W. Budna, G. Knapp, *Fresenius Z. Anal. Chem.* **294** (1979) 122.
34. F. R. Abou-Shakra, M. P. Rayman, N. I. Ward, V. Hotton, G. Bastian, *J. Anal. At. Spectrom.* **12** (1997) 429.
35. H. Hodrejarv, A. Vaarmann, *Anal. Chim. Acta* **396** (1999) 293. Links

36. M. Bettinelli, C. Baffi, G. M. Beone, S. Spezia, *At. Spectrosc.* **21** (2000) 50.
37. A. C. Grillo, *Spectrosc. Int.* **1** (1989) 16.
38. M. Kolb, P. Rach, J. Schäfer, A. Wild, *Fresenius J. Anal. Chem.* **342** (1992) 341. Links
39. L. Dunemann in B. Welz (ed.): *5th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1989, p. 593.
40. L. Dunemann, M. Meinerling, *Fresenius J. Anal. Chem.* **342** (1992) 714. Links
41. J. Hertz, R. Pani, *Fresenius Z. Anal. Chem.* **328** (1987) 487.
42. F. Wahdat, R. Neeb, *Fresenius Z. Anal. Chem.* **335** (1989) 748.
43. J. Begerow, M. Turfeld, L. Dunemann, *J. Anal. Atomic Spectrometry* **11** (1997) 1095.
44. P. Schramel, I. Wendler, P. Roth, E. Werner, *Mikrochim. Acta* **116** (1994) 205. Links
45. D. Florian, G. Knapp, *Anal. Chem.* **73** (2001) 1515. Links
46. L. Dunemann, *Nachr. Chem. Tech. Lab.* **39** (1991) no. 10, M1.
47. M. Würfels, E. Jackwerth, M. Stoeppler, *Fresenius Z. Anal. Chem.* **329** (1987) 459.
48. P. Schramel, S. Haase, G. Knapp, *Fresenius Z. Anal. Chem.* **326** (1987) 142.
49. P. Schramel, G. Lill, R. Seif, *Fresenius Z. Anal. Chem.* **326** (1987) 135.
50. B. Zunk, *Anal. Chim. Acta* **236** (1990) 337. Links
51. L. B. Fischer, *Anal. Chem.* **58** (1986) 261. Links
52. H. Matusiewicz, *Anal. Chem.* **71** (1999) 3145. Links
53. M. Deaker, W. Maher, *J. Anal. At. Spectrom.* **14** (1999) 1193.
54. D. Merton, J. A. C. Broekaert, R. Brandt, N. Jakubowski, *J. Anal. At. Spectrom.* **14** (1999) 1093.
55. R. A. Romero, J. E. Tahán, A. J. Moronta, *Anal. Chim. Acta* **257** (1992) 147. Links
56. M. S. Chaudhary, S. Ahmad, A. Mannan, I. H. Qureshi, *J. Radioanal. Nucl. Chem.* **83** (1984) 387.
57. M. Feinberg, C. Ducauze, *Anal. Chem.* **52** (1980) 207. Links
58. D. Huljev, B. Huljev, Z. Rajkovic-Huljev, *Radiol. Iugosl.* **22** (1988) 403.
59. G. Kaiser, G. Tölg, *Fresenius Z. Anal. Chem.* **325** (1986) 32.
60. P. Barth, R. Caletka, V. Krivan, *Fresenius Z. Anal. Chem.* **319** (1984) 560.
61. S. E. Raptis, G. Kaiser, G. Tölg, *Anal. Chim. Acta* **138** (1982) 93. Links
62. J. Binkowski, P. Rutkowski, *Mikrochim. Acta* **1** (1987) 245.
63. M. Kulke, F. Umland, *Fresenius Z. Anal. Chem.* **288** (1977) 273.
64. M. Fujita *et al.*, *Anal. Chem.* **40** (1968) 2042. Links
65. S. S. Q. Hee, J. R. Boyle, *Anal. Chem.* **60** (1988) 1033. Links
66. D. A. Levaggi, W. Qyung, M. Feldstein, *J. Air. Pollut. Control Assoc.* **21** (1971) 277. Links
67. http://www.anton-paar.com/ap/apinternet/html/default/CXSN-5WXDB5.en.0.jsp
68. G. Schwedt, L. Dunemann, *LaborPraxis*, June 1990, 476.
69. S. E. Raptis, G. Knapp, A. P. Schalk, *Fresenius Z. Anal. Chem.* **316** (1983) 482.
70. D. Behne, P. A. Matamba, *Fresenius Z. Anal. Chem.* **274** (1975) 195.
71. M. Gallorini, E. Orvini, A. Rolla, M. Burdisso, *Analyst (London)* **106** (1981) 328.
72. C. Feldmann, *Anal. Chem.* **55** (1983) 2451.
73. H. Brügmann, T. X. Gian, H. Berge, *Acta Hydrochim. Hydrobiol.* **16** (1988) 457.
74. M. Weidenauer, K. H. Lieser, *Fresenius Z. Anal. Chem.* **320** (1985) 550.
75. T. M. Karadahki, F. M. Najib, F. A. Mohammed, *Talanta* **34** (1987) 995.
76. A. Campiglio, *Mikrochim. Acta* **3** (1987) 425.
77. P. Ostapczuk, M. Froning, M. Stoeppler, *Fresenius Z. Anal. Chem.* **334** (1989) 61.
78. K. Gretzinger, E. Grallath, G. Tölg, *Anal. Chim. Acta* **193** (1987) 1. Links
79. H. Heinrichs, H. J. Brumsack, N. Loftfield, N. Kö-nig, *Z. Pflanzenernähr. Bodenkd.* **149** (1986) 350.
80. S. Sprung, G. Kirchner, W. Rechenberg, *Zem.-Kalk-Gips, Ed. B* **37** (1984) 513.
81. B. Morsches, G. Tölg, *Fresenius Z. Anal. Chem.* **219** (1966) 61.
82. P. C. Lindahl, A. M. Bishop, *Fuel* **61** (1982) 658. Links
83. R. Wickbold, *Angew. Chem.* **64** (1952) 134. Links
84. P. Schramel, S. Hasse, *Mikrochim. Acta*, **116** (1994) 205. Links
85. J. Dolezal, J. Lenz, Z. Sulcek, *Anal. Chim. Acta* **47** (1969) 517. Links
86. G. Tölg, *Pure Appl. Chem.* **44** (1975) 645.
87. J. T. Brenna, G. H. Morrison, *Microbeam Anal.* **19th** (1984) 265.
88. W. van Delft, G. Vos, *Anal. Chim. Acta* **209** (1988) 147. Links
89. L. Vos, R. van Grieken, *Anal. Chim. Acta* **164** (1984) 83. Links
90. M. Bettinelli, U. Baroni, N. Pastorelli, *J. Anal. At. Spectrom.* **2** (1987) 485.
91. F. Koroleff, Baltic Intercalibration Workshop, Kiel March 7 – 14, 1977, p. 30.
92. K. Grasshoff, M. Ehrhardt, K. Kremling (eds.): *Methods of Seawater Analysis*, Verlag Chemie, Weinheim, Germany 1983, p. 164.
93. L. Dunemann, M. Meinerling, unpublished results (1992).

94. W. Rechenberg, *Zem. Kalk Gips* **25** (1972) 410.
95. F. Seel, E. Steigner, J. Burger, *Angew. Chem.* **76** (1964) 532. Links
96. J. Angerer, R. Heinrich-Ramm, G. Lehnert, *Int. J. Environ. Anal. Chem.* **35** (1989) 81. Links
97. E. Schumacher-Wittkopf, J. Angerer, *Int. Arch. Occup. Environ. Health* **49** (1981) 77. Links
98. R. Heinrich, J. Angerer in P. Brätter, P. Schramel (eds.): *Trace Element-Analytical Chemistry in Medicine and Biology,* vol. 2, Walter De Gruyter, Berlin 1983.
99. A. Dornemann, H. Kleist, *Fresenius Z. Anal. Chem.* **300** (1980) 197.
100. B. Griepink, M. Sager, G. Tölg, *Pure Appl. Chem.* **60** (1988) 1425. Links
101. M. Buratti *et al.*, *Clin. Chim. Acta* **150** (1985) 53. Links
102. R. Kobayashi, K. Imaizumi, *Anal. Sci.* **5** (1985) 61.
103. H. Watanabe, *Anal. Chem.* **53** (1981) 738. Links
104. H. Berndt, U. Harms, M. Sonneborn, *Fresenius Z. Anal. Chem.* **322** (1985) 329.
105. P. Burba, P. G. Willmer, *Fresenius Z. Anal. Chem.* **329** (1987) 539.
106. G. Schwedt: *Methoden der Spurenanreicherung anorganischer und organischer Stoffe aus Wässern*, Vogel-Verlag, Würzburg 1988.
107. A. Taylor, S. Branch, H. M. Crews, D. F. Halls, L. M. Owen, M. White, *J. Anal. At. Spectrom.* **12** (1997) 119R.
108. M. Hoenig, A.-M. de Kersabiec, *Spectrochim. Acta B* **51** (1996) 1297. Links
109. J. Begerow, L. Dunemann in H. Günzler (ed.): *Analytiker-Taschenbuch*, vol. **18**, Springer-Verlag, Berlin – Heidelberg – New York 1998, p. 67.
110. T. Mansfeld, H. Biernath, *Anal. Chim. Acta* **406** (2000) 283. Links
111. D. M. Adriá-Cerezo, M. Llobat-Estellés, A. M. Mauri-Aucejo, *Talanta* **51** (2000) 531. Links
112. A. Seubert, G. Petzold, J. W. McLaren, *J. Anal. At. Spectrom.* **10** (1995) 371.
113. I. Jarvis, M. M. Totland, K. E. Jarvis, *Analyst* **122** (1997) 19. Links
114. J. Begerow, M. Turfeld, L. Dunemann, *Anal. Chim. Acta* **340** (1997) 277. Links
115. V. K. Panday, K. Hoppstock, J. S. Becker, H.-J. Dietze, *At. Spectrosc.* **17** (1996) 98.
116. K. E. Laintz, C. M. Wai, C. R. Yonker, R. D. Smith, *Anal. Chem.* **64** (1992) 2875. Links
117. C. M. Wai, S. Wang, J.-J. Yu, *Anal. Chem.* **68** (1996) 3516. Links
118. G. Schulze, O. Elsholz, *Fresenius Z. Anal. Chem.* **335** (1989) 724.
119. M. Agarwal, R. B. Bennett, I. G. Stump, J. M. D'Auria, *Anal. Chem.* **47** (1975) 924. Links
120. M. M. Kingston, I. L. Barnes, T. J. Brady, T. C. Rains, *Anal. Chem.* **50** (1978) 2064. Links
121. P. Burba, K. H. Lieser, *Fresenius Z. Anal. Chem.* **297** (1979) 374.
122. H. J. Fischer, K. H. Lieser, *Fresenius Z. Anal. Chem.* **335** (1989) 738.
123. A. Disam, P. Tschöpel, G. Tölg, *Fresenius Z. Anal. Chem.* **295** (1979) 97.
124. R. Chakrovorty, R. van Grieken, *Int. J. Environ. Anal. Chem.* **11** (1982) 67.
125. B. Andresen, B. Salbu, *Radiochem. Radioanal. Lett.* **52** (1982) 19.
126. C. L. Smith, J. M. Motooka, W. R. Willson, *Anal. Lett.* **17** (1984) 1715. Links
127. J. Bandovskis, M. Vircavs, O. Veveris, A. Pelne, *Talanta* **34** (1987) 179.
128. J. M. Labrecque in H. M. Kingston, L. B. Jassie (eds.): *Introduction to Microwave Sample Preparation*, Am. Chem. Soc., Washington D.C. 1988, p. 203.
129. L. Solujic, E. B. Milosavljevic, M. R. Straka, *Analyst* **124** (1999) 1255. Links
130. R. Sur, J. Begerow, L. Dunemann, *Fresenius' J. Anal. Chem.* **363** (1999) 526.
131. K. Sasaki, G. E. Pacey, *Talanta* **50** (1999) 175. Links
132. C.-C. Huang, M.-H. Yang, T.-S. Shih, *Anal. Chem.* **69** (1997) 3930. Links
133. U. Pichler, A. Haase, G. Knapp, *Anal. Chem.* **71** (1999) 4050. Links
134. B. W. Wenclawiak, T. Hees, C. E. Zöller, H.-P. Kabus, *Fresenius' J. Anal. Chem.* **358** (1997) 471.
135. M. Burguera, J. L. Burguera, O. M. Alarcon, *Anal. Chim. Acta* **179** (1986) 351. Links
136. V. Karanassios, F. H. Li, B. Liu, E. D. Salin, *J. Anal. At. Spectrom.* **6** (1991) 457.
137. B. Welz, M. Schubert-Jacobs in B. Welz (ed.): *5th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1989, p. 327.
138. B. Welz, M. Sperling, X. Yin in B. Welz (ed.): *6th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1992, p. 203.
139. G. Schulze, H. Tessmer, O. Elsholz in B. Welz (ed.): *5th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1989, p. 347.
140. A. Meyer, L. Dunemann in B. Welz (ed.): *6th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1991, p. 115.
141. J. A. Broekaert, *Anal. Proc. (London)* **27** (1990) 336.
142. I. Atsuya, K. Itoh, K. Akatsuka, *Fresenius Z. Anal. Chem.* **328** (1987) 338.
143. P. Fecher, C. Malcherek in B. Welz (ed.): *5th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1989, p. 501.
144. K. H. Tobies, W. Großmann in B. Welz (ed.): *5th Colloquium Atomspektrometrische Spurenanalytik*, Bodenseewerk Perkin-Elmer, Überlingen 1989, p. 513.

145. A. Sugimae, R. M. Barnes, *Anal. Chem.* **58** (1986) 785. Links
146. L. Dunemann, H. Hajimiragha, J. Begerow, *Fresenius' J. Anal. Chem.* **363** (1999) 466.
147. L. Moens, T. de Smaele, R. Dams, P. Van Den Broeck, P. Sandra, *Anal. Chem.* **69** (1997) 1604. Links
148. Y. Jin, F. Liang, H. Zhang, L. Zhao, Y. Huan, D. Song, *Trends Anal. Chem.* **18** (1999) 479.
149. F. Brugnone *et al.*, *Int. Arch. Occup. Environ. Health* **64** (1992) 179. Links
150. J. Rudling, E. Bjoerkholm, B.-O. Lundmark, *Ann. Occup. Hyg.* **30** (1986) 319. Links
151. P. J. H. D. Verkoelen, M. W. F. Nielen, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **11** (1988) 291.
152. M. D. Erickson: *Analytical Chemistry of PCBs*, Ann Arbor Science Publishers, Stoneham, Mass. 1986.
153. B. Bush, J. T. Snow, S. Connor, *J. Assoc. Off. Anal. Chem.* **66** (1983) 258.
154. H.-D. Isengard, U. Bröcker: *Proceedings of the 5th International Conference on Electromagnetic Wave Interaction with Water and Moist Substances*, 2004.
155. ASTM Method E1358-90, *The American Society for Testing and Materials*, Philadelphia 1990.
156. ASTM Method D4643-93, *The American Society for Testing and Materials*, Philadelphia 1993.
157. R. C. Entz, H. C. Hollifield, *J. Agric. Food Chem.* **30** (1982) 84. Links
158. W. Wittfoht, W. J. Scott, H. Nau, *J. Chromatogr.* **448** (1988) 433. Links
159. H. Hajimiragha, U. Ewers, R. Jansen-Rosseck, A. Brockhaus, *Int. Arch. Occup. Environ. Health* **58** (1986) 141. Links
160. J. Angerer, G. Scherer, K. H. Schaller, J. Müller, *Fresenius J. Anal. Chem.* **339** (1991) 740. Links
161. D. G. Patterson *et al.*, *Anal. Chem.* **58** (1986) 705. Links
162. G. A. Junk, M. J. Avery, J. J. Richard, *Anal. Chem.* **60** (1988) 1347. Links
163. A. Bergmann, L. Reutegardh, M. Ahlman, *J. Chromatogr.* **291** (1984) 392.
164. B. Wigilius *et al.*, *J. Chromatogr.* **391** (1987) 169. Links
165. J. Begerow, R. Heinrich-Ramm, J. Angerer, *Fresenius Z. Anal. Chem.* **331** (1988) 818.
166. H. C. Shields, C. J. Weschler, *J. Air Pollut. Control Assoc.* **37** (1987) 1039.
167. J. Begerow, E. Jermann, T. Keles, L. Dunemann, *Fresenius J. Anal. Chem.* (1994), in press.
168. R. Heinrich, J. Angerer, *Fresenius Z. Anal. Chem.* **322** (1985) 766.
169. G. Birner, H. G. Neumann, *Arch. Toxicol.* **62** (1988) 110. Links
170. K. M. Engström, *Scand. J. Work Environ. Health* **10** (1984) 75. Links
171. M. Balikova, J. Kohlicek, *J. Chromatogr.* **497** (1989) 159. Links
172. E. R. Adlard, C. B. Milne, P. E. Tindle, *Chromatographia* **14** (1981) 507. Links
173. J. A. F. de Silva, *J. Chromatogr.* **273** (1983) 345.
174. K. G. Wahlund, T. Arvidson, *J. Chromatogr.* **282** (1983) 527. Links
175. A. M. Rustum, *J. Chromatogr. Sci.* **27** (1989) 18. Links
176. S. G. Colgrove, J. H. Svec, *Anal. Chem.* **53** (1981) 1737. Links
177. J. Lewalter, U. Korallus, *Int. Arch. Occup. Environ. Health* **56** (1985) 179. Links
178. M. Akerblom, G. Alex, *J. Assoc. Off. Anal. Chem.* **67** (1984) 653. Links
179. J. Czuczwa *et al.*, *J. Chromatogr.* **403** (1987) 233. Links
180. T. L. Peters, *Anal. Chem.* **54** (1982) 1913. Links
181. W. G. Jennings, A. Rapp: *Sample Preparation for Gas Chromatographic Analysis*, Hüthig Verlag, Heidelberg 1983.
182. D. A. J. Murray, *J. Chromatogr.* **177** (1979) 135. Links
183. J. F. J. van Rensberg, A. J. Hasset, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **5** (1982) 574.
184. N. F. Wood, *Analyst (London)* **94** (1969) 399.
185. M. Teufel *et al.*, *Arch. Environ. Contam. Toxicol.* **19** (1990) 646. Links
186. S. Hashimoto, H. Ito, M. Morita, *Chemosphere* **25** (1992) 297. Links
187. V. Lopez-Avila, R. Young, J. Benedicto, P. Ho, *R. Kim, Anal. Chem.* **67** (1995) 2096.
188. V. Lopez-Avila, J. Benedicto, C. Charan, R. Young, *Environ. Sci. Technol.* **29** (1995) 2709. Links
189. B. K. Kramer, P. B. Ryan: *Proceedings of the 2000 Conference on Hazardous Waste Research*, http://www.engg.ksu.edu/hsrc/00proceed/kramer.pdf
190. M. Weichbrodt, W. Vetter, B. Luckas, *Journal of AOAC International* **83** (2000) 1334. Links
191. M. Spiro, S. S. Chen, *Journal of Microwave Power and Electromagnetic Energy* **29** (1994) 231.
192. J. L. Luque-García, S. Morales-Muñoz, M. D. Luque de Catro, *Chromatographia* **55** (2002) 117.
193. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids*, 4th ed., McGraw-Hill, New York 1987.
194. J. W. King, *J. Chromatogr. Sci.* **27** (1989) 355. Links
195. K. S. Nam *et al.*: *Proceedings of the International Symposium on Supercritical Fluids*, French Chemical Society: Paris Index, France 1988, p. 743.
196. J. W. King, J. H. Johnson, J. P. Friedrich, *J. Agric. Food Chem.* **37** (1989) 951. Links
197. M. A. Schneiderman, A. K. Sharma, K. R. R. Mahanama, D. C. Locke, *J. Assoc. Off. Anal. Chem.* **71** (1988) 815. Links
198. M. A. Schneiderman, A. K. Sharma, D. C. Locke, *J. Chromatogr. Sci.* **26** (1988) 458. Links
199. K. Sugiyama, M. Saito, *J. Chromatogr.* **442** (1988) 121. Links

200. P. Capriel, A. Haisch, S. U. Khan, *J. Agric. Food Chem.* **34** (1986) 70. Links
201. V. Janda, G. Steenbeke, P. Sandra, *J. Chromatogr.* **479** (1989) 200. Links
202. S. B. Hawthorne, D. J. Miller, *J. Chromatogr. Sci.* **24** (1986) 258. Links
203. B. W. Wright, C. W. Wright, J. S. Fruchter, *Energy Fuels* **3** (1989) 474. Links
204. S. B. Hawthorne, D. J. Miller, *Anal. Chem.* **59** (1987) 1705. Links
205. M. M. Schantz, S. N. Chesler, *J. Chromatogr.* **363** (1986) 397.
206. B. O. Brady, C. P. C. Kao, K. M. Dooley, F. C. Knopf, *Ind. Eng. Chem. Prod. Res. Dev.* **26** (1987) 261.
207. N. Alexandrou, J. Pawliszyn, *Anal. Chem.* **61** (1989) 2770. Links
208. F. I. Onuska, K. A. Terry, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **12** (1989) 357.
209. B. W. Wright, C. W. Wright, R. W. Gale, R. D. Smith, *Anal. Chem.* **59** (1987) 38. Links
210. S. B. Hawthorne, M. S. Krieger, D. J. Miller, *Anal. Chem.* **61** (1989) 736. Links
211. S. Forbes, *Anal. Chim. Acta* **196** (1987) 75. Links
212. U. Juergens, *J. Chromatogr.* **371** (1986) 307. Links
213. J. J. Richard, G. A. Junk, *Mikrochim. Acta* 1986 no. 1, 387. Links
214. G. A. Junk, J. J. Richard, *Anal. Chem.* **60** (1988) 451. Links
215. R. Bagnati, E. Benfenati, E. Davoli, R. Fanelli, *Chemosphere* **17** (1988) 59. Links
216. J. Manes Vinuesa, J. C. Molto Cortes, C. Igualada Canas, G. Font Perez, *J. Chromatogr.* **472** (1989) 365.
217. J. Angerer *et al.*, *Fresenius J. Anal. Chem.* **342** (1992) 433. Links
218. K. I. Tomlins, K. Jewers, R. D. Coker, *Chromatographia* **27** (1989) 67. Links
219. M. R. Khan, C. P. Ong, S. F. Y. Li, H. D. Lee, *J. Chromatogr.* **513** (1990) 360. Links
220. K. E. Savolainen, *J. Pharm. Sci.* **77** (1988) 802. Links
221. R. Schmidt, P. Kupferschmidt, *Clin. Chem. (Winston-Salem N.C.)* **35** (1989) 1352.
222. V. Marko, L. Soltes, K. Radova, *J. Chromatogr. Sci.* **28** (1990) 403. Links
223. V. W. Burse *et al.*, *J. Anal. Toxicol.* **14** (1990) 143. Links
224. W. Butte, C. Fooken, *Fresenius J. Anal. Chem.* **336** (1990) 511. Links
225. H. Bouwman, R. M. Coopan, A. J. Reinecke, *Chemosphere* **19** (1989) 1563. Links
226. P. H. Cramer *et al.*, *Chemosphere* **20** (1990) 821. Links
227. A. K. D. Liem *et al.*, *Chemosphere* **20** (1990) 843. Links
228. P. Fuerst, H. A. Meemken, W. Groebel, *Chemosphere* **15** (1986) 1977. Links
229. H. Beck *et al.*, *Chemosphere* **16** (1987) 1977. Links
230. R. G. Heath *et al.*, *Anal. Chem.* **58** (1986) 463. Links
231. H. F. Walton, R. D. Rocklin: *Ion Exchange in Analytical Chemistry*, CRC Press, Boca Raton Fla. 1990.
232. T. G. Kreindl, H. Malissa, K. Winsauer, *Mikro-chim. Acta* 1986 no. 1, 1.
233. W. E. Turner, S. G. Isaacs, D. G. Patterson Jr., *Chemosphere* **25** (1992) 805. Links
234. D. Barceló, G. Durand, V. Bouvot, M. Nielen, *Environ. Sci. Technol.* **27** (1993) 271.
235. E. R. Brouwer, H. Lingemann, U. A. T. Brinkmann, *Chromatographia* **29** (1990) 415. Links
236. J. Rosenfeld, M. Mureika-Russell, S. Yeroushalmi, *J. Chromatogr.* **358** (1986) 137. Links
237. J. M. Rosenfeld, M. Mureika-Russell, A. Phatak, *J. Chromatogr.* **283** (1984) 127. Links
238. Z. Bardodej, *Arbeitsmed. Sozialmed. Arbeitshyg.* **3** (1968) 141.
239. J. Pawliszyn: *Solid Phase Microextraction: Theory and Practice*, Wiley-VCH, New York 1997.
240. C. Grote, J. Pawliszyn, *Anal. Chem.* **69** (1997) 587. Links
241. I. Valor, C. Cortada, J. C. Moltó, *J. High Resolut. Chromatogr.* **19** (1996) 472.
242. T. Nilsson, F. Pelusion, L. Montanarella, B. Larsen, S. Facchetti, J. Madsen, *J. High Resolut. Chromatogr.* **18** (1995) 617.
243. M. Möder, S. Schrader, U. Franck, P. Popp, *Fresenius' J. Anal. Chem.* **357** (1997) 326.
244. M. T. Sng, F. K. Lee, H. A. Lakso, *J. Chromatogr. A* **759** (1997) 225. Links
245. D. W. Potter, J. Pawliszyn, *Environ. Sci. Technol.* **28** (1997) 298.
246. P. A. Martos, J. Pawliszyn, *Anal. Chem.* **69** (1997) 206. Links
247. F. Mangani, R. Cenciarini, *Chromatographia* **41** (1995) 678. Links
248. B. Tienpont, *et al.*, *J. Pharm. Biomed. Anal.* **32** (2003) 569. Links
249. D. Benanou, F. Acobas, M. R. de Roubin, F. David, P. Sandra, *Anal. Bioanal. Chem.* **376** (2003) 376, 69–77.
250. V. M. León, *J. Chromatogr. A.* **999** (2003) 91. Links
251. J C. R. Demyttenaere *et al.*, *J. Chromatogr. A.* **985** (2003) 221. Links
252. Y. Hayasaka, *et al.*, *Anal. Bioanal. Chem.* **375** (2003) 948.
253. F. E. Ahmed, *Trends Anal. Chem.* **22** (2003) 170.
254. E. Baltussen, C. A. Cramers, P. J. Sandra, *Anal. Bioanal. Chem.* **373** (2002) 3.
255. J. Angerer: “Ameisensäure” in D. Henschler (ed.): *Analysen in biologischem Material,* 4. ed., Verlag Chemie, Weinheim 1980.
256. A. B. Heath, R. R. Black, *J. Assoc. Off. Anal. Chem.* **70** (1987) 862. Links
257. R. L. Brown, C. L. Farmer, R. G. Millar, *J. Assoc. Off. Anal. Chem.* **70** (1987) 442. Links

258. R. A. Minear, L. H. Keith (eds.): *Water Analysis*, vol. **3**, Academic Press, Orlando 1984, p. 84.
259. T. N. Eisenberg, E. J. Middlebrooks: *Reverse Osmosis Treatment of Drinking Water*, Butterworths, London 1986.
260. T. D. Brook: *Membrane Filtration: A User's Guide and Reference Manual*, Science Tech., Madison 1983.
261. M. M. L. Aerts, W. M. J. Beek, U. A. T. Brinkmann, *J. Chromatogr.* **435** (1988) 613.
262. D. C. Turnell, J. D. H. Cooper, *J. Chromatogr.* **395** (1987) 613. Links
263. A. C. Metha, *Tr. AC Trends Anal. Chem. (Pers. Ed.)* **8** (1989) 107.
264. K. Pekari, M. L. Riekkola, A. Aitio, *J. Chromatogr.* **491** (1989) 309. Links
265. G. Machata, J. Angerer: "Gaschromatographie, Headspace-Technik" in D. Henschler (ed.): *Analysen in biologischem Material,* 7th suppl., Verlag Chemie, Weinheim 1983.
266. V. C. Stein, R. S. Narang, *Arch. Environ. Contam. Toxicol.* **19** (1990) 593. Links
267. S. L. Friant, I. H. Suffet, *Anal. Chem.* **51** (1979) 2167. Links
268. L. Dunemann, H. Hajimiragha, *Anal. Chim. Acta* **283** (1993) 199. Links
269. D. L. Ashley *et al.*, *Anal. Chem.* **64** (1992) 1021. Links
270. H. Hajimiragha, U. Ewers, A. Brockhaus, A. Boettger, *Int. Arch. Occup. Environ. Health* **61** (1989) 513. Links
271. K. H. Pannwitz, *Drägerheft* **332** (1985) 10.
272. K. H. Pannwitz, *Drägerheft* **327** (1983) 6.
273. G. A. Reineccius, R. Liardon in R. G. Berger, S. Nitz, P. Schreier (eds.): " Topics in Flavor Research," Proceedings of *International Conference Freising-Weihenstephan,* 1985, p. 125.
274. K. H. Schaller: "Mandelsäure" in: D. Henschler (ed.): *Analysen in biologischem Material,* 2nd suppl., Verlag Chemie, Weinheim 1978.
275. J. Begerow, J. Angerer, *Fresenius J. Anal. Chem.* **366** (1990) 42.
276. M. I. Daneshvar, J. B. Brooks, *J. Chromatogr.* **433** (1988) 248. Links
277. H. B. Lee, T. E. Peart, J. M. Carron, *J. Chromatogr.* **498** (1990) 367. Links
278. S. Jacobbson, A. Larsson, A. Arbin, A. Hagman, *J. Chromatogr.* **358** (1986) 137. Links
279. R. v. Smith, S. Tsai, *J. Chromatogr.* **61** (1971) 29. Links
280. P. Pfäffli, H. Savolainen, H. Keskinen, *Chromatographia* **27** (1989) 483.
281. G. Johanson, H. Kronborg, P. Näslund, M. Byfalt Nordquist, *Scand. J. Work Environ. Health* **12** (1986) 594. Links
282. Y. L. Emeillat, J. Menez, F. Berthou, L. Bardou, *J. Chromatogr.* **206** (1981) 89.
283. R. Patience, J. Thomas, *J. Chromatogr.* **234** (1982) 225. Links
284. E. Grushka, H. Durst, E. Kikta, *J. Chromatogr.* **112** (1975) 673. Links
285. T. Jupile, *J. Chromatogr. Sci.* **17** (1979) 160.
286. W. Steiner, E. Müller, D. Fröhlich, R. Battaglia, *Mitt. Geb. Lebensmitteluntersuch. Hyg.* **75** (1984) 37.
287. M. Petrarulo, S. Pellegrino, M. Marangella, F. Linari, *J. Chromatogr.* **432** (1988) 37. Links
288. M. Petrarulo *et al.*, *J. Chromatogr.* **465** (1989) 87. Links
289. W. Dünges, *Anal. Chem.* **49** (1977) 442. Links
290. W. Dünges, *Chromatographia* **9** (1976) 624.
291. E. Grushka, S. Lam, J. Chassin, *Anal. Chem.* **50** (1978) 1398. Links
292. A. Crozier, J. Zaerr, R. Morris, *J. Chromatogr.* **238** (1982) 157. Links
293. H. Tsuchiya, T. Hayastsi, H. Naruse, N. Takagi, *J. Chromatogr.* **234** (1982) . Links
294. H. D. Winkeler, K. Levsen, *Fresenius Z. Anal. Chem.* **333** (1989) 716.
295. M. A. Carey, H. E. Perisinger, *J. Chromatogr. Sci.* **10** (1972) 537. Links
296. J. K. Lin, J. Y. Yang, *Anal. Chem.* **47** (1975) 1634. Links
297. J. Y. Chang, H. L. Creaser, *J. Chromatogr.* **116** (1976) 215. Links
298. J. K. Lin, C. C. Lai, *Anal. Chem.* **52** (1980) 630. Links
299. K. Karlson, M. Alsandro, M. Novotny, *Anal. Chem.* **57** (1985) 229. Links
300. K. Grob Jr., D. Fröhlich, B. Schilling, *J. Chromatogr.* **295** (1984) 55. Links
301. K. Grob Jr., C. Walder, B. Schilling, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **9** (1986) 95.
302. K. Grob Jr., B. Schilling, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **8** (1985) 726.
303. K. Grob Jr., G. Karrer, M. L. Riekkola, *J. Chromatogr.* **334** (1985) 129. Links
304. H. J. Cortes, E. L. Olberding, J. H. Wetters, *Anal. Chim. Acta* **236** (1990) 173. Links
305. R. Barcaralo, *J. High Resolut. Chromatogr. Chromatogr. Comm.* **518** (1990) 465.
306. B. A. Anderson, *Am. Environ. Lab.* **1** (1989) 41.
307. D. Barcelo *et al.*, *Anal. Chem.* **62** (1990) 1696. Links
308. S. A. Soper, K. L. Ratzlaff, T. Kuwana, *Anal. Chem.* **62** (1990) 1438. Links
309. J. Strojek, S. A. Soper, K. L. Ratzlaff, T. Kuwana, *Anal. Sci.* **6** (1990) 121. Links
310. S. B. Hawthorne, D. J. Miller, J. J. Langefeld, *J. Chromatogr. Sci.* **28** (1990) 2. Links
311. J. R. Wheeler, M. E. McNally, *J. Chromatogr. Sci.* **27** (1989) 534. Links
312. F. I. Onuska, K. A. Tery, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **12** (1989) 527.
313. M. R. Andersen, J. T. Swanson, N. C. Porter, B. E. Richter, *J. Chromatogr. Sci.* **27** (1989) 371. Links
314. C. H. Marvin *et al.*, *J. Chromatogr.* **518** (1990) 242. Links

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Trace Analysis

HELMUT MÜLLER, Universität Halle-Wittenberg, Fachbereich Chemie, Merseburg, Federal Republic of Germany

HEINZ W. ZWANZIGER, Fachhochschule Merseburg, Fachbereich Chemie- und Umweltingenieurwesen, Merseburg, Federal Republic of Germany

JOHANNES FLACHOWSKY, Umweltforschungszentrum, Leipzig-Halle GmbH, Leipzig, Federal Republic of Germany

1. Subject and Scope 2
2. Fields of Work 2
3. Methods of Modern Trace Analysis 3
4. Calibration and Validation 4
4.1. Conceptual Problems 4
4.2. Errors 6
4.3. The Critical Signal Value and Limits of the Procedure 6
4.4. Adequate Calibration Models 8
4.5. Quality Assurance and Standard Reference Materials 11
5. Environmental Analysis 12
5.1. The Problem 12
5.2. Possibilities of Mobile Analysis . . . 13
5.2.1. Definition and Requirements 13
5.2.2. Example of Equipment 14
5.3. Selected Applications of Mobile Analysis 16
5.3.1. GC-MS Screening and IMS monitoring 16
5.3.2. Elemental Analysis by Mobile EDXRF 18
5.4. Conclusions 18
6. References 18

Symbols

1) Symbols of features (quantities, variables):

X	analyte feature (e.g., concentration)
Y	signal, response

2) Symbols of measured or calculated values of features:

x	concentration value
x_{A}	analytical value; analyte concentration
x_{C}	calibration value
x_{decision}	limit of decision
$x_{\mathrm{detection}}$	limit of detection
$x_{\mathrm{quantitation}}$	limit of quantitation
y	signal or response value
y_{blank}	blank value
y_{crit}	critical value

3) Symbols for running indices:

b	with respect to blank values
c	with respect to calibration
n	general running index

4) Symbols for terminal values of indices:

N	number of . . .
N_a	. . . analytical values
N_b	. . . blank values
N_c	. . . calibration values

5) Symbols of statistical features and parameters:
symbols with a bar: mean values
symbols with a hat: calculated values according to a model

a_0	intercept in regression models
a_1	slope; sensitivity; regression coefficient
k	statistical factor (quantile of a distribution to be specified)
q	order of quantile
res	residual value (difference of measured and calculated values)
s	calculated ("sample") standard deviation
s^2	calculated ("sample") variance
sgn (cor_{XY})	sign of the correlation coefficient of X and Y
SS	read as "sum of squares" . . .
	SS_{XX} . . . of centered values of feature X
	SS_{YY} . . . of centered values of feature Y
SS_{XY}	read as "sum of products of centered values of features X and Y
$t, t_{\mathrm{q}}(\nu)$	values of Student's distribution (quantiles)
α	error probability (risk of first kind)
β	error probability (risk of second kind)
ν	degrees of freedom
σ	"population" standard deviation

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_095

1. Subject and Scope

The determination of analytes in very low concentration (traces) plays a fundamental role in many areas of science and technology.

A "trace" means a very small level of a substance that is present alongside a large excess of other components, which causes certain effects and must be detected (with qualitative information) and determined (with quantitative information).

A distinction is made between organic and inorganic "traces." Examples of *organic traces* are dioxins or furans in waste gases from refuse incineration plants, fluorochlorohydrocarbons (FCHCs) in the atmosphere, chlorinated hydrocarbons (CHCs) in water, and many others. *Inorganic traces* are referred to as trace elements, a concept that includes all elements (i.e., metals, semimetals, and nonmetals). Despite the different objects that must be monitored analytically, the methods and instruments in this field have so much in common that a special branch of analysis – trace analysis – has developed [1, 2].

For trace analysis, analytical methods with extreme powers of detection are required, which must often be coupled with enrichment steps. Another requirement follows from the fact that the traces occur alongside a large excess of other substances (matrix). The main components are present in $10^2 - 10^{10}$-fold excess. Therefore, methods must be developed that permit trace determinations without interference by the main component (i.e., specific methods); otherwise, separation of the trace and main components must be carried out before the determination.

The possibility of determining traces of elements arose in the 1920s with the development of spectrophotometry, spectroscopic methods, and polarography. Trace analysis was carried out very intensively and with great sophistication in connection with the development and production of the first atomic bomb. Later, it received an important stimulus from the development of novel materials, especially semiconductors.

More and more areas than ever depend on the results of trace analysis: Nuclear energy, production of semiconductors and ultrapure substances, metallurgy, materials research and production, geology, mineralogy, oceanography, medicine, animal and plant physiology, criminology, ecology, and environmental research.

Ecological research and routine analysis relevant to the environment provide trace analysis with imporant stimuli. At the same time, the centers of attention are shifting more and more from element trace analysis to organic trace analysis.

2. Fields of Work

Modern technology has many possibilities at its disposal for solving problems in trace analysis. A uniform characterization and assessment of trace-analytical methods requires the clearest possible definitions: For reporting a content G, mass relations – of which the simplest is percentage by mass – are unequivocal and independent of any additional information. The contents of the components to be determined can vary within wide limits. Very diverse proposals for the boundaries of these ranges are found. For practical purposes, the following classification has been established:

Main fraction	100 % – 10 %
Minor fraction	10 % – 0.1 %
Traces	less than 0.1 %
Micro fraction range	1000 – 1 µg/g
Nano fraction range	1000 – 1 ng/g

For nanofractions (contents below ng/g) the term "extreme trace analysis" has become generally accepted.

In addition to the specification of content ranges in parts by mass w, parts by volume σ, or parts by amount of substance x, in trace analysis the term ppm (parts per million; 1 ppm = $1 : 10^6 = 10^{-4}$ %) is generally used for parts by mass w. Still lower concentrations are expressed in ppb (parts per billion); errors can arise because the intended reference quantity 10^9 (1 ppb = 10^{-7} %) is called "billion" mainly in the United States, but "milliard" in Europe. This quantity is sometimes represented as ppM (parts per milliard).

In the literature, ppm by mass is still encountered occasionally, and the ppm term is then used without further identification. Unfortunately, the terms ppm and ppb, which preferably should be used only for problems in solid analysis, are also used, instead of mass concentration β (or C) in milligrams or micrograms per liter, for specifying the concentration of trace contents in dissolved samples.

The content of trace components in gases is indicated as volume concentration σ or mass concentration β. Useful units are milliliters per cubic meter (ppm) and milligrams per cubic meter. The conversion is carried out according to

$$\sigma\left[\mathrm{mL/m^3}\right] = \frac{\text{molar volume [L/mol]}}{\text{molar mass [g/mol]}} \cdot \beta\left[\mathrm{mg/m^3}\right]$$

The molar volume is usually related to 25 °C and 101.3 kPa and is 22.47 L/mol. For dusts, the mass concentration β (mg/m^3) or the particle concentration C (particles/m^3 or fibers/m^3) is used.

Table 1 surveys the specification of content data in trace analysis. Limiting factors for the optimum choice of an analytical method for a trace-analytical problem are the portion of sample available and the expected mass concentration β of the analyte.

Table 1. Specification of content units in trace analysis

Unit	Designation	Percentage
Parts by mass w		
mg/kg	1 ppm	10^{-4}
µg/kg	1 ppb	10^{-7}
ng/kg	1 ppt ∗	10^{-10}
pg/kg	1 ppq ∗∗	10^{-13}
Part by volume σ		
mL/m^3	ppm	10^{-4}
µL/m^3	ppb	10^{-7}
nL/m^3	ppt ∗	10^{-10}
pL/m^3	ppq ∗∗	10^{-13}

∗ ppt = Parts per trillion.
∗∗ ppq = Parts per quadrillion.

The sample portions or sample volumes most frequently used are in the range of a minimum of grams or milliliters to micrograms or microliters. Modern methods (e.g., capillary zone electrophoresis, CZE) require sample volumes only in the nanoliter range. This results in a high *mass sensitivity* that leads into the range of detection of individual molecules or atoms. By coupling capillary zone electrophoresis with laser-induced fluorescence detection, an absolute total of 600 molecules has even been detected.

Clearly distinguished from this is the "*concentration sensitivity*" of the individual methods. The current state of the art is ppm analysis of a wide range of analytes in a variety of matrices. The trend to ppb and in a few cases ppt (10^{-10} %) (parts per trillion, 10^{-12}) without enrichment is expected.

3. Methods of Modern Trace Analysis

Modern trace analysis is characterized by two different strategies, whose orientation is more physical or chemical, respectively:

1) Application of instrumental direct methods
2) Development of chemical-analytical compound methods

Instrumental Direct Methods. In instrumental direct methods, after sampling and possibly sample preparation, the sample is analyzed directly. Instrumental direct determination methods are usually matrix-dependent relative methods. A mathematical correction of the matrix effect is possible only in particular cases (e.g., X-ray fluorescence analysis, XRF). To compensate for systematic errors, therefore, standard reference materials are required, which must be very similar in composition to the sample to be analyzed (see Section 4.5). However, an optimum power of detection at high accuracy (freedom from systematic error) can be achieved only when the trace to be determined is present in isolated form in the highest possible mass concentration. In many cases of trace analysis, therefore, more complex multistage (compound) methods are required.

Compound Methods. In compound methods, after dissolution or digestion of the sample, the analyte is separated from the matrix, usually enriched, and then determined (Fig. 1). The calibration is relatively simple because the analyte components have been isolated. The only standards required are analyte solutions of known concentration, since the matrix can no longer interfere with the determination. The advantage of easy calibration is lessened, however, by the fairly large number of operations and the correspondingly greater risk of systematic errors (Fig. 1). This requires very careful and complicated optimization of compound methods to reduce such errors, which today are a central problem of extreme trace analysis. Compound methods are therefore often carried out in closed systems (digestion, enrichment, and/or separation in one apparatus) to prevent errors in the

determination of trace contents resulting from the sample environment and sample preparation steps. The necessary trace-analytical hygiene is as important for samples as for standards, solvents, reagents, vessels, instruments, and the laboratory atmosphere. If the highest purity is required, particle emission by laboratory personnel must also be considered. Work is therefore carried out in clean rooms or clean boxes, through which laminar flow of filtered air occurs.

New operating techniques (e.g., flow methods), in combination with modern instrumental methods, enable the separation and enrichment processes to be automated in quasi-closed systems, and have contributed substantially to the increasing establishment of such composite methods in trace-analytical practice [3].

Despite the recognizable advantages of compound methods, work is increasingly being carried out on the (further) development of multielement methods with high powers of detection, especially for micro samples. Table 2 shows a survey of modern instrumental multianalyte methods.

Particularly for determining organic trace contents, analytical methods are used that combine systems with different measurement principles and therefore different selectivity (coupled or hyphenated techniques). Chromatographic methods, particularly gas chromatography, supercritical fluid chromatography (SFC), high-performance liquid chromatography, and capillary electrophoresis (CE) are often coupled with spectroscopic methods [7, 8]. The most important coupled systems in trace and environmental analysis are the following:

GC – MS, FTIR, AES/AAS, FTIR – MS, ICP – MS,
SFC – MS, FTIR, AES/AAS, FTIR – MS, ICP – MS,
HPLC – MS, FTIR, AES/AAS,
CE – MS, UV

Coupling of Chromatographic Methods
HPLC – GC, HPLC – GC – MS, SFC – GC – MS,
HPLC – TLC – FTIR

The advantages of such coupled techniques lie in a frequently effective separation of the matrix (improvement of the accuracy), an increased power of detection, and a contribution to speciation analysis [9].

4. Calibration and Validation (see also → Chemometrics, Chap. 12.)

4.1. Conceptual Problems

A fully defined trace-analytical procedure is characterized by the concentration range, as well as the calibration function, the precision and accuracy, and the limits of decision, detection, and possibly quantitation. The use of the three last-mentioned terms in the literature unfortunately is not uniform and requires international standardization.

The *limit of decision* is an important criterion in the optimization of a trace-analytical procedure and the comparison of methods. It depends greatly on possible errors in individual steps. It can be determined in many different ways, which may vary in their complexity [10, 11] and whose general applicability on national level is still being elaborated [11, 12]. The precise specification of the means of calculating the limit of decision should therefore be a part of the analytical instructions. Determination of the limit of decision and determination of an analyte are based, except in the case of absolute methods (gravimetry, coulometry), not only on the actual course of the analysis but also essentially on the results of the calibration process.

4.2. Errors

Like any analytical procedure, trace analysis is subject to sources of error that can lead to systematic and random falsification of the observed values or test results. The reliability of results is therefore determined by the accuracy and precision. Measures of the precision are repeatability and reproducibility.

Observed values or results of measurement means a set $\{y_1, y_2, \ldots, y_N\}$ of N recorded values y_n (with $n = 1, 2, \ldots, N$) of a signal quantity Y (e.g., absorbance) that result from N given values x_n ($n = 1, 2, \ldots, N$) of a concentration quantity X or from which an analytical result x_A can be derived (calculated).

Systematic errors lead to one-sided displacement of results. Two cases exist. Case A: If the trace component to be determined is absent from a sample ($x_A = 0$), its presence can be simulated

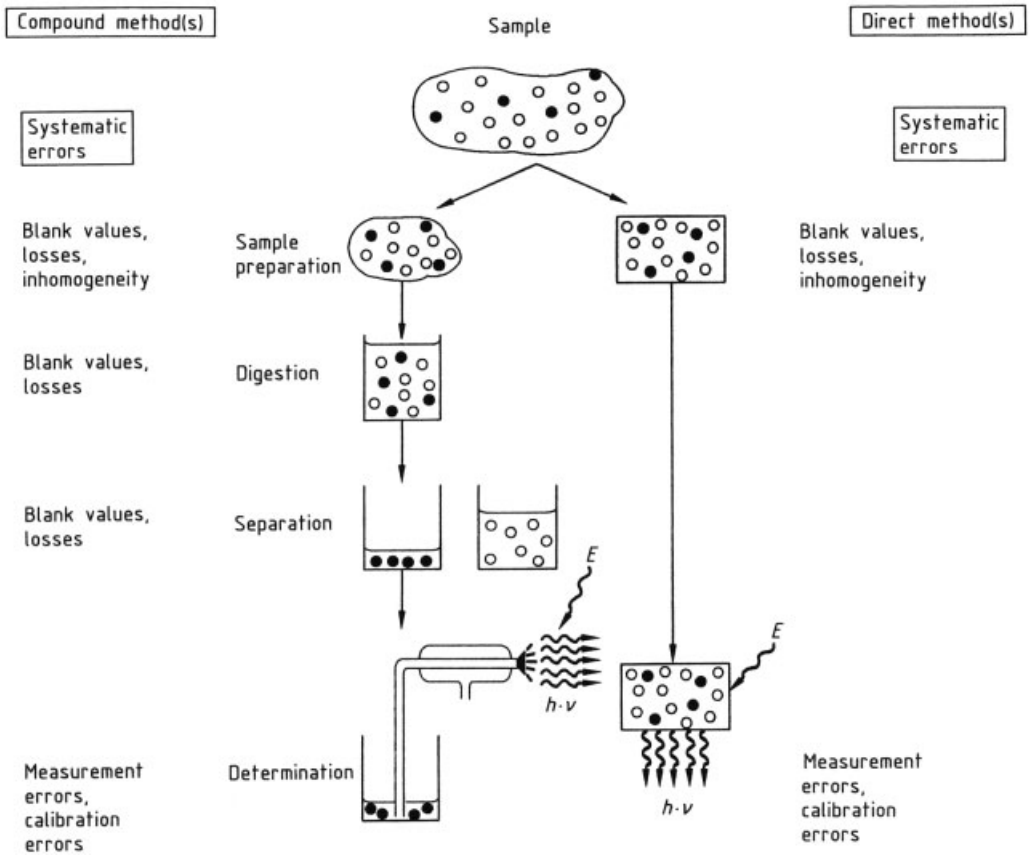

Figure 1. Systematic errors in direct and compound methods

Table 2. Multianalyte methods [4–6]

Method	Concentration range, ppb	Matrix effects [a]	Simultaneous	Sequential
Mass spectrometry (MS)	1 – 10	++	+	
Spark ion source MS (SSMS)				
Secondary ion MS (SIMS)				
Inductively coupled plasma MS (ICP – MS)				
Glow discharge source MS (GD – MS)				
Laser-induced MS (LMS)				
Neutron activation analysis (NAA)	1 – 10	(+)	+	
X-ray fluorescence analysis (XRF)	10 [b] – 1000	++	+	
Total reflection X-ray fluorescence (TRXRF)	10 – 100	++	+	
Atomic emission spectrometry (AES)	10 – 100	(+)		
Inductively coupled plasma AES (AES – ICP)	1 – 10	+++	+	+
Atomic fluorescence spectrometry (AFS)	100 – 1000	++		+
Gas chromatography (GC)	10 – 100 [c]		+	
Capillary GC	1 – 10 [c]		+	
High-performance liquid chromatography (HPLC)	10 – 100 [c]		+	
Capillary zone electrophoresis (CZE)	100 – 1000		+	

[a] + = low . . . +++ = high.
[b] Wavelength-dispersive XRF (WDXRF).
[c] Dependent on detector principle.

(e.g., if the value y_0 of the signal quantity Y specific to the analyte exceeds a critical level y_{crit}). For case B, in the presence of the analyte ($x_{\mathrm{A}} > 0$), two extreme situations are conceivable: The "complete" loss of analyte (which leads to $y_{\mathrm{A}} < y_{\mathrm{crit}}$), or an analyte content that is no longer in the trace range ($y_{\mathrm{A}} \gg y_{\mathrm{crit}}$).

The risk of a wrong decision (in case A to accept the presence of the analyte, or in case B to exclude its presence) can have quite different importance in the two cases. This is not discussed further here, since systematic sources of error can in principle be recognized and eliminated (see Section 4.4).

Unavoidable *random errors* cause a two-sided, random deviation of individual observed values from a particular value (target value, mean, median) and also result in different risks in the assessment of the signal or the reporting of analytical results.

4.3. The Critical Signal Value and Limits of the Procedure

The critical signal value y_{crit} plays a central part in determining the characteristics of the procedure. Via an experimentally supported or calculated calibration relation (e.g., $Y = a_0 + a_1 X$) it leads to the limit of decision $x_{\mathrm{crit}} = x_{\mathrm{decision}}$ (see Fig. 2).

The critical signal value is determined most simply from repeated analysis of several blank analytical portions or from repeated measurement of one blank analytical portion, the latter, however, provides less real characteristic quantities. Also, in the case of an absent analyte ($x = 0$), a total of N_b randomly determined signal values y_b ($b = 1, \ldots, N_b$), are measured, which must not be zero and whose important characteristics can be calculated easily.

Mean $\bar{y}_{\mathrm{blank}} = \frac{1}{N_b} \sum_{b=1}^{N_b} y_b$

Variance $s^2_{\mathrm{blank}} = \frac{1}{N_b - 1} \sum_{b=1}^{N_b} (y_b - \bar{y}_{\mathrm{blank}})^2$

Standard deviation $s_{\mathrm{blank}} = \sqrt{s^2_{\mathrm{blank}}}$

The upper limit of the confidence interval of the mean of the blank value is then regarded as the critical value y_{crit} of the response value.

$$y_{\mathrm{crit}} = \bar{y}_{\mathrm{blank}} + s_{\mathrm{blank}} t_{\mathrm{q}}(\nu) \sqrt{\frac{1}{N_b}}$$

It can be calculated by means of a tabulated value for Student's t distribution with $\nu = N_b - 1$ degrees of freedom and order $q = 1 - \alpha$ (one-sided values) (→ Chemometrics, Chap. 12.).

The notation for the t values is according to [13]. This notation assumes a symmetrical distribution of the observed value signals. However, for trace determinations, this idea, and particularly that of normally distributed values, are frequently unjustified. A logarithmic transformation may be necessary [14, 15].

If a signal value of any test portion falls above this critical value, one can assume, with a certain risk of error α, usually set at 5 % ($\alpha = 0.05$), that the signal value is not a blank value but results from the presence of analyte. At this point the qualitative decision about the presence of an analyte is made with respect to the signal.

The corresponding critical concentration $x_{\mathrm{crit}} = x_{\mathrm{decision}}$, the *limit of decision*, is obtained by combining the calibration relation with the expression for y_{crit}, by assuming that $\bar{y}_{\mathrm{blank}} = a_0$.

$$x_{\mathrm{decision}} = \frac{s_{\mathrm{blank}}}{a_1} t_{\mathrm{q}}(\nu) \sqrt{\frac{1}{N_b}}$$

Optimization of the limit of decision is thus possible by reducing the standard deviation of the blank value, increasing the sensitivity a_1, and increasing the number N_b of blank value measurements. The numerical value of the limit of decision can also be improved by taking future analytical steps into account (i.e., the number N_a of analyses of the analytical portion). If the standard deviation is assumed to be the same for the actual analysis as for the blank value measurements, then

$$x_{\mathrm{decision}} = \frac{s_{\mathrm{blank}}}{a_1} t_{\mathrm{q}}(\nu) \sqrt{\frac{1}{N_b} + \frac{1}{N_a}}$$

A more sophisticated way to obtain the critical signal value is to calculate the upper confidence limit of the intercept a_0 of the calibration relation that was derived from N_c measurements of analytical calibration portions [11]:

$$y_{\mathrm{crit}} = a_0 + s_{\mathrm{res}} t_{\mathrm{q}}(\nu = N_c - 2) \sqrt{\frac{1}{N_c} + \frac{1}{N_a} + \frac{\bar{x}_c^2}{SS_{XX}}}$$

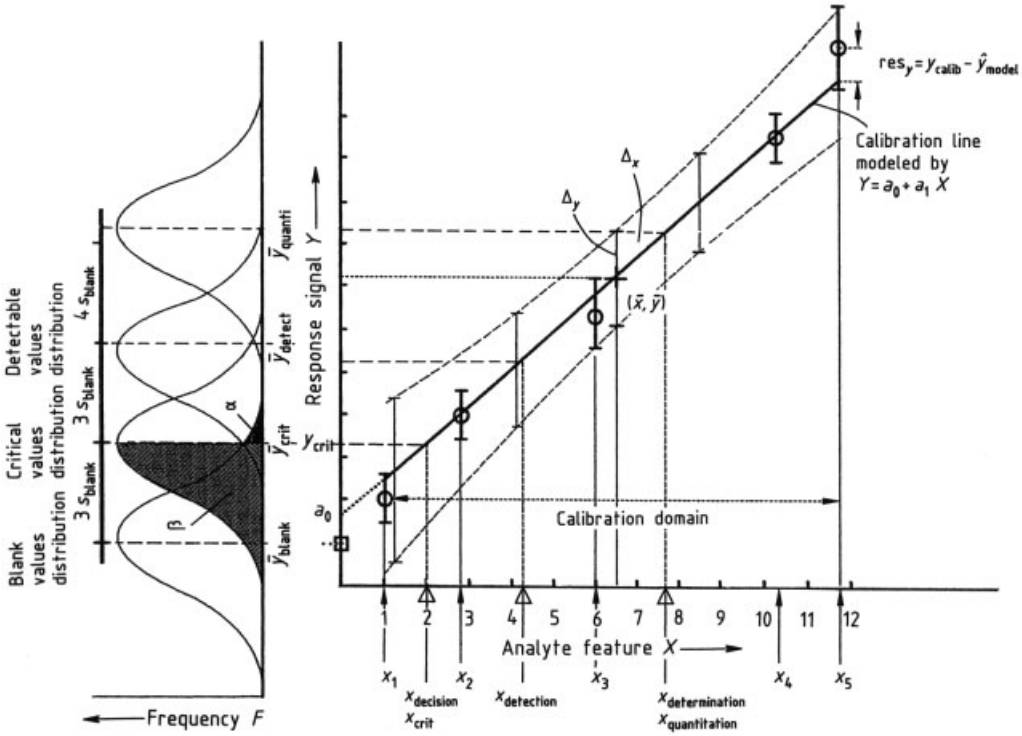

Figure 2. Illustration of problems in simultaneous consideration of the signal and calibration domains

(The residual standard deviation s_{res} of the calibration model and the sum of the squares of deviations SS_{XX} are defined in Section 4.4; $\bar{x}_c^2$ is the mean of the concentration in the calibration step.)

By inserting this expression in the calibration model, the limit of decision is again obtained. This limit could be lowered if the mean of the calibration range lay as close as possible to the limit of decision. In fact, the highest concentration value should be lower than $10 \times x_{\mathrm{decision}}$. Figure 2 shows that the statistical distribution of observed values near the limit of decision provides values that fall with 50 % probability within the spread of the mean blank value. In other words, at this point the risk of wrongly accepting a blank signal as an analyte signal (an error of the first kind with a probability of α) is very low, but the risk of still rejecting an analyte signal as a blank value signal (an error of the second kind with a probability of β) is very high. An analyte signal is thus reliably detected when the last-mentioned risk is as small as possible. For $\alpha = \beta$, therefore, the *limit of detection* is twice the limit of decision (→ Chemometrics).

As can be seen in Figure 2, for a particular observed value $y_{\mathrm{quantitation}}$ an acceptable analyte interval $\pm\Delta x$ (for $x_{\mathrm{quantitation}}$) can be indicated only if $\Delta x \approx s_{\mathrm{res}}/a_1$ is not too large. This quotient is a critical quantity for the practical application of an analytical method. In addition, the relative standard deviation of an analyte result must sometimes not exceed a specified maximum. These two requirements lead to the definition or pragmatic setting of so-called limits of *quantitation*. Since such limits depend on the particular value observed, they cannot be regarded as meaningful characteristics of the method.

Historically, the simple $k\sigma$ criteria used in Figure 2 are based on numerous simplifications. They are usually based on the standard deviation σ of blank value measurements: The limit of decision is defined with $k = 3$, the limit of detection with $k = 6$, and as one possibility, the limit of quantitation with $k = 10$ [16]. (The k values take into consideration the probability α of erroneous statistical and, therefore, erroneous analytical decisions. Thus, by fixing k or α, the purpose of the particular trace-analytical procedure

can be taken into account.) The basic considerations in such definitions of method limits extend back to H. KAISER [14] and G. EHRLICH [17]. KAISER also tried to define characteristic quantities of methods for multicomponent analysis. For reasons of space, simultaneous multicomponent analysis cannot be discussed here [18, 19]. The previous discussion reveals how delicate results of trace analyses are in general. To achieve a responsible discussion in public it should be at least reported together with confidence limits (→ Chemometrics, Chap. 2.) and single values (means) should never be compared to critical target values.

4.4. Adequate Calibration Models

Calibration means the recording of observed values of response feature Y (signal quantity) typical of the method as a function of the analyte feature X (concentration or content) from N_c solid, liquid, or gaseous test portions. If possible, the analytical calibration portions should originate from different standard reference materials (SRMs), whose residual composition (matrix) corresponds largely to that of the expected analytical portions. If only one standard material or one standard portion is available, the analytical calibration portions should be prepared by spiking a blank test sample (standard addition) rather than by dilution of the standard (see Section 4.5). Here, analytical calibration portions for which the sampling error is negligible are assumed. (For nomenclature of sampling steps and of the aliquots arising from sample preparation steps, see [20].)

The objective of the calibration is to set up and statistically ensure the mathematical model of the functional dependence of the signal quantity Y on analyte content X. Linear models of the type

$$Y = a_0 + a_1 X$$

are most widely used.

Calibration in the narrower sense means determination of the sensitivity a_1. If N_c pairs of values (x_n, y_n) are available ($n = 1, 2, \ldots, N_c$), from which the arithmetic means $(\bar{x}, \bar{y})$ have been calculated, the sensitivity a_1 (regression coefficient, slope) can be calculated easily with the sum of squares of deviations from $n = 1$ to N_c, $SS_{XX} = \sum (x_n - \bar{x})^2$ and $SS_{YY} = \sum (y_n - \bar{y})^2$ and the sum of products of deviations $SS_{XY} = \sum (x_n - \bar{x})(y_n - \bar{y})$ where $a_1 = SS_{XY}/SS_{XX}$.

Since all straight lines of calibration pass through the center of gravity of the data $(\bar{x}, \bar{y})$, the constant a_0 is obtained from the model equation.

Strictly speaking, a_0 and a_1 are estimates of the true but unknown characteristic quantities of the calibration model, which represents an unknown relationship between Y and X. For that reason, $\hat{a}_0$ and $\hat{a}_1$ are used. The $\hat{y}_n$ calculated by this model for every given x_n generally differs from the measured y_n by the residual $\text{res}_n = y_n - \hat{y}_n$.

This model is based on several assumptions:

1) The quantity X can be set at points x_n without error
2) The variance of Y (at points x_n) is constant in the calibration range ("homoscedastic", which can be tested with the Cochran test [13, 21]; if the assumption is not confirmed, weighted regression models should be used)
3) The quantity Y (at the points x_n) shows a normal distribution (see Fig. 2, left) and is outlier-free (the latter situation is tested with the Dixon test [13, 21]; if the assumption is not confirmed, a robust regression method can be used, for example)
4) The residuals res_n (see Figure 2) are normally distributed and do not correlate with the values of x_n (which can be tested qualitatively by simply plotting res_n against x_n; if the assumption is not met, a nonlinear analyte – signal relation rather than linear may be valid)

In general, a linear model is naturally assumed to be justified. If tests [for coefficients of determination, goodness of fit (GOF) and lack of fit (LOF); see below] show this to be improbable, a nonlinear calibration must be selected, in the simplest case by using quasi-linear models (consideration of higher powers of X; polynomials), or piecewise linear calibration can be performed [22]. (A survey of various basic regression models can be found in [23]).

In trace-analytical methods the nonconfirmation of item 1 above must be reckoned with. The characteristic quantities a_0 and a_1 can then still

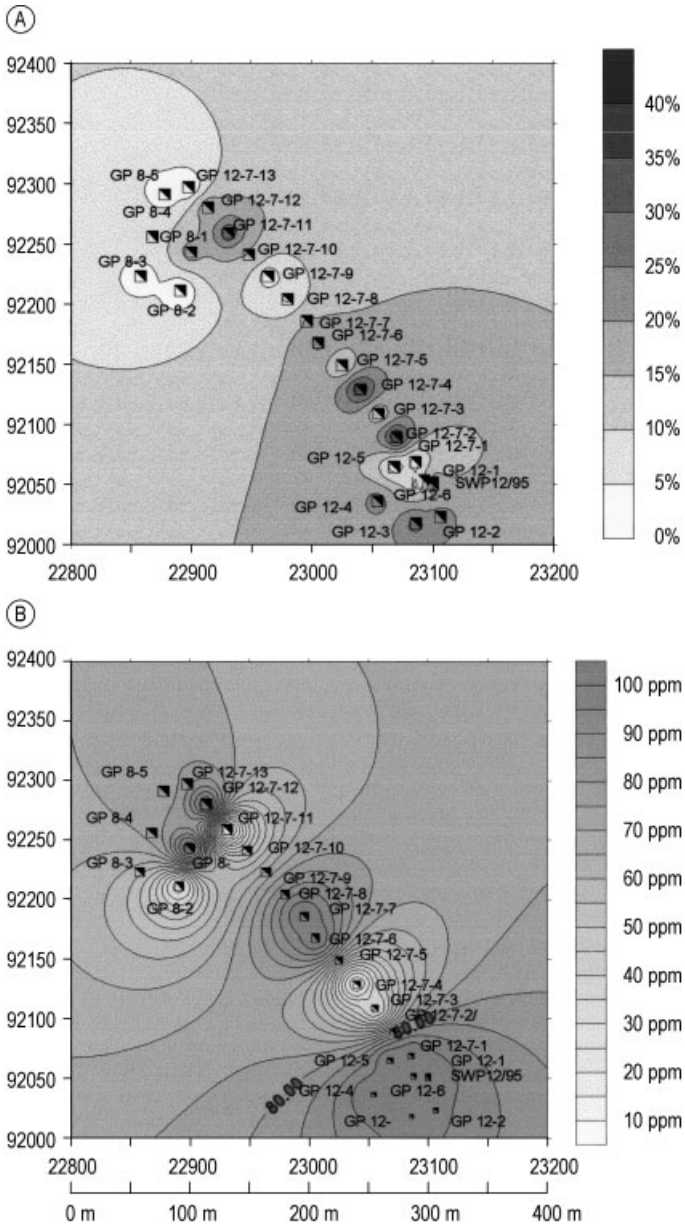

Figure 3. Typical example of the application of hand-held monitoring technique for the soil gas emission of a landfill site (depth of exploration of the drilling points: 1.50 – 1.80 m; visualisation by using SURFER software)
A) Determination of CH_4 distribution with an IR sensor; B) Determination of H_2S distribution with an electrochemical sensor

be calculated within the framework of a linear model:

1) $a_1 = \text{sgn}\ (\text{cor}_{XY})\ \sqrt{SS_{YY}/SS_{XX}}$, if the error ratio (corresponding to the long-term standard deviations) of the two response features Y and X is equal to the slope.
2) $a_1 = \frac{d}{2} + \text{sgn}\ (\text{cor}_{XY})\ \sqrt{\left(\frac{d}{2}\right)^2 + 1}$ where $d = \frac{SS_{YY} - SS_{XX}}{S_{XY}}$ if the errors of Y and X are equal (orthogonal regression)
3) $a_1 = SS_{XY}/(SS_{XX} - N_c \cdot \sigma^2{}_X)$, or
4) $a_1 = (SS_{YY} - N_c \cdot \sigma^2{}_Y)/SS_{XY}$, if the (long-term) variances σ^2 of the response features

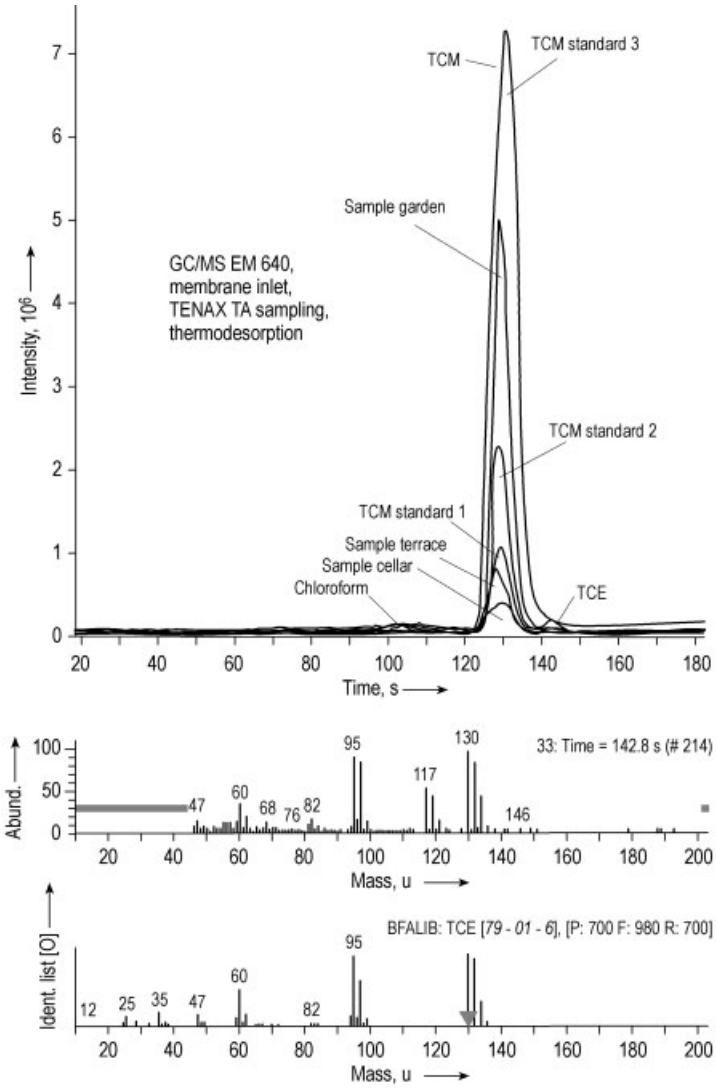

Figure 4. Presentation of some TIC runs of soil gas samples (garden, terrace; bore hole depth 2 m), gas standard and indoor air samples (cellar)
TCM = tetrachloromethane; TCE = trichloroethylene

are known from previous experiments. For more details and alternatives see [24].

The characteristic quantities obtained for the straight lines must be tested for their statistical significance and should be quoted with a confidence interval. Their standard deviation is a function of the residual standard deviation s_{res} of the regression model. This is defined as:

$$s_{\mathrm{res}} = \sqrt{\frac{1}{N_c - 2} \sum_{n=1}^{N_c} \mathrm{res}_n^2}$$

and also determines the width of the confidence band around the straight line of calibration.

If at the points x_n of the calibration range, only single measurements y_n are available, the quality of the calibration model can be evaluated only as a whole (sum of systematic model errors and random experimental errors) usually by testing the goodness of fit and by testing the coefficient of determination.

Only if replicate measurements of y_n have been made for at least one point x_n model errors

and experimental errors can be weighed against each other by LOF and complete analysis of variances, ANOVA. Replicate measurements also reduce the width of the confidence band.

Modeling, tests, calculation of the characteristic quantities of the method, and graphical visualization are supported by computer programs (e.g., [25]), which however usually consider only certain aspects of the calibration. An alternative method of calibration for imperfect response features and concentration quantities is to apply fuzzy theory [26].

4.5. Quality Assurance and Standard Reference Materials

Perhaps the greatest problem in trace analysis is assurance of the accuracy of the results (i.e., the avoidance of systematic errors). Systematic sources of error are possible in every step of an analytical process. The most reliable method for detecting systematic errors is continuous and comprehensive *quality assurance*, particularly by occasional analysis of (certified) standard reference materials. Strictly speaking, an analytical method cannot be calibrated if suitable (i.e., representative) standard reference materials adequately representing the matrix of the expected test samples are not available. However, *internal laboratory reference materials* can then usually be prepared, whose matrix largely resembles the matrix of the test portions expected. If problems occur in the preparation of such reference samples, the standard addition method (SAM) can be applied, in which internal laboratory standards are added stepwise to the test sample (analyte and matrix) and the sum of the two analyte contributions (from sample and standard) is determined.

In these cases, however, the danger exists that the speciation of the analyte in the added standard and that in the real sample are not identical.

Internal quality assurance can also be carried out by comparison of methods. In this case, two methods or procedures that are completely different in analytical principle should be applied to the same analytical problem. Comparison of one's own method with a certified method, is most useful. If the results of the two methods are plotted against each other in a scattergram, in the ideal case all values fall on a straight line with slope $a_1 = 1$ and intercept $a_0 = 0$. Here again, depending on the ratio of the random errors of the methods being compared, the characteristic quantities of the straight line corresponding to the actual data can be calculated as shown above, and interpreted and tested (statistically assured) as systematic additive (a_0) and multiplicative ($a_1 \cdot X$) error contributions.

External quality assurance, usually by participation in interlaboratory (round robin) tests, also contributes to the detection and avoidance of systematic errors.

However, the mere use of standard reference materials offers no guarantee of accurate, error-free calibration. Thus, many solid SRMs are useless for the calibration of solid microanalyses if the danger of microinhomogeneities is not recognized and if the minimum amount of test portion is not chosen accordingly [27].

This illustrates not only the difficulty of carrying out "perfect" quality assurance but once again the problems of reliable trace analysis in general.

5. Environmental Analysis

5.1. The Problem

The task of environmental analysis is the identification and quantification (screening and monitoring) of contaminants [28]. The analytical characterization and evaluation of dangerous wastes from the past are typical examples of applied environmental analysis. Traditionally, for the risk assessments of old waste deposits, analytical methods (in the form of costly laboratory analysis) are used remote from the site of investigation [29]. At abandoned waste deposits and industrial sites, contaminant distributions are extraordinarily heterogeneous in relation to surface, depth, and composition. As a result, by inexpert and not directly verifiable sampling, errors of many hundred percent are caused that are strikingly disproportionate to the complexity of the analysis of the sample material [30]. This inconsistency between precise but cost-intensive single-sample analysis and an incorrect sample can be avoided by applying on-site measurement techniques with hand-held devices, portable devices, or mobile devices with full laboratory capability installed in a vehicle [31]. In the case

of on-site work, the analytical results are mostly of limited precision, but since sampling, sample handling, sample preparation, and analytical work are combined directly at the site the information density is higher.

The cost of surveying and subsequent redevelopment has induced the EPA, within the framework of its SITE program, to start a monitoring and measurement technology program (MMTP). Part of this is the Environmental Technology Verification Program (ETV). The objective of this program is to organize the application of modern analytical techniques for field capability, to create new analytical methods for the field, and to apply these methods in addition to standard EPA specifications for contaminated sites [32].

The EPA has founded an online Field Analytic Technologies Encyclopedia (FATE), available on the Internet at http://fate.clu-in.org/. Summaries of applications of field analytical and site characterization technologies are available at http://clu-in.com. Some activities of the EPA are listed in Table 3. The newest results will be presented yearly at ON-SITE ANALYSIS conferences (http://www.infoscience.com). In Europe, too, there is some activity to establish such field screening methods [33].

Table 3. Some activities of the US EPA in field analytical methods: number of sites by technology (EPA 542-R-97-011, November 1997)∗

Technology	Number of Sites
Immunoassay	43
X-ray fluorescence	39
Cone penetrometer sensor	34
Gas chromatography	24
Fourier-transformed infrared spectrometry	3
Colorimetric test strip	3
Fiber-optic chemical sensor	3
Mercury vapor analyzer	2
Biosensor	1

∗ Some typical reports:
Cone penetrometer coupled with laser induced fluorescence probe
EPA/600/R97/019 (SCAPS-LIF), EPA/600/R97/020 (ROST)
Field portable gas chromatograph/ mass spectrometer
Bruker-Franzen EM 640 (EPA/600/R-97/149)
Viking SpectraTrak 672 (EPA/600/R-97/148)
Field portable X-ray fluorescence spectrometer (EPA/600/R-97-144,-145,-146)
Draft EPA Method 6200, March 1996: "Field Portable X-Ray Fluorescence Spectrometry for the Determination of Elemental Concentrations in Soil and Sediment".

5.2. Possibilities of Mobile Analysis

5.2.1. Definition and Requirements

Mobile analysis is a special form of on-site analysis in the environmental sphere. For this purpose the devices (sampling, sample preparation, measurement devices) are transported in a vehicle. The equipment can be integrated permanently into the car or operated hand-held.

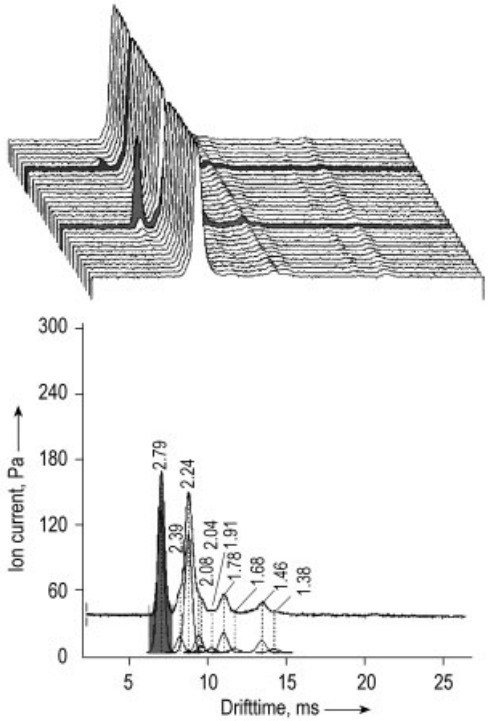

Figure 5. CCl_4 detection in soil gas (sampling depth 1.80 m) by using ion mobility spectrometry (8 bit IMS type RAID 1 from Bruker-Saxonia in the negative mode, push injection of 1 mL soil gas, measurement of the Cl-peak
red peak: sample taken near a walnut tree; Green peak: sample taken at the garden fence
Spectral range with two different sample injections (top) and a typical single spectrum (bottom) with a Cl peak (green)

In the screening mode, it needs methods which can identify unknown components and also quantify them. In the monitoring mode, the contaminant is known but not its distribution. In this case the analytical methods must be selective for the quantification of the contaminant. A typical example is the monitoring of organics (e.g., TNT, PAH) with immunoassay kits accepted by the EPA.

Mostly the field methods have higher detection limits than laboratory methods. Sample preparation and analytical procedure must be fast without extensive clean-up and separation steps.

The large amount of analytical information collected in the field is typical for mobile on-site analysis. The analytical techniques enable immediate analysis at the contaminated site, rapid change of site, and the necessary flexibility (surveying, site description, drill-core analysis, incident and damage analysis) in different measurement positions and at different sites.

Requirements for the use of field measurements are summarized below.

Analytical Requirements
Determination of field and sum parameters, toxic metals, anions and cations,
Organics: volatile organic compounds (VOC), total petroleum hydrocarbons (TPH), polynuclear aromatic hydrocarbons (PAH), polychlorinated biphenyls (PCB), pesticides
in soil, water, and soil vapor

Process Conditions
Sample preparation without or with only little chemical work
Linearity of the analytical signal over several orders of concentration
Large range of variation with small response of the analytical signal (robustness of the analytical method)
Multiparameter analysis with high selectivity
Simple procedures that require little knowledge of analytical chemistry
Low cost of the analytical procedure

Equipment Technology
Rugged design toward mechanical and climatic stress
Low-energy configuration
Minimum space requirement of the equipment with high operating convenience
Autonomous electrical energy generation
Availability of working surfaces and techniques for handling samples
Computer technique, modem connection, global position system device (GPS)
Geo-Environmental software (visualization and grid optimization)

Possible analytical methods for use in the field (ppm range or higher) are listed in what follows:

Field and global parameters
Hand-held devices for the water field parameters [T, pH, electroconductivity (EC), dissolved oxygen (DO), oxidation – reduction potential (ORP)]
Hand-held monitoring techniques for VOC and gases: flame ionization detection (FID), photoionization detector (PID), thermal conductivity detector (TCD), infrared sensor (IR), see Figure 3,
Transportable equipment for measuring adsorbable organic halogen compounds (AOX), chemical oxygen demand (COD), total organic carbon (TOC)

Anions/Cations
Ion-selective sensors, reaction kits, photometer, voltammeter for the determination of ions in water or leachates

Toxic Metals
EDXRF analyzer (hand-held or transportable, X-ray tube or radioactive isotope source) for elements in soils

Organics
GC, preferably with MS coupling and sample preparation [extraction, purge & trap, spray & trap, solid phase extraction (SPE), solid phase microextraction, headspace, thermodesorption, soil air trap], nondispersive infrared spectroscopy (NDIR), FT-IR, TLC, immunoassays and biosensors (for determination of defined substances)

5.2.2. Example of Equipment

Only in some cases are devices designed for work under field conditions. Generally, they are applicable for the determination of inorganic traces in water or leachate and also for the determination of gases and VOC in the soil air. Some special instruments (NDIR) exist for the determination of TPH in soils under field conditions. However, modern laboratory analytical devices are generally so safely transportable that they can be carried along, secured for transport, in vehicles and installed at the site of application. Part of the relevant technology has already

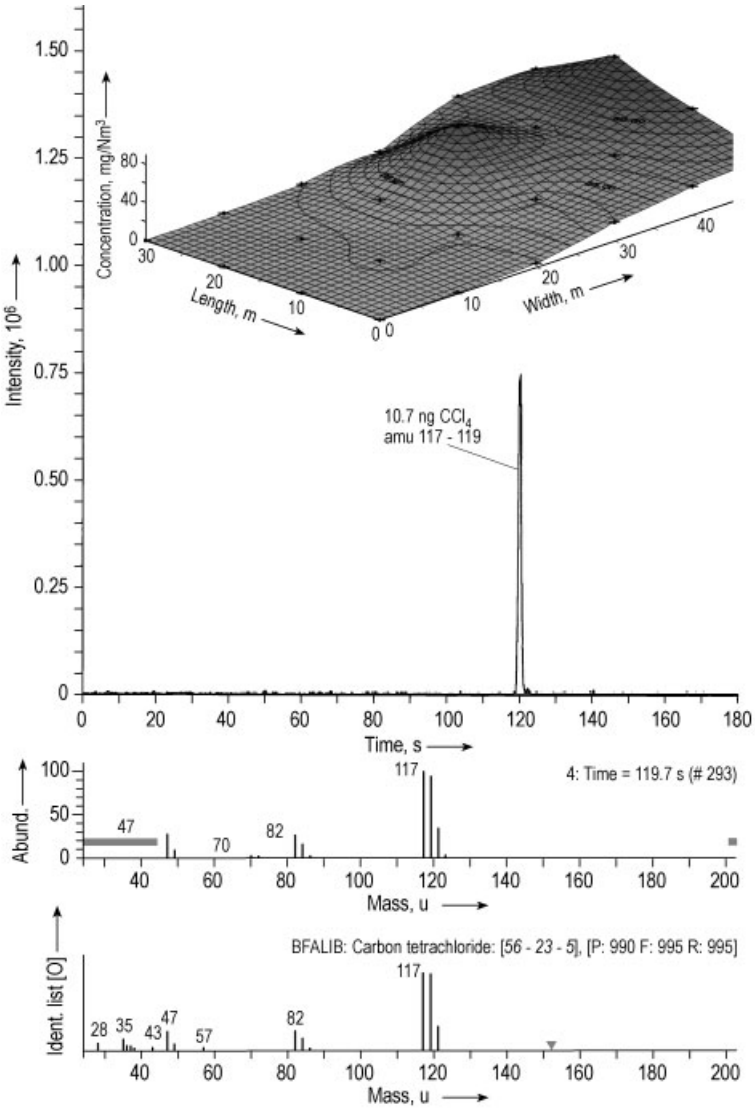

Figure 6. Selective detection of tetrachloromethane (TCM) contaminations by using SIM mode (measuring conditions: Bruker Daltonik GC/MS EM 640S with direct inlet; split 1 : 10; direct injection of 0.2 mL gas samples, 30 m HP 5 ms, $d_i = 0.25$ mm, $f_{th} = 0.25$ mm). The upper part of the figure incorporates the 3D visualisation of TCM distribution of soil gas (sampling depth 1.80 m; combination of IMS and GC-MS data) of part of the contaminated area in the neighborhood of the housing estate

been designed for mobile application. Examples include metal analysis by using EDXRFA (e.g., Spectrace QuanX, SpecTrace 6000, SpecTrace 9000) and the GC/MS EM 640 (membrane inlet and direct inlet) from Bruker Daltonik Bremen and Viking SpectraTrak 573 from Bruker Daltonics. The equipment is also fully suitable for work under laboratory conditions.

The equipment plan of a vehicle is determined largely by the planned task. For a widespread use of mobile on-site analysis, GC/MS and EDXRF are the base instruments. The devices work also with autosamplers. In the case of EDXRFA, the measuring results in the ppm range are obtained in 15 min, including sample preparation (drying, milling). In the case of GC/MS studies it is a function of the analytical problem and the quality, the amount, and the composition of the organics. The upper time

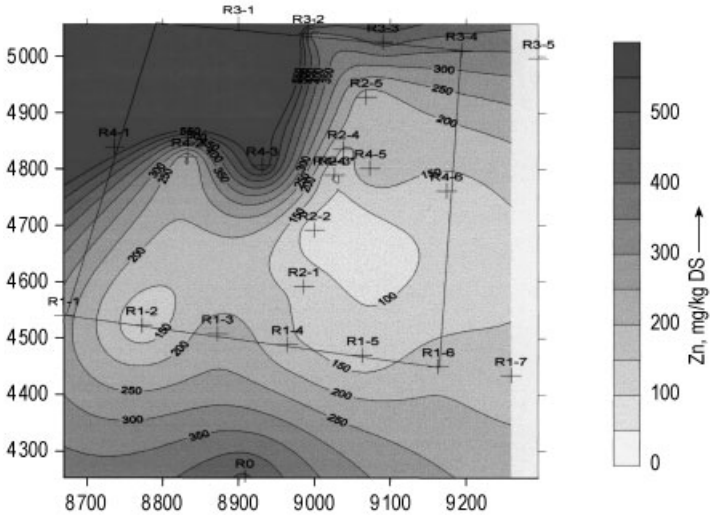

Figure 7. Determination of element distribution (e.g., Zn) in the subsoil of a landfill site by using mobile XRF (SPECTRACE 6000) and geostatistical methods

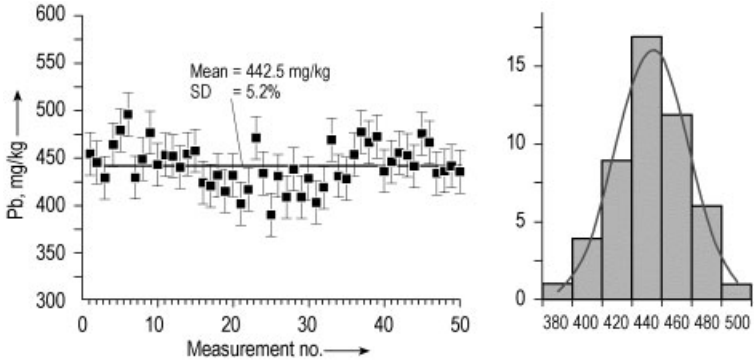

Figure 8. Reproducibility of the field screening EDXRF method [mean value, standard deviation (SD) and gaussian distribution of Pb content in excavated material from a waste disposal site]: determination of the repetition accuracy over a time scale of two months

limit for obtaining the analytical information is not less than one minute.

5.3. Selected Applications of Mobile Analysis

5.3.1. GC-MS Screening and IMS monitoring

A typical example of the application of GC-MS is the field screening of soil gas composition because of its simple sample preparation technique. A case study was to determine the VOC distribution, especially of tetrachloromethane (TCM), in a housing estate in the neighborhood of an old factory site. The first step of such studies is the determination of the soil gas composition and of the major components. The next step consists of choosing a fast, simple, and inexpensive measuring method for monitoring and controlling the major risk component, in this case TCM. Possible methods of TCM detection are gas reaction tubes, a hand-held photoionization detector (PID), a mobile gas chromatograph, a hand-held ion mobility spectrometer (IMS), and a mobile GC-MS. The application of gas reaction tubes is simple, but time consuming and expensive with poor reproducibility of results. The linear range of determination lies in an order of magnitude, and there is a pronounced nonconformity between the different tube types. The

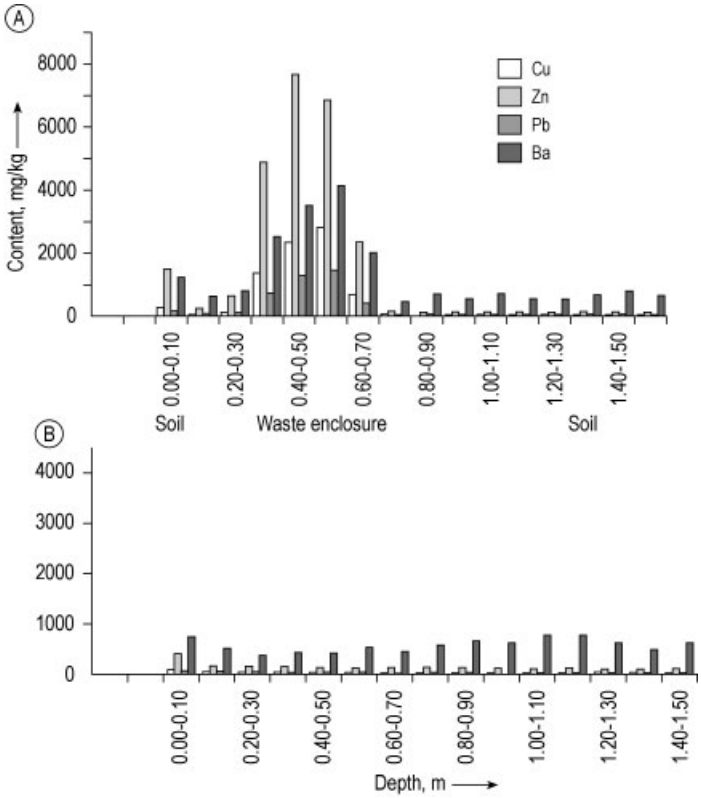

Figure 9. On-site determination of waste enclosure in a flood plain area in neighbourhood of a landfill site by using transportable EDXRF spectrometer (sampling by digging in the depth, waste contamination in profile 1)
A) Digging profile 1; B) Digging profile 2

PID with the typically used 10.6 eV UV lamp is not suitable for TCM determination. The IMS is the fastest instrument (overall analysis time < 2 min). It gives only a sum signal for halogenated hydrocarbons in the negative operation mode (selective detection of a Cl peak) with high sensitivity (< 0.2 vol ppm), but with a small range of linearity of determination and the possibility of extension by using a dynamic gas dilution. The mobile GC-MS system is the best technique in all the cases of field screening of nonpolar compounds with a high dynamic range of more than three orders of magnitude under real field conditions and also allows the detection and identification of other contaminants. The system is expensive and its operation is not simple. A good compromise is to use GC-MS as a master instrument and IMS as a fast monitoring instrument. The accordance of results between GC-MS and IMS is better then 20 %. Figure 4 summarizes the application of GC-MS in this field. After collection (indoor air and soil gas of boreholes) of the volatile organic compounds on TENAX TA the organics are thermodesorped and separated by gas chromatography. The total ion chromatogram (TIC) is analyzed with the aid of a MS spectra data bank and analytical software specially set up for the system (in this case Bruker DataAnalysis software), and the substances are identified. The system is also calibrated for quantification and the concentrations of substances are calculated with the aid of the software.

Figure 5 shows the detection of TCM by ion mobility. The term ion mobility spectrometry refers to the principles, practice, and instrumentation for characterizing chemical substances by means of gas-phase ion mobilities [34]. Ion mobilities are determined from ion velocities that are measured in a drift tube with supporting electronics. Ion mobilities are characteristic of substances and can provide a means for detecting

and identifying vapors. In practice, a vapor sample is introduced into the reaction region of a drift tube in which neutral molecules of the vapor undergo ionization, and the resultant ions, i.e., product ions, are injected into the drift region for mobility analysis. Mobility K is determined from the drift velocity attained by ions in a weak electric field of the drift region at atmospheric pressure. The normalized value K_0 is a characteristic constant of the product ion (in this case $K_0 = 2.79\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$ for the Cl peak).

By using a mass spectrometer with direct inlet (e.g., Bruker EM 640S) it is possible to inject directly the soil air samples and to report only the typical mass fragments of TCM in the SIM mode (Fig. 6). In this way the analytical procedure is faster and more sensitive. The TCM distribution using all results of IMS and GC-MS measuring can be visualized with special software (e.g., SURFER software from Golden Software Inc.). An example of TCM distribution is incorporated in Figure 6.

5.3.2. Elemental Analysis by Mobile EDXRF

Energy-dispersive X-ray fluorescence analysis (EDXRF) has proven to be ideal for on-site analysis [35]. With the availability of Peltier-cooled semiconductor detectors in EDXRF, its mobile application to field screening of toxic metals is unproblematic. In combination with geo-statistical methods or geographic information systems this technique provides important and meaningful information about the distribution of contaminants on hazardous waste sites. In Figure 7 this is illustrated by the determination of the metal distribution in a flood plain area in the neighborhood of a landfill site. A radioisotope is used to excite the characteristic fluorescent X-rays in a sample in hand-held devices. The instruments are rugged, fast, and easy to use. The sensitivity is lower than that of instruments equipped with adjustable X-ray tube sources, which have a detection limit in the ppm range. EDXRF can simultaneously determine 20 or more elements from sodium to uranium within only a few minutes. The instruments are also applicable in the laboratory. Some of these (e.g., the QuanX from Spectrace) also work with a 20-position sampler for automatic operation. The sample preparation is simple and fast and consists of drying and milling (particle size $< 100\ \mu\mathrm{m}$). The accuracy of the results depends on the quality of the used standards and the difference between the matrix of the standards and the sample (Fig. 8). Therefore, the results were calculated with the so-called Fundamental Parameter algorithm supplied with the spectrometer software. This algorithm corrects automatically for interelement and matrix effects. For this reason, no site-specific calibration is required, which is an enormous advantage over empirical calibration models.

Typical for the application of mobile EDRFA is the fast determination of element distributions in drilling cores or in finding hot spots in soil profiles (Fig. 9).

5.4. Conclusions

Field screening methods are used to determine the presence or absence of typical components at a given site. Needs for screening methods exist not only for site characterization, but also for determining the progress of remediation efforts. A driving force in mobile screening technology development and commercialization is the ever-increasing costs associated with environmental compliance. For the different contaminants, there many different field screening instruments. Many companies offer field test kits and a variety of detector tubes. Battery-operated spectrophotometers are available to support field analyses. Gas chromatography is one of the most mature technologies being applied in this field. More recent advances involve the use of kits based on immunochemistry, GC-MS coupling, and XRFA. Attractive features include a high degree of analyte selectivity and ppb to ppm sensitivity in most matrices.

6. References

1. O. G. Koch, G. A. Koch-Dedic: *Handbuch der Spurenanalyse,* 2nd ed., **vol. 1,** 1, Springer Verlag, Berlin 1974.
2. K. Beyermann: *Organische Spurenanalyse,* G. Thieme, Stuttgart 1982.
3. G. Knapp, M. Michaelis: "Chemie in Labor und Biotechnik," *CLB Chem. Labor. Betr.* **44** (1993).

4. G. Tölg, *Analyst (London)* **112** (1987) 365 – 376.
5. B. Sansoni, *Pure Appl. Chem.* **59** (1987) 579 – 610.
6. R. Kellner et al. (ed.): *Analytical Chemistry,* Wiley-VCH, Weinheim 1998.
7. G. Schwedt, A. Meyer, *LaborPraxis* **16** (1992) 712 –721.
8. U. Thiele, *Lab. Trend.* 1999 10 – 14.
9. L. Dunemann, J. Begerow: *Kopplungstechniken zur Elementspeziesanalytik,* VCH-Verlagsgesellschaft, Weinheim 1995.
10. M. Luthardt, E. Than, H. Heckendorff, *Fresenius' Z. Anal. Chem.* **326** (1987) 331 – 339.
11. DIN 32 645 (May 1994): Chemische Analytik; Nachweis-, Erfassungs- und Bestimmungsgrenze; Ermittlung unter Wiederholbedingungen; Begriffe, Verfahren, Auswertung.
12. DIN 55 350, part 34 (Feb. 1991): Begriffe der Qualitätssicherung und Statistik; Erkennungsgrenze, Erfassungsgrenze und Erfassungsvermögen.
13. *ISO Standards Handbook 3: Statistical Methods,* 3rd ed., ISO, Genf 1989.
14. H. Kaiser, *Spectrochim. Acta* **3** (1947) 40 – 67.
15. K. Doerffel, G. Michaelis, *Fresenius' Z. Anal. Chem.* **328** (1987) 226 – 227.
16. G. L. Long, J. D. Winefordner, *Anal. Chem.* **55** (1983) 712A – 724A.
17. G. Ehrlich, *Fresenius' Z. Anal. Chem.* **232** (1967) 1 –17.
18. K. R. Beebe, B. R. Kowalski, *Anal. Chem.* **59** (1987) 1007A – 1017A.
19. C. W. Brown, R. J. Obremski, *Appl. Spectr. Rev.* **20** (1984) 373 – 418.
20. W. Horwitz, *Pure Appl. Chem.* **62** (1990) 1193 –1208.
21. DIN ISO 5725 (April 1988): Präzision von Meßverfahren; Ermittlung der Wiederhol- und Vergleichspräzision von festgelegten Meßverfahren durch Ringversuche.
22. L. M. Schwartz, *Anal. Chem.* **49** (1977) 2062 – 2068.
23. J. D. Hwang, J. D. Winefordner, *Prog. Anal. Spectrosc.* **11** (1988) 209 – 249.
24. C. Hartmann, J. Smeyers-Verbeke, D. L. Massart, *Analysis* **21** (1993) 125 – 132.
25. J. Kramer: *STATCAL,* Umschau Verlag, Frankfurt/Main 1992.
26. Y. Hu, J. Smeyers-Verbeke, D. L. Massart, *Chemom. Intell. Lab. Systems* **8** (1990) 143 – 155.
27. J. Pauwels, C. Vandecasteele, *Fresenius J. Anal. Chem.* **345** (1993) 121 – 123.
28. I. L. Marr, M. C. Cresser, L. J. O. Ottendorfer: *Umweltanalytik,* Thieme, Stuttgart 1988.
29. V. Franzius, R. Stegmann, K. Wolf, E. Brandt: *Handbuch der Altlastensanierung,* **vol. 2,** R. v. Decker's Verlag, Heidelberg 1992 chap. 4.
30. L. H. Keith (ed.): "Principles of Environmental Sampling," *ACS Professional Reference Book,* 2nd ed., American Chemical Society, Washington D.C., 1991.
31. J. Flachowsky: "Mobile Umweltanalytik," in H. Günzler et al. (eds.): *Analytiker-Taschenbuch,* **vol. 18,** Springer-Verlag, Berlin – Heidelberg – New York 1998, pp. 143 – 180.
32. Abstracts of the Eight International Conference ON-SITE ANALYSIS, Lake Las Vegas, NV, January 2000, www.ifpac.com.
33. H. J. Gottlieb et al. (ed.): "Field Screening Europe," Proc. of the first international conference on strategies and techniques for the investigation and monitoring of contaminated sites, Kluwer Academic, Dordrecht 1997.
34. G. A. Eiceman, Z. Karpas: *Ion Mobility Spectrometry,* CRC Press, Boca Raton – Ann Arbor – London – Tokyo 1994.
35. G. Matz, W. Schröder, J. Flachowsky: "On-Site Investigation of Contaminated Soil by GC-MS and EDXRF-Techniques," in F. Arendt, G. J. Annokkée, R. Bosman, W. J. van den Brink (eds.): *Contaminated Soil '93, 4th Int. KfK/TNO Conf. on Contaminated Soil, Berlin May 1993,* vol. I, pp. 657 – 664, Kluwer Academic Publishers, Dordrecht 1993.

Radionuclides in Analytical Chemistry

Rainer P. H. Garten, Max-Planck-Institut für Metallforschung, Dortmund, Federal Republic of Germany

Juraj Tölgyessy, Department of Chemistry, Faculty of Natural Sciences, Matej Bel University, Banská Bystrica, Slovak Republic

1. Introduction 2
1.1. Definition and Purpose 2
1.2. History 2
1.3. General Features 2
1.4. Importance and Trends 4
2. Requirements for Analytical Use of Radionuclides 6
2.1. Safety and Operational Aspects . . . 6
2.2. The Labeled Substance 6
2.3. Activity Measurements 6
2.4. Choice of Radionuclide 7
2.5. Appraisal of Radionuclide Use in Analysis 7
2.5.1. Advantages 7
2.5.2. Disadvantages 8
2.5.3. Sources of Error 8
3. Radiotracers in Methodological Studies 9
3.1. Principles and Importance 9
3.2. Control of Sampling 10
3.3. Control of Contamination and Loss 10
3.4. Separation Procedures 11
3.5. Control of the Determination Stage 11
4. Isotope Dilution Analysis 11
4.1. Direct Isotope Dilution Analysis . . 11
4.2. Reverse Isotope Dilution Analysis . 12
4.3. Derivative Isotope Dilution Analysis 13
4.4. Substoichiometric Isotope Dilution Analysis 13
4.4.1. Substoichiometric Separation by Liquid – Liquid Distribution 14
4.4.2. Redox Substoichiometry 14
4.4.3. Displacement Substoichiometry 14
4.4.4. Applications 14
4.5. Sub- and Superequivalence Method 15
5. Radioreagent Methods 15
5.1. Simple Radioreagent Methods . . . 16
5.1.1. Determination with Labeled Reagents 16
5.1.2. Determination with Labeled Analyte . 17
5.1.3. Determination with Labeled Competing Substances 17
5.2. Method of Concentration-Dependent Distribution 17
5.3. Isotope Exchange Methods 18
5.4. Radioimmunoassay 18
5.5. Radiorelease Methods 18
5.5.1. Radioactive Kryptonates 19
5.5.2. Radioactive Metals 19
5.5.3. Radioactive Salts and Other Radioactive Substances 19
5.6. Radiometric Titration 20
6. References 20

Abbreviations

A	analyte
A_1	activity of component 1
AA	activation analysis
AAS	atomic absorption spectrometry
B	reagent
IDA	isotope dilution analysis
R∗	labeled radioreagent
SRM	standard reference material

1. Introduction

1.1. Definition and Purpose

Radioanalytical chemistry covers the use of radioactive nuclides and nuclear radiation for analytical purposes:

1) In the analytical determination of any chemical species or element
2) In the development, improvement, and quality assurance of analytical procedures for the benefit of analytical principles

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_111

Table 1. The scope of radioanalytical chemistry

Division	Section	Examples
Tracer labeling in organic analysis	labeling with tritium, ^{14}C, and heteroatoms (^{32}P, ^{36}Cl, ^{35}S, ^{125}I, ^{131}I, etc.)	random molecular sites selective molecular sites specific molecular sites
Tracer techniques in inorganic and organic analysis (addition of radioactive compounds to the sample)	chemical methodological studies and quality control of fundamental analytical processes (multistage procedures)	combustion and decomposition
		equilibrium constants
		phase separation efficiencies
		extraction
		volatilization
		precipitation
		redissolving precipitates
		coprecipitation
		electrochemical separation
		stability of solutions
		evaluation of sources of errors in trace and ultra-trace processes
		adsorption
		desorption
		volatilization
		contamination
		yields of complex analytical procedures
	kinetic studies	diffusion and forced diffusion (including electromigration) in solutions and gases
		diffusion, forced diffusion, and self-diffusion [a]
		chemical reactions: from long reaction times down to the ms regime (μs in special cases); macroscopic reaction order and rate
		reactive intermediate studies
		reaction mechanisms
		metabolism in organims and organelles
		rates of emission, immission, and migration of chemical species in environmental systems: natural, anthropogenic, and intentionally introduced radionuclides
	isotope dilution analysis	stoichiometric
		substoichiometric
		sub- and superequivalence method
	radiotracer- and radioreagent based analytical methods and pathway studies	radiochromatography
		radioreagent methods
		radiorelease techniques
		radioimmunoassay
		radiometric titration
		isotope exchange
		autoradiography
		studies of metabolism
		pathways of tracers in the environment (generally short-lived, for safety reasons)
		wear and wear protection in engineering materials
		corrosion and corrosion protection in engineering, construction, and biomaterials

1.2. History

The use of radionuclide techniques in analytical chemistry was first reported in 1913 by G. Hevesy and F. Paneth in a study of the solubility of lead sulfide in water, using the natural lead isotope ^{210}Pb as indicator [67]. Isotope dilution analysis was introduced by O. Hahn in 1923 [68], using ^{231}Pa to determine the yield of ^{234}Pa. The development of radioreagent methods followed, and further development of radioanalytical chemistry has established a range of analytical methods and techniques [1–4, 61, 65, 87, 93, 95, 97]. These include the use of artificial radionuclides (→ Radionuclides, Chap. 3.) and la-

Table 1. Continued

Division	Section	Examples
Activation analysis (AA), (production of radionuclides in the sample by nuclear reactions, their separation and measurement)	discriminating types of AA	Induced by
		thermal neutrons (reactors)
		epithermal neutrons (reactors)
		fast neutrons (generators, cyclotrons, isotope sources)
		protons (accelerators)
		α particles (sources, accelerators)
		fast ions of low to mid-range Z (^{6}Li, ^{9}Be, ^{12}C, ^{14}N, and others from accelerators)
		γ-radiation (reactors, sources, cyclotrons, synchrotrons)
		delayed (ordinary) AA vs. prompt [b] (prompt nuclear reaction analysis NRA) neutron- or charged particle-induced reactions
		fast [c] (spectroscopy of short-lived nuclides) vs. intermediate [d] and long lived [e] AA
		purely instrumental (mostly γ-spectrometric) vs. radiochemical i.e., group separations or specific element separations (including substoichiometric calibration techniques)
	applications	certification analysis of standard reference materials

beled compounds (→ Radionuclides, Chap. 6.), the principles of nuclear activation [4–10, 66] (→ Activation Analysis), and absorption and scattering of radiation (→ Radiation Chemistry) [11, 12]. The most important procedures are shown in Table 1.

1.3. General Features

Radiochemical analysis is based on two outstanding features of radioactivity:

1) The high sensitivity and ease of measurement of radioactive radiation [13, 14] (→ Radionuclides, Chap. 4.)
2) The possibility of labeling chemical compounds with radioactive tracers [15–17, 69, 70] (→ Radionuclides, Chap. 6.)

Radionuclide techniques often have higher sensitivity than other analytical methods. The amounts of nuclides, correlated to an activity of 1000 Bq (see Table 2), can be derived from the law of radioactive decay (→ Radionuclides, Chap. 1.4.). These amounts vary considerably, corresponding to the wide range of radioactive half-lifes. For 90 % of the commonly used nuclides [17], half-lifes range from several minutes to several years, so the corresponding masses are extremely low.

Radionuclides are often diluted with inactive isotopes, but specific activities (i.e., activity per total mass of element) are still very high. Since the background in nuclear spectroscopy is very low, and sensitivity is high, activities as low as 0.2 Bq can be readily detected (^{3}H, ^{125}I, ^{132}I), and even 0.01 Bq (several γ emitters, e.g. ^{24}Na, ^{38}Cl, ^{42}K, ^{46}Sc, ^{59}Fe, ^{60}Co, ^{65}Zn, ^{110m}Ag, ^{182}Ta, ^{187}W, ^{198}Au). Total activities of the order of 0.1 – 20 kBq (3 – 500 nCi) are often sufficient in analytical applications. The precision of the final result can be better than 2 %, if the counting error is kept below 1.5 %. This can be obtained from counting times between 100 s and 5 h.

Table 2. Mass corresponding to 1 kBq for carrier-free radionuclides

Nuclide	Mass [a], ng	Half-live
^{36}Cl	820	3×10^{5} a
^{14}C	6.1	5.8×10^{3} a
^{63}Ni	0.57	120 a
^{85}Kr	6.8×10^{-2}	10.6 a
^{125}I	1.6×10^{-3}	60 d
^{3}H	2.8×10^{-3}	12.4 a
^{131}I	2.2×10^{-4}	8.04 d
^{35}S	6.3×10^{-5}	87.2 d
^{32}P	9.5×10^{-5} *	14.3 d
^{99m}Te	5.1×10^{-6}	6.02 h
^{18}F	2.8×10^{-7}	1.83 h

* Commercially available with maximum specific activity of 2×10^{14} Bq/g.

Table 1. (Continued)

Division	Section	Examples
		basic research in analytical science and technology
		routine control of high-purity materials
		trace elements in biomaterials (including medicine)
		routine production control of commercial high-performance materials
		on-line and in-line process control (e.g., bomb detection)
		trace elements from environmental materials (water, dust, soil, sludge)
		trace elements in plants (pollutants, uptake, storage, mobilization paths)
		trace elements in geochemical and cosmochemical matter
Determination of natural (H, C, K, Th, U, Ra, etc.) and anthropogenic (Cs, Sr, Tc, Np, Pu, etc.) radioactivity		levels and distribution of activity in environmental materials (air, water, soil, etc.)
		dating in archeology, fine art, historical, and environmental research (^{14}C, ^{3}H, ^{7}Be, ^{10}Be, ^{40}K, ^{232}Th, ^{235}U, ^{238}U)
Nuclide radiation sources in analysis (on-line and remote sensing devices)		radionuclide-induced X-ray fluorescence analysis (on-line product flow control, outdoor and remote sensing)
		resonance scattering of γ rays
		resonance absorption of γ rays (→Isotopes, Natural, Chap. 3.2.3.)
		absorption and backscattering of X, α, β and γ radiation (on-line control) :
		mass density monitoring
		product mass flow control
		monitoring of coatings and films
		radiographic testing of materials and devices
		detection of radiation induced radicals
		radiation dosimetry
		particle size distribution
		absorption and moderation of neutrons for light element detection
		remote B- and Cd-sensing
		humidity measurement of product flows
		plasma desorption mass spectrometry (soft ionization technique induced by fast heavy fission nuclei from ^{252}Cf-source)
		prompt (n, γ) reactions (for prospecting, oil and gas well logging)
		α- and β-ionization detector devices (pressure gauges, gas flow metering, smoke sensors, electron capture detectors).

[a] Diffusion of a compound or element in a matrix containing an excess of the same species: (active and inactive) isotopic tracer methods are the only methods available. [b] $\Delta t < 10^6$ s. [c] half-lifes $\tau < 1$ h. [d] half-lifes $\tau < 3$ d. [e] Half-lives $\tau > 20$ d.

Analytical applications of radionuclide techniques rely on the assumption that different isotopes of the same element exhibit the same properties in any macroscopic physical or chemical process, and that radioactive labeling does not influence the other properties of a chemical species. This is generally the case, with deviations below 1 % (with the exception of hydrogen isotopes) owing to isotopic fractionation or radiation effects (→ Isotopes, Natural, Chap. 2.1., → Isotopes, Natural, Chap. 2.2.). For analytical purposes, the radiotracer and the analyte must be present in the same chemical form. This is usually to achieve, but specialized preparative techniques may be necessary for radioactive labeling of more complex organic compounds (see Section 2.2; → Radionuclides, Chap. 6.3.).

1.4. Importance and Trends

Radionuclides are used in many subdivisions of analytical chemistry (see Table 1). Of major importance are radiotracers in methodological and pathway studies, isotope dilution analysis (IDA), radioimmunoassay (→ Diagnostic Reagents, Chap. 3.2.), and nuclear activation analysis (AA) (→ Activation Analysis) [66]. They are all especially suited to analyze the extremely small amounts of substances encountered in ultra-trace analysis or in trace analysis of microsamples.

Over the past two decades, the emphasis has shifted from high detection power in routine analysis toward an independent approach, applying this high detection power to the development of analytical procedures and reference materials.

Radiochemical methods for routine analysis have lost ground to other, primarily spectroscopic, methods (see Chap. 3) [18–21]. Nonradiochemical methods often yield highly reproducible results, but may involve systematic errors.

The need for increased reliability and analytical quality control has emphasized the usefulness of radiochemical methods for the certification of standard reference materials (SRM) [9, 22, 23]. Radioanalytical methods are often suitable for homogeneity testing and distribution analysis of traces in SRMs. Activation analysis, radiotracer techniques, and isotope dilution analysis are becoming increasingly important for assessment of analytical quality.

Radioanalytical methods are well suited to the determination of basic analytical data, such as equilibrium constants, or kinetic data (see Section 3.1). These data are important for development and optimization of new analytical procedures [61].

The most frequent analytical use of tracers is in the biomedical sciences. Radioimmunoassays, in the form of rapid diagnostic test kits have led to the development of nonradioactive assays based on similar principles (→ Diagnostic Reagents, Chap. 3.) [2, 24].

Many important tracer applications can be substituted by other methods. Even IDA and tracer applications in self-diffusion studies can be replaced by inactive isotope tracer methods using mass spectrometry and other methods for isotope ratio determination. However, because of the extremely high sensitivity of IDA, radioactive tracers are of unique usefulness in radioimmunoassays, radiorelease reagents, radiochromatography, AA, and for systematic studies in trace and ultra-trace analysis, physiological chemistry, IDA, diffusion, isotope exchange, and physical chemistry of solids.

Radioanalytical chemistry will continue to occupy an important position, despite increasing criticism of nuclear technology.

Recently, there has been little innovative work in the field of methodology, but the applications field has continued to expand [7, 25, 66]. Highlights include application of isotope enrichment techniques in environmental studies [26, 71, 72], improved accuracy in the characterization of SRM [23, 27, 28], preirradiation chemistry with high-purity reagents under clean room conditions to permit speciation from AA [7, 26] (sometimes termed molecular AA [25]), derivative AA [7], accelerator-based dating methods combined with IDA, short-time and pulsed reactor activation, extension of the sub- and superequivalence method [73–75], in-vivo analysis by prompt-γ neutron AA [7] and by short-lived AA [8, 29], and extended use of the single comparator (k_0) standardization method in neutron AA [30].

2. Requirements for Analytical Use of Radionuclides

2.1. Safety and Operational Aspects

Safety and operational aspects involve the hazards of ingestion of radionuclides and exposure to radiation [76] (→ Radionuclides, Chap. 7.).

2.2. The Labeled Substance

Aspects of the labeled substance involve the following requirements [14, 20, 26, 31–33, 66]:

1) Radionuclide and radiochemical *purity* are essential (→ Radionuclides, Chap. 6.2.3.). Radiochemical purity can be checked by reverse IDA (see Section 4.2). The labeled substance must react identically to the analyte.

2) Complete *homogenization* and *isotope exchange* must be achieved, requiring successive transformation of the active tracer as well as of the inactive analyte into all chemical species that are considered to occur in the complete system [20, 26, 34].
3) *Isotope effects* (→ Isotopes, Natural, Chap. 2.1., → Isotopes, Natural, Chap. 2.2.) must be negligible.
4) For most elements, suitable radionuclides for analytical use are commercially available. There are only seven elements of interest (B, He, Li, N, Ne, O, Mg), for which this is not the case [17].
5) The *preparation* of a specific isotope, of an isotopically modified (i.e., mixture of labeled and unlabeled), isotopically substituted, or even site-specifically labeled compound (→ Radionuclides, Chap. 6.3.) is often mandatory in analytical applications.
6) *Synthesis* (including biosynthesis) of these labeled compounds often requires much more effort than the actual radioanalytical experiment [34]. A variety of labeled compounds and pharmaceuticals are available commerically from stock, and a further range of compounds will be prepared on request by a number of suppliers [35]. In special cases, doubly and triply labeled molecules have been used to follow the reactions of different functional groups.
7) Costs of radionuclides for analytical applications are typically in the range \$ 150–400 per 40 MBq (= 1 mCi), and higher where preparation is difficult or time consuming (e.g., ^{54}Mn, ^{85}Sr, ^{36}Cl: \$ 700–4000 per 40 MBq). Labeled substances from stock usually cost between \$ 2000 (^{14}C) and \$ 200 (^{3}H) per 40 MBq; the price depends on the cost of preparation.

2.3. Activity Measurements

(→ Radionuclides, Chap. 4.3.)

Commercially available instrumentation is generally used, with the exception of some research work and highly specialized routine analytical applications.

Quantitative measurement of radioactive substances is generally straightforward with the usual counting devices and relatively simple with γ-ray, X-ray, and hard β emitters. The *detection limit DL* of the concentration of the analyte is given [36, 77] by:

$$DL = \frac{f}{s}\sqrt{B} \tag{1}$$

where $f = 3$ is appropriate for detection at the 99.9 % confidence level, B is the blank level, in-cluding analytical blank and spectral background of the detection device, and s is the analytical sensitivity given by the slope of the analytical calibration function, i.e., under idealized conditions:

$$s = \varepsilon \eta t a m_{\text{sample}} \tag{2}$$

where ε is the detection efficiency, η is the emission probability, t is the counting time, a is the specific activity of the analyte, and m_{sample} is the mass.

Specimen preparation for radioactivity measurements aims, in general, to obtain the analyte:

1) As completely as possible and with the highest possible degree of purity from extraneous radioactivity and from excess matrix matter
2) Concentrated on the smallest possible area or volume of a solid, liquid, or gaseous (for H, C, S, Ar, Kr, Xe, Rn) target
3) On a specimen support that is compatible with the chemical nature of the species analyzed, with the specific radiation to be measured, and with the detector system, e.g., liquid scintillation cocktails [37] for weak β emitters, or aluminum trays for dried layers of hard β emitters [13].

These preparation requirements are compatible with the more general rules for specimen preparation in trace and ultra-trace analysis by any technique [18, 40, 45]. Requirements are different for measurements of: γ radiation of energy >50 keV; hard β radiation >200 keV; α radiation; weak β radiation below 200 keV; and X radiation below 70 keV. In this respect, specimen preparation follows the general outline of preparation procedures for small counting samples [13, 37].

2.4. Choice of Radionuclide

The choice of radionuclide involves:

1) Measurement, possibly including discrimination between different indicator nuclides or indicator and interfering nuclides (e.g., from byproducts or blank level impurities)
2) Suitable half-life
3) Chemical species

γ *Emitters* are preferred for multitracer experiments (most of all in nuclear AA) for applications where the tracer is dispersed in, or shielded by a material (e.g., in-vivo diagnostics in biology and medicine, measurement from bulky solid material), but also from tracer solutions.

Pure β *emitters* are especially useful where the shorter range of their radiation is utilized, e.g., for autoradiography [39], and for surface and thin film studies, including adsorption, corrosion, and catalysis. Measurements of β-emitting tracer solutions are performed with high efficiency by liquid scintillation.

α *Emitters* are only used as tracers in special cases, either where their very low range is important (e.g., surface and thin film techniques), or where they represent the only available nuclides of the element (e.g., actinides).

Imperfect tracers are often used when labeling of a specific chemical site is inappropriate with isotopic nuclides. Imperfect tracers are widely used to label gas streams with inert gases (^{85}Kr, ^{41}Ar, ^{133}Xe), water with dissolved anionic tracers (^{82}Br, ^{131}I, ^{24}Na), hydrocarbon fuels with dissolved ^{60}Co-naphthenate or 1,2-^{82}Br-dibromoethane. Similarly, loss of metal ions at ultra-trace levels due to adsorption on the vessel walls and by volatilization have been modeled by ^{195}Au, ^{203}Hg, ^{60}Co [18, 20, 40, 78], which are easily measured, and which are available at high specific activity.

2.5. Appraisal of Radionuclide Use in Analysis

2.5.1. Advantages

Utilization of radionuclides in analysis has the following advantages:

1) Cost effectiveness [4, 9, 40–44, 66] (Section 2.2).
2) Element specificity and chemical site specificity (Section 2.2).
3) High dynamic range in a single experimental run. Radionuclide blanks are often extraordinarily low, far below detectability. Multielement analysis by simultaneous use of different tracer nuclides (see Section 2.4, → Radionuclides, Chap. 6.3.), often without separation.
4) Quality and quantity of the signal-producing decay are independent of influences such as temperature, pressure, external fields, and the chemical environment of the nucleus.
5) Locally resolved identification. Distribution detection, e.g., from within an analytical decomposition apparatus, a growing plant, a thin-layer chromatogram, or sections from biological tissue or engineering materials. Lateral resolution as low as 10 µm by autoradiography [39].
6) Predetermination of counting precision [13, 49].
7) Following reaction schemes or metabolites: in vitro and in vivo with fast element detection, often from outside the analytical system at comparable levels of sensitivity and accuracy.
8) No interference with analysis by other methods (nondestructive, conservative).

2.5.2. Disadvantages

Disadvantages of radionuclides in analytical chemistry are as follows:

1) Synthesis of labeled compounds may be difficult
2) Safety considerations (see Section 2.1) are required to avoid incorporation of radiotoxic substances and radiation exposure, specialized laboratory and equipment, qualified and skilled staff, disposal of radioactive waste
3) Boron, aluminum, nitrogen, and oxygen are of analytical interest, but no nuclides are available with half-lifes longer than 10 min
4) Specific sources of error due to radiotracers (see Section 2.5.3) have to be controlled
5) No distinction between different chemical species without additional separation
6) In general, sample material cannot be returned to normal use after analysis. Safe handling may restrict use of some analytical processes

2.5.3. Sources of Error

In analytical applications of radionuclides, specific sources of error can generally be ascribed to either analytical and trace-analytical considerations, or particular features of radionuclides and their measurement.

Difficulties in handling sub-nanogram amounts of analyte are well known [40,45]. Because of the very low concentrations considered, chemical techniques such as solvent extraction, electrodeposition, volatilization, coprecipiration, chromatography, and ion exchange may be necessary. Additionally, surface adsorption on vessel walls, colloidal matter, and dust particles in the system may be a significant source of error.

Errors may be prevented by blank experiments, high acid concentration above 1 mol/L (diminishing adsorption), selection and preconditioning of vessel materials [20,40,44–46]. Since these problems are not specific to radiotracers, they can be most effectively reduced by addition of an inactive isotopic carrier of the same chemical species.

Difficulties in Obtaining Representative Labeling. For a tracer experiment, it is necessary to process the tracer so that it is chemically identical to, and in the same phase as the unlabeled substance. Merely mixing the radiotracer with the analyte sample is generally not sufficient (see Section 2.2) [20,26,33,34,47].

Radiation-induced effects (→ Radionuclides, Chap. 1.5.2.) in labeled molecules occur with compounds of high specific activity, labeled with weak β emitters or even α emitters. Since autoradiolysis takes place mainly via secondary radical reactions of fragments formed from solvents or major components, radiolytic damage can be reduced by cooling and by adding a radical scavenger.

Impurities from decay products (i.e., from radioactive decay and from radiation-induced decay of chemical compounds) can interfere with the activity measurements, or with the chemical reactions employed to process the sample. Radionuclide purity is a function of time.

Limited radiochemical purity may result in detection of radioactivity from chemical species different from the intentionally labeled one. This is of prime importance in specifically labeled enantiomers. Labeled compounds should be purified immediately before use. Chromatography and reverse IDA (see Section 4.2) are important for checking radiochemical purity.

Isotope effects are important in reactions where tritium is involved in the rate-determining step [48], because it reacts more slowly than hydrogen. Similar effects, but to a much lesser degree, are reported for ^{14}C, ^{32}S, and other nuclides [4,17,48].

Errors Due to Measurement Setup. Some sources of error that are specific to radionuclides result from the measuring technique:

1) Errors due to radioactive decay are generally smaller than 0.2 % since half-lifes are accurately known [3,13,30,49].
2) *Statistical uncertainty* results rom counting a limited number N of radioactive decay events. Mostly, there is no problem in obtaining $N \approx 10^4$ with precision at the 1 % level, as long as background is low.
3) *Geometric errors* result from uncertainties in the dimensions of the sample, detector, and the distance between them. They are often difficult to calculate on an absolute scale, and are therefore calibrated empirically. Depending on the calibration precision, on the type of nuclide, sample, and detector, the relative uncertainty can range from 50 to 0.2 %. Calibration on an absolute scale depends on the use of accurate standards, which are rarely available with a certified accuracy better than 2 – 3 %.
4) *Errors from absorption and backscattering* of radiation are generally corrected empirically as well.
5) *Errors from instrumental sources*, *quenching, and luminescence* can be kept very low (ca. 0.1 %) by control of temperature and humidity, uniform count rates from samples and standards to suppress differences in dead time and pulse pile-up, and uniform matrix of samples and standards in liquid scintillation counting [37]. Otherwise, errors as high as

40 % [37, 79] can be reduced to a few percent, either by utilizing fast detectors and electronics, or by empirically correcting for pulse pile-up losses, dead time, and spectrum instability [5, 13, 79].

3. Radiotracers in Methodological Studies

Radioanalytical chemistry offers a variety of accurate and highly sensitive methods. Even with the increased performance of other trace analysis methods [18–21], including electrothermal atomic absorption spectrometry, atomic emission, and fluorescence spectrometry (→ Atomic Spectroscopy), coupling methods, and high-performance sample introduction in organic and inorganic mass spectrometry (→ Mass Spectroscopy), chromatographic separations (→ Gas Chromatography; → Liquid Chromatography), and neutron AA (→ Activation Analysis), there are still a number of areas where radionuclide applications are unsurpassed (see Section 1.4).

Picogram amounts of chemical species can be traced through a complete chemical process, and problems in process performance are often revealed and overcome rapidly and simply [18, 20, 22, 23, 28, 33, 40, 41, 45, 47, 50, 61, 78, 93, 97].

Important information can be obtained rapidly and cheaply by radioindicators. Specific procedures for quantitatie determinations based on radiotracers (see Chaps. 4, 5) involve additional principles.

3.1. Principles and Importance

Radioactive tracer analysis is based on the simple proportionality between the mass m_x of analyte x and the radioactivity A_1 initially added to the sample:

$$m_x = k \cdot A_1 \tag{3}$$

where k is a proportionality constant, measured for an individual experiment. This is based on the condition that the radioactive substance m_1 added to the sample does not increase the mass of the analyte, i.e., $m_x >> m_1$. Otherwise, for determination of m_x from A_1 the formalism of IDA (Chap. 4) is applied, but very often the simple proportionality between mass ($m_x + m_1$) and A_1 is sufficient for a methodological investigation.

By adding a radioactive tracer before the individual process stage, and following it by radioactivity measurements, the recovery, distribution, and loss are measured independently of any additional risk of contamination. By means of nuclear activation, radioactive tracers can be introduced into solid samples and can be used to determine the recovery and loss during decomposition of solids, and separation from solid samples by volatilization, leaching, and solvent extraction [9, 10, 40, 51–64, 67–80, 93, 97].

A variety of examples from the general analytical field [2, 7, 17, 25] and from trace element analysis [20, 33, 44, 46] covers separate stages of combined analytical procedures such as decomposition [20, 44, 45], evaporation [81], coprecipitation, adsorption, and washout of precipitates; contamination and adsorption [18, 33], separation by chromatography, ion-exchange, liquid – liquid distribution, and electrodeposition [4, 33, 34, 38, 40, 41, 78].

Radioanalytical indicator methods (including the radioreagent method, see Chap. 5) experienced a revival in the 1960s and early 1970s. Modern direct instrumental methods [19] underemphasized the importance of radioanalytical methods, which are necessary to establish the accuracy of spectroscopic methods, to avoid systematic errors, and to assess trace content in SRMs. An important advantage is the ability to reveal accidental losses, even at the ultra-trace level.

The most common areas of analytical applications are:

1) Determination of equilibrium constants, including solubility and complex dissociation constants, phase distribution in solvent extraction, adsorption, ion exchange, coprecipitation; and kinetic data, such as rate constants
2) Tracing the distribution of a chemical species in an analytical system for decomposition, chromatography, evaporation, and separation
3) Determination of recovery, yield, and loss of a chemical species in a process

3.2. Control of Sampling

Inaccuracies in sampling procedures occur, owing to contamination, loss, and nonrepresentative sampling. Sampling, pretreatment, storage, etc., all affect the accuracy.

Extended radioindicator studies on trace element sampling (e.g., of Al, Cr, Mn, Fe, Co, Zn, Sr. Cs, As, Hg, Pb) have been reviewed [4,5,20, 32]. In environmental analysis (e.g., solids, water, aerosols, biological tissue), adequate sampling and sample treatment is of primary importance to account for variation in biological activity due to the distribution of chemical species between different phases and solid fractions [5, 16,20,22,26,32,46]. Sample homogeneity may be a significant limitation to the quality of analytical results. It is of increasing importance with decreasing concentrations from the μg/g to the ng/g level, and it limits the amounts of subsamples and the number of replicates. Studies on sample homogeneity, and on sampling and storage of very dilute solutions, are conveniently performed with radiotracers for numerous trace elements [16,18,20,33,44,45].

3.3. Control of Contamination and Loss

Contamination generally poses the most important problem in ultra-trace analysis. It causes fluctuations of the blank values, thus defining the lower limits of detection, but also introduces systematic inaccuracies. Tracer techniques can be used to study the sources of contamination: reagent blanks, vessel walls, airborne pollutants, etc.

Contamination cannot be overcome by radiochemical means. Purification procedures for reagents and vessels and a clean working environment are required and SRMs should be used. Tracer techiques are unique tools in its investigation and control (for reviews see [18,34,40,44]), owing to the advantages mentioned in Chapter 2 and Section 3.1 [18,44]. Radioindicator studies have been conducted on reagent purification by electrolysis, sublimation, sub-boiling distillation, liquid – liquid distribution, ion exchange, and on vessel cleaning by rinsing, leaching, and steaming.

In general, the same principles are applied to radiochemical investigations of losses by adsorption (i.e., during storage, pretreatment, precipitation, filtration), by volatilization (during decomposition, ashing, storage, digestion, drying), and by chemical reaction (complexation, ion exchange, photochemical and redox reaction) using radioindicators [20,40,45].

3.4. Separation Procedures

Separation stages and chemical reactions, e.g., for decomposition or phase transformations, are essential components of combined analytical procedures [52,53]. They should be reproducible, quantitative, selective, and unequivocal.

Separation procedures are based on the principles of volatilization, liquid – liquid distribution, adsorption, diffusion, chromatography, ion exchange, electrophoresis, precipitation, coprecipitation, and electrodeposition. In all of these, radiotracers provide the best tool for methodological investigations, determination of equilibrium constants, kinetic data, and optimization of applied analytical data (yield, interference levels, etc.) [54]. Use of radiotracers in complex multielement separation schemes is reviewed in [4,17,20,41,54], radiochromatography is reviewed in [55,61,93,97].

3.5. Control of the Determination Stage

Radionuclides in methodological studies of determination methods are sometimes useful. These are the measurement processes associated with the determination stage of a combined analytical procedure. Sources of systematic error in atomic absorption spectroscopy, optical emission spectrometry, and electrochemical methods. as well as optimization of the determination procedures have been examined.

Examples are the behavior of trace elements in graphite furnace atomizers [81] and investigations of the double layer structure on analytical electrodes and its exchange reactions with the solution [56].

4. Isotope Dilution Analysis

The principle of isotope dilution analysis (IDA) [31,47,90,97,98] involves measurement of the

change in isotopic ratio when portions of a radiolabeled and nonlabeled form of the same chemical species are mixed. To perform a radioisotope IDA, an aliquot of a radioactive spike substance of known specific activity $a_i = A_i/m_i$ is added to the test sample containing an unknown analyte mass m_x. This analyte mass is then calculated from a determination of the specific activity a_x of the resulting mixture after complete homogenization. The important advantage of IDA is that the analyte need not be isolated quantitatively. This is often significant at the trace concentration level, where quantitative separation is not feasible or inconvenient, or if interferences occur, e.g., in the analysis of mixtures of chemically similar compounds. Only a portion of the analyte is separated, so that the separation reagent is not used in excess, and this generally improves selectivity [31,54,57–59]. Selective determination of *chemical species* in various states (e.g., tri- and pentavalent As, Sb, tri- and tetravalent Ru, Ir) by IDA is favored by the supplementary selectivity gained from the use of substoichiometric amounts of reagent. Losses during purification and other steps are taken into account. A suitable separation procedure is needed to isolate part of the homogenized sample [2,31].

In more advanced variants of IDA, measurements of masses are substituted by volumetric and complementary activity measurements.

The same principle is also applied to inactive stable isotope IDA. Mass spectrometric determination of isotope ratios then replaces the activity measurements [82] (→ Mass Spectrometry).

4.1. Direct Isotope Dilution Analysis

In this most frequently used version of IDA (single IDA), the mass m_x of inactive analyte substance can be determined by using the labeled substance of mass m_1, radioactivity A_1, and the specific activity a_1:

$$a_1 = A_1/m_1 \tag{4}$$

After homogenization with the analyte mass m_x, the specific activity has decreased to:

$$a_2 = \frac{A_1}{m_1 + m_x} = \frac{A_2}{m_2} \tag{5}$$

The total radioactivity of the system is unchanged:

$$(m_1 + m_x)\, a_2 = m_1 a_1 \tag{6}$$

The unknown mass m_x is calculated from:

$$m_x = m_1 \left(\frac{a_1}{a_2} - 1\right) \tag{7}$$

The specific activities a_1 and a_2 are obtained from activities A_1 and A_2 and masses m_1 and m_2 of the fractions isolated from the initial labeled substance (Eq. 4) and from the homogenized solution (Eq. 5).

Equation (7) shows that IDA depends on the relative change in specific activity due to isotope dilution. Low-yield separations may be sufficient. For high precision, a_1 should be high, with mass $m_1 \ll m_x$. Then Equation (7) reduces to

$$m_x \approx m_2 \frac{A_1}{A_2} = \frac{A_1}{a_2} \tag{8}$$

Equation (8) shows that the accuracy of m_x is limited by the physicochemical determination of m_2.

Example. Determination of glycine in hydrolyzed protein [98]: 152.6 mg hydrolyzed protein homogenized with 5.07 mg ^{14}C-labeled glycine of specific activity $a_1 = (96.2 \pm 1.2)$ counts min^{-1} mg^{-1} (relative activities are sufficient). After separation of a portion of glycine from the mixture, specific activity of $a_2 = (51.3 \pm 0.9)$ counts min^{-1} mg^{-1}. Equation (7) gives:

$$m_x = 5.07\,\mathrm{mg} \left(\frac{96.2}{51.3} - 1\right) = (4.44 \pm 0.17)\ \mathrm{mg}$$

The percentage of glycine in the protein is:

$$\frac{4.44}{152.6} \cdot 100 = (2.91 \pm 0.11)\,\%$$

4.2. Reverse Isotope Dilution Analysis

Another application of IDA is the determination of the activity A_y of a radioactive substance y from a complex mixture of radioactive substances by adding a known mass m_c of an inactive carrier substance. This technique is also

based on Equation (7), reversely solved for activity A_1. Homogenization and separation of a portion m_2 (activity A_2, specific activity a_2) gives:

$$A_y = a_2\,(m_c + m_y) \tag{9}$$

from the specific activity a_y of substance y. If the inactive carrier is used in significant excess (i.e., $m_c >> m_y$) the separation yield is:

$$A_y = A_2\,\frac{m_c}{m_2} \tag{10}$$

To obtain m_y, the specific activity a_y is required: $m_y = A_y/a_y$.

Reverse IDA (indirect IDA, or dilution with inactive isotopes) is particularly important in organic analysis and biochemistry [24,32] to test for radiochemical purity and stability of labeled compounds. It is often used in radiochemical nuclear AA for the separation of activated element traces from a variety of interfering radionuclides in the analytical sample, for determination of the isolated activity, in comparison with that of a reference sample (i.e., to determine the radiochemical yield according to Eq. 10).

In contrast to direct IDA, reverse IDA is not limited by the sensitivity of the analytical determination of the isolated mass m_2, but merely by the specific activity a_y of the radioactive substance. By adjusting the carrier mass m_c, the sensitivity and accuracy can be increased, so reverse IDA is suitable for trace and microanalysis. Losses by absorption from carrier-free solutions must be taken into account. In separations from mixtures of radioactive substances, high-quality purification is important.

Double (Multiple) IDA. Since the specific activity a_y is often unknown, double IDA can be used to determine m_y. Two equal aliquots containing the same, but unknown mass m_y are diluted with different known carrier masses, m_c and m'_c. After separation of portions m_2, m'_2 and measurement of specific activities a_2, a'_2 one obtains from Equation (9):

$$a_2 = \frac{A_y}{m_c + m_y} = \frac{a_y m_y}{m_c + m_y} \tag{11}$$

$$a'_2 = \frac{a_y m_y}{m'_c + m_y} \tag{12}$$

Thus:

$$m_y = \frac{a'_2 m'_c - a_2 m_c}{a_2 - a'_2} \tag{13}$$

$$a_y = \frac{(m_c - m'_c)\,a_2 a'_2}{a_2 m_c - a'_2 m'_c} \tag{14}$$

Multiple (direct and reverse) IDA produces a calibration graph $m_x = f\,(1/a_2)$.

4.3. Derivative Isotope Dilution Analysis

If the preparation of labeled substances is too difficult, derivative IDA may be applied. This technique combines a radioreagent and IDA, preferably of low selectivity, to react with one or more analytes. It is mainly used with mixtures of complex organic compounds. The basic stages are:

1) Reaction of analyte A with radioreagent B∗, of known specific activity, to form a radioactive product AB∗. Excess B∗ may be removed in an optional purification step:

$$\mathrm{A} + \mathrm{B}* \longrightarrow \mathrm{AB}* + \mathrm{B}*_{\mathrm{ex}} \tag{15}$$

2) The mass of AB∗ is determined by mixture with a known mass of inactive AB by reverse IDA:

$$\mathrm{AB}* + \mathrm{B}*_{\mathrm{ex}} + \mathrm{AB} \longrightarrow (\mathrm{AB}* + \mathrm{AB}) + \mathrm{B}*_{\mathrm{ex}} \tag{16}$$

3) Separation of diluted product (AB∗ + AB) and determination of its specific activity.

The principal requirements for this technique are:

1) Reaction (Eq. 15) should be quantitative or of known yield.
2) Isotope exchange between substances AB and B∗ must be prevented. Since inorganic substances are quite amenable to such exchange, application of derivative IDA to inorganic analytes is impossible.
3) Purification from excess reagent must be quantitative to avoid significant bias, even from minor impurities.

As a variant, dilution with A∗ is applied prior to the first reaction (Eq. 15) with radioreagent B∗.

Determination of the reaction yield and analysis of inorganic substances are possible with this variant technique [2]. Its main importance is in the determination of trace amounts in complex mixtures. It is routinely used in biochemistry and physiological chemistry [2,31,32,75]. The simpler approaches of radioimmunoassays (→ Diagnostic Reagents, Chap. 3.) are generally preferred.

Further variants are based on combinations with double derivative IDA, or addition of an inactive derivative prior to direct IDA.

4.4. Substoichiometric Isotope Dilution Analysis

Utilization of conventional IDA for trace determinations is limited by the necessity to isolate macroscopic amounts of the analyte substance for determination of m_1 and m_2 by physicochemical methods. The principle of substoichiometry developed by Ružička and Starý [57] avoids this limitation.

Exactly equal (low) masses m_1 and m_2 are isolated from the spike solution of initial specific activity $a_1 = A_1/m_1$ and from the solution after homogenization with the analyte ($a_2 = A_2/m_2$). Equation (7) reduces to:

$$m_x = m_1 \left(\frac{A_1}{A_2} - 1 \right) \tag{17}$$

The masses m_1 and m_2 are isolated using the same amount of the separation reagent, which must be stoichiometrically insufficient and consumed quantitatively (or to exactly the same amount) in this reaction. It must form a product which can easily be separated from the excess of unreacted substance. Separation reagents should have high chemical stability, no tendency to adsorption at very low concentrations, and sufficient selectivity for the analyte. Separations with enrichment factors of $10^4 - 10^5$ are typical. Since the mass m_1 of the added radioactive substance must be accurately known, reverse IDA can be applied.

Error propagation is comprehensively discussed by Tölgyessy and Kyrš [2]. At trace levels below 1 µg, standard deviations of ca. 3 % and accuracies of ca. 6 % are typical; precision and accuracy can be 1.5 % or better at higher concentrations.

4.4.1. Substoichiometric Separation by Liquid – Liquid Distribution

The extraction of metal chelates is suitable for separating equal masses of metals from solutions of different concentrations. Extraction of a metal chelate MA from a metal M^{n+} by an organic reagent HA is generally described by [47, 57,58]:

$$M^{n+} + n\,(HA)_{org} \rightleftharpoons (MA_n)_{org} + nH^+ \tag{18}$$

with the extraction constant K:

$$K = \frac{[MA_n]_{org}\,[H]^n}{[M^{n+}]\,[HA]^n_{org}} \tag{19}$$

where subscript org denotes the organic phase. In trace element analysis by substoichiometric IDA, concentrations of $10^{-6} - 10^{-9}$ mol/L are typically extracted from aqueous samples of volumes V in the range 10 mL to 100 µL into organic phases of smaller volumes (typically one-tenth of the sample volume) of $10^{-5} - 10^{-8}$ mol/L solutions of chelating reagent.

The dependence of Equation (18) on the pH value of the solution makes it necessary to keep within an optimum pH range. The substoichiometric principle requires a higher than 99 % consumption of the initial concentration c_{iHA} in Equation (18), so that $[HA]_{org} < 0.01 c_{iHA}$. Thus the solution must satisfy the condition:

$$pH > -\log\,(0.01 c_{iHA}) - \frac{1}{n}\log K \tag{20}$$

With increasing pH, dissociation of acid HA (dissociation constant K_{HA}) in the aqueous phase increases. This dissociation is negligible at:

$$pH < pK_{HA} + \log q_{HA} + \log \frac{V_{org}}{V} \tag{21}$$

where q_{HA} is the distribution coefficient of the reagent, V is the sample volume, and V_{org} is the volume of the organic phase. Hydrolysis of the metal at increasing pH must not interfere with the separation.

The reagent (e.g., diphenylthiocarbazone, diethyldithiocarbamic acid (2-thenoxy)-3,3,3-trifluoroacetone, 8-hydroxyquinoline) must be stable to decomposition, even at very low concentrations in acid and neutral media, and must form an extractable chelate with a sufficiently high value of K (for further useful ligands see [38,40, 58,83]). Optimum conditions for substoichiometric extraction are calculated from K, K_{HA}, and q_{HA}. Selectivity of the separation can be improved by use of additional masking reagents [54,57].

Substoichiometric extraction of ion-association complexes is generally restricted to higher concentration levels (e.g., determination of Na, P). Water-soluble complexes of more than 20 metals with ethylenediaminotetraacetic acid can be used for substoichiometric determination at concentrations below 1 ng/g, provided that the excess of unreacted metal is separated by liquid–liquid distribution, ion-exchange, carrier coprecipitation, or other suitable means. Selectivity is high and can be further increased by masking reagents.

4.4.2. Redox Substoichiometry

Oxidation or reduction by a substoichiometric mass of reagent is followed by separation, extraction, coprecipitation, etc. [73]. $KMnO_4$, $K_2Cr_2O_7$, and $FeSO_4$ have been used as substoichiometric redox reagents.

4.4.3. Displacement Substoichiometry

Complete extraction of an analyte metal M_1 with excess complexing reagent is followed by removal of the excess and subsequent displacement of M_1 by a substoichiometric amount of another metal M_2, provided that M_2 has a sufficiently higher extraction constant K_2. Further variants are discussed elsewhere [2].

4.4.4. Applications

Applications include trace element determination in rocks, soils, biogenic samples, and pure metals by direct IDA, i.e., added radiotracer [2, 54,60]. Another field of increasing importance is in radiochemical AA for assessment and improvement of accuracy at trace and ultra-trace levels by determination of the radiochemical yield. This applies primarily to biomedical samples, high-purity materials, and environmental materials [7–9,32,84].

4.5. Sub- and Superequivalence Method

The sub- and superequivalence method is based on the isoconcentration principle. Three versions of this method exist: the direct method for the determination of nonradioactive substances [99], the reverse method for the determination of radioactive substances [100], and the universal isotope dilution method for determination of both [101].

Historically, the *reverse method* has been developed first [100]. In the reverse method, two series of aliquots are formed from a solution of the radioactive (labeled) analyte. The first series is formed by equal volumes of the analyte solution [containing $x = (x,..,x,...,x)$ amounts of labeled analyte] while the (identical) aliquots of the second series contain ξ times higher volumes of the same solution ($x = \xi x,..., \xi x,..., \xi x$). The members of the first series are isotopically diluted by successively adding fixed amounts of the same substance in inactive form $y = (y_1,...,y_\xi,...,y_a)$, and so the members of the first series contain $(x+y) = (x+y_1, ..., x+y_\xi,..., x+y_a)$ amounts of the substance. After establishing the same separation conditions for all aliquots of both series, e.g., by adding masking and buffer solutions, and after dilution to the same volume, a fixed amount of reagent is added to all aliquots. Then the separation is performed at the same temperature and pressure. The added amount of reagent is fixed well in the aliquot containing the highest total concentration of analyte (substoichiometric regime), but it is in excess in the aliquot containing the lowest total amount of substance. The ratio of the activities of the isolated fractions from the second series (which are theoretically identical) and the activities of the isolated fractions of particular members of the first series are plotted as a function of the dilution increment and the mass of analyte is determined from this plot [2,73,89].

In the *direct method* two series of aliquots are formed from a solution of the labeled counter-

part of the substance to be analyzed (which is not radioactive). The first series is formed by increasing volumes of this solution containing ξy_1, ξy_2, $\ldots \xi y_a$ amounts of the labeled substance. The second series is formed by increasing, but ξ times lower amounts of the same solution. To the aliquots of the second series constant amounts of the substance to be analyzed (in nonradioactive form) are added. After establishing the same separation conditions for all aliquots of both series, the separation is performed. The evaluation of the analytical results is identical to that in the reverse method.

The *universal method* (two partially different procedures) is (are) the combination of both methods.

Besides a large variety of metals, several chemical species of specific interest, e.g., cyclic 3′,5-adenosine monophosphate, thyroxine and 3,5,3′-trioiodothyronine have been determined by sub- and superequivalence IDA and the results show higher accuracy of the method over that of radioimmunoassay [102, 103]. The principle of the method also allows the determination of certain physicochemical characteristics, e.g., estimation of the stability constants of complexes and the solubility of precipitates.

The method is a calibration curve approach. It does not require that all the separated amounts of the aliquots mentioned are exactly the same, as in substoichiometric IDA (Section 4.4). This is the reason why separation reactions can be used that are not quantitative, a sufficient prerequisite is reproducibility [2, 24, 74].

An alternative calibration curve approach for nonquantitative reactions is to determine the distribution coefficient by a concentration-dependent distribution method [2, 75], which is a variant of the radioreagent method (Section 5.1).

5. Radioreagent Methods

Radioreagent methods (RRM) are based on the use of a radioactive species in a quantitative reaction, and measurement of the change in activity of that species in the course of the reaction. The radioactive species may be a labeled reagent, the analyte, or a substance able to undergo an exchange reaction with some compound of the analyte.

After separation from excess reagent (by liquid-liquid distribution, chromatography, precipitation, etc.), the mass or concentration of this product is determined from activity measurement. The determination is based on a radioactive substance chemically different from the analyte substance (in contrast to IDA), therefore the chemical reaction is of prime importance. By variation of this key reaction, the principle can be adapted to various procedures. The superiority of radioreagent methods over classical separation techniques arises from the use of an inactive carrier and the high sensitivity of the activity measurements, which are not subject to interference by the carrier or other substances [1, 16, 24, 34, 60, 87, 92, 94, 95, 97].

Depending on the type of chemical interaction and relationship of the analyte to the radioactive substance measured, RRM can be divided into three basic group:

1) The radioactive substance is a typical reagent which is able to react with the compound to be determined.
2) The method is based on "nonequivalent" competition; i.e., the radioactive substance is able to exchange with the analyte.
3) The method is based on "equivalent" competition; i.e., the labeled substance is chemically identical with the substance to be determined.

From the practical point of view, it is convenient to divide RRM into the following groups:

1) Simple RRM in which the reactions take place quantitatively, i.e., either the analyte or the reagent is completely consumed and compounds of definite composition are formed.
2) Methods of concentration-dependent distribution utilizing reactions where products of unstable composition are formed, but in which the extent of reaction is determined by the corresponding equilibrium constant or reaction time.
3) Isotope exchange methods based on the exchange of isotopes between two different compounds of one element, with a radioactive isotope in one of the compounds and a nonradioactive isotope in the other, usually the analyte.

4) Radiorelease methods comprising procedures in which the analyte reacts with a radioactive reagent, thus releasing an aliquot of the reagent activity, in most cases into the gas phase.
5) Radiometric titration usually involves use of radioactive reagents to determine the equivalence point, but there are variants that do not use a radioreagent; instead they are based on a change in the intensity of radiation caused by its absorption or scattering in a medium containing the analyte [2,32,87,92,95–97].

5.1. Simple Radioreagent Methods

Simple radioreagent methods (SRRM) utilize a reaction with a suitable reagent, and sometimes with a third substance; one of the reagents is radioactive and the reaction is quantitative [2,32, 87,95]. SRRM can be classified into the following three groups:

1) Determination with labeled reagents,
2) Determination with labeled analyte,
3) Determination with labeled competing substance

5.1.1. Determination with Labeled Reagents

An excess of radioactive reagent solution of known analytical concentration is usually used. The active product is separated from the excess of radioactive reagents by precipitation, formation of extractable chelates, or sorption.

The substance to be determined (n_x mol) reacts with a radioactive reagent (n_R mol), forming a precipitate, extractable compound, or other separable substance. After the separation of the product, the radioactivity of the product (A_P), that of the excess of unreacted reagent (A_E), or both are measured. The following relationship is valid:

$$\frac{A_R}{n_R} = \frac{A_P}{zn_x} = \frac{A_E}{n_{R-zn_x}}$$

Thus

$$n_x = \frac{n_R A_P}{z A_R} = \frac{n_R}{z}\left(1 - \frac{A_E}{A_R}\right)$$

where A_R is the radioactivity of the reagent and z denotes the stoichiometric ratio in the compound formed (the number of moles of the reagent interacting with 1 mol of the test substance). If n is not exactly known, a calibration curve approach can be used. Quantitative isolation of trace amounts of the reaction products poses another important problem (Table 3, [2,41,47, 60]).

In this way, the concentration of chloride in water was determined by precipitation with silver labeled with ^{110m}Ag ;^{14}CO was used for the determination of hemoglobin in blood; and ^{131}I for the determination of bismuth in the presence of an excess of iodide ions using extraction separation [2].

5.1.2. Determination with Labeled Analyte

The first step involves labeling of the analyte. To this end, the radioactive indicator is added to the unknown sample and isotope exchange is allowed to take place; then the reagent is added in a substoichiometric but known quantity. After the reaction is complete, the phases are separated and their respective activities determined. In those cases where the activity of the equilibrium solution is measured, the unknown concentration can be calculated from:

$$x + x_0 = \frac{A_{\text{init}} B}{(A_{\text{init}} - A_{\text{equilib}})}$$

where x is the quantity of analyte (equivalents), x_o the amount of added radioactive indicator (equivalents), B the amount of unlabeled reagent (equivalents), and A_{init} and A_{equilib} represent the total activity of the initial and equilibrium solutions, respectively. In this variant of the RRM, solvent extraction is frequently used.

5.1.3. Determination with Labeled Competing Substances

Analyte M_2 competes with the radioactively labeled substance $^*M_1$ in the formation of a compound with a reagent R, where the compound precipitates from the solution

$$^*M_1R\downarrow + M_2 \rightleftharpoons M_2R\downarrow + {}^*M_1$$

Table 3. Useful concentration regime of the radioreagent method

Analyte	Radionuclide	Labeled reagent	Separation	Concentration regime
H^+	^{131}I	KI	extr. I_2 into pyridine/chloroform	>0.1 µg/mL
H_3BO_3	^{18}F	HF	extr. HBF_4 into 1,2-dichloroethane[b]	≈ 0.1 µg B
NaLS[e]	^{59}Fe	$[Fe(II)(phen)_3]$[c]	extr. $Fe(phen)_3(LS)_2$ into $ChCl_3$	5 – 300 ng/mL Na
Cl^-, Br^-, I^-	^{203}Hg	$C_6H_5Hg^+$	extr. C_6H_5HgCl into benzene	0.5 – 15 mg/mL
Cl^-.I^-	^{110m}Ag	$AgNO_3$	pptn. AgCl; AgI	0.4 – 250 µg/mL
Cl^-	^{203}Hg	$HgNO_3$	pptn. Hg_2Cl_2	0.8 – 13 µg/mL
Bi	^{131}I	KI[d]	extr. $HBiI_4$ into *n*-butylacetate	0.04 – 4 µg
Cationic detergents	^{131}I	Rose Bengal	extr. ion-associate into $CHCl_3$	

[a] Extr. = solvent extraction; pptn. = precipitation. [b] In presence of methylene blue. [c] LS = lauryl sulfate; phen = 1,10 phenanthroline.
[d] In presence of ascorbic acid + Na_2SO_3 in 0.6 mol/L H_2SO_4.

$^*M_1R$ and M_2R occur in a different phase than either substance M_1 or M_2. The equilibrium constant of the reaction, as a rule is >1 so that a significant change takes place.

This method is used when the labeling of reagent R cannot be applied or when the reagent and its compound with the analyte are transferred into the same phase. Generally, a relatively soluble precipitate with one radioactively labeled component (e.g., $^{45}CaCO_3$) is brought into contact with a solution containing the analyte (Pb^{2+}), which displaces the radioactive component ($^{45}Ca^{2+}$) from the precipitate into the solution, and a less soluble precipitate ($PbCO_3$) is formed. The resulting radioactivity of the solution is proportional to the initial amount of lead.

5.2. Method of Concentration-Dependent Distribution

Concentration-dependent distribution (CDD) is based on the utilization of a calibration graph that shows the dependence of the distribution ratio of the substance to be determined in a two-phase system on the total concentration of the substance [2, 32, 87, 92, 95, 96]. Characteristic features of the method of CDD are as follows:

- The analysis is based on the distribution of a radioactive substance between two phases or between several parts of the systems used (paper chromatography, electrophoresis),
- The ratio of the activities in both phases or parts of the system depends strongly on the initial concentration of the unknown substance,
- The distribution is not given by the stoichiometric ratio of the reacting substances, i.e., not by total saturation of one phase or a nearly complete consumption of the reagent, but is given exclusively by the corresponding equilibrium or kinetic constants.

Example. Determination of chloride in water samples. The sample is shaken with an excess of labeled AgCl precipitate and the radioactivity of silver is measured in the equilibrium solution. The higher the chloride concentration the lower is the concentration of silver in solution (owing to the constancy of the AgCl solubility product). The Cl^- ions determined do not react with the AgCl precipitate, they only suppress its dissolution. The amount of chloride in the precipitate has no direct relationship to the concentration of Cl^- ions in the sample solution.

In CCD is useful to define two extreme types of determination: *saturation analysis* and *nonsaturation analysis*.

In *saturation analysis*, the cause of the change in the distribution of the radioactive substance between the phases is an increase in the total saturation of the reagent by the given substance. This term originally appeared in the literature in connection with the determination of biochemically important substances (steroids, hormones, vitamins) and is equivalent to the term radioimmunoassay (→ Diagnostic Reagents, Chap. 3.2.).

In *nonsaturation analysis*, the cause of the change in the distribution coefficient is the shift of the chemical equilibrium due to an increase in the concentration of the substance to be determined, irrespective of the extent of saturation. The determination of some extractants can serve as a typical example. The number of procedures described that use nonsaturation analysis is limited.

5.3. Isotope Exchange Methods

Isotope exchange methods (IEM) are based on the exchange of isotopes between two different compounds of the element X, one of which (XA) is nonradioactive, the other (XB) being labeled with a radioactive isotope [2,61,87,97]. After isotopic equilibrium is reached

$$XA + {}^{*}XB \rightleftharpoons {}^{*}XA + XB$$

the specific activities of the element X in both compounds are equal

$$A_1/m = A_2/m_x$$

where A_1 and A_2 are the equilibrium radioactivities of XB and XA, respectively, and m_x and m are the amounts of X in XA and XB, respectively. The value of m_x can be computed from:

$$m_x = mA_2/A_1$$

or from the calibration graph of $m_x = f(A_2/A_1)$, $m_x = f(A_2)$ or $m_x = f(A_1)$. Isotopic exchange can be carried out in either a heterogeneous or a homogeneous system.

Example. Methylmercury and phenylmercury are determined down to 10 pg/mL by liquid-liquid extraction from 3 mol/L aqueous HCl into benzene, and addition of $K^{131}I$ to the separated organic phase. Inactive chloride is completely displaced by ^{131}I:

$$(CH_3HgCl)_{org} + {}^{131}I^- \longrightarrow (CH_3Hg^{131}I)_{org} + Cl^-$$

The method has been used for determination of methylmercury species in fish and in drinking water [61].

5.4. Radioimmunoassay

Radioimmunoassay is the widely used radioreagent method (→ Diagnostic Reagents, Chap. 3.2.).

5.5. Radiorelease Methods

In radiorelease methods the analyte substance A reacts with a radioactive reagent, so that radioactive R^* is released into a second phase, without being replaced by an inactive analyte [62]. Applications involve either release of radioreagents from solids or liquids into the gas phase, or release from solids into a liquid. Radiorelease methods may be classified according to the type of the radioactive reagents employed, i.e. (1) radioactive kryptonates, (2) radioactive metals, and (3) radioactive salts and other substances.

5.5.1. Radioactive Kryptonates [2,87,92,94,95]

The term radioactive kryptonates is used for substances into which atoms or ions of krypton-85 (^{85}Kr) are incorporated (by diffusion of ^{85}Kr; by bombardment with accelerated krypton ions; by crystallization of the kryptonated substances from a melt, or by placing the solution in an atmosphere of ^{85}Kr, etc.). The radionuclide can be released from the solid lattice by any chemical or physical reaction that breaks down the lattice at the solid surface. The released inert gas is conveniently measured from its main β radiation (0.67 MeV). The half life of 10.27 a is suitable for long-term remote sensing devices.

Applications of Radioactive Kryptonates [32,92,94–96]. The determination of *oxygen* is performed by surface oxidation of copper or pyrographite kryptonate, at elevated temperature, resulting in destruction of the surface layer and release of ^{85}Kr proportional to the oxygen mass. Detection limits are at the 10 ng/m^3 level. Ozone oxidizes copper kryptonate at temperatures below 100 °C, whereas reaction with oxygen starts well above 200 °C, so this detector can detect O_3 and O_2 differentially. Determination of *ozone* in air is feasible over a concentration range of 10^{-7} – 10^{-3} g/m^3 with hydroquinone kryptonate:

$$[C_6H_4(OH)_2]_3[{}^{85}Kr] + O_3 \longrightarrow 3\,C_6H_4O_2 + 3\,H_2O + {}^{85}Kr\uparrow$$

Sulfur dioxide has also been determined by a method based on the mechanism of double release. In the first stage, sulfur dioxide reacts with sodium chlorate to release chlorine dioxide which is a strong oxidizing agent. The chlorine dioxide then oxidizes radioactive hydroquinone kryptonate and gaseous ^{85}Kr is released. The following reactions are involved:

$$SO_2 + 2\,NaClO_3 \longrightarrow 2\,ClO_2 + Na_2SO_4$$
$$ClO_2 + [C_6H_4(OH)_2]_3[^{85}Kr] \longrightarrow {}^{85}Kr\uparrow$$

A radioactive kryptonate of silica has been suggested for the determination of *hydrogen fluoride* in air. *Oxygen dissolved in water* can be measured by the use of thallium kryptonate

$$4\,Tl[^{85}Kr] + O_2 + 2\,H_2O \longrightarrow Tl^+ + 4\,OH^- + {}^{85}Kr\uparrow$$

Thallium is oxidized by oxygen, and the amount of ^{85}Kr released or the decrease in the activity of kryptonated thallium is proportional to the dissolved oxygen concentration in the sample (at constant pH) down to 0.3 µg/mL.

5.5.2. Radioactive Metals

Oxidizing agents in solution can react at the surface of a labeled metal, releasing radioactive ions into the solutions which are used for determination of the oxidants. The decisive factors in the choice of the metal are the following: the metal should not react with water, but with oxidizing agents to yield ions which do not form precipitates in aqueous media. The metal should have a radionuclide with suitable nuclear properties. These conditions are met by thallium (^{204}Tl) and silver (^{110m}Ag) [54, 60, 61].

Dissolved *oxygen* has been determined in seawater, drinking water, and wastewater down to the ng/g level with metallic thallium labeled with ^{204}Tl. Selectivity requires removal or masking of other oxidizing agents. The *vanadate* ion is assayed by acidifying the corresponding sample (pH ≈ 3) and passing it over a column containing radioactive metallic ^{110m}Ag. The labeled silver is oxidized, dissolved, eluted from the column, and detected. The measurement of *dichromate* ion concentration in natural waters can be carried out similarly.

5.5.3. Radioactive Salts and Other Radioactive Substances [2, 32, 92, 96]

Sulfur dioxid is determined on the basis of the reaction:

$$5\,SO_2 + 2\,K^{131}IO_3 + 4\,H_2O \longrightarrow K_2SO_4 + 4\,H_2SO_4 + {}^{131}I_2\uparrow$$

This reaction takes place in an alkaline solution through which air containing sulfur dioxide is bubbled. After completion of the reaction, the solution is acidified and the iodine released is extracted. *Active hydrogen* in organic substances may be determined by reaction with lithium aluminum hydride labeled with tritium (^3H). The activity of released tritium is measured using a proportional counter.

A number of determinations are based on the formation of a soluble complex between the analyte in solution with a radioactively labeled precipitate. In this way it is possible to determine anions forming soluble complexes (e.g., CN^-, $S_2O_3^-$, I^-, F^-). The principle of the determination are given in the following equations:

$$2\,CN^- + {}^{203}Hg(IO_3)_2\downarrow \longrightarrow {}^{203}Hg(CN)_2 + 2\,IO_3^-$$
$$(> 0.50\,\mu g/mL, \pm 55\,\%)$$

$$2\,Cl^- + {}^{203}Hg(IO_3)_2\downarrow \longrightarrow {}^{203}HgCl_2 + 2\,IO_3^-$$
$$(> 1\,\mu g/mL, \pm 5\,\%)$$

$$n\,S_2O_3^{2-} + m\,{}^{110m}AgSCN\downarrow$$
$$\longrightarrow {}^{110m}Ag_m\,(S_2O_3)_2^{(2n-m)-} + m\,SCN^-$$

$$2\,I^- + {}^{203}HgI_2\downarrow \longrightarrow {}^{203}HgI_4^{2-}\ (> 25\,\mu g/mL, \pm 5\,\%)$$

5.6. Radiometric Titration

Radiometric titrations follow the relation between the radioactivity of one component or phase of the solution under analysis and the volume of added titrant. The compound formed during the titration must be easily separable from the excess of unreacted ions. This separation is directly ensured only in the case of precipitation reactions. In other types of reactions, the separation can be accomplished using an additional procedure. The endpoint is determined from the change in the activity of the residual solution or of the other phase.

According to the type of chemical reaction used, methods based on the formation of precipitates and methods based on complex formation can be distinguished. Because of the necessity for handling precipitates, precipitation radiometric titrations are difficult to apply to less than milligram amounts and, therefore, have no special advantages over other volumetric methods. The sensitivity of complexometric titrations

is limited by the sensitivity of the determination of the endpoint. However, the use of radiometric detection can substantially increase the sensitivity of this type of determination. For the separation of the product from the initial component, liquid-liquid distribution, ion-exchange, electrophoresis, or paper chromatography are most often used [2, 63, 88, 93, 97].

This application of radiometric titrations has declined over the past three decades [64]. Their main advantage is where classical methods for detection of the endpoint are either impossible or subject to interference from the titration medium.

Radiopolarography offers highly increased sensitivity and selectivity over polarographic current measurement, without interference from major components of the solution. It measures the amount of labeled ions deposited in single drops in a dropping mercury electrode as a function of potential [85].

6. References

General References

1. IUPAC Commission on Radiochemistry, Nuclear Techniques, Anal. Chem. Div., IUPAC, *Pure Appl. Chem.* **63** (1991) 1269 – 1306.
2. J. Tölgyessy, M. Kyrš: *Radioanalytical Chemistry,* **vol. 1,** Ellis Horwood, Chichester 1989, p. 354 ff.
3. J. Tölgyessy, M. Kyrš: *Radioanalytical Chemistry,* **vol. 2,** Ellis Horwood, Chichester 1989, p. 498 ff.
4. H. A. Das, A. Faanhof, H. A. van der Sloot: "Radioanalysis in Geochemistry," *Dev. Geochem.* **5** (1989) 482 pp.
5. P. Bode, H. T. Wolterbeek, *J. Trace Microprobe Tech.* **81** (1990) 121 – 138.
6. D. De Soete, R. Gijbels, J. Hoste: *Neutron Activation Analysis,* Wiley-Interscience, London 1972.
7. W. D. Ehmann, D. E. Vance, *CRC Crit. Rev. in Anal. Chem.* **20** (1989) 405 – 443.
8. N. M. Spyrou, *J. Trace Microprobe Tech.* **6** (1988) 603 – 619.
9. J. Hoste, F. De Corte, W. Maenhaut in P. J. Elving, V. Krivan, E. Kolthoff (eds.): *Treatise on Analytical Chemistry,* 2nd ed., **part I, vol. 14,** J. Wiley & Sons, New York 1986, p. 645.
10. V. Krivan in W. Fresenius et al. (eds.): *Analytiker Taschenbuch,* **vol. 5,** Springer-Verlag, Berlin 1985, p. 36.
11. Y. Suzuki, *Radioisotopes* **36** (1987) 414 – 421.
12. C. G. Clayton, M. R. Wormald, J. S. Schweitzer in IAEA (ed.): *Proc. Int. Sympos. Nucl. Techn. Explor. Exploit, Energy Miner. Resourc.,* IAEA, Vienna 1991, p. 107.
13. G. F. Knoll: *Radiation Detection and Measurement,* 2nd ed., J. Wiley, New York 1989, p. 754.
14. M. F. L'Annunziata: *Radionuclide Tracers: Their Detection and Measurement,* Academic Press, New York 1987, p. 505.
15. L. I. Wiebe, *Radiat. Phys. Chem.* **24** (1984) 365 – 372.
16. H. Filthuth, *Chem. Anal. (N.Y.)* **108** (1990) 167 – 183.
17. H. J. M. Bowen: "Nuclear Activation and Radioisotopic Methods of Analysis," in P. J. Elving, V. Krivan, E. Kolthoff (eds.): *Treatise on Analytical Chemistry,* 2nd ed., part I, Sect. K, vol. 14, J. Wiley & Sons, New York 1986, p. 193.
18. G. Tölg, *Analyst (London)* **112** (1987) 365 – 376.
19. J. A. C. Broekaert, G. Tölg, *Fresenius' Z. Anal. Chem.* **326** (1987) 495 – 509.
20. V. Krivan: "Activation and Radioisotopic Methods of Analysis," in P. J. Elving, V. Krivan, I. M. Kolthoff (eds.): *Treatise on Analytical Chemistry,* part I, Sect. K, 2nd ed., vol. 14, J. Wiley & Sons, New York 1986, p. 340.
21. M. Linscheid, *Nachr. Chem. Tech. Lab.* **39** (1991) 132 – 137.
22. W. R. Wolf, M. Stoeppler: "Proc. 4th Int. Symp. Biol. Environm. Reference Mater.," *Fresenius' J. Anal. Chem.* **338** (1990) 359 – 581.
23. L. L. Jackson et al., *Anal. Chem.* **63** (1991) 33 R –48 R.
24. J. W. Ferkany, *Life Sci.* **41** (1987) 881 – 884.
25. W. D. Ehmann, J. D. Robertson, St. W. Yates, *Anal. Chem.* **62** (1990) 50 R – 70 R; **64** (1992) 1 R – 22 R.
26. R. Naeumann, E. Steinnes, V. P. Guinn, *J. radioanal. Nucl. Chem.* **168** (1993) 61 – 68.
27. R. R. Greenberg, *J. Radioanal. Nucl. Chem.* **113** (1987) 233 – 247.
28. V. P. Guinn, H. S. Hsia, N. L. Turglio, *J. Radioanal. Nucl. Chem.* **123** (1988) 249 – 257.
29. L. Grodzins, *Nucl. Instrum. Methods Phys. Res. Sect. B* **B 56 – B 57** (1991) 829 – 833.

30. F. De Corte, A. Simonits, *J. Radioanal. Nucl. Chem.* **133** (1989) 43 – 130.
31. J. Tölgyessy, T. Braun, M. Kyrš: *Isotope Dilution Analysis,* Pergamon Press, Oxford 1972.
32. J. Tölgyessy, E. H. Klehr: *Nuclear Environmental Chemical Analysis,* Ellis Horwood, Chichester 1987.
33. V. Krivan, *Sci. Total Environ.* **64** (1987) 21 – 40.
34. J. R. Jones (ed.): *Isotopes – Essential Chemistry and Applications,* 2nd ed., Royal Society of Chemistry, CRC Press, Boca Raton 1988, p. 270.
35. G. R. Choppin, J. Rydberg: *Nuclear Chemistry: Theory and Applications,* Pergamon Press, Oxford 1980, p. 75.
36. J. Mandel in I. M. Kolthoff, P. J. Elving (eds.): *Treatise on Analytical Chemistry,* **part I, vol. 1,** J. Wiley & Sons, New York 1978, p. 243.
37. C. T. Peng: *Sample Preparation for Liquid Scintillation Counting,* Amersham International, United Kingdom, 1977.
38. K. Robards, P. Starr, E. Patsalides, *Analyst (London)* **116** (1991) 1247 – 1273.
39. A. W. Rogers: *Techniques of Autoradiography,* 3rd ed., Elsevier, Amsterdam 1979.
40. A. Mizuike: *Enrichment Techniques for Inorganic Trace Analysis,* Springer-Verlag, Berlin 1983, p. 144.
41. A. Dyer, *Analyst (London)* **114** (1989) 265 – 267.
42. V. Krivan, *Talanta* **29** (1982) 1041 – 1050.
43. Yu. A. Zolotow, M. Grasserbauer, *Pure Appl. Chem.* **57** (1985) 1133 – 1152.
44. J. Versieck, R. Cornelis: *Trace Elements in Human Plasma or Serum,* CRC Press, Boca Raton 1989, p. 23.
45. P. Tschöpel, G. Tölg, *J. Trace Microprobe Tech.* **1** (1982) 1 – 77.
46. R. Cornelis, *Mikrochim. Acta* 1991 (1991) no. III, 37 – 44.
47. *Ullmann,* 4th ed., **5,** 709.
48. L. Melander, W. H. Saunders: *Reaction Rates of Isotopic Molecules,* Wiley-Interscience, New York 1980.
49. K. H. Lieser: "Activation and Radioisotopic Methods of Analysis," in P. J. Elving, V. Krivan, I. M. Kolthoff (eds.): *Treatise on Analytical Chemistry,* part I, Sect. K, 2nd ed., vol. 14, J. Wiley & Sons, New York 1986, p. 1.
50. W. Schmid, V. Krivan, *Anal. Chem.* **57** (1985) 30 – 34.
51. R. Cornelis, J. Versieck: "Activation and Radioisotopic Methods of Analysis," in P. J. Elving, V. Krivan, I. M. Kolthoff (eds.): *Treatise on Analytical Chemistry,* part I, Sect. K, 2nd ed., vol. 14, J. Wiley & Sons, New York 1986, p. 665.
52. *Ullmann,* 4th ed., **5,** 1.
53. G. Tölg et al., *Pure Appl. Chem.* **60** (1988) 1417 –1424.
54. J. Starý, J. Ružička in G. Svehla (ed.): *Wilson and Wilson's Comprehensive Analytical Chemistry,* vol. 7, Elsevier, Amsterdam 1976.
55. D. M. Wieland, M. C. Tobes, T. J. Mangner: *Analytical and Chromatographic Techniques in Radiopharmaceutical Chemistry,* Springer Verlag, New York 1986, p. 300.
56. B. B. Damaskin, O. A. Petrii, V. V. Batrakov in R. Parsons (ed.): *Adsorption of Organic Compounds on Electrodes,* Plenum Press, New York 1971.
57. J. Ružička, J. Starý: *Substoichiometry in Radiochemical Analysis,* Pergamon Press, Oxford 1968.
58. F. Umland, A. Janssen, D. Thierig, G. Wünsch: *Theorie und praktische Anwendung von Komplexbildnern, Methoden der Analyse in der Chemie,* vol. 9, Akademische Verlagsgesellschaft, Frankfurt 1971, p. 759.
59. H. Yoshioka, T. Kambara, *Talanta* **31** (1984) 509 – 513.
60. J. Starý: "Activation and Radioisotopic Methods of Analysis," in P. J. Elving, V. Krivan, I. M. Kolthoff (eds.): *Treatise on Analytical Chemistry,* part I, Sect. K, 2nd ed., vol. 14, J. Wiley & Sons, New York 1986, p. 241.
61. J. Tölgyessy, S. Varga: *Nuclear Analytical Chemistry,* vol. 1 – 3, University Park Press, London 1972.
62. V. Balek, J. Tölgyessy: *Emanation Thermal Analysis and Other Radiometric Emanation Methods,* Elsevier, Amsterdam 1984.
63. T. Braun, J. Tölgyessy: *Radiometrische Titrationen,* Hirzel Verlag, Stuttgart 1968.
64. G. Kraft, J. Fischer: *Indikation von Titrationen,* De Gruyter, Berlin 1972, p. 304.
65. S. J. Parry: *Activation Spectrometry in Chemical Analysis,* Wiley & Sons, New York 1991, p. 243.
66. W. D. Ehrmann, D. E. Vance: *Radiochemistry and Nuclear Methods of Analyses,* Wiley & Sons, New York 1991, p. 531.

Specific References

67. G. Hevesy, F. Paneth, *Z. Anorg. Chem.* **82** (1913) 322–328.
68. O. Hahn, *Z. Phys. Chem. (Leipzig)* **103** (1923) 461.
69. Amersham International plc, Amersham Life Science Products 1993, Amersham/Little Chalfont, Amersham 1992, p. 184.
70. Du Pont de Nemours Biotechnol. Systems, NEN Research Products 1991, Bad Homburg 1991, p. 164.
71. J. T. Van Elteren, H. A. Das, C. L. De Ligny, J. Agterdenbos, *Anal. Chim. Acta* **222** (1989) 159–167.
72. V. Hodge, M. Stalland, M. Koide, E. D. Goldberg, *Anal. Chem.* **58** (1986) 616–620.
73. H. Yoshioka, K. Hasegawa, T. Kambara, *J. Radioanal. Nucl. Chem.* **117** (1987) 47–59.
74. J. Klas, Z. Korenova, J. Tölgyessy, *J. Radioanal. Nucl. Chem.* **102** (1986) 111–120; **109** (1987) 337–351.
75. S. Banerjee, *Anal. Chem.* **60** (1988) 1626–1629.
76. R. H. Clarke: The 1990 Recommendations of ICRP, Supplement to the Radiological Protection Bulletin No. 119, NRPB, Chilton 1991.
77. L. A. Currie, *Anal. Chem.* **40** (1968) 586–593.
78. K. Hoppstock, R. P. H. Garten, P. Tschöpel, G. Tölg, *Fresenius J. Anal. Chem.* **343** (1992) 778–781.
79. G. P. Westphal, *J. Trace Microprobe Tech.* **2** (1984/85) 217–235.
80. D. Gawlik, K. Berthold, F. Chisela, P. Brätter, *J. Radioanal. Nucl. Chem.* **112** (1987) 309–320.
81. W. Schmid, V. Krivan, *Anal. Chem.* **57** (1985) 30–34.
82. K. G. Heumann, *Fresenius Z. Anal. Chem.* **324** (1986) 601–611.
83. M. Schuster, *Fresenius J. Anal. Chem.* **342** (1992) 791–794.
84. K.-H. Thiemer, V. Krivan, *Anal. Chem.* **62** (1990) 2722–2727.
85. C. F. Miranda, R. Muxart, J. Vernois, G. Zuppiroli, *Radiochim. Acta* **19** (1973) 153.
86. K. Hoppstock, R. P. H. Garten, P. Tschöpel, G. Tölg, *Anal. Chim. Acta,* (1994) in press.
87. J. Tölgyessy, E. Bujdosó: *Handbook of Radioanalytical Chemistry,* **vols. 1–2,** CRC Press, Boca Raton 1991.
88. T. Braun, J. Tölgyessy: *Radiometric Titrations,* Pergamon Press, Oxford 1967.
89. J. Klas, J. Tölgyessy, J. Lesný: *Sub-superekvivalentová izotopová zried'ovacia analýza (Sub-superequivalent Isotope Dilution Analysis)* (in Slovak), Veda, Bratislava 1985.
90. J. Tölgyessy, T. Braun, M. Kyrš: *Analiz metodom izotopnogo razbavlenija (Isotope Dilution Analysis)* (in Russian), Atomizdat, Moscow 1975.
91. J. Klas, J. Tölgyessy, E. H. Klehr, *Radiochem. Radioanal. Lett.* **18** (1974) 83–88.
92. J. Tölgyessy, M. Harangozó: "Radio-reagent Methods," in *Encyclopedia of Analytical Science, Radiochemical Methods,* Academic Press, London 1995 pp. 4317–4323.
93. J. Tölgyessy: *Nuclear Radiation in Chemical Analysis* (in Hungarian), Müszaki Kiadó, Budapest 1965, p. 432.
94. J. Tölgyessy: *Radioaktivnye kriptonaty v nauke i technike (Radioactive Kryptonates in Science and Technology), in Buduschee nauki (The Future of Science)* (in Russian), Izdatelstvo Znanie, Moscow 1977.
95. M. Kyrš et al.: *Novye metody radioanaliticheskoi chimii (New Methods of Radioanalytical Chemistry)* (in Russian), Atomizdat, Moscow 1982.
96. J. Tölgyessy, Ju. V. Jakovlev, G. N. Bilimovitch: *Diagnostika okruchajushej sredy radioanaliticheskimi metodami (Diagnostics of the Environments with Radioanalytical Methods)* (in Russian), Energoatomizdat, Moscow 1985.
97. J. Tölgyessy: *Radioanaliticheskaja chimija (Radioanalytical Chemistry* (in Russian), Energoatomizdat, Moscow 1987.
98. S. Aronoff: *Techniques of Radiochemistry,* The Iowa State College Press Building, Ames, IA 1956.
99. V. R. S. Rao, Ch. P. Rao, G. Tataiah, *Radiochem. Radioanal. Lett.* **30** (1977) 365.
100. J. Klas, J. Tölgyessy, E. H. Klehr, *Radiochem. Radioanal. Lett.* **18** (1974) 83.
101. J. Klas, J. Tölgyessy, J. Lesný, *Radiochem. Radioanal. Lett.* **31** (1977) 171–179.
102. J. Lesný, J. Klas, Z. Koreňová, *J. Radioanal. Nucl. Chem., Lett.* **155** (1991) 145–153.
103. J. Lesný, Z. Koreňová, J. Klas, *J. Radioanal. Nucl. Chem., Lett.* **155** (1991) 155–168.

Enzyme and Immunoassays

James N. Miller, Department of Chemistry, Loughborough University, United Kingdom (Chap. 1)
Reinhard Niessner, Technische Universität München, München, Germany (Chap. 2)
Dietmar Knopp, Technische Universität München, München, Germany (Chap. 2)

1. Enzymatic Analysis Methods . . . 2
1.1. Introduction 2
1.2. Enzymes: Basic Kinetics 3
1.3. Enzyme Assays: Practical Aspects 4
1.3.1. General Considerations 4
1.3.2. Spectrometric Methods 5
1.3.3. Other Methods 7
1.3.4. Effects of pH, Buffer, Composition, and Temperature 7
1.3.5. Sources and Activity of Enzymes . 8
1.4. Enzyme Assays: Determination of Substrates 9
1.4.1. General Considerations 9
1.4.2. Determination of Glucose 9
1.4.3. Determination of Ethanol 9
1.5. Enzyme Assays: Determination of Inhibitors 10
1.5.1. Inhibitors of Cholinesterase Enzymes 10
1.5.2. Other Inhibition Methods 10
1.6. Enzyme Assays: Determination of Activators 11
1.6.1. Determination of Metal Ions by Metalloenzymes 11
1.6.2. Other Activator Analyses 11
1.7. Immobilized Enzymes 11
1.7.1. Introduction 11
1.7.2. Properties of Immobilized Enzymes 12
1.7.3. Application of Immobilized Enzymes 13
2. Immunoassays in Analytical Chemistry 14
2.1. Introduction 14
2.2. Polyclonal, Monoclonal, and Recombinant Antibodies 14
2.3. Sample Conditioning 15
2.4. Immunoassays 16
2.4.1. Radioimmunoassay 16
2.4.1.1. Isotopic Dilution Radioimmunoassay 16
2.4.1.2. Immunoradiometric Assay (IRMA) 16
2.4.2. Nonisotopic Homogeneous Immunoassay 17
2.4.2.1. Latex Particle Agglutination Immunoassay 17
2.4.2.2. Enzyme-Multiplied Immunoassay Technique (EMIT) 17
2.4.2.3. Apoenzyme Reconstitution Immunoassay System (ARIS) 17
2.4.2.4. Fluorophore-Labeled Homogeneous Immunoassay 17
2.4.2.5. Homogeneous Fluorescence Polarization Immunoassay (FPIA) . 18
2.4.2.6. Homogeneous Liposome Immunoassay 18
2.4.3. Nonisotopic Heterogeneous Immunoassay 18
2.4.3.1. Enzyme-Linked Immunosorbent Assay (ELISA) 18
2.4.3.2. Sandwich Immunoassay 20
2.4.3.3. Immunoassays using Immobilized Affinity Ligands 20
2.4.3.4. Immunoassays with Biotin and Avidin 21
2.4.3.5. Magnetic Particle Immunoassay . . 21
2.5. Immunoaffinity Techniques 21
2.6. Immunoassays on Test Strips and other Planar Structures 22
2.7. Characterization and Interference 24
2.8. Future Immunological Applications and Techniques . . . 26
3. References 27

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_129.pub2

1. Enzymatic Analysis Methods

1.1. Introduction

Many analyses in biological and clinical chemistry, and in the environmental, food and forensic sciences, involve the study of exceedingly complex multicomponent sample matrices. Moreover, the analytes under study are frequently present in trace amounts. Analytical techniques which combine exceptional selectivity with high sensitivity are thus required. Preeminent among these methods are those where the necessary selectivity is provided by biomolecules with very specific activities. This article surveys two groups of methods, those based on the activities of enzyme and of antibodies (the latter methods are normally called immunoassays).

The catalytic activity of enzymes provides for an enormous range of analytical techniques. Analytes are not restricted to conventional organic molecules, but include virtually all chemical species, including gases and metal ions. The ability of a single enzyme molecule to catalyze the reaction of numerous substrate molecules also provides an amplification effect which enhances the sensitivity of the analyses. A further advantage is that most enzyme-catalyzed reactions can be followed by simple, widely available spectroscopic or electrochemical methods. Enzymes are normally active only in mild conditions—aqueous solutions at moderate temperatures and controlled pH—and this restricts the circumstances in which they can be used, but some enzymes have proved remarkably robust to, e.g., mixed aqueous – organic solvent systems or elevated temperature; immobilized (insolubilized) enzymes often show enhanced stability compared with their solution analogs (see Section 1.7).

Antibodies have been used in analysis for over 60 years, and offer an unexpectedly wide range of techniques and applications. In some cases, the specific combination of an antibody with the corresponding antigen or hapten can be detected directly (e.g., by nephelometry), but more often such reactions are monitored by a characteristic label such as a radioisotope, fluorophore, etc. Since antibody reactions do not have a built-in amplification effect, these labels are frequently necessary to provide sufficient analytical sensitivity. Antibodies of different classes vary greatly in stability, but some are relatively robust proteins, and this contributes significantly to the range of methods available.

Enzymes and antibodies are proteins, with the ability to bind appropriate ligands very strongly and specifically. Enzymes frequently need cofactors or metal ions for (full) activity. Antibodies are multifunctional molecules, whose ability to bind to cell surfaces and to other proteins, at sites distinct from the antigen/hapten-binding sites, extends the range of labeling and detection methods available in immunoassays.

The analytical potential of enzymes and antibodies can be combined. Enzyme immunoassays, in which enzymes act as label groups in antibody-based analyses, are very well established; they are probably the most widely used immunoassay methods. An important development is the production of single molecules which combine antibody and enzyme activities. A critical feature of enzyme activity is that the transition state of the substrate binds very tightly to the enzyme at a position near the amino acid side-chain groups that participate in the catalytic reaction. It is thus possible, by using haptens which are analogs of substrate transition states, to generate antibodies with catalytic activity. These "catalytic antibodies" are of great research interest. Some of them have catalytic rates that are equal or superior to those of naturally occurring enzymes and their potential is emphasized by the possibility that they might be used to develop catalytic activities not present in living organisms. However, they have yet to find widespread application in bioanalysis. Intriguingly, catalytic antibodies have been found to occur naturally, especially in individuals suffering from autoimmune diseases.

This article can only attempt a general survey of enzyme and immunoassay methods, with references to more detailed reviews and books in specific areas. Many important clinical, veterinary, and forensic analyses in which enzymes themselves are the analytes are not discussed; this article considers only those methods where enzymes (and antibodies) are analytical reagents.

1.2. Enzymes: Basic Kinetics

The fundamental features of enzyme catalytic activity are described in detail in textbooks [1]. The simplest mechanism, outlined by MICHAELIS and MENTEN, proposes that the enzyme E reacts reversibly with the substrate S to form a complex, which subsequently decomposes to release the free enzyme and the product molecule P:

$$\mathrm{E+S \rightarrow ES \rightarrow E+P} \quad (1)$$

In this model, the combination reaction of E and S, with a rate constant k_1, occurs at the active site of the enzyme. This is a relatively small region of the E molecule; often a cleft or depression accessible from the surrounding aqueous solution. Its conformation, polarity, and charge distribution are complementary to those of the S molecule. The reverse process, dissociation of the ES complex, has a rate constant k_2, and the decomposition of ES to give E and P has a rate constant k_3. These three rate constants can be used to calculate the Michaelis constant K_M:

$$K_\mathrm{M} = (k_2 + k_3)/k_1 \quad (2)$$

The Michaelis constant clearly has units in mol/L, and in practice values between 10^{-1} and 10^{-7} mol/L. It is also convenient to define the turnover number of an enzyme, k_cat, which is the number of substrate molecules converted to product per unit time by a single enzyme molecule, in conditions where the enzyme is fully saturated with substrate. From the kinetic model $k_\mathrm{cat} = k_3$, but if more complex models are used k_cat depends on several rate constants. Depending on the enzyme and the reaction conditions (see Section 1.3) k_cat varies between ca. 1 s^{-1} and 500 000 s^{-1}.

The kinetic model outlined above is clearly oversimplified (e.g., it should really include an intermediate complex EP, in which the product molecule is bound to the enzyme), but it describes adequately the kinetic behavior of most enzyme-catalyzed reactions. If the steady-state hypothesis is applied to the ES complex in this model, the initial velocity of the enzyme-catalyzed reaction is:

$$V = k_3\ [\mathrm{E_T}]\,[\mathrm{S}]\,/([\mathrm{S}] + K_\mathrm{M}) \quad (3)$$

where squared brackets indicate concentrations, and $[\mathrm{E_T}]$ is the total enzyme concentration, irrespective of whether it is combined with S. When $[\mathrm{S}] >> K_\mathrm{M}$, i.e., when the enzyme active sites are saturated with substrate, V is simply k_3 $[\mathrm{E_T}]$. This is commonly called V_max, so Equation (3) becomes:

$$V = V_\mathrm{max}\ [\mathrm{S}]\,/([\mathrm{S}] + K_\mathrm{M}) \quad (4)$$

Equations (3) and (4) show that, when V is plotted against [S] at constant $[\mathrm{E_T}]$ (Fig. 1), the reaction rate is approximately proportional to [S] at low substrate concentrations, but eventually reaches the plateau level V_max. So provided $[\mathrm{S}] << K_\mathrm{M}$, substrate concentrations can be determined directly from reaction rate measurements. Such analyses are widely used (Section 1.4). Equation (4) also shows that, when $[\mathrm{S}] = K_\mathrm{M}$, $V = V_\mathrm{max}/2$. This provides a method of estimating K_M. In practice, K_M and V_max are usually determined by one of several transformations of Equation (4); the most common is the Lineweaver – Burk double-reciprocal method, in which $1/V$ is plotted against 1/[S] (Fig. 2) to give a straight line of slope $K_\mathrm{M}/V_\mathrm{max}$ and x- and y-axis intercepts $-1/K_\mathrm{M}$ and $1/V_\mathrm{max}$. This gives the following equation:

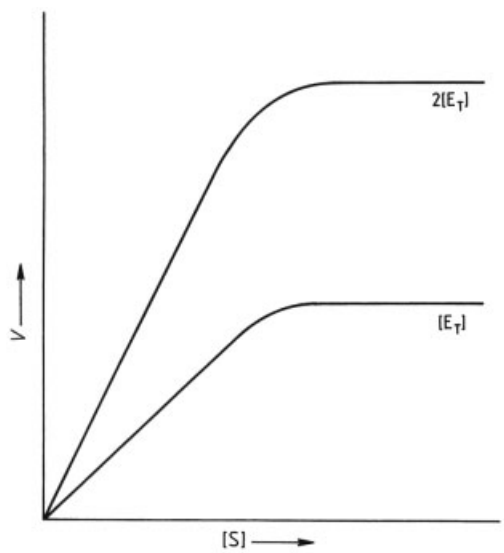

Figure 1. Relationship between reaction velocity V and substrate concentration [S] at different levels of enzyme concentration, $[\mathrm{E_T}]$ and 2 $[\mathrm{E_T}]$

$$(1/V) = (1/V_\mathrm{max}) + (K_\mathrm{M}/\,[\mathrm{S}]\,V_\mathrm{max}) \quad (5)$$

It should be noted that many enzyme-catalyzed reactions are reversible and that, in

common with other types of catalyst, enzymes do not affect the equilibrium position in such cases. In practice this means that additional reagents may have to be added to shift the equilibrium position and allow a practical analysis. For example in the determination of lactate using its oxidation to pyruvate with lactate dehydrogenase, the equilibrium lies on the lactate side of the reaction, so hydrazine or another carbonyl group reactant is added to remove the pyruvate as it is formed and pull the equilibrium in the desired direction.

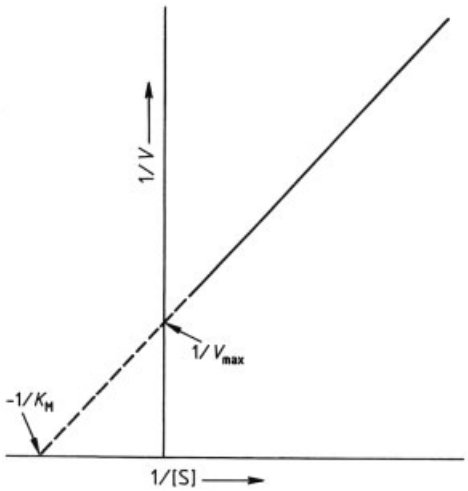

Figure 2. Double-reciprocal enzyme rate plot

An important feature of enzymes is that their active sites can often be occupied by, or react with, molecules other than the substrate, leading to inhibition of enzyme activity. Several inhibition mechanisms are known, but it is necessary only to distinguish between irreversible and reversible inhibition. Irreversible inhibition arises when the inhibitor molecule I dissociates very slowly or not at all from the enzyme active site. The best-known examples occur when I reacts covalently with a critical residue in the active site. Inhibition of cholinesterase enzymes by the reaction of organophosphorus compounds with a serine residue is a case in point. This type of inhibition is said to be noncompetitive—enzyme activity cannot be restored by addition of excess substrate. So although addition of I reduces $V_{\max}$, K_{M} is unaffected. The double-reciprocal plot in such cases has the same x-axis intercept as the plot for the uninhibited enzyme, but greater slope.

Reversible inhibition occurs when the EI complex can dissociate rapidly, just as the ES complex does. The most common examples arise when S and I are chemically or sterically similar, and compete for the active site (competitive inhibition—but note that not all reversible inhibition mechanisms are competitive). Thus, alcohol dehydrogenase enzymes oxidize ethanol to acetaldehyde; but other alcohols, e.g., methanol or ethylene glycol, can act as competitive and reversible inhibitors. The extent of inhibition in such cases depends on the relative concentrations of I and S: if the latter is present in great excess, inhibition is negligible. In these cases, $V_{\max}$ is not affected, but K_{M} may be increased; the double-reciprocal plot has the same y-axis intercept, but greater slope than the plot for the uninhibited enzyme.

When steady-state theory is applied to these mechanisms, in both types of inhibition the initial rate of the enzyme-catalyzed reaction is inversely related to [I], provided that the latter is small. Again, the possibility of a simple quantitative analysis of I by rate measurements is apparent. Such analyses are not so numerous as determinations of substrates, but some are very important (Section 1.5).

A number of enzymes are inactive or weakly active in the absence of activators. The best known examples arise with metalloenzymes, whose active sites include a main-group or transition metal ion. When the active *holoenzyme*, i.e., the complete molecule, including the metal ion, is treated with a complexing agent to remove the metal ion, the resulting inactive *apoenzyme* may provide a sensitive reagent for determination of the metal, enzyme activity being directly proportional to the concentration of activator at low levels of the latter (Section 1.6).

1.3. Enzyme Assays: Practical Aspects

1.3.1. General Considerations

All enzymatic analyses of substrates, inhibitors, and activators involve the determination of reaction rates: since initial rates are required, experiments are usually quite short. It is possible in principle to mix the reactants and measure

the rate directly, by recording an optical or electrochemical signal at regular intervals, the rate being the slope of the resulting signal plotted against time. When many samples are to be analyzed, such an approach is impracticable, and pseudo-rate methods are used. The reaction is allowed to proceed for a fixed time after the initial mixing; it is then stopped abruptly by a large change in pH or temperature (see below), or by addition of a specific enzyme inhibitor, and the extent of substrate depletion or product formation is measured. The signal recorded is then compared with the results of matching experiments on standard analyte solutions, and the test concentrations calculated by interpolation. In simple enzymatic assays, the calibration graphs are usually linear (see above), but in complex systems such as enzyme immunoassays, curved calibration plots are common, requiring more advanced statistical evaluation.

1.3.2. Spectrometric Methods

UV – Visible absorption spectrometry is the most commonly used method in modern enzymatic analysis. Suitable spectrometers are found in every laboratory, and many analyses use simple, robust, relatively cheap instruments. The availability of disposable polystyrene or acrylic cuvettes is a further advantage for conventional solution studies, but most analyses now use microtiter plates, the individual wells of which can be examined in a simple spectrometer. Modern spectrometers with imaging systems give very rapid results with 384- or even 1536-well plates. Continuous flow systems are also simple to set up. Spectrometers interfaced with personal computers routinely offer a range of control and data handling options.

Several generic applications of UV – visible spectrometry are common in enzymology (Section 1.4). Many oxidoreductase enzymes involve NAD/NADH or NADP/NADPH as cofactor systems. Such reactions are readily followed by absorption measurements at 340 nm, where the absorption of NADH and NADPH is strong, but that of NAD and NADP negligible (Fig. 3). This approach is so convenient that it is often applied in the form of a coupled enzyme reaction: if neither substrate nor product in the main reaction of interest is readily determined, one of them may be transformed via a second enzyme-catalyzed reaction involving NAD(P)/NAD(P)H, with detection at 340 nm. This principle has been extended so that three or more reactions occur in sequence, the final one either generating or consuming the readily determined NAD(P)H. Such indirect detection of the primary reaction is much less complex than it seems: often, all the enzymes, and any other reagents needed, can be included in the initial reaction mixture, the specificities of the of enzyme-catalyzed reactions in the sequence ensuring that they do not interfere. Extra sensitivity, and the opportunity to work in the visible region, is provided by using NAD(P)H to reduce colorless tetrazolium salts, in the presence of phenazine methasulfate (PMS), to intensely colored formazans, which absorb at ca. 500 nm. This further step is of particular value when a very rapid qualitative or semiquantitative result is required.

A further approach using UV – visible spectrometry involves colorigenic substrates. These molecules change color when hydrolyzed in the enzymatic reaction under study. A good example is provided by *p*-nitrophenol esters, which are colorless, but are hydrolyzed by appropriate enzymes to yellow *p*-nitrophenol, which can be determined at ca. 405 nm. Many substrates of this type are readily available, e.g., *p*-nitrophenyl phosphate as a phosphatase substrate, the corresponding sulfate as a substrate for aryl sulfatases, *N*-carbobenzoxy-L-tyrosine-*p*-nitrophenyl ester as a substrate for proteolytic enzymes such as chymotrypsin, etc.

Several analytically important enzymes catalyze reactions in which hydrogen peroxide is generated. Such reactions are easily followed colorimetrically, hydrogen peroxide being used in the presence of a second (peroxidase) enzyme to oxidize leuco-dyes to colored products. The same reactions can be used to monitor peroxidases themselves when they are used as labels in enzyme immunoassays; in modern systems, the enzyme-catalyzed reaction is halted by a large pH change, which simultaneously generates a distinctive color in the dye reaction product.

Fluorescence spectrometry is widely used in enzymology, usually because of its extra sensitivity compared with absorption methods. Fluorophores are characterized by two specific wavelengths, absorption and emission. The latter is the longer of the two, excited molecules having

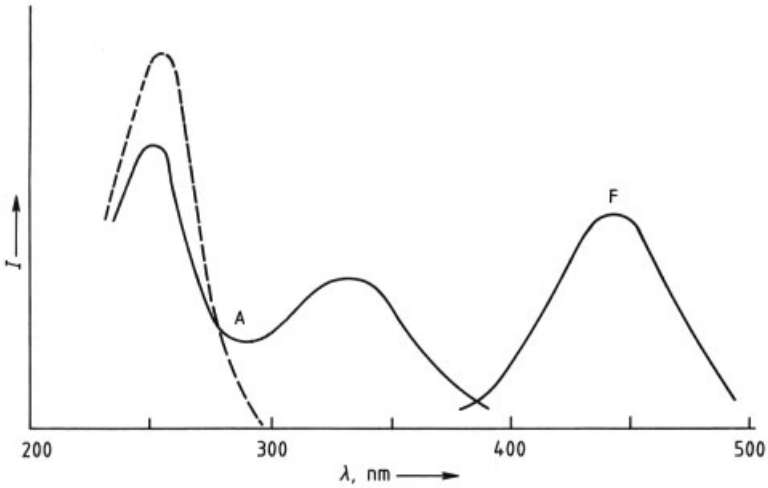

Figure 3. Absorption (A) and fluorescence (F) spectra NAD[P]H (——); and NAD[P] (- - -) (absorption only)

lost some energy between the processes of photon absorption and emission, and this allows the fluorescence to be determined (at 90° to the incident light beam in most instruments) against a dark background. This provides limits of detection unattainable by absorption spectrometry. A second important characteristic of fluorescence spectrometry is its versatility; it is just as easy to study concentrated solutions, suspensions, solid surfaces, flowing systems, etc., as it is to measure dilute solutions.

The generic approaches in absorption spectrometry are mirrored in fluorescence. Thus, NAD(P)H can be determined at an emission wavelength of ca. 460 nm with excitation at ca. 340 nm (Fig. 3). For work at longer wavelengths, NAD(P)H reduces nonfluorescent resazurin to the intensely yellow-fluorescent (ca. 580 nm) resorufin (this can also be determined colorimetrically at ca. 540 nm). Fluorigenic substrates are available for many hydrolytic enzymes, the best known being those based on resorufin esters, and on esters of 4-methylumbelliferone; the esters are nonfluorescent, but are hydrolyzed to fluorescent products. Peroxidase activity can be monitored via the conversion of a resorcin derivative to a fluorescent product with hydrogen peroxide.

Recent years have seen an explosion of interest in chemiluminescence (CL) and bioluminescence (BL) methods in enzymatic analyses. As in fluorescence, these methods measure photon emission from excited molecules, but here the latter are generated chemically, as reaction products. The reactions concerned are oxidations, often in mildly alkaline solutions in the presence of a catalyst. The advantages are:

1) No exciting light source is needed, so scattered light phenomena are minimized
2) It is usual to collect as many as possible of the emitted photons without the use of optical filters or monochromators. Screening devices using photographic detection have been developed. Sensitive Polaroid films allow the convenient qualitative or semiquantitative monitoring of chemiluminescent reactions.
3) Extreme sensitivities are possible
4) The reagents are generally free of hazard, despite offering limits of detection similar to those secured by radiochemical methods

Since the intensity of luminescence is proportional to the rate of generation of the excited reaction product, the principal precautions to be observed are those that apply when reaction rates are measured (control of temperature, pH, etc.), with the important additions that the order and the reproducibility of mixing of the reagents are critical, especially in conditions which yield short bright bursts of light. In such cases, fully automated CL/BL methods, such as those using flow injection analysis, are of obvious value. The common CL/BL techniques mirror those in UV – visible and fluorimetric measurements.

Processes involving NAD(P) and NAD(P)H can be followed with the aid of coupled reactions (in which two or more enzymes catalyzing consecutive reactions are involved), and reactions in which hydrogen peroxide is generated are readily monitored by using the peroxide to oxidize a well-known luminescent compound, such as luminol. Luminogenic substrates, e.g., for phosphatase enzymes, have been developed, and these find particular use in CL immunoassays. Sensitive and versatile microtiter plate readers are now widely available for fluorescence and chemiluminescence detection, so these spectroscopic methods are proving ever more popular in the monitoring of enzyme catalyzed reactions.

1.3.3. Other Methods

Although spectrometric methods still dominate the measurement systems used in biospecific analyses, other methods are important. Some of these are applied in specialized areas. The optical activity of saccharides can be used to follow reactions such as the inversion (hydrolysis) of sucrose, catalyzed by invertase. The products (D-glucose and fructose) have a combined optical rotation (e.g., at 589 nm) quite different from that of sucrose, so a simple polarimeter provides a good method of determining sucrose, the enzyme's substrate, and of other species which inhibit invertase activity. Radiolabeled substrates are often employed in research (^{14}C and ^{3}H are the most common isotopes), but safety precautions preclude the routine use of this approach. Since the radioactivity of the label is not changed in the enzymatic reaction, physical separation is necessary before the radiolabeled reaction product (or residual substrate) can be determined (see Section 2.4).

Electrochemical methods of following enzyme-catalyzed reactions are rapidly growing in popularity, mirroring the growth of these methods across analytical science in general. Particularly important are electrochemical enzyme sensors, portable or disposable devices which combine the specificity of an enzyme reaction with the simplicity and compactness of an electrical transducer. Selective electrodes have been adapted to form such sensors: for example, oxygen electrodes can monitor any enzyme-catalyzed reaction in which oxygen is consumed or produced. Other potentiometric and amperometric methods have achieved impressive successes, exemplified by the variety of glucose sensors now available. These devices use the glucose oxidase reaction (Section 1.4), and a variety of coupled reaction systems which provide an electrochemical response. In early applications, the hydrogen peroxide generated in the oxidation of glucose was used to oxidize hexacyanoferrate(II) to hexacyanoferrate(III), this second reaction being followed amperometrically. More sophisticated methods are now applied, and are generally applicable to all reactions in which peroxide is produced.

The rates of enzyme-catalyzed reactions can be followed, not only by a wide range of instrumental techniques, but also with the aid of a variety of chemical principles. In favorable cases, reaction rates can be determined directly by observing consumption of the substrate or appearance of the product, but if these compounds lack distinctive physical properties, indirect approaches, especially coupled reactions, can be used (see Section 1.3.2) .

1.3.4. Effects of pH, Buffer, Composition, and Temperature

Virtually all enzymatic assays are carried out at 20 – 50 °C in aqueous buffers of known pH and controlled composition. Both temperature and buffer properties affect the rates of enzyme-catalyzed reactions markedly. The effects of temperature can usually be summarized by a bell-shaped curve (Fig. 4 A). At lower temperatures, reaction rates increase with temperature, but beyond a certain point, denaturation (unfolding) of the enzyme molecules begins, so they lose their ability to bind the substrate, and the reaction rate falls. The temperature giving maximum activity varies from one enzyme to another, according to the robustness of the molecule. In some cases, it may be convenient to use a temperature rather below this maximum, otherwise the rate becomes too high to measure precisely. The rates of many enzyme-catalyzed reactions increase by a factor of ca. 2 over a range of 10 °C in the region below the maximum of the curve; it is thus necessary to control the temperature to within ca. ± 0.1 °C.

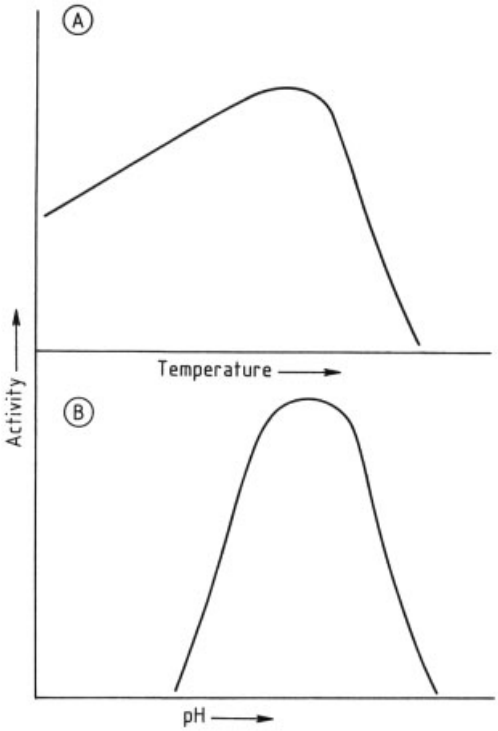

Figure 4. Effects of temperature (A), and pH (B) on the activity of a typical enzyme
The maxima and widths of the curves vary from one enzyme to another.

The pH of the buffer solution in which the analysis takes place must also be controlled. Most enzymes have a pH dependence of their activity of the type shown in Figure 4 B, and pH should be controlled to within $\pm$ 0.02. The optimum pH, and the range over which the enzyme is active, vary widely from one enzyme to another. These phenomena arise from the effects of pH on the structure of the enzyme itself, on the affinity constant between the enzyme and the substrate, on $V_{\max}$, and also on any coupled indicator reaction that is used. The choice of buffer recipe for any given pH may be important. The ionic strength of the buffer and the salts contained in it can influence the rate and mechanism of the main and coupled reactions, sometimes with unpredictable results. It may be necessary to choose a buffer whose properties represent a compromise between the ideal conditions for the main reaction, and the ideal conditions for the indicator reaction(s). Immobilized enzymes (Section 1.7) often have pH (and temperature) dependence significantly different from their soluble counterparts. This all points to the value of establishing and maintaining a well-defined buffer system for enzymatic analyses.

1.3.5. Sources and Activity of Enzymes

A wide range of enzymes is commercially available from numerous suppliers, so it is rarely necessary for the analyst to prepare such materials. Enzymes with the same name, but isolated from different species, may be quite distinct in their chemical and biological properties, and their activity (e.g., their requirements for cofactors). Enzymes are normally supplied as freeze-dried or crystallized proteins, and should be stored carefully at 4 °C or − 20 °C. Repeated freezing and thawing of protein solutions is not advisable, so it may be necessary to divide the dissolved enzyme into small aliquots, each of which is frozen and used just once.

The specificity of many enzyme-catalyzed reactions is a major reason for their use. Many organisms, or separate organs from a particular species, contain only one enzyme capable of catalyzing a given reaction. Thus, it may be unnecessary to use high-purity enzyme preparations; quite crudely purified material is often sufficient. In some cases, however, it is essential to ensure removal of a particular contaminant; enzyme preparations used in reactions involving hydrogen peroxide should be free of catalase, the enzyme which decomposes H_2O_2 to oxygen and water.

In practice many enzymatic analyses are performed with the aid of pre-packed kits. Such kits contain a number of vials which incorporate optimized mixtures of enzymes, cofactors, buffer salts, etc., as required, often in a dried form. Each vial provides a single analysis. The analyst needs only to reconstitute the reagents by the addition of water, add the sample, and perform the measurement after a given time period, recommended by the kit manufacturer.

Since an enzyme used as an analytical reagent may be quite impure, it is essential to know how much substrate a given weight of the enzyme preparation will convert to the corresponding product. This information can be derived from the activity of the material, given in International Units (IU). Unit enzyme activity converts one

micromole of substrate to product per minute at 25 °C and optimal pH. Commercial enzyme preparations are described in terms of IU/mg or, in the case of an enzyme supplied in solution or suspension, IU/mL.

1.4. Enzyme Assays: Determination of Substrates

1.4.1. General Considerations

As shown in Section 1.2, the rate of an enzyme-catalyzed reaction is proportional to the concentration of substrate, if the latter is small compared with K_M. (The guideline $[S] < 0.2\ K_M$ is often used.) This provides a simple approach to substrate determinations, which certainly represent the largest class of enzymatic analyses in modern practice, e.g., the measurement of blood glucose or blood cholesterol. In many instances, the analyses are highly selective, but not absolutely specific for the target substrate.

1.4.2. Determination of Glucose

Several different enzymes have been used to determine β-D-glucose, but by far the most popular approach uses glucose oxidase. Enzymes are often known by their Enzyme Commission numbers: glucose oxidase is E.C. 1.1.3.4. The first digit defines the enzyme as an oxidoreductase, the second as an oxidoreductase in which the hydrogen or electron donor is an alcohol, and the third as one in which the hydrogen or electron acceptor is molecular oxygen. As there are several enzymes exhibiting these three characteristics, the fourth digit purely acts as a means of differentiating between similar substances. This enzyme, normally obtained from *Aspergillus niger,* has a molecular mass of ca. 80 000, uses flavin adenine dinucleotide as cofactor, and catalyzes, at pH 7 – 8, the reaction:

$$\text{D-Glucose} + O_2 \rightarrow \text{D} - \text{Gluconolactone} + H_2O_2$$

This reaction is highly specific (α-D-glucose is oxidized at only 1 % of the rate of the β-anomer), but not perfectly so, as 2-deoxy-D-glucose (which does not occur at significant levels in blood) is also readily oxidized. Scores of different methods have been used to follow the rate of this reaction, almost all based on the properties of the very reactive product, hydrogen peroxide. (Many of these methods can also be used to determine other substrates which are oxidized by oxidoreductase enzymes to yield, among other products, H_2O_2, e.g., cholesterol, xanthine, L- and D-amino acids, D-galactose etc.). Colorimetric methods are based on conversion of a chromogen [e.g., 2,2′-azino-di(3-ethylbenzothiazolinsulfonate) diammonium salt (ABTS)] to a colored product (for ABTS, λ_{max} = 405 nm) with the aid of a second enzyme, horseradish peroxidase. Analogous fluorimetric methods, such as the oxidation of a stable non-fluorescent phenoxazine derivative to yield resorufin, have also been used. The optimum pH for the coupled peroxidase step is ca. 10, but both enzymes show adequate activity at pH ca. 8.5, an example of a compromise pH. Glucose oxidase reactions can also be monitored by the chemiluminescence from luminol stimulated by peroxidase in the presence of hexacyanoferrate(III). Numerous electrochemical methods have also been developed. Many glucose tests are available which use the enzyme in immobilized form. A typical dipstick test for urine samples uses a cellulose matrix which contains immobilized glucose oxidase and peroxidase along with a chromogen. In the presence of glucose, the initially colorless matrix develops a color within a few seconds. (An exactly similar test is available using galactose oxidase to test for D-galactose.)

With such methods, glucose is routinely determined in blood (normal level ca. 4 – 8 mmol/L) and foodstuffs. In all these analyses, it is assumed that molecular oxygen is present in excess, but it is also possible to use the enzyme in an excess of glucose to estimate levels of dissolved oxygen in aqueous or mixed aqueous – organic solvents. This unexpected application exemplifies the broad range of enzymatic analyses.

1.4.3. Determination of Ethanol

Ethanol can be determined with the aid of alcohol oxidase (E.C. 1.1.3.13), but a more usual approach uses alcohol dehydrogenase (ADH, E.C. 1.1.1.1.—the third digit indicates that this oxidoreductase has $NAD(P)^+$ as its hydrogen or electron acceptor). In neutral buffers, this zinc

metalloenzyme, normally obtained from yeast, but also from mammalian liver, catalyzes the reaction:

$$CH_3CH_2OH + NAD^+ \rightarrow CH_3CHO + NADH + H^+$$

The oxidation can be followed by monitoring the appearance of NADH by UV-absorption spectrometry, colorimetry, fluorimetry, or CL/BL methods (Section 1.3.2). Common sample matrices are blood plasma or various foods and beverages; as in the case of glucose, their analyte levels are often quite high, so few sensitivity problems arise. The catalytic action of this enzyme is not very specific: *n*-propanol and *n*-butanol, isobutanol, allyl alcohol, ethylene glycol, glycerol, and methanol (oxidized to formaldehyde, hence its toxicity) are among other alcohols that may interfere. This lack of specificity is one reason for the interest shown in alcohol oxidase as an alternative. The importance of routine alcohol determinations has encouraged the development of many methods using immobilized ADH in the form of "alcohol electrodes" or continuous flow analyzers.

1.5. Enzyme Assays: Determination of Inhibitors

1.5.1. Inhibitors of Cholinesterase Enzymes

Enzyme assays based on inhibition effects are not as commonly employed as substrate determinations, but one or two are very important. Preeminent is the determination of organophosphorus compounds by using their inhibitory effect on cholinesterase enzymes (E.C. 3.1.1.8—the first digit signifies a hydrolase enzyme, the second that the compounds hydrolyzed are esters, and the third that they are phosphoric monoesters). The latter catalyze the conversion of acylcholines to choline and the corresponding acid:

$$\text{Acylcholine} + H_2O \rightarrow \text{Choline} + \text{Acid}$$

This reaction is crucial to many living systems, so the organophosphorus compounds (pesticides, chemical warfare agents) that inhibit it by irreversible binding to the active site are often highly toxic. A range of cholinesterase enzymes is available, the most common being the enzyme isolated from horse serum, which is a pseudocholinesterase or butyrylcholinesterase. The specificity of such an enzyme is broad, and a wide range of esters can be hydrolyzed. This is of value in inhibitor analyses as the "normal" substrate can be replaced by, e.g., a fluorigenic substrate to facilitate detection. Thus, resorufin butyrate, a nonfluorescent ester of resorufin, is hydrolyzed by the horse serum enzyme to the intensely fluorescent resorufin. In the presence of an inhibitor, the expected growth in fluorescence intensity is reduced. Since a free acid is produced in the hydrolysis reaction, electrochemical methods provide viable alternatives, and enzyme reactors using immobilized cholinesterases are commercially available.

Organophosphorus compounds are often used as mixtures in pesticide formulations, and an ingenious application allowed the identification of such mixtures by combining the fluorigenic enzyme assay with thin layer chromatography (TLC). The sample was separated using TLC, and the plate treated sequentially with the enzyme and the fluorigenic substrate. The result was a mostly fluorescent TLC plate, with dark areas marking the positions of the enzyme inhibitors. In addition to organophosphorus compounds, carbamates (also used as pesticides) and the calabar bean alkaloid physostigmine act as inhibitors of cholinesterases, and can also be determined.

1.5.2. Other Inhibition Methods

The literature on enzyme inhibition assays is very large, but many of the methods are of limited value because they lack specificity, and/or do not compete with other techniques. Thus, many enzymes are inhibited to a greater or lesser extent by transition metal ions, but the lack of specificity for individual ions and the preference for alternative techniques, e.g., ion chromatography, has reduced the value of such approaches.

An area of great importance is the search for enzyme inhibitors, especially inhibitors of proteolytic enzymes, in high-throughput screening programs for new drugs. Combinatorial chemistry methods can generate candidate inhibitors at the rate of thousands per working day, so rapid and efficient screening procedures are essential.

Most currently used methods employ microtiter plate formats with fluorescence detection. Fluorescence phenomena are extremely susceptible to changes in the environment of a fluorophore, and variables such as fluorescence lifetimes and polarization values can thus be used conveniently to monitor the binding of small molecules to macromolecules.

1.6. Enzyme Assays: Determination of Activators

1.6.1. Determination of Metal Ions by Metalloenzymes

Activator analyses are the least common application of enzymatic methods, but, as in the case of inhibitor analyses, a small number of methods are extremely valuable. Of great interest are methods where incorporation of metal ions in the active site of an enzyme is critical. Removal of the metal from the complete enzyme, the holoenzyme (e.g., by mildly acidic solutions of EDTA) generates the apoenzyme, which lacks enzyme activity. This apoenzyme then becomes the reagent in the subsequent analysis, along with its natural substrate and any indicator reagents that may be required. In a sample containing the appropriate metal ion, the activity of the enzyme is wholly or partially restored. In contrast to the metal ion inhibition effects mentioned above, these analyses are often highly selective, as the active sites of metalloenzymes accommodate (with generation of activity) only a very few metal ions (ionic radius may be as important as ionic charge).

An example of such analyses is provided by isocitrate dehydrogenase (E.C. 1.1.1.42), normally isolated from pig heart. This enzyme catalyzes the oxidation of isocitrate ions to α-oxoglutarate ions; the reaction is NADP dependent, and results in a decarboxylation:

$$\text{Isocitrate}+\text{NADP}\rightarrow\text{NADPH}+\text{CO}_2+\text{H}^+ +\alpha-\text{Oxoglutarate}$$

The usual methods for following NAD(P)-dependent reactions are available here. The enzyme requires the presence of Mg^{2+} ions for activity, and extensive studies show that, apart from Mg^{2+}, the only other ions that combine with apo-ICDH (apo-isocitrate-dehydrogenase) to restore its activity are Mn^{2+} and to a lesser extent Zn^{2+} and Co^{2+}. This approach has been successfully applied to the determination of Mg^{2+} in blood and other samples at concentrations below 1 μmol/L, and to the analysis of Mn^{2+} at concentrations as low as 5 ppb. The advantage of this method of trace metal determination is that only a single oxidation state of the metal is determined—in the manganese analysis, only Mn(II), with no interference from Mn(VII) or other oxidation states.

1.6.2. Other Activator Analyses

The most obvious activator assays are those in which cofactors are analyzed: all the common cofactors can be determined in this way, and in many cases their own properties provide simple means of measuring reaction rates. The analyses are usually very selective. ATP in red blood cells and many other samples is often determined by using its specific participation in the well-known firefly bioluminescence reaction:

$$\text{Luciferin}+\text{O}_2+\text{ATP}\rightarrow\text{Oxyluciferin}+\text{ADP}+\text{PO}_4^{3-}+\text{Light}$$

This reaction is catalyzed by firefly luciferase, which requires Mg^{2+}, and can also be used to determine this ion. Even in simple instruments, ppb levels of magnesium can be detected.

1.7. Immobilized Enzymes

1.7.1. Introduction

Analysts have attempted to use immobilized enzymes for over 40 years. The original motivations were two-fold: to conserve enzymes that were expensive and difficult to isolate, and to incorporate enzymes in reusable sensors, such as enzyme electrodes and flow analysis systems (see below). During such researches, other advantages and applications of immobilized enzymes were found, and immobilization technologies were extended to other areas of biospecific analysis such as immunoassays (see below) and affinity chromatography. Immobilized enzymes have also been used extensively in manufacturing processes and for therapeutic purposes.

Several distinct approaches to the problem of immobilizing an enzyme in or on a solid or

gel, with optimum retention of biological activity, are available. These include covalent binding, surface adsorption, gel entrapment, encapsulation within a semipermeable membrane, and chemical cross-linking. The first is by far the most popular, and many enzymes are commercially available in covalently immobilized form. Common solid phases include particles of well-known chromatographic media, e.g., agarose, glass and silica, polystyrene, polyacrylamide and cellulose, modern perfusion chromatography phases, and nylon in the form of flow tubing or membranes. The best-known covalent linkage methods include the use of cyanogen bromide (CNBr) to bind $-NH_2$ groups to hydroxylic matrices such as agarose or cellulose, chloro-*sym*-triazinyl derivatives to perform a similar function, and carbodiimides to link amine and carboxylic acid functional groups between enzymes and solid phases. Such covalent linkages are not infinitely stable; even the best immobilized enzyme reactors suffer slow leakage of the enzyme molecules into solution. However, many derivatives remain stable on storage and in use over periods of several months.

The idea of entrapping enzyme molecules within a polymeric matrix of known (average) pore size is attractive and has been frequently studied, polyacrylamide gels being widely used. However the outcomes have often been disappointing. The polymerization process, carried out in the presence of a solution of the enzyme, may generate sufficient heat to denature some of the enzyme; leakage may occur from within the polymer matrix; enzyme activity with respect to large substrates may diminish sharply because of the molecular sieving effect of the polymeric gel; and even the activity to small molecules may be diminished because the enzyme molecules are less available than they are on the surface of a solid particle.

However, interest in this approach to enzyme immobilization has been renewed by the application of silica-based sol-gel entrapment matrices. These provide mild hydrophilic environments with controlled pore sizes. Moreover sol-gels can be formed into optically clear monoliths of varying dimensions, thin films, and fibers. Several applications including the determination of metal ions and of peroxidase have already been described.

Enzymes are readily adsorbed to the surfaces of ion exchange matrices, plastic microtiter plates, hydroxyapatite, even charcoal. Such immobilized molecules may be suitable for one-off assays (i.e., not for repeated use), but they are often readily desorbed by quite small changes in their environment, and enzymes immobilized in this way have not found great application in routine analyses, despite the ease with which adsorption is achieved. However, the adsorption of antibodies on plastic surfaces is routinely used in many immunoassays, including enzyme immunoassays (Section 2.4.3.1).

Bifunctional reagents, such as glutaraldehyde, can cross-link enzyme molecules to produce insoluble aggregates that retain at least some activity, but these cross-linked enzymes often have unsuitable mechanical properties, and they have found little practical use.

Microencapsulation of enzymes (and other biomolecules) has been extensively studied. The microcapsules have semipermeable walls that allow small substrate and product molecules to pass through freely, while large enzyme molecules are retained in solution within the capsules. However, the polymerization process used in forming the capsules generates heat that may damage the enzymes.

1.7.2. Properties of Immobilized Enzymes

Many experiments have shown that the advantages of (usually covalently) immobilized enzymes extend beyond their reusability; their physicochemical properties are often significantly different from those of the corresponding enzyme in free solution. The immobilized enzyme usually has somewhat less activity than the same mass of enzyme in solution. But unless activity losses are severe, this disadvantage is not serious, and can be compensated by using more of the immobilized preparation, knowing that it can be used repeatedly.

A significant advantage of immobilization is the increased thermal stability conferred on many enzymes, allowing their use for longer periods at higher temperatures than would be possible for the soluble molecule. An early experiment showed that immobilized papain (a protease derived from papaya latex) retained over half its room-temperature activity at 80 °C in

conditions in which the soluble enzyme was almost entirely denaturated.

Also interesting and valuable are the changes in the pH – activity curves that often accompany immobilization. A common result is broadening of the curve, i.e., the enzyme is active over a wider pH range than its soluble counterpart. This is normally ascribed to the range of microenvironments of different enzyme molecules in or on the solid matrix. Depending on the charge properties of this matrix, the optimum pH may undergo significant shifts. The optimum pH for an enzyme bound to a negatively charged carrier such as carboxymethylcellulose is shifted to higher values, while immobilization on a cationic matrix such as DEAE-cellulose (diethylaminoethyl-cellulose) has the opposite effect. These effects are ascribed to the change in the enzyme's microenvironment brought about by neighboring charged groups. Immobilization on a neutral carrier is not expected to change the pH optimum. Since these pH effects can be controlled to some degree, immobilization of an enzyme may confer significant advantages, e.g., using two or more enzymes with different pH optima in solution (Section 1.4.2) may be facilitated.

1.7.3. Application of Immobilized Enzymes

Since most of the enzymes routinely used in analytical science are now commercially available in immobilized form, it might be expected that many analyses conventionally carried out in solution are also performed with the immobilized products. In practice, many of the important applications of immobilized enzymes are those in which the enzyme preparation is incorporated into a more specialized analytical system, such as an enzyme electrode or continuous enzyme reactor.

One of the first applications of immobilized enzymes was the construction of an enzyme electrode for glucose, in which a thin layer of glucose oxidase entrapped in a polyacrylamide gel was placed over a conventional Clark oxygen electrode. When the enzyme electrode was immersed in glucose solution, the glucose penetrated the enzyme layer, consuming oxygen and reducing the potential of the Clark electrode. Apart from concerns over the long-term stability of the enzyme layer, the principal drawbacks of this and many other similar devices were their response times (often minutes rather than seconds), and their recovery times, i.e., the rates at which the electrodes could be used with different samples. Some modern devices have largely overcome these problems, but although enzyme electrodes (most based on platinum, oxygen, or other established sensors) have been developed for numerous analytes (amino acids, urea, sugars, simple alcohols, penicillin, cholesterol, carboxylic acids, inorganic anions, etc.), only a very few seem robust enough to survive routine laboratory use over long periods.

On the other hand disposable "one-shot" sensors based on immobilized enzymes are now of major importance in several biochemical and environmental application areas.

A second obvious area of application is in continuous flow analysis or flow injection analysis systems, in which the immobilized molecules form reactors that can be readily inserted and replaced in a flow analysis manifold. The physical form of the enzymes varies widely; packed-bed reactors are often used, but open-tube wall reactors and membrane reactors have also been investigated. A principal advantage of all such systems is that they can use all the optical or electrochemical detectors routinely used in flow analysis. However, the problems of producing stable and robust immobilized enzyme reactors have proved more intractable than many researchers hoped, and other advances (e.g., the use of more sensitive detectors, improved availability of low-cost soluble enzymes) have minimized the advantages of using solid phase enzymes.

Nonetheless, much research continues in this area, particularly work on thin layers or membranes to which enzymes are attached. In such cases, the enzyme kinetics may be favorable, and, of the available detection procedures, fluorescence spectroscopy is particularly suitable for the study of solid surfaces, and is highly sensitive.

2. Immunoassays in Analytical Chemistry

2.1. Introduction

Analytical chemistry has grown tremendously, especially in the fields of environmental and process analysis. Modern measurement techniques have led to lower limiting values in the areas of water, soil, and air analysis. This has been followed by legislation requiring environmentally compatible production processes under constant control. As a result, less expensive (and often faster) monitoring techniques are necessary. The search for "chemical sensors," which, similar to optical and electrical sensors, can be integrated into production monitoring and quality control, is a logical consequence of this ecological and economic pressure.

Immunoassays (IAs) are based on the formation of a thermodynamically stable antigen – antibody complex. These methods play an important role, especially in clinical chemistry, being used for the fast and safe detection of proteins, hormones, and pharmaceutical agents. These techniques promise to close the gap between the cheapest chemical sensors and conventional, expensive, slower analytical methods.

On the other hand, classical analytical chemistry is just starting to accept immunoassays; the reasons for this hesitancy are to be found in the necessity of:

1) Synthesizing an immunogen and coating antigen (hapten linked to different carrier proteins)
2) Isolating an antiserum after immunization, usually of a vertebrate
3) Development, optimization, and synthesis of a labeled analyte (hapten) derivative after isolation and characterization of the first antibodies (AB)
4) Validation of both the antibodies obtained, and the entire test
5) Possibly having to start the immunization all over again, if the antibodies do not fulfill the selectivity and affinity requirements

Every analytically usable method must fulfill the criteria of selectivity, sensitivity, calibration ability, and reproducibility.

The development of immunoassays is frequently based on empirical findings which are an obstacle to certification. In particular, despite the fact that polyclonal antibodies often permit better detection tests, they are not reproducible in any given laboratory. Even the production of monoclonal antibodies, which in principle is easily reproducible, does not aid worldwide acceptance and recognition of immunoassays because of their extremely complex production processes, which are increasingly being patented.

Nevertheless, a steady rapid development of immunotests can be observed, especially in the field of environmental monitoring [1 – 3], where the question of costs has become so important that the establishment of an immunotest can be worthwhile in spite of all the facts mentioned above. In the development of chemical sensors, efforts are also being made to utilize the biochemical recognition principle by coupling with optical, electrochemical, or other transducer (signal transfer) [4 – 6]. However, the slower kinetics involved in molecular biological processes require stricter maintenance of the experimental protocol, e.g., the time of reaction between the antibody and the tracer molecule or analyte (hapten) should be set up in an exactly reproducible manner by means of flow-injection analysis [7].

Immunoassays become important when:

1) Fast measurement and evaluation are required
2) Highest possible detection strength is required
3) Large numbers of samples are to be expected
4) Only complex and expensive analytical methods are otherwise available

The greatest potential for the use of immunoassays in environmental analytical chemistry is in screening, i.e., for the selection of contaminated and uncontaminated samples for further validation analysis.

2.2. Polyclonal, Monoclonal, and Recombinant Antibodies

The isolation of highly selective (specific), high-affinity antibodies is the primary requirement for the successful development of an immune test [8, 9] (→ Monoclonal Antibodies).

In analytical chemistry, especially in the environmental field, analytes of molecular mass <

1000 are often of interest, but molecules of this type (haptens) cannot alone induce the generation of antibodies.

For this reason, it is necessary to produce an immunogen by coupling the hapten to a high molecular carrier protein. The possibilities for immunogen synthesis are restricted by the structure of the hapten. The presence of amino or carboxyl groups, which can be directly coupled to the carrier protein by a peptide bond, is most favorable. If the hapten contains no reactive group, it must be introduced. An example is the reaction of haptens containing hydroxyl groups, which are converted to hemisuccinates with succinic anhydride or *N*-bromosuccinimide. Haptens with keto groups can be converted to oxime derivatives.

Modified haptens or haptens containing reactive groups are then coupled to a high molecular carrier protein. High yields and no interfering side products are usually obtainable from reactions at room temperature in aqueous media and at near-neutral pH. The carrier proteins commonly used are keyhole limpet hemocyanin human or bovine serum albumin, bovine gammaglobulin, or thyroglobulin. Coupling reactions comprise: (1) the carbodiimide-mediated conjugation, (2) the mixed anhydride procedure, (3) the NHS (*N*-hydroxysuccinimide) ester methods, and (4) the glutaraldehyde condensation [10]. The degree of coupling, i.e., the number of hapten molecules per protein molecule, should be determined, using either radiolabeled hapten, UV and visible spectroscopy, matrix assisted laser desorption time-of-flight mass spectrometry (MALDI TOF MS), or indirect methods.

The choice of a suitable position of attachment to link the hapten to the carrier is important because it is responsible for recognition by the antibody produced. Especially in the case of parent substances which are rich in isomers, only some of which are of toxicological interest, e.g., PCBs or dioxins, the reactive positions used for conjugate synthesis must be carefully selected.

For the generation of polyclonal antibodies, the immunogen is applied to a vertebrate (rabbit, sheep, etc.), usually with the aid of adjuvants which enhance immunogenicity, by a mechanism that is not fully understood. The production of polyclonal antibodies is not very reproducible. For this reason, antibody concentration (titer), specificity (selectivity of the antibodies), and affinity must be continually monitored, and the ability of the isolated antiserum to bind to the hapten must be determined. Further, the cross-reactivity of the antibodies should also be determined as early as possible. Cross-reactivity refers to the reaction of an antibody with molecules which are structurally very similar to the desired analyte (hapten); it determines the sensitivity to interfering substances in the final test.

Polyclonal immune sera from an immunization of this type consist of a spectrum of antibodies, which are active against the diverse epitopes of the same antigen and, therefore, possess various affinity constants.

The acceptance of immunoassays depends on the availability of antibodies. Since the amount of polyclonal antibodies is always limited, and even repeated immunizations of the same experimental animal species results in nonreproducible characteristics (specificity, affinity), the production of suitable monoclonal antibodies (they are product of a single cell alone) is most important. (→ Monoclonal Antibodies). The method was developed by Köhler and Milstein and encompasses the immortalization of a single-antibody producing cell (B-lymphocyte) by fusion to a myeloma cell [11]. The resulting hybridoma produces just one species of antibody which can be grown indefinitely and, therefore, offers a means to advise a consistent product in unlimited quantities. Up to now, compared to polyclonal antibodies monclonal antibodies are less frequently applied in environmental analysis.

In recent years, genetic engineering processes have entered the field of antibody production. A number of functional recombinant single-chain antibody fragments (scAbs, recombinant antibodies) which recognizes and binds to small compounds has been expressed and characterized in microorganisms and transgenic plants [12 – 14]. However, these techniques are still at the development stage.

2.3. Sample Conditioning

Sample preparation is of tremendous importance in the application of an immunoassay. Apart from the usual criteria for representative sam-

pling, special attention should be paid to the following influences:

1) pH value
2) Humic substances
3) Surfactants
4) Heavy metals
5) Organic solvents

If enzymatic tracers are used, the actual detection reaction is frequently disturbed by influences exerted on the reactive center of the enzyme, especially those due to interactions of the antibody with polyfunctional groups, e.g., with the humic substances in natural water samples [15]. This should be taken into account, especially in the use of immunoassays for soil screening. In competitive immunoassays, analytes that are sparingly soluble in water can be solubilized by the addition of surfactants [16]. On the other hand, the surfactant can alter the tertiary structure of the antibody, and possibly of the enzyme involved. The use of organic solvents is being intensively studied [17]. In some cases, immunoassays are known to tolerate more than 10 vol % of solvent.

The separation or complexation of humic substances and/or heavy metals is successful in favorable cases, e.g., by the addition of bovine serum albumin (BSA) in a trinitrotoluene enzyme-linked immunosorbent assay (ELISA) [18]. In the screening of soil samples for pesticides, persistent centrifugation at 17 000 rpm, and simple dilution with bidistilled water in the ratio 1 : 10 was sufficient [19].

In each application, the appropriate measurement must be established by independent validation analysis. Nonspecific interactions are very frequent in immunoassays, and usually lead to a slight overdetermination of the analyte.

2.4. Immunoassays

2.4.1. Radioimmunoassay (→ Diagnostic Reagents, Chap. 3.2)

The development of radioimmunoassays (RIA), based on studies by Yalow and Berson, has opened up an area of application especially in clinical chemistry [20, 21].

In RIAs, a radioactive label is used for detection of the formation of an antibody –antigen complex. Thus, a simple, specific, sensitive, and precise determination of the radioactive isotope is available. At present, however, immunoassays are increasingly carried out with nonisotopically labeled antigens or antibodies because of the legal restrictions on the use of radioactive material.

2.4.1.1. Isotopic Dilution Radioimmunoassay

In this classical RIA, a variable amount of unlabeled analyte (hapten) competes with a constant amount of a radiolabeled substance for a limited number of antibody binding sites:

$$\mathrm{Ag^* + Ag} \begin{cases} \nearrow \mathrm{Ag^*\,AB(s) + Ag(b)} \\ \searrow \mathrm{Ag\,AB(s) + Ag^*(b)} \end{cases}$$

where Ag is the antigen (hapten or analyte); Ag* the radiolabeled antigen; AB the antibody; and (b) and (s) indicate species bound or in solution.

The more analyte in the sample, the more labeled antigen is found in solution. To quantify this amount, a separation step must be carried out, usually by subjecting the adsorptively bound antibodies (e.g., on a microtiter plate) to a washing step.

In principle, standard and analyte should have the same ability to displace a tracer molecule from the binding site on the antibody.

2.4.1.2. Immunoradiometric Assay (IRMA)

This process uses radioactively labeled antibody AB*. Two antibodies are required which recognize different binding sites on the antigen; the first is fixed on a microtiter plate or in a test tube, and the second is labeled, e.g., with ^{125}I. Sequential incubation is carried out, first with the sample or the standard analyte, followed by a separation washing step. Subsequent incubation with radiolabeled antibody AB* shows the antigen binding density through sandwich coupling:

$$\mathrm{AB\,(s)\ + Ag \rightarrow AB - Ag\,(s)}$$

$$\mathrm{AB - Ag\,(s) + AB^* \rightarrow AB - Ag - AB^*}$$

Small analyte molecules (haptens) cannot be detected by this technique because of the lack of several binding sites (epitopes), so it cannot be used for environmental analysis.

2.4.2. Nonisotopic Homogeneous Immunoassay

For diverse reasons, such as restrictions on the use of radioactive substances, nonisotopic immunoassays have gained general acceptance, especially in trace analysis and environmental analysis [1 – 3].

Consequently, a sensitive measuring principle other than radioactive decay is required. Furthermore, a distinction must be made between tests that require a separation step to determine the bound component and those that do not. Nonisotopic, homogeneous immunoassay allows the direct observation of the hapten bound to the antibody.

2.4.2.1. Latex Particle Agglutination Immunoassay

A large number of latex agglutination immunoassays have been adopted from clinical chemistry. These assays are based on the visualization of antigen – antibody complexes by the attachment of latex particles or gold colloids. Entities of this type with dimensions in the nanometer or micrometer range can be quantified by turbidimetry, nephelometry, light scattering techniques, and particle counters [22 – 25].

2.4.2.2. Enzyme-Multiplied Immunoassay Technique (EMIT)

In the EMIT [26, 27], the analyte is covalently bound to the enzyme in spatial proximity to the active site and, consequently, the formation of the antibody – antigen complex inactivates the enzyme; addition of hapten results in a reduction of this inactivation.

2.4.2.3. Apoenzyme Reconstitution Immunoassay System (ARIS)

If, however, the antigen is covalently bound to the prosthetic group of an enzyme such as glucose oxidase (E.C. 1.1.3.4) and an aliquot of the coupled antigen to flavin – adenine dinucleotide) is added to determine an analyte, free antibodies prevent the reconstitution of the enzyme. The concentration of the free antibody naturally depends on the analyte concentration in the sample. Similar to the EMIT technique, the ARIS is used in automatic analyzer systems in clinical chemistry [28].

2.4.2.4. Fluorophore-Labeled Homogeneous Immunoassay

At first glance, fluorescent labeling appears to have a much higher detection strength compared to colorimetric detection, but this is not the case. First, the affinity constant generally limits the detection strength of a process. Second, fluorophores are exposed to many influences, such as quenching by impurities, or even adsorption of the fluorophore molecule. However, the fact that the detection can be repeated is advantageous, whereas a chemical reaction is irreversible [29].

The best-known variant is the substrate-labeled fluorescence immunoassay (SLFIA) test [30]. Galactose is linked to the antigen and to the fluorophore, methylumbelliferone. β-Galactosidase (E.C. 3.2.1.23) is capable of hydrolyzing and releasing the fluorophore, as long as the antigen linked to galactose is not stabilized in an antibody – antigen complex. If free antigen is present, the formation of galactose – antigen – fluorophore – antibody complex is reduced. The measuring signal is conveniently proportional to the concentration of free antigen.

In pesticide trace analysis, efforts have been made to observe directly the coupling of antigen – antibody in the quenching of the fluorophore bound to a tracer molecule, by the use of time-resolved, laser-induced fluorescence [31]. On formation of an Ag* – AB complex, a significant change in the fluorescence lifetime is observed. However, the fluorescence quantum yield suffers drastically, and this test has never been used.

2.4.2.5. Homogeneous Fluorescence Polarization Immunoassay (FPIA)

Direct observation of the formation of a hapten – (fluorescently labeled) antibody complex is also possible in polarized light [32]. The presence of free hapten reduces the antibody-tracer complex concentration, and the degree of polarization is lowered. The detection strength of this test is in the µmol/L range and thus not yet high enough for environmental analysis.

2.4.2.6. Homogeneous Liposome Immunoassay

Owing to "cell wall formation" by multilayer membranes, liposomes may encapsulate up to several hundred fluorophore molecules.

The measuring principle is based on the fact that hundreds of hapten molecules are incorporated into the cell membrane of liposomes, e.g., by linkage of phosphatidylethanolamine (a functional part of the multilayer membrane). In the presence of an antibody for the formation of an immune complex, destabilization of the entire vesicle occurs. The contents of the liposome, e.g., fluorophores dyes, spin-labeled molecules, or enzymes, are released. Liposome stability is a general problem associated with this method [33].

2.4.3. Nonisotopic Heterogeneous Immunoassay

The most common type of immunoassay is the nonisotopically labeled heterogeneous test [1]. Either the antibody or the hapten (analyte) to be detected is fixed to a solid interface via a covalent bond or by adsorption. The other significant difference lies in the fact that no radioactive labeling is used for signal production.

2.4.3.1. Enzyme-Linked Immunosorbent Assay (ELISA)

This assay is the most important in trace and environmental analysis. First, antibodies are immobilized on a solid carrier. For this, it may be necessary to take suitable measures, such as ^{60}Co radiation, to make the carrier sorption active. Standard 96-well microtiter plates, polystyrene beads, or test tubes can be used.

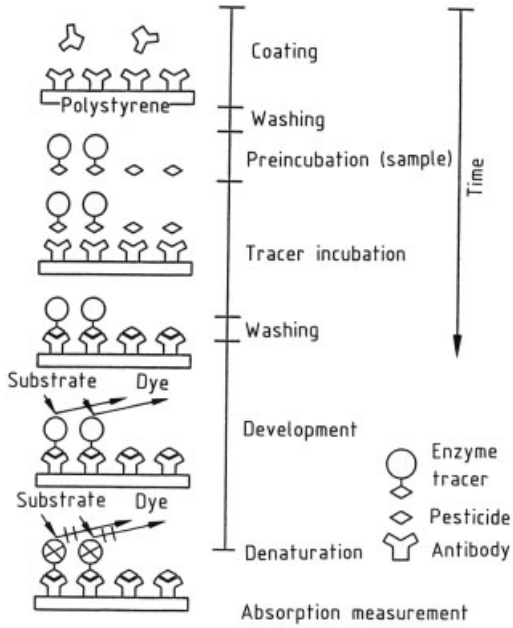

Figure 5. Principle of a direct competitive enzyme immunoassay (ELISA) with photometric detection

In a typical ELISA test on a microtiter plate (Fig. 5), the carrier is first coated with analyte-binding antibody. The sample or standard is added after washing off excess antibody with a surfactant solution. The volume added is typically 100 – 200 µL. After (optional) preincubation, a constant quantity of tracer, e.g., enzyme-labeled hapten, is added to the sample or standard. This initiates a competitive reaction, because only a limited number of antibodies are available for binding. After the tracer incubation period, sample (or standard) and excess reagents are washed away. The bound tracer concentration is inversely proportional to that of the analyte. The amount of bound tracer can be determined via an enzyme substrate reaction with a chromogenic substrate. After a reaction time long enough to produce sufficient dye, the enzyme is denatured (enzyme reaction stopped) by addition of acid. Subsequently, the depth of color formed in the individual cavities of the mi-

crotiter plate can be automatically determined with a "plate reader" (photometer).

The application of a competitive heterogeneous immunoassay is very common. Some important immunoassays for trace amounts of environmentally relevant compounds are presented in Table 1. Especially for agrochemicals, a large number of determination procedures have been established, because of legislation on drinking water which stipulates (as a precautionary measure) a limiting value of 100 ng/L for a single pesticide; this is not based on toxicological findings. The demand for sampling and pesticide analysis has given a sudden worldwide impetus to the development of immunoassays as a fast and cost-effective method of analysis.

Table 1. Selected ELISA tests in environmental analysis

Substance	Detection limit	Reference
Pesticides		
Alachlor	1 ppb	[34]
Aldicarb	300 ppb	[35]
Atrazine	50 ppt	[36]
Benomyl	250 ppb	[37]
Bentazon	2 ppb	[38]
Diflubenzuron	1 ppb	[39]
Endosulfan	3 ppb	[40]
Fenpropimorph	13 ppt	[41]
Hydroxyatrazine	50 ppt	[36]
Metazachlor	10 ppt	[42]
Methabenzthiazuron	50 ppt	[43]
Molinate	30 ppb	[44]
Norflurazon	1 ppb	[45]
Terbuthylazin	60 ppt	[46]
2,4-D	100 ppb	[47]
Aromatic compounds		
Benzo[*a*]pyrene	100 ppt	[48]
1-Nitropyrene	50 ppt	[49]
Pyrene	35 ppt	[50]
Trinitrotoluene	50 ppt	[51]
Halogenated aromatic hydrocarbons		
Pentachlorophenol	30 ppb	[52]
2,3,7,8-Tetrachlorodibenzodioxin	5 ppb	[53]
Heavy metals		
Mercury(II)	500 ppt	[54]
Surfactants		
Triton X-100	2 ppb	[55]

The use of indirect competitive heterogeneous immunoassays (Fig. 6) is also common. Here, the hapten is bound to the surface of the carrier. Since hapten alone cannot be bound very reproducibly to a solid support by adsorption, a protein-linked hapten is used (e.g., to BSA). In principle, however, a conjugate (coating conjugate) different from the immunogen should be employed.

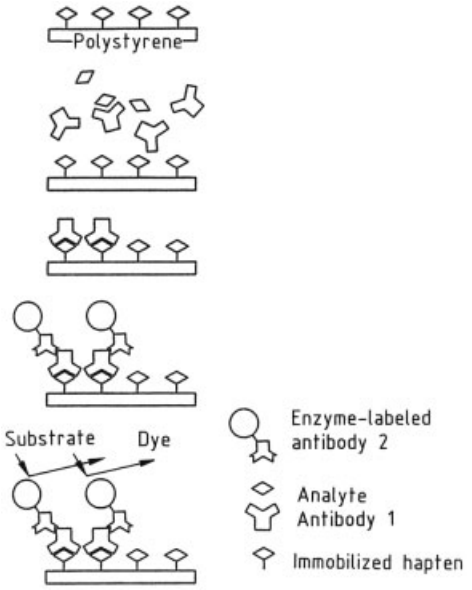

Figure 6. Principle of an indirect competitive enzyme immunoassay (ELISA)
The immobilized hapten corresponds to a hapten – BSA conjugate. AB 1 binds to the analyte, AB 2 binds to foreign antibodies and is covalently linked to an enzyme.

The advantage of this ELISA variant lies in the fact that (1) synthesis of a coating antigen is easier than that of a hapten-enzyme conjugate, (2) interfering matrix constituents are washed out before the addition of the enzyme-labeled second antibody, and (3) a prepared microtiter plate contains only slight amounts of a possibly toxic analyte [important when working with 2,3,7,8-tetrachlorodibenzodioxin (2,3,7,8-TCDD)]. After a washing step, sample and analyte-binding antibody are added. If larger amounts of free analyte are present in the added sample, only a few adsorptively bound analyte molecules react with AB 1. The AB – Ag complexes in solution are removed by a further washing step. The AB 1 molecules bound through hapten – BSA, which contain the analytical information, are detected by means of an enzyme-labeled second antibody. For this purpose, it is possible to use AB_2 that recognizes antibodies (IgG) from rabbits, from which AB 1 originates; these antibodies are commercially

available. Signal production is again inversely proportional to the concentration of analyte. Enzymes commonly used are horseradish peroxidase (E.C. 1.11.1.7) and alkaline phosphatase (E.C. 3.1.3.1).

For environmental analysis, enzyme-labeled tracers are predominantly used because increasingly lower limiting values demand maximum sensitivity. The reason for this choice is based on the enormous turnover rate in chromophore or fluorophore formation by enzymes. In the most common tests, photometric evaluation predominates.

There has been no lack of attempts to reduce the detection limit further by the use of liposomes filled with enzymes [56] or fluorophores [57]. Unfortunately, liposomes are not very stable, and fluorophores with the highest quantum yields are extremely hydrophobic, and subject to quenching by adsorption to vessel walls.

Although they depend on special evaluation techniques, immunoassays based on signal production via chemi- and bioluminescence are successful [58]. In clinical chemistry, this type of signal production has already gained acceptance. This is, however, not the case in environmental analysis because of the often uncontrolled interaction of matrix components with the luminescence-producing enzyme.

Efforts are being made to improve the sensitivity of immunoassays by using long-lived fluorophore labeling and highly intense light sources (e.g., lasers). Labeling with rare earth chelates, which have fluorescence decay times in the range 600 – 1000 µs, permits an especially easy separation of scattered light from the signal [59 – 62]. The high cost of UV laser systems is presently an obstacle to its wider acceptance. However, the first applications at 780 nm, using less expensive semiconductor diode lasers, have already been published [63].

The application of electrically produced time-resolved luminescence without light source represents an extremely interesting variant [64].

2.4.3.2. Sandwich Immunoassay

This type of noncompetitive heterogeneous immunoassay (Fig. 7) is encountered especially in the case of larger antigens with several epitopes [65]. The antibody, immobilized on the microtiter plate in the first step, has the function of binding the analyte (antigen) or standard in the solution. After a washing step, the second antibody, which recognizes the second epitope of the analyte, is added. After another washing step, a labeled (in this case with an enzyme) third antibody, which binds to AB 2, is added. Apart from the large number of steps involved, all of which hinder reproducibility, problems are also posed by the necessity of suppressing nonspecific adsorption in the presence of added reagents. Furthermore, the color formation is not inversely proportional to the analyte concentration. Sandwich immunoassay is not applicable to environmentally relevant haptens because it requires an analyte molecule of a certain minimum size.

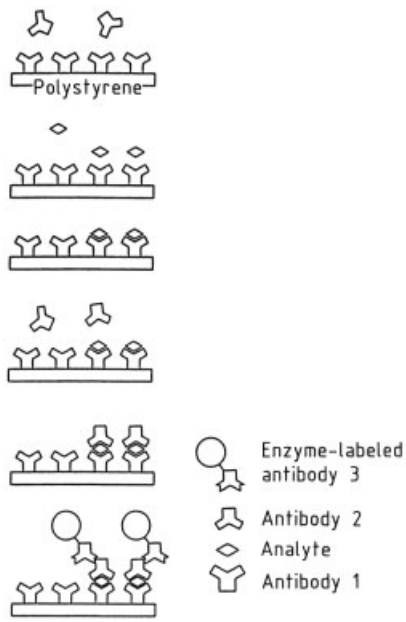

Figure 7. Principle of a noncompetitive immunoassay
AB 1 is a catcher antibody, and binds to the analyte. AB 2 should be able to bind a second (identical or different) antigenic epitope. AB 3 binds to foreign antibodies (of type AB 2) and is covalently linked to an enzyme.

2.4.3.3. Immunoassays using Immobilized Affinity Ligands

New immunoassay concepts were described in which immobilized affinity ligands are used in

unique flow-through systems (flow-injection immunoassay, FIIA). These systems are exclusively heterogeneous. Both immobilization of a specific antibody or a hapten on the solid support is possible. In most cases the scheme of the assay used is based on a sequential competitive enzyme immunoassay procedure [66, 67].

2.4.3.4. Immunoassays with Biotin and Avidin

Avidin (an egg yolk protein) exhibits an extremely strong affinity (ca. 10^{15} L/mol) for biotin (vitamin H). This irreversible interaction can be utilized in the design of heterogeneous immunoassays with biotinylated, antigen-specific antibodies [68, 69]. Biotinylation can be performed without interfering with antibody function. The antibodies are applied in the usual manner in a competitive immunoassay (Fig. 8). In the last step, enzyme-labeled avidin is added in excess. Since there are four possibilities for the binding of avidin to biotin, labeled biotin can also be used to occupy binding sites of the bound avidin to amplify the recognition signal reaction. Streptavidin can be used instead of avidin.

2.4.3.5. Magnetic Particle Immunoassay

An especially elegant heterogeneous immunoassay format depends on coupling antibodies to magnetic particles. The assay is carried out in a test tube. The conjugate of analyte and peroxidase (tracer) is first added to the sample or standard and mixed well. An aliquot of an AB – magnetic particle suspension is then pipetted into the test tube and again mixed well. After a preselected incubation time, the particles are precipitated by a magnet. After pouring off the supernatant, the AB – magnetic particles are resuspended and re-precipitated (washing step). The depth of color, which is inversely proportional to the concentration of analyte in sample, is then determined in the usual manner via an enzyme – substrate reaction. Applications in the field of pesticide analysis have been published [70 – 72].

Instead of the tedious and destructive coupling of antibodies to magnetic particles, the utilization of natural bacterial magnetic particles is also possible [73].

2.5. Immunoaffinity Techniques

An immunosorbens represents an excellent agent for the specific preconcentration of analytes with the aid of antibodies (immunoextraction). Until now, this technique has been used mainly in clinical chemistry. Environmental applications started in 1994 [74]. The appropriate antibody is immobilized on a precolumn and, after a concentration phase, subjected to a desorption step. It is favorable if the eluent used for this purpose is the same as that employed in the subsequent liquid chromatographic separation.

The repeated use of the precolumn sometimes poses problems because desorption must be forcibly effected by denaturation of antibody, and is often reversible. The substantial band broadening in the desorption step represents another problem. The bands can be sharpened only if the analyte can be reconcentrated on a subsequent reversed phase column with a relatively weak eluent (low proportion of organic solvent).

Alternatively, immunoselective desorption can be carried out with a "displacer" having a still higher affinity for the immobilized antibody. In the subsequent chromatography, however, a reduction of the analytical "window" owing to the structural similarity between the displacer and the analyte is to be expected.

In recent years, a rather new technique — the sol-gel technology — was applied for the preparation of new immunosorbents (Fig. 9). Using this rather simple method, the antibodies are encapsulated in a silica network. They retain their activity and, in some cases gain even higher stability, and can react with ligands that diffuse into the highly porous matrix. Several papers were reported on the entrapment of polyclonal and monoclonal antibodies against environmental pollutants [75 – 78].

Preconcentration with immunosorbents, followed by quantification with an immunoassay, is not recommended because no increase in selectivity is achieved by using the same antibody in the preconcentration and in the test.

Immunosorbents are of interest in trace analytical problems which require optimum detection strength and selectivity.

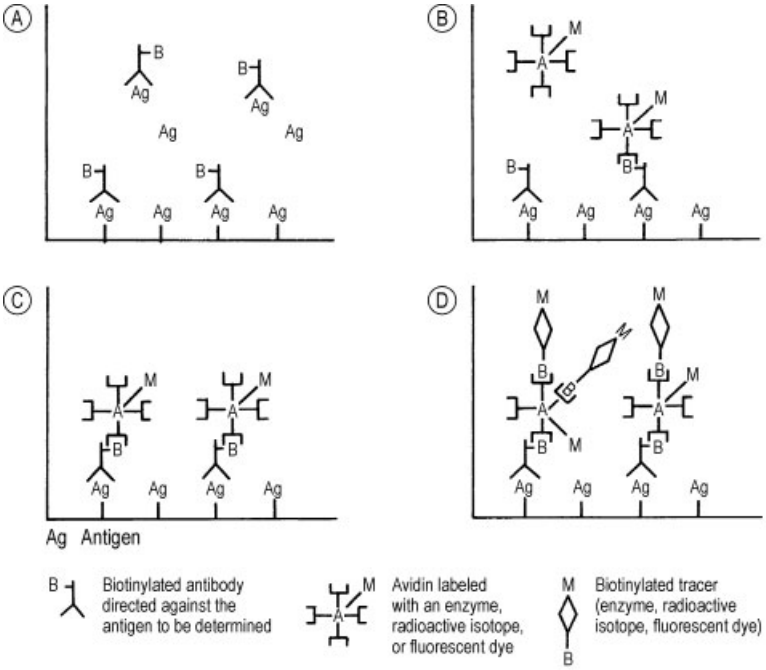

Figure 8. Principle of an immunoassay using the biotin – avidin system for labeling antibodies

2.6. Immunoassays on Test Strips and other Planar Structures

Dry tests or test strips (also called lateral flow devices or immunomigration strips) are popular as quick tests in clinical chemistry. Immunological tests are especially interesting because they can make use of the high selectivity of AB – Ag complex formation [79, 80]. The objective is to accommodate all the reagents required for a quantifiable test on a simple strip, filter, or capillary. It should be possible to dip the strip

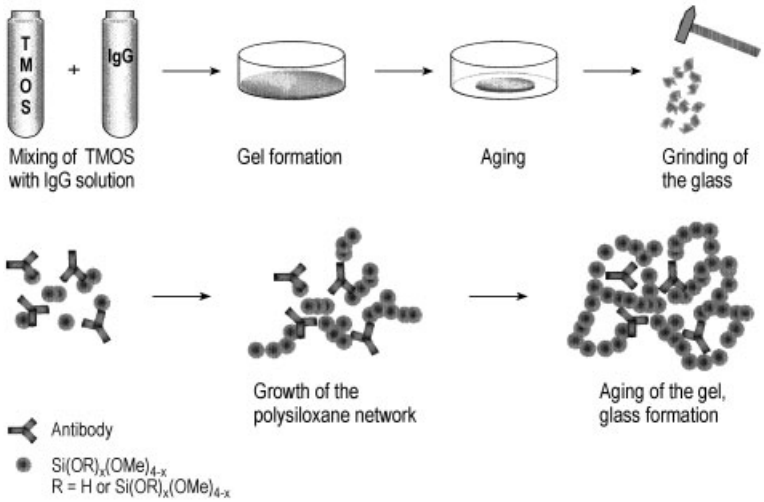

Figure 9. Encapsulation of antibodies in silicate glasses prepared by the sol – gel method. TMOS: tetramethoxysilane; IgG: immunoglobulin G

quickly into the liquid sample or to place a drop of sample on the carrier, and to determine the analyte concentration from the resulting depth of color or length of a colored band. Evaluation can be done visually or spots can be read-out by a pocket reflectometer. There should be tremendous possibilities for this test, especially for on-the-spot environmental analysis.

Two main variants are possible:

1) Accommodation of a homogeneous separation-free immunoassay on a dry matrix. This is possible for SLFIA and for ARIS. Other possibilities are labeling with gold particles or liposomes. The art lies in contact-free accommodation of all the reagents on the test strip, in such a way that no immune reaction takes place during storage.
2) The enzyme-channeling principle is especially suited to test strips. The test contains co-immobilized antibody and glucose oxidase. Two test strips must be used: one contains analyte-specific antibody, and the other is impregnated with peroxidase-specific antibody. The strips are dipped first into the sample solution and then into a color developer. The color developer contains a peroxidase tracer, glucose, and a chromogenic substrate (e.g., 4-chloro-1-naphthol). Channeling occurs by formation of H_2O_2 by glucose oxidase; this forms an insoluble dye on the surface of the test strip with 4-chloro-1-naphthol and peroxidase. In the case of the indicator test strip, the depth of color is influenced by competition with free analyte in the sample. The reference strip gives the standard value.

Enzyme immunochromatography (Fig. 10) represents an especially attractive variant [81]. A narrow strip of paper has the analyte-specific antibody immobilized on it. The sample is mixed with the tracer (analyte-labeled peroxidase and glucose oxidase) and the end of the test strip is dipped into this mixture. By capillary action, the solution is sucked up into the strip matrix. The sample is bound to the immobilized antibody as it travels up the strip, resulting in competition between the tracer and the analyte for the limited binding sites along the strip. A zone is formed which is proportional to the analyte concentration, and occupied by tracer molecules. After this minichromatography step, the entire strip is transferred to a color developing solution which consists of glucose and the peroxidase substrate, 4-chloro-1-naphthol.

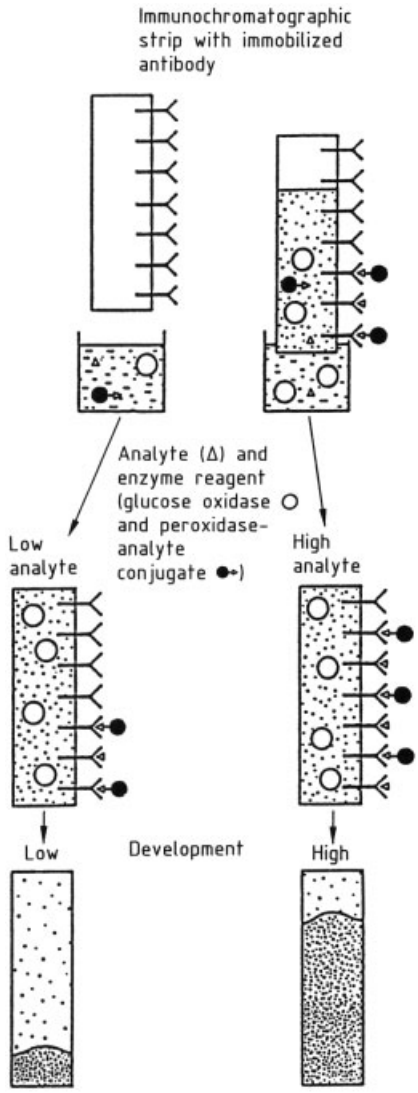

Figure 10. Principle of a test strip using enzyme immunochromatography

A strong color is developed, owing to enzyme channeling only in places where the peroxidase tracer is bound, in addition to the glucose that is present everywhere. Thus, the occupied length, which is proportional to the concentration of analyte, can be visually quantified as in a thermometer.

Another heterogeneous immunoassay, developed for field use [82, 83], employs immunofiltration (Fig. 11). Antibodies are immobilized in the middle of a nylon membrane. The first few drops placed on this membrane are immediately sucked down into a reservoir by capillary action. A few drops of washing buffer are then added, followed by AP- (alkaline phosphatase) labeled antibody to the analyte and, subsequently, a few drops of indoxyl substrate. As in the case of the sandwich immunoassay, the depth of color of the spot formed in the middle is linearly proportional to the analyte concentration; i.e., an ELISA is obtained for determination of aflatoxin [84], atrazin [85 – 87], carbaryl [88] or polychlorinated biphenyls [89].

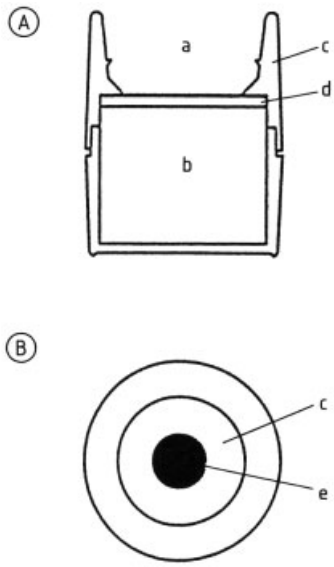

Figure 11. Cross section of the rapid immunofiltration test Hybritech ICON [72]
A) Side view; B) Top view
a) Fluid chamber; b) Absorbent reservoir; c) Membrane; d) Support disk; e) Area of immobilized antibody

2.7. Characterization and Interference

Immunochemical determination seems to be especially suited to the screening of large numbers of samples (e.g., water and soil analyses). Immunoassays are well suited for the analysis of the smallest sample volumes (e.g., in fog water analysis) because they allow analysis without sample preparation in a few microliters [90]. In critical applications, immunoassays do not compete with, but rather supplement conventional trace analytical methods.

In contrast to applications in clinical chemistry, matrix interference must be expected when immunoassays are applied to environmental samples. Disturbances can be caused not only by the presence of high molecular, organic water components (humic acids, lignosulfonic acids), but also by inorganic ions (Al^{3+}, Fe^{3+}, other transition metals) and oxidizing agents (Cl_2, O_3, ClO_2, H_2O_2) used in water treatment.

An exceptional feature of competitive heterogeneous tests, which are at present the only tests that are sufficiently sensitive, is evaluation by photometrical readout of the microtiter plates.

In multiple determinations, the median is calculated for each calibration point (Fig. 12). The margins of error correspond to simple standard deviations. The medians, or the singly determined values, are fitted to a four-parameter function by a simplex process; each curve has seven calibration points. The zero value is not included in the mathematical evaluation, but is used only for control purposes:

$$Y = \frac{(A-D)}{\left[1+\left(\frac{x}{C}\right)^{R}\right]}$$

where x is the concentration of analyte or hapten (µg/L); Y the extinction; A the maximal extinction (upper asymptote); D the minimal extinction (lower asymptote); C the test center point (µg/L); and R a curvature parameter.

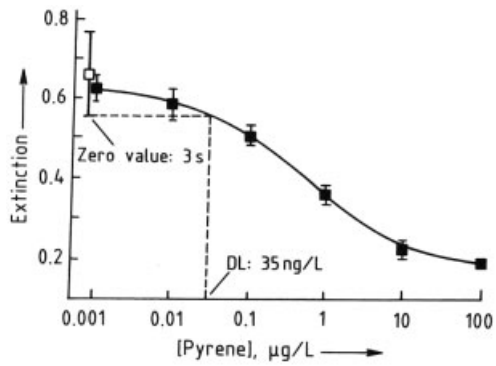

Figure 12. Typical calibration curve of a pyrene ELISA [50]
The detection limit DL is calculated from the blank extinction value plus a three-fold standard deviation: 35 ppt. Test center point, C=0.583 µg/L; Margin of error: 1 Δ; n=12.

The curves are normalized with respect to both A and D:

$$Y_{\mathrm{N}} = \frac{1}{\left[1+\left(\frac{x}{C}\right)^{B}\right]}$$

where Y_{N} approximately corresponds to the signal B/B_0 normalized to the zero value (B represents the extinction of the sample and B_0 the blank reading determined with the plate reader) and enforces convergence of all calibration curves for [analyte] $\rightarrow 0$ and [analyte] $\rightarrow\infty$. This standardization has the advantage that, unlike B/B_0, neither a zero value (associated with a relatively large error) nor an excess value (difficult to obtain owing to solubility or contamination problems) is required.

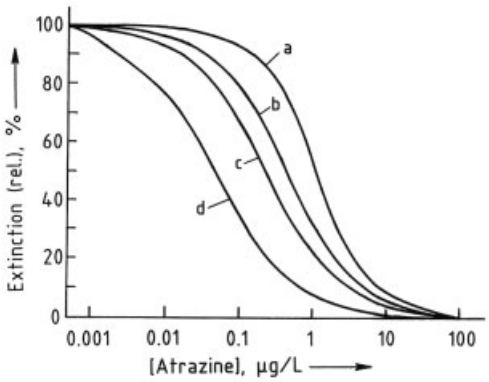

Figure 13. Calibration curves of an atrazine ELISA as a function of the incubation time of the tracer
a) 30 min; b) 10 min; c) 5 min;
Preincubation of unlabeled hapten: 60 min [91].

In the case of an ELISA, the tracer incubation time is important, as illustrated in Figure 13 for a triazine herbicide test [91]. The test center point C can be lowered by a factor of 25 by reducing the incubation time from 30 to 1 min; if it is increased to 40 h, a marked increase in the test center point C is obtained. Thus, the test covers atrazine concentrations of more than five powers of ten by simple variation of the incubation time. It is not known whether polyclonality or a kinetic effect (imbalance between antibody, hapten, and labeled hapten) is responsible for the observed shift of the calibration curves.

In immunoassays, antibodies react to a large number of chemically closely related species with varying sensitivity. This cross-sensitivity, also called cross-reactivity, is an indication of high structural similarity between haptens. Since cross-sensitivities cannot be exactly predicted, they must be experimentally determined for each antibody and each test design. Crossreactivities CR_{50} are normally determined by forming ratios of the 50 % values of the substances to be compared.

However, the actual influence of cross-reacting compounds can depend on their concentrations and, therefore, a single value for the cross-reactivity cannot contain all the information. For unknown samples, the influence of cross-reacting substances as a function of concentration should be determined in advance.

Water components can influence either the tracer (the tracer enzyme in ELISA), or the antibody, or the hapten – AB binding. The concentrations of substances that cause a 10 % inhibition of the horseradish peroxidase normally used as label in enzyme immunoassays are listed in Table 2.

Table 2. Concentrations that cause 10 % inhibition of horseradish peroxidase (N = 12)

Ion	Concentration, mg/L ($\pm$ 2 s)
CN^-	0.001 $\pm$ 0.0002
SCN^-	0.02 $\pm$ 0.004
ClO_3^-	0.03 $\pm$ 0.004
N_3^-	0.04 $\pm$ 0.005
ClO_4^-	0.02 $\pm$ 0.003
F^-	0.15 $\pm$ 0.016
Cu^{2+}	35.40 $\pm$ 4.04
Ni^{2+}	8.70 $\pm$ 0.96
Co^{2+}	17.30 $\pm$ 2.11
Al^{3+}	43.10 $\pm$ 2.97
Fe^{3+}	33.10 $\pm$ 4.74
Fe^{2+}	1.30 $\pm$ 0.08
$[Fe(CN)_6]^{3-}$	0.02 $\pm$ 0.002
$[Fe(CN)_6]^{4-}$	0.17 $\pm$ 0.02
Sn^{2+}	18.00 $\pm$ 2.11

Strongly oxidizing substances, such as ClO_3^- or complexing agents (F^-, SCN^-, CN^-) exert an inhibiting effect. The azide ion, which is used as a bacteriostatic agent in immunology, can also cause inhibition. If horseradish peroxidase is used as a tracer enzyme, NaN_3 can be employed only for the preservation of antibody solutions. No effect is exerted by the following ions: Hg^{2+}, $Cr_2O_7^{2-}$, Na^+, K^+, Li^+, Sr^{2+}, Mg^{2+}, Ba^{2+}, Ca^{2+}, Ag^+, Pb^{2+}, NH_4^+, Cl^-, NO_2^-, NO_3^-, and SO_4^{2-}.

In the presence of Ca^{2+}, the activity of horseradish peroxidase can be increased by up to 40 %, depending on the contact time and the Ca^{2+} concentration [15].

Sample components can interfere with antibodies in diverse ways. Evaluation is possible by means of calibration curves with the addition of various potentially interfering substances. Organic solvents (e.g., acetonitrile, methanol, and ethanol) in relatively low concentrations can affect the test. The maximal extinction falls increases on addition of solvents, the test center point is shifted by up to two powers of ten in the presence of 10 vol % of solvent [15]. In the immunological determination of trace substances (e.g., pesticides), the standard solutions should also be made up with the corresponding solvents.

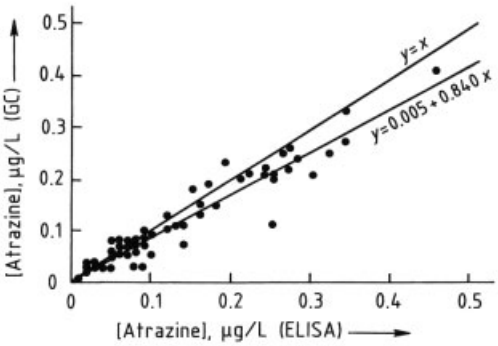

Figure 14. Correlation between two different analytical techniques (GC and ELISA) using atrazine as an example $n = 89$, $r = 0.956$ [15]

An extremely important point in process characterization is validation with, e.g., classical methods, such as gas chromatography (GC) Fig. 14) or liquid chromatography. Atrazine in surface waters was measured by means of ELISA and GC, and the results compared. The results correlate well, and confirm the theory that immunological methods can give reliable values in practice. The results of both methods are equivalent with a probability of 95 %. No false negative values were obtained. The same applies to the use of immunoassays on soil extracts. In comparison with GC analyses, ELISA usually gives higher analyte concentrations, which is put down to the effects of humic acid.

2.8. Future Immunological Applications and Techniques

Apart from the extension of the tests already available to include analytes measurable only by expensive and tedious instrumental analytical techniques, the spectrum of applications is certain to widen.

Until now, water and body fluid (serum, urine) represent the most common test matrices. Soil analysis has already begun to profit tremendously from immunoassays. Very few applications exist in air monitoring, even though active and passive measuring procedures could profit increasingly from immunological techniques for the monitoring of diffuse sources.

The further development of immunolocalizing analytical techniques in conjunction with state-of-the-art optoelectronics and luminescence microscopy also seems worthwhile for the recognition of colloidal or cell-bound haptens.

In chemical sensors, the adaptation of immunological recognition techniques to optical and electrochemical transducers has been under investigation for several years.

An intense discussion presently goes on the usefulness of miniaturized analytical techniques, which may be considered part of the emerging nanotechnology complex. Immunological technology, after hyphenation with chromatographic and electrophoretic techniques, may be transferred to a chip format, obtaining a complete and very efficient "analytical-lab-on-a-chip". If a library of antibodies (an antibody array) is available, exhibiting a different pattern of affinity for a series of structurally related compounds, an appropriate statistical analysis (multivariate statistical techniques, parametric models) has the power to turn the problem of cross-reactivity into an advantage (multianalyte immunoassays). The breakthrough of this technology will mainly depend on the availability of improved solid supports and antibodies, and on the development of more simple instruments [92 – 95].

Increasing activities in validation and standardization of immunochemical tests, initiated by federal government and regulatory agencies of several countries and well-recognized national and international organizations will lead to more transparency and uniformity in method development and evaluation. This will help that

the assays get rid of their image as "dubious biological tests", as they are sometimes called by analytical chemists [96]. Immunological methods cannot be assessed simply as "good" or "bad" but rather as suitable for an application or not.

3. References

1. J. P. Sherry, *Crit. Rev. Anal. Chem.* **23** (1992) 217.
2. J. P. Sherry, *Chemosphere* **34** (1997) 1011.
3. E. P. Meulenberg et al., *Environ. Sci. Technol.* **29** (1995) 553.
4. E. A. H. Hall (ed.): *Biosensoren*, Springer, Berlin 1990.
5. B. Eggins (ed.): *Biosensors*, Wiley and Teubner, Chichester 1996.
6. K. R. Rogers, A. Mulchandani (eds.): *Affinity Biosensors*, Humana Press, Totowa 1998.
7. R. Schmid (ed.): *Flow Injection Analysis (FIA)* Based on Enzymes or Antibodies, VCH Verlagsgesellschaft, Weinheim 1991.
8. P. Tijssen (ed.): *Practice and Theory of Enzyme Immunoassays*, Elsevier, Amsterdam 1985.
9. J. Peters, H. Baumgarten (eds.): *Monoklonale Antikörper*, Springer, Berlin 1990.
10. G. T. Hermanson (ed.): *Bioconjugate Techniques*, Academic Press, San Diego 1996.
11. G. Köhler, C. Milstein, *Nature* **256** (1975) 52.
12. H. A. Lee, M. R. A. Morgan, *Trends in Food Sci. Technol.* **4** (1993) 129.
13. K. Kramer, *Anal. Letters* **31** (1998) 67.
14. C. W. Bell et al., *ACS Symp. Ser.* **586** (1995) 413.
15. T. Ruppert, L. Weil, R. Niessner, *Vom Wasser* **78** (1991) 387.
16. P. Albro et al., *Toxicol. Appl. Pharmacol.* **50** (1979) 137.
17. W. F. W. Stöcklein, A. Warsinke, F. W. Scheller, *ACS Symp. Ser.* **657** (1997) 373.
18. C. Keuchel, L. Weil, R. Niessner, *SPIE* **1716** (1992) 44.
19. R. Schneider, L. Weil, R. Niessner, *Int. J. Environ. Anal. Chem.* **46** (1992) 129.
20. R. Yalow, S. Berson, *J. Clin. Invest.* **39** (1960) 1157.
21. T. Chard (ed.): *An Introduction to Radioimmunoassay and Related Techniques*, Elsevier, Amsterdam 1990.
22. P. Masson et al. in J. Langone, H. van Vunakis (eds.): *Immunochemical Techniques*, Part C, Academic Press, New York 1981, p. 106.
23. Y. V. Lukin et al., *ACS Symp. Ser.* **657** (1997) 97.
24. S. Obata et al., *Anal. Sci. Suppl.* **7** (1991) 1387.
25. H. Weetall, A. Gaigalas, *Anal. Lett.* **25** (1992) 1039.
26. E. Engvall in H. van Vunakis, J. Langone (eds.): *Methods in Enzymology*, vol. **70**, Academic Press, New York 1970, p. 419.
27. D. Morris et al., *Anal. Chem.* **53** (1981) 658.
28. D. L. Morris, R. T. Buckler, *Meth. Enzymol.* **92** (1983) 413.
29. E. Soini, I. Hemmilä, *Clin. Chem. (Winston-Salem, N.C.)* **25** (1979) 353.
30. T. Ngo et al., *J. Immunol. Methods* **42** (1981) 93.
31. R. Niessner et al.: "Application of Time-Resolved Laser-Induced Fluorescence for the Immunological Determination of Pesticides in the Water Cycle," in B. Hock, R. Niessner (eds.): *Immunological Detection of Pesticides and their Metabolites in the Water Cycle,* VCH Verlagsgesellschaft, Weinheim, in press.
32. S. A. Eremin et al., *ACS Symp. Ser.* **586** (1995) 223.
33. S. G. Reeves, S. T. A. Siebert, R. A. Durst, *ACS Symp. Ser.* **586** (1995) 210.
34. J. Rittenburg et al. in M. Vanderlaan et al. (eds.): *Immunoassays for Trace Chemical Analysis*, vol. **451**, American Chemical Society, Washington D.C. 1990, p. 28.
35. J. Brady et al.: "Biochemical Monitoring for Pesticide Exposure," *ACS Symp. Ser.* **382** (1989) 262.
36. J.-M. Schläppi, W. Föry, K. Ramsteiner, "Immunochemical Methods for Environmental Analysis," *ACS Symp. Ser.* **442** (1990) 199.
37. H. Newsome, P. Collins, *J. Assoc. Off. Anal. Chem.* **70** (1987) 1025.
38. Q. Li et al., *J. Agric. Food Chem.* **39** (1991) 1537.
39. S. Wie, B. Hammock, *J. Agric. Food Chem.* **32** (1984) 1294.
40. R. Dreher, B. Podratzki, *J. Agric. Food Chem.* **36** (1988) 1072.
41. F. Jung et al., *J. Agric. Food Chem.* **37** (1989) 1183.
42. H. Scholz, B. Hock, *Anal. Lett.* **24** (1991) 413.
43. S. Kreissig et al., *Anal. Lett.* **24** (1991) 1729.
44. Q. Li et al., *Anal. Chem.* **61** (1989) 819.
45. B. Riggle, B. Dunbar, *J. Agric. Food Chem.* **38** (1990) 1922.
46. P. Ulrich, R. Niessner, *Fres. Environ. Bull.* **1** (1992) 22.

47. C. Hall et al., *J. Agric. Food. Chem.* **37** (1989) 981.
48. A. Roda et al., *Environ. Technol. Lett.* **12** (1991) 1027.
49. D. Knopp et al., *ACS Symp. Ser.* **657** (1997) 61.
50. K. Meisenecker, D. Knopp, R. Niessner, *Analytical Methods and Instrumentation* **1** (1993) 114.
51. C. Keuchel, L. Weil, R. Niessner, *Anal. Sci.* **8** (1992) 9.
52. J. Van Emon, R. Gerlach, *Bull. Environ. Contam. Toxicol.* **48** (1992) 635.
53. B. Watkins et al., *Chemosphere* **19** (1989) 267.
54. D. Wylie et al., *Anal. Biochem.* **194** (1991) 381.
55. S. Wie, B. Hammock, *Anal. Biochem.* **125** (1982) 168.
56. T.-G. Wu, R. Durst, *Mikrochim. Acta* 1990, 187.
57. S. Choquette, L. Locascio-Brown, R. Durst, *Anal. Chem.* **64** (1992) 55.
58. A. Tsuji, M. Maeda, H. Arakawa, *Anal. Sci.* **5** (1989) 497.
59. E. Diamandis, *Clin. Biochem.* **21** (1988) 139.
60. E. Diamandis, T. Christopoulos, *Anal. Chem.* **62** (1990) 1149 A.
61. T. Christopoulos, E. Diamandis, *Anal. Chem.* **64** (1992) 342.
62. Y.-Y. Xu et al., *Analyst (London)* **116** (1991) 1155.
63. T. Imasaka et al., *Anal. Chem.* **62** (1990) 2405.
64. J. Kanakare et al., *Anal. Chim. Acta* **266** (1992) 205.
65. Hybritech, US 4 376 110, 1983 (G. S. David, H. E. Greene).
66. B. B. Kim et al., *Anal. Chim. Acta* **280** (1993) 1991
67. P. M. Kraemer et al., *ACS Symp. Ser.* **657** (1997) 153.
68. M. Wilchek, E. A. Bayer, *Meth. Enzymol.* **184** (1990) 14.
69. F. Eberhard: "Entwicklung von Immunoassays und Markierungsmethoden für die Bestimmung von Schimmelpilzgiften mit immunooptischen Biosensoren," Dissertation, Universität Hannover 1991.
70. F. Rubio et al., *Food Agric. Immun.* **3** (1991) 113.
71. T. Lawruk et al., *Bull. Environ. Contam. Toxicol.* **48** (1992) 643.
72. J. A. Itak et al., *ACS Symp. Ser.* **657** (1997) 261.
73. N. Nakamura, T. Matsunaga, *Anal*. Sci. **Suppl. 7** (1991) 899.
74. G. S. Rule, A. V. Mordehai, J. Henion, *Anal. Chem.* **66** (1994) 230.
75. J. Zuehlke, D. Knopp, R. Niessner, *Fresenius J. Anal. Chem.* **352** (1995) 654.
76. A. Turniansky et al., *J. Sol – Gel Sci. Technol.* **7** (1996) 135.
77. M. Cichna et al., *Chem. Mater.* **9** (1997) 2640.
78. T. Scharnweber, D. Knopp, R. Niessner, *Field Anal. Chem. Technol.* **4** (2000) 43.
79. D. Morris, D. Ledden, R. Boguslaski, *J. Clin. Lab. Anal.* **1** (1987) 243.
80. D. Litman in T. Ngo, H. Lenhoff (eds.): *Enzyme-Mediated Immunoassay*, Plenum Press, New York 1985, p. 155.
81. R. Zuck et al., *Clin. Chem. (Winston-Salem, N.C.)* **31** (1985) 1144.
82. G. Valkiro, R. Barton, *Clin. Chem. (Winston-Salem, N.C.)* **31** (1985) 1427.
83. J. Rittenburg et al.: "Immunochemical Methods for Environmental Analysis," *ACS Symp. Ser.* **442** (1990) 28.
84. E. Schneider et al., *Food Agric. Immun.* **3** (1991) 185.
85. C. Morber, L. Weil, R. Niessner, *Fres. Environ. Bull.* **2** (1993) 151.
86. C. Keuchel, R. Niessner, *Fresenius J. Anal. Chem.* **350** (1994) 538.
87. A. Dankwardt, B. Hock, *Biosens. Bioelectron.* **8** (1993) XX.
88. S. Morias, A. Maquieira, R. Puchades, *Anal. Chem.* **71** (1999) 1905.
89. M. Del Carlo et al. in B. Hock et al. (eds.): *Biosensors and Environmental Diagnostics*, Chap. 2, Teubner Verlag, Stuttgart 1998.
90. F. Trautner, K. Huber, R. Niessner, *J. Aerosol Sci.* **23** (1992) S 999.
91. M. Weller, L. Weil, R. Niessner, *Mikrochim. Acta* **108** (1992) 29.
92. R. Ekins, F. Chu, E. Biggart, *Anal. Chim. Acta* **227** (1989) 73.
93. A. Brecht, G. Gauglitz, *Anal. Chim. Acta* **347** (1997) 219.
94. M. G. Weller et al., *Anal. Chim. Acta* **393** (1999) 29.
95. H. Kido, A. Maquieira, B. D. Hammock, *Anal. Chim. Acta* **411** (2000) 1.
96. J. Rittenburg, J. Dautlick, *ACS Symp. Ser.* **586** (1995) 301.

Basic Principles of Chromatography

GEORGES GUIOCHON, Department of Chemistry, University of Tennessee, Knoxville, TN, 37996-1600 , United States, Division of Chemical and Analytical Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, 37831-6120, United States

1. Introduction 2
2. Historical Development 3
3. Chromatographic Systems 4
3.1. Phase Systems Used in Chromatography 4
3.2. Methods of Implementing Chromatography 5
4. Theory of Linear Chromatography 6
4.1. Plate Models 6
4.1.1. The Craig Model 6
4.1.2. The Martin – Synge Model 7
4.2. Statistical Models 7
4.3. Mass-Balance Models 8
4.3.1. The Equilibrium – Dispersive Model . 8
4.3.2. Lumped Kinetic Models 8
4.3.3. The Golay Equation 9
4.4. The General Rate Model 9
4.5. Moment Analysis 9
4.6. Sources of Band Asymmetry and Tailing in Linear Chromatography 10
5. Flow Rate of the Mobile Phase . . . 10
5.1. Permeability and Porosity of the Packing 11
5.1.1. Column Porosity 11
5.1.2. Column Permeability 11
5.2. Viscosity of the Mobile Phase 11
5.2.1. Viscosity of Gases 11
5.2.2. Viscosity of Liquids 11
6. The Thermodynamics of Phase Equilibria and Retention 12
6.1. Retention Data 12
6.1.1. Absolute Data 12
6.1.2. Relative Data 12
6.2. Partition Coefficients and Isotherms 13
6.3. Gas – Solid Adsorption Equilibria . 13
6.4. Liquid – Solid Adsorption Equilibria 14
6.4.1. Normal-Phase Chromatography 14
6.4.2. Reversed-Phase Chromatography . . . 14
7. Band Broadening 15
7.1. Sources of Band Broadening 15
7.2. The Plate-Height Equation 15
7.3. Resolution Between Peaks 16
7.4. Optimization of Experimental Conditions 16
7.5. Instrumental Requirements 17
7.5.1. Injection Systems 17
7.5.2. Connecting Tubes 17
7.5.3. Detectors 18
7.5.4. Instrument Specifications 18
8. Qualitative Analysis 18
8.1. Comparisons of Retention Data . . . 18
8.2. Precision and Accuracy in the Measurement of Retention Data 19
8.3. Retention and Chemical Properties 19
8.4. Selective Detectors 20
8.4.1. Simple Selective Detectors 20
8.4.2. Chromatography – Mass Spectrometry (. 20
8.4.3. Chromatography – IR Spectrometry . 21
9. Quantitative Analysis 21
9.1. Sampling Problems in Chromatography 22
9.2. Measurement of Peak Area 22
9.3. Calibration Techniques 22
9.4. Sources of Error in Chromatography 23
10. Theory of Nonlinear Chromatography 23
10.1. The Ideal Model of Chromatography 23
10.2. The Equilibrium – Dispersive Model 24
10.3. Moderate Column Overloading . . . 25
10.4. Preparative Chromatography 25
11. Reference Material 25
11.1. Journals 25
11.2. Books 25
12. References 26

Abbreviations

a initial slope of the isotherm
C, C_{m} mobile-phase concentration
C_{s} stationary-phase concentration
d_{f} film thickness
d_{p} particle size (diameter)
D_{a} axial dispersion coefficient

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_155

D_l diffusion coefficient, stationary (liquid) phase
D_m diffusion coefficient, mobile (gas) phase
F phase ratio
F_v mobile-phase flow rate
h reduced plate height
H height equivalent to a theoretical plate (HETP)
k_0 column permeability
k', k'_0 retention factor
K Henry constant
L column length
M_n nth moment of a statistical distribution
N, N_c number of plates
p partial pressure
P probability
P^0 vapor pressure
q Langmuir isotherm
R fraction of sample in the mobile phase; resolution
S_a adsorbent surface area
S_i detector response factor
t_R retention time
u mobile-phase velocity
v reduced velocity
v_m volume per plate of the mobile phase
v_s volume per plate of the stationary phase
V_g specific retention volume
V_G volume of gas
V_l volume of liquid
V_m mobile-phase volume
V_N correlated retention volume
V_R retention volume
α relative retention
γ^∞ activity coefficient at infinite dilution
ε_T total packing porosity
η mobile-phase viscosity
$\mu_n, \bar{\mu}_n$ absolute and central moments
Φ fractional loss of efficiency
σ standard deviation
τ time variance

1. Introduction

Chromatography is the most powerful separation technique available. For reasons that will be made clear later, it is easy to implement chromatographic techniques with small samples, and to carry out separations on an analytical scale. Separated components are usually detected online, and sometimes they are also characterized, after which they are generally discarded. In preparative applications, the purified components are collected for further investigation, although new difficulties may be encountered in an attempt to scale-up the equipment and operate at high concentrations.

Chromatography is a separation method based on differences in equilibrium constants for the components of a mixture placed in a diphasic system. A chromatographic system is one in which a fluid mobile phase percolates through a stationary phase. The stationary phase is often a bed of nonconsolidated particles, but this is not essential. All that is in fact required is two phases in relative motion and excellent contact between these phases so that concentrations in the stationary phase are always very near their equilibrium values. For example, a tubular column through which the mobile phase flows and whose walls have been coated with a layer of stationary phase is a most suitable implementation of chromatography. In all cases the mobile phase is responsible for transporting the sample components through the stationary phase. The velocity of each component, hence its residence time, depends on the both mobile-phase velocity and the distribution equilibrium constant for that component. At least in principle, proper choice of the two phases constituting a chromatographic system permits selective adjustment of the relative migration rates of the mixture components and of the extent of their eventual separation.

The mobile phase is a fluid. The main criteria for its selection are:

1) Good solubility for the analytes
2) Low viscosity

All types of fluids have been used, but the chief types of chromatography are gas, dense-gas (more commonly called supercritical-fluid), and liquid chromatography. In *gas chromatography* (GC; → Gas Chromatography), a quasi-ideal gas is used; pressures are kept low, analyte solubility in the mobile phase depends only on its vapor pressure, and the role of the mobile phase is almost purely mechanical, with minor corrections due to nonideal behavior of the gas phase [1]. Retention adjustment is accomplished primarily by changing the temperature. With denser gases, an analyte's fugacity differs markedly from its vapor pressure, and analyte solubility

depends strongly on the mobile-phase density as well as the temperature. Thus, retention adjustment in supercritical-fluid chromatography can be effected by changing not only the temperature but also the average pressure. In both gas and supercritical-fluid chromatography, the *selectivity* (i.e., the ability to achieve a separation) is relatively insensitive to temperature and density. In *liquid chromatography* (LC), the mobile phase is a mixture of solvents (→ Liquid Chromatography). Retention adjustments and selectivity changes are accomplished by changing the mobile-phase composition and introducing a variety of strong solvents or additives.

The stationary phase may be a solid (i.e., an adsorbent, with a relatively large specific surface area and very accessible pore channels) or a liquid coated on a solid support (to avoid stripping of the stationary phase by the streaming action of the mobile phase). Intermediate systems include immobilized polymers or surface-bonded shorter chemical species (e.g., $C_{18}H_{37}$) of widely varying chain length and molecular mass. Gas – solid, gas – liquid and liquid – solid chromatography are the most popular chromatographic methods. Ion-exchange resins in conjunction with ionic solvents, gels with pores comparable in size to the analyte molecules, and stationary phases containing suitable complexing agents or groups have also served as the basis for important methods, producing selectivities very different from those observed with more classical molecular interactions.

Another useful distinction involves the mechanism by which the mobile phase is transported through the stationary phase. The most popular approach is forced convection by pressurization, based on cylinders of compressed gases in gas chromatography, or a mechanical pump in liquid chromatography. In both cases the stationary phase is placed inside a tube, the *chromatographic column*. Other methods are available as well. For example, in thin layer chromatography (→ Thin Layer Chromatography) capillary forces suffice to draw a liquid stream through a dry, porous bed in the form of a paper sheet or a thin layer of adsorbent coated on a glass, metal, or plastic sheet. More recently, electro-osmosis has been used to force a liquid stream through a column [2]. Although the method is quite similar to standard column liquid chromatography, the resulting separations may be markedly superior.

2. Historical Development

It is not possible to review here the entire history of chromatography and do justice to all the scientists who have contributed to designing and developing the dozens of major separation techniques based on the principle of chromatography. The history of the origin of chromatography has been recounted by SAKODINSKY [3] and ETTRE [4]. Numerous other publications recall landmark contributions in the field.

Although a number of early publications introduced certain ideas related to chromatography, it is clear that the method was conceived and developed by the Russian chemist TSWETT at the beginning of this century [5–7]. TSWETT is best described as a biophysicist. He had a remarkably clear understanding of the interactions between the various phenomena involved in chromatography – some thermodynamic and others kinetic – and of the technical problems that would need to be solved to turn the chromatographic principle into a useful group of analytical methods. Unfortunately, living during the very early stages of the high-technology age, in a country soon to be ravaged by civil war, he was unable to succeed. His ideas remained untried until a group of German chemists who were informed of TSWETT's early work through WILLSTÄTTER [3] succeeded in using chromatography for the extraction and purification of plant pigments.

The original work of TSWETT is most interesting to read – at least the fragments available in the modern literature [3], [4], [8]. Many of the modern problems of high-performance liquid chromatography are already addressed in these early writings. TSWETT knew that small particles should be used for optimum performance, but he had no way to force a mobile-phase stream under pressure through a column. It is interesting also to note that the initial difficulties KUHN's group had to solve in developing their analytical schemes were due to an isomerization of carotenoid derivatives catalyzed by the silica gel used as stationary phase [3], [4]. TSWETT had deliberately used calcium carbonate for these separations, reserving silica gel for the separation of more stable compounds [7].

A. J. P. MARTIN has been associated with several critical developments in chromatography. His Nobel prize paper with SYNGE [9] represents the origin of both partition and thin layer chromatography (TLC). Originally carried out with paper sheets, TLC evolved later under the leadership of STAHL in the direction of consolidated layers of fine particles. Later, MARTIN, GORDON and CONSDEN [10] published the first two-dimensional chromatograms. Here a sample is introduced at one corner of a large, square paper sheet, and two successive developments are conducted in orthogonal directions, using two different solvent mixtures as mobile phases. This method leads to a dramatic increase in separation power. Finally, in cooperation with JAMES, MARTIN invented gas – liquid chromatography [11], a technique ideally suited to the incorporation of many developments in modern instrumentation that occurred at that time. Gas – liquid chromatography led first to a major new instrumentation technique, then to the rejuvenation of conventional liquid chromatography, and finally to the more recent development of capillary-zone electrophoresis.

3. Chromatographic Systems

Chromatography is based on phase-equilibrium phenomena. The components of the analyte sample are caused to equilibrate between two phases, a mobile phase and a stationary phase. Because the mobile phase percolates through the stationary phase, rapid mass transfer takes place, and the mobile phase carries the components through the column to a detector. The velocity of this transfer is related to the equilibrium constant. Hence, only compatible combinations of mobile and stationary phases can be used in practice.

3.1. Phase Systems Used in Chromatography

The mobile phase must be a fluid: A gas, a dense gas, or a liquid, where the rate of mass transfer through the mobile phase, characterized by the diffusion coefficient, decreases in the order listed. To compensate for this marked decrease, finer and finer particles are used as the mobile phase becomes more dense. The practical lack of compressibility of liquids supports this trade-off, since small particles demand higher pressures. However,viscous liquids (e.g.,glycerol,oligomers, or concentrated solutions of polymers) are excluded, because the associated pressure requirements would be unrealistic.

The stationary phase must be compatible with the mobile phase. Lack of solubility of the stationary phase in the mobile phase and a large interfacial area between the two phases are the two major requirements. Most adsorbents are applicable to any of the three types of mobile phase, giving rise, for example, to both gas – solid and liquid – solid chromatography. The specific surface area of the absorbent must be large, not only to enhance column efficiency but also to ensure sufficient retention [12]. This surface must be highly homogeneous as well to avoid the presence of active sites on which the adsorption energy is exceptionally high leading to excessive retention, and a poor band profile. Most of the adsorbents for chromatography have been developed to such a point that they display the required properties; alumina, silica, molecular sieves, and activated or graphitized carbons are the most popular.

Silica comes in two broad types: *Pure* or *underivatized silicas* (all of which have very similar properties, but among which chromatographers can recognize more varieties than there are commercial brands) and the *chemically bonded silicas* [13]. Virtually every conceivable type of chemical group – from alkyl derivatives to sugars and peptides – has been bonded to silica, leading to a large number of phases, each with its specific properties and applications. In liquid – solid chromatography, two types of systems are distinguished: *Normal-phase* systems based on a polar stationary phase and a nonpolar mobile phase (e.g., silica and dichloromethane), and *reversed-phase* systems, in which the stationary phase is nonpolar and the mobile phase polar (e.g., C_{18} silica and water – methanol solutions).

Liquids can also be used as the stationary phase in chromatography, but they must be immiscible with the mobile phase. In gas – liquid chromatography this means that the stationary phase must have a low vapor pressure to avoid its

premature depletion by the mobile phase. Convection would be highly detrimental to the effectiveness of a separation, and this is avoided by coating the liquid phase on an inert solid. The support is usually a pulverulent solid consisting of porous particles. To avoid significant contributions from adsorption (i.e., mixed retention mechanisms), the solid should have a low specific surface area, but it should also be highly porous to permit the use of a large amount of liquid. Alternatively, a stationary liquid can be coated on a nonporous support (e.g., glass beads, or the inner wall of an empty tube). Considerable effort has been devoted to the development of reproducible methods for coating the inner wall of a quartz tube with a homogeneous, stable, thin film of liquid.

While gas – liquid chromatography has become the preferred implementation of gas chromatography because of the excessive adsorption energy associated with gas – solid equilibria, for most compounds but gases, liquid – solid chromatography is the preferred implementation of liquid chromatography. It is difficult to find a stationary liquid that is insoluble in the liquid mobile phase but at the same time not so different from this phase that the partition coefficients are either negligible or nearly infinite. The stability of a liquid coated on an inert support is low, and often the stationary phase is quickly stripped from the support in the form of a dilute emulsion.

Other important stationary phases for liquid chromatography are ion-exchange resins (*ion-exchange chromatography*, which uses anionic or cationic resins) and porous solids that do not adsorb the analytes but instead have pores so small that access to a fraction of the pore volume by some of the analyte molecules is more or less restricted depending on molecular size (*size-exclusion chromatography*).

Finally, chemical groups capable of forming charge transfer complexes or other types of labile complexes with specific components of a sample have also been used successfully as stationary phases. *Affinity chromatography* is one implementation of this concept, in which complex formation between a specific antibody and one analyte results in both exceptional selectivity for a particular component and an inherent lack of flexibility, two striking characteristics of this method.

3.2. Methods of Implementing Chromatography

There are three methods of implementing chromatography: Elution, displacement, and frontal analysis. An immense majority of analytical applications are carried out on the basis of elution. Equilibrium isotherms are often measured using frontal analysis. Finally, preparative separations, though most often accomplished by overloaded elution, also lend themselves to displacement. This has the advantage of producing more concentrated fractions, but it requires special development, choice of an appropriate displacer, and the extra step of column regeneration between the processing of successive batches.

In *elution*, a pulse of sample is injected into the mobile-phase stream before it enters the column. Under the linear conditions most often associated with analytical applications, each pulse of sample component behaves independently, migrating along the column at a velocity proportional to the mobile-phase velocity but dependent also on the equilibrium coefficient between the two phases. If the experimental conditions are properly chosen, pulses of the mixture components arrive separately at the column exit, and the chromatogram consists of a series of peaks, one for each component.

In *frontal analysis*, a stream of a solution of the sample in the mobile phase is abruptly substituted for the mobile-phase stream at the beginning of the experiment. If the components are present at very low concentrations, the front of each component "step" moves at the same velocity as the pulses in an elution experiment, resulting in a "staircase" in which the profile of each individual step can be regarded as an error function. However, in most applications of frontal analysis the component concentrations are in fact rather high, and the isotherms are no longer linear. Accordingly, they are not independent of each other, but are competitive; i.e., the concentration a given compound adsorbed at equilibrium depends on the concentrations of all the components present. The propagation of the various component steps results in breakthrough curves that are steep, and the concentrations of all the components in the eluent change when any one of them breaks through. As a con-

sequence, frontal analysis can be used only to purify the first eluted component of a mixture, or to measure single- component or competitive (mainly binary) isotherms.

In *displacement* chromatography, a pulse of mixture is injected, and this is followed by a step of a single component called a *displacer*, which is adsorbed more strongly than any of the mixture components. After a certain period of time, during which the profiles of each pulse become reorganized, an isotachic pulse train is formed. In the isotachic train, each component forms its own concentration "boxcar," with a height that depends on the component and displacer isotherms as well as on the displacer concentration, and a width proportional to the amount of the corresponding component in the sample [14], [15].

4. Theory of Linear Chromatography

One can distinguish between two basic types of assumption common to all theoretical efforts to understand chromatography [16]. First, it may be assumed that the equilibrium isotherm is linear; alternatively, one might acknowledge that, under the experimental conditions selected, this isotherm is not linear. One can thus study models of either linear or nonlinear chromatography. Analytical applications almost always involve small or dilute samples, and under these conditions any curvature in the isotherm near the origin can be ignored and the isotherm can be regarded as linear. In preparative applications – more generally, in any case of concentrated solutions – this assumption is no longer tenable, and it becomes necessary to consider nonlinear chromatography (see Chap. 10).

Similarly, a decision must be made whether or not to take into account the influence on band profiles of such phenomena as axial dispersion (dispersion in the direction of the concentration gradient in the column) and resistance to mass transfer (i.e., the fact that equilibration between mobile and stationary phases is not instantaneous). These phenomena are responsible for the finite efficiency of actual columns. Neglecting them and assuming the column to have infinite efficiency leads to a model of *ideal* chromatography. Taking them into account results in one of the models of *nonideal* chromatography.

The combination of these two types of consideration leads to four theories of chromatography. The first, *linear, ideal chromatography*, is trivial: each component pulse moves without changes in its profile, and its motion is independent of other pulses. In *linear, nonideal chromatography* one studies the influence of kinetic phenomena on band profiles, assuming each component pulse to move independently of all the others, and at a constant velocity. The chromatogram of a mixture in this case is the sum of the chromatograms for each independent component. This is no longer true in *ideal, nonlinear chromatography*. Here one studies the influence of nonlinear phase-equilibrium thermodynamics on the band profile. The result depends on the amount of each component present in the mixture, and individual elution profiles for the various components cannot be obtained independently (see Chap. 10). Nevertheless, both linear, nonideal and ideal, nonlinear chromatography provide relatively simple solutions. In *nonlinear, nonideal chromatography* it is necessary to take into account both finite column efficiency and a nonlinear, nonindependent behavior of the isotherms, and only numerical solutions are possible.

4.1. Plate Models

These models assume that the chromatographic column can be divided into a series of a finite number of identical plates. Each plate contains volumes v_{m} and v_{s} of the mobile and stationary phases, respectively. The sample is introduced as a solution of known concentration in the mobile phase used to fill the required number of plates. Plate models are essentially empirical, and cannot be related to first principles. Depending upon whether one assumes continuous or batch operation, two plate models can be considered: The Craig model and the Martin and Synge model.

4.1.1. The Craig Model

In this model [17], mobile phase is transported through the column by withdrawing it from the last plate, moving the remaining mobile phase

from each plate to the next one in succession, and refilling the first plate. The process is repeated as often as necessary to elute the material introduced. It is easily shown that the fraction R of the sample in the mobile phase is expressed by

$$R=\frac{vC_{\mathrm{m}}\mathrm{m}}{vC_{\mathrm{m}}+vC_{\mathrm{s}}\mathrm{sm}}=\frac{1}{1+Fa}=\frac{1}{1+k_0'} \tag{1}$$

where C_{m} and C_{s} are the equilibrium concentrations in the mobile and stationary phases, respectively. F is the phase ratio, a is the initial slope of the isotherm, and k_0' is the retention factor. After the moving process has been repeated r times, the probability P of finding a molecule in the plate of rank l is

$$P_{l,r}=\frac{r!}{l!\,(r-l)!}R^l\,(1-R)^{r-l} \tag{2}$$

and the elution profile (at the exit of the last plate, of rank N_{c}) is

$$f\,(r)=R\frac{r!}{N_{\mathrm{c}}!\,(r-N_{\mathrm{c}})!}R^{N_{\mathrm{c}}}\,(1-R)^{r-N_{\mathrm{c}}} \tag{3}$$

When N_{c} exceeds about 50 plates, this profile cannot be distinguished from a Gaussian profile with a retention volume

$$V_{\mathrm{R}}=V_{\mathrm{m}}\left(1+k_0'\right) \tag{4}$$

where $V_{\mathrm{m}}=N_{\mathrm{c}}\,v_{\mathrm{m}}$, and a standard deviation

$$\sigma_v=V_{\mathrm{R}}\sqrt{\frac{1-R}{N_{\mathrm{c}}}} \tag{5}$$

4.1.2. The Martin – Synge Model

This is a continuous plate model [9]. The mobile phase is transferred from one plate to the next, and each plate is regarded as a perfect mixer. Integration of the differential mass balance in every plate leads to a Poisson distribution inside the column, and to the elution profile:

$$f\,(v)=\frac{N}{V_R}\mathrm{e}^{-\frac{Nv}{V_R}}\left(\frac{Nv}{V_R}\right)^{(N-1)}\frac{1}{(N-1)!} \tag{6}$$

As the plate number N increases, just as in the case of the Craig model the profile tends rapidly toward a Gaussian profile:

$$f\,(v)=\frac{1}{\sigma_v\sqrt{2\pi}}\exp\left[-\frac{(v-V_R)^2}{2\sigma_v^2}\right] \tag{7}$$

with a retention volume and standard deviation given by

$$V_R=N\,(v+av\mathrm{sm}) \tag{8a}$$

$$\sigma_v=(v+av\mathrm{sm})\sqrt{N} \tag{8b}$$

Note that, although both models lead to the same profile, the resulting relationships between number of plates in the column and standard deviation differ. The conventional plate number as defined in chromatography is equal to N for the Martin and Synge model, and to $N_{\mathrm{c}}\,(1+k_0')/k_0'$ for the Craig model. In any discussion of column efficiency it is convenient to consider the *height equivalent to a theoretical plate* (or HETP).

$$H=\frac{L}{N} \tag{9}$$

This concept is developed further in Chapter 7.

4.2. Statistical Models

The *random walk* model has been used by Giddings to relate various individual contributions in the mechanism of band broadening to the column efficiency [18]. This leads directly to a Gaussian band profile. The contributions of individual sources of band broadening to changes in the profile are additive. Thus, the column HETP is obtained as the sum of individual contributions

$$H=\sum h_i \tag{10}$$

which can be calculated for each identifiable source of band broadening. Giddings and Eyring [19] have introduced a more sophisticated, stochastic model that treats the chromatographic process as a Poisson distribution process. Assuming molecules undergo a series of random adsorption and desorption steps during their migration along the column, it is possible to derive a distribution for the residence times of the molecules injected in a pulse. This model has been extended by McQuarrie to the case of mixed mechanisms [20]. More recent investigations by Dondi et al. confirm that the model can account for both axial dispersion and mass-transfer resistance [21]. When the column efficiency is not too low, the resulting band profile tends toward a Gaussian distribution.

4.3. Mass-Balance Models

These models involve writing and integrating a differential mass-balance equation. The basic assumptions associated with writing such an equation are discussed in the literature [16]. There are two important models of this type, the equilibrium – dispersive model [22] and the lumped kinetic models, in addition to the general rate model. The differential mass balance equation is

$$\frac{\partial C}{\partial t}+F\frac{\partial q}{\partial t}+u\frac{\partial C}{\partial z}=D_{\mathrm{a}}\frac{\partial^2 C}{\partial z^2} \tag{11}$$

where C and q are the mobile and stationary phase concentrations, respectively, F is the phase ratio [$F=(1-\varepsilon_{\mathrm{T}})/\varepsilon_{\mathrm{T}}$, with ε_{T} the total packing porosity], u is the mobile-phase velocity, and D_{a} is the axial dispersion coefficient. The first two terms on the left of this equation are the accumulation terms, the third is the convective term. The term on the right is the dispersive term. This equation contains two functions of the time (t) and space (z) coordinates, C and q. These must be related by a second equation to fully determine the problem.

4.3.1. The Equilibrium – Dispersive Model

In this model, one assumes that any departure from equilibrium between the two phases of the chromatographic system is very small, and the stationary-phase concentration q is given practically everywhere by the equilibrium isotherm:

$$q=f\,(C) \tag{12}$$

For theoretical discussions, the Langmuir isotherm $q=a\,C/(1+b\,C)$ is convenient. For a review of isotherms used in liquid chromatography, see [23]. With $D_{\mathrm{a}}=0$ this assumption would be equivalent to assuming that the column efficiency is infinite. However, the finite efficiency of an actual column can be taken into account by including in the axial dispersion coefficient the influences on band profiles due to both axial dispersion and the kinetics of mass transfer:

$$D_{\mathrm{a}}=\frac{H\,u}{2}=\frac{L\,u}{2N} \tag{13}$$

where H is the column HEPT and L is the column length. With a linear isotherm, Equation 11 becomes

$$(1+k_0')\,\frac{\partial C}{\partial t}+u\frac{\partial C}{\partial z}=D_{\mathrm{a}}\frac{\partial^2 C}{\partial z^2} \tag{14}$$

where k_0' is the retention factor. Solving this equation requires a choice of initial and boundary conditions [22]. In analytical chromatography, the initial condition is an empty column. Three types of boundary conditions have been studied:

1) An infinitely long column stretching from $z=-\infty$ to $z=+\infty$
2) An infinitely long column stretching from $z=0$ to $z=+\infty$
3) A finite column from $z=0$ to $z=L$

Depending on the boundary condition chosen, different solutions are obtained [16]. In practice, for values of N exceeding 30 – 50 plates there are no noticeable differences between the profiles. The first and second moments of these profiles are

1) $\left(1+\frac{1}{N}\right)t_{\mathrm{R}}$ and $\frac{t_{\mathrm{R}}^2}{N}\left(1+\frac{2}{N}\right)$
2) $\left(1+\frac{1}{2N}\right)t_{\mathrm{R}}$ and $\frac{t_{\mathrm{R}}^2}{N}\left(1+\frac{3}{4N}\right)$
3) t_{R} and $\frac{t_{\mathrm{R}}^2}{N}\left[1-\frac{1}{2N}\left(1-\mathrm{e}^{-2N}\right)\right]$

with $t_{\mathrm{R}}=(1+k_0')\,t_0$, and $t_0=L/u$. The differences relative to a Gaussian curve become rapidly negligible as N increases, ultimately reaching the values commonly achieved.

4.3.2. Lumped Kinetic Models

These models recognize that equilibrium between the mobile and stationary phases can never actually be achieved, and use a kinetic equation to relate $\partial q/\partial t$, the partial differential of q with respect to time, and the local concentrations, q and C. Several different kinetic models are possible depending on which step is assumed to be rate controlling. All are called "lumped kinetic" models because, for the sake of simplicity, the contributions of all other steps are lumped together with the one considered to be most important. These models are discussed and compared in detail in [24]. All such models are equivalent in linear chromatography, but not when the equilibrium isotherm is nonlinear [16]. The kinetic equation can be written as:

$$\frac{\partial q}{\partial t} = k_{\mathrm{m}} \left(C_{\mathrm{s}} - q \right) \tag{15}$$

where C_{s} is the stationary-phase concentration in equilibrium with the mobile-phase concentration C (and the isotherm is $C_{\mathrm{s}} = a\,C$). A general solution of these models has been given by Lapidus and Amundson [25]. Van Deemter et al. [26] have shown that the solution can be simplified considerably if one assumes reasonably rapid mass-transfer kinetics. The profile then becomes Gaussian, and the column HETP is given by:

$$H = \frac{2D_L}{u} + 2\left(\frac{k_0'}{1+k_0'}\right)^2 \frac{u}{k_0' k_{\mathrm{m}}} \tag{16}$$

This is the well-known Van Deemter equation.

4.3.3. The Golay Equation

In the particular case of a cylindrical column the wall of which is coated with a layer of stationary phase, the system of partial differential equations of a simplified kinetic model can be solved analytically [27]. For a noncompressible mobile phase, Golay derived the following equation:

$$H = \frac{2D_{\mathrm{m}}}{u} + \frac{1 + 6k_0' + 11k_0'^2}{96\left(1+k_0'\right)^2} \frac{u d_{\mathrm{p}}^2}{D_{\mathrm{m}}} + \frac{k_0'}{24\left(1+k_0'\right)^2} \frac{u d_{\mathrm{f}}^2}{D_{\mathrm{l}}} \tag{17}$$

where d_{p} is the particle size, d_{f} is the thickness of the liquid film, D_{m} is the diffusion coefficient in the carrier gas, and D_{l} is the diffusion coefficient in the liquid phase. This equation is in excellent agreement with all experimental results so far reported.

4.4. The General Rate Model

In this model, each step of the chromatographic process is analyzed in detail [16]. Separate mass-balance equations are written for the mobile phase that flows through the bed and for the stagnant mobile phase inside the particles. Separate kinetic equations are then written for the kinetics of adsorption/desorption and for mass transfer. Again, if these kinetics are not unduly slow, the band profile tends toward a Gaussian shape (in which case a simpler model is actually more suitable).

4.5. Moment Analysis

The *moments* of a statistical distribution provide a convenient description of this distribution. Since a band profile is the distribution of residence times for molecules injected into a chromatographic system, the use of moments to characterize these profiles is a natural approach. Furthermore, analytical solutions of the general rate model can be obtained in the Laplace domain. The inverse transform to the time domain is impossible to solve in closed form, but the moments of the solution can easily be calculated. Equations are available for the first five moments [28], [29].

The nth moment of a distribution is

$$M_n = \int_0^\infty C(t)\, t^n\, dt \tag{18}$$

It is more practical to use normalized distributions, so the absolute moments, μ_n, $n \geq 1$, and the central moments, $\bar{\mu}_n$, $n > 1$ are most commonly employed:

$$\mu_n = \frac{M_n}{M_0} = \frac{\int_0^\infty C(t)\, t^n\, dt}{\int_0^\infty C(t)\, dt} \tag{19a}$$

$$\bar{\mu}_n = \frac{\int_0^\infty C(t)\,(t-\mu_1)^n\, dt}{\int_0^\infty C(t)\, dt} \tag{19b}$$

These moments have been used to study band profiles. Interpretations are straightforward for the zeroth (peak area), first (retention of the peak mass center), and second ($H = \bar{\mu}_2 L / \mu_1^2$) moments. Higher moments are related to deviations of the band profile from a Gaussian profile [30]. The use of such moments is limited, however, in large part because the accuracy of determination decreases rapidly with increasing order of a moment. Use of the first moment rather than the retention time of a band maximum has been suggested as a way of relating chromatographic and thermodynamic data for an unsymmetrical peak [30], but this is justified only with slow mass-transfer kinetics.

4.6. Sources of Band Asymmetry and Tailing in Linear Chromatography

The theory of chromatography shows that there should be no noticeable difference between a recorded peak shape and a Gaussian profile so long as the number of theoretical plates of the column exceeds 100 [16]. The difference is small even for 25 plates, and careful experiments would be needed to demonstrate that difference. However, it is common to experience in practice peak profiles that are not truly Gaussian.

There are basically four possible causes of unsymmetrical band profiles in chromatography. First, the injection profile may exhibit some tailing. Second, the column may be overloaded such that the isotherm of equilibrium between the two phases is no longer linear. Third, the stationary phase may be inhomogeneous, or mixed retention mechanisms may be operative. Finally, the column packing may not be homogeneous. Slow mass transfer is usually not responsible, contrary to common wisdom, because this would lead to an efficiency much lower than that actually observed.

Injections are rarely carried out under experimental conditions that permit the achievement of narrow, rectangular sample pulses. More often, the injection profile tails markedly. If the length of this tail is significant compared to the standard deviation of the band, the result is tailing of the analytical band. For an exponential injection profile with a time constant τ, the ratio τ/σ should be < 0.10 [31]. Such tailing is usually restricted to high-efficiency bands eluted early, often as a result of poorly designed injection ports or connecting tubes, or an insufficiently high injection-port temperature in gas chromatography. Direct connection of the injection system to the detector permits assessment of this particular source of tailing [32].

When the sample size is too large, the corresponding isotherm can no longer be considered identical to its initial tangent. If the isotherm is convex upward, the retention time decreases with increasing concentration, and the peak tails. Isotherms in gas chromatography are usually convex downward, and peaks from large samples tend to "front." However, peaks eluted early tend to have very sharp fronts and some tail. This is caused by a sorption effect, since the partial molar volume of the vapor is much larger than that of the dissolved solute [33].

In most cases, peak tailing can be explained by heterogeneity of the surface of the stationary phase [16]. Strongly adsorbing sites saturate before weak ones, and they have a much smaller saturation capacity. The efficiency increases when the sample size is reduced, but this observation is often of limited use, because detection limits establish a minimum for the acceptable sample size.

5. Flow Rate of the Mobile Phase

The velocity of the mobile phase is related to the pressure gradient, the mobile phase viscosity η, and the column permeability k_0 by the Darcy law [34]:

$$u = -\frac{k_0}{\eta}\frac{\mathrm{d}p}{\mathrm{d}z} \tag{20}$$

The minus sign here indicates that the flow is directed from high to low pressure. The differential equation can be integrated provided the relationship is known between local volume and pressure. Two equations of state can be used, one for gases (Boyle – Mariotte law), the other for liquids. For dense gases, the equation of state is far more complex [35]. Liquids can be considered as noncompressible for purposes of calculating a pressure gradient. In this case u is constant along the column, and the difference between the inlet and outlet pressures is

$$\Delta p = \frac{u\eta L}{k_0} \tag{21}$$

The retention time for an inert tracer is

$$t_0 = \frac{L}{u} \tag{22}$$

In gas chromatography the carrier gas behaves as an ideal gas from the standpoint of compressibility, so $p\,u = p_0\,u_0$, and the integration of Equation 20 gives

$$u_0 = \frac{k_0 p_0}{2\eta L}\left(p^2 - 1\right) \tag{23}$$

Because of gas-phase compressibility, the average gas velocity $\bar{u}$ is lower than the outlet velocity. It can be calculated that

$$\bar{u}=j\,u_0=\frac{3\left(p^2-1\right)}{2\left(p^3-1\right)}u_0 \tag{24}$$

where j is the James and Martin compressibility factor [11]. The retention time for an inert tracer in this case is

$$t_0=\frac{4\eta L^2\left(p_i^3-p_0^3\right)}{3k_0\left(p_i^2-p_0^2\right)} \tag{25}$$

5.1. Permeability and Porosity of the Packing

5.1.1. Column Porosity

There are three different porosities that must be considered: The *interparticle* or *external* porosity, the *intra particle* or *internal* porosity, and the *total* porosity. The external porosity is the fraction of the overall geometrical volume of the column available to mobile phase that flows between the particles. Chromatographers always strive to achieve a dense column packing as the only known practical way of packing reproducible, stable, and efficient columns, so this porosity tends to be similar for all columns, with a value close to 0.41.

The internal porosity depends on the nature and structure of the particles used to pack the column. It varies considerably from one stationary phase to another, ranging from zero (nonporous beads) to ca. 60 % for certain porous silicas.

The total porosity is the fraction of the bed accessible to the mobile phase; i.e., it is the sum of the internal and external porosities.

5.1.2. Column Permeability

The *permeability* of a column depends on the characteristics of the particular packing material used, and especially on the column external porosity, but it is the same regardless of whether the mobile phase is a gas, a liquid, or a dense gas. This permeability is inversely proportional to the square of the particle diameter:

$$k_0=\frac{d_{\mathrm{p}}^2}{\Phi} \tag{26}$$

For packed columns the numerical coefficient Φ is of the order of 1000, with reported values ranging between 500 and 1200. It tends to be smaller for spherical than for irregular particles. For empty cylindrical tubes, such as those used for open tubular columns, it is equal to 32.

5.2. Viscosity of the Mobile Phase

The viscous properties of gases and liquids differ significantly [36].

5.2.1. Viscosity of Gases

The viscosity of a gas increases slowly with increasing temperature, and is nearly independent of the pressure below 2 MPa. It is approximately proportional to $T^{5/6}$. The rare gases (He, Ar) have higher viscosities than diatomic gases (N_2, H_2) and polyatomic molecules. Hydrogen has the lowest viscosity of all possible carrier gases, and is often selected for this reason. The viscosity of the sample vapor is unlikely to exceed that of the carrier gas.

5.2.2. Viscosity of Liquids

The viscosity of a liquid decreases with increasing temperature, following an Antoine-type equation:

$$\log\eta_{\mathrm{liq}}=A+\frac{B}{T+C} \tag{27}$$

In most cases, however, C is negligible, and if T is expressed in kelvins B is of the order of 1 to 3×10^{-3}. For example, the viscosity of benzene decreases by one-half with an increase in temperature of 120 °C. The viscosity increases slowly and linearly with increasing pressure below 200 MPa, and may increase by 10 – 25 % when the pressure rises from 0 to 20 MPa. The viscosity of a liquid also increases with the complexity of the molecule and its ability to engage in hydrogen bonding. The viscosity of a sample may be much higher than that of the corresponding mobile phase, which presents a serious potential problem in preparative chromatography [37].

6. The Thermodynamics of Phase Equilibria and Retention

Chromatography separates substances on the basis of their equilibrium distributions between the two phases constituting the system. Without retention there can be no separation. Retention data are directly related to the nature of the phase equilibrium and to its isotherm. Although the details depend very much on the nature of the mobile and stationary phases used, a few generalizations are possible.

6.1. Retention Data

It is easy to relate retention times to the thermodynamics of phase equilibria. In practice, however, it is difficult to measure many of the relevant thermodynamic parameters with sufficient accuracy, while retention times can be measured with great precision. Furthermore, it should be emphasized at the outset that even though very precise measurements can indeed be made in chromatography [38], the available stationary phases are often so poorly reproducible that such precision is unwarranted. Column-to-column reproducibility for silica-based phases is reasonable only for columns packed with material from a given supplier. For this reason, chromatographers prefer whenever possible to use relative retention data.

6.1.1. Absolute Data

The experimental data acquired directly in chromatography are retention time, column temperature, and the mobile-phase flow rate. All other data are derived from these primary parameters. The retention time is inversely proportional to the flow rate, so it is often preferable instead to report the retention volume and the retention factor (both of which are independent of the flow rate).

The *retention factor* k' is the ratio of the fraction of component molecules in the stationary phase n_s to the fraction in the mobile phase n_m at equilibrium. The retention factor is thus

$$k' = \frac{n_s}{n_m} = \frac{1-R}{R} \tag{28}$$

This also corresponds to the ratio of the average times spent by a typical molecule in the two phases:

$$k' = \frac{t_R - t_0}{t_0} \tag{29}$$

Accordingly, k' is proportional to the thermodynamic constant for the phase equilibrium at infinite dilution. The *retention volume* V_R (see Eq. 4) is the volume of mobile phase required to elute a component peak. For compressible phases it must be referred to a standard pressure:

$$V_R = t_R F_v f(P) \tag{30}$$

where F_v is the mobile-phase flow rate and $f(P)$ is derived from the equation of state for the mobile phase. For the ideal gases used in GC,

$$f(P) = j = \frac{3(P^2-1)}{2(P^3-1)}$$

(Eq. 24); for liquids, $f(P) = 1$; for dense gases it is a complex function [35].

6.1.2. Relative Data

Because absolute retention data depend on so many controlled (temperature, pressure) and uncontrollable (activity of the stationary phase) factors, and because their quantification requires independent determinations that are time-consuming (e.g., density of the stationary liquid phase in the case of GC) or not very accurate (flow-rate measurements), and also because chromatographers are more often concerned with practical separation problems than with phase thermodynamics, it is more convenient to work with relative retention data. The parameter of choice is the *relative retention* α, which is also the *separation factor:*

$$\alpha_{2,1} = \frac{t_{R,2} - t_0}{t_{R,1} - t_0} = \frac{k'_2}{k'_1} \tag{31}$$

Relative retention data are more convenient than absolute data, but the choice of an appropriate reference may prove difficult. An ideal reference compound

1) Should not be present in mixtures subject to routine analysis

2) Should be introduced into the sample as a standard (though it might also be used as an external standard in quantitative analysis)
3) Should be nontoxic and inexpensive
4) Should elute near the middle of the chromatogram so the relative retention data will be neither too large nor too small
5) Should have chemical properties similar to those of most of the components in the sample mixture, thereby ensuring that relative retention volumes will be insensitive to changes in the level of activation of the stationary phase

For complex mixtures, especially those analyzed with temperature programming (in GC) or gradient elution (in LC), it is impossible to select a single reference compound. The Kovats retention-index system [39], based on homologous series of reference compounds, provides an elegant solution in GC, one which has been widely accepted. In LC, the absence of suitable homologous series, and the fact that retention depends more than in the case of GC on the polarities of compounds and less on molecular weight makes the use of an index system impractical.

Finally, because of its definition, the separation factor for two compounds is related to the difference between the corresponding Gibbs free energies of transfer from one phase to the other:

$$RT\ln\alpha = \Delta\left(\Delta G^0\right) \tag{32}$$

When α is close to unity [40],

$$\alpha - 1 = \frac{\Delta\left(\Delta G^0\right)}{RT}.$$

6.2. Partition Coefficients and Isotherms

From Equation 29 it can be shown that for gas – liquid chromatography:

$$k' = \frac{RTr}{\gamma^\infty P^0 M_s}\frac{V_L}{V_G} = K\frac{V_L}{V_G} \tag{33}$$

where P^0 is the vapor pressure of the analyte, γ^∞ is the activity coefficient at infinite dilution, ρ is the density, and K is the thermodynamic equilibrium constant, or the initial slope of the isotherm. Alternately, the constant K can be derived from measurement of the specific retention volume. The retention volume, $V_R = F_v\, t_R$, is the volume of carrier gas required to elute a particular component peak. The *corrected* retention volume V_N is the retention volume corrected for gas compressibility (see Eq. 24) and for the hold-up time in the mobile phase:

$$V_N = j\,(t_R - t_0)\,F_v = j\,k' t_0 F_v \tag{34}$$

The *specific* retention volume is the corrected retention volume under STP conditions, reported for a unit mass of liquid phase:

$$V_g = \frac{V_N}{m_L}\frac{273}{T_c}\frac{p_0}{P_n} = \frac{K}{\varrho}\frac{273}{T_c} \tag{35}$$

where T_c is the column temperature. In practice, derivation of the constant K from k', V_L, and V_G or from V_g requires the same measurements, and the accuracy and precision are the same. Since V_g is proportional to a thermodynamic constant of equilibrium, it varies with the temperature according to

$$\frac{\mathrm{d}\,(\ln V_g)}{\mathrm{d}\,(1/T)} = -\frac{\Delta H_S}{R} \tag{36}$$

Since the molar enthalpy of vaporization of solute ΔH_S from an infinitely dilute solution is negative, retention volumes decrease with increasing temperature. In liquid – liquid chromatography the retention volume depends on the ratio of activities of the component in the two phases at equilibrium; i.e., on the ratio of the activity coefficients at infinite dilution. As liquid – liquid chromatography has been little used so far (because of instability of the chromatographic system), and activity coefficients are not easily accessible, this relationship has not been extensively explored.

6.3. Gas – Solid Adsorption Equilibria

Gas – solid chromatograhy is used mainly for the separation of gases. In spite of major advantages, such as a high degree of selectivity for geometrical isomers and the high thermal stability of many adsorbents, it is not used for the analysis of organic mixtures.

Two conditions are required for successful analysis by gas – solid chromatography: The isotherm must be close to its initial tangent so that peaks migrate under conditions of linear

equilibrium, and the surface of the adsorbent must be highly homogeneous. Gases are analyzed at temperatures much above their boiling points, so their partial pressures relative to their vapor pressures are very low, and equilibrium isotherms deviate little from their initial tangents. This is impractical with heavier organic vapors that decompose thermally at moderate temperatures. Graphitized carbon black, with a highly homogeneous surface, has been used successfully for the separation of many terpenes and sesquiterpenes, especially closely related geometrical isomers [41]. A number of adsorbents developed for reversed-phase LC have never been tried in gas chromatography despite their apparent potential advantages.

At low partial pressure (more precisely: low values of the ratio p/P^0 of the partial pressure of the analyte to its vapor pressure) the amount of a substance adsorbed on a surface at equilibrium with the vapor is proportional to the partial pressure:

$$m = K_{\mathrm{H}} p \tag{37}$$

where the constant K_{H} is the Henry constant [41]. Retention data are sometimes reported in terms of mass of adsorbent in the column, analogous to the specific retention volume, although the adsorbent mass itself is not actually related to the adsorption equilibrium. Retention data should rather be reported as a function of area S_{a} of adsorbent:

$$K_{\mathrm{H}} = \frac{V_{\mathrm{R}}}{m_{\mathrm{a}} S_{\mathrm{a}}} \tag{38}$$

However, the determination of S_{a} is complex and relatively inaccurate. The Henry constant of adsorption depends on the geometrical structure of the adsorbent as well as on the interaction energy between the surface and chemical groups in the adsorbate molecule. The Henry constant of adsorption decreases rapidly with increasing temperature. Equation 36 applies in gas – solid chromatography as well.

6.4. Liquid – Solid Adsorption Equilibria

Liquid – solid chromatography is the most widely applied method of liquid chromatography. Other modes such as hydrophobic-interaction chromatography are in fact variants of liquid – solid chromatography based on adsorption from complex solutions. The two most important methods in liquid – solid chromatography are *normal phase*, which uses a polar adsorbent (e.g., silica) and a low-polarity mobile phase (e.g., methylene chloride), and *reversed phase*, which uses a nonpolar adsorbent (e.g., chemically bonded C_{18} silica) and a polar eluent (e.g., methanol – water mixtures).

6.4.1. Normal-Phase Chromatography

Retention volumes in normal-phase chromatography depend much on the polarity and polarizability of the analyte molecules, and little on their molecular masses [42]. Thus, all alkyl benzenes, or all alkyl esters, will be eluted in a narrow range of retention volumes. On the other hand, structural changes affecting the polarity of a molecule are easily seen. The amount of water adsorbed by the stationary phase controls its activity, hence the resolution between analytes, and this is in turn a sensitive function of water dissolved in the mobile phase. Thus, control of the water content of the eluent is a primary concern. The influence of the composition of the mobile phase on retention has been discussed by SNYDER, who introduced the concept of *eluotropic strength* [42].

Strong solvents (i.e., those that are retained in a pure weak solvent) compete with components of the sample for adsorption. For this reason, a pulse of a weak solvent generates as many peaks as there are components in the mobile phase [43], [44]. These peaks are called *system peaks* [44]. This property has been used for detecting compounds that themselves produce no response from the detector. Thus, if a compound that does produce a response with the detector is introduced at a constant concentration in the mobile phase, injection of a sample is equivalent to injection of a negative amount of the background compound, leading to a negative peak in the chromatogram.

6.4.2. Reversed-Phase Chromatography

In reversed-phase chromatography the retention volume depends very much on the molecular mass of the analyte. Methylene selectivity

(i.e., the relative retention of two successive homologues) varies over a much wider range in reversed-phase LC than in GC, and can be quite large [45].

Water is the weakest of all solvents in reversed-phase chromatography. The mechanism by which a strong solvent reduces the retention volume of an analyte is quite different in reversed-phase chromatography from that in normal-phase mode. In reversed-phase mode there is no competition. Methanol, acetonitrile, and other organic solvents miscible in water are poorly retained on reversed-phase columns eluted with pure water. Addition of organic solvent instead increases analyte solubility in the mobile phase, thereby decreasing retention. The phenomenon of system peaks is nearly negligible.

7. Band Broadening

As discussed previously (Chap. 4), the band profile in linear chromatography is Gaussian. The relative standard deviation of this Gaussian profile is used as a way of characterizing column efficiency. By definition

$$N = \frac{t_{\mathrm{R}}}{\sigma_t} = 2k^2 Ln2 \frac{t_{\mathrm{R}}}{w_{1/k}} = 5.64 \frac{t_{\mathrm{R}}}{w_{1/2}} \tag{39}$$

where $w_{1/k}$ is the peak width at the fraction $1/k$ of its height, and L is the length of the column. The experimental height equivalent to a theoretical plate is

$$\bar{H} = \frac{L}{N} \tag{40}$$

where N and $\bar{H}$ depend on the experimental conditions. Experimental conditions must be so established that N is at a maximum and $\bar{H}$ at a minimum to ensure that the peaks are as narrow as possible, thereby maximizing the resolution between successive bands.

7.1. Sources of Band Broadening

Various phenomena contribute to the broadening of peaks during their migration along a column [18]. The most important are:

1) Axial diffusion along the concentration gradient of the band profile
2) Dispersion due to inhomogeneous flow distribution between the packing particles
3) Resistance to mass transfer back and forth from the mobile phase stream passing between the particles to the stagnant mobile phase inside the particles
4) Resistance to mass transfer through pores inside the particles
5) The kinetics of the retention mechanism (e.g., of adsorption – desorption)

Calculations of the contributions from these several phenomena have been accomplished by integration of the mass-balance equation (e.g., Eq. 17), or by a statistical approach.

The simplest approach involves the random-walk model, derived first for the study of Brownian motion. The trajectory of a given molecule through the column can be represented as a succession of steps in which the molecule either moves forward at the mobile-phase velocity (u) or stays in the stationary phase, where it is immobile (i.e., it moves backward with respect to the mass-center of the peak, which moves at the velocity $R\,u$). If one can identify the nature of the steps made by the molecule, and then calculate their average length l and number n, it is possible to derive the contribution to the variance from this particular source of band broadening; the result is in fact $n\,l^2$ [18].

7.2. The Plate-Height Equation

Because the plate number is related to particle size, and contributions from axial dispersion and resistance to mass transfer in the mobile phase are related to both the velocity of this phase and the molecular diffusion coefficient, it is convenient to express plate height on the basis of reduced, dimensionless parameters, in the form of a *reduced plate height*, $h = H/d_{\mathrm{p}}$ together with a *reduced velocity*, $v = u\,d_{\mathrm{p}}/D_{\mathrm{m}}$ [46].

Contributions from all the various sources of band broadening listed in the previous section have been calculated [18]. If we assume that these sources are independent, the band variance is the sum of each contribution to the variance, and the column HETP is given by the equation [46]

$$h = \frac{B}{v} + Av^{1/3} + Cv \tag{41}$$

where the coefficients A, B, and C are dimensionless functions of the experimental conditions, but velocity independent. The use of reduced plate height and reduced velocity has become standard practice in LC, but not in GC, probably because the concept arose only in the mid 1970s – although this is not a valid reason for ignoring it. Depending on the type of chromatography under consideration, the plate-height coefficients are obtained from different relationships. The Golay equation [27] is typical. This describes the plate height for an open tubular column. It is characterized by $A = 0$ and C a simple function of k'. In a packed GC column, anastomosis of the flow channels causes the axial dispersion to be more complex, and results in the $A\,v^{1/3}$ contribution in Equation 41 as well as a more complicated equation for the C coefficient than in the Golay equation. It is striking, however, that GC produces $h = f\ (v)$ (i.e., HETP versus velocity) curves that are quite similar to those from LC, with minimum values of h between 2 and 3, and corresponding values of v between 3 and 5. These values are excellent approximations for order-of-magnitude calculations. The only marked difference between experimental results reflects the distinction between open-tubular and packed columns. For the former, $A = 0$, and the curve is a hyperbola. For the latter, the curve has an inflection point; a tangent drawn from the origin has its contact point at a finite velocity.

Identifying those sources of band broadening that contribute significantly to increased peak width, and then calculating their contributions, makes it possible to derive the value of the local plate height; i.e., the coefficient relating an incremental increase in band variance to an incremental distance of migration [12]:

$$\mathrm{d}\left(\sigma^2\right) = H\,\mathrm{d}z \tag{42}$$

The *apparent plate height*, $\bar{H}$, given by Equation 40, is the column length average of the local plate heights

$$\bar{H} = \frac{\int_0^L H\,\mathrm{d}z}{L} \tag{43}$$

This equation can be used to calculate apparent HETPs in both GC and TLC, which are types of chromatography in which the mobile-phase velocity varies quite markedly during the course of peak migration [47].

7.3. Resolution Between Peaks

The *resolution* (R) between two successive Gaussian bands is defined as

$$R = 2\,\frac{t_{\mathrm{R},2} - t_{\mathrm{R},1}}{w_1 + w_2} \tag{44}$$

Assuming the two bands are close (since otherwise the concept of resolution loses most of its relevance), the relationship $t_{\mathrm{R},2} + t_{\mathrm{R},1} \simeq 2\,t_{\mathrm{R},2}$ represents a good approximation, and

$$R = \frac{\sqrt{N}}{4}\,\frac{\alpha - 1}{\alpha}\,\frac{k'}{1 + k'} \tag{45}$$

This equation [48] shows, in the first place, that there is no separation if there is no retention. Separations become very difficult when k' tends toward 0 even if α is large. A case in point is the separation of nitrogen from argon by gas chromatography on molecular sieves (0.5 nm) at room temperature. Furthermore, solving Equation (45) for N gives the column efficiency required for separating two compounds with a given degree of resolution. This shows that the required column length increases rapidly as α approaches 1. With a 100-plate column it is possible to separate with unit resolution two compounds with $\alpha = 2$ and $k' = 3$; a 1 000 000-plate column is required to separate two compounds with $\alpha = 1.005$.

7.4. Optimization of Experimental Conditions

It is quite easy to find the conditions corresponding to maximum column efficiency. The column must be operated at the velocity producing the minimum plate height, and if the efficiency is still insufficient, a longer column must be prepared. In practice, certain constraints are applicable, however. First, there is a maximum pressure at which any column can be operated, a

pressure that depends on the equipment available. Furthermore, analysts are often more interested in performing separations rapidly than in achieving the highest possible resolution. It should be emphasized that the less time a band spends in a column, the narrower it will be, hence the higher its maximum concentration and the lower the detection limit. Optimization for minimum analysis time becomes an important consideration. It is easily shown that minimum analysis time is achieved by operating a column at the reduced velocity for which the ratio h/v is a minimum, with a particle size such that this velocity is achieved at the maximum acceptable column inlet pressure [49–51].

A considerable amount of work has been published on optimizing the experimental conditions for minimum analysis time under various constraints [52]. One complication arises from the definition of reduced mobile-phase velocity. The actual mobile-phase velocity depends largely on the molecular-diffusion coefficient of the analyte. Thus, very small particles can be used for the analysis of high molecular mass compounds, which have low D_{m} values. The actual flow rate required will remain compatible with pressure constraints despite the resulting high pneumatic or hydraulic resistance. Detailed results obviously depend greatly on the mode of chromatography used.

7.5. Instrumental Requirements

A column alone cannot give the information the analyst requires. A complete instrument is needed, one with a mobile-phase delivery system, an injection device, a detector, and a data station. At all stages the requirements and properties of the modules force the analyst to accept compromises and lose resolution. Contributions of the injector and detector to band broadening should be of special concern. The lucid analysis of these problems with respect to GC by Sternberg [53] has become a classic. Most of the conclusions are easily extended to LC.

Band broadening in the equipment results from axial dispersion in valves and connecting tubes, as well as from time delays associated with injection of the sample and detection of the separated peaks. The contributions to band broadening from these sources can be expressed as additional terms in the equation for apparent column HETP:

$$H_{\mathrm{c}} = \frac{\sum \tau_j^2 u^2}{(1+k')^2 L} \tag{46}$$

where τ^2 is the time variance caused by any of the band-broadening sources indicated above.

7.5.1. Injection Systems

There is a dearth of data regarding actual injection profiles for every type of column chromatography. Dependence of the instrument variance on the mobile-phase flow velocity has been studied in liquid chromatography [32]. Probably in part because of the onset of eddy turbulence at various flow rates in different parts of the instrument, this variation is not always simple.

If accurate efficiency data are required, the variance of the injection profile should be measured as a function of the flow rate, with a correction applied by subtracting the first moment of the injection profile from the peak retention time and the variance of the injection profile from the band variance. In actual analytical practice it is often sufficient to minimize this contribution by making sure that the volume of the injection device is much smaller than the volume of the column, that the device is properly swept by the mobile phase, and that actual injection is rapid. The use of a bypass at the column inlet permits a higher flow rate through the injection device than through the column, effectively reducing the band-broadening contribution of the injection system.

7.5.2. Connecting Tubes

Sternberg [53] and Golay and Atwood [32], [54] have studied this particular contribution, showing that for short tubes it is smaller than predicted by application of the Golay equation (Eq. 17) to an empty tube. On the other hand, dead volumes, spaces unswept by the mobile-phase stream but to which the analytes have access by diffusion, act as mixing chambers, and they may have a highly detrimental effect on efficiency. The relative importance of the problem increases with decreasing column diameter.

7.5.3. Detectors

The detector is a concentration sensor appropriate to the analysis in question. As in the case of all sensors, detectors do not react immediately, but instead have a time constant. Thus, they contribute to band broadening through both their cell volume and their response time. The cell-volume contribution can be handled in the same way as the volume of the injection device. The response-time contribution is given by Equation 46, and it is proportional to the square of the mobile-phase velocity. The response time should be less than one-tenth of the standard deviation of a Gaussian peak in order to have a negligible effect on the retention time and profile of the peak [31].

7.5.4. Instrument Specifications

If a fractional loss of efficiency equal to Φ is acceptable, it follows that:

$$\left(\frac{\sigma_{\mathrm{eq}}}{\sigma_{\mathrm{col}}}\right)^2 \leq \Phi \tag{47}$$

where σ^2_{eq} and σ^2_{col} are the variance contributions of the equipment and the column, respectively. Application of this relationship permits easy derivation of equipment specifications as a function of the particular system used and the analysis of interest [52].

8. Qualitative Analysis

Although chromatography is a most powerful method of separation, it affords few clues regarding the identity of the separated eluates. Only a retention time, volume, or factor is available for each peak, and the method cannot even establish whether a given peak corresponds to a pure compound or to a mixture. Supplying information that is both meager and not very accurate, chromatography alone is poorly suited to the identification of single components, although previously identified patterns simplify the matter (e.g., alkane series in petroleum derivatives, or phytane – pristane in crude oils). Innumerable attempts at processing retention data for identification purposes have been reported, including the use of the most refined chemometric methods, but success has been limited. In practice, a combination of chromatographic separation and spectrometric identification has become the preferred method of identification for unknown chemicals.

8.1. Comparisons of Retention Data

The most efficient identification procedure involves the determination in rapid succession of retention data (retention volumes, or preferably retention factors) for the components of a mixture and a series of authentic compounds. This method permits straightforward identification of components known to be present in the mixture, and it can also rule out the presence of other possible components. However, it cannot positively identify the presence of a given, isolated component. It is recommended that such determinations be carried out using similar amounts of compounds, and that a number of spike analyses be performed, where small amounts of authentic compounds are added to the mixture in order to approximately double the sizes of unknown peaks. Comparison of the efficiencies of an unknown peak in the original sample and in the spiked mixtures increases the probability of recognizing differences between the analyte and an authentic compound.

Giddings and Davis [55] have shown that, for complex mixtures, analyte retention data tend to reflect a Poisson distribution. This result permits a simple calculation of the probability of finding a certain number of singlets, doublets, and higher-order multiplets in the course of analyzing a mixture of m components on a column with a peak capacity n. If the ratio m/n is not small, the probability is low. This confirms that there is little chance of separating a complex mixture on the first phase selected [56]. Method development is a long and onerous process because the probability of random success is low.

In practice, qualitative analysis is made easier by the use of retention indices I [39]:

$$I = 100z + \frac{\log\left(t_{\mathrm{R},X} - t_0\right) - \log\left(t_{\mathrm{R},P_z} - t_0\right)}{\log\left(t_{\mathrm{R},P_{z+1}} - t_0\right)} \tag{48}$$

where $t_{\mathrm{R},X}$ is the retention time of component X, and P_z is the n-alkane containing z carbon atoms. The retention index varies little with

temperature, usually following an Antoine-type equation. The retention index increment

$$\Delta I = I^{\mathrm{P}} - I^{\mathrm{A}} \tag{49}$$

is the difference between the retention indices for a compound on two stationary phases: A polar phase, P, and a nonpolar phase, A. Squalane, Apiezon L (a heavy hydrocarbon grease), and various polymethyl siloxanes have been used as nonpolar phases. Kovats and Wehrli [57] have shown that the retention index increment for a compound can be related to its structure.

8.2. Precision and Accuracy in the Measurement of Retention Data

The *precision* of retention data can be quite high [38], [58]. This is related to the precision with which the peak maximum can be located; i.e., it depends on column efficiency and on the signal-to-noise ratio. *Reproducibility* depends on the stability of the experimental parameters (i.e., the mobile-phase flow rate and the temperature). Reproducibility also depends on the stability of the stationary phase, which is, unfortunately, much less satisfactory than often assumed. In gas – liquid chromatography, the stationary liquid phase tends to oxidize, decompose, and become more polar unless extreme precautions are taken to avoid oxygen, which diffuses into the carrier gas through septa and flow-rate controller membranes. In liquid – solid chromatography, impurities in the mobile phase (e.g., water in the case of normal phase LC) are adsorbed, and these slowly modify the activity of the stationary phase; alternatively, the aqueous mobile phase may slowly react with a chemically bonded phase and hydrolyze it.

The *accuracy* of retention data is still more questionable than their reproducibility. Poor batch-to-batch reproducibility for silica-based phases has plagued LC for a generation. For this reason, retention data cannot be compiled in the same way as spectra. Such data collections have limited usefulness, providing orders of magnitude for relative retention rather than accurate information on which an analyst can rely for the identification of unknowns. Precision and accuracy with respect to absolute and relative retention data have been discussed in great detail [58].

8.3. Retention and Chemical Properties

Considerable effort has been invested in elucidating the relationships between chemical structure and retention data in all forms of chromatography. A wide compendium of results has been published in the literature. These studies are much more detailed for GC than LC, probably because work was begun earlier in the former case, before the sterility of the approach for purposes of qualitative analysis was recognized.

A plot of the logarithm of the retention volumes for homologous compounds versus the number of carbon atoms in their chains produces a nearly straight line. Exceptions are found for the first few members of a series and in other rare cases. Over a wide range of carbon-atom numbers, it also appears that such a plot exhibits definite curvature [59]. This result has been observed in all modes of chromatography capable of separating homologues. It should be noted, however, that the approximate linearity of the plots depends on a proper estimate of the hold-up volume, which is not readily defined in liquid chromatography [60]. Lines observed for different functional groups are parallel, or nearly so.

More generally, the Gibbs free energy of retention can be written as a sum of group contributions [61], and increments can be ascribed to different groups or substituents. This permits the approximate calculation of retention data from tables of increments. Unfortunately, multifunctional compounds are subject to interactions between the respective groups, and interaction coefficients must be introduced, so the system rapidly becomes too complex for practical application.

In GC, the retention volumes of nonpolar or slightly polar compounds obey the following correlation [62]:

$$\log V_g = \log \frac{273R}{M_{\mathrm{S}}} + 0.22k\left(\frac{T_{\mathrm{Eb}}}{T} - 1\right) - 2.9 - \log p \tag{50}$$

where M_{S} is the molecular mass of the stationary phase, k is Trouton's constant (ca. 88 J mol^{-1} K^{-1}), and T_{Eb} is the boiling point of the compound. More detailed relationships linking retention-index increments in gas chromatography with the structures of the corresponding components have been suggested [57].

These correlations apply well within groups of related, monofunctional compounds, but difficulties arise in the attempt to extend the relations to multifunctional compounds. However, the identification of complex compounds is the only real challenge remaining in this area. Applying such relationships to a series of standards has facilitated the classification of gas-chromatographic stationary phases [63].

8.4. Selective Detectors

The use of selective detectors is the most promising route to the identification of unknown components in a complex mixture. Chromatography can separate complex mixtures into individual components, but it cannot positively identify them. Most spectrometric methods, on the other hand, supply sufficient information to identify complex compounds, but only if they are so pure that all the spectral clues can be ascribed safely to the same compound. Among the possible combinations, coupling with mass spectrometry, and to a lesser degree IR spectrometry, have been the most fruitful, because the amount of sample easily handled in classical chromatography is well suited to these techniques. Less sophisticated detectors have also been used to advantage.

8.4.1. Simple Selective Detectors

Paradoxically, the gas-density detector, which can supply the molecular mass of an unknown, can be considered to be a selective detector in gas chromatography [58]. This comment effectively illustrates the dearth of convenient selective detectors. At best such detectors help to identify compounds that contain halogen atoms (electron-capture or thermoionic detectors), sulfur, nitrogen, or phosphorus (flame-photometric or thermoionic detectors). Physiological detectors have also been used in certain rare cases (insects that react to sexual pheromones, for example, or the chemist's nose, a dangerous and hazardous application).

The choice of selective detectors in liquid chromatography is still more limited. Electrochemical detectors and the UV-spectrophotometric detector can provide clues, but many types of compounds resist oxidation or reduction in an electrochemical cell, and the paucity of clues provided by a UV spectrum has been a source of disappointment to many chromatographers.

Selective detectors are much more useful for the selective quantitation of a small number of chemicals known to be present in a complex matrix than they are for identification purposes. The use of on-line chemical reactions can be very selective for certain chemical families (e.g., amino acids, sugars). This technique has become widespread in column chromatography, both gas and liquid, and is even the principle underlying many detection schemes in thin layer chromatography [64]. The identification of an unknown presupposes methods supplying far more information and consistent with the ready association of elements of this information with structural details of the corresponding compound.

8.4.2. Chromatography – Mass Spectrometry (→ Mass Spectrometry)

The coupling of gas chromatography and mass spectrometry (GC/MS) was developed during the 1970s, and it has become the generally accepted method for identifying the components of complex mixtures. Its use has considerably influenced the development of modern environmental, clinical, pharmacological, and toxicological analysis. Preferably implemented with open tubular columns and associated with selective derivatization of the sample components, gas chromatography permits the separation and elution of all compounds that have significant vapor pressures at 400 °C and do not decompose, or have derivatives fulfilling these conditions. Study of the resulting mass spectra in conjunction with a rapid computer search in a large spectrum library permits the identification of separated compounds. Quantitation can be accomplished with the aid of the total ion current, currents for selected masses, or the signal from an associated chromatographic detector.

The coupling of liquid chromatography with mass spectrometry (LC/MS) has also been intensively investigated since the early 1980s, but so far without complete success. In GC/MS it is relatively easy to remove the mobile phase, or

to limit and control its effects on the ionization source, but this has proven to be much more difficult in LC/MS. A variety of interfaces has been developed for that purpose;

1) A direct liquid interface, where the mobile phase is nebulized into the source, vaporized, and condensed onto a cold surface
2) Thermo-spray, in which the mobile phase is rapidly heated and vaporized into a spray of fast-moving droplets passing the mass spectrometer inlet, which allows ions to enter the analyzer, but not the unionized solvent
3) Various spraying methods
4) Fast-atom bombardment of liquid flowing out of the column
5) Electro-spray

The variety of these methods, as well as the fact that they produce widely differing spectra and that, for reasons still unclear, one method often appears better suited to a particular group of chemicals, demonstrates that the method is still not a fully developed analytical technique. It nevertheless holds great promise, and has already proven quite useful despite difficulties more serious than those encountered in GC/MS.

8.4.3. Chromatography – IR Spectrometry
(→ Infrared and Raman Spectroscopy)

Infrared (IR) spectrometry is less sensitive than MS, and it therefore requires cells with a long optical path length, but IR spectrometry has also proven quite useful as an identification method in chromatography. Fourier-transform IR (FTIR) spectrometry has become the method of choice because of its sensitivity and the possibility of subtracting most of the contributions due to solvent (in liquid chromatography). It is possible to obtain useful spectra with as little as a few nanograms of a compound, provided it has been chromatographically separated from interfering components. The only remaining source of technical difficulty in coupling a chromatographic separation with an FTIR spectrometer is the size of the sample cell, which is generally excessively large relative to the volume of a band. As a consequence, the resolution between separated components must be greater than in the case of coupling to a mass spectrometer. Columns wider than the conventional 4.6 mm i.d., packed columns could be used with advantage.

9. Quantitative Analysis

Once a mixture has been separated into its components it becomes possible to determine the amount of each [58]. This requires a determination of the size of the concentration profile. Two possible methods are available: Use of either the *peak height* or the *peak area*. The former is more sensitive in the case of GC to fluctuations in details of the injection procedure and the column temperature, and in the mobile-phase composition and column temperature in LC. Peak areas are more sensitive to flow-rate fluctuations in both GC and LC, at least so long as the detector is concentration sensitive. In practice, however, only the peak-area method has been implemented in data packages for chromatographic computers.

With respect to the most frequent case of a concentration sensitive detector, the mass of a particular compound (i) eluted from a chromatographic column is given by [58]

$$m_i = \int_{\infty}^{0} C_i \, \mathrm{d}V \simeq \int_{t_\mathrm{i}}^{t_\mathrm{f}} y_i S_i F_v \, \mathrm{d}t \simeq F_v S_i \int_{t_\mathrm{i}}^{t_\mathrm{f}} y_i \, \mathrm{d}t \qquad (51a)$$

where t_i and t_f are the initial and final times between which integration of the detector profile, $y(t)$, is calculated, and S_i is the appropriate response factor. Thus, the peak area $\int_{t_\mathrm{i}}^{t_\mathrm{f}} y_\mathrm{i} \, \mathrm{d}t$ is inversely proportional to the mobile-phase flow rate. For a mass-flow sensitive detector (e.g., mass spectrometer, flame-ionization detector), the result is simpler:

$$m_i = \int_{0}^{\infty} \frac{\mathrm{d}m_i}{\mathrm{d}t} \, \mathrm{d}t \simeq \int_{t_\mathrm{i}}^{t_\mathrm{f}} \mathrm{d}m_i = S_i \int_{t_\mathrm{i}}^{t_\mathrm{f}} y_i \, \mathrm{d}t \qquad (51b)$$

and the area is independent of the flow rate. The main problems encountered in quantitative analysis are associated with sampling and sample introduction, and with measurement of the peak area and the calibration factors [58]. Chromatographic measurements are usually carried out with a precision of a few percent, but can be much more precise when the sources of error are properly recognized and controlled [65].

9.1. Sampling Problems in Chromatography

As in all analytical techniques, the accuracy of a chromatographic analysis depends on the quality of the sample studied. The techniques involved in ensuring that a representative sample is selected from the system under investigation are discussed elsewhere (→ Sampling). The sample is usually introduced into the mobile-phase stream ahead of the column through a sampling valve, or with a syringe into an injection port. Calibration of the volume of sample injected is difficult, so most analytical work is accomplished by determining only relative concentrations from the actual experimental chromatograms, with internal standards added to permit absolute determinations. A discussion of the repeatability associated with sampling valves can be found in [58]. The approximate volume of an injected sample is most easily measured by injecting a standard acid solution into a flask with the value to be calibrated and titrating.

9.2. Measurement of Peak Area

Peak areas are now measured with the aid of automatic integrators or appropriately programmed computers. These devices incorporate A/D converters and noise filters, and make logical decisions based on the value of the observed signal. One of the major problems is the fact that the devices are closed, and often not transparent to the end user. They are typically programmed by electronic engineers or programmers, who are often poorly informed with respect to the analysts' problems. It is necessary that chemists always have access to raw data when needed, and they should demand that. The nature of the algorithms used to correct for baseline drift, to detect the beginning and end of a peak, and to allocate peak areas between incompletely resolved bands should be clearly indicated, and it should be easy to select the particular algorithm required. Various methods used in automatic integration of chromatographic signals and the related problems have been discussed in detail in [58].

Of special concern is the algorithm used for peak-area allocation when two or more bands are not sufficiently well resolved [58]. In most cases, and if the two peaks are comparable in size, a vertical line is dropped from the bottom of the valley between the peaks. This may cause an error ranging from one to a few percent. When the relative concentration is very different from unity, as in the case of trace analysis, and if the minor component is eluted second in the tail of a major component, an artificial baseline is assumed by establishing a tangent to the major-component profile before and after the minor peak, or the profile of the tail may be interpolated and subtracted on the assumption of an exponential decay. Such a procedure can never be as satisfactory as achieving complete resolution of the peaks, and can in fact cause quite significant errors [58].

9.3. Calibration Techniques

Quantitative analysis requires the determination of individual response factors for the various components of interest in the analyzed mixture [58]. No detector gives the same response factor for two different compounds unless they are enantiomers. In practice, quantitative analysis requires that the detector be calibrated by injection of mixtures of known compositions containing one or two internal standards. Response factors with respect to the internal standards are then determined in the concentration range expected for the analytes. Response factors should never be extrapolated, even if there are good reasons to believe that the detector provides a linear response. The concentration of the analyte C_{an} is

$$C_{\mathrm{an}} = C_{\mathrm{is}} S_{\mathrm{an/is}} \frac{A_{\mathrm{an}}}{A_{\mathrm{is}}} \tag{52}$$

where $S_{\mathrm{an/is}}$ is the relative response factor for the analyte and A_{an} and A_{is} are the observed peak areas for analyte and internal standard. The need to calibrate the detector means that no quantitative analysis is possible for compounds that have not been identified. If no source of authentic compound of sufficient purity is available, calibration becomes difficult, but it is not impossible [58].

Use of an external standard, injected separately from the sample after a predetermined delay selected so as to avoid interferences between peaks of the standard and those of the mixture

components, represents an extremely useful calibration tool, and it also serves as a control for monitoring proper performance of the analytical instrument, especially in process-control analysis [58], [66]. The peak area of the external, deferred standard should remain constant for as long as the chromatograph is in good working order.

9.4. Sources of Error in Chromatography

Any fluctuation of an operating parameter of the chromatograph, including those that control the retention times and band widths of peaks and those that determine response factors, constitutes a source of error in chromatographic analysis [58]. Such parameters include the column temperature and the mobile-phase flow rate, as well as various detector settings [67].

10. Theory of Nonlinear Chromatography

Equilibrium isotherms are not linear as they are assumed to be in classical treatments of analytical chromatography. In the low concentration range used for analysis, observed band profiles are in good agreement with predictions made on the basis of replacing an isotherm by its initial tangent. When large samples are used, however, and the concentrations of mixture components are high, the nonlinear character of the thermodynamics of phase equilibria causes band profiles to deviate significantly from the classical Gaussian shape. Furthermore, the equilibrium composition of the stationary phase at high concentrations depends on the concentrations of all the components. In other words, the concentration of one particular compound in the stationary phase at equilibrium is a function of the mobile-phase concentrations of all the components. Thus, the elution profiles interact. In preparative chromatography, unlike analytical chromatography, the chromatogram obtained for a mixture is not simply the sum of those obtained when equivalent amounts of each component are injected separately [23].

The prediction of such band profiles requires the solution of several problems:

1) How does the nonlinear behavior of isotherms influence band profiles?
2) How do bands of incompletely resolved components interact?
3) What is the relative importance of thermodynamic effects compared to the kinetic sources of band broadening already known to exist from studies in linear chromatography?
4) How should one model competitive isotherms?

The answers to all these questions are now well known. The ideal model of chromatography responds to the first and second questions [68]. The equilibrium – dispersive model speaks to the first three questions, at least so long as the column efficiency under linear conditions exceeds 30 – 50 plates [22]. The last question is outside the scope of this review, but has been abundantly discussed in the literature [23].

Band profiles in chromatography are obtained as solutions to the differential mass-balance equations for the mixture components (Eq. 11). The models differ in the way such balance equations are simplified and solved.

10.1. The Ideal Model of Chromatography

In this model we assume that the column has infinite efficiency [69]. Thus, this model makes it possible to investigate the influence of the nonlinear thermodynamics of phase equilibria independently of the blurring, dispersive effects of axial dispersion and the finite rate of mass transfer. Neglecting one of the two major phenomena controlling band profiles in chromatography, it is as "theoretical" as linear chromatography. The results must be convoluted with those of linear chromatography to lead to a realistic description of actual chromatograms. Note that in linear, ideal chromatography, band profiles are identical to injection profiles, and the bands move at the same velocity as in linear chromatography.

For single-component bands it can be shown that, in the ideal model, a particular migration velocity can be associated with each concentration [22], [69]. However, this assigned velocity is a function of the concentration. Therefore, each

concentration leads to a well-defined retention time, which is a linear function of the differential of the isotherm q:

$$t_{\mathrm{R,C}}=t_0\left(1+F\frac{\mathrm{d}q}{\mathrm{d}C}\right) \tag{53}$$

where F is the phase ratio [see Eq. (11)]. If the isotherm is convex upward, as in most cases found in liquid chromatography, the retention time decreases with increasing concentration. High concentrations cannot pass lower ones, however, as there can be only one concentration at a time at any given point. The concentrations thus pile up at the front, creating a concentration discontinuity or shock. This shock migrates at its own velocity, which is a function of its height. It can be shown [22] that the maximum concentration of the shock is the solution of the equation

$$|q\,(C)-C\frac{\partial q}{\partial C}|=\frac{N}{Ft_0F_v} \tag{54}$$

The retention time is obtained by introducing the maximum concentration into Equation 53. In gas chromatography the isotherms, and especially the gas – liquid isotherms, tend to be convex downward. In this case the concentration shock forms at the rear of the band, which appears to be fronting [33]. However, another effect becomes important. The partial molar volumes of the sample components in the two phases, which are nearly equal in LC, are very different in GC. A sorption effect arises from the much higher velocity of the mobile phase in the band, where the component vapor adds to the mobile phase, relative to the region outside the band [33]. The higher the partial pressure of the vapor, the faster the band travels. Thus, bands eluted early display a front shock, later ones a rear shock. An intermediate band has a shock on both sides [70].

In the case of a binary mixture, an analytical solution exists only if the competitive equilibrium behavior is accounted for by the Langmuir isotherm

$$q_i=\frac{a_iC_i}{1+b_1C_1+b_2C_2} \tag{55}$$

where a_i and b_i are numerical coefficients characterizing the phase system and the compound. This solution has been discussed in detail [71]. An example is provided in Figure 1. Characteristics of this solution include:

1) The existence of two concentration shocks at $t_{\mathrm{R},1}$ and $t_{\mathrm{R},2}$, one for each of the two components
2) The fact that the positive shock of the second component concentration is accompanied by a partial, negative shock from the first component concentration (at $t_{\mathrm{R},2}$)
3) A concentration plateau of length Δt for the second component following the end of the elution profile of the first component
4) The fact that the profile for the end of the second component is the same as the profile of a pure band of that compound

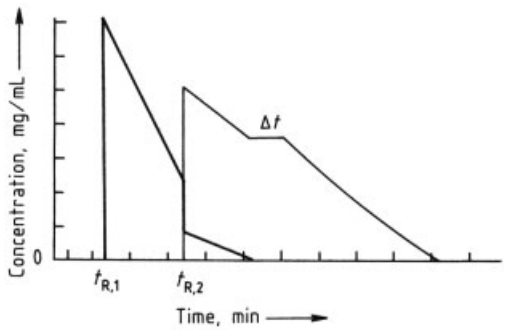

Figure 1. Solution of the ideal model for the elution of a narrow rectangular pulse of a binary mixture

The front of the second component displaces the first component (*displacement effect*), while the tail of the first component drags forward the front part of the second- component profile (*tag-along effect*). These effects have important consequences in preparative chromatography [22], [23].

10.2. The Equilibrium – Dispersive Model

In this model, constant equilibrium is still assumed between the two phases of the system, but finite column efficiency is accounted for by introducing a dispersion term in the mass-balance equation [22]. This term remains independent of the concentration. Solutions of the model must be obtained by numerical calculation. Several possible algorithms have been discussed. Excellent agreement has been reported between such solutions and experimental band profiles [72].

The solutions of the equilibrium – dispersive model exhibit the same features as those of the ideal model at high concentrations, where thermodynamic effects are dominant and dispersion due to finite column efficiency merely smoothes the edges. When concentrations decrease, the solutions tend toward the Gaussian profiles of linear chromatography.

10.3. Moderate Column Overloading

At moderate sample sizes, band profiles depend on the initial curvature of the isotherm, while retention times are functions of both initial slope and curvature [73]. Accordingly, retention varies linearly with increasing sample size [74]. The widespread belief that there is a range of sample sizes over which retention time is independent of sample size is erroneous. What happens is simply that, below a certain size, variation in the retention time is less than the precision of measurement. As the retention time decreases the band front or tail becomes steeper and the band becomes unsymmetrical.

If an isotherm can be reduced to the first two terms of a power expansion (i.e., to a parabolic isotherm), an approximate solution can be obtained by the method of Haarhoff and Van der Linde [75] in the case of a single-component band.

10.4. Preparative Chromatography

Preparative chromatography consists in using chromatography to purify a substance, or to extract one component from a mixture. Unlike analytical chromatography it is not devoted to the direct collection of information. This difference in objective results in marked differences in operation [23]. Production of finite amounts of purified compounds requires the injection of large samples, hence column operation under overloaded conditions. Resolution between individual bands in the chromatogram is no longer necessary provided mixed bands between the components are narrow. Instead, product recovery is the goal, demanding efficient condensers in the case of GC and rotary evaporators for LC. However, equivalent column performance is required in both analytical and preparative applications, and the packing materials must be of comparable quality.

11. Reference Material

11.1. Journals

The literature of chromatography is enormous. Several journals are dedicated entirely to this field, including the *Journal of Chromatography*, *Journal of Chromatographic Sciences*, and *Chromatographia*, while a larger number specialize in various subfields of chromatography (e.g., *Journal of L iquid Chromatography*, *Journal of Planar Chromatography*, *Journal of Microcolumn Separation*, *High Resolution Chromatography*). Several other journals publish one or several papers on chromatography in every issue. Important papers on chromatography are found in journals important in other scientific fields as well, ranging from *Biochemistry Journal* to *Chemical Engineering Science*.

One journal provides only abstracts of chromatographic papers appearing in other journals: *Chromatography Abstracts*, published by Elsevier for the Chromatographic Society. Chemical Abstracts also issues several *CA Selects* dedicated to specific aspects of chromatography.

11.2. Books

Probably the oldest general reference book still available is *Theoretische Grundlagen der Chromatographie* by G. Schay [76]. In the theoretical area, *Dynamics of Chromatography* by J. C. Giddings [12] is still a mine of useful discussions, comments, and ideas. A number of excellent books have been written about gas chromatography. *Gas Chromatography* by A. B. Littlewood [77], though obsolete in many respects, contains excellent fundamental discussions on GC and GC detectors. More practical, *The Practice of Gas Chromatography* by L. S. Ettre and A. Zlatkis [78] supplies much valuable information. Similarly, *Modern Practice of Liquid Chromatography* by J. J. Kirkland and L. R. Snyder [79] has long been the primary reference book for LC.

Chromatography, edited by E. Heftman [80], and *Chromatography Today* by Poole and

POOLE [81] contain excellent practical discussions, covering almost all aspects of implementing chromatographic separations. However, two important areas were virtually ignored: quantitative measurement and preparative separations. With respect to the former, *Quantitative Gas Chromatography* [58] is the only book that discusses in detail the sources of errors, their control, and problems of data acquisition and handling, which should not be left exclusively in the hands of computer specialists. For the latter, *Fundamentals of Preparative and Nonlinear Chromatography* [82] presents all aspects of the theory of chromatography, and discusses the optimization of preparative separations.

12. References

1. D. H. Everett, *Trans. Faraday Soc.* **61** (1965) 1637.
2. T. Tsuda, K. Nomura, G. Nakagawa, *J. Chromatogr.* **248** (1982) 241.
3. K. Sakodynski, *J. Chromatogr.* **73** (1972) 303.
4. L. S. Ettre in Cs. Horváth (ed.): *High Performance Liquid Chromatography: Advances and Perspectives,* **vol. 1,** Academic Press, New York 1980, p. 25.
5. M. S. Tswett, *Ber. Dtsch. Bot. Ges.* **24** (1906) 316.
6. M. S. Tswett, *Ber. Dtsch. Bot. Ges.* **24** (1906) 384.
7. M. S. Tswett: *Khromofilly v Rastitel'nom Zhivotnom Mire (Chromophylls in the Plant and Animal World),* Izd. Karbasnikov, Warzaw, Poland, 1910 (reprinted in part in 1946 by the publishing house of the Soviet Academy of Science, A. A. Rikhter, T. A. Krasnosel'skaya, eds.).
8. V. Berezkin: *M. S. Tswett,* Academic Press, New York 1992.
9. A. J. P. Martin, R. L. M. Synge, *Biochem. J.* **35** (1941) 1358.
10. R. Consden, A. M. Gordon, A. J. P. Martin, *Biochem. J.* **38** (1944) 224.
11. A. J. P. Martin, A. T. James, *Biochem. J.* **50** (1952) 679.
12. J. C. Giddings: *Dynamics of Chromatography,* Marcel Dekker, New York 1965.
13. K. K. Unger: *Porous Silica, Its Properties and Use as Support in Column Liquid Chromatography,* Elsevier, Amsterdam 1979.
14. A. Tiselius, S. Claeson, *Ark. Kemi Mineral. Geol.* **16** (1943) 18.
15. J. Frenz, Cs. Horváth in Cs. Horváth (ed.): *High-Performance Liquid Chromatography, Advances and Perspectives,* vol. 5, Academic Press, New York 1988.
16. S. Golshan-Shirazi, G. Guiochon in F. Dondi, G. Guiochon (eds.): "Theoretical Advancement in Chromatography and Related Separation Techniques," *NATO Adv. Study Inst. Ser. Ser. C* **383** (1992) 61.
17. L. C. Craig, *J. Biol. Chem.* **155** (1944) 519.
18. J. C. Giddings in E. Heftmann (ed.): *Chromatography,* Van Nostrand, New York 1975, p. 27.
19. J. C. Giddings, H. Eyring, *J. Phys. Chem.* **59** (1955) 416.
20. D. A. McQuarrie, *J. Chem. Phys.* **38** (1963) 437.
21. F. Dondi, M. Remelli, *J. Phys. Chem.* **90** (1986) 1885.
22. S. Golshan-Shirazi, G. Guiochon: "Theoretical Advancement in Chromatography and Related Separation Techniques," in F. Dondi, G. Guiochon (eds.): *NATO Adv. Study Inst. Ser. Ser. C* **383** (1992) 35.
23. A. M. Katti, G. Guiochon in J. C. Giddings, E. Grushka, P. R. Brown (eds.), *Adv. Chromatogr. (N.Y.)* **31** (1992) 1.
24. S. Golshan-Shirazi, G. Guiochon, *J. Chromatogr.* **603** (1992) 1.
25. L. Lapidus, N. L. Amundson, *J. Phys. Chem.* **56** (1952) 984.
26. J. J. Van Deemter, F. J. Zuiderweg, A. Klinkenberg, *Chem. Eng. Sci.* **5** (1956) 271.
27. M. J. E. Golay in D. H. Desty (ed.): *Gas Chromatography 1958,* Butterworths, London 1959, p. 36.
28. M. Kubin, *Collect. Czech. Chem. Commun.* **30** (1965) 2900.
29. E. Kucera, *J. Chromatogr.* **19** (1965) 237.
30. E. Grushka, *J. Phys. Chem.* **76** (1972) 2586.
31. L. J. Schmauch, *Anal. Chem.* **31** (1959) 225.
32. J. G. Atwood, M. J. E. Golay, *J. Chromatogr.* **218** (1981) 97.
33. G. Guiochon, L. Jacob, *Chromatogr. Rev.* **14** (1971) 77.
34. H. Darcy: *Les Fontaines Publiques de la Ville de Dijon,* Dalmont, Paris 1856.
35. D. E. Martire, R. E. Boehm, *J. Phys. Chem.* **91** (1987) 2433.
36. R. C. Reid, J. M. Prausnitz, B. E. Poling: *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York 1987.
37. A. Felinger, G. Guiochon, *Biotechnol. Progr.* **9** (1993) 421.

38. C. Vidal-Madjar, M. F. Gonnord, M. Goedert, G. Guiochon, *J. Phys. Chem.* **79** (1975) 732.
39. E. sz Kovats, *Helv. Chim. Acta* **41** (1958) 1915.
40. B. L. Karger, *Anal. Chem.* **39** (1967) no. 8, 24 A.
41. A. V. Kiselev, Ya. I. Yashin: *Gas Adsorption Chromatography,* Plenum Press, New York 1970.
42. L. R. Snyder in C. Horváth (ed.): *High Performance Liquid Chromatography, Advances and Perspectives,* **vol. 1,** Academic Press, New York 1980, p. 280.
43. M. Denkert, L. Hackzell, G. Schill, E. Sjögren, *J. Chromatogr.* **218** (1981) 31.
44. S. Golshan-Shirazi, G. Guiochon, *Anal. Chem.* **62** (1990) 923.
45. W. Melander, C. Horváth in C. Horváth (ed.): *High Performance Liquid Chromatography, Advances and Perspectives,* **vol. 2,** Wiley, New York 1980, p. 114.
46. J. H. Knox, *J. Chromatogr. Sci.* **15** (1977) 352.
47. G. Guiochon, A. Siouffi, *J. Chromatogr. Sci.* **16** (1978) 470.
48. J. H. Purnell, *J. Chem. Soc.* 1960, 1268.
49. B. L. Karger, W. D. Cooke, *Anal. Chem.* **36** (1964) 985.
50. G. Guiochon, *Anal. Chem.* **38** (1966) 1020.
51. G. Guiochon, E. Grushka, *J. Chromatogr. Sci.* **10** (1972) 649.
52. G. Guiochon in C. Horváth (ed.): *High Performance Liquid Chromatography, Advances and Perspectives,* **vol. 2,** Wiley, New York 1980, p. 1.
53. J. C. Sternberg in J. C. Giddings, R. A. Keller (eds.), *Adv. Chromatogr. (N.Y.)* **2** (1966) 205.
54. M. J. E. Golay, J. G. Atwood, *J. Chromatogr.* **186** (1979) 353.
55. J. M. Davis, J. C. Giddings, *Anal. Chem.* **57** (1985) 2178.
56. J. C. Giddings: *Unified Separation Science,* Wiley-Interscience, New York 1991.
57. A. Wehrli, E. sz Kovats, *Helv. Chim. Acta* **42** (1959) 2709.
58. G. Guiochon, C. Guillemin: *Quantitative Gas Chromatography,* Elsevier, Amsterdam 1988.
59. R. V. Golovnya, D. N. Grigoryeva, *Chromatographia* **17** (1983) 613.
60. H. Colin, E. Grushka, G. Guiochon, *J. Liq. Chromatogr.* **5** (1982) 1391.
61. A. J. P. Martin, *Discuss. Faraday Soc.* **7** (1949) 332.
62. D. A. Leathard, B. C. Shurlock: *Identification Techniques in Gas Chromatography,* Wiley-Interscience, New York 1970.
63. L. Rohrschneider, *J. Gas Chromatogr.* **6** (1968) 5.
64. R. W. Frei, J. F. Lawrence: *Chemical Derivatization in Analytical Chemistry, I. Chromatography,* Plenum Press, New York 1981.
65. M. Goedert, G. Guiochon, *J. Chromatogr. Sci.* **7** (1969) 323.
66. C. L. Guillemin et al., *J. Chromatogr. Sci.* **9** (1971) 155.
67. G. Guiochon, M. Goedert, L. Jacob in R. Stock (ed.): *Gas Chromatography 1970,* The Institute of Petroleum, London 1971, p. 160.
68. S. Golshan-Shirazi, G. Guiochon: "Theoretical Advancement in Chromatography and Related Separation Techniques," in F. Dondi, G. Guiochon (eds.): *NATO Adv. Study Inst. Ser. Ser. C* **383** (1992) 1.
69. D. DeVault, *J. Am. Chem. Soc.* **65** (1943) 532.
70. P. Valentin, G. Guiochon, *J. Chromatogr. Sci.* **10** (1975) 271.
71. S. Golshan-Shirazi, G. Guiochon, *J. Phys. Chem.* **93** (1989) 4143.
72. G. Guiochon et al., *Acc. Chem. Res.* **25** (1992) 366.
73. S. Golshan-Shirazi, G. Guiochon, *J. Chromatogr.* **506** (1990) 495.
74. A. Jaulmes, C. Vidal-Madjar, M. Gaspar, G. Guiochon, *J. Phys. Chem.* **88** (1984) 5385.
75. P. C. Haarhof, H. J. Van der Linde, *Anal. Chem.* **38** (1966) 573.
76. G. Schay: *Grundlagen der Chromatographie,* Springer Verlag, Berlin.
77. A. B. Littlewood: *Gas Chromatography,* 2nd ed., Academic Press, New York 1970.
78. L. S. Ettre, A. Zlatkis: *The Practice of Gas Chromatography,* Interscience, New York 1967.
79. J. J. Kirkland, L. R. Snyder: *Modern Practice of Liquid Chromatography,* Wiley-Interscience, New York 1971.
80. E. Heftman (ed.): *Chromatography,* 5th ed., Elsevier, Amsterdam 1993.
81. C. Poole, S. K. Poole: *Chromatography Today,* Elsevier, Amsterdam 1992.
82. G. Guiochon, S. Golshan-Shirazi, A. Katti: *Fundamentals of Preparative and Nonlinear Chromatography,* Academic Press, Boston 1994.

Gas Chromatography

PAT J. F. SANDRA, University of Stellenbosch, Matieland, South Africa; Universiteit Gent, Vakgroep Organische Chemie, Gent, Belgium

1. Introduction 3
2. Instrumental Modules 4
3. The Separation System 4
3.1. Modes of Gas Chromatography . 4
3.1.1. Gas – Liquid Chromatography . . . 4
3.1.2. Gas – Solid Chromatography 4
3.2. Selection of the Carrier Gas 5
3.3. Selection of the Gas Chromatographic Column 6
3.3.1. Packed Columns 6
3.3.2. Capillary Columns 7
3.4. Stationary Phases in Gas Chromatography 10
3.4.1. General Considerations 10
3.4.2. Solid Phases 12
3.4.3. Liquid Phases 14
3.4.3.1. Basic Phases for Capillary GC . . . 14
3.4.3.2. Selectivity Tuning 15
3.4.3.3. Tailor-Made Phases 16
4. Choice of Conditions of Analysis 17
4.1. Temperature-Programmed Analysis 17
4.2. Constant Pressure, Constant Flow, and Pressure Programming 18
5. Sample Inlet Systems 20
5.1. General Considerations 20
5.2. Universal Inlets 21
5.2.1. Packed Column Inlets 21
5.2.2. Capillary Column Inlets 21
5.2.2.1. Split Injection 22
5.2.2.2. Splitless Injection 23
5.2.2.3. Direct Injection 26
5.2.2.4. Cool On-Column Injection 28
5.2.2.5. Programmed-Temperature Injection 31
5.3. Selective Inlets 32
5.3.1. Static and Dynamic Head Space Analysis 33
5.3.2. Purge and Trap Systems 34
5.3.3. Thermal Desorption Units 34
5.3.4. Pyrolysis Gas Chromatography . . . 35
6. Detectors 35
6.1. Classification 35
6.2. Universal Detectors 36
6.2.1. Flame Ionization Detector 36
6.2.2. Thermal Conductivity Detector . . . 37
6.3. Selective Detectors 38
6.3.1. Electron Capture Detector 38
6.3.2. Nitrogen Phosphorus Detector . . . 40
6.3.3. Flame Photometric Detector 40
6.3.4. Overview of Other Selective Detectors 40
6.4. Detectors Allowing Selective Recognition 40
6.4.1. Mass Spectrometry 41
6.4.2. Fourier Transform Infrared Spectroscopy 42
6.4.3. Atomic Emission Detection 44
7. Practical Considerations in Qualitative and Quantitative Analysis 46
7.1. Qualitative Analysis 46
7.2. Quantitative Analysis 48
8. Coupled Systems 48
8.1. Multidimensional Capillary GC . 48
8.2. Multimodal High-Performance Liquid Chromatography – Capillary GC 51
8.3. Multimodal Supercritical Fluid Extraction – Capillary GC 53
9. Applicability 55
9.1. Solute Thermal Stability 55
9.2. Solute Volatility 55
9.3. Comparison of Gas Chromatography, Liquid Chromatography, and Supercritical Fluid Chromatography 56
10. Recent and Future Developments 59
10.1. Fast High Resolution Capillary GC 59
10.2. The Concept of Retention Time Locking 60
10.3. Towards Black Boxes 62
11. References 63

Abbreviations

AED atomic emission detection
CGC capillary gas chromatography
CI chemical ionization
CpSi cyanopropylsilicone
d_f film thickness

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_181

D_M	diffusion coefficient of the mobile phase
d_p	particle diameter
ECD	electron capture detector
EI	electron ionization
ELCD	Hall eletrolytic conductivity detector
EPC	electronic pressure control
FFAP	free fatty acid phase
FID	flame ionization detector
FPD	flame photometric detector
FSOT	fused silica open tubular
GALP	good automated laboratory practice
GLC	gas – liquid chromatography
GLP	good laboratory practice
GSC	gas – solid chromatography
H	column efficiency, plate height
HPLC	high performance liquid chromatography
HS	head space
i.d.	internal diameter
K	Kelvin temperature
k	retention factor
L	column length
MAOT	maximum allowable operating temperature
MDCGC	multidimensional capillary chromatography
MeSi	methylsilicone
MiAOT	minimum allowable operating temperature
N	plate number
NPD	nitrogen phosphorus detector
PCGC	packed column gas chromatography
PID	photoionization detector
PLOT	porous layer open tubular
PT	purge and trap
PTV	programmed-temperature vaporization
r	column radius
RCD	redox chemiluminescence detector
R_s	peak resolution
SFC	supercritical fluid chromatography
SFE	supercritical fluid extraction
SIM	selected ion monitoring
TCD	thermal conductivity detector
TD	thermal desorption
TEA	thermal energy analyzer
TID	thermionic detector
t_M	retention time of solvent
t_R	retention time of analyte
u	mobile phase velocity
WCOT	wall-coated open tubular
α	separation factor

1. Introduction

Gas chromatography (GC) is the separation technique that is based on the multiplicative distribution of the compounds to be separated between the two-phase system solid or liquid (*stationary phase*) and gas (*mobile phase*). Gas chromatography by definition thus comprises all separation methods in which the moving phase is gaseous. Contrary to the other chromatographic techniques, i.e., high-performance liquid chromatography (HPLC) and supercritical fluid chromatography (SFC), the role of the gaseous mobile phase – quasi-ideal inert gases such as nitrogen, helium, or hydrogen – is purely mechanical: they just serve for the transport of solutes along the column axis. The residence time (retention) of solutes is affected only by their vapor pressure, which depends on the temperature and on the intermolecular interaction between the solutes and the stationary phase.

Currently, gas chromatography is one of the most important and definitely the most economic of all separation methods. Its applicability ranges from the analysis of permanent gases and natural gas to heavy petroleum products (up to 130 carbon atoms, simulated distillation), oligosaccharides, lipids, etc. Moreover, as far as chromatographic efficiency and GC system selectivity is concerned, no other separation technique can compete with gas chromatography. The dictum "If the separation problem can be solved by gas chromatography, no other technique has to be tried out" is now generally accepted. As illustration, more than 80 % of the priority pollutants on the lists of the EC and of the EPA in the United States, are amenable to GC analysis.

Gas chromatography was developed in the early 1950s [8]. Tens of thousands of publications and more than 250 textbooks on theory and practice have been published. The purpose of this article is not to write a new textbook on GC but rather to help newcomers in the field to find their way in the present state of the art of modern GC. For the basic aspects of GC, see → Basic Principles of Chromatography. Fundamental equations are given or repeated only to clarify some important statements. The terms, symbols, and nomenclature used are those that have been advised by IUPAC [9].

At the end of the references, some general and specialized textbooks are listed, together with the most important journals dealing with gas chromatography. The reader is advised to consult these books and journals for a more detailed study.

2. Instrumental Modules

A schematic drawing of a modern gas chromatographic system is shown in Figure 1. The basic parts are the carrier gas supply (a), the injector (b), the column (c), and the detector (e). The dashed lines indicate thermostatted regions. The carrier gas is usually supplied from a high-pressure cylinder equipped with a two-stage pressure regulator. Via the controller of the GC instrument (f), the gas flow can be fine-tuned, and preheated carrier gas is delivered to the column in the constant-pressure, constant-flow, or pressure-programmed mode. The sample is introduced into the carrier gas stream via the injector, and the vaporized components are transported into the column, the heart of the system, where separation occurs. The column is placed in an oven, the temperature of which can be kept constant (*isothermal operation*) or programmed (*temperature-programmed operation*). Some applications are performed at subambient temperature and instruments can be equipped with cryogenic units. After separation, the component bands leave the column and are recorded (g) as a function of time (chromatogram) by a detection device (e). The chromatogram provides two types of data. The residence time or retention time of the components in the column is characteristic for the solute – stationary phase interaction and can be used for qualitative interpretation or component identification. The detector's response is proportional to the amount of separated sample component and gives quantitative information on the composition of the mixture.

3. The Separation System

3.1. Modes of Gas Chromatography

Two types of gas chromatography exist: gas – liquid chromatography (GLC) and gas – solid chromatography (GSC). Other classification schemes such as GSC, GLC plus capillary gas chromatography (CGC) are outdated because nowadays GLC and GSC can be performed both in packed columns and in capillary or open tubular columns.

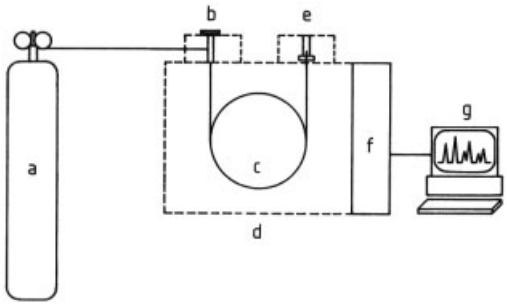

Figure 1. Basic components of a modern GC system
a) Carrier gas supply; b) Injector; c) Column; d) Column oven; e) Detector; f) Controller; g) Recorder

3.1.1. Gas – Liquid Chromatography

In GLC the stationary phase is a liquid acting as a solvent for the substances (solutes) to be separated. The liquid can be distributed in the form of a thin film on the surface of a solid support, which is then packed in a tube (packed column GC, PCGC) or on the wall of an open tube or capillary column (capillary GC, CGC). The term WCOT (*wall-coated open tubular*) is used only for capillary columns coated with a thin film of a liquid. In GLC the separation is based on partition of the components between the two phases, the stationary phase and the mobile phase, hence the term partition chromatography. This mode of GC is the most popular and most powerful one.

3.1.2. Gas – Solid Chromatography

Gas – solid chromatography comprises the techniques with an active solid as the stationary phase. Separation depends on differences in adsorption of the sample components on inorganic adsorbents (i.e., silica, alumina, carbon black) or on organic adsorbents such as styrene – divinylbenzene copolymers. Separation can also occur by a size exclusion mechanism, such as the separation of gases on synthetic

zeolites or molecular sieves. GSC is performed on packed columns or on open tubular columns on the walls of which a thin layer of the porous material is deposited [*porous layer open tubular* (PLOT) columns]. GSC nowadays is used only for special separation problems, and GSC columns are, therefore, referred to as tailor-made columns.

3.2. Selection of the Carrier Gas

Common carrier gases are nitrogen, helium, and hydrogen. The choice of carrier gas depends on the column and the detector used but, above all, on the required speed of analysis and detectability. The influence of the mobile-phase velocity u on column efficiency H has been detailed in → Basic Principles of Chromatography. The relation $H-u$ was derived for packed column GC by Van Deemter, Zuiderweg, and Klinkenberg [10] and adapted to capillary columns by Golay [11]. Practical consequences of the carrier gas selection in capillary GC have been described in [12], [13]. Due to the relatively high density of nitrogen compared to helium and hydrogen, longitudinal diffusional spreading of the solutes in nitrogen is small, but the resistance to mass transfer in the mobile phase is high. In contrast, for low-density gases such as helium and especially hydrogen, longitudinal diffusional spreading is large but resistance to mass transfer small. For columns with small amounts of stationary phases, the resistance to mass transfer in the stationary phase can be neglected, and the minimum plate height $H_{\min}$ is nearly independent of the nature of the carrier gas. The optimum mobile-phase velocity u_{opt}, corresponding to $H_{\min}$ is proportional to the diffusion coefficient of the mobile phase D_{M}; u_{opt} is 10 cm/s for nitrogen, 25 cm/s for helium, and 40 cm/s for hydrogen. In practical GC on thin-film columns, hydrogen is the best choice because the analysis time is reduced by a factor of 4 compared to nitrogen, while the detectability is increased by the same order. Some experimental $H-u$ curves for a *capillary column* 20 m in length and 0.27 mm internal diameter (*i.d.*) coated with a 0.22-µm film of methylsilicone (MeSi) are shown in Figure 2. The experimental data fit the theory perfectly. The fact that the slope (i.e., the increase of plate height with velocity) for nitrogen is much steeper than for the two other gases is noteworthy. Working close to u_{opt} is, therefore, a must with nitrogen, whereas for helium and hydrogen the velocity can be increased to values greater than u_{opt} without losing too much resolution power. For thin-film columns, hydrogen is definitely the best choice. When hydrogen is not allowed for safety reasons – most GC systems nowadays are equipped with hydrogen sensors – helium should be chosen. Because of the important contribution of the resistance to mass transfer in *thick-film columns*, this situation is more complicated. The selection depends on what is considered most important, efficiency or speed. This is dictated by the separation problem at hand. For a given column, the selection of carrier gas and optimum flow rate is best determined experimentally by making $H-u$ plots. This is not difficult and can be carried out in 3 – 4 h in the following way:

1) A solution of a hydrocarbon (i.e., tridecane) in a highly volatile solvent (i.e., hexane) is injected at an inlet pressure P_x of carrier gas A and a column temperature is selected to give a retention factor k of ca. 5.
2) The plate number N is calculated according to

$$N = 5.54\,(t_{\mathrm{R}}/w_{\mathrm{h}})^2 \tag{3.1}$$

   where t_{R} is the retention time of the hydrocarbon and w_{h} is the peak width at half height.
3) The H value is calculated from

$$H = L/N \tag{3.2}$$

   where L denotes the column length.
4) The corresponding u value is given by

$$u = L/t_{\mathrm{M}} \tag{3.3}$$

   where t_{M} is the retention time of the solvent; for thick-film columns, methane is injected instead of hexane to give t_{M}.
5) The H and u values corresponding to P_x of carrier gas A are marked in a $H-u$ plot (Fig. 3).
6) The inlet pressure P is set at values ${P_x}^{-3}$, ${P_x}^{-2}$, ${P_x}^{-1}$, ${P_x}^{+1}$, ${P_x}^{+2}$, ${P_x}^{+3}$, and the corresponding H and u values are drawn in the plot.

7) $H_{\min}$ and u_{opt} are deduced from the plot.
8) The same procedure is repeated for carrier gases B and C.

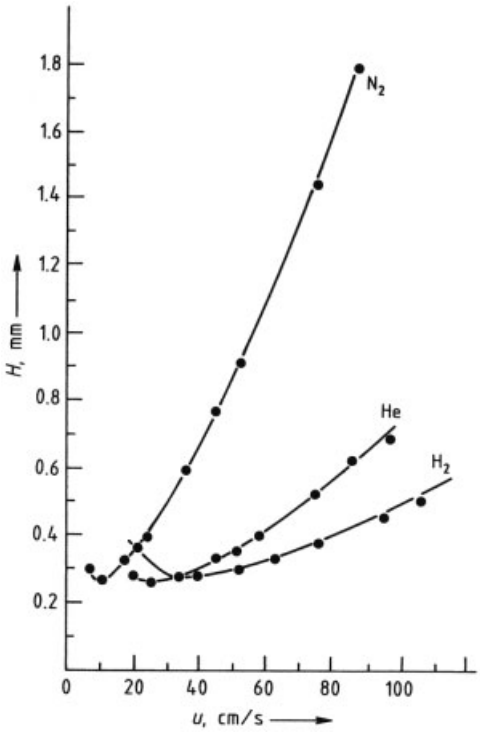

Figure 2. Experimental $H-u$ curves as function of carrier gas

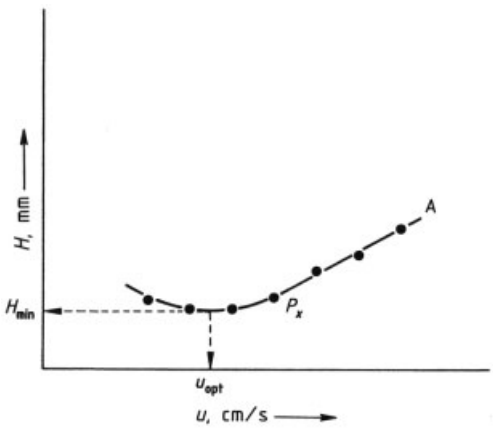

Figure 3. Making a $H-u$ plot

High-purity gases should be used in GC. Most gas suppliers have special GC-grade gases. If lower quality carrier gases are used, purification through a molecular sieve trap to remove moisture and low molecular mass hydrocarbons and through an oxytrap to remove oxygen is recommended.

3.3. Selection of the Gas Chromatographic Column

Two column types are in use, the packed and the capillary column. The selection is often dictated by the nature and complexity of the sample. The present trend is to replace PCGC by CGC whenever possible, because data obtained with the latter technique are much more reliable. The fundamental difference between the two methods is reflected in the resolution equation, the key equation for separation optimization:

$$R_S = \frac{\sqrt{N}}{4}\,\frac{\alpha-1}{\alpha}\,\frac{k}{k+1} \tag{3.4}$$

Anything that increases the column efficiency N, the column selectivity α, or the retention factor k will enhance the separation power of the column. Packed columns are characterized by low plate numbers and PCGC is therefore a low-resolution technique. The lower efficiency is compensated by the high selectivity α of the stationary phase, and this is the main reason why so many different stationary phases have been developed for PCGC. Capillary columns on the other hand have very high plate numbers and, therefore, the number of stationary phases can be restricted because the selectivity is less important. In fact, most separation problems can be handled with four basic stationary phases and half a dozen tailor-made stationary phases. Other important features of capillary columns are their inertness and compatibility with spectroscopic detectors. In the framework of this discussion, emphasis is, therefore, on capillary columns.

3.3.1. Packed Columns

A packed column is defined by the material of which the tube is made, by its length L and internal diameter *i.d.*, by the nature and diameter d_{p} of the particles, and of course by the

film thickness d_f and the nature of the stationary phase. The latter is most important because the stationary phase controls column selectivity (see Section 3.4). All other column parameters represent column efficiency. In packed column GC, H_{min} is controlled mainly by the particle diameter and is always $>2d_p$. A column 1 m long filled with particles of 100 – 120 mesh (150 – 125 µm) can never generate more than 3000 plates. Because of the packed bed, the resistance to the mobile flow is high, and columns longer than 3 m exhibit low permeability and are not practical. This means that a total plate number of 10 000 is the maximum that can be reached with packed columns. Most of the columns for routine analysis in fact have only 3000 to 5000 plates. The column efficiency is independent of column diameter. The column tube can be made of copper, stainless steel, nickel, glass, and polytetrafluoroethylene. Glass and polytetrafluoroethylene are the most inert materials and are preferred for the analysis of polar and thermolabile compounds. In *gas – liquid chromatography*, the support material has a silicon oxide network, such as diatomaceous earth, synthetic spherical silica, or glass beads. The support materials are purified by acid washing and deactivated by silanization, which caps the active sites. However, even the most highly deactivated surface will still exhibit adsorption of polar compounds such as organic acids and bases. A variety of support materials are available commercially, the catalogues of the column manufacturers (Agilent, Chrompack, Supelco, Alltech, etc.) should be consulted for detailed information. The stationary-phase film is homogeneously coated on the particles by dissolving the phase in a suitable solvent and adding the support material to the mixture. The solvent is then removed on a rotavapor (rotary evaporator), and after complete drying under a purified nitrogen flow, the column is packed tightly with the coated support. In PCGC, the contents of liquid coatings are expressed as a percentage of the support material: 1, 3, 5, and 10 wt % are commonly used.

The dimensions of columns used in *gas – solid chromatography* are very similar to those of gas–liquid chromatography. Packing materials (silica, alumina, carbon blacks, zeolites, porous polymers) in proper mesh size are available commercially to fill the columns. GSC columns give lower plate numbers than GLC columns but possess very high selectivities for some typical applications. Prepacked GLC and GSC columns can be purchased commercially and are ready to be installed in the GC instrument.

3.3.2. Capillary Columns

Major progress in gas chromatography, particularly in the area of applications, has been made by the introduction of capillary columns. Invented by M. Golay in 1957 [11], the real breakthrough came only in 1979 with the development of flexible fused silica as tubing material for capillary columns [14] (*fused silica open tubular columns*, FSOT). The inherent strength and flexibility of fused silica made use of capillary GC easy, compared to the glass capillary columns generally used previously. In addition, no other column material can offer the same inertness. A capillary column is characterized by its dimensions (*L, i.d.*) and by the nature of the stationary phase. For *GLC*, a relatively thin stationary (liquid) phase film, with film thickness d_f, is coated directly on the inner wall (WCOT columns). The film thickness controls the retention factor k and the sample capacity of a capillary column. In the case of GSC, the adsorbent with particle diameter d_p is deposited in a thin layer d_l on the capillary wall (PLOT columns). The main advantage of PLOT columns compared to packed GSC columns is the speed of analysis (see Section 3.4.2). Most of the applications (> 90 %) in capillary GC are performed on WCOT columns. These columns can be subdivided in three types according to their internal diameter. Columns with *i.d.* of 0.18 – 0.32 mm are called *conventional capillary columns* because this is the column type most often used for the analysis of complex samples. Columns with an *i.d.* of 0.53 mm are known as *widebore* or *megabore columns*. They have been introduced as an alternative to packed columns for the analysis of less complex mixtures. *Narrowbore columns* with *i.d.*'s ranging from 0.1 to 0.05 mm provide very high efficiencies or short analysis times. This column type is not often used nowadays, but this situation is expected to change in the next few years (see Section 10.1).

The selection of a capillary column depends on the complexity of the sample to be analyzed. The column length, the internal diameter, the stationary phase, and its film thickness determine the separation power (resolution), the sample capacity, the speed of analysis, and the detectability or sensitivity. Theoretical considerations [12], [13] indicate that for capillary columns with thin films (<1 µm), the $H_{\min}$ value is roughly equal to the column diameter. This is illustrated in Figure 4, which shows experimental $H-u$ curves for columns varying in internal diameter. H was calculated for dodecane at 100 °C with hydrogen as carrier gas. $H_{\min}$ measured experimentally is indeed very close to $H_{\min}$ deduced theoretically. By knowing this, the maximum plate number that a capillary column can provide may be calculated without performing any analysis:

$$N=L/H=L/H_{\min}=L/d_c \tag{3.5}$$

The $H_{\min}$ values, the maximum number of plates per meter, and the minimum column length needed for 100 000 plates for different *i.d.*'s of commercially available FSOT columns are listed in Table 1. In contrast to packed columns, the length of a capillary column is not restricted because the permeability of an open tube is very high. To a great extent, the length and internal diameter of a capillary column define the efficiency. The selection of the film thickness for a given stationary phase is also of upmost importance. The resolution equation (Eq. 3.4) shows that low k values result in poor resolution. For the same column efficiency N and selectivity α, increasing the k value from 0.25 to 5 corresponds to a resolution gain of 378 % (R_s increases from 0.37 to 1.40). Increasing the retention factor can be accomplished by increasing the film thickness

$$k=\frac{K\,2d_f}{r_c} \tag{3.6}$$

For a fixed column radius (r_c = constant), the retention factor of a solute at a given temperature (K = constant), doubles when the film thickness is increased by a factor of 2. Guidelines for the selection of film thickness for the stationary phase methylsilicone are listed below (values depend on column length and stationary phase):

≈0.1 µm	high molecular mass compounds	500 – 1500
≈0.2 µm	$M_r \approx 300-500$	
≈0.3 µm	$M_r \approx 150-400$	
≈0.5 µm	$M_r \approx 50-200$	
≈1 µm	$M_r \approx 30-150$	
2 – 5 µm	widebore – high-capacity narrowbore – high volatiles	

Table 1. Characteristics of FSOT columns coated with thin films

d_c, mm	$H_{\min}$, mm	N, m^{-1}	L for 100 000 plates, m
0.05	0.05	20 000	5
0.10	0.10	10 000	10
0.18	0.18	5 556	18
0.22	0.22	4 545	22
0.25	0.25	4 000	25
0.32	0.32	3 125	32
0.53	0.53	1 887	53

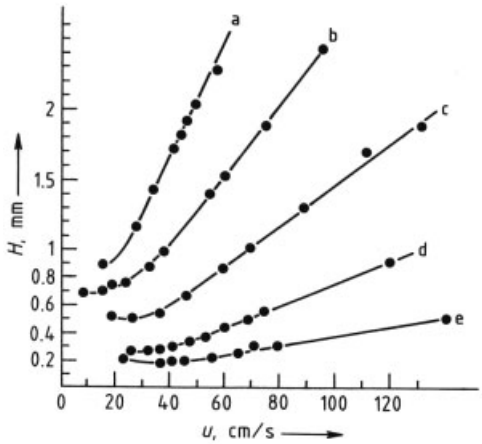

Figure 4. Experimental $H-u$ curves for columns with different *i.d.*'s
a) 0.88 mm; b) 0.70 mm; c) 0.512 mm; d) 0.27 mm; e) 0.18 mm

Conventional Capillary Columns. Capillary columns with *i.d.*'s of 0.18 – 0.32 mm and lengths of 10 – 60 m are the "workhorses" for high-resolution separation. Long columns are preferred for the analysis of very complex samples. The less complex the sample, the shorter the column can be. Figure 5 shows the analysis of the oxygen fraction of the essential oil of black pepper [15]. The profile is extremely complex (more than 300 compounds), and the column efficiency must be high to unravel all

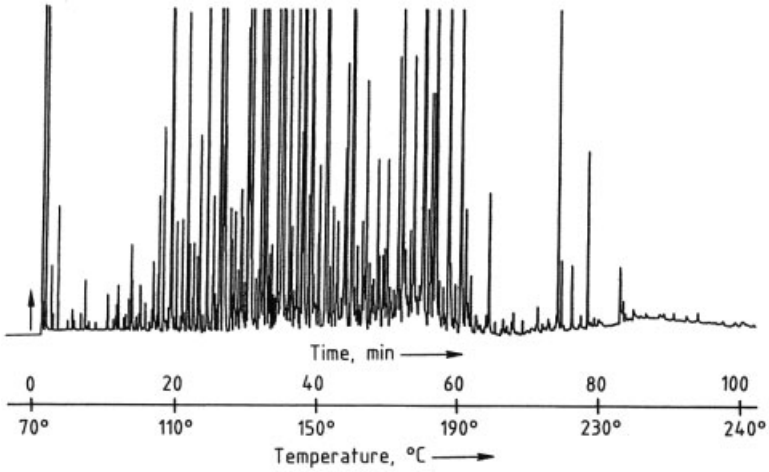

Figure 5. Analysis of oxygen-containing compounds of essential oil of pepper
Column: 40 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f high molecular mass poly(ethylene glycol); temperature as indicated; split injection 1/50; carrier gas hydrogen at 15 psi (≈1 bar); FID

solutes. Long columns, of course, imply long analysis times. Figure 6 shows the carbon number separation of the neutral lipids in palm oil. The separation is less challenging and the column length can be reduced to 5 m, which results in an analysis time of only 3 min [16]. For samples of unknown complexity, column lengths of 25 – 30 m offer a good compromise. Column dimensions are often adapted to the requirements of pre- or postcolumn devices. As an example, for the combination with mass spectroscopy, an internal diameter of 0.18 – 0.25 mm is preferred because in this case the flow of the mobile phase is of the order of 1 mL/min. This flow rate fits well with the pump capacity of the spectrometer so that special interfacing is not required.

Widebore Capillary Columns. FSOT capillary columns with *i.d.*'s of 0.53 mm were introduced as alternatives to packed columns, with the inertness of fused silica as their main advantage. Widebore or megabore columns have higher sample capacities and are applicable for simple packed column-like separations. A serious limitation of small *i.d.* capillary columns indeed is the low sample capacity (25 m $L \times 0.25$ mm *i.d.*; $d_f = 0.2$ µm results in a capacity of roughly 20 ng per compound). Sample capacity is the ability of a column to tolerate high concentrations of solutes. If the capacity is exceeded, the chromatographic performance is degraded as evidenced by leading (overloaded) peaks. Sample capacity is related to the film thickness and column radius. A column of 10 m × 0.53 mm, $d_f = 1$ µm, has a sample capacity >200 ng per compound. Simple inlet devices can be used with widebore columns in the low-resolution mode. In this mode, the carrier gas velocity is far above optimum values (flows of 8 – 12 mL/min), and a 10-m column generates only 5000 – 8000 plates. For simple mixtures, these efficiencies are sufficient. Megabore columns are also available in lengths of 30 – 60 m. At optimum velocities (flows of 1 – 3 mL/min) they offer 50 000 and 100 000 plates, respectively, which is considered high-resolution GC.

Narrowbore Capillary Columns. Narrowbore columns offer very high efficiencies (100 m $L \times 0.1$ mm *i.d.* = 1 000 000 plates) or ultrahigh speed at 100 000 plates but at the expense of sample capacity. This is illustrated in Figure 7 for the analysis of fatty acid methyl esters on a 10 m $L \times 0.1$ mm *i.d.*, d_f 0.1 µm, column of cyanopropylsilicone (CpSi). Total analysis time is only 90 s. Sample injection is critical because the sample capacity is <1 ng. This is the reason for the slow acceptance of narrowbore columns. Normal sample size injection is expected to be possible with the development of new sample

inlets such as programmed temperature vaporization in the solvent venting mode.

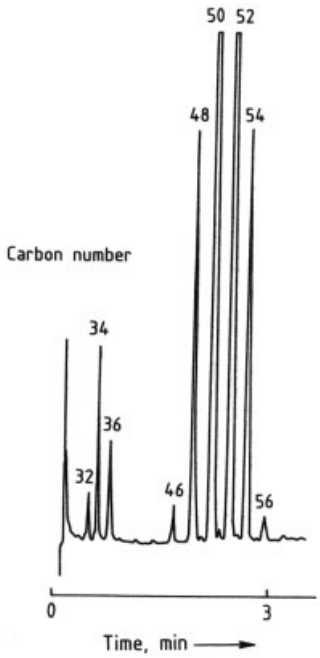

Figure 6. Fast separation of palm oil
Column: 5 m $L \times 0.25$ mm *i.d.*; 0.1 µm d_f methylsilicone; temperature: 290 – 350 °C at 30 °C/min; cool on-column injection; carrier gas hydrogen at 15 psi (≈1 bar); FIDPeaks: 32 – 36 diglycerides, 46 – 56 triglycerides

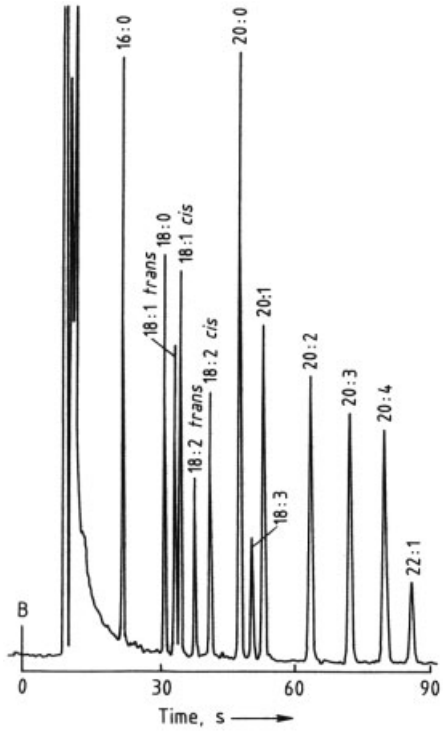

Figure 7. Fast separation of fatty acid methyl esters
Column: 10 m $L \times 0.1$ mm *i.d.*; 0.1 µm d_f cyanopropylsilicone; temperature: 180 °C; split injection 1/500; carrier gas hydrogen at 60 psi (≈ 4 bar); FID
Peaks: as indicated (the figure before the colon denotes the number of carbon atoms; that after the colon, the number of double bonds per molecule)

3.4. Stationary Phases in Gas Chromatography

3.4.1. General Considerations

The importance of the stationary phase in GC can, once again, be deduced from the resolution equation (Eq. 3.4). In Figure 8, the influence of the plate number N, the separation or selectivity factor α, and the retention factor k on peak resolution R_s is represented graphically. The peak resolution of two solutes ($k = 5$) with $\alpha = 1.05$ on a column with a plate number of 20 000 (e.g., 10 m × 0.5 mm coated with 0.2 µm methylsilicone) is 1.4. This value, the crossing point of the curves, is taken as reference. The curves are obtained by changing one of the variables while the other two remain constant. From the curves, the following conclusions can be drawn:

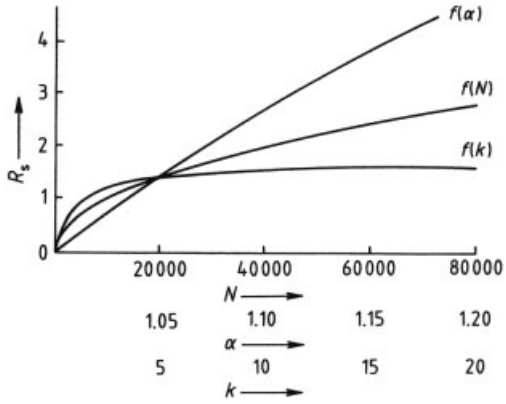

Figure 8. Effect of N, α, and k on R_s

Plate Number *N*. Peak resolution is proportional to the square root of the plate number.

Increasing the plate number by a factor of 2 ($N = 20\,000$ to $N = 40\,000$) improves peak resolution by only 1.4 ($\sqrt{2}$ gives $R_s = 1.40 \rightarrow 1.96$).

Retention Factor *k*. For values > 5, the influence of k on R_s is small. Increasing k from 5 to 20 increases resolution by only 14 % ($R_s = 1.40 \rightarrow 1.60$).

Separation Factor α. Optimization of the separation factor has the most important impact on peak resolution. Increasing α from 1.05 to 1.10 nearly doubles the peak resolution ($R_s = 1.40 \rightarrow 2.68$), while a slight decrease ($\alpha = 1.03$) causes a drastic loss of resolution ($R_s = 1.40 \rightarrow 0.86$). The selection of a "liquid phase" that maximizes the relative retention of solutes is thus of upmost importance. The *x*-scales in Figure 8 have not been selected arbitrarily. Varying plate number and retention factor, respectively, from $N = 20\,000$ to 40 000 and from $k = 5$ to 20 is not so difficult to realize. In contrast, increasing the separation factor is not easy to achieve, especially in gas chromatography. First of all, as already mentioned, the mobile phase is an inert gas and its role is purely mechanical. Second, optimization of the selectivity of the stationary phase for a particular sample means that the nature of the solutes must be known. Optimization then requires good chemical insight and intuition. Several separations have been achieved more by luck than by skill. In 1991 a book was published with very detailed discussion of practically every stationary phase used since the inception of gas chromatography [17]. The reader is referred to this book for more information.

The selectivity of a stationary phase is commonly expressed in terms of the relative retention of a compound pair:

$$\alpha = k_2/k_1 \tag{3.7}$$

The thermodynamic description of Equation (3.7) is

$$\alpha = p_1^\circ/p_2^\circ \cdot \gamma_1^\circ/\gamma_2^\circ \tag{3.8}$$

where p° refers to the saturated vapor pressure of the analyte and γ° is the activity of the analyte at infinite dilution in the stationary phase. Equation (3.8) is split in two contributions: the first term is related to the solutes and depends only on temperature, whereas the second term depends on the selective properties of the stationary phases. Solute – stationary phase interactions are very complex and may include nonspecific interactions (dispersion interactions); specific interactions such as dipole – dipole, dipole – induced dipole, and hydrogen bonds; and chemical interactions such as acid – base, proton donor – proton acceptor, electron donor – electron acceptor, and inclusion. The terms *selectivity* and *polarity* are commonly used to describe a stationary phase. A polarity and selectivity scale of more than 200 stationary phases has been established by McReynolds [18]. The Kovats retention indices (see Chap. 7) of benzene, butanol, 2-pentanone, nitropropane, and pyridine on phase *X* are measured, and the differences in indices in phase *X*, relative to the stationary phase squalane as typical apolar phase, are listed individually (selectivity) and as a sum (polarity). Although the polarity scale is discussed in all textbooks on GC, it has little practical value. The following are by far the most common methods used for selection of a stationary phase: trying a phase with good chromatographic properties, looking for a similar application that has already been published, or asking a colleague with experience (trial and error) for advice. At this stage the difference between PCGC and CGC should be restressed. Consider a packed column and a capillary column coated with the same stationary phase of the following characteristics: *packed column,* 2 m $L \times 4$ mm *i.d.*, packed with 100 – 120 mesh diatomaceous earth particles, 3 % methylsilicone, plate number 5000; *capillary column*, 25 m $L \times 0.25$ mm *i.d.*, coated with 0.25 µm methylsilicone, plate number 100 000. By neglecting the k contribution in the resolution equation, the resolution for two solutes with different α values can be calculated (Table 2). For an α value of 1.05, the resolution on the packed column is insufficient and another stationary phase must be selected. On the capillary column for the same solute – stationary phase interaction the resolution is still 3.6 because of the very high efficiency of a capillary column. Even for α values of 1.02, baseline separation is obtained. In the case of α value 1.01, the resolution is only 0.8, but due to the high permeability of a capillary column, its length can be increased

to provide the required plate number N_{req} on the same stationary phase. From the resolution equation follows:

$$N_{\mathrm{req}} = 16R_{\mathrm{s}}^2\ (\alpha/\alpha-1)^2 = 160000\mathrm{plates} \qquad (3.9)$$

and as $N = L/d_{\mathrm{c}}$, the required length for $R_{\mathrm{s}} = 1$ is 40 m. The analysis of a fatty acid methyl ester mixture on a *packed column* with a highly selective phase for these solutes (ethylene glycol succinate) is shown in Figure 9; several peaks are incompletely separated. The same analysis was performed on a capillary column coated with poly(ethylene glycol), a stationary phase with no particular selectivity for the solutes (Fig. 10) : all peaks are baseline separated.

Table 2. Resolution R_{s} for different α values

α	R_{s}	
	GC	CGC
1.2	2.9	12.5
1.1	1.6	6.8
1.05	0.8	3.6
1.02	↓	1.5
1.01		0.8
	other stationary phase	↓
		other stationary phase or increase of column length

3.4.2. Solid Phases

For certain applications, GSC enjoys some advantages over GLC. Adsorbents are stable over awide temperature range, and column bleed is virtually nonexistent especially for inorganic adsorbents and molecular sieves. *Silica* and *alumina* adsorbents give excellent separation of saturated and unsaturated hydrocarbons with low molar mass. Medium-polarity and polar solutes interact too strongly with the highly adsorptive surface of these adsorbents. *Graphitized carbon blacks* are well suited for separation of structural and geometric isomers. The disadvantage of carbon blacks is that their retention characteristics differ from batch to batch and they seem difficult to prepare in a reproducible way. Inorganic adsorption columns can be modified by coating the material with a small amount of nonvolatile liquid or inorganic salt. The most adsorptive sites will bind preferentially to these modifiers. As a result, the sites are unavailable to participate in retention, and retention times generally decrease with increased reproducibility. *Molecular sieves* (zeolites) are aluminosilicates of alkali metals. The most common types for GC are molecular sieves 5A and 13X. Separation is based mainly on the size of the solutes and the porous structure of the molecular sieve, although adsorptive interactions inside and outside the pores may also contribute. Molecular sieves are used primarily for the separation of permanent and inert gases including low molar mass hydrocarbons. *Porous polymer beads* of different properties have found many applications in the analysis of volatile inorganic and organic compounds. The properties of the commercially available materials vary by chemical composition, pore structure, and surface area. Compared to inorganic adsorbents, polymer beads are stable to water injections, and on some of the materials, polar solutes give perfect peak shapes in reasonable analysis times. The best-known adsorbents are the Porapak series, the Chromosorb 101-108 series, and Tenax. The range of applications on those polymers can be found in catalogues of the column and instrument manufacturers. Tenax (a linear polymer of *para*-2,6-diphenylphenylene oxide) is somewhat unique because of its thermal stability up to 380 °C. Relatively high molecular mass polar compounds such as diols, phenols, or ethanolamines can be analyzed successfully. Some doubt exists as to the mechanism of retention on Tenax. At low temperature, adsorption appears to be the principal retention mechanism, but at higher temperature the surface structure becomes liquid-like and partitioning does occur.

Solid Phases for PLOT Columns. At the moment, GSC is enjoying a revival because of the introduction and commercialization of PLOT columns. The advantages of these columns are higher efficiency, speed of analysis, and better stability and reproducibility over time [19]. To date, PLOT columns are available in different lengths and internal diameters, with aluminum oxide – KCl for the separation of saturated and unsaturated hydrocarbons including low-volatile aromatics; with molecular sieve 5A for the analysis of permanent and inert gases; with porous polymers of the Porapak-type (Q,S,U) (PoraPlot columns) for the analysis of

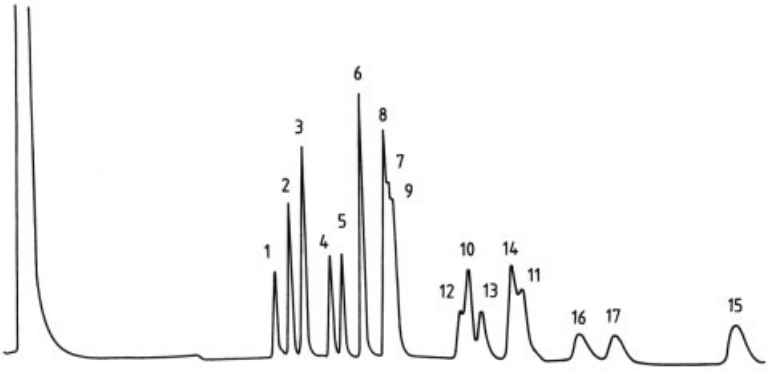

Figure 9. Isothermal analysis (200 °C) of fatty acid methyl esters on a packed column with stationary phase ethylene glycol succinate (EGS-X)
1. $C_{16}:0$; 2. $C_{16}:1$; 3. $C_{17}:0$; 4. $C_{18}:0$; 5. $C_{18}:1$; 6. $C_{18}:2$; 7. $C_{18}:3$; 8. $C_{20}:0$; 9. $C_{20}:1$; 10. $C_{20}:4$; 11. $C_{20}:5$; 12. $C_{22}:0$; 13. $C_{22}:1$; 14. $C_{23}:0$; 15. $C_{22}:6$; 16. $C_{24}:0$; 17. $C_{24}:1$

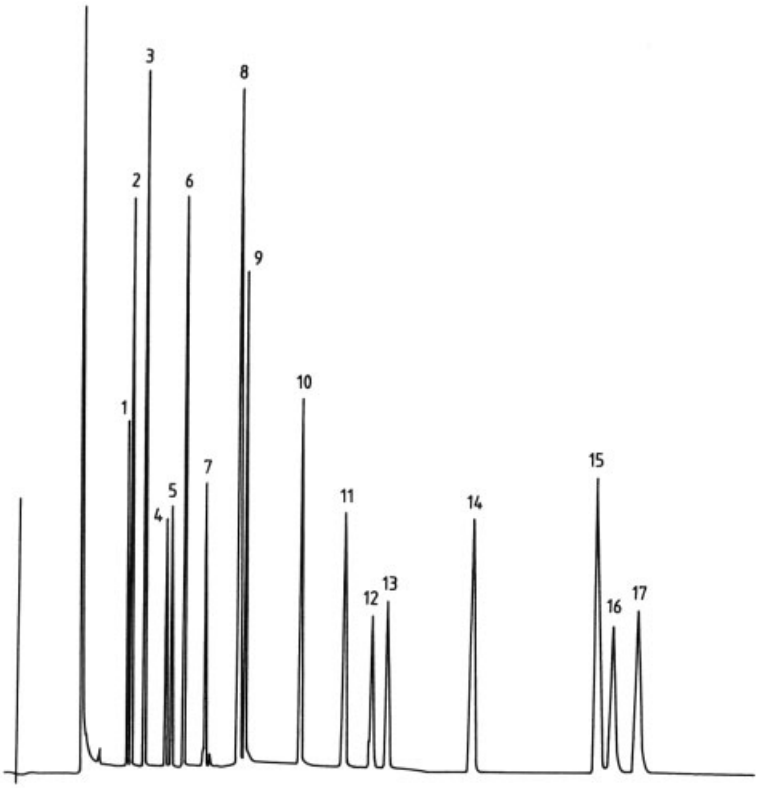

Figure 10. Isothermal analysis (200 °C) of fatty acid methyl esters on a capillary column with high-M_r poly(ethylene glycol) as stationary phase (for designation of peaks, see Fig. 9)

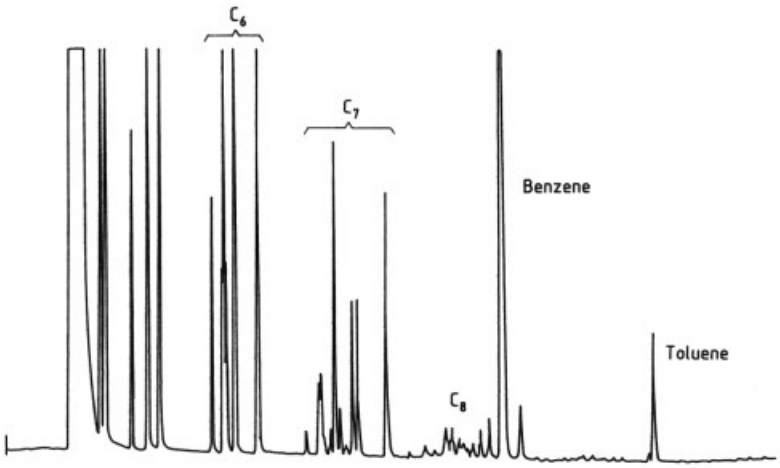

Figure 11. Natural gas analysis on an Al_2O_3 PLOT column
Column: 25 m $L \times 0.25$ mm *i.d.*; Al_2O_3 – KCl; temperature: 75 – 200 °C at 3 °C/min; split injection; carrier gas nitrogen at 5 psi ($\approx 3.5 \times 10^4$ Pa); FID

polar solutes and water; and with carbon black (CarboPlot columns) for volatile apolar solutes in petroleum products and environmental samples. Figure 11 shows the determination of benzene and toluene in natural gas on an Al_2O_3 – KCl PLOT column. The elution of benzene from the C_6 group and of toluene from the C_7 group illustrates the high selectivity.

3.4.3. Liquid Phases

In the past, numerous phases were used in PCGC, leading to what has been called "stationary-phase pollution" [20]. In the beginning, the many phases were a blessing; later, a curse. Today most separations are performed on a dozen preferred phases, which are all specially synthesized for GLC purposes. Some popular PCGC phases, listed according to increasing polarity, are

- Methylsilicone (gum)
- Methylsilicone (fluid)
- Methylphenyl(5 %)silicone
- Methylphenyl(50 %)silicone
- Methyltrifluoropropyl(50 %)silicone
- Methylphenyl(25 %)cyanopropyl(25 %)silicone
- Poly(ethylene glycol), $M_r > 40\,000$
- Cyanopropyl(50 %)phenylsilicone
- Poly(ethylene glycol) esterified with 2-nitroterephthalic acid
- Diethylene glycol succinate
- Cyanopropyl(100 %)silicone
- Ethylene glycol succinate
- 1,2,3-Tris(2-cyanoethoxy)propane

Of this list, by far the most important group are the silicones. The reasons can be summarized as follows: Their diffusion coefficients are high, and thus resistance to mass transfer is low, silicones have a high thermal stability (methylsilicone gum up to 400 °C) and are liquids or gums at low temperature, the structure is well defined, and the polymers are very pure. For practical GC this results in a broad temperature range, which is defined by the minimum allowable operating temperature range (MiAOT) and by the maximum allowable operating temperature range (MAOT), and low column bleeding. Poly(ethylene glycol) with high molecular mass, which has a unique selectivity, is a solid below 60 °C (MiAOT = 60 °C) and bleeds at 240 – 250 °C as a result of the higher vapor pressure compared to silicones.

3.4.3.1. Basic Phases for Capillary GC

As already mentioned, the number of phases in CGC can be reduced further due to the high efficiency those columns offer. Stark et al. [21] studied stationary-phase characteristics in depth and concluded that the order of optimized phases

for CGC, all higher molecular mass crosslinkable gums, is

- Methylsilicone
- Methylphenyl(50 – 70 %)silicone
- Methylcyanopropylsilicones with medium (25 – 50 %) and high (70 – 90 %) cyanopropyl content
- Methyltrifluoropropylsilicone
- High molecular mass poly(ethylene glycol) or a silicone substitute having the same polarity and selectivity

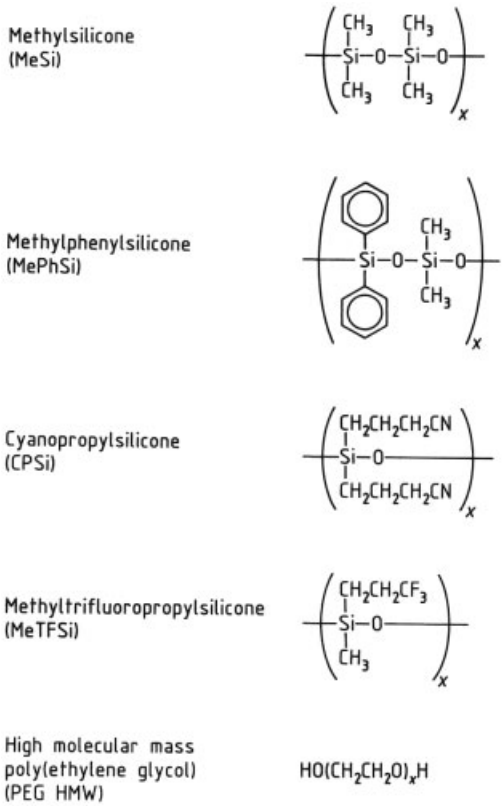

Figure 12. Structures of the basic phases for capillary GC

The functionalities in those phases are the methyl, phenyl, cyano, trifluoro, and hydroxy – ether group, respectively. The similarity with HPLC phases is remarkable. Based on the above, SANDRA et al. [22] proposed five basic phases for capillary GC. The formulas are given in Figure 12. For the majority of high-resolution separations, these stationary phases provide more than adequate performance. For some applications, the functionalities are combined on the same siloxane backbone, or columns are coupled (selectivity tuning). Functionalities can also be modified to provide specific interactions. Optical phases and liquid crystals complete the set of preferred CGC phases. These tailor-made phases are discussed in Section 3.4.3.3.

The most important phase is *methylsilicone*. Separations occur according to the boiling point differences of the solutes. A typical chromatogram is shown in Figure 13. Fatty acid methyl esters are separated according to their boiling point or molar mass. Hardly any separation occurs among oleic (C_{18} : 1 *cis*), elaidic (C_{18} : 1 *trans*), linoleic (C_{18} : 2) and linolenic (C_{18} : 3) acids. *Methylphenylsilicone* is the second most widely used stationary phase. Selectivity is increased due to the polarizable phenyl group. *Methyltrifluoropropylsilicone* possess rather unique selectivity for electron donor solutes such as ketones and nitro groups. Incorporating the *cyano group* with its large dipole moment into the silicone backbone provides strong interactions with dipolar solutes and unsaturated bonds. Cyano phases are widely used for the separation of unsaturated fatty acid methylesters and for the resolution of dioxins and furans. Figure 14 represents the analysis of a sample whose composition is very similar to that in Figure 13. Separation according to the number of double bonds in the alkyl chains is due to $\pi - \pi$ interactions. Methyl oleate (*cis* C_{18} : 1) and methyl elaidate (*trans* C_{18} : 1) differing only in configuration are baseline separated as well. High *molecular mass poly(ethylene glycol*) has a unique selectivity and polarity for the separation of medium-polarity and polar compounds. The most popular of these phases is Carbowax 20 M with average M_r of 20 000.

3.4.3.2. Selectivity Tuning

A solvent with a selectivity intermediate to the selectivities of the five basic phases can, for some applications, offer a better separation or a shorter analysis time. This is illustrated by the following example. On stationary phase *X*, compounds A and B can be separated from C, but they cannot be separated from each other. On another stationary phase *Y*, A and B are separated, but one of them coelutes with C. By using a chromatographic system with selectivity between *X*

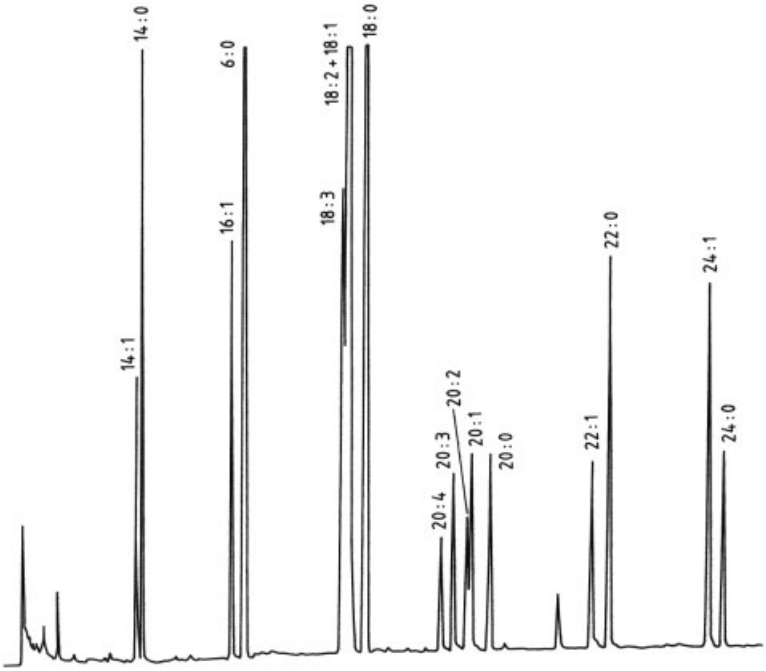

Figure 13. Analysis of fatty acid methyl esters on methylsilicone
Column: 25 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f MeSi; temperature: 100 – 220 °C at 3 °C/min; split injection 1/100; carrier gas hydrogen at 8 psi ($\approx 5.6 \times 10^4$ Pa); FID
(For designation of peaks, see Fig. 7)

and Y, an optimum can be found to separate the three compounds. Intermediate selectivities and polarities can be obtained by selectivity tuning [23]. Selectivity tuning means that the selectivity is adapted to the analytical need by creating a selectivity between two (or more) extreme (basic phase) selectivities. In principle, this can be performed in three different ways: (1) by synthesizing a tuned phase with predetermined amounts of monomers containing the required functional groups; (2) by mixing, in different ratios, two or three basic phases in one column; and (3) by coupling two or more columns of extreme selectivities. Approach (1) and (2) have been commercially used, and columns are available for specific applications. Columns tuned for the analysis of classes of priority pollutants are best known. Method (3) is detailed in Section 8.1.

3.4.3.3. Tailor-Made Phases

The most widely used tailor-made phases are the free fatty acid phase (FFAP), liquid crystals, and the optical phases Chirasil-Val and derivatized cyclodextrins. *FFAP* is produced by condensing 2-nitroterephthalic acid with Carbowax 20M. The phase is recommended for the analysis of acidic compounds such as organic acids and phenols. The phase is not suitable for the analysis of alkaline compounds and aldehydes with which it reacts. *Liquid crystals* can separate isomers as a function of the solute length-to-breadth ratios. High selectivities have been noted for the separation of polycyclic aromatic compounds and dioxin isomers [24]. The most powerful involves the use of a liquid-crystal column as second column in a multidimensional CGC system. Optical isomers (enantiomers) can be separated principally by two different approaches. Enantiomeric mixtures containing a reactive group can be derivatized with an optically pure reagent to produce a mixture of diastereomers that can be separated on nonoptically active stationary phases such as methylsilicone. The direct separation on optically active stationary phases is, however, the current method of choice because it has certain ad-

vantages. Derivatization agents of 100 % enantiomeric purity are not needed, and the method can be applied to enantiomers lacking reactive functional groups. In Chirasil-Val a diamide is incorporated into a silicone backbone, and separation occurs via a triple hydrogen interaction. A number of derivatized α-, β-, and γ-cyclodextrin stationary phases have been introduced for the separation of enantiomeric pairs. Separation is based on the formation of inclusion complexes with the chiral cavity of the cyclodextrins. However, none of the phases is universally applicable, and enantiomer separation is still based on trial and error. The separation of γ- and δ-lactones on a 2,3-*O*-acetyl-6-*tert*-butyldimethylsilyl-β-cyclodextrin stationary phase is shown in Figure 15. Full details on the separation of enantiomers by capillary GC can be found in [25], [26].

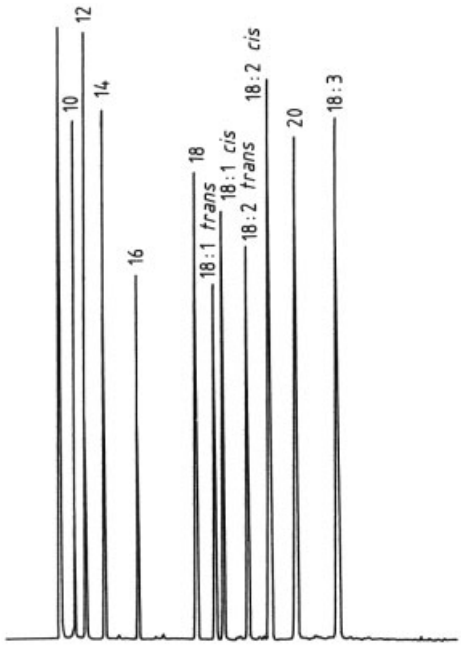

Figure 14. Analysis of fatty acid methyl esters on a cyanopropylsilicone
Column: 25 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f CPSi; temperature: 187 °C; split injection 1/100; carrier gas hydrogen at 8 psi ($\approx 5.6 \times 10^4$ Pa); FID

4. Choice of Conditions of Analysis

Under isothermal and constant-flow conditions, for a homologous series of hydrocarbons, an exponential correlation exists between the residence time in the stationary phase (adjusted retention time) and the solute vapor pressure, boiling point, or number of methylene groups (carbon number).

$$\log t'_R = a\,\mathrm{C}_n + b \tag{4.1}$$

where t'_R is the adjusted retention time; C_n is the number of carbons in the individual homologues; and *a, b* are constants. For a mixture with boiling point range of the solutes less than 100 °C, an optimum isothermal temperature can be selected. For samples in which the boiling point difference of the solutes exceeds ca. 100 °C, working in the isothermal mode is impractical. Usually, early eluting peaks will show poor resolution (low k values), whereas late eluting peaks will be broad when a compromise temperature has been chosen. The problem can be solved by temperature or flow programming.

4.1. Temperature-Programmed Analysis

In temperature-programmed analysis, a controlled change of the column temperature occurs as a function of time. The initial temperature, heating rate, and terminal temperature must be adjusted to the particular separation problem. The initial temperature is chosen so that the low-boiling compounds are optimally separated (k value > 3). Selection of the program rate depends on the nature of the solutes and the complexity of the sample. The final temperature is adjusted to give reasonable total analysis times. Temperature programming has been applied since the beginning of gas chromatography (1960), but only since the end of the 1980s have the possibilities been fully exploited. This might be, in part, because of the time-consuming nature of temperature optimization by trial and error, but above all because of ignorance of the importance of the temperature programming rate to selectivity. This is especially so for polar columns where $\Delta t_\mathrm{R}/\Delta T$ is much more pronounced than on apolar columns. Today, software is available to optimize a temperature program for a particular separation on the basis of only two experimental programmed runs. The theory of temperature-programmed

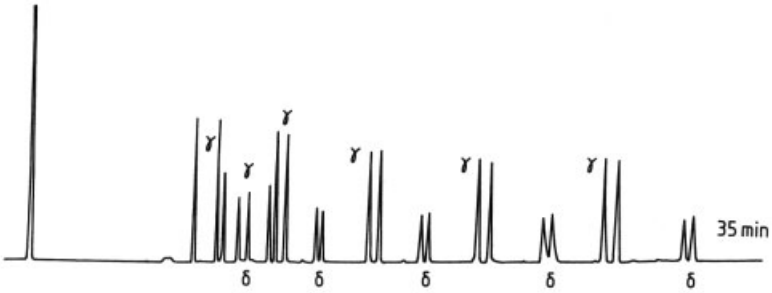

Figure 15. Enantioselective separation of γ- and δ-lactones on 2,3-*O*-acetyl-6-*tert*-butyldimethylsilyl-β-cyclodextrin
Column: 25 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f; temperature: 150 – 185 °C at 1 °C/min; split injection 1/100; carrier gas nitrogen at 18 psi (≈0.13 MPa); FID
Peaks: γ- (6 – 11) and δ-lactones (8 – 12)

Table 3. Retention times (n = 5) in initial runs, and predicted and experimental retention times at the optimum programming rate

Compound		Symbol	Initial runs		Optimized program	
No.	Name		1 °C/min, min	12 °C/min, min	Predicted, min	Experimental, min
1	Methyl butanoate	$C_4:0$	2.95	2.49	2.07	2.21
2	Methyl hexanoate	$C_6:0$	6.30	3.74	2.50	2.73
3	Methyl octadecanoate	$C_8:0$	14.65	5.38	3.01	3.32
4	Methyl decanoate	$C_{10}:0$	28.64	7.09	3.52	3.88
5	Methyl dodecanoate	$C_{12}:0$	45.02	8.69	4.00	4.41
6	Methyl phenylacetate	I.S. ∗	49.51	9.68	4.37	4.85
7	Methyl tetradecanoate	$C_{14}:0$	61.15	10.14	4.44	4.88
8	Methyl hexadecanoate	$C_{16}:0$	76.20	11.45	4.83	5.32
9	Methyl *cis*-9-hexadecanoate	$C_{16}:1$ *cis* 9	79.38	11.85	4.96	5.47
10	Methyl octadecanoate	$C_{18}:0$	90.08	12.65	5.19	5.72
11	Methyl *cis*-9-octadecanoate	$C_{18}:1$ *cis* 9	92.43	12.97	5.31	5.85
12	Methyl *cis*-9,12-octadecadienoate	$C_{18}:2$	97.23	13.47	5.48	6.04
13	Methyl *cis*-6,9,12-octadecatrienoate	$C_{18}:3\ \gamma$	100.56	13.82	5.61	6.19
14	Methyl eicosanoate	$C_{20}:0$	102.95	13.75	5.55	6.11
15	Methyl *cis*-9,12,15-octadecatrienoate	$C_{18}:3\ \alpha$	102.95	14.03	5.68	6.27
16	Methyl docosanoate	$C_{22}:0$	114.64	14.77	5.92	6.49
17	Methyl *cis*-5,8,11,14-eicosatetraenoate	$C_{20}:4$	114.90	15.10	6.11	6.71
18	Methyl *cis*-4,7,10,13,16-eicosa-pentaenoate	$C_{20}:5$	120.20	15.64	6.39	6.89
19	Methyl tetracosanoate	$C_{24}:0$	126.08	15.72	6.34	6.97
20	Methyl *cis*-4,7,10,13,16,19-docosa-hexaenoate	$C_{22}:6$	133.60	16.83	7.15	7.64

∗ I.S. = Internal standard.

optimization is treated in detail in [27], [28]. Figure 16 shows the computer predicted (A) and experimental (B) chromatograms for a complex fatty acid methyl ester mixture (Table 3) at a programmed temperature increase of 40 °C/min [29]. The column was 55 m in length, and a remarkably good separation was achieved in < 8 min. Without the aid of computer simulation, such fast programming rates would never have been considered for this separation.

4.2. Constant Pressure, Constant Flow, and Pressure Programming

Flow programming offers advantages such as reduced separation time, less column bleed, and lower eluting temperature for labile compounds but has not been widely used in gas chromatography because temperature programming is experimentally easier to perform. In 1990, electronic pressure control (EPC) [30] was developed, which allows operation in the

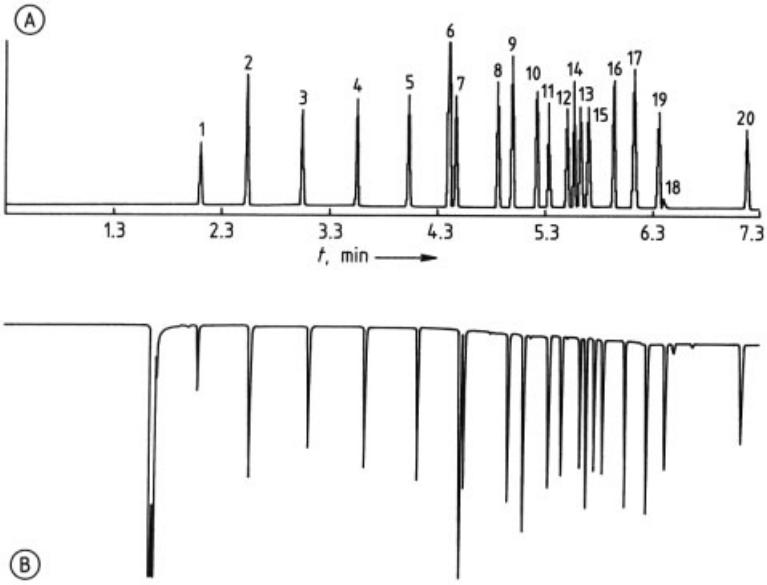

Figure 16. Predicted (A) and experimental (B) chromatograms for fatty acid methyl ester mixture with temperature-programmed analysis at 40 °C/min

constant-pressure, constant-flow, and pressure-programming modes. Whether temperature programming will be replaced by flow programming is questionable, although the features of EPC open new perspectives. Gases play a vital role in GC and are even more important in CGC. The carrier gas is the force moving the solutes through the column, and its velocity or flow controls the chromatographic band broadening and thus the efficiency. Keeping the velocity or flow under control is a prerequisite for good GC practice. Besides this, the carrier gas is often the transport medium of solutes from injection devices such as split, splitless, programmed-temperature vaporizing inlets, purge and trap, thermal desorption units, and pyrolysis units. Often the flows required in the precolumn devices do not correspond to the optimal carrier gas flows, and compromises must be accepted (i.e., split in purge and trap, long residence times of solutes in a splitless liner). On the other hand, optimization of total GC performance also requires precise control of the flows of the detector gases responsible for detector sensitivity and baseline stability. The quality of data generated by the hyphenated spectroscopic methods (see Section 6.4) also depends strongly on the stability of the flows of reagent gases (e.g., in chemical ionization mass spectroscopy and atomic emission detection). In the past, very little attention was paid to the mode of operation of the different gases in a GC system, and all gas supplies were more or less operated in the constant-pressure mode. In the constant-flow mode, all parts of a GC system run under optimum conditions. EPC is likely to become a standard mode of operation in the near future. The CGC applications with EPC published until now all show improved chromatographic performance going from reduced discrimination and large volume introduction in splitless injection by a pressure pulse shortening the residence time in the inlet, to increased sensitivity by EPC control of detector gases independently or in concert with the carrier gas flow. These are just a few examples, and more features are detailed in [31]. Yet more is to be expected because the possibilities of EPC have not yet been fully evaluated and exploited. A typical example can be found in multidimensional CGC (see Section 8.1). One of the reasons for slow acceptance of this extremely powerful technique is the necessary adjustment of the midpoint pressures, which can require several days of fine tuning with conventional pressure control devices. Flow control at the midpoint should solve this shortcoming.

Nevertheless, EPC introduces a new era in good laboratory practice (GLP) and good automated laboratory practice (GALP). Analytical methods can now be stored completely, providing exactly the same flow characteristics of all gases and tremendously enhancing repeatability and accuracy.

5. Sample Inlet Systems

5.1. General Considerations

"If the column is described as the heart of chromatography, then sample introduction may, with some justification, be referred to as the Achilles heel." This statement [32] clearly illustrates that sample introduction is of primary importance in GC and especially capillary GC. The performance of the sample introduction system is crucial for overall chromatographic performance of the complete system. Much progress has been made in recent years, and the understanding of injection phenomena has increased tremendously. Because of the diversity of samples that can be analyzed with modern GC (wide range of component concentrations; from highly volatile to less volatile components; different thermal stabilities; etc.), several injection methods have been developed. A final universal answer to sample introduction, however, has not been provided yet, and whether such an answer can be provided at all is questionable: "There does not now and probably never will exist injector hardware or methodology that is suitable for all samples, under all conditions" [33]. "There is still no such thing as a universal injection system and there probably never will be" [34]. This does not mean that precise and accurate results cannot be generated with the different injection systems that have been optimized over the years. The possibilities and the limitations of available systems, however, must be known exactly. In combination with a sufficient knowledge of the composition of the sample to be analyzed, proper injector choice will guarantee good results.

A qualitative and quantitative analysis means that the chromatographically determined composition of the sample corresponds to its real composition. Possible difficulties encountered in obtaining such data have to be attributed to the sampling technique, various column effects, or a combination of both. A sample introduction system can be "discriminative", meaning that certain components are not introduced quantitatively into the column. However, the column itself can also act to discriminate (reversible and/or irreversible adsorption), the degree of which can, moreover, be injection dependent. The characteristics of the column in use, therefore, must be well established before any conclusion can be drawn with regard to the sampling system.

The basic prerequisite of a sampling system for gas chromatography is that the sample must be introduced into the column as a narrow band, whose composition is identical to the original composition of the sample. The width of the inlet band must be such that its variance does not contribute significantly, or contributes only slightly, to the chromatographic broadening process. The total measured variance σ_{total} is the sum of all contributions to the peak variance:

$$\sigma^2_{\mathrm{total}} = \sigma^2_{\mathrm{column}} + \sigma^2_{\mathrm{injection}} + \sigma^2_{\mathrm{extra}} \tag{5.1}$$

With modern GC instrumentation, the extra peak variance contribution (electrometer, data system, recording) approaches zero. Because of the occurrence of the squared variances in this equation, the contribution of the initial bandwidth to the total peak bandwidth is less dramatic than might be expected, at least for GC on packed columns. For capillary columns, injection bandwidths can be too large compared to the chromatographic band broadening, and narrow initial bands have to be obtained by:

1) Introducing very small sample volumes (1 – 50 nL) into the column. A sample stream splitting device (splitter) accomplishes this task. The sample vapor, formed on injection at elevated temperature, is split into two sample streams of different flow rates.
2) The entire sample is introduced into the column, but the broad initial band is immediately focused into a narrow initial band by sharpening mechanisms such as solvent, thermal, or stationary-phase focusing. In practice, these effects are used with great advantage to focus the solute bandwidths produced by splitless and on-column injection.

In subsequent sections, the most important sample introduction systems will be discussed

in some detail with emphasis on capillary inlets. Because of the much larger σ_{column} in packed columns, injection is less stringent. Inlets are also divided into universal and selective inlets. The aim of universal injection is introduction of the complete sample into the column, whereas selective injection means that only a well-defined fraction enters the column.

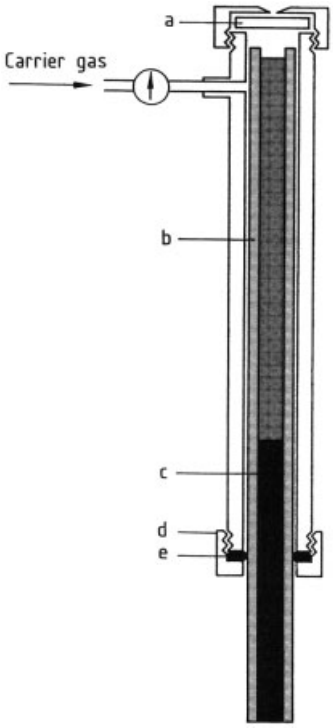

Figure 17. Schematic of a packed column injector
a) Septum; b) Glass wool; c) Packing; d) Nut; e) Ferrule

5.2. Universal Inlets

5.2.1. Packed Column Inlets

Packed column inlets are simple to design and use [35]. Injection is performed directly in the column "on-column" (Fig. 17) or in an in-sert placed before the head of the column. Different inserts are used, depending on the type of packed column (1/8″ = 0.32 cm or 1/4″ = 0.63 cm). Polar and thermolabile compounds are most effectively analyzed by on-column injection on a glass column. The injector zone is heated, and upon manual or automated injection, usually with a microliter syringe, hot mobile-phase gas carries the vaporized sample down to the packing material. Packed column inlets are vaporizing so they can cause sample degradation and needle discrimination. This is discussed in detail for split injection. Sampling valves are often used to sample gases or volatile liquids in constant-flowing streams. The valve is substituted for the syringe and can be used both with packed and with capillary columns. The construction of GC valves is similar to HPLC valves.

5.2.2. Capillary Column Inlets

In the context of this contribution, it is impossible to treat every aspect of sample introduction in capillary GC in depth. Full details on the different sampling methods can be found in [36], [37]. A survey of sampling systems is also given in [34], [38].

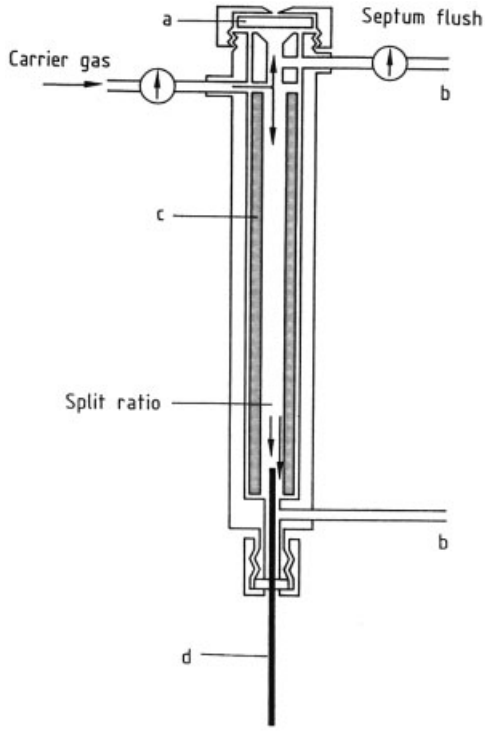

Figure 18. Schematic of a split injector
a) Septum; b) Needle valve; c) Vaporization chamber; d) Capillary column

5.2.2.1. Split Injection

Split sampling was the first sample introduction system developed for capillary GC [39]. The conventional split injector is a flash vaporization device. The liquid plug, introduced with a syringe, is volatilized immediately, and a small fraction enters the column, while the major portion is vented to waste. This technique guarantees narrow inlet bands. A schematic diagram of a split injector is shown in Figure 18. Preheated carrier gas enters the injector and the flow is divided in two streams. One stream of carrier gas flows upward and purges the septum (a). The septum purge flow is controlled by a needle valve (b). Septum purge flow rates are usually between 3 and 5 mL/min. A high flow of carrier gas enters the vaporization chamber (c), which is a glass or quartz liner, where the vaporized sample is mixed with carrier gas. The mixed stream is split at the column inlet, and only a small fraction enters the column (d). A needle valve controls the split ratio. Split ratios (i.e., column flow: inlet flow) range typically from 1 : 50 to 1 : 500 for conventional capillary columns (0.22 – 0.32 mm *i.d.*). Lower split ratios can be used in combination with focusing effects. For columns with a high sample capacity such as widebore or thick-film columns, low split ratios (1 : 5 – 1 : 50) are common. In high-speed capillary gas chromatography, applying 0.1 to 0.05 mm *i.d.* columns, split ratios can exceed 1 : 1000. In split injection, initial bandwidths are only milliseconds. Because split injection is a flash vaporization technique, sample discrimination is difficult to avoid. This is especially the case if the sample contains components in different concentrations and with different volatilities and polarities. Sample discrimination in split injection is caused by inlet-related parameters as well as by operational parameters such as syringe handling. *Inlet-related discrimination* is often referred to as nonlinearity of the splitter device. Linearity in this respect means that the split ratio at the point of splitting is equal to the preset split ratio and equal for all the components in the sample. Linear splitting against varying sample components, be it in concentration, volatility, or polarity, is impossible to achieve, even when a sample is introduced in a nondiscriminative manner into the vaporization chamber. Different mechanisms can cause nonlinear splitting: different diffusion speed of the sample components, incomplete evaporation, and fluctuating split ratio. The different mechanisms and their respective contribution are discussed in [36]. Nonlinearity can be minimized by complete vaporization of the sample, followed by homogeneous mixing with the carrier gas before the sample enters the column. As evident as this may seem, the sample introduced by split injection often arrives at the point of splitting as a mixture of vapor and a nonuniform mist of droplets. Two approaches can be used to minimize this phenomenon: (1) increased injection temperature and (2) optimization of the inlet configuration and glass liners. Different glass liners have been proposed (i.e., empty tube, short glass-wool plug in the splitting region, long and tight glass-wool plug, packing with chromatographic support or glass beads, deformation of cross section, Jennings tube, etc.) with the aim of enhancing efficient heat transfer to the injected sample and thoroughly mixing the vaporized sample. Still, whereas such modifications demonstrate an improvement in linearity for some applications, the same setup may give bad results for others. As a general rule, packed liners should not be used unless bad results have been obtained with an open liner. When discussing nonlinearity, it was taken for granted that sample introduction with a syringe into the vaporization chamber occurs without any alteration to the sample – in other words, no discrimination is caused by the syringe introduction. However, most of the discrimination problems encountered with vaporizing injectors are related to *syringe needle effects*. Upon introducing the syringe needle through the septum, volatiles immediately start to evaporate inside the needle itself, which is heated by the injector. Also after pushing down the plunger, solvent and volatile solutes are evaporated more readily than high-boiling solute material, which partly remains on the needle wall. On removing the needle from the injector body, the nonvolatile components are taken out of the vaporization chamber as well, resulting in a mass discrimination according to volatility. Different methods of syringe manipulation have been studied (e.g., filled needle, cold needle, hot needle, solvent flush, air flush, sandwich method) in combination with fast or slow injection. These different procedures are not discussed here, but rather the best method – "manual hot-needle, fast-sample

introduction" – is described. This method guarantees minimal syringe discrimination, although complete avoidance is impossible when dealing with solutes having a large volatility difference. In the hot-needle method, the sample is taken into the syringe barrel, for example up to the 5-µL mark when a 10-µL syringe is used, without leaving an air plug between sample and plunger. After insertion into the injection zone, the needle is allowed to heat up for 3 – 5 s. This time is sufficient for the needle to be heated to the injector temperature. The sample is then injected by rapidly pushing the plunger down (fast injection), after which the needle is withdrawn from the injector within 1 s. Syringe discrimination is related to the warming up of the needle in the vaporization chamber. Systems can be worked out in which syringe discrimination is reduced or avoided. The best technique is very fast injection. A prerequisite here is *automated cold-needle* sample introduction since all steps of the injection sequence are identical for each injection. The objective of very fast injection is to get the needle into the injection port, inject the sample, and withdraw the needle so quickly that it has no time to warm up, in this way avoiding selective sample evaporation. Fast injection, moreover, implies that the delivered volume of sample equals the preset volume. The effect of the needle *dwell time* in the injection port on sample discrimination has been studied [40]. Dwell time is defined as the interval between the needle tip piercing the septum on the way in and reaching the same point on the way out. Figure 19 shows a plot of $C_X : C_{20}$ area ratios ($X = 10 – 40$) as a function of C-number for different dwell times, with hexane as the solvent. Although these data have been obtained for direct injection, they are also valid for split injection. A dwell time of 500 ms or less shows no noticeable fractionation. Modern autosamplers inject in 100 ms. The previous discussion indicates that obtaining quantitative data by using a split injector can be problematical, but not impossible. For samples in which the solutes do not differ too much in volatility, precision and accuracy can be acceptable if the injection technique has been optimized. Some summarizing guidelines are given. In quantitative analysis, the standard addition or the internal standard method is preferred for manual injections. External standardization in which absolute peak areas are compared can be applied only with fast automated injection. Reproducibility will be enhanced by not varying the injected volume, which typically should be 0.5 – 2 µL. Injector temperature should be adapted to the problem at hand (i.e., close to the boiling point of the last eluting compound). Excessively high injector temperatures should be avoided. The hot-needle, fast-injection method is preferred for manual injection. The use of highly volatile solvents should be avoided whenever possible. If an open liner does not do the job, lose packing with deactivated glass wool or glass beads can provide a solution. However, adsorption and decomposition can occur. One of the main problems associated with split injection is syringe handling. The application of an autosampling system can overcome this difficulty, enhancing both precision and repeatability. For several applications, however, split injection will not give high quantitative accuracy.

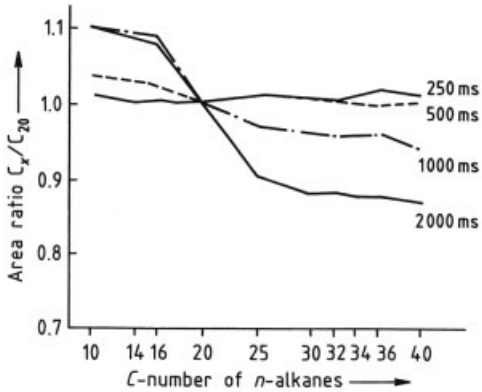

Figure 19. Effect of needle dwell time on sample discrimination

5.2.2.2. Splitless Injection

In splitless injection a conventional split injector is operated in a nonsplitting mode by closing the split valve during injection. The sample is flash vaporized in the vaporizing chamber from which the sample vapors are carried to the column by the flow of the mobile phase. Since this transfer takes several tens of seconds, broad initial bandwidths would be anticipated. However, through optimal use of focusing effects, such as solvent, thermal, and stationary-phase focusing, initial peak broadening can be suppressed. The

main feature of splitless injection lies in the fact that the total injected sample is introduced in the column, thus resulting in much higher sensitivity compared to split injection. For a long time, splitless injection was the only sample introduction technique in capillary GC for trace analysis. The pneumatic configuration is similar to the classical split injector (see Fig. 18). The septum is purged continuously (flow rate 2 mL/min) to maintain the system free of contamination, while a flow of 20 – 50 mL/min passes the split outlet. Prior to injection, a solenoid valve is activated, so that the splitline is closed off while the septum purge is maintained. After waiting a sufficient time for the solvent – solute vapor to be transferred onto the column (i.e., 30 – 80 s), the solenoid valve is deactivated. Residual vapors in the vaporizing chamber are vented to waste via the split line. For this reason, in splitless injection the split line is often referred to as the purge line. The time interval between the point of injection and the activation of the split (purge) line is a function of the characteristics of solvent and solutes, the volume of the vaporizing chamber, the sample size, the injection speed, and the carrier gas velocity. The sample transfer from the vaporizing chamber into the column is a slow process. Solvent vapors especially tend to remain in the inlet for a long time. Purging the insert after injection removes the last traces of vapor from the vaporizing chamber. An important aspect in the design of a splitless injector is the dimension of the vaporizing chamber. Long and narrow inserts are preferred to obtain minimal sample dilution. Internal volumes vary between 0.5 and 1 mL. The column is installed 0.5 cm in the insert, and syringes with long needles are used, creating a distance of 1 – 1.5 cm between the needle tip and the column inlet. Overfilling of the vaporizing chamber is suppressed, and fast sample transfer is achieved. If sample volumes larger than 2 µL are injected, a widebore liner should be used. Because of the relatively long residence time of the solutes in the vaporizing chamber, lower injection temperatures can be used compared to split injection. This can be helpful in minimizing sample degradation. For the same reason, the use of packed inserts should be avoided in splitless injection. An illustration of the effect of glass or fused silica wool in the liner is presented in Figure 20. The organochloropesticide endrin (peak 1) tends to decompose into endrin aldehyde (peak 2) and endrin ketone (peak 3) when active sites are present in the liner. The first chromatogram shows that with an empty liner, 92 % of the endrin is detected, compared to cool on-column injection, which provides 100 % elution. Filling the liner with a 1- cm plug of glass wool or fused silica wool results in the elution of only 83 % and 48 % endrin, respectively. Taking into account the fact that endrin is not the most thermolabile compound, what will happen if polar compounds (i.e., carbamates, barbiturates, etc.) are injected with filled liners can easily be imagined. Electronic pressure control was introduced to reduce the residence time of the solutes in the liner. With EPC, the carrier gas pressure or flow can be programmed. During injection a very high flow rate is applied, which speeds up the transfer liner – column. After injection, the flow is reset to the optimal value for the column in use [41]. Syringe discrimination, which is one of the largest sources of error when applying split injection, also occurs for splitless injection because both are vaporizing inlets. The performance of a splitless injector for a particular application depends on the optimization of experimental variables, the most important of which are sample size, injection speed, purge time, injection temperature, initial column temperature, carrier gas selection, and flow rate. General guidelines cannot be advanced, and some of the variables can be only tuned by trial and error. Whatever the application, however, refocusing of the solutes in the inlet section of the capillary column is necessary. In splitless injection, the initial bandwidths are broadened by two mechanisms: band broadening in time and band broadening in space. *Band broadening in time* is caused by the slow transfer from the vaporizing chamber to the column inlet section, which takes several tens of seconds. Upon their arrival in the column, the solutes spread over a certain column length, mainly by flooding of the sample liquid, thus causing *band broadening in space*. The fundamental difference between band broadening in time and band broadening in space is that, in the first case, solutes are spread equally with respect to gas chromatographic retention time, whereas in the second case, solutes are spread equally with respect to column length. Both phenomena cause distorted elution profiles if the solutes are not refocused before starting the chromatographic distribution process.

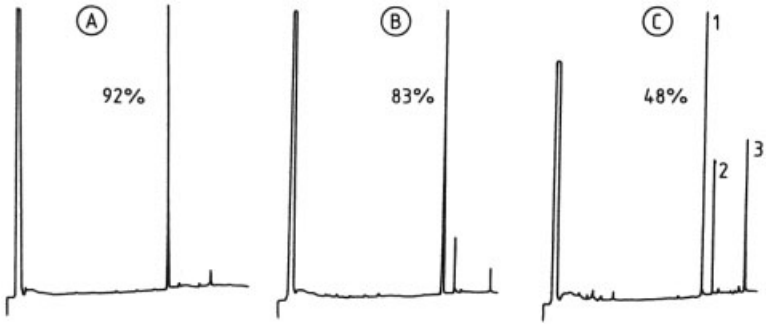

Figure 20. Decomposition of endrin on different splitless liners
A) Empty liner; B) 1-cm plug of glass wool; C) 1-cm plug of fused silica wool

Solvent Focusing, Thermal Focusing, and Stationary-Phase Focusing. Splitless injection most often is combined with column temperature programming, starting with an oven temperature 25 – 30 °C below the boiling point of the solvent. Upon condensation of solvent and solutes, the droplet formed becomes too thick to be stable. The carrier gas pushes the plug further into the column, creating a "flooded zone" (Fig. 21). The solutes now are spread over the full length of the flooded zone, thus creating a solute bandwidth that equals the length of the flooded zone. For 1-µL injections, the length of the flooded zone roughly is 20 cm, provided the stationary phase is perfectly wettable by the solvent [i.e., isooctane solutions on apolar methylsilicone phases, ethyl acetate solutions on poly(ethylene glycol) phases]. On conventional capillary columns, 25 – 30 m in length, *i.d.* 0.32 mm, band broadening will hardly be observed for sample volumes of 1 µL. Only for sample sizes larger than 2 – 3 µL will peak deformation be noted. On the other hand, for polar solvents on apolar stationary phases, the length of the flooded zone can be as large as 1 m, resulting in distorted peaks. Stationary phase focusing then becomes necessary. After creation of the flooded zone, the temperature is increased and the solvent starts to evaporate from the rear of the flooded zone (Fig. 22). Highly volatile solutes evaporate as well and the last traces of solvent act as a barrier since the residual solvent behaves as a thick stationary-phase film for those solutes. High volatiles are thus focused by the *solvent effect*. The low-volatility solutes are cold trapped since the column temperature is still too low to evaporate them. They are *thermally focused* over the length of the flooded zone. When the temperature is high enough to evaporate the low-volatility solutes, they start the chromatographic process at different points. If the flooded zone is homogeneous and short, peak distortion is hardly observed, but when the flooded zone is inhomogeneous and long, which is the case when polar solvents are injected on apolar capillary columns, broad distorted peaks are detected. The only reason for this is that the chromatographic process starts with the solutes distributed over the length of the flooded zone. This can be avoided by removing the stationary phase over the length of the flooded zone or, in other words, by creating a retention gap. Band broadening in space can thus be suppressed by *stationary-phase focusing* via a retention gap (Fig. 23). A retention gap is a definite length of the separation system that is uncoated. All solutes that are spread over the flooded zone are carried with no retention onto the stationary phase where they are retained. In practice, a retention gap is a separate piece of deactivated fused silica, connected to the analytical column through a coupling device such as a press fit connector. The length of the retention gap is a function of the length of the flooded zone, and depends on the sample volume and the nature of the solvent in use. Typical lengths are 0.5 – 1 m for injections of 1 to 2 µL. Longer lengths, however, do not harm the separation, and pieces of 3 m are standard and commercially available. Band broadening in space and the use of a retention gap are also required for cool on-column injection (see Section 5.2.2.3). In recent years, splitless injection has been overshadowed by cool on-column injection. Without any doubt,

the most accurate and precise data are now provided with on-column injection. Notwithstanding this, splitless injection is still used for many routine determinations (i.e., in environmental analysis, pesticide monitoring, drug screening, etc.). In these fields, sample preparation is of primary concern, and cleaning up a sample to such an extent that it is accessible to on- column injection is not always possible or economically justifiable. Traces of nonvolatile or high-boiling components often remain in the sample. Splitless injection is an easy solution to such problems. The "dirty" components remain mainly in the vaporizing inlet, which is easily accessible for cleaning. An illustration of the determination of polychlorinated biphenyls (PCBs) in waste oil via splitless injection is shown in Figure 24. 5 µL of an isooctane solution of the PCBs extracted from waste oil, was injected in the splitless mode on an apolar column equipped with a retention gap of 5 m. A pressure pulse was applied to minimize the residence time of the solutes in the liner [42]. Without a retention gap the flooded zone would have been too large to guarantee good peak shapes. When properly performed, reproducibility in splitless injection can be as high as 1 – 2 % relative standard deviation (RSD). Standard addition or internal standard quantitation is the method of choice for manual injection. The external standard method can be adapted for automated splitless injection. Many capillary GC applications need the best attainable precision and accuracy. Cool on- column injection (see Section 5.2.2.4) then is the method of choice.

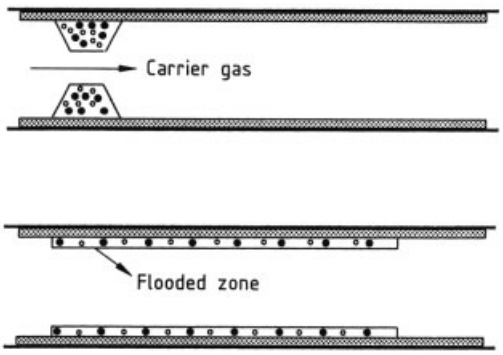

Figure 21. Creation of the flooded zone
o = High-volatile solute; ● = Low-volatile solute

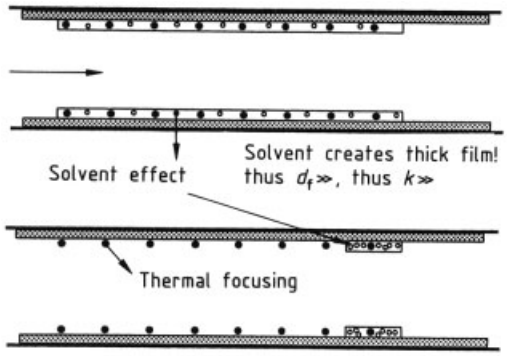

Figure 22. The solvent and thermal focusing effect
o = High-volatile solute; ● = Low-volatile solute

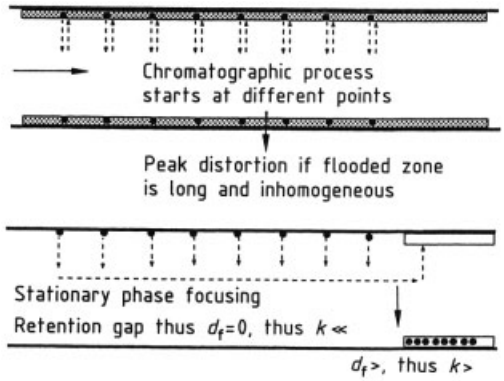

Figure 23. Stationary-phase focusing and the principle of a retention gap
● = Low-volatile solute

5.2.2.3. Direct Injection

The nomenclature of different injection techiques has been created as new developments occur. This often causes confusion with regard to the interpretation of some terms, which do not always stand for what they really mean. Direct injection is an example of such an expression, often confused or even identified with on-column injection. The understanding of the terminology given in this article, based on the chronological order in which techniques were developed, is as follows: *Direct injection* is a flash vaporizing injection method. The inlet system is heated separately and independently from the column oven, and evaporation occurs in the inlet. This inlet can be a glass liner (out-column evaporation) or a part of the column

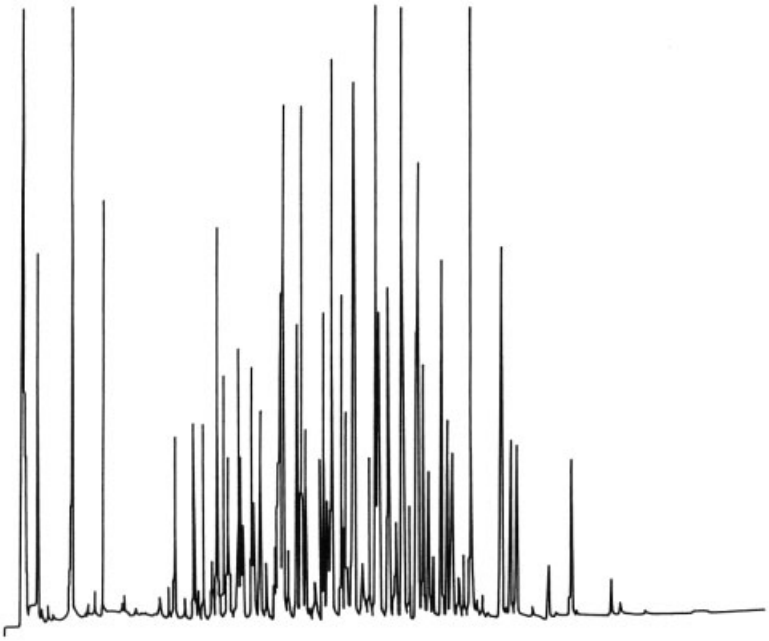

Figure 24. Optimized splitless injection of a PCB sample
Column: 25 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f MeSi; retention gap: 5 m $L \times 0.25$ mm *i.d.*; temperature: 150-280 °C at 5 °C/min; splitless injection with pressure pulse; carrier gas hydrogen; FID

(in-column evaporation). Different direct injection devices are shown in Figure 25. *On-column injection* is a "cold" injection technique. The sample is injected as a liquid, directly on the column. During injection, the injection zone is cooled to avoid needle discrimination. Injection peak broadening is caused only by band broadening in space. The solutes are focused in the inlet section of the column, from where evaporation gradually starts as the oven temperature is raised. Whereas on-column injection can be applied for all conventional capillary columns (*i.d.* 0.25, 0.32, 0.53 mm), direct injection is restricted to widebore or megabore columns.

Direct injection into widebore columns was described in 1959 [43]. At that time, column efficiency was all important. Through the introduction of widebore fused-silica capillary columns with *i.d.*'s of 0.53 mm, the interest in direct sampling has been renewed. Injection on a 0.53-mm *i.d.* FSOT column, operated in the high-resolution (flow rates 1 – 3 mL/min) or the low-resolution (flow rates 8 – 12 mL/min) mode, as a packed column alternative, is said to be as easy as installing a simple glass liner in the packed column inlet and connecting the column to it. This view is slightly oversimplified since the direct injection technique is based on the evaporation of the sample at elevated injection temperature, all previous statements about syringe discrimination and band broadening effects remain valid. Direct injection must definitely be optimized according to the problem at hand. Some experience-based guidelines on direct injection into widebore columns are advanced. However, when widebore capillary columns are used in the high-resolution mode, they should be connected to conventional capillary injectors (split, splitless, on- column, programmed temperature vaporization). For direct injection on widebore columns operated at high flow rates, the sample size should be kept as small as possible, and the sample should preferably be injected with an autosampler. For manual injections, the fast hot-needle injection technique is preferred to minimize syringe discrimination. To avoid sample losses due to overfilling the insert with sample vapors, the point of injection should be close to the column entrance. Long and narrow inserts are preferred to obtain minimum sample dilu-

tion (band broadening in time). Band broadening should be compensated by refocusing the solutes in the inlet section of the column. The principle of the solvent effect and of thermal focusing have been discussed in Section 5.2.2.2. Injections of microliter volumes at oven temperatures (far) above the boiling point of the solvent should be avoided, unless the sample is known to contain heavy compounds only, which will be refocused by cold trapping at the chosen column temperature. Figure 26 shows the analysis of the oxygenated fraction of hop oil dissolved in dichloromethane on a narrowbore column with split injection (Fig. 26 A) and on a widebore column with direct injection (Fig. 26 B) [44]. Both columns (25 m×0.25 mm and 50 m× 0.50 mm) offer the same efficiency. Peak broadening did not occur on the widebore column with direct injection, for which only 0.2 µL was injected.

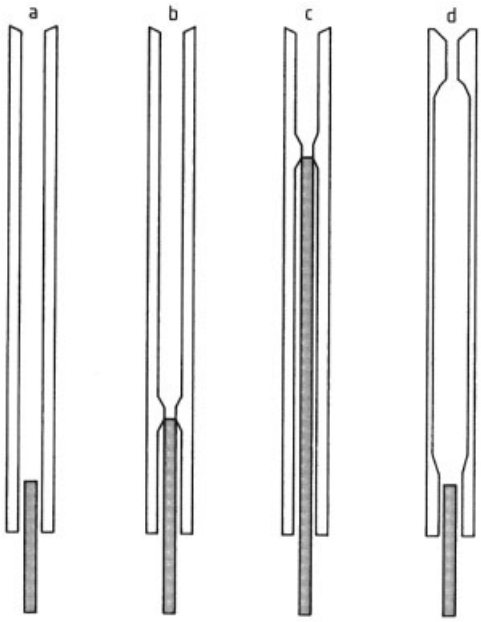

Figure 25. Different injection devices for direct injection a) Open liner; b) Open liner with conical contraction; c) Direct on-column liner; d) Open liner with expansion volume

5.2.2.4. Cool On-Column Injection

Syringe on-column injection with small-diameter capillary columns was first described in 1978 [45], [46]. Special on-column devices (a micro and a macro version) were introduced in 1977 by Schomburg et al. [47], but the technical requirements were stringent and the devices lacked practical flexibility. Since this pioneering work, all instrument manufacturers have introduced cool on-column injectors. As an example, the cool on-column injector developed by Hewlett–Packard is shown in Figure 27. It can be used for manual and automated injection. The injector has a low thermal mass, which facilitates cooling. A key part of the injector for manual operation is the duck bill valve (b). The duck bill valve, made out of a soft elastomer, is a passive element in that it has no moving parts. It consists simply of two surfaces pressed together by the column inlet pressure. During injection, the syringe needle slips between the two surfaces to maintain the seal. For automated cool on-column injection, the system is modified with a disk septum (h). Today, cool on-column injection systems can be operated in the constant-pressure, constant-flow, or pressure-programmed mode. These features allow reduction of analysis time and elution of high molar mass compounds. The advent of cool on-column injection has extended the range of applicability of capillary GC to many classes of compounds that heretofore were difficult if not impossible to analyze. The technique of introducing the sample directly on the column without prior vaporization offers many advantages, such as elimination of sample discrimination and sample alteration, high analytical precision, and data operator independence. As with all other sampling techniques, the operational parameters strongly affect the chromatographic data. Parameters such as initial column temperature, solvent nature, injection rate, injected volume, and boiling point range of the sample components are interrelated. Once again, all aspects cannot be treated in detail in this article. *On-column Injection in Capillary GC* by K. Grob contains 590 pages [37] ! The characteristics of the technique are highlighted here, and some important aspects for daily practical use are discussed. For sample sizes of the order of 0.5–2 µL, injection should be performed as quickly as possible, with the column oven temperature below or equal to the boiling point of the solvent. The oven temperature in on-column injection may be higher than in splitless injection. In splitless injection, the solvent has to recondense; whereas in on-

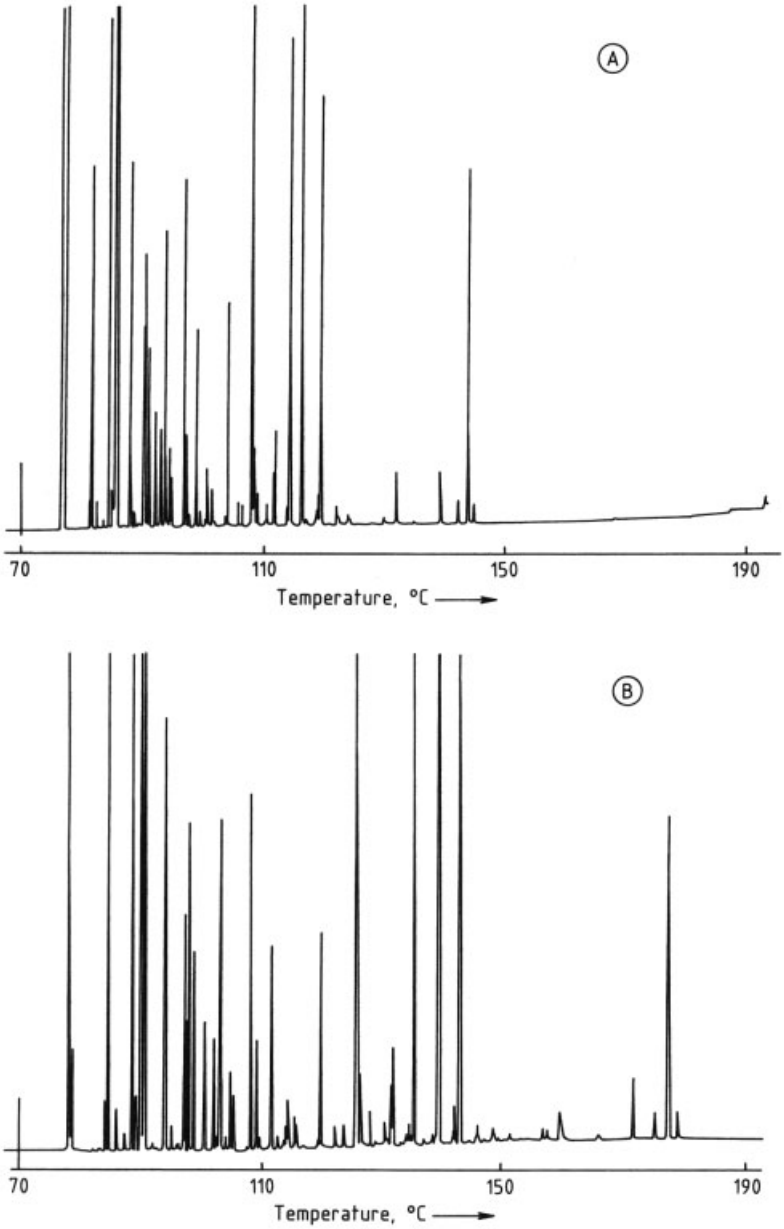

Figure 26. Analysis of the oxygenated fraction of hop essential oil on a conventional and a widebore capillary column
A) Column: 25 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f high-M_r poly(ethylene glycol); temperature: 70 – 190 °C at 2 °C/min; split injection 1/20; carrier gas hydrogen; FID; B) Column: 50 m $L \times 0.5$ mm *i.d.*; 0.25 µm d_f high-M_r poly(ethylene glycol); temperature: 70 – 190 °C at 2 °C/min; direct injection; carrier gas hydrogen; FID

column injection, the liquid is introduced directly into the column. As in splitless injection, the length of the flooded zone must be reduced because solutes are distributed over its entire distance. The width of the initial band equals the length of the flooded zone, i.e., band broadening in space (band broadening in time does not occur in on- column injection). The length of the flooded zone is a function of the column diameter, the volume injected, the actual inlet temperature, the stationary-phase thickness, and, most important, the affinity of the solvent for the stationary phase. For perfect wettability (i.e., apolar solvents on apolar silicone phases), the length of the flooded zone is about 20 cm per microliter injected. This length easily increases by a factor of five or ten if the wettability is poor, such as for methanol injections on apolar silicone phases. If the solutes are not refocused, band broadening in space makes qualitative and quantitative analysis impossible if large sample volumes are injected or if the solvent does not properly wet the stationary phase. The latter phenomenon is illustrated in Figure 28 with the analysis of a fraction collected from a step-elution HPLC analysis of the essential oil of *Valeriana celtica* L., with hexane and methanol. The fraction of the polar compounds was contained in 1 mL of hexane and 6 mL of methanol in a two-phase system. Both phases were injected on-column (1 µL) onto a 20 m×0.3 mm *i.d.* methylsilicone column with a film thickness of 0.3 µm. Figure 28 A shows the chromatogram of the hexane layer and Figure 28 B that of the methanol layer. Methanol causes peak splitting, whereas hexane does not [48] . The poor wettability (polar – apolar) in the case of methanol results in a long, nonhomogeneous flooded zone. This artifact can be bypassed via a retention gap (see Section 5.2.2.2). On-column injection provides the most accurate and precise results because syringe discrimination, which is one of the main sources of error in the quantitative analysis of samples covering a wide range of molar masses, is completely avoided. Moreover, inlet-related discrimination does not occur either, because the liquid is introduced directly onto the column. The automated on- column analysis of free fatty acids from acetic to decanoic acid is shown in Figure 29. Relative standard deviations on absolute peak areas are < 1 % and on relative peak areas < 0.4 %, for 20 injections ($n = 20$). The analysis

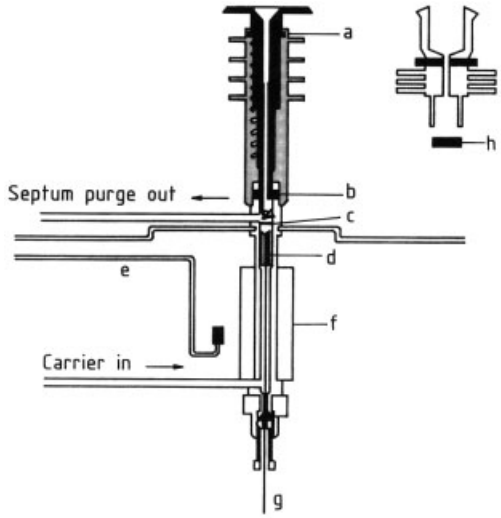

Figure 27. Schematic of a cool on-column injector
a) Cool tower, needle guide; b) Duck bill valve (isolation valve); c) Spring; d) Insert; e) Cryogenic cooling (optional); f) Heater block; g) Column; h) Septum

of polymer additives covering a broad range of functionalities and molecular masses is shown in Figure 30 [49]. To elute Irganox 1010 (M_r 1176) in a reasonable time, cool on-column injection was performed in the constant-flow mode. Both the external standard and the internal standard method can be applied for quantitation. Current literature data contain overwhelming evidence for the superior features of cool on- column injection in quantitative work. A second important feature of cold on- column injection is the elimination of sample alterations. Thermally labile components are not exposed to thermal stress: they begin the chromatographic process at relatively low temperature. Decomposition and rearrangement reactions are nearly completely eliminated. This feature allows analyses that heretofore were impossible with gas chromatography. Figure 31 shows the analysis of some retinoids [50]. Quantitative analysis in plasma samples was possible only with cool on- colum injection. The only disadvantage of cool on-column injection is that since the sample is introduced directly onto the column, relatively "clean" samples have to be prepared. Nonvolatile and less volatile material collects at the head of the column. The presence of this material causes a loss of separation efficiency or introduces adsorptive sites in the column inlet. Proper sample cleanup is thus very important. The use of a precolumn

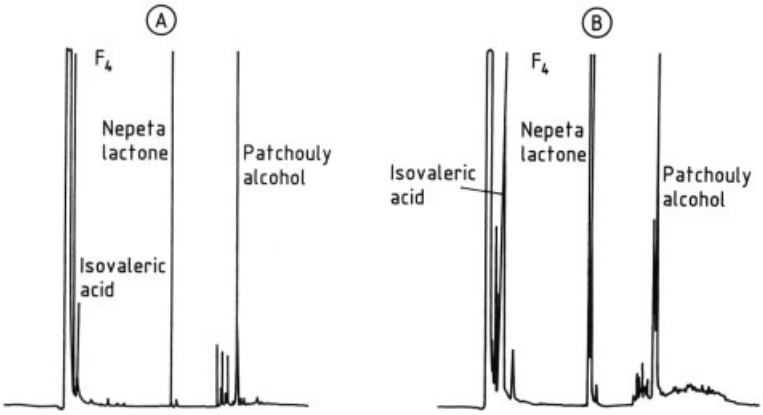

Figure 28. Illustration of peak splitting in cool on-column injection
A) Hexane as solvent; B) Methanol as solvent

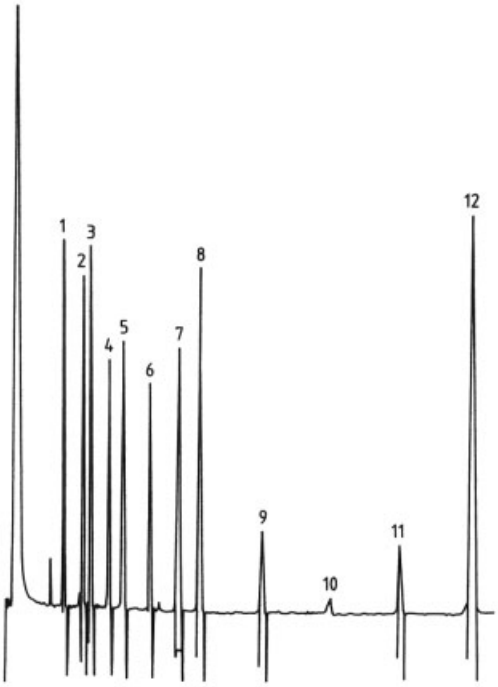

Figure 29. Analysis of free fatty acids on a widebore capillary column
Column: 10 m $L \times 0.53$ mm *i.d.*; 0.25 µm d_f FFAP; temperature: 60 – 100 °C balistically, to 140 °C at 2 °C/min; cool on-column injection; carrier gas hydrogen 4 mL/min; FIDPeaks: (1) acetic acid, (2) propionic acid, (3) isobutyric acid, (4) butyric acid, (5) isovaleric acid, (6) valeric acid, (7) isocaproic acid, (8) caproic acid, (9) heptanoic acid, (10) octanoic acid, (11) nonanoic acid, (12) decanoic acid

of 3 m, which at the same time acts as a retention gap, is therefore recommended. When the chromatographic performance decreases, 5 to 10 cm of the inlet section of the pre-column is cut off, restoring the original performance of the system.

5.2.2.5. Programmed-Temperature Injection

In 1979, Vogt and coworkers [51] described an injection method for the introduction of sample volumes up to 250 µL. The sample was introduced slowly into a cold glass insert packed with glass wool. The low-boiling solvent was evaporated continuously and vented through the split exit. The solutes, which remained in the inlet, were transferred to the capillary column by rapidly heating (30 °C/min) the inlet. During this transfer, the split line was closed (splitless injection). Based on this idea, the groups of Schomburg [52] and Poy [53] almost simultaneously developed programmed-temperature injection. Their aim, however, was not the injection of large sample volumes but rather the elimination of syringe needle discrimination, which at that time was the subject of interest. Different programmed-temperature vaporization (PTV) injection devices are now commercially available, offering a broad range of possibilities (i.e., hot or cold split injection, hot or cold splitless injection, cool on- column injection, direct injection, etc.). Through these possibilities, together with features such as the injection of large sample volumes, concentration by multiple injection, solvent venting, etc., PTV injection is claimed to be the most universal sample introduction system. It is, how-

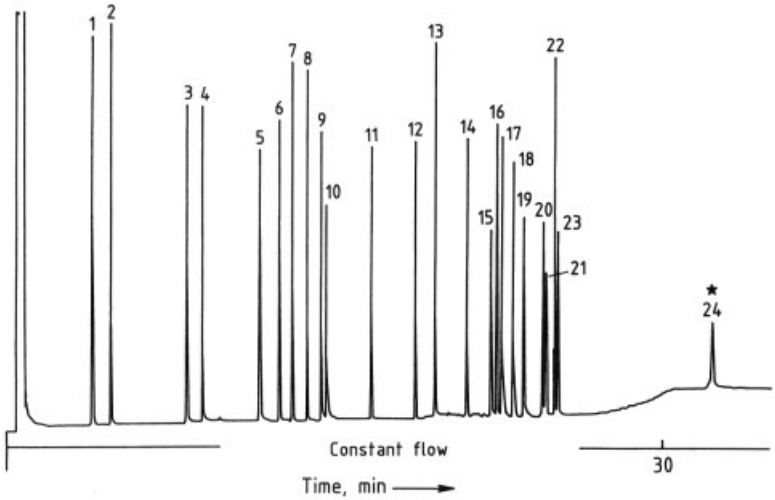

Figure 30. Analysis of polymer additives by high-temperature capillary GC
Column: 25 m $L \times 0.32$ mm *i.d.*; 0.17 µm d_f MeSi; temperature: 80 – 380 °C at 10 °C/min; cool on-column injection; carrier gas hydrogen 50 kPa constant flow; FID
Peaks: (1) butyl-4-methoxyphenol, (2) diethyl phthalate, (3) dibutyl phthalate, (4) Tinuvin P, (5) triphenyl phosphate, (6) dicyclohexyl phthalate, (7) dioctyl phthalate, (8) Tinuvin 327, (9) benzophenone UV, (10) erucamide, (11) Tinuvin 770, (12) Irgaphos 168, (13) Irganox 1076, (14) Tinuvin 144, (15) Irganox 245, (16) Irganox 259, (17) Irganox 1035, (18) Irganox 565, (19) crodamide, (20) Irganox 1098, (21) Irganox 3114, (22) Irganox 1330, (23) Irganox PS 802, (24) Irganox 1010

ever, not the best choice for all applications. The upper part of a PTV inlet resembles a classical split – splitless device, including carrier gas inlet and septum flush. The sample is injected into a glass liner with a low thermal mass while the injector body is cold. After withdrawal of the syringe needle, the vaporizing tube is heated rapidly to volatilize the solvent and the solutes. The heat can be provided electrically or by means of preheated compressed air. Depending on its construction, heating of the device can be performed ballistically or linearly at selected rates (for example 2 – 12 °C/s) and even to consecutive temperature levels. These facilities allow optimization of conditions for the analysis of thermally labile compounds, operating in the solvent vent mode when working with specific detection such as electron capture detection (ECD) or mass spectrometry, concentration through multiple injection, etc. Split or splitless injection can be achieved through regulation of the split valve. During or after the chromatographic run, the vaporizing chamber is cooled by air or carbon dioxide, in preparation for the next injection. In an interesting mode of operation (i.e., *solvent elimination injection*), the sample is introduced into the cold injector with the split valve open. Conditions are selected such that only the solvent evaporates while the solutes of interest remain in the liner. The applicability of this technique is, therefore, restricted to the analysis of relatively high-boiling compounds. In this way, large sample volumes can be introduced to enrich traces. The principle has been applied to detect dioxins and furans down to the 100-fg/µL level with capillary GC – low-resolution mass spectroscopy [54].

5.3. Selective Inlets

A number of selective inlets are available in GC. Best known are static and dynamic head space (HS) injection (selectivity based on vapor pressure), purge and trap (PT) samplers (selectivity based on solubility of the solutes in water), thermal desorption (TD) units (selectivity based on characteristics of the adsorbent), and pyrolysis units (selectivity based on the pyrolysis or cracking temperature). These inlets are discussed in Sections 5.3.1 – 5.3.4. More sophisticated systems, which can also be considered as sample introduction systems, such as solid-phase extrac-

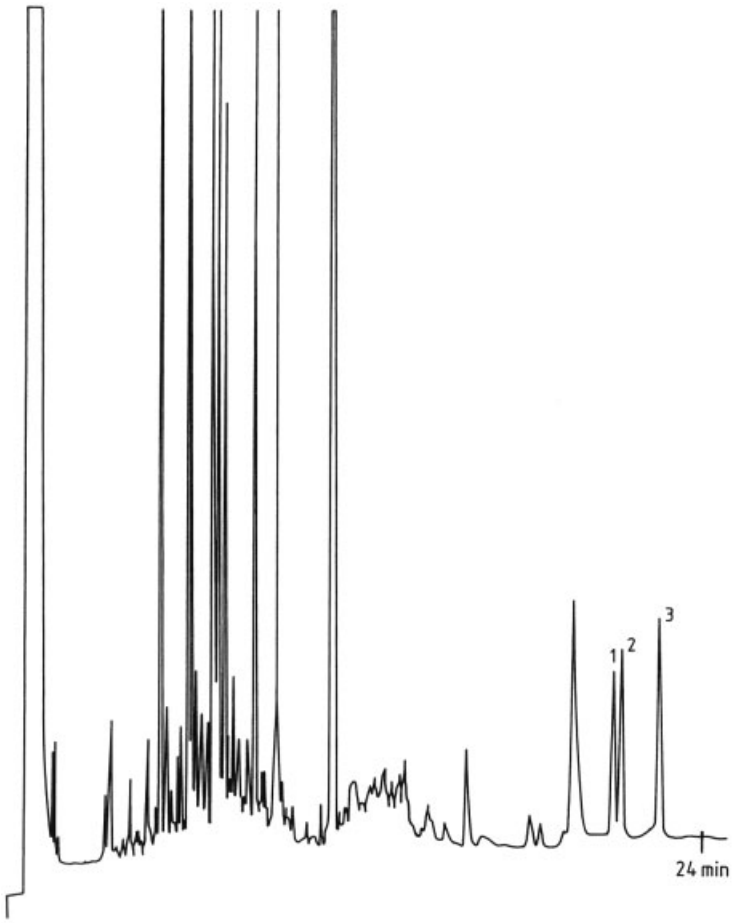

Figure 31. Illustration of the application of cool on-column injection to the analysis of thermolabile compounds
Column: 12.5 m $L \times 0.25$ mm *i.d.*; 0.25 µm d_f CPSi; temperature: 60 – 190 °C at 17 °C/min, to 220 °C at 6 °C/min and to 235 °C at 1.5 °C/min; cool on-column injection; carrier gas hydrogen; FIDPeaks: (1) *trans*-acitretin methyl ester, (2) *cis*-acitretin methyl ester, (3) *cis*-acitretin propyl ester

tion (SPE; selectivity based on the polarity of both adsorbent and desorbing liquid), supercritical fluid extraction (SFE; selectivity based on the solubility in the supercritical medium, the variables being density, temperature, pressure, and polarity), multidimensional capillary GC (MDCGC), the combinations of capillary GC with liquid chromatography (HPLC – CGC) or with supercritical fluid chromatography (SFC – CGC) allowing preseparation according to molar mass (size-exclusion chromatography, SEC), functionality (normal phase LC), hydrophobicity (reversed phase LC), or solubility (SFC) are treated in Chapter 8. In addition, *on-line sample preparation units* became available in 1993, providing functions such as dispensing (addition of reagents or internal standards), diluting (concentration adjustments and multipoint calibration standards), heating (derivatization, solubilization), evaporating (solvent exchange, concentration), solid-phase extraction, and filtering.

5.3.1. Static and Dynamic Head Space Analysis

Head space analysis is used to analyze volatiles in samples for which the matrix is of no interest (i.e., water, soil, polymers, etc.). Typical applications include monitoring of volatiles in soil and water (environmental); determination of

monomers in polymers, aromas in food and beverages, residual solvents in drugs (pharmaceuticals), etc. Different head space autosamplers are commercially available, and they are based on the principle of static or dynamic head space. The difference between these modes is depicted in Figure 32. In *static head space* (Fig. 32 A), the sample (soil, water, solid) is transferred in a head space vial that is sealed and placed in a thermostat to drive the desirable components into the head space for sampling. Via a gas-tight syringe (a) or a sample loop of a gas sampling valve, an aliquot of the vapor phase is introduced in the GC system, which can be equipped with a packed or a capillary column. Because of the much lower sample loadability of a capillary column, split injection is usually applied in this case. Static head space implies that the sample is taken from one phase equilibrium. To increase detectability, *dynamic head space analysis* has been developed. The phase equilibrium is continuously displaced by driving the head space out of the vial via an inert gas. The solutes are focused by cold trap (e) or an adsorbent such as Tenax and after enrichment the trap is heated and the solutes are introduced into the GC system. Whereas static head space is applied to analyze ppm amounts (1 part in 10^6), dynamic headspace allows determination of ppb amounds (1 part in 10^9) of volatiles. Sample pretreatment can often help to increase sensitivity and enhance reproducibility. The best-known methods are salting out of water samples with sodium sulfate, adding water to solid samples, and adjusting the pH to drive organic acids or bases into the head space.

5.3.2. Purge and Trap Systems

Purge and trap samplers have been developed for analysis of apolar and medium-polarity pollutants in water samples. The commercially available systems are all based on the same principle. Helium is purged through the sample placed in a sealed system, and the volatiles are swept continuously through an adsorbent trap where they are concentrated. After a selected time (10 – 20 min), purging is stopped, the carrier gas is directed through the trap via a six-way valve, and the trap is heated rapidly to desorb the solutes. The method has proven to be highly accurate for environmental monitoring of ppb amounts of volatile aromatics, trihalomethanes, etc.

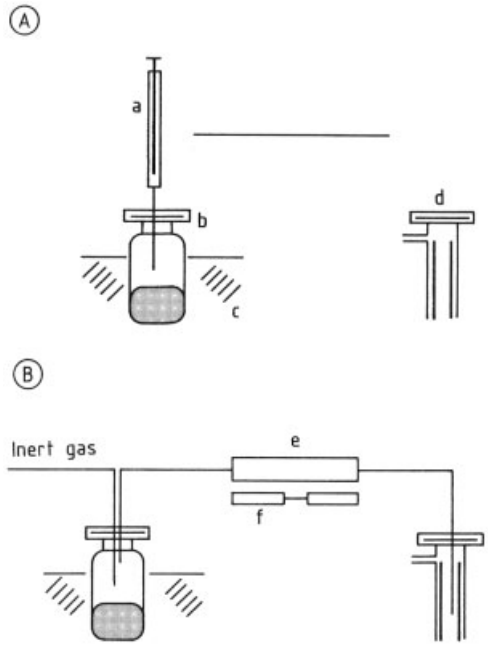

Figure 32. Principle of static (A) and dynamic head space (B)
a) Syringe;b) Head-space bottle; c) Thermostat;d) Injector; e) Cold trap; f) Solid-phase extraction

5.3.3. Thermal Desorption Units

Thermal desorption is the method of choice to monitor air pollution. A known amount of air is drawn through an adsorbent tube filled with activated carbon, Tenax, silicagel, or mixtures of these. Analytes of interest are trapped and concentrated on the adsorbent. The adsorbent tube is sealed and transported to the laboratory where it is installed in the thermal desorption unit. An example of a thermal desorption unit is shown in Figure 33. After the adsorbent tube is placed in the desorption module, the carrier gas is directed over the adsorbent, which is heated from room temperature to 250 – 300 °C. Since thermal desorption is a slow process, the solutes are focused by cold trapping on a fused silica trap. The cryogenic trap is then heated rapidly to transfer the sample to the column. The method works well for apolar and medium-polarity components in air. Highly polar solutes are very hard to desorb, and other sampling methods must be selected.

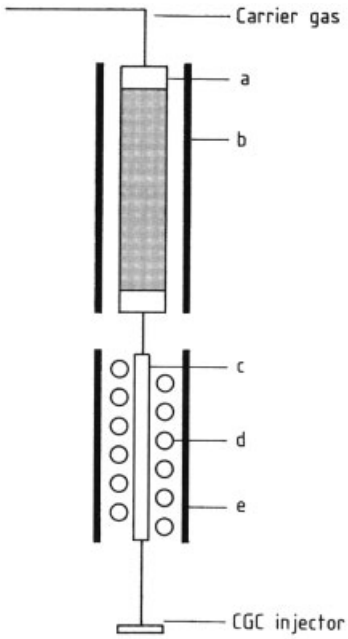

Figure 33. Schematic of a thermal desorption unit
a) Adsorbent tube; b) Heating device; c) FSOT trap; d) Cooling; e) Heating

5.3.4. Pyrolysis Gas Chromatography

Pyrolysis involves the thermal cleavage (cracking) of large molecules, which are not amenable to GC, into small volatile fragments. Pyrolysis GC is an excellent technique for identifying polymers and microorganisms because the pyrolysis profiles are very typical fingerprints. Most important in this respect is the pyrolysis temperature, typically between 400 and 1000 °C. The higher the temperature, the greater is the degree of fragmentation. Several types of analytical pyrolyzers are available. The most common types are the *platinum resistively heated* and the *Curie point pyrolyzers*. Figure 34 shows an example of a resistively heated pyrolyzer. The pyroprobe (c) consists of an injector cap (a), a quartz tube (d), and a platinum wire (f). The sample is placed in the quartz tube, and the pyroprobe is introduced in the injection port of the GC. The carrier gas enters the system, and after equilibration, the platinum wire is heated to a preselected temperature. The material decomposes and the volatiles are analyzed.

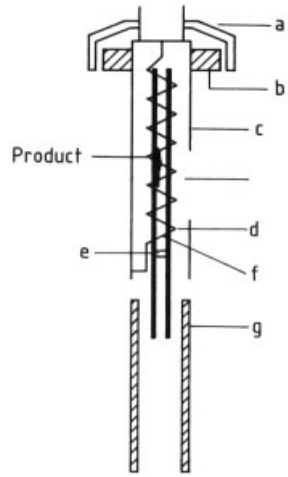

Figure 34. Principle of pyrolysis GC
a) Injector cap; b) Septum; c) Pyroprobe; d) Quartz tube; e) Glass wool; f) Platinum wire; g) Inlet liner

6. Detectors

Detectors in gas chromatography sense the presence of analytes in the carrier gas and convert this information in an electrical signal, which is recorded. Numerous detection devices have been developed in GC [55]; here, only the currently most common detectors are dealt with. An excellent book describing detectors for capillary chromatography has been published [56].

6.1. Classification

GC detectors are classified mainly on the basis of response or of detector selectivity. Detectors whose response is proportional to the concentration of a sample component in the mobile phase (g/mL) are called *concentration-sensitive detectors*, whereas detectors whose response is proportional to the amount of sample component reaching the detector in unit time (g/s) are called *mass-flow detectors*. *Universal detectors* respond to every component in the mobile phase; *selective detectors* respond only to a related group of substances; and *specific detectors* respond to a single sample component or to a limited number of components with similar chemical characteristics. The differentiation of

selective versus specific is confusing, and therefore in this article, selective is used for both to distinguish them from universal. Another classification is destructive versus nondestructive detectors. Any detector can be coupled in series with a nondestructive detector. The most important characteristics for practical work are sensitivity, dynamic range, linear range, detector response factors, and selectivity.

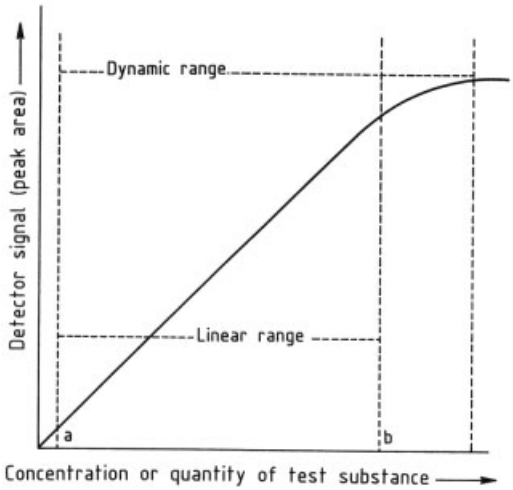

Figure 35. Dynamic range of a GC detector
a) Minimum detectability; b) Upper limit of linearity

Sensitivity is the signal output per unit concentration or unit mass of a substance in the mobile phase and is best expressed by the minimum detectable quantity, which is defined as the concentration or the amount of analyte for which the peak height is four times the intrinsic noise height of the detector (signal-to-noise ratio $S/N = 4$). The *dynamic range* of a detector is that range of concentration or quantity over which an incremental change produces an incremental change in detector signal. Figure 35 represents a plot used for the determination of the dynamic range of a detector. That part of the curve for which the response increases proportionally with increased concentration or quantity is called the *linear range.* For quantitation, work in the linear range is preferable, it is expressed numerically as the ratio of the upper limit of linearity (b) to the minimum detectability (a), both measured for the same substance. The *detector response factor* f_i is a relative term expressing the sensitivity of a detector to a given compound i, relative to its sensitivity to a standard compound:

$$f_i = f_{st}\ (A_i/A_{st}) \tag{6.1}$$

where A_i and A_{st} are the respective peak areas and f_{st} is the detector response factor for the standard; f_{st} is usually assigned a value of 1.000.

Detector selectivity is the ratio of the signal obtained for a compound to which the detector has special selectivity versus the signal obtained for a compound to which the detector is not particularly sensitive. Both signals must be measured for equal amounts. A compilation of selectivity, minimum detectability, and linear range of the most common GC detectors is given in Table 4.

6.2. Universal Detectors

6.2.1. Flame Ionization Detector

The flame ionization detector (FID) is the most widely used GC detector. Its operation is simple, sensitivity is of the order of 10 pg carbon per second with a linear range of 10^7, the response is fast, and detector stability is excellent. The carrier gas is mixed with hydrogen, and this mixture is combusted in air at the exit of a flame jet. Ions are formed that are collected at an electrode producing a current that is proportional to the amount of sample compound in the flame. Analytes such as permanent gases, nitrogen oxides, sulfur oxides, carbon oxides, carbon disulfide, water, formic acid, formaldehyde, etc., do not provide a significant FID response. The flows of carrier and combustion gas should be set properly for optimal FID operation. Typical flow rates are 1 : 1 : 12 (30 : 30 : 360 mL/min), respectively, for nitrogen as carrier gas, and the combustion gases hydrogen and air. If hydrogen is applied as carrier gas in packed column GC, the hydrogen combustion gas is replaced by nitrogen. For capillary GC where the carrier gas flows are much lower (e.g., 1 – 3 mL/min), hydrogen combustion gas is adjusted to a total hydrogen flow of 30 mL/min. Nitrogen can be added as make-up gas, which increases sensitivity and stability slightly. Detector temperature must always be kept above 150 °C to prevent condensation of water produced in the combustion process. Flame tip and electrodes should be cleaned

Table 4. Characteristics of common GC detectors

Name	Type	Selectivity	Minimum detectability	Linear range
Flame ionization detector (FID)	universal [c]	[a]	10 pg C/s	10^7
Thermal conductivity detector (TCD)	universal [c]	[b]	1 ng/mL mobile phase	10^6
Electron capture detector (ECD)	selective	compounds capturing electrons, e.g., halogens	0.2 pg Cl/s	10^4
Nitrogen phosporus detector (NPD)	selective	N and P	1 pg N/s 0.5 pg P/s	10^4
Flame photometric detector (FPD)	selective	P and S	50 pg S/s	10^3
			2 pg P/s	10^4
Photoionization detector (PID)	selective	aromatics	5 pg C/s	10^7
Electrolytic conductivity detector (ELCD)	selective	halogens and S	1 pg Cl/s	10^6
			5 pg S/s	10^4
Fourier transform infrared spectroscopy (FTIR)	universal [c]	molecular vibrations	1 ng (strong absorber)	10^3
Mass spectroscopy (MS)	universal [c]	characteristic ions	1 ng full-scan mode	10^5
			1 pg ion-monitoring mode	
Atomic emission detection (AED)	universal [c]	any element	0.2 – 50 pg/s depending on element	10^4

[a] FID responds to all organic compounds that ionize in a flame.
[b] TCD responds if thermal conductivity is different from carrier gas.
[c] Spectroscopic detectors that can be operated in a universal or a selective mode.

regularly because certain compounds can cause deposits that must be removed mechanically to prevent loss in sensitivity and detector stability. Bleeding of stationary phases and the use of derivatization agents are important sources of sensitivity loss and spikes. The flame shape for capillary columns should be optimized, and therefore flame tips with different bores are available. Small-bore jets produce the greatest signal for capillary GC.

6.2.2. Thermal Conductivity Detector

The thermal conductivity detector (TCD), also called the *hot-wire detector* or *katharometer*, measures the change in the thermal conductivity of the mobile-phase stream. Helium is the recommended carrier gas because of its high thermal conductivity. When an analyte is present in the carrier gas, the thermal conductivity will drop and less heat is lost to the cavity wall. A filament in the detector cell operated under constant voltage will heat up and its resistance will increase. This resistance is measured and recorded.

In the *original TCD* design the resistance is incorporated in a Wheatstone bridge circuit. The difference in thermal conductivity between the column effluent R_1 and a reference flow R_2 is recorded. This original design is still used in packed column GC (e.g., in process control). Nowadays, *single-filament,* flow-modulated thermal conductivity detectors are available for capillary work. The cell size is reduced to 3.5 µL to minimize detector band broadening, and the single filament measures the reduction in voltage required to maintain a constant filament temperature. A diagram of a commercially available TCD cell is presented in Figure 36. In Figure 36 A, the switching flow pushes the column effluent through the filament channel. On changing the switching flow (Fig. 36 B), the column effluent passes through the empty channel while the filament channel fills with the pure reference gas. Switching between column effluent and reference gas occurs every 100 ms. An important parameter in practical work is the detector temperature. The lower the temperature, the higher is the sensitivity and the longer the lifetime of the filament. The temperature must, however, be high enough to avoid condensation of the solutes in the detector cell. Oxygen-contaminated carrier gas, column bleeding, and active components such as strong acids and bases all decrease sensitivity and reduce the lifetime of the filament. Note also that if helium is the carrier gas, and hydrogen together with some other solutes must be measured, the hydrogen peak will be negative because its thermal conductivity is higher than that of helium, while the other solutes give positive peaks.

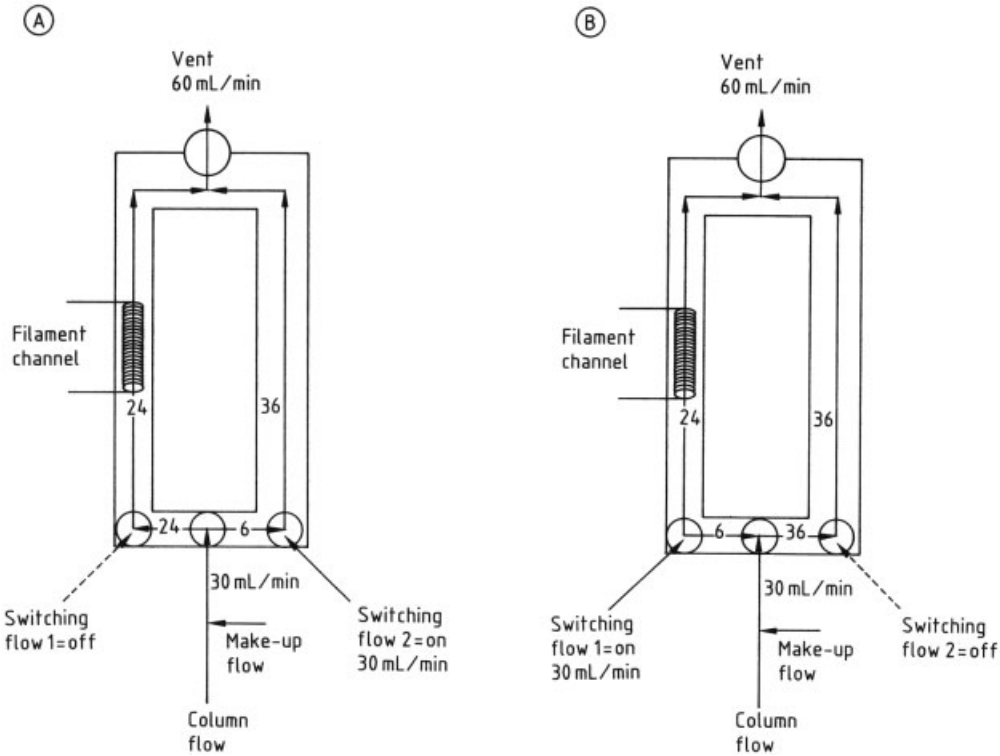

Figure 36. Diagram of a miniaturized TCD cell
A) Measurement of column effluent flow; B) Measurement of reference flow

6.3. Selective Detectors

6.3.1. Electron Capture Detector

One of the most popular selective GC detectors in use today is the electron capture detector (ECD). The ECD is an extremely powerful tool in environmental and biomedical studies. The principle is based on the fact that electronegative organic species (RX) react with low-energy thermal electrons to form negatively charged ions. The loss of electrons from the electron stream that is measured is related to the quantity of analyte in the carrier gas stream. The capturable electrons originate from ionization of the carrier gas or a make-up gas (usually nitrogen or an argon – methane mixture 95 : 5 or 90 : 10) by β particles from a radioactive source such as ^{63}Ni placed in the cell. The electron flow results in a current that is measured. Electronegative species in the carrier gas capture electrons and decrease the current. The highest sensitivity is noted for organic polyhalides such as trihalomethanes, organochloropesticides, PCBs, dioxins, and furans. Sensitivity increases in the order F < Cl < Br < I. The selectivity depends strongly on the compound's affinity for electrons and for polyhalides it is as high as 10^6 compared to hydrocarbons, and 10^4 compared to ketones, anhydrides, amines, etc. By chemical derivatization, the applicability of ECD can be expanded for trace analysis of species that do not or only weakly capture electrons. The best-known reagents are pentafluorobenzoic acid anhydride and pentafluorobenzoyl chloride for the derivatization of alcohols, amines, and phenols. (The fluorocompounds are chosen because of their high volatility, even though the detector is less sensitive for fluorine.) The derivatives formed are generally stable and have good chromatographic properties. As an example, the EPA method for the determination of chlorophenols in water samples includes a derivatization step with pentafluorbenzoyl chloride followed by ECD detection, in this way enhancing both

sensitivity and selectivity. For proper use of the ECD detector, the carrier gas and the make-up gas must be very clean and dry. In packed column GC, nitrogen or argon – methane is used as carrier gas and no make-up gas is required at the column outlet. In capillary GC, helium or hydrogen is the preferred carrier gas because of the speed of analysis, and 25 – 30 mL/min nitrogen or argon – methane is added as make-up gas to produce the thermal electrons. The column should also be preconditioned before connection to the detector. High stability and sensitivity are guaranteed only with the detector in constant operation. Contamination is normally corrected for by baking out the detector at 350 °C over a 24-h period. Some typical ECD chromatograms are shown in Figure 37 for the direct analysis of trihalomethanes in drinking water [57] and in Figure 38 for the determination of organochloropesticides.

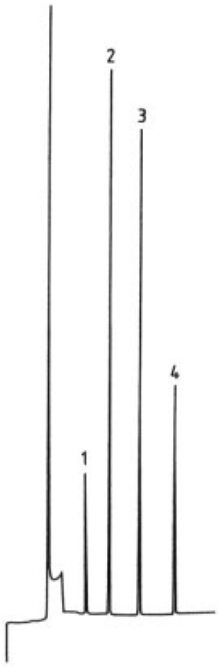

Figure 37. Direct analysis of trihalomethanes in drinking water
Column: 30 m $L \times 0.53$ mm *i.d.*; 2.65 µm d_f MeSi; retention gap 1.5 m $L \times 0.53$ mm *i.d.*; temperature: 50 – 150 °C at 10 °C/min; cool on-column injection; carrier gas 10 kPa hydrogen; ECD
Peaks: (1) chloroform, (2) dichlorobromoform, (3) dibromochloroform, (4) bromoform

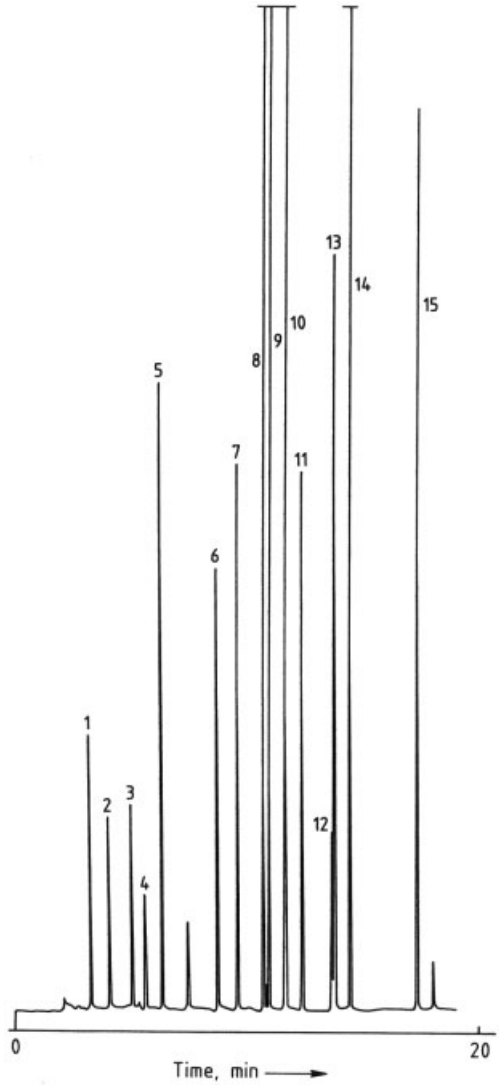

Figure 38. Analysis of organochloropesticides with ECD detection
Column: 20 m $L \times 0.32$ mm *i.d.*; 0.25 µm d_f 7 % CPSi, 7 % PhSi, 84 % MeSi (tailor-made called 1701); temperature: 180 – 240 °C at 4 °C/min; carrier gas hydrogen 8 psi ($\approx 5.6 \times 10^4$ Pa); ECD
Peaks: (1) HCB, (2) hexachlorocyclohexane, (3) lindane, (4) heptachlor, (5) aldrin, (6) heptachlorepoxide, (7) α-endosulfan, (8) 1,1-dichloroethenylidenbis(4-chlorobenzene), (9) dieldrin, (10) endrin, (11) 2,4-DDT, (12) β-endosulfan, (13)1,1′-(2,2-dichloroethylidene)bis(4-chlorobenzene),(14) 4,4′-DDT, (15) methoxychlor

6.3.2. Nitrogen Phosphorus Detector

The nitrogen phosphorus detector (NPD), also called the *thermionic detector* (TID) is similar in construction to the FID. The detector is made element specific for N and P because a source of an alkali salt – salts of potassium and rubidium are more commonly used – is positioned above the flame jet. By careful control of the combustion gases hydrogen and air, whose flow rates are less than those for the normal FID, the ionization of compounds containing no nitrogen or phosphorus is minimized, while the ionization of N- and P- containing compounds is increased. The exact mechanism is still unknown. Small changes in detector gas flows have a significant effect on sensitivity and selectivity. Optimal gas flows can be found by analyzing a mixture containing azobenzene, heptadecane, and malathion in the ratio 1 : 1000 : 2 (100 pg/µL, 100 ng/µL, and 200 pg/µL, respectively). Prominent responses shoin recent years uld be obtained for azobenzene and malathion, whereas the hydrocarbon should hardly be detected. Optimal flow rates on the order of 4 – 10 mL/min for hydrogen and 60 –100 mL/min for air are typical. For some versions of the NPD detector the use of a make-up gas such as helium increases the N or P response. The NPD detector is a valuable tool in clinical and pharmaceutical analysis, and especially in environmental monitoring of organophosphorus- and organonitrogen- containing pesticides.

6.3.3. Flame Photometric Detector

The flame photometric detector (FPD) uses the principle that sulfur- and phosphorus-containing compounds emit light by chemiluminescence when burned in a hydrogen-rich flame. For sulfur, for example, electronically excited diatomic sulfur ($S*_2$) is formed, which upon relaxation emits light over the wavelength region 320 – 450 nm, with an emission maximum at 394 nm. An optical filter at 394 nm permits the light to enter a photomultiplier, and a signal is produced. For phosphorus the interference filter is 526 nm. Although widely used mainly as a sulfur-selective detector in the petroleum industry, for the analysis of flavors and fragrances, in environmental analysis, etc., the FPD detector has far from ideal characteristics as a GC detector. It has many shortcomings; the most important one is that the intensity of light emitted is not linear with concentration, but approximately proportional to the square of the sulfur atom concentration. Careful calibration is a must when quantitation is aimed at. Other negative aspects are that the flame extinguishes when high amounts of solvents are introduced (this problem is eliminated when a dual-flame burner is used), hydrocarbon quenching occurs, and self-quenching occurs at high concentrations of the heteroatom species. Moreover, optimization of the different flows is very critical for optimal sensitivity and selectivity. Therefore, for the analysis of sulfur-containing compounds the FPD detector is often replaced by the atomic emission detector, the Hall electrolytic conductivity detector, or the chemiluminescence detector based on ozone reactions, whereas for organophosphorus compounds the NPD is preferred.

6.3.4. Overview of Other Selective Detectors

Two of the more recently developed detectors, namely, the Hall electrolytic conductivity detector (ELCD) and the photoionization detector (PID) are recommended by the EPA for the analysis of volatile and semivolatile halogenated organic compounds and low molar mass aromatics. Chemical emission based detectors, such as the thermal energy analyzer (TEA) for its determination of nitrosamines and the redox chemiluminescence detector (RCD) for the quantitation of oxygen- and sulfur- containing compounds, are becoming more and more important for some typical applications.

6.4. Detectors Allowing Selective Recognition

The selective detectors discussed in the previous sections often do not provide enough information to elucidate with 100 % probability the nature of the eluting solutes. For this reason, data with selective detectors can be erratic. The future in this respect definitely belongs to the spectroscopic detectors that allow selective recognition of the separated compounds. Today, the hyphenated techniques CGC – mass spectroscopy

(CGC – MS), CGC – Fourier transform infrared spectroscopy (CGC – FTIR), and CGC – atomic emission detection (CGC – AED) are the most powerful analytical techniques available. They provide sensitive and selective quantitation of target compounds and structural elucidation or identification of unknowns. The applicability and ease of use of the hyphenated techniques were greatly increased by the introduction of fused silica wall coated open tubular columns. The main reason for this is that because of the low flows of capillary columns, no special interfaces are required and columns are connected directly to the different spectrometers. The introduction of relatively inexpensive benchtop hyphenated systems has enabled many laboratories to acquire such instrumentation, which in turn has expanded their applicability ever further.

The hyphenated techniques CGC – MS, CGC – FTIR, and CGC – AED are generally used as stand-alone units. Due to the nondestructive character of FTIR, CGC – FTIR – MS units (Fig. 39) are possible and have been commercialized. The software then allows simultaneous recording of the infrared and mass spectra of the eluting compounds. In principle, CGC – FTIR – MS – AED is also possible if an open split interface is applied for the CGC – FTIR – MS combination and the split-line is directed into the AED detector. The fundamental aspects of CGC – MS, CGC – FTIR and CGC – AED are discussed in [58], [59], and [60].

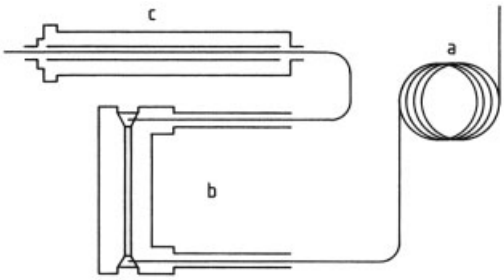

Figure 39. Diagram of a capillary GC – FTIR – MS system a) Capillary GC; b) FTIR; c) MS

6.4.1. Mass Spectrometry

The principle of CGC – MS is illustrated with a quadrupole mass analyzer as an example. As sample molecules exit the chromatographic column, they are introduced in the ion source of the mass spectrometer and ionized. Depending on the ionization method, molecular ions or fragment ions are formed, which are accelerated and separated from each other according to their mass-to-charge ratios (m/z) by a mass analyzer such as a quadrupole (c) (Fig. 40). Other mass analyzers are magnetic sector instruments, ion trap detectors, and time-of-flight mass spectrometers. The separated ions are then detected by a electron multiplier (d), and the resulting mass spectrum is a line spectrum of intensity of ions (y-axis) versus their m/z ratio (x-axis).

Two different ionization methods are available in CGC – MS, namely, electron ionization (EI) and chemical ionization (CI). In *electron ionization*, the molecules are bombarded with electrons of 70 eV emitted from a rhenium or tungsten filament. A molecular radical ion is formed ($M^{\bullet +}$) with sufficient amount of energy accumulated in its bonds to dissociate into typical fragment ions, radicals, and neutral species.

$$\begin{aligned} ABCD \longrightarrow ABCD^{+\bullet}(M^{+\bullet}) + e^- & \\ \longrightarrow & \; AB^{+\bullet} + CD \\ \longrightarrow & \; AB^{+} + CD^{\bullet} \\ \longrightarrow & \; ABC^{+} + D^{\bullet} \\ & \quad \longrightarrow AB^{+} + C \\ & \quad \longrightarrow A^{+} + BC \end{aligned}$$

Since low ion source pressures are employed (10^{-5} torr $\approx 1.33 \times 10^{-2}$ Pa), reactions in the ion source are unimolecular, and association between molecule and fragment ions does not occur. A disadvantage of electron ionization is that energy in the molecule is so large that complete fragmentation can occur and the molecular ion is absent. Less fragmentation is obtained with *chemical ionization*, which is a soft technique of ionizing molecules through gas-phase ion – molecule reactions. In chemical ionization, a reagent gas such as methane is introduced into the ion source and a relatively high source pressure is maintained (1 torr $\approx$ 133 Pa). The reagent gas is ionized by the electrons, and the ions formed interact with the sample molecules. Since these reactions are low in energy, abundant quasi-molecular ions, most often $M^{\bullet +} + 1$ are observed. Figures 41 and 42 show the electron-impact and the chemical-ionization spectra of trimethylsilylated clenbuterol. Chemical ionization is very useful to elucidate the molecular ion of unknowns and is often the ionization method

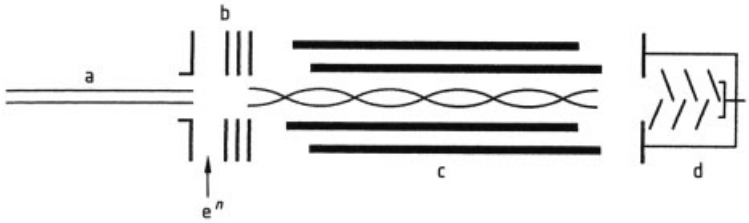

Figure 40. Schematic of a capillary GC – MS system
a) Capillary column; b) Ion source; c) Quadrupole mass filter; d) Multiplier

of choice for quantitative mass spectroscopy via ion monitoring. After ionization, the positive ions are accelerated and separated in the mass analyzer by tuning the voltage on the quadrupole rods. An ion of unit mass will pass the rods and reach the detector only if the voltages are properly adjusted. To detect all ions, the voltage is varied in time, and the time to detect all ions is called a scan. This means that if all ions between mass 20 and 400 are to be detected in 1 s, the voltages are on the exact value for each ion to pass during 1/380 of a second. This mode, called the *full-scan mode*, is necessary to identify compounds via their fragmentation pattern. Data are acquired by continuous repetitive scanning of the column eluate. The reconstructed chromatogram of a mass spectrometer is a plot of the total ion current as a function of time or scan number. All recorded spectra or scans are stored in the computer and can be recalled for identification. Nowadays this is performed mostly by comparison with a spectral library. The information in the stored mass scans can provide, in a simple way, valuable information on specific compounds or compound classes. This is illustrated in Figure 43, which shows the reconstructed total ion chromatogram of a solvent mixture (Fig. 43 A) and the reconstructed chromatogram of three specific ions typical for C_3 (trimethyl or methylethyl), C_4 (tetramethyl, dimethylethyl, diethyl, etc.), C_5 (pentamethyl, trimethylethyl, dimethylpropyl, etc.) benzenes (Fig. 43 B – D). This is called *mass chromatography*, and only data stored in the computer of the full-scan mode are used to provide selectivity. *Mass fragmentography* or *selected ion monitoring* (SIM), on the other hand, is a method to quantify target compounds with high selectivity and sensitivity. In this mode of operation, the voltages on the quadrupole rods are adjusted stepwise to detect only two, three, or four ions. The time the rods are on a particular voltage to let an ion pass is much longer than in the case of the full-scan mode, which results in enhanced sensitivity. As an illustration, in the previous case for the full-scan mode, each ion was seen for only 1/380 of a second. Suppose clenbuterol has to be quantified. By selecting the ions 259 and 349 only (Fig. 42), each ion is detected in a 1-s scan for 1/2 s. Compared to the full-scan mode, the residence time is 190 times higher; thus the sensitivity increases by a factor 190. With modern CGC – MS systems, picogram amounts can easily be quantified when operated in the ion-monitoring mode. An interesting feature of SIM is that internal standards can be selected that chemically, physically, and chromatographically behave exactly the same as the compound to be measured. In the case of clenbuterol, for example, the best internal standard is D_6- clenbuterol with typical ions at *m/z* 265 and 355. In routine analysis, D_6- clenbuterol is added to the sample (i.e., urine) before any pretreatment. Sample preparation and cleanup are then performed, and the four ions 259 and 349 for clenbuterol and 265 and 355 for the internal standard D_6- clenbuterol are recorded. Quantitation is very easy.

6.4.2. Fourier Transform Infrared Spectroscopy

In the combination CGC – FTIR, IR spectra of the eluting peaks are recorded as they emerge sequentially from the column outlet. The eluate is introduced in a light pipe where compounds absorb radiation at well-defined frequencies (i.e., frequencies characteristic of the bonds within a molecule). The absorption is sensed and recorded. The sensitivity of FTIR depends on the

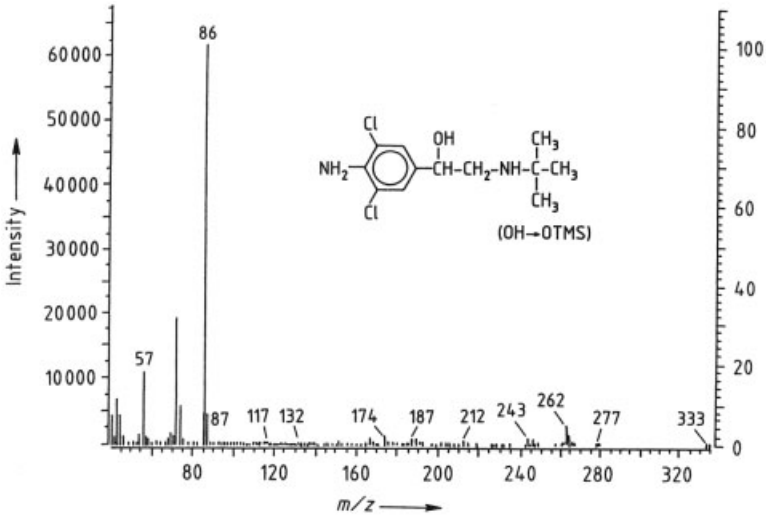

Figure 41. Electron ionization spectrum of TMS – clenbuterol

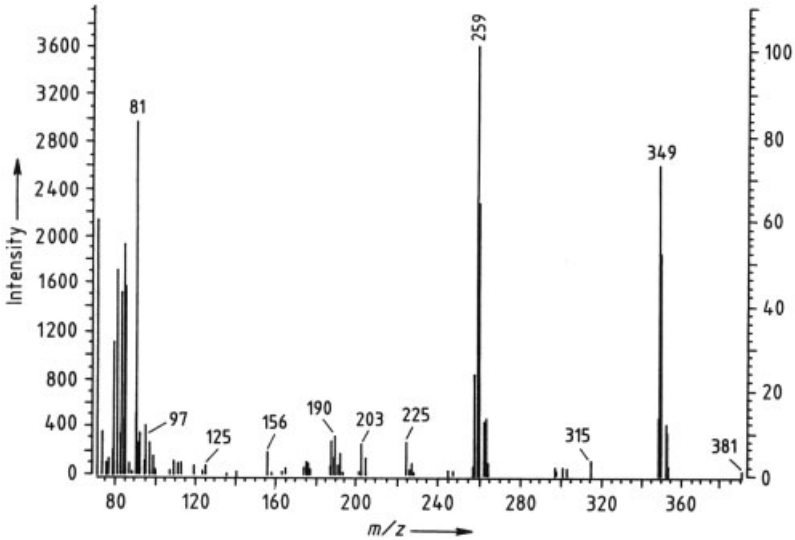

Figure 42. Chemical ionization spectrum (reagent gas isobutane) of TMS – clenbuterol

functionalities within a molecule, and for strong IR absorbers, 1 ng yields a good spectrum. Current FTIRs include computers, and the recorded spectra can be compared with a library of spectra to aid in compound identification. Monitoring of specific wavelengths allows the elucidation of classes of compounds (e.g., aldehydes, ketones, alcohols, ethers). On-line coupling of CGC – FTIR and CGC – MS can be realized in series since the FTIR is nondestructive. The IR spectra are complementary in nature to MS spectra, especially for the elucidation of isomers for which MS is not very informative. A typical example is the determination of impurities in α-angelica lactone. Figure 44 shows the total response chromatogram obtained with FTIR and the total ion chromatogram, recorded by the mass spectrometer. Both chromatograms correspond very well, and at any time MS and IR data can be recalled and compared. The mass spectra of peak 1 and 2 are very similar and do not allow the differentiation of structural differences. The infrared

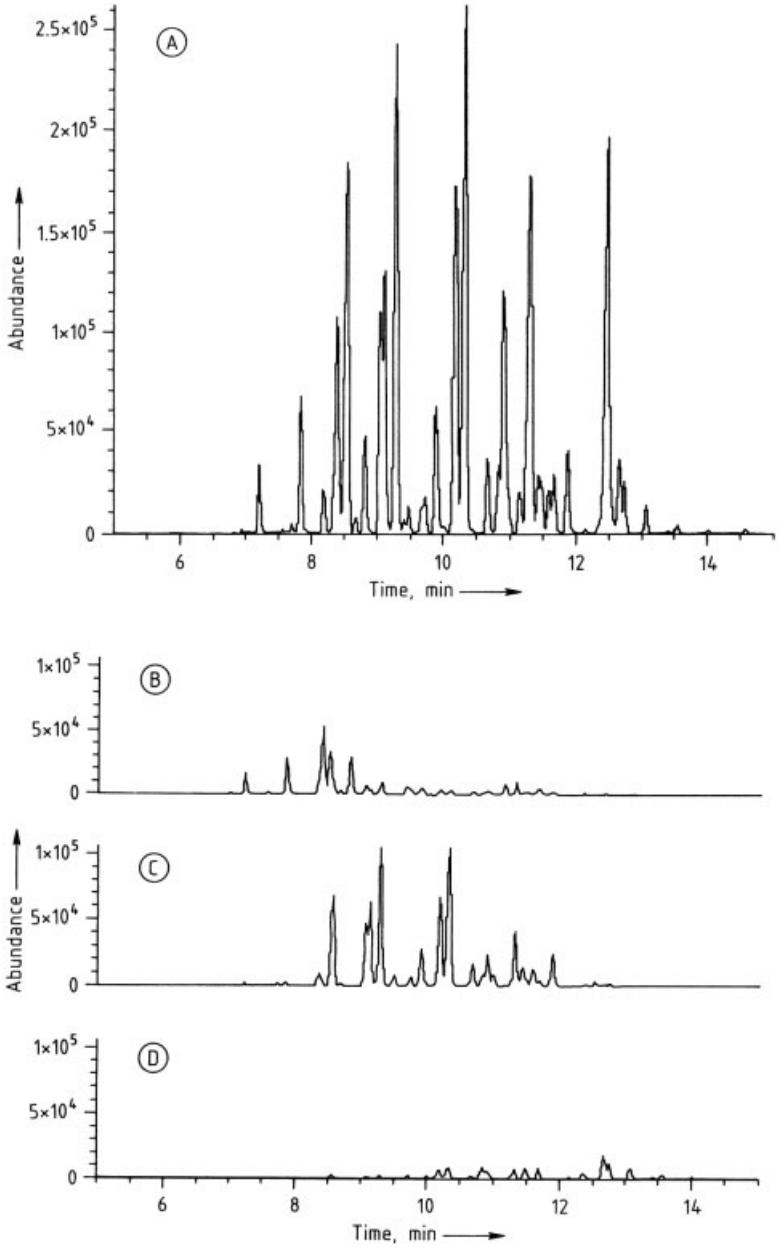

Figure 43. Illustration of mass chromatography
A) Solvent mixture; B) C_3-Benzene; C) C_4-Benzene; D) C_5-Benzene

spectra of peak 1 and 2 are compared in Figure 45. Some important differences are noted. The shift of the carbonyl band from 1833 cm^{-1} (peak 1) to 1806 cm^{-1} (peak 2) indicates that peak 2 corresponds to β-angelica lactone (α, β-unsaturated instead of β, γ-unsaturated).

6.4.3. Atomic Emission Detection

One of the latest developments in hyphenated techniques is the atomic emission detector. The availability of a bench-top model CGC–AED

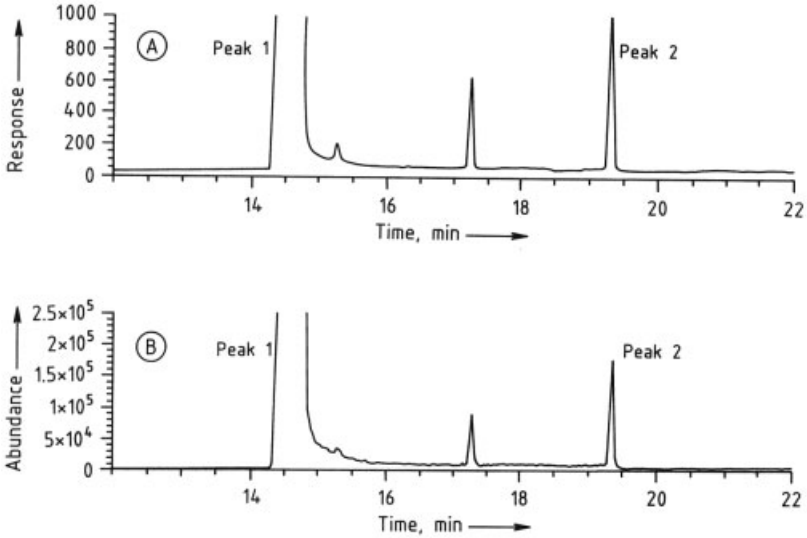

Figure 44. Capillary GC – FTIR – MS chromatograms of α-angelica lactone
A) Total response chromatogram (FTIR); B) Total ion chromatogram (MS)

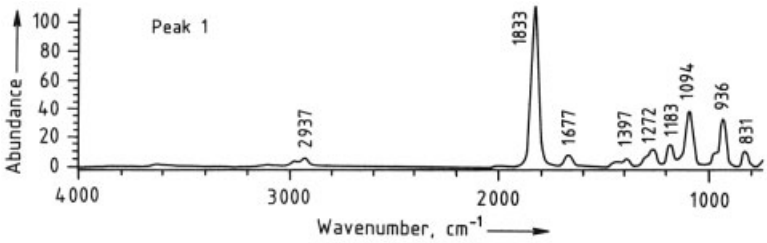

Figure 45. Infrared spectra of peak 1 and 2 of chromatogram in Figure 44

enables the use of this powerful technique in routine analysis. Detectabilities of this element-specific detector are of the order of 0.1 pg/s for organometallic compounds, 0.2 pg/s for carbon (more sensitive than FID), 1 pg/s for sulfur, and 15 pg/s for nitrogen, to mention only a few. The power of the technique lies in its supreme selectivity: all elements can be detected selectively. As opposed to ECD, AED allows differentiation between various halogenated organic compounds (e.g., fluoro, chloro, and bromo components). Multielement analysis can be carried out by simple preselection of the atoms to be monitored. In AED, the solutes eluting from the column are atomized in a high-energy source; the resulting atoms are in the excited state and emit light as they return to the ground state. The various wavelengths of the emitted light are then dispersed in a spectrometer and measured by using a photodiode array. Each element has a typical emission spectrum, with emission lines commonly occurring in clusters. The relative intensities of lines within these clusters are constant. Figure 46 shows the emission spectrum of a sulfur- containing compound, and the 180.7-, 182.0-, and 182.6-nm cluster serves as conclusive proof of the presence of sulfur in a peak.

Multiwavelength detection became possible through the introduction of diode array technology. Quantitative treatment of the data obtained in a multielement analysis moreover allows the calculation of empirical formulas, yielding information that is complementary to and helpful in the interpretation of mass spectral data. Quantitation is simplified considerably, since calibration of the detector no longer depends on the type of components to be quantified. Nontoxic compounds can thus be used to quantify harmful chemicals. Figure 47 illustrates the sensitivity of the detector for organometallic compounds.

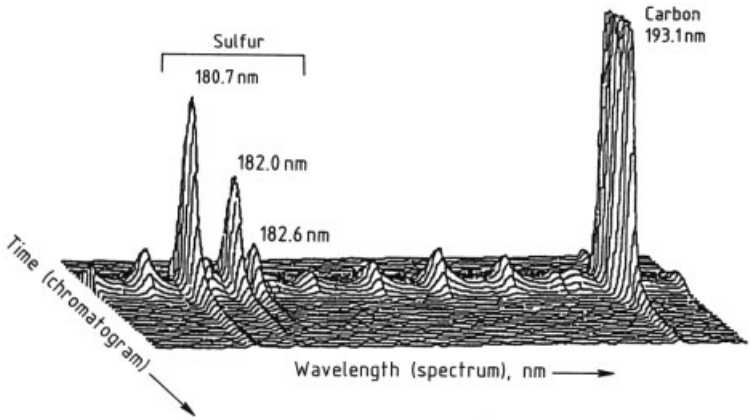

Figure 46. Emission spectrum of a sulfur-containing compound

Organolead halides are derivatized with butylmagnesium bromide into tetraalkylated solutes, which are amenable to CGC analysis. By selecting a wavelength of 406 nm, high selectivity and sensitivity (10 pg per compound) are guaranteed. Another important application concerns direct monitoring of oxygenates in petroleum products. Figure 48 shows the direct analysis of a light naphtha containing low concentrations (ppm) of oxygenated compounds. The concentrations for methanol and ethanol are 100 and 50 ppm, respectively.

7. Practical Considerations in Qualitative and Quantitative Analysis

A chromatogram provides information on the complexity (number of peaks), identity (retention time), and quantity (peak area or height) of the components in a mixture. This information can be considered suspect if the quality of separation is not optimal. Capillary columns have greatly enhanced the use of gas chromatography as a qualitative and quantitative tool.

7.1. Qualitative Analysis

Modern capillary GC is characterized by very good precision in retention time, and this allows the use of retention indices for peak identification. In a retention index system, the retention behavior of a particular solute is expressed in a uniform scale determined by a series of closely related standard substances. In the retention index scale developed by Kovats [61] for isothermal separations and by Van den Dool and Kratz [62] for temperature-programmed analyses, the fixed points of the scale are normal alkanes. The Kovats retention index system is illustrated in Figure 49. Compound X elutes between the hydrocarbons hexane and heptane with carbon numbers 6 and 7, respectively. For convenience, the carbon numbers are multiplied by 100, thus 600 and 700. The retention index I_x of compound X is 673.8. Indices are not graphically measured but calculated by Equation (7.1)

$$I_x = 100z + 100\,\frac{\log t'_{Rx} - \log t'_{Rz}}{\log t'_{Rz+1} - \log t'_{Rz}} \qquad (7.1)$$

where z is the carbon number of the n-alkane eluting before X; $z+1$ is the carbon number of the n-alkane eluting after X; and t'_R is the adjusted retention time.

To perform the interpolation, the adjusted retention times of the n-alkanes must be known. This is often done by analyzing a separate mixture containing n-alkanes. This requires, however, automated injection. For manual injection, n-alkanes are added to the sample. In Equation (7.1), logarithmic interpolation is necessary because the adjusted retention times increase exponentially with carbon number (Eq. 4.1). In linear temperature-programmed runs, the retention times of a homologous series (i.e., the n-alkanes) increase linearly with carbon number, and the

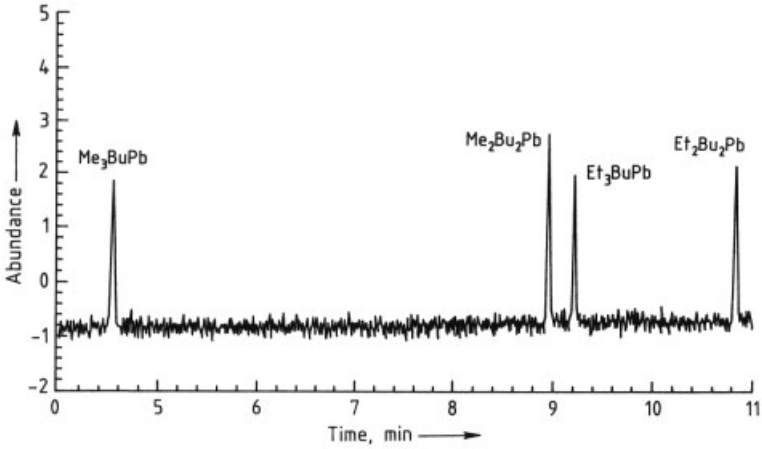

Figure 47. AED analysis of organolead compounds

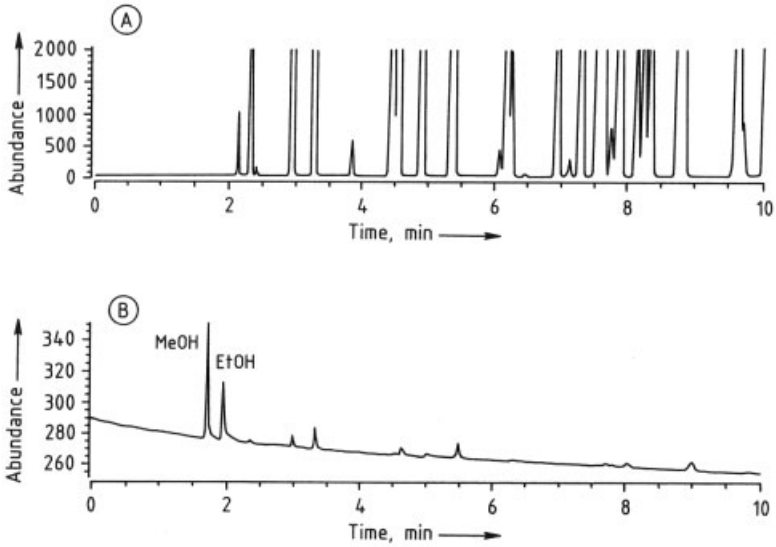

Figure 48. AED analysis of light naphtha in the C (A) and the O (B) emission line

equation for temperature-programmed analysis is simplified to

$$I_x = 100z + 100\,\frac{t_{\mathrm{R}x} - t_{\mathrm{R}z}}{t_{\mathrm{R}z+1} - t_{\mathrm{R}z}} \tag{7.2}$$

Unadjusted retention times may be used. The isothermal indices depend on the stationary phase and film thickness and on the temperature. Specification of those variables is a must. The programmed indices, often called linear retention indices, require specification of the temperature profile as well. Once determined, the indices can be used to identify compounds from libraries with tabulated indices [63], [64]. The general utility of those libraries for identification is questionable. Indices are best used within a laboratory for specific samples (i.e., petroleum products, essential oils, steroids, etc.). Greater confidence in identification is obtained by measuring indices on more than one stationary phase. Unequivocal compound elucidation requires spectroscopic characterization as provided by mass spectroscopy, Fourier transform infrared spectroscopy, and atomic emission detection.

Recently, the concept of retention time locking was introduced, and this opens new perspec-

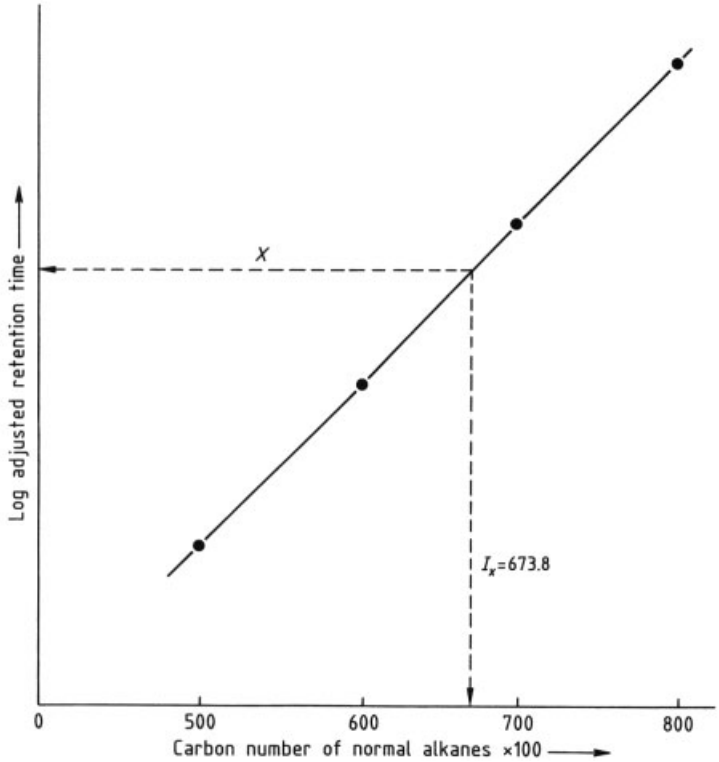

Figure 49. The Kovats retention index system

tives in capillary GC for solute elucidation (see Section 10.2).

7.2. Quantitative Analysis

Four techniques are commonly used in quantitative analysis: the normalization method, the external standard method, the internal standard method, and the standard addition method. Whatever method is used, the accuracy often depends on the sample preparation and on the injection technique. Nowadays these are two main sources of error in quantitative analysis. The quantitative results produced by PCGC and CGC are comparable.

8. Coupled Systems

Coupled systems include multidimensional and multimodal systems. *Multidimensional chromatography* involves two columns in series preferably two capillary columns, with different selectivity or sample capacity, to optimize the selectivity of some compounds of interest in complex profiles or to provide an enrichment of relevant fractions. In *multimodal systems*, two chromatographic methods or eventually a sample preparation unit and a chromatographic method are coupled in series. Coupled systems that received much interest in recent years are multidimensional CGC (MDCGC), the combination of high-performance liquid chromatography with CGC (HPLC – CGC) and the on- or off-line combination of supercritical fluid extraction with CGC (SFE – CGC). Multidimensional and multimodal techniques in chromatography are described in detail in [65].

8.1. Multidimensional Capillary GC

For highly complex samples, the separation power of a single capillary column is insufficient to achieve complete resolution for the compounds of interest. Even after optimization of

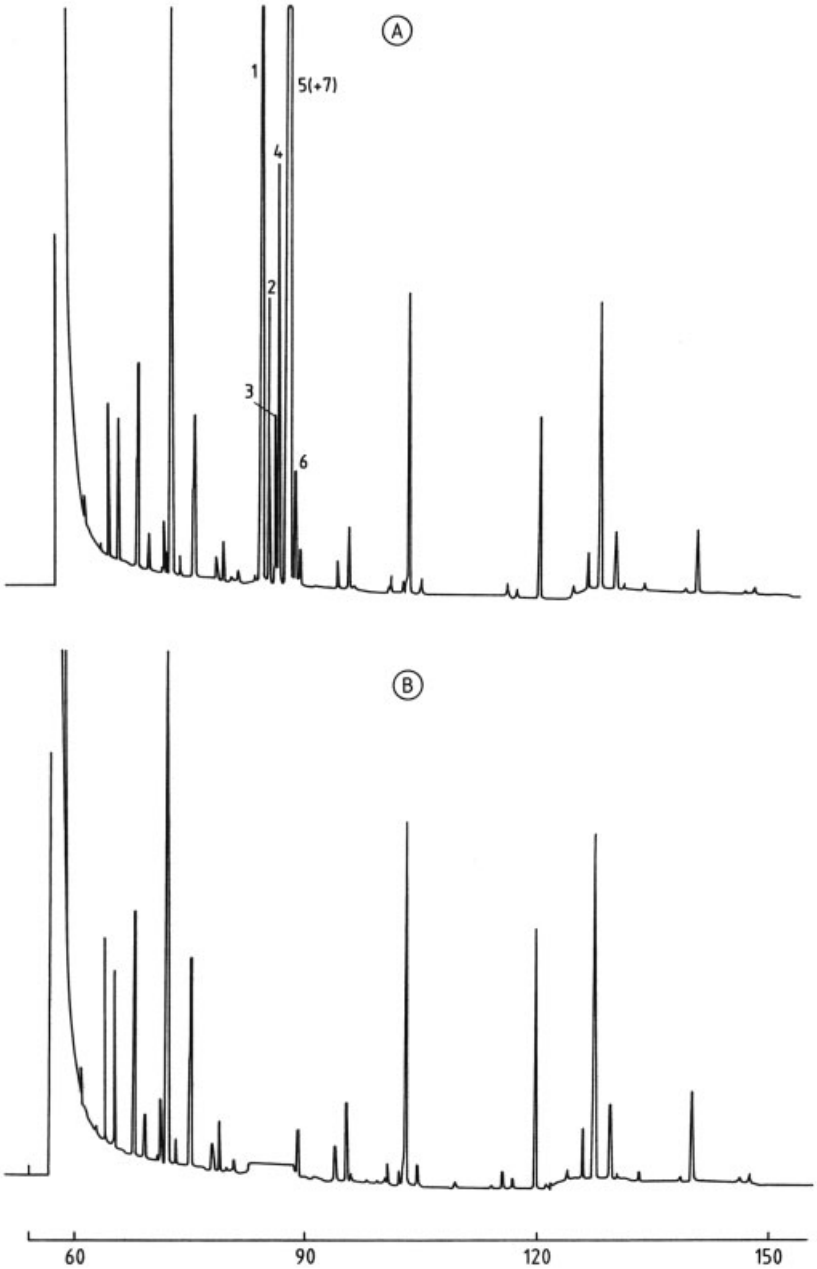

Figure 50. Multidimensional CGC analysis of pepermint oil, two column system constant pressure – constant pressure
A) Separation on MeSi; B) Analysis on MeSi after heartcut ; C) Analysis of heartcut fraction on column B with intermediate trapping; D) Analysis of heartcut fraction on column B without intermediate trapping (selectivity tuning)
Column A : 20 m $L \times 0.32$ mm *i.d.*; 0.24 µm d_f MeSi
Column B : 20 m $L \times 0.32$ mm *i.d.*; 0.24 µm d_f high-M_r PEG
Temperature: 60 °C (2 min) to 180 °C at 3 °C/min; split injection 1/50; carrier gas hydrogen; FID

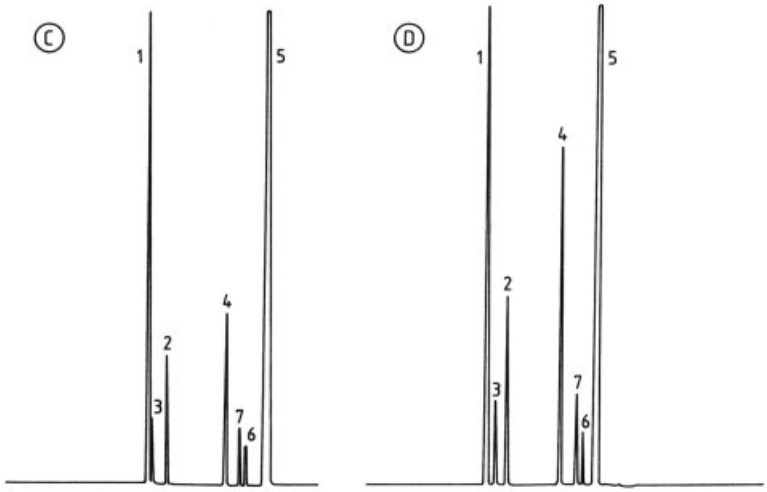

Figure 50. (Continued)

the selectivity, important compounds will still coelute, since the better separation of one pair of compounds is likely to be counteracted by the overlapping of another pair of compounds present in the sample. The only solution is the combination of more than one column system. In multidimensional CGC, a group of compounds that has not been separated on a first column is transferred (heartcut) to a second column, where complete resolution is achieved. Two operational modes can be distinguished. If the transferred fraction is coldtrapped before analysis on the second column (intermediate trapping), the selectivities of both columns are decoupled. The elution of the transferred solutes will be independent of the selectivity of the first column and will be the same as if the pure sample compounds were analyzed on the second column only. Alternatively, when no intermediate trapping is applied, the elution pattern observed after analysis on the second column will be the same as the elution pattern for analysis on a mixed-phase column or a coupled column system with an intermediate selectivity between the selectivities of both columns. Both principles are illustrated in Figure 50. Peppermint oil was analyzed on a methylsilicone capillary column, and the separation of the menthone – menthol fraction (peak 1 – 7) is incomplete (Fig. 50 A). This fraction was heartcut (Fig. 50 B) and further analyzed on a poly(ethylene glycol) capillary column. With intermediate trapping (Fig. 50 C) the separation is drastically improved, but the highest resolution, in this case, is without intermediate trapping (Fig. 50 D). This is an illustration of the power of multidimensional CGC in combination with selectivity tuning.

Developments in multidimensional CGC diverge in two main directions: the coupling of narrow to narrowbore capillaries to enhance resolution, on the one hand, and the coupling of widebore precolumns to narrowbore capillaries as a selective sample introduction method, on the other. Some inlet devices, such as the programmed temperature vaporization injector operated in the solvent vent mode, can also be classified as multidimensional systems. The solvent is removed through the split vent, while the solutes of interest remain in the liner. The multidimensional approach with selective sampling is of considerable interest in routine analysis. It allows preseparation of relatively large amounts of sample on the precolumn operated under constant-flow conditions. Fine-tuning of the separation can be achieved on the analytical column, operated in the constant-pressure mode. The system is easy to operate and can be installed without problems in every GC instrument. Systems based on constant pressure – constant pressure and constant flow – constant flow offer more flexibility because column dimensions are less critical than in the constant flow – constant pressure system; however, they are expensive and require time-consuming optimization procedures

when new applications are carried out. The possibilities of multidimensional capillary GC as a sampling method are illustrated with the analysis of volatiles emitted by living plants [66]. The atmosphere surrounding a plant (headspace) is sucked in an open tubular trap (3 m in length, 530 µm *i.d.*) coated with a 15-µm film of methylsilicone where the volatiles are retained by the normal chromatographic partitioning process at ambient temperature. The advantages over traps filled with adsorbents are that breakthrough volumes can be calculated, the inertness is high, and thermal desorption of polar compounds is feasible. After sampling, the trap is installed as precolumn in a multidimensional CGC system (Fig. 51). The volatiles are thermally desorbed, reconcentrated in a cold trap (e), and then analyzed on-line on the analytical capillary column (f). Figure 52 shows the headspace CGC – MS analyses of a healthy and a fungus-infected *Mentha arvensis* plant. Menthol (peak 15) is the main compound in the healthy plant (Fig. 52 A), and the highly volatile fraction is relatively poor. In the infected plant (Fig. 52 B), menthyl acetate (peak 21) is the main compound, while the relative percentage of menthone (peak 13) is increased and the highly volatile fraction is more abundant. This hyphenated combination live headspace sampling – multidimensional capillary GC – MS provides interesting results in phytochemical research. In 1992, a more sophisticated system, PTV injection –MDCGC with three columns –MS was described for the analysis of dioxins and furans [54].

8.2. Multimodal High-Performance Liquid Chromatography – Capillary GC

The on-line coupling of HPLC to capillary GC has received much attention since the mid-1980s. Pioneering work was performed by K. Grob, Jr. [67]. The coupling of HPLC to capillary GC can be considered as an on-line sample preparation and enrichment step. Preseparation can be carried out according to molecular mass (SEC), functionality (straight-phase LC), or hydrophobicity (reversed-phase LC). Similar results are obtained with off-line techniques (i.e., solid-phase extraction or LC off-line fractionation), but advantages of the on-line approach include easier automation, fewer artifacts, and higher sensitivity because the total elution volume of the fraction of interest is transferred to the capillary GC column.

A fully automated and flexible on-line LC – GC system has been described [68]. The system is based on standard LC and capillary GC instrumentation. The on-line LC – GC system was evaluated for the determination of pesticides in orange oil. The essential oil was fractionated by normal-phase LC, resulting in a separation according to polarity. The fraction containing pesticides was transferred and analyzed by capillary GC – NPD. The LC – GC interface is a modified large-volume sampler (Gerstel, Mulheim a/d Ruhr, Germany). The heart of the system is a specially designed flow cell that replaces a normal vial. The mobile phase leaving the LC detector is directed via a capillary tube with well-defined dimensions into a T-shaped flow cell. The cell is equipped with a septumless sampling head through which a syringe needle can be introduced. The sampler is completely computer controlled. To transfer a selected LC fraction, the transfer start and stop times, measured on the LC chromatogram, are introduced into the controller software. The time delay between the LC detector and the flow cell is automatically calculated from the dimensions of the connecting tubing and the LC flow rate. When the LC fraction of interest passes through the flow cell, the syringe needle penetrates the septumless head and samples the LC fraction at a speed equal to the LC flow rate. Up to 2 mL can be collected in the syringe. After collection, the needle is withdrawn from the flow cell, which rotates away from the PTV inlet, and a large-volume injection is made by using the PTV in the solvent-vent mode. Depending on the fraction volume and solvent type, the sample introduction parameters (inlet temperature, vent flow, vent time, injection speed, etc.) are calculated by the PTV calculator program. For fractionation of the orange oil sample, the following HPLC parameters were used: column 250 mm × 4.6 mm i.d. × 10 µm Lichrospher 100 DIOL, injection volume 20 µL, mobile phase in a gradient from 100 % hexane for 10 min, to 40 % isopropyl alcohol at 20 min (2 min hold) at a flow rate of 1 mL/min and UV detection at 220 nm. The fraction eluting from 4.4 – 4.9 min (volume = 0.5 mL) was automatically transferred to the GC inlet. The LC – GC interface was programmed to take the sample at a

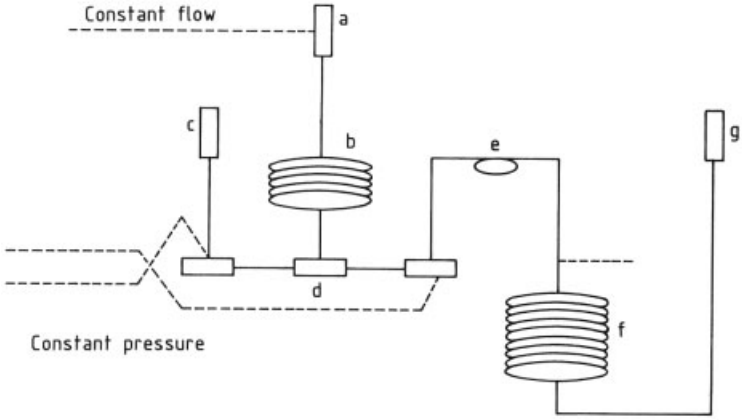

Figure 51. MDCGC system constant flow – constant pressure
a) Injector; b) Precolumn; c) Detector 1; d) Switching device; e) Cold trap; f) Analytical column; g) Detector 2

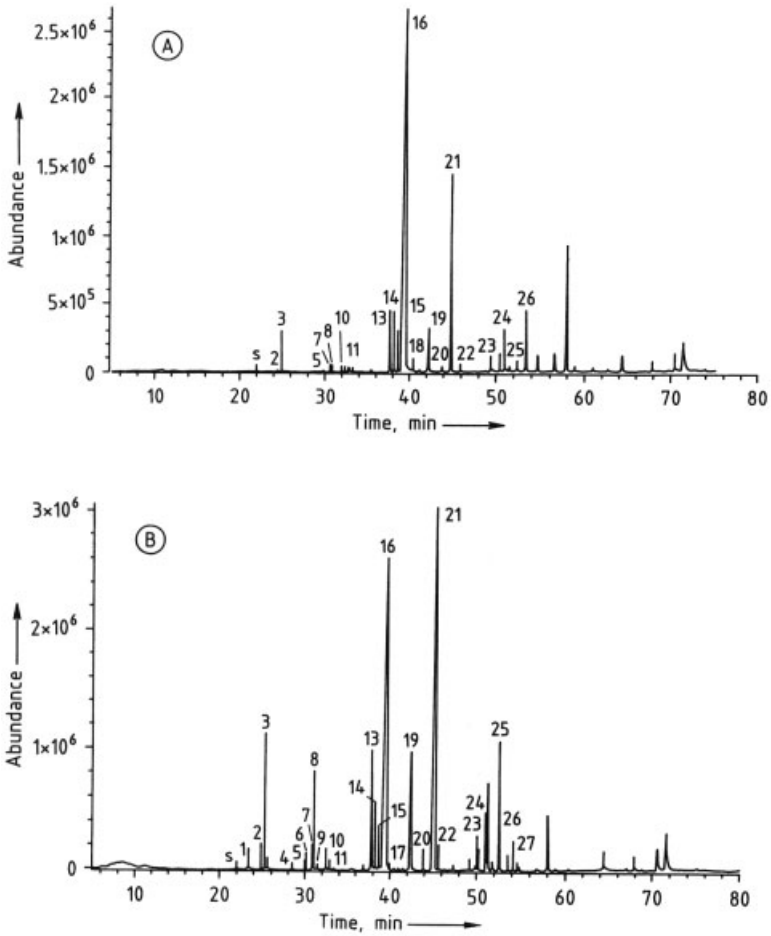

Figure 52. CGC pattern of the head space of a healthy (A) and an infected (B) *Mentha arvensis* plant
Thermal desorption and cryofocusing

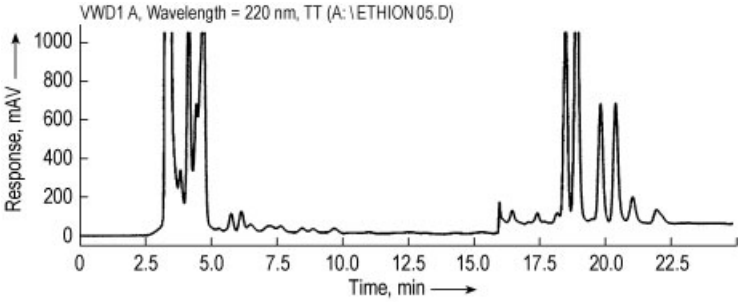

Figure 53. LC fractionation of orange oil on Lichrosorb 100 DIOL

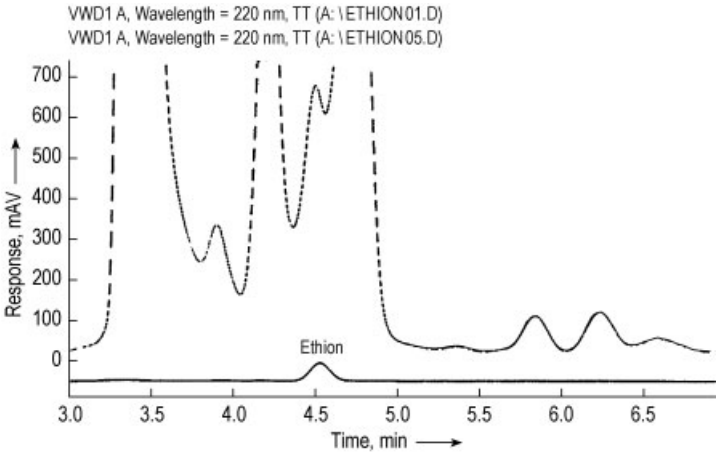

Figure 54. Overlay of relevant parts of LC chromatograms of a 1 ppm ethion standard (bottom trace) and orange oil (top trace)

sampling speed of 1000 µL/min (the same as the LC flow rate). The complete fraction of 500 µL was injected in the PTV inlet at 250 µL/min. This injection speed corresponds to the injection speed calculated by the PTV software program. Capillary GC – NPD analysis was performed on a 30 m × 0.25 mm i.d. × 0.25 µm MeSi column and the oven was programmed from 70 °C (2 min) to 150 °C at 25 °C/min and then to 280 °C at 8 °C/min (10 min). The detector was set at 320 °C with 3 mL/min hydrogen, 80 mL/min air, and 30 mL/min helium make-up flow.

The LC profile for the orange essential oil is shown in Figure 53. The apolar mono- and sesquiterpenes elute first, followed by the terpenoids and after 16 min also the flavanoids are eluted. These last compounds in particular give most interference in GC analysis because they have similar molecular weights to the pesticides. Using the same conditions, ethion and chlorpyriphos elute at 4.6 – 4.7 min. Figure 54 shows a comparison of the analysis of a 1 ppm ethion reference sample and the essential oil analysis.

The fraction eluting from 4.4 to 4.9 min was transferred to the CGC-NPD and analyzed. Figure 55 shows the resulting GC – NPD chromatogram and both ethion and chlorpyriphos (t_R 15.79 min) are detected. The chromatograms are very clean and no interferences are present. This demonstrates the excellent selectivity of the LC – GC combination. The concentrations of ethion and chlorpyriphos were calculated by external standard analysis and are 1.9 ppm for chlorpyriphos and 0.8 ppm for ethion.

8.3. Multimodal Supercritical Fluid Extraction – Capillary GC

Because supercritical fluids have excellent characteristics for the rapid extraction of organic

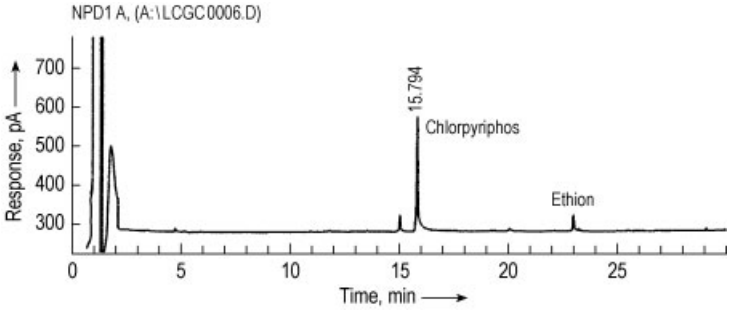

Figure 55. LC – Capillary GC – NPD analysis of orange oil

compounds, combining supercritical fluid extraction with CGC is an interesting approach to the enrichment of organic solvents [69], [70]. Since the density of supercritical fluids –usually carbon dioxide –can be varied (pressure, temperature), a high degree of selectivity can be introduced in the sample preparation procedure (i.e., prior to the analytical system). The selectivity can be further enhanced by making use of modifiers that may be polar (e.g., CO_2 – MeOH) or nonpolar (CO_2 – hexane). The main applications of SFE – CGC are related to solid samples, such as soil or sediment samples, and plant materials. For aqueous solutions (e.g., wastewater, urine), samples are enriched by solid-phase extraction. The extraction cartridges are then submitted to SFE – CGC, thus providing all the advantages of SFE (i.e., selective extraction and ease of automation). There is some controversy in the literature whether on-line SFE – CGC is the best approach to the application of analytical supercritical fluid extraction. Indeed off-line SFE has some advantages, and the use of robotics allows automated transfer of the sample to the capillary GC instrument. Selection of on-line or off-line depends mainly on the specific problem. Off-line is preferred for method development and on-line for routine analysis. On-line can, however, be recommended only when the matrix from which the solutes have to be extracted is always the same (e.g., in chemical composition, particle size, pore size). Moreover, the literature on SFE extraction times and recoveries is often incorrect because spiked samples are analyzed. Applied to real contaminated samples, SFE is found to be much more difficult. Figure 56 shows the optimized off-line SFE – CGC analysis of pesticides from sediment: 500 mg sediment was placed in the extraction vessel, and 250 µL of methanol was added on top of the sediment. SF extraction was performed at 60 °C with pure CO_2 at 35 MPa for 30 min. The concentration of the main compound dieldrin is 20 ppb. Classical Soxhlet extraction gave the same concentration.

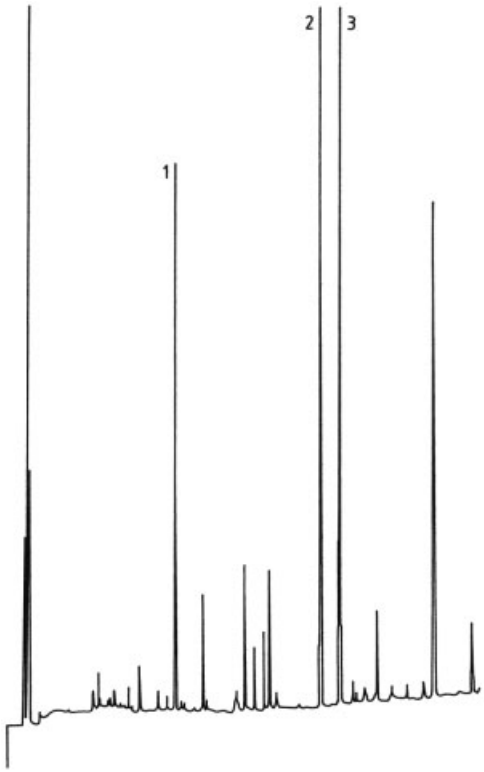

Figure 56. Analysis of SFE extract of a sediment with ECD detection
Peaks: (1) aldrin, (2) dieldrin, (3) endrin

9. Applicability

9.1. Solute Thermal Stability

Solutes subjected to GC must be thermally stable. The thermostability of organic compounds depends strongly on their nature, on the activity of the environment in which thermal stability is measured, and on the thermal stress given to the solutes. Capillary GC nowadays is performed in a completely inert system, i.e., highly pure carrier gas, purified stationary phases, fused silica with less than 0.1 ppm metal traces and specially deactivated, etc. Moreover, thermal stress can be avoided by applying cool on-column injection. Lipids serve as a good illustration. When oils or fats are used in food preparation, they decompose (formation of aldehydes etc.) or polymerize (formation of dimers, trimers, etc.) as a function of time, which make them no longer useful for cooking. These alterations are caused by the presence of water and oxygen. When heated under inert conditions, fats and oils are stable and evaporate. Figure 57 compares the thermogravimetric profiles recorded for triolein ($M_r = 886$) under a stream of pure nitrogen gas (a) and air (b). In present state-of-the-art GC, the systems are operated under circumstances shown in curve a.

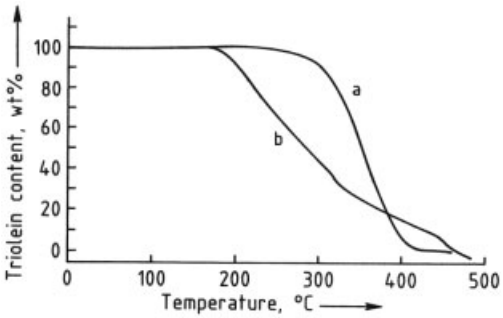

Figure 57. Thermogravimetric analysis of triolein
Ramp from 40 – 500 °C at 5 °C/min
a) Pure nitrogen; b) Air

9.2. Solute Volatility

Volatility is related to the vapor pressure (boiling point) of the compounds. Polydimethylsiloxanes with molecular masses as high as 5000 are volatile enough to be analyzed by GC, whereas poly(ethylene glycols) with M_r of ca. 1500 are not volatile at all. Volatility is thus related not only to the molecular mass, but to the polarity of the functional groups. Polydimethylsiloxane (M_r 5000) has approximately 68 apolar dimethylsiloxane units, whereas poly(ethylene glycol) (M_r 1500) has 114 polar ethoxy units besides the terminating hydroxyl groups. Some thermolabile compounds, although able to be volatilized, undergo partial decomposition at high temperature. Derivatization can be employed to impart volatility and to yield a more stable product, thereby improving chromatographic performance and peak shape. Three general types of analytical derivatization reactions are used in GC: alkylation, acylation, and silylation [71], [72]. The ideal derivatization procedure must be simple, fast, and complete. Alkylation, acylation, and silylation are used to modify compounds containing functional groups with active hydrogens such as −COOH, −OH, and $-NH_2$. The tendency of these functional groups to form intermolecular hydrogen bonds affects the inherent volatility of compounds containing them, their interaction with the stationary phase, and their stability. Figure 58 shows the analysis of sucrose in molasses with capillary GC. Trehalose is used as internal standard, and the sample is derivatized into the oxime – trimethylsilyl derivatives.

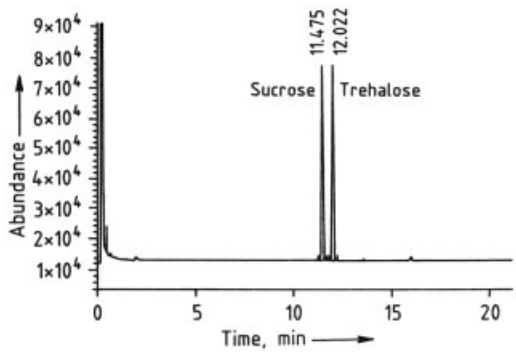

Figure 58. Capillary GC analysis of sugar derivatives (oxime – TMS)
Column: 10 m $L \times 0.53$ mm *i.d.*; 0.1 µm d_f MeSi; temperature: 80 – 300 °C at 10 °C/min; cool on-column injection; carrier gas hydrogen 4 mL/min; FID

9.3. Comparison of Gas Chromatography, Liquid Chromatography, and Supercritical Fluid Chromatography

When a scientist is faced with deciding which separation technique to use for a specific analytical problem, the primary question is, which chromatographic technique best suits this particular separation. The initial answer is dictated by the chemical nature of the analytes themselves. For the analysis of volatile components, gas chromatography is preferred. For analyzing strongly polar or ionic compounds, liquid chromatography may be the best choice. In some cases, supercritical fluid chromatography has distinct advantages over GC and LC. However, for many applications – such as analyses of polycyclic aromatic hydrocarbons, polychlorinated biphenyls, lipids, fatty acids, steroids, fat-soluble vitamins, mycotoxins, etc., – several chromatographic techniques can be considered.

In such cases, strange as it may seem, the choice of a chromatographic technique is often biased on secondary considerations such as the analyst's know-how and skill or the availability of a particular instrument. This choice is usually justified with supplementary arguments, especially those concerning the simplicity of a technique. Recent technical developments in nearly all branches of chromatography, however, have largely overcome any "advantage of simplicity" (to some, the "disadvantage of complexity") of various techniques. The extended possibilities of automation further weaken arguments of this kind.

More important, especially when the analytical method must be applied routinely, is its cost. This will be governed mainly by the price of the instrument, columns, and mobile phases, as well as maintenance costs. In this respect, the chromatographic techniques can be ranked in increasing order of the prices of analysis: GC < LC < SFC.

From a practical viewpoint that disregards costs, the choice of method must be based on three fundamental aspects: separation efficiency, accuracy of the analytical method, and total required analysis time. For some applications, the specificity and sensitivity (detectability) of a technique are also important.

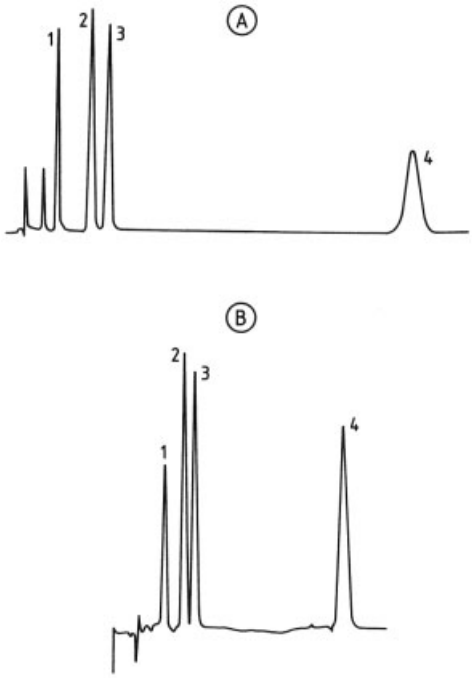

Figure 59. HPLC analysis of steroid esters
A) Column: 10 cm $L \times 0.3$ cm *i.d.*, 5 µm ODS, flow 0.5 mL/min methanol – water 85 : 15, UV at 254 nm; B) see A, light scattering detection
Peaks: (1) testosterone propionate, (2) testosterone isocaproate, (3) testosterone phenylpropionate, (4) testosterone decanoate

A study was conducted in which the performance of three techniques – high-performance LC (HPLC), capillary GC (CGC), and capillary SFC (CSFC) – was compared in the analysis of the percent composition of steroid esters in a pharmaceutical hormone formulation [73]. The study used standard chromatographic conditions, applicable in every laboratory, and the conditions were optimized so as to yield comparable analysis times of <30 min.

High-Performance Liquid Chromatography. In the pharmaceutical industry, analysis of steroid ester formulations is performed routinely by using HPLC. Figure 59 A shows the HPLC chromatogram of the sample mixture applying UV detection at 254 nm. The steroid esters are baseline separated in 24 min. The column plate number, calculated for testosterone decanoate, is 4500, which corresponds to 45 000 plates per meter. With UV detection, allowance must be made for different response factors. Calibration

Table 5. Calculated percent composition of steroid esters ∗

Compound	HPLC		CSFC	CGC on-column	Composition in mixture, %
	UV	MD ∗∗			
Testosterone propionate	15.13	11.93	10.98	11.98	12.00
Testosterone isocaproate	27.77	23.68	23.57	25.20	24.00
Testosterone decanoate	33.25	40.43	44.33	39.99	40.00
Testosterone phenylpropionate	23.38	23.96	21.11	23.83	24.00
Mean RSD, %	1.43	1.08	2.58	0.58	
Analysis time, min	22.50	25.00	21.50	21.50	

∗ All values calculated from uncorrected peak areas with response factors = 1.0; mean values of five experiments.
∗∗ MD = mass detector.

graphs, however, can be recorded only if pure compounds are used, which was not the case in this study. A response factor of 1.0 was, therefore, applied for the various steroid esters, giving the quantitative data listed in Table 5. Very often, in practice, unknown compounds are present in HPLC chromatograms, and quantitation is impossible using UV detection.

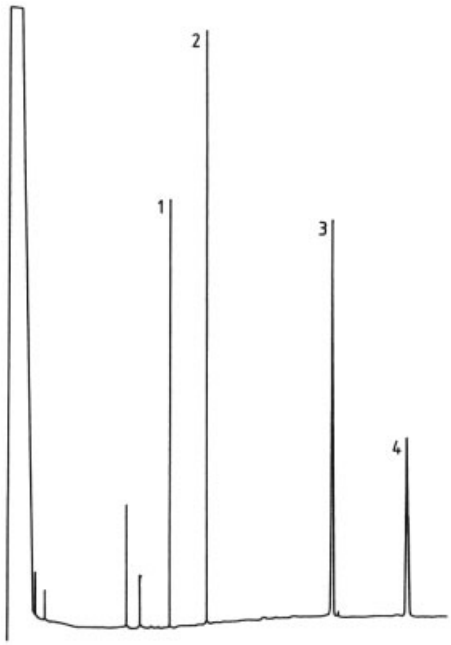

Figure 60. Capillary SFC analysis of steroid esters
Column: 20 m $L \times 0.1$ mm *i.d.*; 0.1 µm d_f MeSi; temperature: 130 °C; pressure 200 bar CO_2; FID
Peaks: (1) testosterone propionate, (2) testosterone isocaproate, (3) testosterone decanoate, (4) testosterone phenylpropionate

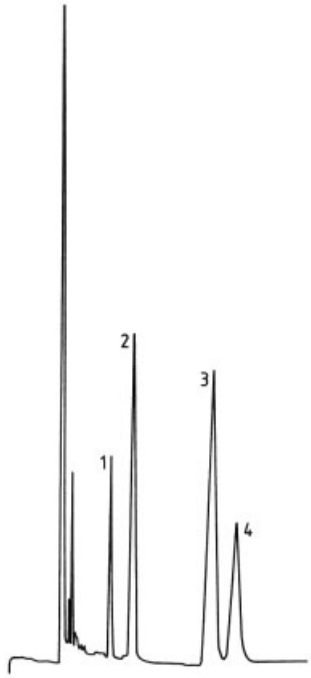

Figure 61. Capillary GC analysis of steroid esters
Column: 25 m $L \times 0.25$ mm *i.d.*; 0.15 µm d_f MePhSi; temperature: 60 – 300 °C at 10 °C/min; cool on-column injection; carrier gas hydrogen 15 psi (≈1 bar); FID
Peaks: see Figure 60

Calibration can be avoided by using universal HPLC detection systems, such as refractive index detection or light-scattering detection. Compared to UV, however, both of these detectors are less sensitive. Figure 59 B shows the analysis using light-scattering detection; the sample is five times more concentrated than in the UV trace in Figure 59 A. The chromatographic efficiency was reduced by 20 %, to 3600 plates, because of a detector contribution to band broadening.

The quantitative data, calculated using response factors of 1, are listed in Table 5 in the column labeled MD (mass detector). The calculated composition corresponds well to the known sample composition.

Capillary Gas Chromatography. Steroid esters are thermally stable. They can be analyzed easily by CGC. Figure 60 shows the analysis of the sample with cold on-column injection. The plate number for this conventional CGC column operated at 130 °C was 85 000 plates; analysis time was 22 min. The quantitative results (response factors = 1.0) are given in Table 5. The data for cold on-column injection correspond very well with the known composition of the sample, and % RSD is < 0.8 % ($n = 5$) for all the steroid esters.

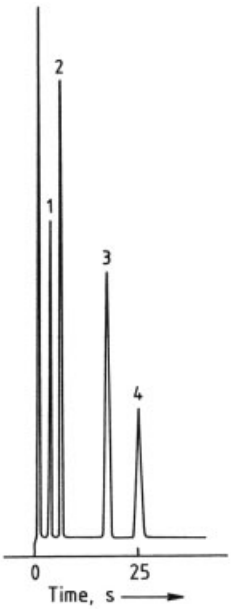

Figure 62. High-speed capillary GC analysis ofsteroid esters
Column: 2 m $L \times 0.25$ mm *i.d.*; 0.1 µm d_f MePhSi; temperature: 260 °C; split injection 1/100; carrier gas hydrogen 0.1 psi (≈ 700 Pa); FID

Capillary Supercritical Fluid Chromatography. In the mid-1980s, considerable progress has been made in CSFC, and the technique is now accessible for routine chromatography. CSFC is claimed to have advantages compared to GC. Because of the high solvating power of CO_2, high molar mass compounds elute at much lower temperature, allowing better analysis of thermally labile compounds. Compared with HPLC, for the same column, the speed of analysis for SFC is roughly four times faster. In addition, because of the lower viscosity of the mobile phase, much longer columns can be used, increasing the effective plate number. Universal GC detectors can be applied, and combination with hyphenated techniques such as mass spectrometry or Fourier transform infrared spectrometry is easier. Analysis of the steroid ester sample on a 20 m × 0.1 mm column is shown in Figure 61. In this run at constant density, the plate number is only 5000. This illustrates one of the disadvantages of CSFC. Theoretically the column should offer 150 000 plates, but at a mobile phase velocity of only 0.2 – 0.3 cm/s. This corresponds to an unretained peak time of 2.5 h, and an analysis time for a compound eluting with a k value of 3.5 (peak 3, testosterone decanoate) of 7.3 h. Such analysis times are, of course, unacceptable. Therefore the column is operated far above the optimal velocity (Figure 61, at 6 cm/s). The slope of the $H - u$ curve is very steep for the 20-m column, and the plate height at 6 cm/s is of the order of 6 mm. Much faster separations can be obtained on short capillary columns with internal diameters of ≈20 µm, but because of technical problems (especially with sample introduction), quantitation is very difficult. Quantitation of the steroid mixture with CSFC (Table 5) yields higher % RSD values and less accurate results than either CGC or HPLC. The reason for this lies mainly in the injection method (valve split discrimination). Compared to conventional HPLC, the primary advantage of SFC is the use of the universal FID detector. The results presented in Table 5 show that all three techniques can be applied to analysis of the steroid sample. The difference in chromatographic behavior among HPLC, CGC, and CSFC is determined mainly by the mobile-phase properties such as density, diffusivity, viscosity, and polarity. When using dimensionless parameters, equations can be derived that are valid for the three different types of mobile phases and independent of particle size (packed columns) or column diameter (capillary columns) [74]. The reduced parameters will not be discussed in detail but show clearly, as far as efficiency and speed of analysis are concerned, no other chromatographic methods can compete with GC. Baseline sepa-

ration of the steroid esters can be obtained on a 2-m capillary column offering 7000 plates in 25 s (Figure 62). This is not a plea for gas chromatography. It is just an example to illustrate the possibilities of different chromatographic techniques for the application selected. Often, however, one might be surprised to see the enormous effort spent in trying a separation with a particular technique, while the use of another technique seems straightforward. Each chromatographic technique has its own possibilities, and a good understanding of their performance is a prerequisite before selection for a particular application can be justified.

10. Recent and Future Developments

10.1. Fast High Resolution Capillary GC

Since the introduction of gas chromatography, there has been an ongoing interest in improving analysis speed. Increasing analysis speed in capillary GC in the first instance is dictated by the problem at hand, the primary objective being the complete separation of the compounds in a mixture. In many cases, the plate number of a capillary column is too high for a given separation problem, and the resolution may be impaired. Typical ways to shorten the analysis time in this case are decreasing the column length or increasing the carrier gas flow rate far above the optimum. Recent alternatives include the use of multichannel columns or multicapillaries and flash GC.

Increasing analysis speed for complex profiles without impairing resolution can only be realized by reduction of the internal diameter and the length of the capillary column. A 10 m × 0.1 mm i.d. column offers the same resolution as a 25 m × 0.25 mm i.d. column. Because the column is 2.5 times shorter, the analysis time is reduced drastically. Moreover, since the optimum carrier gas velocity is higher and the H-u curves are flatter for narrow bore columns, higher average carrier gas velocities can be used without loss of resolution. Presently, capillary columns with internal diameters on the order of 100 µm are in the picture for routine operation because state-of-the-art capillary GC instrumentation allows installation of such small diameter columns. This is illustrated with the analysis of a diesel sample with baseline separation of the biomarkers pristane and phytane from the preceeding normal hydrocarbons on a standard GC instrument (Fig. 63). The diesel sample was diluted 10 times in cyclohexane, and 1 µL was injected in the split mode (1/1000 split ratio). The column was a 10 m × 0.1 mm i.d. × 0.1 µm MeSi. The oven was programmed from 100 to 325 °C at 75 °C/min and held for 1 min (4 min program). FID detection with a signal data acquisition rate of 100 Hz was used.

An important obstacle to the use of narrow bore columns, however, is method development and validation. When using narrowbore columns, different operational conditions (inlet pressure, split ratio, temperature program) have to be used. Since little information is yet available on the use of fast capillary GC, the transfer of standard validated operating procedures developed for conventional capillary columns into operating procedures for narrow bore columns might be difficult and will definitely hamper their use in a routine environment. In this respect, the development of method-translation software [75], [76] was very helpful for translating a standard operating procedure for a conventional column (whatever its dimensions and stationary phase film thickness) to an operating procedure for a narrowbore column (coated with the same stationary phase). After performing the analyses on the standard column, the optimized conditions are introduced into the method-translation program, and all operational conditions for the new column are calculated so as to obtain the same resolution. The gain in analysis time is also predicted. The method-translation software is available free of charge from the Agilent website. The principle is illustrated with the analysis of an essential oil. Quality control of essential oil samples is routinely done on a 50 – 60 m × 0.25 mm i.d. column with a slow (2 °C/min) temperature program, which results in analysis times on the order of 2 – 3 h. The same peak capacity can be obtained on shorter narrowbore capillary columns, but, although analysis times can be reduced drastically, quality control laboratories hesitate to use this approach because changing the column dimensions implies different operational conditions, which results in different selectivities. Details of well-known fingerprints can be lost. The application of the method-translation software

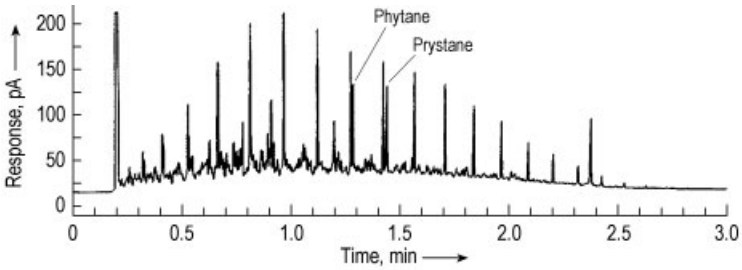

Figure 63. Separation of diesel oil in 2.5 min on standard GC instrumentation

allows the translation from conventional to narrowbore columns with hardly any change in resolution, selectivity, and hence fingerprint. This is illustrated with the analysis of nutmeg oil.

The analysis was first performed on a "standard" column used for detailed essential oil profiling (60 m × 0.25 mm i.d. × 1 µm MeSi). The operational conditions optimized for routine QC were applied. Secondly, the analyses were repeated on a 20 m × 0.1 mm i.d. × 0.4 µm MeSi. The operational conditions for the narrowbore column were calculated by using the method-translation software. The most important operational conditions are summarized in Table 6. From the method-translation software program, a speed gain factor of 5.9 is predicted. Note that the carrier gas in the analyses are helium and hydrogen for the 0.25 mm i.d. and the 0.1 mm i.d. columns, respectively.

The chromatograms obtained on the respective columns are compared in Figure 64 top (nutmeg oil, standard column) and bottom (nutmeg oil, narrow bore column). From these chromatograms, it is obvious that the resolution is exactly the same on both columns. The analysis on the narrowbore column is much faster. For the last-eluting peak, a speed gain factor of 5.7 is measured, which is close to the predicted speed gain of 5.9. Since for the calculation, the nominal column lengths were used, it might be expected that the correlation could be even better if actual column lengths are measured and used in the method-translation software.

10.2. The Concept of Retention Time Locking

Recent developments in capillary GC, i.e., electronic pneumatic control (EPC) of the carrier gas, improved oven temperature stability, and excellent reproducibility in column making have led to the concept of retention time locking (RTL) [77], [78]. With retention time locked data bases, absolute retention times instead of retention indices can be used to elucidate the structures of eluting solutes. Moreover, retention time locking can be used in combination with different injectors and detectors. Exact scaling of capillary GC – FID, capillary GC – MS, and capillary GC – AED chromatograms is feasible.

The concept is illustrated with the analysis of some organochlorine pesticides in wine. Figure 65 shows the capillary GC – AED trace in the chlorine emission line (837 nm) of a wine extract. RTL conditions were adapted to the analysis, and the identities were searched via a pesticide RTL database. A retention time window of 0.5 (± 0.25 min from the measured retention time) and the presence of chloor were introduced in the RTL dialogue box. From the search for the main peak eluting at 21.93 min (Table 7) it is clear that the smallest deviation is obtained for procymidone. The presence of this compound was confirmed by using RTL conditions in capillary GC – MS. The other chlorine compounds, detected by capillary GC – AED analysis, could not be identified directly by capillary GC – MS. Owing to the complexity of the chromatogram no clear spectra were obtained. For these compounds, the result screener was used. Hereby, the TIC is searched for specific ions for all pesticides present in the database (567 pesticides) in their elution window. The peak at 25.64 min in the AED trace was identified as *p,p′*-DDD. The peak at 28.40 min was identified as iprodione. The peak at 29.76 min could not be identified since no ions corresponding to a pesticide present in the database and eluting in this time window could be detected.

Table 6. Experimental conditions for essential oil analysis on a conventional and on a narrowbore column. The conditions in italics for the narrowbore column were calculated by the method-translation software. The standard operating conditions (SOP) on the conventional column apply broader conditions than needed for the essential oil analysed here.

Column	60 m × 0.25 mm i.d. × 1 µm HP-1	20 m × 0.1 mm i.d. × 0.4 µm HP-1
Injection	split, 1 µL, 1/50 split ratio, 250 °C	split, 1 µL, *1/500* split ratio, 250 °C
Carrier type	helium	hydrogen
Carrier pressure	209 kPa, constant pressure	*411 kPa,* constant pressure
Carrier flow	1.74 mL/min at 50 °C	*0.87 mL/min* at 50 °C
Velocity	29.5 cm/s (average at 50 °C)	*58 cm/s* (average at 50 °C)
Hold-up time	3.39 min	*0.57 min*
Oven program	50 °C – 2 °C/min – 275 °C – 40 min	*50 °C – 11.88 °C/min – 275 °C – 7 min*
Detection	FID	FID
Signal data rate	10 Hz	50 Hz
Analysis time	152.5 min	*25.94 min*

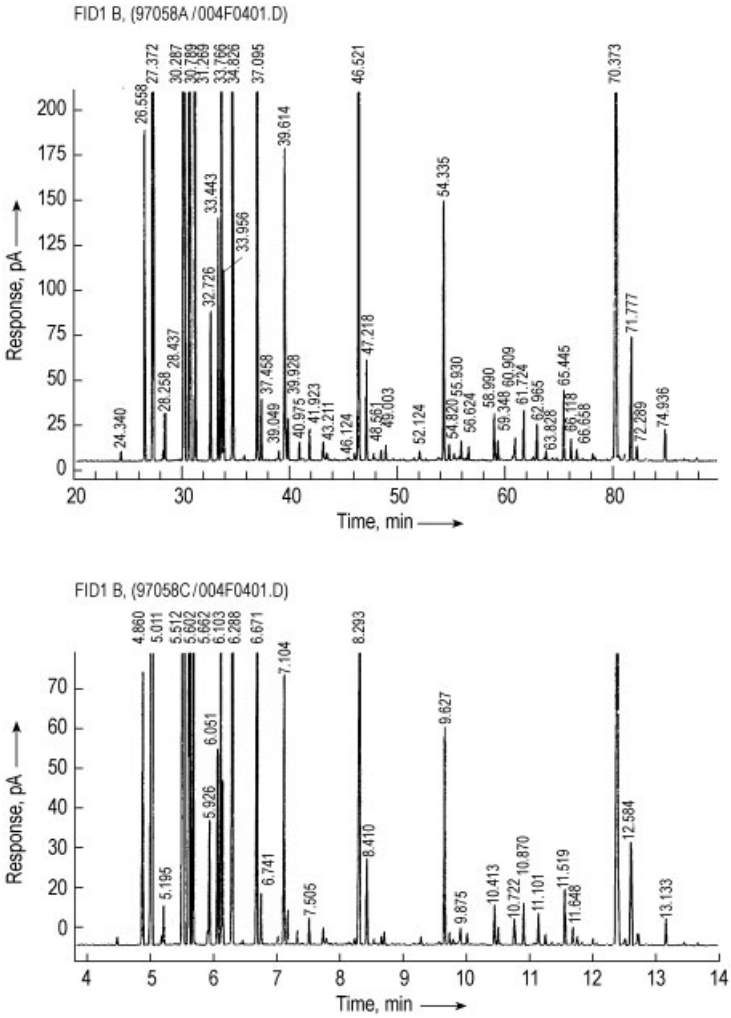

Figure 64. Analyses of nutmeg oil on a 60 m × 0.25 mm i.d. × 1 µm MeSi column (top trace) and on a 20 m × 0.1 mm × 0.4 µm MeSi column (bottom trace)

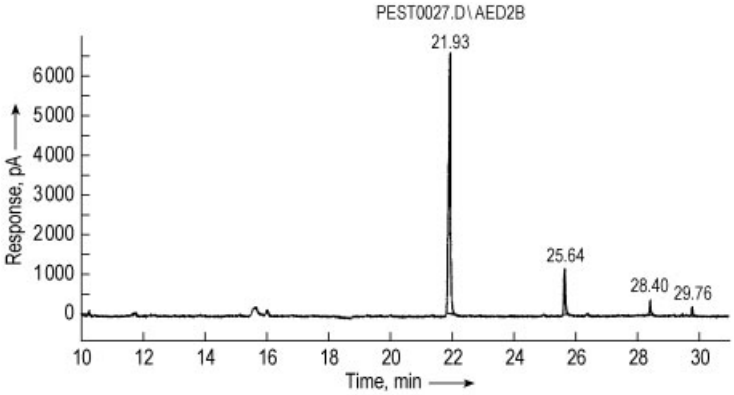

Figure 65. Capillary GC – AED chromatogram at the 837 nm emission line of a white wine extract recorded under RTL conditions

This example clearly demonstrates the complementary nature of capillary GC – AED and capillary GC – MS. The selectivity of the AED results in simple profiles and allows fast screening of pesticides in samples. Capillary GC – MS in combination with the result screener allows confirmation of the detected solutes. Analysis of the samples with capillary GC – MS only results in complex chromatograms and more time consuming data interpretation is needed. Retention time locking also allows the correct matching of the capillary GC – AED chromatograms with the capillary GC – MS chromatograms. The corresponding peaks elute at virtually the same retention times.

10.3. Towards Black Boxes

State-of-the-art GC offers column efficiencies ranging from 5000 to 500 000 plates (packed or short megabore columns to long narrowbore columns), with sample capacities from microgram to subnanogram amounts. Selectivity can be introduced in all parts of the GC system. Column selectivity can be adapted to the specific need by selecting the most suitable *stationary phase*. To enhance resolution of critical pairs, column selectivity can be optimized by *selectivity tuning* between two columns coated with different stationary phases or by *two-dimensional CGC*. *Multidimensional* CGC, HPLC – CGC, SFE – CGC, etc., provide preseparation and enrichment of relevant fractions. In addition to universal inlets a number of *selective inlets* are available in GC. GC enjoys universal detection and sensitive, selective detection. Selective recognition can be performed routinely by CGC – mass spectroscopy, CGC – Fourier transform infrared spectroscopy, and CGC – atomic emission detection, to provide sensitive and selective quantitation of target compounds and structural elucidation or identification of unknowns.

There is much more to come, such as comprehensive capillary GC, hyphenation with time-of-flight MS (capillary GC – TOFMS) and inductively coupled plasma MS (capillary GC – ICP-MS) . . . to mention a few recent technological realizations.

Nothwithstanding all these capabilities of GC, the data are often disappointing, as evidenced by results from round-robin tests. Deviations can be as high as 100 ppm for BTX measurements in air in the 100 ppm level, as high as 30 ppb for trihalomethanes at the 50 ppb level in drinking water, as high as 60 ppt (ppt = 1 part in 10^{12}) for organochloropesticides in the 70 ppt level in wastewater, etc. How should data such as 3.34 ppt 2,3,7,8 TCDD be interpreted?

What are the reasons for this? Two main aspects can be identified. First of all, there is an educational problem. Technicians in the laboratory often lack experience and expertise, and moreover time and money for training are unavailable. On the other hand, shift of good staff now occurs very quickly without time to transfer the know-how. The second aspect is precolumn errors (e.g., sample introduction and sample preparation), and both parts of the total ana-

Table 7. Results of retention time locked pesticide data base search

Search results for 21.939 +/− 0.250 min
Contains elements: Cl
Does not contain elements: {no restriction}
RTT file searched: C:\\HPCHEM\\RTL\\HPGCPST.RTT

FID_RT	Compound name	MW	Formula	Δ *RT* (min)
21.825	Chlorbenside	269.19	C:13,H:10,Cl:2,S:1,	− 0.114
21.971	Procymidone	284.14	C:13,H:11,Cl:2,N:1,O:2,	+ 0.032
22.037	trans-Chlordane	409.78	C:10,1,Cl:6,H:8,	+ 0.098
22.95	Chlorflurecol-Me- ester	274.70	C:15,H:11,Cl:1,O:3,	+ 0.156

Search results for 25.649 +/− 0.250 min
Contains elements: Cl
Does not contain elements: {no restriction}
RTT file searched: C:\\HPCHEM\\RTL\\HPGCPST.RTT

FID_RT	Compound name	MW	Formula	Δ *RT* (min)
25.409	Chlorobenzilate	325.19	C:16,H:14,Cl:2,O:3,	− 0.240
25.434	Chloropropylate	339.22	C:17,H:16,Cl:2,O:3,	− 0.215
25.581	Diniconazole	326.23	C:15,H:17,Cl:2,N:3,O:1,	− 0.068
25.654	Cyprofuram	279.72	C:14,H:14,Cl:1,N:1,O:3,	+ 0.005
25.703	*p,p′*-DDD	320.05	C:14,H:10,Cl:4,	+ 0.054
25.763	Etaconazole	328.20	C:14,H:15,Cl:2,N:3,O:2,	+ 0.114
25.786	*o,p′*-DDT	354.49	C:14,H:9,Cl:5,	+ 0.137
25.823	Flamprop-isopropyl	363.82	C:19,H:19,Cl:1,F:1,N:1,O:3,	+ 0.174

Search results for 28.404 +/− 0.250 min
Contains elements: Cl
Does not contain elements: {no restriction}
RTT file searched: C:\\HPCHEM\\RTL\\HPGCPST.RTT

FID_RT	Compound name	MW	Formula	Δ *RT* (min)
28.239	Endrin ketone	380.91	C:12,H:8,Cl:6,O:1,	− 0.165
28.362	Benzoylprop ethyl	366.24	C:18,H:17,Cl:2,N:1,O:3,	− 0.042
28.424	Iprodione	330.17	C:13,H:13,Cl:2,N:3,O:3,	+ 0.020
28.436	Dichlorophen	269.13	C:13,H:10,Cl:2,O:2,	+ 0.032
28.545	Leptophos oxon	396.00	C:13,H:10,Br:1,Cl:2,O:3,P:1,	+ 0.141
28.572	Chlorthiophos sulfoxide	377.24	C:11,H:15,Cl:2,O:4,P:1,S:2,	+ 0.188

Search results for 29.768 +/− 0.250 min
Contains elements: Cl
Does not contain elements: {no restriction}
RTT file searched: C:\\HPCHEM\\RTL\\HPGCPST.RTT

FID_RT	Compound name	MW	Formula	Δ *RT* (min)
29.699	Phosalone	367.80	C:12,H:15,Cl:1,N:1,O:4,P:1,S:2,	− 0.069
29.754	Leptophos	412.06	C:13,H:10,Br:1,Cl:2,O:2,P:1,S:1,	− 0.014
29.848	Mirex	545.55	C:10,Cl:12,	+ 0.080

lytical procedure are interrelated. Future developments in GC are therefore expected in the direction of fully automated sample preparation and sample introduction systems. Some instrumentation nowadays is already tailor-made for a given application. More can be expected in the future and development may result in specific instruments, such as a pesticide analyzer, a PCB analyzer, or a sugar analyzer.

11. References

General textbooks on chromatography and gas chromatography

1. J. C. Giddings: *Unified Separation Science,* J. Wiley & Sons, New York 1991.
2. C. F. Poole, S. K. Poole: *Chromatography Today,* Elsevier, Amsterdam 1991.
3. W. Jennings: *Analytical Gas Chromatography,* Academic Press, San Diego 1987.
4. G. Schomburg: *Gas Chromatography, A Pratical Course,* VCH Verlagsgesellschaft, Weinheim 1990.

5. D. Rood: *A Practical Guide to the Care, Maintenance and Troubleshooting of Capillary Gas Chromatographic Systems,* Hüthig Verlag, Heidelberg 1991.
6. M. L. Lee, F. J. Yang, K. D. Bartle: *Open Tubular Column Gas Chromatography, Theory and Practice,* J. Wiley & Sons, New York 1984.
7. A. Van Es: *High Speed, Narrow-Bore Capillary Gas Chromatography,* Hüthig Verlag, Heidelberg 1992.

Specific References

8. A. T. James, A. J. P. Martin, *Biochem. J.* **50** (1952) 679.
9. L. S. Ettre, *Pure Appl. Chem.* **65** (1993) no. 4, 819.
10. J. J. Van Deemter, F. J. Zuiderberg, A. Klinkenberg, *Chem. Eng. Sci.* **5** (1956) 271.
11. M. J. E. Golay in D. Detsy (ed.) : *Gas Chromatography,* Butterworths, London 1958, p. 36.
12. P. Sandra, *J. High Resolut. Chromatogr.* **12** (1989) 82.
13. P. Sandra, *J. High Resolut. Chromatogr.* **12** (1989) 273.
14. R. D. Dandeneau, E. H. Zerenner, *J. High Resolut. Chromatogr. Comm.* **2** (1979) 351.
15. M. Verzele, P. Sandra, *J. Chromatogr.* **158** (1978) 111.
16. P. Sandra, *LC.GC* **5** (1987) 236.
17. H. Rotzsche: *Stationary Phases in Gas Chromatography,* Elsevier, Amsterdam 1991.
18. W. O. McReynolds, *J. Chromatogr. Sci.* **8** (1970) 685.
19. J. de Zeeuw, R. de Nijs, L. Henrich, *J. Chromatogr. Sci.* **25** (1987) 71.
20. J. R. Mann, S. T. Preston, *J. Chromatogr. Sci.* **11** (1973) 216.
21. T. J. Stark, P. A. Larson, R. D. Dandaneau, *J. Chromatogr.* **279** (1983) 31.
22. P. Sandra et al., *J. High Resolut. Chromatogr. Chromatogr. Commun.* **8** (1985) 782.
23. P. Sandra, F. David in H. J. Cortes (ed.) : *Multidimensional Chromatography, Techniques and Applications,* Marcel Dekker, New York 1990, p. 145.
24. K. E. Markides et al., *Anal. Chem.* **57** (1985) 1296.
25. W. A. König: *The Practice of Enantiomer Separation by Capillary Gas Chromatography,* Hüthig Verlag, Heidelberg 1987.
26. W. A. König: *Gas Chromatographic Enantiomer Separation with Modified Cyclodextrins,* Hüthig Verlag, Heidelberg 1992.
27. D. E. Bautz, J. W. Dolan, W. D. Raddatz, L. R. Snyder, *Anal. Chem.* **62** (1990) 1560.
28. G. N. Abbay et al., *LC.GC* Int. 4 (1991) 28.
29. E. Sippola, F. David, P. Sandra, *J. High Resolut. Chromatogr.* **16** (1993) 95.
30. B. W. Hermann et al., *J. High Resolut. Chromatogr.* **13** (1990) 361.
31. S. S. Stafford (ed.) : *Electronic Pressure Control in Gas Chromatography,* Hewlett-Packard, Little Falls 1993.
32. V. Pretorius, W. Bertsch, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **6** (1983) 64.
33. R. Jenkins, W. Jennings, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **6** (1983) 228.
34. P. Sandra in P. Sandra (ed.) : *Sample Introduction in Capillary Gas Chromatography,* Hüthig Verlag, Heidelberg 1985, p. 1.
35. M. S. Klee: *GC Inlets, An Introduction,* Hewlett-Packard, Little Falls 1990.
36. K. Grob: *Classical Split and Splitless Injection in Capillary GC,* Hüthig Verlag, Heidelberg 1986.
37. K. Grob: *On-Column Injection in Capillary GC,* Hüthig Verlag, Heidelberg 1987.
38. P. Sandra in K. Hyver (ed.) : *High Resolution Gas Chromatography,* 3rd ed., Hewlett-Packard, Little Falls 1987.
39. D. Desty, A. Goldup, H. Whyman, *J. Inst. Pet.* **45** (1959) 287.
40. W. D. Snyder, Technical Paper no. 108, Hewlett-Packard, Little Falls 1987.
41. P. Wyllie, K. J. Klein, M. Q. Thompson, B. W. Hermann, *J. High Resolut. Chromatogr.* **14** (1991) 361.
42. F. David, P. Sandra, S. S. Stafford, B. Slavica, Hewlett-Packard Application Note 228 – 223, Little Falls 1993.
43. A. Zlatkis, H. R. Kaufman, *Nature (London)* **184** (1959) 4010.
44. P. Sandra in F. Bruner (ed.) : *The Science of Chromatography,* Elsevier, Amsterdam 1985, p. 381.
45. K. Grob, K. Grob Jr., *J. Chromatogr.* **151** (1978) 311.
46. K. Grob, *J. High Resolut. Chromatogr. Chromatogr. Commun.* **1** (1978) 263.
47. G. Schomburg et al., *J. Chromatogr.* **142** (1977) 87.
48. P. Sandra, M. Van Roelenbosch, M. Verzele, C. Bicchi, *J. Chromatogr.* **279** (1983) 287.

49. F. David, L. Vanderroost, P. Sandra, S. Stafford, *Int. Labmate* **17** (1992) no. 3, 13.
50. E. Meyer, A. P. De Leenheer, P. Sandra, *J. High Resolut. Chromatogr.* **15** (1992) 637.
51. W. Vogt, K. Jacob, H. W. Obwexer, *J. Chromatogr.* **174** (1979) 437.
52. G. Schomburg in R. E. Kaiser (ed.), *Proc. Int. Symp. Capillary Chromatogr. 4th* 1981, 371, 921 A.
53. F. Poy, S. Visani, F. Terrosi, *J. Chromatogr.* **217** (1981) 81.
54. F. David, P. Sandra, A. Hoffmann, J. Gerstel, *Chromatographia* **34** (1992) no. 5 – 8, 259.
55. R. Buffington, M. K. Wilson: *Detectors for Gas Chromatography, A Practical Primer,* Hewlett-Packard, Little Falls 1987.
56. H. H. Hill, D. G. McMinn (eds.) : *Detectors for Capillary Chromatography,* J. Wiley & Sons, New York 1992.
57. I. Temmermann, F. David, P. Sandra, R. Soniassy, Hewlett-Packard Co., Application Note 228 – 135, Little Falls 1991.
58. F. W. Karasek, R. E. Clement: *Basic Gas Chromatography – Mass Spectroscopy,* Elsevier, Amsterdam 1988.
59. W. Herres: *Capillary Gas Chromatography – Fourier Transform Infrared Spectroscopy,* Hüthig Verlag, Heidelberg 1987.
60. R. Buffington: *GC-Atomic Emission Spectroscopy Using Microwave Plasmas,* Hewlett-Packard, Little Falls 1988.
61. E. Kovats, J. C. Giddings, R. A. Keller in: *Advances in Chromatography,* vol. 1, Marcel Dekker, New York 1965, chap. 7.
62. H. Van den Dool, P. D. Kratz, *J. Chromatogr.* **11** (1963) 463.
63. W. Jennings, T. Shibamoto: *Qualitative Analysis of Flavor and Fragrance Volatiles by Gas Chromatography,* Academic Press, New York 1980.
64. J. F. Sprouse, A. Varano, *Am. Lab. (Fairfield, Conn.)* **16** (1984) 54.
65. H. J. Cortes (ed.) : *Multidimensional Chromatography, Techniques and Applications,* Marcel Dekker, New York 1990, p. 145.
66. C. Biichi, A. D'Amato, F. David, P. Sandra, *J. High Resolut. Chromatogr.* **12** (1989) 316.
67. K. Grob: *On-Line Coupled LC-GC,* Hüthig Verlag, Heidelberg 1991.
68. F. David, P. Sandra, D. Bremer, R. Bremer, F. Rogies, A. Hoffmann, *Labor Praxis* **21** (1997) 5.
69. M. Verschuere, P. Sandra, F. David, *J. Chromatogr. Sci.* **30** (1992) 388.
70. F. David, M. Verschuere, P. Sandra, *Fresenius J. Anal. Chem.* **344** (1992) 479.
71. K. Blau, G. S. King: *Handbook of Derivatization for Chromatography,* Heyden (Wiley & Sons), New York 1977.
72. J. Drozd: *Chemical Derivatization in Gas Chromatography,* Elsevier, Amsterdam 1981.
73. P. Sandra, F. David, *LC.GC* **7** (1989) 746.
74. L. S. Ettre, J. V. Hinshaw: *Basic Relationships of Gas Chromatography,* Advanstar, Cleveland 1993.
75. B. D. Quimby, V. Giarrocco, M. S. Klee, Hewlett-Packard Application Note 228-294, February 1995.
76. F. David, D. R. Gere, F. Scanlan, P. Sandra, *J. Chromatogr. A* **842** (1999) 309.
77. V. Giarrocco, B. D. Quimby, M. S. Klee, Hewlett-Packard Application Note 228-392, December 1997.
78. B. D. Quimby, L. M. Blumberg, M. S. Klee, P. L. Wylie, Hewlett-Packard Application Note 228-401, May 1998.

Gas Cleaning → Gas Production
Gas Cleaning → Natural Gas

Liquid Chromatography

PETER LEMBKE, Belovo-Spain, S.L., Reus, Spain (Chaps. 1, 3 – 10, Sections 2.1 – 2.6.7)

GÜNTER HENZE, Institut für Anorganische und Analytische Chemie der Technischen Universität Clausthal (Section 2.6.8)

KARIN CABRERA, Merck KGaA, Darmstadt, Federal Republic of Germany (Chap. 11)

WOLFGANG BRÜNNER, formerly Merck KGaA, Darmstadt, Federal Republic of Germany (Chap. 11)

EGBERT MÜLLER, Merck KGaA, Darmstadt, Federal Republic of Germany (Chap. 11)

1. General 3
1.1. History 3
1.2. Definition and Theoretical Background 4
2. Equipment 6
2.1. Filters and Connecting Tubing . 6
2.2. Containment of Mobile Phase . 7
2.3. Pumps 7
2.4. Sample Introduction Units 8
2.4.1. Syringe Injectors 8
2.4.2. Sample Valve Injectors 8
2.4.3. Automatic Introduction Systems . 9
2.5. Columns 9
2.6. Detectors 9
2.6.1. Concentration-Sensitive and Mass-Flow-Sensitive Detectors 11
2.6.2. Differential Refractive Index (RI) Detector 11
2.6.3. UV – VIS Absorption Detector . . 11
2.6.4. Diode-Array Detector 12
2.6.5. Fluorescence Detector 13
2.6.6. Evaporative Light-Scattering Detector 14
2.6.7. Infrared Detectors – Fourier Transform Infrared Spectroscopy 14
2.6.8. Electrochemical Detection 14
3. Solvents (Mobile Phase) 24
3.1. Purification and Pretreatment of LC Solvents 24
3.2. Elutropic Order 24
3.3. Selection of Mobile Phase 25
4. Column Packing (Stationary Phase) 26
4.1. Silica 27
4.2. Alumina 27
4.3. Chemically Bonded Phases . . . 27
4.4. Column Packing Techniques . . 29
4.5. Column Specifications and Column Diagnosis 29
5. Separation Processes 30
5.1. Adsorption Chromatography . . 30
5.1.1. Polar Adsorbents and Their Properties 30
5.1.2. Polar Chemically Bonded Phases . 31
5.1.3. Nonpolar Adsorbents – Reversed Phase 31
5.1.4. Other Adsorbents 32
5.2. Partition Chromatography (LLC) 33
5.3. Ion-Exchange Chromatography (IEX) 34
5.4. Size Exclusion Chromatography 36
6. Gradient Elution Technique . . . 39
7. Quantitative Analysis 40
7.1. External Standard Method . . . 40
7.2. Internal Standard Method 41
7.3. Error Sources in HPLC 41
8. Sample Preparation and Derivatization 42
8.1. Sample Cleanup and Enrichment 42
8.2. Pre- and Postcolumn Derivatization 43
8.2.1. Precolumn Derivatization 43
8.2.2. Postcolumn Derivatization 45
9. Coupling Techniques 47
9.1. Column Switching 47
9.1.1. Front-, Heart-, and End-Cut Techniques 48
9.1.2. On-Line Sample Cleanup and Trace Enrichment 48
9.1.3. Backflush Technique 48
9.1.4. Multidimensional HPLC 49
9.1.5. Working with Precolumns 49
9.2. HPLC – Mass Spectrometry . . . 49
10. Supercritical Fluid Chromatography 50
10.1. History 50
10.2. Theoretical Background 51
10.3. SFC Compared to GC and LC . 52
10.4. Equipment 53
10.4.1. Pumps 54
10.4.2. Oven 54

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_237

10.4.3. Sample Introduction 54
10.4.4. Detectors 55
10.4.5. Restrictors 56
10.5. Packed-Column SFC (pc-SFC) and Capillary SFC (c-SFC) . . . 56
10.5.1. Capillary SFC (c-SFC) 56
10.5.2. Packed-Column SFC (pc-SFC) . . 57
10.6. Applications 58
11. Affinity Chromatography 59
11.1. Preparation of an Affinity Matrix 60
11.1.1. Selection of a Carrier 60
11.1.2. Immobilization of the Ligand . . . 61
11.2. Ligands 62
11.2.1. Group-Specific Ligands 63
11.2.1.1. Low-Molecular Ligands 63
11.2.1.2. Macromolecular Ligands 64
11.2.2. Monospecific Ligands 64
11.3. Elution 65
12. References 66

Abbreviations

2′,5′-ADP	adenosine 2′,5′-diphosphate
5′-AMP	adenosine 5′-monophosphate
A_{ref}	area or height of the reference peak in the standard solution
$A_{\mathrm{s}i}$	area or height of component i in the sample
AUFS	absorption units full scale
C_{C}	cell capacitance
C_{D}	double-layer capacitance
CRD	chemical reaction detector
DAD	diode-array detector
D_{m}	diffusion coefficient
dp	particle size
E	molecular extinction coefficient
ECD	electron capture detector
ELSD	evaporative light-scattering detector
ES	electrospray
F	flow rate
FAB	fast ion bombardment
FD	field desorption
FID	flame ionization detector
FSD	fast-scanning detector
FTIR	Fourier transform IR spectroscopy
GFC	gel filtration chromatography
GPC	gel permeation chromatography
H	plate height, band height
HPIC	high-performance ion chromatography
HPSEC	high-performance size exclusion chromatography
$I*$	radiation
I_0/I	transmittance
IEC	ion exclusion chromatography
IEX	ion-exchange chromatography
IMAC	immobilized metal ion affinity chromatography
ISE	ion-selective electrode
K	distribution coefficient
k'	capacity ratio
K_{ij}^{Pot}	selectivity factor
L	length
LC	liquid chromatography
LIF	laser-induced fluorescence
LLC	liquid – liquid chromatography
LSC	liquid – solid chromatography
LSD	light-scattering detector
MSD	mass-selective detector
MWD	multiwavelength detector
N	plate number
NAD	nicotinamide – adenine dinucleotide
NADP	nicotinamide – adenine dinucleotide phosphate
NARP	nonaqueous reversed phase
NP	normal phase
PC	paper chromatography
RI	refractive index
RIG	refractive index gradient
RIU	refractive index units
RP	reversed phase
R_{S}	electrolyte resistance
Rs	resolution
SAX	strong anion exchanger
SCX	strong cation exchanger
SEC	size exclusion chromatography
SFC	supercritical fluid chromatography
SFE	supercritical fluid extraction
t_0	time analyte spends in the mobile phase
TID	thermoionic detector
TOF	time of flight
t_{R}	retention time
$t_{\mathrm{R}'}$	time analyte spends in the stationary phase
u	velocity
W	peak width
WAX	weak anion exchanger

WCX	weak cation exchanger
WE	working electrode
z_i	atomic number of ion *i*
z_j	atomic number of interfering ion *j*
α	selectivity
Γ	correction factor
λ	correction factor, wavelength
$\varkappa$	conductivity
ω	angular frequency, correction factor

1. General

Liquid chromatography is a separation technique based on a different distribution rate of sample components between a stationary and a liquid mobile phase. Depending on the stationary phase, a distinction is made between thin layer chromatography (TLC), paper chromatography (PC), and liquid column chromatography (LC). Liquid column chromatography can be divided into open and closed systems. Generally an *open system* consists of a glass column packed with large-sized (≈30 µm and larger) particles of the stationary phase onto which the mobile phase is poured. Because of the large particles, the mobile phase can pass the stationary phase only by means of gravity, sometimes assisted by suction. Therefore, no solvent pump is necessary.

However, to increase separation efficiency, high velocities and small-diameter (≤10 µm) particles are essential as the stationary phase. Unfortunately, this leads to an increase in back-pressure. To overcome this back-pressure, a pump is necessary for transport of the mobile phase (high-pressure- or high-performance liquid chromatography, HPLC). Separation is performed in a *closed system.*

HPLC is the LC technique with the highest efficiency (ability to separate different sample compounds in a given time with a given resolution). The selectivity of this technique is based not only on the different types of stationary phases but also on the mobile phase. An advantage of gas chromatography (GC) is that selectivity in this technique is based only on the stationary phase. (In GC, interactions occur only between the solute and the stationary phase. In LC, additional interactions between the solute and the liquid mobile phase are found.) Furthermore GC can be used only for samples that can be volatized without decomposition at temperatures up to ca. 400 °C. This limits the use of GC to about 20 % of the samples of interest [1]. The limit of LC is set by the solubility of the sample in the mobile phase. Since nearly every substance is soluble in some kind of a solvent, the application range of LC is very broad. In contrast, capillary GC (→ Gas Chromatography, Chap. 3.3.2.) has a far higher separation efficiency than HPLC.

Neither technique can be said to be better than the other. Both GC and HPLC have their advantages for particular separation problems.

In LC, two sample introduction techniques can be distinguished the continuous and the discontinuous mode. *Continuous sample introduction*, also known as front analysis or adsorptive filtration, is often used for solvent cleanup. In this technique the sample solution is introduced continuously onto the column; thus only the least-retained compound is obtained in the pure state.

Discontinuous Sample Introduction. Discontinuous sample introduction can divided into elution analysis and displacement analysis. In *elution analysis*, the sample is injected into a continuous mobile phase. If the composition of the mobile phase remains unchanged during the analysis this is called the *isocratic mode.* In *gradient elution* the composition of the mobile phase is changed in such a way that the elution strength is increased (or sometimes also decreased) during analysis. In the discontinuous sample introduction mode, an elution diagram (chromatogram) is obtained in which, ideally, each compound is separated by a pure solvent zone from the next eluting compound.

*Displacement analysis*is often used in preparative chromatography. After sample injection, the composition of the mobile phase is changed, and the interaction between the mobile and the stationary phases becomes so dominant that the retained sample compounds are displaced by solvent molecules and move with identical velocity (i.e., that of the eluent).

1.1. History

In 1905, Tswett, a Russian botanist, published the first paper on liquid chromatography [2], in

which he described the separation of plant pigments by means of open-column liquid chromatography. The importance of this new technique was recognized only after KUHN (Nobel prize, 1938) and LEDERER used it for the preparative separation of α- and β- carotin from carrots in 1931 [3,4]. In 1941 MARTIN and SYNGE [5] (Nobel prize, 1951) described the technique of partition chromatography (liquid – liquid chromatography, LLC), and in 1949 SPEDDING [6] and TOMPKINS [7] published the first paper on ion-exchange chromatography (IEX). In 1959, PORATH and FLODIN [8] introduced size exclusion chromatography (SEC). One of the first papers on high-performance liquid chromatography was published by PIEL [9] in 1966. Since then, HPLC has become one of the most frequently applied analytical separation methods. Especially because of the rapid development of instrumentation (pumps, detectors, stationary phase, gradient elution, etc.) and column packings during the 1970s and 1980s, very high selectivity and efficiency can be obtained. Today, HPLC is one of the most versatile techniques in the analytical (e.g., trace analysis) and preparative fields.

1.2. Definition and Theoretical Background (→Basic Principles of Chromatography)

In a chromatographic column, two processes occur simultaneously:

1) The separation process, which is based mainly on the interactions between solute and stationary phase
2) Dilution of the solute in the mobile phase due to diffusion processes; this effect counteracts separation

To understand these effects, several important chromatographic definitions are discussed in this section. Figure 1 shows a typical chromatogram with the most important parameters characterizing a chromatographic separation.

The *retention time* t_R of a particular component is defined by:

$$t_R = t_0 + t_{R'}$$

where t_0 is the time the analyte spends in the mobile phase and $t_{R'}$ is the time the analyte spends in the stationary phase. For given HPLC equipment and a constant flow rate (milliliters per minute), all analytes reside in the mobile phase for the same time t_0. They differ only in the time $t_{R'}$ they spend in the stationary phase. To become independent of column dimension and mobile phase flow the *capacity ratio* k' was introduced:

$$k' = \frac{t_R - t_0}{t_0} = \frac{t_{R'}}{t_0} \frac{\text{time spent in stationary phase}}{\text{time spent in mobile phase}}$$

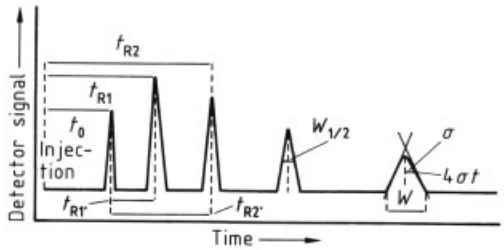

Figure 1. Characteristic information obtainable from a chromatogram

Another important chromatographic quantity is the relative retention or selectivity of a chromatographic system. The *selectivity* α is defined by equation:

$$\alpha = \frac{t_{R2} - t_0}{t_{R1} - t_0} = \frac{k_{2'}}{k_{1'}} = \frac{t_{R2'}}{t_{R1'}}$$

This represents the ability of a given chromatographic system to separate two components. If $\alpha = 1$, no separation is possible, no matter how high the separation efficiency may be. As in GC, the selectivity can be influenced by the choice of the stationary phase, but in LC the selection of the mobile phase is also a factor. In preparative chromatography the selectivity α is the most important chromatographic parameter to be optimized.

A further quantitative measure for determining the separation of two components in a given sample is the *resolution Rs.*

$$Rs = \frac{2\,(t_{R2} - t_{R1})}{(W_1 + W_2)}$$

where W is the peak width. A resolution of $Rs = 1.5$ (baseline resolution) is ideal for quantification of particular sample peaks. Yet even

with a resolution of $Rs = 1.0$, quantification is possible because only ca. 2 % of the peak areas overlap. Chromatograms with resolutions of less than 1.0 should not be used for quantitative analysis. On the other hand, an $Rs > 1.5$ is not necessary, since no advantage is achieved in quantification, but analysis time is prolonged. The above values for Rs are valid only for Gaussian peaks having the same size. Nevertheless, they can assist and guide the chromatographer.

Other important parameters in HPLC are the plate height H and the plate number N. *Plate height* is calculated from a chromatogram by

$$H = \frac{L}{5.54}\left(\frac{W_{0.5}}{t_{\mathrm{R}}}\right)^2$$

where $W_{0.5}$ is the peak width at half maximum peak height; t_{R} the retention time; and L the length of the column.

Well-packed HPLC columns should have a plate height of $2-4\,dp$ (dp = stationary phase particle size).

The *plate number* is an indicator of column quality. The higher it is, the better is the peak capacity (number of peaks separated in a given time) of the column. It is calculated by:

$$N = \frac{L}{H}$$

The plate number can be increased either by reducing the particle size of the stationary phase dp and thus the achievable plate height or by lengthening the column. Both have the disadvantage of increasing the pressure drop over the column as well. Unfortunately, the equipment in most HPLC pumps, does not work at back-pressures exceeding 40 – 50 MPa ("high-pressure shutdown").

The relationship among most of the aforementioned chromatographic parameters is given by the following equation:

$$Rs = \frac{1}{4}\,\frac{\alpha-1}{\alpha}\,\frac{k'_2}{1+k'_2}\sqrt{N}$$

This is the most important equation in LC because it combines the chromatographic separation parameters selectivity, capacity ratio, and plate number. Since these three terms are more or less independent, one term can be optimized after the other. The selectivity term can be optimized by changing the stationary or the mobile phase; k' can be influenced by the solvent strength of the mobile phase, and N can be adjusted by changing the column length, particle diameter, or solvent linear velocity u.

$$u = \frac{L}{t_0}$$

The different processes contributing to the overall band broadening are described by the Van Deemter equation [10]

$$H = A + B/u + Cu$$

where A is the eddy diffusion term ($\approx 2\,\lambda dp$), B the longitudinal (axial) diffusion term ($\approx 2\,\gamma D_{\mathrm{m}}$), C the mass transfer resistance term ($\approx \omega dp^2/D_{\mathrm{m}}$), and u the linear mobile phase velocity.

$$H = 2\lambda dp + 2\gamma D_{\mathrm{m}}/u + \left(dp^2\omega/D_{\mathrm{m}}\right)u$$

where dp is the mean particle diameter; D_{m} the diffusion coefficient in the mobile phase, and λ, γ, ω are correction factors.

For liquid chromatography the following conclusions can be drawn from the Van Deemter equation: In contrast to gas chromatography, the C term (which increases tremendously with the linear velocity of the mobile phase) is very important in liquid chromatography because the diffusion coefficients in the fluid phase are much lower than in GC. A good HPLC column should have as low a C as possible (i.e., a small slope in the H versus u curve; see Fig. 2). In LC the B term is not as important as in GC because of the low diffusion coefficients. The B term decreases with increasing linear velocity u. The A term is independent of linear velocity; it depends only on particle size dp. Particle size is very important in LC (the A and C terms). The smaller the particle size of the stationary phase, the smaller is the achievable plate height and the higher is the plate number of a particular column. The development of uniform small particles for stationary phases in LC is one reason for the booming development of HPLC since the 1970s. The smallest particle size currently available is 1.5 µm (standard sizes: 5, 7, and 10 µm). Particles < 3 µm are not so popular because the back-pressure of the column increases dramatically with decreasing particle diameter, while packing the column becomes very difficult. As a rule, reducing the particle diameter by

half increases the back-pressure of the column to the second power whereas the resolution *Rs* improves by only a factor of 1.4. Nevertheless, short columns (2 cm) packed with 1.5 µm material do allow extremely fast analysis combined with low consumption of mobile phase and can be very useful for laboratories that have to deal with a large number of samples every day. Because of their nonporous nature, these small particles have a very low capacity, which may occasionally result in column-overload problems (peak tailing, poor resolution, etc.).

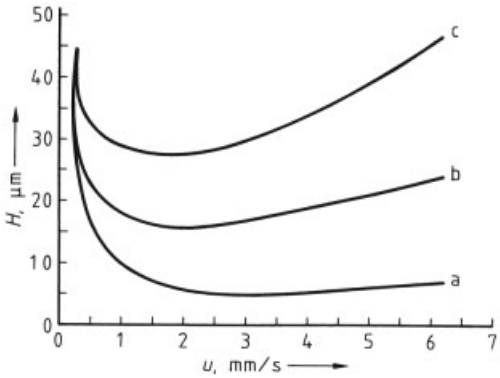

Figure 2. Typical Van Deemter *H/u* curves
a) 3 µm; b) 5 µm; c) 10 µm

2. Equipment

To obtain separations sufficient for quantitative analysis in liquid chromatography, good equipment is essential. Only the most important parameters are discussed in this chapter. For further information, see [11–15] and other monographs on LC. Figure 3 shows a schematic of a simple HPLC system.

2.1. Filters and Connecting Tubing

The connecting tubing between solvent reservoir and pump is usually made of polytetrafluoroethylene (PTFE) with an inner diameter i.d. of 2 – 4 mm. The end of each pump – inlet tube should be fitted with an inlet-line filter (frit), which keeps out large, inadvertent contaminants and holds the inlet line at the bottom of the reservoir so that no air (bubbles) is drawn into the pump head. These frits, which are normally made of stainless steel (but also of glass or PTFE, usually have a pore diameter of 5 – 10 µm and can be cleaned easily with 5 % nitric acid.

The tubing connecting the sample introduction unit, column, and detector consists of small-volume stainless steel capillaries with an inside diameter of 0.25 mm and an outer diameter of 1.25 mm (1/16 inch). For bioseparations that are sensitive to stainless steel, tubing made from polyetheretherketone (PEEK) is recommended. All connecting tubing, especially between sample introduction system and column, should have a small volume (small inside diameter and shortest length possible), to minimize extra column band broadening and loss in separation efficiency.

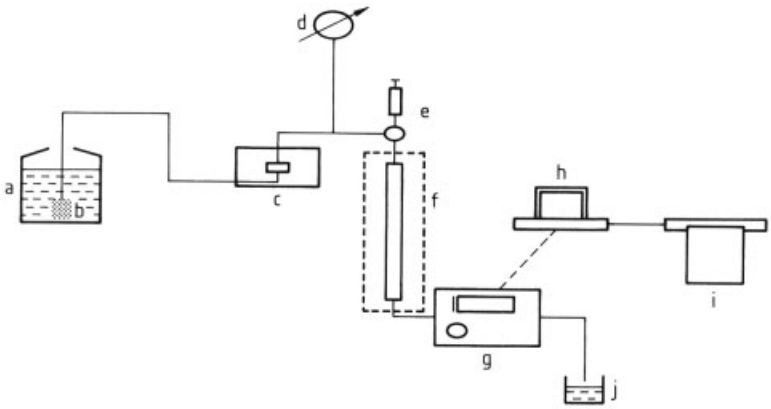

Figure 3. Diagram of a simple HPLC system
a) Solvent reservoir; b) Frit; c) Pump; d) Manometer; e) Injection port; f) Separation column with optional oven; g) Detector; h) Data processing unit; i) Integrator; j) Waste

2.2. Containment of Mobile Phase

Depending on the physicochemical properties of the mobile phase (eluent), different containers are used. The most common are brown laboratory bottles for the protection of light-sensitive eluents. If inorganic ions are to be monitored, polyethylene containers should be used because additional inorganic ions may be dessolved from the glass, thus leading to false results. Safety regulations concerning toxic solvents, fire prevention, etc., must also be taken into account [16].

2.3. Pumps

The pump is one of the most important parts of an HPLC system. It is responsible for constant flow of the mobile phase, which is necessary for the quantification and reproducibility of an analysis. The retention time t_R can be correlated with the retention volume V_R and the flow rate F:

$$V_R = F t_R$$

Therefore a change in the flow rate or strong pulsation changes the retention time of an analyte, making identification by t_R impossible, because the results are not reproducible. Two types of pump pulsation occur that should be kept as low as possible: The *long-time pulsation* (causes irreproducible retention times) and the *short-time pulsation* (causes irreproducible peak areas). The effects of flow instabilities due to pump pulsation and their measurement are discussed in [17].

Good HPLC pumps should be able to deliver a flow rate of 0.5 – 10 mL/min, with a maximum pressure of ca. 40 – 50 MPa, a constant flow rate with variation of $< 0.5\,\%$ (see also [18]), and a rest pulsation of $< 0.1\,\%$. In addition, they should be equipped with adjustable high- and low-pressure limits to avoid damage to the pump or column; should have a solvent compressibility control; must be easy to maintain; and should be chemically inert. The pump heads should not warm up (evaporation of solvent and bubbles), and the pumps should, preferably, not be too noisy.

Basically, two types of pumps exist: Constant-pressure pumps and constant-flow pumps. Pressure- constant pumps are often used to pack HPLC columns.

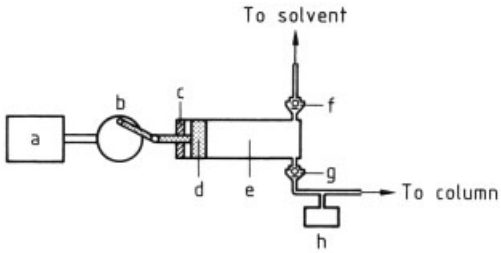

Figure 4. Diagram of a simple reciprocating pump [19]
a) Motor; b) Camshaft; c) Seal; d) Piston; e) Chamber; f) Inlet check valve; g) Outlet check valve; h) Damper

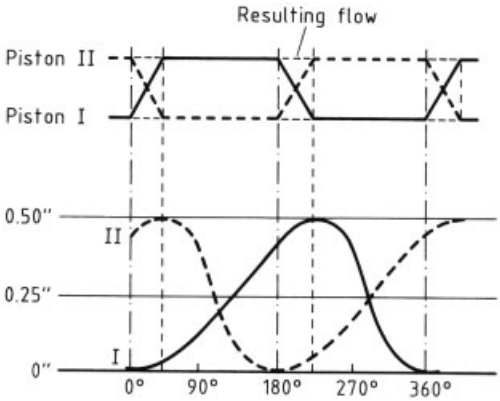

Figure 5. Flow profile of a double-head reciprocating pump (phase shift = 180°)
Courtesy of Laboratory Control/Milton Roy

Modern HPLC systems work with pumps that deliver a constant flow independent of the back-pressure of the column. These *constant-flow pumps* can be divided into continuous and discontinuous systems. In principle, discontinuous pumps work like a huge syringe and have lost their importance in HPLC. Currently, most HPLC pumps work in the continuous mode to deliver a constant flow. In these reciprocating pumps a plunger piston causes suction of the eluent during the back stroke and delivers the solvent under pressure during the forward stroke. To ensure proper flow direction, the pump head is usually equipped with ball check valves (see Fig. 4 f, g). Single-head, one-and-a-half-head, double-head and triple-head reciprocating pumps are available commercially. To reduce the pulsation caused by reciprocating pumps, many different techniques are applied. For double- and triple-head reciprocating pumps, two or three pistons are used at the same time with a 180° or 120° phase shift between

each. The resulting mobile phase flow is more or less periodic (see Fig. 5). The rest pulsation remaining can be decreased further if piston movement is not sinusoidal. Because of an eccentric disk, the second piston starts to deliver before the first has reached its end position. At the end position, the second piston has reached its maximum delivery output, while the first piston is refilled during its back stroke.

2.4. Sample Introduction Units

Good column performance is correlated directly with proper sample introduction onto the column in form of a narrow plug. Basically, the sample can be introduced into the high-pressure system in two ways: By syringe injectors or by sample valves.

2.4.1. Syringe Injectors

Syringe injection is the simplest way to introduce a sample to the system. The technique is very similar to that applied in GC. The sample is transferred with a syringe, through a self-sealing elastomeric septum, into the pressurized column. A specially designed pressure-resistant microsyringe must be used. One disadvantage of this technique is that column inlet plugging through particles of the septum can occur and the reproducibility of sample injection is often less than 2 % [14]. Therefore, this technique is used rarely, and most LC systems are equipped with sample valves.

2.4.2. Sample Valve Injectors

Today, the most common sample valves operate in the loop filling mode and are called *high-pressure sample injectors.* This means that the sample is transferred at atmospheric pressure from a syringe into a sample loop. By means of valving action the sample loop is then connected to the high-pressure mobile phase stream, which carries the sample onto the column (Fig. 6). The valving action can be carried out either manually or automatically by electric or pneumatic actuators.

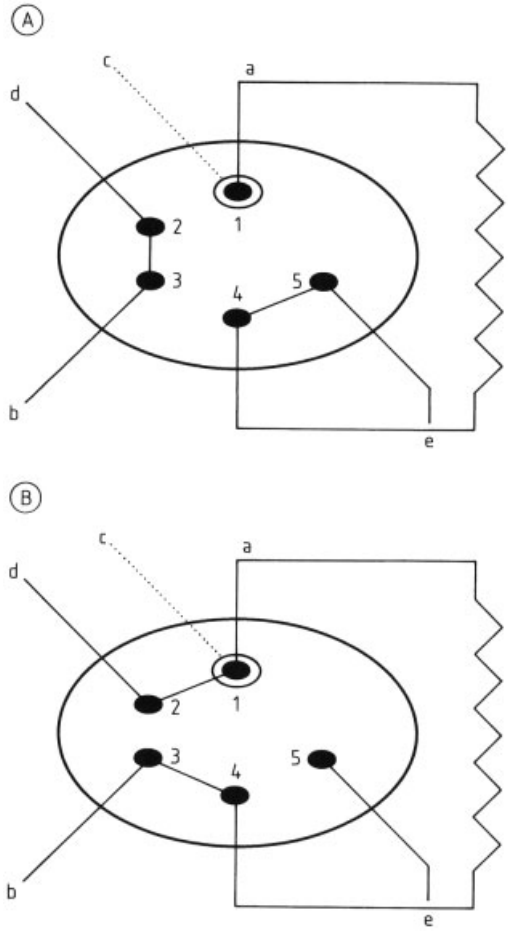

Figure 6. Diagram of a sample loop injection system [20]
A) Load position; B) Inject position
a) Loop; b) Column; c) Injection port; d) Pump; e) Waste

Commercial sample loops are available from 0.2 µL (for microcolumns) to 5 mL or more. On the other hand, a sample loop can be made easily by cutting an appropriate length of capillary (with known i.d.) and connecting it to the sample injector.

Normally the complete loop filling method is preferred to the partial filling method because the former has excellent volumetric precision. When loop volumes exceed 5 – 10 µL, this can be better than 0.1 % RSD. Nevertheless, the partial filling method still has a volumetric precision (dependent on syringe repeatability) better than 1 % relative standard deviation (RSD) [20].

2.4.3. Automatic Introduction Systems

Automatic sample introduction units (auto samplers) are often used for routine analysis (e.g., product quality control). A modern, commercially available HPLC auto sampler can handle ca. 100 samples; has a variable injection volume (e.g., from 1 to 1000 µL); can heat or cool the samples (e.g., 4 – 40 °C); and can be programmed while they are waiting to be injected; and can be programmed individually for each sample. The programmed run of the auto sampler should be able to be interrupted for an urgent analysis and continued later without any problems.

The principle of injection is the same as with sample loop injectors (see Section 2.4.2).

With an automatic sample introduction system, very efficient work is possible because sample cleanup, pretreatment (e.g., precolumn derivatization), and sample programming can performed during the day, while the actual analyses are run automatically at night.

2.5. Columns

Column Construction. HPLC columns (Fig. 7) are usually manufactured of stainless steel. Current standard HPLC columns have a 4.0- or 4.6-mm inner diameter and a length of 60 – 250 mm. Columns with smaller inner diameter (e.g., 1 or 2 mm – microbore columns) are also available. Although eluent consumption can be reduced with these columns, sensitivity is usually not increased and undesirable "wall effects" and dead volume effects become more noticeable with decreasing column diameter. Microbore columns are preferred only if the sample volume is very limited.

For preparative applications, columns with much larger dimensions (i.d. up to 900 mm and more) are available. Generally all parts of the column that do not contain the packing material, should be as small as possible to minimize dead volume. Often, unnecessarily large dead volumes result from false connection of the end fittings.

To ensure good separation, the inner column wall must be well polished and inert to the various mobile phases. Normal glass columns that are inert and have a very smooth inner wall can unfortunately be operated only at pressures less than ca. 5 MPa. To overcome this drawback, glass columns with a surrounding stainless steel jacket were developed that ensure stability at higher operating pressure. Pressure-stable columns made of PEEK which show the same positive properties as glass columns and are gaining interest for metal-free bioseparations.

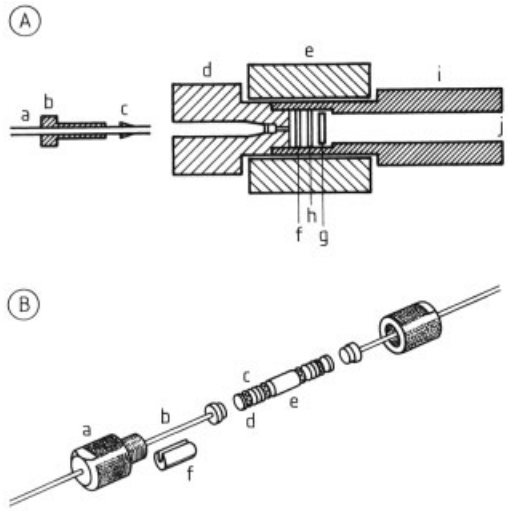

Figure 7. Diagram of the two most common column types for HPLC
A) Conventional HPLC column terminator for 1/16″ fittings
a) Capillary; b) Nut; c) Ferrule; d) End fitting with internal thread; e) Coupling body to column; f) Metal frit; g) Filter paper; h) Sealing ring; i) Stainless steel tubing; j) Column packing
B) Cartrige system column
a) Compression nut; b) Column fitting adapter; c) Sieve sandwich; d) PTFE soft seal; e) Column of precolumn; f) Spacer (required when working with precolumns)
Courtesy of Bischoff-Analysentechnik, Leonberg, Germany

2.6. Detectors

Another important instrument required in modern LC is a sensitive or selective detector for continuous monitoring of the column effluent. In GC, the differences in physical properties of the mobile phase (carrier gas) and the sample are great enough for universal detectors with good *sensitivity* to be used (e.g., flame ionization detector, thermal conductivity detector, → Gas Chromatography, Chap. 6.2.1.). The problem in LC is that the physical properties of the mo-

Table 1. Summary of the most important HPLC detectors

Detector	Type	Gradient possible	Maximum linear range	Maximum sensitivity, g/mL	Sensitive to changes in
UV – VIS	selective	yes	$10^3 - 10^5$	$10^{-9} - 10^{-10}$	pressure (temperature)
Diode-array detector (DAD)	selective	yes	$10^3 - 10^4$	$<2\times10^{-5}$	flow, pressure (temperature)
Refractive index (RI) detector	universal	no	$10^3 - 10^4$	10^{-7}	temperature (10^{-4} RIU °C), pressure
Fluorescence – phosphorescence detector	selective	yes	$10^2 - 10^4$	$<10^{-9}$	flow (temperature)
Fast-scanning detector	selective	yes	10^3	$<10^{-5}$	flow, temperature
Evaporative light scattering detector (ELSD)	universal∗	yes	$10^2 - 10^4$	10^{-8}	flow
Amperometric detector	selective	yes	10^6	10^{-12}	flow, temperature (1.5 %/°C)
FTIR detector	selective	yes	10^4	10^{-4}	flow
Mass selective detector (MS)	universal	yes	$10^2 - 10^4$	$<10^{-9}$ ∗∗	flow

∗ Only universal for nonvolatile solutes.
∗∗ Grams per second.

bile phase and the sample are often very similar, which makes the use of a universal detector impossible. Nevertheless, presently available LC detectors are very sensitive, are generally selective, and have a relatively wide range of applications (see Table 1).

To compare detectors from different manufacturers, characteristics must be found that determine the performance of the equipment. Besides instrument specifications such as cell geometry, cell volume, dimension of the connecting capillary, or additional special outfitting (e.g., integration marker, auto zero), analytical characteristics such as noise, linearity, sensitivity, and response must be considered. A survey of the different commercially available detectors is given in [21]. A good LC detector should have the following characteristics:

1) High sensitivity and detection limit
2) Good selectivity
3) Fast response
4) Wide range of linearity
5) No contribution to column band broadening (small dwell volume)
6) Reliability and convenience

Detector Sensitivity and Detector Noise. *Detector sensitivity* is the quantitative response of the detector signal to the amount of solute introduced (slope of the plot sample input M versus signal output A: sensitivity = dA/dM).

Noise originates from the electronics associated with the detector and from fluctuations of the physical parameters in the detector environment (e.g., flow pulsation or temperature change in refraction detectors, or light intensity changes in UV detectors). For further information, see [11].

Detector Linearity. Detectors differ in their linear dynamic range. The response of an ideal detector should be proportional to either the concentration or the mass flow rate of the solute (see Section 2.6.1). In reality, deviations from linearity usually occur at high concentrations (saturation effect). An easy way to determine the linear dynamic range of a detector is to plot the signal-to- concentration (or sample size) ratio versus the logarithm of concentration (sample size) (see Fig. 8).

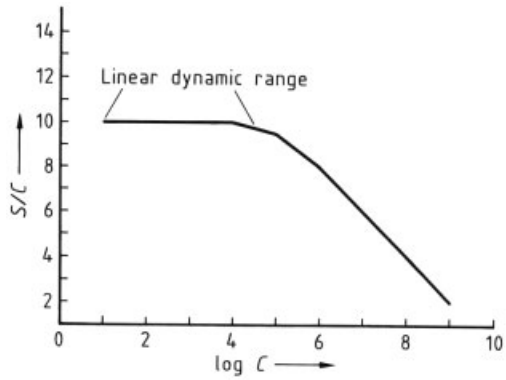

Figure 8. Plot used to determine detector linearity

2.6.1. Concentration-Sensitive and Mass-Flow-Sensitive Detectors

Two kinds of detectors exist, those that respond to solute concentration in the mobile phase and those that respond to changes of the solute mass flow rate into the detector cell. *Concentration-sensitive detectors* are used most frequently in LC. They are independent of the mass flow rate of the solute and are generally nondestructive (e.g., UV and differential detectors). (An easy way to find out whether a detector is concentration sensitive or not is to stop the solvent flow during a run. If the signal remains constant, it is a concentration-sensitive detector.)

The signal of *mass flow-rate-sensitive detectors* is proportional to the product of the solute concentration and the mobile phase flow rate. Consequently, if the flow rate is zero the signal is also zero, whatever the concentration may be. Detectors of this group are usually destructive (e.g., electrochemical detectors and mass spectrometers).

2.6.2. Differential Refractive Index (RI) Detector

The RI detector was developed in 1942 by Tiselius [22] and was one of the first commercially available on-line detectors. It is concentration sensitive and nondestructive, which makes it interesting for preparative chromatography. The detection limit of 10^{-6} g/mL is not as good as that of light absorption detectors (e.g., fluorescence or UV detectors), but the RI detector (RID) is a universal detector. To enhance the detection limit, a mobile phase should be chosen that differs as much as possible in its refraction index from the sample. A great disadvantage of the RID is that no eluent gradient can be run and the detector is very sensitive to even slight temperature variations [1 °C causes a change in the refractive index of ca. 10^{-4} refractive index units (RIU)]. Therefore, very good thermostating must be provided.

The four general methods of measuring refractive index are the Christiansen, the Frensel, the beam deflection, and the interferometric methods. Today, the most commonly applied technique in HPLC is the beam deflection method.

Beam Deflection Refractive Index Detector. A schematic of the beam deflection refractive index detector is shown in Figure 9. Here, the deflection of a light beam is measured in a single compact cell. A light beam from the incandescent lamp (g) passes through the optical mask (f), which confines the beam within the region of the sample (b) and the reference cell (c). The lens (d) collimates the light beam, and the parallel beam passes through the cells to a mirror (a). The mirror reflects the beam back through the sample and reference cell to the lens, which focuses it onto a photo cell (h). The location of the focused beam, rather than its intensity, is determined by the angle of deflection resulting from the difference in refractive index between the two parts of the cell. As the beam changes location on the photo cell, an output signal is generated that is amplified and sent to the recorder.

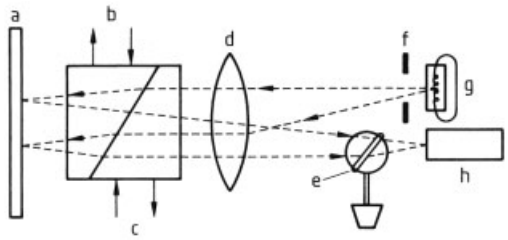

Figure 9. Diagram of a beam deflection (refractive index) detector [24]
a) Mirror; b) Sample cell; c) Reference cell; d) Lens; e) Optical zero-point adjustment; f) Mask; g) Lamp; h) Detector

2.6.3. UV – VIS Absorption Detector

The UV – VIS detector is the most frequently used detector in HPLC (Fig. 10). It is simple to handle, concentration sensitive, selective, and nondestructive. The latter also makes it suitable for preparative chromatography.

When light passes through a liquid (mobile phase), the intensity of absorption is proportional to the concentration of the analyte in the mobile phase C and the optical path length L (Lambert – Beer law):

$$\log I_0/I = ELC$$

where I_0 is the intensity of the light beam passing through the detector cell without the sample; I the intensity of the light beam passing through

the detector cell with the sample; I_0/I the transmittance; E the molar extinction coefficient; L the optical path length; and C the concentration of analyte in the mobile phase.

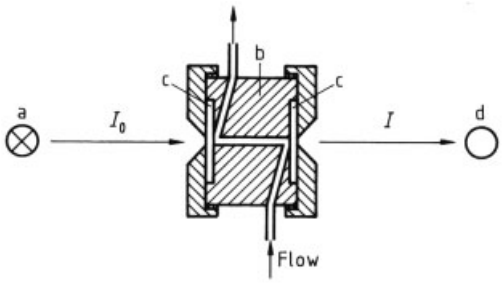

Figure 10. Diagram of a simple UV detector Z-pattern flow cell
a) Lamp; b) Flow cell; c) Quartz window; d) Photo detector

The Lambert – Beer law is theoretically valid only for measurements at the absorption peak maximum and for monochromatic light; it assumes that only one species present, absorbs light of the wavelength used. Only then does the Lambert – Beer law predict a linear relationship between the concentration and the response of the detector. The chromatographer should be aware of the fact that these optimal conditions are usually not fulfilled with real samples and the linearity of the Lambert – Beer law is not always obeyed. This is often the case when working with highly concentrated sample solutions. Generally, a sample peak should not be quantified if the absorption exceeds 1 AUFS (absorption units full scale). In working with UV – VIS detectors, many solvents also absorb – especially in the UV range (see eluotropic order) – which decrease the sensitivity significantly.

The analytes to be monitored must contain a chromophore; otherwise, UV – VIS detection is impossible. If no chromophore is present, one can often be introduced into the molecule by pre- or postcolumn derivatization (see Section 8.2).

Besides fixed-wavelength UV detectors, which usually operate at 254 nm, modern UV – VIS detectors have an adjustable wavelength from 190 to ca. 600 nm. This enables individual adjustment of the measuring wavelength according to the specific absorption characteristic of the solute. An additional advantage of the adjustable wavelength UV – VIS detector is that a UV – VIS spectrum of the solute can be obtained by simply stopping the mobile phase flow while the solute is inside the detector cell and scanning the sample over a specific range of wavelength. This can be very helpful in determining optimum wavelength at which a particular sample can be detected or identifying an analyte by its characteristic spectrum.

Some UV – VIS detectors, the so-called *multiwavelength detectors* (*MWDs*), are capable of monitoring several different wavelengths simultaneously, thereby enabling the sensitive detection of various analytes that differ in their adsorption maximum.

2.6.4. Diode-Array Detector

The diode-array detector (DAD) enables the simultaneous measurement of absorption versus analysis time (chromatogram) and absorption versus wavelength (spectrum), without having to stop the mobile phase flow. Diode-array detectors can record spectra simultaneously every few milliseconds at different points of an elution band (peak) during a chromatographic run. Usually the spectra are taken at the beginning, the maximum, and the end of an eluting peak. Comparison of the spectra obtained supplies important information concerning the identity and purity of the monitored peak.

Like modern UV – VIS detectors, the DAD can also monitor a sample at more than one wavelength at a time, thus enhancing sensitivity dramatically if the absorption maxima of the test compounds differ. Figure 11 (see next page) shows a schematic diagram of a typical DAD. Here, an achromatic lens system focuses polychromatic light into the detector flow cell. The beam then disperses on the surface of a diffraction grating and falls onto the photo diode array.

At present, good commercially available DADs have 512 diodes to cover a 600-nm wavelength range (ca. 190 – 800 nm). With decreasing number of diodes (for a set wavelength range) the resolution of the obtained spectra decreases as well, which results in a significant loss of information.

The drawback of a DAD is that because the light is split, up in front of the diodes, the resulting sensitivity is not as good as that obtained with classical UV – VIS detectors. Modern DADs are approximately a factor of 2, and older DADs a

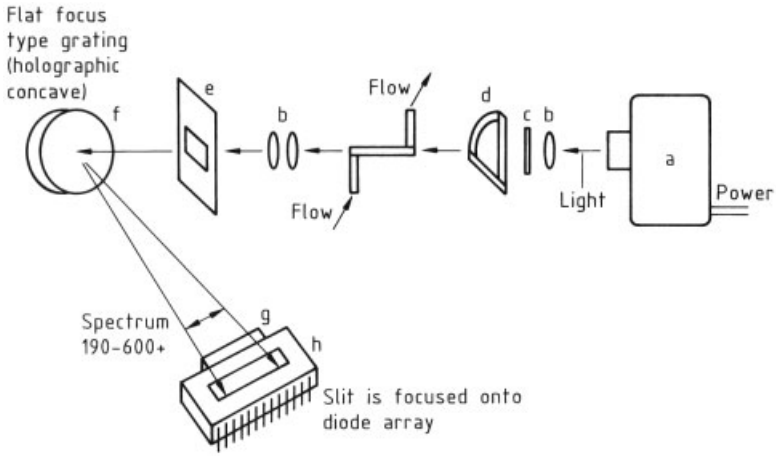

Figure 11. Diagram of a diode-array detector (DAD)
a) D_2 lamp; b) Lens; c) Aperture; d) Shutter; e) Slit; f) Flat focus-type grating (holographic concave); g) Second-order filter; h) Diode array
Courtesy of Waters-Millipore

factor of 10 or more, less sensitive than normal UV – VIS detectors.

Fast-Scanning Detector (FSD). The fast-scanning detector is a relatively inexpensive alternative to the diode-array detector in LC. Compared to the DAD it has only one diode (the reason why it is cheaper) and functions in principle as follows: Polychromatic light incident on a rotating grated mirror is split up into single wavelengths, which are then registered one after another by the diode. The required scan time is longer (ca. 1 s) than with a DAD (millisecond range). For liquid chromatography, this does not present a major problem because an average peak width is ca. 20 s, so enough time is available to monitor several spectra per peak for purity control, for example. In capillary electrophoresis (CE) this slowscan time can become a problem because peak width can be extreme short.

Indirect UV Detection. At the end of the 1980s, a new variation of UV detection was developed. This so-called indirect UV detection uses the same equipment as the classical DAD or UV – VIS detector. The difference between these methods is that in the classical mode the analyte contains a chromophore. In indirect UV detection a chromophore is added to the mobile phase so that a continuous, positive baseline signal is generated. If an analyte that has no chromophore passes the detector cell, the adsorption of the mobile phase is decreased and a negative peak is recorded. The advantage of this method is that analytes without chromophores (e.g., alkali and alkaline-earth ions) can be detected; the present disadvantage is its relatively poor detection limit (between 0.1 and 1 mg/L).

2.6.5. Fluorescence Detector

Fluorescence detection provides high selectivity and sensitivity, which is the reason for its use in trace analysis. The high selectivity results from the fact that only a few analytes are able to fluoresce. Typically, two types of spectra are obtained from fluorescing molecules: *excitation spectra* and *emission spectra.* Both spectra are characteristic for each analyte and therefore a great help in identification.

The high selectivity is due to the different detection technique applied (Fig. 12) compared to absorption spectroscopy (e.g., UV detector). The sample that passes the detector cell is irradiated by a light beam with a certain intensity I_0 and excitation wavelength λ_{ex}. Fluorescing solutes are able to emit a radiation $I*$ with a characteristically higher wavelength λ_{em} which is recorded and converted into an electric signal.

Since the emited radiation is not monochromatic, selectivity can be enhanced by inserting a monochromator into the beam. In this case the

increase in selectivity is paid for by a decrease in sensitivity.

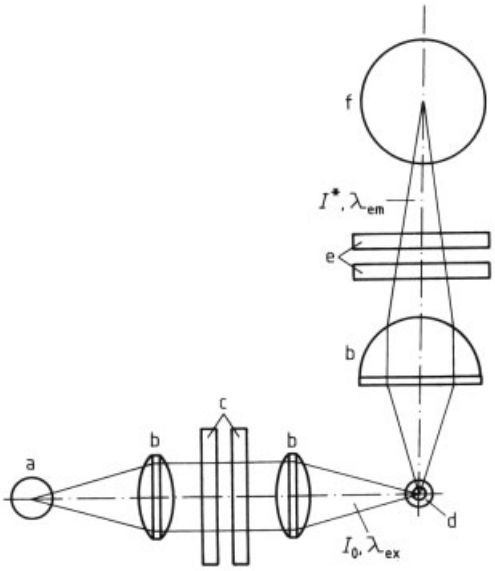

Figure 12. Diagram of a fluorescence detector
a) Lamp; b) Lens; c) Excitation filter; d) Flow cell; e) Emission filter; f) Photomultiplier

With help of pre- or postcolumn derivatization (Section 8.2), nonfluorescing analytes can be transformed into detectable ones, with very high selectivity and detection limits.

Since 1990, *laser-induced fluorescence* (*LIF*) detection has become increasingly important. A good review of the principles and applications of LIF detection is given in [25]. For more detailed information on fluorescence detection, see [26] [27].

2.6.6. Evaporative Light-Scattering Detector

The evaporative light-scattering detector (ELSD), sometimes wrongly called a mass detector, is influenced by many variables. Schulz and Engelhardt [28] showed that its specific response in addition to the analyte's molecular size also depends on the linear flow rate of the mobile phase; consequently, this detector is mass sensitive. Furthermore, they found that the ELSD behaves in certain situations more like a selective and not a universal detector. For example, it can detect only those solutes that are nonvolatile under the applied detection conditions. Nevertheless, ELSD can be a good alternative to the RI detector. Its sensitivity is better by one to two orders of magnitude, and a gradient can be run that would be impossible when measuring the refractive index. Another advantage of the ELSD compared to the RI and UV (<200 nm) detectors, is the much better signal-to-noise ratio and a reduced baseline drift tendency. Typical applications of the ELSD are in the analysis of sugars and other carbohydrates, polymers, resins, steroids, and bile acids.

2.6.7. Infrared Detectors – Fourier Transform Infrared Spectroscopy

Infrared detectors offer the possibility of scanning IR spectra during a chromatographic run. IR detection is limited to a mobile phase that does not absorb in the same range as the analyte.

Modern interfaces joining LC and Fourier transform infrared spectroscopy (FTIR) function in two ways: The first technique is based on a flow cell ("on-line"), in which sensitivity is limited by the dilution of the analyte in the mobile phase. In the second technique the mobile phase is evaporated, leaving the analyte as a small spot that is scanned ("off-line") by FTIR. The advantage of this approach is a significant enhancement in sensitivity. Compared to the refractive index detector the IR detector has a lower sensitivity but theoretically offers the possibility of gradient runs.

2.6.8. Electrochemical Detection

In liquid chromatography, electrochemical detection is a superior alternative to optical detection for many analytical problems [29]. Electrochemical detectors monitor currents, electrode potentials, or electrolytic conductivities. In all cases the signal is directly proportional to the concentration or mass flow rate over relatively wide linear ranges. With electronic amplification and simple signal processing, electrical quantities can be measured with high sensitivity.

Conductometric Detectors. Conductometric detectors are sensitive to ionic solutes in a

medium with low conductivity, for instance water or another polar solvent [30]. The response is relatively nonspecific since the conductivity of all ions present in the eluate is determined.

The reciprocal of the resistance is the conductance G (Ω^{-1}); it is expressed in the unit siemens (S). The conductivity $\varkappa$ is the product of the conductance G times the quotient of the cell electrode spacing D and area A (the cell constant):

$$\kappa = G\,(D/A)$$

The conductivity $\varkappa$ has the dimension microsiemens per centimeter.

The cell constant is important only for absolute measurements; it is determined by calibration with solutions of known conductivity. For relative measurements (i.e., to track eluate conductivity changes on a chromatographic column), the cell constant of the detector need not be known.

Conductivity detectors are important mainly for ion-exchange chromatography. In analytical practice, ion-exchange chromatography with the background conductivity suppressed is used chiefly for the highly sensitive determination of inorganic and organic anions and of alkali and alkaline-earth cations (see Section 5.3). The calibration curves (response peak height or area versus concentration) are linear over a wide range, up to three orders of magnitude; the limits of detection are in the nanogram-per-gram range [12].

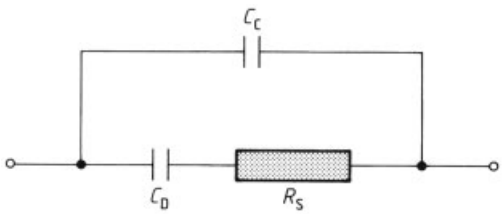

Figure 13. Equivalent circuit of conductivity measuring cell

Conductivity is measured in a cell containing two plates as electrodes (usually made of platinum) through which the column effluent flows. The electrodes have a known surface area and are located in the cell at a fixed distance. The cell is connected in an a.c. circuit and is characterized by the double-layer capacitance C_D, the cell capacitance C_C, and the electrolyte resistance R_S of the sample solution (see the equivalent circuit diagram in Fig. 13).

For the electrolyte resistance R_S to be determined, the effect of C_C must be negligible:

$$\frac{1}{2\cdot\pi\cdot f\cdot C_C} \gg R_S$$

where f is the frequency in Hz of the alternating current.

In practice, this condition is satisfied whenever very dilute solutions or solutions of weak electrolytes are measured with low-frequency alternating currents ($f = 40-80$ Hz) or when solutions having conductivities of about 100 µS/cm are measured with frequencies in the kilohertz range. In this way, R_S can be determined fairly accurately with a Wheatstone bridge circuit, shown in simplified form in Figure 14.

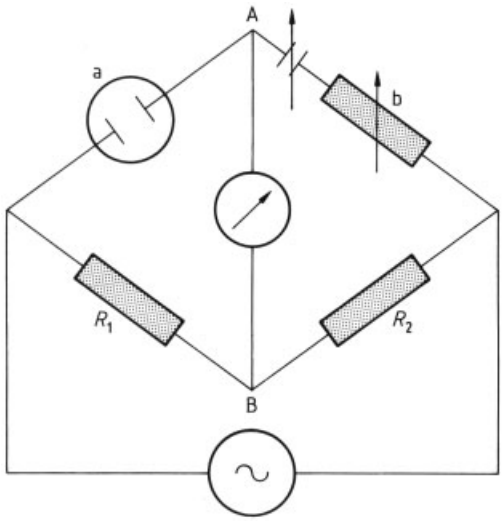

Figure 14. Wheatstone bridge circuit
a) Measuring cell; b) Balancing leg

Because the impedance of the measuring cell has a capacitive component, the balancing leg in the bridge circuit must include not only the variable resistance but also a variable capacitance. The bridge is balanced or nulled when the a.c. voltage between points A and B is zero. Direct-reading instruments are used to check the balance; for higher measurement accuracy the signals may be amplified first if necessary.

If the conductivity cell is used in combination with a chromatographic column, the mea-

suring electrodes have the form of capillaries. This makes it possible to keep the cell volume very small, as required; the volume is ca. 1–2 µL. Figure 15 illustrates the principle of an HPLC conductivity detector with two measuring electrodes. The cell must be thermally insulated because the conductivity is temperature dependent.

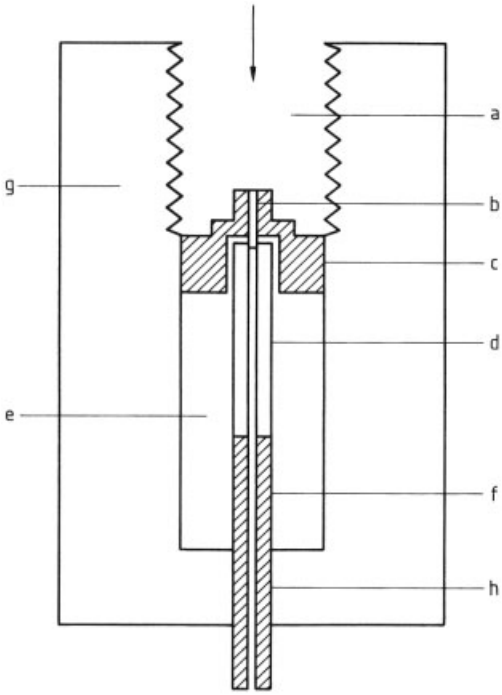

Figure 15. HPLC conductivity detector [31]
a) Connection to column; b) Platinum capillary; c) Electrical connection; d) Polytetrafluoroethylene capillary; e) Polytetrafluoroethylene; f) Electrical connection; g) Plastic detector housing; h) Stainless steel capillary
Measurement is performed between (b) and (h)

Four- and six-electrode cells have been designed for precision measurements [13, 14].

Potentiometric Detectors. Essentially, potentiometric detectors consist of ion-selective electrodes (ISEs) to monitor ionic species in chromatographic effluents. Examples of ISEs are glass, solid-state, liquid-membrane, and enzyme electrodes.

The detector response is the potential of the electrode, which depends on the activity of the analyte. This potential is measured with respect to a reference electrode. The response can be described by [32, 33]

$$E = E_0 + \frac{s}{z_i} \log \left(a_i + \sum K_{ij}^{\mathrm{Pot}} a^{z_i/z_j j} \right)$$

where E is the electromotive force of the cell assembly in volts; z_i and a_i are the charge number of the primary ion i and its activity in the sample solution; z_j and a_j are the charge number of the interfering ion j and its activity in the sample solution; K_{ij}^{Pot} is the selectivity factor, a measure of the sensor's preference for interfering ion j relative to measured ion i (an ideally selective membrane electrode would have all $K_{ij}^{\mathrm{Pot}} = 0$); E_0 is the potential difference equal to a constant (for a given temperature) plus the liquid junction potential; and $s = 2.303\, RT/z_i F = 59.16\,\mathrm{mV}/z_i$ (at 25 °C for the ideal case).

The activity of an ion is related to its concentration c_i (moles per liter) by $a_\mathrm{i} = c_\mathrm{i} \cdot \gamma_\mathrm{i}$, where γ_i is the activity coefficient of the ion.

Ion-selective electrodes are less important for monitoring chromatographic eluates than for automated flow injection analysis (FIA).

Major characteristics of potentiometric detectors are high specificity, low sensitivity, and slow response.

Only a few examples can be found in the literature in which ISEs have been used as detectors for ion chromatography. They have been employed to detect nitrate in the presence of nitrite [32] and to detect alkali cations [33, 34] and a variety of anions [35, 36] after isolation by ion-exchange liquid chromatography.

Potentiometric detectors vary widely in cell design. One example worth noting is a microtype membrane cell with two compartments separated by an ion-exchange membrane. The column eluate is passed through one compartment of the detector, and the reference solution with the same components of the eluate is passed through the other. This device is useful for monitoring ion-exchange as well as reversed-phase ion-pair chromatography; it offers detection limits in the nanomole range, together with a linear response over about two to three orders of magnitude [34].

The design of a picoliter-volume cell potentiometric detector for open tubular column liquid chromatography is shown in Figure 16. The microelectrode has a tip diameter of 1 µm and is inserted in the open tubular column. A

cation-exchange solution can be used as membrane. The reference electrode, located in a micropipette, is a chlorinated silver wire in 3-mol/L KCl. The response time of the microdetector is less than 10 ms [33].

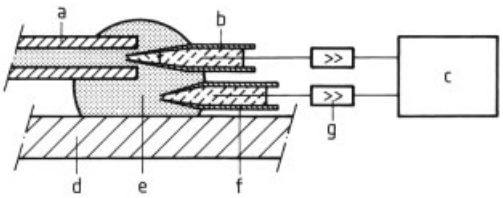

Figure 16. Potentiometric detector with ion-selective microelectrodes [33]
a) Capillary column; b) Ion-selective microelectrode; c) Voltmeter-recorder; d) Glass plate; e) Eluent drop; f) Reference electrode; g) Impedance converter

Amperometric Detectors. Amperometric detectors are the most commonly used electrochemical detectors for highly sensitive and selective determinations in HPLC [30, 31]. The frequently used synonym "electrochemical detector" is not very precise in relation to the detection mode.

The signal current is the result of an electrochemical conversion of the analyte by oxidation or reduction at the working electrode (Fig. 17).

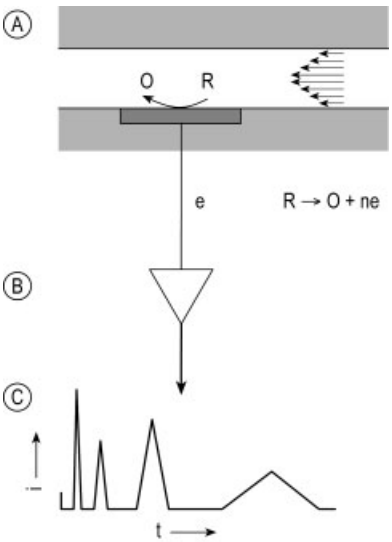

Figure 17. Schematic view of thin-layer amperometric detection (conversion of reactant (R) to product (O) on the surface of the working electrode)
A Detector Cell; B Amplifier; C Chromatogram

The currents are in the pA – μA range and are measured by the three-electrode technique (→ Analytical Voltammetry and Polarography, Chap. 3.). As a result extremely low detection limits at the 0.1-pmol level can be achieved.

The simplest and most common detection mode is current measurement at constant potential. The reference and counter electrodes should be located downstream of the working electrode so that reaction products at the counter electrode or leakage from the reference electrode do not interfere with the working electrode. The current is amplified and plotted versus time to yield a chromatogram.

Because the electrode process takes place solely at the surface of the working electrode, amperometric flow cells can determine only local (interfacial) concentrations. Conductometric and potentiometric measuring cells, on the other hand, can track the average volumetric concentrations in the eluate. The same holds for UV detection, which is preferred for routine HPLC studies (see Section 2.6.3).

In amperometric detector cells, the analyte is transported to the working electrode surface by diffusion as well as convection; migration is suppressed by the supporting electrolyte. Electrolyte concentrations in the eluent of 0.01 – 0.1 mol/L are sufficient. Several cell geometries employing solid electrodes have been designed and tested. The two types shown in Figure 18 (see next page) are the most commonly used.

In the *thin-layer cell*, the eluate is forced through a thin, rectangular channel, normally 0.05 – 0.5 mm thick. The working electrode (usually a disk electrode, 2.4 mm in diameter) is embedded in the channel wall. In the "wall-jet" design, the solution is introduced into the cell through a small diameter jet or nozzle and impinges perpendicularly onto the planar electrode. In both cases, the volume is ca. 1 – 2 µL. The advantages of these cells in comparison with other detector designs are good hydrodynamic properties and fast response [37, 38].

The thin-layer cell for HPLC was originally described by Kissinger and coworkers in 1973 [39]. The wall-jet principle has been known since 1967 and was originally introduced for HPLC in 1974 [40]. The development is based on Kemula's work with a mercury drop as detector for column chromatography [41].

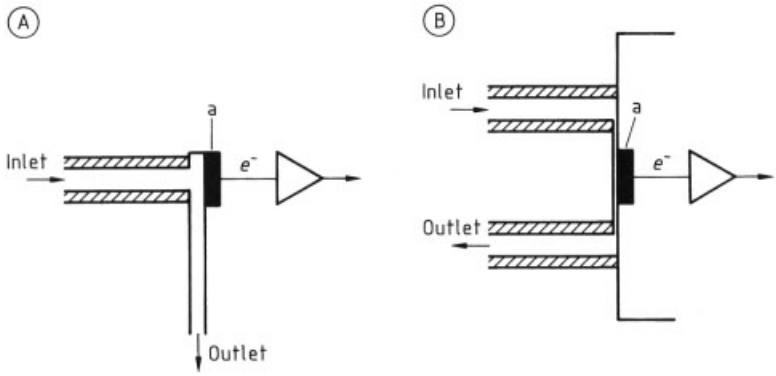

Figure 18. Schematics of the most common amperometric detector cells
A) Wall-jet cell; B) Thin-layer cell
a) Working electrode

When mercury is used as the electrode material, detection is possible over a particularly wide cathodic potential range. The anodic range, however, is limited. It is thus commonly used for the detection of reducible species. Detectors which utilize a static mercury drop are particularly attractive as the drop (working electrode) is easily renewed for each chromatogram; this is quite an advantage for the reproducibility of response [42]. The construction and operation of detectors based on mercury drop electrodes have been reviewed [42, 43].

Solid electrodes, which have a larger anodic range than mercury-based electrodes, include different types of carbon and noble-metal electrodes. The preferred electrode material for commercial detectors is glassy carbon. Its electrochemical properties allow measurements in a wide potential range of about − 0.8 to +1.2 V. The effective range (chiefly the cathodic voltage range) depends on the composition, mainly on the pH value, of the eluate. In contrast to carbon-paste electrodes, which have been proposed for amperometric detection – particularly because of the low background response and the good signal-to-noise ratio – glassy carbon electrodes offer better chemical resistance. They are stable in all solvents employed, such as acetonitrile and methanol as mobile phase.

Noble-metal (Au, Pt) electrodes have been employed for special studies, and gold amalgam electrodes as well as mercury-film electrodes have been used for reductive determinations [44–46].

The basis of the good selectivity of amperometric detection is that only compounds with electroactive functional groups (reducible or oxidizable) can undergo reaction at the electrode. The potential at which these electrode reactions occur depends on the redox behavior of the compounds of interest and the working conditions (working and reference electrodes, composition and pH of the mobile phase). The most suitable detection potential is best determined experimentally from the cyclic voltammogram or the hydrodynamic voltammogram of the analyte.

A *hydrodynamic voltammogram* is a current – potential curve which shows the dependence of the chromatographic peak height on the detection potential. The technique used to obtain the necessary information is voltammetric flow injection analysis, in which an aliquot of the analyte is injected into the flowing eluent prior to the detector and the peak current recorded. This is repeated many times, the detector potential being changed after each injection, until the peak current – potential plot reaches a plateau or a maximum, as shown for some polycyclic aromatic hydrocarbons (PAHs) in Figure 19.

Whereas indeno[1,2,3-*cd*]pyrene, benzo[*ghi*]perylene, benzo[*k*]fluoranthene, and benzo[*a*]pyrene give a maximum detection signal at potentials between + 1.30 and + 1.35 V (Figure 19 A), the current – potential curves of acenaphthalene, acenaphthene, fluoranthene and benzo[*b*]fluoranthene do not reach a maximum even at + 1.45 V (Figure 19 B). A measurable signal is first obtained at + 1.35 V

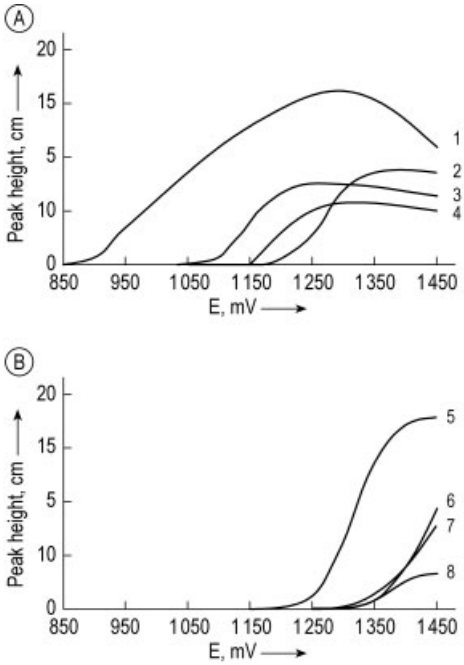

Figure 19. Hydrodynamic voltammograms of polycyclic aromatic hydrocarbons (Eluent: methanol/water 85/15) [47,67]. 1) Benzo[*a*]pyrene; 2) Benzo[*k*]fluoranthene; 3) Benzo[*ghi*]perylene; 4) Indeno[1,2,3-*cd*]pyrene; 5) Acenaphthene; 6) Fluoranthene; 7) Acenaphthalene; 8) Benzo[*b*]fluoranthene.

for acenaphthalene benzo[*b*]fluoranthene, and fluoranthene, whereas the other PAHs are oxidized at less positive potentials. Based on these results, it is possible to selectively determine benzo[*a*]pyrene in a mixture of PAHs by using a detection potential of + 1.0 V. At more positive potentials, other peaks appear in the chromatogram until finally, at a potential of + 1.45 V, peaks for all the PAHs are recorded, as shown in the chromatograms in Figure 20.

When using amperometric detectors, the composition of the eluent should remain as constant as possible (isocratic elution). If gradient elution is used to reduce the analysis time, then base line drift is observed, resulting in a loss of sensitivity. This is particularly so with glassy carbon working electrodes and is due to the slow adjustment of the electrochemical equilibrium at the electrode surface to the changing eluent composition.

Working electrodes made of other materials such as the ultratrace graphite electrode (porous graphite structure impregnated with epoxy resin, produced by Metrohm, Switzerland), have much shorter lead-in times. The relatively rapid response of this kind of electrode arises from its semimicro electrode properties [48].

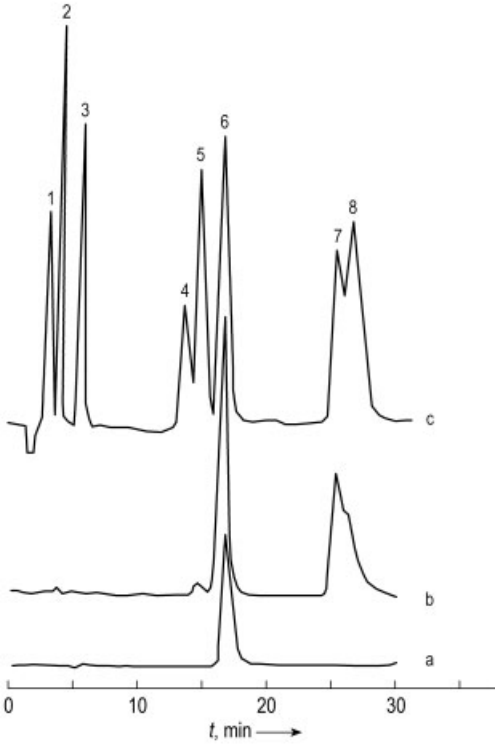

Figure 20. Determination of polycyclic hydrocarbons by HPLC with amperometric detection.
a) Detection potential + 1.0 V : selective determination of 10 ng benzo[*a*]pyrene (6). b) Determination of 10 ng benzo[*a*]pyrene (6) and 10,5 ng benzo[*ghi*]perylene (7) at +1.2 V. c) Determination of 8.6 ng acenaphthalene (1), 11.4 ng acenaphthene (2), 51,5 ng fluoranthene (3), 9.5 ng benzo[*a*]fluoranthene (4), 10.5 ng benzo[*k*]fluoranthene (5), 10 ng benzo[*a*]pyrene (6), 10.5 ng benzo[*ghi*]perylene (7) and 10 ng indeno[1,2,3-*cd*]pyrene at a detection potential of + 1.45 V.
Eluent : methanol/water 85/15 with 2 g/L trichloracetic acid as supporting electrolyte. Working electrode: glassy carbon; reference electrode: Ag/AgCl in methanolic KCI. The detection and determination limits were in the range 100 – 300 pg per injection [47,67].

When used as the working electrode in amperometric detection following gradient elution, e.g., in the analysis of beno[*a*]pyrene (BaP) and its metabolites, the hydroxybenzo[*a*]pyrenes, which result from the enzymatic degradation of BaP in animals and humans after uptake in food or from the air, rapid adjustment to steady state

conditions is an important requirement and leads to a more stable base line (Figure 21 B). For comparison, the chromatogram using the much slower isocratic elution is shown in Figure 21 A. The detection limits for the analysis of benzo[*a*]-pyrene and its metabolites lie between 0.2 ng and 0.4 ng and the determination limits between 0.3 ng and 0.6 ng when either the glassy carbon electrode or the ultratrace electrode is used.

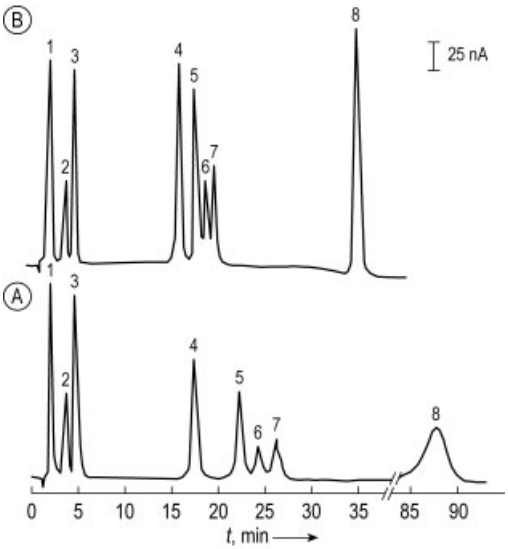

Figure 21. Chromatograms of BaP and metabolites (each 90 ng). Amperometric detection at + 1.35 V (vs Ag/AgCl, 3 M KCl) using an ultratrace electrode with isocratic elution (A) and a glassy carbon electrode with gradient elution (B) [49].
1) Benzo[*a*]pyrene-*trans*-9,10-dihydrodiol; 2) Benzo[*a*]-pyrene-*trans*-4,5-dihydrodiol; 3) Benzo[*a*]pyrene-*trans*-7,8-dihydrodiol; 4) 9-Hydroxybenzo[*a*]pyrene; 5) 7-Hydroxybenzo[*a*]pyrene; 6) 1-Hydroxybenzo[*a*]pyrene; 7) 3-Hydroxybenzo[*a*]pyrene; 8) Benzo[*a*]pyrene.

Amperometric detection after HPLC separation has been used to detect easily oxidizable phenols, aromatic amines, indoles, phenothiazines, thiols, and polyaromatic hydrocarbons (Table 2) and easily reducible nitro compounds and quinones (Table 3). These methods have been used in environmental, pharmaceutical, biochemical, and food analysis, as well as for determining hormone residues in soil and water samples and in plant and animal products. The main application is in the analysis of phenols and aromatic amines such as the catecholamines, which can be detected with high sensitivity in the lower picogram region in body fluids and biological samples [51–55].

While the current is measured at a constant voltage in amperometric detectors, in voltammetric detectors it is measured while the voltage is changing rapidly. Either linear scan, staircase, square-wave, pulse or a.c. voltammetry can be used for the measurement. When compared to stationary methods, nonstationary methods with superimposed voltage pulses often result in higher selectivity and sensitivity and fewer problems from the adsorption of impurities and electrode reaction products on the electrode surface [53–58].

A technique for cleaning the surface of solid working electrodes to improve the reproducibility of the measurements (removal of electrochemical reaction products) is to alternate the potential repeatedly [59].

For the "fast-scan" technique in HPLC, in which the potential is varied rapidly across its full range, the method of square-wave voltammetry (→ Analytical Voltammetry and Polarography, Chap. 2.2.) has provided the basis for the development of a "rapid-scan square-wave voltammetric detector" [60]. In addition, a scanning microvoltammetric detector for open tubular liquid chromatography has been described [61]. This detector consists of a 9-µm (outer diameter) carbon fiber as electrode, inserted into the end of a 15-µm (inner diameter) capillary column. The potential can be scanned at rates up to 1 V/s over a range of 0.0 to +1.5 V. A distinct advantage of this detection method is its three-dimensional resolving power.

Figure 22 shows a conventional chromatogram of five compounds (ascorbic acid, epinephrine, tyrosine, dopamine, and hydroquinone) detected at +1.0 V. The third peak in the chromatogram consists of both tyrosine and dopamine, because these are not resolved chromatographically. This is revealed by the three-dimensional chromatogram of Figure 23 (both see page 23).

No single potential can be used to detect the presence of the two overlapping peaks. However, a potential scan yields a three-dimensional chromatogram of the mixture, in which the two peaks are quite obvious.

Coulometric Detectors. Coulometric detectors are based on potentiostatic coulometry [30].

Table 2. Organic analytes detected amperometrically in the oxidative mode [50]

Analyte with electroactive group	Structural formula	Operating potential, mV	Examples
Aromatic amines			
Anilines	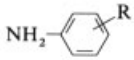	+1000	phenylenediamine, 3-chloroaniline, aniline, methylaniline, naphthylamine, diphenylamine
Benzidines	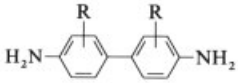	+ 600	3,3′-dichlorobenzidine
Sulfonamides	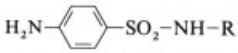	1200	
Aromatic hydroxyls			
Phenols, halogenated phenols	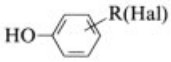	+1200	salicylic acid (and derivatives) benzofuranols, chlorophenols (and metabolites), paracetamol, hesperidin, tyrosine
Flavones	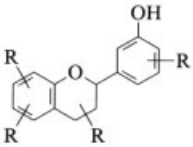	+1100	flavonoles, procyanidins
Catecholamines	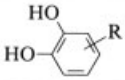	+ 800	DOPA, dopamine nor-epinephrine, epinephrine, and their metabolites
Estrogens	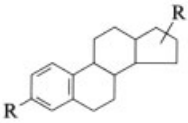	+1000	estriol, catecholestrogens
Antioxidants	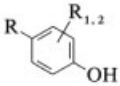	+1000 (200)	*tert*-butylhydroxyanisole, *tert*-butylhydroxytoluene
Hydroxycumarins	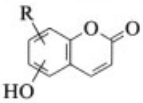	+1000	esculin
Aliphatic amines (*sec*- and *tert*-)		+1200	clonidine, diphenhydramine
Indoles	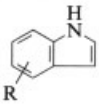	+1000 + 450	tryptophan, pindolol, bopindolol metabolit
Thiols	R−SH	+ 800	cysteine, glutathione, penicillamine
Phenothiazones	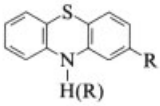	+1000	thioridazine, chlorpromazine

Table 2. Continued.

Pyrimidines	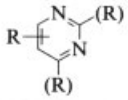	+1200	trimethoprim
Purines	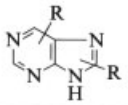	+1000	uric acid
Hydrazines	$R-NH-NH_2$	+ 600	hydralazine
Miscellaneous			
Vitamines		+ 800–1000	vitamins A, E, C
Abusive drugs		+ 700–1200	morphine and derivatives, benzodiazepines, cocaine, hallucinogens, tetrahydro-cannabinols, tricyclic anti-depressants, neuroleptics
Pesticides, herbicides		+1100–1300	phenylurea derivatives, carbamates

Pyrimidines	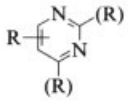	+1200	trimethoprim
Purines	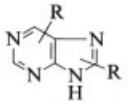	+1000	uric acid
Hydrazines	$R-NH-NH_2$	+ 600	hydralazine
Miscellaneous			
Vitamines		+ 800–1000	vitamins A, E, C
Abusive drugs		+ 700–1200	morphine and derivatives, benzodiazepines, cocaine, hallucinogens, tetrahydro-cannabinols, tricyclic anti-depressants, neuroleptics
Pesticides, herbicides		+1100–1300	phenylurea derivatives, carbamates

The signal of the constant-potential measurements, as in amperometric detection, is the current resulting from an electron-transfer process (reduction or oxidation of the analyte arriving with the eluent) while the working electrode is held at constant potential.

In both flow-through detection techniques, the instantaneous measured current i_t is given by

$$i_t = n \cdot F \frac{dN}{dt}$$

where n is the number of electrons per mol exchanged in the reaction, F is the Faraday constant, and dN/dt is the number of molecules per unit time reacting at the working electrode.

The difference between coulometric and amperometric detection lies in the percentage of analyte converted at the working electrode. Of the molecules located in the measuring cell at a given time, only ca. 1 – 10 % take part in the electrode reaction in amperometric detection, whereas this conversion should be > 95 % in coulometric detection.

To achieve nearly 100 % conversion, the ratio of electrode area to cell volume must be very large in coulometric detectors. The work-

Table 3. Organic analytes detected amperometrically in the reductive mode [50]

Analyte with electroactive group	Structural formula	Operating potential, mV	Examples
Benzodiazepines (and -4-oxides)	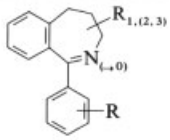	−500 to −1000	chlordiazepoxide, nitrazepam, diazepam
Steroid hormones	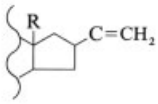	−500	ethinylestradiol, levonorgestrel
Aromatic and aliphatic nitro compounds	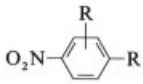	−400 to −800	explosives (TNT, nitroglycerin), chloramphenicol, nitropolyaromatics, parathion derivatives, dinitrophenol pesticides
Quinones	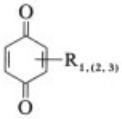		vitamin K
Miscellaneous Nitrosamines, thioamides, organometallic compounds, peroxides, azo compounds			prothionamide, benzoyl peroxides, triphenyllead, pyridylazo resorcinol

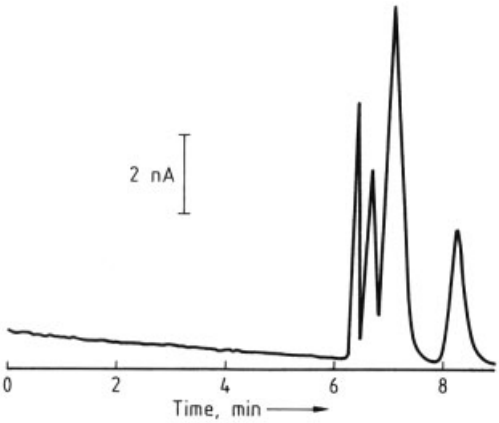

Figure 22. Single-potential chromatogram at +1.0 V versus Ag – AgCl of (in order of elution) 0.1-mmol/L ascorbic acid, 0.1-mmol/L DL-epinephrine, 0.2-mmol/L tyrosine, 0.1-mmol/L dopamine, 0.1-mmol/L hydroquinone
Scan rate was 1 V/s, flow rate 0.59 cm/s; electrode length, 1 mm [61]

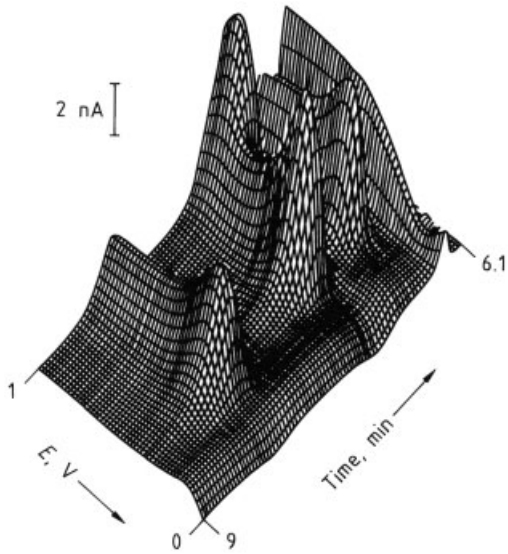

Figure 23. Partial threedimensional chromatovoltammogram of mixture shown in Figure 22, [61]

ing electrodes employed in coulometric detectors must therefore have a much greater surface area (ca. 5 cm^2) than those used for amperometric detectors. Large electrode areas, however, lower the signal-to-noise ratio and thus increase the detection limits [62]. For this reason, and also because of problems with electrode passivation, coulometric detection has not become an important technique in HPLC, even though the same classes of compounds can be determined as with amperometric detection.

A coulometric detector with two working electrodes and separate reference electrodes, configured one downstream of the other, is called a *dual-electrode cell* [15,63,64]. If different potentials are applied to the two working electrodes, two separate chromatograms can be recorded for the same eluate. If the cell is appropriately sized and large-area electrodes (e.g., porous graphite) are employed, current (conversion) efficiencies close to 100 % can be attained. The design of a dual-electrode coulometric detector is illustrated in Figure 24.

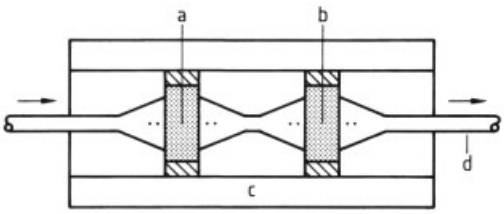

Figure 24. Coulometric detector with dual porous electrodes and total cell volume $< 5\,\mu L$ (Coulochem 5100 A electrochemical detector)
a) Test electrode 1; b) Test electrode 2; c) Stainless steel tubing

Substances with reversible electrode behavior can be determined more sensitively with dual detectors than with conventional coulometric detectors. One example is the anodic and cathodic detection (redox mode) of catecholamines [65]. The determination can also be made more sensitive if the two detector signals are fed to a difference circuit so as to eliminate the background that is due chiefly to contaminants in the solvent (differential mode). Dual detectors operated at appropriate electrode potentials can also block out the signals of undesirable compounds (screen-mode detection).

3. Solvents (Mobile Phase)

3.1. Purification and Pretreatment of LC Solvents

Especially in trace analysis, purification and pretreatment – even of commercially available HPLC-grade solvents – are often necessary (e.g., by filtration or distillation). Occasionally, organic solvents have to be dried as well (normal phase chromatography), for instance, with a molecular sieve. Dissolved atmospheric oxygen in the mobile phase is a major source of trouble in LC. Oxygen increases the UV absorption below 230 nm, thus disturbing sensitive UV detection at lower wavelengths significantly. Dissolved air furthermore, leads to air bubbles inside the chromatographic system (often noticed in "low-pressure" gradient systems), which are responsible for pulsation of the pump or can cause "ghost" peaks when entering the detector cell (e.g., refractive index detector or UV detector). To prevent these problems, the mobile phase has to be degassed. The best method for degassing the eluent is to rinse it with an inert gas (e.g., helium). Often, however, applying vacuum degassing or simply degassing the eluent by ultrasonic treatment for several minutes before use is sufficient.

3.2. Elutropic Order

The more strongly a solvent is adsorbed to the stationary phase, the higher is its elution power. Because the solvent molecules outnumber the analyte molecules with respect to the available adsorption sites of the stationary phase, a relative weak solvent (mobile phase) is still strong enough to elute even quite polar sample compounds. In classical normal phase (NP) adsorption chromatography, the different solvents are listed according to increasing elution power (elution strength), resulting in their elutropic order (see Table 4). The elutropic order was developed for normal phase chromatography; consequently, water is the strongest solvent. Today, most separations are carried out in the reversed-phase (RP) chromatography, so the elutropic order is inversed, turning water into the weakest mobile phase (see Section 4.3).

Table 4. Important characteristics of the most frequently used mobile phases in LC ∗

Solvent	Polarity	Solvent strength	Dielectric constant	UV cutoff, nm	Refractive index	Viscosity at 25 °C, mPa · s
Heptane	0.2	0.01	1.92	195	1.385	0.40
Hexane	0.1	0.01	1.88	190	1.372	0.30
Pentane	0.0	0.00	1.84	195	1.355	0.22
Cyclohexane	− 0.2	0.04	2.02	200	1.423	0.90
Toluene	2.4	0.29	2.40	285	1.494	0.55
Ethyl ether	2.8	0.38	4.30	218	1.350	0.24
Benzene	2.7	0.52	2.30	280	1.498	0.60
Methylene chloride	3.1	0.42	8.90	233	1.421	0.41
1,2-Dichloroethane	3.5	0.44	10.40	228	1.442	0.78
Butanol	3.9	0.70	17.50	210	1.397	2.6
Propanol	4.0	0.82	20.30	240	1.385	1.9
Tetrahydrofuran	4.0	0.57	7.60	212	1.405	0.46
Ethylacetate	4.4	0.58	6.00	256	1.370	0.43
Isopropanol	3.9	0.82	20.30	205	1.384	1.9
Chloroform	4.1	0.40	4.80	245	1.443	0.53
Dioxane	4.8	0.56	2.20	215	1.420	1.20
Acetone	5.1	0.56		330	1.356	0.30
Ethanol	4.3	0.88	24.60	210	1.359	1.08
Acetic acid	6.0		6.20		1.370	1.10
Acetonitrile	5.8	0.65	37.50	190	1.341	0.34
Methanol	5.1	0.95	32.70	205	1.326	0.54
Nitromethane	6.0	0.64		380	1.380	0.61
Water	10.2		80.00		1.333	0.89

∗ Data taken from elutropic order in [19].

With the help of the elutropic order, appropriate solvents can be chosen, or various solvent mixtures created, that have the same elution strength but different selectivity.

3.3. Selection of Mobile Phase

The mobile phase in LC consists of one or more solvents, which additionally may contain modifiers. Generally, the composition of the eluent should be kept as simple as possible. When working with gradients, solvated mobile phase modifiers (e.g., salts) might precipitate in the chromatographic system because of the dramatic change in eluent composition. The following factors must be considered in choosing an appropriate mobile phase for a specific LC problem:

1) Solubility of the sample in the eluent
2) Polarity of the eluent
3) Light transmission of the eluent
4) Refractive index of the eluent
5) Viscosity of the eluent
6) Boiling point of the eluent
7) Purity and stability of the eluent
8) Safety regulations
9) pH

Items 1 and 2 must be seen in the context of the separation system applied (see Chap. 5).

The *light transmission* of the eluent is of importance in using a UV detector. When monitoring a sample at a wavelength of 230 nm, acetone, for instance, (which absorbs up to 330 nm) cannot be used as a mobile phase. As long as the absorption of the eluent does not exceed 0.5 AUFS, it can be tolerated. Table 4 shows some important characteristics and the UV cutoff from often used eluents. Organic or inorganic trace compounds (impurities) in the mobile phase often shift the UV absorption to higher wavelengths.

When working with a refractive index detector, the difference in the *refractive index* of the sample and the mobile phase should be as great as possible to achieve better sensitivity and lower detection limits.

The *viscosity* of the mobile phase is responsible for the back-pressure of a given LC system. In addition, the higher viscosity of the eluent results in a lower diffusion coefficient for the sample, which leads to reduced mass transfer and peak broadening. Especially during gradient elution with two or more relatively polar organic solvents, unexpected changes in viscosity can occur [66].

The possibilities for influencing the viscosity of the mobile phase are limited. Working at

higher temperatures can be an advantage in some cases (e.g., high-temperature liquid chromatography).

The *boiling point* of the solvents should also be considered. For preparative chromatography, a high-purity and low-boiling mobile phase is preferred. On the other hand, the boiling point should not be too low because that could cause gas bubbles in the pump or detector.

The pH of the mobile phase often has a dramatic influence on the selectivity and resolution of investigated samples. For many pharmaceuticals, the optimal pH of the mobile phase is between pH 4 and 5. During method development and optimization, it is recommended to run a sequence of samples between pH 2.5 and 7.5 in pH steps of 0.5. Plotting resolution and selectivity against mobile-phase pH then allows the optimum eluent pH to be identified.

Of course, the eluent must also be sufficiently stable toward thermal or chemical influence. Most organic solvents are easily flammable and have a high vapor pressure at room temperature. Therefore, LC instrumentation should always be placed in a well-ventilated area.

The pH of the mobile phase has a strong influence on the solute retention and the peak shape of eluting ionogenic compounds. It is a critical factor for the reproducibility and ruggedness of a HPLC method (see also [67]).

4. Column Packing (Stationary Phase)

A very important factor for successful HPLC separation is the correct choice of column packing. In the past, many different packing materials were tested, for instance, powdered sugar, diatomaceous earth, aluminum oxide, calcium carbonate, calcium hydroxide, magnesium oxide, Fuller's earth, and silica. Today, silica is generally used as the basis for adsorbent chromatography; aluminum oxide is also used to a certain extent. Both materials are pressure stable and are unaffected by pH over a wide range.

Particle Size, Particle-Size Distribution. Standard particle size for analytical HPLC columns is 3, 5 or 10 µm. The smallest commercially available particle size is 1.5 µm. The particle-size distribution should be as narrow as possible so that "good" packings and therefore good separations are obtainable. The smaller the particle-size distribution of the stationary phase, the higher is its quality and therefore its price. However, if the size of stationary phase particles is absolutely identical, this can cause problems in packing the column.

Pore Diameter, Specific Surface Area. In adsorption chromatography a packing material with a very high specific surface area is required. The larger the surface area, the more active sites are available for (adsorptive) interaction. In the case of spherical particles (10-µm i.d.) $< 1\,\%$ of the geometrical surface contributes to the total surface area. The dominant remainder of the surface area originates from the pores of the particles. A normal silica phase has a specific surface area of ca. 300 m^2/g, a mean pore diameter of 10 nm, and a mean pore volume of ca. 0.5 – 1.0 mL/g.

Most analytes have no access to pores with a diameter smaller than 3 – 4 nm. Therefore the packing material should show a narrow pore-size distribution so that the total number of very small pores (to which the analytes have no access) is as low as possible. Generally, the pore diameter should be about 3 – 4 times larger than the diameter of the analyte in solution, to ensure sufficient accessibility.

Nonporous packings with very small particle diameters have developed into useful tools for fast and economic separations. Due to the reduced resistance to mass transfer of the nonporous material, separation efficiency is increased at high flow rates and consequently analysis times are reduced. To overcome the high back-pressure at increased mobile phase velocities, it is recommended to increase the mobile phase operating temperature. This lowers the mobile phase viscosity and helps to reduce the back-pressure generated over the separation column. However, the boiling point of the mobile phase at the corresponding operating pressure should always be taken into consideration to avoid gas bubbles in the pump or in the detector (see also Chapter 3.3). More detailed information regarding nonporous column packing material can be found in [68–71].

Short analysis times can be achieved by using short conventional columns (4.5 mm i.d. and in

20 to 100 mm in length) combined with higher mobile phase flow rates (e.g., 3 mL/min) and, if possible, by applying sharp gradient profiles, as described in [72].

4.1. Silica

Silica is the traditional and most common material used in normal phase chromatography and serves as the basis for many chemically bonded reverse phases. It is a highly porous, amorphous silicic acid in the form of small, rigid particles. A distinction is made between spherical and irregular particles. Traditionally, *irregular silica* was, and to a certain extent still is, prepared by precipitation of water glass with sulfuric acid (→Silica, Chap. 7.2.). Modern manufacturing techniques for *spherical silica* are, for example, sol – gel processing or spray drying of a silica sol; the latter is apparently the most common and economical technique today [73].

Commercially available silicas differ in their physical (specific surface area, mean pore diameter, pore volume, etc.) and chemical (contamination of different metals such as Al, Mg, Fe) characteristics. These metal contaminations can cause secondary retention effects (e.g., strong tailing of basic solutes). Additionally, commercial silicas differ widely in their Na content. The higher the Na content of the silica, the more alkaline its behavior in contact with water. This often acts as a catalyst for the hydrolysis of the Si – C bonds of chemically modified silica phases (see Section 4.3). Good silicas have an Na content of < 30 ppm. Excessively high contents of metal contaminants could easily be removed by the manufacturer by means of a thorough acid-washing step, which would, however, slightly increase the production costs of the material.

The chemical properties of the silica surface are determined by silanol (SiOH) and siloxane (Si–O–Si) groups. There are ca. 7 – 8 µmol SiOH groups per square meter of surface area (i.e., the average distance between silanol groups is 0.4 – 0.5 nm). These silanol groups are responsible for adsorptive interaction with the solute and are the active sites for the derivatization into chemically bonded phases (see Section 4.3). Silica is mildly acidic, with a pH of ca. 5. Due to its mildly acidic behavior, problems may occur with very sensitive solutes (acid-catalyzed reaction). At pH > 8, silica goes into solution and the column bed is destroyed.

4.2. Alumina

Although alumina is rarely used as the stationary phase in HPLC, in certain applications it could be preferred to silica (e.g., separation of isomeric aromatic hydrocarbons). In contrast to silica, natural alumina is a basic material (basic alumina) with a pH of ca. 12 and a specific surface area of ca. 200 m^2/g (depending on the pore-size distribution). Because of the high pH of the packing, acidic analytes (e.g., carboxylic acids) may show strong peak tailing and are sometimes retained irreversibly. Problems with base-sensitive samples on alumina columns have been reported [74]. Through special treatment, basic alumina can be converted into acidic or neutral alumina; both are commercially available. For adsorption chromatography, neutral alumina is preferred. Acidic alumina can be used as a weak anion exchanger; basic alumina, as a weak cation exchanger.

4.3. Chemically Bonded Phases

Chemically bonded phases are usually made by derivatization of the silanol groups of silica (see Fig. 25 A). Remaining unreacted SiOH groups can be removed by end capping with trimethylchlorosilane (TMCS) or hexamethyldisilazane (HMDS) (Fig. 25 B). The positive effect of end capping on the separation of basic solutes on a reversed-phase column is shown in Figure 26. A further advantage of end-capped material is that it is often more stable and therefore has a prolonged service life.

Depending on the size and reactivity of the silica matrix, a maximum of 50 % of all silanol groups present reacts with the silane. This is very important for RP packings because the interactions of basic analytes and the remaining silanol groups can result in severe peak tailing (see Fig. 26 A). Peak tailing can be suppressed by end capping or by the addition of a relatively basic component or a salt to the mobile phase.

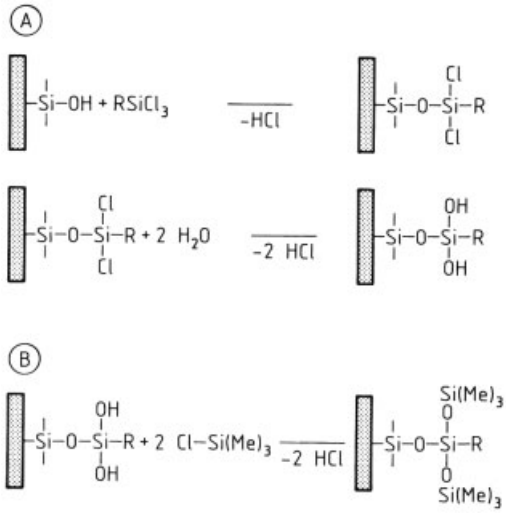

Figure 25. Example (A) derivatization of silica and (B) end capping

A variety of different chemically bonded phases is available (see Table 5). All of them have a stable Si–C bonding. If their organic residue is bonded to a function group (e.g., $-OH$, $-NH_2$, $-CN$, or $-NO_2$) the phases are often called selective polar phases or chemically bonded phases, and are used in both the normal phase and the reversed-phase chromatographic modes.

To ensure stable bonding, especially Si–C bonds, the polar functional group, which is responsible for the selectivity of the material, is connected to a spacer [$-(CH_2)_3$ group] that is fixed onto the silica. The result is a nonpolar stationary phase with a relatively polar surface. If the organic residue of a chemically bonded phase consists only of a hydrocarbon chain ($C_1 - C_{18}$), the stationary phase is very nonpolar or hydrophobic. Here, the elution order of hydrocarbon samples is reversed compared to normal phase chromatography (where retention decreases with increasing number of carbon atoms). There fore, these phases are called reversed phase and the chromatographic mode *reversed-phase chromatography*. Commercially available packing materials for RP columns have a chain length between C_1 and C_{18}. Most common is RP-8 and RP-18 material. RP columns of the same type (e.g., RP-18) from different manufactures, or even from different production batches of the same manufacturer, can differ dramatically in their chromatographic behavior because a different packing material (silica) is of-

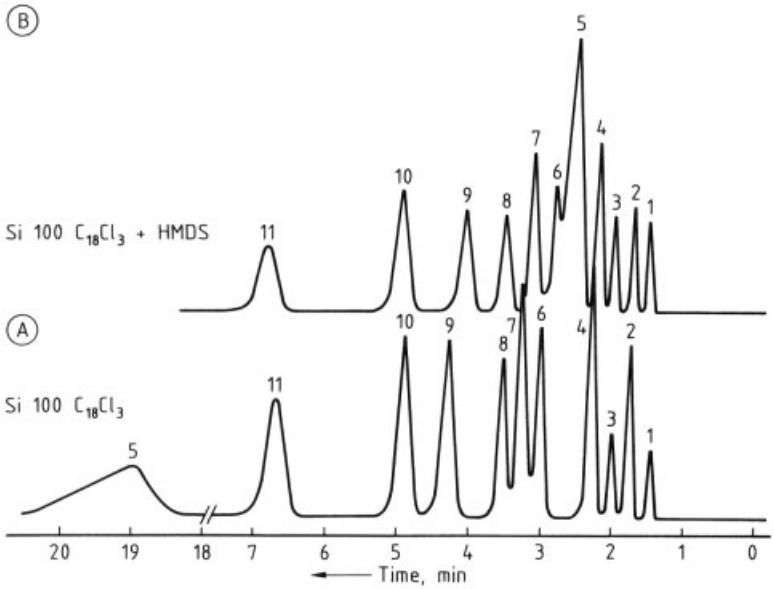

Figure 26. Separation of column performance test on a reversed-phase column
(A) Without end capping; (B) After end capping or basic component or salt to mobile phase
1) H_2O; 2) Benzamide; 3) Phenol; 4) 2-Phenylethanol; 5) Aniline; 6) Nitrobenzene; 7) Methyl benzoate; 8) Benzene; 9) Ethyl benzoate; 10) Toluene; 11) Ethylbenzene
Eluent: H_2O-CH_3OH 34.6 : 65.4 wt %

Table 5. Examples of chemically bonded phases and their application range

Chemical modification	Functional group*	Application range
Butyl	$-(CH_2)_3-CH_3$	RP and ion-pairing chromatography
Octyl	$-(CH_2)_7-CH_3$	RP and ion-pairing chromatography
Octadecyl	$-(CH_2)_{17}-CH_3$	RP and ion-pairing chromatography
Phenyl–diphenyl	$-C_6H_5-(C_6H_5)_2$	RP and ion-pairing chromatography, additional selectivity for $\pi-\pi$ interactions
Nitro	$-NO_2$	analytes containing double bonds
Nitrile (cyano)	$-CN$	NP and RP chromatography
Diol	$-CH_2-CH(CH_2-OH)-OH$	NP and RP chromatography
Amino	$-NH_2$	NP, RP, and weak anion exchanger
Dimethylamino	$-N(CH_3)_2$	NP, RP, and weak anion exchanger
Quaternary ammonium	$-\overset{+}{N}(CH_3)_3$	strong anion exchanger
Sulfonic acid	$-SO_3^{-+}$	strong cation exchanger

* Usually a spacer; for example, a propyl group $[-(CH_2)_3-]$ is placed between the silica surface and the functional group.

ten used. For instance, a higher natural alumina contamination of the silica increases the acidity of the Si–OH groups, which again strongly influences the retention of basic analytes (peak tailing).

The development of RP columns by Halasz and Sebastian [75] was the breakthrough for modern HPLC. Through that development, application of this technique was extended significantly (especially for biological materials), and because of the fast re-equilibration of the column, gradient runs could be performed easily, even in routine analysis.

4.4. Column Packing Techniques

Even if excellent equipment is used, but the sample is passed over a "badly packed" column, separation efficiency is poor. Therfore, the goal is to pack the chromatographic bed as densely and uniformly as possible. Care must be taken that no cracks or channels develop and that no fractionation of particles occurs during the packing process. Because experience and know-how are required to pack columns well, especially with small particles (<10 µm), purchase of a prepacked column is recommended.

If columns with relatively coarse particles (>30 µm) have to be packed, the *dry-fill packing process* [19] can be applied. Here vertical tapping and simultaneous rapping of the column during the packing process are necessary for well-packed chromatographic beds.

All analytical HPLC columns with particles < 20 µm (usually 5 or 10 µm) are packed by the *slurry technique* [19], which can be further classified into the high-viscosity technique and the balanced-density technique. In the latter, the particles are suspended in a fluid that has a density similar to theirs, so particle segregation by sedimentation decreases. In both techniques, the suspending fluid must be chosen in such a way that particle dispersion is maintained without aggregation and particle agglomeration is avoided by proper selection of the polarity of the slurry [19, 24].

The suspension is pressed into the empty column, which is fitted with a frit on its far side, by means of a constant-pressure pump (pressure filtration). The usual packing pressure for a standard analytical column (250×4 mm i.d.) is between 40 and 60 MPa. In this case, 1.7–2.0 g of stationary phase has to be suspended in ca. 30–50 mL of suspension fluid. Detailed discussion of column packing techniques is given in [73, 19, 76].

4.5. Column Specifications and Column Diagnosis

The plate number N is often considered the most important column performance specification. A typical plate number for a 250×4 mm i.d. column with 10-µm particles is ca. $N = 10\,000$. The plate number required for a specific separation can be achieved in two ways: By using either

Table 6. Typical reversed-phase column performance tests

Test name	Characterization of	Test compounds	Eluent	pH	Detection
DMD∗ test	$C_8 - C_{18}$ packing hydrophobic interactions	diphenylhydramine, 5-(*p*-methylphenyl)-5-phenylhydantoin, diazepam	acetonitrile – potassium dihydrogenphosphate/phosphoric acid butter, 156 : 340 wt %	2.3	UV (220 nm)
Engelhardt – Jungheim test	$C_8 - C_{18}$ packing hydrophobic interaction, behavior of basic and acidic analytes (silanophilic interactions)	aniline, phenol, *o*-,*m*-,*p*-toluidine, *N*,*N*-dimethylaniline, toluene, ethylbenzene, methyl benzoate	methanol – water, 50 : 50 wt %		UV (254 nm)
Aromatics test	$C_8 - C_{18}$ packing hydrophobic interaction, selectivity	benzene, naphthalene, fluorene, anthracene, benzanthracene	acetonitrile – water, 75 : 25 wt %		UV (254 nm)

∗ DMD = Diphenylhydramine,5-(*p*-methylphenyl)-5-phenylhydantoin, diazepam

a short column with small particles or a long column with larger particle diameter. As an example, the same separation efficiency can be obtained with 10-µm particles in a 250-mm-long column or with 3-µm particles in 50 – 100-mm-long columns.

Even though the plate number is a very useful column specification, itis certainly not sufficient to evaluate the suitability of a tested column for a given separation problem. Especially for RP columns, a variety of column performance tests have been developed [77]. These tests give important information about the quality of the column bed (well packed or not), the kinetic and thermodynamic performance, basic and acidic behavior, and the selectivity and stability (lifetime) of the column (see Table 6)

Especially in routine analysis, an individual test mixture, consisting of characteristic compounds of everyday samples, is often sufficient. This test mixture should be injected onto every new column (to see if selectivity and k' are unchanged) and again from time to time to follow the aging process of the column (peaks get broader, resolution gets worse, k' get lower, peaks become asymmetric, etc.). If one or more of these characteristics are observed, the old column should be replaced by a new one.

5. Separation Processes

5.1. Adsorption Chromatography

Adsorption chromatography or liquid – solid chromatography (LSC) is the most frequently applied LC mode. Here, specific interactions occur among the solute, the stationary phase, and the mobile phase. Examples are hydrophobic, dipole – dipole, hydrogen bond, charge-transfer, and $\pi - \pi$ interactions.

Every chromatographic separation should give sufficient resolution *Rs* and good reproducibility, which have great importance for qualitative and quantitative analysis. Good reproducibility in LSC can be achieved only in the range of a linear sorption isotherm. In this case the retention time t_R is independent of sample size and sample volume. Generally, a sample size of < 0.1 mg of analyte per gram of stationary phase ensures a linear sorption isotherm and, therefore, reproducible results.

Two groups of stationary phases are used in adsorption chromatography: Polar and nonpolar adsorbents.

5.1.1. Polar Adsorbents and Their Properties

The most frequent used polar adsorbents in LSC are silica and alumina (see Sections 4.1 and 4.2). Polar stationary phases are very selective with respect to the separation of configurational isomers. Figure 27 shows the separation of β-carotin isomers on an alumina column in NP chromatography.

Generally, polar adsorbents show a greater selectivity (higher retention) toward polar solutes, whereas the retention k' of nonpolar analytes decreases with increasing hydrophobicity of the analyte.

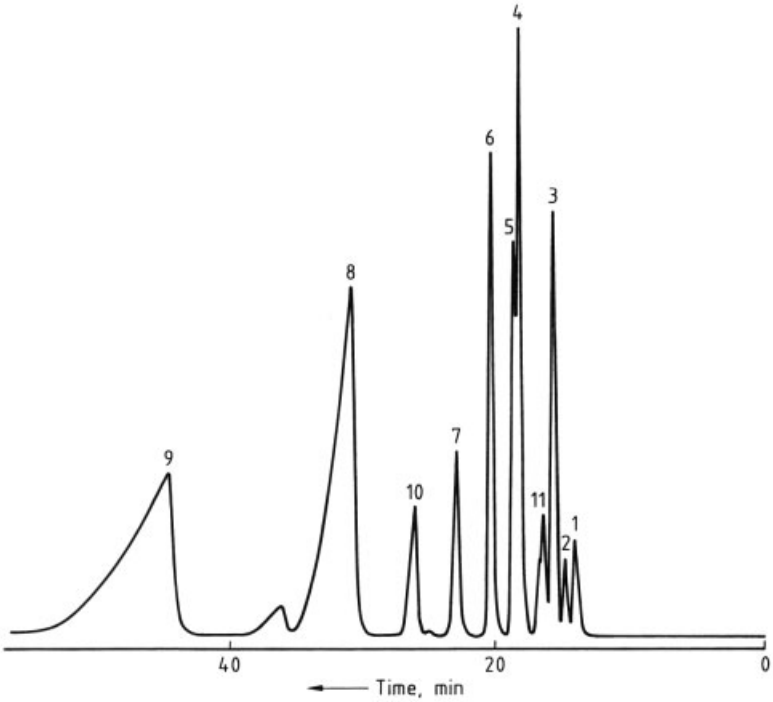

Figure 27. Separation of β-carotin isomers on an alumina column in normal phase chromatography [78]
1) 13,13′-di- c*is*; 2) 9,13,13′-tri- c*is*; 3) 9,13′-di- c*is*; 4) 15- c*is*; 5) 9,13-di- c*is*; 6) 13- c*is*; 7) 9,9′-di- c*is*; 8) all-*trans*; 9) 9- c*is*; 10) α-Carotin; 11) 3 a, b, c

Working with polar stationary phases and nonpolar (hydrophobic) eluents is the classical form of LC (which TSWETT [2] used in 1903) and is called *normal phase chromatography.* Currently, only ca. 20 % of analytical separations are solved by NP chromatography – for instance, chiral analyses – because re-equilibration of the column can be very time consuming (up to 12 h), and in certain cases, exact water-content control is necessary to obtain reproducible results. In NP chromatography, even traces of water (<100 mg/L) in the solvent cause a dramatic decrease in k'.

5.1.2. Polar Chemically Bonded Phases

Diol phases are often used as "polar" column packings. They belong to the group of polar chemically bonded phases, which also includes nitrile, nitro, amino, and similar polar groups. The diol phase is less polar than silica but is therefore easily wettable with water. Because of their rapid equilibration, these packings are much more suitable for gradient elution than unmodified silica. Diol columns are typically used for the separation of peptides, proteins, and generally components that can build up hydrogen bonds.

5.1.3. Nonpolar Adsorbents – Reversed Phase

In contrast to NP, in RP chromatography a nonpolar stationary phase and a polar solvent are used. In the reversed-phase mode the elution-strength order of the solvents is the opposite of that of the elutropic row (see Section 4.3 and Table 4).

Some separation problems (e.g., triglyceride separation [76]) require a hydrophobic stationary phase and a nonpolar (nonaqueous) mobile phase. This chromatographic mode is called *nonaqueous reversed-phase* (*NARP*) *chromatography.* The most common commercially available reversed-phase materials are C_2, C_4, C_8, C_{18}, and phenyl phases.

As shown in Figure 28 A the retention of a particular hydrophobic analyte increases with growing chain length of the RP material. The selectivity, on the other hand, reaches a plateau near the C_8 column (Fig. 28 B). This means that under identical conditions, the selectivity for the same sample is more or less the same on an RP-8 as on an RP-18 column. The only difference is that the compounds on the RP-18 column are more strongly retained than those on the RP-8 material. Therefore, many separations, which can be performed on an RP-18 column can also be achieved on an RP-8 column by adjusting the elution strength of the mobile phase (in this case by adding more water to the eluent) and vice versa.

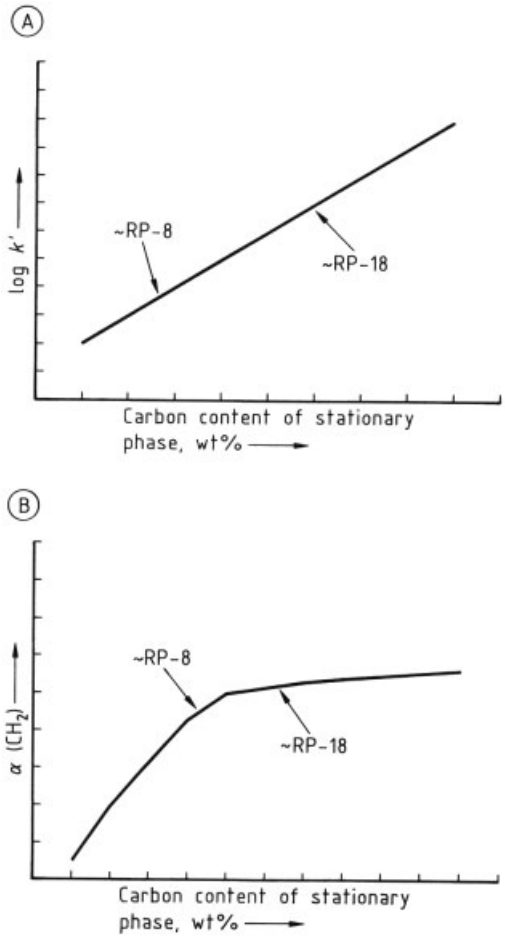

Figure 28. Retention behavior of hydrophobic analytes in RP chromatography
A) Retention; B) Selectivity

Generally, packing materials with shorter alkyl chains (e.g., RP-4 or RP-8) are more suitable for the separation of moderately to highly polar solutes, whereas those with longer alkyl chains (RP-18) are preferred for the separation of moderately polar to nonpolar analytes. Reversed-phase columns are also used for ion-pairing chromatography.

Today, between 50 and 80 % of HPLC applications are run in the RP mode and only 10 to 20 % in the NP mode [79]. This can be explained by the following reasons:

1) The reproducibility of the elution strength of the mobile phase is easier in RP chromatography (more reproducible separations are possible)
2) Re-equilibration between the mobile and stationary phases is much faster in the RP mode (after three to five column volumes) than in NP chromatography, especially if water was added
3) Because of item 2, gradient elution is often not used in NP chromatography
4) RP columns combined with gradient elution are able to separate a wide range of analytes (from hydrophobic to polar components)

Other modes, e.g., IEX (Section 5.3) and SEC (Section 5.4) are used to a much lesser extent.

The application range of RP chromatography is so large that it is impossible to mention it all. In principle, the technique is applied for nearly all separations in the pharmaceutical, environmental, and biological fields.

5.1.4. Other Adsorbents

Polyamide as a stationary phase is used mainly for the separation of compounds that can form hydrogen bonds to the amide group. A typical example is the separation of phenolic compounds on polyamide, due to hydrogen bonds between the phenolic hydroxyl and the amide groups [80]. The drawback of these polyamide packings is that they are not as pressure stable as silica and that undesired swelling of the phase can occur with certain solvents.

Magnesia and magnesia – silica (florisil) are also sometimes used in adsorption chromatography. Magnesia – silica has the disadvantage

of retaining many solutes irreversibly (e.g., aromatics, amines, and ester). Possible applications are in the separation of lipids or chlorinated pesticides [80].

Titanium dioxide is also an interesting packing material because of its good mechanical and chemical (pH) stability. Its main drawback is the relatively small surface area (ca. 50 m^2/g). However, this is partially compensated by its high packing density of 1.5 g/mL, which compares to only 0.6 g/mL for silica.

5.2. Partition Chromatography (LLC)

Partition chromatography, or liquid – liquid chromatography (LLC), was first mentioned in 1941 in a paper by Martin and Synge [5]. This is a well-known technique in gas chromatography, where the separation is based on distribution of the analyte between a "liquid" stationary phases and a gaseous mobile phase. In liquid chromatography, the distribution takes place between two liquids: The liquid stationary phase, which is dispersed on an inert support (e.g., silica), and the liquid mobile phase. As in the distribution of solutes between two mobile phases in a separatory funnel, the stationary and mobile phases in LLC must be immiscible. However, liquids are always miscible to a certain extent, which results in a washing out of the stationary phase form the column (" column bleeding").

A major difference between adsorption and partition chromatography is that in LLC the separation of a sample is based not on different adsorption energies of the analytes but on different solubilities of the sample components in the liquid mobile and stationary phases. This can be a disadvantage if binary systems (stationary and only one solvent as mobile phase) are used. If the solute is much more soluble in the liquid stationary phase, the retention will be strong, resulting in increased dilution and band broadening; thus, the sensitivity is poor. On the other side, if the solute is more soluble in the mobile phase, the retention will be very weak and the solute peak will elute shortly after the inert peak. To reduce these relatively large differences in polarity (solubility) between the two liquids, a third liquid ("modifier") can be introduced to form a ternary system. In this case the mobile phase is a mixture of two solvents. Now, the more strongly retained solutes elute earlier, with better peak symmetry and smaller elution volumes, thereby increasing the sensitivity significantly. A drawback of this ternary system technique is that the stationary phase also becomes more "soluble" in a solvent containing a modifier, which reduces the column lifetime. Furthermore, even a slight temperature increase can cause total miscibility of all three liquids, which destroys the chromatographic system instantly. Such well-thermostated chromatographic equipment (which also includes the solvent reservoirs) is essential (temperature variations 0.1 – 0.2 °C). If a UV detector is used, neither the mobile nor the liquid stationary phase must absorb at the detector wavelength. A further disadvantage of LLC is that gradient elution is impossible. So if a sample contains components with widely different k' ranges, adsorption chromatography is preferred.

The advantages of partition chromatography compared to adsorption chromatography are as follows:

1) Easy exchange of the stationary phase (in situ) and longer column life
2) High loadability (10 times higher than LSC, ca. $10^{-3} - 10^{-2}$ g of analyte per gram of liquid stationary phase)
3) Reproducible results because the supporting material of the column is not, or at least should not be, involved in the separation process
4) Broad range of selectivity due to the large number of different mobile and stationary phase liquids available (with the condition that the liquids are immiscible)

When selecting the stationary phase in LLC, liquids with lower viscosity should be preferred because the use of high-viscosity liquids leads to poorer separation efficiencies due to imparted diffusion and thus sample transport (mass transfer). In addition, the mobile phase must be saturated with stationary phase to reduce column bleeding. Column life can be further extended by the use of a precolumn for proper equilibration and exclusion of oxygen from both phases.

Column Packings and Preparation. Inert materials with relatively large pores (10 – 50 nm) are used (usually silica) as liquid stationary phase supports. Material with larger

pores is not suitable because, unlike in narrow pores, the stationary liquid phase is fixed less strongly to the support by capillary forces. Wide-pore supports start to bleed at low flow rates (<1 mL/min). Materials with small pores (<50 nm) are usually stable up to flow rates of 5 mL/min.

Typical stationary phases are, for instance, β,β'-oxydipropionitrile (BOP) or different types of poly(ethylene glycol). Other examples of stationary and mobile phases for LLC are given below:

Stationary phase	Mobile phase
Normal phase LLC	
β,β'-Oxydipropionitrile	pentane, cyclopentane, hexane, heptane, ...
Carbowax	modified with up to 10 – 20 % chloroform, THF
Triethylene glycol	dichloromethane, acetonitrile, ...
Ethylene glycol	nitromethane
Ethylenediamine	hexane
Water	butanol
Reversed-phase LLC	
Polydimethylsiloxane	acetonitrile – water
Heptane	methanol – water

The stationary phase in LLC can be prepared in different ways, depending on the support particle size chosen. An often-applied preparation technique is the in situ coating method. In this case the stationary phase liquid is pumped through the dry packed (slurry packing) column. Later, the column is flushed with mobile phase to displace excess stationary phase liquid between the particles.

Separation Modes. Like adsorption chromatography, partition chromatography can also be used in the normal and reversed-phase mode.

In *normal phase LLC* the stationary phase is polar and the mobile phase consists of nonpolar solvents. This mode is suited for more polar, water-soluble analytes. The elution order is similar to that in normal phase adsorption chromatography because nonpolar solutes prefer the moving phase and elute with low k' values. Solutes preferring the polar stationary phase elute later with high k'.

In *reversed-phase LLC* the two phases are interchanged and the elution order reversed. This mode is more suitable for hydrophobic analytes that have poor water solubility. For reversed-phase LLC, the support should be hydrophobic to ensure appropriate fixing of the stationary phase liquid. This can be achieved, for instance, through silanization of the silica.

Applications. LLC can be a good alternative to LSC for analytes that are sensitive to catalytic decomposition on the active surface of the column packings. Typical applications of LLC are in the separation of compounds containing metal ions, radioactive metals, and free and underivatized steroids, as well as in preparative chromatography.

Because of the high loadability of the column (10^{-3} – 10^{-2} g of solute per gram of stationary liquid), LLC is often used for preparative chromatography. Since the mobile phase is saturated with stationary phase, high-boiling stationary phases may be difficult to remove from separated fractions and volatile stationary phases are preferred.

To ensure good chromatographic behavior (as in adsorption chromatography), samples should not be injected in solvents with a higher solution power than the mobile phase. Further, the injected sample volume should not exceed one-third the elution volume of th first peak of interest.

5.3. Ion-Exchange Chromatography (IEX) (see also, →Ion Exchangers)

Ion-exchange chromatography (IEX) can be regarded as a special form of adsorption chromatography resulting from the ionic interactions taking place during the separation process. It was first described by SMALL et al. [81] in 1975. Even in the 1980s, IEX was not very well accepted in HPLC because of unstable column packings and detection problems. These drawbacks have been overcome, and today IEX is a very powerful tool in water analysis, amino acid analysis, and the separation cleanup of proteins, for example. Generally all charged solutes (inorganic ions, organic acids, bases, salts, etc.) can be separated by this chromatographic method.

The following abbreviations are often used in ion chromatography (IC):

SCX	strong cation exchanger
WCX	weak cation exchanger
SAX	strong anion exchanger
WAX	weak anion exchanger

IEX	ion-exchange chromatography
IEC	ion exclusion chromatography
IPC	ion-pair chromatography
HPIC	high-performance ion chromatography

The principle of ion-exchange chromatography is as follows: Ions differing in their charges interact with the functional groups of the exchanger with different intensities, and can be selectively eluted from the matrix by changing the elution strength of the mobile phase (variation of ion concentration, i.e., "salt gradient" or pH gradient).

During the ion-exchange process an equilibrium between, for example, the cation in solution and the anion fixed on the matrix is formed. For a particular ion this equilibrium is described by the equilibrium constant K_D. The equilibrium constant depends on the ion size, its charge, and its polarizability; i.e., the affinity for the stationary phase (ion exchanger) increases with

1) Increasing charge
2) Increasing polarizability
3) Decreasing size of the ion investigated

Usually this equilibrium is shifted toward the side of the fixed ions on the exchanger material. Therefore the effluent in IEX always contains ions (buffers, acids, bases) that compete with the analyte for active sites on the matrix so that the analyte can be desorbed from the matrix. The ionic strength of a mobile phase in IEX is a measure of its elution strength. In addition, the pH and dissociation rate of the analyte also have a great influence on retention. Furthermore, the technique shows a much better selectivity for cations differing in their charges than for anions.

Instrumentation. A schematic of an ion-exchange chromatograph is given in Figure 29. In principle, this is the same setup as for a normal HPLC system consisting of a pump (b), sample introduction system (c), separation column (d), and detector (e). IEX differs from normal HPLC in the "suppressor column" that is installed between the separation column and the detector (e.g., conductivity detector). This suppressor column has two tasks: First, it reduces the high background conductivity of the effluent; second, it converts the analytes into species with higher conductivity. The function of the suppressor column can best be explained by a practical example:

The Cl^- content of a water sample is to be determined. The mobile phase consists of an $Na_2CO_3 - NaHCO_3$ buffer solution. The reactions taking place in the suppressor column – which is a strong cation exchanger (SCX) in the protonated form – are the following:

1) Background conductivity is reduced by exchanging the sodium ions of the buffer with protons. The resulting H_2CO_3 is nonconducting
2) At the same time, NaCl is converted into the strong contucting HCl :

$$\text{Resin}-SO_3^-H^+ + NaHCO_3 \longrightarrow \text{Resin}-SO_3^-Na^+ + H_2CO_3$$

$$\text{Resin}-SO_3^-H^+ + NaCl \longrightarrow \text{Resin}-SO_3^-Na^+ + HCl$$

This technique is often referred to as *suppressed conductivity detection*.

Because of the corrosive solvents used in IEX, conventional stainless steel HPLC instrumentation is unsuitable. The pump-head, capillaries, and column should be made of inert material such as PEEK, PTFE, tetrafluoroethylene – ethylene copolymer, polypropylene, ruby and sapphire. Otherwise, contamination problems (especially iron contamination) can occur in the mobile phase, decreasing the separation efficiency and capacity, and increasing the noise of the amperometic detectors.

Detectors. In ion chromatography, suppressed conductivity detection is still the most frequently used technique. Indirect UV detection (Section 2.6.3) can also be a powerful tool for the detection of ions, although with respect to sensitivity, conductivity detection (detection power 10 – 100 ng/g) cannot be surpassed. In amino acid analysis, UV and fluorescence detection (after post- or precolumn derivatization) is often applied.

Column Packings. Modern column packings for ion-exchange chromatography are made of a polystyrene matrix with different functional groups, depending on whether it is a cation or an anion-exchange resin.

The following functional groups are used:

SCX:	$-SO_3^-$
WCX:	$-COO^-$
SAX:	$-CH_2NR_3^+$
WCX:	$-CH_2-NHR_2$ or $-CH_2-NHR$

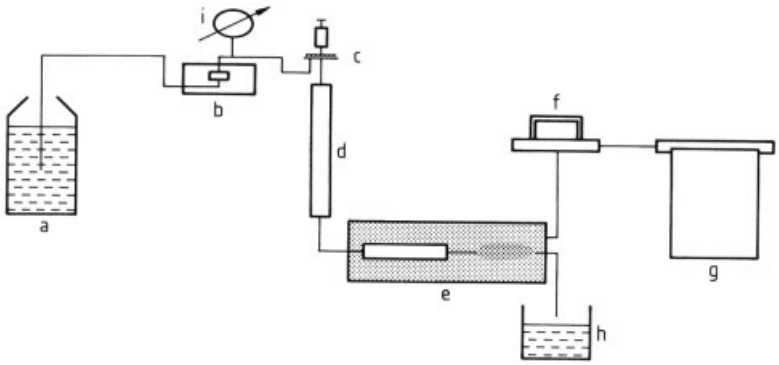

Figure 29. Schematic setup of a simple ion-exchange chromatographic system
a) Solvent; b) Pump; c) Injection port; d) Separation column; e) Detector with suppressor column; f) Control unit; g) Recorder; h) Waste; i) Manometer

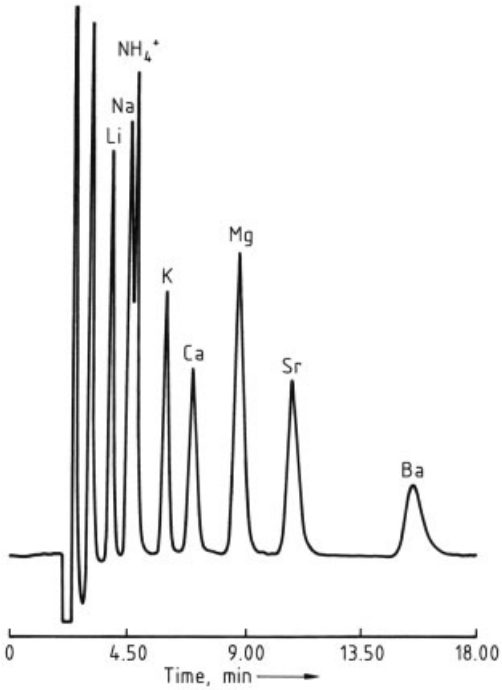

Figure 30. Simultaneous isocratic analysis of alkali and alkaline-earth cations with ion-exchange chromatography
Column: Polyencap WCX; eluent: 4 mmol/L citric acid; 0.7 mmol/L dipicolinic acid in water, 1 mL/min; standard solution: 2.5 – 40 µL/L in water; detection: conductivity

Polystyrene column packings are also available with small particle sizes (<10 µm). Compared to inorganic ion exchangers (silicates or zeolites), they have the advantage of being pH resistant over the entire pH range. Many protein separations must be carried out at a pH of ca. 10. A silica-based exchanger would be destroyed at this basic pH, but polystyrene-based exchangers are not.

A disadvantage of the polystyrene matrix for high-performance ion chromatography is the fact that it tends to swell or shrink if certain solvents are used.

An important characteristic of every ion-exchange material is its "total exchange capacity," expressed in equivalents per unit weight (microequivalents per gram) (→Ion Exchangers, Chap. 3.2.). One equivalent corresponds to 1 mol charge. For weak ion exchangers the capacity is a function of pH, whereas the capacity of strong ion exchangers is independent of pH. Figure 30 shows a simultaneous isocratic separation of alkali and alkaline-earth cations obtained by IEX.

5.4. Size Exclusion Chromatography

Although separation due to size exclusion has been known for a long time [82], systematic investigation of this method began only after the development of synthetic matrices [83, 84]. If the mobile phase consists of or contains water, size exclusion chromatography (SEC) is often called *gel filtration chromatography* (GFC). If the mobile phase is an organic solvent, the term *gel permeation chromatography* (GPC) is often used. Another frequently used abbreviation is HPSEC, which stands for high-performance size exclusion chromatography.

In SEC an analyte is, by definition, separated only according to its molecular size, not its

molecular mass. This is important to remember since the same solute (with its specific molecular mass) can have a different structure, and therefore different sample size, in different solvents. If all other interactions (ionic interactions, adsorption, ect.) with the stationary phase are suppressed, the selectivity of the system depends entirely on the different accessibility of the pores of the matrix for the solutes. Very small molecules (usually solvent molecules) have access to the entire intra particle volume V_i of the stationary phase. In contrast to the moving interparticle volume V_m (which is the effluent volume outsite the particle pores) the intra particle volume V_i is called the stagnant mobile phase. The total solvent volume V_0 inside the column is therefore

$$V_0 = V_m + V_i$$

Since solvent molecules usually have the smallest size, they elute at the end of an SEC chromatogram with a retention time t_0. In other words, they completely permeate the stationary phase ("total permeation"). In all other chromatographic modes the solvent elutes first, if no exclusion has appeared. In SEC, molecules that are so large that they cannot get into the pores stay only in the faster-moving interparticle volume V_m and elute as the first peaks of the chromatogram. For these large analytes the column behaves as if it were filled with nonporous glass beads. The difference between the retention volumes of the largest (excluded) and smallest (solvent peak) molecules is the pore volume or interparticle volume V_i of the stationary phase. Medium-sized molecules permeate the stationary phase only partially, and consequently elute between the total permeation t_0 and the totally excluded solutes (dead or interparticle volume). Generally, the smaller a molecule, the higher is the accessible pore volume and therefore the longer is the time required to pass through the column.

In SEC the retention volume V_R is used instead of the retention time t_R. V_R is defined as the sum of the interparticle volume V_m and the accessible pore volume V_p for the particular solute:

$$V_R = V_m + V_p$$

For very small molecules (e.g., solvent) V_R is

$$V_R = V_m + V_p = V_0$$

For large and therefore excluded particles, V_R is

$$V_R = V_m$$

To become independent of various packings, column volume, and pore volume, and to enable a comparison of data from different packings, a distribution coefficient K_0 was introduced

$$K_0 = \frac{V_R - V_m}{V_i}$$

The above equation shows that the solvent (last eluting peak) has a K_0 of 1. Consequently, the entire sample must elute with $K_0 < 1$. Because all components of a sample elute so early, band broadening is reduced, which results in high and symmetric peak shapes. Thus, even relatively nonsensitive but therefore universal detectors such as the refractive index detector can be used. Since the solvent peak usually elutes at the end of an chromatogram the analysis time is predictable.

Column Packings in SEC. Two types of packings exist: Rigid packings, which are essential for HPSEC because of their stability, and nonrigid gels. The latter are not resistant to higher pressure and are therefore used in open-column SEC.

In *HPSEC* the most common column packings consist of pure or chemically modified silica or, highly cross-linked macroporous styrene–divinylbenzene copolymer. Due to the differing production processes for these polymer phases (e.g., amount of inserted cross-links), various pore sizes are available with small (5 and 10 µm), pressure-stable (maximum pressure 10–15 MPa), spherical particles [85]. The advantage of these polymer-based columns compared to silica-based columns is that under certain conditions (e.g., pH > 4–5), silica can act as a weakly acidic ion exchanger. Interactions between the sample and the active silanol groups (e.g., adsorption) are also possible and must be suppressed by proper choice of effluent. Ion exchange, for instance, is suppressed by increasing the ionic strength of the eluent (e.g., inserting a salt or buffer, 0.1–0.5 mol/L). Adsorption effects resulting from silica can be re-

duced by modifying the silica with chlorotrimethylsilane [86]. An interaction between the analytes and the column packing is indicated either by peak tailing (uncertain) or by the elution of any components after the solvent peak ($t_R > t_0$ or $K_0 > 1$). Nevertheless, silica-based SEC column also have their advantages. Here, pore size and pore volume are independent of the effluent. A silica column can be calibrated and characterized, for instance, with a polystyrene standard in an organic solvent (e.g., CH_2Cl_2); then, the organic solvent is replaced by a water-containing effluent, water-soluble polymers can be investigated without recalibration. In contrast, the pore size and thus the pore volume of polymer-based materials often change with the mobile phase applied. In some cases, a change of solvent can lead to a collapse of the particles, which results in a significant decrease of performance. When polystyrene gels are used the mobile phase must be degassed very well, because air bubbles can damage the packing. In addition, aromatic sample compounds may also form $\pi - \pi$ interactions with the styrene – divinylbenzene copolymer.

To obtain high separation efficiency, similar to other LC techniques, small particles ($\leq 5\,\mu m$) and a narrow pore-size distribution are necessary. Common pore sizes in SEC for low molar mass samples are between 6 – 12 nm, whereas high molecular mass samples (e.g., proteins) require a pore diameter from 30 – 100 nm or wider (e.g., 400 nm). Since separation capacity is correlated directly with pore size and therefore pore volume, the latter should be as large as possible. To enlarge the pore volume of an SEC system, either the sample can be reintroduced into the same column after the first separation (recycle SEC) or several columns with the same packing can be connected in series. Doubling the column length doubles the elution volume and the analysis time. If the sample contains molecules with a molecular mass that exceed the separation range of the column, several columns with differing sep-aration ranges can be connected for the "first analysis" of an virtually unknown sample, to obtain maximum information about the molar mass distribution. Two or more column packings differing in pore size should not be inserted into the same column because particle sizing can occur, which decreases the separation efficiency. A problem with columns connected in series is that band broadening may occur. This can be reduced by avoiding any unnecessary dwell volumes, by using small particles and solvents with low viscosities (higher diffusion), and by working at higher temperature (high-temperature gel permeation chromatography).

For choice of the appropriate solvent in SEC the following rules apply:

1) The sample must be soluble in the mobile phase
2) The stationary phase must be resistant to the mobile phase
3) Possible interactions between solute and column packing should be suppressed by appropriate additives to the mobile phase or by control of its pH, ionic strength, etc.
4) The mobile phase should be more strongly adsorbed to the column packing than the solute
5) The maximum permitted solute concentration in SEC should not exceed 0.25 wt %; a concentration of 0.1 wt % is sufficient in most cases

Applications. If all interfering interactions between solute and stationary phase are suppressed, SEC is a very simple, reliable, and fast technique for the separation of many samples, especially oligomers and polymers. The only limitation is that the solutes to be separated must differ by at least 10 % in their molecular masses. Typical applications of SEC:

1) Determination of molar mass distribution of polymers
2) Determination of pore-size distribution of column packings
3) Cleanup technique prior to other high-resolution methods
4) Fingerprint analysis
5) Biopolymer separations (e.g., proteins)

To determine the elution (retention) volume V_R that corresponds to a certain molecular mass, the column must be calibrated with a standard mixture of known molar mass distribution. These standard, high-purity mixtures are available from various column suppliers. Remember that molecule size can change in different solvents and that different solutes (even with comparable molecular masses) can show differing molecular structure and size in the same solvent. In both cases, the retention volume obtained gives an incorrect molar mass. Therefore,

the standard must be chosen in such a way that it corresponds as much as possible to the sample structure. Figure 31 shows a typical schematic calibration curve obtained in SEC.

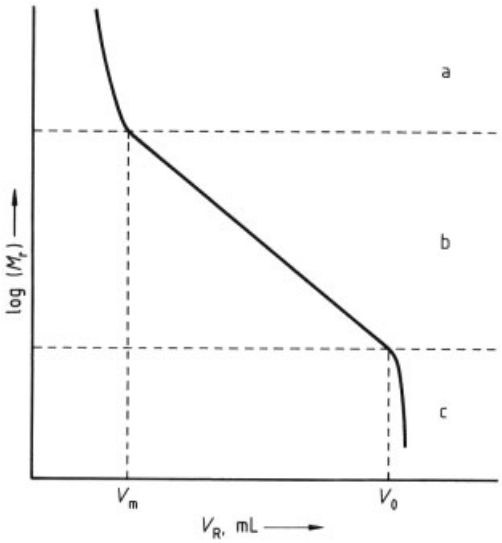

Figure 31. Typical calibration curve in SEC
a) Molecules are too large and elute with interparticle volume; b) Linear range for separation according to the molecular mass of the sample; c) Molecules are too small and penetrate the entire packing

Because retention volume is the most important factor in SEC (it is related to the logarithm of the molecular mass, even small inaccuracies in the determination of V_R are magnified), the pump must deliver a very constant flow which can be controlled by a continuous flow measurement system, for example. Further information on SEC can be found in [19, 24, 87, 88].

6. Gradient Elution Technique

In gradient elution the mobile phase is usually composed of two solvents, A and B. The characteristic of gradient elution is that the composition of the mobile phase is changed during the chromatographic run. Usually the concentration of the stronger solvent B (percent) is increased during the run. In the reversed-phase systems that are generally used, solvent A is normally water.

The three main reasons for using gradient elution are:

1) Separation of sample components showing a wide range of k'
2) Fast evaluation of optimum eluent composition for subsequent isocratic analysis
3) Enhancement of detection limit

In addition, the technique can be applied to remove strongly retained solutes or impurities from the column, thus enhancing sensitivity and column life.

Gradient elution is usually applied in reversed-phase chromatography, although for certain separation problems in normal phase chromatography it can also be a very useful tool. Ideally, the solvent strength of the mobile phase should be adjusted in such a way that the k' range of the sample compounds lies between 2 and 10 (although k' values between 1 and 20 are often acceptable). For small molecules in reversed-phase liquid chromatography a 10 % increase in the percentage of B (organic solvent) decreases the k' value of every solute by a factor of ca. 2.

Compared to the isocratic elution mode, the equipment necessary for gradient elution is extended by a second solvent reservoir, an additional pump (in the case of a high-pressure-mixing gradient system), a mixing chamber, and a proportioning valve that is regulated by a gradient control system.

Two types of gradient systems are possible (see Fig. 32 on next page): The low-pressure-mixing gradient system and the high-pressure-mixing gradient system. In the first case, only one HPLC pump is necessary because the two solvents (A and B) are mixed before the pump (i.e., in the low-pressure area of the system). With the high-pressure gradient system, two pumps are required because solvents A (e.g., water) and B (e.g., organic solvent) are mixed after leaving the pump in the high-pressure area of the system.

Many different gradient profiles are possible (see Fig. 33); the most frequent and useful are the linear and segmented (multilinear) profiles. Besides binary gradients (two solvents), some separations require ternary gradients in which the mobile phase consists of three solvents (A, B, C). With these ternary gradients, special selectivity effects can be achieved [89].

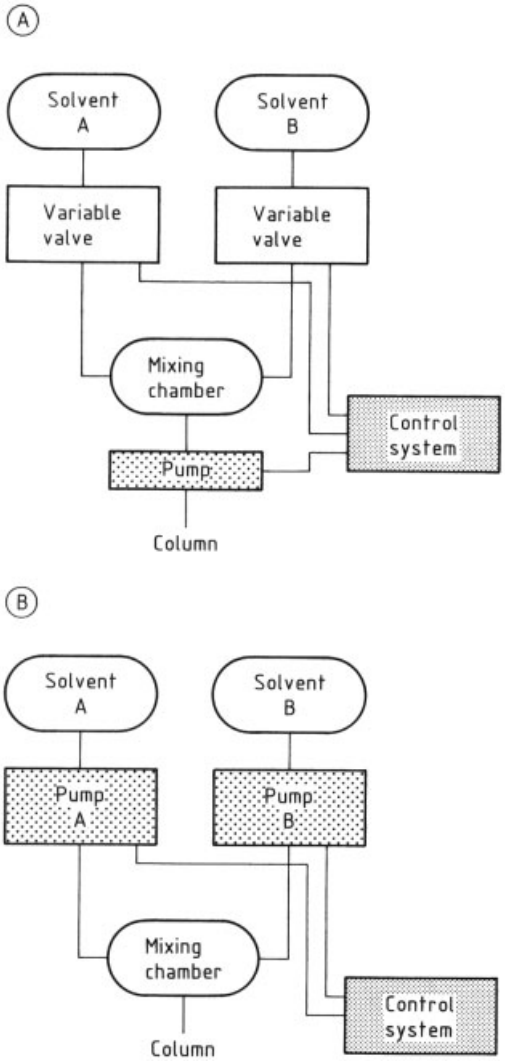

Figure 32. Gradient elution technique
A) Low-pressure system, B) High-pressure system

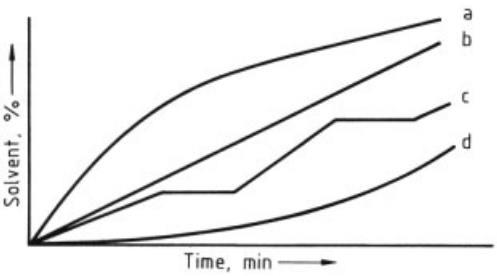

Figure 33. Different possible gradient profiles
a) Convex; b) Linear; c) Step; d) Concave

7. Quantitative Analysis

Quantitative analysis should deliver reproducible and reliable information about the composition of the sample investigated. Since end of the 1980s, quantitative analysis has become extremely important for product quality control (e.g., food industry, pharmaceutical industry) and environmental analysis. It is of great economic importance, and an erroneous quantitative analysis can result in fatal consequences. Therefore, this chapter introduces different quantification methods for HPLC and describes some often-occurring sources of error.

Correct sampling and sample storage are essential to obtain exact and reliable analytical results. After a possible cleanup step, the sample is injected into the chromatographic system and a strip chart is obtained from the recorder or electronic integrator, on which the intensity of the detector signal is plotted against analysis time (chromatogram). To be able to make any quantitative statements with the help of this chromatogram, the system must have been previously calibrated. The two most important methods are the external standard method and the internal standard method.

7.1. External Standard Method

The external standard method is the technique used most frequently to gather quantitative information from a chromatogram. In this case, a pure reference substance (ideally the same compound as the one to be determined in the sample) is injected in increasing concentrations and the peak areas or peak heights obtained are plotted versus the concentration (calibration curve; →Chemometrics, Chap. 12.). These calibration curves should show a constant slope (linear curves), and the intercept should be as close to zero as possible. Since the calibration curves usually show nonlinear behavior and flatten off at higher concentrations (see Section 2.6), the quantification should be carried out only within the linear part of the curve.

A calibration curve must be recorded for each component of interest. The peak areas/heights

of the sample chromatogram are then compared with those in the calibration curve. All chromatographic conditions must remain absolutely constant, for both sample and reference substances. If the volume of injected sample and injected reference solution is constant (which is usually the case if a sample loop is applied), the calculation can be carried out as follows:

$$\frac{A_{\mathrm{s}i}C_{\mathrm{ref}}}{A_{\mathrm{ref}}} = C_{\mathrm{s}i} \tag{1}$$

where $A_{\mathrm{s}i}$ is the area or height of component i in the sample; A_{ref} the area or height of the reference peak in the standard solution; C_{ref} the concentration of reference substance in the standard solution in milligrams per milliliter; and $C_{\mathrm{s}i}$ the concentration of component i in the sample solution (milligrams per milliliter).

The amount (percent) of component i in the sample is:

$$\frac{A_{\mathrm{s}i}C_{\mathrm{ref}}V_{\mathrm{s}}}{A_{\mathrm{ref}}M_{\mathrm{s}}} \cdot 100 = \%C_{\mathrm{s}i} \tag{2}$$

where V_{s} is the volume of the sample solution in milliliter; M_{s} the mass of sample in volume V_{s} in milligrams; and $\%\,C_{\mathrm{s}i}$ is the percentage component of i in the sample.

7.2. Internal Standard Method

The internal standard method is well known from GC (→ Gas Chromatography, Chap. 7.) and should always be used if exact control of the injection volume is not possible. A defined amount of a chosen compound (internal standard) is added to the sample, so that two peaks (that of the internal standard and that of the compound of interest) must be measured to obtain one result. An internal standard must fulfill several requirements: It must not already be present in the sample; it should elute close to the peak of interest; and the two peak sizes should not differ too much. If possible, peaks should not overlap, and no chemical interaction between the internal standard and other sample components should occur. The internal standard must also have good storage stability and be available in high purity. This method is slightly less accurate than the external standard, but it has several significant advantages:

1) Compensation of varying injection volumes
2) Compensation of recovery of sample during sample cleanup, especially in trace analysis
3) Subjection of the standard to the same chemical reaction (e.g., amino acid analysis) in postcolumn reactions

Calculations for the internal standard method are similar to those mentioned above for the external standard. More detailed information concerning the external and internal standard methods, as well as other methods of evaluation [e.g., area normalization method (100 % method), addition method] is given in [11].

7.3. Error Sources in HPLC

(→Chemometrics)

Error sources in HPLC are so manifold that they cannot all be discussed completely. Nevertheless, one can distinguish between random errors and systematic errors. Whereas random errors are relatively easy to recognize and avoid, systematic errors often lead to false quantitative interpretation of chromatograms. Two simple ways of recognizing a systematic error in a newly developed (or newly applied) analytical method are either to test this new method with certified materials or – if an "old" (i.e., validated) method already exists – to compare the result from the old and the new methods. If the new method shows good reproducibility, a correction factor can be calculated easily by

$$C_{\mathrm{f}} = \frac{A_{\mathrm{certified}}}{A_{\mathrm{new}}} \tag{3}$$

where C_{f} is the quantitative correction factor; $A_{\mathrm{certified}}$ the peak area/height or concentration of component i in the certified reference sample or the "old" method; and A_{new} the peak area/height or concentration of component i obtained with the "new" method. Multiplying the calculated amount of component i (see Eq. 1) in the sample by the correction factor C_{f} compensates for the difference that can be caused by one or more systematic errors. (If possible, however, the systematic error should be identified and eliminated.)

Sample Injection. For every quantitative HPLC analysis, especially in the external standard mode, a constant volume of injected sample solution is essential. This can be achieved only by using sample loops (5, 10, 20 µL, etc.) or automatic sample introduction system (see Section 2.4). The constancy in injection volume of these systems has to be checked from time to time. This is done by injecting the same sample several times and comparing the peak areas and peak heights generated. If a constant injected sample volume cannot be ensured, the internal standard method should be applied to compensate for these variations. A variation of the peak area can also result from short-time pulsation of the pump.

Another error in quantification can be caused by overloading the column (rule of thumb: A maximum of 1 mg of analyte per gram of stationary phase). The overloading of a chromatographic column is shown by the appearance of decreasing peak symmetry (fronting or tailing), peak broadening, and decreasing retention time. Obviously, loss of the sample through leakages or strong adsorption to the stationary phase must be eliminated. The latter is often the reason for a baseline rise or for nonlinear behavior of the calibration curve in the lower concentration range.

Pump. A constant flow rate of the mobile phase is another essential condition for quantitative analysis. Especially for peak area quantification, even small variations in flow can have a disastrous effect because the recorder or integrator plots the concentration (signal intensity) versus analysis time (not versus flow rate). Therefore a variation in flow automatically influences the recorded peak area. Quantification over peak height is not so strongly influenced by fluctuations in the flow of the mobile phase.

As a rule of thumb, peak area calculations are sensitive to baseline variations, and therefore many trace analyses are quantified over the peak height. On the other hand, peak height is influenced strongly by retention time (the longer the retention time, the greater is the dilution of the sample and therefore the weaker is the detector output signal). Generally every variation of flow (even very slow variations – often recognizable by a baseline drift) must be avoided for quantitative analysis.

Detector. If possible, the use of highly selective detectors (e.g., fluorescence detector) should be preferred in quantitative analysis because sensitivity is usually enhanced and, due to the high selectivity, peak overlapping is often avoided. However, most HPLC detectors (even the selective ones) are photometric detectors that convert the light absorption of an analyte in the mobile phase into an electrical signal, which is recorded as a function of time. Therefore, their linearity is limited by the Lambert – Beer law (see Section 2.6). The influence of temperature on detection (UV, fluorescence, refractive index, etc.) should not be neglected. The solute concentration in the detector cell is indirectly a function of temperature because the density of the mobile phase is a function of temperature [90]. Additionally, the position of the absorption maxima changes with temperature. Generally, for quantitative analysis, temperature variations must be avoided (in hot countries, daily temperature variations of up to 10 °C are quite common) by thermostating the chromatographic system including the detector cell.

Processing of Data. The chromatogram obtained contains qualitative (t_R and k') and quantitative (peak area or height) information. The quantitative information can be evaluated manually or automatically (by an integrator or computer system).

If the baseline shows any drift, *peak heights* are often more correctly determined manually than by an automatic system (important for trace analysis). The *peak area* is very seldom calculated manually, and if it is, a peak area equivalent is used, which is the product of peak height and peak width at baseline. In older publications, the peak area was often calculated by cutting out the peak and weighing it [88]. This method also gave very good results.

8. Sample Preparation and Derivatization

8.1. Sample Cleanup and Enrichment

Usually, before a sample is injected into the chromatographic system it has to undergo a

cleanup procedure. In the simplest case, cleanup consists of only a filtration step. Each sample solution should be filtered routinely before injection onto the column. This prevents sample loop, capillary, or column plugging by small particles in the sample solution and also prolongs column life significantly. Besides filtration, more complex samples require further cleanup and often an additional enrichment step. No general rule for sample cleanup or sample enrichment can be given; each sample matrix is different, and the particular cleanup must be adjusted to the sample in question. Nevertheless, water-soluble (polar) samples can be cleaned up or enriched simply by liquid – solid extractions (e.g., small silica columns) or liquid – liquid extractions in a separatory funnel. The latter is also very important for the isolation of hydrophobic analytes [e.g., polyaromatic hydrocarbons (PAHs), pesticides] from a polar matrix (i.e., water). Isolation of hydrophobic solutes from a polar matrix is done easily by pumping the solution over a lipophilic stationary phase (e.g., RP-18). Here, the hydrophobic analytes are retained while the more polar components do not interact with the stationary phase. After a certain amount of sample solution (depending on the separation problem and the capacity of the solid phase) has been pumped over the stationary phase (either an open-column "solid-phase extraction" or a small HPLC column), the mobile phase is changed to a hydrophobic (organic) solvent, which elutes the retained analytes. After evaporation of the organic solvent, the analytes can be dissolved in a known amount of solvent for analysis.

Another very interesting and promising cleanup technique is supercritical fluid extraction (SFE) (→ Liquid – Solid Extraction, Chap. 4.). This method, which usually employs supercritical carbon dioxide as an extraction medium, shows a high selectivity toward nonpolar solutes. As a rule of thumb, if a solute is soluble in hydrophobic solvents such as heptane or sometimes even in more polar ones such as dichloromethane, it should also be soluble in supercritical carbon dioxide. Through addition of modifiers (e.g., methanol, water, formic acid), even relatively polar molecules are extractable, although the increase in solvent strength is paid for by a loss in selectivity. The selectivity and thus the solvent strength of the supercritical fluid can also be adjusted by variation of the extraction pressure, density of the extractant, or temperature.

SFE with carbon dioxide offers several advantages compared to classical solvent extraction methods: (1) it is very selective; (2) it is fast, because of lower viscosity and approximately 10 times higher diffusion coefficients; and (3) CO_2 is neither flammable nor explosive and is nontoxic. SFE is suitable for oxidizable and temperature-sensitive samples. In addition, no reactions occur between extraction medium and sample compounds and on-line coupling is possible with both supercritical fluid chromatography (SFC) and gas chromatography.

Nevertheless, this technique also has some disadvantages, especially in collecting the extracted analytes. High molecular mass analytes tend to precipitate inside the restrictor (see Section 10.4.5) and cause it to plug, during the expansion of the supercritical fluid (expansion of supercritical fluids results in a decrease of density and therefore solvent power). Low molecular mass samples (volatile samples) can be carried away in form of an "analyte fog" to a certain extent during the expansion procedure, resulting in insufficient recovery. Further information on this promising sample cleanup method is given in [91,92].

8.2. Pre- and Postcolumn Derivatization

Sample derivatization is often applied to enhance detection selectivity or sensitivity. This technique is also applied in GC and TLC. Although derivatization is a very helpful tool in chromatography it has the disadvantage that derivatization reactions make the individual sample compounds more similar in adsorption behavior, which can decrease separation selectivity.

In LC the derivatization can be performed either before the chromatographic separation (precolumn derivatization) or after the analytes have passed the column (postcolumn derivatization).

8.2.1. Precolumn Derivatization

Precolumn derivatization is the most frequently applied derivatization technique for many rea-

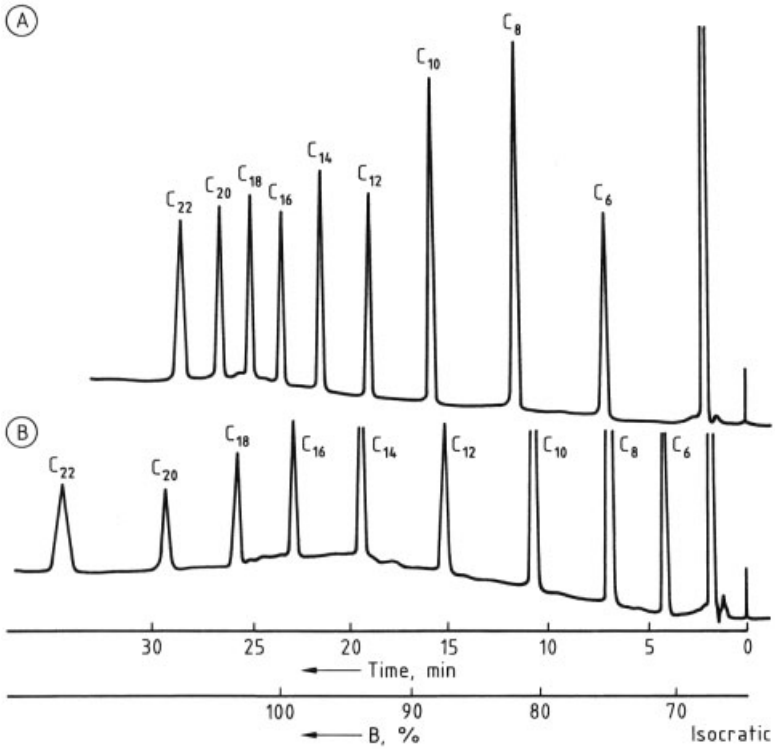

Figure 34. Separation of fatty acid phenacyl esters on an RP-18 column (300×4.2 mm i.d.) with flow = 2 mL/min and t_g = 20 min linear [95]
A) 70 % Methanol in water to 100 % methanol; B) 70 % Acetonitrile in water to 100 % Acetonitrile

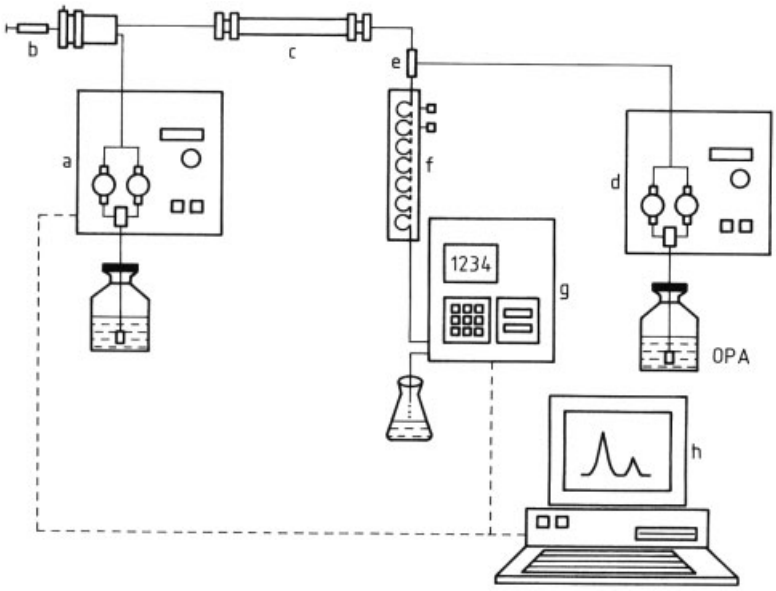

Figure 35. Diagram of postcolumn derivatization system
a) Pump; b) Sample introduction; c) Separation column; d) Reagent pump; e) Mixing chamber; f) Reactor; g) Detector; h) Data control unit

sons. One very important factor is that the chromatographic equipment need not be changed because precolumn derivatization is performed off-line. Furthermore, the reaction time is of minor importance. Nevertheless, the derivatization should be quantitative, and in certain cases, excess derivatization reagents can lead to poor chromatographic separation.

The most frequent application of precolumn derivatization is in amino acid analysis. Here, the amino acids are derivatized for instance with 9-fluorenyl methoxycarbonyl chloride (FMOC) or *o*-phthalaldehyde (OPA) to form strongly fluorescing compounds, which can be detected easily with a fluorescence detector [93, 94]. Another typical application for precolumn derivatization is in the analysis of fatty acids. Fatty acids have no chromophore, so UV detection is impossible. Even highly unsaturated fatty acids can be detected only in the very far UV range (< 200 nm) with poor sensitivity. Other detectors, such as the refractive index detector, have a disadvantage because they cannot be used in gradient elution, which is essential for fatty acid analysis. Therefore, the fatty acids are reacted with UV-absorbing reagents (e.g., phenacyl bromide) to form fatty acid phenacyl esters, which are easily detectable at 254 nm [95] (see Fig. 34).

8.2.2. Postcolumn Derivatization

In postcolumn derivatization, also known as chemical reaction detector (CRD), the sample compounds are derivatized on-line after separation. Derivatization takes place in a reactor situated between the column outlet and the detector (see Fig. 35). The chromatographic equipment is therefore extended by an additional reagent pump, which delivers the reagent into a zero-void-volume mixing chamber (Fig. 36). Here the reagent is mixed thoroughly with the mobile phase containing the solute. (An ordinary three-way unit is not sufficient as a mixing chamber in postcolumn derivatization.) The mixture is then pumped through a reactor where actual derivatization occurs before the sample reaches the detector. The derivatization reagent must be delivered continuously, with as little pulsation as possible.

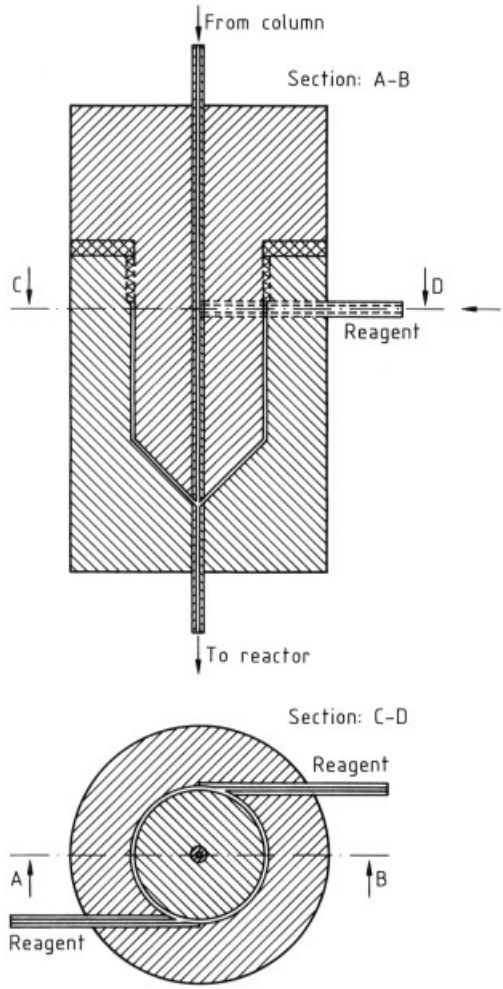

Figure 36. Example of a postcolumn derivatization mixing chamber (cyclon type) [94]

The reaction time required for a particular derivatization is the limiting factor. It must be short enough so that derivatization can take place during passage of the solutes through the thermostated reactor. Of course, the required reaction time can be prolonged to a certain extent by enlarging the reactor, but this is paid for by increasing back-pressure and additional peak dispersion. Even if the required reaction time for a particular derivatization (i.e., for 100 % conversion) cannot be achieved with a reactor, this is often no problem in postcolumn derivatization since the derivatization need not to be quantitative, merely reproducible.

The general rule of avoiding any additional void volume between the column outlet and the detector cell cannot be fulfilled in postcolumn derivatization. Therefore, loss of separation efficiency to a certain extent, due to peak dispersion, is unavoidable. To reduce this drawback, several reactor types have been developed. The most important ones are:

1) Packed-bed reactors
2) Open tubes with optimal geometric orientation
3) Open tubes with liquid or gas segmentation
4) Coiled open tubes

Many excellent papers have been published that compare these different types of reactors [94, 96]; therefore, only the most important reactors (open tubes with optimal geographic orientation) are discussed here in more detail.

Coiled Open Tubes. Initially, WALKLING [97] and HALASZ [98] showed that systematic defor-mation of open tubes can significantly reduce peak dispersion, compared to a simply coiled tube with uniform i.d. A detailed discussion of the different parameters influencing peak dispersion (e.g., relationship of axis, coiling diameter, knot distance, kinematic viscosity of the mobile phase, inner surface of the tube) is given in [99]. The breakthrough in the optimal geometric orientation of open tubes came with the introduction of PTFE capillaries (i.d. = 0.2 – 0.5 mm). With this easily deformable material (compared to glass or stainless steel capillaries), not only could coiled and waved reactors be made [99], or figure eights formed around two rods [100], but the capillary could be knitted [100, 101]. Capillaries are knitted in such a way that each right loop is followed by a left one (with small coiling diameter) so that each loop is bent out of the plane of the preceding one (Fig. 37). Because of the small coiling diameter and continuous change in direction, centrifugal forces are effective even at low flow rates, initiating a secondary flow, that is ultimately responsible for the reduced band broadening. These so- called knitted open tubes (KOTs) have the best performance in terms of peak dispersion. With these KOTs, peak dispersion could be made independent of linear velocity over a large velocity range (even at very low flow rates) (Fig. 38).

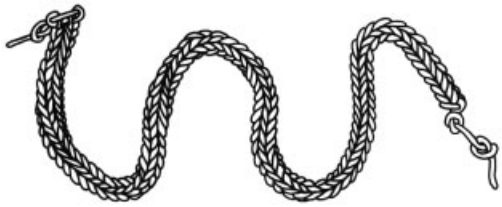

Figure 37. Knitted open tube (KOT) reactor

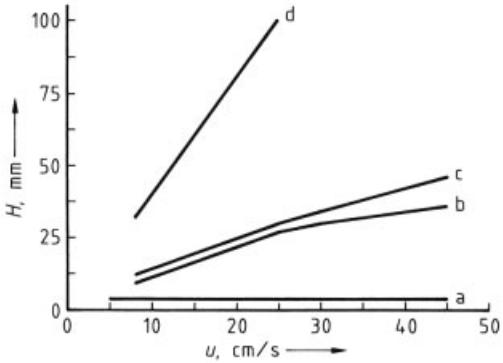

Figure 38. Peak dispersion in open tubes
a) Straight capillary; b) $R = 2.6$ cm; c) $R = 0.35$ cm; d) Knitted open tube (KOT)

Packed-Bed Reactors. Packed-bed reactors are packed preferably with small glass beads (20 – 45 µm). The reactor length and therefore the reaction time are limited by the generated back-pressure of the packed-bed reactor. This is the reason it is generally used only for fast derivatization reactions.

Open Tubes with Liquid or Gas Segmentation. Another often-used reactor type (especially for relatively long-lasting derivatization reactions) is the open tube with liquid or gas segmentation [94]. Segmentation prevents the formation of a parabolic flow profile (axial dispersion), and peak dispersion is due mainly to the "phase separator," which is responsible for removal of the segments prior to detection. The principle of segmentation was first used by MOORE and STEIN [102], who developed an amino acid autoanalyzer (ninhydrin reaction) in 1958.

During the development of a new postcolumn derivatization for a particular separation problem, the following points should be considered:

Table 7. Examples of postcolumn derivatization techniques in HPLC

Compound	Reagent,∗ reaction type	Chromatography	Reference
Amino acids	ninhydrin	IEX	[103]
	fluorescamine	IEX	[104]
	OPA	IEX	[105]
Peptides	fluorescamine	RP	[106]
Mono-, oligosaccharides	3,5-dihydroxytoluene	IEX	[107]
	catalytic hydrolysis/ABAH	IEX	[108]
Carbamate pesticides	alkaline hydrolysis/OPA	RP	[109]
Glyphosate (herbicide)	hypochlorite oxidation/OPA	RP	[110]
Vitamin B_1	thiochrome	RP	[111]
Vitamin B_6	semicarbazide	RP	[112]
Vitamin C	oxidation/*o*-phenylenediamine	RP	[113]

∗ OPA = *o*-phthaladehyde; ABAH = 4-aminobenzoic acid hydrazide.

1) An appropriate derivatization reaction should be sought; these are often described in the literature for classical off-line derivatization
2) The reagent concentration should be optimized: Highly concentrated reagents should be delivered with a low flow rate to prevent any unnecessary sample dilution in the mobile phase
3) Reaction time should be as short as possible; the longer the required reaction time is, the longer must the reactor be, which results in increased back-pressure and peak dispersion
4) To shorten the required reaction time, the reactor temperature should be as high as possible (an increase of 10 °C doubles the reaction speed)
5) An ca. 80 % conversion of the original solute into the derivatized compound is sufficient to achieve maximum detection sensitivity even for quantitative analysis

Table 7 shows a few derivatization reagents and their typical applications.

9. Coupling Techniques

Many different coupling techniques in HPLC were applied since the 1970s but only a few proved to be a significant advantage for the analytical chemist. Even simple HPLC equipment can pose severe handling and reproducibility problems for the beginner. Each additional coupling technique naturally increases these problems, making sophisticated coupling techniques more interesting for academic use than for routine analysis. Nevertheless, coupling techniques such as HPLC – GC and HPLC – MS (mass spectrometry) have become very important tools for analytical chemistry. Another often-applied coupling technique is column switching.

9.1. Column Switching

Column switching was described in the late 1960s by Deans [114] for multidimensional GC (→ Gas Chromatography, Chap. 8.1.). In principle, itis the coupling of different columns (usually with different selectivities) by means of a manual or automatically driven high-pressure valve. The first application of column switching in HPLC was described by Huber et al. [115] in 1973. They used this technique as a simple and inexpensive alternative to gradient elution. Gradient systems were very expensive in those days. However, in the following years, good and reasonably priced gradient systems came onto the market, making column switching unnecessary as an alternative.

New applications for this technique were found especially in the analysis of very complex samples (e.g., drug analysis in blood serum), where the analyte is selectively "cut out" of the sample matrix. According to the position of the "cut" in a chromatogram, front-cut, heart-cut, and end-cut techniques can be distinguished. Other applications of column switching are online sample cleanup, backflush technique, and multidimensional chromatography. Figure 39 shows different examples of switching arrangements.

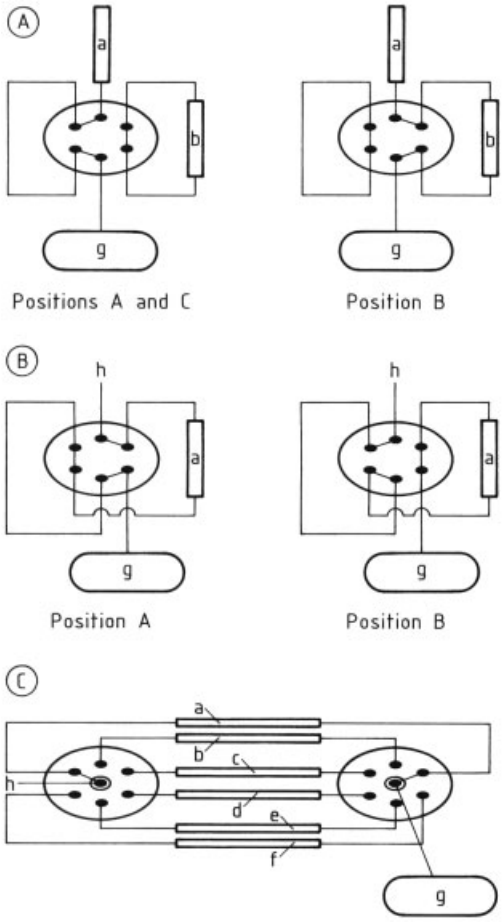

Figure 39. Arrangements for column switching technique
A) Heart-, front-, and end-cut techniques; B) Backflush technique; C) Multidimensional chromatography
a) – f) Column 1 – 6; g) Detector; h) Injection valve

The high-pressure switching valve is the most important part of the instrumentation in this technique. It must work reliably even after several thousand switches under high pressure, must have a very small dwell volume, and must be chemically inert to both the sample and the mobile phase. Normal switching valves have six ports, but four- and up to ten-port valves are commercially available as well. Their setup is very similar to the usual sample valve injectors, e.g., "rheodyne valves" (Section 2.4.2). They can be driven by hand or automatically either by pneumatic force or electronically. Pneumatically or electronically driven valves have the advantage that the entire analysis can be automated.

9.1.1. Front-, Heart-, and End-Cut Techniques

A sample is usually preseparated on a short precolumn and then directed onto an analytical column (with the same or different selectivity) where actual separation occurs. The *heart-cut technique* is most often used when the solute coelutes with another sample or when it is situated on the flank of another peak. In this case the interesting part of the chromatogram is cut out and "parked" on the analytical column, while the rest of the sample is eluted from the precolumn into the waste. Thereafter, the valve is switched back to the analytical column position, and the parked sample can now be baseline separated (Fig. 39 A). The *front- and end-cut techniques* are basically the same, differing only in the position of the cut.

9.1.2. On-Line Sample Cleanup and Trace Enrichment

Column switching is also very suitable for sample cleanup and can be combined with trace enrichment. A typical example of this technique is the on-line sample cleanup and enrichment of hydrophobic pesticides from aqueous solutions. Here the sample (water, blood, etc.) is pumped over a small RP column, where only the hydrophobic components are retained. Since the amount of sample can be chosen in any size or volume, hydrophobic trace enrichment takes place. After the entire sample volume has passed the precolumn, the high-pressure valve is switched to the analytical column position and an eluent (having a stronger elution strength for hydrophobic analytes e.g., dichloromethane) is passed through the RP precolumn, eluting the hydrophobic components onto the analytical column where separation occurs.

9.1.3. Backflush Technique

The switching arrangement for the backflush technique is shown in Figure 39 B. This technique is used mainly if the components of interest in a particular sample are very strongly retained by the stationary phase. Mobile phase

flow is maintained in one direction until all early-eluting components have left the column. At this stage, the more strongly retained components are still near the head of the column. Then, the flow direction is reversed (backflush) to allow faster elution of these analytes. The switching arrangement for the backflush technique can be used in the routine analysis of very complex samples that contain components having a high affinity for the stationary phase. In this case, after a certain number of chromatographic runs the flow direction is reversed to flush the analytical column and elute these strongly retained components [116]. This lengthens column life significantly.

9.1.4. Multidimensional HPLC

In multidimensional HPLC, several columns with different selectivities are switched parallel to each other with the help of two high-pressure switching valves (see Fig. 39 C). This technique is used for samples that contain components with a wide range of polarity (from low to high k' values). In this case, one should determine whether gradient elution would be more efficient.

All of the aforementioned column switching techniques can be combined with each other in various ways to provide analytical chemists with important tools for their investigations.

9.1.5. Working with Precolumns

In column switching, often at least one column is a precolumn. In the majority of cases the precolumn does not require high separation efficiency. Its purpose is to retain or enrich a certain part of a sample. The actual separation is performed on a second column (the analytical column), which possesses the appropriate separation efficiency. Therefore, the precolumn can be packed with inexpensive, coarse particles. Even a well-packed column bed is unnecessary. If the analyte of interest is retained on the precolumn in a relatively wide zone, this large elution band is reconcentrated as soon as it reaches the head of the analytical column (as long as the analytical column has a better selectivity and higher separation efficiency than the precolumn). Precolumns are usually short stainless steel columns (ca. 20×4 mm i.d.), packed with coarse material (30 – 60 µm), which need not have the small particle-size distribution required for analytical columns. Therefore, these precolumns can easily be homemade, and if purchased, they are inexpensive. If the precolumn is used as a "protection" for the analytical column (its original purpose), it should be well packed and the packing material should be the same as that in the analytical column.

9.2. HPLC – Mass Spectrometry

The coupling of high-performance liquid chromatography with mass spectrometry (HPLC – MS) has become a very important tool for the reliable identification and quantification of complex samples. In the pharmaceutical industry, for instance, the metabolic pathways or pharmacological effects of new products are often studied with this system.

The main problem in coupling the LC system with the mass spectrometer is the fact that the mass spectrometer is under a high vacuum and the mobile phase from the HPLC system must be removed before the analyte can enter the MS. This is the reason why HPLC – MS coupling is usually done with microbore or even capillary columns. In using the "normal" 4.0 – 4.6-mm i.d. columns, too much solvent would have to be removed. Initially, solvent removal was performed by the *moving belt* technique, for example. Today, the *direct liquid inlet* technique is often utilized. In this case, solvent eluting from the LC system is sent through an ca. 5-µm hole into the vacuum of the MS. The resulting aerosol comes into a "desolvation chamber," where the solvent is completely removed. With the *thermospray techniqe*, the analyte containing the mobile phase is sent through a heated (60 – 200 °C), narrow pipe (ca. 0.1-mm i.d.) directly into the MS vacuum of a specially designed ion source. This technique has the advantage that no split is necessary after the separation column, so the entire mobile phase can be sent into the MS system, which enhances detection limits significantly. Further, "normal" HPLC columns (4.0 – 4.6-mm i.d.) can be used in this technique, with flow rates between 0.8 and 1.5 mL/min.

For the HPLC – MS systems, many different ionization techniques have been described in the past. Various interface, ionization methods, and

Table 8. Overview of advantages and disadvantages of various LC – MS interfaces ∗ [117]

LC – MS type	LC flow range	Detection limit	Solvent type range	Nonvolatile thermally labile samples	Semivolatile samples	Molecular mass range
Moving belt	+++++	++	+++++	+	+++++	+
Particle beam	+++	++	++++	+	+++++	+
Thermospray	+++++	++	++++	+++	+++++	+++
LC – FAB	+	++	+++	+++++	++++	++++
Electrospray	+	++++	++	+++++	++++	+++++

∗ Increasing numbers of "+" indicate improved performance.

operating techniques applicable to LC – MS are discussed in [117] for instance: Thermospray, particle beam, electrospray (ES), field desorption (FD), fast atom bombardment (FAB), time of flight (TOF), etc. The electrospray technique produces a soft ionization for thermally labile compounds, while FAB has the advantage that higher molecular mass samples can be introduced into the mass spectrometer. Table 8 offers a rough guide to the applicability of various LC – MS interfaces. For more detailed information on LC – MS, see [118].

In HPLC – MS coupling, careful sample preparation is recommended. The "dilute-and-shoot" approach is not possible, especially when handling real-world samples such as environmental and biological samples. In these cases sample preparation techniques (see also Chaps. 8 and 9) such as liquid – liquid extraction, solid-phase extraction [119], or even affinity techniques (see also Chap. 11) directly coupled with LC – MS [120] can be useful.

10. Supercritical Fluid Chromatography

10.1. History

In 1822, Cagniard de la Tour first described the critical state [121]; In 1869, Andrew [122] published the first paper about a gas – liquid critical point. In 1879, Hannay and Hogarth described the solubility of cobalt and iron chlorides in supercritical ethanol [123]. Lovelock (1958) first suggested the use of supercritical fluids as a mobile phase in chromatography [124], and in 1962, Klesper et al. [125] published the first paper on supercritical fluid chromatography. In this paper they described the separation of nickel porphyrins using supercritical chlorofluoromethanes as the mobile phase. This work was done before HPLC was introduced by Piel in 1966 [9]. In the following years, Sie, Rijnders, and Giddings [126–134], through their practical and theoretical studies, made significant contributions to understanding of the separation process with supercritical fluids. Sie and Rijnders at the Shell Laboratories in Amsterdam were the first to use the term supercritical fluid chromatography in 1967 [135].

In the 1970s, SFC declined in popularity, partly because of the very fast development of HPLC and because, in those days, working with supercritical fluids was not so easy due to several technical problems. Another reason SFC declined in popularity was that expectations had been set far too high in the early use of this technique.

Nevertheless, the development of SFC continued. In 1969 – 1970, Jentoft and Gouw [136, 137] were the first to work with pressure programming and to use modifiers to increase the elution strength of the mobile phase. Schneider and coworkers [138–140] investigated the physicochemical aspects of supercritical fluids and thereby contributed much to the understanding of the technique.

The introduction of capillary columns in SFC by Novotny and Lee in 1981 [141] led to a revival of this technique, which was followed by a large number of papers on new applications. At the same time (1981) the first commercial packed-column SFC instrument came on the market. Four years later, in 1985, the first capillary column SFC instrument was available commercially.

Since the late 1980s, capillary and packed-column SFC have become more and more popular, as shown by the growing number of published papers, new books, and international meetings devoted exclusively to SFC and SFE.

The reason for this recent gain in popularity is that significant progress has been made in improving the instrumentation, which expanded the applications to many problems such as quantitative environmental analysis (although generally the sensitivity is not as good as in GC, for instance). Another reason is that modern chromatographers have begun to realize that the various analytical separation techniques (HPLC, GC, TLC, SFC, CE, etc.) should not compete against each other. Each technique has its advantages, and which technique is preferred should be decided on a case-to-case basis.

10.2. Theoretical Background

If a gas is compressed above its critical pressure p_c at temperatures exceeding the critical temperature t_c, then neither a gas nor a liquid is present but a so- called supercritical fluid, which combines properties of both. Table 9 shows the critical parameters of the most commonly used fluids in SFC.

Table 9. Critical parameters of the most frequently used supercritical fluids [142]

Supercritical fluid	p_c, MPa	t_c, °C	ϱ_c, g/cm^3
CO_2	7.39	31.3	0.47
N_2O	7.34	36.5	0.45
SF_6	3.76	45.5	0.74
Xe	5.92	16.6	1.10
NH_3	11.4	132.5	0.24
CCl_2F_2	4.12	111.8	0.56

Table 10 compares some important physicochemical properties of a liquid (HPLC), a gas (GC), and a supercritical fluid. The density – and thus the solvation characteristics of the supercritical fluid – are very similar to those of a liquid. On the other hand, the viscosity is comparable to a gas, and the diffusion coefficient is about ten times higher than that of a liquid. Therefore in SFC (especially packed-column SFC), analyses can be performed at much higher flow rates than in HPLC without loss of too much efficiency (see Fig. 40), which results in a significantly shorter analysis time. Because of the low viscosity of supercritical fluids, several packed columns can be combined in series, even with small particles (5 µm), without generating too high back-pressure. If several columns packed with the same stationary phase are connected in series, the separation efficiency can be improved dramatically. The combination of different stationary phases (columns) and the variation of their column length result in totally new and interesting selectivity ("selectivity tuning").

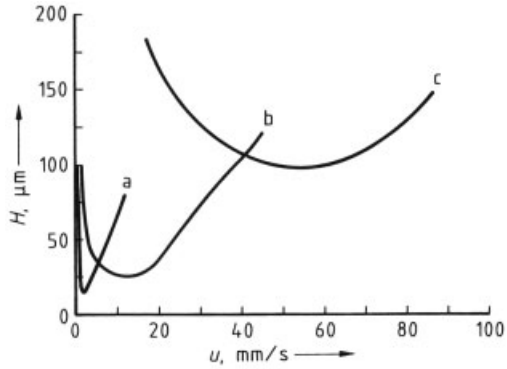

Figure 40. Comparison of Van Deemter curves in GC, HPLC, and SFC
All packed columns (particle size 10 µm); mobile phase in HPLC: methanol, in GC: N_2, in SFC: CO_2; samples in HPLC: benzene, in GC and SFC: methanea) HPLC; b) SFC; c) GC

In general, a mobile phase for SFC should exhibit the following characteristics:

1) The critical parameters (temperature and pressure) should be achievable with commercial equipment
2) The fluid should be available in high purity and not be too expensive
3) The fluid should be chemically compatible with the sample and SFC equipment
4) The fluid should be miscible with many organic solvents (modifiers)

Table 10. Comparison of physical properties of a liquid, a supercritical fluid, and a gas

	Liquid	Supercritical fluid	Gas
Viscosity η, Pa · s	$<10^{-1}$	$10^{-4}-10^{-3}$	10^{-4}
Density ϱ, g/cm^3	0.8 – 1.0	0.2 – 0.9	10^{-3}
Diffusion coefficient D_m, cm^2/s	$<10^{-5}$	$10^{-3}-10^{-4}$	10^{-1}

Because of the low critical temperature and pressure of carbon dioxide and its compatibility with much equipment, especially with the flame ionization detector (FID), supercritical carbon dioxide is the most frequently used mobile phase in SFC.

The elution strength of a supercritical fluid is often described by its dielectric constant *DC*. Figure 41 shows that the influence of pressure on the *DC* at different temperatures is very similar its influence on the density, which is presented in Figure 42. This diagram is one of the most important for understanding supercritical fluid chromatography and extraction. Generally, the solubility of a solute in a supercritical fluid increases with increasing density of the fluid. Simultaneously, the elution strength of the fluid increases as well, which results in shorter retention times. Figure 42 shows that a small variation of pressure slightly above the critical parameters (p_c and t_c) causes a large change in density and therefore in the elution strength of the fluid. On the other hand, a temperature increase at constant pressure causes a decrease in the density and consequently in the elution strength. In other words, in contrary to gas chromatography, an increase in temperature at constant pressure results in increased retention and therefore analysis time. This explains why negative temperature gradients are often run in SFC.

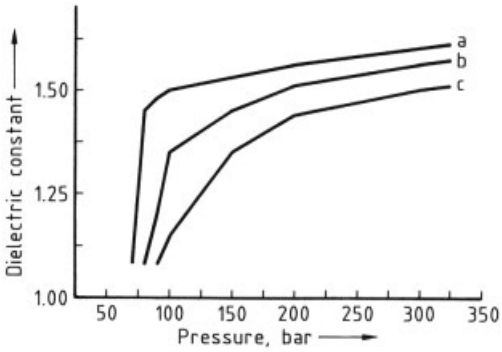

Figure 41. Influence of pressure on the dielectric constant of carbon dioxide at different temperatures
a) $t = 25\,°C$; b) $t = 40\,°C$; c) $t = 60\,°C$

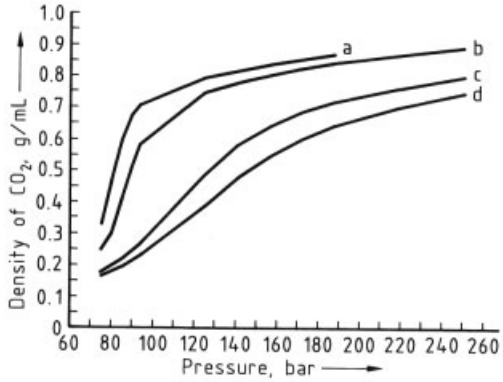

Figure 42. Influence of pressure on the density of carbon dioxide at different temperatures
a) $t = 35\,°C$; b) $t = 40\,°C$; c) $t = 60\,°C$, d) $t = 70\,°C$

Further discussions concerning the theory and basics of SFC are given, for example, in [124, 142, 143].

10.3. SFC Compared to GC and LC

As mentioned above, SFC should not be regarded as competing with GC or HPLC. SFC is complementary to these two analytical separation techniques, and one should decide from case to case which method is preferred.

Because of its intermediate position, SFC combines several advantages of both GC and HPLC. One significant advantage of SFC compared to HPLC is the fact that besides the normal LC detectors (e.g., UV detector), sensitive universal or selective GC detectors (e.g., FID, electron capture detector) are applicable as long as supercritical carbon dioxide is utilized as mobile phase. Furthermore in packed-column SFC, the same variety of stationary phases (selectivities) can be used as in LC (in contrast to GC), with the additional advantage that the analysis time is significantly shorter.

SFC is superior to GC if thermolabile solutes must be separated. However, this is true only as long as the supercritical fluid has a relatively low critical temperature like carbon dioxide (31 °C), for example.

Nevertheless, SFC also has its drawbacks; for example, only a limited number of samples are soluble in supercritical fluids. In supercritical carbon dioxide, for instance, only relatively nonpolar solutes are soluble. As a rule of thumb, so-

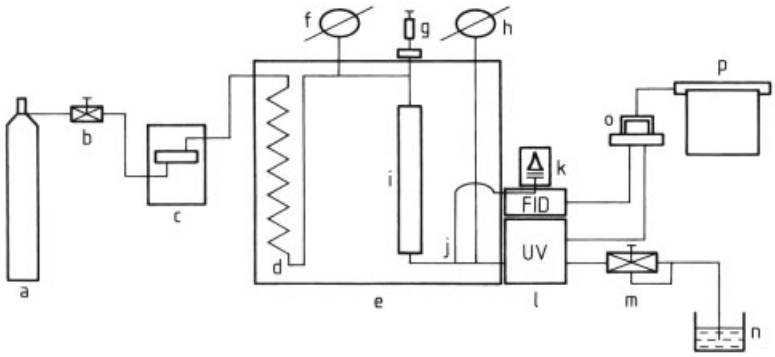

Figure 43. Schematic diagram of an SFC apparatus
a) Gas supply; b) Valve; c) Pump; d) Equilibration coil; e) Oven; f) Manometer; g) Sample introduction unit; h) Manometer; i) Separation column; j) T-unit; k) Flame ionization detector (FID); l) UV detector; m) Back-pressure valve, restrictor, or pressure and flow control unit; n) Fraction collector; o) Data control unit; p) Recorder

lutes that are soluble in organic solvents having a polarity less than or equal to that of *n*-heptane are usually also soluble in supercritical carbon dioxide. Samples that are soluble in water are insoluble in CO_2.

If the elution strength of a supercritical fluid is not sufficient, it can be increased by adding an organic modifier (see Section 8.1). In this way, especially in combination with capillary columns, the applicability of SFC is extended to more polar analytes. Methanol is the most commonly used modifier because it shows the widest solubility range for samples that are insoluble in CO_2. If selectivity enhancement is required, a more nonpolar solvent should be chosen. The most common modifiers for SFC are discussed in [144]. Several papers have been published on the mechanism of modifiers in SFC systems [142]. Apparently, the modifier has at least two effects: First, it dramatically alters the solvation properties of the fluid (usually increases the solubility of the sample compound); second, it blocks the most active retention sites of the column packing so that the retention strength of an analyte is decreased as is its analysis time. Even though modifiers have significant advantages, they limit the choice of suitable detectors (e.g., methanol as modifier interferes in FID). Whereas in packed- column SFC, modifier concentrations of 1 – 5 % are usually sufficient, in capillary SFC, modifier concentrations up to 20 % can be necessary. For the separation of acidic or basic samples an additive to the modifier or mobile phase can often be helpful. This improves peak shape and often shortens analysis time considerably. A typical additive for acidic samples is trifluoroacetic acid (TA) whereas for basic samples, primary amines are often used. The function of an additive is not yet clear, although it seems to homogenize the stationary phase (effectively blocking off silanol groups), to improve the solubility of the solute in both phases, and to suppress solute ionization.

A disadvantage of SFC is the fact that even though the same detectors are used as in GC, the detector sensitivity is not as good. In capillary SFC, this results from the very small sample volumes (10 – 250 nL) that must be injected into the column. In the case of packed columns, larger sample volumes can be injected but the column outlet going into the combustion detector (e.g., FID, NPD) often has to be split. If not split is applied, the extremely high flow rates, caused by the expanding supercritical fluid, tend to extinguish the flame of the combustion detector.

10.4. Equipment

The equipment for packed-column SFC is very similar to that for HPLC, whereas the instrumentation for capillary SFC is more reminiscent of gas chromatography. In Figure 43 a schematic diagram of packed-column SFC instrumentation is shown. It consists of a high-pressure pump, sample introduction system, oven, detector, restrictor, and recorder. In addition, a microcomputer is installed that controls temperature, flow, pressure, and density programming.

10.4.1. Pumps

The pump systems used most frequently in open tubular (capillary) and packed-column SFC are the reciprocating piston pumps and the syringe pumps (see Section 2.3). The syringe pumps have the advantage of pulseless flow and easy pressure control, while the reciprocating pumps have the advantage of continuous flow. The main difference between the LC pumps described in Section 2.3 and SFC pumps is that the latter need a pump head cooling system (temperature of pump head $< 5\,^{\circ}$C). Reciprocating pumps can deliver only liquids efficiently, not highly compressible supercritical fluids. A further difference is that in SFC a very effective pulse dampener is required to ensure reproducible separations.

The maximum pressure of a modern SFC pump should be as high as possible, at least 40 – 45 MPa, with a flow range from several microliters per minute for open tubular columns (capillary columns) to ca. 5 – 10 mL/min for packed columns.

10.4.2. Oven

The oven of an SFC system should meet the same requirements as a normal GC oven. A constant temperature (variation $\pm 0.1\,^{\circ}$C) must prevail in the entire oven at any time of a positive or negative temperature gradient. This is very important for reproducible capillary column SFC analysis. These columns are very sensitive to even slight variations in temperature, which can result in peak shape deformation, peak splitting, or irreproducible retention times.

10.4.3. Sample Introduction

For packed columns and capillary columns, different sample introduction systems are used.

In the case of capillary columns, finding appropriate solutions for sample introduction is much more challenging than with packed columns. Since packed columns have dimensions similar to normal HPLC, optimal sample introduction can be achieved with ordinary sample loop injectors (see Section 2.4.2).

Onto a normal capillary column (10 m $\times$ 50 µm i.d.), < 100 mL should be injected to prevent more than 1 % loss in resolution [142]. To achieve this, different techniques of sample introduction have been used. In 1988, Greibrokk et al. [145] described a technique that involves separation of the solvent from the solutes on a precolumn. The solutes are then transferred to the analytical column where they are focused at the column head. Other solvent elimination techniques based on gas purging and solvent backflush were reported by Lee et al. [146]. The most common reproducible techniques to inject such small volumes of sample onto capillary columns are the dynamic split and the time split technique. In the *dynamic split mode*, which is the most frequently applied technique in open tubular column SFC, the split ratio is determined by the ratio of the flow out of the column to the flow out of the split restrictor. Thus, if the flow out of the split is 500 times faster than the flow out of the column, the split ratio is 1 : 500. The disadvantage of the dynamic split technique is that obtaining reproducible results can be difficult with certain samples. The *time split injection* technique is based on early work done by J. C. Moore [147] for LC. In 1986, Richter [148] was the first to report on this injection technique in SFC. A standard injection valve, which is pneumatically driven by a low-viscosity gas such as helium, is connected directly to the analytical column (packed column or capillary column) The valve is electronically driven by high-speed pneumatics to ensure very fast switching times (from load position to inject position and back in the millisecond range). The shorter the switching time is, the smaller is the injected sample volume. Many papers have been published comparing the reproducibility of the time split and the dynamic split techniques, some of them differing significantly in their results. Whereas Richter et al. [149] showed that time split is superior to dynamic split, Schomburg et al. [150] saw no significant difference between the precision of the two techniques. Because of problems in reproducibility with these two injection techniques, the use of an internal standard is strongly recommended for quantitative analysis.

Table 11. Detection possibilities with SFC ∗

Detector	Type	Capillary SFC	Packed- column SFC	Modifier possible	Preferred mobile phase
FID	U/D	++	+ ∗∗	yes	CO_2, SF_6, N_2O
NPD	S/D	++	+ ∗∗	yes	CO_2, SF_6
UV	S/ND	+(−)	++	yes	CO_2
ECD	S/D	++	+∗∗	yes	CO_2, Xe
LSD	U/(N)	+	++	yes	CO_2
FTIR	U/N	+	++	no	Xe, CO_2, SF_6
MS	U/D	++	+ ∗∗	yes	CO_2

∗ U = universal detection; S = selective detection; D = destructive; N = nondestructive; + = detection possible; ++ = preferred column type for the particular detector; − = column type not recommended.
∗∗ Split usually necessary.

10.4.4. Detectors

As mentioned before, a unique advantage of SFC is the fact that a wide variety of detection methods can be applied (see Table 11). Besides the traditional LC detection (e.g., UV) the use of GC detectors, especially, enables relatively sensitive, universal (FID), as well as selective [e.g., electron capture detectors (ECD), thermoionic detectors (TID)] methods of detection.

As in LC, both destructive and nondestructive detectors are used in SFC. The detector most commonly used for this technique is the flame ionization detector. It is a destructive, sensitive, and universal detector, which is well known from gas chromatography. Even though Richter et al. [151] described, in 1989, a new FID detector with a detection limit (5 pg of carbon per second) and a linear dynamic range (10^6) very similar to that achieved in FID – GC systems, the ordinary FID – SFC system is still a factor of 10 – 1000 poorer in sensitivity, which consequently calls for relatively concentrated sample solutions. If very sensitive detection is essential, the FID – SFC detection limit can be increased by a factor of ca. 1000 by replacing carbon dioxide with nitrous oxide, which has similar physicochemical characteristics (see Table 9). However, one should be aware of the fact that N_2O can explode under higher pressure and temperature.

A further drawback of the SFC – FID system is the fact that the amount of organic modifier that can be added to the fluid is limited (methanol ca. 3 – 5 vol %) because the FID responds to organic molecules. Water can be an alternative, but its solubility in supercritical carbon dioxide is low (<1 %), and the elution strength does not increase too much.

The UV detector is also very frequently used in SFC, mainly in packed column SFC, because of problems in sensitivity with capillary columns. Since pure carbon dioxide shows little absorbance between 200 and 800 nm, UV detection in packed- column SFC is very sensitive. In general, the characteristics of UV detectors discussed in Section 2.6.3 are also valid for UV – SFC detectors, except that the latter must be equipped with a pressure-resistant flow cell. A normal UV detector flow cell in LC can resist a maximum pressure of ca. 7 – 10 MPa. In SFC, the UV detector cell must be under the same separation conditions (temperature and pressure) as the column (if the pressure in the detector cell were lower, the fluid would loose its solvation power), so it should resist pressure ≤ 40 MPa.

A nondestructive and universal detector in SFC is the FTIR detector (see Section 2.6.7). Although its sensitivity is not too high, this technique is highly informative and often helps to identify analytes by their characteristic functional groups or to solve molecular structure problems.

Compared to LC, SFC – FTIR systems have the significant advantage that carbon dioxide is transparent in a large part of the IR region, allowing FTIR flow cell detection [152]. In LC, the mobile phase (especially organic solvents) must usually be removed prior to FTIR detection, because it absorbs in the required IR region. Consequently, organic modifiers used in SFC – FTIR cause the same problem [153, 154], making flow cell detection impossible. The "ideal" fluid for SFC – FTIR flow cell detection is supercritical xenon. It is transparent over the entire IR range and shows sufficient solvation power [155]. Unfortunately, the high price of xenon limits its application in routine analysis. Many other detection techniques are described for SFC. An ex-

cellent discussion of the various detection possibilities in SFC is given in [142].

10.4.5. Restrictors

The restrictor is an important part of all SFC equipment. It maintains constant pressure in the system and ensures, or at least should ensure, uniform, pulse-free flow of the supercritical fluid; thus, it is also responsible for reproducible analysis. Two types of restrictors exist: Fixed restrictors (see Fig. 44) and variable restrictors. *Fixed restrictors* have the disadvantage that pressure programming automatically also changes the flow. Therefore, independent pressure and flow programming can be achieved only with computer-driven variable restrictors, which often function in principle like a needle valve. Usually the restrictor must be heated since, due to the expansion of the supercritical fluid to atmospheric pressure, strong cooling and formation of solid carbon dioxide result (Joule – Thomson effect), which causes restrictor plugging and therefore irregularities in the flow and reproducibility of the analysis.

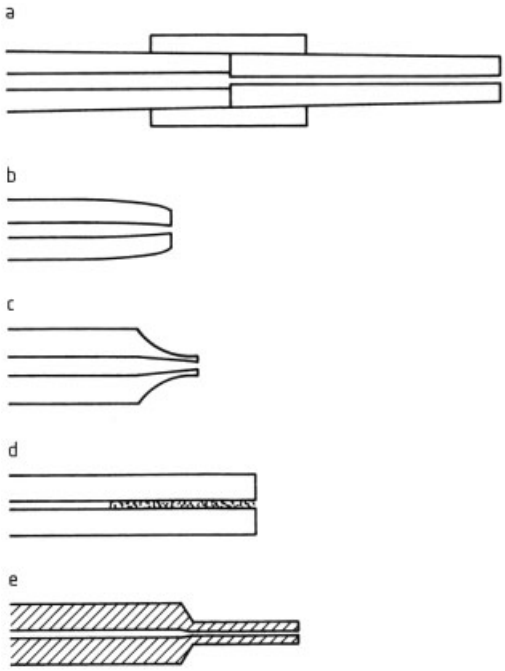

Figure 44. Fixed restrictors
a) Linear capillary restrictor; b) Polished integral restrictor; c) Fast-drawn capillary restrictor; d) Porous frit restrictor; e) Pinched restrictor

10.5. Packed-Column SFC (pc-SFC) and Capillary SFC (c-SFC)

The question of whether packed-column or capillary SFC is preferred, is very dificult to answer. Both techniques have their advantages and disadvantages, depending on the particular separation problem. Lately the trend is more toward the packed-column technique because reproducibility and higher sensitivity are ensured. In addition, independent pressure and flow control is possible only in packed-column SFC.

10.5.1. Capillary SFC (c-SFC)

Capillary SFC columns are usually manufactured from fused silica, with an internal diameter of 25 – 100 µm and a total column length between 1 and 35 m. The most commonly used are 3 – 10-m, 50-µm-i.d., fused-silica columns. A column like this has a very low supercritical fluid flow rate, from several microliters per minute to ca. 0.01 mL/min which corresponds to a gas flow rate of about 1 mL/min (see Table 12).

c-SFC has a wider application range than the packed-column technique because the phase ratio is much more advantageous. In other words, in c-SFC the absolute amount of stationary phase interacting with the solute (and therefore also the retention strength) is much lower than in pc-SFC. Thus, the separation in c-SFC (as in GC) is dominated by the volatility of the solutes and their solubility in the mobile phase and not so much by the selectivity gained from the interaction between the analyte and the stationary phase. Nevertheless, because of the very low linear velocity (millimeters per second) and therefore low flow rates (millimeter per minute) in c-SFC, the analysis time is usually much longer than in pc-SFC. Even though the separation efficiency is theoretically higher in capillary SFC than in packed-column SFC, in practice, this difference is often not so pronounced. The coupling of c-SFC with mass spectrometry (c-SFC – MS) is much easier than for LC, and many articles on this hyphenated technique have been published [156, 157].

Table 12. Column dimensions and characteristics in SFC

Column type	Internal diameter,∗ mm	Length,∗ m	Preferred particle size, µm	Supercritical fluid flow (0.4 g/mL), mL/min	Efficiency per meter,∗ N/m	Efficiency per time,∗ N/s
Open tubular capillary	0.025 – 0.1	1 – 35	50 ∗∗	0.0005 – 0.002	2000 – 22 000	13 – 33
Packed capillary	0.1 – 0.5	0.05 – 0.5	3, 5, 10			
Packed columns	2.0 – 4.6	0.03 – 0.25	3, 5, 10	0.008 – 2.7	33 000 – 51 000	31 – 83
Microbore packed	0.5 – 2.0	0.03 – 0.25	3, 5, 10			

∗ Data from [142].
∗∗ 50-µm-i.d. open tubular (capillary) column.

The disadvantages of capillary SFC are that the number of commercially available, stable stationary phases showing different selectivities is limited. Furthermore, problems with reproducibility can occur because of technical difficulties with quantitative sample introduction. Finally, because of the very low flow rates in c-SFC, analysis time is relatively long and independent pressure programming is impossible; the latter can be done only in combination with a flow program.

10.5.2. Packed-Column SFC (pc-SFC)

Typical dimensions of packed columns and flow rates in pc-SFC are given in Table 12. This technique is becoming more and more popular compared to capillary SFC. A wide spectrum of stable stationary phases, showing different selectivities, is available commercially. Reproducible, quantitative sample introduction by means of the HPLC sample loop technique is no problem, nor is independent pressure or flow control.

Very high separation efficiencies (more than 100 000 plates) can be achieved by coupling several packed columns in series (e.g., Fig. 45 shows a separation of triazine pesticides with ten 200×4.6 mm i.d. hypersil silica (5 µm) columns in series – note that the pressure drop over all these columns is only 12.4 MPa. This would not be possible in HPLC since, because of the higher viscosity of the solvents, the back-pressure over the column would become too high (high-pressure shutdown).

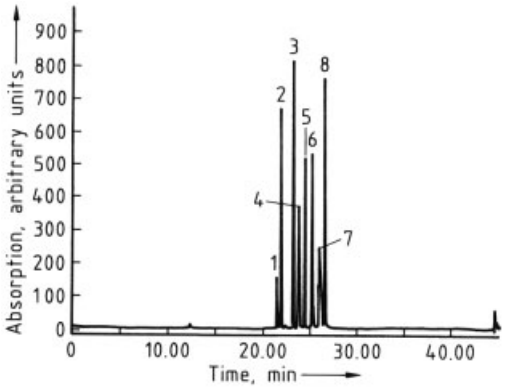

Figure 45. pc-SFC separation of triazine pesticides
Ten 200×4.6 mm i.d. columns hypersil silica columns (particle size 5 µm) in series: program: 2 % methanol, 8 MPa for 5 min, 500 kPa/min + 1 % MeOH/min to 13 MPa and 12 MeOH %; 60 °C, 2.0 mL/min; δp = 12.4 MPa detector: FID
1) Propazine; 2) Atrazine; 3) Anilazine; 4) Simazine (4-doublet); 5) Prometryne; 6) Ametryne; 7) Bladex; 8) Prometon

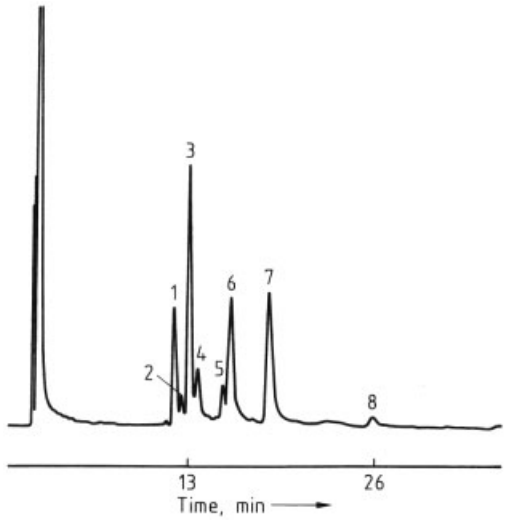

Figure 46. pc-SFC separation of lipids in human blood serum
Columns: 250×4.6 mm ICN silica (3 – 6 µm); 200× 4.6 mm Nucleosil NH_2 (5 µm); 125×4.6 mm Select B (RP-8) column (5 µm) in series; pure CO_2, 40 °C, 14.5 MPa; detector: FID
1) Phytanic acid; 2) 14 : 0; 3) 16 : 3; 4) 18 : 0; 5) 16 : 1; 6) 18 : 1; 7) 18 : 2; 8) 20 : 4 (the figure before the colon denotes the number of carbon atoms in the chain, that after the colon, the number of double bonds in the molecule)

Several columns with different lengths and selectivities can also be connected in series. This enables individual adjustment of the total selectivity of the system depending on the separation problem (selectivity tuning). Figure 46 (see previous page) shows the separation of fatty acid ethyl esters from human blood serum. The phytanic acid ester (ethyl 3,7,11,15-tetramethylhexadecanate), an indicator for a rarely occurring inherited disease (Refsums syndrome) can be separated and quantified from other coeluting fatty acid esters only with the column sequence shown.

The pc-SFC separation technique is rapid (ca. five times faster than HPLC with the same column dimensions). Preparative application and coupling with a mass spectrometer (pc-SFC – MS) are possible. However, capillary SFC, or at least narrow-bore columns are preferred for MS coupling because pc-SFC using conventional 100 – 250 mm × 4 – 4.6 mm i.d. columns usually has flow rates that are too high.

The retention mechanisms of capillary and packed-column SFC differ from one another. Whereas packed column SFC with supercritical carbon dioxide shows a similar retention mechanism to HPLC (separation due mainly to specific interactions between the analyte and the stationary phase), capillary SFC behaves more like capillary GC (separation in pc-SFC according to "solubility" and "volatility" of solute; separation in c-SFC according to "volatility" of solute). Investigations concerning the retention mechanism in reversed-phase packed-column SFC showed [158] that both a hydrophobic and a polar retention mechanism apply. This can be of significant advantage, as shown in Figure 47. Here, a mixture of salmon oil fatty acid ethyl esters is separated by pc-SFC on an aminopropyl column [159]. A combination of an RP mechanism (retention increases with increasing number of carbon atoms of the esters) and an NP mechanism (retention increases with the growing number of double bonds of the fatty acid esters) can be observed. This separation is an example of the potential of pc-SFC to obtain very fast separation of complex sample mixtures. If plain silica columns are used, the retention mechanism corresponds to the normal phase mechanism in LC, with the advantage that pc-SFC is significantly faster than the NP – HPLC technique.

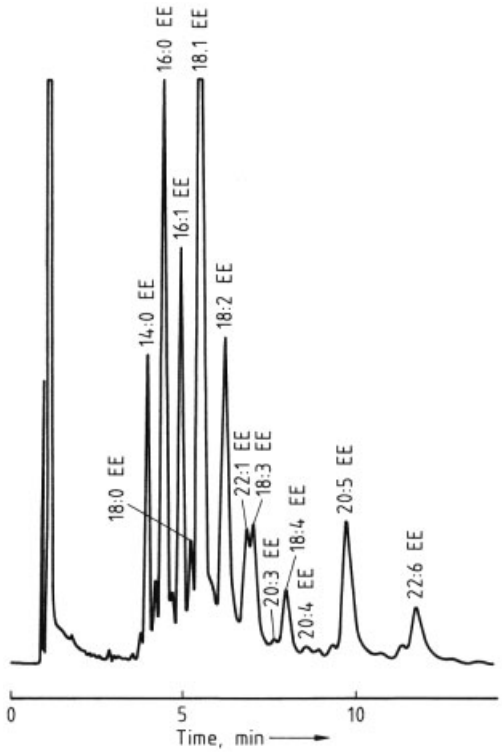

Figure 47. Separation of salmon oil fatty acid ethyl esters with pc-SFC
Column: aminopropyl (particle size 5 µm), 200×4.6 mm i.d.; 34 °C, 1.05 GPa; detector: FID [159]

10.6. Applications

Although the application of SFC is limited to those solutes that are soluble in the particular supercritical fluid (generally carbon dioxide), many interesting separations have been reported in various fields. Polymer analysis (M_r up to ca. 5000 – 8000; even higher with modifier) is certainly one of the most important [160, 161]. Figure 48 shows a typical application in polymer analysis. In the petroleum industry, SFC is becoming more and more important in simulated distillation under mild, nondegradative conditions [142, 162] for analysis of hydrocarbons and of the carcinogenic polycyclic aromatic compounds found in fossil fuels [162] (see Fig. 49). Other typical applications for pc- and c-SFC are the separation of explosives [163], pesticides [164], drugs [165], and pharmaceuticals [166].

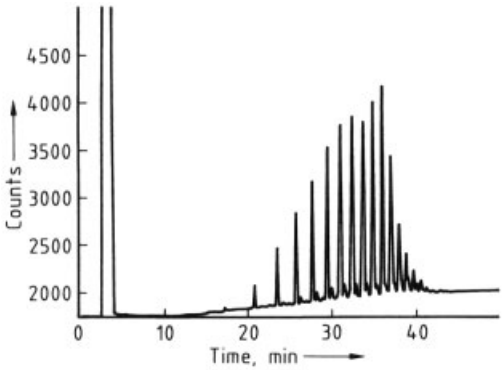

Figure 48. Capillary column SFC separation of Triton X-100 (surfactant)
Column: 50 µm × 10 m SB-methyl-1000 (0.35 µm); mobile phase: pure CO_2; temperature: 120 °C; inlet pressure 10 MPa for 10 min, 1 MPa/min to 35 MPa

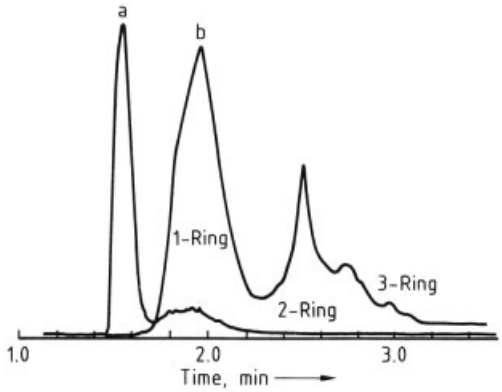

Figure 49. Determination of aromatics in diesel fuel
a) FID detection for saturated compounds and olefins; b) UV detection (at 254 nm) for aromatics
Column: 250 × 4.6 mm Lichrosorb Si-60; CO_2 isobaric, 15 MPa, 28 °C

Figure 50 shows the potential fo packed-column SFC for an efficient and fast control of the chiral purity of pharmaceuticals. Because of the excellent solubility of lipids in supercritical carbon dioxide, SFC in combination with the universal FID detector is a very helpful tool in lipid analysis.

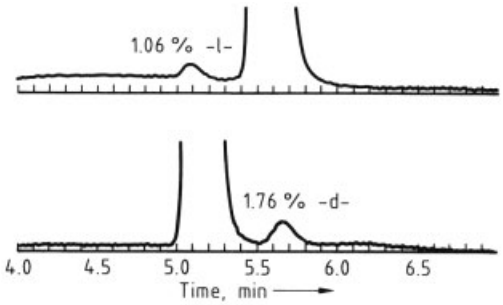

Figure 50. Chiral purity analysis of carbobenzyloxyd, l-alanine at 220 nm with packed-column SFC
Column: 250 × 4.6 mm Ciralcel-OD; 20 % ethanol (+0.2 % TPA/CO_2) 1 mL/min

Although all the applications of pc- and c-SFC would be impossible to mention, their potential in preparative applications should be noted. The latter will certainly become more and more interesting in the future, especially for the food and pharmaceutical industries because very pure extracts are obtainable that are not contaminated with any toxic organic solvents.

In 1997, the worlds first industrial-scale SFC production plant, equipped with separation columns with an internal diameter of over 50 cm and a length of 2 m was designed and started up [167] for the production of high-purity (> 95 %) polyunsaturated fatty acids from fish oil [167].

11. Affinity Chromatography

Affinity chromatography is a special type of adsorption chromatography for the isolation and purification of biologically active macromolecules. It was used successfully for the first time in 1968 for the purification of enzymes [168]. Since then, innumerable proteins (e.g., enzymes, receptors, or antibodies) have been isolated with high enrichment (100 – 10 000 fold) from complex raw extracts in a single chromatographic run. The proteins still possess their full biological activity after separation. The routine application of affinity chromatography has been described in many review articles [169–177].

The high selectivity of affinity chromatography is based on biospecific interactions, such as those occurring between two molecules in natural biological processes. One interaction partner (ligand) is covalently attached (immobilization) to an insoluble carrier, while the corresponding partner (frequently a protein) is reversibly adsorbed by the ligand because of its complementary biospecific properties.

In laboratory practice, an affinity matrix (stationary phase) is tailor-made for the protein to

be purified and filled into a chromatography column. For process applications many different affinity gels are commercially available. The raw extract is then passed through the column by using a physiological buffer as the mobile phase. In this process, the desired product is adsorbed selectively by the ligand and all unwanted components are washed away.

The chemical composition of the mobile phase is then changed to permit desorption and isolation (elution) of the protein (Fig. 51).

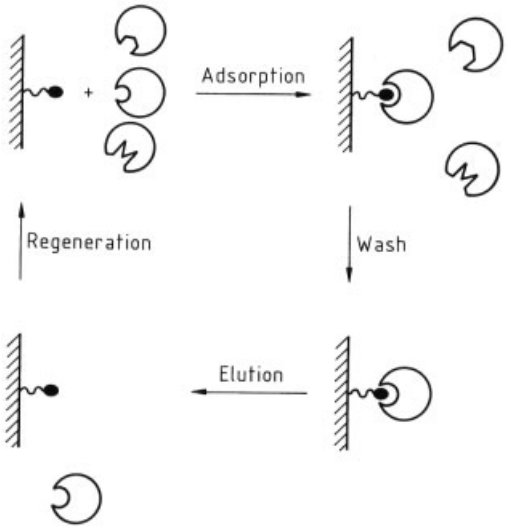

Figure 51. Schematic representation of affinity chromatography

11.1. Preparation of an Affinity Matrix

11.1.1. Selection of a Carrier [178, 179]

The preparation of an affinity matrix requires a suitable carrier that covalently attaches the selected ligand (immobilization). An ideal carrier should have the following properties:

1) A hydrophilic neutral surface – avoidance of nonspecific adsorption
2) Macroporosity – good permeability
3) Mechanical stability – high flow rates, stability to pressure
4) Chemical and biological stability – insolubility in various solvents, resistance to microbial attack
5) Functional groups on the surface – covalent binding of the ligand

The carriers used can be divided into three groups:

Polysaccharides	agarose
	dextran
	cellulose
Synthetic polymers	polyacrylamides
	polyacrylates
	poly(vinyl alcohols)
Inorganic materials	SiO_2, ZrO_2, TiO_2
	magnetic particles

Agarose, a polysaccharide consisting of galactose units, is so far the carrier of choice because of its favorable chemical and physical properties. However, mechanical instability sometimes results in short column lifetime. In comparison, synthetic polymers and inorganic materials are mechanically more stable.

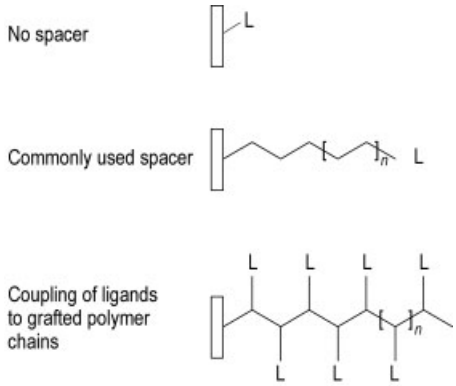

Figure 52. Three general possibilities for the covalent fixation of ligands to a matrix surface: direct coupling, coupling to spacers, and coupling to grafted polymer chains. L = ligand.

The spatial availability of immobilized ligands is often substantially improved by coupling ligands to spacers that keep them at a distinct distance from the matrix [180]. Basically, there are three general possibilities for the covalent fixation of ligands to the surface of appropriate support materials (Fig. 52):

1) Direct coupling to the surface (i.e., with no spacers)
2) Commonly used spacer arms : linear aliphatic hydrocarbons, 6-aminohexanoic acid, hexamethylenediamine, poly(ethylene glycol) [181]. One of the groups (often a primary amine, $-NH_2$) is attached to the matrix, while the group at the other end is selected

on the basis of the ligand to be bound. The latter group which also is called the terminal group is usually a carboxyl (–COOH) or amino group.

3) Coupling of ligands to grafted polymer chains which freely rotate (i.e., in a "tentacle" arrangement) [182].

11.1.2. Immobilization of the Ligand

Covalent binding of a ligand to a carrier requires the presence of functional groups on both binding partners.

Generally, the following groups are available for immobilization of the ligand: $-NH_2$ (amino), $-COOH$ (carboxyl), $-CHO$ (aldehyde), $-SH$ (thiol), or $-OH$ (hydroxyl). High-molecular ligands (e.g., proteins) are usually bound through ε-NH_2 groups of lysine residues or occasionally through carboxyl groups. The functional group of the ligand must be mainly located in the nonspecific region of the molecule that is of no significance for reversible complex formation with the complementary molecule to be purified.

The functional groups of the carrier are either already present in the natural structure (e.g., OH groups of polysaccharides) or introduced later by the chemical functionalization of the surface:

$-\overset{\overset{O}{/\ \backslash}}{CH-CH_2}$	epoxide (a)
$-CHO$	aldehyde (b)
$-NCS$	isothiocyanate (c)
$-SO_2-CH=CH_2$	vinyl sulfone (d)
$-OH$	hydroxyl (e)
$-COOH$	carboxyl (f)
$-NH_2$	amine (g)

A carrier with reactive functional groups (a) – (d) can be used directly for the immobilization of ligands (e.g., via their NH_2 groups). In the case of epoxides (a), however, the coupling yields obtained are low. In the case of aldehydes (b), chemically instable ligand bonds (Schiff base) are formed that require further reduction (e.g., with boron hydride). Isothiocyanates (c) and vinyl sulfones (d) have the disadvantage that after immobilization, chargeable ligand bonds (thiourea and amine derivatives) are formed. These groups can act as weak ion exchangers and cause additional nonspecific adsorptions.

Functional groups (e) – (g) must be converted to a chemically reactive form (activation) before coupling to the ligand can occur. After immobilization has been carried out, any unreacted reactive groups on the carrier must be deactivated to prevent chemical reactions with the raw extract. This is frequently effected by additional coupling to a low-molecular ligand (e.g., ethanolamine) or by hydrolysis of the remaining activated groups.

Methods of Activation. Functional groups such as hydroxyl, carboxyl, or amino groups (e) – (g) must be converted to an activated form before they can be attached covalently to the ligand.

The classical method for activating *hydroxyl groups* involves reaction with cyanogen bromide [171, 183] to give a reactive cyanate ester (Fig. 53 A, see next page). However, the coupling of ligands (e.g., via ε-NH_2 groups of lysine residues) results in chemically unstable isourea linkages that can cause bleeding of the ligand during affinity chromatography. This disadvantage can be avoided by activation with other reagents such as 1,1′-carbonyldiimidazole (Fig. 53 C) [184], *p*-nitrophenyl and *N*-hydroxysuccinimide chloroformates (Fig. 53 B) [185, 186], or sulfonyl chlorides [187], such as 4-toluenesulfonyl chloride and 2,2,2-trifluoroethanesulfonyl chloride (Fig. 53 D). The resulting imidazoyl, carbonate, and sulfonyl derivatives react, for example, with the amino groups of the ligand, with high coupling yields, and the ligand is attached to the carrier through chemically stable carbamate or amine bonds. The carbamate bond also has the advantage of being neutral and, consequently, causing no undesirable, nonspecific interactions. Unreacted activated groups can be hydrolyzed easily to reform the original neutral hydroxyl groups.

Carboxyl groups can be converted to active esters (e.g., with *p*-nitrophenol or *N*-hydroxysuccinimide), which react with the ligand with high coupling yields. Furthermore, carboxyl groups can be activated with water-soluble carbodiimide.

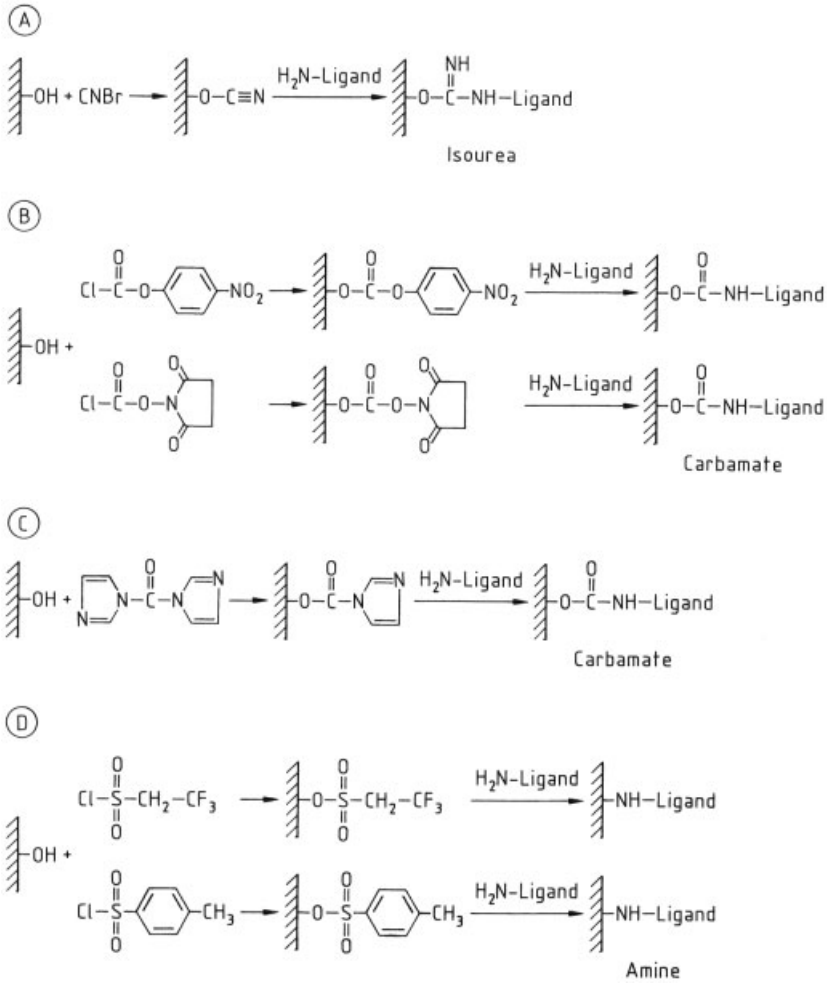

Figure 53. Activation and coupling of ligands to base supports with hydroxyl functions
A) Cyanogen bromide; B) Chloroformates; C) Carbonyldiimidazole; D) Sulfonyl chlorides

However, unreacted carboxyl groups act as weak ion exchangers, leading to unwanted nonspecific adsorptions. This disadvantage is also encountered with base matrices that are functionalized with amino groups.

11.2. Ligands

A ligand suited to the production of an affinity matrix should have the following properties:

1) Ability to form reversible complexes with the protein to be isolated and purified
2) Suitable specificity for the planned application
3) Large enough complexation constant to obtain stable complexes or sufficient delays during the chromatographic process
4) Complex that can be dissociated easily by simple changes in the mobile phase without irreversible loss of the biological activity of the protein

A distinction is made between low-molecular and high-molecular ligands as well as between group-specific and monospecific ligands. Group-specific ligands exhibit binding affinity for a range of substances that have similar properties with regard to function or structure. Typical examples are the triazine dyes Cibacron blue F 3 G-A and Procion red HE-3 B, which are used

as ligands for many dehydrogenases, kinases, phosphatases, and other proteins. Monospecific ligands (e.g., antibodies) are "tailor-made" and can be used for only one application. In general, the binding of monospecific ligands to their complementary substance is substantially stronger (i.e., unfavorable elution conditions are occasionally required).

11.2.1. Group-Specific Ligands

11.2.1.1. Low-Molecular Ligands

The low-molecular, group-specific ligands belong to the largest class of ligands used in affinity chromatography:

Ligand	*Target molecule*
Cibacron blue F 3 G-A	kinases, phosphatases, dehydrogenases albumin, clotting factors II and IX, interferon, lipoproteins
Procion red HE-3 B	dehydrogenases, carboxypeptidase G, interferon, plasminogen, dopamine β-monooxygenase, inhibin
5′-AMP	NAD^+-dependent dehydrogenases, ATP (adenosine 5′-triphosphate)-dependent kinases
2′,5′-ADP	$NADP^+$-dependent dehydrogenases
NAD^+ and $NADP^+$	dehydrogenases
Phenylboric acid	nucleosides, nucleotides, carbohydrates, glycoproteins
Benzamidine	trypsin, urokinase, pronase E, glu- and lys-plasminogen

The coenzymes nicotinamide – adenine dinucleotide (NAD) and nicotinamide – adenine dinucleotide phosphate (NADP), their reduced forms, and adenosine 5′-monophosphate (5′-AMP) and adenosine 2′,5′-diphosphate (2′,5′-ADP) represent well-known examples of this class. They can be considered for the purification of ca. 25 % of the known enzymes (especially dehydrogenases).

Moreover, the triazine dyes (e.g., Cibacron blue F 3 G-A and Procion red HE-3 B) [188, 189] mentioned above are also used as ligands for the purification of these proteins. Triazine dyes can replace ligands of the NAD group because they have similar spatial structures. In addition, they bind a series of blood proteins (e.g., albumins, lipoproteins, clotting factors, and interferon).

Thiophilic ligands [190, 191] form a separate class of group-specific ligands that possess high selectivity with respect to immunoglobulins. They are low-molecular compounds, and contain sulfone and thioether groupings (see below), which show a strong affinity for mono- and polyclonal antibodies in the presence of ammonium sulfate (ca. 1 mol/L).

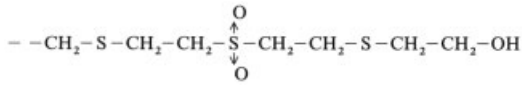

Immobilized metal ion affinity chromatography (IMAC) uses gels with group-specific ligands which display selectivity for certain amino residues (histidine, cysteine, tryptophan) exposed at the outer surface of proteins [192]. The ligands are immobilized chelating agents [e.g., iminodiacetic acid (IDA) or tris (carboxymethyl)ethylenediamine (TED)] with bound transition metal ions. The metal ions are mostly by Cu^{2+}, Ni^{2+}, Zn^{2+}, and Co^{2+}, with decreasing complex stability and affinity [193]. The amino acids at the protein surface are bound to a free coordination site of the immobilized complex (Fig. 54) [194].

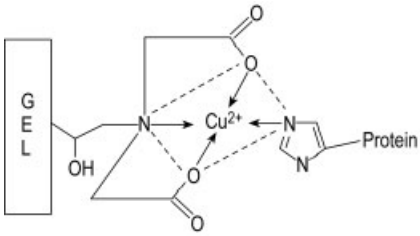

Figure 54. Schematic of binding of a histidine protein residue to an immobilized IDA copper(II) complex

Adsorbed proteins can be recovered by lowering the pH to 3 or 4 or by competitive elution with an increasing concentration of glycine, imidazole, or histidine [195]. Through advances in molecular biology, histidine residues can be inserted into the protein sequence (fusion protein), resluting in improved separation in IMAC. The modified protein has a high affinity to the immobilized metal ions due to the addition of the histidine residues in the primary sequence. The fusion part can cleaved chemically or with a specific proteinase after purification.

11.2.1.2. Macromolecular Ligands

Macromolecular group-specific ligands include proteins that are used widely in the purification of biologically relevant macromolecules. Important macromolecular ligands are lectins (for the isolation of glycoproteins), protein A and protein G (for the purification of immunoglobulins), calmodulin (for Ca-dependent enzymes), and the polysaccharide heparin (for the purification of clotting factors and other plasma proteins).

Lectins [196]. Lectins exhibit a marked affinity for sugar molecules. They are perfectly suited to the isolation and purification of soluble and membrane-bound glycoproteins and polysaccharides (e.g., enzymes, hormones, plasma proteins, antigens, antibodies, and blood group substances). Among the available lectins, Con A (from *Canavalia ensiformis*) is the most widely used because of its specificity with respect to the frequently occurring α-D-glucose and α-D-mannose structural units. The desorption of the component to be purified from the ligand elution is often achieved by the addition of these low-molecular sugars to the mobile phase (see Section 11.3).

Lentil lectin (from *Lens culinaris*) has the same specificity as Con A, but a lower binding constant. For this reason, lentil lectin is suited to the purification of membrane-bound glycoproteins, which often have an exceedingly strong binding affinity for Con A.

Proteins A and G [197, 198]. Apart from the thiophilic ligands mentioned above, proteins A and G from bacterial cell walls are also suited to the purification of immunoglobulins. Protein A binds different subclasses of immunoglobulins to different extents. Accordingly, elution is often achieved with a decreasing stepwise, pH gradient. Mouse immunoglobulins (Igs) are eluted, for example, with 0.1 mol/L citrate buffer at the following pH values:

	IgG 1	IgG 2a	IgG 2b	IgG 3
pH	6.0	5.0	4.0	4.5

Protein G in its natural form can bind not only immunoglobulins, but also albumin. For this reason, it is applied in a genetically engineered form in which the albumin binding domain is eliminated (Fig. 55). Immunoglobulins generally display a higher binding affinity for protein G than for protein A.

11.2.2. Monospecific Ligands

Highly specific protein – protein interactions play a large role in many natural biological processes. For instance, the transport and transmission of information proceed via highly specific receptors that can frequently interact with only one molecular species (e.g., hormone or transmitter). In spite of this, the affinity chromatographic isolation of receptors and other highly specific proteins is rather an exception. However, antibodies occupy a special position as monospecific macromolecular ligands.

Antibodies and Antigens (Immunoaffinity Chromatography) [199, 200]. Immunoaffinity chromatography refers to the immobilization of antibodies (monoclonal or polyclonal) or antigens on an insoluble carrier (immunosorbent) for isolation and purification of the corresponding partner.

Antibodies are ligands of special interest because they show high biospecificity and strong binding to the molecules that elicited their synthesis, the antigens. The antibody – antigen interaction takes place between a particular part of the antibody (the antigen-binding site) and its complementary part on the antigen (the antigenic determinant or epitope).

Polyclonal antibodies represent a wide spectrum of antibody molecules that all have specificities with respect to different epitopes and, therefore, different avidities. The association constants of an antigen and its corresponding antibody may range from 10^4 to 10^{11} L/mol. As a result, multipoint binding of the antigen to two or more different antibodies can occur on an immunosorbent prepared from polyclonal antibodies, resulting in such tight binding that dissociation and recovery of the antigen may be difficult.

In contrast to polyclonal antibodies, *monoclonal antibodies* show absolute specificity for a single epitope, the smallest immunologically recognized submolecular group of an antigen.

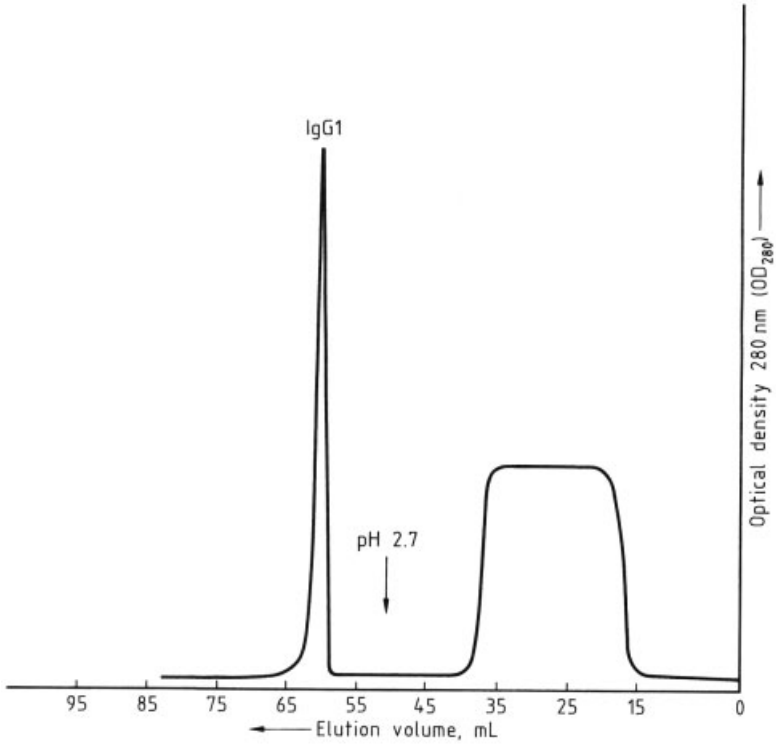

Figure 55. Isolation and purification of IgGl with genetically engineered protein G immobilized onto Sepharose

Table 13. Proteins of potential therapeutic value purified by immunoaffinity chromatography

Protein	Potential use
Interferons	
α-Interferon	antiviral, antitumor drug
β-Interferon	
γ-Interferon	
Interleukin 2	antitumor drug; also used against immuno-deficiency syndrome
Anticlotting factor	
Urokinase	dissolving blood clots
Urokinase-type plasminogen activator	
Zymogen of plasminogen activator	
Plasminogen activator	
Tissue plasminogen activator	
Human protein C	
Clotting factor	
Factor VIII	blood coagulant
Factor IX	
α- and β-Kallikrein	
Insulin-like growth hormone	growth hormone
Vaccines	
Human hepatitis B surface antigen vaccine	
Rubella virus antigen	
Tetanus toxin	
Polio virus	

Therefore, from a group of monoclonal antibodies the one most suited to antigen purification can be selected. Consequently, immunoaffinity chromatography has gained importance because monoclonal antibodies can be produced in large quantities by a hybridoma technology developed by KÖHLER and MILSTEIN [201] (→Monoclonal Antibodies, Chap. 1.). This allows the use of immunoaffinity chromatography even for industrial-scale applications. Several proteins of potential therapeutic value are now produced by bioengineering techniques and purified with immunoaffinity chromatography (Table 13).

11.3. Elution [176]

Above all, successful affinity chromatography requires suitable desorption conditions (i.e., conditions that permit elution of the adsorbed substance with full biological activity). In principle, desorption is achieved by cleaving the binding sites responsible for ligand – protein interaction. These are mainly electrostatic, hydrophobic, and hydrogen bonds that can be specifi-

cally or nonspecifically disrupted. A *nonspecific method* of eluting an adsorbed substance is by altering the pH or ionic strength of the mobile phase. A pH decrease to pH 2 – 4 or an increase in ionic strength through a NaCl gradient, for instance, typically causes the dissociation of mainly electrostatic complex binding site. Ligand – protein complexes with high binding constants, such as those encountered in immunoaffinity chromatography, can often be eluted only under drastic conditions that can lead to denaturation. For this purpose, chaotropic salts (e.g., KSCN, KI) are used in concentrations of 1 – 3 mol/L. Urea or guanidinium chloride (4 – 6 mol/L) is also employed for the elution of strongly adsorbed substances.

In the case of *specific elution*, low-molecular substances with a certain affinity for the ligand or the adsorbed protein are added to the mobile phase. They displace the adsorbed protein from the ligand by competitive binding. This method of elution is frequently used with group-specific ligands (e.g., lectins). Glycoproteins can be eluted gently from immobilized lectins by the addition of low concentrations (10 – 100 mmol/L) of a monosaccharide to the mobile phase. Another example is the elution of dehydrogenases from coenzyme affinity supports by using NAD and its derivatives. In general, specific elution is a gentle method that usually ensures retention of biological activity.

12. References

1. J. J. Kirkland, *Anal. Chem.* **43** (1971) no. 12, 36 A.
2. M. Tswett, *Tr. Protok. Varshav. Obshch. Estestvoistpyt. Otd. Biol.* **14** (1903, Publ. 1905) 20 (Reprinted and translated in G. Hesse, H. Weil: *Michael Tswett's erste chromatographische Schrift,* Woelm, Eschwege 1954, pp. 37.
3. R. Kuhn, E. Lederer, *Naturwissenschaften* **19** (1931) 306.
4. R. Kuhn, E. Lederer, *Ber. Dtsch. Chem. Ges.* **64** (1931) 1349 – 1357.
5. A. Martin, R. Synge, *Biochem. J.* **35** (1941) 1358 – 1368.
6. F. Spedding, *Discuss. Faraday Soc.* **7** (1949) 214 – 231.
7. E. Tompkins, *Discuss. Faraday Soc.* **7** (1949) 232 – 237.
8. J. Porath, P. Flodin, *Nature (London)* **183** (1959) 1657 – 1659.
9. E. Piel, *Anal. Chem.* **38** (1966) 370 – 372.
10. J. J. Van Deemter, F. J. Zuiderweg, A. Klinkenberg, *Chem. Eng. Sci.* **5** (1956) 271.
11. H. Engelhardt (ed.): *Practice of High Performance Analysis,* Springer Verlag, Berlin 1986.
12. F. C. Smith, R. C. Chang, *CRC Crit. Rev. Anal. Chem.* **9** (1980) 197.
13. P. L. Bailey, *Anal. Chem.* **50** (1978) 698 A.
14. K. Stulik, V. Pacakova, *CRC Crit. Rev. Anal. Chem.* **14** (1982) 297.
15. C. J. Blank, *J. Chromatogr.* **117** (1976) 35.
16. L. Roth: *Sicherheitsdaten/MAK-Werte,* Verlag moderne Industrie, W. Dummer & Co., München 1978.
17. H. Schrenker, *Int. Lab.* (1978) 67.
18. I. Halasz, P. Vogtel, *J. Chromatogr.* **142** (1977) 241.
19. L. R. Snyder, J. K. Kirkland: *Introduction to Modern Liquid Chromatography,* 2nd ed., John Wiley and Sons, New York 1979.
20. *Rheodyne, Products for Liquid Chromatography,* Catalog no. 3, Cotati, Calif., USA, 1986.
21. J. Meister, H. Engelhardt, *Nachr. Chem. Tech. Lab.* **40** (1992) no. 10.
22. A. Tiselius, D. Claesson, *Ark. Kemi Mineral. Geol.* **15** (1942) no. 22.
23. Refractive gradient detector
24. H. Engelhardt: *Hochdruck-Flüssigkeits-Chromatographie,* Springer Verlag, Berlin 1975.
25. U. A. T. Brinkmann, G. J. De Jong, C. Gooijer, *Methodol. Surv. Biochem. Anal.* **18** (1988) 321 – 338.
26. T. M. Vickrey (ed.): *Liquid Chromatography Detectors, Chromatographic Science,* vol. 23, Marcel Dekker, New York 1983.
27. R. P. W. Scott, *J. Chromatogr. Libr. Liquid Chromatography Detectors* **11** (1977).
28. R. Schulz, H. Engelhardt, *Chromatographia* **29** (1990) nos. 11/12, 517 – 522.
29. M. Varadi, J. Balla, E. Pungor, *Pure Appl. Chem.* **51** (1979) 1177.
30. G. Henze, R. Neeb: *Elektrochemische Analytik,* Springer Verlag, Berlin 1986.
31. G. Henze: *Polarographie und Voltammetrie; Grundlagen and analytische Praxis,* Springer Verlag, Berlin 2000.
32. F. A. Schultz, D. E. Mathis, *Anal. Chem.* **46** (1974) 2253.
33. A. Manz, W. Simon, *J. Chromatogr. Sci.* **21** (1983) 326.

34. R. S. Deelder, H. A. J. Linssen, J. G. Koen, A. J. B. Beeren, *J. Chromatogr.* **203** (1981) 153.
35. H. Akaiwa, H. Kawamoto, M. Osumi, *Talanta* **29** (1982) 689.
36. T. Deguchi, T. Kuma, H. Nagai, *J. Chromatogr.* **152** (1978) 379.
37. T. H. Ryan: *Electrochemical Detectors,* Plenum Press, New York 1984.
38. K. Brunt: *Electrochemical Detectors in Trace Analysis,* vol. 1, Academic Press, New York 1981.
39. P. T. Kissinger, C. Refshauge, R. Dreiling, R. N. Adams, *Anal. Lett.* **6** (1973) 465.
40. B. Fleet, C. J. Little, *J. Chromatogr. Sci.* **12** (1974) 747.
41. W. Kemula, *Rocz. Chem.* **26** (1952) 281.
42. P. Just, M. Karakaplan, G. Henze, F. Scholz, *Fresenius J. Anal. Chem.* **32** (1993) 345.
43. W. W. Kubiak, *Electroanalysis NBS* (1989) 379.
44. J. Frank, *Chimia* **35** (1981) 24.
45. P. T. Kissinger, C. S. Brunett, K. Bratin, J. R. Rice, *Spec. Publ. (U.S.)* **519** (1979) 705.
46. S. Yao, A. Meyer, G. Henze, *Fresenius J. Anal. Chem.* **339** (1991) 207.
47. H.-P. Nirmaier, E. Fischer, G. Henze, *GIT Labor-Fachzeitschrift* **42** (1998) 12.
48. J. Schiewe, A. M. Bond, G. Henze, in G. Henze, M. Köhler, J. P. Lay (eds.): *Voltammetrische Spurenanalyse mit Mikroelektroden in Umweltdiagnostik mit Mikrosystemen,* Wiley-VCH Verlag, Weinheim 1999, p. 121.
49. H.-P. Nirmaier, E. Fischer, G. Henze, *Electroanalysis,* **10** (1998) 187.
50. R. Wintersteiger, G. Berlitz, *GIT Suppl. 3* (1989) 19.
51. J. P. Hart, *Electroanalysis of Biologically Important Compounds,* Ellis Horwood, Chichester 1990.
52. D. M. Radzik, S. M. Lunte: "Application of Liquid Chromatography/Electrochemistry in Pharmaceutical and Biochemical Analysis," *CRC Critical Reviews in Analytical Chemistry* **20** (1989) no. 5, 317.
53. K. Štulik, V. Pacáková: "Electrochemical Detection in High-Performance Liquid Chromatography," *CRC Critical Reviews in Analytical Chemistry* **14** (1982) no. 4, 297.
54. K. Štulik, V. Pacáková: "Electrochemical Detection Techniques in High-Performance Liquid Chromatography," *J. Electroanal. Chem.* **129** (1981) 1.
55. G. Patonay, HPLC Detection, Newer Methods, VCH **1992.**
56. K. Štulik, V. Pacáková: "Electrochemical Detection Techniques in High-Performance Liquid Chromatography," *J. Electroanal. Chem.* **129** (1991) 1.
57. R. J. Rucki: "Electrochemical Detectors for flowing Liquid Systems," *Talanta* **27** (1980) 147.
58. D. C. Johnson, S. G. Weber, A. M. Bond, R. M. Wightman, R. E. Shoup, J. S. Krull: "Electroanalytical Voltammetry in Flowing Solutions," *Anal. Chim. Acta* **180** (1986) 187.
59. D. S. Austin-Harrison, D. C. Johnson: "Pulsed Amperometric Detection Based on Direct and Indirect Anodic Reactions," *Electroanalysis* **1** (1989) 189.
60. R. Samuelsson, J. O'Dea, J. Osteryoung, *Anal. Chem.* **52** (1980) 2215.
61. J. G. White, R. L. St. Claire, J. W. Jorgenson, *Anal. Chem.* **58** (1986) 293.
62. P. T. Kissinger, *Anal. Chem.* **49** (1977) 447 A.
63. M. Goto et al., *J. Chromatogr.* **226** (1981) 33.
64. E. L. Craig, P. T. Kissinger, *Anal. Chem.* **55** (1983) 1458.
65. J. Dutrieu, Y. A. Delmotte, *Fresenius J. Anal. Chem.* **314** (1983) 416.
66. C. L. Putzig et al., *Anal. Chem.* **64** (1992) 270 R – 302 R.
67. H.-P. Nirmaier, E. Fischer, A. Meyer, G. Henze, *J. Chromatography A,* 1996, **730**, 169.
68. K. K. Unger, G. Gilge, J. N. Kinkel, M. T. W. Hearn, J. Chromatogr. **359**, 61 (1986).
69. H. Chen, C. Horvath, J. Chromatogr. **705**, 3 (1995).
70. K. Stulik, V. Pacáková, J. Suchánková, H. Claessens, Anal. Chim. Acta **352**, 1 (1997).
71. R. Ohnmacht, B. Boros, I. Kiss, L. Jelinek, *Chromatographia* **50** (1999) no. 1/2.
72. I. M. Mutton, *Chromatographia* **47** (1998) no. 5/6, 291 – 298.
73. K. K. Unger (ed.): "Packings and Stationary Phases in Chromatographic Techniques," *Chromatographic Science Series,* vol. 47, Marcel Dekker, New York 1990.
74. J. Persek, H. Lin, *Chromatographia* **28** (1989) 565.
75. I. Halasz, I. Sebastian, *Angew. Chem. Int. Ed. Engl.* **8** (1969) 453.
76. B. G. Herslöf, *HRC & CC, J. High Resolut. Chromatogr. Chromatogr. Commun.* **4** (1981) 471 – 473.
77. A. Sándy et al., *Chromatographia* **45** (1997) 206 – 214.
78. M. Vecchi, G. Englert, R. Maurer, V. Meduna, *Helv. Chim. Acta* **64** (1981) 2746.

79. E. R. Majors, *LC-GC Int.,* **8** (1995) no. 7, 368 – 374.
80. K. Aitzetmüller, *J. Chromatogr.* **113** (1975) 231 – 266.
81. H. Small, T. S. Stevens, W. C. Baumann, *Anal. Chem.* **47** (1975) 1801.
82. J. W. BcBain, *Kolloid Z.* **40** (1926) 1.
83. I. Porath, P. Flodin, *Nature (London)* **183** (1959) 1657.
84. J. C. Moore, *J. Polymer Sci. Part A* **2** (1964) 835.
85. Macherey & Nagel & Co.: "HPLC", Düren 1991.
86. K. K. Unger: "Adsorbents in Column Liquid Chromatography," in [24]
87. Z. Deyl, K. Macek, J. Janak (eds.): "Liquid Column Chromatography," *J. Chromatogr. Libr.* **3** (1975).
88. V. Meyer: *Praxis der Hochdruck-Flüssigkeits-Chromatographie,* Diesterweg-Salle-Sauerländer, Frankfurt 1979.
89. K. Aitzetmüller, J. Koch, *J. Chromatogr.* **145** (1987) 195.
90. J. E. Campbell, M. Hewins, R. J. Lych, D. D. Shrewsbury, *Chromatographia* **16** (1982) 162 – 165.
91. S. A. Westwood (ed.): *Supercritical Fluid Extraction and its Use in Sample Preparation,* Blackie Academic & Professional, London 1993.
92. K. Jinno (ed.): "Hyphenated Techniques in Supercritical Fluid Chromatography and Extraction," *J. Chromatogr. Libr.* **53** (1992).
93. H. Godel, Th. Graser, P. Földi, *J. Chromatogr.* **297** (1984) 49.
94. I. S. Krull (ed.): "Reaction Detection in Liquid Chromatography," *Chromatographic Science Series,* **vol. 34,** Marcel Dekker, New York 1986.
95. H. Engelhardt, H. Elgass, *Chromatographia* **22** (1986) no. 1 – 6, June.
96. A. H. M. T. Scholten, U. A. T. Brinkman, R. W. Frei, *J. Chromatogr.* **205** (1981) 229.
97. P. Walkling, Ph. D. Thesis, Frankfurt 1968.
98. I. Halasz, P. Walkling, *Ber. Bunsenges. Phys. Chem.* **74** (1970) 66.
99. K. Hofmann, I. Halasz, *J. Chromatogr.* **199** (1980) 3.
100. H. Engelhardt, U. D. Neue, *Chromatographia* **15** (1982) 403.
101. U. D. Neue, Ph. D. Thesis, Saarbrücken 1976.
102. D. Moore, W. H. Stein, D. H. Speckman, *Anal. Chem.* **30** (1958) 1190.
103. D. H. Speckman, W. H. Stein, S. Moore, *Anal. Chem.* **30** (1958) 1190.
104. M. Weigele, S. L. De Bernardo, J. P. Tengi, W. Leimgruber, *J. Am. Chem. Soc.* **94** (1972) 5927.
105. M. Roth, *Anal. Chem.* **43** (1971) 880.
106. S. Stein et al., *Arch. Biochem. Biophys.* **155** (1973) 202.
107. D. Gottschalk, H. Körner, J. Puls in H. Engelhardt, K.-P. Hupe (eds.): *Kopplungsverfahren in der HPLC,* GIT Verlag, Darmstadt 1985.
108. P. Vratny, U. A. T. Brinkman, R. W. Frei, *Anal. Chem.* **57** (1985) 224.
109. L. Nondek, R. W. Frei, U. A. T. Brinkman, *J. Chromatogr.* **282** (1983) 141.
110. H. A. Moye, C. J. Miles, S. J. Scherer, *J. Agric. Food Chem.* **31** (1983) 69.
111. J. Schrijver et al., *Ann. Clin. Biochem.* **19** (1982) 52.
112. J. Schrijver, A. J. Speek, W. H. P. Schreurs, *Int. J. Vitam. Nutr. Res.* **51** (1981) 216.
113. J. T. Vanderslice, D. J. Higgs, *J. Chromatogr. Sci.* **22** (1984) 485.
114. Deans-säulenschalten GC.
115. J. F. K. Huber et al., *J. Chromatogr.* **83** (1973) 267.
116. S. Wielinski, A. Olszanowski, *Chromatographia* **50** (1999) no. 1/2, 109 – 112.
117. A. P. Bruins, *Adv. Mass Spectrom.* **11 A** (1989) 23 – 31.
118. A. L. Burlingame, D. S. Millington, D. L. Norwood, D. H. Russel, *Anal. Chem.* **62** (1990) 268 R – 303 R.
119. H. Kataoka, J. Pawliszyn, *Chromatographia* **50** (1999) no. 9/10, 532 – 538.
120. M L. Nedved, S. Habbi-Gondarzi, B. Ganem, J.-D. Henion, *Anal. Chem.* **68** (1996) 4228 – 4236.
121. Cagnaird de la Tour, *Ann. Chim. Phys.* **21** (1822) 127.
122. T. Anrews, *Philos. Trans. R. Soc. London* **159** (1869) 575.
123. J. B. Hannay, J. Hogarth, *Proc. R. Soc. London* **30** (1880) 178.
124. C. M. White: *Modern Supercritical Fluid Chromatography,* Hüthig, Heidelberg 1988.
125. E. Klesper, A. H. Corwin, D. A. Turner, *J. Org. Chem.* **27** (1962) 700.
126. S. T. Sie, W. van Beersum, G. W. A. Rijnders, *Sep. Sci.* **1** (1966) 459.
127. S. T. Sie, G. W. A. Rijnders, *Sep. Sci.* **2** (1967) 699.
128. S. T. Sie, G. W. A. Rijnders, *Sep. Sci.* **2** (1967) 729.

129. S. T. Sie, G. W. A. Rijnders, *Sep. Sci.* **2** (1967) 755.
130. S. T. Sie, J. P. A. Bleumer, G. W. A. Rijnders in C. L. A. Harbourn (ed.): *Gas Chromatography 1968,* The Institute of Petroleum, London 1969, p. 235.
131. J. C. Giddings, *Sep. Sci.* **1** (1966) 73.
132. J. C. Giddings, M. N. Myers, L. McLaren, R. A. Keller; *Science* (*Washington, D.C. 1883)* **162** (1968) 67.
133. J. C. Giddings, M. N. Myers, J. W. King, *J. Chromatogr. Sci.* **7** (1969) 276.
134. J. C. Giddings, L. M. Bowman, N. M. Myers, *Anal. Chem.* **49** (1977) 243.
135. S. T. Sie, G. W. A. Rijnders, *Anal. Chim. Acta* **38** (1967) 31.
136. R. E. Jentoft, T. H. Gouw, *J. Chromatogr. Sci.* **8** (1970) 138.
137. R. E. Jentoft, T. H. Gouw, *J. Polym. Sci. Polym. Lett. Ed.* **7** (1969) 811.
138. D. Bartmann, G. M. Schneider, *J. Chromatogr.* **83** (1973) 135.
139. U. van Wasen, G. M. Schneider, *Chromatographia* **8** (1975) 274.
140. U. van Wasen, G. M. Schneider, *J. Phys. Chem.* **84** (1980) 229.
141. M. Novotny et al., *Anal. Chem.* **53** (1981) 407 A.
142. M. L. Lee, K. E. Markides (eds.): *Analytical Supercritical Fluid Chromatography and Extraction,* Chromatography Conferences, Inc., Provo, Utah, 1990.
143. K. P. Johnston, J. M. L. Penninger (eds.): "Supercritical Fluid Science and Technology," *ACS Symp. Ser.* **406** (1989).
144. M. L. Lee, *J. Microcol. Sep.* **4** (1992) 91 – 122.
145. A. F. Buskhe, B. E. Berg, O. Gyllenhaal, T. Greibrokk, *HRC & CC, J. High Resolut. Chromatogr.* **11** (1988) 16.
146. M. L. Lee et al., *J. Microcol. Sep.* **1** (1989) 7.
147. J. C. Moore, *J. Polym. Sci. Part* **A 2** (1964) 835.
148. B. E. Richter: *1986 Pittsburgh Conference and Exposition on Analytical Chemistry and Applied Spectroscopy,* Atlantic City, NJ, March 10 – 14, 1986, paper 514.
149. B. E. Richter et al., *HRC & CC, J. High Resolut. Chromatogr.* **11** (1988) 29.
150. G. Schomburg et al., *HRC & CC, J. High Resolut. Chromatogr.* **12** (1989) 142.
151. B. E. Richter et al., *J. Chromatogr. Sci.* **27** (1989) 303.
152. P. Morin, M. Caude, H. Richard, R. Rosset, *Chromatographia* **21** (1986) 523.
153. J. W. Jordan, L. T. Taylor, *J. Chromatogr. Sci.* **24** (1980) 82.
154. R. C. Wieboldt, G. E. Adams, D. W. Later, *Anal. Chem.* **60** (1988) 2422.
155. S. B. French, M. Novotny, *Anal. Chem.* **58** (1986) 164.
156. J. D. Pinkston, D. J. Bowling in [43]
157. M. Smith (ed.): "Supercritical Fluid Chromatography," *RSC Chromatography Monographs,* 1989.
158. J. Zapp, Ph. D. Thesis, Saarbrücken 1990.
159. K. D. Pharma GmbH: DE P 42 065 39.9, 1992 (P. Lembke, H. Engelhardt, R. Krumbholz).
160. Y. Hirata, F. Nakata, *J. Chromatogr.* **295** (1984) 315.
161. H. H. Hill, M. A. Morrissey in G. M. White (ed.): *Modern Supercritical Fluid Chromatography,* Huethig, Heidelberg 1988.
162. P. Morin, M. Caude, H. Richard, R. Rossel, *Analysis* **15** (1987) 117.
163. J. M. F. Douse, *J. Chromatogr.* **445** (1988) 244.
164. B. W. Wright, R. D. Smith, *J. High Resolut. Chromatogr.* **9** (1986) 73.
165. J. B. Crowther, J. D. Henion, *Anal. Chem.* **57** (1985) 2711.
166. R. M. Smith, *J. Chromatogr.* **505** (1990) 147.
167. P. Lembke, in K. Anton, C. Berger (eds.): *Supercritical Fluid Chromatography with Packed Columns,* Marcel Dekker, New York 1998, pp. 429 – 443.
168. P. Cuatrecasas, M. Wilchek, C. Anfinsen, *Proc. Natl. Acad. Sci. U.S.A.* **61** (1968) 636.
169. W. H. Scouten: *Affinity Chromatography,* John Wiley & Sons, New York 1981.
170. W. Brümmer in R. Bock, W. Fresenius, H. Günzler, W. Huber, G. Tölg (eds.): *Analytiker-Taschenbuch,* **vol. 2,** Springer Verlag, Berlin 1981, pp. 63 – 96.
171. M. Wilchek, T. Miron, J. Kohn in W. B. Jacoby (ed.): *Methods of Enzymology,* **vol. 104,** Academic Press, New York 1984, pp. 3 – 55.
172. I. R. Birch, C. R. Hill, A. C. Kenney in P. N. Cheremisinoss, R. P. Ovellette (eds.): *Biotechnology Applications and Research,* Techcromic Publishing Company Inc., Lancaster 1985, p. 594.
173. P. Mohr, K. Pommering: "Affinity Chromatography, Practical and Theoretical Aspects," in J. Cazes (ed.): *Chromatography Science,* vol. 33: Marcel Dekker, New York 1986.
174. K. Ernst-Cabrera, M. Wilchek in R. Burgess (ed.): *Protein Purification: Micro to Macro, UCLA Symposia on Molecular and Cellular Biology, new series,* **vol. 68,** A.R. Liss Inc., New York 1987, pp. 163 – 175.

175. St. Ostrove in J. N. Abelson, M. I. Simon (eds.): *Methods of Enzymology,* **vol. 182,** Academic Press, New York 1990, pp. 357 – 371.
176. J. Carlsson, J. C. Janson, M. Sparrmann in J. C. Janson, L. Ryden (eds.): *Protein Purification,* VCH Publishers, New York 1989, pp. 275 – 329.
177. G. T. Hermanson, A. K. Mallia, P. K. Smith (eds.): *Immobilized Affinity Ligand Techniques,* Academic Press, New York 1992.
178. R. F. Taylor, *Anal. Chim. Acta* **172** (1985) 241.
179. K. Ernst-Cabrera, M. Wilchek, *Makromol. Chem., Makromol. Symp.* **19** (1988) 145 – 154.
180. C. R. Lowe, M. J. Harvey, D. B. Craven, P. D. G. Dean, *Biochem. J.* **133** (1973) 499 – 506.
181. P. Cuatrecasas, *Nature* **228** (1970) 1327.
182. E. Müller, "Coupling Reactions" in M. Kastner (eds.): *Protein Liquid Chromatography,* Elsevier, Amsterdam 2000.
183. J. Kohn, M. Wilchek, *Enzyme Microb. Technol.* **4** (1982) 161 – 163.
184. G. S. Bethell, J. S. Ayers, M. T. W. Hearn, W. S. Hancock, *J. Chromatogr.* **219** (1981) 361.
185. T. Miron, M. Wilchek, *Appl. Biochem. Biotechnol.* **11** (1985) 445 – 456.
186. K. Ernst-Cabrera, M. Wilchek, *J. Chromatogr.* **397** (1987) 187.
187. K. Nilsson, K. Mosbach in W. B. Jacoby (ed.): *Methods of Enzymology,* **vol. 104,** Academic Press, New York 1984, p. 56.
188. E. Stellwagen in J. N. Abelson, M. I. Simon (eds.): *Methods of Enzymology,* **vol. 182,** Academic Press, New York 1990, pp. 343 – 357.
189. St. J. Burton in A. Kenney, S. Fowell (eds.): *Methods in Molecular Biology,* **vol. 11,** The Humana Press Inc., Totowa, N.Y., 1992, pp. 91 – 103.
190. J. Porath, F. Maisano, M. Belew, *FEBS Lett.* **185** (1985) 306 – 310.
191. B. Nopper, F. Kohen, M. Wilchek, *Anal. Biochem.* **180** (1989) 66 – 71.
192. J. Porath, J. Carlson, I. Olsson et al., *Nature* **258** (1975) 598 – 599.
193. Z. Horvath, G. Nagydiosi, *J. Inorg. Nucl. Chem.* **37** (1975) 767 – 769.
194. V. A. Davankov, A. V. Semechkin, *J. Chrom.* **141** (1977) 313 – 353..
195. E. Sulkowski, *Trends Biotechn.* **3** (1985) 1 – 7.
196. P. J. Hogg, D. J. Winzor, *Anal. Biochem.* **163** (1987) 331 – 338.
197. J. J. Langone, *Adv. Immunol.* **32** (1982) 157 – 252.
198. A. Jungbauer et al., *J. Chromatogr.* **476** (1989) 257 – 268.
199. K. Ernst-Cabrera, M. Wilchek, *Med. Sci. Res.* **16** (1988) 305 – 310.
200. M. L. Yarmush et al., *Biotechnol. Adv.* **10** (1992) 413 – 446.
201. G. Köhler, C. Milstein, *Nature (London)* **256** (1975) 495 – 497.

Thin Layer Chromatography

Joseph C. Touchstone, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104-6080, United States

1. Introduction 1
2. Choice of the Sorbent Layer 1
2.1. Properties of Presently Available Sorbents 2
2.2. Binders 3
2.3. Suppliers 3
3. Sample Cleanup 4
4. Sample Application 7
5. The Mobile Phase 9
6. Development 11
6.1. Apparatus 11
6.2. Development Techniques 13
7. Visualization 14
7.1. Nondestructive Visualization Methods 15
7.2. Destructive Visualization Methods . . 15
7.3. Reactive Visualization Methods 16
7.4. Application of Visualization Reagents 16
8. Quantitation 17
9. References 19

1. Introduction

The following constitutes a general guide to the use of thin layer chromatography (TLC). Pertinent reference sources on the subject include [1–3]. The discussion here is focused on precoated plates, since these have largely supplanted homemade layers. For further information on the relationship between TLC and other types of chromatography → Basic Principles of Chromatography.

2. Choice of the Sorbent Layer

Precoated plates for TLC have been commercially available since 1961. The sorbents may be coated on glass, plastic, or aluminum supports. Sorbents with and without binder, and with and without UV indicators, are available in a variety of layer thicknesses, ranging from 100 µm in the case of plastic plates and high-performance layers to 2311 µm for preparative layers. The largest selection is presented by glass plates, coated to a thickness of 250 µm in the case of "analytical layers."

Important considerations with respect to all TLC plates are that the layers be uniform in thickness and surface, and that the plates yield reproducible results from package to package. These are difficult characteristics to duplicate with homemade plates, and are of prime concern in analytical work. The availability of high-quality commercial precoated plates is in fact one of the reasons why quantitative thin layer chromatography has been so successful in many applications.

Plates with a plastic support have an additional distinguishing feature relative to glass-backed plates: They can be cut with scissors. This is advantageous, for example, when small plates are desired but only 20×20 cm plates are on hand. With plastic-backed plates it is also easy to elute separated substances from the developed layer by cutting out an appropriate region and immersing the cut piece in a suitable extraction solvent. This is an especially important consideration when working with radioactive substances (→ Radionuclides in Analytical Chemistry). Measuring the radioactivity of such a substance is facilitated by simply immersing the portion of plate containing the radioactive zone in a scintillation vial containing counting fluid. Because the layer need not be scraped, potential contamination and losses caused by flying dust are minimized.

Considering only the cost of the sorbent and plates, precoated plates are obviously more expensive than homemade plates. However, the total expense of homemade plates also includes the cost of the coating apparatus and drying–storage racks, as well as the time involved. After these factors are taken into account it will be seen that precoated plates are highly desirable.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_301

2.1. Properties of Presently Available Sorbents (→ Adsorption, Chap. 3.)

Silica Gel. Silica gel is prepared by the precipitation of a silicate solution with acid, or by hydrolysis of silicon derivatives. The surface area and diameter of the silica gel particles depend on the method of precipitation. Variations in pH during precipitation can produce silica gels with surface areas ranging from 200 – 800 m^2/g. It has been shown [4] that silica gel provides three types of surface hydroxyl groups: Bound, reactive, and free. Relative reactivity and adsorption follow the order bound > free > reactive. Thus, control of the distribution of surface functions can have a significant effect on the chromatographic properties of a silica.

Special methods have been developed for preparing spherical particles with specific pore characteristics. Thermal treatment can also affect the pore size. Many of the silicas currently in use have a pore size of 6 nm.

Alumina. Alumina has seen limited use as a sorbent in TLC. Three types of alumina are available: Acidic, basic, and neutral. The amount of water present greatly affects the chromatographic behavior of alumina. For control purposes the plates may be heated to a specific temperature prior to use.

Kieselguhr. Kieselguhr is a purified diatomaceous earth usually characterized by wide pores. It is frequently used as a *preadsorbent layer* on other types of TLC plates. Commercial kieselguhr preparations include Celite, Filter Cel, and Hyflo Super Cel.

Cellulose. Cellulose has been used as a sorbent since the earliest days of chromatography, originally in the form of paper chromatography. Most methods of preparing celluloses for TLC are based on procedures for preparation of this medium for other uses. The types used for TLC are powdered (fibrous) or microcrystalline celluloses. The fiber length is shorter with TLC cellulose than in paper chromatography, which leads to less diffusion of the spots relative to paper chromatography.

Polyamides. Polyamides, previously used almost exclusively in column chromatography, have seen increasing application as sorbents for TLC. Certain separations are accomplished much more readily on polyamide than on conventional layers, since this medium lends itself to the analysis of water-soluble samples.

Ion Exchange. Ion exchange provides greater selectivity than most other forms of TLC, and recent developments have broadened its applicability. Ion exchange can be utilized with either organic or inorganic substances. Within these classifications, a further division can be made into anionic or cationic processes according to the nature of the ions subject to exchange. Several sorbents are now available for ion-exchange TLC, and the number will grow as development continues.

Reversed-Phase Sorbents. Reversed-phase (RP) sorbents can be classified into (1) those in which a liquid is impregnated in a solid support, and (2) chemically bonded sorbents, where the liquid phase is bonded chemically instead of physically. Chemical bonding provides layers that are more stable, preventing the stationary phase from being removed by solvent. Three types of reactions are used for bonding: Esterification, reaction of a chlorosilane with a selected organic substituent, and chlorination of the support followed by reaction with an organolithium compound to produce Si−C bonds. These bonding procedures can lead to various types of structures depending on the mode of attachment to the silica and the nature of substituents present at the other end of the alkyl side chain. The major bonding types have the structures Si−O−C, Si−O−Si−C, and Si−C−O−Si−N. The length of commercially available chains ranges from 2 to 18 carbon atoms.

Gels. Gels provide an entirely different type of medium for separation on thin layers. Sephadex was one of the first gels used as a stationary phase in the resulting *steric exclusion* TLC.

Cyclodextrin. More recently, cyclodextrin-modified layers have been introduced for use in the separation of enantiomers. The increasing

importance of isomer separation in pharmaceutical applications has led to a demand for these specially modified layers. Chiral layers can take several different forms. Most are C_{18} layers impregnated with a chiral-selecting protein derivative together with copper ions. Others consist of layers in which a sorbent silica has been chemically bonded to a cyclodextrin derivative.

2.2. Binders

Sorbents are prepared both with and without binding substances. A binder changes the properties of a sorbent such that the latter will adhere to its support and not readily flake off during handling, chromatography, visualization, or documentation. Nevertheless, binding agents are not always necessary, nor are they always desirable. Cellulose, for example, does not usually require a binder, since it has good inherent properties of adhesion and strength. Binders are obviously undesirable if they interfere with development or visualization. Calcium sulfate, a commonly used binder, is slightly soluble in water, and it is therefore not recommended in conjunction with aqueous solvent systems. Layers employing starch or polymer binders (so-called organic binders) should not be used when very high-temperature (>150 °C) visualization techniques are contemplated. They are also not suitable with universal visualization reagents such as sulfuric acid spray. Starch is appropriate only when it does not interfere, and when tough, strong layers are essential.

Calcium sulfate binder concentrations range from ca. 5 wt % in some silica gel preparations to 9 wt % in aluminum oxide G sorbent, 13 wt % in the widely used silica gel G, and up to 30 wt % in certain silica gel preparations for preparative (high capacity) TLC, where a "G" in a sorbent name denotes the presence of calcium sulfate (gypsum).

Commercial precoated plates are sometimes made with a small amount of organic polymeric binder, which makes the layer very hard and resistant to abrasion. This binder does not dissolve in relatively strong, polar solvents or in organic solvents, which increases the range of possible applications for separation but precludes certain detection procedures. Such binders, which are of a proprietary nature, absorb less moisture from the atmosphere than calcium sulfate, and for this reason the plates usually do not require activation prior to use.

2.3. Suppliers

A number of distributers and manufacturers supply precoated TLC plates, but not all carry a complete selection of sorbent/backing-plate combinations in the four major sizes (5×20, 10×20, 20×20, and 20×40 cm):

Analtech Inc., Newark, Del, United States	full line
J. T. Baker Chemical Co., Phillipsburg, N.J., United States	full line
Brinkmann Instrument Inc., Westbury, N.Y., United States	full line
Desaga GmbH, Heidelberg, Germany	relatively broad line
Eastman Kodak, Rochester, N.Y., United States	mylar backed layers
EM Science, Darmstadt, Germany	full line
Ingenur F. Heidenreith, Oslo, Norway	relatively broad line
Macherey-Nagel GmbH, Düren, Germany	sorbents and layers
Schleicher & Schuell GmbH, Kassel, Germany	specialty layers
Whatman Chemical Separations, Clifton, N.J., United States	full line

Some commercial 20×20 cm plates have prescored lines on the back at 5, 10, and 15 cm so they can easily be broken into smaller plates.

There is little literature available comparing plate-layer characteristics. To the benefit of the method, only two or three major manufacturers produce each sorbent type, so diversity with respect to a given sorbent is kept to a minimum. However, different manufacturers of the plates themselves employ different coating techniques, which in turn has an effect on the characteristics of sorbent that has been coated on the plates. Thus, even though two precoated-plate companies buy their sorbent from the same manufacturer, the resulting plates may not display identical separation characteristics.

The *hydrophilic modified* precoated plates have broadened the spectrum of selectivity available for TLC. These plates are referred to as "HPTLC-NH_2 precoated plates," in which γ-aminopropyl groups are bonded to the surface of a silica-gel skeleton, and "HPTLC-CN precoated plates," with γ- cyanopropyl moieties as the functional groups. In terms of polarity the two hydrophilic modifications can be classified as follows: Normal $SiO_2 > NH_2 > CN > RP$. Since the surface NH_2 and CN groups are of

medium polarity, no wetting problems are encountered with the corresponding plates even at the maximum level of modification. Any mobile phase can be used, without restriction.

The presence of the functional group means that an NH_2-modified support can act as a weakly basic ion exchanger, useful, for example, in separating charged substances. Thus, various nucleotide building-blocks can be distinguished on the basis of differing charge numbers, with substances retained more strongly as the charge numbers increase. Other retention mechanisms can also be involved in sample retention on an NH_2 plate. In the case of oligonucleotides, the observed selectivity arises through differences in hydrophobicity among the various nucleobases.

A CN-modified layer can be used for either normal-phase or reversed-phase chromatography depending on the polarity of the solvent system.

The *hardness* of a layer varies as a function of any binder present, and this in turn affects the speed of development and the resulting R_f for a given substance. Factors such as activation and developing- chamber saturation are also important influences in this regard.

Generalizations regarding cellulose layers are easier to formulate. As noted previously, the two most popular forms of cellulose are fibrous and microcrystalline. The major supplier of fibrous cellulose is Macherey, Nagel and Co., whose main product is MN 300 cellulose. The major supplier of microcrystalline cellulose is FMC Corporation, under the trade name Avicel. Microcrystalline cellulose layers usually develop more rapidly and produce zones with lower R_f values that are more compact in size relative to fibrous cellulose layers.

3. Sample Cleanup

Proper preparation of the sample is the most important preliminary step in conducting any analysis (→ Sample Preparation for Trace Analysis), and this is especially true in chromatography. Of particular concern is the amount of extraneous material present in the sample solution. The greater the amount of "junk," the more sample preparation that will be required.

The Sample. Before undertaking the preparation of any sample, an aliquot of material representative of that sample must be obtained. Otherwise the analysis itself may be meaningless. The best way to assay a solid sample is to collect random aliquots from the whole and then mix them together before removing the extraction aliquot. If a package of tablets or capsules is to be analyzed, random portions from the whole should similarly be mixed together prior to selecting an aliquot for analysis.

The Solvent for the Sample. An appropriate solvent must of course be selected for dissolving the sample. When a sample must be derivatized, solubility characteristics may change drastically, or there may not be enough sample remaining for performing general solubility tests. The solubility of a crystalline material can be readily determined with a microscope. A few crystals of the sample are placed on a slide in the field of the microscope and carefully observed as a drop of solvent is added. If the crystals do not dissolve, the solvent can be evaporated and a new one selected for further testing. Testing with a number of solvents greatly expedites selection of the one that is optimal for experimental analysis.

Extraction of the compounds of interest from a complex sample and subsequent dissolution in the proper solvents requires different strategies depending on the compound class under consideration. For example, polar compounds do not dissolve well in nonpolar solvents, and nonpolar compounds will not dissolve in polar solvents. Ionic compounds dissolve best in a polar solvent to which the counter ion has been added. Satisfactory sample application on a TLC layer requires that the solvent be highly volatile and as nonpolar as possible. Nonvolatile solvents spread through the plate-layer matrix during the period of contact, taking sample with them. Polar solvents likewise spread solute through the matrix as a consequence of both high solvent power and low volatility.

Sample Extraction. Several alternatives are available for extracting the sample from its solvent. These include:

1) *Direct application* of the solution, with extraction taking place on the TLC layer

2) Application of a solution of the sample on a *preadsorbent layer,* followed by direct development
With both of these techniques proper selection of the mobile phase can ensure that extraneous materials are left at the chromatographic origin. In some cases the layer can be predeveloped in such a way that analyte is deposited on a line which can be regarded as an origin for subsequent development with the correct mobile phase for separation.
3) *Liquid Extraction of the Sample.* Some samples may be too impure for simple extraction by one of the first two methods, but control of the pH often improves the situation. For example, organic acids are converted into water-soluble salts at high pH. Lowering the pH restores the unionized acids, which are then extractable into organic solvents. At high pH, basic compounds are extracted into organic solvents, returning to the water phase as their salts at low pH.
4) *Solid Phase Extraction.* Columns and microcolumns (also referred to as cartridges) prepacked with sorbents can be used for *solid phase extraction* (*SPE*) of samples. Such columns are available commercially, but one can easily pack a simple tube with the required sorbent, such as those in Table 1. Some extraction columns are made to fit onto a syringe, whereby the column becomes an extension of the syringe barrel, and the plunger is used to force the sample and eluting solvents through the packing. Table 2 lists a number of United States suppliers of packed SPE columns.
SPE columns require minimal amounts of solvent. Eluants of increasing strength can be used to elute analytes with different polarities in separate fractions. Thus, silica-gel columns release low-polarity substances upon elution with a nonpolar solvent, whereas the analyte is retained as a consequence of its polarity. The analyte is then eluted with a more polar solvent. The least polar solvent capable of eluting the analyte should always be used to ensure that very polar impurities remain behind. The eluant is subsequently evaporated; alternatively, if an analyte concentration is sufficiently high, an aliquot can be taken for application to the TLC plate. This methodology presents unlimited possibilities for analyte isolation.
5) *Liquid – Liquid Extraction Using Microcolumns.* Special microcolumns buffered at pH 4 – 9 and packed with diatomaceous earth have been developed for liquid – liquid partitioning. These can be utilized for sample volumes ranging from one to several hundred milliliters. The principle of liquid – liquid partitioning is also the basis of a device known as an Extube. This is a disposable plastic syringe packed with an inert fibrous matrix presenting a large surface area. A sample and an extraction solvent are poured into the tube, where liquid – liquid partitioning occurs. Analyte is extracted into the organic phase, which passes through the matrix, while impurities and other polar components, including water, are retained. The eluate is then concentrated by evaporation prior to TLC.

Table 1. Sorbents for use in microcolumns

Nonpolar	Polar	Ion exchange
C_{18}	cyanopropyl	carboxylic acid
C_8	diol	propylsulfonic acid
C_2	silica gel	benzenesulfonic acid
Phenyl	aminopropyl	diethylaminopropyl
Cyclohexyl	propyl	quaternary or secondary amine

General Guidelines for the Use of SPE Columns [5]. The SPE cartridge must always be *conditioned* with the particular solvent that will be used. This is necessary both to activate the sorbent and to clean the column. Conditioning procedures vary depending on whether normal or reversed-phase sorbents are employed. Ion-exchange conditioning is a function of the polarity of the sample solvent.

For normal-phase packings the first conditioning wash is often conducted in the solvent in which the analyte is dissolved. With most SPEs 2 mL of solvent is sufficient. An SPE can also be conditioned with 3 – 5 mL of a nonpolar solvent such as hexane. This can be adjusted by adding methylene chloride, but the final mix should always be miscible with the sample solvent.

With reversed-phase sorbents, methanol is used for the first wash. A typical cartridge requires 3 – 5 mL of methanol, which is followed

by the solvent used to prepare the sample. A film of water-miscible solvent remains on the sorbent after the aqueous solvent has been added, improving the interaction between aqueous sample and hydrophobic sorbent.

Table 2. Suppliers of SPE columns in the United States

Company	Column designation	Accessories available
Alltech Inc., Deerfield, Ill.	"Extract-Clean"	manifold
Analtech Inc., Newark, Del.	"Spice" cartridges	manifold
J. T. Baker Chem. Co., Phillipsburg, N.J.	"Baker-Bond"	manifold
Applied Separations, Bethlehem, Pa.	"Spe-ed" cartridges	various accessories
Bodman Chemicals, Aston, Pa.	"Pure-Sep" cartridge	various accessories
Isolab Inc., Akron, Ohio	"Quick-Sep"	manifold
J & W Scientific, Folsom, Ca.	"Accuford" tubes	manifold
Versa-Prep Lab Systems Inc., Nanuet, N.Y.	"Tef-Elutor"	various accessories
Varian Inc., Sunnyvale, Ca.	"Bond-Elute"	various accessories
Waters Associates, Milford, Mass.	"Sep-Pak"	various accessories
Worldwide Monitoring Corp., Hatfield, Pa.	"Cleanscreen" columns	full line of sorbents

For ion-exchange sorbents, deionized water is usually used first, but if the sample is in a nonpolar solvent, the SPE should be conditioned with 2 – 5 mL of this solvent instead. Aqueous samples require use of 2 mL of methanol followed by 2 mL of water. The sorbent is then adjusted to the organic solvent content and salt concentration of the sample solution. Sorbent must not be allowed to dry between conditioning and addition of the sample.

The sample solution is subsequently added to the top of the cartridge and drawn through the sorbent bed by application of a vacuum. Special vacuum manifolds are available for this purpose, some of which handle 12 or more tubes, thus increasing sample throughput.

The next step is *washing* of the column (and sample). For reversed-phase work, washing with a polar solvent generally ensures that the analyte will be retained. Normal-phase applications require a wash with some nonpolar solvent that leaves the analyte on the sorbent. An ion-exchange column is washed with water, again so that the analyte will be retained.

Washing is followed by *elution* of the analyte. In reversed-phase mode a nonpolar solvent is used to remove the analyte; in normal phase, it is a polar solvent that removes the analyte. Adjustment of the ionic strength or modification of the eluting solvent is used to free the solute from an ion-exchange column. In each case, two volumes of 100 – 500 µL should suffice for complete elution.

The eluant is then *evaporated,* and the residue (extract) is *reconstituted* in a solvent suitable for introduction into the required chromatographic system. Typically, 5 – 200 µL of solvent is used for this purpose, 2 – 10 µL of which is taken for analysis. If the concentration is suitable, an aliquot of the eluant can be analyzed directly without the need for evaporation and reconstitution.

4. Sample Application

Numerous devices are available for applying samples to TLC plates. *Microliter syringes*, such as those manufactured by Hamilton, have frequently been used for applying 5 to 10 µL amounts of samples. Application of small volumes of sample with a syringe results in relatively narrow or small application areas, since little solvent is present to cause diffusion. Sucessful use of these devices takes practice, however, and it necessitates washing of the delivery device between sample applications. *Microcapillaries* for this purpose are available from many suppliers as an alternative, and these offer the best volume-delivery precision. Capillaries are also disposable, ruling out contamination between samples. These capillaries with holders, as illustrated in Figure 1, are available in a wide range of fixed volumes, and they have proven to be ideal for TLC applications.

Automated or *semiautomated sample applicators* are available as well. These devices apply consistent and reproducible sample spots, but it is a misconception that they are necessary for quantitative work. With proper technique, manual methods of sample application can provide results entirely comparable to those from automatic devices. Commercial automated units employ syringes or rows of microcapillaries to ap-

ply a spot or band of sample, and some actually "spray" the sample onto the layer. Many are designed for preparative separations, applying large amounts of sample as streaks across the sorbent layer.

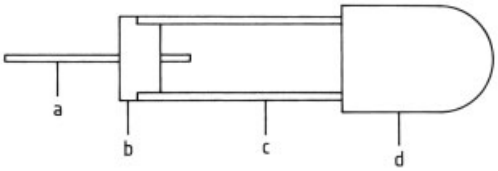

Figure 1. Microcapillary spotting device for TLC
a) Capillary tube; b) Septum; c) Support tube; d) Rubber bulb

The Sample. In the case of silica-gel layers the sample should be dissolved in the least polar solvent in which it is soluble. With a solution of the proper concentration, 1 – 5 µL should suffice to permit detection of the analyte of interest after chromatography. Concentrating may prove to be the best procedure with a dilute solution. However, dilute solutions can also be applied in the form of several successive aliquots. If the sample is too dilute, spreading of the application area can be prevented by accelerating evaporation of solvent with a stream of air blown over the area in question. A hair dryer is often used for this purpose. Correct application ensures that the separated zones will be both symmetrical and compact.

Sample Amounts. The amounts of the sample and the analyte are critical factors in achieving good results, since applied load affects the shapes and positions of developed zones. Application of a sample that is too concentrated can result in overloading. Overloaded zones produce comet-like vertical streaks after development, which can cause overlap of previously separated components.

TLC is a sensitive technique, and it is important to "think small" with respect to sample size. The highest resolution will always be obtained with the smallest sample commensurate with the ability to visualize the analyte. This is also one of the prerequisites for reproducibility, especially if densitometry is to be used for quantitation (see Chap. 8). Zones that are too concentrated result in integrated curve areas that behave in a nonlinear way on regression plots. Depending on the nature of the analyte, sample sizes as small as picogram amounts are not uncommon, especially in high-performance thin layer chromatography (HPTLC). This is especially true in the analysis of fluorescent substances. Thus, for ultimate sensitivity it is sometimes desirable that a fluorescent derivative of the analyte be prepared.

With regard to preadsorbent plates, the applied sample volume has an effect on preadsorbent efficiency. Up to certain levels, increasing the sample volume appears to increase plate efficiency, since a larger volume prevents compacted zones, and uniformity of the starting line after predevelopment is maximized.

Using Micropipettes for Manual Application of Analytical Samples. Care should be taken to ensure that separate spots are placed on a straight line horizontally across the plate. With hard-layer plates a guideline can be drawn on the plate surface itself with a soft pencil to act as a spotting guide. A white sheet of paper with a dark line drawn on it can also serve as a guide, since such a line will be visible through a layer positioned over it. Use of a guideline makes it easier to keep the sample origins evenly spaced across the plate and parallel to the edge. In the case of plates with a preadsorbent zone no guide or guideline is required, because the sample can be applied anywhere within the preadsorbent zone. This zone is inert, so it will not retain the analyte, permitting samples to move with the solvent front until they reach the adsorbent zone. A preadsorbent provides a convenient means of applying a dilute sample without the need for evaporation. No matter how the sample is applied, the result after predevelopment will be a fine line at the juncture between the adsorbent and the preadsorbent material.

Samples are applied to a 20 × 20 cm TLC plate 2 cm from the bottom. The two outermost samples should be 1.5 to 2 cm from the edges. Normally, individual "lanes" are kept about 1 cm apart. The origin should be located 2 cm from the bottom to ensure that sample will not be eluted inadvertently into the mobile-phase reservoir. Standards are interspersed at every third or fourth position. A single plate can provide space for 20 or more samples.

Usually 2 – 10 µL of sample solution is applied to the TLC plate. A fixed-volume mi-

cropipette is completely filled with the solution, and the end of the capillary is then touched to the surface of the TLC adsorbent, causing solution to be "wicked" into the adsorbent, and producing a round "spot." The capillary is most conveniently held by thrusting it through a hole in a septum, which is in turn inserted into a larger piece of glass tubing. Some manufacturers provide a rubber bulb (similar to the top of a medicine dropper) that fits over this larger tube. The bulb has a hole in it to equilibrate pressure while the capillary is draining (see Fig. 1).

Since it is known that small samples lead to improved separation, a method known as *overspotting* is sometimes used for sample application. Thus, a 10 μL sample might be applied to the plate as five spots of 2 μL each, all delivered at the same place. Successive spots must not be applied until solvent from the previous application has completely evaporated. Failure to observe this precaution can result in distorted zones. Sample application in this way can be used to form a single spot, or a series of spots can be joined to form a band. Application of a sample as a band gives greater resolution [6]. Treatment between applications with a hair dryer or a stream of filtered air facilitates solvent evaporation. The zone of application should be kept as narrow as possible compatible with the degree of concentration required. Spots or zones that are too concentrated result in poor separation, since the mobile phase tends to travel around the concentrated center, which leads to distorted zones or tailing.

Applying Preparative Samples. The preferred method for introducing a sample onto a preparative layer is to apply it as a narrow streak across the entire width of the plate. Both manual and automatic means are available for applying such a streak. The streak should be made as straight (horizontally) and as narrow as possible. Most commercial streakers are capable of producing a sample zone only 1 – 3 mm wide. A layer 1000 μm (1 mm) thick can readily handle 100 mg of sample.

Some successful preparative work has been reported based on the same manual techniques developed for applying analytical spots. In these cases a small (i.e., 2 – 4 μL) pipette is used to apply successive round spots in direct contact with each other across a plate, but the method is time-consuming and tedious, and it requires great care. Any spotting irregularities are greatly magnified in the separation process, and this often makes separation of closely spaced components very difficult. However, if the separated components are widely spaced, then satisfactory results are achievable in this way.

The use of preparative plates with a preadsorbent zone reduces the amount of care necessary to produce a narrow band. Sample in this case is applied to the relatively inert preadsorbent layer as a crude "streak" across the plate. Even so, it is best to exercise some care and apply the sample evenly across the plate, because a high local sample concentration can cause an overloading effect in the corresponding region of the plate.

The Importance of Sample Drying. Thorough drying of applied samples is a necessary step that is frequently ignored. Errors can result if a sample carrier solvent is not allowed to evaporate completely from the TLC plate. Many chromatographers erroneously assume that the sample spot is sufficiently dry when solvent is no longer visible on the plate. The preferred practice is to allow a drying time of five to ten minutes, perhaps assisted by the air stream from a hair dryer.

5. The Mobile Phase

Proper selection of the mobile phase is essential to achieving optimal TLC results. Indeed, the great strength of TLC lies in the wide range of applicable mobile-phase solvents, extending from the extremely nonpolar to the extremely polar. A highly polar mobile phase is generally used when the separation involves very polar substances, whereas a less polar mobile phase is best for less polar components. A very polar substance is strongly attracted to silica gel, so a polar solvent is required to displace it. Chromatographic conditions cited in references may be useful as starting points, but some knowledge of solvent – solute characteristics is required for optimal separation.

Selecting a Mobile Phase. It is quite rare for nothing whatsoever to be known about the chemical nature and solubility characteristics of a sample. If one knows something about what

constitutes a "good" solvent for the sample, this may imply what will be a "poor" solvent. When a "good" solvent is used as the carrier, adsorbed materials will move in the solvent front, while a "poor" solvent will probably not move anything. What is required for effective TLC is a "mediocre" solvent, one that will cause partial movement and some measure of selective discrimination. Usually the first step is to find a solvent or mobile phase that will move the analyte with an R_f of 0.5. In difficult separations, a mobile phase with a changing composition may be necessary.

Interaction forces involved in TLC include: Intramolecular forces, inductive forces, hydrogen bonding, charge transfer, and covalent bonding [7]. Understanding selectivity thus requires that something be known about the nature of these various types of interaction.

Selecting a Mobile Phase Based on Solvent – Solute Characteristics. The mobile phase in TLC is the source of mobility, but more importantly it produces selectivity. Selectivity – the degree of interaction between components of the mobile phase and analytes in the solute – determines the extent of any resolution that will be achieved. Solvent strength is the factor that determines mobility. Knowing the characteristics of a solute makes it easier to decide which solvents to try during the search for a mobile phase that will produce the desired separation.

Important general characteristics of mobile-phase solvents important in TLC include:

1) *Water content;* this can vary widely, and it is a common source of reproducibility problems, since traces of water may lead to large changes in solvent strength and selectivity
2) *Autoxidation* sometimes occurs during solvent storage, especially over prolonged periods
3) Solvents kept in metal containers may be subject to *decomposition* and unknown chemical reactions, which may change the solvent characteristics
4) Solvent *impurities* may differ from batch to batch, and such impurities may interact with the solute, the solvent, or the sorbent
5) *Chloroform* often contains ethanol as a preservative, and since ethanol has high solvent strength, small amounts may have a dramatic effect

The optimal composition for a mobile phase depends on the nature and number of functional groups in the material to be analyzed. A molecule containing many oxygen functions will display high interactivity, whereas a hydrocarbon with no functional groups will show little interaction. On the other hand, a hydrocarbon solvent will dissolve other hydrocarbons, but not molecules characterized by numerous polar functional groups. Functional groups affect the solvent – solute interaction in the order $RH < ROCH_3 < RN(CH_3) < ROR < RNH_2 < ROH < RCONH_2 < RCO_2H$. This sequence also applies to components of the mobile phase [8].

The level of interaction between the solute and the mobile phase as opposed to the sorbent determines the distribution of the solute between these two phases. The mobile phase is generally less reactive than the sorbent. If the analyte interacts more strongly with the mobile phase it will prefer to remain in the mobile phase, and the separation will be poor. If the solute has a more powerful interaction with the sorbent the separation potential will be greater. Chromatography thus consists of an interplay between solute –sorbent and solute – mobile-phase interactions (→ Basic Principles of Chromatography).

The *solvent strength parameter* (see below) provides a quantitative measure of the relationship between solvent strength and the type of sorbent used, although solvent behavior differs greatly in normal- vs. reversed-phase chromatography. The interaction of a solvent with the sorbent is thus a function of the type of sorbent used. Methanol is more polar than chloroform. As polarity increases, so does elution strength with respect to a normal-phase sorbent such as silica gel. Depending on the analyte, methanol will lead to more rapid analyte migration up the layer relative to a less polar solvent. The opposite is the case with a reversed-phase system. Here, the more polar the mobile phase the greater the tendency toward retention by the sorbent. Thus, water will not readily move analytes in a reversed-phase system, but will do so easily under normal-phase conditions. Even so, if the analyte does not dissolve, nothing will happen. Hexane, a nonpolar solvent, is at the opposite end of the scale from methanol.

Solvent	E°, Al_2O_3
Fluoroalkanes	− 0.25
n-Pentane	0.00
Hexane	0.00
Isooctane	0.01
Petroleum ether, Skellysolve B, etc.	0.01
n-Decane	0.04
Cyclohexane	0.04
Cyclopentane	0.05
Diisobutene	0.06
1-Pentene	0.08
Carbon disulfide	0.15
Carbon tetrachloride	0.18
Amyl chloride	0.26
Butyl chloride	0.26
Xylene	0.26
Diisopropyl ether	0.28
2-Chloropropane	0.29
Toluene	0.29
1-Chloropropane	0.30
Chlorobenzene	0.30
Benzene	0.32
Ethyl bromide	0.37
Ethyl ether	0.38
Ethyl sulfide	0.38
Chloroform	0.40
Methylene chloride	0.42
Methyl isobutyl ketone	0.43
Tetrahydrofuran	0.45
Ethylene dichloride	0.49
Methyl ethyl ketone	0.51
1-Nitropropane	0.53
Acetone	0.56
Dioxane	0.56
Ethyl acetate	0.58
Methyl acetate	0.60
Amyl alcohol	0.61
Dimethyl sulfoxide	0.62
Aniline	0.62
Diethyl amine	0.63
Nitromethane	0.64
Acetonitrile	0.65
Pyridine	0.71
Butyl cellosolve	0.74
Isopropyl alcohol, *n*-propyl alcohol	0.82
Ethanol	0.88
Methanol	0.95
Ethylene glycol	1.11
Acetic acid	large
Water	larger
Salts and buffers	very large

Perhaps the most widely recognized characteristic solvent property is the *solvent strength parameter* E° [7]. This was defined by SNYDER as the adsorption energy for a solvent per unit of standard sorbent. The higher the solvent strength, the greater will be the R_f value of an analyte in normal-phase TLC. The opposite is true in the case of reversed-phase sorbents. Solvent strength parameters are listed below for a variety of common solvents [7]. The values cited are based on heats of adsorption with alumina, but a similar sequence and similar values apply to silica gel. The same principles also hold true for reversed phases, but the corresponding solvent-strength values are reversed.

Solvent *selectivity* refers to the ability of a solvent to establish interaction differences for two different analytes. Resolution is dependent on three factors:

1) The nature of the sorbent
2) The migration rate of the two solutes, which in turn depends on the solvent strength
3) Separation factors, which are determined by interactions between the solvent and the mobile phase

If two solutes cannot be separated by a particular mobile phase of predetermined strength, it may still be possible to achieve separation with a different mobile phase of the same strength. The above list of solvent strengths can be used to illustrate application of the solvent selectivity concept. Selectivity arises from the fact that a given solvent will interact differently with two solutes of different functional makeup. The solvent-strength list includes several examples of solvents with identical strength values. Ethyl acetate and diethylamine for example, have approximately the same solvent strength, but entirely different chemical structures, and thus different reactivities with respect to other molecules. In particular, two solutes that cannot be separated with a given mobile-phase system containing ethyl acetate may separate when diethylamine is substituted for the ethyl acetate despite a minimal change in overall mobility.

The mechanisms of retention on chemically bonded reversed phases are not clearly understood, but observed R_f values for a series of solutes usually turn out to be reversed compared to the corresponding sequence on silica gel provided water constitutes a large fraction of the mobile phase. It is also possible to achieve separations on RP plates using an entirely organic mixture as the mobile phase.

Most RP chromatography is conducted with two-solvent mixtures composed of water together with an alcohol, acetonitrile, acetone, or dioxane. Methanol – water (8 : 2) is a convenient starting point for work with RP layers. The solvent strength is then varied by changing the wa-

ter–organic modifier ratio, and selectivity is improved by using a different modifier. More polar samples require relatively more water. Small changes in mobile-phase strength have less effect on R_f values in RPTLC than in normal-phase TLC, making mobile-phase selection easier in the former case.

Mobile phases for gel TLC must be capable of dissolving the sample and at the same time causing the gel network to swell. Organic solvents such as tetrahydrofuran (THF) are common for gel permeation TLC on organic polymer layers (e.g., styrene – divinylbenzene), while aqueous buffers are employed for gel filtration chromatography on layers such as dextran.

Useful general rules related to selectivity and solvent strength include the following:

1) Increasing the solvent strength will decrease separation and increase R_f.
2) Decreasing the solvent strength will increase separation, but decrease R_f.
3) Selectivity is greatest when the concentration of the stronger component of the mobile phase is greater than 5 vol % and less than 50 vol %. In other words, more favorable solute – solvent interactions result when the solvent of greater strength is not dominant.
4) Introduction of a solvent with greater hydrogen-bonding potential may increase resolution.
5) With respect to (3), it is generally best to examine the effect of small changes in the amount of the stronger modifying solvent.
6) If the observed R_f value is too high, a primary solvent of lower solvent strength should be tried rather than drastically changing the proportions of a mixture, and vice versa.

6. Development

6.1. Apparatus

Most TLC development is accomplished by allowing the mobile phase to ascend a plate standing in a suitable-sized chamber or tank containing the appropriate mobile phase. Glass chambers are by far the most common. Simple chambers can be made from wide-mouthed screw- cap, battery, or museum jars. Chambers specifically made for TLC are available in a variety of sizes, both rectangular and circular, accommodating all sizes and formats of TLC plates. Aluminum racks are available for supporting and developing several conventional plates at one time in a single tank. One company has introduced a special tank with a raised glass hump running lengthwise on the inside bottom to act as a tank divider. The latter requires only 20 mL of solution per side, and makes it possible for two large plates to be developed simultaneously. The divided bottom also allows the tank to be saturated with two mobile-phase systems simultaneously, placed on opposite sides of the divider, so that a plate can be introduced on either side for development. Small chambers have been designed specifically for use with 10×10 cm or HPTLC plates. These small chambers occupy much less bench space, and require considerably less solvent.

A *sandwich chamber* for use with 10×10 cm plates offers several advantages, including small solvent volume (6 mL), rapid chamber saturation, and rapid development. A glass-backed TLC plate forms one wall of the chamber, with the adsorbent facing inward toward a second glass plate. The sides are clamped together with metal clips and the open end of the chamber is inserted in a solvent trough containing the developing solvent. Other commercially available sandwich chambers accommodate larger plates (20×20 and 10×20 cm). Such a chamber can be purchased commercially from a number of suppliers for about the same price as a normal glass TLC tank.

Ascending or Linear Development. The separation process in TLC occurs in a three-phase system consisting of stationary, mobile, and gas phases, all of which interact with each other until equilibrium is reached. If a plate is developed in a chamber whose atmosphere is not saturated with the mobile phase, flow up the plate may occur in an unsymmetrical way because some mobile phase evaporates from the plate into the chamber atmosphere in an attempt to establish equilibrium. The entire chromatographic process then depends on the volatility of the mobile phase, the temperature, and the rate of development. It is for this reason that the development chamber should be made as small

as conveniently possible based on the plate size and the number of plates to be developed at one time.

Reproducible results are achieved only with a constant environment. Laboratory benches exposed to the sun during part of the day should be avoided because of temperature variation. Most development chambers are of such a size that they may conveniently be placed in an environmentally controlled space, room, or bath as a way of maintaining consistency. It is advisable to weigh down the lid of the chamber so that it cannot be dislodged by vapor pressure or heat produced by the mobile phase, which would disturb the equilibrium within the system. Some newer chambers include latches that actually "lock down" the lids.

Chamber Saturation [9]. Saturation of the developing chamber atmosphere is facilitated by lining the inner walls with blotting paper or thick chromatography paper (e.g., Whatman 3 MM), or by using saturation pads specially designed for TLC, although some separations can be accomplished in unlined tanks. The individual components of the mobile phase should be mixed thoroughly before they are introduced into the developing chamber. Reported solvent proportions are assumed to be by volume unless stated otherwise. It is advantageous to pour the mixed mobile phase over the lining paper as it is added to the developing chamber, since this hastens the process of saturating the paper and the chamber. Enough time should be allowed for complete equilibration of the chamber.

Plate development may be described in terms of either time or distance. With timed development a plate is allowed to develop for a specific period of time, after which it is carefully lifted out of the chamber. The location of the solvent front is then marked with a pencil or a sharp object such as a syringe needle. If mobile phase is instead expected to travel a fixed distance (e.g., 10 cm), this distance should be marked on the plate at the time of sample application. It then becomes easy to observe when the mobile phase has reached the desired point so the plate can be removed promptly from the chamber. A fixed development distance is used to standardize R_f values; because R_f values by definition fall between 0.0 and 1.0, 10 cm is a convenient standard distance.

6.2. Development Techniques

Standard Development. The standard method of development has already been described in the preceding paragraphs. If the TLC characteristics of the analyte are unknown, a simple single-component mobile phase is preferable for preliminary experiments aimed at establishing separation behavior. If no single-component mobile phase provides the separation desired, however, multicomponent systems should be examined.

Multiple Development. If a single development of a plate suggests that a partial separation has been achieved, the plate should be allowed to dry, after which it can be reinserted into the developing chamber for redevelopment. This procedure may be repeated as many times as necessary to achieve a satisfactory separation. Repeated development increases the effective length of the layer, thereby increasing the likelihood of separation, because the various molecules will have an opportunity to interact over a greater distance, perhaps culminating in complete resolution.

With a mixture containing compounds that differ considerably in their polarities, development with a single mobile phase may not provide the desired separation. However, separation may result if such a plate is developed successively with various mobile phases that have different selectivities or strengths. Multiple mobile phases can be used in a number of ways. For example the least polar phase might be introduced first, followed by a more polar phase, or vice versa. More than two developments can also be tried, and the development distance on the plate might be changed with each development. For example, a very polar mobile phase, such as methanol, could be used first with a given sample mixture in order first to separate the polar substances. After the methanol had developed part way up the layer the plate would be removed from the chromatography tank and dried. Then a second mobile phase, a nonpolar solvent such as cyclohexane, would be allowed to develop the entire length of the plate in an attempt to resolve in the upper portion those nonpolar substances that were not resolved in the previous

(polar-phase) development. Depending on the circumstances, the same procedure might also be carried out with the reverse sequence.

Continuous Development. Continuous development involves a continuous flow of mobile phase along the length of the plate, with the mobile phase being permitted to evaporate as it reaches the end. The chromatography tank in this case is not as tall as a normal tank, which means that a normal-sized TLC plate extends beyond the top. Mobile phase then evaporates from the exposed end of the plate, causing a continuous flow to be maintained.

Continuous development offers the advantage of greater separation without the need for changing the mobile phase. Solvent selection is simplified because one can exploit the high separation power of a low-polarity solvent without suffering the usual consequence of reduced R_f. Sensitivity may be increased as well, because zone diffusion is reduced as a result of the shorter distance traveled by the analyte. The technique is essentially equivalent to development on a very long TLC plate without the requirement for special equipment.

Two-Dimensional Development. This developing technique involves the application of normal, multiple, or stepwise development in two successive dimensions. A very versatile method, it has been widely used in paper chromatography as well as in certain TLC separations. A single sample is applied at one corner of a plate and developed in the usual way. The plate is then removed from the developing chamber and dried, after which it is rotated by 90° and placed in a chamber containing a different mobile phase so that it will develop again, this time in a direction perpendicular to that of the first development. The line of partially resolved sample components from the first development becomes the origin for the second development (see Fig. 2). It is important that initial sample application not be too close to the edge of the plate, since otherwise the resulting line of developed spots might fall below the level of solvent in the second developing chamber. A distance of 2.5 – 3 cm from the edge is usually sufficient. The versatility of the technique arises not only from the second development with a different mobile phase, but also from the opportunity provided for treating or modifying the layer and the sample prior to the second development. Standards can be applied to the second origin before the second development to permit direct comparisons and calculation of R_f values after development and visualization. The technique is also applicable to plates containing two different sorbents, such as silica gel on one half and RP_{18} on the other.

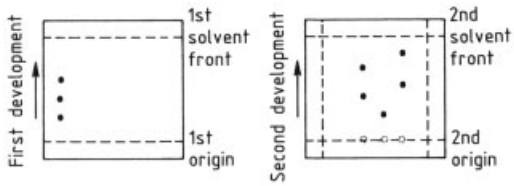

Figure 2. An example of two-dimensional development

Circular and Anticircular Development. Circular and anticircular development have been accomplished in a number of ways. Because of the need for solvent-delivery control in such systems, chromatography is best carried out in equipment specially designed for the purpose. It is also important that the chamber be kept level. Camag, Analtech, and Anspec all offer chambers for circular development. Samples are applied at the center, and mobile phase is wicked to the surface. Development causes the analyte mixture to separate into a series of concentric rings. Figure 3 illustrates these devices.

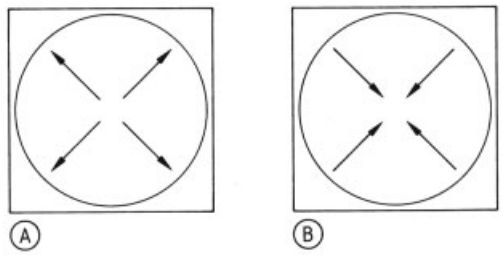

Figure 3. Circular and anticircular TLC development
A) Circular development; a wick is used to apply mobile phase to the center, causing radial flow outward; B) Anticircular development, in which solvent is applied with a wick to the outer ring, causing solvent migration inward toward the center

Reversed-Phase Partition Development. Reversed-phase development employs a rela-

tively nonpolar stationary phase, which means that the layer is hydrophobic rather than hydrophilic as in normal partition TLC. The corresponding mobile phase is relatively polar, and may be water or an alcohol. Most of the common TLC sorbents have been used in conjunction with partition development, including cellulose, silica gel, and kieselguhr. This technique is generally applicable to the separation of polar compounds, which often remain at the origin during conventional development. With reversed-phase development, nonpolar compounds are the ones that usually remain near the origin. The need for polar eluents results in development times longer than those normally encountered with ordinary silica gel and less polar eluents.

Horizontal Development. Linear development can also be conducted from two directions simultaneously, with mobile phase entering from two opposing sides of a single plate. Such development requires the use of a horizontal developing chamber. The main advantage of the procedure, which was originally designed for use with 10×10 HPTLC layers, is that the number of samples per plate can be doubled. In closed systems and with small samples it is possible to separate in this way as many as 72 samples at the same time.

7. Visualization

Once a sample has been transferred to a TLC plate, all the components of that sample will be located somewhere on the layer. If the analyte is colored, localization is simple. Most organic compounds are not colored, however, so special techniques are required to change the materials in such a way that some optical contrast is established in the visible or the ultraviolet region. Reagents useful in color formation have been listed in [1].

A satisfactory visualization technique must take account of the following considerations:

1) Significant contrast is essential between the separated zone and the background.
2) A high degree of sensitivity is important to ensure detection of minute amounts of analyte.
3) Many compounds are not stable once separated on a sorbent; fluorescent compounds in particular are often unstable. This may not be a serious concern however, if only qualitative information is required. Nonetheless, the stability characteristics of the solutes should be ascertained.
4) A universal reagent capable of detecting all the solutes present is most desirable.
5) Satisfactory detection methods always result in well-defined zones of separation. Spraying or dipping procedures can be problematic as a result of dispersion of the separated solutes.
6) An ideal detection method would permit quantitative analysis based on proportional absorbance intensities (see Chap. 8).

Methods for inducing detectability may be either *nondestructive* or *destructive*, where nondestructive methods are ones that do not destroy the analyte.

7.1. Nondestructive Visualization Methods

Ultraviolet absorbance can be utilized for detection purposes in several ways. For example, ultraviolet light often makes it possible to locate analytes separated on a fluorescent TLC plate. Thus, a UV-absorbing component may quench the background fluorescence, appearing as a dark spot against a luminescent field. Zinc silicate is a common phosphor in plates marked "F" for fluorescent, which fluoresce under the influence of light with a wavelength of 254 nm. Plastic fluorescent plates sometimes contain the fluorescent dye Rhodamine 6 G.

However, this fluorescence quenching is not a very sensitive detection method. As a rule it is recommended that visualization of the thin layer chromatogram take place in a dark room under the influence of lamps with emission maxima at 254 and 366 nm.

Inherent fluorescence can also be used for detecting separated zones, although the commonly available 254- and 366-nm lamps are suitable only for detecting compounds that fluoresce within the output range of the lamps themselves. Inherent fluorescence is the most sensitive aid to detection, occasionally revealing as little as 1 pg of an analyte.

Nondestructive detection is sometimes possible simply by spraying a plate with water. Free polar adsorbents (silica gel and alumina) are able to adsorb large amounts of water, whereas regions on the plate occupied by spots of organic material adsorb much less moisture, and appear as fatty zones against an almost translucent white background.

Detection with iodine vapor is one of the most widely applicable visualization techniques. If iodine crystals are placed in the bottom of a covered glass development chamber, within a few hours the atmosphere in the chamber will be saturated with the vapors. Analyte components on a thin layer chromatogram placed in such an iodine chamber tend to adsorb iodine, subsequently appearing as dark brown spots against a lighter tan background [10]. This method is characterized by either destructive or nondestructive detection depending on the analyte. After the developed layer in an iodine tank has been examined for evidence of reactive zones, the plate can be removed and areas of interest can be marked. Once the iodine has evaporated, analyte can sometimes be recovered for use in another procedure – provided it has not reacted with the iodine. Some analytes do undergo reaction with iodine, however, in which case the detection process is nonreversible. For example, phenolic compounds are known to react irreversibly under these conditions.

The iodine reaction can be surprisingly sensitive in some cases. Thus, morphine and codeine have been detected in this way at the 0.1-µg level. Iodine reagent can also be applied by spraying with a 1 % solution of iodine in an inert, volatile solvent such as hexane.

7.2. Destructive Visualization Methods

Charring is a destructive method that can be regarded as essentially universal. When an organic material on a TLC plate is heated in the presence of acid, an easily distinguishable black or brownish-black spot is formed. In many cases, quantitation of such a spot is possible by densitometry (see Chap. 8).

Acid solutions frequently used for TLC charring include:

sulfuric acid : acetic anhydride (1 : 3)
sulfuric acid : sodium dichromate (the well known "chromic acid cleaning agent")
sulfuric acid : nitric acid (1 : 1)
sulfuric acid : copper acetate or copper sulfate
phosphoric acid : copper sulfate or copper acetate

The plate under investigation is first treated with the acid solution and then heated. Charring is accomplished at 120 – 280 °C over a period of 1 – 40 min. Initial experiments should be conducted at 120 °C for 1 min, since this may produce dehydration products that fluoresce with different colors. For example, steroids generally produce fluorescent products when sprayed with 1 % sulfuric acid in ethanol followed by heating to as little as 100 °C. As noted previously, this type of reagent has the inherent advantage of offering universal detection; most analytes produce some type of reaction, and if no reaction is initially apparent the sample on the layer can be heated to increasingly higher temperatures until charring occurs. The obvious disadvantage is destruction of the analyte (to say nothing of surrounding structures and clothing if care is not exercised!).

7.3. Reactive Visualization Methods

This category includes the preparation of derivatives, either before or after development of the chromatogram. It is sometimes feasible to improve the detectability of an analyte by subjecting it to a chemical reaction prior to chromatography. The separation that follows such a reaction is obviously not of the original materials, but rather of the reaction products. Confirmatory identification should be based on R_f values for both the unchanged and the reacted product. Reactive derivatization often has the additional advantage of improving the resolution of compounds that are otherwise difficult to separate.

7.4. Application of Visualization Reagents

The most common approach to application of a visualization reagent is probably spraying, using a spray gun powered either by compressed

air from a laboratory line or compressed gas in a canister. The usual device incorporates a propellant container mounted on a screw cap that fits the top of a glass reagent jar. Reagent is drawn out of the jar by the venturi effect, and is dispersed as fine droplets in the emerging gas "cone." Spraying should be conducted in a fume hood to eliminate or minimize the exposure of personnel to reactive and toxic materials. Many distributers offer a wide variety of specially designed plate supports, spraying racks, and disposable or washable spray booths.

A TLC plate should be sprayed as evenly as possible, combining side-to-side and up-and-down motions of the spray equipment. Careful work is required to ensure uniform distribution of the sprayed reactant. Quantitation becomes particularly uncertain if spot intensity is dependent upon both the concentration of the spot component and the amount of reagent locally applied.

Application of reactive solutions is also possible by immersion of the TLC plate in some type of "dipping tank." The chief disadvantage of dipping techniques, especially with highly reactive agents, is that the back of the plate also becomes wet, and must be wiped in order to remove excess reagent. Care must also be taken when attempting to dip "soft" plates, since a soft coating is easily damaged by the forces associated with moving the plate in a surrounding liquid. One alternative is allowing the reactive reagent to ascend the plate by capillary action, just as in development.

If the need arises to spray a large number of plates it may be advantageous to consider setting up a *vapor charring* system, in which a plate is heated in a chamber through which SO_3 gas is passed, which in turn induces charring. The gaseous approach has the advantage that it distributes charring reagent evenly over the entire plate. The procedure utilizes a square Pyroceram dish approximately $10 \times 10 \times 25$ cm in size and equipped with a curved glass cover. These are readily available as accessories for microwave cooking. The TLC plate is placed in the bottom of the vessel, adsorbent side up. A small amount of fuming sulfuric acid is then swabbed around the glass cover about 2.5 – 5 cm from the edge. The prepared cover is placed over the base, and the entire assembly is transferred to a hot plate or an oven. Appropriate conditions of time and temperature must be precisely established for each required analysis. It must also be emphasized that extreme precautions are necessary withrespect to the fuming acid, particularly when opening the vessel after charring. The advantages, of course, are very even charring and highly reproducible results.

Both Camag and Analtech also distribute sealed chambers commercially, utilizing ammonium hydrogencarbonate for visualization purposes. Vapors formed as the salt is heated tend to char or otherwise cause visual changes in a plate which is heated simultaneously. Many compounds form fluorescent derivatives under these conditions.

8. Quantitation

Modern approaches to *quantitative* TLC require the use of densitometry. Alternative visual quantitation techniques will not be discussed here, nor will elution techniques and the associated spectrophotometric methodology. For a discussion in this area one should consult the literature, for instance the "Practice of Thin Layer Chromatography" by TOUCHSTONE [1] now in its third edition. The focus here will be densitometric methodology, which is clearly becoming the method of choice. For a detailed discussion of theoretical aspects of densitometry see POLLAK [11].

In situ methods based on densitometers and spectrodensitometers are now available as well, and these can be very precise, with relative standard deviations of less than 5 %. Instruments on the market (1994) for in situ scanning of TLC layers include the following:

Ambias Inc., San Diego, Ca., USA	scanner for combined densitometry and isotope detection
Analtech Inc., Newark, Del., USA	one of the only video scanners designed for TLC
Camag, Muttenz, Switzerland	linear scanner used in many laboratories; equipped with a monochromator as a standard feature
Desaga GmbH, Heidelberg, FRG	linear scanner with monochromator
Helena Laboratories, Beaumont, Tex., USA	densitometer originally designed for electrophoresis scanning
Shimadzu Scientific Instruments, Columbia, Md. USA	linear scanner with monochromator

Video Scanners. Video scanners (such as the Uniscan device from Analtech) permit very rapid data acquisition. An entire TLC plate can beanalyzed at once with this device. Three detector modes are available: transmittance, reflectance, and fluorescence. The scanner incorporates its own computer with a video digitizer, light sources (visible and UV), and an imaging detector. Bandpass filters in front of the lens are used to select specific emission wavelengths, which can extend the usefulness of the instrument. In the case of fluorescent compounds separated by HPTLC, quantities in the low picogram range are readily assayed. For rapid quantitation in many regions, multichannel detector instruments offer a great advantage over conventional detection systems. Inexpensive instruments now permit digital storage of large amounts of data, opening the way to mathematical methods of data analysis. The most satisfactory instruments take advantage of a charge-coupled discharge video or standard camera, which can capture a complete image of a TLC layer in fractions of a second. The Analtech Uniscan comes with an enclosed view box featuring visible and short- or long-wave UV illumination. The image is projected on a monitor, and it is processed with the aid of a cursor for setting scan limits. Signals are processed in such a way as to provide both a chromatogram and tabulated quantitative data in the form of a printout. The quantitative information is calibrated on the basis of data from serial applications of standards that have been separated on the same chromatogram.

Linear Scanners. Linear scanners usually rely on a fixed light source under which the TLC layer must be moved. Changes in density are recorded by a computer programmed to convert the signal into a chromatogram, which is than evaluated on the basis of parameters so adjusted as to provide optimal scanning. Both Shimadzu and Camag market linear scanners of this type, which are also capable of determining absorption spectra of separated materials suitably positioned unter the light beam from a monochromator. The same instruments can be used to scan fluorescent compounds as well. Judicious use of filters can enhance the specificity of an analysis. For example, a 400-nm cut-off filter provides useful screening in the case of fluorescence analyses.

In *double-beam* operation the primary signals are amplified and fed directly into logarithmic converters. At the output, the reference signal is substracted from the sample signal, and the result is fed into an integrator. From there the signal passes to a conventional chart recorder that records the ratio of the two beam signals.

Important properties to consider in evaluating an instrument include the nature of the light source, the lines of highest emission energy, and the response curve of the detecting element, which may be either a photomultiplier or a cadmium sulfide diode. For example, the following spectral characteristics were observed with one instrument based on a xenon – mercury lamp (Hanovia 901.B-11) and photomultiplier detectors: Mercury bands from the xenon – mercury lamp were apparent at 312 – 313 nm and 365 – 366 nm, with the band at 365 – 366 nm by far the most intense. The lamp had an intensity curve rising gradually from 200 nm to a hump at ca. 500 nm, after which it leveled off. The photomultipliers showed a maximum response to light at 380 nm. This wavelength was preceded by a shoulder at the end of a steady rise. Above 380 nm the response declined gradually, becoming almost zero at 570 nm. The peaks of highest emission of the lamp coincided with the bands of greatest activation for the fluorescence of quinine sulfate separated on a silica-gel layer. This is significant, because it is usually possible to obtain greater sensitivity by using light with a wavelength close to the lines of maximum emission of the lamp.

It is important to scan at the wavelength of maximum absorption of the solute in the sorbent, where the absorption maximum for a compound adsorbed on silica gel particles is typically not the same as that for the same compound in an alcohol solution. Precise and sensitive determination requires that the absorption maximum be determined in situ on the plate. This is possible with an instrument equipped with sources of variable wavelength or monochromators. Narrow-bandpass filters are available for instruments based instead on filters.

Single-beam recordings show the expected background variations, together with an increased noise level in both transmission and reflectance measurements. (The background is much smoother with a double-beam system.) Over short distances near the center of the plate

the background variations and noise experienced in the course of reflectance measurements in single-beam operation may appear acceptable, but in reflectance mode the peak response for a given spot is substantially lower than in transmission mode, so the S/N value (i.e., precision) should be expected to decrease. Peak signals obtained in transmission are about five times as large as the corresponding signals in reflectance, whereas the noise level is only twice as great.

Errors in sample application are among the most significant sources of inaccuracy. The technique of spotting can have a marked effect on delivery volumes between 1 and 5 µL. A constant *starting zone size* should be maintained so that a given amount of substance applied in the form of different volumes will yield spots with identical areas, whereas different amounts spotted at constant volume will exhibit a linear relationship in a plot of peak area (from recordings of densitometer scans) versus mass.

Errors introduced during movement of the solute through the sorbent layer include lateral diffusion during development, and problems can also arise from structural variations for different plates or variations between tanks. Lateral diffusion is a function of the relative rates of diffusion of solute and mobile phase (→ Basic Principles of Chromatography). Errors due to lateral diffusion when substantially different amounts of solute are applied next to each other can be prevented by scoring the plate into strips ca. 1 cm wide prior to spotting. In this way two dissimilar samples can safely be developed side by side. Variation in structure from one layer to another leads to incorrect extrapolation of data obtained when a known quantity of standard is compared to the quantity of an unknown. The ratio between the quantity in a final spot and the quantity spotted varies for different layers, so it is recommended that a standard be applied to each test plate. Factors such as the spot size before and after development and the R_f value also become important in this context.

There are several possible sources of error in quantifying the results obtained from a densitometer or radiochromatogram scan. Important considerations include sampling methods, separation variables, variations in detector response, and calibration and measurement techniques.

Many of the errors in quantitation originate in the way a sample is applied to the layer, since this affects the shape of the spot after chromatogram development. The chromatographic process itself determines whether there will be tailing or overlap of the spots. High R_f values result in broad peaks that are not compatible with reproducible results from scanning in situ.

Another source of error is sample decomposition during chromatography. Evidence of decomposition can be obtained by developing chromatograms with standards of known serial dilutions. If the peak area does not increase linearly as a function of concentration, sample decomposition should be suspected. Errors due to tailing and overlap are quite common. These can be eliminated by changing the mobile phase to one that produces better separation.

The peak-area calibration procedure is usually less subject to influence from changes in instrumental parameters than calibration based on peak height. In addition, it offers improved results with peaks whose shapes are not Gaussian. However, results obtained with the peak-area quantitative method are adversely affected by neighboring peaks. Peak areas should always be used when precision is important.

The electronic integrators used in GC analysis are generally satisfactory for measuring peak areas in TLC scans as well. An electronic digital integrator measures peak areas automatically and converts them into numerical data, which are then printed. Sophisticated versions of such devices also correct for baseline drift.

Thin layer chromatography peak areas derived from scans can also be measured conveniently with computer systems developed for GC applications, often without any change. Most devices of this type are programmed to print out a complete report, including compound names, retention times, peak areas, area correction factors, and weight-percent composition with respect to the various sample components.

9. References

1. J. C. Touchstone: *Practice of Thin Layer Chromatography,* 3rd ed., Wiley, New York 1992.
2. J. Sherma, B. Fried: *Handbook of Thin-Layer Chromatography,* Marcel Dekker, New York 1991.

3. F. Geiss: *Fundamentals of Thin Layer Chromatography,* Hüthig-Verlag, New York 1987.
4. T. E. Beesley in J. C. Touchstone, D. Rogers (eds.): *Thin Layer Chromatography, Quantitative Environmental and Clinical Applications,* J. Wiley, New York 1980, p. 10.
5. M. Zief, R. Riser, *Am. Lab. (Fairfield, Conn.)* **6** (1990) 70.
6. D. E. Jaenchen, H. Issaq, *J. Liq. Chromatogr.* **11** (1988) 1941.
7. L. R. Snyder: *Principles of Adsorption Chromatography,* Marcel Dekker, New York 1968.
8. L. R. Snyder, J. J. Kirkland: *Introduction to Modern Liquid Chromatography,* Wiley, New York 1979.
9. R. A. de Zeeuw, *J. Chromatogr.* **33** (1968) 222.
10. G. C. Barrett in J. C. Giddings, R. A. Keller (eds): *Advances in Chromatography,* **vol. 11,** Marcel Dekker, New York 1974, p. 151.
11. V. Pollak in J. C. Giddings et al. (eds.): *Advances in Chromatography,* Marcel Dekker, New York 1979, pp. 1 – 50.

Electrophoresis

PIER GIORGIO RIGHETTI, Department of Agricultural and Industrial Biotechnologies, University of Verona, I-37134 Verona, Italy

CECILIA GELFI, ITBA, CNR, L.I.T.A., Segrate 20090 (Milano), Italy

1. Introduction 1
2. Basic Principles 2
3. Electrophoretic Matrices 3
3.1. Cellulose Acetate 3
3.2. Agarose Gels 3
3.3. Polyacrylamide Gels 5
4. Discontinuous Electrophoresis 7
5. Isoelectric Focusing 7
6. Sodium Dodecyl Sulfate Electrophoresis 11
7. Porosity Gradient Gels 11
8. Two-Dimensional Maps (Proteome Analysis) 12
9. Isotachophoresis 14
10. Immunoelectrophoresis 16
10.1. Rocket Immunoelectrophoresis . . . 17
10.2. Crossed Immunoelectrophoresis . . . 17
10.3. Tandem Crossed Immunoelectrophoresis 17
10.4. Intermediate Gel Crossed Immunoelectrophoresis 17
10.5. Fused-Rocket Crossed Immunoelectrophoresis 18
11. Staining Techniques and Blotting . . 18
12. Immobilized pH Gradients 18
13. Capillary Zone Electrophoresis . . . 20
14. Preparative Electrophoresis 21
14.1. Preparative Isoelectric Focusing in Granulated Gel Layers 21
14.2. Continuous-Flow, Recycling Isoelectric Focusing 22
14.3. The Rotofor 23
14.4. Recycling Free-Flow Focusing (RF3) 24
14.5. Multicompartment Electrolyzers with Isoelectric Membranes 25
15. References 26

1. Introduction

The theory of electrophoresis has been adequately covered in the excellent textbooks of GIDDINGS [1] and ANDREWS [2] as well as in specific manuals [3,4]. For discussion on electrophoresis in free liquid media, e.g., curtain, free-flow, endless belt, field-flow-fractionation, particle, and cell electrophoresis the reader is referred to a comprehensive review by VAN OSS [5] and to a book largely devoted to continuous-flow electrophoresis [6]. Here the focus is mostly on electrophoresis in a capillary support, i.e., in gel-stabilized media, and discussion is limited to protein applications.

Electrophoresis is based on the differential migration of electrically charged particles in an electric field. As such, the method is applicable only to ionic or ionogenic materials, i.e., substances convertible to ionic species (a classical example: neutral sugars, which form negatively charged complexes with borate). In fact, with the advent of capillary zone electrophoresis (CZE) it has been found that a host of neutral substances can be induced to migrate in an electric field by inclusion in charged micelles, e.g., of anionic (sodium dodecyl sulfate, SDS) or cationic (cetyltrimethylammonium bromide, CTAB) surfactants [7]. Even compounds that are not ionic, ionogenic, or complexable can often be analyzed by CZE as they are transported past the detector by the strong electrosmotic flow on the capillary walls [8].

Electrophoretic techniques can be divided into four main types: zone electrophoresis (ZE), moving-boundary electrophoresis (MBE), isotachophoresis (ITP), and isoelectric focusing (IEF). This classification is based on two criteria: (1) initial component distribution (uniform vs. sharply discontinuous); and (2) boundary permeability (to all ions vs. only to H^+ and OH^-). Alternatively, electrophoretic techniques may be enumerated in chronological order, as follows: moving-boundary electrophoresis, zone electrophoresis, disc electrophoresis, isoelectric focusing, sodium dode-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_319

cyl sulfate – polyacrylamide gel electrophoresis (SDS – PAGE), two-dimensional (2-D) maps, isotachophoresis (ITP), staining techniques, immobilized pH gradients (IPG), and capillary zone electrophoresis.

Zone electrophoresis became practicable when hydrophilic gels (acting as an anticonvective support) were discovered. GRABAR and WILLIAMS in 1953 first proposed the use of an agar matrix (later abandoned in favor of a highly purified agar fraction, agarose) [9]. They also combined, for the first time, electrophoresis on a hydrophilic support with biospecific detection (immunoelectrophoresis). Barely two years after that, SMITHIES [10] applied another gel, potato starch. The starch blocks were highly concentrated matrices (12 – 14 % solids) and introduced a new parameter in electrophoretic separation: molecular sieving. Human sera, which in cellulose acetate or paper electrophoresis were resolved into barely five bands, now produced a spectrum of 15 zones.

2. Basic Principles

In a free buffer solution, in the absence of molecular sieving, the velocity $\boldsymbol{v}$ of a particle is proportional to the field strength $\boldsymbol{E}$, multiplied by its electrophoretic mobility μ:

$$\boldsymbol{v}=\mu\boldsymbol{E}$$

The velocity and $\boldsymbol{E}$ are both vectors, while μ is a scalar, being positive for cations and negative for anions. A distinction is made between absolute mobility (the μ value of species at infinite dilution and complete ionization) and effective mobility (the μ value of species under the actual experimental conditions, e.g., as a result of sieving in gel matrices, or in the presence of a strong electrosmotic flow). In deriving an expression for μ which takes into account parameters such as charge, size, and solvent effects, one can consider the different forces acting on a species. At infinite dilution, the electric field is exerting a force on the charge Q of the particle and this force is opposed by the frictional force, given by the friction factor f_c multiplied by the velocity $\boldsymbol{v}$. At steady state:

$$Q\boldsymbol{E}=f_c\boldsymbol{v}$$

For a spherical particle obeying Stokes' law, f_c is given by:

$$f_c=6\pi\eta r$$

where η is the solvent viscosity and r the particle radius. By substituting and rearranging one can derive the expression for the absolute mobility:

$$\mu=\boldsymbol{v}/\boldsymbol{E}=Q/(6\pi\eta r)$$

According to the Debye – Hückel theory, the total charge Q of a particle is correlated to the dielectric constant ε and to the zeta potential by the expression:

$$Q=\varepsilon r\zeta$$

whereby the expression for μ follows:

$$\mu=\varepsilon\zeta/(6\pi\eta)$$

The mobility is directly proportional to the dielectric constant and zeta potential and inversely related to the solvent viscosity. Thus, it follows that μ is strongly reduced in organic solvents, is zero at the isoelectric point of an amphoteric species (where the net charge is zero), and is greatly reduced in any form of gel media, which can be viewed as highly viscous solutions.

3. Electrophoretic Matrices

Only three matrices are in use today: cellulose acetate, agarose, and polyacrylamide.

3.1. Cellulose Acetate

Cellulose acetate is a cellulose derivative in which each hexose ring of the polysaccharide chain has two hydroxyl groups esterified to acetate (in general, in the C-3 and C-6 positions). This medium was developed for electrophoresis largely through the work of KOHN [11], who also designed the original tank, produced by Shandon Southern Instruments as Kohn's Electrophoresis Apparatus Model U77 and imitated throughout the world.

Cellulose acetate is still very popular in clinical chemistry, as it offers a ready-made support

that can be equilibrated with buffer in a few seconds and produces good separations of proteins from biological fluids in ca. 20 min. Completely automatic systems (for sample loading, staining, destaining, and densitometry) have been built around it, so that this system has become the first example of combined electrophoresis and robotics, an integration which had its testing ground and underwent development in clinical chemistry and is now spreading to basic research laboratories.

Cellulose acetate allows migration of even large serum proteins, such as pre-β-lipoprotein ($M_r > 5 \times 10^6$) and is amenable to immunofixation in situ [12, 13] without any need for prior transfer by blotting (as required, e.g., after SDS-PAGE). In addition, the immunoprecipitate, obtained directly on the cellulose acetate, can be enhanced by use of a second antibody, conjugated with, e.g., alkaline phosphatase [14]. Moreover, owing to their large porosity and their availability as thin films, cellulose acetate matrices can be stained with gold micelles, increasing the detection sensitivity to the subnanogram range [15], whereas agarose and polyacrylamide are incompatible with this gilding process [16]. However, the use of cellulose acetate is limited to clinical electrophoresis and is not much in vogue in basic research for high-resolution runs, even though focusing and two-dimensional techniques on cellulose acetate membranes have been described [17].

3.2. Agarose Gels

Agarose is a purified linear galactan hydrocolloid, isolated from agar or recovered directly from agar-bearing marine algae, such as the Rhodophyta. Genera from which agarose is extracted include *Gelidium, Gracilaria, Acanthopeltis, Ceramium, Pterocladia*, and *Campylaephora*. Agar is extracted from red algae by boiling in water and is then isolated by filtering off the particulates, gelling and freeze-thawing the colloid to remove water-soluble impurities, and precipitation with ethanol [18]. The resultant product is a mixture of polysaccharides, which are alternating copolymers of 1,4-linked 3,6-anhydro-α-L-galactose and 1,3-linked β-D-galactose. The repeating disaccharide is referred to as agarobiose [19]. Of course, this is an idealized structure, since even purified agarose contains appreciable amounts of the following substituents: sulfate, pyruvate, and methoxyl groups. Sulfate is generally esterified to the hydroxyl at C-4 of β-D-galactopyranose, and C-2 and C-6 of 3,6-anhydro-α-L-galactose. Pyruvate is linked to both C-4 and C-6 of some residues of β-D-galactopyranose. The resulting compound is referred to as 4,6-*O*-carboxyethylidene [20]. In addition, the hydroxyl group on C-6 of some residues of β-D-galactopyranose can also be methylated.

Agarose was first used to form gels for electrophoresis by HJERTÉN [21] and also for immunodiffusion [22] and chromatography [23]. Agarose can be separated from agaropectin by several procedures, including preferential precipitation of agaropectin by a quaternary ammonium salt, followed by fractionation of agarose with polyethylene glycol. By using a combination of polyethylene glycol precipitation and anion-exchange chromatography, an agarose with no detectable pyruvate and 0.02 % sulfate has been prepared [24]. A further treatment consists of alkaline desulfation in the presence of $NaBH_4$ followed by reduction with $LiAlH_4$ in dioxane [25]. When an agarose solution is cooled, during gelation, single strands of agarose (average molecular mass 120 kDa or ca. 392 agarobiose units) [26] dimerize to form a double helix. The double helix has a pitch of 1.9 nm, and each strand has threefold helical symmetry [27]. However, formation of the double helix is not sufficient to produce a gel; it is assumed that these helices aggregate laterally during gelation to form suprafibers (Fig. 1). In fact, the width of an agarose double helix is no more than 1.4 nm, whereas in gelled agarose, fibers one order of magnitude wider than this have been observed by electron microscopy and light scattering. The greater the number of aggregated double helices (i.e., the thicker the diameter of the "pillar"), the larger is the "pore" defined by the supporting beam [28]. Thus, it is reasonable to assume that it is the presence of these suprafibers that causes agarose gels to have strengths and pore sizes greater than those of gels that do not form suprafibers, such as starch and 2.5 – 5 % cross-linked polyacrylamide gels. In electrophoresis, agarose gels are used for fractionation of several types of particles, including serum lipoproteins [29], nucleic acids

[30], virus and related particles [31], and subcellular organelles, such as microsomes. However, a suspension of 0.14 % agarose particles is used for the latter, rather than a coherent gel mass [32]. More recently, by exploiting ultradilute gels (barely 0.03 % agarose, a special high-strength gel, Seakem LE, from Marine Colloids, Rockwille, ME, United States) SERWER et al. [33] have even been able to fractionate *E. coli* cells. Previously, this had been thought to be impossible, and in fact RIGHETTI et al. [34] reported a maximum pore diameter, for even the most dilute agaroses (0.16 %) of 500 – 600 nm. According to SERWER et al. [33] the effective pore radius P_E of a 0.05 % agarose gel is 1.2 µm; hence, the average pore of a 0.03 % gel should be several micrometers wide.

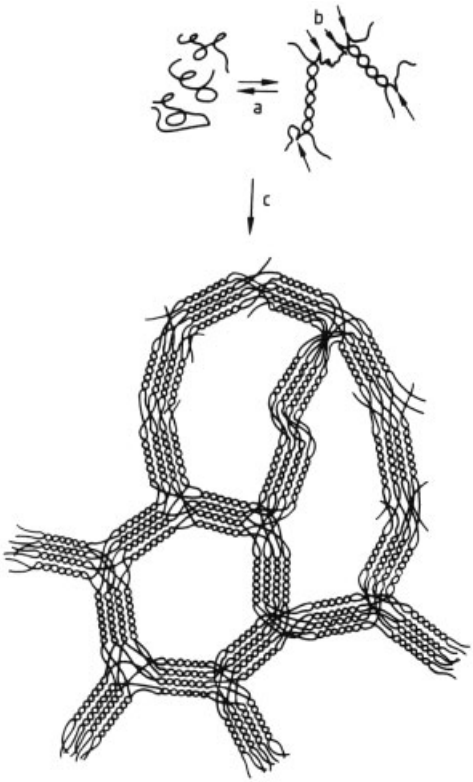

Figure 1. Formation of an agarose gel
Strands of agarose join to form double helices (a). At kinks (b) the strands exchange partners. Higher-order aggregation results in suprafibers (c). In hydroxyethylated agarose, it appears that the pillar looses its structure and more of the double helices are dispersed into the surrounding space: this results in a decrease in average pore size, i.e., higher sieving for the same % matrix as compared with underivatized agarose [31].

FMC BioProducts (formerly Marine Colloids, a division of FMC Corporation) of Rockland, ME, United States, has produced a number of highly sophisticated agaroses to meet the diverse requirements of electrophoresis users; the reader is referred to an extensive review on this topic [35]. Mixed-bed gels, containing 0.8 % agarose and variable amounts (2.5 – 9 %T) of acrylamide (%T indicates the total amount of monomers) have also been produced [36, 37]. Such gels were quite popular in the 1970s for nucleic acid fractionation, as they allowed fine resolution of lower M_r DNAs and RNAs, but they are seldom used in modern electrokinetic methods. However, these gels (under the trade name Ultrogels) in a spherical form are still popular for chromatographic purposes. Five types of bead exist, with a fractionation range of 1 – 1200 kDa [38].

3.3. Polyacrylamide Gels

The most versatile of all matrices are the polyacrylamide gels. Their popularity stems from several fundamental properties: (1) optical clarity, including ultraviolet (280 nm) transparency; (2) electrical neutrality, due to the absence of charged groups; and (3) availability in a wide range of pore sizes. Their chemical formula, as commonly polymerized from acrylamide and *N,N′*-methylenebisacrylamide (Bis), is shown in Figure 2, together with that of the two most widely employed catalysts, persulfate (ammonium or potassium) and *N,N,N′,N′*-tetramethylethylenediamine (TEMED) [39]. Four groups simultaneously and independently developed polyacrylamide gel electrophoresis: RAYMOND and WEINTRAUB [40], DAVIS [41], ORNSTEIN [42], and HJERTÉN [43]. In addition, HJERTÉN [44] introduced polyacrylamide beads as a matrix for chromatography. Normal polyacrylamide gels (i.e., cross-linked with standard amounts of 3 – 5 % Bis) have porosities that decrease linearly with the monomer content (%T). Their porosities are substantially smaller than those of corresponding agarose gels. However, highly cross-linked gels (> 30 %C) can have porosities comparable to or even greater than those of agarose matrices (%C is the mass in grams of cross-linker per 100 g total monomers) [45, 46]. However, high-C polyacrylamide ma-

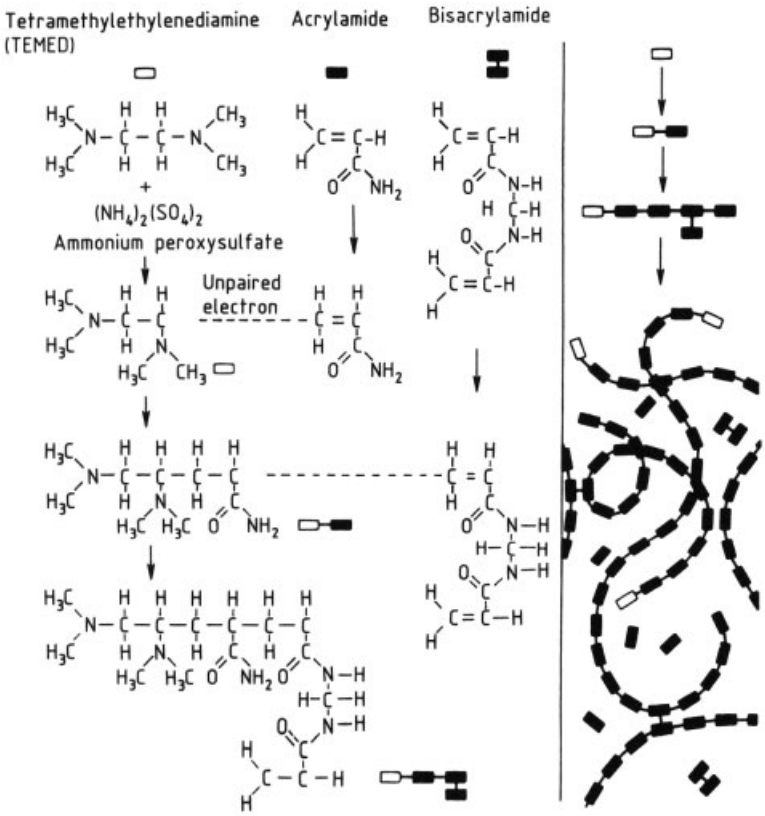

Figure 2. The polymerization reaction of acrylamide
The structure of acrylamide, *N,N′*-methylenebisacrylamide (Bis), and of a representative segment of cross-linked polyacrylamide are shown. Initiators shown are peroxysulfate and *N,N,N′, N′*-tetramethylethylendiamine (TEMED).

trices are brittle and opaque, and tend to collapse and exude water above 30 %C [34]. Thus, at present their use is limited.

Hydrophilic gels are considered to be a network of flexible polymer chains with interstices into which macromolecules are forced to migrate by an applied potential difference, according to a partition governed by steric factors. Large molecules can only penetrate into regions where the meshes in the net are large, while small molecules find their way into tightly knit regions of the network. Perhaps the best way to envision a hydrophilic matrix is to consider it as being composed of two interlaced fluid compartments, with different frictional coefficients [47]. In general, agaroses are more porous than polyacrylamides and these two matrices are used to complement each other.

The versatility of polyacrylamide gels is also shown by the large number of cross-linkers, besides Bis, that can be used to cast gels with particular properties for different fractionation purposes. *N,N′*-(1,2-dihydroxyethylene)bisacrylamide (DHEBA) can be used for casting reversible gels (i.e., gels that can be liquefied), since the 1,2-diol structure of DHEBA renders them susceptible to cleavage by oxidation with periodic acid [48]. The same principle should also apply to *N,N′*-diallyltartardiamide (DATD) gels [49]. Alternatively, ethylenediacrylate (EDA) gels [50] may be used, since this cross-linker contains ester bonds that undergo base-catalyzed hydrolytic cleavage. The poly(ethylene glycol) diacrylate cross-link belongs to the same series of ester derivatives [34]. As a further addition to the series, *N,N′*-bisacrylylcystamine (BAC) gels, which contain a disulfide bridge cleavable by thiols, have been described [51]. Practically any cross-linker can be used, but DATD and

N,N′,N″-triallylcitrictriamide (TACT) should be avoided, since, being allyl derivatives, they are inhibitors of gel polymerization when mixed with compounds containing acrylic double bonds. Their use at high %C is simply disastrous [52, 53].

Other monomers, in addition to acrylamide, have been described. One of them is Trisacryl, *N*-acryloyl-2-amino-2-hydroxymethyl-1,3-propanediol. The Trisacryl monomer creates a microenvironment that favors the approach of hydrophilic solutes (proteins) to the gel polymer surface, since the polyethylene backbone is buried underneath a layer of hydroxymethyl groups. In addition, owing to the much larger M_r of the monomer, as compared with acrylamide, gels made with the same nominal matrix content (%T) should be more porous, since they would have on average thicker fibers than the corresponding polyacrylamide matrix. This type of gel is extensively used by IBF Reactifs (VilleneuveLa-Garenne, France) for preparing gel filtration media or ion exchangers and also for electrophoresis [54, 55]. However, this monomer degrades with zero-order kinetics and hence seems to be of limited use in electrophoresis [56].

In 1984, Artoni et al. [57] described two novel monomers that impart peculiar properties to polyacrylamide-type gels: acryloylmorpholine (ACM) and bisacrylylpiperazine (BAP). These gels exhibited some interesting features: owing to the presence of the morpholino ring on the nitrogen atom involved in the amido bond, the latter was rendered stable to alkaline hydrolysis, which bedevils conventional polyacrylamide matrices. In addition, such matrices are fully compatible with a host of hydro-organic solvents, thus allowing electrophoresis in mixed phases. The use of the BAP cross-linker has been reported to have beneficial effects when silver staining polyacrylamide matrices.

More recently, novel monomers combining high hydrophilicity with extreme hydrolytic stability have been described. One of them is *N*-acryloylaminoethoxyethanol (AAEE). As a free monomer, it has ten times higher resistance to hydrolysis than plain acrylamide, but as a polymer its stability (in 0.1 N NaOH, 70 °C) is 500 times higher [58]. The strategy: the ω-OH group in the *N*-substituent was sufficiently remote from the amido bond so as to impede formation of hydrogen bonds with the amido carbonyl group; if at all, the oxygen in the ethoxy moiety of the *N*-substituent would act as a preferential partner for hydrogen-bond formation with the ω-OH group. The second is acryloylaminopropanol (AAP), in which the ether group on the *N*-substituent was removed and the chain shortened [59]. As a result, this monomer was even more hydrophilic than AAEE and highly resistant to hydrolysis as compared to free acrylamide (hydrolysis constant of AAP $0.008\,\mathrm{L\,mol^{-1}\,min^{-1}}$ versus $0.05\,\mathrm{L\,mol^{-1}\,min^{-1}}$ for acrylamide, in an alkaline milieu). The AAP monomer was found to have a unique performance in DNA separations when used as a sieving liquid polymer in capillary zone electrophoresis (CZE) [60]. In addition, poly(AAP) has excellent performance in CZE at high temperatures (60 °C and higher), which are necessary for screening for DNA point mutations in temperature-programmed CZE [61]. It is known that acrylamide is a neurotoxin and possibly even (at high doses) a carcinogen. Now the situation could have dramatically changed: if we assess toxicity in terms of the ability of these monomers to alkylate free SH groups in proteins (thus forming a cysteinyl-*S*-β-propionamide adduct) and evaluate this reaction by high-resolution, delayed extraction MALDI-TOF MS (matrix-assisted laser desorption ionization time of flight mass spectrometry) both in the linear and reflectron mode, this picture will change substantially: at one end of the scale the major offenders remain indeed acrylamide and DMA (strongly alkylating, hence highly toxic agents), whereas at the opposed end (minimally alkylating) one can locate only AAP [62]. Tables 1 and 2 list a number of monomers and cross-linkers, respectively, in use today.

4. Discontinuous Electrophoresis

In 1959 Raymond and Weintraub [40] described the use of polyacrylamide gels (PAG) in ZE, offering UV and visible transparency (starch gels are opalescent) and the ability to sieve macromolecules over a wide range of sizes. In 1964, Ornstein [42] and Davis [41] created discontinuous (disc) electrophoresis by applying a series of discontinuities to PAG (lead-

Table 1. Monomers utilized for preparing polyacrylamide gels

Name	Chemical formula	M_r
N,N-Dimethylacrylamide (DMA)	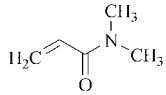	99
N-Acryloyl-2-amino-2-hydroxymethyl-1,3-propane (Trisacryl)	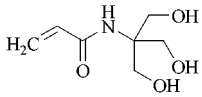	175
N-Acryloylmorpholine	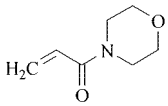	141
N-Acryloyl-1-amino-1-deoxy-D-glucitol	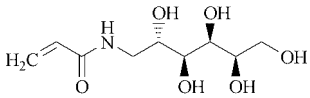	235
N-Acryloylaminoethoxyethanol	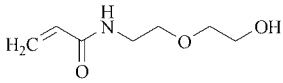	159
N-Acryloylaminopropanol	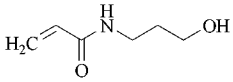	129
N-Acryloylaminobutanol	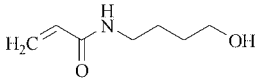	143

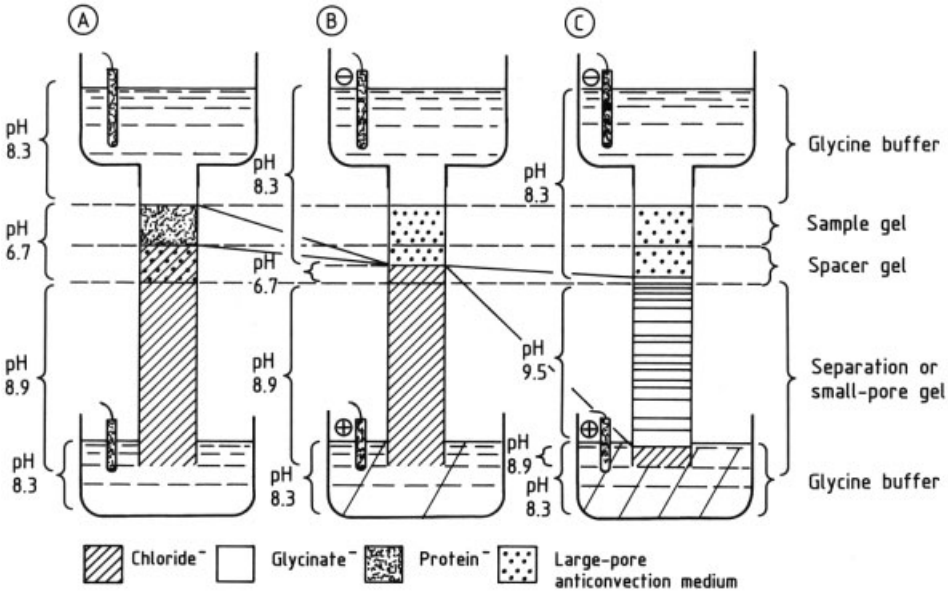

Figure 3. Principle of discontinuous electrophoresis
A) Sample in sample gel; B) Sample concentration in stacking gel; C) Sample separation in small-pore, separation gel [42]

Table 2. Cross-linkers described for gelling polyacrylamide matrices

Name	Chemical formula	M_r	n*
N,N′-Methylene bisacrylamide (Bis)	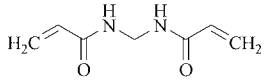	154	9
Ethylene diacrylate (EDA)	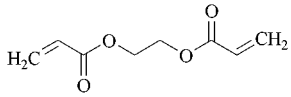	170	10
N,N′-(1,2-Dihydroxyethylene) bisacrylamide (DHEBA)	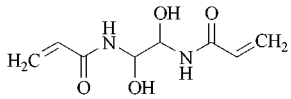	200	10
N,N′-Diallyltartardiamide (DATD)	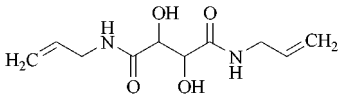	228	12
N,N′,N′′-Triallyl citric triamide (TACT)	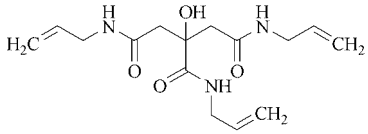	309	12 – 13
Polyethylene glycol diacrylate 200 ($PEGDA_{200}$)	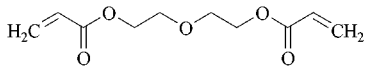	214	13
N,N′-Bisacrylylcystamine (BAC)	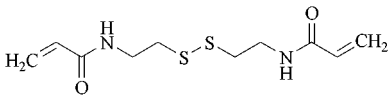	260	14
Polyethylene glycol diacrylate 400 ($PEGDA_{400}$)	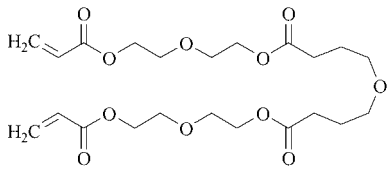	400	25
N,N'-Bis-acrylylpiperazine (BAP)	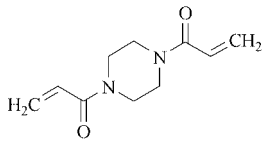	194	10

∗ Chain length.

ing and terminating ions, pH, conductivity, and porosity), thus further increasing the resolving power of the technique. In disc electrophoresis (Fig. 3, see previous page) the proteins are separated on the basis of two parameters: surface charge and molecular mass. FERGUSON [63] showed that one can distinguish between the two by plotting the results of a series of experiments with polyacrylamide gels of varying porosity. For each protein under analysis, the slope of the curve log m_T (electrophoretic mobility) versus gel density (%T) is proportional to molecular mass, while the *y*-intercept is a measure of surface charge [64]. Nonlinear Ferguson plots have been reported [65], related to the reptation mode of DNA in sieving media.

5. Isoelectric Focusing

New routes were explored toward a simpler assessment of charge and mass of proteins. As early as 1961 Svensson [66] experimented with isoelectric focusing (IEF), an electrophoretic technique by which amphoteric compounds are fractionated according to their isoelectric point (pI) along a continuous pH gradient. In contrast to ZE, where the constant pH of the separation medium establishes a constant charge density at the surface of the molecule and causes it to migrate with constant mobility (in the absence of sieving), the surface charge of an amphoteric compound in IEF keeps changing and decreasing according to its titration curve, as it moves along the pH gradient toward its steady-state position, i.e., the region where the pH matches its pI. There its mobility is zero and the molecule comes to a stop.

The gradient is created and maintained by the passage of current through a solution of amphoteric compounds with closely spaced pI values, encompassing a given pH range. The electrophoretic transport causes these buffers (carrier ampholytes, CA) to stack according to their pI values, and a pH gradient, increasing from anode to cathode, is established. This process is shown in Figure 4, which gives the formulae of the CA buffers and portions of their titration curves in the proximity of the pI value. When the system has achieved a steady state, maintained by the electric field, no further mass transport is expected except from symmetric to-and-fro micromovements of each species about its pI, generated by the action of two opposite forces, diffusion and voltage gradient, acting on each focused component. Thus, this diffusion – electrophoresis process is the primary cause of the residual current under isoelectric steady-state conditions.

The technique applies only to amphoteric compounds and, more precisely, to good ampholytes with small — pI-pK_1 — values, i.e.,

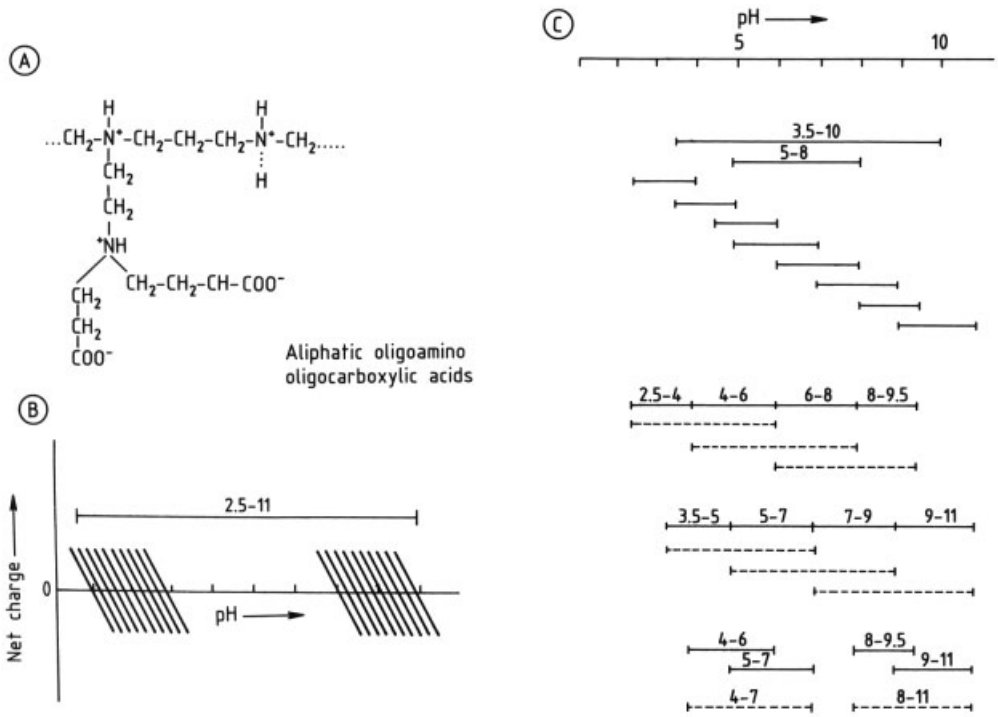

Figure 4. Composition of carrier ampholytes for isoelectric focusing
A) Representative chemical formula (oligoamino oligocarboxylic acids); B) Portions of titration curves of carrier ampholytes near pI; C) Narrow pH intervals, obtained by subfractionation of the wide pH 3.5 – 10 gradient components

with a steep titration curve around their pI (Fig. 4); the essential condition for any compound to focus in a narrow band. In practice, notwithstanding the availability of CAs covering the pH range 2.5 – 11, the limit of CA-IEF is in the pH 3.5 – 10 range, due to solvent conductivity outside this interval. When a restrictive support, such as polyacrylamide, is used, a limit also exists for the size of the largest molecules exhibiting an acceptable mobility through the gel. A conservative evaluation sets an upper limit around 750 kDa for standard techniques. The molecular form in which the proteins are separated depends strongly on the presence of additives, such as urea and/or detergents. Moreover, supramolecular aggregates or complexes with charged ligands can be focused only if their K_d is lower than 1 μmol/L and if the complex is stable at pH = pI. Aggregates with higher K_d are easily split by the pulling force of current, whereas most chromatographic procedures are unable to modify the native molecular form.

The general properties of the carrier ampholytes, i.e., of the amphoteric buffers used to generate and stabilize the pH gradient in IEF, can be summarized as follows:

1) Fundamental "classical" properties: (a) zero buffering ion mobility at pI; (b) good conductance; (c) good buffering capacity
2) Performance properties: (a) good solubility; (b) no influence on detection system; (c) no influence on sample; (d) separable from sample
3) "Phenomena" properties: (a) plateau effect, drift of the pH gradient; (b) chemical change in the sample; (c) complex formation

In chemical terms, CAs are oligoamino, oligocarboxylic acids, available from different suppliers under different trade names (Ampholine from LKB Produkter, Pharmalyte from Pharmacia Fine Chemicals, Biolyte from Bio-Rad, Servalyte from Serva, Resolyte from BDH). There are two basic synthetic approaches: (1) the Vesterberg approach [67], allowing different oligoamines (tetra-, penta-, and hexa-amines) to react with acrylic acid; and (2) the Pharmacia synthetic process [68], which involves the copolymerization of amines, amino acids, and dipeptides with epichlorohydrin. The wide-range synthetic mixture (pH 3 – 10) seems to contain hundreds, possibly thousands of different amphoteric chemicals, having pI values evenly distributed along the pH scale. CAs, from any source, should have an average molecular mass around 750 Da (range 600 – 900 Da, the higher values referring to the more acidic CA species). Thus, they should be easily separable from macromolecules by gel filtration, unless they are hydrophobically complexed to proteins. A further complication arises from the chelating effect of acidic CAs, especially toward Cu^{2+} ions, which could inactivate metalloenzymes. In addition, focused CAs represent a medium of very low ionic strength (less than 1 mequiv./L at steady state) [3,69]. Since the isoelectric state involves a minimum of solvation and thus of solubility for the protein macroion, there could be a tendency for some proteins (e.g., globulins) to precipitate during IEF close to the pI. This is a severe problem in preparative IEF, but in analytical procedures it can be lessened by reducing the total amount of sample applied.

6. Sodium Dodecyl Sulfate Electrophoresis

Sodium dodecyl sulfate (SDS) electrophoresis fractionates polypeptide chains essentially on the basis of their size. It is therefore a simple, yet powerful and reliable method for molecular mass determination. In 1967 Shapiro et al. [70] first reported that electrophoretic migration in SDS is proportional to the effective molecular radius and, thus, to the M_r of the polypeptide chain. This means that SDS must bind to proteins and cancel out differences in molecular charge, so that all components migrate solely according to size. Surprisingly large amounts of SDS appear to be bound (an average of 1.4 g/g protein). This means that the number of SDS molecules bound is approximately half the number of amino acid residues in a polypeptide chain. This amount of highly charged surfactant molecules is sufficient to overwhelm effectively the intrinsic charges of the polymer coil, so that their net charge per unit mass becomes approximately constant. If migration in SDS (and disulfide reducing agents, such as 2-mercaptoethanol, in the denaturing step, for proper unfolding of the proteins) is proportional only to M_r, then, in addition to canceling out charge differences, SDS

also equalizes molecular shape differences (e.g., globular vs. rod-shaped molecules). This seems to be the case for protein – SDS mixed micelles. These complexes can be assumed to behave as ellipsoids of constant minor axis (ca. 1.8 nm) and a major axis proportional to the length of the amino acid chain (i.e., molecular mass) of the protein. The rod length for the 1.4 g/g protein complex is of the order of 0.074 nm per amino acid residue. For further information on detergent properties, see HELENIUS and SIMONS [71].

In SDS electrophoresis, the proteins can be prelabeled with dyes that covalently bind to their NH_2 residues. The dyes can be conventional (e.g., the blue dye Remazol) or fluorescent, such as dansyl chloride, fluorescamine, *O*-phthaldialdehyde, and MDPF (2-methoxy-2,4-diphenyl-3[2*H*]-furanone). Prelabeling is compatible with SDS electrophoresis, as the size increase is minimal, but would be disastrous in disc electrophoresis or IEF, as it would generate a series of bands of slightly altered mobility or pI from an otherwise homogeneous protein.

For data treatment the sample and M_r standards are electrophoresed side-by-side in a gel slab. After detection of the polypeptide zones, the migration distance (or R_F) is plotted against log M_r to produce a calibration curve [72] from which the M_r of the sample can be calculated. Note that, in a gel of constant %T, linearity is obtained only in a certain range of molecular size. Outside this range a new gel matrix of appropriate porosity should be used. Two classes of protein show anomalous behavior in SDS electrophoresis: glycoproteins (because their hydrophilic oligosaccharide units prevent hydrophobic binding of SDS micelles) and strongly basic proteins, e.g., histones (because of electrostatic binding of SDS micelles through their sulfate groups). The first can be partially alleviated by using alkaline Tris/borate buffers [73], which will increase the net negative charge on the glycoprotein and thus produce migration rates well correlated with molecular size. The migration of histones can be improved by using pore gradient gels and allowing the polypeptide chains to approach the pore limit [74].

7. Porosity Gradient Gels

When macromolecules are electrophoresed in a continuously varying matrix concentration (which results in a porosity gradient), rather than in a gel of uniform concentration, the protein zones are compacted along their track, because the band front is, at any given time, at a gel concentration somewhat higher than at the rear of the band, so that the former is decelerated continuously. A progressive band sharpening thus results. There are other reasons for resorting to gels of graded porosity. Disc electrophoresis separates macromolecules on the basis of both size and charge differences. If the influence of molecular charge could be eliminated, then clearly the method could be used, with a suitable calibration, for measuring molecular size. This has been accomplished by overcoming charge effects in two main ways. In one, a relatively large amount of charged ligand, such as SDS, is bound to the protein, effectively swamping the initial charges present on the protein molecules and giving a quasi-constant charge/mass ratio. However, in SDS electrophoresis proteins are generally dissociated into their constituent polypeptide subunits, and the concomitant loss of functional integrity and antigenic properties cannot be prevented. Therefore, the size of the original, native molecule must be evaluated in the absence of denaturing substances.

The second method for M_r measurements relies on a mathematical canceling of charge effects, following measurements of the mobility of native proteins in gels of different concentrations. This is the so-called Ferguson plot [63], discussed in Chapter 4. A third method for molecular size measurements uses gels of graded porosity. This method is characterized by high resolving power and relative insensitivity to variations in experimental conditions. Under appropriate conditions (at least $10\,kV \cdot h$), the mobility of most proteins becomes constant and eventually ceases as each constituent reaches a gel density region in which the average pore size approaches the diameter of the protein (pore limit) [75]. Thus, the ratio of the migration distance of a given protein to that of any other becomes constant after the proteins have all entered a gel region in which they are subjected to drastic sieving conditions. This causes the electrophoretic pattern to become constant af-

ter prolonged migration in a gel gradient. The gel concentration at which the migration rate for a given protein becomes constant is called the "pore limit": If this porosity is properly mapped with the aid of a suitable set of marker proteins, it is possible to correlate the migration distance to the molecular mass of any constituent in the mixture.

After electrophoresis is over and the proper experimental data are gathered, they can be handled by two-step or one-step methods. Among the former, the most promising approach appears to be that of LAMBIN and FINE [76], who observed that there is a linear relationship between the migration distance of proteins and the square root of electrophoresis time, provided that time is kept between 1 and 8 h. The slopes of the regression lines of each protein are an indication of molecular size. When the slopes of the various regression lines thus obtained are plotted against the respective molecular masses, a good linear fit is obtained, which allows molecular mass measurements of proteins between 2×10^5 and 10^6 Da. The shape of the proteins (globular or fibrillar), their carbohydrate content (up to 45 %), and their free electrophoretic mobilities (between 2.1 and $5.9 \times 10^{-5}\ \mathrm{cm^2\ V^{-1}\ s^{-1}}$) do not seem to be critical for proper M_r measurements by this procedure. One-step methods have been described by ROTHE and PURKHANBABA [77] who found that, when $\log M_\mathrm{r}$ is plotted against either D (distance migrated) or %T (acrylamide + Bis), a nonlinear correlation is always obtained. However, when $\log M_\mathrm{r}$ is plotted against $\sqrt{\%T}$or $\sqrt{D}$, a linear regression line is obtained, which allows the accurate determination of M_r values of proteins (standard deviation $\pm$ 3.7 %; Fig. 5). The correlations $\log M_\mathrm{r} - \sqrt{\%\mathrm{T}}$ or $\log M_\mathrm{r} - \sqrt{D}$are not significantly altered by the duration of electrophoresis. Therefore, a constant M_r value should be obtained for a stable protein, no matter how long electrophoresis has been going on. ROTHE and

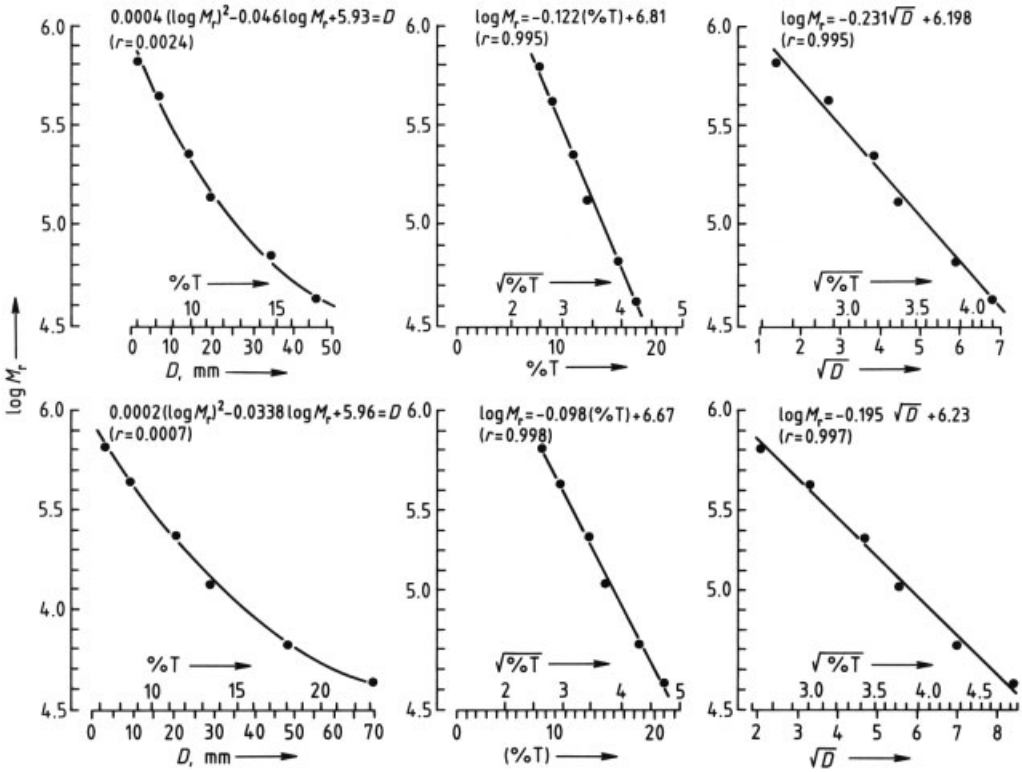

Figure 5. Typical plots of log M_r against migration distance D or gel composition %T after pore gradient electrophoresis Note that these plots are nonlinear, whereas when log M_r is plotted against Dor $\%T$a linear relationship is obtained [77].

MAURER [78] have demonstrated that the relationship log $M_r - \sqrt{D}$is also applicable to SDS electrophoresis in linear polyacrylamide gel gradients.

8. Two-Dimensional Maps (Proteome Analysis)

By sequentially coupling pure charge (IEF) to pure size fractionation (SDS-PAGE; the latter orthogonal to the first), one can distribute the polypeptide chains two-dimensionally, with charge and mass as coordinates (IEF-SDS or ISO-DALT, according to ANDERSON's nomenclature: ISOelectric for charge and DALT for mass separation) [79]. When the first dimension is performed in immobilized pH gradients, the technique is called IPG-DALT [4]. The technique was first reported by BARRETT and GOULD [80] and described in more detail by O'FARRELL [81], KLOSE [82], and SCHEELE [83]. Large gels (e.g., 30 × 40 cm) [84] and prolonged exposure to radiolabeled material (up to two months) have allowed resolution of as many as 12 000 labeled peptides in a total mammalian cell lysate. Thus, there is a good chance that, in a properly prepared 2-D map, a spot will represent an individual polypeptide chain, uncontaminated by other material. On this assumption, and provided that enough material is present in an individual spot (about 1 µg), it is possible to blot it onto a glass fiber filter and perform sequencing on it [85].

Two-dimensional electrophoresis (2-DE) excited grandiose projects, like the Human Protein Index System of Anderson & Anderson [79], who initiated the far-reaching goal of mapping all possible phenotypes expressed by all the different cells of the human organism. An Herculean task, if one considers that there may be close to 75 000 such phenotypes (assuming a total of ca. 250 differentiated cells, each expressing > 3000 polypeptides, of which 300 are specific to a given cell). This started a series of meetings of the 2-DE group, particularly strong in the field of clinical chemistry [86, 87]. A number of books were devoted to this 2-DE issue [88, 89], and a series of meetings was started [90]. 2-DE has found a proper forum in the journal *Electrophoresis*, which began hosting individual papers dealing with variagate topics in 2-D maps. Starting in 1988, *Electrophoresis* launched special issues devoted to 2-D maps, not only in clinical chemistry and human molecular anatomy, but in fact in every possible living organism and tissue, the first being dedicated to plant proteins [91]. Soon, a host of such "Paper Symposia" appeared, collecting databases on any new spots of which a sequence and a function could be elucidated, under the following editorships: CELIS [92–101]; DUNN [102–105]; LOTTSPEICH [106]; TÜMMLER [107]; WILLIAMS [108]; HUMPHERY-SMITH [109]; APPEL et al. [110, 111] and CASH [112]. This collection of "Paper Symposia" is a gold mine of new information and novel evolutionary steps on the IEF technique (such as new solubilizers, new staining procedures, sample pretreatment before IEF and the like). The world "proteome" is a recent neologism that refers to PROTEins expressed by the genOME or tissue. Just as genome has become a generic term for "big-science" molecular biology, so proteome is becoming a synonym of bioinformatics, since this project requires the building and continuous updating of a vast body of information, such as the amino acid sequence and biological function of the newly discovered proteins. An idea of the complexity of this field is given in [113, 114]. A partial list of the many sites available on the WWW for consulting databases and for visualizing codified 2-D maps is given in the following:

1) A 2D PAGE protein database of *Drosophila melanogaster*
   http://tyr.cmb.ki.se/
2) ExPASy Server (2D liver, plasma, etc. SWISS-(PROT, 2DPAGE, 3DIMAGE), BIOSCI, Melanie software)
   http://expasy.ch/
3) CSH QUEST Protein Database Center (2D REF52 rat, mouse embryo, yeast, Quest software)
   http://siva.cshl.org/
4) NCI/FCRDC LMMB Image Processing Section (GELLAB software)
   http://www-lmmb.ncifcrf.gov/lemkin/gellab.html
5) 2-D Images Meta-Database
   http://www-lmmb.ncifcrf.gov/2dwgDB/
6) *E. coli* Gene-Protein Database Project-ECO2DBASE (in NCBl repository)
   ftp://ncbi.nlm.nih.gov/repository/ECO2DBASE/

7) Argonne Protein Mapping Group Server (mouse liver, human breast cell, etc.)
http://www.anl.gov/CMB/PMG/
8) Cambridge 2D PAGE (a rat neuronal database)
http://sunspot.bioc.cam.uk/Neuron.html
9) Heart Science Centre, Harefield Hospital (Human Heart 2D Gel Protein DB)
http://www.harefield.mthames.nhs.uk/nhli/protein/
10) Berlin Human Myocardial 2D Electrophoresis Protein Database
http://www.chemie.fu-berlin.de/user/pleiss/
11) Max Delbruck Ctr. for Molecular Medicine-Myocardial 2D Gel DB
http://www.mdc-berlin.de/~emu/heart/
12) The World of Electrophoresis (EP literature, ElphoFit)
http://www.uni-giessen.de/~gh43/electrophoresis.html
13) Human Colon Carcinoma Protein Database (Joint Protein Structure Lab)
http://www.ludwig.edu.au/www/jpsl/jpslhome.html
14) Large Scale Biology Corp (2D maps: rat, mouse, and human liver)
http://www.lsbc.com/
15) Yeast 2D PAGE
http://yeast-2dpage.gmm.gu.se/
16) Proteome Inc. YPD Yeast Protein Database
http://www.proteome.com/YPDhome.html
17) Keratinocyte, cDNA Database (Danish Centre for Human Genome Research)
http://biobase.dk/cgi-bin/celis/
18) Institut de Biochemie et Genetique-Yeast 2D gel DB
http://www.ibgc.u-bordeaux2.fr/YPM/
19) Swiss-Flash newsletter
http://www.expasy.ch/swiss-flash/
20) Phosphoprotein Database
http://www-lmmb.ncifcrf.gov/phosphoDB/

Figure 6 (see next page) shows the complexity of one such 2-D map, downloaded from the ExPASy site, representing a partial panorama (limited to the pH gradient 5.5 – 8.5) of the thousands of polypeptide spots present in the human liver.

9. Isotachophoresis

The theoretical foundations of isotachophoresis (ITP) were laid in 1897 by Kohlrausch [115], who showed that, at a migrating boundary between two salt solutions, the concentrations of ions were related to their effective mobilities (Kohlrausch autoregulating function) [116]. The term isotachophoresis underlines the most important aspect of this technique, namely the identical velocities of the sample zones at equilibrium [117]. ITP will take place when an electric field is applied to a system of electrolytes, consisting of:

1) A leading electrolyte, which must contain only one ionic species, the leading ion L^-, having the same sign as the sample ions to be separated, and an effective mobility higher than that of any sample ions
2) A second, terminating electrolyte, which contains one ion species, the terminating ion T^-, having the same sign as the sample ions to be separated, and an effective mobility lower than that of any sample ions
3) An intermediate zone of sample ions, having the same sign as the leading and terminating ions and intermediate mobilities

The three zones are juxtaposed, with the proviso that sharp boundaries must be created at the start of the experiment. The polarity of the electric field must be such that the leading ion migrates in front of the ITP train at all times. When the system has reached the steady state, all ions move with the same speed, individually separated into a number of consecutive zones, in immediate contact with each other, and arranged in order of effective mobilities. Once the ITP train is formed, the ionic concentration in a separated sample zone adapts itself to the concentration of the preceding zone. The Kohlrausch function, which is given at the leading/terminating ion boundary, in fact gives the conditions at any boundary between two adjacent ions A^-, B^-, with one common counterion R^+ when the boundary migrates in the electric field.

There are two fundamental properties of ITP built into the autoregulating function: the concentrating effect and the zone-sharpening effect (Fig. 7). Suppose component A^- is introduced at very low concentration, even lower than that of the terminating ion T^-. Since the mobility of A^- is intermediate between that of L^- and T^-, its concentration must also be intermediate between those of L^- and T^-. This will result in A^- being concentrated (decrease in zone

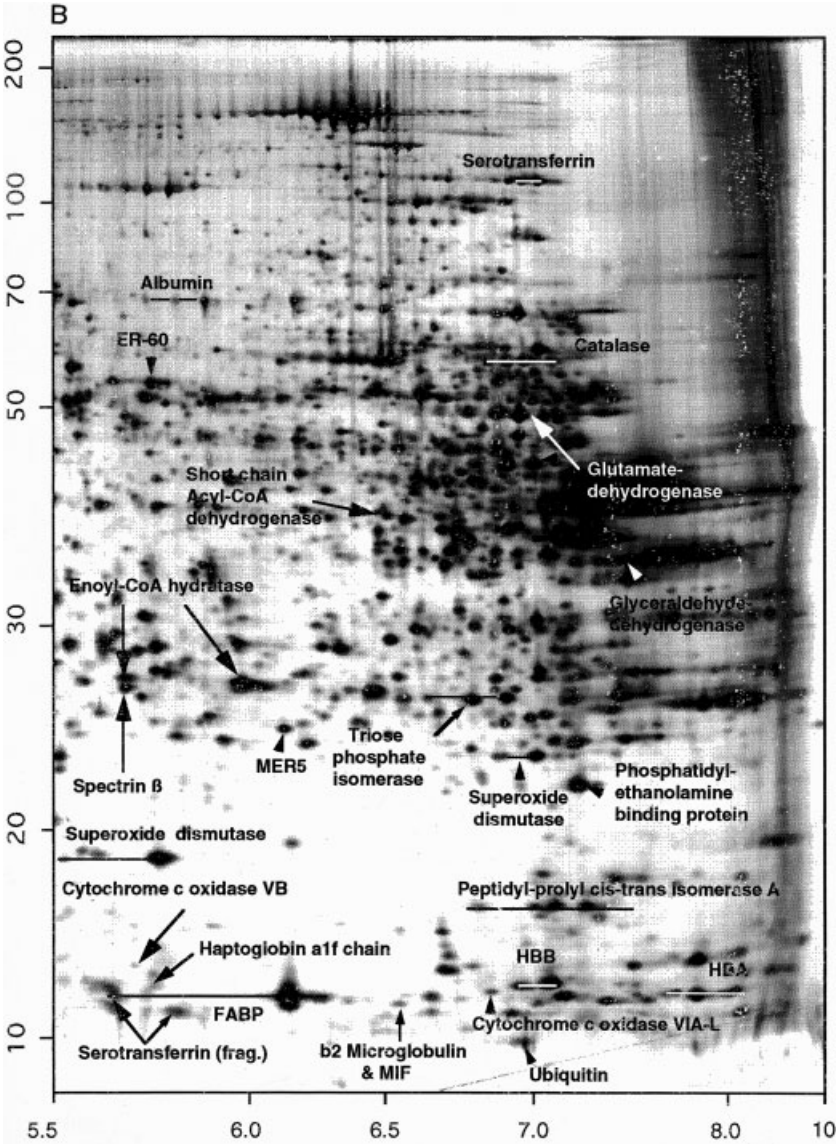

Figure 6. Partial 2-D map of human liver proteins, as resolved by IEF in the first dimension (*x*-axis, pH 5.5 – 8.5) and SDS-PAGE in the second dimension (*y*-axis, molecular mass).

length) until it reaches the theoretically defined concentration. Conversely, if the concentration of A^- is too high (even higher than that of L^-), the A^- zone will be diluted (increase in zone length) until the correct equilibrium concentration is reached. This is an example of the zone-sharpening effect: if A^- diffuses into the B^- zone, it is accelerated, since it enters a zone of higher voltage gradient, therefore, it automatically migrates back into its zone. Conversely, if it enters the L^- zone, it finds a region of lower voltage, is decelerated, and thus falls back into its zone (Fig. 7). This applies to all ions in the system. Note that the first few minutes of disc electrophoresis, the period of migration in the sample and stacking gels, represent in fact an ITP

migration. This is what produces the spectacular effects of disc electrophoresis: stacking into extremely thin starting zones (barely a few micrometers thick, from a zone up to 1 cm in thickness, a concentration factor of 1000 – 10 000) and sharpening of the zone boundaries. A characteristic of ITP is that the peaks, unlike those in other separation techniques (except for displacement chromatography), are square-shaped as a result of the Kohlrausch autoregulating function, i.e., the concentration of the substance within a homogeneous zone is constant from front to rear boundary [118]. In fact, under ideal conditions, the diffusive forces, which cause an eluted peak to spread into a Gaussian shape, are effectively counteracted. Although ITP is not much in vogue today, it is now used for stacking purposes as a transient step in CZE, for concentrating and sharpening dilute sample zones [119].

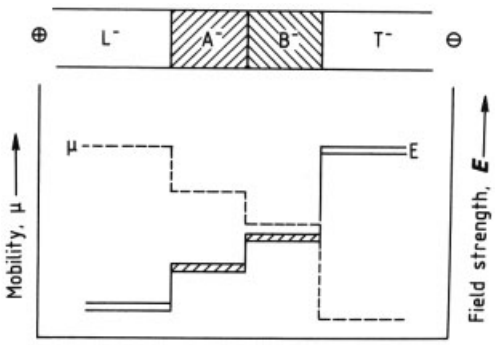

Figure 7. Scheme of an isotachophoretic process, with adjacent zones of leading ion L^-, samples A^- and B^- and terminating ion T^-
The respective mobilities (dotted line) and voltage gradients (solid line) in each zone are represented.

10. Immunoelectrophoresis

The basic technique of immunoelectrophoresis was first described by Grabar and Williams in 1953 [9]. The antigen sample is applied to a hole in the middle of a glass plate, coated with a 1-mm layer of agarose gel. Electrophoresis separates the antigen mixture into various zones. A longitudinal trough, parallel to the long edge of the plate, is then made in the gel and filled with the antiserum to the antigen mixture. A passive, double-diffusion takes place; the antiserum diffuses into the gel (advancing with a linear front) while the antigens diffuse radially in all directions from the electrophoresis zones. The antigen – antibody complexes are formed at equivalence points, the number of precipitates formed corresponding to the number of independent antigens present. The precipitates form a system of arcs, resulting from the combination of linear and circular fronts. A number of variants have been described over the years, but in this chapter only some of the most popular immunoelectrophoretic techniques are reviewed. For a more extensive treatise, see [126–128].

10.1. Rocket Immunoelectrophoresis

In rocket immunoelectrophoresis, the antigen – antibody reaction occurs during the electrophoresis of an antigen in an antibody-containing medium. Both antigen and antibody move according to their electrophoretic mobilities, and they also react with each other, resulting in flame-shaped precipitation zones of antigen – antibody complexes (Laurell's rocket technique) [120]. Under the influence of the electric field, unbound antigen within the peak of the flame-shaped precipitate migrates into the precipitate, which redissolves in the excess antigen. Thus, the leading edge of the flame is gradually displaced in the direction in which the antigen is migrating. The amount of antigen within the leading boundary edge is successively diminished because of the formation of soluble antigen – antibody complexes. When the antigen is diminished to equivalence with the antibody, the complexes can no longer be dissolved and a stable precipitate forms at the leading edge, which is thereafter stationary. The distance finally traveled by the peak depends on the relative excess of antigen over antibody and can be used as a measure of the amount of antigen present. Every precipitation band featured as a flame represents an individual antigen [120].

10.2. Crossed Immunoelectrophoresis

In the case of polydisperse samples, crossed immunoelectrophoresis (CIE) as described by Clarke and Freeman [121] is employed. First, a mixture of protein antigens is subjected to conventional electrophoresis in agarose. Then the

agarose gel is cut into strips, which are transferred to a second glass plate. Melted agarose (1 % w/v) containing the antiserum is then cast to form a gel of the same thickness as that of the first-dimension gel, and a secure junction between the two layers is achieved by melting to fuse their edges. During the second-dimension electrophoresis, performed perpendicular to the first, each antigen migrates independently, and precipitation zones are formed which resemble the Laurell rockets but have a wider base. Remarkably, two antigens present in a single zone, i.e., possessing equal electrophoretic mobilities, can be distinguished by the second-dimension electroimmunoassay as the shape and height of the respective rockets generally do not coincide exactly. In a typical product of the Clarke and Freeman method, ca. 50 rockets can be counted in human serum, and their relative abundance can be assessed by measuring the areas of the respective peaks [121]. Peak assessment and evaluation have been greatly facilitated by the development of a computer system for the specific analysis of CIE patterns [122].

10.3. Tandem Crossed Immunoelectrophoresis

It would be quite difficult to identify a single antigen in the complex patterns of CIE. One of the proposed methods for complex mixtures is the tandem CIE technique. Before the electrophoresis in the first dimension, two wells rather than one are cut into the gel strip, positioned one after the other in the direction of electrophoretic migration. One well is loaded with a mixture of the antigens to be analyzed, while the pure antigen whose peak is to be identified in the mixture is introduced into the other. The remaining manipulations are performed according to classical CIE. In the final pattern there will be two peaks which fuse smoothly and are separated by exactly the distance between the two wells. These two peaks may be of different heights (owing to difference in antigen concentration in the sample and reference wells), whereas fusion between them is indicative of the fact that they are caused by the same protein antigen. It is this canceling of the inner flanks of the two rockets and the fusion process that allow the unknown antigen to be located. By repeating this process with different purified antigens, it is possible to map, one by one, the components of a heterogeneous mixture [123].

10.4. Intermediate Gel Crossed Immunoelectrophoresis

Tandem CIE aimed at the identification of a single antigen in a mixture requires the availability of a pure marker antigen (see Section 10.3). In an alternative approach, identification can also be made when monospecific antiserum to the antigen is available. This approach is called intermediate gel CIE. After the first-dimension electrophoresis, the slab is prepared as for conventional CIE. Then a 1- to 2-cm strip of polyspecific antiserum-containing agarose gel nearest to the first-dimension gel is excised and a monospecific antiserum-containing agarose is cast instead. In the course of the second-dimension electrophoresis, the antigen to be detected will precipitate with the specific antiserum just as it enters the intermediate gel, while other antigens will pass through this gel without being retarded. The bases of their precipitation peaks, formed in the polyspecific antiserum-containing agarose, will be positioned on the borderline between the two latter gels. The antigen under study is thus distinguishable from the others [124].

10.5. Fused-Rocket Crossed Immunoelectrophoresis

In contrast to the Laurell rocket technique, this method is not used for quantitative determinations. Rather, it is applied to detect heterogeneity in seemingly homogeneous protein fractions, obtained by gel chromatography or ion exchange chromatography. A set of wells is punched in a checkerboard pattern in a strip of antibody-free agarose gel. Aliquots of each fraction eluted from the column are placed in the wells and the gel is left in a humid chamber for 30 – 45 min to allow the proteins in the wells to migrate and fuse, thus reproducing the continuous elution profile on an extended scale (hence the term "fused rockets"). On the remainder of the plate a thin layer of agarose is cast containing a polyvalent antiserum, and electrophoresis is performed

as usual. If the collected chromatographic peak is indeed homogeneous, a single fused rocket will appear, while inhomogeneity will be revealed by an envelope of subpeaks within the main eluate fraction [125].

11. Staining Techniques and Blotting

MERRIL et al. [129] described a silver-staining procedure in which the sensitivity – which in Coomassie Blue is merely on the order of a few micrograms per zone – is increased to a few nanograms of protein per zone, thus approaching the sensitivity of radioisotope labeling. In the gilding technique of MOEREMANS et al. [130] polypeptide chains are coated with 20-nm particles of colloidal gold and detected with a sensitivity of <1 ng/mm^2. Proteins can also be stained with micelles of Fe^{3+}, although with a sensitivity about one order of magnitude lower than with gold micelles [131]. These last two staining techniques became a reality only after the discovery of yet another electrophoretic method, the so-called Southern [132] and Western [133] blots, in which nucleic acids or proteins are transferred from hydrophilic gels to nitrocellulose or any of a number of other membranes, where they are immobilized by hydrophobic adsorption or covalent bonding. The very large porosity of these membranes makes them accessible to colloidal dyes. In addition, transfer of proteins to thin membranes greatly facilitates detection by immunological methods. This has resulted in new, high-sensitivity methods called "immunoblotting." After saturation of potential binding sites, the antigens transferred to the membrane are first made to react with a primary antibody, and then the precipitate is detected with a secondary antibody, tagged with horseradish peroxidase, alkaline phosphatase, gold particles, or biotin, which is then allowed to react with enzyme-linked avidin [134]. In all cases the sensitivity is greatly augmented. It appears that immunoblotting will pose a serious threat to quite a few of the standard immunoelectrophoretic techniques [135]. In terms of colloidal staining, a direct staining method for polyacrylamide gels with colloidal particles of Coomassie Blue G-250 is said to be as sensitive as silver staining [136]. For an extensive review covering all aspects of staining techniques, together with a vast bibliography on polypeptide detection methods, see [37].

12. Immobilized pH Gradients

In 1982, immobilized pH gradients (IPG) were introduced, resulting in an increase in resolution by one order of magnitude compared with conventional IEF [138]. Conventional IEF was besieged by several problems: (1) very low and unknown ionic strength; (2) uneven buffering capacity; (3) uneven conductivity; (4) unknown chemical environment; (5) nonamenable to pH gradient engineering; (6) cathodic drift (pH gradient instability); and (7) low sample loadability. In particular, a most vexing phenomenon was the near-isoelectric precipitation of samples of low solubility at the isoelectric point or of components present in large amounts in heterogeneous samples. The inability to reach stable steady-state conditions (resulting in a slow pH gradient loss at the cathodic gel end) and to obtain narrow and ultranarrow pH gradients, aggravated matters. Perhaps, most annoying was the lack of reproducibility and linearity of pH gradients produced by the so-called carrier ampholyte buffers [3]. IPGs proved able to solve all these problems.

IPGs are based on the principle that the pH gradient, which exists prior to IEF itself, is copolymerized, and thus rendered insoluble, within the fibers of a polyacrylamide matrix (Fig. 8). This is achieved by using, as buffers, a set of eight commercial chemicals (known as Immobilines, by analogy with Ampholine, produced by Pharmacia-LKB Biotechnologies, Uppsala, Sweden) having pK values distributed over the pH range 3.1 – 10.3. They are acrylamido derivatives with the general formula $CH_2{=}CH{-}CO{-}NH{-}R$, where R denotes one of three weak carboxyl groups, with pK values of 3.1, 3.6, and 4.6, for the acidic compounds, or one of five tertiary amino groups, with pK values of 6.2, 7.0, 8.5, 9.3, and 10.3, for the basic buffers [139]. Two additional compounds are needed: a strongly acidic (pK 1.0) and a strongly basic (p$K > 12$) titrant for producing linear pH gradients covering the entire pH range 3 – 10 [140]. Computer simulations had shown that, in the absence of these two titrants, extended pH intervals would exhibit strong deviations from linearity at the two extremes, as the most acidic and most

basic of the commercial Immobilines would act simultaneously as buffers and titrants [141]. The 2-acrylamidoglycolic acid (p*K* 3.1) species [142] is useful for separating strongly acidic proteins, since it extends the pH gradient to as low as pH 2.5. *N,N′*-Diethylaminopropylacrylamide (p*K* 10.3) has been used for analysis of strongly alkaline proteins [143]. Given the fairly evenly spacing of the p*K* values along the pH scale, it is clear that the set of ten chemicals proposed here (eight buffers and two titrants) is quite adequate to ensure linear pH gradients in the range pH 2.5 – 11 (the ideal ΔpK for linearity would be 1 pH unit between two adjacent buffers). The rule $\Delta pK = 1$ is fairly well obeyed, except for two "holes" between p*K* 4.6 and 6.2 and between p*K* 7.0 and 8.5. For a more detailed treatise on how to use an IPG gel, and IPG recipes, the reader is referred to an extensive manual [4] and to reviews [144–148]. Owing to the much increased resolution of IPG, quite a number of so-called electrophoretically silent mutations (bearing amino acid replacements with no ionizable groups in the side chains) have now been fully resolved.

13. Capillary Zone Electrophoresis

Capillary zone electrophoresis (CZE) appears to be a most powerful technique, perhaps equaling the resolving power of IPG. If one assumes that longitudinal diffusion is the only significant source of zone broadening, then the number of theoretical plates *N* in CZE is given by [149]:

$$N = \mu V / 2D$$

where μ and *D* are the electrophoretic mobility and diffusion coefficient, respectively, of the analyte, and *V* is the applied voltage. This equation shows that high voltage gradients are the most direct way to high separation efficiencies. For nucleic acids, it has been calculated that *N* could be as high as 10^6 theoretical plates. Figure 9 is a schematic drawing of a CZE system, illustrating one of the possible injection systems, namely gravity feed [150]. The fused-silica capillary has a diameter of 50 – 100 µm and a length up to 1 m. It is suspended between two reservoirs, connected to a power supply that is able to deliver up to 30 kV (typical operating currents 10 – 100 μA). One of the simplest ways

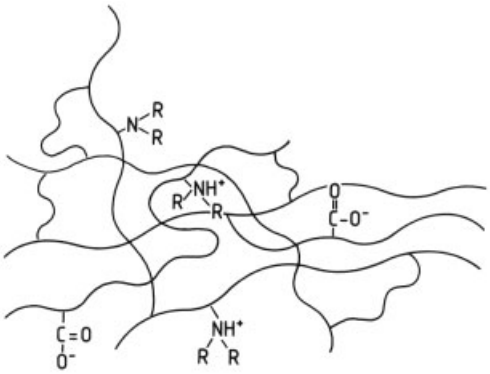

Figure 8. Scheme of a polyacrylamide matrix with a grafted pH gradient, depicted as bound negative and positive charges The ratio of such charges in each gel layer between anode and cathode defines a unique pH value.

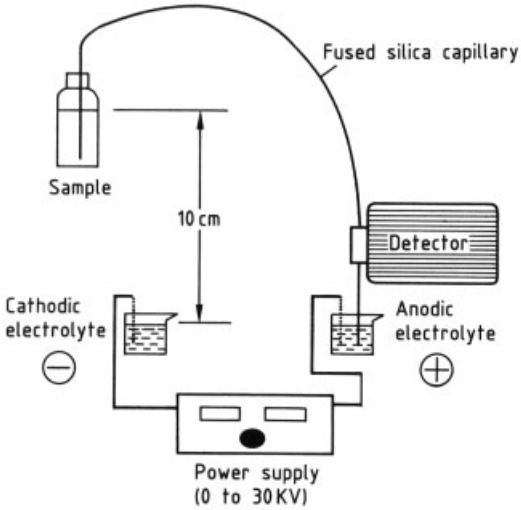

Figure 9. Scheme of a capillary zone electrophoresis apparatus
The detector includes a high-pressure mercury – xenon arc lamp, oriented perpendicular to the migration path, at the end of the capillary. The sample signal is measured with a photomultiplier and a photometer connected to the analog/digital converter of a multifunction interface board, connected to a microcomputer.

to introduce the sample into the capillary is by electromigration, i.e., by dipping the capillary extremity into the sample reservoir, under voltage, for a few seconds. Detection is usually accomplished by on-column fluorescence and/or UV absorption. Conductivity and thermal detectors, as usually employed in ITP, exhibit too low a sensitivity in CZE. The reason for this stems from the fact that the flow cell where sample monitoring occurs has a volume of barely 0.5 nL, allowing sensitivities down to the femtomole level. In fact, with laser-induced fluorescence detection, a sensitivity on the order of 10^{-21} mol is claimed [151]. By forming a chiral complex with a component of the background electrolyte (copper aspartame) it is possible to resolve racemates of amino acids [8]. Even neutral organic molecules can be made to migrate in CZE by complexing them with charged ligands, such as SDS. This introduces a new parameter, a hydrophobicity scale, in electrokinetic migrations. For more on CZE, readers are referred to the Proceedings of the International Symposia on High-Performance Capillary Electrophoresis [152–162], which give an account of the evolution of and new developments in the CZE technique, spanning a ten-year period. Other special issues dedicated to CZE can be found in *J. Chromatogr. B* [163] and *J. Biochem. Biophys. Methods* [164]. In addition, the journal Electrophoresis has devoted a number of Paper Symposia to various aspects of CZE [165–171]. Two books which offer a broad coverage of the variegated aspects of CZE are [172] and [173].

It is of interest to note here some aspects of the CZE technique, which offers a unique performance, unrivalled by other electrokinetic methods, except for 2-D maps. These aspects regard the use of sieving liquid polymers instead of true gels for separation of macromolecules, notably proteins and nucleic acids. Any polymer above the entanglement threshold can exhibit sieving and separate macro-analytes almost as efficiently as true gels. Such polymers embrace the classical polyacrylamides, as well as a number of others, such as celluloses, poly(vinyl alcohol), and poly(vinyl pyrrolidone). The advantages: after each run, the polymers can be expelled from the lumen of the capillary and thus each new run can be performed under highly controlled and reproducible starting conditions, in the absence of carryover from previous runs. In addition, all celluloses are transparent to low wavelengths, down to 200 nm, thus permitting on-line detection of proteins and peptides via the adsorption of the peptide bond, with much increased sensitivity. Other interesting, developments regard the use of isoelectric buffers, i.e., amphoteric species as the sole buffering ion at pH = pI. Such buffers, due to their very low conductivity, allow delivery of high voltage gradients, with much increased resolution and very short analysis times. Examples on separations of nucleic acids in isoelectric His [174] and of peptides in isoelectric Asp [175] are given in Figures 10 and 11 (see next page), respectively.

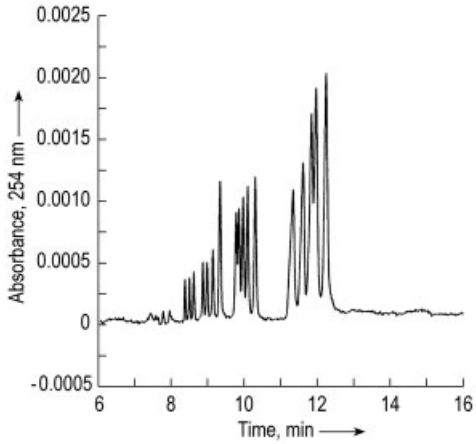

Figure 10. CZE of DNA marker V in isoelectric buffers. Conditions: 150 mM His buffer, pH=pI=7.6 in presence of 1.5 % poly(DMA); poly(AAP)-coated capillary, i.d. 100 µm, total length 30 cm (25.4 cm to the detector). UV absorbance detection at 254 nm. Run at 25 °C and 100 V/cm [174].

14. Preparative Electrophoresis

Several variants of electrophoresis have been described for preparative runs. They utilize either completely fluid beds or gel phases. This chapter restricts itself to those techniques exploiting isoelectric focusing principles, both in carrier ampholytes or in immobilized, non-amphoteric buffers, since they couple, in general, high loading capacities to a high resolving power. A classical example involves vertical columns filled with a sucrose density gradient, which is then emptied and collected in fractions at the end of the IEF run.

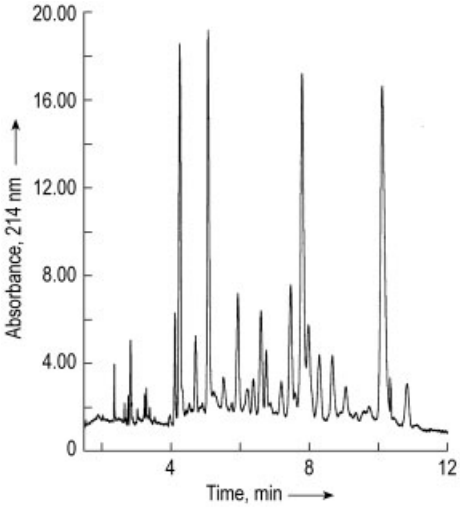

Figure 11. CZE of a tryptic digest of β-casein in a $37 \times 100\,\mu$m i.d. capillary, bathed in 50 mM isoelectric aspartic acid (pH approximating the pI value of 2.77) added with 0.5 % hydroxyethyl cellulose. Run: 600 V (current 58 μA). The three major peptides (pI 6.1, 6.93, and 3.95) were eluted and analyzed by mass spectrometry [175].

This was RILBE's original idea for performing IEF [176]; however, even though LKB has sold more than 3000 columns over the years, they are seldom used today. The following sections describe a few methods which appear to be more accessible; a complete survey of all other variants is available [3].

14.1. Preparative Isoelectric Focusing in Granulated Gel Layers

RADOLA [177, 178] described a method for large-scale preparative IEF in troughs coated with a suspension of granular gels, such as Sephadex G-25 superfine (7.5 g/100 mL), Sephadex G-200 superfine (4 g/100 mL), or Bio Gel P-60 (4 g/100 mL). The trough consists of a glass or quartz plate at the bottom of a Lucite frame (Fig. 12, see next page). Various sizes of trough may be used; typical dimensions are 40×20 cm or 20×20 cm, with a gel layer thickness of up to 10 mm. The total gel volume in the trough varies from 300 to 800 mL. A slurry of Sephadex (containing 1 % Ampholine) is poured into the trough and allowed to dry in air to the correct consistency. Achieving the correct consistency of the slurry before focusing seems to be the key to successful results.

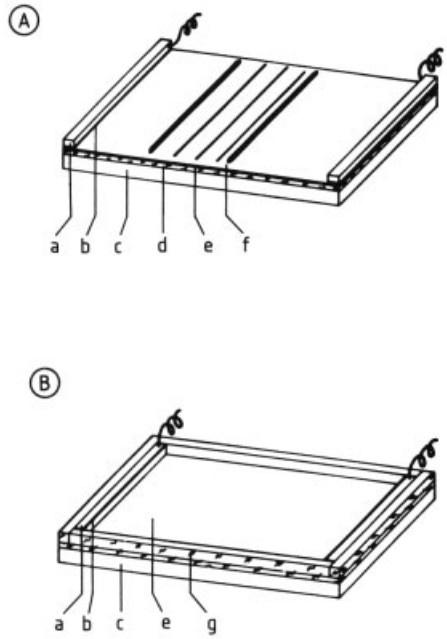

Figure 12. Preparative IEF in granulated Sephadex layers
A) Small-scale separation; B) Large-scale separation
a) Electrode; b) Filter paper pad soaked in electrode solution; c) Cooling block; d) Glass plate; e) Gel layer; f) Focused proteins; g) Trough [178].

Best results seem to be obtained when 25 % of the water in the gel has evaporated and the gel does not move when the plate is inclined at 45°. The plate is run on a cooling block maintained at 2 – 10°C. The electric field is applied via flat electrodes or platinum bands which make contact with the gel through absorbent paper pads soaked in 1 mol/L sulfuric acid at the anode, and 2 mol/L ethylenediamine at the cathode. In most preparative experiments, initial voltages of 10 – 15 V/cm and terminal voltages of 20 – 40 V/cm are used. Power gradients as high as 0.05 W/cm are well tolerated in 1-cm thick layers. Samples may be mixed with the gel suspension or added to the surface of preformed gels, either as a streak or from the edge of a glass slide. Larger sample volumes may be mixed with the dry gel (ca. 60 mg Sephadex per mL) and poured into a slot in the gel slab. Samples may be applied at any position between the electrodes. After focusing, proteins are located by the paper print technique. Focused proteins in the surface gel layer are absorbed onto a strip of filter paper which is then dried at 110°C. The proteins may

then be stained directly with dyes of low sensitivity, such as Light Green SF or Coomassie Violet R-150, after removing ampholytes by washing in acid. The pH gradient in the gel can be measured in situ with a combination microelectrode sliding on a calibrated ruler. Radola's technique offers the advantage of combining high resolution, high sample load, and easy recovery of focused components. As much as 5 – 10 mg/mL protein per milliliter gel suspension can be fractionated over wide pH ranges. Purification of 10 g pronase E in 800 mL gel suspension has been reported. At these high protein loads, even colorless samples can be easily detected, since they appear in the gel as translucent zones. As there is no sieving effect for macromolecules above the exclusion limits of the Sephadex, high molecular mass substances, such as virus particles, can be focused without steric hindrance. The system has a high flexibility, since it allows analytical and small- and large-scale preparative runs in the same trough, merely by varying the gel thickness.

For recovery of protein bands at the end of the IEF run, a fractionation grid pressed onto the gel layer allows recovery of 30 fractions, i.e., one fraction each 8 mm along a separation axis of 25 cm (Fig. 13). After scraping off each fraction with a spatula, the protein is eluted by placing the gel into a syringe equipped with glass wool as bottom filter, adding a buffer volume, and ejecting the eluate with the syringe piston.

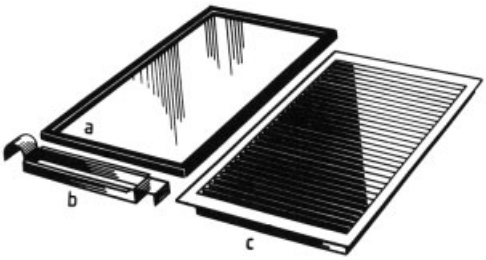

Figure 13. Trough (a), sample applicator (b), and fractionation grid (c) for preparative IEF in granulated gel layers (LKB Multiphor 2117 apparatus)

14.2. Continuous-Flow, Recycling Isoelectric Focusing

Another interesting approach, called recycling IEF, has been described by Bier's group [179, 180]. It is well known that continuous-flow techniques, which appear essential for large-scale preparative work, are disturbed by parabolic and electrosmotic flows, as well as by convective flows due to thermal gradients. Bier et al. have improved this system by separating the actual flow-through focusing cell, which is miniaturized, from the sample and heat-exchange reservoir, which can be built up to any size. Minimization of parabolic flow, electrosmosis, and convective liquid flow is achieved by flowing the sample to be separated through a thin focusing cell (the actual distance from anode to cathode is only 3 cm) built of an array of closely spaced filter elements oriented parallel to the electrodes and parallel to the direction of flow. Increased sample load is achieved by recirculating the process fluid through external heat-exchange reservoirs, where the Joule heat is dissipated. During each passage through the focusing cell, only small sample migrations toward their eventual pI are obtained, but through recycling a final steady state is achieved. The IEF cell has ten input and output ports for sample flow-through, monitored by an array of ten miniaturized pH electrodes and ten UV sensors. The entire system is controlled and operated by a computer. A scheme of the entire apparatus can be seen in Figure 14 (see next page). By activating pumps at two extreme channels, the computer can alter the slope of the pH gradient, counteracting any effect of cathodic drift, which results in a net migration of the sample zones towards the cathode.

14.3. The Rotofor

The Rotofor cell is a preparative-scale, free-solution, isoelectric focusing apparatus that provides a useful alternative to the recycling system described above. This cell [181] consists of a cylindrical focusing chamber that rotates around its axis, achieving liquid stabilization, not by recycling, but by gravity stabilization, as reported in 1967 by Hjertén [182]. The inside of the cylindrical focusing apparatus is divided by 19 parallel, monofilament, polyester membrane screens into 20 discrete compartments. Figure 15 gives a schematic view of the focusing chamber with the assembled Nylon screens supported by a cooling tube. There are several advantages of such a horizontal, rotating system.

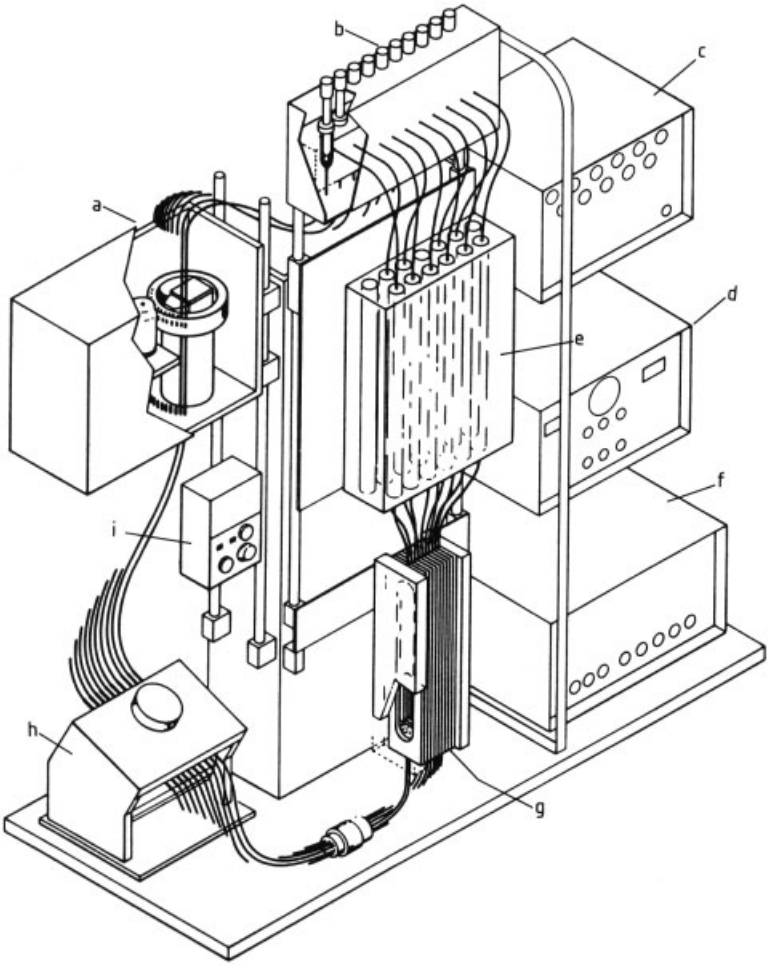

Figure 14. Scheme of the recycling IEF apparatus [180]
a) UV monitor; b) pH meter; c) Data interface; d) Power supply; e) Heat-exchange reservoir; f) UV control; g) Focusing cell; h) Pump; i) Pump control

First of all, in the classical vertical density gradient IEF column of VESTERBERG et al. [176] isoelectric protein precipitates (not at all uncommon at the high protein loads of preparative runs) used to sediment along the supporting density gradient, thus contaminating other focused zones along the separation path. The Rotofor cell overcomes this problem, since, by focusing in a horizontal axis, potential protein precipitates stay focused in their compartment and do not disturb the remainder of the gradient. Secondly, the vertical density gradient column for IEF had a weak point in the elution funnel: all the liquid in the column was emptied via a bottom outlet, and this resulted in severe remixing of the separated protein zones and in contamination by diffusion, owing to the long times required for harvesting the column content (usually 30 – 60 min, depending on column size). These problems are eliminated in the Rotofor apparatus, since the 20 chambers inside the column are emptied simultaneously and instantaneously by an array of 20 collection needles. The cylindrical focusing cell holds a total of 55 ml of solution for preparative protein purification; during operation, the chamber is rotated at 10 rpm around its axis to stabilize against convective disturbances. The run is usually performed at 12 W constant

power (up to 2000 V maximum voltage); focusing is in general accomplished in as little as 4 h. Once the system is at steady state, collection is quickly performed by using an array of 20 collection needles that lead to 20 test tubes nested in a specially designed vacuum collection box. The needles simultaneously pierce a tape sealing 20 small holes, each at the bottom of an individual chamber in the Rotofor. By applying vacuum to the collection box, the 20 compartments are simultaneously emptied. In common with other preparative IEF techniques, the Rotofor can be used in a cascade set-up, that is, the contents of a single cell, containing the protein of interest, can be spread over the entire 20-cell assembly, and focused a second time over a shallower pH gradient, upon dilution with an appropriate range of carrier ampholytes.

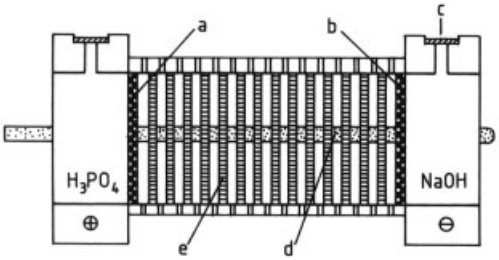

Figure 15. Scheme of the Rotofor cell
a) Cation-exchange membrane; b) Anion-exchange membrane; c) Vent caps; d) Ceramic cold finger; e) Screening material (6-µm polyester)

14.4. Recycling Free-Flow Focusing (RF3)

In the recycling IEF apparatus, stabilization of fluid flow is performed by screen elements; in the Rotofor, stabilization is achieved by screen elements and by rotation. In the RF3 apparatus shown schematically in Figure 16, fluid stabilization is achieved by means of rapid flow through narrow gaps, with continuous recycling. The key to stability of laminar flow through the cavity is its shallow depth (0.75 mm), combined with the rapid flow of process fluid (the latter necessary for avoiding electrohydrodynamic distortion). The lateral electrodes are separated from the focusing cavity by ion-permselective membranes. This instrument can also be used in the electrodialysis mode, by having the cation-selective membrane facing the cathodic compartment and the anion-selective membrane in front of the anodic chamber [183].

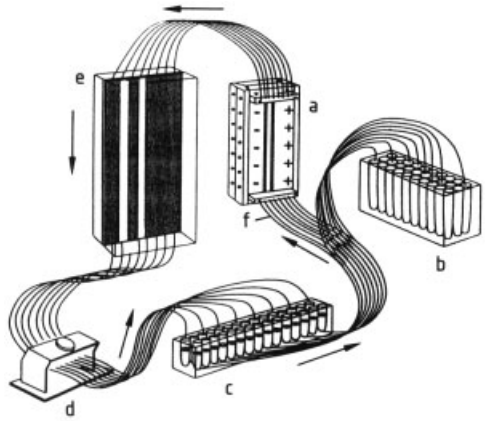

Figure 16. Scheme of the recycling free-flow focusing (RF3) apparatus
The arrows indicate the fluid flow pattern. At the end of the process, the flow pattern is switched to the harvesting position toward the fraction collector [183]
a) Focusing cell; b) Fraction collector; c) Bubble trap; d) Peristaltic pump; e) Heat exchanger; f) 30-Channel manifold

14.5. Multicompartment Electrolyzers with Isoelectric Membranes

Perhaps the final evolution of preparative IEF is the concept of a membrane apparatus, in which the entire IPG gel is reduced to isoelectric membranes delimiting a series of flow chambers. Figure 17 gives an exploded view of this novel apparatus [184, 185]: it consists of a stack of chambers sandwiched between an anodic and a cathodic reservoir. The apparatus is modular and in the present version can accommodate up to eight flow chambers. Figure 17 shows a stack of three chambers already assembled to the left, a central compartment, and a thinner chamber to the right for connection to the other electrode. All flow chambers are provided with inlet and outlet for sample or electrolyte recycling, an O-ring for ensuring flow-tight connections, and four holes for threading four long metal rods which can be tightened by hand-driven butterfly nuts for assembling the apparatus. Sev-

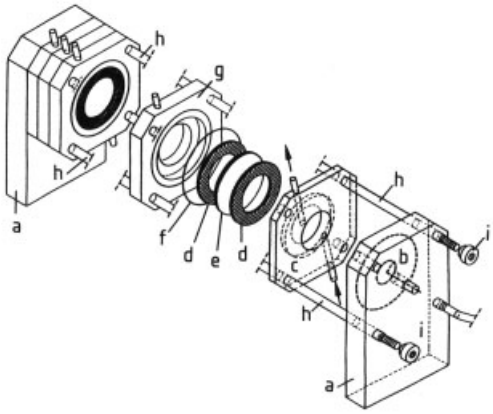

Figure 17. Exploded view of the multicompartment electrolyzer
a) Rectangular supporting legs; b) Pt electrode; c) Thin terminal flow chamber; d) Rubber rings for supporting the membrane; e) Isoelectric Immobiline membrane cast onto the glass-fiber filter; f) O-ring; g) One of the sample flow chambers; h) Threaded metal rods for assembling the apparatus; i) Nuts for fastening the metal bolts [128]

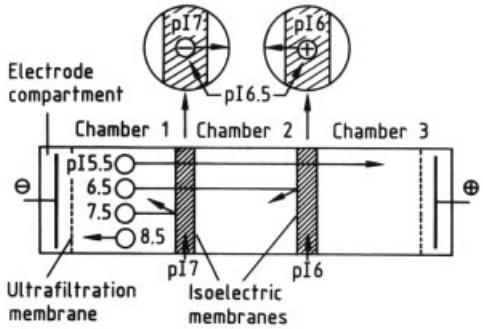

Figure 18. Mechanism of the purification process in the multicompartment electrolyzer
In the case shown, a protein with pI 6.5 is trapped between the pI 6 anodic and pI 7 cathodic membranes, whereas proteins with higher pI move toward the cathode and a species with lower pI (5.5) crosses the two membranes and collects at the anode [128].

eral versions of these cells have been built, capable of housing Immobiline membranes from 4.7 cm (the present apparatus) up to 9 cm diameter. The pH-controlling membranes are housed in the central depression between two 1-cm-wide caoutchouc rings. After assembling and tightening the apparatus, each compartment is flow-tight, so that no net liquid bulk flow ensues (except, when applicable, that generated by electrosmosis). The Pt electrodes are housed in two rectangular Perspex mountings, which also act as legs on which the electrolyzer stands. The distance between adjacent cells is only 10 mm, so that each chamber holds ca. 5 ml of liquid. The reason why this system works is shown in Figure 18: two isoelectric membranes facing each flow chamber act by continuously titrating the protein of interest to its isoelectric point. They can be envisaged as highly selective membranes, which retain any protein having pIs in between their limiting values, and which allow transmigration of any nonamphoteric, nonisoelectric species. The only condition required is that $pI_{cm} > pI_p > pI_{am}$, where the subscripts cm and am denote cathodic and anodic membranes, and p is the protein having a given isoelectric point between the two membranes [184]. For this mechanism to be operative, the two isoelectric membranes must possess good conductivity and good buffering capacity, so as to effectively titrate the protein present in the flow chamber to its pI, while ensuring good current flow through the system. WENGER et al. [186] had in fact synthesized amphoteric, isoelectric Immobiline membranes and demonstrated that they are good conductors and good buffers at their pI.

15. References

1. J. C. Giddings: *Unified Separation Science,* Wiley-Interscience, New York 1991, chap. 8.
2. A. T. Andrews: *Electrophoresis: Theory, Techniques and Biochemical and Clinical Applications,* Clarendon Press, Oxford 1986.
3. P. G. Righetti: *Isoelectric Focusing: Theory, Methodology and Applications,* Elsevier, Amsterdam 1983.
4. P. G. Righetti: *Immobilized pH Gradients: Theory and Methodology,* Elsevier, Amsterdam 1990.
5. C. J. Van Oss, *Sep. Purif. Methods* **8** (1979) 119 – 198.
6. P. G. Righetti, C. J. Van Oss, J. W. Vanderhoff (eds.): *Electrokinetic Separation Methods,* Elsevier, Amsterdam 1979.
7. B. L. Karger, *Nature (London)* **339** (1989) 641 – 642.
8. P. Gozel, E. Gassmann, H. Michelsen, R. N. Zare, *Anal. Chem.* **59** (1987) 44 – 49.
9. P. Grabar, C. A. Williams, *Biochim. Biophys. Acta* **10** (1953) 193 – 201.
10. O. Smithies, *Biochem. J.* **61** (1955) 629 – 636.
11. J. Kohn, *Nature (London)* **180** (1957) 986 – 987.

12. L. P. Cawley, B J. Minard, W. W. Tourtellotte, P. I. Ma, C. Clelle, *Clin. Chem. (Winston Salem N.C.)* **22** (1976) 1262 – 1268.
13. A. A. Keshgegian, P. Peiffer, *Clin. Chim. Acta* **108** (1981) 337 – 340.
14. J. Kohn, P. Priches, J. C. Raymond, *J. Immunol. Methods* **76** (1985) 11 – 16.
15. P. G. Righetti, P. Casero, G. B. Del Campo, *Clin. Chim. Acta* **157** (1986) 167 – 174.
16. P. Casero, G. B. Del Campo, P. G. Righetti, *Electrophoresis (Weinheim Fed. Repub. Ger.)* **6** (1985) 373 –376.
17. T. Toda, T. Fujita, M. Ohashi in R. C. Allen, P. Arnaud (eds.): *Electrophoresis '81,* De Gruyter, Berlin 1981, pp. 271 – 280.
18. P. Grabar, *Methods Biochem. Anal.* **7** (1957) 1 – 38.
19. C. Araki, *Proceedings 5th International Seaweed Symposium,* Halifax, 1965, pp. 3 – 17.
20. M. Duckworth, W. Yaphe, *Carbohydr. Res.* **16** (1971) 189 – 197.
21. S. Hjertén, *Biochim. Biophys. Acta* **53** (1961) 514 – 517.
22. S. Brishammar, S. Hjertén, B. Van Hofsten, *Biochim. Biophys. Acta* **53** (1961) 518 – 521.
23. S. Hjertén, *Arch. Biochem. Biophys.* **99** (1962) 466 – 475.
24. M. Duckworth, W. Yaphe, *Anal. Biochem.* **44** (1971) 636 – 641.
25. T. J. Låås, *J. Chromatogr.* **66** (1972) 347 – 355.
26. T. G. L. Hickson, A. Polson, *Biochim. Biophys. Acta* **168** (1965) 43 – 58.
27. S. Arnott, A. Fulmer, W. E. Scott, I. C. M. Dea, R. Moorhourse, D. A. Rees, *J. Mol. Biol.* **90** (1974) 269 – 284.
28. P. G. Righetti in R. C. Allen, R. Arnaud (eds.): *Electrophoresis '81,* De Gruyter, Berlin 1981, pp. 3 – 16.
29. J. J. Opplt in L. A. Lewis, J. J. Opplt (eds.): *Handbook of Electrophoresis,* CRC Press, Boca Raton 1980, pp. 151 – 180.
30. N. C. Stellwagen, *Adv. Electrophor.* **1** (1987) 177 – 228.
31. P. Serwer, *Electrophoresis (Weinheim Fed. Repub. Ger.)* **4** (1983) 375 – 382.
32. S. Hjertén, *J. Chromatogr.* **12** (1963) 510 – 526.
33. P. Serwer, E. T. Moreno, G. A. Griess in C. Schafer-Nielsen (ed.): *Electrophoresis '88,* VCH, Weinheim 1988, pp. 216 – 222.
34. P. G. Righetti, B. C. W. Brost, R. S. Snyder, *J. Biochem. Biophys. Methods* **4** (1981) 347 – 363.
35. P. G. Righetti, *J. Biochem. Biophys. Methods* **19** (1989) 1 – 20.
36. J. Uriel, J. Berges, *C. R. Hebd. Seances Acad. Sci. Ser. 3* **262** (1966) 164 – 170.
37. A. C. Peacock, C. W. Dingman, *Biochemistry* **7** (1968) 668 – 673.
38. E. Boschetti in P. G. D. Dean, W. S. Johnson, F. A. Middle (eds): *Affinity Chromatography,* IRL Press, Oxford 1985, pp. 11 – 15.
39. A. Chrambach, D. Rodbard, *Science (Washington D.C.)* **172** (1971) 440 – 451.
40. S. Raymond, L. Weintraub, *Science (Washington D.C.)* **130** (1959) 711 – 712.
41. B. J. Davis, *Ann. N.Y. Acad. Sci.* **121** (1964) 404 – 427.
42. L. Ornstein, *Ann. N.Y. Acad. Sci.* **121** (1964) 321 – 349.
43. S. Hjertén, *J. Chromatogr.* **11** (1963) 66 – 70.
44. S. Hjertén, *Anal. Biochem.* **3** (1962) 109 – 118.
45. S. Hjertén, *Arch. Biochem. Biophys. Suppl. 1* (1962) 276 – 282.
46. S. Hjertén, S. Jersted, A. Tiselius, *Anal. Biochem.* **27** (1969) 108 – 129.
47. J. H. Bode in B. J. Radola (ed.): *Electrophoresis '79,* De Gruyter, Berlin 1980, pp. 39 – 52.
48. P. B. H. O'Connell, C. J. Brady, *Anal. Biochem.* **76** (1976) 63 – 76.
49. H. S. Anker, *FEBS Lett.* **7** (1970) 293 – 296.
50. P. N. Paus, *Anal. Biochem.* **42** (1971) 327 – 376.
51. J. N. Hansen, *Anal. Biochem.* **76** (1976) 37 – 44.
52. C. Gelfi, P. G. Righetti, *Electrophoresis (Weinheim Fed. Repub. Ger.)* **2** (1981) 213 – 219.
53. A. Bianchi-Bosisio, C. Loherlein, R. S. Snyder, P. G. Righetti, *J. Chromatogr.* **189** (1980) 317 – 330.
54. P. G. Righetti, C. Gelfi, M. L. Bossi, E. Boschetti, *Electrophoresis (Weinheim Fed. Repub. Ger.)* **8** (1987) 62 – 70.
55. B. Kozulic, K. Mosbach, M. Pietrzak, *Anal. Biochem.* **170** (1988) 478 – 484.
56. C. Gelfi, P. De Besi, A. Alloni, P. G. Righetti, *J. Chromatogr.* **608** (1992) 333 – 341.
57. G. Artoni, E. Gianazza, M. Zanoni, C. Gelfi, M. C. Tanzi, C. Barozzi, P. Ferruti, P. G. Righetti, *Anal. Biochem.* **137** (1984) 420 – 428.
58. M. Chiari, C. Micheletti, M. Nesi, M. Fazio, P. G. Righetti, *Electrophoresis* **15** (1994) 177 – 186.
59. E. Simò-Alfonso, C. Gelfi, R. Sebastiano, A. Cittero, P. G. Righetti, *Electrophoresis* **17** (1996) 723 – 731; **17** (1996) 732 – 737; **17** (1996) 738 – 743.

60. C. Gelfi, M. Perego, F. Libbra, P. G. Righetti, *Electrophoresis* **17** (1996) 1342 – 1347.
61. E. Simò-Alfonso, C. Gelfi, M. Lucisano, P. G. Righetti, *J. Chromatogr. A* **756** (1996) 255 – 262.
62. E. Bordini, M. Hamdan, P. G. Righetti, *Rapid Commun. Mass Spectrom.* **13** (1999) 2209 – 2215.
63. K. A. Ferguson, *Metab. Clin. Exp.* **13** (1964) 985 –995.
64. H. R. Maurer: *Disc Electrophoresis,* De Gruyter, Berlin 1972.
65. A. Chrambach in C. Schafer-Nielsen (ed.): *Electrophoresis '88,* VCH, Weinheim 1988, pp. 28 – 40.
66. H. Svensson, *Acta Chem. Scand.* **15** (1961) 325 – 341; **16** (1962) 456 – 466.
67. O. Vesterberg, *Acta Chem. Scand.* **23** (1969) 2653 –2666.
68. L. Söderberg, D. Buckley, G. Hagström, *Protides Biol. Fluids* **27** (1980) 687 – 691.
69. C. Schafer-Nielsen in C. Schafer-Nielsen (ed.): *Electrophoresis '88,* VCH, Weinheim 1988, pp. 41 – 48.
70. A. L. Shapiro, E. Vinuela, J. V. Maizel, Jr., *Biochem. Biophys. Res. Commun.* **28** (1967) 815 – 822.
71. A. Helenius, K. Simons, *Biochim. Biophys. Acta* **415** (1975) 29 – 79.
72. J. L. Neff, N. Munez, J. L. Colburn, A. F. de Castro in R. C. Allen, P. Arnauds (eds.): *Electrophoresis '81,* De Gruyter, Berlin 1981, pp. 49 – 63.
73. J. F. Poduslo, *Anal. Biochem.* **114** (1981) 131 – 140.
74. P. Lambin, *Anal. Biochem.* **85** (1978) 114 – 124.
75. J. Margolis, K. G. Kenrick, *Anal. Biochem.* **25** (1968) 347 – 358.
76. P. Lambin, J. M. Fine, *Anal. Biochem.* **98** (1979) 160 – 168.
77. G. M. Rothe, M. Purkhanbaba, *Electrophoresis (Weinheim Fed. Repub. Ger.)* **3** (1982) 33 – 42.
78. G. M. Rothe, W. D. Maurer in M. J. Dunn (ed.): *Gel Electrophoresis of Proteins,* Wright, Bristol 1986, pp. 37 – 140.
79. N. G. Anderson, N. L. Anderson, *Clin. Chem. (Winston Salem N.C.)* **28** (1982) 739 – 748.
80. T. Barrett, H. J. Gould, *Biochim. Biophys. Acta* **294** (1973) 165 – 170.
81. P. O'Farrell, *J. Biol. Chem.* **250** (1975) 4007 – 4021.
82. J. Klose, *Humangenetik* **26** (1975) 231 – 243.
83. G. A. Scheele, *J. Biol. Chem.* **250** (1975) 5375 – 5385.
84. R. A. Colbert, J. M. Amatruda, D. S. Young, *Clin. Chem. (Winston Salem N.C.)* **30** (1984) 2053 – 2058.
85. R. H. Aebersold, J. Leavitt, L. E. Hood, S. B. H. Kent in K. Walsh (ed.): *Methods in Protein Sequence Analysis,* Humana Press, Clifton 1987, pp. 277 – 294.
86. D. S. Young, N. G. Anderson (eds.): Special Issue on Two Dimensional Electrophoresis, *Clin. Chem.* **28** (1982) 737 – 1092.
87. J. S. King (ed.): Special Issue on Two Dimensional Electrophoresis, *Clin. Chem.* **30** (1984) 1897 – 2108.
88. J. E. Celis, R. Bravo (eds.): *Two Dimensional Gel Electrophoresis of Proteins,* Academic Press, Orlando 1984, pp. 1 – 487.
89. B. S. Dunbar: *Two-Dimensional Electrophoresis and Immunological Techniques,* Plenum Press, New York 1987, pp. 1 – 372.
90. M. J. Dunn (ed.): *2-D PAGE '91,* Zebra Printing, Perivale 1991, pp. 1 – 325.
91. C. Damerval, D. de Vienne (eds.): Paper Symposium: Two Dimensional Electrophoresis of Plant Proteins, *Electrophoresis* **9** (1988) 679 – 796.
92. J. E. Celis (ed.): Paper Symposium: Protein Databases in Two Dimensional Electrophoresis, *Electrophoresis* **10** (1989) 71 – 164.
93. J. E. Celis (ed.): Paper Symposium: Cell Biology, *Electrophoresis* **11** (1990) 189 – 280.
94. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **11** (1990) 987 – 1168.
95. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **12** (1991) 763 – 996.
96. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **13** (1992) 891 – 1062.
97. J. E. Celis (ed.): Paper Symposium: Electrophoresis in Cancer Research, *Electrophoresis* **15** (1994) 307 – 556.
98. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **15** (1994) 1347 – 1492.
99. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **16** (1995) 2175 – 2264.
100. J. E. Celis (ed.): Paper Symposium: Two Dimensional Gel Protein Databases, *Electrophoresis* **17** (1996) 1653 – 1798.
101. J. E. Celis (ed.): Genomics and Proteomics of Cancer, *Electrophoresis* **20** (1999) 223 – 243.

102. M. J. Dunn (ed.): Paper Symposium: Biomedical Applications of Two-Dimensional Gel Electrophoresis, *Electrophoresis* **12** (1991) 459 – 606.
103. M. J. Dunn (ed.): 2D Electrophoresis: from Protein Maps to Genomes, *Electrophoresis* **16** (1995) 1077 – 1326.
104. M. J. Dunn (ed.): From Protein Maps to Genomes, Proceedings of the Second Siena Two-Dimensional Electrophoresis Meeting, *Electrophoresis* **18** (1997) 305 – 662.
105. M. J. Dunn (ed.): From Genome to Proteome: Proceedings of the Third Siena Two-Dimensional Electrophoresis Meeting, *Electrophoresis* **20** (1999) 643 – 845.
106. F. Lottspeich (ed.): Paper Symposium: Electrophoresis and Amino Acid Sequencing, *Electrophoresis* **17** (1996) 811 – 966.
107. B. Tümmler (ed.): Microbial Genomes: Biology and Technology, *Electrophoresis* **19** (1998) 467 – 624.
108. K. L. Williams (ed.): Strategies in Proteome Research, *Electrophoresis* **19** (1998) 1853 – 2050.
109. I. Humphrey-Smith (ed.): Paper Symposium: Microbial Proteomes, *Electrophoresis* **18** (1997) 1207 – 1497.
110. R. D. Appel, M. J. Dunn, D. F. Hochstrasser (eds): Paper Symposium: Biomedicin and Bioinformatics, *Electrophoresis* **18** (1997) 2703 – 2842.
111. R. D. Appel, M. J. Dunn, D. F. Hochstrasser (eds): Paper Symposium: Biomedicin and Bioinformatics, *Electrophoresis* **20** (1999) 3481 – 3686.
112. P. Cash (ed.): Paper Symposium: Microbial Proteomes, *Electrophoresis* **20** (1999) 2149 – 2285.
113. M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser (eds.): *Proteome Research: New Frontiers in Functional Genomics,* Springer, Berlin 1997.
114. S. M. Hanash, in B. D. Hames (ed.): *Gel Electrophoresis of Proteins: a Practical Approach,* Oxford University Press, Oxford 1998, pp. 189 – 211.
115. F. Kohlrausch, *Ann. Phys. Chem.* **62** (1897) 209 – 234.
116. F. M. Everaerts, J. L. Becker, T. P. E. M. Verheggen: *Isotachophoresis: Theory, Instrumentation and Applications,* Elsevier, Amsterdam 1976.
117. P. Gebauer, V. Dolnik, M. Deml, P. Bocek, *Adv. Electrophor.* **1** (1987) 281 – 359.
118. S. G. Hjalmarsson, A. Baldesten, *CRC Crit. Rev. Anal. Chem.* **18** (1981) 261 – 352.
119. L. Krivankovà, P. Bocek, in M. G. Khaledi (ed.): *High Performance Capillary Electrophoresis,* John Wiley & Sons, New York 1998, pp. 251 – 275.
120. C. B. Laurell, *Anal. Biochem.* **15** (1966) 45 – 52.
121. H. G. M. Clarke, T. Freeman, *Protides Biol. Fluids* **14** (1967) 503 – 509.
122. I. Sondergaard, L. K. Poulsen, M. Hagerup, K. Conradsen, *Anal. Biochem.* **165** (1987) 384 – 391.
123. J. Kroll in N. H. Axelsen, J. Kroll, B. Weeke (eds.): *A Manual of Quantitative Immunoelectrophoresis,* Universitetsforlaget, Oslo 1973, pp. 61 – 67.
124. P. J. Svendsen, N. H. Axelsen, *J. Immunol. Methods* **2** (1972) 169 – 176.
125. P. J. Svendsen, C. Rose, *Sci. Tools* **17** (1970) 13 – 17.
126. N. H. Axelsen, J. Kroll, B. Weeke (eds.): "A Manual of Quantitative Immunoelectrophoresis," *Scand. J. Immunol. Suppl.* **1** (1973) no. 2, pp. 1 – 169.
127. N. H. Axelsen (ed.): "Quantitative Immunoelectrophoresis," *Scand. J. Immunol. Suppl.* **2** (1975) pp. 1 – 230.
128. N. H. Axelsen (ed.), *Scand. J. Immunol. Suppl.* **10** (1983) no. 17, pp. 1 – 280.
129. C. R. Merril, R. C. Switzer, M. L. Van Keuren, *Proc. Natl. Acad. Sci. U.S.A.* **76** (1979) 4335 – 4339.
130. M. Moeremans, G. Daneels, J. De Mey, *Anal. Biochem.* **145** (1985) 315 – 321.
131. M. Moeremans, D. De Raeymaeker, G. Daneels, J. De Mey, *Anal. Biochem.* **153** (1986) 18 – 22.
132. E. M. Southern, *J. Mol. Biol.* **98** (1975) 503 – 510.
133. H. Towbin, T. Staehelin, J. Gordon. *Proc. Natl. Acad. Sci. U.S.A.* **76** (1979) 4350 – 4355.
134. H. Towbin, J. Gordon, *J. Immunol. Methods* **72** (1984) 313 – 340.
135. O. J. Bjerrum, N. H. H. Heegaard, *J. Chromatogr.* **470** (1989) 351 – 367.
136. V. Neuhoff, R. Stamm, H. Eibl, *Electrophoresis* **6** (1985) 427 – 448.
137. C. R. Merril, K. M. Washart, in: *Gel Electrophoresis of Proteins, a Practical Approach,* Oxford University Press, Oxford 1998, pp. 53 – 92, 319 – 343.
138. B. Bjellqvist, K. Ek, P. G. Righetti, E. Gianazza, A. Görg, W. Postel, R. Westermeier, *J. Biochem. Biophys. Methods* **6** (1982) 317 – 339.

139. M. Chiari, E. Casale, E. Santaniello, P. G. Righetti, *Appl. Theor. Electrophor.* **1** (1989) 99 – 102; **ibid. 1** (1989) 103 – 107.
140. E. Gianazza, F. Celentano, G. Dossi, B. Bjellqvist, P. G. Righetti, *Electrophoresis* **5** (1984) 88 – 97.
141. G. Dossi, F. Celentano, E. Gianazza, P. G. Righetti, *J. Biochem. Biophys. Methods* **7** (1983) 123 – 142.
142. P. G. Righetti, M. Chiari, P. K. Sinha, E. Santaniello, *J. Biochem. Biophys. Methods* **16** (1988) 185 –192.
143. C. Gelfi, M. L. Bossi, B. Bjellqvist, P. G. Righetti, *J. Biochem. Biophys. Methods* **15** (1987) 41 – 48; P. K. Sinha, P. G. Righetti, **ibid. 15** (1987) 199 – 206.
144. P. G. Righetti, E. Gianazza, C. Gelfi, M. Chiari, P. K. Sinha, *Anal. Chem.* **61** (1989) 1602 – 1612.
145. P. G. Righetti, C. Gelfi, M. Chiari, in B. L. Karger, W. S. Hancock (eds.): *Methods in Enzymology: High Resolution Separation and Analysis of Biological Macromolecules, Part A: Fundamentals,* **vol. 270,** Academic Press, San Diego 1996, pp. 235 – 255.
146. P. G. Righetti, A. Bossi, *Anal. Biochem.* **247** (1997) 1 – 10.
147. P. G. Righetti, A. Bossi, *J. Chromatogr. B* **699** (1997) 77 – 89.
148. P. G. Righetti, A. Bossi, C. Gelfi, in B. D. Hames (ed.): *Gel Electrophoresis of Proteins,* 3rd ed., Oxford University Press, Oxford 1998, pp. 127 – 187.
149. J. W. Jorgenson in J. W. Jorgenson, M. Phillips (eds.): "New Directions in Electrophoretic Methods," *ACS Symp. Ser. 335,* Am. Chem. Soc., Washington, 1987, pp. 70 – 93.
150. D. J. Rose, Jr., J. W. Jorgenson, *J. Chromatogr.* **447** (1988) 117 – 131.
151. F. Foret, P. Bocek, *Advances Electrophor.* **3** (1989) 271 – 347.
152. B. L. Karger (ed.): "Proceedings of the 1st International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr.* **480** (1989) 1 – 435.
153. B. L. Karger (ed.): "Proceedings of the 2nd International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr.* **516** (1990) 1 –298.
154. J. W. Jorgenson (ed.): "Proceedings of the 3rd International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr.* **559** (1991) 1 –561.
155. F. M. Everaerts, T. P. E. M. Verheggen (eds.): "Proceedings of the 4th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr.* **608** (1992) 1 – 429.
156. B. L. Karger (ed.): "Proceedings of the 5th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **652** (1993) 1 – 574.
157. B. L. Karger, S. Terabe (eds.): "Proceedings of the 6th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **680** (1994) 1 – 689.
158. H. Engelhardt (ed.): "Proceedings of the 7th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **716** (1995) 1 – 412; **717** (1995) 1 – 431.
159. B. L. Karger (ed.): "Proceedings of the 8th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **744** (1996) 1 – 354; **745** (1996) 1 – 303.
160. W. S. Hancock (ed.): "Proceedings of the 9th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **781** (1997) 1 – 568.
161. S. Fanali, B. L. Karger (eds.): "Proceedings of the 10th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **817** (1998) 1 – 382.
162. E. S. Yeung (ed.): "Proceedings of the 11th International Symposium on High-Performance Capillary Electrophoresis," *J. Chromatogr. A* **853** (1999) 1 – 576.
163. A. M. Krstulovic (ed.): "Capillary Electrophoresis in the Life Sciences," *J. Chromatogr. B* **697** (1997) 1 – 290.
164. Y. Baba (ed.): "Analysis of DNA by Capillary Electrophoresis," *J. Biochem. Biophys. Methods* **41** (1999) 75 – 165.
165. B. L. Karger (ed.): Paper Symposium: Capillary Electrophoresis, *Electrophoresis* **14** (1993) 373 – 558.
166. N. J. Dovichi (ed.): Paper Symposium: Nucleic Acid Electrophoresis, *Electrophoresis* **17** (1996) 1407 – 1517.
167. F. Foret, P. Bocek (eds.): Paper Symposium: Capillary Electrophoresis: Instrumentation and Methodology, *Electrophoresis* **17** (1996) 1801 – 1963.
168. J. P. Landers (ed.): Paper Symposium: Capillary Electrophoresis in the Clinical Sciences, *Electrophoresis* **18** (1997) 1709 – 1905.

169. Z. El Rassi (ed.): "Capillary Electrophoresis and Electrochromatography," *Electrophoresis* **18** (1997) 2123 – 2501.
170. B. R. McCord (ed.): Paper Symposium: Capillary Electrophoresis in Forensic Science, *Electrophoresis* **19** (1998) 3 – 124.
171. Z. El Rassi (ed.): Paper Symposium: Capillary Electrophoresis and Electrochromatography Reviews, *Electrophoresis* **20** (1999) 2989 – 3328.
172. P. G. Righetti (ed.): *Capillary Electrophoresis: an Analytical Tool in Biotechnology,* CRC Press, Boca Raton, FL 1996.
173. M. G. Khaledi (ed.): *High Performance Capillary Electrophoresis: Theory, Techniques and Applications,* John Wiley & Sons, New York 1998.
174. P. G. Righetti, C. Gelfi, *J. Biochem. Biophys. Methods* **41** (1999) 75 – 90.
175. P. G. Righetti, A. Bossi, E. Olivieri, C. Gelfi, *J. Biochem. Biophys. Methods* **40** (1999) 1 – 16.
176. O. Vesterberg, T. Waldström, K. Vesterberg, H. Svensson, B. Malmgren, *Biochim. Biophys. Acta* **133** (1967) 435 – 445.
177. B. J. Radola, *Ann. N.Y. Acad. Sci.* **209** (1973) 127 –143.
178. B. J. Radola, *Biochim. Biophys. Acta* **295** (1973) 412 –428.
179. M. Bier, N. Egen in H. Haglund, J. C. Westerfeld, J. T. Ball Jr. (eds.): *Electrofocus '78,* Elsevier, Amsterdam 1979, pp. 35 – 48.
180. M. Bier, N. Egen, T. T. Allgyer, G. E. Twitty, R. A. Mosher in E. Gross, J. Meienhofer (eds.): *Peptides: Structure and Biological Function,* Pierce Chem. Co., Rockford 1979, pp. 79 – 89.
181. N. Egen. W. Thormann, G. E. Twitty, M. Bier in H. Hirai (ed.): *Electrophoresis '83,* De Gruyter, Berlin 1984, pp. 547 – 550.
182. M. Bier, T. Long, *J. Chromatogr.* **604** (1992) 73 –83.
183. P. G. Righetti, E. Wenisch, M. Faupel, *J. Chromatogr.* **475** (1989) 293 – 309.
184. P. G. Righetti, E. Wenisch, A. Jungbauer, H. Katinger, M. Faupel, *J. Chromatogr.* **500** (1990) 681 –696.
185. P. G. Righetti, E. Wenisch, M. Faupel, *Advances Electrophor.* **5** (1992) 159 – 200.
186. P. Wenger, M. de Zuanni, P. Javet, P. G. Righetti, *J. Biochem. Biophys. Methods* **14** (1987) 29 – 43.

Electrophoretic Coating → Paints and Coatings
Electrophotography → Imaging Technology
Electrostatic Sealing → Glass

Structure Analysis by Diffraction

ERICH F. PAULUS, Institute of Mineralogy and Crystallography of the University of Frankfurt, Frankfurt/Main, Federal Republic of Germany

ALFRED GIEREN, Institute for Analytical Chemistry and Radiochemistry, University of Innsbruck, Innsbruck, Austria

1. General Principles 2
2. Structure Analysis of Solids 2
2.1. Diffraction by Crystal Lattices . . . 2
2.1.1. Materials and Methods 2
2.1.2. Introduction to X-Ray Crystallography; History 5
2.1.3. Experimental Principles, Applications 7
2.1.4. Crystal Growth for X-Ray Structure Analysis 10
2.1.5. X-Rays, Neutrons, and Electrons . . . 11
2.1.6. Cameras and Diffractometers 13
2.1.7. Safety 21
2.1.8. Instrument Manufacturers 22
2.2. The Phase Problem 23
2.2.1. Patterson (Vector) Methods 24
2.2.2. Direct Methods 26
2.2.3. Trial-and-Error Method, R Test 29
2.2.4. Experimental Phase Determination . . 29
2.3. Least-Squares Refinement 29
2.4. Determination of Absolute Configuration 30
2.5. Example of a Single-Crystal X-Ray Structure Analysis Solved by Direct Methods 30
2.6. Diffraction by Polycrystalline Specimens 34
2.6.1. General Considerations 34
2.6.2. Qualitative and Quantitative Analyses of Powder Specimens 36
2.6.3. Solid Solutions 38
2.6.4. Rietveld Method 38
2.6.5. Texture 39
2.7. Crystallographic Databanks 40
2.8. Noncrystallinity 40
2.8.1. Crystal Quality Analysis 40
2.8.2. Crystallite Size 41
2.8.3. Defects 41
2.8.4. X-Ray Crystallinity 41
2.8.5. Elastic Stress 42
2.8.6. Radial Electron Density Distribution . 42
3. Synchrotron Radiation 43
4. Neutron Diffraction 43
5. Electron Diffraction 44
5.1. Diffraction in the Electron Microscope 44
5.2. Electron Diffraction in the Gas Phase 44
6. Future Developments 44
7. References 45

1. General Principles

When electromagnetic or corpuscular radiation is diffracted by matter, i.e., when it is involved in an elastic interaction with matter, a scattering diagram or diffraction pattern is produced; this pattern is the Fourier transform of the diffracting object. If the inverse transform is applied, going from "reciprocal" to "direct" space, an image of the object is obtained. This inverse transformation is called Fourier synthesis; the transformation from direct to reciprocal space is called Fourier analysis. The principle of Fourier analysis (diffraction) followed by Fourier synthesis (imaging) controls the classical process of magnification in an optical microscope. Here, an object is irradiated, and elastic scattering takes place. The objective lens causes the scattered rays to interfere, generating a magnified image. To apply this principle at atomic resolution (ca. 10^{-10} m), the first thing needed is radiation with wavelength of the same order of magnitude (maximum resolution is $\lambda/2$) [55]. A further condition is that the destruction caused by the interaction of the radiation with the object should be as slight as possible, so that the image is not substantially degraded by radiation damage.

The chief kinds of radiation used are X-rays, neutrons, and electrons. X-rays and neutrons cause relatively little damage, but electrons have a scattering power several orders of magnitude greater, because of their charge, which leads to

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_341

more radiation damage. Electrons have the advantage that they can be imaged with lenses in a microscope configuration. The use of a lens for imaging (Fourier synthesis) is not, however, essential. Rapid advances in electronic computers, make it possible to replace the lens by a computer. However, this is possible only if all the relevant parameters of the scattering diagram, including the phases of the scattered waves, are known. As a rule, the phases of X-ray and neutron radiation are not accessible to direct measurement, so that computer-intensive techniques are required to obtain them.

X-ray and neutron structure analysis generally deal with crystalline matter, either single crystals or crystalline powders. Nonperiodic objects, such as glasses, play only a secondary role.

2. Structure Analysis of Solids

2.1. Diffraction by Crystal Lattices

With a few exceptions, this article deals with the crystalline solid state; other aggregations such as liquid crystals are not covered.

Naturally occurring crystals differ from ideal crystals in two main aspects:

1) They are finite with surfaces
2) They contain structural defects and vacancies, e.g., the mosaic structure of real crystals

The crystals available in practice thus lie somewhere between amorphous and ideally crystalline, but nearer the ideal end of the scale ("ideal imperfect crystals"). A crystalline specimen may be a single crystal (smallest dimension ca. 0.05 – 0.5 mm) or a polycrystalline material with crystallite size ca. 0.04 mm.

High-polymer plastics occupy an intermediate position. According to the two-phase model [56] the amorphous phase is present along with the crystalline phase. The unit cells in the crystalline portion may be oriented in some preferred direction (as in fibers), or there may be no preferred direction.

2.1.1. Materials and Methods

"Single crystals" are the classical objects investigated by crystallography. The diffraction of X-rays by crystals was discovered in 1912 by VON LAUE, FRIEDRICH, and KNIPPING in a study of copper sulfate pentahydrate crystals. It was later found that diffraction also takes place from crystalline powders. The methodological principles are explained here for single crystals and to a certain extent also for crystalline powder.

Single-crystal structure analysis, the determination of the structure of the asymmetric unit of a crystal, is so far the only analytical method that can yield direct and unambiguous information about the three-dimensional atomic structure of a substance. Crystalline powder structure analysis needs in the average also a model.

Simple Inorganic Compounds. The class of simple inorganic compounds includes substances such as metals and salts, that contain very few atoms in the asymmetric unit. In such cases, the complete crystal structure can often be derived as soon as the crystal symmetry (space group) has been determined. If several possibilities exist, the correct structure can be arrived at by trial and error. When this approach also fails, a Patterson synthesis (see Eq. 27) can be used to solve the structure.

Complex Inorganic and Organic Compounds. Two methods can be used for the solution of a more complex molecular and crystal structure problem:

1) Patterson synthesis followed by successive Fourier syntheses of the electron density
2) "Direct" determination of the phases of the structure factors, possibly followed by successive Fourier syntheses of the electron density (direct methods)

These methods complement each other in many cases. When there are several possible solutions for the Patterson function, some may be eliminated if the direct methods yield a possible solution by which the Patterson synthesis can be interpreted. The contrary can also occur, when a Patterson synthesis is used to decide which phase set is correct.

A prerequisite for the use of Patterson synthesis is that concrete assumptions about the complete molecular structure exist, or at least that parts of the structure are known. The position

and orientation of the known molecular part can then be determined, for example with the aid of the convolution molecule method [57] or minimum search programs. In the case of heavy-atom derivatives, the known molecular part consists of one or several heavy atoms [58].

Direct methods are most commonly used to solve crystal structures. A number of advanced program systems exist and, in many cases, make solution of the phase problem a routine matter for the expert. The proposed structure, which often proves to be correct and complete, can be output directly in graphical form (Fig. 1). The numbers represent atomic positions; a small numerical value indicates a larger electron density maximum and a large value indicates a smaller one. The program determines which electron density maxima should be connected to each other to obtain the structure; tables are generated that show the positions of the atoms in the cluster (see also Table 3). However, if this method does not succeed immediately, generally much experience, including an understanding of the principles of direct methods, is required to solve the phase problem. Given sufficient time, virtually all structure problems (ca. 95 %) satisfying certain minima criteria are solvable; these include structures with up to 100 and many more nonhydrogen atoms in the asymmetric unit (the atomic cluster from which the entire unit cell can be generated by symmetry operations), and correspond to those structures of greatest interest to the synthetic chemist. If the data set is improved by cooling the crystal (below − 100 °C) to reduce the thermal vibrations of the atoms, this limit can be extended in special cases to more than 300.

Proteins. In these much larger molecules, it is often possible to determine the phases by the method of isomorphous replacement. Most commonly, this means allowing ions of heavy atoms to diffuse into the solvent-filled vacancies in protein crystals, where they adopt well-defined positions. Thus, this represents the incorporation of atoms in completely new positions, rather than an atom-replacement process. One exception is insulin, in which zinc ions are replaced by lead or mercury ions. Newer methods allow the structure to be solved by direct methods when excellent protein crystals are available.

A number of techniques have been devised for determining the positions of heavy atoms in the unit cells of proteins [6, pp. 337 – 371], but these are not discussed here.

In the noncentrosymmetric case, which always holds for proteins, phase determination of the structure factors requires at least two heavy-atom derivatives to estimate the phases for the native protein from the heavy atom positions [6, pp. 337 – 371]. As a rule, three or four derivatives are investigated in order to reduce phase errors.

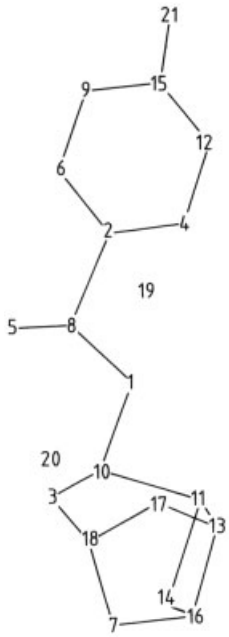

Figure 1. Computer-generated molecular structure diagram after solution of a crystal structure by direct methods

All reflections whose phases are to be determined have to be measured on the native protein and the isomorphous crystals. A computer-controlled diffractometer is used. Alternatively, photographic data and automatic photometers can be employed; this has the advantage that the crystals do not have to be exposed to X-rays for such a long time, reducing radiation damage. More recently, films have been replaced by area detectors which produce the intensities in digital form. If a synchrotron is used as radiation source, and the pattern is recorded by the Laue method (see Section 2.1.5), data acquisition takes only a few minutes.

The extent to which the Fourier synthesis of the electron density can be interpreted depends on the resolution with which the data could be measured. Resolution is said to be low when the smallest lattice plane spacings to which measured reflections can be assigned are $> 5 \times 10^{-10}$ m; aside from the dimensions and overall shape of the molecule, no detail can be extracted from the Fourier synthesis. At a resolution of 3×10^{-10} m, the shape of polypeptide chains can be identified; at 1.5×10^{-10} m, single atoms can be distinguished and side chains can be identified without knowledge of the amino acid sequence. The atomic positions cannot be refined by conventional least-squares methods, because the computational problem becomes immense and the number of observed data seldom exceeds the number of parameters. For special methods, see [6, pp. 420 – 442].

Supplementary Structure Investigations by Neutron Diffraction. Single-crystal X-ray structure analysis cannot determine the positions of hydrogen atoms in many cases. Even with error-free measurements, there are essentially two reasons:

1) Because hydrogen atoms generally occupy peripheral locations in the molecule, they are smeared out by static and/or dynamic disorder (small conformational differences), and their scattering contribution at higher scattering angles is not significantly higher than the background.
2) The compound under study may contain not only light atoms, but also heavy atoms, such as bromine and iodine. The intensity measurements are not accurate enough to detect the small relative intensity differences due to hydrogen atoms.

In most such cases, well-known stereochemical rules make it possible to calculate the positions of most hydrogen atoms from those of the other atoms. But if one wishes to investigate, say, hydrogen bonds more closely, the use of neutron diffraction is indicated, because H or, better, D atoms (because of the smaller incoherent scattering) are strong neutron scatterers.

Neutron diffraction analysis may also be necessary when elements near one another in the periodic table need to be distinguished. A well-known example is differentiation between the isoelectronic ions Si^{4+} and Al^{3+}.

By virtue of their magnetic moment, neutrons can also be used to investigate magnetic structures [5, pp. 395 – 511], [59].

Chemical bonds cause a displacement of the bonding electrons. Because X-rays are also scattered by these electrons, a slight distortion of the apparent atomic positions is unavoidable. For example, X-ray diffraction always gives carbon –hydrogen bond lengths that are too short, because the hydrogen electron is pulled closer to its bonding partner. Neutron diffraction analysis, in contrast, determines the position of the atomic nucleus, not the maximum of the electron density distribution. The deformation density resulting from chemical bonding may be directly revealed by difference Fourier syntheses of electron densities from X-ray diffraction data, and the structure factors calculated from the atomic positions obtained by neutron diffraction analysis (X – N method).

The measurements must be performed with the utmost care, because the effects are comparable to the standard deviations. X-ray and neutron diffraction patterns should be recorded at the lowest possible temperatures, to minimize thermal motion.

For reasons of cost, neutron diffraction is virtually always preceded by single-crystal X-ray structure analysis. In many cases, the deformation densities can be determined by X-ray measurements alone, in view of the fact that the scattering contribution of the bonding electrons vanishes at large scattering angles (X – X method).

2.1.2. Introduction to X-Ray Crystallography; History

Democritos of Abdera (born ca. 465 B.C.) wrote, "It is customary to say that there is color, sweetness, bitterness; in actuality, however, there are atoms and void" [60]. In this sentence, the atom, as the fundamental building block of matter, was introduced into the discussion of the structure of matter. In 1807, Dalton had only to invoke this hypothesis in order to impose order on the experimental facts known at the time. For example, a crystalline structure consisting of identical particles was also in agreement with these principles.

NIKOLAUS STENO, as early as 1669, formulated the law of constant angles, which states that two like edges of crystals of the same substance always make equal angles with each other. In 1823, FRANZ ERNST NEUMANN discovered the law of simple rational indices. Planes drawn through faces of a crystal intersect three coordinate axes, constructed parallel to crystal edges, at minima lengths a, b, c that are in a constant ratio to one another. The other faces of the crystal can be characterized by planes drawn through them that cut the axes at lengths a_i , b_i , c_i which are in whole-number (possibly negative) ratios to the intercepts a, b, c ($a_i : b_j : c_i = u \cdot a : v \cdot b : w \cdot c$) (Fig. 2).

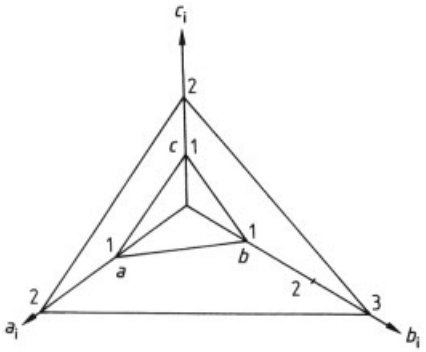

Figure 2. The law of simple rational indices
Planes shown are: $u = v = w = 1$; and $u = 2$, $v = 3$, $w = 2$

About 1890, FEDOROV, SCHOENFLIES, and BARLOW more or less simultaneously showed that for crystals built up from discrete particles in a three-dimensionally ordered manner, there can be no more than 230 different combinations of elements, the 230 space groups. In 1912, on the basis of their key diffraction experiment, which yielded discrete X-ray reflections from a crystal, LAUE, FRIEDRICH, and KNIPPING demonstrated both the periodic construction of crystals from atoms and the wave nature of X-rays.

W. H. and W. L. BRAGG were the first to show that the models proposed by BARLOW for simple compounds, such as NaCl, CsCl, and ZnS are in agreement with the scattered X-ray intensities. PAUL NIGGLI found that space groups could be determined by X-ray methods via the systematic extinction laws.

The Bragg equation:

$$\sin\theta = n\frac{\lambda}{2d} \tag{1}$$

$$n\lambda = 2d\sin\theta \tag{2}$$

describes the diffraction of X-rays by a crystal as reflection from the net planes. Constructive interference occurs only if the difference in path length between two rays is an integer multiple of the wavelength λ (Fig. 3). Here θ is the glancing angle ($2\,\theta$ is the diffraction angle), and d is the net plane spacing.

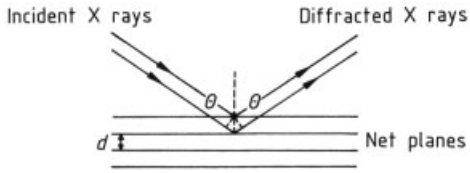

Figure 3. Derivation of the Bragg equation

A construction due to EWALD illustrates the importance of the reciprocal lattice in X-ray crystallography. As Figure 4 shows, the Bragg equation is satisfied where the reflection sphere is cut by a lattice point of the reciprocal lattice constructed around the center of the crystal. Rotating the crystal together with the reciprocal lattice around a few different directions in the crystal fulfills the reflection condition for all points of the reciprocal space within the limiting sphere. The reciprocal lattice vector $\boldsymbol{S}$ is perpendicular to the set of net planes, and has the absolute magnitude $1/d$. In vector notation:

$$\boldsymbol{S} = \boldsymbol{s} - \boldsymbol{s}_0 \tag{3}$$

Vectors $\boldsymbol{s}$ and $\boldsymbol{s}_0$ each have modulus $1/\lambda$.

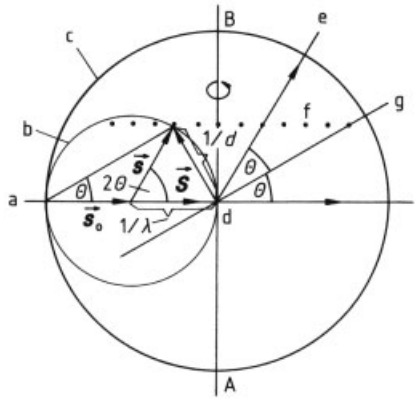

Figure 4. Ewald construction
a) Incident beam; b) Reflection sphere; c) Limiting sphere; d) Crystal; e) Diffracted beam; f) Lattice points on a plane of the reciprocal lattice perpendicular to the plane of the diagram; g) Reflecting lattice plane

The reciprocal lattice provides a simple way to illustrate the positions of diffraction points in space. Figure 4 also shows that the Ewald limiting sphere determines how many reflections can be measured. All reflections inside the limiting sphere with radius $2/\lambda$ are obtained only if the crystal axis AB can take up any orientation. This is not so in many data acquisition methods, e.g., the rotating crystal method. If the crystal is rotated about AB, reciprocal lattice points lying on reciprocal lattice planes perpendicular to AB follow circular paths, cutting the reflection sphere. Every such cut corresponds to a reflection.

The portion of reciprocal space inside the Ewald limiting sphere that can be measured with a single crystal setting is shown in Figure 5. Almost all lattice points of the limiting sphere are registered when measurements are done with two crystal settings about axes forming an angle ca. 90°.

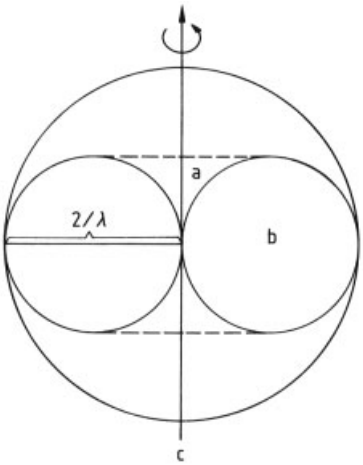

Figure 5. The region of reciprocal space that can be measured for a single setting of a rotating crystal
a) Blind spot; b) Toroidal region of reciprocal space covered for a single setting; c) Crystal axis AB

It is generally true that a vector of the reciprocal lattice is perpendicular to a plane in the direct lattice and, conversely, a vector of the direct lattice is perpendicular to a plane of the reciprocal lattice.

The deeper connection between the real and reciprocal lattices becomes clear when it is shown (see Eqs. 13, 18) that the absolute magnitudes of the functional values of the Fourier transform of the entire crystal are equal to the square roots of the measured intensities. The position coordinates of these functional values are coordinates in reciprocal space (i.e., the indices of the reflections) if nonzero functional values are allowed only at reciprocal lattice points. A reflection can then be defined completely generally as a reciprocal lattice point. Further details are presented in the textbooks cited at the end of this article.

2.1.3. Experimental Principles, Applications

If suitable single crystals of a compound are available, it is a routine matter to determine the cell constants. The crystalline specimen should not be much smaller than 0.3 – 0.5 mm. A crystal volume of 0.005 mm^3 is adequate for measurements on organic compounds; much smaller crystals can be used in the case of inorganic compounds with strongly scattering atoms. Air-sensitive substances can be sealed in borosilicate capillaries.

The following method has rapidly gained acceptance where it is possible or desirable to measure the crystals at low temperature (ca. − 100 °C). The crystalline specimen is embedded in an oil that is liquid at the embedding temperature, but solidifies at low temperature [61]. Measurements can be performed on extremely air-sensitive substances, and on those that are liquid at room temperature. The method has the drawback that the crystal employed for data collection is normally not recoverable at the end of the measurement.

Before the introduction of laboratory computers, virtually the only way of determining cell constants was the photographic method. With the Weissenberg and precession cameras, as well as the Space Explorer, a "mixture" of cell constants in both reciprocal and direct space is obtained. The relationships between these cell constants are expressed by the following equations:

$$a = \frac{1}{V^*} b^* c^* \sin\alpha^* \tag{4}$$

$$b = \frac{1}{V^*} c^* a^* \sin\beta^* \tag{5}$$

$$c = \frac{1}{V^*} a^* b^* \sin\gamma^* \tag{6}$$

$$\cos\alpha = (\cos\beta^* \cos\gamma^* - \cos\alpha^*) \,/\, (\sin\beta^* \sin\gamma^*) \tag{7}$$

$$\cos\beta = (\cos\alpha^* \cos\gamma^* - \cos\beta^*) \,/\, (\sin\alpha^* \sin\gamma^*) \tag{8}$$

$$\cos\gamma = (\cos\alpha^* \cos\beta^* - \cos\gamma^*) / (\sin\alpha^* \sin\beta^*) \quad (9)$$

$$V^* = a^* b^* c^* \cdot \sqrt{1 + 2\cos\alpha^* \cos\beta^* \cos\gamma^* - \cos^2\alpha^* - \cos^2\beta^* - \cos^2\gamma^*} \quad (10)$$

Symbols with asterisks relate to the reciprocal lattice; a, b, c are lengths, α, β, γ are angles, and V is the cell volume. If the left-hand sides of these equations are replaced by the cell constants of the reciprocal lattice, the right-hand sides contain the corresponding lattice values in direct space.

The determination of cell constants is particularly convenient with the precession camera, in which the crystal has to be set on a goniometer head; one reciprocal lattice axis must lie parallel to the goniometer head axis. Two settings of the goniometer spindle yield two principal planes in the reciprocal lattice. The third angle of the reciprocal lattice, obtained in terms of the corresponding angle in direct space, is the difference between the two spindle settings. A photographic cone record (see also Fig. 25) represents a complete projection of the reciprocal lattice; this avoids mistakes in the choice of the cell dimensions due to missing planes of the reciprocal lattice. The first two layer lines supply the remainder of the data needed for space group determination.

Equations 4 – 10 make it possible to calculate the cell constants in direct and reciprocal space. From the symmetries present and the systematic extinction of reflections, the crystal system, the lattice type (Bravais lattice), and the space group can be derived [4, pp. 124 – 159]. Because the diffraction pattern always has a center of symmetry (Friedel's law, which holds when anomalous scattering is neglected), ambiguities in the space group determination are still possible (11 Laue groups instead of 32 crystal classes). In principle, the space group is not definite until the structure has been completely solved, although in most cases there is a high prior probability for a particular space group.

If the slightest doubt exists whether the crystal system is right, no time should be lost in carrying out a cell reduction by the Azaroff – Buerger method [4]; many computer programs are on the market for this operation, but it is also reasonably easy to perform "manually".

If a computer-controlled single-crystal diffractometer is available, the cell constants, and space group can be determined by the computer. Three methods exist, differing only in the means of determining the reflection coordinates in reciprocal space. In "peak-hunting" programs, the region of reciprocal space within specified boundaries is systematically searched for reflections [62]; the Acedeva program makes a "systematically unsystematic" search for reflections [63], with a random number generator determining the directions in which to drive the diffractometer. Normally, either of these two methods yields a number of usable reflections in reciprocal space, from which the three shortest noncoplanar vectors can be extracted and used to index all the reflections. Finally, the reduced cell and the Niggli matrix are calculated, and thus the crystal system and Bravais lattice can be derived [4]. The third method employs an oscillation photograph, made on the diffractometer by using a polaroid film, from which two-dimensional coordinates of selected reflections are obtained; the search for about ten reflections is based on these coordinates [64].

A diffractometer which performs only point measurements, does not give the cell constants unambiguously, since it deals only with discrete points in the reciprocal space. "Continuous" measurement in the reciprocal space is necessary to make certain that no essential points for unit cell determination have been missed. This fact, together with pedagogical considerations, is the reason that film techniques are still used in the X-ray diffraction laboratory. A modern four-circle diffractometer is equipped to perform the necessary photographic measurements which a three-circle diffractometer is unable to do.

Electron diffraction studies on tiny crystals ($6 \leq 10^{-8}$ m) are possible in the electron microscope. The resulting single-crystal diffraction patterns can be indexed similarly to precession photographs (Fig. 6), provided multiple orientations of the same specimen relative to the electron beam can be set up. The results should always be combined with those of X-ray powder patterns. This technique has the advantage that it does not require the laborious growth of single crystals, but its greatest disadvantage is that the intensities are often totally falsified (this is, however, of no importance for unit cell determination).

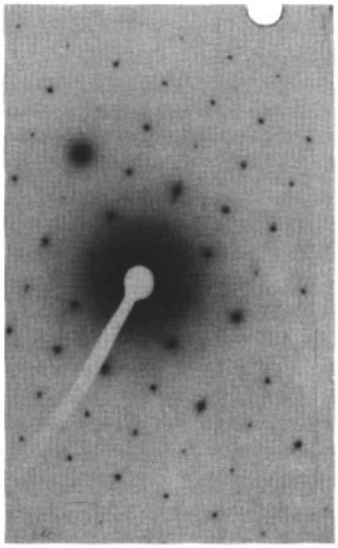

Figure 6. Part of the electron diffraction pattern of a single crystal (MoO_3)

Up to this point, only the geometry of diffraction has been considered. However, the intensity ratios of the reflections are the basis for determining the structure of the unit cell contents.

For the intensity ratios to be quantified, scattering must be discussed in more detail. Consider the scattering of X-rays by electrons in the atomic electron shells. The scattering power I_e of an electron at point A is given by Maxwell's theory as:

$$I_e = \frac{I_0 e^4}{r^2 m^2 c^4} \sin^2 \alpha \tag{11}$$

where I_0 is the incident X-ray intensity, e is the electron charge, m is the electron mass, r is the distance from the scattering electron to the observer, c is the speed of light, and α is the angle between the electric field vector $\boldsymbol{E}$ and the direction to the observer (Fig. 7).

If the incident X-rays are unpolarized, $\sin^2 \alpha$ must be replaced by the factor $(\frac{1}{2} + \frac{1}{2} \cos^2 2\theta)$, the polarization factor. Here, 2θ is the deflection angle.

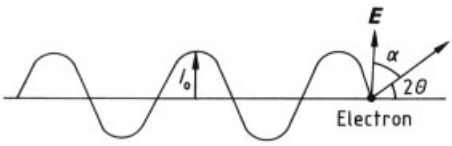

Figure 7. Scattering of X-rays by an electron

X-rays employed in structure analysis normally have wavelengths $0.5 - 2.3 \times 10^{-10}$ m, the same order of magnitude as the size of atoms and their electron shells. Only at a scattering angle of 0° can e and m in Equation (11) be replaced by Ze and Zm, for an atom with atomic number Z. At other angles, the X-rays scattered by the electrons interfere, and the total intensity is reduced as a function of scattering angle. The loss of intensity is greater, the higher the glancing angle θ and the larger the electron shell are.

The best-known example demonstrating the effect of atomic radius is the case of O^{2-} and Si^{4+} (important in the analysis of silicates). As θ increases, the scattered intensity falls off much faster for O^{2-} than it does for the smaller Si^{4+}. In this way, it is possible to distinguish between these two atomic species (Fig. 8). Atomic form factors f for atoms and ions [29, vol. III, pp. 201 – 246] are a function of the scattering power of the corresponding atom, and are expressed in units of a free electron.

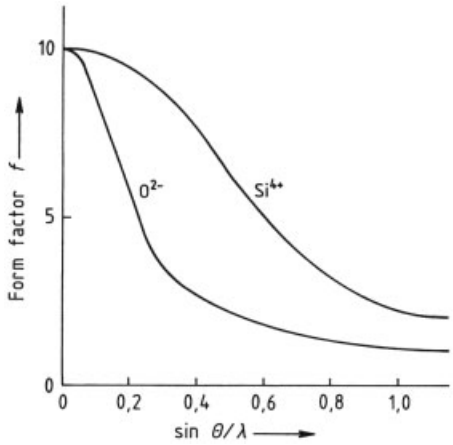

Figure 8. Atomic form factors f for O^{2-} and Si^{4+}

If the frequency of the incident X-rays is close to the frequency of the K absorption edge of the irradiated atom, scattering can no longer be regarded as elastic; some of the incident energy is used to excite the electrons in the atom. The atomic form factors have to be corrected for anomalous dispersion; the tables in [29] include this correction. The significance of these corrections for the determination of the absolute configuration is discussed in Section 2.4.

Among the key properties of electron beams relating to their scattering by atoms, the elec-

tron charge is particularly important. Electrons are scattered by inhomogeneities in the electrostatic potential distribution, chiefly at atomic nuclei. The process is described by the equation:

$$f^{\mathrm{e}} = (Z - f)\,\frac{e^2 m}{2h^2} \cdot \frac{\lambda^2}{\sin^2\theta} \tag{12}$$

where Z is the atomic number, f is the atomic form factor for X-rays, f^{e} is the atomic form factor for electron beams, and e and m are the electron charge and mass, respectively.

Neutrons are scattered only by atomic nuclei (provided there is no magnetic scattering due to unpaired $3\,d$ or $4f$ electrons). The size of the atomic nucleus is orders of magnitude smaller than the wavelength of thermal neutrons, and so no dependence of scattered amplitude on glancing angle can be observed.

The X-ray intensities I of reflections produced by lattice planes having Miller indices (hkl) are measured. These are related in the following way to the absolute magnitudes of the structure factors — F —, i.e., the structure amplitudes:

$$|F\,(hkl)| = k\,\sqrt{I\,(hkl)} \tag{13}$$

Similarly to Equation (11), the following can be derived: Given a crystal of volume V, z unit cells per unit volume, X-rays with wavelength λ, and rotation of the crystal through the reflection position at rotational frequency ω; the intensity I of a reflection having structure factor F is:

$$I = \frac{e^4}{m^2 c^4}\, I_0\, |F|^2\, \frac{\lambda^3 z^2 V}{\omega}\, \frac{1 + \cos^2 2\theta}{2}\, \frac{1}{\sin 2\theta} \tag{14}$$

where c is the speed of light. The last two factors depend on the recording technique, and are commonly combined in the Lorentz-polarization (LP) factor. Strictly speaking, Equation (14) holds only if the coherent interaction between incident and diffracted X-rays is neglected (primary extinction), and if one ignores the fact that the incident X-ray beam is attenuated by diffraction as it propagates through the crystal (secondary extinction), so that not all portions of the crystal contribute equally to diffraction.

All these falsifications become negligibly small if the crystal has an "ideal mosaic" structure, and the "mosaic particles" are smaller than 1 µm. Generally:

$$I \sim |F|^n \tag{15}$$

where n is a number between 1 and 2. If $n = 1$, Equation (14) gives:

$$I = \frac{8}{3\pi}\, \frac{e^2}{mc^2}\, I_0\, |F|\, \frac{\lambda^2 z}{\omega}\, \frac{1 + \left|\cos^2 2\theta\right|}{2 \sin 2\theta} \tag{16}$$

Diffraction is controlled by Equation (14) if the kinematic theory is viewed as completely valid ($n = 2$), and by Equation (16) if the dynamic theory ($n = 1$) is applied. Normally, neither of these extremes is exactly fulfilled, and there is a combination of both. The kinematic theory is usually adopted, with "dynamic" contributions being handled by an extinction correction, although this (like the absorption correction) often proves unnecessary for measurements of standard level.

The structure factor F is needed to determine the electron density ϱ at point x, y, z (in fractional units of the cell constants) and thus to determine the crystal structure:

$$\varrho\,(xyz) = \frac{1}{V} \sum_{h,k,l=-\infty}^{+\infty} F\,(hkl)\, \exp\left[-2\pi i\,(hx + ky + lz)\right] \tag{17}$$

$$F\,(hkl) = \sum_{n=1}^{N} f_n \exp\left[2\pi i\,(hx_n + ky_n + lz_n)\right] \tag{18}$$

The summation in Equation (17) is over all measurable reflections (hkl), and that in Equation (18) is over all atoms N in the unit cell; V is the cell volume; x_n, y_n, z_n are the coordinates of atom n; and f_n is its atomic form factor.

The significance of systematic extinctions for space-group determination can be seen from Equation (18). In space group $\mathrm{P}\,2_1$, for example, for an atom at x, y, z there is another at $-x$, $y + 1/2$, $-z$. The effect on the ($0k0$) reflections is that the contributions of these two atoms to the structure factor differ in phase by 180°. If k is odd they therefore cancel each other, but are additive if k is even. The following equation represents the situation for space group $\mathrm{P}\,2_1$:

$$F\,(0k0) = \sum_{n=1}^{N/2} f_n \left[\exp\,(2\pi\, i\, k y_n) + \exp\left(2\pi i k \left(y_n + \frac{1}{2}\right)\right)\right] \tag{19}$$

Corresponding relations have been tabulated for all space groups [29], [30].

Equations (17) and (18) further show a relationship between real and reciprocal space. The function $F(hkl)$ is the Fourier transform of the unit cell contents, expressed in the reciprocal space coordinates h, k, and l. Because the symmetry operation of translation holds for all three spatial directions in crystals, the Fourier transform of the entire crystal is zero, except at reciprocal lattice points.

2.1.4. Crystal Growth for X-Ray Structure Analysis

While the growing of single crystals from the melt or the gas phase may sometimes be important for industrial purposes (as in the crucible pulling method used for semiconductors such as Si and Ge), in X-ray structure analysis such methods are secondary to crystal growth from solution, and they are employed only when other techniques are not possible.

In principle, suitable crystals can be grown in either of two ways:

1) The first approach uses the temperature dependence of solubility. In accordance with the maxim "like dissolves like", the solvent employed is a compound with the same or similar parent structure, and/or having the same or similar functional groups. Normally, compounds are more soluble at higher temperature. Slow cooling, which can be effected under programmed control, yields crystals suitable for structure analysis in most instances (maximum dimension 0.3 – 0.8 mm). A simple way to bring about slow cooling is to place the solution in a Dewar flask filled with a liquid that can be heated to the boiling point of the solvent without itself vaporizing to any great extent.
2) The second approach depends on raising the concentration. The simplest method is to evaporate the solvent; generally, the more slowly this is done, the better are the crystals. A straightforward way of doing this is to pierce one or more small holes in the stopper of the crystallization flask.

Evaporation is not the only way of raising the concentration. Another possibility is to use the diffusion of two liquids into each other; the liquids should be miscible, and the substance to be crystallized should have greatly different solubilities in them. The diffusion can take place in the liquid or the gas phase. In the first case, a layer of the "weaker" solvent (lower solubility) is formed on top of the solution, and the first crystal nuclei will form at the interface between the two solvents. In the simplest example of diffusion via the gas phase, an open vessel containing the solution with the "stronger" solvent (higher solubility) is placed inside a vessel containing the "weaker", and the whole assembly is hermetically sealed.

In principle, the handling and growing of crystals at low temperature can be done by the same methods, with appropriate modifications [61]. It is even possible to grow a crystal in the measuring instrument itself [65].

2.1.5. X-Rays, Neutrons, and Electrons

X-rays of sufficient intensity for diffraction studies are obtained from an X-ray generator, which essentially consists of two portions, the high-voltage generator and the tube housing. The generator transforms 220 V alternating current to a d.c. voltage of ≤ 60 kV. The total power is generally in the range 3 – 6 kW, but the output of the sealed X-ray tubes (Fig. 9) used for diffraction purposes is preferably 2.5 kW. In general, there is no improvement in luminance with a 3 kW tube; the only difference is that the focus is longer.

In "normal-focus" tubes, the focal spot measures 1×10 mm. If this spot is viewed at a 6° angle in the direction of its longest dimension, the effective spot size is 1×1 mm (point focus); in the perpendicular direction, the effective size is 0.1×10 mm (line focus). A typical X-ray tube has four windows, two with point focus and two with line focus. Specimens extended in two dimensions are irradiated with the line focus; point, sphere, or needle-shaped specimens, with the point focus. Fine-focus tubes (0.4×8 mm) are commonly used for single-crystal studies; these offer markedly higher luminance at the same energy output.

The residual gas pressure in the envelope of an X-ray tube (where the hot cathode and the anode are located) is $< 10^{-1}$ Pa. The appropriate

voltage for the anode material is applied between the cathode and the anode (Table 1). Electrons emitted by the hot cathode are accelerated by the applied voltage, then braked when they strike the anode. If the electrons have sufficient kinetic energy to cause internal ionization of the atoms in the anode, a small part of the energy released (ca. 1 %) is consumed in ejecting electrons from the K shell of the atoms. The vacant energy levels are filled, chiefly by electrons from the L shell (L III and L II levels) and also from the M shell (M III level). The energy thus gained is emitted as X-rays (Fig. 10). The quantum transitions differ in probability, which can be varied to some extent through the operating voltage (see Table 1).

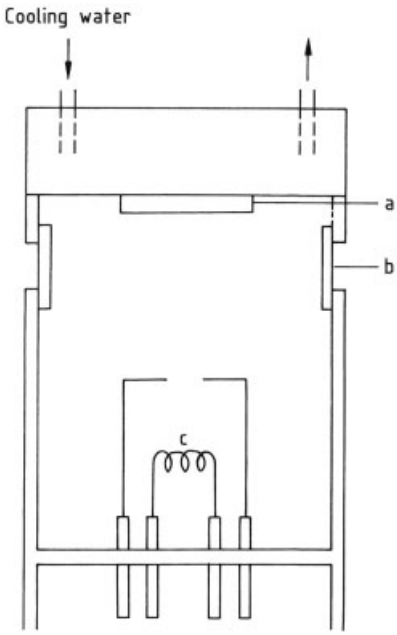

Figure 9. X-ray tube
a) Anode; b) Be window; c) Hot cathode

Given the proper tube voltage, the intensities of the principal components are in the ratio:

$$\mathrm{K}\alpha_1 : \mathrm{K}\alpha_2 : \mathrm{K}\beta_1 = 10 : 5 : 2 \tag{20}$$

A mixture of $\mathrm{K}\alpha_1$ and $\mathrm{K}\alpha_2$ radiation can be used for most diffraction studies; the $\mathrm{K}\beta_1$ is removed by a filter or a crystal monochromator. The main constituent of the filter material is an element with an atomic number one less than the anode material; the X-ray spectrum of the filter material then has an absorption edge at the wavelength of the $\mathrm{K}\beta$ radiation to be removed. When a monochromator is used, the various wavelengths are separated by diffraction through a crystal, usually a graphite or α-quartz single crystal. The Bragg equation then gives different glancing angles θ for the several wavelengths, so that the unwanted radiation components can be eliminated.

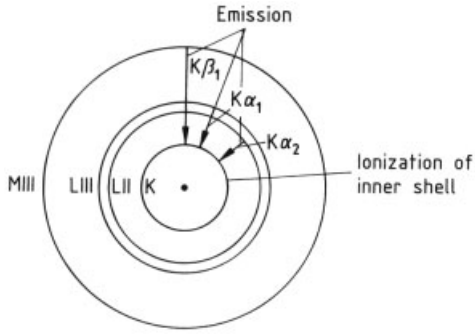

Figure 10. Generation of X-rays

The most common anode materials for X-ray tubes are copper and molybdenum; silver, iron, and chromium are also used. Compounds containing only light atoms (C, N, and O) show only small absorption for molybdenum radiation, while heavier elements have correspondingly higher absorption, but a much more reduced one compared to Cu K_α radiation. Because the scattered intensity is directly proportional to λ^3 (Eq. 14), a copper anode has to be used for extremely small and weakly scattering specimens. The opposite is the case for larger specimens: The low absorption of Mo radiation shifts the intensity ratios in favor of molybdenum. X-ray tubes with chromium anodes generally have low output, air scattering is very high, and absorption problems are difficult or impossible to cope with, at least in single-crystal studies.

Contrasted with these sealed X-ray tubes are rotating anode tubes. The tube is not sealed and the vacuum must be maintained by a pump. The rotation of the anode and the more intensive cooling of the anode material mean that the specific energy loading of the anode can be made far higher; e.g., a generator of power 1.2 kW can yield a specific output of 12 kW/mm^2 with focal spot 0.1×1 mm. For comparison, a fine-focus tube at 90 % of maximum loading has a specific output of only 0.56 kW/mm^2. Rotating anode tubes have now become so reliable that sealed

Table 1. Data on X-ray generation

Anode material	Excitation voltage, kV	Operating voltage, kV	K α_1, nm	K α_2, nm	K α, nm	K β_1, nm	K β filter
Cr	6.0	20 – 25	0.22896	0.22935	0.22909	0.20848	V
Fe	7.5	25 – 30	0.19360	0.19399	0.19373	0.17565	Mn
Co	7.8	30 – 35	0.17889	0.17928	0.17902	0.16207	Fe
Ni	8.5	30 – 35	0.16578	0.16617	0.16591	0.15001	Co
Cu	9	35 – 40	0.15405	0.15443	0.15418	0.13922	Ni
Mo	20	50 – 60	0.07093	0.07135	0.07107	0.06323	Zr
Ag	25.5	60	0.05594	0.05638	0.05609	0.04970	Pd

X-ray tubes may disappear from practical use in the near future.

Another kind of X-ray generators are the microfocus X-ray sources. The focus has an area of only 30 μm^2, and with an operating power of 24 W it attains the same brightness as a 5 kW rotating anode generator. The future will show whether its promise holds for very small crystals. A two orders of magnitude higher X-ray intensity is obtained by using synchrotron radiation. This highly monochromatic, parallel, and brilliant X-ray radiation should only be used when the experiments demand this very expensive X-ray source (small crystals, protein crystals, high scattering angle resolution, very special experiments).

Neutron Beams. The neutron was discovered by CHADWICK in 1932. The equation:

$$mv = \frac{h}{\lambda} \tag{21}$$

where h is Planck's constant, m is the mass of the particle, λ is the wavelength, and v is the particle velocity, makes it possible to regard any moving particle as a wave. The neutron has not only mass but also magnetic moment; accordingly, it can interact with atomic nuclei and with unpaired electrons. Practical neutron sources are high-flux nuclear reactors such as those at Grenoble, Oak Ridge, and Brookhaven, which offer neutron fluxes of roughly $1.5 \times 10^{15}\ cm^{-2}\ s^{-1}$. Reactors available in Germany have fluxes about an order of magnitude lower (BERII in HMI Berlin: $1.2 \times 10^{14}\ cm^{-2}\ s^{-1}$).

The wavelength of neutron radiation is controlled by the absolute ambient temperature T:

$$\frac{1}{2}mv^2 = \frac{2}{3}kT \tag{22}$$

where k is Boltzmann's constant, and Equation (21) gives the value of λ. Neutrons in equilibrium at room temperature (thermal neutrons) have a wavelength of ca. 10^{-10} m. Monochromators for neutrons are generally diffracting crystals. Time-of-flight spectrometers make use of the fact that neutrons with a certain velocity correspond to neutron radiation with a certain wavelength.

Electron Beams. Electrons, like neutrons, are a form of particle radiation. If matter absorbs neutrons far less than X-rays, exactly the opposite holds for electrons. They are so strongly absorbed by air that electron diffraction by the solid phase must generally be carried out in the high vacuum of the electron microscope. The wavelength, and thus the energy of electron radiation are controlled by the voltage drop V through which the electrons pass:

$$\frac{1}{2}mv^2 = eV \tag{23}$$

where e is the elementary charge, m is the electron mass, and v is the electron velocity. This, in combination with Equation (21), gives the wavelength:

$$\lambda = 12.236 V^{-1/2}\ \left(10^{-10}\,\mathrm{m}\right) \tag{24}$$

For example, if the voltage drop is 100 kV, the electron beam has a wavelength of 0.039×10^{-10} m. Because electrons do not penetrate to a great depth, they are particularly suited to the study of thin films and surface structures. The scattering length is four orders of magnitude larger than for X-rays and neutrons. This leads not only to a correspondingly higher absorption, but also to an extremely high scattering power of the atoms.

2.1.6. Cameras and Diffractometers

Three types of detector are available for measuring the X-rays diffracted by a crystal: Point detectors, linear and area detectors. The film method involves a two-dimensional detector, which can provide simultaneous information about every point in the region of reciprocal space investigated. The measurements are not immediately available in digital form, but powerful computer-controlled photometers can digitize them.

This technology is sometimes still used in protein crystallography, but the electronically controlled area counter is being introduced in this field. This type of position-sensitive counter directly supplies digital values as a function of position in reciprocal space; it also has a dynamic range several orders of magnitude wider than the film. In conjunction with a rotating anode X-ray source, it forms a data acquisition system with which not just proteins, but also other large molecules with poorly scattering atoms or very small crystals can be measured in an optimal way [66], [67]. One-dimensional position-sensitive detectors are available for powder diffractometry [68].

Point measurements with scintillation or proportional counters yield equally precise values. However, important parts of the information can be lost if it is not borne in mind that only those points can be measured that are moved into reflection position with complete prior knowledge of that position.

Recording methods also differ with which clompleteness a region of reciprocal space can be measured. In the analysis of crystalline powders, where the orientations of the lattice planes are expected to show a statistical distribution, the largest measurable glancing angle controls the extent of reciprocal space that can be covered. The situation is similar with the four-circle single-crystal diffractometer; every net plane can be brought to reflection position, provided a maximum glancing angle (characteristic of the instrument) is not exceeded. In all other single-crystal goniometers, there are "blind regions" in the Ewald sphere that cannot be measured.

Film Methods. The prerequisite for using the film method is a suitable film material. The Committee for Crystallographic Apparatus of the International Union of Crystallography has tested a variety of X-ray films [69]. The properties of a usable X-ray film can be characterized as: Fine grain and homogeneous emulsion; low sensitivity to temperature and moisture; low susceptibility to fogging; wide range of linearity between film blackening and logarithm of exposure time; and high sensitivity to the wavelengths produced by X-ray tubes.

Nowadays it can be difficult to obtain film material of such quality because the medical purposes for which it was primarily intended have mostly vanished. However, the conditions can be relaxed because quantitative intensity measurements are no longer made by using films.

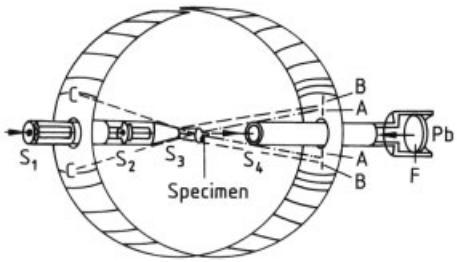

Figure 11. Debye – Scherrer camera

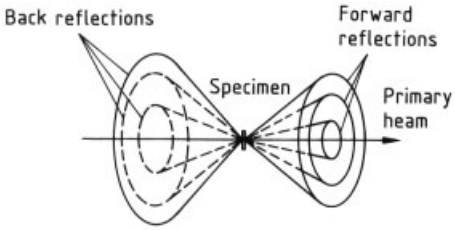

Figure 12. Diffraction cones in the crystalline powder method

Debye – Scherrer Camera. Figure 11 illustrates the principle of the Debye – Scherrer camera, while Figure 12 shows schematically how the diffraction pattern is generated. A specimen of crystalline powder (ca. 0.1 mg) is placed in a borate glass capillary of diameter 0.3 – 1 mm and length ca. 1 cm, and the capillary is rotated about a specified axis. The film fits snugly inside the camera. In a camera with internal diameter 114.6 mm, 1 mm on the film corresponds exactly to a diffraction angle of 1°. The radiation passes through the collimator (with three diaphragms S_1, S_2, S_3) to reach the specimen; S_1 and S_2

Figure 13. Debye – Scherrer diffraction pattern of LiF

have the same diameter, while S_3 is somewhat larger. The primary beam is absorbed in the primary beam trap, which carries diaphragm S_4, and can be looked at through the lead-glass window, Pb, on the fluorescent screen F for setting purposes. The position and size of screen S_4 limit the smallest measurable diffraction angle 2θ, and thus the resolution power of the camera (A – A) in the small angle region. The position of S_3 relative to S_2 determines the range of angles within which scattered radiation from S_2 is observed (B – B). S_3 also masks backscattering, so that no diffracted X-rays can reach the region C – C. Figure 13 shows a Debye – Scherrer pattern. The intersections of the diffraction cones in Figure 12 with the film give the Debye – Scherrer rings.

The camera should be evacuated while the pattern is being recorded, to minimize the background due to air scattering.

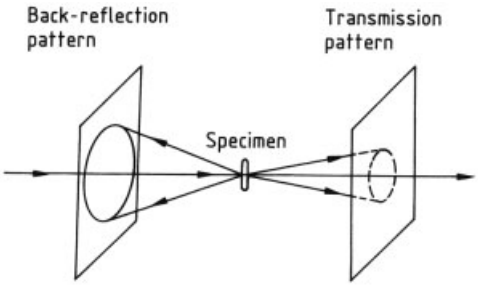

Figure 14. Laue camera

This method has the advantages that a small specimen can be used, operation is simple, and there is only slight falsification of the intensities due to nonstatistical distribution of crystallite orientations in the specimen. Disadvantages of the method are that exposure times are quite long (up to and beyond 12 h) and interpretation of measurements is inconvenient, especially when there are many specimens, because the results are not in digital form.

Laue Camera. This flat-film camera can record both transmission and back-reflection patterns. The principle is illustrated in Figure 14; Figure 15 shows a flat-film pattern. Transmission patterns with polychromatic radiation yield information about the symmetry of a crystal (Laue symmetry) or with monochromatic radiation the orientation of the crystallites (preferred orientation, texture); such patterns are recorded, for example, from fibers. Back-reflection patterns permit the nondestructive examination of workpieces.

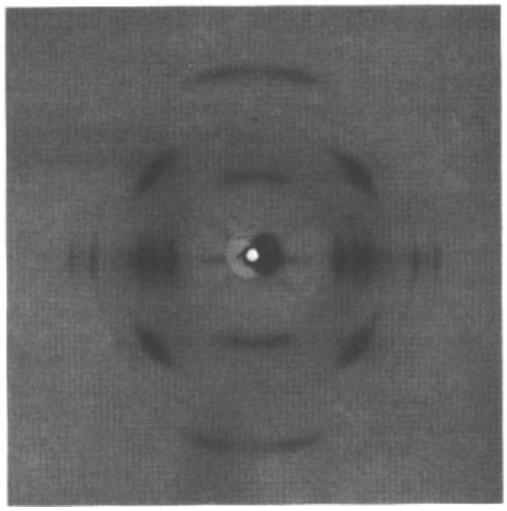

Figure 15. Flat-film diffraction pattern of polypropylene fibers

In the Laue method, single crystals are irradiated with polychromatic X-rays, so that the crystal need not be rotated. All net planes have wavelengths for which the reflection conditions are satisfied. The patterns are employed for the alignment of single crystals and the identification of their symmetries.

Guinier Camera. The Guinier camera is a focusing goniometer; for a certain geometric configuration of specimen and X-ray tube focus F, the X-rays are focused, so that higher angular resolution is achieved (Fig. 16). The radiation is reflected from a bent quartz monochromator (Johansson principle). The focussing circle (Seemann – Bohlin circle) is coupled with it in such a way that both circles have the same

focal axis [70]. In addition to the symmetrical transmission configuration shown, three other relative positions of the monochromator and Seemann – Bohlin circles are possible, so that the entire X-ray diffraction spectrum can be recorded; these are the asymmetric transmission position, the symmetric reflection position, and the asymmetric reflection position (Fig. 17).

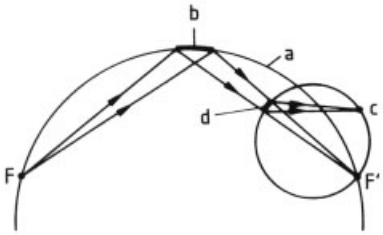

Figure 16. Principle of the Guinier method
a) Seemann – Bohlin circle; b) Monochromator; c) Reflection; d) Specimen

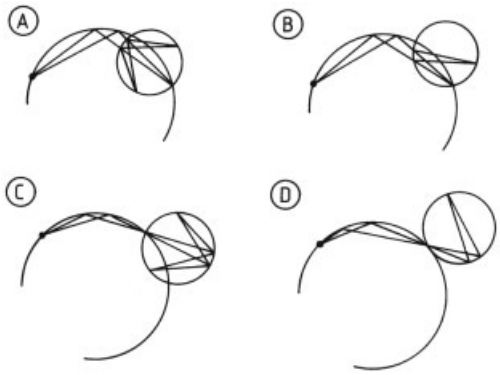

Figure 17. Guinier method, showing the four possible relative positions of specimen and monochromator
A) Symmetrical transmission; B) Asymmetrical transmission; C) Symmetrical reflection; D) Asymmetrical reflection

The specimen, in the form of a crystalline powder, is applied to the substrate in the thinnest possible layer. With strongly scattering inorganic substances, the specimen can be dusted onto a film of grease. With organic substances, it is necessary to work with a layer at least 0.1 – 0.3 mm thick, which may be sandwiched between foils, with the result, that the resolution of the system is somewhat less.

Weissenberg Camera. Figure 18 shows the camera schematically, while Figure 19 illustrates the registration of a diffraction pattern. This camera is used in recording diffraction data from single crystals [10, pp. 221 – 225]. If the layer-line screens are removed, an overall record of the reciprocal lattice is obtained (rotating crystal pattern; Fig. 20). With these screens in place, the individual diffraction cones of Figure 19 (the layer lines) can be isolated. Coupling the crystal rotation with a translation of the film holder in the direction of the rotation axis spreads the one-dimensional layer line of the rotating crystal pattern into a layer plane of the reciprocal space. If the crystal is rotated about one of the crystallographic axes a, b, c, the number of the layer line gives the index of the corresponding reciprocal axis. The diffraction points can be uniquely indexed, even though the reciprocal lattice is recorded in distorted form (Fig. 21). The method is superbly suited to intensity measurements, since an entire film pack can be placed in the film cassette and exposed to the X-ray beam; the films are then exposed simultaneously, but exposure varies with the penetration depth of the X-rays.

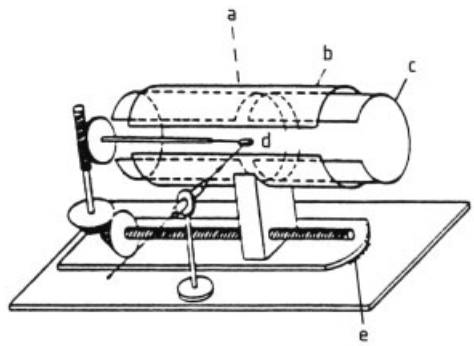

Figure 18. Principle of the Weissenberg camera
a) Equi-inclination axis; b) Film; c) Layer-line screen; d) Crystal; e) Equi-inclination angle setting

The normal technique is the equi-inclination method, in which the direct and diffracted X-ray beams make equal angles with the layer line being measured. In addition to increasing the number of measurable reciprocal lattice points, this also simplifies the necessary intensity corrections. Figure 22 shows that the maximum length of the reciprocal lattice vector parallel to the rotation axis is:

$$\zeta_{\max} = \frac{2}{\lambda} \sin\mu_{\max} \tag{25}$$

The maximum equi-inclination angle $\mu_{\max}$ is normally 30° ($\zeta_{\max} = 1/\lambda$), so that $\zeta_{\max}$ is the

same as in the normal rotating crystal pattern, except that there is no blind region near the rotation axis (Figs. 5 and 22).

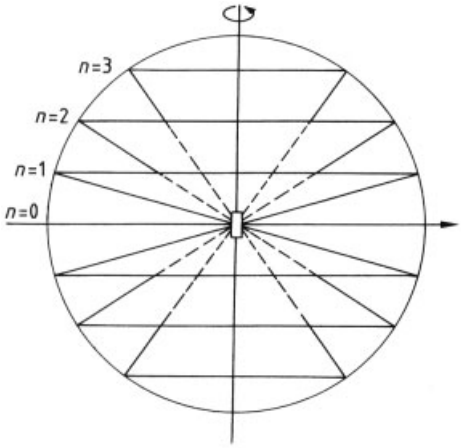

Figure 19. Diffraction cones in the rotating-crystal method (n = number of layer line)

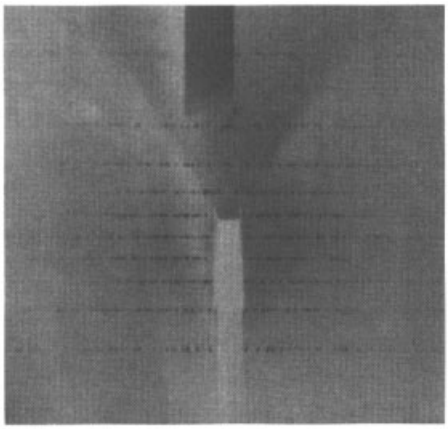

Figure 20. Rotating crystal pattern

Precession (*Buerger*) *Camera.* In contrast to the Weissenberg camera, the crystal carrier is now coupled to the film cassette so that both move in exactly the same way while the pattern is being recorded (Fig. 23). A direction in the crystal precesses (precession angle μ) about the incident X-ray beam, so that reciprocal lattice points lying in a plane perpendicular to this crystal direction cut the reflection sphere. A precession pattern represents the reciprocal lattice in undistorted form (Fig. 24); the reflections are easily indexed and the pattern is easily interpreted [9, pp. 15 – 68]. The cone pattern of Figure 25, which shows the layer lines measurable with a given crystal setting, corresponds to the rotating crystal pattern.

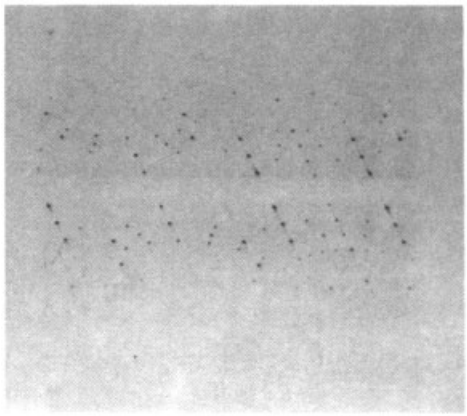

Figure 21. Weissenberg pattern

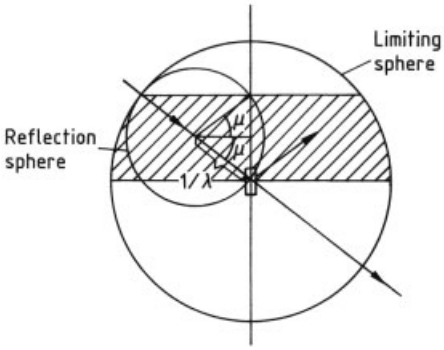

Figure 22. The portion of the limiting sphere that is accessible with the equi-inclination method

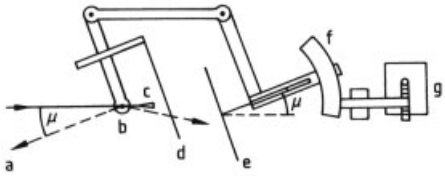

Figure 23. Precession camera
a) Vector in real space; b) Crystal; c) Primary beam stop; d) Layer-line screen; e) Film plane; f) Precession angle setting; g) Motor

Figure 26 illustrates the region of reciprocal space whose lattice points can be measured with a single spindle setting (shaded area). The maximum ζ is $2/\lambda \cos \mu$. One further spindle setting, which should be as close as possible to 90° away

from the first, will eliminate most of the blind spot about the ζ-axis.

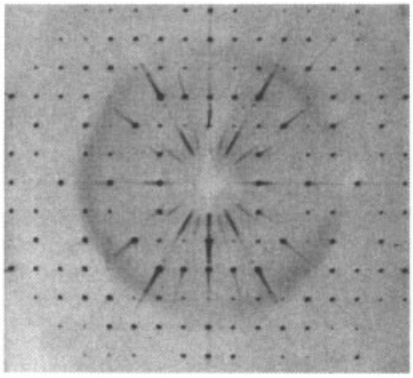

Figure 24. Precession pattern

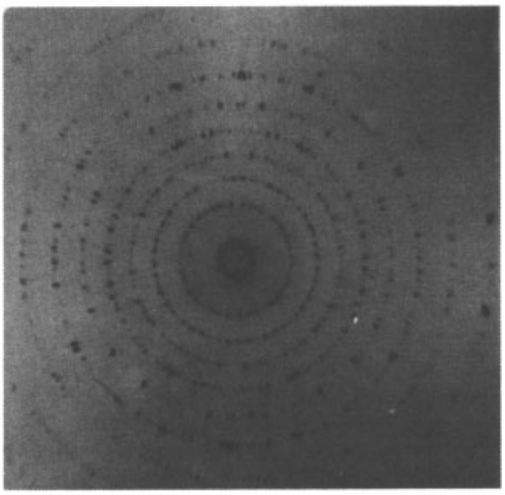

Figure 25. Cone pattern

Space Explorer. The Weissenberg camera records only reciprocal lattice planes perpendicular to the crystal axis that is parallel to the goniometer head axis, and the precession camera can measure only lattice planes that have a common reciprocal lattice axis parallel to the goniometer head axis and lattice planes which are parallel to these [24, pp. 98 – 167]. With the Space Explorer both types of series can be recorded. This facilitates the determination of cell constants and space groups in many cases.

These film methods may appear old fashioned, but this is wrong. Any serious X-ray laboratory will have a few of these machines, which are always necessary to give a first impression of the diffraction problem on a relatively cheap apparatus, especially when the expensive computer-controlled diffractometers are not immediately available or when there are problems and a three-dimensional inspection of the reciprocal space is necessary to gain an impression of the difficulties.

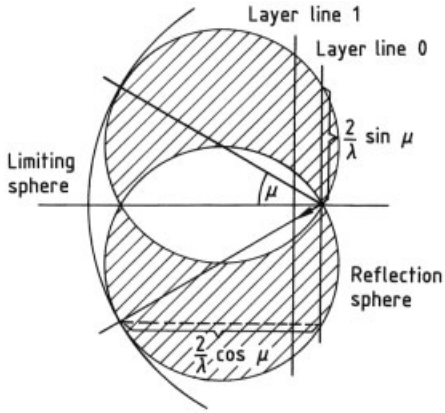

Figure 26. Region of the limiting sphere covered by the precession method

Counter Tube Method. The term "counter tube" means a detector that moves on a circle, the equator of the diffractometer. Only point measurements along this circle are possible. If one wishes to measure reflections that do not lie on this plane, the specimen must be oriented to position the reflections to be measured on the equator.

The use of this type of counter tube demands complicated electronics that make it possible to count the absorbed X-ray quanta. In the proportional counter, every X-ray quantum initiates a gas discharge; the voltage surge is electronically amplified and recorded as a pulse. In the scintillation counter, the X-ray quanta absorbed by the scintillator crystal are transformed into light (fluorescence), and amplified by a photomultiplier. The two types of counter tube have very similar properties, and the pulses, once amplified, can be processed by the same electronics. A device called an electronic discriminator can block a given energy range ("window") and thus render the beam more monochromatic, greatly increasing the ratio of peak to background. Unfortunately, the use of a filter or a monochromator crystal to separate the K α and K β radiation cannot be dispensed with, because the resolution

of an ordinary discriminator is not sufficient to do so.

The situation is different if a semiconductor detector (lithium-doped silicon crystal) and the proper electronics are available. This setup can separate K α and K β radiation, so that the beam does not have to be passed through a monochromator. The counter tube itself has nowadays become easier to operate, for the semiconductor crystal no longer has to be maintained at liquid nitrogen temperature – the problem has been solved electronically. Efforts continue to find broader applications for this type of detector, since replacement of the monochromator by electronics means that there is no longer a 50 % loss of primary intensity through diffraction by the monochromator crystal or absorption by the β filter.

The diffractometer must have a precision mechanical system for the positioning of specimen and detector, with a reproducibility of at least 0.01°. Digital evaluation makes it relatively easy to perform a profile analysis of the reflections.

In comparison with film cameras, diffractometers are roughly a factor of 10 more costly. They are still more expensive if the point counter is replaced by a one- or two-dimensional position-sensitive detector. The data acquisition time is, however, shortened by a factor of 50 – 100, or even more, depending on the geometry of the crystal lattice measured.

There are four types of position-sensitive detectors: Wire detectors (one- and two-dimensional), video detectors, CCD (charge coupled device chip), and image-plate detectors. The wire detector is based on the proportional counter principle; voltage pulses produced by quanta incident at different points along the wire take different times to reach its two ends. The travel times are derived from the profile of the sum of the two pulses. In the video counter, a phosphor layer generates light flashes where X-ray quanta are incident, and the flashes are picked up by a television camera. The CCD chip is covered by a scintillation foil, which is specific for a certain X-ray wavelength. The generated light flashes are counted by the CCD. The image-plate has a coating of rare-earth oxides, in which metastable states are produced by X-ray quanta. Transitions back to the ground state, induced by laser light, can be observed and quantified.

These position-sensitive detectors permit the use of simpler mechanical systems in diffractometers. A one-dimensional counter makes it possible to leave out one circle; a two-dimensional counter, two circles. In principle, then, a single-crystal diffractometer equipped with a two-dimensional counter needs only one circle, while a powder diffractometer equipped with a one-dimensional counter does not require a circle at all. Experience has shown that the greatest flexibility is achieved by incorporating these counters without reducing the number of circles.

A direct oscillation photograph, for example, can only be made by a four-circle diffractometer equipped with a two-dimensional detector.

Two-circle powder diffractometer fitted with a point counter. There are three powder diffractometer geometries with focusing optics, whereby the monochromator, the specimen, or both, can have a focusing characteristics. Usually, they are two-circle diffractometers.

The Bragg – Brentano two-circle diffractometer has a Seemann – Bohlin focusing circle, on which the X-ray tube anode, the counter tube window, and the specimen are located (Fig. 27). Only the specimen is focusing.

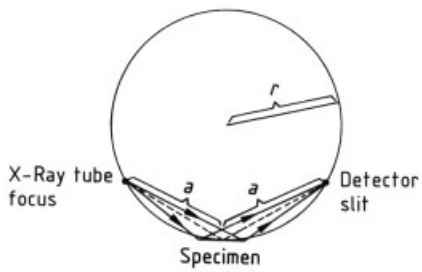

Figure 27. Beam path of the Bragg – Brentano two-circle powder diffractometer

The planar specimen is placed tangentially to the focusing circle. Because theoretically this can be the case only along a line parallel to the line focus, there are residual geometric beam-spreading effects. The radius r of the focusing circle decreases with increasing θ, from infinity at 0° to $a/2$ at 90°. In order to maintain the focusing conditions while a diffraction pattern is being recorded, the detector is moved at twice the angular velocity of the specimen carrier. Be-

cause the specimen extends in two dimensions, it is possible, and necessary, to work with the line focus of the X-ray tube (perpendicular to the plane of Fig. 27); this makes very effective use of the X-ray tube anode surface.

As an alternative to a β filter, beam monochromatization can be achieved by a graphite monochromator between specimen and detector; this arrangement removes not only K β radiation, but also fluorescence emitted by the specimen. The effect is particularly beneficial in the study of iron-containing specimens with Cu K α radiation.

An $x-y$ plotter connected to the instrument (if it is not inherently computer-controlled) makes it possible to output the scattered X-ray intensity in analog form as a function of diffraction angle (Fig. 28). The intensity of the diffracted beam can also be output in digital form. An especially economical approach is to fit the instrument with an automated specimen changer, record the measured X-ray intensities on a mass storage medium (magnetic tape, diskette, or preferably yet a hard disk), and control the operation of the diffractometer by computer. Data measured, say, overnight or over the weekend can then be interpreted during regular working hours with the same computer.

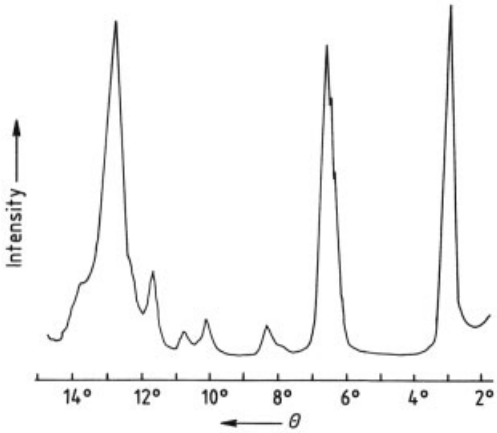

Figure 28. X-ray diffraction pattern of the γ-phase of quinacridone, recorded in a Bragg – Brentano diffractometer with Cu K α radiation

In the *Debye – Scherrer diffractometer* with monochromator (Fig. 29), the line focus of the X-ray tube, the monochromator crystal, and the detector window all lie on the focusing circle of the curved monochromator crystal (zero setting). The specimen is measured in transmission. For inorganic substances with relatively high scattering power, a specimen dusted onto a foil with the thinnest possible coating of grease is sufficient. This technique is generally not adequate for organic specimens; instead, the specimen is contained between two thin foils, separated by a spacer. It is useful to do a transmittance measurement in the primary beam before performing the diffraction measurement itself. A transmittance of about 70 % should be obtained. For substances that do not absorb too strongly, the specimen can be placed in a capillary; among the advantages of this technique is that the measurement can be carried out in a controlled atmosphere.

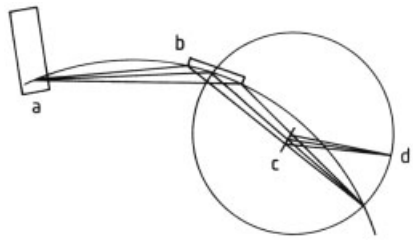

Figure 29. Beam path of the Debye – Scherrer diffractometer
a) X-ray tube; b) Monochromator; c) Specimen; d) Detector and detector slit

Figure 29 further shows that the specimen position does not contribute to focusing. In a fixed specimen position, only absorption varies somewhat with diffraction angle. This setup is thus suitable when a curved one-dimensional position-sensitive detector is employed. There are detectors on the market that can measure up to $2\theta = 120°$ at the same time. When such a detector is used, no diffractometer is needed – only a base plate support. This technique has proven itself useful, particularly with programmed heating of specimens, and with in situ recording of diffraction diagrams.

The principle of the Guinier camera, described in section page 13, has been developed by replacing the film by a counter tube to give the *Guinier diffractometer* (Fig. 30). The reflection intensities can be digitally evaluated; this is a major advantage for quantitative phase and profile analyses. The resolution obtained is much better than with the Bragg – Brentano method (compare Figs. 28 and 31), but specimen prepa-

ration is more complicated, and no automatic sample changer is available.

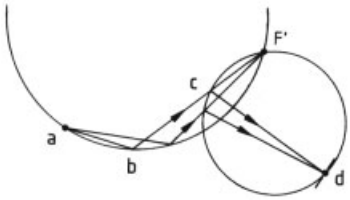

Figure 30. Beam path of the Guinier diffractometer
a) Line focus of X-ray tube; b) Monochromator; c) Specimen; d) Detector slit

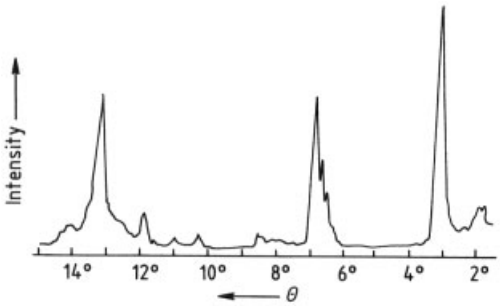

Figure 31. X-ray diffraction pattern of the γ phase of quinacridone, recorded by the Guinier method with Cu K α_1 radiation

Single-Crystal Diffractometer. Most intensity measurements are now carried out with a counter tube; measurements are more accurate, and the results are obtained directly in digital form.

A relatively simple and inexpensive diffractometer is the single-crystal instrument with Weissenberg geometry [24, pp. 98 – 167] (see Fig. 18). Measurements can be automatically controlled by computer, but the counter tube must be realigned manually for each layer line. The instrument has the advantage of low cost, but frequent realignment is needed (semiautomatic instrument), and the individual reflections are not all measured in the same geometrical arrangements (Weissenberg geometry).

Four-circle diffractometers (Figs. 32 and 33) are free of these disadvantages. The ϕ and χ circles set the crystal in such a way that the reciprocal lattice vector of the reflection being measured lies in the equatorial plane of the instrument, while the counter tube moves along this circle (2θ circle). The θ and 2θ circles make it possible to set up the Bragg relation. The χ circle, called the Eulerian cradle, can be either closed or open (as in Fig. 32). A closed Eulerian cradle has the advantage that any reciprocal lattice vector can, in principle, be brought into the equatorial plane; with an open 90° Eulerian cradle, this is possible only for half the Ewald sphere, though normally this is sufficient. The open Eulerian cradle permits a much sturdier construction, and "shading" by the closed Eulerian cradle ceases to be a problem. Complete control and monitoring by computer is possible with either form.

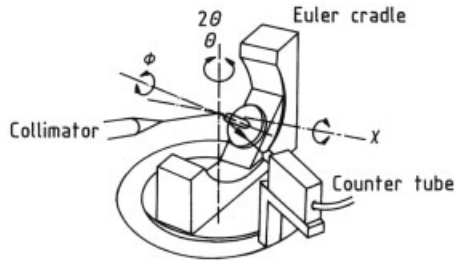

Figure 32. Four-circle diffractometer with open Eulerian cradle (χ circle)

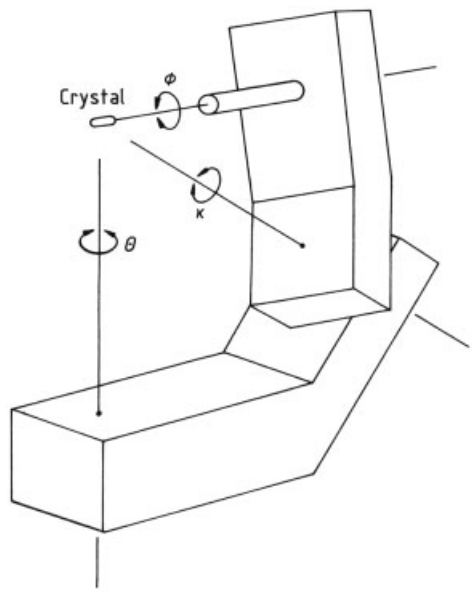

Figure 33. CAD 4 diffractometer

An important advantage of the closed Eulerian cradle is that in low-temperature measurements the cold nitrogen stream can keep its lamellar characteristics when it is passing the crystal on the rod, provided the nitrogen stream

and the axis of the goniometer head remain always coaxial, what can be easily arranged with a closed Eulerian cradle.

An elegant version of the four-circle diffractometer is the CAD 4 diffractometer made by Nonius. Here the Eulerian cradle or χ circle is replaced by the $\varkappa$ circle, whose axis intersects that of the θ circle at an angle of 50° (Fig. 33). The axis of the ϕ circle likewise makes an angle of 50° with that of the $\varkappa$ circle, so that the ϕ axis can lie anywhere on a 100° cone.

Data Processing by Computer. Because minicomputers have become so inexpensive, it is an obvious step to use the computer not only to control the diffractometer, but also to solve the structure problem. In fact, it is common to use one computer for control, and a second for data processing. While the diffractometer, for example, is at work measuring the diffracted intensities of the computer-adjusted crystal of a compound, another structure can be solved, by either the same minicomputer or a second one. Nowadays, a personal computer is used for the least-squares refinement of the full matrix of the atomic parameters even for large structures. The same holds for the solution of the phase problem.

The combination of the rotating anode X-ray generators which are nowadays very stable, the robust one-, two-, three-, or four-circle single-crystal diffractometers, the area detectors (image plate, wire counters, or CCD), and powerful PCs make solving the crystal structure of a medium-sized molecule a job of a few hours, of a few days, or exceptionally of a week for crystals that are relatively small ($< 50\,\mu$m in one or all dimensions) and/or show greater disorder.

2.1.7. Safety

For electron radiation, the very high absorption in air means that no special safety rules have to be followed; but the X-rays produced by electrons do call for precautions. The manufacturers of most electron microscopes have solved this problem. Neutron diffraction experiments are not performed in the chemist's own lab, and the host institute therefore has the responsibility of complying with safety regulations. In this section, only practices relating to X-rays are discussed.

The regulations, in force in Germany are set forth in the "X-Ray Regulation" of 8, January 1987 [71], which supersedes the similarly titled regulation of 1, March 1973. The new regulation is based on recommendations issued in 1958 by the International Committee for Radiation Protection [72], [73]. The code prescribes, in part, as follows:

1) A radiation protection supervisor must be designated for the operation of X-ray instruments other than fully shielded instruments. If necessary, the radiation protection supervisor designates radiation protection officers, who must demonstrate familiarity with the requisite specialist knowledge.
2) X-ray equipment other than a fully shielded instrument may be operated only in a room enclosed on all sides and designated as an X-ray room. The controlled access area is defined as an area where persons can receive more than 15 mSv (1.5 rem) in a calendar year as whole-body exposure; the corresponding figure for the monitored area is 5 mSv.
3) Workers occupationally exposed to radiation may not exceed a whole-body dose of 50 or 15 mSv in a calendar year. In the first case, much more stringent provisions apply to the medical supervision of the worker. Special provisions apply for pregnant women and persons under 18 years of age.
4) The individual dose is measured by two independent methods. The first method is performed with dosimeters, which are to be submitted, at intervals not to exceed 1 month, to the government agency having jurisdiction. The second measurement must permit the determination of dose at any time, and may be put into use at the direction of the radiation protection supervisor or of the radiation protection officer or at the request of the worker.

Exposure to the primary beam from an X-ray tube (ca. 10^5 R/min) must be avoided under all circumstances.

2.1.8. Instrument Manufacturers

What follows is a list of companies trading in Germany, together with the instruments they market. No claim for completeness is made.

Nonius B. V., Röntgenweg 1, P. O. Box 811, 2600 AV Delft, The Netherlands. X-ray sources: Sealed tubes or rotating anode generators, microsource generators; goniometer heads; high- and low-temperature devices; four-circle K-axis goniometers; one- and two-dimensional detector systems, CCD and imaging plate detectors.

Freiberger Präzisionsmechanik GmbH, Hainichener Strasse 2 a, 09599 Freiberg, Germany. X-ray sources; X-ray tubes; Debye – Scherrer cameras; two-circle powder diffractometers; precession cameras; Weissenberg cameras.

Huber Diffraktionstechnik GmbH, Sommerstraße 4, 83253 Rimsting/Chiemsee, Germany. X-ray diffraction equipment, goniometer heads: Laue, Debye-Scherrer, Buerger-precession, and Weissenberg cameras; imaging-plate Guinier cameras; X-ray diffractometers; high- and low-temperature attachments; monochromators; Eulerian cradles, rotation stages, linear stages, multiaxis goniometers for synchrotron and neutron facilities.

Philips GmbH, Miramstrasse 87, 34123 Kassel, Germany. X-ray sources; X-ray tubes; goniometer heads; high- and low-temperature stages; position-sensitive detectors; texture goniometers; two-circle powder diffractometers; thin-film high-resolution diffractometers.

Molecular Structure Corporation Europe, Unit D2, Chaucer Business Park, Watery Lane, Kemsing, Sevenoaks, Kent TN15 6YU, United Kingdom. Imaging plate detectors; CCD detectors; partial-Chi circle; low-temperature devices; rotating anode generators; peripheral devices.

Bruker AXS GmbH, Östliche Rheinbrückenstrasse 50, 76187 Karlsruhe, Germany. Rotating-anode and sealed-tube sources; ceramic X-ray tubes; multilayer optics; mono/poly capillaries; two-circle powder diffractometer; reflectometer; texture/stress diffractometer; attachments for high/low temperature, high pressure, humidity control, reactive gas; micro diffractometer; 2D SAXS camera; process automation diffractometer; chemical and biological 2D-CCD single-crystal systems.

Rich. Seifert & Co GmbH Röntgenwerk, Freiberger Präzisionsmechanik GmbH, Bogenstraße 41, 22926 Ahrensburg, Germany. Four-circle diffractometer, Theta/Theta powder diffractometer, low- and high-temperature sample stages, phase-texture-stress-diffractometer for laboratory and automotive industry, thin-film texture/stress Eta-diffractometer, reflectometer, high-resolution diffractometer, wafer ingot orientation systems for industrial applications, Rayflex Windows Software, X-ray generators, X-ray tubes, X-ray optics, parabolic multilayers, monochromators, film cameras, position-sensitive detectors PSD, energy dispersive detectors, electroluminescence detectors LUX.

Stoe & Cie GmbH, P.O. Box 101302, Hilpertstrasse 10, 64295 Darmstadt, Germany. Goniometer heads; Debye – Scherrer cameras; Guinier cameras; Weissenberg cameras; precession cameras; Reciprocal Lattice Explorers; powder diffractometers with scintillation counter, position-sensitive detectors or imaging plates; special sample holders; high-temperature and texture sample stages; four-circle single-crystal diffractometers also with CCD detector; one- and two-circle single-crystal diffractometers with imaging plates; crystallographic software.

2.2. The Phase Problem

In general, $F(hkl)$ is a complex number; when plotted in the Gaussian plane (Fig. 34), it has not only a modulus $|F|$ but also a phase angle φ.

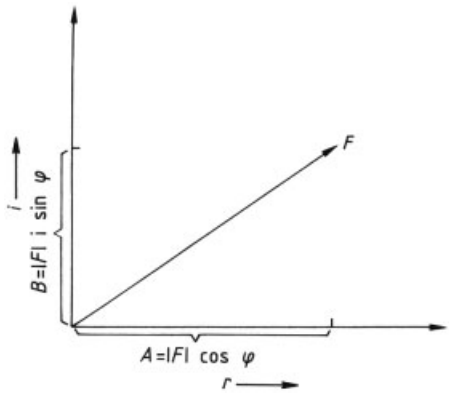

Figure 34. Gaussian plane
The real part of the complex number F is plotted along the r-axis; the imaginary part along the i-axis

By the Euler – Moivre theorem

$$|F| \exp(i\varphi) = |F| \cos\varphi + i\,|F| \sin\varphi = A + i\,B \qquad (26)$$

A structure factor is obtained not by simple scalar addition of the scattering contributions atom-by-atom, but by vector addition, as the plot in the Gaussian plane makes clear (Fig. 35). The angle between a scattering contribution and the real axis is given by $2\pi(hx_n + ky_n + lz_n)$ (Eq. 18). The phase angles of the structure factors give the angles by which the sine waves of the individual structure factors in the Fourier synthesis must be shifted with respect to the origin in order to yield the electron density.

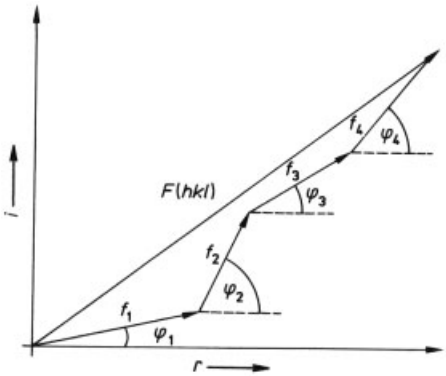

Figure 35. Addition of structure factor contributions in the complex plane

An important special case is where B in Equation (26) vanishes, and the only two phase angles possible for $F(h,k,l)$ are 0° and 180°. The phase problem then becomes a matter of adding signed quantities. According to Equations (18) and (26), this occurs whenever an atom with coordinates x, y, z always has an atom of the same species at coordinates $-x$, $-y$, $-z$ corresponding to it. In other words, the phase problem becomes a signed addition problem in all centrosymmetric space groups.

In the methods now commonly used for measuring diffracted X-ray intensities, it is generally not possible to measure the phase angle of a structure factor. The phase is said to be "lost" in the intensity measurement. In the early days of X-ray structure analysis, it was thought that phases could not be deduced from measured intensities.

2.2.1. Patterson (Vector) Methods

Not until A. L. PATTERSON's work [74] did it gradually become clear that the measured intensities also contain phase information. The function:

$$P(uvw) = \frac{1}{V} \sum_{hkl=-\infty}^{+\infty} |F(hkl)|^2 \exp[2\pi i(hu + kv + lw)] \qquad (27)$$

does not give the electron density at point x, y, z but does yield maxima in a vector space. Each such maximum has a position (relative to the origin) corresponding to the vector between two atoms; its weight is proportional to the product of the electron numbers of the two atoms. The vector space equation requires knowledge only of the structure amplitudes — F —, not of the phases. It is possible to solve a structure that is not too complicated, if all the maxima of a Patterson synthesis are known.

Multiple solutions may exist for one and the same set of vectors in extremely rare cases (homometric structures). No case is known in which two homometric structures are both chemically meaningful.

The complete interpretation of a Patterson synthesis is generally made much more difficult because the maxima are not resolved, and cannot be explicitly identified. Figure 36 illustrates this situation for the two-dimensional case. Corresponding to the 6 maxima in the crystal space, there are 19 maxima in the vector space; these must be placed in the same unit cell volume that accommodates the 6 atoms. What is more, the maxima in the vector space require more volume than atoms in the crystal space, because the indeterminacies of each pair of atoms add at the Patterson maximum. The half-width values of the maxima in the vector space are thus roughly twice those in the crystal space. As a consequence, the peaks overlap strongly. The relative weights of the maxima are indicated in Figure 36.

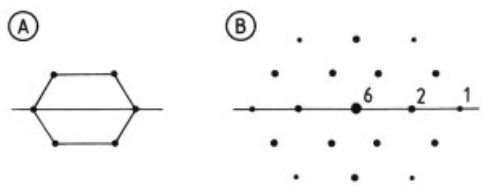

Figure 36. A) Crystal space; B) Vector space

Patterson synthesis by itself cannot solve even moderately large molecular structures (five or six equally heavy atoms) unless concrete structure information already exists so that the vector space can be systematically searched for a presumed structure or substructure. A search of this kind might be done with a computer, perhaps by the method of pattern-seeking functions, or the convolution molecule method [57], [75]. These techniques, however, are not so often used today, having been supplanted by direct phase-determination methods (Section 2.2.2), which make it possible to reach the goal quickly on the basis of relatively little prior knowledge [76]. Patterson synthesis may return to prominence when used to identify sets of starting phases (Section 2.2.2).

An exception of that is PATSEE [77], which uses one or two known fragments of the structure, whose orientation is found by a real-space Patterson rotation search and its translation in the cell by direct methods (Section 2.2.2), starting from random positions. Figures of merit related to those of the direct methods are used to refine only promising phase sets.

Heavy Atom Method. The principal use of Patterson synthesis at present is in the heavy atom method. If a few heavy atoms, such as bromine or iodine, are present in a molecule along with light atoms, such as carbon, nitrogen, or oxygen, the Patterson technique makes it possible to analyze the X-ray diffraction pattern into heavy- and light-atom components. For a structure to be solved reliably, however, the light and heavy atoms must have sufficiently different scattering powers, but the sum of the heavy-atom scattering powers (proportional to the sum of the squares of the atomic numbers Z) must not be much different from that of the light atoms:

$$\sum Z_{\mathrm{H}}^2 \geq \sum Z_{\mathrm{L}}^2 \tag{28}$$

Given these conditions, the heavy-atom vectors in the Patterson synthesis have markedly higher maxima associated with them than the heavy/light- or light/light-atom vectors. In centrosymmetric structures, this holds even when the ratio of heavy to light atoms is less favorable, since in this case the knowledge of a small structural component is enough to establish with fair confidence the sign trend of a large structure factor.

If, say, a bromine atom has been incorporated into an organic molecule to facilitate X-ray structure analysis, the weights of bromine – bromine, bromine – carbon, and carbon – carbon vectors are in the ratio:

$$Z_{\mathrm{Br}}^2 : Z_{\mathrm{Br}} Z_{\mathrm{C}} : Z_{\mathrm{C}}^2 = 1225 : 210 : 36 \tag{29}$$

Thus, it should not be hard to identify the bromine – bromine vectors, given the sizes of the maxima. Further, heavy-atom maxima are not expected to overlap, because they have virtually the entire cell available, while carbon – carbon maxima and, to a great extent, bromine – carbon maxima will recede into the background noise.

Example. Vectors between heavy atoms are especially easy to identify if just one heavy atom is located in the asymmetric unit. This is shown here for the centrosymmetric space group P 2_1/c. From the coordinates in the crystal space:

$$x, y, z; x, \frac{1}{2} - y, \frac{1}{2} + z;\ -x, -y, -z;\ -x, \frac{1}{2} + y, \frac{1}{2} - z$$

the coordinates in the vector space can be found by forming the 16 possible vectors, as shown in Table 2. Such a table should always be prepared for difficult cases.

The space group in the vector space is P 2/m, which is the Laue group of P 2_1/c. Only 21 such "Patterson groups" are possible with conservation of the Bravais lattice. The space group of the vector space (Patterson group) is derived from that of the crystal by adding a center of symmetry (if none is present) and removing translation from the group of symmetry elements.

Four maxima are located at the origin; four at 0, $\frac{1}{2} \pm 2y$, $\frac{1}{2}$; four at $\pm 2x$, $\frac{1}{2}$, $\frac{1}{2} \pm 2z$; two at $\pm 2x$, $\pm 2y$, $\pm 2z$; and two at $2x$, $\pm 2y$, $2z$. For the 0, $\frac{1}{2} \pm 2y$, $\frac{1}{2}$ maximum, only y is variable; the maximum lies on a "Harker line". At $\pm 2x$, $\frac{1}{2}$, $\frac{1}{2} \pm 2z$, the section cuts through the cell at $y = \frac{1}{2}$ (Harker section). The first condition yields the y coordinate; the second yields the x and z coordinates of the heavy atom. The maxima at $2x$, $2y$, $2z$ (and the symmetry-related positions) simply provide confirmation of the values found.

Maxima on Harker lines and sections are always doubly occupied, so that they are still easier to find.

If there is more than one heavy atom in the asymmetric unit, there are also vectors between

the heavy atoms in the asymmetric unit. A table should be prepared in order to make clear the weighting of individual maxima. The convolution technique can lead to a solution in a very direct way [58].

All of this work is nowadays carried out by computer programs (e.g., SHELX-97 [49]). It is advisable, however, to do at least part of it by hand, so that one is familiar with the theory when there are difficulties.

From the heavy-atom positions obtained by this procedure, the structure factors are calculated from Equation (18), and the signs or phases are assigned to the measured structure amplitudes. The result is a Fourier synthesis of electron density. Because the phases depend only on the positions of the heavy atoms, this first Fourier synthesis often fails to reveal the whole structure of the molecule; instead, it yields only a partial structure, but this is generally enough to lead to better phases for the next Fourier synthesis. The structure should be solved after two or three (at most six or seven) of these successive Fourier syntheses of electron density. This step can be automated. The maxima of a Fourier synthesis can be analyzed by computer; the only maxima accepted as atoms are those that form a cluster with chemically meaningful geometries [73]. The combination of this technique with direct methods is discussed further on.

Special difficulties arise when the heavy-atom structure, taken separately, contains one more symmetry element (usually a center of symmetry) than the overall structure. The first Fourier synthesis of the electron density then shows two light-atom structures along with the heavy-atom structure. In order to calculate the second Fourier synthesis, the heavy-atom structure is then used together with only that part of the light-atom structure that can be assigned with high confidence to just one of the two light-atom structures indicated in the Fourier synthesis. The additional symmetry element vanishes in the next Fourier synthesis, and the structure analysis proceeds in the usual way.

2.2.2. Direct Methods

As Equation (17) implies, the distribution of the electron density in a crystal structure is unambiguously determined once the structure factors have been obtained. A structure factor comprises the structure amplitude and the phase angle; as has already been shown, all that can be determined by experiment is generally the structure amplitude, not the phase angle φ. The preceding section discussed a fairly straightforward way of using the heavy-atom method to gain knowledge of a substructure, and using this to determine the φ values indirectly from the Patterson synthesis.

As the theory of Patterson synthesis shows, it must be generally possible to infer the phases, given the associated structure amplitudes.

The best-known equation for direct phase determination is that of David Sayre [78]:

$$F(h) = \frac{1}{V}\frac{f}{g}\sum_{h'} F(h')\,F(h-h') \tag{30}$$

where $(1/V)(f/g)$ can be regarded as merely a kind of scaling factor and h, h' and $h-h'$ are vectors in reciprocal space, i.e., each is a triplet of indices (h, k, l). Equation (30) is not a probability relation, but a true equation; all atoms are assumed identical, and all combinations of structure factors are required. Sayre used this formula to elucidate a centrosymmetric projection of the crystal structure of hydroxyproline [78].

Multiplying both sides of Equation (30) by $F(-h)$ yields:

$$|F(h)|^2 = \frac{1}{V}\frac{f}{g}\sum_{h'} F(-h)\,F(h')\,F(h-h') \tag{31}$$

For large $F(h)$, the right-hand side is also large, real, and positive. If $F(h)$, $F(h')$, and $F(h-h')$ are large, the following probability statements then hold:

$$s(h)\,s(h')\,s(h-h') \approx +1 \tag{32}$$

in the centrosymmetric case, and:

$$\varphi(-h) + \varphi(h') + \varphi(h-h') \approx 0 \pmod{2\pi} \tag{33}$$

in the noncentrosymmetric case. Here, $s(h)$ denotes the sign and $\varphi(h)$ the phase of $F(h)$, and similarly for the other terms. Given the signs or phases of two structure factors, it is thus possible to infer the sign or phase of a third.

The form in which the Sayre equation is commonly used is due to Hughes [79]:

$$E(h) = N^{1/2}\langle E(h')\,E(h-h')\rangle_{h'} \tag{34}$$

where N is the number of atoms in the unit cell and $\langle\rangle_{h'}$ denotes the average over all h'; $E(h)$ is the normalized structure factor defined as:

$$|E(h)|^2 = \frac{|F(h)|^2}{\sum_{j=1}^{N} f_j^2} = \frac{I(h)}{\bar{I}} \tag{35}$$

where the sum of the squares of the atomic form factors f is taken over all N atoms in the unit cell. This is the expected value of the intensity of a reflection; $—F(h)—$ is the structure amplitude after absolute scaling is applied (see Eq. 58); $I(h)$ is the intensity of the reflection (h) and I the average intensity at the glancing angle of (h).

Equation (32) is a probability relation. The probability that the positive sign is correct is:

$$P_+(h) = \frac{1}{2} + \frac{1}{2}\tanh\left(N^{-1/2}\left|E(h)E(h')E(h-h')\right|\right) \tag{36}$$

It should be noted that the probability of the calculated phases depends not on the absolute intensities, but on the relative intensities (Eq. 35) that describe the scattering power of a structure with point atoms. Basically, the highest E values are the reflections containing the largest amount of information. The more atoms in the unit cell, the lower the probability that relation (32) holds. For structures where the unit cell contains atoms differing in atomic number:

$$P_+(h) = \frac{1}{2} + \frac{1}{2}\tanh\left(\sigma_3\sigma_2^{-3/2}|E(h)|\sum_{h'} E(h')E(h-h')\right) \tag{37}$$

for all products $E(h')E(h-h')$ available for $E(h)$, and:

$$\sigma_n = \sum_{j=1}^{N} Z_j^n \tag{38}$$

Using Equation (37) for the probability of relation (32) takes care of the fact that not all products required for the Sayre equation are available. A strict equation has become a probability relation. One of the corresponding equations for the probability that relation (33) is correct in the noncentrosymmetric case reads as follows:

$$P(\varphi(-h)+\varphi(h-h')+\varphi(h')) = \frac{\exp\left[2\sigma_3\sigma_2^{-3/2}|E(-h)E(h')E(h-h')|\cos(\varphi(-h)+\varphi(h-h')+\varphi(h'))\right]}{2\pi I_0 2\sigma_3\sigma_2^{-3/2}|E(-h)E(h')E(h-h')|} \tag{39}$$

where I_0 is a Bessel function of the second kind [80].

The Hughes form of the Sayre equation (34) is for the centrosymmetric case; in the noncentrosymmetric case, at least as far as physical information is concerned, it corresponds to the tangent formula. In contrast to Equation (33), here the sum is weighted with the normalized structure factors, and all pairs of structure factors possible in a special case are taken care of:

$$\tan\varphi(h) = \frac{\sum_{h'} K(hh')\sin[\varphi(h')+\varphi(h-h')]}{\sum_{h'} K(hh')\cos[\varphi(h')+\varphi(h-h')]} \tag{40}$$

where:

$$K(hh') = 2\sigma_3\sigma_2^{-3/2}\left|E(-h)E(h')E(h-h')\right| \tag{41}$$

A measure of the reliability of a phase obtained in this way is $\alpha(h)$, where:

$$[\alpha(h)]^2 = \left[\sum_{h'} K(hh')\cos(\varphi(h')+\varphi(h-h'))\right]^2 + \left[\sum_{h'} K(hh')\sin(\varphi(h')+\varphi(h-h'))\right]^2 \tag{42}$$

These equations are also used in the centrosymmetric case for determining the signs of structure factors (only the cosine terms are required).

In summary, it is probable that the sum of the phases for two reflections equals the phase of a third reflection, if the sums of the indices of the first two reflections are the indices of the third, and if all three reflections are relatively strong. The probability that this holds can be calculated. It is also possible to determine the probability for one phase when many phase combinations exist.

Table 2. Patterson maxima∗ for the space group $P2_1/c$ with one heavy atom in the asymmetric unit

Atomic coordinates	x	x	$-x$	$-x$
	y	$\frac{1}{2}-x$	$-y$	$\frac{1}{2}+y$
	z	$\frac{1}{2}+z$	$-z$	$\frac{1}{2}-z$
x	0	0	$2x$	$2x$
y	0	$\frac{1}{2}+2y$	$2y$	$\frac{1}{2}$
z	0	$\frac{1}{2}$	$2z$	$\frac{1}{2}+2z$
x	0	0	$2x$	$2x$
$\frac{1}{2}-y$	$\frac{1}{2}-2y$	0	$\frac{1}{2}$	$-2y$
$\frac{1}{2}+z$	$\frac{1}{2}$	0	$\frac{1}{2}+2z$	$2z$
$-x$	$-2x$	$-2x$	0	0
$-y$	$-2y$	$-\frac{1}{2}$	0	$-\frac{1}{2}-2y$
$-z$	$-2z$	$-\frac{1}{2}-2z$	0	$-\frac{1}{2}$
$-x$	$-2x$	$-2x$	0	0
$\frac{1}{2}+y$	$\frac{1}{2}$	$2y$	$\frac{1}{2}+2y$	0
$\frac{1}{2}-z$	$\frac{1}{2}-2z$	$-2z$	$\frac{1}{2}$	0

∗ Vector space coordinates are differences between pairs of atomic coordinates.

As Equation (34) shows, phases must be known or assumed at the start of a phase determination, before the equations can be used at all. Depending on the space group or Bravais lattice type, the phases of up to three reflections can be freely chosen; this fixes the origin of the coordinate system [81–84]. In the case of noncentrosymmetric space groups or those without mirror symmetry, the phase of one further reflection must be restricted to a 180° angular range in order to fix one of the two enantiomers. Other phases can be represented by letter symbols (symbolic addition) [85], [86] or simply by angle values (multisolution) [87–89]. In the case of a center of symmetry, for every additional reflection with freely selected phase, only two possibilities (0° and 180°) must be taken into account; this is done by letting each resulting phase combination serve as the starting point for determination of a new phase set. In noncentrosymmetric space groups, on the other hand, an infinite number of possibilities between 0° and 360° would have to be examined, but experience has shown that four values per structure factor (45°, 135°, 225°, 315°) are enough to place an adequate restriction on one phase for the determination of other phases. If there are n additional reflections with freely selected phases, the number of phase sets will be 2^n in the centrosymmetric case and 4^n in the noncentrosymmetric case. If the phases have been calculated in symbolic form, the stated possibilities must be permuted over all the symbols to yield the same number of phase sets as in the latter cases (unless relations can be established between the symbols, as is very often the case).

In another direct phasing method, a random number generator is used to supply phases for a relatively large number of reflections. "Figures of merit" are calculated in order to decide whether the calculation of a phase set should be completed. With present-day fast computers, 20 000 or more such phase sets can be calculated. Because one starting phase set contains a great number of reflections, there is no longer any need to fix the origin, or select one enantiomer through the phases of some reflections [90].

Ultimately, the only way to decide whether a phase set is the right one is to look for a chemically meaningful interpretation of the Fourier synthesis of the electron density. But it would be rather tedious to check, say, 64 Fourier syntheses in this way (256 or more in troublesome cases). An attempt is therefore made to find figures of merit [91] that will make one phase set more probable than the others, even before the Fourier synthesis is calculated. Such a figure of merit might refer to the internal consistency of a phase set. The determination of phases or signs by direct methods then involves the following steps:

1) Calculate the magnitudes — E — of the normalized structure factors
2) Rank the E values (decreasing order)
3) List pairs of reflections that can figure in a Sayre relation with a given reflection
4) For the starting phase set, select the reflections defining the origin and a few additional reflections
5) Determine the phases of each phase set
6) Select the phase sets that appear best under given criteria (figures of merit)
7) Calculate the Fourier synthesis of the electron density (E map) and interpret it

Different systems of weighting for phase determination and an almost countless number of figures of merit have been developed [92–94].

H is now possible to solve rather large crystal structures, including small proteins with more than 1000 non-hydrogen atoms, by real/reciprocal space Fourier recycling methods [95–99]. These are very lenghtly procedures, and a powerful PC may require four weeks or more without breaks. These major problems usually involve measuring synchrotron data.

2.2.3. Trial-and-Error Method, R Test

The trial-and-error method used to be employed for simple inorganic structures, but it is no longer important.

The R value:

$$R_1 = \frac{\sum |F(hkl)|_{\mathrm{obs}} - |F(hkl)|_{\mathrm{calc}}}{\sum |F(hkl)|_{\mathrm{obs}}} 100\ [\%] \tag{43}$$

indicates how close the solution approaches the correct structure. The sums are taken over all measured structure factor magnitudes. The $—F—_{\mathrm{obs}}$ are observed structure amplitudes, while the $—F—_{\mathrm{calc}}$ are the amplitudes calculated from Equation (18). If a least-squares fit is also done, a second R test becomes important; this includes the weights w, a function based on the counting statistics of the measurements, assigned to the squares of the differences between observed and calculated structure amplitudes:

$$R_2 = \tag{44}$$

$$\left[\frac{\sum w\left(|F(hkl)|_{\mathrm{obs}} - |F(hkl)|_{\mathrm{calc}}\right)^2}{\sum w\,|F(hkl)|^2_{\mathrm{obs}}}\right]^{1/2} 100\ [\%]$$

A value of $R < 50\ \%$ generally indicates that one is on the right path toward a solution of the structure. In particularly simple cases, space-filling considerations and the requirements of a highly symmetric space group may limit the number of possible solutions.

2.2.4. Experimental Phase Determination

High-quality crystals, together with intense X-ray sources, such as synchrotron radiation (and, in especially simple cases, rotating anode generators), make it possible to find the phase angles of structure invariants (triple products or reflections with even indices) by experimental means. This supplemental information can be derived from intensity fluctuations that occur when a crystal is rotated about the reciprocal lattice vector of a net plane in reflection position, and another reciprocal lattice point additionally passes through the Ewald sphere. The results can no longer be interpreted by kinematic lattice theory; the dynamic theory must be employed [100], [101].

2.3. Least-Squares Refinement

The least-squares method is used to achieve the best possible fit of the structure model, found after phase determination, to the measured data. For every atom, in general, the least-squares calculation refines three positional parameters as well as one (isotropic) or six (anisotropic) temperature coefficients. There are also "overall" parameters such as the absolute scaling factor. Special cases may require further parameters, for example a factor for enantiomorphism.

Formerly the general practice was to minimize the quantity:

$$\sum \left(|F_{\mathrm{obs}}| - |F_{\mathrm{calc}}|\right)^2 \tag{45}$$

This approach has the disadvantage that no significance is accorded to "negative" intensities [103], which means that the zero point was too high.

Modern least-squares computer programs minimize:

$$\sum \left(F^2_{\mathrm{obs}} - F^2_{\mathrm{calc}}\right)^2 \tag{46}$$

Among other advantages of this procedure is that even low-intensity data sets can result in useful parameter refinements [50].

Least-squares refinement yields not only the structure parameters, but also their standard deviations. The R values defined above can serve as figures of merit if w represents the weighting function employed in the refinement. A more important point for assessing the quality of a structure analysis, however, is that the results fit with the wealth of experience of crystallographers and chemists, and deviate from prior knowledge only in justifiable cases.

Spectroscopic data must also be taken into account.

Constraints (fixed parameters) and restraints (partly fixed parameters with a specified allowance of deviation) have become important in the last few years [51]. In this way it is possible to extract as much "truth" as possible out of relatively bad data (originating from the "best" crystal which was available).

Another important problem is twinning, which usually gets apparent during parameter refinement. This "model" is very often more "true" than any model of disorder [102].

2.4. Determination of Absolute Configuration

Friedel's law states that:

$$|F(hkl)| = |F(-h-k-l)| \tag{47}$$

even for noncentrosymmetric space groups. When there is anomalous dispersion, however, Friedel's law no longer holds rigorously. The atomic form factors adopt complex values:

$$f = f_0 + f' + \mathrm{i}f'' \tag{48}$$

where f_0 is the atomic form factor with no allowance for anomalous scattering, f' is the real part of the anomalous scattering, and f'' is its imaginary part.

For the noncentrosymmetric case, the calculated magnitudes of $F(h,k,l)$ and $F(-h,-k,-l)$ are different, so that the absolute configuration of a molecule or atomic group can be determined. The effect is the greater the more strongly an atom scatters and the closer (on the long-wavelength side) the absorption edge of an atom is to the X-ray wavelength. Normally, when extremely accurate intensity measurements are not possible, an atom heavier than oxygen must be present in the molecule (or must be incorporated synthetically) for the configuration of the molecule to be unambiguously determinable. Naturally, the synthesis must exclude racemization.

There are several methods for determining the absolute configuration. After the structure has been solved, one can calculate which pairs of reflections $(h,k,l/-h,-k,-l)$ differ most in intensity, and then measure these reflections more accurately. This approach makes it relatively easy to decide between enantiomers. D. W. Engel has used the technique to determine the absolute configuration for compounds that do not contain atoms heavier than oxygen [104].

A more convenient method that does not call for further measurements is to obtain the best possible least-squares fits of both enantiomers to the measurements. The R_2 values (Eq. 44) are subjected to a χ^2 test, which yields the statistical confidence level for one of the two configurations. The confidences are normally well over 99 %. This method was proposed and developed by W. C. Hamilton [105].

Another method is to attach a factor to the imaginary part in Equation (48) and to introduce this factor as a parameter in the least-squares fit. If it refines to + 1, the configuration is correct; to − 1, the other configuration is [106]. A new formalism has been introduced [107].

2.5. Example of a Single-Crystal X-Ray Structure Analysis Solved by Direct Methods

To illustrate the technique, consider a compound for which two structures have been proposed (Fig. 37). Spectroscopic measurements (NMR, mass spectroscopy) have not yielded decisive results. The coupling constants and chemical shifts are too unusual, especially for proposed structure **2**, which further involves bond angles so far from the norm that no Dreiding model could be built. At this point in the structure analysis, if not earlier, X-ray examination becomes essential.

A glass-clear crystal measuring $0.45 \times 0.45 \times 0.2$ mm was mounted on the inside wall of an capillary with a tiny amount of grease, and the capillary was sealed. The capillary was glued to the glass fiber of a goniometer head in such a way that the primary beam avoided the fiber. Alignment patterns of a precession camera showed a very high probability that the specimen was a single crystal, because it yielded sharp reflections.

It is important to consider what type of X-rays to use. Photographic recording of patterns was done with Cu Kα radiation; diffractometer measurements were performed with Mo Kα radiation. The absorption coefficient μ was calculated as:

$$\mu = [0.68\,(\mu/\varrho)_{\mathrm{C}} + 0.06\,(\mu/\varrho)_{\mathrm{H}} + 0.26\,(\mu/\varrho)_{\mathrm{O}}]\,\varrho = 0.788\,\varrho \tag{49}$$

where the weighting corresponds to the mass fractions of the atoms in the compound.

Values of μ/ϱ by element and wavelength are given in the International Tables [29, vol. III, pp. 157–200]. The density ϱ is calculated from the molecular mass M, the number of molecules Z in the cell, Avogadro's number (6.022×10^{23}), and the cell volume V (in 10^{-30} m^3):

$$\varrho = \frac{10MZ}{6.022V}\,(\mathrm{g\,cm^{-3}}) \tag{50}$$

With a computer-controlled single-crystal diffractometer, the cell constants and the setting matrix for the diffractometer were automatically determined on 98 reflections [63]; an appropriate program for controlling the diffractometer was employed, and the measurements were subjected to least-squares refinement. The result was the triclinic space group P $\bar{1}$ with cell constants:

$a = 10.561\,(5)$; $b = 6.134\,(3)$;
$c = 10.418\,(6)\,(10^{-10}$ m)
$\alpha = 89.37°\,(4)$; $\beta = 64.08°\,(4)$; $\gamma = 89.46°$

The numbers in parentheses are the standard deviations derived from the least-squares calculations. The density, determined by suspension in aqueous K_2HgI_4, was 1.35 g/cm^3; this implies that the unit cell contains two molecules, i.e., the X-ray measurements give a density of 1.347 g/cm^3. Accordingly, the absorption coefficient is 1.06 cm^{-1}. The fraction of the radiation not absorbed can be calculated:

$$I/I_0 = \exp\,(-\mu d) \tag{51}$$

It is 93 % for the maximum path length d (0.67 mm) in the crystal, and 98 % for the minimal path length (0.2 mm). Thus, no absorption correction is required with Mo K α radiation; the lower intensity of the reflections compared with Cu K α radiation must, however, be taken into consideration (Eq. 14).

Diffracted X-ray intensities were measured on a computer-controlled diffractometer with a $\theta/2\,\theta$ scan mode [108]. The reference reflection was remeasured at the beginning and end of each cycle of 20 reflections. This reflection had to be carefully selected because it formed the basis for the relative scaling. At a maximum glancing angle of 28°, the resolution was 0.76×10^{-10} m [$=\lambda/(2\sin\theta)$].

Of the 2930 reflections to be measured, 2391 had X-ray intensities significantly above the background.

The five-point measurement used here (Fig. 38) is not so often employed today. Large mass storage devices make it possible to store all the measured points for a reflection profile, so that laborious corrections do not have to be applied during the measurements.

The standard deviation σ of the intensity of a reflection is derived from the counting statistics:

$$\sigma\,(hkl) = \sqrt{\mathrm{meas}_1 + \mathrm{meas}_3 + \mathrm{meas}_5 + 4\,\mathrm{meas}_2 + 4\,\mathrm{meas}_4} \tag{52}$$

The relative error $r\,(hkl)$ is given by

$$r\,(hkl) = \frac{\sigma\,(hkl)}{I\,(hkl)_{\mathrm{obs}}} \tag{53}$$

In the least-squares calculations, the reciprocal of $r\,(hkl)$ is needed for the weight function $w\,(hkl)$ in order to fit the structure parameters to the measured data.

The Lorentz–polarization factor correction had to be applied to the diffracted intensities:

$$I\,(hkl) = (\mathrm{LP})^{-1}\,I\,(hkl)_{\mathrm{obs}} \tag{54}$$

For the case under consideration:

$$\mathrm{LP} = \frac{1 + \cos^2 2\theta}{2\sin 2\theta} \tag{55}$$

The absolute structure amplitude is then calculated as:

$$|F\,(hkl)| = c\,\sqrt{I\,(hkl)} \tag{56}$$

The use of direct methods for phase determination requires not just the relative structure amplitudes, but also the absolute structure amplitudes (Eq. 56). These can be calculated with the aid of Wilson statistics [109]:

$$\bar{I} = \exp\left(-2B\sin^2\theta/\lambda^2\right) k \sum_{n=1}^{N} f_n^2 \tag{57}$$

$$\ln\left(\bar{I}/\sum_{n=1}^{N} f_n^2\right) = -2B\sin^2\theta/\lambda^2 + \ln k \quad (58)$$

The theory says that the average $\bar{I}$ of the diffracted X-ray intensities for a given glancing angle equals the sum of the squares of the atomic form factors of the N atoms in the unit cell. If the left-hand side of Equation (58) is plotted against $\sin^2\theta/\lambda^2$, the slope of the regression line is twice the overall temperature coefficient B, and the vertical-axis intercept is the natural logarithm of the scaling factor k (Fig. 39). The resulting values were $B = 4.3\times10^{-20}\ \mathrm{m}^2$ and $k = 0.32$. After parameter refinement, a scaling factor of 0.28 and an overall temperature coefficient of $4.8\times10^{-20}\ \mathrm{m}^2$ were obtained. Factor B can be described by:

$$B = 8\pi^2\bar{u}^2 \quad (59)$$

where $\bar{u}^2$ is the mean-square displacement of atoms from their average positions perpendicular to the diffracting net planes.

From Equation (35), the normalized structure amplitudes $|E|$ were obtained. These provided the basis for direct phase determination.

The subsequent calculations were done with the MULTAN 77 program system [110], which comprises six programs linked through data files. From the corrected intensities $I(hkl)$, NORMAL generates normalized structure amplitudes $|E(hkl)|$. Wilson statistics (or Debye statistics [111] if parts of the molecule are already known) are used to determine the scaling factor and the overall temperature factor, and to work out the statistics of the scattered X-ray intensities by reflection groups. The actual phase determination takes place in the MULTAN program. The 393 largest E values were used for phase determination, the maximum E being 3.5 and the minimum 1.0. First, all combinations of two reciprocal lattice vectors for each reflection were found: $H = H' + (H - H')$, where H stands for an index triplet (h, k, l). Processing continued with the 2000 largest triple products of $|E|$ values.

The critical question in phase determination is which reflections are assigned phases in the starting set. The case under study was centrosymmetric, so that only signs had to be allocated. The origin of the coordinate system was fixed by the reflections (7, 3, 3), (1, 1, 4), and (0, 1, 2). The four reflections (1, 1, 2), (8, 2, 2), (5, 0, 2), and (0, 5, 2) gave 16 sets of signs. The reflections were suggested by the program. The basic criteria for including a reflection in the starting set were the number, size, and type of triple-product relations. In more difficult cases this selection must be performed manually.

The program used the tangent formula (Eq. 40) to generate from the initial sets, sign sets with the fewest contradictions. Statistical criteria (figures of merit) were then applied to choose the most probable set of signs. The most probable set was that in which only (1, 1, 2) among all the starting reflections had negative sign. From these results, and the normalized structure amplitudes $|E|$, the programs EXPAND, FFT, and SEARCH were used to calculate an E map (a Fourier synthesis of the electron density of point scatterers) and to interpret it. Figure 1 and Table 3 present the result. Maximum 19 in Table 3 is markedly smaller than those which precede it. Comparison with Figure 37 reveals that strucural formula **2** is in agreement with the computer suggestion. The 18 nonhydrogen atoms of proposed structure **2** were assigned to the 18 strongest maxima of the Fourier synthesis. A structure factor calculation (Eq. 18) gave $R_1 = 28.6\,\%$ (Eq. 43).

The structure model obtained is chemically meaningful. The chemist's experience thus comes into play only when a possible solution to the phase problem has already been found. In contrast to indirect structure analysis methods (such as spectroscopy), the analysis does not start with structure model.

In the least-squares refinement the structure parameters (one isotropic temperature coefficient and three position parameters per atom, one scaling factor) had to be fitted to the 2391 observed reflections; unit weights were assigned to all reflections. When R_1 reached 14.3 %, every atom was assigned six temperature coefficients. The expression for the anisotropic temperature factor is:

$$\exp-\left(b_{11}h^2 + b_{22}k^2 + b_{33}l^2 + 2b_{12}hk + 2b_{13}hl + 2b_{23}kl\right) \quad (60)$$

After further refinement cycles, hydrogen atoms in chemically meaningful positions were assigned to the largest maxima in the difference

Table 3. Maxima in the E map plotted by computer

Peak	Peak height	Initial coordinates			Interpretation	Coordinates after least-squares refinement		
		x	*y*	*z*		*x*	*y*	*z*
1	999	0.0914	0.2955	0.1963	O (3)	0.0881	0.3031	0.2044
2	873	0.3008	0.4824	0.1738	C (9)	0.2904	0.4925	0.1766
3	871	−0.1507	0.4079	0.2851	O (2)	−0.1478	0.4049	0.2865
4	844	0.3264	0.3231	0.2409	C (14)	0.3287	0.3208	0.2401
5	842	0.1360	0.6191	0.0820	O (4)	0.1401	0.6100	0.0727
6	790	0.3636	0.6667	0.1563	C (10)	0.3715	0.6809	0.1391
7	771	−0.3413	0.1716	0.4126	O (1)	−0.3412	0.1781	0.4128
8	747	0.1515	0.5000	0.1563	C (8)	0.1671	0.4802	0.1443
9	743	0.4954	0.6820	0.1616	C (11)	0.4886	0.6936	0.1655
10	728	−0.0462	0.2833	0.1889	C (7)	−0.0403	0.2748	0.1869
11	703	−0.0724	0.0370	0.2020	C (1)	−0.0691	0.0334	0.2044
12	672	0.4485	0.3440	0.2635	C (13)	0.4467	0.3348	0.2647
13	561	−0.1418	−0.0400	0.3657	C (2)	−0.1393	−0.0384	0.3634
14	554	−0.2104	−0.0157	0.1946	C (4)	0.2051	−0.0301	0.1941
15	550	0.5288	0.5165	0.2304	C (12)	0.5265	0.5216	0.2272
16	536	−0.2746	−0.0438	0.3560	C (3)	−0.2813	−0.0325	0.3561
17	453	−0.1515	0.1667	0.4688	C (5)	−0.1499	0.1433	0.4641
18	426	−0.2326	0.3038	0.4336	C (6)	−0.2334	0.3031	0.4214
19	322	0.2147	0.3104	0.2263				
20	309	−0.0538	0.4083	0.0974				
21	225	0.6667	0.5556	0.2188				

Fourier synthesis of electron density. This is a Fourier synthesis in which the Fourier coefficients are the differences between the observed and calculated structure factors. When the hydrogen coordinates had been fitted to the measured data, the isotropic temperature coefficients of the hydrogen atoms were also refined. The R_1 was now 5.3 %. Finally, weighting the squared deviations in accordance with the counting statistics gave $R_1 = 4.7$ %. The weighted R_2 from Equation (44) was 3.6 %.

At this point, the least-squares refinement could be regarded as complete because the largest change in a parameter was smaller than 10 % of the corresponding standard deviation. In a difference Fourier synthesis of the electron density, the values for the 10 largest maxima were 0.12 – 0.16 electrons/10^{-30} m^3.

The software used was MULTAN 77 [110]. There are a number of more up-to-date program systems:

SIR 92 [42]
DIRDIF [43]
MULTAN 88 [44]
NRCVAX [46]
XTAL [47]
ORTEP II [48]
SHELXS 86 [49]
SHELXL 92 [50]
SHELX 97 [51]
CRYSTALS [52]

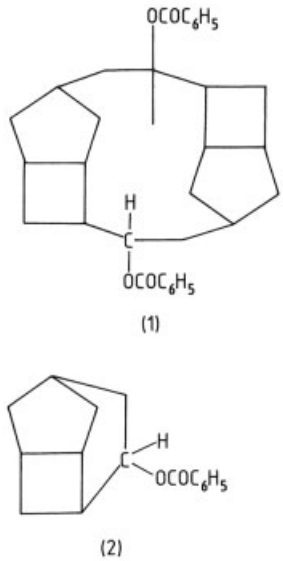

Figure 37. Two suggested structures for the compound under study

Representations of the molecule are shown in Figures 40 (configuration and bond lengths), 41 (stereoscopic view), and 42 (projection of

unit cell content along the *a*-axis). If the bond and torsion angles (not shown here) are considered, along with the bond lengths, it becomes clear that major difficulties arise in determining the molecular structure by spectroscopic means alone. The results of the structure analysis have been published elsewhere [112].

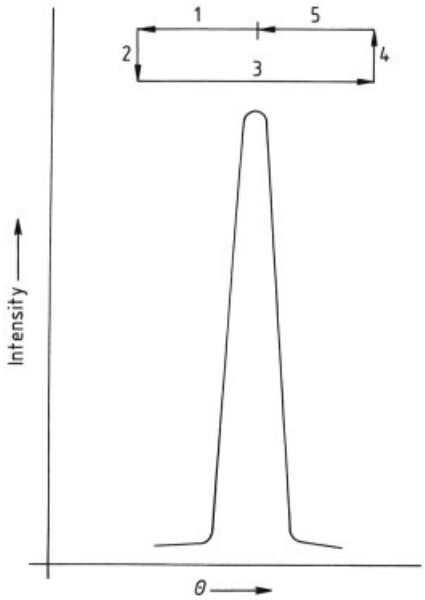

Figure 38. Steps in the five-point measurement

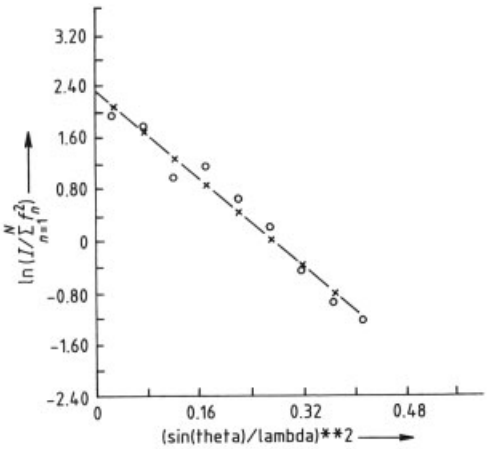

Figure 39. Computer plot of results from Wilson statistics

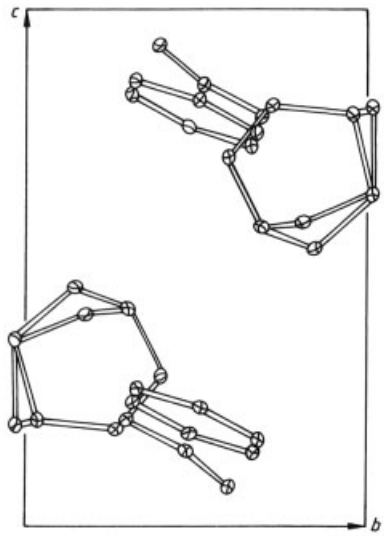

Figure 42. Projection of unit cell along the *a*-axis

2.6. Diffraction by Polycrystalline Specimens

A substance is characterized as polycrystalline if its crystalline particles cannot be distinguished with the naked eye (<0.01 mm). Larger particles, if any, must be ground before the measurement, provided this does not adversely affect the crystal structure of the substance.

2.6.1. General Considerations

If the apparatus ordinarily used to record diffraction patterns of polycrystalline specimens is compared with that used for single-crystal diffractometry, the following essential differences are seen:

1) Instruments for powder diffractometry are much simpler. If an "ideally polycrystalline" specimen is assumed, and if a one- or two-dimensional detector with appropriately large angular aperture is employed, there is no need to rotate either the specimen or the detector. Given an ideal statistical distribution of crystallites in the specimen, together with not too small a specimen size, it can be assumed that every orientation of a crystallite relative to the incident X-ray

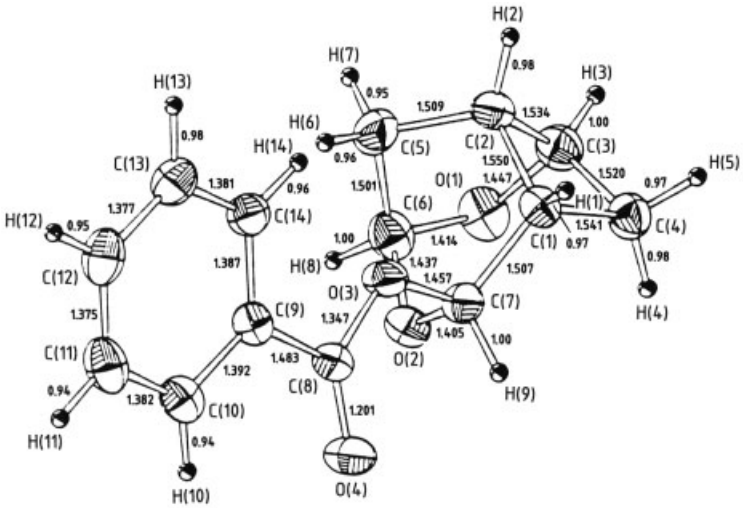

Figure 40. Molecular structure of (1 RS, 3 SR, 4 RS, 6 SR, 8 SR)-5,9-dioxatricyclo[4.2.1.0^{3,8}]-non-4-yl benzoate. Lengths in 10^{-10} m. Standard deviations $\bar{\sigma}$: 0.002×10^{-10} m for O – C, 0.002×10^{-10} m for C – C, 0.01×10^{-10} m for C – H

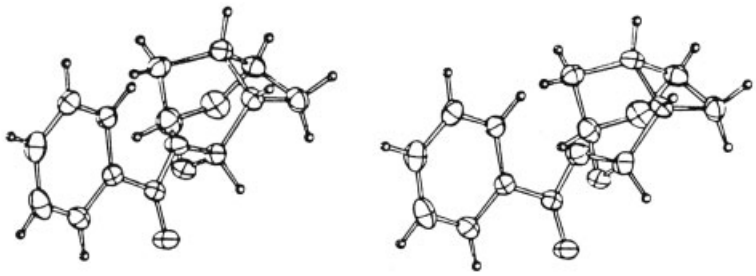

Figure 41. Stereoscopic view

beam occurs in the specimen (as a statistical average). A rotation about one axis is enough to get a quite substantial increase in the scattered X-ray intensities, because then a large number of crystallites contribute to the intensity for a selected net plane. Single-axis goniometers used for this purpose are the Debye – Scherrer and the transmission diffractometers.

2) The specimen sizes employed are much larger than those in single-crystal studies. Because ideally there must be a statistical distribution of crystallites, only a small fraction of the specimen can contribute to a given reflection.

One of the problems in X-ray diffraction by a polycrystalline powder is that the specimen is not "uniformly polycrystalline". The crystallites vary widely in size and their relative orientations are not random. The predominance of one habitus or another is crucial in the preparation of the specimen; for example, needles line up parallel to one another, while plates align themselves parallel to the specimen surface. In principle, this effect can be diminished by grinding the specimen and screening out the coarse particles, but there are two disadvantages to this:

1) Organic substances, in particular, are so sensitive that the crystal structure is disrupted by grinding, the specimen undergoes a change in degree of crystallinity, or there may even be a phase transformation
2) The specimen thus treated is no longer in the as-delivered condition that was to be characterized

If one has a phase mixture, demixing is possible, because the components often differ in their grinding characteristics.

Far better than excessive size reduction is to rotate the specimen appropriately and integrate over all directions in space. These ideal conditions can be approximated if an area detector is employed.

2.6.2. Qualitative and Quantitative Analyses of Powder Specimens

The X-ray diffraction pattern of a polycrystalline specimen is a plot of diffracted X-ray intensity against diffraction angle. Because the diffracted intensity is a single-valued function of the unit cell contents (Eqs. 15, 18), and the diffraction angle is a single-valued function of the unit-cell dimensions (Eq. 2), an identical unit cell with identical contents must also yield an identical diffraction pattern. The diffraction pattern is thus a genuine fingerprint.

But the fingerprint also contains information on the macroscopic properties of a specimen. The half-value widths of the reflections are inversely proportional to the crystallite sizes, and directly proportional to the concentration and magnitude of structural defects. The coherent scattered component of the background results from amorphous constituents in the specimen, while the incoherent scattering is chiefly due to inelastic atomic scattering.

Determination of Cell Constants, Crystal System, and Space Group. The requirements on specimen preparation are a bit less demanding when a powder pattern is used to determine the cell constants only as long as lines are not fully missing, because preferred orientation is so heavy. Unfortunately, an unknown unit cell cannot always be determined in this simple way. The method is most suitable for the cubic case. Squaring the Bragg equation gives:

$$\sin^2\theta = \frac{\lambda^2}{4d^2} \tag{61}$$

For the cubic crystal system:

$$1/d^2 = \frac{h^2+k^2+l^2}{a^2} \tag{62}$$

where a is the cell constant. It follows that:

$$\sin^2\theta = \frac{\lambda^2}{4a^2}\left(h^2+k^2+l^2\right) \tag{63}$$

For a cubic substance, the constant $\lambda^2/4\,a^2$ can normally be identified directly if the $\sin^2\theta$ values of the reflections are ranked in ascending order of magnitude. The value of $\lambda^2/4\,a^2$ can be recognized and all $\sin^2\theta$ values can be obtained by multiplying by a sum of at most three squares.

Unambiguous results are, however, obtained only if the errors of measurement are small. It can never be excluded with utter confidence that slight deviations from, say, the cubic system occur. Absolutely certain results come only from single-crystal patterns, which additionally reveal the symmetry.

The powder patterns of substances belonging to other crystal systems can be indexed on the basis of analogous regularities, and thus assigned to a unit cell. The method of T. Ito even makes it possible to index the diffraction spectra of triclinic compounds [113]. The precondition is that many discrete reflections must be present, and their glancing angles must be determinable with great accuracy. Organic compounds commonly do not fulfill these conditions. Large cell constants lead to overlapping of reflections at relatively small glancing angles; many crystal structure defects and large temperature factors have the same effect because they increase the half-value widths of the reflections.

Special difficulties arise in determining the cell constants of high polymers. Using single crystals instead of crystalline powders is not ordinarily an option because in most cases there are principally no single crystals. In most cases, however, filaments can be drawn from the substance; after appropriate heat treatment, such filaments tend to take on a single-crystal character. The reflections are then located on layer lines in flat-film patterns. With some luck, trial and error will lead to not only the translational period in the fiber direction, but also the other cell constants [1, pp. 405 – 406].

If provisional cell constants have been determined for a specimen, all reflections can be indexed. There must be no common divisor for all the indices h, k, and l (primitive cell). Now the reduced cell (whose edges are the three shortest noncoplanar translational periods of the lattice) can be calculated. From the Niggli matrix:

$$\begin{pmatrix} \boldsymbol{a}\cdot\boldsymbol{a} & \boldsymbol{b}\cdot\boldsymbol{b} & \boldsymbol{c}\cdot\boldsymbol{c} \\ \boldsymbol{b}\cdot\boldsymbol{c} & \boldsymbol{c}\cdot\boldsymbol{a} & \boldsymbol{a}\cdot\boldsymbol{b} \end{pmatrix} \tag{64}$$

which is formed from the scalar products of the cell constants, it is possible to deduce the crystal system and Bravais lattice [4, pp. 124 – 159].

With the final cell constants, the crystal system derived from them, and the Bravais lattice, the reflections can be indexed. The space group can be identified from systematic absences [19, pp. 21 – 53].

As the preceding discussion implies, the determination of cell constants from powder patterns involves large uncertainties, even more so for the determination of the crystal system and space group, because the diffraction pattern does not directly lead to the symmetries, which are instead deduced from measured angles which include experimental errors, e.g., 90° for one of the cell constants. The case is similar with systematic extinctions: Lack of resolution can result in uncertainty or a wrong conclusion.

A better procedure is the use of powder patterns for crystalline phases whose cell constants are known. Changes in the cell under given conditions can be tracked in a convenient and reliable way. The cell data can be fitted to the measured data by least-squares methods.

Phase Analysis. If a specimen is crystalline, an obvious way to study and identify it is by using X-ray diffraction patterns. The X-ray diffraction pattern of a crystalline compound functions as a fingerprint that should permit unambiguous identification. Isomorphous crystalline compounds often create major problems; elemental analysis can be of aid here.

The case is even more difficult when the specimen for analysis is a mixture of crystalline phases differing markedly in the content of the individual components. This problem may be so aggravated that one of the crystalline phases is represented solely by its strongest reflection. Other methods of investigation are then needed in order to verify this crystal phase. All that X-ray diffraction can do in such instances is to make the presence of a crystalline phase probable.

If two isomorphous crystalline compounds cannot be distinguished from the X-ray diffraction pattern, neutron diffraction is an option. The atomic form factors for neutron diffraction and X-ray diffraction are completely different, so that the reflections can be expected to differ in intensity.

The successful use of X-ray diffraction to analyze unknown substances and mixtures of substances requires a knowledge of many X-ray diffraction patterns. The International Center for Diffraction Data (ICDD, 1601 Park Lane, Swarthmore, PA 19081, United States) has taken on the job of compiling X-ray diffraction patterns. The tabulated data can be obtained in the form of data cards, microfiches, compact disc (CD) or magnetic tape, together with a search program. In the first two cases, the X-ray diffraction data can be searched by means of tables ordered via lattice plane distances and relative intensities. The continually updated compilation concentrates on inorganic substances. The collection of data for organic compounds is far less complete. Furthermore, analysis is made more difficult by the poorer crystal quality and larger cell constants of organic compounds. Both these factors tend to push the principal reflections to relatively small glancing angles, so that the net plane distances must be assigned relatively large errors.

Quantitative Analysis of Crystalline Phases. The intensity of a reflection is directly proportional to the quantity of the crystalline phase in a mixture, but it is also a function of the crystal structure. Comparing the intensities of two reflections is therefore not enough to determine the mass ratio of two phases in a mixture; the constant of proportionality must also be found.

The following technique can be used in the case of a two-phase mixture in which the percentage content of one phase is to be determined. In a number of specimens of known composition, the intensity ratio for the two crystalline phases is plotted against the mass ratio. The dependence is linear only if both crystalline phases have similar absorption coefficients, which is roughly so for, say, crystal modifications of the same compound.

A method for the absolute quantitative determination of crystalline phases in a multiphase mixture is to mix a standard substance into the specimen. Calibration curves are necessary for mixtures of the standard with each of the phases to be determined. The components can then be determined independently of each other. The critical point is to determine accurately the integral intensity of a sufficiently strong reflec-

tion that does not have any other reflection superposed on it. With a diffractometer linked to a computer, it is possible to separate overlapping reflections and determine their intensities, although some idealizations are required.

When a phase mixture contains just a few percent ($< 8\,\%$) or even fractions of a percent of the phase being analyzed a more accurate result can be obtained by recording X-ray diffraction patterns for specimens that lie in the composition range of interest. These allow the content of the strongly deficient phase to be estimated. The results are often distorted because it proves quite difficult to find a suitable standard for the specimen being measured. The standard should have the same crystal quality and crystallite size as the substance to be determined; otherwise, errors in the results are inevitable. These idealized conditions are scarcely ever fully achievable.

2.6.3. Solid Solutions

Compounds are normally miscible in all proportions in the gas phase, but in the liquid phase two or more substances must have certain chemical properties in common if they are to be miscible. The situation is much more complicated in the solid phase. Here, miscibility demands not just chemical affinity, but also a similar manner of space filling. For the formation of a continuous series of solid solutions, a molecule or atom of one phase must be arbitrarily replaceable by a molecule or atom of the other one. The formation of solid solutions has a relatively great effect on the cell volume, and this effect can be used for the very exact determination of a compound in trace element analysis (e.g., quantitative determination of trace impurities in semiconductor crystals).

Very often, solid solutions can be formed only over a limited range. The host lattice tries to adapt, more or less, to the guest until a point is reached where the stability limit of the host lattice is passed. Variations of the lattice parameters in the stability range are often controlled by the Vegard rule, which says that variations of parameters are proportional to the content of foreign components. The validity of this rule can, conversely, be regarded as evidence of solid-solution formation. It is generally sufficient if proportionality can be demonstrated for a few net plane distances.

The use of computers to control the diffractometer and interpret the measurements has proven itself here in particular. Not only can glancing angles be determined accurately, but any new reflections that appear can be identified with sufficient statistical reliability.

2.6.4. Rietveld Method

Measurements on crystalline powders lead to a solution of only relatively small and simple structures: Most of the reflections must be resolved, so that intensities can be associated with them unequivocally. Even theoretically, this is often impossible, e.g., when several reflections occur at precisely the same glancing angle. In the cubic crystal system this is so when the sum of the squares of the indices does not permit a single indexing; e.g., (221) and (300), where the sum of the squares of the indices is 9 for each case. Measurements on crystalline powders that give unambiguous values for all reflections should be inherently preferable to single-crystal measurements, because many experimental difficulties (e.g., extinction and absorption) either do not occur at all, or are easier to handle.

This is theoretically correct, but this ideal case never occurs. The atomic resolution is always lower. In general a single-crystal structure analysis is always better and gives more significant data, nevertheless the single crystal might be very poor. The powder method is used when it is not possible to obtain sufficiently good single crystals.

The strength of the Rietveld method lies in the refinement of the structure parameters of the whole powder sample not just of one single crystal. Rietveld [114] recognized that it is by no means necessary to allocate the measured intensities to individual reflections in a one-dimensional powder pattern. Instead, it is better to measure a diffraction pattern as exactly as possible and use it in its entirety as the basis for refinement of the structure parameters. Parameters such as the atom coordinates, temperature coefficients, and scaling factors are fitted to the diffraction data by a least-squares procedure.

This method is particularly important for neutron diffraction analysis, where in general a pre-

vious single-crystal X-ray structure analysis has led to a structure model that is satisfactory within certain limits. The powder method can then be used, for example, to examine the magnetic properties of a substance in more detail, without having to grow large single crystals.

Rietveld analysis is especially valuable for the quantitative determination of components in polytypic mixtures. If the structures of the polytypes are known, the only parameters to be determined are the temperature coefficients, the scaling factors, and the cell constants. Figure 43 shows an analysis of SiC. The specimen contains 11.2 % of polytype 4 H, 85.2 % of 6 H, and 3.6 % of 15 R. The profile *R* value is ca. 6.7 %. Another very important application is the study of solid-solution systems. As Figure 44 shows, the content of an additional phase can be estimated when there are gaps in the series of solid solutions [115]. The profile *R* value in this case was 2.67 %.

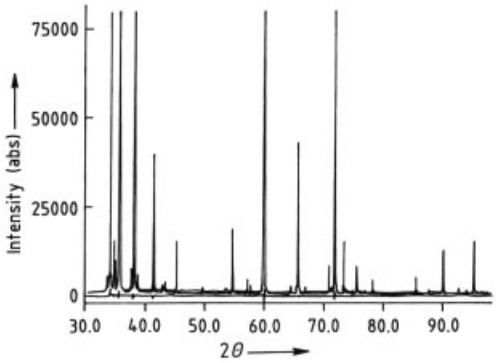

Figure 43. Diffraction pattern of a polytypic mixture of SiC

In the past this method could only be used when the overall crystal structure was known and only special characteristics had to be explored. Now many methods have been developed to perform quasi-a-priori crystal structure analyses of crystal powder samples. Computer programs for single-crystal work often have an option for powder structure analysis (e.g., [42], [50]). They need the successful extraction of intensities of as many as possible single reflections out of the powder diagram, what is often performed by the method of Le Bail [117]. This method is also used by POSIT, in which only low-order reflections are necessary to determine a model from known fragments of the unknown structure for the Rietveld refinement [118]. The computer program Powder Cell is a complete program system for solving crystal structures from powder data [53]. Fully another approach is Monte Carlo/simulated annealing [119]. Only an approximate starting model with a few internal degrees of freedom is necessary. Thousands of diffraction diagrams are calculated and compared with the experimental one. The software package Powder Solve is fully integrated within the Cerius^2 molecular modeling environment [54]. A further approach, quite different from the above, is the ab initio prediction of possible molecular crystal structures by energy minimization. No cell determinations, no indexing, and no extraction of intensities of single reflections are necessary [53], [120–122], [124].

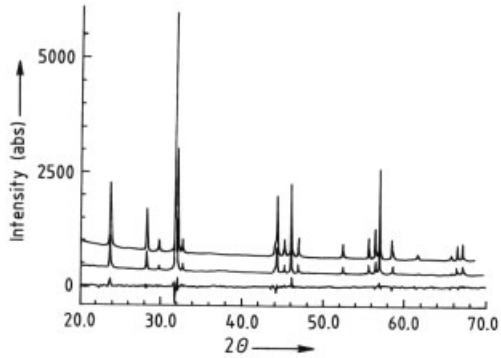

Figure 44. Diffraction pattern of a solid solution and a mixed phase: 12.9 % $Nd_{0.4}Ce_{0.6}O_{1.8}$, 87.1 % $Nd_{1.85}Ce_{0.15}CuO_4$

2.6.5. Texture

Up to this point, two distinct types of specimen, single crystals and polycrystalline powders, have been discussed. In the analysis of a single crystal, it is essential that the specimen contains no polycrystalline material; in polycrystalline powders, one condition for X-ray examination is that the directions of the net planes have a uniform statistical distribution in space, otherwise more or less complicated corrections must be applied, perhaps by preparation or by computer programs. An intelligent use of preferred orientation for structure analysis is recording the diffraction diagram by an area detector. The reflections are partly spread out into two dimensions, many more reflections

are resolved, more reflections can be used for determining the cell constants, and overlap of intensities of reflections is diminished [116].

The properties of many substances, however, exhibit a direction dependence, which is similar to a single crystal, but at the same time indicates that the substance has only partial single-crystal character. Such a solid is said to have texture. The individual crystallites are fully or partly aligned; they are said to have preferred orientations.

Texture in a polycrystalline structure can arise in quite varied ways, e.g., deformation texture, casting texture, recrystallization texture, plate texture, or fiber texture.

Texture studies on wires and plates are very important. Initial information can be acquired from transmission and reflection patterns recorded with a flat-film camera (see Fig. 15). The type and degree of texture can be determined more or less exactly, depending on the object of investigation and the type of instrument used. One of the best-developed instruments is the Lücke texture goniometer [125]. With this apparatus, the distribution of given net planes can be directly imaged in a pole figure.

The fully automated, computer-controlled X-ray texture goniometer sold by Seifert allows reflection and transmission measurements; the background is measured and the appropriate correction applied automatically. Other corrections, e.g., absorption, are also performed automatically. The pole figure is output to a terminal. Several reflections can be measured in a fully automated programmed sequence.

2.7. Crystallographic Databanks

X-ray studies typically generate large amounts of data. Specimen synthesis and preparation, data acquisition, and interpretation of results are generally so expensive that repeating a structure determination is a waste of resources, unless there are compelling reasons to do so. For this reason, databanks have been set up and made available to scientists worldwide. Unfortunately, using these databanks generally incurs high costs; the importance of structure studies to the advance of the natural sciences suggests that the scientific community ought to be trying to make such data freely available to all. The following databanks deserve mention:

Powder Patterns. PDF (Powder Diffraction File) Databases [38]

Crystal Structures of Organic and Organometallic Compounds. CSDS (Cambridge Structural Database System) [37]

Crystal Structures of Inorganic Compounds. Inorganic Crystal Structure Database [39]; MDAT (minerals) [40]

Crystal Structures of Proteins. Protein Data Bank [41]

2.8. Noncrystallinity

The diffraction of X-rays by solid matter is only possible when a certain regularity of structure exists. Scattering does not become diffraction unless the material contains equal and repeated distances. These conditions are fulfilled by a crystal lattice. If the crystallites are too small, or the crystals contain structural defects, the coherent scattering domains become smaller and the reflections become broader; this broadening may be so great that it leaves only an amorphous halo, often with a half-value width of several degrees, as is the case with quartz glass, for example. In addition, the halo is often centered near the highest intensity reflection of the corresponding crystalline phase. A similar effect occurs in solids that do not have this continuous gradation from crystalline to amorphous, but in which similar atomic distances recur. Some amorphous substances give no halo at all. If the particles in a uniform host matrix are large ($20–1000 \times 10^{-10}$ m and above) scattering takes place at small angles ($< 0.5°$), where profile forms can yield information about the particle shape. The precondition for such studies is that the particles and the enclosing matrix differ sufficiently in electron density. Several problem areas are dealt with in somewhat more detail.

2.8.1. Crystal Quality Analysis

The relations discussed above hold for an ideal crystal; in practice, crystals are ideal only to a

limited extent. In the following section, an attempt is made to indicate where a particular specimen should be placed in the range from ideal crystal to amorphous matter. Along with structural defects in the interior of the crystallites, or at their surfaces, the crystallite size also has to be estimated. A crystal may also contain stresses, or it may be that the single-crystal ordering is only partial, or applies in just one or two dimensions. Each of these factors affects the X-ray diffraction pattern.

2.8.2. Crystallite Size

A real crystal is not infinitely large. In the case of X-rays and neutrons, size does not begin to affect the diffraction pattern until the crystallites (or more precisely the coherent scattering domains) become smaller than ca. 0.5 µm. When electrons are used, this limit is nearly two orders of magnitude smaller (0.01 µm).

SCHERRER, in 1918, derived the following equation between the "physical half-value width" β (in degrees 2θ), i.e., that part of the broadening of a reflection due solely to inadequate crystallite size, and the crystallite size L:

$$\beta = \frac{K\lambda}{L\cos\theta} \tag{65}$$

Theoretically, L is the dimension of the crystallite perpendicular to the diffracting net planes; K is a constant, close to unity (often taken as 0.9); and λ is the wavelength of the radiation employed. The observed half-value width B is commonly assumed to be the sum of the instrumental broadening b and the physical broadening β. Some authors also give the formula:

$$B = \sqrt{\beta^2 + b^2} \tag{66}$$

It should be kept in mind that the error of determination is 20 – 30 %.

The instrumental broadening b can be obtained by calibrating with a substance with a large crystallite size (> 0.5 µm) [126].

2.8.3. Defects

Like excessively small crystallite size, crystal structure defects also increase the half-value width of a reflection. Such defects make the coherent scattering domains smaller. The amount of broadening depends on the type of defect, because in general newly generated scattering domains remain coherent with one another across the new disruptions.

It is very difficult to separate the broadening of a reflection into components due to small crystallite size and to structural defects. Equation (65) implies that $\beta\cos\theta$ should be constant for all glancing angles, if the broadening results entirely from small crystallite size. Because L is the dimension perpendicular to the diffracting net planes, this statement cannot hold rigorously, except for the diffraction orders of one and the same net plane. If line broadening is the consequence of crystal structure defects alone, the width increases with θ.

2.8.4. X-Ray Crystallinity

The use of X-ray methods to determine crystallinity in fibers is treated in → Fibers, 2. Structure. This can be extended to high polymers with no fiber structure, and so the discussion here is brief.

The two-phase model generally accepted for high polymers has amorphous and crystalline domains occurring next to each other. If the crystallinity obtained by X-ray examination isstated as a percentage, this figure should not be regarded as a serious value; it serves only for comparison within a series of experiments. In no case should comparison be made between the X-ray crystallinities of high polymers derived from different monomers.

It has become customary to extend the term X-ray crystallinity to compounds other than high polymers, in which case numerical figures should be treated with even more caution. For example, the properties of organic pigments depend very strongly on the crystal quality of the primary particles, i.e., the coherent scattering domains. Figure 45 presents X-ray diffraction patterns of organic pigments consisting of the same compound. Between A and C, a continuous transition (such as case B) is possible. It therefore appears meaningful to speak of a crystalline component, but numerical values can never be assigned in such a case, as they can for high

polymers, where portions of the X-ray diffraction pattern can be attributed to amorphous and crystalline parts (Fig. 46).

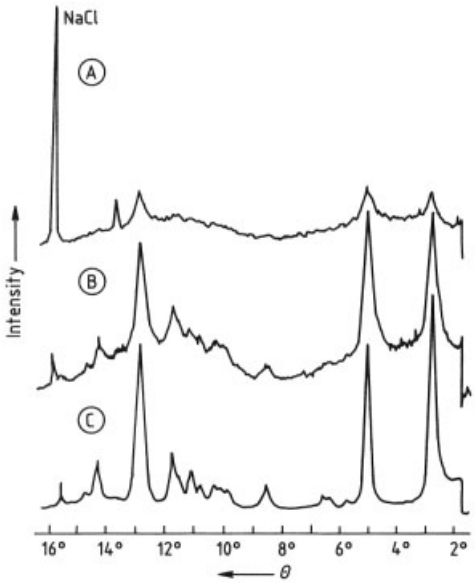

Figure 45. Organic pigments consisting of the same compound, but differing in crystal quality (patterns recorded with Cu K α radiation)
A) "Over-ground" pigment; B) Moderately well-finished pigment; C) Well-finished pigment

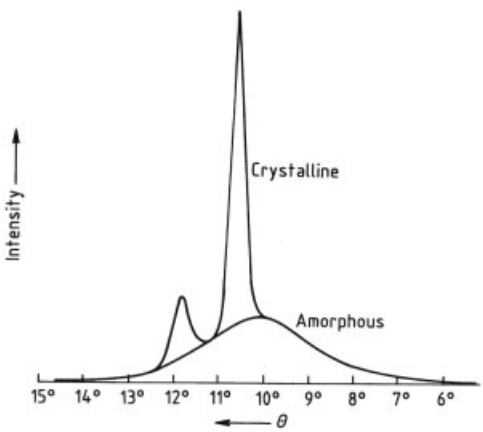

Figure 46. Amorphous and crystalline components in the X-ray diffraction pattern of a high polymer, recorded with Cu K α radiation

2.8.5. Elastic Stress

The measurement of elastic stress is especially important in the examination of steels. Elastic stress (residual macroscopic stress) expands the lattice in the direction of the tensile stress and contracts it in the perpendicular direction. The corresponding shifts toward lower and higher glancing angles, respectively, can be measured by X-ray methods. These lattice changes can be correlated with the elastic moduli E:

$$\sigma_\varphi = \frac{d_\psi - d_0}{d} \frac{E}{1+\nu} \frac{1}{\sin^2\psi} \tag{67}$$

where d is the net-plane distance in the stress-free material; d_ψ is the distance in the stressed material, where the lattice planes make an angle of ψ with the surface, and similarly d_0 for 0°; ν is the Poisson ratio (ratio of transverse contraction to longitudinal extension); and σ_φ is the stress in direction φ in the specimen surface (the X-ray beam direction projected onto the specimen).

The macrostress should be distinguished from the residual microstress, which is due to internal stresses caused by factors such as the hardening of steels.

In contrast to other methods of measuring elastic stress, the X-ray method is not subject to distortion of the results by plastic deformation. Furthermore, highly inhomogeneous stress fields can be measured, because the irradiated surface can be as small as 1 – 10 mm^2. The technique has the disadvantage that X-rays penetrate only some 10^{-2} cm into steels, so that only stresses at the surface can be determined.

The elastic stress in a given direction can be obtained from Equation (67), by making measurements at a range of incident angles. If a photographic method is used, objects such as bridge piers can be tested nondestructively in situ.

2.8.6. Radial Electron Density Distribution

The systems discussed up to this point have been crystalline in the common meaning of that term. As a rule of thumb, there must be more than 10 translational periods in every direction in space before a material can be deemed crystalline (a one- or two-dimensional state of order does not qualify). The expression "noncrystalline" here means the same as "X-ray amorphous".

The more noncrystalline a sample becomes, the less possible it is to give precise statements about individual atomic distances. In these cases, only a radial distribution function (RDF) of density $\varrho(r)$ of electrons or atoms can be given, according to the equation [17, p. 795]

$$4\pi r^2 \varrho(r) = 4\pi r^2 \varrho_0 + \frac{2r}{\pi} \int_0^\infty S\, i(S) \sin r S \, \mathrm{d}S \qquad (68)$$

where ϱ_0 is a function of the density of the solid, S equals $4\pi \sin\theta/\lambda$, and $i(S)$ is a function of the diffraction pattern. This equation can only be applied to samples with only one kind of atom.

3. Synchrotron Radiation

Synchrotron radiation is generated when electrons travel along curved trajectories at speeds close to that of light. The curvature is produced by magnetic fields. The nature and intensity of these fields determine the spectral distribution of the electromagnetic radiation produced. In general, a spectrum ranging from short-wavelength X-rays ($< 0.5 \times 10^{-10}$ m) up to the ultraviolet (200 nm) is available. In contrast to X-rays from rotating anode tubes, the synchrotron source offers high intensity, but also a broad spectrum, only slight divergence, and well-defined state of polarization. The radiated power (whole spectrum) obtained from an electron current of 100 mA is around 50 kW, several orders of magnitude greater than the output of conventional X-ray sources [127].

For elastic interactions at large angles ($>5°$), scattering by an amorphous material (short-range density fluctuations) is distinguished from diffraction, where interference effects play a key role. Studies in this field employ the powder, rotating crystal, and diffractometer methods, but the principal method and the most developed is also the oldest – the Laue method. Protein crystal structures can be examined in the time-resolved mode (submillisecond range) with this technique. Because of the high degree of polarization of the radiation, magnetic structures can also be investigated. At small angles, the Bragg reflections of structures with very large cells (including long periods) are accompanied by scattering, which yields information about domain sizes, defects, microcracks, and clusters and segregations of all kinds – provided there are regions differing in electron density. Time-resolved studies are possible in all these cases because of the high intensity of the X-ray beam used.

The inelastic interaction between synchrotron radiation and solid matter is also utilized by crystallographers, although spectroscopic transitions rather than diffraction phenomena are involved (EXAFS, XANES, UPS, XPS). These techniques measure the fine structure of absorption edges, to yield information about the bonding of the atom under study.

4. Neutron Diffraction

Many experiments can be performed with either X-rays or neutrons. Neutron radiation is some orders of magnitude more expensive and so is not a candidate for studies that can be done equally well or better (because of the higher intensity) with X-rays [127], [128].

An important application of neutron diffraction is based on the fact that the dependence of scattering power on atomic number is different for neutrons and X-rays. For X-rays, the atomic form factor is directly proportional to the atomic number, but the relationship is more complicated for neutron radiation, and there is a correlation with the energy states of isotopes. Between hydrogen and deuterium, for example, there is an appreciable difference in scattering power. The fact that neutrons, unlike X-ray quanta, are scattered by atomic nuclei means that the scattering power is virtually independent of the diffraction angle. A further advantage of the use of neutrons is the lower absorption.

By virtue of these factors, diffraction of neutrons by polycrystalline matter yields far more information than that of X-rays. Thicker layers of material can be penetrated, and preferred orientations can be eliminated. Useful intensities can be measured out to large diffraction angles.

For most structure problems investigated with neutrons, an X-ray structure analysis already exists. The task is to improve or refine certain parameters. There are four essential problems:

1) To determine the coordinates of hydrogen atoms

2) To distinguish between atoms of similar atomic number
3) To perform Rietveld profile analyses on powder samples to determine special structure parameters or achieve structure refinement
4) To analyze magnetic structures

5. Electron Diffraction

5.1. Diffraction in the Electron Microscope

Because of their strong interaction with matter, electrons have little penetrating power, and are thus particularly suited to surface studies. No line broadening is expected up to a crystallite size of 100×10^{-10} m for electrons with a wavelength of 0.05×10^{-10} m (accelerating voltage 60 kV). If the crystallite orientation has a statistical distribution, a diffraction spectrum comparable to the Debye – Scherrer diagram is obtained. This can be utilized for phase analysis (like an X-ray powder pattern); in the case of reflection measurements, the analysis is valid for layers near the surface (Fig. 47).

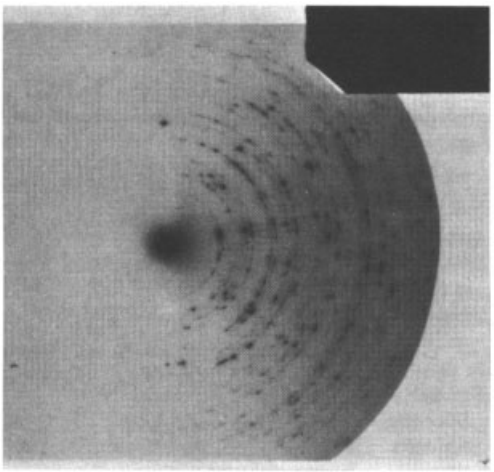

Figure 47. Electron diffraction pattern of titanium sheet, measured in reflection immediately after annealing

Surface regions just a few atomic layers thick can be examined by diffraction of electrons that have been accelerated by a low voltage (150 V). Epitaxy can be dealt with in this way. Because of absorption, deeper layers contribute nothing to the diffraction.

Often, the distribution of the vectors of the net planes in a powder specimen is not statistical. Because the curvature of the Ewald sphere is small relative to the dimensions of the reciprocal lattice, many net planes are unequally well in their reflection position, resulting in a correspondingly unequal degree of error in the diffracted electron intensities. Electron diffraction results should therefore be verified by comparative X-ray measurements whenever possible.

If the X-ray diffraction pattern offers a fingerprint with two parameters, the glancing angle and the reflection intensity, the electron diffraction pattern gives only the glancing angle, and one cannot speak of a fingerprint, especially as even strong reflections may be absent altogether. Transmission electron diffraction data, however, can be very similar to the corresponding X-ray patterns, and may contain a similar level of information (see Figs. 48 and 13). This is only the case when the distribution of net plane vectors is statistical.

5.2. Electron Diffraction in the Gas Phase

The discussion of this technique has been kept short because the diffraction spectra are very similar to Debye – Scherrer patterns; the method is becoming very important for molecular structure analysis. An electron beam in a high vacuum (0.1 – 10 Pa) collides with a molecular beam, and the electrons are diffracted by the molecules. The film reveals washed-out Debye –Scherrer rings, from which a radial electron density distribution function for the molecule can be derived. Together with spectroscopic data, the distribution makes it possible to infer the molecular structure [129].

6. Future Developments

Many of the techniques discussed here have become routine. X-ray diffraction data of organic pigments, for example, are employed for product surveillance in industry. Not just the fingerprint as such, but also the half-value widths of the reflections and the level of the background,

are utilized. The situation is similar in the fiber industry, as well as many other areas of chemical production.48

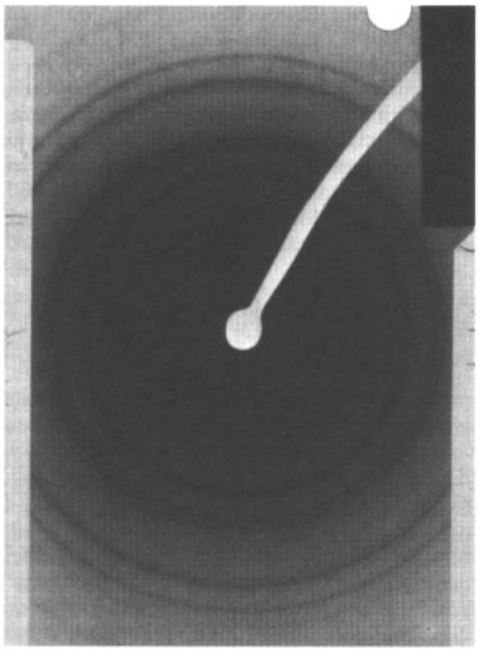

Figure 48. Transmission electron diffraction pattern of LiF

The most spectacular successes of X-ray methods, however, are in molecular and crystal structure analysis. Examples are the structures of insulin [130], hemoglobin [131], and vitamin B_{12} [132]. Today, single-crystal X-ray structure analyses of relatively complex compounds (up to ca. 40 – 200 nonhydrogen atoms) are performed with computer-controlled diffractometers whose computers can be used simultaneously to solve the phase and structure problem within a few days or just a few hours.. The method has gained parity in investigation time with spectroscopic methods for the solution of relatively complex structure problems as far as the compounds are crystallizable. What is more, X-ray structure analysis is far superior to all other methods with respect to the reliability of the results. A single-crystal X-ray structure analysis of a compound yields everything that one could "see" if it were possible to examine a single molecule under the microscope. But in the end it is very important to remove all the discrepancies between diffraction and spectroscopic methods.

However, it can be quite laborious to solve the phase problem, even for relatively simple crystal structures. There is not yet any general method; a special solution of the phase problem must be found for every structure. The methods have been developed to the point where the solution can often be achieved by a computer alone; if this fails, however, it is usually not clear why, considerable ingenuity may be needed to reach the solution.

The fundamental reason for this situation is that the relationships in direct phase-determination methods are nothing more than probability relations. If these can be turned into true equations, single-crystal X-ray structure analyses would become a routine analytical procedure. H. Hauptmann, co-creator of the direct methods of phase determination, expressed the dilemma as "One good idea and we are out of a job".

One attempt to replace the probability relations with true equations for phase determination was that of R. Rothbauer [133], based on earlier work by Sayre [78], and Woolfson [134]. The future must show whether the equations derived can be used for practical structure determination [135]. Their use for the determination of the absolute scaling factor and the overall temperature factor has already been tested in practice [136]. A publication by R. Rothbauer shows, that solving the phase problem generally is still up to date.

7. References

General References

Text Books

1. L. E. Alexander: *X-Ray Diffraction Methods in Polymer Science,* Wiley, New York 1969.
2. K. W. Andrews, D. J. Dyson, S. R. Keown: *Interpretation of Electron Diffraction Patterns,* Hilger & Watts, London 1967.
3. U. W. Arndt, B. T. M. Willis: *Single Crystal Diffractometry,* University Press, Cambridge 1966.
4. L. V. Azaroff, M. J. Buerger: *The Powder Method in X-Ray Crystallography,* McGraw-Hill, New York 1958.
5. G. E. Bacon: *Neutron Diffraction,* 3rd ed., Clarendon Press, Oxford 1975.

6. T. L. Blundell, L. N. Johnson: *Protein Crystallography,* Academic Press, New York 1976.
7. M. J. Buerger: *Vector Space and Its Application in Crystal-Structure Investigation,* Wiley, New York 1959.
8. M. J. Buerger: *Crystal-Structure Analysis,* Wiley, New York 1960.
9. M. J. Buerger: *The Precession Method in X-Ray Crystallography,* Wiley, New York 1964.
10. M. J. Buerger: *X-Ray Crystallography,* 7th ed., Wiley, New York 1966.
11. M. J. Buerger: *Kristallographie, eine Einführung in die geometrische und röntgenographische Kristallkunde,* De Gruyter, Berlin 1977.
12. J. D. Dunitz: *X-Ray Analysis and the Structure of Organic Molecules,* Cornell University Press, Ithaca 1979.
13. C. Giacovazzo et al.: "Fundamentals of Crystallography," in C. Giacovazzo (ed.): *International Union of Crystallography,* Oxford University Press, Oxford 1992.
14. J. P. Glusker, K. N. Trueblood: *Crystal Structure Analysis, a Primer,* Oxford University Press, New York 1985.
15. E. F. Kaelble (ed.): *Handbook of X-Rays,* McGraw-Hill, New York 1967.
16. A. J. Kitaigorodsky: *Molecular Crystals and Molecules,* Academic Press, New York 1973.
17. H. P. Klug, L. E. Alexander: *X-Ray Diffraction Procedures for Polycrystalline and Amorphous Materials,* J. Wiley & Sons, New York 1974.
18. H. Krebs: *Grundzüge der anorganischen Kristallchemie,* Enke-Verlag, Stuttgart 1968.
19. H. Lipson, W. Cochran: *The Determination of Crystal Structures,* 3rd ed., G. Bell & Sons, London 1966.
20. P. Luger: *Modern X-Ray Analysis on Single Crystals,* De Gruyter, Berlin 1980.
21. D. McKie, C. McKie: *Essentials of Crystallography,* Blackwell Scientific Publ., Oxford 1986.
22. H. Neff: *Grundlagen und Anwendungen der Röntgenfeinstrukturanalyse,* 2nd ed., R.-OldenbourgVerlag, München 1962.
23. G. H. Stout, L. H. Jensen: *X-Ray Structure Determination,* Macmillan, New York 1968.
24. E. R. Wölfel: *Theorie und Praxis der Röntgenstrukturanalyse,* Vieweg, Braunschweig 1975.
25. M. M. Woolfson: *An Introduction to X-Ray Crystallography,* University Press, Cambridge 1970.
26. W. Borchardt-Ott: Kristallographie, eine Einführung für Naturwissenschaftler; 5. Auflage; Springer-Verlag, Berlin 1997.
27. J. P. Glusker, M. Lewis, M. Rossi: Crystal Structure Analysis for Chemists and Biologists; New York, Weinheim, Cambridge (UK) 1994
28. J. S. Rollet (ed.): *Computing Methods in Crystallography.* Pergamon Press, Oxford 1965.

General References
Reference Works

29. *International Tables for X-Ray Crystallography,* Kynoch Press, Birmingham, N. F. M. Henry, K. Lonsdale (eds.): "Symmetry Groups," vol. I, 1952; J. S. Kasper, K. Lonsdale (eds.): *"Mathematical Tables,"* **vol. II,** 1959; C. H. MacGillavry, G. D. Rieck (eds.): *"Physical and Chemical Tables,"* **vol. III,** 1962; J. A. Ibers, W. C. Hamilton (eds.): *"Revised and Supplementary Tables,"* **vol. IV,** 1974.
30. International Tables for Crystallography, Kluwer Academic Publishers, Dordrecht: Th. Hahn (ed.), Vol. A: Space-Group Symmetry, fifth revised edition, 2000; U. Shmueli (ed.), Vol. B: Reciprocal Space, second edition, 2000; A. J. C. Wilson, E. Prince (eds.), Vol. C: Mathematical, Physical & Chemical Tables, second edition, 1999.

General References
Databanks

31. J. D. H. Donnay, H. M. Ondik (eds.): *Crystal Data: Determinative Tables,* 3rd. ed., **vol. 1, "Organic Compounds"; vol. 2, "Inorganic Compounds,"** U.S. Dep. of Commerce – Nat. Bur. of Standards – Joint Committee on Powder Diffraction Standards, 1972.
32. *Landolt-Börnstein,* 7th ed., **III,** 5; III 6; III 7; III 8.
33. A. R. Hölzel: *Systematics of Minerals, Systematik in der Mineralogie,* Ober-Olm, Germany 1990.
34. O. Kennard, F. H. Allen, D. G. Watson (eds.): *Molecular Structures and Dimensions, Guide to the Literature 1935 – 1976, Organic and Organometallic Crystal Structures,* Crystallographic Data Centre, Cambridge 1977.
35. O. Kennard et al.: *Interatomic Distances, 1960 – 1965, Organic and Organometallic Crystal Structures,* Crystallographic Data Centre, Cambridge 1972.

36. R. W. G. Wyckoff: *Crystal Structures,* 2nd ed., **vols. 1 – 6,** Wiley, New York 1963 – 1971.
37. F. H. Allen, et al., *J. Chem. Inf. Comput. Sci.* **31** (1991) 187 – 204.
38. L. G. Berry (ed.): *Powder Diffraction File (Organic and Inorganic),* International Centre for Diffraction Data (ICDD), Swarthmore, USA.
39. G. Bergerhoff et al., ICSD, Inorganic Crystal Structure Database, University Bonn.
40. A. R. Hölzel: MDAT, Datenbank der Mineralien, Systematik in der Mineralogie, Ober-Olm, Germany 1992.
41. Protein Data Bank, Brookhaven National Laboratory, USA.

General References

Computer Programs

42. A. Altomare et al., SIR 92, User's Manual, University of Bari and Perugia, Italy 1992 [138]
43. P. T. Beurskens et al., DIRDIF, a Computer Program System for Crystal Structure Determination by Patterson Methods and Direct Methods applied to Difference Structure Factors, 1992.
44. T. Debaerdemaeker et al., MULTAIN 88, a System of Computer Programs for the Automatic Solution of Crystal Structures from X-Ray Diffraction Data, Universities of Ulm (Germany), Louvain (Belgium), and York (England) 1988.
45. A. L. Spek, PLATON, a Multi-Purpose Crystallographic Tool, University of Utrecht, the Netherlands (1998)
46. E. J. Gabe, A. C. Larson, F. L. Lee, Y. Lepage: NRCVAX, Crystal Structure System, National Research Council, Ottawa, Canada.
47. S. R. Hall, J. M. Stewart (eds.): XTAL 3.0, Reference Manual, University of Western Australia, Australia, and Maryland, USA.
48. C. K. Johnson: ORTEP II, Report ORNL 5138, Oak Ridge National Laboratory, Tennessee, USA.
49. G. M. Sheldrick: SHELXS 86, Program for the Solution of Crystal Structures, University of Göttingen, Germany 1986.
50. G. M. Sheldrick: SHELXL 93, Program for the Refinement of Crystal Structures, University of Göttingen, Germany 1993.
51. G. M. Sheldrick: SHELX-97, Computer program for the solution and refinement of crystal strucutres. University of Göttingen, Germany 1997.
52. D. J. Watkin, J. R. Carruthers, P. W. Betteridge: Crystals User Guide, Chemical Crystallography Laboratory, University of Oxford, England 1985.

Specific References

53. R.-G. Kretschmer, G. Reck, POSIT in W. Kraus, G. Noke, Computer program PowderCell, BAM Berlin (1996).
54. G. E. Engel, S. Wilke, O. König, K. D. M. Harris, F. J. J. Leusen, PowderSolve (1999) in Cerius^2 (Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA).
55. E. Keller, *Chem. unserer Zeit* **16** (1982) 71 – 88, 116 –123.
56. R. Bonart, R. Hosemann, *Makromol. Chem.* **39** (1960) 105 – 118.
57. W. Hoppe, *Z. Elektrochem.* **61** (1957) 1076 – 1083.
58. W. Hoppe, E. F. Paulus, *Acta Crystallogr.* **23** (1967) 339 – 342.
59. H. Fuess, *Chem. Br.* **14** (1978) 37 – 43.
60. H. Diehls: *Die Fragmente der Vorsokratiker,* 8th ed., **vol. 2,** Weidmannsche Verlagsbuchhandlung, Berlin 1952, p. 97.
61. T. Kottke, D. Stalke, *J. Appl. Crystallogr.* **26** (1993) 615 – 619.
62. *Sci. Ind. (Eindhoven)* **18** (1972) 22 – 28.
63. D. Kobelt, E. F. Paulus, *Siemens Power Eng.* **1** (1979) 269 – 271.
64. R. A. Sparks, *Abstracts of Internat. Summer School on Crystallography Computing,* Prague, July, 28 –Aug. 5, 1975, pp. 452 – 467.
65. R. Boese, M. Nussbaumer: "In situ Crystallization Techniques," in D. W. Jones (ed.). *Organic Crystal Chemistry,* IUCR, Cryst. Symposia, vol. VI, Oxford University Press, Oxford 1993.
66. E. Pohl et al., *Acta Crystallogr.* **D 49** (1993) in press.
67. G. M. Sheldrick, E. F. Paulus, L. Vértesy, F. Hahn, Acta Crystallogr. B51, 89 – 98.
68. B. Baumgartner et al.: *Modern Fast Automatic X-Ray Powder Diffractometry (Stoe & Cie GmbH, company brochure),* Darmstadt 1988.
69. Commission of Crystallographic Apparatus of the Internat. Union of Crystallography, *Acta Crystallogr.* **9** (1956) 520 – 525.
70. Robert Huber Diffraktionstechnik: *Manual Guinier-System 600 (company brochure),* Rimsting, Germany.
71. Verordnung über den Schutz vor Schäden durch Röntgenstrahlen (Röntgenverordnung) January 8, 1987 (BGBl. I S. 114). Textausgabe

mit Anm. von E. Witt. C. Heymanns-Verlag, Köln 1987.
72. Permissible Dose from External Sources of Ionising Radiations, NBS Handbook 59, pp. 2, 3, addendum, April 15, 1958, Government Printing Office, Washington.
73. Recommendations of the International Commission on Radiological Protection, Pergamon Press, New York 1959.
74. A. L. Patterson, *Phys. Rev.* **46** (1934) 372 – 376.
75. R. Huber, *Acta Crystallogr.* **19** (1965) 353 – 356.
76. J. Karle, *Adv. Struct. Res. Diffr. Methods* **1** (1964) 55 – 88.
77. E. Egert, G. M. Sheldrick, *Acta Crystallogr.* **A43** (1985) 262 – 268.
78. D. M. Sayre, *Acta Crystallogr.* **5** (1952) 60 – 65.
79. E. W. Hughes, *Acta Crystallogr.* **6** (1953) 871.
80. G. N. Watson: *The Theory of Bessel Functions,* 2nd ed., University Press, Cambridge 1958, p. 77.
81. H. Hauptmann, J. Karle: "Solution of the Phase Problem. I, The Centrosymmetric Crystal," *Monogr.* no. 3, Polycrystal Book Service, Pittsburgh 1953.
82. H. Hauptmann, J. Karle, *Acta Crystallogr.* **12** (1959) 93 – 97.
83. H. Hauptmann, J. Karle, *Acta Crystallogr.* **9** (1956) 45 – 55.
84. J. Karle, H. Hauptmann, *Acta Crystallogr.* **14** (1961) 217 – 223.
85. I. L. Karle, J. Karle, *Acta Crystallogr.* **16** (1963) 969 –975.
86. I. L. Karle, J. Karle, *Acta Crystallogr.* **17** (1964) 835 –841.
87. G. Germain, M. M. Woolfson, *Acta Crystallogr.* **B 24** (1968) 91 – 96.
88. M. M. Woolfson, *Acta Crystallogr.* **A 33** (1977) 219 – 225.
89. P. Main, *Acta Crystallogr.* **A 34** (1978) 31 – 38.
90. G. M. Sheldrick, *Acta Crystallogr.* **A 46** (1990) 467 – 473.
91. G. Germain, P. Main, M. M. Woolfson, *Acta Crystallogr.* **A 27** (1971) 368 – 376.
92. T. Debaerdemaeker, C. Tate, M. M. Woolfson, *Acta Crystallogr.* **A41** (1985) 268 – 290.
93. M. M. Woolfson, *Acta Crystallogr.* **A43** (1987) 593 – 612.
94. H. Hauptman, *Science* **233** (1986) 178 – 183.
95. R. Miller, G. T. DeTitta, R. Jones, D. A. Langs, C. M. Weeks, H. A. Hauptman, *Science* **259** (1993) 1430 – 1433.
96. R. Miller, S. M. Gallo, H. G. Khalak, C. M. Weeks, *J. Appl. Cryst.* **27** (1994) 613 – 621.
97. G. M. Sheldrick in S. Fortier (ed.): *Direct Methods for Solving Macromolecular Structurs,* Kluwer Academic Publishers, Dordrecht 1998, pp. 401 – 411.
98. I. Usón, G. M. Sheldrick, *Curr. Op. Struct. Biol.* **9** (1999) 643 – 648.
99. M. Schäfer, Th. R. Schneider, G. M. Sheldrick, *Structure* **4** (1996) 1509 – 1515.
100. B. Post, *Acta Crystallogr.* **A 39** (1983).
101. E. Weckert, W. Schwegle, K. Hümmer, *Proc. R. Soc. London A* **424** (1993) 33 – 46.
102. R. Herbst-Irmer, G. M. Sheldrick, *Acta Crystallogr.* **B54** (1998) 443 – 449.
103. F. L. Hirshfeld, D. Rabinovich, *Acta Crystallogr.* **A 29** (1973) 510 – 513.
104. D. W. Engel, *Acta Crystallogr.* 28 (1972) 1496 – 1509.
105. W. C. Hamilton, *Acta Crystallogr.* **18** (1965) 502 –510.
106. D. Rogers, *Acta Crystallogr.* **A 37** (1981) 734 – 741.
107. H. D. Flack, *Acta Crystallogr. Sect. A: Found Crystallogr.* **A 39** (1983) 876 – 881.
108. W. Hoppe, *Angew. Chem.* **77** (1965) 484 – 492.
109. A. J. C. Wilson: *Nature (London)* **150** (1942) 151 –152.
110. P. Main et al.: *MULTAN 77, a System of Computer Programmes for the Automatic Solution of Crystal Structures from X-Ray Diffraction Data,* University of York, England 1977.
111. P. Debye, *Ann. Phys. (Leipzig)* **46** (1915) 809 – 823.
112. D. Reuschling, E. F. Paulus, H. Rehling, *Tetrahedron Lett.* 1979, 517 – 520.
113. T. Ito: *X-Ray Studies on Polymorphism,* Maruzen Co., Tokyo 1950, pp. 187 – 228.
114. H. M. Rietveld, *Acta Crystallogr.* **22** (1967) 151 –152.
115. E. F. Paulus et al., *Solid State Commun.* **73** (1990) 791 – 795.
116. R. Hedel, H. J. Bunge, G. Reck, *Material Science Forum* **157 – 162** (1994) 2067 – 2074.
117. A. Le Bail, Structure Determination from Powder Diffraction-Database (1993 – 1999). http://www.cristal.Org/iniref.html.
118. G. Reck, R.-G. Kretschmer, L. Kutschabsky, W. Pritzkow, *Acta crystallogr.* **A44** (1988) 417 – 421.
119. G. E. Engel, S. Wilke, O. König, K. D. M. Harris, F. J. J. Leusen, *J. Appl. Cryst.* **32** (1999) 1169 – 1179.

120. R. J. Gdanitz in A. Gavezzotti (ed.): *Theoretical Aspects and Computer Modeling,* John Wiley & Sons 1997, pp. 185 – 201.
121. M. U. Schmidt, R. E. Dinnebier, *J. Appl. Cryst.* **32** (1999) 178 – 186.
122. M. U. Schmidt in D. Braga, G. Orpen (eds.): *Crystal Engineering: From Molecules and Crystals to Materials,* Kluwer Academic Publishers, Dordrecht 1999, pp. 331 – 348.
123. K. D. M. Harris, M. Tremayne, *Chem. Mater.* **8** (1996) 2554 – 2570.
124. P. Erk in D. Braga, G. Orpen (eds.): *Crystal Engineering: From Molecules and Crystals to Materials,* Kluwer Academic Publishers, Dordrecht 1999, pp. 143 – 161.
125. G. Ibe, K. Lücke, *Texture (London)* **1** (1973) 87 – 98.
126. R. C. Rau, *Adv. X-Ray Anal.* **5** (1962) 104 – 115.
127. A. K. Cheetham, A. P. Wilkinson, *Angew. Chem.* **104** (1992) 1594 – 1608.
128. H. Fuess: "Application of Neutron Scattering in Chemistry. Pulsed and Continuous Sources in Comparison," in M. A. Carrondo, G. A. Jeffrey (eds.): *Chemical Crystallography with Pulsed Neutrons and Synchrotron X-Rays,* D. Reidel Publ., Dordrecht 1988, pp. 77 – 115.
129. L. Schäfer, *Appl. Spectrosc.* **30** (1976) 123 – 149.
130. T. Blundell, G. Dodson, D. Hodgkin, D. Mercola, *Adv. Protein Chem.* **26** (1972) 279 – 402.
131. M. F. Perutz, H. Muirhead, J. M. Cox, L. G. C. Goaman, *Nature (London)* **219** (1968) 131 – 139.
132. C. Brink-Shoemaker et al., *Proc. R. Soc. London A* **278** (1964) 1 – 26.
133. R. Rothbauer, *Acta Crystallogr.* **A 33** (1977) 365 – 367.
134. M. M. Woolfson, *Acta Crystallogr.* **11** (1958) 277 –283.
135. R. Rothbauer, *Z. Kristallogr.* **209** (1994) 578 – 582 .
136. R. Rothbauer, *Acta Crystallogr.* **A 34** (1978) 528 – 533.
137. R. Rothbauer, *Z. Kristallogr.* **215** (2000) 158 – 168.
138. A. Altomare, G. Cascarano, C. Giacovazzo, A. Guagliardi, M. C. Burla, G. Polidori, M. Camalli, *J. Appl. Cryst.* **27** (1994) 435 – 436.

Ultraviolet and Visible Spectroscopy

GÜNTER GAUGLITZ, Universität Tübingen, Institut für Physikalische und Theoretische Chemie, Tübingen, Federal Republic of Germany

1. Introduction . . . 3
1.1. Comparison with Other Spectroscopic Methods . . . 3
1.2. Development and Uses . . . 3
2. Theoretical Principles . . . 4
2.1. Electronic States and Orbitals . . . 4
2.2. Interaction Between Radiation and Matter . . . 4
2.2.1. Dispersion . . . 5
2.2.2. Absorption . . . 5
2.2.3. Scattering . . . 6
2.2.4. Reflection . . . 6
2.2.5. Band Intensity . . . 7
2.3. The Lambert – Beer Law . . . 7
2.3.1. Definitions . . . 7
2.3.2. Deviations from the Lambert – Beer Law . . . 8
2.4. Photophysics . . . 8
2.4.1. Energy Level Diagram . . . 8
2.4.2. Deactivation Processes . . . 8
2.4.3. Transition Probability and Fine Structure of the Bands . . . 10
2.5. Chromophores . . . 10
2.6. Optical Rotatory Dispersion and Circular Dichroism . . . 11
2.6.1. Generation of Polarized Radiation . . . 11
2.6.2. Interaction with Polarized Radiation . . . 11
2.6.3. Optical Rotatory Dispersion . . . 12
2.6.4. Circular Dichroism and the Cotton Effect . . . 12
2.6.5. Magnetooptical Effects . . . 13
3. Optical Components and Spectrometers . . . 13
3.1. Principles of Spectrometer Construction . . . 13
3.1.1. Sequential Measurement of Absorption . . . 14
3.1.2. Multiplex Methods in Absorption Spectroscopy . . . 15
3.2. Light Sources . . . 16
3.2.1. Line Sources . . . 16
3.2.2. Sources of Continuous Radiation . . . 16
3.2.3. Lasers . . . 17
3.3. Selection of Wavelengths . . . 17
3.3.1. Prism Monochromators . . . 18
3.3.2. Grating Monochromators . . . 18
3.4. Polarizers and Analyzers . . . 19
3.5. Sample Compartments and Cells . . . 19
3.5.1. Closed Compartments . . . 19
3.5.2. Modular Arrangements . . . 19
3.5.3. Open Compartments . . . 19
3.6. Detectors . . . 20
3.7. Optical Paths for Special Measuring Requirements . . . 21
3.7.1. Fluorescence Measurement . . . 21
3.7.2. Measuring Equipment for Polarimetry, ORD, and CD . . . 22
3.7.3. Reflection Measurement . . . 23
3.7.4. Ellipsometry . . . 23
3.8. Effect of Equipment Parameters . . . 23
3.9. Connection to Electronic Systems and Computers . . . 26
4. Uses of UV – VIS Spectroscopy in Absorption, Fluorescence, and Reflection . . . 27
4.1. Identification of Substances and Determination of Structures . . . 27
4.2. Quantitative Analysis . . . 27
4.2.1. Determination of Concentration by Calibration Curves . . . 27
4.2.2. Classical Multicomponent Analysis . . . 29
4.2.3. Multivariate Data Analysis . . . 29
4.2.4. Use in Chromatography . . . 29
4.3. Fluorimetry . . . 30
4.3.1. Inner Filter Effects . . . 31
4.3.2. Fluorescene and Scattering . . . 31
4.3.3. Excitation Spectra . . . 31
4.3.4. Applications . . . 31
4.4. Reflectometry . . . 32
4.4.1. Diffuse Reflection . . . 32
4.4.2. Color Measurement . . . 32
4.4.3. Regular Reflection . . . 32
4.4.4. Determination of Film Thickness . . . 33
4.4.5. Ellipsometry . . . 33
4.5. Resonance Methods . . . 34
4.5.1. Surface Plasmon Resonance . . . 34
4.5.2. Grating Couplers . . . 34
4.6. On-Line Process Control . . . 35
4.6.1. Process Analysis . . . 35
4.6.2. Measurement of Film Thicknesses . . . 35
4.6.3. Optical Sensors . . . 35

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_383

4.7. Measuring Methods Based on Deviations from the Lambert – Beer Law 36
5. Special Methods 36
5.1. Derivative Spectroscopy 36
5.2. Dual-Wavelength Spectroscopy . . . 39
5.3. Scattering 40
5.3.1. Turbidimetry 40
5.3.2. Nephelometry 40
5.3.3. Photon Correlation Spectroscopy . . . 41
5.4. Luminescence, Excitation, and Depolarization Spectroscopy, and Measurement of Lifetimes 41
5.5. Polarimetry 42
5.5.1. Sugar Analysis 42
5.5.2. Cellulose Determination 42
5.5.3. Stereochemical Structural Analysis . . 42
5.5.4. Use of Optical Activity Induced by a Magnetic Field 42
5.6. Photoacoustic Spectroscopy (PAS) . 43
6. References 44

Abbreviations

a	linear (decadic) absorption coefficient, concentration
A	absorption
$\boldsymbol{b}$	matrix of signals
$\boldsymbol{B}$	magnetic field vector
c	velocity of light, concentration
d	thickness, pathlength
E	energy, absorbance, extinction
$\boldsymbol{E}$	electric field vector
f	spectral distribution
F	spontaneous emission
g	grating constant
h	Planck's constant
H	magnetic field
I	fluorescence intensity
k	Boltzmann constant, absorption coefficient
l	path length
n	refractive index
N	number of molecules
p	polarization
P	photochemistry
$\boldsymbol{Q}$	inverse of the matrix of the absorption coefficient
R	reflectance of the sample
s	scattering coefficient
S	singlet state
T	triplet state
v	velocity
$[\alpha$	]specific rotation
α	angle, absorptance
β	angle
Γ	geometric factor
δ	phase shift
ε	extinction coefficient, molar (decadic) absorption coefficient
η	fluorescence quantum yield
ϑ	angle
Θ	ellipticity
λ	wavelength
μ	dipole moment
ν	frequency
$\tilde{\nu}$	wavenumber
τ	turbidity coefficient
φ	phase shift
Φ	angle
$[\Phi$	]molecular rotation
ω	Verdet constant

1. Introduction

1.1. Comparison with Other Spectroscopic Methods

Spectroscopy in the ultraviolet and visible regions of the spectrum (UV – VIS spectroscopy) involves observation of the excitation of electrons and is, therefore, often referred to as "electron spectroscopy". The term electron spectroscopy does not mean spectroscopy in which excitation is performed by electrons, but spectroscopy of the electronic states (excitation of the electrons).

An electron is excited if the frequency of the incident electromagnetic radiation matches the difference in energy between two electronic states. This energy difference depends on the electronic structure of the molecule being investigated and its environment. In addition, vibrational and rotational transitions can be excited. Therefore, bands in UV – VIS spectroscopy are generally broad. These bands provide little information about the molecular structure and functional groups present, especially for spectroscopy of liquids and solids [1–4].

1.2. Development and Uses

The present state of UV – VIS measurement technology has developed from its starting point *colorimetry*. In colorimetry, the intensity of a color was first taken as a measure of concentration, the sample being compared visually with standards, indicator paper being a typical example. This mainly qualitative procedure was then improved by the use of filter-photometers. These measure the attenuation of light by the sample in a narrow spectral region (photometry) and compare it with solutions containing only pure solvent. Replacement of the filter by a monochromator with a wavelength scanning mechanism then resulted in *spectrometry*.

Whereas IR, NMR, and mass spectroscopy are used mainly for the elucidation of structure and the identification of substances, UV – VIS spectroscopy enables quantitative determinations to be carried out much more precisely and reproducibly. Therefore, its primary areas of application are in quantitative analysis and clinical medicine, in the determination of drug concentrations, in the quantification of pharmaceuticals and as detectors in chromatographic processes (HPLC, TLC) (→ Liquid Chromatography, Chap. 2.6.3.) [4,5]. Furthermore, mixtures as well as pure substances can be studied and the components determined quantitatively by methods of multicomponent analysis [6]. Since modern spectrometers operate very rapidly and can be constructed in the form of photodiode arrays, they have the advantage over other analytical methods of being usable not only for observing stationary systems, but also for carrying out repeated determinations very rapidly, sometimes within milliseconds [3,4].

2. Theoretical Principles

2.1. Electronic States and Orbitals

The energy of a molecule is given by its movement through space (translational energy), its rotary motion (rotational energy), the oscillatory movement of the atoms with respect to each other (vibrational energy), and the distribution of its electron density (electronic energy) [7–9]. The corresponding energy states can be calculated as eigenvalues by the Schrödinger equation through use of the electronic, vibrational, and rotational eigenfunctions, and the boundary conditions given by the molecular structure [10].

The energy levels or states in the UV – VIS spectral region correspond to electron density distributions (orbitals). Some typical orbitals for an organic molecule are shown schematically in Figure 1 [3]. Also shown are possible transitions of varying intensity corresponding to the symmetry of the electron density distributions of the initial and final states.

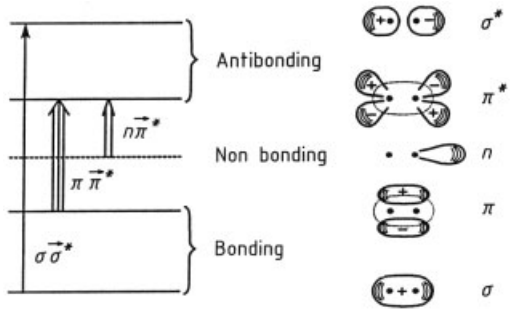

Figure 1. Bonding and antibonding orbitals, showing symmetries (+/− : sign of the wave function) and positions of atomic nuclei (●) for a diatomic molecule (HOMO = π or n; LUMO = $\pi*$)
Thickness of vertical arrows indicates intensity of the transition

2.2. Interaction Between Radiation and Matter

In electromagnetic radiation, either the particle (photons) or the wavelike character is more noticeable. In the ultraviolet and visible wavelength region, the latter is the case. The electric field vector of the radiation induces charge separation (polarization) within the molecules. In the UV – VIS region, the sign of the electric field vector changes at ca. 10^{15} Hz, causing the electron density to be polarized at this frequency. The extent of this polarization depends on the dielectric properties of the molecules, their environment, and the wavelength of the radiation [11–13]. The Bohr – Einstein relationship

$$\Delta E = E_2 - E_1 = h\nu$$

$$h\nu = hc\tilde{\nu} = hc/\lambda \quad (1)$$

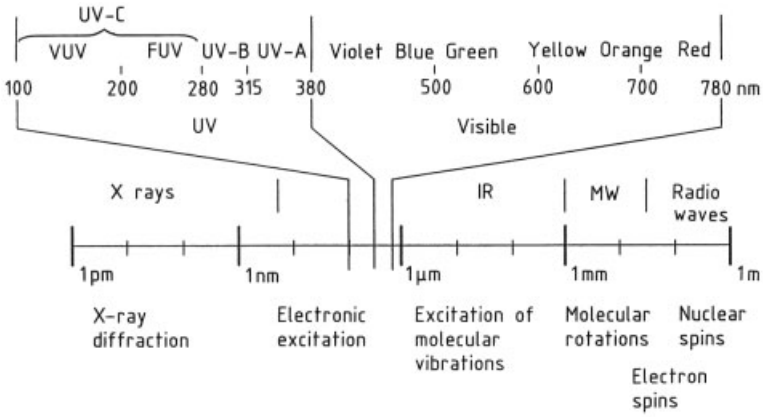

Figure 2. Location of UV and visible regions within the complete electromagnetic radiation spectrum, showing visible colors and types of excitation
VUV = Vacuum UV; FUV = Far UV;; MW = Microwaves

connects the discrete atomic or molecular electron energy states with the frequency ν of the exciting electromagnetic radiation, where $\tilde{\nu} = 1/\lambda$ is the wavenumber, c is the velocity of light in a vacuum ($c = 2.99 \times 10^8$ m/s), and h is Planck's constant (6.62×10^{-34} J · s). The location of this spectral region within the complete electromagnetic radiation spectrum is shown in Figure 2, which also indicates the colors of the visible range.

2.2.1. Dispersion

The simple model of a harmonic oscillator can account for the processes of polarization, dispersion, absorption, and scattering. In *polarization*, interaction of the radiation with matter (dielectric) induces a reversing dipole in each molecule. The effect is proportional to the number of particles, so that polarization is a volume-based quantity. The induced dipole moment represents a Hertz dipole reversing with the excitation frequency, thus behaving as a radiation source. Because of the fluctuation of the charge distribution that fluctuates with the dipole's eigenfrequency. As shown in Figure 3, this radiation is emitted perpendicularly to the axis of polarization.

When radiation propagates through a medium, its frequency remains the same due to the conservation of energy, but its velocity decreases [11, 13]. The ratio of propagation rates is the refractive index n ($n \geq 1$). The variation of refractive index with frequency or wavelength is the dispersion [11, 13]. Two cases must be distinguished – absorption and scattering.

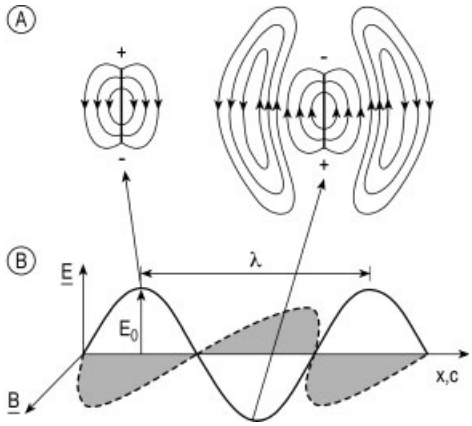

Figure 3. A) Electrical fields for a Hertz dipole (**E** || *dipole moment* μ); B) Electrical and magnetic field vectors in the propagation of electromagnetic radiation

2.2.2. Absorption

If the frequency of the radiation corresponds closely to the energy difference of the transition between two energy states, as given by Equation (1), this leads to a resonance excitation, a change in the electron density distribution, and an electronic transition (see Fig. 1) from the highest occupied molecular orbital (HOMO) to the lowest unoccupied molecular orbital (LUMO). In solids and metals, HOMO

corresponds to the valence, and LUMO to the conduction band [12,14,15]. The energy required for this transition is provided by the radiation. This process is known as (induced) absorption. In the range of an absorption band, anomalous dispersion is observed (refractive index increases with wavelength). Depending on the molecular environment and possible pathways of deactivation, the new excited state can exist for 10^{-13} to 10^{-3} s.

Only gaseous atoms give line spectra. Molecules in the gaseous state and in solution produce bands (sometimes with vibrational structure), because the energy for the electronic transition is very high, and this transition is usually accompanied by vibrational and rotational excitation. Line broadening in solution is caused by interaction between the solvent and the molecules, leading to a reduction in the lifetime of the excited state. According to Heisenberg's uncertainty principle

$$\Delta E \Delta \tau \geq h/2\pi \tag{2}$$

the energy uncertainty of the excited state increases at the expense of its lifetime. The relationship between refractive index and absorption coefficient can be expressed by the Kramers–Kronig equations [16].

2.2.3. Scattering

If the frequency of incident radiation does not correspond to the energy difference between two electronic levels, the polarization does not lead to a new stable (stationary) electron density distribution. After a very short delay, the molecule emits the energy that it absorbed previously. In this frequency range normal dispersion is observed (i.e., refractive index decreases with wavelength). When refractive index is plotted as a function of wavelength a dispersion curve is obtained. It characterizes the molecule together with the polarizability. With the aid of the dispersion curve, relationships between the molecular structure and the spectrum can be established. This gives good results in the infrared, but only poor results in the UV–VIS region [17].

Since the energy causing the polarization is reemitted immediately in the nonresonant case (see Fig. 3), no energy is absorbed. Even the smallest molecules show scattering effects. However, since the intensity of scattered radiation increases proportionally to the sixth power of the length of the induced dipole moment, the radiation becomes observable only for molecular diameters of $\geq \lambda/10$ (Rayleigh scattering) and/or extremely high radiation intensities (lasers) [18,19].

In the case of Mie scattering (larger molecules with a diameter of $\lambda/4 - \lambda/2$), a molecule contains several scattering centers. Thus, the different scattering centers emit waves that interfere to varying extents in various spatial directions, depending on the location of the scattering centers. Hence, the interference pattern contains information about the size and shape of the molecule. The intensity distribution of Rayleigh and Mie scattering assuming many molecules with statistical orientations overlap is shown in Figure 4 [18].

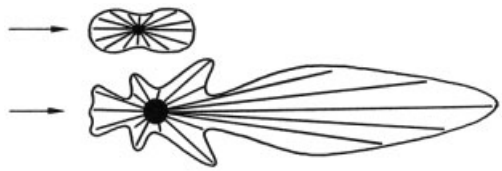

Figure 4. Intensity distribution of Rayleigh (above) and Mie (below) scattered radiation projected onto a plane. Excitation radiation is in the direction of the arrow. The black spot represents the superposition of many molecular dipoles in all spatial directions.

2.2.4. Reflection

Whereas in gases or liquids the particles are disordered and cause random scattering, atoms or molecules in the surface layer of solids (especially crystals and metals) are very highly ordered. Especially metals exhibit extremely polarizable electron density distributions. The many Hertz dipoles show constructive interference in certain directions. At solid surfaces this leads to directional reflection, which encloses the same angle with the optical normal to the surface as the incident radiation [11,20] (Fig. 5). Reflectance R is given by the Fresnel equation, which, for the simplest case of normal incidence and a transparent medium, is [11,13]

$$R = \left[\frac{(n_1 - n_2)}{(n_1 + n_2)}\right]^2 \quad (3)$$

When radiation enters a material with a higher refractive index n, it is refracted toward the optical axis (Snell's law):

$$n_1 \sin\alpha = n_2 \sin\beta \quad (4)$$

For a material with a lower refractive index, the angle between reflected radiation and the optical axis becomes larger. Under these conditions, total internal reflection occurs at the critical angle β_{cr} (see Fig. 5) [20]. Techniques based on reflection have become very important in analysis.

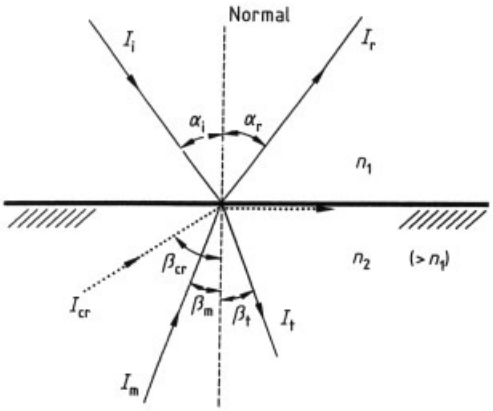

Figure 5. Ray paths for internal, external, and total internal reflection
Indices i, r, t = Incident, reflected, transmitted; I = Intensity; cr = Critical angle; m = Ray out of medium 2 corresponding to incident ray I_i, forms I_r

2.2.5. Band Intensity

Regarding transitions from ground to excited state and vice versa, the interaction of the radiation with the molecules are in principle the same. For this reason, the Einstein coefficients, which are a measure of probability of the two transitions, are the same for both (induced) absorption and induced emission [11, 12, 14, 15]. Which of the two transitions is more effective for the interaction with radiation depends only on the relative distribution of the molecules between the two states.

In the normal case, according to Boltzmann's law,

$$N_2 = N_1 e^{-\Delta E/kT} \quad (5)$$

where N_1 is the number of molecules in the ground state and N_2 that in the excited state; the ground state is occupied to a considerably greater extent than the excited state (not strictly true for rotational levels) so absorption mainly occurs (ΔE = energy difference, k = Boltzmann's constant, T = temperature in K). However, if intermediate energy levels are produced by high-intensity radiation (optical pumping), leading to an inversion in the populations of the two energy levels, induced emission predominates. This principle is utilized in lasers [21].

2.3. The Lambert – Beer Law

Many years ago, Bouguer, Lambert, and Beer discovered a relationship between the number of particles in a sample, their properties, the path length of the sample, and the observed attenuation of light [22, 23].

2.3.1. Definitions

When radiation penetrates a medium, an exponential relationship exists among the "extinction" (attenuation of transmitted radiation), the concentration of absorbing particles in the medium, and path length. The extinction is the sum of the effects of absorption and scattering. The proportionality constant is the *molar* (decadic) *absorption coefficient* ε_λ in the case of absorption and the *turbidity coefficient* in the case of scattering [24, 25]. By taking the logarithm of the ratio of the radiant powers, i.e., of the nonabsorbing reference [$\Phi(\lambda,0)$] to that observed passing a sample [$\Phi(\lambda, d)$], a linear relationship is obtained:

$$\begin{aligned} E(\lambda) &= \log \frac{\Phi(\lambda, 0)}{\Phi(\lambda, d)} \\ &= \varepsilon_\lambda c d \\ &= a_\lambda d \end{aligned} \quad (6)$$

In the definition of extinction (E, decadic internal absorbance), the sample thickness d is expressed in centimeters, and the concentration c

of the solutions in moles per liter. IUPAC definitions of these quantities are given in Table 1, with their meanings and formal relationships [23, 26, 27].

Table 1. Physical quantities relating to radiation, with symbols and units in accordance with the IUPAC standard

Quantity	Symbol	Unit
Radiant power (radiant energy Q)	$\Phi_i = dQ/dt$	W
Radiant power transmitted	Φ_t	W
Radiant power reflected	Φ_r	W
Radiant power absorbed	Φ_a	W
Radiant intensity	I	W/sr, W
Reflectance or reflection factor	$\varrho = \Phi_r / \Phi_0$	1
Absorptance or absorbance factor	$\alpha = 1 - \tau$	1
Internal transmittance	τ_i	1
Internal absorptance	α_i	1
(Decadic) internal absorbance	$A = -\log_{10} \tau_i(\lambda)$	1
Linear (decadic) absorption coefficient	$a(\lambda) = A/d$	cm^{-1}
Molar (decadic) absorption coefficient	$\varepsilon = A/(c\,d)$	$L\,mol^{-1}\,cm^{-1}$
Molar concentration of absorber	c	mol/L

For investigation of solid bodies or thin films, molar concentrations such as those used for solutions cannot be quoted. In these cases, absorption coefficients a, with dimensions of cm^{-1}, are used instead, which correspond to the product of the molar absorption coefficient and the concentration of the particles.

2.3.2. Deviations from the Lambert – Beer Law

The Lambert – Beer law assumes that monochromatic radiation falls perpendicularly onto the sample. Since the absorption coefficient is a molecule-specific quantity, any chemical change of the molecule or interaction with the solvent, with other components or dissociation/association leads to false results when the Lambert – Beer law is applied (chemical deviations). On the other hand physical deviations are observed if the natural bandwidth of the absorption spectrum of the components in the sample is narrower than the spectral bandwidth (see Section 3.3) of the "monochromatic" radiation produced by the equipment [3, 22]. If more than one component in the sample absorbs radiation, the total effect is the sum of the molar absorption coefficients of the different components times their concentration.

2.4. Photophysics

2.4.1. Energy Level Diagram

The energy level diagram for a typical organic molecule is shown in Figure 6, which illustrates the energy levels and transitions described above. In the absorption (A) of electromagnetic radiation at the resonance wavelengths, various vibrational states are reached in the excited electronic state (S_1), and are coupled with the electronic transition (within 10^{-15} s). Many vibrational and rotational levels exist in each electronic state (S_n), but these are not shown for reasons of clarity. Normally, only the first excited electronic state is important in spectroscopy, because all higher states have very short lifetimes (relaxation times).

In organic molecules in the ground state, electron spins are usually paired (singlet states). However, in inorganic transition-metal complexes in the ground state, triplets, and other states with a larger number of unpaired electrons occur, in accordance with Hund's rule. Therefore, several energy level systems usually exist side by side as solutions of the Schrödinger equation. A direct excitation of the singlet electronic ground state to the excited triplet state (T_1 parallel unpaired spins) is forbidden in principle because of spin inversion, so that the process of induced absorption from singlet states always leads to excited singlet states (S_n) [3, 5, 7, 28].

2.4.2. Deactivation Processes

Figure 6 shows deactivation as well as excitation processes. After excitation to a higher level, the molecule usually relaxes to the vibrational ground state of the S_1 level by thermal equilibration (te). This S_1 level is the starting point for several competitive processes, the preferred one depending on the type of molecule, temperature, and its environment [29]. These include (1) radiationless deactivation (rd), which consists of internal conversion (ic) followed by thermal equilibration (te); (2) intercombination or intersystem crossing (isc); (3) spontaneous emission (F,

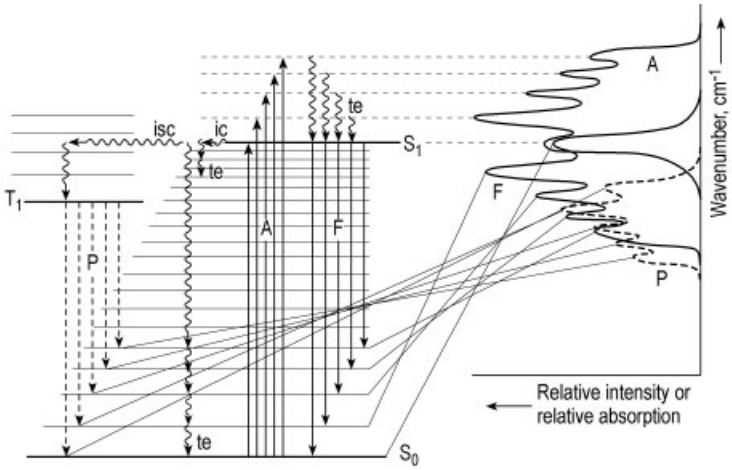

Figure 6. Schematic diagram of energy levels of an organic molecule with singlet ground state (S_0), singlet excited state (S_1), and triplet excited state (T_1) with possible photophysical transitions
The spectra resulting from these transitions are shown on the right. A = Absorption; F = Fluorescence; P = Phosphorescence

P); and (4) photochemical processes (see survey in Table 2) [3,28–31]. The vibrational ground states in electronically excited states have a relatively long lifetime (lifetime $S_1 = 10^{-8}$ s, lifetime $T_1 = 10^{-3}$ s).

In competition with radiationless deactivation, energy can also be lost in the form of radiation (F, P) [3,28]. Fluorescence occurs with molecules that either (1) have extended π-systems (such as polycondensed aromatics); (2) do not permit deactivation by torsional or rotational motion of parts of the molecule; or (3) have no heavy atoms as substituents [32, 33]. In addition to these molecular properties, the environment also plays a part. Thus, the fluorescence intensity increases at low temperature and in solid matrices. This is even more important in phosphorescence, where the $T_1 \rightarrow S_0$ transition is in fact spin-forbidden. If a higher vibrational level ($v' > 0$) is occupied in T_1, in accordance with the Boltzmann equation (Eq. 5), so-called delayed fluorescence [28] can occur by backward intersystem crossing [$T_1 \rightarrow S_1$ ($v = 0$) $\rightarrow S_0$].

Alternatively, during this relatively long lifetime, the molecule can undergo an internal structural change or form new reaction products by an activated collision (photochemistry). This deactivation route is promoted by high light intensities and is used in photochemical synthesis, but it can lead to undesired side effects (e.g., in fluorescence spectroscopy) because photochemical processes reduce the number of molecules that can fluoresce (photobleaching) [5].

2.4.3. Transition Probability and Fine Structure of the Bands

A wave function can be assigned to each energy state. In the induced transition between two energy levels, the overlapping of these wave functions is important and strongly influences the intensity. For this reason, all transitions do not have the same intensity; rather, they have varying transition moments, so that absorption spectrum A on the right of Figure 6 shows a vibrational fine structure with individual peaks of varying intensities [14,31].

Depending on the positions of the nuclei at their equilibrium distances in the molecule in the ground and electronically excited states, these wave functions overlap to varying extents and lead to differing intensity distributions for the vibrational transitions within the entire band. Since atomic nuclei hardly move during the electronic transition (Born – Oppenheimer approximation), the shape of the entire absorption band, according to the Franck – Condon principle, is given mainly by the relative positions of the equilibrium distances of the atoms in the electronic ground and excited states. The lower the degree of interaction between the molecules and their environment, the larger the vibrational splitting becomes. Absorption, fluores-

Table 2. Possible deactivation processes from the vibrational ground state of the first excited singlet state (see also Fig. 6)

Process type	Abbreviation	Name of process	Process description
Radiationless (rd)	te	thermal equilibration	relaxation from high vibrational level in the same electronic state lifetime of excited vibrational level determined by: – external effects (collisions, medium) – inner deactivation (transfer of energy in torsional vibrations to heavy substituents)
	ic	internal conversion	isoelectronic transition within the same energy level system from the vibrational ground state of a higher electronic state into the very high energy vibrational state of a lower electronic state
	isc	intersystem crossing, intercombination	isoelectronic transition into another energy level system ($S \leftrightarrow T$), usually from the vibrational ground state such a radiative transition is forbidden because of the spin inversion prohibition (except for heavy nuclei)
Radiation transfers (spontaneous emission)	F	fluorescence	without spin inversion from S_n (provided that lifetime is 10^{-8} s) within singlet system
	P	phosphorescence	out of triplet into singlet system (provided that lifetime is 10^{-3} s) very low probability, only possible at low temperatures, or in a matrix
Photochemical reactions			photoinduced reaction leading to ionization, cleavage, or a bimolecular step, provided that lifetime in excited state is relatively long

cence, and phosphorescence spectra are shown schematically in the right-hand side of Figure 6.

2.5. Chromophores

Unlike IR spectra (normal vibrations), electron density distributions in UV – VIS spectra do not depend so much on molecular structure or on interaction with solvent molecules. However, changes in the electronic ground or excited state lead to a shift in the relative position of the energy levels and to a change in polarizability. Therefore, even in the UV – VIS region, molecules may be identifiable by the position and height of the absorption bands. Both are influenced, e.g., by extension of conjugated π-systems by substitution or, in the case of aromatics, by electron-attracting or electron-repelling substituents in the *ortho, para*, or *meta* position and by the effects of solvents and of pH. This is especially important in pharmaceutical chemistry [5].

As shown in Figure 7, $\pi \rightarrow \pi*$- and $n \rightarrow \pi*$-transitions often behave differently. If cyclohexane (b) is replaced by methanol (a) as the solvent, the polarity increases considerably, so that the $\pi \rightarrow \pi*$ band is shifted by 10 – 20 nm towards the red (bathochromic) and the $n \rightarrow \pi*$-band by 10 nm towards the blue (hypsochromic). This shift gives an indication of the effect not only of the solvent polarity on the behavior of the bands, but also of the structure of the system. A change in intensity of the absorption band is described as hyperchromic (intensity increase) or hypochromic (intensity decrease) [5, 31].

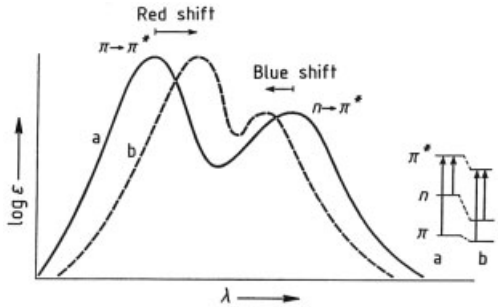

Figure 7. Schematic diagram showing effect of the solvent polarity on positions of the maxima of the $\pi \rightarrow \pi*$ and $n \rightarrow \pi*$ transitions
a) Methanol; b) Cyclohexane

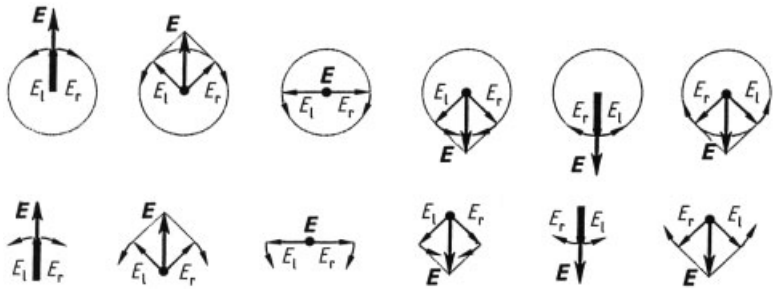

Figure 8. Production of plane-polarized light from two superposed beams of left and right circular-polarized light

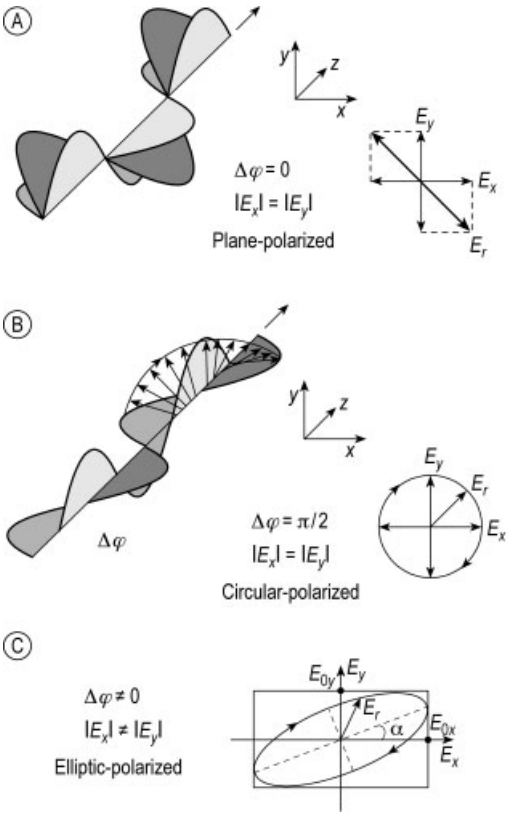

Figure 9. Production of plane- (A), circular- (B), and elliptic- (C) polarized light

2.6. Optical Rotatory Dispersion and Circular Dichroism

2.6.1. Generation of Polarized Radiation

For substances that are optically active or for molecules that are arranged in thin films or with a given orientation, additional information can be obtained with polarized electromagnetic radiation [11, 34]. Light can be plane, circular, or elliptic polarized. Plane-polarized light can be obtained with either polarizing films or special anisotropic crystals. Other possibilities are shown in Figures 8 and 9 [11, 13, 35]. Plane-polarized light is obtained from two circular-polarized beams that rotate in opposite directions and have the same intensity and phase (see Fig. 8).

As shown in Figure 9 A plane-polarized light is formed from two plane-polarized beams whose phase difference $\Delta\varphi$ equals zero. If a phase shift of 90° ($\Delta\varphi = \pi/2$) is chosen, as shown in Figure 9 B, spiral propagation of the resulting field vector is produced, giving circular-polarized light. If the phase difference between two plane-polarized light beams is chosen arbitrarily, or if the amplitudes of the two field vectors are different, elliptic-polarized light (Fig. 9 C) is obtained, which represents the most general form of polarization.

2.6.2. Interaction with Polarized Radiation

The plane of polarization of plane-polarized light can be rotated to the left or the right by interaction with optically active substances. By convention, an observed rotation in a counter clockwise direction when viewed against the direction of propagation of radiation is called levorotatory and is given a negative sign (the opposite holds true for clockwise – dextrorotatory – rotation). Optical activity reveals structural properties usually referred to as chirality (→ Optically Active Compounds, Chap. 1.). These structural properties remain unchanged in molten materials, in solution, and also in complexes. For chiral substances, *quantitative polarimetry* is a very sen-

sitive technique; it is used mainly in pharmacy and medicine to determine concentration.

The two circular-polarized beams of the incident radiation (l, levorotatory; r, dextrorotatory) not only can be influenced with respect to their direction of rotation, but also, in the region of an absorption band, can be absorbed to a different extent by the sample, so that in addition to so-called optical rotatory dispersion (ORD), circular dichroism (CD) or the Cotton effect is observed [35–38]. Along with these classical methods of analysis, modern methods for the analytical investigation of surfaces and boundary layers have become very important, including the use of polarized light in ellipsometry (→ Surface and Thin-Film Analysis) and surface plasmon resonance. Under the influence of external forces (e.g., a magnetic field), even optically inactive substances can be caused to produce magnetooptic rotation.

2.6.3. Optical Rotatory Dispersion

The rotation α caused by optically active bodies is proportional to the thickness d (in millimeters) of the substance through which light passes. The specific rotation $[\alpha]$ for wavelength λ is defined by [36,38]:

$$[\alpha]^{\lambda} = \frac{\alpha^{\lambda}}{d} \tag{7}$$

If an optically active substance is dissolved in an inactive solvent, α is given by Biot's law:

$$\alpha^{\lambda} = [\alpha]^{\lambda} \frac{cd}{100} \tag{8}$$

Here, because of the relatively small rotation of solutions, sample thickness is given in decimeters, and concentration c in grams per 100 mL of solvent. The sodium D line ($\lambda = 589.3$ nm) is often used as incident radiation, and measurement is performed at 25 °C. These parameters are indicated by quoting indices in the form $[\alpha]^{\mathrm{D}}_{25}$. The *molar* or *molecular rotation* $[\Phi]^{\lambda}$ (sometimes $[\alpha]^{\lambda}{}_{\mathrm{M}}$) is given by

$$[\Phi]^{\lambda} = [\alpha]^{\lambda} \frac{M}{100} = \frac{100\alpha^{\lambda}}{dc} \tag{9}$$

where M is the molar mass, c concentration in moles per liter, and d path length in centimeters. Specific rotation is a characteristic quantity for a substance, but it also depends on experimental conditions (temperature, concentration, solvent, wavelength of light) and perhaps time. Along with concentration and temperature, intermolecular forces also play a part. A change of solvent can even lead to a reversal of the direction of rotation.

If optically active substances undergo chemical conversion (e.g., sucrose in acid solution), the specific rotation becomes time dependent (*mutarotation*). For example, when dextrorotatory sucrose reacts until chemical equilibrium is reached, a levorotatory mixture of glucose and fructose is formed (inversion of sucrose).

The specific rotation as a function of wavelength is known as *optical rotatory dispersion* [39]. The angle of rotation usually increases as wavelength decreases. This *normal rotatory dispersion* is found in broad spectral regions that are sufficiently remote from absorption bands of the optically active substance. Rotation can be represented by Drude's equation

$$\alpha^{\lambda} = \sum_{i} \frac{A_i}{\lambda^2 - \lambda_i^2} \tag{10}$$

where A_i is a constant and λ_i is the wavelength of the nearest absorption maximum. This region corresponds to the normal dispersion curve.

2.6.4. Circular Dichroism and the Cotton Effect

The different refractive indices n_l and n_{r} for left- and right-polarized light lead to a specific rotation. In the region of an absorption band the left- and right-hand polarized beam are absorbed to different extents, and the originally plane-polarized light beam becomes ellipticpolarized ($\varepsilon_l \neq \varepsilon_{\mathrm{r}}$). This phenomenon is called circular dichroism. Circular dichroism is coupled with peaks and troughs in the ORD curve (anomalous dispersion) [35,38]. Both anomalous ORD and CD are grouped under the term *Cotton effect* [34].

Figure 10 shows the relationships between the spectral behavior of the two polarized beams with respect to rotation, absorption, and refraction. Only optically active substances interact differently with two polarized beams, so that this difference leads to observable effects. Whereas

in the region of anomalous rotatory dispersion the ORD curve shows a point of inflection, the CD curve has a maximum or minimum [38]. The extreme of the CD curve coincides approximately with the point of inflection of the ORD curve. In analogy to molar rotation, the molar ellipticity of the CD curve is given by

$$[\Theta]^{\lambda} = 100\,\Theta^{\lambda}\,\frac{d}{c} \tag{11}$$

Where Θ is the ellipticity of the transmitted beam in angular degrees, d is expressed in centimeters, and c in moles per liter. In the SI system, molar ellipticity should be expressed in $\deg\,\mathrm{m}^2\,\mathrm{mol}^{-1}$.

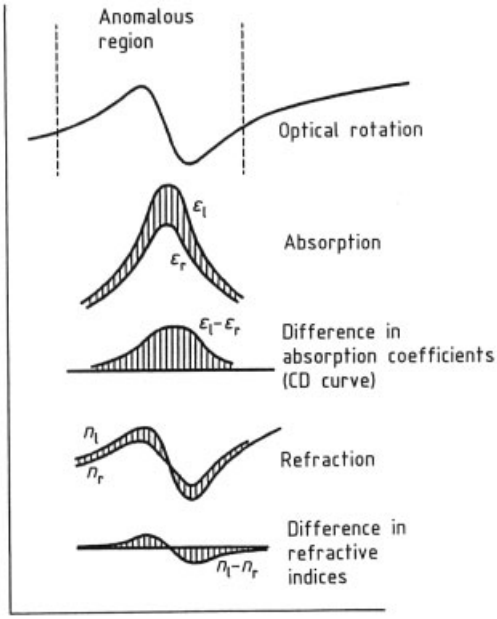

Figure 10. Rotation and circular dichroism optical rotatory dispersion in normal and anomalous spectral regions [38]

The relationship between rotation and refractive index or between ellipticity and absorption coefficient can be represented, in analogy to dispersion theory, by the model of coupled linear oscillators or by quantum mechanical methods. ORD and CD are related to one another by equations analogous to the Kramers – Kronig equations [34].

2.6.5. Magnetooptical Effects

If plane-polarized radiation enters a transparent isotropic body parallel to a magnetic field, the plane of polarization is rotated (magnetooptical rotation, MOR, or Faraday effect) [40]. The angle of rotation α is given by

$$\alpha = \omega l H \tag{12}$$

where H is the magnetic field strength, l the path length, and ω the Verdet constant [38] . The degree of rotation depends on the substance, the temperature, and the wavelength of the radiation. The wavelength dependence is known as *magnetooptical rotatory dispersion* (MORD). Since the Faraday effect is practically instantaneous, it is used in the construction of rapid optical shutters and for the modulation of optical radiation paths. If Zeeman splitting of atomic and molecular terms takes place in the magnetic field, this leads to *magnetic circular dichroism* (MCD) [38,40].

All of these effects are very important in laser technology for modulation and, as an extension of absorption spectroscopy, for the assignment of electronic transitions.

3. Optical Components and Spectrometers

3.1. Principles of Spectrometer Construction

Photometers or spectrometers in the ultraviolet and visible spectral region either are built only for the measurement of light absorption, fluorescence, reflection, or scattering, or have a modular construction so that they can be used for several measuring operations. Spectrometers always include a polychromatic light source, a monochromator for the spectral resolution of the beam, a sample holder, and a detector for measuring the radiation that has been modified by its passage through the sample. Depending on the measuring problem, these components can be modified, and the light path changed [3,4].

3.1.1. Sequential Measurement of Absorption

In the measurement of absorption, the radiation is measured against a reference for each wavelength (monochromatic) after attenuation by a sample. By this reference measurement, such effects as reflections at the cell windows or other optical surfaces can be corrected for, if sample and reference differ only with respect to the substance being examined and are otherwise optically identical. The equipment can be of the single- or double-beam type. In *single-beam equipment*, radiation of the selected wavelength is passed alternately through the reference cell and the sample-containing cell before it strikes the detector. In older, more simple equipment, the position of the sample holder in the cell compartment is adjusted manually. Usually, only a small number of wavelengths are measured because a spectrum cannot be recorded automatically.

Photometers can be of the *broadband type* (broad spectral band light source + filter) or the *narrow-band type* (line spectrum light source + filter) [41,42]. In more expensive equipment, continuous radiation sources (broadband) are used, and the filter is replaced by a monochromator, so that a complete spectrum can be observed point by point. Only a very good monochromator can give resolution (monochromaticity) as good as that from a combination of a line source and filter since in the latter system, the very small width of the natural line determines the bandwidth of the radiation used [41, 42].

Single-Beam Equipment. Modern single-beam equipment first records a spectrum of the standard in the reference optical path and then the spectrum of the sample. The internal microprocessor serves not only for control but also for calculating the intensity ratio according to the Lambert – Beer law (see Eq. 6). The optical light path of such equipment is shown in Figure 11 A [43]. This equipment has fewer optical components and is less mechanically complex (no beam splitter) than the double-beam spectrometer, but the time required for each sample measurement is longer (successive measurements of dark, reference, and test spectra) [44]. Good stability of the light source and of the electronics is necessary to avoid any drift between measurements.

Double-Beam Equipment. In double-beam equipment, the two light paths (through the sample and the reference) are automatically continuously interchanged during the wavelength scan. A rotating mirror, oscillating mirror, or half-transparent mirror splits the beam after it emerges from the monochromator. However, for accurate measurement, at each wavelength, the "dark" signal must be observed as well as that for the reference cell and the sample. The optical paths in equipment with single and double monochromators are shown in Figure 11 A and 11 B.

High recording rates lead to errors in the spectral resolving power (which is fixed by the slit width of the monochromator) because sample and reference values are no longer determined at the same wavelength [3]. In many types of equipment, a rotating mirror, which is divided into four segments, is used to measure two dark periods, one reference, and the sample. Usually the mirror wheel rotates at 50 Hz, so that a maximum of 50 points per second is observed. Wavelength scan rates can be as high as > 1200 nm/min or > 20 nm/s. Thus, at 50 Hz and this extreme scan rate, the reading is taken only every 0.4 nm (with a very short signal integration time); in addition, the wavelength differs from the reference wavelength by 0.2 nm (restricting spectral resolution to ca. 1 nm). Therefore, a considerably slower recording rate is used for high-resolution spectra. Alternatively, it ispossible to stop at each measured wavelength ("stop and go") [45,46].

In double-beam spectrometers, fewer problems of stability of the light source and the electronics arise. Also, a spectrum can be recorded considerably more rapidly than with single-beam equipment. Disadvantages include the greater number of optical components (each optical component – i.e., mirrors and lenses – reduces the available light intensity due to reflection); the polarization of the light beams because of movement of the optical components; and the cost and complexity of construction [3].

Radiation in the UV – VIS region is of high energy. Therefore, in sequential spectrometers the monochromator is positioned before the sample. In IR equipment, in which the sample

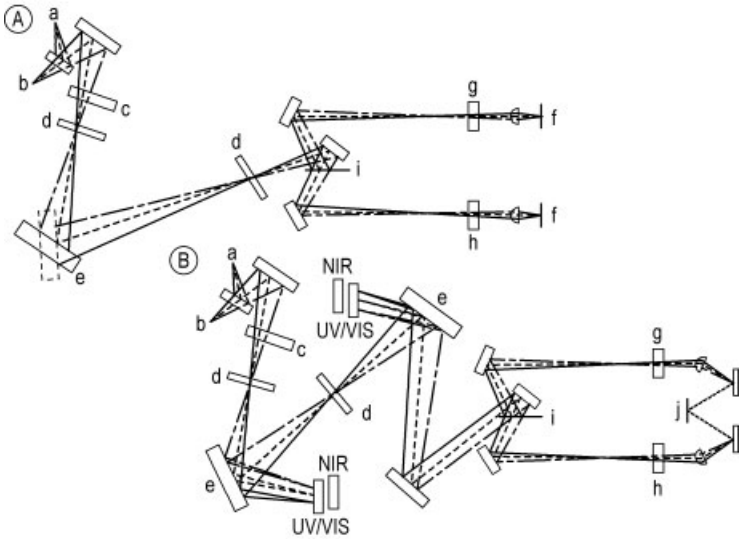

Figure 11. Optical light paths in double beam spectrometers [43]
A) Apparatus with single monochromator and two detectors; B) Research apparatus with double monochromator and beam splitter with one detector
a) Halogen lamp; b) Deuterium lamp; c) Filter wheel; d) Slit; e) Monochromator; f) Photodiode; g) Reference; h) Sample; i) Beam splitter; j) Detector

itself emits radiation, the reverse arrangement is used [3].

3.1.2. Multiplex Methods in Absorption Spectroscopy

With multiplex detectors, the radiation of the various wavelengths no longer need to be measured sequentially (one after the other). Instead, by parallel imaging of many wavelengths on a spatially resolving detector they can be measured simultaneously so that a complete spectrum is recorded. This can be done by using linear photodiode arrays or, more recently, so-called charge-coupled devices (CCDs) [2, 3, 47, 48].

For this equipment, the classical single-beam arrangement is used. However, the light striking the sample is not monochromatic, but white. After passing through the sample, light is spectrally dispersed by a monochromator without an exit slit (polychromator) and then strikes an array of diodes. A possible arrangement of such a polychromator with a multiplex detector is shown in Figure 12. The resolution is determined by the specification of the grating and the number of diode elements in the array (256 – 1024).

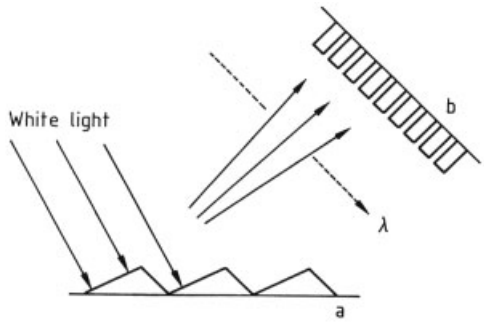

Figure 12. Diagram of multiplex detector (diode array) combined with a grating for spectral dispersion of white light
a) Grating; b) Diode array

By this method, complete spectra can be recorded in fractions of a second (1 – 100 ms, depending on the equipment). Within these times, a reference measurement is impossible so, as with modern microprocessor-controlled single-beam equipment, the reference and sample can only be observed consecutively (except when using a double diode array configuration). The sample is exposed to white light, which can result in

photoinduced processes. However, this type of equipment has the advantage of recording hundreds of spectra within a few seconds, so many dynamic processes can be observed in real time. This requires that neither the diode array nor the light source drifts within the time necessary for measurement, which is true to only a limited extent. In fact, the problems are the same as those with microprocessor-controlled single-beam equipment. The application of this type of equipment for routine measurements, in kinetics, and as a detector in chromatography is continually increasing [49].

Hadamard Spectroscopy. If, instead of the exit slit, an irregular arrangement of slits of various widths is used, and an arrangement in the form of a comb-like second slit is moved in front of this, for each relative position a different summation of intensities at various wavelengths is obtained on a nonmultiplex detector. By using convolution functions, spectra can be computed rapidly for a cycle of different slit positions. Since for mathematical deconvolution so-called Hadamard matrices are used, the principle is known as Hadamard spectroscopy. It has gained some importance in waste gas analysis [50].

Fourier Transform Spectroscopy. After the development of diode arrays, the use of Fourier transform spectroscopy, which is very successful in the IR, NIR, and FIR regions, has not been very widespread for routine analyis in the UV–VIS region. In Fourier transform UV spectroscopy [11], the polychromatic light beam is split in a Michelson interferometer, 50 % falling on a fixed mirror and 50 % on a second moving mirror, and then recombined by the beam splitter. Each position of the moving mirror gives a phase relationship between the two partial beams. After recombination of the component beams the modulation of intensity caused by the position change depending on time can be transformed by Fourier analysis to spectral information (wavenumber) [51–54]. Although modern computers calculate such spectra in fractions of a second, they can give high resolution only if the moving mirror is positioned exactly by using laser interference ($\Delta x \approx 1/\Delta\lambda$).

3.2. Light Sources

3.2.1. Line Sources

In filter photometers, the spectral bandwidth can be improved considerably if line sources are combined with filters. Mercury line sources [56] emit various wavelengths, as shown in Figure 13. Such equipment is used mainly in clinical analysis, because in general, single wavelengths are only used for photometry or in simple kinetic experiments. Mercury vapor arcs are also employed in polarimeters. By using these, rotations at ca. 10 wavelengths can be measured. In polarimetry, the use of sodium vapor arcs is also common. These emit the doublet 589.0 and 589.6 nm (sodium D line). In the past, only these wavelengths were used to determine optical rotation. The sodium lamp is rarely used today for routine measurements; however, the sodium D line is still indispensable for calibration and reference measurements.

3.2.2. Sources of Continuous Radiation

In spectrometers, a light source supplying continuous radiation must be available. Normally, a high-radiant power source that emits in the visible range (tungsten – halogen lamp) is combined with an ultraviolet emitter (hydrogen or deuterium lamp) to provide a wider spectral range.

Tungsten – Halogen Lamps. By addition of iodine to incandescent lamps a higher filament temperature can be reached, and consequently a higher radiation density at the shortwave end of the spectrum. The emission maximum of a tungsten – halogen lamp is in the green part of the spectrum. At wavelengths < 300 nm, the intensity decreases rapidly unless the lamps are operated at high current (high filament temperature), because the intensity maximum is shifted toward shorter wavelengths by an amount proportional to T^4 (Wien's law). However, the lifetime of these light sources is thereby decreased drastically [57, 58].

Hydrogen and Deuterium Lamps. Deuterium lamps, whose emission maximum is at ca. 200 nm, are used mainly in sequential spectrometers. The emission maximum of hydrogen

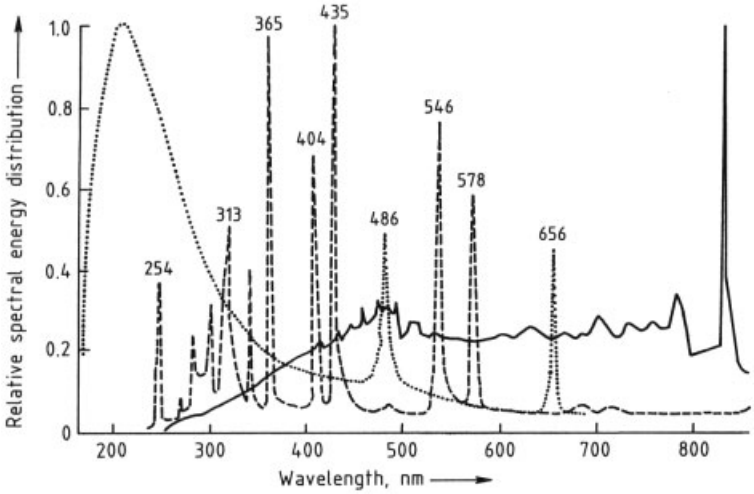

Figure 13. Line and continuous spectra from various light sources
—· = Xenon lamp; – – – = Mercury vapor arcs; · · · · · = Hydrogen lamp

lamps is at a longer wavelength. At > 400 nm, both types have rather low intensities, and show the familiar Fraunhofer lines at 486 and 656 nm whereby the lines of the two isotopes differ by fractions of a nanometer [57,59]. These lines can be used for wavelength calibration. The user must realize that equipment that is fitted only with deuterium lamps can be used only to a limited extent for measurement in the visible range.

This is a very important consideration, particularly for diode array spectrometers, since these, like sequential routine equipment, must often cover the range between 200 and 800 nm. Despite the higher costs and greater complexity of the high-voltage supply, *xenon lamps* are preferable because they cover the entire 200–800-nm range with sufficient intensity, as shown in Figure 13. However, in xenon lamps, a marked structure at 490 nm is present in the continuum so that balancing the two optical paths is difficult in this range. These light sources are also used in ORD and CD equipment.

3.2.3. Lasers (→ Lasers)

Lasers represent a special type of light source [16,21,60,61]. They are used in trace analysis by fluorescence measurement or laser-induced fluorescence (LIF) (→ Laser Analytical Spectroscopy, Chap. 3.) [62–64], in high-resolution spectroscopy, and in polarimetry for the detection of very small amounts of materials. Lasers can be of the gas, solid, or dye type [21]. In dye lasers, solutions of dyes are pumped optially by another laser or a flash lamp and then show induced emission in some regions of their fluorescence bands. By tuning the resonator the decoupled dye laser line can be varied to a limited extent, so that what may be termed sequential laser spectrometers can be constructed [65]. In modern semiconductor lasers, pressure and temperature can also be used to "detune" the emission wavelength by 20–30 nm [66,67].

3.3. Selection of Wavelengths

In simple equipment, only relatively wide transmission bands are selected by means of colored glass or interference filters. A working beam of narrow bandwidth is obtained by combining this with a mercury source. Interference filters give spectral bandwidths in the 3–10-nm range [42, 68]. These operate on the principle of multiple reflections between layers of dielectric material that have been produced in a definite sequence by vacuum deposition. The wave character of light leads to interference of the reflected beams at many interfaces. A suitable choice of thicknesses of the evaporated films enables a narrow band containing a small number of wavelengths to be selected.

3.3.1. Prism Monochromators

Prisms are relatively expensive and very difficult to manufacture in the highest optical quality. Furthermore, dispersion by diffraction is not linear, so that so-called cam disks must be used to give a wavelength scaling that is as linear as possible over the entire range. Prism monochromators are mechanically very complex and therefore very expensive. They were at one time used widely in the UV region. Today they are often used as premonochromators (predispersion devices) since the decomposition of white light into rainbow colors by a prism gives only one spectrum order [42].

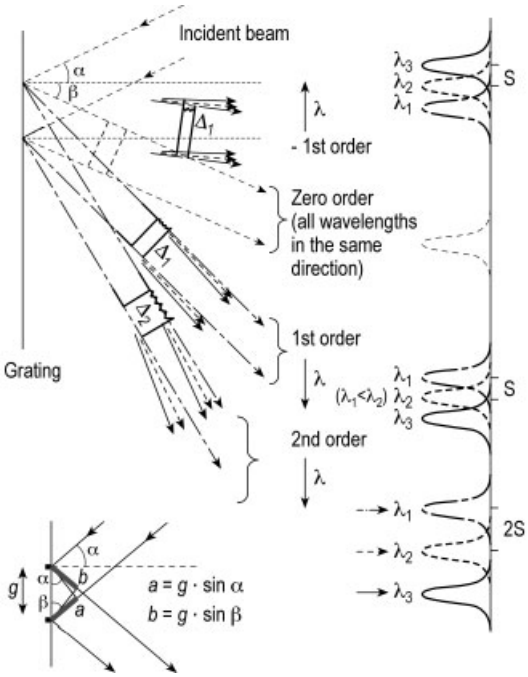

Figure 14. Principle of the grating
Beams reflected at the grating grooves overlap, giving constructive interference at an angle β for each wavelength except the zero-order case. The separation (resolution) of wavelength maxima S on a projection surface for the second order is twice that for the first order.

3.3.2. Grating Monochromators

In the dispersion spectrum produced by gratings, several wavelengths of different orders appear at the same place. As shown in Figure 14, incident white light is reflected at a number of grooves that are either ruled mechanically on a plane surface or produced holographically by superposition of two laser rays. The reflected beams interfere in different ways, depending on their wavelength, the distance between lines of the grating, and the angle of reflection, so that constructive or destructive interference can occur. Thus, reinforcement at a given wavelength takes place only in a certain direction. However, several orders are produced (see Fig. 14) so that, e.g., the first-order reflection of 400 nm and the second-order reflection of 200 nm form an image at the same place. The greater the number of grooves, the more do the partial beams interfere, and the more significant is the destructive interference at certain wavelengths [11,42].

The dispersion of the grating is almost linear. However, as already mentioned, individual orders overlap so that either an order filter must be fitted or a predispersion device (prism) must be used. If the grooves are blazed, reflected intensities of mainly one order can be collected. This is at a maximum for a particular angle (blaze angle) and a particular wavelength (blaze wavelength).

An inlet slit is usually located in front of the prism or grating symmetrical to the exit slit. If the inlet and exit slits are coupled and have the same width, the slit function (see below) forms a triangular intensity distribution. In high-quality equipment, the first monochromator is followed by a second. These double monochromators have considerably better resolving power, although less light intensity is available since part of the light is cut off by the nonparallel beam through each slit. However, in a qualitatively high-grade double monochromator, a central slit must also be present, adjusted synchronously with the inlet and outlet slits; otherwise slit function is impaired [42,69].

The slit function can be observed well in the imaging of line sources when the inlet and outlet slits (and the middle slit if present) are moved synchronously. If the slit is too wide the line of a mercury lamp does not appear as the expected Gaussian-shaped curve intensity distribution, but as a triangle. If the inlet and exit slits are different, a trapezoid is obtained. Triangular shapes in the spectrum indicate defective adjustment of slit widths in the equipment. These effects are observed mainly in spectrometers in which only a small number of preset slit widths

can be selected. Distortion of a spectral band is negligible only if the spectral bandwidth of the monochromator (slit width) is not more than one-tenth the natural bandwith of the sample [3, 42, 70].

3.4. Polarizers and Analyzers

The polarizer and analyzer separate electromagnetic radiation into an ordinary and extraordinary beam. The so-called Glan – Thompson prism is suitable for this. The path of the ordinary beam, whose mode is parallel to the plane of incidence (incident beam normal to prism edge), is refracted at the edge of the prism, while the extraordinary beam continues along the axis [11, 34]. Simple polarization filters (polarizing films) are also often used, in which two mutually perpendicular, plane-polarized beams are absorbed to different extents. Such polarizing films can be obtained by dyeing highly stretched (oriented) plastic film with dichroic dyes. Additional information (e.g., phase dependence) can be obtained by modulation. This technique is very important, especially in equipment for spectral measurements, particularly ellipsometry, because automatic balancing is possible. Modulation can be achieved by either rotational or oscillatory movement of the analyzer or polarizer, or by superposition with the inclusion of a Faraday modulator, which utilizes the Faraday effect described in Section 2.6.5 [11, 38]. Since spectral polarimeters (ORD equipment) are also used in the shortwave UV region, prism assemblies that include Canada balsam adhesive (which absorbs in this spectral range) must be excluded. Calcite can also cause problems because it absorbs at wavelengths below ca. 220 – 250 nm [38]. For these reasons, quartz Rochon prisms are used mainly today [34, 71].

3.5. Sample Compartments and Cells

3.5.1. Closed Compartments

Light from the monochromator is usually convergent and is focused on a minimum beam diameter at the position of the cell. The less convergent beam, the easier can measurements be performed in cells having long optical paths (10 cm) without the light grazing the sidewall of the cell. To prevent this grazing, the position of the cell in the path of the beam should be checked. In the visible region of the spectrum, a simple sheet of paper is suitable for this (monochromator set at ca. 550 nm and with a very wide slit) [3] . It should be placed perpendicularly to the beam to monitor the green spot with respect to the cell position. Furthermore, the edges of the cell and the regions of adhesive at the edges should be inspected for any radiation. In the UV region, paper containing optical brighteners that emit blue light can be used.

Depending on spectral region, the correct material must be selected from appropriate catalogues, for example, because cells of glass or polymer are not transparent to UV radiation. The plastic cells often used in routine analysis differ in their optical properties. This holds true for cells from the same batch and from different batches [72]. Because of the manufacturing process used, their interior surfaces are not parallel, so the reproducibility and potential applications of these cells are limited.

In the sample compartment, as much space as possible should be available to allow for thermostatically controlled sample holders or other equipment (Ulbricht globes, praying mantis devices, multicell holders, sippers). A secondary cooling circuit is useful to protect the optics and electronics from high temperature. Large sample compartments require well-thought-out, carefully calculated optical paths. Photochemical reactions can be followed directly in a spectrometer by special accessories [45].

3.5.2. Modular Arrangements

In modular construction, monochromators, light sources, and detectors can be arranged in various ways adjacent to the sample compartment, so that absorption, fluorescence, and reflection measurements are possible by using the same set of equipment components.

3.5.3. Open Compartments

Diode arrays combined with optical fiber arrangements are very well suited to this modular concept. The possible variety of such arrangements is illustrated by the combination of

a diode array with optical fibers in Figure 15 [55]. The optical path between the light source (a) and the sample (b), and between the sample and the detection system (f), is realized by means of glass or quartz fibers, so that with one experimental arrangement, absorption (c), fluorescence (d), and reflection (e) measurements are possible by a variable arrangement of optical components. This type of equipment is very important for color measurements and becomes increasingly applied in optical sensor technology [55, 73]. In principle, the use of sequential recording equipment with designs based on optical fibers is also possible [74, 75].

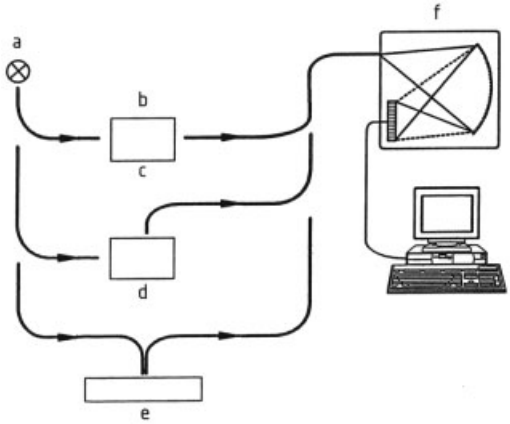

Figure 15. Combination of diode array and glass fibers in various arrangements for measuring absorption, fluorescence, or reflection

3.6. Detectors

Detectors are of mainly two types, based on either the external or the internal photoelectric effect [59, 76–78].

Photomultipliers. In photomultipliers, the external photoelectric effect is utilized (→ Photoelectricity, Chap. 2.). Photons incident on a photocathode, which have an energy greater than the work function of cathode material expel electrons, which are captured by an anode that is positively charged relative to the cathode. The number of electrons produced is proportional to light intensity. They can be measured by the current flowing between electrodes.

Dynodes are often placed between the photocathode and the anode as intermediate electrodes. High-quality photomultipliers have 10–14 dynodes (see Fig. 16). This internal amplification is critical because in addition to electrons from the light-induced process, other thermally activated electrons give rise to a dark current. Photomultipliers are therefore often cooled for the measurement of low light intensities.

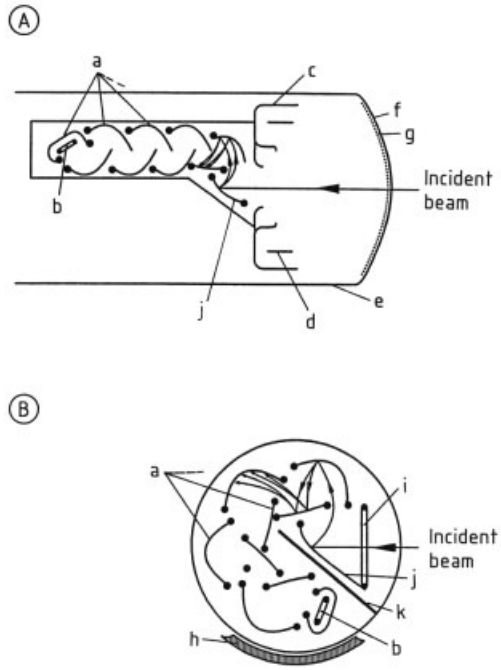

Figure 16. Types of photomultiplier
A) Head-on arrangement; B) Side-on arrangement (on a larger scale)
a) Dynodes; b) Anode; c) Focusing electrode; d) Focusing ring; e) Conductive internal coating; f) Front plate; g) Semitransparent photocathode; h) Envelope; i) Grill; j) Opaque photocathode; k) Shield

The head-on arrangement (Fig. 16 A) has a larger inlet angle and is extended lengthwise, with a correspondingly greater distance separating the dynodes (lower field strengths). It can accommodate more dynodes than the side-on arrangement, and therefore has a greater range of amplification. However, response times are longer. The side-on arrangement (Fig. 16 B) is considerably more compact, the distances between dynodes are shorter and the response times are relatively short. However, the "capture an-

gle" is smaller. For detecting only a small number of photons, the more slowly reacting head-on photomultiplier is preferable [79].

Photodiodes, on the other hand, use the internal photoelectric effect (→ Photoelectricity, Chap. 3.) [77] . When these diodes are arranged in arrays, they are known as diode arrays or diode matrices. Multiplex detectors are used in simultaneous spectrometers [80].

An example is the Reticon array shown in Figure 17. Also, CCD arrays (charge coupled devices), which were developed for television cameras and fax machines, have also become important in spectroscopy (Fig. 18) [4, 49, 81–83]. Apart from the advantage of being multiplex devices, these charge-transfer components have characteristics more similar to those of photomultipliers than normal photodiodes.

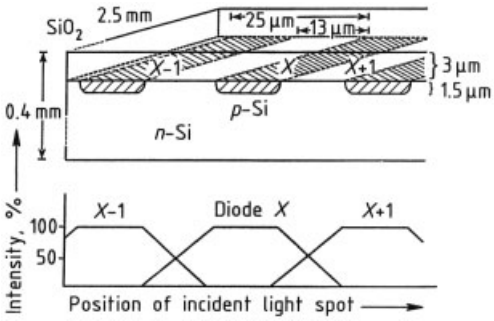

Figure 17. Linear photodiode array [3, 81]
The *p*-doped islands can be found in the *n*-silicon under a layer of SiO_2. In the region of these *p*-islands, photons can produce charges that discharge the capacitors arranged in parallel.

3.7. Optical Paths for Special Measuring Requirements

3.7.1. Fluorescence Measurement

Unlike transmission spectroscopy, in fluorescence measurement the sample is observed perpendicularly to the incident excitation beam. In modern equipment, monochromators are placed both in the path of the excitation beam and the observation beam (see Fig. 19). Since no direct comparative measurement is possible in fluorescence, the intensity of the excitation light beam is usually controlled at least by a semitransparent mirror and a photodiode. To measure a *fluorescence spectrum*, the wavelength of the excitation beam is kept constant and the wavelength of the observation monochromator is varied, so that the excitation monochromator can be replaced by a filter in simple equipment.

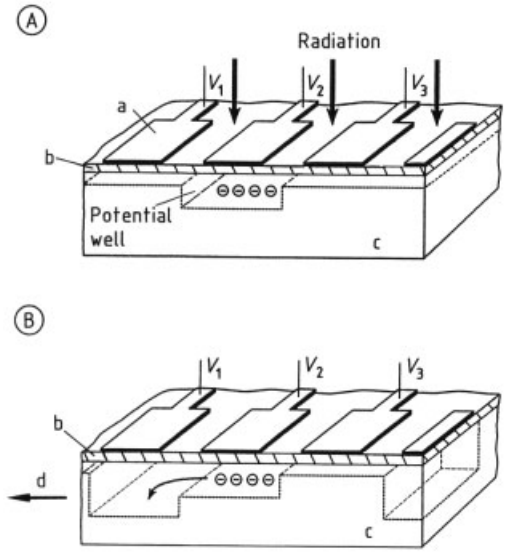

Figure 18. Charge-coupled device (CCD) [55]
A) Free charges are produced by incident light and localized in potential wells (countervoltage) ; B) Periodic application of different voltages causes charges to be displaced from the potential well like a shift register
a) Metallic electrode (gate) ; b) SiO_2 (insulator) ; c) *p*-Type silicon

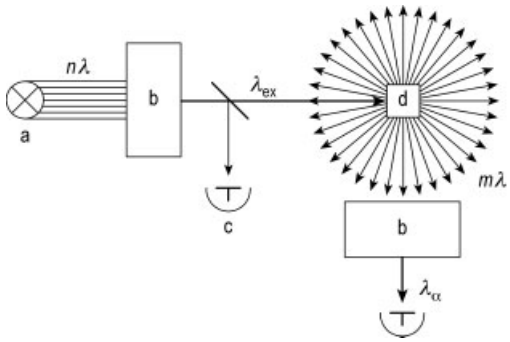

Figure 19. Diagram of optical light paths in a fluorescence experiment
a) Light source; b) Monochromator; c) Photomultiplier; d) Sample

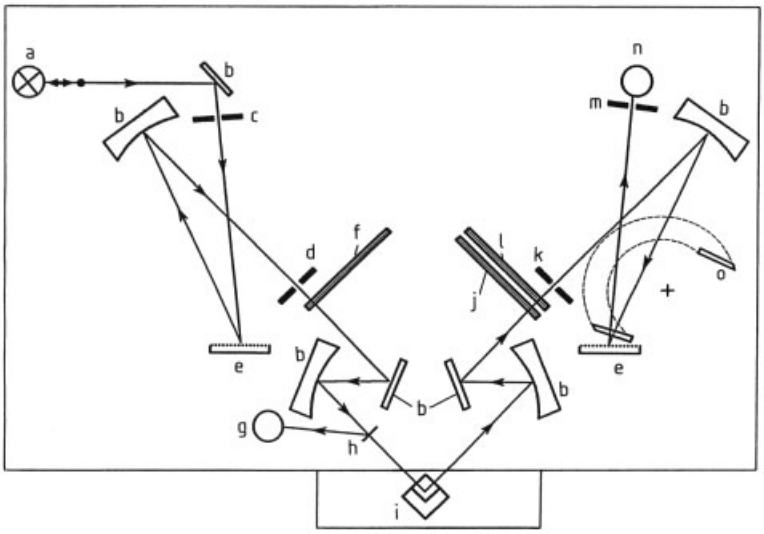

Figure 20. Light paths in commercial fluorimeter from Perkin – Elmer [43]
a) Light source; b) Mirror; c) Slit excitation entry; d) Slit excitation exit; e) Grating; f) Excitation filter wheel with excitation polarizer accessory; g) Reference photomultiplier; h) Beam splitter; i) Sample; j) Emission polarizer accessory; k) Slit emission entry; l) Emission filter wheel; m) Slit emission exit; n) Sample photomultiplier; o) Zero-order mirror accessory

For observation of *fluorescence excitation spectra*, the monochromator at the observation end is adjusted to a wavelength of highest possible fluorescence intensity, and the wavelength of the monochromator on the excitation side is varied. Hence, for these spectra, a shape analogous to an absorption spectrum is obtained [84]. If the absolute fluorescence spectra must be measured, the equipment must be calibrated for all wavelengths by using fluorescence standards [85]. This operation is highly complex and can be performed only with special equipment and appropriate software. Commercial equipment of this type is illustrated in Figure 20 [43].

3.7.2. Measuring Equipment for Polarimetry, ORD, and CD

Simple *polarimeters* consist of a polarizer, several slits, a light source (usually the sodium D line), and a detector, which in the simplest case is the eye of the observer. To adjust an optimal minimum of intensity passing the analyzer half-shaded apparatus are used. The measuring field is separated into two adjacent areas which are adjusted to equal brightness for an adequate analyzer position. Equipment with self-equalizing polarimeters is more complex. Self-equalizing polarimeters contain an additional Wollaston prism and a Faraday modulator, as shown in Figure 21. The polarizer (a) and analyzer (e) are crossed with respect to each other. Radiation from the polarizer is split in the Wollaston prism into two beams of plane-polarized light, whose planes of polarization are perpendicular to each other. The prism (c) is positioned such that the plane of polarization of the incident light forms an angle of 45° to the two beams passing the prism (without sample). The two split beams are of equal intensity under such conditions. Passage through Faraday modulator causes the direction of rotation of the two partial beams to oscillate, and they combine to give a constant radiation flux. This situation is disturbed when an optically active substance is placed in front of the Wollaston prism. An additional oscillating radiation flux falls on the photomultiplier. By use of a feedback system, the polarizer is rotated until this signal disappears.

In *spectral polarimeters*, another monochromator is introduced in front of the polarizer. In this type of ORD equipment, the Wollaston prism and polarizer are often combined by pivoting the polarizer to and from in a defined angle. CD equipment is often provided as an additional device for classical absorption spectrometry. With these devices, absorption of left and

right circular-polarized light can be measured independently. For this, combinations of Rochon prisms (as polarizers) and Fresnel quartz rhombic prisms (for production of circular-polarized light) are used [34]. Equipment of this type is shown in Figure 22.

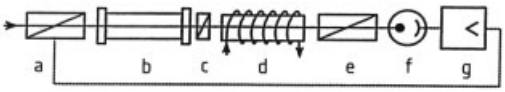

Figure 21. Schematic arrangement of self-balancing polarimeter [38]
a) Polarizer (rotatable) ; b) Sample cell; c) Wollaston prism; d) Faraday modulator; e) Analyzer; f) Photomultiplier; g) Amplifier
– = Feedback between detector and polarizer

3.7.3. Reflection Measurement

Reflection can be either regular (at smooth surfaces) or diffuse (by material whose small particle size causes scattering). *Diffuse reflection* is used mainly in thin-film chromatography [87–89], to evaluate the Kubelka – Munk function at an angle of $< 60°$ from the Lambert cosine squared law [4, 90]. Diode arrays can also be used as detector [91].

To observe *regular (specular) reflection*, goniometer equipment is used in combination with optical fibers and spectrometers. A modular system of this type is shown in Figure 15.

3.7.4. Ellipsometry

If measuring equipment for ORD and CD (Section 3.7.2) is combined with a reflection system, either ellipsometry at one wavelength or spectral ellipsometry can be carried out at defined angles (Fig. 23). In this apparatus, a rotating polarizer is used to produce rotating plane-polarized light. After reflection, an analyzer is used to measure the phase and attenuated relative intensity of the two beams (i.e., ellipticity). From the modulated signal, Hadamard matrices can be used to calculate the refractive index and absorption coefficient of the sample as a function of wavelength [93–95].

3.8. Effect of Equipment Parameters

In many spectrometers, some parameters (e.g., slit width, time constant, and amplification) cannot be chosen freely. Hence, their effects on measurement must be known. These settings depend on the type of spectrum being observed (e.g., a spectrum for routine analysis; a rapid scan spectrum, or high-resolution spectrum).

The following properties of the sample or the equipment can affect measurement [3, 5, 70, 96, 97]:

1) After being switched on, all equipment requires time to warm up until the electronics, light sources, and optical components are in thermal equilibrium.
2) The number of digits that can be read correctly from the display is affected by noise. The number of correct digits does not necessarily correspond with those displayed.
3) The cell compartment as well as the cell holder should be temperature controllable to protect the electronics and optics from high measuring temperature.
4) To prevent any effects of fluctuation in voltage supply on the "bright value" (100 % signal), such fluctuations should be controllable even under load.
5) Measurements in the spectral range < 190 nm can be performed accurately, only if the entire optical light path and sample compartment can be spilled by nitrogen.
6) Blaze wavelengths of ca. 320 nm ensure good measurements down to 200 nm even if the transmission of the optical components and sensitivity of the detector decrease in this range. However, this short blaze wavelength cannot usually be compensated for by higher detector sensitivity if wavelengths of 800 – 1000 nm are measured.
7) If cells with a long path length are used, they must be filled to the minimum level, and the geometry of the measuring beam must be correct.
8) If the intensity of the measuring beam in the UV is too high, it can lead to photochemical reactions in the sample.
9) In principle, the wavelength scan should be stopped for each measurement to prevent integration of the changing signal with the wavelength. Some equipment includes this

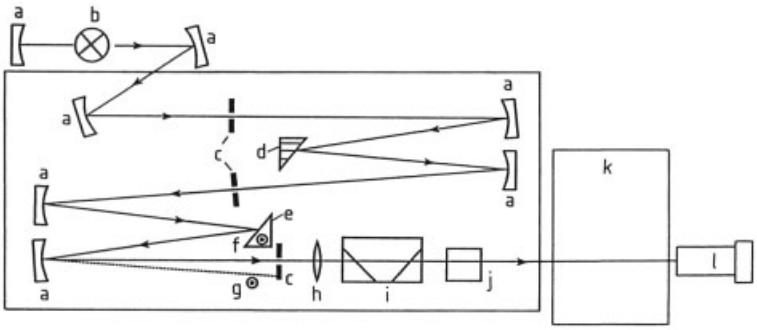

Figure 22. Light path of CD apparatus from Jasco company [86]
a) Mirror; b) Light source; c) Slit; d) Prism (vertical optical axis); e) Prism (horizontal optical axis); f) Ordinary ray; g) Extraordinary ray; h) Lens; i) Filter; j) CD modulator; k) Sample compartment; l) Photomultiplier

facility, but the overall scan rate is reduced [45,46].

Damping (time constant), amplification, spectral bandwidth, and scan rate have a considerable effect on the signal obtained and are interdependent. For these reasons, depending on the spectral characteristics of the sample and the purpose of the experiment (exploratory, routine, or high-resolution spectrum), parameters must be optimized relative to each other. Some criteria are given below. Other information on standards and calibration methods can be found in the literature [98,99].

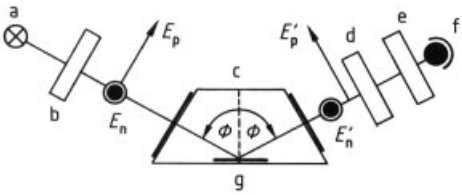

Figure 23. Design of MOSS ES4G ellipsometer from SOPRA [92]
a) Light source; b) Rotating polarizer; c) Measuring cell; d) Analzyer; e) Monochromator; f) Detector; g) Sample

Reproducibility and Accuracy of Wavelengths. In both grating and prism equipment, mechanical wandering of the wavelength scale can occur over time. Therefore, the accuracy of the wavelength must be monitored, and adjusted if necessary, using holmium perchlorate solution, holmium chloride solution, or filters, or by using the lines of a mercury high pressure arc. The simplest method is to use the lines of the hydrogen or deuterium lamp that is part of the equipment.

Wavelength reproducibility can be simply checked by an overlay of repetitively recorded line emitter or holmium perchlorate solution spectra. Since wavelength reproducibility is especially important in kinetic investigations, a chemical reaction can also be used as standard, for which the absorbance at certain wavelengths does not change during reaction (isosbestic points). An example is shown in Figure 24, in which repeated measurements were carried out on a thermal reaction of 2-hydroxy-5-nitrobenzylsulfonic acid [3] .

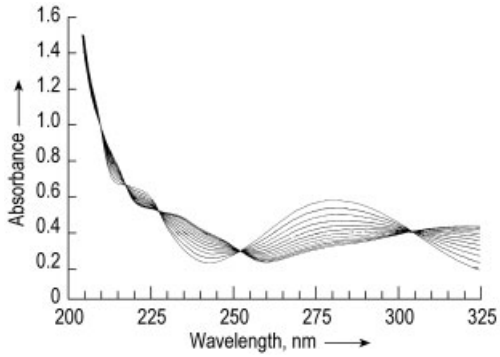

Figure 24. Sulfone reaction spectrum for characterization of wavelength reproducibility of a spectrometer

Photometric Reproducibility and Accuracy. Standards for absolute determination of photometric accuracy are available from, for example, national standards offices (Physikalisch-Technische Bundesanstalt, PTB; NBS). However, the accuracy of these standards (both filters and standard solutions) is not greater than

ca. 0.3 %, and even that can be achieved only by expert operators.

A simpler determination of photometric reproducibility is often carried out with the aid of a solution of potassium dichromate in 0.005 mol/L sulfuric acid [100, 101]. The Lambert – Beer law requires a linear relationship between the signal and concentration. Deviations can be due to several reasons provided chemical effects can be excluded: (1) the recording rate or time constant may be too high; (2) the spectral bandwidth chosen may be too large compared to the natural bandwidth of the sample; (3) too much stray light may be present. As shown in Figure 25, if the spectral bandwidth used is too large, this leads not only to a reduction in the peak height, but also to a displacement of the maximum. A good rule of thumb is that the spectral bandwidth of the equipment should be only one-tenth the natural bandwidth of the substance [3]; otherwise, peak height will be > 0.5 % less than the true value. Therefore, equipment with fixed slit widths is of limited use.

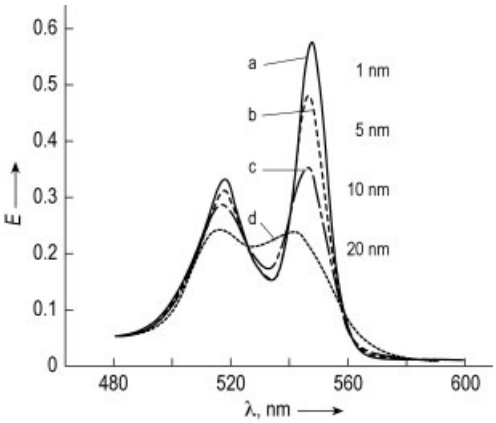

Figure 25. Effect of slit width on spectrum for various ratios of spectral slit width to natural bandwidth
a) 1 : 20 (1 nm); b) 1 : 4 (5 nm); c) 1 : 2 (10 nm); d) 1 : 1 (20 nm)

Stray Light. In addition to radiation of the desired wavelength, every monochromator passes a proportion of so- called stray light (light of different wavelength). If radiation at the extremes of the wavelength range (< 250 nm or > 500 nm) is strongly attenuated owing to high absorption properties of the sample, even stray light of low intensity can lead to severe signal distortion because it usually has a wide bandwidth, is scarcely attenuated by the sample, and strikes the detector, which is not wavelength selective. This effect is most noticeable at absorbances > 2. The transmission of the sample appears higher, a false value of absorbance is recorded, and the absorption maxima are reduced. As shown in Figure 26, totally absorbing salt solutions with cutoff filter properties can be used to reveal stray light [3, 102].

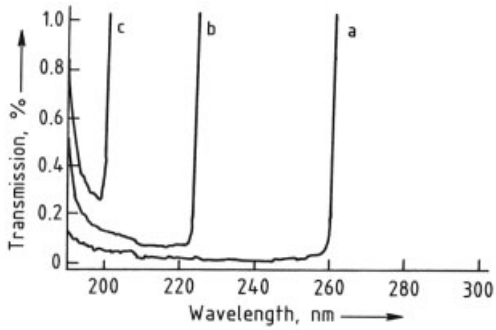

Figure 26. Cutoff filter properties for various totally absorbing salt solutions
a) NaI (cutoff at 260 nm); b) NaBr (cutoff at 220 nm); c) KCl (cutoff at 200 nm)

Another possibility, which is successful mainly in the visible region, is provided by comparative measurements of colored glasses of different thicknesses. As shown in Figure 27, the bands have a tendency to even out in the high-absorption region. In an extreme case, this can be manifest by an intermediate minimum in the resulting spectrum which can lead to misleading interpretations [3, 103].

In principle, the following rules of thumb can be given for recording various types of spectra:

1) For exploratory spectra (recorded at high speed), high amplification and small time constants are used, with a large spectral slit width.
2) For routine spectra, a medium-sized spectral slit width is used, depending on the natural bandwidth, and the scan rate selected

should not be too high. A medium time constant with a medium degree of amplification avoids noisy spectra.

3) For high-resolution spectra, a very low recording rate with large time constants, high amplification, and an extremely small slit width is used.

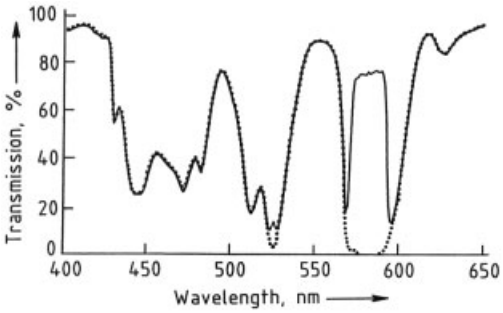

Figure 27. Determination of amount of stray light by a difference measurement of colored glass of various thicknesses
· · · · · = Colored glass 1 mm thick with respect to silica glass 1 mm thick (without chromophore); —·— = Colored glass 4 mm thick measured against colored glass 3 mm thick

The difference between these three types of spectra is illustrated in Figure 28, which shows results obtained on benzene vapor under various conditions by using a spectrometer capable of high resolution.

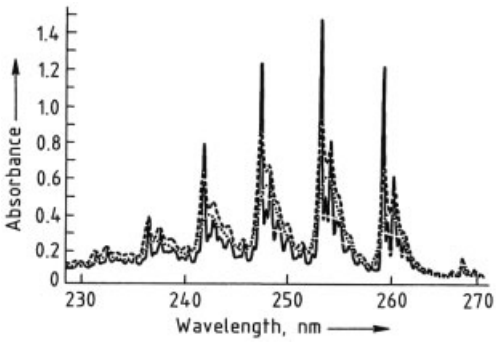

Figure 28. Benzene vapor spectrum recorded with the double monochromator equipment Specord M-500 (C. Zeiss) at various settings of spectral slit widths
· · · = > 2.0 mm; – – – = 0.5 mm; —·— = 0.2 mm

Cleaning the Optical Cell. The type of material used for the cell and the cleaning method are very important for measurement quality. Standard cleaning procedures recommended by cell manufacturer should be used [104]. Under no circumstances should finger marks be left on the cell because this seriously impairs the refractive properties of the cell walls. Alteration of the optical path and the refractive properties of a cell with scratched walls are revealed by a change in the signal of an empty or solution-filled cell when the cell is rotated through 180°.

Since plastic cells generally do not have two plane-parallel walls, fluctuations can occur even in different cells from the same batch [72]. Cells with polytetrafluoroethane stoppers should be used for quantitative measurements because, with other types of cells, the organic solvent can creep over the edges or evaporate.

3.9. Connection to Electronic Systems and Computers

Modern spectrometers have digital readouts and integrated microprocessors for purposes of calibration, control of the wavelength drive, slit adjustment, and sometimes for the interchange of the light sources. They also include interfaces to personal computers sometimes called data stations. The trend is now toward spectrometers with a limited amount of built-in intelligence, which are controlled by personal computers that can also readout data. For this purpose, software for control, data collection, and evaluation is available, operating under commonly used operating systems. In this way, available additional development costs are reduced. Built-in analog-to-digital converters should be provided with preamplifiers to ensure that a wide dynamic range is available.

In the UV – VIS range, personal computers are used in data evaluation, especially in multicomponent analysis. When connected to detection systems in chromatography, they are also used for peak detection, peak integration, and eluent optimization [105]. Although in UV – VIS spectrometry, libraries of spectra do not play an important role, standard spectra are used in chromatography for identification [106].

In larger laboratories, personal computers are linked via a local area network (LAN) to a laboratory information and management system (LIMS) (→ Laboratory Information and Management System) [107].

4. Uses of UV – VIS Spectroscopy in Absorption, Fluorescence, and Reflection

4.1. Identification of Substances and Determination of Structures

Although the main strength of UV – VIS spectroscopy does not lie in the identification of substances, it can often be used as a quick, inexpensive method for identifying certain classes of materials [108–111]. This is especially true in the area of pharmaceutical chemistry, where many otherwise difficult-to-identify substances can be identified with certainty, especially at different pH values and via color reactions. Other examples are aromatic carboxylic acids [111], where the position of the absorption maxima is characteristic and depends on the substituents, for example, with benzadiazepines [112]; and also pregnenones and androstenones [108, 109].

Since steroids contain α, β-unsaturated keto groups absorption maxima can even be predicted by increment rules [108]. In this case, identification is easier by UV – VIS spectroscopy than by the more sophisticated NMR method. The position of methyl or hydrogen substituents relative to the keto groups in the ring enables the wavelength of absorption peaks to be calculated. The spectra of antidiabetics such as carbutamide, tolbutamide, etc. [111], can be calculated and classified readily, although they sometimes differ only in the *para*-substituents on the aromatic sulfonamide ring. Since carbutamide contains an amino group, a significant pH effect occurs that is not observed with the other compounds.

4.2. Quantitative Analysis

4.2.1. Determination of Concentration by Calibration Curves (→ Chemometrics, Chap. 7.2.)

Molar decadic absorption coefficients can be determined from the slope of the absorbance-versus-concentration curve. This slope is known as the sensitivity. Any intersections with axes or nonlinearities can be caused either by impurities inside or outside the cell, by limited photometric linearity of the instrument, or by deviations from the

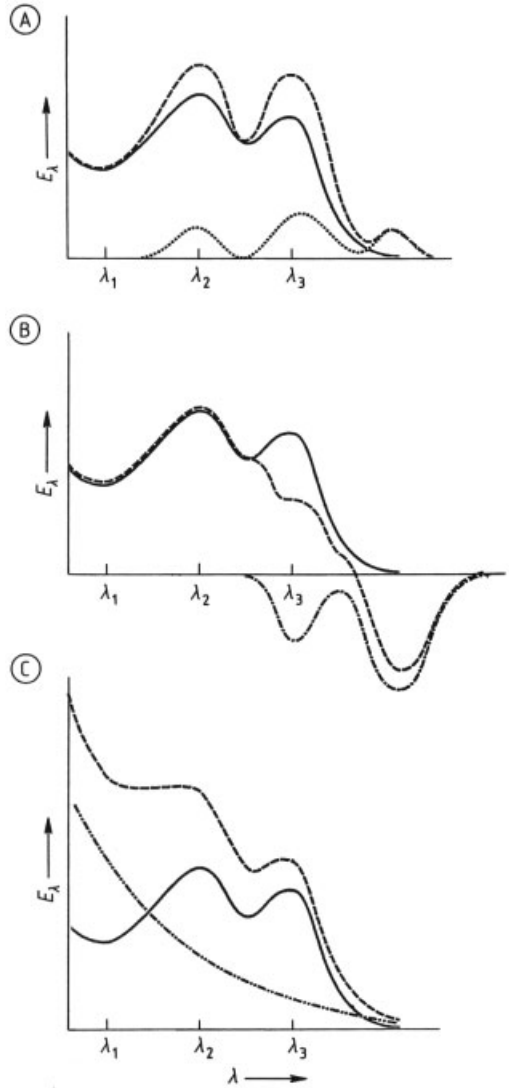

Figure 29. Diagram of absorption spectrum of a pure substance (–)
A) Superimposed on spectrum of an impurity (· · ·); B) With a fluorescent component (– · – · – · –); C) With a scattering substance (– – – – – – – – – – – –)
– – – = Measured spectrum

Lambert – Beer law. Influences on the spectrum of a pure substance are shown in Figure 29 (A – C). These changes are caused by the presence of an impurity, fluorescence of the absorbing substance, or scattering. Three measuring wavelengths (λ_1, λ_2, λ_3) are also indicated in the figure. *Scattering* is apparent at all three measuring wavelengths, although to different extents.

The *impurity* affects only the spectrum of the pure substance at wavelengths λ_2 and λ_3, and fluorescence appears only at wavelength λ_3. By drawing extinction – concentration diagrams for the three cases (see Fig. 30), instead of the "true" absorbance marked by the symbol dot in circle, a false signal is used. Calibration at one wavelength can lead to incorrect determination. The measured absorbance of a sample can lead to various apparent concentrations, depending on what is causing the false reading.

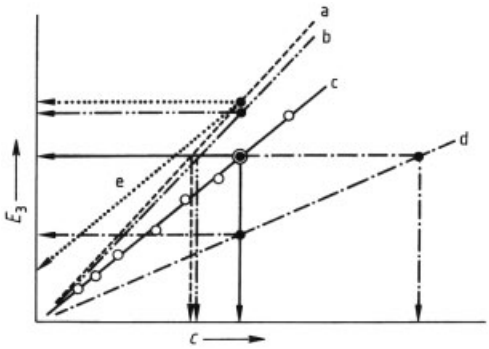

Figure 30. Absorbance plotted as a function of concentration to give a calibration line
Various signals are observed for the same concentration. Same absorbance value corresponds to different concentrations.
a) Substance + impurity; b) Substance + scattering; c) Pure substance; d) Substance + fluorescence; e) Calibration curve if, e.g., finger marks occur on outside faces of cell
○ ○ ○ = Measured points

To ensure correctness of calibration, this operation must always be carried out at several wavelengths that are carefully selected from the spectrum. A simple graphical method of detecting errors in calibration and measurement, and also of recognizing their causes, is given in Figure 31 (A – C). Absorbances are compared for various wavelength combinations by using the same sample. These show a pattern that depends on the type of influence, thereby enabling various types of disturbance to be distinguished. The diagram obtained (see Fig. 31) is known as an extinction diagram [3, 30]. If the influence is due not to the presence of an additional substance in solution, but to an impurity on the cell wall (e.g., finger marks or cells that are not equally clean), this appears in the calibration diagram as an intersection with an axis (i.e., a parallel shift of the calibration line with offset), as shown by curve e in Figure 30 [3].

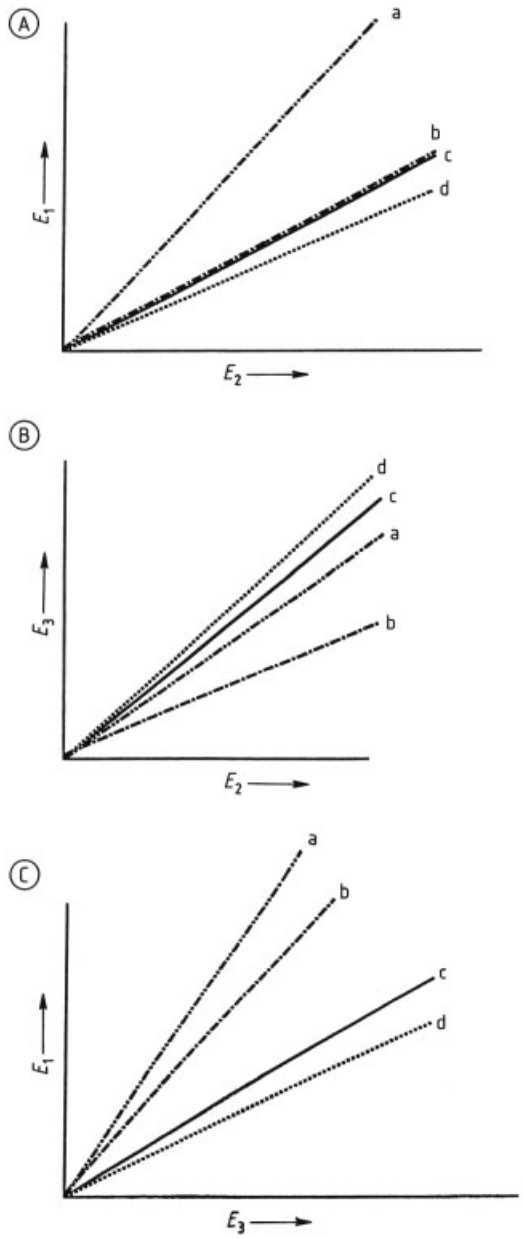

Figure 31. Extinction diagram in various combinations for detection of errors in calibration measurement
a) Substance + scattering; b) Substance + fluorescence; c) Pure substance; d) Substance + impurity

4.2.2. Classical Multicomponent Analysis

Calibration at several wavelengths also enables multicomponent analysis to be carried out in mixtures of substances by using the spectra of pure substances or defined mixtures for calibration (see Fig. 32) [113]. In this case, the Lambert – Beer law for a pure substance is modified, giving the relationship [3]

$$E_\lambda = d \sum_i \varepsilon_{\lambda,i}\, c_i \tag{13}$$

Thus, for three components A, B, C, a linear equation system is obtained, which could take the form

$$\begin{aligned} E_{500} &= d(\varepsilon_{500,\mathrm{A}}\, a + \varepsilon_{500,\mathrm{B}}\, b + \varepsilon_{500,\mathrm{C}}\, c) \\ E_{400} &= d(\varepsilon_{400,\mathrm{A}}\, a + \varepsilon_{400,\mathrm{B}}\, b + \varepsilon_{400,\mathrm{C}}\, c) \\ E_{300} &= d(\varepsilon_{300,\mathrm{A}}\, a + \varepsilon_{300,\mathrm{B}}\, b + \varepsilon_{300,\mathrm{C}}\, c) \end{aligned}$$

$$\boldsymbol{E} = \boldsymbol{\varepsilon}\, \boldsymbol{c}\, d \tag{14}$$

if three wavelengths are being measured. In classical multicomponent analysis, this linear equation system is solved by matrix operations and normal equations. However, in this so-called K-matrix method [6, 114], errors that arise in calibration and measurement must be taken into account. These are represented by the error vector $\boldsymbol{e}$, so that the following matrix equation is obtained

$$\boldsymbol{E} = \boldsymbol{\varepsilon}\, \boldsymbol{c}\, d + \boldsymbol{e} \tag{15}$$

The number of components that can be determined independently in such multicomponent analysis is normally limited to three or four if individual components are not situated favorably in the spectrum with respect to their wavelengths and bandwidths. The ε-matrix is obtained from the calibration measurements. Inversion of this matrix often leads to numerical problems [3, 6, 115].

4.2.3. Multivariate Data Analysis

To take account of interactions between individual components (association, nonlinearities), calibration using multivariate data analysis is often also carried out with mixtures rather than pure substances. Despite this fact, limitations to this method of assessment are encountered quickly. Therefore, the so-called inverse method using the $\boldsymbol{Q}$-matrix is employed, and either principal component regression (PCR) or the partial least squares (PLS) method is used [6, 114, 116] . In both methods, calibration is carried out not with pure substances, but with various mixtures, which must cover the expected concentration range of all components. Within limits, this can allow for nonlinearities:

$$\boldsymbol{c} = \boldsymbol{Q}\boldsymbol{b} \tag{16}$$

Principal component analysis first gives the number of significant components, which avoids a typical error of classical multicomponent analysis, which distributes an unconsidered component among the others. However, principal component analysis reveals it as an additional "impurity." Then a large number of regression steps is carried out with respect to each of the mixtures. In PCR, all possible combinations of factors must be considered since the algorithm does not produce these factors in systematic order whichis the intended aim (quantitative multiple component determination with minimal error). By using the Lanczos bidiagonalization method (PLS), this validation is simplified considerably, since only combinations of the first sequence of factors need be considered [113, 117] .

Another method is the so-called Kalman filter [118, 119]. Iterative calculations enable the spectrum of an unknown component to be determined using a sufficiently large number of calibrating wavelengths. Methods of multicomponent analysis by fuzzy logic have not proved very successful [106]. In contrast, a method using Fourier transforms followed by multicomponent analysis is available commercially [120]. Methods using neuronal networks [6, 121] are used much less in classical multicomponent analysis [122] than in sensor arrays with relatively low selectivity, where they are employed to determine relative concentrations in gas mixtures or liquids [123, 124]. Another wide field of application, in the infrared, is in foodstuffs analysis and pharmaceutical chemistry [125].

4.2.4. Use in Chromatography

In principle, multicomponent analysis can also be used in chromatography if peaks cannot be separated completely. By using a diode array

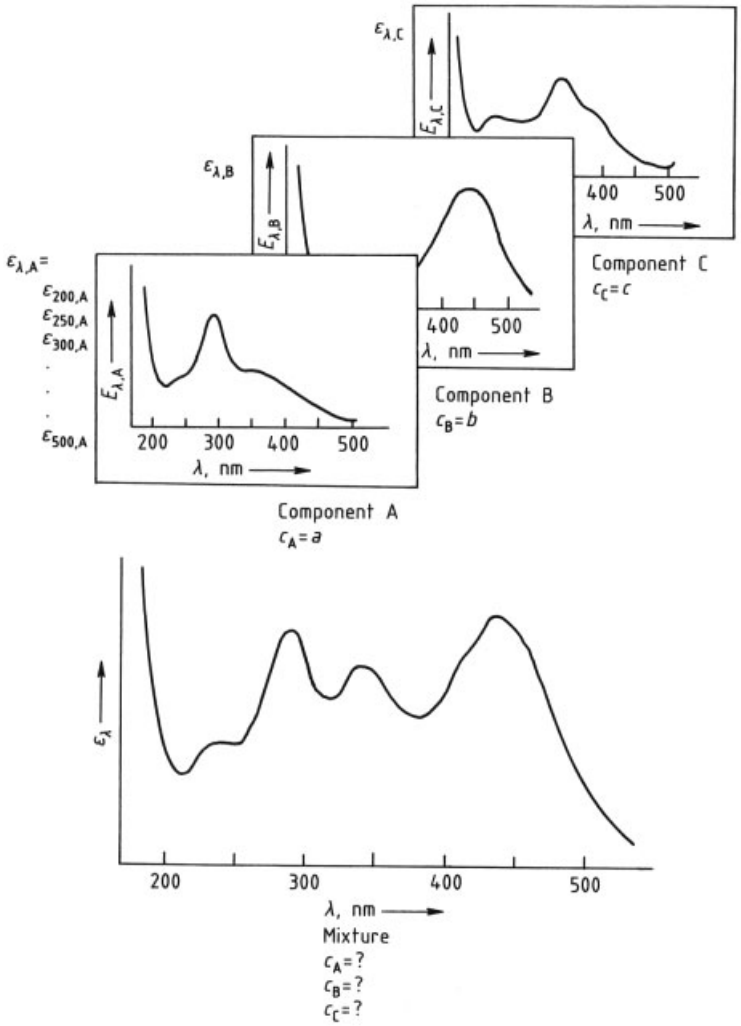

Figure 32. Diagram of pure spectra of three components A, B, and C for calibration, and spectrum of an unknown mixture

while recording a chromatogram, components hidden within a peak can be detected by factor analysis if they are spectrally distinguishable [126]. If their calibration spectra are known, their relative concentrations can be determined.

Every multicomponent analysis should be assessed critically by the analyst, and different algorithms should be used on the same data to improve the level of significance of the evaluation.

4.3. Fluorimetry

As explained in Section 2.4, fluorescence takes place in competition with radiationless deactivation. Fluorescent emission is normally spherically isotropic and of longer wavelength than the exciting radiation. It is extinguished by heavy atoms in the molecule, by the excitation of free rotation about chemical bonds and torsional vibrations, by high concentrations because of increased collision probabilities, or by the presence of paramagnetic oxygen in the sample so-

lution. This can affect reproducibility in quantitative measurements [28, 84, 127].

The observed fluorescence intensity I_α^F depends not only on the absorbing power of the molecule at the excitation wavelength $\varepsilon(\lambda)$, but also on the observed wavelength α within the fluorescence spectrum, depending on its spectral distribution f_α. Measuring fluorescence does not involve any reference measurement, so any fluctuations in the lamps or the electronics, especially long-term drifts, distort the measured signal. An advantage of fluorescence measurement is that by increasing the intensity of the excitation light source, trace analysis can be carried out even with samples that fluoresce only slightly [28, 84, 127]. However, high intensities of the excitation light source can also cause photochemical decomposition reactions in the sample. Even laser dyes have such poor photo stability that two fluorescence spectra of these dyes observed consecutively can give different results [128].

4.3.1. Inner Filter Effects

At high analyte concentrations, inner filter effects can occur. These lower fluorescence intensity by self-absorption if the observed wavelength is incorrectly chosen (within the absorption band). To prevent this quenching due to high concentrations, measurements are taken at the lowest possible concentrations (usually $< 10^{-7}$ mol/L). A rule of thumb is that absorbance at the excitation wavelength should be < 0.02 to prevent errors $> 2.5\,\%$. Under these conditions, fluorescence intensity I_α^F is proportional to the concentration c of the fluorophor [3]:

$$I_\alpha^F = \Gamma I_0(\lambda)\,\eta\, f_\alpha c \varepsilon(\lambda)\,\frac{1-10^{-E(\lambda)}}{E(\lambda)} \qquad (17)$$

Here, η is the fluorescence quantum yield, $E(\lambda)$ the absorbance, and Γ a geometric factor. In practice (cuboid-shaped cell), this factor corrects for nonisotropic intensity distribution. The factor $(1-10^{-E(\lambda)})/E(\lambda)$ describes the finite absorption and becomes E^{-1}, if $E \ll 1$ [3, 28]. A linear relationship exists between observed fluorescence intensity and concentration [3, 28].

4.3.2. Fluorescene and Scattering

If fluorescence is measured in the direction of excitation radiation and analyte concentration is low, much unattenuated excitation radiation falls on the detector. Therefore, measurements are usually made at 90° to the path of the excitation beam. Despite this, the fluorescence signal is usually mixed with scattered light, especially if the solution is cloudy and the substances show only weak fluorescence.

Particularly in trace analysis, artifacts are recorded that occur because of the so-called Raman bands of the solvent. These result in incoherent scattering by the solvent at high excitation intensities (→ Infrared and Raman Spectroscopy, Chap. 2.2.). To avoid this, in good-quality equipment, excitation radiation is produced with the aid of a monochromator rather than a filter, so that these additional Raman lines can be cut off by the second (observation) monochromator. In fluorescence measurements, the spectrum of pure solvent should always be recorded as a "blank" before the sample is investigated.

4.3.3. Excitation Spectra

Whereas absorption spectra describe the relative position of the vibrational level of the first excited state, fluorescence spectra give the position of the vibrational level of the ground state.

If the observation monochromator is kept at the wavelength of the fluorescence maximum and the wavelength of the excitation monochromator is varied, a so-called excitation spectrum is obtained that resembles the absorption spectrum in its spectral distribution. If a mixture of substances contains only one component capable of fluorescence, a fluorescence excitation spectrum can be used to obtain spectral information about this specific component [28, 33].

4.3.4. Applications

Fluorometry is used to determine elements such as boron, silicon, aluminum, beryllium, and zirconium [7, 129, 130], as well as organic compounds [131] (e.g., vitamins [132]). An important application in the field of trace analysis is

the determination of aromatic hydrocarbons in wastewater [133]. Here, the limits of detection can be improved considerably [64] if fluorescence is excited by a laser (LIF). This is very important for the measurement of polycyclic aromatic hydrocarbons [134].

Other applications, especially those involving the use of sensors, have been reviewed [135, 136].

4.4. Reflectometry

4.4.1. Diffuse Reflection

Depending on the surface of the substrate, incident electromagnetic radiation can be reflected either diffusely or regularly (see Fig. 5). If the surface is rough or consists of many small crystallites, diffuse reflection is observed. For a pulverized sample with a very large layer thickness (so that no appreciable reflection of substrate material on which the powder is deposited is observable), the reflectance R_∞ can be related to both the scattering coefficient s and the absorption coefficient k. The reflection behavior of a powdered sample can be described by the so-called Kubelka – Munk function [90, 137]:

$$F\,(R_\infty) = \frac{k}{s} = \frac{(1-R_\infty)^2}{2\,R_\infty} \tag{18}$$

Diffuse reflection has a wide range of uses, e.g., in research, where it is used in the observation of molecules adsorbed on a surface [138] (especially in catalysis), and also in analysis (e.g., thin-film chromatography [89, 139]). To "capture" as much of the diffusely reflected radiation as possible, an Ulbricht sphere is often used. This type of arrangement, in conjunction with optical fibers, is used in production lines in the automobile industry for quality control of bodywork paint [140].

4.4.2. Color Measurement

An important use of reflection spectroscopy is in color measurement [141]. Visual comparisons are replaced by methods based on spectral quantities and the C.I.E. standard colorimetric system (DIN 5033). These techniques are very important in the paint industry [142]. Subtractive and additive mixing of several components can be represented in the color space by means of vectors. Proportions of standard color values based on standard spectral curves, together with the physiological color stimulus specification, give a color stimulus corresponding to the color equation [143]. This method of measuring paint colors is very important, especially in the automobile manufacturing industry.

4.4.3. Regular Reflection

Depending on the change in refractive index at an interface, internal and external reflection can occur. In *external reflection* (transition to a medium of higher refractive index, i.e., into an optically denser medium), reflection occurs at all angles of incidence. The reflection can be calculated from Fresnel's laws (see Eq. 3).

In *internal reflection*, at angles of incidence larger than the critical angle, electromagnetic radiation is totally reflected (attenuated total reflectance, ATR, see Section 2.2.4 and Fig. 5). This special case is very important in analysis for two approaches. First, simple transportation of radiation within the fiber (or a waveguide). Second, in total reflection, an evanescent field appears in which the electrical field vector decays exponentially in the optically less dense medium. Every change within the medium with lower refractive index influences the field vector coupled to the field in the optically denser medium. Therefore, the totally reflected radiation contains information about effects on the other side of the phase boundary (the medium with lower refractive index) [20, 144]. Various principles to interogate this effect are known and used in evanescent field sensors.

If ATR experiments are combined with dichroitic measurements, additional information on orientation and lateral spacing can be obtained. This is especially important if molecules are oriented perpendicularly or parallel at interfaces owing to self-organizing effects [145].

Both approaches of internal reflection (transportation of radiation, application of evanescent fields) are discussed in the following.

4.4.4. Determination of Film Thickness

If reflection occurs at not just one interface but at least two adjacent boundaries, the regularly reflected beams can superimpose and interfere. Interference occurs in certain preferred directions, at certain wavelengths, and with certain substrate and superstrate materials (see Fig. 33 A). If the partial beams extinguish each other, it is called destructive interference; if they reinforce each other, constructive interference [11]. If white light is used, well-resolved interference spectra are obtained with film thicknesses of 1 – 5 µm [146]. Bifurcated fibers are used for incident radiation and its collection after reflection at the phase boundaries.

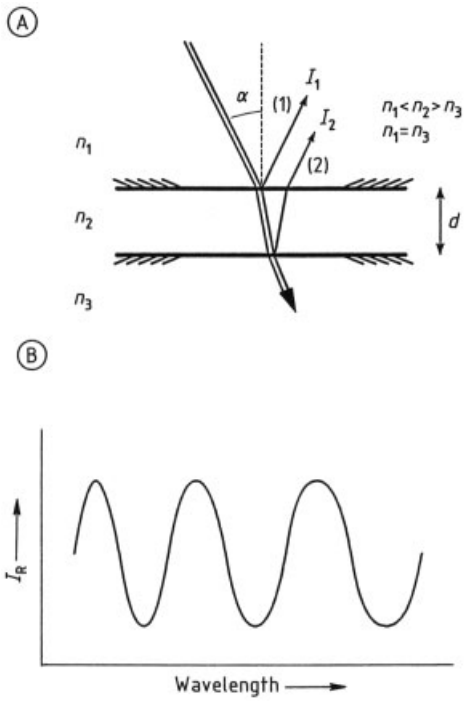

Figure 33. A) Reflected beams at a multilayer system (upper diagram) in which two beams (1) and (2) interfere constructively or destructively; B) Resulting interference spectrum (lower diagram)

An interference pattern of this type is shown in Figure 33 B; the modulation is given by a $\cos(1/\lambda)$ function. Reflection at the phase boundaries leads to reflected intensities I_1 and I_2 which are of the order of a few percent of the total intensity. According to wave theory, these intensities superimpose as follows [11]:

$$I_R = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(2\pi\frac{2n_2 d}{\lambda} + \delta\right) \quad (19)$$

where δ corrects for a phase shift on reflection at the optically denser medium; n_2 is the refractive index of the film; and d is the thickness of the film at whose phase boundaries the reflections take place. According to Equation (19), maxima in the interference spectrum occur where $m = 0, 1, 2, \ldots$

$$2\pi\frac{2n_2 d}{\lambda} = 2\pi m \quad (20)$$

Whereas absolute film thicknesses can be determined only by complete curve fitting of the many distances between maxima of the calculated dispersion curve [124], relative film thickness changes can be determined simply and accurately. According to [11] and [147], from the equation ($\Delta m = 1$)

$$2n_2 d = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \quad (21)$$

the optical film thicknesses can be determined from the spacing of the extrema at the two wavelengths. Minimal changes in refractive index lead to a shift of extrema in the interference pattern.

Improvement of the signal-to-noise ratio and refinement of the evaluation algorithm [148] enable film thickness changes of ca. 1 pm to be observed. The method is dynamic and has response times of < 1 s if diode arrays are used to obtain interference spectra [149]. The method is called reflectometric interference spectroscopy (RIfS) [147]. Similar procedures are also followed in the IR region [116].

4.4.5. Ellipsometry

The method described in Section 3.7.4 for performing ellipsometric measurements using polarized light (see Fig. 23) [150] is very important in microchip production [151]. This method is also used in the characterization of alloys and the investigation of biological films or membranes [152].

The disadvantage of ellipsometry is undoubtedly that it does not represent a direct analytical method, but attempts to match functional relationships as well as possible to the experimental

data by the development of models [93]. On the other hand, ellipsometry is one of the few non-destructive analytical methods that allows measurement of the monolayer or submonolayer. However, even when a "microspot" of measuring light is used, it represents several square millimeters in area. In this surface area, the product $n\,d$ (optical film thickness) is averaged laterally.

Compared to analyses carried out with monochromatic light at a fixed angle, the result is improved by measurement either as a function of angle or as a function of wavelength at a fixed angle [94]. Samples with weak absorbing properties cause problems [153]. Complex refractive indices must then be used for calculations.

However, in contrast to simple reflection measurements or by surface plasmon spectroscopy at one angle or wavelength [153], spectral measurements with ellipsometry enable n and d to be determined independently.

4.5. Resonance Methods

4.5.1. Surface Plasmon Resonance

Surface plasmon resonance (SPR) is a typical method to interrogate evanescent field effects. The measuring system consists of a thin metal film produced by vacuum deposition onto a glass prism (or half cylinder). Depending on wavelength, angle of incidence, and refractive index at the phase boundary, resonance conditions are produced if electromagnetic radiation leaving the prism excites plasmons in the metal to coupled vibrations [153–156].

Generally, monochromatic radiation is used and the angle of incidence is varied. Resonance leads to attenuation of the reflected beam because some of its energy is used for plasmon excitation. The second possibility, spectral excitation by white light at a preset angle, is also of interest. Here, the resonance wavelength is detected by a diode array [94]. An apparatus of this type, with a flow cell in the path of the signal, is shown in Figure 34. Gold, silver, and occasionally copper are used to form the films.

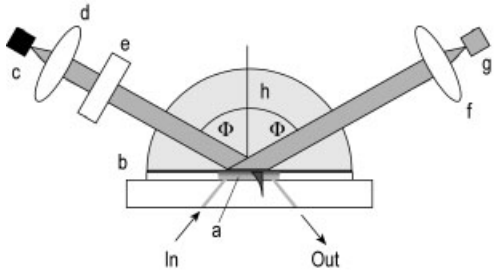

Figure 34. Diagram of system for measuring spectral surface plasmon resonance with a flow cell and resonance point
a) Flow cell; b) Silver layer; c) Light source; d) Collimator lens; e) Polarizer; f) Lens; g) Optical fiber (to the diode array) ; h) Glass cylinder

SPR is used to characterize biological membranes and Langmuir – Blodget films, as well as a sensor in bioanalysis [157, 158]. SPR application to sensors is reviewed in [159].

4.5.2. Grating Couplers

Another resonance technique involves the use of grating couplers, introduced by Lukosz [160, 161]. These can be obtained from the ASI company in Zürich. Grating patterns are imprinted on thin films of highly refractive material that is supported by a carrier material. Depending on the refractive index ratios at the grating, a resonance condition is produced for the film waveguide under the grating, leading to coupling or decoupling of the interfering beams at the grooves of the grating. The principle is shown in Figure 35. Both modes of electromagnetic radiation (transversal electric, TE, and transversal magnetic, TM) can be used, so that a difference is produced [162]. A large number of uses have been published in the field of gas sensors and for rapid analysis with biosensors [163].

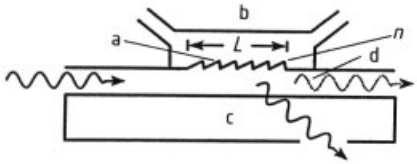

Figure 35. Principle of grating coupler [160]
a) Grating; b) Flow cell; c) Carrier; d) Waveguide

4.6. On-Line Process Control

4.6.1. Process Analysis

By use of flow cells, normal spectrometers can be employed for routine measurement. Since recording a spectrum by sequential spectrometers required a certain amount of time, the absorption was usually observed continuously at one wavelength. The development of diode arrays has enabled this system of measurement using photometers to be superseded. Also, these new types of measuring equipment use optical fibers, enabling measuring cells to be installed at a considerable distance from the spectrometer, so that continuous measurements are possible in dangerous or explosion-protected spaces. Equipment designs such as those shown in Figure 15 are very suitable for such tasks, enabling absorption, reflection, and fluorescence measurements to be carried out with the aid of glass fibers.

By using optical fibers, biotechnological processes in fermentors can be monitored continuously with spectral measurements. In simple cases, transmission measurements are performed [164]; or for monitoring reactions, scattering phenomena are employed. These techniques are also used in wastewater analysis by continuous measurement of turbidity. If multicomponent systems are to be observed (e.g., in the analysis of foods), methods using the near IR are often more suitable [125].

Many analytical problems can be solved by the use of flow cells. If these are combined with flow injection analysis equipment [165], the limits of detection in water analysis can be improved [166] or the scope of biochemical analysis increased [147].

4.6.2. Measurement of Film Thicknesses

When beams of radiation reflected from the two surfaces of a film superimpose, constructive or destructive interference occurs when visible light is used if the film thickness is in the region of a few micrometers, since one partial beam has a longer path length to travel than the other. The resulting phase difference depends on the physical distance between film interfaces and the refractive index of the film (see also Fig. 33). From the interference pattern, the so-called optical film thickness $n\,d$ can be calculated. Use of the diode arrays previously described, in combination with optical fibers, enables rapid measurements of optical film thicknesses to be performed even on films produced at high speed in coating machines. The possibility of monitoring coating thickness at various parts of the strip by means of glass fibers, and of reacting to production problems within a few seconds, makes this technique extremely useful in process control.

4.6.3. Optical Sensors

For a general definition of sensors, see → Chemical and Biochemical Sensors, Chap. 1. The term sensor denotes a small specialized device that operates selectively on several analytes [167] . Optical sensors are certainly competitive with ion-selective electrodes because they are now less complex than in the past and miniaturization promises further success in the future. Furthermore, optical sensors have advantages [168] in remote sensing and by utilization of the spectral information, especially in the case of sensor arrays.

Along with simple optical sensors that measure transmission, so-called optodes [169] play an important role. In these, either the color change of an indicator dye in a membrane is monitored by using diffuse reflection [170], or the effect of an analyte on the fluorescence spectrum is observed. In addition, many other effects are used, such as surface plasmon resonance (see Section 4.5.1), and grating couplers (see Section 4.5.2). In most cases, interaction of the analyte with the evanescence field of a glass fiber, a slab or strip waveguide plays the important part, influencing the wave propagation of the coupled electromagnetic field by changing the refractive index of the superstrate [171]. Furthermore, determination of the thickness of thin films by interferometric methods (see Section 4.4.4) can be used for analysis of both gases [146, 172, 173] and liquids [147, 174]. Here, the amount of swelling of polymer film due to the permeation by the analyte is determined. Accumulations of material caused by antigen – antibody interactions [147] can also be observed by methods sensitive to changes in layer thickness. Label-free

detection of biomolecular interaction by optical sensors has been reviewed [175].

Direct optical detection of DNA hybridisation by SPR [176] or RIfS [177] has been achieved. In routine sensing, assays with labeled compounds are preferred using different fluorescent techniques [159].

Optical sensors have become very important, are used more and more widely, and have been reviewed comprehensively in several monographs [178].

4.7. Measuring Methods Based on Deviations from the Lambert – Beer Law

Effects that lead to deviations from the Lambert – Beer law are given below. On the left-hand side, the factors that influence the spectrum are listed, on the right the chemical and physical aspects are given:

Changes to the spectrum due to	Caused by
Concentration	self-association, dissociation
pH change	acids/bases proteolytic reaction, equilibrium
Solvent	interaction between solvent and solute
Presence of additional component B	complex formation
Time	thermal chemical reaction (including solvolysis)
Incident light	photochemical reaction, photochromism
Temperature	thermochromism

The most important of these are the temperature dependence of the molar absorption coefficient (due to density change); the effect of association and dissociation equilibria, which change with temperature, concentration, and pH; and, not least, concentration changes caused by chemical and photochemical reactions.

Like typical titrations in which electrodes are used, photometric titrations can be carried out by using indicator dyes. Here, the pH is plotted not against the concentration of anions or dissociated components, but against the absorbance at one or more wavelengths [179]. From these diagrams, pK values of even polyvalent acids and bases or of complex equilibria can be determined. Furthermore, it is possible to linearize the diagram, which allows better control of the model of the number of steps of titration [180, 181].

UV – VIS spectroscopy has also found wide use in monitoring thermal and photochemical reactions. Absorbance is measured at several wavelengths during the reaction, and dynamic multicomponent analysis is carried out by using the Lambert – Beer law (see Section 4.2). In the simplest case, wavelengths can be found at which just one component absorbs. The absorbance signal is then directly proportional to the concentration of one reactant and can be evaluated by the usual kinetic equations [182, 183]. An example is given in Figure 24, in which case the thermal reaction has also been used to check the reproducibility of the wavelength of a spectrometer.

The interpretation of photochemical reactions is more complicated if they have to be followed spectroscopically. An example is given in Figure 36, which illustrates the photochemical isomerization of *trans*-stilbene to *cis*-stilbene followed by ring closure to form the end product phenanthrene via dihydrophenanthrene, an intermediate that is not detectable by conventional methods. The diagram shows a photochemical reaction spectrum. The measured data can be evaluated by using various methods [184]. When investigating such reactions, a process-controlled system is usually employed in which the irradiation equipment and the spectrometer are combined [185].

5. Special Methods

5.1. Derivative Spectroscopy

If absorption bands of two compounds overlap (one broad and the other narrow), derivative spectroscopy can be used for resolution. Herewith even trace amounts can be analyzed quantitatively [186]. Derivative spectroscopy is also used to investigate turbid solutions such as industrial wastewaters or biological samples. This method is also suitable for the analysis of shoulder-shaped spectra and overlapping absorption bands. The derivative of transmission with respect to wavelength is not proportional to concentration, but the derivative of absorbance with respect to wavelength is. Generally, the first and second derivatives are used [115]. Higher derivatives present noise problems, as the amount of noise increases with each

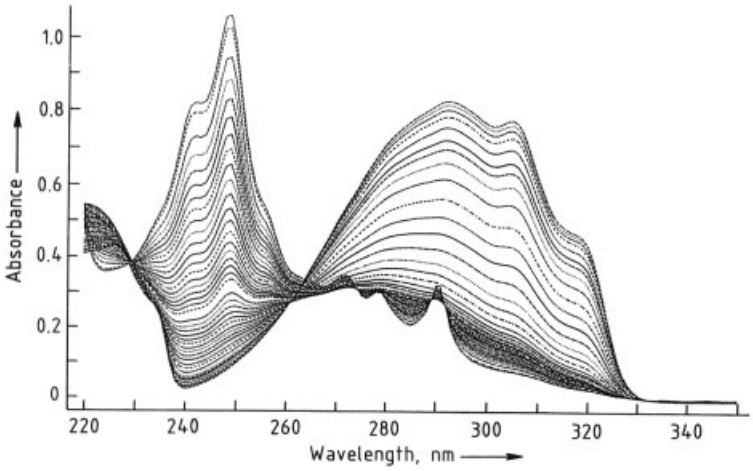

Figure 36. Reaction spectrum of the photoreaction of *trans*-stilbene [183]

Table 3. Various methods of obtaining derivatives, with their advantages and disadvantages

Method	Details of method	Advantages	Disadvantages
Optical	frequency modulation of the light beam by oscillating gratings, slits, or mirrors determination of Fourier coefficients equipment with constant wavelength distance	no effect on the spectrum of variations in the intensity of the light source very good signal to noise ratio	higher derivatives not obtainable special optical construction necessary
Electronic	derivative obtained from the analog signal by differentiation and attenuation circuits	direct on-line spectra higher derivatives possible	derivative as a function of time and not of wavelength
Computed	the digitalized data are stored and the derivative is calculated	direct derivative as a function of wavelength original spectrum is not lost; use of various algorithms optimization of the selected parameters	necessity for an integrated computer or data station

derivative. Higher derivatives have been claimed to be useful for special applications [187].

Table 4. Various mathematical algorithms for smoothing curves and producing their derivatives, with their advantages and disadvantages

Numerical algorithm	Advantages	Disadvantages
Differential coefficient $\Delta E/\Delta\lambda$	simple, quick	signal-to-noise ratio is made worse by subtraction of two measured values that are subject to error
Polynomial smoothing	derivative obtained directly from polynomial coefficients number of restart points determines the degree of smoothing	loss of measured points at the beginning and end of the spectrum problems with sharp peaks
Cubic spline	treatment of the entire spectrum in one step good linear dependence on concentration	large influence of the parameters

Optical, electronic, and mathematical methods are available [3, 188], whose advantages and disadvantages are listed in Table 3. The algorithms generally used for derivation are listed in Table 4, along with their advantages and disadvantages. For three overlapping Gaussian peaks, the normal spectra and the first, second, third, and seventh derivatives are shown schematically in Figure 37. For both the first and the second derivative, by forming the difference fora neighboring minimum and maximum, a linear relationship can be found between these values and concentration. On differentiation, the peaks become narrower and the slopes steeper. Thus, in the second derivative, a steep slope occurs in the calibration curve, and the extreme values are further apart. The derivative signals are less marked if the distance between supporting points is comparable to, or wider than, the width at half peak value.

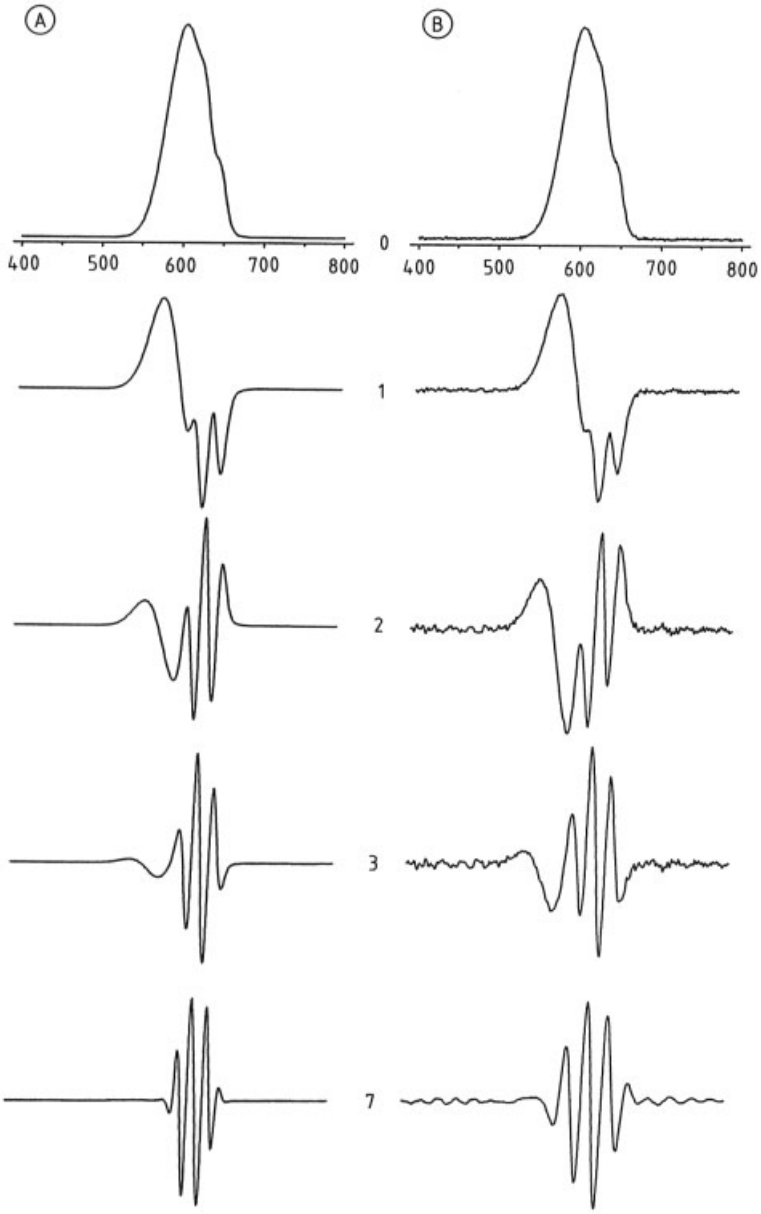

Figure 37. Schematic diagram of various derivatives of three overlapping Gaussian peaks
A) Simulated spectra; B) Spectra with 0.3 % errors superimposed

However, Figure 37 shows not only that the number of peaks increases with higher derivatives, but that when several bands overlap, the number of minima and maxima increases so that the shape of the curve becomes unclear. If real, noisy, raw data are used instead of simulated data, interpretation with the aid of derivatives becomes much more critical (see Fig. 37 B). If too much smoothing is used, information is lost. If too little smoothing is employed, significantly higher derivatives are impossible.

Each differentiation step leads in principle to an increase in noise, so that no additional information can be obtained by this mathematical procedure, although in some cases this form of "spectral manipulation" offers a possible way of

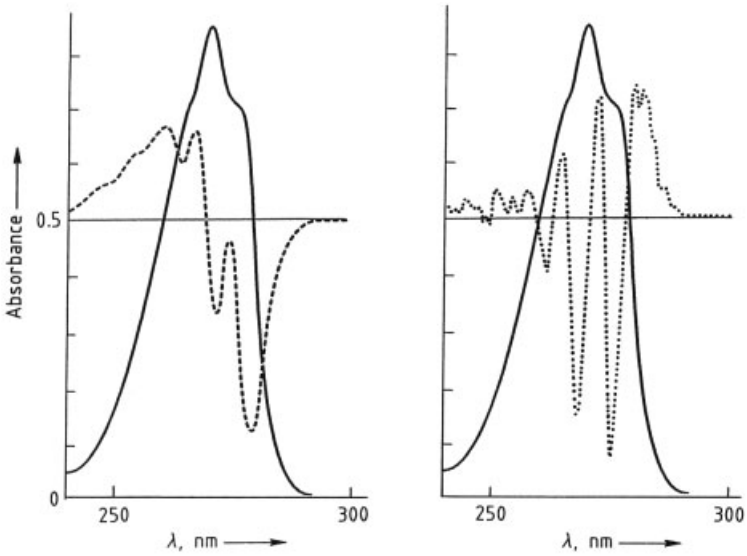

Figure 38. Absorption spectrum of phenol in water (–), showing first (– – –) and second (· · · ·) derivatives of the spectrum [190]

revealing information hidden in raw data. However, as implied by the word "manipulation", the final result is affected by the procedure used.

The potential and limitations of derivative spectroscopy are described in many literature references, which also include several examples of its use [7, 186, 189].

Derivative spectroscopy can be applied to fluorescence and fluorescence excitation spectra [188]. It is also used in chromatography in the search for peak maxima, and in interferometric measurements for determining minima and maxima in interference spectra [124]. A typical application is the determination of phenol in water (Fig. 38) [190] .

5.2. Dual-Wavelength Spectroscopy

Another possible method of investigating biological samples that cause light scattering is by dual-wavelength spectroscopy [191, 192], although this method is limited to very special applications. The number of types of equipment used is therefore also limited. First, normal spectrometers can be used, the signals for two wavelengths being taken from the observed spectra. Second, so-called bifrequency equipment can be used, which operates with two beams at different wavelength at the same time. This is possible either with two monochromators, which cause light of two wavelengths to fall on the sample simultaneously, or with one monochromator and a modulated slit, which provide light of two wavelengths that alternate continuously.

This method has the advantage of avoiding problems from differences in cell positions, differences in the cells themselves due to spurious absorptions or impurities inside or outside the cell, scattering effects, or turbidity. Two techniques are used, first the so-called *equiabsorption method* in which two wavelengths are chosen for which one of the components has the same absorption. The second component must absorb at only one of the two wavelengths. This enables sodium nitrite to be determined, for example, in the presence of sodium nitrate (see also Fig. 39).

The *signal amplification method* is of general application. It is used either if no region occurs in which only one component absorbs or if the interfering component has no absorption maximum. With this type of equipment, the absorption can be corrected by a factor at one wavelength to give the same value as that obtained at the other wavelength (the uncharacteristic spectrum is "evened out"). This enables the concentration of the desired component to be determined from the absorbance difference at the two wavelengths if its absorbance coefficients at

the two measuring wavelengths are known. This method has been used in medicine and for biological samples [192]. However, it has recently decreased in importance.

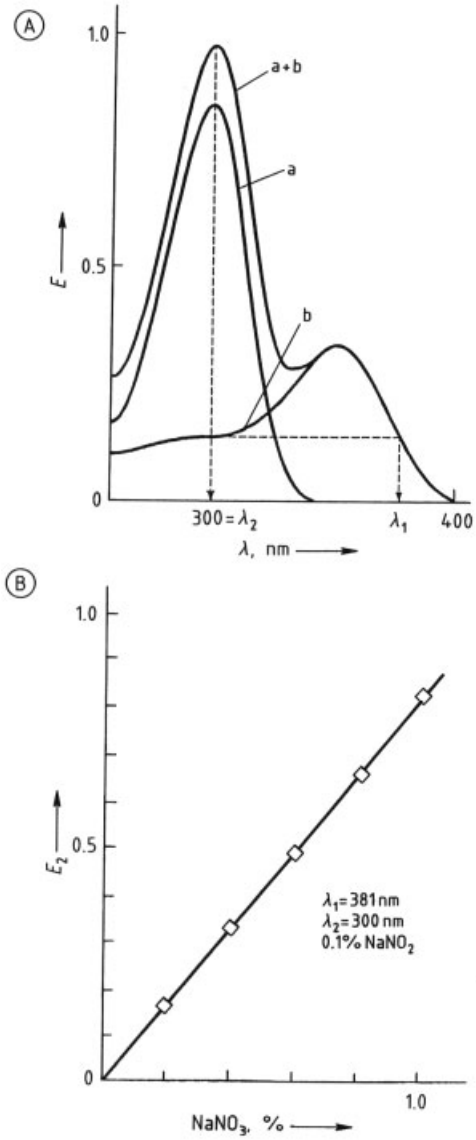

Figure 39. Determination of sodium nitrate and sodium nitrite in the presence of each other [192]
A) Spectra of 1 % $NaNO_3$ (a) and 0.1 % $NaNO_2$ (b) ;B) Calibration curve

5.3. Scattering

As described in Section 2.2.3, scattering of aqueous solutions increases considerably when larger suspended particles such as soot, mist droplets, fat droplets, or bacteria are present. In analogy to the Lambert – Beer law, attenuation of the incident light by a scattering medium is described in the ideal case by

$$I = I_0 e^{-\tau l} \tag{22}$$

where I is the intensity after passage through the medium, I_0 is the incident intensity, l is the optical path length, and τ is the so-called turbidity coefficient. This equation is valid only for scattering particles whose diameter is small compared to the wavelength of the light. The intensity of scattered light depends on the number of scattering particles per unit of volume, the angle at which the scattered light is observed, the size and shape of the scattering particles, and the wavelength of the light [18, 193]. These various relationships can be utilized in analysis [194].

5.3.1. Turbidimetry

Measurement of the attenuation of an incident light beam on passing through a scattering medium is known as turbidimetry and can be carried out with simple colorimeters or photometers. However, spectrometers in which the wavelengths of the incident radiation can be chosen are used for more exact determinations. These are preferable for measuring either relatively low or very high turbidity. Measuring errors can become very large in both cases.

5.3.2. Nephelometry

Nephelometry is used if turbidity is slight. Here, the intensity of scattered radiation is measured, not the difference between incident and attenuated radiation. Under these conditions, errors due to absorption are less likely to occur in nephelometry. The intensity of light deflected to the side is determined; this can be carried out by using fluorometers. Since measurement at an angle of 90° to the incident light direction does not always give the best results (even with particles having a diameter of 1 µm, ca. 90 % of the scattered light is in a cone at an angle of only 10°

to the incident light beam), the equipment must include a goniometer so that measurements can be carried out at various angles [195, 196].

Synthetic turbidity standards are generally used in the quantitative measurement of turbidity. Alternatively, calibration curves can be obtained by precipitating known amounts of the substance to be measured. However, grain size depends not only on precipitation conditions, but also on the time delay between precipitation and measurement, so conditions must be standardized carefully to ensure reproducible results. Nephelometric and turbidimetric methods have also become important in continuous automatic on-line process control (e.g., of wastewater).

5.3.3. Photon Correlation Spectroscopy

Turbidimetry and nephelometry give no information on particle shape. However, if the principle of nephelometry is combined with time-dependent measurement, photon correlation spectroscopy can provide information on the shape of the scattering particle. This dynamic light scattering has become very important in colloid chemistry [197, 198].

5.4. Luminescence, Excitation, and Depolarization Spectroscopy, and Measurement of Lifetimes

If the molecules and their direction of polarization are not distributed statistically in the sample (anisotropy), a preferred spatial direction for the fluorescence exists. If polarized light is used for excitation [199], only those molecules whose axis of polarization is parallel to the axis of polarization of incident light are excited substantially. This results in a preferred direction of polarization for the fluorescence radiation. However, if molecules can move or rotate freely within the lifetime of the excited state, this preferred direction of fluorescence polarization is disturbed (depolarization effects) [200]. These effects enable statements to be made about the environment of the molecules. The degree of polarization is given by

$$p = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \quad p = \frac{3\cos^2\alpha - 1}{\cos^2\alpha + 3} \tag{23}$$

where α is the angle between the transition moments during absorption and emission processes.

Since the lifetime of the excited state is also characteristic of individual substances, and particularly of the environment of the molecule, time-dependent fluorescence measurements with polarized light, with variation of the excitation and polarization wavelength, can be performed. They give important information about the orientation and dynamics of the sample. Many of these methods are successfully applied in biosensing and bioanalytics [159].

The method of *single photon counting* [201, 202] is generally used for such lifetime determinations. A weak pulsed light source causes excitation of a small number of molecules. Spread over a period of time (spontaneous fluorescence is a statistical process), single fluorescence photons are emitted and are "counted" only in a small time window which proceeds between pulses by a ramp technique. By following thousands of pulses, the statistically emitted photons are counted in the channels (time windows), giving the decay curve shown in Figure 40. The time constants are determined by curve fitting.

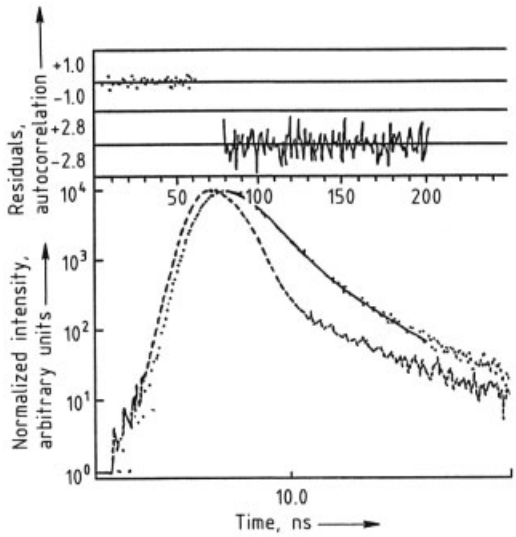

Figure 40. Logarithmic plot of measured intensity of fluorescence against time in a single photon counting experiment Top diagram shows autocorrelation and residuals between fluorescence intensity decay curve (· · ·) and excitation light pulse (– – –) [138]

If several deactivation processes occur at the same time, an incorrectly chosen number of overlapping exponential functions for fitting can lead to physical misinterpretation. Figure 40 shows the intensity of the excitation light source, the decay of fluorescence intensity, and curve fitting. If application of the residuals leads to systematic deviation from the mean zero value, this can indicate a poor curve fit or an incorrectly chosen number of exponential curves.

5.5. Polarimetry

5.5.1. Sugar Analysis

Saccharimetry is very important in the sugar industry. An aqueous solution of sugar is analyzed by determining the optical rotation α, using the green mercury line (546 nm) or yellow sodium line (589 nm). However, the effects of evaporation, the unstable nature of sugar solutions, and temperature effects can easily lead to errors in the determination of optical rotation.

In sugar manufacture, automatic quality control is important, and in yeast production the molasses stream must be determined to have optimum sugar content. Here, too, the optical rotation of molasses is measured automatically in a continuous flow cell [38].

5.5.2. Cellulose Determination

Whereas conventional gravimetric cellulose determination is complicated and time-consuming, cellulose concentrations can be determined rapidly in ammoniacal copper solution by measuring the optical rotation of the levorotatory solution. However, the solutions must be completely free of turbidity [138].

5.5.3. Stereochemical Structural Analysis

Optical rotatory dispersion can be used as a routine procedure, especially for steroids, since the asymmetry of spatial orientation of a molecule is also affected by its conformation. Cholestan-1-one, cholestan-2-one, and cholestan-3-one, for example, can be distinguished because of their different ORD curves (see Fig. 41). Molecular asymmetry (dissymmetry) is a prerequisite of optical activity. A dissymmetric disturbance of the electron system of the keto group caused by a change in the stereochemical arrangement is the case of other ketosteroids (such as androsterone). Here, ORD and CD curves often give information only on relative configuration.

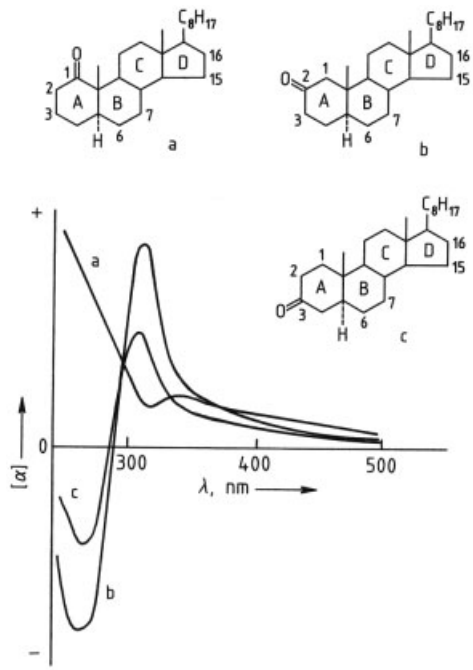

Figure 41. ORD curves of cholestan-1-one (a), cholestan-2-one (b), and cholestan-3-one (c), all dissolved in methanol [38]

Such ORD and CD curves are used for purity control of optically active substances in pharmaceutical chemistry.

5.5.4. Use of Optical Activity Induced by a Magnetic Field

In solid-state inorganic chemistry, MCD is an important technique for the determination of color centers in alkali-metal halide crystals, rare earths, and highly symmetric compounds such as hexahalides of platinum or ruthenium. In organic chemistry, MCD is observed with, e.g., metal porphyrins, pyrimidine bases, and nucleoside derivatives.

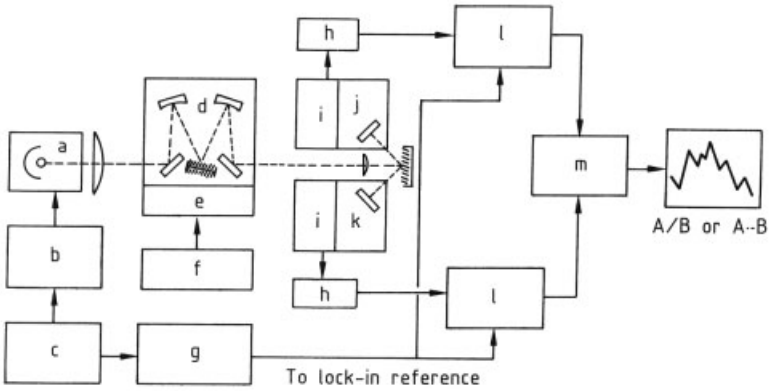

Figure 42. Block diagram of commercial photoacoustic equipment
a) 300-W Xenon lamp; b) Lamp power supply; c) Sine-wave modulator; d) Monochromator (two gratings) ; e) Wavelength drive stepper motor; f) Monochromator controls; g) Modulation frequency display; h) Preamplifier; i) Microphone; j) Cell A ; k) Cell B ; l) Lock-in amplifier; m) Ratiometer/difference amplifier

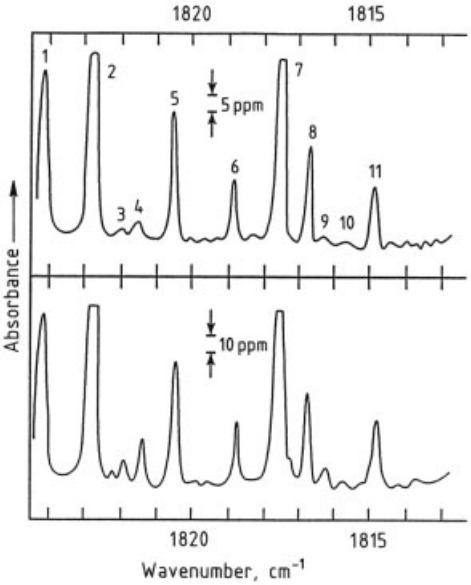

Figure 43. Absorption spectra of gas samples
Upper diagram shows calibration spectra for 20 ppm NO in N_2 . Lines 1, 5, 6, 8, and 11 are due to NO, and other lines to water vapor; lower diagram shows spectrum of automobile exhaust gas

5.6. Photoacoustic Spectroscopy (PAS)

Photophysical processes can take place in combination with the excitation of molecules to more highly excited electronic states (see Section 2.4.2). In the subsequent radiationless transitions, the energy of excited molecules is taken up by the environment (neighboring molecules). Thermal equilibrium is thus reestablished and the temperature of the environment is increased. This is associated with a volume change, which is appreciable for a gas but less pronounced for liquids and solids.

If periodically modulated visible radiation is used, the periodic volume changes cause pressure variations and sound waves that can be detected (e.g., by a microphone). The photoacoustic or optoacoustic effect is based on this principle [203, 204]. Although this has been known for a long time, it has become more important since the late 1970s as an adjunct to reflection spectroscopy if samples are either very strongly or very weakly absorbing. The "heating" of the sample or its environment depends on the extent to which the modulated electromagnetic radiation can be absorbed. This corresponds to the spectral distribution of the molar absorption coefficients. If the wavelength of excitation radiation is varied, the amplitudes of pressure fluctuations are proportional to dynamic changes in the temperature of the system, and thus to the absorption spectrum [205].

Photoacoustic spectra can be observed with a simple microphone and optical apparatus. A diagram of a modern piece of equipment is given in Figure 42, in which there is a blackbody sample at the second input of a look-in amplifier. Corrected photoacoustic spectra can be recorded with this type of monochromator equipment. Frequency-modulated xenon lamps are used as

white light sources, so that a chopper is not needed for modulation.

Photoacoustic spectroscopy is used with gases [206–208], liquids [209], and solid samples [210–212]. Simple solutions of the thermal diffusion equation can be used only under certain conditions. Areas of use [213] include investigation of waste gases (see Fig. 43), absolute measurements of fluorescence quantum yields [214], determination of dye concentrations in films, investigation of thin films in catalysis [210, 215], and biological and physiological problems such as the investigation of leaves and samples of skin or blood. Despite the wide area of application, use of this equipment is very limited so few types of equipment are available on the market. However, individual instances of the application of PAS in sensor technology have been described [216].

6. References

1. J. D. Ingle, Jr., S. R. Crouch: *Analytical Spectroscopy,* Prentice-Hall, Englewood Cliffs, NJ, 1988.
2. G. Svehla (ed.): "Analytical Visible and Ultraviolet Spectrometry," in: *Comprehensive Analytical Chemistry,* vol. XIX, Elsevier, Amsterdam 1986.
3. G. Gauglitz: *Praktische Spektroskopie,* Attempto Verlag, Tübingen 1983.
4. D. A. Skoog, J. J. Leary: *Principles of Instrumental Analysis,* Saunders College Publishing, Fort Worth 1992.
5. G. Gauglitz, K.-A. Kovar in E. Nürnberg, P. Surmann (eds.): *Hagers Handbuch der pharmazeutischen Praxis,* 5th ed., **vol. 2,** Methoden, Springer Verlag, Berlin 1991, p. 157.
6. D. L. Massart, B. G. M. Vandeginste, S. N. Demmig, Y. Michotte, L. Kaufman: Data Handling in Science and Technology 2: *Chemometrics: a Textbook,* Elsevier, Amsterdam 1988.
7. H.-H. Perkampus: *UV-VIS Spectroscopy and its Applications,* Springer Verlag, Berlin 1992.
8. H. H. Jaffé, M. Orchin: *Theory and Applications of Ultraviolet Spectroscopy,* Wiley, New York 1962.
9. M. Klesinger, J. Michl: *Lichtabsorption und Photochemie organischer Moleküle,* VCH Verlagsgesellschaft, Weinheim 1989.
10. P. W. Atkins: *Molecular Quantum Mechanics,* 2nd ed., Oxford University Press, Oxford 1983.
11. L. Bergmann, C. Schaefer: *Lehrbuch der Experimentalphysik,* vol. 6: Optik, De Gruyter, Berlin 1987.
12. M. Born, E. Wolf: *Principles of Optics,* Pergamon, New York 1980.
13. E. Hecht, A. Zajac: *Optics,* Addison-Wesley, Reading 1974.
14. P. W. Atkins: *Physikalische Chemie,* VCH Verlagsgesellschaft, Weinheim 1990.
15. R. P. Feynman, R. B. Leighton, M. Sands: *The Feynman Lectures on Physics,* Bilingna Edition, R. Oldenbourg, München, Addison-Wesley, Reading 1977.
16. W. Demtröder, "Laser Spectroscopy," Springer Series in Chemical Physics 5, Springer Verlag, Berlin 1981.
17. H. A. Stuart: *Molekülstruktur,* Springer Verlag, Berlin 1967.
18. H. C. van der Hulst: *Light Scattering by Small Particles,* Wiley, New York 1957.
19. K. A. Stacey: *Light-Scattering in Physical Chemistry,* Butterworths, London 1956.
20. N. J. Harrick: *Internal Reflection Spectroscopy,* Harrick Scientific Corporation, New York 1979.
21. H. Weber, G. Herziger: *Laser,* Physik-Verlag, Weinheim 1972.
22. G. Kortüm: *Kolorimetrie, Photometrie und Spektrometrie,* Springer Verlag, Berlin 1962.
23. M. Krystek in W. Erb (ed.): *Leitfaden der Spektroradiometrie,* **INSTAND vol. 6,** Springer Verlag 1989, p. 321.
24. H.-H. Perkampus: *Lexikon der Spektroskopie,* VCH Verlagsgesellschaft, Weinheim 1993.
25. K. Laqua, W. H. Melhuish, M. Zander: "Molecular Absorption Spectroscopy, Ultraviolet and Visible (UV/VIS)," *Pure Appl. Chem.* **60** (1988) 1449 – 1460.
26. V. A. Fassel et al., *Pure Appl. Chem.* **45** (1976) 105.
27. J. Inczedy, T. Lengyel, N. Ure: *Compendium of Analytical Nomenclature,* Blackwell Science, Oxford 1998.
28. C. A. Parker: *Photoluminescence of Solutions,* Elsevier, Amsterdam 1969.
29. W. H. Melhuish, *Pure Appl. Chem.* **56** (1984) 231.
30. H. Mauer, G. Gauglitz: "Photokinetics; Theoretical Fundamentals and Applications," in R. G. Compton, G. Hancock (eds.): *Comprehensive Chemical Kinetics,* **vol. 36,** Elsevier, Amsterdam 1998.

31. H. A. Staab: *Einführung in die theoretische organische Chemie,* Verlag Chemie, Weinheim 1966.
32. G. G. Guilbault: *Practical Fluorescence; Theory, Methods, and Techniques,* Marcel Dekker, New York 1973.
33. Th. Förster: *Floureszenz organischer Verbindungen,* Vandenhoek, Göttingen 1951.
34. J. Michl, E. W. Thulstrup: *Spectroscopy with Polarized Light,* VCH Verlagsgesellschaft, Weinheim 1986.
35. H. Rau in H. Naumer, W. Heller (eds.): *Untersuchungsmethoden in der Chemie; Optische Aktivität und Polarimetrie,* Thieme-Verlag, Stuttgart 1990.
36. C. Djerassi: *Optical Rotatory Dispersion Applications to Organic Chemistry,* McGraw Hill, New York 1960.
37. G. Snatzke: *Optical Rotatory Dispersion and Circular Dichroism in Organic Chemistry,* Heyden & Son, London 1967.
38. *Ullmann,* 4th ed., **5,** 628.
39. F. Ciardelli, P. Salvadori (eds.): *Fundamental Aspects and Recent Developments in Optical Rotatory Dispersion and Circular Dichroism,* Heyden & Son, London 1973.
40. J. Michl: "Magnetic Circular Dichroism of Cyclic π-Electron Systems," *J. Am. Chem. Soc.* **100** (1978) 6801, 6812, 6819.
41. DIN Norm 5030 Part 3 (1983), DIN Norm 32 635, Beuth-Verlag, Berlin 1982.
42. A. Reule in W. Erb (ed.): *Leitfaden der Spektroradiometrie,* **INSTAND vol. 6,** Springer Verlag, Berlin 1989, p. 179.
43. Perkin-Elmer, company information.
44. W. Kaye, D. Barber, R. Marasco: "Design of a Microcomputer-Controlled UV – VIS Spectrophotometer," *Anal. Chem.* **52** (1980) 437A.
45. G. Gauglitz, T. Klink, A. Lorch, *Fresenius Z. Anal. Chem.* **319** (1984) 364.
46. Application information by Varian, Darmstadt, Germany.
47. P. M. Epperson, J. V. Sweedler, R. B. Bilhorn, G. R. Sims, M. B. Denton, *Anal. Chem.* **60** (1988) 327A.
48. J. V. Sweedler, R. B. Bilhorn, P. M. Epperson, M. B. Denton, *Anal. Chem.* **60** (1988) 282A.
49. D. G. Jones, *Anal. Chem.* **57** (1985) 1057A.
50. J. A. Decker: "Hadamard Transform Spectrometry: a New Analytical Technique," *Anal. Chem.* **46** (1974) 1803.
51. R. Geick, *Chem. Labor Betr.* **23** (1972) 193, 250, 300.
52. L. Genzel, *Fresenius Z. Anal. Chem.* **273** (1975) 391.
53. A. G. Marshall, F. R. Verdon: *Fourier Transform in NMR, Optical, and Mass Spectrometry,* Elsevier, New York 1990.
54. P. R. Griffiths: *Chemical Fourier Transform Spectroscopy,* Wiley, New York 1975.
55. O. S. Wolfbeis, G. Boisdé, G. Gauglitz in W. Göpel (ed.): *Sensors and Biosensors,* VCH Verlagsgesellschaft, Weinheim 1992, p. 575.
56. J. Kiefer (ed.): *Ultraviolette Strahlen,* De Gruyter, Berlin 1977.
57. L. Endres, H. Frietz in W. Erb (ed.): *Leitfaden der Spektroradiometrie; Strahler,* **INSTAND vol. 6,** Springer Verlag, Berlin 1989, p. 23.
58. H. A. Strobel, W. R. Heinemann: *Chemical Instrumentation: a Systematic Approach,* Wiley & Sons, New York 1989.
59. J. F. Rabeck: *Experimental Methods in Photochemistry and Photophysics,* Wiley, Chichester 1982.
60. J. Wilson, J. F. B. Hawkes: *Lasers, Principles and Applications,* Prentice-Hall, Englewood Cliffs, NJ, 1987.
61. D. Engelage: *Lichtwellenleiter in Energie und Automatisierungsanlagen,* Hüthig-Verlag, Heidelberg 1986.
62. J. Pfab, *Anal. Proc. (London)* **28** (1991) 415.
63. K. Niemax, *Fresenius J. Anal. Chem.* **337** (1990) 551.
64. U. Panne, F. Lewitzka, R. Nießner, *Analusis* **20** (1992) 533.
65. F. P. Schäfer in F. P. Schäfer (ed.): *Dye Lasers, Topics in Applied Physics,* **vol. 1,** Springer Verlag, Berlin 1973, pp. 1 – 90.
66. F. K. Kneubühl, M. W. Sigrist: *Laser,* Teubner Studienbücher Physik, Stuttgart 1991, p. 331.
67. J. Eichler, H.-J. Eichler: *Laser,* Springer Verlag, Berlin 1999 0, p. 173.
68. H. A. Macleod: *Thin-film Optical Filters,* Adam Hilger Ltd., Bristol 1986.
69. A. G. Reule: "Errors in Spectrometry and Calibration, Procedures to Avoid Them," *J. Res. Nat. Bur. Stand. Sect. A* **80** (1976) 609.
70. Optimum Parameters for Spectrophotometry, Varian Instruments, OPT-720A.
71. L. Velluz, M. Legrand, M. Grosjean: *Optical Circular Dichroism,* Verlag Chemie, Weinheim 1965.
72. U. B. Seiffert, D. Janson: "Zur optischen Qualität von Photometerküvetten aus Plastik," *J. Clin. Chem. Clin. Biochem.* **19** (1981) 41.
73. G. Gauglitz: "Opto-Chemical and Opto-Immuno Sensors," in H. Baltes, W. Göpel, J. Hesse (eds.): *Sensors Update,* **vol. 1,** VCH Verlagsgesellschaft, Weinheim 1996.

74. W. Böhme, H. W. Müller, W. Liekmeier, *Int. Lab.* (1991) Jan/Feb.
75. W. Böhme, K. Horn, D. Meissner, *LaborPraxis* **11** (1987) 628.
76. E. L. Dereniak, D. G. Crowe: *Optical Radiation Detectors,* Wiley, New York 1984.
77. K. Möstl in W. Erb (ed.): *Leitfaden der Spektroradiometrie,* **INSTAND vol. 6,** Springer Verlag, Heidelberg 1989, p. 101.
78. F. Grum, C. J. Bartleson: *Optical Radiation Measurements,* **vol. 4**: Physical Detectors of Optical Radiation, Academic Press, New York 1983.
79. Photomultiplier Handbook, Burle Ind., Lancaster 1989.
80. C. Zeiss: Produktbeschreibung Simultanspektrometer, Oberkochen.
81. Firma Reticon, München, data sheets.
82. Y. Talmi: *Multichannel Image Detectors,* **vols. 1, 2,** Am. Chem. Soc., Washington, DC, 1979, 1982.
83. J. V. Sweedler, K. L. Ratzlaff, M. B. Denton (eds.): *Charge-Transfer Devices in Spectroscopy,* VCH Verlagsgesellschaft, Weinheim 1994.
84. E. L. Wehry (ed.): *Modern Fluorescence Spectroscopy,* Plenum, New York 1981.
85. E. Lippert, W. Nägele, I. Seibold-Blankenstein, U. Steigert, W. Voss, *Fresenius Z. Anal. Chem.* **170** (1959) 1.
86. Operation manual CD-spectrometer Jasco.
87. R. Hamilton, S. Hamilton: *Thin Layer Chromatography,* Wiley, New York 1987.
88. D. Oelkrug, A. Erbse, M. Plauschinat, *Z. Phys Chem.* **96** (1975) 283.
89. U. Hezel: "Direkte quantitative Photometrie an Dünnschichtchromatogrammen," *Angew. Chem.* **85** (1973) 334 ; *GIT Fachz. Lab.* **21** (1977) 694.
90. P. Kubelka, *J. Opt. Soc. Am.* **38** (1948) part I, 448.
91. S. Bayerbach, G. Gauglitz, *Fresenius J. Anal. Chem.* **335** (1989) 370.
92. Operation manual of MOSS ES4G ellipsometer, SOPRA Paris, France.
93. R. M. A. Azzam, N. M. Bashera: *Ellipsometry and Polarized Light,* North Holland, New Amsterdam 1977.
94. Ch. Striebel, A. Brecht, G. Gauglitz: "Characterization of Biomembranes by Spectral Ellipsometry, Surface Plasmon Resonance and Interferometry with regard to Biosensor Application," in *Biosen. Bioelectr.,* **9** (1994) 139.
95. H. Arwin: "Ellipsometry," in A. Baszkin, W. Norde (eds.): *Physical Chemistry of Biological Interfaces,* Marcel Dekker, New York 2000.
96. "Maintaining Optimum Spectrophotometer Performance," Perkin-Elmer, Application Data Bulletin; *Am. Soc. Clin. Pathol. Techn. Improv. Ser.* **27.**
97. Recommended Practices for General Techniques of Ultraviolet Quantitative Analysis, E169-63, ASTM in Manual of Recommended Practices in Spectroscopy, Philadelphia: American Society of Testing Materials Committee E13, 1969, 10 – 34.
98. C. Burgess, K. D. Mielenz (eds.): *Advances in Standards and Methodology in Spectrophotometry,* Elsevier, Amsterdam 1987.
99. A. Knowless, C. Burgess: *Practical Absorption Spectrometry,* Chapman and Hall, London 1984.
100. C. Burgess, A. Knowless: *Standards in Absorption Spectrometry,* Ultraviolet Spectrometry Group, vol. 1, Chapman and Hall, London 1981.
101. R. W. Burke, E. R. Deardorff, O. Menis: "Liquid Absorbance Standards," *J. Res. Nat. Bur. Stand. Sect. A* **76** (1972) 469.
102. R. E. Poulson: "Test Methods in Spectrophotometry: Stray Light Determination," *Appl. Opt.* **3** (1964) 99.
103. W. Luck, *Z. Elektrochem.* **64** (1960) 676.
104. Instruction manual, Hellma, Mühlheim/Baden.
105. S. Tesch, K. H. Rentrop, M. Otto, *Fresenius J. Anal. Chem.* **344** (1992) 206.
106. H. Bandemer, M. Otto, *Mikrochim. Acta* 1986, no. 2, 93.
107. K. Schuchardt, *Nachr. Chem. Tech. Lab.* **40** (1992) M1 – M20; H. J. Majer, *GIT Fachz. Lab.* **37** (1993) 881.
108. D. H. Williams, I. Flemming: *Strukturaufklärung in der organischen Chemie,* Thieme-Verlag, Stuttgart 1985.
109. G. Rücker, M. Neugebauer, G. G. Williams: *Instrumentelle pharmazeutische Analytik,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1988.
110. M. Hesse, H. Meier, B. Zeeh: *Spektroskopische Methoden in der organischen Chemie,* Thieme Verlag, Stuttgart 1985.
111. G. Gauglitz, K.-A. Kovar in E. Nürnberg, P. Surmann (eds.): *Hagers Handbuch der pharmazeutischen Praxis,* 5th ed., **vol. 2,** Methoden, Springer Verlag, Berlin 1991, p. 471.
112. I. Barret, W. F. Smyth, I. E. Davidson, *J. Pharm. Pharmacol.* **25** (1973) 387.

113. H. Martens, T. Naes: *Multivariate Calibration,* J. Wiley & Sons, Chichester 1993.
114. E. R. Malinowski, D. G. Howery: *Factor Analysis in Chemistry,* Wiley, New York 1980.
115. S. Ebel, E. Glaser, S. Abdulla, U. Steffens, V. Walter, *Fresenius Z. Anal. Chem.* **313** (1982) 24.
116. H. M. Heise, *Fresenius J. Anal. Chem.* **345** (1993) 604.
117. R. Marbach, H. M. Heise, *Chemometr. Intel. Lab. Syst.* **9** (1990) 45.
118. S. D. Brown, *Chemometr. Intel. Lab. Syst.* **10** (1991) 87.
119. M. Mettler, G. Gauglitz, *TrAc Trends Anal. Chem.* **11** (1992) 203.
120. J. A. Weismüller, A. Chanady, *TrAc Trends Anal. Chem. (Pers. Ed.)* **11** (1992) 86.
121. T. Kohonen: *Self-Organization and Associative Memory,* Springer Verlag, Berlin 1989.
122. C. Schierle, M. Otto, *Fresenius J. Anal. Chem.* **344** (1992) 190.
123. J. Göppert et al., Proc. Neuro Nimes 1992.
124. G. Kraus, G. Gauglitz, *Fresenius J. Anal. Chem.* **346** (1993) 572.
125. K. Molt, M. Egelkraut, *Fresenius J. Anal. Chem.* **327** (1987) 77.
126. W. Windig, *Chemometr. Intel. Lab. Syst.* **16** (1992) 1.
127. J. Eisenbrand: *Fluorimetrie,* Wissenschaftliche Verlagsgesellschaft, Stuttgart 1966.
128. G. Gauglitz, R. Goes, W. Stooss, R. Raue, *Z. Naturforsch. A Phys. Phys. Chem. Kosmophys.* **40 A** (1985) 317.
129. G. Elliott, A. Radley, *Analyst (London)* **86** (1961) 62.
130. S. Udenfried: *Fluorescence Assay in Biology and Medicine,* Academic Press, New York, vol. I, 1962, vol. II, 1969.
131. I. B. Berlman: *Handbook of Fluorescence Spectra of Aromatic Compounds,* Academic Press, New York 1965.
132. R. Strohecker, H. M. Hemming: *Vitaminbestimmungen,* Verlag Chemie, Weinheim 1963.
133. F. P. Schwarz, St. P. Wasik, *Anal. Chem.* **48** (1976) 524.
134. U. Panne, R. Nießner, *Vom Wasser* **79** (1992) 89.
135. O. S. Wolfbeis (ed.): *Fluorescence Spectroscopy, New Methods and Applications,* Springer Verlag, Berlin 1993.
136. U. Schobel, C. Barzen, G. Gauglitz: "Immunoanalytical techniques for pesticide monitoring based on fluorescence detection," *Fresenius J. Anal. Chem.* **366** (2000) 646 – 658.
137. G. Kortüm: *Reflexionsspektroskopie,* Springer Verlag, Berlin 1969.
138. H. Wilsing, *Leitz-Mitt. Wiss. Tech.* **1** (1960) 133.
139. A. Zlatkis, R. E. Kaiser, *J. Chromatogr. Lib.* **9** (1977) .
140. *Accessory Lambda-Series* Perkin-Elmer, Überlingen.
141. H. Arens: *Farbmetrik,* Akademie Verlag, Berlin 1957.
142. M. Richter: *Einführung in die Farbmetrik,* De Gruyter, Berlin 1981.
143. H. Naumer, W. Heller (eds.): *Untersuchungsmethoden in der Chemie,* Thieme-Verlag, Stuttgart 1990, p. 348.
144. R. Th. Kersten: *Einführung in die optische Nachrichtentechnik,* Springer Verlag, Berlin 1983.
145. V. Hoffmann et al., *J. Mol. Struct.* **293** (1993) 253.
146. G. Gauglitz, A. Brecht, G. Kraus, W. Nahm, *Sens. Actuators B* **11** (1993) 21.
147. A. Brecht, G. Gauglitz, W. Nahm, *Analusis* **20** (1992) 135.
148. G. Kraus, G. Gauglitz, *Fresenius J. Anal. Chem.* **344** (1992) 153.
149. G. Gauglitz, J. Krause-Bonte, H. Schlemmer, A. Matthes, *Anal. Chem.* **60** (1988) 2609.
150. D. E. Aspnes: "The Accurate Determination of Optical Properties by Ellipsometry," in: E. D. Palik (ed.): *Handbook of Optical Constants of Solids,* Academic Press, London 1986, pp. 89 ff.
151. D. E. Aspnes, A. A. Studna, *Phys. Rev. B Condens. Matter* **27** (1983) 985.
152. H. Arwin, D. E. Aspnes, *Thin Solid Films* **138** (1986) 195.
153. G. Kraus, A. Brecht, G. Gauglitz, Ch. Striebel, Lecture held at the Bunsentagung, Leipzig 1993.
154. H. Raether: Surface Plasmons on Smooth and Rough Surfaces and on Gratings, Springer Tracts in Modern Physics, vol. 111, Springer Verlag, Berlin 1988.
155. E. Kretchmann, H. Raether, *Z. Naturforsch. A Astrophys. Phys. Phys. Chem.* **23A** (1968) 2135.
156. A. Otto, *Z. Phys.* **216** (1968) 398.
157. B. Liedberg, C. Nylander, I. Lundström: "Surface Plasmon Resonance for Gas Detection and Biosensing," *Sens. Actuators* **4** (1983) 299.

158. P. B. Daniels, J. K. Deacon, M. J. Edowes, D. G. Pedley: "Surface Plasmon Resonance Applied to Immunosensing," *Sens. Actuators* **15** (1988) 11.
159. J. Homola, S. S. Yee, G. Gauglitz: "Surface plasmon resonance sensors: review," *Sensors & Actuators* **B 54** (1999) 3 – 15.
160. W. Lukosz, K. Tiefenthaler, *Opt. Lett.* **8** (1983) 537.
161. W. Lukosz, Th. Brenner, V. Briguet, Ph. Nellen, P. Zeller, *Proc. SPIE Int. Soc. Opt. Eng.* **1141** (1989) 192.
162. Ch. Stamm, W. Lukosz, *Sens. Actuators* **B11** (1993) 177.
163. D. Clerc, W. Lukosz, *Sens. Actuators* **B11** (1993) 177.
164. W. Böhme, P. Pospisil, K. Horn, H. Morr, *Spectrosc. Int.* **3** (1991) 36.
165. U. Ruzicka, E. H. Hansen: *Flow Injection Analysis,* John Wiley, New York 1981.
166. H. Müller, B. Frey, W. Böhme, *Fresenius J. Anal. Chem.* **341** (1991) 647.
167. W. Göpel et al. (eds.): *Sensors,* VCH Verlagsgesellschaft, Weinheim 1992.
168. O. S. Wolfbeis, *Fresenius J. Anal. Chem.* **325** (1986) 387.
169. D. W. Lübbers, N. Opitz, *Z. Naturforsch. C. Biosci.* **C 30** (1975) 532.
170. G. Gauglitz, M. Reichert, *Sens. Actuators* **B6** (1992) 83.
171. G. Gauglitz, J. Ingenhoff, *Ber. Bunsen-Ges. Phys. Chem.* **95** (1991) 1558.
172. M. A. Butler, R. Buss, A. Galuska, *J. Appl. Phys.* **70** (1991) 2326.
173. W. Nahm, G. Gauglitz, *GIT Fachz. Lab.* **34** (1990) 889.
174. G. Kraus, A. Brecht, V. Vasic, G. Gauglitz, *Fresenius J. Anal. Chem.* **348** (1994) 598.
175. H.-M. Haake, A. Schütz, G. Gauglitz, "Label-free detection of biomolecular interaction by optical sensors," *Fresenius J. Anal. Chem.* **366** (2000) 576 – 585.
176. C. Mischiati et al.: "Interaction of the human NF-kappaB p52 transcription factor with DNA-PNA hybrids mimicking the NF-kappaB binding sites of the human immunodeficiency virus type 1 promoter," *J. Biol. Chem.* **274** (1999) 33114 – 33122.
177. M. Sauer et al.: "Interaction of Chemically Modified Antisense Oligonucleotides with Sense DNA: A Label-Free Interaction Study with Reflectometric Interference Spectroscopy," *Analytical Chemistry* **71** (1999) 2850 – 2857.
178. W. Göpel, J. Hesse, J. N. Zemel (eds.): "Optical Sensors," in *Chemical and Biochemical Sensors,* VCH Verlagsgesellschaft, Weinheim 1992.
179. J. Polster, H. Lachmann: *Spectrometric Titrations,* VCH Verlagsgesellschaft, Weinheim 1989.
180. R. Blume, H. Lachmann, H. Mauser, F. Schneider, *Z. Naturforsch. B Anorg. Chem. Org. Chem.* **B 29** (1974) 500.
181. H. Lachmann, *Z. Anal. Chem.* **290** (1978) 117.
182. G. Gauglitz: "Physical Properties of Photochromic Substances," in H. Dürr, H. Bois-Laurant (eds.): *Photochromic Systems,* Elsevier, Amsterdam 1990.
183. G. Gauglitz in H. G. Seiler, A. Sigel, H. Sigel (eds.): *Handbook on Metals in Clinical Chemistry,* Marcel Dekker, New York, in press.
184. H. Mauser, G. Gauglitz: *Principles and Applications of Photokinetics,* Elsevier, Amsterdam, in preparation.
185. G. Gauglitz, E. Lüddecke, *Z. Anal. Chem.* **280** (1976) 105.
186. T. C. O'Haver, G. L. Green, *Anal. Chem.* **48** (1976) 312, *Am. Lab. (Fairfield, Conn.)* March (1975) 15.
187. G. Talsky, L. Mayring, H. Kreuzer, *Angew. Chem.* **90** (1978) 840.
188. G. L. Green, T. C. O'Haver, *Anal. Chem.* **46** (1974) 2191.
189. S. Shibata, M. Furukawa, T. Honkawa, *Anal. Chim. Acta* **81** (1976) 206.
190. *Ullmann,* 4th ed., **5,** 269.
191. S. Shibata, *Angew. Chem.* **88** (1976) 750.
192. A. Schmitt: *Labor-Praxis,* vol. 9, Vogel-Verlag, Würzburg 1979.
193. G. Stella, *Chem. Phys. Lett.* **39** (1976) 176.
194. H. P. Engelbert, R. Lawaczek, *Chem. Phys. Lipids* **38** (1985) 365.
195. J. C. Sternberg, *Clin. Chem.* **23** (1977) 1456.
196. G. Träxler, *Medizintechnik (Stuttgart)* **99** (1979) 79.
197. K. Oka et al., *Int. Lab.* March (1986) 24.
198. E. Lüddecke, D. Horn, *Chem. Ing. Tech.* **54** (1982) 266.
199. F. Dörr: "Polarized Light in Spectroscopy and Photochemistry," in A. A. Lamola (ed.): *Creation and Detection of the Excited State,* vol. 1, Part A, Dekker Inc., New York 1971, pp. 53 ff.
200. Th. Förster: *Fluoreszenz organischer Verbindungen,* Vandenhoeck G. Ruprecht, Göttingen 1982.

201. D. V. O'Connor, D. Phillips: *Time-Correlated Single Photon Counting,* Academic Press, New York 1984.
202. A. E. W. Knight, B. K. Selinger, *Aust. J. Chem.* **26** (1973) 1.
203. R. B. Somoano, *Angew. Chem.* **90** (1978) 250.
204. Yok-Han Pao (ed.): *Optoacoustic Spectroscopy and Detection,* Academic Press, New York 1977.
205. L. B. Kreuzer, *J. Appl. Phys.* **42** (1971) 2934.
206. L. G. Rosengreen, *Appl. Opt.* **14** (1975) 1960.
207. C. F. Dewey, R. D. Kamm, C. E. Hackett, *Appl. Phys. Lett.* **23** (1973) 633.
208. K. Kaya et al., *J. Am. Chem. Soc.* **97** (1975) 2153.
209. J. F. McClelland, R. N. Knisely, *Appl. Phys. Lett.* **28** (1976) 467; *Appl. Opt.* **15** (1976) 2658.
210. A. Rosencwaig: *Optoacoustic Spectroscopy and Detection,* Academic Press, New York 1977, p. 193.
211. J. C. Murphy, L. C. Aamodt, *Appl. Phys. Lett.* **31** (1977) 728.
212. A. Rosencwaig, A. Gersho, *J. Appl. Phys.* **47** (1976) 64.
213. W. R. Harshberger, M. B. Robin, *Acc. Chem. Res.* **6** (1973) 329.
214. M. G. Rockley, K. M. Wangle, *Chem. Phys. Lett.* **54** (1978) 597.
215. A. Rosencwaig, *Science (Washington, D.C.)* **181** (1973) 657.
216. A. Petzold, R. Niessner, SPIE 1716 (1992) 510.

Unsaturated Polyesters → Polyester Resins, Unsaturated

Infrared and Raman Spectroscopy

HANS-ULRICH GREMLICH, Novartis AG, Basel, Switzerland

1. Introduction 2
2. Techniques 3
2.1. Fourier Transform Infrared Technique 3
2.2. Dispersive and Fourier Transform Raman Techniques 5
3. Basic Principles of Vibrational Spectroscopy 7
3.1. The Vibrational Spectrum 7
3.2. Quantitative Analysis 8
3.3. The Symmetry of Molecules and Molecular Vibrations; Selection Rules 10
4. Interpretation of Infrared and Raman Spectra of Organic Compounds 11
4.1. The Concept of Group Frequencies 11
4.2. Methyl and Methylene Groups . . 12
4.3. Alkene Groups 15
4.4. Aromatic Rings 15
4.5. Triple Bonds and Cumulated Double Bonds 16
4.6. Ethers, Alcohols, and Phenols . . . 18
4.6.1. Ethers 18
4.6.2. Alcohols and Phenols 19
4.7. Amines, Azo, and Nitro Compounds 19
4.7.1. Amines 19
4.7.2. Azo Compounds 20
4.7.3. Nitro Compounds 20
4.8. Carbonyl Compounds 20
4.8.1. Ketones 20
4.8.2. Aldehydes 21
4.8.3. Carboxylic Acids 21
4.8.4. Carboxylate Salts 23
4.8.5. Anhydrides 23
4.8.6. Esters 23
4.8.7. Lactones 24
4.8.8. Carbonate Salts 24
4.8.9. Amides 24
4.8.10. Lactams 24
4.9. Sulfur-Containing Compounds . . 25
4.9.1. Thiols 25
4.9.2. Sulfides and Disulfides 25
4.9.3. Sulfones 26
4.10. Computer-Aided Spectral Interpretation 26
5. Applications of Vibrational Spectroscopy 27
5.1. Infrared Spectroscopy 27
5.1.1. Transmission Spectroscopy 27
5.1.2. External Reflection Spectroscopy . . 28
5.1.3. Internal Reflection Spectroscopy . . 30
5.1.4. Diffuse Reflection Spectroscopy . . 31
5.1.5. Emission Spectroscopy 33
5.1.6. Photoacoustic Spectroscopy 33
5.1.7. Chromatography/Fourier Transform Infrared Spectroscopy 34
5.1.8. Thermogravimetry/Fourier Transform Infrared Spectroscopy 36
5.1.9. Vibrational Circular Dichroism . . . 36
5.2. Raman Spectroscopy 37
5.3. Fourier Transform Infrared and Raman Microspectroscopy 37
5.4. Time-Resolved FT-IR and FT-Raman Spectroscopy 39
5.5. Vibrational Spectroscopic Imaging 39
5.6. Infrared and Raman Spectroscopy of Polymers 40
6. Near-Infrared Spectroscopy 40
6.1. Comparison of Mid-Infrared and Near-Infrared Spectroscopy 40
6.2. Applications of Near-Infrared Spectroscopy 41
7. References 42

1. Introduction

The roots of infrared spectroscopy go back to the year 1800, when WILLIAM HERSCHEL discovered the infrared region of the electromagnetic spectrum. Since 1905, when WILLIAM W. COBLENTZ ran the first infrared spectrum [1], vibrational spectroscopy has become an important analytical tool in research and in technical fields. In the late 1960s, infrared spectrometry was generally believed to be a dying instrumental technique that was being superseded by nuclear magnetic resonance and mass spectrometry for structural determination, and by gas and liquid chromatography for quantitative analysis.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_429

However, the appearance of the first research-grade Fourier transform infrared (FT-IR) spectrometers in the early 1970s initiated a renaissance of infrared spectrometry. After analytical instruments (since the late 1970s) and routine instruments (since the mid 1980s), dedicated instruments are now available at reasonable prices. With its fundamental multiplex or Fellgett's advantage and throughput or Jacquinot's advantage, FT-IR offers a versatility of approach to measurement problems often superior to other techniques. Furthermore, FT-IR is capable of extracting from samples information that is difficult to obtain or even inaccessible for nuclear magnetic resonance and mass spectrometry. Applications of modern FT-IR spectrometry include simple, routine identity and purity examinations (quality control) as well as quantitative analysis, process measurements, the identification of unknown compounds, and the investigation of biological materials.

Raman and infrared spectra give images of molecular vibrations which complement each other; i.e., the combined evaluation of both spectra yields more information about molecular structure than when they are evaluated separately. The 1990s witnessed a tremendous growth in the capabilities and utilization of Raman scattering measurements. The parallel growth of Fourier transform Raman (FT-Raman) with charge-coupled device (CCD) based dispersive instrumentation has provided the spectroscopist with a wide range of choice in instrumentation. There are clear strengths and weaknesses that accompany both types of instrumentation. While FT-Raman systems virtually eliminate fluorescence by using near-IR laser sources, this advantage is paid for with limited sensitivity. On the other hand, fluorescence interferences always play a role in dispersive measurements with higher sensitivity.

As an alternative to infrared spectroscopy, Raman spectroscopy can be easier to use in some cases; for example, whereas water and glass are strong infrared absorbers they are weak Raman scatterers, so that it is easy to produce a good-quality Raman spectrum of an aqueous sample in a glass container.

General References. Details of vibrational theory are given in the books by Herzberg [2], [3]; Wilson, Decius, and Cross [4]; Steele [5]; Koningstein [6]; and Long [7]. The technique and applications of FT-IR spectroscopy are discussed in detail by Griffiths and de Haseth [8], and by Mackenzie [9]. An introduction to infrared and Raman spectroscopy, dealing with theoretical as well as experimental and interpretational aspects, is provided by Colthup, Daly, and Wiberley [10] and by Schrader [11]. Books about Raman spectroscopy in particular are those of Kohlrausch [12]; Baranska, Labudzinska, and Terpinski [13]; and Hendra, Jones, and Warnes [14]. Practical FT-IR and Raman spectroscopy is covered by Ferraro and Krishnan [15]; Pelletier [16]; and Grasselli and Bulkin [17]. For the interpretation of IR and Raman spectra of organic compounds, see Socrates [18]; Bellamy [19], [20]; Dollish, Fateley, and Bentley [21], and Lin-Vien, Colthup, Fateley, and Grasselli [22]. Vibrational spectroscopy of inorganic and coordination compounds is described by Nakamoto [23]. Biological applications of IR and Raman spectroscopy are covered by Gremlich and Yan [24]. Various collections of vibrational spectra are available. Both Raman and IR spectra of about 1000 organic compounds are offered as an atlas by Schrader [25]. The most comprehensive reference-spectra libraries, digital and hardcopy, can be obtained from Sadtler [26]. A book on modern techniques in applied optical spectroscopy was edited by Mirabella [27].

2. Techniques

2.1. Fourier Transform Infrared Technique

The basis of Fourier transform infrared (FT-IR) spectroscopy is the two-beam interferometer, designed by Michelson in 1891 [28], and shown schematically in Figure 1A. Broadband infrared radiation is emitted by a thermal source (globar, metal strips, Nernst glower) and falls onto a beam splitter which, in the ideal case, transmits half the radiation and reflects the other half. The reflected half, after traversing a distance L, falls onto a fixed mirror M1. The radiation is reflected by M1 and, after traversing back along distance L, falls onto the beam

splitter again. The transmitted radiation follows a similar path and also traverses distance L; however, the mirror M2 of the interferometer can be moved very precisely along the optical axis by an additional distance x. Hence, the total path length of the transmitted radiation is $2(L+x)$. On recombination at the beam splitter the two beams possess an optical path difference of $\Delta = 2x$. Since they are spatially coherent, the two beams interfere on recombination. The beam, modulated by movement of the mirror, leaves the interferometer, passes through the sample cell and is finally focussed on the detector. The signal registered by the detector, the interferogram, is thus the radiation intensity $I(x)$, measured as a function of the displacement x of the moving mirror M2 from the distance L (Fig. 1B). The mathematical transformation, a Fourier transform [8], of the interferogram, performed by computer, initially provides a so-called single-beam spectrum. This is compared with a reference spectrum measured without the sample to obtain a spectrum analogous to that measured by conventional dispersive methods. This spectrum is stored digitally in the computer, from which it can be retrieved for further use (Fig. 1C). In addition to Michelson interferometers, which often have air bearings for the moving mirror and thus require a source of dry compressed air, mechanical interferometer concepts such as frictionless electromagnetic drive mechanisms or refractively scanned interferometers have been developed by instrument manufacturers.

Two major kinds of detector are used in mid-infrared FT-IR spectrometers, the deuterated triglycine sulfate (DTGS) pyroelectric detector and the mercury cadmium telluride (MCT) photodetector. The DTGS type, which operates at room temperature and has a wide frequency range, is the most popular detector for FT-IR instruments. The MCT detector responds more quickly and is more sensitive than the DTGS detector, but it operates at liquid-nitrogen temperature and has both a limited frequency range and a limited dynamic range; therefore, it is used only for special applications [8].

The resolution of an interferometer depends mainly on two factors: firstly, the distance x the moving mirror travels or the maximum optical path difference $\Delta = 2x$ (Fig. 1), and secondly, the apodization function used in computing the spectrum.

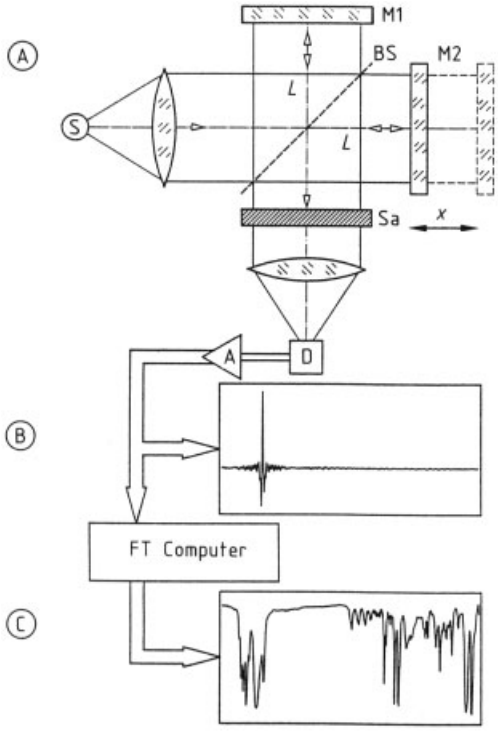

Figure 1. A) Schematic diagram of a Michelson interferometer; B) Signal registered by the detector D, the interferogram; C) Spectrum obtained by Fourier transform (FT) of the interferogram
S = Radiation source; Sa = Sample cell; D = Detector; A = Amplifier; M1 = Fixed mirror; M2 = Movable mirror; BS = Beam splitter; x = Mirror displacement

The first point is derived from the Rayleigh criterion of resolution, which states that to resolve two spectral lines separated by a distance d, the interferogram must be measured to an optical path length of at least $1/d$ [29].

The second factor is related to the fact that in real life the interferogram is truncated at finite optical path difference. In addition, in the fast Fourier transform (FFT) algorithm, according to Cooley and Tukey [30], which is used to perform the Fourier transform faster than the classical method, certain assumptions and simplifications are made. The result is that the FFT of a monochromatic source is not an infinitely narrow line, corresponding to infinite optical path difference and hence to infinite resolution (see above), but a definable shape known as instru-

mental line shape (ILS) function [8]. The ILS of a perfectly aligned Michelson interferometer has the shape of a $(\sin x)/x$ or sinc x function (Fig. 2), exhibiting strong side lobes or feet, which may even have negative intensities. To attenuate these spurious feet, so-called apodization functions are used as weighting functions for the interferogram. Numerous such functions exist (e.g., triangular, Happ – Genzel, or Blackman – Harris) which produce instrumental line shapes with lower intensity side lobes, but also with broader main lobes than the sinc function. Hence, side-lobe suppression is only possible at the cost of resolution, because the full width at half-height of the ILS defines the best resolution achievable with a given apodization function [31]. In contrast to dispersive spectroscopy, where the triangular apodization function is determined by the slit of the grating spectrometer, in FT-IR spectroscopy, a choice between highest resolution (no apodization) and best side lobe suppression (Blackman – Harris apodization function) is possible.

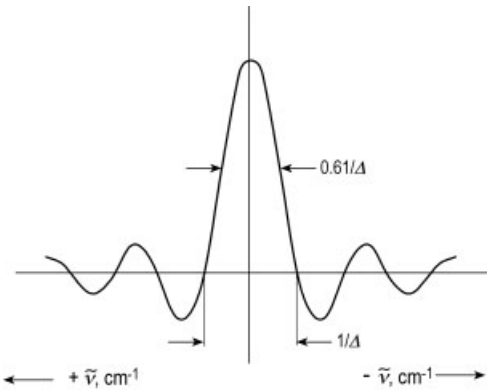

Figure 2. The sinc x function: instrumental line shape function of a perfectly aligned Michelson interferometer with no apodization
$\tilde{\nu}$ = wavenumber, cm^{-1}; Δ = maximum retardation, cm

In comparison with conventional spectroscopy, the FT-IR method possesses significant advantages. In grating spectrometers, the spectrum is measured directly by registering the radiation intensity as a function of the continuously changing wavelength given by a monochromator [10]. Depending on the spectral resolution chosen, only a very small part (in practice less than 0.1 %) of the radiation present in the monochromator reaches the IR detector. In the FT-IR spectrometer, all frequencies emitted by the IR source reach the detector simultaneously, which results in considerable time saving and a large signal-to-noise ratio advantage over dispersive instruments; this is the most important advantage, known as the multiplex or Fellgett's advantage [8]. A further advantage stems from the fact that the circular apertures employed in FT-IR spectrometers have larger surface areas than the linear slits of grating spectrometers and hence permit radiation throughputs at least six times greater; this is known as the throughput or Jacquinot's advantage [8]. In FT-IR spectroscopy, the measuring time is the time required to move the mirror M2 the distance necessary to give the desired resolution; since the mirror can be moved very rapidly, complete spectra can be obtained within fractions of a second. By contrast, the measuring time for a conventional spectrum is on the order of minutes. In an FT spectrum, the accuracy of each wavenumber is coupled to the accuracy with which the position of the moving mirror is determined; by using an auxiliary HeNe laser interferometer the position of the mirror can be determined to an accuracy better than 0.005 µm. This means that the wavenumbers of an FT-IR spectrum can be determined with high accuracy ($< 0.01\ cm^{-1}$). In other words, FT-IR spectrometers have a built-in wavenumber calibration of high precision; this advantage is known as accuracy or Connes' advantage [29]. As a result, it is possible to carry out highly precise spectral subtractions and thus very reliably detect slight spectral differences between samples.

2.2. Dispersive and Fourier Transform Raman Techniques

In modern dispersive Raman spectrometers, arrays of detector elements such as photodiode arrays or charge-coupled devices (CCDs) are arranged in a polychromator. Here, each element records a different spectral band at the same time, thus making use of the multichannel advantage. In dispersive instruments, the very strong radiation at and near the exciting line, the Rayleigh radiation, may produce stray radiation in the entire spectrum with an intensity much higher than that of the Raman lines. In order to remove this unwanted radiation, line filters

in the form of so-called notch filters or holographic filters are used which specifically reflect the Rayleigh line. In conventional, dispersive Raman spectroscopy visible lasers are usually used to excite the Raman effect (see Section 3.1), while in FT-Raman spectroscopy near-IR lasers are employed. Excitation with visible lasers often produces strongly interfering fluorescence originating from the sample or impurities; however, this is eliminated by using a neodymium yttrium aluminum garnet (Nd : YAG) laser providing a discrete line excitation at 1.064 μm, as at this wavelength in the near-IR region no electronic transitions are induced.

An FT-Raman spectrometer is shown in schematic form in Figure 3A. The beam of the laser source passes to the sample compartment where it is focused, typically to a spot of 100 μm diameter, onto the sample, either in 90° or 180° geometry (Fig. 3B). The scattered Raman light is collected by an aspherical lens and passed through a filter module to remove the Rayleigh line and Rayleigh wings [32]. It is then directed through the infrared input port into the interferometer which is optimized for the near-IR region as is the liquid-nitrogen-cooled germanium or room-temperature indium gallium arsenide (InGaAs) detector [33]. The procedure to obtain the FT-Raman spectrum from the interferogram, the signal registered by the detector, is essentially the same as that described above for an FT-IR spectrum.

Before FT-Raman spectrometers could be successfully used several difficulties had to be overcome. As the intensity of Raman scattering is proportional to the fourth power of the frequency of excitation a loss in sensitivity by a factor between 20 and 90 must be compensated for when the exciting radiation is shifted from the visible to the near-IR region [32]. Moreover, the noise equivalent power (NEP) of near-IR detectors is greater by about one to two orders of magnitude than that of a photomultiplier tube, the usual detector for conventional Raman spectroscopy. However, both disadvantages are compensated for by the throughput (Jacquinot) and multiplex (Fellgett) advantages of Fourier transform spectroscopy (see Section 2.1). Another major problem is the fact that Raman scattering is always accompanied by very strong radiation, the Rayleigh line and Rayleigh wings, occurring at and near the frequency of excitation with six to ten orders of magnitude higher intensity than the Raman signal. The shot noise of the Rayleigh line and Rayleigh wings produces a background noise which would completely mask all Raman lines. Therefore, an effective filter that removes the Rayleigh line and Rayleigh wings in the range 750 – 100 cm^{-1} is one of the most important components of a powerful FT-Raman instrument [32]. In addition, optimization of the sampling technique is indispensable in order to convert the flux of exciting radiation into a maximum flux of Raman radiation to be received by the optical system of the spectrometer [34]. The 180° backscattering arrangement (Fig. 3B), rather than the 90° arrangement used in visible, dispersive Raman technique, is undoubtedly the most efficient way of collecting the Raman signal and provides the best quality data in the near-IR FT-Raman experiment.

3. Basic Principles of Vibrational Spectroscopy

3.1. The Vibrational Spectrum

To explain the origins of a vibrational spectrum the vibration of a diatomic molecule is considered first. This can be illustrated by a molecular model in which the atomic nuclei are represented by two point masses m_1 and m_2, and the interatomic bond by a massless spring. According to Hooke's law, the vibrational frequency ν (in s^{-1}) determined by classical methods is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{f}{\mu}} \tag{1}$$

where f is the force constant of the spring in N/m, and μ is the reduced mass in kg:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \tag{2}$$

Thus, the vibrational frequency is higher when the force constant is higher, i.e., when the bonding of the two atoms is stronger. Conversely, the heavier the vibrating masses are, the smaller is the frequency ν. The frequency of fundamental vibrations is on the order of magnitude of 10^{13} s^{-1}. Quantum-mechanical methods give

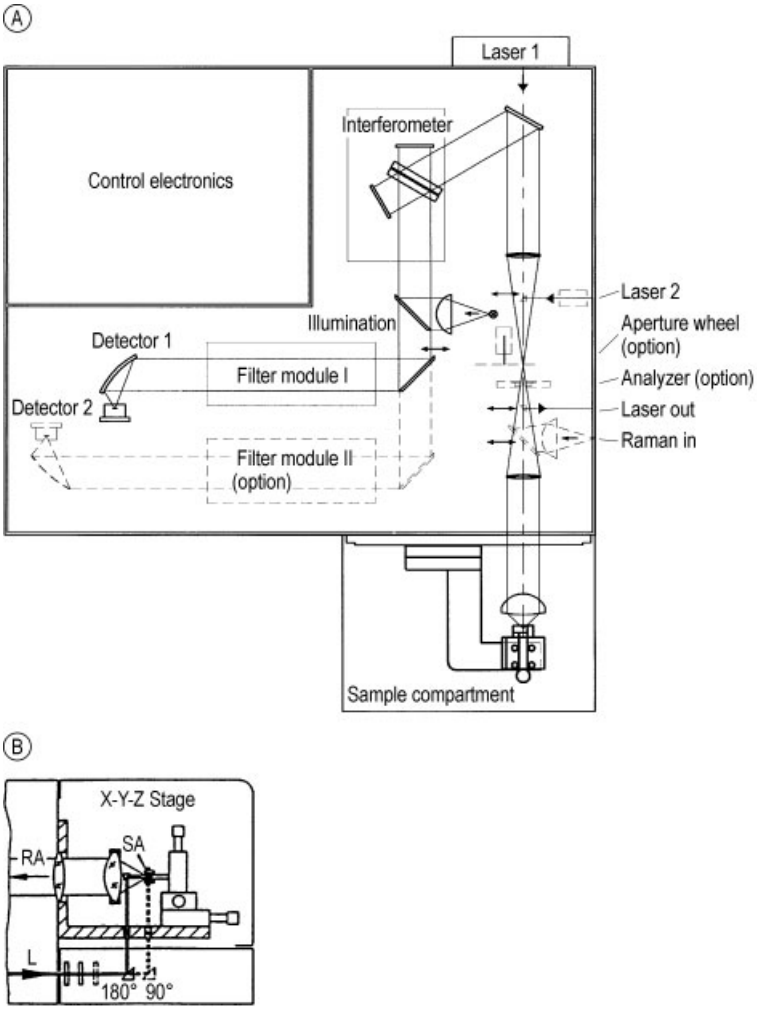

Figure 3. A) Schematic diagram of an FT-Raman spectrometer (Bruker RFS 100); B) Schematic diagram of the sample compartment
L = laser; RA = Raman; SA = Sample
(Reproduced by permission of Bruker Optik GmbH, D-76275 Ettlingen, Germany)

the following expression for the vibrational energy of a Hooke's oscillator in a state characterized by the vibrational quantum number v:

$$E_{\mathrm{vib}}=\frac{h}{2\pi}\sqrt{\frac{f}{\mu}}\left(v+\frac{1}{2}\right)=h\nu_0\left(v+\frac{1}{2}\right) \tag{3}$$

where f and μ have the same definition as in Equation (1) and ν_0 is the vibrational frequency of the ground state. This relation is valid with good accuracy for vibrational transitions from the ground state ($v=0$) to the first excited state ($v=1$). In free molecules (gas state), vibrational transitions are always accompanied by rotational transitions. The corresponding energies are given by:

$$E_{\mathrm{vib,rot}}=h\nu_0\left(v+\frac{1}{2}\right)+BhJ\,(J+1) \tag{4}$$

where J is the rotational quantum number and B is the rotational constant:

$$B=\frac{h}{8\pi^2 I} \tag{5}$$

where I is the moment of inertia. In the real world, molecular oscillations are inharmonic.

Therefore, the transition energy decreases with increasing vibrational quantum number until the molecule finally dissociates. The quantum mechanical energy of a diatomic inharmonic oscillator is given in good approximation by:

$$E_{\mathrm{vib}}=h\nu_0\left[\left(v+\frac{1}{2}\right)-x_0\left(v+\frac{1}{2}\right)^2\right] \tag{6}$$

where x_0 is the inharmonicity constant.

A complex molecule is considered as a system of coupled inharmonic oscillators. If there are N atomic nuclei in the molecule, there will be a total of $3N$ degrees of freedom of motion for all the nuclear masses in the molecule. Subtracting the pure translations and rotations of the entire molecule leaves $3N-6$ internal degrees of freedom for a nonlinear molecule and $3N-5$ internal degrees of freedom for a linear molecule. The internal degrees of freedom correspond to the number of independent normal modes of vibration; their forms and frequencies must be calculated mathematically [10]. In each normal mode of vibration all the atoms of the molecule vibrate with the same frequency and pass through their equilibrium positions simultaneously. The true internal motions of the molecule, which constitute its vibrational spectrum, are composed of the normal vibrations as a coupled system of these independent inharmonic oscillators. Thus, a molecule is unambiguously characterized by its vibrational spectrum. Real-world molecules are best described by the model of inharmonic oscillators because here transitions are allowed which are forbidden to the harmonic oscillator: transitions from the ground state ($v=0$) to states with $v=2, 3, \ldots$ (overtones), transitions from an excited level of a vibration (difference bands), and transitions between states belonging to different types of normal mode (combination bands).

In vibrational spectroscopy, instead of the frequency ν (in s^{-1}) the so-called wavenumber $\tilde{\nu}$ expressed in cm^{-1} (reciprocal centimeters) is mostly used; this signifies the number of waves in a 1 cm wavetrain:

$$\tilde{\nu}=\frac{\nu}{(c/n)}=\frac{1}{\lambda} \tag{7}$$

where c is the velocity of light in a vacuum (2.997925×10^{10} cm/s), c/n is the velocity of light in a medium with refractive index n in which the wavenumber is measured (the refractive index of air is 1.0003), and λ is the wavelength in centimeters. In the infrared region of the electromagnetic spectrum the practical unit of wavelength is 10^{-4} cm or 10^{-6} m (i.e., µm), and wavenumber and wavelength are related as follows:

$$\tilde{\nu}/\mathrm{cm}^{-1}=\frac{10^4}{\lambda/\mu\mathrm{m}} \tag{8}$$

In infrared spectroscopy, both wavenumber and wavelength are used, whereas in Raman spectroscopy only wavenumber is employed.

The infrared and microwave regions of the electromagnetic spectrum are shown in Table 1. In the near-infrared region, overtones and combination bands are observed. Since these transitions are only possible because of inharmonicity, their probability is diminished. A good rule of thumb is that the intensity of the first overtone is about one-tenth that of the corresponding fundamental vibration. The fundamentals occur in the mid-infrared range, which is the most important infrared region. In the far-infrared region, the fundamentals of heavy, single-bonded atoms and the absorptions of inorganic coordination compounds [23] are found.

Infrared spectroscopy is based on the interaction of an oscillating electromagnetic field with a molecule, and it is only possible if the dipole moment of the molecule changes as a result of a molecular vibration. While the absorption frequency depends on the molecular vibrational frequency, the absorption intensity depends on how effectively the infrared photon energy is transferred to the molecule. It can be shown [3] that the intensity of infrared absorption is proportional to the square of the change in the dipole moment μ, with respect to the change in the normal coordinate q describing the corresponding molecular vibration:

$$I_{\mathrm{IR}}\sim\left(\frac{\partial\mu}{\partial q}\right)^2 \tag{9}$$

A second way to induce molecular vibrations is to irradiate a sample with an intense source of monochromatic radiation, usually in the visible or near-infrared region, this leads to the Raman effect, which can be regarded as an inelastic collision between the incident photon and the molecule. As a result of the collision, the vibrational or rotational energy of the molecule

Table 1. Infrared and microwave regions of the electromagnetic spectrum

Region	Wavelength λ, µm	Wavenumber $\tilde{\nu}$, cm^{-1}	Frequency ν, s^{-1}
Infrared			
near	7.8×10^{-1} to 2.5	12 800 to 4000	3.8×10^{14} to 1.2×10^{14}
mid	2.5 to 5.0×10^{1}	4000 to 200	1.2×10^{14} to 6.0×10^{12}
far	5.0×10^{1} to 1.0×10^{3}	200 to 10	6.0×10^{12} to 3.0×10^{11}
Microwave	1.0×10^{3} to 1.0×10^{6}	10 to 0.01	1.0×10^{12} to 3.0×10^{8}

is changed by an amount ΔE_{m}. For energy to be conserved, the energy of the scattered photon $h\nu_{\mathrm{s}}$ must differ from that of the incident photon $h\nu_{\mathrm{i}}$ by an amount equal to ΔE_{m}:

$$h\nu_{\mathrm{i}}-h\nu_{\mathrm{s}}=\Delta E_{\mathrm{m}}=h\nu_{\mathrm{m}} \tag{10}$$

If the molecule gains energy, then ΔE_{m} is positive and ν_{s} is smaller than ν_{i}, giving rise to so-called Stokes lines in the Raman spectrum. If the molecule loses energy, ΔE_{m} is negative and ν_{s} is larger than ν_{i}, giving rise to so-called anti-Stokes lines in the Raman spectrum. The dipole moment μ induced by the Raman effect can be related to the electric field E of the incident electromagnetic radiation as follows:

$$\mu=\alpha E \tag{11}$$

where α is the polarizability of the molecule. It can be derived from classical Raman theory [3] that for a molecular vibration to be Raman active it must be accompanied by a change in the polarizability of the molecule. The intensity of Raman absorption is proportional to the square of the change in polarizability α, with respect to the corresponding normal coordinate q:

$$I_{\mathrm{RA}}\sim\left(\frac{\partial\alpha}{\partial q}\right)^{2} \tag{12}$$

Although it follows from classical theory that Stokes and anti-Stokes lines should have the same intensity, according to quantum theory and in agreement with experiment (Fig. 4) anti-Stokes lines are a much less intense since the number of molecules in the initial state $v=1$ of anti-Stokes lines is only $e^{-(h\nu_{\mathrm{m}}/kT)}$ times the number of molecules in the initial state $v=0$ of the Stokes lines [3]. The Raman shift in cm^{-1}, i.e., the difference $\Delta\tilde{\nu}$ between the wavenumber $\tilde{\nu}_{\mathrm{e}}$ of the exciting laser and the wavenumber $\tilde{\nu}_{\mathrm{s}}$ of the scattered Raman light, is indicated on the abscissa of each Raman spectrum.

3.2. Quantitative Analysis

Quantitative analysis is well established not only in ultraviolet/visible and in near-infrared spectroscopy, but it is also very important in mid-infrared measurements. The general prerequisite for spectrometric quantitative analysis is defined as follows [35]: information derived from the spectrum of a sample is related in mathematical terms to changes in the level(s) of an individual component, or several components within the sample or series of samples, i.e., the spectral response of an analyte can be related by a mathematical function to changes in concentration of the analyte. In the ideal situation, the measured spectroscopic feature varies linearly with concentration. In reality, however, true linearity is not always obtained, but this is not important provided the measured function is reproducible. Most practical analyses are not absolute measurements, and normally measurements are made on a given instrument within a fixed working environment; under such set circumstances, reproducibility and consistency of the measurement are the most important factors.

If I_0 is the intensity of monochromatic radiation entering a sample and I is the intensity transmitted by the sample, then the ratio I/I_0 is the transmittance T of the sample. The percent transmittance ($\%T$) is $100\,T$. If the sample cell has thickness b and the absorbing component has concentration c, then the fundamental relation governing the absorption of radiation as a function of transmittance is:

$$T=\frac{I}{I_0}=10^{-abc} \tag{13}$$

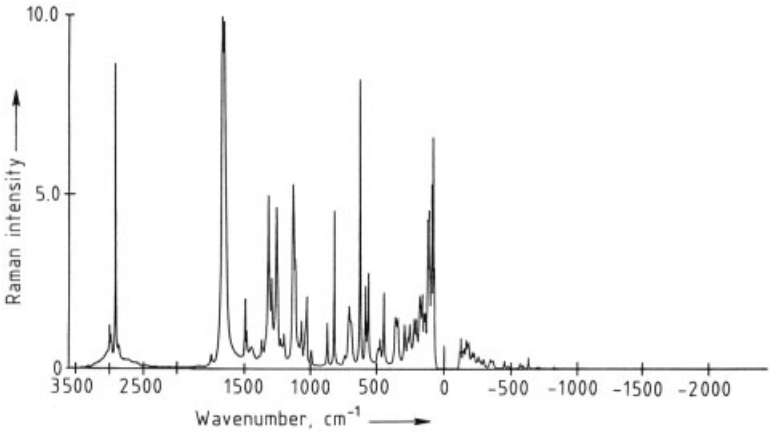

Figure 4. FT-Raman spectrum of ascorbic acid recorded on a Bruker FT-Raman accessory FRA 106 directly interfaced to a Bruker FT-IR spectrometer IFS 66
The accessory was equipped with a germanium diode cooled by liquid nitrogen. Resolution was $2\,cm^{-1}$ and laser output power was 300 mW. Stokes shift: 3600 to $50\,cm^{-1}$, anti-Stokes shift: -100 to $-2000\,cm^{-1}$.

The constant a is called the absorptivity and is characteristic of a specific sample at a specific wavelength. As the transmittance T does not vary linearly with the concentration of an absorbing species the equation above is usually transformed by taking the logarithms of both sides of the equation and replacing I/I_0 with I_0/I to eliminate the negative sign:

$$\log \frac{I_0}{I} = abc \tag{14}$$

The term $\log(I_0/I)$ is the absorbance A, which changes linearly with changes in concentration of an absorbing species:

$$A = abc \tag{15}$$

This fundamental equation for spectrometric quantitative analysis is known as Beer–Lambert–Bouguer law, sometimes shortened to Beer's law or Beer–Lambert law.

In practice, several steps are involved in the development of a quantitative method [35]. The first and probably most crucial for the ultimate success of the analytical method is to understand the system, i.e., to obtain reference spectra of the analyte and all other components. In the second step, the best method of sampling should be determined. Then, with standards having been prepared, the system must be calibrated. The last step before analyzing samples is to prepare validation samples and to evaluate the method.

With modern sampling techniques, good quantitative infrared analysis with virtually every type of sample is practicable; however, liquids are ideal for this purpose, being measured in a liquid cell of fixed thickness, either as 100 % sample or diluted with solvent. In this connection it should be taken into account that there are no ideal solvents for infrared spectroscopy [35]. In addition, because absorption bands and path length can be influenced by the temperature of the transmission cell, it is advisable to control the temperature.

Polymers are usually analyzed as pressed films, and solids are often difficult to examine quantitatively by infrared methods; in these cases, absorption band ratios [10] give the best results. In fact, the main application of multicomponent quantitative infrared analysis is gas analysis. If the difficulty in handling the gases is overcome, multicomponent analysis can be readily performed.

A simple chemical system can consist of a pure single component, or of a single component or more than one component in a mixture with no spectral interference; it is assumed that the radiation absorption by one component is not affected by the presence of other components. In simple systems, absorbance peak height measurements, directly or by using a selected baseline [35], are often employed for calibration and analysis. Because of intrinsic instrumental errors the practical limit for usable absorbance values is about

three. Peak height measurements are also sensitive to changes in instrumental resolution and can vary considerably from instrument to instrument. To circumvent these problems, an alternative method is the use of integrated absorbance or peak area [10].

Complex chemical systems are composed of one or more components in a mixture with a significant degree of spectral interference, or of several components with a large amount of mutual physical and/or chemical interaction. In these cases, quantitative analysis is best performed by statistical methods such as principal component regression (PCR) or partial least squares (PLS) [36]; these are offered in the software packages of instrument manufacturers and software suppliers. Artificial neural networks (ANNs) should be primarily used when a data set is nonlinear [37].

3.3. The Symmetry of Molecules and Molecular Vibrations; Selection Rules

All molecules show symmetry properties and they all possess at least one (trivial) symmetry element, the identity. The symmetry of a molecule is important in spectroscopy because changes in symmetry during molecular vibration determine whether a vibrational dipole moment μ occurs or not. As mentioned, a vibration is infrared active if μ changes, if not, it is infrared inactive; analogously, this is true of the polarizability α and Raman activity.

In infrared and Raman spectroscopy, the symmetry of molecules is usually discussed in terms of five symmetry elements and their corresponding five symmetry operations [38]. If a wide variety of molecules is investigated, it will be found, as can be proved by mathematical group theory [39], that only certain combinations of symmetry elements are possible. Such restricted combinations of symmetry elements, known as point groups, are used for the classification of molecules. Each point group is associated with a set of normal vibrations. These in turn are classified according to the symmetry of vibration. From this it is possible to predict whether a normal vibration is infrared or Raman active or neither. These are the so-called selection rules. Many vibrations of molecules with low symmetry are both infrared and Raman active, and it is chiefly in the band intensities that the two types of spectra differ, sometimes markedly so. On the other hand, when a center of symmetry is present, i.e., for molecules with a high degree of symmetry, bands that are infrared active are Raman inactive and vice versa. By using laser excitation (i.e., linearly polarized light) (see Section 2.2), Raman spectroscopy also provides a means of recognizing totally symmetrical vibrations [3]. The intensity of a Raman line depends on the direction of the exciting electric field in relation to the orientation of the analyzer (Fig. 5). The latter can be either perpendicular or parallel to the direction of the electric field. The depolarization ratio ϱ is the ratio between the intensities of scattered light measured at each of these two positions:

$$\varrho = \frac{I_{\perp}}{I_{\parallel}} \tag{16}$$

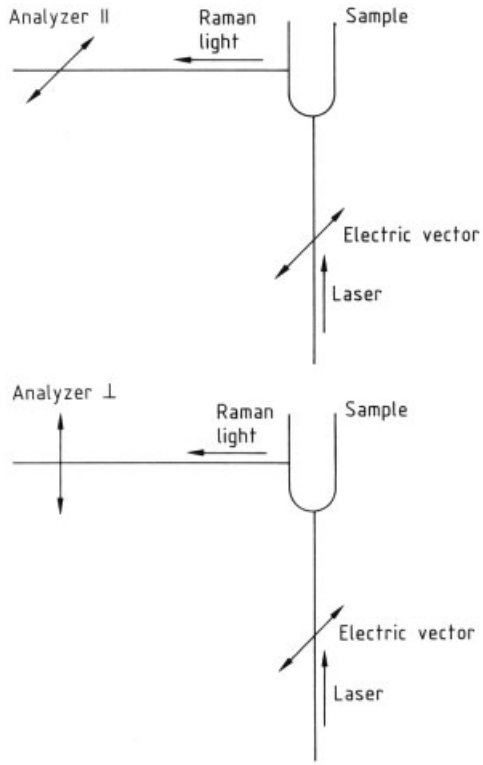

Figure 5. Orientation of the analyzer in relation to the direction of the electrical field of the exciting laser

The depolarization ratio of a Raman line depends on the symmetry of the molecular vibra-

tion involved (i.e., the change in molecular symmetry induced by the corresponding vibration); the maximum value of depolarization observed with linearly polarized light is $\varrho_1^{max} = 3/4$. If a Raman line shows this extent of depolarization, it is said to be depolarized, whereas, if the degree of depolarization is smaller, the line is polarized. It can be shown [3] that only Raman lines corresponding to totally symmetric vibrations can have a degree of depolarization smaller than the maximum value, that is, can be polarized.

4. Interpretation of Infrared and Raman Spectra of Organic Compounds

4.1. The Concept of Group Frequencies

A complex molecule can be considered as a system of coupled inharmonic oscillators. Empirically it is found that vibrational coupling is restricted to certain submolecular groups of atoms. This coupling is relatively constant from molecule to molecule, so that the submolecular groups produce bands in a characteristic frequency region of the vibrational spectrum. These bands – the characteristic group frequencies – are readily predictable and so form the empirical basis for the interpretation of vibrational spectra. Their position, intensity, and width are decisive for the correlation of an absorption band with a certain submolecular group. For example, the vibrational spectra of *n*-heptane, *n*-octane, and *n*-nonane have a number of bands in common; these are the group frequencies for normal alkanes. This concept of group frequencies corresponds to the concept of chemical shift in nuclear magnetic resonance spectroscopy. These spectra also have a number of bands which are not in common; these so-called fingerprint bands are characteristic of the individual chemical compound and are used to distinguish one compound from another. In Table 2, commonly used symbols and descriptions of different vibrational forms are given, and Table 3 lists usual abbreviations for the classification of vibrational absorption bands. For practical applications, spectral interpretation is mainly based on personal experience.

Table 2. Commonly used symbols and descriptions or different vibrational forms

Symbol	Designation	Example	Illustration
s	symmetric	ν_s	
as	antisymmetric (asymmetric)	ν_{as}	
ip	in-plane (ip)	δ	
oop	out-of-plane (oop)	γ	
ν	stretching		
δ	ip deformation		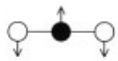
γ	oop deformation		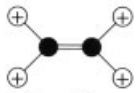
ω	wagging		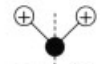
τ	twisting		
ϱ	rocking		

Table 3. Usual abbreviations for the classification of vibrational absorption bands

Abbreviation	Signification
vs	very strong
s	strong
m	medium
w	weak
vw	very weak
sh	shoulder
b	broad
sr	sharp
v	variable

In the following sections, dealing with spectrum – structure correlations, characteristic group frequencies having diagnostic value are given in the form of charts. These correlation charts show the positions of the characteristic group frequencies which are the same in infrared and Raman spectra. The two types of spectra mainly differ in the band intensities, which, however, are not indicated in the correlation charts. To provide this information, examples of infrared and Raman spectra of most of the functional groups discussed are shown. Fourier transform Raman spectra (Fig. 4) are usually

recorded between 3600 and 50 cm^{-1} (Stokes shift) and − 100 to − 2000 cm^{-1} (anti-Stokes shift). For comparison with mid-infrared spectra, in the following sections the Fourier transform Raman spectra are presented in the range between 4000 and 400 cm^{-1}. The relative intensities of vibrational absorption bands, which are evident in the model spectra presented, are important for appropriate spectral interpretation. Specific differences between infrared and Raman spectra are given special mention.

In the figures in this chapter (Figs. 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44), the upper curve is the infrared spectrum, with intensity increasing from the bottom to the top of the diagram. The lower curve is the Raman spectrum with an ordinate linear in relative intensity units increasing from the bottom to the top of the diagram. The Raman spectra have not been corrected for fluctuations in the sensitivity of the spectrometer. The infrared spectra were recorded with a Bruker IFS 66 FT-IR spectrometer; for recording of the Raman spectra a Bruker FRA 106 FT-Raman accessory was used.

4.2. Methyl and Methylene Groups

Figure 6 shows characteristic group frequencies, and Figure 7 illustrates typical vibrations of alkanes. In Figure 8, infrared and Raman spectra of 2-methylbutane (isopentane) are given as an example of alkane spectra. The most intense bands are those of the antisymmetric and symmetric CH_3 or CH_2 stretching vibrations between 2960 and 2850 cm^{-1}. Typical of isopropyl and tertiary butyl groups is that the symmetric CH_3 deformation band, usually observed at 1375 cm^{-1}, is split into two bands near 1390 and 1370 cm^{-1}. In the Raman spectra, the symmetric CH_3 deformation band near 1375 cm^{-1} is generally relatively weak in alkanes, but when the CH_3 group is next to a double or triple bond or an aromatic ring the Raman intensity is noticeably enhanced [10]. This is demonstrated in Figure 9, which shows the spectra of 2-methyl-2-butene.

The CH_2 wagging bands are spread over a region between 1350 and 1180 cm^{-1}, occurring as a characteristic progression of weak bands. They are best seen in the solid-phase spectra of long straight-chain compounds such as fatty acids [10].

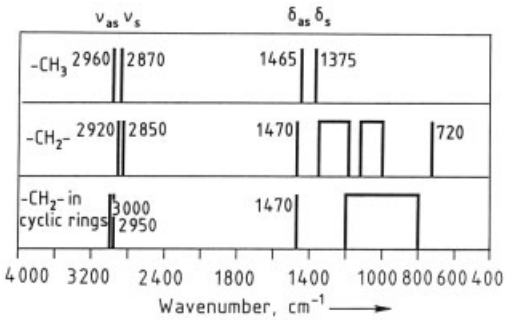

Figure 6. Characteristic group frequencies of alkanes

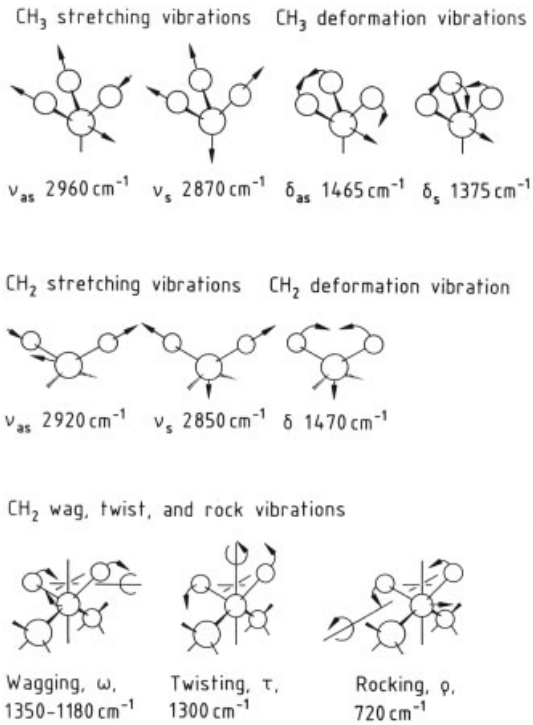

Figure 7. Typical vibrations of alkanes

The CH_2 twisting vibrations in CH_2 chains have frequencies dispersed over the same region as the wagging bands, as can be seen in the spectra of polyethylene [25], for example. The infrared intensity of these bands is weak, whereas in the Raman spectrum at about 1300 cm^{-1} the in-phase CH_2 twist vibration is a useful group frequency.

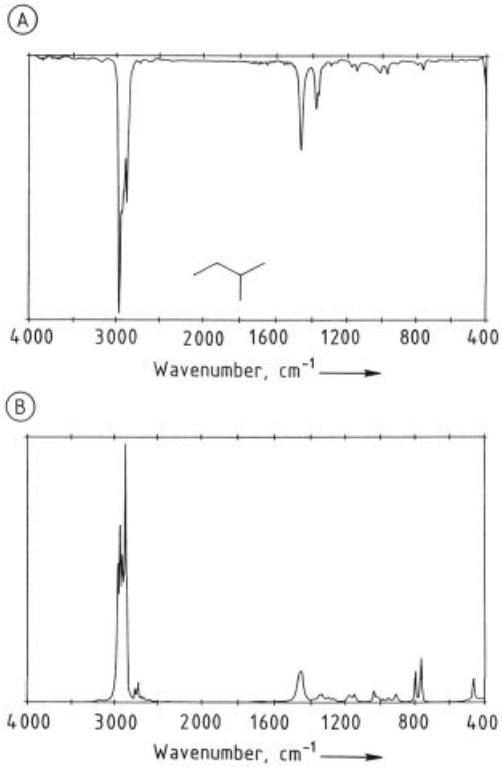

Figure 8. FT-infrared (A) and FT-Raman (B) spectra of 2-methylbutane

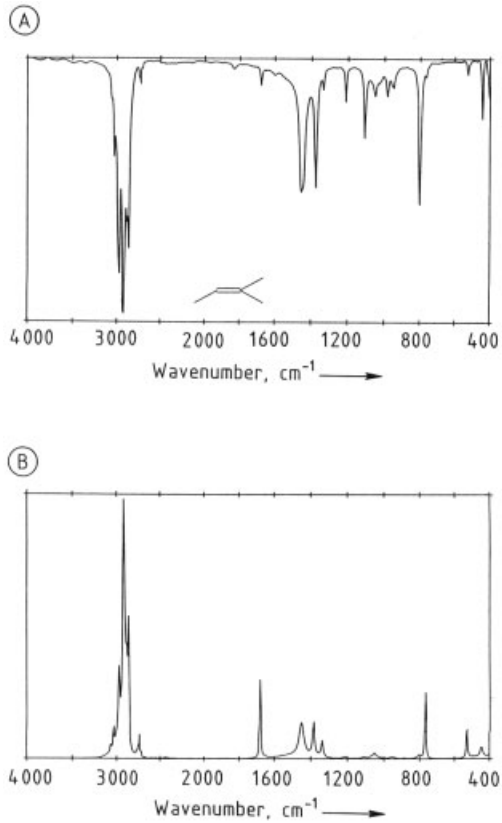

Figure 9. FT-infrared (A) and FT-Raman (B) spectra of 2-methyl-2-butene

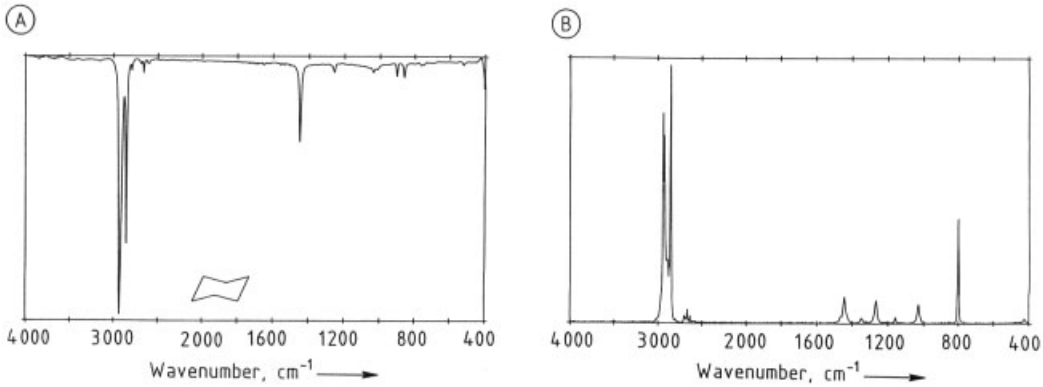

Figure 10. FT-infrared (A) and FT-Raman (B) spectra of cyclohexane

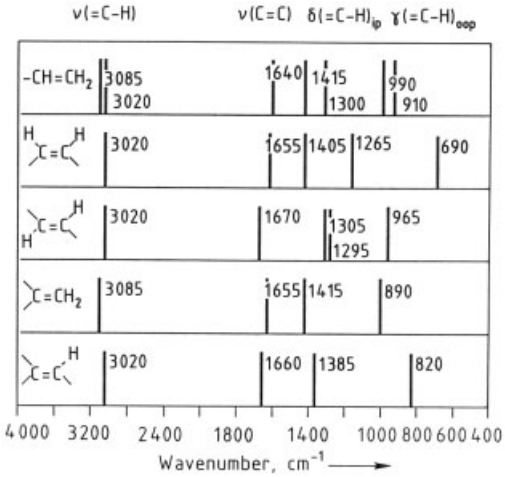

Figure 11. Characteristic group frequencies of alkenes

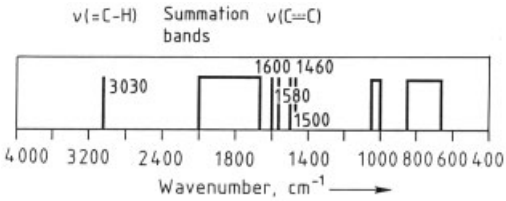

Figure 13. Characteristic group frequencies of aromatic rings

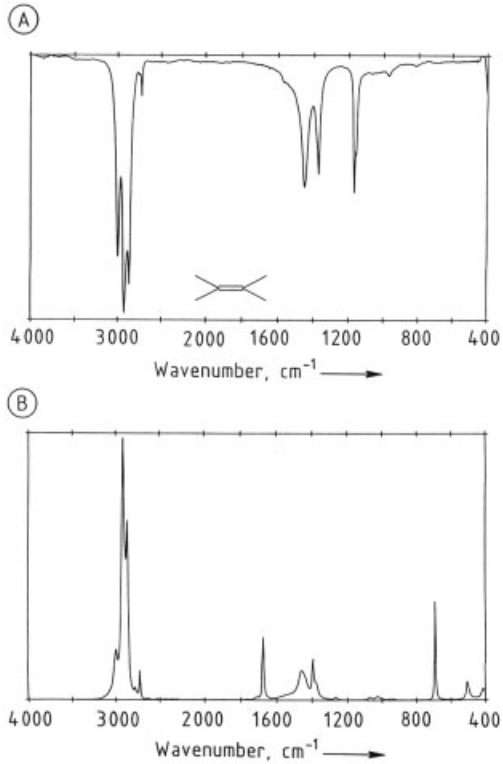

Figure 12. FT-infrared (A) and FT-Raman (B) spectra of 2,3-dimethyl-2-butene

The C−C stretching vibration of CH_2 chains is seen only in the Raman spectrum, at 1131 and 1061 cm^{-1} [40], while the CH_2 rocking vibration at 720 cm^{-1}, which is characteristic for longer $(CH_2)_n$ chains with $n \geq 4$, is observed only in the infrared spectrum (e.g., polyethylene) [25].

Aliphatic ring compounds are best characterized by their Raman spectra, since in the infrared only very weak characteristic ring frequencies due to C−C ring stretching vibrations are observed near 1000 cm^{-1} (Fig. 10). Raman spectra, in contrast, show prominent lines due to ring stretching vibrations: cyclopropane 1188 cm^{-1} [25], cyclobutane 1001 cm^{-1} [40], cyclopentane 886 cm^{-1} [25], and cyclohexane 802 cm^{-1} (Fig. 10).

4.3. Alkene Groups

Characteristic group frequencies of alkenes are shown in Figure 11. In alkenes, the =CH stretching vibration generally occurs above 3000 cm^{-1}. In symmetrical *trans* or symmetrical tetrasubstituted double-bond compounds, the C=C stretching frequency near 1640 cm^{-1}, usually a medium intensity band, is infrared inactive because, in this case, no change of dipole moment occurs as a result of the vibration. However, the C=C stretching vibration gives rise to a strong Raman signal in the region between 1680 and 1630 cm^{-1} in all types of alkenes (Fig. 12).

In vinyl and vinylidine groups, the $=CH_2$ in-plane deformation gives rise to a medium intensity band in both the infrared and Raman spectrum near 1415 cm^{-1}. In the case of the vinyl group, an additional infrared and Raman deformation band is observed near 1300 cm^{-1} [10]. In *cis*-1,2-dialkyl ethylenes, the in-plane deformation band appears near 1405 cm^{-1} in the infrared and at ca. 1265 cm^{-1} in the Raman spectrum. In the corresponding *trans* compounds, the in-plane bending vibrations occur at 1305 cm^{-1} in the Raman and at 1295 cm^{-1} in the infrared spectrum. In these *trans* compounds, the =CH

out-of-plane bending vibration gives rise to a strong infrared band, of high diagnostic value, occurring between 980 and 965 cm^{-1}.

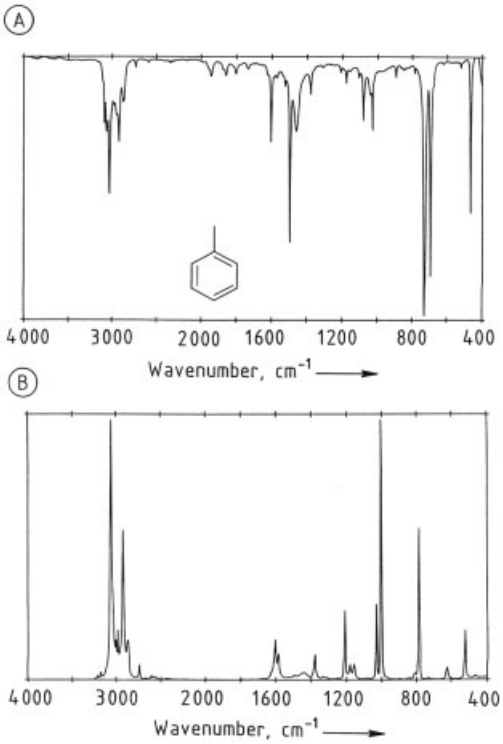

Figure 14. FT-infrared (A) and FT-Raman (B) spectra of toluene

4.4. Aromatic Rings

Figure 13 shows characteristic group frequencies of aromatic rings. The bands between 1600 and 1460 cm^{-1} are ring modes involving C–C partial double bonds of the aromatic ring. The band at 1500 cm^{-1} is usually the strongest of these bands, which generally show medium to strong infrared but weak Raman intensities. The use of both infrared and Raman spectra gives reliable information about the type of substitution of an aromatic ring: in infrared spectra, intense bands in the 850 – 675 cm^{-1} region, due to out-of-plane CH wagging and out-of-plane sextant ring bending vibrations, are indicative of the type of substitution [10]. In the Raman spectra, there are very strong lines at ca. 1000 cm^{-1} involving ring stretching and ring bending vibrations. As shown in Figures 14, 15, 16, 17, a very strong signal at 1000 cm^{-1} occurs for mono and *meta* substitution, a strong line is observed between 1055 and 1015 cm^{-1} for *ortho* substitution, while no signal is observed around 1000 cm^{-1} for *para* substitution. In infrared spectra, the out-of-plane CH wagging vibrations give rise to relatively prominent summation bands in the 2000 – 1650 cm^{-1} region. As the summation band patterns are approximately constant for a particular ring substitution, they can also be used to determine the type of substitution [10].

4.5. Triple Bonds and Cumulated Double Bonds

Triple bonds (X≡Y) and cumulated double bonds (X=Y=Z) absorb roughly in the same region, namely, 2300 – 2000 cm^{-1} (Table 4). The bands are very diagnostic because in this region of the spectrum no other strong bands occur.

Table 4. Characteristic group frequencies of triple bonds and cumulated double bonds

Bond	ν (≡C−H), cm^{-1}	ν (X≡Y) or ν_{as}(X=Y=Z), cm^{-1}
−C≡C−H	3300	2260 – 2100
−C≡N		2260 – 2200
−N≡C		2165 – 2110
−C≡N→O		2300 – 2290
$-N^{+}\equiv N$		2300 – 2140
−S−C≡N		2180 – 2140
−N=C=S		2150 – 2000
−N=C=O		2275 – 2230
$-N=N^{+}=N^{-}$		2250 – 2100
−N=C=N−		2150 – 2100
−C=C=O		2155 – 2130
$-C=N^{+}=N^{-}$		2100 – 2010
$-CO-C=N^{+}=N^{-}$		2100 – 2050

Monosubstituted acetylenes have a relatively narrow ≡CH stretching band near 3300 cm^{-1} which is strong in the infrared and weaker in the Raman. In monoalkyl acetylenes, the C≡C stretching vibration shows a weak infrared absorption near 2120 cm^{-1}, while no infrared C≡C absorption is observed in symmetrically substituted acetylenes because of the center of symmetry. In Raman spectra, this vibration always occurs as a strong line. Nitriles are characterized by a strong C≡N stretching frequency at

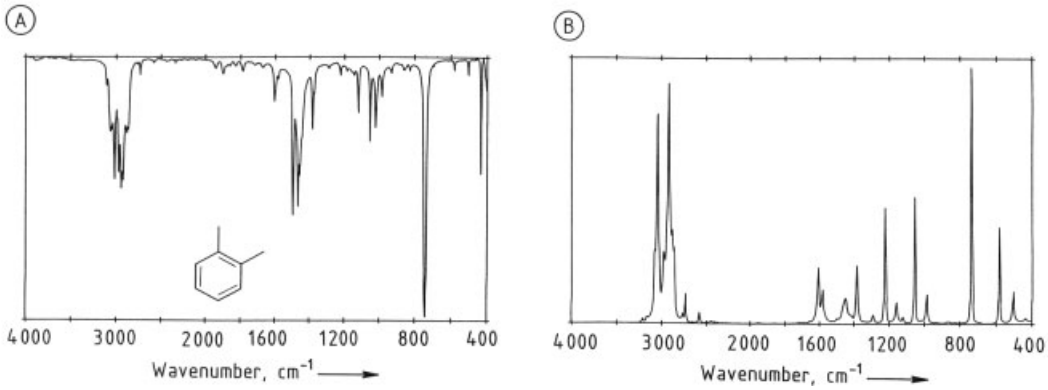

Figure 15. FT-infrared (A) and FT-Raman (B) spectra of *o*-xylene

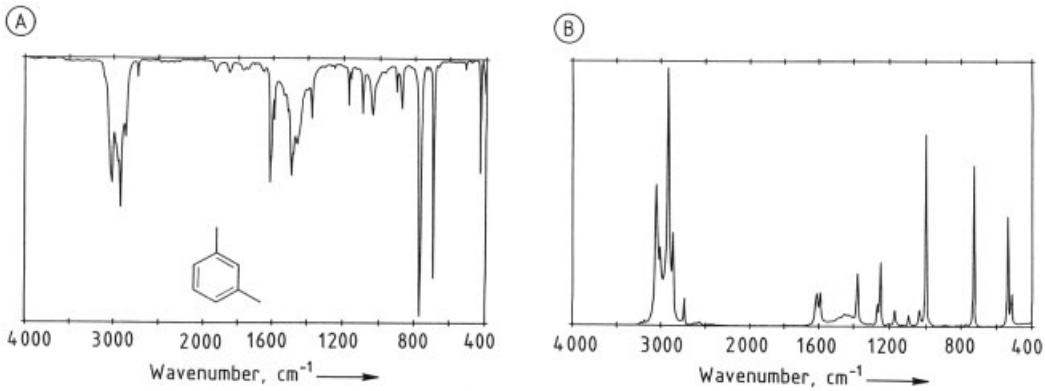

Figure 16. FT-infrared (A) and FT-Raman (B) spectra of *m*-xylene

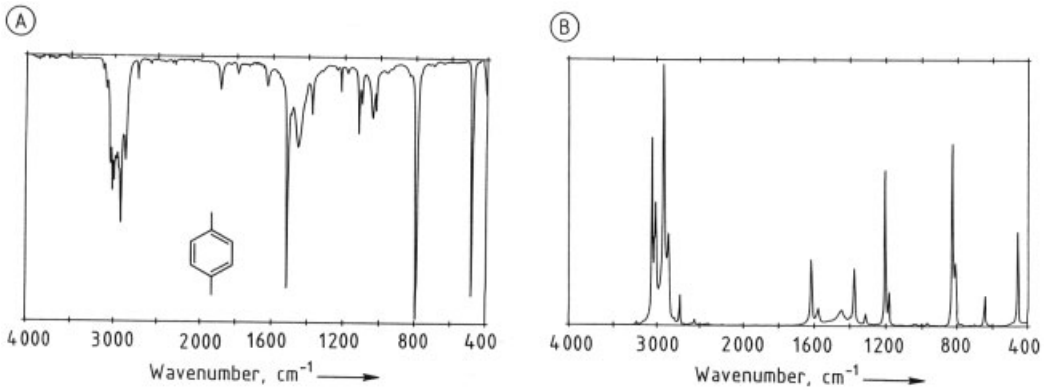

Figure 17. FT-infrared (A) and FT-Raman (B) spectra of *p*-xylene

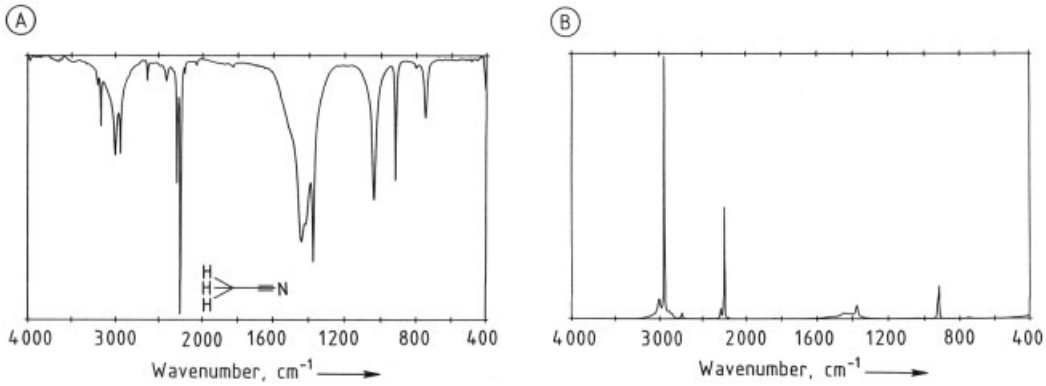

Figure 18. FT-infrared (A) and FT-Raman (B) spectra of acetonitrile

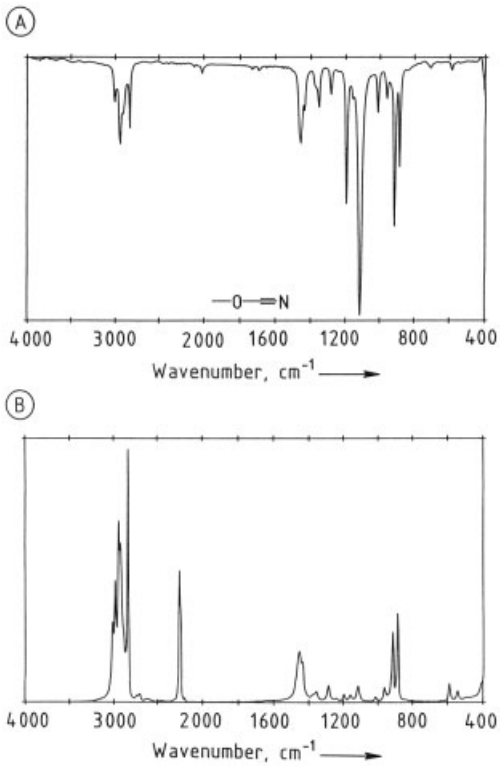

Figure 19. FT-infrared (A) and FT-Raman (B) spectra of methoxy-acetonitrile

$2260-2200\,cm^{-1}$ in the infrared and in the Raman (Fig. 18). Upon electronegative substitution at the α-carbon atom, the infrared intensity of the C≡N stretching vibration is considerably weakened so that only the Raman spectrum, where a strong signal is always observed, can be used for identification (Fig. 19).

4.6. Ethers, Alcohols, and Phenols

4.6.1. Ethers

Characteristic group frequencies of ethers are shown in Figure 20. The key band for noncyclic ethers is the strong antisymmetric C−O−C stretching vibration between 1140 and $1085\,cm^{-1}$ for aliphatic, between 1275 and $1200\,cm^{-1}$ for aromatic, and between 1225 and $1200\,cm^{-1}$ for vinyl ethers. With the exception of aliphatic–aromatic ethers, the symmetric C−O−C stretching band at ca. $1050\,cm^{-1}$ is too weak to have diagnostic value. In cyclic ethers, both the antisymmetric ($1180-800\,cm^{-1}$) and the symmetric ($1270-800\,cm^{-1}$) C−O−C stretching bands are characteristic. The antisymmetric C−O−C stretch is usually a strong infrared band and a weak Raman signal, the symmetric C−O−C stretching vibration generally being a strong Raman line and a weaker infrared band (Fig. 21).

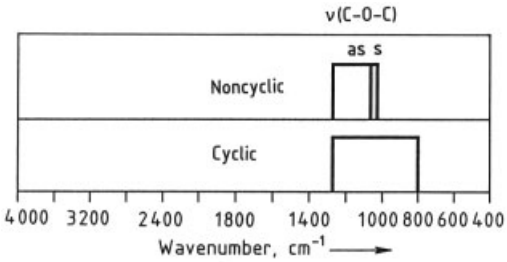

Figure 20. Characteristic group frequencies of ethers

4.6.2. Alcohols and Phenols

Characteristic group frequencies of alcohols and phenols are shown in Figure 22. Because of the strongly polarized OH group, alcohols generally show a weak Raman effect (and therefore are excellent solvents for Raman spectroscopy). In the liquid or solid state, alcohols and phenols exhibit strong infrared bands due to O−H stretching vibrations. Due to hydrogen bonding of the OH groups, these bands are very broad in the pure liquid or solid, as well as in concentrated solutions or mixtures. The position of the similarly intense C−O stretching band permits distinction between primary, secondary, and tertiary alcohols, as well as phenols.

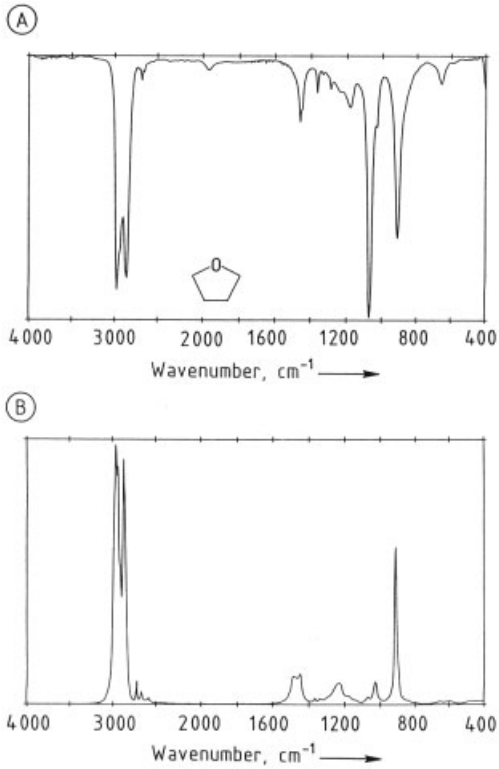

Figure 21. FT-infrared (A) and FT-Raman (B) spectra of tetrahydrofuran

4.7. Amines, Azo, and Nitro Compounds

4.7.1. Amines

Characteristic group frequencies of amines are shown in Figure 23. In the infrared and Raman, key bands are the NH stretching bands, although they are often not very intense. Primary amines exhibit two bands due to antisymmetric (3550 – 3330 cm^{-1}) and symmetric (3450 – 3250 cm^{-1}) stretching, while for secondary amines only one NH stretching band is observed. The intensities of NH and NH_2 stretching bands are generally weaker in aliphatic than in aromatic amines. The C−N stretching vibration is not necessarily diagnostic, because in aliphatic amines it gives rise to absorption bands of only weak to medium intensity between 1240 and 1000 cm^{-1}. In aromatic amines, strong C−N stretching bands are observed in the range 1380 – 1250 cm^{-1}.

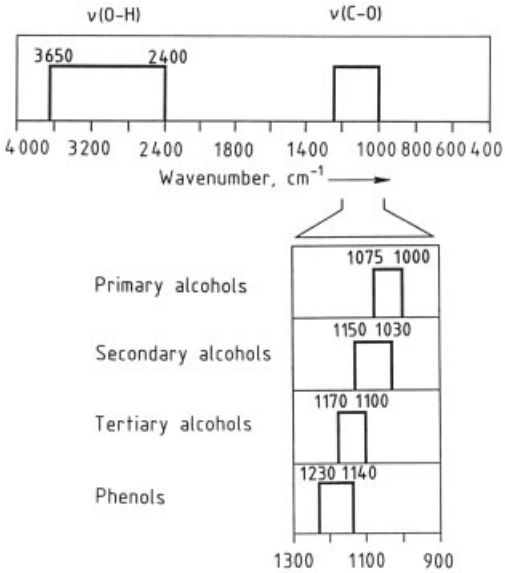

Figure 22. Characteristic group frequencies of alcohols and phenols

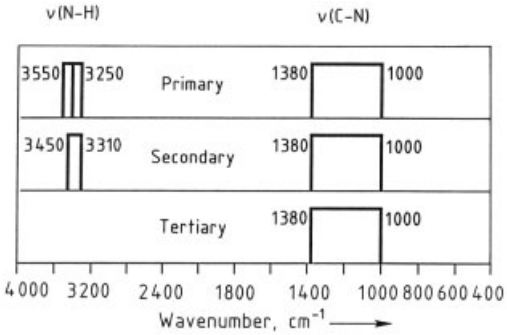

Figure 23. Characteristic group frequencies of amines

4.7.2. Azo Compounds

Characteristic group frequencies of azo compounds are shown in Figure 24. In the infrared spectrum, similarly to symmetrical *trans*-substituted ethylenes, the N=N stretching vibration of the *trans* symmetrically substituted azo group does not occur, whereas all azo compounds show a medium (aliphatic compounds) or strong (aromatic compounds) N=N stretching Raman line. This band occurs between 1580 and 1520 cm^{-1} for purely aliphatic or aliphatic – aromatic, and between 1460 and 1380 cm^{-1} for purely aromatic azo compounds. In azoaryls, an additional strong Raman line involving aryl – N stretch is observed near 1140 cm^{-1}. As an example, the spectra of azobenzene are shown in Figure 25.

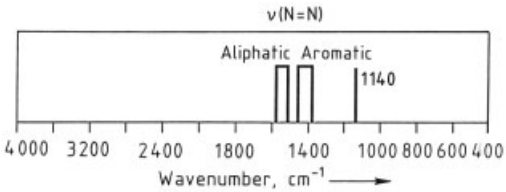

Figure 24. Characteristic group frequencies of azo compounds

4.7.3. Nitro Compounds

Characteristic group frequencies of nitro compounds are shown in Figure 26. The nitro group, with its two identical N–O partial double bonds, gives rise to two bands due to antisymmetric and symmetric stretching vibrations. In the infrared, the antisymmetric vibration causes a strong absorption at 1590 – 1545 cm^{-1} in aliphatic, and at 1545 – 1500 cm^{-1} in aromatic nitro compounds. In comparison, the infrared intensities of the symmetric stretching vibration between 1390 and 1355 cm^{-1} for aliphatic, and 1370 and 1330 cm^{-1} for aromatic nitro compounds are somewhat weaker. However, a very strong Raman line due to symmetric vibration is observed in this region (see Fig. 27).

4.8. Carbonyl Compounds

Because of the stretching of the C=O bond, carbonyl compounds generally give rise to a very strong infrared and a medium to strong Raman line between 1900 and 1550 cm^{-1}. The position of the carbonyl band is influenced by inductive, mesomeric, mass, and bond-angle effects [10].

4.8.1. Ketones

Characteristic group frequencies of ketones are shown in Figure 28. Ketones are characterized by a strong C=O band near 1715 cm^{-1}, often accompanied by an overtone at ca. 3450 cm^{-1}. Ketones in strained rings show considerably higher C=O frequencies, whereas in di-*tert*-butyl ketone, for example, the carbonyl frequency is observed at 1687 cm^{-1}; a steric increase in the C–C–C angle causes this lower value.

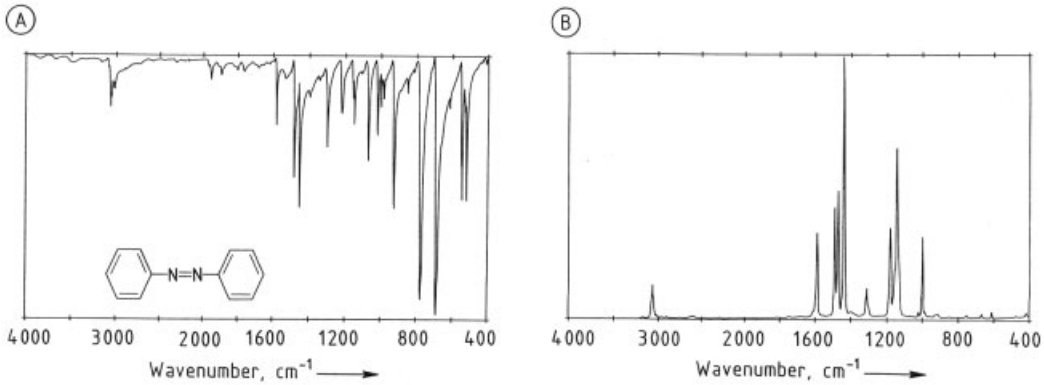

Figure 25. FT-infrared (A) and FT-Raman (B) spectra of azobenzene

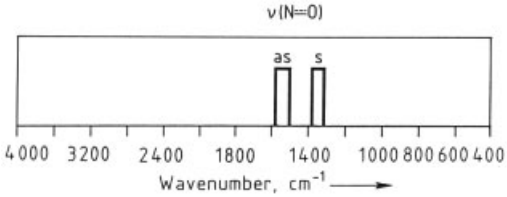

Figure 26. Characteristic group frequencies of nitro compounds

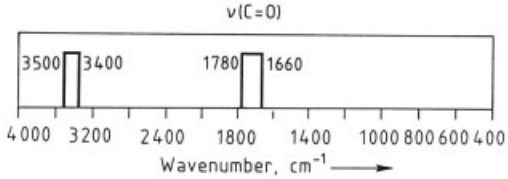

Figure 28. Characteristic group frequencies of ketones

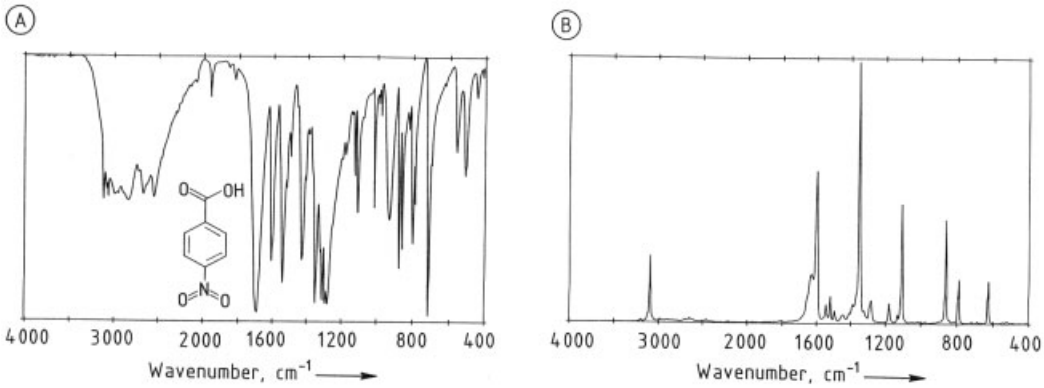

Figure 27. FT-infrared (A) and FT-Raman (B) spectra of 4-nitrobenzoic acid

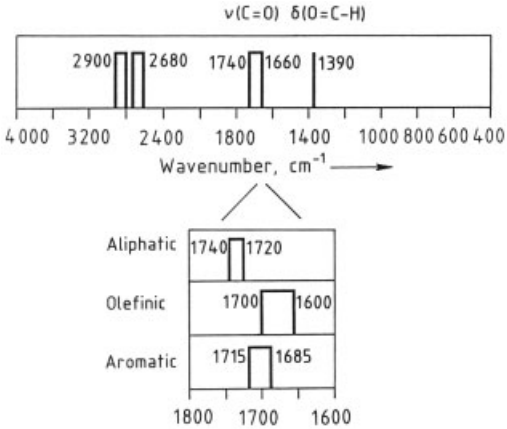

Figure 29. Characteristic group frequencies of aldehydes

4.8.2. Aldehydes

Characteristic group frequencies of aldehydes are shown in Figure 29. Together with the carbonyl band, the most intense band in the infrared spectrum, aldehydes are characterized by two CH bands at 2900 – 2800 cm^{-1} and 2775 – 2680 cm^{-1}. These bands are due to Fermi resonance [3], i.e., interaction of the CH stretch fundamental with the overtone of the O=C–H bending vibration near 1390 cm^{-1}.

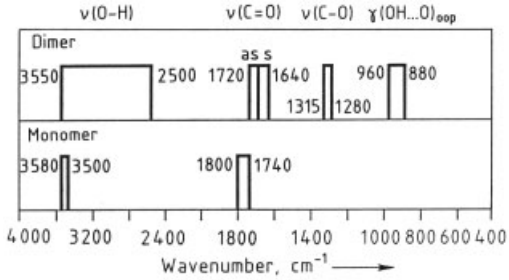

Figure 30. Characteristic group frequencies of carboxylic acids

4.8.3. Carboxylic Acids

Characteristic group frequencies of carboxylic acids are shown in Figure 30. Because carboxylic acids have a pronounced tendency to form

hydrogen-bonded dimers, in the condensed state they are characterized by a very broad infrared OH stretching band centered near 3000 cm^{-1}. As the dimer has a center of symmetry, the antisymmetric C=O stretch at 1720 – 1680 cm^{-1} is infrared active only, whereas the symmetric C=O stretch at 1680 – 1640 cm^{-1} is Raman active only; this is shown in Figure 31 using acetic acid as an example. In the infrared, a broad band at 960 – 880 cm^{-1} due to out-of-plane OH · · · O hydrogen bending is diagnostic of carboxyl dimers, as is the medium intense C−O stretching band at 1315 – 1280 cm^{-1}. On the other hand, monomeric acids have a weak, sharp OH absorption band at 3580 – 3500 cm^{-1} in the infrared, and the C=O stretching band appears at 1800 – 1740 cm^{-1} in both infrared and Raman.

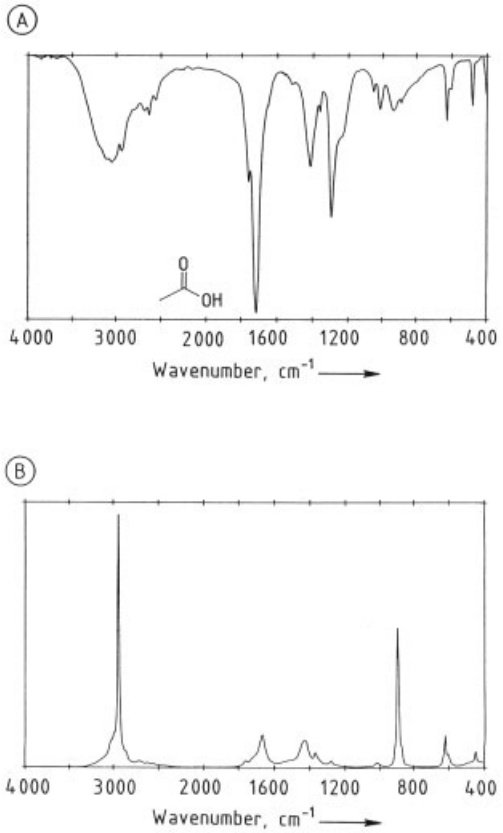

Figure 31. FT-infrared (A) and FT-Raman (B) spectra of acetic acid

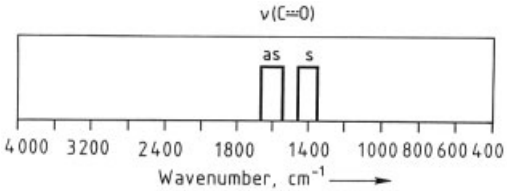

Figure 32. Characteristic group frequencies of carboxylate salts

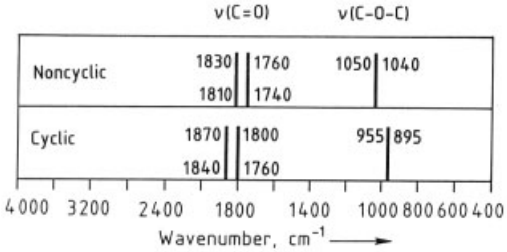

Figure 33. Characteristic group frequencies of anhydrides

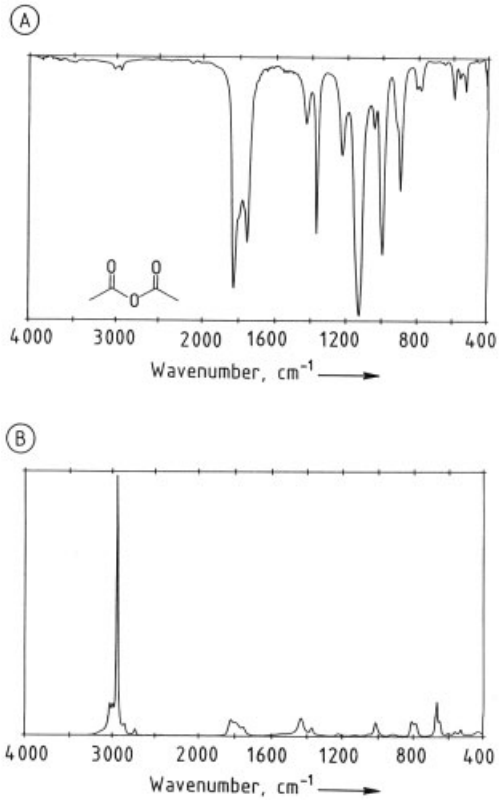

Figure 34. FT-infrared (A) and FT-Raman (B) spectra of acetic anhydride

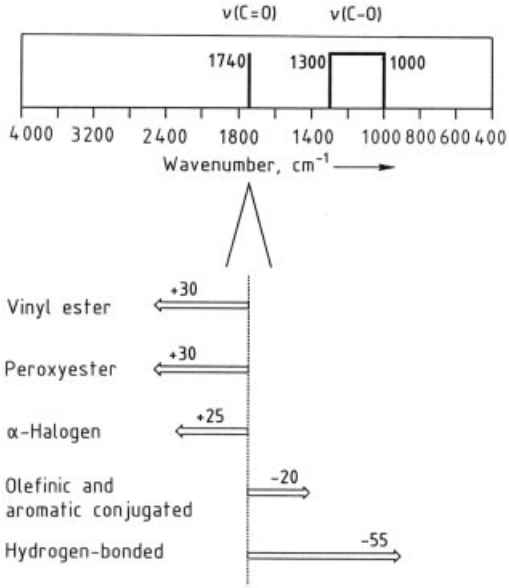

Figure 35. Characteristic group frequencies of esters

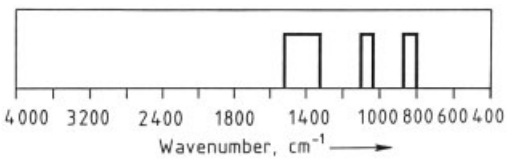

Figure 36. Characteristic group frequencies of carbonate salts

4.8.4. Carboxylate Salts

Characteristic group frequencies of carboxylate salts are shown in Figure 32. In carboxylate salts, the two identical C–O partial double bonds give rise to two bands: the antisymmetric CO_2 stretching vibration at $1650-1550\,cm^{-1}$ is very strong in the infrared and weak in the Raman, while the corresponding symmetric vibration at $1450-1350\,cm^{-1}$ is somewhat weaker in the infrared but strong in the Raman.

4.8.5. Anhydrides

Characteristic group frequencies of anhydrides are shown in Figure 33. Anhydrides are characterized by two C=O stretching vibration bands, one of which occurs above $1800\,cm^{-1}$. These bands are strong in the infrared but weak in the Raman. In noncyclic anhydrides, the infrared C=O band at higher wavenumber is usually more intense than the C=O band at lower wavenumber, whereas in cyclic anhydrides the opposite is observed. In the Raman, the C=O line at higher wavenumber is generally stronger. Unconjugated straight-chain anhydrides have a strong infrared absorption, due to the C–O–C stretching vibration, at $1050-1040\,cm^{-1}$ (except for acetic anhydride: $1125\,cm^{-1}$), while cyclic anhydrides exhibit this infrared band at $955-895\,cm^{-1}$. In Figure 34, the spectra of acetic anhydride are shown as an example.

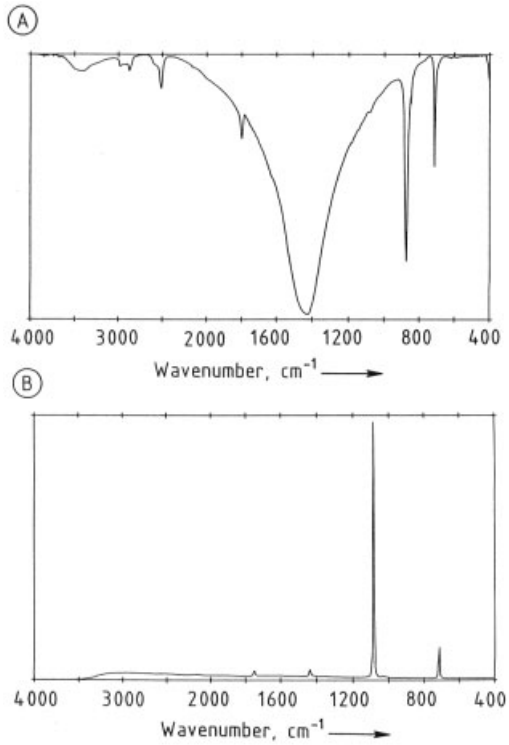

Figure 37. FT-infrared (A) and FT-Raman (B) spectra of calcium carbonate

4.8.6. Esters

Figure 35 shows the characteristic group frequencies of esters. In esters, the C=O stretching vibration gives rise to an absorption band at $1740\,cm^{-1}$ which is very strong in the infrared and medium in the Raman. Its position is strongly influenced by the groups adjacent to

the ester group. The C–O stretching band at 1300 – 1000 cm^{-1} often shows an infrared intensity similar to the C=O band.

4.8.7. Lactones

In lactones, the position of the C=O band is strongly dependent on the size of the ring (Table 5). In lactones with unstrained, six-membered and larger rings the position is similar to that of noncyclic esters.

4.8.8. Carbonate Salts

Characteristic group frequencies of carbonate salts are shown in Figure 36. In inorganic carbonate salts, the antisymmetric CO_3 stretching vibration gives rise to a very broad absorption band at 1520 – 1320 cm^{-1}, which is very strong in the infrared but only weak in the Raman. The symmetric CO_3 stretching vibration, on the other hand, gives rise to a very strong Raman line at 1100 – 1030 cm^{-1}, which is normally inactive in the infrared. Another characteristic medium sharp infrared band at 880 – 800 cm^{-1} is caused by out-of-plane deformation of the CO_3^{2-} ion (Fig. 37).

4.8.9. Amides

Figure 38 shows characteristic group frequencies of amides. In amides, the C=O stretching vibration gives rise to an absorption at 1680 – 1600 cm^{-1} known as the amide I band; it is strong in the infrared and medium to strong in the Raman. In primary amides, a doublet is usually observed in the amide I region, the second band being caused by NH_2 deformation. A characteristic band for secondary amides, which mainly exist with the NH and C=O in the *trans* configuration, is the amide II band at 1570 – 1510 cm^{-1}, involving in-plane NH bending and C – N stretching. This band is less intense than the amide I band.

As in amines, the NH stretching vibration gives rise to two bands at 3550 – 3180 cm^{-1} in primary amides and one near 3300 cm^{-1} in secondary amides; these are strong both in infrared and Raman. In secondary amides a weaker band appears near 3100 cm^{-1} due to an overtone of the amide II band.

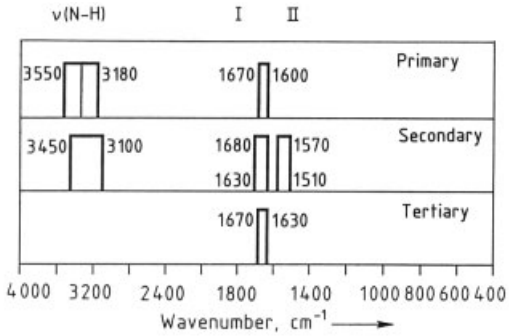

Figure 38. Characteristic group frequencies of amides

4.8.10. Lactams

Characteristic group frequencies of lactams are shown in Figure 39. As in lactones, the size of the ring influences the C=O frequency in lactams. In six- or seven-membered rings the carbonyl vibration absorbs at 1680 – 1630 cm^{-1}, the same as the noncyclic *trans* case. Lactams in five-membered rings absorb at 1750 – 1700 cm^{-1}, while β-lactams absorb at 1780 – 1730 cm^{-1}. These bands are strong in the infrared and medium in the Raman.

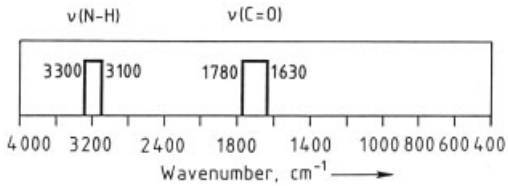

Figure 39. Characteristic group frequencies of lactams

Because of the cyclic structure the NH and C=O groups are forced into the *cis* configuration so that no band comparable to the amide II band in *trans* secondary amides occurs in lactams. A characteristically strong NH stretching absorption near 3200 cm^{-1} occurs in the infrared, which is only weak to medium in the Raman. A weaker infrared band near 3100 cm^{-1} is due to a combination band of the C=O stretching and NH bending vibrations.

Table 5. Characteristic group frequencies of lactones

Lactone		Structure	ν (C = O), cm^{-1}
β-Lactones			1840
γ-Lactones:	saturated		1795 – 1775
	α, β-unsaturated		1785 – 1775 and 1765 – 1740
	β, γ-unsaturated		1820 – 1795
δ-Lactones:	saturated		1750 – 1735
	α, β-unsaturated		1760 – 1750 and 1740 – 1715
	β, γ-unsaturated		

4.9. Sulfur-Containing Compounds

Sulfur-containing functional groups generally show a strong Raman effect, and groups such as C–S and S–S, for example, are extremely weak infrared absorbers. Hence, Raman spectra of sulfur-containing compounds have a much greater diagnostic value than their infrared spectra.

4.9.1. Thiols

Characteristic group frequencies of thiols are shown in Figure 40. The S−H stretching vibration of aliphatic and aromatic thiols gives rise to a medium to strong Raman line at 2590 – 2530 cm^{-1}; this band is weak or very weak in the infrared.

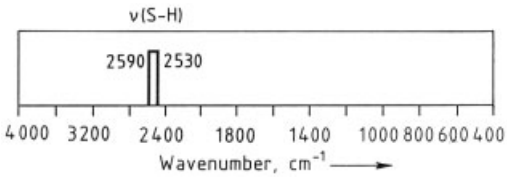

Figure 40. Characteristic group frequencies of thiols

4.9.2. Sulfides and Disulfides

Figure 41 shows characteristic group frequencies of sulfides and disulfides. The stretching vibration of C–S bonds gives rise to a weak infrared but a strong Raman signal at 730 – 570 cm^{-1}. Similarly, the S–S stretching vibration at 500 cm^{-1} is a very strong Raman line but very weak infrared band (Fig. 42). Both Raman signals are very diagnostic in the conformational analysis of disulfide bridges in proteins.

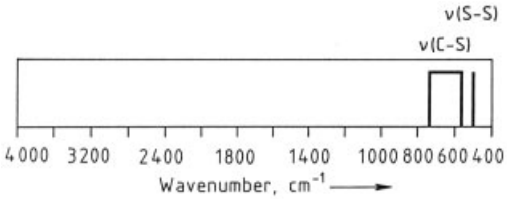

Figure 41. Characteristic group frequencies of sulfides and disulfides

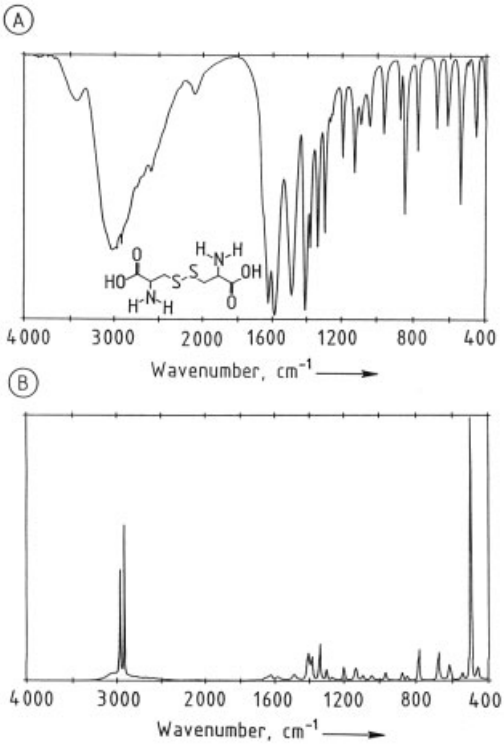

Figure 42. FT-infrared (A) and FT-Raman (B) spectra of L-cystine

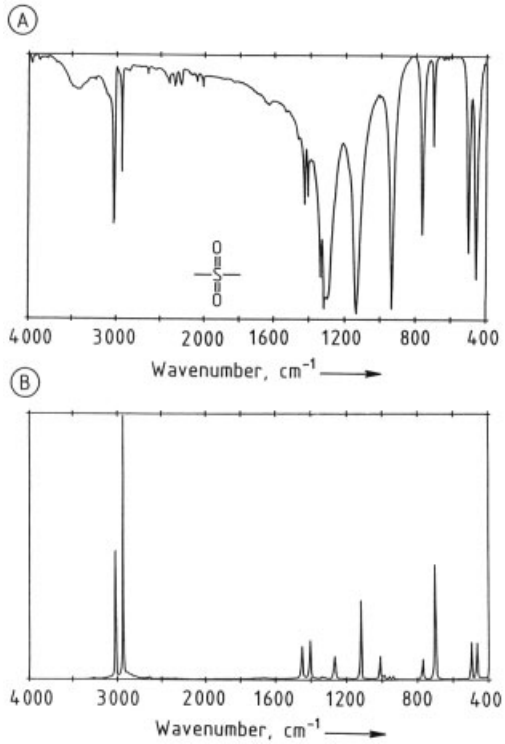

Figure 44. FT-infrared (A) and FT-Raman (B) spectra of dimethyl sulfone

4.9.3. Sulfones

Similar to nitro groups and carboxyl salts, sulfones are characterized by two bands due to antisymmetric and symmetric stretching of the SO_2 group, the former at 1350 – 1290 cm^{-1} and the latter at 1160 – 1120 cm^{-1} (Fig. 43). Both bands are strong in the infrared, while only the symmetric vibration results in a strong Raman line (Fig. 44).

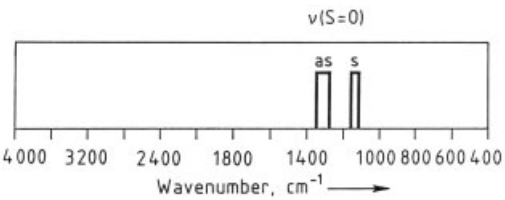

Figure 43. Characteristic group frequencies of sulfones

4.10. Computer-Aided Spectral Interpretation

After his first experiences an interpreting vibrational spectra the spectroscopist soon realizes the limited capacity of the human brain to store and selectively retrieve all required spectral data. Powerful micro- and personal computers are now available together with fast and efficient software systems to help the spectroscopist to identify unknown compounds. IR Mentor [41] is a program that resembles an interactive book or chart of functional group frequencies. Although the final interpretation remains in human hands, this program saves the spectroscopist time by making tabular correlation information available in computer form, and, moreover, it is also an ideal teaching tool. To facilitate the automated identification of unknown compounds by spectral comparison numerous systems based on library search, e.g., SPECTACLE [42], GRAMS/32 [43], or those offered

by spectrometer manufacturers, are used. These systems employ various algorithms for spectral search [44], and the one that uses full spectra according to the criteria of LOWRY and HUPPLER [45] is the most popular. The central hypothesis behind library search is that if spectra are similar then chemical structures are similar [46]. In principle, library search is separated into identity and similarity search systems. The identity search is expected to identify the sample with only one of the reference compounds in the library. If the sample is not identical to one of the reference compounds the similarity search presents a set of model compounds similar to the unknown one and an estimate of structural similarity. The size and contents of the library are crucial for a successful library search system, especially a similarity search system. A smaller library with carefully chosen spectra of high quality is more useful than a comprehensive library containing spectra of all known chemical compounds or as many as are available because, in similarity search, the retrieval of an excessive number of closely similar references for a particular sample only increases output volume without providing additional information [46]. A critical discussion of the performance of library search systems is presented by CLERC [46]. While library search systems are well-established, invaluable tools in daily analytical work, expert systems (i.e., computer programs that can interpret spectral data) based on artificial neural networks (ANNs) are more and more emerging [47], [48].

5. Applications of Vibrational Spectroscopy

5.1. Infrared Spectroscopy

5.1.1. Transmission Spectroscopy

Transmission spectroscopy is the simplest sampling technique in infrared spectroscopy, and is generally used for routine spectral measurements on all kinds of samples. It requires only simple accessories, such as infrared-transparent windows [e.g., potassium bromide, sodium chloride, or thallium bromoiodide (KRS-5)] for gases and liquids, or a sample holder for solids. The sample is placed in the light beam of an infrared spectrometer, and the intensity of the incident beam is compared with that transmitted by the sample. According to the fundamental relation governing the absorption of radiation as a function of transmittance (see Section 3.2), parameters that can be determined in transmission spectroscopy are the thickness or the concentration of the sample. The thickness can range from micrometers for solids and liquids to even kilometers for gas samples.

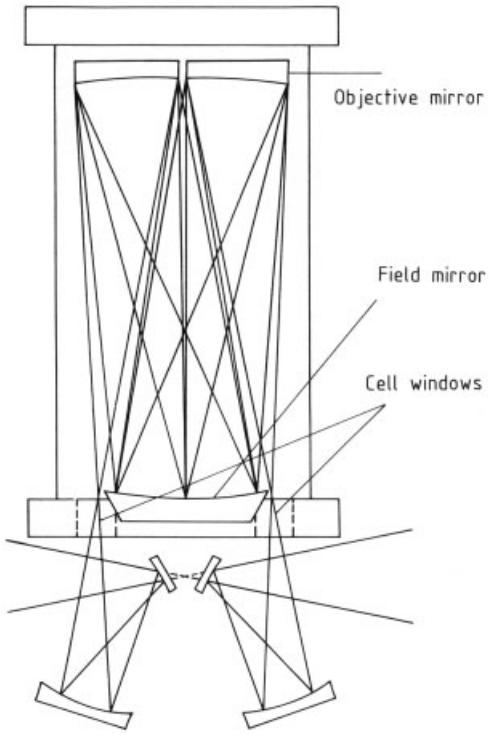

Figure 45. Optical diagram of a White gas cell (A mode) showing eight passes, to change the path length only one objective mirror is rotated, the rotation is nonlinear and is very small at long path lengths. For B mode operation, the light beam enters and exits on one and the same side of the field mirror, the objective mirrors are coupled and are rotated together linearly over the entire operational range
(Reproduced by permission of Harrick Scientific Corporation, Ossining, NY 10562)

Apart from general laboratory gas analysis, gas-phase transmission spectroscopy is used for the characterization of air pollutants as well as the monitoring of stack gas and of air quality in work places. As air quality has become a major concern for environmentalists, industrial manufacturers and many others, spectroscopic mea-

surement in the gas phase is receiving increasing attention. However, two major problems must be overcome: absorptions in the gas phase are many times weaker than those in the liquid or solid phase, and the species to be measured usually makes up only a small fraction of the total gas volume. The most popular way of surmounting these problems is to increase the path length by reflecting the beam several times from precisely controlled mirrors. Such multiple-pass gas cells are all based on the White design [49] or variations thereof. Figure 45 shows an optical diagram of a White cell aligned for eight passes. The length of the gas cell and the number of passes employed determines the optical path length, which may vary from centimeters to kilometers. The number of passes is adjusted in four pass increments by rotating one of the objective mirrors. In the A mode of operation, only one objective mirror is rotated, while in the B mode the objective mirrors are coupled and rotated together linearly over the entire operational range to change the path length. To prepare a sample, the gas cell is evacuated, the sample is bled into the cell, and the cell is sealed.

Liquid samples are easily studied with the aid of a wide variety of liquid cells, including heatable, flow through, and variable path length cells. These cells are constructed of two infrared transparent windows with a spacer between them, thereby forming a cavity for the sample. As the samples are often strong infrared absorbers, for this spectral range the liquid transmission cell must be constructed with short optical path lengths (0.025 – 1 mm).

Polycrystalline or powdered samples can be prepared as a suspension in mineral oil (Nujol mull), as a potassium bromide disk (pellet), or as thin films deposited on infrared-transparent substrates. The potassium bromide pellet is the most common way of preparing powder samples; in this method a small amount, usually 1 mg, of finely-ground solid sample is mixed with powdered potassium bromide, usually 300 mg, and then pressed in an evacuated die under high pressure. The resulting disks are transparent and yield excellent spectra. The only infrared absorption in the potassium bromide matrix is due to small amounts of adsorbed water, which can, however, be confused with OH- containing impurities in the sample. To minimize adsorbed water the potassium bromide is ground as little as possible. The sample is ground separately and then mixed with potassium bromide, after which the mixture is ground only so far as is necessary to achieve good mixing.

5.1.2. External Reflection Spectroscopy

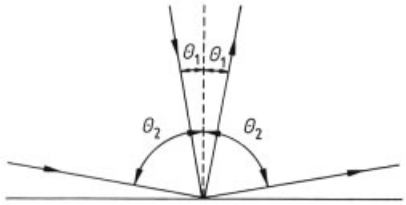

Figure 46. Near-normal (θ_1) and grazing angle (θ_2) incidence

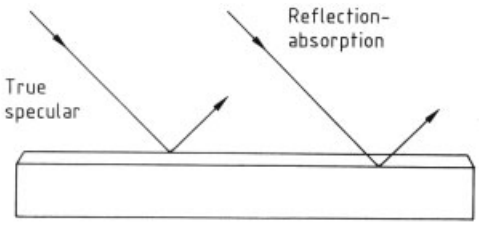

Figure 47. Types of specular reflection

In external reflection spectroscopy, electromagnetic radiation is reflected at the interface of two materials differing in refractive index or from a smooth surface, a conductor such as metal, a semiconductor, a dielectric, or a liquid. The intensity of light reflected is measured to obtain information about the surface chemistry and structure. The angle of collection is the same as the angle of incidence and is measured from the normal to the surface (Fig. 46). Small-angle incidence (θ_1, usually 10°) is called near-normal incidence while that at large angles (θ_2, usually 80°) is known as grazing angle incidence. A nondestructive method requiring no sample preparation, external reflection spectroscopy comprises two different types of measurements, both being termed specular reflectance. In the first type, the reflection from a smooth mirrorlike surface, e.g., free-standing film, single crystal face, or surface of an organic material, is referred to as true specular reflection or simply specular reflection (Fig. 47). In the second, a film of sample is deposited on a highly reflecting substrate, usually a

metal, from which the infrared beam is reflected, the radiation is transmitted through the sample layer, reflected from the substrate surface, and finally transmitted back through the sample. This technique is called reflection – absorption (RA), infrared reflection-absorption spectroscopy (IR-RAS), or transflectance.

A material's object properties are characterized by its complex refractive index, $n' = n + ik$, where n is the refractive index and k is the absorption index, both of which are dependent on the frequency of incident light. Unlike normal transmission spectra, true specular reflection spectra often have distorted, derivative bands because specular reflectance from a surface frequently has a large refractive index component. However, by using a mathematical operation known as the Kramers – Kronig transformation [50] a transmission-like spectrum showing the expected absorptivity information can be calculated from a measured, distorted spectrum. As true specular reflectance is often measured at or near normal incidence, the reflected energy is small, usually between 0 and 12 % for most organic materials. However, when this low signal-to-noise ratio is overcome by using more sensitive cooled MCT detectors the true specular reflection technique has the important advantage of permitting the noncontact analysis of bulk solids and liquids. Depending on the thickness of the film of sample, two types of reflection – absorption measurement are used: near-normal and grazing angle. In near-normal reflection – absorption, the thickness of the sample usually ranges between 0.5 and 20 µm and the angle of incidence varies between 10 and 30°. Hence, the amount of energy reflected from the substrate is much greater than the amount reflected from the front surface of the sample film, and a completely transmission-like reflection – absorption spectrum is obtained. Grazing angle reflection – absorption measurements are more sensitive than near-normal reflectance and are used to analyze sample films less than 1 µm thick, and even monolayers, on highly reflective materials. The improved sensitivity is due to the enhanced electromagnetic field that can be produced at the reflection surface by using polarized light at large angles of incidence [51], [52]. As the plane of incidence contains the incident and reflected rays, as well as the normal to the reflection surface, the electrical field of p-polarized light is parallel to it whereas that of s-polarized light is perpendicular to it. For s-polarized light, the phase of the reflected beam is shifted 180° from the incident beam at all angles of incidence (Fig. 48), and the two beams interfere at the surface to produce a node. Hence little interaction occurs between the infrared radiation and films thinner than one-quarter of the wavelength. However, for p-polarized light, the phase shift is approximately 90° at large angles of incidence, and the combination of incident and reflected beams at the surface produces a standing wave, giving an intense electrical field oriented normal to the surface. For metal substrates the greatest enhancement of the electrical field is produced by using p-polarization at large angles of incidence (e.g., 80°).

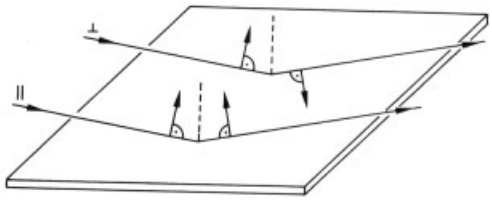

Figure 48. Phase shift of the reflected beam occurring at grazing angle incidence with perpendicular (180° shift) and parallel (90° shift) polarized light

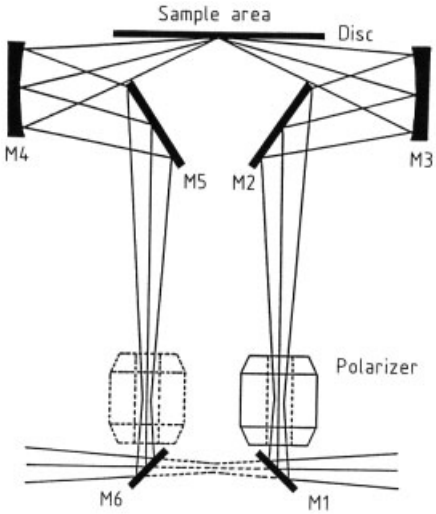

Figure 49. Ray diagram of the wafer/disk checker attachment for recording spectra of thin films on large samples, e.g., lubricants on hard magnetic disks;
$M_1 - M_6$ = mirrors
(Reproduced by permission of Harrick Scientific Corporation, Ossining, NY 10562)

Reflection – absorption measurements are used in the analysis of a wide range of industrial products, including polymer coatings on metals, lubricant films on hard magnetic disks (Fig. 49), semiconductor surfaces, or the inner coating of food and beverage containers. In addition, grazing angle reflectance is especially suited to the study of molecular orientations at surfaces [53]. External reflectance at different angles of incidence can be employed to determine both the thickness and refractive index of a film of sample by measuring the interference fringe separation [54]; Figure 50 shows one of the various external reflection attachments which are available.

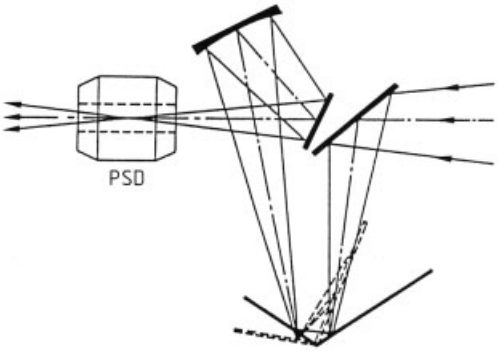

Figure 50. Ray diagram of the versatile reflection attachment (VRA) with single diamond polarizer (PSD) (Reproduced by permission of Harrick Scientific Corporation, Ossining, NY 10562)

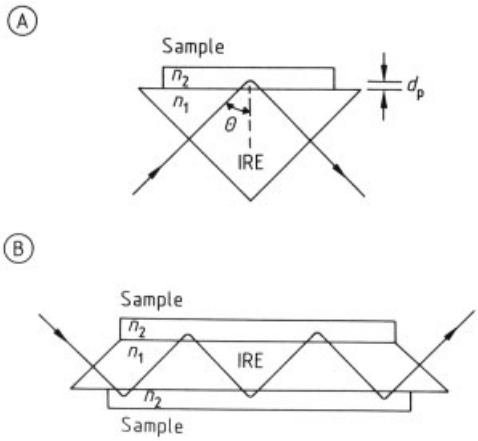

Figure 51. Schematic representation of total internal reflection with: A) Single reflection; B) Multiple reflection
IRE (internal reflection element)
n_1 = Refractive index of the internal reflection element; n_2 = Refractive index of the sample with $n_2 < n_1$; θ = Angle of incidence; d_p = Depth of penetration

5.1.3. Internal Reflection Spectroscopy

Internal reflection spectroscopy, also known as attenuated total reflectance (ATR) or multiple internal reflectance (MIR), is a versatile, nondestructive technique for obtaining the infrared spectrum of the surface of a material or the spectrum of materials either too thick or too strongly absorbing to be analyzed by standard transmission spectroscopy.

The technique goes back to Newton [55] who, in studies of the total reflection of light at the interface between two media of different refractive indices, discovered that an evanescent wave in the less dense medium extends beyond the reflecting interface. Internal reflection spectroscopy has been developed since 1959, when it was reported that optical absorption spectra could conveniently be obtained by measuring the interaction of the evanescent wave with the external less dense medium [56], [57]. In this technique, the sample is placed in contact with the internal reflection element (IRE), the light is totally reflected, generally several times, and the sample interacts with the evanescent wave (Fig. 51) resulting in the absorption of radiation by the sample at each point of reflection. The internal reflection element is made from a material with a high refractive index; zinc selenide (ZnSe), thallium iodide – thallium bromide (KRS-5), and germanium (Ge) are the most commonly used. To obtain total internal reflection the angle of the incident radiation θ must exceed the critical angle θ_c [58]. The critical angle is defined as:

$$\theta_c = \sin^{-1} \frac{n_2}{n_1} \tag{17}$$

where n_1 is the refractive index of the internal reflection element and n_2 is the refractive index of the sample. What makes ATR a powerful technique is the fact that the intensity of the evanescent wave decays exponentially with the distance from the surface of the internal reflection element. As the effective penetration depth is usually a fraction of a wavelength, total internal reflectance is generally insensitive to sample thickness and so permits thick or strongly absorbing samples to be analyzed. The depth of penetration d_p, defined as the distance required for the electrical field amplitude to fall to e^{-1} of its value at the interface, is given by:

$$d_p = \frac{\lambda_1}{2\pi\left(\sin^2\theta - n_{21}^2\right)^{1/2}} \tag{18}$$

where $\lambda_1 = \lambda/n_1$ is the wavelength in the denser medium, and $n_{21} = n_2/n_1$ is the ratio of the refractive index of the less dense medium divided by that of the denser [56].

Although ATR and transmission spectra of the same sample closely resemble each other, differences are observed because of the dependency of the penetration depth on wavelength: longer wavelength radiation penetrates further into the sample, so that in an ATR spectrum bands at longer wavelengths are more intense than those at shorter ones.

The depth of penetration also depends on the angle of incidence; hence, an angle of 45°, which allows a large penetration depth, is generally used to analyze organic substances, rather than an angle of 60°, which results in a substantially weaker spectrum due to the decreased depth of penetration.

The degree of physical contact between sample and internal reflection element determines the sensitivity of an ATR spectrum. To achieve this, a horizontal ATR accessory such as FastIR [58], in which the top plate is the sampling surface, is used [59]; reproducible contact is ensured by a special sample clamp or powder press. Good quality spectra are thus obtained for many materials that present problems of analysis with routine transmission methods, e.g., powders, pastes, adhesives, coatings, rubbers, fibers, thick films, textiles, papers, greases, foams, and viscous liquids. Possible methods of obtaining a spectrum from a variety of samples are discussed in [60]; in situ ATR spectroscopy of membranes is described in [61], [62]. Liquid samples are also well suited to ATR analysis. Most liquids require a very short path length; aqueous samples, for instance, are measured at path lengths of no more than ca. 15 µm, which makes the design of transmission cells difficult because flow of liquids is hindered; they also exhibit interference fringes because of the small spacing between the high refractive index infrared windows. These problems are eliminated by using liquid ATR cells, a variant of solid ATR, in which the internal reflection element is surrounded by a vessel into which the liquid is poured. Various such liquid cells are available, e.g., the Circle [63], the Prism Liquid Cell (PLC) [58], the Squarecol [64] and the Tunnel Cell [65]. With optimized optical design and fixed internal reflection elements these cells provide a highly reproducible path length which permits the quantitative analysis of liquids and aqueous solutions [66]. Liquid ATR cells are uniquely suited to fully automatable, on-line process-monitoring of liquids and viscous fluids. In addition, with suitable optical transfer modules such as the Axiot system [67], it is possible to operate liquid cells outside the sample compartment of the FT-IR spectrometer. This is important for on-line applications in which it is not practicable to pipe the sample to the spectrometer because the analysis can be carried out at the most desirable location, even in a fume hood or on a process line. With an ATR sensing head, immersion probes [68] offer the possibility of in situ FT-IR spectroscopy, for monitoring batch process reactions (Fig. 52), laboratory process development, the identification of hazardous waste, the verification of the contents of storage drums, and the inspection of incoming raw materials.

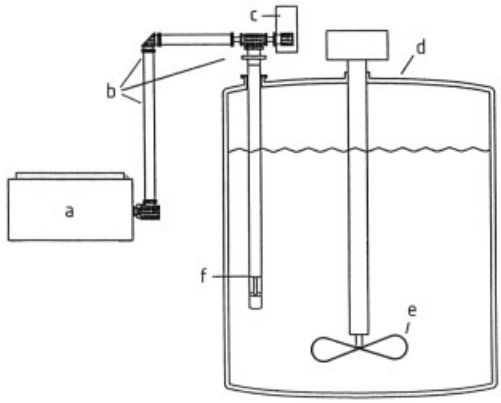

Figure 52. Schematic drawing of an FT-IR measurement system utilizing the Deep Immersion Probe Model DPR-124 mounted in a batch reaction vessel [68]
a) FT-IR spectrometer; b) Optical transfer elements; c) Detector assembly; d) Reaction vessel; e) Mixing blade; f) ATR sensing head
(Reproduced by permission of Axiom Analytical, Inc., Irvine, CA 92614)

5.1.4. Diffuse Reflection Spectroscopy

Diffuse reflection spectroscopy, also known as diffuse reflectance infrared Fourier transform

(DRIFT) spectroscopy, enables the analysis of many samples for which traditional techniques fail, to be made with little or no sample preparation. Many substances in their natural state, especially powders but also any solid with a rough surface, such as dyed textiles and printed papers, exhibit diffuse reflection, i.e., incident light is scattered in all directions, unlike specular reflection (see Section 5.1.2) where the angle of incidence equals the angle of reflection. Diffuse reflectance spectra are in practice complex, because of spectral distortions caused by the mixing of both absorbance and reflectance features owing to contributions from transmission, internal, and specular components in the measured radiation. Hence, DRIFT spectra are affected by particle size, packing, and dilution of the sample.

In diffuse reflectance spectra large glossy particle faces or crystal surfaces cause specular reflection which produces inverted bands, known as reststrahlen bands; this effect is reduced by fine grinding of the sample. Reststrahlen bands may also occur in neat samples with a rough surface, such as dyed textiles, printed papers, and agricultural and botanical specimens [69], measured by direct nondestructive, DRIFT spectroscopy. In this case, roughening the surface with emery paper reduces the disturbing bands because with smaller, thinner particles on the roughened surface the transmission component of the collected light increases, thus decreasing the reststrahlen effect. In strongly absorbing samples, the slight penetration of the incident beam produces mainly specular reflection at the surface, which also results in strong reststrahlen bands. By diluting the sample with a nonabsorbing powder such as KBr the effect is minimized or even eliminated, as this ensures deeper penetration of the incident beam and, therefore, an increased contribution of transmission and internal reflection components to the DRIFT spectrum. The diffuse reflectance spectrum is usually calculated as the ratio of sample and reference reflectance, with the pure diluent being taken as reference. Typically, a mixture of 90 – 95 % diluent and 10 – 5 % sample is used. Although the diluted spectra and the bulk reflectance spectra of weakly absorbing compounds have a similar appearance, they show enhanced intensity at lower wavenumbers compared to transmission spectra.

The interpretation of diffuse reflectance is based on a theory developed by Kubelka and Munk [70] and extended by Kortüm, Braun, and Herzog [71] for diluted samples in nonabsorbing matrices. The diffuse reflectance R_∞ of a diluted sample of infinite thickness (i.e., a sample for which an increase in thickness does not appreciably change the spectrum) is linearly related to the concentration c of the sample, as given by the Kubelka – Munk (K – M) equation:

$$f(R_\infty) = \frac{(1-R_\infty)^2}{2R_\infty} = \frac{2.303ac}{s} \tag{19}$$

where a is the absorptivity and s is the scattering coefficient. Because the scattering coefficient depends on both particle size and degree of sample packing, the linearity of the K – M function can only be assured if particle size and packing method are strictly controlled. If these conditions are fulfilled, accurate quantitative diffuse reflectance measurements can be obtained [72], [73].

Initiated by the work of Griffiths et al. [74], diffuse reflectance accessories have been developed that pay special attention to the optimal reduction of the disturbing specular component. For example, the Praying Mantis diffuse reflectance attachment (Fig. 53) uses an off-line collection angle incorporating two 6 : 1 90° off-axis ellipsoidal mirrors. One ellipsoidal mirror focuses the incident beam on the sample, while the radiation diffusely reflected by it is collected by the other. Both ellipsoidal mirrors are tilted forward; therefore, the specular component is deflected behind the collecting ellipsoid, permitting the diffusely reflected component to be collected. As diffuse reflectance is a very sensitive technique for the analysis of trace quantities of materials, all available attachments allow for microsampling [75]. The analysis of thin-layer chromatograms by DRIFT spectroscopy, in situ or after transfer, is described in [76], [77]. Catalytic reactions can be monitored by using a controlled reaction chamber attached to a DRIFT accessory [78], [79]. Intractable samples, such as paint on automobile panels, can be analyzed by using silicon carbide abrasives [80]: a silicon carbide disk is rubbed against the sample, a small amount of which adheres to the disk. The DRIFT spectrum of the sample is then quickly obtained by placing the disk in the diffuse reflectance attachment.

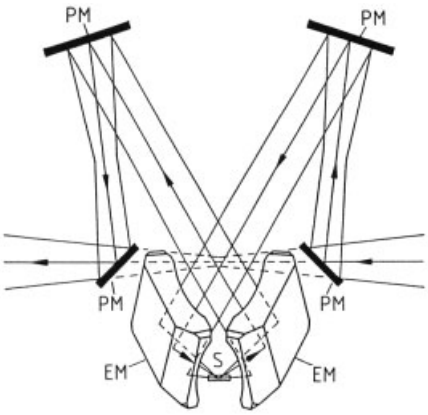

Figure 53. Ray diagram of the Praying Mantis diffuse reflectance attachment
EM = Ellipsoidal mirror; PM = Planar mirror; S = Sample
(Reproduced by permission of Harrick Scientific Corporation, Ossining, NY 10562)

5.1.5. Emission Spectroscopy

Infrared emission spectroscopy is a useful and effective technique for studying the surface of liquid and solid organic, inorganic, and polymeric materials, and for the in situ process-monitoring of high-temperature reactions and thermal transformations [81], [82]. In contrast to the conventional transmission technique, in emission spectroscopy the sample itself is the infrared source. The temperature of the sample is raised, which increases the Boltzmann population of the vibrational energy states. Infrared radiation is emitted on return to the ground state. In theory, emission spectra can always be measured provided the sample is at a different temperature to the detector. Like transmission spectra, emission spectra are obtained by dividing the emission of the sample by that of the reference, which is often a blackbody source at the same temperature as the sample. The main reason for the lack of any extensive work in this technique before 1970 was the low intensity of infrared emission, which results in spectra with poor signal-to-noise ratio. However, the availability of FT-IR spectrometers with the potential for low-energy sampling and for the development of highly efficient FT-IR emission spectroscopy accessories [83], [84] has led to a resurgence of interest in infrared emission spectroscopy.

Because of the phenomenon of self-absorption the ideal sample for conventional emission studies is a thin layer (e.g., a polymer film), on both metal and semiconductor surfaces [81]. A sample is usually heated from below the emitting surface, the lower surface thus having a higher temperature than the upper one. Therefore, radiation emitted from below the upper surface is absorbed before it reaches the surface, and this self-absorption of previously emitted light severely truncates and alters features in the emission spectra of optically thick samples. This problem is overcome by using a laser for controlled heat generation within a thin surface layer of the sample, self-absorption of radiation thus being minimized. These methods, known as laser-induced thermal emission (LITE) spectroscopy [85], [86] and transient infrared emission spectroscopy (TIRES) [87], [88] can produce analytically useful emission spectra from optically thick samples. Quantitative applications of infrared emission spectroscopy are described in [89–91].

5.1.6. Photoacoustic Spectroscopy

Photoacoustic spectroscopy (PAS) [92–95] is a fast, nondestructive method for analyzing various materials in the gas, liquid, or solid state with virtually no sample preparation. In cases where samples are insoluble, difficult to grind into a powder, or of irregular shape – such as coal, carbon-filled and conducting polymers, pharmaceutical preparations, spin coatings on fibers, and coatings on irregular surfaces – photoacoustic spectroscopy complements other established techniques such as attenuated total reflectance (ATR) and diffuse reflectance (DRIFT). Moreover, measurements on samples that show the Christiansen effect [96] when prepared as KBr pellets for standard transmission spectroscopy can be better performed by the photoacoustic method. As most photoacoustic measurements are made with solid or semisolid samples, the sample is sealed in a small-volume cell containing a nonabsorbing gas such as helium and having windows for optical transmission (Fig. 54). The sample is illuminated with modulated radiation from the monochromator of a dispersive instrument or the interferometer of a Fourier transform infrared spectrometer. At

wavelengths where the sample absorbs a fraction of the incident radiation, a modulated temperature fluctuation at the same frequency, but not necessarily with the same phase, as that of the incident radiation is generated in the sample, and the surrounding inert gas produces periodic pressure waves in the sealed cell. These fluctuations in pressure can be detected by a microphone because the modulation frequency of the incident beam usually lies in the acoustic range, e.g., between 100 and 1600 Hz. The low power available from dispersive instruments results in spectra with a rather poor signal-to-noise ratio; hence, photoacoustic mid-infrared spectra of condensed substances are preferably obtained by using an FT-IR spectrometer with a photoacoustic cell accessory. In a rapid-scanning Michelson interferometer with a mirror velocity of v (in cm/s), radiation of wavenumber $\tilde{\nu}$ (in cm^{-1}) is modulated with a frequency of $2\,v\,\tilde{\nu}$ Hz [8]. Therefore, FT-IR spectrometers are ideal for photoacoustic measurements in the mid-infrared, where mirror velocities on the order of $0.05-0.2\,cm\,s^{-1}$ provide modulation frequencies in the acoustic range. As carbon black absorbs all incident radiation, the sample single-beam spectrum is usually ratioed against a carbon black single-beam spectrum to obtain an absorbance-like spectrum.

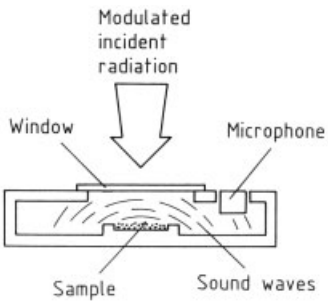

Figure 54. Schematic of a photoacoustic cell

Before the first Fourier transform infrared photoacoustic spectrum was reported in 1978 [97], most infrared photoacoustic experiments were performed by using high-intensity tunable laser sources. However, as the source is monochromatic, this type of experiment is only useful for monitoring a single line in the spectrum. Highly-developed photoacoustic accessories are now available from FT-IR manufacturers or other companies [98], which have a high sensitivity also for microsamples. As the signal is inversely proportional to the cell volume, cells optimized for the measurement of condensed samples have a small volume, typically well under $1\,cm^3$, to enhance the signal amplitude [99]. Water vapor or carbon dioxide generate good photoacoustic signals, so even traces of these impurities cause serious interference in the photoacoustic spectroscopy signal; therefore, the purity of the gas in the cell is essential for good-quality photoacoustic spectra.

Important applications of photoacoustic spectroscopy are near-surface analysis [100] and depth profiling of polymers and polymer layers [101–103]. The theory of depth profiling is described in detail in [92]. It is based on the fact that thermal diffusion length (the distance a thermal wave travels in the sample before its intensity is reduced by 1/e in magnitude) and hence intensity of the photoacoustic signal are increased by lowering the modulation frequency. In a rapid-scanning interferometer this is achieved by decreasing the mirror velocity. To date, rapid-scanning interferometers that provide modulation frequencies which continuously change with wavenumber have generally been used. However, step-scan interferometers accommodating any desired modulation frequency, constant over the entire wavelength range, produce improved photoacoustic spectra and are thus more effective in depth profiling [104–107].

Saturation effects, which can occur for stronger absorption bands, are the main limitation of photoacoustic spectroscopy. These distortions can be overcome by using thin samples [93] or by analyzing the phase of the photoacoustic signal [108]. The latter method is particularly useful for quantitative analysis.

5.1.7. Chromatography/Fourier Transform Infrared Spectroscopy

The combination of chromatographic separation with Fourier transform infrared spectroscopy has significantly improved the analysis of complex mixtures [109]. In chromatography/FT-IR systems, an infrared detector providing information on the structure of separated species is used instead of standard bulk chromatography detectors such as thermal conductivity, flame ioniza-

tion, ultraviolet, or fluorescence; these detectors can be used for quantitative analysis when the identities of the mixture components are known.

In these systems the following factors must be taken into consideration: the physical state of the matter to be analyzed, the optimization of the interface between chromatograph and FT-IR, and the specific requirements and constraints of all three parts. In gas chromatography (GC)/FT-IR spectroscopy [110], one type of interface is composed of temperature-controlled, long, thin flow-through gas cells or "light pipes" coated with gold for high reflectivity. As the light-pipe techniques have a detection limit in the nanogram range, subnanogram limits can be routinely attained with the matrix-isolation technique (GC/MI/FT-IR) [109], and even better detection limits are achieved with the subambient trapping GC/FT-IR technique [109], [111]. Rapid-scanning interferometers able to record the entire spectrum in less than one second at low resolution (usually 4 or 8 cm^{-1}), together with highly sensitive detectors (e.g., cooled MCT detectors), are essential in all column chromatography/FT-IR spectroscopic techniques, with capillary GC/FT-IR requiring the most rapid data acquisition. In comparison to vapor phase species, rotations of matrix-isolated species are hindered, which results in sharper and more intense bands. Hence, for accurate spectra representation, GC/MI/FT-IR spectra are usually measured at 1 cm^{-1} resolution.

The Gram – Schmidt vector orthogonalization method [112] is the most sensitive and computationally most efficient and, hence, the most popular procedure for the reconstruction of the total chromatographic trace from interferometric infrared data. Selective detection is obtained by functional-group-specific chromatograms (chemigrams [113]), which are produced by computing absorbance spectra during chromatographic separation and plotting integrated absorbance for prespecified wavelength windows as a function of separation time, representing the quantity of a specific functional group that elutes as a function time. In chromatography/FT-IR spectral evaluation, in order to classify unknowns as to compound type, library searching (see Section 4.10) has proved to be very useful.

The addition of a complementary detector significantly increases the power of GC/FT-IR analysis. Mass spectrometry (MS) is an ideal choice [114] because the disadvantages of each method – infrared spectroscopy often cannot distinguish between long-chain homologues, and mass spectrometry frequently fails to distinguish isomeric species – are offset by the other.

The crucial point in liquid chromatography (LC)/FT-IR methods, however, is the interferences caused by the liquid phase. Actually, two viable LC-IR systems are commercially available: the LC-Transform [115] and the InfraRed Chromatograph (IRC) [116], [117]. While the LC-Transform is a solvent-elimination interface, the IRC provides full FT-IR spectra of the eluent stream from an LC column. The LC-Transform system utilizes a high-performance ultrasonic nebulizer or a pneumatic linear capillary nozzle to remove the mobile phase from samples as they elute from the chromatograph. The capillary is surrounded by hot, flowing sheath gas which provides sufficient thermal energy to evaporate the mobile phase and to contain the nebulized spray in a tight, focused cone as the spray emerges from the nozzle. The sample is deposited on a germanium sample-collection disk as spots with diameters between 0.5 and 1 mm. In a consecutive, uncoupled step, good-quality infrared spectra can be obtained from these spots, which contain microgram to submicrogram amounts, by the use of a specially designed optics module, comprising a beam condenser. While this optics module can be used in any FT-IR spectrometer uncoupled from the LC-Transform interface, the IRC includes an FT-IR spectrometer, a data station for driving the spectrometer and for data-processing, and a vacuum interface for depositing and scanning the sample. Being a real-time direct deposition interface, the IRC uses an ultrasonic nebulizer, a desolvation tube and a vacuum chamber to evaporate the solvent from the LC stream. The sample residue is then deposited onto a moving zinc selenide window. The data station collects a continuous array of spectra as the dry sample track on the window moves through the focused beam of the spectrometer.

LC-IR provides powerful identification possibilities, based on molecular structure. A high degree of confidence can be placed on substance identification that simultaneously matches infrared spectral characteristics and chromato-

graphic elution time. This is especially so when molecules that have various conformational isomers are to be distinguished.

Supercritical fluid chromatography (SFC)/FT-IR spectroscopy, generally with carbon dioxide as the mobile phase, bridges the gap between GC/FT-IR and LC/FT-IR, and is particularly useful for separating nonvolatile or thermally labile materials not amenable to gas chromatographic separation [109]. Flow cells, mobile phase elimination and matrix-isolation techniques are used as SFC/FT-IR interfaces.

Applications of chromatography/FT-IR spectroscopy involve toxins and carcinogens, wastewater constituents, sediment extracts, airborne species, pesticides and their degradation products, fuels and fuel feedstocks, flavors and fragrances, natural products, pharmaceuticals, biomedicine, and polymers.

5.1.8. Thermogravimetry/Fourier Transform Infrared Spectroscopy

Thermogravimetry (TG), also called thermogravimetric analysis (TGA), measures changes in sample mass as a function of increasing temperature and has become an important tool for the analysis and characterization of materials [118], [119]. To correlate weight changes with chemical changes, it is necessary to identify gases evolved from the thermogravimetric analyzer furnace. This is made possible by coupling thermogravimetry with FT-IR spectroscopy, which provides the necessary speed and sensitivity.

A system has been described in which the infrared beam of the FT-IR spectrometer is led directly into the TG equipment [120], but in commercially available systems [121], [122] the evolved gases are conducted from the TG system to the spectrometer by carrier gas flow via a transfer line. The interface is heated to prevent recondensation and incorporates the transfer line and a flow-through gas cell with nitrogen or helium as carrier gas.

After infrared spectra at low resolution (usually $4\,cm^{-1}$) have been recorded as a function of time and temperature, the further procedure is similar to GC/FT-IR (see Section 5.1.7) and, in fact, GC/FT-IR software, such as Gram–Schmidt thermogram reconstruction and library search with a vapor-phase spectra database is commonly used. Helpful for the identification in the vapor phase is the virtual absence of interaction between molecules, and therefore the nondestructive IR detection enables in many cases an easy and fast interpretation of single and overlapping absorption bands.

Providing structural information for sample components down to nanogram levels, TG/FT-IR applications include decomposition studies of polymers and laminates [123], the analysis of coal, oil shales, and petroleum source rocks [124], [125], and the determination of activation energies [126] and thermal decomposition of energetic materials [127].

The IR detection of evolved gases opens new possibilities in the quantification of thermogravimetric analysis. Besides measuring the mass loss of a TG step, the total amount can also be calculated after calibration of a single component and integration over time [128]. A further improvement in quantitative TG/FT-IR is achieved by PulseTA [129]. Here, a certain amount of gas or liquid is injected into the system and used as internal calibration standard for the quantification of the unknown amount of evolved gas.

5.1.9. Vibrational Circular Dichroism

Vibrational circular dichroism (VCD) extends the functionality of electronic circular dichroism (CD) into the infrared spectral region, with VCD bands of enantiomeric pairs of chiral molecules having opposite signs. Early VCD spectrometers were dispersive instruments equipped with a CaF_2 photoelastic modulator for creating left and right circularly polarized radiation and a liquid-nitrogen-cooled InSb detector. FT-IR-VCD measurements are carried out by using the so-called double modulation technique [130] in which the VCD interferogram is carried by the polarization modulation frequency in the region of 50 kHz, and the IR transmission interferogram occurs in the usual range of Fourier frequencies, typically below 5 kHz.

Combining the structural specificity of FT-IR spectroscopy with the stereosensitivity of circular dichroism, VCD allows the determination of optical purity without enantiomeric separation or derivatization, as well as of absolute

configuration without crystallization. Simultaneous monitoring of the optical purity of multiple chiral species, such as reactant and product molecules in a reaction process is also possible, as is the determination of conformations of biological molecules such as proteins, peptides, and DNA [131], [132].

5.2. Raman Spectroscopy

In mid-infrared spectroscopy, Fourier transform instruments are used almost exclusively. However, in Raman spectroscopy both conventional dispersive and Fourier transform techniques have their applications, the choice being governed by several factors [133], [134]. Consequently, a modern Raman laboratory is equipped with both Fourier transform and CCD-based dispersive instruments. For a routine fingerprint analysis, the FT system is generally used, because it requires less operator skill and is quicker to set up; the FT system is also be tried first if samples are highly fluorescent or light sensitive. However, if the utmost sensitivity is required, or if Raman lines with a shift smaller than $100\,cm^{-1}$ are to be recorded, conventional spectrometers are usually preferred.

Raman spectroscopy offers a high degree of versatility, and sampling is easy because no special preparation is required. As it is a scattering technique, samples are simply placed in the laser beam and the backscattered radiation is analyzed.

Glass gives only a weak Raman signal, so that FT-Raman spectra from liquids are recorded by using a capillary cell [34], a sapphire sphere [34], or glass cuvettes with the back surface silvered for the favored 180° backscattering arrangement (see Section 2.2). Moreover, the study of aqueous samples in glass or quartz vessels, which is difficult in the mid-infrared region owing to strong infrared absorption bands of water and quartz, is possible in Raman spectroscopy because there are no significant bands causing interference in the solution spectrum. Powders are either measured by using microspherical cells [33] or are filled into cells consisting of a brass cylinder with a 3 mm hole drilled through, into which a reflective metal rod is placed to compress the powder [14]. The Raman intensity is related to the sample thickness, but once the thickness exceeds 2 mm the intensity does not increase. Optical fibers are used for remote sampling, and for FT-Raman spectroscopy this is especially advantageous as the transmission by the fibers, excited by a near-infrared Nd : YAG laser [33], [135], can be at its highest just in the range of the Raman spectrum. Guidelines for optimizing FT-Raman recording conditions are given in [34]. Conventional dispersive Raman spectroscopy has not been widely used for quantitative measurements, mainly because of difficulties with band reproducibility. However, by using an FT-Raman spectrometer, spectra with excellent reproducibility, both in band position and intensity, are obtained, so that this technique has great potential for quantitative analysis [136–138].

Along with general reviews [139–142] and surveys of major areas of Raman spectroscopy [16], [17], various reviews covering special applications of Raman spectroscopy, such as polarization measurements [143], polymers [144], proteins and nucleic acids [145], pharmaceuticals and biomaterials [146–150], paint systems [151], [152], inorganic and organometallic materials [153], and thin-layer chromatography [154], have been published. For further applications and special Raman techniques, such as resonance Raman (RR), surface-enhanced Raman spectroscopy (SERS), coherent anti-Stokes Raman scattering (CARS), and stimulated Raman scattering (SRS), see [14], [155].

5.3. Fourier Transform Infrared and Raman Microspectroscopy

In conventional infrared microsampling, beam size is reduced four to six times by using beam condensers. These accessories, which permit submicrogram samples of between 2 and 0.5 mm in diameter to be measured, have, however, drawbacks such as heating effects, since the entire power of the infrared beam is condensed on the sample, and the fact that the sample cannot be visibly located, so that a skilled operator is needed for this time-consuming work. These disadvantages are largely overcome by coupling an optical microscope to an FT-IR spectrometer. As infrared lenses are poor in performance, infrared microscopes use reflecting optics such as Cassegrain objectives instead of the glass or

quartz lenses employed in conventional optical microscopes. Figure 55 shows the beam path of a typical infrared microscope with both transmission and reflection modes. The infrared beam from the FT-IR instrument is focussed onto the sample placed in a standard microscope, and the beam that passes through the sample is collected by a Cassegrain objective with 15- or 36-fold magnification that produces an image of the sample within the barrel of the microscope. The radiation passes through the variable aperture placed in the image plane and is focussed by another Cassegrain objective on a small area ($250 \times 250\,\mu m$ or less) of the specially matched MCT detector. As the visible optical train is collinear with the infrared light path, it is possible to position the sample and to isolate and aperture the area for analysis visually. This is the main advantage of an FT-IR microscope over a beam condenser, and it enables the measurement of, for example, microcontaminants that cannot be removed from the sample, or the individual layers of a polymer laminate. FT-IR microscopy requires virtually no sample preparation. A further advantage of the technique is its unsurpassed sensitivity, which permits samples in the picogram range to be analyzed, with the diffraction limit of infrared radiation (10–20 µm) as the limiting condition for the size of measurement spot. To minimize stray light due to diffraction being collected by the microscope, particularly when the size of the area investigated is approximately equal to the wavelength employed, a second aperture is introduced into the optics that focus the radiation onto the sample [156].

Uses of FT-IR microspectroscopy include general characterization of particulate matter, dichroic measurements with polarized light, polymer characterization, semiconductor measurements, the identification of contaminants, as well as forensic, biological, and pharmaceutical applications [157–159].

By using diamond cells with the beam condenser or microscope, the thickness of a sample (e.g., paint chips) can be adjusted by squeezing. An alternative to this is an ultrasmall sample analyzer [160], which allows nondestructive internal reflectance studies of microgram and nanogram samples to be performed.

In principle, Raman spectroscopy is a microtechnique [161] since, for a given light flux of a laser source, the flux of Raman radiation is inversely proportional to the diameter of the laser-beam focus at the sample, i.e., an optimized Raman sample is a microsample. However, Raman microspectroscopy able to obtain spatially resolved vibrational spectra to ca. 1 µm spatial resolution and using a conventional optical microscope system has only recently been more widely appreciated. For Raman microspectroscopy both conventional [162] and FT-Raman spectrometers [163], [164] are employed, the latter being coupled by near-infrared fiber optics to the microscope.

Applications of Raman microspectroscopy include the analysis of a wide variety of organic and inorganic materials, e.g., semiconductors, polymers, single fibers, molecular crystals, and minerals [165–168].

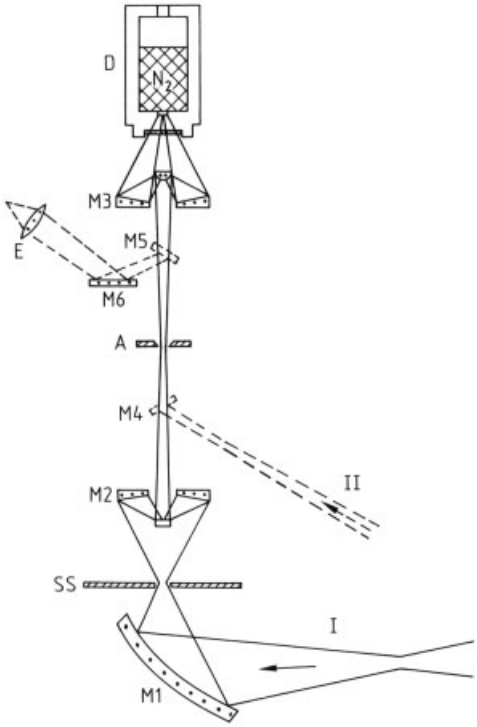

Figure 55. Beam path of a typical infrared microscope
I = Infrared transmittance beam; II = Infrared reflectance beam; M1 = Condensing mirror; M2 and M3 = Cassegrain objectives; M4 = Movable, semi-transparent mirror; M5 = Movable mirror; M6 = Mirror; SS = sample stage; A = Aperture; E = Eyepiece; D = MCT-Detector, liquid N_2 cooled
(Reproduced by permission of Bruker Optik GmbH, D-76275 Ettlingen, Germany)

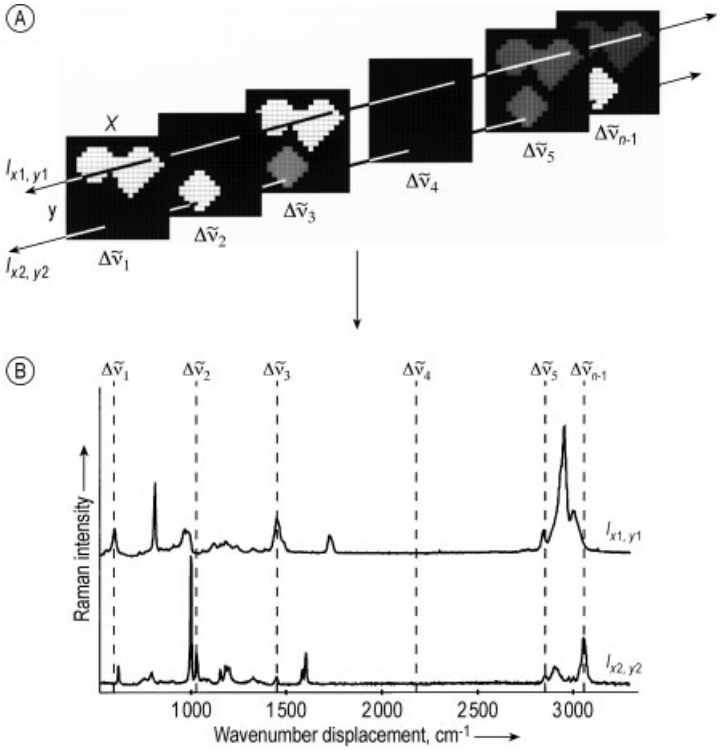

Figure 56. Conceptualization of a three-dimensional hyperspectral image data cube
(A) Images as a function of wavenumber; B) Selected spectra associated with pixels (i, j) in the image
(Reproduced by permission of Marcel Dekker, Inc., 270 Madison Avenue, New York, NY 10016-0602)

5.4. Time-Resolved FT-IR and FT-Raman Spectroscopy

Time-resolved FT-IR and FT-Raman techniques allow studies of molecular reaction mechanisms with time resolutions up to nanoseconds [169], which is a prerequisite for the understanding of biological processes at the molecular level. The major problem in measuring biological reactions consists in selecting small absorption bands of the molecular groups undergoing reactions from the large background absorption of water and the entire biological molecule. This difficulty is met by obtaining difference spectra of two reaction states of the molecules under investigation. The intrinsic advantages of FT spectroscopy, i.e., the multiplex, the throughput, and the accuracy (see Section 2.1) enable such small absorption changes to be reliably detected. With operation of the FT spectrometer in the step-scan mode [169], a time resolution of a few nanoseconds is achieved. Biological applications of time-resolved FT-IR and FT-Raman spectroscopy are described in [169].

5.5. Vibrational Spectroscopic Imaging

Although vibrational spectroscopic imaging methods are relative newcomers to the panoply of vibrational techniques, they inherit the versatility of the traditional single-point infrared and Raman approaches and, in combination with digital imaging technology, allow for new perspectives in the interpretation of the spectra. Imaging advantages are derived, in part, from the ability of an investigator to process and discern effectively a complex two-dimensional spatial representation of a sample in terms of a variety of spectrally related molecular parameters. In particular, a single spectroscopic image, or chemical map, is capable of summarizing and conveying a wealth of spatial, chemical, and molecular data of benefit to both specialists and nonspe-

cialists. Regardless of the specific method used to collect data during a vibrational spectroscopic imaging experiment, the final representation is typically expressed as a three-dimensional image cube, or hypercube, consisting of two spatial dimensions and one spectral dimension. The image cube concept, shown in Figure 56, is interpreted in one of two ways, either (1) a series of images collected as a function of wavenumber (Figure 56A), or (2) a vibrational spectrum corresponding to each spatial location, or pixel, within the image plane (Figure 56B). The ability to merge and recall, within a single analytical technique, information corresponding to the spatial and spectral axes of the image cube provides extraordinary flexibility in probing and extracting structural and compositional information from samples. Thus, a data set may be either summarized as a single image or expressed as a series of images derived from one or more spectroscopic parameters, such as spectral peak intensity, band frequency, or linewidth. The chemical and morphological interpretations of the resulting images are then based on the spatial variations of the intrinsic molecular property being highlighted and their relevance to the biological questions being pursued. In contrast to images generated by conventional point mapping, which generally suffer from low spatial resolution due to relatively few sampling points, direct imaging approaches using the advantages inherent in two-dimensional array detectors derive an enormous benefit from the highly parallel nature of the data collection. That is, a single data set which typically contains many tens of thousands of independent spectral channels may be analyzed and interpreted by using a variety of statistical algorithms, allowing subtle spatial and spectral variations that are often overlooked or misinterpreted in traditional spectroscopic studies to be tested and revealed.

Vibrational spectroscopic imaging has been widely applied to diverse materials, including polymers, semiconductors, pharmaceuticals, cosmetics, consumer products, and biological materials. The strengths and adaptability of vibrational spectroscopic imaging rest not only on the ability to determine chemical compositions and component distribution within a sample, but also on being able to extract localized molecular information relevant to the sample architecture. For example, vibrational spectra are sensitive to alterations in protein secondary and tertiary structures. Spectral shifts could therefore be used to image specifically the distributions of specific structural moieties within a biological sample as opposed to measuring simply an overall distribution of a general class of molecules. Multivariate approaches may also be used to generate a composite metric indicative of either disease progression or biological function. Although these metrics can not be readily interpreted in terms of biochemical changes, they may often be statistically correlated with disease. In addition, by simultaneously recording data on a two-dimensional focal-plane array detector from a myriad of spatial locations within a biological matrix, pooled spectra may be treated statistically. In this manner, analysis of variance can be employed to more robustly test the significance of observed spectral changes appearing, for example, because of a perturbed or diseased state.

Raman and infrared imaging techniques as well as their biological applications are described in detail in [170].

5.6. Infrared and Raman Spectroscopy of Polymers (see also → Plastics Analysis, Chap. 4.)

As polymer analysis is a discipline of its own beyond the scope of this article, the reader is referred to the comprehensive works by Hummel [171], Koenig [172], and Zerbi et al. [173].

6. Near-Infrared Spectroscopy

6.1. Comparison of Mid-Infrared and Near-Infrared Spectroscopy

Near-infrared spectroscopy as a valuable tool for qualitative and quantitative analysis [174], [175] has experienced a revival through the development of chemometric methods and the introduction of fiber optics during the 1980s, and high-quality near-infrared spectra are provided by improved conventional and by Fourier transform instruments.

Various aspects differentiate a mid-infrared spectrum from the corresponding near-infrared

spectrum (Table 6). In the mid-infrared region fundamental molecular vibrations always occur between 4000 and 200 cm^{-1}, while in the near-infrared region overtones and combination bands of these fundamentals are observed between 12 800 and 4000 cm^{-1}. Near-infrared bands are due primarily to hydrogenic stretches of C−H, N−H, and O−H bonds whose fundamental vibrations give rise to strong bands between 4000 and 2000 cm^{-1}, so that their overtones and combinations occur above 4000 cm^{-1} as the most intense absorption bands in the near-infrared spectrum. Since the near-infrared absorptions of polyatomic molecules thus mainly reflect vibrational contributions from very few functional groups, near-infrared spectroscopy is less suitable for detailed qualitative analysis than mid-infrared spectroscopy, which shows all (active) fundamentals and the overtones and combinations of low-energy vibrations. The near-infrared overtone and combination bands depend more on their environment than does the fundamental of the same vibration; slight perturbation in the bonding scheme causes only small changes in the fundamental but significant frequency shifts and amplitude changes in the near-infrared. In going from the fundamental to the first overtone the intensity of an absorption band decreases by a factor of about 10–100, so that the sensitivity of near-infrared spectroscopy is reduced in comparison with the mid-infrared. This is a disadvantage when gases are to be measured, but for liquids, it is a considerable advantage as regards sample handling because cells with convenient path lengths between 1 mm and 10 cm can be used. Glass or quartz is used as window material. Near-infrared bands of liquids and solids have relatively large bandwidths between 30 and 60 cm^{-1}, so that they strongly overlap and direct assignment of bands is generally not possible for larger, complex molecules. This is why, in the near-infrared, easy quantitative analysis using isolated bands, as in the mid-infrared, is not possible. Chemometrics [37], [176–178], however, by the application of mathematical and statistical procedures, generates correlations between experimental data and the chemical composition or physical properties of the sample, and can be used in a general manner to solve quantitative and qualitative analytical problems.

Table 6. Comparison of mid-infrared and near-infrared spectroscopy

Mid-Infrared	Near-Infrared
4000 – 200 cm^{-1}	12 800 – 4000 cm^{-1}
Fundamentals	overtones and combinations
High intensity	low intensity
High sensitivity (trace analysis possible)	low sensitivity (not suited for trace analysis)
Sharp, separated bands	strongly overlapping bands
Easy quantitation with isolated bands	quantitation complex, chemometrics necessary
FT-instruments	grating, filter, FT and acousto-optical tunable scanning instruments

6.2. Applications of Near-Infrared Spectroscopy

Applications of near-infrared spectroscopy include chemistry [179], [180], the oil industry [181], clinical analysis [182], biological and medical analysis [183], and the pharmaceutical industry [184], [185].

Since the intensities of characteristic near-infrared bands are independent of or only slightly dependent on the state of the system, near-infrared spectroscopy is widely applicable for the quantitative study of liquid and compressed gaseous systems, including fluid systems, up to high pressures and temperatures.

The application of near-infrared data to routine quantitative analysis was initiated by Norris [186], who used diffuse reflectance measurements to quantitatively determine major components, such as moisture, protein and fat, in agricultural commodities. In comparison with mid-infrared, near-infrared diffuse reflectance [187] is able to measure powdered samples with minimal sample preparation, and lends itself extremely well to quantitative analysis because the smaller absorptivities and larger scattering coefficients at shorter wavelengths result in a larger scattering/absorption ratio. Although near-infrared reflectance analysis was initially developed for agricultural and food applications [188], in combination with chemometrics it is now applied to many other areas, e.g., polymers [189], pharmaceuticals [190], organic substances [191], geological samples [192], and thin-layer chromatography [193], [194].

The excellent transmittance of quartz in the near-infrared range has led to a further enhancement of the potential of near-infrared spectroscopy by the introduction of fiber optics. Fiber-optic waveguides are used to transfer light from the spectrometer to the sample and, after transmission or reflection, back to the spectrometer. Most fiber optic cables consist of three concentric components: the inner part, through which the light propagates, is called the core, the middle layer is the cladding, and the outer protective layer is the jacket. Crucial to the light throughput of the fiber optic cable are the optical properties of the core and cladding. The difference in refractive index of the two materials defines the highest angle at which the core/cladding interface exhibits total internal reflection, and hence allows the light to propagate through the cable. This angle is related to the numerical aperture (N.A.) of the fiber by the following equation:

$$\sin\alpha = \text{N.A.} = \left(n_{\text{core}}^2 - n_{\text{cladding}}^2\right)^{1/2} \tag{20}$$

where α is the half-angle of acceptance and n is the refractive index of the material. The greater the difference in refractive index the greater the half-angle of acceptance. Any light entering at an angle greater than the half-angle of acceptance is lost to the cladding through absorption. The core and cladding are usually of quartz (of different refractive index), while the outer protective layer is a polymer; other fiber optic materials are zirconium fluoride (ZrF) and chalcogenide (AsGeSe) [195]. The obvious advantage of the optical fiber technique is that the location of measurement can be separated from the spectrometer by between two and several hundred meters, depending on the length of the waveguide. Since almost all practically relevant substances show characteristic near-infrared absorption bands, quantitative analysis by near-infrared spectroscopy is generally applicable to on-line concentration measurements in connection with chemical reactions, chemical equilibria, and phase equilibria. Various probes can be integrated into different reactors or bypasses and offer numerous remote-sensing and on-line process control applications [196]. In the analysis of hazardous or toxic materials this has proved of outstanding importance. In the refining, petrochemical, polymer, plastics, and chemicals industries, typical applications of near-infrared remote spectroscopy are the determination of the octane number and oxygenates in gasoline, aromatics in fuels, the composition of solvent mixtures [197], [198], the hydroxyl number of polyols, low contents of water in solvents, and for polymer analysis.

7. References

1. W. W. Coblentz: *Investigations of Infrared Spectra,* Carnegie Institution, Washington 1905, re-published, The Coblentz Society, Norwalk 1962.
2. G. Herzberg: *Molecular Spectra and Molecular Structure I,* Van Nostrand, Princeton 1953.
3. G. Herzberg: *Molecular Spectra and Molecular Structure II,* Van Nostrand, Princeton 1960.
4. E. B. Wilson, Jr., J. C. Decius, P. C. Cross: *Molecular Vibrations,* McGraw-Hill, New York 1955, re-published Dover, Mineola 1980.
5. D. Steele: *Theory of Vibrational Spectroscopy,* Saunders, Philadelphia 1971.
6. J. A. Koningstein: *Introduction to the Theory of the Raman Effect,* Reidel, Dordrecht 1972.
7. D. A. Long: *Raman Spectroscopy,* McGraw-Hill, New York 1977.
8. P. R. Griffiths, J. A. de Haseth: *Fourier Transform Infrared Spectroscopy,* J. Wiley & Sons, New York 1986.
9. M. W. Mackenzie: *Advances in Applied Fourier Transform Infrared Spectroscopy,* J. Wiley & Sons, New York 1988.
10. N. B. Colthup, L. H. Daly, S. E. Wiberley: *Introduction to Infrared and Raman Spectroscopy,* 3rd ed., Academic Press, San Diego 1990.
11. B. Schrader (ed.): *Infrared and Raman Spectroscopy,* VCH Verlagsgesellschaft, Weinheim 1995.
12. K. W. F. Kohlrausch: *Ramanspektren,* Akademische Verlagsgesellschaft, Leipzig 1943, re-published Heyden & Son, London 1973.
13. H. Baranska, A. Labudzinska, T. Terpinski: *Laser Raman Spectroscopy,* J. Wiley & Sons, New York 1983.
14. P. Hendra, C. Jones, G. Warnes: *Fourier Transform Raman Spectroscopy,* Ellis Horwood, Chichester 1991.

15. J. R. Ferraro, K. Krishnan (eds.): *Practical Fourier Transform Infrared Spectroscopy,* Academic Press, San Diego 1990.
16. M. J. Pelletier (ed.): *Analytical Applications of Raman Spectroscopy,* Blackwell Science, Oxford 1999.
17. J. G. Grasselli, B. J. Bulkin (eds.): *Analytical Raman Spectroscopy,* J. Wiley & Sons, New York 1991.
18. G. Socrates: *Infrared Characteristic Group Frequencies,* J. Wiley & Sons, New York 1980.
19. L. J. Bellamy: *The Infrared Spectra of Complex Molecules,* 3rd ed., vol. 1, Chapman & Hall, London 1975.
20. L. J. Bellamy: *The Infrared Spectra of Complex Molecules,* 2nd ed., vol. 2, Chapman & Hall, London 1980.
21. F. R. Dollish, W. G. Fateley, F. F. Bentley: *Characteristic Raman Frequencies of Organic Compounds,* Wiley Interscience, New York 1974.
22. D. Lin-Vien, N. B. Colthup, W. G. Fateley, J. G. Grasselli: *The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules,* Academic Press, Boston 1991.
23. K. Nakamoto: *Infrared and Raman Spectroscopy of Inorganic and Coordination Compounds,* 5th ed., J. Wiley & Sons, Chichester 1997.
24. H.-U. Gremlich, B. Yan (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000.
25. B. Schrader: *Raman/Infrared Atlas of Organic Compounds,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1989.
26. Sadtler IR Digital Spectra Libraries, Sadtler Division, Bio-Rad Laboratories, Inc., 3316 Spring Garden Street, Philadelphia, PA 19104.
27. F. M. Mirabella (ed.): *Modern Techniques in Applied Molecular Spectroscopy,* J. Wiley & Sons, New York 1998.
28. A. A. Michelson, *Philos. Mag.* **31** (1891) 256 – 259.
29. W. Herres, J. Gronholz, *Comput. Appl. Lab.* **4** (1984) 216 – 220.
30. J. W. Cooley, J. W. Tukey, *Math. Comput.* **19** (1965) 297 – 301.
31. J. Gronholz, W. Herres, *Instrum. Computers* **3** (1985) 10 – 19.
32. B. Schrader, A. Hoffmann, A. Simon, J. Sawatzki, *Vib. Spectrosc.* **1** (1991) 239 – 250.
33. B. Schrader in J. R. Ferraro, K. Krishnan (eds.): *Practical Fourier Transform Infrared Spectroscopy,* Academic Press, San Diego 1990, p. 167.
34. B. Schrader, A. Hoffmann, S. Keller, *Spectrochim. Acta* **47 A** (1991) 1135 – 1148.
35. J. Coates in W. O. George, H. A. Willis (eds.): *Computer Methods in UV, Visible and IR Spectroscopy,* The Royal Society of Chemistry, Cambridge 1990, p. 95.
36. P. S. Wilson in W. O. George, H. A. Willis (eds.): *Computer Methods in UV, Visible and IR Spectroscopy,* The Royal Society of Chemistry, Cambridge 1990, p. 143.
37. F. Despagne, D. L. Massart, *Analyst* **123** (1998) 157R – 178R.
38. G. M. Barrow: *Introduction to Molecular Spectroscopy,* McGraw-Hill, New York 1962.
39. S. R. La Paglia: *Introductory Quantum Chemistry,* Harper & Row, New York 1971.
40. *Ullmann,* 4th ed., **5,** 303.
41. Available from Sadtler Division, Bio-Rad Laboratories, Inc., 3316 Spring Garden Street, Philadelphia, PA 19104.
42. Available from LabControl GmbH, D-50858 Köln 40, Max-Planck-Str. 17 a, Germany.
43. Available from Galactic Industries Corporation, 395 Main Street, Salem NH 03079.
44. J. P. Coates, R. W. Hannah in T. Theophanides (ed.): *Fourier Transform Infrared Spectroscopy,* D. Reidel Publ., Dordrecht 1984, p. 167.
45. S. R. Lowry, D. A. Hupplar, *Anal. Chem.* **53** (1981) 889 – 893.
46. J. T. Clerc in W. O. George, H. A. Willis (eds.): *Computer Methods in UV, Visible and IR Spectroscopy,* The Royal Society of Chemistry, Cambridge 1990, p. 13.
47. D. A. Cirovic, *Trends Anal. Chem.* **16** (1997) 148 – 155.
48. H. Yang, P. R. Griffiths, *Anal. Chem.* **71** (1999) 751 – 761.
49. J. U. White, *J. Opt. Soc. Am.* **32** (1942) 285 – 288.
50. R. T. Graf, J. L. Koenig, H. Ishida, *Appl. Spectrosc.* **39** (1985) 405 – 412.
51. R. G. Greenler, *J. Chem. Phys.* **44** (1966) 310 – 315.
52. R. G. Greenler, *J. Chem. Phys.* **50** (1969) 1963 – 1968.
53. D. L. Allara, J. D. Swalen, *J. Phys. Chem.* **86** (1982) 2700 – 2704.
54. N. J. Harrick, *Appl. Opt.* **10** (1971) 2344 – 2349.
55. I. Newton: *Opticks,* Dover, New York 1952.

56. N. J. Harrick: *Internal Reflection Spectroscopy,* Harrick Scientific Corporation, Ossining, New York 1979.
57. J. Fahrenfort in A. Mangini (ed.): *Adv. Mol. Spectrosc. Proc. Int. Meet. 4th 1959,* **vol. 2,** 1962, p. 701.
58. Available from Harrick Scientific Corporation, 88 Broadway, Ossining, NY 10562.
59. N. J. Harrick, M. Milosevic, S. L. Berets, *Appl. Spectrosc.* **45** (1991) 944 – 948.
60. S. Compton, P. Stout, FTS 7/IR Notes no. 1, Bio-Rad, Digilab Division, Cambridge, MA 02139, 1991.
61. U. P. Fringeli in F. M. Mirabella (ed.): *Internal Reflection Spectroscopy: Theory and Application,* Marcel Dekker, New York 1992, p. 255.
62. U. P. Fringeli, *Chimia* **46** (1992) 200 – 214.
63. Available from Spectra-Tech Inc., Stamford, CT 06906.
64. Available from GRASEBY SPECAC Ltd., Lagoon Road, St. Mary Cray, Orpington, Kent BR5 3QX, U.K.
65. W. M. Doyle, *Appl. Spectrosc.* **44** (1990) 50 – 69.
66. B. E. Miller, N. D. Danielson, J. E. Katon, *Appl. Spectrosc.* **42** (1988) 401 – 405.
67. Available from Axiom Analytical, Inc., 17751-C Sky Park Circle, Irvine, CA 92614.
68. W. M. Doyle, *Process Control Quality* **2** (1992) 11 –41.
69. N. L. Owen, D. W. Thomas, *Appl. Spectrosc.* **43** (1989) 451 – 455.
70. P. Kubelka, F. Munk, *Z. Tech. Phys.* **12** (1931) 593 –601.
71. P. Kortüm, W. Braun, G. Herzog, *Angew. Chem., Int. Ed. Engl.* **2** (1963) 333 – 341.
72. P. H. Turner, W. Herres, BRUKER FT-IR Application Note 23, Bruker Analytische Meßtechnik GmbH, D-76153 Karlsruhe, Germany.
73. D. Reinecke, A. Jansen, F. Fister, U. Schernau, *Anal. Chem.* **60** (1988) 1221 – 1224.
74. M. P. Fuller, P. R. Griffiths, *Anal. Chem.* **50** (1978) 1906 – 1910.
75. M. Fuller in *Nicolet FT-IR Spectral Lines,* Nicolet Analytical Instruments, Madison, WI 53711, Spring/Summer 1990, p. 15.
76. G. Glauninger, K.-A. Kovar, V. Hoffmann, *Fresenius J. Anal. Chem.* **338** (1990) 710 – 716.
77. H. Yamamoto, K. Wada, T. Tajima, K. Ichimura, *Appl. Spectrosc.* **45** (1991) 253 – 259.
78. S. A. Johnson, R. M. Rinkus, T. C. Diebold, V. A. Maroni, *Appl. Spectrosc.* **42** (1988) 1369 – 1375.
79. K. A. Martin, R. F. Zabransky, *Appl. Spectrosc.* **45** (1991) 68 – 72.
80. Si-Carb Sampler, available from Spectra-Tech Inc., Stamford, CT 06906.
81. F. J. DeBlase, S. Compton, *Appl. Spectrosc.* **45** (1991) 611 – 618.
82. A. M. Vassallo, P. A. Cole-Clarke, L. S. K. Pang, A. J. Palmisano, *Appl. Spectrosc.* **46** (1992) 73 – 78.
83. M. Handke, N. J. Harrick, *Appl. Spectrosc.* **40** (1986) 401 – 405.
84. R. T. Rewick, R. G. Messerschmidt, *Appl. Spectrosc.* **45** (1991) 297 – 301.
85. L. T. Lin, D. D. Archibald, D. E. Honigs, *Appl. Spectrosc.* **42** (1988) 477 – 483.
86. A. Tsuge, Y. Uwamino, T. Ishizuka, *Appl. Spectrosc.* **43** (1989) 1145 – 1149.
87. R. W. Jones, J. F. McClelland, *Anal. Chem.* **61** (1989) 650 – 656.
88. R. W. Jones, J. F. McClelland, *Anal. Chem.* **61** (1989) 1810 – 1815.
89. R. J. Pell, B. C. Erickson, R. W. Hannah, J. B. Callis, B. R. Kowalski, *Anal. Chem.* **60** (1988) 2824 –2827.
90. J. A. McGuire, B. Wangmaneerat, T. M. Niemczyk, D. M. Haaland, *Appl. Spectrosc.* **46** (1992) 178 – 180.
91. B. Wangmaneerat et al., *Appl. Spectrosc.* **46** (1992) 340 – 344.
92. A. Rosencwaig: *Photoacoustics and Photoacoustic Spectroscopy,* J. Wiley & Sons, New York 1980.
93. J. F. McClelland, *Anal. Chem.* **55** (1983) 89 A – 105 A.
94. J. A. Graham, W. M. Grim III, W. G. Fateley in J. R. Ferraro, L. J. Basile (eds.): *Fourier Transform Infrared Spectroscopy,* **vol. 4,** Academic Press, Orlando 1985, p. 345.
95. H. Coufal, J. F. McClelland, *J. Mol. Struct.* **173** (1988) 129 – 140.
96. G. Laufer, J. T. Huneke, B. S. H. Royce, Y. C. Teng, *Appl. Phys. Lett.* **37** (1980) 517 – 519.
97. G. Busse, B. Bullemer, *Infrared Phys.* **18** (1978) 631 –634.
98. MTEC Photoacoustics, Inc., Ames, IA 50010.
99. R. O. Carter III, S. L. Wright, *Appl. Spectrosc.* **45** (1991) 1101 – 1103.
100. C. Q. Yang, *Appl. Spectrosc.* **45** (1991) 102 – 108.
101. J. C. Donini, K. H. Michaelian, *Infrared Phys.* **24** (1984) 157 – 163.
102. M. W. Urban, J. L. Koenig, *Appl. Spectrosc.* **40** (1986) 994 – 998.

103. C. Q. Yang, R. R. Bresee, W. G. Fateley, *Appl. Spectrosc.* **41** (1987) 889 – 896.
104. R. Rubinovitz, J. Seebode, A. Simon, *Bruker Report* **1** (1988) 35 – 38.
105. M. J. Smith, C. J. Manning, R. A. Palmer, J. L. Chao, *Appl. Spectrosc.* **42** (1988) 546 – 555.
106. R. A. Crocombe, S. V. Compton, FTS/IR Notes no. 82, Bio-Rad, Digilab Division, Cambridge, MA 02139, 1991.
107. K. H. Michaelian, *Appl. Spectrosc.* **45** (1991) 302 –304.
108. L. Bertrand, *Appl. Spectrosc.* **42** (1988) 134 – 138.
109. R. White: "Chromatography/Fourier Transform Infrared Spectroscopy and Its Applications," *Practical Spectroscopy Series,* vol. 10, Marcel Dekker, New York 1990.
110. W. Herres: *HRGC-FTIR: Capillary Gas Chromatography-Fourier Transform Infrared Spectroscopy,* Hüthig Verlag, Heidelberg 1987.
111. S. Bourne et al., FTS/IR Notes no. 75, Bio-Rad, Digilab Division, Cambridge, MA 02139, 1990.
112. R. L. White, G. N. Giss, G. M. Brissey, C. L. Wilkins, *Anal. Chem.* **53** (1981) 1778 – 1782.
113. P. Coffey, D. R. Mattson, J. C. Wright, *Am. Lab.* **10** (1978) 126 – 132.
114. K. S. Chiu, K. Biemann, K. Krishnan, S. L. Hill, *Anal. Chem.* **56** (1984) 1610 – 1615.
115. Available from Lab Connections, Inc., 201 Forest Street, Marlborough, MA 01752.
116. Available from Bourne Scientific, Inc., 65 Bridge Street, Lexington, MA 02421-7927.
117. S. Bourne, *Am. Lab.* **16** (1998) 17F – 17J.
118. C. M. Earnest, *Anal. Chem.* **56** (1984) 1471 A –1486 A.
119. D. Dollimore, *Anal. Chem.* **64** (1992) 147 R – 153 R.
120. J. A. J. Jansen, J. H. van der Maas, A. Posthuma de Boer, *Appl. Spectrosc.* **46** (1992) 88 – 92.
121. K. Ichimura, H. Ohta, T. Tajima, T. Okino, *Mikrochim. Acta* **1** (1988) 157 – 161.
122. R. C. Wieboldt, S. R. Lowry, R. J. Rosenthal, *Mikrochim. Acta* **1** (1988) 179 – 182.
123. J. Mullens et al., *Bull. Soc. Chim. Belg.* **101** (1992) 267 – 277.
124. R. M. Carangelo, P. R. Solomon, D. J. Gerson, *Fuel* **66** (1987) 960 – 967.
125. J. K. Whelan, *Energy Fuels* **2** (1988) 65 – 73.
126. L. A. Sanchez, M. Y. Keating, *Appl. Spectrosc.* **42** (1988) 1253 – 1258.
127. M. D. Timken, J. K. Chen, T. B. Brill, *Appl. Spectrosc.* **44** (1990) 701 – 706.
128. E. Post, S. Rahner, H. Möhler, A. Rager, *Thermochim. Acta* **263** (1995) 1 – 6.
129. M. Maciejewski, F. Eigenmann, A. Baiker, E. Dreser, A. Rager, *Thermochim. Acta* (2000) in press.
130. L. A. Nafie in M. W. Mackenzie (ed.): *Advances in Applied FTIR Spectroscopy,* John Wiley and Sons, New York 1988, p. 67.
131. L. A. Nafie, T. B. Freedman in H.-U. Gremlich, B. Yang (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000, p. 15.
132. T. A. Keiderling in H.-U. Gremlich, B. Yang (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000, p. 55.
133. B. Schrader, S. Keller, *Proc. SPIE Int. Soc. Opt. Eng. 1575 (Int. Conf. Fourier Transform Spectrosc., 8th, 1991)* (1992) 30 – 39.
134. N. J. Everall, J. Lumsdon, *Spectrosc. Eur.* **4** (1992) 10 – 21.
135. C. G. Zimba, J. F. Rabolt, *Appl. Spectrosc.* **45** (1991) 162 – 165.
136. T. Jawhari, P. J. Hedra, H. A. Willis, M. Judkins, *Spectrochim. Acta* **46 A** (1990) 161 – 170.
137. F. T. Walder, M. J. Smith, *Spectrochim. Acta* **47 A** (1991) 1201 – 1216.
138. H. Sadeghi-Jorabchi, R. H. Wilson, P. S. Belton, J. D. Edwards-Webb, D. T. Coxon, *Spectrochim. Acta* **47 A** (1991) 1449 – 1458.
139. L. A Lyon et al., *Anal. Chem.* **70** (1998) 341R – 361R.
140. B. Chase in A. M. C. Davies, C. S. Creaser (eds.): *Anal. Appl. Spectrosc. 2,* The Royal Society of Chemistry, Cambridge 1991, p. 13.
141. K. P. J. Williams, S. M. Mason, *TRAC* **9** (1990) 119 –127.
142. P. J. Hendra, *J. Mol. Struct.* **266** (1992) 97 –114.
143. A. Hoffmann et al., *J. Raman Spectrosc.* **22** (1991) 497 – 503.
144. P. J. Hendra, C. H. Jones, *Makromol. Chem. Macromol. Symp.* **52** (1991) 41 – 56.
145. G. J. Thomas, *Ann. Rev. Biophys. Biomol. Struct.* **28** (1999) 1 – 27.
146. A. M. Tudor et al., *J. Pharm. Biomed. Anal.* **8** (1990) 717 – 720.
147. P. R. Carey, *J. Biol. Chem.* **274** (1999) 26 625 – 26 628.
148. J. Góral, V. Zichy, *Spectrochim. Acta* **46 A** (1990) 253 – 275.
149. I. W. Levin, E. N. Lewis, *Anal. Chem.* **62** (1990) 1101 A – 1111 A.

150. S. Nie et al., *Spectroscopy* **5** (1990) 24 – 32.
151. G. Ellis, M. Claybourn, S. E. Richards, *Spectrochim. Acta* **46 A** (1990) 227 – 241.
152. D. Bourgeois, S. P. Church, *Spectrochim. Acta* **46 A** (1990) 295 – 301.
153. T. N. Day, P. J. Hendra, A. J. Rest, A. J. Rowlands, *Spectrochim. Acta* **47 A** (1991) 1251 – 1262.
154. N. J. Everall, J. M. Chalmers, I. D. Newton, *Appl. Spectrosc.* **46** (1992) 597 – 601.
155. D. L. Gerrard, J. Birnie, *Anal. Chem.* **64** (1992) 502 R – 513 R.
156. B. Yan, *Acc. Chem. Res.* **31** (1998) 621 – 630.
157. R. G. Messerschmidt, M. A. Harthcock (eds.): *Infrared Microspectroscopy: Theory and Applications,* Marcel Dekker, New York 1988.
158. K. Krishnan, S. L. Hill in J. R. Ferraro, K. Krishnan (eds.): *Practical Fourier Transform Spectroscopy,* Academic Press, San Diego 1990, p. 103.
159. S. Compton, J. Powell, FTS 7/IR Notes no. 2, Bio-Rad, Digilab Division, Cambridge, MA 02139, 1991.
160. N. J. Harrick, M. Milosevic, S. L. Berets, *Appl. Spectrosc.* **45** (1991) 944 – 948.
161. B. Schrader, *Fresenius J. Anal. Chem.* **337** (1990) 824 – 829.
162. J. D. Louden in D. J. Gardiner, P. R. Graves (eds.): *Practical Raman Spectroscopy,* Springer Verlag, Berlin 1989, p. 119.
163. J. Sawatzki, *Fresenius J. Anal. Chem.* **339** (1991) 267 –270.
164. A. Hoffmann, B. Schrader, R. Podschadlowski, A. Simon, *Proc. SPIE-Int. Soc. Opt. Eng. 1145 (Int. Conf. Fourier Transform Spectrosc., 7th, 1989)* (1989) 372 – 373.
165. F. Bergin, *Proc. Annu. Conf. Microbeam Anal. Soc.* **25** (1990) 235 – 239.
166. B. W. Cook in A. M. C. Davies, C. S. Creaser (eds.): *Anal. Appl. Spectrosc.* **2,** The Royal Society of Chemistry, Cambridge 1991, p. 61.
167. A. J. Sommer, J. E. Katon, *Appl. Spectrosc.* **45** (1991) 527 – 534.
168. D. Gernet, W. Kiefer, H. Schmidt, *Appl. Spectrosc.* **46** (1992) 571 – 576.
169. K. Gerwert in H.-U. Gremlich, B. Yang (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000, p. 193.
170. M. D. Schaeberle, I. W. Levin, E. N. Lewis in H.-U. Gremlich, B. Yang (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000, p. 231.
171. D. O. Hummel: *Atlas of Polymer and Plastics Analysis,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1991.
172. J. L. Koenig: *Spectroscopy of Polmers,* American Chemical Society, Washington, D.C. 1992.
173. G. Zerbi et al.: *Modern Polymer Spectroscopy,* Wiley-VCH, Weinheim 1998.
174. J. Workman, Jr., *J. Near Infrared Spectrosc.* **1** (1993) 221.
175. I. Murray, I. A. Cowe (eds.): *Making Light Work: Advances in Near Infrared Spectroscopy,* VCH Verlagsgesellschaft, Weinheim 1992.
176. P. Geladi, E. Dabbak, *J. Near Infrared Spectrosc.* **3** (1995) 119.
177. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbecke: *Handbook of Chemometrics and Qualimetrics, Part A,* Elsevier, Amsterdam 1997.
178. D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbecke: *Handbook of Chemometrics and Qualimetrics, Part B,* Elsevier, Amsterdam 1998.
179. S. M. Donahue, C. W. Brown, B. Caputo, M. D. Modell, *Anal. Chem.* **60** (1988) 1873 – 1878.
180. J. Lin, C. W. Brown, *Vibr. Spectrosc.* **7** (1994) 117 – 123.
181. S. R. Westbrook, *SAE Tech. Paper Ser.* **930734** (1993) 1 – 10.
182. H. M. Heise in H.-U. Gremlich, B. Yan (eds.): *Infrared and Raman Spectroscopy of Biological Materials,* Marcel Dekker, New York 2000, p. 259.
183. R. J. Dempsey, D. G. Davis, R. G. Buice, Jr., R. A. Lodder, *Appl. Spectrosc.* **50** (1996) 18A – 34A.
184. J. D. Kirsch, J. K. Drennen, *Appl. Spectrosc. Rev.* **30** (1995) 139 – 174.
185. W. Plugge, C. van der Vlies, *J. Pharm. Biomed. Anal.* **14** (1996) 891 – 898.
186. I. Ben-Gera, K. H. Norris, *Israel J. Agric. Res.* **18** (1968) 125 – 132.
187. D. L. Wetzel, *Anal. Chem.* **55** (1983) 1165 A – 1176 A.
188. P. Williams, K. Norris (eds.): *Near-Infrared Technology in the Agricultural and Food Industries,* AACC, St. Paul 1987.
189. C. E. Miller, *Appl. Spectrosc. Rev.* **26** (1991) 277 –339.
190. B. F. MacDonald, K. A. Prebble, *J. Pharm. Biomed. Anal.* **11** (1993) 1077 – 1085.
191. L. G. Weyer, *Appl. Spectrosc. Rev.* **21** (1985) 1 – 43.

192. D. E. Honigs, T. B. Hirschfeld, G. M. Hieftje, *Appl. Spectrosc.* **39** (1985) 1062 – 1065.
193. E. W. Ciurczak, L. J. Cline-Love, D. M. Mustillo, *Spectrosc. Int.* **3** (1991) no. 5, 39 – 42.
194. E. W. Ciurczak, W. R. Murphy, D. M. Mustillo, *Spectrosc. Int.* **3** (1991) no. 7, 39 – 44.
195. M. J. Smith, T. E. May, *Int. Lab.* **22** (1992) 18 – 24.
196. L. P. McDermott, *Ad. Instrum. Control.* **45** (1990) 669 – 677.
197. Guided Wave, Inc., Application Note no. A3-987, El Dorado Hills, CA 95630.
198. Guided Wave, Inc., Application Note no. A4-188, El Dorado Hills, CA 95630.

Ink-Jet Printing → Imaging Technology
Inks → Drawing and Writing Materials
Inorganic Fibers → Fibers, 5. Synthetic Inorganic

Nuclear Magnetic Resonance and Electron Spin Resonance Spectroscopy

Reinhard Meusinger, Institute of Organic Chemsitry, University of Mainz, Mainz, Federal Republic of Germany

A. Margaret Chippendale, ZENECA Specialties, Manchester, United Kingdom

Shirley A. Fairhurst, AFRC Institute for Plant Science Research, Brighton, United Kingdom

1. Introduction 2
2. Principles of Magnetic Resonance . 3
2.1. Nuclear and Electronic Properties . 3
2.2. Nuclei and Electrons in a Stationary Magnetic Field 3
2.3. Basic Principles of the NMR and ESR Experiments 4
2.4. Relaxation 6
3. High-Resolution Solution NMR Spectroscopy 6
3.1. The NMR Experiment 6
3.1.1. Continuous Wave Methodology 8
3.1.2. Fourier Transform Methodology . . . 9
3.2. Spectral Parameters 10
3.2.1. Chemical Shift 11
3.2.2. Spin – Spin Coupling 11
3.2.3. Signal Intensity 13
3.2.4. Relaxation Times 13
3.3. NMR and Structure 14
3.3.1. Hydrogen (^{1}H and ^{2}H) 14
3.3.2. Carbon (^{13}C) 18
3.3.3. Fluorine (^{19}F) 19
3.3.4. Phosphorus (^{31}P) 20
3.3.5. Nitrogen (^{14}N and ^{15}N) 21
3.3.6. Oxygen (^{17}O) 22
3.3.7. Silicon (^{29}Si) 22
3.4. Double Resonance Techniques . . . 24
3.4.1. Homonuclear Spin Decoupling 24
3.4.2. Heteronuclear Spin Decoupling 24
3.4.3. NOE Difference Spectroscopy 25
3.5. One-Dimensional Multi-Pulse FT Experiments 26
3.5.1. T_1 Measurement 27
3.5.2. T_2 Measurement 28
3.5.3. Spectral Editing Experiments 28
3.6. Multi-Dimensional NMR 28
3.6.1. Basic Principles 29
3.6.2. J-Resolved Spectra 29
3.6.3. Homonuclear Chemical Shift Correlation (COSY) 30
3.6.4. Heteronuclear Chemical Shift Correlation (HETCOR, HMQC) 31
3.6.5. Homonuclear NOE Correlation 32
3.7. NMR Spectral Collections, Databases, and Expert Systems . . . 33
3.8. Applications 35
3.8.1. Chemical Structure Determination . . 35
3.8.2. Quantitative Chemical Analysis by NMR 36
3.8.3. Rate Processes and NMR Spectra . . . 37
3.8.4. NMR Methods Utilized in Combinatorial Chemistry and Biochemistry . . . 38
4. NMR of Solids and Heterogeneous Systems 39
4.1. High-Resolution NMR of Solids . . 39
4.2. Low Resolution ^{1}H NMR of Heterogeneous Systems 40
5. NMR Imaging 40
6. ESR Spectroscopy 41
6.1. The ESR Experiment 41
6.1.1. Continuous Wave ESR 41
6.1.2. ENDOR and Triple Resonance 41
6.1.3. Pulse ESR 43
6.1.4. ESR Imaging 43
6.2. Spectral Parameters 44
6.2.1. g-Factor 44
6.2.2. Nuclear Hyperfine Interaction 45
6.2.3. Quantitative Measurements 45
6.3. ESR in the Liquid State 46
6.3.1. Slow Molecular Tumbling 46
6.3.2. Exchange Processes 47
6.4. Computer Simulation of Spectra . . 47
6.5. Specialist Techniques 48
6.5.1. Spin Trapping 48
6.5.2. Spin Labeling 48
6.5.3. Oximetry 49
6.5.4. Saturation Transfer 50
7. References 50

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_471

Nuclear Magnetic Resonance and Electron Spin Resonance

Symbols

a	ESR isotropic hyperfine coupling constant; ESR total hyperfine interaction
AT	acquisition time
B	magnetic field
DW	dwell time
g	electronic or nuclear g factor
h	Planck constant
I	nuclear spin quantum number
J	spin – spin coupling constant
k	Boltzmann constant
K	intensity of ESR response
M	magnetization vector
m_I, m_S	magnetic quantum numbers
N	Avogadro number
s	shielding constant
S	electronic spin quantum number
SW	spectrum width
T	ESR anisotropic hyperfine coupling
T_1	spin – lattice relaxation time
T_2	spin – spin relaxation time
T_d	relaxation delay
β	electronic Bohr or nuclear magneton
γ	magnetogyric ratio
δ	chemical shift
μ	magnetic dipole moment
τ	rotational correlation time, pulse width
ν	resonance frequency
ω_0	Larmor frequency

Units

ppm	parts per million (unit of the dimensionless δ scale of the chemical shift)
T	tesla (unit of magnetic field strength, $1\,\mathrm{T} = 10^4$ gauss)

1. Introduction

It is now about 50 years since the phenomena of electron spin resonance (ESR) and nuclear magnetic resonance (NMR) were first discovered. During the last thirty years NMR spectroscopy has developed into perhaps the most important instrumental measuring technique within chemistry. This is due to a dramatic increase in both the sensitivity and the resolution of NMR and ESR instruments. NMR and ESR spectroscopy are used today within practically all branches of chemistry and in related sciences, such as physics, biology, and medicine, both at universities and in industrial laboratories. NMR has its most important applications in the determination of molecular structures. It can today be applied to a wide variety of chemical systems, from small molecules to proteins and nucleic acids. ESR has played a key role in the study of free radicals, many transition metal compounds, and other species containing unpaired electrons.

The continuing development of each technique has resulted from a series of step changes. Both methods were first developed using continuous wave methodology. By the late 1960s the pulsed Fourier transform techniques were developed and additionally the NMR spectroscopists began to use magnets based on superconducting material. On grounds of these developments the quality of spectra, expressed both in terms of sensitivity and resolution, improved quickly during the 1970s. The increase in sensitivity has made it possible to study small amounts of material as well as chemically interesting isotopes of low natural occurrence, e.g., carbon-13 (^{13}C NMR), routinely. Pulsed ESR techniques have since followed.

Since the 1970s two-dimensional NMR methods and more recently three- and multidimensional varieties have allowed ever more complex molecules to be studied. In the last two decades an additional increase of sensitivity was obtained by inverse detection techniques simultaneously combined with a drastic shortening of measuring time by gradient enhanced spectroscopy. In the area of solid states NMR techniques have been developed that obtain solution-like spectra from solids. Superconducting magnet systems for NMR frequencies up to 900 MHz (21.13 T) allow now the elucidation of the structure of large proteins and make NMR spectroscopy invaluable for the development of new drugs.

Since the mid-1970s NMR and ESR imaging has also been developed. The construction of whole-body magnets with magnetic field strength of up to 8.0 T and with magnet bores of up to 90 cm has paved the way for investigations on humans and NMR imaging today is an indispensable aid for medical diagnosis.

2. Principles of Magnetic Resonance [1–6]

Many of the principles of magnetic resonance are common to both ESR and NMR spectroscopy. To exhibit magnetic resonance the system concerned must have a magnetic moment μ. For NMR this is provided by a nucleus with nonzero nuclear spin, and for ESR by an unpaired electron in an atom, radical, or compound. The two phenomena differ in the signs and magnitudes of the magnetic interactions involved. When matter is placed in a strong magnetic field, these systems (nuclei or electrons) behave like microscopic compass needles. According to the laws of quantum mechanics these tiny compass needles can orient themselves with respect to the magnetic field only in a few ways. These orientations are characterized by different energy levels. If the sample is exposed to electromagnetic waves of an exactly specified frequency the nuclear spins or the unpaired electrons can be forced to jump between different energy levels. In magnetic fields with a field strength of a few teslas, the resonance frequencies for the NMR and ESR are in the radio and microwave frequency regions, respectively. During the experiment the frequency is varied. When it exactly matches the characteristic frequency of the nuclei or the electrons an electric signal is induced in the detector. The strength of the signal is plotted as a function of this resonance frequency in a diagram called the NMR or ESR spectrum, respectively.

2.1. Nuclear and Electronic Properties

The nuclear spin I of an atomic nucleus has values of 0, 1/2, 1, 3/2, etc. in units of $h/2\pi$, where h is the Planck constant. The value for a given nucleus is dependent on the mass and atomic numbers of the nucleus (Table 1). Thus some important, common nuclei such as ^{12}C and ^{16}O which have both even mass and atomic numbers have zero spin. An electron has a spin S of 1/2. Nuclear (μ_{N}) and electronic (μ_{e}) magnetic moments are proportional to the magnitude of the spin:

$$\mu_{\mathrm{N}} = \gamma_{\mathrm{N}} \frac{hI}{2\pi}$$

where γ_{N} is the magnetogyric ratio of the nucleus. In SI units μ_{N} is in amperes meter squared and γ_{N} in radians per tesla per second. Alternatively the magnetic moment can be expressed as:

$$\mu_{\mathrm{N}} = g_{\mathrm{N}} \beta_{\mathrm{N}} I$$

where g_{N} is a dimensionless constant (nuclear g factor) and β_{N} the nuclear magneton:

$$\beta_{\mathrm{N}} = \frac{eh}{4\pi Mc}$$

where e and M are the charge and mass of the proton, and c the velocity of light. The quantities g_{N} and I distinguish nuclei from one another.

Table 1. Nuclear properties

Mass number	Atomic number	Nuclear spin I
Odd	even or odd	1/2, 3/2, 5/2 . . .
Even	even	0
Even	odd	1, 2, 3, . . .

The magnetic moment of an electron is given by an analogous expression:

$$\mu_{\mathrm{e}} = -g\beta S$$

where g is the dimensionless electron g factor, and β the electronic Bohr magneton:

$$\beta = -\frac{eh}{4\pi mc}$$

where $-e$ and m are the charge and mass of the electron. Because the electron is negatively charged, its moment is negative.

2.2. Nuclei and Electrons in a Stationary Magnetic Field

When a nucleus with nonzero spin or an electron is placed in a strong magnetic field B the orientation of the spin becomes quantized, with each possible orientation having a different energy. A nucleus of spin I has $2I+1$ possible orientations, with associated magnetic quantum numbers m_I of $-I, -I+1, \ldots I-1, I$. Hydrogen nuclei (protons) have a spin quantum number $I = 1/2$. Consequently the proton can exist in only two spin states, which are characterized by their magnetic quantum numbers $m_I = 1/2$ and

$m_I = -1/2$. The proton can therefore be pictured as a magnetic dipole, the z component μ_z of which can have a parallel or antiparallel orientation with respect to the z direction of the coordinate system (Fig. 1). In the following discussion the denotations α and β are used for these two orientation states, where α denotes the lower, more stable energy level and β the higher one. For a single electron with $S = 1/2$, allowed values for the magnetic quantum number m_S are also + 1/2 and − 1/2.

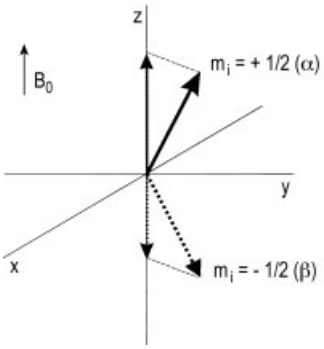

Figure 1. Orientation of the nuclear magnetic moments in the magnetic field B_0.

In the absence of a magnetic field the α and β states have the same energy. Only in a static magnetic field B_0 is this degeneracy destroyed as a result of the interaction of the nuclear or electronic moment μ with B_0 and an energy difference for the two orientation states results. The energy of interaction is proportional to the nuclear or electronic moment and the applied field. For a nucleus and electron, respectively, this can be written as:

$$E_{\mathrm{I}} = -g_{\mathrm{N}}\beta_{\mathrm{N}} m_I B_0 = \frac{-\gamma_{\mathrm{N}} h m_I B_0}{2\pi}$$

$$E_{\mathrm{S}} = g\beta m_S B_0$$

Energy level diagrams for nuclei with spins of 1/2 and 3/2 are given in Figure 2.

The energy levels diagram for an electron looks exactly like that of the nucleus in Figure 2 A), but with exchanged signs of the magnetic quantum numbers m_{S}. The selection rule for NMR and ESR allows transitions for which m_{I} and m_{S} change by one unit. Thus:

$$\Delta E_{\mathrm{I}} = g_{\mathrm{N}}\beta_{\mathrm{N}} B_0 = \frac{\gamma_{\mathrm{N}} h B_0}{2\pi}$$

$$\Delta E_{\mathrm{S}} = g\beta B_0$$

2.3. Basic Principles of the NMR and ESR Experiments

To observe transitions, radiation given by $\Delta E = h\nu$ must be applied. The fundamental resonance conditions for NMR and ESR experiments respectively are:

$$h\nu = g_{\mathrm{N}}\beta_{\mathrm{N}} B_0 = \frac{\gamma_{\mathrm{N}} h B_0}{2\pi} \tag{1}$$

and

$$h\nu = g\beta B_0$$

where ν is the resonance frequency. For a proton in a magnetic field of 2.349 T, ν is 100 MHz. The resonance frequency of a free electron in the same field is 65.827 MHz.

Resonance absorption can only be detected if the spin levels involved differ in population. For a very large number N of protons in thermal equilibrium, the distribution of the spin 1/2 nuclei between the ground (α) state and the excited (β) state is given by the Boltzmann relation:

$$\frac{N_\beta}{N_\alpha} = \exp\left(-h\nu/kT\right)$$

The excess of spins in the lower state is only ca. 1 in 10^5 for NMR at a magnetic field strength of 9 T and at room temperature. This is the fundamental reason for the low sensitivity of NMR compared with IR and UV spectroscopy and even ESR.

The magnetic resonance experiment itself is best explained and visualized in classical as opposed to quantum mechanical terms. How is the energy transferred to nuclei or electrons aligned in a magnetic field, and how is the energy absorbed measured? The following description is for nuclei of spin 1/2 but is equally applicable to electrons. In a magnetic field B_0, applied in the z-direction, the nuclei have an orientation which is either aligned or opposed to the direction of B_0. Additionally, the nuclei precess randomly around the z-axis with a frequency ω_0, known as the Larmor frequency, where $\omega_0 = 2\pi\nu$. A macroscopic magnetization vector M along the

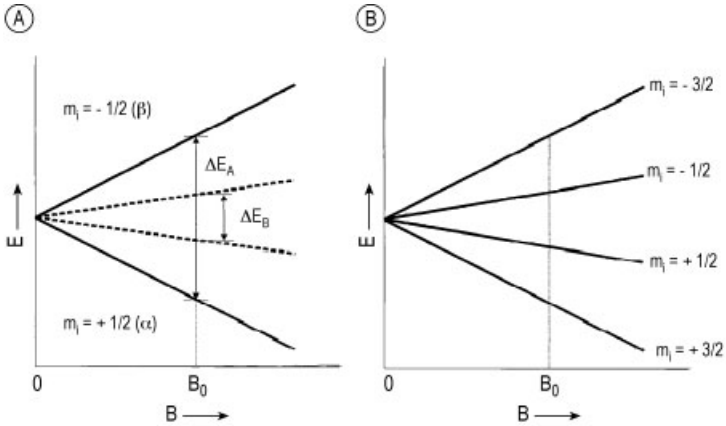

Figure 2. Energy levels for a nucleus with a spin of A) $I = 1/2$, and B) $I = 3/2$. On the left-hand side the levels for two different isotopes A and B with different magnetogyric ratios are represented ($\gamma_A > \gamma_B$). For the nuclei of the isotope B an obviously smaller energy difference ΔE results at the same magnetic field strength B_0. Only the value and not the direction of the magnetic field B_0 is plotted on the abscissas.

z-axis but not in the x-y plane results from the individual nuclear magnetic moments μ_z defined in Figure 1. The slight excess of nuclei aligned with the magnetic field as shown above, gives rise to the alignment of this magnetization vector M_z in direction of B_0 (Fig. 3).

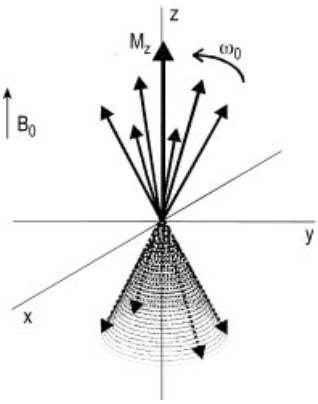

Figure 3. Motion of spin 1/2 nuclei in a magnetic field B_0. Note that more nuclei are arranged parallel than antiparallel to the direction of the static magnetic field B_0.

The quantum description of the lower and higher energy levels can be equated with the classical description of precessing nuclei aligned with and against the applied field. Up to here the description of the nuclear motions was referred to in the stationary or laboratory frame of reference. If the laboratory could be rotated at the Larmor frequency ω_0, the nuclei would no longer appear to precess but would become stationary and coincident with the magnetic field axis B_0 as shown in Figure 4; x' and y' are the rotating x- and y-axes. The magnetic behavior is now completely described by a magnetization vector M_z acting along B_0. Use of the rotating frame system greatly simplifies the description of pulsed NMR experiments (see Section 3.1.2).

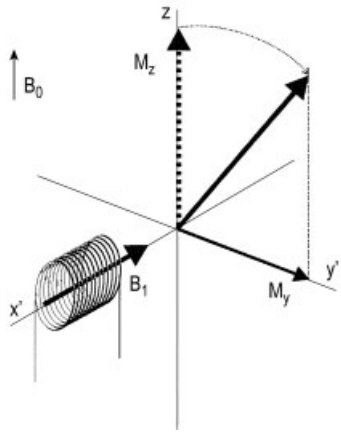

Figure 4. Motion of spin 1/2 nuclei in the rotating frame and generation of transverse magnetization M_y by rotation of the vector of the macroscopic magnetization M_z. The B_1 field caused by the transmitter coil along the x-axis deflects M_z into the $x'-y'$-plane.

In the magnetic resonance experiment the longitudinal net magnetization M_z is tipped toward the $x-y$ plane. Radio-frequency electro-

magnetic energy is applied such that its magnetic component B_1 is at right angles to B_0 and is rotating with the precessing nuclei. The RF irradiation is applied along the x-axis and, at the resonance frequency, tilts the net magnetization towards the $x-y$ plane. This creates an x-y component, the so-called transverse magnetization, which induces a signal in the detector coil. Conventionally the receiver is located on the y-axis of the Cartesian coordinates (Fig. 4).

To obtain a spectrum, either the RF frequency or the magnetic field B_0 can be scanned over a narrow range. Historically, this mode, known as continuous wave (CW), was employed in both NMR and ESR instruments. For NMR it has been almost completely superceded by the more sensitive and sophisticated pulsed Fourier transform (FT) method. Pulsed methods are now becoming more widely used in ESR.

2.4. Relaxation

In the absence of any mechanism enabling the nuclear or electronic spins to return to the ground state, all the excess population in the lower energy state will eventually be raised to the higher energy state, and no more energy will be absorbed. However, such mechanisms do exist, the so-called relaxation phenomena. That is, the magnitudes of both longitudinal and transversal magnetization are time dependent . One is the spin – lattice or longitudinal relaxation process, which occurs by loss of energy from excited nuclear spins to the surrounding molecules. By this process the component of magnetization along the z-axis relaxes back to its original value M_z, with an exponential decay characterized by a time constant T_1, the spin – latice relaxation time.

After application of the RF energy the magnetization has components in the $x-y$ plane, and these also decay at least as fast as the spin – lattice relaxation returns the magnetization to the z-axis. However, $x-y$ magnetization can also be lost by other processes, which cause the $x-y$ magnetism to dephase to give a net x – y magnetization of zero. The rate constant in the $x-y$ plane for this process is T_2, the spin – spin or transverse relaxation time. Such relaxation involves transfer of energy from one nucleus to another. An important feature of spin – spin relaxation is that it determines the natural linewidths of the signals in a spectrum, which is defined in terms of the half-height linewidth Δ. In hertz one has:

$$\Delta = \frac{1}{\pi T_2}$$

if T_2 is measured in seconds. However, the observed linewidth is always larger because of a contribution from the magnetic field inhomogeneity.

3. High-Resolution Solution NMR Spectroscopy [7–24]

Knowing the gyromagnetic ratio γ of a nucleus enables the resonance frequency from Equation (1) to be calculated for any magnetic field strength. Table 2 lists useful, commonly studied nuclei and their magnetic properties. The frequencies are for a 2.35 T applied magnetic field, i.e., relative to ^{1}H at 100 MHz. Table 2 gives the relative sensitivity for equal numbers of each nucleus and their relative receptivities, i.e., the product of the relative sensitivity and natural abundance.

The abundant nuclei, ^{1}H, ^{19}F, and ^{31}P, with $I = 1/2$ were the first to be studied by CW NMR. Owing to sensitivity problems, less abundant isotopes of important spin 1/2 nuclei such as ^{13}C and ^{15}N require FT techniques for observation in natural abundance. Nuclei with spin $I \geq 1$ have a quadrupole moment in addition to their magnetic moment. The interaction of the nuclear quadrupole moment with the electric field gradient at the nucleus provides a very efficient process for nuclear relaxation. Often, the quadrupolar mechanism is dominant, the resulting relaxation times are very short, and the signals broad. In extreme cases (e.g., covalently bonded ^{35}Cl or ^{37}Cl), the signals are too broad to detect. However, nuclei with very small quadrupole moments such as ^{2}H and ^{11}B can be observed in the usual manner as the broadening is much less severe.

3.1. The NMR Experiment

The first successful NMR experiments were reported in 1945 by two independent groups

Table 2. Magnetic properties of some nuclei

Isotope	Spin	Natural abundance, %	NMR frequency,∗ MHz	Relative sensitivity	Relative receptivity
^{1}H	1/2	99.985	100.000	1.00	1.00
^{2}H	1	0.015	15.351	9.65×10^{-3}	1.45×10^{-6}
^{11}B	3/2	80.42	32.084	0.17	0.13
^{13}C	1/2	1.108	25.144	1.59×10^{-2}	1.76×10^{-4}
^{14}N	1	99.63	7.224	1.01×10^{-3}	1.01×10^{-3}
^{15}N	1/2	0.37	10.133	1.04×10^{-3}	3.85×10^{-6}
^{17}O	5/2	0.037	13.557	2.91×10^{-2}	1.08×10^{-5}
^{19}F	1/2	100	94.077	0.83	0.83
^{23}Na	3/2	100	26.451	9.25×10^{-2}	9.25×10^{-2}
^{29}Si	1/2	4.70	19.865	7.84×10^{-3}	3.69×10^{-4}
^{31}P	1/2	100	40.481	6.63×10^{-2}	6.63×10^{-2}

∗ At $B = 2.35$ T.

in the USA [25], [26]. The discoveries by Bloch at Stanford and by Purcell at Harvard were awarded a Nobel Prize in Physics in 1952. Already in September 1952 the first commercial NMR spectrometer was installed. Early instruments used permanent magnets or electromagnets with fields from 0.94 up to 2.35 T, corresponding to 40 up to 100 MHz for proton resonance. Permanent magnets give very stable magnetic fields and are often used without further stabilization. Electromagnets need additional control mechanisms to provide the necessary stability. The commonest way to do this is to use the NMR signal of another nucleus in the sample to provide a field-frequency locking mechanism. The other nucleus may be the reference nucleus (homo-lock) or a different nucleus (hetero-lock), e.g., the deuterium resonance of a deuterated solvent. The need for instruments with higher resolution and sensitivity has led to the development of commercially available systems operating at up to 900 MHz for protons. All instruments operating above 100 MHz use helium-cooled superconducting solenoids to provide the magnetic field. High-resolution NMR requires both the magnetic field and the RF source to be homogeneous and stable to better than 1 part in 10^8. To achieve this performance, which allows to obtain the maximum information possible from an NMR experiment, the magnetic field profile has to be made extremely uniform in the region of the sample. The two parameters which quantify these performance characteristics are magnetic field "homogeneity" and field "drift". The magnets are designed to minimize any magnetic field variations in time. Inevitable imperfections are introduced during manufacture of the wire of the solenoid and also caused by different samples, temperature variations, etc. They are eliminated using correction coils and so-called shim coils to remove residual field gradients. Also the magnetic field stability is of major importance. In conventional magnets the superconducting wire is bathed in liquid helium under atmospheric pressure where it has a boiling point of 4.2 K. By reducing the pressure the boiling point of helium can be lowered to 2.3 K. By doing this the performance of the superconducting wire is enhanced and higher magnetic fields can be achieved (cryostabilization). In Figure 5 a schematic diagram of a superconducting magnet is shown.

The sample, normally a solution in a deuterated solvent in a 5 or 10 mm glass tube (smaller tubes have been employed for specific applications for some years as well), is placed in the instrument probe between the poles of an electromagnet/permanent magnet or inside the solenoid of a superconducting magnet. The probe contains the RF transmitter and receiver coils and a spinner to spin the tube about its vertical axis. Sample spinning is used to average out magnetic field inhomogeneity across the sample. Multidimensional NMR spectra are measured in modern instruments without rotating.

The ideal NMR solvent should dissolve the sample, be inert, nonpolar, liquid over the temperature range of interest, inexpensive, and not give rise to large interfering peaks in the spectrum being recorded. Deuterated solvents such as deuterochloroform and deuterodimethyl sulf-

oxide are most commonly used. They offer the advantage of being "transparent" in proton spectra and provide a deuterium signal which can be used to stabilize and lock the spectrometer system. A clear, mobile solution is required for a high-quality spectrum. Solid particles or viscous solutions degrade the resolution. Traces of paramagnetic impurities in a solution can dramatically reduce relaxation times and cause peak broadening. With increasing field strength the demands both on the sample and on the used NMR tubes are usually growing.

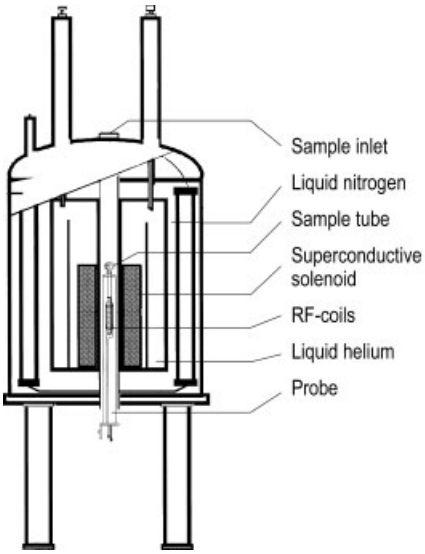

Figure 5. Schematic diagram of a cryomagnet. The superconducting coil is immersed in liquid helium. The sample tube is placed from above into the middle of the solenoid, where the field is strongest and has the best homogeneity. The lines of the static magnetic field are parallel to the long axis of the sample tube.

Spectra are recorded either by the CW scan or pulsed FT method. The amount of sample required depends on the method of detection (CW or FT) and the receptivity of the nucleus being studied. For most routine CW proton spectrometers 10 – 20 mg of compound of M_r 200 – 300, dissolved in 0.4 mL of solvent should give a reasonable spectrum. On a 500 MHz FT instrument it should be possible to obtain a proton spectrum from less than one milligram of such a compound. Even smaller amounts can be measured with specific techniques, e.g., special NMR tube inserts. For specially designed probes amounts of few nanograms are sufficient [114], [144].

To measure nuclei with low receptivity, such as ^{13}C, a more concentrated solution must be used, in order to obtain a spectrum in as short a time as possible. Larger diameter sample tubes (10 or 15 mm) and, consequently, larger sample volumes can be used to increase the amount of sample. Recently the so-called inverse detection techniques were used for observation of nuclei with small sensitivities (Section 3.6.4).

3.1.1. Continuous Wave Methodology

Until the late 1960s all NMR instruments were of the CW type. Figure 6 shows a block diagram of such a spectrometer. To obtain a spectrum from a sample, either the magnetic field or the RF frequency is slowly varied. When the resonance condition (Eq. 1) is met for the nuclei being observed, the sample absorbs RF energy and the resulting signal is detected by the receiver coils, amplified, and recorded. Spectra are recorded on precalibrated charts relative to a reference compound.

The CW method is suitable for recording spectra of abundant nuclei from relatively large amounts of sample. As a result, early NMR work was essentially limited to the ^{1}H, ^{19}F, and ^{31}P nuclei. Nowadays the use of CW instruments, which are relatively inexpensive, is mainly limited to ^{1}H observation. Such instrumentation is usually to be found in university teaching or synthetic organic chemical laboratories rather than specialist NMR facilities.

One of the main limitations of NMR spectroscopy is its inherent lack of sensitivity. As a result, the CW method is unsuitable for recording spectra from nuclei of low natural abundance or even from abundant nuclei in solutions of very low concentration. One way of improving the signal-to-noise (S/N) ratio is to record several spectra from a sample and add them together. The noise averages out, and the S/N ratio increases by a factor equal to the square root of the number of spectra accumulated. Using a computer to store and add the spectra allows many thousands of spectra to be accumulated in this way. The main drawback to this method on a CW instrument is the time to acquire each spectrum, typically ca. 4 min, resulting in excessive and often prohibitive total experiment times. A technique was required which could simultane-

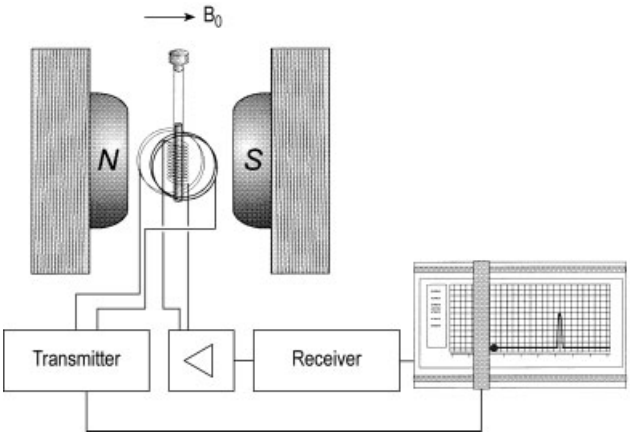

Figure 6. Block diagram of a CW NMR spectrometer.

ously excite the whole spectral region of interest. The pulsed FT technique, which first became available on commercial instruments in the late 1960s, provided the answer.

3.1.2. Fourier Transform Methodology

In 1966 Ernst and Anderson discovered that the sensitivity of NMR spectra could be increased dramatically if the sample was not exposed to the slow radio frequency sweep but to short and intense radio frequency pulses. The intensity of the signal was measured as a function of time after this pulse. The next pulse and signal acquisition were started after a few seconds, and the signals after each pulse were summed in a computer. If the RF frequency ν is turned on and off very rapidly to obtain a pulse τ seconds long, this is equivalent to irradiating the sample with a range of frequencies centered about ν over a frequency range $\nu \pm 1/\tau$. Thus if $\tau = 50\,\mu$s a range of $\nu \pm 20\,000$ Hz is covered.

If an RF pulse, at the resonant frequency ν_0, is applied along the x-axis in the laboratory frame, this is equivalent to applying a static field $\boldsymbol{B}_1$ along the x'-axis of the rotating frame. This drives the sample magnetization about the x'-axis by an angle θ, as shown in Figure 4. Spectrometers are normally designed to detect signals along the y'-axis. The component of magnetization along this direction is given by $\boldsymbol{M}_0 \sin\theta$, and the maximum signal is obtained when $\theta = 90°$. A pulse producing this effect is known as a 90° pulse, and the time for which the pulse must be applied to achieve this is known as the 90° pulse time. On modern FT spectrometers 90° pulse times are usually significantly less than 50 μs. Since the pulse applied is very short it must be powerful enough to excite nuclei in the time available across the whole spectral region of interest. To do this it must satisfy the condition:

$$\gamma \boldsymbol{B}_1 \geq 2\pi\,(\mathrm{SW})$$

where SW is the spectral width required.

At the end of the RF pulse, the spin system begins to relax back towards its equilibrium condition by means of the two relaxation mechanisms described in Section 2.4. After a time $5\,T_1$ the magnetization along y' – the signal detection axis – has decayed essentially to zero. For the remainder of this article the rotating frame is assumed and the axes are referred to simply as x, y, and z.

In the CW experiment, desrcibed in Section 3.1.1, the intensity of the NMR signal is recorded as function of frequency. However, in the pulsed experiment, the intensity of the decaying signal, following the RF pulse, is detected and recorded as a function of time. The signal is known as the free induction decay (FID) and it is not amenable to a simple interpretation. In Figure 7 a pulse experiment is represented as diagram in the usual notation. After the delay period T_d the excitation occurs through a pulse with the pulse width (length) τ and the subsequent acquisition. The resultant FID is also illustrated in Figure 7 together with the conventional NMR spectrum, for

a sample which gives a spectrum with a single peak. For a spectrum containing more than one peak, the FID is an interference pattern resulting from the frequencies corresponding to all the individual peaks in the spectrum. The FID is often referred to as the time domain signal, whereas the corresponding NMR spectrum provides a frequency domain signal.

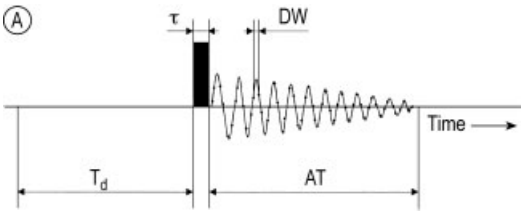

Figure 7. Schematic presentation of a pulse or FT NMR experiment (A); the free induction decay (FID) (B); and the appendant Fourier-transformed signal (NMR spectrum) (C). In A the FID is shown in a simplified manner also in the digital form, which will be generated by an analog-to-digital converter (ADC).

Time and frequency domain data can be interconverted by the mathematical process of Fourier transformation (FT), which can be performed rapidly by the computer. The intensity of the FID is measured at *n* intervals equally spaced by the so-called dwell time *DW* and stored in digital form. In other words, the dwell time is the time used to produce a particular data point. The total time to acquire the data is referred to as the acquisition time *AT*. In order to obtain a spectral width SW the following equation must be satisfied:

$$AT = \frac{n}{2\,(SW)} = n \cdot DW$$

Thus, typical values for obtaining a 100 MHz ^{13}C NMR spectrum would be:

SW = 20 000 Hz (equivalent to 200 ppm)
n = 16 384 (stored in 16 K computer words)
DW = 25 μs
AT = 0.41 s

Fourier transformation of the 16 K data points gives a spectrum with 8 K points since the transform contains both real and imaginary components, the real component being identical to the conventional CW spectrum.

The principal advantage of the FT method is that full spectral data can be obtained in a few seconds acquisition time as compared with minutes for the CW method. Signal averaging, by adding together the FIDs obtained after each of a series of RF pulses is therefore a much more efficient process than its CW analogue. The introduction of FT instruments made the routine study of less abundant isotopes, and in particular ^{13}C, possible.

FT instruments are available with magnetic fields from 1.9 to 21.1 T, i.e., 80 to 900 MHz for ^{1}H observation. Most such instruments are available with a range of probeheads for studying different nuclei. Some probeheads are selectively tuned for observation of a single nucleus at maximum sensitivity. Dual- and trial-tuned probeheads for the convenient observation of ^{1}H and other fixed nuclei (e.g., ^{13}C, ^{15}N, ^{31}P) on a single probehead are now widely used, as are tunable probeheads which enable a wide range of nuclei to be studied. However, the most extensive development occurred in the case of the pulse sequences. Several hundred pulse sequences were described in the last twenty years and it is impossible to know all. A very good overview of the 150 most important ones is given in [24]. Furthermore, numerous information about basics and new developments are to be found in the Internet. Some of the most impressive web pages are listed in [115], [116].

The relatively high sensitivity and computing facilities associated with modern FT instruments have led to the development of fully automated systems with robotic sample changers. This enables unattended, round-the-clock use of expensive high-field instrumentation. A typical sample changer can accommodate from 6 up to 100 samples, and data is automatically acquired, processed, and plotted.

3.2. Spectral Parameters

Around 1950, it was discovered that nuclear resonance frequencies depend not only on the nature of the atomic nuclei, but also on their chemical environment. The possibility of using NMR as a tool for chemical analysis soon became apparent. All applications of solution NMR spectroscopy make use of one or more of the spectral parameters chemical shift, spin – spin coupling, signal intensity, and relaxation time. In the following sections examples from ^{1}H and ^{13}C NMR are used but the concepts described apply to magnetic nuclei in general.

3.2.1. Chemical Shift

Equation 1 might suggest that only a single peak would be obtainable from each magnetic isotope. However, when the molecule under observation is placed in the magnetic field, the electrons in the molecule shield the nuclei from the applied magnetic field to a small extent. For a particular nucleus within a molecule the degree of shielding depends on the density of the electrons circulating about that nucleus. Protons in different chemical environments therefore have different shieldings. Figure 8 shows diagramatically what happens to spherically symmetrical *s* electrons around a nucleus under the influence of an external magnetic field B_0. They circulate around the nucleus, producing their own magnetic field which opposes the applied field. The effective field B at the nucleus is then given by

$$B = B_0\,(1-\sigma)$$

Thus, to obtain the resonant condition (Eq. 1) it is necessary to increase the applied field above that for an isolated nucleus. The shielding constant σ is a small fraction, usually of the order of parts per million. For electrons in *p*-orbitals where there is no spherical symmetry the situation is more complex since shielding and deshielding effects depend on the orientation of the molecule with respect to the applied field. Such effects are considered in more detail in Section 3.3.1.

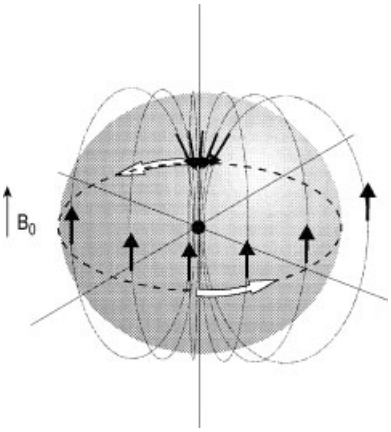

Figure 8. Electronic shielding of a nucleus (•) by circulating *s*-electrons (.....) in a magnetic field B_0

The magnitude of the nuclear shielding is proportional to the applied magnetic field. The difference in the peak position of a particular nucleus from the peak position of a reference nucleus is called the chemical shift δ. In ^{1}H NMR if two peaks are separated by 80 Hz in a magnetic field corresponding to 100 MHz they will be separated by 320 Hz in a field corresponding to 400 MHz. Chemical shifts δ can be expressed in dimensionless units as parts per million (ppm) of the applied frequency in MHz as follows:

$$\delta = \frac{(\nu_{\text{sample}} - \nu_{\text{reference}})}{\text{Applied frequency}}$$

Thus for the example given above

$$\text{At 100 MHz} \quad \delta = \frac{80\,\text{Hz}}{100\times10^6\,\text{Hz}} = 0.8\,\text{ppm}$$

$$\text{At 400 MHz} \quad \delta = \frac{320\,\text{Hz}}{400\times10^6\,\text{Hz}} = 0.8\,\text{ppm}$$

The chemical shift δ is therefore a molecular parameter that does not depend on the spectrometer frequency. ^{1}H and ^{13}C nuclei in organic molecules give peaks which cover a range of approximately 20 ppm and 220 ppm, respectively.

Tetramethylsilane (TMS) is the most commonly used chemical shift reference compound for ^{1}H and ^{13}C NMR. In both types of spectra TMS gives a single sharp resonance at lower frequency than most other proton or carbon resonances. The TMS is usually added to the sample solution as an internal reference. The accepted sign convention is for chemical shifts to high and low frequency of the reference peak to be positive and negative, respectively. Figure 9 shows the 250 MHz ^{1}H spectrum of methyl acetate referenced to TMS; the peaks at 2.0 δ and 3.6 δ arise from the acetyl and methyl protons, respectively.

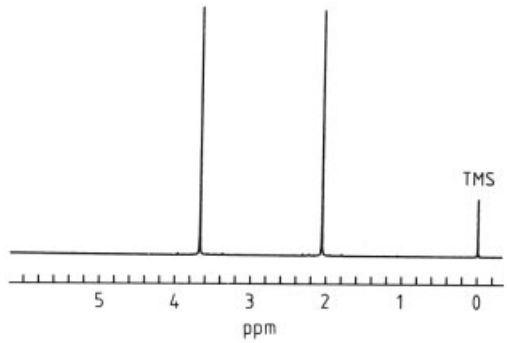

Figure 9. 250 MHz ^{1}H NMR spectrum of methyl acetate

3.2.2. Spin – Spin Coupling

Chemical shift is one source of information on fine structure in NMR spectra. The sec-

ond is spin – spin coupling. Figure 10 shows the aromatic region of the proton spectrum of 3-methoxy-6-nitro-*o*-xylene. One might expect two peaks from the two different aromatic protons in the molecule. In fact four peaks are observed. This results from the indirect coupling of proton spins through the intervening bonding electrons. In Figure 11 this is shown schematically for the two diastereotopic protons in a methylene group. The magnetic B_0 field experienced by nucleus A has a small component arising from nucleus B. Nucleus B has two equally probable orientations in the magnetic field, $m_B = +1/2$ or α and $m_B = -1/2$ or β. As a result nucleus A gives two equally intense peaks corresponding to the two orientations of B. By the same reasoning B itself also gives two equally intense peaks with the same separation as those for A. This splitting is called the spin – spin coupling J, where J_{AB} is the coupling between nuclei A and B. The magnitude of the coupling, which is measured in hertz, is independent of the applied magnetic field. Couplings between protons may be observed from nuclei separated by up to five bonds. Their magnitude rarely exceeds 20 Hz.

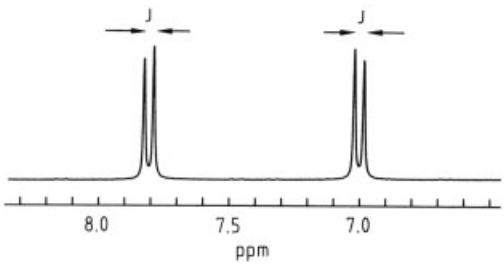

Figure 10. Aromatic region of the 250 MHz ^{1}H NMR spectrum of 3-methoxy-6-nitro-*o*-xylene

If more than two nuclei are interacting with each other all possible orientations of the coupling nuclei must be considered. When three nuclei have different chemical shifts, as for the olefinic protons of vinyl acetate, then each nuclear signal is split by coupling to both of the other nuclei. Thus each resonance can be envisaged as being split into a doublet by the first coupling, each line of which is further split into a doublet to give a quartet pattern for each nucleus. When two of the three nuclei are equivalent, the total number of lines in the spectrum is reduced because some of the couplings are identical. Chemically equivalent protons (e.g., in a methyl group) do not exhibit splitting from coupling among each other.

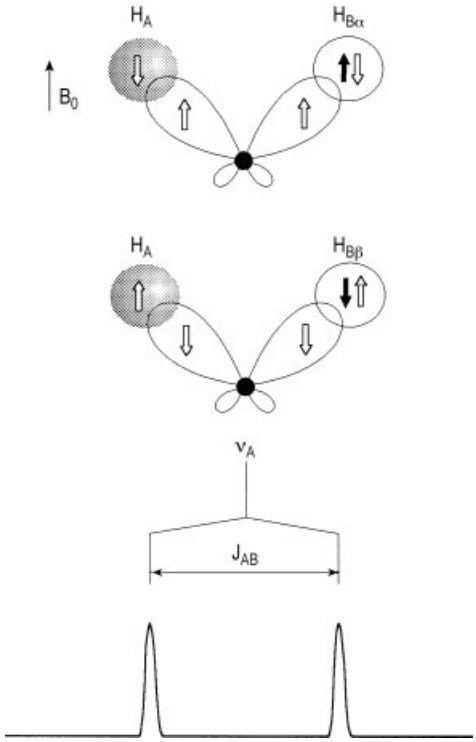

Figure 11. Nuclear spin-spin-interaction through the bonding electrons in a CH_2-group. The low-energy state corresponds to the antiparallel orientation of the magnetic moments of the nuclei (black arrows) and electrons (white arrows). The two possible states of the magnetic moments of hydrogen nucleus B are shown. As a result the resonance signal of nucleus A shows a splitting into a doublet. The effect is only observable in a methylene group in the case of diastereotopic, i.e., chemically inequivalent protons.

Simple splitting patterns, for nuclei of spin 1/2, obey the following general rules:

1) If a nucleus interacts with n other nuclei with a different coupling to each then the signal will consist of 2^n lines of equal intensity
2) If a nucleus A couples to n equivalent nuclei B then the pattern for A comprises $(n+1)$ lines with relative intensities given by the coefficients of the binomial expansion of $(1+x)^n$

These simple rules are only strictly applicable to so- called first-order spectra. For spectra to be first order a single criterion must be obeyed. The chemical shift difference in hertz $\Delta\nu$ for the

coupled nuclei must be much larger than the coupling between them ($\Delta\nu/J \geq 10$). If this condition is not met the spectra may show more lines than the rules predict, and their intensities are distorted, with a build up of intensity towards the middle of the coupling pattern. It is beyond the scope of this article to describe these second-order spectra.

Because the magnitude of the spin – spin coupling interaction is independent of the applied magnetic field, whereas the chemical shift separation is proportional to it, a complex NMR spectrum can be simplified by using higher applied fields.

All the above examples of spin – spin coupling involve homonuclear coupling in proton spectra. In natural abundance ^{13}C spectra homonuclear coupling is not normally observed since the probability of two adjacent ^{13}C atoms is extremely low (10^{-4}). For organic molecules the most commonly encountered couplings are those to protons. One bond couplings range from about 110 to 250 Hz. Most longer range couplings are < 10 Hz. Proton- coupled ^{13}C spectra often show complex overlapping multiplets. To simplify the spectra and obtain them in a shorter time, routine ^{13}C measurements are usually proton noise decoupled (see Section 3.4.2). Thus, in the absence of other coupling nuclei such as ^{31}P or ^{19}F, a single sharp peak is observed for each chemically inequivalent ^{13}C atom. Figure 12 shows the proton-coupled and -decoupled ^{13}C NMR spectra of ethylbenzene.

3.2.3. Signal Intensity [117]

NMR is an inherently quantitative technique. For any nucleus the NMR signal obtained, at least for a single-scan CW experiment or a single-pulse FT experiment, is directly proportional to the number of nuclei producing it. For FT experiments in general to be quantitative the spin system must be given sufficient time to relax completely between successive pulses. Most proton spectra are recorded under essentially quantitative conditions and are integrated to give the relative number of protons contributing to each resonance. Figure 13 shows an integrated proton spectrum of camphor with relative peak areas of approximately 1 : 1 : 1 : 1 : 1 : 2 : 3 : 3 : 3 for the CH-protons, the diastereotopic methylene, and the methyl protons, respectively.

^{13}C spectra and those of other low natural abundance nuclei are not normally recorded under quantitative conditions. For structural identification work with such nuclei, conditions are usually selected for optimum sensitivity, not quantification. There is an additional problem in spectra such as ^{13}C where broadband proton decoupling is usually used. Apart from collapsing the multiplet structure, the decoupling irradiation causes an increase in signal intensity due to nuclear Overhauser enhancement (NOE, see Section 3.4.2). The enhancement factor is often different for the various ^{13}C resonances in the spectrum. Quantitative ^{13}C work is possible but the experimental conditions must be carefully selected [117].

3.2.4. Relaxation Times

Spin – lattice and spin – spin relaxation, the two processes by which a perturbed spin system returns to equilibrium, are outlined in Section 2.4. The various mechanisms that contribute to these processes are discussed in [8]. This section is concerned with the effect variations of T_1 and T_2 have on the appearance of a spectrum. Proton spectra are usually little affected by differences in relaxation times for different protons in a molecule because most protons have short relaxation times (a few seconds). For ^{13}C and many other nuclei this may not be the case. ^{13}C relaxation times in organic materials range from a few milliseconds to > 100 s, depending largely on the molecular mass of the molecule and the environment of a particular ^{13}C atom in the molecule.

For spin 1/2 nuclei the principle T_1 mechanism is often dipole – dipole relaxation, which results from interactions between two nuclei with magnetic spins. Its efficiency is inversely proportional to the sixth power of the internuclear distance. In ^{13}C spectra of organic molecules, protons provide the main source of such relaxation. Where a ^{13}C nucleus is directly bonded to a proton, T_1 is usually quite short, whereas a quaternary carbon atom in the same molecule will probably have a significantly longer value. In a FT ^{13}C spectrum this will result in a relatively low intensity for a quaternary

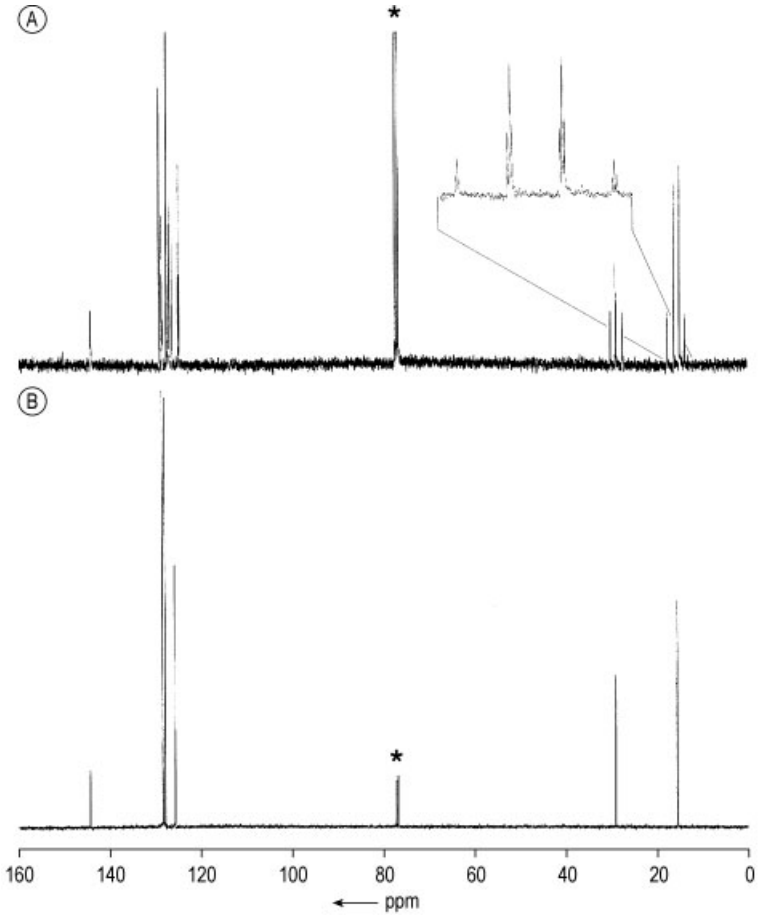

Figure 12. Proton-coupled (A) and proton broadband decoupled (B) ^{13}C NMR spectra of 10 % ethylbenzene in $CDCl_3$ (∗). Applied frequency 100.6 MHz; puls width 13.5 μs; repetition time 5 s; number of accumulations 1024 and 128; measuring time 85 min and 11 min, respectively.

resonance unless the interpulse delay is sufficiently long to allow complete relaxation of all nuclei. Such effects can be used as an aid in spectral assignment.

For polymers and other high molecular mass molecules, T_1 can be very short, in the range 10^{-3} to 1 s. For organic molecules with molecular masses < 1000, values for proton-bonded carbons lie typically in the range 0.1 – 10 s, and for nonproton-bonded carbons, 10 – 300 s.

The most important effect of the spin – spin relaxation time T_2 is that it determines the natural width of the lines in the spectrum. However, for most ^{1}H and ^{13}C spectra the observed linewidth is determined by the homogeneity of the applied magnetic field.

3.3. NMR and Structure

In this section the relationship between chemical structure and the NMR parameters of chemical shift and spin – spin coupling are discussed for some of the more commonly studied nuclei.

3.3.1. Hydrogen (^{1}H and ^{2}H)

The two naturally occurring isotopes of hydrogen are the proton and the deuteron with natural abundances of 99.985 % and 0.015 %, respectively. Whilst the former with $I = 1/2$ is the most commonly studied of all magnetic nuclei,

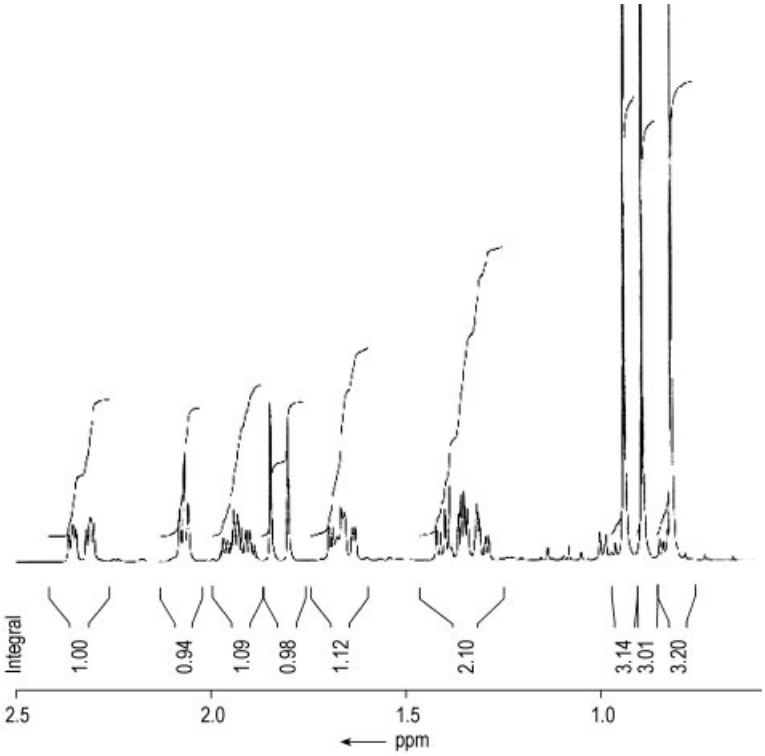

Figure 13. Integrated 400 MHz ^{1}H NMR spectrum of camphor. The reason for the differences between the experimental and the theoretical integral values are impurities and spectrum processing effects.

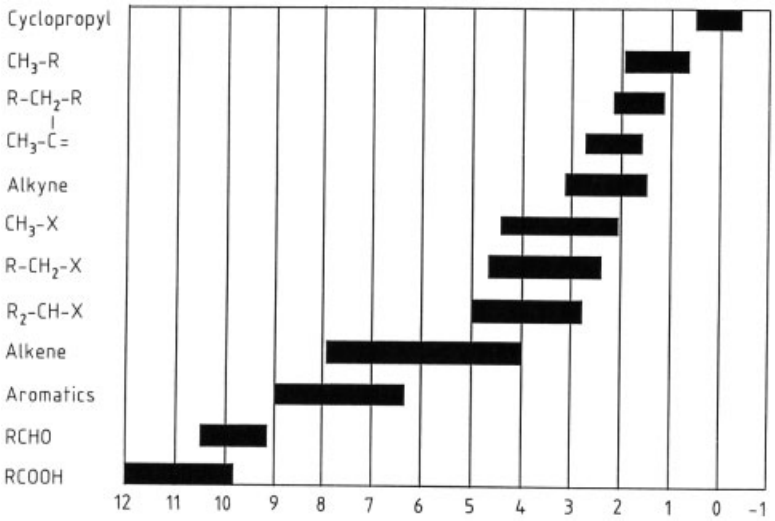

Figure 14. ^{1}H NMR chemical shift ranges
X = Halogen, $-N<$, $-S-$, $-O-$; R=Alkyl

the latter with $I = 1$ is rarely studied in natural abundance and then only for a few specialist applications. One such application, the determination of deuterium levels in samples, is discussed in Section 3.8.2. If small isotope effects are disregarded, corresponding ^{1}H and ^{2}H chemical shifts are the same. ^{2}H spectra are less well dispersed than the corresponding ^{1}H spectra by a factor of 6.51, the ratio of the magnetogyric ratios. The smaller magnetogyric ratio also results in correspondingly smaller coupling constants. Whilst the following discussion is exemplified by reference to the proton, most of the content is equally applicable to the deuteron. However, since deuterium homonuclear coupling constants are very small, ^{2}H spectra with broadband ^{1}H decoupling give a single peak from each chemically inequivalent deuteron even for deuterium enriched samples.

Chemical Shifts. The shielding effect of spherically symmetrical *s* electrons is discussed in Section 3.2.1. This diamagnetic upfield shift affects all nuclei since every molecule has *s* electrons. For electrons in *p*-orbitals there is no spherical symmetry and the phenomenon of diamagnetic anisotropy is used to explain some otherwise anomalous chemical shifts. Diamagnetic anisotropy means that shielding and deshielding depend on the orientation of the molecule with respect to the applied magnetic field.

Table 3. ^{1}H chemical shifts of CH_3X compounds

X	δH, ppm
$SiMe_3$	0.0
H	0.13
Me	0.88
CN	1.97
$COCH_3$	2.08
NH_2	2.36
I	2.16
Br	2.68
Cl	3.05
OH	3.38
F	4.26

The hydrogen atom is a special case since it has no *p* electrons. A consequence of this is that in proton NMR the direct influence of the diamagnetic term can be seen. For example in substituted methanes CH_3X, as X becomes more electronegative, the electron density round the protons decreases and they resonate at lower fields (Table 3). For the same reason acidic protons in carboxylic acid groups resonate at very low fields. Figure 14 shows the chemical shift ranges for a selection of organic groups.

Whilst the electronegativity concept can be used to explain chemical shifts for saturated aliphatic compounds, the effects of diamagnetic anisotropy must also be considered for some other classes of compounds. For example, aromatic compounds such as benzene are strongly deshielded, the protons of acetylene are much more strongly shielded than those of ethylene (acetylene 1.48 ppm compared with ethylene 5.31 ppm), and aldehydic protons resonate at very low fields. In the case of benzene the deshielding results from the so-called ring current effect. When a benzene molecule is orientated perpendicular to the applied magnetic field B_0, there is a molecular circulation of π electrons about the field, as shown in Figure 15. The resulting ring current produces an additional magnetic field which opposes the applied magnetic field along the sixfold axis of the benzene ring but adds to it at the protons on the benzene ring, giving a low-field shift. The ring current is only induced when the applied magnetic field is perpendicular to the benzene ring. In practice only a proportion of the rapidly tumbling molecules are perpendicular to the magnetic field, but the overall averaged shift is affected by them.

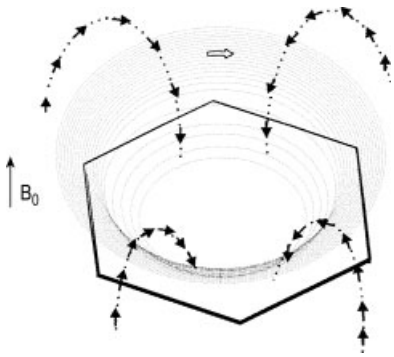

Figure 15. Ring current effects in benzene. The overlapping of the $2p_z$ orbitals leads to two current loops, above and below the *s*-bond plane of the benzene molecule. Only one loop is represented here. A magnetic field (black arrows) is induced by the ring current (white arrows), changing the shielding constant.

Similar arguments can be invoked to explain the apparently anomalous shifts of acetylenic

and aldehydic protons. The free circulation of electrons which gives rise to the diamagnetic effects in spherically symmetric atoms and the benzene ring can also occur around the axis of any linear molecule when the axis is parallel to the applied field. In the case of acetylene the π electrons of the bond can circulate at right angles to the applied field, inducing a magnetic field that opposes the applied field. Since the protons lie along the magnetic axis they are shielded and an upfield shift results. In the case of the aldehydic proton, the effect of the applied magnetic field is greatest along the transverse axis of the C=O bond. This anisotropy causes deshielding of protons lying in a cone whose axis is along the C=O bond, and shielding outside this cone. The aldehydic proton, which lies within the cone, thus experiences a low-field shift and consequently resonates at low fields (9.5 – 10.0 ppm).

The effect of substituents on the proton chemical shifts of hydrocarbons has been extensively investigated. Schoolery's rules [7], [137] enable the prediction of the chemical shift of any CH_2XY and CHXYZ proton. The chemical shift is the shift for methane plus the sum of substituent contributions, which are largely determined by the electronegativity of the substituent. In aromatic compounds strongly electron-withdrawing groups such as nitro and carboxyl deshield all the protons, but the effect is largest at the *ortho* and *para* positions. The opposite is true for strongly electron-donating groups such as NH_2 and OH, while the halogens show smaller effects. For the various types of compounds the substituent increments can be used in an additive manner for compounds with more than one substituent. However, the accuracy of prediction decreases as the number of substituents increases.

Spin – Spin Coupling. The origin of spin – spin coupling and the resulting multiplet structure in proton NMR spectra is introduced in Section 3.2.2. The magnitude of couplings plays an important role in chemical structure determination, particularly in conformational analysis, stereochemistry, and in differentiating between isomeric structures. In proton NMR couplings are observed between protons and to other magnetic nuclei in the molecule such as ^{19}F and ^{31}P. Couplings to rare nuclei are only observed in accordance with their small abundance as so-called "satellites" in ^{1}H NMR spectra. A vicinal proton – proton coupling over three bonds is denoted by $^{3}J_{HH}$, whereas $^{1}J_{PH}$ represents a one bond coupling between a proton and a phosphorus atom. Depending on whether the coupling occurs over two, three, ore more bonds it is designated geminal, vicinal, or long-range coupling.

The magnitude of geminal couplings (^{2}J) depends to a first approximation on the H–C–H angle between the coupled protons. For tetrahedrally substituted carbon atoms a coupling in the range 12 – 15 Hz is typical. In aliphatic systems where there is free rotation about bonds geminal couplings are only observed where methylene protons are diastereotopic, that is, when the molecule contains a chiral center (C∗):

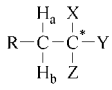

In such compounds H_a and H_b cannot be interchanged by symmetry operations or by rapid rotation and therefore may have different chemical shifts and exhibit coupling to each other. The most commonly encountered coupling in NMR is the vicinal proton-proton coupling. For protons on vicinal carbon atoms in saturated aliphatic groups the coupling constant depends primarily on the dihedral angle Φ between the CH bonds:

$$^{3}J_{(H-C-C-H)} = A + B\cos\Phi + C\cos^2\Phi$$

where A, B and C are constants with the values 4.22, – 0.5, and 4.5, respectively. The graph of this function-the so-called Karplus curve is shown in Figure 16. A series of important regularities is explained by the Karplus curve, e.g., those for couplings in six-membered rings of fixed geometry. An axial-axial coupling J_{aa} in a cyclohexane ring where $\Phi = 180°$ would be ca. 10 Hz according to the Karplus equation. Experimentally, J_{aa} varies from 8 to 13 Hz, J_{ae} from 2 to 4 Hz, and J_{ee} from 1 to 3 Hz, depending on the nature of the substituents present. In acyclic systems the same angular dependence exists, with *gauche* and *trans* couplings typically in the ranges of 2 – 4 and 8 – 12 Hz, respectively. However, in systems with free rotation around the intervening C-C bond, an average coupling of ca. 7 Hz is observed.

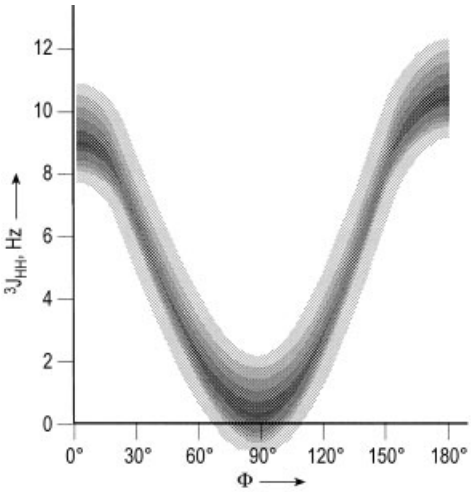

Figure 16. The Karplus curve shows the dependence of vicinal proton-proton coupling $^3J_{\mathrm{HH}}$ on the dihedral angle Φ. The shaded area symbolizes the range of the empirical results.

In olefinic compounds the $^3J_{\mathrm{HH}}$ coupling constant is used to differentiate *cis* and *trans* isomers since $J_{trans} > J_{cis}$. Ranges for *trans* and *cis* couplings are 12 – 18 and 6 – 12 Hz, respectively. Geminal couplings in olefins lie in the range 0 – 3 Hz. In aromatic compounds the observed couplings are also characteristic of the disposition of the hydrogen atoms. For benzene derivatives the following ranges are typical:

$^3J_{ortho} = 6 - 10$ Hz
$^4J_{meta} = 1 - 3$ Hz
$^5J_{para} = 0 - 1$ Hz

Coupling of protons to other nuclei are discussed in the sections concerned with those nuclei.

3.3.2. Carbon (^{13}C) [27–32]

Although present in natural abundance at only 1.1 %, the ^{13}C nucleus has, since the advent of FT methodology, been extensively studied. Only the approximately 5700 times more sensitive ^{1}H NMR is more widely used. In organic chemistry the information obtained from the two nuclei is often complementary for solving structural problems. The large ^{13}C chemical shift range and the small line width of the ^{13}C signals increases spectral resolution effectively. So, ^{13}C NMR is the method of choice for structural investigations of complex organic molecules, of complex mixtures of compounds, as well as biological oligomers and macromolecules.

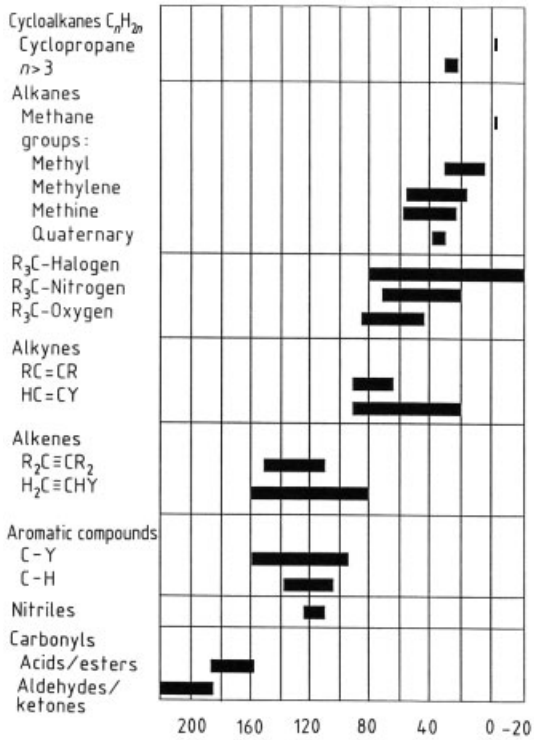

Figure 17. ^{13}C NMR chemical shift ranges
R = H, alkyl; Y = All substituents

Chemical Shifts. For organic chemicals most shifts are spread over a range of 0 – 200 ppm – about twenty times that for the proton. Figure 17 shows chemical shift ranges for some common organic groups. By comparison with Figure 14 it can be seen that there is an overall similarity in ordering on going downfield from TMS to that observed for proton chemical shifts. However, in ^{13}C NMR the effects of substitution are often significant several bonds away from the point of substitution. In simple alkanes α- and β- carbons result in deshielding effects of approximately 9 ppm each whilst a γ- carbon shields by about 2.5 ppm. As a result the C-1 shifts for methane, ethane, and propane are -2.3, 5.7, and 15.4 ppm, respectively. For alkanes and most alkenes the effects are additive and a set of rules have been derived [33]. Sim-

ilarly, substitution leads to typical stereochemical effects in cycloalkanes as also in open-chain compounds. This is of importance for conformational analysis, where the stereochemistry of different conformers can be assessed through their ^{13}C data.

Polar substituents have large effects as can be seen by comparing the shifts for *n*-pentane and 1-pentanol.

13.7 22.6 34.5 22.6 13.7
$CH_3-CH_2-CH_2-CH_2-CH_3$

13.9 22.9 28.5 32.7 62.0
$CH_3-CH_2-CH_2-CH_2-CH_2-OH$

Substituent chemical shifts in benzene derivatives have been extensively studied [34]. Such shifts can be used with caution in an additive manner to predict shifts and hence probable assignments for multi-substituted aromatics.

Spin – Spin Coupling. Carbon – proton couplings are usually sacrificed by broadband proton decoupling in the interests of spectral simplification and obtaining an acceptable spectrum more quickly. However, such couplings can provide valuable structural information.

One-bond carbon – proton couplings J_{CH} are of particular interest. At the simplest level the resulting multiplet provides information about the type of carbon atom; for example, a methyl group gives a quartet resulting from coupling to three equivalent hydrogen atoms. However, such information can be more conveniently obtained from editing experiments (see Section 3.5.3). $^1J_{CH}$ values vary from ca. 110 to 320 Hz, increasing with increasing *s* character of the C–H bond. Thus the couplings for methane (sp^3), ethylene (sp^2), and acetylene (sp) are 125, 156, and 248 Hz, respectively. Polar substituents also have significant effects on the magnitudes of couplings.

$^2J_{CH}$ values in aliphatic hydrocarbons range from 5 to 60 Hz. Couplings are often characteristic of substitution and/or bond hybridization. Thus whilst $^2J_{CH}$ for ethane is 4.5 Hz, the coupling to the methyl carbon of acetaldehyde is 26.7 Hz. Acetylene has a very large $^2J_{CH}$ of 49.3 Hz.

$^3J_{CH}$ couplings show orientational effects analogous to those observed in the corresponding H–H couplings. Thus $^3J_{cis}$ is always less than $^3J_{trans}$ in olefins, and the same Karplus type of dihedral angle dependence exists for saturated aliphatic groups. In aromatic rings $^3J_{CH}$ values are characteristically larger than $^2J_{CH}$ values. For benzene itself, $^3J_{CH} = 7.4$ Hz and $^2J_{CH} = 1.0$ Hz.

Note that heteronuclear couplings to magnetic nuclei other than ^1H (e.g., ^{19}F, ^{31}P, and ^2H) are observed in proton decoupled ^{13}C NMR spectra. Thus coupling to deuterium is seen in signals from deuterated solvents. Some representative C–X couplings constants are given in Table 4.

Table 4. Representative $^{13}C-X$ coupling constants (Hz)

Compound	X	$^1J_{CX}$	$^2J_{CX}$	$^3J_{CX}$	$^4J_{CX}$
C_6H_5F	F	245.3	21.0	7.7	3.3
n-$C_6H_{13}F$	F	166.6	19.9	5.3	
$(C_6H_5)_3P$	P	12.4	20.0	6.7	
$(n$-$C_4H_9)_3P$	P	10.9	11.7	12.5	
C_6H_6	H	157.5	1.0	7.4	
C_6D_6	D	25.5			
$CHCl_3$	H	209.0			
$CDCl_3$	D	31.5			

3.3.3. Fluorine (^{19}F) [21], [35], [37–39]

With 100 % natural abundance and $I = 1/2$, ^{19}F is ideally suited to NMR investigations. Since the earliest days of NMR spectroscopy ^{19}F applications in organic, inorganic, and organometallic chemistry have been widespread. As in the case of the ^{13}C nucleus broadband proton decoupling is often used to simplify spectra.

Chemical Shifts. ^{19}F NMR spectroscopy has the advantage of a very large chemical shift range. Organofluorine compounds alone cover over 400 ppm, whilst the range is extended to some 700 ppm if inorganic compounds are included. $CFCl_3$ is now the generally adopted internal reference. Most organofluorine resonances are high field shifted to the $CFCl_3$ reference peak. In NMR spectroscopy the accepted convention is for shifts to high field of the reference to be assigned negative values. Table 5 gives some representative ^{19}F chemical shifts. The shifts are sensitive not only to directly attached substituents but also to long range effects and thus are very good probes of changes in molecular structure. The chemical shifts given

for the following compound provide an idea of the wide chemical shift range observed for some fluorocarbons.

Table 5. Representative ^{19}F chemical shifts∗

Compound	δ, ppm	Compound	δ, ppm
CH_3F	−271.9	CF_2Cl_2	−6.9
C_2H_5F	−211.5	CF_3CO_2H	−78.5
CF_3CF_3	−88.2	$C_6H_5CF_3$	−63.9
$FCH{=}CH_2$	−114.0	F_2	422.9
$F_2C{=}CH_2$	−81.3	AsF_3	−43.5
$F_2C{=}CF_2$	−135.2	AsF_5	−67.2
C_6F_6	−162.9	SF_6	48.5
C_6H_5F	−113.2	PF_3	−36.2
$C_6H_5SO_2F$	65.5	PF_5	−77.8

∗ Referenced to $CFCl_3$.

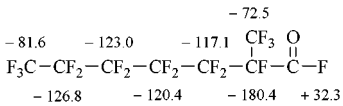

Spin – Spin Coupling. Large compilations of data are available for both fluorine – fluorine and fluorine – hydrogen coupling constants. The presence of several relatively large couplings can lead to complicated spectra even from relatively simple molecules if they are highly fluorinated. However, the spectra are often amenable to first-order analysis, and in general the coupling constants are very sensitive to structure and substituents.

Three-bond fluorine – hydrogen couplings show orientational effects similar to those observed in the corresponding H−H couplings, i.e., the dihedral angle dependence in saturated aliphatics and that $^3J_{cis}$ is less than $^3J_{trans}$ in olefins. Typical coupling constant ranges are given in Table 6.

3.3.4. Phosphorus (^{31}P) [40], [41], [118]

The ^{31}P nucleus has 100 % natural abundance and $I = 1/2$. However, a relatively small magnetogyric ratio results in the sensitivity of the ^{31}P nucleus being only 0.066 that of the proton. Nevertheless, with Fourier transform methodology, the nucleus is easy to detect even at low levels. There is much data in the literature covering applications involving inorganic compounds and complexes, organophosphorus chemistry, and biological and medical studies. In addition to these, even the correlation between ^{31}P and ^{13}C is observable with modern spectrometers in 2D correlation experiments [119].

Table 6. Selected F – F and F – H coupling constant ranges

Group	Coupling	J_{FX} coupling constant range, Hz	
		X = F	X = H
	^2J	40 – 370	40 – 80
	^3J	40 – 80	0 – 40
	^2J	0 – 110	70 – 90
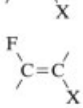	^3J	100 – 150	10 – 50
	^3J	0 – 60	0 – 20
Fluorobenzenes:			
ortho-X	^3J	18 – 35	7 – 12
meta-X	^4J	0 – 15	4 – 8
para-X	^5J	4 – 16	0 – 3

Chemical Shifts. The accepted standard for ^{31}P NMR work is 85 % H_3PO_4, which is normally used as an external reference. Phosphorus compounds are often reactive and no suitable internal reference standard has been found for general use. Relative to H_3PO_4, ^{31}P shifts cover the range from ca. 250 ppm to − 470 ppm. Phosphorus(III) compounds cover the complete range, showing that the chemical shifts are very substituent dependent, whilst phosphorus(V) compounds cover a much smaller range of − 50 ppm to 100 ppm. To a first approximation the substituents directly attached to the phosphorus atom determine the chemical shift. Three variables are important: the bond angles, the electronegativity of the substituent, and the π-bonding character of the substituent. Some representative ^{31}P shifts are given in Table 7.

The phosphorus nuclei are well suited for biological and medical investigations. Relatively simple high resolution ^{31}P NMR spectra are obtained for the chemically different phosphorus atoms of cell metabolism products. The quality

of the spectra which can be obtained from living systems (in vivo ^{31}P NMR spectroscopy) allows to detect metabolic changes. Furthermore, the position of the ^{31}P resonance of inorganic phosphate in the cell is a sensitive indicator for the intracellular pH value [120].

Table 7. Representative ^{31}P shifts*

P^{III} compounds		P^{V} compounds	
Compound	δ, ppm	Compound	δ, ppm
PF_3	97	PF_5	−35
PCl_3	219	PCl_5	−80
PBr_3	227	PBr_5	−101
$P(OMe)_3$	141	$P(OEt)_5$	−71
$P(OPh)_3$	127	$P(OPh)_5$	−86
PH_3	−240	$(MeO)_3P{=}O$	2
$MePH_2$	−164	$(PhO)_3P{=}O$	−18
Me_2PH	−99	$Me_3P{=}S$	59
Me_3P	−62	$Me_3P{=}O$	36
Ph_3P	−6	$Ph_3P{=}O$	25

* Referenced to 85 % H_3PO_4.

Spin – Spin Coupling. Coupling constants have been studied between ^{31}P and a wide variety of other nuclei. Couplings to protons, fluorine, and other phosphorus nuclei have been extensively used for structure elucidation. One-bond phosphorus – hydrogen couplings depend on the oxidation state of the phosphorus atom. Thus for P^{III} compounds ^1J (^{31}P, ^1H) values are on the order of 200 Hz, whilst for P^V compounds they are usually in the range 400 – 1100 Hz. By contrast, one-bond phosphorus – fluorine coupling constants for both P^{III} and P^V compounds have a common range of ca. 800 to 1500 Hz. Phosphines are widely used as ligands in inorganic complexes and therefore phosphorus couplings to some transition metals are important aids to structure determination.

3.3.5. Nitrogen (^{14}N and ^{15}N) [42], [43], [21], [39]

There are two magnetic isotopes of nitrogen, ^{14}N and ^{15}N, with natural abundances of 99.63 % and 0.36 %, respectively. In a magnetic field of 9.4 T the ^{14}N resonance frequency is 28.9 MHz, while that for ^{15}N is 40.5 MHz. There are problems associated with observing both nuclei. The ^{14}N nucleus possesses a quadrupole moment since $I = 1$. As a result, although ^{14}N signals can often be detected easily, the lines can be very broad with widths often between 100 and 1000 Hz. With a 9.4 T applied magnetic field this is equivalent to 3.5 – 34.6 ppm. This can cause resolution problems when a molecule contains more than one type of nitrogen atom.

Although $I = 1/2$ for the ^{15}N nucleus its very low natural abundance results in low receptivity. It also has a negative magnetogyric ratio which leads to a negative NOE when broadband proton decoupling is used. This reduces the intensity of peaks and can result in negative and even null resonances. It has therefore been necessary to use experimental procedures which minimize or eliminate the NOE. The two principal ones are inverse-gated proton decoupling (see Section 3.4.2) and the use of a paramagnetic relaxation reagent such as [Cr (acac)$_3$], which reduces T_1 and consequently the NOE. With modern FT equipment it is now possible to obtain ^{15}N spectra from most small and medium-sized nitrogen-containing molecules at natural abundance level. However, for work with large biological macromolecules ^{15}N enrichment is commonly used.

Chemical Shifts. In principle the use of two isotopes raises the possibility of a primary isotope shift, but in practice ^{14}N and ^{15}N shieldings differ by no more than a few tenths of a ppm and are interchangeable for practical purposes. Nitromethane is used as reference standard for calibration purposes and the convention of assigning negative values to signals occurring high field of the reference is used here also. However, the chemical shift of nitromethane is somewhat solvent and concentration dependent. The total nitrogen chemical shift range covers about 1000 ppm, and ranges for a selection of chemical classes are shown in Figure 18.

Spin – Spin Coupling. It is unusual to observe coupling in ^{14}N spectra where resonances are very broad as a result of quadrupolar relaxation. Coupling is, however, often resolved for ammonium salts. The relative magnitudes of the corresponding ^{14}N and ^{15}N coupling constants are given by the ratio of their magnetogyric ratios. The discussion here is restricted to ^{15}N coupling. The observation of coupling to low-abundance nuclei such as ^{13}C or another ^{15}N nucleus normally requires isotopic enrichment.

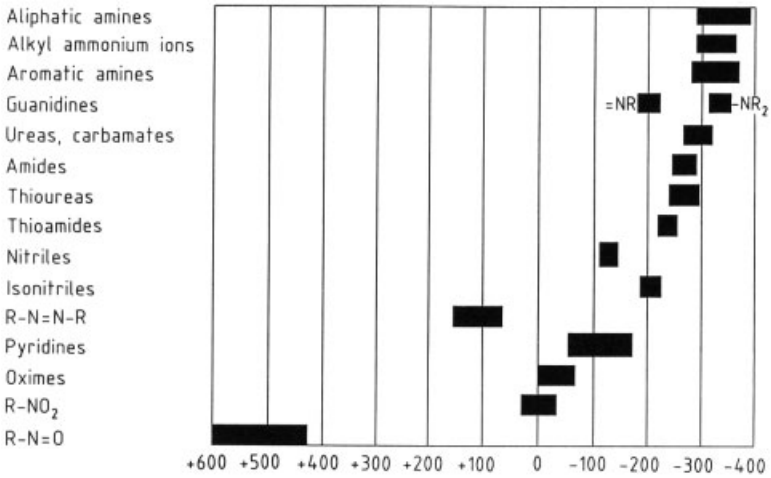

Figure 18. $^{14}N/^{15}N$ NMR chemical shift ranges
R = alkyl, aryl

Couplings to abundant nuclei such as ^{1}H, ^{19}F, and ^{31}P are more readily observed, as in the case of ^{13}C spectroscopy.

One-bond couplings to hydrogen depend on hybridization and electronegativity. Thus the ^{1}J (^{15}N, ^{1}H) values for NH_4^+, $O{=}C(NH_2)_2$, and $HC{\equiv}NH^+$ are 73.3, 89.0, and 134.0 Hz, respectively. One-bond couplings to ^{13}C also show a reasonable correlation with hybridization. The magnitudes of one- and two-bond couplings to ^{13}C are often comparable.

3.3.6. Oxygen (^{17}O) [121], [122]

The only magnetic isotope of oxygen is ^{17}O. With a natural abundance of only 0.037 %, a receptivity that is 0.61 that of ^{13}C, and a quadrupole moment ($I = 5/2$), ^{17}O is difficult to study. Nevertheless, because of the importance of oxygen in both inorganic and organic chemistry as well as biology, the ^{17}O nucleus has been quite widely studied both in natural abundance and in isotopically enriched samples. For small molecules linewidths are typically in the range of several tens to several hundreds of hertz. With large molecules or viscous solutions much larger linewidths may be encountered.

Chemical Shifts. Water is the usual chemical shift reference, being readily available in enriched form. A chemical shift range of over 1000 ppm affords wide spectral disperison. Figure 19 shows representative chemical shift ranges for some organic groups. There is a clear distinction between single and double carbon – oxygen bonds. However, a single resonance is observed from a carboxylic group at the average position expected for the carbonyl and OH groups as a result of rapid proton exchange facilitated by dimeric species. Primary alcohols and ethers resonate close to the water position, whilst lower field shifts are observed for secondary and tertiary species.

Spin – Spin Coupling. Since ^{17}O resonances are usually broad, the majority of reported values are for couplings between directly bonded nuclei. ^{1}J (^{17}O, H) in water and alcohols are ca. 80 Hz. The ^{1}J (P, ^{17}O) couplings involving P=O groups cover the range 145 – 210 Hz. In contrast, couplings involving P–O groups are usually smaller (90 – 100 Hz). Some two-bond couplings have been reported. Examples of two-bond hydrogen – oxygen couplings are 38, 7.5, and 10.5 Hz for the O–CH, O–CH_3, and HC=O interactions in methyl formate.

3.3.7. Silicon (^{29}Si) [44], [123]

^{29}Si, with an abundance of 4.7 %, is the only naturally occurring isotope of silicon with nonzero spin ($I = 1/2$). Whilst its receptivity is 2.1 times

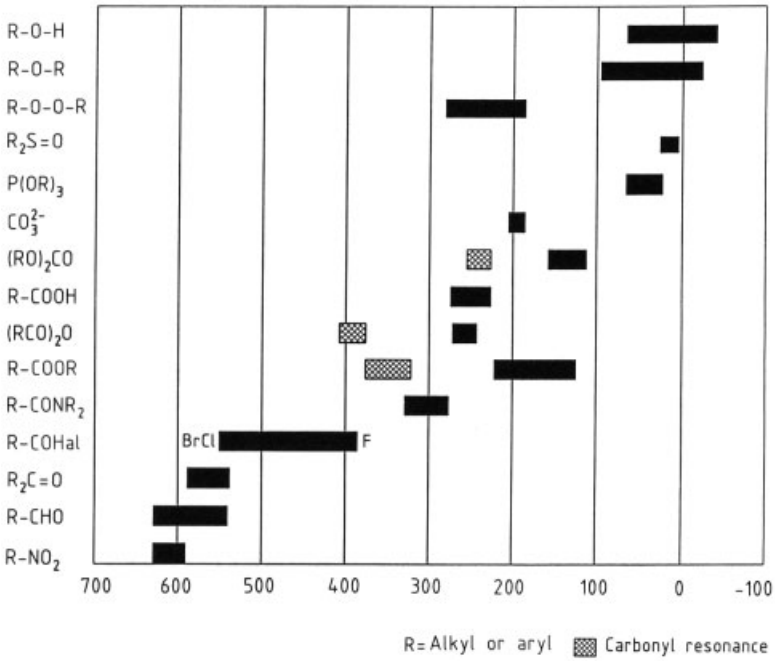

Figure 19. ^{17}O NMR chemical shift ranges
R = Alkyl, aryl

that of the ^{13}C nucleus there are problems associated with its detection. In common with the ^{15}N nucleus a negative magnetogyric ratio results in negative NOEs when broadband proton decoupling is employed, as is usually the case. Inverse-gated decoupling and relaxation reagents are used to circumvent the problem. Broad signals from the glass in the NMR probehead or tube can also cause problems. The glass can be replaced by alternative materials such as teflon or the spectrum obtained from a blank run (i.e., without sample) can be subtracted from the spectrum of the sample. ^{29}Si is very important for solid state NMR, e.g., for analysis of glasses and zeolites [128].

Chemical Shifts. The ^{29}Si resonance of TMS is the usual chemical shift reference. ^{29}Si chemical shifts cover a range of ca. 400 ppm. The large dispersion arising from structural effects makes ^{29}Si NMR a valuable tool for molecular structure determination. The nature of the atoms directly bonded to silicon is of particular importance. Table 8 lists chemical shifts for a series of compounds SiX_nMe_{4-n}. ^{29}Si NMR has proved to be an ideal tool for structure determination of polysiloxane macromolecules.

Table 8. ^{29}Si chemical shifts for SiX_nMe_{4-n} (ppm)

X	$n = 1$	$n = 2$	$n = 3$	$n = 4$
H	−15.5	−37.3	−65.2	−93.1
Et	1.6	4.6	6.5	8.4
Ph	− 5.1	− 9.4	−11.9	−14.0
OMe	17.2	− 2.5	−41.4	−79.2
NMe_2	5.9	− 1.7	−17.5	−28.1
Cl	30.2	32.3	12.5	−18.5

Spin – Spin Coupling. Couplings to abundant nuclei such as the proton and fluorine can be measured either in a coupled ^{29}Si spectrum or by observing the ^{29}Si satellites in the spectrum of the abundant nucleus. The latter are relatively weak compared with peaks from the main ^{28}Si isotopomer. Nevertheless many determinations were made in this way during the 1960s by using CW spectrometers. In a well-resolved ^{1}H NMR spectrum the satellites due to a $^2J_{SiH}$ coupling of 6.8 Hz are readily observed at the base of the TMS reference peak. One-bond proton – silicon coupling constants in silanes are

usually in the range 150 – 380 Hz, the magnitude of J increasing with substituent electronegativity. The literature contains data about coupling to a range of other nuclei, including ^{31}P, ^{13}C, ^{15}N, and ^{29}Si itself.

3.4. Double Resonance Techniques

In Section 3.2.2 reference is made to the use of proton noise (or broadband) decoupling to simplify ^{13}C spectra and improve the signal-to-noise ratio. This is just one example of a large number of double resonance techniques which are available to either simplify spectra or as spectral interpretation tools to aid structure elucidation.

3.4.1. Homonuclear Spin Decoupling

The principles of homonuclear double resonance can be illustrated by considering two coupled protons, A and B, as shown in Figure 11. In the double resonance experiment nucleus A is observed in the usual way in a CW or FT experiment whilst simultaneously selectively irradiating nucleus B with a second much stronger RF field at the "decoupling" frequency. This irradiation induces transitions between the two spin states of nucleus B. If sufficient irradiating power is applied, B flips between the α and β spin states so rapidly that nucleus A can no longer distinguish between the two orientations of B and perceives an average orientation. The coupling J_{AB} disappears and the A signal doublet collapses to give a single line. In the same way irradiation of an arbitrary signal in a complex NMR spectrum results in the obvious collapsing of all multiplets which are coupled with the irradiated proton. At time before 2D NMR was feasible, this method was often used to prove the connectivities between different protons in a molecule.

3.4.2. Heteronuclear Spin Decoupling

In the case of heteronuclear spin decoupling the decoupled and observed nuclei are of different chemical types. The experiments are illustrated with examples in which protons are decoupled and the effect observed on the ^{13}C spectrum. However the methodologies are equally applicable to other heteronuclear spin systems.

Selective spin decoupling is analogous to the homonuclear decoupling experiment described above. In general higher RF power is required than for homonuclear decoupling but lower than is generally used for broadband heteronuclear decoupling. An individual proton signal is selected and irradiated to remove all couplings to that specific proton in the ^{13}C spectrum.

Broadband Spin Decoupling. If the proton decoupler frequency is set to the center of the proton spectrum and modulated by using a noise generator with a bandwidth wide enough to cover the complete proton region, then this is equivalent to simultaneously irradiating every proton frequency. This effectively decouples all the protons in the molecule producing the effect shown in Figure 12 B. Two problems are, however, associated with this type of broadband decoupling. The high RF power required heats the sample and, secondly, decoupling is ineffective if the bandwidth to be irradiated is large. The latter is more of a problem on higher field spectrometers and where broadband decoupling of other nuclei such as ^{19}F or ^{31}P, with their large spectral widths, is required. Alternative methods such as WALTZ-16 [45] based on composite pulse decoupling have largely overcome these problems.

Broadband decoupling also produces a change in intensity, known as nuclear Overhauser enhancement, of the observed signals. For ^{13}C the increase in intensity can be as much as 200 %.

Gated Decoupling. When decoupling is turned on or off, coupling effects appear and disappear instantaneously. The NOE effects however, grow and decay at rates related to some relaxation process. As a result it is possible by computer control of the decoupler to completely separate the two effects.

Sometimes it is necessary to record broadband proton decoupled spectra without any NOE. The NOE may be undesirable either because quantitative information is needed about the relative number of atoms contributing to each

peak in, for example, a ^{13}C spectrum, or because the nucleus being studied has a negative NOE (e.g., ^{15}N or ^{29}Si). Such spectra are also used as controls when measuring NOE values. In other instances coupled spectra with NOE are used as a means of improving sensitivity.

Figure 20 shows a FT ^{13}C experiment with three possible broadband decoupling schemes. The repetitive pulse experiment consists of three parts: a relaxation delay T_d, a ^{13}C RF pulse, and the acquisition time AT during which time the FID is sampled. AT and T_d are on the order of seconds, whilst the pulse duration is a few microseconds. For conventional broadband decoupling (Fig. 20 C) the decoupler is on throughout the experiment. If the decoupler is gated off just before the RF pulse as in Figure 20 D, NOE effects can be observed without any decoupling. Conversely, if the decoupler is switched on when the pulse is applied and off as soon as the acquisition is complete, as shown in (E), the resultant decoupled spectrum does not exhibit an NOE. The latter experiment is referred to as inverse gated decoupling.

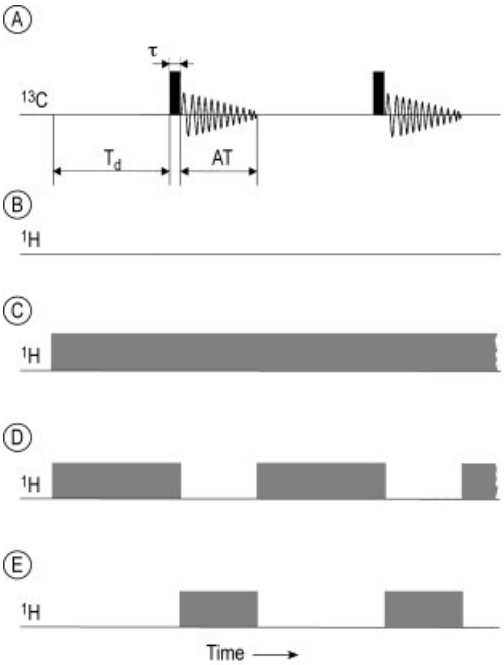

Figure 20. Schematics of an ^{13}C NMR experiment (A) without (B) and with different ^{1}H decoupling techniques: C) broadband decoupling; D) gated decoupling and E) inverse gated (igated) decoupling

Off-Resonance Decoupling. Whilst the broadband proton decoupling described above produces considerable simplification of ^{13}C and other spectra, it removes all the coupling information. As a result the spectra may be difficult to assign. In the early days of ^{13}C NMR a technique known as off-resonance decoupling was used to circumvent this problem. Off-resonance decoupling is achieved by setting the central frequency of the broadband proton decoupler 1000 – 2000 Hz outside the proton spectrum. This results in residual couplings from protons directly bonded to ^{13}C atoms, whereas longer range couplings are lost. The magnitude of the residual coupling, which is smaller than the true coupling J, is determined by the amount of the offset and the power of the decoupler. In such spectra it is usually possible to determine the multiplicities of all the individual carbon atoms, i.e., methyl carbons appear as quartets, methylenes as triplets, etc. Nowadays the use of gated decoupling for multiplicity determination has largely been replaced by spectral editing experiments (Section 3.5.3).

3.4.3. NOE Difference Spectroscopy [46], [47]

The phenomenon of nuclear Overhauser enhancement was introduced in Section 3.2.3 to explain the increase in intensity of ^{13}C peaks when broadband proton decoupling is used. In general terms the NOE is defined as the change in signal intensity of a nucleus when a second nucleus is irradiated. It is beyond the scope of this article to give a theoretical treatment of the origin of the effect, which is caused by through space nuclear relaxation by dipolar interaction. Among other factors the magnitude of the enhancement has a $1/r^6$ dependence on the distance between the two nuclei. Therefore, homonuclear NOEs between protons have been used extensively to obtain information about internuclear distances and thereby to distinguish between possible structures.

The most convenient way of observing such enhancements, which rarely exceed 20 % in the most favorable cases and are sometimes < 1 %, is by a technique known as NOE difference spectroscopy. In such an experiment two spectra are

recorded. The first is a conventional ^{1}H spectrum, while in the second a chosen resonance is selectively irradiated during the interpulse delay of an FT experiment. The second spectrum is then subtracted from the first to give a "difference" spectrum.

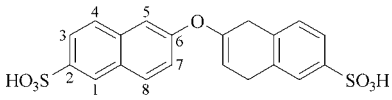

1

Figure 21 A shows the ^{1}H NMR spectrum for **1**. From this spectrum alone it is not obvious whether the ether linkage is at C-6 or C-7. The peak at 8.2 ppm is clearly due to H-1 whilst that at ca. 8.0 ppm is from H-5 or H-8, depending on which isomer is present. Figure 21 B shows the difference spectrum obtained when the resonance at 8.0 ppm is selectively irradiated. The irradiated resonance, which is partially or fully saturated, gives a negative signal in the difference spectrum. Positive signals are obtained from those resonances which exhibit an enhancement. Since a positive enhancement is observed for H-1 this is a good indication that H-1 and H-8 are close together. Measurable effects can be observed between atoms up to about 0.4 nm apart. Since the above description is rather simplistic it should be stressed that because dipolar relaxation of a particular nucleus may involve interaction with more than one other nucleus a basic understanding of the processes involved is a prerequisite to applying the technique successfully. Also spin-spin coupling may complicate NOE measurements. If irradiation is applied to spin multiplets, NOE measurements may fail if selective population transfer is present and special precautions have to be taken in such cases. Furthermore, the NOE effect depends on the molecular correlation time τ. During slow molecular motions frequencies dominate the relaxation process. For macromolecules or small molecules in viscous solvents the NOE may completely vanish or become negative. A positive NOE effect is observable, therefore, only under condition $\omega\tau \ll 1$, which is always true for small and isotropic molecules of low molecular weight compounds ($M_r < 500$) in solutions of low viscosity. In the last years the two-dimensional NMR technique has been used more frequently for measurement of the NOE effect (NOESY spectroscopy, see Section 3.6.5).

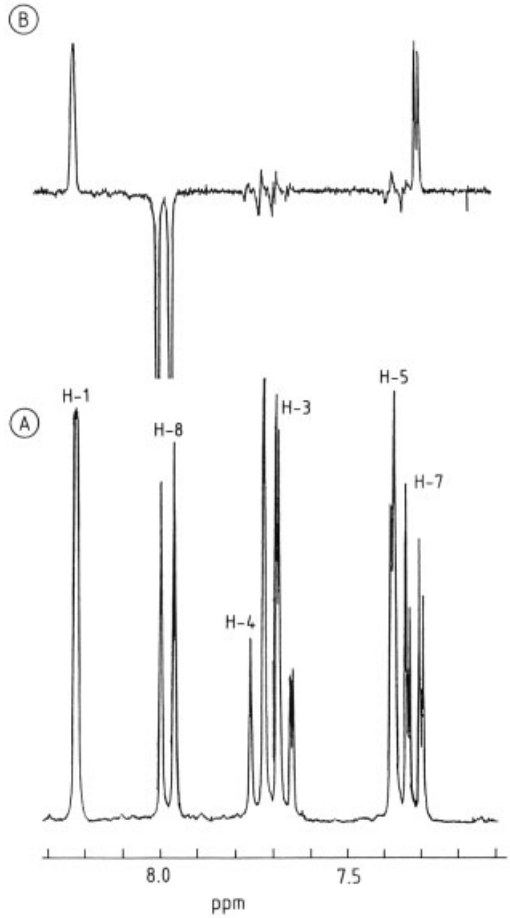

Figure 21. (A) 250 MHz ^{1}H NMR spectrum of **1**; (B) NOE difference spectrum on irradiation of H-8

3.5. One-Dimensional Multi-Pulse FT Experiments

The FT experiment described in Section 3.1.2. and depicted in Figure 7 is a single-pulse experiment. This means that only one RF pulse is applied prior to recording the FID response. A large number of experiments have been developed, for a variety of purposes, in which more than one pulse is applied prior to recording the FID. One of these, the inversion recovery method of determining T_1, is described below in some detail. Others will be mentioned only in so far as their application is concerned.

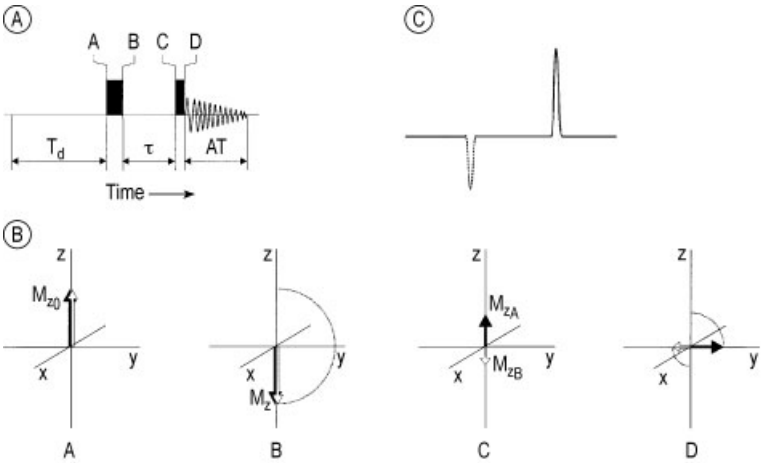

Figure 22. The principle of the inversion-recovery experiment for T_1 measurement: Pulse sequence (A) and position of the macroscopic magnetization M_z (B) during the experiment at four selected points (A-D), shown for two nuclei A and B with different T_1 values, resulting in the spectrum (C)

3.5.1. T_1 Measurement

Spin – lattice relaxation is described briefly in Section 3.2.4. T_1 values can provide valuable information about a molecule. Each of the magnetic nuclei in a different environment in a molecule has a different T_1. Thus each type of hydrogen or carbon atom will have a different value. A knowledge of these values and their relative sizes can give an insight into such things as the type of atom (e.g., whether it is a quaternary or proton-bonded carbon atom), the mechanism responsible for the relaxation, and the molecular motion of that part of the molecule. A knowledge of T_1 values is also required for using NMR as an accurate quantitative tool (see Section 3.8.2.).

The most common method of measuring T_1 is the inversion recovery experiment, which uses the pulse sequence: $[T_d - 180° - \tau - 90° \text{ (FID)}]_n$. The principle is pictured in Figure 22. Whereas a 90° pulse rotates the magnetization vector, which at equilibrium is M_{z0} and lies along the z-axis (Fig. 22 B), to the y-axis of the rotating frame, a 180° pulse which is twice as long inverts the magnetization to the $-z$-axis. Immediately after the pulse the magnetization vector $\boldsymbol{M_z}$ begins to relax back to its equilibrium value M_{z0} according to the following equation:

$$\boldsymbol{M_z} = M_{z0}\,[1 - 2\exp(-\tau/T_1)] \qquad (2)$$

With time, $\boldsymbol{M_z}$ becomes less negative, passes through zero, and eventually relaxes back to M_{z0}. This process proceeds at a different rate for each peak in a spectrum. At a time τ after the 180° pulse, a 90° pulse is applied which tips the magnetization vector onto the $-y$- or $+y$-axis, depending on whether $\boldsymbol{M_z}$ was negative or positive. The delay T_d required before the sequence can be repeated should be at least five times the longest T_1 being measured to enable the system to relax completely. As in the case of the single pulse experiment the sequence is repeated n times and the FID added together. The resulting FID can then be transformed in the usual way. The sign and the amplitude of the resulting peaks depends on the length of the interpulse delay τ and T_1 for each nucleus (Fig. 22 C).

Rearranging Equation (2) and taking logarithms gives:

$$\ln(\boldsymbol{M_z} - M_{z0}) = -\ln(2M_{z0}) - \tau/T_1$$

Thus a plot of $\ln(\boldsymbol{M_z} - M_{z0})$ against τ for each peak gives a straight line with a gradient of $-1/T_1$. In practice a series of experiments is carried out with increasing values of τ. The value of M_{z0} is determined by performing an experiment in which τ is very long.

3.5.2. T_2 Measurement

It is much more difficult to measure T_2 than T_1. One of the main reasons is that imhomogeneities in the magnetic field can make a significant contribution to the apparent spin – spin relaxation time. Complex, so- called spin-echo pulse sequences [49], [50] have been developed for such measurements.

3.5.3. Spectral Editing Experiments [48]

The most frequently used multi-pulse experiments are probably those for multiplicity determination in ^{13}C NMR. Both the J-modulated spin-echo [51] and the distortionless enhancement by polarization transfer (DEPT) [52] experiments are extensively used for this purpose.

Figure 23 shows the pulse sequence used for the *J-modulated spin-echo* experiment which includes the three periods characteristic of all modern pulse experiments, namely preparation, evolution, and detection. The preparation period is a relaxation delay ending with a 90°_x pulse along the x-axis of the rotating frame. The decoupler is switched off, and during the first half of the evolution period the magnetization in the $x-y$ plane evolves under the influence of the proton – carbon couplings. At the beginning of the second period the decoupler is switched on again and a 180°_x pulse refocusses all effects other than J-modulation. In the detection period a proton-decoupled FID signal with NOE is observed. If $t = 1/J$, signals from quaternary and CH_2 carbons are in anti-phase to those from CH and CH_3 carbons. Figure 24 shows such a J-modulated spin-echo ^{13}C spectrum for camphor (**2**).

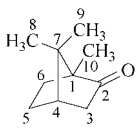

2

When $t = 1/2\,J$ only signals from quaternary carbons are observed. Whilst widely used, this experiment has two shortcomings for multiplicity assignment. Firstly, signals from CH and CH_3 carbons cannot be distinguished, and secondly it is sensitive to variations in J values. One-bond carbon – proton couplings are typically ca. 125 and ca. 160 Hz for saturated aliphatic groups and aromatic/olefinic carbon atoms, respectively. For such compounds use of a compromise value for t works well. However, for carbon atoms with much larger couplings such as alkynyl carbons and some unsaturated heterocyclic carbons it is less successful.

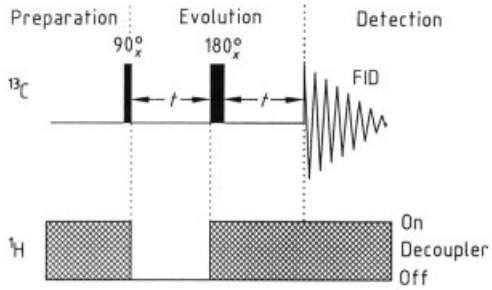

Figure 23. J-modulated spin-echo pulse sequence

In such instances new methods for ^{13}C assignment are more suitable. Compared with the J-modulated spin-echo experiment, the *DEPT* (Distortionless Enhancement by Polarization Transfer) pulse sequence is insensitive to variations in $^1J_{CH}$. The method is based on the use of polarization transfer, whereby magnetization is transferred from protons to ^{13}C nuclei, and as a result quaternary carbon atoms are not detected in such experiments. This experiment involves pulsing both the ^{1}H and ^{13}C channels. The final proton pulse is along the y-axis and by performing two experiments with angles for this pulse of 90° and 135°, full multiplicity data can be obtained. The 135° experiment gives negative CH_2, but positive CH and CH_3 carbon signals whilst in the 90° experiment only CH signals are observed.

3.6. Multi-Dimensional NMR [124]

A wide range of two-dimensional (2 D) experiments are now available as routine tools, and more recently three- and four-dimensional methods have been developed. In general the experiments are used to extract information from complex spectra (e.g., about which nuclei are J coupled to each other) and to measure J couplings. In this article it is only possible to give a very

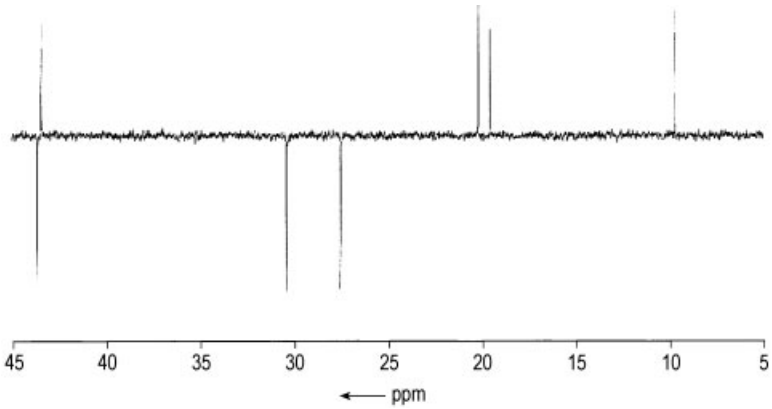

Figure 24. Part of the J-modulated ^{13}C NMR spin-echospectra of camphor. Note the distinction between the C-3 methylene and the C-4 methine carbons with a shift difference of only 0.3 ppm.

brief introduction to two-dimensional methodology together with a few examples of its application in ^{1}H and ^{13}C NMR. Further information about this important field is available in [14], [18], [23], [24], [53], [54].

3.6.1. Basic Principles

The common feature of the 1 D multipulse experiments described above was the time sequence preparation – evolution – detection, whereby the detected signal is only a function of the detection time t_2. The important difference in 2 D NMR is that the evolution time t_1 is now a variable. In a 2 D experiment n separate experiments are performed with incremented values of t_1. For each experiment a free induction decay $S(t_2)$ is measured. In this way a matrix $S(t_1, t_2)$ is built up. Fourier transformation of this matrix with respect to t_2 gives a series of spectra $S(t_1, f_2)$. Further transformation with respect to t_1 gives a matrix $S(f_1, f_2)$ which is a spectrum in two independent frequency dimensions. The dispersion of the signals as a function of f_1 depends on the effect of the pulse sequence used on the spin system as t_1 is varied. In general there are two distinct types of 2 D experiments, referred to as J-resolved and correlated spectroscopy. In a *J-resolved spectrum* one frequency axis (f_1) contains spin – spin coupling information and the other (f_2) chemical shift information. In *correlated spectroscopy* both axes contain chemical shift information, the connection between the two being via spin – spin coupling, NOE, or exchange effects.

3.6.2. *J*-Resolved Spectra

J-Resolved spectra provide the possibility to observe the resonance frequencies δ and the coupling constants J on two distinct frequency axes separately. This is achieved by a pulse sequence $[T_d - 90° - t_1/2 - 180° - t_1/2 - \mathrm{FID}(t_2)]_n$. After excitation by a 90° pulse the evolution time t_1 is divided by a 180° pulse. The signal (FID) is then detected in t_2. This experiment, known as 2 D J-resolved or J,δ-spectroscopy is mainly used in analysis of crowded spectra. The overlapping absorptions in a 1D proton spectrum can sometimes be resolved by this experiment, in which chemical shifts are presented on one axis and coupling constants on the other. The projection of the 2 D spectrum onto the δ-axis is effectively a "^{1}H broad-band-decoupled" proton spectrum. However, the relatively long measuring time, which are usually larger than that of other 2 D methods and a number of artifacts, shown in strongly coupled spin systems, caused a decrease of the application frequency.

In heteronuclear *J*-resolved spectra chemical shifts of an arbitrary nucleus X which couples with protons (this is mostly ^{13}C) are presented on one axis and proton-X J couplings on the other. The information content is equivalent to that in a proton-coupled ^{13}C spectrum

(Fig. 12) but without the severe overlap of multiplets which is usually encountered in the latter. In common with off-resonance proton decoupling, J-modulated spin echo, and DEPT experiments, it facilitates multiplicity determination. In addition, it enables proton-X coupling constants to be measured.

3.6.3. Homonuclear Chemical Shift Correlation (COSY)

The most frequently applied 2 D technique is proton homonuclear correlation using one of the many variants of the COSY (COrrelated SpectroscopY) experiment. The basic COSY experiment consists of two 90°_x pulses separated by a time t_1, as shown in Figure 25. Figure 26 shows the aliphatic part of the COSY spectrum obtained from camphor (**2**).

The spectrum is presented as a contour plot in which intensities above a chosen threshold level are plotted like height contours on a map. Since both frequency domains are proton chemical shifts the matrix is square. Contours along the diagonal of the square correspond to the peaks in a one-dimensional spectrum. Homonuclear couplings give rise to off-diagonal contours or cross peaks. The one-dimensional spectrum is also plotted as a projection along both axes. The cross-peaks provide the same information about proton – proton connectivities as can be obtained from a series of homonuclear decoupling experiments, but all in a single spectrum.

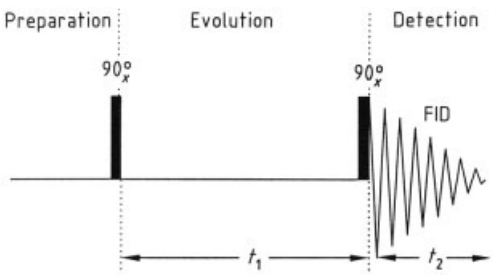

Figure 25. Basic COSY pulse sequence

The connectivities for a particular proton can be extracted by drawing a horizontal line from the relevant diagonal peak that intercepts cross peaks corresponding to correlations. In Figure 26 the multiplet at 2.35 ppm is due to H-3_{exo}, and from the cross-peaks it shows correlations to three other multiplets. By drawing a vertical line from each cross-peak back to the diagonal it can be determined with which contour the first diagonal contour is correlated (i.e., coupled). In this case correlations are observed at 2.1, and 1.84 ppm from H-4 and the attached H-3_{endo}, respectively, and a weak one at 1.9 ppm from the H-5_{exo}, four bonds away from the H-3_{exo}. Note that the cross-peaks are found symmetrically on both sides of the diagonal.

2 D methodology is inherently much less sensitive than 1 D, principally because of the requirement to perform a large number of individual experiments. However, because of the relatively high sensitivity of the proton, a few milligrams of sample are sufficient to yield a good COSY spectrum on a modern high-field instrument. Even shorter measuring time have become possible by the application of gradient accelerated or enhanced COSY spectroscopy, a new technique where the traditional phase cycling techniques will be replaced by the use of pulsed field gradients. Specially designed probe-heads and electronics are necessary to employ the gradients. Depending on the manufacture of the spectrometer, either one or three additional coils are required to generate the *z*- or the *x-, y-, z*-field gradients within the sample. There exists, however, no real enhancement by field gradients, and the abbreviation *gs-COSY* (gradient selected) become increasingly common. Indeed, gradients select desired and undesired coherences. Simplified, a coherence describes all possible mechanism for the exchange of spin population between different energy levels in a NMR diagram. This task was previously performed by time-consuming phase cycling. The selection of the desired coherence already happens in the probe, and only one single transient is sufficient, provided that enough substance is available. Thus, a typical gradient enhanced COSY experiment with 256 time increments can be recorded in a few minutes.

In addition, several other modifications of the COSY pulse sequence exist: COSY-45 for the reduction of the diagonal signals, long-range COSY to emphasize small couplings, relayed-COSY for the observation of remote protons, and the double quantum filter DQF-COSY for elimination of singlet signals. The last one is very important for suppression of water signals in biological samples [125].

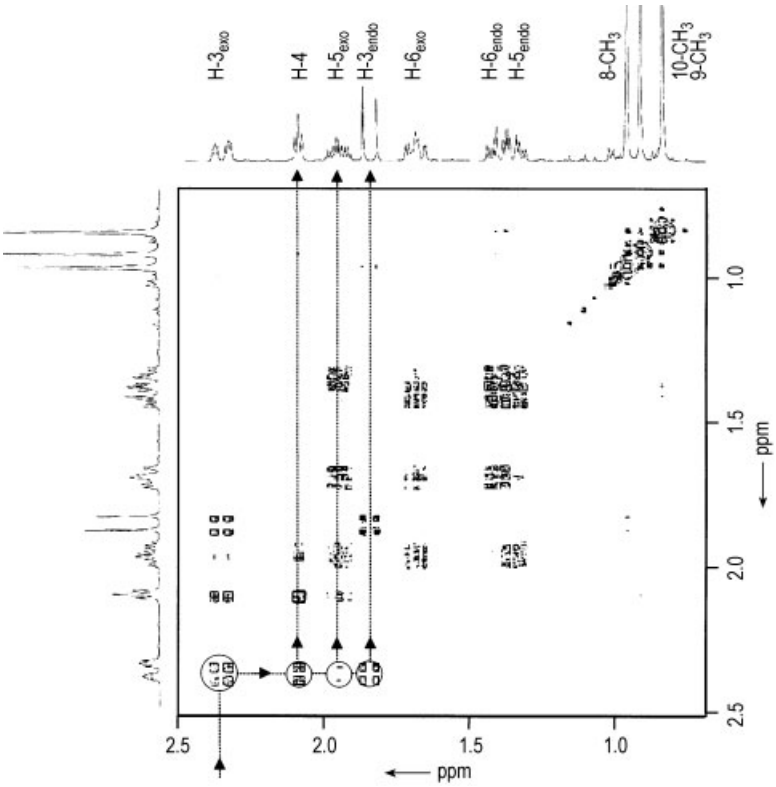

Figure 26. ^{1}H COSY spectrum of camphor

3.6.4. Heteronuclear Chemical Shift Correlation (HETCOR, HMQC)

In the *HETCOR* experiment the peaks of an insensitive nucleus (^{13}C, ^{15}N) are correlated with those of a sensitive nucleus (^{1}H, ^{19}F, ^{31}P). In Figure 27 the aliphatic part of the HETCOR spectrum of camphor (**2**) shows the specific resonances of the protons which are attached to each ^{13}C nucleus. The relevant parts of the corresponding 1 D spectra are plotted along the axes. A correlation is observed as a cross-peak at the intersection of two lines drawn from a proton resonance and from a carbon peak, respectively. The three pairs of diastereotopic methylene protons H-3$_{\mathrm{endo/exo}}$, H-5$_{\mathrm{endo/exo}}$ and H-6$_{\mathrm{endo/exo}}$ give individual cross peaks at the same carbon resonance, respectively. Correlations are not observed for quaternary carbon atoms. The technique is an important tool for chemical shift assignment and thus structure elucidation.

The cross peaks in the basic HETCOR experiment are generated by a magnetization transfer from the sensitive (^{1}H) to the adjacent insensitive nucleus (^{13}C). A number of variants have been developed to increase the sensitivity and to enable observation of long range proton-carbon correlations.

A dramatic increase in sensitivity has been obtained by introduction of the reverse or inverse shift correlation, where the sensitive nucleus (^{1}H) is used for signal detection [126]. This method is known by the acronym *HMQC* (heteronuclear multiple quantum coherence). In theory, a ^{1}H-excited and ^{1}H-detected H,C-correlation experiment is 32 times more sensitive (S/N) compared to a normal ^{13}C detection. In practice, an average S/N-enhancement of 2 – 3 for a HMQC compared to a proton-decoupled 1 D ^{13}C spectrum will be found. For detection of long-range connectivities between carbons and protons the *COLOC sequence* (correlation via long -range couplings) and an inverse technique with the highest possible intensity, the *HMBC* (heteronuclear shift correlations

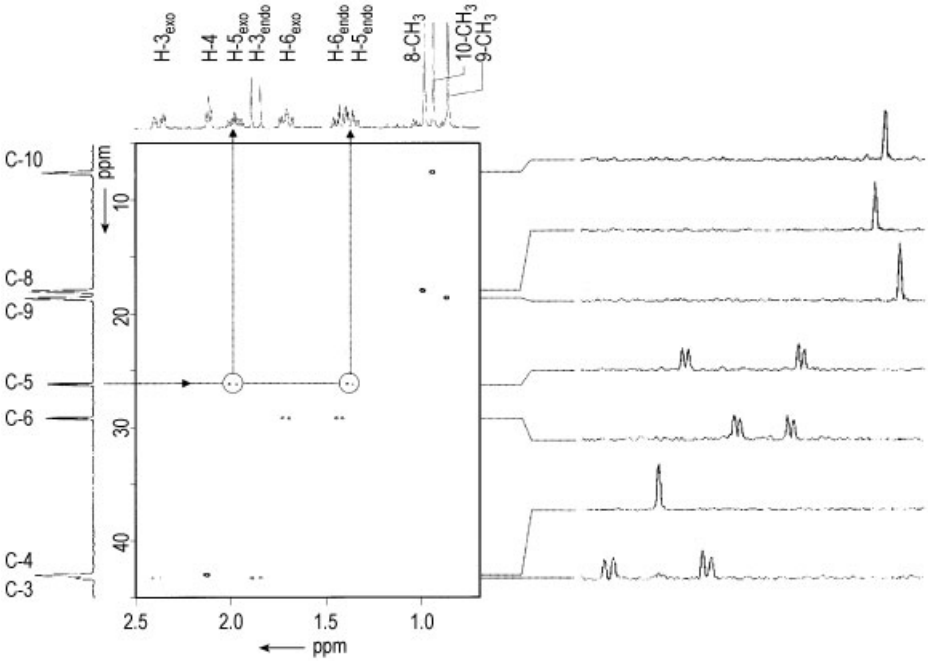

Figure 27. Aliphatic part of the heteronuclear chemical shift correlated (HETCOR) spectrum of camphor. Contour diagram (left) and the traces in the ^{1}H dimension (right) of all C-H-systems.

via multiple bond connectivities) experiment, were developed [127]. In Figure 28 the gradient selected HMBC of camphor are shown. The corresponding 1D ^{1}H and ^{13}C spectra are plotted on the axis analogous to the HMQC spectrum, but in contrast cross peaks are shown here only for H,C-couplings over two (^{1}H-C-^{13}C) or three bonds (^{1}H-C-C-^{13}C). That is illustrated clearly for the carbons C-1 and C-7 in Figure 28. Both quaternary carbons show cross peaks to all three methyl groups whereas the couplings over one bond are suppressed, as shown for the C-10 methyl carbon. Therefore, the interpretation of a HMBC spectrum requires the previous analysis of the HETCOR or HMQC spectrum.

However, there are some problems connected with inverse correlation experiments that should be mentioned. In contrast to the HETCOR, in the inverse experiment the ^{13}C resolution does not depend on the acquisition time AT, but on the number of time increments, because of the interchange of the f_1 and f_2 frequency axes (see Section 3.6.1). Another large problem is the suppression of the undesired signals of protons bound to ^{12}C. In theory, phase cycling should remove all non-^{13}C bonded protons. But a perfect cancellation can never be achieved. Thus, a severe so-called t_1 noise ridge is usually shown at all ^{1}H chemical shift frequencies. In this article it is only possible to name the two methods that help to reduce these problems dramatically: The BIRD (bilinear rotational decoupling) sequence and the pulsed field gradients (PFG, see Section 3.6.3). With PFG the homogeneity of the magnetic field can be destroyed in a controlled way. This yields artifact-free correlation spectra in a fraction of the time need previously (e.g., gs-HMQC). With modern spectrometers, the 2 D gs-HMQC is a good alternative to the normal 1 D ^{13}C NMR spectrum and there is no reason to run conventional HETCOR experiments anymore.

3.6.5. Homonuclear NOE Correlation

The 2 D NOESY (nuclear Overhauser and exchange spectroscopy) technique in principal enables all proton – proton NOE effects to be assembled in a single spectrum. Such spectra are comparable to the respective COSY spectrum. The cross peaks arise from through-space proton-proton interactions between nonbonded

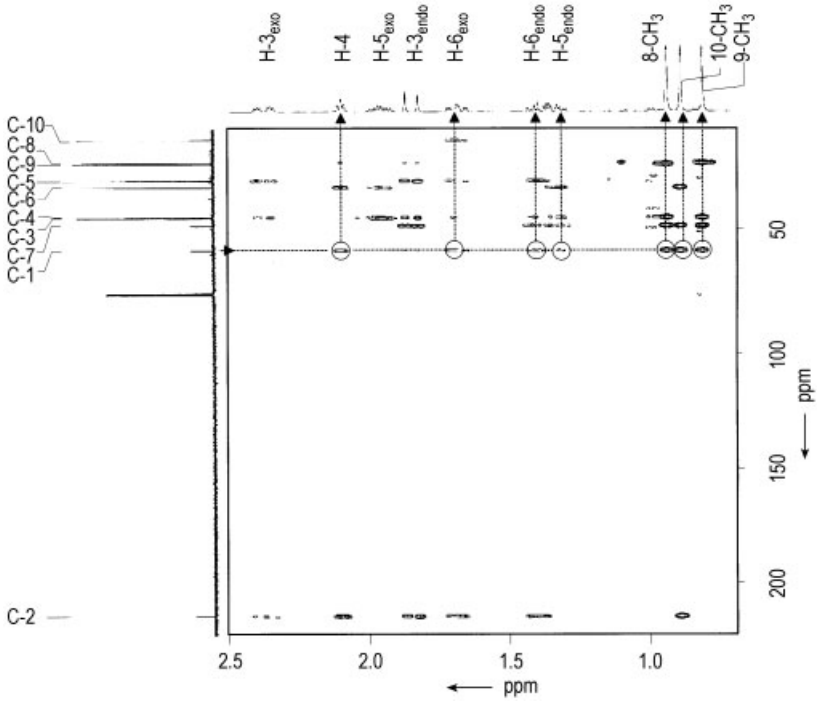

Figure 28. HMBC (*h*eteronuclear shift correlations via *m*ultiple *b*ond *c*onnectivities) spectrum of camphor

protons that are nearby in space. The technique is extensively used in the determination of three-dimensional structures and conformations of molecules. This is shown in Figure 29 with a part of the NOESY spectrum of camphor (**2**). The ^{1}H NMR signals at 0.83 ppm (^{13}C 19.8 ppm) and 0.97 ppm (^{13}C 19.2 ppm) can be unambiguously assigned to the methyl groups 8 and 9 by use of the NOE effects. The high-field proton methyl signal shows an obvious NOE to the H-3$_{\mathrm{exo}}$ proton what is only possible from the 9-position whereas the low-field signal shows effects to the H-5$_{\mathrm{exo}}$and H-6$_{\mathrm{exo}}$ protons.

3.7. NMR Spectral Collections, Databases, and Expert Systems

NMR is a powerful tool for the determination of structures from first principles and the chemical shift is the most important NMR parameter in structural analysis. For estimating the relationship between chemical structures and chemical shifts three possibilities exist: the calculation of the chemical shift values by empirical methods [137], the computation by quantum chemical procedures, e.g., with the IGLO-method (Individual Gauge for Localized Orbitals [129]), or the use of large compilations of NMR spectra and the associated chemical structures. The access to relevant reference data for identical or similar compounds can facilitate the assignment process enormously. Reference data may assist by reducing the amount of experimental and/or interpretive effort required or increase confidence in the suggested structure.

Many data compilations exist, containing significant amounts of reference data, especially those dedicated to a single nucleus. In addition there are many spectral libraries, some available in hard-copy format only. One of the first ^{1}H NMR libraries was published by Varian Associates always in the early 1960s [55]. Such older hard-copy libraries contain only low-field data and are normally searchable by parameters such as name, molecular formula, chemical shift value, or compound type [56–60], [137]. However, at the state of the art are search softwares in computer-readable form. Computer-searchable libraries, especially those which al-

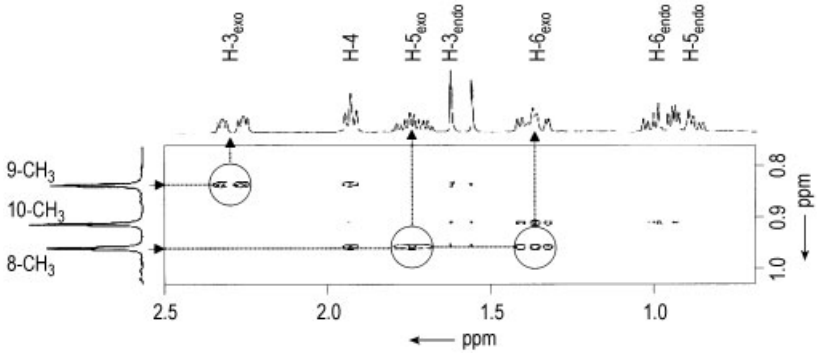

Figure 29. Part of the NOESY spectrum of camphor. Only the NOE effects of the three methyl groups (left) to all other protons (top) are shown here.

low substructure searching of compound data in conjunction with spectral data searching, are much more powerful than hard-copy libraries. In 1973 Bremser published a procedure for the description of the chemical environment of nuclei by their hierarchically ordered spheres of environment (HOSE code) [130]. Today several databases exist containing hundreds of thousands of chemical shift values in particular for ^{13}C nuclei and the corresponding information about the chemical environment of the individual carbons. Only some examples of electronically stored databases should be mentioned here: SpecInfo [61], CSEARCH [132], and CNMR [133]. All three systems listed above contain assigned NMR data which can be searched by molecular formula, name, peak positions/spectrum similarity, and structure/substructure. Users are also able to add their own spectral data to the libraries. All databases can be used to identify known substances contained within the database, and they are an efficient tool for elucidating the structure of unknown substances.

The SpecInfo system which was originally developed in-house by Bremser et al. at BASF and was also marketed in a searchable microfiche version [62], is the premier spectroscopic archive and interpretation system for ^{13}C, ^{1}H, hetero-NMR, IR, UV/VIS, Raman, and mass spectra. In a more automated way the structure assignment can be achieved by using a structure generator, which assembles all possible isomers for a given molecular formula and performs an automatic ranking of the candidate structures (SPECTACLE [131]). In 2000 a PC program was developed on basis of the SpecInfo database for the calculation of ^{13}C NMR spectra of any proposed organic molecular structure [134]. The spherically encoded chemical environments of more than 500 000 carbons were used here to train artificial neural networks which allow the fast determination of ^{13}C chemical shift values.

CSEARCH spectrum estimation is based on scripts written by R. Bobrovsky, G. Löffler ,and W. Robien and can be accessed via an e-mail based server system at the University of Vienna [132]. Four different versions are available. Optionally the stereochemistry of the molecules can be considered and/or neural networks can be used for the fast estimation of chemical shifts.

The Advanced Chemistry Development (ACD) presents a toolset of PC and web-based software for NMR prediction, processing, and database management also for ^{19}F ($>$ 11 500 spectra), ^{31}P NMR ($>$ 18 500 spectra) and for MS, IR, UV-Vis and chromatographic databases. The operation of an actual NMR spectrometer can be simulated here, allowing to choose among different spectra modes (off-resonance, J-modulation, DEPT) just as the operating frequency, the solvent and the concentration of the solute.

Another web-based integrated spectral data base system for organic compounds (SDBS) was developed by the National Institute of Materials and Chemical Research of Tsukuba (Japan) [135]. This system includes six different types of spectra: ^{13}C NMR (ca. 11 000 spectra), ^{1}H NMR (ca. 13 500 spectra), MS, IR, laser-Raman, and ESR (ca. 2000 spectra). The studies on the SDBS started in 1970. This system only allows

the search for available structures and not the determination of spectra from unknown compounds.

However, spectra databases are statistical tools to establish the relationships between NMR spectral parameters and chemical environment of individual atoms. So, the results can only be offered with a statistical probability, depending on the quantity and quality of the available database entries. In other words, the accuracy of the predicted data cannot be more precise than the stored data. Usually, ^{13}C NMR spectra can be calculated for almost any drawn organic structure to an accuracy of ± 3 ppm or better, apart from stereochemical problems which can not be considered by some databases. Table 9 summarizes the results of the determination of the ^{13}C NMR chemical shifts for the ten carbon atoms in camphor (**2**) obtained with different methods.

Many publications on a variety of expert systems and computer-based structure determinations will be found in the literature [136]. In general they are not associated with the large commercially available databases but they are useful for the elucidation of a completely unknown structure. However, in most practical cases the more common type of structure determination is structure verification. Here, the structure information achieved via the chemical shift is usually sufficient.

Table 9. ^{13}C chemical shift estimation of the camphor carbons with different databases

C	CH_n	exp	SpecTools [a] [153]	C_shift [a] [134]	SBSD [135]	ACD [133]
1	C	57.7	64.7	60.3	57.6	57.2
2	C	219.7	213.9	216.5	219.3	217.3
3	CH_2	43.3	33.1	42.9	43.3	43.7
4	CH	43.1	36.3	46.6	43.1	43.3
5	CH_2	27.1	22.1	26.5	27.1	27.3
6	CH_2	30.0	26.3	29.0	29.95	30.1
7	C	46.8	31.8	47.6	46.8	46.6
8	CH_3	19.2	21.4[b]	20.2[b]	19.15	18.7
9	CH_3	19.8	21.4[b]	20.2[b]	19.8	20.7
10	CH_3	9.3	12.7	11.8	9.25	9.5

[a] In contrast to the databases which fall back on the experimental spectra, the shift values were calculated with this PC programs by use of additional increments (SpecTools) or neural networks (C-shift).
[b] Exchangeable since stereochemistry is not considered here.

3.8. Applications

Solution NMR spectra usually contain a wealth of information. Previous sections have provided an overview of experimental techniques, the parameters which can be determined from NMR spectra and techniques which can be used for the assignment of signals in spectra. Here some of the applications of solution NMR spectroscopy are briefly summarized.

3.8.1. Chemical Structure Determination

NMR spectroscopy is the most powerful tool available for the unambiguous determination of molecular structures of compounds in solution. In a modern chemical research laboratory the principal tools for structure determination are usually mass spectrometry, to obtain molecular mass data, and NMR spectroscopy which provides detailed information about the groups present and how they are assembled. The abundant nuclei ^{1}H, ^{19}F, and ^{31}P have been used for over 40 years to study the structures of organic and inorganic chemicals. In the early 1970s the advent of FT methodology enabled a much wider range of nuclei to be studied and in particular resulted in an explosive growth in the use of ^{13}C NMR as a complementary technique to ^{1}H NMR in the identification of organic compounds. The most useful NMR parameters are, in general, chemical shifts, coupling constants, and integrals from which structural information can be deduced.

In addition to the study of monomeric species, NMR is a long established tool for studying the structures of synthetic polymers [63], [138] (see also → Plastics, Analysis, Chap. 6.1.). In the case of copolymers the various groups present can be identified, their relative concentrations determined, and often information about sequence distributions obtained. The latter provides a measure of the random/block character of the polymer. Whilst ^{1}H and ^{13}C are the most commonly studied nuclei for such work several others such as ^{19}F and ^{29}Si have been used where appropriate. For vinyl polymers ^{13}C and ^{1}H NMR spectra provide information on tacticity. Chain branching and end groups can also be identified from NMR spectra of polymers.

During the last decade NMR has developed as a powerful tool for biochemical structure determination. The most impressive applications are in the determination of protein structures with molecular masses up to the 25 000 – 30 000 range [64], [139]. Such work has been made possible by the development of multidimensional NMR techniques. 2 D and 3 D techniques are widely used for such work and 4 D experiments are being developed. 3 D and 4 D methodologies are used in conjunction with ^{15}N and/or ^{13}C labelling of the amino acid residues to reduce resonance overlap. To date NMR protein structures have usually been calculated on the basis of interproton distances and dihedral angles derived from NOE and J coupling measurements, respectively. Each structure determination necessitates a large number of sophisticated experiments, so that data acquisition and subsequent analysis may take several months. The value of the solution structure lies in the fact that it provides a starting point for the study of protein/substrate interactions, a knowledge of which may then aid drug design.

3.8.2. Quantitative Chemical Analysis by NMR [65]

NMR spectroscopy offers several important advantages for quantitative analysis over other techniques, including LC and GC chromatography. Firstly a single technique can be used to unambiguously confirm the identity of the components and quantify them. For chromatographic analysis spectroscopic techniques such as MS and NMR are generally required initially to identify components. Secondly and most importantly pure samples of the compounds of interest are not required to calibrate the response of the instrument in an NMR experiment. This is a result of the fact that, given due attention to the experimental NMR conditions, the integrated resonance intensity is directly proportional to the concentration of nuclei giving rise to the resonance. Relative concentrations can be obtained directly from relative resonance intensities and absolute concentrations by adding a known amount of another compound as an internal intensity standard. The nondestructive nature of the NMR experiment offers an additional advantage. On the negative side, the NMR experiment is inherently much less sensitive than other spectroscopic and chromatographic methods. Therefore the technique is rarely suitable for quantifying components present at very low concentrations even when the more sensitive nuclei such as ^{1}H, ^{19}F, or ^{31}P are used.

In a modern NMR laboratory, pulsed FT techniques are normally used for quantitative analysis. For such work careful selection of the experimental parameters is required to obtain accurate intensity relationships between the resonances in a spectrum. Whilst ^{1}H NMR spectra are normally integrated for structural identification work, the accuracy required to identify the relative numbers of different types of protons in a molecule is much less than for quantitative analysis.

For proton NMR the main consideration is ensuring complete relaxation between successive pulses for all the different types of hydrogen atoms present. This requires the interpulse delay to be at least 5 times the longest T_1. In addition, for nuclei such as ^{13}C, where broadband decoupling is usually required, the inverse gated technique (Section 3.4.2) should be used to prevent the occurrence of NOE effects. A further consideration in the case of spectra from nuclei such as ^{13}C and ^{19}F, which may have very wide spectral widths, is whether the RF pulse has sufficient power to irradiate all the nuclei equally effectively. The digital resolution and data processing requirements for a particular application also require careful selection.

The wide range of chemical, biochemical, and clinical applications include strength determinations, mixture analyses, polymer analyses, and reaction monitoring. Choice of nucleus obviously depends on the individual application. For organic chemical applications, the proton would normally be the preferred nucleus because of its relatively high sensitivity. However, proton spectra of mixtures are often highly congested and the greater dispersion afforded by ^{13}C NMR may make it the nucleus of choice. Two specialist applications are described briefly below.

Isotope Content Determination. A highly developed example of isotope content determination is the site-specific natural isotope fractionation (SNIF) NMR of deuterium [66]. The natural abundance of deuterium at different sites

in a natural product can vary significantly depending on its biochemical origin. This pattern of natural abundances of deuterium is often quite characteristic of the source of the material, as is the case for alcohol in wine. By using a standard with a known deuterium content, ^{2}H NMR can be used to determine the deuteration levels in the alcohol. This data provides information about the origin of the wine and is a means of detecting watering down or artificial enrichment.

Enantiomeric Purity Determination. NMR has become an increasingly important tool for the determination of enantiomeric purity [67]. Many pharmaceuticals and agrochemicals are chiral, with only one enantiomer having the required effect. At best the other is inactive, at worst it may have undesirable properties. This situation has led to a surge in enantioselective synthesis, the products of which require analysis. Chromatographic methods which use chiral columns are usually used for quality control purposes with an established process, but at the research/development stage NMR is often the method of choice. In an achiral medium, enantiomers cannot be distinguished by NMR because their spectra are identical. Diastereoisomers, however, which have different physical properties, can be distinguished by NMR. Therefore the determination of enantiomeric purity by NMR requires the use of a chiral auxiliary that converts a mixture enantiomers into a mixture of diastereoisomers. Chiral lanthanide shift reagents [68] and chiral solvating agents form diastereomeric complexes in situ and can be used directly for NMR analysis. A third method involves reaction with chiral derivatising agents prior to NMR examination.

Lanthanide shift reagents (LSR) are *tris* (β-diketonate) complexes. Further complexation with organic compounds may result in large shifts in the NMR resonances of the latter. This is caused by the magnetic properties of the lanthanide ions. Compounds containing a wide range of functional groups, including alcohols, amines, and ketones, form complexes with LSRs. Prior to the widespread introduction of high-field instruments such materials were widely used to increase chemical shift dispersion in spectra. If an LSR chelate containing an optically pure ligand interacts with a pair of optical isomers, two diastereoisomers result which can be distinguished by NMR. Many of the ligands used in such reagents are derivatives of camphor.

Chiral solvating agents (CSA) form diastereoisomeric solvation complexes with solute enantiomers via rapidly reversible equilibria in competition with the bulk solvent.

Where quantitative chiral purity determinations are required, the minimum level of detection for the minor component and accuracy achieved depend on factors which include the resolution between the resonances from the diastereoisomers formed and the signal-to-noise ratio in the spectrum.

3.8.3. Rate Processes and NMR Spectra [69]

NMR is an important tool for the study of certain types of rate processes. In deciding whether a process is amenable to study by NMR the rate must be compared with the "NMR timescale," which refers to lifetimes of the order of 1 s to 10^{-6} s. An example of a process which can be studied in this way is rotation about the C–N bond in an amide such as *N,N*-dimethylacetamide (3).

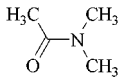

3

Rotation causes the two *N*-methyl groups to exchange positions. In the room-temperature ^{1}H NMR spectrum separate peaks are observed for the two *N*-methyl groups. If the temperature is raised, the rate of rotation increases and the *N*-methyl resonances broaden until a temperature is reached when the two peaks coalesce into a single broad resonance. This peak sharpens as the temperature is raised further until finally the fast exchange limit is reached, above which no further change is detected. The temperatures at which slow exchange, coalescence, and fast exchange are observed for a particular process depend on the rate of the process and the frequency difference between the exchanging lines. Thus at higher field coalescence will take place at a higher temperature. Analysis of such spectra enables exchange rates to be determined over a range of temperatures. Examples of other processes which have been studied in this way in-

clude keto – enol tautomerism, ring inversion, and proton-exchange equilibria.

3.8.4. NMR Methods Utilized in Combinatorial Chemistry and Biochemistry [114], [140]

Combinatorial chemistry (→ Combinatorial Chemistry) provides a powerful means of rapidly generating the large numbers of structurally diverse compounds necessary for the biological screening required by drug discovery. Thus, combinatorial chemistry has become an indispensable tool in pharmaceutical research. The diversity of techniques being employed has caused the adaptation of a number of NMR methods for combinatorial chemistry purposes. These methods include the NMR analysis of reaction intermediates and products in solution or in the gel state and the analysis of ligands interacting with their receptors. High-throughput screening strategies (→ High-Throughput Screening) were developed which involve the acquisition of 2 D spectra of small organic molecules in a few minutes. Libraries of more than 200 000 compound can be tested in less than one month. There are many advantages of high-throughput NMR-based screening compared to conventional assays, such as the ability to identify high-affinity ligands for protein targets. This suggests that the method will be extremely useful for screening the large number of targets derived from genomics research [145].

Shuker et al. described a NMR method in which small organic molecules that bind to proximal subsites of a protein and produce high-affinity ligands are identified. This approach was called "SAR by NMR" because structure-activity relationships (SAR) are obtained here from NMR. This technique involves a series of 2 D spectra of a labeled receptor protein in the presence and absence of potential ligands. The method reduces the amount of chemical syntheses and time required for the discovery of high-affinity ligands and appears particularly useful in target-directed drug research [141].

NMR has become established also as a valuable technique in clinical biochemistry where ^{1}H NMR in particular is used to study complex biochemical fluids such as plasma, urine, and bile. The major problem associated with such work is the removal or reduction of the very large water signal. One method involves freeze-drying the sample and redissolving it in deuterated water. NMR techniques are also available for selective suppression of the water signal. These methods, such as presaturation of the water, are required to reduce the water signal such that acceptable suppression can be attained from phase cycling. The same effect can be accomplished with magnetic field gradients alone because the removal of the undesirable resonances is accomplished in a single scan.

Under the conditions used, only low molecular mass constituents are observable, the signals from proteins and other large molecules are not resolved. Applications include screening and monitoring for metabolic diseases, the study of biochemical mechanisms associated with disease processes, and the identification of drug metabolites.

Translational diffusion measurements are a useful tool for studying supramolecular complexes and for characterizing the association state of molecules that aggregate at NMR concentrations. NMR spectra quality can provide a direct measure of the mobility of organic polymeric compounds and pulsed field gradients can be used to measure diffusion coefficients. Therefore, diffusion measurements can play an important role in the study of the properties and dynamics of the resins used in combinatorial chemistry. *Diffusion-ordered NMR* (DOSY) is another powerful tool for the analysis of complex mixtures. It allows to resolve the NMR signals of discrete compounds in a mixture on basis of variance of their molecular diffusion coefficients [142].

One of the most powerful methods for performing analysis of complex mixtures represents the redevelopment of HPLC-NMR coupling techniques [143]. HPLC-NMR has been widely applied in drug metabolism and also used for combinatorial chemistry. Data collection usually occurs with the system operating in either on-flow or stopped-flow mode. In the latter, up to three dozen fractions may be collected within storage loops. If the differences in the retention time of individual compounds are smaller than the required NMR measuring time, the storage loop acts like a sample changer.

4. NMR of Solids and Heterogeneous Systems [70]

Solids give very broad NMR signals devoid of the fine structure which provides detailed information about chemical structures. The restriction of NMR spectroscopy to the liquid, solution, or gas phase, however, imposes considerable limitations on its applicability since some compounds are insoluble or have very low solubility. Moreover, a compound may experience a change in molecular structure on dissolution (e.g., by tautomerism). Studying samples as solids should provide some insight about how molecules pack as well as about molecular structure. Since the 1970s developments in experimental design and spectrometer hardware have made it possible in some situations to obtain spectra from solids in which the resolution approaches that obtainable in a solution experiment.

This section provides a brief explanation of the origins of the broadening effects, followed by an overview of the techniques employed to obtain high-resolution NMR spectra from solids and examples of their applications. Only spin 1/2 nuclei are considered. There are two principal factors which cause broadening in the NMR spectra of spin 1/2 nuclei obtained from solid samples. The first is chemical shift anisotropy: the chemical shifts are orientation dependent and the different crystallites in a sample have a range of orientations to the applied magnetic field. Broadening resulting from dipole – dipole interactions is the second factor. Both of these interactions are present in solution but are usually averaged out by molecular tumbling. A further problem results from the fact that nuclei in solids generally have very long relaxation times. This means that they take a long time to relax back to the equilibrium magnetization, which in turn affects the amount of time required to obtain a spectrum with a reasonable signal-to-noise ratio.

4.1. High-Resolution NMR of Solids

The study of "dilute" nuclear spin systems such as ^{13}C has been more amenable to obtaining high-resolution solid-state spectra than that of abundant nuclei such as the proton. The reason for this is that the problems associated with homonuclear dipolar interactions are essentially eliminated in a dilute spin system. Dilution may be the result of low natural abundance, as in the case of ^{13}C, or of abundant isotopes present in low concentration (e.g., ^{31}P in organophosphates). In either case homonuclear dipolar interactions are negligible, since dipolar interaction has a $1/r^3$ dependence. Heteronuclear dipolar interactions, usually to protons, remain but these can be removed effectively by the use of high-power proton decoupling. A technique known as magic angle spinning (MAS) is used to remove the effects of chemical shift anisotropy. This involves spinning the sample at an angle of 54°44′ to the applied magnetic field. To be effective the rate of rotation must be comparable to the frequency range being observed (i.e., several kilohertz). To overcome the problems associated with low sensitivity a technique known as cross-polarization (CP) is used. This involves the transfer of magnetization from highly polarizable abundant spins such as protons to dilute spins such as ^{13}C.

The combined application of CP, MAS, and high-power proton decoupling can give high-resolution spectra from powders and amorphous solids, comparable in quality to those obtained in solution. The ^{13}C, ^{29}Si, ^{15}N, and ^{31}P nuclei have been widely studied in this way in the last 20 years [128], [146], [147] . There are two principal reasons for obtaining a high-resolution NMR spectrum from a solid. Either it is insoluble or information is being sought about the solid-state structure. In the latter case crystallographic information is sought either because a single-crystal X-ray structure is not available or to enable a comparison of solid and solution state structures/conformations to be made when spectra are available in both phases. Crystallographic effects frequently give rise to splitting of lines in spectra of solids when compared with those obtained in solution. For a particular type of atom to yield a single line in a spectrum obtained from a solid, all the carbons of that type must be related by symmetry in the crystal. Nonequivalence is not an uncommon occurrence and may be intramolecular or intermolecular. Figure 30 shows the solid-state and solution spectra of 4,4′-dimethoxybiphenyl (**4**). In solution a single peak is observed for C-3 and C-5, whereas for the solid two peaks separated by 6.4 ppm are present. This is attributed [71] to the fact that in the solid the compound exists

in the locked conformation **4** whereas in solution rotation about the C-4–O bond causes an average shift to be observed.

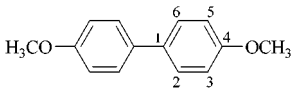

4

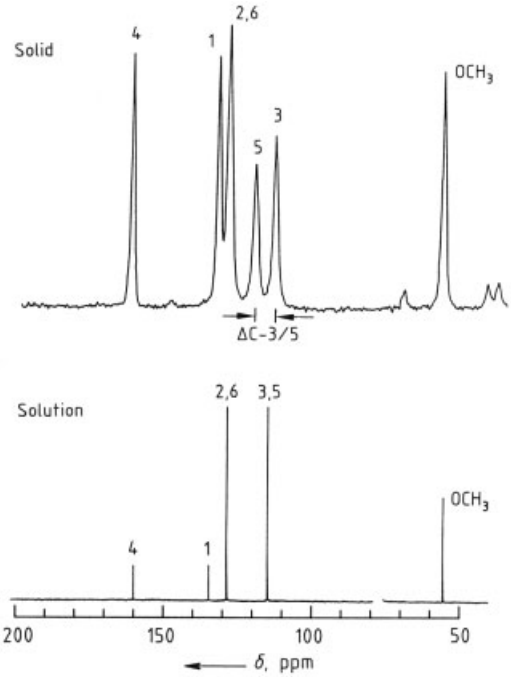

Figure 30. ^{13}C NMR spectra of 4,4′-dimethoxybiphenyl (**4**)

Solid-state NMR spectroscopy, particularly ^{13}C, has been widely used to study polymorphism since polymorphs invariably give spectra which are discernably different.

4.2. Low Resolution ^{1}H NMR of Heterogeneous Systems

Thus far the discussion has assumed a requirement for the highest resolution NMR spectra. However, there are important analytical applications where it is not necessary to resolve the signals from all the different hydrogen atoms in a sample. As in the case of high-resolution NMR, early instruments used CW methodology. Pulsed NMR, which is much better suited to and more versatile for this type of measurement, is now used. After excitation the bulk magnetization of a sample is studied. The initial amplitude of the signal is proportional to the total number of hydrogen nuclei in the sample. Each component of the signal decays at a different rate, so, for example, signals from hydrogen atoms in solid phases decay far more rapidly than those from liquids. There are three types of measurements on such instruments. For many years the methodology has been used to measure solid/liquid ratios (e.g., to determine the solid fat content in margarine). An absolute measurement method is also available where the absolute quantity of protons in a sample can be measured and expressed as a percentage by weight. The third more sophisticated method involves the measurement of bulk relaxation times, which can be correlated with a property of the sample. For instance the decrease in the spin–spin relaxation time T_2 has been used to follow the styrene polymerization reaction [72].

5. NMR Imaging [148–151]

Since 1970 a technique known as magnetic resonance imaging (MRI) has been developed which has revolutionized diagnostic medicine. The technique produces an NMR picture or image. In addition to the NMR methodology described above, MRI also uses magnetic field gradients in the *x, y*, and *z* directions to make the resonance frequency a function of the spatial origin of the signal. First, a slice at height *z* and thickness d*z* of the object is selected. This is done by means of a selective RF pulse, combined with a field gradient in *z* direction. Next, the spatial coding for the *x* direction is obtained by means of an *x*-gradient. Finally, the spatial coding for the second direction, i.e., the *y* direction, is obtained by means of a *y*-gradient. In medical imaging the protons in water are usually detected. While the water content in different types of tissue may show little variation, the T_1 and T_2 relaxation times are different. Pulse sequences are therefore used which produce contrast that reflects the different relaxation times. The technique has become particularly important in the diagnosis of cancer, since cancerous tissue has a longer relaxation time than healthy tissue. This method

acquired special importance in neurochemistry [152].

Equipment is available for a range of applications from whole human body imagers operating at field strengths of up to ca. 8 T to equipment for imaging small samples up to 11.7 T. The achievable resolution depends on the magnitude and duration of the field gradients applied. For small objects the best resolution which has been achieved is of the order of 5 – 10 µm. For whole body in-vivo applications it is much lower.

The techniques developed for biomedical applications to study the liquid state have also been used in materials science. Such applications include the study of solvents ingressed into polymers, liquids absorbed onto porous media, and polymerization reactions [73], all of which can be studied as a function of space and time.

6. ESR Spectroscopy

Electron spin resonance (ESR) spectroscopy is also known as electron paramagnetic resonance (EPR) spectroscopy or electron magnetic resonance (EMR) spectroscopy. The main requirement for observation of an ESR response is the presence of unpaired electrons. Organic and inorganic free radicals and many transition metal compounds fulfil this condition, as do electronic triplet state molecules and biradicals, semiconductor impurities, electrons in unfilled conduction bands, and electrons trapped in radiation-damaged sites and crystal defect sites.

The principles of ESR and general applications are covered in various monographs [74–79]. Reviews of technique developments and applications have been published [80], [81]; industrial applications have been reviewed [82]; and dosimetry [83] and biological applications described [84–88].

6.1. The ESR Experiment

The resonance condition for ESR, for a system having a spin value of 1/2, is:

$$\Delta E = h\nu = g\beta B \tag{3}$$

where ΔE is the separation of energy levels produced by the application of an external magnetic field B, and β is the Bohr magneton (9.274×10^{24} J/K). Most ESR spectrometers operate at a fixed frequency ν and record an ESR spectrum by sweeping the external field B. The most convenient frequency is ca. 9 GHz, which is in the microwave region. It corresponds to a wavelength of ca. 3 cm and is known as X-band. Spectrometers have been built which operate at other microwave frequencies: L-band (1.0 – 1.2 GHz), S-band (3.8 – 4.2 GHz), K-band (24 GHz), and Q-band (34 GHz); and some even operate at a few megahertz in the earth's field. The lower frequencies are used when "lossy" samples, that is those with high dielectric constants such as aqueous or biological samples, are examined. The higher frequencies are used when greater dispersion is required, e.g., when improved separation of anisotropic g components in solid samples is required to reduce overlap of lines. Unfortunately, at the higher frequencies the sample size is restricted and noisier spectra often result.

6.1.1. Continuous Wave ESR

The basic components of a continuous wave (CW) ESR spectrometer are a frequency source (klystron or Gunn diode), an electromagnet (having a field of ca. 350 mT for X-band) and a resonant cavity (or loop gap resonator) in which the sample is placed (Fig. 31). To improve the signal-to-noise ratio, it is customary to modulate the microwave frequency at 100 kHz (other frequencies are also used) and to detect at this frequency by using a phase sensitive detector. This results in the usual recorder tracing from the spectrometer being the first-derivative instead of absorption as found in NMR spectrometers. A simple ESR spectrum is shown in Figure 32.

6.1.2. ENDOR and Triple Resonance

Electron – nuclear double resonance (ENDOR) offers enhanced resolution compared with conventional ESR. This is achieved mainly because the technique is ESR-detected NMR spectroscopy. Another advantage is that only two ENDOR lines arise from each group of equivalent nuclei, resulting in a simplification in the spectra when compared with ESR. ENDOR has

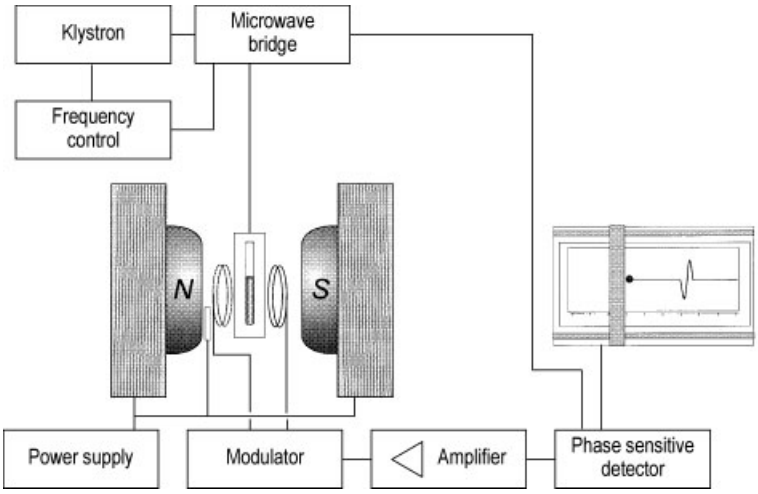

Figure 31. Block diagram of a CW ESR spectrometer

been applied to the study of organic [89] and inorganic radicals and transition metal complexes [84], [90] in both the solid (powdered and crystalline) and liquid states.

The ENDOR experiment involves setting the spectrometer on a chosen line in the ESR spectrum and then increasing the microwave power to saturate the transition while sweeping the selected NMR frequency.

Two NMR transitions are observed at the frequencies:

$$\nu = |\nu_n \pm a/2| \tag{4}$$

where ν_n is the NMR frequency for the applied magnetic field of the ESR spectrometer and a is the ESR hyperfine coupling constant. Equation (4) applies when $\nu_n > -a/2-$; however, when $\nu_n < -a/2-$, the two lines are centred at $-a/2-$ and are separated by $2\,\nu_n$.

Other experiments performed with an ENDOR spectrometer include special and general electron-nuclear-electron triple resonance, where three frequencies are employed. In the special triple experiment, two NMR transitions belonging to one set of equivalent nuclei are irradiated. This results in improved spectral resolution compared with ordinary ENDOR. In the general triple experiment, NMR transitions belonging to different sets of nuclei are irradiated. This allows the relative signs of the hyperfine coupling constants to be determined. As an example the ENDOR spectrum, and the corresponding ESR, general triple and special triple spectra, of bacteriochlorophyll a (**5**) are shown in Figure 33.

Apart from the improvement in resolution already described, ENDOR spectroscopy has several additional advantages over CW ESR, namely:

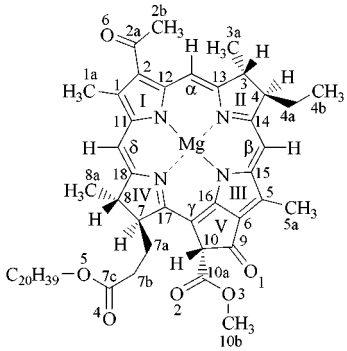

5

1) Simplification of the spectrum when several equivalent nuclei are present
2) Direct measurement of hyperfine splittings
3) Unambiguous assignment of the hyperfine splittings of magnetic nuclei since the spectrometer is set at the appropriate NMR frequency for that particular nucleus
4) Direct determination of quadrupolar splittings

5) Determination of relative signs of hyperfine and quadrupolar interactions
6) Spectra of individual species in a mixture can be determined
7) Very small hyperfine coupling constants can be measured

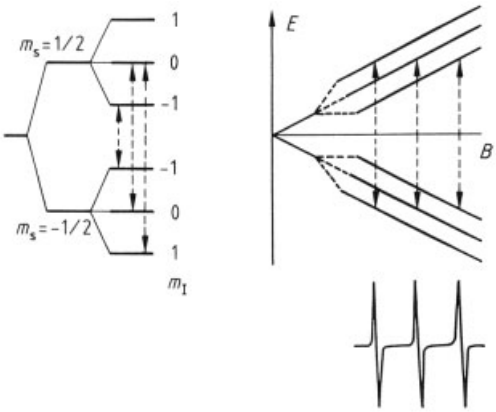

Figure 32. Simple ESR spectrum
P. L. Nordio in [86] (with permission)

6.1.3. Pulse ESR [91–95]

Although the basic principles of ESR and NMR are similar, practical difficulties mean that pulse methods are less useful in ESR. This is because the pulse power required to produce the frequency span of a typical ESR spectrum would be several kilowatts and the pulse very short (nanoseconds). The pulse or FT (Fourier transform) ESR spectrometer is usually based on a standard CW instrument because it is often useful to record a standard ESR spectrum before carrying out pulse experiments. The pulse microwave source is usually a travelling wave tube; other essentials include a travelling wave tube amplifier, a programmable pulse generator, and fast digitizer.

SCHWEIGER [94] has reviewed pulse ESR and pulse ENDOR methods and HOFF [88] includes many applications. Echo methods are similar to those used in NMR. The electron spin echo (ESE) method involves the application of two or three microwave pulses to a spin system in a constant magnetic field. The ESE signal decay due to relaxation is accompanied by periodic rises and falls in the echo amplitude. These modulation effects depend on electron – nuclear and electron – electron couplings in the spin system, and are termed electron spin echo envelope modulation (ESEEM) [92], [93]. ESEEM is used to investigate hyperfine interactions in magnetically diluted solids, and most experiments are performed at temperatures between 4 and 77 K. ESEEM has been applied to the study of surface complexes, coordination of water in metmyoglobin, structures of ligands in copper proteins, and magnetic properties of electronic triplets and solitons in polyacetylene. Pulses and detection need to be on a very short time scale since a typical spin – lattice relaxation time is about 1 μs.

Spin dynamic studies, including saturation and inversion recovery, Hahn echo and stimulated echo decay, as well as Carr – Purcell – Meiboom – Gill sequences, can be performed, yielding relaxation times. When rapid reaction kinetics are being measured, information is also gained on transient phenomena such as chemically induced electron polarization (CIDEP).

Pulsed ENDOR has several advantages over conventional ENDOR for solids [94], [95]. This means that pulsed ENDOR and ESE – ENDOR can be used at any temperature, provided that a spin echo can be detected.

6.1.4. ESR Imaging

ESR imaging [96] is used for both biological and materials research and, as with the more commonly used NMR imaging, it is based on the application of field gradients in the *x, y,* and *z* directions, to allow volume elements to be selectively studied. Large field gradients are required because of the wide field range of the ESR spectrum, typically ca. 3 mT.

Imaging has been performed at various irradiation frequencies. X-band (9 GHz) imaging is limited to samples with a maximum diameter of 10 mm and is unsuitable for most biological samples because of their high water content. Larger or biological samples are usually examined at lower frequency. Small-animal studies have been performed at L-band (1.0 – 1.2 GHz), and the use of even lower frequencies permits samples as large as 100 mm diameter to be examined.

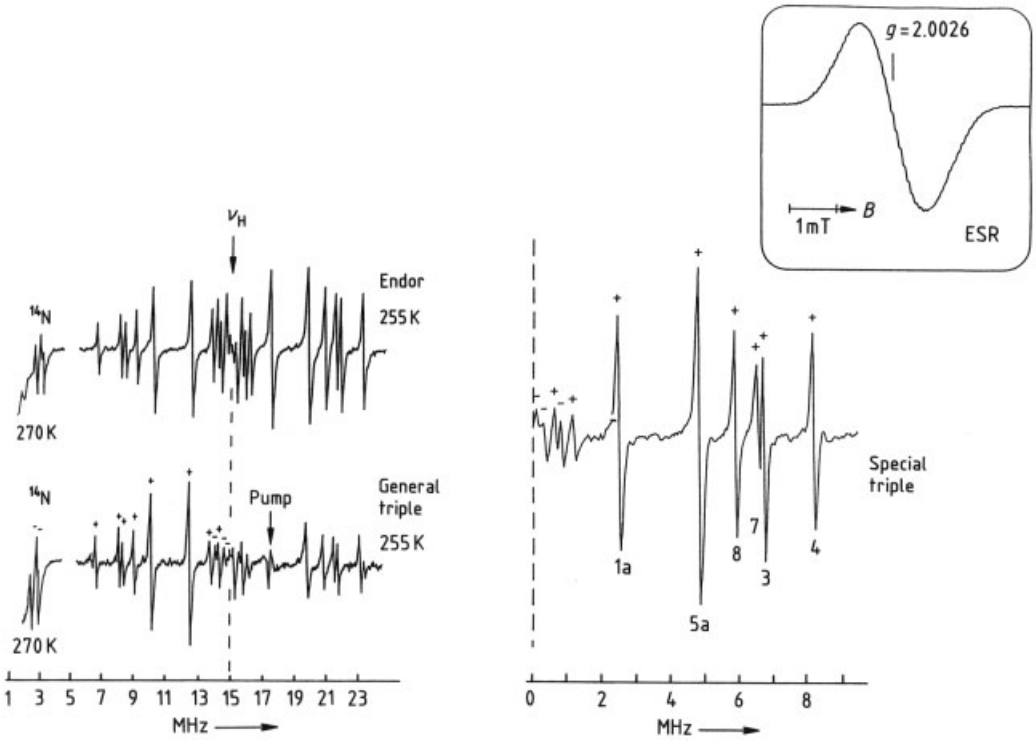

Figure 33. Typical ESR, ENDOR, general triple and special triple spectra of bacteriochlorophyll **a** (**5**) cation radical (iodine oxidation in 6 : 1 CH_2Cl_2 : CH_3OH)
K. Mobius, W. Lubitz and M. Plato in [88] (with permission)

ESR microscopy (at X-band) has been applied to a variety of materials including in situ studies of the oxidation of coals; defects in diamonds; radiometric dosimeters; polymer swelling and to the diffusion of oxygen in model biological systems. The resolution of ESR imaging depends on the field gradient and line width, and the best resolution to date from an X-band imager is 1 µm.

6.2. Spectral Parameters

6.2.1. g-Factor

The ESR g-factor is also known as the Landé g-factor or spectroscopic splitting factor and depends on the resonance condition for ESR (Eq. 3) and is independent of both applied field and frequency. The g-factor of a free electron g_e is 2.002322, while the g-factors of organic free radicals, defect centers, transition metals, etc. depend on their electronic structure. The g-factors for free radicals are close to the free electron value but may vary from 0 to 9 for transition metal compounds. The most comprehensive compilations of g-factors are those published in [97], [98]. The magnetic moments and hence g-factors of nuclei in crystalline and molecular environments are anisotropic, that is the g-factor (and hyperfine interactions) depend on the orientation of the sample. In general, three principal g-factors are encountered whose orientation dependence is given by:

$$g^2 = (g_{xx})^2 (l_{xx})^2 + (g_{yy})^2 (l_{yy})^2 + (g_{zz})^2 (l_{zz})^2 \quad (5)$$

where l_{xx}, l_{yy}, and l_{zz} are the direction cosines between the direction of B and the principal g-factors, g_{xx}, g_{yy}, and g_{zz}. When the site shows threefold or fourfold symmetry only two g-factors, $g_{\parallel}$ (g-parallel) and $g_{\perp}$ (g-perpendicular) are observed, where

$g_\perp = (g_{xx} + g_{yy})/2$. For solutions, free molecular tumbling leads to averaging and the isotropic g-factor g_{iso} is observed:

$$g_{iso} = 1/3\,(g_{xx} + g_{yy} + g_{zz}) \quad (6)$$

or

$$g_{iso} = 1/3\,(g_{\parallel} + 2g_\perp) \quad (7)$$

Figure 34 shows the single crystal, powder and solution spectra expected for an aminoxyl free radical (**6**): this sample has anisotropic g and hyperfine interactions (see Section 6.2.2).

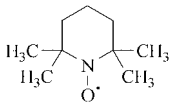

6

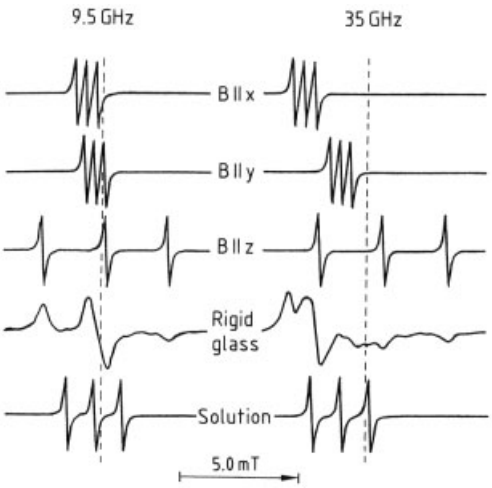

Figure 34. Simple crystal, powder and solution spectrum of a typical aminoxyl radical such as TEMPO (2,2,6,6-tetramethylpiperidine-*N*-oxyl) (**6**).
O. Hayes Griffith, P. Jost in [86] (with permission)

6.2.2. Nuclear Hyperfine Interaction

If the interaction of an electron with a magnetic field were the only effect operative, then all ESR spectra of free radicals would consist of one line. When the nuclear spin quantum number I is nonzero a nuclear hyperfine interaction $\boldsymbol{A}$ is observed. When several equivalent nuclei are present (e.g., $\cdot CH_3$, $\cdot C_2H_5$), the number of lines in the spectrum is given by $\Pi_i\,(2\,n_i\,I_i + 1)$, where n is the number of magnetically equivalent nuclei i.

The isotropic, or Fermi, hyperfine interaction a arises from the presence of nonzero electron density at the nucleus. Interaction between electrons and nuclear dipoles gives rise to the anisotropic, or dipolar, hyperfine coupling $\boldsymbol{T}$. This interaction is orientation dependent, but averages to zero in freely tumbling solution. The total hyperfine interaction $\boldsymbol{A}$ is the sum of a and $\boldsymbol{T}$, where $\boldsymbol{A}$ and $\boldsymbol{T}$ are tensors.

The orientation dependence is given by:

$$A^2\,g^2 = (A_{xx})^2\ (g_{xx})^2\ (l_{xx})^2 + (A_{yy})^2\ (g_{yy})^2\ (l_{yy})^2 + (A_{zz})^2\ (g_{zz})^2\ (l_{zz})^2 \quad (8)$$

where A_{xx}, A_{yy}, and A_{zz} are the principal hyperfine constants. As with the g-factor, an axially symmetric spectrum with only $A_\parallel$ and $A_\perp$ is often observed.

6.2.3. Quantitative Measurements

***g*-Factor and Hyperfine Splitting.** The determination of g-factor and hyperfine splitting depends on accurate determination of the external magnetic field at the sample. This can be done directly by using a NMR magnetometer. The

$$A^2\,g^2 = (A_{xx})^2\ (g_{xx})^2\ (l_{xx})^2 + (A_{yy})^2\ (g_{yy})^2\ (l_{yy})^2 + (A_{zz})^2\ (g_{zz})^2\ (l_{zz})^2 \quad (8)$$

microwave frequency is measured directly with a frequency counter, and the g-factor can then be found from Equation (3). Alternatively, the g-factor and magnetic field sweep of the spectrometer can be determined by comparison with secondary standards. Several authors list suitable materials [74], [99].

Spin Concentration. To determine the electron spin concentration it is essential to obtain, as accurately as possible, the intensity K of the ESR response. Preferably, this is done by double integration of the first derivative spectrum using ESR data acquired in a computer; however, if this is not available or if there is considerable

overlap with lines from a second species then the relation:

$$K = hw^2 \tag{9}$$

is used, where h is the peak-to-peak amplitude and w is the peak-to-peak line width. The latter relationship works satisfactorily if relative intensities are required for signals with similar line shapes. Generally, concentration determinations are made relative to a concentration standard. Some of the points which must be considered in concentration determinations including the choice of concentration standard are noted in [74], [99], [100]. Examples of ESR concentration standards include solutions of α,α'-diphenylpicrylhydrazyl (DPPH), potassium peroxylamine disulfonate ($K_2NO(SO_3)_2$), nitroxide radicals, copper sulfate pentahydrate ($CuSO_4 \cdot 5\,H_2O$), and manganese sulfate monohydrate ($MnSO_4 \cdot H_2O$). Solid standards include single crystals of ruby and $CuSO_4 \cdot 5\,H_2O$, or powders of weak and strong pitch in potassium chloride; DPPH; $MnSO_4 \cdot H_2O$; F-centers in alkyl halides and defects in magnesium oxide. Some of the latter are suitable as g-factor standards also.

For magnetically dilute samples the concentration of the unknown C_X is given by:

$$C_x = C_{std}\, K_x\, (g_{std})^2\ [S\,(S+1)]_{std} / K_{std}\, (g_x)^2\ [S\,(S+1)]_x \tag{10}$$

where S is the total spin (e.g., $S = 1/2$ for organic free radicals, 5/2 for manganese(II) which has five d electrons), and the unknown (X) and standard samples (std) have been examined under identical conditions (temperature, spectrometer gain, modulation and scan range), without microwave power saturation. For neat materials the following applies:

$$M_p = g^2 \beta^2 N_0 S\,(S+1)\, B_r / 3k\,(T-\theta) \tag{11}$$

where M_p is the magnetization of a nondilute paramagnetic material, β is the Bohr magneton, N_0 is the Avogadro number, B_r is the applied field at resonance, k is the Boltzmann constant, T is the absolute temperature, and θ is the Curie temperature.

6.3. ESR in the Liquid State

In solution, molecular motion leads to a simplification of the spectra due to the averaging of anisotropic interactions to zero, resulting in the observation of isotropic spectra.

There have been many studies on organic and inorganic free radicals in solution, including their role as reaction intermediates, mechanistic studies, and studies of molecular dynamics.

6.3.1. Slow Molecular Tumbling

When molecular tumbling leads to incomplete averaging, perhaps because of high solution viscosity, variations are observed in the line width of the hyperfine lines. This variation is shown in Figure 35 for vanadyl acetylacetonate (**7**) dissolved in 1 : 1 toluene : chloroform at room temperature.

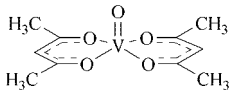

7

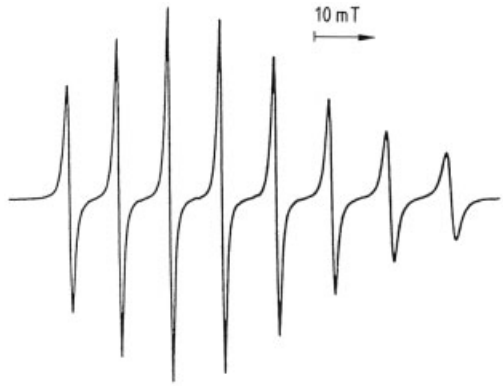

Figure 35. ESR spectrum of vanadyl acetylacetonate (**7**) in 1 : 1 toluene : chloroform, showing asymmetric line broadening

Asymmetric line broadening is commonly observed for aminoxyl radicals, which are therefore used in a wide variety of spin label and spin probe studies (see Section 6.5.2).

The variation of line width $1/T_2$ with m_I has been studied extensively [101], [102]. For isotopic tumbling with correlation times of 10^{-11}

to 10^{-9} s the following expression is generally used:

$$[T_2(m_I)]^{-1} = A + Bm_I + Cm_I^2 + Dm_I^3 \quad (12)$$

The parameters *A, B, C,* and *D* depend on the sign and magnitudes of the g and hyperfine anisotropies and upon the rate of molecular reorientations in the liquid. *D* is usually small compared with *B* and *C* and only terms up to m_I^2 need normally be retained.

In Equation (12) the *A* term includes broadening contributions arising from instrumental factors and mechanisms which are independent of the magnetic anisotropies. The term in m_I progressively broadens the line across the spectrum, transitions with the largest negative m_I value being broadened the least.

The term in m_I^2 broadens the lines at the extremities of the spectrum more than those in the middle, but its effect is symmetrical. The B and C parameters are given by:

$$B = 4/15\ (b\Delta\gamma B_0\tau_c) \quad (13)$$

$$C = b^2\tau_c/8 \quad (14)$$

where B_0 is the operating field of the spectrometer,

$$b = 4\pi\,[A_{zz} - (A_{xx} + A_{yy})/2]/3 \quad (15)$$

$$\Delta\gamma = -\beta_e h^{-1}\,[g_{zz} - (g_{xx} + g_{yy})/2] \quad (16)$$

and τ_c is the rotational correlation time.

According to the Debye diffusion model, a measure of the length of time a molecule remains in a given orientation τ_c can be calculated from the expression:

$$\tau_c = 4\pi\eta r^3/3kT \quad (17)$$

where r is the spherical radius of the molecule and η is the viscosity of the solvent. This method has been used to estimate radical radii and the effects of hindered environments [103].

Correlation times with time scales of 10^{-11} to 10^{-9} s are determined by the above method and are said to be in the fast tumbling region. In the slow tumbling region (10^{-9} to 10^{-7} s) more sophisticated theoretical approaches [101], [102] are used. The very slow region (10^{-6} to 10^{-3} s) cannot be studied by conventional ESR and requires special techniques such as saturation transfer spectroscopy (see Section 6.5.4).

6.3.2. Exchange Processes

Generally, it is assumed that the electrons in free radicals and paramagnetic materials are completely independent and noninteracting. Several effects become important when electron spins interact magnetically and chemically with each other and with their environment.

Electron spin exchange is a bimolecular reaction in which the unpaired electrons of two free radicals exchange their spin states. At low concentration (ca. 10^{-3} mol/L) the usual isotropic spectrum is observed. As the radical concentration is increased the lines gradually broaden until at high concentration ($> 10^{-1}$ mol/L) the lines coalesce to a single line. This single line sharpens further as the concentration is increased.

The spectrum is said to be exchange narrowed since the electron spins are exchanging so fast that the time average of the hyperfine field is close to zero. Electron spin exchange must be avoided if narrow lines are desired in solution and this is readily achieved if the concentration is maintained below ca. 10^{-3} mol/L.

Exchange narrowed spectra (where hyperfine splittings are coalesced) are observed for many pure solid free radicals, such as DPPH and $CuSO_4 \cdot 5\,H_2O$. For this reason, single crystal spectra are obtained by incorporating about 1 % of the paramagnetic compound into an isomorphous diamagnetic host crystal.

6.4. Computer Simulation of Spectra

Isotropic Spectra. Numerous computer programs have been written for the simulation of solution state (isotropic) ESR spectra. Often these include the facility to simulate spectra from mixtures of radicals and to allow for contributions from satellite lines arising from the presence of low-abundance magnetic nuclei such as ^{13}C, ^{15}N, and ^{33}S.

Time-dependent phenomena, including the alternating line width effect [74], [104] and asymmetric line broadening [101], [102], require simulated spectra to obtain the ESR parameters and rates.

Anisotropic Spectra. Programs are available for the simulation of single crystal, partially oriented, and powder simulation (sometimes in the same program) and in several cases allowance for noncoincidence of the axes of the $\boldsymbol{g}$ and $\boldsymbol{A}$ tensors can be included. Various programs are described in detail in [75], [78], [79].

6.5. Specialist Techniques

6.5.1. Spin Trapping

A limitation to the application of ESR spectroscopy to the study of short-lived transient radicals is the difficulty of producing sufficient free radicals for direct observation. Spin trapping provides a simple and efficient means of overcoming this problem [105], [106].

A diamagnetic compound (spin trap) is introduced into the radical producing system to give a relatively stable ESR-observable free radical (spin adduct). Typical spin traps are nitrone compounds such as phenyl-*N*-*tert*-butylnitrone (PBN) (**8**) and 5,5-dimethyl-1-pyrroline-*N*-oxide (DMPO) (**9**) and nitroso compounds such as 2-methyl-2-nitrosopropane (MNP) (**10**) and 2,4,6-tri-*tert*-butylnitrosobenzene (TBN) (**11**)

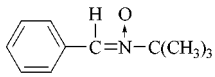

PBN (**8**)

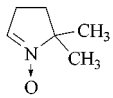

DMPO (**9**)

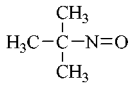

MNP (**10**)

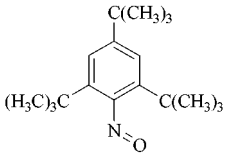

TBN (**11**)

Spin trapping of a radical X with nitrone and nitroso compounds occurs as follows:

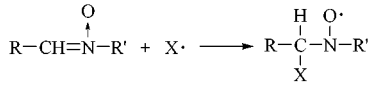

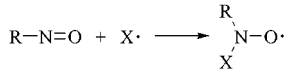

In favorable cases, the original free radical can be identified from the g-factor and hyperfine coupling constants of the spin adduct.

When a methyl radical is trapped by MNP (**10**) the resulting ESR spectrum is a 1 : 1 : 1 triplet of 1 : 3 : 3 : 1 quartets. However, if PBN (**8**) is used, the hyperfine structure is not indicative of the attached radical and its identity must be deduced from hyperfine splittings originating from the nitrogen and lone proton already present in PBN.

The main use of spin trapping is in identifying radical intermediates in organic and inorganic reactions. The reactions can be thermally, electrochemically, or radiation induced. Probably the most important application of spin trapping is the study of radicals in biological systems.

6.5.2. Spin Labeling [86], [87], [107]

Spin labeling is a powerful means of studying biochemical and biophysical characteristics of lipids, proteins, and amino acids. Many investigations have been devoted to the synthesis of radicals having a suitable reactive functional group for coupling selectively to the biological system whilst causing minimum perturbation of biological activity.

In spin labeling, a stable free radical is chemically attached to a macromolecule, and the increase in molecular mass of the radical reduces its rate of tumbling, which can be easily monitored by ESR spectroscopy, as described in Section 6.3.1. The most suitable class of free radicals for the purpose are aminoxyls (also referred to as nitroxides and nitroxyls) because of their stability and because their anisotropic $\boldsymbol{g}$ and $\boldsymbol{A}$ tensors are sufficiently different in the x, y, and z directions to provide high sensitivity to a change

in the rate of molecular tumbling. Rates of tumbling can be measured by the change from a completely isotropic spectrum (rates of about $10^{10}\ s^{-1}$) to a near-powder spectrum (rates of about $10^{6}\ s^{-1}$).

Thousands of aminoxyl spin labels have been synthesized and some examples are given below (compounds **12**–**15**) [107].

A technique related to spin labeling is that of using stable free radicals as spin probes to monitor molecular motions. In this technique a radical, usually an aminoxyl, is synthesized that mimics as closely as possible the structure of the molecule under investigation. The major difference compared with spin labeling is that the radical is not chemically attached to the molecule under investigation.

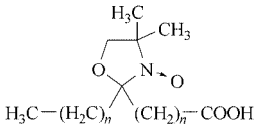

12

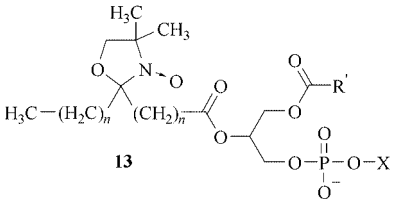

13

X = $(CH_2)_2\overset{+}{N}(CH_3)_3$	(phosphatidylcholine)
X = $(CH_2)_2\overset{+}{N}H_3$	(phosphatidylethanolamine)
X = $CH_2CH(OH)CH_2OH$	(phosphatidylglycerol)
X = $CH_2CH(\overset{+}{N}H_3)COO^-$	(phosphatidylserine)
X = H	(phosphatidic acid)

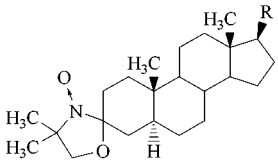

14

R = 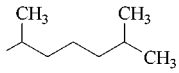	(cholestane derivative)
R = —OH	(androstanol derivative)

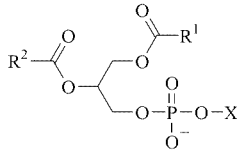

15

X = 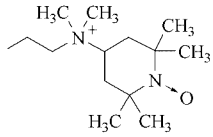

X = 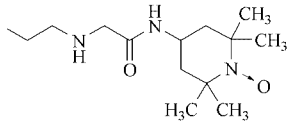

X = 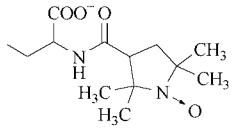

6.5.3. Oximetry

Molecular oxygen is paramagnetic and this property is used to provide an indirect method for the detection and measurement of oxygen by oximetry [108]. Bimolecular collisions of oxygen with free radicals results in broadening of the ESR spectrum. By using radicals with sharp lines and measuring the extent of broadening produced by oxygen it is possible to determine the oxygen concentration. In principle, dissolved oxygen concentrations as low as 10^{-7} mol/L can be rapidly measured.

Probably the most important application of this technique is the measurement of oxygen concentrations in cells and animals, both in vivo and in vitro [109]. Aminoxyl radicals can be used, the most commonly used radical being 3- carboxamido-2,2,5,5-tetramethylpyrrolidine-*N*-oxyl, CTPO (**16**) and its deuterated analogue 3-carbamido-2,2,5,5-tetra-[$^2H_3$]methylpyrrolidine-*N*-oxyl:

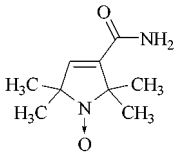

16

Another approach is to use a solid which is pervious to oxygen and whose ESR spectrum has a single very sharp line. This can be achieved with a small crystal of lithium phthalocyanine, which, in the absence of oxygen has a peak-to-peak line width of less than 0.002 mT [110]. A less good but cheaper material is a type of coal called fusinite.

6.5.4. Saturation Transfer

Saturation transfer is a technique that is sensitive to very slow motions with correlation times from 10^{-6} to 10^{-3} s – a region where conventional ESR is not sensitive [111]. It has been used extensively in spin label studies of biological systems [112], [113]. Spectra are recorded under saturation conditions and high modulation amplitudes, as second harmonic in-phase and quadrature spectra. The method measures the transfer of saturation throughout a spectrum by reorientational molecular motions. The line shape is sensitive to molecular motions and the saturation properties of the spin label.

7. References

1. A. Abragam: *The Principles of Nuclear Magnetism,* Clarendon Press, Oxford 1961.
2. D. M. Grant, R. K. Harris (eds.): *Encyclopedia of Nuclear Magnetic Resonance,* J. Wiley, Chichester 1996.
3. C. P. Slichter: *Principles of Magnetic Resonance,* 3rd ed., Springer Verlag, Berlin 1996.
4. R. S. Macomber: *A Complete Introduction to Modern NMR Spectroscopy,* Wiley, New York 1998.
5. R. Freeman: *A Handbook of Nuclear Magnetic Resonance,* 2nd ed., Addison Wesley Longman, Essex 1997.
6. S. W. Homans: *Dictionary of Concepts in NMR,* Oxford University Press, Oxford 1992.
7. R. J. Abraham, J. Fisher, P. Loftus: *Introduction to NMR Spectroscopy,* J. Wiley & Sons, Chichester 1988.
8. R. K. Harris: *Nuclear Magnetic Resonance Spectroscopy,* Longman Scientific & Technical, Harlow 1986.
9. J. Mason: *Multinuclear NMR,* Plenum Publishing, New York 1987.
10. D. Shaw: *Fourier Transform NMR Spectroscopy,* 2nd ed., Elsevier, Amsterdam 1984.
11. T. C. Farrar: *An Introduction to Pulse NMR Spectroscopy,* Farragut Press, Chicago 1987.
12. A. E. Derome: *Modern NMR Techniques for Chemistry Research,* Pergamon Press, Oxford 1987.
13. J. K. M. Sanders, B. K. Hunter: *Modern NMR Spectroscopy,* Oxford University Press, Oxford 1987.
14. R. R. Ernst, G. Bodenhausen, A. Wokaun: *Principles of Nuclear Magnetic Resonance in One and Two Dimensions,* Clarendon Press, Oxford 1990.
15. D. M. S. Bagguley: *Pulsed Magnetic Resonance,* Oxford Univ. Press, Oxford 1992.
16. J. W. Akitt: *NMR and Chemistry: An Introduction to the Multinuclear Fourier Transform Era,* 3rd ed., Chapman & Hall, London 1992.
17. P. J. Hore: *Nuclear Magnetic Resonance,* Oxford University Press, Oxford 1995.
18. H. Günther: *NMR Spectroscopy: Basic Principles,Concepts and Applications in Chemistry,* J. Wiley & Sons, New York 1995.
19. R. Freeman: *Spin Choreography, Basic Steps in High Resolution NMR,* Freeman, Basingstoke 1996.
20. D. Canet: *Nuclear Magnetic Resonance: Concepts and Methods,* Wiley, Chichester 1996.
21. S. Berger, S. Braun, H. O. Kalinowski: *NMR-Spectroscopy of the Non-metallic Elements,* John Wiley, Chichester 1996.
22. G. Batta, K. Kovir, L. Jossuth, C. Szankey: *Methods of Structure Elucidation by High-Resolution NMR,* Elsevier, London 1997.
23. H. Friebolin: *Basic One and Two-Dimensional NMR Spectroscopy,* Wiley-VCH, Weinheim 1998.
24. S. Braun, H.-O. Kalinowski, S. Berger: *150 and More Basic NMR Experiments,* Wiley-VCH, Weinheim 1998.
25. F. Bloch, W. W. Hansen, M. E. Packard, *Phys. Rev.* **69** (1946) 127.

26. E. M. Purcell, H. C. Torrey, R. V. Pound, *Phys. Rev.* **69** (1946) 37.
27. G. C. Levy, R. L. Lichter, G. L. Nelson: *Carbon-13 Nuclear Magnetic Resonance for Organic Chemists,* 2nd ed., Wiley, New York 1980.
28. F. W. Wehrli, A. P. Marchand, S. Wehrli: *Interpretation of Carbon-13 NMR Spectra,* 2nd ed., Wiley, New York 1988.
29. E. Breitmaier, W. Voelter: *Carbon-13 NMR Spectroscopy,* 3rd ed., VCH Publishers, New York 1987.
30. H.-O. Kalinowski, S. Berger, S. Braun: *Carbon-13 NMR Spectroscopy,* Wiley, Chichester 1988.
31. K. Pihlaja, E. Kleinpeter: *Carbon-13 NMR Chemical Shifts in Structural and Stereochemical Analysis,* VCH, Weinheim 1994.
32. Attar-ur-Rahman, V. U. Ahmad: *13C NMR of Natural Products,* **Vol. 1 and 2,** Plenum Press, New York 1992.
33. L. P. Lindeman, J. Q. Adams, *Anal. Chem.* **43** (1971) 1245.
34. D. E. Ewing, *Org. Magn. Reson.* **12** (1979) 499.
35. E. F. Mooney: *An Introduction to ^{19}F NMR Spectroscopy,* Heyden-Sadtler, London 1970.
36. J. W. Emsley, L. Phillips, *Prog. Nucl. Magn. Reson. Spectrosc.* **7** (1971) 1 – 526.
37. V. Wray, *Annu. Rep. NMR Spectrosc.* **10 B** (1980) 1.
38. V. Wray, *Annu. Rep. NMR Spectrosc.* **14** (1983) 1.
39. Landolt Bernstein: *III/35B Nuclear Magnetic Resonance (NMR)Data: Chemical Shifts and Coupling Constants for 19F and 15N,* Springer Verlag, Berlin 1998.
40. D. G. Gorenstein: *Phosphorus-31 NMR: Principles and Applications,* Academic Press, New York 1984.
41. J. Emsley, D. Hall: *The Chemistry of Phosphorus,* Harper & Row, New York 1976.
42. G. C. Levy, R. L. Lichter: *Nitrogen-15 NMR Spectroscopy,* J. Wiley & Sons, New York 1979.
43. M. Witanowski, L. Stefaniak, G. A. Webb, *Annu. Rep. NMR Spectrosc.* **18** (1986) 1.
44. E. A. Williams, J. D. Cargioli, *Annu. Rep. NMR Spectrosc.* **9** (1979) 221.
45. A. J. Shaka, J. Keeler, R. Freeman, *J. Magn. Reson.* **53** (1983) 313.
46. J. H. Noggle, R. E. Schirmer: *The Nuclear Overhauser Effect – Chemical Applications,* Academic Press, New York 1971.
47. D. Neuhaus, M. P. Williamson: *The Nuclear Overhauser Effect in Structural and Conformation Analysis,* VCH Publishers, New York 1989.
48. J. K. M. Sanders, B. K. Hunter: *Modern NMR Spectroscopy,* Oxford University Press, Oxford 1987.
49. H. Y. Carr, E. M. Purcell, *Phys. Rev.* **94** (1954) 630.
50. S. Meiboom, D. Gill, *Rev. Sci. Instrum.* **29** (1958) 688.
51. C. LeCocq, J. Y. Lallemand, *J. Chem. Soc. Chem. Commun.* 1981, 150.
52. D. M. Doddrell, D. T. Pegg, M. R. Bendall, *J. Magn. Reson.* **48** (1982) 323.
53. G. A. Morris, *Magn. Reson. Chem.* **24** (1986) 371.
54. R. Benn, H. Günther, *Angew. Chem. Int. Ed. Engl.* **22** (1983) 350.
55. Varian Associates, Palo Alto: *High Resolution NMR Spectra Catalogue,* **vol. 1, 1962; vol. 2, 1963.**
56. The Sadtler Collection of High Resolution Spectra, Sadtler Research Laboratories, Philadelphia.
57. C. J. Pouchert: *Aldrich Library of NMR Spectra,* vols. 1 and 2, Aldrich, Milwaukee, WI, 1983.
58. W. Brugel: *Handbook of NMR Spectral Parameters,* vols. 1 – 3, Heyden, Philadelphia 1979.
59. L. F. Johnson, W. C. Jankowski: *Carbon-13 NMR Spectra, a Collection of Assigned, Coded, and Indexed Spectra,* Wiley, New York 1972.
60. E. Breitmaier, G. Haas, W. Voelter: *Atlas of C-13 NMR Data,* vols. 1 – 3, Heyden, Philadelphia 1979.
61. SPECINFO, Chemical Concepts, Weinheim.
62. W. Bremser et al.: *Carbon-13 NMR Spectral Data (microfiche),* 4th ed., VCH Publishers, New York 1987.
63. J. L. Koenig: *Spectroscopy of Polymers,* American Chemical Society, Washington 1992.
64. A. Bax, S. Grzesiek, *Acc. Chem. Res.* **26** (1993) 131.
65. D. L. Rabenstein, D. A. Keire, *Pract. Spectrosc.* **11** (1991) 323.
66. G. J. Martin, X. Y. Sun, C. Guillou, M. L. Martin, *Tetrahedron* **41** (1985) 3285.
67. D. Parker, *Chem. Rev.* **91** (1991) 1441.
68. T. C. Morrill (ed.): *Lanthanide Shift Reagents in Stereochemical Analysis,* VCH Publishers, New York 1986.

69. J. I. Kaplan, G. Fraenkel: *NMR of Chemically Exchanging Systems,* Academic Press, New York 1980.
70. C. A. Fyfe: *Solid State NMR for Chemists,* CFC Press, Guelph 1983.
71. A. M. Chippendale et al., *Magn. Reson. Chem.* **24** (1986) 81.
72. Bruker Minispec Application Note 30, Bruker Analytische Messtechnik GmbH, Karlsruhe.
73. P. Jackson et al., *Poly. Int.* **24** (1991) 139.
74. J. E. Wertz, J. R. Bolton: *Electron Spin Resonance; Elementary Theory and Practical Applications,* Chapman & Hall, New York 1986.
75. J. A. Weil, J. R. Bolton: *Electron Paramagnetic Resonance Spectroscopy, Theory and Examples,* Wiley, New York 1993.
76. N. M. Atherton: *Principles of Electron Spin Resonance,* Ellis Horwood, Chichester 1993.
77. A. Abragam, B. Bleaney: *Electron Paramagnetic Resonance of Transition Ions,* Clarendon Press, Oxford 1970.
78. J. R. Pilbrow: *Transition Ion Electron Paramagnetic Resonance,* Clarendon Press, Oxford 1990.
79. F. E. Mabbs, D. Collison: *Electron Paramagnetic Resonance of d Transition Metal Compounds,* Elsevier, Amsterdam 1992.
80. M. C. R. Symons (ed.): "Specialist Periodical Reports," *Electron Spin Resonance,* **vols. 1 – 13,** Royal Society of Chemistry, London 1965 – 1993.
81. J. A. Weil (ed.): *Electron Magnetic Resonance of the Solid State,* Canadian Society for Chemistry, Ottawa 1987.
82. S. A. Fairhurst: "Specialist Periodical Reports," in M. C. R. Symons (ed.): *Electron Spin Resonance,* vol. 13 A, Royal Society of Chemistry, London 1992.
83. M. I. Ikeya: *New Applications of Electron Spin Resonance Dating, Dosimetry and Microscopy,* World Scientific Publ., New York 1993.
84. H. Sigel (ed.): "ENDOR, EPR and Electron Spin Echo for Probing Co-ordination Spheres," *Metal Ions in Biological Systems,* vol. 22, Marcel Dekker, New York 1987.
85. L. J. Berliner, J. Reuben (eds.): "Biological Magnetic Resonance," *EMR of Paramagnetic Molecules,* vol. 13, Plenum Press, New York 1993.
86. L. J. Berliner (ed.): *Spin Labeling, Theory and Applications,* Academic Press, New York 1976.
87. L. J. Berliner, J. Reuben (eds.): "Biological Magnetic Resonance," *Spin Labeling Theory and Applications,* **vol. 8,** Plenum Press, New York 1989.
88. A. J. Hoff (ed.): *Advanced EPR: Applications in Biology and Biochemistry,* Elsevier, Amsterdam 1989.
89. H. Kurreck, B. Kirste, W. Lubitz: *Electron Nuclear Double Resonance Spectroscopy of Radicals in Solution; Applications to Organic and Biological Chemistry,* VCH Publishers, New York 1988.
90. L. Kevan, L. D. Kispert: *Electron Spin Double Resonance Spectroscopy,* Wiley, New York 1976.
91. C. P. Keijzers, E. J. Reijerse, J. Schmidt (eds.): *Pulsed EPR: A New Field of Applications,* North Holland, Amsterdam 1989.
92. L. Kevan, M. K. Bowman (eds.): *Modern Pulsed and Continuous-Wave Electron Spin Resonance,* Wiley, New York 1990.
93. S. A. Dikanov, Y. D. Tsvetkov: *Electron Spin Echo Envelope Modulation (ESEEM)* Spectroscopy, CRC Press, Boca Raton 1992.
94. A. Schweiger, *Angew. Chem. Int. Ed. Engl.* **30** (1991) 265.
95. A. Schweiger, *Pure Appl. Chem.* **64** (1992) 809.
96. G. R. Eaton, S. S. Eaton, K. Ohno: *EPR Imaging and In vivo EPR,* CRC Press, Boca Raton 1991.
97. *Landolt-Börnstein,* **New Series Group II, vols. 1, 9 a, b, c1, c2, d1, d2.**
98. *Landolt-Börnstein,* **New Series Group II, vols. 2, 8, 10, 11.**
99. I. B. Goldberg, A. J. Bard: *Treatise on Analytical Chemistry,* part 1, vol. 10, Wiley Interscience, New York 1983.
100. R. G. Kooser, E. Kirchmann, T. Matkov, *Concepts Magn. Reson.* **4** (1992) 145.
101. J. H. Freed in L. J. Berliner (ed.): *Spin Labeling Theory and Applications,* Academic Press, New York 1976.
102. J. H. Freed in L. J. Berliner, J. Reuben (eds.): *Spin Labeling Theory and Applications,* vol. 8: "Biological Magnetic Resonance," Plenum Press, New York 1989.
103. R. F. Boyer, S. E. Keinath: *Molecular Motions in Polymers by ESR,* Harwood Academic Press, Chur 1980.
104. P. D. Sullivan, J. R. Bolton in J. S. Waugh (ed.): *Advances in Magnetic Resonance,* vol. 4, Academic Press, New York 1970.
105. M. J. Perkins in V. Gold, D. Bethell (eds.): *Advances in Physical Organic Chemistry,* vol. 17, Academic Press, London 1980.

106. E. G. Janzen, D. L. Haire in D. D. Tanner (ed.): *Advances in Free Radical Chemistry,* vol. 1, JAI Press, London 1990.
107. D. Marsh, *Tech. Life Sci. B4/II* **B 426** (1982) 1 – 44.
108. J. S. Hyde, W. K. Subczynski in L. J. Berliner, J. Reuben (eds.): "Biological Magnetic Resonance," *Spin Labeling Theory and Applications,* **vol. 8,** Plenum Press, New York 1989.
109. R. K. Woods et al., *J. Magn. Reson.* **85** (1989) 50.
110. X.-S. Tang, M. Mousavi, G. C. Dismukes, *J. Am. Chem. Soc.* **113** (1991) 5914.
111. J. S. Hyde, L. R. Dalton in L. J. Berliner (ed.): *Spin Labeling Theory and Applications,* Academic Press, New York 1979.
112. M. A. Hemminga, P. A. Jager: "Biological Magnetic Resonance," in L. J. Berliner, J. Reuben (eds.): *Spin Labeling Theory and Applications,* vol. 8, Plenum Press, New York 1989.
113. P. F. Knowles, D. Marsh, H. W. E. Rattle: *Magnetic Resonance of Biomolecules; An Introduction to the Theory and Practice of NMR and ESR in Biological Systems,* Wiley-Interscience, New York 1976.
114. P. A. Keifer, *Drugs of the Future* **23** (1998) 301.
115. J. P. Hornak, http://www.cis.rit.edu/htbooks/nmr/inside.htm, Rochester Institute of Technology, Rochester, NY.
116. P. A. Petillo, http://www-chem430.scs.uiuc.edu/lessons/lessons.htm, University of Illinois, Urbana, IL.
117. D. J. Cookson, B. E. Smith, *J. Magn. Reson.* **57** (1984) 355.
118. L. D. Quin, J. G. Verkade: *Phosphorus-31 NMR Spectral Properties in Compound Characterization and Structural Analysis,* VCH Publishers, New York 1994.
119. S. Berger, T. Faecke, R. Wagner, *Magn. Reson. Chem.* **34** (1996) 4.
120. D. G. Gadian: *Nuclear Magnetic Resonance and its Application to Living Systems,* Clarendon Press, Oxford 1982.
121. D. W. Boykin: *17O NMR Spectroscopy,* CRC Press, Boca Raton 1990.
122. S. Kuroki: *O-17 NMR in I. Ando, T. Asakura (ed.): Solid State NMR of Polymers,* Elsevier Science Publ, Amsterdam 1998, p. 236.
123. Y. Takeuchi, T. Takayama: "Si-29 NMR spectroscopy of organosilicon compounds" in Z. Rappoport, Y. Apeloig (eds.): *Chemistry of Organic Silicon Compounds,* **vol. 2,** J. Wiley & Sons, Chichester 1998, p. 267.
124. F. J. M. van de Veen: *Multidimensional NMR in Liquids: Basic Principles and Experimental Methods,* VCH, Weinheim 1995.
125. A. E. Derome, M. P. Williamson, *J. Magn. Reson.* **88** (1990) 177.
126. A. A. Maudsely, L. Müller, R. R. Ernst, *J. Magn. Reson.* **28** (1977) 463.
127. W. Bermel, C. Griesinger, H. Kessler, K. Wagner, *J. Magn. Reson. Chem.* **25** (1987) 325.
128. G. Engelhardt, D. Michel: *High Resolution Solid State NMR of Silicates and Zeolites,* Wiley, Chichester 1987.
129. M. Schindler, W. Kuttzelnigg, *J. Chem. Phys.* **76** (1982) 1919.
130. W. Bremser, *Anal. Chim. Acta* **103** (1973) 355.
131. LabControl Scientific Consulting and Software Development GmbH, S*pectroscopic Knowledge Management System – SPECTACLE*, Köln, Germany.
132. V. Purtuc, V. Schütz, S. Felsinger, W. Robien, Institute for Organic Chemistry, University of Vienna, Vienna, Austria, http://felix.orc.univie.ac.at/
133. Advanced Chemistry Development Inc., *CNMR data base*, Toronto, Canada.
134. J. Meiler, R. Meusinger, M. Will, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1169.
135. K. Hayamizu, M. Yanagisawa, O. Yamamoto, *Integrated Spectral Data Base System for Organic Compounds (SDBS)*, National Institute of Materials and Chemical Research, Tsukuba, Japan, http://www.aist.go.jp/RIODB/SDBS/menu-e.html
136. E. M. Munk, *J. Chem. Inf. Comput. Sci.* **38** (1998) 997.
137. E. Pretsch, J. T. Clerc: *Spectra Interpretation of Organic Compounds,* VCH, Weinheim 1997.
138. F. A. Bovey, P. A. Mireau: *NMR of Polymers,* Academic Press, San Diego 1996.
139. D. G. Reid: *Protein NMR Techniques,* Humana Press, Totowa, N.J. 1997.
140. M. J. Shapiro, J. S. Gounarides, *Prog. NMR Spectr.* **35** (1999) 153.
141. S. B. Shuker, P. J. Hajduk, R. P. Meadows, S. W. Fesik, *Science* **274** (1996) 1531.
142. H. Barjat, G. A. Morris, S. Smart, A. G. Swanson, S. C. R. Williams, *J. Magn. Reson. Ser. B.* **108** (1995) 170.
143. M. Spraul, M. Hoffman, P. Dvortsak, J. K. Nicholson, I. D. Wilson, *Anal. Chem.* **65** (1993) 327.

144. D. L. Olson, T. L. Peck, A. G. Webb, R. L. Magin, J. V. Sweedler, *Science* **270** (1995) 1967.
145. P. J. Hajduk, T. Gerfin, J. M. Boehlen, M. Häberli, D. MArek, S. W. Fesik, *J. Med. Chem.* **42** (1999) 2315.
146. I. Ando, J. Austin: *Solid State NMR Spectroscopy,* Elsevier, Amsterdam 1995.
147. E. O. Stejskal, J. D. Memory: *High Resolution NMR in the Solid State,* Oxford University Press, New York 1994.
148. B. Hills: *Magnetic Resonance Imaging in Food Science,* Wiley-VCH, Weinheim 1998.
149. R. Kimmich (ed.): *NMR Tomography, Diffusometry, Relaxometry,* Springer Verlag, Berlin 1997.
150. R. R. Edelmann, J. R. Hesselink, M. B. Zlatkin (eds.): *Clinical Magnetic Resonance Imaging,* W. B. Saunders Comp. 1996.
151. M. T. Vlaardingerbroek, J. A. den Boer: *Magnetic Resonance Imaging,* Springer Verlag, Berlin 1996.
152. H. Bachelard (ed.): *Magnetic Resonance Spectroscopy and Imaging in Neurochemistry,* Plenum Press, New York 1997.
153. A. Gloor, M. Cadisch, R. Bürgin-Schaller, M. Farkas, T. Kocsis, J. T. Clerc, E. Pretsch, R. Aeschimann, M. Badertscher, T. Brodmeier, A. Fürst, H.-J. Hediger, M. Junghans, H. Kubinyi, M. E. Munk, H. Schriber, D. Wegmann: *SpecTool 1.1, A Hypermedia Book for Structure Elucidation of Organic Compounds with Spectroscopic Methods,* Chemical Concepts GmbH, Weinheim 1995.

Mass Spectrometry

MICHAEL LINSCHEID, Department of Chemistry, Humboldt University, Berlin, Germany

1. Introduction . . . 2
2. General Techniques and Definitions 3
2.1. Resolution . . . 3
2.2. Tools for Structure Elucidation . . . 4
2.2.1. Full Spectra: Low Resolution . . . 4
2.2.2. Elemental Compositions of Ions . . . 4
2.3. Fragmentation in Organic Mass Spectrometry . . . 5
2.4. Quantitative Analysis . . . 6
3. Sample Inlets and Interfaces . . . 8
3.1. Direct Probe . . . 8
3.2. Batch Inlets . . . 8
3.3. Pyrolysis . . . 8
3.4. GC/MS Interfaces . . . 9
3.5. LC/MS Interfaces . . . 10
3.5.1. Moving Belt . . . 10
3.5.2. Continuous Flow FAB . . . 10
3.5.3. Direct Liquid Introduction . . . 10
3.5.4. Supercritical Fluid Interface . . . 11
3.5.5. Particle Beam Interface . . . 11
3.5.6. Chemical Ionization at Atmospheric Pressure . . . 11
3.5.7. Thermospray . . . 12
3.5.8. ESI Interface . . . 13
3.6. TLC/MS . . . 13
4. Ion Generation . . . 13
4.1. Electron Impact . . . 13
4.2. Chemical Ionization . . . 14
4.3. Negative Chemical Ionization (Electron Capture) . . . 14
4.4. Field Ionization (FI) . . . 14
4.5. Plasma Ionization . . . 15
4.6. Thermal Ionization . . . 16
4.7. Optical Methods . . . 16
4.8. Desorption Methods . . . 17
4.8.1. Secondary Ion Mass Spectrometer (SIMS) . . . 17
4.8.2. Field Desorption (FD) . . . 17
4.8.3. Fast Atom Bombardment (FAB) . . . 18
4.8.4. ^{252}Cf Plasma Desorption . . . 18
4.8.5. Laser Desorption/Ionization . . . 19
4.9. Electrospray . . . 19
5. Analyzers . . . 20
5.1. Electromagnetic Sector Fields . . . 20
5.2. Quadrupoles . . . 21
5.3. Time-of-Flight (TOF) Mass Spectrometer . . . 22
5.4. Fourier Transform Mass Spectrometry . . . 23
5.5. Ion Traps . . . 24
5.6. Isotope Mass Spectrometer . . . 24
5.7. Accelerator Mass Spectrometry (AMS) . . . 25
5.8. Other Analyzers . . . 25
6. Metastable Ions and Linked Scans . 25
6.1. Detection of Metastable Ions . . . 25
6.2. Mass Selected Ion Kinetic Spectrometry . . . 26
6.3. Linked Scans . . . 27
7. MS/MS Instrumentation . . . 27
7.1. Triple Quadrupoles . . . 28
7.2. Multiple-Sector Instruments . . . 28
7.3. Hybrids (Sector – Quadrupoles) . . 28
7.4. Ion-Storage Devices (FTMS, Ion Traps) . . . 30
7.5. Quadrupole Time-of-Flight Tandem Mass Spectrometers . . . 30
7.6. Others . . . 30
8. Detectors and Signals . . . 30
8.1. Faraday Cage . . . 30
8.2. Daly Detector . . . 30
8.3. Secondary Electron Multiplier (SEM) . . . 32
8.4. Microchannel Plates . . . 32
8.5. Signal Types . . . 33
9. Computer and Data Systems . . . 33
9.1. Instrument Control, Automation . . 33
9.2. Signal Processing . . . 33
9.3. Data Handling . . . 35
9.4. Library Searches . . . 35
9.5. Integration into Laboratory Networks . . . 35
9.6. Integration in the World Wide Web 36
10. Applications . . . 37
10.1. Environmental Chemistry . . . 37
10.2. Analysis of Biomedical Samples . . 38
10.3. Determination of High Molecular Masses . . . 41
10.4. Species Analysis . . . 44
10.5. Determination of Elements . . . 44
10.6. Surface Analysis and Depth Profiling . . . 45
11. References . . . 46

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_515

Abbreviations

AMS	accelerator mass spectrometry
API	atmospheric pressure ionization
APCI	atmospheric pressure chemical ionization
CAD	collision activated dissociation
CIMS	chemical ionization mass spectrometry
CZE	capillary zone electrophoresis
DADI	direct analysis of daughter ions
EIMS	electron impact (ionization) mass spectrometry
ESI	electrospray ionization
FABMS	fast atom bombardment mass spectrometry
FDMS	field desorption mass spectrometry
FIMS	field ionization mass spectrometry
FTMS	Fourier transform mass spectrometry
GC/MS	gas chromatography/mass spectrometry
GDMS	glow discharge mass spectrometry
HPLC	high performance liquid chromatography
ICPMS	inductively coupled plasma mass spectrometry
ICR	ion cyclotron resonance
IDMS	isotope dilution mass spectrometry
IMMA	ion microprobe mass analyzer
IRMS	isotope ration mass spectrometry
LC/MS	liquid chromatography/mass spectrometry
LD	laser desorption
LAMMA	laser microprobe mass analyzer
LAMS	laser ablation mass spectrometry
LSIMS	liquid secondary ion mass spectrometry (see FAB)
MALDI	matrix assisted laser desorption
MID	multiple ion detection
MIKES	mass analyzed ion kinetic energy spectrometry
MS/MS	mass spectrometry/mass spectrometry
MUPI	multiphoton ionization
NCI	negative chemical ionization
PD	plasma desorption
PFK	perfluorokerosene
REMPI	resonance multi photon ionization
RIMS	resonance ionization mass spectrometry
SEM	secondary electron multiplier
SIMS	secondary ion mass spectrometry
SIR	selected ion recording
SNMS	sputtered neutral mass spectrometry
TCDD	tetrachlorodibenzodioxin
TIMS	thermal ionization mass spectrometry
TOFMS	time-of-flight mass spectrometry

1. Introduction

A mass spectrometer can be defined as any instrument capable of producing ions from neutral species and which provides a means of determining the mass of those ions, based on the mass-to-charge ratio (*m/z*, where *z* is the number of elemental charges) and/or the number of ions. Therefore, a mass spectrometer has an ion source, an analyzer, and a detector. All further details are largely dependent on the purpose of the mass spectrometric experiment.

Historically, mass spectroscopy was developed to separate atoms and to determine the masses of isotopes accurately [1]. The predecessor of magnetic sector field mass spectrometers was built in 1918 by Dempster [2]. Later, some of the first instruments for the analysis of molecules [3] were used by analytical chemists in refineries to determine hydrocarbons (the first article in the first issue of the journal *Analytical Chemistry* dealt with mass spectrometry of a mixture of hydrocarbons [4]). Combinations of infrared and mass spectroscopy were reported shortly afterwards [5].

Today, mass spectrometers may be used to determine the isotopic distribution of an element, the elemental or molecular composition of a sample, or the structure of a compound or its molecular mass. They also make it possible to study the kinetics and thermodynamics of gasphase processes or the interactions at phase boundaries. Mass spectrometers can also serve to accurately determine physical laws and natural constants.

In recent years, mass spectrometry has become one of the key technologies in proteome and genome research because of the rapid development of new components – from the ion sources to the detector – and intergration into the World Wide Web has also helped to make this possible.

2. General Techniques and Definitions

2.1. Resolution

Generally, resolution is defined by $R = m/\Delta m$, where m is the mass under consideration. This can be separated from a mass $m + \Delta m$, assuming a peak overlap of 10 %. The equation can be used either to calculate the resolution of two signals in a spectrum or to calculate the resolution necessary to separate two species of different elemental composition. For example, consider C_7H_{16} and $C_6H_{12}O$. The mass difference at $m/z = 100$ is 0.0364 amu. Thus, a resolution of ca. 3000 is required to resolve the two signals. The same result is obtained when the peak width at 5 % overlap is used for the calculation.

A different definition uses the peak width at half height as Δm (50 % definition) for the calculation (see Fig. 1). This approach is used often for Fourier transform mass spectroscopy (FTMS) and time-of-flight (TOF) instruments. In modern mass spectrometers, the resolution is calculated directly by the computer from the peak width.

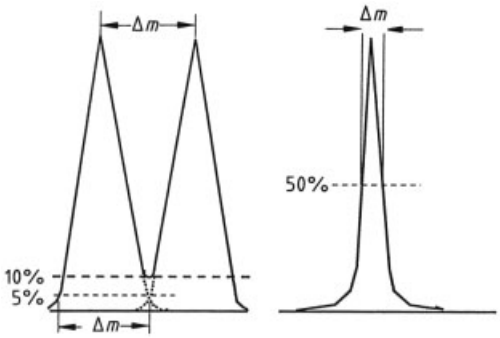

Figure 1. Resolution of two peaks, defined for the 10 % height of the valley, for one signal at the 5 % (sector field MS), or the 50 % (all others) height

More information can be extracted if the elemental composition of fragments or molecular ions can be determined. For this purpose two different requirements have to be fulfilled. First, the signal must be homogeneous, i.e., even isobaric ionic species have to be separated (resolved) and, secondly, the accuracy of the mass measurement has to be sufficient to allow a calculation with small errors.

High Resolution. The separation of isobaric ionic species (i.e., high resolution, with $R > 10\,000$), is a relative requirement. It depends on various factors, such as the type of heteroatoms, but with $R = 10\,000$ most common species can be resolved. Exceptions can be ions containing, e.g., sulfur. Increasing the resolution requires reducing the width of the ion beam in the mass spectrometer. Thus, in theory, the height of the signal is inversely proportional to the resolution. In reality, this relation is not linear over the full range of possible resolutions, and generally the loss of peak height becomes much greater when the resolution is increased above 10 000. Merely instruments like the ion trap (with very large molecules) or especially the FT mass spectrometers are capable of achieving resolution exceeding 10^5, even up to 10^6 under optimal circumstances without deterioration in detection power. The scan speed and the scan width may be affected. Thus, for applications in analytical problem solving, the resolution should be set as high as necessary yet as low as possible.

Accurate Mass Measurements. For accurate mass measurements a signal containing sufficient ions must be recorded on a precisely calibrated mass scale. In the past this was a cumbersome undertaking because each peak had to be determined separately by comparison to a known reference signal. The accelerating voltage U_0 was changed by the amount necessary to shift the signal of the mass m_x to the position of the reference m_0 on the monitor. This can be performed very accurately by switching between the two voltages U_0 and U_x (for the unknown). Since $U_0\, m_0 = U_x m_x$, m_x can be calculated from $m_x = m_0 \cdot (U_0/U_x)$. This technique, known as peak matching, is still the most precise method, and computerized versions are available.

Techniques were then developed that allow accurate mass measurement of many signals in a spectrum simultaneously. A reference compound such as perfluorokerosene (PFK) is introduced into the ion source together with the analyte in such a concentration that signals of both can be acquired simultaneously. The reference compound is chosen so that its signals are

easilyresolved from those of the analyte. The computer then searches for the reference signals in the spectrum, interpolates between them, and calculates the correct position of the unknowns on the mass scale. Sufficient precision can be achieved to allow the calculation of possible elemental compositions for the signal, provided the number of ions is sufficient. It has been shown [6] that the number of ions N required to define the peak position is a function of resolution R, where σ is the standard deviation of the mass measurement in parts per million

$$N = \frac{1}{24}\left(\frac{10^6}{R\sigma}\right)^2$$

Thus, to obtain an accuracy of 2 ppm, the number of ions required for $R = 10\,000$ is $N = 105$, but for $R = 5000$, $N = 420$. As long as the reduction of signal intensity with increasing resolution is linear, higher resolution can be beneficial, because fewer ions are required for a given accuracy. However, if the intensity of the signal is also of relevance (e.g., for quantitative determinations) more ions must be detected. Thus, in this case the optimal resolution may be lower to allow better detection power.

2.2. Tools for Structure Elucidation

2.2.1. Full Spectra: Low Resolution

In the structure determination of natural products, electron impact ionization not only produces ions of the compound itself ("molecular ions") but also results in formation of fragments, often with typical patterns (Fig. 2).

The spectra can be used to identify a compound or at least the class of compound. The spectrum of, for example, a steroid can reflect its chemical nature to such an extent that the probability of correct identification of the type of compound or even the compound itself may be acceptable. To achieve this it is sufficient to register the spectrum of fragments with integer precision. However, this means that isobaric molecules containing different heteroatoms are not separated. The success of this method for the elucidation of structures of, say, terpenes, steroids, complex lipids, and saccharides is documented in many publications and textbooks [7], [8]. The typical features of such spectra are described, and explanations are given for the characteristic fragmentation patterns. This is still the most widely used application of mass spectrometry, for which there are several reasons. First, most mass spectrometers are low resolution instruments such as quadrupoles. Second, the most common inlet systems or interfacing techniques, namely the direct probe and GC, are compatible with electron impact, giving highly reproducible spectra, largely independent of the type of mass spectrometer. Even the more modern interface for supercritical fluid chromatography or the particle-beam LC/MS interface can be used. Finally, modern library searches or intelligent data handling systems make efficient use of such data and can retrieve much of the information.

2.2.2. Elemental Compositions of Ions

The determination of the elements contained in an ion is a major aid in structure analysis. This information could be derived in principle from the isotope pattern of the molecular ion, because the isotope distribution is typical for each element. The ratios of $^{12}C/^{13}C$ and sometimes $^{16}O/^{18}O$ are the most significant parameters as well as the distinctive patterns of sulfur, silicon, the halogens, and metals in the mass range below ca. 1000 amu. Generally, however, the accuracy of the intensity measurement is not sufficient to distinguish between the various possibilities because the differences may be very small. As a rough indication, the height of the $M+1$ signal can be used to estimated the number of carbon atoms, since ca. 1.1 % of carbon is ^{13}C. Only elements such as sulfur, silicon, bromine, and chlorine with typical isotope patterns can be identified easily. For the calculation of the isotope pattern from a known elemental composition, many programs are available and can be very helpful.

Another way of obtaining information about the elemental composition is to accurately determine the mass of an ion. It is then possible to propose possible elemental compositions. These proposals can be used to support suggestions from the low-resolution data and from other means or even to elucidate the structure of a compound. However, with increasing mass (above about $m/z = 500$) the total number of possible elemental compositions for the measure-

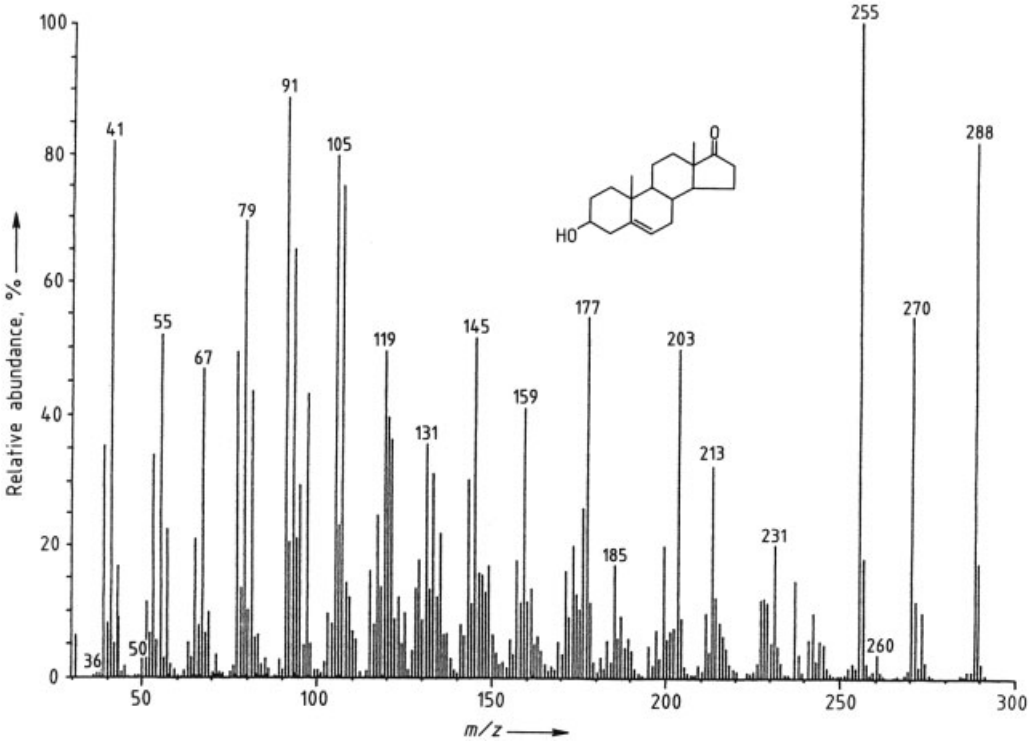

Figure 2. Low resolution ($R = 1000$) spectrum of dehydro-epi-androsteron, direct inlet using a VG 7070E, electron impact mass spectrometer (EIMS)

ments even under the assumption of very small errors becomes so great that without further information interpretation becomes impossible.

In such cases, information about the unknown compound from other sources is required. Other spectroscopic techniques or chemical evidence may be used, because the knowledge about the absence of certain elements can reduce the number of possible combinations to a manageable level. Generally, ions above $m/z = 1000$ cannot be analyzed in this way, even when the errors are very small, because the number of possible elemental compositions is far too large. There are many successful examples in the literature, mostly involving the structure elucidation of natural compounds. Accurate mass measurement has also been used in GC/MS analyses with a resolution of 10 000; in one example the structures of photoadducts of psoralens to nucleosides were determined. Since a portion of the molecules was known (thymine was the nucleoside) it was possible to ascertain the nature of the previously unknown product [9].

2.3. Fragmentation in Organic Mass Spectrometry

The formation of fragments is one of the most useful characteristics of organic mass spectrometry. This becomes particularly apparent when no fragments are available, as with soft-ionization techniques, which produce almost exclusively molecular ions with negligible fragmentation. With electron impact, still the most important ionization technique, the majority of molecules form a large number of fragments.

Organic mass spectrometry can be regarded as gas-phase chemistry. Knowledge about the chemical reactions leading to specific fragments is largely based on isotopic labeling studies, often employing metastable ions, linked scans, or MS/MS experiments. These reactions are described by the so-called quasi-equilibrium theory. This statistical theory was formulated originally in 1952 [10] to characterize qualitatively

the fragmentation behavior of gas-phase ions by using physical parameters. In Figure 3 a brief summary is given in terms of the so-called Wahrhaftig diagram. A treatment based on thermal reactions gave a similar description, often referred to as the Rice – Ramsberger – Kassel – Marcus (RRKM) theory [11]. The most important unimolecular reactions can be explained in terms of reaction kinetics and thermochemistry.

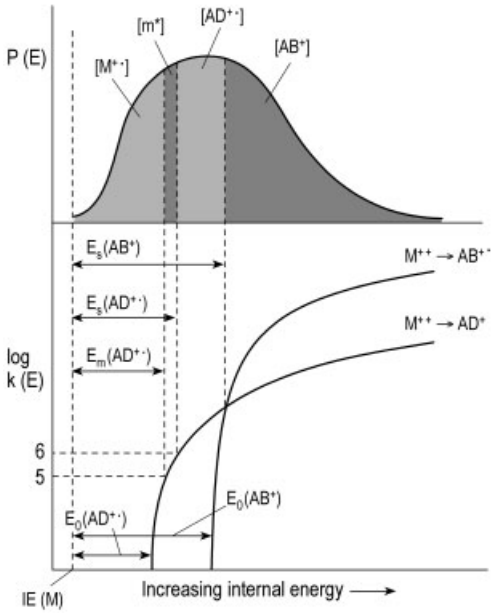

Figure 3. The Wahrhaftig diagram shows the relationship of probability function $P(E)$ (upper half) and two hypothetical functions of the rate constants $k(E)$ (lower half) for unimolecular descompositions of a molecule $ABCD^+$
$IE(M)$ = Ionization energy; $E_0(AD^+\cdot)$ = Critical reaction value for the rearrangement $M^+\cdot$ to $AD^+\cdot$; and $E_0\,(AB^+)$ = Critical reaction value for the decomposition $M^+\cdot$ to AB^+
For energies between $E_m(AD^+\cdot)$ and $E_s\ (AD^+\cdot)$ corresponding to values between $\log k = 5$ and 6, metastable ions are formed (lifetimes longer than 10^{-6} s; from then on the rearrangement dominates until the function for the decomposition crosses at $E_s(AB^+)$; the formation of the fragment AB^+ is now favored.
Adapted from [13] with permission

More simplified concepts for explaining the fragmentation patterns of complex molecules such as steroids and terpenes are based on charge- or radical-induced reactions ("charge localization"). By considering the stability of intermediate products it is possible to construct reasonable fragmentation pathways, even though the true energetic, kinetic, and thermochemical background is unclear. The greatest problem here is that many rearrangements take place, some of which are well known and discernable [e.g., the McLafferty rearrangement (Fig. 4), retro-Diels – Alder rearrangement, hydrogen shifts], but others are not obvious in the mass spectra (e.g., hydrogen scrambling, skeleton rearrangements, silyl group shifts), since the mass-to-charge ratio remains the same. Some of them can be studied by using techniques such as isotope labeling.

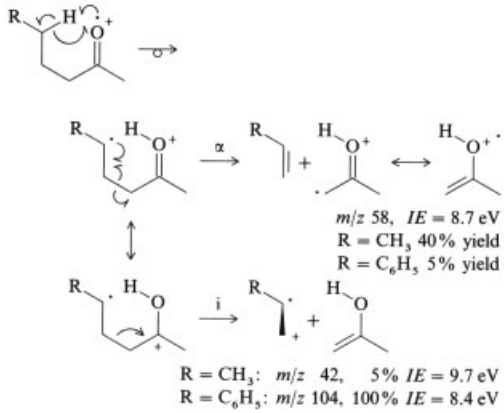

Figure 4. The McLafferty rearrangement with examples for R = methyl or phenyl and the corresponding ionization energies IE for charge retention (α cleavage) and charge migration (inductive fission) [13]

The large number of rearrangement reactions, controlled by small differences in activation energy, make the prediction of mass spectra and the understanding of the bidirectional relation between spectra and structures impossible. For a thorough discussion of these problems and related topics see [12] and [13].

2.4. Quantitative Analysis

Generally stated, mass spectrometers are unstable detectors. The rather high noise level of the baseline and the signal is a problem. The noise can be of chemical origin (i.e., undefined ions throughout the spectrum) or may have its origin in electronic components such as the voltages of the detection system. In the long term, drift of the calibration, contamination of the source or the analyzer, and decreasing efficiency of the detector can slowly degrade the performance of the instrument, resulting in poor reproducibil-

ity over long periods. In addition, especially for small signals, the ion statistics may be poor. If N is the number of ions in a peak, than the percentage variation coefficient for the measurement of the area is

$$v = 100\sqrt{\frac{1}{N}}$$

Even with 1000 ions in the signal, the area can only be measured with a variation coefficient of 1 %. Clearly the only way to improve the ion statistics for a given ion current is to increase the time spent at the top of the peak. This is possible if the associated chromatography allows slow scanning or the use of selected ion recording. For quantitative determinations, often the peak areas are used. If the separation of signals is not complete, peak heights rather than peak areas may give better results, because the overlap should be smaller.

Generally speaking, quantitative measurements require most meticulous attention to obtain high precision and good reproducibility. This requires carefully chosen acquisition conditions and, in most cases, use of internal standards.

Selected Ion Recording. Most mass spectrometers in practical work are scanning type instruments, which screen the mass range for ions by varying the voltage, frequency, or magnetic field. Therefore, a major fraction of the signal remains unrecorded, giving away the sensitivity necessary to obtain signals with a satisfactory number of ions. Techniques are available which circumvent this problem by observing only a few ions or even only one type of ion. This increases the sensitivity by two orders of magnitude but also results in a considerable loss in specificity, since a signal with a single *m/z* value makes distinction of unknown compounds impossible. Therefore, chromatographic information from the GC or LC is necessary. The technique, termed selected ion recording (SIR) or multiple (single) ion detection (MID), is in general applicable for known compounds with known chromatographic properties. In some cases, the observation of isotopic patterns can improve the certainty of identification and may help to rule out cross contamination. When the mass of one of the ions in the mass spectrum is accurately known, high resolution can be used. This brings back some specificity, although the power of detection of the mass spectrometer is reduced. Often, the improved signal-to-noise ratio due to the reduction of chemical noise more than makes up for the reduction in detection power. In cases where the compound shows strong fragmentation it may be possible to record the metastable transition signal for this reaction. The technique, termed "reaction monitoring," is specific for a given compound.

Internal and External Standards. Quantitative determinations require standards, either external or internal. External standardization using samples with known amounts of analyte ("standard addition") is possible, if a reproducibility and accuracy of ca. 10 % is acceptable. In pharmaceutical studies or routine environmental work such an accuracy is sufficient, since individual variations are much larger. When the required precision must be higher, internal standards must be used. As internal standard a compound with similar chromatographic and, if possible, similar mass spectrometric properties as the analyte can be employed. This may be a homologue, an analogue with, e.g., a different heteroatom, or a positional isomer. The best choice is an isotopically labeled compound containing stable isotopes.

Isotope Dilution. This technique is based on the fact that most elements have several stable isotopes; if only one stable isotope exists, as for phosphorus or fluorine, the technique cannot be used. Let the ratio of two isotopes of an element in the sample be R_s, and the ratio for the same isotopes in the internal standard R_I. From the ratios, the percentage peak heights can be calculated for the sample (s_1 and s_2) and the standard (i_1 and i_2). Then, a known amount I of the standard is added to the sample S and the two signals are measured (Fig. 5). The ratio of the two peaks is then

$$R = \frac{Ss_1 + Ii_1}{Ss_2 + Ii_2}$$

Rearranging for the unknown amount S of the sample gives

$$S = I\frac{i_1 - Ri_2}{Rs_2 - s_1}$$

Isotope dilution for the analysis of organic compounds is often accomplished by deuterium labeling at positions not prone to rapid exchange in the ion source. The synthesis of such compounds is normally straightforward (e.g., by exchange processes). Care must be taken, however, where the mass difference between the labeled and the unlabeled compound must be large because the isotope cluster is broad (several mass units). If many hydrogens are replaced with deuterium, the labeled compound may have a different chromatographic behavior and the retention times can differ. If deuterium labeling is difficult or not possible, ^{13}C, ^{15}N, or ^{18}O enrichment or combinations thereof can be used although the synthesis may be laborious.

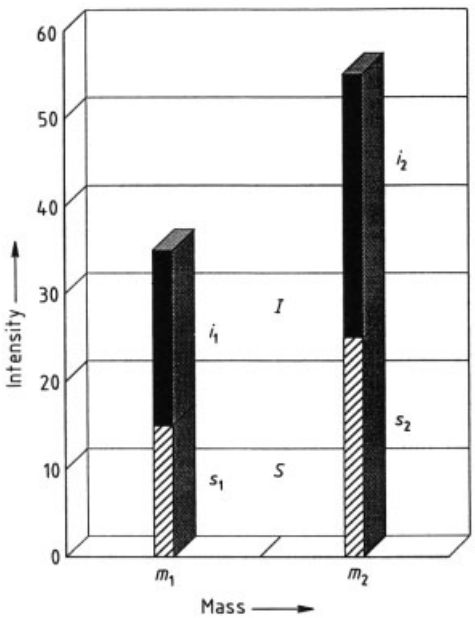

Figure 5. Isotope dilution

3. Sample Inlets and Interfaces

3.1. Direct Probe

For many years the direct probe for introducing samples into the spectrometer has dominated all the others. The sample is transferred into a small cup with a volume of a few microliters, made of gold, silica, or aluminum. Some of them have a cap with a small hole to exploit the Knudsen effect for controlled evaporation. The cup is placed in the tip of the removable probe. The probe is introduced into the ion source housing with the sample cup inside the ion volume without interfering with the electron beam. It can be heated and cooled directly under computer control, when necessary as a function of the total ion current or of a single ion current.

3.2. Batch Inlets

Batch inlets used to be important for the analysis of gases, volatile liquids, and mixtures such as hydrocarbons in oil samples. The inlet consists of a capillary, which connects a heated reservoir with a volume of several hundred milliliters, normally glass-lined to prevent cracking and pyrolysis, to the ion source. The capillary acts as a throttle, controlling the flow into the ion source; alternatively, a diaphragm can be used. The sample is introduced in the reservoir through a septum by means of a syringe or via a small sample inlet connected to the reservoir. This type of inlet has been largely replaced by GC/MS interfaces.

3.3. Pyrolysis

Pyrolysis is a powerful tool for the analysis of very high molecular mass samples, where there is no possibility for separating mixtures and determining the components or where the compounds are too complex to give mass spectra [14]. Examples are the attempts to characterize different fractions of humic acids, strains of bacteria, and unusual components in intact DNA of a variety of animals. In all such cases, mixtures of smaller molecules are prepared by carefully controlled pyrolysis (e.g., Curie point pyrolysis). The distribution of the resulting mixture of thermally stable compounds is characteristic of the starting material. Since the identity of the pyrolytic decomposition products is not known in most instances and the chemistry involved is complex, only the appearance of the particular distribution is significant, and techniques for cluster analysis based on pattern recognition must be used. Field desorption, field ionization, direct chemical ionization, and low-energy EI are mainly used with pyrolysis, because molecular ions are mostly formed and the pattern is not

further complicated by fragmentation in the ion source. The desorption process in FDMS (see Section 4.8) can be considered a pyrolysis technique. Often, the pyrolysis products are injected into a GC, separated, and soft ionization techniques such as FIMS, low energy EI, or photoionization [15] are used to detect the components of the mixture.

3.4. GC/MS Interfaces

The most successful hyphenated method continues to be GC/MS, generally with EI ionization, but particularly the inductively coupled plasma MS (ICPMS) has become a successful detector as well. In the form of benchtop instruments, this combination is found in the chromatography laboratory rather than in mass spectrometry laboratories. Instruments such as mass selective detectors and ion traps are dedicated devices for gas chromatography and have developed into robust, dependable instruments with high detection power and selectivity. With modern computerized instruments and the extensive storage capabilities of data systems, series of spectra can be acquired rapidly, and elaborate data extraction procedures have been devised for the identification of important spectra with relevant data.

The reasons for this success are numerous: Gas chromatography is a well-developed technique with high separation power and excellent stability, and its specificity can easily be adjusted to analytical problems (see → Gas Chromatography). Further, the gas flow in capillary GC is a reasonable match to the pumping requirements in the mass spectrometer. The GC/MS interface has now been developed to such an extent that compromises are no longer necessary, allowing both to be optimized separately. Today, the mass spectrometer can be regarded as a GC detector with high detection power and specificity. Compared with flame ionization or electron capture detectors, for example, it has many advantages, and the disadvantages such as higher expenses and lack of long-term stability have become less significant.

Separators. As long as packed columns with high gas-flow rates were used, it was necessary to separate the carrier gas from the analytes. This was accomplished by a variety of devices called separators; examples include the Biemann – Watson separator and the jet separator. The latter was subsequently used in LC/MS interfaces. Due to the replacement of packed GC columns by capillary columns, separators are no longer used; they are described in detail in [16].

Direct Coupling. In direct coupling, the GC capillary is introduced directly into the ion source. This technique has the advantage of simplicity in construction and it avoids losses of analyte. The disadvantage is that the chromatographic separation is influenced by the vacuum of the ion source. Furthermore, the flow of gas into the ion source is a function of the GC oven temperature. The ion source pressure and the sensitivity may vary with the temperature program. Recently, this problem has been addressed using flow and temperature programming simultaneously. It is therefore not straightforward to transfer a separation developed for GC alone to the GC/MS system, because the parameters may be vastly different. But the simplicity of construction makes this version very popular.

Open Coupling. In open coupling, the end of the capillary from the chromatograph is inserted into a small glass tee piece and the interface capillary, with a slightly wider inner diameter, is located opposite to it, separated by a small gap. Through the third entrance of the tee piece, an additional flow of inert gas (the make-up gas) protects the gap from air and adds to the flow into the spectrometer. If the dimensions are right, the complete effluate from the separation column, together with some make-up gas, enters the ion source and no loss of analyte occurs. Since the end of the GC capillary remains at atmospheric pressure, the separation integrity is maintained and analytical procedures can be transferred.

Reaction Interfaces. For special purposes it can be advantageous to add a reaction zone between the GC and the MS to convert the analytes into species that give characteristic mass spectra or more precise quantitative information. Examples are the hydrogenation of double bonds in unsaturated hydrocarbons and the combustion of organic materials to CO_2 for the accurate determination of the $^{13}C/^{12}C$ ratio. Generally, the analyte is allowed to contact a reactive surface or

catalyst, or a gaseous reagent is mixed with the gas stream from the GC. The analyzer can be any conventional sector or quadrupole spectrometer. For accurate isotope analysis, specially designed versions can be used. Since this method is important in analytical applications such as the determination of the origin of sugars in wine, dedicated instruments are commercially available. For a recent review with well chosen instructive examples, see reference [17].

3.5. LC/MS Interfaces

LC/MS interfaces can be divided into three groups: the first comprises sputtering and desorption processes, the second group has an aerosol generating step and separate ionization by EI or CI, and the third group – thermospray and electrospray – has an integrated ionization/aerosol generation step. The areas of applications for the three groups are: natural products for the first; smaller molecules, metabolites, and industrial compounds for the second; and biological oligomers and polymers for the third. Recent years have seen dramatic developments towards very stable and sensitive interfaces based on electrospray and atmospheric pressure ionization, whereas all others, except for the particle beam interface, have disappeared almost completely. The latter is still in use in industrial laboratories and for environmental studies where a detector for less polar compounds, is required.

3.5.1. Moving Belt

The moving belt system, invented by W. H. McFadden, was one of the first LC/MS interfaces. The stream of eluent from the LC is directed onto the surface of a belt made of pure nickel or polyimide, which runs through a series of vacuum stages into the ion source of the mass spectrometer. To minimize memory effects (signals from not completely removed compounds) and to clean the belt, infrared heaters and washing steps are included. The analyte was transferred into the source by thermal desorption using direct or indirect heating, by SIMS, FAB, or laser desorption. However, the technical requirements and the rather high costs (additional vacuum systems, maintenance of the belt), and the high background and memory effects limited the number of users from the beginning. The interface has the advantage of a complete separation of chromatography and mass spectrometry, allowing a wide range of chromatographic techniques (HPLC, micro HPLC, SFC) to be used.

3.5.2. Continuous Flow FAB

The continuous-flow FAB interface (online FAB, frit FAB) was developed by R. Caprioli et al. [18] and Y. Ito et al. [19]. The technique, based on the conventional FAB source (see Section 4.8.3), is straightforward and is compatible with many mass spectrometers. The exit of a capillary, which is coupled to a micro HPLC system, is connected to an FAB target and the effluent is sputtered from the target as usual. Since the gas load to the ion source is high due to the evaporating effluent, glycerol, and the FAB gas, the sensitivity is enhanced by adding high pumping capability to the source housing using an additional cryopump. Glycerol is used to enhance the desorption, as with FAB; it can be added directly to the eluent (1 – 5 %) or after separation, by means of concentric capillaries [20]. Flow rates of 15 µL/min are acceptable, but the performance is better with lower rates. Recently, capillary zone electrophoresis (CZE) has been interfaced [21], [22], although serious problems arise due to the very small amounts of sample and narrow peaks from the electrophoretic separation. Mass spectrometers with simultaneous detectors compare very favorably [23] (see Section 8.4).

3.5.3. Direct Liquid Introduction

Direct introduction of liquids was probably the first LC/MS interface to be used in modern organic mass spectrometry [24], [25]. A flow of typically 10 – 20 µL/min is allowed to enter the ion source without separation of eluent and analyte. The construction is relatively simple, and conventional electron impact ion sources can be used, even without modifying the pumping systems [26]. The main disadvantage is the inflexibility with respect to the ionizing method (for

CI only the eluent is available as reactant gas), while the compounds, that can be analyzed are similar to those amenable to GC/MS.

3.5.4. Supercritical Fluid Interface

The SFC/MS interface is a direct-introduction system with careful temperature control up to the very tip of the capillary [27]. The end of the capillary contains a frit or a diaphragm that keeps the phase supercritical until it leaves the capillary and enters the ion source. Otherwise, the dissolved analyte would clog the end of the capillary. Since the gas load is rather high, chemical ionization is normally used with capillary SFC, although EI is possible. When packed columns are used, momentum separators can reduce the pressure in the ion source, making electron impact possible. Limitations of the method with respect to classes of compounds are mainly due to SFC itself, not to the interface or the MS.

3.5.5. Particle Beam Interface

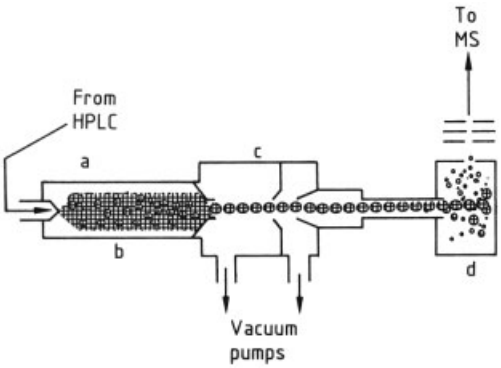

Figure 6. Particle beam interface
a) Nebulizer; b) Desolvation chamber; c) Momentum separator; d) Ion source

The particle beam interface (Fig. 6) was created under the acronym MAGIC (monodisperse aerosol generator interface for chromatography) [28]. Now, the aerosol is produced by a variety of means (with auxiliary gas, thermospray, or ultrasonic nebulizers) at atmospheric pressure and a uniform distribution of the droplets results in particles of a narrow size distribution, which can be handled more efficiently by the separator. The droplets are dried to particles in a heated expansion chamber, and a momentum separator isolates the particles from the gas. In the source, the particles are destroyed by impact and the sample is released and ionized by using EI, CI, or even FAB. The appearance of the EI spectra is almost identical to conventional EI spectra obtained by direct probe or GC/MS. Therefore, library searches are possible, which is the major advantage of this interface.

The main problem of the technique is its unsatisfactory basic power of detection. Also, sensitivity can vary widely even for similar compounds, and the difficulties in obtaining a linear response over a wider range of concentrations make quantitative determinations difficult. In addition, components eluting at the same time may interfere with one another, leading to unpredictable effects. For applications where the sensitivity is not of importance (e.g., monitoring of synthesis), this technique can be employed with success, because typical, interpretable mass spectra can be obtained.

3.5.6. Chemical Ionization at Atmospheric Pressure

Atmospheric-pressure chemical ionization (APCI) was described by Horning et al. in 1960 [29], and they recognized the potential of this approach for LC/MS applications. An aerosol is generated by spraying an LC eluate within a heated sheath gas flow to give a spray stable at atmospheric pressure. In the spray, ions are generated via a chemical ionization plasma, created by a corona discharge from a needle. The analyte is ionized by the reactant ions in the plasma and subsequently transferred into the mass spectrometer. This is achieved by a sampler – skimmer system with an intermediate pumping stage to separate the high vacuum from the atmosphere. A curtain of drying gas is used in front of the sampler to reduce the background ions.

In general, basic compounds with high proton affinity can be detected with the highest limit of detection [30], and in some cases, a strong temperature dependence has been observed [31]. In principle, APCI is compatible with every separation technique which allows the effluent to be sprayed in an aerosol. Thus, not only HPLC but also SFC can be interfaced to this ion source.

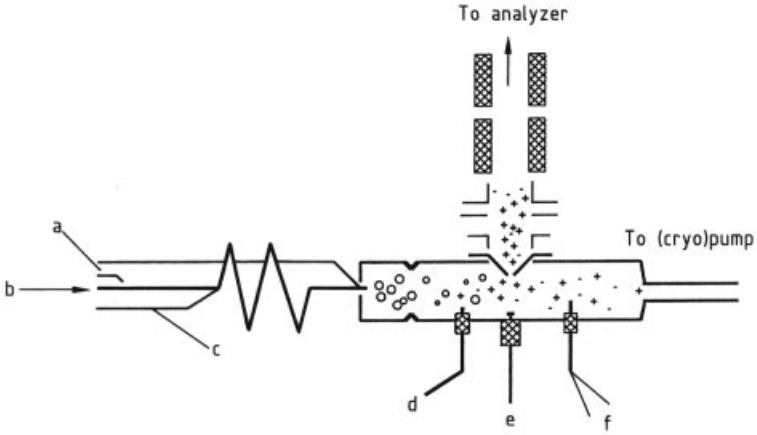

Figure 7. Thermospray interface/ion source
a) Connections for direct heating; b) HPLC flow; c) Thermocouple; d) Townsend discharge; e) Repeller; f) Thermocouple (aerosol)

3.5.7. Thermospray

Thermospray, developed in 1980 by M. VESTAL [32] rapidly found its way into routine applications, but now the atmospheric pressure interfaces have replaced it completely. The eluent of an LC is rapidly heated in a steel capillary either by direct resistive heating or by indirect heating using a cartridge system. A fine aerosol of charged droplets is generated in the ion source at reduced pressure, although the details of the formation of ions in the source is still subject to discussion [33–36]. The ions are transferred into the mass spectrometer by a sampling cone reaching into the center of the inner chamber of the ion source, but most of the solvent is rapidly pumped away. The source has to be very tight to stabilize the pressure inside and the vacuum outside. The method is shown schematically in Figure 7 [37].

It is possible to include a Townsend discharge or a filament and a repeller electrode to allow plasma chemical ionization to be performed. For some compounds, the sensitivity, selectivity, and fragmentation behavior can be modified. The advantage of the technique is the simplicity of transferring previously developed LC procedures since the normal flow rates and some of the common buffers (e.g., ammonium acetate, formate) can be used.

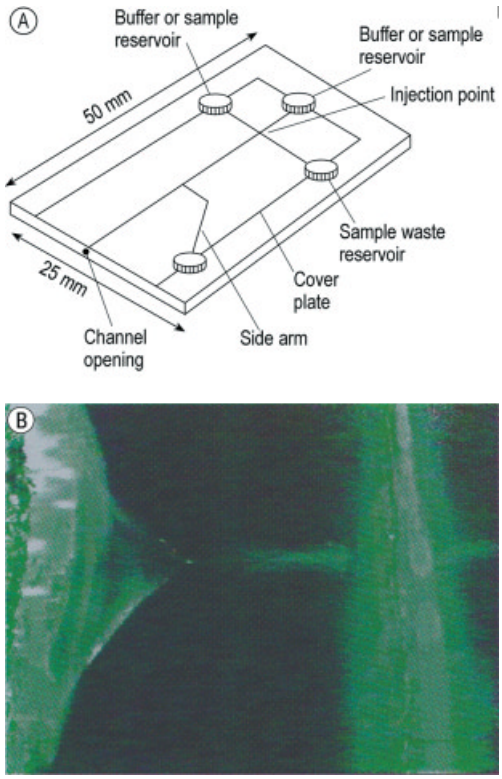

Figure 8. (A) Schematic diagram of microchip used to pump fluids electroosmotically and to generate the electrospray; (B) Photomicrograph of the Taylor cone and electrospray generated from a 60 % H_2O – 40 % methanol solution that was electroosmotically pumped using a microchip (reproduced from RAMSEY [51])

3.5.8. ESI Interface

The combination of the electrospray ion source with HPLC has without a doubt become "the" LC/MS interface in recent years. It is a particularly powerful combination, since this ionization technique covers a wide range of samples [38] that are commonly separated by HPLC [39] or electrophoresis [40–44]. The ESI source exhibits concentration-dependent behavior and thus gives optimal signals at most flow rates. The principle of the ionization process is discussed in Section 4.9. The most important feature of this interface is aspray needle which can be connected directly to the separation column, if the flow rates are compatible. Initially the major limitation was that only low flow rates (a few μL/min) could be used, but now flow rates of 1 mL/min or more are possible by using heated sprayers or ultrasonic devices. Splitting of the flow is possible as well, allowing two detectors to be used simultaneously. Since buffers can be used as long as they are volatile and not too concentrated, a sheath flow of methanol can be added to enhance the ion yield and reduce background due to buffer salt clusters. Interfacing with capillary zone electrophoresis (CZE) has been demonstrated [45] and this combination appears to be a powerful tool for the analysis of trace organic compounds [46], although the amount of analyte which can be separated in a CZE capillary and transferred to the mass spectrometer is very small and may even lie below the detection capabilities of the mass spectrometer [44], [47–50]. The near future will see devices used in many fields based on a chip with separation by CE or HPLC and an integrated sprayer, especially in biochemical and clinical analysis [51], [52]. In Figure 8 such a device is shown (a), and a photograph of the Taylor cone developed from the chip (b).

3.6. TLC/MS

Two-dimensional separations such as thin layer chromatography (TLC) or gel electrophoresis in combination with SIMS [53], FAB [54], and laser desorption [55] have given impressive results and are promising for the future [56]. In principle the technique is identical to FAB or laser desorption, with the crucial difference being that the target is the device used for the separation of the sample mixture. There are problems associated with this difference: the sample may be rather dilute, not in an appropriate matrix and, because the layers are rather thick, much of the sample is protected from the sputtering beam. It is therefore necessary to optimize the TLC procedure, e.g., by using micro rough surfaces for the separation or blotting techniques to transfer the analyte to a different matrix. To allow the scanning of larger TLC plates, the incorporation of sample stages with x,y-adjustments would be necessary. The major advantage of the technique is that method development of TLC methods is rapid, efficient, and cost effective, and the plates can be investigated with several spectroscopic techniques sequentially after separation, making this setup very attractive for routine analysis (e.g., in environmental chemistry) [57], [58]. For biochemistry, matrix assisted laser desorption of peptides, proteins, and oligonucleotides separated in thin-layer electrophoresis and often transferred to a more suitable matrix by blotting techniques appears promising, e.g., for screening purposes and the identification of modifications and mutants [59–62], even though excision of the spots and off line analysis is often favored. This strategy allows direct enzymatic digest ion in the gel avoiding sample loss in the course of transfers.

4. Ion Generation

4.1. Electron Impact

Electron impact is still the most widely used technique for generating ions. It uses electrons, emitted from a rhenium or tungsten wire or ribbon and accelerated with typically 70 V, to ionize the vapor of a compound. In this process, molecular ions are formed together with a pair of electrons after impact. The molecular ion may be stable or it may dissociate immediately into a neutral and an ionic species (dissociative ionization) or into two ions of opposite charge (ion-pair formation). The positive ions may have sufficient internal energy to undergo further fragmentation. Generally, the fragments are formed in the ion source within a few microseconds and are responsible for the normal mass spectrum. If

ions are stable enough to dissociate later in one of the field-free regions between analyzers, they appear as "metastable ions" or may be detected by using MS/MS experiments. These metastable ions have been studied in great detail, since they facilitate the understanding of many gas-phase processes. The EI ion source is a small chamber with a filament (cathode), an electron trap to regulate the emission, and a pusher or repeller, which controls the space charge distribution in the chamber. Small permanent magnets perpendicular to the electron beam broaden the electron cloud, which enhances the sensitivity. The ions are extracted by means of a set of extraction plates and focused with several lenses and steering plates onto the entrance of the analyzer.

4.2. Chemical Ionization

Chemical ionization can be carried out either at atmospheric pressure or under reduced pressure in an ion source chamber; in both cases ion/molecule reactions are used to produce the ions of the analyte [63]. At atmospheric pressure, the plasma is generated with a discharge needle and the ions come mainly from the surrounding air, water, and from components of the analyte spray, which can contain organic solvents and buffers. Mainly protonation occurs. Under reduced pressure, a reactant gas must be added to the mixture. For this purpose, a gas such as methane, isobutane, ammonia, or water is introduced into a gas-tight ion source. This gas is then ionized by electron impact (or, sometimes, by a Townsend discharge) to give a reactive plasma. The nature of the plasma ions determines which type of spectrum is formed; therefore, the different gases can be used to control the specificity of the ionization process. The reactions most often observed are proton transfer from the reagent ion to the analyte, charge exchange, electrophilic addition of a plasma ion to the analyte, and abstraction of a negative leaving group from the analyte. When ammonia is used, only substances with high proton affinity are protonated, and hydrocarbons, for example, do not show up in the mass spectra. With CO_2 as reagent gas, the spectrum resembles a normal EI spectrum due to the high energy associated with the electron transfer in the CI plasma. The important feature of chemical ionization with respect to analytical applications is that the ratio of molecular ions to fragments is usually much higher than with EI, thus yielding information about the molecular mass.

4.3. Negative Chemical Ionization (Electron Capture)

Negative chemical ionization is similar to CI only in so far as the same gases are used in the same ion source, although the pressure is usually lower than in CI. The purpose of the gas is to moderate the energy of the electrons emitted from the filament to thermal energies. With such slow electrons, electron capture becomes possible, resulting in the formation of negative ions. If the analyte can stabilize the negative charge either in electronegative moieties, by charge delocalization, or by dissociation to give a stable fragment, the cross section of the ionizing process can be high, and the technique very sensitive. For the analysis of substances unable to accommodate negative charge, special derivatization techniques such as electrophoric labeling are used. An example is the derivatization of adducts of carcinogenic metabolites to nucleosides with perfluorobenzyl bromide yielding GC/MS detection limits in the attomole range [64].

4.4. Field Ionization (FI)

Field ionization, developed by H. D. Beckey [65], was of importance only for a short time. In the ion source, a strong electric field gradient (10^7 V/cm) is formed by means of a potential applied to a sharp edge, a tip, or an emitter with whiskers; the potential difference between the acceleration voltage and field potential is ca. 11 kV. In the gradient, the vaporized molecules of the analytes travel to the regions of increasing field strength. When the probability of an electron leaving the molecule by tunneling through the internal energy barrier becomes sufficiently high, ions can be observed. The ionization process is very soft. Since the technique is technically difficult and needs skilled operators, only a few laboratories have used FI for practical work. Its strength is that only molecular ions are

formed, allowing, for example, the distribution analysis of a series of homologous hydrocarbons in oil samples. A special version of the technique, called field ion kinetics, has been used for some time to study the very fast fragmentation processes of several classes of compounds [65].

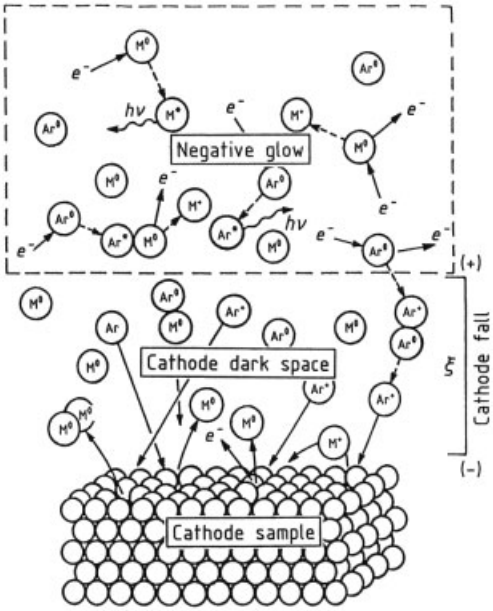

Figure 9. Principles of GDMS illustrating the main processes of sputtering, excitation, and ionization
M^0 = Sample atom; M∗/Ar∗ = Excited states; M^+ = Sample atom
Only neutrals can cross into the negative glow, where ions are formed via electron impact or in collisions with excited argon molecules (Reproduced with permission from [68])

4.5. Plasma Ionization

Plasma ionization can, in general, only be applied to the analysis of elements, since the energy of the plasma is too high for most molecules to survive. However, molecules formed from the plasma gas and the analytes in the plasma are one of the main obstacles of this ion source type. Interferences in the lower mass regions – others come from doubly charged ions and oxides – complicate the spectra and may even obscure some elements completely if only trace amounts are present. High-resolution mass spectrometers can be used to solve this problem. Three major types of plasma source are in use: the glow discharge source, the inductively coupled plasma source, and sputtered neutral mass spectrometry [66].

Glow Discharge (GDMS). Glow discharge ion sources have been built for quadrupole and sector field instruments. The solid sample is part of the source chamber and operates as cathode. The source chamber serves as anode and is filled with noble gas (generally argon) at a reduced pressure (1 kPa) that allows a stable plasma to be formed. A high voltage is applied between cathode and anode resulting in a self-sustaining discharge. Two main regions exist in the plasma: the cathode fall next to the sample and the negative glow (Fig. 9).

In the cathode fall, argon ions are accelerated toward the sample surface and atoms from the sample are sputtered into the plasma. Upon electron impact, the sputtered atoms are ionized and extracted into the mass spectrometer by application of a potential [67]. The technique is considered to suffer only minor matrix effects thus giving reliable quantitative results. Since the sample is one of the discharge electrodes, the thermal and electrical conductivity of the sample may influence the plasma and the sputtering process. For the analysis of nonconducting materials it may be necessary to mix the sample with a good conductor.

Sputtered Neutral MS. In sputtered neutral MS (SNMS) an electron cyclotron resonance in a static magnetic field is used. Under these conditions, an inductively coupled RF plasma is formed in high vacuum with argon as plasma gas; since no electrodes are involved the electrical properties of the sample do not influence the plasma, allowing the analysis of metals, particulate materials, and ceramics. The sample is connected to an adjustable high voltage, and argon ions from the plasma are accelerated onto the surface of the sample. The intensity of the sputtering process can be varied by means of the high voltage and thus the depth resolution obtainable can be optimized for the analytical problem. Since only atoms can penetrate into the potential – sputtered ions being reflected – the electron impact ionization step is virtually independent of the sputtering process and the technique is therefore free of matrix effects. In addition, ions formed at plasma potential are suppressed by an

energy filter. Therefore, the technique gives reliable quantitative results when calibrated with standard reference materials. The best resolution that has been achieved is 4 nm. Since the sputtering process is slow, the usable range for maximum depth is in the micrometer range. Consequently, the technique has some attraction for the analysis of semi- and superconductors [69].

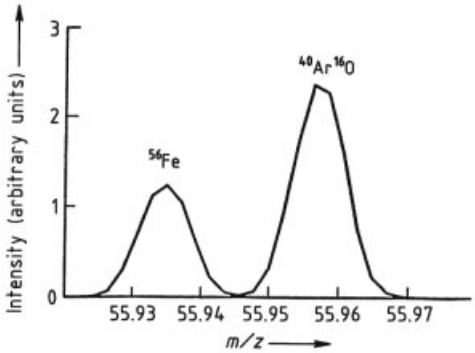

Figure 10. The spectrum shows the ^{56}Fe and $^{40}Ar^{16}O$, resolved with a resolution of $R = 3000$, using a laboratory prototype of the ELEMENT, Finnigan MAT (high resolution ICPMS) [75]

Inductively coupled plasma mass spectrometry (ICPMS) is very sensitive and allows the analysis of solutions of metal ions at pg/mL levels [66]. In principle, the ion source consists of an aerosol generator, a desolvation unit, the inductively coupled plasma, and the ion transfer and focusing optics. Several types of aerosol generator, such as the Babington and Meinhard types, ultrasonic nebulizers, and thermospray are used [70] . For interfacing ICPMS to LC/MS, efficient drying of the aerosol is necessary to stabilize the plasma. The most effective method appears to be heating and subsequent cooling of the aerosol, sometimes supported by a counterflow of gas. The ions are generated in an argon plasma, which burns in the high-frequency field of the induction coil just in front of the transfer optics. Since this all takes place at atmospheric pressure, the sampling into the mass spectrometer by means of a skimmer sampler arrangement, similar to that of other API sources, requires very effective pumping to ensure sufficient sensitivity, and the outer skimmer must withstand the temperature of the plasma. ICPMS instruments are available with quadrupole and sector field analyzers [71] ; the latter is of interest, when higher resolution (typically ca. 7000 – 10 000) is necessary to separate molecular interferences due to oxides and argides [66], [72–74] (see Table 1 and Fig. 10).

4.6. Thermal Ionization

Thermal ionization is a technique for elemental analysis that uses thermal energy to ionize elements. The sample is deposited on a ribbon which can be heated to very high temperature, and a second ribbon a few millimeters from the first is used to ionize the atoms in the gas phase. This ribbon is made from rhenium or tungsten, sometimes treated with carbon or thorium, depending on the nature of the element under investigation. Heavy metals such as lead, uranium, thorium, and others have been determined as positive ions in very different matrices such as geological samples, microelectronic chips, and ice cores from the antarctic, in conjunction with isotope dilution (IDMS) techniques. Iodine, bromine, chlorine, selenium, and also some heavy metals have been determined by using negative thermal ionization and IDMS [76]. Thermal ionization with isotope dilution is considered one of the reference techniques allowing the most precise quantitative determination of elements.

4.7. Optical Methods

Photo-Ionization. Photo-ionization has been used mainly to study basic principles of the ionization process. The light source is usually a helium lamp emitting 21.22 eV photons; other noble gases can be used to produce other photon energies [77]. When the energy deposited in the molecules is sufficient to exceed the first ionization potential, ions of the analyte are formed. By analyzing the energy distribution of the photoelectrons, the energy distribution of the molecular ions can be deduced. It has been shown that the distribution formed by electron impact may be similar.

Multiphoton Ionization. The specificity of multiphoton ionization (MUPI, sometimes termed resonance multiphoton ionization,

Table 1. Some ICPMS interferences and the resolution required to resolve the signals

Analyte species	Exact mass	Interfering species	Exact mass	Resolution required for separation
^{55}Mn	54.938	$^{40}Ar^{14}N^{1}H$	55.090895	359
^{28}Si	27.976929	$^{14}N_2$	28.006146	960
^{31}P	30.9737634	$^{14}N^{16}O^{1}H$	31.005813	966
^{48}Ti	47.9479467	$^{32}S^{16}O$	47.986986	1 228
^{44}Ca	43.95549	$^{12}C^{16}O^{16}O$	43.98982	1 280
^{32}S	31.9720727	$^{16}O_2$	31.989828	1 802
^{56}Fe	55.934938	$^{40}Ar^{16}O$	55.957299	2 504
^{51}V	50.944	$^{35}Cl^{16}O$	50.96376	2 580
^{62}Ni	61.9283	$^{46}Ti^{16}O$	61.94754	3 220
^{75}As	74.921596	$^{40}Ar^{35}Cl$	74.9312358	7 771
^{80}Se	79.91647	$^{40}Ar_2$	79.92477	9 638
^{48}Ti	47.94795	^{48}Ca	47.95253	10 466
^{58}Ni	57.9353	^{58}Fe	57.9333	29 000
^{115}In	114.903871	^{115}Sn	114.903346	218 900
^{87}Rb	86.90916	^{87}Sr	86.908892	325 000

REMPI) is based on the existence of well-defined electronic states at low temperature [78], [79] . If a molecule is cooled to a few kelvin, a first photon can be absorbed upon resonance and excites the electrons to a state below the ionization potential. This absorption process is specific for a compound or class of compounds; for example, isomeric methylphenols can be distinguished with such a technique. A second photon is necessary to excite the molecules above the ionization threshold. By varying the laser power, the fragmentation can be controlled (see the example given in [80]). The more energy is available the greater the fragmentation.

In the ion source, a supersonic beam of noble gas is generated and the sample is seeded into it (e.g., by laser desorption). Together with the beam, the sample is carried through a skimmer system that allows only the isentropic core to enter the vacuum chamber where the ions are generated by two tunable lasers. The technique is still in the development stage.

4.8. Desorption Methods

Desorption methods have been developed since around 1970. They have not only expanded the scope of mass spectrometry, but have also changed the strategies of sample introduction and brought attention to solid and liquid phase chemistry as well as to particle and gas phase chemistry. They have also stimulated the rapid development of new ion sources, mass analyzers, and detectors with higher mass range capabilities and improved sensitivity.

4.8.1. Secondary Ion Mass Spectrometer (SIMS)

Secondary ion mass spectrometry, pioneered by A. Benninghoven et al. [81], can be regarded as one of the first desorption techniques. A primary ion beam (typically ions as Cs^+ and O^-) is focused onto a target; the sample can be deposited on the surface of the target or the surface is the sample itself. Upon impact, the sample is sputtered and ionized simultaneously, and the secondary ions are extracted into a time-of-flight instrument. Since the primary beam can be focused to a small spot, the technique can be used in principle to achieve spatial resolution as well (ion micro probe) [82]. SIMS is used in material science to analyze the surface of metals, but due to the combination of the sputtering and ionization processes, matrix effects can be strong and atomic mixing occurs. Although this approach is used primarily for the determination of elements, the investigation of organic surfaces is also possible [83], and organic materials deposited on a appropriate surface such as silver have been studied [84].

4.8.2. Field Desorption (FD)

Field desorption is technically closely related to field ionization, although the desorp-

tion/ionization process is different [85]. The sample is deposited directly on the emitter to produce the high field gradient and, therefore, materials with virtually no vapor pressure can be investigated, in contrast to field ionization. The emitter can be heated either by direct current or by using external sources (e.g., lasers). The sample is generally desorbed as so-called "preformed ions" (protonated or with alkali cations attached, both or either type are always present in samples). The ion currents are generally very weak and often erratic; therefore, integrating devices (photoplates, multichannel analyzers) have been used to overcome these problems. Since FDMS and FIMS are similar methods, they have the same technical problems. FDMS was the first desorption technique to allow mass spectrometric analysis of peptides, oligonucleotides, and polysaccharides. Today the technique is used for the analysis of industrial polymers.

4.8.3. Fast Atom Bombardment (FAB)

Fast atom bombardment, developed by M. Barber in 1980 [86], was the first technique which could be used with virtually every mass spectrometer, since the ion source can be retrofitted to a given mass spectrometer without problems. Furthermore, sample preparation is straightforward. The technique is based on SIMS, with the major difference being that the sample on the target is dissolved in a liquid matrix, usually glycerol. In the meantime, many other matrices have been examined for the analysis of different classes of compounds. The most widely used are glycerol, thioglycerol, "magic bullet" (a 5:1 mixture of dithiothreitol/dithioerythritol), 3-nitrobenzyl alcohol, 2-nitrophenyl octyl ether and triethanolamine. It can be beneficial to add acid (e.g., phosphoric acid, sulfonic acids) to glycerol to increase the abundance of analyte ions. The second difference to SIMS, the use of accelerated atoms instead of ions, is of importance only in so far as an atom gun can be added to the source housing of even sector field instruments. Because atoms can penetrate the high electric field but cannot be focused on a small area, the atom gun has often been replaced by a high-energy ion gun in a technique known as liquid SIMS (LSIMS). The advantage is that the primary ions can be focused and deliver a higher ion density in comparison to the saddle field atom guns commonly used with FAB MS.

4.8.4. ^{252}Cf Plasma Desorption

Plasma desorption (PD), first described as a tool for protein analysis in 1976 [87], exploits the fact that ^{252}Cf "explodes" spontaneously into smaller nuclides, the lighter ones centered around ^{108}Tc, and the heavier ones around ^{144}Cs. The products from each unsymmetric fission contain most of the high energy as kinetic energy and travel in opposite directions. The californium source is located between a thin foil loaded with the sample and a target, which serves to start the clock of a detector in a time-of-flight mass spectrometer [88] (see Fig. 11). When the heavy atom hits the foil, a cloud of atoms and neutrals, including intact large molecules, are desorbed in a very short pulse. The ions observed are preformed ions (protonated or with alkali metal ions), which are either desorbed directly or formed in the dense cloud of desorbed material.

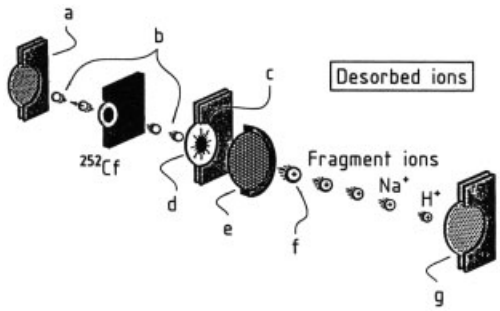

Figure 11. Plasma desorption MS
a) Fission fragment detector; b) Fission fragments; c) Excitation area (100 nm^2); d) Sample foil; e) Acceleration grid; f) Molecular ion; g) Desorbed ions detector

The technique gives poor resolution, but good mass accuracy and high detection capability, although strong background problems must be solved. The time required to record a spectrum depends on several factors (sample preparation, amount of ^{252}Cf available) and can vary from a few minutes to several hours. The technique has already experienced the same fate as FDMS. When it was developed, it was highly praised

for its unprecedented high molecular mass determination capabilities, its sensitivity, and simplicity. The results for peptides [89] and other high mass polymers were equally impressive. This type of work is now largely done by matrix-assisted laser desorption and electrospray.

4.8.5. Laser Desorption/Ionization

Lasers have been used in mass spectrometry for many years. Trace elements in biological samples [90] can be determined by using laser microprobes (LAMMA, laser microprobe mass analyzer) or a combination of laser ablation with ICPMS. For the analysis of bulk materials, techniques such as resonance ionization mass spectrometry (RIMS) and laser ablation MS (LAMS) are employed; for a review see [91].

The desorption of organic substances from surfaces without destruction is difficult and the duration of ion production is short. Therefore, the mass limitation of this approach without a matrix is around 1000 amu [92]. Recently, molecular ions of complete proteins with $M_r > 200\,000$ [93] have been generated, which is mainly due to the use of appropriate matrices for matrix-assisted laser desorption/ionization [94]. For sample preparation, the analytes are dissolved and crystallized in a matrix on a target. The matrix for the first experiment using a UV laser for protein analysis was nicotinic acid [95], but the search for other matrices is continuing, since the nature of the matrix seems to be crucial. A similar approach with a metal powder (10 nm diameter) mixed with a liquid matrix (glycerol) to form a colloid gave comparable results [96]. The crucial point is the absorption of the matrix, which protects the large molecules from the laser energy. Apparently, a variety of ions and radicals are formed initially from the matrix by photoionization and, in the dense expanding plume of desorbed material, protons become attached to the analyte molecules. These ions are then extracted in the TOF analyzer. The first lasers used were UV lasers tuned to the absorption of nicotinic acid, but later it was shown that IR lasers give similar spectra [97] . Due to the pulsed ion generation, TOF instruments are the analyzers of choice, although more recent attempts have used sector field instruments with simultaneous detectors. The signals obtained are rather broad, probably due to matrix – molecule adducts of the analytes, but the detection potential is very good. Since the accuracy of the mass measurement is ca. $1\times10^{-3} - 5\times10^{-3}$, yielding an error of ca. 100 – 500 amu at $M_r = 100\,000$, the precision of the technique is superior to other methods such as gel chromatography or centrifugation. Today, benchtop instruments are available with easy sample handling and short analysis times, making the technique suitable for biochemical laboratories. The development of delayed extraction has improved the resolution considerably [98–100] and with post-source decay [101–103] the detection of fragments has also become possible.

4.9. Electrospray

Electrospray was described in 1968 by Dole et al. [104] and the ions were detected by ion mobility. Fenn and coworkers [105], [106], and, independently, Alexandrov et al. [107] readopted the idea as a means of creating ions for mass spectrometry and since then rapid development of this technique has occurred. The fundamental difference to thermospray is that the spray is formed by charging droplets to such an extent that they explode by coulomb repulsion into smaller and smaller droplets. The droplets are further decreased in size by collisions with gas and, finally, highly charged ions are liberated. The process in which ions are created from the droplets is still not clear and several partial models exist [108].

The electrospray source is shown schematically in Figure 12. A capillary is held under high voltage (3 – 5 kV) at atmospheric pressure to generate a spray of charged droplets. To obtain a dry aerosol, either a heated steel capillary [109] or a hot gas curtain [110] is used in conjunction with collisions in the first vacuum stage. These collisions are necessary to decluster the ionic species, but may also be used to make fragments in a process equivalent to collision-induced dissociations for MS/MS experiments.

For quadrupole mass spectrometers, the ions from the first stage are extracted and pass directly into the mass spectrometer through a skimmer. For sector field mass spectrometers,

a second vacuum stage is required to minimize

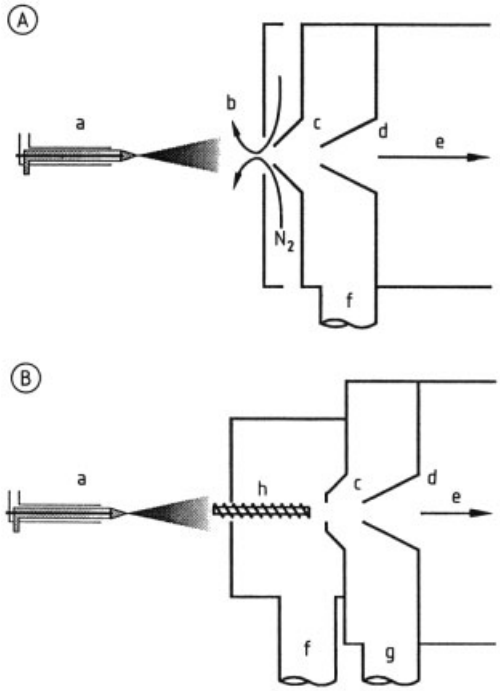

Figure 12. Electrospray ion source types
A) Desolvation region with gas curtain; B) Desolvation region with heated capillary
a) Sprayer with three concentric capillaries: the inner flow is the eluent from the separation (CZE, HPLC), the next flow is sheath liquid such as methanol or isopropanol to enhance desolvation, and the outer flow is sheath gas, normally nitrogen; b) Gas curtain; c) Nozzle; d) Skimmer; e) To ion optics and analyzer; f) To roughing pump; g) To turbomolecular pump; h) Heated capillary

high-energy collisions in the acceleration region, which could destroy the ions. The main advantage of electrospray is that even molecules with very high molecular mass can be detected at low *m/z* values, because the ions are highly charged (charges of up to 100 have been observed). Thus, the use of inexpensive quadrupole instruments is possible, but the more powerful analyzers such as FTMS or the new hybrid spectrometers have demonstrated superior performance (vide infra). The range of commercially available ion sources with new design features such as the orthogonal sprayer, developed by Hewlett Packard, (Fig. 13) or similar devices allow connection with HPLC even with high salt loads. The new low flow sprayer has increased the sensitivity for microHPLC and capillary electrophoresis and the nanospray sources for biomedical work [111] have become pivotal in proteomic research.

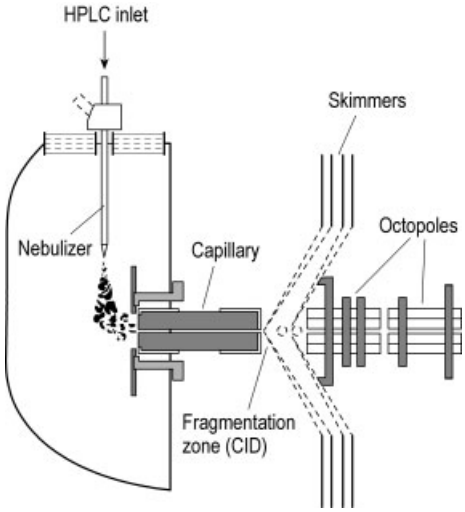

Figure 13. Schematic of the orthogonal sprayer assembly ESI Ion Source and transfer optics (Hewlett Packard)

5. Analyzers

5.1. Electromagnetic Sector Fields

The first mass spectrometer [112] was a single focusing instrument with a magnet deflecting the ions through 180°. Since then, magnetic fields alone and in combination with electrostatic fields have been used most successfully to separate ions. When an ion of mass *m* has the charge *ze* where *z* is the number of charges and *e* is the elemental charge, its kinetic energy after acceleration through a potential *V* is

$$\frac{1}{2}mv^2 = zeV$$

where *v* is the velocity after acceleration. In a magnetic field of strength *B*, the force is *Bzev*. For the stable radius *r*

$$Bzev = \frac{mv^2}{r}$$

Combining the equations yields

$$\frac{m}{z} = \frac{B^2 r^2 e}{2V}$$

Thus, for any given *m/z* value a certain *r* is valid, if *B* and *V* are constant. Scanning the magnetic

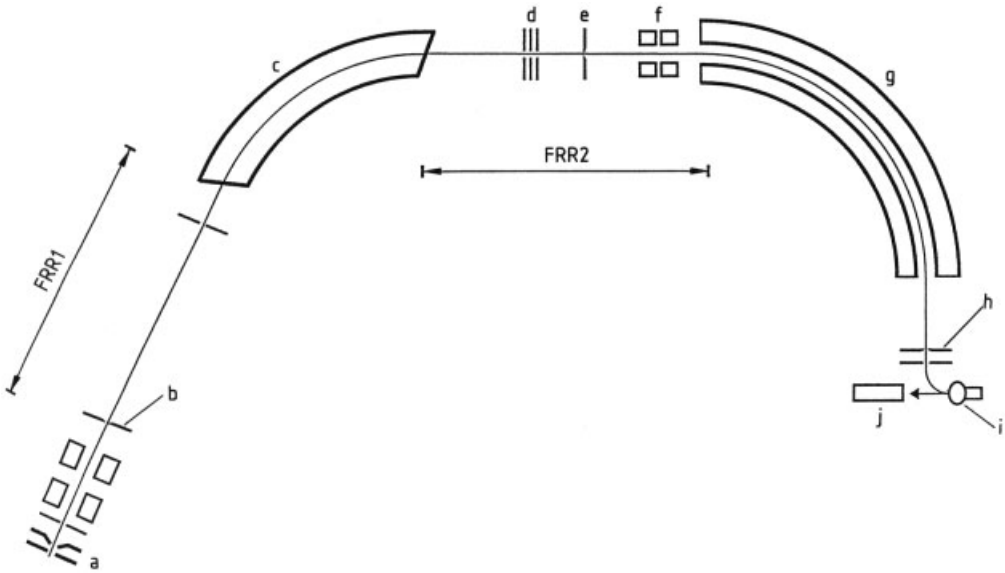

Figure 14. Double focusing MS (reversed Nier – Johnson geometry)
a) Ion source; b) Entrance slit; c) Magnetic sector; d) Focusing elements; e) Energy slit; f) Beam Shaping static quadrupoles; g) Electrostatic analyzer; h) Exit slit; i) Conversion dynode; j) Multiplier; FFR1, FFR2, Field-free regions

field or the voltage, on the other hand, allows different ions to travel the same path through the magnet and reach the detector consecutively. The resolution attainable with a magnet is limited to ca. 5000 by the initial energy spread of ions due to the Boltzmann distribution and as a result of inhomogeneous and fringe fields. It can be increased by adding an electrostatic field so that the ions are focussed in both direction and energy ("double focusing"). An example is given in Figure 14.

Various designs exist, with the electrostatic field preceding or following the magnet [Nier – Johnson (BE), reversed geometry (EB), Herzog – Mattauch], and all of them differ in performance. The advantage in all cases is a far superior resolution (10 000 – 100 000) compared to magnets alone. But another degree of complexity is added to the mass spectrometer and there are some other handicaps as well. In comparison, the instruments are expensive, and more difficult to operate and maintain. In addition the long-term stability of calibration is rather poor and some of the interfacing techniques are quite cumbersome to carry out. Nevertheless, sector field instruments are, generally speaking, still the most versatile mass spectrometers with strengths in detection power, mass accuracy, resolution, and flexibility [113].

5.2. Quadrupoles

Quadrupoles are the most successful concept from a broad range of so-called dynamic mass spectrometers [114]. They differ from static analyzers in that they use a high-frequency voltage to disperse the ions. In a quadrupole four poles (rods) form a hyperbolic field through which the ions travel with low kinetic energies. The superposition of high frequency and a variable d.c. voltage enables the quadrupole to separate the ions according to their mass. The regions of stability and instability are shown in Figure 15. The coefficients a and q are derived from the Matthieu equations. They are functions of the d.c. voltage U and the a. c. voltage $V_0 \cos wt$. For a fixed frequency, m is proportional to both U and V_0 and when the ratio $a/q = 2U/V_0$ is held constant, the masses follow stable paths and a spectrum may be recorded.

Although mathematical treatment of the separation process is complicated, the operation of such a spectrometer is simple. They have quickly become established in GC/MS systems, since

they can be very compact, easy to handle, and less expensive than other instruments. In addition, computer control is simple. This advantage became even more significant with the development of triple quadrupole instruments (see Section 7.1).

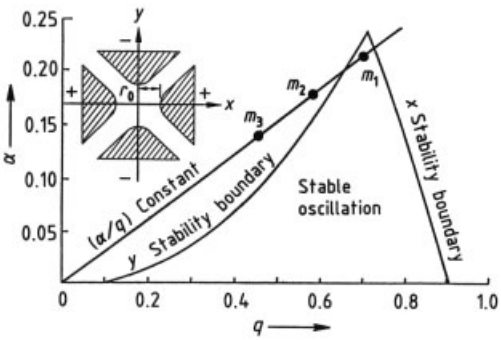

Figure 15. Quadrupole stability diagram indicating the conditions for stable and unstable oscillations

5.3. Time-of-Flight (TOF) Mass Spectrometer

A few years ago, time-of-flight instruments were used almost exclusively for surface analysis. Then, the development of plasma desorption and laser desorption/ionization changed the situation dramatically. It is now conceivable that in the future TOF analyzers may become routine analyzers. The mass analysis is based on the fact that after uniform acceleration in the ion source, small ions arrive earlier at the detector than heavy ions. This is because all particles acquire the same energy resulting in different velocities. If the acceleration voltage is V, the energy is eV. For an analyzer of length L, the time to travel to the detector is

$$t=\sqrt{\left(\frac{m}{z}\right)\left(\frac{1}{2eV}\right)}L$$

and, solved for *m/z*

$$\frac{m}{z}=\frac{2eVt^2}{L^2}$$

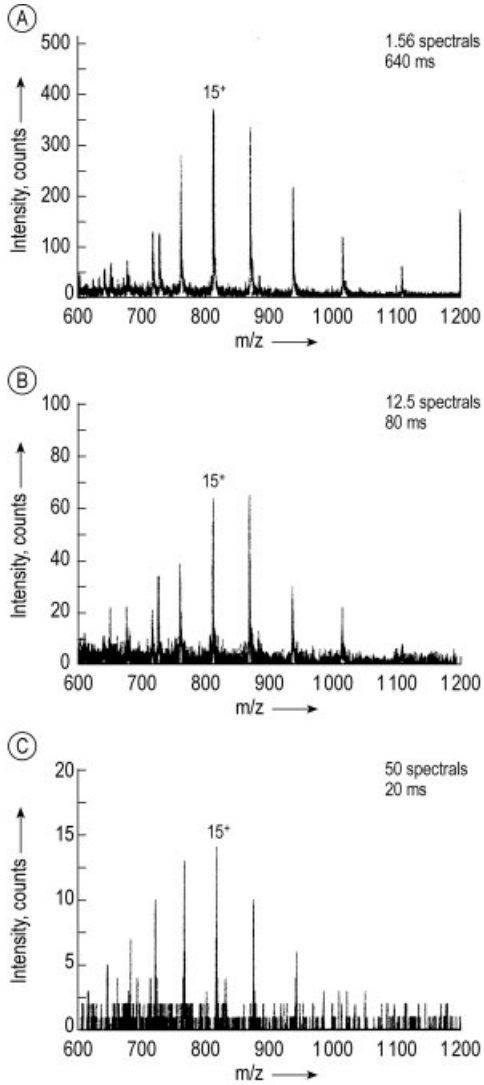

Figure 16. TOFMS of cytochrome *c*, concentration 0.1 μM in the solvent; Data aquisition: 5000 Hz; Spectrum A 3200, B 400, and C 100 summed scans

The detectors, normally multichannel arrays, are capable of resolving the resulting small differences in time. Almost every ion leaving the source can be recorded, giving TOF instruments excellent sensitivity. To enhance resolution, a reflector device can be added to the linear flight tube. This reflectron focuses the ion beam on the detector, compensating for the initial energy spread of the ions, because the faster ions penetrate more deeply into the reflector than the slower ions and thus have a somewhat longer path to the detector. Resolutions in the range

of 5000 – 10 000 have been obtained and ions with $m/z > 100\,000$ have been recorded. Alternatively, the reflector can be used to determine the sequence of biopolymers by means of laser desorption MS. Many metastable ions dissociate after acceleration, and fragments can be generated by collisions (psd, post-source decay) . These normally arrive simultaneously with the precursor ions, since no field is present to induce separation. If the voltage of the reflectron is then changed stepwise, such ions can be separated prior to arrival at the detector [115], since their masses are different. Recently orthogonal injection of ions into the TOF instrument was used to interface electrospray and electron impact sources, also giving excellent performance (for a review see [116]). The acquisition speed of the TOF instruments renders them excellent detectors for rapid separations, such as microCE. An impressive example has been published recently [117] and in Figure 16 spectra of such a combination are shown.

5.4. Fourier Transform Mass Spectrometry

Fourier Transform MS was developed on the bases of ion cyclotron (ICR) mass spectrometry [118]. The principle is shown in Figure 17. The cubic cell comprises two trapping plates (T), two excitation plates (E), and two detector plates (D) and is located in the center of a high-field magnet, generally a superconducting magnet with up to 7 Tesla. Trapping voltages applied to two opposite plates keep the ion in the center. After application of a radiofrequency (RF) voltage at two plates perpendicular to the others, the ions are excited and travel on circular trajectories with the same radius, but with frequencies that depend on their mass. The third pair of plates detects the RF image current induced by the rotating ions. The different frequencies and amplitudes in the image current signal can be extracted by Fourier transformation and converted into mass spectra. Theoretically, FT ICR MS has almost unmatched high-mass and high-resolution capabilities, the latter at the expense of very long acquisition times. CsI clusters up to $m/z = 31\,830$ have been observed [119], with rather low resolution, using an external ion source and quadrupole transfer ion optics; with a different instrument, a resolution of 60 000 at $m/z = 5922$ has been demonstrated using laser desorption [120]. For small molecules, resolutions in excess of 10^8 have been established. Further, MS/MS is rather easy to perform and operation is straightforward [121]. FTMS is the only competitor for sector field instruments with respect to resolution and accuracy.

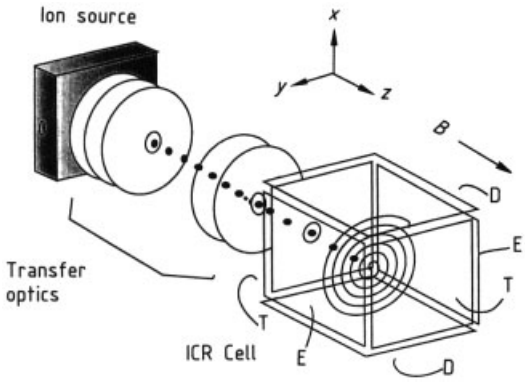

Figure 17. Schematic of a Fourier transform ion cyclotron resonance mass spectrometer with external ion source, transfer optics, and cubic FTMS cell
Ions can be generated using EI (even in combination with GC), FAB (LC/MS), or laser desorption/ionization; the transfer optics allow the required vacuum to be maintained.

However, its usefulness in analytical mass spectrometry still seems uncertain, partly because the very high cost prevents widespread use. Furthermore, optimal vacuum conditions ($< 10^{-6}$ Pa) are necessary in the ICR cell, which is difficult to maintain during GC/MS analyses. Attempts to solve the vacuum problem have been made with dual cell systems, where the ions are transported from an ionizer cell to an analyzer cell [122]. Two quadrupoles for mass selection and an RF-only quadrupole for focused injection [123] have been used, as has a time-of-flight arrangement [124]. They all serve to isolate the "dirty" ion source or the sample matrix from the analyzer to maintain optimal pressure conditions. With the last mentioned arrangement, the GC inlet replaced by a FAB target, and an atom gun, a spectrum of an enkephalin (Tyr-Gly-Gly-Phe-Leu) was obtained [125] with normal FAB features, but with almost no matrix background. Newly designed instruments seem to have solved the technical problems to a great

extent and both, MALDI and electrospray spectra have been published with exorbitantly high resolutions (in excess of 500 000) and with impressive sensitivity.

5.5. Ion Traps

Another ion storage device in rapid development is the ion trap. Although it has long been used by physicists [126] it was only much later that it was converted into a powerful GC detector [127]. In this case, only high frequencies and no magnetic field are necessary to create a cage for the ions in a closed quadrupole field. The ions are ejected several times per second from the trap, hitting the detector sequentially. The principles of operation are depicted in Figure 18. The first commercial version had a limited mass range of 600 amu and low resolution, but quite convincing detection capabilities.

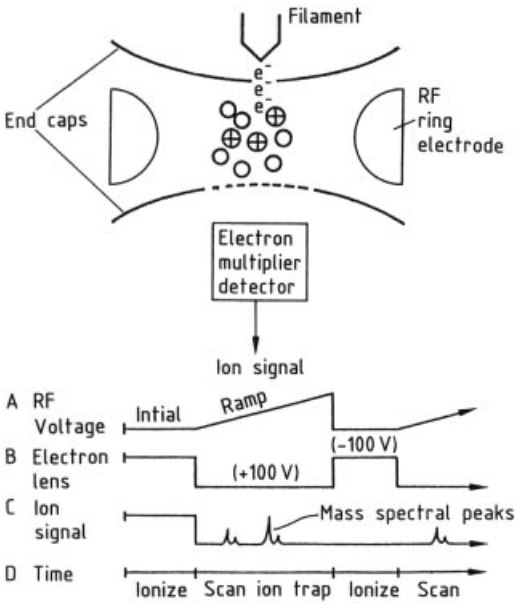

Figure 18. Schematic representation of the ion trap components, the scan function, the electron beam, and the resulting ion signals on the time axis (Adapted from [113] and [127])

The mass range has since been dramatically increased (up to 40 000 amu), although the resolution and particularly the mass accuracy is still fairly poor. In more recent versions, an additional scan ramp is used to evaluate the ion population of the trap for optimization of the residence time. However, high-resolution experiments using electrospray have now been described [128] and the detection capabilities have been steadily improved by optimizing the ion transfer into the trap. A modern version that is used as a detector for HPLC is shown in Figure 19. A weak point is still quantitative reproducibility. The storage capability is limited because when many ions are stored, coulomb repulsion destroys the performance of the trap. Therefore, software has been developed to monitor the number of ions in the trap and adjust the storage time accordingly to avoid unwanted ion – molecule reactions. Ion traps are the analyzer of choice in all cases where structure elucidation is the major application and several commercial versions from different vendors are available.

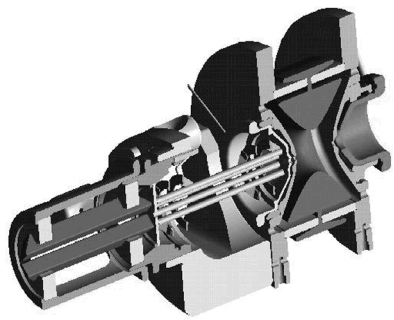

Figure 19. Cut view of the LCQ_{Deca} (Thermoquest)

5.6. Isotope Mass Spectrometer

The reliable determination of isotope ratios of elements normally needs specialized instrumentation. The ion sources depend on the type of element or molecule; for many elements the thermal ionization source has been employed, but for others and for small molecules, EI can be used as well. Although conventional analyzers can be used in such instruments, for precise determination, simultaneous detection of more than one ion beam is necessary and the instruments

must be optimized for this purpose. Isotope ratio mass spectrometry (IRMS) is a specialized technique designed for precise determination of isotope ratios such as $^{12}C/^{13}C$, $^{1}H/^{2}H$ in organic compounds. The analysis requires the separation of compounds by GC, combustion of the compound to small molecules such as CO_2 and H_2O, ionization and accurate measurement of the signals due to the isotopically pure fragments. The analyzers allow simultaneous detection of the signals at low levels with high precision by using several Faraday cups [129] (Fig. 20). With such instruments, an abundance sensitivity greater than 10^{-8} was obtained for $^{230}Th/^{232}Th$ ratios [129].

To obtain the highest precision, further precautions must be taken with sample handling, instrumental parameters, and electronics to minimize the noise levels. This makes the procedure highly complicated, and only a few laboratories in the world are able to perform such precise measurements. An example is the correction of the Avogadro number, based on isotope mass spectrometry with silicon isotopes [130].

5.7. Accelerator Mass Spectrometry (AMS)

Accelerator mass spectrometry (Fig. 21) is designed for the most precise atom counting of cosmogenic radionuclides such as ^{10}Be, ^{14}C, ^{26}Al, ^{32}Si, and ^{36}Cl. It can measure these elements with unparalleled detection power in, for example, polar ice cores, deep sea sediments, or wood. The nuclide ^{14}C is used in the geosciences for dating of organic matter. The ions from the source (different sources may be used) are selected with a 90° magnet, accelerated toward the stripper foil, and converted into highly charged positive ions. These ions are focused again into a magnetic sector and detected within a gas-filled chamber by means of secondary ions produced there. The differentiation of, for example, $^{14}C^{6+}$ and N^{7+} is possible, since both species penetrate into the chamber with a different depth.

The detector is designed to identify species of nearly the same mass and velocity but different charges with very high selectivity.

5.8. Other Analyzers

The number of additional analyzer systems similar to those discussed above, but with only minor modifications, and others making use of different principles, is large, but most of them have no practical application. An exception may be the Wien filter, an analyzer with crossed electric and magnetic fields [132]. It has high-mass capabilities and high sensitivity but very poor resolution, which has inhibited its widespread application as an analyzer, although it is used to control ion energies in some instruments. The use of Wien filters in hybrid MS/MS instrumentation has been proposed [133].

Aside from quadrupoles and ion traps, other dynamic analyzers [114] such as the monopoles have been merely of theoretical interest.

6. Metastable Ions and Linked Scans

The determination of metastable ions has found much interest, because of their structural and theoretical significance. Several techniques have been developed and some of them are discussed here. Some of the techniques can be made more powerful by means of forced fragmentation in a field-free region (collision-induced dissociation). This process is essential for MS/MS experiments (see Chapter 7). Metastable ions are gaining new relevance in time-of-flight instruments with reflectrons (see Section 5.3).

6.1. Detection of Metastable Ions

In a magnetic sector instrument the signals of metastable ions appear as broad, mostly gaussian peaks in normal mass spectra. The location $m*$ may be calculated from the precursor ion mass m_p and the product ion mass m_d using the relation:

$$m^* = \frac{m_d^2}{m_p}$$

These signals are normally lost in double focusing instruments, although in Nier – Johnson type mass spectrometers, reactions in the field-free region between electrostatic analyzer and magnet deliver signals at the same positions as well. In one experiment the acceleration voltage

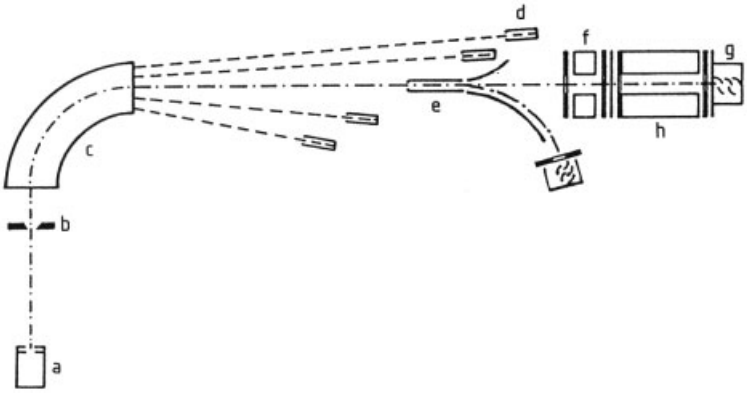

Figure 20. Isotope mass spectrometer with variable faraday cups and a quadrupole mass filter for MS/MS experiments where a deceleration lens is necessary between the analyzers
a) Ion source; b) Aperture slit; c) Magnet; d) Variable Faraday cups; e) Center slit; f) Deceleration lenses; g) Counter-SEM; h) Quadrupole mass filter

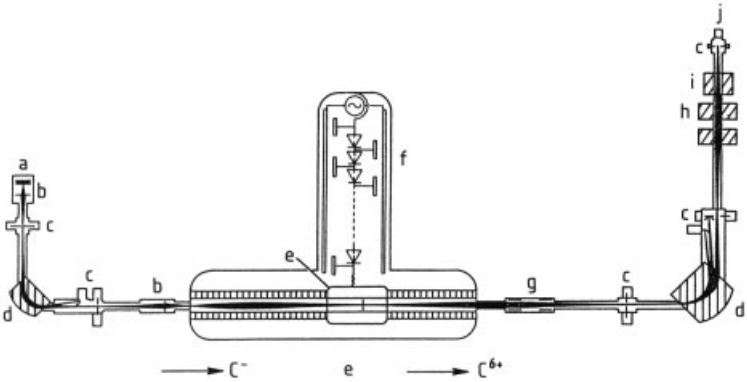

Figure 21. Accelerator mass spectrometry
a) Ion source; b) Lens; c) Slits; d) Beam bending magnet; e) Stripper foil charged to 2.5×10^6 V; f) High voltage source; g) Electrostatic lens; h) Magnetic lens; i) Velocity filter; j) Ion detector
(Reproduced with permission from [131])

is scanned. To detect a metastable transition, the spectrometer is focused for the product ion m_d with reduced acceleration voltage V_d. The acceleration voltage is then increased to V_p without changing the electrostatic analyzer and the magnet until the signal of the unknown precursor mass m_p appears. The ratio V_d/V_p is equal to the ratio of the masses m_d/m_p and m_p can be calculated. The resolution obtained with this so-called Barber – Elliott technique [134] or accelerating voltage scan is low and the instrument is continuously detuned, but for processes with rather small losses the detection capabilities can be great and kinetic energy release information is retained. This means that part of the excess energy of the ion is released as kinetic energy, leading to a range of translational energies yielding broad signals.

6.2. Mass Selected Ion Kinetic Spectrometry

Double-focusing instruments can be regarded as a combination of two mass spectrometers. For instruments with a reversed Nier – Johnson geometry, the magnet can serve as the first mass spectrometer (MS-1) that isolates one ion species from the mixture. This ion may undergo collisions in the field-free region following the magnet to form fragments, which can be separated by means of the electrostatic sector (MS-

2). This technique, called mass selected ion kinetic spectrometry (MIKES) or direct analysis of daughter ions (DADI), has only the very limited resolution of the energy analyzer for the fragments and requires reversed geometry. Therefore, the number of applications in analytical problems is small; its importance lies in determining the energy released during fragmentation.

6.3. Linked Scans

A simple and inexpensive way of obtaining MS/MS information is the use of linked scans in sector field instruments. The advantage over the MIKES technique is that the resolution is generally better and that all double-focusing instruments can be used. The two sectors and the acceleration voltage are linked together by functions that allow only specific fragment types to pass both analyzers. These may be the product ions of a precursor, all precursor ions of a particular product, or all ions with a preselected mass difference of precursor and product ions (neutral loss scan). In Figure 22, the relation between the most common linked scans is shown.

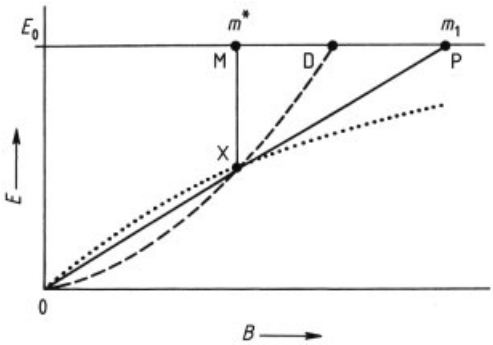

Figure 22. The *B/E* plane
—- for $B/E = \text{const.}$; – – – for $B^2/E = \text{const.}$; . . . . for neutral loss scan $B\,(1\text{-}E)^{1/2}/E = \text{const.}$

Product (Daughter Ion) Scans. When the ratio of magnetic field *B* and acceleration voltage *E* is held constant ($B/E = 1/v$, with v as velocity) during a mass scan downward from a precursor ion, only those fragments of this preselected ion which are formed in the first field-free region are transmitted through the electrostatic sector. This is because the velocity is unique for each precursor and thus for all the fragments formed after acceleration. The resolution of the precursor ion is low, but the resolution of the fragment ions can be quite high and the kinetic energy release information is lost. If a collision cell is available, high-energy collisions may be employed to induce fragmentation.

This type of scan is often used to analyze the fragmentation pattern of a substance. The spectra obtained may serve as a "finger print" of an ion and may allow a fragmentation scheme to be established. For a known reaction of a known compound the instrument can be set to detect only this one signal on this particular $B/E = \text{const.}$ surface, enhancing the specificity of detection (see "reaction monitoring," Chap. 2).

Precursor (Parent) Ion Scans. When a product ion appears with high frequency and the precursor is to be identified, a different scan is necessary. This precursor or parent ion scan requires a scan law in which $B^2/E = m$ with constant m. The velocity spread of the ions is not filtered, thus having poor resolution for the precursors, but good detection capabilities. This approach has been employed successfully for the screening analysis of drugs and their metabolites (metabolic profiling).

Neutral Loss Scans. Among several other possibilities for linking the fields and voltages in a double-focusing mass spectrometer, the neutral loss scan is the only one which deserves a brief discussion here. The scan function is

$$\frac{B^2\,(1-E)}{E^2} = \text{const.}$$

This type of scan has a distinct analytical potential because the loss of a common neutral fragment can be taken as an indication of a particular class of compounds. Therefore, it has been applied in the search for specific compounds derived from a certain precursor compound (drug/metabolite monitoring).

7. MS/MS Instrumentation

The combination of two or more analyzers, commonly known as MS/MS or tandem mass spectrometry, is a highly specific means of separating mixtures, studying fragmentation processes,

and analyzing gas-phase reactions. The field is reviewed thoroughly in [135]. With the first analyzer, one ion is isolated from all the others. In the next, reactions of that ion are studied further. Fragments may be formed in unimolecular reactions from metastable ions, but generally fragmentation is induced by a variety of means. Collisions with gases in a collision chamber (CAD: collision-activated decomposition), laser photodissociation [136], or surface collisions [137] can be employed to induce fission. Since MS/MS can be combined with chromatographic separation techniques, it has found a broad range of applications in analytical chemistry. Indeed, LC/MS interfaces such as thermospray or the API techniques depend largely on collision-induced dissociations and MS/MS. Since they produce only molecular ions, little structural information is provided. With CAD and MS/MS these details are obtainable. A review gives a fairly detailed discussion of the thermochemical aspects of the various fragmentation induction methods used in mass spectrometry [138].

7.1. Triple Quadrupoles

Triple quadrupoles for analytical MS/MS experiments have been built since the pioneering work of Yost and Enke in 1978 [139] and they have found applications in all areas of mass spectrometry. The first and the third quadrupole are MS-1 and MS-2, whereas the second in the middle is an RF-only quadrupole acting as collision cell. In the collision cell, the transmitted ions undergo low energy collisions with a gas. For the product ion scan (daughter scan), only one precursor (parent) is selected and the second analyzer scans for the product ions. For the precursor scan, the first quadrupole is scanned and only the selected product ion is recorded. In the neutral loss scan, both analyzers are scanned with the selected mass difference, and for reaction monitoring only one precursor and one product ion species is permitted to travel through MS-1 and MS-2, respectively. In the meantime, instruments with different collision regions have been built. It was surprising that fragments are formed under the low energy regime in the collision cell. The fragment ion reactions are similar to those from high-energy collisions; exceptions include side-chain reactions of amino acids in peptides. The collision efficiency is quite high, particularly when the collision cell is an RF-only quadrupole [140] or octapole with nearly 100 % transmission [141]. The resolution of both quadrupoles as MS-1 and MS-2 can be set independently allowing, for example, unit resolution for precursor and product ions. Many different experiments with scan functions similar to the linked scans have been made and the set up and computer control is quite straightforward. Figure 23 shows schematically the widely used MS/MS experiments. The triple quadrupole is the most common MS/MS instrument.

7.2. Multiple-Sector Instruments

In cases where high-resolution MS/MS experiments are required, multiple-sector field instruments are the only choice. They have the advantages of potentially high resolution in both stages, a wide mass range, and high sensitivity. However, there is a considerable trade off due to complexity, space requirements, and high cost. This is true particularly for the four-sector field mass spectrometers. As an example, the schematic of a four-sector field instrument is given in Figure 24.

Such mass spectrometers are highly flexible since the possible combinations of sectors and the number of field-free regions allow many different experiments to be performed [142]. The limitation is that the fragmentation process becomes less probable with increasing mass and the upper limit seems to be around 2500 amu. To increase the dissociation efficiency not only collision cells, but also photodissociation and surface-collision-induced dissociation have been investigated as well as simultaneous recording with a zoom lens integrated into the second MS [143].

7.3. Hybrids (Sector – Quadrupoles)

The apparently logical combination of the high resolution capabilities of sector instruments with the simplicity of quadrupoles in hybrid instruments is complicated in practice. Since the quadrupole only accepts ions of low kinetic energy, it must be floated at the potential of the sector instrument or, if high-energy collisions are to

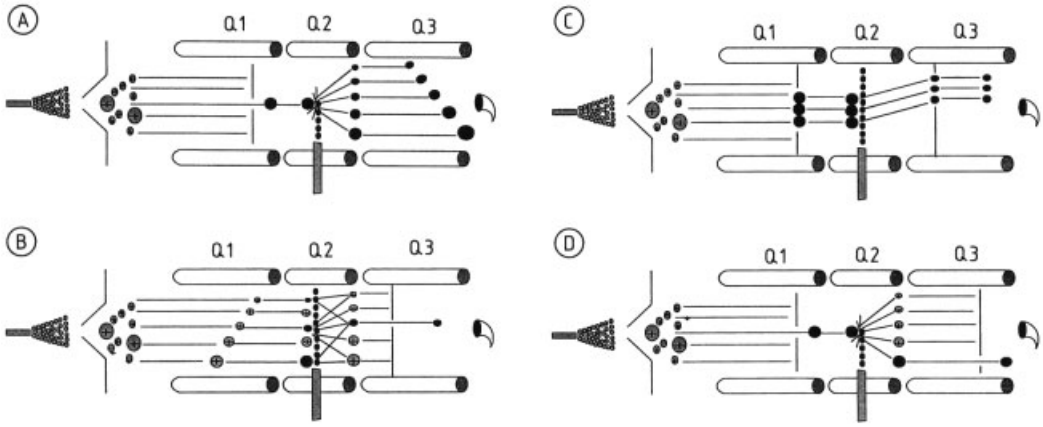

Figure 23. MS/MS experiments with the triple quadrupole MS
A) Daughter scan; B) Parent scan; C) Neutral loss scan; D) Reaction monitoring
Q1, Q3: mass analyzers; Q2: collision cell

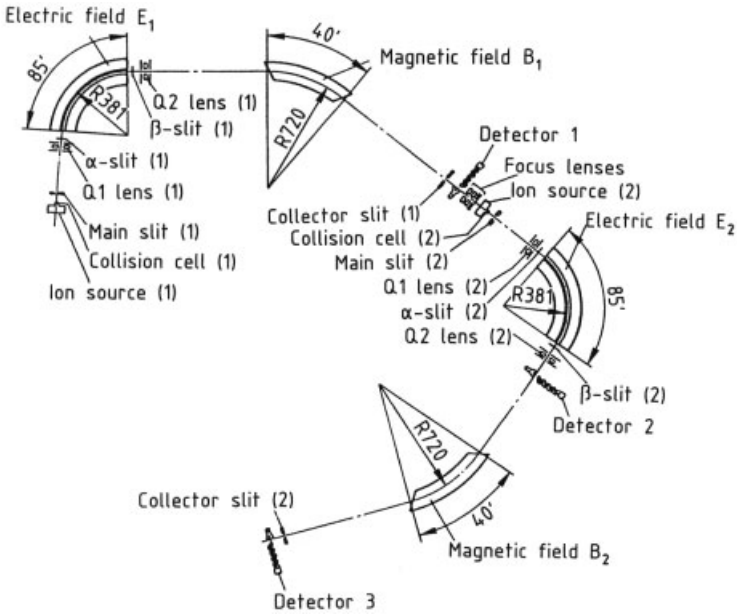

Figure 24. The schematic diagram and the ion optical components of the four-sector field mass spectrometer JEOL HX110/HX110 (Reprinted from [142])

be possible, a zoom lens allowing the deceleration of the ions must be incorporated between the mass spectrometers. The latter has been realized in only one commercial version, whereas the simpler first combination has been made by several manufacturers. Mostly the sequence is sector – quadrupole, but it is possible to arrange the spectrometers the other way around. Combinations of double focusing spectrometers with time-of-flight instruments [156] or with ion traps have been built and both are hybrids with great analytical potential. However, in general the sector instruments made these combinations too complicated and too expensive.

7.4. Ion-Storage Devices (FTMS, Ion Traps)

Two types of ion-storage device (ion traps) are in use today as mass spectrometers, as discussed already, with built in MS/MS capabilities: the magnetic trap as ion cyclotron resonance spectrometer, which is, in its modern version, the fourier transform mass spectrometer [144–147] and the electric field version, called ion trap [148–152]. Both types of instrument allow all but the one selected ion to be ejected, which then undergoes ion – molecule reactions, collisions, or photodissociation to yield fragments. This process may be repeated, allowing not only MS/MS but also MS/MS/MS and even MS^n, which allows the detailed study of fragmentation reactions. However, currently only daughter-ion scans are possible.

7.5. Quadrupole Time-of-Flight Tandem Mass Spectrometers

The development of combination sector field/TOF led almost instantaneously to the combination of quadrupoles with TOF spectrometers, since at the same time the TOF spectrometers were converted into high-resolution high-precision instruments (Fig. 25). A new aspect of this is that the ion injection into the TOF spectrometer is orthogonal, yielding an MS/MS combination with extremely high sensitivity for the fragments, without compromising scanning speed, resolution, mass range or the mass accuracy [153].

Thus this spectrometer has become one of the most successful new developments in the history of mass spectrometry (produced by two vendors) and today the instrument is one of the key tools in protein, peptide, and DNA research all over the world [154].

7.6. Others

Several TOF/TOF instruments have recently been described with impressive sensitivities, although the resolution in the second spectrometer is low due to the energy spread (similar to the situation in MIKE) [155]. The same is true of BE/TOF combinations, which seem to be highly suitable for electrospray and MALDI techniques [156], but it is too early to predict the future importance of these instruments.

8. Detectors and Signals

The detectors in mass spectrometry can be divided into two general types. One group is designed to detect one ion beam at a single location only (e.g., Faraday cage, secondary ion multiplier), and the second group can detect several ion beams simultaneously (e.g., photoplates and multiplier arrays). It is clear that the second group can be adapted only to specific analyzers, but the advantage in terms of detection power is so great that such instruments have been developed. The most common types of detector are shown in Figure 26.

8.1. Faraday Cage

One of the first detectors in mass spectrometry was the faraday cage. In principle it is simply a cup, which converts the current induced by impacting ions into a voltage pulse, the height of which is a measure of the number of ions reaching the detector per unit time. Since at that stage no amplification is achieved, its detection power is limited, whereas its linearity over the full dynamic range is still unsurpassed. The latter characteristic makes it even today a valuable tool, when high precision must be maintained (e.g., isotope mass spectrometry). In early instruments it was possible to switch between a faraday cup and a secondary electron multiplier. Conversion dynodes normally included into modern spectrometers, may be used as faraday cups as well, if they are positioned in line with the ion beam [157].

8.2. Daly Detector

The Daly detector [158] is still in use; the first version was able to distinguish between metastable and normal ions. The ions are converted into photons by means of a scintillator and the photons are measured by a photomultiplier which is located outside the vacuum

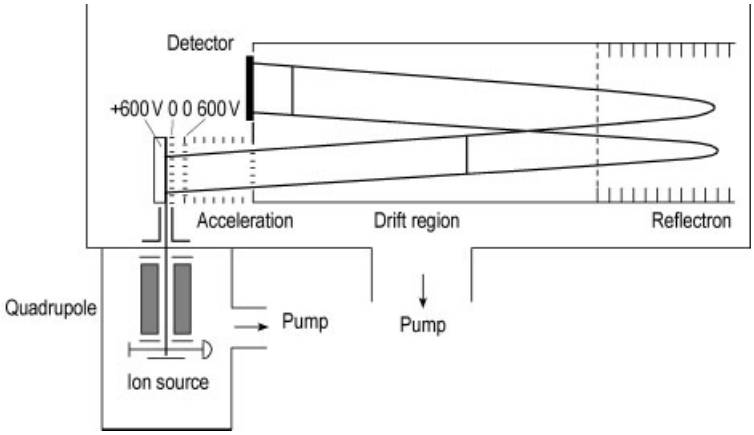

Figure 25. Q -TOF tandem mass spectrometer, schematic.

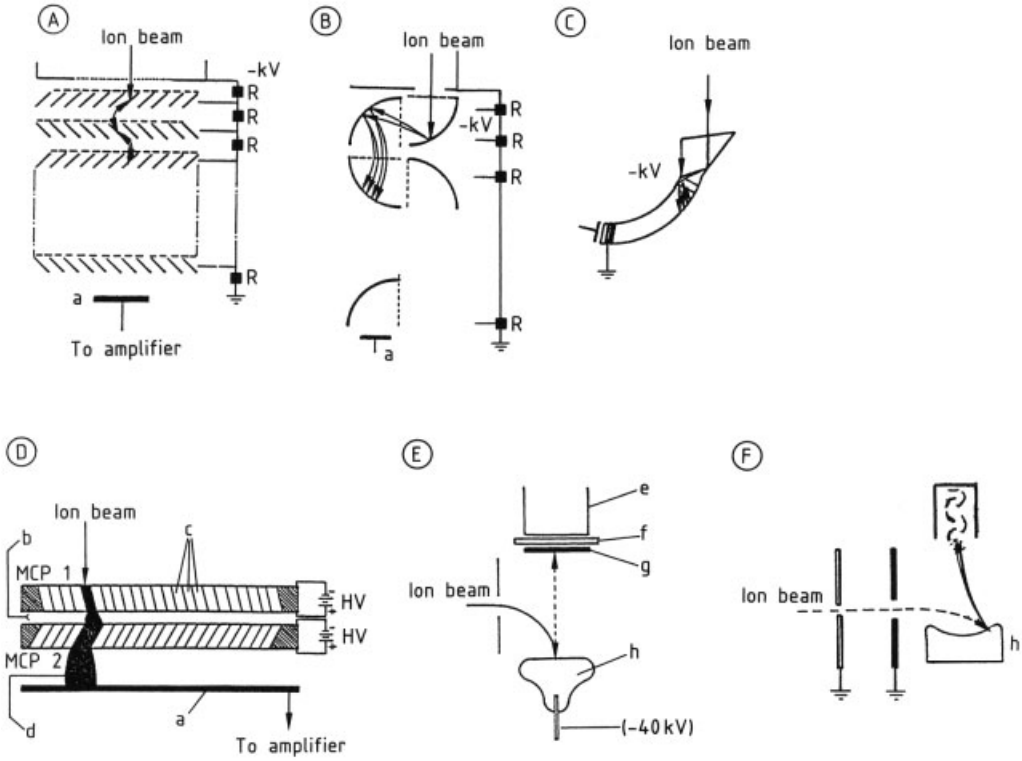

Figure 26. Types of detectors
A) Venetian blind multiplier; B) Box and grid multiplier; C) Channeltron multiplier; D) Microchannelplates; E) Daly detector; F) Postacceleration/conversion dynode
a) Anode; b) Chevron; c) Channels; d) Electrons; e) Photomultiplier; f) Window; g) Scintillator (+ 10 kV); h) Dynode

(see Fig. 26E). It is easy to integrate a post-acceleration/conversion dynode opposite to the scintillator plate to enhance the sensitivity. The originalversion was arranged in beam and had an additional plate in front of the scintillator, serving as conversion dynode. With a potential at the ion energy, the normal ion would only gently touch the scintillator, giving no signal. Metastable ions with less energy will be reflected to the intermediate plate, giving secondary ions which are in turn accelerated toward the scintillator; thus, it is possible to distinguish metastable ions from "normal" ions.

8.3. Secondary Electron Multiplier (SEM)

The most common detector is the secondary electron multiplier, which amplifies the first impulse from the impacting ion by orders of magnitude. Several different types have been developed, with specific properties. Multipliers with open dynodes, the surfaces of which are made from beryllium oxide, have high amplification, can be baked out and, depending on the number of stages, can have short blind times allowing ion counting. This is the period of time during which the electron cascade runs through the dynodes and when no new impact is seen. Similar are the venetian blind multipliers, which, however, cannot be used for ion counting. Each dynode consists of slats arranged like a venetian blind with a light grid in front to ensure a homogeneous field in the dynode. Due to the circular, wide open active area, the positioning is not critical, although the ion beam should not be too narrow (see Fig. 26A). Multipliers based on semiconducting surfaces (channeltrons) are widely used in high-current systems and are incorporated in quadrupole and ion-trap instruments. Instead of discrete dynodes with resistor chains, the channeltron uses a semiconducting material for the same effect and many different types and shapes are in use, giving similar performance as the dynode types (see Fig. 26C).

When heavy molecules had to be detected, the need for post-acceleration to increase the probability for secondary electrons arose. Since then, conversion dynodes have been included in the detection systems to accelerate slow heavy ions (typical potentials between 15 and 30 kV), and thus enhance the sensitivity for both positive and negative ions.

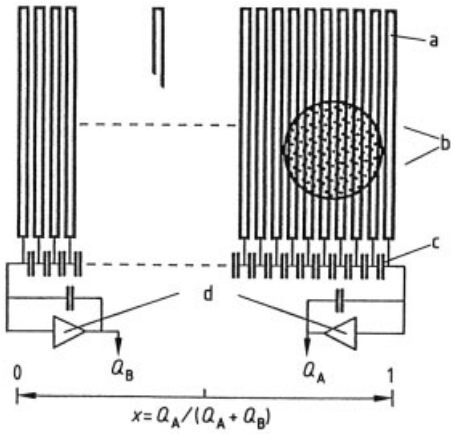

Figure 27. The Patric detector
a) Conductive strips; b) Electron cloud; c) Condenser chain; d) Charge sensitive amplifiers (position and time information for every ion)
(Reproduced with permission)

8.4. Microchannel Plates

Microchannel plates are based on the same principle as the channeltron (Fig. 26C) and can be regarded as an array of small channeltrons of rather short length (a few mm instead of cm) arranged side by side in one plate similar to the photoplates of Mattauch – Herzog instruments. They are used in TOF instruments and for the simultaneous detection of ions in sector field mass spectrometers. The major problem to be solved in the latter case was the fast readout of the signals, and various techniques have been developed for this purpose. A first, very elaborate version of such a device was described as early as 1974 [159]; more modern ones suitable for tandem mass spectrometers use fiber optics to transmit the signals to photodiode arrays [160]. Detectors with zoom capabilities [143], static or scanning modes, and switching between analog and counting modes have been designed and successfully used [161], [162]; the gain in sensitivity can be two orders of magnitude depending on the system employed, although the resolution and sometimes the dynamic range is limited. The Patric detector (see Fig. 27) is a simultaneous detector with scanning option. The detector

is capable of locating the impinging ion on the surface as long as the events are not exactly simultaneous; therefore, the dynamic range of the device is somewhat limited. The resolution obtainable is ca. 5000; since the mass spectrometer can switch between analog recording with a conventional SEM and the Patric detector even in a scan, the enhancement in sensitivity (ca. 100) can be fully utilized.

8.5. Signal Types

The shape of the signals depends largely on the type of instrument. Signals in sector field instruments are ideally triangular when the magnetic field is scanned and resolution as well as intensity are under optimal conditions. A rectangular beam profile (from the entrance slit) is moved across the exit slit, the width of which should be adjusted to match the width of the ion beam. This means in practice that the height of the signal is adjusted by the entrance slit. Then, the width of the final triangle is optimized by means of the exit slit without reducing the intensity further.

In quadrupole instruments the signals are in principle rectangular, since no slit can be used to define the beam shape precisely. In Fourier transform mass spectrometers, the peak profile should be gaussian when the signals are detected by resonance phenomena. The signals in time-of-flight instruments depend on the shape of the primary pulse (ions, neutrals, or laser photons), but even more on the desorption characteristics of the sample. Since the time for desorption and ionization of the sample is usually much longer than the primary pulse, the resulting signals have rather wide, poorly defined peak shapes with poor resolution. It is possible to shape the secondary pulse by means of ion optics to enhance the performance of the TOF instruments with respect to resolution.

9. Computer and Data Systems

The first applications of computers in mass spectrometry were merely the recording of signals from the multipliers, conversion to spectra, and some data handling. Today, powerful computers are integrated in the spectrometer. They control the scan, define the scan function, set the required voltages, tune the source, and supervise the experiments. The data are stored and interpreted with advanced tools, although at that stage the fully automated procedures are at an end.

9.1. Instrument Control, Automation

Modern mass spectrometers are completely dependent on computers and data systems. All the instrumental parameters of the inlet systems such as direct probe, GC, or LC (temperatures, gradients, times) are supervised by the computer. The ion source is tuned by automated procedures, the instrumental parameters (voltages, frequencies, resolution, etc.) are set by software procedures, and even the design of experiments (EI, CI, positive/negative ions, linked scans, MS/MS, etc.) can be started automatically, if specified conditions are met (e.g., signal height, retention times). Under such circumstances, the spectrometers can run a large number of samples unattended, since the analytical methods can be specified first, stored and activated upon request of the software. One advantage of this development is that a single man–machine interface is sufficient to work with such an MS. Further, most of the information necessary to describe an experiment from an instrumental point of view is stored automatically and is therefore available for reports and documentation.

9.2. Signal Processing

Most mass spectrometers produce analog signals at the detector. Thus it is necessary to convert the analog current or voltage signal into digital signals by means of analog-to-digital converters. This step has been a problem in the past, because the dynamic range and the speed of such devices was not always sufficient to maintain the integrity of resolution and dynamic range delivered by SEMs. In cases where ion counting is possible, different problems must be solved: the maximum counting speed is the limiting factor for the dynamic range. Since the design of the

multiplier defines the pulse width and blind period, the dynamic range is a function of the performance of the multiplier and the pulse counting device. When pulses overlap, errors may occur, although correction is possible if the frequency is not to high.

Analog-to-Digital Conversion and Mass Calibration. As mentioned above, the analog signal is digitized by fast and accurate A/D converters (typically above 200 kHz) with sufficient dynamic range ($\geq$16 bit). The limiting factor for the dynamic range is often the noise of the baseline. The digitized information about the ion current is fed into the computer, where the signals are analyzed and converted from "raw data" into peaks or noise, the intensities are calculated and placed on a mass scale. The conversion can be performed on the basis of time, magnetic field, or frequency, depending on the type of spectrometer. The mass scale must be calibrated to achieve the required precision. The most reliable way to calibrate the mass scale for all mass spectrometers is by means of reference compounds, whose mass spectrum in terms of accurate masses and approximate intensities should be known to the software. In addition, they should have signals evenly spaced over a large mass range.

Some reference compounds for mass calibration are listed below:

EI, CI, positive and negative
Perfluorokerosene (PFK, up to 1000 amu)
Perfluorotributylamine (FC-43, heptacosa, M_r 671)
Tris(perfluoroheptyl)-*s*-triazine (up to 1200 amu)
Ultramark 3200F (Fomblin oil, perfluoro polyethers, up to 3200 amu)
Ultramark 1621 (perfluorinated phosphazine, up to 1600 amu)

LC/MS (Thermospray)
Polyethylene glycols (up to 2000 amu)
Polypropylene glycols (up to 2500 amu)
Acetic acid cluster (up to 1500 amu)

FDMS
Polystyrenes (up to 10 000 amu)

FABMS
Glycerol (useful up to 1000 amu)
CsI (wide mass range, rather wide spacings)
CsI/RbI and NaI/CsI/RbI (wide mass range)
Ultramark 3200F (up to 3200 amu)
Phosphoric Acid (5 %) in glycerol (up to 1500 amu, negative ions)

Electrospray
Polypropylene glycols (up to 2500 amu)

After acquisition of the spectrum, the computer tries to identify some starting pattern and then searches for all the other peaks. When the number of signals identified is sufficient, a calibration table is created which allows raw data to be converted into mass/intensity pairs. If high-resolution data have been acquired and accurate masses are to be calculated, additional internal standards are necessary for precise calculation of the signal location relative to known signals of the internal standard.

Peak Detection. One of the problems to be solved is the detection of true signals. It is necessary to determine the base line and to discriminate between noise signals of low intensity or insufficient width and real peaks. For this purpose, sophisticated software has been developed. Several parameters must be defined under a given set of instrumental conditions to ensure that no signal is overlooked and that as few noise signals as possible are recognized as peaks, obscuring the information contained in the mass spectrum. To do this the peak width, the minimum height and minimal number of counts, and, for unresolved signals, the depth of the valley between two adjacent peaks are calculated for a given scan speed, scan function, and number of digital points ("counts"). Only signals meeting all such requirements are accepted as peaks and if a problem is encountered during assignment, the peak can be flagged for classification.

Dynamic Range. The dynamic range of a multiplier in a mass spectrometer covers several orders of magnitude (up to 1×10^6) and is defined as the ratio between the smallest and the largest signal which can be produced. The digitalization system in the computer interface has to comply with this adequately, meaning the dynamic range of the A/D converter and its speed must be sufficient. With modern converters and sample and hold devices coupled with autoranging systems, this problem is solved. The simultaneous recording devices have rather low dynamic ranges due

to saturation effects and readout problems. Since they are specifically designed to acquire signals of very small ion currents, this limitation is acceptable.

9.3. Data Handling

When the mass spectral data have been stored in the computer, the interpretation of the data is the next step. In most cases the data are spectra with peak intensities on a *m/z* scale. However, chromatograms from selected ion recordings and data from multichannel analyzers must also be handled. Sophisticated software has been written to develop increasingly automated procedures for quantitative analyses, mathematical treatment of data, display of spectra and structures, and to transport the data into laboratory data systems for integration into reports. While this aspect of data handling can be considered rather satisfactory today, the extraction of information contained in the spectra is under development. The deduction of possible structures of unknown compounds from mass spectra is demanding and often fails, and the calculation of spectra based on the structure is in its infancy. It is still uncertain whether this is at all possible, because the competitive reactions of the gas-phase ions are controlled by minute differences in energy between intermediate structures, and the kinetic behavior is difficult to model for complex molecules. Initial attempts have been published for both the modeling of reactions in the mass spectrometer [163] and the prediction of mass spectra from structures [164]. In both cases the results are far from application for analytical purposes.

9.4. Library Searches

So far, the most successful approach for extracting structural information from spectra is the library search. Libraries have been collected from different sources with large numbers of spectra; the most widely used are the Wiley data base with ca. 160 000 and the NIST data base with ca. 60 000 spectra; others, specializing in topics such as toxicology, agricultural chemicals, or foodstuffs, are also available [165]. The quality of the data contained in some of the large databases is still a substantial problem and must be improved, since spectra often strongly reduced in the number of peaks, incomplete, or incorrect have been taken from old collections and included in the new compilations. It is not trivial to identify and to replace them with better spectra, because wrong spectra may be never found (similar spectra may no longer show up) and the definition of better spectra is also difficult. Quality is a relative attribute of a spectrum and sometimes two completely different spectra must be maintained, if they were recorded under diverse conditions. Library searches, with a variety of algorithms, have been developed, and the best known are SISCOM [166], [167], PBM [168], [169], and the INCOS type search [170]. The SISCOM approach also allows the identification of unknowns and the deduction of structural features, thus giving information about similar compounds. This is illustrated by the example given in Figure 28.

More recently the chemical structures of the compounds have been integrated as a means to enhance the power of the information retrieval. It has become possible to identify not only similar spectra, but also the spectra of similar structures, revealing much of the complex relation between spectra and structures [171].

9.5. Integration into Laboratory Networks

Since in many laboratories, particularly in industrial environments, mass spectrometers are merely one tool among others such as NMR and IR spectrometers, it is becoming increasingly necessary to integrate the spectrometers into laboratory networks [172]. Advantages of this approach include:

1) The customer as a producer of a sample can use his terminal in the network to inspect all the spectroscopic data of a compound submitted for analysis or structure elucidation.
2) "Intelligent" interpretation, aided by computer software, can be initiated prior to examination.
3) It is possible to follow the fate of sample not only on the spectroscopic environment, but in other analytical, biological, or toxicological examinations as well. In the following

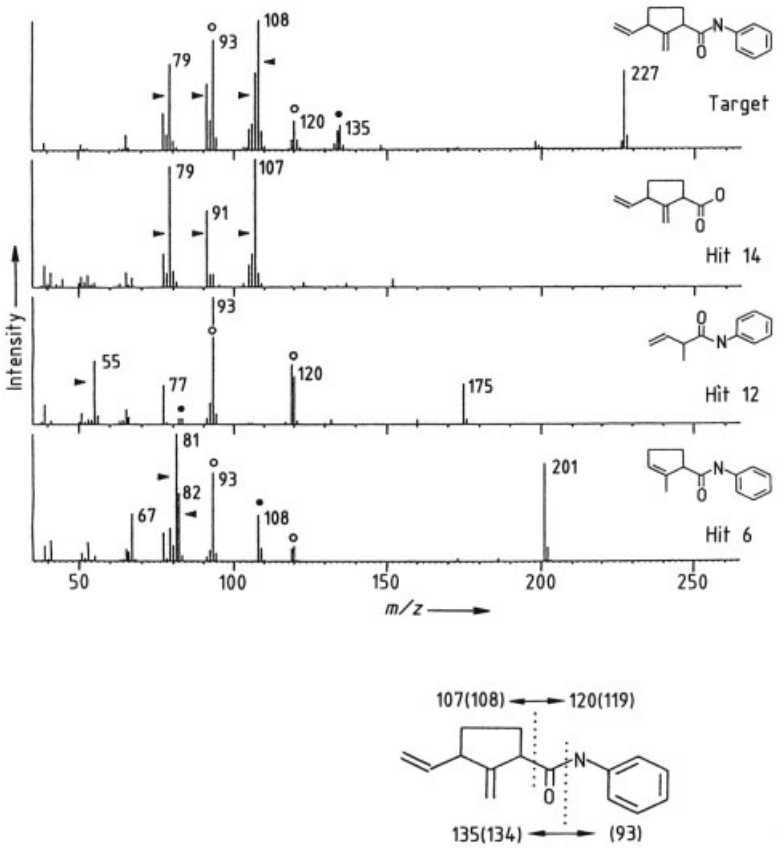

Figure 28. Spectra of similar compounds found by a similarity search for structures, not for spectra
► = Hydrocarbon fragment; ○ = Anilide moiety; ● = Loss of aniline radical or aniline molecule
The fragmentation pattern (bottom) explains why the analog reaction in Hit 6 yields *m/z* 82 instead of *m/z* 108; *m/z* 108 in spectrum Hit 6 must have a different origin. (Adapted from [171], with permission).

step, all the data can be combined and the information may be integrated into a single report.

4) The thorough documentation and the assurance of quality becomes possible. The integrity of the data from the sample to the final report becomes increasingly important and is only possible with adequate use of the data systems.

9.6. Integration in the World Wide Web

As with many others examples, mass spectrometric data have been published through the WWW from many publishers and it is fairly easy today to access new bibliographic data via the internet. The homepages of most publishers provide either the bibliographic data or even the abstracts free and often the costs for the online publications are more or less the same as for the printed versions. Even more relevant may be the fact that in protein and nucleic acid research the most important databases are accessible over the net. MS datasystems have built-in software to search the databases directly or even via an automated strategy for known sequences in peptides or oligonucleotides, thus allowing the rapid identification of known proteins based on MALDI data of peptide mixtures and electrospray MS/MS data ("sequence tags"). It can be stated that the stunning speed in proteomic research is largely due to the immediate access to such databases and the retrieval software.

10. Applications

Typical areas for the application of mass spectrometry are:

1) Analysis of trace elements and compounds in environmental samples
2) Problem solving in biomedical research (protein identification/DNA) , particularly the structure elucidation of natural compounds and determination of molecular mass
3) Control of synthetic procedures by rapid analysis of intermediates and products
4) Analysis of bulk materials and surfaces

One of the major driving forces for the rapid development of mass spectrometry during the last years is trace analysis in environmental chemistry and biochemistry. The search for sensitive, reliable methods that allow the determination of compounds at very low levels together with the option for mass spectra of highly polar and labile compounds motivated the development of new interfaces, ion sources, analyzers, and detectors. The second impetus is the search for techniques for the determination of very high molecular masses with better precision than is attainable by other known methods.

10.1. Environmental Chemistry

Mass spectrometry is a powerful tool in environmental chemistry because of its detection power, specificity, and structure analysis capabilities. Techniques have been developed for the determination of catabolites and metabolites of chemicals in practically all relevant matrices since no other comparable tool is available, even though technical problems can occur. Generally, sample preparation uses at least one type of chromatography, connected with MS either off-line or on-line. Of the on-line combinations ("hyphenated techniques"), GC/MS is the most successful, eventhough LC/MS is rapidly catching up in this area. The development of benchtop GC/MS and LC/MS instruments has made it possible to use mass spectrometry in routine laboratories and even in the field [173].

Air. The investigation of reactions in the atmosphere is one of the major fields of modern environmental chemistry. The processes responsible for the formation of smog in cities [174] as well as the interactions of man-made or natural hydrocarbons with ozone [175] are examples where mass spectrometric techniques have been applied. Usually, the chemicals under investigation must be trapped on a stationary phase such as charcoal or tenax. The samples are then transferred by thermodesorption to a GC/MS system [176]. For the investigation of the organic contents of single particles from aerosols, laser desorption has been employed successfully [177]. The potential of the new ion sources such as the atmospheric pressure ion (API) source for monitoring reactions directly or for detecting air pollutants at very low levels has yet to be fully explored; however, initial experiments are encouraging [178]. The process of gas to particle conversion has been investigated using an adapted API source [179]. Volatile organic compounds in air and in water have been successfully analyzed by using membranes in various configurations to give enrichment factors in the range 40–780 [180]; with nitrogen as matrix, several halogenated or aromatic compounds have been detected in the ppb range. Recently, the specificity and sensitivity of REMPI mass spectrometry has been used to detect organic compounds in gases directly without trapping [181–183]; one additional benefit of this technique is the superior time resolution allowing rapid changes in concentration, to be followed.

Water. For water the situation is similar in that usually in the first step pollutants must be extracted from the matrix. Derivatization is often necessary to make MS analysis possible. Extraction and preconcentration is the critical step in the procedure, because the nature of the chemicals to be monitored can vary widely. LC/MS techniques that avoid derivatization have been used successfully. Pesticides such as triazines, organophosphorus compounds, and phenylureas have been determined in water samples by thermospray LC/MS [184]. Even though the well-developed GC/MS procedures are mostly used, they fail to detect some classes of compounds, such as plasticizers in drinking water, where they were first found by LC/MS [185]. A broad range of pollutants have been identified in wastewaters by particle beam LC/MS [186]. For the direct analysis of organic compounds in water an in-

strument has been designed with a membrane inlet, which allows certain chemicals to pass directly into the detection system, in this case, an ion trap. The detection limits for some of the compounds are low enough (e.g., benzene 1 ppb, toluene 10 ppb, ether 10 ppb; acetone and tetrahydrofuran 100 ppb) for routine monitoring of water [187].

For the determination of elements in water, ICPMS has demonstrated superior performance due to its sensitivity and multielement capabilities (see Section 10.6).

Soil. The analysis of herbicides, insecticides, and fungicides in soil is commonly based on extraction with liquids or supercritical fluids. The samples are then concentrated, derivatized if necessary, and transferred into the detection system. Often, GC/MS methods are the first step, although MS can be replaced with a less complicated detector in the final routine version of the procedure. For developing new procedures and in cases where the identification of a compound is uncertain due to contamination, GC/MS may be the only choice. One of the most important applications is the analysis of tetrachlorodioxins (TCDDs), especially the toxic isomer 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, and related compounds. These are formed as byproducts of pesticide production, and occur in fly ashes from municipal incinerators and other sources [188]. All standard procedures for the detection of such compounds have as the key step GC/MS, normally with high resolution (10 000) [189], [190] to increase the specificity. Mass spectrometers have been designed solely for this purpose by several manufacturers and, since the detection limits are very low and the reliability of the procedures must be high, the use of reference compounds with ^{13}C labels is mandatory [191]. Another method of enhancing the specificity is the use of API sources [192]. For the investigation of metabolism or catabolism of herbicides (e.g., by photooxidation), LC/MS has become increasingly important [193]. The fate of such compounds (e.g., phenoxyalkanoic acids and their derivatives [194]) can be monitored in soil samples by analysis of their stability and by the emergence of catabolic and metabolic products. A further example is the analysis of phenols and their biological conjugates (e.g., sulfates, glucuronides) with LC/MS [195], which was previously very difficult and time consuming. The fate of elements in soil is difficult to investigate, since the interactions with fulvic and humic acids are complex and not well understood.

Waste Materials and Dump Sites. The major difficulty in MS analysis of wastes is often the highly complex matrix, which must be excluded from the analysis. One possibility is the use of multiple chromatographic steps prior to MS analysis. Alternatively the compounds of interest may have very different physical properties from the matrix. This is the case with polycyclic aromatic compounds, which have been determined directly by using thermal desorption with GC/MS [196] detection. Recently, small portable quadrupole mass spectrometers have been built as a means to identify contaminated sites such as abandoned industrial plant locations and old waste dumps.

10.2. Analysis of Biomedical Samples

In the last few years the biomedical field has experienced a revolution of mass spectrometric possibilities [197].

MS has long been important for the detection of trace amounts of, for example, steroids [198], [199], drugs [200], and prostaglandins [201] in biological fluids such as urine, plasma, or saliva. Another area with many applications is the structure analysis of a wide variety of compounds, sometimes after derivatization or chemical modification. Micro-scale methylation [202], acetylation, and silylation have been used for the MS analysis of large molecules. For oligosaccharides, often not only the sequence and the type of monomer must be analyzed, but also the position of the glycosidic linkage and derivatization/cleavage schemes have been developed for this purpose [203], [204]; for details of the Lindberg method [205], see [197]. The sequence of peptides can be determined by cleavage to short oligomers and GC/MS, often in combination with FABMS for the determination of large fragments made enzymatically [206]. The reactions of the protonated ions produced with FAB (or liquid SIMS, LSIMS) have been studied extensively. Chemically labile bonds have the tendency to break more often, leading to analytically significant ions. Some of the most

important are shown in Figure 29. The same type of ions are observed with electrospray CID (MS/MS) or MALDI PSD spectra which have revolutionized the analysis of proteins in the last few years, and proteomics without mass spectrometry would be unthinkable. The general strategy applied is shown in Figure 30.

N-terminal ions

a_n: $H-(NH-CHR-CO)_{n-1}-\overset{+}{N}H=CHR_n$

or

$H-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-NH-\overset{CR_n^aR_n^b}{\overset{\|}{C}H}$

a_n+1: $H-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-NH-\overset{R_n}{\overset{|}{C}H}\cdot$

b_n: $H-(NH-CHR-CO)_{n-1}-NH-CHR_n-C\equiv O^+$

c_n: $H-\overbrace{(NH-CHR-CO)_n}^{H^+}-NH_2$

d_n: $H-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-NH-\overset{CR_n^b}{\overset{\|}{C}H}$

C-terminal ions

v_n: $HN=CH-CO-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-OH$

w_n: $\overset{CR_n^b}{\overset{\|}{C}H}-CO-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-OH$

x_n: $^+O\equiv C-NH-CHR_n-CO-(NH-CHR-CO)_{n-1}-OH$

or

$O=C=N-\overbrace{CHR_n-CO-(NH-CHR-CO)_{n-1}}^{H^+}-OH$

y_n: $H-\overbrace{(NH-CHR-CO)_n}^{H^+}-OH$

y_n-2: $\overbrace{HN=CR_n-CO-(NH-CHR-CO)_{n-1}}^{H^+}$

z_n: $\overset{CR_n^aR_n^b}{\overset{\|}{C}H}-CO-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-OH$

z_n+1: $\cdot CHR_n-CO-\overbrace{(NH-CHR-CO)_{n-1}}^{H^+}-OH$

Figure 29. Most common peptide fragments from positive ions
R = Side chain of an amino acid; R^a and R^b = Beta substituents of the *n*th amino acid (From [207], with permission)

Since such ionizing techniques are compatible with many MS/MS experiments, sequence information can be deduced from the spectra in great detail. When high-energy collisions are used, even leucine and isoleucine can be distinguished due to the different fragmentation of the side chain.

The analysis of complex lipids such as glycosphingolipids [208] and lipopolysaccharides [209] became even more complete with the advent of desorption techniques. But the localization of the position of double bonds in unsaturated fatty acids from complex lipids is still difficult. Several techniques were conceived using pyrolidides [210] or dimethylsulfides [211] and EIMS, CIMS [212], or MS/MS techniques [213]. As another example, electrophoretic labeling (perfluoroacylation) allows the determination of modifications in nucleobases at the attomole level [64] or the characterization of prostaglandins in biological fluids by means of negative chemical ionization (NCI) [214].

Field desorption mass spectrometry (FDMS) indicated for the first time [215] that the analysis of large, labile, and highly polar compounds such as nucleotides [216] and oligosaccharides [217] might become possible. The next step was the development of fast atom bombardment, although it seems in retrospect that plasma desorption was the first method to allow molecular ions of very large molecules to be observed. Thus the whole range of biological polymers became accessible to mass spectrometry. This expectation was indeed largely fulfilled with laser desorption and electrospray, which appeared shortly afterwards. As an instructive early example, the spectrum of an intact protein is shown in Figure 31.

With these methods, large proteins, enzymes, or polynucleotides yield molecular ions, allowing the precise determination of molecular mass. In combination with chemical modification of amino acid side chains, the secondary and even the tertiary structure of proteins could be studied [218]. The synthesis of peptides, for example, became much more efficient, because the time-limiting step was formerly the analysis of the products, which can now be performed in minutes.

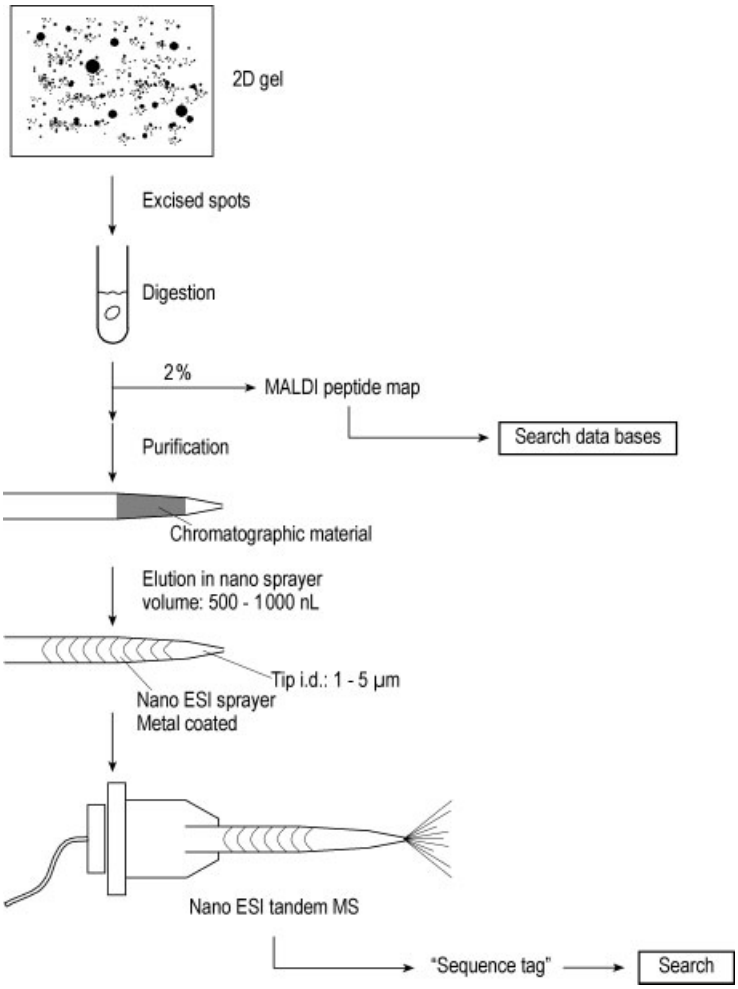

Figure 30. Strategy for the identification of proteins (after Protana, Odense)

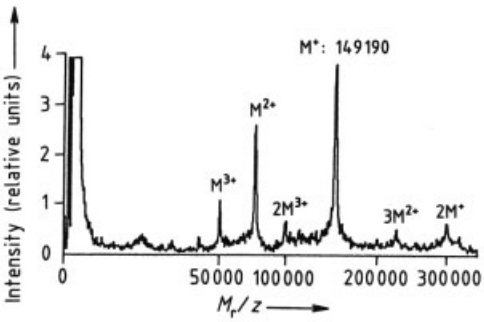

Figure 31. Matrix-assisted laser desorption (MALDI) spectrum of a carefully purified monoclonal antibody (IgG of mouse against a specific human lymphokine)
(Reproduced with permission)

In addition, electrospray is also well suited to the analysis of small, highly polar substances (e.g., drugs and metabolites [219]). Even non-covalent, highly specific interactions [220] have been studied. Apparently, the conformation of some biological polymers can be retained in the gas phase, allowing receptor – ligand complexes to survive the ionizing step. Examples include the interaction of the synthetic macrolide FK506 with its cytoplasmic receptor [221] and the detection of selfcomplementary DNA duplexes [222].

Since some of the techniques can be interfaced to liquid chromatography (LC/MS) or capillary zone electrophoresis (CZE/MS), the analysis of complex mixtures is also possible.

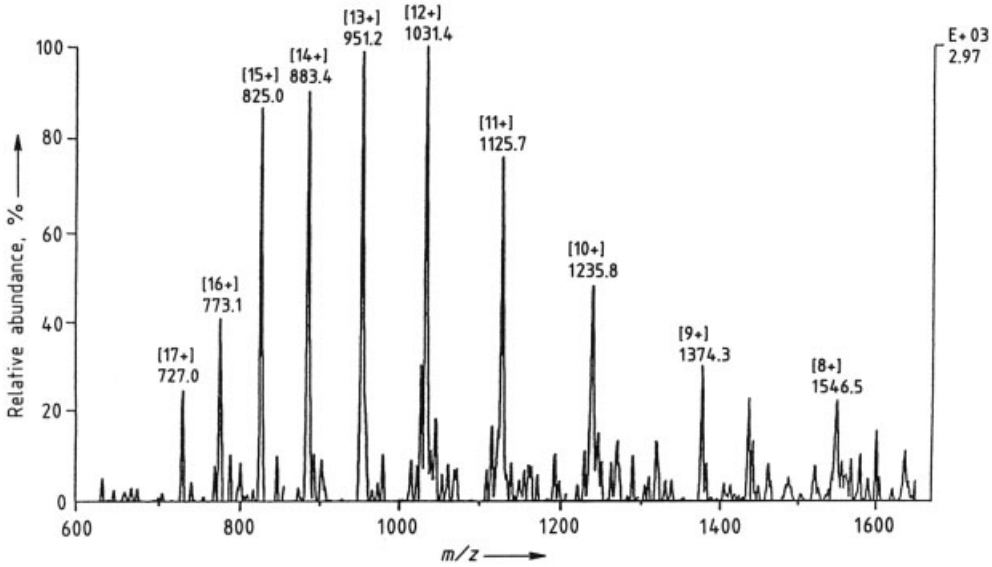

Figure 32. Electrospray spectrum of ca. 180 pmol cytochrome c using an MAT 90 sector field MS; homemade ESI source

Thus, these methods are suitable for the efficient design of pharmaceuticals and agrochemicals where careful analysis of the metabolic processes is necessary, but was formerly difficult and time consuming.

The detection capability of this type of instrumentation allows the reactions of xenobiotic compounds and drugs with macromolecules such as proteins or nucleotides [223] to be studied and very low concentrations of platinum-drug-modified biopolymers [224] in living organisms to be investigated. The metabolic or catabolic fate of drugs can be studied (e.g., the abuse of steroid hormones by athletes).

10.3. Determination of High Molecular Masses

One of the most important applications of mass spectrometry has always been the determination of molecular mass. Since this can be difficult with EIMS, the search for soft ionization techniques is almost as old as organic mass spectrometry. The first step was chemical ionization and particularly direct chemical ionization (DCI), which allowed the samples to be desorbed directly from a thin wire into the CI plasma. It has been successfully used for natural compounds such as glycerides. Later, field ionization (FI) and especially field desorption (FD) opened up new possibilities for the analysis of labile natural compounds such as saccharides, peptides, and nucleotides. When fast atom bombardment (FAB), plasma desorption (PD), electrospray ionization (ESI), and laser desorption (LD) mass spectrometry came into use, molecular ions of labile and very large molecules in excess of 200 000 amu could be determined with unparalleled accuracy. The electrospray spectrum of a protein allows its molecular mass to be calculated with high precision from a series of differently charged molecular ions (Fig. 32).

In this context, the discussion about resolution became a new relevance [225]. The use of nominal integral masses (e.g., 12 for C, 1 for H, 16 for O) is meaningless here as is true for the monoisotopic masses (e.g., 1.0078 for ^{1}H, 14.0030 for ^{14}N, 15.9949 for ^{16}O). Even with resolutions exceeding 10 000 it is impossible to resolve all the possible isobaric species in the molecular ion pattern of molecules such as peptides or oligonucleotides with hundreds of carbons and hydrogens, and tens of oxygen, nitrogen, and sulfur atoms. This is even more difficult with multiply charged ions as occurs regularly with electrospray. In such a case, the averaged chemical mass of the elements can be used for the calculation such as 12.011 for C, 1.008 for

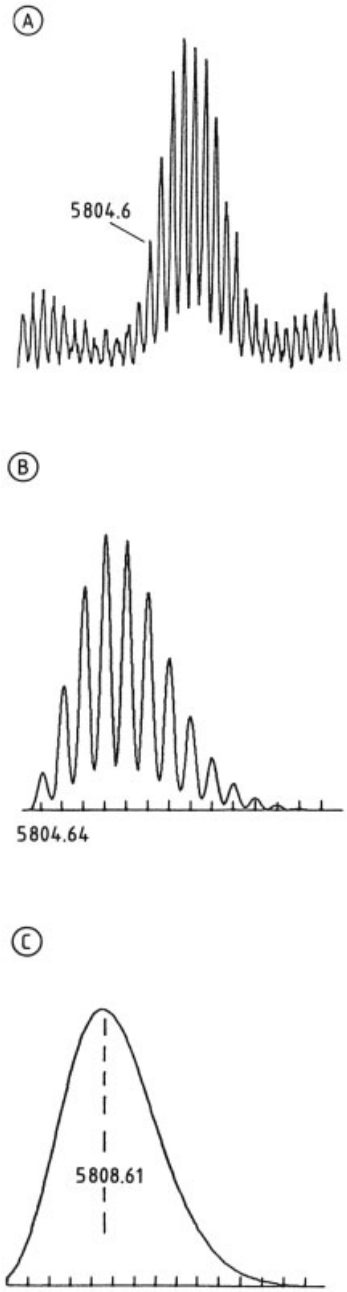

Figure 33. Effect of resolution on peak shapes
A) The protonated molecular cluster of human insulin ($C_{257}H_{384}N_{65}O_{77}S_6$) recorded with FABMS; B) Calculated at a resolution of 6000 (10 % valley definition) with the monoisotopic mass at $m/z = 5804.6455$; C) The unresolved cluster with an average mass at $m/z = 5808.61$
(Reproduced from [226], with permission)

H, 14.007 for N, 15.999 for O, 32.060 for S, and 30.974 for P. The resolution power of the instrument should always be set to give maximum sensitivity and, if possible, only sufficient separation power to resolve the major peaks due to carbon isotopes for singly charged molecules (see Fig. 33). For multiply charged species, even this is impossible but, nevertheless, information about the chemical molecular mass can be deduced from such spectra (all the isotopes are essentially merged into one peak). It is only necessary to distinguish between different biological modifications (e.g., phosphorylation, glycosylation).

If instruments can, however, achieve much higher resolution (100 000 or more) without sacrificing sensitivity, the gain in information can be striking: since from m/z, the distance between isotope signals is $1/z$, the charge state of a multiply charged ion signal can be read directly simply by counting the peaks within one mass unit. For example, 10 peaks at m/z 800–801 means that this ion carries 10 charges. The distance between the isotope signals is, therefore, $^1/^{10}$ of the mass. This can be very important for MS/MS work, since the charge state of fragments from multiply charged ions is generally unknown and the information about the "true" molecular mass is obscured. When the charge is known, sequencing of peptides on the basis of MS/MS analysis of electrospray ions becomes feasible. This is the reason why in the last few years FTMS with MALDI and electrospray have begun to be surprisingly employed in biomedical laboratories. The results reported on large peptides [145–147], [227–229] as well as on large oligonucleotides [147], [230–233] are so promising that even the high costs of such instruments cannot stop the development and not only research institutes but also industry have bought the FTMS instruments. As an illustration, data from the second isotope signal of a peptide show a resolution of 420 000, separating the isotope signals in this peak (Fig. 34).

In Figure 35, the product ions from a doubly charged tryptic digestion product illustrate the use. The product ion spectrum of the doubly protonated oligo-peptide Phe-Ser-Trp-Gly-Ala-Glu-Gly-Glu-Arg ($m/z = 520$) contains several significant singly charged sequence ions; this type of data is used for "mass tagging" in protein identification.

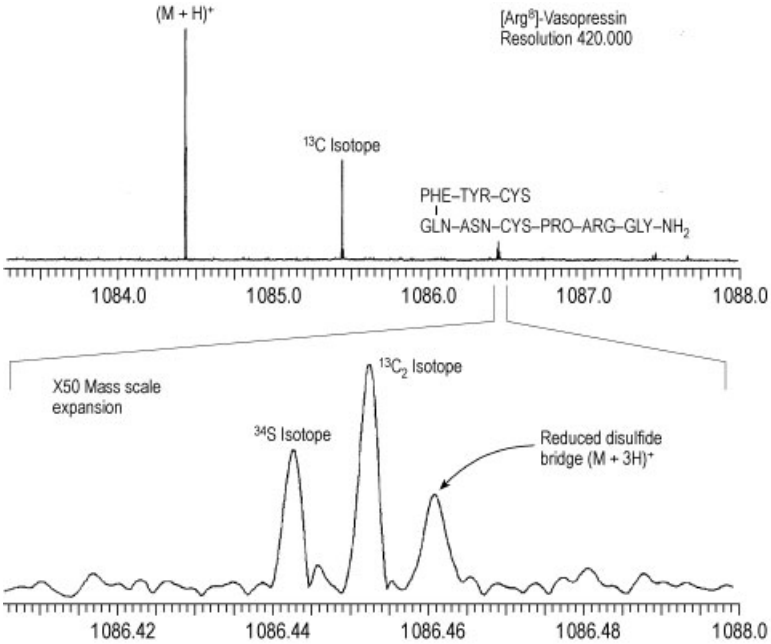

Figure 34. Top: 1 pmol of [arg8]-Vasopressin MALDI MS at a resolution of 420 000. Bottom: isobaric 34S (*m*/*z* 1086.4415) and 13C2 (*m*/*z* 1086.4524) and a third peak due to the reduced disulfide bridge are clearly distinguished

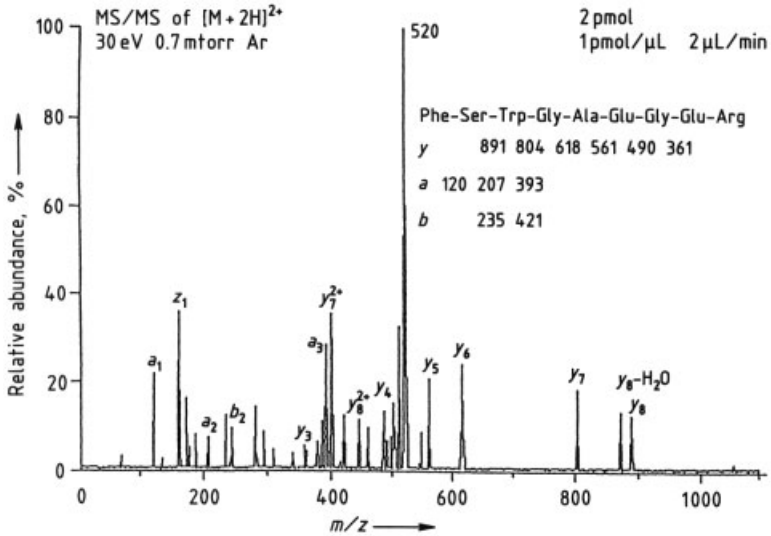

Figure 35. Doubly charged ions from tryptic digests of proteins
The fragments are labeled according to the scheme given in Figure 29 (Reproduced with permission from [238])

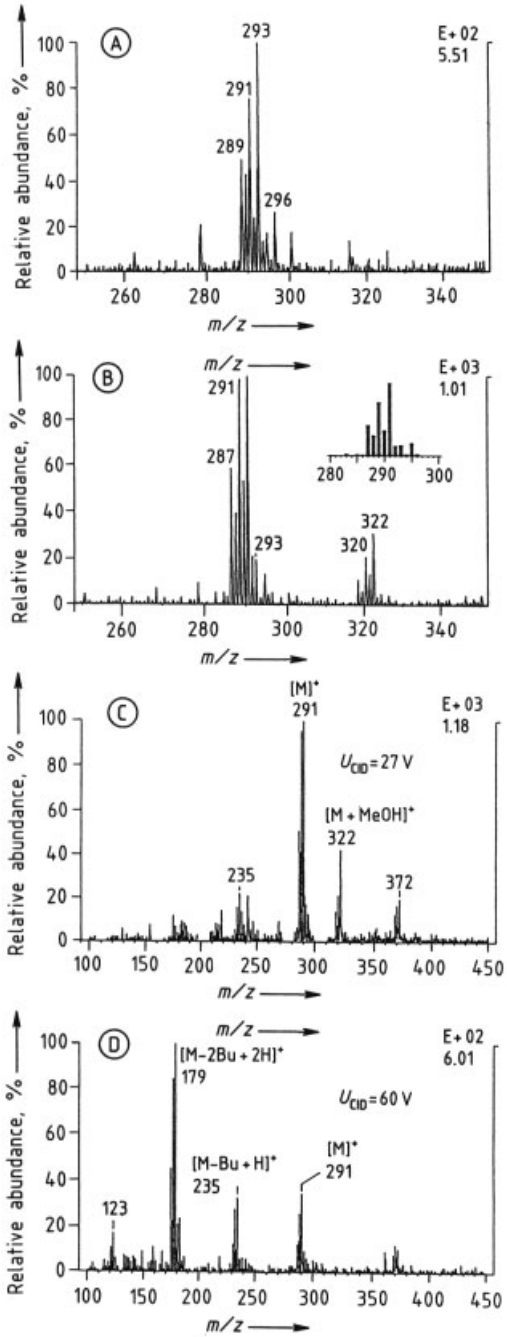

Figure 36. ESI spectra of some tin compounds
A) Dibutyl tin acetate (DBT) $(C_4H_9)_2Sn^{2+}(OCOCH_3)^-$; B) Tributyl tin (TBT) $(C_4H_9)_3Sn^+$ cations with theoretical distribution given in the inset; C) TBT with small CAD voltage; D) TBT with large CAD voltage showing fragments due to loss of the side chains
Instrument: MAT 90, electrospray MS, flow injection
Upper spectrum: CAD voltage 27 V; Lower spectrum: CAD voltage 60 V

10.4. Species Analysis

The storage, transport, and action of metals in the environment is largely dependent on the chemical form such as oxidation state and association with organic ligands. Heavy metals in the environment are stored in complexes with humic acids, can be converted by microbes into different complexes, and may be transported into living animals and human beings. This is true for many elements such as lead, mercury [234], arsenic [235], [236], astatine, tin, and platinum [237]. Several mass spectrometric techniques have been employed in the study of the fate of metals in the human body and other organic environments by the determination of the species formed in the biological or environmental matrices.

For example, tin and lead alkylates have been determined in soil, water, or muscle tissue with GC/MS after exhaustive alkylation or with thermospray [239], API [240], and ICP – LC/MS [244] methods. The ESI spectra of some tin compounds is given in Figure 36. The APIMS techniques have proven to be very successful, even with metals bound to proteins and enzymes [241] and, interfaced to microseparation techniques, even elements such as iodine [242] or phosphorus [243] can be determined quantitatively.

10.5. Determination of Elements

Mass spectrometry as a multielement technique has the advantage that many metal ions can be detected and quantitatively determined at once. In bulk materials such as steel [245] or refractory metals [246], elements can be determined by means of low-resolution glow-discharge mass spectrometry. High-resolution GDMS has been used to study semiconducting materials [247]. GDMS is regarded as being virtually free of matrix effects [248]. The state of the art in glow discharge mass spectrometry has been reviewed, and from the data presented there it is evident that the technique is a mature tool for the material sciences [249]. If part of the material can be dissolved prior to mass spectral anal-

Table 2. Detection limits in ICPMS [65]

Element	Detection limit, pg/mL	Element	Detection limit, pg/mL
Li	100	Sn	60
B	400	Sb	60
F	1.1×10^5	Te	80
Mg	130	I	10
Al	160	Cs	20
P	2×10^4	Ba	150
S	1×10^5	La	10
Cl	1×10^3	Ce	10
Ti	320	Pr	10
V	80	Nd	45
Cr	10	Sm	30
Mn	30	Eu	15
Co	10	Gd	50
Ni	40	Tb	10
Cu	20	Dy	35
Zn	10	Ho	10
Ge	20	Er	35
As	40	Tm	10
Se	800	Yb	30
Br	1×10^3	Lu	10
Y	50	Hf	10
Zr	30	Ta	50
Mo	40	W	50
Ru	60	Au	60
Ag	30	Hg	20
Cd	30	Pb	10
In	60	Th	20
		U	10

ysis, ICPMS or, for high precision isotope determination, thermal ionization mass spectrometry (TIMS) can be applied. Detection limits in ICPMS are listed in Table 2.

ICPMS and TIMS are also suitable for the analysis of solutions such as biological fluids or water samples. In combination with laser ablation, bulk materials or small biological compartments can be analyzed with ICPMS (see Section 4.5). The desorbed neutrals, ions, and clusters are entrained in the gas, which carries the aerosol directly into the plasma. One important area of application is geochronology by ICPMS and AMS; for a comprehensive description of this field see [250].

10.6. Surface Analysis and Depth Profiling

The characterization of material surfaces and of the underlying layers is one of the most important problems in modern material science, since the properties of a material are not only dependent on the bulk concentration of elements, but also on the distribution on the surface, in the upper layers, and at grain boundaries. Therefore, techniques have been developed which can analyze the surface and the next layers with very high resolution. The first two monolayers, ca. 0.5 nm in depth, can be investigated by secondary ion mass spectrometry (SIMS) [81]. Positive and negative ions from the elements in the surface can be detected with a sensitivity in the ppm range. It has found many applications in the analysis of metals, alloys, semiconductors, ceramics, and glasses. The problem with SIMS is that matrix effects can cause difficulties in quantitation. It is possible to employ SIMS for the analysis of organic compounds on a metal surface. This technique, called static SIMS, makes use of the sputtered organic ions from material deposited on the surface (typically Ag), in monolayers. Since the ion current is weak and lasts for only a few seconds, time-of-flight instruments are generally used.

For the depth profiling of a metal, of particulate matter, or a ceramic with a resolution of

a few nanometers, sputtered neutral mass spectrometry (SNMS) [251] is the best choice since the sputtering process is completely decoupled from the ionization, and quantitation is possible by calibrating the instrument with standard reference materials. The technique has found many applications for the characterization of semiconductors and implanted atoms. Figure 37 gives an example compared to a result using GDMS.

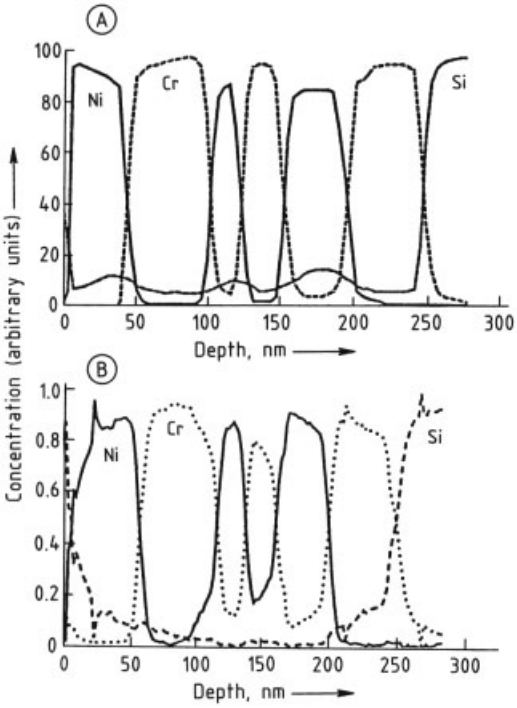

Figure 37. Depth profiles of a multilayer sample (Ni, Cr 50 nm, Ni, Cr 25 nm, Ni, Cr 50 nm)
A) SNMS; B) GDMS
(Reproduced with permission from [252])

The determination of the spatial distributions of elements is becoming increasingly important since the grain boundaries in ceramics or the metal distribution in microelectronic devices have a major influence on the stability of the material and the performance of the device. With a highly focused primary ion beam it is possible to scan the surface in two dimensions and the distribution of the elements on the surface. Combinations of sputtering processes and microprobing allow this distribution to be investigated in depth. Instruments used for this type of analysis are called ion microprobe mass analyzers (IMMA). The spot of the primary beam has a diameter of less than 10 µm, thus defining the resolution obtainable.

11. References

1. F. W. Aston: *Mass Spectra and Isotopes,* Edward Arnold & Co, London 1933.
2. A. J. Dempster, *Phys. Rev.* **11** (1918) 316.
3. H. W. Washburn, H. F. Wiley, S. M. Rock, *Ind. Eng. Chem. Anal. Ed.* **15** (1943) 541 – 547.
4. R. A. Brown, R. C. Taylor, F. W. Melpolder, W. S. Young, *Anal. Chem.* **20** (1948) 5 – 9.
5. M. J. O'Neal, Jr., *Anal. Chem.* **22** (1950) 991 – 995.
6. A. J. Campbell, J. S. Halliday in R. M. Elliott (ed.): *13th Annual Symposium on Mass Spectrometry,* vol. 2, Pergamon Press, Oxford 1962.
7. H. Budzikiewicz, C. Djerassi, D. H. Williams: *Structure Elucidation of Natural Products by Mass Spectrometry,* Holden Day, San Francisco 1964.
8. G. R. Waller (ed.): *Biochemical Applications of Mass Spectrometry,* Wiley Interscience, New York 1972.
9. D. Kanne, K. S. Straub, J. E. Hearst, H. Rapoport, *J. Am. Chem. Soc.* **103** (1982) 6754 – 6764.
10. H. M. Rosenstock, M. B. Wallenstein, A. L. Wahrhaftig, H. Eyring, *Proc. Natl. Acad. Sci. USA* **38** (1952) 667.
11. R. A. Marcus, O. K. Rice, *J. Phys. Colloid. Chem.* **55** (1951) 894.
12. K. Levsen: *Fundamental Aspects of Organic Mass Spectrometry,* VCH Verlagsgesellschaft, Weinheim 1978.
13. F. W. McLafferty, F. Turecek: *Interpretation of Mass Spectra,* University Science Books, Mill Valley, California 1993.
14. J. J. Boon, *Int. J. Mass Spectrom. Ion Processes* **118/119** (1992) 755 – 787.
15. W. Genuit, J. J. Boon, O. Faix, *Anal. Chem.* **59** (1987) 508 – 513.
16. W. H. McFadden: *Techniques of Combined Gas Chromatography/Mass Spectrometry,* Wiley & Sons, New York 1973.
17. W. A. Brand, *J. Mass Spectrom.* **31** (1996) 225.
18. R. M. Caprioli, T. Fan, *Biochem. Biophys. Res. Commun.* **141** (1986) 1058.
19. Y. Ito, T. Takeuchi, D. Ishii, M. Goto, *J. Chromatogr.* **346** (1985) 161.
20. M. A. Moseley et al., *Anal. Chem.* **61** (1989) 1577.

21. R. M. Caprioli et al., *J. Chromatogr.* **480** (1989) 247.
22. M. A. Moseley, L. J. Deterding, K. B. Tomer, J. W. Jorgenson, *Rapid Commun. Mass Spectrom.* **3** (1989) 87.
23. N. J. Reinhoud et al., *J. Chromatogr.* **516** (1990) 147.
24. M. A. Baldwin, F. W. McLafferty, *Org. Mass Spectrom.* **7** (1973) 1111.
25. V. L. Tal'roze, G. V. Karpov, I. G. Gordetskii, V. E. Skurat, *Russ. J. Phys. Chem. (Engl. Transl.)* **42** (1968) 1658.
26. J. D. Henion, *J. Chromatogr. Sci.* **19** (1981) 57.
27. R. D. Smith, W. D. Felix, J. C. Fjedsted, M. L. Lee. *Anal. Chem.* **54** (1982) 1885.
28. R. C. Willoughby, R. F. Browner, *Anal. Chem.* **56** (1984) 2625.
29. E. C. Horning, M. G. Horning, D. I. Carroll, R. N. Stillwell, I. Dzidic, *Life Sci.* **13** (1973) 1331.
30. J. Sunner, G. Nicol, P. Kebarle, *Anal. Chem.* **60** (1988) 1300.
31. J. Sunner, M. G. Ikonomou, P. Kebarle, *Anal. Chem.* **60** (1988) 1308.
32. C. R. Blakley, J. C. Carmody, M. L. Vestal, *Clin. Chem. (Winston Salem N.C.)* **26** (1980) 1467.
33. J. V. Iribarne, B. A. Thomson, *J. Chem. Phys.* **64** (1976) 2287.
34. B. A. Thomson, J. V. Iribarne, *J. Chem. Phys.* **71** (1979) 4451.
35. G. Schmelzeisen-Redeker, L. Bütfering, F. W. Röllgen, *Int. J. Mass Spectrom. Ion Processes* **90** (1989) 139.
36. M. L. Vestal, *Int. J. Mass Spectrom. Ion Phys.* **46** (1983) 193.
37. J. Hau, W. Nigge, M. Linscheid, *Org. Mass Spectrom.* **28** (1993) 223 – 229.
38. M. Linscheid, D. Westmoreland, *Pure Appl. Chem.,* 66 (1994).
39. M. A. Niessen, J. van der Greef: *Liquid Chromatography-Mass Spectroscopy,* Marcel Dekker, New York 1992.
40. R. D. Smith et al., *Anal. Chem.* **65** (1993) 574A –584A.
41. W. M. Niessen, *J. Chromatogr. A* **856** (1999) 179.
42. W. M. A. Niessen, A. P. Tinke, *J. Chromatogr. A* **703** (1995) 37.
43. M. R. Emmett, F. M. White, C. L. Hendrickson, S.-H. Shi, A. G. Marshall, *J. Am. Soc. Mass Spectrom.* **9** (1997) 333.
44. I. M. Lazar, A. L. Rockwood, E. D. Lee, J. C. H. Sin, M. L. Lee, *Anal. Chem.* **71** (1999) 2578.
45. R. D. Smith, J. A. Olivares, N. T. Ngyen, H. R. Udseth, *Anal. Chem.* **60** (1988) 436 – 441.
46. C. G. Edmonds et al., *Biochem. Soc. Trans.* **19** (1991) 943.
47. S. A. Hofstadler, F. D. Swanek, D. C. Gale, A. G. Ewing, R. D. Smith, *Anal. Chem.* **67** (1995) 1477.
48. J. C. Severs, S. A. Hofstadler, Z. Zhao, R. T. Senh, R. D. Smith, *Electrophoresis* **17** (1996) 1808.
49. I. M. Johannson, E. C. Huang, J. D. Henion, J. Zweigenbaum, *J. Chromatogr.* **554** (1991) 311.
50. A. L. Burlingame, J. A. McCloskey, *Biol. Mass Spectrom., Proc. Int. Symp. Mass Spectrom. Health Life Sci.*, Elsevier, Amsterdam 1990.
51. R. S. Ramsey, J. M. Ramsey, *Anal. Chem.* **69** (1997) 1174.
52. R. D. Oleschuk, D. J. Harrison, *Trends Anal. Chem.* **19** (2000) 379.
53. K. L. Busch, *J. Planar Chromatogr.-Mod. TLC* **2** (1989) 355.
54. A. Hayashi, T. Matsubara, Y. Nishizawa, T. Hattori, *Iyo Masu Kenkyukai Koenshu* **11** (1986) 147.
55. L. Li, D. M. Lubman, *Anal. Chem.* **61** (1989) 1911.
56. K. L. Bush, *TrAc Trends Anal. Chem. (Pers. Ed.)* **11** (1992) 314.
57. I. D. Wilson, *J. Chromatogr. A* **856** (1999) 429.
58. J. T. Mehl, D. M. Hercules, *Anal. Chem.* **72** (2000) 68.
59. Y. Liu, J. Bai, X. Liang, D. M. Lubman, *Anal. Chem.* **67** (1995) 3482.
60. M. Schreiner, K. Strupat, F. Lottspeich, C. Eckerskorn, *Electrophoresis* **17** (1996) 954.
61. C. Eckerskorn et al., *Anal. Chem.* **69** (1997) 2888.
62. D. Schleuder, F. Hillenkamp, K. Strupat, *Anal. Chem.* **71** (1999) 3238.
63. A. G. Harrison: *Chemical Ionization Mass Spectrometry,* CRC Press, Boca Raton 1982.
64. G. B. Mohamed et al., *J. Chromatogr.* **314** (1984) 211 –217.
65. H. D. Beckey: *Principles of Field Ionization and Field Desorption Mass Spectrometry,* Pergamon Press, Frankfurt 1977.
66. G. M. Hieftje, L. A. Norman, *Int. J. Mass Spectrom. Ion Processes* **118/119** (1992) 519 – 573.
67. W. W. Harrison, *J. Anal. At. Spectrom.* **7** (1992) 75 –79.
68. W. W. Harrison et al., "Glow Discharge Techniques in Analytical Chemistry," *Anal. Chem.* **62** (1990) 943A – 949A.

69. H. Oechsner, *Pure Appl. Chem.* **64** (1992) 615.
70. A. Montaser, H. Tan, I. Ishii, S. H. Nam, M. Cai, *Anal. Chem.* **63** (1991) 2660.
71. N. Jakubowski, L. Moens, F. Vanhaecke, *Spectrochim. Acta B: At. Spectrosc.* **53** (1998) 1739.
72. M. Morita, H. Ito, M. Linscheid, K. Otsuka, *Anal. Chem.* **66** (1994) 1588.
73. K. E. Milgram et al., *Anal. Chem.* **69** (1997) 3714.
74. J. R. Eyler, R. S. Houk, C. H. Watson, K. E. Milgram: "Development and Application of Inductively Coupled Plasma (ICP)-FTICR Mass Spectrometry," *46th ASMS Conference*, Orlando, Fl 1998.
75. N. Jakubowski, D. Stuewer, unpublished results.
76. K. G. Heumann, *Int. J. Mass Spectrom. Ion Processes* **118/119** (1992) 575 – 592.
77. W. Genuit, Chen He-Neng, A. J. H. Boerboom, J. Los, *Int. J. Mass Spectrom. Ion Phys.* **51** (1983) 207 – 213.
78. U. Boesl, J. Grotemeyer, K. Walter, E. W. Schlag, *Anal. Instrum. (N. Y.)* **16** (1987) 151.
79. J. Grotemeyer, U. Boesl, K. Walter, E. W. Schlag, *Org. Mass Spectrom.* **21** (1986) 645.
80. J. Grotemeyer, E. W. Schlag, *Acc. Chem. Res.* **22** (1989) 399.
81. A. Benninghoven, *Z. Phys.* **230** (1970) 403.
82. L. Schwieters et al.: "Secondary Ion Mass Spectrometry," in A. Benninghoven, K. T. F. Janssen, J. Tümpner, H. W. Werner (eds.): *Proc. 8th Int. Conf. (SIMS VIII)* **1992,** 497.
83. P. M. Thompson, *Anal. Chem.* **63** (1991) 2447 – 2456.
84. B. Hagenhoff et al., *J. Phys. D, Appl. Phys.* **25** (1992) 818.
85. F. W. Roellgen in A. Benninghoven (ed.): *Ion Formation from Organic Solids,* Springer-Verlag, Berlin 1983, p. 2.
86. M. Barber, R. S. Bordoli, R. D. Sedgwick, A. N. Tyler, *J. Chem. Soc. Chem. Commun.* 1981, 325 –327.
87. R. D. McFarlane, D. F. Torgerson, *Science (Washington, D.C. 1883)* **191** (1976) 920.
88. R. D. McFarlane: "Methods in Enzymology," vol. 193 in J. A. McCloskey (ed.): *Mass Spectrometry,* Academic Press, San Diego 1990, pp. 263 – 280.
89. G. Jonsson et al., *Rapid Commun. Mass Spectrom.* **3** (1989) 190.
90. A. H. Verbueken, F. J. Bryunseels, R. E. van Grieken, *Biomed. Mass Spectrom.* **12** (1985) 438.
91. J. S. Becker, H. J. Dietze, *Fresenius J. Anal. Chem.* **344** (1992) 69 – 86.
92. F. Hillenkamp: "Ion Formation from Organic Solids II," in A. Benninghoven (ed.): *Springer Series in Chem. Phys.* **1983,** no. 25, p. 190.
93. M. Karas, U. Bahr, A. Ingendoh, F. Hillenkamp, *Angew. Chem.* **101** (1989) 805 – 806.
94. M. Karas, D. Bachmann, F. Hillenkamp, *Anal. Chem.* **57** (1985) 2935.
95. M. Karas, F. Hillenkamp, *Anal. Chem.* **60** (1988) 2299 – 2301.
96. K. Tanaka et al., *Rapid Commun. Mass Spectrom.* **8** (1988) 151.
97. F. Kirpekar, S. Berkenkamp, F. Hillenkamp, *Anal. Chem.* **71** (1999) 2334.
98. R. Kaufmann, T. Wingerath, D. Kirsch, W. Stahl, H. Sies, *Anal. Biochem.* **238** (1996) 117.
99. P. Juhasz et al., *Anal. Chem.* **68** (1996) 941.
100. U. Bahr, J. Stahl-Zeng, E. Gleitsmann, M. Karas, *J. Mass Spectrom.* **32** (1997) 1111.
101. R. Kaufmann, *J. Biotechnol.* **41** (1995) 155.
102. B. Sprengler, *J. Mass Spectrom.* **32** (1997) 1019.
103. P. Chaurand, F. Luetzenkirchen, B. Spengler, *J. Am. Soc. Mass Spectrom.* **10** (1999) 91.
104. M. Dole et al., *J. Chem. Phys.* **49** (1968) 2240.
105. M. Yamashita, J. B. Fenn, *J. Phys. Chem.* **88** (1984) 4451 – 4459.
106. C. M. Whitehouse, R. N. Dreyer, M. Yamashita, J. B. Fenn, *Anal. Chem.* **57** (1985) 675 – 679.
107. M. L. Alexandrov, L. N. Gall', N. V. Krasnov, V. I. Nikolaev, V. A. Shkurov, *Zh. Anal. Khim.* **40** (1985) 1227 – 1236.
108. L. Tang, P. Kebarle, *Anal. Chem.* **65** (1993) 3654 –3668.
109. S. K. Chowdhury, V. Katta, B. T. Chait, *Rapid Commun. Mass Spectrom.* **4** (1990) 81 – 87.
110. A. P. Bruins, T. R. Covey, J. D. Henion, *Anal. Chem.* **59** (1987) 2642 – 2646.
111. M. Wilm, M. Mann, *Anal. Chem.* **68** (1996) 1.
112. A. J. Dempster, *Phys. Rev.* **11** (1918) 316.
113. C. Brunnée, *Int. J. Mass Spectrom. Ion Processes* **76** (1987) 121 – 237.
114. E. W. Blauth: *Dynamische Massenspektrometer,* Vieweg & Sohn, Braunschweig 1965.
115. R. Kaufmann, B. Spengler, F. Lützenkirchen, *Rapid Commun. Mass Spectrom.* **7** (1993) 902.
116. I. V. Chernushevich, W. Ens, K. G. Standing, *Anal. Chem.* **71** (1999) 452A.
117. I. M. Lazar, R. S. Ramsey, S. Sundberg, J. M. Ramsey, *Anal. Chem.* **71** (1999) 3627.

118. M. B. Comisarow, A. G. Marshall, *Chem. Phys. Lett.* **26** (1974) 489; M. B. Comisarow, A. G. Marshall, *J. Chem. Phys.* **64** (1976) 110.
119. C. B. Lebrilla, D. T. S. Wang, R. L. Hunter, R. T. McIver, Jr., *Anal. Chem.* **62** (1990) 878 – 880.
120. C. F. Ijames, C. F. Wilkins, *J. Am. Chem. Soc.* **110** (1988) 2687 – 2688.
121. A. G. Marshall, *Acc. Chem. Res.* **18** (1985) 316 – 322.
122. C. Giancaspro, F. R. Verdun, *Anal. Chem.* **58** (1986) 2097.
123. R. T. McIver, R. L. Hunter, W. D. Bowers, *Int. J. Mass Spectrom. Ion Processes* **64** (1985) 67.
124. P. Kofel, M. Allemann, H. Kellerhals, K. P. Wanczek, *Int. J. Mass Spectrom. Ion Processes* **72** (1986) 53.
125. P. Grossmann, P. Caravatti, M. Allemann, H. P. Kellerhals, *Proc. 36th ASMS Conf. Mass Spectrom. All. Top.* 1988, 617.
126. W. Paul, H. Steinwedel, *Z. Naturforsch. A* : Phys. Phys. Chem. Kosmophys. 8 (1953) 448.
127. G. C. Stafford et al., *Int. J. Mass Spectrom. Ion Processes* **60** (1984) 85.
128. B. D. Nourse, R. G. Cooks, *Anal. Chim. Acta* **228** (1990) 1.
129. H.-J. Laue, K. Habfast, *Proc. 37th ASMS Conf. Mass Spectrom. All. Top.* 1989, p. 1033.
130. P. De Bievre, *Fresenius J. Anal. Chem.* **337** (1990) 767.
131. R. E. M. Hodges in J. F. J. Todd (ed.): *Proc. 10th Int. Mass. Spectrom. Conf. Swansea 1985,* Wiley, London 1986, p. 185.
132. W. Aberth, *Biomed. Mass Spectrom.* **7** (1980) 367.
133. H. Matsuda, H. Wollnik, *Int. J. Mass Spectrom. Ion Processes* **86** (1988) 53.
134. M. Barber, R. M. Elliott, paper presented at ASTM E-14 Conference on Mass Spectrometry, Montreal 1964.
135. K. L. Busch, G. L. Glish, S. A. McLuckey: *Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry,* VCH Verlagsgesellschaft, Weinheim 1988.
136. M. A. Seeterlin, P. R. Vlasak, D. J. Beussmann, R. D. McLane, C. G. Enke, *J. Am. Soc. Mass Spectrom.* **4** (1993) 751 – 754.
137. S. R. Horning, J. M. Wood, J. R. Gord, B. S. Freiser, R. G. Cooks, *Int. J. Mass Spectrom. Ion Processes* **101** (1990) 219.
138. S. A. McLuckey, D. E. Goeringer, *J. Mass Spectrom.* **32** (1997) 461.
139. R. A. Yost, C. G. Enke, *J. Am. Chem. Soc.* **100** (1978) 2274.
140. R. E. March, R. J. Hughes, *Quadrupole Storage Mass Spectrometry,* Wiley Interscience, New York 1989.
141. C. Hagg, I. Szabo, *Int. J. Mass Spectrom. Ion Processes* **73** (1986) 295.
142. K. Sato et al., *Anal. Chem.* **59** (1987) 1652.
143. J. A. Hill et al., *Int. J. Mass Spectrom. Ion Processes* **92** (1989) 211.
144. I. J. Amster, *J. Mass Spectrom.* **31** (1996) 1325.
145. V. C. M. Dale et al., *Biochem. Mass Spectrom. Biochem. Soc. Trans.* **24** (1996) 943.
146. M. W. Senko, J. P. Speir, F. W. McLafferty, *Anal. Chem.* **66** (1994) 2801.
147. E. R. Williams, *Anal. Chem.* **10** (1998) 179A.
148. A. Ingendoh et al., *J. Mass Spectrom. Soc. Jpn.* **45** (1997) 247.
149. R. E. March, *J. Mass Spectrom.* **32** (1997) 351.
150. R. J. Strife, M. Bier, J. Zhou, *Proc. 43rd ASMS Conference on Mass Spectrometry and Allied Topics, (Atlanta),* (1995) 1113.
151. J. C. Schwartz, M. Bier, D. M. Taylor, J. Zhou, *Proc. 43rd ASMS Conference on Mass Spectrometry and Allied Topics, (Atlanta)*, (1995) 1114.
152. E. R. Badman, J. M. Wells, H. A. Bui, R. G. Cooks, *Anal. Chem.* **70** (1998) 3545.
153. H. R. Morris et al., *Rapid Commun. Mass Spectrom.* **10** (1996) 889.
154. C. Borchers, C. E. Parker, L. J. Deterding, K. B. Tomer, *J. Chromatogr. A* **854** (1999) 119.
155. T. J. Cornish, R. J. Cotter, *Anal. Chem.* **65** (1993) 1043 – 1047.
156. F. H. Strobel, L. M. Preston, K. S. Washburn, D. H. Russell, *Anal. Chem.* **64** (1992) 754 – 762.
157. M. Linscheid, *Fresenius Z. Anal. Chem.* **337** (1990) 648 – 661.
158. N. R. Daly, *Rev. Sci. Instrum.* **34** (1963) 1116.
159. C. E. Griffin, H. G. Boettger, D. D. Norris, *Int. J. Mass Spectrom. Ion Phys.* **15** (1974) 437.
160. J. S. Cottrell, S. Evans, *Anal. Chem.* **59** (1987) 1990 –1995.
161. C. E. D. Ouwerkerk, A. J. H. Boerboom, T. Matsuo, T. Sakurai, *Int. J. Mass Spectrom. Ion Phys.* **70** (1986) 79 – 96.
162. R. Pesch, G. Jung, K. Rostand, K.-H. Tietje, *Proc. 37th ASMS Conf. Mass Spectrom. All. Top.* 1989, 1079.
163. S. Bauerschmidt, W. Hanebeck, K.-P. Schulz, J. Gasteiger, *Anal. Chim. Acta* **265** (1992) 169 – 182.
164. J. Gasteiger, W. Hanebeck, K.-P. Schulz, *I. Chem. Inf. Comp. Sci.* **32** (1992) 264 – 271.

165. K. Pfleger, H. H. Maurer, A. Weber: *Mass Spectral and GC Data of Drugs, Poisons, Pesticides, Pollutants and their Metabolites,* 2nd ed., VCH Verlagsgesellschaft, Weinheim 1992.
166. H. Damen, D. Henneberg, B. Weimann, *Anal. Chim. Acta* **103** (1978) 289.
167. D. Henneberg, *Anal. Chim. Acta* **150** (1983) 37.
168. G. M. Pesyna, R. Venkataraghavan, H. G. Dayringer, F. W. McLafferty, *Anal. Chem.* **48** (1976) 1362.
169. B. L. Atwater, D. B. Stauffer, F. W. McLafferty, D. W. Peterson, *Anal. Chem.* **57** (1985) 899.
170. S. Sokolov, J. Karnovsky, P. Gustafson, *Finnigan Appl. Rep.* no. 2, 1978.
171. D. Henneberg, B. Weimann, U. Zalfen, *Org. Mass Spectrom.* **28** (1993) 198 – 206.
172. M. J. Hayward, P. V. Robandt, J. T. Meek, M. L. Thomson, *J. Am. Soc. Mass Spectrom.* **4** (1993) 742 –750.
173. T. Kotiaho, *J. Mass Spectrom.* **31** (1996) 1.
174. P. Ciccioli, A. Cecinato, E. Brancaleoni, M. Frattoni, A. Liberti, *HRC CC J. High Resolut. Chromatogr.* **15** (1992) 75 – 84.
175. T. Hoffmann, P. Jacob, M. Linscheid, D. Klockow, *Int. J. Environ. Anal. Chem.* **52** (1993) 29 – 37.
176. A. Robbat, C. J. Liu, T. Y. Liu, *J. Chromatogr.* **625** (1992) 277 – 288.
177. P. J. McKeown, M. V. Johnston, D. M. Murphy, *Anal. Chem.* **63** (1991) 2073 – 2075.
178. K. G. Asano, G. L. Glish, S. A. McLuckey, *Proc. 36th ASMS Conf. Mass Spectrom. All. Top.* 1988, 637.
179. T. Hoffmann, R. Bandur, U. Marggraf, M. Linscheid, *J. Geophys. Res. D: At Armospheres* **103** (1998) 25.
180. M. A. LaPack, J. C. Tou, C. G. Enke, *Anal. Chem.* **62** (1990) 1265 – 1271.
181. R. Zimmermann, H. J. Heger, A. Kettrup, U. Boesl, *Rapid Commun. Mass Spectrom.* **11** (1997) 1095.
182. R. Zimmermann et al., *Rapid Commun. Mass Spectrom.* **13** (1999) 307.
183. U. Boesl, *J. Mass Spectrom.* **35** (2000) 289.
184. G. Durand, N. de Bertrand, D. Barcelo, *J. Chromatogr.* **554** (1991) 233 – 250.
185. H. Fr. Schröder, *J. Chromatogr.* **554** (1991) 251 – 266.
186. M. A. Brown, I. S. Kim, F. I. Sasinos, R. D. Stephens: "Liquid Chromatography/Mass Spectrometry, Applications in Agricultural, Pharmaceutical and Environmental Chemistry," in M. A. Brown (ed.): *ACS Symp. Ser.* **420** (1990) 199 – 214.
187. A. K. Lister, R. G. Cooks, *Proc. 36th ASMS Conf. Mass Spectrom. All. Top.* 1988, 646.
188. C. Rappe, S. Marklund, R.-A. Berqvist, M. Hanson in G. Choudhary, L. H. Keith, C. Rappe (eds.): *Chlorinated Dioxins and Dibenzofurans in the Total Environment,* Butterworth, Stoneham 1983, chap. 7.
189. R. M. Smith et al., *Environ. Sci. Res.* **26** (1983) 73 –94.
190. L. Q. Huang, A. Paiva, H. Tond, S. J. Monson, M. L. Gross, *J. Am. Soc. Mass Spectrom.* **3** (1992) 248 – 259.
191. E. R. Barnhart et al., *Anal. Chem.* **59** (1987) 2248 –2252.
192. W. A. Korfmacher, G. F. Moler, R. R. Delongchamp, R. K. Mitchum, R. L. Harless, *Chemosphere* **131** (1984) 669 – 685.
193. M. Linscheid, *Int. J. Environm. Anal. Chem.* **49** (1992) 1 – 14.
194. T. L. Jones, L. D. Betowski, J. Yinon: "Liquid Chromatography/Mass Spectrometry, Applications in Agricultural, Pharmaceutical and Environmental Chemistry," in M. A. Brown (ed.): *ACS Symp. Ser.* **420** (1990) 62.
195. W. M. Draper, F. R. Brown, R. Bethem, M. J. Miille: "Liquid Chromatography/Mass Spectrometry, Applications in Agricultural, Pharmaceutical and Environmental Chemistry," in M. A. Brown (ed.): *ACS Symp. Ser.* **420** (1990) 253.
196. A. Robbat, T. Y. Liu, B. M. Abraham, *Anal. Chem.* **64** (1992) 1477 – 1483.
197. J. A. McCloskey (ed.): "Mass Spectrometry," *Methods in Enzymology,* **vol. 193,** Academic Press, San Diego 1990.
198. H. M. Leith, P. L. Truran, S. J. Gaskell, *Biomed. Environm. Mass Spectrom.* **13** (1986) 257 – 261.
199. C. H. L. Shackleton in H. Jaeger (ed.): *Glass Capillary Chromatography in Clinical Medicine and Pharmacology,* Marcel Dekker, New York 1985.
200. A. Frigerio, C. Marchioro, A. M. Pastorino in D. D. Breimer, P. Speiser (eds.): *Top. Pharm. Sci. Proc. Int. Congr. Pharm. Sci. F.I.P.* **45th** (1985) 297 –310.
201. H. Schweer, C. O. Meese, O. Fuerst, P. K. Gonne, H. J. W. Seyberth, *Anal. Biochem.* **164** (1987) 156 –163.
202. H. R. Morris, D. H. Williams, R. P. Ambler, *Biochem. J.* **125** (1971) 189.
203. V. N. Reinhold in S. J. Gaskell (ed.): *Mass Spectrometry in Biomedical Research,* J. Wiley & Sons, Chichester 1986, pp. 181.

204. D. Rolf, G. R. Gray, *Carbohydr. Res.* **137** (1985) 183 – 196.
205. H. Björndal, C. G. Hellerqvist, B. Lindberg, S. Svensson, *Angew. Chem.* **82** (1979) 643.
206. K. Biemann, *Pract. Spectrosc.* **8** (1990) 3 – 24.
207. K. Biemann in J. A. McCloskey (ed.): "Mass Spectrometry," *Methods in Enzymology,* **vol. 193,** Academic Press, San Diego 1990, pp. 886 – 887.
208. S. A. Carr, V. N. Reinhold, *Biomed. Mass Spectrom.* **11** (1984) 633 – 642.
209. A. Dell, C. E. Ballou, *Carbohydr. Res.* **120** (1983) 95 – 111.
210. B. A. Andersson, W. W. Christie, R. T. Holman, *Lipids* **10** (1975) 215 – 219.
211. P. Sperling, M. Linscheid, S. Stöcker, H.-P. Mühlbach, E. Heinz, *J. Biol. Chem.* **268** (1993) 26935 –26940.
212. M. Suzuki, T. Ariga, M. Sekine, E. Araki, T. Miyatake, *Anal. Chem.* **53** (1981) 985 – 988.
213. N. J. Jensen, K. B. Tomer, M. L. Gross, *J. Am. Chem. Soc.* **107** (1985) 1863.
214. H. Schweer, H. W. Seyberth, C. O. Meese, O. Fürst, *Biomed. Environm. Mass Spectrom.* **15** (1988) 143 –151.
215. H.-R. Schulten in H. R. Morris (ed.): *Soft Ionization of Biological Substrates,* Heyden & Sons, London 1981.
216. M. Linscheid, G. Feistner, H. Budzikiewicz, *Isr. J. Chem.* **17** (1978) 163.
217. M. Linscheid, J. D'Angona, A. L. Burlingame, A. Dell, C. Ballou, *Proc. Natl. Acad. Sci. USA* **78** (1981) 1471 – 1475.
218. D. Suckau, M. Mak, M. Przybylski, *Proc. Natl. Acad. Sci. USA* **89** (1992) 5630 – 5634.
219. L. Weidolf, T. R. Covey, *Rapid Commun. Mass Spectrom.* **6** (1992) 192 – 196.
220. R. D. Smith, K. J. Light-Wahl, *Biol. Mass Spectrom.* **22** (1993) 493 – 501.
221. B. Ganem, Y.-T. Li, J. D. Henion, *J. Am. Chem. Soc.* **113** (1991) 6294 – 6296.
222. B. Ganem, Y.-T. Li, J. D. Henion, *Tetrahedron Lett.* **34** (1993) 1445 – 1448.
223. M. G. Ikonomou, A. Naghipur, J. W. Lown, P. Kebarle, *Biomed. Environm. Mass Spectrom.* **19** (1990) 434 – 446.
224. C. E. Costello, K. M. Comess, A. S. Plaziak, D. P. Bancroft, S. J. Lippard, *Int. J. Mass Spectrom. Ion Processes* **122** (1992) 255 – 279.
225. J. Yergey, D. Heller, G. Hansen, R. J. Cotter, C. Fenselau, *Anal. Chem.* **55** (1983) 353 – 356.
226. K. Biemann in J. A. McCloskey (ed.): "Mass Spectrometry," *Methods in Enzymology,* vol. 193, Academic Press, San Diego 1990, pp. 295 – 305.
227. J. Yao, M. Dey, S. J. Pastor, C. L. Wilkins, *Anal. Chem.* **67** (1995) 3638.
228. Y. Li, R. T. McIver, R. L. Hunter, *Anal. Chem.* **66** (1994) 2077.
229. C. C. Stacey et al., *Rapid Commun. Mass Spectrom.* **8** (1994) 513.
230. Y. Li, K. Tang, D. P. Little, H. Köster, R. L. Hunter, R. T. M. Jr, *Anal. Chem.* **68** (1996) 2090.
231. E. A. Stemmler, M. V. Buchanan, G. B. Hurst, R. L. Hettich, *Anal. Chem.* **67** (1995) 2924.
232. E. A. Stemmler, M. V. Buchanan, G. B. Hurst, R. L. Hettich, *Anal. Chem.* **66** (1994) 1274.
233. H. Yoshida, R. L. Buchanan, *Radiat. Res.* **139** (1994) 271.
234. R.-D. Wilken, R. Falter, *Appl. Organomet. Chem.* **12** (1998) 551.
235. L. Ebdon, A. Fischer, N. B. Roberts, M. Yaqoob, *Appl. Organomet. Chem.* **13** (1999) 183.
236. Y. Inoue et al., *Appl. Organomet. Chem.* **13** (1999) 81.
237. C. Siethoff, B. Pongraz, M. Linscheid, *Proc. 44th ASMS Conference on Mass Spectrometry and Allied Topics* (Portland, USA), (1996) 987.
238. I. Jardine in J. Villafranca (ed.): *Current Research in Protein Chemistry,* Academic Press, San Diego 1991.
239. N. Blaszkiewicz, G. Baumhoer, B. Neidhart, R. Ohlendorf, M. Linscheid, *J. Chrom.* **439** (1988) 109 – 119.
240. K. W. M. Siu, G. J. Gardner, S. S. Berman, *Rapid Commun. Mass Spectrom.* **2** (1988) 201 – 204.
241. Y. Morita, K. Takatera, T. Watanabe, *Seisan Kenkyu* **43** (1991) 367.
242. B. Michalke, P. Schramel, *Electrophoresis* **20** (1999) 2547.
243. C. Siethoff, I. Feldmann, N. Jakubowski, M. Linscheid, *J. Mass Spectrom.* **34** (1999) 421.
244. I. Tolosa, J. M. Bayona, J. Albaigés, L. F. Alencastro, J. Taradellas, *Fresenius J. Anal. Chem.* **339** (1991) 646 – 653.
245. N. Jakubowski, D. Stuewer, W. Vieth, *Anal. Chem.* **59** (1987) 1825.
246. N. Jakubowski, D. Stuewer, W. Vieth, *Mikrochim. Acta* **1** (1987) 302.
247. A. P. Mykytiuk, P. Semeniuk, S. Berman, *Spectrochim. Acta Rev.* **13** (1990) 1.
248. D. Stüwer in G. Holland, A. Eaton (eds.): *Application of Plasma Source Mass Spectrometry,* Royal Society of Chemistry, Cambridge 1990, pp. 1 ff.

249. F. L. King, J. Teng, R. E. Steiner, *J. Mass Spectrom.* **30** (1995) 1061.
250. J. R. d. Laeter, *Mass Spectrom. Rev.* **17** (1998) 97.
251. H. Oechsner, W. Gerhard, *Phys. Lett.* **40A** (1972) 211.
252. N. Jakubowski, D. Stuewer in D. Littlejohn, D. T. Burns (eds.): *Rev. Anal. Chem., Euroanalysis VIII,* Royal Society of Chemistry, (1994) 121 – 144.

Atomic Spectroscopy

José A. C. Broekaert, Universität Dortmund, Fachbereich Chemie, Dortmund, Federal Republic of Germany

E. Hywel Evans, University of Plymouth, Plymouth, United Kingdom

1. Introduction . . . 3
2. Basic Principles . . . 3
2.1. Atomic Structure . . . 3
2.2. Plasmas . . . 5
2.3. Emission and Absorption of Radiation . . . 6
2.4. Ionization . . . 9
2.5. Dissociation . . . 11
2.6. Sources and Atom Cells . . . 12
2.7. Analytical Atomic Spectrometry . . . 15
3. Spectrometric Instrumentation . . . 16
3.1. Figures of Merit of an Analytical Method . . . 16
3.2. Optical Spectrometers . . . 18
3.2.1. Optical Systems . . . 18
3.2.2. Detectors . . . 23
3.2.3. Nondispersive Spectrometers . . . 26
3.3. Mass Spectrometers . . . 28
3.3.1. Types of Mass Spectrometer . . . 28
3.3.2. Ion Detection . . . 30
3.3.3. Ion Extraction . . . 31
3.4. Data Acquisition and Processing . . . 32
4. Sample Introduction Devices . . . 34
4.1. Pneumatic Nebulization . . . 34
4.2. Ultrasonic Nebulization . . . 36
4.3. Hydride Generation . . . 38
4.4. Electrothermal Evaporation . . . 38
4.4.1. The Volatilization Process . . . 39
4.4.2. Types of Electrothermal Device . . . 39
4.4.3. Temperature Programming . . . 40
4.4.4. Analytical Performance . . . 41
4.5. Direct Solid Sampling . . . 41
4.5.1. Thermal Methods . . . 41
4.5.2. Slurry Atomization . . . 42
4.5.3. Arc and Spark Ablation . . . 43
4.5.4. Laser Ablation . . . 44
4.6. Cathodic Sputtering . . . 45
5. Atomic Absorption Spectrometry . . . 48
5.1. Principles . . . 48
5.2. Spectrometers . . . 49
5.2.1. Spectrometer Details . . . 49
5.2.2. Primary Sources . . . 50
5.3. Flame Atomic Absorption . . . 51
5.3.1. Flames and Burners . . . 51
5.3.2. Nebulizers . . . 52
5.3.3. Figures of Merit . . . 52
5.4. Electrothermal Atomic Absorption . . . 53
5.4.1. Atomizers . . . 54
5.4.2. Thermochemistry . . . 55
5.4.3. Figures of Merit . . . 55
5.5. Special Techniques . . . 56
5.5.1. Hydride and Cold Vapor Techniques . . . 56
5.5.2. Direct Solid Sampling . . . 57
5.5.3. Indirect Determinations . . . 58
5.6. Background Correction . . . 58
5.6.1. Deuterium Lamp Technique . . . 58
5.6.2. Zeeman Effect Technique . . . 59
5.6.3. Smith – Hieftje Technique . . . 60
5.6.4. Coherent Forward Scattering . . . 60
5.7. Fields of Application . . . 62
6. Atomic Emission Spectrometry . . . 63
6.1. Principles . . . 63
6.2. Spectrometers . . . 66
6.3. Flame Emission . . . 67
6.4. Arcs and Sparks . . . 67
6.5. Plasma Sources . . . 70
6.5.1. Direct Current Plasmas . . . 70
6.5.2. Inductively Coupled Plasmas . . . 71
6.5.3. Microwave Plasmas . . . 75
6.6. Glow Discharges . . . 76
6.6.1. Hollow Cathodes . . . 77
6.6.2. Furnace Emission Spectrometry . . . 77
6.6.3. Glow Discharges with Flat Cathodes . . . 77
6.6.4. New Developments . . . 79
6.7. Laser Sources . . . 79
7. Plasma Mass Spectrometry . . . 80
7.1. Principles . . . 80
7.1.1. Instrumentation . . . 80
7.1.2. Analytical Features . . . 81
7.1.3. Applications . . . 87
7.2. Glow Discharge Mass Spectrometry . . . 87
7.2.1. Instrumentation . . . 88
7.2.2. Analytical Performance . . . 88
7.2.3. Applications . . . 89
8. Atomic Fluorescence Spectrometry . . . 90
8.1. Principles . . . 90
8.2. Instrumentation . . . 91
8.3. Analytical Performance . . . 92
9. Laser-Enhanced Ionization Spectrometry . . . 93
10. Comparison With Other Methods . . . 96
10.1. Power of Detection . . . 96

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_559

10.2. Analytical Accuracy 97
10.3. Cost 97
11. References 98

Symbols :

A absorbance, $A = \log I_0/I$ with I_0 the intensity of the incident beam and I the intensity of the sorting beam
A_{qp} Einstein transition probability for spontaneous emission of radiation corresponding with a transition from the higher level q to the lower level p (s^{-1})
B_{pq} Einstein transition probability for absorption corresponding with a transition from the lower level p to the higher level q (s^{-1})
B_{qp} Einstein transition probability for stimulated emission corresponding with a transition from the higher level q to the lower level p (s^{-1})
c velocity of light in vacuum (3×10^{10} m/s)
c_L detection limit in concentration units or in terms of absolute amounts of analyte
c_U background equivalent concentration, calibration constant in optical atomic spectrometry for trace analysis
C capacitance (Farad)
d_0 Sauter diameter; droplet diameter for an aerosol at which the volume-to-surface ratio of the particles equals the volume-to-surface ratio for the whole aerosol
D_{hkl} lattice constant of crystal
f_k focal length of collimator (m)
g_q statistical weight of an atomic (molecular) energy level q
G optical conductance
h isotopic abundance in %
j total internal quantum number
k Boltzmann's constant (1.38×10^{-23} J/K)
k_{12} velocity constant for collisional excitation (s^{-1})
k_{21} velocity constant for collisional deexcitation (s^{-1})
l orbital quantum number
L inductance (Henry)
L Lorentz unit ($e/4\pi m \cdot c^2$)
m_1 magnetic quantum number
M magnetic quantum number
n principle quantum number
$P(\nu)$ profile function of a line
q sputtering rate (µg/s)
QE quantum efficiency of a photocathode
R Rydberg constant (109.677 cm^{-1})
R_0 theoretical resolving power of optical spectrometer
RSF relative sensitivity factor in elemental mass spectrometry
s spin quantum number
s_a exit slit width (µm)
s_e entrance slit width (µm)
s_{eff} effective slit width of an optical spectrometer in µm; $s_{eff}^2 = s_e^2 + s_L^2 + s_0^2 + s_z^2$ with s_L and s_z the contributions of the physical width of the line measured and the optical aberrations, respectively (all in µm)
s_0 diffraction slit width (µm)
$s_r(c_x)$ relative standard deviation of the estimate
S blackening of a photographic emulsion
S sputtering yield in sputtered ions/incident ion
$S_{nj}(T)$ Saha function for neutrals n of an element j
S/N signal-to-noise ratio
t_r response time of a decay process for an excited level (s)
t_p pulse duration (s)
U potential (V)
ν' wavenumber (cm^{-1})
$\Delta\nu$ frequency broadening contribution (cm^{-1})
w penetration rate (µm/s)
Z charge of nucleus or ion (in elementary charge units)
Z_{ij} partition function of an element j in the ith ionization state
α_D divergency of laser beam with $\alpha_D = \lambda/D$ with λ the wavelength and D the beam diameter
γ contrast of photographic emulsion
$\Delta\lambda_L$ physical halfwidth of a spectral line (nm or pm)
$\theta(V)$ amplification of photomultiplier
$\lambda/\Delta\lambda$ practical resolving power
ϱ_ν radiant density of a spectral feature with frequency ν
$\sigma(v)$ collision cross section for species with velocity v

φ radiant flux $\varphi = \tau B G$ with τ the transmittance and B the radiant density of the source in $W\,m^{-2}$ $sterad^{-1}$

1. Introduction

Atomic spectroscopy is the oldest principle of instrumental elemental analysis, going back to the work of BUNSEN and KIRCHHOFF in the mid-nineteenth century [1]. The optical radiation emitted from flames is characteristic of the elements present in the flame itself or introduced by various means. This discovery depended on the availability of dispersing media, such as prisms, which spectrally resolved the radiation to produce the line spectra of the elements. It was also found that radiation with the same wavelength as the emitted lines is absorbed by a cold vapor of the same element. This observation is the basis for atomic emission and absorption spectrometry.

Flames proved to be suitable sources for the analysis of liquids, but the need for direct chemical analysis of solids soon arose, including the need for rapid analysis of more than one element in a solid sample. The development of electrical discharges fostered the growth of atomic spectrometry; arc and spark sources proved useful for analyte ablation and excitation. Low-pressure discharges proved to be powerful radiation sources (a similar development has been observed more recently, with the availability of laser sources).

The limitations stemming from the restricted temperatures of flames led to the development of high-temperature plasma sources for atomic emission spectrometry. The development of high-frequency inductively coupled plasmas and microwave plasmas has led to the widespread use of these methods for routine analytical work.

Parallel with these developments, there have been advances in spectrometer design, from the spectroscope, through photographic spectrographs, sequential spectrometers, and simultaneous multichannel spectrometers with solid-state detectors.

The use of flames as atomizers for atomic absorption spectrometry (AAS) has also made its impact on analytical methodology, largely as the result of work by WALSH [2]. L'VOV and MASSMANN [3], [4] introduced the graphite furnace atomizer, now widely available in commercially available AAS with electrothermal atomization, especially for ultratrace analysis.

The sources used for optical atomic spectrometry are also powerful sources for elemental mass spectrometry, from the classical spark source mass spectrometer to present-day plasma mass spectrometric methods such as glow discharge and inductively coupled plasma mass spectrometry.

2. Basic Principles

2.1. Atomic Structure

The basic processes in atomic spectrometry involve the outer electrons of atoms. Accordingly, the possibilities and limitations of atomic spectroscopy depend on the theory of atomic structure. On the other hand, the availability of optical spectra was decisive in the development of the theory of atomic structure and even for the discovery of a number of elements.

In 1885, BALMER reported that for a series of atomic emission lines of hydrogen, a simple relationship existed between the wavelengths:

$$\lambda = k\frac{n^2}{n^2-4} \tag{1}$$

where $n = 2, 3, 4, \ldots$ for the lines $H_\alpha, H_\beta, H_\gamma, \ldots$ Equation (1) can be written in wavenumbers:

$$\nu' = \frac{1}{\lambda} = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right) \tag{2}$$

where ν' is the wavenumber (cm^{-1}) and R the Rydberg constant ($109.677\,cm^{-1}$). The wavenumbers of all series in the hydrogen spectrum are given by:

$$\nu' = R\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \tag{3}$$

where n_2 is a series of numbers $> n_1$, with $n_1 = 1, 2, 3, 4, \ldots$ for the Lyman, Balmer, Paschen, and Brackett series, respectively. Rydberg extended the Balmer formula to:

$$\nu' = RZ^2\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \tag{4}$$

where Z is the effective charge of the atomic nucleus, allowing calculation of wavelengths for

elements other than hydrogen. The wavenumber of each line in an atomic spectrum can thus be calculated from a formula involving two positive numbers, called spectral terms. The significance of the spectral terms emerges from Bohr's theory of atomic structure, in which the atom has a number of discrete energy levels related to the orbits of the electrons around the nucleus. As long as an electron is in a defined orbit no energy is emitted, but when a change in orbit occurs, another energy level is reached and the excess energy is emitted in the form of electromagnetic radiation. The wavelength is given by Planck's law:

$$E = h\nu = hc/\lambda \tag{5}$$

Here $h = 6.626 \times 10^{-34}$ J · s, ν is the frequency (s^{-1}), $c = 3 \times 10^{10}$ cm/s is the velocity of light, and λ is the wavelength (cm). Accordingly:

$$\nu' = \frac{1}{\lambda} = \frac{E}{hc} = \frac{E_1}{hc} - \frac{E_2}{hc} = T_1 - T_2 \tag{6}$$

where T_1 and T_2 are the Bohr energy levels, the complexity of emission spectra reflects the complex structure of atomic energy levels. For an atom with nuclear charge Z and one valence electron, the energy of this electron is given by:

$$E = -\frac{2\pi Z^2 e^4 \mu}{n^2 h^2} \tag{7}$$

where $\mu = mM/(m+M)$, with m the mass of the electron and M the mass of the nucleus; n is the principal quantum number ($n = 1, 2, 3, \ldots$) on which the order of the energy levels depends. For the orbital angular momentum $\boldsymbol{L}$:

$$|\boldsymbol{L}| = \frac{h}{2\pi}\sqrt{l\,(l+1)} \tag{8}$$

l is the orbital quantum number and has values of: $0, 1, 2, \ldots, (n-1)$; $l = 0$ for a circular orbit, $l = 1, 2, \ldots$ for elliptical orbits. The possible orientations of the elliptical orbits with respect to an external electric or magnetic field:

$$\boldsymbol{L}_z = \frac{h}{2\pi} m_1 \tag{9}$$

where $\boldsymbol{L}_z$ is the component of the orbital angular momentum along the field axis and $m_1 = \pm 1, \pm(l-1), \ldots, 0$ is the magnetic quantum number; for each value of l it has $(2l+1)$ values. When a spectral line source is subject to a magnetic field, the spectral lines display hyperfine structure (Zeeman effect). In order to explain hyperfine structure it is postulated that the electron rotates on its axis with spin angular momentum $\boldsymbol{S}$:

$$|\boldsymbol{S}| = \frac{h}{2\pi}\sqrt{S\,(S+1)} \tag{10}$$

The spin quantum number m_s determines the angles between the axis of rotation and the external field as:

$$s_z = \frac{h}{2\pi} m_s \tag{11}$$

where $m_2 = \pm 1/2$. The orbital and spin angular momenta determine the total angular momentum of the electron $\boldsymbol{J}$:

$$\vec{\boldsymbol{J}} = \vec{\mathrm{L}} + \vec{\mathrm{S}} \quad \text{with } |\boldsymbol{J}| = \frac{h}{2\pi}\sqrt{j\,(j+1)} \tag{12}$$

where $j = l \pm s$ is the total internal quantum number. Atomic spectral terms differ in their electron energies and can be characterized by the quantum numbers through the term symbols:

$$n^m l_j \tag{13}$$

Here $l = 0, 1, 2, \ldots$ and the corresponding terms are given the symbols s (sharp), p (principal), d (diffuse), f (fundamental), etc., relating originally to the nature of different types of spectral lines; n is the principal quantum number, m is the multiplicity ($m = 2s \pm 1$), and j is the total internal quantum number. The energy levels of each element can be given in a term scheme, in which is also indicated which transitions between energy levels are allowed and which forbidden. This is reflected by the so-called selection rules; only those transitions are allowed for which Δn has integral values and at the same time $\Delta l = \pm 1$, $\Delta j = 0$ or ± 1, and $\Delta s = 0$. The terms for an atom with one outer (valence) electron can easily be found, e.g., for Na ($1s^2 2s^2 2p^6 3s^1$), in the ground state: $3^2S_{1/2}$ [$l = 0\,(s)$, $m = 2\,(1/2) + 1 = 2\,(s = 1/2)$, and $j = 1/2$, ($j = — l \pm s —$)]. When the $3s$ electron is promoted to the $3p$ level, the term symbol for the excited state is: $3^2P_{1/2,3/2}$ ($l = 1\,(p)$, $m = 2\,(1/2) + 1 = 2$, and $j = 1/2, 3/2$). The term schemes of the elements are well documented in the work of Grotrian [5]. The term scheme for Na is shown in Figure 1.

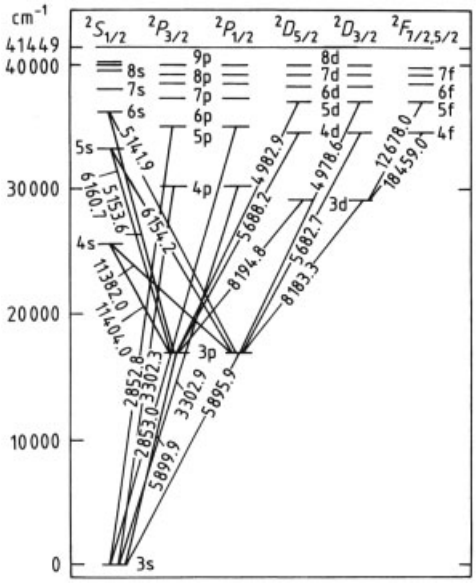

Figure 1. Atomic energy level diagram for the sodium atom [5]

For atoms with more than one valence electron, Russell – Saunders ($L-S$) coupling applies. The orbital momenta of all electrons have to be coupled to the total orbital momentum, like the spin momentum. The fact that each electron has a unique set of quantum numbers is known as the Pauli exclusion principle. The total quantum number L is obtained as $L=\sum l$, $S=\sum s$, and $J=L-S,\ldots,L+S$. The term symbol accordingly becomes:

$$^{M}L_{J} \tag{14}$$

For the case Mg ($1s^2 2s^2 2p^6 3s^2$), the ground state is 3^1S_0 ($L=0$ as $l_1=0$ and $l_2=0$, $S=0$ as $s_1=1/2$ and $s_2=-1/2$, and $J=0$ as both L and $S=0$). The first excited state ($1s^2 2s^2 2p^6 3s3p$) is characterized by the terms: 3^1P_1 ($L=1$ as $l_1=0$ and $l_2=1$, $S=0$ as $s_1=1/2$ and $s_2=-1/2$, and $J=-L\pm 5-=1$), but also 3^3P_2, 3^3P_1, and 3^3P_0 (as for the parallel spins $s_1=1/2$ and $s_2=1/2$, $S=1$, and further $J=0$, 1, 2). As the number of electrons increases, the coupling becomes more complex, increasing the number of spectral terms and thus the number of lines in the spectrum.

2.2. Plasmas [6]

Partially ionized gases are usually referred to as plasmas. They contain molecules, radicals, atoms, ions, and free electrons, and result from the interaction of energy with matter in the gaseous state. Just as for atoms, the radicals, molecules, and ions in a plasma can be in their ground or excited states and radiation can be emitted or absorbed when transitions occur from one state to another. The wavelengths of spectral lines can be obtained from Planck's law, but their intensities depend on the number densities of the species and states involved.

Energy transfer in a plasma results from nonradiative processes (involving collisions) as well as from radiative processes (emission, absorption, and fluorescence). The efficiency of collision processes is described by the cross section $\sigma(v)$. It reflects the change in momentum of a particle with mass m and velocity v when it collides with a particle with mass M. Apart from the cross section, the velocity distribution for a given species affects energy transfer in a plasma.

For a Maxwell distribution, the velocity distribution is given by:

$$\frac{\mathrm{d}n}{n}=\frac{2}{\sqrt{\pi}}\sqrt{u'}\exp\left(-u'\right)\mathrm{d}u' \tag{15}$$

For the so-called Druyvenstein distribution which pertains at low pressure:

$$\frac{\mathrm{d}n}{n}=1.039\sqrt{u'}\exp\left(-0.548u'^2\right)\mathrm{d}u' \tag{16}$$

where $u'=E/kT$, E is the mean energy of the particles, and T the absolute temperature. For a plasma contained in a closed system and in thermal equilibrium, the population of the excited levels for all particles is given by Boltzmann's law:

$$\frac{n_q}{n_0}=\frac{g_q}{g_0}\exp\left(-E_q/kT\right) \tag{17}$$

where n_q is the number density of particles in the excited state, n_0 is the number density of particles in the ground state, g_q and g_0 are the statistical weights of the corresponding levels, E_q is the excitation energy of the state q, k is Boltzmann's constant (1.38×10^{-23} J/K), and T is the absolute temperature. Equation (17) can be transformed when n_q is expressed as a function of the total number density of the species. Then:

$$\frac{n_q}{n}=\frac{g_q \exp\left(-E_q/kT\right)}{\sum_m g_m \exp\left(-E_m/kT\right)} \tag{18}$$

as $n=\sum_m n_m$. The sum $Z_m=\sum_m g_m \exp\left(-E_m/kT\right)$ is the partition function. It is a function of temperature and its coefficients for a large number of neutral and ionized species are listed in the literature, e.g., [7]. When E_q is expressed in electron volts, Equation (18) can be written:

$$\log n_{aq}=\log n_a+\log n_q-\left(5040\right)/TV_q-\log Z \tag{19}$$

2.3. Emission and Absorption of Radiation [8]

In a steady-state plasma, the number of particles leaving an energy level per unit time equals the number returning to this level. The most important energy exchange processes in a plasma are:

1) Collisions where atoms are excited to a higher level by collision with energetic neutrals (collisions of the first kind)
2) Collisions where excited neutrals lose energy through collisions without emission of radiation (collisions of the second kind)
3) Excitation by collision with electrons
4) Deexcitation where energy is transferred to electrons
5) Excitation of atoms by the absorption of radiation
6) De-excitation of atoms by spontaneous or stimulated emission

When n is the number density of one type of particle and N is that of a second species, present in excess ($n \ll N$), the following equilibria are to be considered:

$$\alpha N n_0=\beta N n_q \tag{20}$$

$$\alpha_e n_e n_0=\beta_e n_e n_q \tag{21}$$

$$B' \varrho_\nu n_0=\left(A+B\varrho_\nu\right) n_q \tag{22}$$

where n_e is the electron number density; A, B, and B' are the Einstein transition probabilities for spontaneous emission, stimulated emission, and absorption, respectively; and α_e, α, β_e, and β are functions of the cross sections for the respective processes, as well as of the velocity distribution of the particles involved. The radiation density at frequency ν is ϱ_ν. When the system is in thermodynamic equilibrium, the neutrals and electrons have the same Maxwell velocity distribution and at temperature T:

$$\begin{aligned} n_q/n_0&=\alpha/\beta=\alpha_e/\beta_e=B'/\left(A/\varrho_\nu+B\right)\\ &=\frac{g_q}{g_0}\exp\left(-E_q/kT\right) \end{aligned} \tag{23}$$

Thus, each process is in equilibrium with its inverse and the Boltzmann distribution of each state is maintained by collisions of the first and second kind, including collisions with electrons.

In a real radiation source, emission and absorption of radiation have to be considered as well, but contribute little to the energy balance. The system is in so-called local thermal equilibrium (LTE) and:

$$\begin{aligned} \alpha N n_0+\alpha_e n_e n_0+B' \varrho_\nu n_0&=\beta N n_q+\beta_e n_e n_q\\ &+\left(A+B\varrho_\nu\right) n_q \end{aligned} \tag{24}$$

from which n_q/n_0 can be calculated. The population of excited states is determined by the excitation processes in the radiation source, as reflected by the coefficients in Equation (24). For a d.c. arc, $\alpha N \gg \alpha_e n_e + B' \varrho_\nu$ and $\beta N \gg \beta_e n_e + (A + B \varrho_\nu)$. This leads to:

$$\frac{n_q}{n_0}=\frac{\alpha}{\beta}=\frac{g_q}{g_0}\exp\left(-E_q/kT\right) \tag{25}$$

As the radiation density is low, the d.c. arc plasma can be considered to be in thermal equilibrium. The simplification which leads to Equation (24) does not apply for low-pressure discharges, where collisions with electrons are very important as are radiative processes; moreover, the velocity distributions are described by the Druyvenstein equation. These sources are not in thermal equilibrium.

Excited states are prone to decay because of their high energy. This can take place by collisions with molecules, atoms, electrons, or ions, or by emission of electromagnetic radiation. In the latter case, the wavelength is given by Planck's law. For levels q and p, the number of spontaneous transitions per unit time is given by:

$$-\mathrm{d}N_q/\mathrm{d}t = A_{qp}N_q \tag{26}$$

where A_{qp} is the Einstein coefficient for spontaneous emission (s^{-1}). When several transitions can start from level q, Equation (26) becomes:

$$-\mathrm{d}N_q/\mathrm{d}t = N_q \sum_p A_{qp} = N_q \nu_q \tag{27}$$

where ν_q is the lifetime of excited state q. If decay can take place by an allowed radiative transition, the lifetime is of the order of 10^{-8} s. When no radiative transitions are allowed levels are metastable (e.g., Ar 11.5 and 11.7 eV), which can decay only by collisions. In low-pressure discharges such levels may have very long lifetimes (up to 10^{-1} s). For absorption of electromagnetic radiation with frequency ν_{qp} and radiation density ϱ_ν, the number density of N_q increases as:

$$\mathrm{d}N_q/\mathrm{d}t = B_{qp}N_q\nu_\varrho \tag{28}$$

For stimulated radiation, atoms in excited state q decay only when they interact with radiation of wavelength λ_{qp} and:

$$-\mathrm{d}N_q/\mathrm{d}t = B_{qp}N_q\varrho_\nu \tag{29}$$

For thermal equilibrium:

$$g_q B_{qp} = g_p B_{pq} \tag{30}$$

where g_p and g_q are the degeneracies of the respective levels. The intensity I_{qp} of an emitted spectral line (a) is proportional to the number density of atoms in state q:

$$I_{qp} = A_{qp} n_{aq} h\nu_{qp} \tag{31}$$

or after substitution of n_{aq} (n_q for the atomic species according to Eq. 18):

$$I_{qp} = A_{qp} h\nu_{qp} n_a \left(g_q/Z_a\right) \exp\left(-E_q/kT\right) \tag{32}$$

where T is the excitation temperature which can be determined from the intensity ratio for two lines (a and b) of the same ionization state of an element as:

$$T = \left[5040\left(V_a - V_b\right)\right] / \left\{\log\left[(gA)_a / (gA)_b\right] - \log\left(\lambda_a/\lambda_b\right) - \log\left(I_a/I_b\right)\right\} \tag{33}$$

Often the line pair Zn 307.206/Zn 307.59 nm is used. This pair is very suitable because ionization of zinc is low as a result of its relatively high ionization energy, the wavelengths are close to each other, which minimizes errors introduced by changes in the spectral response of the detector, and the ratio of the gA values is well known. The excitation temperatures can also be determined from the slope of the plot $\ln\left[I_{qp}/(g_q A_{qp}\,\nu_{qp})\right]$ or $\ln\left(I_{qp}\,\lambda/gA_{qp}\right)$ against E_q, which is $-1/kT$. The λ/gA values for a large number of elements and lines are available [8]. Spectroscopic measurement of temperatures from line intensities may be hindered by deviations from ideal thermodynamic behavior in real radiation sources, and by inaccuracies of transition probability estimates. Determination of excitation temperatures in spatially inhomogeneous plasmas is treated extensively by Boumans [9].

According to classical dispersion theory, the relation between absorption and the number density of the absorbing atoms is given by:

$$\int K_\nu \,\mathrm{d}\nu = \frac{\pi \mathrm{e}^2}{mc} N_\nu f \tag{34}$$

where K_ν is the absorption coefficient at frequency ν, m is the mass and e the charge of the electron, c is the velocity of light, N_ν is the density of atoms with frequency between ν and $\nu + \mathrm{d}\nu$ (practically the same as N), and f is the oscillator strength. This relation applies strictly to monochromatic radiation, so use of a primary source which emits very narrow lines is advantageous. The relation between absorption A and the concentration of the absorbing atoms in an atomizer is given by the Lambert–Beer law. If I_0 is the intensity of the incident radiation, l the absorption path length, and I the intensity of the transmitted radiation, the change in intensity dI resulting from absorption within the absorption path length dl is:

$$-dI = kI_0 c\mathrm{d}I \tag{35}$$

or

$$\int_{t_0}^{t} -\mathrm{d}t/t = kc \int_0^l \mathrm{d}l \tag{36}$$

$A = \log\left(I_0/I\right) = \log\left(1/T\right)$, in which A is the absorbance and T the transmittance. Accordingly, Equation (36) becomes:

$$A = kcl \tag{37}$$

The absorbances are additive. The Lambert–Beer law, however, is valid only within a restricted concentration range; in atomic absorption spectrometry deviations from linearity are common.

Line Broadening. Atomic spectral lines have a physical width resulting from several broadening mechanisms [10]. The *natural width* of a spectral line results from the finite lifetime of an excited state, τ. The corresponding half-width in terms of frequency is:

$$\Delta\nu_{\mathrm{N}} = 1/(2\pi\tau) \tag{38}$$

This results for most spectral lines in a halfwidth of the order of 10^{-2} pm. The *Doppler spectral width* results from the velocity component of the emitting atoms in the observation direction. The corresponding half-width is:

$$\Delta\nu_{\mathrm{D}} = \left[2\sqrt{(\ln 2)}/c\right]\nu_0\sqrt{[2(RT)/M]} \tag{39}$$

where c is the velocity of light, ν_0 is the frequency of the line maximum, R is the gas constant, and M the atomic mass. Doppler broadening thus depends strongly on temperature. Accordingly, it is often referred to as temperature broadening and reflects the kinetic energy of the radiating atoms, ions, or molecules. For the Ca 422.6 nm line in the case of a hollow cathode discharge at pressures of the order of 0.1 kPa, the Doppler broadening at 300 K is 0.8 pm whereas at 2000 K it is 2 pm [11].

Lorentz or pressure broadening results from the interaction between the emitting atoms and atoms of other elements. The half-width is:

$$\Delta\nu_{\mathrm{L}} = (2/\pi)\,\sigma_{\mathrm{L}}^2 N\sqrt{\left[2\pi RT\left(\frac{1}{M_1}+\frac{1}{M_2}\right)\right]} \tag{40}$$

where M_1 and M_2 are the atomic masses, N is the concentration of the foreign atoms, and σ_{L} is their cross section. Pressure broadening is low in low-pressure discharges. For the Ca 422.6 nm line, this type of broadening at 300 K and 0.9 kPa is only 0.02 pm. At atmospheric pressure, however, this type of line broadening is predominant.

Isotopic and *hyperfine structure, resonance broadening* (resulting from interaction between radiating and nonradiating atoms of the same species), and *Stark broadening* (resulting from interaction with electric fields), contribute to the widths of spectral lines.

Natural and Lorentz broadening have a Lorentzian distribution:

$$I(\nu) = I_0/\left\{1 + [2(\nu-\nu_0)/\Delta\nu]^2\right\} \tag{41}$$

Doppler broadening has a Gaussian distribution:

$$I(\nu) = I_0 \exp - \left\{[2(\nu-\nu_0)/2\Delta\nu_{\mathrm{D}}]\sqrt{(\ln 2)}\right\} \tag{42}$$

The combination of both types of profile (normally both pressure and temperature broadening are important) results in a so-called Voigt profile:

$$V(\alpha,\omega) = \alpha/\pi \cdot \int_{-\infty}^{+\infty}\left\{[\exp(-y^2\,\mathrm{d}y)]/\left[\alpha^2 + (\omega-y)^2\right]\right\} \tag{43}$$

where $\omega = 2(\nu-\nu_0)/\Delta\nu_{\mathrm{D}}\sqrt{(\ln 2)}$ and $\alpha = (\Delta\nu_{\mathrm{L}}/\Delta\nu_{\mathrm{D}})\sqrt{(\ln 2)}$, when the contribution of the natural width is neglected. From the widths of spectral lines, in most cases between 1 and 20 pm, and knowledge of the distribution functions, the contributions of the different broadening processes can be calculated by deconvolution methods. This allows calculation of the so-called gas temperature from the Doppler broadening component, giving an estimate of the kinetic energy of neutrals and ions. The profiles of spectral lines from plasma sources at atmospheric pressure are illustrated by the high-resolution spectra of a number of rare-earth elements in an inductively coupled plasma (ICP) (Fig. 2) [12].

Self-Absorption. The radiation emitted by the source is absorbed by ground state atoms of the same species. This phenomenon is known as self-absorption [13]. As the chance that an absorbed photon is re-emitted is less than unity, the observed radiation is reduced. When I_0 is the intensity emitted at the line maximum and $P_{\mathrm{E}}(\nu)$ is the profile function, the intensity distribution for an emission line is $I_0 P_{\mathrm{E}}(\nu)$ and the intensity observed after the radiation passes through a layer with number density of absorbing atoms of n_{A} is:

$$I(\nu) = I_0 P_{\mathrm{E}}(\nu)\exp\left[-\varrho P_{\mathrm{A}}(\nu)/P_{\mathrm{A}}(\nu_0)\right] \tag{44}$$

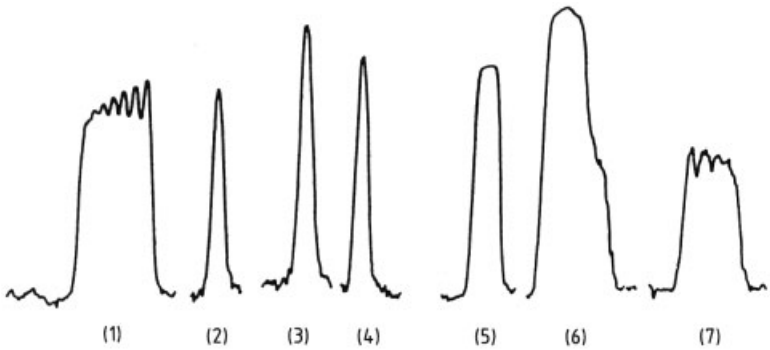

Figure 2. Profiles of some rare earth atomic emission lines in inductively coupled plasma atomic emission spectrometry (photographic measurements obtained with a high-resolution grating spectrograph; theoretical resolving power 460 000 [12]

where ν_0 is the frequency at the line center, $P_{\mathrm{A}}(\nu)$ is the absorption profile function, $P_{\mathrm{A}}(\nu_0)$ is its value at the line center, and ϱ is an absorption parameter:

$$\varrho \sim BP_{\mathrm{A}}(\nu)\, n_{\mathrm{A}} \tag{45}$$

ϱ increases with the transition probability for absorption and thus is larger for resonance lines which involve transitions to the ground state. It is also larger in sources with high analyte number density n_{A}. As the absorption maximum is in the center of the line, self-absorption always leads to flatter line profiles. When a minimum occurs in the absorption profile, the line undergoes so-called self-reversal and $\varrho > 1$. Self-reversal only occurs when there is a strong temperature gradient in the radiation source and when the analyte number densities in both the hot and the cooler zones of the plasma are high.

2.4. Ionization

If sufficient energy is transferred to a plasma, atoms may be ionized. This depends on the temperature of the plasma, but also on the ionization energy of the elements. As ions have discrete energy levels between which transitions are possible, ionic spectra are also important when considering the emission of radiation by a plasma. The ionization of atoms a of the element *j* into ions i is an equilibrium:

$$n_{\mathrm{a}j} \leftrightharpoons n_{\mathrm{i}j} + n_{\mathrm{e}} \tag{46}$$

and the equilibrium constant $S_{nj}(T)$, known as the Saha constant, is:

$$S_{nj}(T) = \left(n_{\mathrm{i}j} n_{\mathrm{e}}\right) / n_{\mathrm{a}j} \tag{47}$$

The degree of ionization α_j for element *j* is:

$$\alpha_j = n_{\mathrm{i}j}/n_j = n_{\mathrm{i}j} / \left(n_{\mathrm{a}j} + n_{\mathrm{i}j}\right) \tag{48}$$

where $n_{\mathrm{a}j}$ and $n_{\mathrm{i}j}$ are the concentrations of atoms and ions and can be expressed as a function of the total number of atoms n_j by:

$$n_{\mathrm{a}j} = (1-\alpha_j)\, n_j$$

and

$$n_{\mathrm{i}j} = \alpha_j n_j \tag{49}$$

In the notation used by Boumans [9], the intensity of an atom line can be written:

$$I_{qp} = A_{qp} h\nu_{qp} \left(g_q/Z_{\mathrm{a}j}\right)(1-\alpha_j)\, n_j \exp\left(-E_q/kT\right) \tag{50}$$

and the intensity of an ion line is:

$$I^+_{qp} = A^+_{qp} h\nu^+_{qp} \left(g^+_q/Z_{\mathrm{i}j}\right) \alpha_j n_j \exp\left(-E^+_q/kT\right) \tag{51}$$

These expressions for the intensities contain three factors which depend on temperature; i.e., the degree of ionization, the Boltzmann factor, and the partition functions. The degree of ionization can be written as a function of the electron number density and the Saha function:

$$\left[\alpha_j / (1-\alpha_j)\right] n_{\mathrm{e}} = S_{nj}(T) \tag{52}$$

However, the latter is also given by the Saha equation. When the Saha function is expressed in partial pressures, the Saha equation is:

$$S_{pj}(T) = [p_{ij}p_e] / [p_{aj}] = \frac{\left[(2\pi m)^{3/2}\ (kT)^{5/2}\right]}{h^3\ (2Z_{ij}/Z_{aj})\ [\exp(-E_{ij}/kT)]} \quad (53)$$

The factor 2 is the statistical weight of the free electron (two spin orientations), $k = 1.38\times10^{-23}$ J/K, the mass of the electron $m = 9.11\times10^{-28}$ g, $h = 6.62\times10^{-34}$ J · s, and 1 eV = 1.6×10^{-19} J. This leads to the following expression for the Saha equation:

$$\log S_{pj} = \frac{5}{2}\log\frac{T-5040}{T}V_{ij} + \log\frac{Z_{ij}}{Z_{aj}} - 6.18 \quad (54)$$

where V_{ij} is the ionization energy (in eV). The Saha equation is valid only for a plasma in local thermal equilibrium; then the temperature in the equation is the ionization temperature. When this condition is not fulfilled, the equilibrium between the different states of ionization is given by the so-called Corona equation [14].

Accordingly, the degree of ionization in a plasma can be determined from the intensity relation of atom and ion lines of the same element:

$$\log[(\alpha_j)/(1-\alpha_j)] = \log\left(I^+_{qp}/I_{qp}\right) - \log\left[\left(g^+_q A^+_{qp}\nu^+_{qp}\right)/(g_q A_{qp}\nu_{qp})\right] + (5040/T)\left(V^+ - V_q\right) + \log\left(Z_{ij}/Z_{aj}\right) \quad (55)$$

This method can only be applied for a plasma in local thermal equilibrium at known temperature. The partition functions Z_{aj} and Z_{ij} for the atom and ion species are functions of the temperature, and their coefficients have been calculated for many elements [7]. The accuracy of the gA values and of the temperatures affects the accuracy of the degree of ionization. The line pairs Mg II 279.6 nm/Mg I 278.0 nm, and Mg II 279.6 nm/Mg I 285.2 nm are often used for determinations of the degree of ionization of an element in a plasma.

Once α_j is known, the electron pressure can be determined. From

$$\log[\alpha_j/(1-\alpha_j)] = \log[S_{pj}(T)/p_e] \quad (56)$$

one can calculate:

$$\log p_e = -\log[\alpha_j/(1-\alpha_j)] + \log S_{pj}(T) \quad (57)$$

Taking into account Equations (54) and (55), this results in:

$$\log p_e = -\log\left(I^+_{qp}/I_{qp}\right) + \log\left(g^+_q A^+_{qp}\nu^+_{qp}/g_q A_{qp}\nu_{qp}\right) - (5040/T)\left(V_{ij} + V^+_q - V_q\right) + 5/2\,(\log T) - 6.18 \quad (58)$$

This reflects the fact that the intensity ratio of the atom and ion lines of an element changes considerably with the electron pressure in the plasma, particularly for elements with low ionization energy, such as Na. This is analytically very important as it is the cause of so-called ionization interferences in classical d.c. arc emission spectrometry, atomic absorption, and plasma optical emission, as well as mass spectrometry.

When the plasma is not in LTE, the electron number densities cannot be determined from the Saha equation, but have to be derived directly from the Stark broadening of the H_β line or of a suitable argon line. This contribution to broadening can be written [15]:

$$\delta\lambda = 2\,[1 + 1.75\alpha\,(1 - 0.75\varrho)]\,\omega \quad (59)$$

where ϱ is the ratio of the distance between the ions and the Debye path length, ω is the broadening due to the interaction of the electrons ($\sim n_e$), and α is the contribution of the interaction with the quasi-static ions ($\sim n_e^{1/4}$). The value of $\delta\lambda$ can be calculated as a function of n_e. The electron number density can be determined directly from the widths of the Ar I 549.59 or Ar II 565.07 nm lines, which are mainly due to Stark broadening. The temperature can also be determined by combining measurements of the intensities of atom and ion lines of the same element:

$$\log n_i/n_a = -\log n_e + 3/2\,(\log T) - (5040/T)\,V_{ij} + \log\left(Z_{ij}/Z_{aj}\right) + 15.684 \quad (60)$$

which can be combined with Equation (58). From Equations (50) and (51), which give the intensity of a line, and from the Saha equation (54), it can be seen that, for each spectral line emitted by a plasma source, there is a temperature where its emission intensity is maximum. This is the so-called standard temperature. To a first approximation [16], it can be written:

$$T = \left(0.95V_{ij}\times10^3\right)/\left[1 - 0.33\left(V_{ij}/10\right) + 0.37\log\left(V_{ij}/10\right) - 0.14\log P^*_e\right] \quad (61)$$

where V_{ij} is the ionization energy and V_e the excitation energy; $P*_e$ is the electron pressure (in atm; 1 atm ≈ 101 kPa) and is a function of T and n_e. In the case under discussion, the standard temperature for a line of an element which is present as an impurity in a plasma (e.g., in a noble gas), dilution in the plasma must also be considered. The standard temperatures indicate which types of line will be optimally excited in a plasma of given temperature, electron pressure, and gas composition, and are thus important indications for line selection. Atom lines often have standard temperatures below 4000 K, especially when analyte dilution in the plasma is high, whereas ion lines often have values of 10 000 K. Atom and ion lines are called "soft" and "hard" lines, respectively. Because of the interaction of free and bound electrons in a plasma continuous radiation is also emitted over a wide wavelength range. The intensity distribution for this background continuum can be expressed:

$$\begin{aligned} &I_\nu \mathrm{d}\nu = K n_e n_r r^2 / \left(T_e^{1/2}\right) \cdot \\ &\exp\left(-h\nu/kT_e\right) \cdot \mathrm{d}\nu \,(\text{free} - \text{free}) \\ &+K' \left(1/j^3\right) n_e n_Z Z^4 / \left(T_e^{3/2}\right) \cdot \\ &\exp\left(U_j - h\nu\right) / \left(kT_e\right) (\text{free} - \text{bound}) \end{aligned} \tag{62}$$

where n_r is the concentration of atoms with charge r, Z is the nuclear charge, and U_j is the ionization energy of the level with quantum number j. This background continuum consists of so-called bremsstrahlung, caused by the interaction of free electrons with each other, and of recombination radiation, which is due to the interaction of free and bound electrons. The latter is especially important in the UV region. It is continuous radiation with spectral lines superimposed. As T_e is the electron temperature, absolute measurements of the background continuum emission in a plasma, e.g., for hydrogen, allow the determination of the electron temperature in a plasma, whether in local thermal equilibrium or not.

2.5. Dissociation

The dissociation of molecular plasma gases or analyte molecules within the radiation source is an equilibrium reaction. Accordingly, highly stable radicals or molecules are always present in a radiation source. They emit molecular bands which are superimposed on the atomic and ionic line spectra in the emission spectrum. Radicals and molecules may also give rise to cluster ions which may be detected in mass spectra. Common species in plasma gases are: CN, NH, NO, OH, and N_2 (or N_2^+). From the analytes, highly stable oxides may persist (e.g., AlO^+, TiO^+, YO^+). A thorough treatment of molecular spectra is available [17], [18].

Vibrational and Rotational Hyperfine Structures. Molecules or radicals have various electronic energy levels ($^1\sum$, $^2\sum$, $^2\Pi, \ldots$), which have vibrational fine structure ($\nu = 0, 1, 2, 3, \ldots$), which, in turn, have rotational hyperfine structure ($J = 0, 1, 2, 3, \ldots$). The total energy of a state may be written:

$$E_i = E_{e1} + E_{vib} + E_{rot} \tag{63}$$

E_{e1} is of the order of 1 – 10 eV, the energy difference between two vibrational levels of the same electronic state is of the order of 0.25 eV, and the separation of rotational levels is of the order of 0.005 eV. When the rotational levels considered belong to the same electronic level, the emitted radiation is in the infrared region. When they belong to different electronic levels, they occur in the UV or visible region. Transitions are characterized by the three quantum numbers of the states involved: n', ν', j' and n'', ν'', j''. All lines which originate from transitions between rotational levels belonging to different vibrational levels of two electronic states form the band: $n', \nu' \longrightarrow n'', \nu''$. For these band spectra the selection rule is $\Delta j = j' - j'' = \pm 1, 0$. Transitions for which $j'' = j' + 1$ give rise to the P-branch, $j'' = j' - 1$ to the R-branch, and $j' = j''$ to the Q-branch of the band. The line corresponding to $j' = j'' = 0$ is the zero line of the band. When $\nu' = \nu'' = 0$ it is also the zero line of the system. The difference between the wavenumber of a rotation line and the wavenumber of the zero line in the case of the P- and the R-branch is a function of the rotational quantum number j and the rotational constant B_ν for which:

$$E_j / (hc) = B_\nu j (j+1) \tag{64}$$

The functional relation is quadratic known as the Fortrat parabola. As in the case of atomic spectral lines, the intensity of a rotational line can be written:

$$I_{nm} = (N_m A_{nm} h\nu_{nm}) / 2\pi \quad (65)$$

where N_m is the population of the excited level and ν_{nm} the frequency of the emitted radiation. The transition probability for dipole radiation is:

$$A_{nm} = \left(64\pi^4 \nu_{nm}^3\right) / (3k) \left[1/(g_m)\right] \sum |Rn_i m_k|^2 \quad (66)$$

where i and k are degenerate levels of the upper (m) and the lower state (n), $Rn_i m_k$ is a matrix element of the electrical dipole moment and g_m is the statistical weight of the upper state; N_m is given by the Boltzmann equation:

$$N_m = N(g_m) / Z(T) \exp(-E_r/kT) \quad (67)$$

where E_r is the rotational energy of the excited electronic and vibrational level, given by:

$$E_r = hcB_{\nu'} J' \left(J'+1\right) \quad (68)$$

where $B_{\nu'}$ is the rotational constant and J' is the rotational quantum number of the upper state m. For a $^2\sum_g - {}^2\sum_u$ transition, between a so-called "gerade" (g) and "ungerade" (u) level, the term $\sum — Rn_i m_k —^2 = J' + J'' + 1$, where J' and J'' are the rotational quantum numbers of the upper and lower state. Accordingly:

$$I_{nm} = \left(16\pi^3 cN\nu_{nm}^4\right) / 3Z(T) \left(J'+J''+1\right) \exp\left(-hcB_{\nu'} J' \left(J'+1\right) / kT\right) \quad (69)$$

or

$$\ln\left[I_{nm} / \left(J'+J''+1\right)\right] = \ln\left[16\pi^3 cN\nu_{nm}^4\right] / \left[3Z(T)\right] - \left[hcB_{\nu'} J' \left(J'+1\right)\right] / kT \quad (70)$$

By plotting $\ln[I_{nm}/(J'+J''+1)]$ against $J'(J'+1)$ for a series of rotational lines a so-called rotational temperature can be determined, reflecting the kinetic energy of neutrals and ions in the plasma.

Molecular Bands. Spectral lines of molecular bands emitted by molecules and radicals in a plasma often interfere with atomic emission lines. In atomic absorption spectrometry, absorption by molecular bands arising from undissociated molecules in the atomizer may also interfere. Therefore, it is important to study the dissociation of molecular species in high-temperature radiation sources or atomizers. In plasma sources, band-emitting species may stem from the working gas. In this respect, N_2, N_2^+, CN, OH, and NH band emission are important. Undissociated sample and analyte species are also present in the plasma. Highly stable molecules such as Al_2O_3, La_2O_3, BaO, AlF_3, CaF_2, MgO, may be present in atomic spectrometric sources. It is important to know their dissociation as a function of plasma temperature and composition. This dependence can be described by a dissociation equation, similar to the Saha equation:

$$K_n = \left[\left(2\pi/h^2\right)\left(m_X m_Y / m_{XY}\right)(kT)^{3/2}\right] \cdot \left(Z_X Z_Y / Z_{XY}\right)\left[\exp\left(-E_d/kT\right)\right] \quad (71)$$

where:

$$K_n = \left(n_X n_Y\right) / n_{XY} \quad (72)$$

Thus, for a metal oxide the degree of dissociation can be calculated when the plasma temperature, the partial pressure of oxygen in the plasma, and the dissociation constant are known.

2.6. Sources and Atom Cells

For the purposes of the following discussion it is useful to distinguish between the terms "source" and "atom cell". Sources are the generators of the radiation (narrow line or continuum) which is eventually to be measured by the detector. Atom cells are the devices for producing the population of free atoms (or ions) from a sample. In atomic absorption and fluorescence spectrometry these two devices are separate. In atomic emission spectrometry the source is also the atom cell. In atomic mass spectrometry the terms source and atom cell are often used interchangeably to denote the "ion source".

In atomic spectrometry, the sample is introduced, by means of a sampling device, into a high-temperature source or atom cell (plasma, flame, etc.). Here, the sample is vaporized, e.g., by thermal evaporation or sputtering. It is important to supply as much energy as possible, so that the volatilization processes, which involve a physical or chemical equilibrium, result in complete atomization, irrespective of the state of aggregation, solid state structure, or chemical composition of the sample. This is very important, both to ensure maximum sensitivity and

to minimize matrix interference in the analysis. The effectiveness of the volatilization processes involved, the plasma temperature, and the number densities of the various plasma components will all influence sample atomization.

Rotational temperatures are relevant for all processes in which molecules, radicals, and their dissociation products are involved. They can be obtained from the intensity distribution for the rotational lines in rotation – vibration spectra. Molecules such as OH or CN have often been used as "thermometers" (Eqs. 63 – 70) [19–21].

The *gas temperature* is determined by the kinetic energy of neutral atoms and ions. It can be estimated from the Doppler broadening (Eq. 39). However, the contributions of Doppler and temperature broadening have to be separated, which involves the use of complicated deconvolution procedures [22].

The *electron temperature* is a measure of the kinetic energy of the electrons. It is relevant in the study of excitation and ionization by collisions with electrons, which is an important process for analyte signal generation. The electron temperature can be determined from the intensity of the recombination continuum or of the bremsstrahlung (Eq. 62).

The *excitation temperature* describing the population of the excited levels of atoms and ions – important in studies of the dependence of analyte line intensities on various plasma parameters in analytical emission spectrometry. It can be determined from the intensity ratio of two atomic emission lines of the same element and state of ionization (Eq. 33) or from plots of the appropriate function for many atomic emission lines against their excitation energies.

The *ionization temperature* is relevant for all phenomena involving equilibria between analyte atoms, ions, and free electrons in plasmas. In the case of thermal equilibrium, it occurs in the Saha equation (Eqs. 53, 54) and can be determined from the intensity ratio of an ion and an atom line of the same element. In all other cases, ionization temperatures can be determined from the n_e value obtained from Stark broadening (Eqs. 59, 60). Temperatures for the most important sources in atomic spectrometry are listed in Table 1.

In a plasma which is at least in local thermal equilibrium, all these temperatures are equal. This implies that the velocity distribution of all types of particle in the plasma (molecules, atoms, ions, and electrons) at any energy level can be described by the Maxwell equation (15). For all species, the population of the different levels is then given by the Boltzmann equation (Eq. 17). Further, the ionization of atoms, molecules, and radicals can be described by the Saha equation (Eqs. 53, 54) and the related equations of chemical equilibrium. Finally, the radiation density in the source obeys Planck's law, and the exchange of both kinetic energy between particles and electromagnetic radiation are in equilibrium with each other. Real plasma sources used in atomic spectrometry are, at best, approximations to local thermal equilibrium. Although processes between the particles may be in thermal equilibrium, those involving electromagnetic radiation are not, as the plasma cannot be considered as a closed system. Moreover, real plasma sources are extremely inhomogeneous with respect to temperature and number density distributions. Accordingly, the equilibria occur only within small volume elements of the sources, in which gradients can be neglected. Many plasmas, however, have cylindrical symmetry and can be observed laterally, integrating the information from many volume elements along the observation direction. From this integral value, the values at a defined radial distance from the plasma center can be calculated. This

Table 1. Temperatures of sources and atom cells in atomic spectrometry

Source	Temperature, K				State
	Rotational, T_{rot}	Excitation, T_{ex}	Electron, T_e	Ion T_i	
Arc (d.c.)	5000	5 000	5 500	5 000	LTE
Spark		20 000	20 000	20 000	LTE
Inductively coupled plasma	4800	5 000	6 000	6 000	≈ LTE
Microwave plasma	2000	4 000	6 000	6 000	non-LTE
Low-pressure discharge	600	20 000	30 000	30 000	non-LTE

necessitates side-on observation along several axes, which are equidistant with respect to each other, and a so- called Abel inversion. If $I'(x)$ is the first derivative of function $I(x)$, describing the variation of a measured value as a function of the lateral position x during side-on observation (Fig. 3), the radial values at distance r are:

$$I(r) = -(1/\pi) \int_{r}^{\infty} [I'(x)\,\mathrm{d}x] / \sqrt{(r^2 - x^2)} \qquad (73)$$

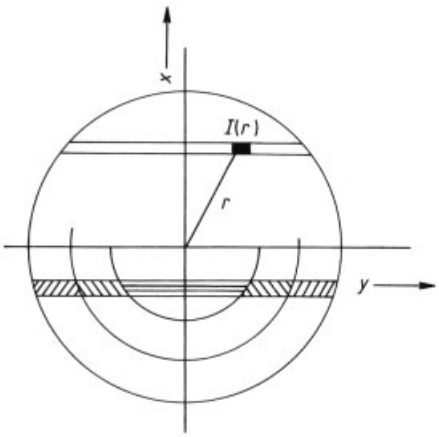

Figure 3. Abel inversion procedure for the determination of radial distributions in plasma sources
$I(r)$ = Intensity at radial distance r in volume element $\mathrm{d}x \times \mathrm{d}y$

This integral, known as the Abel integral, can only be solved numerically, writing each laterally measured intensity as a sum of the different volume elements with coefficients denoting their contribution. For certain numbers of lateral observation positions (3, 10, . . .) these Abel coefficients are listed in the literature [23]. By inversion of the corresponding matrix the radial values are obtained. This allows determination of radial distributions of emissivities, absorbances, temperatures, or number density distributions in plasmas. Repeating this procedure at different heights in a plasma is equivalent to plasma tomography. Similar results, which do not require the assumption of cylindrical symmetry in the source can be obtained by imaging spectrometry with two-dimensional detectors [24].

In real plasmas, departures from thermal equilibrium often occur. In the extreme case, as encountered in plasmas under reduced pressure, emission or absorption of radiation becomes so important that there is no longer a clear relation between the mean kinetic energies of the species and the excitation temperatures. The latter then lose the physical meaning of temperature. The absence of local thermal equilibrium in these non-LTE plasmas relates to the existence of high field gradients or a.c. fields; only the electrons, not the heavy atoms and ions, can follow the field changes and fully take up the dissipated energies. Accordingly, the mean kinetic energy of the electrons and thus the electron temperature is much higher than the gas temperature.

Sources and atom cells for atomic spectrometry include flames, arcs, sparks, low-pressure discharges, lasers as well as d.c., high-frequency, and microwave plasma discharges at reduced and atmospheric pressure (Fig. 4). They can be characterized as in Table 2. *Flames* are in thermal equilibrium. Their temperatures, however, are at most 2800 K. As this is far below the standard temperature of most elemental lines, flames have only limited importance as sources for atomic emission spectrometry, but are excellent atom cells for absorption, fluorescence, and laser-enhanced ionization work. *Arcs and sparks* are well-known sources for atomic emission spectrometry. The high temperature obtained in spark sources leads to the excitation of ion lines, for which the standard temperatures are often beyond 10 000 K, whereas in arc sources atom lines predominate. In *plasma sources* under reduced pressure the gas kinetic temperatures are low and their atomization capacity is limited. When these sources are used as atomic emission sources or as primary sources in atomic absorption, line widths are very narrow. Moreover, especially with gases having high ionization energy such as helium, high-energy lines such as those of the halogens can be excited. Low-pressure discharges are valuable ion sources for mass spectrometry as analyte ionization takes place. Because of the low pressure, coupling with a mass spectrometer is possible. So-called plasmajet and plasma sources at atmospheric pressure are especially useful for the emission spectrometric analysis of solutions. Their gas kinetic temperatures are high enough to achieve complete dissociation of thermally stable oxides and both atom and ion lines occur in the spectra. As reflected by the fairly high ionization temperature, they are powerful ion sources for mass spectrometry; plasma mass spectrometry is now

one of the most sensitive methods in atomic spectrometry. Lasers are very suitable for ablation of solids. Owing to their high analyte number densities, plasmas are subject to high self-absorption so it is better to use them solely for volatilization, and to introduce the ablated material in a second source for signal generation.

Table 2. Applications of sources and atom cells for atomic spectrometry

Source	Spectroscopy	Sample	Concentration range [a]
Flames	AES [b]	liquid	t (alkali)
	AAS [c]	liquid (solid)	t
	AFS [d]	liquid	ut
	LEI [e]	liquid	ut
Arc (d.c.)	AES [b]	solid, liquid	t
	MS [f]	solid	ut
Spark	AES [b]	solid	t
	MS [f]	solid	ut
Glow discharge	AES [b]	solid, liquid, gas	t
	AAS [c]	solid, liquid, gas	t
	AFS [d]	solid, liquid, gas	ut
	MS [f]	solid, gas	ut
	LEI [e]	solid, gas	ut
Laser	AES [b]	solid	m
	AAS [c]	solid	t
	AFS [d]	solid	ut
	MS [f]	solid	ut
	LEI [e]	solid	ut
Inductively coupled plasma	AES [b]	liquid (solid, gas)	t
	AFS [d]	liquid (solid, gas)	t
	MS [f]	liquid (solid, gas)	ut
Microwave plasma	AES [b]	gas (liquid)	t
	MS [f]	gas (liquid)	ut
Furnace	AES [b]	liquid, solid	t
	AAS [c]	liquid, solid	ut

[a] m = Minor; t = Trace; ut = Ultra-trace.
[b] AES = Atomic emission spectroscopy.
[c] AAS = Atomic absorption spectroscopy.
[d] AFS = Atomic fluorescence spectroscopy.
[e] LEI = Laser-enhanced ionization.
[f] MS = Mass spectrometry.

2.7. Analytical Atomic Spectrometry

Analytical atomic spectrometry nowadays includes the use of flames and plasma discharges for optical and for mass spectrometry. These are used directly as emission sources, as atomizers for atomic absorption or atomic fluorescence, or for ion production. In a source for atomic spectrometry atomic vaporization and signal generation take place together. The first processes require high energy for complete atomization, whereas the signal generation processes often require discrete, selective excitation. Therefore, the use of so-called tandem sources is preferred, where analyte vapor generation and signal generation take place separately [26].

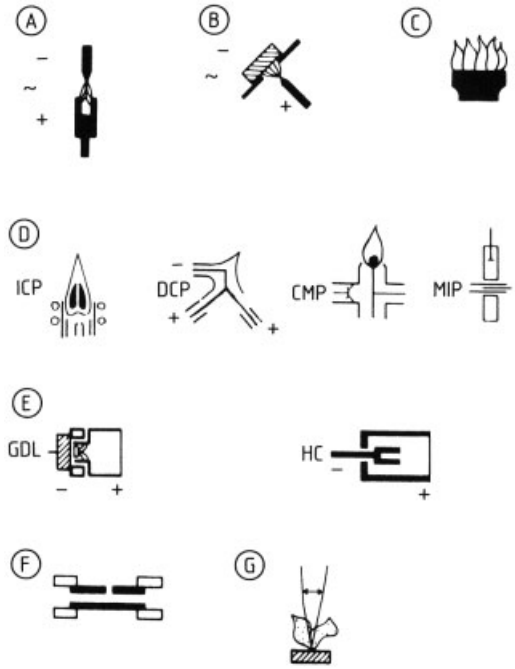

Figure 4. Sources for atomic spectrometry [25]
A) Arc; B) Spark; C) Flame; D) Plasma sources; E) Low-pressure discharges; F) Graphite furnace; G) Laser plume
ICP = Inductively coupled plasma; DCP = Direct current plasmajet; CMP = Capacitively-coupled microwave plasma; MIP = Microwave-induced plasma; GDL = Glow discharge lamp; HC = Hollow cathode

For atomic emission spectrometry, selectivity is achieved by isolation of the spectral line at the exit slit of the spectrometer. This puts high demands on the optical quality of the spectral apparatus. In atomic absorption and fluorescence, selectivity is partly controlled by the radiation source delivering the primary radiation, which in most cases is a line source (hollow cathode lamp, laser, etc.). Therefore, the spectral bandpass of the monochromator is not as critical as it is in atomic emission work. This is especially true for laser-based methods, where in some cases of atomic fluorescence a filter is sufficient, and in laser-induced ionization spectrometry where no spectral isolation is required at all. In the case of

glow discharges and inductively coupled high-frequency plasmas, ion generation takes place in the plasmas. In the first case, one can perform direct mass spectrometry on solids, and in the second case on liquids or solids subsequent to sample dissolution. In the various methods of atomic spectrometry, samples have to be delivered in the proper form to the plasma source. Therefore, in the treatment of the respective methods attention will be given to techniques for sample introduction.

3. Spectrometric Instrumentation

Atomic spectrometric methods of analysis essentially make use of equipment for spectral dispersion to achieve their selectivity. In optical atomic spectrometry, this involves the use of dispersive as well as nondispersive spectrometers, whereas in the case of atomic spectrometry with plasma ion sources mass spectrometric equipment is used. In both cases, suitable data acquisition and processing systems are built into the instruments.

In the design of instruments for atomic spectrometry the central aim is to achieve the full potential of each method. This involves the power of detection and its relation to precision, freedom from spectral interference, and the price – performance ratio.

3.1. Figures of Merit of an Analytical Method

In practice, atomic spectrometric methods are relative methods and need to be calibrated. The calibration function describes the relationship between the analytical signals (absorbances, absolute or relative radiation intensities, or ion currents). In its simplest form, the calibration function can be written (→ Chemometrics, Chap. 12.):

$$Y = ac + b \tag{74}$$

The inverse function is often used:

$$c = a'Y + b' \tag{75}$$

This is known as the analytical evaluation function. Calibration curves are often nonlinear, so a polynomial function is needed:

$$Y = a_0 + a_1 c + a_2 c^2 + \ldots \tag{76}$$

where $a_0, a_1, a_2, \ldots$ are determined by multivariate regression. Normally, a second-degree polynomial is sufficient to describe the calibration function over a large concentration range. Alternatively, segmented calibration curves can be used if the calibration function is nonlinear. For a series of analytical signal measurements for a well-defined analyte concentration, a statistical uncertainty exists which stems from fluctuations in the analytical system. The precision achievable is an important figure of merit of an analytical procedure, for the statistical evaluation of analytical data see, → Chemometrics, and classical textbooks [27–29]. The precision of a measurement procedure is expressed in terms of the standard deviation:

$$\sigma = \sqrt{\left[\sum (Y_{\mathrm{m}} - Y_i)^2 / (n-1)\right]} \tag{77}$$

where $Y_{\mathrm{m}} = \sum (Y_i/n)$ is the mean value, Y_{i} an individual measurement, and n the number of measurements. The precision of a certain measurement system depends on the noise in the system. Different types of noise can be distinguished [30].

Fundamental or Random Noise. This is statistically distributed and its amplitude as a function of frequency can be written as a sum of sinusoidal functions. This type of noise is related to the discrete nature of matter and the quantization of radiation, and cannot be completely eliminated.

Nonfundamental, Flicker, or Excess Noise. Here, the sign or the magnitude can correlate with well-defined phenomena.

Interference Noise. This is observed at well-defined frequencies and mostly stems from components in the system.

Both nonfundamental and interference noise can often be eliminated by appropriate filtering.

In a noise spectrum (Fig. 5) the noise amplitude is plotted as a function of frequency [31]. White noise (Fig. 5 A) occurs over all frequencies and is almost always fundamental in origin,

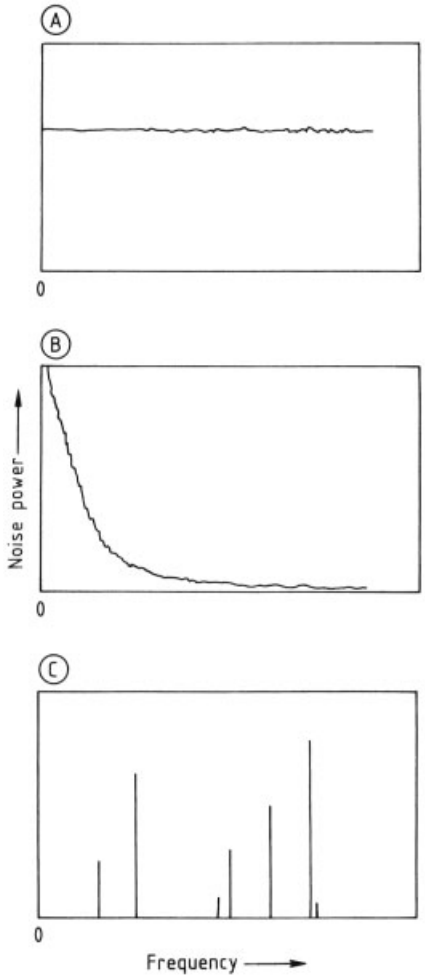

Figure 5. Noise power spectra commonly found in chemical instrumentation
A) White noise; B) Flicker (1/*f*) noise; C) Interference noise [31]

whereas for 1/*f* noise (Fig. 5 B) the amplitude decreases with the frequency and it is nonfundamental in origin. Discrete noise bands with well-defined causes may also occur (Fig. 5 C). These may stem from the source, caused by gas flow dynamics or contributions from the vacuum line, etc. Noise spectra are a powerful diagnostic tool to trace the sources of noise, and to study instrumental limitations to the power of detection. In atomic spectrometry, it is important to determine if noise from the detector is predominant; if so, it can be described by a Poisson distribution:

$$\sigma^2 = n \tag{78}$$

where n is the number of events per unit time and σ the standard deviation. Alternatively, the background noise of the source may be much more important, or Flicker noise or frequency-dependent noise are predominant. In the latter case, overtones often occur.

For the precision of an analytical method, not only the reproducibility of single measurements, but also calibration errors have to be considered. This is a complex problem as, depending on the nature of the noise, precision may vary considerably with concentration.

When a linear regression is performed, the standard deviation of the regression can be defined:

$$s(Y) = \sqrt{\left[\sum (Y_i - Y)^2 / (n-2)\right]} \tag{79}$$

where Y_i is the signal obtained for a standard sample with concentration c from the regression equation. The latter is calculated by a least-squares procedure from the pairs (c, Y), where Y are the measured signals, and n is the total number of measurements. The standard deviation for the concentration of the analytical sample c_x can be calculated through propagation of errors:

$$s_{\mathrm{r}}(c_x) = \ln_{10} s(c_x) = \ln_{10} a' s(Y) \tag{80}$$

where a' is the slope of the calibration curve. The magnitude of the concentrations of the analytical sample with respect to those of the standard samples has to be considered, and can be included in Eq. (80):

$$s_{\mathrm{r}}(c_x) = \ln_{10} a' s(Y) \cdot \sqrt{\left\{(1/n) + (1/m) + (c_x - c_{\mathrm{m}})^2 / \left[\sum (c - c_{\mathrm{m}})^2\right]\right\}} \tag{81}$$

where m is the number of replicates for the analytical sample and c_{m} is the mean of all the standard concentrations measured. Equations (74) and (75) are valid only for a limited concentration range, known as the linear dynamic range. This range is limited at the upper end by physical phenomena, such as detector saturation, and at the lower end by the limit of determination. This limit is typical of a given analytical procedure and is the lowest concentration at which a determination can be performed with a certain precision [32] (→ Chemometrics, Chap. 12.; → Trace Analysis, Chap. 4.2., → Trace Analysis, Chap. 4.3.). For 99.86 % certainty, and

provided the fluctuations of the limiting noise source can be described by a normal distribution, the lowest detectable net signal Y_L is three times the relevant standard deviation:

$$Y_L = 3\sigma^* \tag{82}$$

The net signal is determined from the difference between a background signal and a gross signal including analyte and background contributions, so a factor of $\sqrt{2}$ has to be introduced. In many cases the limiting standard deviation is often the standard deviation obtained for a series of blank measurements [33]. From the calibration curve the detection limit is:

$$c_L = a' \left(3\sqrt{2}\right) \sigma \tag{83}$$

The detection limit is thus closely related to the signal-to-background and the signal-to-noise ratios. It is the concentration for which the signal-to-background ratio is $3\sqrt{2}$ times the relative standard deviation of the background, or at which the signal-to-noise ratio is $3\sqrt{2}$. The signal-to-noise ratio itself is related to the types of noise occurring in the analytical system. From a knowledge of the limiting noise sources, well-established signal acquisition measures can improve the signal-to-noise ratio and hence the power of detection.

The analytical signals measured often include contributions from constituents other than the analyte (e.g., matrix constituents). This is known as spectral interference and can be corrected by subtracting contributions calculated from the magnitude of interference and the concentration of the interfering species. A special type of interference influences the background signal on which the analyte signals are superimposed; a number of correction methods exist. Freedom from interference is an important figure of merit for an analytical method.

3.2. Optical Spectrometers

In optical atomic spectrometry, the radiation emitted by the radiation source, or the radiation which comes from the primary source and has passed through a separate atomizer, must be fed into a spectrometer; this radiation transfer should be as complete as possible. The amount of radiation passing through an optical system is expressed by its optical conductance G_0:

$$G_0 = \int_A \int_B (\cos\alpha_1 \cos\alpha_2 \, dA \, dB) / a_{12}^2 \tag{84}$$

$$\sim (AB) / a_{12}^2 \tag{85}$$

where dA and dB are surface elements of the entrance and exit apertures, a_{12} is the distance between them, and α_1, α_2 are the angles between the normals of the aperture planes and the radiation. If n is the refractive index of the medium, the optical conductance is:

$$G = G_0 n^2 \tag{86}$$

The radiant flux through an optical system is:

$$\varphi = \tau B G \tag{87}$$

where τ is the transmittance, determined by reflection or absorption losses at the different optical elements, and B is the radiant density of the source (W m^{-2} sterad^{-1}). For optimal optical illumination of a spectrometer, the dispersive element should be completely illuminated to obtain full resolution. However, no radiation should bypass the dispersive element, as stray radiation may produce anomalous signals. Further, the optical conductance at every point in the optical system should be maximum.

3.2.1. Optical Systems

The type of illumination system needed to fulfill these conditions depends on the source and detector dimensions, the homogeneity of the source, the need to fill the detector homogeneously with radiation, the distance between the source and the entrance aperture of the spectrometer, and the focal length of the spectrometer.

In conventional systems, lenses and imaging mirrors are used. Glass lenses can be used only at wavelengths above 330 nm. For quartz, radiation with wavelengths down to 165 nm is transmitted, but evacuation or purging the illumination system and spectrometer with nitrogen or argon is required below 190 nm, as a result of the absorption of short-wavelength radiation by oxygen. At wavelengths below 160 nm, the system must be evacuated and MgF_2 or LiF optics must be used. With lenses, three main illumination systems are used.

Imaging on the Entrance Collimator. In this system a lens is placed immediately in front of the entrance slit to image the relevant part of the radiation source on the entrance collimator (Fig. 6 A). This has the advantage that the entrance slit is homogeneously illuminated; however, stray radiation may occur inside the spectrometer. The distance between the source and the entrance slit a is given by the magnification required:

$$x/W = a/f_k \tag{88}$$

where x is the width of the source, W is the width of the entrance collimator, and f_k its focal length. The f-number of the lens is:

$$1/f = 1/a + 1/f_k \tag{89}$$

The focal length of a lens depends on the wavelength:

$$f(\lambda_1)/F(\lambda_2) = n_1/n_2 \tag{90}$$

where n_1, n_2 are the refractive indices of the lens material at wavelengths λ_1, λ_2. For quartz, a factor of 0.833 has to be applied when passing from the Na 583 nm D-line to 200 nm.

Illumination with Intermediate Image. In the second illumination system, a field lens is used to produce an intermediate image on a diaphragm. A suitable zone can be selected to fully illuminate the collimator mirror with the aid of a lens placed immediately in front of the exit slit. A third lens is used to illuminate the entrance slit homogeneously (Fig. 6 B). The magnification is divided over all three lenses, so that chromatic aberrations are minimized, but the set-up is inflexible. The distances between the three lenses must be selected to achieve the required magnifications:

$$x/I = a_1/a_2 \tag{91}$$

$$D/s_\mathrm{h} = a_2/a_3 \tag{92}$$

and

$$I/W = a_3/f_k \tag{93}$$

where I is the diameter of the intermediate image, D the diameter of the field lens, s_h the entrance slit height, and a_1, a_2, a_3 are the distances between the lenses. Further

$$a_1 + a_2 + a_3 = A \tag{94}$$

is mostly fixed by constructional factors, and:

$$1/f_1 = 1/a_1 + 1/a_2 \tag{95}$$

$$1/f_2 = 1/a_2 + 1/a_3 \tag{96}$$

$$1/f_3 = 1/a_3 + 1/f_k \tag{97}$$

give the f-numbers of the respective lenses. Only one parameter (e.g., the width of the intermediate image) can be freely selected because x, W, s, A, and f_k are fixed.

Image on the Entrance Slit. The third illumination system produces an image on the entrance slit with the aid of one lens; the structure of the source is imaged on the entrance collimator and the detector. This allows spatially resolved line intensity measurements when a detector with two-dimensional resolution, such as a photographic emulsion or an array detector (see Section 3.2.2), is employed. This type of imaging is often used for diagnostic studies.

Fiber optics are very useful for radiation transmission. Small lenses are necessary to accommodate the entrance angle of the fiber, which depends on the refractive index of the material (usually quartz), and is typically 30 – 40°. A typical illumination of a spectrometer with an optical fiber (Fig. 7) uses a lens (diameter d) for imaging the source on the fiber. Bundles of fiber beams with diameter $D = 600\,\mu\mathrm{m}$ are often used. The magnification x/D as well as the entrance angle $\alpha = \arctan(d/2)/a_2$ and the lens formula determine the f-number of the lens and d. At the fiber exit, a lens allows the radiation to enter the spectrometer, without causing stray radiation.

With quartz fibers, radiation is efficiently fed into spectrometer, but the transmittance decreases seriously below 220 nm. This may lead to detector noise limitations for analytical lines at lower wavelengths.

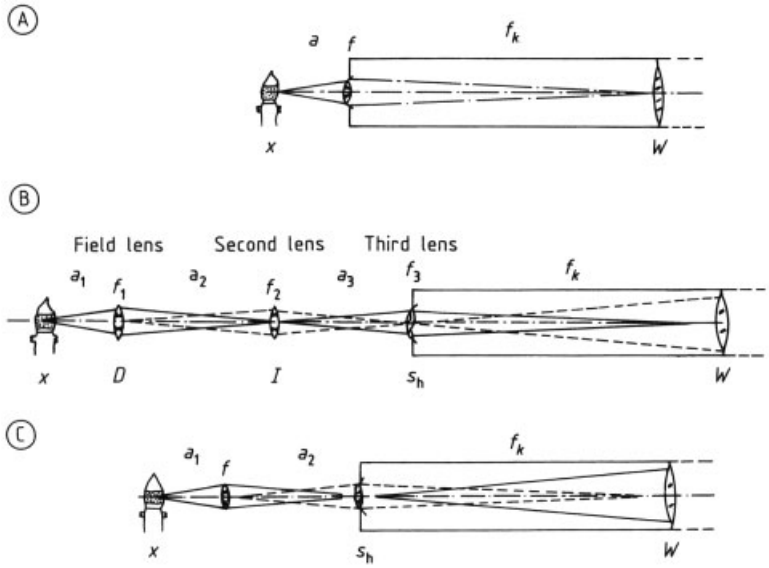

Figure 6. Illumination of the optical spectrometer with lenses
A) Imaging on the entrance collimator; B) Illumination with intermediate image; C) Imaging on the entrance slit
For explanation of symbols see text

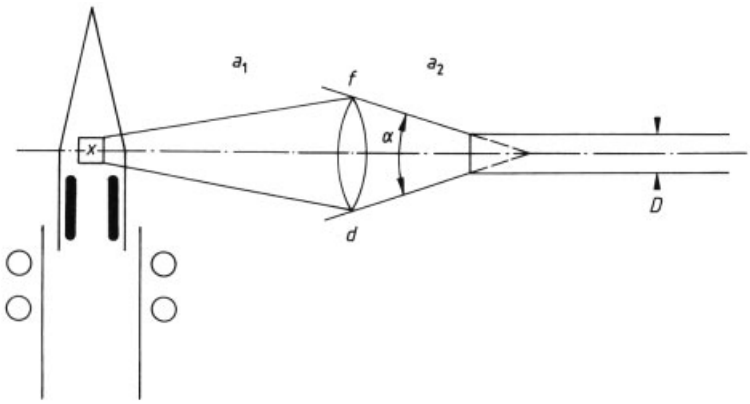

Figure 7. Use of optical fibers for illumination of an optical spectrometer
α = Entrance angle of the fiber; d = Lens diameter; D = Fiber diameter; x = Diameter of the zone in the radiation source to be selected; α_1 = Distance between lens and radiation source; α_2 = Distance between fiber entrance and lens

Spectrometers are used to produce the spectrum or to isolate narrow spectral ranges, without further deconvolution. In dispersive apparatus, the spectrum is produced with a prism or diffraction grating. In nondispersive apparatus, spectral regions are isolated from the radiation beam by reflection, interference, or absorption in interferometers or filter monochromators. The latter are of use only in flame emission spectrometry.

A *dispersive spectrometer* contains an entrance collimator, a dispersive element, and an exit collimator [34], [35].

The entrance collimator produces a quasi-parallel beam from the radiation coming through the entrance aperture, of width s_e and height h_e. The entrance collimator has focal length f_k and width W. The diffraction slit width s_0 and height h_0 are the half-widths:

$$s_0 = \lambda f / W$$

and

$$h_0 = \lambda f / h \tag{98}$$

The entrance aperture dimensions should not be smaller than s_0 and h_0, so that diffraction does not limit resolution. The value of f/W is a measure of the amount of radiant energy entering the spectrometer. The exit collimator images the monochromatized radiation leaving the dispersive element on the exit slit, giving a series of monochromatic images of the entrance slit. In a monochromator, a single exit slit allows isolation of one line after another by turning the dispersive element. In a polychromator the dispersive element is fixed and there are many exit slits placed where the images of the lines of interest occur. They are often on a curved surface with a radius of curvature R (Rowland circle). Simultaneous measurement of several lines and thus simultaneous multielement determinations are possible. With a spectrographic camera, lines are focused in a plane or on a slightly curved surface, where a two-dimensional detector can be placed. With such a detector (photographic plate, diode array detector, etc.), the signal intensities of part of the spectrum over a certain wavelength range or intensities of a spectral feature at several locations in the source can be simultaneously recorded. The energy per unit area at the detector is given by the irradiance:

$$E = \varphi \cos\alpha / A \tag{99}$$

Diffraction Gratings. Diffraction gratings are used almost exclusively as the dispersive element in modern apparatus; prisms are still used as predispersers and cross-dispersers. Both plane and concave gratings are used; the latter have imaging qualities. Because of the profile of the grooves, holographically ruled gratings have a rather uniform radiant output over a large spectral area compared with mechanically ruled gratings which always have a so-called blaze angle, and hence a blaze wavelength where radiant energy is maximum. Modern spectrometers tend to use reflection gratings. As a result of interference, a parallel beam of white radiation at incident angle φ_1 with the grating normal is dispersed through angle φ_2 ; radiation with wavelength λ is sorted (Fig. 8) according to Bragg's equation:

$$sin\varphi_1 + \sin\varphi_2 = m\lambda / a \tag{100}$$

where m is the order of interference and $a = 1/n_G$ is the grating constant, where n_G is the number of grooves per unit length. When B is the grating width, the total number of grooves $N = B\, n_G$ determines the theoretical resolving power R_0:

$$R_0 = B n_G m \tag{101}$$

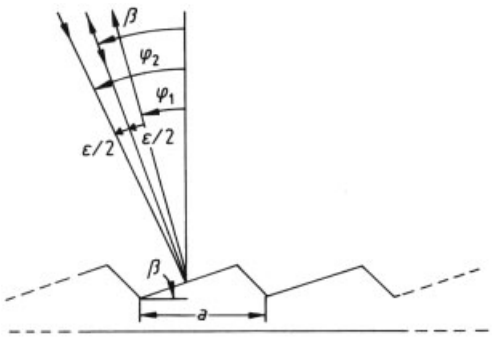

Figure 8. Spectral dispersion at a diffraction grating α = Grating constant; β = Blaze angle; φ_1 = Angle of incident radiation; φ_2 = Angle of diffracted radiation with wavelength λ; $\varepsilon = \varphi_1 - \varphi_2$

The angular dispersion can be obtained by differentiation of Equation (101) with respect to λ:

$$\mathrm{d}\varphi_2/\mathrm{d}\lambda = m/\,(a\cos\varphi_2) \tag{102}$$

The reciprocal linear dispersion is:

$$\begin{aligned} \mathrm{d}x/\mathrm{d}\lambda &= (\mathrm{d}x/\mathrm{d}\varphi_2)\,(\mathrm{d}\varphi_2/\mathrm{d}\lambda) \\ &= \left(f_\mathrm{k}/\cos\theta'\right)(m/a\cos\varphi_2) \end{aligned} \tag{103}$$

where θ' is the slope of the surface of the radiation detector. The form and depth of the grooves determine the blaze angle β defined earlier. The blaze wavelength for order m is:

$$\sin\beta = \lambda_\mathrm{B} m / \left[2a\cos\left(\varepsilon/2\right)\right] \tag{104}$$

where $(\varphi_1 - \varphi_2) = \varepsilon$. In stigmatic spectrometers the height and width of the slit are imaged in the same plane. Several mountings (Fig. 9) are used with a plane gratings.

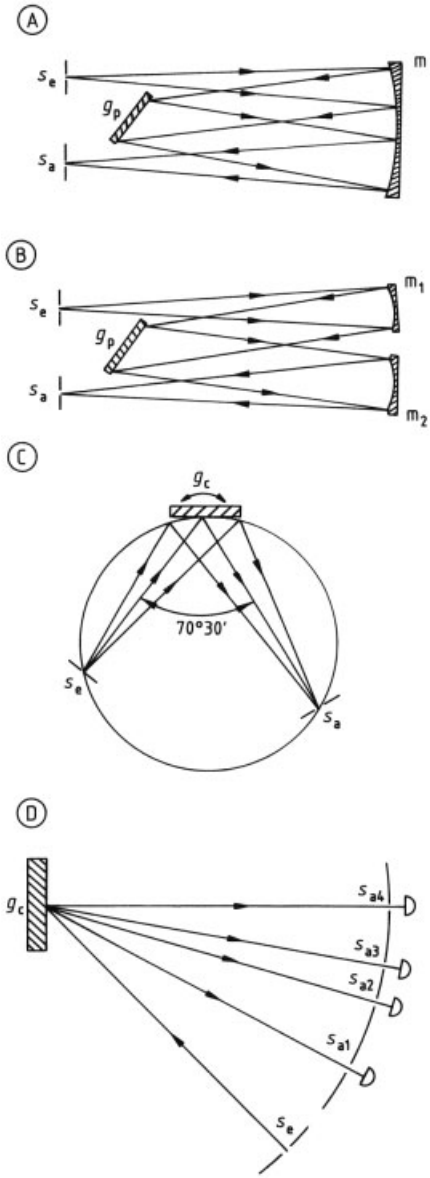

Figure 9. Optical mountings for optical spectrometers with a plane grating
A) Ebert; B) Czerny – Turner; C) Seya – Namioka; D) Paschen – Runge
s_e = Entrance slit; s_a = Exit slit; g_P = Plane grating; g_c = Concave grating; m = Mirror

In the Czerny – Turner mounting (Fig. 9 B), two spherical mirrors with slightly different focal lengths are positioned at slightly different angles to correct for spherical aberration. This mounting is used for high-luminosity monochromators with short focal length and highly dispersive gratings (more than 1800 grooves per millimeter). In the Ebert mounting (Fig. 9 A), there is only one mirror serving for both entrance and exit collimator; accordingly, aberrations occur and require the use of curved slits, to achieve maximum resolution. The Fastie – Ebert mounting is used for large spectrographs. The entrance collimator lies under the exit collimator plane. The spectrum is focused in a plane and the f/W value can be as low as 1/30, which may require the use of a cylindrical lens to increase the detector irradiance. Photographic plates or films are used as detectors.

Normally, gratings are used at low orders; different orders can be separated by special photomultipliers. With a solar-blind photomultiplier only radiation with wavelength below 330 nm can be detected. This allows separation of the first order radiation at 400 nm from the second order at 200 nm. This can be used in polychromators to double the effective resolution. So-called echelle gratings have low groove density (up to 1/100 mm) and therefore enable the use of orders up to 70 [36]. Here, the orders overlap and must be separated by a second dispersive element (e.g., a prism) either with its axis parallel to that of the echelle grating or in the so-called crossed-dispersion mode. In the latter case, the spectrum occurs as a number of dots of height equal to that of the entrance slit. The radiant flux is:

$$\varphi = (B_\lambda \tau s h W H \cos\beta) / f^2 \qquad (105)$$

where B_λ is the spectral radiance, τ is the transmittance of the optics, s is the slit width and h is height, W is the width, H the height of the grating, and β is the angle between the collimator plane and the radiation beam. Echelle spectrometers often use an Ebert mounting for high resolution at low focal length. They are thermally very stable and are used as both monochromators and polychromators. For concave gratings, the radius of the grating determines the Rowland circle. In the direction of dispersion spectral lines are focused on the Rowland circle as monochromatic images of the entrance slit. In the Paschen – Runge mounting (Fig. 9 D), grating, entrance slit, and all the exit slits are fixed on the Rowland circle. This mounting is most often used in large simultaneous polychromators with

photoelectric detection. In the Eagle mounting, the grating can be simultaneously turned around its axis and moved along the radiation direction. This mounting can be easily housed in a narrow tank; it is mechanically rather simple and is much used in vacuum monochromators. In the Seya – Namioka mounting (Fig. 9 C), the directions of incident and dispersed radiation beams are kept constant at an angle of 70°30′. Wavelength selection at the exit slit is performed by rotating the grating. At large angles serious defocusing occurs and the aberrations are larger than in the Eagle mounting. The Seya – Namioka system is often used in vacuum monochromators.

In most sequential spectrometers, switching from one line to another is achieved by turning the grating. From Equation (100) it can be seen that, as the spectral lines have widths of 1 – 3 pm, angle selection must be very accurate. For a grating with $a = 1/2400$ mm and a line width of 1 pm, the grating must be positioned with an accuracy of 0.0001°.

It is possible to step through the line profile or to achieve random access by using angle encoders. Computer-controlled stepper motors are used to turn the grating. Their angular resolution is often above 1000 steps per turn. For a so-called sinebar drive, the number of steps performed by the motor is directly proportional to the wavelength displacement. A pen recorder is running off on the spindle driven by the motor. A further approach is to use a Paschen – Runge spectrometer with equidistant exit slits and a detector moving in the focal plane. Fine adjustment of the lines is by computer-controlled displacement of the entrance slit.

3.2.2. Detectors

Photographic emulsions or photoelectric devices can be used as detectors for electromagnetic radiation between 150 and 800 nm. Among the latter, photomultipliers are the most important, but solid-state devices are now becoming very common.

Photographic emulsions allow recording of the whole spectrum simultaneously and, accordingly, they are useful for multielement survey analysis. However, their processing is slow, precision is low, and they are not capable of on-line data processing. Quantitative treatment of photographically recorded spectra is of limited importance for analytical use and current interest in photographic detection is limited to research in which documentation of the whole spectrum is required. Therefore, it is treated only briefly.

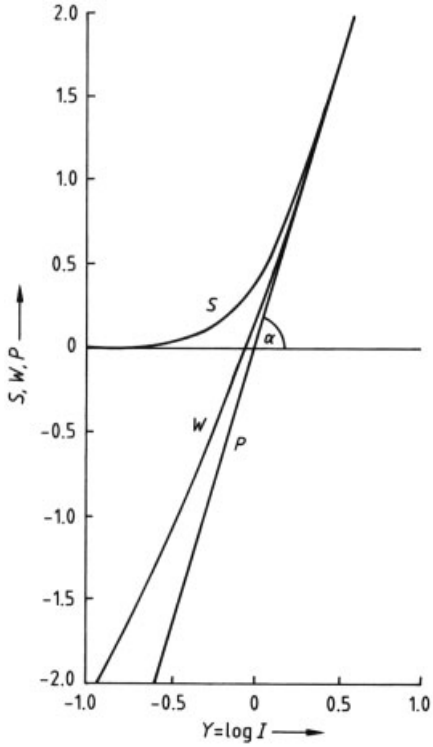

Figure 10. Emulsion characteristic (S) and transformation functions (P, W)

When a photographic emulsion is exposed to radiation blackening S occurs, which is a function of the radiant energy accumulated during the exposure time. It is given by:

$$S = \log(1/\tau) = -\log\tau = \log\varphi_0/\varphi \qquad (106)$$

where τ is the transmission, φ_0 is the flux through a nonirradiated part of the emulsion, and φ is the flux for an exposed part of the emulsion, when, in a densitometer, a beam of white light passes through the emulsion. The emulsion characteristic for a selected illumination time gives the relation between the logarithm of the intensity ($Y = \log I$) and the blackening, and is S-shaped (Fig. 10). Originally, only the linear part of the characteristic was used for quantitative work. However,techniques have been de-

veloped,which linearize the characteristic at low exposures. The P-transformation has been described by KAISER [37]:

$$P=\gamma\,(Y-Y_0)=S-\varkappa D \tag{107}$$

where $D=\log\,[1/(1-\tau)]$.

The Seidel function is:

$$W=\log\,(10^s-1) \tag{108}$$

and thus:

$$P=(1-\varkappa)\,S+\varkappa W \tag{109}$$

where Y_0 is the inertia (speed) of the emulsion and constants $\varkappa$ and γ describe the properties of the emulsion, but also depend on the densitometer. In most cases, Y_0 is unknown so ΔP or ΔY values are used. $\gamma=\tan\alpha$ is the contrast of the emulsion. The emulsion can be calibrated with the aid of a step filter placed in front of the spectrometer [38]. When the ratio of the intensities passing through two sectors of the filter is:

$$\Delta Y_m=\log\left(I_{m(1)}/I_{m(2)}\right) \tag{110}$$

the transformation equation becomes:

$$\Delta S=\gamma\Delta Y_m+\varkappa\Delta D \tag{111}$$

where ΔD is the independent variable and ΔS the function. A linear regression for a number of line pairs gives $\varkappa$ as slope and $\gamma\,\Delta Y_m$ as intercept on the x-axis. Values of $Y=\log I$ for a given blackening can be calculated:

$$Y=(\varkappa/\gamma)-\log\,(1-\tau)-(1/\gamma)\log\tau+Y_0 \tag{112}$$

and

$$\Delta Y=-\,(\varkappa/\gamma)\,\Delta D+(1/\gamma)\,\Delta S \tag{113}$$

Errors in the intensity measurement stem from the graininess of the emulsion. According to SIEDENTOPF [39], the standard deviation at low exposure is:

$$\sigma\,(S)=\sqrt{[0.5\,(a/F)\,S]} \tag{114}$$

where a is the surface area of a grain and F the area of the densitomer slit (e.g., 10 µm×1 mm). From $\sigma\,(S)$ it is possible to calculate the relative standard deviation of the intensities $\sigma_r\,(I)$:

$$\sigma_{\mathrm{r}}\,(I)=\ln 10\sigma\,(S)\,(1/\gamma\,\{1+[\varkappa/\,(10^s-1)]\}) \tag{115}$$

Typical values for γ are 1–2.5 and for $\varkappa$, 0.5–2. Nonsensitized emulsions can be used in the wavelength range 220–450 nm. At longer wavelengths green- or red-sensitized emulsions can be used. They have a large γ and accordingly a small dynamic range. At UV or vacuum UV wavelengths the gelatin of the emulsion absorbs significantly, and low-gelatin emulsions (which are rather insensitive) are used; alternatively, radiation can be transformed to longer wavelengths with the aid of a scintillator, such as sodium salicylate.

Photomultipliers [40] are most commonly used for precise measurements of radiant intensities. The photomultiplier has a photocathode and a series of dynodes. The radiation releases photoelectrons from the photocathode as a result of the Compton effect. These photoelectrons impact after acceleration on the dynodes so that a large number of secondary electrons is produced, which after successive action on a number of dynodes lead to a measurable photocurrent I_{a} at the anode. For a photon flux N_φ through the exit slit of the spectrometer and impacting on the photocathode, the flux of photoelectrons produced N_{e} is:

$$N_{\mathrm{e}}=N_\varphi\,\mathrm{QE}\,(\lambda) \tag{116}$$

where QE is the quantum efficiency (up to 20 %). The cathode current I_{c} is:

$$I_{\mathrm{c}}=N_{\mathrm{e}}\,\mathrm{e} \tag{117}$$

where $\mathrm{e}=1.9\times10^{-19}$ A · s and:

$$I_{\mathrm{a}}=I_{\mathrm{c}}\theta\,(V) \tag{118}$$

where $\theta\,(V)$ is the amplification factor (up to 10^6). The dark current I_{d} is the value of I_{a} measured when no radiation enters the photomultiplier (ca. 1 nA). The quantities QE, $\theta\,(V)$, and I_{d} are characteristics for a certain type of photomultiplier. The photocurrents are fed into a measurement system. First, I_{a} is fed to a preamplifier containing a high-ohmic resistor. The voltage produced is fed to an integrator. By selection of the resistors and capacitors, a high dynamic measurement range can be realized. When I_{a} is accumulated in a capacitor with capacitance

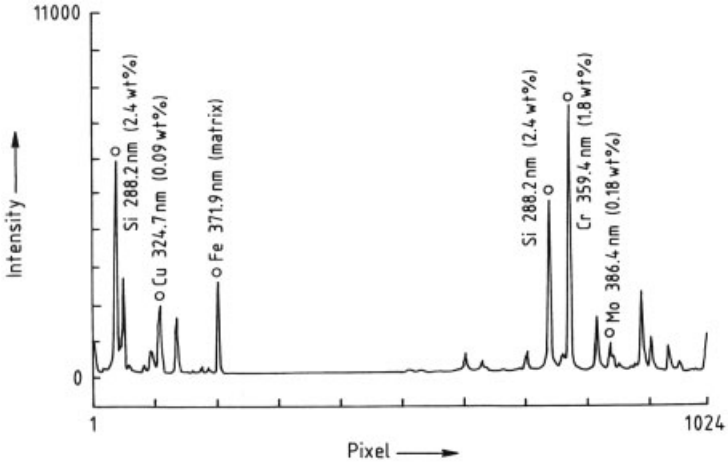

Figure 11. Segmented spectrum of glow discharge [43]
Grimm-type glow discharge lamp with floating restrictor 8 mm diameter, 50 mA, 3.5 torr, 2 kV [44], plasma-array spectrometer (LECO), steel sample 217 A

C, during an integration time t the potential obtained is:

$$U = I_a t / C \tag{119}$$

This voltage is digitized. The dynamic range is limited at the lower end by the dark current of the photomultiplier and at the high end by saturation at the last dynode.

For a photomultiplier, the relative standard deviation of the measured intensities $\sigma_r(I)$ is determined by the noise of the photoelectrons and the dark current electrons. The quantum efficiency and the amplification at the dynodes are also important:

$$\sigma_r(I) \approx 1.3 \left(N_{e,d}\right)^{-1/2} \tag{120}$$

where $N_{e,d}$ is the total number of photoelectrons and dark current electrons in measurement time t. Photomultipliers for the spectral range 160–500 nm usually have an S 20 photocathode; QE $(\lambda) \approx 5-25\,\%$, the amplification factors are 3–5 at dynode voltages 50–100 V, and there are 9–15 dynodes mounted in head-on or side-on geometry. For selected types, the dark current may be below 100 photoelectrons per second ($I_a < 10^{-10}$ A). Red-sensitive photomultipliers have a so-called bialkali photocathode. Their dark current is higher, but can be decreased by cooling. For wavelengths below 160 nm, photomultipliers with MgF_2 or LiF windows are used. Solar-blind photomultipliers are only sensitive for short-wavelength radiation (e.g., below 330 nm).

Solid-state detectors are increasingly important for optical atomic spectrometry. They are multichannel detectors and include vidicon, silicon-intensified target (SIT) vidicon, photodiode array detectors, and image dissector tubes [41]. Charge-coupled devices (CCD) and charge-injection devices (CID) [42] have been introduced. Photodiode arrays (PDA) may consist of matrices of up to 512 and even 1024 individual diodes (e.g., Reticon) with individual widths of 10 µm and heights of up to 20 mm. The charge induced by photoelectric effects gives rise to photocurrents which are sequentially fed into a preamplifier and a multichannel analyzer. They are rapid sequential devices, with the advantage that memory effects due to the incidence of high radiant densities on individual diodes (lag) as well as cross-talk between individual diodes (blooming) are low. Their sensitivity in the UV and vacuum UV is low compared with photomultipliers. Nevertheless, they are of interest for atomic emission spectrometry, especially as they can be coupled with microchannel plates, giving additional amplification. Signal-to-noise ratios can be considerably improved by cooling the array with the aid of a Peltier element or liquid nitrogen. Photodiode arrays have been used

successfully in a segmented echelle spectrometer. Here, spectral segments around the analytical lines are sorted out by primary masks subsequent to dispersion with an echelle grating, and are detected after a second dispersion on a diode array. More than ten analytical lines and their spectral background contributions can be measured simultaneously. This is shown for a glow discharge lamp in Figure 11 [43], but restrictions arise due to detector noise limitations at low wavelengths.

Image dissector tubes (Fig. 12) make use of an entrance aperture (d) behind the photocathode (a), by which photoelectrons from different locations on the photocathode can be scanned and measured after amplification in the dynode train, as in a conventional photomultiplier. Though used in combination with an echelle spectrometer with crossed dispersion for flexible rapid sequential analyses [45], the system has not been successful. This may be due to its limited cathode dimensions, but also to stability problems.

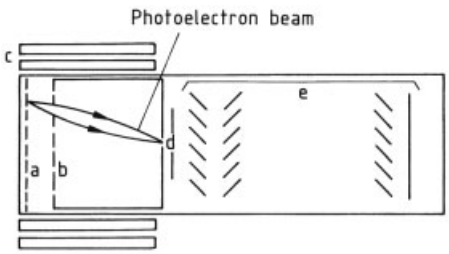

Figure 12. Image dissector tube
a) Photocathode; b) Accelerating electrode (mesh); c) Deflection coils; d) Aperture; e) Electron multiplier

Charge transfer devices (CTDs) are solid-state, silicon-based detectors which store charge as a result of photons impinging on the individual CTD elements, called pixels, and can be used to monitor large portions of the spectrum in multiple orders by taking an "electronic photograph". There are two main types, namely, charge injection device (CID) and charge-coupled device detectors (CCD). They differ in that the CID can be read randomly and nondestructively, whereas the CCD can only be read sequentially and destructively. A typical CCD detector [46] is the segmented array CCD used in the Perkin-Elmer Optima ICP-AES instrument. This consists of 224 linear photodetector arrays on a silicon chip with a surface area of 13 mm by 18 mm (Fig. 13). The array segments detect three to four analytical lines of high analytical sensitivity and large dynamic range and which are free from spectral interferences. Each subarray is comprised of pixels. The pixels are positioned on the detector at $x-y$ locations that correspond to the locations of the desired emission lines, usually generated by an echelle spectrometer. The emission lines are detected by means of their location on the chip, and more than one line can be measured simultaneously. The detector can then be electronically wiped clean and the next sample analyzed. A number of newer CCD detectors are capable of a pixel resolution of 0.009 nm and can record the continuous first-order spectrum from 175 to 800 nm, giving simultaneous access to 10 000 emission lines. Compared with many PMTs, a CCD offers improvement in quantum efficiency and lower dark current.

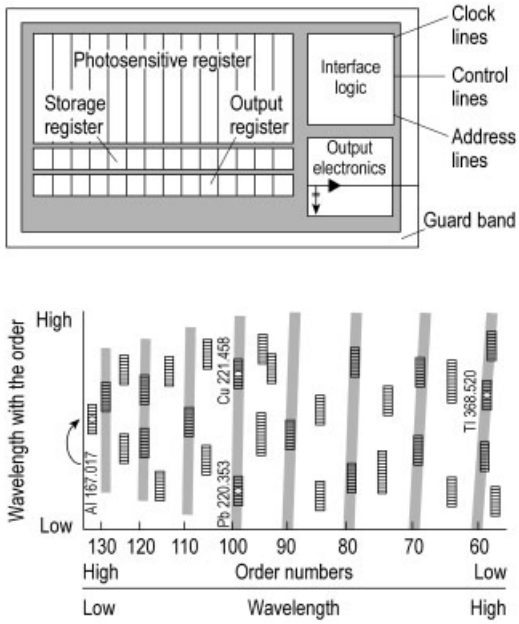

Figure 13. Schematic of a CCD detector (top) and a segmented two-dimensional detector array with respect to wavelength and order (bottom)

3.2.3. Nondispersive Spectrometers

Filter monochromators are of use only for flame photometry. They make use of interference filters, which often have a spectral bandpass of a few nanometers or less. Multiplex spectrome-

Table 3. Evaluation of detectors for atomic emission spectrometry

Detector	Dimensions, mm	Spectral region, nm	Sensitivity vs. photomultiplier
Vidicon [41]	12.5×12.5 [a]	with scintillator down to 200	poorer, especially at < 350 nm; dynamics $< 10^2$; lag/blooming
Diode array [47]	up to 25 (up to 1024 diodes)	especially for > 350	poorer in UV; dynamics 10^3
Diode array MCP [b] [41]	12.5	200 – 800	similar; linear; dynamic range $> 10^3$
Dissector tube [45]	up to 60 mm [a]	200 – 800	similar; linear; dynamic range $> 10^3$
CCD [42]	up to 13×18	200 – 800	similar; linear; dynamic range $> 10^5$

[a] Coupling with crossed-dispersion echelle spectrometer possible.
[b] MCP = Microchannel plate.

ters include Hadamard transform spectrometers and Fourier transform spectrometers, and are especially useful where very stable sources are needed. Hadamard transform instruments make use of a coding of the spectrum produced by recombining the information with the aid of a slit mask which scans the spectrum [48].

Fourier transform spectrometry [49] often uses a Michelson interferometer (Fig. 14) to produce the interferogram. With the aid of a beam splitter, the radiation is split into two parts which are each directed to a mirror. Shifting the mirror in one of the side arms, gives the interference for each wavelength:

$$I(x) = B(\sigma)\left[1 + \cos(2\pi\sigma x)\right] \tag{121}$$

where x is the optical path difference, σ is the wavenumber of the radiation, I is the intensity measured with the detector, and B is the radiation density of the source. A polychromatic radiation source gives an interferogram where the intensity of each point is the sum of all values resulting from Equation (121). The central part contains the low-resolution information and the ends contain high-resolution information. The resolution depends on the recording time, the spectral bandwidth, and the number of scans. By applying a Fourier transform to the signal for each point of the interferogram:

$$I(x) = \int_{-\infty}^{+\infty} B(\sigma)\,\mathrm{d}\sigma + \int_{-\infty}^{+\infty} B(\sigma)\cos(2\pi\sigma x)\,\mathrm{d}\sigma \tag{122}$$

$$= C + \int_{-\infty}^{+\infty} B(\sigma)\cos(2\pi\sigma x)\,\mathrm{d}\sigma \tag{123}$$

where C is a constant which must be subtracted before the transformation. The end result is:

$$B(\sigma) = \int_{-\infty}^{+\infty} I(x)\cos(2\pi\sigma x)\,\mathrm{d}x \tag{124}$$

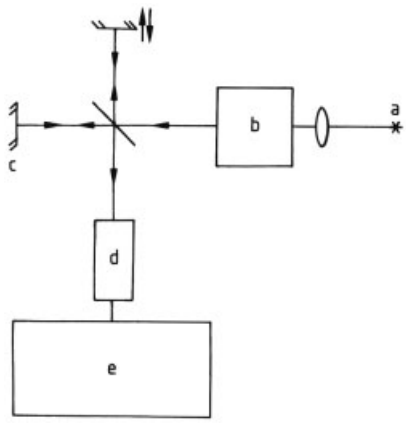

Figure 14. Fourier transform emission spectrometry
a) Source; b) Predispersion; c) Interferometer; d) Detector; e) FT in computer

By digitizing the interferogram, rapid Fourier transform is possible with a powerful computer. For complex spectra, this is possible by small, but highly accurate stepwise displacements of the side arm. Repetitive scanning intensifies the image, and a reference laser is used to make the mirror positioning reproducible. The technique is well known in infrared spectroscopy and can now be used in the visible and UV regions. Fourier transform atomic emission spectrometry is suitable for sources with a low radiance with a high stability detector to avoid detector noise. Wavelength calibration must be very

accurate to achieve maximum resolution. Reasonable signal-to-noise ratios are only achieved with low-noise sources.

3.3. Mass Spectrometers

In the sources used in optical atomic spectrometry, considerable ionization takes place, so they are useful ion sources for mass spectrometry [50]. Although, an overall treatment of instrumentation for mass spectrometry is given elsewhere (→ Mass Spectrometry), the most widely used types of mass spectrometer are briefly outlined here. In particular, the new types of elemental mass spectrometry have to be considered, i.e., glow discharges, and inductively coupled and microwave plasmas. In contrast to classical high voltage spark mass spectrometry [51] or thermionic mass spectrometry [52], which are beyond the scope of this article, the plasma sources are operated at considerably higher pressure between some 10 Pa and atmospheric pressure than the pressure in the mass spectrometer itself (10^{-3} Pa). Consequently, there has to be an interface with the appropriate apertures and high capacity vacuum pumps to bridge the pressure difference between source and spectrometer.

In commercial instrumentation both low-cost quadrupole and time-of-flight (TOF) mass spectrometers, and expensive double-focusing sector field mass spectrometers are used [53].

3.3.1. Types of Mass Spectrometer

In all mass spectrometers a vacuum of 10^{-3} Pa or better is maintained to avoid further ion formation from residual gas species or collisions of analyte ions with these species. Nowadays, turbomolecular pumps are preferred over diffusion pumps as their maintenance is easier and oil backflow does not occur.

Quadrupole Mass Spectrometers. In a quadrupole instrument (Fig. 15), the spectrometer consists of four equidistant, parallel rods (diameter 10 – 12 mm) between which a d.c. voltage and a high-frequency field (up to 1 MHz) are maintained. The d.c. voltage at the quadrupole should be slightly below the energy of the entering ions (usually below 30 eV). The ions enter through the so- called skimmer and have to pass the ion lenses and then a mask which prevents UV from entering the spectrometer (beam-stop). By changing the voltages at the ion lenses, at the beam-stop, at the quadrupole rods, and on the spectrometer walls, both the transmission of the spectrometer and the resolution of ions of a given mass and energy, can be optimized. As these parameters are all interdependent, optimization is complex. By changing the quadrupole field, the transmission for a certain ion changes. Accordingly, ions of a certain mass can be filtered out by manual setting of the field, or the spectrum can be scanned. The scanning velocity is usually limited to 30 000 mass numbers per second, because of the high-frequency and d.c. components in the spectrometer. Accordingly, the mass interval of 0 – 300 can be scanned in about 30 ms. Mass resolution is mainly determined by the quality of the field, and in quadrupoles for plasma mass spectrometry it is optimally about 1 atomic mass unit. Line profiles are approximately triangular, but side-wings occur as well; together with the isotopic abundances they determine the magnitude of spectral interference. Quadrupole mass spectrometers are rapid sequential measurement systems. In the case of isotopic dilution techniques for calibration, both precision and figures of merit can be limited by noise in the source. The price of quadrupole instruments is reasonable and their transmission is high. This is important because it allows their use in plasma mass spectrometry, where the source itself can be kept at earth potential. This is an advantage in inductively coupled plasmas.

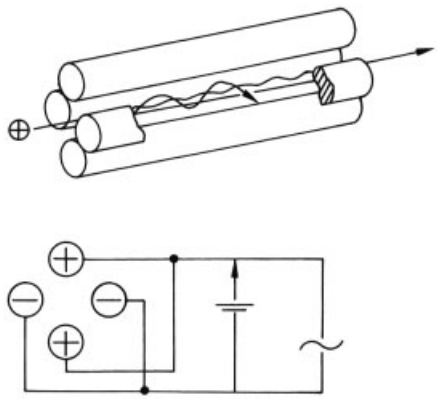

Figure 15. Quadrupole mass spectrometry

Time-of-Flight (TOF) Mass Spectrometers. In quadrupole and sector field mass analysers the ion signal is a continuous beam. In TOF mass spectrometry the ion beam is pulsed so that the ions are either formed or introduced into the analyzer in "packets". These ion packets are introduced into the field-free region of a flight tube 30 – 100 cm long. The principle behind TOF analysis is that, if all ions are accelerated to the same kinetic energy, each ion will acquire a characteristic velocity dependent on its *m*/*z* ratio [54]. The ion beam reaches its drift energy (2700 eV) in less then 2 cm. The ions are then accelerated down the TOF tube with whatever velocity they have acquired. Because all ions have essentially the same energy at this point, their velocities are inversely proportional to the square roots of their masses. As a result, ions of different mass travel down the flight tube at different speeds, thereby separating spatially along the flight tube with lighter, faster, ions reaching the detector before the heavier ions. Hence, the *m*/*z* ratio of an ion, and its transit time (t, in microseconds) through a flight distance (L, in cm) under an acceleration voltage (V) are given by

$$t = L\sqrt{\left[\left(\frac{m}{z}\right)\left(\frac{1}{2V}\right)\right]} \; or \; \frac{m}{z} = \frac{2Vt^2}{L^2} \tag{124a}$$

TOF mass analyzers are calibrated by using two ions of known mass, so exact values of L and V need not be known. The mass calibration is based on the conditions of the analyzer during the entire period of measurement.

While simple in theory, the TOF analyzer caused numerous problems when coupled with a plasma source such as an ICP. The spread in ion kinetic energies caused by the ion sampling process resulted in the ions' entering the field-free region of the flight tube at different angles [55–57]. Another difficulty with ICP-TOF-MS is that ions have to be introduced in "packets". This can be achieved, for example, by using an orthogonal interface with a pulsed repeller plate. Background noise can be reduced by using a combination of quadrupole ion optics and an energy discriminator before the detector.

Sector Field Mass Spectrometers. In a magnetic sector analyzer, ions are subjected to a magnetic field which causes them to be deflected along curved paths. The ions are introduced into the analyzer via a series of electrostatic slits which accelerate and focus them into a dense ion beam. The velocity of the ions is controlled by the potential V applied to the slits. As the ions enter the magnetic sector analyzer they are subjected to a magnetic field parallel to the slits but perpendicular to the ion beam. This causes the ions to deviate from their initial path and curve in a circular fashion. A stable, controllable magnetic field H separates the components of the beam according to momentum. This causes the ion beam to separate spatially, and each ion has a unique radius of curvature, or trajectory R, according to its *m*/*z*. Only ions of a single *m*/*z* value will posess the correct trajectory that focuses the ion on the exit slit to the detector. By changing the magnetic field strength, ions with differing *m*/*z* values are brought to focus at the detector slit.

The ion velocity v in the magnetic field is given by:

$$\frac{1}{2}mv^2 = zV \; or \; v = \sqrt{\frac{2zV}{m}} \tag{124b}$$

As the ions enter the magnetic field, they are subjected to a magnetic force at right angles to both the magnetic lines of force and their line of flight. This leads to a centrifugal force and hence to curvature of the ion beam.

$$\frac{mv^2}{R} = Hzv \tag{124c}$$

The radius of curvature of the flight path is proportional to its momentum and inversely proportional to the magnetic field strength:

$$R = \frac{mv}{zH} \tag{124d}$$

By eliminating the velocity term (v) between Equations (124b) and (124d) we get

$$R = \frac{1}{H}\sqrt{2V\left(\frac{m}{z}\right)} \tag{124e}$$

Hence, ions accelerated through an electrostatic field, uniform in nature, and then subjected to a uniform magnetic field have different radii of curvature. This leads to ions of a specific *m*/*z* value being focused onto the detector slit while all other ions hit the side of the analyzer. Thus, the magnetic field classifies and separates ions

into a spectrum of beams with each part of the spectrum having a different m/z ratio, where:

$$\frac{m}{z} = \frac{H^2 R^2}{2V} \tag{124f}$$

To obtain a complete mass spectrum from a magnetic sector analyzer, either the accelerating voltage V or the magnetic field strength H is varied. Each m/z ion from light to heavy is sequentially focused on the detector, and hence a mass spectrum is obtained.

Single-focusing magnetic sectors, as described above, have the disadvantage that ion energies vary depending on their point of formation in the ion source. The difference in ion energy is accentuated by the accelerating voltage, which leads to peak broadening and low resolution in the single-focusing mass analyser.

Double focusing magnetic/electrostatic sector instruments use magnetic and electrical fields to disperse ions according to their momentum and translational energy (Fig. 16). An ion entering the electrostatic field travels in a circular path of radius R such that the electrostatic force acting on it balances the centrifugal force. The equation of motion or transmission is:

$$\frac{mv^2}{R} = Ez \tag{124g}$$

where E is the electrostatic field strength. Hence the radius of curvature of the ion path in the electrostatic sector is dependent on its energy and not its mass [58]. A narrow slit placed in the image plane of the electrostatic sector can be used to transmit a narrow band of ion energies. If this sort of analyzer were placed in front of a magnetic sector analyzer an increase in resolution would result, but a decrease in detection would be inevitable due to the decrease in ions exiting the electrostatic sector in comparison with ions exiting the ion source. This loss in sensitivity can be compensated for by the choice of a suitable combination of electrostatic and magnetic sectors, such that the velocity dispersion is equal and opposite in the two analyzers. The narrow ion energy range transmitted the full length of a sector instrument would suggest a loss in sensitivity, compared with quadrupole mass analyzers when both are operated at the same resolution, but this has not been observed in practice. This has been mainly attributed to the ion-focusing requirements for the different analyzers. Double-focusing mass analyzers require a high energy (3 – 8 kV) ion beam for effective ion transmission and resolution. However, if such a high energy beam was focused down a quadrupole, the ions would have obtained a velocity profile too great to be affected by the hyperbolic energy fields of the quadrupole analyzer, which generally requires ion energies of less than 10 eV for effective mass separation. The higher energy ion beam is less affected by space – charge interference, which causes scattering of the ion beam, and this has led to increased sensitivity of double-focusing instruments.

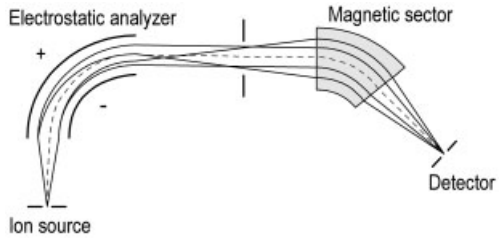

Figure 16. Sector field mass spectrometer

Other Mass Spectrometers. Spectrometers such as ion traps and ion cyclotron resonance have also been used in atomic mass spectrometry, but have not yet been incorporated into a commercial instrument. The former instrument has the advantage of ion storage and ion – molecule reaction capability, while the latter instrument is capable of extremely high resolution.

3.3.2. Ion Detection

For ion detection, several approaches are possible. Classical spark source mass spectrometry has developed the use of photographic plates. Very hard emulsions with low gelatin content are required. Emulsion calibration is similar to that described in optical emission spectrography, but usually varying exposure time are applied instead of using optical step filters. Provided an automated microdensitometer is used, mass spectrography is still a useful tool for survey analysis of solids down to the sub-µg/g level.

Faraday cup detectors consist of a collector electrode which is surrounded by a cage. The electrode is positioned at an angle with respect to the ion beam, so that ions exiting the analyzer strike the electrode, but secondary emissions are reflected away from the detector entrance. The function of the cage is to prevent detected ions and secondary electrons escaping from the detector. The electrode is connected to ground via a resistor. The ion current striking the electrode is neutralized by electron flow from ground through the resistor. This causes a potential drop which is amplified to create a signal. Currents as low as 10^{-15} A have been successfully measured in this way.

Electron Multipliers. For currents of less then 10^{-15} A an electron multiplier is necessary for detection. When the ion beam exits the analyzer it strikes a conversion plate which converts ions into electrons. The ions are drawn towards the plate by a strong voltage applied to the conversion plate. On striking the conversion plate the ions stimulate the ejection of electrons, which are accelerated by the voltage applied to the plate. The electrons are multiplied by continuous or discrete dynodes. A continuous dynode channel electron multiplier is shown in Figure 17. This consists of a curved glass tube of approximately 1 mm in internal diameter with an inner resistive coating and a flared end. The multiplier can be operated in one of two modes. In pulse counting mode – the most sensitive mode of operation – a high voltage of between -2600 and -3500 V is applied to the multiplier, which attracts ions into the funnel opening. When a positive ion strikes the inner coating the collision results in the ejection of one or more secondary electrons from the surface, which are accelerated down the tube by the potential gradient, and collide with the wall, resulting in further electron ejection. Hence, an exponential cascade of electrons rapidly builds up along the length of the tube, eventually reaching saturation towards the end of the tube, resulting in a large electron pulse and a consequent gain of $10^7 - 10^8$ over the original ion collision. The electron pulses are read at the base of the multiplier and are approximately 50 – 100 mV and 10 ns in duration. Continuous dynode multipliers consist of a leaded-glass tube which contains a series of metal oxides. The tube is curved and electrons are drawn down the tube by the potential gradient established by the resistivity of the glass. The tube is curved to stop electron feedback. A $10^5 - 10^7$ increase in signal is expected; however, the major constricting factor is the background noise of the system. Alternatively, the multiplier can be operated in analogue mode with a gain of only $10^3 - 10^4$ so that the multiplier does not become saturated and the pulses vary greatly in size. In this mode the applied voltage is between -500 and -1500 V and the electron pulses are read at the collector electrode where they are amplified and averaged over a short time interval to allow rapid data acquisition. The greatest sensitivity is achieved with the detector in pulse-counting mode, but the detector becomes saturated at counting rates above 10^6 Hz, which are encountered when the analyte is at a high concentration in the sample. If the detector is switched to analogue mode it is less sensitive, but can be used for analyte concentrations which are much higher, typically up to three orders of magnitude higher than for pulse counting. Such dual-mode operation results in an extremely large linear dynamic range of up to nine orders of magnitude. A discrete dynode detector consists of an array of discrete dynode multipliers, usually containing 15 – 18 dynodes, coated with a metal oxide that has high secondary electron emission properties. The dynodes are placed in one of two configurations, either venetian blind or box and grid fashion. Secondary electrons emitted by the metal oxide are forced to follow a circular path, by a magnetic field, so they strike successive dynodes, thereby multiplying the signal.

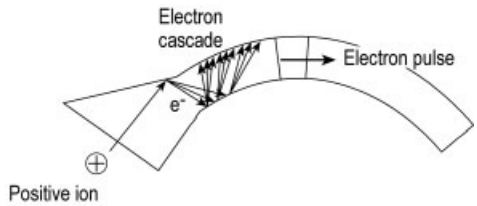

Figure 17. Continuous dynode electron multiplier

3.3.3. Ion Extraction

An important aspect of plasma mass spectrometry is the sampling interface. Ions are physi-

cally extracted from the plasma into a mass spectrometer, which is required to be at extremely low pressure, so the sampling interface must be in direct contact with the plasma (usually an ICP). The problem of extracting ions from an extremely hot plasma at atmospheric pressure into a mass spectrometer at ca. 10^{-4} Pa (10^{-9} atm) is overcome by making use of a series of differentially pumped vacuum chambers held at consecutively lower pressures. A schematic diagram of the ICP – MS sampling interface is shown in Figure 18. The ICP is aligned so that the central channel is axial with the tip of a water-cooled, sampling cone, made of nickel or copper, which has an orifice of approximately 1 mm in diameter. The pressure behind the sampling cone is reduced, by means of a vacuum pump, to ca. 200 Pa (2×10^{-3} bar) so the plasma gases, together with the analyte ions, expand through the sampling orifice to form a shock-wave structure. This expansion is isentropic (i.e., no change in the total entropy) and adiabatic (i.e., there is no transfer of energy as heat), resulting in a supersonic expansion accompanied by a fall in temperature. This supersonic expansion takes the form of a cone with a shock-wave stucture at its base called a Mach disk. The region within the expansion cone is called the "zone of silence" which is representative of the ion species to be found in the ICP, i.e., the ionization conditions have been "frozen". The skimmer cone is another metal cone, the tip of which has an orifice of approximately 0.7 mm in diameter, that protrudes into the "zone of silence", and is axially in line with the sampling orifice. The ions from the zone of silence pass through the orifice in the skimmer cone, into a second intermediate vacuum chamber held at $< 10^{-2}$ Pa (10^{-7} bar), as an ion beam. The ion beam can then be focused by means of a series of ion lenses, which deflect the ions along a narrow path and focus them onto the entrance to the mass analyzer.

3.4. Data Acquisition and Processing

Modern optical and mass spectrometers are controlled by computers, responsible for wavelength or ion mass presetting, parameter selection, as well as control of the radiation or ion source and the safety systems. The software for data acquisition and processing may comprise routines for the calculation of calibration functions with the aid of linear or higher order regression procedures, for calculation of sample concentrations by standard addition techniques or by calibration with synthetic samples, for calculation of statistical errors, and drift correction. Computers also offer graphical display of parts of the spectrum, or superimposition of spectra of several samples for recognition of spectral interference or background correction.

In most cases, a linear calibration function can be used (see Section 3.1). If c_i is the concentration of a standard sample and x_i is the radiation intensity, absorption, or ion current for the element to be determined, the linear regression has the form:

$$x_i = a c_i \tag{125}$$

where a is the sensitivity, i.e., the slope of the calibration curve. To overcome source fluctua-tions and for maximum precision, it may be useful to plot signal ratios as the analytical signal:

$$x_i / R = a c_i \tag{126}$$

where R is the signal obtained for a reference element, which may be the matrix element, a major component, or a foreign element added in known concentration to all samples to be analyzed. This approach is known as internal standardization. If the calibration function is nonlinear, it may be described by a second- or third-degree polynomial:

$$x_i / R = a_0 + a_1 c_i + a_2 c_i^2 \tag{127}$$

or

$$x_i / R = a_0 + a_1 c_i + a_2 c_i^2 + a_3 c_i^3 \tag{128}$$

The coefficients a are calculated by regression procedures. Alternatively, different linear calibration functions may be used for different concentration ranges.

Interference or matrix effects occur when the analytical signals depend not only on the analyte element concentrations, but also on the concentrations of other sample components. Additive interference may be avoided by estimation of the magnitude of the interfering signal from a scan of the spectral background in the vicinity of the analytical line. Mathematical interference correction expresses the net analytical signal as:

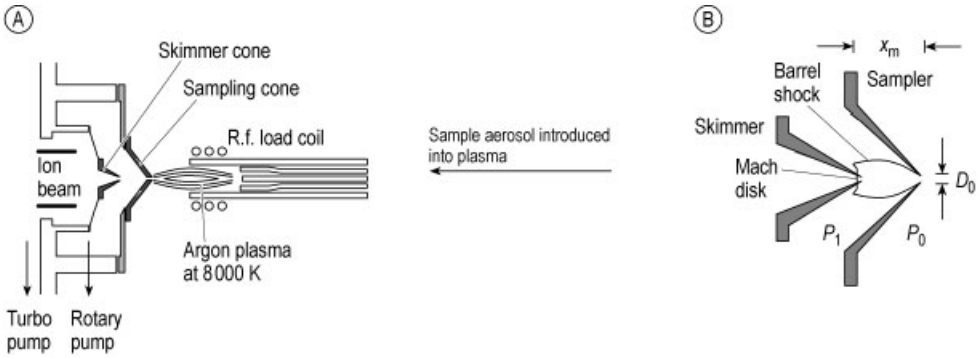

Figure 18. Ion extraction interface for ICP – MS
A) orientation of ICP and interface; B) schematic of sampler and skimmer cones showing barrel shock region
$D_0 \approx 1.0$ mm, $P_0 \approx 1 \times 10^3$ Pa, $P_1 \approx 2 \times 10^2$ Pa , X_m = 7 – 15 mm (depending on operating conditions)

$$x_i/R = (x_i/R)_{\text{measured}} - [\mathrm{d}\,(x_i/R)\,/\mathrm{d}c_{\text{interferent}}]\,x c_{\text{interferent}} \tag{129}$$

where $c_{\text{interferent}}$ must be determined separately and $\mathrm{d}(x_i/R)/\mathrm{d}c_{\text{interferent}}$ is the interference coefficient for a certain sample constituent, determined from a separate series of measurements. One can also express the calibration function as:

$$x_i/R = a_1 c_1 + a_2 c_2 \tag{130}$$

where a_1 and a_2 are determined by regression, which is straightforward for a linear relationship. Certain matrix constituents may produce signal enhancement or reduction. These are multiplicative in nature and can be due to influences on sampling efficiency, analyte transport into the source, or generation of the species delivering the analytical signals. They can be written as:

$$Y = a\,(c_{\text{interferent}})\,c \tag{131}$$

where the sensitivity a is a function of the concentration of the interferent $c_{\text{interferent}}$. This relation can be linear or of higher order. Calibration by standard addition allows correction of errors arising from signal reduction or enhancement insofar as the spectral background is fully compensated. When calibrating by standard addition, the concentration of the unknown sample can be determined graphically. This is done by extrapolating to zero the curve of the signals from samples to which known amounts of analyte have been added plotted against the added amounts. Standard addition can be very useful in atomic absorption and plasma mass spectrometry, which are zero-background methods. However, for atomic emission, where the lines are superimposed on a spectral background from the atomic emission source itself, and which is highly dependent on the sample matrix, it is more difficult. Here, calibration by standard addition can be used only when a highly accurate background correction is applied. It is important that the curve remains within the linear dynamic range of the method. This can be problematic atomic absorption spectrometry.

The precision obtainable in standard addition methods is optimum when the added amounts of analyte are of the same order of magnitude as the analyte concentrations present in the sample. An estimate of the analytical error can be made from the propagation of the errors of the measured intensities [64]. If X is a function of n measured values, each resulting from N replicate measurements, the standard deviation of the estimate $s\,(X)$ can be calculated from their estimates x_k ($k = 1, 2, 3, \ldots, n$) and standard deviations s_k:

$$s^2\,(X) = \sum_{k=1}^{n} [(\delta X/\delta x_k)\,s_k]^2 \tag{132}$$

For one standard addition [65], the intensities I_x and $I_{x+\Delta}$ for the unknown sample c_x and the standard addition sample $c_{x+\Delta}$ are:

$$c_x = a I_x \tag{133}$$

$$c_x + \Delta c = a I_{x+\Delta} \tag{134}$$

and

$$c_x = \Delta c I_x / (I_{x+\Delta} - I_x) \quad (135)$$

The standard deviation is obtained by applying Equation (132) to c_x. If I_x and $I_{x+\Delta}$ scatter independently and the added concentration Δc is free of error [66]:

$$s^2(c_x) = \{c_x [I_{x+\Delta}/I_x] [s(I_x)] / [I_{x+\Delta} - I_x]\}^2 + \{c_x [s(I_{x+\Delta})] / [I_{x+\Delta} - I_x]\}^2 \quad (136)$$

or

$$s_r^2(c_x) = [I_{x+\Delta}^2 / (I_{x+\Delta} - I_x)] \cdot [s_r^2(I_x) + s_r^2(I_{x+\Delta})] \quad (137)$$

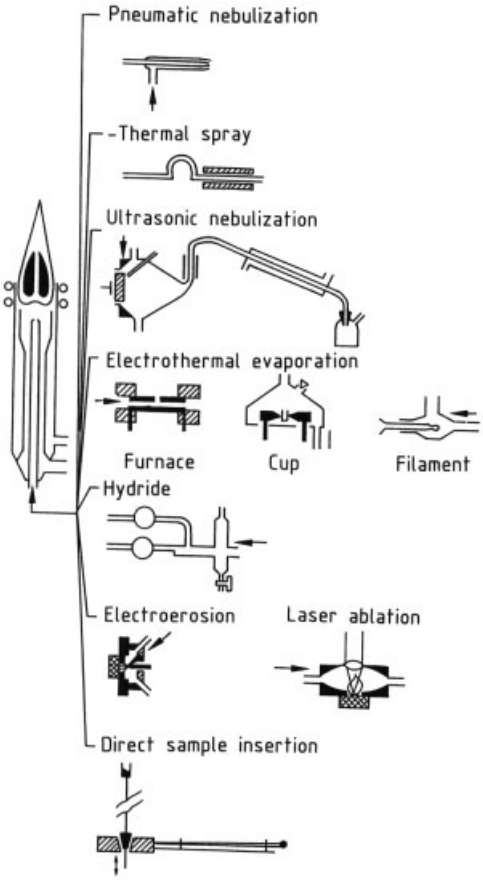

Figure 19. Sample introduction in plasma spectrometry [72]

4. Sample Introduction Devices

In atomic spectrometry, the sample must be converted to aerosol form and atomized before excitation. Accordingly, sample introduction is a very important step in all atomic spectrometric methods. The approach varies for different types of sample (size, state of aggregation stability, etc.) [67–71].

If sample volatilization and signal generation take place in a single source, both thermal evaporation and cathodic sputtering may occur. The latter is especially important in low-pressure discharges, as the ions can reach high energies when passing through the high-energy zones of the discharge and energy losses are small, because of the low number of collisions. In plasma sources at atmospheric pressure or flames, aerosol generation is often performed in a separate device. Figure 19 shows a number of sampling techniques for the analysis of solutions and solids. The techniques used for plasma spectrometry [73] and for sampling by cathodic sputtering, as performed in low-pressure discharges, are treated in detail in what follows.

4.1. Pneumatic Nebulization

For analysis of solutions, pneumatic nebulization is well known from early work on flame emission spectrometry. Pneumatic nebulizers must produce a fine aerosol with a gas flow, low enough to transport the aerosol droplets through the source with low velocity; this is essential for complete atomization. Nebulization of liquids is based on the viscosity drag forces of a gas flow passing over a liquid surface, producing small, isolated droplets. This may occur when the liquid is passed through a capillary tube, with the exit gas flows concentric around the tube, or perpendicular to the liquid stream. It may also occur at a frit, which is continuously wetted and through which the gas passes. Nebulizers are mounted in a nebulization chamber, which separates off the larger particles. Impact surfaces are often provided, fractionate the aerosol droplets further, and increase the aerosol production efficiency. The liquid flow is typically a few milliliters per minute and the maximum efficiency (a few percent) can be achieved at liquid flows of 2 mL/min; even below this flow rate, droplet diameters of the order of 10 µm and injection velocities of the order

of 10 m/s are obtained. A number of nebulizer types are used. For flame emission, atomic absorption, and for plasma spectrometry, they include concentric nebulizers, cross-flow nebulizers, Babington nebulizers, and fritted-disk nebulizers (Fig. 20) [68].

Concentric nebulizers date back more than a century, to GOUY [75], and in principle are self-aspirating. They can be made of metal, plastic, or glass. In atomic absorption spectrometry, Pt – Ir capillaries are often used to allow the aspiration of highly acid solutions. In plasma atomic spectrometry, concentric glass nebulizers (Meinhard) [74] are well known. They use a rather lower gas flow (1 – 2 L/min) than atomic absorption types (up to 5 L/min). In both cases, aspiration rates are 1 – 2 mL/min and aerosol efficiencies 2 – 5 %. Direct injection nebulizers are a useful alternative for the introduction of small-volume samples [76].

Cross-flow nebulizers are less sensitive to salt deposits at the nozzle. They are used in plasma spectrometry and flame work, and require a peristaltic pump for sample feeding. Cross-flow nebulizer capillaries can also be made of metal, glass, or plastic and have similar characteristics to those of concentric nebulizers. Materials such as poly(*p*-phenylene sulfide) (Ryton) enable aspiration of solutions containing HF.

In a *Babington nebulizer*, the sample solution is pumped through a wide tube and over a small gas orifice, so slurries can be analyzed without the risk of clogging.

Fritted-disk nebulizers have high nebulization efficiencies and low sample consumption (down to 0.1 – 0.5 mL/min). They are especially useful for plasma spectrometry of organic solutions, or for coupling with liquid chromatography [77]. However, in comparison with other types of nebulizer they suffer from memory effects; rinsing times are much higher than the 10 – 20 s required for the other pneumatic nebulizers.

With pneumatic nebulizers, samples can be taken up continuously. After a few seconds of aspiration, a continuous signal is obtained, and can be integrated over a certain time. Alternatively, small sample aliquots can be injected directly into the nebulizer [78–81] or into a continuous flow of liquid. The latter is known as *flow injection analysis* (*FIA*) [82]. The signals in both cases are transient; they have a sharp rise and some tailing, resulting from displacement of the aerosol cloud at the signal onset and dilution of the aerosol at its end. Signals last for a few seconds and with sample volumes 10 – 20 µL are capable of giving sufficient signal intensities. These techniques enable analysis of small samples (down to some 10 µL), reduce salt deposits, and maximize the absolute power of detection. FIA are also amenable to automatation and on-line sample preparation steps, such as preconcentration and matrix removal by coupling to column chromatography [83–85].

Optimization of pneumatic nebulizers is concentrated on selection of the working conditions for optimum droplet size and efficiency. The so- called Sauter diameter d_0, i.e., the diameter for which the volume-to-surface ratio equals that of the complete aerosol, is given by the Nukuyama – Tanasawa equation [86]:

$$d_0 = (C/v_{\mathrm{G}})\,(\sigma/\varrho)^{0.5} + C'\left\{\eta/\left[(\sigma\varrho)^{0.5}\right]\right\}^{C''} \cdot \left[1000\,(Q_{\mathrm{L}}/Q_{\mathrm{G}})\right]^{C'''} \tag{138}$$

where v_{G} is the gas velocity, Q_{G} the gas flow, Q_{L} the liquid flow, η the viscosity, ϱ the density, σ the surface tension of the liquid, and C, C', and C'' and C''' are constants. When the nebulizer gas flow increases, d_0 becomes smaller, the sample introduction efficiency increases, and so do the signals. However, as more gas is blown through the source, it is cooled, and the residence time of the droplets decreases, so that atomization, excitation, and ionization also decrease. These facts counteract increase of the signals as a result of the improved sampling; maximum signal intensity and power of detection are achieved at a compromise gas flow.

The physical properties of the sample liquid (σ, η, ϱ) also influence d_0, the efficiency, and the analytical signals. Differences in the physical properties of the sample liquids thus lead to so- called nebulization effects. Viscosity has a large influence through the second term in Equation (138). However, in free sample aspiration, it also influences d_0 through Q_{L}, which is given by the Poiseuille law:

$$Q_{\mathrm{L}} = \left[\left(\pi R_{\mathrm{L}}^4\right)/(8\eta l)\right]\Delta P \tag{139}$$

where R_{L} is the diameter and l the length of the capillary. ΔP is the pressure difference:

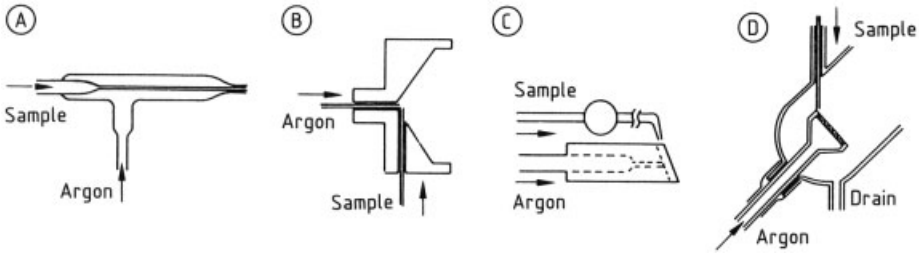

Figure 20. Pneumatic nebulizers for plasma spectrometry [68]
A) Concentric glass nebulizer [74]; B) Cross-flow nebulizer; C) Babington nebulizer; D) Fritted-disk nebulizer

$$\Delta P = \Delta P_{\mathrm{U}} - gh\varrho - f(\sigma) \tag{140}$$

where ΔP_{U} is the difference between the pressure inside and outside the nebulizer, h is the height and $f(\sigma)$ a correction factor. In order to minimize nebulization effects, forced feeding with a peristaltic pump can be applied, so that Q_{L} is no longer a function of η. The nebulizer can also be operated at high gas flow, so that Equation (138) reduces to:

$$d_0 = C/v_{\mathrm{G}}\,(\sigma/\varrho)^{0.5} \tag{141}$$

when $Q_{\mathrm{G}}/Q_{\mathrm{L}} > 5000$. However, the gas flow may cool the source, so that incomplete atomization, or signal reduction/enhancement may lead to even stronger interference. Aerosol droplet size distributions can be determined by stray-light measurements [87], [88]. Data obtained from sampling with cascade impactors may suffer from evaporation of the droplets on their way through the device, especially as sampling may require the application of underpressure [89]. Data for different nebulizers are available [90]. Particle size distributions depend on the nebulizer and its working conditions, on the liquid nebulized, and on the nebulization chamber. These relations have been studied in depth for organic solutions [91], [92], for which, at low gas flow, pneumatic nebulizers often produce an aerosol with a lower mean particle size compared with aqueous solutions, and thus a higher efficiency and power of detection.

Apart from pneumatic nebulization by gas flow over a liquid surface, other methods have been developed. It is possible to use the impact of a high pressure jet of liquid on a surface (*jet impact nebulization*) [93], or of the expansion of a liquid jet under high pressure at a nozzle (*high-pressure nebulization*), as described by Berndt [94] (Fig. 21). Much higher nebulization efficiencies can be obtained, even for solutions with high salt content or oils. This is due to the small droplet size, ca. 1 – 10 µm [95]. The system is ideally suited for speciation work as it is the appropriate interface for HPLC and atomic spectrometry [96]. A further advantage lies in the fact that the aerosol properties are independent of the physical properties of the liquid. In plasma spectrometry, however, it is necessary to desolvate the aerosol, as high water loadings cool the plasma too much and decrease the analytical signals; they may also lead to increased matrix effects. Desolvation can be done by heating the aerosol followed by solvent condensation, as applied in atomic absorption by Syti [97]. New techniques involve the use of Peltier elements for cooling in aerosol desolvation. High-pressure nebulization is very promising for on-line preenrichement or separation by high-performance liquid chromatography.

4.2. Ultrasonic Nebulization

By the interaction of sufficiently energetic acoustic waves of suitable frequency with the surface of a liquid, vortices are formed which produce an aerosol. The diameter of the aerosol droplets depends on the frequency and the physical properties of the liquid. For water, aerosol droplets formed at a frequency of 1 MHz have a diameter of ca. 4 µm. The energy can be focused with the aid of a liquid lens on the surface of the sample solution, or the sample liquid can be pumped continuously over the transducer, which must be cooled efficiently (Fig. 22).

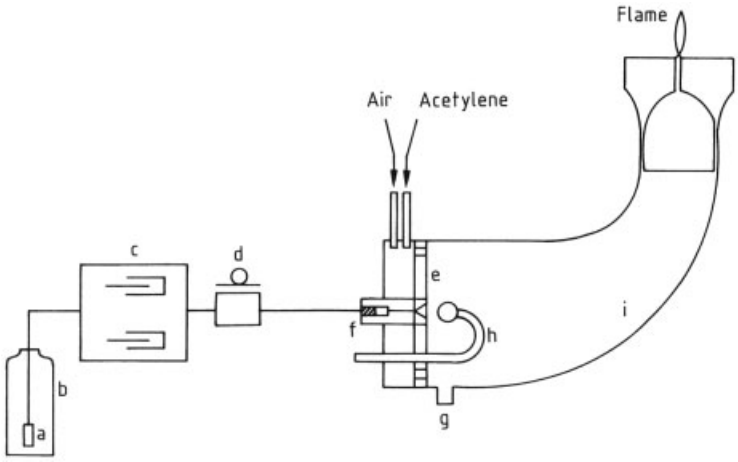

Figure 21. Hydraulic high-pressure nebulization system [94]
a) Solvent filter (20 µm); b) Solution reservoir (water, organic solvents;c) High-pressure pump; d) Sample feed valve/sample loop; e) Plate with concentric holes; f) High-pressure nebulization nozzle (10 – 30 µm) with integrated protection filter; g) Drain; h) Impact bead; i) Gas mixing chamber

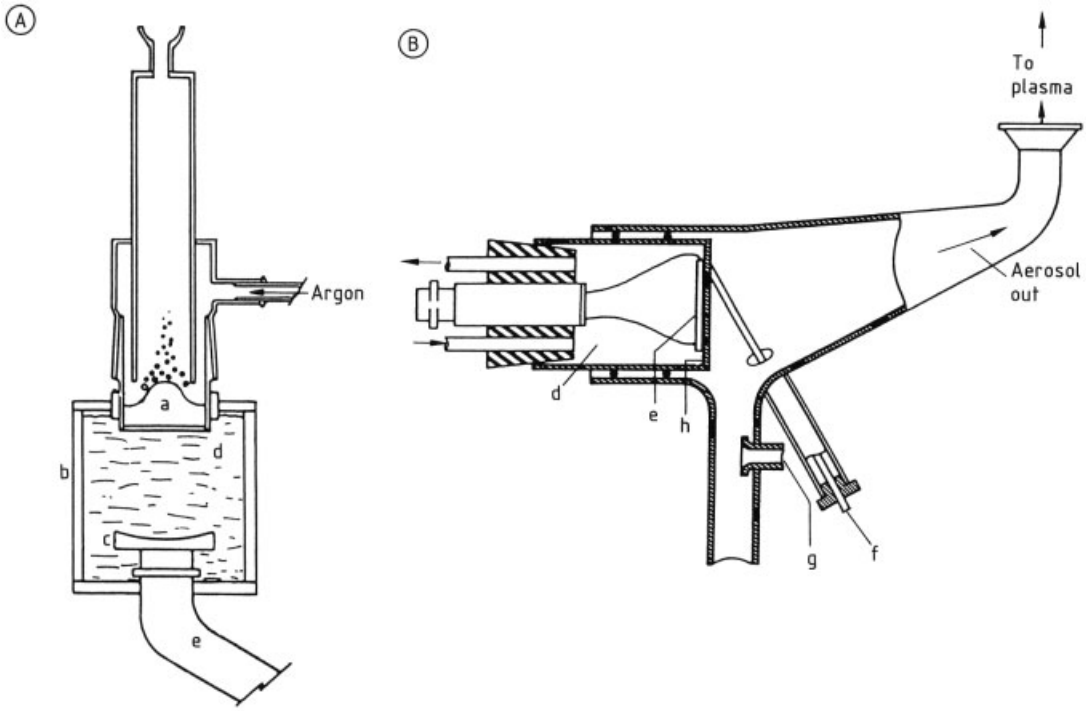

Figure 22. Ultrasonic nebulization
A) Discontinuous system [231]; B) System with liquid flowing over the transducer [99]
a) Sample solution; b) Focusing chamber; c) Focusing lens; d) Cooling water; e) Ultrasonic transducer; f) Sample introduction; g) Gas inlet; h) Transducer plate

Ultrasonic nebulization has two advantages over pneumatic nebulization. The aerosol particles have a lower diameter and a narrower particle size distribution compared with pneumatic nebulization (< 5 compared with $10-25\,\mu m$). Therefore, aerosol production efficiency may be up to 30 %, and analyte introduction efficiency is high. No gas flow is required for aerosol production, the transport gas flow can be selected completely freely. However, when applying ultrasonic nebulization to plasma spectrometry, it is necessary to desolvate the aerosol, to prevent too intensive cooling of the plasma. After this measure, ultrasonic nebulization leads to an increase in the power of detection.

Memory effects in ultrasonic nebulization are generally higher than in the case of pneumatic nebulization. The nebulization of solutions with high salt concentrations may lead to salt deposition; special attention has to be paid to rinsing, and precision is generally lower than in pneumatic nebulization. Its state of the art is summarized in [98].

4.3. Hydride Generation

For elements with volatile hydrides, sampling efficiency can be increased by volatilizing the hydrides. This applies to elements As, Se, Sb, Te, Bi, Sn, and some others. By in-situ generation of the hydrides of these elements from sample solutions, sampling efficiency can be increased from a few percent in the case of pneumatic nebulization to virtually 100 %.

Hydride generation can be efficiently performed by reduction with nascent hydrogen, produced by the reaction of zinc or $NaBH_4$ with dilute acids. In the latter case, the sample can be brought into contact with a solid pellet of $NaBH_4$, which may prove useful as a microtechnique [100]. Alternatively, a stream of $NaBH_4$ solution stabilized with NaOH can be added to an acidified sample. This can be done in a reaction vessel, as in atomic absorption work [101], or in a flow cell, as in plasma atomic emission spectrometry [102], [103]. The hydrides produced are separated from the liquid and subsequently transported by a carrier gas into the source. The hydrides are accompanied by excess hydrogen. In weak sources, such as microwave discharges, this may disturb discharge stability. To avoid this the hydrides can be separated off, e.g., by freezing them out and sweeping them into the source by subsequent heating [104]. Permselective membranes may also be useful [105].

Hydride generation increases the power of detection of atomic spectrometric methods for the determination of certain elements, and allows their matrix-free determination. However, the technique is prone to a number of systematic errors. First, the hydride-forming elements must be present as inorganic compounds in a well-defined valence state. This may require sample decomposition prior to analysis. In water analysis, treatment with H_2SO_4/H_2O_2 may be effective [106]. Traces of heavy metals such as Cu^{2+} may have a catalytic influence on the formation and dissociation of the hydrides, as investigated by Welz et al. [107] in atomic absorption with quartz cuvettes. These interferents can be masked by complexation with tartaric acid or coprecipitated with $La(OH)_3$. Calibration by standard addition is advisable.

Mercury compounds can also be reduced to metallic mercury with nascent hydrogen or Sn^{II}. The released Hg can be transferred to an absorption cell or plasma source. It can also be trapped on a gold substrate, from which it can be swept into the source after release by heating. This approach allows effective preconcentration and, in the case of optical and mass spectrometry, leads to very sensitive determinations of Hg.

4.4. Electrothermal Evaporation

Thermal evaporation of analyte elements from the sample has long been used in atomic spectrometry. In 1940, Preuss [108] evaporated volatile elements from a geological sample in a tube furnace and transported the released vapors into an arc source. Thermal evaporation has also been used in double arc systems, where selective volatilization gives many advantages for direct solid analysis. Electrothermal evaporation became especially important with the work of L'vov [3] and Massmann [4], who introduced electrothermally heated systems for the determination of trace elements in dry solution residues by atomic absorption spectrometry of the vapor cloud.

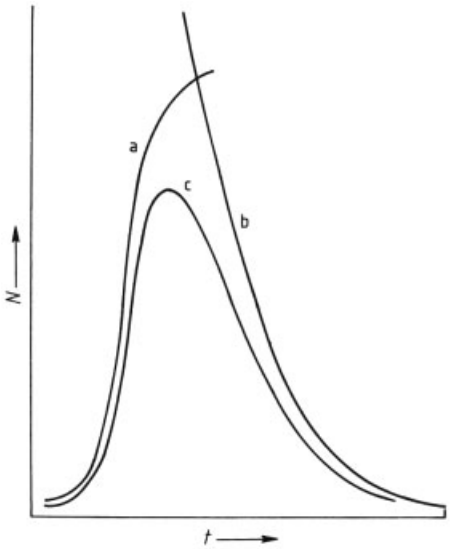

Figure 23. Signal form in graphite furnace atomic absorption spectrometry
a) Supply function; b) Removal function; c) Overall response function

4.4.1. The Volatilization Process

Electrothermal evaporation can be performed with dry solution residues as well as with solids. In both cases, the analyte evaporates and the vapor is held for a long time inside the atomizer, from which it diffuses. The high analyte concentration in the atomizer results from a formation and a decay function. The formation function is related to the production of the vapor cloud. After matrix decomposition, elements are present in the furnace as salts (nitrates, sulfates, etc.). They dissociate into oxides as a result of the temperature increase. In a device made of carbon (graphite), the oxides are reduced in the furnace:

$$MO_{(s/l)} + C_{(s)} \rightleftharpoons M_{(g)} + CO_{(g)}$$

However, a number of metals tend to form carbides, which can be very stable (thermodynamic control) or refractory (kinetic control). In this case, no analyte is released into the vapor phase. Decay of the vapor cloud is influenced by several processes:

1) Diffusion of the sample liquid into the graphite, which can often be prevented by the use of tubes, pyrolytically coated with carbon
2) Diffusion of the sample vapor
3) Expansion of the hot gases during the temperature increase (often at a rate of more than 1000 K/s)
4) Recombination processes, minimized by transferring the sample into the electrothermal device on a carrier platform with low heat capacity [109]
5) Action of purging gases

Therefore, in electrothermal devices, transient signals are obtained. They have a sharp increase and a more or less exponential decay, lasting 1 – 2 s. Their form has been studied for volatilization processes [110]. The signal form (Fig. 23) is also influenced by adsorption and resorption of analyte inside the electrothermal device.

4.4.2. Types of Electrothermal Device (Fig.24)

Graphite Furnaces. In most cases, the furnace is made of graphite, which has good thermal and corrosion resistance [3], [111]. As a result of its porosity, it can take up the sample without formation of appreciable salt deposits at the surface. Atomizers made of refractory metals such as tungsten have been used [112], [113].

Graphite tubes with internal diameters of some 4 mm, wall thickness 1 mm, and a length up to 30 – 40 mm are mostly used. Filaments enabling the analysis of very small sample volumes and mini-cups of graphite held between two graphite blocks have also been described [111]. When graphite tube furnaces are used for optical spectrometry, the optical path coincides with the axis of the tubes, whereas in filaments and cups, the radiation passes above the atomizer. Furnace currents up to 100 A and potentials of a few volts are used, giving temperatures of 3000 K or more. The tube is shielded by an argon flow, assuring stable working conditions for up to several hundreds of firings. With filaments, a lower power is used and the system can be placed in a quartz enclosure to keep out the air. Cups and tubes are nowadays often made of pyrographite, which prevents the analyte solution entering the graphite. Accordingly, chromatographic effects leading to selective volatilization are avoided, and the formation of refractory carbides, which

hampers the volatilization of elements such as Ti or V, decreases.

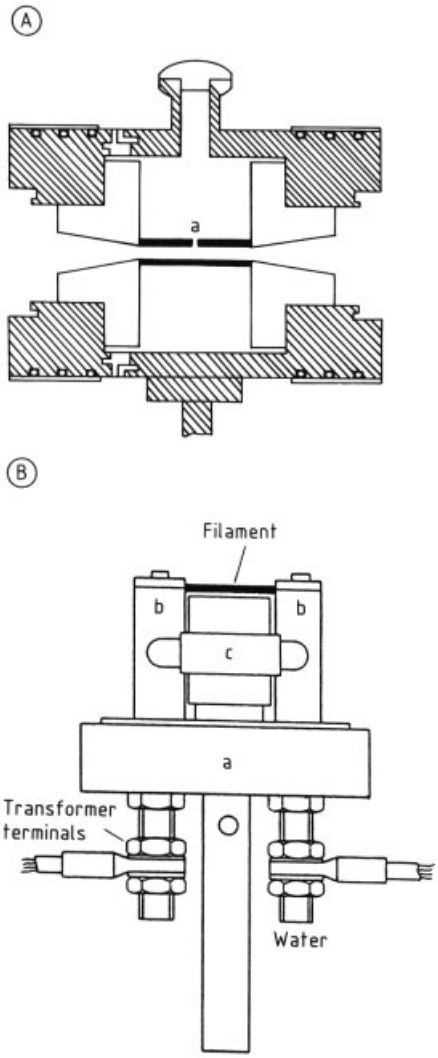

Figure 24. Graphite atomizers used in atomic spectrometry A) Original graphite tube furnace according to MASSMANN [4]; B) Carbon-rod atomizer system according to WEST [109]
a) Sampling hole; b) Heating blocks; c) Graphite rod or cup

Refractory Metal Furnaces. Electrothermal furnaces made of refractory metals (tungsten especially) have been described by SYCHRA et al. [112] for use in atomic absorption work. Their heat capacity is generally smaller that of graphite tubes, which results in steeper heating and cooling. This may be extreme in the case of tungsten probes and cups [113], which are mechanically more stable than graphite probes. The signals are extremely short and the analyte is released in very short times, which leads to high signal-to-background ratios and extremely low detection limits, as in wire loop atomization in atomic absorption spectrometry [114] or wire loops used for microsample volatilization in plasma spectrometry [115]. Therefore, metal devices used for electrothermal evaporation find it hard to cope with high analyte loading. Small amounts of sample are easily lost during heating, as the sample is completely at the surface and thermal effects cause stress in the salt crystals, formed during the drying phase. In graphite, the sample partly diffuses into the graphite, suppressing this effect. Graphite is further advantageous, as for many elements, reduction occurs, which leads to free element formation as required for atomic absorption spectrometry.

4.4.3. Temperature Programming

When electrothermal evaporation is used for sample introduction, the development of a suitable temperature program is of prime importance. In liquid samples, a small sample aliquot (10 – 50 µL) is transferred to the electrothermal device with a syringe or with the aid of an automated sampler and several steps are performed.

Drying. The solvent is evaporated and the vapor allowed to escape, e.g., through the sampling hole in the case of a graphite furnace. This step may last from tens of seconds to several minutes, and may take place at a temperature near the boiling point, e.g., 105 °C for aqueous solutions. This procedure often comprises several steps, advisable in the case of serum samples to avoid splashing.

Matrix Destruction. During this step, the matrix is decomposed and removed by volatilization. Chemical reactions are often used to enable volatilization of matrix constituents or their compounds. The temperature must be selected so that the matrix, but not the analyte, is removed. This is often achieved by applying several temperatures or even by gradually increasing the temperature at one or more ramping rates. Temperatures during this step are between

100 and 1000 °C, depending on the matrix to be removed and the analyte elements. Thermochemical reagents are often used to assist matrix destruction chemically and to achieve complete matrix removal at a lower temperature. Quaternary ammonium salts are often employed for the destruction of biological matrices at relatively low temperatures. Matrix decomposition is very important to avoid the presence of analyte in different chemical compounds, which can lead to transient peaks with several maxima.

Evaporation. The temperature selected for analyte evaporation depends strongly on the analyte elements and may range from 1000 K for volatile elements (e.g., Cd, Zn) to 3000 K for refractory elements (Fe). This step normally lasts no longer than 10 s, to maximize the number of firings which can be performed with a single tube.

Heating. The temperature is increased to its maximum (e.g., 3300 K over ca. 5 s with graphite tubes) to remove any sample residue from the evaporation device and to minimize memory and cross-over effects.

Direct solid sampling in electrothermal evaporation can be performed by dispensing an aliquot of a slurry prepared from the sample into the furnace. The analytical procedure is completely analogous to that with solutions. Powders can also be sampled with special dispensers, e.g., as described by Grobenski et al. [116]. They allow sampling of a few milligrams of powder reproducibly. Analyses with such small amounts make high demands on sample homogeneity. In direct powder sampling, the temperature program often starts with matrix decomposition and proceeds as for dry solution residues. In modern atomic spectrometers, sampling temperature programming, selection of gas flow, and visualization of the temperature program and signals are mostly controlled by computer.

The development of the temperature program is most important in the methodology of a spectrometric method using electrothermal evaporation. It should be documented completely in an analytical prescription for the determination of a given series of elements in well-defined samples.

4.4.4. Analytical Performance

Electrothermal atomization, because of its high analyte vapor generation efficiency (in theory 100 %), can yield extremely high absolute and relative detection power with any type of atomic spectrometry. In two-stage procedures, where the analyte vapor has to be transported into the signal generation source, diffusional losses of analyte vapor may occur. This has been described in detail for Cd [117], but is a general problem.

Owing to the transient nature of the analytical signals, precision is generally lower than with nebulization of liquids. Relative standard deviations are 3 – 5 % for manual injection of microaliquots, and 1 – 3 % for automated sample dispensation, whereas in pneumatic nebulization they are below 1 %.

Interference in electrothermal evaporation may stem from differences in the physical properties of the sample liquids from one sample to another. This may influence the wetting of the graphite or the metal of the electrothermal device. When the latter is temperature dependent, it leads to differences in volatilization. Differences in the anions present may severely influence evaporation (chemical matrix effect). The boiling points of the compounds formed dictate the volatilization. Accordingly, the occurrence of double peaks is easily understood. In the case of Cd, this is documented for a sample rich in chloride ($CdCl_2$ evaporates at 480 °C) and for a sample rich in nitrate (Cd volatilizes first at 900 °C and then at 1200 °C as CdO; Fig. 25) [66].

4.5. Direct Solid Sampling

These techniques are especially useful in two-stage procedures for solid analysis. A separate source is used for the generation of a sufficiently concentrated sample vapor cloud, the composition of which truly reflects the sample. Several approaches may be used according to sample properties.

4.5.1. Thermal Methods

For powders, electrothermal evaporation may be used. It is often hampered by sample inhomo-

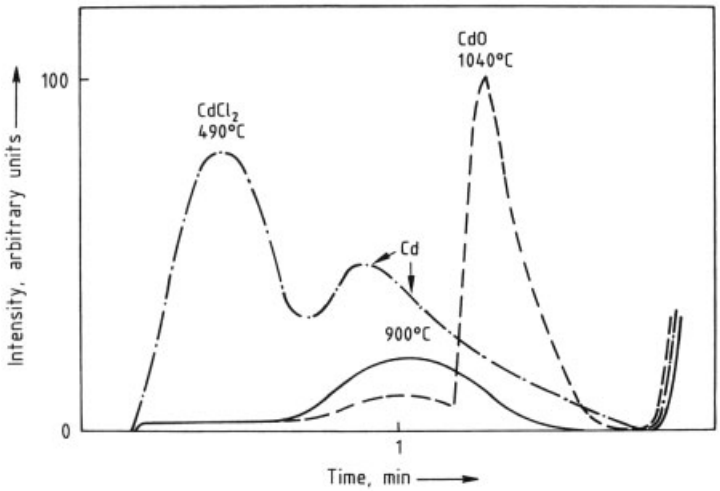

Figure 25. Thermochemical behavior of Cd and its compounds in graphite furnace evaporation [66]
ICP3kW argon – nitrogen; samples 10 µg/mL $Cd(NO_3)_2$ in aqueous solution (solid line), added to 0.1 g/mL of SRM 1571 (dashed line) and SRM 1577 (NaCl) (dash-dotted line); Cd II 226.5 nm line; sampling volume 50 µL; temperature program: drying 95 °C, gradual temperature increase from 350 – 2400 °C (600 °C/min)

geneity, as it is generally a micro method, as well as by anion effects, causing chemical interference [118]. Transport losses may also occur. These vanish when the sample is inserted directly in the signal generation source and evaporated thermally. This approach is known from work with graphite or metal probes in atomic absorption, where tungsten wire cups and loops are used. In spectrometry with inductively coupled plasmas, the technique is used both in atomic emission [119–121] and in mass spectrometry [122]. Its absolute power of detection is extremely high and the technique can be used both for the analysis of dry solution residues as well as for the volatilization of micro-amounts of solids.

In-situ thermal destruction of the matrix or selective volatilization can be applied. The latter has proved useful in geological samples [123]. Selective volatilization of volatile elements from refractory matrices is useful; spectral interference from matrices with complex spectra can be avoided. The approach can also be used for the volatilization of refractory oxide- or carbide-forming elements, which form volatile halogenides (e.g., Ti). Here, substances such as AgCl or polytetrafluoroethylene (PTFE) powder can be used as thermochemical reagents [121]. In the case of Ti in the presence of C and PTFE, the following reactions have to be considered:

$$TiO_2 + 2\,C \rightleftharpoons Ti + 2\,CO$$

$$Ti + C \rightleftharpoons TiC$$

$$Ti + 4\,F \rightleftharpoons TiF_4$$

At high temperature, the latter reaction is favored thermodynamically as well as kinetically (TiF_4 is volatile).

4.5.2. Slurry Atomization

Powders often form stable slurries in suitable dispersing fluids. The latter should be able to wet the powder and to form stable suspensions, both of which depend on their physical properties (viscosity, density, surface tension). Surface tension, which is generally a function of pH, is known to influence the surface charge of suspended particles and hence the stability of the slurry (for ZrO_2 powders see [124]). The stability of suspensions depends greatly on the particle size of the powder and its tendency to form agglomerates. The latter depends on the preparation method applied, as known from synthetic ceramic powders [125], [126]. The particle size should be in the low micrometer range at

most. The particle size distribution can be determined by laser diffractometry or by automated electron probe microanalysis [127], and is particularly critical when the aerosol is produced by pneumatic or ultrasonic nebulization. Measurements with powders of different particle size have shown that particles larger than 12 – 15 µm cannot be held in aerosol form with gas flows $\leq$ 2 L/min, as in the nebulizers used in plasma spectrometry [198]. In the other techniques of sample introduction, particle size is not so critical. This applies when slurries are analyzed by graphite furnace atomic absorption spectrometry. Volatile elements can be evaporated selectively from some types of sample, as shown for ceramic powders [128]. In any case, it is often necessary to apply ultrasonic stirring to destroy agglomerates and to disperse powders optimally, prior to slurry analyses. The addition of surface active substances such as glycols [129] has been proposed, but may introduce contamination when trace determinations are required.

In pneumatic nebulization, it is advisable to stir the sample continuously during slurry aspiration. A Babington nebulizer should be used, as other types are prone to clogging. The aerosol loaded with the solid particles, with or without a solvent layer, is transported to the source. Complete drying and evaporation on its way through the source is necessary for calibration by standard addition with aqueous solutions of the analyte. From knowledge of the temperature distribution in the source, the trajectory of the particles through the source, and the thermochemical properties of the sample material, it is possible to calculate the degree of volatilization for particles of a given size and composition at a certain point in the source, or the maximum size for particles to be completely volatilized [126], [130], [131].

Slurry analysis can also be applied to compact samples which are difficult to dissolve after pulverization. Attention must be paid to possible contamination resulting from abrasion in the mill. For mills of very hard materials such as tungsten carbide abrasion in the case of ceramics may amount to serveral percent of the sample mass. Slurry sampling has been reviewed comprehensively in [132].

4.5.3. Arc and Spark Ablation

Arc ablation has long been used for producing an aerosol at the surface of electrically conducting samples, in combination with various sources. In the version described by Johnes et al. [133], the metal sample acts as the cathode of a d.c. arc discharge. With an open circuit voltage of 600 V and currents of 2 – 8 A, a broad pulse spectrum (mean frequency up to 1 MHz) is observed and rapid movement of the discharge across the cathode produces uniform sampling over a well-defined area. A gas stream can then transport the aerosol particles to the signal generation source. Remote sampling at up to 20 m makes the technique attractive for the analysis of large samples. On the other hand, it is also useful for precision analyses, both by atomic absorption and plasma spectrometric methods. The analysis of electrically insulating powders is possible. A mixture of the powder to be analysed and copper powder in a ratio of about 1: 5 can be made and formed into a pellet at pressures up to 80 t/cm^2 (ca. 80 MPa) [134]. The mass ratio of analyte to copper may differ considerably from one sample to another.

Spark Ablation. When sparks are used, the ablation of electrically conducting solids is less dependent on matrix composition. This applies both to high-voltage sparks (up to 10 kV) and to medium-voltage sparks, as used for emission spectrometric analysis of metals. Voltages of 500 – 1000 V with repetition rates up to 1 kHz are used. Increase in voltage and spark repetition rate leads to a considerable increase of the ablation rate. For aluminum samples, trapping the ablated material in concentrated HNO_3 increases the ablation rate from 1 – 5 µg/min for a 25-Hz spark. However, the particle diameter also increases significantly hampering volatilization of the material in a high-temperature source such as ICP, and leading to flicker noise [135]. It is better to use a medium-voltage spark at a high sparking frequency [136]. After aerosol sampling on Nucleopore filters, X-ray fluorescence spectrometry shows that the composition of the aerosol produced in aluminum samples agrees with the composition of compacted samples [137]. The particle size in the case of 400-Hz spark is 1 – 2 µm and the particles are mostly spherical, which argues for their forma-

tion by condensation outside the spark channel. In the case of supereutectic Al – Si alloys ($c_{Si} > 11$ wt %) some elements (especially Si) are enriched in the smaller particles.

Spark ablation is very abrasive when performed under liquid. BARNES and MALMSTADT [138] used this effect to reduce errors stemming from sample inhomogeneity in classical spark emission spectrometry. The approach is also useful for the dissolution of refractory alloys, which are highly resistant to acids [139]. When sparking under liquid, ablation rates up to ca. 3 mg/min can easily be achieved. In high-alloy Ni – Cr steels, electron probe micrographs show that selective volatilization can become problematic. Very stable colloids can be formed when a ligand such as ethylenediaminetetraacetic acid is present in the liquid.

4.5.4. Laser Ablation (→ Laser Analytical Spectroscopy)

The interaction of the radiation from high-power laser sources with solids can lead to evaporation. As this is independent of the electrical conductivity of the sample, laser ablation is increasingly important for solid analysis [140–142].

Sources. Laser sources make use of population inversion. When radiation enters a medium, both absorption and stimulated emission of radiation may occur and the change in flux at the exit is:

$$\mathrm{d}F = \sigma F\,(N_2 - N_1)\,\mathrm{d}z \tag{142}$$

where dz is the length of the volume element in the z direction, σ is the cross section for stimulated emission or absorption and F is the flux. The sign of $(N_2 - N_1)$ is normally negative, as the population of the excited state N_2 is:

$$N_2 = N_1 \exp\left[-\left(E_2 - E_1\right)/kT\right] \tag{143}$$

and is smaller than the population of the ground state N_1. In the case of population inversion $N_2 > N_1$, the medium acts as an amplifier. When it is positioned between two mirrors, one of which is semitransparent, energy can leave the system. If R_1 and R_2 are the reflectances of the mirrors, the minimum theoretical population inversion is:

$$(N_2 - N_1)_{\mathrm{th}} = 1/\left(2\sigma l\right) \ln\left[1/\left(R_1 R_2\right)\right] \tag{144}$$

as the losses by reflection in a double pathway $2\,l$ are compensated. Population inversion requires the existence of a three- or four-level energy diagram. Laser radiation is monochromatic and coherent. The beam has a radiance – divergency relation α_{D} ($\approx \lambda/D$, where D is the beam diameter). The medium is in a resonator of which the length d determines the resonance frequency $\nu = n\,c/(2\,d)$, where n is the mode. Solid state lasers are especially of interest for laser ablation. The laser medium is a crystal or a glass, doped with a transition metal. The medium is pumped optically by flash-lamps (discharges of 100 – 1000 J over a few ms) or continuously with a tungsten – halogen lamp. The resonator may be the space between two flat mirrors, or an ellipsoid, with the laser rod at one focus and the flash-lamp at the other. The ruby laser emits in the visible region (694.3 nm). It is thermally very robust, but requires a high pumping energy and its energy conversion efficiency is low. The neodymium – yttrium aluminum garnet (Nd – YAG) laser is very widely used. Its wavelength is 1.06 µm and it has a much higher energy conversion. Gas lasers with CO_2 or Ar can be pumped electrically have a high power output, whereas dye and semiconductor lasers are mostly used for selective excitation.

Lasers can be operated in the Q-switch mode, so as to deliver their energy spikes very reproducibly. Optoacoustic or electroacoustic switches are often used. The first make use of a pressure-dependent change of refractive index of gases such as SF_6, which may be produced periodically at a suitable frequency (Boissel switch), whereas the latter are based on periodic changes of the transmittance of crystal in an a.c. electric field. As the interaction of laser radiation with solids highly depends on wavelength, frequency doubling, resulting from nonlinear effects, e.g., in $LiNbO_3$, is often used.

Interaction with Solids [142], [143]. When a laser beam with small divergence impinges on a solid surface, part of the energy (10 – 90 %) is absorbed and material evaporates locally. The energy required varies between ca. 10^4 W/cm^2 (for biological samples) and 10^9 W/cm^2 (for glasses). A crater is formed, of minimum diam-

eter determined by diffraction of the laser radiation:

$$d \approx 1.2 f \alpha_{\mathrm{D}} \nu$$

where f is the focal length of the lens used for focusing the laser radiation on the sample (5–50 mm, minimum dictated by the risk of material deposits) and α_{D} is the divergency (2–4 mrad). Typical crater diameters are of the order of 10 µm. They also depend on the energy and the Q-switch used. The depth of the crater relates to the laser characteristics (wavelength, radiant density, etc.) as well as to the properties of the sample (heat conductance, evaporation energy, reflectance, etc.). The ablated material is ejected from the surface at high velocity (up to 10^4 cm/s); it condenses and is volatilized again by absorption of radiation. In this way, a laser vapor cloud is formed of which the temperature and the optical density are high, and the expansion velocity, composition, and temperature depend on the laser parameters and the gas atmosphere and pressure. Favorable working conditions in laser ablation work involve reduced pressures (1–10 kPa) [144–146]. Optical and mass spectrometry can be performed directly on the laser vapor cloud, or the ablated material can be transferred to another source.

Analytical Performance. Laser ablation is a microsampling technique and thus enables local distribution analyses. Laterally resolved measurements can be made with a resolution of ca. 10 µm [144]. Sampling depth can be varied from 1 to ca. 10 µm by adjusting the laser. The amounts of material sampled can be of the order of 0.1–10 µg. As a result of the high signal generation efficiencies possible with mass spectrometry or laser spectroscopy, its absolute power of detection is very high. The laser sources now available allow high precision (relative standard deviations ca. 1 %), limited only by variations in reflectivity along the sample surface, and sample homogeneity.

4.6. Cathodic Sputtering

In low-pressure discharges, the atoms and ions undergo few collisions in the gas phase. Therefore, an electric field can induce in them the high energies required to remove material from solid lattices by impact and momentum transfer. Positive ions in particular can acquire the necessary energies; the phenomenon takes place when the sample is used as cathode and is known as cathodic sputtering. Its nature can be understood from the properties of low-pressure discharges. Models developed in physical studies reveal the analytical features of cathodic sputtering [147].

Low-Pressure Discharges. When a voltage is applied across two electrodes in a gas under reduced pressure, the ionization in the vicinity of the electrodes produces electrons, which may gain energy in the field and cause secondary ionization of the gas by collision. Secondary electron emission occurs when they impact on the electrodes. Field emission may take place near the electrodes (at field strengths above 10^7 V/cm) and, when the cathode is hot, glow emission can take place as well. In such a discharge, energy exchange takes place as a result of elastic and inelastic collisions, of which the latter may lead to ionization and recombination. With increasing voltage across the electrodes, a current builds up.

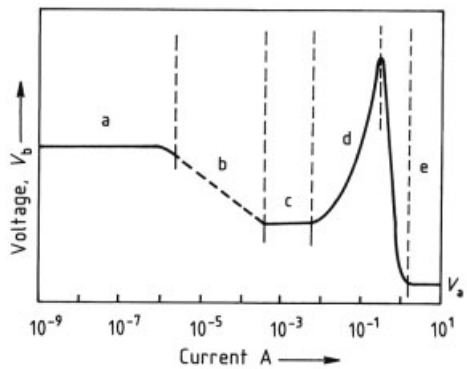

Figure 26. Current–voltage characteristic of a self-sustaining d.c. discharge
a) Townsend discharge; b) Corona range; c) Glow discharge with normal cathode fall; d) Glow discharge with abnormal cathode fall; e) Arc discharge
V_{a} = Arc voltage [148]; V_{b}=Breakdown voltage

The characteristic (Fig. 26) includes the region of the corona discharge and the normal region, where the discharge starts to spread over the electrodes. Here, the current can increase at practically constant voltage, whereas once the discharge covers the entire electrodes (restricted discharge), the current can only increase when

the voltage is increased drastically (abnormal part of the characteristic). Here, the burning voltage, the positive space charge, and the field gradient in front of the cathode are very high.

Once the cathode has been heated to a sufficiently high temperature, thermal emission may start and the characteristic enters the arc discharge region, where burning voltage and space charge decrease rapidly. Such a discharge displays a cathode dark space where the energies are too high for efficient collisions, the cathode layer where intensive emission takes place, a further dark space, and the negative glow where the negative space charge is high (Fig. 27). The plasma is not in local thermal equilibrium, as the number of collisions is too low to thermalize the plasma. Electron temperatures are high (5000 K for the electrons involved in recombination and $>10\,000$ K for the high-energy electrons responsible for excitation by electron impact), but the gas temperatures are low (<1000 K). In radio frequency (r.f.) discharges, which are increasingly important for atomic spectrochemical analysis, a barrier-layer and a bias voltage are built up in the vicinity of a metal electrode as a result of the different mobilities of electrons and ions in the r.f. field. Accordingly, the energy situation becomes very similar to that of d.c. discharges and similar sputtering phenomenona can occur.

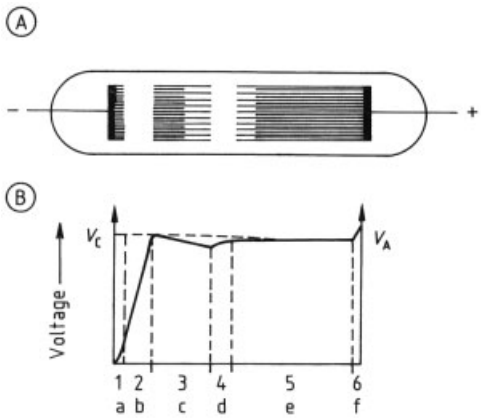

Figure 27. Geometry (A) and potential distribution (B) of a d.c. electrical discharge under reduced pressure
a) Aston dark space; b) Hittorf dark space; c) Negative glow; d) Faraday dark space; e) Positive column; f) Anode region
V_A = Anode fall; V_C = Cathode fall

Ablation. The models developed for cathodic sputtering start from ideal solids, i.e., single, perfect crystals; whereas real samples in atomic spectrochemical analyses are polycrystalline or even chemically heterogeneous. The available models are valid only for monoenergetic beams of neutrals impacting on the sample, whereas in fact both ions and atoms of widely different energies impact at different angles.

The ablation is characterized by a sputtering rate q (µg/s) and a penetration rate w (µm/s). The latter is the thickness of the layer removed per unit time and relates to the sputtering rate as:

$$w = \left(10^{-2} q\right) / (\varrho s) \tag{145}$$

where s is the target area (cm^2) and ϱ its density (g/cm^3). The sputtering yield S indicates the number of sputtered atoms per incident particle:

$$S = \left(10^{-6} q N_{\mathrm{A}} \mathrm{e}\right) / \left(M i^{+}\right) \tag{146}$$

where N_{A} is the Avogadro constant, e the charge on the electron (C), M the atomic mass, and i^{+} the ion current (A). Classical sputtering experiments with a monoenergetic ion beam in high vacuum show that the sputtering yield first increases with the mass of the incident ions and the pressure, but then decreases. For polycrystals, it is maximum at incident angle 30°. For single crystals, it is a maximum in the direction perpendicular to a densely packed plane. The results of these experiments can only be explained by the impulse theory. According to this theory a particle can be removed from a lattice site when the displacement energy E_{d} (the sum of the covalent or electrostatic binding energies) is delivered by momentum transfer from the incident particles. The maximum fraction of the energy transferrable from an incident particle is:

$$E_{\max}/E = 4mM/\left[(m+M)^2\right] \tag{147}$$

The ablation rate is thus proportional to the number of particles which deliver an energy equal to the displacement energy. It should, however, be taken into account that some incident particles are reflected (f_{r}) or adsorbed at the surface. Particles of low mass can penetrate into the lattice and be captured (f_{p}). Other particles enter the lattice and cause a number of collisions until their energy is below the displacement energy.

The overall sputtering yield, accounting for all processes mentioned, is:

$$S = [(\alpha E)/E_d]^{1/2} \; (f_p + A f_r)\, \varphi$$
$$\alpha = 2mM/(m+M)^2$$
and
$$\varphi = f(m, M) \tag{148}$$

Accordingly, cathodic sputtering increases with the energy of the incident particles and is inversely proportional to the displacement energy E_d. It is maximum when $m = M$. This explains why sputtering by analyte particles which diffuse back to the target is very efficient.

The dependence of sputtering yield on the orientation of the target with respect to the beam can be easily explained. In a single crystal, there is a focusing of momentum along an atom row in a direction of dense packing. If D_{hkl} is the lattice constant and d the smallest distance between atoms (or ions) during a collision, the angle θ_0, at which particles are displaced from their lattice sites relates to the angle θ_1, between the direction of the atom row and the line between the displaced atom and its nearest neighbor in the next row, according to the equation:

$$\theta_1 = \theta_0 \, (D_{hkl}/d - 1) \tag{149}$$

This focusing of momentum, referred to as Silsbee focusing, takes place when:

$$f = \theta_{n+1}/\theta_n$$

or

$$D_{hkl}/d < 2$$

Analytical Performance. The model described explains the features of cathodic sputtering as a technique for sample volatilization and is very helpful for optimizing the performance of cathodic sputtering sources [149]. When using sputtering for sample volatilization, it should be noted that some features can be realized only when working at sputtering equilibrium. On initiating a discharge, the burning voltage is normally high in order to break through the layer of oxides and of gases adsorbed at the electrode surface. When these species are sputtered off and the breakdown products are pumped away, the burning spot (crater) can start to penetrate with a constant velocity in the sample, so that the composition of the ablated material may become constant. The time required to reach sputtering equilibrium (burn-in time) depends on the nature and pretreatment of the sample as well as on the filler gas used and its pressure. All measures which increase ablation rate will shorten burn-in times.

The topography of the burning crater depends on the solid-state structure of the sample. It reflects the graininess, chemical homogeneity, and the degree of crystallinity. Inclusions and defects may locally disturb the sputtering. These effects can be observed on micrographs, comparing the craters obtained with a glow discharge and a spark (Fig. 28) [150]. The roughness of the burning crater can be measured with sensing probes; it imposes the ultimate limitation of the depth resolution obtainable when applying sputtering to study variations of sample composition with depth from the sample surface.

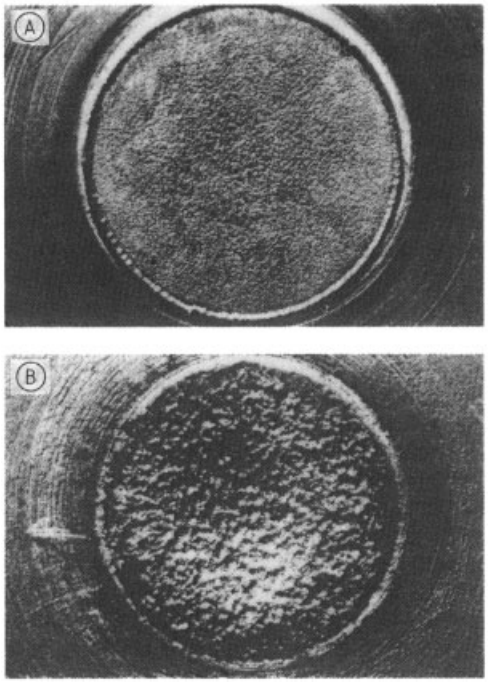

Figure 28. Burning spots obtained with a glow discharge (A) and a medium-voltage spark (B)
Sample: aluminum [150]

Achievable ablation rates depend on the sample composition, the discharge gas, and its pres-

sure. As filler gas, a noble gas is normally used. With nitrogen or oxygen, chemical reactions at the sample surface occur and disturb the sputtering, as electrically insulating oxide or nitride layers are formed. Reactions with the ablated material produce molecular species which emit molecular band spectra in optical atomic spectrometry, or produce cluster ion signals in mass spectrometry. In both cases, severe spectral interference can hamper the measurement of the analytical signals. The relation between ablation rates and sample composition can be understood from impulse theory. In most cases argon ($m = 40$) is used as sputtering gas, and the sequence of the ablation rates follows the mass sequence:

$$C < Al < Fe < steel < Cu < brass < Zn$$

Helium is not suitable, as its small mass renders its sputtering efficiency negligible. The sputtering rates further increase in the sequence

$$Ne < Ar < Kr < Xe$$

The last two gases are rarely used because of their price. Neon is attractive because of its high ionization potential; argon may cause spectral interference. Gas pressure has a very important influence on electrical characteristics. At low gas pressure the burning voltage is high, as is the energy of the incident particles. At high pressure, the number density of potential charge carriers is higher and the voltage decreases. The number of collisions also increases, so the energy of the incident particles decreases. The resulting decrease of sputtering rate with gas pressure for a glow discharge with planar cathode and abnormal characteristic [151], can be described:

$$q = c/\sqrt{p} \tag{150}$$

where c is a constant and p is the pressure. Many studies have been performed on volatilization by cathodic sputtering in analytical glow discharges used as sources for atomic emission, atomic absorption, and mass spectrometry [152]. Also, studies on the trajectories of the ablated material have been performed, as described, e.g., for a pin glow discharge [153].

5. Atomic Absorption Spectrometry

5.1. Principles

As a method for elemental determinations, atomic absorption spectrometry (AAS) goes back to the work of Walsh in the mid-1950s [2]. In AAS, the absorption of resonant radiation by ground state atoms of the analyte is used as the analytical signal. This process is highly selective as well as very sensitive, and AAS is a powerful method of analysis, used in most analytical laboratories. Its methodological aspects and applications are treated in several textbooks [154–156].

A primary source is used which emits the element-specific radiation. In the beginning, continuous sources were used with a high-resolution spectrometer to isolate the primary radiation. However, due to the low radiant densities of these sources, detector noise limitations occurred, or the spectral band width was too large for sufficiently high sensitivity.

Generally, a source which emits discrete lines was found more suitable, with low-cost monochromators to isolate the radiation. It was also necessary in early work to use sources in which the physical widths of the emitted analyte lines are low. This is necessary to get high absorbances (Fig. 29). When the band width of the primary radiation is low with respect to the absorption profile of the line, a higher absorption results from a given amount of analyte as compared with a broad primary signal. Narrow lines in the primary radiation are obtained with low-pressure discharges, as in hollow cathode lamps or low-pressure r.f. discharges. The availability of narrow-band, tunable laser sources, such as the diode lasers, has opened new perspectives [157], [158]. Only the analytical line is present and no monochromator is necessary. If tuning is applied, it is possible to scan across the absorption profile, to increase the dynamic range or to perform background correction. Switching from one analyte element to another becomes feasible, certainly when several diode lasers are provided. However, laser sources are restricted by the limited spectral range which can be covered – at the time of writing, only down to the green.

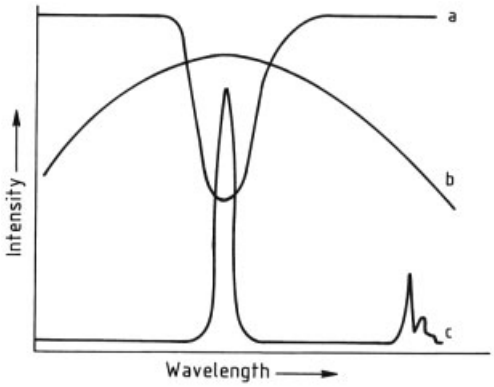

Figure 29. Importance of physical line widths in atomic absorption spectrometry
a) Absorption signal for elemental line; b) Spectral bandpass of monochromator; c) Emission of hollow cathode lamp

The use of line sources generally leads to high analytical selectivity, sensitivity, and power of detection as well as a high dynamic range within which the Lambert – Beer law (Eq. 38) applies. A linear relationship between absorption and concentration can only apply when all radiation passing to the detector is absorbed to the same extent by the analyte atoms (ideal case). In the normal case, the curve of concentration and absorption falls off toward the concentration axis, as a result of the presence of nonabsorbed radiation. For a primary source emitting narrow lines (widths below one-fifth of those of the absorption lines) this consists mainly of contributions from nonabsorbed lines of the cathode material or of the filler gas, which fall within the spectral bandwidth of the monochromator. At high concentrations, a decrease in dissociation gives rise to lower absorbances, and suppression of ionization leads to higher absorbances.

For AAS, the analyte must be present as an atomic vapor, i.e., an atomizer is required. Both flames and furnaces are used, and the corresponding methodologies are known as flame AAS and graphite furnace AAS. Special methods of atomization are based on volatile compound formation, as with the hydride technique (Sections 4.3 and 5.5). AAS is generally used for the analysis of liquids, so solids must first be dissolved. Wet chemical decomposition methods are of use, but involve all the care required in trace analysis to prevent losses or contamination, and the resultant systematic errors. For direct solid analysis, a few approaches exist and are discussed in Chapter 4.

5.2. Spectrometers

Atomic absorption spectrometers contain a primary source, an atomizer with its sample introduction system, and a monochromator with a suitable detection and data acquisition system (Fig. 30).

5.2.1. Spectrometer Details (Fig. 30)

Radiation from the primary source (a) is led through the absorption volume (b) and subsequently into the monochromator (c). As a rule, radiation densities are measured with a photomultiplier and processed electronically. Czerny – Turner or Ebert monochromators with low focal length (0.3 – 0.4 m) and moderate spectral bandpass (normally not below 0.1 nm) are frequently used.

In most instruments, the radiant flux is modulated periodically. This can be achieved by modulating the current of the primary source, or with the aid of a rotating mirror (d) in the radiation beam. Accordingly, it is easy to differentiate between the radiant density emitted by the primary source and that emitted by the flame by lock-in amplification. Both single- and dual-beam instruments are used. In the latter, part of the primary radiation is fed directly into the monochromator, while the rest passes first through the flame. In this way, fluctuations and drift can be compensated insofar as they originate from the primary radiation source or the measurement electronics.

The spectrometer can be equipped for quasi-simultaneous measurement of the line and background absorption [159]. Radiation from a second, continuous source, such as a deuterium arc lamp, is rapidly switched with the radiation of the primary source and fed through the absorption volume, by a semitransparent mirror. The radiant flux of the continuous source is not decreased significantly by atomic absorption, owing to the low spectral resolution of the monochromator; however, it will be weakened by broad-band absorption of molecules, or stray radiation. By this means, nonelement-specific background absorption can be compen-

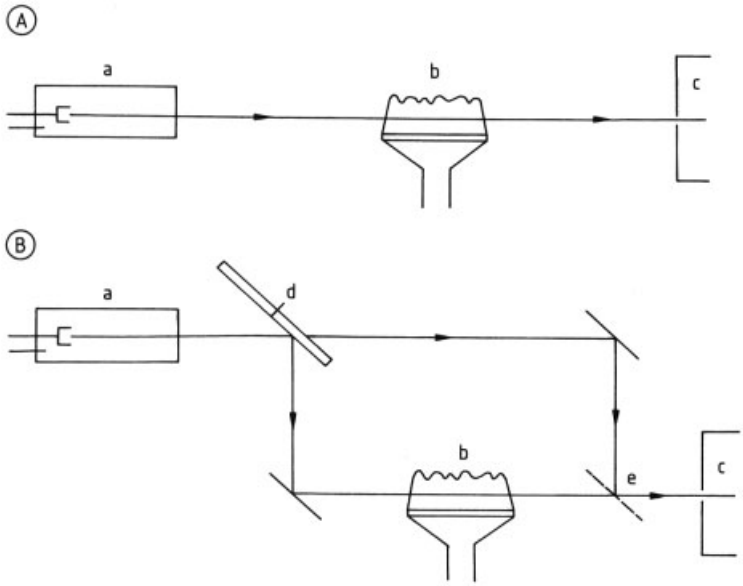

Figure 30. Flame atomic absorption spectrometer
A) Single beam; B) Dual-beam system
a) Source; b) Burner; c) Monochromator; d) Rotating mirror; e) Half-silvered mirror

sated. This and other approaches to background correction are discussed in detail in Section 5.6.

Dual-Channel Instruments. Apart from single-channel instruments, with which only measurements at a single wavelength in one channel can be performed, dual-channel instruments are also used. They contain two independent monochromators which enable simultaneous measurements at two wavelengths. They are of use for the simultaneous determination of two elements of which one element may be the analyte and the second a reference element. It is also possible to use two lines with widely different intensity to determine one element over a wide concentration range.

Multichannel spectrometers which allow the simultaneous determination of a large number of elements, as in atomic emission spectrometry have not encountered a breakthrough in AAS yet. However, over a number of years, work with high-intensity continuous sources and high-resolution echelle spectrometers for multielement AAS determinations has aroused some interest [160]. Fourier transform spectrometry and multichannel detection with photodiode arrays has opened new perspectives for the simultaneous detection of larger numbers of spectral lines, and high-intensity sources have improved considerably.

5.2.2. Primary Sources

The primary radiation sources used in AAS have to fulfill several conditions:

1) They should emit the line spectrum of one or more analyte elements with line widths smaller than those of the absorption lines in the atomizer
2) They should have high spectral radiance density (radiant power per unit area, solid angle, and unit wavelength) in the center of the spectral line
3) Their optical conductance (radiant area multiplied by usable solid angle) must be high
4) The radiances of the analytical lines must be constant over a long time

These conditions are fulfilled by low-pressure discharge such as hollow cathode discharges and, for some elements, high-frequency discharges.

In a commercially available hollow cathode lamp (Fig. 31), the cathode (a) has the form of a hollow cylinder closed at one side. The lamp is sealed and contains a noble gas at a pressure of the order of 100 Pa. At discharge currents up to 10 mA (at ca. 500 V), a glow discharge is sustained between this cathode and a remote anode (b). The atomic vapor is produced by cathodic sputtering and excited in the negative glow contained within the cathode cavity. Lines of the discharge gas are also emitted, which may lead to interference in AAS. In most cases, high-purity argon or neon are used. For mechanical reasons, it may be necessary to manufacture the cathode mantle from a material other than that of the internal part of the hollow cathode, which is usually made of the analyte, so a further atomic spectrum may be emitted.

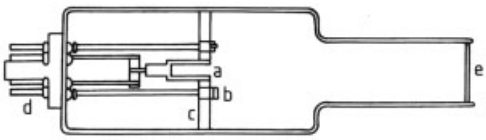

Figure 31. Hollow cathode source for atomic absorption spectrometry
a) Hollow cathode; b) Anode; c) Mica insulation; d) Current supply; e) Window (e.g., quartz)

For a number of elements, lamps in which the hollow cathode consists of several elements may also be used. The number of elements contained in one lamp is limited because of the risk of spectral interference.

For a few elements, electrodeless discharge lamps are to be preferred over hollow cathode lamps. This applies for volatile elements such as As, Se, and Te. In the hollow cathode lamps of these relatively volatile elements self-absorption at low discharge currents may be considerable, and self-reversal may even occur. This is not the case with electrodeless discharge lamps. The latter consist of a quartz balloon in which the halide of the element is present. The analyte spectra are excited with the aid of a high-frequency (MHz) or microwave field (GHz), which may be supplied through an external antenna.

5.3. Flame Atomic Absorption

In flame AAS, the sample solutions are usually nebulized pneumatically in a spray chamber and the aerosol produced is transferred, together with a mixture of a combustion gas and an oxidant, into a suitable burner.

5.3.1. Flames and Burners

Flames may use propane, acetylene, or hydrogen as combustion gases, with air or nitrous oxide as oxidant. With these mixtures, temperatures of 2000 – 3000 K can be obtained. Temperatures of flames (in kelvin) used in AAS are given below [157]:

Air – propane	1930
Air – acetylene	2300
Air – hydrogen	2045
Nitrous oxide – acetylene	2750
Oxygen – acetylene	3100

The *air – acetylene flame* is most often used. Its temperature is high enough to obtain sufficient atomization for most elements which can be determined by AAS, but not so high that ionization interference becomes significant. The latter occur only with the alkali and alkaline earth metals. The analytical conditions can be optimized by changing the composition of the gas mixture. The sensitivity for the noble metals (Pt, Ir, Au, etc.) can be improved by increasing the oxidative properties of the flame (excess air), which also lowers interference. The use of a reducing flame is more advantageous for determination of the alkali metals.

For more than 30 elements, however, the temperature of this flame is too low. This applies for elements which have very stable oxides, such as V, Ti, Zr, Ta, and the lanthanides. For these elements, *nitrous oxide – acetylene flame*, as introduced in flame AAS by Amos and Willis in 1966 [161], is more suitable. However, this flame has a higher background emission, especially due to the radiation of the CN bands.

The *hydrogen – air flame* is of litte use. For the determination of Sn, however, its sensitivity is higher than that of the acetylene – air flame. Because of its higher transmission at short UV wavelengths the hydrogen – air flame is also used for the determination of As (193.7 nm) and Se (196.0 nm).

The *propane – air flame* has the lowest temperature and is used for the determination of easily atomized elements such as Cd, Cu, Pb, Ag, Zn, and especially the alkali metals.

In flame work, the burners have one or more parallel slits of length 5 – 10 cm, providing an absorption path for the primary radiation. In special laminar burners, the density of the analyte atoms at relatively high observation positions (up to more than 5 cm) can be kept high, which is advantageous as measurements are then possible in these zones where more complete atomization occurs. Multislit burners may be useful to prevent deposits in solutions with high salt content. High-temperature flames require the use of special burners. Detailed information on flames is available in classical textbooks [162–164].

5.3.2. Nebulizers

The nebulizer is mounted in the spray-chamber burner assembly, as shown for the nitrous oxide or air – acetylene burner system (Fig. 32). Air or nitrous oxide are fed to the mixing chamber through the nebulizer. The combustion gas is fed directly to the burner. A self-aspirating concentric nebulizer is often used, but other types, including the Babington and cross-flow types (Section 4.1) are also applied. Nebulizers may be made of glass, corrosion-resistant metals such as Pt – Ir, or plastics such as Ryton. Sample solution consumption usually amounts to ca. 5 mL/min in the case of gas flows of 1 – 5 L/min.

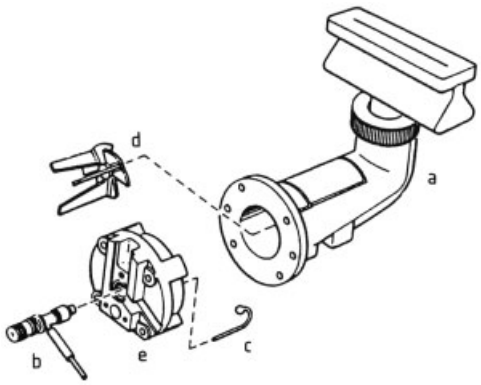

Figure 32. Burner – nebulizer assembly for flame atomic absorption spectrometry
a) Burner head with mixing chamber; b) Nebulizer; c) Impactor bead; d) Impact surfaces; e) Nebulizer socket

In flame AAS, the nebulization system must be able to generate small droplets (<10 µm), as only such droplets can be transported and completely vaporized in the flame. This is a prime condition for the realization of high sensitivity and low volatilization interference. Nebulizers for AAS generate not only small droplets, but also larger ones (up to 25 µm or more). For fragmentation of the latter, the nebulizer should be positioned in a mixing chamber provided with impact surfaces. In pneumatic nebulization, the sampling efficiency remains restricted (a few percent) as fragmentation is limited, and all particles larger than 10 µm are led to waste. In flame AAS with pneumatic nebulization, the sample solution is usually aspirated continuously. After 5 –7 s, a stable signal is obtained and signal fluctuations depending on concentration are of the order of 1 %.

Apart from continuous sample aspiration, flow injection and discrete sampling can be applied (see Section 4.1), which both deliver transient signals. In the latter, 10 – 50 µL aliquots can be injected manually or with a sample dispenser into the nebulization system [78], [79]. This approach is especially useful for preventing clogging in the case of sample solutions with high salt content, for the analysis of microsamples as required in serum analysis, or when aiming for maximum absolute power of detection in combined analytical procedures for trace analysis.

5.3.3. Figures of Merit

Power of Detection. The lowest elemental concentrations which can be determined by flame AAS are often given in terms of the so-called characteristic concentrations (µg/mL). For aqueous solutions of such concentration an absorption of 1 % corresponds to an extinction of 0.0044. The noise comprises contributions from the nebulizer as well as from the flame. Detector noise limitations may also occur. The latter can be minimized by operating the hollow cathode lamp at sufficiently high current.

In order to obtain maximum power of detection, the atomization efficiency should be as high as possible. Optimization of the geometry of the spray chamber and the nebulizer gas flow is required. The primary radiation should be well isolated by the monochromator and the amount of nonabsorbed radiation reaching the detector should be minimized by selection of the appropriate observation zone with the aid of a suitable illumination system.

The detection limits of flame AAS are especially low for rather volatile elements, which do not form thermally stable oxides or carbides, and have high excitation energies, such as Cd and Zn. Apart from these and some other elements such as Na and Li, the detection limits in flame AAS are higher than in ICP – AES (see Table 9).

Analytical Precision. By integrating the extinction signals over 1 – 3 s, relative standard deviations down to 0.5 % may be achieved. When injecting discrete samples into the nebulizer or applying flow injection analysis, slightly higher relative standard deviations are obtained. They increase rapidly on leaving the linear part of the calibration curve and applying linearization by software.

Interference. Interference in flame AAS comprises spectral, chemical, and physical interference.

Spectral Interference. Spectral interference of analyte lines with other atomic spectral lines is of minor importance compared with atomic emission work. It is unlikely that resonance lines emitted by the hollow cathode lamp coincide with an absorption line of another element present in the atomizer. However, it may be that several emission lines of the hollow cathode are within the spectral band width or that flame emission from bands or continuum occurs. Both contribute to the nonabsorbed radiation, and decrease the linear dynamic range. Nonelement-specific absorption (Section 5.6) is a further source of spectral interference.

Chemical Interference. Incomplete atomization of the analyte causes chemical interference, due to the fact that atomic absorption can only occur with free atoms. Reactions in the flame which lead to the formation of thermally stable species decrease the signal. This is responsible for the depression of calcium signals in serum analysis by proteins, as well as for the low sensitivities of metals which form thermally stable oxides or carbides (Al, B, V, etc.) in flame AAS. A further example of chemical interference is the suppression of the extinction of alkaline earth metals as a result of the presence of oxyanions (OX) such as aluminates or phosphates. This well-known "calcium phosphate" interference is caused by the reaction:

$$\mathrm{MOX} \overset{\text{excess}}{\rightleftharpoons} \mathrm{X} + \mathrm{MO} \overset{\text{high } T}{\rightleftharpoons} \mathrm{M} + \mathrm{O} \overset{+\mathrm{C}}{\rightleftharpoons} \mathrm{M} + \mathrm{CO}$$

In hot flames, such as the carbon-rich flame, the equilibrium lies to the right, but with excess oxyanions it is shifted to the left, and no free M atoms are formed. This can be corrected by adding a metal R, the so-called releasing agent, which forms yet more stable oxysalts and releases metal M again. La and Sr compounds can be used:

$$\mathrm{MOX} + \mathrm{RY} \overset{\text{excess}}{\rightleftharpoons} \mathrm{ROX} + \mathrm{MY}$$

When $LaCl_3$ is added to the sample solutions, phosphate can be bound as $LaPO_4$.

For alkali metals, the free atom concentrations in the flame can decrease as a result of ionization, especially in hot flames. This leads to a decrease of extinction for alkali metals. It may also lead to false analysis results, as the ionization equilibrium for the analyte element is changed by changes in concentration of easily ionized elements. In order to suppress these effects, ionization buffers can be added. The addition of excess (low ionization potential) is most effective to suppress ionization changes of other elements, as it provides a high electron number density in the flame.

Physical Interference. Physical interference may arise from incomplete volatilization, and occurs especially in strongly reducing flames. In steel analysis, the depression of Cr and Mo signals as a result of excess Fe is well known. It can be reduced by adding NH_4Cl. Further interference is related to nebulization effects, and arises from the influence of the concentration of acids and salts on the viscosity, density, and surface tension of the analyte solutions. Changes in physical properties from one sample solution to another influence aerosol formation efficiencies and the aerosol droplet size distribution, as discussed in Section 4.1. However, related changes of nebulizer gas flow also influence the residence time of the particles in the flame.

5.4. Electrothermal Atomic Absorption

The use of furnaces as atomizers for quantitative AAS goes back to the work of L'vov, and led to

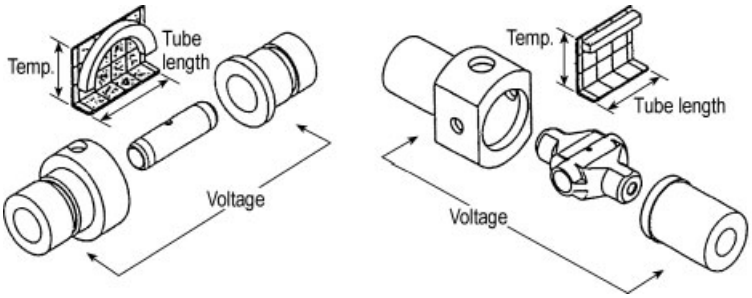

Figure 33. Temperature distribution in a (a) longitudinally and (b) transversely heated graphite furnace

the extension of AAS to very low absolute detection limits. In electrothermal AAS, graphite or metallic tube or cup furnaces are used and, with resistive heating, temperatures are achieved at which samples can be completely atomized. For volatile elements, this can be accomplished at 1000 K, whereas for more refractory elements the temperature should be up to 3000 K.

The high absolute power of detection of electrothermal AAS is due to the fact that the sample is completely atomized and vaporized, and the free atoms spend a long time in the atomizer. The signals obtained are transient, as discussed in Section 4.4.1.

5.4.1. Atomizers

Graphite furnaces, cups, and filaments are all used as atomizers in electrothermal AAS [165]. The models originally proposed by L'vov et al. [3] and by Massmann [4] are described in Section 4.4.2. In the most widely used type the optical beam is led centrally through a graphite tube, closed on both sides, with a quartz viewing port mounted in the cooled tube holders. Sample aliquots are introduced with the aid of a micropipette or a computer-controlled dispenser, through a sampling hole in the middle of the tube. The normal temperature profile of a graphite furnace leads to differences in spreading of the analyte over the graphite surface and may produce changes in volatilization from one sample to another. This effect can be avoided by using a transversely heated furnace, where the temperature is constant over the whole tube length (Fig. 33). Especially in electrothermal AAS, problems may arise from recombination of the atomized analyte with oxygen and other nonmetals, as the free atom concentration measured in AAS decreases. This may occur when the volatilized analyte enters a fairly cool plasma, as is normally the case in a tube furnace. This can be prevented [109] by dispensing the sample on a low-mass graphite platform located in the tube furnace. The low heat content then allows very rapid heating and volatilization of the analyte. The analyte is surrounded by a furnace plasma of about the same temperature as the sample carrier, which lowers the risk of recombination. Tungsten furnaces (see Section 4.4.2) [112] are useful for the determination of refractory carbide-forming elements, but they are more difficult to use with real samples, and reduction of analyte oxides does not occur. Berndt et al. [166] have shown that tungsten coils are suitable atomizers for dry solution aliquots, especially for matrix-free solutions obtained in combined analytical procedures involving separation of the analyte from the matrix elements. Owing to the relatively large coil surface area (Fig. 34), salt problems in real samples may be lower than with a metal tube furnace.

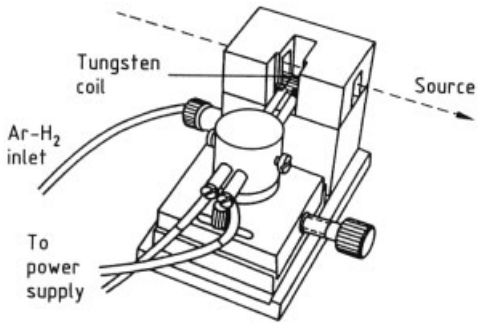

Figure 34. Tungsten filament atomizer [166]

5.4.2. Thermochemistry

The dissociation equilibria between the analyte elements and their compounds are very important, as they determine the fraction of analyte available as free atoms for AAS. They affect both the analytical sensitivity, and the possibility of systematic errors. The thermochemical behavior of the sample is of prime importance for the height of the extinction signal as well as its form.

The acids present in the sample solution are normally removed during drying and matrix decomposition. Residues present during the absorption measurement lead to nonelement-specific absorption, especially in the case of acids with high boiling points, such as $HClO_4$, H_2SO_4, or H_3PO_4. The absorption stems from ClO, SO, SO_2, or PO molecular bands. These species may also be produced by the dissociation of salts. Further problems may be caused by the oxides of the analytes, which result from salt dissociation. This nondissociated oxide is lost for AAS, and may give rise to nonelement-specific absorption. When salt dissociation and reduction of analyte oxides is not complete before the absorption measurement, several peaks may occur in the absorption signal. This can often be avoided by the platform technique, which produces a sharp rise in heating of the furnace and lowers the risks of analyte deposition at the cool parts of the furnace.

Thermochemistry is especially important reliable matrix destruction and removal of matrix elements without analyte loss. Thermochemical reagents such as quarternary ammonium salts ($R_4N^+Cl^-$) mineralize organic samples at temperatures below 400 °C, and prevent losses of elements which are volatile or form volatile, perhaps organic, compounds. This may be helpful for Cd and Zn, which form volatile compounds in a number of organic matrices. Removal of NaCl, present in most biological samples, may be helpful to prevent matrix interference; this must be done at low temperature to prevent analyte loss, and can be achieved by the addition of NH_4NO_3:

$$NH_4NO_3 + NaCl \rightleftarrows NH_4Cl + NaNO_3$$

Excess NH_4NO_3 dissociates at 350 °C, NH_4Cl sublimates and $NaNO_3$ decomposes below 400 °C, so NaCl is completely removed at temperatures below 400 °C. Without the addition of NH_4NO_3, this would be possible only at the volatilization temperature of NaCl (1400 °C), with inevitable analyte loss. In the graphite furnace, several elements (Ti, V, etc.) form thermally stable compounds such as carbides, producing negative errors some of the analyte is bound and does not contribute to the AAS signal. The use of pyrolytically coated graphite tubes is helpful, as diffusion of the analyte solution into the graphite decreases the risk of carbide formation. Flushing the furnace with nitrogen can be helpful. In the case of Ti, a nitride is formed which, in contrast to the carbide, can be easily dissociated. Other thermochemical means to decrease interference are known as matrix modification. Addition of Pd compounds, $Mg(NO_3)_2$, etc., has been successful for matrix-free vapor cloud formation [167]. The mechanisms involved also relate to surface effects in the furnace [168] and are an active field of research.

Development of the temperature program is vital in furnace AAS. Selection of the different temperature steps, and the use of all kinds of thermochemical effect are most important to minimize matrix influence without analyte loss.

Radiotracers are helpful for understanding and optimizing analyte volatilization in furnace AAS; element losses and their causes at all stages of the atomization process can be followed quantitatively. This has been studied in detail for elements such as As, Pb, Sb, and Sn by Krivan et al. [169].

5.4.3. Figures of Merit

The *analytical sensitivity* and *power of detection* of electrothermal AAS are orders of magnitude higher than those of flame AAS. This is due to the fact that, in the furnace, a higher concentration of atomic vapor can be maintained compared with flames. Dilution of the analyte by the solvent is avoided, as it has already evaporated before atomization and large volumes of combustion gases are not involved. For most elements, the characteristic mass, i.e., the absolute amount for which an extinction of 0.0044 (1 % absorption) is obtained, are orders of magnitude lower than in flame AAS [165].

Detection limits in flame AAS: 0.1 – 1 ng/mL (Mg, Cd, Li, Ag, Zn); 1 – 10 ng/mL (Ca, Cu, Mn, Cr, Fe, Ni, Co, Au, Ba, Tl); 10 – 100 ng/mL (Pb, Te, Bi, Al, Sb, Mo, Pt, V, Ti, Si, Se); > 100 ng/mL (Sn, As)

Detection limits in furnace AAS, (20 µL): 0.005 – 0.05 ng/mL (Zn, Cd, Mg, Ag); 0.05 – 0.5 ng/mL (Al, Cr, Mn, Co, Cu, Fe, Mo, Ba, Ca, Pb); 0.5 – 5 ng/mL (Bi, Au, Ni, Si, Te, Tl, Sb, As, Pt, Sn, V, Li, Se, Ti, Hg)

When sample aliquots of 20 – 50 µL are used, the concentration detection limits are often in the sub-ng/mL range. They are especially low for elements such as Cd, Zn, and Pb which have high volatility and high excitation energies, but also for Mg, Cu, Ag, Na, etc.

Interference in furnace AAS is much higher than in flame AAS. This applies both for physical and chemical interference.

Physical Interference. Physical interference may stem from differences in viscosity and surface tension, resulting in different wetting of the graphite tube surface with the sample solution, and changes in diffusion of the sample solution into the graphite. The first may be suppressed by the addition of surfactants to the sample solutions, and the second can be avoided by using pyrolytically coated graphite tubes, where diffusion of the analyte into the tubes is lower.

The influences of concomitants on volatilization can be minimized by the use of platform techniques and of an isothermal furnace. Both measures are particularly helpful in direct solid sampling.

Undissociated molecules, which may be oxides formed by dissociation of ternary salts (MgO, ZnO, etc.) or radicals and molecular species from solvent residues (OH, PO, SO_2, etc.) may cause nonelement-specific absorption. Rayleigh scattering of primary radiation by nonevaporated solid particles may occur, which also leads to radiation losses. Both phenomena necessitate the application of suitable techniques for background correction for most analyses of real samples by furnace AAS (Section 5.6). Emission by the furnace itself may give continuous radiation, which may lead to systematic errors.

Chemical Interference. Chemical interference may be caused by losses of analyte during ashing, especially in the case of volatile elements such as As, Sb, Bi, Cd, but also for elements forming volatile compounds (e.g., halides). In this respect, removal of NaCl, which is present in most biological samples, is critical. Further element losses may occur from the formation of thermally stable compounds such as carbides or oxides which cannot be dissociated and hamper atomization of the analyte element; elements such as Ti or V are especially difficult to determine by furnace AAS. Some of these sources of systematic error can be reduced in furnace AAS by the use of thermochemical reactions.

Differences in the volatility of elements and their compounds may also be used for speciation work. It is possible to determine organolead compounds such as $Pb(CH_3)_4$ and $Pb(C_2H_5)_4$ directly by furnace AAS [170], [171] or after online coupling with chromatographic separation. This approach can be used for other species, but needs to be carefully worked out in each case.

Calibration in furnace AAS is usually done by standard addition, as many of the sources of interference change the slope of the calibration curve.

Recommendations to be followed to enable determinations to be as free from interferences as possible include: (i) isothermal operation; (ii) the use of a matrix (chemical) modifier; (iii) an integrated absorbance signal rather than peak height measurements; (iv) a rapid heating rate during atomization; (v) fast electronic circuits to monitor the transient signal; (vi) the use of a powerful background correction system such as the Zeeman effect. Most or all of these recommendations are incorporated into virtually all analytical protocols nowadays and this, in conjunction with the transversely heated tubes, has decreased the interference effects observed considerably.

5.5. Special Techniques

5.5.1. Hydride and Cold Vapor Techniques

Determination of traces and ultra-traces of Hg, As, Se, Te, As, and Bi is often by formation of either volatile mercury vapor, or of the volatile hydrides. This gives a high sampling efficiency and power of detection. The absorption measurement is often performed in a quartz cuvette.

Hg, for instance, can be transported after reduction by means of a carrier gas. For the absorption measurement no heating is required. The other elements are reduced to the volatile hydride, which is then transported with a carrier gas flow (argon or nitrogen) into a cuvette. In order to dissociate the hydrides thermally, the quartz cuvette must be heated to a temperature of 600 – 900 °C.

In both cases, flow cells enable continuous formation of the volatile compound and subsequent separation of the reaction liquid from the gaseous compound [172]. This is very useful (Section 4.3) for routine work and has practically replaced earlier systems, where $NaBH_4$ solution was reacted with a given volume of acidified sample [173]. In any case, trapping of the analyte vapor released may increase the absolute power of detection. In the case of Hg, this may be achieved by passing the Hg vapor over a gold sponge and binding it as an amalgam [174]. After collection, Hg can be liberated completely by heating. In the case of the hydrides, cold trapping can be applied. Here, the hydrides can be condensed by freezing in a liquid nitrogen trap, and freed from the excess hydrogen formed during the reaction [175]. The isolated hydrides can be released by warming. This removes hydrogen and gives trapping enrichment factors well above 100.

Hot trapping can also be used [176] as the hydrides have decomposition temperatures lower than the boiling points of the elements. Trapping can thus be performed in the graphite furnace itself. It is only necessary to heat the furnace to ca. 600 – 700 °C to decompose the volatile hydrides, and the elements condense in the furnace. Here they can be preenriched and volatilized at temperatures above 1200 °C during absorption measurement.

The cold vapor technique for Hg gives detectin limits < 1 ng with 50 mL of sample, and this can be further improved by trapping. The hydride technique gives detection limits below the ng/mL level for As, Se, Sb, Bi, Ge, Sn, etc., levels appropriate for analyses of drinking water. Hydride techniques, however, are prone to interference (Section 4.3). In AAS, interference may occur not only as a result of influences on the hydride formation reaction, but also as a result of influences of concomittants on the thermal dissociation of the hydride. Interference by other volatile hydride-forming elements may also occur [177].

5.5.2. Direct Solid Sampling

In both flame and furnace AAS, direct solid sampling is possible. In flames, sampling may involve a boat or cup, as introduced by Delves [113]. New approaches such as the combustion of organic samples and the introduction of the vapors released into a flame are possible [178]. For volatile elements such as Pb, detection limits are in the μg/g range. Direct powder sampling in furnaces was introduced in atomic spectrometry by Preuss in the 1940s, and in furnace AAS it can be easily done with a powder sampling syringe [116]. For direct powder analyses with flame and furnace AAS, however, it is preferable to work with slurries. Slurry atomization was introduced in flame AAS work by Ebdon and Cave [179] in 1982. It has been shown to be useful for flame work, and to improve powder sampling in furnace AAS. The latter has been successfully used for direct analysis of SiC powders [180]. Direct solid sampling, however, should be used very carefully. The amount of sample used is often of the order of a few milligrams, and sampling errors may occur as a result of sample inhomogeneity. Therefore, the use of furnaces in which larger powder amounts can be handled has been proposed. In direct powder analysis, calibration may be troublesome because the particle size affects analyte volatilization and transport.

In compacted samples, direct analysis by AAS is feasible. The use of cathodic sputtering combined with AAS of the atomic vapor cloud produced has been recently reviewed [181], with jet-enhanced sputtering to give high analyte number densities. The approach has been made commercially available [182]. Its feasibility for direct analyses of steels has been shown, especially for samples which are difficult to dissolve, in which refractory oxide-forming elements such as Zr need to be determined [183]. For both electrically conducting and insulating samples, laser ablation combined with AAS may be useful. AAS measurements can be performed directly in the laser plume. Measurement of nonelement-specific absorption is very important, because of the presence of particles, molecules, and radicals, and the emission of con-

tinuous radiation. The absorption measurement should be performed in the appropriate zone. When using laser ablation for direct solid sampling, the atomic vapor produced can be fed into a flame for AAS work, as described by KANTOR et al. [184].

5.5.3. Indirect Determinations

For the determination of elements of which the most sensitive lines are in the vacuum UV conventional AAS cannot be applied. For this, and for the determination of chemical compounds, indirect methods can be used. For instance, sulfates can be determined by precipitation as $BaSO_4$ and a determination of excess barium with AAS. A similar approach can be used for the determination of halides after precipitation with silver.

Further indirect determinations can be performed by making use of chemical interference. Calcium phosphate interference in flame AAS can be used to determine phosphate by measuring the decrease of the absorption signal for calcium. However, for real samples, indirect techniques should be used very carefully in view of possible systematic errors.

5.6. Background Correction

In AAS, systematic errors are often due to nonelement-specific absorption, which necessitates the use of background correction procedures. They can only be omitted in the analysis of sample solutions with low matrix concentrations by flame AAS, and are especially necessary in furnace AAS. Determination of nonelement-specific absorption can be performed in several ways.

5.6.1. Deuterium Lamp Technique

This approach was introduced by KOIRTHYOHANN and PICKETT in 1965 [159] and is now provided in almost every AAS system. The total absorption resulting from the presence of the element and the background absorption are measured with hollow cathode lamp radiation, but, in addition, a continuum source is used which measures only the background absorption. This is possible as the monochromators used in AAS have a large spectral band width compared with the physical width of the resonance line emitted by the hollow cathode source, and the width of the absorption line. Accordingly, all radiation emitted by the hollow cathode lies within the absorption profile of the line, and elemental absorption does not contribute significantly to the absorption of the continuous source (Fig. 35), which extends homogeneously over the whole spectral band with. As the continuous source, an electrical discharge in deuterium can be used, as it emits a rather smooth continuum in the UV range up to 400 nm.

Alternative measurements of the absorption of the hollow cathode lamp radiation and of the radiation of the continuous source can be performed in a rapidly alternating mode to give a quasi-simultaneous measurement of line and background absorption, as required with transient signal. The latter can be realized by a rotating mirror and is rapid enough for high time resolution [185].

This type of background correction assumes that the background absorption is continuous within the spectral band width of the monochromator. This is not the case when the background absorption stems from molecular bands, which have a rotation – vibration hyperfine structure. They can arise from radicals produced by dissociation of the solvent (OH, SO_2, SO_3, N_2^+, CN, etc.) or from molecular oxides (MO). As such contributions occur especially with the complex matrices of real samples, background correction with a D_2 lamp often leads to systematic errors, which cannot be corrected by standard addition methods. Furthermore, the radiant density of the D_2 lamp over a large part of the spectrum is rather low. Accordingly, there are limits to the number of analytical lines which can be used and the number of elements which can be determined. As the spectral radiance of the D_2 lamp is lower than that of the hollow cathode, the latter must be operated at low current, which leads to detector noise limitations and poor detection limits. Finally, the two primary radiation sources are difficult to align, so they have to pass through the same zone of the atomizer, and this may lead to further systematic errors.

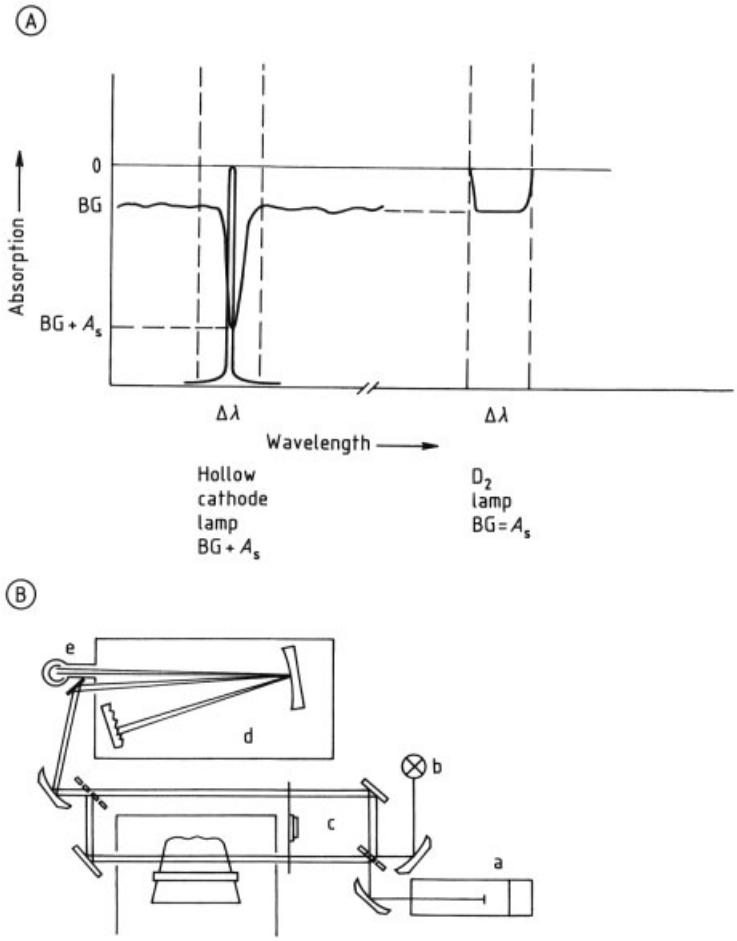

Figure 35. Background correction with the D_2 lamp technique
A) $\Delta\lambda$ = Spectral bandwidth of monochromator; B) Optical diagram of the spectrometer (Model 2380, Bodenseewerk Perkin-Elmer)
a) Hollow cathode lamp; b) D_2 lamp; c) Chopper; d) Monochromator; e) Photomultiplier
BG = Nonelement-specific background absorption; A_S = Element-specific absorption signal

5.6.2. Zeeman Effect Technique

Zeeman AAS makes use of the splitting of atomic spectral lines by a magnetic field. When a magnetic field B (≤ 1 T) is applied, the shift in wavenumber (ΔT_m) of the so-called σ-components, where the π-components remain constant, is:

$$\Delta T_m = M g L B \tag{151}$$

where M is the magnetic quantum number, L is the Lorentz unit (e/(4 πmc^2), and g the Landé factor, a function of the total quantum number J, the orbital momentum quantum number L, and the spin quantum number S. The intensities of the σ-components (for which $\Delta M = \pm 1$) and the π-components (for which $\Delta M = 0$) are a function of ΔJ (0, 1) and ΔM (0, $\pm$ 1) for the transitions. In the normal Zeeman effect, which occurs for singlet transitions (e.g., with alkaline earth metals and metals of group 12 such as Cd and Zn) $g = 1$, whereas in all other cases multiplets occur (anomalous Zeeman effect). For a transverse magnetic field (perpendicular to the observation direction) there is a π-component at the original wavelength. For this component, $\Delta M = 0$ and it is polarized parallel to the field. There are two σ-components (σ^+ and σ^-) for

which $\Delta M = \pm 1$. They are polarized perpendicular to the field. For a longitudinal field (parallel to the observation direction) there is no π-component ($\Delta M = 0$) and the σ-components ($\Delta M = \pm 1$) are circularly polarized. In order to use the Zeeman effect for background correction, several approaches are possible (Fig. 36) [186]. A constant transverse field may be used, measuring the π- and the σ-components alternately with the aid of a polarizer and a rotating analyzer. Alternatively, an a.c. longitudinal field can be used, however, measuring σ-components only with the aid of a static polarizer, at zero and at maximum field strength.

In both cases, substracting one signal from the other eliminates the background absorbance. This assumes that both signals have constant intensities through the whole analytical system and that both have the same absorption coefficients for the background. The latter technique has the additional advantage that the magnetic field can be controlled since it is produced by an electromagnet. Accordingly, the splitting can be optimized with respect to the element determined and to the background structure. The magnetic field can be applied in the atomizer, which is possible both with permanent and electromagnets. In principle, the magnetic field can also be placed at the primary source, which is optimal in the case of a permanent magnet; both flames and furnaces can be used as atomizers and the exchange is easier. Moreover, a larger furnace can be used which is very useful for direct solid sampling.

Zeeman AAS has several analytical advantages. The accuracy of correction of a structured background is better than with the D_2 lamp technique. However, when the background structure stems from molecular bands, it must be remembered that molecular bands may also display a Zeeman effect. Systematic errors resulting from this fact may be larger when one line component is measured in an alternating field.

The *detection limits* in Zeeman AAS should be lower than in the case of background correction with a D_2 lamp. The system uses only one source; it can be operated at high intensity, so detector noise limitations are avoided. This advantage is most pronounced when one component is measured in an alternating field.

Another consequence of the use of one primary source is improved stability. The analytical sensitivity in Zeeman AAS, however, is inferior to that of conventional AAS, is the linear dynamic range. The latter relates to the fact that a difference between two absorbances is involved, which may even lead to bending of the calibration curve. These effects are less pronounced when measuring one component in an alternating field. Zeeman AAS is widely used, for instance, for trace determinations in biological samples.

5.6.3. Smith – Hieftje Technique [187]

This technique of background correction makes use of the fact that resonant atomic spectral lines emitted by a hollow cathode lamp may display self-reversal when the lamp is operated at high discharge current. A hollow cathode lamp operates during the first part of the measurement cycle, at low current. Self-reversal does not occur and the resonant radiation is absorbed both by the analyte atoms and by background-producing species. In the second part of the measurement cycle, the current is briefly pulsed to > 500 mA, and a pronounced self-reversal occurs. The intensity at the analytical wavelength decreases, but the intensities in the side-wings remain high (mostly background absorption). By subtraction of the absorbances, the net atomic absorption signal is obtained.

Similar to Zeeman AAS, the Smith – Hieftje technique can be used for lines over the entire spectral range, and uses only one primary radiation source, so that alignment and stability are optimum. The technique is simple and much cheaper than Zeeman AAS. It does not suffer from limitations due to Zeeman splitting of molecular bands. However, as self-reversal is not complete, the technique can be used only for rather low background absorbance, sensitivity is decreased as the self-absorption is 40 % at most, and special provisions have to be made for pulsing the lamps at high currents.

5.6.4. Coherent Forward Scattering

The intensity of scattered radiation is particularly high when monochromatic radiation is used as primary radiation, the wavelength of which is equal to that of a resonance line of

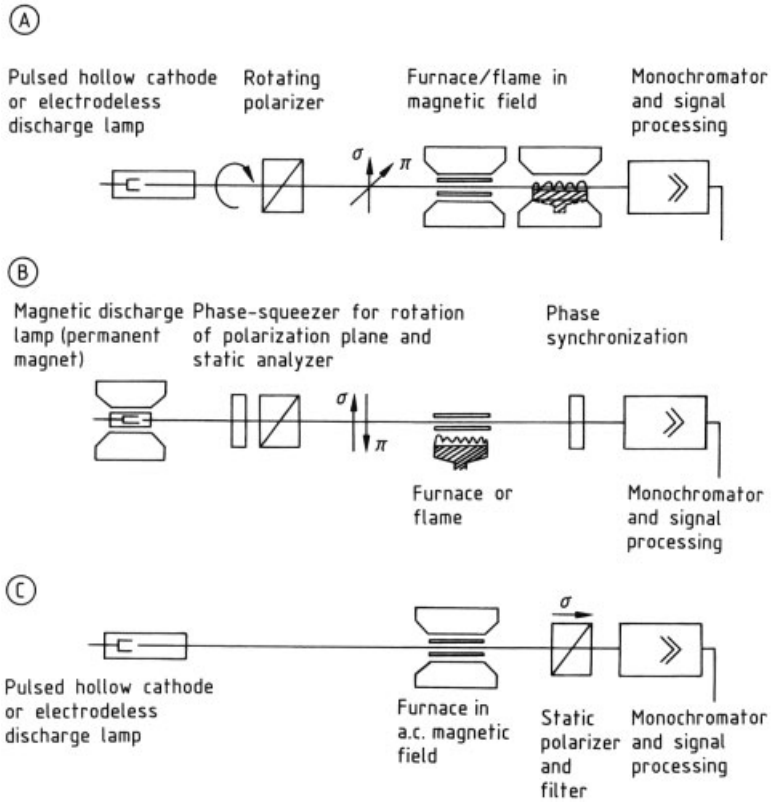

Figure 36. Zeeman atomic absorption
A) Rotating polarizer and permanent magnet applied to the atomizer; B) Permanent magnet around the primary source; C) Longitudinal field of a.c. magnet applied to the atomizer [186]

the scattering atoms. When the latter are subjected to a magnetic field, the scattered radiation becomes coherent in the direction of the primary beam and the scattering atomic vapor becomes optically active (magneto-optic effect). Depending on whether a transverse or longitudinal magnetic field is used, a Voigt or Faraday effect is observed. In coherent forward scattering, radiation from the primary source is led through the atomizer across which a magnetic field is applied. When the atomizer is placed between crossed polarizers, scattered signals for the atomic species occur against zero background. When a line source, such as a hollow cathode lamp or a laser, is used, determinations of the respective element can be performed. For a continuous source, such as a xenon lamp, and a multichannel spectrometer, simultaneous multielement determinations can be performed. The method is known as coherent forward scattering (CFS) atomic spectrometry [188–190].

The *intensities of the scattering signals* in both Voigt and Faraday effects can be calculated [188]. In the *Faraday effect*, two waves are polarized parallel to the magnetic field. If n^+ and n^- are the refractive indices, $n_m = (n^+ + n^-)/2$ and $\Delta n = (n^+ - n^-)/2$, the intensity $I_F(k)$ at wavenumber k is:

$$I_F(k) = I_0(k)\, F[\sin(kl\Delta n)] \exp(kln_m) \qquad (152)$$

where $I_0(k)$ is the intensity of the primary radiation at wavenumber k, and l is the length of the atomizer. The sinusoidal term relates to the rotation of the polarization plane and the exponential term to the atomic absorption. As both n_m and Δn are functions of the density of the scattering atoms, $I_F(k)$ is proportional to the square of the density of scattering atoms N:

$$I_F(k) = I_0 N^2 l^2 \qquad (153)$$

In the *Voigt effect*, the scattered radiation has two components. One is polarized parallel to the magnetic field (normal component) and the other perpendicular (abnormal component). If n_0 and n_e are the respective refractive indices, their intensities can be calculated as for the Faraday effect.

Coherent forward scattering atomic spectrometry is a multielement method. The instrumentation required is simple and consists of the same components as a Zeeman AAS spectrometer. As the spectra contain only a few resonance lines, a spectrometer with low spectral resolution is sufficient. The detection limits depend considerably on the primary source and on the atomizer. With a xenon lamp as primary source, multielement determinations can be performed, but the power of detection is low as the spectral radiances are low compared with those of a hollow cathode lamp. By using high-intensity laser sources, the intensities of the signals and the power of detection can be considerably improved. When furnaces are used as atomizers, typical detection limits in the case of a xenon arc are: Cd 4 ng; pb 0.9; Tl 1.5; Fe 2.5; and Zn 50 ng [188]. These are considerably higher than in furnace AAS.

The sensitivity of CFS atomic spectrometry is high as the signals are proportional to the square of the atom number densities. The dynamic range is similar to atomic emission spectrometry, of the order of three decades. Information on matrix effects in real samples is still scarce. As scattering by molecules and undissociated species is expected to be low, background contributions may be low compared with AAS.

5.7. Fields of Application

The different methods of AAS are very powerful for analysis of solutions. The instruments are simple and easy to operate; they are in use in almost any analytical laboratory. Especially when one or only a few elements have to be determined in a large number of samples, as is the case in clinical or food analysis, AAS methods are of great importance compared with other methods of elemental analysis.

When solids need to be analyzed, samples have to be dissolved. Sample decomposition methods range from simple dissolution in aqueous solutions to treatment with strong and oxidizing acids. In all sample dissolution and pretreatment work for ASS, attention must be paid to all the problems which be set trace elemental analytical chemistry. This includes precautions for avoiding contamination from the reagents, the vessels used, and from the laboratory atmosphere (→ Trace Analysis, Chap. 4.5.). All factors which may cause analyte loss as a result of adsorption or volatilization must be studied in detail [191].

Clinical chemistry, the metals industry, the analysis of high-purity chemicals and pure substances, environmental analysis, and the life sciences are important fields for the application of AAS. Applications in all these areas are regularly reviewed [192].

Clinical Chemistry. In clinical analysis, flame AAS is very useful for serum analysis. Here, Ca and Mg can be directly determined in serum samples after 1 : 50 dilution, even in micro-aliquots of 20 – 50 µL [193]. In the case of Ca, La^{3+}, or Sr^{2+} is added to avoid phosphate interference. Na and K are mostly determined in the flame emission mode, which can be carried out with any flame AAS instrument. Here, the burner head is often turned to shorten the optical path, avoiding self-reversal. For the direct determination of Fe, Zn, and Cu, flame AAS can be used, but at lower sample dilution. For trace elements such as Al, Cr, Co, Mo, and V, determination with flame AAS often requires preenrichment, but in serum and other body fluids as well as in biological matrices, some elements can be directly determined with furnace AAS. This applies to toxic elements such as Ni, Cd, Pb, etc., in screening for workplace exposure.

For *metal samples*, flame and furnace AAS are important analysis methods. They find applications in the characterization of incoming material, and for product analysis. In combination with matrix removal, they are indispensable for the characterization of laboratory standards. The latter are used to calibrate direct methods such as X-ray spectrometry and spark emission spectrometry for production control. However, relevant elements such as B, the rare earths, Hf, Zr, etc., are much better determined by plasma emission spectrometry. For the determination of environmentally relevant elements in slags and

ores, dissolution by fusion and subsequent AAS may be applied.

In the analysis of *high-purity substances*, matrix removal is often very important for preconcentration. For this, separation techniques such as ion exchange, liquid – liquid extraction of metal complexes with organic solvents, fractional crystallization, precipitation, coprecipitation, and electrochemical methods may be used [194].

For analytical problems in the *life sciences* and *environmental analyses*, AAS is widely used. For the analysis of wastewater, flame and furnace AAS are complementary to plasma emission spectrometry. The latter enables determination of a large number of elements per sample, but is more expensive than flame AAS in terms of instrumentation, and less sensitive than furnace AAS. For analysis of drinking water, furnace AAS is very useful, but because of its single-element character, it is increasingly being replaced by plasma mass spectrometry, especially in large laboratories.

Atomic absorption spectrometry is a useful tool for *speciation* work. This applies for combinations of species-specific extraction procedures and subsequent determination of the respective species by AAS. Such a method has been worked out for Cr(III) and CR(VI) compounds [195]. On-line coupling of gas or liquid chromatography with AAS is increasingly used [196]. In furnace AAS, coupling with gas chromatography has been shown to be very useful for determination of organolead [170] and organotin compounds [197]. By combining the cold vapor technique with HPLC, a very sensitive determination of Hg species at the sub-nanogram level becomes possible [198]. For speciation, coupling of flow injection analysis and column chromatography with flame AAS, and direct coupling of HPLC with flame AAS (possible with high-pressure nebulization) are very powerful, as shown by the speciation of Cr(III) and Cr(VI) for wastewater analysis [96].

Manufacturers. The most important manufacturers of AAS equipment are:

Baird Atomic, USA: arc/spark AES, ICP-AES, AAS
Analytic Jena AG, Germany: AAS
Fisher Scientific, USA: AAS
GBC, Australia: TOF-ICP-MS, AAS
Hitachi, Japan: ICP-AES, AAS, ICP-MS
PerkinElmer Instruments (incorporating Princeton Applied Research, ORTEC, Signal Recovery and Berthold products), USA: AAS, ICP-OES, ICP-MS, GC, GC/MS
Shimadzu, Japan: arc/spark AES, ICP-AES, AAS
TJA Solutions (incorporating Thermo Jarrell Ash, VG Elemental and Unicam AA), USA: AAS, ICP-AES, ICP-MS, SF-ICP-MS, GD-AES, GD-MS
Thermoquest (incorporating Finnigan), USA: SF-ICP-MS, AAS
Varian Associates, Australia: ICP-AES, ICP-MS, AAS

6. Atomic Emission Spectrometry

6.1. Principles

Atomic emission spectrometry (AES) goes back to Bunsen and Kirchhoff, who reported in 1860 investigations of the alkali and alkalineearth metals with the aid of their spectroscope [1], which is still the heart of any dispersive spectrometer. The new elements cesium, rubidium, thorium, and indium were discovered from their atomic emission spectra. By the beginning of the twentieth century, Hartley, Janssen, and De Gramont further developed spectrochemical analysis, and flames, arcs, and sparks became powerful analytical tools. To overcome the difficulties arising from unstable excitation conditions, Gerlach in 1925 introduced the principle of the internal standard [199], where the ratio of intensities of the analytes to that of the matrix or of an added element is used. This was the breakthrough in atomic spectroscopy for elemental analysis. In AES, a reproducible and representative amount of the sample is introduced into a radiation source. Its role in AES is similar to the dissolution procedures of wet chemical analysis; continuous endeavors are made to improve existing radiation sources, and to develop new sources, to perform analyses with enhanced figures of merit. The role of the radiation source in AES does not end at the atomization stage (as in AAS), but includes excitation of states from which emission or even ionization occur.

Qualitative analyses can be performed with AES; the unambiguous detection and identification of a single, interference-free atomic spectral line of an element is sufficient to confirm its presence. The most intense line under a given set of working conditions is known as the most sensitive line. Elemental lines for different elements are situated in widely different spectral regions and may differ from one radiation source to another, as a result of the excitation and ionization processes. Despite the fact that spectral lines are very narrow (a few picometers), line coincidence can occur. High-resolution spectrometers are necessary for the qualitative evaluation of spectra, and more than one line should be used to decide on the presence of an element. Qualitative emission spectral analysis is easy when the spectra are recorded photographically or with a scanning monochromator. For the first case, atlases are available, in which the spectra are reproduced with the most sensitive lines indicated. Spectral line tables are also very useful. They are available for arc and spark sources [200–202], and in much less complete form for newer radiation sources such as glow discharges [203] and inductively coupled plasmas [204–206].

Quantitative AES. In quantitative AES, the intensity of an elemental atomic or ionic line is used as analytical signal. As AES is a relative method, calibration has to be performed.

Calibration Function. The determination of the calibration function is important; it relates the intensity of a spectral line I to the concentration c of an element in the sample. SCHEIBE [207] and LOMAKIN [208], propose the following relation between absolute intensities and elemental concentrations:

$$I = ac^b \tag{154}$$

where a and b are constants. Absolute intensities are practical only in flame work or in plasma spectrometry, where very stable radiation sources are used. Normally, the intensity ratio of an analyte line to a reference signal is used, as proposed by GERLACH. This leads to calibration functions of the form:

$$I/I_{\mathrm{R}} = a' c^{b'} \tag{155}$$

The inverse calibration function is the analytical evaluation function:

$$c = c_{\mathrm{R}}\,(I/I_{\mathrm{R}})^{\eta} \tag{156}$$

or:

$$\log c = \log c_{\mathrm{R}} + \eta \log\,(I/I_{\mathrm{R}}) \tag{157}$$

In trace analyses, the slope of the analytical evaluation curve (in logarithmic form) is usually 1; at higher concentrations η may be > 1 as a result of self-reversal. In trace analysis by AES, the intensity of the spectral background I_{U} can be used as reference signal and c_{R} should be replaced by c_{U}, the concentration at which the line and background intensities are equal (background equivalent concentration or BEC value):

$$c_{\mathrm{U}} = [(I_{\mathrm{X}}/I_{\mathrm{U}})\,/c]^{-1} \tag{158}$$

The ratio $(I_{\mathrm{X}}/I_{\mathrm{U}})$ and c_{U} for a given element and line depend on the radiation source and its working conditions, and on the spectral apparatus [209], [210]:

$$c_{\mathrm{U}} = \frac{(\mathrm{d}B_{\mathrm{U}}/\mathrm{d}\lambda)\,\Delta\lambda_{\mathrm{L}}}{B_0}\,\frac{s_{\mathrm{eff}} R_{\mathrm{L}}}{s_0 R_0}\,\cdot$$
$$[(\pi/(\ln 16))]^{1/2}\,\left\{\left[(s_{\mathrm{a}}/s_{\mathrm{eff}})^4 + 11\right]\right\}^{1/4}$$
$$c_{\mathrm{U}} = c_{\mathrm{U},\infty} A_1 A_2 \tag{159}$$

where $c_{\mathrm{U},\infty}$ describes the influence of the radiation source, $(\mathrm{d}B_{\mathrm{U}}/\mathrm{d}\lambda)$ reflects the spectral distribution of the background intensity, B_0 is the radiant density for an analytical line at $c = 1$, and $\Delta\lambda_{\mathrm{L}}$ is the physical width of the analysis line. The term A_1 describes the influence of the spectral apparatus, and:

$$s_{\mathrm{eff}} = \left(s_{\mathrm{e}}^2 + s_0^2 + s_{\mathrm{L}}^2 + s_{\mathrm{z}}^2\right)^{1/2} \tag{160}$$

where s_{L} is the slit width corresponding to the physical width of the spectral line $\Delta\lambda_{\mathrm{L}}$, and $R_{\mathrm{L}} = \lambda/\Delta\lambda_{\mathrm{L}}$, the resolution required to resolve the line; s_{z} is the slit width corresponding to the optical aberrations in the spectral apparatus:

$$s_{\mathrm{z}} = \Delta\lambda_{\mathrm{z}}\,\mathrm{d}x/\mathrm{d}\lambda \tag{161}$$

where $\Delta\lambda_{\mathrm{z}}$ can be determined by measuring the practical resolution of the monochromator from deconvolution of the two components of the Hg 313.1 nm doublet and subtraction of the contributions of the diffraction slit width $s_0 = \lambda f/W$ and the entrance slit width s_{e}. The contribution

of the natural width of the Hg lines can be neglected, as it is very low in the case of a hollow cathode lamp. In monochromators, s_z predomates, but at very high resolution it becomes less significant and $A_1 \longrightarrow 1$. The factor A_2 describes the influences of the profile of the analysis line and the effective measurement slit. For photoelectric measurements, the latter is the exit slit width s_a of the spectrometer. This contribution is relevant only when $s_a/s_{eff} < 2$, which is practically the case only for photographic measurements where the line profile is scanned with a very narrow densitometer slit. For photoelectric measurements, $A_2 = s_a/s_{eff} >> 1$, as the exit slit width must be larger than the effective line width because of the thermal and mechanical stability of the system. The detection limit is a relevant figure of merit for an analytical method (Section 3.1). It can be written for AES:

$$c_L = c_U \left(I_X/I_U\right) \tag{162}$$

where I_X/I_U is the smallest line-to-background ratio which can be measured. When the definition is based on a 99.86 % probability, it is three times the relative standard deviation of the background intensities. For photographic or diode array registration, the intensities of the line and the spectral background are recorded simultaneously, and:

$$\sigma^2 = \left[\sigma_r^2 \left(I_X + I_U\right) + \sigma_r^2 \left(I_U\right)\right]^{1/2} \tag{163}$$

As the intensities of the line and the spectral background are almost equal near the detection limit, their standard deviations are equal and the relevant relative standard deviation reduces to:

$$\sigma^2 = \sigma_r \left(I_U\right) 3\sqrt{2} \tag{164}$$

Its value can be measured directly, but in photographic measurements it can be calculated from the standard deviation of the blackening with the aid of Equation (115). The latter shows that for trace analyses, emulsions with a high γ must be used. In photoelectric measurements, σ is the relative standard deviation of the background intensity, measured at a wavelength adjacent to the spectral line or in a blank sample at the wavelength of the analytical line. It may comprise several contributions:

$$\sigma^2 \left(I_U\right) = \sigma_P^2 + \sigma_D^2 + \sigma_A^2 + \sigma_V^2 \tag{165}$$

where σ_P is the photon noise of the source or the noise of the photoelectrons and $\sigma_P \approx \sqrt{n}$, where n is the number of photoelectrons; σ_D is the dark current noise of the photomultiplier and $\sigma_D \approx I_D$, where I_D is the dark current; σ_A is the flicker noise of the radiation source and $\sigma_A \approx \varphi$, where φ is the radiant density; and σ_V is the amplifier noise. In most cases, σ_V can be neglected. For very constant sources, σ_A is small. Owing to the proportionality between σ_D and I_D it is advisable to select a photomultiplier with a low dark current, to avoid detector noise limitations. When these points are taken care of, the photon noise limits the power of detection.

When signals include a blank contribution, the latter also has fluctuations, which limit the power of detection. If I_X is the lowest measurable signal without blank contribution, I'_X the lowest analytical signal including a considerable blank value, I_U the background signal, and I_{Bl} the blank signal:

$$\frac{\left(I'_X + I_{Bl} + I_U\right) - \left(I_{Bl} + I_U\right)}{3\sqrt{\left[\sigma^2\left(I'_X + I_{Bl} + I_U\right) + \sigma^2\left(I_{Bl} + I_U\right)\right]}} =$$

and

$$\underline{\left(I_X + I_U\right) - I_U} = 3\sqrt{\left[\sigma^2\left(I_X + I_U\right) + \sigma^2\left(I_U\right)\right]} \tag{166}$$

This leads to:

$$\frac{\left(I'_X/I_U\right)}{\left[\sigma_r\left(I_{Bl} + I_U\right)/\sigma_r\left(I_U\right)\right]} = \frac{\left(I_X/I_U\right)\left(1 + I_{Bl}/I_U\right)}{} \tag{167}$$

and the detection limit becomes [211]:

$$c'_L = c_U \left\{\underline{\left(I_X/I_U\right)}\left(1 + I_{Bl}/I_U\right)\left[\sigma_r\left(I_{Bl} + I_U\right)/\sigma_r\left(I_U\right)\right]\right\}$$

or:

$$c_L \left[\sigma_r\left(I_{Bl} + I_U\right)/\sigma_r\left(I_U\right)\right]\left(1 + I_{Bl}/I_U\right) \tag{168}$$

To obtain maximum power of detection in AES, the standard deviations of the background, the blank contributions, and the c_U values must be minimized. The latter are a function of the radiation source, the elements, and the lines (reflected in $c_{U,\infty}$), and of the spectrometer (through A_1 and A_2). In order to keep A_1 and A_2 as close to unity as possible, it is necessary to use a spectrometer with high resolving power (high R_0), a narrow entrance slit ($s_e = 1.5\, s_0$), and to take

$s_a = s_{eff}$. On the other hand, the spectral background intensities must still be measurable. Indeed, the background intensities obtained are proportional to the entrance slit width, and detector noise limitations can arise when slit widths become too narrow. Also thermal and mechanical stabilities become limiting at narrow slit widths.

Apart from high power of detection, maximum analytical accuracy is very important. This relates to the freedom from interference. Whereas interference from influences of the sample constituents on sample introduction or volatilization and excitation in the radiation source differ widely from one source to another, most sources emit complex spectra and the risks of spectral interference in AES are much more severe than in absorption or fluorescence. Therefore, it is advisable to use high-resolution spectrometers. This is especially the case when trace determinations are performed in matrices emitting complex spectra. Knowledge of the atomic spectra is also very important so as to be able to select interference-free analysis lines for a given element in a well-defined matrix at a certain concentration level. This requirement is also much more stringent than in absorption or fluorescence.

6.2. Spectrometers

As all elements present in the radiation source emit their spectra at the same time, AES is by definition a multielement method. As well as simultaneous determinations, so-called sequential analyses are possible, provided the analytical signals are constant. Sequential and simultaneous multielement spectrometers have their own advantages and limitations.

Sequential spectrometers usually include Czerny – Turner or Ebert monochromators with which lines can be rapidly selected and measured one after another. Owing to the complex emission spectra, focal lengths are often up to 1 m and a grating with width up to 100 mm and spacing of 1/2400 mm is used. The instrument requires very accurate wavelength presetting, which is difficult for random access. The half-width of a spectral line corresponds to a grating angle of 0.0001°. Therefore, it is possible to work in a scanning mode and integrate stepwise at different locations across the line profile; alternatively, optical encoders and direct grating drives may be used. Systems employing fixed optics with a multislit exit mask, a movable photomultiplier, and fine wavelength adjustment by computer-controlled displacement of the entrance slit are also used. Sequential spectrometers are flexible because they can measure any elemental line and any background wavelength within the spectral range (usually 200 – 600 nm); however, they are slow when many elements need to be determined. They are especially of use in the case of stable radiation sources such as plasmas at atmospheric pressure or glow discharges operated at steady-state sputtering conditions.

Simultaneous Spectrometers. Modern simultaneous spectrometers use photoelectric radiation detection rather than the photographic plates formerly used for qualitative and semiquantitative analyses. Simultaneous photoelectric spectrometers mostly have a Paschen – Runge optical mounting. They have many exit slits and are especially suitable for rapid determinations of many elements in a constant, well-characterized matrix; the achievable precision is high. However, the analysis program is fixed and simultaneous spectrometers are not suitable for the analysis of random samples. Owing to stability requirements, larger exit slits are often used to overcome thermal drift, but spectral resolution is reduced. Especially for trace analysis, background correction may be required. This can be achieved by computer-controlled displacement of the entrance slit or by rotation of a quartz refractor plate behind the entrance slit.

Echelle spectrometers [36] are also often used. By a combination of an order-sorter and an echelle grating, in either parallel or crossed-dispersion mode, high practical resolution (up to 300 000) can be realized with an instrument of rather low focal length (down to 0.5 m). Therefore, stability and luminosity are high. By using an exit slit mask with a high number of preadjusted slits or a solid-state, two-dimensional array detector highly flexible and rapid multielement determinations are possible.

As the most sensitive lines of a number of elements such as S, P, C, As, are in the vacuum, it is desirable to work in this spectral range, which necessitates the elimination of oxygen from the entire optical path. This can be achieved by continuous evacuation of the spectrometer, purging with nitrogen or argon, or by the use of sealed gas-filled spectrometers. It is often helpful to measure low wavelengths in the second order, to increase spectral resolution and wavelength coverage (into the region where the alkali metals have their most sensitive lines). Solar blind photomultipliers can be used to filter out long-wavelength radiation of the first order.

6.3. Flame Emission

The oldest spectroscopic radiation sources operate at low temperature (Section 5.3.1), but have good spatial and temporal stability. They readily take up wet aerosols produced by pneumatic nebulization. Flame atomic emission spectrometry [162] is still a most sensitive technique for determination of the alkali metals, e.g., for serum analysis. With the aid of hot flames such as the nitrous oxide – acetylene flame, a number of elements can be effectively excited, but cannot be determined at low concentration. Interference arising from the formation of stable compounds is high.

6.4. Arcs and Sparks

Arc and spark AES have been widely used for analysis of solids and remain important methods for routine analysis.

Electrical arc sources for AES use currents between 5 and 30 A and burning voltages between the electrodes, for an arc gap of a few millimeters, of 20 – 60 V. Direct current arcs can be ignited by bringing the electrodes together, which causes ohmic heating. Under these conditions, d.c. arcs have a normal (decreasing) current – voltage characteristic owing to the high material ablation by thermal evaporation. A d.c. arc displays the cathode fall near the cathode (cathode layer) the positive column, and the anode fall area. The cathode glow delivers a high number of electrons carrying the greater part of the current. This electron current heats the anode to its sublimation or boiling temperature (up to 4200 K in the case of graphite) and the cathode reaches temperatures of ca. 3500 K as a result of atom and ion bombardment. An arc impinges on a limited number of small burning areas. The plasma temperature in the carbon arc is of the order of 6000 K and it can be assumed to be in local thermal equilibrium. Arcs are mostly used for survey trace analysis requiring maximum power of detection, but they may be hampered by poor precision (relative standard deviation $\geq 30\,\%$).

Direct Current Arc Spectrometry. The sample is presented to an electrode with suitable geometry and volatilizes as a result of the high electrode temperature. Cup-shaped electrodes made of graphite are often used. In the case of refractory matrices, anodic evaporation may be applied and the positive column or, for very sensitive methods, the cathode layer, where the volatilized sample ions are enriched (cathode layer effect) is selected as analytical zone.

Solid samples must be powdered and sample amounts up to 10 mg (powder or chips) are brought to the electrode after mixing with graphite powder and pelleting. Thermochemical reagents such as NH_4Cl may be added, to create similar volatilization conditions for all samples. Spectrochemical buffers such as CsCl may be added for controlling the arc temperature or the electron pressure. In most cases internal standards are used. The plasma is operated in noble gas atmospheres or in O_2 or CO_2 to create favorable dissociation and excitation conditions or to avoid intense bands such as those of CN. Procedures for the analysis of oxide powders by d.c. arc analysis are well known [212]; d.c. arcs are also of use for the analysis of metals, as is still done for copper with the so-called globule arc. Here distillation of volatile elements from the less volatile matrix is applied.

Stabilized d.c. arcs may be obtained by sheathing gas flows or by wall stabilization [213]. Magnetic fields can be used to produce rotating arcs, which have higher precision.

Dry solution analysis can be performed with metal electrodes (e.g., copper) or with pretreated graphite electrodes.

Direct current arc spectrography is still a most powerful method for trace determinations in solids, even with difficult matrices such as U_3O_8. Detection limits for many elements are

in the sub-µg/g range [214]. It is still in use for survey analysis.

Alternating Current Arc Spectrometry. Alternating current arcs have alternating polarity at the electrodes. Arc reignition is often by a high-frequency discharge applied at the a.c. discharge gap. Thermal effects are lower and the burning spot changes more frequently, so reproducibility is better than in d.c. arcs. Relative standard deviations of the order of 5 – 10 % may be obtained.

Spark Emission Spectrometry. Spark emission spectrometry is a standard method for direct analysis of metallic samples and is of great use for production as well as for product control in the metals industry [215], [216].

Sparks are short capacitor discharges between at least two electrodes. The spark current is delivered by a capacitor charged to voltage U_C by an external circuit (Fig. 37 A). When the voltage is delivered directly by an a.c. current, an a.c. spark is produced. The voltage can also be delivered after transforming to a high voltage and rectifying, to produce d.c. spark discharges. In both cases, the voltage decreases within a very short time from U_C to U_B, the burning voltage. During the formation of a spark, a spark channel is first built up and a vapor cloud is produced at the cathode. This plasma may have a temperature of 10 000 K or more, whereas in the spark channel even 40 000 K may be reached. This plasma is in local thermal equilibrium, but its lifetime is short and the afterglow is soon reached. Here the temperature and the background emission are low, but owing to the longer lifetime of atomic and ionic excited states, line emission may still be considerable.

In a discharge cirucit with capacitance *C*, inductance *L*, and resistance *R*, the maximum discharge current is:

$$I_0 = (U_0 - U_B)\sqrt{(C/L)} \approx U_0\sqrt{(C/L)} \quad (169)$$

The spark frequency within the spark train is:

$$f = 1/(2\pi)\sqrt{[1/LC - R^2/(4L^2)]} \quad (170)$$

or, when *R* is small:

$$f = 1/(2\pi)\sqrt{(1/LC)} \quad (171)$$

The capacitor may store energy:

$$Q = 1/2U_C^2C \quad (172)$$

of which 20 – 60 % is dissipated in the spark and the rest in the so-called auxiliary spark gaps in series with the analytical spark, or as heat. Depending on *R, L*, and *C* the spark has different properties; e.g., when *L* increases the spark becomes harder. As shown in Figure 37 B, there may be an oscillating (a), critically damped (b), or an overcritically damped spark (c). High-voltage sparks with $10 < U_C < 20$ kV were frequently used in the past. Their time cycle was regulated mechanically or electrically. Medium-voltage sparks ($0.5 < U_C < 1.5$ kV) are now used for metal analysis. They need a high-frequency discharge for ignition and are mostly unipolar discharges in argon. Accordingly, band emission is low, the most sensitive lines of elements such as C, P, S, and B can be excited, and the vacuum UV radiation is not absorbed. Often, a high energy is used to prespark, by which the sample melts locally, is freed from structural effects, and becomes homogeneous. Working frequencies are often up to 500 Hz, so high energies can be dissipated and high material ablation occurs.

Analytical Features. Spark emission spectrometry enables rapid and simultaneous determination of many elements in metals, including C, S, B, P, and even gases such as N and O. Therefore, spark emission AES is complementary to X-ray spectrometry for metallurgical analyses [217]. Analysis times including presparking (5 s) may be down to 10 – 20 s and the method is of great use for production control. Special sampling devices are used for liquid metals, e.g., from blast furnaces or convertors, and special care must be taken that samples are homogeneous and do not include cavities. The detection limits obtained in metal analysis differ from one matrix to another, but for many elements they are in the µg/g range (Table 4). The power of detection of spark analysis is lower than that of d.c. arc analysis. Analytical precision is high. By using a matrix line as internal standard, relative standard deviations at the 1 % level and even lower are obtained.

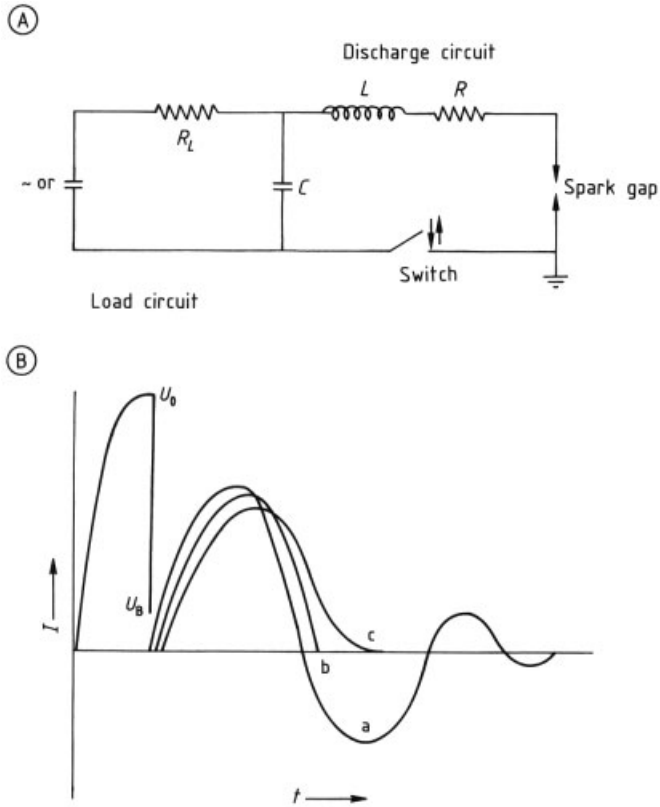

Figure 37. Circuitry in spark generator for optical emission spectrometry (A) and types of spark discharge (B)
a) Oscillating spark ($R<2L/C$); b) Critically damped spark ($R= 2L/C$); c) Overcritically damped spark ($R> 2L/C$)

Table 4. Detection limits (µg/g) for steels in spark and glow discharge AES and glow discharge MS

	Spark AES [a]	GD – AES [b]	GD – MS [c]
Al	0.5	0.1	
B	1	0.3	
Cr	3	0.05	
Cu	0.5	0.3	
Mg	2	0.9	
Mn	3	0.2	
Mo	1	0.8	
Nb	2	0.6	
Ni	3	0.1	
Pb	2		0.004
Si	3	0.4	
Ti	1	0.6	
V	1	1	
Zr	2	1.5	

[a] Spark discharge in argon with polychromator [215].
[b] Grimm-type glow discharge with 1 m Czerny – Turner monochromator [290].
[c] Grimm-type glow discharge and quadrupole mass spectrometer [334].

Matrix interference is high and relates to the matrix composition and the metallurgical structure of the sample. Therefore, a wide range of standard samples and matrix correction procedures are used (Section 5.2.1). Better selection of the analytical zone of the spark is now possible with the aid of optical fibers and may decrease matrix interference.

Acquisition of the spectral information from each single spark is now possible with advanced measurement systems, using integration in a small capacitor and rapid computer-controlled readout combined with storage in a multichannel analyzer. With the aid of statistical analysis, sparks with outlier signals can be rejected and the precision improves accordingly. Pulse differential height analysis enables discrimination between dissolved metals and excreted oxides for a number of elements such as Al and Ti. The first give low signals and the latter high intensities as they stem from locally high concentra-

tions [218]. By measurements in the afterglow of single sparks, or by the application of cross-excitation in spark AES [219], there is still room for improvement of the power of detection.

Spark AES has also been used for oil analysis. The oil is sampled by a rotating carbon electrode. Electrically insulating powders can be analyzed after mixing with graphite powder and pelleting.

6.5. Plasma Sources

In order to overcome the disadvantages related to the low temperature of chemical flames, but to have sources with similarly good temporal stability and versatility with respect to analyte introduction, efforts were directed toward electrically generated plasmas. At atmospheric pressure, these discharges have a temperature of at least 5000 K and, provided their geometry is properly optimized, they allow efficient introduction of analyte aerosols. In most cases, they are operated in a noble gas to avoid chemical reactions between analyte and working gas.

Among these plasma sources, d.c. plasmajets, inductively coupled plasmas (ICP), and microwave discharges have become important sources for AES, mainly for the analysis of solutions [220], [221] .

6.5.1. Direct Current Plasmas

Direct current plasmajets (DCPs) were first described as useful devices for solution analysis by Korolew and Vainshtein [222] and by Margoshes and Scribner [223]. Sample liquids were converted to aerosols by pneumatic nebulization, and arc plasmas operated in argon at temperatures of ca. 5000 K.

Types of Plasmajets. It is customary to distinguish between so-called current-carrying and current-free plasma jets. With *current-carrying plasmajets*, cooling as a result of changes in aerosol introduction has little influence. The current decrease resulting from a lower number density of current carriers can be compensated by an increase of the burning voltage, and this stabilizes the energy supply. The discharge, however, is very sensitive to changes in the electron number density, caused by varying concentrations of easily ionized elements. When the latter increase, the burning voltage energy supply and break down.

In *current-free* or *transferred plasmas*, the observation zone is situated outside the current-carrying zone. This can be achieved by the use of a supplementary gas flow directed perpendicular to the direction of the arc current, and observation of the tail flame. In this observation zone, there is no current. This type of plasma reacts strongly to cooling as no power can be delivered to compensate for the temperature drop. Therefore, it is rather insensitive to the addition of easily ionized elements. They do not cause a temperature drop, but shift the ionization equilibrium and give rise to so-called ambipolar diffusion. When the ionization in the observation zone increases, the pressure (proportional to the total particle number density – ions and electrons) increases, and ions as well as electrons move toward the outer zones of the plasma. Owing to the higher mobility of the electrons, a field is built up which amplifies the force on the positive ions directed to the outer plasma zones and thus increases the plasma volume. The d.c. plasmajet described by Margoshes and Scribner [223] is a current-carrying plasma, whereas the plasma described by Kranz [224] is a transferred plasma.

Three-Electrode Plasmajet. The three-electrode d.c. plasmajet developed by Spectrometrics (Fig. 38) [225] uses two graphite anodes protected by sheathing gas flows, each having a separate supply (2×10 A, 60 V). There is a common shielded cathode made of tungsten. The sample solution is usually nebulized by a cross-flow nebulizer, and the aerosol is aspirated by the gas flows around the anodes. The observation zone is small ($0.5 \times 0.5\ mm^2$), but very intense and located just below the point where the two positive columns meet. Optimally, this plasma is combined with a high-resolution echelle monochromator with crossed dispersion, as the latter has an entrance slit height ca. 0.11 mm. It has the advantages of a current-free plasma, as the observation zone is just outside the current-carrying part of the discharge, but still close enough to be compensated for strong cooling. The plasma temperature is of the order

of 6000 K, $n_e = 10^{14}$ cm^3 [226], and the plasma is near local thermal equilibrium.

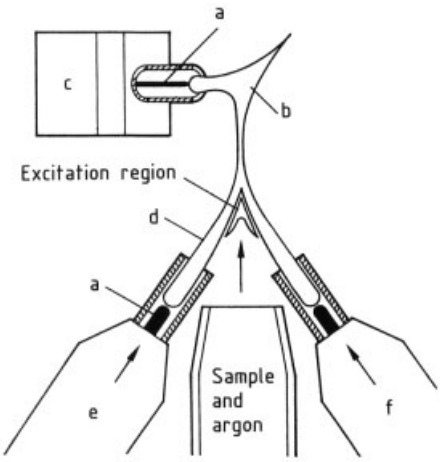

Figure 38. Three-electrode plasmajet [225]
a) Electrode; b) Plume; c) Cathode block; d) Plasma column (background) continuum; e) Argon anode block front; f) Argon anode block back

Detection limits for most elements are of the order of 5 – 100 ng/mL. For elements with very sensitive atomic lines such as As, B, and P, the detection limits are slightly lower than in ICP – AES [227]. The high power of detection also relates to the high resolution of the echelle spectrometer. Different concentrations of alkali metals, however, cause higher matrix effects than in ICP – AES and even may necessitate the use of spectrochemical buffers. The analytical precision achievable is high and relative standard deviations below 1 % can be reached. The system can cope with high salt content (>100 g/L), and has found considerable use in water analysis [228]. Its multielement capacity is limited, as the optimal excitation zones for most elements differ slightly; as a result of the small observation zones, considerable losses in power of detection occur. Owing to the high resolution of the echelle spectrometer, DCP – AES is very useful for the analysis of materials with complex emission spectra, such as high-alloy steels, tungsten, molybdenum, and other refractory metals.

6.5.2. Inductively Coupled Plasmas

Inductively coupled plasmas in their present form date back to the work of Reed [229], who used these sources for crystal growth in the 1960s. They were introduced in spectrochemical analysis by Greenfield et al. in 1964 [230] and Wendt and Fassel [231], who used them in their present form as sources for AES.

In inductively coupled plasmas, the energy of a high-frequency generator (6 – 100 MHz) is transferred to a gas flow at atmospheric pressure (usually argon) in a quartz tube system, with the aid of a working coil (Fig. 39). The electrons take up energy and collide with atoms to form a plasma with a temperature of up to 6000 K. At a suitable gas flow, torch geometry, and frequency, a ring-shaped toroidal plasma occurs, where the sample aerosol passes centrally through the hot plasma. The burner consists of three concentric quartz tubes. The aerosol is fed, with its carrier-gas, through the central tube. Between the outer and intermediate tubes, a gas flow is introduced tangentially. It takes up the high-frequency energy, and prevents the torch from melting. The Fassel torch has a diameter of ca. 18 mm and can be operated at 0.6 – 2 kW with 10 – 20 L/min of argon. In so-called minitorches, gas consumption is ca. 6 L/min, by using a special gas inlet or by further reducing the torch dimensions [232]. Greenfield used a larger torch (outer diameter up to 25 mm), in which an ICP can be operated at higher power and where argon, oxygen, or air can be used as outer gas flows (up to 40 L/min) with an intermediate argon flow (ca. 8 L/min). With the Fassel torch, an intermediate gas flow (ca. 1 L/min) is useful when analyzing solutions of high salt content or organic solutions, to prevent salt or carbon deposition.

The ICP is near local thermal equilibrium. Excitation temperatures (from atomic line intensity ratios) are about 6000 K [233] and rotation temperatures (from the rotation lines in the OH bands) are 4000 – 6000 K [234], [235]. From the broadening of the H_β-line, an electron number density of 10^{16} cm^{-3} is obtained, whereas from the intensity ratio of ionic and atomic lines of the same element the electron number density is 10^{14} cm^{-3}. Measured line intensity ratios of ionic to atomic lines were a factor of 100 higher than those calculated for a temperature of 6000 K and an electron number density of 10^{16} cm^{-3}. This indicates the existence of overionization. The excitation processes taking place include:

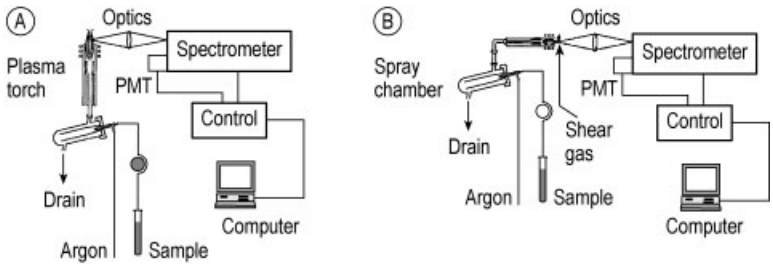

Figure 39. ICP-AES with (A) side and (B) axial viewing of the plasma

1) Excitation through electrons or processes involving radiation:

$$Ar + e \longrightarrow Ar*, Ar^{m} \text{ (metastable levels } E \approx 11.7\,\text{eV})$$

and

$$Ar^{+} \ (E_i = 15.7\,\text{eV})$$
$$Ar + h\nu \longrightarrow Ar^{+} \text{ (radiation trapping)}$$
$$M + e \longrightarrow M*, M^{+}* + e \text{ (electron impact)}$$
$$M^{+} + e \longrightarrow M + h\nu \text{ (radiative recombination)}$$

2) Excitation through argon species:

$$Ar + M \longrightarrow Ar + M*$$
$$Ar^{m} + M \longrightarrow Ar + M^{+} \text{ (Penning effect)}$$
$$Ar^{+} + M \longrightarrow Ar + M^{+}* \text{ (charge transfer)}$$

Accordingly, an overpopulation of the argon metastable levels would explain both the overionization as well as the high electron number density in the ICP. Indeed, it may be that argon metastables act as ionizers, but at the same time are easily ionized [236]. This could explain the rather low interference caused by easily ionized elements, and the fact that ionic lines are excited very efficiently, despite the fact that their standard temperatures are much higher than the plasma temperatures. However, discrepancies are not encountered to such a great extent when the temperatures from the Saha equation are used in the calculations [237]. Nevertheless, a number of processes may be considered to predominate in well-defined zones of the plasma, as indicated by spatially resolved measurements of various plasma parameters [238].

Instrumentation (Fig. 39). Inductively coupled plasma generators are $R-L-C$ circuits producing a high-frequency current at high power. As end-stage a triode or transistor is used. The high-frequency energy can be led to the remote plasma stand by coaxial cable. This requires the use of a matching unit to tune the plasma with the generator before and after ignition. In other systems, the working coil may be an integral part of the oscillator. Generators with a fixed frequency and generators where the frequency may change with load are used. In the first case, a change of load changes the ratio of the forward power and the power reflected to the generator, and requires retuning; in the second case, a small frequency shift occurs. Both types of generator may perform equally, provided the power supply to the plasma, usually ca. 50 % of the energy dissipated in the generator, remains constant. Both the power stability and constant gas flows are essential to obtain maximum analytical precision.

In plasma spectrometry, the sample must be introduced into the plasma as a fine aerosol, usually by pneumatic nebulization. The aerosol should enter the hot plasma zones with low injection velocity for residence times in the millisecond range, required for efficient atomization and excitation. Therefore, the carrier gas flow should not exceed 1 – 2 L/min. With advanced pneumatic nebulizers for ICP spectrometry, aerosols with a mean droplet diameter ca. 5 µm can be obtained with an efficiency of a few percent [90], [91] for a gas flow of 1 – 2 L/min and a sample uptake of 1 – 2 mL/min. The concentric glass Meinhard nebulizer is self-aspirating, the cross-flow nebulizer may have capillaries made of different materials and is pump-fed, but less sensitive to clogging in the case of high-salt solutions. The Babington nebulizer can be used for slurry nebulization [179]. It has been used in ICP–AES for many types of powder analyses, including ceramic powders [125]. The fritted disk nebulizer has not achieved widespread use. New approaches for ICP – AES lie in the use of the jet-impact nebulizer [93], direct injection nebulization [94] (Section 4.1). In pneu-

matic nebulization for ICP – AES, continuous sample feeding requires a sample aspiration time of ca. 30 s, to give a stable signal, a measurement time of some 5 – 10 s, and a minimum rinsing time of 30 s. Discrete sampling is also possible with injection systems developed for flame AAS [78], [79], and by flow injection. Work with sample aliquots down to 10 µL becomes possible – especially useful with microsamples [91], or for the analysis of high-salt solutions [239].

For ICP – AES, sequential, simultaneous, and combined instruments are used. In sequential spectrometers, special attention is given to the speed of wavelength access and, in simultaneous spectrometers, to the provision of background correction. In combined instruments, a number of frequently used channels are fixed and a moving detector or an integrated monochromator gives access to further lines. Echelle instruments allow maximum resolution at low focal length and can be combined with advanced detectors such as CCDs.

Analytical Performance.

Power of Detection. The commonest configuration for an ICP-AES instrument is for the monochromator and detector to view the plasma side-on as shown in Figure 39 A. This means that there is an optimum viewing height in the plasma which yields the maximum signal intensity, lowest background, and least interferences, and it is common practise to optimize this parameter to obtain the best performance from the instrument. Criteria of merit commonly used include the signal-to-noise ratio (SNR), signal-to-background ratio (SBR) and net signal intensity. Different elements, and different analytical lines of the same element, will have different optimum viewing heights so a set of compromise conditions is usually determined which gives satisfactory performance for a selected suite of analytical lines. An alternative configuration which is gaining popularity is end-on viewing of the plasma (Fig. 39 B). This has the advantage that an optimum viewing height does not have to be selected because the analytical signal from the whole length of the central channel is integrated. However, any interferences present are also summed, so that interference effects are expected to be more severe with this configuration.

The nebulizer gas flow and power, and the observation height are also important parameters to optimize. Such multivariate optimizations can be done by single-factor studies or Simplex optimization [240], [241]. The nebulizer gas influences the aerosol droplet size, nebulizer efficiency, and plasma temperature. For each element and line there is a rather sharp optimum for the aerosol gas flow where the SBR is maximum. The power determines mainly the plasma volume and is optimum for soft lines (atomic and ionic lines of elements with low ionization potential) at rather low values (1 – 1.2 kW), whereas for hard lines (lines of elements with high excitation potential and high-energy ionic lines) and for organic solutions, the optimum value is higher (1.5 – 1.7 kW). The optimum observation height is a compromise between the analyte number densities (highest in the lower zones) and completeness of atomization and excitation (a few millimeters above the working coil). Only elements such as the alkali metals are measured in the tail flame. As the analytical lines in ICP – AES have a physical width of 1 – 20 pm and the spectra are complex, a high-resolution spectrometer is required.

The detection limits for most elements are at the 0.05 ng/mL (Mg, Ca, etc.) to 10 ng/mL level (As, Th, etc.). Especially for elements which have refractory oxides (Cr, Nb, Ta, Be, rare earths, etc.), fairly low ionization potential, and sensitive ionic lines, the detection limits are much lower than in AAS. For P, S, N, O, and F, the most sensitive lines are at vacuum wavelengths.

Interference. Most interference in ICP – AES is additive and relates to coincidences of spectral background structures with the elemental lines used. Interference is minimized by the use of high-resolution spectrometers and background correction procedures. In the wavelength range 200 – 400 nm more than 200 000 atomic emission lines of the elements are listed [200] and molecular bands also occur. Accordingly, coincidences often occur within the spectral line widths themselves. In order to facilitate the selection of interference-free spectral lines in ICP – AES, Boumans worked out a system which gives the interference for about 900 prominent lines [205] for spectrometers with a given spectral slit width, in terms of critical concentration ratios. Further, line tables for

ICP – AES [242] and atlases [204], [206] are available, although incomplete. In ICP – AES multiplicative interference also occurs. This may have several causes:

1) Nebulization effects: as discussed earlier, differences in physical properties of the sample solutions lead to variations in the aerosol droplet size and thus in the efficiency of nebulization and sample introduction. This effect is strongest in the case of free sample aspiration and relatively low nebulizer gas flow, and can be minimized (Section 4.1).
2) Easily ionized elements have a complex influence. First they may cause a decrease in excitation temperature, as energy is consumed for ionization. Further, they may shift the ionization equilibrium for partially ionized elements as their easy ionization can influence the electron number densities. They may also cause changes in plasma volume as a result of ambipolar diffusion or they may even quench excited species such as metastables [243]. After correcting for the influence of easily ionized elements on the spectral background, signal enhancements or suppressions are negligible up to concentrations of 500 µg/mL for Na, K, Ca, or Mg. At larger differences in concentration matrix matching of the standards should be applied.

Analytical Precision. As sample introduction in the plasma is very stable, relative standard deviations in ICP – AES are ca. 1 % and the limiting noise is proportional, including flicker and nebulizer noise. Below 250 nm detector noise limitations may occur, especially with high-resolution spectrometers. At integration times beyond 1 s, the full precision is normally achieved.

With the aid of an internal standard, fluctuations in sample delivery and, to a certain extent, changes in excitation conditions can be compensated. To gain the full benefit of internal standardization, the intensities for the analytical and internal standard lines (and in trace determinations that of the spectral background near the analytical line) should be measured simultaneously. The selection of line pairs for ICP – AES has been investigated by Fassel et al. [248] and the background correction by Meyers and Tracy [249]. With internal standardization relative standard deviations below 0.1 % can be obtained.

Multielement Capacity. The multielement capacity of both sequential and ICP – AES is high, as is the linear dynamic range (up to five orders of magnitude). It is limited at the lower end by the limits of determination and at higher concentrations by the onset of self-reversal. This starts at analyte concentrations of ca. 20 000 µg/mL, as analyte atoms hardly occur at all in the cooler plasma zones.

Special Techniques. The analytical performance of ICP – AES can be considerably improved by using alternative techniques for sample introduction.

Organic solutions (methyl isobutyl ketone, xylene, kerosene) can be nebulized pneumatically and determinations in these solvents can be performed directly with ICP – AES [250]. The addition of oxygen or the use of higher power is helpful in avoiding carbon deposits. Oils can be analyzed after dilution 1 : 10 with xylene ($c_L = 0.5 - 1\ \mu$/g) and heavy metals can be extracted from wastewater as complexes with ammonium pyrrolidinedithiocarbonate (APDTC) and separated from more easily ionized elements [251].

Ultrasonic nebulization has been used right from the beginning of ICP – AES [99]. Nebulizer types where the sample liquid flows over the nebulizer transducer crystal and types where ultrasound at 1 MHz is focused through a membrane on the solution have both been used. With aerosol desolvation the power of detection of ICP –AES can be improved by a factor of ten by using ultrasonic nebulization. This applies especially to elements such as As, Cd, and Pb, which are of environmental interest. However, because of the limitations discussed in Section 4.2, the approach is of practical use only for dilute analytes, as in water analysis [98]. New approaches such as thermal spray nebulization [252] and high-pressure nebulization [94] increase analyte introduction efficiency and thus the power of detection, but again only when aerosol desolvation is applied. They are especially interesting for speciation by coupling of ICP – AES to HPLC.

Electrothermal evaporation (ETV) and direct sample insertion increase sampling efficiency so that work with microsamples becomes possible, and the power of detection can be fur-

ther improved. ETC – ICP – AES has been applied with graphite cups [117], graphite furnaces [66], and tungsten filaments [253]. Direct sample insertion enables the direct analysis of waste oils [124], and of microamounts of biological samples [122], as well as the determination of volatile elements in refractory matrices, as shown for Al_2O_3 [121]. Difficulties lie in calibration and in signal acquisition. In trace analysis, the latter necessitates simultaneous, time-resolved measurement of the transient signals for the line and background intensities.

Hydride generation with flow cells [102] decreases the detection limits for As, Se, etc., by 1.5 orders of magnitude down to the ng/mL level – useful in water analysis. However, possible systematic errors from heavy metals and analytes in nonmineralized form remain.

Gas and liquid chromatography coupled to ICP – AES are useful for speciation work [196], [254] , e.g., for Cr^{III}/Cr^{VI} speciation.

Direct solid sampling is of special interest, even though ICP – AES is of use mainly for the analysis of liquids. In a number of cases the same precision and accuracy can be obtained as in solutions, but without the need for time-consuming sample dissolution, involving analyte dilution. For compacted solids, arc/spark ablation is a viable approach in the case of metals [255], [135]. Aerosols with particle sizes at the μm level [137] and detection limits at the μg/g level are obtained [256]. Owing to the separate ablation and excitation, matrix influences are especially low, as shown for aluminum [135] and steels [256]. Laser ablation can be used with similar figures of merit [257], even for electrically insulating samples with lateral resolution (up to ca. 10 μm).

High-frequency discharges at low power and with capacitive power coupling, e.g., stabilized capacitively coupled plasmas (SCP) [255] for element-specific detection in chromatography may be a useful alternative, especially when advanced sample introduction techniques are used.

Applications. Nowadays, ICP – AES is a routine analytical method, of special use when a large number of elements has to be determined in many samples (solutions or dissolved solids). In many cases it is complementary to AAS. ICP – AES is of special interest for the analysis of geological samples, environmental analysis, clinical analysis, food analysis, metals analysis, chemical analysis, and in the certification of reference materials [220], [221].

In the analysis of *geological samples* [243], sample dissolution is important with respect to both nebulization effects and ionization interference.

The analysis of *environmentally relevant samples* is a major field of application [243] and many standard methods have been developed, e.g., by bodies such as the U.S. EPA [244], [245]. This covers procedures for sample preparation and determination of 25 metals in waters, wastes, and biological samples. Details of quality assurance procedures, and interference correction are included.

In *clinical analysis*, Ca, Fe, Cu, Mg, Na, and K can be directly determined in serum samples even of microsize [211]. Electrothermal evaporation and direct sample insertion are particularly useful from this point of view. Generally, however, the power of detection of ICP – AES is too low for most analytical problems encountered in the life sciences.

In the field of industrial products, especially *metals analysis* deserves mention [246]. Apart from solution analysis, direct metal analysis by spark ablation is very useful [256]. Analysis of refractory powders by slurry nebulization, even for ZrO_2 matrices [125] or analysis subsequent to sample decomposition, is very powerful. Laser ablation of solid samples is an important technique which has benefitted from the advent of solid-state detectors [247].

6.5.3. Microwave Plasmas

Microwave plasmas are operated at 1 – 5 GHz. They are mostly produced in a magnetron. Electrons emitted by a glowing cathode are led through a chamber with a series of radially arranged resonance chambers to which an ultrahigh-frequency field is applied. The currents produced, can be coupled to the anode and transported by coaxial cable. The microwave current is led to a waveguide or a cavity. Only when the latter have dimensions below certain critical values can microwave radiation be transported. The transport efficiency can be regulated with the aid of movable walls and screws. In a cavity, a standing wave is produced and the microwave energy is coupled to the resonator by a

loop or antenna; $\lambda/2$, $\lambda/4$, etc., cavities have been proposed [267] and allow tuning with respect to a favorable ratio of forward to reflected power.

The role of microwave plasmas in optical spectrometry has recently been reviewed [268].

Capacitively-Coupled Microwave Plasmas. This type of single-electrode discharge goes back to the work of MAVRODINEANU and HUGHES [269]. The microwave radiation is led to a pin-shaped electrode, and the surrounding burner housing acts as counter-electrode. A bush-shaped discharge burns at the top of the electrode which enters the sample aerosol concentrically. The plasma, which can be operated with 2 – 4 L/min argon, helium or nitrogen (at 400 –800 W), is not in local thermal equilibrium (T_{ex} = 5000 – 8000 K, T_g = 3000 – 5700 K, n_e = 10^{14}) and it is a current-carrying plasma. As the sample enters, with fairly low efficiency, into a hot zone where the background emission is also high, the detection limits are poor, especially for elements with thermally stable oxides [270], e.g., Na 589 nm, 0.05 µg/mL; Be 234.9 nm, 0.006 µg/mL; Mg 285.2 nm, 0.04 µg/mL; Al 396.1 nm, 0.16 µg/mL. Alkali and alkaline-earth metals cause high matrix effects; e.g., for 500 µg/mL they amount to 100 % in the case of Mg. Therefore, spectrochemical buffers such as CsCl are often used. Because it can work with He, the source is also useful for the speciation of widely diverse elements [271].

Microwave-Induced Plasmas. These electrodeless discharges are operated in a noble gas at rather low power (usually < 200 W). They became important as sources for atomic spectrometry when BEENAKKER et al. [272] succeeded in operating a stable plasma at atmospheric pressure with less than 1 L/min argon or helium and power below 100 W (Fig. 40). The so-called TM_{010} resonator is used. It has a cylindrical resonant cavity (diameter ca. 10 cm) where the power is introduced with a loop or antenna and the filament plasma is contained in a centrally located capillary. At higher power toroidal argon [273] and diffuse helium [274] discharges can be obtained. The plasma is not in local thermal equilibrium and metastable argon or helium species are thought to contribute to the excitation. MOUSSANDA et al. [275] studied the spectroscopic properties of various microwave-induced plasma (MIP) discharges and reported rotational temperatures of ca. 2000 K. These plasmas are very useful for coupling with electrothermal evaporation, where detection limits in the upper picogram range can be obtained. This can be achieved with both graphite furnace [81] and tungsten filament atomizers [115]. Excitation of desolvated aerosols is also possible. Wet aerosols can only be taken up with the toroidal argon MIP. This allows element-specific detection in liquid chromatography [276]. Helium MIPs are excellent for element-specific detection in gas chromatography, and are commercially available [277]. Not only the halogens and other elements relevant in pesticide residue analysis, but also organolead and organotin compounds can be determined down to low concentrations. This makes MIP – AES very useful for speciation work [197]. The MIP is also useful for excitation of volatile hydride-forming elements after removal of excess hydrogen. With different trapping techniques detection limits down to the sub-nanogram level are easily obtained [104]. A further useful type of MIP is the so-called surfatron described by HUBERT et al. [278] (Fig. 40 B). Here, the power is introduced by a perpendicular antenna directed near the plasma capillary; a plasma is formed by microwave coupling through a slit and power propagates along the surface of the discharge. Compared with the MIP obtained in a TM_{010} resonator this plasma is more stable and allows more variability in sampling, as shown by direct comparative studies with the same tungsten filament atomizer [279]. The MIP torch described by QUEHAN JIN et al. [280] can operate both argon and helium discharges and make use of hydride generation, electrothermal evaporation, etc.

A strip-line microwave plasma has been described by BARNES and RESZKE [281], [282]. A particular advantage of this design is that it is possible to sustain more than one plasma in the cavity at the same time. Such a plasma has compared favorably with other designs [282], though it has not yet been widely used.

6.6. Glow Discharges

Glow discharges have long been recognized as unique sources for AES [283]. Their special fea-

tures relate to the possibility of analyte introduction by sputtering as well as to the advantages of plasmas not in local thermal equilibrium, enabling the discrete excitation of terms, even those of high energy. These advantages are realized in the hollow cathode and related sources, in glow discharges with flat cathodes, with d.c. or r.f. powering, and in special sources such as gas-sampling glow discharges.

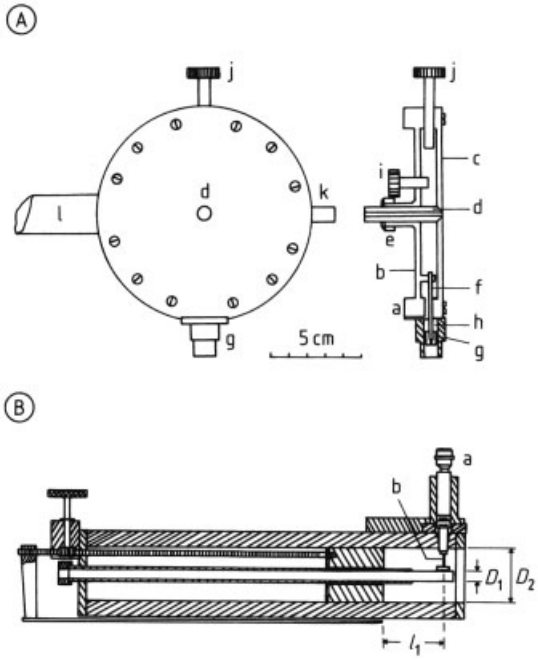

Figure 40. Microwave-induced plasma discharges
A) MIP in a TM_{010} resonator according to BEENAKKER [272]
a) Cylindrical wall; b) Fixed bottom; c) Removable lid; d) Discharge tube; e) Holder; f) Coupling loop; g) PTFE insulator; h) Microwave connector; i, j) Tuning screw; k) Fine-tuning stub; l) Holder
B) Surfatron MIP, original version 20 mm $< l_1 <$ 250 mm, D_1 11 mm, D_2 40 mm, cut-off frequency 2 GHz [278]
a) Standard coaxial connector; b) Coupler

6.6.1. Hollow Cathodes

Here, the energetically important parts of the discharge (cathode layer, dark space, and negative glow) as well as the sample, are inside the cathode cavity. The volatilization results from cathodic sputtering and/or thermal evaporation and the analyte is excited in the negative glow. The characteristic depends on the discharge gas (at 100 – 500 Pa Ar, $i = 0.2 - 2$ A, $V = 1 - 2$ kV; at 1 – 2 kPa He, $i = 0.2 - 2$ A, $V < 1$ kV), and also on the cathode mounting. In a cooled cathode, the characteristic is abnormal and the analyte volatilizes by cathodic sputtering only, whereas in a hot hollow cathode, thermal evaporation also takes place, and the characteristic becomes normal especially at high currents.

Owing to the long residence times in the excitation zone and the high line-to-background ratios, resulting from the nonequilibrium character of the plasma (gas temperatures of 600 –1000 °C [19] and much higher electron temperatures), the hollow cathode is the radiation source with the lowest absolute detection limits. For this reason, it is widely used for dry solution residue analysis, which is useful subsequent to matrix separation in the analysis of high-purity substances [284]. The Doppler widths of the lines are low and this source can even be used for isotopic analysis (determination of $^{235}U/^{238}U$). It is still employed for the determination of volatile elements in a nonvolatile matrix, such as As, or Sb in high-temperature alloys [285], where detection limits below 1 µg/g can be obtained. A treatment of the hollow cathode lamp as an AES source is given in [286].

6.6.2. Furnace Emission Spectrometry

Similar to hollow cathodes, electrically heated graphite furnaces in a low-pressure environment can be used for sample volatilization. The released analyte can then be excited in a discharge between the furnace and a remote anode. This source is known as furnace atomic nonresonance emission spectrometry (FANES), and was introduced by FALK et al. [287]. Owing to the separation of volatilization and excitation, its absolute detection limits are in the picogram range. For real samples, however, volatilization and excitation interference may be considerable.

6.6.3. Glow Discharges with Flat Cathodes

With this source, which is normally a d.c. discharge, compact electrically conducting samples can be analyzed. They must be flat, are used as the cathode, and ablated by cathodic sputtering. The sputtered material is excited in the negative

glow of the discharge, which is usually operated at 100 – 500 Pa of argon. As the sample is ablated layer-by-layer, both bulk and in-depth profiling analysis are possible.

Glow Discharge Sources. A useful device was first described by GRIMM in 1968 [288]. The anode is a tube (diameter 8 – 10 mm) with the distance from its outer side to the cathode block and the sample less than the mean free path of the electrons (ca. 0.1 mm). The anode – cathode interspace is kept at a lower pressure than the discharge lamp. According to the Paschen curve, no discharge can take place at these locations, and the glow discharge is restricted to the cathode surface (Fig. 41). In later versions, a lamp with floating restrictor and remote anode is used, so that no second vacuum is necessary. The Grimm lamp has an abnormal characteristic, the current is set at 40 – 200 mA and the burning voltage is usually below 1 kV. In both krypton and neon, the burning voltage is higher than in argon, but it also depends on the cathode material. The ablated material is deposited both at the edge of the burning crater, limiting the discharge time to a few minutes, and inside the anode tube, making regular rinsing necessary. In AES, the glow discharge is viewed end on.

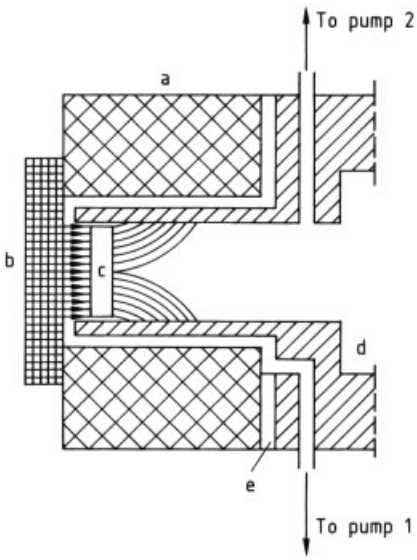

Figure 41. Classical GRIMM glow discharge lamp [149] a) Cathode body; b) Sample; c) Negative glow; d) Anode body; e) PFTE sheath (0.2 mm thick)

Characteristics. After ignition of the discharge, a burn-in time is required to achieve sputtering equilibrium. First, the surface layers have to be removed, but once equilibrium is reached the discharge penetrates into the sample at constant velocity. Burn-in times are usually up to 30 s for metals (at 90 W in argon, zinc 6 s; brass 3 – 5 s; steel 20 s; Al 40 s), but depend on the sputtering conditions (shorter at high voltage, low pressure, etc.). The burning crater has a structure depending on the material structure or inclusions, and the electrical field may induce a small curvature, especially in the classical Grimm lamp [289]. Material volatilization is of the order of mg/min and increases in the sequence: C < Al < Fe < steel < Cu <brass < Zn. It also depends on the gas used, normally a noble gas. Helium is not used; because of its low mass sputtering is too limited and the sputtering rate increase in the sequence: neon < argon < krypton.

Sputtering is proportional to $1/\sqrt{p}$ [151], where p is the gas pressure. The sputtering rates of alloys can be predicted to a certain extent from the values for the pure metals [149].

For excitation, collision with high-energy electrons, but also other processes such as charge transfer or excitation by metastables are important. Electron temperatures are 5000 K (slow electrons for recombination) and > 10 000 K (high-energy electrons for excitation), and gas temperatures are below 1000 K, so line widths are low. Owing to the low spectral background intensities, the limits of detection arising from the source are low and, for metals, range down to µg/g in a classical Grimm lamp [290]. Elements such as B, P, S, C and As can also be determined. The plasma is optically thin, but for resonance lines of matrix elements, self-reversal may occur. However, the linear dynamic range is high compared with arc or spark sources.

Owing to the stable nature of the discharge, noise is low. There is hardly any flicker noise and only a few frequency-dependent components from the vacuum line are superimposed on a white noise background [44]. Relative standard deviations with no internal standard can be below 1 %. However, the low radiant densities may lead to shot noise limitations. Because of its stable nature, the glow discharge source can be coupled with Fourier transform spectrometry. As sample volatilization is due to cathodic sput-

tering only, matrix interference as a result of the thermochemical properties of the elements do not occur.

Applications. Glow discharge AES is widely used for *bulk analysis of metals.* This applies less to production control (the burn-in time required is long compared with sparks) than to product control. Here, the easy calibration due to the absence of volatilization interference and low spectral interference, as well as the high linear dynamic range are advantageous. For elements with high sputtering rates, such as copper, the method is especially useful. Samples have to be available as flat and vacuum-tight disks (thickness 10 – 30 mm), but with special holders threads and metal sheets can be analyzed. The samples must be polished and freed from oxide layers prior to analysis.

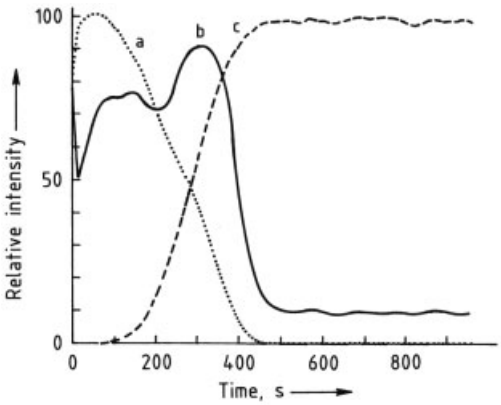

Figure 42. Depth profile through a galvanized steel sheet surface. The Zn coating (a) is ca. 18 µm thick, and the Al (b) concentration in the Fe bulk (c) is 0.049 %, the voltage increases from 510 V in the Zn to 740 V in the bulk steel [291]

Depth-profile analysis is also possible, as the sample is ablated layer-by-layer with a penetration rate of 1 – 3 µm/min. Here, the intensities of the analyte lines are measured as a function of time. However, the sputtering rates of alloys with varying composition must be known to convert the time scale into a depth scale. The intensities must be related to concentrations, which can be done by using theoretical models and sputtering constants [291]. The power of detection may be quite good and depth resolution is of the order of 5 nm, when elemental concentrations > 0.1 % are monitored. Depth profile analysis is now a main field of application of glow discharge emission spectrometry in the metals industry (Fig. 42) [291].

Electrically insulating powders may be analyzed after mixing with Cu powder and pelleting [134]. Performance depends greatly on the particle size of the powders, but detection limits in the µg/g range are possible for the light elements (Al, Be, B, etc.).

6.6.4. New Developments

Since it was found that a great part of the analyte released by sputtering consists of ground-state atoms, cross-excitation by d.c. [292], high-frequency [293], and microwave discharges [290] was investigated. The latter was found to improve the power of detection by up to a factor of ten, e.g., for steels. Further, it was found that simple types of d.c. discharges were capable of exciting vapors and even molecular species such as volatile hydrides [294].

Radio-frequency glow discharges are very useful sources for atomic spectrometry. By means of a bias potential in the vicinity of the sample, insulating samples such as ceramics can be directly ablated and analyzed by AES [295].

6.7. Laser Sources (→ Laser Analytical Spectroscopy)

High-power lasers have proved useful sources for the direct ablation of solids. In AES, ruby and Nd – YAG lasers have been used since the 1970s for solid ablation. When laser radiation interacts with a solid, a laser plume is formed. This is a dense plasma, containing both atomized material and small solid particles, evaporated from the sample. The processes occurring and the figures of merit in terms of ablation rate, crater diameter (ca. 10 µm), depth at various energies, and the expansion velocities of the plume are well known. The plume is optically thick and seems of little interest as an emission source. By applying cross-excitation with a spark or microwave energy, however, the power of detection is considerably improved [141], [142]. Absolute detection limits go down to 10^{-15} g, but relative standard deviations are no better than 30 %. The method, however, can be applied for identification of mineral inclusions and microdomains in

metallurgical analysis, forensic samples, archeology, and remote analysis in the nuclear industry.

The method has evoked renewed interest, mainly due to the availability of improved Nd–YAG laser systems. Different types of detector, such as microchannel plates coupled to photodiodes and CCDs in combination with multichannel analyzers, make it possible to record an analytical line and an internal standard line simultaneously, so that analytical precision is considerably improved. By optimizing the ablation conditions and the spectral observations, detection limits obtained with the laser plume as source for AES are in the range 50–100 µg/g with standard deviations ca. 1 % [142].

Manufacturers. Equipment for atomic emission spectrometry now is available from many different manufacturers. Several types of sources – arc, spark, ICP, MIP, glow discharge (GD), etc. – are available:

Agilent Technologies (formerly Hewlett Packard), USA: ICP-MS, MIP-AES
Baird Atomic, USA: arc/spark AES, ICP-AES, AAS
Analytic Jena AG, Germany: AAS
Hitachi, Japan: ICP-AES, AAS, ICP-MS
Instruments SA/Jobin Yvon Emission, France: ICP-AES, GD-AES
LECO Corporation, USA: TOF-ICP-MS, GD-AES
Leeman Labs Inc., USA: ICP-AES
PerkinElmer Instruments (incorporating Princeton Applied Research, ORTEC, Signal Recovery and Berthold products), USA: AAS, ICP-AES, ICP-MS, GC, GC/MS
Shimadzu, Japan: arc/spark AES, ICP-AES, AAS Spectro Analytical Instruments, Germany: ICP-AES, arc/spark AES, ICP-MS
TJA Solutions (incorporating Thermo Jarrell Ash, VG Elemental and Unicam AA), USA: AAS, ICP-AES, ICP-MS, SF-ICP-MS, GD-AES, GD-MS
Varian Associates, Australia: ICP-AES, ICP-MS, AAS

7. Plasma Mass Spectrometry

The plasma sources developed for AES have also proved very suitable ion sources for mass spectrometry. This is especially true for electrical discharges at pressures in the 100 – 500 Pa range and sources at atmospheric pressure with the advent of powerful vacuum systems with which to bridge the pressure difference between the mass spectrometer (of the order of 10^{-4} Pa) and the source. Elemental mass spectrometry, however, goes back to the use of high-vacuum arcs and sparks, for ultratrace and survey analyses of metal samples [50]. Spark source mass spectrography, with high-resolution sector field mass spectrometers, though still useful for characterization of electrically conducting solids down to the ng/g level, needs expensive instruments, and the analytical precision is low. In this chapter, only the newer mass spectrometric methods, which make use of plasma sources from optical atomic spectrometry are treated, namely, ICP mass spectrometry (ICP – MS) and glow discharge mass spectrometry (GD – MS). For a detailed treatment of mass spectrometric methods see → Mass Spectrometry.

7.1. Principles

Plasma mass spectrometry developed on the basis of the experience gathered from the extraction of ions from flames, as performed for diagnostic purposes in the mid-1960s [296]. In 1980 Gray, Houk, and coworkers [297] first showed the analytical possibilities of ICP – MS. The method has the same ease of sample introduction as ICP – AES, but has much lower detection limits, covers more elements, and allows the determination of isotope concentrations; ICP – MS rapidly became an established analytical method.

The principles and practice of ICP – MS have been treated in a number of excellent texts [220], [298], [299].

7.1.1. Instrumentation (Fig. 43)

In ICP – MS, the ions formed in the ICP are extracted with the aid of a sampler into the first vacuum stage, where a pressure of the order of 100 Pa is maintained. Here, a supersonic beam is formed and a number of collision processes take place as well as an adiabatic expansion. These

processes, together with their influence on the ion trajectories in the interface and behind the second aperture (skimmer), are very important for the transmission of ions and for related matrix interference [300]. Today's commercially available instrumentation falls into two groups: quadrupole and high-resolution instruments.

7.1.2. Analytical Features

The analytical features of ICP – MS are related to the highly sensitive ion detection and the nature of the mass spectra themselves.

ICP Mass Spectra. In low-resolution ICP – MS, resolution is at best 1 atomic mass units and, especially in the lower mass range, signals from cluster ions occur, which may cause spectral interference with analyte ions [301]. Cluster ions may be formed from different types of compounds:

1) Solvents and acids: H^+, OH^+, H_2O^+, NO^+, NO^{2+}, Cl^+, SO^+, SO_2^+, SO_3H^+ (when residual H_2SO_4 is present)
2) Radicals from gases in the atmosphere: O_2^+, CO^+, CO_2^+, N_2^+, NH^+, NO^+
3) Reaction products of the above-mentioned species with argon: ArO^+, $ArOH^+$, $ArCl^+$, Ar_2^+

Cluster ions give strong signals in the mass range up to 80 atomic mass units and may hamper the determination of light elements as a result of spectral interference. At higher masses, singly charged ions of the heavier elements occur, as well as doubly charged ions of the light elements, especially those with low ionization energies.

Signals from compounds of the analyte with various other species occur, such as MO^+, MCl^+, MOH^+, MOH_2^+. These ions may be formed by the dissociation of nitrates, sulfates, or phosphates in the plasma, or they may result from reactions between analyte ions and solvent residuals or oxygen in the plasma, or at the interface between the plasma and the mass spectrometer. Selection of the sampling zone in the plasma and the aerosol carrier gas flow may strongly influence the signal intensities of cluster ions and doubly charged analyte ions [302], [303].

For each element, signals from its different isotopes are obtained. Their intensity ratios correspond to the isotopic abundances in the sample. This fact can be used in calibration by isotopic dilution with stable isotopes, and in tracer experiments. Isotopic abundance is also useful for the recognition of spectral interference. The spectra in ICP – MS are simpler than in ICP – AES.

In ICP – MS, the background intensities are mostly low and are mainly due to the dark current of the detector, and to signals produced by ions scattered inside the mass spectrometer as a result of collisions with residual gas species or reflection. In order to shield the direct UV radiation, a beam stop is supplied often in the mass spectrometer. The plasma contributes little to the background, in contrast to ICP – AES where the background continuum stems from interactions between electrons and ions in the plasma, from molecular bands, from the wings of broad matrix lines, and from stray radiation.

A particular problem with quadrupole instruments is that they have insufficient resolution to resolve molecular ion interferences. Resolution is defined as:

$$R = \frac{M}{\Delta M} \tag{172a}$$

where R is resolution; M is mass (strictly speaking m/z); and ΔM is peak width at 5 % peak height. The resolution obtainable with quadrupoles is limited by the stabiliy and uniformity of the r.f./d.c. field, and by the spread in ion energies of the ions. Quadrupoles used in ICP – MS are typically operated at resolutions between 12 and 350, depending on m/z, which corresponds to peak widths of between 0.7 and 0.8. In comparison, sector field instruments (Fig. 44) are capable of resolution up to 10 000, resulting in peak widths of 0.008 at $m/z = 80$. For most applications the resolution provided by a quadrupole is sufficient, but for applications where spectroscopic interferences cause a major problem, the resolution afforded by a magnetic sector may be desirable. Table 5 gives examples of some common spectroscopic interferences that may be encountered, and the resolution required to separate the element of interest from the interference. For example, a particular problem is the determination of arsenic in a matrix which contains chloride (a common

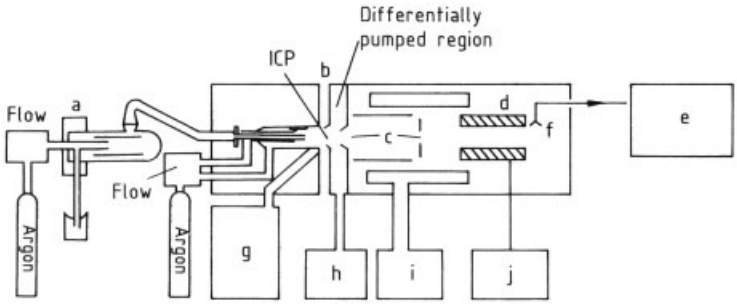

Figure 43. Instrumentation for low-resolution ICP–MS
a) Nebulizer; b) Sampler, skimmer; c) Ion lenses; d) Quadrupole mass filter; e) Computer; f) Electron multiplier; g) Radiofrequency supply; h) Mechanical pump; i) Cryosystem; f) Quadrupole supply

component of most biological or environmental samples). Arsenic is monoisotopic with $m/z = 75$ and the chloride matrix gives rise to an interference at $m/z = 75$ due to $^{40}Ar^{35}Cl^{+}$, so an alternative isotope is not available for analysis. A quadrupole has insufficient resolution to separate the two species but a magnetic sector could do so easily (Table 5). The ion transmission of sector field instruments is much greater, compared with a quadrupole, for comparable resolution, so sensitivity is much higher and detection limits are lower by three orders of magnitude or so (i.e., detection limits of $1-10\,\mathrm{fg\,ml^{-1}}$ providing sufficiently clean reagents are available). Also, simultaneous isotopic detection is possible using an array of detectors, so extremely precise isotope ratio measurements can be made, allowing application of the technique for geochronology.

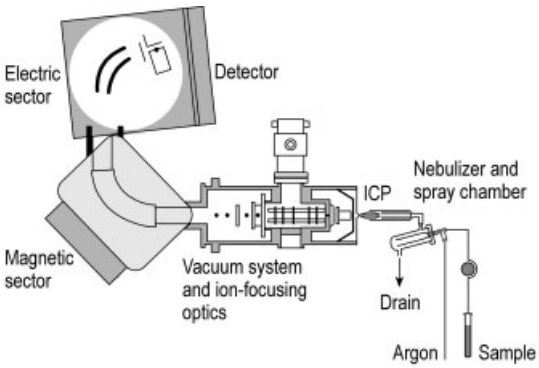

Figure 44. A sector-field ICP – MS instrument

Recently, ICP time-of-flight (ICP – TOF) mass spectrometers have been commercialized (Figure 45). The TOF analyser has a number of advantages over other analyzers used for ICP – MS. The fact that all masses in the spectrum can be monitored simultaneously is the main attraction for many ICP – MS users. Even for simple multielement analysis, scanning analyzers suffer from a decrease in sensitivity as the number of ions to be monitored is increased. The fact that over 30 000 simultaneous mass spectra per second can be obtained by the TOF analyzer has many advantages for monitoring transient signals, such as those produced by laser ablation, chromatographic, capillary electrophoresis, and electrothermal vaporization sample introduction techniques. Also, improved isotope ratio measurements have been reported owing to the elimination of noise due to temporal fluctuations in the plasma and sample introduction systems.

Optimization. For the optimization of ICP – MS with respect to maximum power of detection, minimal spectral interference, signal enhancement or depression, and maximum precision, the most important parameters are the power of the ICP, its gas flows (especially the nebulizer gas), the plasma, the position of the sampler, and the ion optical parameters. These parameters determine the ion yield and the transmission, and thus the intensities of analyte and interference signals. At increasing nebulizer gas flow, the droplet size decreases (Section 4.1) and thus the analyte introduction efficiency goes up, but at the expense of the residence time in the plasma, the plasma temperature, and the ionization [304]. However, changes of the nebulizer gas flow also influence the formation and breakdown of cluster ions as well as the ion energies (as shown for $^{63}Cu^{+}$ and ArO^{+} [305]), and the

Table 5. Examples of spectroscopic interferences caused by molecular ions, and the resolution that would be necessary to separate the analyte and interference peaks in the mass spectrum

Analyte ion		Interfering ion		
Nominal isotope	Accurate *m/z*	Nominal isotope	Accurate *m/z*	Resolution required
^{51}V	50.9405	$^{16}O^{35}Cl$	50.9637	2580
^{56}Fe	55.9349	$^{40}Ar^{16}O$	55.9572	2510
^{63}Cu	62.9295	$^{40}Ar^{23}Na$	62.9521	2778
^{75}As	74.9216	$^{40}Ar^{35}Cl$	74.9312	7771

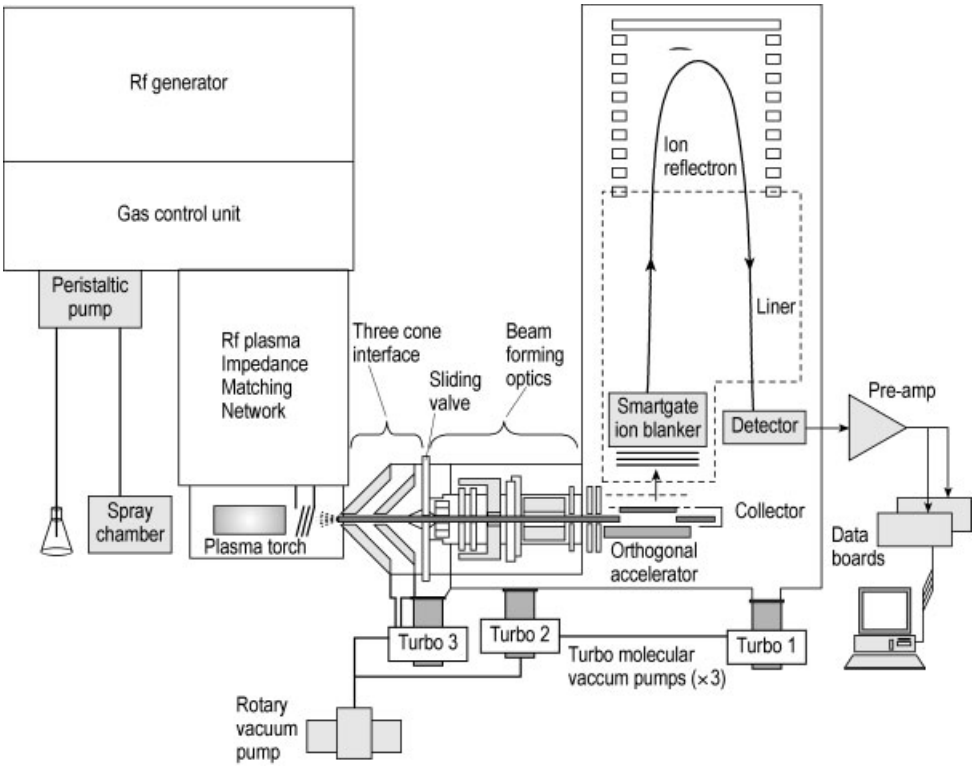

Figure 45. Orthogonal TOF-ICP-MS instrument

geometry of the aerosol channel. Normally, the aerosol gas flow is 0.5 – 1.5 L/min. It must be optimized together with the power, which influences the plasma volume, the kinetics of the different processes taking place, and the position of the sampler. By changing the voltages of the different ion lenses, the transmission for a given ion can be optimized, enabling optimization of its detection limit and minimization of interference. In multielement determinations, a compromise must always be selected.

Power of Detection. For optimum power of detection, the analyte density in the plasma, the ionization, and the ion transmission must be maximized. The necessary power is 0.6 – 2 kW with the sample ca. 10 – 15 mm above the tip of the injector. The detection limits, obtained at single element optimum conditions, differ considerably from those at compromise conditions, but are still considerably lower than in ICP – AES (Table 6). For most elements they are in the same range, but for some they are limited by spectral interference. This applies to As ($^{75}As^+$ with $^{40}Ar^{35}Cl^+$), Se ($^{80}Se^+$ with $^{40}Ar^{40}Ar^+$), and Fe ($^{56}Fe^+$ with $^{40}Ar^{16}O^+$). The acids present in the measurement solution and the material of which the sampler is made (Ni, Cu, etc.) may have

considerable influence on these sources of interference and the detection limits for a number of elements. The detection limits for elements with high ionization potential may be even lower when they are detected as negative ions (for Cl^+ $c_L = 5$ ng/mL and for Cl^- $c_L = 1$ ng/mL).

Table 6. Detection limits in ICP – AES and quadrupole ICP – MS

Element	Detection limit, ng/mL	
	ICP – AES [204]	ICP – MS [156]
Ag	7	0.005
Al	20	0.05
As	50	0.005
Au	20	0.005
B	5	0.05
Cd	3	0.005
Ce	50	0.001
Co	6	0.001
Cr	6	0.005
Fe	5	0.05
Ge	50	0.05
Hg	20	0.001
In	60	0.001
La	10	0.005
Li	80	0.001
Mg	0.1	0.05
Mn	1	0.001
Ni	10	0.005
Pb	40	0.001
Se	70	0.05
Te	40	0.05
Th	60	0.001
Ti	4	0.05
U	250	0.001
W	30	0.005
Zn	2	0.005

Precision and Memory Effects. The constancy of the nebulizer gas flow is of prime importance for the precision achievable in ICP – MS. After stabilizing the nebulizer gas flows, relative standard deviations may be below 1 %. They can be further improved by internal standardization [306]. The tolerable salt concentrations (1 – 5 g/L) are much lower than in ICP – AES. Memory effects are low and 1 – 2 min rinsing times are appropriate.

Interference. ICP – MS suffers from two principal types of interference [301]. Spectroscopic interferences arise when an interfering species has the same nominal *m/z* as the analyte of interest. The interfering species can be either an isotope of another element (which are well documented and hence easily accounted for), or a molecular ion formed between elements in the sample matrix, plasma gas, water, and entrained atmospheric gases. The molecular ions are less easy to correct for since they will vary depending on the nature of the sample matrix. Some common molecular ion interferences are shown in Table 7. Many of these interferences can be overcome by choosing an alternative isotope of the analyte which is free from interference, though a sacrifice in sensitivity may result. If a "clean" isotope is not available then one recourse is to separate the analyte from the matrix before the analysis by chemical extraction and chromatography, or use a magnetic sector instrument which is capable of resolving the interfering species from the analyte. Many of these interferences are thought to be formed in the interface due to a secondary discharge. This discharge can be eliminated by operating the plasma at low power, typically 600 W, and modifying the torch by inserting a grounded shield between it and the coil, or by using a center-tapped r.f. coil. Under these so-called cool-plasma conditions interferences due to ArO^+, for example, can be eliminated, though the sensitivity for refractory elements and those with high ionization potentials may be reduced. An alternative method of reducing molecular ions is to use a collision cell prior to the quadrupole. A collision gas, such as helium, ammonia, or water vapor, is introduced into the cell, where it either reacts with, collisionally dissociates, or neutralizes the polyatomic species or precursors. It remains to be seen whether new interferences, formed from the matrix species and the gas, outweigh the benefits.

Table 7. Some commonly occurring spectroscopic interferences caused by molecular ions derived from plasma gases, air, and water, and the sample matrix

Molecular ion interference	Analyte ion interfered with	Nominal *m/z*
O_2^+	S^+	32
N_2^+	Si^+	28
NO^+	Si^+	30
NOH^+	P^+	31
Ar^+	Ca^+	40
ArH^+	K^+	41
ClO^+	V^+	51
CaO^+	Fe^+	56
ArO^+	Fe^+	56
ArN^+	Cr^+, Fe^+	54
$NaAr^+$	Cu^+	63
Ar_2^+	Se^+	80

Nonspectroscopic interferences are caused by the sample matrix, and are manifested as an

apparent enhancement or suppression in the analyte signal in the presence of a concentrated sample matrix. Such effects are thought to be caused primarily by space charge in the ion beam, whereby positive analyte ions are repelled from the ion beam by the high poisitive charge of the matrix ions, with low mass ions being relatively more affected than high mass ions. Such interferences are usually compensated for by using an internal standard, or by separating the matrix from the analyte before analysis. Because ICP – MS has such low detection limits it is also possible to dilute the sample to such an extent that the interference becomes negligible.

Isotope Ratio Analysis. Isotope ratio measurements are performed whenever the exact ratio, or abundance, of two or more isotopes of an element must be known. For example, the isotopic ratios of lead are known to vary around the world, so it is possible to determine the source of lead in paint, bullets, and petrol by knowing the isotopic abundances of the four lead isotopes 204, 206, 207, and 208. Another example is the use of stable isotopes as metabolic tracers, where an animal is both fed and injected with an element having artificially enriched isotopes and the fractional absorption of the element can be accurately determined. An important field is that of geological dating, where extremely precise isotope ratio measurements must be made.

To perform the isotope ratio experiment correctly it is necessary to compensate for a number of biases in the instrumentation. Mass spectrometers and their associated ion optics do not transmit ions of different mass equally. In other words, if an elemental solution composed of two isotopes with an exactly 1 : 1 molar ratio is analyzed by ICP – MS, then a 1 : 1 isotopic ratio will not necessarily be observed. This so-called mass bias will differ depending on mass, and even very small mass biases can have deleterious effects on the accuracy of isotope ratio determinations, so a correction must always be made by using an isotopic standard of known composition, as shown in Equation (173):

$$C = \frac{R_\mathrm{t}}{R_\mathrm{o}} \tag{173}$$

where C is the mass bias correction factor, R_t is the true isotopic ratio for the isotope pair, and R_o the observed isotope ratio for the isotope pair. For best results, such a correction should be applied to each individual isotopic pair that is to be ratioed, though this is not always possible in practice since a large number of isotopic standards would be needed to cover every eventuality. Also important is the effect of detector dead time. When ions are detected using a pulse counting (PC) detector the resultant electronic pulses are approximately 10 ns long. During and after each pulse there is a period of time during which the detector is effectively "dead" (i.e., it cannot detect any ions). The dead (D) time is made up of the time for each pulse and recovery time for the detector and associated electronics. Typical dead times vary between 20 and 100 ns. If dead time is not taken into account there will be an apparent reduction in the number of pulses at high count rates, which would cause an inaccuracy in the measurement of isotope ratios when abundances differ markedly. However, a correction can be applied as shown in Equation (174):

$$R_\mathrm{t} = \frac{R_\mathrm{o}}{1 - R_\mathrm{o} \cdot D} \tag{174}$$

where R_t is the true count rate and R_o is the observed count rate. The relative abundances of the isotopes and operating conditions will affect the precision of any isotope ratio measurement. The rate at which the mass spectrometer scans or hops between masses the time spent monitoring each mass (the dwell time), and the total time spent acquiring data (total counting time) are important parameters which affect precision. The best precision achievable with a quadrupole instrument is ca. 0.1 % for a 1 : 1 isotope ratio at high count rate. This can be reduced by at least an order of magnitude if a magnetic sector instrument with an array of detectors is used to facilitate simultaneous detection. A long total counting time is desirable, because the precision is ultimately limited by counting statistics and the more ions that can be detected the better.

Isotope Dilution Analysis. A result of the ability to measure isotope ratios with ICP – MS is the technique known as isotope dilution analysis. This is done by spiking the sample to be analyzed with a known concentration of an enriched isotopic standard, and the isotope ratio is measured by mass spectrometry. The observed isotope ratio R_m of the two chosen isotopes can

then be used in the isotope dilution equation (Eq. 174a) to calculate the concentration of the element in the sample:

$$C_x = \frac{C_s W_s M_x}{W_x M_s} \times \frac{A_s - R_m B_s}{R_m B_x - A_x} \quad (174a)$$

where R_m is the observed isotope ratio of A to B; A_x and B_x the atom fractions of isotopes A and B, respectively, in the sample; A_s and B_s the atom fractions of isotopes A and B, respectively, in the spike; W_x the weight of sample; W_s the weight of spike; C_x and C_s the concentration of element in sample and spike respectively; M_x and M_s the molar mass of element in sample and spike, respectively. The best precision is obtained for isotope ratios near one, but if the element to be determined is near the detection limit then the ratio of spike isotope to natural isotope should be between 3 and 10 so that noise contributes only to the uncertainty of natural isotope measurement. Errors also become large when the isotope ratio in the spiked sample approaches the ratio of the isotopes in the spike (overspiking), or the ratio of the isotopes in the sample (underspiking). The accuracy and precision of the isotope dilution analysis ultimately depends on the accuracy and precision of the isotope ratio measurement, so all the precautions that apply to isotope ratio analysis also apply in this case. Isotope dilution analysis is attractive because it can provide very accurate and precise results. The analyte acts as its own de facto internal standard. For instance, if the isotopic spike is added prior to any sample preparation then the spike will behave in exactly the same way as the analyte because it is chemically identical, providing it has been properly equilibrated by a series of oxidation/reductions. Hence, any losses from the sample can be accounted for because the analyte and spike will be equally affected.

Alternative Methods for Sample Introduction. Apart from continuous pneumatic nebulization, all sample introduction techniques known from ICP – AES have been used for ICP – MS. Similar to ICP – AES, the analysis of organic solutions is somewhat more difficult [308].

The use of *ultrasonic nebulization* as in ICP – AES increases sampling efficiency. The high water loading of the plasma has to be avoided by desolvation, not only to limit the cooling of the ICP, but also to minimize the formation of cluster ions and the related spectral interference and it should also be borne in mind that increased matrix loading in the plasma will occur. The use of high-pressure nebulization in ICP – MS has similar advantages [309], and is suitable for coupling ICP – MS to HPLC.

With the formation of *volatile hydrides*, the detection limits for elements such as As, Se, Sb, and Pb can be improved [310]. The improvement is due to enhanced analyte sampling efficiency and to the decrease of cluster ion formation.

Electrothermal evaporation (ETE), in addition to its advantages for microanalysis and improving sample transport, has the additional advantage in ICP – MS of introducing a dry analyte vapor into the plasma. Accordingly, it has been found useful for elements of which the detection limits are high as a result of spectral interference with cluster ions. In the case of ^{56}Fe, which is interfered by $^{40}ArO^+$, Park et al. [311] showed that the detection limit could be considerably improved by ETE. For similar reasons, the direct insertion of samples in ICP – MS maximizes the absolute power of detection [312], [313].

Although ICP – MS is mainly a method for the analysis of liquids and dissolved solids, techniques for *direct solid sampling* have been used [142]. They are especially needed when samples are difficult to dissolve or are electrically insulating and thus difficult to analyze with glow discharge or spark techniques, or spatial analysis over the sample surface is required. In particular, the use of laser ablation for geological applications with UV or frequency-doubled lasers is attracting considerable interest [314], [315]. For powders such as coal, slurry nebulization with a Babington nebulizer can be used in ICP – MS as well [316]. For the direct analysis of metals, spark ablation can be applied and the detection limits are in the ng/g range, as shown for steels [317]. For the analysis of metals [318] as well as of electrically insulating samples, laser ablation combined with ICP – MS is very useful. A Nd – YAG laser with a repetition rate of 1 – 10 pulses per second and an energy of ca. 0.1 J has been used. For ceramic materials such as SiC, the ablated sample amounts are of the order of 1 ng, and the detection limits are down to the 0.1 µg/g level.

7.1.3. Applications

Applications for ICP – MS are broadly similar to those for ICP-AES, though the better sensitivity of the former has resulted in a greater number of applications in which ultralow levels of detection are required. Sample preparation methods are similar to those used for FAAS and ICP-AES, but nitric acid is favored for sample digestion since the other mineral acids contain elements that cause spectroscopic interferences.

For geological applications ICP – MS is extremely well suited to the determination of the lanthanide elements because of its relatively simple spectrum. When coupled with UV laser ablation, it is an extremely powerful technique for spatial analysis and the analysis of geological inclusions [319], especially with the extremely low limits of detection afforded by a sector field instrument. If simultaneous array detection is used then extremely precise isotope ratios can be measured, suitable for geochronology [320]. In environmental analysis the sensitivity of ICP – MS makes it ideal for applications such as the determination of heavy metals [243] and long-lived radionuclides [321] in the environment, and the determination of transition metals in seawater with on-line preconcentration [322]. Because of its capability for rapid multielement analysis ICP – MS is particularly suited to sample introduction methods which give rise to transient signals. For example, electrothermal vaporization, flow injection, and chromatographic methods for the speciation of organometallic and organohalide compounds [299]. A major attraction is the ability to perform isotope dilution analysis [320] which is capable of a high degree of accuracy and precision, and approaches the status of a primary method of analysis.

ICP – MS is the method of choice for routine analysis of samples containing ppb levels of analyte, and for speciation studies. Most quadrupole instruments are now benchtop-sized, and sector-field instruments are becoming easier to use and maintain, though high-resolution operation still requires considerable skill and experience.

7.2. Glow Discharge Mass Spectrometry

Glow discharges [152] are known from their use as radiation sources for AES, and have been recognized as powerful ion sources for mass spectrometry. This development started from spark source mass spectrometry, where there was a continuous search for more stable sources to reduce the matrix dependence of the analyte signals [51].

In glow discharge mass spectrometry (GD – MS), the analyte is volatilized by sputtering and the ions are produced in the negative glow of the discharge, mainly as a result of collisions of the first kind with electrons and Penning ionization. Subsequently, an ion beam of composition representative of the sample is extracted. Between the glow discharge and the spectrometer a reduction in gas pressure is required. As the glow discharge is operated at a pressure around 100 Pa, a two-stage vacuum system is required. Cathodic extraction can be done at the cathode plume, by making the sample the cathode and drilling a hole in it. The aperture should reach the hot plasma center, and in most cases the extraction is done anodically.

Owing to impurities in the filler gas and the complexity of the processes taking place, not only analyte and filler gas ions, but many others may occur in the mass spectrum. Spectral interference may occur from several causes:

1) Isobaric interference from isotopes of different elements, e.g., $^{40}Ar^{+}$ with $^{40}Ca^{+}$, or $^{92}Zr^{+}$ with $^{92}Mo^{+}$
2) Interference of analyte ion signals with doubly charged ions, e.g., $^{56}Fe^{2+}$ with $^{28}Si^{+}$ This type of interferences is much rarer than in earlier work with high-vacuum sparks
3) Interference of analyte ions with cluster ions formed from analyte and gas species, e.g., $^{40}Ar^{16}O^{+}$ with $^{56}Fe^{+}$
4) Interference by signals from residual gas impurities, e.g., $^{14}N^{16}O^{1}H^{+}$ with $^{31}P^{+}$

Separation of the interfering signals depends on instrument resolution. In the case of quadrupoles, the mass resolution is at best unity, and interference can be corrected only by mathematical procedures. However, with high-resolution instruments with resolution of 5000 interference remains, especially for hydrides in the mass range 80 – 120.

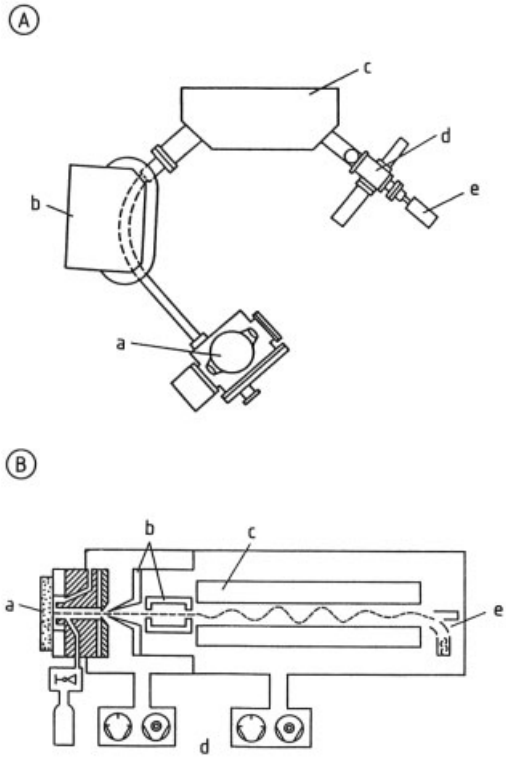

Figure 46. Glow discharge mass spectrometry
A) Sector field instrument VG 9000 [324]
a) Source with sample introduction; b) Magnetic sector field; c) Electrostatic energy analyzer; d) Daly detector; e) Faraday cup
B) Quadrupole-based, low-resolution glow discharge mass spectrometer [325]
a) Source; b) Ion optics; c) Quadrupole mass filter; d) Differential pumping system; e) Detector

7.2.1. Instrumentation

Glow discharges operated at pressures in the 10 – 100 Pa range can be coupled to mass spectrometers by using an aperture between both chambers with a size of ca. 1 mm. The glow discharge may use pin-shaped samples which can be introduced into the source without admitting air. BENTZ and HARRISON [323] coupled discharges with a pin cathode and hallow cathode plumes to a high-resolution sector feld mass spectrometer. Various electrode geometries were studied and diagnostic studies performed. This approach dates back to the VG 9000 spectrometer (Fig. 46 A), a double-focusing instrument in so-called reversed Nier – Johnson geometry, with a spectral resolution of ca. 10^4. The mass range to 280 can be covered with an acceleration voltage of 8 kV. As detectors, a Faraday cup and a so-called Daly multiplier are used, with the possibility of continuous switching. The vacuum of 10^{-4} Pa is maintained with the aid of oil diffusion or turbomolecular pumps. To reduce background, the source housing can be cooled to act as a cryopump.

Quadrupole mass spectrometers have also been used to detect the ions produced in a glow discharge. A high voltage is not required to achieve adequate ion extraction, and rapid scans can easily be performed, which enables quasi-simultaneous, multielement measurements. An electrostatic analyzer is often used in front of the spectrometer as an energy filter. JAKUBOWSKI et al. [325] described a combination of a Grimm-type glow discharge with a quadrupole mass spectrometer (Fig. 46 B) and studied the basic influence of the working and construction parameters. Glow discharge mass spectrometers using quadrupoles are commercially available.

7.2.2. Analytical Performance

The discharge is usually operated at about 100 Pa, a gas flow rate of a few mL/min, current ca. 2 – 5 mA, with a burning voltage of 500 – 1500 V. Normally the power is stabilized and pressure, current, voltage, or power are taken as optimization parameters. All electrically conducting samples, and even semiconductors, can be directly analyzed. After a burn-in phase, the discharge can be stable for hours. Detection limits down to the ng/g range are obtained for most elements.

Mass spectrometric methods are relative methods, and need to be calibrated with known reference samples; so-called relative sensitivity factors (RSF) are used:

$$\mathrm{RSF} = (x_{\mathrm{el}}/x_{\mathrm{ref}})\,(c_{\mathrm{ref}}/c_{\mathrm{el}}) \tag{175}$$

where x_{el} and x_{ref} are the isotope signals for the element to be determined and a reference element, and c are their concentrations. When matrix effects are absent RSF = 1. In Gd – MS, RSFs are much closer to unity as compared

for instance with spark source mass spectrometry [326] and secondary ion mass spectrometry. Especially in quadrupole systems, spectral interference may limit the power of detection as well as the analytical accuracy. This is shown by a comparison of high- and low-resolution scans of a mass spectrum (Fig. 47) [326], [327]. In Figure 47 A the signals of $^{56}Fe^{2+}$, $^{28}Si^{+}$, $^{12}C^{16}O^{+}$, and $^{2}N_2^{+}$ are clearly separated, but not in the quadrupole mass spectra. In the latter (Fig. 47 B), interference can be recognized from comparisons of the isotopic abundances. Physical means, such as the use of neon as discharge gas [328], can be used to overcome the spectral interference problem.

7.2.3. Applications

Glow discharge mass spectrometry is of use for the direct determination of major, minor, and trace bulk concentrations in electrically conducting and semiconductor solids down to concentrations of 1 ng/g [329]. Because of limited ablation (10 – 1000 nm/s), the analysis times are long, especially when samples are inhomogeneous. The technique has been applied especially for the characterization of materials such as Al, Cd, Ga, In, Si, Te, GaAs, and CdTe. Both for high-purity substances [330], and semiconductor-grade materials [331], up to 16 elements can be determined in 5 N and 6 N samples. At high resolution, determination of Ca and K is possible [332]. In high-purity Ga, the detection limits are below the blank values of chemical analysis [333]; SiC analysis results agree with those of neutron activation analysis [331]. In Al samples, Th and U concentrations of 20 ng/g can be determined [330]. The capabilities of low-resolution GD – MS have been shown in steel analysis [333], where detection limits are down to 1 ng/g and as many as 30 elements can be determined. Electrically insulating powder samples can be analyzed after mixing with a metal powder such as Cu, Ag, or Au and pelleting, as known from glow discharge – optical emission spectrometry (GD – OES). However, degassing and oxygen release may necessitate long burn-in times and induce instabilities. Blanks arising from the metal powder may be considerable, e.g., for Pb, in the case of copper powder.

Depth profiles can be recorded just as in GD – OES, but with such high resolution that much thinner multilayers can be characterized [334]. Conversion of the intensity scale to a concentration scale is easier and the sensitivity higher. By the use of different working gases [335], the scope of the method can be continuously adapted.

Dry solution residue analysis can be performed with GD – MS. This has been shown with the aid of hollow cathodes, in which Sr, Ba, and Pb can be determined [336] and where in samples of 10 – 20 µL, up to 70 elements can be determined with detection limits in the picogram range [337]. With a Grimm-type discharge, solution residues can be formed on the cathode by evaporating microvolumes of analyte solutions with low total salt content. With noble metals [338], cementation can be used to fix the analyte on a copper target, so that it is preconcentrated and can be sputtered reproducibly. This has been found useful in fixing picogram amounts of Ir on a copper plate prior to analysis with GD – MS.

Manufacturers. Inductively coupled plasma (ICP) and glow discharge (GD) mass spectrometers with high (h) and low (l) resolution are available from several manufacturers:

Agilent Technologies (formerly Hewlett Packard), USA: ICP-MS, MIP-AES
Extrel, USA: GD-MS
GBC, Australia: TOF-ICP-MS, AAS
Hitachi, Japan: ICP-AES, AAS, ICP-MS
LECO Corporation, USA: TOF-ICP-MS
Micromass, UK: S-ICP-MS, ICP-MS
Nu Instruments Ltd., UK: SF-ICP-MS
PerkinElmer Instruments (incorporating Princeton Applied Research, ORTEC, Signal Recovery and Berthold products), USA: AAS, ICP-OES, ICP-MS, GC, GC/MS
Spectro Analytical Instruments, Germany: ICP-AES, arc/spark AES, ICP-MS
TJA Solutions (incorporating Thermo Jarrell Ash, VG Elemental and Unicam AA), USA: AAS, ICP-AES, ICP-MS, SF-ICP-MS, GD-AES, GD-MS
Thermoquest (incorporating Finnigan), USA: SF-ICP-MS, AAS
Varian Associates, Australia: ICP-AES, ICP-MS, AAS

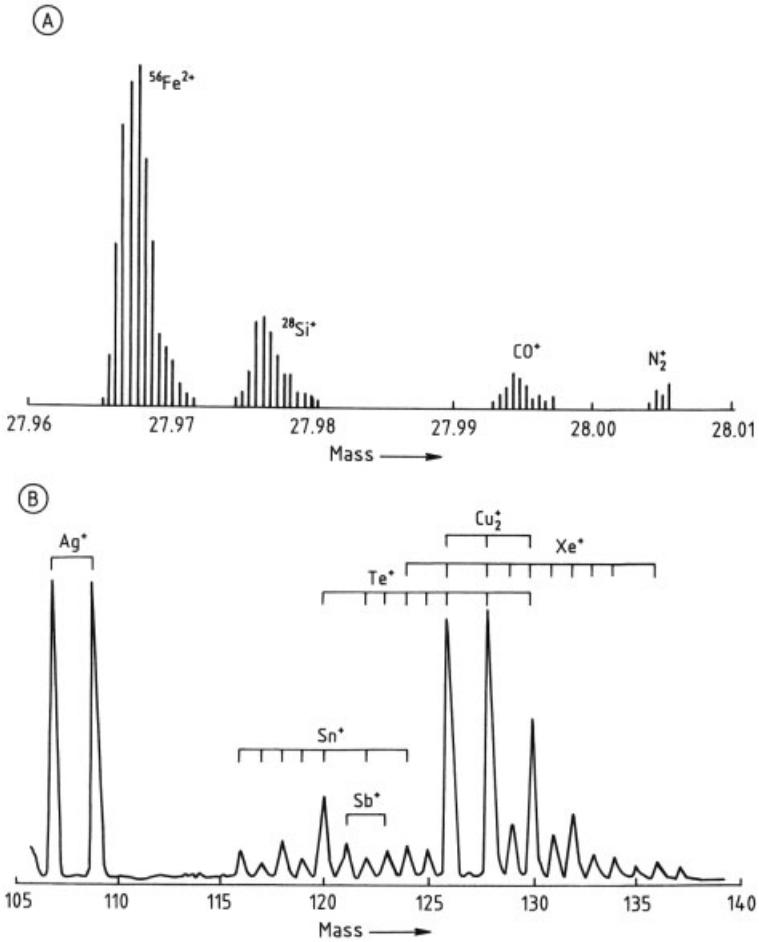

Figure 47. Glow discharge mass spectrum
A) High resolution [326]; B) Low resolution [327]

8. Atomic Fluorescence Spectrometry

In atomic fluorescence spectrometry (AFS), the analyte is introduced into an atomizer (flame, plasma, glow discharge, furnace) and excited by monochromatic radiation emitted by a primary source. The latter can be a continuous source (xenon lamp) or a line source (hollow cathode lamp, electrodeless discharge lamp, or tuned laser). Subsequently, the fluorescence radiation, which may be of the same wavelength (resonance fluorescence) or of longer wavelength (nonresonance fluorescence), is measured.

8.1. Principles [339]

For a two-level system (Fig. 48), with excitation only as a result of absorption of radiation with radiant density ϱ_ν, the population of the excited level n_2 is:

$$\mathrm{d}n_2/\mathrm{d}t = n_1 B_{12} \varrho_\nu - n_2 \left[A_{21} + k_{21} + B_{21} \varrho_\nu\right] \qquad (176)$$

where B_{12} and B_{21} are the transition probabilities for absorption and stimulated emission, ν the frequency, and n_1 the population of the lower level; A_{21} is the transition probability for spontaneous emission, and k_{21} is the coefficient for decay by collisions. The exciting radiation is supposed to be constant over the absorption profile. When the statistical weights of n_1 and n_2 are

equal, $B_{21} = B_{12}$, and when $n_T = n_2 + n_1$, Equation (176) can be solved:

$$n_2 = n_T B \varrho_\nu t_r \left[1 - \exp\left(-t/t_r\right)\right] \tag{177}$$

where:

$$t_r = \left[2B\varrho_\nu + A_{21} + k_{21}\right]^{-1} \tag{178}$$

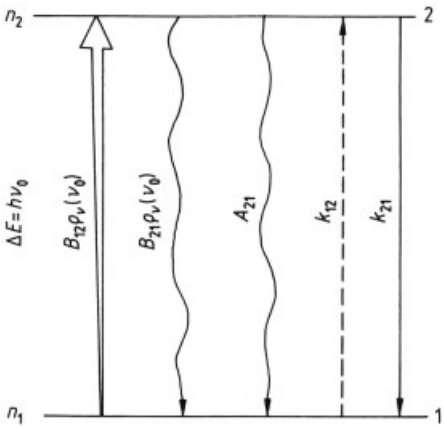

Figure 48. Transitions in a two-level system for atomic fluorescence spectrometry [339]

Accordingly, when the absorption of radiation increases to a certain value, A_{21} and k_{21} become negligible. The $n_2 = n_T/2$, and is independent of the radiant density. This state of saturation can be obtained when lasers are used as primary sources. As the fluorescence intensities are low, especially in the case of nonresonant AFS, special precautions must be taken for signal acquisition. Signal-to-noise ratios can be considerable improved by using pulsed signals and phase-sensitive amplification. The noise is partly due to flickering (proportional to the product of measurement time t and the mean radiance B^{s}_{ave} of the source and to background contributions (proportional to background mean intensity B^{b}_{ave} and time t) and:

$$S/N = K B^{s}_{ave} / \left(B^{s}_{ave} + B^{b}_{ave}\right)^{1/2} \tag{179}$$

By pulsing the source (with pulse duration t_p, frequency f, and pulse intensity B^{s}_{p}) as in the case of a laser, the signal-to-noise ratio is:

$$S/N = K B^{s}_{p} f t_p / \left(B^{s}_{p} f t_p + B^{b}_{ave}\right)^{1/2} \tag{180}$$

but by phase-sensitive amplification, in the case of background noise limitations, it can be improved to:

$$S/N' = K B^{s}_{p} f t_p / \left[f t_p \left(B^{s}_{p} + B^{b}_{ave}\right)\right]^{1/2} = S/N \left(f t_p\right)^{1/2} \tag{181}$$

This improvement can be considerable when "boxcar" integrators with small time windows are used, to give $f t_p \approx 10^{-8}$.

8.2. Instrumentation

An AFS system comprises a primary source, an atomizer, and a detection system for measuring the fluorescence radiation, which in some cases can be done without spectral dispersion.

Primary Source. As primary sources, continuous sources, such as tungsten halogenide or deuterium lamps, can be used. They have the advantage that multielement determinations are possible, but owing to the low radiant densities saturation is not obtained and the power of detection not fully exploited. With line sources, such as hollow cathode sources and electrodeless discharge lamps, much higher radiances can be obtained. Even ICPs into which a concentrated solution is introduced can be used as primary sources, so that multielement determinations become possible. With laser sources, saturation can be obtained. In most cases, dye lasers are used. They have a dye (mostly a substituted coumarin) as active medium. This liquid is continuously pumped through a cuvette and has a broad wavelength band (ca. 100 nm), within which a wavelength can be selected by a grating or prism. By using several dyes, the whole wavelength range down to 440 nm can be covered. The laser is often pumped by a second laser (e.g., an excimer laser) and operated in a pulsed mode. As the energy output is very high, it is possible to apply frequency doubling and Raman shifting, to access the wavelength range down to 200 nm. Accordingly, saturation can be flexibly achieved for almost any term. In the case of resonant transitions (with an oscillator strength: 0.1 – 1) the energy required is 10 – 100 mW, whereas for nonresonant transitions (oscillator strength 10^{-3}) it can reach 1 W.

Tunable dye lasers can deliver frequencies up to ca. 10 Hz and the pulse length can be down to 2 – 10 ns.

Atomizers. As fluorescence volumes, chemical flames are very useful, because of their good sample uptake and temporal stability. There has been interest in the low radiative background, low quenching, argon – hydrogen diffusion flame. The temperature of this flame is too low to prevent severe chemical interferences and therefore the argon separated air – acetylene flame has been most widely used. The hot nitrous oxide – acetylene flame (argon separated) has been used where atomization requirements make it essential. In all cases, circular flames, sometimes with mirrors around them, offer the preferred geometry. These problems do not occur when an ICP is used as atomizer [340]. In the ICP, measurements should be made in the higher part of the tail flame, otherwise the ionized and excited analyte fraction is too high. To minimize the formation of oxides and sideways diffusion of analyte, a burner with an extended outer tube should be used for AFS. In a graphite furnace, the atomization efficiency is high and, owing to the limited expansion of the vapor cloud, the residence time of the analyte atoms is long; accordingly, graphite furnace atomization laser AFS has extremely low absolute detection limits. Low-pressure discharges can also be used as the fluorescence volume. They have the advantage that quenching of the fluorescent levels, which may limit the power of detection and cause matrix effects, is lower [341]. Laser-produced plasmas may be used as the fluorescence volume, [145], and are especially useful for direct trace determinations in electrically insulating solids.

Detection of fluorescence differs in resonant and nonresonant AFS. In *resonant AFS*, the radiation is measured perpendicular to the direction of the exciting radiation. The system suffers from stray radiation and emission of the flame. The latter can be eliminated by using pulsed primary sources and phase-sensitive detection. In *nonresonant fluorescence*, stray radiation problems are not encountered, but fluorescence intensities are lower, which necessitates the use of lasers as primary sources, and spectral apparatus to isolate the fluorescence radiation. A set-up for laser-excited AFS (Fig. 49) may make use of a pulsed dye laser pumped by an excimer laser. Here, pulses of 2 – 10 ns occur at a frequency up to 25 Hz. A mirror can be positioned behind the source to pass the exciting radiation twice through the fluorescence volume and to lead the largest possible part of the fluorescence radiation into the monochromator. The photomultiplier signal is then amplified and measured with a box-car integrator with a nanosecond window synchronized with the pumping laser.

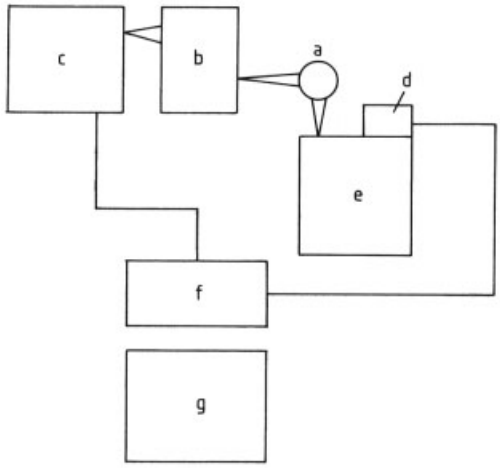

Figure 49. Experimental set-up for laser-induced atomic flourescence
a) Flame, ICP, etc.; b) Dye laser; c) Pumping laser; d) Photomultiplier; e) Monochromator; f) Boxcar integrator; g) Data treatment and display

8.3. Analytical Performance

Detection limits in AFS are low, especially for elements with high excitation energies, such as Cd, Zn, As, Pb, Se, and Tl.

Elements such as As, Se, and Te can be determined by AFS with hydride sample introduction into a flame or heated cell followed by atomization of the hydride. Mercury has been determined by cold-vapor AFS. A nondispersive system for the determination of Hg in liquid and gas samples using AFS has been developed commercially. Mercury ions in an aqueous solution are reduced to mercury using tin(II) chloride solution. The mercury vapor is continuously swept out of the solution by a carrier gas and fed to the fluorescence detector, where the fluorescence ra-

diation is measured at 253.7 nm after excitation of the mercury vapor with a high-intensity mercury lamp (detection limit 0.9 ng/L). Gaseous mercury in gas samples (e.g., air) can be measured directly or after preconcentration on an absorber, for example, gold-coated sand. By heating the absorber, mercury is desorbed and transferred to the fluorescence detector. This preconcentration method can also be applied to improve the detection limits when aqueous solutions (or dissolved solid samples) are analyzed.

In *flame AFS*, elements which form thermally stable oxides, such as Al, Mg, Nb, Ta, Zr, and the rare earths, are hampered by insufficient atomization. This is not the case when an ICP is used as the fluorescence volume. Detection limits in the case of laser excitation and nonresonant fluorescence are lower than in ICP – AES (Table 8) [340]. For both atomic and ionic states, ICP – AFS can be performed [342].

Table 8. Detection limits in laser atomic fluorescence and laser-enhanced ionization spectrometry

	Detection limit, ng/mL		
	Laser AFS ICP [340]	Laser AFS furnace [341]	LII flame [348]
Ag		0.003	0.07
Al	0.4		0.2
B	4		
Ba	0.7		0.2
Ca			0.03
Co		0.002	0.08
Cu		0.002	0.7
Eu		10	
Fe		0.003	0.08
Ga	1		0.04
In			
Ir		0.2	
K			0.1
Li			0.0003
Mn		0.006	0.02
Mo	5		10
Na		0.02	0.003
Pb	1	0.000025	0.09
Pt		4	
Rb			0.1
Si	1		40
Sn	3		0.3
Ti	1		1
Tl	7		0.008
V	3		0.9
Y	0.6		10
Zr	3		

In the case of a *graphite furnace* as fluorescence volume, and laser excitation, detection limits are very low; e.g., Ag 3×10^{-3} ng/mL; Co 2×10^{-3} ng/mL; Pb 2×10^{-3} ng/mL [339], [343].

Dry solution residue analysis with laser-excited AFS and a *hollow cathode discharge* atomizer offers very sensitive determinations of Mo [347]. Bolshov et al. [345] showed that very low levels of lead in Antarctic ice samples could be determined by laser-excited AFS of dry solution residues with graphite furnace atomization.

When performing laser-excited AFS at a *laser plume*, it seems useful to produce the laser plasma below atmospheric pressure, as then the ablation depends little on the matrix [146], [346].

In AFS, spectral interference is low as the fluorescence spectra are not very complex. The linear dynamic range of AFS is similar to that of OES, or even larger. There is no background limitation nor a serious limitation by self-reversal, both of which limit the linear dynamic range in OES, at the lower and at the higher end, respectively. It well may occur that the exciting radiation gets less intense when penetrating into the fluorescence volume, and that fluorescence radiation is absorbed by ground-state atoms before leaving the atomizer. These limitations, however, are less stringent than self-reversal in OES.

The analytical application of laser-excited AFS has so far, been seriously hampered by the complexity and cost of tunable lasers; this may change with the availability of less expensive but powerful and tunable diode lasers.

9. Laser-Enhanced Ionization Spectrometry (→ Laser Analytical Spectroscopy, Chap. 3.2.)

Laser-enhanced ionization spectrometry (LEI) is based on the optogalvanic effect, known from early work on gas discharges. Here, an increase of the current occurs in a low-pressure discharge when irradiating with monochromatic radiation at a wavelength corresponding to a transition in the energy level diagram of the discharge gas. As an analytical method, LEI is based on measurement of the increase of analyte ionization in an atomizer, resulting from selective excitation by laser radiation [347], [348].

Alternatively, two lasers can be used to both excite and photoionize the analyte. This is known as direct laser ionization (DLI) and

the two methods (LEI and DLI) are collectively termed laser-induced ionization (LII).

Ionization of atoms in an atomizer may result from collisions with high-energy atoms of another species (Fig. 50) and the rate constant for this reaction can be written:

$$k_{1\mathrm{i}} = n_{\mathrm{X}} \left[8kT/\left(\pi m_{\mathrm{r}}\right)\right]^{1/2} Q_{\mathrm{MX}} \exp\left(-E_{\mathrm{i}}/kT\right) \quad (182)$$

where X refers to analyte atoms, Q_{MX} is the cross section for collisions between X and M (another element), k is Boltzmann's constant, E_{i} is the ionization energy, and T is the temperature in the atomizer. When the atoms X are in an excited state, being 1 eV above the original level $k_{1\mathrm{i}}$ changes to:

$$k_{2\mathrm{i}} = \exp\left(1/kT\right) k_{1\mathrm{i}} \quad (183)$$

and the increase in ionization for a flame with $T = 2500\,\mathrm{K}$ is $\mathrm{e}^5 = 150$. In order to get sufficiently high population of the excited level, a laser must be used. For a closed system with two levels n_1 and n_2, the equilibrium is characterized by:

$$B_{12}\varrho_\nu n_1 = B_{21}\varrho_\nu n_2 + A_{21}n_2 + k_{21}n_2 \quad (184)$$

where:

$$B_{21} = \left(g_1/g_2\right) B_{12}$$

and:

$$n_{\mathrm{T}} = n_1 + n_2 + n_{\mathrm{i}}$$

B_{12} and B_{21} are the transition probabilities for absorption and stimulated emission, and n_1, n_2, n_{i}, and n_{T} the analyte atom densities for the lowest state, the excited state, the ionized state, and the total number densities; g_1 and g_2 are the statistical weights, A_{21} is the transition probability for spontaneous emission and k_{21} the rate coefficient for collisional decay. Accordingly:

$$n_2 = \frac{B_{12}\varrho_\nu\left(n_{\mathrm{T}} - n_{\mathrm{i}}\right)}{\left[\left(1 + g_1/g_2\right) B_{12}\varrho_\nu + A_{21} + k_{21}\right]} \quad (185)$$

and n_2 will increase until $A_{21} + k_{21}$ becomes negligible and $n_2/n_1 = g_2/g_1$. The required radiation densities are ca. 10 – 100 mW/cm^3 for a resonant transition (with oscillator strength 0.1 – 1) and 2 W/cm^3 for a nonresonant transition (with oscillator strength 0.001). The use of a continuous wave laser maximizes n_{i}, and the use of multistep excitation may be helpful, especially when high-energy terms (Rydberg states) have to be populated.

As an analytical technique, LEI requires a tunable laser, an atomizer, a galvanic detector and read-out electronics, including a boxcar integrator. It was first realized experimentally with a flame, with pneumatic nebulization for sample uptake and atomization [347].

A dye laser is used in most cases, as the selection of different dyes, frequency doubling, and Raman shifting allow coverage of the spectral range down to 200 nm. In the case of a flame (Fig. 50), the laser enters a few millimeters above the burner head, and a large collector electrode is situated a similar distance above the laser beam. This allows high-efficiency extraction of the ions formed. Apart from flames, furnaces and glow discharges are also very useful atomizers. The measurement unit includes phase-sensitive amplification to give high signal-to-noise ratios, but does not need any spectrometric equipment as only an ion current has to be measured.

The sensitivity in LEI depends on the properties of the laser system, the terms populated, the atomizer, and the detection system. It can be written:

$$S = cX_{\mathrm{A}}\beta\left(A\right)\left[\exp\left(-E_{\mathrm{i}}^{*}/kT\right) B_{12}\left(A\right)\varrho_\nu^{\mathrm{P}}\left(A\right)\right] \quad (186)$$

where c is a constant, X_{A} is the number density of analyte atoms A in the lower level according to Boltzmann, β is the atomization efficiency in the flame, $\varrho^{\mathrm{P}}{}_\nu\left(A\right)$ is the spectral radiation density per laser pulse, $E*_{\mathrm{i}}$ is the ionization energy of the excited level, and S is the sensitivity in nA ng^{-1} mL^{-1}. The noise includes contributions from flicker noise, nebulizer noise, and detector noise, and with phase-sensitive amplification, detection limits are in the sub-ng/mL range (Table 8) [347], [350]. Because of the maximum atom number densities and residence times, detection limits are lowest for furnace atomizers. In pulsed lasers, the radiation densities favor sensitivity, whereas in continuous wave lasers the n_{i} values increase as a result of improved excitation efficiencies. In the thermionic diode, a heated collector electrode is placed in the vicinity of the laser beam and isolated from the vessel walls. No extraction voltage is applied and the ion current flowing as a result of decreasing negative space charge around the collector,

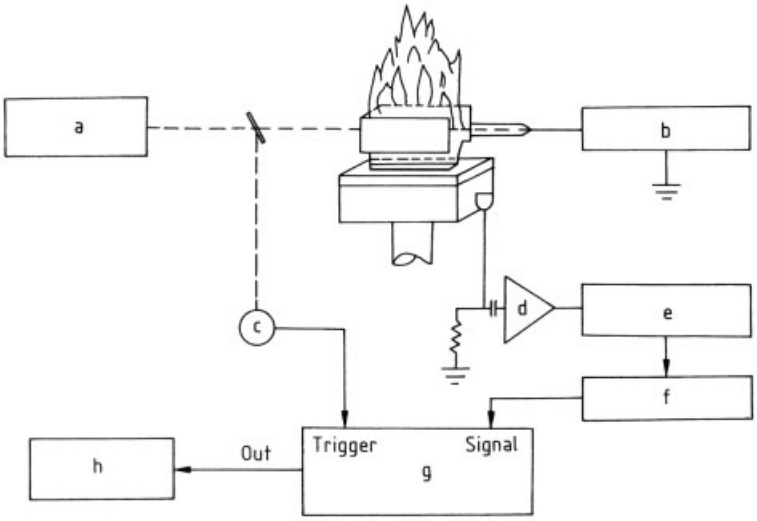

Figure 50. Flame laser-enhanced ionization spectrometry [349]
a) Flashlamp/dye laser; b) High voltage; c) Trigger photodiode; d) Preamplifier; e) Pulse amplifier; f) Active filter; g) Boxcar averager; h) Chart recorder

caused by laser ionization, is taken as analytical signal. The whole system is usually operated at reduced pressure and can detect as few as 10^2 atoms [351]. In LEI, matrix effects are low. This applies especially for samples with complicated atomic spectra, as spectral interference is absent. Elements with low ionization energies may cause matrix effects, possibly due to a change of the ion current as a result of changes in the electron number density in the flame. This can be compensated by measuring alternately with and without laser irradiation. Ionization of the matrix by the laser may also falsify the ion currents measured. This can be eliminated by modulation of the frequency of the exciting laser radiation around the frequency ν, which corresponds to $E_2 - E_1 = h\nu$ when E_2 is the intermediate level (Rydberg state) and E_1 the ground state. Ion production as a result of laser irradiation may influence the collection efficiency at the electrode as the electrical field around the electrode changes. This effect may be minimized by using a large electrode and by collecting all the ions produced. This sets limits to the pulsing frequency. For a distance of 5 mm between the collector and the burner, a field of 100 V/cm and a mobility of $20\,\mathrm{cm^2\,V^{-1}\,s^{-1}}$ (typical for a flame in which $T = 2500$ K), the highest allowed pulsing rate is 4000 Hz. It has been found that for Na concentrations of 100 µg/mL, matrix effects are still negligible.

The linear dynamic range in flame LEI is of the order of three decades. However, the multielement capacity is limited as monochromatic laser radiation is required.

When using two lasers and applying two-photon spectroscopy, only those atoms which do not have a velocity component in the observation direction will undergo LEI. The absorption signals become very narrow (Doppler-free spectroscopy), enhancing the selectivity and the power of detection, and making isotope detection feasible.

The LEI technique has been applied successfully for certification purposes [352], [353], and represents a powerful approach for combinations of laser evaporation with all other atomization techniques. Apart from the galvanic detection of the ion currents, direct mass spectrometric detection of the ions can be applied, as in resonance ionization mass spectrometry [354]. Here, ionization can also be performed by multiphoton absorption, which requires very intense primary sources.

Table 9. Figures of merit of analytical methods

Method	Power of detection	Matrix effects	Cost
Atomic absorption			
Flame	++	+	+
Furnace	+++	+++	++
Atomic emission			
d.c. Arc	++	+++	+++
Spark	+	++	++
d.c. Plasma	++	++	+
ICP	++	+	++
MIP	++	+++	+
Glow discharge	++	+	++
Laser	+	+	++
Atomic fluorescence			
(laser)	+++	++	+++
LEI	+++	+	+++
X-ray spectrometry			
XRF	++(+[a])	+	++
Electron microprobe	++	++	+++
PIXE	+++	++	+++
Auger electron			
spectroscopy	++	++	+++
Mass spectrometry			
Spark	+++	+++	+++
ICP	+++	++	+++
Glow discharge	+++	+	+++
SIMS	+++	++	+++
Laser	+++	++	+++
Activation analysis	+++	+	+++
Electrochemistry	+++	+++	+
Spectrophotometry	++	+++	+
Spectrofluorimetry	+++	+++	+
Chromatography			
Gas	+++	+++	++
Liquid	++	+++	++

+ Low; ++ medium; +++ very high
[a] TYXRF = Total reflection X-ray fluorescence.

10. Comparison With Other Methods

Atomic absorption, optical emission, atomic fluorescence, plasma atomic spectrometry, and newer approaches such as laser-enhanced ionization now represent powerful tools for elemental analysis, including speciation, and are to be found in many analytical laboratories. Their power of detection, reliability in terms of systematic errors, and cost are compared with those of other methods of analysis in Table 9.

10.1. Power of Detection

Atomic absorption spectrometry using furnaces remains a most sensitive method for easily volatilized elements with high excitation energies, and this also applies to AFS using laser excitation. The development of diode lasers should make it possible to eliminate the use of spectrometers in AAS. Dynamic range problems can be decreased by measuring in the wings of the absorption lines, and the background measurement can be performed by tuning the laser [157]. This requires lasers with analytical wavelengths down to the vacuum UV. Laser-enhanced ionization is a further promising approach with respect to power of detection, especially when applying Doppler-free spectroscopy. However, these methods have limited multielement capacity. From this point of view, plasma atomic spectrometric methods offer the better possibilities; ICP – OES is likely to remain a powerful tool for multielement determinations in liquid samples and its power of detection is between that of flame and furnace AAS. It is the appropriate method for tasks such as wastewater analysis. Inductively coupled plasma – mass spectrometry offers further progress, provided the matrix loading of the sample solutions remains low. It may well have limitations in the low mass range

(below 80), owing to interference by cluster ions. The use of sample introduction techniques for generating dry aerosols has great potential, e.g., for Fe, As, Se, etc. Further progress is possible by the use of high-resolution mass spectrometry with sector field instruments. It is likely to occur as now lower cost instruments make the technique available to a wider circle of users.

Whereas ICP atomic spectrometry is certainly a most useful method for multielement determinations in liquid samples at low concentrations, glow discharges are very useful for the analysis of solids. Glow discharge atomic emission has established itself, especially for depth profiling analysis, and is useful for bulk analysis, together with standard methods such as arc and spark emission spectrometry. Glow discharge mass spectrometry first found a role in the analysis of ultra-pure metals, as requird in microelectronics. However, its quantification is difficult and requires certified samples for calibration. In dry solution residue analysis, GD – MS has extremely low absolute detection limits. This approach is very useful for many applications in the life sciences, as it combines high power of detection with the possibilities of isotope dilution. In mass spectrometry, this approach offers new possibilities, as more classical thermionic techniques are only applicable to elements which form volatile compounds. Optical atomic and mass spectrometric methods can be used for the determination of light elements, which is an advantage over X-ray spectrometric methods. Apart from this restriction, total reflection X-ray fluorescence (TRXRF) offers high powers of detection, accuracy, and multielement capacity, with a minimum of sample preparation [355].

10.2. Analytical Accuracy

The analytical accuracy of methods can only be discussed in view of the complete analytical procedure applied. It is necessary to tune sample preparation and trace – matrix separations to the requirements of the analytical results in terms of accuracy, power of detection, precision, cost, number of elements, and, increasingly, the species to be determined. However, the intrinsic sensitivity of the different determination methods to matrix interference remains important. In optical emission and mass spectrometry, spectral interference remains an important limitation to the achievable analytical accuracy. In atomic emission, this applies especially to the heavier elements, as they have the more complex atomic spectra. Especially when they are present as the matrix (as is often the case in metals analysis) the necessitity of trace matrix separation is obvious when trace analyses are to be performed. In order to overcome limitations due to spectral interference, high-resolution echelle spectrometers are increasingly used. They are compact and thus combine high resolution with excellent stability, and further enable multielement determinations, by using advanced detector technology, e.g., CCDs.

In mass spectrometry, spectral interference especially limits the accuracy in the low mass range. Progress may be expected from the use of sector field and time-of-flight mass spectrometry. However, in the first case transmission and in the second case dynamic range problems must be given attention. Signal depression and enhancement are a further cause of interference in ICP – MS. They can be taken care of by standard addition, as the sprectral background is low. Internal standardization may compensate for the effects of easily ionized element. This is more difficult in ICP – AES, where the spectral background (especially in trace analysis) is considerable and may be strongly influenced by changes in the concentrations of easily ionized elements.

In AAS and AFS, limitations in the analytical accuracy are mostly related to physical and chemical interference and less to spectral interference. Especially in furnace AAS, thermochemical processes limit the achievable accuracy and necessitate carefully worked out temperature programs. In AFS and LEI, it is necessary to control matrix influences relating to quenching when analyzing real samples.

10.3. Cost

The power of detection and accuracy of analytical methods cannot be discussed without considering the costs of instruments and their operation, including the costs of laboratory personnel. Methods allowing multielement analyses and high throughput are advantageous for routine analyses. The ICP spectrometric methods offer particularly good possibilities, despite

the high instrument costs and the high consumption of power and gases. In many cases, however, the costs arising from sample preparation are decisive, and favor X-ray spectrometric methods, provided the limitations mentioned earlier are not encountered. Further progress is likely to depend on the availability of on-line sample treatment, e.g., with flow injection and on-line sample dissolution, which now becomes possible with microwave-assisted heating. In any case, the method selected has to be discussed in detail for any new analytical task.

11. References

1. G. R. Kirchhoff, R. Bunsen, *Philos. Mag.* **20** (1860) 89 – 98.
2. A. Walsh, *Spectrochim. Acta* **7** (1955) 108 – 117.
3. B. V. L'vov, *Spectrochim. Acta* **17** (1961) 761 – 770.
4. H. Maßmann, *Spectrochim. Acta Part B* **23 B** (1968) 215 – 226.
5. W. Grotrian: *Graphische Darstellung der Spektren von Atomen mit ein, zwei und drei Valenzelektronen,* Springer Verlag, Berlin 1928.
6. H. R. Griem: *Plasma Spectroscopy,* McGraw-Hill, New York 1964.
7. L. de Galan, R. Smith, J. D. Winefordner, *Spectrochim. Acta Part B* **23 B** (1968) 521 – 525.
8. W. J. Pearce, P. J. Dickerman (eds.): *Symposium on Optical Spectrometric Measurements of High Temperatures,* University of Chicago Press, Chicago 1961.
9. P. W. J. M. Boumans: *Theory of Spectrochemical Excitation,* Hilger & Watts, London 1966.
10. J. Junkes, E. W. Salpeter, *Ric. Spettrosc. Lab. Astrofis. Specola Vaticana* **2** (1961) 255.
11. C. F. Bruce, P. Hannaford, *Spectrochim. Acta Part B* **26 B** (1971) 207 – 235.
12. J. A. C. Broekaert, F. Leis, K. Laqua, *Spectrochim. Acta Part B* **34 B** (1979) 73 – 84.
13. W. Lochte-Holtgreven (ed.): *Plasma Diagnostics,* North-Holland Publ., Amsterdam 1968.
14. G. Elwert, *Z. Naturforsch. A: Astrophys. Phys. Phys. Chem.* **7 A** (1952) 703 – 711.
15. J. M. Mermet in P. W. J. M. Boumans (ed.): *Inductively Coupled Plasma Emission Spectroscopy,* Part II, Wiley-Interscience, New York 1987, p. 353.
16. R. Diermeier, H. Krempl, *Z. Phys.* **200** (1967) 239 –248.
17. G. Herzberg: *Molecular Spectra and Molecular Structure,* 2nd ed., Van Nostrand Reinhold, New York 1950.
18. R. W. B. Pearce, A. G. Gaydon: *The Identification of Molecular Spectra,* Chapman and Hall, London 1953.
19. J. A. C. Broekaert, *Bull. Soc. Chim. Belg.* **86** (1977) 895 – 906.
20. B. Raeymaekers, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **43 B** (1988) 941 – 949.
21. I. Ishii, A. Montaser, *Spectrochim. Acta Part B* **46 B** (1991) 1197 – 1206.
22. N. P. Ferreira, H. G. C. Human, L. R. P. Butler, *Spectrochim. Acta Part B* **35 B** (1980) 287 – 295.
23. K. Bockasten, *J. Opt. Soc. Am.* **51** (1961) 943 – 947.
24. G. M. Hieftje, *Spectrochim. Acta Part B* **47 B** (1992) 3 – 25.
25. J. A. C. Broekaert, *Anal. Chim. Acta* **196** (1987) 1 – 21.
26. M. Borer, G. M. Hieftje, *Spectrochim. Acta Reviews* **14** (1991) 463 – 486.
27. K. Doerffel: *Statistik in der analytischen Chemie,* Verlag Chemie, Weinheim 1984.
28. L. Sachs: *Statistische Auswertungsmethoden,* Springer Verlag, Berlin 1969.
29. V. V. Nalimov: *The Application of Mathematical Statistics to Chemical Analysis,* Pergamon Press, Oxford 1963.
30. J. D. Ingle, Jr., S. R. Crouch: *Spectrochemical Analysis,* Prentice-Hall International, Englewood Cliffs 1988.
31. G. M. Hieftje, *Anal. Chem.* **44** (1972) 81 A – 88 A.
32. H. Kaiser, *Spectrochim. Acta* **3** (1947) 40 – 67.
33. H. Kaiser, H. Specker, *Fresenius Z. Anal. Chem.* **149** (1956) 46 – 66.
34. E. L. Grove: *Analytical Emission Spectrometry,* vol. 1, Part I and II, Decker Inc., New York 1971.
35. C. S. Williams, O. A. Becklund: *Optics,* Wiley-Interscience, New York 1972.
36. P. N. Keliher, C. C. Wohlers, *Anal. Chem.* **48** (1976) 333 A – 340 A.
37. H. Kaiser, *Spectrochim. Acta* **3** (1948) 159 – 190.
38. P. W. J. M. Boumans, *Colloq. Spectrosc. Int. Proc. 16th* **II** (1971) 247 – 253.
39. H. Siedentopf, *Z. Phys.* **35** (1934) 454.

40. V. K. Zworykin, E. G. Ramberg: *Photoelectricity and Its Application,* Wiley, New York 1949.
41. Y. Talmi: "Multichannel Image Detectors", *ACS Symp. Ser.* **102** (1979).
42. J. V. Sweedler, R. F. Jalkian, M. B. Denton, *Appl. Spectrosc.* **43** (1989) 954 – 962.
43. J. A. C. Broekaert, K. R. Brushwyler, G. M. Hieftje, unpublished work.
44. J. A. C. Broekaert, C. A. Monnig, K. R. Brushwyler, G. M. Hieftje, *Spectrochim. Acta Part B* **45 B** (1990) 769 – 778.
45. A. Danielson, P. Lindblom, E. Södermann, *Chem. Scr.* **6** (1974) 5 – 9.
46. R. B. Bilhorn, P. M. Epperson, J. V. Sweedler, M. B. Denton, *Appl. Spectrosc.* **41** (1987) 1114 – 1124.
47. N. Furuta, K. R. Brushwyler, G. M. Hieftje, *Spectrochim. Acta Part B* **44 B** (1989) 349 – 358.
48. P. J. Treado, M. D. Morris, *Anal. Chem.* **61** (1989) 723 A – 734 A.
49. L. M. Faires, *Anal. Chem.* **58** (1986) 1023 A – 1043 A.
50. F. Adams, R. Gijbels, R. Van Grieken (eds.): *Inorganic Mass Spectrometry,* Wiley, New York 1988.
51. J. R. Bacon, A. Ure, *Analyst (London)* **109** (1984) 1229 – 1254.
52. K. G. Heumann, F. Beer, H. Weiss, *Mikrochim. Acta I* 1983, 95 – 108.
53. C. Brunnee, *Int. J. Mass Spectrom. Ion Processes* **76** (1987) 125 – 237.
54. H. H. Willard, L. L. Merrit, A. J. Dean, A. F. Settle: *Instrumental Methods of Analysis,* 7th ed., Wadsworth, California 1998, Chap. 16.
55. D. P. Myers, G. L. P. Yang, G. M. Hieftje, *J. Am. Soc. Mass Spectrom.* **5** (1994) 1008.
56. D. P. Myers, G. Li, P. P. Mahoney, G. M. Hieftje, *J. Am. Soc. Mass Spectrom.* **6** (1995) 411.
57. D. P. Myers, G. Li, P. P. Mahoney, G. M. Hieftje, *J. Am. Soc. Mass Spectrom.* **6** (1995) 400.
58. J. R. Chapman: *Practical Organic Mass Spectrometry,* 2nd ed., John Wiley & Sons, Chichester 1993.
59. A. Benninghoven, F. G. Rüdenauer, H. W. Werner: *Secondary Ion Mass Spectrometry,* Wiley, New York 1986.
60. E. Denoyer, R. Van Grieken, F. Adams, D. F. S. Natusch, *Anal. Chem.* **54** (1982) 26 A – 41 A.
61. D. P. Meyers, G. M. Hieftje: "Use of a Time-of-Flight Mass Spectrometer for Plasma Mass Spectrometry", *1992 Pittsburgh Conference,* New Orleans, March 9 – 13, 1992, Paper 606.
62. A. L. Gray, A. R. Date, *Analyst (London)* **108** (1983) 1033 – 1050.
63. F. E. Lichte, A. L. Meier, J. G. Crock, *Anal. Chem.* **59** (1987) 1150 – 1157.
64. F. Kohlrausch: *Praktische Physik,* 21st ed., vol. 1, Taubner Verlagsgesellschaft, Stuttgart 1961.
65. F. Rosendahl, *Spectrochim. Acta* **10** (1957) 201 – 212.
66. A. Aziz, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **37 B** (1982) 369 – 379.
67. J. Sneddon (ed.): *Sample Introduction in Atomic Spectroscopy,* Elsevier, Amsterdam 1990.
68. J. A. C. Broekaert: "Sample Introduction Techniques in ICP-AES", in P. W. J. M. Boumans (ed.): *Inductively Coupled Plasma Emission Spectroscopy,* vol. I, Wiley, New York 1987, pp. 296 – 357.
69. A. Montaser, D. W. Golightly (eds.): *Inductively Coupled Plasmas in Analytical Atomic Spectrometry,* 2nd ed., VCH Publishers, Weinheim 1992.
70. R. F. Browner, A. W. Boorn, *Anal. Chem.* **56** (1984) 786 A – 798 A.
71. R. F. Browner, A. W. Boorn, *Anal. Chem.* **56** (1984) 875 A – 888 A.
72. J. A. C. Broekaert, G. Tölg, *Fresenius' J. Anal. Chem.* **326** (1987) 495 – 509.
73. J. A. C. Broekaert, *Anal. Chim. Acta* **196** (1987) 1 – 21.
74. J. E. Meinhard in R. M. Barnes (ed.): *Applications of Plasma Emission Spectroscopy,* Heyden, London 1979, p. 1.
75. G. L. Gouy, *Ann. Chim. Phys.* **18** (1887) 5.
76. E. L. Kimberley, G. W. Rice, V. A. Fassel, *Anal. Chem.* **56** (1984) 289 – 292.
77. L. R. Layman, F. E. Lichte, *Anal. Chem.* **54** (1982) 638 – 642.
78. E. Sebastiani, K. Ohls, G. Riemer, *Fresenius' Z. Anal. Chem.* **264** (1973) 105 – 109.
79. H. Berndt, W. Slavin, *At. Absorpt. Newsl.* **17** (1978) 109.
80. S. Greenfield, P. B. Smith, *Anal. Chim. Acta* **59** (1972) 341.
81. A. Aziz, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **37 B** (1982) 381 – 389.
82. J. Ruzicka, E. H. Hansen: *Flow Injection Analysis,* Wiley, New York 1988.
83. S. D. Hartenstein, J. Ruzicka, G. D. Christian, *Anal. Chem.* **57** (1985) 21 – 25.

84. Z. Fang: *Flow Injection Separation and Preconcentration,* VCH Verlagsgesellschaft, Weinheim 1993.
85. P. Schramel, L. Xu, G. Knapp, M. Michaelis, *Fresenius J. Anal. Chem.* **345** (1993) 600 – 606.
86. S. Nukuyama, Y. Tanasawa, *Nippon Kikai Cakkai Ronbunshu* **4** (1938) 86, **5** (1939) 68.
87. R. Nießner, *Angew. Chem.* **103** (1991) 542 – 552.
88. T. Allen: *Particle Size Measurement,* Chapman & Hall, London 1981.
89. W. Van Borm et al., *Spectrochim. Acta Part B* **46 B** (1991) 1033 – 1049.
90. S. D. Olsen, A. Strasheim, *Spectrochim. Acta Part B* **38 B** (1983) 973 – 975.
91. A. W. Boorn, R. Browner, *Anal. Chem.* **54** (1982) 1402 – 1410.
92. G. Kreuning, F. J. M. J. Maessen, *Spectrochim. Acta Part B* **42 B** (1987) 677 – 688.
93. M. P. Doherty, G. M. Hieftje, *Appl. Spectrosc.* **38** (1984) 405 – 411.
94. H. Berndt, *Fresenius' J. Anal. Chem.* **331** (1988) 321 – 323.
95. J. Posta, H. Berndt, *Spectrochim. Acta Part B* **47 B** (1992) 993 – 999.
96. J. Posta, H. Berndt, S. K. Luro, G. Schaldach, *Anal. Chem.* **65** (1993) 2590 – 2595.
97. A. Syti, *CRC Crit. Rev. Anal. Chem.* **4** (1974) 155.
98. V. A. Fassel, B. R. Bear, *Spectrochim. Acta Part B* **41 B** (1986) 1089 – 1113.
99. P. W. J. M. Boumans, F. J. De Boer, *Spectrochim. Acta Part B* **30 B** (1975) 309 – 334.
100. N. W. Barnett, L. S. Chen, G. F. Kirkbright, *Spectrochim. Acta Part B* **39 B** (1984) 1141 – 1147.
101. D. C. Manning, *At. Absorpt. Newsl.* **10** (1971) 123.
102. M. Thompson, B. Pahlavanpour, S. J. Walton, G. F. Kirkbright, *Analyst (London)* **103** (1978) 568 – 579.
103. J. A. C. Broekaert, F. Leis, *Fresenius' Z. Anal. Chem.* **300** (1980) 22 – 27.
104. E. Bulska, P. Tschöpel, J. A. C. Broekaert, G. Tölg, *Anal. Chim. Acta* **271** (1993) 171 – 181.
105. H. Tao, A. Miyazaki, *Anal. Sci.* **7** (1991) 55 – 59.
106. C. S. Stringer, M. Attrep, Jr., *Anal. Chem.* **51** (1979) 731 – 734.
107. B. Welz, M. Melcher, *Analyst (London)* **108** (1983) 213 – 224.
108. E. Preuß, *Z. Angew. Mineral.* **3** (1940) 1.
109. B. V. L'vov, L. A. Pelieva, A. I. Sharnopolsky, *Zh. Prikl. Spektrosk.* **27** (1977) 395.
110. J. McNally, J. A. Holcombe, *Anal. Chem.* **59** (1987) 1105 – 1112.
111. T. S. West, X. K. Williams, *Anal. Chim. Acta* **45** (1969) 27.
112. V. Sychra et al., *Anal. Chim. Acta* **105** (1979) 105.
113. H. T. Delves, *Analyst (London)* **95** (1970) 431 – 438.
114. H. Berndt, J. Messerschmidt, *Spectrochim. Acta Part B* **34 B** (1979) 241 – 256.
115. E. I. Brooks, K. J. Timmins, *Analyst (London)* **110** (1985) 557 – 558.
116. Z. Grobenski, R. Lehmann, R. Tamm, B. Welz, *Mikrochim. Acta I* 1982, 115 – 125.
117. D. L. Millard, H. C. Shan, G. F. Kirkbright, *Analyst (London)* **105** (1980) 502 – 508.
118. J. A. C. Broekaert, F. Leis, *Mikrochim. Acta II* 1985, 261 – 272.
119. D. Sommer, K. Ohls, *Fresenius' Z. Anal. Chem.* **304** (1980) 97 – 103.
120. E. D. Salin, G. Horlick, *Anal. Chem.* **51** (1979) 2284 –2286.
121. G. Zaray, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **43 B** (1988) 241 – 253.
122. V. Karnassios, G. Horlick, *Spectrochim. Acta Part B* **44 B** (1989) 1345 – 1360.
123. V. Karnassios, M. Abdullah, G. Horlick, *Spectrochim. Acta Part B* **45 B** (1990) 119 – 130.
124. L. Blain, E. D. Salin, *Spectrochim. Acta Part B* **47 B** (1992) 205 – 217.
125. R. Lobinski et al., *Fresenius' J. Anal. Chem.* **342** (1992) 563 – 568.
126. B. Raeymaekers et al., *Spectrochim. Acta Part B* **43 B** (1988) 923 – 940.
127. B. Raeymaekers, P. Van Espen, F. Adams, *Mikrochim. Acta II* 1984, 437 – 454.
128. B. Docekal, V. Krivan, *J. Anal. At. Spectrom.* **7** (1992) 521 – 528.
129. H. Min, S. Xi-En, *Spectrochim. Acta Part B* **44 B** (1989) 957 – 964.
130. R. M. Barnes, S. Nikdel, *J. Appl. Phys.* **47** (1976) 3929.
131. G. M. Hieftje, R. M. Miller, Y. Pak, E. P. Wittig, *Anal. Chem.* **59** (1987) 2861 – 2872.
132. L. Ebdon, M. Foulkes, K. Sutton, *J. Anal. At. Spectrom.* **12** (1997) 213 – 229.
133. J. L. Johnes, R. L. Dahlquist, R. E. Hoyt, *Appl. Spectrosc.* **25** (1971) 629 – 635.
134. S. ElAlfy, K. Laqau, H. Maßmann, *Fresenius' Z. Anal. Chem.* **263** (1973) 1 – 14.

135. A. Aziz, J. A. C. Broekaert, K. Laqua, F. Leis, *Spectrochim. Acta Part B* **39 B** (1984) 1091 – 1103.
136. J. A. C. Broekaert, F. Leis, K. Laqua: "A Contribution to the Direct Analysis of Solid Samples by Spark Erosion Combined to ICP-OES", in B. Sansoni (ed.): *Instrumentelle Multielementanalyse,* Verlag Chemie, Weinheim 1985.
137. B. Raeymaekers, P. Van Espen, F. Adams, J. A. C. Broekaert, *Appl. Spectrosc.* **42** (1988) 142 – 150.
138. R. M. Barnes, H. V. Malmstadt, *Anal. Chem.* **46** (1974) 66 – 72.
139. N. Bings, H. Alexi, J. A. C. Broekaert, Universität Dortmund, unpublished results.
140. N. Omenetto: *Analytical Laser Spectroscopy,* Wiley, New York 1979.
141. L. Moenke-Blankenburg: *Laser Micro Analysis,* Wiley, New York 1989.
142. J. Sneddon, T. L. Thiem, Y. Lee: *Lasers in Analytical Spectrometry,* VCH, New York 1997.
143. K. Laqua: "Analytical Spectroscopy Using Laser Atomizers", in S. Martellucci, A. N. Chester (eds.): *Analytical Laser Spectroscopy,* Plenum Publishing Corp., New York 1985, pp. 159 – 182.
144. F. Leis, W. Sdorra, J. B. Ko, K. Niemax, *Mikrochim. Acta II* 1989, 185 – 199.
145. W. Sdorra, A. Quentmeier, K. Niemax, *Mikrochim. Acta II* 1989, 201 – 218.
146. J. Uebbing et al., *Appl. Spectrosc.* **45** (1991) 1419 –1423.
147. M. Kaminsky: *Atomic and Ionic Impact Phenomena on Metal Surfaces,* Springer Verlag, Berlin 1965.
148. F. M. Penning: *Electrical Discharges in Gases,* Philips Technical Library, Eindhoven 1957, p. 41.
149. P. W. J. M. Boumans, *Anal. Chem.* **44** (1973) 1219 –1228.
150. J. A. C. Broekaert, *J. Anal. At. Spectrom.* **2** (1987) 537 – 542.
151. M. Dogan, K. Laqua, H. Maßmann, *Spectrochim. Acta Part B* **27 B** (1972) 631 – 649.
152. R. Marcus (ed.): *Glow Discharge Spectrometries,* Plenum Press, New York 1993.
153. M. Van Straaten, A. Vertes, R. Gijbels, *Spectrochim. Acta Part B* **46 B** (1991) 283 – 280.
154. B. Welz: *Atomic Absorption Spectrometry,* Verlag Chemie, Weinheim 1985.
155. W. J. Price: *Spectrochemical Analysis by Atomic Absorption,* Heyden, London 1979.
156. S. J. Hill, A. Fisher: in L. Ebdon, E. H. Evans (eds.): *An Introduction to Analytical Atomic Spectrometry,* Wiley, Chichester 1998.
157. A. Zybin, C. Schneurer-Patschan, M. A. Bolshov, K. Niemax, *Trends. Anal. Chem.* **7** (1998) 513.
158. N. Omenetto, *J. Anal. At. Spectrom.* **13** (1998) 385 – 399.
159. R. S. Koirthyohann, E. E. Pickett, *Anal. Chem.* **37** (1965) 601 – 603.
160. G. P. Moulton, T. C. O'Haver, J. M. Harnly, *J. Anal. At. Spectrom.* **5** (1990) 145 – 150.
161. M. D. Amos, J. B. Willis, *Spectrochim. Acta* **22** (1966) 1325 – 1343.
162. R. Hermann, C. T. J. Alkemade: *Chemical Analysis by Flame Photometry,* 2nd ed., Interscience Publishers, New York 1983.
163. C. T. J. Alkemade, T. Hollander, W. Snelleman, P. J. Th. Zeegers: *Metal Vapours in Flames,* Pergamon Press, Oxford 1982.
164. J. A. Dean, T. C. Rains (eds.): *Flame Emission and Atomic Absorption Spectrometry,* vol. 1, Marcel Dekker, New York 1969.
165. W. Slavin, *Anal. Chem.* **58** (1986) 589 A – 597 A.
166. H. Berndt, G. Schaldach, *J. Anal. At. Spectrom.* **3** (1988) 709 – 713.
167. G. Schlemmer, B. Welz, *Spectrochim. Acta Part B* **41 B** (1986) 1157 – 1166.
168. B. Welz et al., *Spectrochim. Acta Part B* **41 B** (1986) 1175 – 1201.
169. V. Krivan, *J. Anal. At. Spectrom.* **7** (1992) 155 –164.
170. W. R. A. De Jonghe, F. C. Adams, *Anal. Chim. Acta* **108** (1979) 21 – 30.
171. D. Chakraborti et al., *Anal. Chem.* **56** (1984) 2692 – 2697.
172. M. Yamamoto, M. Yasuda, Y. Yamamoto, *Anal. Chem.* **57** (1985) 1382 – 1385.
173. E. Jackwerth, P. G. Willmer, R. Höhn, H. Berndt, *At. Absorpt. Newsl.* **18** (1978) 66.
174. G. Kaiser, D. Götz, P. Schoch, G. Tölg, *Talanta* **22** (1975) 889 – 899.
175. J. Piwonka, G. Kaiser, G. Tölg, *Fresenius' Z. Anal. Chem.* **321** (1985) 225 – 234.
176. R. E. Sturgeon, S. N. Willie, S. S. Berman, *Anal. Chem.* **57** (1985) ff 2311.
177. B. Welz, M. Melcher, *Anal. Chim. Acta* **131** (1981) ff 17.
178. H. Berndt, *Spectrochim. Acta Part B* **39 B** (1984) 1121 – 1128.
179. L. Ebdon, M. R. Cave, *Analyst (London)* **107** (1982) 172 – 178.

180. B. Docekal, V. Krivan, *J. Anal. At. Spectrom.* **7** (1992) 521 – 528.
181. D. S. Gough, *Spectrochim. Acta Part B* **54 B** (1999) 2067 – 2072.
182. A. E. Bernhard, *Spectroscopy* **2** (1987) 118.
183. K. Ohls, *Fresenius' Z. Anal. Chem.* **327** (1987) 111 –118.
184. T. Kantor, L. Polos, P. Fodor, E. Pungor, *Talanta* **23** (1976) 585 – 586.
185. H. Berndt, D. Sopczak, *Fresenius' Z. Anal. Chem.* **329** (1987) 18 – 26.
186. M. T. C. De Loos-Vollebregt, L. De Galan, *Prog. Anal. At. Spectrom.* **8** (1985) 47 – 81.
187. S. B. Smith, Jr., G. M. Hieftje, *Appl. Spectrosc.* **37** (1983) 419 – 414.
188. M. Yamamoto, S. Murayama, M. Ito, M. Yasuda, *Spectrochim. Acta Part B* **35 B** (1980) 43 – 50.
189. P. Wirz, H. Debus, W. Hanle, A. Scharmann, *Spectrochim. Acta Part B* **37 B** (1982) 1013 – 1020.
190. G. M. Hermann, *Anal. Chem.* **64** (1992) 571A – 579A.
191. P. Tschöpel, G. Tölg, *J. Trace Microprobe Tech.* **1** (1982) 1 – 77.
192. Analytical Spectrometry Updates, *J. Anal. At. Spectrom.*
193. H. Berndt, E. Jackwerth, *J. Clin. Chem. Clin. Biochem.* **19** (1979) 71.
194. A. Mizuike: *Enrichment Techniques for Inorganic Trace Analysis,* Springer Verlag, Berlin 1983.
195. K. S. Subramanian, *Anal. Chem.* **60** (1988) 11 – 15.
196. J. Szpunar, *Analyst* **125** (2000) 963 – 988.
197. W. M. R. Dirkx, R. Lobinski, F. C. Adams, *Anal. Sci.* **9** (1993) 273 – 278.
198. C. E. Oda, J. D. Ingle, *Anal. Chem.* **53** (1981) 2305 –2309.
199. W. Gerlach, E. Schweitzer: *Die chemische Emissionsspektralanalyse,* vol. 1, L. Voss, Leipzig 1930.
200. G. R. Harrison: *Wavelength Tables,* The M.I.T. Press, Cambridge, Mass., 1991.
201. A. N. Saidel, V. K. Prokofiev, S. M. Raiski: *Spektraltabellen,* V.E.B. Verlag Technik, Berlin 1955.
202. A. Gatterer, J. Junkes: *Atlas der Restlinien,* **vols. 3,** Specola Vaticana, Citta del Vaticano, 1945, 1947, 1959.
203. E. W. Salpeter: *Spektren in der Glimmentladung von 150 bis 400 nm,* vols. 1 – 5, Specola Vaticana, Citta del Vaticano 1973.
204. R. K. Winge, V. A. Fassel, V. J. Peterson, M. A. Floyd: *ICP Atomic Emission Spectroscopy,* Elsevier, Amsterdam 1985.
205. P. W. J. M. Boumans: *Line Coincidence Tables for ICP-OES,* vols. 1 and 2, Pergamon Press, Oxford 1984.
206. B. Huang et al.: *An Atlas of High Resolution Spectra for Rare Earth Elements for ICP – AES,* Royal Society of Chemistry, Cambridge 2000.
207. G. Scheibe: *Chemische Spektralanalyse, physikalische Methoden der analytischen Chemie,* **vol. 1,** 1933.
208. A. Lomakin, *Z. Anorg. Allg. Chem.* **75** (1930) 187.
209. *Ullmann,* 4th ed., **5,** 441 – 500.
210. K. Laqua, W.-D. Hagenah, H. Wächter, *Fresenius' Z. Anal. Chem.* **221** (1967) 142 – 174.
211. A. Aziz, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **36 B** (1981) 251 – 260.
212. N. W. H. Addink: *DC Arc Analysis,* MacMillan, London 1971.
213. M. Z. Riemann, *Fresenius' Z. Anal. Chem.* **215** (1966) 407 – 424.
214. R. Avni in E. L. Grove (ed.): *Applied Atomic Spectroscopy,* vol. 1, Plenum Press, New York 1978.
215. K. Slickers: *Die automatische Atom-Emissions-Spektralanalyse,* Brühlsche Universitätsdruckerei, Gießen 1992.
216. K. Slickers: *Automatic Atomic Emission Spectroscopy,* Brühlsche Universitätsdruckerei, Gießen 1993.
217. K. H. Koch, *Spectrochim. Acta Part B* **39 B** (1984) 1067 – 1079.
218. K. Tohyama, J. Ono, M. Onodera, M. Saeki: *Research and Development in Japan,* Okochi Memorial Foundation, 1978, pp. 31 – 35.
219. D. M. Coleman, M. A. Sainz, H. T. Butler, *Anal. Chem.* **52** (1980) 746 – 753.
220. A. Montaser: *Inductively Coupled Plasmas in Analytical Atomic Spectrometry,* 2nd ed., Wiley-VCH, New York 1998.
221. S. J. Hill (ed.): *Inductively Coupled Plasma Spectrometry and its Applications,* Sheffield Academic Press, Sheffield 1999.
222. V. V. Korolev, E. E. Vainshtein, *J. Anal. Chem. USSR (Engl. Transl.)* **14** (1959) 658 – 662.
223. M. Margoshes, B. Scribner, *Spectrochim. Acta* **15** (1959) 138 – 145.
224. E. Kranz in: *Emissionsspektroskopie,* Akademie-Verlag, Berlin 1964, p. 160.
225. J. Reednick, *Am. Lab.* 1979, no. 5, 127 – 133.

226. R. J. Decker, *Spectrochim. Acta Part B* **35 B** (1980) 19 – 35.
227. F. Leis, J. A. C. Broekaert, H. Waechter, *Fresenius' Z. Anal. Chem.* **333** (1989) 2 – 5.
228. G. W. Johnson, H. E. Taylor, R. K. Skogerboe, *Spectrochim. Acta Part B* **34 B** (1979) 197 – 212.
229. T. B. Reed, *J. Appl. Phys.* **32** (1961) 821 – 824.
230. S. Greenfield, I. L. Jones, C. T. Berry, *Analyst (London)* **89** (1964) 713 – 720.
231. R. H. Wendt, V. A. Fassel, *Anal. Chem.* **37** (1965) 920 – 922.
232. R. Rezaaiyaan et al., *Appl. Spectrosc.* **36** (1982) 626 –631.
233. D. J. Kalnicki, V. A. Fassel, R. N. Kalnicky, *Appl. Spectrosc.* **31** (1977) 137 – 150.
234. B. Raeymaekers, J. A. C. Broekaert, F. Leis, *Spectrochim. Acta Part B* **43 B** (1988) 941 – 949.
235. I. Ishii, A. Montaser, *Spectrochim. Acta Part B* **46 B** (1991) 1197 – 1206.
236. P. W. J. M. Boumans, F. J. de Boer, *Spectrochim. Acta Part B* **32 B** (1977) 365 – 395.
237. L. De Galan, *Spectrochim. Acta Part B* **39 B** (1984) 537 – 550.
238. M. Huang, D. S. Hanselmann, P. Y. Yang, G. M. Hieftje, *Spectrochim. Acta Part B* **47 B** (1992) 765 –786.
239. J. A. C. Broekaert, F. Leis, *Anal. Chim. Acta* **109** (1979) 73 – 83.
240. F. H. Walters, L. R. Parker, Jr., S. L. Morgan, S. N. Deming: *Sequential Simplex Optimization,* CRC Press, Inc., Boca Raton 1991.
241. G. L. Moore, P. J. Humphries-Cuff, A. E. Watson, *Spectrochim. Acta Part B* **39 B** (1984) 915 – 929.
242. M. L. Parsons, A. R. Forster, D. Anderson: *An Atlas of Spectral Interferences in ICP Spectroscopy,* Plenum Press, New York 1980.
243. M. R. Cave, O. Butler, J. M. Cook, M. S. Cresser, L. M. Garden, D. L. Miles, *J. Anal. At. Spectrom.* **15** (2000) 181 – 235.
244. U.S. Environmental Protection Agency: *Methods for the Determination of Metals in Environmental Samples,* EPA600491010, Washington DC 1991.
245. U.S. Environmental Protection Agency: *Methods for the Determination of Metals in Environmental Samples,* Supplement 1, EPA600R94111, Washington DC 1994.
246. B. Fairman, M. Hinds, S. M. Nelms, D. M. Penny, P. Goodall, *J. Anal. At. Spectrom.* **14** (1999) 1937 – 1969.
247. R. E. Russo, X. L. Mao, O. V. Borisov, *Trends Anal. Chem.* **17** (1998) 461.
248. W. B. Barnett, V. A. Fassel, R. N. Kniseley, *Spectrochim. Acta Part B* **23 B** (1968) 643 – 664.
249. S. A. Meyers, D. H. Tracy, *Spectrochim. Acta Part B* **38 B** (1983) 1227 – 1253.
250. P. W. J. M. Boumans, M. C. Lux-Steiner, *Spectrochim. Acta Part B* **37 B** (1982) 97 – 126.
251. J. A. C. Broekaert, F. Leis, K. Laqua, *Talanta* **28** (1981) 745 – 752.
252. K. Vermeiren, C. Vandecasteele, R. Dams, *Analyst (London)* **115** (1990) 17 – 22.
253. K. Dittrich et al., *J. Anal. At. Spectrom.* **3** (1993) 1105 – 1110.
254. A. G. Cox, L. G. Cook, C. W. McLeod, *Analyst (London)* **110** (1985) 331 – 333.
255. H. G. C. Human, R. H. Scott, A. R. Oakes, C. D. West, *Analyst (London)* **101** (1976) 265 – 271.
256. A. Lemarchand, G. Labarraque, P. Masson, J. A. C. Broekaert, *J. Anal. At. Spectrom.* **2** (1987) 481 – 484.
257. T. Ishizuka, Y. Uwamino, *Spectrochim. Acta Part B* **38 B** (1983) 519 – 527.
258. I. B. Brenner et al., *Spectrochim. Acta Part B* **36 B** (1981) 785 – 797.
259. J. R. Garbarino, M. E. Taylor, *Appl. Spectrosc.* **33** (1979) 167 – 175.
260. U.S. Environmental Protection Agency: Inductively Coupled Plasma-Atomic Emission Spectrometric Method for Trace Element Analysis of Water and Wastes, Washington, DC, 1979.
261. DIN 38 406,Bestimmung der 24 Elemente Ag, Al, B, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Na, Ni, P, Pb, Sb, Sr, Ti, V, Zn und Zr durch Atomemissionsspektrometrie mit induktiv gekoppeltem Plasma (ICP – AES) (E 22), Beuth Verlag, Berlin 1987.
262. M. Thompson, B. Pahlavanpour, S. J. Walton, G. F. Kirkbright, *Analyst (London)* **103** (1978) 705 – 713.
263. A. Miyazaki, A. Kimura, K. Bansho, Y. Umezaki, *Anal. Chim. Acta* **144** (1982) 213 – 221.
264. H. Berndt, U. Harms, M. Sonneborn, *Fresenius' Z. Anal. Chem.* **322** (1985) 329 – 333.
265. M. Hiraide et al., *Anal. Chem.* **52** (1980) 804 – 807.
266. J. A. C. Broekaert, S. Gucer, F. Adams: *Metal Speciation in the Environment,* Springer Verlag, Berlin 1990.

267. R. K. Skogerboe, G. N. Coleman, *Anal. Chem.* **48** (1976) 611 A – 622 A.
268. A. E. Croslyn, B. W. Smith, J. D. Winefordner, *Crit. Rev. Anal. Chem.* **27** (1997) 199 – 255.
269. R. Mavrodineanu, R. C. Hughes, *Spectrochim. Acta* **19** (1963) 1309 – 1307.
270. A. Disam, P. Tschöpel, G. Tölg, *Fresenius' Z. Anal. Chem.* **310** (1982) 131 – 143.
271. S. Hanamura, B. W. Smith, J. D. Winefordner, *Anal. Chem.* **55** (1983) 2026 – 2032.
272. C. I. M. Beenakker, *Spectrochim. Acta Part B* **32 B** (1977) 173 – 178.
273. D. Kollotzek, P. Tschöpel, G. Tölg, *Spectrochim. Acta Part B* **37 B** (1982) 91 – 96.
274. G. Heltai, J. A. C. Broekaert, F. Leis, G. Tölg, *Spectrochim. Acta Part B* **45 B** (1990) 301 – 311.
275. P. S. Moussanda, P. Ranson, J. M. Mermet, *Spectrochim. Acta Part B* **40 B** (1985) 641 – 651.
276. D. Kollotzek et al., *Fresenius' Z. Anal. Chem.* **318** (1984) 485 – 489.
277. P. Uden: *Element-Specific Chromatographic Detection by Atomic Emission Spectroscopy,* American Chemical Society, Washington 1992.
278. J. Hubert, M. Moisan, A. Ricard, *Spectrochim. Acta Part B* **33 B** (1979) 1 – 10.
279. U. Richts, J. A. C. Broekaert, P. Tschöpel, G. Tölg, *Talanta* **38** (1991) 863 – 869.
280. Q. Jin, C. Zhu, M. W. Borer, G. M. Hieftje, *Spectrochim. Acta Part B* **46 B** (1991) 417 – 430.
281. R. M. Barnes, E. E. Reszke, *Anal. Chem.* **62** (1990) 2650.
282. K. A. Forbes, E. E. Reszke, P. C. Uden, R. M. Barnes, *J. Anal. At. Spectrom.* **6** (1991) 57.
283. R. K. Marcus (ed.): *Glow Discharge Spectroscopies,* Plenum Press, New York 1993.
284. K. I. Zil'bershtein: *Spectrochemical Analysis of Pure Substances,* A. Hilger, Bristol 1977, pp. 173 – 227.
285. B. Thelin, *Appl. Spectrosc.* **35** (1981) 302 – 307.
286. S. Caroli: *Improved Hollow Cathode Lamps for Atomic Spectroscopy,* Ellis Horwood, Chichester 1985.
287. H. Falk, E. Hoffmann, C. Lüdke, *Spectrochim. Acta Part B* **39 B** (1984) 283 – 294.
288. W. Grimm, *Spectrochim. Acta Part B* **23 B** (1968) 443 – 454.
289. J. B. Ko, *Spectrochim. Acta Part B* **39 B** (1984) 1405 –1423.
290. F. Leis, J. A. C. Broekaert, K. Laqua, *Spectrochim. Acta Part B* **42 B** (1987) 1169 – 1176.
291. A. Bengtson, *Spectrochim. Acta Part B* **40 B** (1985) 631 – 639.
292. R. M. Lowe, *Spectrochim. Acta Part B* **31 B** (1978) 257 – 261.
293. N. P. Ferreira, J. A. Strauss, H. G. C. Human, *Spectrochim. Acta Part B* **38 B** (1983) 899 – 911.
294. J. A. C. Broekaert, R. Pereiro, T. K. Starn, G. M. Hieftje, *Spectrochim. Acta Part B* **48 B** (1983) 1207 –1220.
295. M. R. Winchester, C. Lazik, R. K. Marcus, *Spectrochim. Acta Part B* **46 B** (1991) 483 – 499.
296. T. M. Sugden in R. L. Reed (ed.): *Mass Spectrometry,* Academic Press, New York 1965, p. 347.
297. R. S. Houk et al., *Anal. Chem.* **52** (1980) 2283 –2289.
298. A. Montaser (ed.): *Inductively Coupled Plasma Mass Spectrometry,* Wiley-VCH, New York 1998.
299. E. H. Evans, J. J. Giglio, T. Castillano, J. A. Caruso: *Inductively Coupled and Microwave Induced Plasmas for Mass Spectrometry,* Royal Society of Chemistry, Cambridge 1995.
300. H. Niu, R. S. Houk, *Spectrochim. Acta Part B* **51 B** (1996) 119 – 815.
301. E. H. Evans, J. J. Giglio, *J. Anal. At. Spectrom.* **8** (1993) 1 –8.
302. A. L. Gray, R. S. Houk, J. G. Williams, *J. Anal. At. Spectrom.* **2** (1987) 13 – 20.
303. M. A. Vaughan, G. Horlick, *Appl. Spectrosc.* **40** (1986) 434 – 445.
304. G. Horlick, S. H. Tan, M. A. Vaughan, C. A. Rose, *Spectrochim. Acta Part B* **40 B** (1985) 1555 – 1572.
305. N. Jakubowski, B. J. Raeymaekers, J. A. C. Broekaert, D. Stüwer, *Spectrochim. Acta Part B* **44 B** (1989) 219 – 228.
306. C. Vandecasteele, M. Nagels, H. Vanhoe, R. Dams, *Anal. Chim. Acta* **211** (1988) 91 – 98.
307. J. S. Crain, R. S. Houk, F. G. Smith, *Spectrochim. Acta Part B* **43 B** (1988) 1355 – 1364.
308. D. Hausler, *Spectrochim. Acta Part B* **42 B** (1987) 63 – 73.
309. N. Jakubowski, I. Feldman, D. Stüwer, H. Berndt, *Spectrochim. Acta Part B* **47 B** (1992) 119 – 129.
310. X. Wang, M. Viczian, A. Lasztity, R. M. Barnes, *J. Anal. At. Spectrom.* **3** (1988) 821 – 828.

311. C. J. Park, J. C. Van Loon, P. Arrowsmith, J. B. French, *Anal. Chem.* **59** (1987) 2191 – 2196.
312. D. Boomer, M. Powell, R. L. A. Sing, E. D. Salin, *Anal. Chem.* **58** (1986) 975 – 976.
313. G. E. M. Hall, J. C. Pelchat, D. W. Boomer, M. Powell, *J. Anal. At. Spectrom.* **3** (1988) 791 – 797.
314. T. E. Jeffries, S. E. Jackson, H. P. Longerich, *J. Anal. At. Spectrom.* **13** (1998) 935 – 940.
315. E. K. Shibuya, J. E. S. Sarkis, J. Enzweiler, A. P. S. Jorge, A. M. G. Figueiredo, *J. Anal. At. Spectrom.* **13** (1998) 941 – 944.
316. J. G. Williams, A. L. Gray, P. Norman, L. Ebdon, *J. Anal. At. Spectrom.* **2** (1987) 469 – 472.
317. N. Jakubowski, I. Feldman, B. Sack, D. Stüwer, *J. Anal. At. Spectrom.* **7** (1992) 121 – 125.
318. P. Arrowsmith, *Anal. Chem.* **59** (1987) 1437 – 1444.
319. D. L. Miles: "Geological Applications of Plasma Spectrometry", in S. J. Hill (ed.): *Inductively Coupled Plasmas in Analytical Atomic Spectrometry,* 2nd ed., Wiley-VCH, New York 1999.
320. F. Vanhaecke, L. Moens, P. Taylor: "Use of ICP – MS for Isotope Ratio Measurements", in S. J. Hill (ed.): *Inductively Coupled Plasmas in Analytical Atomic Spectrometry,* 2nd ed., Wiley-VCH, New York 1999.
321. J. B. Truscott, L. Bromley, P. J. Jones, E. H. Evans, J. Turner, B. Fairman, *J. Anal. At. Spectrom.* **14** (1999) 627 – 631.
322. LM. J. Bloxham, S. J. Hill, P. J. Worsfold, *J. Anal. At. Spectrom.* **9** (1994) 935 – 938.
323. W. W. Harrison, B. L. Bentz, *Prog. Anal. At. Spectrosc.* **11** (1988) 53 – 110.
324. D. J. Hall, K. Robinson, *Am. Lab.* 1987, no. 8, 14.
325. N. Jakubowski, D. Stüwer, G. Tölg, *Int. J. Mass Spectrom. Ion. Processes* **71** (1986) 183 – 197.
326. P. M. Charalambous, *Steel Res.* **5** (1987) 197.
327. W. Vieth, J. C. Huneke, *Spectrochim. Acta Part B* **46 B** (1991) 137 – 154.
328. N. Jakubowski, D. Stüwer, *Fresenius' Z. Anal. Chem.* **335** (1988) 680 – 686.
329. A. Bogaerts, R. Gijbels, *Spectrochim. Acta Part B* **53** (1998) 1.
330. VG Instruments, Technical Information GD 012, Manchester, U.K.
331. VG Instruments, Application Note 02.681, Manchester, U.K.
332. VG Instruments, Technical Information GD 701, Manchester, U.K.
333. N. Jakubowski, D. Stüwer, W. Vieth, *Anal. Chem.* **59** (1987) 1825 – 1830.
334. N. Jakubowski, D. Stüwer, W. Vieth, *Fresenius' Z. Anal. Chem.* **331** (1988) 145 – 149.
335. M. Hecq, A. Hecq, M. Fontignies, *Anal. Chim. Acta* **155** (1983) 191 – 198.
336. W. A. Mattson, B. L. Bentz, W. W. Harrison, *Anal. Chem.* **48** (1976) 489 – 491.
337. G. O. Foss, H. J. Svec, R. J. Conzemius, *Anal. Chim. Acta* **147** (1983) 151 – 162.
338. N. Jakubowski, D. Stüwer, G. Tölg, *Spectrochim. Acta Part B* **46 B** (1991) 155 – 163.
339. N. Omenetto, J. D. Winfordner, *Prog. Anal. At. Spectrom.* **2** (1979) 1 – 183.
340. N. Omenetto, H. G. C. Human, P. Cavalli, G. Rossi, *Spectrochim. Acta Part B* **39 B** (1984) 115 – 117.
341. V. A. Bolshov, A. V. Zybin, II. Smirenkins, *Spectrochim. Acta Part B* **36 B** (1981) 1143 – 1152.
342. N. Omenetto et al., *Spectrochim. Acta Part B* **40 B** (1985) 1411 – 1422.
343. B. W. Smith, P. B. Farnsworth, P. Cavalli, N. Omenetto, *Spectrochim. Acta Part B* **45 B** (1990) 1369 – 1373.
344. S. Grazhulene, V. Khvostikov, M. Sorokin, *Spectrochim. Acta Part B* **46 B** (1991) 459 – 465.
345. M. A. Bolshov, C. F. Boutron, A. V. Zybin, *Anal. Chem.* **61** (1989) 1758 – 1762.
346. W. Sdorra, K. Niemax, *Spectrochim. Acta Part B* **45 B** (1990) 917 – 926.
347. G. C. Turk, J. C. Travis, J. R. DeVoe, *Anal. Chem.* **51** (1979) 1890 – 1896.
348. R. B. Green, M. D. Seltzer: "Laser Induced Ionization Spectrometry", in J. Sneddon (ed.): *Advances in Atomic Spectroscopy,* **vol. I,** JAI Press, London 1992.
349. A. R. Date, A. L. Gray, *Spectrochim. Acta Part B* **40B** (1985) 115 – 122.
350. G. J. Havrilla, S. J. Weeks, J. C. Travis, *Anal. Chem.* **54** (1982) 2566 – 2570.
351. K. Niemax, *Appl. Phys. B* **B 38** (1985) 147 – 157.
352. N. Omenetto, T. Berthoud, P. Cavalli, G. Rossi, *Anal. Chem.* **57** (1985) 1256 – 1261.
353. G. C. Turk, M. DeMing, in W. F. Koch (ed.): *NBS Special Publication 260 – 106,* US Government Printing Office Washington DC 1986, pp. 30 – 33.
354. J. P. Young, R. W. Shaw, D. H. Smith, *Anal. Chem.* **61** (1989) 1271 A – 1279 A.
355. R. Klockenkämper, J. Knoth, A. Prange, H. Schwenke, *Anal. Chem.* **64** (1992) 1115 A – 1123 A.

Laser Analytical Spectroscopy

MICHAEL A. BOLSHOV, Institute of Spectroscopy, Russian Academy of Sciences, Troitzk, Russia; Institute of Spectrochemistry and Applied Spectroscopy, Dortmund, Germany

YURI A. KURITSYN, Institute of Spectroscopy, Russian Academy of Sciences, Troitzk, Russia

1. Introduction 1
1.1. Properties of Laser Light 2
1.1.1. Tunability 2
1.1.2. Intensity 2
1.1.3. Time Resolution 2
1.1.4. Temporal Coherence 2
1.1.5. Spatial Coherence 2
1.2. Types of Laser Analytical Techniques 2
1.3. Peculiarities of Resonance Laser Radiation Absorption 3
2. Tunable Lasers 4
2.1. Dye Lasers 4
2.2. Semiconductor Diode Lasers 5
3. Laser Techniques for Elemental Analysis 6
3.1. Laser-Excited Atomic Fluorescence Spectrometry 7
3.1.1. Physical Principles 7
3.1.2. Atomizers 7
3.1.3. Limits of Detection 8
3.1.4. Real Sample Analysis 9
3.2. Laser-Enhanced Ionization 9
3.2.1. Physical Principles 9
3.2.2. Atomizers 11
3.2.3. Limits of Detection 12
3.2.4. Real Sample Analysis 12
3.3. Resonance Ionization Spectroscopy 13
3.3.1. Physical Principles 13
3.3.2. Atomizers 14
3.3.3. Limits of Detection: Real Samples . . 15
3.4. Diode Lasers in Elemental Analysis 16
3.4.1. Physical Principles of Absorption Spectrometry 16
3.4.2. Atomic Absorption Spectrometry with Diode Lasers 16
3.4.3. Element Specific Detection by DLAAS 18
3.4.4. Isotope Selective Detection with Diode Lasers 18
3.4.5. Commercial System 18
4. Laser Techniques for Molecular Analysis 19
4.1. Molecular Absorption Spectroscopy with Diode Lasers 19
4.1.1. Detection of Molecular Gas Impurities with Lead-Salt Diode Lasers 19
4.1.2. Near IR Gas Sensors 20
4.1.3. Advanced Developments 20
4.2. Laser Optoacoustic Spectroscopy . 21
4.3. Thermal Lens Spectroscopy 22
4.4. Fluorescence Analysis of Organic Molecules in Solid Solutions 23
5. Laser Ablation 26
6. References 27

1. Introduction

Modern technologies in electronics, in high-purity materials and chemical reagents, in toxicology, biology, and environmental control and protection need extremely sensitive analytical techniques. Impurity concentrations at and below pg/mL (pg/g) levels are state-of-the-art for many analyses. In many cases the sample volume is very small; if an impurity concentration of the order of ng/g has to be detected in a sample of mass 1 µg, the analytical technique should have an absolute detection limit of ca. 1 fg (10^{-15} g). Even techniques such as atomic absorption spectrometry (AAS), optical emission spectrometry with inductively coupled plasma (ICP-OES), neutron activation spectroscopy (NAS), and X-ray fluorescence spectroscopy (XRFS) can not detect directly trace elements at such low levels of concentration and/or absolute mass. In most cases, these techniques need elaborate chemical pretreatment of samples. Contamination or loss of analyte are very probable during any chemical procedure at such low concentration.

During the last 30 years the development of laser technology has opened up new possibilities in analytical spectroscopy, both by improvements in traditional analytical techniques,

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_653

such as absorption and fluorescence or optoacoustic spectroscopy, and by the introduction of new techniques such as multistep photoionization spectroscopy, thermal lens spectroscopy, and site-selection spectroscopy.

1.1. Properties of Laser Light

1.1.1. Tunability

Practically all laser analytical techniques are based on systems with tunable wavelength. By using different types of lasers and the methods of nonlinear optics, one can obtain intense tunable laser radiation over the spectral range of 0.2 – 40 µm and in some regions of the vacuum ultraviolet and far infrared. Almost any strong atomic or molecular transition is thus accessible to laser excitation.

1.1.2. Intensity

The intensities of lasers exceed by many orders of magnitude those of conventional light sources. At high intensity levels, nonlinear interaction of light with atomic and molecular systems becomes pronounced. New analytical techniques involving multistep and multiphoton excitation and ionization, optical saturation, and excitation of forbidden transitions become possible. High radiation intensity also increases the sensitivity of laser analytical techniques.

1.1.3. Time Resolution

Pulsed lasers with extremely short pulse durations (down to 10^{-15} s) are now available, enabling investigations of chemical reaction kinetics, energy and electron transfer in complex excited molecules, etc. The efficiency of excitation of higher states increases because the duration of the laser pulse is often shorter than the decay time of the intermediate state of atoms or molecules.

1.1.4. Temporal Coherence

The spectral linewidth of modern lasers is often less than the homogeneous linewidth of a given atom or molecule. This results in very high spectral intensity and thus very efficient excitation. It also allows highly selective excitation of a particular atom or molecule in a complex mixture. High selectivity is an important advantage of laser analytical techniques.

1.1.5. Spatial Coherence

The ability to focus a laser beam to a very small spot (a few µm diameter) provides high spatial resolution in sample analysis. High spatial coherence also enables collimated laser beams to be used for remote sensing.

A review of the physical fundamentals and main characteristics of lasers is given in → Lasers.

1.2. Types of Laser Analytical Techniques

Laser analytical techniques can be classified on the basis of the physical process caused by the interaction of laser radiation with matter (Fig. 1).

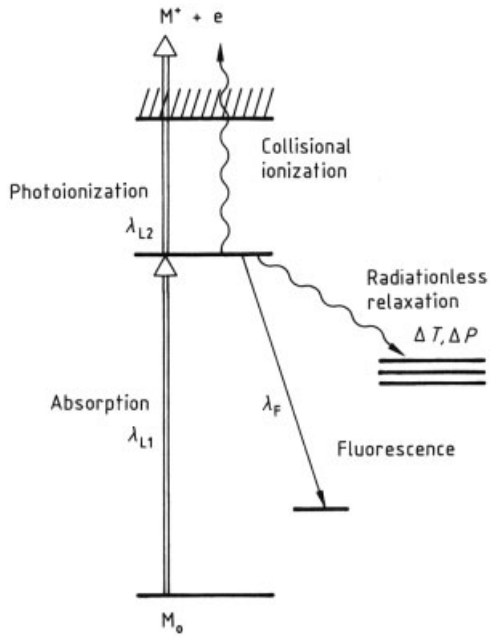

Figure 1. The basic laser analytical techniques

The primary process is absorption of photons by atoms or molecules. If the absorption lines are strong and the number density of the absorbing species is sufficient, the difference between incident and transmitted energies may be measurable for a single pass of the laser beam through the sample. In this case the signal can be measured by a simple light detector, such as a photodiode. For weaker absorption lines and lower number densities of the absorbing species multipass schemes are used.

Another set of analytical techniques (optocalorimetric methods) is based on indirect measurement of the energy absorbed by the sample. Absorbed energy is transformed into kinetic energy of the sample particles followed by an increase in the local temperature of the sample. The resultant changes in pressure and refractive index can be measured by optoacoustic methods or by laser scattering techniques.

The absorption of laser photons leads to excitation of atoms or molecules. In many cases, radiative decay of the excited state is very efficient, i.e., quantum yields of luminescence or fluorescence are very high. Measurement of the intensity of emitted photons as a function of analyte concentration is a widespread analytical method with conventional light sources. The use of intense laser sources greatly increased the sensitivity and selectivity of traditional luminescence methods.

Excitation of atoms or molecules may be followed by collisional or multistep (multiphoton) ionization of the excited species. The charged particles so formed may be recorded by conventional methods. Several techniques involve a combination of optical excitation of a neutral particle and its subsequent ionization.

1.3. Peculiarities of Resonance Laser Radiation Absorption

The interaction of intense laser radiation with matter differs from that of conventional radiation sources. The high incident irradiances available with lasers can lead to a significant depletion of the initial level population. As a result, the optical absorption coefficient decreases as a function of the excitation energy and the absorbed energy tends towards a constant value. This so called optical saturation effect leads to a nonlinear dependence of the absorption signal on the light intensity. Because the amount of energy released by radiative and nonradiative relaxation processes depends on the amount of energy absorbed, nonlinear intensity-dependent signals are also obtained in other spectrometric techniques, such as fluorescence, optoacoustic, thermal lens, etc.

In the simplest case for the interaction of a two-level system with cw (continuous wave) laser radiation (steady-state mode), the number of transitions per unit time from level 1 to level 2 (absorption) is equal to the number of transitions from level 2 to level 1 (stimulated and spontaneous emission):

$$\sigma\,(N_1 - N_2)\,I/hv = N_2/\tau$$

where σ_0 is the cross-section of the radiation (stimulated) transition, N_1 and N_2 are the number densities of particles at the levels 1 and 2, respectively, I is the intensity of laser radiation, hv is the energy of the resonance photon, τ is the radiative decay time of upper state 2 (assuming that non-radiative transitions due to collisions are negligible). Both N_1 and N_2 are dependent on laser intensity, so that

$$N_2 = \frac{1}{2} N \frac{\frac{I}{I_{\mathrm{sat}}}}{1 + \frac{I}{I_{sat}}},\ N_1 = N - N_2$$

where N is the total number density of particles, $I_{\mathrm{sat}} = hv/2\sigma_0\tau$

is the saturation intensity, defined as the intensity for which the number density of the excited particles N_2 reaches one half of its maximum value $1/2\,N$.

If the intensity of laser radiation is much lower than I_{sat}, the absorption energy and the number density of the excited particles N_2 are directly proportional to the light intensity, and hence the analytical signal for different techniques (absorption, optocalorometric, fluorescence, etc.) has a linear dependence on the intensity of the radiation. This linear regime is typical for conventional radiation sources used in elemental analysis: hollow cathode lamps (HCLs), electrodeless discharge lamps (EDLs) and for low intensity lasers.

In the opposite case, when the intensity of the resonance laser radiation is high enough ($I \gg I_{\mathrm{sat}}$), the absorbed laser energy reaches its

maximal value and the population of the upper level 2 approaches *1/2 N*. The analytical signal for the absorption technique

$$\Delta I \approx \sigma_0 (N_1 - N_2) lI = \frac{\sigma_0 NlI}{1 + I/I_{\mathrm{sat}}}$$

(where l is the absorption pathlength) tends towards a constant value. Because the shot noise of radiation is directly proportional to $\sqrt{I}$, then optimum detection of the absorbing particles is achieved at $I = I_{\mathrm{sat}}$.

If deep saturation is achieved ($I \gg I_{\mathrm{sat}}$), two very important advantages of different laser analytical techniques (fluorescence, optocalorimetric, ionization) are gained: Maximum analytical signal and minimal influence of intensity fluctuations of the excitation source on the precision of the analysis. The ultimate sensitivity down to single atom/molecule detection has been demonstrated under saturation conditions in many analytical applications with lasers.

In real analytes the scheme of the energy levels is rather complicated. Thus, there may be many channels of radiative and collisional transfer of the absorbed laser energy. The complex structure of the energy levels scheme offers different combinations of excitation and detection of a particular species, but nevertheless the simplest scheme described above gives correct qualitative presentation of the analytical signal behavior.

Details on the optimization of laser parameters for different analytical techniques can be found in [1], [2], and [3].

2. Tunable Lasers

In most analytical applications, tunable lasers are used. Tunability allows selective excitation of a particular analyte in a complex mixture and relatively simple "switching" from one analyte to another. For conventional sources with fixed spectral lines it is often necessary to use different lamps for different analytes.

For a few laser analytical techniques (molecular fluorimetry of liquids, spontaneous Raman scattering), fixed wavelength lasers are used. The following discussion is limited to the two most widely used types of tunable laser.

2.1. Dye Lasers [4]

For a long time the dye lasers were the most popular type for atomic and molecular analysis. The active medium of these lasers is an organic dye in an appropriate solvent. About 20 different dyes cover the spectral range from ca. 320 to ca. 900 nm.

Typically, dye lasers are pumped by an intense fixed-frequency laser, the most often used being nitrogen (337 nm), argon ion (488.0, 514.5 nm), XeCl excimer (308 nm), and frequency-doubled and frequency-tripled Nd-YAG (530, 353.3 nm).

Dye molecules have broad emission bands. The optical resonator of a dye laser includes a spectrally selective element (grating, prism, interferometer, Lyot filter, etc.) which narrows the dye laser spectral line. By appropriate adjustment (e.g., rotation of the grating) one can smoothly tune the laser wavelength within the luminescence band of a particular dye. The tuning range may be 30 – 40 nm. At the extremes of the tuning range the energy of a dye laser falls off.

The organic dyes used for lasers can be classified on the basis of molecular structure: terphenyls (spectral range: 312 – 400 nm); coumarines (420 – 580 nm); rhodamines (570 – 720 nm), oxazines (680 – 860 nm). Each class consists of several dyes with slightly modified chemical structure and, hence, slightly shifted absorption and luminescence bands. There are special dyes for UV and near IR.

The most utilized solvents are ethanol, methanol, cyclohexane, ethylene glycol, and dioxane. For different solvents, the tuning curves of the same dye may be shifted by 10 – 15 nm.

The dyes for the UV spectral range (below 380 – 400 nm) are not photochemically stable. In most analytical applications, intense UV radiation below 360 nm is obtained by nonlinear optics, i.e., the generation of harmonics and sum frequencies of the visible dye laser radiation in nonlinear crystals. The most widely employed crystals are KDP, ADP, LiF, $LiIO_3$, and BBO. For example, the second harmonic of a rhodamine dye laser (fundamental band near 640 nm) is a more practical way to obtain tunable UV radiation near 320 nm rather than direct lasing of an unstable UV dye laser.

The efficiency of energy conversion of the pumping laser radiation in commercially available lasers varies from a few percent to ca. 20 %, depending on the type of dye and the spectral range.

2.2. Semiconductor Diode Lasers [5]

In the mid-1960s, tunable semiconductor diode lasers (DLs) were introduced into spectroscopic and analytical research. In conventional DLs the laser action can occur when a forward bias current is applied to a $p-n$ junction. Photons are generated in the process of electron – hole recombination. The wavelength of laser radiation is determined by the energy gap between conduction and valence bands, which, in turn, is determined by the chemical composition of the semiconductor. Wavelength tuning is possible by varying the injection current and by the temperature of the laser.

At first, DLs based on the IV – VI (Pb-salt) materials were used for analytical applications. These lasers operate in the 3 – 30 µm region, where the fundamental absorption bands of most molecules lie. Linewidths as small as $2 \times 10^{-6}\,\mathrm{cm}^{-1}$ were observed for the DLs, but in the commercial systems used for analytical applications the DL linewidth is determined by the quality of temperature stabilization. Usually the temperature is stabilized at a level of 10^{-2} – 10^{-3} K, which corresponds to a linewidth of $\sim 10^{-3}\,\mathrm{cm}^{-1}$. This width is much smaller than the widths of molecular absorption lines at atmospheric pressure and even smaller than Doppler broadening for light molecules. The IV – VI DLs operate at cryogenic temperatures. In laboratory high-resolution spectrometers operating temperatures of 10 – 120 K are usually provided by closed-cycle Stirling microrefrigerators. Now lasers operating above 77 K are available at most wavelengths, and in analytical instruments temperature-controlled liquid-N_2 Dewars are the cooling method of choice. Typical emitting powers are 100 – 300 μW, and cw lasers with powers of 1 mW can also be fabricated.

The characteristics of DLs operating in the visible/near-IR (NIR) region have been greatly improved. High quality lasers, made from the III-V group of semiconductor materials, are now commercially available in a number of discrete spectral ranges between 630 nm and 2.0 µm. In principle, lasers can be made at any wavelength, but because of the market demand the DL are commercialized mainly for specific spectral ranges. These lasers can work in cw mode at temperatures accessible with thermoelectric coolers, generally − 40 to 60 °C. The typical linewidth of a free-running DL is 20 – 100 MHz. Commercially available edge-emitting Fabry – Perot (FP) DLs yield singe-mode radiation up to 150 mW in the 670 – 870 nm spectral range and up to 30 mW in the 630 – 670 nm range. DLs, working in the 630 – 780 nm region, have been shown to be very attractive for atomic absorption analyzers.

Compared with other lasers, the output radiation from DLs is highly divergent. The emission angles are typically 10° and 30° parallel and perpendicular to the laser junction, respectively, and may vary from sample to sample. These beam features can be corrected with appropriate collimating and anamorphic beam-conditioning optics.

The drawbacks of the FP-DLs are the frequency drift and mode jumping, therefore other types of lasers such as distributed feedback (DFB), distributed Bragg reflector (DBR), external cavity (ECDL), and vertical cavity surface emitting lasers (VCSELs) have been developed. In a typical DFB laser, an index grating is fabricated directly on a chip alongside the active layer, providing distributed feedback only for a selected wavelength specified by the grating. A DBR laser uses the same principle, but the grating is placed beyond the active region. DFB/DBR lasers provide single-mode operation, can be continuously tunable by current and temperature, and exhibit no mode hops. The drawbacks are limited tuning range and higher cost compared with simple FP lasers.

Wavelengths longer than 1 µm are attractive for molecular analysis. Commercially available DLs for this region include: InGaAsP/InP (1.3, 1.55 µm FP structures, 1.2 – 1.8 µm DFB and EC lasers). For wavelengths longer than 1.8 µm, DLs are made from compounds containing antimonides such as AlGaAsSb, InGaAsSb, and InAsSb. Lasing in the 2 – 2.4 µm region at room temperature is available, and the spectral range up to 3.7 µm at liquid N_2 is achieved.

Vertical cavity surface emitting lasers (VCSELs), in contrast to conventional FP lasers,

emit from their top surface. A conventional VCSEL consists of a multi-quantum-well (MQW) active region placed in a cavity between two DBR mirrors. These mirrors control the wavelength and beam divergence. For typical VCSELs, the far field beam characteristic will be a near Gaussian shape, with a beam divergence less than 12°. The narrow beam divergence, circular symmetry, and lack of astigmatism greatly simplify the optical design with these type of DLs. VCSELs operating in cw mode at room temperature in the visible/NIR range up to 2.3 µm can be produced.

An alternative approach for tunable single-mode operation is to build an external cavity. In this case, a FP laser is placed into an optical cavity, and the wavelength is tuned by using a dispersive element, such as a grating. The wavelength can be tuned continuously over many wavenumbers ($20-50\,\mathrm{cm}^{-1}$) by rotating the grating. The linewidth depends on the construction of the external cavity and its stability and is usually less than 1 MHz.

To obtain generations in the regions of mode-hops for conventional FP lasers a simple design with a small external cavity was proposed. A glass plate mounted in front of the diode chip acts as the low-reflectivity mirror of an extended resonator and helps to select particular longitudinal modes within gain profile, even where laser action is not possible without the feedback.

Recently developed multi-electrode lasers are very promising devices for multielement sensors. These lasers separate the wavelength selection, gain, and phase control regions of the laser cavity and are available in a number of different configurations. A tuning range over 100 nm around 1.5 µm has been demonstrated. These tuning ranges are equivalent to EC devices, thereby simplifying present multi-laser sensor concepts by allowing comparable wavelength coverage with a single laser.

Frequently conversion is used to extend the available spectral range of DLs. Second harmonic generation (SHG) in bulk nonlinear crystals, such as $LiIO_3$ and $KNbO_3$, can provide up to 0.1 μW of radiation in the 335 – 410 nm range and up to 1 – 3 μW in the 410 – 430 nm range. The generation of sum frequencies (SFG) in nonlinear crystals can open the way to tunable radiation in the UV/visible part of the spectrum. Over the past several years, progress in the development of quasi-phase matching structures (QPMS) and waveguides, based on nonlinear crystals, has been achieved. Conversion efficiencies in both types of such nonlinear converters are much higher than in bulk crystals, thus single-mode powers up to some mW in the UV/blue range can be produced. Mid-IR radiation can be also produced by difference-frequency mixing of two NIR lasers. Difference-frequency generation in periodically poled $LiNbO_3$ and bulk $AgGaSe_2$ or $AgGaS_2$ has been demonstrated for generating tunable radiation between 3.2 and 8.7 µm.

In the mid-IR, very promising substitutes to lead-salt lasers are quantum-cascade lasers (QCLs), made using III-V compounds. In QCL the active region consists of several gain quantum well (QW) layers separated by the digitally graded electron-injecting regions. Electrons are injected into the high energy level of the first gain layer, fall down to the lower level with photon emission, then through the digitally graded region are injected into the high level of the next gain layer and the process repeats. Quantum efficiencies higher than 100 % are possible in such structures. It is important that such cascades can be organized between quantized states in the conduction band only. The close spacing of these quantized states allows light to be emitted in the mid-IR range. Currently, room temperature lasing up to 8.2 µm in structures fabricated from AlInAs/InGaAs can be obtained. QCLs with wavelength as long as 11 µm at cryogenic temperatures have been demonstrated.

In 1999 blue DLs became commercially available. InGaN MQW DLs were demonstrated to have a lifetime of more than 10 000 h under conditions of room temperature cw operation. Single-mode emissions were produced in the region of 390 – 450 nm. Maximum output power is about 30 mW. With an external cavity the narrowband 1 MHz laser linewidth can be coarsely adjusted over a tuning range of 2.5 nm and then fine-tuned by a piezo stack over a more than 20 GHz range without mode hops.

3. Laser Techniques for Elemental Analysis

In this chapter we shall discuss laser analytical techniques for elemental analysis, i.e., for detec-

tion of impurity atoms of a particular element in a sample.

3.1. Laser-Excited Atomic Fluorescence Spectrometry [2], [3], [6]

3.1.1. Physical Principles

The principles of the laser-excited atomic fluorescence (LEAF) technique are very simple. A liquid or solid sample is atomized in an appropriate device. The atomic vapor is illuminated by laser radiation tuned to a strong resonance transition of an analyte atom. The excited analyte atoms spontaneously radiate fluorescence photons and a recording system registers the intensity of fluorescence (or total number of fluorescent photons). The extremely high spectral brightness of lasers makes it possible to saturate a resonance transition of an analyte atom. Therefore, the maximum fluorescence intensity of the free analyte atoms can be achieved while the effect of intensity fluctuations of the excitation source are minimized. Both factors provide the main advantage of LEAF – extremely high sensitivity. The best absolute detection limits achieved in direct analysis by LEAF lie in the femtogram-range, which corresponds to concentrations down to pg/mL.

Several different excitation – recording schemes used in LEAF are shown in Figure 2. The choice of a particular scheme is dictated by the structure of the atomic energy levels, the sources of background and noise, and spectral interference with the major sample components.

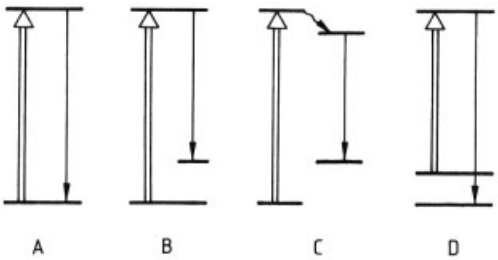

Figure 2. Classification of LEAF schemes
A) Resonance fluorescence; B) Direct-line fluorescence; C) Collisionally assisted direct-line fluorescence; D) Anti-Stokes direct-line fluorescence

The resonance scheme (Fig. 2A) has the great advantage of a possible cyclic interaction of analyte atoms with the laser radiation and, hence, a high yield of fluorescence photons. The highest sensitivity of LEAF, the detection of a single free atom, has been demonstrated with Na atomic beams and cw laser excitation. This scheme has a serious disadvantage – a high level of scattered laser light and, hence, high background. For a few elements (Na, Cd, etc.) the resonance scheme is the only possible one. For the most elements this scheme is not optimal in real analytical conditions when pulsed dye lasers are used for excitation.

The Stokes direct-line or collisionally assisted Stokes direct-line schemes (Fig. 2B, C) are often employed (Tl, Pb, Au, Ag, Co, etc.). Their main advantage is that different wavelengths are used for excitation and fluorescence. Because of this, the laser scattered light can be easily separated from the weak fluorescence radiation by cut-off optical filters or monochromators. The analytical potential of collisionally assisted detection schemes depends strongly on the efficiency of collisional mixing between the neighboring levels. Typically this efficiency is high enough for energy gaps between the levels of ca. 1000 cm^{-1}. For some elements with low-lying excited states (Ga-type) the anti-Stokes direct-line scheme (Fig. 2D) is very useful. An advantage of this scheme is a blue shift of the fluorescence and an appropriate reduction in background, caused by scattered laser light. In some cases, double-resonance excitation by two lasers with different wavelengths is used.

The possibility to select the optimal excitation – detection scheme often simplifies the problem of spectral interference with the major components, providing the second advantage of LEAF – high selectivity.

3.1.2. Atomizers

Graphite electrothermal atomizers (GETA), flames, glow discharges, and laser plumes are the most widely used types of atomizers in modern LEAF analytical practice.

GETA. Graphite electrothermal atomizers are of two types: open and closed. The main unit is a graphite crucible (cup, profiled rod, boat, furnace) held between massive water- cooled electrodes. Typical sample volumes are 2 – 100 µL.

The maximum temperature of a crucible depends on its mass, the electrical power supply, and the type of graphite used. In most applications, it is 2700 – 2900 °C. In open and closed atomizers an inert gas atmosphere is used to minimize quenching of the excited analyte atoms and sublimation of the hot graphite crucible. Graphite furnaces have proved preferable because of the high temperature of the analytical zone and, hence, less chemical interference. Low-pressure noble gas in a closed GETA has been used to minimize matrix interference.

Flames. The flames used in LEAF are very similar to those used in AAS. In most applications, cylindrically symmetric burners are used since the dimensions of the analytical zone in LEAF do not exceed 10 mm. Quenching of the excited atoms by components of the flame is a serious problem. As oxygen is one of the most effective quenching agents, flames with minimum oxygen concentration should be used (e.g., an oxygen – argon – hydrogen flame). The high temperature of the analytical zone increases not only the fluorescence signal, but also the background, caused by flame emission. The optimal position of the analytical zone on the flame axis maximizes the signal-to-noise ratio. The background radiaton is usually much higher in flames as compared with GETA because of fluctuations in the optical density of the analytical zone (Rayleigh scattering of the excited laser radiation), luminescence of organic compounds in the flame, Mie scattering by dust particles, etc. As a result detection limits are typically much worse for flame-LEAF than for GETA-LEAF. On the other hand, the relative simplicity of design and the stationary analytical regime are definite advantages of flames.

Glow Discharges. Glow discharges (GD) have proved very useful atomizers for direct analysis of solid samples by LEAF. In GD the surface of a solid sample is bombarded by ions of a noble gas, which is continuously pumped through the atomizer at low pressure. Depending on the type of GD, the gas pressure ranges from 10^{-1} Pa to 10 kPa. At such pressures collisions between analyte or matrix atoms and atoms in the gas phase are relatively scarce, which substantially reduces the probability of gas phase reactions and fluorescence quenching. The sample to be analyzed is used as the cathode. The anode – cathode voltage is typically several hundred volts, the energy range of the bombarding ions is $10^{-17} - 10^{-15}$ J (0.1 – 10 keV). Both direct current and pulsed discharges are used. In the latter case the discharge is synchronized with the laser pulse, with an appropriate time delay, depending on the distance between the sample surface and the analytical zone and the mean energy of the sputtered atoms.

The process has substantial advantages over thermal atomization. The sputtering yields of different elements differ by no more than one order of magnitude, whereas their vapor pressures may differ by several orders of magnitude. Sputtering yields are much less sensitive to the chemical composition of the matrix as compared to thermal atomization efficiencies. Ion sputtering avoids the drawbacks of GETA (e.g., chemical interference, recombination of free atoms into molecules and clusters, carbide formation). The problem of surface contamination can be easily solved by preliminary cleaning in a discharge. After the surface layer is removed, sample sputtering stabilizes and the bulk composition of a homogeneous sample or the layer-by-layer composition of an inhomogeneous one can be analyzed.

3.1.3. Limits of Detection

Like most spectroscopic analytical techniques, LEAF needs a calibration procedure. In common with other modern analytical techniques, determination of limits of detection (LOD) poses a new problem. Commonly, LOD are defined as the concentration (or absolute amount) of analyte that gives an analytical signal three times the standard deviation of the blank. The analyte concentration in the purest solvent used for blank measurements can often be detected by modern analytical techniques. One should differentiate the LOD determined by the limited purity of the solvent from the limit determined by the intrinsic characteristics of the analytical technique (nature of background, electronic noise, recording sensitivity, etc.). The first LOD is measured with the laser tuned to the resonance line of the analyte (on-line LOD). The second is measured with the laser detuned (off-line LOD). The detection power of most laser analytical techniques is very

high and most published LOD are off-line values.

LEAF LODs are listed in Table 1 [7].

3.1.4. Real Sample Analysis

Trace amounts of Pb, Cd, and Bi at femtogram levels have been directly determined in ancient ice and recent snows in the Antarctic by LEAF – GETA. Toxic metal concentrations in deep Antarctic ice cores give unique information about the preindustrial tropospheric cycles of these metals and about recent anthropogenic sources of heavy metals. Concentrations of Pb and Cd in the samples analyzed were 0.3 – 30 pg/mL and 0.08 – 2 pg/mL, respectively. The values measured by LEAF – GETA were in good agreement with those measured by isotope dilution mass spectrometry (Pb) and graphite furnace AAS (Cd). Both traditional techniques required elaborate chemical pretreatment and large sample volumes (50 – 200 mL). The sample volume in LEAF experiments was 20 – 50 µL.

The most successful analysis of real samples by LEAF was done with sample atomization in a graphite furnace. The high detection power (absolute LODs at and below femtogram levels) and wide linear range (up to 5 – 7 orders of magnitude) of LEAF were combined with modern furnace technology: Fast heating, platform atomization, matrix modification, and direct solid sampling. As a result low LODs, reasonable accuracy and precision, less matrix interference, and simpler sample pretreatment (if necessary) were achieved (Table 2).

The potential of LEAF with glow discharge atomizers (LEAF-GD) have been demonstrated. The LODs for Pb and Ir as low as 0.1 and 6 ng/mL, respectively, were achieved with hollow cathode GD and aqueous standards. With a similar atomizer 40 ng/g LOD was obtained for direct analysis of Pb traces in Cu-based alloys. With the more efficient magnetron GD even lower LODs were achieved: 3 ng/g for Pb in Cu alloys, 0.8 ng/g and 2 ng/g for Si in In and Ga, respectively.

Analysis of pure gases by LEAF is much less developed. Nevertheless, some very impressive results have been demonstrated. Traces of Pb and Tl in the air of a normal laboratory and a class-100 clean room were detected by LEAF. Air from the room was pumped through a tantalum jet, fixed in a hole inside a graphite tube. The heavy particles impacted onto the tube wall opposite the jet. After pumping a certain volume of ambient air in a particular time interval the impacted metal traces were detected by graphite furnace LEAF. LODs of 0.1 pg/m^3 and 0.01 pg/m^3 for Pb and Tl, respectively, were achieved. The Pb LOD is two orders of magnitude better than the LOD for graphite furnace AAS. The measured concentrations of Pb and Tl in laboratory air were 1.25 ng/m^3 and 4.3 pg/m^3, respectively. The same values in clean-room air were about 0.3 –0.6 ng/m^3 and 0.9 – 1.0 pg/m^3. The measured concentrations of the same metals inside a clean bench, in the clean room were 6.5 pg/m^3 and 0.035 pg/m^3.

Detection of atoms and molecules which have transitions in the vacuum UV region is a serious challenge for any spectroscopic technique. One way to solve the problem is excitation of a species, not from the ground state, but from a higher metastable state which can be efficiently populated by electron impact in a gas discharge. The population of a metastable level by electron impact followed by LEAF detection was successfully applied for the detection of Ne and N_2 traces in He, and for detection of NO and NO_2 traces in He, Ar, N_2, and air. LODs of ca. $10^{-7} - 10^{-8}$ vol % were obtained. These LODs are 100 – 1000 times better as compared with any traditional analytical technique.

3.2. Laser-Enhanced Ionization [3]

3.2.1. Physical Principles

Laser-enhanced ionization (LEI) spectroscopy is based on the difference in collisional ionization rates of atoms in their ground and excited states. If a d.c. voltage is applied across the atom reservoir and analyte atoms are optically excited by laser radiation, the enhanced probability of collisional ionization can be detected as a change of current (or a change of voltage across the load resistor). The LEI signal is linearly proportional to the ionization rate of the excited analyte atom, which in turn depends on its ionization potential, the nature of its collision partners, and the temperature of the analytical zone. The maximum

Table 1. Best detection limits for LEAF-GETA [7]

Element	λ_{11}, nm	$\lambda_{12}*$, nm	λ_{fl}, nm	Limit of detection	
				fg	pg/mL**
Ag	328.1		338.3	8.0	0.4
Al	308.2		394	100	5.0
Au	242.8		312.3	10	0.5
Bi	306.8		472.2	800	40
Cd	228.8	643.8	361	18	0.9
	228.8		228.8	3.5	0.175
Co	308.4		345.4	4.0	0.2
Cu	324.7		510.5	60	3.0
Eu	287.9		536.1	3.6×10^5	1.8×10^4
Fe	296.7		373.5	70	3.5
Ga	403.3	641.4	250.0	1.0	0.05
In	410		451	20	1.0
Ir	295.1		322.1	6×10^3	300
Li	670.8		670.8	10	0.5
Mn	279.5		279.5	90	4.5
Mo	313.3		317.0	10^5	5×10^3
Na	589.6		589.6	30	1.5
Ni	322.2		361.9	10^3	50
Pb	283.3		405.7	0.2	0.01
Pt	265.9		270.2	10^3	50
Rb	780		780	20	1.0
Sb	212.7		259.8	10	0.5
Sn	286.3		317.5	30	1.5
Te	214.3		238	20	1.0
Tl	276.8		353	0.1	0.005
V	264.8		354.4	2.2×10^6	1.1×10^5
Yb	398.8	666.7	246.4	220	11

* λ_{12} is the wavelength of the second laser in a two-step excitation scheme.
** The pg/mL values are calculated assuming a sample volume of 20 µL.

sensitivity of LEI occurs when both the probabilities of ionization and ion collection are maximal (close to unity). The ionization efficiency depends exponentially on the energy difference between the upper excited state and ionization potential of the analyte atom. The main advantage of LEI as compared with LEAF is the very high probability of ion collection (close to unity in the optimal scheme); the collection efficiency of spontaneously emitted fluorescence photons is much lower.

Different excitation schemes for LEI are shown in Figure 3. The simplest and most widely used is the one-step excitation scheme (Fig. 3A). Elements with low ionization potential (mainly alkali metals) can be efficiently ionized by the one-step scheme with dye lasers in the visible range. For a number of elements (Cd, Mg, Pb, etc.) the one-step scheme can be efficient with excitation in the UV (harmonics of visible dye lasers). Excitation by UV photons below 300 nm may increase the background caused by nonselective one- or two-step photoionization of the thermally excited species in the analytical zone. The two-step excitation scheme (Fig. 3B) is more complicated – thus more expensive – as it utilizes two tunable lasers, but it has significant advantages. Ionization efficiency increases up to three orders of magnitude compared with one-step excitation. The background does not change significantly if both lasers operate in the visible range. Selectivity increases dramatically as the selectivities of the single steps are multiplied in the two-step process. LEI is an efficient technique only if collision rates are high. Thus, the excitation – ionization scheme (Fig. 3C), with noncoinciding levels of the first and second steps, might be efficient owing to effective collisional coupling of both levels.

Table 2. Real sample analysis by LEAF

Analyte	Matrix [a]	Certified value	Measured value	LEAF LOD (aqueous standard)	Comments
Co	ASRMs	ng/g	ng/g	fg	digested, diluted samples
	wheat grains (SBMP)	60 ± 20	45 ± 2	60	
	dried potato (SBMK)	100 ± 30	90 ± 5		
	dried grass (SBMT)	60 ± 20	48 ± 2		
	Sn alloys			pg	vacuum atomizer
	OSCh		4.4 ± 1.5	4	
	OVCh		2.5 ± 0.8		
Mn	ASRMs	µg/g	µg/g	fg	slurried samples
	citrus leaves (SRM 1572)	23.0 ± 2	23.4 ± 0.7	100	
	bovine liver (SRM 1577 a)	9.9 ± 0.8	9.4 ± 0.5		
	milk powder (SRM 1549)	0.26 ± 0.06	0.29 ± 0.01		
Pb	Ni alloys	µg/g	µg/g	fg	direct solid sampling
	SRM 897	11.7 ± 0.8	12.6 ± 1.2	1	
	SRM 898	2.5 ± 0.6	2.2 ± 0.4		
	SRM 899	3.9 ± 0.1	3.7 ± 0.7		
	ASRMs				slurried samples
	citrus leaves (SRM 1572)	13.3 ± 2.4	14.1 ± 0.5		
Tl	ASRMs	ng/g [b]	ng/g	fg	dissolved samples
	pine needles (SRM 1575)	50	48 ± 2	3	
	tomato leaves (SRM 1573)	50	51 ± 4		
	bovine liver (SRM 1577 a)	3 ± 0.3	3.2 ± 0.2		
	Ni alloys	µg/g	µg/g	fg	direct solid sampling
	SRM 897	0.51 ± 0.03	0.56 ± 0.07	10	
	SRM 898	2.75 ± 0.02	2.5 ± 0.5		
	SRM 899	0.252 ± 0.003	0.24 ± 0.03		
Te	Ni alloys	µg/g	µg/g	fg	direct solid sampling
	SRM 897	1.05 ± 0.07	0.97 ± 0.15	20	
	SRM 898	0.54 ± 0.02	0.52 ± 0.05		
	SRM 899	5.9 ± 0.6	5.4 ± 0.7		
Sb	Ni alloys	µg/g [c]	µg/g	fg	dissolved samples
	PW 1 A	5	4.8 ± 0.5	10	
	PW 2 A	9	8.3 ± 0.5		
	PW 3 A	17	16.5 ± 0.8		
P	Ni alloys	µg/g	µg/g	pg	digested, diluted samples
	SRM 349	30 ± 10	23 ± 3	8	
	SRM 865	120 ± 30	100 ± 20		
	SRM 126 c	40 [c]	37 ± 5		
	ASRMs	mg/g	mg/g		
	SRM 1577 a	11.1 ± 0.4	11.6 ± 0.9		
	SRM 1549	10.6 ± 0.2	10.9 ± 0.9		
	SRM 1575	1.2 ± 0.2	1.1 ± 0.2		

[a] ASRM = agricultural standard reference materials (Russian State Reference Materials). [b] Noncertified value. [c] Reference value.

3.2.2. Atomizers

The most commonly used atomizers for LEI are atmospheric flames. Flames have relatively high temperatures in the analytical zone and, hence, high efficiency of analyte atomization. Thermally excited molecular compounds in flames provide efficient ionization of the optically excited analyte atoms. All types of flames used in atomic absorption and emission spectrometry have been investigated in LEI. Relatively hot hydrocarbon flames such as C_2H_2/air or C_2H_2/N_2O are preferable for LEI because of the high degree of atomization of the elements. A large number of elements can be detected by LEI with C_2H_2/air flames; the hotter C_2H_2/N_2O flame is more suitable for the detection of refractory elements.

Few numerical values can be given for LEI in flames. In a flame at 2500 K, the rate of ionization of an element is enhanced ca. 100-fold for each electron volt (1.6×10^{-19} J) of laser excitation energy. If the energy gap between the excited level and the ionization continuum is less

than 1 eV, the probability of ionization within 10 ns is close to unity. If cw laser excitation is used 100 % ionization can be achieved for an energy gap of ca. 3 eV (and less).

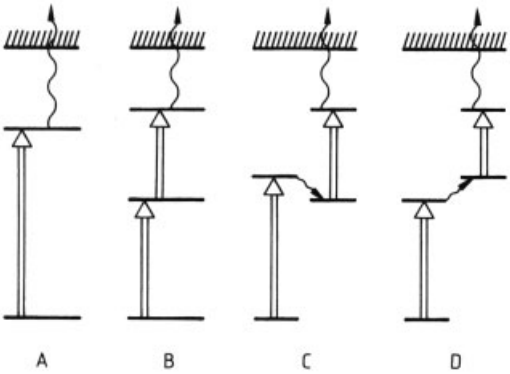

Figure 3. Classification of LEI schemes
A) One-step excitation; B) Two-step excitation; C, D) Collisionally assisted two-step excitation

Other atomizers such as graphite furnaces, ICP, and GD have also been investigated for LEI. Argon buffer gas, typically used in graphite furnaces, is a much poorer collisional partner than the molecular constituents of flames, so collisional ionization is less efficient. On the other hand, good analytical results have been obtained with a combined flame – rod atomizer. A small volume of sample (positioned in a hole in a graphite rod) combined with the high temperature of the analytical zone (located above the rod in the surrounding flame) provided a low absolute detection limit in the picogram – femtogram range. ICP and GD proved to be electrically noisy for LEI.

Table 3. Best detection limits for one- and two-step flame LEI of aqueous solutions

Element	λ_1, nm	λ_2, nm	Flame∗	Detection limit, ng/mL
Ag	328.1		AA	1
	328.1	421.1	AA	0.05
Al	309.3		AN	0.2
As	278.0		AA	3000
Au	267.7		AA	1.2
	242.8	479.3	AA	1
Ba	307.2		AA	0.2
Bi	227.7		AA	0.2
Ca	227.6		AA	0.006
	422.6	585.7	HA	0.03
Cd	228.8		AA	0.2
	228.8	466.2	AA	0.1

Table 3. (Continued)

Element	λ_1, nm	λ_2, nm	Flame∗	Detection limit, ng/mL
Co	240.8		AA	0.06
	252.1	591.7	AA	0.08
Cr	240.9		AA	0.2
	298.6	483.6	AA	0.3
Cs	455.5		AA	0.002
Cu	324.8		AA	3
	324.8	453.1	AA	0.07
Fe	302.1		AA	0.08
	271.9	468.1	AA	0.1
Ga	294.4		AA	0.01
In	271.0		AA	0.001
	451.1	571.0	AA	0.0009
K	766.5		AA	0.1
Li	670.8		AA	0.001
	670.8	610.4	AA	0.0002
Mg	285.2		AA	0.003
	285.2	470.3	AA	0.001
Mn	279.5		AA	0.04
	279.5	521.5	AA	0.02
Mo	319.4		AN	10
Na	285.3		AA	0.001
	589.0	568.8	AA	0.0006
Ni	229.0		AA	0.02
	282.1	501.4	AA	0.04
Pb	283.3		AA	0.2
	283.3	600.2	AA	0.09
Rb	780.0		HA	0.09
Sb	287.8		AA	50
Si	288.2		AN	40
Sn	284.0		AN	0.4
	284.0	597.0	HA	0.3
Sr	230.7		AA	0.003
	460.7	554.3	AA	0.2
Ti	320.0		AN	1
Tl	276.8		AA	0.006
	291.8	377.6	AA	0.008
V	318.5		AN	0.9
W	283.1		AA	300
Yb	267.2		AA	1.7
	555.6	581.2	AA	0.1
Zn	213.9	396.6	AA	1

∗ AA = acetylene – air; AN = acetylene – nitrous oxide; HA = hydrogen – air.

3.2.3. Limits of Detection

The best LODs for one- and two-step LEI in flames are listed in Table 3.

3.2.4. Real Sample Analysis

The potential of LEI has been investigated by analysis of Standard Reference Materials (SRMs) at the National Institute of Standards and Technology (NIST). Traces of Co, Cu, Cd,

Mn, Ni, and Pb were measured in different SRMs of water; Mn and Ni traces in bovine serum; Mn, Ni, and Fe traces in apple and peach leaves; Tl and Ni traces in river sediments; and Na, K, Ca, and Mg traces in rainwater.

The concentrations of the alkali metals Cs, Li, and Rb were detected by LEI in different types of reference rock samples. Rock samples were dissolved by standard methods and analyses performed without preconcentration, with pure aqueous standards for calibration. Concentrations were determined by both calibration curve and standard additions methods. The results obtained by the two techniques coincided within experimental errors. Concentrations varied in the range 0.4 – 17.0 ppm and were in a good agreement with the certified values for the reference samples.

The LEI technique has been applied to the analysis of pure materials for the semiconductor industry. The concentrations of Cr, Fe, and Ni in Cr-doped GaAs disks were measured with the simplest pretreatment (dissolution in high-purity nitric acid and dilution with pure water). After dilution, a concentration of 0.1 ng/g in solution corresponded to 1 µg/g in the solid. Impurity concentrations were 4, 8, and 2.3 µg/g, respectively.

The rod – flame atomizer avoids the general disadvantage of flame-based analytical techniques – the necessity to dissolve the solid sample. The potential of two-step LEI with the rod – flame atomizer was investigated by analysis of Cu traces in Ge and In traces in CdHgTe. The LODs as low as 1 pg/g for In and 0.5 ng/g for Cu were achieved in direct analysis of 10-mg solid samples. Measured values in real samples varied in the range 30 – 70 ng/g for Cu and 2.6 – 3.4 ng/g for In.

Laser-enhanced ionization is a powerful, versatile and relatively simple technique which provides LODs in the ng/g – pg/g range with large linear dynamic range. A serious drawback of the technique is its susceptibility to interference: Thermal ionization of easily ionized matrix components (mainly alkali metals); photoionization of the thermally excited major components (especially if UV laser light is used for excitation); thermal ionization from levels excited by the spectral wing of an intense laser line; and ionization of the molecular components of the flame (especially NO). Interference may be greatly reduced or avoided by the appropriate choice of excitation wavelength, by the two-step excitation scheme, by specially constructed collecting electrodes; by optimal location of the analytical zone within the flame. To reduce interference one should investigate precisely the origins of noise and background for a particular analyte and matrix.

3.3. Resonance Ionization Spectroscopy [2]

3.3.1. Physical Principles

Resonance ionization spectroscopy (RIS) is based on resonance multistep excitation of high-lying levels of free analyte atoms and their subsequent ionization. Ionization may be caused by the photons of the final laser step (photoionization) or by an external electric field (field ionization). Sometimes the latter technique is referred to as field ionization laser spectroscopy (FILS). Multistep excitation of atoms may be effective if the rates of stimulated transitions to each intermediate level are much higher than the total decay rates (radiative and collisional) of any intermediate level. In contrast to LEI, RIS is efficient only if negligible collisions occur in the analytical zone. As in the case of LEI, the high sensitivity of RIS is based on the high efficiency of ion collection (close to unity). Thus, the overall efficiency of neutral atom detection by RIS depends on the efficiency of stepwise excitation and final ionization. Maximum sensitivity is achieved if all the intermediate transitions are saturated and the ionization efficiency is close to unity. In optimal conditions, a single free atom may be detected with a probability close to unity.

The three most widely used schemes for RIS are shown in Figure 4. They differ from each other in the final ionization step. Nonresonant photoionization to a continuum (Fig. 4A) can be efficient if the energy flux of the ionizing laser is high enough (0.01 – 1 J/cm^2), as the nonresonant ionization cross sections are relatively low (10^{-17} – 10^{-19} cm^2). For comparison, typical cross-sections of resonance optical transitions are 10^5 – 10^6 times higher.

The photoionization scheme via the so-called autoionization state (AIS) is shown in Figure 4B. The AISs are discrete states of the inner excited

electrons with energies above the ionization potential of the atom. On the one hand, the cross section of a transition to AIS may be several orders of magnitude higher than direct photoionization to the continuum (Fig. 4A); on the other hand, the AIS may decay to the continuum (i.e., the atom is ionized) with a typical lifetime of a few nanoseconds. Thus, in the field of an intense laser pulse with a duration of ca. 10^{-8} s, the probability of excited AIS depletion and ionization may be close to unity. Ionization via AIS is highly efficient for some elements (Gd, Lu).

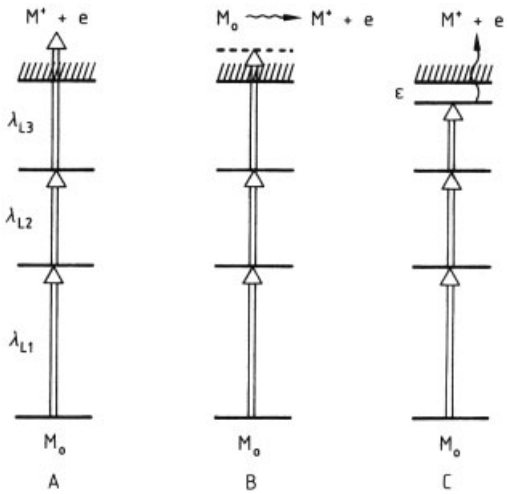

Figure 4. Classification of RIS schemes
A) Nonresonance photoionization to a continuum; B) Photoionization via an autoionization state; C) Electric field ionization of a Rydberg state

Ionization through high-lying Rydberg states (Fig. 4C) has been found to be very effective. Rydberg states close to the ionization limit can be efficiently ionized by an electric field pulse of moderate amplitude. The dependence of ionization yield of a particular Rydberg state on the electric field amplitude has a characteristic Z-profile with a pronounced threshold E_{crit}. For electric fields above E_{crit} the ionization yield is nearly constant and close to unity. Both the excitation cross section and E_{crit} depend strongly on the principal quantum number of the Rydberg state ($n*$). For higher $n*$, E_{crit} is lower (an advantage), but the excitation cross section is much lower (a disadvantage). Therefore, compromise values of $n*$ are used, which for most elements lie between 15 and 30. For such $n*$, excitation cross sections are of the order 10^{-14} – $10^{-15}\,\mathrm{cm}^2$ and E_{crit} values are 1 – 10 kV/cm. For analytical applications, ionization via Rydberg states has proved to be an ideal scheme for RIS. Because of the relatively high excitation cross sections, maximum sensitivity can be realized with tunable lasers of moderate pulse duration and average power, and with reasonable electric fields. The effectiveness of AIS is limited by the absence of AIS near the ionization potential of many elements and the lack of appropriate spectroscopic information.

The fact that RIS measures charged particles is an important advantage. Highly efficient signal ion collection is possible by appropriate electromagnetic focusing. Second, discrimination of signal ions from background ions can be achieved by mass separation methods. These two factors improve both the sensitivity of RIS and its accuracy.

In most analytical RIS applications two- or three-step excitation is used. This provides extremely high selectivity (up to 10^{10} – 10^{12}) as the overall selectivity of a multistep process is a product of one-step selectivities.

3.3.2. Atomizers

Vacuum or low-pressure atomizers are used in RIS to obtain collisionless interaction of the free analyte atoms with resonance laser radiation. Different types of GD and vacuum electrothermal atomizers (VETA) are most commonly used in RIS. In some specific analytical applications (detection of traces of rare radionuclides) RIS is used on-line with fast atomic beams.

In many applications of RIS – VETA the generated ions are detected by a secondary-emission multiplier incorporated in the atomizer. The vacuum should be better than 10^{-3} Pa throughout the analytical volume and the path of free ions from the interaction zone to the multiplier. This limits the speed of sample evaporation and dictates the efficiency of the pumping system. Both unified vacuum systems and systems with differential pumping of the analytical volume and free flight zones are used. In these respects the strategy of electrothermal atomization of a sample in RIS differs significantly from the traditional one used in graphite furnace AAS. In the latter technique fast atomization (1 – 10 s) gives the best

analytical results. In RIS, very slow atomization (up to several minutes) is typical.

In some applications, a proportional counter operating near atmospheric pressure is used. In this case, the electrons created in the process of atomic resonance ionization are detected. By appropriate choice of gas mixture and accelerating voltage, an electron cascade with a magnification factor of about 10^4 can be achieved. This gain is high enough for single photoelectron detection.

For maximum sensitivity and isotope selectivity RIS is often combined with time-of-flight (RIS-TOF) or quadrupole mass spectrometry (RIMS).

3.3.3. Limits of Detection: Real Samples

The high sensitivity of RIS has been demonstrated in the analysis of pure semiconductor materials, biological samples, seawater, and environmental samples. In some cases, direct analysis was possible; for some matrices, sample pretreatment was used.

Traces of Na, Al, and B in high purity Ge were detected by RIS with direct electrothermal atomization of solid Ge chips in vacuum. For all these elements, two-step excitation of Rydberg states and electric field ionization were used. The measured concentrations of Na, Al, and B in real samples were 0.2, 1.0, and 0.2 ng/g while the LODs were 0.05, 0.01, and 0.05 ng/g, respectively. In the case of B, the strong ion background seriously limited the sensitivity and selectivity of traditional RIS. This background was caused by efficient two-step nonresonance photoionization of Ge atoms (the energy difference between the excited $2s^2 3s$ level of B and the $4s^2 4p5s$ level of Ge is only $20\,\mathrm{cm}^{-1}$). Both analytical characteristics were greatly improved by a special technique, with two sequential electrical pulses of different duration, polarity, and voltage.

The distribution of ruthenium and rhodium traces in the ocean has been investigated by RIS–VETA. Both elements are rare and their distribution is of great geological and geochemical importance. The abundance of Ru and Rh in the Earth's crust is estimated to be 1 – 100 pg/g and can not be directly measured by current analytical techniques. The fire assay method and concentration by polymer sorbents were used prior to RIS analysis. Concentrations of Ru varied from 1.3 pg/mL in ocean water to 7.6 ng/g in Red Sea sediments. Concentrations of Rh in ocean water were 3 – 10 pg/mL.

The detection of ^{81}Kr, produced by cosmic rays, is of great importance for dating samples from the Earth's crust, groundwater, and polar ice cores. Measurement of ^{81}Kr in groundwater is a unique analytical problem as its concentration is ca. 1400 atoms per liter. The problem is further complicated by isotopic interference of stable krypton isotopes (^{82}Kr/^{81}Kr is about 10^{12}). Prior to final analysis by RIS – TOF [8] several consecutive steps of field sampling and isotope concentration were necessary. All steps required extreme care to avoid sample contamination. Water samples of 50 – 100 L were necessary to detect ^{81}Kr above the background. This value is more than 100 times less than for the routine technique of ^{39}Ar dating.

The environmentally important radionuclides ^{90}Sr and ^{210}Pb have been detected by RIMS – VETA. Two-step excitation of Rydberg states by two cw dye lasers, followed by nonresonant ionization by a cw infrared CO_2 laser, was used. The solid samples were dissolved by standard methods. The LODs were 60 ag (6×10^{-17} g) for ^{210}Pb and 30 ag for ^{90}Sr. Absolute ^{210}Pb concentrations of 0.12 – 0.14 fg were measured in brain tissue. Isotope selectivity of 10^9 for ^{90}Sr/^{89}Sr was achieved.

Diode lasers were successfully used in a multiphoton RIMS scheme for isotopically selective detection of strontium. The two-step resonance excitation of ^{90}Sr was performed with two single-mode cw DLs ($\lambda_1 = 689.45$ nm, $\lambda_2 = 688.02$ nm). The excited ^{90}Sr atoms were photoionized by a powerful Ar^+ laser ($\lambda = 488.0$ nm). The ions were detected in a commercial quadrupole MS. Using graphite crucible atomization the absolute LOD of 0.8 fg and overall (optical + mass spectrometer) isotopic selectivity of $> 10^{10}$ in the presence of stable ^{88}Sr was achieved for ^{90}Sr. No other non-radiochemical technique can realize such sensitivity and selectivity of ^{90}Sr detection.

A similar scheme was used for the spectroscopic studies on hyperfine components and isotope shifts of rare stable isotopes of calcium and the radionuclide ^{41}Ca. Calcium atoms in an atomic beam were excited with single-frequency

cw dye and titanium sapphire lasers and then photoionized with the 363.8 nm or 514.5 nm line of an Ar^+ laser.

The use of RIMS in chemical vapor deposition diagnostics has been investigated.

3.4. Diode Lasers in Elemental Analysis [6], [8]

3.4.1. Physical Principles of Absorption Spectrometry

Quantitative absorption measurements are based on the Lambert – Beer law:

$$P_t(\lambda) = P_0 \exp[-k(\lambda) l]$$

where P_t is the laser power at wavelength λ, transmitted through a layer with absorption coefficient $k(\lambda)$ and absorption pathlength l. The absorbed power ΔP is $P_0 - P_t = P_0[1 - \exp(-k(\lambda) l)]$.

For small values of the absorption $k(\lambda)l \ll 1$, this is approximated with $\Delta P(\lambda) \approx P_0 k(\lambda) l$.

Theoretically, the minimum detectable absorption (MDA) is determined by the shot noise of detected radiation. In single pass measurements ($l = L$, where L is a sample absorption pathlength)

$$(kL)_{\min} = \Delta P_{\min}/P_0 = (2Bhv/\eta P_0)^{1/2}$$

where η is the quantum efficiency of the photodetector, and B is a detection bandwidth. To obtain noise equivalent relative power variations, $\Delta P_{\min}/P_0$, smaller than 2×10^{-8} with a measurement time of 1 s, the laser photon flux should be at least 5×10^{15} photons/s. This corresponds to laser powers of 1 mW at 1 µm and 100 μW at 10 µm. In practice there are additional noise factors affecting the detection sensitivity, such as intensity fluctuations of the laser (laser excess noise) and technical fluctuations. Thus, for typical DLs, the value of $\Delta P_{\min}/P_0 = 1 \times 10^{-8}$ can be considered as the lowest, fundamentally limiting one.

Usually, absorption measurements are not considered to be very sensitive since the signal $P_0 k(\lambda) l$ is a subtle variation of a large baseline P_0. In order to have zero baseline, sample or wavelength/frequency modulation (WM/FM) techniques can be used. In the case when λ is modulated with frequency f, a signal at the n-th harmonics of f is simply proportional to $P_0 k_{nf}(\lambda) l$, where k_{nf} is the amplitude of the corresponding harmonics of k. As $k(\lambda) = \sigma(\lambda) N$,

where $\sigma(\lambda)$ is the absorption cross section, $\sigma(\lambda) = \sigma_0(\lambda) / (1 + I/I_{\text{sat}})$

(see Section 1.3), then the signal after the phase-sensitive detector is directly proportional to the concentration of absorbing particles N. Another important advantage of WM/FM is the detection of a signal at high frequency (> 10 kHz), where the low-frequency flicker noise is greatly reduced.

Besides reducing noise, improvements in the MDA can be achieved by enhancing the signal. This can be done by increasing the absorption pathlength using multipass cells. By an appropriate mirror arrangement a DL beam can be directed many times through the absorbing layer, increasing the overall absorption with an appropriate decrease in MDA. The number of passes is limited by the quality of the mirrors, which determines the cell transmission, or by interference effects.

3.4.2. Atomic Absorption Spectrometry with Diode Lasers [6], [8]

UV/visible/NIR semiconductor diode lasers (DLs) as sources of tunable narrowband resonance radiation have been discussed in Section 2.2. The practical and technical advantages of visible/NIR DLs are ease of operation, small size, room-temperature operation, and low price. In comparison with hollow cathode lamps (HCLs), they have such advantages as high spectral power, spatial coherence, and tunability. Owing to these features, detection limits, dynamic range, and selectivity of analysis are all significantly improved.

The presently available powers of DLs are several orders of magnitude higher than the powers of the best commercial HCLs. In contrast to an HCL, a DL emits a prominent single narrow line. This feature dramatically simplifies the construction of an analytical device because one does not need a monochromator, which is used in commercial AAS instruments for isolation of the analytical line from the spectral lines of the

HCL buffer gas and unwanted lines of the cathode material. In the case when optimal experimental conditions are realized, more than 1 mW of laser radiation power can be obtained at the photodetector. It has been shown that a near fundamental shot-noise limit can be achieved and relative power variations $\Delta P_{\min}/P_0$, which can be measured with DLs, can be as low as 10^{-8}.

The LODs of diode laser atomic absorption spectrometry (DLAAS) have been greatly improved using a WM technique. In conventional atomic absorption spectrometry with HCLs, low-frequency (10 – 60 Hz) amplitude modulation (mainly by mechanical choppers) is used. In contrast, the wavelength of the DL can be easily modulated with much higher frequencies, in DLAAS usually up to 10 – 100 kHz. By recording the analytical signal at the harmonics of the modulation frequency, usually the second one ($2f$-mode), one can dramatically suppress the $1/f$ noise, thus decreasing the LOD. Even in the first simple measurements with a DL power of ca. 100 μW, a modulation frequency of 5 kHz, and with $2f$-mode of recording, absolute LODs in the femtogram range have been achieved experimentally. These LODs correspond to measured values of $\Delta P_{\min}/P_0 \approx 10^{-4} - 10^{-5}$ (compared with $10^{-2} - 10^{-3}$ for conventional AAS).

Further improvement of WM-DLAAS has been demonstrated for the case when both the radiation source and the absorption are modulated. Such a situation was realized with a microwave induced plasma (MIP) or a direct current plasma (DCP) as the atom reservoir. The double-beam double-modulation approach with logarithmic signal processing (LSP) and $2f$-mode recording permitted a complete elimination of the background, caused by non-selective absorption, low-frequency fluctuations, and drift of the baseline and laser intensity. A shot-noise limited MDA as low as 3×10^{-8} for a 1 s time constant was experimentally realized. The linear dynamic range can be also substantially extended with this technique, in the best cases up to six – seven orders of magnitude as compared with less than two orders of magnitude for conventional AAS.

Applications of DLs in analytical atomic spectrometry are collected in the following list (numbers in parentheses are analytical wavelengths in nanometers):

- Diode laser atomic absorption spectrometry (DLAAS):
  - Graphite furnace
    - Al (396.15), Ba (791.13), Cs (894.35), K (769.90), La (657.85), Li (670.78), Rb (780.03), Sr (689.26)
  - Flame
    - Cs (425.44, 852.11), Cr (425.44), K (766.49), Li (670.78), Rb (780.03), Ti (399.97, 842.65)
  - Discharge (direct current plasma (DCP) or microwave induced plasma (MIP))
    - Cl (837.60), F (739.87), O (777.18)
- Laser induced fluorescence (LIF):
  - Li (670.78), Rb (780.03)
- Isotope selective detection:
  - Li, Pb, wavelength modulation diode laser atomic absorption spectrometry (WMD-LAAS)
  - $^{235}U/^{238}U$ in glow discharge, optogalvanic detection
  - $^{235}U/^{238}U$ in solid samples, laser ablation + LIF with DL excitation
- Physical vapor deposition process control:
  - Y (668), external cavity diode laser (ECDL), high frequency modulation spectroscopy
  - Al (394), second harmonic generation (SHG) of ECDL
- Control of semiconductor films growth:
  - Al (394), Ga (403), In (410), SHG in bulk crystals
- Specific detection of haloform components:
  - Cl (837.60), Br (827.24), MIP or DCP, GC-WMDLAAS, double-modulation
- Environmental analysis:
  - Cr^{III}/Cr^{VI} speciation in deionized and drinking water
    - Cr (427.48), SHG in $LiIo_3$, flame, HPLC-WMDLAAS
  - Cr^{VI} in tap water
    - Cr (427.48), SHG in $KNbO_3$, flame, HPLC-WMDLAAS, double-beam, LSP
  - Chlorinated hydrocarbons in oil from plastic material recycling
    - Cl (837.60), MIP, GC-WMDLAAS
  - Chlorophenols in plant extracts
    - Cl (837.60), MIP, GC-WMDLAAS, double-beam, logarithmic signal processing (LSP)

Speciation of methylcyclopentadienyl manganese carbonyl (MMT) in gasoline, human urine and tap water
Mn (403.1), flame, SHG in $LiIO_3$, HPLC-WMDLAAS, 4*f*, LSP
Analysis of chlorine in polymers
Cl (837.60), MIP, laser ablation + WMDLAAS, double-modulation, LSP

Resonance ionization mass spectrometry (RIMS):
La (657.85), three-step RIMS, DL – first step
Sr (689.45, 688.02), double-resonance excitation by two DLs

The detection capabilities for various species differ considerably depending on the type of atomizer, oscillator strength of a transition, detection technique used. Multielement analysis with DLs has also been investigated. Several laser beams may be collimated and directed simultaneously through an absorbing layer in a conventional atomizer, either by a system of mirrors, or by optical fibers. The monochromator – photomultiplier system can be replaced by a simple semiconductor photodiode. Appropriate modulation of a wavelength of a particular DL, followed by Fourier transform or lock-in amplification analysis of the photodiode signal, makes it possible to process analytical signals of individual analytes. A six-element analysis with DLs and a commercial graphite tube atomizer has been demonstrated.

3.4.3. Element Specific Detection by DLAAS

Diode laser based analytical devices can be used as element selective detectors for gas/liquid chromatography (GC/LC). The atomic constituents of different molecular species in a sample can be measured in an atomizer, such as a flame or low pressure plasmas, e.g., an MIP or DCP. The GC-MIP/DCP in combination with DLAAS enables the detection of non-metals, such as H, O, S, noble gases, and halogens. Most of these elements have metastable levels, which are efficiently populated in a plasma, and which have strong absorption transitions in the red – NIR spectral region.

A trend in elemental analysis is from detection of the total content of a heavy metal in a sample to the detection of its different chemical compounds or oxidation states (speciation). This is because the essentiality or toxicity of the metals should not be attributed to an element itself but to element compounds with distinct biological, physical, and chemical properties. A compact and relatively cheap DL based instrument could be a reasonable alternative to the powerful but expensive ICP-MS, presently used for speciation. The potential of DLAAS as an element selective detector for speciation has been demonstrated by its coupling with high performance liquid chromatography (HPLC). This system has been used for speciation of chromium (III)/(VI) and organo-manganese compounds. An LOD of 30 pg/mL for Cr^{III} and Cr^{VI} was obtained by use of 3 μW power frequency-doubled laser light, a double-beam configuration, and logarithmic processing of the signals. In the case of chromium, the LOD of HPLC-DLAAS is several times lower, than the LOD for HPLC-ICP-MS.

3.4.4. Isotope Selective Detection with Diode Lasers

The narrow spectral line of a DL enables isotope selective analysis. For light and heavy elements (such as Li and U) the isotope shifts in spectral lines are often larger than the Doppler widths of the lines, in this case isotopically selective measurements are possible using simple Doppler-limited spectroscopy – DLAAS or laser induced fluorescence (LIF). For example, ^{235}U and ^{238}U ratios have been measured by Doppler-limited optogalvanic spectroscopy in a hollow cathode discharge. DLAAS and LIF techniques have been combined with laser ablation for the selective detection of uranium isotopes in solid samples. This approach can be fruitful for development of a compact analytical instrument for rapid monitoring of nuclear wastes.

3.4.5. Commercial System

The DLAAS devices are now commercially available from Atomica Instruments (Munich, Germany). The first system was prepared for the measurement of ultra-trace levels of Al in the semiconductor industry. The commercial laser diode module includes the DL with a heat sink for temperature tuning, the microoptics, and the

nonlinear crystal for SHG. The module has the size of a HCL. Since delivery, the module has successfully worked routinely under the conditions of an industrial analytical laboratory without any repair or maintenance. The semiconductor industry in particular is interested in the measurements of light elements, such as Al, K, Ca, Na, in ultra pure water or chemicals. For example, the module designed for the detection of K provides an LOD of 0.5 pg/mL (10 µL aliquot). Similar compact and sensitive devices with electrothermal atomizers, flames or micro-plasma are available for a number of elements, such as Mn, Ni, Pt, Pd, Ca, Cu, Fe, Ga, the alkali elements, the halogens, C, O, and H. In forthcoming years, one can also expect the commercialization of specific element selective DLAAS detectors for GC or LC.

4. Laser Techniques for Molecular Analysis [6], [8]

4.1. Molecular Absorption Spectroscopy with Diode Lasers

4.1.1. Detection of Molecular Gas Impurities with Lead-Salt Diode Lasers [2], [9]

The Pb-salt-based DLs (Section 2.2) efficiently generate tunable laser radiation in the region 3 – 30 µm, where the great majority of molecules have intense vibrational rotational absorption lines. As the Doppler-broadened absorption linewidths of molecules are of the order of (1 – 3) $\times 10^{-3}$ cm^{-1}, DLs provide the best selectivity for linear absorption techniques.

Both cw and pulsed (sweep) modes of DL generation have been widely tested in analytical applications. Many well-developed modulation techniques have been used with consequent improvement of analytical sensitivity. The best results were achieved with a high FM technique ($f \sim 100 - 1000$ MHz). The minimal shot-noise-limited relative power variations $\Delta P_{\min}/P_0$ which can be experimentally detected with FM DLs are about 10^{-7}. Typical values of $\Delta P_{\min}/P_0 \approx 10^{-5}$ (1 Hz detection bandwidth) are mainly determined by instrumental factors (in most cases by interference effects). For this minimal detectable value of $\Delta P_{\min}/P_0$ and for molecular species at atmospheric pressure the corresponding concentration LODs are ca. 0.1 – 10 ppb m (0.1 – 10 parts per billion by volume over an optical path of 1 m). Evidently, if 10 m absorption length could be realized, the minimal detectable concentration of absorbing molecules would be 10 times lower.

The most important analytical applications of mulecular absorption spectroscopy with mid-IR DLs are environmental monitoring and atmospheric studies [9]. Two different approaches are used: (1) measurements in sample cells at reduced pressure, when the absorption linewidths are of ca. (1 – 3) $\times 10^{-3}$ cm^{-1}, and (2) open-path measurements at atmospheric pressure, when the pressure broadening leads to ca. 10^{-1} cm^{-1} linewidths.

In a cell, optimal gas pressure and appropriate dilution of primer gas by a buffer can be used. A number of mulecules have been measured with lead-salt DLs: CO, NO, NO_2, N_2O, HNO_3, NH_3, SO_2, O_3, HCl, H_2CO, CH_4, C_2H_4, C_2H_6, CF_2Cl_2, etc. The effective absorption pathlength of 20 – 300 m was realized with multipass cells with an appropriate decrease of the LOD down to a record value of 10 ppt (5 min integration time). Several Pb-salt DL-based gas monitors have been used to perform *in situ* trace gas measurements in the troposphere and lower stratosphere from airborne or balloon-borne platforms. The systems were designed to measure CH_4, CO, N_2O, NO_2, N_2O, HCl and H_2O with LODs at sub-ppb concentration levels (3 – 30 s integration time). A typical system includes a liquid-N_2 Dewar for diode lasers and detectors, a multipass cell, and can simultaneously measure up to four gases.

Owing to pressure broadening, ground-based remote detection of trace gases with DLs was not as effective as measurements in sample cells. Long-path atmospheric gas analyzers have been exploited mainly to monitor carbon monoxide concentrations. Mirror retroreflectors have been used to reflect the DL probing beam back to the detector providing a round-trip absorption pathlength of up to 1 km. Sensitivities achieved are in the range from 0.06 to 40 ppm · m.

Besides atmospheric trace measurements, Pb-salt DLs have successfully been used for detection of impurities in the process control of semiconductor materials, analysis of high purity

gases, sensing of toxic gases, explosive detection, isotope analysis of gases, and medical diagnosis of various diseases by analysis of expired air. The use of a DL provides high time resolution (0.1 – 1 μs), facilitating sensitive control of transient processes (e.g., intermediates in chemical reactions, discharges, molecular beams). So far the high price, complexity, and limited reliability of the Pb-salt DL spectrometers limit their applicability to the research laboratories when utmost sensitivity and selectivity have to be realized.

4.1.2. Near IR Gas Sensors

Advances in room-temperature NIR lasers have led to a new generation of gas sensors. A number of DL-based instruments have been developed and are now commercially available.

These sensors have rather simple constructions, may be very compact, have low cost, are reliable, and allow completely autonomous operation for a long time.

Peculiarities of developing NIR-DL sensors are defined by characteristics of molecular absorption in this spectral region. There exists a large quantity of absorption lines that arise from overtones and combination bands. Typical linestrengths of these transistions are two to three orders weaker than those for the fundamental bands in the mid-IR. The strongest absorptions in the 1.3 – 1.6 µm region are provided by the H_2O, HF, HCl, and C_2H_2 molecules. The LODs which are achieved for these molecules in pure gas can be at the ppt level. However, if the absorption of the matrix gas is larger than that of the detected impurity, the LOD will be determined by the interfering absorption of the matrix gas molecules.

There are a number of quite successful applications of NIR-DL sensors for monitoring of gas molecules, especially H_2O. A few types of hygrometers based on a 1.4 µm InGaAsP-laser are commercially available. A very small LOD of 65 ppt for water in air was achieved in laboratory test measurements with a hygrometer, designed to measure moisture in semiconductor feed gas. Another DL sensor, capable of measuring low levels of water vapor in the presence of methane, is designed to be attached to a sampling station where pipeline gas is extracted at near-ambient pressures (1 – 2 atm).

Flight versions of NIR-DL instruments for trace water vapor measurements have been demonstrated on research aircraft with a sensitivity of the order of 10 ppm. Examples of such systems flown on NASA aircraft include both open-path and closed-path configurations designed for fast-response measurements of H_2O from the Earth's surface to an altitude of 20 km.

Similar instruments developed for monitoring O_2, NH_3, HCl, HF, H_2S, CO, CO_2 are suited to industrial stack monitoring, process and environmental control, where high temperatures, continuous vibration, thermal cycling, or high-speed air flow are present. The automated sensors provide fast and accurate ppm-level measurements and are used for optimizing incineration, sewage and waste treatment, power generation, and so on. Fiber optics are often preferred for their simplicity and ability to place the instrument at distances hundreds of meters from the hazardous zone.

4.1.3. Advanced Developments

Rapid improvements of the DL characteristics make these devices very attractive for absorption spectrometry not only of gas species (atomic or molecular) but also of liquid and condensed-phase samples. Much attention is now focused on using NIR-DLs for molecular fluorescence spectrometry, Raman spectroscopy, thermal lens spectrometry, medical diagnostics, development of fiber sensors, and so on.

It is unlikely that the NIR-DLs will permit measurement of all molecular trace constituents in the atmosphere with sensitivities sufficient for environmental control. New, room temperature operated lasers have been tested for trace measurements in mid-IR region. Potentially the most significant technology is quantum-cascade lasers (QCLs, see Section 2.2). Spectroscopic characteristics and noise properties of these structures enable measurements of trace impurities with LODs achieved using cryogenic Pb-salt lasers.

New absorption methods, like intracavity spectroscopy, cavity-ring-down and cavity-enhanced spectroscopy, have demonstrated very high sensitivities in laboratory measurements

with DLs. An ultrasensitive technique that combines external cavity enhancement and FM spectroscopy has been developed. This method, which has been called NICE-OHMS, or noise-immune cavity-enhanced optical heterodyne molecular spectroscopy, is based on frequency modulation of the laser at the cavity free-spectral-range frequency or its multiple. The MDA of 5×10^{-13} ($k_{\min} = 1 \times 10^{-14}\,\mathrm{cm}^{-1}$) in the detection of narrow saturated absorption spectra of C_2H_2 and C_2HD has been reported, with a cavity finesse of 10^5. This result is one of the most sensitive absorption measurement ever reported.

4.2. Laser Optoacoustic Spectroscopy

[1], [2], [10]

The optoacoustic effect, based on partial nonradiative dissipation of the absorbed light energy, followed by generation of an acoustic wave, was discovered about a hundred years ago. The effect is very weak and, before the advent of lasers, it was applied mainly for spectroscopic investigations of gases. Lasers and improvements in acoustic vibration detection have led to the development of laser optoacoustic spectroscopy (LOAS). Several physical processes are involved in LOAS: (1) absorption of laser radiation by one or more compounds in a medium; (2) radiationless conversion of a part of the absorbed energy into local heating of the sample; (3) formation of an acoustic wave due to a temperature gradient; and (4) detection of the acoustic vibrations.

The main advantages of LOAS are high sensitivity, applicability to samples of any state of aggregation, and relatively simple instrumentation. Its drawback is poor selectivity, due to spectral interference of the molecular compounds in a complex mixture. To improve selectivity, LOAS is often combined with a preselecting technique such as gas or liquid chromatography. This discussion is limited to the applications of LOAS for sensitive analysis of simple mixtures.

The LOAS instrument consists of the following main units: A tunable laser, a sample cell with an acoustic transducer (sometimes called a spectrophone or SP), an amplifier, and a recording system. The acoustic response of the medium can be stimulated by amplitude or frequency modulation of the laser light or by Stark or Zeeman modulation of the absorption line of the analyte. Both pulsed and cw lasers with mechanical or electro-optical amplitude modulation are often employed in LOAS analytical applications.

The key component of an LOAS instrument is the spectrophone. Its construction depends critically on the state of aggregation of the sample. The SP for gas samples consists of a gas cell with input and output windows for laser radiation and a microphone, fixed on an inside or outside wall of the cell. As in the case of MAS with SDLs, the sensitivity can be improved by the use of multipass SPs. Plane capacitor or electronic microphones are the most widely used. The sensitive element of these microphones is a thin (1 – 10 μm) elastic membrane made of Mylar, Teflon, or metallic foil. The membrane serves as one electrode of a dielectric capacitor (membranes made of dielectric materials have metallic coatings). Thus, the acoustic vibrations of a gas mixture in the cell can be directly converted to an electrical signal. Commercial microphones with sensitivities of 5 – 50 mV/Pa are utilized in routine applications of LOAS; specially constructed and optimized ones are used for ultrasensitive analyses.

Resonant and nonresonant SPs can be used in LOAS. In the former type, the acoustic resonances are formed inside the cell. Resonant SPs provide higher sensitivity, but they are less flexible, as the acoustic resonances depend on the physical dimensions of the cell, the type of gas mixture, its temperature, and pressure. The modulation frequency of cw laser radiation should be matched to the acoustic resonance frequency (typically of the order of several kilohertz). The dimensions of nonresonant cells are commonly 5 – 10 mm diameter and 5 – 20 cm length, while some resonant cells have a diameter up to 10 cm.

For the analysis of liquid and solid samples, more complicated SPs are used. For liquid samples, the simplest and most sensitive detection of acoustic waves is by direct contact of the microphone and a liquid. The acoustic vibrations of irradiated solid samples can be transduced to a microphone via a buffer gas or a "pure" liquid. Piezoelectric microphones are the most widely used type for solids and liquids.

Some important molecular impurities in pure gases have been successfully detected by LOAS. Typical gas pressures range from 1 – 100 kPa;

sample volumes 0.5 – 3 cm^3; routine accuracy 5 –15 %; linear dynamic range up to 10^5. The most impressive LODs are: 0.1 ppb for SO_2 (multipass, resonant SP); 10 ppb for NO_2 (resonant SP); 1 ppm for HF (nonresonant SP, pulse laser, time resolution); 100 ppb for CH_4 (nonresonant SP); about 100 ppb for CO (resonant SP); 0.1 ppb for NO (nonresonant cell, six microphones); 10 ppb for SF_6 (resonant SP).

The best sensitivity of LOAS for liquid samples (up to $10^{-6}\,cm^{-1}$) was achieved with pulsed lasers of ca. 1 mJ pulse energy. With modulated cw lasers of ca. 1 W average power, a minimum absorption of 10^{-5} – $10^{-6}\,cm^{-1}$ was detected. A few examples of real analyses include the LODs for β- carotene, Se, and Cd in chloroform of 0.08, 15, and 0.02 ng/mL, respectively; the LODs for bacteriochlorophyll, chlorophyll b, and hematoporphyrin in ethanol of 1, 0.3, and 0.3 ng/mL, respectively; the LODs for cytochrome c, vitamin B_{12}, uranium(IV), and uranium(VI) in water of 30 ng/mL, 4 ng/mL, 8×10^{-7} mol/L and 10^{-6} mol/L, respectively. Typically, the LODs for direct analysis of liquids are worse as compared with "pure" gases because of higher background absorption of the solvent. To reduce this background, frequency modulation of lasers and differential two-beam detection schemes are used.

LOAS has been successfully applied for the detection of thin films and layers deposited on a solid substrate. Both direct measurement by a piezoelectric microphone connected to the substrate and detection of acoustic waves via an intermediate gas or liquid were used. For example, the absorption of thin Al_2O_3 layers (up to monolayer) on an aluminum substrate was measured with a cw CO_2 laser, modulated by a mechanical chopper. Measurements of thin layers of liquids and powders as well as aerosols, suspensions, colloids, and small dispersed particles are important applications of LOAS.

4.3. Thermal Lens Spectroscopy [1], [3]

Thermal lens spectroscopy (TLS) is another laser-based technique, associated with dissipation of absorbed laser energy followed by the formation of temperature gradients in a medium. It was observed for the first time in 1964 when a 1- cm Brewster cell containing benzene was placed inside the resonator of a cw He – Ne laser. Contrary to the expected behavior, the laser intensity began to fluctuate and in a few seconds, when the steady state was reached, the beam diameter at the resonator mirrors had increased. The benzene cell acted as a spherical negative lens with a focal length of ca. 1 m.

The physical principles of TLS are rather simple. If a laser beam with a Gaussian-like intensity distribution propagates in a weakly absorbing medium the local temperature will increase and the transverse temperature profile will match the intensity profile of the laser beam. Since most liquids have a negative temperature coefficient of their refractive index ($dn/dT < 0$) the induced transverse profile of the refractive index acts as a negative (diverging) lens. The focal length of the lens depends on the parameters of the laser beam (power, diameter) and the characteristics of the medium (absorption coefficient, thermal conductivity, thermal diffusivity, dn/dT, cell length). Calculations show that in benzene a 0.01-W laser beam of 0.1 cm radius induces a negative thermal lens with focal length -1.14×10^4 cm. If a beam of the same power is focused to radius 0.01 cm, the focal length is -2.27×10^2 cm [2].

The absorption coefficient of a medium and the concentration of the absorbing species can be evaluated from the experimentally measured focal length value of an induced thermal lens. The simplest way to determine focal length is to measure the "far field" diameter of a laser beam with and without the sample. Since the optical strength of a thermal lens is not very high, the distance between the sample and the plane of measurement should be several meters. The maximum TLS signal can be detected optimally when the laser beam is focused just before the sample cell. This distance (the confocal length) is defined by the parameters of the laser beam and the focal length of the lens.

An optimal TLS scheme detects the fraction of laser beam power P passed through a pinhole in a screen, located in the farfield, being detected by a photomultiplier (or vacuum photodiode) and lock-in amplifier (Fig. 5). The TLS signal can be expressed as a normalized fractional change in the detected laser power:

$$[P(t=0) - P(t=\infty)]/P(t=\infty)$$

where $P(t=0)$ is the detected part of a laser beam not affected by the thermal lens effect, and $P(t=\infty)$ is the equilibrium value of P in the presence of the thermal lens. The absorption coefficient of a sample can be directly evaluated from this ratio and the parameters of the sample and laser beam. Both single- and two-beam versions of TLS are widely used in analytical applications. In the latter version, one beam of appropriate power is used as the heating beam while the second, low intensity, stable probe beam is used for signal detection.

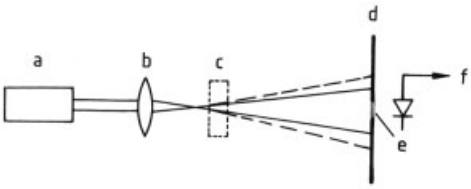

Figure 5. Schematic diagram of a single-beam thermal lens spectrometer
a) Laser; b) Lens; c) Sample; d) Screen; e) Pinhole; f) Detector
The dashed lines show the beam divergence due to the thermal lens effect

From the introduction of TLS, cw lasers with discrete wavelengths (e.g., He – Ne, He – Cd, Ar^+, Kr^+, CO_2) are the most widely used. These lasers have the best power stability and beam profile. In most cases, the absorption bands of the trace species of interest do not coincide with a discrete wavelength of the heating laser. Chemical pretreatment is necessary for formation of an analyte derivative with an appropriate absorption band. As in the case of LOAS, TLS has poor selectivity because of spectral interference of sample components. Selectivity depends mainly on the appropriate choice of reagent for analyte derivative formation.

The signal also depends on the solvent. Carbon tetrachloride, benzene, and acetone are preferred. For example, the TLS signal of a given analyte in water is 45 times smaller than in carbon tetrachloride. Suspended dust and air bubbles in a sample may perturb the laser beam, affecting the signal. To minimize these perturbations, liquids should be filtered and allowed to equilibrate in the cell before beginning the experiment. Absorption by the cell windows followed by a heating of the sample should also be minimized, as well as the formation of interference fringes due to the reflections of the laser beam by the optical components of the apparatus.

Originally, TLS was expected to be an absolute (standardless) technique because of the direct relationship between absorption coefficient and the experimental parameters. It was later recognized that the accuracy of the experimental and molecular parameters is not sufficient for absolute quantitative evaluation of the trace concentration of an analyte; calibration by standard reference solutions is necessary for quantitative analysis.

The minimum absorption coefficients measured by TLS are of the order of10^{-6} – $10^{-7}\,cm^{-1}$, which corresponds, for strong absorbers ($\varepsilon \geq 10^5\,M^{-1}\,cm^{-1}$), to minimal concentrations of ca. 10^{-11} – 10^{-12} M. Some data from real sample analysis by TLS are listed in Table 4.

4.4. Fluorescence Analysis of Organic Molecules in Solid Solutions [2]

Fluorescence is one of the most sensitive techniques for the detection of complex organic molecules such as aromatic hydrocarbons, heterocyclic compounds, porphyrins (including chlorophyll and its derivatives), organic dyes, etc. Its high sensitivity depends on the relatively high fluorescence quantum yields for these classes of molecules and well-developed methods of photon detection. If a pure, single-component solution is analyzed, extremely low LODs can be achieved, e.g., highly luminescent dyes in aqueous solution can be easily detected with LODs down to 10^{-13} M. For many organic compounds in liquid or solid solution form, single molecules can be detected by modern experimental techniques. In many cases, the main analytical problem is not sensitivity, but selectivity. At room temperature, the absorption and luminescence spectra of organic molecules consist typically of one or more bands whose spectral width is hundreds or thousands of wavenumbers; severe spectral interference is a key problem in the analysis of multicomponent mixtures.

The Shpol'skii method is a powerful improvement. The spectra of many organic molecules embedded in a crystalline matrix of short-chain *n*-paraffins consist, at helium temperature, of well-defined quasi-lines with

Table 4. Best detection limits for pure solutions and some examples of real sample analysis by TLS

Analyte	Solvent	Sample∗	Laser	Detection limit, wt %
Co^{2+}	$CHCl_3$	PS	He – Ne	1.2×10^{-6}
Cu^{2+}	water	PS	He – Ne	3.8×10^{-6}
Fe^{2+}	$CHCl_3$	PS	Ar^+	1.7×10^{-10}
Nd^{3+}	D_2O	PS	dye	3.0×10^{-5}
Pr^{3+}	water	PS	dye	7.2×10^{-3}
Tb^{3+}	butanol	PS	Ar^+	7.0×10^{-6}
UO_2^{2+}	water	NW	dye	4.7×10^{-7}
U^{6+}	$HClO_4$	NW	dye	7.0×10^{-6}
Am^{3+}	$HClO_4$	PS	dye	7.0×10^{-7}
Pu^{4+}	HCO_3^-	PS	dye	2.5×10^{-6}
Po_4^{3-}	water – acetone	PS	dye	5.0×10^{-10}
AsO_3^{3-}	2-butanol	NW	He – Ne	5.0×10^{-8}
SO_3^{2-}	water	PS	Ar^+	1.3×10^{-7}
S^{2-}	2-butanol	water, gas	He – Ne	5.0×10^{-8}
I_2	CCl_4	PS	Ar^+	1.0×10^{-16}∗∗
Formaldehyde	water	PS	dye	4.5×10^{-8}
Nitroaniline	methanol – water	PS	He – Cd	1.2×10^{-9}∗∗
Trimethylpentane	CCl_4	PS	He – Ne	8.0×10^{-9}
Azulene	liquid CO_2	PS	Kr^+	3.0×10^{-10}
NO_2	air	air	Ar^+	5.0×10^{-7}
Methanol	N_2	PS	CO_2	1.2×10^{-6}
CCl_2Fe_2	Ar	PS	CO_2	1.0×10^{-7}
CS_2	Ar	PS	CO_2	1.0×10^{-7}

∗ PS = pure solution; NW = natural water.
∗∗ Absolute detection limit, g.

linewidth 2 –20 cm^{-1}. In spite of many successful analytical applications the Shpol'skii technique is not universally applicable because many organic molecules have low solubility in *n*-paraffins and, moreover, many compounds in *n*-paraffin matrices have broad-band spectra even at helium temperature.

The discovery that in many cases the spectra of organic compounds at low temperature are inhomogeneously broadened was the next and very important step in the understanding of the spectroscopy of complex molecules. In most cases, the broad spectra have an intrinsic line structure, which can be revealed by selective laser excitation. This new technique, laser fine-structure selective spectroscopy of complex molecules in frozen solutions is often called site-selection spectroscopy (SSS). A physically more realistic term is energy selection spectroscopy. By now the SSS technique is so far developed that inhomogeneous broadening of organic molecules at low temperature can be eliminated by selective laser excitation [2].

The fundamentals of SSS are based on the theory of impurity centers in a crystal. The optical spectrum of an organic molecule embedded in a matrix is defined by electron – vibrational interaction with intramolecular vibrations (vibronic coupling) and interaction with vibrations of the solvent (electron – phonon coupling). Each vibronic band consists of a narrow zero-phonon line (ZPL) and a relatively broad phonon wing (PW). ZPL corresponds to a molecular transition with no change in the number of phonons in the matrix (an optical analogy of the resonance γ-line in the Mössbauer effect). PW is determined by a transition which is accompanied by creation or annihilation of matrix phonons. The relative distribution of the integrated intensity of a band between ZPL and PW is characterized by the Debye – Waller factor:

$$\alpha = I_{\mathrm{ZPL}} / (I_{\mathrm{ZPL}} + I_{\mathrm{PW}})$$

and depends critically on the temperature and the strength of electron – phonon coupling. At liquid helium temperature, if this coupling is not too strong, the vibronic bands of an isolated impurity molecule consist of a strong, narrow ZPL and a broad, weak PW. The observed broad band of an assembly of organic molecules in solution at low temperature is an envelope of a great number of narrow ZPLs (accompanied by PWs) of

individual molecules which are surrounded by different local fields (mainly because of small variations of the relative orientation of an analyte molecule and the solvent molecules). Statistical variations cause shifts in ZPL positions and, hence, inhomogeneous broadening of a spectral band. A narrow-band laser excites only a small fraction of the molecules with an appropriate ZPL position. Thus, well-separated, narrow ZPLs, accompanied by PWs, are observed in the fluorescence spectra of a multicomponent cryogenic solid mixture. This effect is often called fluorescence line narrowing (FLN). Informative spectra may be obtained for many types of molecule (polar and nonpolar), in different media (glasses, polymers, and crystals). The efficiency of FLN for a particular compound is well defined by the factor α. The higher the value of α, the better is excitation selectivity. For most systems, α varies within the interval 0.2 – 0.8. For compounds with $\alpha \ll 1$, selective excitation is inefficient. Examples of such molecules are ionic dyes (proflavines and rhodamines in ethanol).

Fluorescence line narrowing depends strongly on temperature. For most organic molecules, ZPL disappears completely at temperatures above 40 – 50 K. This is the main disadvantage of FLN for analytical applications; the sample temperature should be below 15 – 30 K.

The main components of FLN instruments for analysis are a cryogenic system, a tunable laser, a medium resolution spectrometer, and a recording system. The cryogenic system consists of a liquid He optical cryostat with a He circulation system. In some cases, commercially available closed-cycle cryostats can be used. The optical system consists of a sample holder, optics for the exciting beam input, and a high aperture collector of the fluorescent radiation.

The most intense 0 – 0 ZPLs of organic molecules of interest occupy the spectral range 650 – 250 nm. Thus, lasers for FLN should be tunable within this range. The laser linewidth should match the FLN vibronic linewidth of the analytes, typically of the order of 1 – 3 cm^{-1}. To avoid sample heating, saturation effects, and/or stimulated two-photon photoreactions, the average laser power should not exceed 1 – 10 mW. The cw dye lasers, pumped by Ar^+ or Kr^+ ion lasers are relatively cheap and reliable sources for the visible spectrum. For UV, frequency-doubled dye lasers pumped by copper or quasi-continuous Nd-YAG lasers (typical repetition rates 1 –10 kHz) may be preferable.

As the distance between the wavelengths of excitation and fluorescence in FLN is ca. 100 – 200 cm^{-1}, background scattering of the excitinglaser radiation may be a serious problem. Therefore, high-quality double- or even triple-grating spectrometers are used in FLN analysis. New multichannel recording systems (photodiode arrays, image intensifiers with linear CCDs) are potentially very useful detectors for FLN.

Polycyclic aromatic hydrocarbons (PAH) and heterocyclic aromatic compounds are the most widely analyzed organic compounds by FLN techniques. Both types of compound are widespread in nature. Many have strong carcinogenic and mutagenic properties and these properties depend on the isomeric structure of the compound. Analysis of PAH in a glassy matrix by FLN has many advantages compared with other techniques; simple sample preparation, high optical quality of the frozen matrices, etc., provide higher sensitivity and selectivity. LODs down to 10^{-12} – 10^{-13} g/cm^3 have been achieved for single-component solutions. For example, a concentration of perylene in ethanol of 10^{-12} g/mL (ca. 10^5 molecules in the sample) was easily detected in a frozen solution by FLN. The calibration curve was linear over the concentration range 10^{-6} – 10^{-9} g/cm^3. Well-characterized fine-structure FLN spectra were detected and identified in an artificial mixture of 14 PAHs in a water – glycerol solution. In another experiment, a mixture of 7 PAHs with concentrations of individual components ranging from 10^{-7} to 10^{-9} g/cm^3 was quantitatively analyzed and all compounds were correctly identified.

In some cases, direct analysis of raw products (without sample pretreatment or dissolution) is possible. To avoid self-absorption and energy transfer effects, low analyte concentrations and low optical density of the sample are necessary. Several PAH were successfully detected in gasoline when a frozen sample (T = 4.2 K) was selectively excited by laser radiation of appropriate wavelength. The same sample excited by a conventional Hg-lamp exhibited a smooth, unresolved spectrum with no possibility of identifying different compounds. As an example, the

active carcinogen 3,4-benzpyrene was easily detected in gasoline by FLN. The concentration of 10^{-7} g/mL was measured by the standard additions method. The estimated LOD depended on the background luminescence of a gasoline sample which is mainly determined by its composition. The LODs varied within 10^{-10} – 10^{-11} g/mL for different gasoline samples. By the same technique, perylene at concentration 5×10^{-8} g/mL was detected in solid paraffin.

5. Laser Ablation [3], [6], [11]

Laser ablation (LA) is a technique in which focused laser radiation is used to release material from a sample. The advantages of LA are now well known: reduced or no sample preparation is needed; conducting or non-conducting samples may be analyzed; spatial resolution up to a few micrometers can be obtained.

Different analytical techniques are used for detection of the elemental composition of the ablated material. The simplest one is direct detection of the emission of a plasma plume formed above a sample surface. This technique is generally referred to as LIBS (laser induced breakdown spectroscopy). Strong continuous background radiation of the hot plasma plume does not allow detection of atomic and ionic lines of specific elements during the first few hundred nanoseconds of plasma evolution. One can achieve reasonable signal-to-noise ratio for the measurement of atomic and ionic spectral line intensities by optimization of experimental parameters: type and pressure of a buffer gas, time delay, and distance from the sample surface for detection of line intensities. The LODs in (mg – μg)/g range have been realized for direct analysis of metals, glasses, and ceramics by LIBS techniques for a variety of trace elements. Better LODs were demonstrated when LA was followed by detection of the ablated atoms or ions by laser atomic absorption or fluorescence spectroscopy (LAAS, LIF). Even though matrix-effects are significant and detection limits are not as good as for LA-ICP-MS, LIBS has one major advantage. It allows fast and on-line analysis in industrial environment, including the nuclear industry. When fibers are used both for excitation and emission collection, a very robust system can be built that allows on-line quality control.

The intrinsic drawback of LIBS is a short duration (less than a few hundred microseconds) and significant non-stationary conditions of a laser plume. Much higher sensitivity was realized by transport of the ablated material into a secondary atomic reservoir such as an MIP. Owing to the much longer residence time of ablated atoms and ions in a stationary MIP (typically several ms as compared with at most a hundred microseconds in a laser plume) and due to additional excitation of the radiating upper levels in the low pressure plasma, the line intensities of atoms and ions are greatly enhanced. Because of both factors the LODs of the LA-MIP technique have been improved by 1 – 2 orders of magnitude as compared with LIBS.

Analytical capabilities of LIBS and LA-MIP-OES were noticeably improved by use of an advanced detection scheme based on an Echelle spectrometer combined with a high sensitivity ICCDs (intensified charge-coupled devices) detector. This detection scheme enables simultaneous detection of a large spectral range from the UV to NIR in a single laser shot. It allows estimation of temperature of a laser plume by constructing Boltzman plots and correction for plasma temperature variations. The advantages of this technique are: complete sample analysis in a single laser shot, improved accuracy and precision, the possibility to detect sample inhomogeneities.

The sensitivity, accuracy, and precision of solid sample analysis were greatly improved by coupling of LA with ICP-OES/MS. The ablated species are transported with a carrier gas (usually argon) into the plasma torch. Additional atomization, excitation and ionization of the ablated species in a stationary hot plasma provide a dramatic increase in the sensitivity of emission detection (LA-ICP-OES) or detection of ions (LA-ICP-MS). The efficiency of the transport of ablated species into an ICP strongly depends on the size of the particles. The optimal conditions for ablation in the case of LA-ICP differ significantly from the optimal conditions for LIBS because the efficient transport of the ablated matter to an ICP requires a fine aerosol (with solid particle diameters less than a few micrometers), whereas direct optical emission spectroscopy of the laser plume needs excited atoms and ions.

The LODs in the sub-μg/g – ng/g range were realized for different materials (metals, glasses,

polymers) by LA-ICP-OES. Even higher sensitivity was achieved for LA-ICP-MS due to the high efficiency of ion collection and detection. Under optimal conditions of ablation (choice of gas, diameter of a crater, flow rate of a carrier gas) the LODs in the ng/g – pg/g range are now routinely realized for most elements of the periodic table. Today the LA-ICP-MS technique is accepted as a most powerful technique for direct analysis of solid samples.

As a result of its attractive characteristics, LA-ICP-MS is used for a large variety of applications, such as the *in situ* trace element analysis of glasses, geological samples, metals, ceramic materials, polymers, and atmospheric particulates. Fingerprinting (characterization of the trace element composition) of diamonds and gold and glass and steel using LA-ICP-MS has been reported to be very useful for provenance determination and for forensic purposes, respectively.

So far the pulsed lasers with a pulse duration of some nanoseconds have been most widely used for LA – Nd: YAG (1064, 532, 354.7, 266 nm), excimer lasers XeCl (308 nm), KrF (248 nm). For these lasers the mechanisms responsible for the material removal are thermal melting and evaporation or some kind of explosive evaporation. As a result, so called fractional evaporation could occur – the elements with different melting and boiling temperatures evaporate from a melt at different rates. Because of this composition of a laser plume does not match the bulk composition. This was a serious problem as inadequate probing of a sample could not be corrected for by any sensitive detection scheme. The optimal conditions for laser ablation and advanced methods of data processing were found to avoid the problem of fractionation for different classes of samples in the case of nanosecond pulses.

The use of the lasers with femtosecond duration opens the way to overcome the problem of fractionation in the most radical way. The first generation of all-solid-state femtosecond lasers is now commercially available, although expensive. For such a fast energy deposition the non-thermal mechanisms of material removal (e.g., ion repulsions due to Coulomb forces) can be significant. The non-thermal character of material removal with fs pulses enables the depth profiling of a solid sample with a few nm resolution, which was impossible with ns pulses because of melting and mixing of different layers.

6. References

1. D. S. Kliger (ed.): *Ultrasensitive Laser Spectroscopy,* Academic Press, New York 1983.
2. V. S. Letokhov (ed.) : *Laser Analytical Spectrochemistry,* Adam Hilger, Bristol 1986.
3. J. Sneddon, T. L. Thiem, Y. Lee (eds.): *Lasers in Analytical Atomic Spectrometry,* John Wiley & Sons, New York 1997.
4. F. P. Schäfer (ed.): *Dye Lasers,* Springer Verlag, Heidelberg 1973.
5. M. Ch. Amann, J. Buus: *Tunable Laser Diodes,* Artech House, Boston 1998.
6. B. Smith: "Lasers in Analytical Chemistry," *Trends Anal. Chem.* **17** (1998) nos. 8 + 9.
7. S. Sjöström, *Spectrochimica Acta Rev.* **13** (1990) 407 – 465.
8. K. Niemax: "Diode Laser Spectroscopy," *Spectrochim. Acta Rev.* **15** (1993) no. 5.
9. H. I. Schiff, G. I. Mackay, J. Bechara, in M. W. Sigrist (ed.): *Air Monitoring by Spectroscopic Techniques,* John Wiley & Sons, New York 1994.
10. S. Martellucci, A. N. Chester (eds.): *Optoelectronics for Environmental Science,* Plenum Press, London 1991.
11. R. Russo: "Laser Ablation," *Appl. Spectr.* **49** (1995) no. 9, 14 A – 28 A.

Laser Diodes → Electroluminescent Materials and Devices

X-Ray Fluorescence Spectrometry

Ron Jenkins, International Centre for Diffraction Data, Newtown Square, Pennsylvania, United States

1. Introduction . . . 1
1.1. Properties of X-Rays . . . 2
1.2. Scattering and Diffraction . . . 2
1.3. Absorption of X-Rays . . . 3
1.4. X-ray Fluorescence . . . 3
2. Historical Development of X-ray Spectrometry . . . 3
3. Relationship Between Wavelength and Atomic Number . . . 4
3.1. Characteristic Radiation . . . 4
3.2. Selection Rules . . . 4
3.3. Nomenclature for X-ray Wavelengths 5
4. Instrumentation . . . 5
4.1. Sources . . . 5
4.2. Detectors . . . 6
4.3. Types of Spectrometer . . . 7
4.4. Wavelength Dispersive Systems . . . 8
4.5. Energy Dispersive Systems . . . 8
5. Accuracy . . . 9
5.1. Counting Statistical Errors . . . 9
5.2. Matrix Effects . . . 9
6. Quantitative Analysis . . . 9
6.1. Internal Standards . . . 10
6.2. Type Standardization . . . 10
6.3. Influence Correction Methods . . . 10
6.4. Fundamental Methods . . . 11
7. Trace Analysis . . . 11
7.1. Analysis of Low Concentrations . . . 11
7.2. Analysis of Small Amounts of Sample 12
8. New developments in Instrumentation and Techniques . . . 12
8.1. Total Reflection Spectrometry . . . 12
8.2. Spectrometers with Capillary Optics 13
8.3. Applications of the Synchrotron . . . 13
9. References . . . 14

1. Introduction

X-ray spectrometric techniques provided important information for the theoretical physicist in the first half of the 19th century, and since the early 1950s they have found increasing use in the field of materials characterization. While most of the early work in X-ray spectrometry was carried out with electron excitation [1], today most stand-alone X-ray spectrometers use X-ray excitation sources rather than electron excitation. X-ray fluorescence spectrometry typically uses a polychromatic beam of short wavelength X-radiation to excite longer wavelength characteristic lines from the sample to be analyzed. Modern X-ray spectrometers use either the diffracting power of a single crystal to isolate narrow wavelength bands, or a proportional detector to isolate narrow energy bands, from the polychromatic radiation (including characteristic radiation) excited in the sample. The first of these methods is called wavelength dispersive spectrometry and the second, energy dispersive spectrometry. Because the relationship between emission wavelength and atomic number is known, isolation of individual characteristic lines allows the unique identification of an element, and elemental concentrations can be estimated from characteristic line intensities.

1.1. Properties of X-Rays

X-rays are a short wavelength form of electromagnetic radiation discovered by Wilhelm Roentgen in the late 19th century [2]. When a high-energy electron beam is incident upon a specimen, one of the products of the interaction is the emission of a broad-wavelength band called the continuum, also referred to as white radiation or bremsstrahlung. This white radiation is produced as the atomic electrons of the elements making up the specimen decelerate the impinging high-energy electrons. Characteristic X-ray photons are produced following the ejection of an inner orbital electron from an excited atom and subsequent transition of atomic orbital electrons from states of high to low energy.

A beam of characteristic X-rays passing through matter is subject to three processes, absorption, scatter, and fluorescence. The absorption of X-rays involves mainly inner orbital atomic electrons and varies as the third power of

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_675

the atomic number of the absorber. X-rays are scattered mainly by the loosely bound outer electrons of an atom. Scattering of X-rays may be coherent (same wavelength) or incoherent (longer wavelength). Coherently scattered photons may undergo subsequent mutual interference, leading in turn to the generation of diffraction maxima. Fluorescence occurs when the primary X-ray photons are energetic enough to create electron vacancies in the specimen, leading in turn to the generation of secondary (fluorescence) radiation produced from the specimen. This secondary radiation is characteristic of the elements making up the specimen. The technique used to isolate and measure individual characteristic wavelengths following excitation by primary X-radiation is called X-ray fluorescence spectrometry.

1.2. Scattering and Diffraction

Scattering occurs when an X-ray photon interacts with the electrons of the target element. Where this interaction is elastic, i.e., no energy is lost in the collision process, the scattering is referred to as coherent (Rayleigh) scattering. Since no energy change is involved, the coherently scattered radiation retains exactly the same wavelength as that of the incident beam. It can also happen that the scattered photon gives up a small part of its energy during the collision. In this instance, the scatter is referred to as incoherent (Compton scattering).

X-ray diffraction is a combination of two phenomena – coherent scatter and interference. At any point where two or more waves cross one another, they are said to interfere. Under certain geometric conditions, wavelengths that are exactly in phase may add to one another, and those that are exactly out of phase may cancel each other. Under such conditions, coherently scattered photons may constructively interfere with each other, giving diffraction maxima. Since a crystal lattice consists of a regular arrangement of atoms, with layers of high atomic density throughout the crystal structure, a crystal can be used to diffract X-ray photons. Since scattering occurs between impinging X-ray photons and the loosely bound outer orbital atomic electrons, when a monochromatic beam of radiation falls onto the high atomic density layers, scattering will occur. In order to satisfy the requirement for constructive interference, it is necessary that the scattered waves originating from the individual atoms, that is, the scattering points, be in phase with one another. The geometric conditions for this are that:

$$n\lambda = 2d\sin\theta \tag{1}$$

where λ is the wavelength, d the interplanar spacing of the crystal, and n is an integer. Equation (1) is a statement of Bragg's law. The diffraction phenomenon is the basis of wavelength dispersive spectrometry, since by using a crystal of fixed $2\,d$, each wavelength is diffracted at a unique diffraction (Bragg) angle θ. Hence, by measuring the diffraction angle, knowledge of the d-spacing of the analyzing crystal allows the determination of the wavelength.

1.3. Absorption of X-Rays

When a beam of X-ray photons of intensity $I_0\,(\lambda)$ falls on a specimen, a fraction of the beam passes through, this fraction being:

$$I\,(\lambda) = I_0\,(\lambda)\,\exp\,(\mu\varrho x) \tag{2}$$

where μ is the mass attenuation coefficient of absorber for wavelength λ, ϱ is the density of the specimen, and x the distance traveled by the photons through the specimen. It can be seen from Equation (2) that a number $(I_0 - I)$ of photons has been lost in the absorption process. Although a significant fraction of this loss may be due to scatter, by far the greater loss is due to the photoelectric effect. Photoelectric absorption occurs at each of the energy levels of the atom; thus the total photoelectric absorption is determined by the sum of each individual absorption effect within a specific shell. Where the absorber is made up of a number of different elements, as is usually the case, the total absorption is the sum of the products of the individual elemental mass attenuation coefficients and the weight fractions of the respective elements. This product is referred to as the total matrix absorption. The value of the mass attenuation in Equation (2) is a function of both the photoelectric absorption and the scatter. However, the photoelectric absorption influence is usually large in comparison with the scatter,

and to all intents and purposes the mass absorption coefficient is equivalent to the photoelectric absorption. A plot of the mass attenuation coefficient as a function of wavelength contains a number of discontinuities, called absorption edges, at wavelengths corresponding to the binding energies of the electrons in the various subshells. Between absorption edges, as the wavelength of the incident X-ray photons increases, the absorption increases. This effect is very important in quantitative X-ray spectrometry, because the intensity of a beam of characteristic photons leaving a specimen is dependent on the relative absorption effects of the different atoms making up the specimen. This is called a matrix effect and is one of the reasons why a curve of characteristic line intensity as a function of element concentration may not be a straight line.

1.4. X-ray Fluorescence

X-ray fluorescence spectrometry provides the means for the identification of an element by measurement of its characteristic X-ray emission wavelength or energy. The method allows the quantitative estimation of a given element by first measuring the emitted characteristic line intensity and then relating this intensity to elemental concentration. The inherent simplicity of characteristic X-ray spectra make the process of allocating atomic numbers to the emission lines relatively easy, and the chance of making a gross error is rather small. Within the range of the conventional spectrometer, each element gives, on average, only half a dozen lines. A further benefit of using the X-ray emission spectrum for qualitative analysis is that, because transitions arise from inner orbitals, the effect of chemical combination, or valence state is almost negligible.

2. Historical Development of X-ray Spectrometry

While the roots of the X-ray spectrometric method go back to the early part of this century, it is only during the last thirty years or so that the technique has gained major significance as a routine means of elemental analysis. The first use of the X-ray spectrometric method dates back to the classic work of Henry Moseley [3]. One of the first papers on the use of X-ray spectroscopy for real chemical analysis appeared in 1922, when Hadding [4] described its use for the analysis of minerals. The use of X-rays, rather than electrons, to excite characteristic X-radiation [5] avoids the problem of heating of the specimen. Use of X-rays rather than electrons represented the beginnings of X-ray fluorescence as we know it today. The fluorescence method was first employed on a practical basis in 1928 by Glocker and Schreiber [6]. However, widespread use of the technique had to wait until the mid-1940s when X-ray fluorescence was rediscovered by Friedman and Birks [7]. The basis of their spectrometer was a diffractometer, originally designed for the orientation of quartz oscillator plates. A Geiger counter was used to measure the intensities of the diffracted characteristic lines, and quite reasonable sensitivity was obtained for a very large part of the atomic number range. The first commercial X-ray spectrometers became available in the early 1950s and, although these operated only with an air path, they were able to provide qualitative and quantitative information on all elements above atomic number 22 (titanium). Later versions allowed use of helium or vacuum paths that extended the lower atomic number cut-off.

3. Relationship Between Wavelength and Atomic Number

3.1. Characteristic Radiation

If a high-energy particle, such as an electron, strikes a bound atomic electron, and the energy of the particle is greater than the binding energy of the atomic electron, it is possible that the atomic electron will be ejected from its atomic position. The electron departs from the atom with a kinetic energy $E - \varphi$, equivalent to the difference between the energy E of the initial particle and the binding energy φ of the atomic electron. When the exciting particles are X-ray photons, the ejected electron is called a photoelectron, and the interaction between primary X-ray photons and atomic electrons is called the photoelectric effect. As long as the vacancy in the shell exists, the atom is in an unstable state,

and there are two processes by which it can revert to its original state. The first of these involves a rearrangement that does not result in the emission of X-ray photons, but in the emission of other photoelectrons from the atom. This is known as the Auger effect [8], and the emitted photoelectrons are called Auger electrons.

The second process by which the excited atom can regain stability is by transfer of an electron from one of the outer orbitals to fill the vacancy. The energy difference between the initial and final states of the transferred electron may be given off in the form of an X-ray photon. Since all emitted X-ray photons have energies proportional to the differences in the energy states of atomic electrons, the lines from a given element are characteristic of that element. The relationship between the wavelength of a characteristic X-ray photon and the atomic number Z of the element was first established by MOSELEY. Moseley's law is written:

$$1/\lambda = K\,(Z-\sigma)^2 \tag{3}$$

in which K is a constant that takes different values for each spectral series; σ is the shielding constant with a value just less than unity. The wavelength of the X-ray photon λ in Å is inversely related to the energy E of the photon in keV according to the relationship:

$$\lambda = 12.4/E \tag{4}$$

Since there are two competing routes by which an atom can return to its initial state, and since only one of these processes produces a characteristic X-ray photon, the intensity of an emitted characteristic X-ray beam depends on the relative effectiveness of the two processes within a given atom. As an example, the number of quanta of K series radiation emitted per ionized atom is a fixed ratio for a given atomic number, this ratio being called the fluorescent yield. The fluorescent yield varies as the fourth power of atomic number and approaches unity for higher atomic numbers. Fluorescent yield values are several orders of magnitude less for very low atomic numbers.

3.2. Selection Rules

An excited atom can revert to its ground state by transferring an electron from an outer atomic level to fill the vacancy in the inner shell. An X-ray photon is emitted from the atom as part of this de-excitation step, the emitted photon having an energy equal to the energy difference between the initial and final states of the transferred electron. The selection rules for the production of normal (diagram) lines require that the principal quantum number n must change by at least one, the angular quantum number l must change by only one, and the J quantum number [the total momentum J of an electron is given by the vector sum of $(l+s)$ where s is the spin quantum number] must change by zero or one (Table 1). Transition groups can now be constructed, based on the appropriate number of transition levels. Figure 1 shows the lines that are observed in the K series. Three groups of lines are indicated with the normal lines being shown on the left-hand side of the figure. While most of the observed fluorescent lines are normal, certain lines may also occur in X-ray spectra that at first sight do not fit the basic selection rules. These lines are called forbidden lines and are shown in the center portion of the figure. Forbidden lines typically arise from outer orbital levels where there is no sharp energy distinction between orbitals. A third type of line may also occur, namely, satellite lines, which arise from dual ionizations. Neither forbidden transitions nor satellite lines have great analytical significance; they may cause confusion in qualitative interpretation of spectra, and may even be misinterpreted as coming from trace elements.

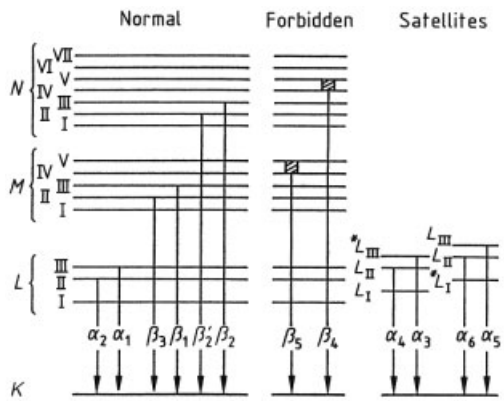

Figure 1. Observed lines in the K series (asterisks indicate ionized states)

Table 1. Atomic structures of the first three principal shells

Shell (electrons)	n	m	l	s	Orbitals	J
K (2)	1	0	0	$\pm\frac{1}{2}$	$1s$	$\frac{1}{2}$
L (8)	2	0	0	$\pm\frac{1}{2}$	$2s$	$\frac{1}{2}$
	2	0	0	$\pm\frac{1}{2}$	$2p$	$\frac{1}{2}$
	2	-1	1	$\pm\frac{1}{2}$	$2p$	$\frac{1}{2}, \frac{3}{2}$
	2	+1	1	$\pm\frac{1}{2}$	$2p$	$\frac{1}{2}, \frac{3}{2}$
M (18)	3	0	0	$\pm\frac{1}{2}$	$3s$	$\frac{1}{2}$
	3	1	0	$\pm\frac{1}{2}$	$3p$	$\frac{1}{2}$
	3	-1	1	$\pm\frac{1}{2}$	$3p$	$\frac{1}{2}, \frac{3}{2}$
	3	+1	1	$\pm\frac{1}{2}$	$3p$	$\frac{1}{2}, \frac{3}{2}$
	3	0	0	$\pm\frac{1}{2}$	$3d$	$\frac{1}{2}$
	3	-1	1	$\pm\frac{1}{2}$	$3d$	$\frac{1}{2}, \frac{3}{2}$
	3	+1	1	$\pm\frac{1}{2}$	$3d$	$\frac{1}{2}, \frac{3}{2}$
	3	-2	2	$\pm\frac{1}{2}$	$3d$	$\frac{3}{2}, \frac{5}{2}$
	3	+2	2	$\pm\frac{1}{2}$	$3d$	$\frac{3}{2}, \frac{5}{2}$

3.3. Nomenclature for X-ray Wavelengths

The classical nomenclature system for the observed lines is that proposed by Siegbahn in the 1920s. Figure 2 shows plots of the reciprocal of the square root of the wavelength as a function of atomic number for the K, L, and M series. As indicated by Moseley's law (Eq. 3), such plots should be linear. A scale directly in wavelength is also shown, to indicate the range of wavelengths over which a given series occurs. In practice, the number of lines observed from a given element depends on the atomic number of the element, the excitation conditions, and the wavelength range of the spectrometer. Generally, commercial spectrometers cover the K series, the L series, corresponding to transitions to K, L, and M levels respectively. There are a much larger number of lines in the higher series and for a detailed list of all of the reported wavelengths the reader is referred to the work of Bearden [9]. In X-ray spectrometry most analytical work is carried out using either the K or the L series wavelengths. The M series may also be useful, especially in the measurement of higher atomic numbers.

4. Instrumentation

4.1. Sources

Several different types of source have been employed for the excitation of characteristic X-radiation, including those based on electrons, X-rays, γ rays, protons, and synchrotron radiation. Sometimes a bremsstrahlung X-ray source is used to generate specific X-radiation from an intermediate pure element sample, called a secondary fluorescer. By far the most common source today is the X-ray photon source. This source is used in primary mode in the wavelength and primary energy dispersive systems, and in secondary fluorescer mode in secondary target energy dispersive spectrometers. A γ-source is typically a radioisotope that is used either directly, or in a mode equivalent to the secondary fluorescer mode in energy dispersive spectrometry. Most conventional wavelength dispersive X-ray spectrometers use a high-power (2 – 4 kW) X-ray bremsstrahlung source. Energy dispersive spectrometers use either a high- or low-power (0.5 – 1.0 kW) primary source, depending on whether the spectrometer is used in the secondary or primary mode. In all cases, the primary source unit consists of a very stable high-voltage generator, capable of providing a potential of typically 40 – 100 kV. The current from the generator is fed to the filament of the X-ray tube, which is typically a coil of tungsten wire. The applied current causes the filament to glow and emit electrons. A portion of this electron cloud is accelerated to the anode of the X-ray tube, which is typically a water-cooled block of copper with the required anode material plated or cemented to its surface. The impinging electrons produce X-radiation, a significant portion of which passes through a thin beryllium window to the specimen.

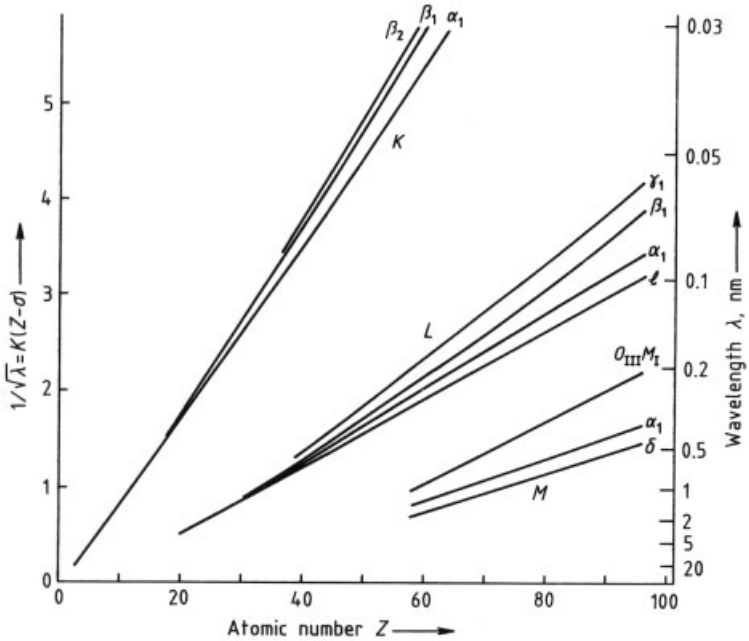

Figure 2. Moseley diagrams for K, L, and M series

4.2. Detectors

An X-ray detector is a transducer for converting X-ray photon energy into voltage pulses. Detectors work by a process of photoionization, in which interaction between the incoming X-ray photon and the active detector material produces a number of electrons. The current produced by these electrons is converted to a voltage pulse by a capacitor and resistor, such that one digital voltage pulse is produced for each entering X-ray photon. In addition to being sensitive to the appropriate photon energies, i.e., being applicable to a given range of wavelengths or energies, there are two other important properties that an ideal detector should possess: proportionality and linearity. Each X-ray photon entering the detector produces a voltage pulse, and if the size of the voltage pulse is proportional to the photon energy, the detector is said to be proportional. Proportionality is needed where the technique of pulse height selection is to be used. Pulse height selection is a means of electronically rejecting pulses of voltage levels other than those corresponding to the characteristic line being measured. X-ray photons enter the detector at a certain rate and, if the output pulses are produced at the same rate, the detector is said to be linear. Linearity is important where the various count rates produced by the detector are to be used as measures of the photon intensity for each measured line.

In the case of wavelength dispersive spectrometers, the gas flow proportional counter is generally employed for the measurement of longer wavelengths, and the scintillation counter for shorter wavelengths. Neither of these detectors has sufficient resolution to separate multiple wavelengths on its own, and so they have to be employed along with an analyzing crystal. However, in the case of energy dispersive spectrometry, where no dispersing crystal is employed, a detector of higher resolution must be used, generally the Si(Li) detector.

The Si(Li) detector consists of a small cylinder (ca. 1 cm diameter and 3 mm thick) of *p*-type silicon, compensated by lithium to increase its electrical resistivity. A Schottky barrier contact on the front of the silicon disk produces a *p-i-n*-type diode. To inhibit the mobility of the lithium ions and to reduce electronic noise, the diode and its preamplifier are cooled to liquid-nitrogen temperature. Incident X-ray photons interact to produce a specific number of

electron – hole pairs. The charge produced is swept from the diode by the bias voltage to a charge-sensitive preamplifier. A charge loop integrates the charge on a capacitor to produce an output pulse. Special energy dispersive detectors have been used in miniaturized instrumentation for in vivo applications [10] and for in situ analysis [11].

4.3. Types of Spectrometer

The basic function of the spectrometer is to separate the polychromatic beam of radiation coming from the specimen in order that the intensity of each individual characteristic line can be measured. A spectrometer should provide sufficient resolution of lines to allow such data to be taken, at the same time providing a sufficiently large response above background to make the measurements statistically significant, especially at low analyte concentrations. It is also necessary that the spectrometer allow measurements over the wavelength range to be covered. Thus, in the selection of a set of spectrometer operating variables, four factors are important: resolution, response, background level, and range. Owing to many factors, optimum selection of some of these characteristics may be mutually exclusive; as an example, attempts to improve resolution invariably cause lowering of absolute peak intensities. There is a wide variety of instrumentation available today for the application of X-ray fluorescence techniques and it is useful to break the instrument types down into three main categories: wavelength dispersive spectrometers (sequential and simultaneous); energy dispersive spectrometers (primary or secondary); and special spectrometers (including total reflection, synchrotron source, and proton-induced). The wavelength dispersive system was introduced commercially in the early 1950s and probably around 25 000 such instruments have been supplied commercially. Energy dispersive spectrometers became commercially available in the early 1970s and today there are around 15 000 of these units in use. There are far fewer of the specialized spectrometers in use and these generally incorporate an energy dispersive rather than a wavelength dispersive spectrometer.

X-ray spectrometers differ in the number of elements measurable at one time, and the speed at which they collect data. All of the instruments are, in principle at least, capable of measuring all elements from $Z = 9$ (fluorine) upwards, and most modern wavelength dispersive spectrometers can do useful measurements down to $Z = 6$ (carbon). Most systems can be equipped with multisample handling facilities and can be automated by use of minicomputers. Typical spectrometer systems are capable of precision on the order of a few tenths of a percent, with sensitivities down to the low ppm level. Single-channel wavelength dispersive spectrometers are typically employed for routine and nonroutine analysis of a wide range of products, including ferrous and nonferrous alloys, oils, slags and sinters, ores and minerals, thin films, etc. These systems are very flexible, but somewhat slower than multichannel spectrometers. Multichannel wavelength dispersive instruments are used almost exclusively for routine, high-throughput analyses where the great need is for fast, accurate analysis, but where flexibility is of no importance. Energy dispersive spectrometers have the great advantage of being able to display information on all elements at the same time. They lack somewhat in resolution compared with wavelength dispersive spectrometers, but the ability to reveal elements absent, as well as elements present, makes the energy dispersive spectrometer ideal for general trouble-shooting problems. They have been particularly effective in the fields of scrap alloy sorting, in forensic science, and in the provision of elemental data to supplement X-ray powder diffraction data.

4.4. Wavelength Dispersive Systems

A wavelength dispersive spectrometer may be a single-channel instrument in which a single crystal and a single detector are used for the sequential measurement of a series of wavelengths; or a multichannel spectrometer in which many crystal – detector sets are used to measure elements simultaneously. Of these two basic types, the sequential systems are the most common. A typical sequential spectrometer system consists of the X-ray tube, a specimen holder, a primary collimator, an analyzing crystal, and a detector. A portion of the characteristic fluorescence radiation from the specimen is passed via

a collimator or slit onto the surface of an analyzing crystal, where individual wavelengths are diffracted to the detector in accordance with the Bragg law. A goniometer is used to maintain the required $\theta - 2\theta$ relationship between crystal and detector. Typically, about six different analyzing crystals and two different collimators are provided in this type of spectrometer, giving the operator a wide choice of dispersion conditions. The separating power of a crystal spectrometer depends on the divergence allowed by the collimators, which to a first approximation determines the width of the diffracted lines, and the angular dispersion of the crystal [12]. While the maximum wavelength covered by traditional spectrometer designs is about 2 nm, more recent developments allow extension of the wavelength range significantly beyond this value.

The output from a wavelength dispersive spectrometer may be either analog (a) or digital (d). For qualitative work, an analog output is traditionally used; the digital output from the detector amplifier is fed through a d/a converter, called a rate meter, to an *x*/*t* recorder synchronously coupled with the goniometer scan speed. The recorder thus records an intensity – time diagram, in terms of an intensity – 2θ diagram. It is generally more convenient to employ digital counting for quantitative work, and a timer – scaler combination is provided which allows pulses to be integrated over a period of several tens of seconds and then displayed as count or count rate. In more modern spectrometers, a scaler – timer may take the place of the rate meter, using step-scanning. Some automated wavelength dispersive spectrometers provide the user with software programs for the interpretation and labeling of peaks [13].

4.5. Energy Dispersive Systems

The energy dispersive spectrometer consists of the excitation source and the spectrometer/detection system. The spectrometer/detector is typically a Si(Li) detector. A multichannel analyzer is used to collect, integrate, and display the resolved pulses. While similar properties are sought from the energy dispersive system as with the wavelength dispersive system, the means of selecting these optimum conditions are very different. Since the resolution of the energy dispersive system equates directly to the resolution of the detector, this feature is of paramount importance. The output from an energy dispersive spectrometer is generally displayed on a cathode ray tube, and the operator is able to dynamically display the contents of the various channels as an energy spectrum. Provision is generally made to allow zooming in on portions of the spectrum of special interest, to overlay spectra, to subtract background, etc. Generally, some form of minicomputer is available for spectral stripping, peak identification, quantitative analysis, and other useful functions.

All of the earlier energy dispersive spectrometers were operated in what is called the primary mode and consisted simply of the excitation source, typically a closely coupled low-power, end-window X-ray tube, and the detection system. In practice there is a limit to the maximum count rate that the spectrometer can handle and this led, in the mid-1970s, to the development of the secondary mode of operation. In the secondary mode, a carefully selected pure element standard is interposed between primary source and specimen, along with absorption filters, such that a selectable energy range of secondary photons is incident on the sample. This allows selective excitation of certain portions of the energy range, thus increasing the ratio of useful to unwanted photons entering the detector. While this configuration does not completely eliminate the count rate and resolution limitations of the primary system, it certainly does reduce them.

5. Accuracy

5.1. Counting Statistical Errors

The production of X-rays is a random process that can be described by a Gaussian distribution. Since the number of photons counted is nearly always large, typically thousands or hundreds of thousands, rather than a few hundred, the properties of the Gaussian distribution can be used to predict the probable error for a given count measurement. The random error $\sigma(N)$ associated with a measured value of N, is equal to $\sqrt{N}$. For example, if 10^6 counts are taken, the standard deviation σ is $\sqrt{10^6} = 10^3$, or 0.1 %. The measured parameter in wavelength dispersive X-ray spectrometry is generally the counting rate

R, and the magnitude of the random counting error associated with a given datum can be expressed as:

$$\sigma\,(\%) = 100/\sqrt{N} = 100/\sqrt{(Rt)} \tag{5}$$

Care must be exercised in relating the counting error (or indeed any intensity-related error) with an estimate of the error in terms of concentration. Provided that the sensitivity of the spectrometer in count rate (*c*/*s*) per percent analyte is linear, the count error can be directly related to a concentration error. However, where the sensitivity of the spectrometer changes over the range of measured response, a given fractional count error may be much greater when expressed in terms of concentration. In many analytical situations, the peak lies above a significant background and this adds a further complication to the counting statistical error.

5.2. Matrix Effects

In the conversion of net line intensity to analyte concentration, it may be necessary to correct for absorption and/or enhancement effects. Absorption effects include both primary and secondary absorption. Primary absorption occurs because all atoms of the specimen matrix absorb photons from the primary source. Since there is competition for these primary photons by the atoms making up the specimen, the intensity – wavelength distribution of the photons available for excitation of a given analyte element may be modified by other matrix elements. Secondary absorption refers to the absorption of characteristic analyte radiation by the specimen matrix. As characteristic radiation passes out from the specimen in which it was generated, it is absorbed by all matrix elements by amounts relative to their mass attenuation coefficients. The total absorption of a specimen depends on both primary and secondary absorption. Enhancement effects occur when a nonanalyte matrix element emits a characteristic line, of energy just in excess of the absorption edge of the analyte element. This means that the nonanalyte element is able to excite the analyte, giving characteristic photons in addition to those produced by the primary continuum. This gives an enhanced signal from the analyte.

6. Quantitative Analysis

The simplest quantitative analysis situation to handle is the determination of a single element in a known matrix. In this instance, a simple calibration curve of analyte concentration versus line intensity is sufficient for quantitative determination. A slightly more difficult case might be the determination of a single element where the matrix is unknown. Three basic methods are commonly employed in this situation: use of internal standards, addition of standards, and use of a scattered line from the X-ray source. The most complex case is the analysis of all, or most, of the elements in a sample, about which little or nothing is known. In this case a full qualitative analysis would be required before any attempt is made to quantitate the matrix elements. Once the qualitative composition of the sample is known, one of three general techniques is typically applied: type standardization, influence coefficient methods, or fundamental parameter techniques. Both the influence coefficient and fundamental parameter techniques require a computer for their application.

In principle, an empirical correction procedure can be described as the correction of an analyte element intensity for the influence of interfering elements, using the product of the intensity from the interfering element line and a constant factor, as the correction term [14]. This constant factor is today generally referred to as an influence coefficient, since it is assumed to represent the influence of the interfering element on the analyte. Commonly employed influence coefficient methods may use either the intensity or the concentration of the interfering element as the correction term. These methods are referred to as intensity correction and concentration correction methods, respectively. Intensity correction models give a series of linear equations which do not require much computation, but they are generally not applicable to wide ranges of analyte concentration. Various versions of the intensity correction models found initial application in the analysis of nonferrous metals where correction constants were applied as look-up tables. Later versions [15] were supplied on commercially available computer-controlled spectrometers, and were used for a wider range of application. A concentration model (e.g., [16])

requires the solving of a series of simultaneous equations by regression analysis or matrix inversion techniques. This approach is more rigorous than the intensity models, and so it became popular in the early 1970s as suitable low-cost minicomputers became available.

6.1. Internal Standards

One of the most useful techniques for the determination of a single analyte element in a known or unknown matrix is to use an internal standard. The technique is one of the oldest methods of quantitative analysis and is based on the addition of a known concentration of an element that gives a wavelength close to that of the analyte. The assumption is made that the effect of the matrix on the internal standard is essentially the same as the effect of the matrix on the analyte element. Internal standards are best suited to the measurement of analyte concentrations below about 10 %. Care must also be taken to ensure that the particle sizes of specimen and internal standard are about the same, and that the two components are adequately mixed. Where an appropriate internal standard cannot be found, it may be possible to use the analyte itself as an internal standard. This method is a special case of standard addition, and it is generally referred to as spiking.

6.2. Type Standardization

Provided that the total specimen absorption does not vary significantly over a range of analyte concentrations, and provided that enhancement effects are absent, and that the specimen is homogeneous, a linear relationship will be obtained between analyte concentration and measured characteristic line intensity. Where these provisos are met, type standardization techniques can be employed. Type standardization is probably the oldest of the quantitative analytical methods employed, and the method is usually evaluated by taking data from a well-characterized set of standards and, by inspection, establishing whether a linear relationship is indeed observed. Where this is not the case, the analyte concentration range may be further restricted. The analyst of today is fortunate in that many hundreds of good reference standards are commercially available. While the type standardization method is not without its pitfalls, it is nevertheless extremely useful, especially for quality control applications where a finished product is compared with a desired product.

6.3. Influence Correction Methods

Influence coefficient correction procedures can be divided into three basic types: fundamental, derived, and regression. Fundamental models are those which require starting with concentrations, then calculating the intensities. Derived models are based on some simplification of a fundamental method but still allow concentrations to be calculated from intensities. Regression models are semiempirical in nature and allow the determination of influence coefficients by regression analysis of data sets obtained from standards. All regression models have essentially the same form and consist of: a weight fraction term W (or concentration C); an intensity (or intensity ratio) term I; an instrument-dependent term which essentially defines the sensitivity of the spectrometer for the analyte in question; and a correction term, which corrects the instrument sensitivity for the effect of the matrix. The general form is:

$$W = I\,(\text{instrument})\,[1 + \{\text{model}\}] \tag{6}$$

Equation (6) shows that the weight fraction W of analyte is proportional to the product of the measured intensity I of the analyte, corrected for instrumental effects, and a matrix correction term. Different models vary only in the form of this correction term.

The major advantage of influence coefficient methods is that a wide range of concentration ranges can be covered by using a relatively inexpensive computer for the calculations. A major disadvantage is that a large number of well-analyzed standards may be required for the initial determination of the coefficients. However, where adequate precautions have been taken to ensure correct separation of instrument- and matrix-dependent terms, the correction constants are transportable from one spectrometer to another and, in principle, need only be determined once.

6.4. Fundamental Methods

Since the early work of SHERMAN, there has been a growing interest in the provision of an intensity – concentration algorithm which would allow the calculation of the concentration values without recourse to the use of standards. Sherman's work was improved upon, first by the Japanese team of SHIRAIWA and FUJINO [17] and later, by the Americans, CRISS and BIRKS [18], [19] with their program NRLXRF. The same group also solved the problem of describing the intensity distribution from the X-ray tube [20]. The problem for the average analyst in the late 1960s and early 1970s, however, remained that of finding sufficient computational power to apply these methods. In the early 1970s, DE JONG suggested an elegant solution [21], in which he proposed the use of a large main-frame computer for the calculation of the influence coefficients, then use of a minicomputer for their actual application using a concentration correction influence model.

7. Trace Analysis

One of the problems with any X-ray spectrometer system is that the absolute sensitivity (i.e., the measured *c/s* per % of analyte element) decreases significantly as the lower atomic number region is approached. The three main reasons for this are: first, the fluorescence yield decreases with atomic number; second, the absolute number of useful long-wavelength X-ray photons from a bremsstrahlung source decreases with increasing wavelength; and third, absorption effects generally become more severe with increasing wavelength of the analyte line. The first two of these problems are inherent to the X-ray excitation process and to constraints in the basic design of conventional X-ray tubes. The third, however, is a factor that depends very much on the instrument design and, in particular, upon the characteristics of the detector.

7.1. Analysis of Low Concentrations

The X-ray fluorescence method is particularly applicable to the qualitative and quantitative analysis of low concentrations of elements in a wide range of samples, as well as allowing the analysis of elements at higher concentrations in limited quantities of materials. A measurement of a line at peak position gives a counting rate which, in those cases where the background is insignificant, can be used as a measure of the analyte concentration. However, where the background is significant, the measured value of the analyte line at the peak position now includes a count-rate contribution from the background. The analyte concentration in this case is related to the net counting rate. Since both peak and background count rates are subject to statistical counting errors, the question now arises as to the point at which the net peak signal is statistically significant. The generally accepted definition for the lower limit of detection is that concentration is equivalent to two standard deviations of the background counting rate. A formula for the lower limit of detection LLD can now be derived [22]:

$$\mathrm{LLD} = 3/m\,\sqrt{R_\mathrm{b}/t_\mathrm{b}} \tag{7}$$

where R_b is the count rate at the background angle, t_b the time spent counting the background, and m the sensitivity of the spectrometer.

It is important to note that not only does the sensitivity of the spectrometer vary significantly over the wavelength range of the spectrometer, but so does the background counting rate. In general, the background varies by about two orders of magnitude over the range of the spectrometer. From Equation (7), the detection limit is optimal when the sensitivity is high and the background is low. Both the spectrometer sensitivity and the measured background vary with the average atomic number of the sample. While detection limits over most of the atomic number range lie in the low ppm range, the sensitivity of the X-ray spectrometer falls off quite dramatically towards the long-wavelength limit of the spectrometer due mainly to low fluorescence yields and the increased influence of absorption. As a result, poorer detection limits are found at the long-wavelength extreme of the spectrometer, which corresponds to lower atomic numbers. Thus, the detection limits for elements such as fluorine and sodium are at the levels of hundredths of one percent rather than parts per million. The detection limits for very low atomic number elements such as carbon ($Z = 6$) and oxygen ($Z = 8$) are very poor, typically on the order of 3 – 5 %.

7.2. Analysis of Small Amounts of Sample

Conventional X-ray fluorescence spectrometers are generally designed to handle rather large specimens with surface areas on the order of several square centimeters, and problems occur if the sample to be analyzed is limited in size. The modern wavelength dispersive system is especially inflexible in the area of sample handling, mainly because of geometric constraints arising from the need for close coupling of the sample to the X-ray tube and the need to use an airlock to bring the sample into the working vacuum. The sample to be analyzed is typically placed inside a cup of fixed external dimensions that is, in turn, placed in the carousel. This presentation system places constraints not only on the maximum dimensions of the sample cup, but also on the size and shape of samples that can be placed into the cup itself. Primary source energy dispersive systems do not require the same high degree of focusing, and to this extent are more easily applicable to any sample shape or size, provided that the specimen fits into the radiation-protected chamber. In some instances, the spectrometer can even be brought to the object to be analyzed. Because of this flexibility, analyses of odd-shaped specimens have been almost exclusively the purview of the energy dispersive systems.

In the case of secondary target energy dispersive systems, while the geometric constraints are still less severe than those of the wavelength system, they are much more critical than in the case of primary systems. Where practicable, the best solution for the handling of limited amounts of material is invariably found in one of the specialized spectrometer systems.

8. New developments in Instrumentation and Techniques

8.1. Total Reflection Spectrometry

One of the major problems that inhibits the obtaining of good detection limits in small samples is the high background due to scatter from the sample substrate support material. The suggestion to overcome this problem by using total reflection of the primary beam was made as long ago as 1971 [23], but the absence of suitable instrumentation prevented real progress being made until the late 1970s [24], [25]. The TRXRF method is essentially an energy dispersive technique in which the Si(Li) detector is placed close to (ca. 5 mm) and directly above the sample. Primary radiation enters the sample at a glancing angle of a few seconds of arc. The sample itself is typically presented as a thin film on the optically flat surface of a quartz plate. A beam of radiation from a sealed X-ray tube passes through a fixed aperture onto a pair of reflectors placed very close to each other. Scattered radiation passes through the first aperture to impinge on the sample at a very low glancing angle. Because the primary radiation enters the sample at an angle barely less than the critical angle for total reflection, this radiation barely penetrates the substrate medium; thus scatter and fluorescence from the substrate are minimal. Because the background is so low, picogram amounts can be measured or concentrations in the range of a few tenths of a ppb. can be obtained without recourse to pre-concentration [26]. Ideally, the radiation falling onto the specimen surface should be monochromatic. Unfortunately, in earlier executions of TXRF spectrometers the use of monochromatic source radiation was not possible because of the high intensity loss. This problem was solved with he advent of multilayer monochromators, since these exhibit high reflectivity and produce an intense almost parallel beam [27]. A TXRF spectrometer has been designed to operate with a standard X-ray tube and a multilayer monochromator, and this was found to give a two- to threefold increase in intensity over the normal geometry [28].

8.2. Spectrometers with Capillary Optics

Curved glass capillaries with channel diameters on the order of a few microns can efficiently scatter X-rays through large angles by multiple grazing incidence reflection. This technology generates a very high specific intensity, yielding an increase in intensity of two to three orders of magnitude [29]. The use of capillary optics has opened up tremendous possibilities in energy dispersive X-ray fluorescence. A new

type of X-ray spectrometer has now evolved, broadly categorized as microscopic X-ray fluorescence (MXRF). Earlier versions of MXRF systems simply used an aperture X-ray guide to control the size of the irradiated spot. Use of a monolithic polycapillary optic in place of the original aperture was found to provide more than two orders of magnitude increase in intensity [30]. There are two basic types of MXRF based on capillary optics: the first is based on classical sealed X-ray tube sources, and the second on the use of the synchrotron. As an example, one low-power, portable XRF analyzer, based on this technology [31], allowed the analysis of specimens as small as 30 mm in diameter with detection limits down to 5 x 10^{-5} %. Gormley [32] has described a series of experiments, based on the use of a CCD camera, to better quantify the focusing and transmission characteristics of these capillaries. Another instrument was designed with an 80 mm^2detector [33] and allowed the analysis of small (5 mm) areas of bulky objects. Each of these spectrometers was designed to work with low-power X-ray tubes. At the other end of the source power scale, Vekemans et al. described an instrument based on a rotating anode X-ray tube. This high-efficiency, microscopic X-ray fluorescence system has been utilized for the analysis of 16-20th century brass statues originating from Northern India and Nepal and for the imaging analysis of enamel paints on 17th century Chinese porcelain [34].

8.3. Applications of the Synchrotron

The availability of intense, linearly polarized synchrotron radiation beams [35] has prompted workers in the fields of X-ray fluorescence [36] to explore what the source has to offer over more conventional excitation media. In the synchrotron, electrons with kinetic energies on the order of several billion electron volts (typically 3 GeV at this time) orbit in a high-vacuum tube between the poles of a strong (ca. 1 T) magnet. A vertical field accelerates the electrons horizontally, causing the emission of synchrotron radiation. Thus synchrotron source radiation can be considered as magnetic bremsstrahlung as opposed to the normal electronic bremsstrahlung that is produced when primary electrons are decelerated by the electrons of an atom. It has been found that because the primary source of radiation is so intense, it is possible to use a high degree of monochromatization between source and specimen, giving a source that is wavelength- and, therefore, energy-tunable, as well as being highly monochromatic. There are several different excitation modes that can be used for synchrotron-source X-ray fluorescence spectroscopy (SSXRF) including direct excitation with continuum, excitation with absorber-modified continuum, excitation with source crystal monochromatized continuum, excitation with source radiation scattered by a mirror, and reflection and transmission modes.

Most SSXRF is carried out in the energy dispersive mode. Since all solid-state detectors are count-rate limited, the tremendous radiation flux available from the synchrotron is, in general, not directly useful. The real power of the synchrotron comes from the use of highly conditioned beams, which provide a highly focused, monochromatic beam at the sample. A newer development in the area of beam conditioning for the synchrotron is the use of laterally graded multilayers [37]. The use of X-ray optical multilayers in X-ray instrumentation became widespread in the 1990s, and newer technology based on pulsed laser deposition is resulting in the production of multilayers of uniform thickness with a minimum of columnar thin-film growth. The intensity of the synchrotron beam is probably four to five orders of magnitude greater than the conventional bremsstrahlung source sealed X-ray tubes. This, in combination with its energy tunability and polarization in the plane of the synchrotron ring, allows very rapid bulk analyses to be obtained on small areas. Because the synchrotron beam has such a high intensity and small divergence, it is possible to use it as a microprobe of high spacial resolution (about 10 µm). Absolute limits of detection around 10^{-14} µg have been reported for such an arrangement [38]. Use of monolithic polycapillary X-ray lenses has added an even wider dimension to synchrotron microbeam analysis, and beam sizes down to 4 – 10 µm have been reported [34]. Such systems have been successfully applied in trace element determination [39].

Synchrotron source X-ray fluorescence has also been used in combination with TRXRF. The small divergence of the continuous beam can

be caused to strike the sample at less than the critical angle. Very high signal-to-background ratios have been obtained on employing this arrangement for the analysis of small quantities of aqueous solutions dried on the reflector, with detection limits of < 1 ppb, i.e., 1 pg. [40]. Additional advantages accrue because synchrotron radiation is highly polarized, and background due to scatter can be greatly reduced by placing the detector at 90° to the path of the incident beam and in the plane of polarization. A disadvantage of the SSXRF technique is that the source intensity decreases with time, but this can be overcome by bracketing analytical measurements between source standards and/or by continuously monitoring the primary beam.

9. References

1. G. v. Hevesey: *Chemical Analysis by X-Rays and its Application,* McGraw-Hill, New York 1932.
2. W. C. Roentgen, *Ann. Phys. Chem.* **64** (1898) 1.
3. H. G. I. Moseley, *Philos. Mag.* **26** (1912) 1024; **27** (1913) 703.
4. A. Hadding, *Z. Anorg. Allg. Chem.* **122** (1922) 195.
5. D. Coster, J. Nishina, *Chem. News J. Ind. Sci.* **130** (1925) 149.
6. R. Glocker, H. Schreiber, *Ann. Phys. (Leipzig)* **85** (1928) 1085.
7. L. S. Birks: *History of X-Ray Spectrochemical Analysis,* American Chemical Society Centennial Volume, ACS, Washington, D.C., 1976.
8. P. Auger, *Compt. R.* **180** (1925) 65; *Journal de Physique* **6,** 205.
9. J. A. Bearden: *X-Ray Wavelengths,* U.S. Atomic Energy Commission Report NYO-10586, 1964, 533 pp.
10. J. Iwanczyk, B. E. Patt, *Adv. X-ray Anal.,* **41** (1997) 951 – 957.
11. W. T. Elam, et.al., *Adv. X-ray Anal.,* **42** (1999) 137 – 145.
12. R. Jenkins: *An Introduction to X-Ray Spectrometry,* Chap. 4, Wiley/Heyden, London 1974.
13. M. F. Garbauskas, R. P. Goehner, *Adv. X-Ray Anal.* **26** (1983) 345.
14. M. J. Beattie, R. M. Brissey, *Anal. Chem.* **26** (1954) 980.
15. H. J. Lucas-Tooth, C. Pyne, *Adv. X-Ray Anal.* **7** (1964) 523.
16. G. R. Lachance, R. J. Traill, *Can. Spectrosc.* **11** (1966) 43.
17. T. Shiraiwa, N. Fujino, *Bull. Chem. Soc. Jpn.* **40** (1967) 1080.
18. J. W. Criss, L. S. Birks, *Anal. Chem.* **40** (1968) 1080.
19. J. W. Criss, *Adv. X-Ray Anal.* **23** (1980) 93.
20. J. V. Gilfich, L. S. Birks, *Anal. Chem.* **40** (1968) 1077.
21. W. K. de Jongh, *X-Ray Spectrom.* **2** (1973) 151.
22. R. Jenkins, J. L. de Vries: *Practical X-Ray Spectrometry,* 2nd ed., Springer Verlag, New York 1970.
23. Y. Yoneda, T. Horiuchi, *Rev. Sci. Instrum.* **42** (1971) 1069.
24. H. Schwenke, J. Knoth, *Nucl. Instrum. Methods Phys. Res.* **193** (1982) 239.
25. J. Knoth, H. Schwenke, *Fresenius Z. Anal. Chem.* **301** (1978) 200.
26. H. Aiginger, P. Wobrauschek, *Adv. X-Ray Anal.* **28** (1985) 1.
27. E. Zeigler, O. Hignette, M. Lingham, A. Souvorov, *SPIE*, vol. 2856/61 (1096).
28. R. Schwaiger, P. Wobrauschek, C. Streli, *Adv. X-ray Anal.,* **41** (1997) 804 – 811.
29. M. A. Kumakhov, *Nucl. Instr. Methods,* **B48** (1990) 283.
30. G. Worley, G. Havrilla, N. Gao, Q. F. Xiao, *Adv. X-ray Anal.* **42** (1998) 26 – 35.
31. M. A. Kumakhov, *Adv. X-ray Anal.* **41** (1997) 214 – 226.
32. J. Gormley, *Adv. X-ray Anal,* **41** (1997) 239 – 242.
33. J. A. Nicolosi, E. Scrugge, *Adv. X-ray Anal.* **41** (1997) 227 – 233.
34. *Adv. X-ray Anal.* **41** (1997) 278 – 290.
35. C. J. Sparks Jr. in H. Winnick, S. Doniach (eds.): *Synchrotron Radiation Research,* Plenum Press, New York 1980, p. 459.
36. J. V. Gilfrich et al., *Anal. Chem.* **55** (1983) 187.
37. T. Holz et al., *Adv. X-ray Anal.* **41** (1997) 346 – 355.
38. W. Petersen, P. Ketelsen, A. Knoechel, R. Pausch, *Nucl. Instrum. Methods Phys. Res. Sect. A* **246** (1986) no. 1 – 3, 731.
39. K. W. Jones, B. M. Gordon, *Anal. Chem.* **61** (1989) 341A.
40. A. Iida, Y. Gohshi, *Adv. X-ray Anal.* **28** (1985) 61 – 68.

Activation Analysis

Richard Dams, Gent University, Laboratory Analytical Chemistry, Gent, Belgium

Karel Strijckmans, Gent University, Laboratory Analytical Chemistry, Gent, Belgium

1. Introduction 1
2. Neutron Activation Analysis 2
2.1. Basic Principles 2
2.2. Methods of Neutron Activation 3
2.3. Standardization 6
2.4. Sources of Errors 7
2.5. Measuring Equipment 8
2.6. Procedures for Activation Analysis . . 11
2.7. Sensitivity and Detection Limits 12
3. Photon Activation Analysis 14
4. Charged-Particle Activation Analysis 14
5. Applications 15
5.1. High-Purity Materials 15
5.2. Environmental Materials 16
5.3. Biological Materials 16
5.4. Geo- and Cosmochemistry 17
5.5. Art and Archaeology 17
6. Evaluation of Activation Analysis . . . 17
7. References 18

In addition to the standard symbols defined in the front matter of this volume the following symbols and abbreviations are used.

A activity (count rate, decay rate)
AA activation analysis
B background
C, D correction terms
CPAA charged-particle activation analysis
E element
E energy
ENAA epithermal neutron activation analysis
FNAA fast-neutron activation analysis
FWHM full width at half maximum
I beam intensity
I_0 resonance integral
INAA instrumental neutron activation analysis
IPAA instrumental photon activation analysis
l length
MCA multichannel analyzer
n particle density
N number of nuclides, counts
NAA neutron activation analysis
PAA photon activation analysis
R reaction rate; range
RBS Rutherford backscattering
S saturation function; signal; stopping power
SCA single-channel analyzer
TNAA thermal neutron activation analysis
γ relative intensity of gamma radiation
ε detection efficiency
Θ abundance of isotope
κ gamma quantum; particle
λ disintegration constant
σ cross section
τ dead time of detector
φ flux density of radiation (fluency rate)
Φ flux (fluency)

1. Introduction

Activation analysis has, in a few decades, become one of the most important methods for determination of minor (0.1 – 10 mg/g), trace (1 – 100 µg/g), and ultratrace (1 – 1000 ng/g) elements in solid samples. The main advantages of activation analysis are its accuracy and sensitivity (detection limit). Moreover, it is an independent method, i.e., not subject to the same systematic errors as other, more commonly used analytical methods. When a radiochemical separation is applied (after irradiation and prior to measurement) it is not subject to systematic errors due to reagent blanks, sample contamination (after irradiation), and non-quantitative (and even non-reproducible) yield. Precision goes down to a few percent in most favorable cases. The method is applied in the semiconductor industry, medicine, biology, criminology, archaeology, geochemistry, and environmental studies or control. Activation analysis is based on the exposure of a sample to a flux of activating parti-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_689

cles or radiation. Irradiation is performed mainly with thermal neutrons (90 %), but fast neutrons (3 %), charged particles (4 %), or photons (3 %) can also be used (figures based on publications in Chemical Abstracts over the last two decades). This induces a nuclear reaction, whereby an excited intermediate is formed, which deexcites in 10^{-14} s by emission of prompt gamma rays, which can be measured, or the prompt emission of another particle [neutron(s), proton, alpha] i.e., nuclear reaction analysis (NRA). However, in activation analysis (AA) the resulting radioactive nuclide is monitored by means of the emitted alpha, beta, gamma or X-ray quanta, or delayed neutrons. The radiation emitted during the decay is measured for the identification and quantification of the target elements.

2. Neutron Activation Analysis

For a number of reasons, neutrons have found the widest applications as bombarding particles in activation analysis. First, the cross sections for reactions of nuclides with neutrons, yielding primarily (n, γ) reactions, but also (n, α); (n, p), (n, 2 n), and (n, n′) reactions, are often large. The type of reaction depends on the energy of the neutron and is chosen individually for each case, on the basis of the following considerations:

1) A suitable reaction product with not too long a half-life (preferably below one year) and with decay parameters compatible with the available counting equipment
2) A sufficiently high cross section
3) The absence of important interfering reactions yielding the same radionuclide
4) Easy treatment of the sample after irradiation (limited matrix activity)

Second, intense neutron sources (nuclear reactors, neutron generators, isotopic sources) were generally available. Third, the high penetration depth of neutrons allows a nearly homogeneous flux density of the entire sample to be analyzed.

2.1. Basic Principles

If a target with n^A nuclides of A per cubic centimeter is placed in a neutron beam φ (number of neutrons $cm^{-2}\ s^{-1}$) then the number of interactions to form nuclei B is given by the reaction rate R ($cm^{-3}\ s^{-1}$)

$$R = \varphi \sigma n^{\mathrm{A}} \tag{1}$$

where σ is the cross section. The general nuclear reaction can be written as follows:

$$\mathrm{A}\,(n, \kappa) \xrightarrow{\lambda} \mathrm{C}$$

where n is a neutron; κ the promptly emitted gamma quantum or particle; λ is the disintegration constant of the radioactive nuclide B produced (equal to $\ln 2/t_{1/2}$; $t_{1/2}$ denotes the half-life of the radionuclide B); and C is the nuclide produced after radioactive decay of B.

From Equation (1) the definition of the cross section σ for this reaction is obtained:

$$\sigma = \frac{\text{Number of interactions}}{\varphi n^{\mathrm{A}}} \tag{2}$$

The cross section is expressed as an area in square centimeters or, more practically, in barn (1 b = 10^{-24} cm^2).

The number N^A of target nuclei is assumed to remain constant, as is the flux density φ throughout the target. Since the nuclides B decay at a rate equal to $-\lambda N^B$,

$$\frac{\mathrm{d}N^{\mathrm{B}}}{\mathrm{d}t} = \varphi \sigma N^{\mathrm{A}} - \lambda N^{\mathrm{B}} \tag{3}$$

If no nuclides B were present before irradiation ($N^B = 0$ for $t = 0$), integration gives, after an irradiation time t,

$$N^{\mathrm{B}} = \frac{\varphi \sigma N^{\mathrm{A}}}{\lambda} \left(1 - \mathrm{e}^{-\lambda t}\right) \tag{4}$$

The absolute decay rate A_0, in terms of the number of decays per second, at the end of the irradiation is given by

$$A_0 = N^{\mathrm{B}} \lambda = \varphi \sigma N^{\mathrm{A}} \left(1 - \mathrm{e}^{-\lambda t}\right) \tag{5}$$

The expression

$$S = 1 - \mathrm{e}^{-\lambda t} = 1 - \mathrm{e}^{\ln(2)\, t/t_{1/2}} \tag{6}$$

is termed the saturation function. After an irradiation time equal to one half-life of B, a saturation of 50 % is obtained. For irradiation times that are much lower than the half-life, the saturation function can be considered approximately linear:

$$S \approx \ln(2)\, t/t_{12} \quad (\text{for } t < 0.1 t_{1/2}) \tag{7}$$

A correction term ($D = \mathrm{e}^{-\lambda t_d}$) must be applied for the decay during the time t_d between the end of irradiation and the measuring time. When the counting (measuring) time is not very short ($t_\mathrm{m} > 0.3\, t_{1/2}$) compared with the half-life, a correction C for decay during the counting period t_m can also be applied. It equals

$$C = \frac{1 - \mathrm{e}^{-\lambda t_\mathrm{m}}}{\lambda t_\mathrm{m}} \tag{8}$$

Finally the counting equipment has a detection efficiency ε, which is usually less than unity. Thus the number of counts N measured after an irradiation time t, a waiting time t_d, and during a counting time t_m becomes

$$N = \varphi \sigma N^\mathrm{A} \left(1 - \mathrm{e}^{-\lambda t}\right) \mathrm{e}^{-\lambda t_\mathrm{d}} \frac{1 - \mathrm{e}^{-\lambda t_\mathrm{m}}}{\lambda} \varepsilon \tag{9}$$

$$= \frac{\varphi \sigma m \theta N_\mathrm{A} S D C \varepsilon}{A_\mathrm{r}} t_\mathrm{m} \tag{10}$$

where m is the mass of the element A; A_r is its relative atomic mass; θ is the abundance of the isotope yielding the radioactive isotope; and N_A is the Avogadro constant.

2.2. Methods of Neutron Activation

Reactor Neutrons. Nuclear reactors are the largest and most often used neutron sources. Neutrons formed in the reactor have a continuous energy spectrum extending from nearly 0 to 15 MeV. Conventionally the neutron spectrum is divided in three components:

1) *Fast or fission neutrons*, produced by the fission of ^{235}U with energies ranging from 0.1 to 15 MeV. The mean energy is ca. 2 MeV. These neutrons give rise to so-called threshold reactions, e.g., (n, p), (n, α), and (n, 2 n).
2) *Epithermal or resonance neutrons* have energies varying from 0.1 MeV down to about 0.5 eV. In the ideal case the epithermal flux Φ_epi is inversely proportional to the neutron energy E

$$\Phi_\mathrm{epi}(E) = \frac{\Phi_\mathrm{epi}}{E} \tag{11}$$

The highest resonance neutron flux is in the reactor core.

3) *Thermal or moderated neutrons* are in thermal equilibrium with the atoms of the moderator, and their velocities of 2200 m/s at 20 °C correspond to an energy of 0.0253 eV.

In some research reactors, pulses of very high neutron flux may be generated by removing the control rods. The neutron flux increases immediately to several times 10^{16} cm^{-2} s^{-1} and falls after ca. 10 ms to a very low value.

Table 1 shows that at most irradiation facilities the highest fluxes belong to thermal neutrons. Fortunately for activation analysis, they are the most useful ones. As shown in Figure 1 for the lower-energy neutrons (up to several hundred electronvolts) the decrease in cross section with increasing energy is roughly proportional to $1/v$ (where v is the velocity of the neutrons). Sometimes superimposed on this smoothly dropping curve are sharp resonance peaks in the epithermal region just above 1 eV. The total cross section for capture of epithermal neutrons is, therefore, the sum of the $1/v$ probability and that associated with these superimposed resonance peaks, and is called the resonance integral I_0. In neutron activation analysis, thermal and epithermal flux are separated by the so-called cadmium cutoff energy E_Cd of 0.55 eV. Cadmium in fact strongly absorbs neutrons with energies of less than 0.55 eV.

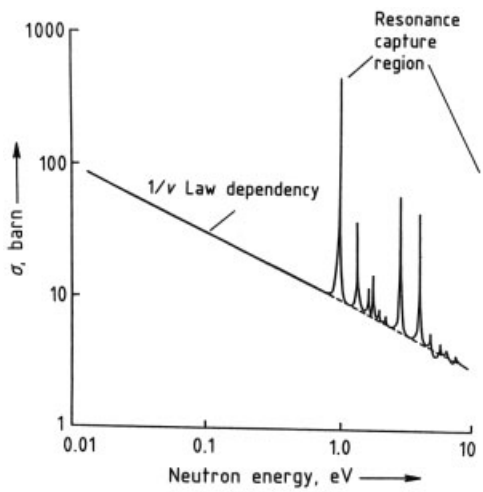

Figure 1. Typical excitation function for neutron-induced reaction, illustrating $1/v$ dependency and resonance peaks in epithermal region

Table 1. Typical neutron fluxes at irradiation facilities in research reactors

Irradiation site	Neutron flux, $cm^{-2}\ s^{-1}$		
	Thermal	Epithermal	Fast
Large reactor			
Reactor core	2×10^{14}	3×10^{13}	4×10^{13}
Surrounding moderator	$1-2\times10^{14}$	$0.5-2\times10^{13}$	$0.2-2\times10^{13}$
Small reactor			
Reactor core	5×10^{12}	4×10^{11}	3×10^{12}
Reflector	$0.3-2\times10^{12}$	$0.2-8\times10^{10}$	$0.3-30\times10^{10}$

In Equation (1) and following, the reaction rate can thus be split up according to the Høgdahl convention [29]

$$R=N\left(\Phi_{\mathrm{th}}\sigma_0+\Phi_{\mathrm{epi}}I_0\right) \tag{12}$$

where σ_0 is the tabulated thermal neutron cross section for the corresponding energy (0.0253 eV). The conventional thermal flux is, with a thermal neutron density n

$$\varphi_{\mathrm{th}}=nv0$$

The second term in Equation (12) is the somewhat more complicated solution for the integration of the reaction probability in the epithermal region. The resonance integrals I_0 have been tabulated:

$$I_0=\int_{E_{\mathrm{Cd}}}^{\infty}\sigma_{\mathrm{epi}}(E)\,\frac{\mathrm{d}E}{E} \tag{13}$$

Values of φ_{th} and φ_{epi} can be found by irradiating a nuclide (monitor) with known σ_0 and I_0 in a cylindrical cadmium box with 1-mm wall thickness. Often, gold is used as a monitor. The cross sections for slow neutron reactions are often on the order of several barns and can in some cases be as large as 10^4 barn. The predominant reaction for slow neutrons is the (n, γ) reaction.

Equation (10) allows calculation, to a first approximation, of the detection limit attainable by neutron activation analysis. If a count rate of $A=1\ s^{-1}$ can be detected and $A_r=100$, $S=0.5$, $\sigma=1$ barn, $\varphi=3\times10^{12}\ cm^{-2}\ s^{-1}$, $\varepsilon=0.33$, with θ, D, and C close to unity, a mass of 10^{-9} g of a typical element can be detected; for many elements with larger cross sections or when irradiating at high neutron fluxes, much lower masses can be detected.

When determining a specific element, the *half-life* of the radionuclide to be measured is a very important parameter. For the production of a radioisotope with a half-life of only a few hours or less, the irradiation time should be limited to a saturation of ca. 0.5 – 0.75, to avoid production of much longer-lived radionuclides. The short-lived radionuclides must be measured soon after the end of irradiation. Also, the counting period should not greatly exceed the half-life of the radionuclide. This however, results in poor counting statistics. For really short-lived radionuclides, this drawback can to some extent be overcome by *cyclic activation analysis* which is based on the use of repetitive short irradiations and counting periods, and summing of the γ-ray spectra. Also *pulsed neutron irradiation* enhances the detection efficiency of radionuclides with $t_{1/2}<20$ s. In contrast, to determine elements producing longer-lived radionuclides (e.g., $t_{1/2}>5$ d), much longer irradiation times are required although activation to a high saturation factor is not meaningful. Treatment of the sample and counting should be postponed several days to allow for decay of the shorter-lived radionuclides.

An additional selectivity factor is the neutron energy. By covering the sample with cadmium or boron carbide, the thermal neutrons are filtered out and the sample is activated largely with epithermal neutrons. (Boron has a nearly pure $1/v$ excitation function and is therefore also chosen to filter out thermal low-energy resonance neutrons.) A distinction is made between *thermal* (*TNAA*) and *epithermal neutron activation analysis* (*ENAA*). In the latter, selective activation of a number of elements relative to other interfering elements can be enhanced. ENAA is often advantageous because most of the matrix ele-

ments in geological and environmental samples (e.g., Na, Al, P, K, Ca, Sc, Cr, Mn, and Fe) are activated primarily with thermal neutrons (σ_0), whereas a number of trace or ultratrace elements (e.g., As, Sb, Se, Sn, Co, Ni, Ga, Mo, Ag, Pd, In, Cd, Br, I, Rb, Sr, Cs, Ta, W, Th, U, Pt, and Au) have large resonance integrals (I_0). Not only is the relative activation of many trace elements enhanced, but the overall activity is also generally reduced so that the sample can be measured much closer to the detector, which improves the counting geometry and thus the precision. Since fast neutrons induce threshold reactions, *fast-neutron activation analysis* (*FNAA*) is used primarily for the determination of light elements (O, N, F, P, S, Cl, Si, Mg, Al, Fe), several of which cannot (or can barely) be determined by thermal neutron activation. Although the reactor neutron spectrum covers a wide range of fast-neutron energies, neutron generators are much more appropriate for FNAA.

Neutron Generators. Neutrons can also be obtained by bombarding an appropriate target with ions accelerated to high kinetic energies. In cyclotrons, a typical neutron flux of 2×10^{11} s^{-1} is obtained with a 10-μA beam of 30-MeV deuterons on a thick beryllium target. Muchmore important, however, are the Cockroft – Walton accelerators, called 14-MeV neutron generators, in which neutrons are produced by the ^{3}H (d, n) ^{4}He reaction. The neutrons obtained have an energy of 14.7 MeV and are emitted nearly isotropically from the target. The total neutron yield varies from 10^8 to 10^{12} s^{-1}, which gives a usable flux density for irradiation ca. 20 – 100 times lower. Since neutron generators are rather compact, are not too expensive, and do not need very large shielding, they are used in many research and industrial laboratories. Cross sections, especially for light elements, often range from 10 mb to a few hundred millibarn for (n, p) and (n, α) reactions on light elements, and even up to 2 – 3 b for some (n, 2 n) reactions on heavy elements. With a neutron flux of 10^9 cm^{-2} s^{-1} typical detection limits of 10^{-4} to 10^{-6} g, and for a few elements (F, Si, P, Cu, Br, Ag, Sb) ca. 10^{-7}, can be reached. Disadvantages of FNAA with neutron generators are the many possibilities for producing the same indicator radionuclide from different elements and the large flux gradients.

The determination of oxygen by the reaction ^{16}O (n, p) ^{16}N is most important. The half-life of the indicator radionuclide ^{16}N is only 7.1 s, and it emits high-energy gamma rays ($>$ 6 MeV), which can be measured without the interference of other radionuclides. A detection limit of about 0.005 % can be obtained by using large samples (10 – 100 g).

Isotopic Neutron Sources [30]. Neutrons can also be produced from isotopic sources. They have the unique advantage of being small, portable, and completely reliable, and of having a constant neutron flux and low cost. There is no need for complicated controlling mechanisms, and the shielding can be very limited. Isotopic neutron sources offer interesting analytical perspectives for accurate determination of minor and major constituents.

Their operation is based on releasing neutrons from a target nucleus either by the action of an α particle or a γ quantum, or during spontaneous fission of a transuranic element. Beryllium is mostly used as target material. As an α source, historically ^{226}Ra was used; it has now been replaced by ^{227}Ac, ^{241}Am, or ^{238,239}Pu, and in the future possibly ^{224}Cm. As a γ source, ^{124}Sb is employed ($t_{1/2}$ = 60.2 d). The neutrons emitted have a broad energy spectrum that peaks at ca. 3 – 6 MeV. Paraffin wax is often used as a moderator. The neutron outputs range from 10^6 to 1.5×10^7 s^{-1} Ci^{-1}. In practice, sources of 10^8 neutrons/s are available commercially. The isotope ^{252}Cf disintegrates not only by alpha decay, but also by spontaneous fission, releasing several neutrons in the process. It has a number of advantages over (α, n) sources: the small size of the source, and the close similarity of the neutron spectrum to that obtained from the fission of uranium. The main disadvantage is obviously its relatively short effective half-life of 2.65 a.

Since the obtainable neutron flux density is generally less than 2×10^6 cm^{-2} s^{-1}, elements with cross sections $>$ 10 barn and half-lifes of less than 5 d have typically been determined. Examples are In, Dy, Rh, V, Ag, La, Eu, Mn, I, Br, Co, Ga, Cu, Ir, and Au. They are determined preferentially in matrices that do not activate very much so that nondestructive analyses are possible. Attainable precisions are better than 1 %, sometimes as good as 0.1 %.

2.3. Standardization

Standardization (i.e., calibration) is performed by Equation (10), which shows that m, the mass of the element to be determined, can be calculated from the measured activity. In practice, this is rarely done because although the values for θ, N_A, A_r, S, D, $C(\lambda)$, and t_m are known with good accuracy for all radionuclides considered, this is not the case for φ (φ_{th} and φ_{epi}), σ (σ_0 and I_0), and ε. Therefore a *relative or comparator method* is normally applied, although more recently absolute methods have also been used. The relative method is essentially very simple. Each sample is coirradiated with a standard containing known amounts of the elements to be determined. When samples and standards are counted under identical conditions, all these factors may be assumed to cancel out, leaving only the following very simple equation:

$$\frac{m_x}{m_{st}} = \frac{A_x}{A_{st}} \cdot \frac{D_{st}}{D_x} = \frac{A_x}{A_{st}} \frac{\left(e^{-\lambda t_d}\right)_{st}}{\left(e^{-\lambda t_d}\right)_x} \tag{14}$$

where the subscripts x and st denote the unknown and the standard, respectively. This simple relationship is valid only when all other factors are the same or are held constant. For short-lived radionuclides, the sample and the standard are irradiated sequentially. Since the neutron flux cannot be relied on to remain entirely constant with time, the activities measured must be corrected for any flux variation by coirradiating flux monitors. Second, the sample and the standard, even when irradiated simultaneously, are not always exposed to the same neutron flux because vertical or lateral flux gradients may occur. This can be corrected by means of flux-gradient monitors. The standards containing known amounts of the elements to be determined can be mixtures of pure elements or compounds of known stoichiometry. Since in NAA the samples to be irradiated are mostly solid, the mixed standard solutions are spotted on clean paper filters or ultrapure graphite, dried, and finally pressed into pellets with similar geometry to the sample. To avoid problems associated with the preparation of these standards, reference materials or other "in-house standard materials" are often used in NAA.

k_0-Method. The k_0-method has been developed to overcome the labor-intensive and time-consuming work of preparing such multielement standards when routine multielement or panoramic analyses are required. Single comparators have been used for a long time, but they are applicable only to constant experimental parameters of activation and counting. Therefore, the k_0-method is being implemented increasingly in NAA laboratories. It is intended to be an absolute technique in which uncertain nuclear data are replaced by a compound nuclear constant, the k_0-factor, which has been determined experimentally for each radionuclide. This k_0 is given by

$$k_0 = \frac{A_r^* \theta \sigma_0 \gamma}{A_r \theta^* \sigma_0^* \gamma^*} \tag{15}$$

where γ is the relative intensity of the gamma radiation in the decay scheme of the radionuclide and * denotes the comparator. For determination of the k_0-factor, gold was used as comparator, which was coirradiated as dilute Au – Al wire. This Au – Al wire should be coirradiated with each sample. It can, however, be converted to any comparator that is found suitable for coirradiation and that has been coirradiated with gold before. To determine the epithermal-to-thermal flux ratio, coirradiation of a zirconium monitor is suitable. The k_0-factors for 68 elements and their relevant gamma lines of 135 analytically interesting radionuclides have been determined and published by De Corte and coworkers. For nearly all of them, the uncertainly is < 2 %. An overview is given in Table 2 [31–33].

Table 2. Elements and their analyte radionuclides that can be determined by k_0 INAA

F	^{20}F
Na	^{24}Na
Mg	^{27}Mg
Al	^{28}Al
Si	^{31}Si
S	^{37}S
Cl	^{38}Cl
K	^{42}K
Ca	^{47}Ca
	^{47}Ca $\rightarrow$ ^{47}Sc
	^{49}Ca
Sc	^{46}Sc
Ti	^{51}Ti
V	^{52}V
Cr	^{51}Cr
Mn	^{56}Mn
Fe	^{59}Fe
Co	^{60}Co
Ni	^{65}Ni

Table 2. (Continued)

Cu	^{64}Cu
	^{66}Cu
Zn	^{65}Zn
	^{69m}Zn
	^{71}Zn
Ga	^{72}Ga
Ge	^{75m}Ge
	^{75m}Ge $\rightarrow$ ^{75}Ge
	^{77m}Ge
	^{77m}Ge $\rightarrow$ ^{77}Ge
As	^{76}As
Se	^{75}Se
	^{77m}Se
Br	^{80m}Br
	^{80m}Br $\rightarrow$ ^{80}Br
	^{82m}Br $\rightarrow$ ^{82}Br
Rb	^{86m}Rb $\rightarrow$ ^{86}Rb
	^{88}Rb
Sr	^{85m}Sr
	^{85m}Sr $\rightarrow$ ^{85}Sr
	^{87m}Sr
Y	^{90m}Y
Zr	^{95}Zr
	^{95}Zr $\rightarrow$ ^{95m}Nb
	^{95}Zr $\rightarrow$ ^{95m}Nb $\rightarrow$ ^{95}Nb
	^{97}Zr
	^{97}Zr $\rightarrow$ ^{97m}Nb
	^{97}Zr $\rightarrow$ ^{97m}Nb $\rightarrow$ ^{97}Nb
Nb	^{94m}Nb
Mo	^{99}Mo
	^{99}Mo $\rightarrow$ ^{99m}Tc
Mo	^{101}Mo
	^{101}Mo $\rightarrow$ ^{101}Tc
Ru	^{97}Ru
	^{103}Ru
	^{105}Ru
	^{105}Ru $\rightarrow$ ^{105m}Ru
	^{105}Ru $\rightarrow$ ^{105m}Ru $\rightarrow$ ^{105}Ru
Rh	^{104m}Rh
	^{104m}Rh $\rightarrow$ ^{104}Rh
Pd	^{109m}Pd
	^{109m}Pd $\rightarrow$ ^{109}Pd
	^{111m}Pd
Ag	^{108}Ag
	^{110m}Ag
	^{110}Ag
Cd	^{115}Cd
	^{115}Cd $\rightarrow$ ^{115m}In
In	^{114m}In
	^{116m}In
Sn	^{113m}Sn $\rightarrow$ ^{113}Sn $\rightarrow$ ^{113m}In
	^{117m}Sn
	^{123m}Sn
	^{125m}Sn
	^{125m}Sn $\rightarrow$ ^{125}Sn
	^{125m}Sn $\rightarrow$ ^{125}Sn $\rightarrow$ ^{125}Sb
Sb	^{122m}Sb $\rightarrow$ ^{122}Sb
	^{124m}Sb $\rightarrow$ ^{124m1}Sb $\rightarrow$ ^{124}Sb
	^{124m2}Sb $\rightarrow$ ^{124m1}Sb
Te	^{131m}Te $\rightarrow$ ^{131}Te $\rightarrow$ ^{131}I
I	^{128}I
Cs	^{134m}Cs
	^{134m}Cs $\rightarrow$ ^{134}Cs

Table 2. (Continued)

Ba	^{131m}Ba $\rightarrow$ ^{131}Ba
	^{133m}Ba
	^{139}Ba
La	^{140}La
Ce	^{141}Ce
	^{143}Ce
Pr	^{142m}Pr $\rightarrow$ ^{142}Pr
	^{109m}Pd $\rightarrow$ ^{109}Pd $\rightarrow$ ^{109m}Ag
Nd	^{147}Nd
	^{149}Nd
	^{149}Nd $\rightarrow$ ^{149}Pm
	^{151}Nd
	^{151}Nd $\rightarrow$ ^{151}Pm
Sm	^{153}Sm
	^{155}Sm
Eu	^{152m}Eu
	^{152m}Eu $\rightarrow$ ^{152}Eu
	^{154m}Eu $\rightarrow$ ^{154}Eu
Gd	^{153}Gd
	^{159}Gd
	^{161}Gd
Tb	^{160}Tb
Dy	^{165m}Dy
	^{165m}Dy $\rightarrow$ ^{165}Dy
Ho	^{166}Ho
Er	^{171}Er
Tm	^{170}Tm
Yb	^{169}Yb
	^{175}Yb
	^{177}Yb
Lu	^{177}Lu
	^{176m}Lu
Hf	^{175}Hf
	^{179m}Hf
	^{180m}Hf
	^{181}Hf
Ta	^{182m}Ta $\rightarrow$ ^{182}Ta
W	^{187}W
Re	^{186}Re
	^{188m}Re
	^{188m}Re $\rightarrow$ ^{188}Re
Os	^{185}Os
	^{191m}Os $\rightarrow$ ^{191}Os
	^{193}Os
Ir	^{194}Ir
Pt	^{199m}Pt $\rightarrow$ ^{199}Pt $\rightarrow$ ^{199}Au
Au	^{198}Au
Hg	^{197m}Hg
	^{203}Hg
	^{205}Hg
Th	^{233}Th $\rightarrow$ ^{233}Pa
U	^{239}U
	^{239}U $\rightarrow$ ^{239}Np

2.4. Sources of Errors

Besides typical analytical errors in sample preparation (contamination, inhomogeneity, inaccurate standards) and errors during counting

(pulse pileup, instability, spectrometric interference, electronic failure, etc.), some systematic bias may also occur during the irradiation step, such as neutron shielding and nuclear interference. By taking the proper measures, these can be avoided or corrected for. *Flux gradients* can be determined experimentally and corrected for. For accurate work, vertical and horizontal gradients should be measured for the different types of flux (e.g., thermal, epithermal, and fast). When relatively large samples or standards with large absorption cross sections or resonance integrals are irradiated, nonnegligible *neutron shadowing* or *self-shielding* may also occur.

In addition, *flux hardening* (preferential absorption of low-energy neutrons) and *self-moderation* (unmoderated neutrons are further moderated inside the sample) may occur. Since the correction is difficult to apply for samples with mixed composition and irregular shape, the effect is often avoided by using as small a sample and a standard as possible or by diluting the sample with a material having a low absorption cross section (graphite, cellulose). An internal standard can also be applied.

In thermal neutron activation, the determination of an element ${}^{M}_{Z}\mathrm{E}$ is often based on an (n,γ) reaction producing the radionuclide ${}^{M+1}_{Z}\mathrm{E}$. An (n,p) reaction on nuclide ${}^{M+1}_{Z+1}\mathrm{E}'$ or an (n,α) reaction on nuclide ${}^{M+3}_{Z+2}\mathrm{E}''$ will produce the same indicator radionuclide, resulting in positive errors. This *nuclear interference* depends on the concentration ratios of the elements in the sample, the fast-to-thermal flux ratio, and the cross sections involved, and can be estimated by a simple calculation. The cross sections for (n,p) and (n,α) reactions are usually much smaller than the (n,γ) cross sections. Well-known examples are ${}^{27}\mathrm{Al}\,(\mathrm{n},\gamma)\,{}^{28}\mathrm{Al}$ interference by ${}^{28}\mathrm{Si}\,(\mathrm{n},\mathrm{p})\,{}^{28}\mathrm{Al}$, and ${}^{55}\mathrm{Mn}\,(\mathrm{n},\gamma)\,{}^{56}\mathrm{Mn}$ interference by ${}^{56}\mathrm{Fe}\,(\mathrm{n},\mathrm{p})\,{}^{56}\mathrm{Mn}$.

If the interference is not too important it can be corrected for. If uranium or other fissionable material is present in the sample, the fission products with high fission yields (Sr, Mo, Zr, Ce, Ba) can induce important positive errors. The interference can easily be estimated when the uranium content is known. In FNAA, *secondary interference reactions* may occur when fast neutrons interact with other elements and produce particles that induce a nuclear reaction that forms the same indicator nuclide. These particles are usually protons ejected by fast neutrons from a matrix with a high hydrogen content. Examples are:

${}^{19}\mathrm{F}\,(\mathrm{n},2\,\mathrm{n})\,{}^{18}\mathrm{F}$ interference by: ${}^{17}\mathrm{O}\,(\mathrm{p},\gamma)\,{}^{18}\mathrm{F}$ and ${}^{18}\mathrm{O}\,(\mathrm{p},\mathrm{n})\,{}^{18}\mathrm{F}$

${}^{14}\mathrm{N}\,(\mathrm{n},2\,\mathrm{n})\,{}^{13}\mathrm{N}$ interference by: ${}^{13}\mathrm{C}\,(\mathrm{p},\mathrm{n})\,{}^{13}\mathrm{N}$

2.5. Measuring Equipment

For identification and quantification of the radionuclides produced during the activation step a radiation detector is needed. The major way to select a specific radionuclide is of course by *radiochemical separation.* At the end of the irradiation a nonradioactive carrier of the element to be measured can be added prior to separation. A second parameter used for selection is the *half-life* of the radionuclide. Selection on the basis of the *energy of the emitted radiation* is, however, the most powerful technique. Obviously, detection of the γ- and X-rays is much more suitable for identification and selective for quantification purposes than β-ray detection. Alpha emitters are almost never produced in activation analysis.

Gas-filled detectors, such as ionization chambers, proportional counters, and Geiger–Müller counters (→ Radionuclides, Chap. 4.2.2.1.) are mainly sensitive to β-radiation. They rarely allow any selection on the basis of energy. Gamma counting with a selection for energy is called gamma spectrometry and is performed by means of *scintillation* or *semiconductor detectors* (→ Radionuclides, Chap. 4.2.2.2., → Radionuclides, Chap. 4.3.1.). Currently the gamma spectrometric measurements in neutron activation analysis are only performed with semiconductor detectors. Scintillation counters will therefore be discussed only briefly.

Interaction of γ- or X-rays with matter leads to three effects (→ Radionuclides, Chap. 1.5.1., → Radionuclides, Chap. 4.1., → Radionuclides, Chap. 4.2.): (1) A photoelectric effect in which the entire energy is absorbed by the detector and the result is a photopeak in the spectrum, with a pulse height proportional to the full energy of the gamma radiation. (2) Compton scattering results in a partial absorption of the γ-ray en-

ergy. In the spectrum a continuous background is seen, which can interfere with the measurement of other lower-energy γ- or X-rays. (3) Pair production may result in the escape of one or two annihilation photons, each with an energy of 0.511 MeV. The *efficiency* ε of a detector refers to the ratio of the number of radiations actually detected to the number emitted by the source. In addition a photopeak-efficiency ε_p may be defined as the number of counts recorded under the photopeak. The *energy resolution R* is the ability of the detector to discriminate between two γ- or X-rays of different energy. This resolution is defined as the full width ΔE of the peak at half-maximum (FWHM) (→ Radionuclides, Chap. 4.2.2.3.). For scintillation detectors, it is expressed mostly as a percentage of the energy E corresponding to the centroid of the peak. For semiconductor detectors, it is expressed in energy units (kiloelectronvolts or electronvolts). Finally, the *dead time* τ of the detector is the amount of time required before the detector can recover from an incoming radiation and respond to the next event. In gamma spectrometry the main cause of dead time is not the detector itself but the electronic equipment and, more specifically, the amplifier (i.e. pulse pile up) and the multichannel analyzer (MCA).

Scintillation Detectors. In scintillation detecors the incoming radiation interacts with the material by ionization or excitation. The excited atoms or molecules undergo deexcitation by the emission of a photon of light. This light passes through the optically transparent scintillator material and is guided through a light pipe to a photomultiplier, which converts the light to an electrical signal. The signal is then amplified and fed to a counting or storage system. This can be a single-channel analyzer (SCA) or a multichannel analyzer (MCA). The detection efficiency is excellent. The major drawback of NaI(Tl) scintillators is their poor energy resolution, which is typically 7 % for the 661.6-keV line of ^{137}Cs (Fig. 2). Dead times are important at high count rates. More selective counting can be performed with NaI(Tl) detectors when coincidence or anticoincidence spectrometry is applied by using two or even more detectors.

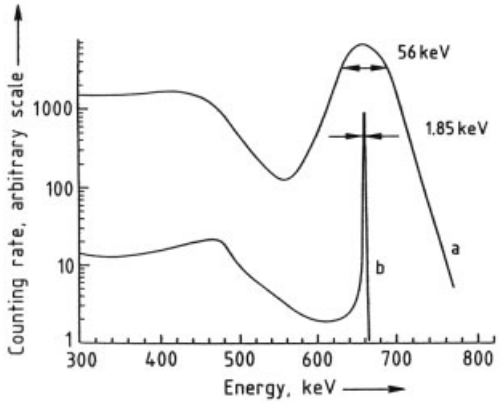

Figure 2. Comparison of gamma spectrum of ^{137}Cs (gamma energy 661.6 keV) obtained with 7.5×7.5-cm NaI(Tl) scintillation detector (a) and with a 50-cm^3 Ge(Li) semiconductor detector (b)

In charged-particle activation analysis (Chapter 4) pure positron emitters have to be measured by their annihilation photons, emitted simultaneously in a 180° geometry. The sample is then introduced between two NaI detectors, coupled to an SCA to select the photopeak of the 511-keV annihilation photons. The fast SCA output is fed into a coincidence circuit, which selects only coincident signals (typically within 40 ns). The poor energy resolution of the NaI detector is no drawback, while the high detection efficiency (as compared with semiconductor detecors) is an advantage. The NaI detectors are much cheaper than semiconductor detectors (for the same detection efficiency) and do not require liquid nitrogen cooling. The signal-to-noise ratio is much higher for the NaI coincidence set-up, than for a Ge spectrometer, mainly because the noise (background) is much lower.

Semiconductor Detectors. These have a $p-i-n$ diode structure in which the intrinsic region is formed by depletion of charge carriers by applying a reverse bias across the diode. The energy of an electron in a semiconductor material is confined to the conduction and the valence bands, both states separated by forbidden energies. The passage of radiation through the semiconductor may inject sufficient energy to raise an electron from the valence band to the conduction band, thus creating an electron–hole pair. In order for the material to be used as a radiation detector, this electrical charge must be collected. Therefore, several thousand volts

are applied to the detector. Owing to the small amount of energy needed to create an electron–hole pair, a large number of pairs (>300 000 pairs/MeV) is created, with small relative statistical fluctuations. Since the energy resolution of the detector depends on this fluctuation, which can be predicted by Poisson statistics, much better energy resolution is obtained than with scintillation counters. For a highly efficient detector, a large, nearly intrinsic or depletion region must be present. Therefore, the $p-n$ junction is created within a single p or n crystal. Ways to achieve this are the surface barrier detector (used only for detection of charged particles because the depletion region is too small), the lithium-drifted detector [Ge (Li) and Si(Li)], and at present the high-purity or intrinsic germanium detector (HP-Ge). *Ge (Li) detectors* must be cooled constantly at liquid nitrogen temperature, they are still in use but are no longer available. They are more efficient for γ-ray detection than Si (Li) detectors, owing to their higher atomic number Z. *Si (Li) detectors* are sufficiently efficient for detection of X-rays and low-energy γ-rays and must be cooled only during operation to ensure optimum energy resolution. Typical resolutions of Si(Li) detectors are 140–160 eV for the 5.9-keV Mn K line from a ^{55}Fe source.

HP-Ge detectors have the advantage that they can be stored at room temperature and are cooled only during operation to reduce the problem of thermal excitation of electrons. Coaxial p-type detectors and n-type detectors are available. In the *n-type detector* the inner contact is made by diffused lithium, while the outer contact is achieved by ion implantation, which ensures a very thin entrance window (0.3 µm). Thin *planar germanium detectors* (5–20 mm) can also be used to measure low-energy radiation and have the advantage of not stopping the more energetic γ-rays, thereby reducing interference effects such as Compton radiation. *Well-type detectors* are also available with well sizes up to 15 mm in diameter and a depth of 40 mm. For cooling, a cold finger and a large cryostat are needed, which makes semiconductor detectors bulky and difficult to maneuver.

Germanium detectors are characterized by three parameters: resolution, peak-to-Compton ratio, and efficiency. The *resolution* is typically given for the 1332-keV ^{60}Co line and varies from 1.8 keV for the very best to 2.3 keV for the very large detectors. The *peak-to-Compton ratio* is measured as the ratio of the number of counts in the 1332-keV peak to the number of counts in a region of the Compton continuum. Values vary from 30 to 90 for the most expensive model. The efficiency is expressed as a relative efficiency compared with the 7.5×7.5-cm NaI(Tl) scintillation detector. Relative efficiencies of HP-Ge detectors vary from 10 % up to 150 %. The dead time of semiconductor detectors is low, so the count rate is limited largely by the electronic circuit.

Gamma spectrometry is performed mainly by using MCAs, and the relation between the obtained pulse height or corresponding channel number and gamma energy must be established. When an absolute method of standardization is applied, the peak detection efficiency ε_{p} of the detector with surrounding material must also be measured.

For *peak area calculation*, two procedures can be applied in principle: integration or fitting. In *integration*, the number of counts recorded in the channels in which about 90 % of the peak is located is corrected for Compton contribution from other peaks by subtracting the content of an equal number of channels before and after the peak. The method fails, of course, when other peaks are located in the immediate neighborhood or in the case of superimposed peaks. *Fitting procedures* assume a defined shape, either Gaussian or with non-Gaussian distortions (determined from the shape of a well-defined peak in the spectrum).

The overall procedure for the evaluation of γ-ray spectra as part of activation analysis consists of several steps as defined by Erdtmann und Petri [34]:

1) Search for the peaks
2) Calculation of γ-ray energy
3) Calculation of peak areas
4) Calculation of γ-ray emission rates
5) Identification of radionuclides
6) Calculation of decay rates
7) Calculation of amounts of elements

The amount of data to be handled during this evaluation procedure requires the use of computers. At present, the multichannel analyzer not only carries out pulse height analysis but also does most of the data processing. The manufac-

turers of computer-based MCAs usually provide built-in or software programs for the different aforementioned steps.

Obviously the use of a personal computer is an important step toward automation of the analysis. The computer can control the automatic sample changer, the timing of the analysis, counting, storing of the spectra, and finally, data processing.

2.6. Procedures for Activation Analysis

Sample Preparation. As in any other analytical technique, the quality of the final results depends strongly on the care taken during sampling (→ Sampling). Since sample sizes can often be small (2 mg up to 20 g), owing to the high sensitivity of NAA, a well-homogenized material is required. In a number of cases, some preirradiation treatment will be necessary, such as: (1) the removal of water by drying, freeze drying, or ashing for biological materials, or evaporation for liquid samples, to avoid pressure buildup due to evaporation or radiolysis during irradiation; (2) the removal of surface contamination; or (3) perhaps even preconcentration. All of these treatments should be performed in clean environments.

For the irradiation itself, the sample is packed in polyethylene vials for total neutron doses of less than 10^{17} cm^{-2} and in high-purity quartz ampoules for longer irradiations at higher fluxes. The packed samples are transferred to the reactor in containers (rabbits) of aluminum, polyamide, graphite, or high-pressure polyehtylene for low-dose irradiation.

The major advantage of activation analysis, however, is that after irradiation, contamination with nonradioactive material does not introduce more errors. On the contrary, when radiochemical separations are required, an amount of nonradioactive carrier of the element to be separated is added, after irradiation.

Instrumental Neutron Activation Analysis (INAA). A purely instrumental approach is applicable for the determination of many elements (i.e., after irradiation, the sample can be unpacked and the induced radioactivity measured without further treatment). Such a technique has a number of advantages. After decay of the induced activities, the sample is essentially unchanged (nondestructive activation analysis), and is therefore available, in principle, for further investigation. In addition, short-lived radionuclides can be used for the analysis. Tedious chemical separations are not required, which implies reduction of analysis time and avoidance of errors associated with separation. INAA becomes feasible when the activity induced in the sample matrix is not prohibitively high and no single major activity is produced that overshadows that of the other radionuclides. Of course, the differences in half-lifes of various radionuclides can also be exploited by limiting the saturation factor S of interfering long-lived activities or by allowing the short-lived matrix activity to decay before counting. Virtually all instrumental multielement activation analysis is based on high-resolution gamma spectrometry using semiconductor detectors after thermal or epithermal activation of the sample in a reactor or a neutron generator.

INAA is favored especially for many organic, biological, and environmental samples, because the elements O, H, C, N, S, and P are not activated or do not produce gamma emitters. The most abundant minor or trace elements such as Si, Fe, Al, Ca, Mg, and Cl have relatively small cross sections or short half-lifes, the nuclides with the highest activity often being ^{24}Na($t_{1/2}$ = 15.0 h) or ^{82}Br ($t_{1/2}$ = 36 h). Also, a number of metallic matrices can be analyzed purely instrumentally with thermal neutrons. Examples of these metals are Be, Mg, Al, Ca, Ti, V, Ni, Y, Zr, Nb, Rh, Pb, and Bi, some of them after decay of the short-lived matrix activity.

In multielement analysis, choosing the optimal irradiation, cooling, and counting conditions for each element is obviously not feasible, so a compromise must be found. The balance between working up a reasonable number of samples and the desired quality of the analytical results will dictate the mode of the irradiation–counting scheme. A typical scheme has been described for atmospheric aerosols (Fig. 3) [35]. For the *short-lived radionuclides*, irradiation of a 2- to 25-mg sample for a few minutes, together with a flux monitor, is employed. Typically, 10 – 15 elements are determined from this irradiation. *Longer irradiation* of a number of samples, together with a multielement standard

or a flux monitor, is followed by two to three counts after decay times varying from 1 to 30 d. In total, the scheme allows detection of up to 45 elements in favorable cases although, more realistically, about 30 elements can be determined with precisions varying from 2 to 10 %. Shorter and less complicated schemes can be applied at the expense of precision and the number of elements determined. The selectivity for the determination of some elements can be enhanced by applying epithermal, pulsed, or cyclic activation and by counting with low-energy photon detectors or anticoincidence spectrometers.

Radiochemical Neutron Activation Analysis. The ultimate sensitivity and the lowest detection limits are generally obtained when radiochemical separations are performed after irradiation. Chemical treatment of the samples and the standards is aimed at separating the radionuclides formed. The separated fractions are then used in a chemical and physical form suitable for counting. Because the actual amount of radionuclides is very small, a chemical carrier is added before the decomposition step to ensure chemical equilibration with the radionuclide. The next stage is chemical separation using classical methods such as precipitation, distillation, solvent extraction, chromatography, ion exchange, or electrodeposition. Since the amount of carrier added is known and greatly exceeds the amount of the element originally present, the yield of the chemical separation can be determined. Correction of the activity measured is possible.

If clean separation of one radionuclide is achieved, scintillation counting or even β-counting can be performed in principle. Nevertheless, high resolution HP-Ge spectrometry is applied to check the absence of any interfering radionuclide. For most types of samples, such as geological, environmental, and biological materials, typical group separation schemes have been developed and can be found in the literature. For biological samples, the emphasis is on the removal of major activities from ^{32}P, ^{24}Na, and ^{82}Br. In geological samples, rare earth elements or noble metals are often determined after separation. Sometimes carriers or so-called scavengers such as hydroxides of Fe^{3+}, Mn^{4+}, and Sn^{4+} are used to remove interfering activities or to separate the analyte radionuclides.

2.7. Sensitivity and Detection Limits

The smallest amount of an element that can be determined depends on the specific activity produced and the minimum activity measurable with sufficient precision. From Equation (10) the activity produced and measured per gram of an element can be calculated. However, the minimum activity that can be measured depends not only on the decay properties of the radionuclide and on the counting equipment, but also on the background of the detector or the Compton continuum on which the photopeak must be detected in gamma spectrometry. Thus the detection limit of an element for specific irradiation – counting conditions is not immutable since it depends on the presence of other radionuclides.

For measurement results giving a Poisson distribution, as in radioactive decay, the equation used to calculate the standard deviation σ is simply the square root of the number of counts recorded when the number is large. The net signal S is obtained as the difference between the measured signal $(S+B)$ minus the background B (e.g., Compton continuum in gamma spectrometry). Thus the statistically derived standard deviation is given as

$$\sigma_s = (S+B+B)^{1/2} \tag{16}$$

The limit of detection depends on whether the background is well known or is not well known as in gamma spectrometry. Curie [36] defines three specific quality levels for a measurement (→ Chemometrics):

1) *Decision limit* L_c is the net signal above which an "observed" signal can be accepted (95 % probability) as detected; L_c equals $1.64\,\sigma_B$ or $2.33\,\sigma_B$; for well-known or not well-known background, respectively
2) *Detection limit* L_D is the "true" net signal that a priori may be expected to be detected; L_D equals $3.29\,\sigma_B$ or $4.65\,\sigma_B$
3) *Determination limit* L_Q is the net signal on which quantitive determination with a precision of 10 % can be performed; L_Q equals $10\,\sigma_B$ or $14.1\,\sigma_B$

Because the sources of errors in NAA are generally known, they can be kept under control. When good laboratory practice is applied,

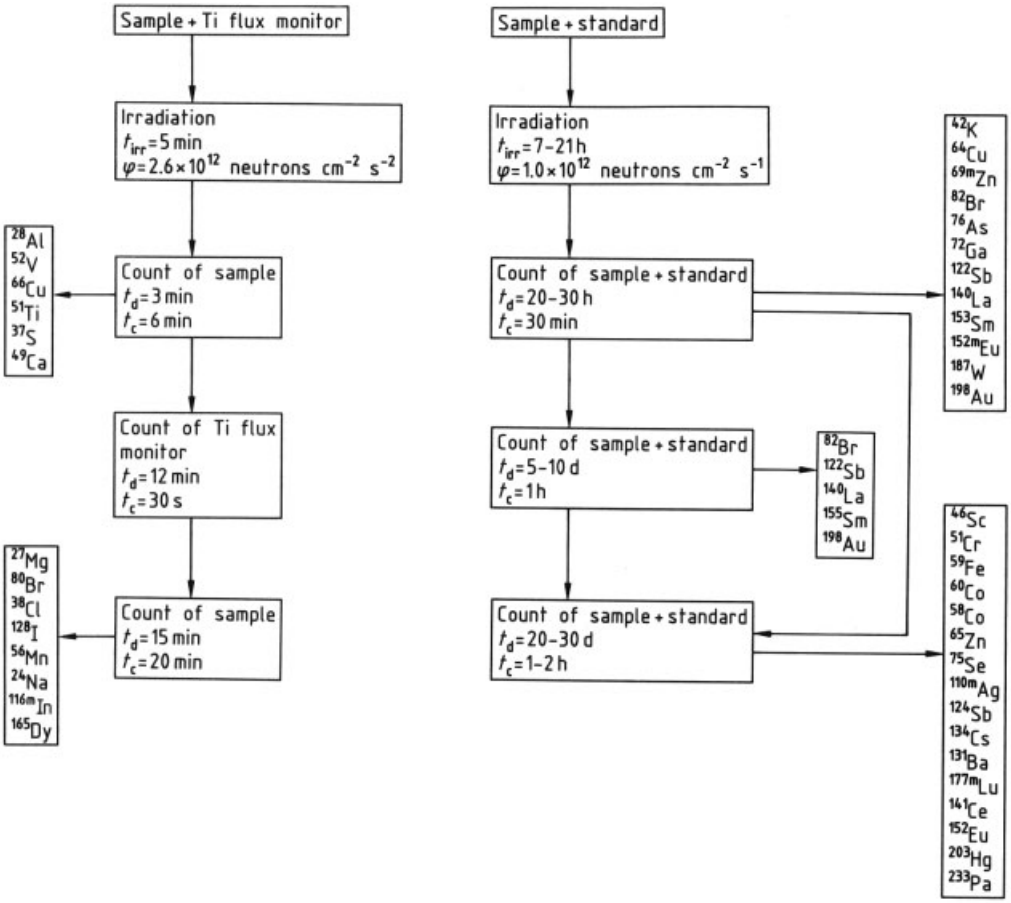

Figure 3. Irradiation – counting scheme for INAA of air particulates collected on filter paper [35]
φ = Thermal neutron flux; t_{irr} = Irradiation time; t_{d} = Decay time; t_{c} = Counting time

Table 3. Detection limits (units: 10^{-12} g) for INAA under ideal conditions of gamma spectrometry [135 cm^3 Ge(Li) detector, lead shielded] [a] [35]

Na	0.1	Ge	30	Sb	0.02	Tm	2
Si	900 000	As	0.03	Te	3	Y	0.06
Cl*	10 000	Se	3	I*	1500	Lu	0.03
Ar*	30	Br	0.05	Cs	0.5	Hf	0.4
K	5	Rb	8	Ba	40	Ta	0.4
Ca	1 500	Sr*	50	La	0.02	W	0.04
Sc	0.04	Y	1 500	Ce	1.5	Re	0.015
Ti	600	Zr	15	Pr	0.5	Os	0.8
Cr	3	Nb	15 000	Nd	5	Ir	0.006
Mn*	0.4	Mo	0.5	Sm	0.002	Pt	1
Fe	300	Ru	2	Eu	0.003	Au	0.001
Co	0.8	Pd	40	Gd	0.9	Hg	1
Ni	200	Ag	2	Tb	0.15	Tl	3 000
Cu	2	Cd	0.6	Dy*	2	Pb	30 000
Zn	6	In*	4	Ho	0.3	Th	0.2
Ga	0.06	Sn	100	Er	0.5	U	0.06

[a] For detection of radionuclides with $t_{1/2} > 1$ h, irradiation: 48 h at 2×10^{14} cm^{-2} s^{-1} (cadmium ratio =11, i.e., ratio obtained by irradiation without and with cadmium cover); decay time 10 h; counting time 64 h. For detection of radionuclides with $t_{1/2} < 1$ h (marked with *), irradiation: 2 h at 3×10^{11} cm^{-2} s^{-1}; decay time 1 h; counting time 3 h.

the reproducibility and precision can be nearly as good as the counting statistics, in favorable cases as low as 1 – 2 %. Detection limits obtainable by multielement INAA were determined by VERHEYKE and RATH [37] for compromise irradiation – counting conditions but for ideal gamma spectrometry conditions, namely, calculated from a background (Compton continuum) spectrum obtained from very pure irradiated graphite (Table 3). For several elements (Ar, Mn, Sr, Y, Dy) a better sensitivity can be calculated for a shorter decay time. For some elements (F, Mg, Al, V, and Rh) giving rise to radionuclides with half-lifes of less than 10 min no limits are given, although they can be determined with low detection limits after a short irradiation and a fast count. Elements such as S, P, and Bi produce only β-emitters and are therefore absent from the table.

3. Photon Activation Analysis [23],[24]

Photons – often obtained as bremsstrahlung produced in an electron accelerator – with energies varying from 10 to 45 MeV can overcome the threshold of (γ, n), (γ, p), $(\gamma, 2\,\mathrm{n})$, and (γ, α) reactions. Inelastic scattering with low-energy photons leads to short-lived metastable isotopes by (γ, γ') reaction. The light elements C, N, O, and F are determined by (γ, n) reactions, but the reaction leaves the nuclides proton rich, so that a positron emitter is formed. Radiochemical separation or decay curve analysis is needed. Detection limits of $< 1\,\mu\mathrm{g}$ are obtainable. *Instrumental photon activation analysis* (*IPAA*) is possible for heavier elements in the presence of an excess of low atomic number material, because when a bremsstrahlung spectrum is used, the yields of the reactions increase monotonically with increasing atomic number. Typical detection limits range from 0.01 to 1 µg. IPAA has been applied to the analysis of many environmental samples. Typically, about 15 elements can be determined precisely. Photon activation analysis was considered a valuable complement to NAA for the determination of light elements and some heavy elements before charged-particle activation analysis (CPAA) was developed. The major drawback is the limited availability of high-energy electron beams with sufficient intensity.

4. Charged-Particle Activation Analysis

Although NAA and PAA have a few characteristics in common with charged-particle activation analysis (CPAA), there are fundamental differences, owing to the electric charge of the bombarding particle.

In NAA there are endo-ergic reactions [e.g., (n, 2n) or (n, α) in FNAA] as well as exo-ergic reactions [e.g., (n, γ) in TNAA] and this is also the case for CPAA. However, owing to the Coulombic repulsion between a charged particle (CP) and the target nucleus, CPs always require a minimum energy, even for endo-ergic reactions. For CPAA, mono-energetic particles with energies between 5 and 50 MeV are utilized, leading to more nuclear reactions, and more possible nuclear interferences. For a proton energy below 10 MeV, e.g, in general (p, n) and (p, α) reactions are possible. Below 20 MeV also (p, 2n) (p, d) (p, t), (p, ^{3}He), (p, αn) ... reactions have to be considered. For TNAA only (n, γ) reactions have to be considered, nuclear interference from a (n, p) reaction is fairly infrequent.

CPs are slowed down in matter and are all stopped at depth, called the range, being 0.1 to 1 mm. As a consequence the beam intensity is hardly reduced, in contrast to neutron shadowing. Moreover, the induced activity is not related to the mass of the analyte, but to its mass fraction. So standardization (calibration) is very different from NAA.

The *stopping power (S)* is defined as the energy loss (– dE) per unit pathlength (dl) for a particular CP in a particular target:

$$S = \frac{-\mathrm{d}E}{\mathrm{d}l} \tag{17}$$

It depends on the CP energy. As a result the CPs are stopped at a depth, called *range R*. The range is characteristic for a CP, a target, and the CP energy, being mono-energetic.

The equivalent of $\varphi\,\sigma\,N_\mathrm{A}$ in Equations (3) – (10), i.e., the reaction rate in s^{-1}, can be written for an infinitively thin target

$$I\sigma n\mathrm{d}l \tag{18}$$

with I = beam intensity (number of CPs per s), n = number of target nuclides per cm^3, and dl =

thickness (cm) of the infinitively thin target. For a target thicker than the range R, Equation (18) is integrated

$$In \int_0^R \sigma_l \mathrm{d}l = In \int_{E_\mathrm{i}}^0 \frac{\sigma_E \mathrm{d}E}{\mathrm{d}E/\mathrm{d}l} = In \int_{E_\mathrm{t}}^{E_\mathrm{i}} \frac{\sigma_E \mathrm{d}E}{S_E} \tag{19}$$

with E_i the incident CP energy. The sample is supposed to be homogenous. The CP beam intensity is nearly unchanged when penetrating the target, while the CP energy decreases, and consequently the reaction cross section. As the reaction cross section σ is only known as a function of the energy, the stopping power S is used. For a target for which the thickness D is less than the range R, Equation (19) can be rewritten

$$In \int_0^D \sigma_l \mathrm{d}l = In \int_{E_\mathrm{i}}^{E_0} \frac{\sigma_E \mathrm{d}E}{\mathrm{d}E/\mathrm{d}l} = In \int_{E_0}^{E_\mathrm{i}} \frac{\sigma_E \mathrm{d}E}{S_E} \tag{20}$$

with E_0 = outgoing energy. It is clear that the induced activity is proportional to the mass fraction ($\sim n$ in cm^{-3}) of the analyte, in contrast to NAA where it is proportional to the mass of the analyte ($\sim N^A$ dimensionless) in Equations (3) – (10). This difference in dimensions may be suprising for two different modes of activation analysis. However, it is related to (1) the analysed depth, being limited to the range (cm) and (2) the use of beam intensity (in s^{-1}) versus fluency rate (in $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$) in NAA.

As relative standardization is applied in CPAA, there is no need for absolute activity and beam intensity measurements, nor for absolute data of the reaction cross section σ and stopping power S. If the (relative) cross section is not accurately known, approximations have been developed. The stopping power of any elemental matter can be calculated accurately; for mixtures or compounds the major matrix should be known. If not, an internal standardization method has been developed.

CPAA has been applied successfully for the determination of (trace) elements in the bulk of solid samples, such as pure metals and alloys, semiconductors, geological and solid environmental materials. The method has proven its unique capabilities and outstanding performance in the determination of light elements such as B, C, N, and O. Activation analysis has the unique possibility of removing surface contamination after irradiation and prior to measurement. CPAA has been a milestone, e.g., in the assessment of the bulk oxygen concentration in aluminum.

Recently CPAA has been extended to surface characterization. If it is possible to determine a trace element A in a matrix B at the µg/g level in the bulk of a sample, i.e., in a 1 mm thick layer, then it is also possible to determine the thickness of a mono-elemental layer A on a substrate B (that does not contain A) down to 1 nm. Alternatively, one can determine the composition of such a layer if the thickness is known, and consequently the stoichiometry without any foreknowledge of the thickness.

The principles of CPAA for bulk analysis are summarized in [25] and [26]; for surface characterization in [27] and in [28]. Applications of CPAA for bulk analyses are reviewed in [38]; the feasibility study on surface characterization in [39].

5. Applications

5.1. High-Purity Materials

Since the 1960s, activation analysis has been used intensively for determining trace elements in high-purity materials. Analysis of semiconductor materials, as well as of high-purity metals, obtained by zone refining, and of materials used in nuclear technology, has been one of the major fields of interest in activation analysis. Impurities are known to affect the properties of these materials, even when present in concentrations so low that they can barely be detected by other methods.

INAA is especially suited for the multielement analysis of matrices that, on neutron irradiation, do not produce intense gamma radiation or produce only short-lived ones. Examples of these matrices are graphite, silicon, and aluminum, or their oxides. A typical irradiation counting scheme allows the determination of 55 elements at submicrogram-per-gram concentrations in wafers of silicon [37]. Also, other semiconductor or alloying materials such as Se, Ge, Sb, Te, and ultrapure metals such as Cu, Fe, Ti, Mo, Nb, Ni, Ga Zn, Sn, Zr, Bi, Pb, and noble

metals have been analyzed. Often a postirradiation separation of the matrix activity is required, or a radiochemical separation scheme must be applied based on ion exchange, extraction, or precipitation. Matrix activation can sometimes be avoided by irradiation in a well-thermalized neutron flux or under the cover of cadmium (ENAA).

Photon activation (*PAA*) is used for elements not activated by (n, γ) reaction, with the emphasis on carbon, nitrogen, and oxygen. Detection limits of the order of 20 µg can be obtained.

Even lower limits of detection can be obtained for determining the light elements in semiconductor materials or ultrapure metals by *CPAA*.

The major advantage of activation analysis over nearly all other techniques of analysis – namely, the absence of contamination problems after irradiation – is extremely valuable in this field of application. Preirradiation treatments should be carried out in clean environments, and to further avoid surface contamination, postirradiation washing or etching is highly recommended. Assessing the impurity levels of these typical industrial products by activation analysis is often used in the quality control of manufacturing companies and has therefore been entirely automated and computerized by using a single comparator for calibration.

5.2. Environmental Materials

Neutron activation analysis is one of the major techniques for the determination of many minor and trace elements in a large variety of solid environmental and pollution samples, such as atmospheric aerosols, particulate emissions, fly ash, coal, incineration ash, and sewage sludge. *Instrumental neutron activation analysis* of total, inhalable, or respirable airborne particulate matter collected on a cellulose or membrane filter, or in a cascade impactor on some organic substrate, allows the determination of up to 45 elements by an irradiation – counting scheme similar to the one given in Figure 3. *Radiochemical NAA* is applied only when extremely low limits of determination are required. *Instrumental photon activation analysis* is also complementary to INAA.

The problem in charged-particle activation analysis related to a matrix of unknown composition and degradation (votalization) of the organic matrix during irradiation, has been solved by the development of an internal standardization method.

When applied to water samples, NAA suffers from some severe drawbacks compared with other multielement techniques. Irradiation of water results in radiolysis. Evaporation, freeze drying, adsorption on charcoal, and ion exchange can be used to remove water. A final problem is the interference of intense radiation induced in matrix elements such as Na, K, Cl, Br, and P. Saline water in particular, yields difficult samples for INAA. The preconcentration step may include separation from these elements. A wealth of literature exists describing these environmental applications [40], [41].

5.3. Biological Materials

Activation analysis, particularly with neutrons, is a very effective method for elemental trace analysis of biological materials and plays an important role in studies of trace elements related to health. Because of the limited sensitivity of classical analytical methods, the trace and ultratrace elements that play an important role in biological systems can rarely be determined by the usual methods. Radiochemical NAA has contributed significantly to establishing the normal levels of these elements in blood (total blood, serum, packed cells, and erythrocytes), liver, kidney, lung, muscles, bones, teeth, nerves, and hair and in studies involving disease [42]. It is the method of choice for the determination of Sb, As, Cs, Co, Mn, Mo, Rb, Se, Ag, and V, and is very reliable for Br, Cr, Cu, Hg, I, and Zn, but it lacks sensitivity for Cd, Ni, and Sn [43]. Since radiochemical separations are generally required, a number of group separations using ion exchange or distillation procedures have been developed to remove at least the interfering activities from P, Na, and Br. For a limited number of elements (Co, Cs, Fe, Rb, Se, and Zn), INAA is possible only after a long decay period [44]. Activation analysis offers many advantages for the analysis of foods and plants. INAA of vegetables, meat, fish, and poultry allows the detection of about 20 trace elements [43].

In the last decade speciation (the determination of the species of an element, i.e., its oxidation state or its compounds) has become more important than trace element determination. Speciation analysis is obtained by hyphenated techniques, i.e., coupling of, for example, a chromatographic separation technique to a very sensitive elemental detector. As NAA cannot measure on-line, the role of NAA becomes less important. However, for the development of separation and preconcentration techniques, the use of radiotracers with very high specific activity is an outstanding tool, as these techniques can be developed at realistic concentration levels. Radiotracers produced by charged-particle activation, provide just such a specific activity, as the radionuclide formed is not an (radio)isotope of the element irradiated.

5.4. Geo- and Cosmochemistry

Activation analysis has probably most often been applied in the analysis of geological samples. The NAA results on rare earth elements in meteorites, rocks, and sediments were a significant contribution to the development of modern geo- and cosmochemistry. Also in geological surveys, hundreds of analyses are performed on soil and stream sediments by this technique. Silicate rock, nonsilicate rock, meteorites, minerals, and marble can often be analyzed purely instrumentally. *Epithermal activation* enhances the detectability of a number of interesting elements. When, however, up to 40 elements are to be determined – including the rare earths and the platinum-group elements – a preconcentration or some group separations based on ion exchange and liquid – liquid extraction are applied. A *14-MeV activation* can be applied for the determination of some major elements (O, F, Mg, Al, Si, P, Fe, Cu, Zn). Radioanalysis in geochemistry has been described in [45].

5.5. Art and Archaeology

Archaeologists and museum directors hope that small samples of art objects can be analyzed with high sensitivity, preferentially in a nondestructive way. INAA can usually achieve this. The wealth of information on the concentration of trace elements, obtained by the technique, can often serve as so-called fingerprints for identification or classification purposes. The concentrations may give clues to the provenance of the artifact, the methods of treatment, or the cultural or technological context in which it was made. The artifacts of archaeological interest analyzed range from metals (coins, medals, statuettes, utensils) to nonmetallic artifacts (clays, pottery, marble, obsidian, and paintings). An overview is given in [46].

6. Evaluation of Activation Analysis

When activation analysis is compared with other analytical techniques, it is most suited for the multielemental analysis of minor, trace, and ultratrace elements in solid samples. Often a purely instrumental approach is possible, making the technique nondestructive and leaving the sample intact for other investigations. About 65 % of the elements of the Periodic Table can be determined by *INAA* at concentrations lower than micrograms per gram and, in favorable cases, down to nanograms per gram. The detection of another 20 % is less sensitive, and about 10 – 15 % (mostly light elements) can rarely be detected at all. These can, however, be detected by *FNAA* (F) or *CPAA* (B, C, N, O). The lowest detection limits are always obtained after radiochemical separation. These low detection limits are possible because the technique can be made highly selective. Adjustable parameters are irradiation and waiting time before counting, the nature and energy of the bombarding particle or radiation, and above all the high resolving power of the measurement. A *major benefit* of activation analysis is that it can provide very accurate results for trace concentrations. Contamination from reagents or the laboratory environment is excluded after irradiation. Radiochemical separations can be performed after the addition of inactive carrier, avoiding the need for working with trace amounts. The absorption effects of neutrons during irradiation and of gamma rays during counting are small and generally negligible or can be corrected for. The information obtained is independent of matrix or chemical form. Other sources of error, such as interfering reactions, can generally be calculated and thus

accounted for or avoided. Therefore the reproducibility and the precison can be as good as the counting statistics when good laboratory practice is applied. In favorable cases and for homogeneous samples, it can be as good as 1 – 2 %. For the detection of major contents, a precision of > 0.5 % is rarely achievable. The accuracy depends to a large extent on the calibration procedure. When appropriate standards are used, very accurate results can be obtained, making this a technique preeminently suited for certification purposes. The *major drawbacks* are probably that it is expensive, requires access to nuclear facilities such as a reactor, and is generally slow. Rapid analysis based on short-lived isotopes is possible only in exceptional cases. The cost of the counting equipment is of the order of or much less than most other instruments for trace analysis [atomic absorption spectrometry (AAS), inductively coupled plasma optical emission spectrometry (ICP-OES) and mass spectrometry (ICP-MS), X-ray fluoroescence (XRF), etc.]. To work with radioisotopes, a laboratory must meet legal requirements for radiological safety. An additional disadvantage is the lack of information on the chemical form of the element, which makes speciation studies difficult. Finally, analysis of strongly activated matrices renders the instrumental approach impossible and necessitates the application of important shielding material for the radiochemical separations.

7. References

General References

1. W. D. Ehmann, D. E. Vance, J. D. Winefordner, I. M. Kolthoff (eds.): *Radiochemistry and Nuclear Methods of Analysis,* Wiley-Interscience, New York 1991.
2. G. Erdtmann, H. Petri in I. M. Kolthoff, P. J. Elving, V. Krivan (eds.) *Treatise on Analytical Chemistry,* 2nd ed., part I, vol. 14, Wiley-Interscience, New York 1986, pp. 419 – 643.
3. J. Hoste et al. in I. M. Kolthoff, P. J. Elving, V. Krivan (eds.): *Treatise on Analytical Chemistry,* 2nd ed., part I, vol. 14, Wiley-Interscience, New York 1986, pp. 645 – 775.
4. S. J. Parry in J. D. Winefordner, I. M. Kolthoff (eds.): *Activation Spectrometry in Chemical Analysis,* Wiley-Interscience, New York 1991.
5. Z. B. Alfassi (ed.): *Activation Analysis,* vols. I and II, CRC-Press, Boca Raton 1989.
6. J. Tölgyessy, M. Kyrš: *Radioanalytical Chemistry,* vols. 1 and 2, Ellis Horwood, Chichester 1989.
7. D. De Soete et al. (eds.): *Neutron Activation Analysis,* Wiley-Interscience, London 1972.
8. Z. B. Alfassi (ed.): *Determination of Trace Elements,* VCH Verlagsgesellschaft, Weinheim 1994.
9. Z. B. Alfassi (ed.): *Chemical Analysis by Nuclear Methods,* John Wiley, Chichester 1994.
10. M. D. Glascock: Activation Analysis, in: Z. B. Alfassi (ed.): *Instrumental Multi-element Chemical Analysis,* Kluwer, Dordrecht 1999.
11. O. Navrátil et al. (eds.): *Nuclear Chemistry,* Ellis Horwood, New York 1992.
12. W. Seelmann-Eggebert, G. Pfennig, H. Münzel, H. Klewe-Nebenius: *Karlsruher Nuklidkarte – Charts of the Nuclides,* 5th ed., Kernforschungszentrum, Karlsruhe 1981.
13. J. R. Parrington, H. D. Knox, S. L. Breneman, E. M. Baum, F. Feiner: *Chart of the Nuclides,* 15th ed., KAPL, Knolls Atomic Power Lab., New York 1996.
14. R. B. Firestone, V. S. Shirley, C. M. Baglin, S. Y. F. Chu, J. Zipkin: *Table of Isotopes,* 8th ed. (book and CD-ROM + yearly update), John Wiley, Chichester 1996.
15. M. D. Glascock: *Tables for Neutron Activation Analysis,* 4th ed., University of Missouri, Columbia 1996.
16. http://nucleardata.nuclear.lu.se/database
17. http://www.nndc.bnl.gov/nndc/nudat
18. http://www.nndc.bnl.gov/nndcscr/pc_prog/ *Only* NDTxx_16.EXE *and* RADTIONS.ZIP *are required.*
19. http://www.nndc.bnl.gov/nndc/exfor/
20. http://www-nds.iaea.or.at/exfor/
21. http://www.nea.fr/html/dbdata/x4/welcome.html *Only for users registered* at http://www.nea.fr/html/signon.html
22. http://nucleardata.nuclear.lu.se./database/toi/
23. C. Segebade, H. P. Weise, G. L. Lutz: *Photon Activation Analysis,* De Gruyter, Berlin 1987.
24. A. P. Kushelevsky: "Photon Activation Analysis", in Z. B. Alfassi (ed.): *Activation Analysis,* **vol. II,** CRC Press, Boca Raton 1990, pp. 219 – 2237.
25. K. Strijckmans: "Charged Particle Activation Analysis", in Z. B. Alfassi (ed.): *Chemical*

Analysis by Nuclear Methods, Chap. 10, John Wiley, Chichester 1994, pp. 215 – 252.
26. K. Strijckmans: "Charged Particle Activation Analysis", in A. Townshend, R. Macrae, S. J. Haswell, M. Lederer, I. D. Wilson, P. Worsfold (eds.): *Encyclopedia of Analytical Science,* Academic Press, London 1995, pp. 16 – 25.
27. K. Strijckmans: "Charged Particle Activation Analysis", in D. Brune, R. Hellborg, H. J. Whitlow, O. Hunderi (eds.): *Surface Characterisation: a User's Sourcebook,* Wiley-VCH, Weinheim 1997, pp. 169 – 175.
28. K. Strijckmans: "Charged Particle Activation Analysis", in R. A. Meyers (ed.): *Nuclear Methods – Theory and Instrumentation of the Encyclopedia of Analytical Chemistry (EAC), Part 2: Instrumentation and Applications,* Section: Nuclear Methods, John Wiley, New York 2000.

Specific References

29. O. T. Høgdahl in IAEA (ed.): *Radiochem. Methods Anal. Proc. Symp. 1964* **I** (1965) 23.
30. J. Hoste: "Isotopic Neutron Sources for Neutron Activation Analysis," *IAEA-TECDOC* 1988, 465.
31. F. De Corte, F. Simonits, A. De Wispelaere, J. Hoste, *J. Radioanal. Nucl. Chem.* **113** (1987) 145 – 161.
32. F. De Corte et al., *J. Radioanal. Nucl. Chem.* **169** (1993) 125 – 158.
33. S. Van Lierde, F. De Corte, D. Bossus, R. Van Sluijs, S. S. Pommé, *Nucl. Instr. Meth.* **A422** (1999)874 – 879.
34. G. Erdtmann, H. Petri in I. M. Kolthoff, P. J. Elving, V. Krivan (eds.): *Treatise on Analytical Chemistry,* 2nd ed., **part I, vol. 14.** J. Wiley & Sons, New York 1986, p. 446.
35. R. Dams, J. A. Robbins, K. A. Rahn, J. W. Winchester, *Anal. Chem.* **42** (1970) 861 – 866.
36. L. A. Curie, *Anal. Chem.* **40** (1968) 586 – 593.
37. M. L. Verheyke, H. J. Rath in M. Grasserbauer, H. W. Werner (eds.): *Analysis of Microelectronic Materials and Devices,* J. Wiley and Sons, New York 1991, pp. 1 – 39.
38. G. Blondiaux, J. L. Debrun, C. J. Maggiore: "Charged Particle Activation Analysis", in J. R. Temer, M. Nastasi, J. C. Barbour, C. J. Maggiore, J. M. Mayer (eds.): *Handbook of Modern Ion Beam Materials Analysis,* Material Research Society, Pittsburgh, PA 1995.
39. K. De Neve, K. Strijckmans, R. Dams: "Feasibility Study on the Characterization of Thin Layers by Charged-Particle Activation Analysis," *Anal. Chem.* **72** (2000) 2814 – 2820.
40. J. Tolgyessy, E. H. Klehr: *Nuclear Environmental Chemical Analysis,* Ellis Horwood, Chichester 1987.
41. R. Dams in I. M. Kolthoff, P. J. Elving, V. Krivan (eds.): *Treatise on Analytical Chemistry,* 2nd ed., part I, vol. 14, J. Wiley & Sons, New York 1986, p. 685.
42. J. Versieck, R. Cornelis: *Trace Elements in Human Plasma or Serum,* CRC Press, Boca Raton 1989.
43. W. C. Cunningham, W. B. Stroube, *Sci. Total Environ.* **63** (1987) 29 – 43.
44. R. Cesareo (ed.): *Nuclear Analytical Techniques in Medicine,* Elsevier, Amsterdam 1988.
45. H. A. Das, A. Faanhof, H. A. van der Sloot: *Radioanalysis in Geochemistry,* Elsevier, Amsterdam 1989.
46. J. Op de Beeck in I. M. Kolthoff, P. J. Elving, V. Krivan (eds.): *Treatise on Analytical Chemistry,* 2nd ed., **part I, vol. 14,** J. Wiley & Sons, New York 1986, p. 729.

Analytical Voltammetry and Polarography

GÜNTER HENZE, Institut für Anorganische und Analytische Chemie der Technischen Universität Clausthal, Clausthal-Zellerfeld, Germany

1. Introduction 1
2. Techniques 4
2.1. Direct Current Polarography 4
2.2. Pulse Techniques 7
2.3. Alternating Current Polarography . . 11
2.4. Linear-Sweep and Cyclic Voltammetry 13
2.5. Chronopotentiometry 14
2.6. Stripping Techniques 15
3. Instrumentation 20
4. Evaluation and Calculation 25
5. Sample Preparation 27
6. Supporting Electrolyte Solution 29
7. Application to Inorganic and Organic Trace Analysis 31
8. References 40

1. Introduction

The term *voltammetry* is used to classify that group of electoanalytical techniques in which the current (ampere) that flows through an electrochemical cell is measured as the potential (volt) applied to the electrodes in the cell is varied. The term is derived from the units of the electrical parameters measured – *volt-am* (pere)-*metry*.

The essential difference between voltammetric and other potentiodynamic techniques, such as constant current coulometry, is that in voltammetry an electrode with a small surface area ($< 10\,\mathrm{mm}^2$) is used to monitor the current produced by the species in solution reacting at this electrode in response to the potential applied. Because the electrode used in voltammetry is so small, the amount of material reacting at the electrode can be ignored. This is in contrast to the case in coulometry where large area electrodes are used so that all of a species in the cell may be oxidized or reduced.

When mercury is used as the electrode in the form of small drops falling slowly from a fine capillary tube in the test solution, the technique has the special name *polarography*. This name is derived from the fact that the electrode can be polarized. An electrode is said to be polarized when no direct current flows across its interface with the solution even though there is a potential difference across this interface. In his work published in 1922 on "Electrolysis with a dropping mercury electrode", JAROSLAV HEYROVSKÝ referred to this phenomenon and called the recorded current – potential polarization curves *polarograms* [1].

The recommendation of IUPAC is that voltammetry is the general term to be used when current – potential relationships are being investigated and that only when a flowing conducting liquid electrode (such as a dropping mercury electrode) is used as the working electrode should the term polarography be used. Polarography, the original technique, is thus a special case of voltammetry.

The *voltammetric cell* is a multi-phase system, in the simplest form designed with two electronic conductors called *electrodes* immersed in an electrolytic conductor (ionic cell solution). The electrolytic conductor consists of the sample solution with the electrochemically active *analyte* and an excess of an inert *supporting electrolyte*. The analyte is an inorganic or organic species and can be present as a cation, an anion or a molecule.

The application of a voltage to the electrodes produces, as the result of an electrode process, an electrical response – a current – from the analyte in the cell solution. The nature and magnitude of this response may be used to both identify and quantify the analyte. Most polarographic and voltammetric procedures utilize electrode processes in which electrons are exchanged between the two phases. Such a process is referred to as a *charge transfer reaction*, because of the flow of charge, i.e., a current, through the electrode.

The small electrode used to monitor the response of the analyte is known as the *working*

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b05_705

electrode (WE). Even though only a negligible amount of material is involved in the processes occurring at the working electrode, its small size ensures that a high current density develops at its surface. Consequently, it is the processes which occur at this small electrode that control the current flow through the cell. The WE may be constructed from a wide variety of conduction materials, preferably of mercury and graphite, or alternatively of gold and platinum. For polarographic experiments the dropping mercury electrode (DME) is important and for voltammetric investigations the thin mercury film electrode (TMFE), the hanging mercury drop electrode (HMDE), the glassy carbon electrode (GCE), the carbon paste electrode (CPE), the rotating platinum electrode (RPE), and chemically modified electrodes (CME's) may be used.

The second electrode in the simple voltammetric cell, called the *counter electrode* (CE), serves two purposes. It is used to control the potential applied to the working electrode and to complete the circuit for carrying the current generated at the WE. In the former role it must act as a reference electrode.

In modern measuring systems the current carrying role of the counter electrode is separated from its potential control role by introducing the *auxiliary electrode* (AE) as a third electrode of the cell. The addition of the auxiliary electrode means that the counter electrode is now used only to control the potential of the working electrode and so becomes a true *reference electrode* (RE). Two electrodes which are commonly used as reference electrodes for the precise control of the working electrode potential in aqueous media are the silver – silver chloride electrode in a solution of fixed chloride concentration and the saturated calomel electrode or SCE (a mercury – mercurous chloride electrode in a saturated KCl solution). These electrodes are robust, easily constructed and maintain a constant potential. The three electrode system will be discussed in more detail in Chapter 3.

When the voltage applied to the electrodes (WE and AE) is such that no charge transfer reactions are occuring, the working electrode is said to be "polarized". At higher voltages, the polarization of the working electrode disappears as soon as the conditions exist for a charge transfer reaction, $Ox + ne^- \rightleftarrows Red$, to occur at the interface between the electrolyte solution and the WE. The current flow that results from the oxidation or reduction of the analyte is known as a *Faradaic current*. Its magnitude depends on the concentration of the analyte in the sample solution and on the kinetics of all steps in the associated electrode process. Because of the small surface area of the working electrode ($1 - 10\,mm^2$) the current flow is normally in the nA to μA range.

The temporal decrease of the analyte concentration in the interfacial region of the electrode, as a result of reduction or oxidation, is balanced chiefly by diffusion of the analyte from the bulk of the solution. The transport of charged analytes by migration is insignificant in the presence of an excess of supporting electrolyte in the test solution. Normally migration can be neglected if the supporting electrolyte concentration is at least 10^3 times higher than the analyte concentration. However, mass transport to the interface is assisted by convection, facilitated by movement of the solution relative to the electrode, e.g., by stirring or by the action of the mercury dropping from the capillary of the DME.

According to Nernst, diffusion occurs in a stable layer of thickness δ at the interface of the working electrode, i.e., from the electrode surface to some distance into the solution (Fig. 1). Within the Nernst diffusion layer, the decrease of analyte concentration c_a in the sample solution to c_s at the electrode surface is linear and voltage dependent. Convection within the layer is negligible.

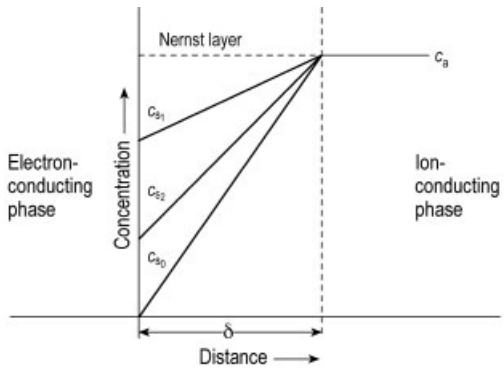

Figure 1. Variation of concentration in the Nernst diffusion layer

For evaluating the diffusion-limited electron transfer current, the following equations are important:

1) The general equation for the cell current as a function of the number of moles converted:

$$i = \frac{\mathrm{d}N}{\mathrm{d}t} nFA \tag{1}$$

where N is the number of moles, t the time, n the number of electrons involved in the electrode reaction, F the Faraday constant, and A the surface area of the working electrode;

2) Fick's first law

$$\frac{\mathrm{d}N}{\mathrm{d}t} = D\left(\frac{\mathrm{d}c}{\mathrm{d}x}\right) \tag{2}$$

where D is the diffusion coefficient, and dc represents the change in concentration over the distance dx from the electrode surface. According to the ideas behind the Nernst diffusion layer, dc/dx can be replaced by $(c_\mathrm{a} - c_s)/\delta$ (Fig. 1). From Equations (1) and (2):

$$i(t) = DnFA\frac{c_\mathrm{a} - c_s}{\delta} \tag{3}$$

The concentration gradient increases with increasing voltage (see c_{s_1} and c_{s_2} in Fig. 1) and reaches its highest value when all active species reaching the electrode by diffusion immediately participate inthe electron transfer reaction. Then $c_s = c_{s_0} = 0$ and Equation (3) becomes

$$i(t) = DnFA\frac{c_\mathrm{a}}{\delta} \tag{4}$$

In this case the cell current is also called the *limiting diffusion current* i_d, whose value remains constant with further increase of voltage; i_lim is accordingly the maximum value of i and, being proportional to the concentration, is important for analytical voltammetry. However, these relations apply only if the thickness of the diffusion layer remains constant during the cell reaction (stationary state). This is true if the electrode is moved at constant velocity (rotating disk or ring electrode) or if the sample solution is stirred or flows past the electrode at constant velocity (flow cell). In such cases and under ideal experimental conditions, when i is plotted against the voltage E, a line parallel to the E axis is obtained on reaching the limiting diffusion current (Fig. 2 A). The relationships change if both electrode phases remain in the stationary state. Without additional convection the thickness of the diffusion layer, and hence dc/dx, changes with time (Fick's second law). In the course of the electrode reaction, as a result of the electrochemical conversion, the diffusion layer thickness increases as $\sqrt{\pi Dt}$. This means that the diffusion current, after reaching the limiting value i_p, decreases with time according to the Cottrell equation:

$$i(t) = DnFA\frac{c_\mathrm{a}}{\sqrt{\pi Dt}} \tag{5}$$

(Fig. 2 B).

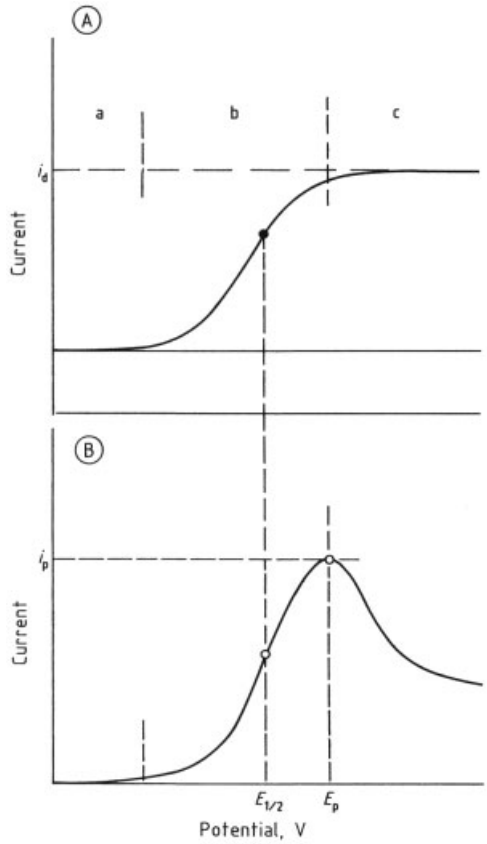

Figure 2. Current – voltage curve at electrodes
A) Stationary; B) Nonstationary diffusion layer thickness
Potential ranges for a) Electrode polarization; b) Electrode reaction; c) Limiting diffusion-controlled current i_d and peak current i_p
E_p = Peak potential; $E_{1/2}$ = Half-wave potential

In *polarography* with the DME, the relationships are even more complicated as both the con-

centration gradient and the diffusion layer width change with time. This is a consequence of the growth of the drop and hence the electrode surface area being in the opposite direction to the diffusion direction. With regard to the electrode reaction the two effects are superimposed, the time average being a constant limiting diffusion-controlled current corresponding to Figure 2 A. Further information on i_{lim} is provided by the Ilkovič equation (Section 2.1, Eq. 8).

The signals proportional to the analyte concentration, i_{d} and i_{p}, are analytically important.

Aside from the diffusion current, currents can occur in voltammetric/polarographic studies which are controlled by reaction kinetics, catalysis, and adsorption.

Kinetic Currents. In kinetic currents, the limiting current is determined by the rate of a chemical reaction in the vicinity of the electrode, provided this precedes the cell reaction. Electrochemically inactive compounds are converted into reducible or oxidizable forms (time-dependent protonation and deprotonation processes, formation and decomposition of complexes, etc.). Conversely, during a chemical reaction after the cell reaction, the product of the electrode reaction is converted to an electrochemically inactive form without influence on the current. However, owing to the changed equilibria between the concentrations of the oxidized and reduced forms at the electrode surface, the half-wave and peak potentials are shifted (Section 2.1). In evaluating kinetic effects, cyclic voltammetry can be helpful (Section 2.4).

Catalytic Currents. For analytical studies, catalytic currents, of which two types are known, are of great interest. In the first case the product of the cell reaction is returned, by a chemical reaction, to the initial state of the analyte. As a result, the analyte concentration at the electrode surface is always high, which results in a considerable increase of the limiting current.

The currents based on the catalytic evolution of hydrogen are the best known. The catalysts are electrochemically inactive heavy metals (primarily salts of the platinum group metals) or organic compounds. They lower the hydrogen overvoltage on the mercury working electrode and so shift hydrogen evolution to more positive potentials. The catalytic current associated with this depends, within limits, on the catalyst concentration. Because of the large number of substances that can lower the hydrogen overvoltage, catalytic currents can only be used to determine concentration under controlled conditions.

Adsorption currents arise when the analyte, in its original form or after chemical or electrochemical conversion, is adsorbed on the working electrode. Adsorption and desorption lead to capacitance changes in the electrochemical double layer and so to analytically usable current signals (Section 2.3).

The theoretical principles of voltammetry and polarography are covered in several monographs [1–7].

2. Techniques

Voltammetric measurement techniques have continued to improve. Some of the improvements have had to await the advance of electronics before being used in practical analysis. Methods with modulated direct current (d.c.) ramps, e.g., differential pulse polarography, as well as alternating current (a.c.) and square-wave polarography, etc., are worth mentioning. New possibilities for controlling experiments and evaluating current signals have followed from digital electronics. The basic principles of analytical voltammetry and polarography are still best explained by way of classical d.c. polarography.

2.1. Direct Current Polarography

Direct current polarography (DCP), introduced by Heyrovský in 1923, is characterized by the recording of current – voltage curves in an unstirred sample solution with the DME as the working electrode. With a slow linear change of voltage, stepwise polarograms are obtained, with marked oscillations caused by the growth and fall of the mercury drop. For better evaluation, the curves can be damped with electronic filters (Fig. 3).

The mercury flows with a mass flow rate m (mg/s) and a periodic drop time t_{d} (s) from a glass capillary. The surface area A of the mercury drop is greatest at the end of the drop time and, at time t during the growth of the drop is:

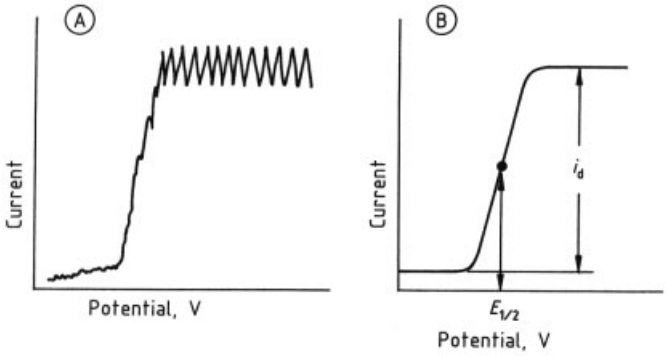

Figure 3. Direct current polarograms
A) Undamped; B) Damped

$$A = k m^{2/3} t^{2/3} \tag{6}$$

With this expression and the Cottrell equation (5), the following relation is obtained for the magnitude of the instantaneous current at time t_d for an individual drop:

$$i = k' m^{2/3} t_d^{1/6} n D^{1/2} c_a \tag{7}$$

The constant k' includes all constant quantities, including the Faraday constant, and also takes account of the conditions for spherical diffusion. In 1935, it was calculated by ILCOVIČ to have the numerical value 0.708.

The mean of the limiting diffusion current in the DC polarogram is obtained from Equation (7) by integration over the entire drop time:

$$i_d = \frac{1}{t_d} \int_0^{t_d} i \cdot dt = 0.607 \cdot n \cdot m^{2/3} \cdot D^{1/2} \cdot t_d^{1/6} \cdot c_a \tag{8}$$

This is the Ilcovic equation, and applies in the voltage range in which all analyte species reaching the electrode surface by diffusion are either reduced or oxidized. The product $m^{2/3} t_d^{1/6}$ is dependent on the particular glass capillary used and is called the capillary constant. The characteristic shape of the DC polarogram can be described by the Nernst equation:

$$E = E'_0 + \frac{RT}{nF} \ln \frac{c_{a_{ox}}}{c_{a_{red}}} \tag{9}$$

on condition that, for the particular electrode potential E, the mean current magnitude at an individual mercury drop is considered. For the case of reduction $Ox + e^- \rightarrow Red$:

$$i = k \left(c_{a_{ox}} - c^0_{a_{ox}} \right) \tag{10}$$

where $c^0_{a_{ox}}$ indicates the concentration of the analyte on the electrode surface (corresponding to c_s in Eq. 3). The same applies for the concentration $c^0_{a_{red}}$, which increases as $c^0_{a_{ox}}$ decreases; i.e., the sum $c^0_{a_{ox}} + c^0_{a_{red}}$ corresponds to the total analyte concentration $c_{a_{ox}}$. For the limiting diffusion current, at which $c^0_{a_{ox}} = 0$ and $c^0_{a_{red}} = c_{a_{ox}}$, it follows from Equation (10) that:

$$i_d = k c_{a_{ox}} \tag{11}$$

Thus:

$$c^0_{a_{ox}} = c_{a_{ox}} - \frac{i}{k} = \frac{i_d}{k} - \frac{i}{k} \tag{12}$$

and:

$$c^0_{a_{red}} = \frac{i}{k} \tag{13}$$

When the expressions for $c^0_{a_{ox}}$ (Eq. 12) and $c^0_{a_{red}}$ (Eq. 13) are inserted in the Nernst equation (9):

$$E = E'_0 + \frac{RT}{nF} \ln \frac{i_d - 1}{i} \tag{14a}$$

or:

$$E = E'_0 + \frac{0.059}{n} \log \frac{i_d - 1}{i} \tag{14b}$$

For $i = i_d/2$, $E = E_{1/2}$ by definition. $E_{1/2}$, the half-wave potential in the DC polarogram is determined by the electrochemical properties of the redox system concerned.

According to Equation (14), the slope of the current curve depends on the electron transfer number n. Conversely, n can be calculated from the slope. However, the slope also depends on the reversibility of the electrode reaction.

If the activity coefficients f_{ox} and f_{red} for $c_{\mathrm{a_{ox}}}$ and $c_{\mathrm{a_{red}}}$ are inserted into the Nernst equation (9) as well as the diffusion coefficients $D_{\mathrm{ox}}^{1/2}$ and $D_{\mathrm{red}}^{1/2}$ (see Eq. 8), there follows for the half-wave potential:

$$E_{1/2} = E_0 - \frac{RT}{nF} \ln \frac{f_{\mathrm{red}} D_{\mathrm{ox}}^{1/2}}{f_{\mathrm{ox}} D_{\mathrm{red}}^{1/2}} \tag{15}$$

As f and D are determined by the composition of the supporting electrolyte solution, the half-wave potential also depends on this composition.

Direct current polarography is a basic method, from which improved analytical techniques of polarography and voltammetry have been developed. With DCP, inorganic and organic analytes can be analyzed with a sensitivity of about 10^{-5} mol/L. For determining several analytes simultaneously, their half-wave potentials must be at least 100 mV apart.

On closer examination of DC polarograms, the current measured in the voltage range of the charge transfer reaction consists of two parts, the true measurement signal and an interference or noise signal. The measurement signal is the diffusion current, whose value depends on the electrode process and which, because of its origin, is called the Faradaic current i_{F}. The interference signal is the charging current or capacitive current i_{C}, which is a consequence of charging the electrochemical double layer. In the case of the DME, the charging current is given by the equation

$$I_{\mathrm{C}} = (\mathrm{d}E/\mathrm{d}t)\, C_{\mathrm{D}} A + (\mathrm{d}A/\mathrm{d}t)\, C_{\mathrm{D}} E \tag{16}$$

where C_{D} is the double layer capacitance per unit area, E the potential applied to the cell and A the area of the electrode. The first term of this equation allows for the effect of the rate of change of the electrode potential on the charging current, while the second term gives the increase in i_{C} due to the growing mercury drop.

With deceasing analyte concentration i_{F} becomes smaller and approaches i_{C}. The $i_{\mathrm{F}} : i_{\mathrm{C}}$ signal-to-noise ratio is the determining factor for the sensitivity of DCP. If more sensitive polarographic determinations are to be performed, this ratio must be increased by using techniques in which either the Faradaic current is enhanced or the charging current is minimized, or where both can be achieved.

During drop growth (drop life t_{d}), i_{F} and i_{C} are subject to changes in opposite directions. As shown in Figure 4, i_{F} increases with increasing surface area of the mercury drop, while i_{C} becomes smaller.

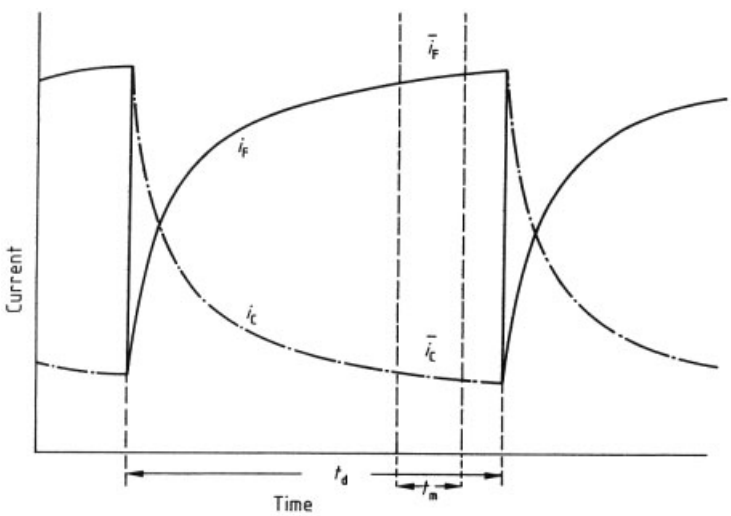

Figure 4. Change with time of i_{F} and i_{C} during the drop life t_{d} with the measurement or sampling interval t_{m} and the mean values $\bar{i}_{\mathrm{F}}$ and $\bar{i}_{\mathrm{C}}$ over the sampled range

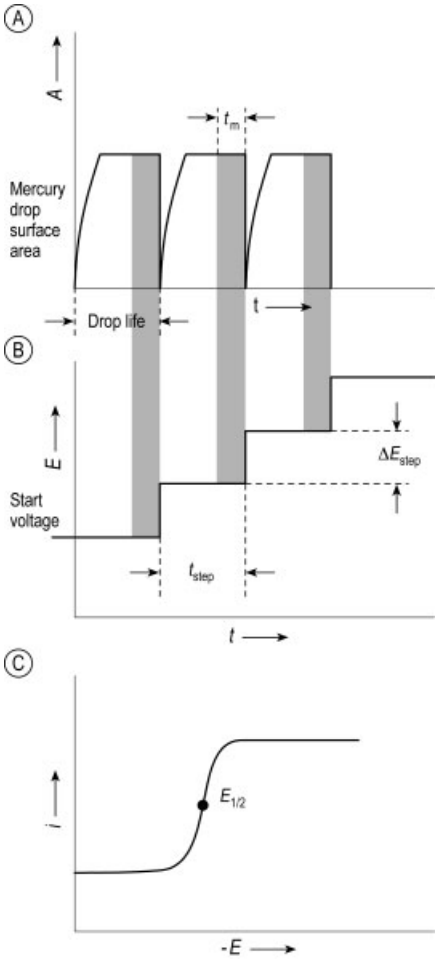

Figure 5. Measurement technique of current sampled DC polarography with the static mercury drop electrode as working electrode
(a) Variation of surface area with time for the SMDE; (b) Potential – time profile for the staircase voltage applied to the electrode; (c) Current – potential response (polarogram). t_m Measuring interval; t_{step} Time interval of the voltage step; ΔE_{step} Voltage step of the applied staircase waveform

If the current is not measured over the whole drop life (as in classical DCP), but over a short period of time at the end of the life of a mercury drop where the growth of surface area is least the current averaged over the measurement (sampling) interval t_m has an improved ratio of $i_F : i_C$. This is the basis of the *current sampled DC-polarography* using a linear scan.

Developments in the measurement technique have resulted in a major reduction in the charging current. Firstly, with the introduction of digitally controlled instruments, the linear voltage ramp applied to the electrodes has been replaced with a stepped (staircase) waveform. The voltage steps are synchronized with the drop formation and the current is sampled at the end of the drop life. Since the voltage is constant when the current is measured, there is no contribution to the charging current from this source (see Eq. 16). The second innovation which led to a reduction in the charging current is the *static mercury drop electrode* (SMDE). In this electrode the mercury drop is extruded rapidly and its growth then stopped after which the area of the mercury electrode remains constant. The area – time profile for the SMDE, the potential – time waveform with the sampling period, and the resulting DC polarogram is shown in Figure 5. The current sampled DC polarography (sometimes called Tast Polarography) is not only more sensitive than classical DCP (ca. 5×10^{-7} mol/L), but also leads to polarograms without oscillations.

2.2. Pulse Techniques

With voltammetric pulse methods, Barker and Gardner [31] found very effective means of reducing the unwanted capacitive (noise) current, markedly improving the sensitivity of polarographic and voltammetric determinations.

A periodically changing square-wave potential with constant or increasing amplitude ΔE_p is applied to the working electrode. During the pulse duration t_p, the currents i_F and i_C that flow in response to the potential change ΔE_p decay in different ways. The Faradaic current decreases according to $t^{-1/2}$, (Eq. 5) while the capacitive current decreases according to

$$i_C = \frac{\Delta E_p}{R} \exp\left(-\frac{t}{R \cdot C_D}\right) \tag{17}$$

(Eq. 17 corresponds to the capacitor formula with discharging resistance R and double-layer capacitance C_D). Figure 6 illustrates the variations of i_F and i_C during the duration t_p of the square pulse ΔE_p. If the current is measured only toward the end of the pulse, essentially only i_F is recorded since i_C has effectively decayed to zero.

The methods based on the application of square voltage pulses differ in the frequency and

height of the pulses applied and in the principle of measurement.

Pulse methods may be used with the DME, the HMDE, the SMDE as well as with solid electrodes and the different types of modified electrodes. Except for square-wave polarography, normally only one pulse is applied to each drop in analytical studies with the DME or SMDE. This pulse is synchronized with the drop period and is applied in each case towards the end of the drop life.

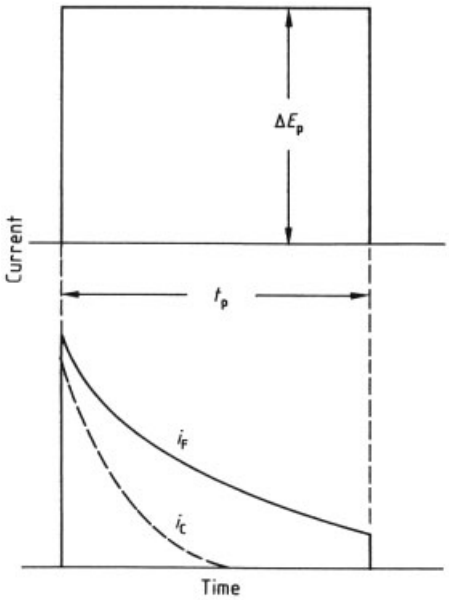

Figure 6. Principle of the pulse methods, according to BARKER

Square-Wave Polarography. In the oldest technique of square-wave polarography (SWP), a periodic rectangular alternating voltage with a frequency of 125 Hz and an amplitude in the range of $\Delta E_{\mathrm{p}} = 5 - 30\,\mathrm{mV}$ is superimposed on the linear rising d.c. ramp. The current is sampled over a measurement interval t_{m} near the end of each square-wave half cycle. This current is then amplified and recorded as a function of the applied d.c. voltage. Peak-shaped polarograms are obtained in which the peak current corresponds to the DC polarographic half-wave potential.

Sampling the current near the end of each half cycle ensures that the charging current arising from the sudden step in potential at the beginning of the half cycle has decayed away and does not contribute to the measured current. However, when using the DME, there is still a charging current flowing during t_{m} as a result of the slight increase in area of the mercury drop (see Eq. 16). This is minimized by using a tilted square-wave as shown in Figure 7.

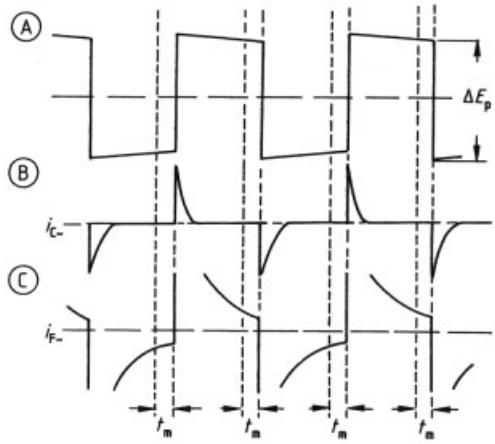

Figure 7. Square-wave alternating voltage and current signals in square-wave polarography
A) Waveform with pulse tilt; B) Capacitive current $i_{\mathrm{C}\sim}$; C) Faradaic current $i_{\mathrm{F}\sim}$
t_{m} = Measuring interval

In the modern version of square-wave polarography the linear d.c. potential ramp is replaced by a d.c. staircase potential and a SMDE replaces the DME. The drop life is synchronized with each step in the staircase potential. The constant potential of each voltage step is modulated with a small amplitude (ΔE_{p}) alternating square-wave voltage of frequency f towards the end of the step as shown in Figure 8.

The current is measured over a number of cycles in the measurement period t_{m}, twice in each cycle, at point 1 at the positive end of the pulse and at point 2 at the negative end. The average of the differences between the current values $i_1 - i_2$ for each measurement cycle is plotted against the potential of the voltage ramp and gives a peak-shaped polarogram, as before.

For reversible electrode processes, the peak current is dependent on the superimposed rectangular voltage:

$$i_{\mathrm{p}} = kn^2 D^{1/2} \Delta E_{\mathrm{p}} c_{\mathrm{a}} \tag{18}$$

where k is constant at a given frequency. The square-wave polarograms shown in Figure 9

highlight the influence of n and ΔE_p on the height and width of the peaks.

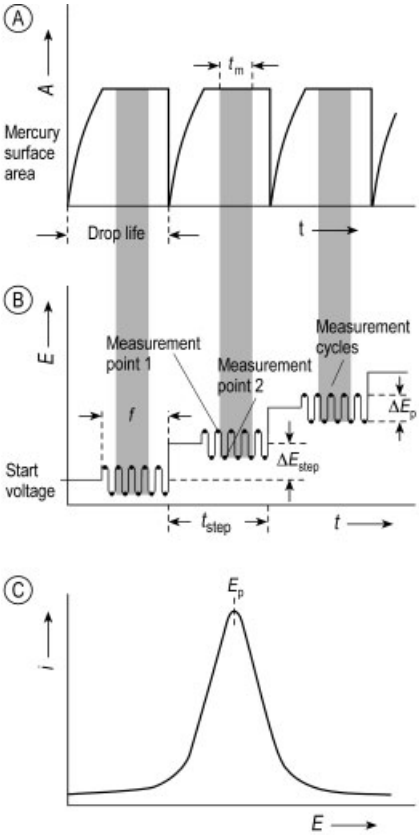

Figure 8. Digitally controlled SWP using the SMDE
a) Variation of the mercury drop surface area with the measurement period t_m; b) Potential – time waveform with the measuring points 1 and 2; c) Current potential response (polarogram). Typical values of parameters: $\Delta E_p \sim 25$ mV, $f \sim 125$ Hz (in 40 ms), $\Delta E_{step} \sim 10$ mV, $t_{step} \sim 10$ s

For reversible processes, sensitivities of about 10^{-8} mol/L can be achieved with SWP and the peak-to-peak resolution is 40 – 50 mV. The sensitivity decreases rapidly with increasing irreversibility of the electrode process.

Normally the voltage scan rate of SWP measurements is about 20 mV/s, but much higher scan rates up to 1000 mV/s can readily be programmed making the method ideal for fast voltammetric measurements.

A high scan rate is the characteristic feature of square-wave voltammetry (SWV), first described by Osteryoung and Osteryoung [32]. In this technique the whole measurement process can be carried out with an extremely rapid change of voltage on a single mercury drop. SWV thus enables very short analysis times and is particularly important as a detection method in flow systems.

A relatively large square wave signal of amplitude $\Delta E_p = 50$ mV is superimposed on a stepped voltage ramp (staircase) with voltage steps of about 10 mV. The duration of the square wave cycle is equal to that of the staircase voltage steps t_{step} and is usually within the range 5 – 10 ms.

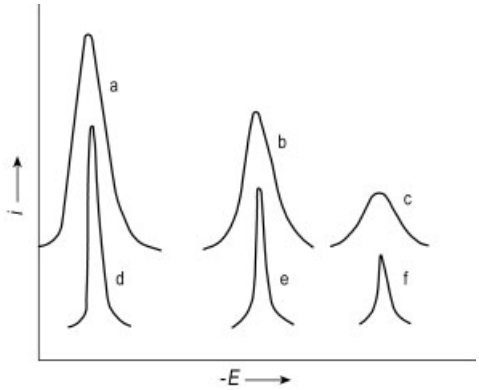

Figure 9. Dependence of the peak height on the square-wave voltage amplitude ΔE_p and on the number n of electrons transferred, n [4]
a) – c) $n = 1$ (thallium); d) – f) $n = 3$ (bismuth)
a), d) $\Delta E_p = 40$ mV; b), e) $\Delta E_p = 20$ mV; c), f) $\Delta E_p = 10$ mV

As can be seen from Figure 10, two values of current are determined for each square wave cycle. Measurements are made at point 1 at the positive end of the cycle and point 2 at the negative end. The difference between the two current values $i_1 - i_2$ is plotted against the potential of the voltage ramp and gives a peak-shaped current –voltage curve, as in SWP.

The pulse times, which are in the millisecond range, enable the potential to be scanned at extremely high rates of up to 1200 mV/s. Only one mercury drop is required for each measurement process. However, this rapid procedure is achieved at the cost of sensitivity, as the ratio of Faradaic to capacitive currents is lowered by the short pulse times.

The state of development and capability of square-wave polarography/voltammetry has been reviewed [8], [33].

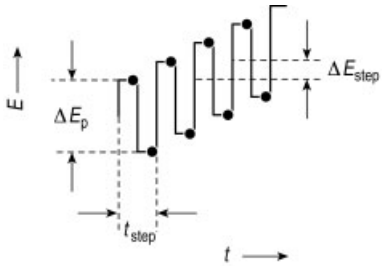

Figure 10. Measurement technique of square-wave voltammetry
ΔE_{step} = Voltage step of the staircase voltage ramp, (ca. 10 mV) ΔE_{p} = Amplitude of superimposed pulse, ca. 50 mV; t_{step} = wave form period, ca. 5 – 10 ms; 1 and 2 Measurement points

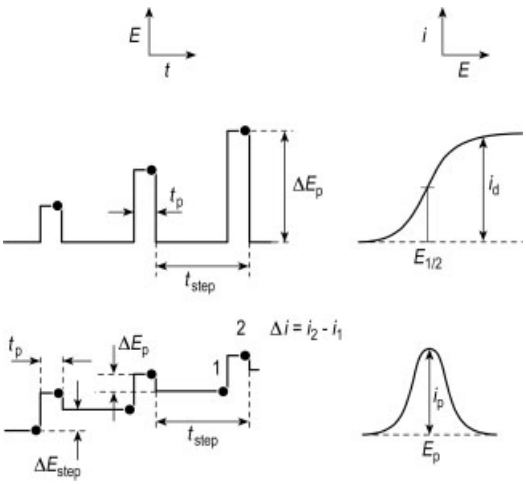

Figure 11. Potential – time waveforms and current – potential responses of A) normal pulse polarography and B) differential pulse polarography (typical values of parameter see text)

Normal Pulse Polarography. In normal pulse polarography (NPP), the voltage of the working electrode is changed, not by means of a direct voltage ramp (as in SWP), but by increasing rectangular pulses superimposed on an initial, constant voltage. The application of the pulses is synchronized with the mercury drop and occurs just before the end of the drop life. Only one voltage pulse with a duration of 30 – 60 ms is applied to each drop. The amplitude ΔE_{p} increases from one pulse to the next and reaches a maximum of 1000 mV. To eliminate i_{C} the current measurement is made ca. 10 – 15 ms before the end of the pulse (Fig. 11).

The measured current determined in pulse polarography is plotted against the pulse amplitude and, as in DCP, yields a curve in the form of a wave. The height of the wave is proportional to the analyte concentration, and the half-wave potential corresponds to the DCP value. Peak-shaped curves can be obtained when the current differences between successive pulses are plotted against the voltage. Sensitivities are ca. 10^{-7} mol/L; the peak resolution is reported to be ca. 100 mV.

Differential Pulse Polarography. The most important pulse method is differential pulse polarography (DPP). In this technique a stepped voltage rise (modern staircase technique) is used and a rectangular voltage pulse with constant amplitude ΔE_{p} of 10 – 100 mV is applied to each mercury drop at the end of its drop time; the pulse duration t_{p} is 40 – 60 ms.

As shown in Fig. 11 B two current measurements are made in DPP: the first i_1 at measurement point 1, immediately before the application of voltage pulse and the second i_2 at point 2 near the pulse end. Both measurements are made on the same mercury drop and also over the same surface area, when the SMDE is used as the working electrode. The difference $\Delta i_1 = i_2 - i_1$ is plotted against the applied d.c. voltage (staircase) and gives a peak-shaped polarogram, because $\Delta i_1/\Delta E$ reaches a maximum in the region of $E_{1/2}$.

Determinations by DPP are on average ten times more sensitive than determinations with NPP. The sensitivity of DPP is about 10^{-8} mol/L and for irreversible processes it falls off only slightly to 5×10^{-8} mol/L. The loss of sensitivity by irreversibility is thus smaller than in the case of other pulse methods. For reversible processes the peak height in DP (differential pulse) polarograms is proportional to the concentration to be analyzed, c_{a}:

$$i_p = \frac{n^2 F^2}{4RT} A c_a \Delta E_p \sqrt{\frac{D}{\pi t_p}} \quad (19)$$

and depends not only on the quantities already mentioned, but on the amplitude of the voltage pulse ΔE_p and the pulse duration t_p. The DP polarograms shown in Figure 12 highlight the increase of peak heights with increasing ΔE_p.

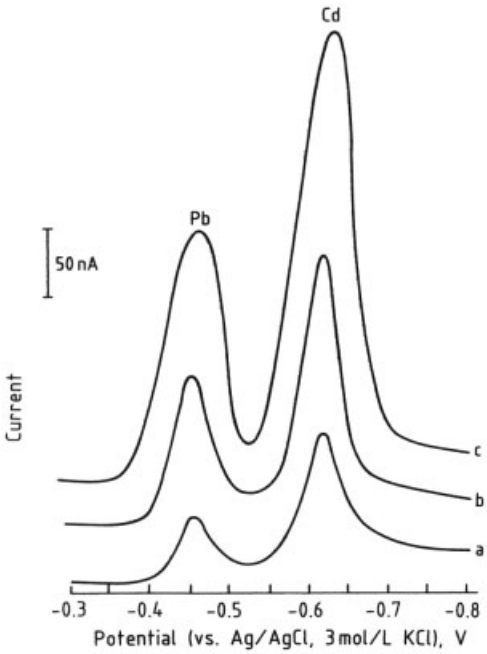

Figure 12. Influence of the pulse amplitude on DP polarograms; Pb and Cd (10 mg/L) in acetate buffer solution pH 2.8
a) $\Delta E_p = 10$ mV; b) $\Delta E_p = 25$ mV; c) $\Delta E_p = 75$ mV

The theory, measurement techniques, and analytical applications of pulse methods have been reported [9].

2.3. Alternating Current Polarography

In this technique, a linear voltage ramp is modulated with a sinusoidal alternating voltage of small amplitude ($\Delta E_\sim = 10 - 100$ mV) and low frequency ($f = 5 - 100$ Hz). The superimposed alternating voltage causes an alternating current $i_\sim$, whose size depends on the instantaneous value of the linear voltage ramp, and is greatest at the half-wave potential. On plotting the selectively measured alternating current against the applied voltage, the peak-shaped AC polarogram is obtained (Fig. 13).

In reversible processes, the peak potential of an AC polarogram is identical with the DC polarographic half-wave potential. The peak current is

$$i_p = \frac{n^2 \cdot F^2}{4RT} (2\pi \cdot D)^{1/2} \cdot A \cdot \Delta E_\sim \cdot f^{1/2} \cdot c_a \quad (20)$$

and is thus dependent on the usual variables as well as being influenced by the frequency f, the amplitude $\Delta E_\sim$ of the superimposed alternating voltage and the area of the electrode A. The value of i_p also depends on the reversibility of the cell reaction and decreases with increasing irreversibility. Owing to the high capacitive current component in AC polarography, caused by the periodic charging and discharging of the double layer, the sensitivity is limited to about 10^{-5} mol/L.

Compared with the alternating voltage applied to the electrode, the capacitive current shows a phase shift of 90° ($\pi/2$). Since the Faradaic alternating current $i_{F\sim}$ has a phase shift of only about 45°, the ratio $i_{F\sim} : i_{C\sim}$ and therefore the sensitivity of AC polarographic determinations can be improved by phase-selective rectification. Instruments designed for AC polarography generally have provision for phase selective current measurements to be made (*AC1 polarography*).

By using phase selective measurement of a higher-harmonic alternating current (e.g., 2nd harmonic with a frequency $2f$) increased sensitivity (10^{-7} mol/L) is obtained because of the marked reduction in the capacitive current component of the higher harmonic (*AC2 or second harmonic wave polarography*).

A phase-selective second harmonic wave polarogram is shown in Figure 14. In contrast to the AC1 polarograms, it has two current peaks. For an invariant residual current (Fig. 14 A), the distance between the two current peaks $i_p^+ + i_p^- = i_p$ can be used for analytical purposes. For a rising residual current (Fig. 14 B), the current components i_p^+ and i_p^- are measured over the extrapolated course of the residual current.

Tensammetry is a variant of AC polarography. In this method, the capacitive current component, whose value is determined by the

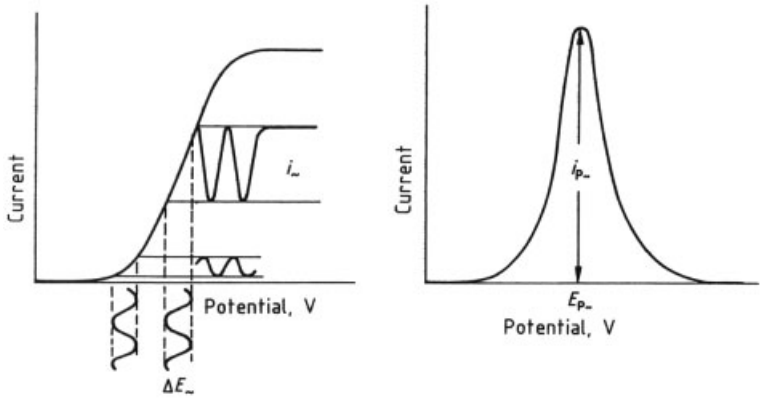

Figure 13. Waveform and response of AC polarography

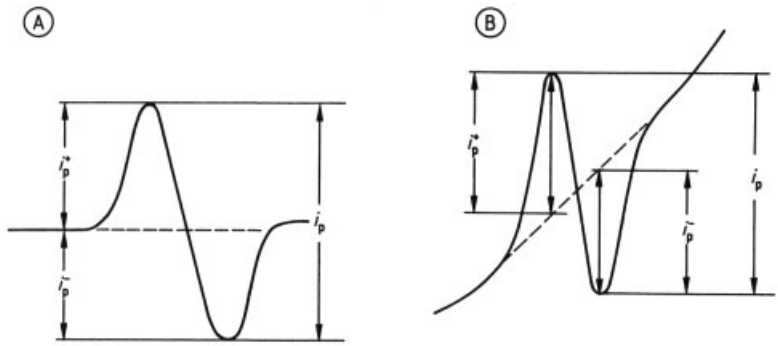

Figure 14. Current profile and evaluation of second harmonic wave polarograms
A) Horizontal residual current; B) Rising residual current

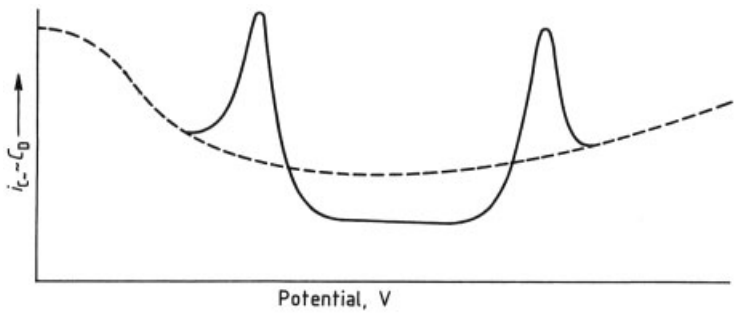

Figure 15. Tensammogram
Alternating current polarogram of the supporting electrolyte solution (– – –); Formation of current peaks after addition of a surfactant (—-); $i_{c\sim}$ = Capacitive current; C_D = Double layer capacitance

double layer capacitance, is measured instead of the Faradaic alternating current. Capacitance changes are caused by the adsorption and desorption of surface active compounds at the electrode surface. Commonly, two almost symmetrical current peaks that are particularly well defined and known as tensammetric peaks are obtained. In the potential region between the peaks the compound is adsorbed. As a result, the capacitance of the electrochemical double layer is lowered and the capacitive current is smaller than the current of the supporting electrolyte solution (Fig. 15).

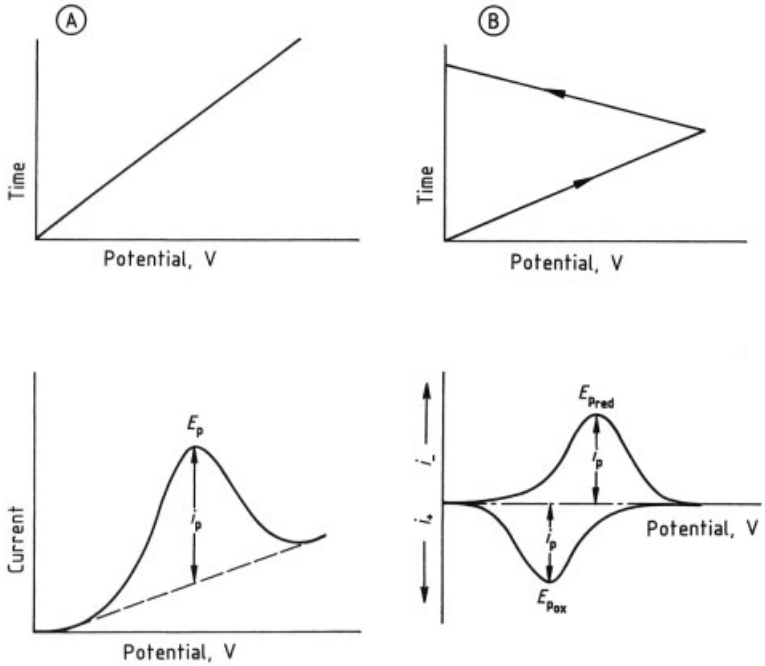

Figure 16. Principle of linear sweep voltammetry (A) and cyclic voltammetry (B)
Top: Variation of voltage with time; Bottom: Resulting current – potential curves

Tensammetric studies are important for the characterization of interfacial problems and for the analysis of surface-active compounds (e.g., surfactants) that are neither reducible nor oxidizable.

2.4. Linear-Sweep and Cyclic Voltammetry

Both linear-sweep voltammetry (LSV) and cyclic voltammetry (CV) are based on recording the current during a linear change of voltage at a stationary working electrode. The rate of change of voltage $v = \mathrm{d}E/\mathrm{d}t$ is relatively high, in the range 1 – 100 mV/s. The difference is that in the case of CV the voltage is returned to the starting potential; CV is therefore also known as triangle-sweep voltammetry.

Figure 16 compares the variation of voltage with time and the response for the two methods. The peak currents, for both LSV and CV, are proportional to the analyte concentration:

$$i_{\mathrm{p}} = k c_{\mathrm{a}} v 1/2 \tag{21}$$

and are influenced by the rate of change of voltage v. The sensitivity of LSV is about 10^{-7} mol/L and the resolution is about 50 mV.

Cyclic voltammetry is mainly used for studying the reversibility of electrode processes and for kinetic observations, and only sometimes for analytical purposes. The voltage cycle illustrated in Figure 16 ensures that the reaction products formed at the potential $E_{\mathrm{p_{red}}}$ on the cathodic path are reoxidized at $E_{\mathrm{p_{ox}}}$ in the anodic sweep.

For a reversible redox process:

$$\Delta E_{\mathrm{p}} = E_{\mathrm{p_{red}}} - E_{\mathrm{p_{ox}}} = \frac{-57}{n}\,\mathrm{mV} \tag{22}$$

the position of the current peak in this case being independent of the voltage scan rate. The two peaks have equal heights. With increasing irreversibility, ΔE_{p} becomes greater. For quasi-reversible processes and for a slow change of voltage, the difference is about (60/n) mV, but it becomes greater for a faster sweep. For totally irreversible processes, the reduction product is not reoxidized, so the anodic current peak is not seen.

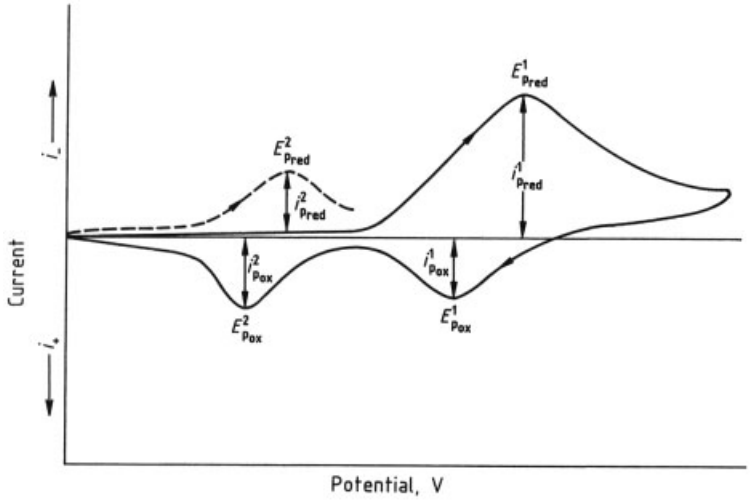

Figure 17. Cyclic voltammogram of a reaction with reversible charge transfer and subsequent chemical reaction
First sweep (—-); Second sweep (– – –)

If the reversible charge transfer is followed by a chemical reaction during which an electrochemically active product is formed, a cyclic voltammogram of the type shown in Figure 17 is obtained.

In the first sweep the two peaks $E^1_{\mathrm{p_{red}}}$ and $E^1_{\mathrm{p_{ox}}}$ appear. The ratio of the peak heights is not, however, 1: 1, as would be expected for a reversible process. Rather, $E^1_{\mathrm{p_{ox}}}$ is smaller than $E^1_{\mathrm{p_{red}}}$, since part of the electrolysis product is chemically converted and therefore no longer available for reoxidation. Since the formed product is electrochemically active, a second peak $E^2_{\mathrm{p_{ox}}}$, which corresponds to the oxidation of this product, appears in the complete anodic sweep. In the second sweep an additional cathodic peak $E^2_{\mathrm{p_{red}}}$, which indicates the reduction of the previously oxidized compound, can also be recorded.

Cyclic voltammetry gives information on the redox behavior of electrochemically active species and on the kinetics of electrode reactions as well as offering the possibility of identifying reactive intermediates or subsequent products [10].

2.5. Chronopotentiometry

In voltammetric studies, the current flowing through the working electrode is measured as a function of the applied voltage and the rate at which the voltage is scanned. In chronopotentiometry, the change of potential of the working electrode with time at constant current is followed. Chronopotentiometry has also been called "galvanostatic voltammetry" and can, therefore, be grouped with the methods described in this chapter.

Determinations are usually carried out under stationary conditions. The electrolytic processes occurring when current flows decide the ratio $c_{\mathrm{a_{ox}}} : c_{\mathrm{a_{red}}}$ of the analyte on the electrode surface, which leads to changes of potential according to the Nernst Equation (9). The potential – time plot is illustrated in Figure 18, and depends on the diffusion behavior and analyte concentration.

The time between the steep sections of the curve is called the transition time τ, which, in diffusion- controlled reversible electrode processes, is described by the Sand equation:

$$\tau^{1/2} = \frac{\pi^{1/2} n F D^{1/2}}{2 j_0} c_{\mathrm{a}} \tag{23}$$

where j_0 is the current density (A/cm^2), and $\tau^{1/2}$ corresponds to the limiting diffusion current in DC polarography. At constant j_0, the potential – time function is:

$$E = E_{\tau/4} + \frac{RT}{nF} \ln \frac{\tau^{1/2} - t^{1/2}}{t^{1/2}} \tag{24}$$

with $E_\tau/4 = E_{1/2}$ for a reversible reaction. Chronopotentiograms were traditionally evaluated graphically, which with small transition times is difficult and often inaccurate. It is more convenient to measure the time values – which are proportional to concentration – electronically via the first or second derivative. In this way empirical time values are obtained which always differ to some degree from the theoretical transition times [34].

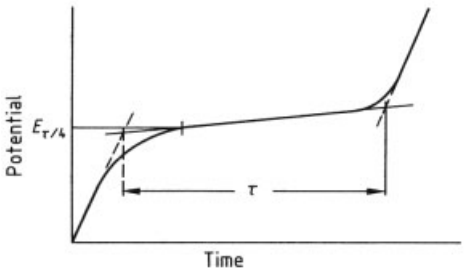

Figure 18. Chronopotentiometric potential – time curve with the transition time τ

Normal chronopotentiometry has little importance for analysis. Determinations are possible in the range 10^{-4} – 10^{-5} mol/L at best. However, the technique is of use in understanding the methods of stripping chronopotentiometry and potentiometric stripping analysis, which are important in trace analysis.

2.6. Stripping Techniques

The stripping techniques are so-called multistage combined procedures, with the characteristic that the voltammetric or chronopotentiometric determination is preceded by an electrochemical accumulation of analyte. The accumulation occurs on the surface of a stationary working electrode and leads to considerably improved performance. The methods enable analytical studies in the pico-trace range [11–13].

Stripping Voltammetry. The analyte is accumulated, as the metal or in the form of a known compound, by controlled potential electrolysis or adsorption and with constant convection of the sample solution. The subsequent stripping of the preconcentrated species from the electrode occurs voltammetrically as a linear voltage ramp is applied to the electrode, generally using one of the DC, AC, or DP measurement modes. The current peaks in the voltammograms result from either the reduction or oxidation of the accumulated products. Depending on this electrode reaction, a distinction is made between *anodic stripping voltammetry* (ASV) and *cathodic stripping voltammetry* (CSV).

According to the Randles – Sevcik equation:

$$i_p = kn^{3/2}AD^{1/2}vc1/2 \quad (25)$$

the peak height i_p in the stripping voltammogram is (analogous to i_p in the linear sweep and cyclic voltammogram) proportional to the concentration c and is also influenced by the working conditions, particularly by the surface area A of the working electrode and the rate of change of voltage v during the stripping process. The proportionality to concentration relates to the amount of accumulated material or to that part of the electrolysis product which takes part in the stripping process. Consequently, i_p depends on two further influences: the accumulation time and the degree of stripping of the electrolysis product. By careful selection of all parameters, limits of determination can be well below 10^{-8} mol/L. Various means of trace accumulation and determination are used in stripping voltammetry. In the simplest case the analyte is reduced to the metal and accumulated as an amalgam at a stationary mercury electrode (HMDE or TMFE). The determination step proceeds in the reverse direction to the accumulation and is based on the anodic stripping of the metal (reoxidation). This process of ASV is also known as *inverse voltammetry* and can be illustrated as follows:

$$M^{n+} + ne^- + (Hg) \underset{\text{stripping}}{\overset{\text{accumulation}}{\rightleftharpoons}} M^0\,Hg$$

The current profile and the sequence of steps for inverse voltammetric determination of the analyte M^{n+} are shown in Figure 19.

The analyte is deposited from the stirred sample solution at a voltage E_{acc}, as metal on the mercury electrode. The accumulation time is followed by a rest period, during which the solution is unstirred and the cathodic current falls to a small residual value. During the determination step, the voltage is ramped to more positive values (linear sweep, potential scan rate $v = \Delta E/\Delta t$) and, at the appropriate potential, the accumulated metal is reoxidized. The resulting

current – potential curve displays a peak, with peak potential E_p. The height of the peak is proportional to the analyte concentration (Eq. 25).

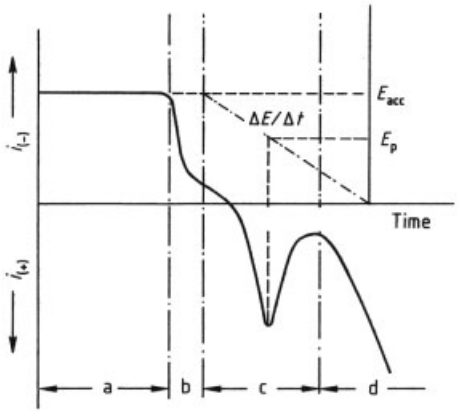

Figure 19. Stripping voltammetric determination
a) Accumulation time; b) Rest period; c) Determination or stripping step; d) Anodic dissolution of mercury

The results of stripping voltammetric determinations are reproducible if the operating conditions are maintained exactly. This includes reproducible renewal of the electrode surface, uniform stirring of the solution, reproducible electrolysis and rest times, and exact adjustment of the electrolysis voltage and its rate of change.

The electrolysis time, adjusted to the analyte concentration to be determined, is usually 1 – 5 min.

Useful guidance on the selection of a suitable electrolysis voltage E_{acc} is given by the DC or DP voltammogram of the analyte in the particular supporting electrolyte solution. For an electrode reaction with $n = 1$, the voltage should be about 0.15 V more negative than the half-wave or peak potential of the analyte. Consequently, E_{acc} should lie in the voltage range of the limiting diffusion current.

The potentiostatic principle of accumulation electrolysis enables analytes present simultaneously to be determined either individually or together by adjustment of the accumulation potential. Simultaneous determinations require a difference of peak potentials $\Delta E_p > 100$ mV.

Those metals which can be determined are soluble in mercury and form amalgams (Pb, Cu, Cd, Sb, Sn, Zn, Bi, In, Mn, and Tl). Nobler metals such as Hg, Ag, Au, and Pt are also determinable by ASV using inert solid electrodes (metal or carbon electrodes; Chap. 3).

Metal ions such as arsenic(III), selenium, and tellurium (Me_a^{n+}) may be determined by stripping voltammetry after adding a second metal, such as copper (Me_b^{m+}), to the test solution and co-electrolysing the two metals onto the surface of the HMDE. The copper acts as a co-deposition agent and facilitates the deposition of the analyte, Me_a, on the electrode surface as an intermetallic compound. The analyte may then be stripped from the electrode either by oxidation (ASV) or by further reduction (CSV) to an anionic species according to the following reaction scheme.

$$M_a^{n+} + M_b^{m+} + (m+n)\,e^- + (Hg) \longrightarrow M_a^0 M_b^0{}_{(Hg)} \qquad \text{Accumulation}$$

$$M_a^0 M_b^0{}_{(Hg)} \xrightarrow{CSV} M_a^{p-} + M_b^0{}_{(Hg)} - p\,e^-$$

$$M_a^0 M_b^0{}_{(Hg)} \xrightarrow{ASV} M_a^{n+} + M_b^{m+} + (Hg_2^{2+}) + (m+n+2)e^- \qquad \text{Stripping}$$

Arsenic, selenium, and tellurium are three such elements which may be determined by cathodic stripping voltammetry after having been reductively co-deposited with copper [38]. A characteristic of the cathodic stripping voltammograms of these three elements is that only a single current peak, which arises from the further reduction of the deposited analyte to As^{3-}, Se^{2-} or Te^{2-}, respectively, is observed. In this case the determination by CSV is more selective than by ASV, since in the anodic dissolution, additional current signals are obtained, which arise from the oxidation of the copper and possibly also of mercury [35]. Typical stripping voltammograms for the determination of arsenic by both ASV and CSV after deposition from a copper containing solution are shown in Figure 20. The determination limits for the CSV determination of As, Se or Te in the presence of copper were found to be 0.5 µg/L for As and 0.2 µg/L for Se and Te. Note that only As(III) is determined since under these experimental conditions As(V) is not electroactive. The small amount of arsenic present in water samples is mainly in the + 5 oxidation state as a result of oxidation by oxygen. If the test solution contains D-mannitol, As(V) can be reduced electrolytically to As(0). This is the basis of a method for

determining As(V) in the presence of As(III) by cathodic stripping voltammetry [36].

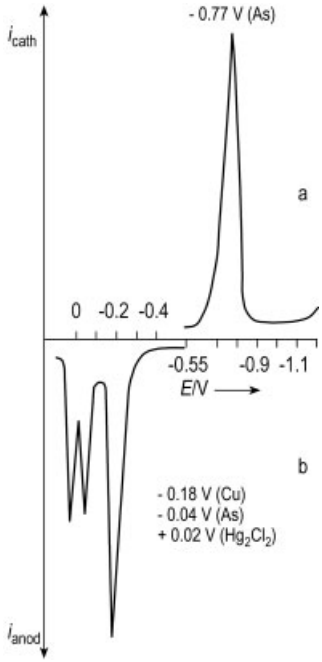

Figure 20. The differential pulse cathodic (a) and anodic (b) stripping voltammograms of arsenic. The cell solution was 0.1 mol/L HCI containing 2×10^3 mol/L Cu^{2+} and 10 µg/L As. Voltage scan rate = 100 mV/s, $E_{acc} = -0.55$ V; $t_{acc} = 1$ min

The technique of CSV can be applied to the determination of anions forming as accumulation products sparingly soluble Hg(I) salts on the electrode surface. These include the halides, pseudohalides, and oxometallates (vanadate, chromate, tungstate, molybdate).

During the subsequent cathode stripping step, the Hg_2^{2+} ion in the sparingly soluble compound is reduced with formation of a voltammetric current peak (Fig. 21). The peak height is proportional to the accumulated amount of the anion to be determined.

Organic substances can also be determined by this indirect procedure, provided they form insoluble Hg_2^{2+} or Hg^{2+} compounds. These are principally thiols, thioureas, thiobarbiturates, dithiocarbamates, and thioamides.

In adsorptive stripping voltammetry (AdSV) metal chelates and organic molecules are accumulated by adsorption at the surface of the working electrode. If these compounds are electrochemically active, i.e., if they are reducible or oxidizable, their subsequent voltammetric determination is possible. By this principle of so-called adsorptive stripping voltammetry, organic and organometallic compounds are determined in the ultra-trace range. This technique is particularly important for the trace analysis of metals that are not readily deposited as the element on mercury electrodes [13].

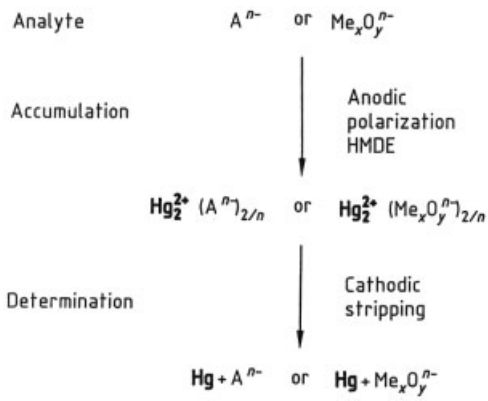

Figure 21. Determination of anions by CSV

Two approaches have been used to effect the adsorption of metal ions onto the electrode surface as the metal chelate. The simplest is to add an excess of a suitable complexing agent to the test solution prior to the accumulation step. This is the most common approach and is used preferably with mercury or glassy carbon working electrodes. A selection of frequently used complexing agents is listed in Table 1.

An alternative approach is to modify the surface of a glassy carbon, carbon paste or graphite electrode with the complexing agent. The metal ion is then accumulated by reaction with this modified surface (see page 23).

The potentiostatically accumulated species is determined voltammetrically by reduction or oxidation of the central atom or of the ligand of the metal chelate complex or via catalytic hydrogen evolution. Adsorbed organic molecules can be determined in an analogous way by the oxidation or reduction of their electroactive functional groups.

The quantity of the adsorbed species c_{ad} present after the accumulation time t_{acc} on the surface of the mercury drop area A and with the radius r determines the value of the voltammetric peak current $i_{p.}$ Moreover, c_{ad} is propor-

tional to the concentration c_a of the analyte in solution for low surface coverage. For diffusion-controlled processes, which are assumed here in order to simplify the relationships, the following dependence results for the peak in the stripping voltammogram:

$$i_p = kAc_{ad} = kAc_a \left(\frac{D}{r} t_{acc} + 2\sqrt{\frac{D}{\pi}} t_{acc}^{1/2} \right) \quad (26)$$

The peak current increases linearly with $t_{acc}^{1/2}$ until the electrode surface is saturated, and reaches its maximum value at:

$$i_{p(max)} = kAc_{ad(max)} \quad (27)$$

$c_{ad(max)}$ being proportional to the accumulation time $t_{acc(max)}^{1/2}$, and corresponds to the maximum surface coverage. The linear proportionality

$$i_p \sim c_a t_{acc}^{1/2} \quad (28)$$

applies in AdSV only for the lower and middle µg/L range and for short accumulation times $< t_{acc(max)}^{1/2}$. In routine AdSV practice, departures from linearity at fairly high analyte concentrations are controlled, either by dispensing with the usual stirring of the sample solution during accumulation or by dilution of the sample solution. The voltage dependence of the adsorption of neutral molecules is obviously connected with the interaction studied by A. N. Frumkin between the adsorption energy of a surface-active molecule and the value of the electrode potential determined by the surface charge. The zero charge potential in the electrocapillary curve of the mercury is critical for the adsorption of neutral molecules. Negatively charged surface-active ions are chiefly adsorbed at potentials corresponding to the positive branch of this curve, and the cations at potentials that lie in the negative range (see textbooks on electrochemistry).

Stripping Chronopotentiometry (SCP). With stripping chronopotentiometry, as with stripping voltammetry, the preliminary electrolytic accumulation of the analyte significantly increases the sensitivity of the technique. The difference between the two stripping techniques is that, in stripping chronopotentiometry the change of potential of the working electrode with time during the stripping of the electrolysis product (e.g., of a deposited metal) is recorded. Stripping is performed by application of an anodic current. The output is an "inverse" potential – time curve from which the transition time τ is obtained, as this is a function of concentration, analogous to the determination of the peak current i_p in the stripping voltammogram. From a comparison of the two curves in Figure 22 it can be seen that $E_{\tau/4}$ in the stripping chronopotentiogram corresponds to the peak potential E_p in the stripping voltammogram.

Table 1. Complexing agents used for the determination of metal ions by adsorptive stripping voltammetry

Complexing agent	Element
Reduction of the central metal ion	
1,2-Dihydroxybenzene (catechol)	U, Cu, Fe, V, Ge, Sb, Sn, As
1,3-Butanedionedioxime (dimethylglyoxime)	Co, Ni, Pd
8-Hydroxyquinoline (oxine)	Mo, Cu, Cd, Pb, U
2-Hydroxy-2,4,6-cycloheptatriene (tropolone)	Mo, Sn
2,5-Dichloro-3,6-dihydroxy-1,4-benzoquinone (chloranilic acid)	U, Mo, Sn, V, Sb
N-Nitroso-*N*-phenylhydroxylamine (cupferron)	U, Mo, TI
Reduction of the ligand	
o-Cresolphthalexone (OCP)	Ce, La, Pr
5-Sulfo-2-hydrobenzeneazo-2-naphthol (Solochrome Violet RS - SVRS)	Al, Fe, Ga, Ti, Y, Zr, V, TI, Mg, alkali and alkaline earth metals
1,2-Dihydroxyanthraquinone-3-sulfonic acid (Alizarin Red S)	Al
Chromazurol B (MB9, Mordant Blue 9)	Th, U
Catalytic hydrogen evolution	
Formazone	Pt

For the accumulation step, the HMDE and the TMFE are mainly used. The transition times for determinations by stripping chronopotentiometry at these electrodes are approximately:

$$\tau_{HMDE} = \frac{n \cdot F \cdot c_{amalg} \cdot r_d}{j_0} \quad (29)$$

and

$$\tau_{TMFE} = \frac{nFc_{amalg} \cdot l}{j_0} \quad (30)$$

Here, c_{amalg} is the amalgam concentration after accumulation, r_d the radius of the mercury drop, l the thickness of the mercury film on the electrode surface, and j_0 the current density during the stripping process.

In addition to the time for which the electrolysis potential is applied during the accumulation step, the current density during stripping is also

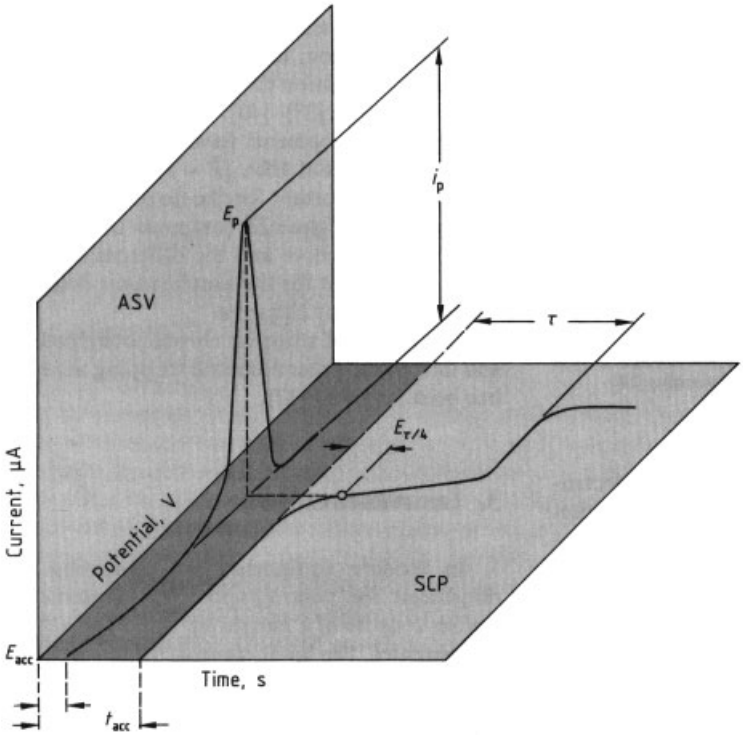

Figure 22. Comparison of anodic stripping voltammetry and stripping chronopotentiometry

of crucial importance for the limits of detection: as j_0 decreases, the determinations become more sensitive. Limits of detection of ca. 10^{-8} mol/L can be achieved [14].

Figure 23 illustrates the form of the curve for the determination of Cd, Tl, and Pb by stripping chronopotentiometry. The normal potential – time curve (A) is compared with the second derivative (B), from which the transition times can be more simply and accurately obtained. A TMFE coated in situ is used for the accumulation.

For the highest possible sensitivities in stripping chronopotentiometry, and to avoid systematic errors, the sample solutions must not contain oxidizing impurities. Care must therefore be taken to remove the oxygen present in the solution.

This precondition does not apply if the anodic stripping of the accumulated elements is replaced by chemical oxidative stripping. With this variant the technique is known as *Potentiometric Stripping Analysis* (PSA). The working electrode chiefly used for this is the TMFE; the oxidants are Hg^{2+} salts or oxygen dissolved in the sample.

Potentiometric stripping analysis is carried out in several stages. After electrochemical generation of the mercury film on a graphite substrate, the elements to be determined are accumulated by electrolysis at constant potential. The next stage is the oxidation of the deposited elements by the oxidant present in the solution. For this, the current circuit is disconnected. The deposited analytes are stripped in the order of their electrochemical potentials. Anodically deposited precipitates can similarly be stripped by chemical reduction. In all cases, potential – time curves with transition times proportional to concentration result [39–41].

Computer-assisted differential PSA $[E = f\,(\mathrm{d}t/\mathrm{d}E)]$ has become important for the detection of trace elements [42]. Figure 24 compares the normal potential – time curve and the differential curve calculated from it for the simultaneous determination of the four elements.

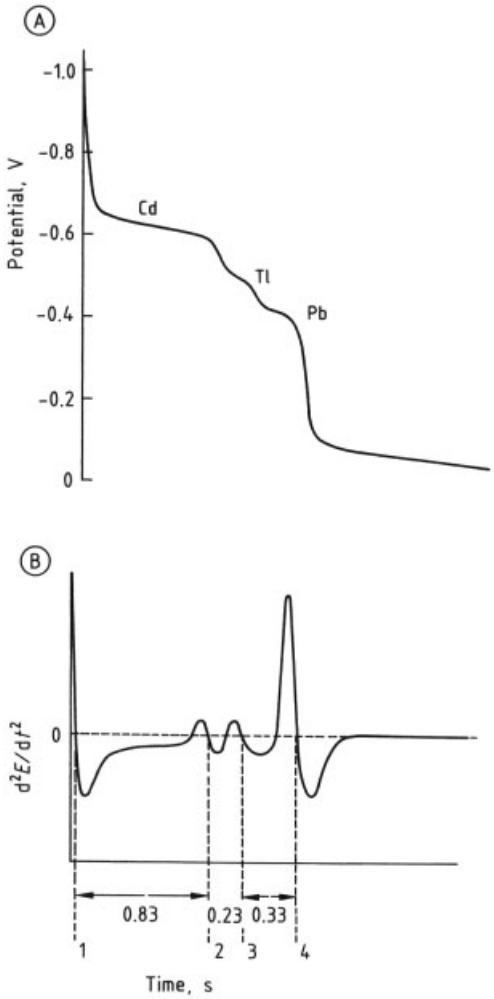

Figure 23. Simultaneous determination of Cd (5×10^{-7} mol/L), Tl (2×10^{-7} mol/L), and Pb (3×10^{-7} mol/L) in 0.05 mol/L HCl + 0.5 mol/L NaCl + Hg^{2+} 100 mg/L by stripping chronopotentiometry at a TMFE (electrolysis time 10 min; stripping current strength 10 μA) A) Potential – time curve; B) Second derivative [14]

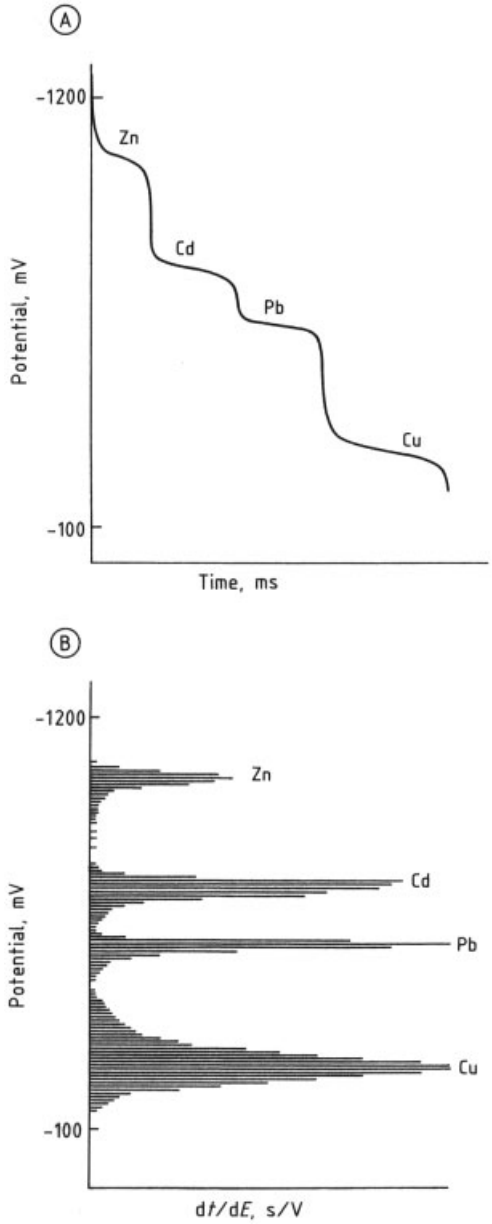

Figure 24. Potential – time curve (A) and differential plot (B) for the determination of zinc, cadmium, lead, and copper by PSA [42]

The theory of stripping chronopotentiometry and differential potentiometric stripping analysis has been reviewed [43].

3. Instrumentation

In modern equipment, the measuring arrangement for polarographic or voltammetric studies consists of a measuring stand and a digital computer. The latter controls the more or less auto-

mated techniques and also analyzes the current signals.

The measuring stand includes the potentiostat locked into analog electronics, the current measurement amplifier, and the digital – analog and analog – digital converters. It also provides a holder for the electrochemical cell, which consists of measuring vessel, electrodes, stirrer, etc.

The *potentiostat* is the voltage source for the DC supply (ca. 4 V) as well as for the superimposition of voltage pulses (NPP and DPP modes) and a sine-wave or square wave alternating voltage (ACP and SWP mode). The basic circuit diagram is shown in Figure 25.

When recording the voltammograms or polarograms using three electrodes, the current is flowing between the working electrode and auxiliary electrode. Compared with the working electrode, the auxiliary electrode has a much greater surface area and thus a smaller current density. The reference electrode is an electrode of the second kind, and the potential of the working electrode is measured against this reference electrode. Potential data for voltammetric and polarographic measurements are reported relative to calomel or silver – silver chloride reference electrodes.

The potentiostatic measuring arrangement with three electrodes (Fig. 25) has superseded the classical measuring arrangement with two electrodes, as with the latter the ohmic potential drop in the sample solution cannot be compensated adequately.

In addition to carbon and the noble metals, the principle material used for the working electrode is mercury, especially for capillary and film electrodes. Because of the high overvoltage of hydrogen on mercury, the latter electrodes ensure a relatively wide potential range for voltammetric/polarographic determinations in various supporting electrolyte solutions. If capillary electrodes are used, the electrode surface can be renewed easily and reproducibly as with the DME, the SMDE and the HMDE.

Dropping Mercury Electrode. The DME, originally proposed by HEYROVSKÝ and used for polarographic studies, consists of a glass capillary, 10 – 20 cm long, with internal diameter 0.04 – 0.1 mm, and is connected by a tube to the mercury storage vessel. The distance between the capillary opening and the mercury level determines the drop time and influences to different degrees the types of current useful for polarographic determinations [1]. The drop time also depends on the electrode potential and the surface tension of the mercury. The surface tension is mainly influenced by the medium, and particularly the solvent, into which the mercury flows. The drop time must be held constant during the measurement and should be 2 – 8 s. In modern instruments the mercury drop is knocked off mechanically by tapping on the capillary at a predetermined constant time.

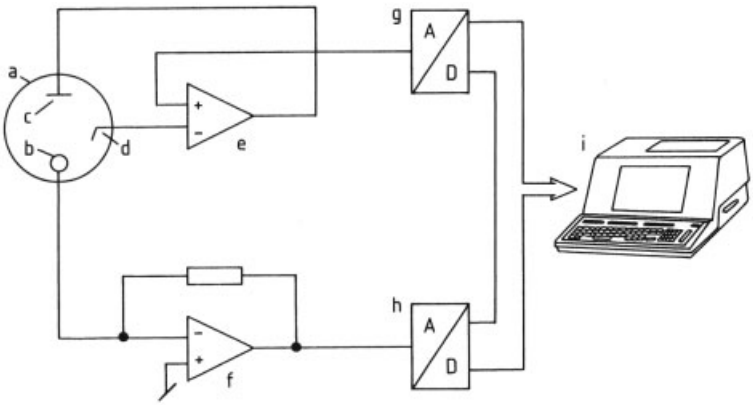

Figure 25. Circuit diagram of a polarograph for various polarographic/voltammetric measuring techniques (Courtesy of Metrohm)
a) Measuring cell; b) Working electrode; c) Auxiliary electrode; d) Reference electrode; e) Potentiostat; f) Current measurement amplifier; g) Digital – analog converter; h) Analog – digital converter; i) Computer

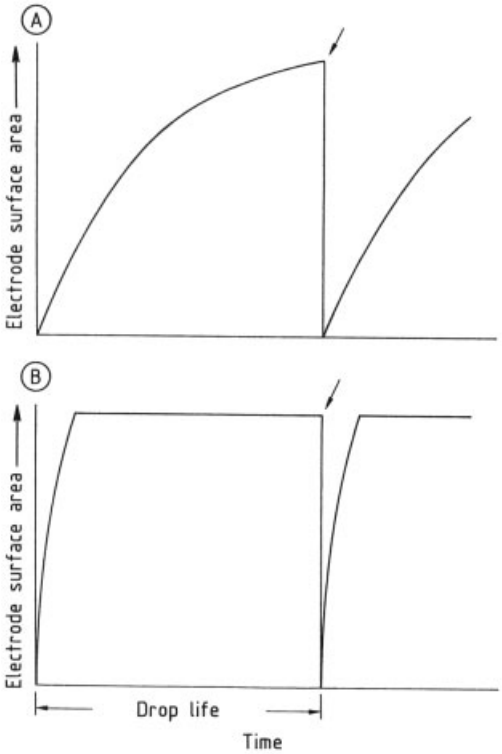

Figure 26. Drop growth as a function of time; comparison of the DME (A) with the SMDE (B)
Drop fall indicated by arrows

Static Mercury Drop Electrode. The static mercury drop electrode (SMDE) is the preferred electrode for polarographic analysis. Compared with the DME, the drops from this type of electrode can be renewed more frequently, about once or twice per second. A microvalve with opening times of 20 – 200 ms controls the mercury inflow to the capillary. During each open period a mercury drop forms, whose surface area remains constant after the valve closes.

In Figure 26 the increase of the electrode surface area with time during the drop life is compared for the DME and the SMDE. The maximum electrode surface area is reached much more rapidly with the SMDE than with the DME so that the measurement can be carried out at constant electrode surface area. This results in a reduction of the capacitive current and increased analytical sensitivity.

The greatest advantage of this new type of electrode is its ability to function as a stationary and a nonstationary mercury electrode, as well as the possibility of synchronizing the electrode (drop) formation with the programm-control arrangement. The reproducibility of the mercury metering of an SMDE is important for its dependability in operation. In a commercial model (Multi-Mode electrode, Metrohm AG, Switzerland), the mercury is pushed through the capillary by a nitrogen overpressure and a needle valve controls the mercury flow.

Figure 27 shows the construction of the pneumatically controlled Multi-Mode electrode. The electrode is suitable for all voltammetric/polarographic work for which a DME, SMDE, or HMDE is required.

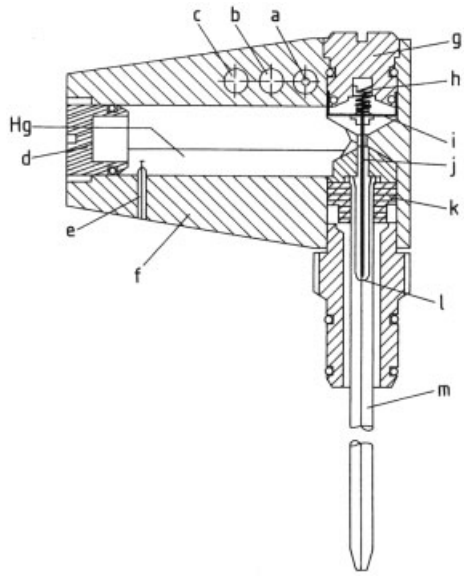

Figure 27. Construction of a static mercury electrode (Courtesy of Metrohm)
a) N_2 supply for membrane control; b) N_2 supply for overpressure; c) Blind hole; d) Stopper for topping up; e) Electrode connection; f) Hg reservoir; g) Adjusting screw; h) Steel spring; i) PTFE membrane; j) Steel needle; k) Rubber packing (damper); l) Seal; m) Capillary

Thin mercury film electrodes (TMFE) are used chiefly for ultra-trace analysis by stripping voltammetry. Glassy carbon has proved itself to be the best substrate onto which a mercury film can be formed. The mercury is deposited by electrolysis, either in a separate operation before the voltammetric determination or, in the case of stripping voltammetry, together with the amalgam-forming metal during the accumulation by electrolysis (in situ coated TMFE, Section 2.6; the mercury film generated should be

1 – 100 µm thick). Voltammetric determinations with the TMFE can be more sensitive than with the HMDE, with improved resolution of neighboring peaks.

Carbon and graphite electrodes are employed principally in the anodic potential range. Glassy carbon electrodes (GCE) and carbon paste electrodes (CPE) are mostly used and are available commercially. They can be used either as stationary or rotating electrodes. *Glassy carbon electrodes* consist of pins or rods 2 – 8 mm diameter, cemented or pressed into a glass or plastic holder. *Carbon paste electrodes* consist of mixtures of spectroscopically pure carbon or graphite powder and water-insoluble, highly viscous organic liquids, e.g., nujol, silicone oil, or naphthalene derivatives. The mixture is ground to a paste, transferred to a glass tube, and expressed from the tube with a piston. The surplus carbon paste must be rubbed off and the surface polished. In voltammetric studies the composition of CPE influences the peak potential and the peak current. Glassy carbon electrodes and carbon paste electrodes are also used as working electrodes in voltammetric detectors for liquid chromatography (→ Liquid Chromatography).

Noble Metal Electrodes. Working electrodes of noble metals are often used for studies in the anodic potential range, or for determining elements with which mercury electrodes interfere (e.g., voltammetric Hg determination at an Au electrode). The electrochemical properties of noble metal electrodes and the reproducibility of the analytical values obtained with them depends on their physical and chemical pretreatment. In comparison with mercury and carbon electrodes, noble metal electrodes are of secondary importance for voltammetric determinations.

Modified Electrodes. When the surface of a working electrode has been chemically changed or treated in order to change its electrochemical properties or to improve its sensitivity and selectivity in an analytical determination, the electrode is referred to as a *modified electrode*. The modification of a voltammetric working electrode, usually a carbon electrode, is done by fixing (immobilizing) reagents on the electrode surface. The reagent that is used to modify the surface is referred as the *modifier*. One of the reasons for the development of modified electrodes is to use them to replace mercury electrodes (for environmental reasons) and so facilitate the further development of voltammetry, and in particular stripping voltammetry, as a "leading edge" technique. A major role of the modifier is to improve the selectivity of the electrode and aid the binding and enrichment of the analyte on the electrode surface.

There are several methods used to modify electrode surfaces: adsorption of reagents on the surface, covalent bonding by reaction between the modifier and functional groups on the electrode surface, incorporation of the modifier within a gel or an electrically conducting polymer film, physically coating the electrode surface with the modifier, e.g., an enzyme, or mixing the modifier with carbon paste.

Carbon paste electrodes have the advantage that the surface can be quickly renewed by simply removing the used surface layer with a tissue or knife. Organic complexing agents, ion exchangers, solid phase extractants and biological materials may be used as modifiers. The modifiers must have a low solubility in the analyte solution and not be electroactive in the working potential range [44].

Modified screen printed electrodes may be prepared either by incorporating the modifier in the carbon containing "ink" or by fixing it to the screen printed electrode surface. A mercury film, electrolytically deposited on a screen electrode has been used to determine heavy metals by ASV or AdSV (e.g., uranium [45]).

Thick-film graphite electrodes can be modified in a similar way to screen printed electrodes to give *thick film modified graphite electrodes* (TFMGEs). After modifying the surface in different ways the electrodes can be used for the stripping voltammetric determination of a number of heavy metals [46].

Microelectrodes. In general, the carbon and mercury working electrodes commonly used in voltammetry have surface areas in the mm^2 range. This means that the linear dimensions of these electrodes are an order of magnitude greater than the thickness of the diffusion layer and the transport of the analyte to the electrode takes place mainly by linear diffusion. However, if the lateral dimensions of the electrode

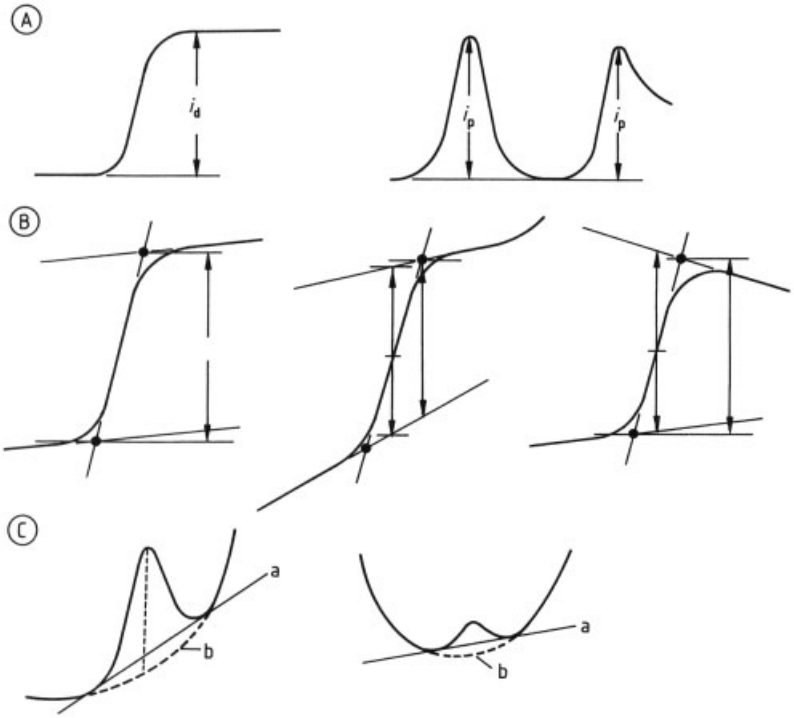

Figure 28. Interpretation of polarograms and voltammograms
A) Ideal conditions: linear trend of residual current; B) Evaluation of wave heights under nonideal conditions; C) Evaluation of peak currents with strongly curved residual current.
a) Tangent method; b) Evaluation with French curve [5]

are similar to or less than those of the diffusion layer (< 0.1 mm), the diffusion of the analyte to the electrode is then spherical rather than linear. Electrodes with such small dimensions in the lower µm range are referred to as *microelectrodes*. Microelectrodes are becoming more important as working electrodes in voltammetric trace analysis.

The miniaturization of the working electrode results in a higher current density at the electrode surface, but also in small currents flowing in the electrochemical processes. The small area means that the charging current i_c, which is proportional to the area of the electrode (Eq. 16), is also small. Thus as the electrode size decreases, the signal to noise ratio (or sensitivity) increases. In addition, note the very fast response to potential changes (response time) at a microelectrode.

A method of overcoming the problem of measuring the small currents is to join many microelectrodes in parallel and so obtain a much larger current output. An arrangement with many (up to 10 000) single microelectrodes is called a *microelectrode array*. By wiring n single electrodes in any array together in parallel, the output current will be n times that of a single microelectrode, i.e., it can be in the μA range instead of in the pA – nA range. Such an array, while having the advantages of a single microelectrode, produces a current which is large enough to be measured with conventional polarographic equipment.

With microelectrode arrays (mainly made from platinum or carbon by various techniques) it is possible to extend the application of analytical voltammetry to the direct determination of analytes in non-aqueous samples without adding a supporting electrolyte. In stripping voltammetry, the accumulation of the analyte on the electrode can be achieved in a very short time without stirring the solution because of the higher mass transport [47].

Reference and Auxiliary Electrodes. The function of the reference electrode is to monitor the potential of the working electrode and the function of the auxiliary electrode is to complete the electrical circuit for the current generated at the working electrode. The voltage of the auxiliary electrode (Pt wire or graphite rod) with respect to the working electrode is electronically adjusted via the potentiostat to the desired voltage plus the *iR* drop in the solution so that the potential difference between the working and reference electrode is maintained at the required value. Usually the area of the auxiliary electrode is at least 50 times greater than the working electrode.

The potential of the reference electrode must be constant and independent of the composition of the solution. Therefore the reference electrodes used in polarographic and voltammetric practice are electrodes of the second kind, of which the silver – silver chloride electrodes (with potassium chloride of a defined concentration, for example Ag/AgCl, 3 M KCl) are the most important. Previously calomel electrode (for example the saturated calomel electrodes, SCE) were used.

Measuring Cells. The commonest cells have a maximum volume of 25 mL. The vessels are generally made of glass; for demanding trace analyses they consist of quartz or plastic (principally polytetrafluoroethylene, PTFE). Micro-cells with a capacity of 1 mL or less are used for special microanalytical purposes and are constructed individually. Flow cells for analyses in connection with liquid chromatography or for flow injection analysis (FIA) are becoming increasingly important.

4. Evaluation and Calculation

Voltammograms and polarograms are evaluated either graphically or by calculation. For accurate determination of the current value – which is proportional to concentration – from the current wave or from the current peak, the residual current must be subtracted from the signal. For routine analysis, it is often sufficient to interpolate the residual current in the potential range of the voltammetric/polarographic current signal.

Figure 28 compares various practical curves and the methods for their manual interpretation. Such current – voltage curves, and others, can be evaluated more rapidly and with improved accuracy with computer-controlled instruments. The best values are obtained when a polynomial describing the approximate shape of the baseline is used. The coefficients of the polynomial are determined from the experimental points which lie on both sides of the current peak [4].

This evaluation becomes straightforward with equipment in which the current – potential curves are recorded digitally at a high data rate. The first step in the data processing consists of the automatic search for stray experimental data points. These are then eliminated and replaced by the mean value of both neighboring data points. The line of best fit of the current – potential curve is then smoothed within a sliding window along the voltage axis. The smoothing factor that can be applied essentially depends on the number of data points available. The principle of the curve smoothing is illustrated in Figure 29.

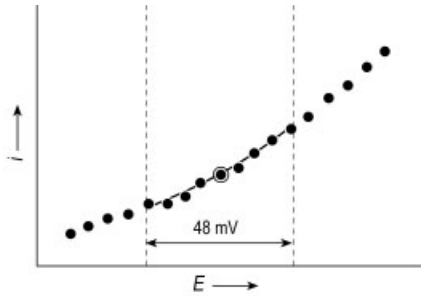

Figure 29. Curve smoothing using a voltage window with 9 data points

After smoothing, the curve is automatically differentiated (first derivative) and in this form used for peak recognition. This is achieved by seeking out successive maxima and minima in the differentiated curve (Fig. 30). In the usual voltammetric presentation (increasing negative potentials to the right), a maximum followed by a minimum indicates a reduction peak, whereas the reverse order indicates an oxidation peak. The potentials of the maximum and minimum in the d*i*/d*E* versus *E* plot are noted and the peak width is the difference between these values while the peak potential is the mid-point between them. The qualitative identification of the

analyte is achieved by comparing the peak potential (± a pre-set tolerance) with data stored in the computer memory. For a peak to be identified it must be greater than a pre-determined minimum peak height. After the peak is recognized a baseline is constructed and its value at the peak potential is subtracted from the maximum value of current to give the peak height. The result is presented as a peak or wave height, or as the peak area, all of which are proportional to the concentration of the analyte.

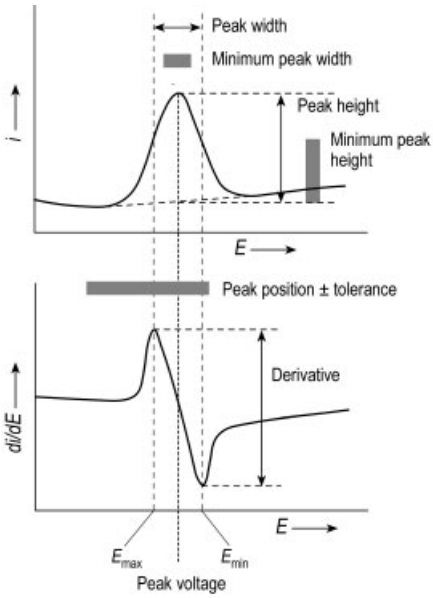

Figure 30. Peak recognition and characterization (Courtesy of Metrohm).
Peak potential = $(E_{max} + E_{min})/2$
Peak width = $E_{min} - E_{max}$

When the baseline is curved the polynomial which approximately describes this baseline is used to establish it mathematically over the potential region across the base of the peak. This potential region is determined from the potentials at which the change in the gradient (becomes more positive) of the current – potential curve occur. The same procedure is used for the evaluation of closely adjacent peaks.

The peak or wave height (h) or peak area (A) calculated by the above procedure can be converted into the desired concentration units either by a *calibration curve* (c versus h or c versus A) or by the method of *standard additions.* When using a calibration curve its slope and linearity must be checked at least daily, also when experimental parameters are changed.

When using the method of standard additions, an aliquot, v, of a standard solution of known analyte concentration c is added (manually with a micropipette or automatically with a motorized burette) to a Volume V of the sample solution in which the analyte concentration is c_a. If h is the peak or wave height of the sample before the standard addition, and H the height after that, then the analyte concentration is given by

$$C_a = hcv/H\,(V + v) - Vh \tag{31}$$

The highest accuracy can be obtained by using several (at least three) standard additions, provided the total concentration always lies within the linear range of the standard curve, when plotting h versus c, and extrapolating the straight line plot back to the c axis. The distance from the origin to the intercept corresponds to c_a.

The reproducibility of voltammetric/polarographic determinations depends not only on the evaluation, but also on the quality of the electrode and the measurement setup. With modern equipment, reproducibility is 0.5 – 2.0 %.

A characteristic source of error is the temperature dependence of voltammetric currents. All equations for the proportionality of polarographic currents to concentration contain the diffusion coefficient, which increases with temperature. The rate constants of preceding and succeeding chemical reactions are also affected by temperature and this must be taken into account when evaluating kinetic and catalytic currents. Electrode reactions that are associated with analyte adsorption processes likewise show a specific temperature dependence. The temperature influence differs and leads to different temperature coefficients for the individual voltammetric methods [48].

In simple diffusion-controlled electrode reactions, deviations between + 1.5 and + 2.2 %/ °C occur. For kinetic and catalytic currents, the temperature coefficients can be up to + 70 %/ °C. Currents that are based on adsorption processes can have negative temperature coefficients, because of decreasing adsorption with increasing temperature.

5. Sample Preparation

Before performing the voltammetric/polarographic determination, in some cases the analyte must be separated from the matrix and from other interfering components, and converted into a measurable form. For total element determinations, sample digestion is normally required. To separate organic analytes from the matrix, extraction methods and chromatographic separation are mainly used.

Heavy metals in association with organic materials are usually in a bound form, from which they must be liberated prior to analysis. This is carried out by dry or wet ashing, whereby the organic matrix must be completely destroyed. To avoid systematic errors, care must be taken in all forms of sample treatment to ensure that the elements to be determined remain completely in the sample and that sample contamination from digestion reagents and vessels is avoided.

Wet digestions are carried out only with high-purity acids. If there is a risk that an element will volatilize, the treatment is carried out under reflux, or better in pressure digesters with quartz or plastic inserts; heating is carried out electrically or by microwave radiation. Sample digestion with nitric acid at high temperatures is favored owing to the increased oxidation power. Figure 31 shows a high-pressure ashing apparatus for the digestion of organic and inorganic samples with acid (HNO_3, HCl, $HClO_4$, H_2SO_4, HF, or mixed acids) in sealed quartz vessels under high pressure (up to 13 MPa) and high temperature (up to 320 °C). The digester is contained in a heated autoclave and filled with nitrogen (ca. 10 MPa) in order to compensate the internal pressure resulting from the digestion reaction.

For the digestion of biological samples, combustion in oxygen (*dry ashing*) can also be useful under suitable conditions. The use of oxygen of highest purity ensures low contamination digestion. Some forms of equipment operate with reflux condensation to avoid losses of elements by volatilization; cooling with liquid nitrogen is particularly suitable. Organic samples can be digested at relatively low temperature in high-frequency oxygen plasmas.

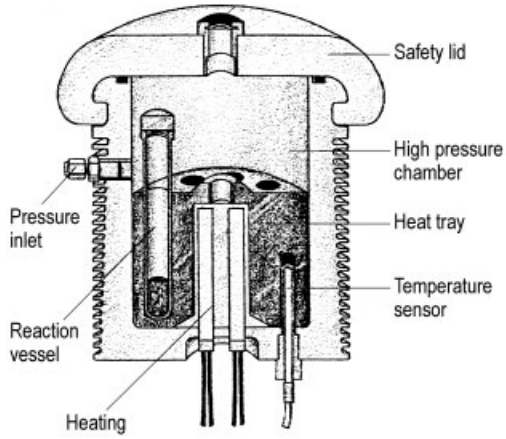

Figure 31. High-pressure ashing apparatus HPA-S (Courtesy of Perkin Elmer)

Voltammetry and polarography are particularly important for the analysis of water samples for heavy metals. In surface waters with organic contamination, they can only be analyzed after digestion. Sampling and storage is of prime importance for correct analytical results.

Water samples are stored in flasks of polyvinyl chloride (PVC) or PTFE, cleaned with 50 % nitric acid before use. The flask is subsequently pre-rinsed with the sample water. Contaminated samples are filtered through membrane filters directly into the sample flask and acidified, e.g., with high purity nitric acid to pH 1 – 2; the filter residue is analyzed separately. Water samples with little or no organic contamination are mixed with the chosen supporting electrolyte and immediately placed in the voltammetric/polarographic set-up.

Contaminated water samples are prepared for electrochemical heavy metal trace analysis either by UV photolysis or by oxidative microwave digestion.

Ultraviolet photolysis is based on the generation of OH radicals, which react with organic compounds and degrade them. Hydrogen peroxide is used as an initiator. The acidified water sample plus hydrogen peroxide is irradiated in quartz vessels. An arrangement for the digestion of one to six samples is shown in Figure 32.

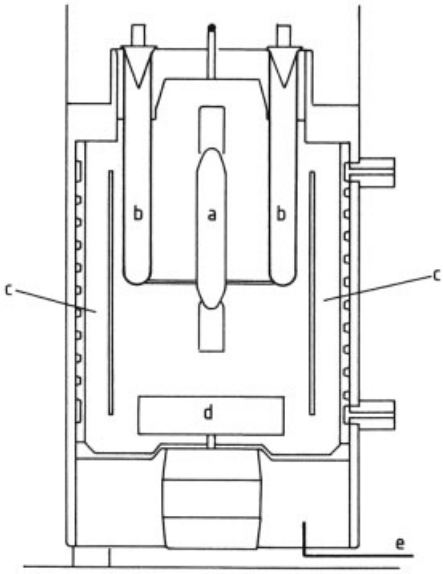

Figure 32. Ultraviolet digester (Courtesy of Metrohm) a) Lamp; b) Digester vessels with stopcocks; c) Cooling chamber with water inlet and outlet; d) Fan; e) Cables

When nitrate-containing samples are UV-irradiated, nitrite is formed by reduction. The voltammetrically active nitrite causes a broad current – voltage peak that can cause interference (e.g., with the determination of zinc). This must be taken into account when examining water samples acidified with HNO_3. Another acid should be used or the nitrite formed should be reduced to nitrogen with sulfamic acid (NH_2SO_3H) immediately after the end of the photolysis.

The complete destruction of organic substances in water samples by UV photolysis depends on the nature and extent of the contamination, and on the intensity of the irradiation. The irradiation time required for waste water samples is usually 1 h (see Fig. 33).

Microwave Digestion. Oxidative microwave digestion is suitable for the rapid degradation of interfering organic substances in surface waters. For this purpose the samples are heated with the oxidizing agent Oxisolv (Merck, Germany) in sealed PTFE vessels in a microwave oven. During the thermal treatment, the pH value of the sample passes through the wide range of pH values (from 11 to 2). Wastewater samples are generally completely digested after 50 – 60 s, so that the voltammetric/polarographic metal trace analysis can be carried out completely free of interference.

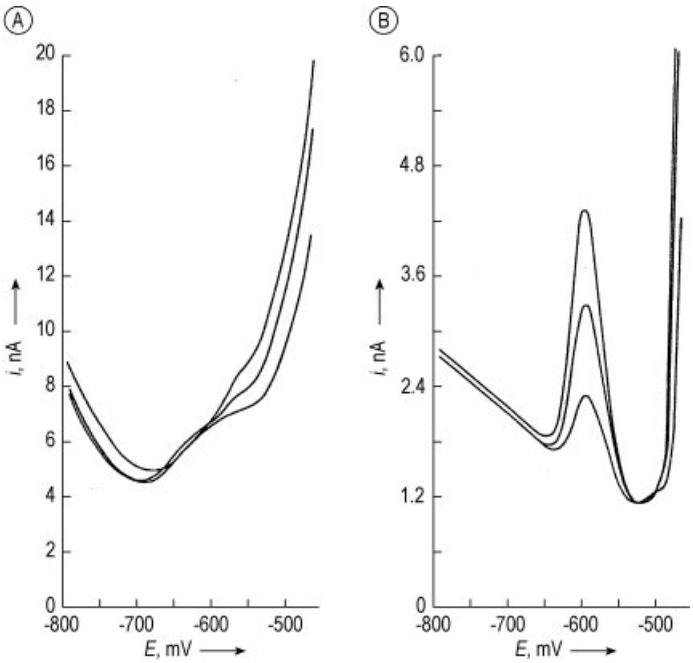

Figure 33. Determination of Cd in municipal wastewater by DPASV before (A) and after (B) 1 h UV irradiation [49]

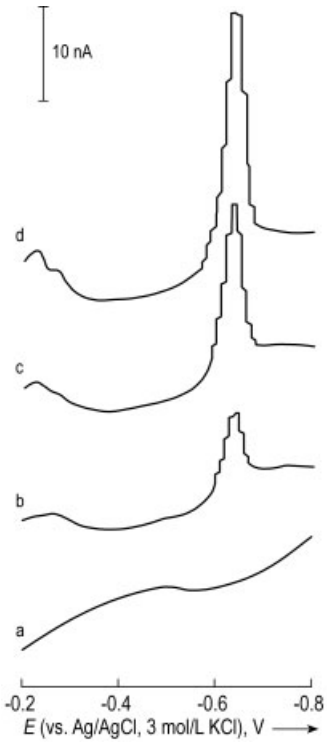

Figure 34. Adsorptive stripping voltammograms for Mo in river water after microwave digestion [50]
a) Untreated sample; b) After microwave digestion + Oxisolv; c) + 2 ng/mL Mo; d) + 4 ng/mL Mo Conditions: 10^{-3} mol/L chloranilic acid, pH 2.7, stirred 1 min, accumulation at −0.2 V
Calculated sample concentration of Mo 1.82 ng/mL

If the metals are determined by the highly sensitive method of adsorptive stripping voltammetry, the digestion time must be increased by 1 – 2 min. Figure 34 shows the voltammograms for the determination of molybdenum in a river water sample; the digestion time was 2×50 s at 260 W microwave power [50].

6. Supporting Electrolyte Solution

The solution for the voltammetric/polarographic determination of elements and organic compounds always contains a supporting electrolyte in order to minimise the migration current and reduce the ohmic voltage drop. The supporting electrolytes used are mainly chlorides, nitrates, and sulfates of Li, Na, K, perchlorates of Li, Na, and salts of tetraalkylammonium bases ($NR_4^+X^-$; R = methyl, ethyl, butyl; X = Cl, ClO_4, Br, I). Moreover, acids (HCl, H_2SO_4), bases (LiOH, NaOH, $NR_4^+OH^-$), and frequently buffer solutions are used as they simultaneously provide for regulation of pH during electrode reactions. Supporting electrolyte concentrations are commonly between 0.1 and 1 mol/L.

If complexation occurs between the analyte and the supporting electrolyte or an added ligand, crucial changes in electrochemical behavior can occur. As a result of complexation, the analyte can become electrochemically inactive, or its behavior may change in such a way that the value of the polarographic half-wave or peak potential becomes more negative. For the simple complexing reaction of a hydrated cation $M(H_2O)_x$ with ligand L and coordination number i:

$$M(H_2O)_x + i\,L \rightleftharpoons ML_i + xH_2O$$

and with the stability constant K of the complex:

$$K = \frac{[ML_i]}{[M]\,[L]^i}$$

the shift of the half-wave or peak potential is given by:

$$\Delta E_{1/2} = \Delta E_p = -\frac{0.059}{n}\log K - \frac{0.059}{n} i \log\,[L] \quad (32)$$

Stability constants and coordination numbers of complexes can be determined from the shift of the polarographic half-wave or peak potentials.

The complexing behavior of supporting electrolytes is analytically important for the resolution of overlapping current signals and for masking interfering sample components. Specific additions of suitable complexing agents are therefore used to improve the selectivity of polarographic determinations, as demonstrated in Figure 35 by the separation of the DPP peaks of Tl/Pb and Cd/In. Similar phenomena are known in multielement determinations by stripping voltammetry. By changing the composition of the supporting electrolyte solution after the simultaneous accumulation by electrolysis and use of suitable complexing agents, the metals are stripped at different potentials and can be determined from their separated peaks [5].

Simple ligands are used as complexing agents, e.g., CN^-, OH^-, NH_3, oxalate, halides, as well as stronger complexing agents, such as ethylenediaminetetraacetic acid (EDTA), citrate, tartrate, etc.

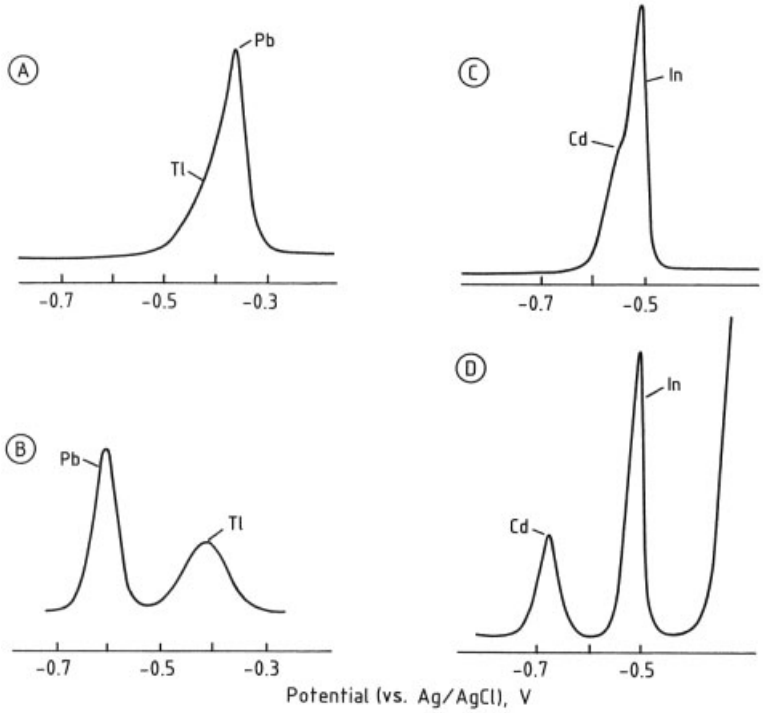

Figure 35. Influence of complexation on the peak separation in differential pulse polarographic curves [5]
10^{-4} mol/L Pb^{2+} and Tl^{+} in 0.1 mol/L KCl (A) and 0.1 mol/L KCl + 0.1 mol/L NaOH (B)10^{-4} mol/L Cd^{2+} and In^{3+} in 0.1 mol/L HCl (C) and 0.1 mol/L HCl + 2 mol/L KI (D)

A high salt content, which can arise after sample digestion or as a result of preceding separation operations, changes the viscosity η of the supporting electrolyte solution and hence the diffusion behavior of the analyte. With increasing salt content the polarographic currents become smaller and displaced to more negative potentials. The peaks in the stripping voltammograms are subject to the same influence, except that the potentials become more positive. Small amounts of inert salts, as can occur in natural samples, generally have no influence on polarographic or voltammetric determinations. Sometimes, as in the case of heavy metal determination in seawater by stripping voltammetry, electrolyte addition becomes superfluous.

The useful voltage range for voltammetric/polarographic determinations depends significantly on the supporting electrolye used. The reduction of the cation of the supporting electrolyte restricts the cathodic working range, as hydrogen evolution from acidic supporting electrolyte solutions depends on the working electrode material. The high overvoltage of H^{+} discharge on mercury is therefore particularly advantageous.

The anodic range is determined by oxidation of water or of the supporting electrolyte anion, as well as by oxidation of the electrode material. Anodic dissolution of mercury can even begin at ca. -0.4 V if anions that form poorly soluble Hg(I) salts are present.

Oxygen must be removed (by deaeration) from the supporting electrolyte solution before each voltammetric/polarographic determination. Oxygen is present in aqueous solutions at ca. 10^{-3} mol/L, depending on the solubility conditions, and is reduced at the working electrode in two stages. The course of the reduction is pH dependent and can be described by the following equations.

1) Reduction in acid solution:

$$O_2 + 2H^+ + 2e^- \longrightarrow H_2O_2$$
$$H_2O_2 + 2H^+ + 2e^- \longrightarrow 2H_2O$$

2) Reduction in alkaline solution:

$$O_2 + 2H_2O + 2e^- \longrightarrow H_2O_2 + 2OH^-$$
$$H_2O_2 + 2e^- \longrightarrow 2OH^-$$

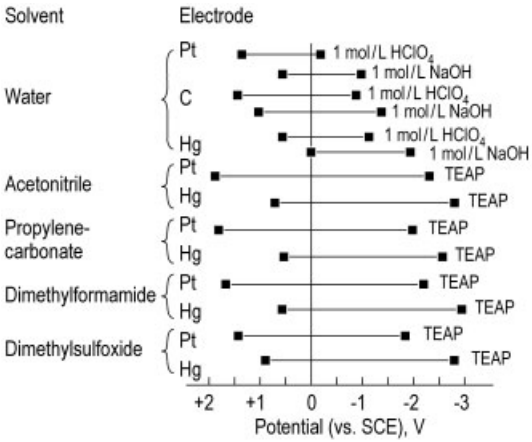

Figure 36. Polarographically usable voltage ranges with Hg, C, and Pt electrodes in various solvents and supporting electrolytes [16]
TEAP = Tetraethylammonium perchlorate

In a DC polarogram the first stage lies between − 0.1 and − 0.3 V, depending on the pH value, and the second stage occurs between − 0.7 and − 1.3 V.

The currents caused by the reduction of oxygen interfere with the current – voltage signals of many analytes. In addition, there is interference due to possible subsequent reactions of the H_2O_2 and of the OH^- ions formed.

The dissolved oxygen is removed by introducing pure nitrogen or argon for 5 – 10 min. Oxygen can also be removed from alkaline solutions by addition of Na_2SO_3.

Inorganic ions are determined almost exclusively in aqueous supporting electrolyte solutions. For organic and organometallic compounds, either aqueous supporting electrolyte solutions mixed with organic solvents or pure organic solvents are used (e.g., alcohols, acetonitrile, propylene carbonate, dimethylformamide, and dimethylsulfoxide). Figure 36 illustrates the potential ranges that are available when using these solvents for anodic and cathodic studies at various working electrodes.

Organic solvents are also used for the extraction of metal chelates, which can then be analyzed in the extract, directly or after adding a polar solvent (in order to raise the solubility of the supporting electrolyte). The procedure known as *extractive voltammetry/polarography* combines the selectivity of partition methods with the possibility of carrying out the determination in organic solvents for many analytes. Using this procedure, the analyte can be simultaneously accumulated and separated from interfering components [15].

7. Application to Inorganic and Organic Trace Analysis

Voltammetry and polarography are used primarily for the trace analysis of inorganic and organic substances. Because of the introduction of the powerful pulse methods and stripping techniques, and of advances in instrumentation, these measurement techniques are also of interest for routine analysis. With modern polarographs the most important voltammetric or polarographic methods can be selected by simple switching. Investigations are therefore possible with the same measuring instrument in a concentration range extending over at least six decades. Ions and molecules can be determined and the resulting signals give qualitative and quantitative information as well as providing insight into the bonding state of the analyte.

Elements and Inorganic Ions. The determination of elements and inorganic ions is based chiefly on reduction and oxidation processes that can take place in combination with kinetic, catalytic, or adsorptive processes. The reduction of simple or complexed ions of metals that are soluble in mercury is usually reversible and leads to readily interpretable current – potential curves. Examples of this are the polarographic determinations of cadmium, copper, thallium, zinc, and lead. Less important for routine analysis is the possibility of determining mercury, silver, and gold voltammetrically at noble metal or carbon electrodes.

Multivalent ions are often reduced in several steps, the mechanism and degree of reversibility of the electrode reaction in most cases being dependent on the composition of the supporting electrolyte solution. Two-step current – potential curves are obtained with

chromium(VI), copper(II), iron(III), molybdenum(VI), and vanadium(V), among others, when suitable supporting electrolyte solutions are used. In a multi-step process, the shape and height of the polarographic wave or peak associated with each step is determined by the number of electrons involved and the reversibility of that step. For analytical purposes, the best sensitivity is obtained when a supporting electrolyte can be found in which the multi-electron exchange occurs in one step at a single potential, resulting in a large single wave or peak in the polarogram. This is illustrated in the DP-polarographic reduction of Cr(VI) shown in Figure 37. In 0.1 mol/L KNO_3 supporting electrolyte (curve A), Cr(VI) is reduced in two steps – firstly to Cr(III) at -0.3 V and then to Cr(II) at -1.2 V. In 0.1 mol/L Na_2EDTA solution (curve B), only one peak is observed corresponding to the four electron reduction of Cr(VI) to Cr(II). Because the three-electron reduction to Cr(III) is irreversible the first peak in curve A is smaller than second peak for the reversible one-electron reduction of Cr(III) to Cr(II). The largest peak, obtained for the four-electron reduction of Cr(VI) to Cr(II) in EDTA solution (curve B), gives a determination limit of 10^{-6} mol/L for the determination of chromium.

For those systems which give well behaved catalytic currents, fairly sensitive polarographic and voltammetric analyses are possible. Examples are the determination of Cr(VI) or Mo(VI) in the presence of nitrate ions or of Fe(III) in the presence of bromate ions. For example, the determination limit for Cr(VI) using DPP can be lowered to ca. 10^{-8} mol/L when Na_2EDTA containing nitrate ions is used as the supporting electrolyte [62]. The Cr(II) formed by the reduction of Cr(VI) at the working electrode is immediately oxidized back to the trivalent state by the nitrate ions and is available to be reduced again in the electrode reaction. Titanium, tungsten, uranium and vanadium can also be determined via catalytic currents in the presence of nitrate or chlorate ions with limits of detection in the µg/L (ca. 10^{-8} mol/L) range [4], [5], [63].

Voltammetric/polarographic methods are suitable both for the simultaneous determination of several analytes and for the determination of single analytes in a sample solution. The possibilities are a consequence of the different potentials at which the peak currents occur and the selection of a suitable voltage range for the analysis. Superimposition or overlap of responses can be eliminated by using another supporting electrolyte. Complexing compounds that are expected to change drastically the redox behavior of the analyte or of the interfering component are suitable for the purpose.

The half-wave potentials determined for selected elements in various supporting electrolyte solutions are compared in Figure 38.

Voltammetric/polarographic methods are not very important for *anion analysis*. Only bro-

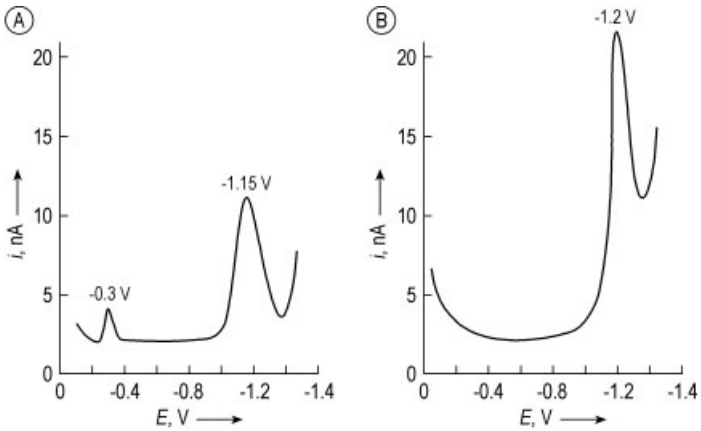

Figure 37. DP-polarogram of chromium(VI) in different supporting electrolytes
A) 0.5 mg/L Cr(VI) in 0.1 M KNO_3; B) 0.5 mg/L Cr(VI) in 0.1 M Na_2EDTA Electrode – SMDE; Potential scan from -0.1 V to -1.4 V (vs. Ag/AgCl, 3 M KCl)

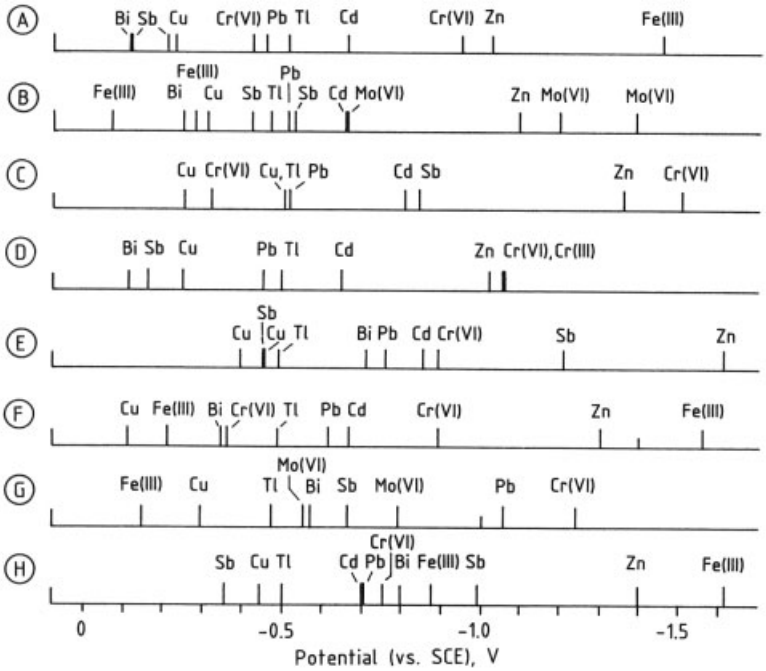

Figure 38. Polarographic half-wave potentials of selected elements measured in the following supporting electrolyte solutions A) 1 mol/L KCl; B) 2 mol/L CH_3COOH + 2 mol/L CH_3COONH_4; C) 1 mol/L NH_3 + 1 mol/L NH_4Cl; D) 1 mol/L HCl; E) 1 mol/L NaOH; F) 0.5 mol/L Na tartrate (pH 9); G) 0.1 mol/L Na_2EDTA; H) 0.1 mol/L Na citrate + 0.1 mol/L NaOH

mate, iodate, and periodate can be determined by means of polarographic reduction currents. Indirect determinations are possible for those anions that form compounds of low solubility or stable complexes with the Hg_2^{2+} ions formed in the anodic oxidation of the electrode mercury. The current caused by the dissolution of the mercury is proportional to the concentration of these anions in the sample solution. Halides, sulfide, and cyanide can be determined in this way [2], [5].

The most important and most successful application of voltammetry/polarography is the determination of traces of heavy metals in aquatic environmental samples. In only lightly polluted water samples, e.g., in rainwater, drinking water, and seawater, the determinations can often be carried out without further sample preparation, i.e., directly after sampling and adjustment to the required pH value. More heavily polluted samples must be digested, which is usually performed by UV irradiation or in a microwave oven (Chap. 5). There is particular interest in determining those metals whose concentration limits in natural waters or wastewaters are regulated by law and accordingly must be controlled (regulations governing purity of drinking water and wastewater discharge). In DIN 38 406, Part 16, for example (German Standard Method), the Zn, Cd, Pb, Cu, Tl, Ni, and Co concentrations are monitored with the aid of polarographic and voltammetric methods of determination: the procedure is shown in Figure 39.

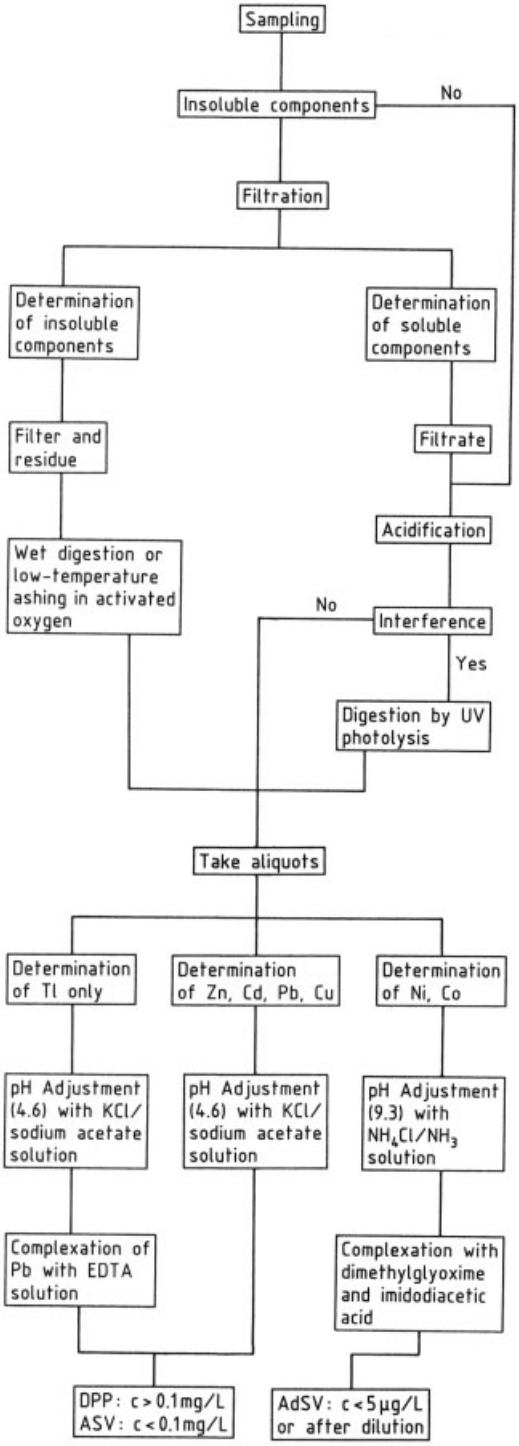

Figure 39. Flow diagram of the procedure for the polarographic/voltammetric determination of heavy metals in drinking water, groundwater, and surface water according to DIN 38 406 Part 16

The method illustrated in Figure 39 is suitable for determinations between the lower µg/L and middle mg/L ranges. It is used for monitoring the heavy metal content in polluted surface waters and identifies the metal content of solutions as well as the constituents that occur in association with suspended matter or colloidal particles. The generally very small amounts of Co and Ni are determined as dimethylglyoxime complexes by AdSV, while for the other metals DPP or ASV is used, depending on the concentration. Sample preparation depends on the state of the sample.

The methods of voltammetry/polarography are important for determining the concentrations of metals in different oxidation states and for characterizing physicochemical forms of bonding (*speciation analysis*) [17], [18].

The metal content determined by ASV in untreated water samples at natural pH values or in buffered solutions is usually lower than the total content and is described as the biologically available fraction. This is the fraction that is present in the form of labile inorganic complexes or hydrated ions. The metal content combined with organic ligands can only be determined voltammetrically after digestion by UV photolysis. In this way labile and nonlabile or combined metal species can be distinguished.

For the ultra-trace analysis of heavy metals in water the AdSV method, in addition to ASV, is especially powerful [19–21]. The limits of determination are generally in the middle to upper ng/L range; Table 2 gives a survey of the best-known methods.

Most AdSV determinations are carried out, after sample digestion by UV irradiation, by recording cathodic stripping voltammograms. Oxidative microwave digestion with Oxisolv (Chapter 5) is also suitable for sample preparation. Only rarely can trace concentrations of metals in natural samples such as seawater be determined without sample digestion. An example is the determination of uranium using chloranilic acid (2,5-dichloro-3,6-dihydroxy-1,4-benzoquinone) (Fig. 40) as complexing agent. This method enables interference-free determinations in river water samples to be carried out [64].

In contrast, the determination of *platinum* by AdSV with formazone (a condensation product

Table 2. Ultra-trace analysis of metals in natural waters by AdSV

Analyte	Ligand	Electrolyte	Sample pretreatment	Detection limit, ng/L	Reference
Aluminum	DASA [c]	BES [d] pH 7.1	UV irradiation (not seawater)	30	[52]
Antimony	chloranilic acid	pH 1.0 adjusted with HCl	UV irradiation	550	[53]
Gallium	solochrome violet RS	acetate buffer pH 4.8	UV irradiation (not seawater)	80	[54]
Molybdenum	chloranilic acid	pH 2.7 adjusted with HCl	UV irradiation (not seawater)	20	[50]
Platinum	formazone	formaldehyde + hydrazine + 0.5 mol/L H_2SO_4	UV irradiation	0.01	[55]
Thorium	mordant blue S	acetate buffer pH 6.5	UV irradiation	100	[56]
Tin	pyrocatechol	acetate buffer pH 4.2	UV irradiation (not seawater)	5	[57]
Titanium	mandelic acid	$KClO_3 + NH_3$ pH 3.2	UV irradiation	0.05	[58]
Uranium	pyrocatechol	acetate buffer pH 4.7	separation by ion exchange	240	[59]
		HEPES [a] + 0.5 mol/L NaOH pH 6.8	UV irradiation (not seawater)	70	[60]
	oxine	0.01 mol/L PIPES [b] + 0.5 mol/L NaOH pH 6.7 – 6.9	UV irradiation (not seawater)	500	[61]
	chloranilic acid	pH 2.5 adjusted with HCl	unnecessary	25	[64]
Vanadium	pyrocatechol	PIPES [b] pH 6.9	UV irradiation	5	[65]

[a] HEPES = *N*-2-hydroxyethylpiperazine-*N*-2-ethanesulfonic acid. [b] PIPES = piperazine-*N,N′*-bis(2-ethanesulfonic acid) monosodium salt. [c] DASA = 1,2-dihydroxyanthraquinone-3-sulfonic acid. [d] BES = *N,N′*-bis-(2-hydroxyethyl)-2-aminoethanesulfonic acid

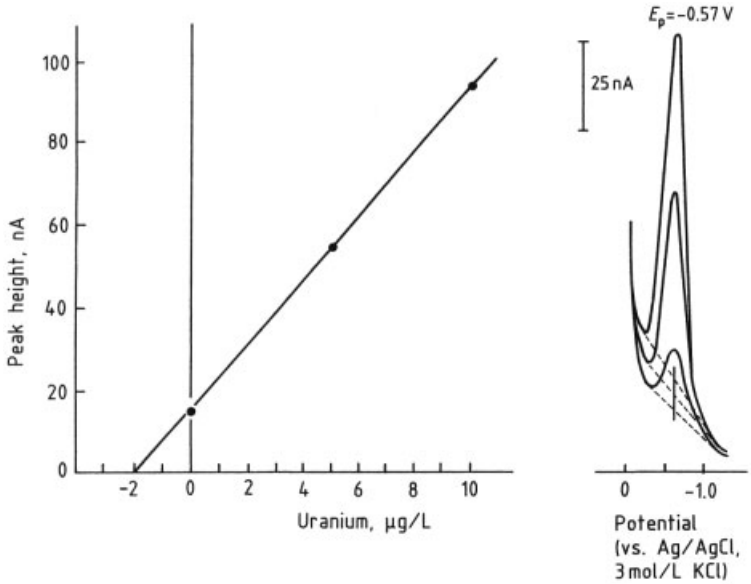

Figure 40. Direct determination of low concentrations of uranium in river water by AdSV [64]
Sample contains 2.4×10^{-4} mol/L chloranilic acid, adjusted with HCl to pH 2.5; $E_{acc} = +0.15$ V; $t_{acc} = 90$ s; standard additions method

of formaldehyde and hydrazine) suffers interference from surface-active substances which can be present even in the purest drinking water at trace level. The unusual feature of the extremely sensitive determination of platinum by AdSV is that the complex of Pt(II) with formazone, after adsorptive accumulation at the HMDE, reduces the hydrogen overvoltage on the mercury surface. At the same time, catalytic currents are developed that can be used for the trace analysis of platinum in the ng/L range [55]. The Pt determination is carried out in water samples after prior UV photolysis, and in biological materials after digestion with HNO_3 in a high-pressure asher.

A further field of application of AdSV is the determination of heavy metal traces in high-purity chemicals. The methods originally developed for water analysis (Table 2) can be used without modification for the purity control of salts of the alkali and alkaline earth metals without having to take account of interference by the salt matrix [66].

Arsenic, selenium, and tellurium can be determined at the trace level by cathodic stripping voltammetry (CSV) after having been reductively co-deposited with copper. A characteristic of the stripping technique of these three elements is that only a single current peak which arises from the further reduction of the deposited analyte to As^{3-}, Se^{2-}, or Te^{2-} is observed. This method is suitable for the determination and speciation of As(V) and As(III) in water samples [36], selenium in minerals [38] and biological samples [68], and for the simultaneous determination of selenium and tellurium (Fig. 41) [37].

Potentiometric stripping analysis (PSA) and chronopotentiometric stripping analysis (CPS) are an alternative techniques that may be used for the determination of metals such as Bi, Cd, Cu, Pb, Sn, Tl, and Zn in water samples [67]. With microprocessor controlled equipment, limits of detection are in the sub-µg/L range. The main advantage of PSA and CPS is that atmospheric oxygen dissolved in an aqueous sample does not have to be removed and that in many cases, including those of wine, urine, and blood, direct determination of cadmium and lead is possible [69–71].

The determination of trace concentrations of heavy metals by stripping voltammetry in body fluids and organ samples is used to assess their physiological function. Methods have been developed for determining toxic metals by ASV and PSA in blood, serum, and urine. Stripping methods are also important for monitoring the heavy metal content in food and drink. The samples are mineralized by dry ashing or by wet digestion. The choice of the appropriate decomposition procedure depends on the type of test and the form of the elements to be determined [4], [5].

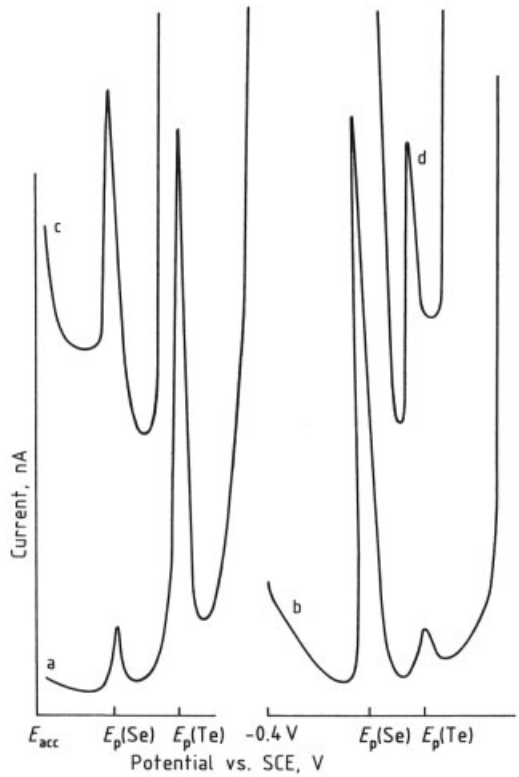

Figure 41. Cathodic stripping voltammograms of selenium and tellurium [37]
Supporting electrolyte solution: 0.5 mol/L $(NH_4)_2SO_4$ + 4×10^{-3} mol/L EDTA + H_2SO_4(pH 4.5) + 1 mg/L Cu^{2+}; t_{acc} = 1 min, E_{acc} = −0.4 V, sensitivity 200 nA fc
a) 1 µg/L Se, 10 µg/L Te; b) 1 µg/L Te, 10 µg/L Se; c) 1 µg/L Se standard addition to 1 mg/L Te; d) 1 µg/L Te standard addition to 1 mg/L Se

Organic compounds. The polarographic and voltammetric determination of organic compounds usually depends on the reduction or oxidation of functional groups. In addition there are some applications where the determination is based on the reaction of organic molecules with the electrode material (mercury) or on the

change in the double layer capacitance associated with the organic species being adsorbed on the electrode surface.

Organic functional groups which are electroactive and the oxidations or reductions of which are the bases of the analytical determinations are listed in Table 3 together with the most important classes of compounds containing these groups. The ranges for the half-wave or peak potentials of selected groups of organic compounds are presented in Figure 42.

Compounds with reducible functional groups predominate. Polyaromatic hydrocarbons, aromatic hydroxy compounds and amines, as well as amides and various nitrogen heterocyclic compounds can be determined anodically. These processes are in many cases pH dependent. Determinations based on redox processes are therefore carried out in buffered solutions. The composition and concentration of the buffer systems generally has no influence on the position of the half-wave or peak potentials. If its concentration is sufficiently high, the buffer simultaneously performs the function of the supporting electrolyte.

The polarographic analysis of aldehydes, organic acids, monosaccharides, and substituted pyridine derivatives is based on kinetic currents, if a rate-determining chemical reaction precedes or follows the actual electrode reaction. Alkaloids, proteins, and many organic sulfur compounds reduce the overvoltage of the hydrogen reaction on the mercury drop electrode and are determined via catalytic hydrogen waves. In native protein molecules, the sulfhydryl and disulfide groups are masked and therefore do not catalyze this electrode process. Not until denaturation by the action of acids, bases, UV radiation, etc., do the catalytic waves appear as a result of exposure of these groups. This phenomenon can be used to determine denatured DNA in natural DNA samples [22].

Indirect polarographic methods are used to determine nitrilotriacetic acid (NTA) and other *polyamino acids*, e.g., EDTA (auxiliary agents for detergents and cleansers). The methods are based on the formation of stable heavy metal complexes with polarographic properties that are different from those of the acids. Bismuth and cadmium salts are used, added in small excess to the sample. For determination of the acids, use is made of the current signals that are caused either by the excess of the metal ions or by a newly formed complex. Both NTA and EDTA are determined according to DIN 38 413 Part 5 after addition of bismuth(III) nitrate to the sample acidified to pH 2 with nitric acid. The polarographic curve is shown in Figure 43.

The determination of *surface-active substances* by tensammetry is particularly sensitive owing to adsorption on the electrode surface and the associated change of the double layer capacitance [23].

Voltammetric trace analysis of *alkaloids and other pharmaceuticals* is carried out after adsorptive accumulation at the electrode surface at constant potential [72]. The compounds that can be determined by AdSV in the submicromolar and nanomolar concentration range also include various pesticides [73–75]. The importance of AdSV for the trace analysis of adsorbable molecules has been summarized [24].

Organic substances that form slightly soluble or coordination compounds with mercury ions (Table 3) are determined polarographically via anodic waves or voltammetrically after anodic accumulation. Agrochemicals containing thiourea are also determined by this CSV technique; the limits of determination are reported to be $10^{-7} - 10^{-8}$ mol/L [76].

Electrochemically inactive organic compounds can be converted into an electrochemically active state by nitration, nitrosation, oxidation, hydrolysis, or some other preliminary chemical reaction, and can then be determined polarographically or voltammetrically. Examples of this are polarographic determinations of alkylbenzenesulfonates in natural water samples after extraction and nitration of the aromatic ring [77] and determinations of beta-receptor blocking agents in tablets after conversion to the corresponding *N*-nitroso derivatives [78].

Aliphatic and aromatic *nitro compounds* are reduced at relatively positive potentials via the hydroxylamines to the corresponding amines. The carcinogenic 4-nitroquinoline-*N*-oxide is determined by DPP in the presence of 4-hydroxyaminoquinoline-*N*-oxide and 4-aminoquinoline-*N*-oxide via the reduction of the nitro group [79]. Nitrazepam, parathion, nitrofurantoin, and the nitroimidazoles in blood plasma or urine are also determined via the reduction of the respective nitro groups.

Table 3. Functional groups for the polarographic and voltammetric determination of organic compounds

Reducible groups	Compounds
$>C=C<$	unsaturated aliphatic and (poly)aromatic hydrocarbons with conjugated double or triple bonds and with multiple double bonds
$-C\equiv C-$	
$-\overset{\vert}{\underset{\vert}{C}}-X$	halogen-substituted aliphatic and aromatic hydrocarbons with the exception of fluoro compounds
$>C=O$	aliphatic and aromatic aldehydes, ketones, and quinones
$-O-O-$	aliphatic and aromatic peroxides and hydroperoxides
$-NO_2$	aliphatic and aromatic nitro and nitroso compounds
$-NO$	
$-N=N-$	azo compounds
$>C=N-$	benzodiazepines, pyridines, quinolines, acridines, pyrimidines, triazines, oximes, hydrazones, semicarbazones
$-S-S-$	disulfides
$>C=S$	thiobenzophenones
$-SO-$	diaryl and alkylaryl sulfoxides
$-SO_2-$	sulfones
$-SO_2NH-$	sulfonamides
$-\overset{\vert}{\underset{\vert}{C}}-Me$	organometallic compounds
Oxidizable groups	**Compounds**
$A-\overset{\vert}{\underset{\vert}{C}}-H$	poly(aromatic) hydrocarbons (A = aryl)
$-OH$	phenols
$-NH_2$	aromatic amines
$-CO-N<$	amides
Groups reacting with mercury ions	**Compounds**
$-SH$	thiols
$-N-C(=S)S^-$	dithiocarbamates
$-NH-C(=S)-NH-$	thioureas, thiobarbiturates
$-C(=S)NH_2$	thioamides
$-NH-C(=O)-NH-$	derivatives of barbituric acid and of uracil

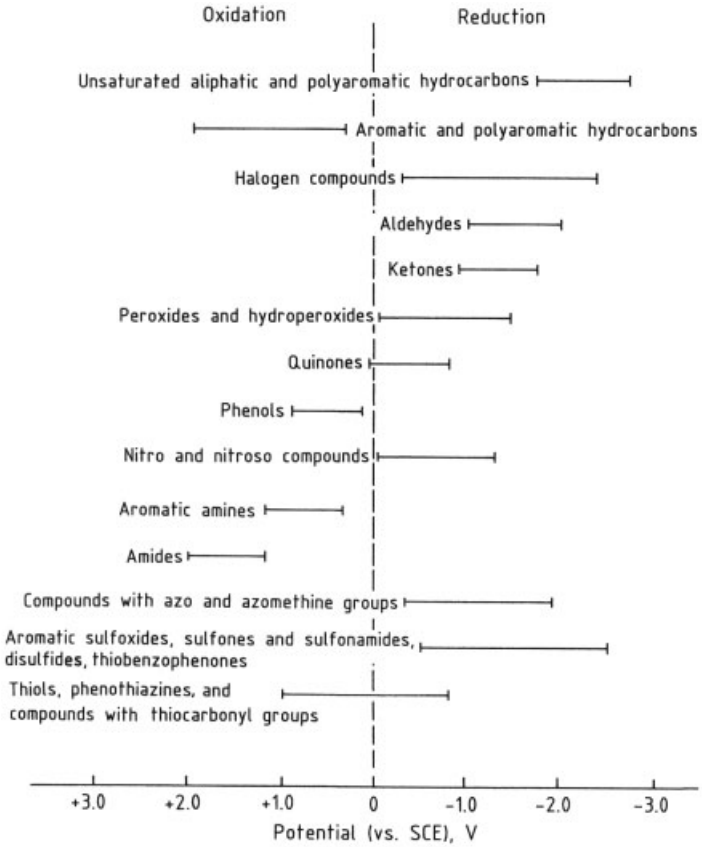

Figure 42. Potential ranges for the polarographic and voltametric half-wave and peak potentials of organic compounds

The *nitroso group* is generally more easily reducible than the nitro group. *N*-Nitrosamines, e.g., *N*-nitroso-*N*-methylaniline in blood, serum, or albumin, are determined in the µg/kg range by DPP [80]. The polarographic and voltammetric behavior of the nitrosamines, quinones, steroid hormones, and imidazoles is of interest, particularly in connection with cancer research [25].

Knowledge about the bio-transformation of *drugs* and their excretion is important pharmacologically. Numerous methods for the polarographic and voltammetric determination of pharmaceuticals and their metabolites have been developed for the assessment of such relationships. The determination is generally preceded by separation from the matrix by liquid – liquid extraction with diethyl ether or ethyl acetate.

1,4-Diazepines are determined polarographically via their azomethine groups. In addition to flurazepam, further representatives of this class of pharmaceutically active agents can be determined polarographically. These include diazepam, chlordiazepoxide, bromazepam, lorazepam, and chlorazepam; the limits of determination with DPP are in the µg/L range [4], [5].

Organic compounds with oxidizable functional groups (see Table 3) are mostly determined, after separation by liquid (column) chromatography or by high-performance liquid chromatography, with the aid of a voltammetric (amperometric) detector (→ Liquid Chromatography). These compounds, which can be determined in liquid and solid biological samples, in foods, or in aquatic environmental samples, include catecholamines, sulfonamides, phenothiazines, dopamines, estrogens and other hormones, tocopherols, and various groups of agricultural pesticides. Liquid – liquid extraction or solid phase extraction are also used for sample preparation; the determinations are carried out

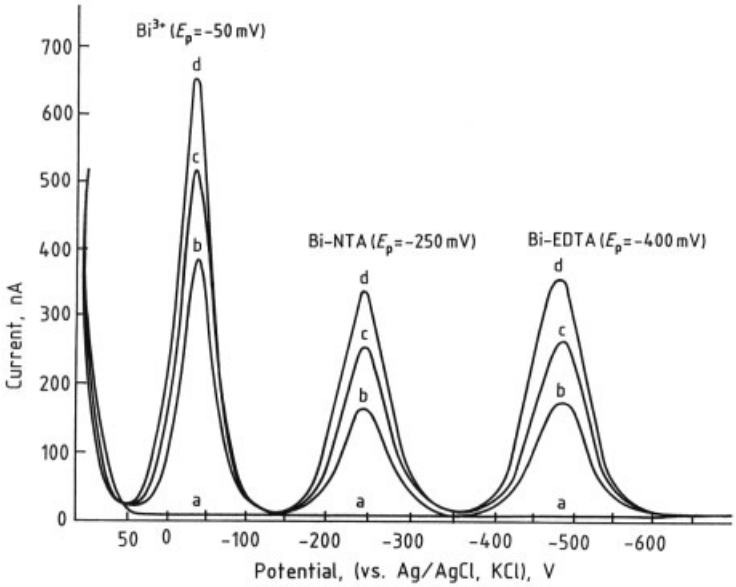

Figure 43. Polarographic determination of NTA and EDTA according to DIN 38 413 Part 5, standard additions method
a) Residual current; b) After Bi^{3+} addition; c) First addition of Bi-NTA/EDTA solution; d) Second addition of Bi-NTA/EDTA solution

in the ng/L range (→ Liquid Chromatography, Chap. 8.1.).

Polarography and voltammetry are of importance in the *analysis of organometallic* compounds in environmental samples. In contrast to spectroscopic methods which only give information on the total metal contents in a sample, the voltammetric techniques are capable of characterizing and determining the various forms of those compounds as they exist in the sample (speciation).

Representative of the organometallic compounds are the R – Me – compounds where Me = As, Hg, P, Pb, Sb, Se, Sn, Te, and Ti; the peak potentials are listed in [30].

Organomercury compounds of the type RHgX generally are reduced in two steps according to the mechanism.

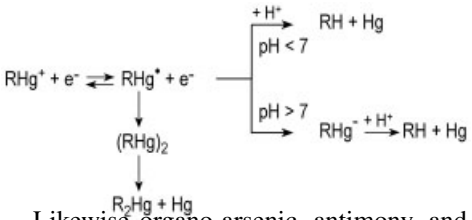

Likewise organo-arsenic, antimony, and tin compounds either singly or in admixture can be determined in the µg range by polarography and particularly mono-, di-, and tributyltin in the ng range, based on solid phase extraction and adsorptive cathodic stripping voltammetry [81].

The polarographic or voltammetric determination of organic compounds of biological, pharmaceutical, or environmental significance have been summarized [27–29].

8. References

General References

1. J. Heyrovský, J. Kuta: *Grundlagen der Polarographie,* Akademie Verlag, Berlin 1965.
2. A. M. Bond: *Modern Polarographic Methods in Analytical Chemistry,* Marcel Dekker, New York 1980.
3. F. G. Thomas, G. Henze: *An Introduction to Analytical Voltammetry – Theory and Practice,* CSIRO, Melbourne (in press).
4. G. Henze: *Polarographie und Voltammetrie; Grundlagen und analytische Praxis,* Springer Verlag, (in press).
5. G. Henze, R. Neeb: *Elektrochemische Analytik,* Springer Verlag, Berlin 1986.
6. J. Koryta, J. Dvorak: *Principles of Electrochemistry,* J. Wiley & Sons, New York 1987.

7. M. R. Smyth, J. G. Vos: *Analytical Voltammetry,* Elsevier, New York 1992.
8. G. C. Baker, A. W. Gardner: "Forty Years of Square-Wave Polarography," *Analyst (London)* **117** (1982) 1811.
9. J. G. Osteryoung, M. M. Schreiner: "Recent Advances in Pulse Voltammetry," *CRC Crit. Rev. Anal. Chem.* **19 (Suppl. 1)** (1988) 1.
10. J. Heinze: "Cyclovoltammetrie – die Spektroskopie des Elektrochemikers," *Angew. Chem.* **96** (1984) 823.
11. R. Neeb: *Inverse Polarographie und Voltammetrie,* Verlag Chemie, Weinheim 1969.
12. J. Wang: *Stripping Analysis* VCH Verlagsgesellschaft, Weinheim 1985.
13. R. Kalvoda, M. Kopanica: "Adsorptive Stripping Voltammetry in Trace Analysis," *Pure Appl. Chem.* **61** (1987) 97.
14. G. Henze: "Stripping-Chronopotentiometrie," *Fresenius Z. Anal. Chem.* **315** (1983) 438.
15. G. K. Budnikow, N. A. Ulakhovich: "Extraction Polarography and its Analytical Application," *Russ. Chem. Rev. (Engl. Transl.)* **49** (1980) 74.
16. P. Rach, H. Seiler: *Polarographie und Voltammetrie in der Spurenanalyse,* Hüthig Verlag, Heidelberg 1985.
17. G. Henze: "Application of Polarographic and Voltammetric Techniques in Environmental Analysis," *NATO ASI Ser. G* **23** (1990) 391.
18. T. M. Florence: "Analytical Application to Trace Element Speciation in Water," *Analyst* **111** (1986) 489.
19. C. M. G. van den Berg: "Adsorptive Cathodic Stripping Voltammetry of Trace Elements in Water," *Analyst* **114** (1989) 1527.
20. C. M. G. van den Berg: "Electroanalytical Chemistry of Sea-Water," *Chem. Oceanogr.* **9** (1988) 195.
21. C. M. G. van den Berg: "Potentials and Potentialities of Cathodic Stripping Voltammetry of Trace Elements in Natural Water," *Anal. Chim. Acta* **250** (1991) 265.
22. E. Palecek: "Polarographic Techniques in Nucleic Acid Research," in W. F. Smyth (ed.): "Electroanalysis in Hygiene, Environmental, Clinical and Pharmaceutical Chemistry," *Anal. Chem. Symp.* **2** (1980) 79.
23. P. M. Bersier, J. Bersier: "Polarographic Adsorption Analysis and Tensammetry: Toys or Tools," *Analyst* **113** (1988) 3.
24. R. Kalvoda: "Polarographic Determination of Adsorbable Molecules," *Pure Appl. Chem.* **59** (1987) 715.
25. "Electrochemistry in Cancer Research," *Anal. Proc. (London)* **17** (1980) 278 (Conference Proceedings)
26. J. P. Hart: "Electroanalysis of Biologically Important Compounds," in *Analytical Chemistry,* Ellis Harwood Series.
27. J. Wang: *Electroanalytical Techniques in Clinical Chemistry and Laboratory Chemistry,* VCH Verlagsgesellschaft, Weinheim 1988.
28. W. F. Smyth: "Voltammetric Determination of Molecules of Biological, Environmental, and Pharmaceutical Importance," *CRC Crit. Rev. Anal. Chem.* **18** (1987) 155.
29. W. F. Smyth, M. Smyth: "Electrochemical Analysis of Organic Pollutants," *Pure Appl. Chem.* **59** (1987) 245.

Specific References

30. W. F. Smyth: *Voltammetric Determination of Molecules of Biological Significance,* John Wiley & Sons, New York 1992.
31. G. C. Barker, A. W. Gardner, *Fresenius' Z. Anal. Chem.* **113** (1960) 73.
32. J. G. Osteryoung, R. A. Osteryoung, *Anal. Chem.* **57** (1985) 101 A.
33. E. J. Zachowski, M. Wojciechowski, J. Osteryoung, *Anal. Chim. Acta* **183** (1986) 47.
34. G. Henze, R. Neeb, *Fresenius' J. Anal. Chem.* **310** (1982) 111.
35. G. Henze, A. P. Joshi, R. Neeb, *Fresenius J. Anal. Chem.* **300** (1980) 267.
36. G. Henze, W. Wagner, S. Sander, *Fresenius J. Anal. Chem.* **358** (1997) 741.
37. G. Henze et al., *Fresenius' Z. Anal. Chem.* **295** (1979) 1.
38. G. Henze, *Mikrochim. Acta* 1981, no. II, 343.
39. D. Jagner, K. Aren, *Anal. Chim. Acta* **100** (1978) 375.
40. D. Jagner, *Anal. Chem.* **50** (1978) 1924.
41. L. Kryger, *Anal. Chim. Acta* **120** (1980) 19.
42. A. Graneli, D. Jagner, M. Josefson, *Anal. Chem.* **52** (1980) 2220.
43. T. Garai et al., *Electroanalysis* **4** (1992) 899.
44. K. Kalcher, "Chemically Modified Carbon Paste Electrodes in Voltammetric Analysis," *Electroanalysis* **2** (1990) 419.
45. J. Wang, T. Boamin, R. Setiadji, *Electroanalysis* **6** (1994) 317.
46. Kh. Brainina, G. Henze, N. Stojko, N. Malakhova, Chr. Faller, "Thick-film Graphite Electrodes in Stripping Voltammetry," *Fresenius J. Anal. Chem.* **364** (1999) 285.

47. J. Schiewe, A. M. Bond, G. Henze: "Voltammetrische Spurenanalyse mit Mikroelektroden" in G. Henze, M. Köhler, J. P. Lay (eds.): *Umweltdiagnostik mit Mikrosystemen,* Wiley-VCH, Weinheim 1999, p. 121.
48. H. Y. Chen, R. Neeb, *Fresenius' Z. Anal. Chem.* **319** (1984) 240.
49. M. Kolb, P. Rach, J. Schäfer, A. Wild, *Fresenius' J. Anal. Chem.* **342** (1992) 341.
50. M. Karakaplan, S. Gücer, G. Henze, *Fresenius' J. Anal. Chem.* **342** (1992) 186.
51. D. J. Myers, J. Osteryoung, *Anal. Chem.* **45** (1973) 267.
52. C. M. G. van den Berg, K. Murphy, J. P. Riley, *Anal. Chim. Acta* **188** (1986) 177.
53. W. Wagner, S. Sander, G. Henze, *Fresenius J. Anal. Chem.* (in press).
54. J. Wang, J. M. Zadeii, *Anal. Chim. Acta* **185** (1986) 229.
55. C. M. G. van den Berg, G. S. Jacinto, *Anal. Chim. Acta* **211** (1988) 129.
56. J. Wang, J. M. Zadeii, *Anal. Chim. Acta* **188** (1986) 187.
57. S. B. Adeloju, *Anal. Sci.* **7** (1991) 1099.
58. K. Yokoi, C. M. G. van den Berg, *Anal. Chim. Acta* **245** (1991) 167.
59. N. K. Lam, R. Kalvoda, M. Kopanica, *Anal. Chim. Acta* **154** (1983) 79.
60. C. M. G. van den Berg, Z. Qiang Huang, *Anal. Chim. Acta* **164** (1984) 209.
61. C. M. G. van den Berg, M. Nimmo, *Anal. Chem.* **59** (1987) 924.
62. J. Zarebski, *Chemia Analityczna* **22** (1977) 1037.
63. L. Meites: *Polarographic Techniques,* 2nd ed., Interscience, New York 1961.
64. S. Sander, G. Henze, *Fresenius J. Anal. Chem.* **349** (1994) 654.
65. C. M. G. van den Berg, Z. Q. Huang, *Anal. Chem.* **56** (1984) 2383.
66. R. Naumann, W. Schmidt, G. Höhl, *Fresenius J. Anal. Chem.* **343** (1992) 746.
67. J. M. Estela, C, Tomas, A. Cladera, V. Cerda, *Crit. Rev. Anal. Chem.* **25** (1995) 91.
68. S. B. Adeloju, A. M. Bond, *Anal. Chem.* **56** (1984) 2397.
69. D. Jagner, M. Josefson, S. Westerlund, K. Aren, *Anal. Chem.* **53** (1981) 1406.
70. H. Huiliang, D. Jagner, L. Renman, *Talanta* **34** (1987) 539.
71. C. Hua, D. Jagner, L. Renman, *Anal. Chim. Acta* **197** (1987) 25.
72. R. Kalvoda, *Anal. Chim. Acta* **138** (1982) 11.
73. M. Pedrero, F. J. Manuel de Villena, J. M. Pingarron, L. M. Polo, *Electroanalysis* **3** (1991) 419.
74. Ch. Li, B. D. James, R. J. Magee, *Electroanalysis* **2** (1990) 63.
75. M. A. Goicolea, J. F. Arrenz, R. J. Barrio, Z. G. de Balugera, *Fresenius J. Anal. Chem.* **339** (1991) 166.
76. M. R. Smyth, J. G. Osteryoung, *Anal. Chem.* **49** (1977) 2310.
77. J. P. Hart, W. F. Smyth, B. J. Birch, *Analyst* **104** (1979) 853.
78. M. A. Korany, H. Riedel, *Fresenius' Z. Anal. Chem.* **314** (1983) 678.
79. J. S. Burmich, W. F. Smyth, *Anal. Proc. (London)* **17** (1980) 284.
80. H. M. Pylypiw, G. W. Harrington, *Anal. Chem.* **53** (1981) 2365.
81. J. Schwarz, G. Henze, F. G. Thomas, *Fresenius J. Anal. Chem.* **352** (1995) 474, 479.

Thermal Analysis and Calorimetry

Stephen B. Warrington, Anasys, IPTME, Loughborough University, Loughborough, United Kingdom (Chap. 1)

Günther W. H. Höhne, Polymer Technology (SKT), Eindhoven University of Technology, Eindhoven, The Netherlands (Chap. 2)

1. Thermal Analysis 1
1.1. General Introduction 1
1.1.1. Definitions 1
1.1.2. Sources of Information 2
1.2. Thermogravimetry 2
1.2.1. Introduction 2
1.2.2. Instrumentation 2
1.2.3. Factors Affecting a TG Curve 3
1.2.4. Applications 3
1.3. Differential Thermal Analysis and Differential Scanning Calorimetry 5
1.3.1. Introduction 5
1.3.2. Instrumentation 5
1.3.3. Applications 6
1.3.4. Modulated-Temperature DSC (MT-DSC) 7
1.4. Simultaneous Techniques 7
1.4.1. Introduction 7
1.4.2. Applications 8
1.5. Evolved Gas Analysis 8
1.6. Mechanical Methods 9
1.7. Less Common Techniques 10
2. Calorimetry 10
2.1. Introduction 10
2.2. Methods of Calorimetry 11
2.2.1. Compensation of the Thermal Effects 11
2.2.2. Measurement of a Temperature Difference 12
2.2.3. Temperature Modulation 12
2.3. Calorimeters 13
2.3.1. Static Calorimeters 14
2.3.1.1. Isothermal Calorimeters 14
2.3.1.2. Isoperibolic Calorimeters 14
2.3.1.3. Adiabatic Calorimeters 16
2.3.2. Scanning Calorimeters 16
2.3.2.1. Differential-Temperature Scanning Calorimeters 17
2.3.2.2. Power-Compensated Scanning Calorimeters 18
2.3.2.3. Temperature-Modulated Scanning Calorimeters 19
2.3.3. Chip-Calorimeters 19
2.4. Applications of Calorimetry 20
2.4.1. Determination of Thermodynamic Functions 20
2.4.2. Determination of Heats of Mixing . 21
2.4.3. Combustion Calorimetry 22
2.4.4. Reaction Calorimetry 22
2.4.5. Safety Studies 24
3. References 24

In Chapter 2 the basics of calorimetry are presented. After the introduction of static and dynamic methods to measure heats or heat flow rates, different types of calorimeters are described, both of classical and of modern design. A section on applications of calorimetry in science and technology follows, where some essential fields such as determination of thermodynamic functions and safety studies are presented in more detail.

1. Thermal Analysis

1.1. General Introduction

1.1.1. Definitions

Thermal analysis (TA) has been defined as "a group of techniques in which a physical property of a substance and/or its reaction products is measured as a function of temperature while

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_001.pub2

the substance is subjected to a controlled temperature programme" [1]. The formal definition is usually extended to include isothermal studies, in which the property of interest is measured as a function of time.

The definition is a broad one, and covers many methods that are not considered to fall within the field of thermal analysis as it is usually understood. The present chapter will be restricted to major techniques. Since all materials respond to heat in some way, TA has been applied to almost every field of science, with a strong emphasis on solving problems in materials science and engineering, as well as fundamental chemical investigations. TA is applicable whenever the primary interest is in determining the effect of heat upon a material, but the techniques can also be used as a means of probing a system to obtain other types of information, such as composition.

The aim of this chapter is to give an overview of the main TA methods and their applications, with sufficient references to provide reader access to more detailed information. The following list summarizes properties subject to investigation and the corresponding thermal analysis techniques:

Mass	Thermogravimetry (TG)
Temperature	Differential thermal analysis (DTA)
Enthalpy	Differential scanning calorimetry (DSC)
Evolved gas	Evolved gas analysis (EGA)
Physical dimensions	Thermodilatometry (TD)
Mechanical properties	Thermomechanical analysis (TMA)
	Dynamic mechanical analysis (DMA)
Optical properties	Thermoptometry
Electrical properties	Thermoelectrometry
Magnetic properties	Thermomagnetometry

1.1.2. Sources of Information

Two journals (the *Journal of Thermal Analysis and Calorimetry* and *Thermochimica Acta*) devote their contents entirely to TA; the *Proceedings* of the (now) four-yearly Conferences of the International Confederation for Thermal Analysis and Calorimetry (ICTAC) constitute an excellent additional source of research papers. Specific information regarding *Proceedings* volumes for the nine ICTAC Conferences between 1965 and 1991 is available in [1]. The most complete listing of worldwide TA literature is also found in the ICTAC handbook [1], which in addition gives addresses for national TA societies and important equipment suppliers. The most useful textbooks include [2 – 5].

There is a wealth of information available on the Internet; see, for example,
http://www.users.globalnet.co.uk/∼dmprice/-ta/links/index.htm

1.2. Thermogravimetry

1.2.1. Introduction

Thermogravimetry (TG) is used to measure variations in mass as a function of temperature (or time). Processes amenable to study in this way are listed in Table 1. TG is one of the most powerful TA techniques from the standpoint of quantitative data, and for this reason it is often employed in combination with other measurements.

Table 1. Processes that can be studied by thermogravimetry

Process	Weight gain	Weight loss
Ad- or absorption	*	
Desorption		*
Dehydration/desolvation		*
Sublimation		*
Vaporization		*
Decomposition		*
Solid – solid reactions		*
Solid – gas reactions	*	*

1.2.2. Instrumentation

The instrument used is a *thermobalance*. The schematic diagram in Figure 1 presents the main components of a typical modern unit. Component details vary according to the design, and choice of a particular instrument is usually dictated by requirements of the problem under investigation (temperature range, sensitivity, etc.). The balance mechanism itself is usually of the null-deflection type to ensure that the sample's position in the furnace will not change. The balance transmits a continuous measure of the mass of the sample to an appropriate recording system, which is very often a computer. The resulting plot of mass vs. temperature or time is called a *TG curve*. Balance sensitivity is usually

of the order of one microgram, with a total capacity of as much as a few hundred milligrams. Furnaces are available that operate from subambient (e.g., −125 °C) or room temperature up to as high as 2400 °C. A furnace programmer normally supports a wide range of heating and cooling rates, often in combination, as well as precise isothermal control. The programming functions themselves are increasingly being assumed by computers. Most applications involve heating rates of 5 – 20 K/min, but the ability to heat a sample as rapidly as 1000 K/min can be useful in the simulation of certain industrial processes, for example, or in flammability studies. Special control methods, grouped under the term *controlled-rate thermal analysis* (CRTA), are receiving increasing attention [6], and offer advantages in resolving overlapping processes and in kinetic studies.

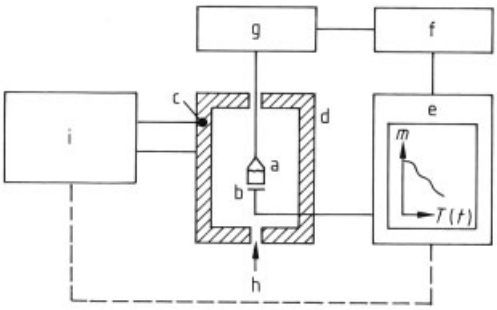

Figure 1. Schematic diagram of a thermobalance
a) Sample; b) Sample temperature sensor; c) Furnace temperature sensor; d) Furnace; e) Recorder or computer, logging sample mass, temperature, and time; f) Balance controller; g) Recording microbalance; h) Gas; i) Furnace temperature programmer
The computer may also control the furnace programmer

The quality of the furnace atmosphere deserves careful attention. Most commercial thermobalances operate at atmospheric pressure. Vacuum and high-pressure studies normally require specialized equipment, either commercial or home-made, as do experiments with corrosive gases [7]. The ability to establish an inert (oxygen-free) atmosphere is useful, as is the potential for rapidly changing the nature of the atmosphere.

The mode of assembly of the components varies; for example, the furnace might be above, below, or in line with the balance. Sample containers also vary widely in design; cylindrical pans are common, typically 5 – 8 mm in diameter and 2 – 10 mm high, though flat plates and semisealed containers may be used to investigate the effects of atmospheric access to a sample. Compatibility between the construction materials and the system under investigation must be carefully considered. Materials commonly available include aluminum, platinum, alumina, and silica. Temperature indication is normally provided by a thermocouple located near the sample container. Because of inevitable thermal gradients within the apparatus, an indicated temperature can never be taken as an accurate reflection of the temperature of the sample. Reproducible location of the thermocouple is vital; recommended calibration procedures have been described in [8]. Only in simultaneous TG – DTA instruments is direct measurement of the temperature possible.

1.2.3. Factors Affecting a TG Curve

Wendlandt [2] has listed 13 factors that directly affect a TG curve, both sample- and instrument-related, some of which are interactive. The primary factors are heating rate and sample size, an increase in either of which tends to increase the temperature at which sample decomposition occurs, and to decrease the resolution between successive mass losses. The particle size of the sample, the way in which it is packed, the crucible shape, and the gas flow rate also affect the progress of a thermal reaction. Careful attention to consistency in experimental details normally results in good repeatability. On the other hand, studying the effect of deliberate alterations in such factors as heating rate can provide valuable insights into the nature of observed reactions. All these considerations are equally applicable to other techniques as well, including DTA (Section 1.3).

1.2.4. Applications

TG has been applied extensively to the study of analytical precipitates for gravimetric analysis [9]. One example is calcium oxalate, as illustrated in Figure 2. Information such as extent of hydration, appropriate drying conditions, stability ranges for intermediate products, and reac-

tion mechanisms can all be deduced from appropriate TG curves. Figure 2 also includes the first derivative of the TG curve, termed the *DTG curve*, which is capable of revealing fine details more clearly.

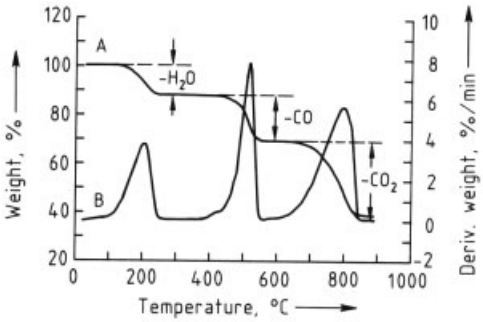

Figure 2. TG (A) and DTG (B) curves for calcium oxalate monohydrate
Sample mass 30 mg, heating rate 20 K/min, argon atmosphere

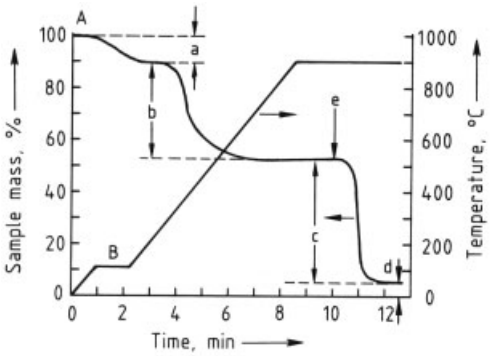

Figure 3. TG curve (A) illustrating a programmed-temperature method for the proximate analysis of coal, where the line B defines the temperature program
a) Moisture; b) Volatiles; c) Fixed carbon; d) Ash; e) Point at which the atmosphere was changed to air

A TG method has also been devised for the proximate analysis of coal, permitting four samples per hour to be analyzed with results of the same precision and accuracy as BS and ASTM methods [10]. A typical curve is shown in Figure 3. The approach taken here can be generalized to the compositional analysis of many materials [11]. TG also facilitates comparisons of the relative stabilities of polymers (Fig. 4). An analysis of curves prepared at different heating rates makes it possible to extract kinetic data for estimating the service lifetimes of such materials. In another coal-related application, an empirical correlation has been established between characteristic temperatures obtained from DTG curves of coals examined under air and the performance of the same coals in steam-raising plants [12], permitting the rapid assessment of new fuels (Fig. 5). The simulation of industrial processes, particularly heterogeneous catalysis, is another important area of application.

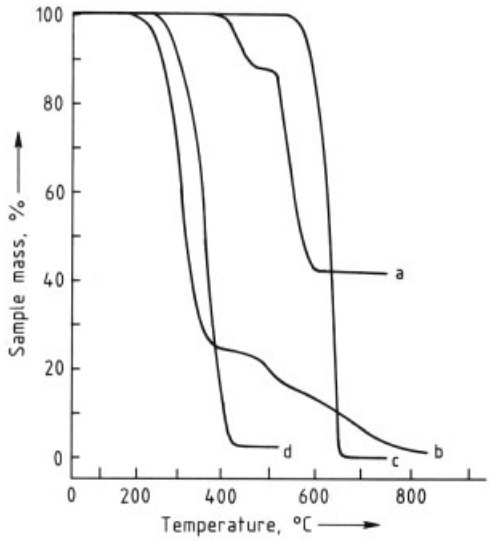

Figure 4. TG curves for four polymers in air, showing their relative stabilities
a) Silicone rubber; b) PVC; c) PTFE; d) Perspex

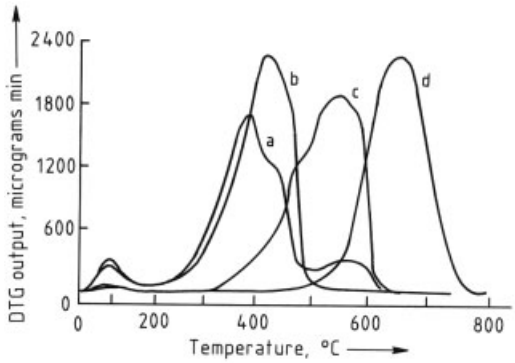

Figure 5. DTG curves for four ASTM standard coal samples heated in air [12]
a) Lignite "A"; b) Sub-bituminous "A"; c) High-volatility bituminous "A"; d) Buckwheat anthracite

1.3. Differential Thermal Analysis and Differential Scanning Calorimetry

1.3.1. Introduction

Both differential thermal analysis (DTA) and differential scanning calorimetry (DSC) are concerned with the measurement of energy changes, and as such are applicable in principle to a wider range of processes than TG. From a practical standpoint DSC may be regarded as the method from which quantitative data are most easily obtained. The use of DSC to determine absolute thermodynamic quantities is discussed in Sections 2.3.2 and 2.4.1. Types of processes amenable to study by these methods are summarized in Table 2.

Table 2. Processes that can be studied by DTA/DSC

Process	Exothermic	Endothermic
Solid – solid transitions	*	*
Solid – liquid transitions	*	*
Vaporization		*
Sublimation		*
Adsorption	*	
Desorption		*
Desolvation		*
Decomposition	*	*
Solid – solid reactions	*	*
Solid – liquid reactions	*	*
Solid – gas reactions	*	*
Curing	*	
Polymerizations	*	
Catalytic reactions	*	*

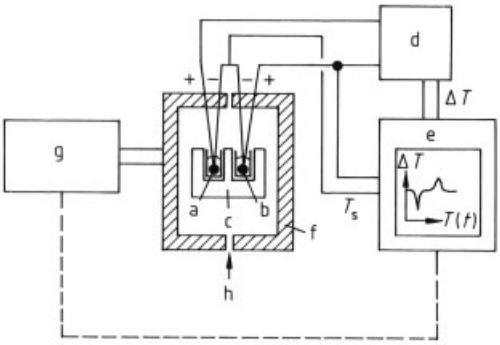

Figure 6. Schematic diagram of a classical DTA apparatus a) Reference thermocouple; b) Sample thermocouple; c) Heating block; d) ΔT amplifier; e) Recorder or computer, logging Ts, ΔT, and time; f) Furnace; g) Temperature programmer, which may be linked to the computer; h) Gas inlet
Ts, sample temperature; Tr, reference temperature; $\Delta T = T\text{s} - T\text{r}$

1.3.2. Instrumentation

Figure 6 shows the components of a typical DTA apparatus. The sample and an inert reference material (commonly alumina) are subjected to a common temperature program while the temperature difference between the two is monitored. In an ideally designed instrument, the temperature difference ΔT remains approximately constant if no reaction is taking place in the sample. When an endothermic process such as melting occurs, the sample temperature T_s lags behind that of the reference T_r; $\Delta T = T_\mathrm{s} - T_\mathrm{r}$ is thus negative, so a (negative) peak is produced on the corresponding *DTA curve*, which is a record of ΔT against either temperature or time (Fig. 7). Exothermic processes lead to peaks in the opposite direction. The peak *area* is related to both the energy change and the sample size according to the relationship

$$\Delta H \cdot m = K \int \Delta T \cdot \mathrm{d}t$$

where m is the sample mass, and ΔH is the enthalpy of reaction. The proportionality constant K is temperature-dependent, and with classical DTA it is also influenced by thermal properties of the sample and container. Certain apparatus designs minimize these influences, so that once the characteristics of $K = f\ (T)$ have been determined the ΔT signal can be conditioned (electronically or digitally) to give an output signal calibrated directly in power units; i.e., the sensitivity is nominally independent of temperature. This is the principle underlying heat-flux DSC. For a more complete description of this technique, as well as power-compensated DSC, see Section 2.3.2. A wide variety of commercial devices covers the temperature range from ca. −150 – 2500 °C, but high-temperature instruments are less sensitive than low-temperature designs due to constraints imposed by thermocouple technology. DSC instruments are usually limited to a maximum temperature of 700 °C, though heat-flux DSC has been extended to 1400 °C. DTA head designs vary enormously, but most accommodate samples in the range 5 – 200 mg in metal or ceramic containers. The thermocouples usually take the form of beads that fit into recesses in the containers, or plates upon

which the containers rest. Nowadays the thermocouple is never actually placed in the sample, in contrast to the "classical" arrangement shown in Figure 6.

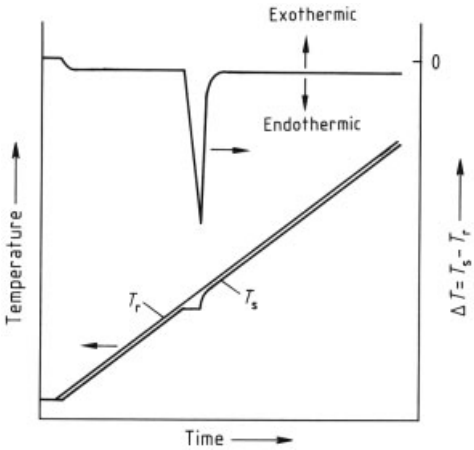

Figure 7. Typical DTA curve for the melting of a pure material

1.3.3. Applications

DSC is used extensively in polymer science [13] (→ Plastics, Properties and Testing). A generalized DSC curve for a polymer is shown in Figure 8. Most polymers display a *glass transition*, in the course of which the material passes from a glassy to a rubbery state with a simultaneous increase in specific heat capacity (→ Plastics, Properties and Testing). Glass-transition temperature measurements are used for material characterization and comparison, often providing information about such factors as thermal history, crystallinity, extent of cure, and plasticizer content of a polymer. Amorphous polymers usually exhibit a crystallization exotherm, and thermosetting polymers a curing exotherm, both of which can be analyzed kinetically to establish the thermal treatment required to obtain desired properties in a particular product. The melting endotherm is a measure of the extent of crystallinity, an important parameter related to mechanical properties. The temperature at which exothermic degradation begins under an oxidizing atmosphere is used to compare oxidative stabilities of different polymers or different stabilizers. Discrimination can be improved by isothermal measurements, in which the material under investigation is heated initially in an inert gas, after which the atmosphere is changed to air or oxygen and the time lapse prior to exothermic deflection is measured. The same approach is applicable to studies of oils, greases, fats, etc., though high-pressure cells may be necessary to prevent volatilization.

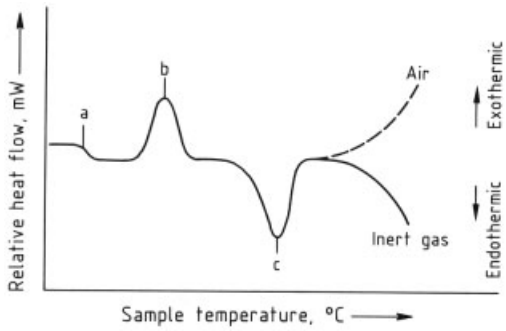

Figure 8. Generalized DSC curve for a polymer
a) Glass transition; b) Crystallization; c) Melting
The subsequent reactions may be endothermic or exothermic

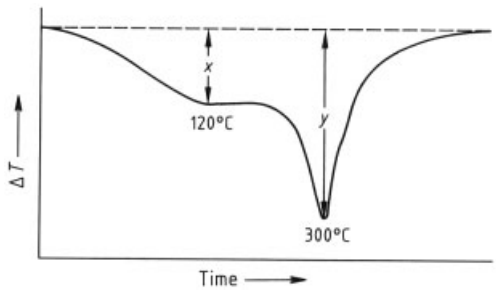

Figure 9. DTA curve for a partially converted high-alumina cement concrete from a test drilling
Degree of conversion $=100\ y/(x + y)$

The particular temperatures at which peaks are observed can be used for the identification of components in a mixture, and the size of a particular peak can be used for quantitative evaluation. Examples include the determination of quartz in clays and analysis of the constituents of polymer blends, both of which would be difficult to achieve by other means. DTA has been used to determine the presence and degree of conversion of high-alumina cement concrete (HAC), a material subject to severe loss of strength under certain atmospheric conditions due to solid-state

reactions in the cement matrix. Figure 9 shows a DTA curve for partially converted HAC together with the way the data are used for a conversion measurement.

An analysis of the shape of the fusion peak from a relatively pure material (e.g., a pharmaceutical agent) can, with certain restrictions, lead to an estimate of purity [14]. Figure 10 shows an idealized melting curve for an impure material. By calculating the fraction (F) of material melted at a series of temperatures, and applying several simplifying assumptions, the mole fraction of an impurity can be calculated from the expression:

$$T_s = T_0 - \frac{RT_0^2 X_2}{\Delta H_f} \cdot \frac{1}{F}$$

where T_s = sample temperature, T_0 = melting point of the pure material, R = gas constant, X_2 = mole fraction of impurity, ΔH_f = enthalpyof fusion, and F = the fraction melted at temperature T_s. The calculations are readily performed by commercial software.

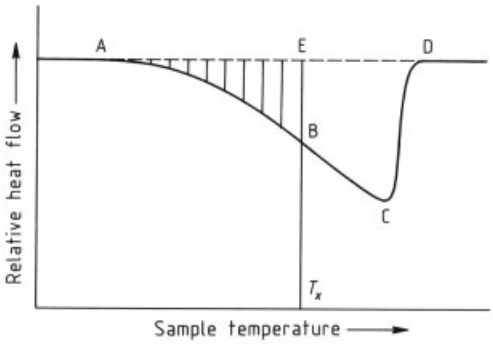

Figure 10. Generalized DSC curve for the melting of an impure material
Fraction melted at temperature Tx = (area ABE)/(area ADC)

DTA curves can be used to construct phase diagrams for mixtures. Measurement of a single point on the phase diagram may be sufficient for quality-control purposes. Figure 11 shows the DTA curve for a nickel-based superalloy, where the solvus temperature (the temperature corresponding to complete solution of one solid phase in another) of the γ' phase characterizes the heat treatment to which the alloy has been subjected.

Solid – solid reactions have also been successfully studied by these techniques, as have decompositions of high explosives. The applicability of DTA/DSC to very small samples is of obvious value here. Similarly, preliminary screening of potentially hazardous reaction mixtures is conveniently carried out in this way as a guide to the likelihood of exothermic reaction, the corresponding temperature range and magnitude, and possible effects of pressure and atmospheric conditions (see also Section 2.4.5).

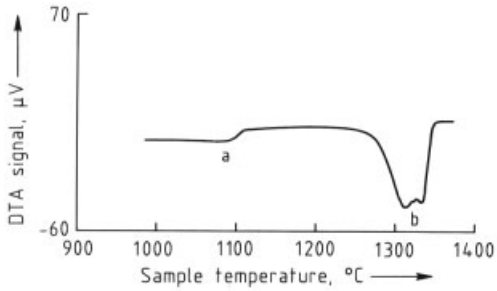

Figure 11. DTA curve for a nickel-based superalloy
a) Solution of the γ' phase; b) Melting 250 mg sample heated at 10 °C min^{-1} in argon

1.3.4. Modulated-Temperature DSC (MT-DSC)

MT-DSC is an important development (see also Section 2.2.3) which offers significant advantages in the quality of results and gives additional information, particularly for polymers [15]. A sinusoidal modulation is superimposed on the normally linear temperature ramp, and the resulting signals are analyzed by a Fourier Transform technique. This allows the separation of "reversing" and "nonreversing" thermal events and simplifies the analysis of overlapping transitions such as those often found in the glass transition region. Other benefits include the ability to measure heat capacity changes pseudo-isothermally, as in, e.g., curing systems, and improved signal-to-noise ratio and baseline linearity.

1.4. Simultaneous Techniques

1.4.1. Introduction

TG or DTA/DSC alone rarely gives sufficient information to permit a complete interpretation

of the reactions in a particular system; results must usually be supplemented by other thermal methods and/or general analytical data. Alternative thermal methods are best applied simultaneously, leading, for example, to TG and DTA information from the same sample under identical experimental conditions. This avoids ambiguity caused by material inhomogeneity, but also problems attributable to different experimental conditions with different instruments, which sometimes markedly affects the correspondence between TG and DTA curves. Other advantages include: (1) indication of the thermal stability of materials examined by DTA, which in turn makes it possible to correct measured heats of reaction for partially decomposed samples; (2) accurate temperature measurement in TG work, which is vital in kinetic studies; and (3) the detection of unsuspected transitions in a condensed phase, which may help explain puzzling features of the corresponding TG curve.

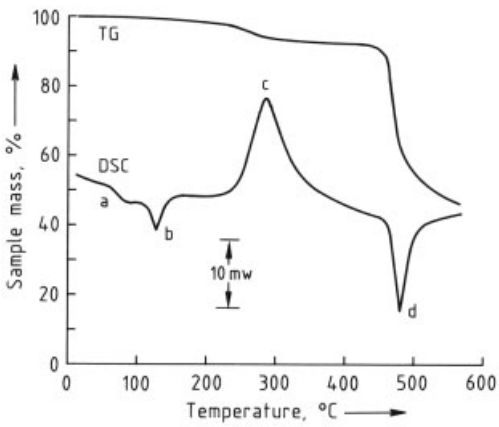

Figure 12. Simultaneous TG – DSC curves for an uncured polyimide resin
a) Glass transition; b) Melting; c) Curing; d) Degradation
10 mg sample heated at 10 °C min^{-1} in argon

The range of simultaneous techniques has been reviewed by Paulik [16]. TG – DTA and TG – DSC are the commonest simultaneous methods, followed by evolved gas analysis (Section 1.5).

1.4.2. Applications

TG – DSC curves showing the curing and decomposition of a polyimide resin are presented in Figure 12. Following a glass transition (a) and subsequent fusion of the material at ca. 120 °C (b), which are detected by DSC, the TG curve reveals the initial stages of an exothermic curing reaction (c) that occurs with a 7 % mass loss. Endothermic decomposition of this resin does not take place until ca. 430 °C, as shown again by TG (d), a finding that permitted assignment of the appropriate baseline for measuring the heat of curing.

1.5. Evolved Gas Analysis

Evolved gas analysis (EGA) measures the nature and/or quantity of volatile products released on heating. In practice it is often performed simultaneously with TG or TG – DTA, and is particularly useful in furnishing direct chemical information to augment the physical measurements of TG or DTA. Additional advantages that have emerged include specificity – a single decomposition can be followed against a background of concurrent processes – and sensitivity, which can be far greater than with TG alone [17, 18].

Instrumentation. Almost every imaginable type of gas detector or analyzer has been utilized in EGA, including hygrometers, nondispersive infrared analyzers, and gas chromatographs. Absorption of the products into solution permits analysis by coulometry, colorimetry, ion selective electrode measurements, or titrimetry. The most important analyzers are Fourier transform infrared (FTIR) spectrometers and, preeminently, mass spectrometers [19]. The two latter methods can be used to record spectra repetitively, thereby producing a time-dependent record of the composition of the gas phase, from which EGA curves can be constructed for selected species.

Applications. Simultaneous TG – MS curves for a brick clay are shown in Figure 13, which simulates a set of firing conditions [20]. After an initial loss of moisture, combustion of organic matter at ca. 300 °C gives characteristic peaks for CO_2, H_2O, and SO_2. More SO_2 results from the oxidation of iron sulfides. Clay dehydroxylation occurs in the range 350 – 600 °C, and calcite dissociates to give the CO_2 peak observed at higher temperatures. EGA is of great value in the interpretation of complex TG and DTA curves.

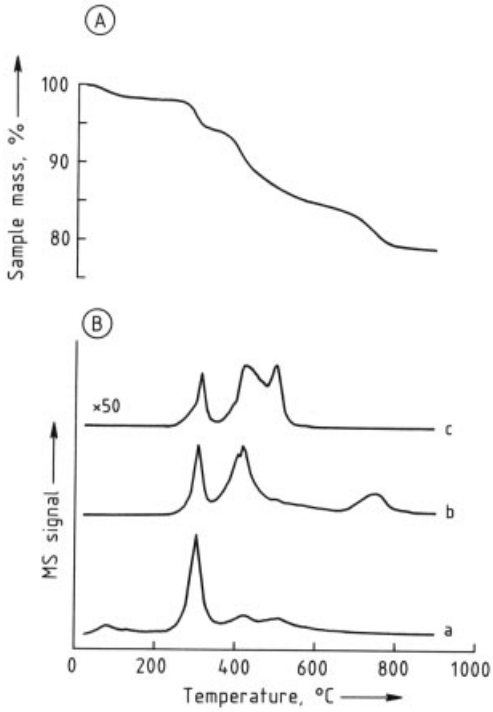

Figure 13. Simultaneous TG – MS curves for a brick clay
A) TG curve; B) MS data
Evolution of a) Water (M, 18); b) Carbon dioxide (M, 44); c) Sulfur dioxide (M, 64)
40 mg sample heated at 15 °C min^{-1} in air

In another example (Fig. 14), the evolution of benzene from two samples of PVC, one containing a smoke retardant, was compared quantitatively by injecting known amounts of benzene into the purge gas stream before and after the experiments, thereby providing a standard for estimating the amount of benzene produced by the polymer [20]. The smoke retardant was found in this case to reduce benzene evolution by a factor of ca. 20.

1.6. Mechanical Methods

This section covers a family of techniques leading to dimensional or mechanical data for a material as a function of temperature or time. For a succinct review of the application areas see [22].

Thermodilatometry (TD) measures dimensional changes as a function of temperature in materials subject to negligible loads. A probe, which is held in light contact with the heated sample, is connected to a sensitive position sensor, usually a linear variable differential transformer (LVDT). In addition to providing expansion coefficients the technique can also indicate phase changes, sintering, and chemical reactions. Major application areas include metallurgy and ceramics.

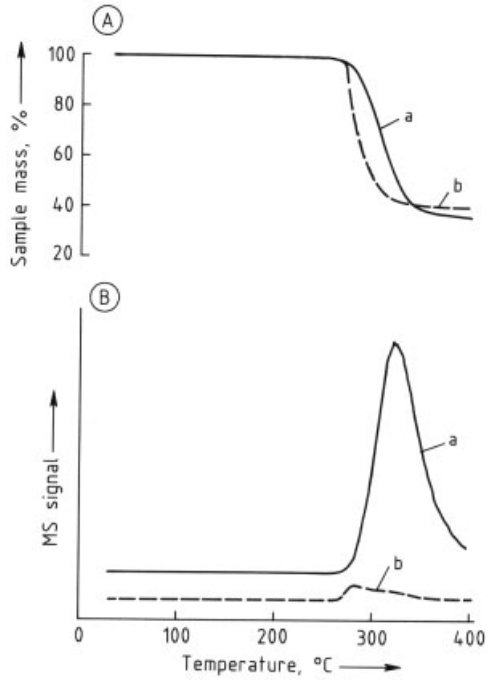

Figure 14. Simultaneous TG – MS curves showing the evolution of benzene (M_r, 78)
A) TG curve; B) MS curve
a) PVC without a smoke retardant; b) PVC with a smoke retardant (4 % MoO_3)
15 mg samples heated at 10 °C min^{-1} in argon

Thermomechanical Analysis. The equipment used in thermomechanical analysis (TMA) is similar in principle to that for TD, but provision is made for applying various types of load to the specimen, so that penetration, extension, and flexure can be measured. This approach to analyzing such modes of deformation is illustrated schematically in Figure 15. The technique finds most use in polymer studies, as in the determination of glass-transition and softening temperatures for thin films and shrinkage characteristics of fibers.

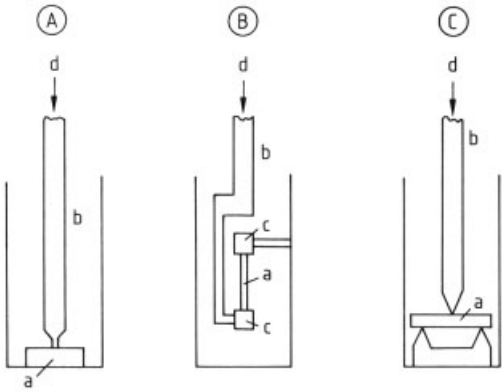

Figure 15. Modes of deformation subject to examination by TMA
A) Penetration; B) Extension; C) Flexure
a) Sample; b) Probe; c) Clamps; d) Load

Dynamic Mechanical Analysis. Polymers exhibit both elastic and viscous behavior under stress. In dynamic mechanical analysis (DMA) a sinusoidally modulated stress, often flexural (though measurement is also possible under tensile, compressive, shear, or torsional conditions) is applied to a specimen of material maintained under a specified temperature regime. Displacement transducers measure strain induced in-phase with the stress, as well as strain that lags behind. The former gives a measure of the sample's modulus, or stiffness, and the latter reflects damping characteristics. Like most other thermoanalytical techniques, DMA can be used to generate quantitative physical data, but it is also invoked in a comparative sense: to monitor the effects of additives, for example, or in quality control. It is particularly well suited to the examination of engineering composites, since gelation and curing behavior is sometimes more easily followed by DMA than by alternative techniques such as DSC.

1.7. Less Common Techniques

Thermoptometry. This group of techniques includes:

1) *Thermomicroscopy*, in which a material is observed under reflected, transmitted, or polarized transmitted light. The corresponding apparatus is referred to as a *hot stage*, which may also incorporate a DSC sensor for additional information.
2) *Thermophotometry*, in which the intensity of reflected or transmitted light is measured.
3) *Thermoluminescence*, which provides a measure of the intensity of light emitted by the material itself.

Electrical and Magnetic Techniques. Thermomagnetometry is used to study the magnetic properties of materials by revealing apparent weight changes that accompany phase transitions in the presence of an applied magnetic field, or variations in the amount of a magnetic substance.

Thermoelectrometry comprises methods for the measurement of resistance or capacitance during heating. Some equipment in this category provides remote probes for in situ monitoring of the curing of thermosets.

Microthermal analysis (μ-TA) enables a range of thermal methods to be applied to regions of a material only a few micrometers wide. They are based on the use of an atomic force microscope (AFM) with a heatable tip. Many materials are heterogeneous, and the microstructure can influence the material properties on the macroscale. After acquiring the conventional AFM topographic image, selected features can be subjected to DTA and TMA with the probe tip. The tip can also be used to pyrolyze the selected area, and the vapors are then analyzed by GC-MS [23].

2. Calorimetry

2.1. Introduction [24, 36 – 44]

Calorimetry in the broadest sense means the quantitative measurement of energy exchanged in the form of heat during a reaction of any type. By contrast, thermal analysis (Chap. 1) is concerned only with the measurement and recording of temperature-induced changes or temperature differences in dependence on time. Since all chemical reactions and many physical changes (e.g., deformation, phase transformations) are associated with the uptake or release of heat, the quantitative investigation of heat exchange is a

relatively simple and universal method for characterizing particular processes both in an overall sense and with respect to time.

Nevertheless, only in recent decades has calorimetry emerged from the laboratories of a few thermodynamicists and specialists to become a widespread, convenient analytical method. The development of commercial calorimeters since the 1950s has led to rapid dissemination and application of the method even beyond the bounds of universities.

The explanation of this phenomenon, as with many other apparatus-based developments, lies in the refinement in measurement techniques made possible by modern electronics. As a rule, heat cannot be measured directly, but must be determined instead on the basis of a temperature change in the system under investigation. Very accurate classical thermometers tend to be very slow measuring devices. Resistance thermometers and thermocouples respond much more rapidly, but they require electronic amplification, and amplifiers with the required precision have become available only in recent decades.

A further explanation for the increasing importance of calorimetry is the development of personal computers, which relieve the user of the often laborious task of evaluating the experimental results, thereby opening the way for the method to become a routine laboratory technique.

In the sections that follow, calorimetry and the present state of instrument development will be described in such a way as to reveal both the possibilities and the limits of the method. The scope of the chapter rules out exhaustive treatment of the subject, so the goal is rather to delineate the method and guide the interested reader to other relevant technical literature.

2.2. Methods of Calorimetry

As noted above, heat that is released or consumed by a process cannot be measured directly. Unlike material quantities such as amount of substance, which can be determined with a balance, or the volume of a liquid, which can be established with a liter measure, the quantity heat must be measured indirectly through its effect on a substance whose temperature it either raises or lowers. The fundamental equation for all of calorimetry describes the relationship between heat exchanged with a calorimeter substance and a corresponding temperature change:

$$\Delta Q = C\,(T) \cdot \Delta T \tag{1}$$

where ΔQ = exchanged heat (J), ΔT = observed temperature change (K), and C (T) = heat capacity (J/K).

For heats (and therefore temperature changes) that are not too large, the temperature change is directly proportional to the exchanged heat. The proportionality constant is the heat capacity of the calorimeter substance (previously referred to as the "water value"). However, if the temperature change exceeds a few Kelvin, the temperature dependence of heat capacity stands in the way of a linear relationship, and a knowledge of the temperature function of the particular heat capacity in question is required in order to determine heat on the basis of a measured temperature difference.

The relationship above leads directly to the standard calorimetric methods for determining the heat of a process: either temperature is held constant by appropriate compensation for the heat effect, and the required compensation power is measured, or a temperature change is determined and used to calculate a corresponding value for exchanged heat. A precondition for the latter approach is accurate knowledge of the heat capacity as well as the heat transport properties of the measurement system.

2.2.1. Compensation of the Thermal Effects

In the first method of heat measurement, temperature changes in the calorimeter substance are avoided by supplying or dissipating heat exactly equal in magnitude (but opposite in sign) to that associated with the process under investigation. As a rule, electrical energy is used to provide this compensation, either by introduction in the form of Joule's heat or dissipation through the Peltier effect. A combination approach is also possible, with an appropriate constant level of cooling and simultaneous controlled electrical heating. The requisite amount of compensation power can today be determined with a high degree of precision. It was once customary to invoke for compensation purposes "latent" heat derived from a phase change; in other words, heat released in

a process was measured by using it to melt a substance such as ice, with subsequent weighing of the resulting water and conversion of the data into energy units based on the known heat of fusion for ice.

The method of compensation is advantageous, since it permits measurements to be carried out under quasi-isothermal conditions, thereby avoiding heat loss from the calorimeter to the surroundings by heat transport processes. Furthermore, there is no need for a calibrated thermometer, only a sensor sufficiently sensitive to temperature changes to provide adequate control over the compensation power.

2.2.2. Measurement of a Temperature Difference

In this alternative method of heat measurement, which is also indirect, a measured temperature difference is used to calculate the amount of heat exchanged. A distinction can be made between temporal and spatial temperature-difference measurements. In *temporal* temperature-difference measurement the temperature of a calorimeter substance is measured before and after a process, and a corresponding heat is calculated on the basis of Equation (1) (which presupposes an accurate knowledge of the heat capacity). In the *spatial* method a temperature difference between two points within the calorimeter (or between the calorimeter substance and the surroundings) is the quantity of interest. The basis for interpretation in this case is from Fourier's law the equation for stationary heat conduction (Newton's law of cooling):

$$\Phi = \lambda(T) \cdot A \cdot \Delta T \cdot l^{-1} \tag{2}$$

where Φ = heat flow rate (W), $\lambda(T)$ = thermal conductivity (W m^{-1} K^{-1}), A = cross-sectional area (m^2), ΔT = temperature difference (K), and l = length (m).

From this expression it follows that the heat flow rate through a heat-conducting material is proportional to the corresponding temperature difference. If this temperature difference is recorded as a function of time, then for a given thermal conductivity one acquires a measure of the corresponding heat flow rate, which can be integrated to give a total heat. The technique itself is very simple, but the result will be correct only if the measured temperature difference accurately reflects the total heat flow rate and no heat is lost through undetected "heat leaks". Careful calibration and critical tests of the calorimeter are absolutely necessary if one hopes to obtain reliable results.

A fundamental disadvantage of the temperature-difference method is the fact that nonisothermal conditions prevail, so the precondition of a stationary state is fulfilled only approximately. The potential therefore exists for nonlinear phenomena in the heat-transport process, which might result in conditions outside the range of validity for Equation (2). Furthermore, with nonisothermal studies it is impossible to rule out completely the presence of undetected leaks. It must also be pointed out that in the case of nonisothermal studies the thermodynamic quantities of interest cannot be established according to the methods of reversible thermodynamics, at least not without further consideration. Strictly speaking, *irreversible thermodynamics* should be applied to nonisothermal calorimetry. Although the distinction is generally unimportant in normal calorimetric practice, it becomes essential in cases where time-dependent processes are involved with timescales which are comparable to those of the experiment, e.g., in the *temperature-modulated* mode of operation.

2.2.3. Temperature Modulation

The temperature-modulated mode of operation has been well known for many decades in calorimetry [45], but became well established only during the 1990s, when commercial DSC was modified this way [46]. The idea is to examine the behavior of the sample for periodic rather than for isothermal or constant-heating-rate temperature changes. In this way it is possible to obtain information on time-dependent processes within the sample that result in a time-dependent generalized (apparent) heat capacity function or, equivalently, in a complex frequency-dependent quantity. Similar complex quantities (electric susceptibility, Young's modulus) are known from other dynamic (dielectric or mechanical) measurement methods. They are widely used to investigate, say, relaxation processes of the material.

In calorimetry there are three established temperature modulation methods:

1) The sample is heated with a periodically changing heat flow from a electrical heater [47] or chopped light beam [48], and the temperature change of the sample (magnitude and phase shift) is measured as the response signal (*AC calorimetry*).
2) A thermal wave is sent into the sample from a metallic film that is evaporated onto its surface and which heats it periodically. The temperature change is measured by using the same metal film as a resistance thermometer. The third harmonic of the resulting temperature oscillations is proportional to the power input and depends on the product of the heat capacity and thermal conductivity of the sample in a characteristic way (*3ω method* [49]).
3) The sample (or furnace) temperature is controlled to follow a set course with superimposed periodical changes, and the heat flow rate is measured via the differential temperature between sample and reference (*temperature-modulated differential scanning calorimetry, TMDSC* [34, 50, 51]).

In all three cases, both the magnitude of the periodic part of the response signal and its phase shift with respect to the stimulating signal are measured, and this results in a complex quantity (temperature, heat flow rate, and heat capacity, respectively). As heat transport always requires time, the frequency of temperature modulation is normally limited to a maximum of 0.2 Hz to allow the heat to flow through the sample properly. For the 3ω method, however, frequencies of several kilohertz are possible because the heat source and the temperature probe are identical in this case. At the lower end, the frequency range is only limited by the noise threshold (sensitivity) of the sensors in question. As a rule only two decades are available. Consequently, only time-dependent processes with timescales within this window can be followed in the calorimeter.

For quasistatic conditions it follows from Equation (1) that:

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = \Phi = C_p \cdot \frac{\mathrm{d}T}{\mathrm{d}t} \tag{3}$$

where the real quantities heat flow rate Φ and temperature change (heating rate) are strictly proportional, with the real heat capacity C_p as proportionality factor. This implies an infinite propagation velocity of thermal waves and an infinitely fast excitation time of the vibrational states that characterize the (static) heat capacity of a sample. Both conditions are fulfilled for normal solids within the above-mentioned frequency range. In cases where processes are involved which need time, the heat flow rate and the temperature change are no longer proportional, and the process in question causes the heat flow to lag behind the temperature. This results formally in a time-dependent apparent heat capacity, and Equation (3) no longer holds; instead:

$$\Phi(t) = C_p(t) * \frac{\mathrm{d}T(t)}{\mathrm{d}t} \tag{4}$$

This appears very similar, but the operator * stands for the convolution product, which is an integral operator rather than a product [31, 52].

Equation (4) can be Fourier transformed and then reads:

$$\Phi(\omega) = C_p(\omega) \cdot \mathrm{Fourier}\left(\frac{\mathrm{d}T(t)}{\mathrm{d}t}\right) \tag{5}$$

The (complex) heat flow rate is given as a normal product of the complex heat capacity and the (complex) Fourier-transformed heating rate in the frequency domain. For a given $T(t)$ course (normally periodical with underlying constant heating rate), the complex heat capacity can be determined from the measured complex (magnitude and phase) heat flow rate function. The resulting C_p function is given either as real and imaginary part or as magnitude (absolute value) and phase. The two forms are mathematically equivalent. From the frequency dependence of these quantities the $C_p(t)$ behavior can be derived [31].

2.3. Calorimeters

There are many diverse types of calorimeters, differing with respect to measuring principle, mode of operation, and general construction. Measuring principles have been described in the preceding sections. In terms of mode of operation, two approaches can be distinguished: *static calorimeters*, in which the temperature remains constant or changes only in consequence of the heat of reaction, and *dynamic calorimeters*, in

which the temperature of the calorimeter substance is varied deliberately according to a prescribed program (usually linear with time).

With regard to construction, a distinction can be made between *single* and *differential calorimeters*. Twin design is characteristic of the latter: that is, all components are present in duplicate, and the two sets are arranged as identical as possible. The reaction under investigation takes place in one side of the device, while the other side contains an inert reference substance. The output signal represents the difference between signals originating in the sample and reference side. With such a difference signal all symmetrical effects cancel (e.g., the heat leaks to the surroundings).

2.3.1. Static Calorimeters

Considering first the static calorimeters, three types can be distinguished depending upon whether they are operated in an *isothermal, isoperibolic* (isoperibolic = uniform surroundings), or *adiabatic* manner.

2.3.1.1. Isothermal Calorimeters

In the *isothermal* mode of operation it is imperative that all thermal effects be somehow compensated. This is achieved either electrically or with the aid of a phase transition of some pure substance. Only phase-transition calorimeters can be regarded as strictly isothermal. In this case thermodynamics ensures that the temperature will remain precisely constant since it is controlled by a two-phase equilibrium of a pure substance. The most familiar example is the *ice calorimeter*, already in use by the end of the 18th century and developed further into a precision instrument about 100 years later by Bunsen (Fig. 16). The liquid – gas phase transition has also been used for thermal compensation purposes; in this case a heat of reaction can be determined accurately by measuring the volume of a vaporized gas.

Phase-transformation calorimeters are easy to construct, and for this reason they are not available commercially. Such a device permits very precise determination of the heat released during a process. An important drawback is the fact that the isothermal method is limited to the few fixed temperatures corresponding to phase transitions of suitable pure substances.

This disadvantage does not apply to an isothermal device based on electrical compensation. Nevertheless, calorimeters of the latter type operate only in a quasi-isothermal mode, since electronic control systems depend for their response upon small deviations from an established set point, and a certain amount of time is required for changing the prevailing temperature. The use of modern circuits and components ensures that errors from this source will be negligible, however. Electrical compensation makes it possible to follow both endothermic and exothermic processes. In both cases the compensation power is readily measured and recorded or processed further with a computer. Isothermal calorimeters are used quite generally for determining heats of mixing and solution. Commercial devices are available that also support the precise work required for multiphase thermodynamics. As a rule such calorimeters are of the single rather than the differential type.

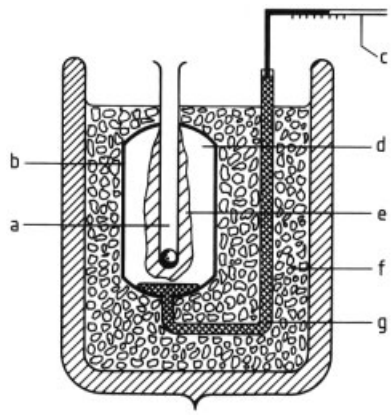

Figure 16. Ice calorimeter (based on [53])
a) Sample holder; b) Calorimeter vessel; c) Capillary with mercury thread; d) Water; e) Ice; f) Ice – water mixture; g) Mercury

2.3.1.2. Isoperibolic Calorimeters

An *isoperibolic calorimeter* is distinguished by the fact that it is situated within a thermostated environment. As a result, the temperature of the surroundings is kept constant, as is the effect of

the surroundings on the calorimeter vessel. A well-defined thermal conduction path is maintained between the calorimeter vessel and its surroundings so that the temperature difference between the calorimeter vessel and the environment will be proportional to the heat flow rate. Temperature equilibration usually adheres to an exponential time function, and an integral taken over the entire course of this function is proportional to the heat of reaction. Such instruments often have exceptionally large time constants; that is, temperature equilibration occurs very slowly. Indeed, the greater the equilibration time, the more sensitive is the calorimeter. It is of course difficult with an exponential function to identify precisely the end point, so the integral over the measurement function and therefore the heat of interest are subject to corresponding degrees of uncertainty. The advantages lie in simplicity of construction and ease of producing a measurement signal (a straightforward temperature difference), which might for example be derived from a differential thermocouple and a simple set of amplifiers.

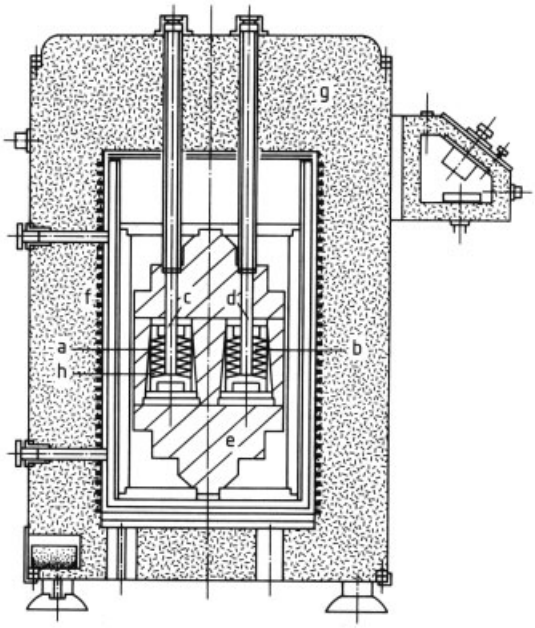

Figure 17. Tian – Calvet calorimeter
a, b) Heat conduction paths between measuring cells and block; c, d) Sample and reference containers; e) Isoperibolic block; f) Thermostatic jacket; g) Thermal insulation; h) Thermopiles

The best-known calorimeter of this type was developed by Tian and Calvet (Fig. 17). Here the defined heat-conduction path to the thermostated surroundings (a large aluminum block) consists of a large number of differential thermocouples coupled in series (thermopile). This arrangement permits optimum determination of the heat flow rate to the surroundings, and such an instrument can be very sensitive (*microcalorimeter*).

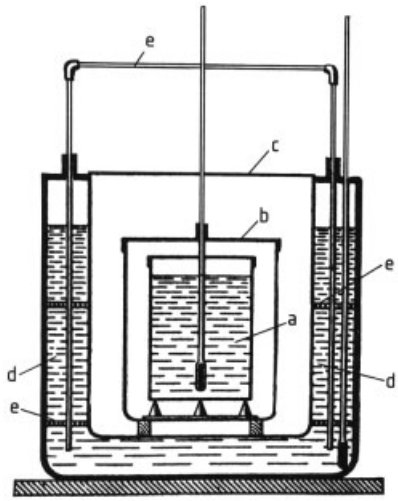

Figure 18. A classical mixing calorimeter (from Meyers Conversationslexikon, 1893)
a) Calorimeter vessel; b) Convection and radiation shield; c) Cover; d) Jacketing vessel (thermostated); e) Lifting mixer

If heat transport to the surroundings is restricted by suitable insulation, temperature equilibration becomes so slow that – with small corrections – temperature changes in the calorimeter substance alone can serve as the basis for determining the heat of a reaction. Calorimeters in this category include the classical *mixing calorimeter* (with heat exchange) in which the sample substance is brought into contact with a (usually) colder calorimeter substance, whereupon an intermediate "mixing temperature" is established (Fig. 18). Other examples of mixing calorimeters include the simpler types of *combustion calorimeters*, in which the calorimeter substance is water, and the *drop calorimeter* (see Fig. 24). The latter consists of a metal block whose temperature change subsequent to introduction of a sample of accurately known tem-

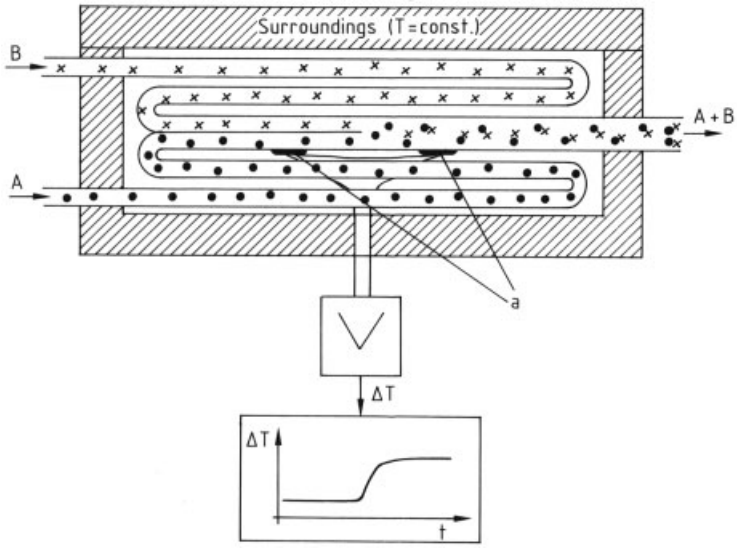

Figure 19. Flow calorimeter, with a chemical reaction occurring in the liquid
A, B reactants, A + B reaction product; a) Differential thermocouple

perature provides information about the sample's heat capacity. The category of isoperibolic calorimeters also includes *flow calorimeters*, in which two fluid media (usually liquids containing the reactants) are first brought to the temperature of the thermostated surroundings and then combined. The ensuing reaction causes a measurable temperature change, the extent of which is proportional to the heat of the reaction (Fig. 19).

2.3.1.3. Adiabatic Calorimeters

An *adiabatic calorimeter* is equipped with an electronic control system designed to ensure to the greatest extent possible that the surroundings of the calorimeter cell remain at all times at precisely the same temperature as the substance under examination. The object is to prevent virtually all heat transport to the surroundings, thereby making it possible to establish the total heat effect from the change in temperature of the calorimeter substance. Good adiabatic calorimeters are difficult to construct, and they are not available commercially. Nevertheless, precision instruments of this type permit quantities such as specific heat capacities, as well as associated anomalies, to be measured very accurately (Fig. 20). The method entails an extraordinary level of effort, and it is practiced only by specialists working in a few laboratories in the world [54].

Simpler *adiabatic calorimeters* are distributed commercially in the form of adiabatic mixing and combustion calorimeters. These constitute very satisfactory routine instruments, especially since the adiabatic mode of operation (which entails measurement of only a single temperature difference) facilitates automated data acquisition and interpretation.

2.3.2. Scanning Calorimeters

Scanning calorimeters are devices in which the sample temperature is deliberately changed during the course of an experiment according to a prescribed program. As a rule, heating or cooling of the sample is caused to occur linearly with time, and the heat flow rate required to accomplish the desired change is determined and output as a function of time or temperature. Any chemical or physical change in the sample at a particular temperature will manifest itself as a change in the heat flow rate, leading to a "peak" in the measurement curve whose area is proportional to the heat of the process (see also Section 1.3.2).

Such instruments have been on the market since the 1950s. Their great advantage is that

they permit heats of reaction and transformations of every type to be measured both rapidly and with reasonable precision. Heating rates usually fall in the range 5 – 20 K/min, so a measurement extending over 300 K can be completed within an hour, providing a reliable indication of all changes in the sample over this temperature range, including heats of reaction. The rapid development of calorimetry in recent years is due largely to the perfection of this type of calorimeter.

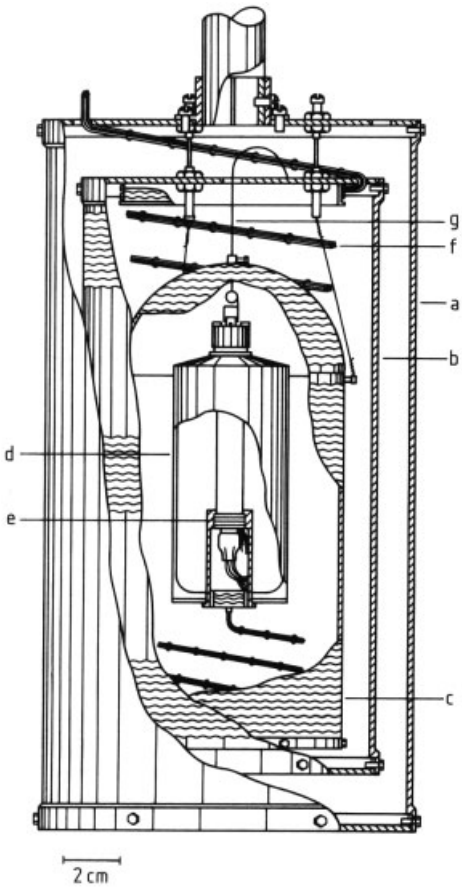

Figure 20. Construction of an adiabatic calorimeter (based on [55])
a) External radiation shield; b) Internal radiation shield; c) Adiabatic jacket; d) Calorimeter vessel; e) Sample suspension platform with resistance thermometer and heater; f) Electrical leads; g) Calorimeter suspension system

Scanning calorimeters are designed almost without exception as differential devices. With regard to the principle of measurement, two varieties can be distinguished: *differential-temperature* and *power-compensated calorimeters*. Although the two types rely upon very different approaches to determining heat flow rates with respect to the sample, and are therefore subject to different sources of error, they are marketed under the single heading *differential scanning calorimeter* (DSC). They can even be operated in *temperature-modulated* mode.

2.3.2.1. Differential-Temperature Scanning Calorimeters

In a differential-temperature scanning calorimeter the heat necessary for raising the sample temperature reaches the sample position via one well-defined heat conduction path (cf. Fig. 21). The measured signal is derived from a temperature difference generated along this path, which is proportional to the heat flow rate [see Eq. (2)]. This measuring and operating principle leads to the alternative designation *heat-flux calorimeter* for such a device. In order to minimize errors from heat leaks and other influences of the surroundings, calorimeters of this type are always based on a differential design, with full duplex construction. Many commercial models are available.

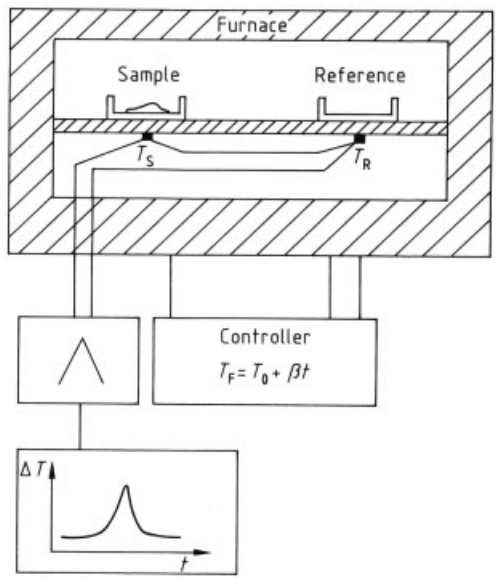

Figure 21. Operating principle of a temperature-difference scanning calorimeter (heat-flux calorimeter)
T_F furnace temperature, T_S sample temperature, T_R reference temperature, T_0 initial temperature, ΔT differential temperature $T_S - T_R$, β heating rate, t time

Any temperature difference between the sample and the reference is directly proportional to the corresponding differential heat flow rate, and so this is the quantity to be measured. The appropriate functional relationship follows directly from Equation (2):

$$\Phi = K(T) \cdot \Delta T \tag{6}$$

The heat flow rate of interest, Φ (in W), is proportional to the measured temperature difference ΔT (in K). Unfortunately, the proportionality constant K (the calibration factor for the calorimeter in K W^{-1}) is itself a function of the thermal conductivity of the sample [see Eq. (2)], and therefore temperature dependent. Moreover, temperature differences between the sample and the reference result in variable amounts of heat loss to the surroundings via inevitable heat leaks. Heat exchange with the environment also occurs via radiation and convection. Since the two transport processes are linked to temperature differences in a nonlinear way, the calibration factor in the presumably linear calorimeter equation necessarily depends on ΔT, which is the quantity to be measured. Consequently, the calibration factor depends upon such sample parameters as mass, thermal conductivity, and specific enthalpy difference [43], so heats measured with these instruments must realistically be assigned an uncertainty of about five percent unless considerable effort has been expended in calibration [56]. Repeatability of the results is generally many times better, but the user should not make the mistake of assuming an equally high level of certainty. Another source of systematic error may develop during the measurement if some reaction causes a change in the heat capacity and/or the extent of heat transfer to the sample holder. Both phenomena contribute to changes in the baseline, with the result that the course of the latter cannot be plotted accurately (Fig. 22). This in turn means that the area corresponding to the heat of reaction also cannot be determined exactly, leading to uncertainty in the associated heat [43]. Despite these restrictive comments, differential-temperature scanning calorimeters offer the advantages of a simple, transparent mode of operation and relatively low cost, and for this reason they account for a large share of the market. With appropriately critical supervision they are capable of fulfilling their assignment admirably. In 2002 a new type of heat-flux calorimeter with "Tzero® technique" [57] became commercially available. This DSC uses the power of modern computer technology to correct for some of the mentioned sources of error resulting in a higher certainty and better resolution of the measured curves and an excellent (almost zero) baseline.

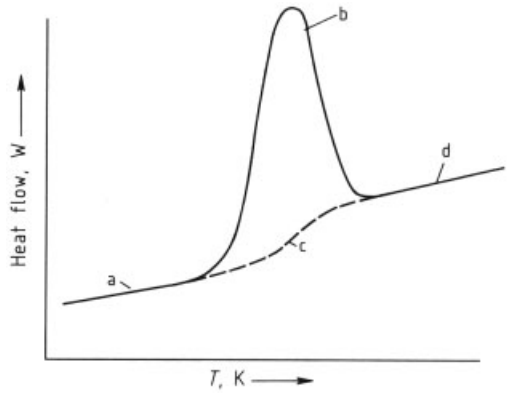

Figure 22. Output curve from a scanning calorimeter showing a shift in the baseline caused by variability in either c_P or the coefficient of heat transfer
a) Original baseline; b) Peak; c) Unknown baseline path; d) Final baseline

2.3.2.2. Power-Compensated Scanning Calorimeters

Like differential-temperature devices, *power-compensated scanning calorimeters* always feature duplex construction, with sample and reference holders heated in such a way that at every instant the temperature of each corresponds almost exactly to a programmed set temperature. Any temperature differences that arise between sample and reference by reason of thermal events in the sample are compensated immediately by appropriate changes in the electrical heating (Fig. 23). The output signal is proportional to the instantaneous differential heat flow rate. Since temperature differences between sample and reference are essentially compensated at once, measurement errors due to undetected heat leaks can be largely avoided. Instruments of this type, the most widely distributed of which are manufactured by Perkin – Elmer instruments, require a very complicated electronic control system, and are therefore rather

expensive. The principal disadvantage lies in the fact that the electronic system is a black box for the average layman, thereby obscuring possible sources of error. These instruments also require that there be some deviation, albeit minimal, from the set temperature to guide the control circuit, so just as in the preceding case some systematic error must be anticipated (although certainly to a lesser degree), associated once again with temperature differences between sample and reference and nonlinear heat exchange. Accordingly, the level of systematic uncertainty for heats measured by power-compensated DSC can be estimated at 1 – 3 %.

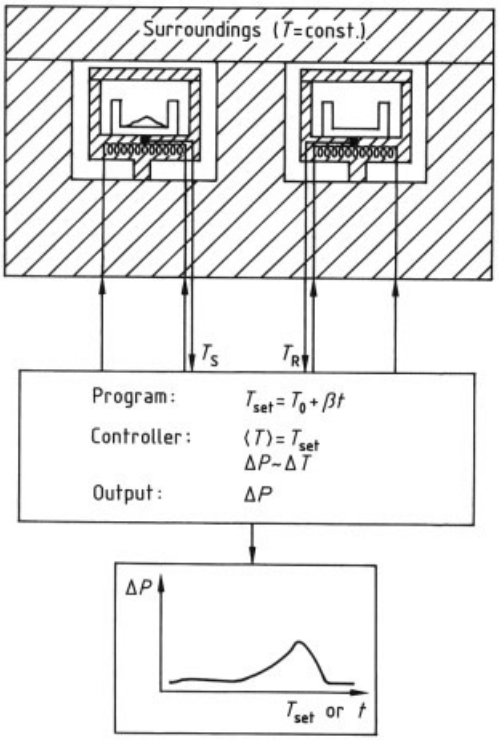

Figure 23. Operating principle of a power-compensated differential scanning calorimeter (Perkin – Elmer)
T_0 initial temperature; T_S sample temperature; T_R reference temperature; $\langle T \rangle$ mean temperature between sample and reference; ΔP differential power $P_S - P_R$; β heating rate; t time

2.3.2.3. Temperature-Modulated Scanning Calorimeters

Differential-temperature and power-compensated scanning calorimeters are now commercially available to run in the temperature modulated mode (TMDSC). In both cases the temperature is controlled to follow the normal (linear in time) course but with an additional periodical temperature change of particular amplitude and frequency. The measured heat flow rate function is the sum of the "underlying" part and the periodical part. The former can be extracted by calculating the "gliding average" integral over a period along the measured curve. It is identical to the curve that would be obtained in this DSC when the temperature modulation is switched off. The periodical part, on the other hand, is the difference of the measured and the underlying curves [31]. From the periodic part the complex C_p can be calculated by using Fourier analysis or other mathematical techniques. This enables time dependent-processes to be investigated in cases where their timescale falls within the frequency window of the TMDSC method [31].

For precise measurements it is necessary to correct (calibrate) the measured heat flow rate magnitude and phase shift for influences from the apparatus, as well as from the heat transport, which also is a time-dependent process and therefore influences the measured heat flow rate function [31, 58, 59].

2.3.3. Chip-Calorimeters

In the early 2000s a huge progress has been made in miniaturization of electronic equipment, in particular in the field of integrated circuits. It became more and more easy to produce "chips" with thin Si or SiN membranes as basis which can be used as micro- or nanocalorimeters. Such *chip-calorimeters* have on the one hand a very fast response time (in the milliseconds region) and open the possibility to measure very small samples (nanograms) at heating and cooling rates up to 5000 K/s [60] (for details see [61] and other articles in this special issue of *Thermochimica Acta*). This technology opens very new fields of research on thermodynamic behavior of small particles as well as membranes on the one hand and the detection of very small amounts of substances with the help of specific chemical reactions on the other hand. The latter can be a large step toward the development of an "artificial nose" which may serve as a sensor to detect, say, dangerous gases in the surroundings.

2.4. Applications of Calorimetry

Modern calorimeters permit relatively rapid and truly precise measurement of heat exchanges in a wide variety of reactions. Since heat evolution is proportional to the conversion (extent of reaction) in a chemical, physical, or biological reaction, calorimetric measurement constitutes one method for quantitative evaluation of the reaction itself. Measurement is possible not only of the total heat (and therefore the total conversion) of a reaction but also of the course of the reaction with respect to time. This opens the way to studies of kinetics, problems related to stability and safety, and other areas in which temperature and time-dependencies in a reaction play a part.

The sections that follow describe several possible applications of calorimetry, although these should be regarded only as illustrative, since the scope of the present article rules out any treatment that is even approximately exhaustive. No such limits apply to the reader's imagination, however, and any process that generates (or consumes) heat can in principle be followed with a calorimeter. Enzymatic and bacterial manufacturing processes have assumed increasing importance in industry, and interest in biocalorimetry is growing accordingly [62 – 64].

2.4.1. Determination of Thermodynamic Functions

Various *thermodynamic potential functions* (enthalpy, entropy, free energy, etc.) can be determined by integration on the basis of *specific heat capacities*. The accurate establishment of specific heat capacities at constant pressure, c_p, as a function of temperature is thus of fundamental importance. Different types of calorimeters are used for determinations of c_p depending upon accuracy requirements and the temperature range in question.

The most accurate results are obtained with adiabatic calorimeters, which are useful over a temperature range from fractions of a Kelvin to about 1000 K. These instruments permit specific heat capacities to be determined with very small uncertainties (<1 %). Nevertheless, such measurements are so complex that they are rarely applicable to normal laboratory practice in industry, especially given the lack of commercial instruments.

The standard approach to c_p measurement therefore relies on the less precise but more easily constructed and operated mixing calorimeters, in particular drop calorimeters (Fig. 24). A (mean) specific heat capacity is calculated in this case from the heat released when a sample of known temperature is introduced (dropped) into an isoperibolic calorimeter. Charging temperatures may exceed 2500 K. Determining the temperature dependence of c_p requires many individual measurements at various charging temperatures. The uncertainty in such measurements is usually 3 – 5 %.

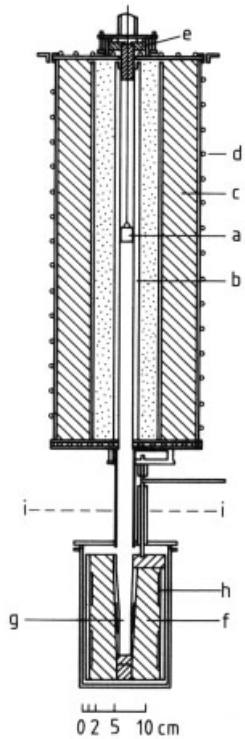

Figure 24. High-temperature drop calorimeter (based on [66])
a) Sample; b) Heated corundum tube; c) Magnesia insulation; d) Cooling coils; e) Sample mounting device; f) Copper block; g) Hole for receiving sample; h) Resistance thermometer; i) Level of the thermostated oil bath

A truly elegant method for determining c_p at <1000 K takes advantage of a scanning calorimeter [43], preferably a power-compensated DSC. The procedure involves raising the temperature of a sample in the calorimeter at a defined rate (typically 10 K/min). The

same measurement is subsequently repeated, this time with empty crucibles. The difference between the two measurement curves provides a record of the heat flow rate into the sample, from which the specific heat capacity can be calculated. This method makes it possible conveniently to measure the $c_p(T)$ of a material over a temperature range of several hundred Kelvins within the course of a single day. However, it does demand a very high level of repeatability (baseline stability) for the DSC in question. In the case of differential-temperature (heat-flux) DSC – even with careful calibration – uncertainties of 5 – 10 % in the results must be anticipated. With power-compensated DSC the systematic uncertainties tend to be in the range of 3 – 5 %. Uncertainties in the calculated thermodynamic potential functions of interest are comparable.

The TMDSC enables another elegant possibility of c_p (magnitude) determination from the amplitude of the modulated part of the measured heat flow rate function both in the isothermal and scanning modes of operation. This method is especially advantageous in cases of noisy signals with low sample masses or low heating rates. Precise calibration of the heat flow rate amplitude is a prerequisite for obtaining reliable results [31, 59, 60].

However, a calorimeter operating dynamically always indicates a temperature that deviates to a greater or lesser extent from the true sample temperature, resulting in further uncertainty. It must in any case be borne in mind that potential functions determined on the basis of thermodynamic equilibrium, and thus defined statically, may differ from functions determined dynamically, since equilibrium conditions are not maintained in DSC, and in the case of a reaction even stationary conditions are lacking.

If time-dependent processes are involved, then the process is clearly outside the scope of classical thermodynamics. The time-dependent (apparent) heat capacity, measured with, say, TMDSC would lead to time-dependent potential functions which must be interpreted in terms of *irreversible thermodynamics*. In such cases, a nonzero imaginary part of the complex heat capacity exists which is linked to the entropy production of the process in question (for details see [44]). Thus, temperature-modulated calorimetry makes it possible to determine time-dependent (irreversible) thermodynamic quantities.

2.4.2. Determination of Heats of Mixing

For liquid systems consisting of several components, multiphase thermodynamics provides the theoretical background for computing phase diagrams from heats of mixing for the relevant components. Phase diagrams in turn are interesting not only to thermodynamicists; they are also of great importance in chemical process engineering practice.

The determination of heats of mixing is a classical field of investigation for calorimetry, one that at the required level of precision is today incomparably less laborious than previously. Enthalpies of mixing were formerly determined in some cases indirectly via vapor pressure measurements, but today they are measured directly in (quasi)isothermal or isoperibolic calorimeters. Recently also chip-calorimeters have been used for this purpose [65].

The art of conducting such an experiment successfully lies in rigorously excluding vaporization during the measurement, since specific heats of vaporization are orders of magnitude larger than the effects of mixing that are to be measured. As excess quantities (which characterize deviations from ideal behavior), heats of mixing are invariably small, and they must be determined with high precision in order to permit the construction of sufficiently precise phase diagrams. Isothermal calorimeters are therefore definitely to be preferred, since here the problem of incomplete determination of a heat of mixing – owing perhaps to undetected heat leaks – is less severe than with an isoperibolic instrument. Moreover, the isothermal heat of mixing is defined more precisely, and it lends itself more readily to mathematical treatment than in the case of parallel temperature changes.

Phase transitions in solid mixtures can be followed directly in a scanning calorimeter, since vapor pressures associated with the components are here so low that vaporization plays virtually no role and therefore does not interfere. Corresponding phase diagrams (i.e., liquidus – solidus curves) can be inferred from heat flow rate curves obtained by controlled heating and cooling of appropriate mixtures (cf. Section

1.3.3). Thermodynamically relevant quantities can in turn be determined indirectly by comparing measured phase diagrams with calculated diagrams.

2.4.3. Combustion Calorimetry

The determination of heats of combustion for fuels of all types is prescribed in detail by government regulation, since the "calorific value" or "heat content" of a particular coal, oil, or natural or manufactured gas determines its economic value. Combustion calorimeters or "bomb calorimeters" (usually operated adiabatically) are used for such measurements, whereby the sample is combusted in a "calorimetric bomb" (likewise standardized; cf. Fig. 25) in pure oxygen at $3 \cdot 10^6$ Pa to a defined set of end products. Methods of measurement and interpretation are also standardized, and they have been simplified to such an extent that the entire procedure can be automated.

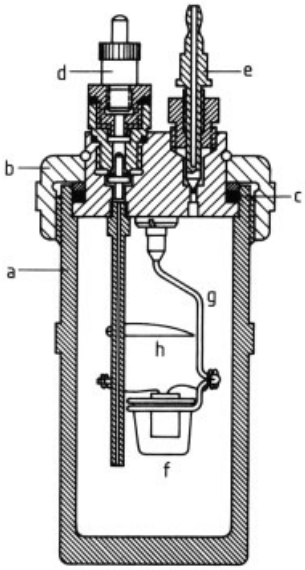

Figure 25. Example of a calorimetric bomb for use in a combustion calorimeter
a) Pressure vessel; b) Screw cap; c) Sealing ring; d) Inlet valve; e) Outlet valve; f) Sample crucible with sample; g) Crucible holder; h) Heat shield

The same instruments are useful for determining heats of combustion for a wide variety of (usually organic) substances, from which standard enthalpies of formation can be determined. For these calorimetric methods errors in measurement are very small (< 1 ‰), with a level of precision that is also appropriate for analytical purposes (e.g., purity or composition checks on batches of chemicals). Combustion calorimetry is useful in biology as well. For example, in determinations of the nutrient content of leaves and fruits as a function of season and place based on appropriate heats of combustion.

Many combustion calorimeters are available commercially. Instruments utilizing water as the calorimeter substance generally operate adiabatically, thereby avoiding the somewhat complicated Regnault – Pfaundler procedure (cf. [41]) for determining and interpreting prior and subsequent temperature patterns. "Dry" combustion calorimeters that avoid the use of water or other liquids have also come onto the market.

2.4.4. Reaction Calorimetry

Every chemical process is associated with some type of heat effect. Accurate knowledge of heat released or consumed is of great importance in process engineering. Moreover, heat evolution is strictly coupled with the particular course of a reaction. If one is able to determine unambiguously the time dependence of the heat flow rate associated with a particular sample, one also gains access to the corresponding reaction rate law, which in turn opens the way to the reaction kinetics [43]. Knowledge of the appropriate kinetic parameters permits one to predict the course of the reaction under other sets of conditions, which is of great practical value. Heat of reaction determinations as well as kinetic studies are best carried out under isothermal (and isobaric) conditions, and for this reason it is advisable to select a (quasi)isothermal calorimeter with compensation for effects of the measurement process. One practical difficulty lies in the fact that reaction begins immediately after introduction of the sample at the appropriate temperature, even though the calorimeter itself is not yet at thermal equilibrium (and therefore not ready for use). One way out of this dilemma is to devise a system for bringing the components of the reaction together only after isothermicity

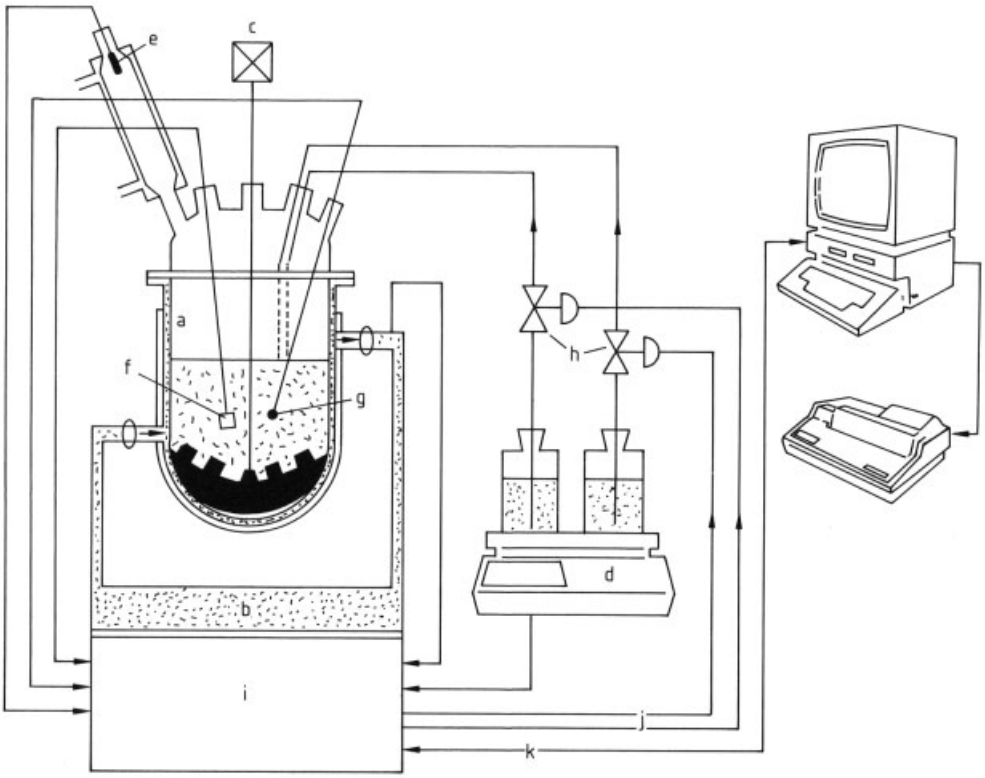

Figure 26. Operating principle of a reaction calorimeter (Contraves)
a) Reaction vessel; b) Heating and cooling bath; c) Stirrer; d) Balance; e) Pressure sensor for internal pressure; f) pH probe; g) Temperature sensor; h) Dosing valve; i) Control unit; j) Control and monitoring signals; k) Measured signal

has been achieved. With liquids the challenge is not particularly great, since pumps and stirrers can ensure rapid (and thorough) mixing of the reactants at any desired time. The piercing of a membrane (or an ampoule) is also a common expedient. Another possibility is the use of a specially designed reaction calorimeter. With one device of this type the first reactant is present from the outset in a suitable glass or metallic vessel, to which a second reactant is added with mixing after stationary conditions have been established (Fig. 26). With the aid of electronic temperature control the reaction can then be carried out in any way desired: isothermally, adiabatically, or under the influence of some temperature program. Measurement of the relevant quantities, process control, and thermochemical data interpretation are all accomplished with the aid of a personal computer. A second type of special reaction calorimeter, suitable only for liquid components, is so constructed that each reactant is pumped in through its own system of tubing, with the streams uniting after temperature equilibration (Fig. 27). The ensuing reaction causes a temperature change, from which the corresponding heat of reaction can be calculated.

Transport and mixing of liquid or gaseous components is accomplished rather easily, but the situation is more complex for solid substances or mixtures. Here there is little hope of satisfactorily mixing the components inside a normal calorimeter, so one is forced to resort to a scanning calorimeter. The calorimeter is charged at a temperature below that at which the substances react, after which the temperature is increased linearly with time until reaction is complete. From a heat flow rate – time function acquired at a well-defined heating rate it is possible to establish both the heat of reaction and, if need be, kinetic parameters for the process. One prerequisite, however, is that the recorded function reproduces in an unbiased way the true heat production of the sample. This presents serious problems, since heat must in-

evitably flow from the point at which it is generated to the site of the temperature sensors before any measurement signal can appear. In the case of a fairly rapid chemical process, where the rate constant is comparable to the time constant of the calorimeter, the measurement signal from the recording device must be treated mathematically ("desmeared") [43] before it can be utilized for a kinetic analysis.

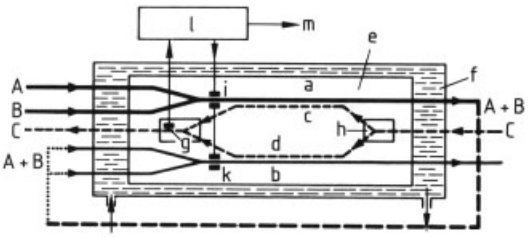

Figure 27. Schematic diagram of a flow calorimeter after P. PICKER (Setaram)
A, B) Reactants; A + B) Reaction mixture; C) Heat-exchange liquid
a) Reaction tube (sample tube); b) Reference tube; c, d) Heat-exchange tubes; e) Vacuum jacket; f) Thermostated liquid; g) Thermistor; h) Flow divider; i, k) Heating elements; l) Measurement and control electronics; m) Output signal

2.4.5. Safety Studies

Calorimetric methods applicable to safety studies are also concerned with clarifying the kinetic behavior of substances, but in this case from a safety engineering viewpoint.

The vast majority of chemical substances are endothermic materials; that is, they release heat when they decompose, and their decomposition is accelerated by heat. Chemical reaction kinetics teaches that any reaction is capable of proceeding at temperatures considerably lower than the temperature at which the maximum reaction rate is achieved. Usually, however, the reaction rate under these conditions is so low that the process is not apparent. The more unstable a substance, the lower is its decomposition temperature, and therefore the greater its decomposition rate at room temperature. In the event that heat evolved during storage – attributable to traces of reaction occurring under storage conditions – exceeds the level at which heat is emitted to the surroundings, then the decomposition rate may increase very rapidly, possibly to the point of an explosion.

Thorough knowledge of decomposition kinetics is therefore essential for ensuring safe storage and handling. The requisite information can, in principle, be gained through the methods described in Section 2.4.4, but it is often impossible to avoid an uncontrolled outcome and simultaneous destruction of the measuring system. This obviously rules out the use of expensive apparatus, which has led to the development of special safety calorimeters that are either inexpensive and disposable or designed to withstand the effects of violent reactions.

Simple arrangements of this type clearly preclude precision measurements, but they are entirely adequate from the specific viewpoint of assessing storage safety at particular temperatures or with particular types of packaging. The simplest example is an exceedingly primitive isoperibolic calorimeter consisting of a glass ampoule, either open or fitted with a pressure-tight closure, mounted together with a thermocouple inside a small furnace – a device that anyone can easily assemble and calibrate. More ambitious designs differ from standard calorimeters in that the sample holder (or measuring head) is incorporated into an autoclave of sufficient stability to permit studying the thermal behavior even of explosives. Included in this category are the so-called accelerating rate calorimeters (ARC), in which a sample is heated very slowly under quasiadiabatic conditions until the reaction intensifies by itself as a result of exothermic decomposition. The resulting temperature – time function can then be used to establish the corresponding heat effect and stability behavior.

3. References

1. J. O. Hill (ed.): "For Better Thermal Analysis and Calorimetry," 3rd ed., International Confederation for Thermal Analysis and Calorimetry, 1991.
2. W. Wendlandt: *Thermal Analysis*, 3rd ed., Wiley-Interscience, New York 1986.
3. M. E. Brown: *Introduction to Thermal Analysis*, Chapman and Hall, London 1988.
4. R. C. Mackenzie (ed.): *Differential Thermal Analysis*, Academic Press, London, vol. 1, 1970, vol. 2, 1972.

5. M. E. Brown (ed.): *Handbook of Thermal Analysis and Calorimetry*, **vol. 1** , Elsevier, Amsterdam 1998.
6. M. Reading in [5], pp. 423 – 443.
7. S. P. Wolsky, A. W. Czanderna (eds.): *Microweighing in Controlled Environments*, Elsevier, Amsterdam 1980.
8. ASTM Standard E 914–83.
9. C. L. Duval: *Inorganic Thermogravimetric Analysis*, Elsevier, Amsterdam 1953.
10. M. R. Ottaway, *Fuel* **61** (1982) 713.
11. C. M. Earnest (ed.): *Compositional Analysis by Thermogravimetry*, ASTM STP 997, Philadelphia 1988.
12. J. W. Cumming, J. McLaughlin, *Thermochim. Acta* **57** (1982) 253.
13. E. Turi (ed.): *Thermal Characterization of Polymeric Materials*, Academic Press, San Diego 1997.
14. R. L. Blaine, C. K. Schoff (eds.): *Purity Determination by Thermal Methods*, ASTM STP 838, Philadelphia 1984.
15. M. Reading, *Trends Polym. Sci.* **1** (1993) 248.
16. F. Paulik, J. Paulik, *Analyst (London)* **103** (1978) 417.
17. W. Lodding (ed.): *Gas Effluent Analysis*, Marcel Dekker, New York 1967.
18. S. B. Warrington in E. L. Charsley, S. B. Warrington (eds.): *Thermal Analysis – Techniques and Applications*, Royal Society of Chemistry, London 1992.
19. M. R. Holdiness, *Thermochim. Acta* **75** (1984) 361.
20. E. L. Charsley, C. Walker, S. B. Warrington, *J. Therm. Anal.* **40** (1993) 983 (Proc. 10th ICTAC Conf., Hatfield, 1992).
21. V. Balek, J. Tolgyessy in C. L. Wilson, D. W. Wilson (eds.): *Comprehensive Analytical Chemistry,* vol. 12, part C, Elsevier, Amsterdam 1984.
22. M. Reading in E. L. Charsley, S. B. Warrington (eds.): *Thermal Analysis – Techniques and Applications*, Royal Society of Chemistry, London 1992.
23. T. Lever et al., *American Laboratory* **30** (1998) 15.

General References for Chapter 2

24. W. Zielenkiewicz, E. Margas: *Theory of Calorimetry*, Kluwer Academic Publishers 2002.
25. P. Haines, J. P. Haines (eds.): *Principles of Thermal Analysis and Calorimetry*, The Royal Society of Chemistry, London 2002.
26. *Handbook of Thermal Analysis and Calorimetry*, Elsevier Science, Amsterdam 1998.
27. IUPAC Series: *Experimental Thermodynamics*, Vol. IV, Blackwell Science, Oxford 1994.
28. IUPAC Series: *Experimental Thermodynamics*, Vol. VI, Elsevier, Amsterdam 2003.
29. IUPAC Series: *Experimental Thermodynamics*, Vol. VII, Elsevier, Amsterdam 2005.
30. J. E. Ladbury, B. Z. Chowdhry: *Biocalorimetry*, John Wiley and Sons, New York 1998.
31. J. E. Ladbury, M .L. Doyle: *Biocalorimetry II*, John Wiley and Sons, 2004.
32. M. Sorai: *Comprehensive Handbook of Calorimetry and Thermal Analysis*, John Wiley and Sons, New York 2004.
33. A. E. Beezer: *Biological Microcalorimetry*, Academic Press, London 1980.
34. S. Gaisford, G. Gaisford: *Pharmaceutical Isothermal Calorimetry*, CRC Press, Florida 2006.
35. R.-Y. Zhu (ed.): *Calorimetry in Particle Physics*, World Scientific Publishing Company, New Jersey, London 2003.
36. W. Hemminger, G. Höhne: *Calorimetry, Fundamentals and Practice*, Verlag Chemie, Weinheim, Germany 1984.
37. E. Koch: *Non-Isothermal Reaction Analysis*, Academic Press, London 1977.
38. J. P. McCullough, D. W. Scott (eds.): *Experimental Thermodynamics,* vol. 1: "Calorimetry of Non-reacting Systems," Butterworths, London 1968.
39. B. Le Neindre, B. Vodar (eds.): *Experimental Thermodynamics,* vol. 2, Butterworths, London 1975.
40. H. Spink, I. Wadsö: "Calorimetry as an Analytical Tool in Biochemistry and Biology," in D. Glinck (ed.): *Methods of Biochemical Analysis,* vol. 23, J. Wiley & Sons, New York 1976.
41. S. Sunner, M. Månsson (eds.): "Combustion Calorimetry," *Experimental Chemical Thermodynamics*, vol. 1, Pergamon Press, Oxford 1979.
42. H. Weber: *Isothermal Calorimetry*, Herbert Lang, Bern, Peter Lang, Frankfurt 1973.
43. G. W. H. Höhne, W. Hemminger, H. J. Flammersheim, *Differential Scanning Calorimetry*, 2nd ed. Springer-Verlag, Berlin, 2003.
44. C. Schick, G. W. H. Höhne (eds.): "Temperature-Modulated Calorimetry", *Thermochim. Acta* **304/305** (1997) special issue.

Specific References

45. E. Gmelin: "Classical Temperature-Modulated Calorimetry: A Review", *Thermochim. Acta* **304/305** (1997) 1 – 26.
46. P. S. Gill, S. R. Sauerbrunn, M. Reading, *J. Thermal Anal.* **40** (1993) 931 – 939.
47. P. F. Sullivan, G. Seidel *Phys. Rev.* **173** (1968) 679.
48. P. Handler, D. E. Mapother, M. Rayl, *Phys. Rev. Lett.* **19** (1967) 356.
49. O. M. Corbino, *Phys. Z.* **11** (1910) 413; **12** (1911) 292.
50. H. Gobrecht, K. Hamann, G. Willers, *J. Physics E: Sci. Instr.* **4** (1971) 21 – 23.
51. M. Reading, D. J. Hourston: *Modulated-Temperature Differential Scanning Calorimetry*, Kluwer Academic Publication, 2006.
52. J. E. K. Schawe, G. W. H. Höhne, *Thermochim. Acta* **287** (1996). 213 – 223.
53. R. W. Bunsen, *Ann. Phys.* **141** (1870) 1.
54. S. Stølen, R. Glöckner, F. Grønvold, *J. Chem. Thermodyn.* **28** (1996) 1263.
55. E. D. West, E. F. WestrumJr. in J. P. McCullough, D. W. Scott (eds.): *Experimental Thermodynamics*, **vol. 1**, Butterworths, London 1968, p. 354.
56. S. M. Sarge et al.: "The Caloric Calibration of Scanning Calorimeters," *Thermochim. Acta* **247** (1994) 129 – 168.
57. R. L. Danley, *Thermochim. Acta* **395** (2003) 201–208.
58. G. W. H. Höhne, M. Merzlyakov, C. Schick, *Thermochim. Acta* **391** (2002) 51–67.
59. M. Merzlyakov, G.W.H. Höhne, C. Schick, *Thermochim. Acta* **391** (2002) 69–80.
60. S. Adamovsky, C. Schick, *Thermochim. Acta* **415** (2004) 1–7.
61. G. W. H. Höhne, W. Winter, *Thermochim. Acta* **432** (2005) 196–176.
62. J. Lamprecht, W. Hemminger, G. W. H. Höhne (eds.): "Calorimetry in the Biological Sciences," *Thermochim. Acta* **193** (Special Issue) (1991) 1 – 452.
63. R. B. Kemp, B. Schaarschmidt (eds.): "Proceedings of Biological Calorimetry", *Thermochim. Acta* **251/252** (1995) special issues.
64. J. A. McLean, G. Tobin: *Animal and Human Calorimetry*, Cambridge University Press, 1988.
65. J. Lerchner, R. Kirchner, J. Seidel, D. Waehlisch, G. Wolf, *Thermochim. Acta* **415** (2004) 27–34.
66. T. B. Douglas, E. G. King in J. P. McCullough, D. W. Scott (eds.): *Experimental Thermodynamics*, **vol. 1**, Butterworths, London 1968, p. 297.

Surface and Thin-Film Analysis

Henning Bubert, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Dortmund, Germany (Chaps. 1 and 2)

John C. Rivière, Harwell Laboratory, AEA Technology, Didcot, United Kingdom (Chaps. 1, 2; Sections 2.5, 3.4 and 4.4)

Heinrich F. Arlinghaus, Physical Institute, Westfälische Wilhelms-University, Münster, Germany (Sections 3.1.1 and 3.2.3)

Herbert Hutter, Institute of Analytical Chemistry, University of Technology, Vienna, Austria (Section 3.1.2)

Holger Jenett, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Dortmund, Germany (Sections 3.2.1 and 3.2.2)

Peter Bauer, Institute of Experimental Physics, Johannes Kepler University, Linz, Austria (Section 3.3)

Leopold Palmetshofer, Institute of Semiconductor and Solid State Physics, Johannes Kepler University, Linz, Austria (Section 3.3)

Laszlo Fabry, Wacker Siltronic AG, Burghausen, Germany (Section 4.1)

Siegfried Pahlke, Wacker Siltronic AG, Burghausen, Germany (Section 4.1)

Alfred Quentmeier, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Dortmund, Germany (Section 4.2)

Karsten Hinrichs, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Berlin, Germany (Chap. 4.3.1)

Wieland Hill, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Dortmund, Germany (Section 4.3.2)

Bernd Gruska, Sentech Instruments GmbH, Berlin, Germany (Chap. 4.3.3)

Arthur Röseler, Institute of Spectrochemistry and Applied Spectroscopy (ISAS), Berlin, Germany (Chap. 4.3.3)

Gernot Friedbacher, Institute of Analytical Chemistry, University of Technology, Vienna, Austria (Chap. 5)

1. Introduction 5
2. Electron Detection 8
2.1. X-Ray Photoelectron Spectroscopy (XPS) 8
2.1.1. Principles 8
2.1.2. Instrumentation 10
2.1.2.1. Vacuum Requirements 10
2.1.2.2. X-Ray Sources 11
2.1.2.3. Synchrotron Radiation 12
2.1.2.4. Electron Energy Analyzers 13
2.1.2.5. Spatial Resolution 14
2.1.3. Spectral Information and Chemical Shifts 15
2.1.4. Quantification, Depth Profiling, and Imaging 16
2.1.4.1. Quantification 16
2.1.4.2. Depth Profiling 17
2.1.4.3. Imaging 18
2.1.5. The Auger Parameter 19
2.1.6. Applications 19
2.1.6.1. Catalysis 20
2.1.6.2. Polymers 21
2.1.6.3. Corrosion and Passivation 21
2.1.6.4. Adhesion 22
2.1.6.5. Superconductors 23
2.1.6.6. Interfaces 24
2.2. Ultraviolet Photoelectron Spectroscopy (UPS) 25
2.3. Auger Electron Spectroscopy (AES) 26
2.3.1. Principles 26
2.3.2. Instrumentation 27

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_023

2.3.2.1. Vacuum Requirements 27
2.3.2.2. Electron Sources 27
2.3.2.3. Electron Energy Analyzers 28
2.3.3. Spectral Information 29
2.3.4. Quantification and Depth Profiling 30
2.3.4.1. Quantification 30
2.3.4.2. Depth Profiling 33
2.3.5. Applications 33
2.3.5.1. Grain Boundary Segregation . . . 33
2.3.5.2. Semiconductor Technology 34
2.3.5.3. Thin Films and Interfaces 35
2.3.5.4. Surface Segregation 36
2.4. Scanning Auger Microscopy (SAM) 37
2.5. Other Electron-Detecting Techniques 38
2.5.1. Auger Electron Appearance Potential Spectroscopy (AEAPS) . . 38
2.5.2. Electron Energy Loss Methods . 38
2.5.2.1. Electron Energy Loss Spectroscopy (EELS) and Core-Electron Energy Loss Spectroscopy (CEELS) 38
2.5.2.2. High-Resolution Electron Energy Loss Spectroscopy (HREELS) . . 40
2.5.3. Diffraction Methods 40
2.5.3.1. Low-Energy Electron Diffraction (LEED) 40
2.5.3.2. Reflection High-Energy Electron Diffraction (RHEED) 41
2.5.4. Ion-Excitation Method 42
2.5.4.1. Ion (Excited) Auger Electron Spectroscopy (IAES) 42
2.5.4.2. Ion-Neutralization Spectroscopy (INS) 42
2.5.4.3. Metastable Quenching Spectroscopy (MQS) 42
2.5.5. Inelastic Electron Tunneling Spectroscopy (IETS) 43
3. Ion Detection 43
3.1. Secondary Ion Mass Spectrometry 43
3.1.1. Static Secondary Ion Mass Spectrometry (SSIMS) 43
3.1.1.1. Principles 44
3.1.1.2. Instrumentation 44
3.1.1.2.1. Ion Sources 44
3.1.1.2.2. Mass Analyzers 45
3.1.1.3. Quantification 48
3.1.1.4. Spectral Information 49
3.1.1.5. Applications 50
3.1.1.5.1. Oxide Films 50
3.1.1.5.2. Interfaces 51
3.1.1.5.3. Polymers 53
3.1.1.5.4. Biosensors 54
3.1.1.5.5. Surface Reactions 55
3.1.1.5.6. Imaging 55
3.1.1.5.7. Ultrashallow Depth Profiling . . . 57
3.1.2. Dynamic SIMS 57
3.1.2.1. Principles 57
3.1.2.2. Instrumentation 59
3.1.2.2.1. Ion Sources 59
3.1.2.2.2. Mass Analyzers 60
3.1.2.2.3. Detectors 61
3.1.2.3. Spectral Information 61
3.1.2.4. Quantification 61
3.1.2.5. Mass Spectra 62
3.1.2.6. Depth Profiles 62
3.1.2.7. Imaging 64
3.1.2.8. Applications 66
3.1.2.8.1. Implantation Profiles 66
3.1.2.8.2. Layer Analysis 67
3.1.2.8.3. 3D Bulk Element Distribution . . 68
3.2. Secondary Neutral Mass Spectrometry (SNMS) 68
3.2.1. General Principles 68
3.2.2. Electron-Beam and HF-Plasma SNMS 70
3.2.2.1. Principles 70
3.2.2.2. Instrumentation 71
3.2.2.3. Spectral Information 73
3.2.2.4. Quantification 73
3.2.2.5. Element Depth Profiling 75
3.2.2.6. Applications 75
3.2.3. Laser-SNMS 76
3.2.3.1. Principles 76
3.2.3.1.1. Nonresonant Laser-SNMS 76
3.2.3.1.2. Resonant Laser-SNMS 77
3.2.3.1.3. Experimental Setup 77
3.2.3.1.4. Ionization Schemes 78
3.2.3.2. Instrumentation 79
3.2.3.3. Spectral Information 79
3.2.3.4. Quantification 79
3.2.3.5. Applications 80
3.2.3.5.1. Nonresonant Laser-SNMS 80
3.2.3.5.2. Resonant Laser-SNMS 80
3.3. Ion-Scattering Techniques . . . 83
3.3.1. Rutherford Backscattering Spectroscopy (RBS) 83
3.3.1.1. Principles 83
3.3.1.2. Instrumentation 85
3.3.1.3. Spectral Information 86
3.3.1.4. Quantification 87
3.3.1.5. Applications 87
3.3.2. Low-Energy Ion Scattering (LEIS) 89
3.3.2.1. Principles 89
3.3.2.2. Instrumentation 90

3.3.2.3. Information 91
3.3.2.4. Quantification 93
3.3.2.5. Applications 93
3.4. Other Ion-Detecting Techniques 95
3.4.1. Desorption Methods 95
3.4.1.1. Electron-Stimulated Desorption (ESD) and Electron-Stimulated Desorption Ion Angular Distribution (ESDIAD) 95
3.4.1.2. Thermal Desorption Spectroscopy (TDS) 96
3.4.2. Glow Discharge Mass Spectroscopy (GDMS) 96
3.4.3. Fast Atom Bombardment Mass Spectroscopy (FABMS) 97
3.4.4. Atom Probe Microscopy 97
3.4.4.1. Atom Probe Field-Ion Microscopy (APFIM) 97
3.4.4.2. Position-Sensitive Atom Probe (POSAP) 98
4. Photon Detection 98
4.1. Total-Reflection X-ray Fluorescence Analysis (TXRF) 98
4.1.1. Principles 98
4.1.2. Instrumentation 100
4.1.3. Spectral Information 101
4.1.4. Quantification 102
4.1.5. Applications 104
4.1.5.1. Particulate and Film-Type Surface Contamination 104
4.1.5.2. Semiconductors 104
4.1.5.2.1. Depth Profiling by TXRF and Multilayer Structures 105
4.1.5.2.2. Vapor Phase Decomposition (VPD) and Droplet Collection . . 106
4.2. Glow Discharge Optical Emission Spectroscopy (GD-OES) . . 107
4.2.1. Principles 107
4.2.2. Instrumentation 109
4.2.3. Spectral Information 110
4.2.4. Quantification 110
4.2.5. Depth Profiling 111
4.2.6. Applications 111
4.3. Surface-Sensitive IR and Raman Spectroscopy; Ellipsometry 113
4.3.1. Reflection – Absorption IR Spectroscopy (RAIRS) 113
4.3.1.1. Principles 113
4.3.1.2. Instrumentation and Applications 114
4.3.2. Surface Raman Spectroscopy . . 115
4.3.2.1. Principles 115
4.3.2.2. Surface-Enhanced Raman Scattering (SERS) 116
4.3.2.2.1. Instrumentation 116
4.3.2.2.2. Spectral Information 117
4.3.2.2.3. Applications 118
4.3.2.3. Nonlinear Optical Spectroscopy . 118
4.3.3. UV – VIS – IR Ellipsometry (ELL) 119
4.3.3.1. Principles 119
4.3.3.2. Instrumentation 120
4.3.3.3. Applications 122
4.4. Other Photon-Detecting Techniques 125
4.4.1. Appearance-Potential Methods . 125
4.4.1.1. Soft X-Ray Appearance-Potential Spectroscopy (SXAPS) 125
4.4.1.2. Disappearance-Potential Spectroscopy (DAPS) 126
4.4.2. Inverse Photoemission Spectroscopy (IPES) and Bremsstrahlung Isochromat Spectroscopy (BIS) 126
4.4.3. Ion-Beam Spectrochemical Analysis (IBSCA) 126
5. Scanning Probe Microscopy . . 127
5.1. Atomic Force Microscopy (AFM) 127
5.1.1. Principles 127
5.1.2. Instrumentation 129
5.1.3. Applications 130
5.2. Scanning Tunneling Microscopy (STM) 132
5.2.1. Principles 132
5.2.2. Instrumentation 139
5.2.3. Lateral and Spectroscopic Information 140
5.2.4. Applications 141
6. Summary and Comparison of Techniques 142
7. Surface Analytical Equipment Suppliers 142
8. References 143

The current update was coordinated by Henning Bubert

1. Introduction

Wherever the properties of a solid surface are important, it is also important to have the means

Table 1. Surface-specific analytical techniques ∗ using particle excitation. The acronyms printed in bold belong to methods discussed in more details in this article.

Detection	Excitation ∗∗					
	Electrons e^-		Ions, neutrals A^+, A^-, A^0		Photons $h\nu$	
e^-	**AES** AEAPS EELS HREELS LEED	**SAM** CEELS RHEED	IAES INS MQS		**XPS**	**UPS**
A^+, A^-, A^0	ESD	ESDIAD	**SIMS** GDMS **LEIS**	**SNMS** FABMS **RBS**		
$h\nu$	SXAPS IPES	DAPS BIS	**GD-OES** IBSCA		**TXRF** **RAIRS** **ELL** SHG	 **SEIRA** **SERS**

∗ For meanings of acronyms, see Listing 1. ∗∗ Some of these techniques have angle-resolved variants, with the prefix AR (e.g. ARUPS), or they use Fourier transform methods, with the prefix FT (e.g. FT-RAIRS).

to measure those properties. The surfaces of solids play an overriding part in a remarkably large number of processes, phenomena, and materials of technological importance. These include catalysis; corrosion, passivation, and rust; adhesion; tribology, friction, and wear; brittle fracture of metals and ceramics; microelectronics; composites; surface treatments of polymers and plastics; protective coatings; superconductors; and solid surface reactions of all types with gases, liquids, or other solids. The surfaces in question are not always external; processes occurring at inner surfaces such as interfaces and grain boundaries are often just as critical to the behavior of the material. In all the above cases, the nature of a process or of material behavior can be understood completely only if information about both surface composition (i.e., types of atoms present and their concentrations) and surface chemistry (i.e., chemical states of the atoms) is available. Occasionally, knowledge of the arrangement of surface atoms (i.e., the surface structure) is also necessary.

First of all, what is meant by a *solid surface*? Ideally the surface should be defined as the plane at which the solid terminates, that is, the last atom layer before the adjacent phase (vacuum, vapor, liquid, or another solid) begins. Unfortunately such a definition is impractical because the effect of termination extends into the solid beyond the outermost atom layer. Indeed, the current definition is based on that knowledge, and the surface is thus regarded as consisting of that number of atom layers over which the effect of termination of the solid decays until bulk properties are reached. In practice, this decay distance is on the order of 5 – 20 nm.

By a fortunate coincidence, the depth into the solid from which information is provided by the techniques described here matches the above definition of a surface almost exactly. These techniques are therefore *surface-specific*; in other words, their information comes only from that very shallow depth of a few atom layers. There are other techniques that can be *surface-sensitive*, in that they would normally be regarded as techniques for bulk analysis but have sufficient sensitivity for certain elements that they can be analyzed even if they are present only on the surface.

Why should surfaces be so important? The answer is twofold. First, because the properties of surface atoms are usually different from those of the same atoms in the bulk; second, because in any interaction of a solid with another phase, the surface atoms are the first to be encountered. Even at the surface of a *perfect single crystal* the surface atoms behave differently from those in the bulk simply because they do not have the same number of nearest neighbors; their electronic distributions are altered and hence their reactivity. Their structural arrangement is often

also different. When the surface of a *polycrystalline or glassy multielemental solid* is considered, such as that of an alloy or a chemical compound, the situation can be very complex. The processes of preparation or fabrication may produce a material whose surface composition is quite different from that of the bulk, in terms of both constituent and impurity elements. Subsequent treatments (e.g., thermal and chemical) will almost certainly change the surface composition to something different again. The surface is highly unlikely to be smooth, and roughness at both micro- and macrolevels may be present, leading to the likelihood that many surface atoms will be situated at corners and edges and on protuberances (i.e., in positions of increased reactivity). Surfaces exposed to the atmosphere, which include many of those of technological interest, will acquire a contaminant layer 1 – 2 atom layers thick, containing principally carbon and oxygen but also other impurities present in the local environment. Atmospheric exposure may also cause oxidation. Because of all these possibilities the surface region must be considered as a separate entity, effectively a separate quasi-two-dimensional phase overlaying the normal bulk phase. Analysis of the properties of such a quasi-phase necessitates the use of techniques in which the information provided originates only or largely within the phase (i.e., the surface-specific techniques described in this article).

Nearly all these techniques involve interrogation of the surface with a particle probe. The function of the probe is to excite surface atoms into states giving rise to emission of one or more of a variety of secondary particles such as electrons, photons, positive and secondary ions, and neutrals. Since the primary particles used in the probing beam may also be either electrons or photons, or ions or neutrals, many separate techniques are possible, each based on a different primary – secondary particle combination. Most of these possibilities have now been established, but in fact not all the resultant techniques are of general application, some because of the restricted or specialized nature of the information obtained and others because of difficult experimental requirements. In this article, therefore, most space is devoted to surface analytical techniques that are widely applied and readily available commercially, whereas much briefer descriptions are given of the many others whose use is less common but which – in appropriate circumstances, particularly in basic research – can provide vital information.

Because the various types of particle can appear in both primary excitation and secondary emission, most authors and reviewers have found it convenient to group the techniques in a matrix, in which the columns refer to the nature of the exciting particle and the rows to the nature of the emitted particle [1–9]. Such a matrix of techniques is given in Table 1, with the acronyms now accepted. The meanings of the acronyms, together with some of the alternatives that have appeared in the literature, are in Listing 1.

A few techniques, including one or two important ones, cannot be classified according to the nature of the exciting particle, since they do not employ primary particles but depend instead on the application either of heat or a high electric field. These techniques are listed in Table 2.

Listing 1. Meanings of surface analysis acronyms, and their alternatives, that appear in Tables 1 and 2.

1) Electron Excitation

AEAPS	Auger electron appearance potential spectroscopy
AES	Auger electron spectroscopy
BIS	Bremsstrahlung isochromat spectroscopy
CEELS	Core-electron energy loss spectroscopy (or ILS: ionization loss spectroscopy)
DAPS	Disappearance potential spectroscopy
EELS	Electron energy loss spectroscopy
ESD	Electron-stimulated desorption (or EID: electron-induced desorption)
ESDIAD	Electron-stimulated desorption ion angular distribution
HREELS	High-resolution electron energy loss spectroscopy (or LEELS: low-energy electron loss spectroscopy)
IPES	Inverse photoemission spectroscopy
LEED	Low-energy electron diffraction
RHEED	Reflection high-energy electron diffraction

SXAPS Soft X-ray appearance potential spectroscopy (or APS appearance potential spectroscopy)
SAM Scanning Auger microscopy

2) Ion Excitation

GDMS Glow discharge mass spectrometry
GD-OES Glow discharge optical emission spectroscopy
IAES Ion (excited) Auger electron spectroscopy
IBSCA Ion beam spectrochemical analysis (or SCANIIR: surface composition by analysis of neutral and ion impact radiation, or BLE: bombardment-induced light emission)
INS Ion neutralization spectroscopy
LEIS Low energy ion scattering (or ISS: ion scattering spectroscopy)
MQS Metastable quenching spectroscopy
RBS Rutherford backscattering spectroscopy (or HEIS: high-energy ion scattering)
SIMS Secondary-ion mass spectrometry (SSIMS: static secondary-ion mass spectrometry) (DSIMS: dynamic secondary-ion mass spectrometry)
SNMS Secondary neutral mass spectrometry

3) Photon Excitation

ELL Ellipsometry
RAIRS Reflection – absorption infrared spectroscopy (or IRRAS: infrared reflection – absorption spectroscopy) (or IRAS: infrared absorption spectroscopy) (or ERIRS: external reflection infrared spectroscopy)
SEIRA Surface-enhanced IR absorption spectroscopy
SERS Surface-enhanced Raman scattering
SHG Optical second harmonic generation
TXRF Total reflection X-ray fluorescence analysis
UPS Ultraviolet photoelectron spectroscopy
XPS X-ray photoelectron spectroscopy (or ESCA: electron spectroscopy for chemical analysis)

4) Neutral Excitation

FABMS Fast-atom bombardment mass spectrometry

5) Thermal Excitation

TDS Thermal desorption spectroscopy

6) High-Field Excitation

APFIM Atom probe field-ion microscopy
IETS Inelastic electron tunneling spectroscopy
POSAP Position-sensitive atom probe
STM Scanning tunneling microscopy
STS Scanning tunneling spectroscopy

7) Mechanical Force

AFM Atomic force microscopy

Table 2. Surface-specific analytical techniques ∗ using nonparticle excitation. The acronyms printed in bold belong to methods discussed in more detail in this article.

Detection	Excitation		
	Heat kT	High electrical field F	Mechanical force
A^+	TDS	APFIM POSAP	
A^-	TDS		
e^-		IETS **STM**, **STS**	
(Displacement)			**AFM**

∗ For meanings of acronyms, see Listing 1.

2. Electron Detection

2.1. X-Ray Photoelectron Spectroscopy (XPS)

X-ray photoelectron spectroscopy is the most widely used surface analytical technique, and is therefore described here in more detail than any of the other techniques. At its inception by Siegbahn et al. [10] it was called ESCA (electron spectroscopy for chemical analysis), but the name ESCA is now considered too general, since many surface electron spectroscopies exist, and the name given to each one must be precise. Nevertheless, the name ESCA is still used in many places, particularly in industrial laboratories and

their publications. Briefly, the reasons for the popularity of XPS are the exceptional combination of compositional and chemical information that it provides, its ease of operation, and the ready availability of commercial equipment.

2.1.1. Principles

The surface to be analyzed is irradiated with soft X-ray photons. When a photon of energy $h\nu$ interacts with an electron in a level with binding energy E_{B} (E_{B} is the energy E_{K} of the K-shell in Figure 1), the entire photon energy is transferred to the electron, with the result that a photoelectron is ejected with kinetic energy

$$E_{\mathrm{kin}} = h\nu - E_{\mathrm{B}} - \Phi_S \tag{2.1}$$

where Φ_{S} is a small, almost constant, work-function term.

Clearly $h\nu$ must be greater than E_{B}. The ejected electron may come from a core level or from the occupied portion of the valence band, but in XPS most attention is focused on electrons in core levels. Since no two elements share the same set of electronic binding energies, measurement of the photoelectron kinetic energies provides an *elemental analysis*. In addition, Equation (2.1) indicates that any changes in E_{B} are reflected in E_{kin}, which means that changes in the chemical environment of an atom can be followed by monitoring changes in the photoelectron energies, leading to the provision of *chemical information*. XPS can analyze all elements in the periodic table except hydrogen and helium.

Although XPS is concerned principally with photoelectrons and their kinetic energies, ejection of electrons by other processes also occurs. An ejected photoelectron leaves behind a core hole in the atom. The sequence of events following the creation of the core hole is shown schematically in Figure 1 (right). In the example, the hole has been created in the K shell, giving rise to a photoelectron whose kinetic energy would be ($h\nu - E_{\mathrm{K}}$), and is filled by an electronic transition from the unresolved L_2 shell. The energy $E_{\mathrm{K}} - E_{\mathrm{L}_{23}}$ associated with the transition can then either be dissipated as a characteristic X-ray photon or given up to an electron in the same or a higher shell, also shown in this example as the L_{23} shell. The second of these possibilities is called the Auger process after its discoverer [11], and the resultant ejected electron is called an Auger electron and has an energy given by

$$E_{\mathrm{kin}}\,(\mathrm{KL_1L_{23}}) = E_{\mathrm{K}} - E_{\mathrm{L}_1} - E_{\mathrm{L}_{23}} - f\,(\mathrm{L_1L_{23}}) + R - \Phi_S \tag{2.2}$$

where $f(\mathrm{L_1L_{23}})$ is the interaction energy between the holes in the L_1 and L_{23} shells, and R is the sum of the intra- and extra-atomic relaxation energies. X-ray photon emission (i.e., X-ray fluorescence) and Auger electron emission are obviously competing processes, but for the shallow core levels involved in XPS and AES the Auger process is far more likely.

Thus, in all X-ray photoelectron spectra, features appear due to both photoemission and Auger emission. In XPS, the Auger features can be useful but are not central to the technique, whereas in AES (see Chap. 2.3), Equation (2.2) forms the basis of the technique.

At this point the nomenclature used in XPS and AES should be explained. In XPS the *spectroscopic notation* is used, and in AES the *X-ray notation*. The two are equivalent, the different usages having arisen for historical reasons, but the differentiation is a convenient one. They are both based on the so-called j–j coupling scheme describing the orbital motion of an electron around an atomic nucleus, in which the total angular momentum of an electron is found by summing vectorially the individual electron spin and angular momenta. Thus if l is the electronic angular momentum quantum number and s the electronic spin momentum quantum number, the total angular momentum for each electron is given by $j = l + s$. Since l can take the values 0, 1, 2, 3, 4, ... and $s = \pm\frac{1}{2}$, clearly $j = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, etc. The principal quantum number n can take values 1, 2, 3, 4, ... In *spectroscopic notation*, states with $l = 0$, 1, 2, 3, ... are designated s, p, d, f, ..., respectively, and the letter is preceded by the number n; the j values are then appended as suffixes. Therefore, one obtains 1s, 2s, $2\mathrm{p}_{1/2}$, $2\mathrm{p}_{3/2}$, 3s, $3\mathrm{p}_{1/2}$, $3\mathrm{p}_{3/2}$, etc.

In *X-ray notation*, states with $n = 1, 2, 3, 4, \ldots$ are designated K, L, M, N, ..., respectively, and states with various combinations of $l = 0$, 1, 2, 3, ... and $j = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$, are appended as the suffixes 1, 2, 3, 4, ... In this way one arrives at K,

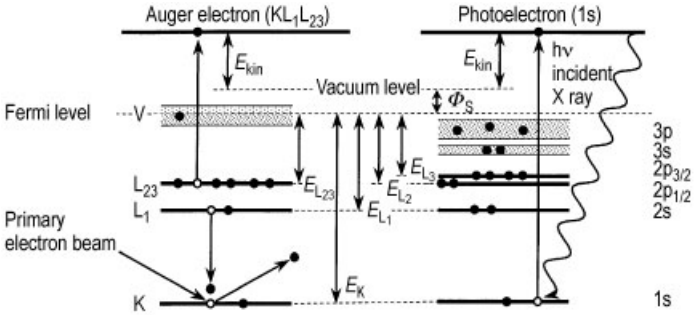

Figure 1. Schematic diagram of electron emission processes in solids. Left: Auger process; Right: photoelectron process. Electrons involved in the emission processes are indicated by open circles.

L_1, L_2, L_3, M_1, M_2, M_3, etc. The equivalence of the two notations is set out in Table 3.

Table 3. Spectroscopic and X-ray notations

Quantum numbers			Spectroscopic	X-ray	X-ray
n	l	j	state	suffix	state
1	0	1/2	1s	1	K
2	0	1/2	2s	1	L_1
2	1	1/2	$2p_{1/2}$	2	L_2
2	1	3/2	$2p_{3/2}$	3	L_3
3	0	1/2	3s	1	M_1
3	1	1/2	$3p_{1/2}$	2	M_2
3	1	3/2	$3p_{3/2}$	3	M_3
3	2	3/2	$3d_{3/2}$	4	M_4
3	2	5/2	$3d_{5/2}$	5	M_5
	etc.		etc.	etc.	etc.

In X-ray notation the Auger transition shown in Figure 1 would therefore be labeled KL_2L_3. In this coupling scheme, six Auger transitions would be possible in the KLL series. Obviously, many other series are possible (e.g., KLM, LMM, MNN). These are discussed more fully in Section 2.3.

The reasons why techniques such as XPS and AES, which involve measurement of the energies of ejected electrons, are so surface-specific, should be examined. An electron with kinetic energy E moving through a solid matrix M has a probability of traveling a certain distance before losing all or part of its energy as a result of an inelastic collision. Based on that probability, the average distance traveled before such a collision is known as the inelastic mean free path (IMFP) $\lambda_M(E)$. The IMFP is a function only of M and of E. Figure 2 shows a compilation of measurements of λ made by Seah and Dench [12], in terms of atomic monolayers as a function of kinetic energy. Note that both λ and energy scales are logarithmic. The important consequence of the dependence of λ on kinetic energy is that in the ranges of secondary electron kinetic energies used in XPS and AES, the values of λ are very small. In XPS, for example, typical ranges are 250 – 1500 eV, corresponding to a range of λ from about four to eight monolayers, while in AES, the energy range is typically 20 to 1000 eV, in which case λ would range from about two to six monolayers. What this means in practice is that if the photoelectron or the Auger electron is to escape into a vacuum and be detected, it must originate at or very near the surface of the solid. This is the reason why the electron spectroscopic techniques are surface specific.

2.1.2. Instrumentation

2.1.2.1. Vacuum Requirements

Electron spectroscopic techniques require vacua on the order of 10^{-8} Pa for their operation. This requirement arises from the extreme surface specificity of these techniques. With sampling depths of only a few atom layers, and elemental sensitivities down to 10^{-5} atom layers (i.e., one atom of a particular element in 10^5 other atoms in an atomic layer), the techniques are clearly very sensitive to surface contamination, most of which comes from the residual gases in the vacuum system. According to gas kinetic theory, for sufficient time to be available to make a surface-analytical measurement on a surface that has just been prepared or exposed, before contamination from the gas phase interferes, the base pressure

should be 10^{-8} Pa or lower, that is, in the region of ultrahigh vacuum (UHV).

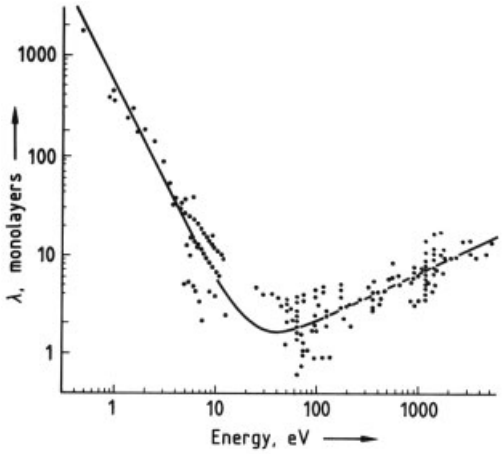

Figure 2. Compilation by SEAH and DENCH [12] of measurements of inelastic mean free path as a function of electron kinetic energy
The solid line is a least-squares fit

The requirement for UHV conditions imposes restrictions on the types of material that can be used for the construction of surface analytical systems, or inside the systems, because UHV can be achieved only by accelerating the rate of removal of gas molecules from internal surfaces by raising the temperature of the entire system (i.e., by baking). Typical baking conditions are 150 – 200 °C for several hours. Inside the system, any material is permissible that does not produce volatile components either during normal operation or during baking. For example, brass, which contains the volatile metal zinc, can not be used. The principal construction material is stainless steel, with mu-metal (76 % Ni, 5 % Cu, 2 % Cr) used occasionally where magnetic screening is needed (e.g., around electron energy analyzers). For the same reasons, metal seals, not elastomers, are used for the demountable joints between individual components – the sealing material usually being pure copper although gold is sometimes used. Other materials that may be used between ambient atmosphere and UHV are borosilicate glass or quartz for windows, and alumina for electrical insulation for current or voltage connections.

Table 4. Energies and linewidths of some characteristic low-energy X-ray lines

Line	Energy, eV	Width, eV
$Y_{M\zeta}$	132.3	0.47
$Zr_{M\zeta}$	151.4	0.77
$Nb_{M\zeta}$	171.4	1.21
$Mg_{K\alpha}$	1253.6	0.70
$Al_{K\alpha}$	1486.6	0.85
$Si_{K\alpha}$	1739.5	1.00
$Y_{L\alpha}$	1922.6	1.50
$Zr_{L\alpha}$	2042.4	1.70

2.1.2.2. X-Ray Sources

The most important consideration in choosing an X-ray source for XPS is energy resolution. Equation (2.1) gives the relationship between the kinetic energy of the photoelectron, the energy of the X-ray photon, and the binding energy of the core electron. Since the energy spread, or linewidth, of an electron in a core level is very small, the linewidth of the photoelectron energy depends on the linewidth of the source if no undue broadening is introduced instrumentally. In XPS the analyst devotes much effort to extracting chemical information by means of detailed study of individual elemental photoelectron spectra. Such a study needs an energy resolution better than 1.0 eV if subtle chemical effects are to be identified. Thus the linewidth of the X-ray source should be significantly smaller than 1.0 eV if the resolution required is not to be limited by the source itself.

Other considerations are that the source material, which forms a target for high-energy electron bombardment leading to the production of X-rays, should be a good conductor to allow rapid removal of heat, and it should also be compatible with UHV.

Table 4 lists the energies and linewidths of the characteristic X-ray lines from a few possible candidate materials. In practice $Mg_{K\alpha}$ and $Al_{K\alpha}$ are the two used universally due to their line energies and width and their simple use as anode material.

For efficient production of X-rays by electron bombardment, exciting electron energies that are at least an order of magnitude higher than the line energies must be used, so that in Mg and Al sources accelerating potentials of 15 kV are employed. Modern sources are designed with dual anodes, one anode face being coated with

magnesium and the other with aluminum, and with two filaments, one for each face. Thus, a switch from one type of X irradiation to the other can be made very quickly.

To protect the sample from stray electrons from the anode, from heating effects, and from possible contamination by the source enclosure, a thin (ca. 2 µm) window of aluminum foil is interposed between the anode and the sample. For optimum X-ray photon flux on the surface (i.e., optimum sensitivity), the anode must be brought as close to the sample as possible, which means in practice a distance of ≈ 2 cm. The entire X-ray source is therefore retractable by means of a bellows and a screw mechanism.

The X radiation from magnesium and aluminum sources is quite complex. The principal $K\alpha$ lines are in fact unresolved doublets and should correctly be labeled $K\alpha_{1,2}$. Besides the $K\alpha_{1,2}$ lines, a series of further lines, so-called satellite lines, also exist, of which the more important ones are $K\alpha_{3,4}$. The energy separations of the satellite lines for Mg and Al and their intensities relative to $K\alpha_{1,2}$ are listed in Table 5.

Removal of satellites, elimination of the bremsstrahlung background, and separation of the $K\alpha_{1,2}$ doublet can be achieved by monochromatization, shown schematically in Figure 3. The X-ray source is positioned at one point on a spherical surface, called a Rowland sphere, and a quartz crystal is placed at another point. X-rays from the source are diffracted from the quartz, and by placing the sample at the correct point on the Rowland sphere, the $K\alpha_1$ component can be selectively focused on it. Quartz is a very convenient diffracting medium for $Al_{K\alpha}$ since the spacing between the $10\bar{1}0$ planes is exactly half the wavelength of the X radiation. Since the width of the $Al_{K\alpha_1}$ line is < 0.4 eV, the energy dispersion needed around the surface of the sphere implies that the Rowland sphere should have a diameter of at least 0.5 m. Although an XPS spectrum will be much "cleaner" when a monochromator is used, because satellites and background have been removed, the photon flux at the sample is much lower than that from an unmonochromatized source operating at the same power. Against this must be set the greatly improved signal-to-background level in a monochromatized spectrum.

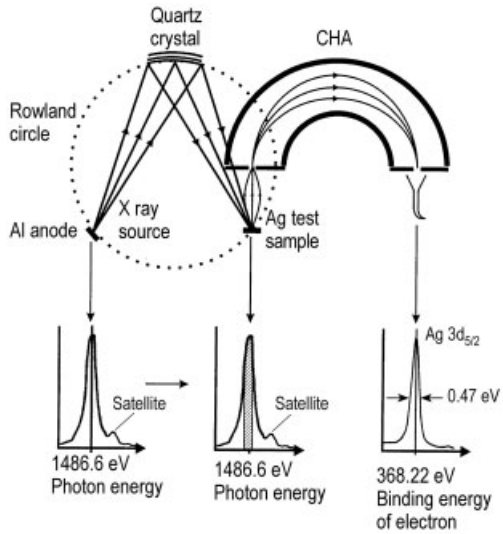

Figure 3. Schematic of X-ray monochromatization to remove satellites, eliminate bremsstrahlung background, and separate the Al $K\alpha_{1,2}$ doublet
Courtesy of Kratos Analytical

2.1.2.3. Synchrotron Radiation

The discrete line sources described above for XPS are perfectly adequate for most applications, but some types of analysis require that the source be tunable (i.e., that the exciting energy be variable). The reason is to allow the photoionization cross section of the core levels of a particular element or group of elements to be varied, which is particularly useful when dealing with multielement semiconductors. Tunable radiation can be obtained from a synchrotron.

In a synchrotron, electrons are accelerated to near-relativistic velocities and constrained magnetically into circular paths. When a charged particle is accelerated, it emits radiation, and when the near-relativistic electrons are forced into curved paths, they emit photons over a continuous spectrum. The general shape of the spectrum is shown in Figure 4. For a synchrotron with an energy of several gigaelectronvolts and a radius of some tens of meters, the energy of the emitted photons near the maximum is on the order of 1 keV (i.e., ideal for XPS). As can be seen from the universal curve, plenty of usable intensity exists down into the UV region. With suitable monochromators on the output to select a particular wavelength, the photon en-

Table 5. Satellite lines of magnesium and aluminum [13]

X-ray line	Mg		Al	
	Separation from $K\alpha_{1,2}$, eV	Relative intensity, % ($K\alpha_{1,2} = 100\,\%$)	Separation from $K\alpha_{1,2}$, eV	Relative intensity, % ($K\alpha_{1,2} = 100\,\%$)
$K\alpha'$	4.5	1.0	5.6	1.0
$K\alpha_3$	8.4	9.2	9.6	7.8
$K\alpha_4$	10.0	5.1	11.5	3.3
$K\alpha_5$	17.3	0.8	19.8	0.4
$K\alpha_6$	20.5	0.5	23.4	0.3
$K\beta$	48.0	2.0	70.0	2.0

ergy can be tuned continuously from about 20 to 500 eV. The available intensities are comparable to those from conventional line sources. The main advantage is the small beam diameter of the synchrotron radiation, which allows high spatial resolution.

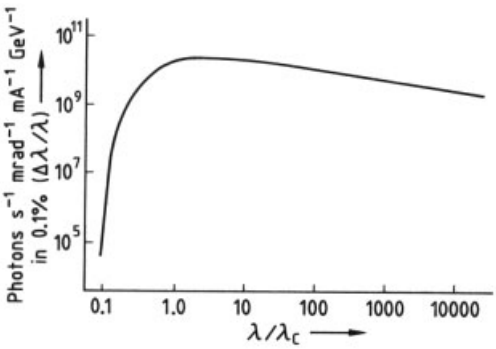

Figure 4. Normalized spectrum of photon energies emitted from a synchrotron
λ_c = wavelength characteristic of the individual synchrotron

2.1.2.4. Electron Energy Analyzers

In electron spectroscopic techniques – among which XPS is the most important – analysis of the energies of electrons ejected from a surface is central. Nowadays the concentric hemispherical analyzer (CHA) is universally employed.

Energy analyzers cannot be discussed without a discussion of energy resolution, which is defined in two ways. *Absolute resolution* is defined as ΔE, the full width at half-maximum (FWHM) of a chosen peak. *Relative resolution* is defined as the ratio R of ΔE to the kinetic energy E of the peak energy position (usually its centroid), that is, $R = \Delta E/E$. Thus absolute resolution is independent of peak position, but relative resolution can be specified only by reference to a particular kinetic energy. In XPS, closely spaced peaks at any point in the energy range must be resolved, and this requires the same absolute resolution at all energies.

The CHA is shown in schematic cross section in Figure 5 [14]. Two hemispheres of radii r_1 (inner) and r_2 (outer) are positioned concentrically. Potentials $-V_1$ and $-V_2$ are applied to the inner and outer hemispheres, respectively, where V_2 is greater than V_1. The source S and the focus F are in the same plane as the center of curvature, and r_0 is the radius of the equipotential surface between the hemispheres. If electrons of energy $E = eV_0$ are injected at S along the equipotential surface, they will be focused at F if

$$V_2 - V_1 = V_0\,(r_2/r_1 - r_1/r_2) \tag{2.3}$$

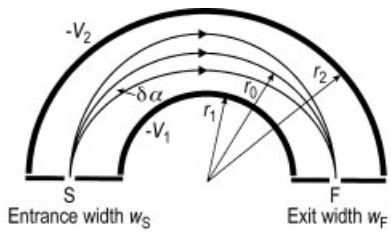

Figure 5. Schematic of a concentric hemispherical analyzer [14]

If electrons are injected not exactly along the equipotential surface, but with an angular spread $\delta\alpha$ about the correct direction, then the energy resolution is given by

$$\Delta E/E = (w_S + w_F)\,/4r_0 + (\delta\alpha)^2 \tag{2.4}$$

where w_S and w_F are the respective widths of the entrance and exit slits. In most instruments, for convenience in construction, $w_S = w_F = w$, whereupon the resolution becomes

$$\Delta E/E = w/2r_0 + (\delta\alpha)^2 \quad (2.5)$$

In XPS the photoelectrons are retarded to a constant energy, called the pass energy, as they approach the entrance slit. If this were not done, Equation (5) shows that to achieve an absolute resolution of 1 eV at the maximum kinetic energy of about 1500 eV (with $Al_{K\alpha}$ radiation), and with a slit width of 2 mm, would require an analyzer with an average radius of about 300 cm, which is impracticable. Pass energies are selected in the range 20 – 100 eV for XPS, which allows the analyzer to be built with a radius of 10 – 15 cm.

Modern XPS spectrometers employ a lens system on the input to the CHA, which transfers an image of the analyzed area on the sample surface to the entrance slit of the analyzer. The detector system on the output of the CHA consists of several single channeltrons or a channel plate. Such a spectrometer is illustrated schematically in Figure 6.

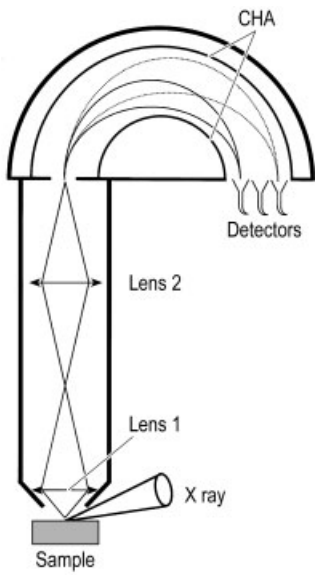

Figure 6. Typical configuration in an XPS spectrometer

2.1.2.5. Spatial Resolution

A principal disadvantage of conventional XPS was the lack of spatial resolution; the spectral information came from an analyzed area of several square millimeters and was therefore an average of the compositional and chemical analyses from that area. However, many technological samples are inhomogeneous on a scale much smaller than that of conventional XPS analysis, and obtaining chemical information on the same scale as the inhomogeneities would be very desirable.

Improved XPS spatial resolution was achieved by using the focusing properties of the CHA described above, and this led to the development of the ESCASCOPE, now developed into the present ESCALAB 250 (VG Scientific, East Grinstead, UK), which acts as an XPS microscope. Here, a Fourier-transformed image is generated at the entrance slit of the CHA by introducing a third lens into the transferring lens system. The CHA disperses this image energetically, so that a Fourier-transformed image exists at the exit slit, but only for the energy selected by the setting of the CHA. This image is then inverted into a real image, which can be detected in the plane of a channel plate and a spatial resolution of about 3 µm is obtained.

A further development using the focusing properties of the transfer lens system is realized in the AXIS instruments (Kratos Analytical, Manchester, UK). Here, the principle plane of the first lens is moved towards the specimen by introducing a magnetic lens below the specimen holder. The spatial resolution is improved to about 15 µm without loss of intensity due to the simultaneous enhancement of the angle of acceptance.

A third development is realized in the Quantum 2000 (Physical Electronics, Eden Prairie, USA). Here, high-energy electrons of an electron gun are scanned over the surface of an Al anode, thereby generating an X-ray beam. This beam is then focused onto the specimen surface by an ellipsoidal monochromator and scans the specimen surface in the same rhythm as the electron beam is scanned. The X-ray beam is less than 10 µm in diameter, and the scanned area is 1.4×1.4 mm.

The thrust of development is toward ever-better spatial resolution, which means the smallest possible spot size on the sample compatible with adequate signal-to-noise ratio, for acquisition within a reasonable length of time.

2.1.3. Spectral Information and Chemical Shifts

Figure 7 shows a *wide-scan* or *survey XPS spectrum*, that is, one recorded over a wide range of energies, in this case 1000 eV. The radiation used was unmonochromatized $Al_{K\alpha}$, at 1486.6 eV, and the surface is that of almost clean copper. Such a spectrum reveals the major, or primary, features to be found, but to investigate minor or more detailed features, spectra are acquired over much more restricted energy ranges, at better energy resolution; the latter ones are called *narrow-scan spectra*.

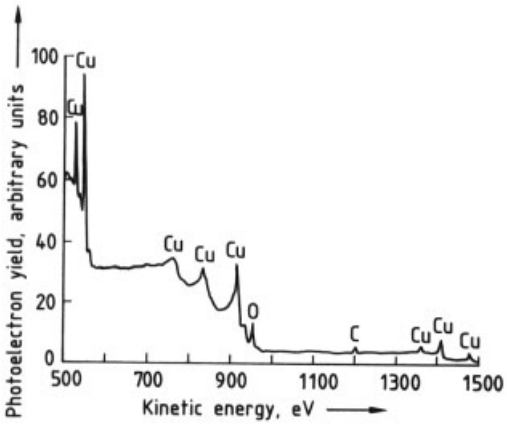

Figure 7. Wide-scan spectrum from almost clean copper, recorded with $Al_{K\alpha}$ radiation

The primary features in Figure 7 are peaks arising from excitation of core level electrons according to Equation (2.1). At the low-energy end, two intense peaks are found at about 553 and 533 eV, corresponding to photoelectrons from the $2p_{3/2}$ and $2p_{1/2}$ levels of copper, respectively, the separation of 20 eV being the spin-orbit splitting. At the high kinetic energy end, there are three other copper photoelectron peaks at about 1363, 1409, and 1485 eV, which arise from the 3s, 3p, and 3d levels, respectively. The other primary features associated with copper are the three Auger peaks $L_3M_{4,5}M_{4,5}$, $L_3M_{2,3}M_{4,5}$, and $L_3M_{2,3}M_{2,3}$at 919, 838, and 768 eV, respectively. As pointed out in Section 2.1.1, the creation of a core hole by any means of excitation can lead to ejection of an Auger electron, so that in XPS, Auger features make a significant contribution to the spectrum. If Auger and photoelectron peaks happen to overlap in any spectrum, they can always be separated by changing the excitation (e.g., from $Al_{K\alpha}$ to $Mg_{K\alpha}$, or vice versa) since the Auger peaks are invariant in energy (Eq. 2.2) whereas the photoelectron peaks must shift with the energy of the exciting photons according to Equation (2.1).

In addition to primary features due to copper in Figure 7, there are small photoelectron peaks at kinetic energies of 955 and 1204 eV that arise from the oxygen and carbon 1s levels, respectively, due to the presence of some contamination on the surface. Secondary features are X-ray satellite and ghost lines, surface and bulk plasmon energy loss features, shake-up lines, multiplet splitting, shake-off lines and asymmetries due to asymmetric core levels [15].

Chemical shift is the observed shift in energy of a photoelectron peak from a particular element when the chemical state of that element changes. When an atom enters into combination with another atom or group of atoms, an alteration occurs in the valence electron density, which can be positive or negative according to whether charge is accepted or donated, and this causes a consequent alteration in the electrostatic potential affecting the core electrons. Therefore the binding energies of the core electrons change, giving rise, according to Equation (2.1), to shifts in the corresponding photoelectron peaks. Tabulation of the chemical shifts experienced by any one element in a series of pure compounds of that element thus enables its chemical state to be identified during analysis of unknown samples. Many such tabulations have appeared, and a major collection of them can be found in [2]. The identification of chemical state in this way is the principal advantage of XPS over other surface analytical techniques.

An example of a spectrum exhibiting a chemical shift is that of the tin 3d peaks in Figure 8. A thin layer of oxide on the metallic tin surface allows photoelectrons from both the underlying metal and the oxide to appear together. Resolution of the doublet $3d_{5/2}$, $3d_{3/2}$ into the components from the metal (Sn^0) and from the oxide Sn^{n+} is shown in Figure 8 B. The shift in this case is 1.6 – 1.7 eV. Curve resolution is an operation that can be performed routinely in data-processing systems associated with photoelectron spectrometers.

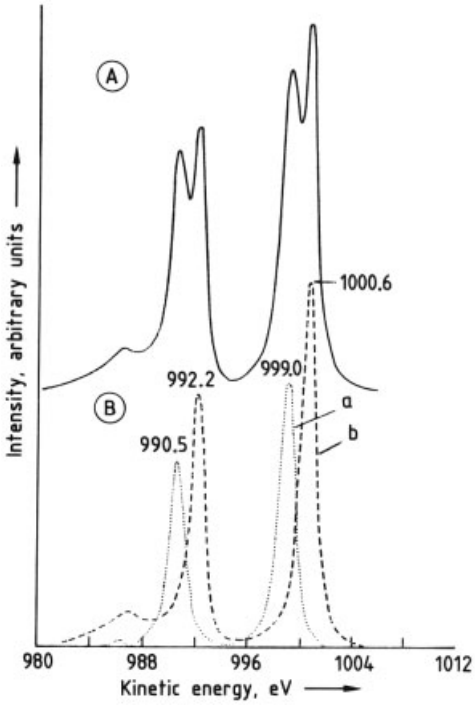

Figure 8. Example of a chemical shift in the Sn 3d peak, exhibited by a very thin layer of Sn oxide on Sn metal
A) Spectrum after linear background subtraction; B) Spectrum resolved into its respective components
a) Sn^{n+}; b) Sn^0

2.1.4. Quantification, Depth Profiling, and Imaging

2.1.4.1. Quantification

If an X-ray photon of energy $h\nu$ ionizes a core level X in an atom of element A in a solid, the photoelectron current I_A from X in A is

$$I_A(X) = K\,\sigma_A(h\nu,X)\,\beta_A(h\nu,X)\,\bar{N}_A\,\lambda_M(E_A)\cos\theta \qquad (2.6)$$

where $\sigma_A(h\nu, X)$ is the photoelectric cross section for ionization of X (binding energy E_B) by photons of energy $h\nu$; $\beta_A(h\nu, X)$ the asymmetry parameter for emission from X by excitation with photons of energy $h\nu$; $\bar{N}_A$ the atomic density of A averaged over depth of analysis; $\lambda_M(E_A)$ the inelastic mean free path in matrix M containing A, at kinetic energy E_A, where $E_A = h\nu - E_B$; and θ the angle of emission to surface normal. K is a constant of proportionality that contains fixed operational parameters such as incident X-ray flux, transmission of the analyzer at kinetic energy E_A, and efficiency of the detector at kinetic energy E_A, kept fixed during any one analysis.

Values of the total cross section σ_A for $Al_{K\alpha}$ radiation, relative to the carbon 1s level, have been calculated by Scofield [16], and of the asymmetry parameter β_A by Reilman et al. [17]. Seah and Dench [12] (Fig. 2) have compiled many measurements of the inelastic mean free path, and for elements the best-fit relationship that they found was

$$\lambda_M(E) = 538E^{-2} + 0.41\,(aE)^{1/2}\ \text{monolayers} \qquad (2.7)$$

where a is the monolayer thickness in nanometers.

In principle, therefore, the surface concentration of an element can be calculated from the intensity of a particular photoelectron emission, according to Equation (2.6). In practice, the *method of relative sensitivity factors* is in common use. If spectra were recorded from reference samples of pure elements A and B on the same spectrometer and the corresponding line intensities are I_A^∞ and I_B^∞, respectively, Equation (2.6) can be written as

$$\frac{I_A/I_A^\infty}{I_B/I_B^\infty} = \left(\frac{\lambda_M(E_A)\,\lambda_B(E_B)}{\lambda_M(E_B)\,\lambda_A(E_A)}\right)\left(\frac{R_B^\infty}{R_A^\infty}\right)\left(\frac{N_A N_B^\infty}{N_B N_A^\infty}\right) \qquad (2.8)$$

where R_A^∞ and R_B^∞ are factors for the reference samples, depending on the roughness of the surfaces, and N_A^∞ and N_B^∞ are the atomic densities of the reference samples. From the above equation the ratio N_A/N_B can be obtained from the measured intensity ratio I_A/I_B. However the accuracy of quantitative results is only as good as the accuracy of the ingoing parameter and the intensity values I_A and I_B. These intensities are simply the areas under the line peaks after background subtraction. Three different procedures are commonly in use: straight line subtraction, Shirley's method, and Tougaard's method. With the last-mentioned method, the measured spectrum is deconvoluted into the primary (true) excited spectrum by taking into account that the emitted electrons suffer an energy loss due to scattering during travelling in the solid.

A further critical point are the intensities I^{∞} correlated to spectra of the pure elements. Calculated and experimentally determined values can diverge considerably, and the best data sets for I^{∞} measured on pure reference samples still show a scatter of up to 10 %. The use of an internal standard or a simultaneously measured external standard seems to be the most successful way to decrease the inaccuracy to below 10 %. (For a more detailed discussion of background subtraction and quantification, see [18].)

2.1.4.2. Depth Profiling

Very often, in addition to analytical information about the original surface, information is required about the distribution of chemical composition to depths considerably greater than the inelastic mean free path. Nondestructive methods such as variation of emission angle or variation of exciting photon energy are possible, but they are limited in practice to evaluation of chemical profile within depths of only ca. 5 nm. Obtaining information from greater depths requires destructive methods, of which the one used universally is removal of the surface by ion bombardment, also called sputtering. The combination of removal of the surface by sputtering and of analysis by XPS, AES, or SSIMS is termed depth profiling. Instrumentally, this is performed by means of ion guns which are described elsewhere (see Section 3.1.1.1.1). Depth profiling in XPS is mostly performed with a noble gas to minimize chemical effects. In compounds, preferential sputtering, e.g., of O from TiO_2 surfaces and of Be from Cu – Be alloys, may occur. This is easily comprehensible in terms of the different surface binding energies and different momenta transferred to atoms with different masses. Preferential sputtering frequently results in a chemically modified surface, even if the primary ion from the ion gun is not chemically reactive after its (immediate) neutralization in the solid.

The depth resolution achievable during profiling depends on many variables, and the reader is referred to the comprehensive discussion in [18].

In XPS, chemical information is comparatively slowly acquired in a stepwise fashion along with the depth, by using alternate cycles of sputtering and analysis. Examples of profiles through oxide films on pure iron and on Fe – 12Cr – 1Mo alloy are shown in Figure 9, in which the respective contributions from the metallic and oxide components of the iron and chromium spectra have been quantified [19]. In these examples the oxide films were only ≈5 nm thick on iron and ≈3 nm thick on the alloy.

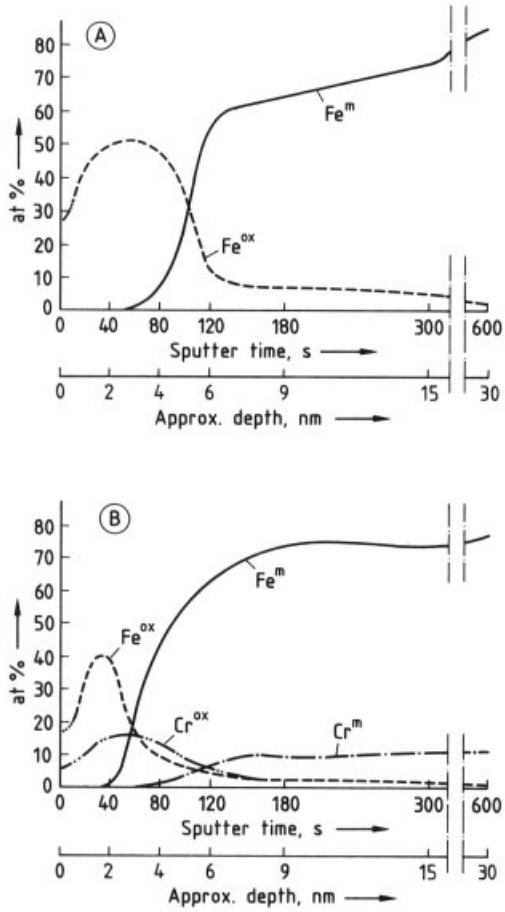

Figure 9. Depth profiles of thin (3 – 5 nm) oxide films [19] on
A) Pure iron; B) Fe – 12 Cr – 1 Mo alloy

Current practice in depth profiling is to use positively charged argon ions at energies between 0.5 and 10 keV, focused into a beam of 2 – 5 µm diameter, which is then rastered over the area to be profiled. The area bombarded by the primary ion beam, which is typically finely focused and rastered, should be significantly greater than the analyzed area, so that only the flat bottom of the sputtered crater is analyzed. Ion current densities are variable between 5 and

50 μA/cm^2. Uniformity of sputtering, and therefore the depth resolution of the profile, can be improved by rotating the sample during bombardment.

During the sputtering process, many artifacts are introduced that can affect quantification; some of these derive from the nature of the sample itself (e.g., roughness, crystalline structure, phase distribution, electrical conductivity), and others are radiation-induced (e.g., atomic mixing, preferential sputtering, compound reduction, diffusion and segregation, sputter-induced topography). This is an extensive subject in its own right; a discussion of all aspects is given in [20].

State of the art for data evaluation of complex depth profile is the use of factor analysis. The acquired data can be compiled in a two-dimensional data matrix in a manner that the n intensity values $N(E)$ or, in the derivative mode $\mathrm{d}N(E)/\mathrm{d}E$, respectively, of a spectrum recorded in the ith of a total of m sputter cycles are written in the ith column of the data matrix $\boldsymbol{D}$. For the purpose of factor analysis, it now becomes necessary that the $(n \times m)$-dimensional data matrix $\boldsymbol{D}$ can be expressed as a product of two matrices, that is, the $(n \times k)$-dimensional *spectrum matrix* $\boldsymbol{R}$ and the $(k \times m)$-dimensional *concentration matrix* $\boldsymbol{C}$, in which $\boldsymbol{R}$ contains the spectra of k components in k columns, and $\boldsymbol{C}$ contains the concentrations of the respective m sputter cycles in k rows, i.e.

$$\boldsymbol{D} = \boldsymbol{R} \cdot \boldsymbol{C} + \boldsymbol{E} \tag{2.9}$$

$\boldsymbol{E}$ is an error matrix taking errors of measurement (e.g., random noise) into consideration. The term *component* describes such chemical or physical states whose spectra cannot be generated by a linear combination of the other components. Thus, components can be elements, chemical compounds – stoichiometric or nonstoichiometric – or even states induced by physical processes, provided that the spectra differ significantly, for example, in line shapes or line shifts.

Since the introduction of factor analysis for evaluation of depth profile data by Gaarenstroom [21], many papers have been published [22–29]. With the help of factor analysis, three results can directly be achieved: (1) the number k of relevant components which are necessary to create the data matrix, (2) the so-called abstract spectrum matrix $\boldsymbol{R}*$ and the abstract concentration matrix $\boldsymbol{C}*$, and (3) a new data matrix $\boldsymbol{D}*$, which can be generated from $\boldsymbol{R}*$ and $\boldsymbol{C}*$ by leaving out the so-called noise components.

If the spectra of the k relevant components are known, the abstract matrices can be transformed into interpretable spectrum and concentration matrices, respectively, and as a final result, one obtains the required *component depth profile*. If one or more of the spectra of the k relevant components is not known, a quantified component depth profile cannot be constructed.

New developments which have still to be checked for their usability in data evaluation of depth profiles are *artificial neural networks* [25], [30–34], *fuzzy clustering* [35], [36], and *genetic algorithms* [37].

2.1.4.3. Imaging

As already mentioned in Section 2.1.2.5 the ESCALAB acts as an XPS microscope and the result is called an *elemental map* or *image*. An example of *chemical state imaging* is given in Figure 10 [38]. The material is a contaminated fluoropolymer, the contamination being in the form of dark spots of 10 – 80 μm in size. Both images in Figure 10 are in the C 1s photoelectron peak, but of different chemical states, that is, different binding energies. The upper is in that contribution arising from carbon bound to fluorine, and the lower from carbon bound to carbon (i.e., graphitic in nature). The complementarity of the images indicates that the contamination is graphitic in nature.

If the analyzer is set to accept electrons of an energy characteristic of a particular element, and the incident X-ray beam is rastered over the surface to be analyzed, then a visual display whose intensity is modulated by the peak intensity will correspond to the distribution of that element over the surface. The result is also an *image*, and this technique is realized with the Quantum 2000.

When the whole sample surface is irradiated by the exciting X-rays, an image can be obtained in a different way: The spot accepted by the transferring lens system in front of the input of the CHA is rastered by introducing deflector plates in front of the lens system. Again,

only electrons of a characteristic energy can pass the analyzer. This technique is realized with the AXIS series.

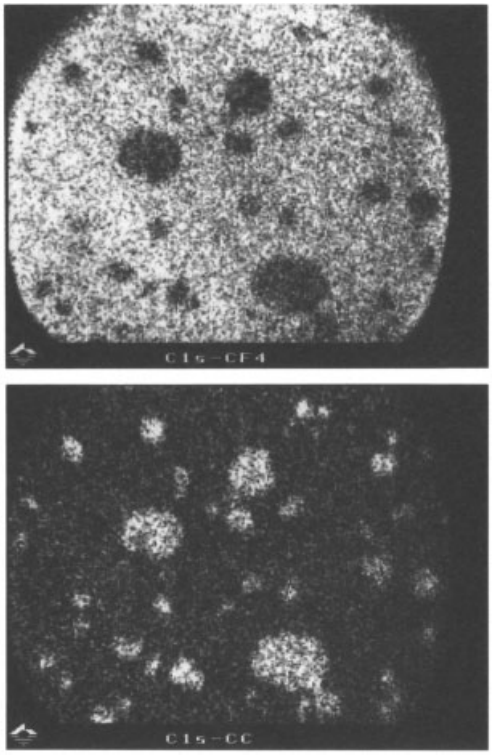

Figure 10. Chemical state images obtained with an ESCASCOPE, from a contaminated fluoropolymer [198]
A) Image in contribution to C 1s from C–F bonding; B) Image in contribution to C 1s from C–C bonding

Tonner et al. have recorded scanning XPS microscographs at the Advanced Light Source synchrotron radiation center of Lawrence Berkeley National Laboratory [15]. They investigated a polished and sputter-cleaned surface of mineral ilmenite with the nominal composition of $FeTiO_3$, and used the Fe 3p and Ti 3p lines for imaging. Using synchrotron radiation they demonstrated a spatial resolution of about 0.25 µm.

Care must be taken in interpreting the intensity distribution, since the electron intensity depends not only on the local concentration of the element but also on the topography, because surface roughness can affect the inelastic background underneath the line. Therefore, elemental maps are customarily presented as variations of the ratio of peak intensity divided by the magnitude of the background on one or both sides of the line, which can be carried out easily by computer.

2.1.5. The Auger Parameter

Section 2.1.3 shows that in an XPS spectrum, X-ray excited Auger peaks are often as prominent as the photoelectron peaks themselves. For many elements, the chemical shifts in Auger peaks are actually greater than the shifts in photoelectron peaks. The two shifts can be combined in a very useful quantity called the Auger Parameter $\alpha*$, first used by Wagner [39] and defined in its modified form [40] as

$$\alpha* = E_{\mathrm{kin}}\,(XYZ) + E_{\mathrm{B}}\,(W) \tag{2.10}$$

where $E_{\mathrm{kin}}\,(XYZ)$ is the kinetic energy of the Auger transition XYZ, and $E_{\mathrm{B}}\,(W)$ is the binding energy of an electron in level i. The Auger and photoelectron peak energies must be measured in the same spectrum.

The parameter $\alpha*$ is independent of photon energy and has the great advantage of being independent of any surface charging. It is therefore measurable for all types of material from metals to insulators. Extensive tabulations of $\alpha*$ have appeared, based on a large number of standard materials, and the most reliable have been collected by Wagner [41].

As defined in Equation (2.10), $\alpha*$ is purely empirical; any prominent and conveniently situated Auger and photoelectron peaks can be used. If $\alpha*$ is defined slightly more rigorously, by replacing $E_{\mathrm{B}}\,(W)$ by $E_{\mathrm{B}}\,(X)$ in Equation (2.10), it is related closely to a quantity called the extra-atomic relaxation energy R^{ea}, since $\Delta\alpha*$ can be shown to equal $2\,\Delta R^{\mathrm{ea}}\,(X)$. In other words, the change in Auger Parameter between two chemical states is equal to twice the change in the extraatomic relaxation energies in the two states, associated with the hole in the core level X. Thus $\alpha*$ has an important physical basis as well as being useful analytically.

2.1.6. Applications

XPS has been applied in virtually every area in which the properties of surfaces are important. The most prominent areas can be drawn

from conferences on surface analysis, especially from the ECASIA, which is held every two years. These areas are: adhesion, biomaterials, catalysis, ceramics and glasses, corrosion, environmental problems, magnetic materials, metals, micro- and optoelectronics, nanomaterials, polymers and composite materials, superconductors, thin films and coatings, and tribology and wear. Furthermore, the proceedings of these conferences give a representative survey on current surface-analytical problems and studies [42], [43]. Given below are a few selected examples from the areas mentioned above; for more comprehensive discussions of the applications of XPS, see [1], [3–9], [44–49].

2.1.6.1. Catalysis

An understanding of the basic mechanisms of catalysis has been one of the major aims of XPS from its inception. To this end the technique has been used in two ways: (1) in the study of "real" catalysts closely resembling those used industrially; (2) in the study of "model" catalysts, where the number of parameters has been reduced by the use of well-characterized crystal surfaces.

For industrial catalysts the chemical information obtainable from XPS is all important, since the course and efficiency of the catalyzed reaction will be governed by the states of oxidation of the elements at the surface. Hence, XPS is used to analyze both elemental and chemical composition as a function of bulk composition, of surface loading where the active material is deposited on a support, of oxidation and reduction treatments, and of time in reactor under operating conditions. Figure 11 demonstrates the effect on the chemical state of molybdenum in the surface of Mo – Sm – O catalysts of various compositions [50]. For the highest concentration of molybdenum (spectrum d) the Mo 3d doublet is very similar to that of MoO_3, but at the samarium-rich composition (spectrum a), it is more like that of a molybdate or a mixed oxide. The effect of successive reduction treatments on a cobalt – molybdenum – alumina catalyst is shown in the changes in the Mo 3d spectra of Figure 12 [51]. The air-fired catalyst in spectrum a contains entirely Mo^{VI} at the surface, but with repeated hydrogen reduction the spectrum changes radically and can be curve-resolved into Mo^{VI}, Mo^{V}, and Mo^{IV} contributions. The chemical state of the surface in spectrum f contains approximately equal proportions of these three oxidation states.

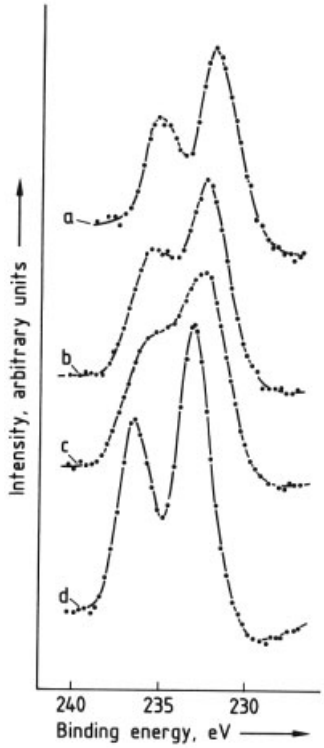

Figure 11. The Mo 3d XPS spectra of model catalysts in the Mo – Sm – O system [39]
Mo : Sm ratios are a) 0.25; b) 1.33; c) 4.00; d) 8.00

PIJPERS et al. [52] reviewed the application of XPS in the field of catalysis and polymers. Other applications of XPS to catalytic problems deal with the selective catalytic reduction of NO_x on Pt- and Co-loaded zeolites. Although the Al 2p line (Al from zeolite) and Pt 4f line interfere strongly, the two oxidation states Pt^0 and Pt^{II} can be distinguished after careful curve fitting [53]. The influence of Zn deposition on Cu(111) surfaces in methanol synthesis by hydrogenation of CO_2 shows that Zn creates sites that stabilize the formate intermediate and thus promote the hydrogenation process [54]. Further publications deal with methane oxidation by various layered rock-salt-type oxides [55]; poisoning of vanadia in VO_x/TiO_2 by K_2O leading to a lower reduction capability of vanadia due to the formation of V^V [56], and interaction of SO_2 with Cu, Cu_2O,

and CuO showing temperature-dependent SO_2 absorption or sulfide formation [57].

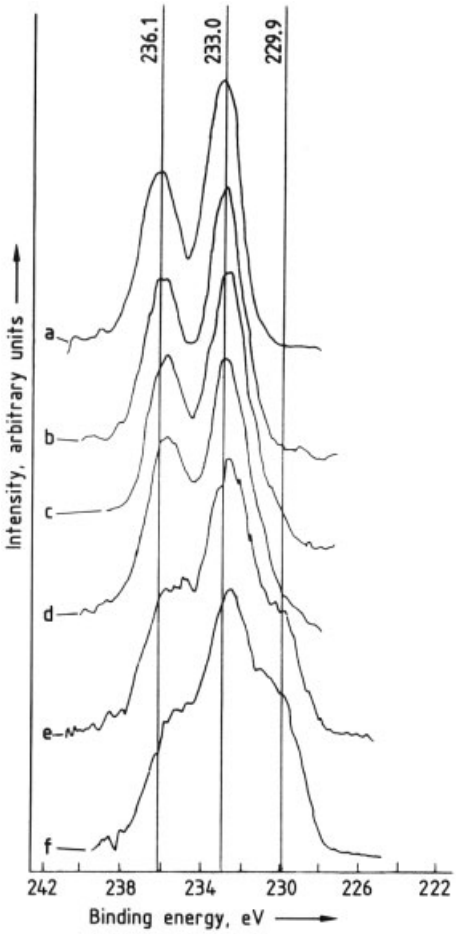

Figure 12. Changes in the Mo 3d XPS spectrum from a cobalt – molybdenum – alumina catalyst during successive reduction treatments in hydrogen at 500 °C [51]
a) Air-fired catalyst; b) Reduction time 15 min; c) 50 min; d) 60 min; e) 120 min; f) 200 min

2.1.6.2. Polymers

Polymers have been the most prominent objects of investigation since XPS became commercially available in the late 1960s. Nowadays XPS is still one of the most widely applied methods for characterizing polymer surfaces. "Bibles" of high resolution XP spectra of more than a hundred different organic polymers have been published by Beamson and Briggs [58] and Briggs [59]. A crucial problem is the X-ray degradation that takes place in a more or less pronounced manner on virtually all polymer surfaces. Therefore, one has to exercise caution in investigating polymer surface reactions and respective quantification; in practice this means reducing the radiation time of the exciting X-rays.

Since polymers are typically nonconductive, sample charging can occur and has to be compensated carefully, for example, by a low-energy electron flood gun, to avoid lineshape distortion and misinterpretation of the measurements.

The photoelectron line of main interest is C 1s. Different bonding environments of the carbon atom lead to very small chemical shifts of this line. Therefore, high-resolution XPS is required, and monochromatic radiation should be applied to avoid overlaps with satellite lines.

Many publications appear every year dealing with the analysis of polymers, and a few examples are given here. Butyl rubbers are mainly used in tire production, and to study the influence of different polymers on the surface chemistry and to determine the adhesion of bromobutyl blends as a function of temperature, valence-band XPS was applied to a variety of pure polymers and different butyl – elastomer blends. These investigations show the value of valence-band XPS for polymers, which are not easy to analyze by core-level XPS [60]. In a study dealing with the enhancement of the adhesion strength of the interface Cu – polyimide, polyimide surfaces were subjected to a plasma treatment, and the modified surfaces were studied by TOF-SIMS/XPS [61]. More than 80 review articles appeared since 1995 dealing with special problems in polymer analysis, for example, polymer membranes [62], polymeric biomedical and optoelectronic applications [63], and the current status of XPS instrumentation in polymer analysis [64].

2.1.6.3. Corrosion and Passivation

Besides the application of XPS in catalysis, the study of corrosion mechanisms and corrosion products is another main area of application. Special attention has to be devoted to artefacts as a consequence of X-ray irradiation. For example, reduction of metal oxides (e.g., $CuO \rightarrow Cu_2O$) can take place, loosely bound

water or hydrates may be desorbed in the spectrometer vacuum, and decomposition of hydroxides may occur. Thorough investigations are supported by other surface-analytical and/or microscopic techniques like AFM, which is gaining increasing importance.

Corrosion products that are formed as thin layers on metal surfaces in aqueous or gaseous environments, along with the nature and stability of passive and protective films on metals and alloys, have also been major areas of XPS application. XPS has been used in two ways, one in which materials corroded or passivated in the natural environment are analyzed, and another in which well-characterized, usually pure metal surfaces are studied after exposure to controlled conditions.

More effort has probably been devoted to study of the corrosion and passivation properties of Fe – Cr – Ni alloys such as stainless steel and other transition-metal alloys than to any other metallic system [52], [54, 65, 66]. The type of spectral information obtainable from an Fe – Cr alloy of technical origin, carrying an oxide and contaminant film after corrosion, is shown schematically in Figure 13 [67].

Corrosion-inhibiting organic films on steel were studied by XPS [68], and investigations on inhibiting films formed in sea water [69] showed their stability, that is, no pitting corrosion was observed. Protective layers, frequently TiN, deposited on different materials are of major interest. Investigations on TiN coatings exposed to different humid SO_2-containing aggressive environments show that the most aggressive atmosphere can penetrate through pores towards the interface TiN – Fe and cause corrosion [70]. The corrosion behavior of sandwich-structured TiN – Ti layers was also investigated [71].

Biocorrosion of stainless steel occurs due to exopolymer-producing bacteria. It was shown that Fe is accumulated in the biofilm [72]. The influence of bacteria on the corrosion behavior of Mo metal surface was also investigated by XPS [73]. These last two mentioned investigations indicate a new field of research in which XPS can be employed successfully. XPS has also been applied to studying the corrosion of glasses [74], of polymer coatings on steel [75], of tooth-filling materials [76], and to investigate the role of surface hydroxyl groups of oxide films on metals [77] or other passive films.

2.1.6.4. Adhesion

The adhesion of metal and ink to polymer, as well as the adhesion of paint and coating to metal, are of vital importance in several technologies. In the aircraft industry, aluminum-to-aluminum adhesion is employed. The strength and durability of an adhesive bond depend absolutely on the way in which the adhesive compound interacts with the surfaces to which it is supposed to adhere, which in turn often involves pretreatment of the surfaces to render them more reactive. The nature and degree of the reactivity are functions of the chemical states of the adhering surfaces; these states can be monitored by XPS.

In electronic packaging technology, metal – polymer adhesion has been important for some years. Many studies have been devoted to the interactions between metals and polymers, and that of Chin-An Chang et al. [78] is a good example. They deposited 200-nm-thick films of copper, chromium, titanium, gold, and aluminum on various fluorocarbon polymer films, and found that adhesion in terms of peel strength was highest for titanium. Figure 14 shows the C 1 s spectra after metal deposition; they reveal that for titanium, and to a lesser extent chromium, a strong carbide bond is formed, as evidenced by the peak at 282 – 283 eV. With the other metals, very low adhesion was found, and no carbide-like peaks occurred.

The path of failure of an adhesive joint can give information about the mechanism of failure if analysis of the elemental and chemical composition can be carried out along the path. Several authors have performed such analyses by loading the adhesive joint until it fractures and then using XPS to analyze each side of the fracture.

The complementary use of XPS and TOF-SIMS was demonstraded by Fitzpatrick et al. and Watts et al. [79], [80] who studied the failure of hot-dipped galvanized steel pretreated with a zinc phosphate conversion coating and joined with epoxy. They found that the observed electrochemical activity is due to degradation at the zinc – polymer interface and is responsible for the ingress of aggressive substances. Furthermore Wolany et al. [81] investigated the adhesion of untreated and plasma-modified poly-

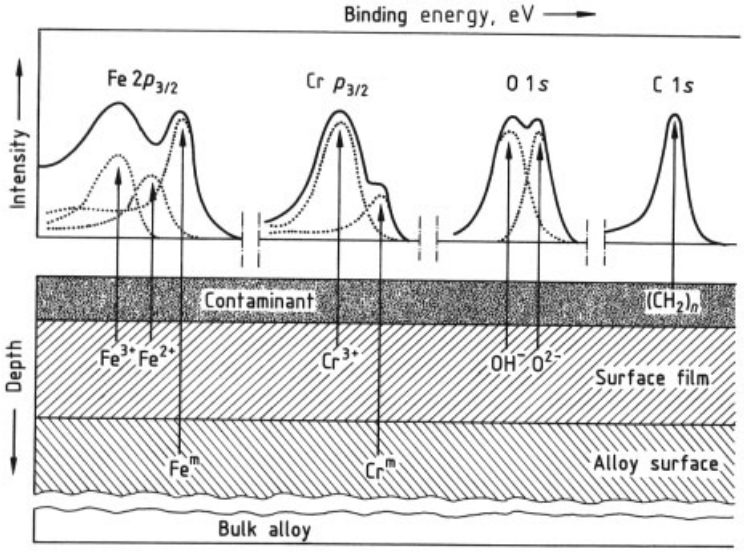

Figure 13. Schematic of type of information obtainable from XPS spectra from an Fe – Cr alloy with an oxide film underneath a contaminant film [67]

imide by XPS and static SIMS. They showed that the number and kind of functional groups available for adhering bonds were enhanced by plasma treatment and that the adhesion strength for Cu deposition was improved. This was also corroborated by SIMS.

Recently, XPS was used to study the effects of thermal treatment on the morphology and adhesion of the interface between Au and the polymer trimethylcyclohexane-polycarbonate [82]; the composition of the surfaces and interfaces of plasma-modified Cu – PTFE and Au – PTFE, as well as the surface structure and the improvement of adhesion [83]; the influence of excimer laser irradiation of the polymer on the adhesion of metallic overlayers [84]; and the behavior of the Co-rich binder phase of WC – Co hard metal and the deposition of diamond thereon [85].

2.1.6.5. Superconductors

In the years since the discovery of superconducting oxides in 1986, many papers have been published describing the application of XPS to superconductors. One reason for this frenzied activity has been the search for the precise mechanism of superconductivity in these materials.

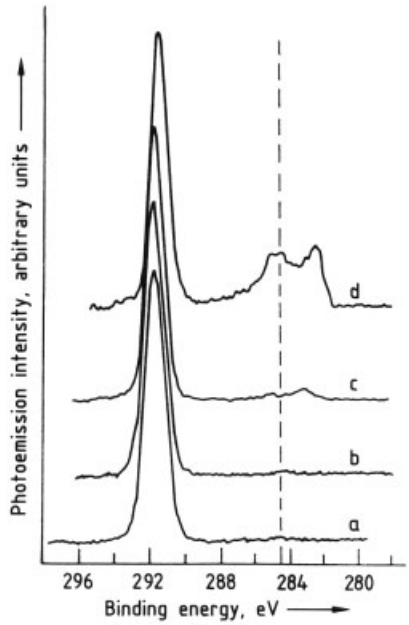

Figure 14. The C 1s XPS spectra recorded by Chin-An Chang et al. [78] from perfluoroalkoxy polymer (PFA)
a) Before deposition; b) Deposition of copper; c) Deposition of chromium; d) Deposition of titanium

An essential piece of information concerns the chemical states of the constituent elements, particularly copper, which is common to all of them. Some theories predict that small but

significant amounts of the Cu^{3+} state should occur in superconductors, but other models disagree. Comparison by STEINER et al. [86] of the Cu $2p_{3/2}$ spectra from copper compounds containing copper as Cu^0 (metal), Cu^+, Cu^{2+}, and Cu^{3+}, with the spectra from the superconducting oxide $YBa_2Cu_3O_7$ and from the base material La_2CuO_4, showed conclusively that the Cu^{3+} contribution to the superconducting mechanism must be negligible. Some of their comparative spectra are shown in Figure 15. The positions of the Cu $2p_{3/2}$ peak in $YBa_2Cu_3O_7$ and La_2CuO_4, along with the satellite structure, are very similar to that in CuO (i.e., Cu^{2+}, rather than in $NaCuO_2$, which contains Cu^{3+}).

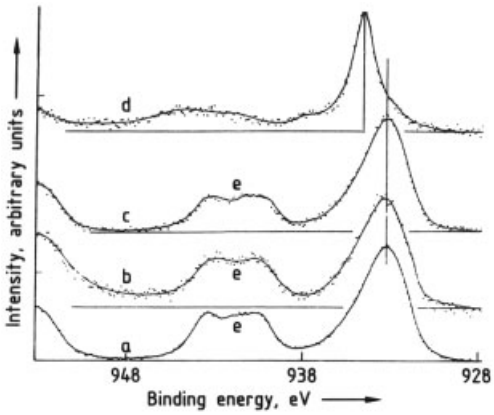

Figure 15. Comparison of the Cu $2p_{3/2}$ and satellite XPS spectra from several copper compounds with the spectrum from the superconducting oxide $YBa_2Cu_3O_7$ [86]
a) CuO, $\Gamma = 3.25$ eV; b) La_2CuO_4, $\Gamma = 3.30$ eV; c) $YBa_2Cu_3O_7$, $\Gamma = 3.20$ eV; d) $NaCuO_2$, $\Gamma = 1.60$ eV; e) Cu^{2+} satellites

Much effort in superconductor research has been devoted to fabricating thin films with the same superconducting properties as the bulk material, and to prepare suitable electrical contact to the surfaces of superconductors. XPS has been of great use in both areas through its ability to monitor both composition and chemical state during the processes either of thin superconducting film formation, or of the deposition of thin conducting and semiconducting films onto a superconducting surface. Typical of the latter studies is that of ZIEGLER et al. [87], who deposited increasing amounts of silicon onto epitaxial films of $YBa_2Cu_3O_{7-x}$. Figure 16 shows the Si 2p and Ba 4d spectra during the deposition of silicon and after heating the films in oxygen at 250 °C once 6 nm of silicon had accumulated [86]. The Ba 4d spectrum at the bottom can be resolved into three $4d_{5/2,3/2}$ doublets, corresponding – with increasing energy – to barium in the bulk superconductor, in the surface of the superconductor, and as carbonate. With increasing silicon thickness the barium spectrum attenuates but does not change in character, but silicon does not appear in the spectrum until nearly 5 nm has been deposited, suggesting migration into the superconductor at room temperature. At a thickness of 6 nm, the Si 2p spectrum can be separated into suboxide contributions, indicating reaction of silicon with the $YBa_2Cu_3O_{7-x}$ surface. After heating in oxygen the Si 2p spectrum changes to that characteristic of SiO_2 and silicates, while the Ba 4d spectrum also changes by increasing in intensity and conforming mostly to that expected of a barium silicate. As a result of the latter changes the superconducting properties of the film were destroyed. The Y 3d and Cu 2p spectra establish that yttrium and copper oxides are formed as well.

Other surface reactions studied have been those of gold [88] and silver [89] with $Bi_2Sr_2CaCu_2O_{8+y}$, of niobium with $YBa_2Cu_3O_{7-x}$ [90], and of lead with $YBa_2Cu_3O_{7-x}$ [91], [92]. Characteristics of superconductors in the form of thin films prepared in several ways have been recorded by XPS for $YBa_2Cu_3O_{7-x}$ [93–97], $GdBa_2Cu_3O_{7-x}$ [98], $Nd_{2-x}Ce_xCuO_{4-y}$ [99], $Rb_xBa_{1-x}BiO_3$ [100], $(Bi_{0.9}Pb_{0.1})_{2.3}Sr_2Ca_{1.9}Cu_3O_{10+x}$ [101], and $Tl_2Ca_2Ba_2Cu_3O_{10}$ [102]. Detailed electronic studies have been made on the bulk superconductors $Bi_2Sr_2Ca_{1-x}Nd_xCu_2O_y$ [103], [104], $Nd_{2-x}Ce_xCuO_{4-y}$ [105–109], $ErBa_2Cu_4O_8$ [110], and $Tl_2Ba_2CaCu_2O_8$ [111].

In a number of applications depth profiles were recorded with the aim of investigating the extent of reaction of various materials with superconductors, or analyzing the uniformity and concentration during preparation of superconductors. Such investigations were carried out with XPS as well as AES.

2.1.6.6. Interfaces

The chemical and electronic properties of elements at the interfaces between very thin films

and bulk substrates are important in several technological areas, particularly microelectronics, sensors, catalysis, metal protection, and solar cells. To study conditions at an interface, depth profiling by ion bombardment is inadvisable since both composition and chemical state can be altered by interaction with energetic positive ions. The normal procedure is therefore to start with a clean or other well-characterized substrate and deposit the thin film onto it slowly at a chosen temperature while XPS monitors the composition and chemical state by recording selected characteristic spectra. The procedure continues until no further spectral changes occur, as a function either of film thickness, of time elapsed since deposition, or of changes in substrate temperature. A good example is the study of the reaction at the interface of the rare-earth element thulium with the silicon (111) surface [112]. Figure 17 shows the Si 2p spectrum from the substrate as thulium is deposited on it at room temperature. With the $Mg_{K\alpha}$ radiation used for the analysis, the kinetic energy of the Si 2p electrons is about 1155 eV, corresponding to an inelastic mean free path of about 2 nm. The Si 2p peak eV is still plainly detectable after 12.5 nm of thulium has been deposited, and in addition a new feature appears on the low binding energy side of the peak, increasing in magnitude and shifting to lower energy with thickness. Raising the temperature in stages for the thickest film and use of X-ray diffraction established that the energetic final position labeled R_1 corresponded to Tm_5Si_3, and the initial energetic position found at a thickness of 0.5 – 1.0 nm, to $TmSi_2$. However, the magnitude of the residual Si 2p peak was too great at all stages to be accounted for by compound formation alone, indicating the formation of thulium clusters followed by their coalescence into islands. Measurement of the Tm $4d_{5/2}$ binding energy during the interaction also showed that the oxidation state of thulium was +3 at all temperatures and coverages.

Studies of interfaces by XPS have been carried out on U – Si(111)-7×7 surface [113], on interfaces forming during the oxidation of $Ti_{1-x}Al_xN$ ($x = 0.2 - 0.65$) [114], on $ZrO_2 - Y_2O_3$ on steel [115], on naphthalocyanine on Si(111)-7×7 and Si(100)-2×1 faces [116], and on a phenylacetylene – *p*-nitrophenylacetylene copolymer on Cr [117]. These five examples show that in the different areas of application discussed so far, interfaces play a dominant role, because they determine the chemical and electronic properties of the effectiveness of adhesives, the adhesion of protective or decorative coatings, the functionality of microelectronics and sensors, and the behavior of catalysts.

Other important areas of application of XPS have included lubricating films [118], [119], glasses [120–124], ion implantation [125–129], and cleaning [130–132] and passivation [133], [134] of semiconductor surfaces.

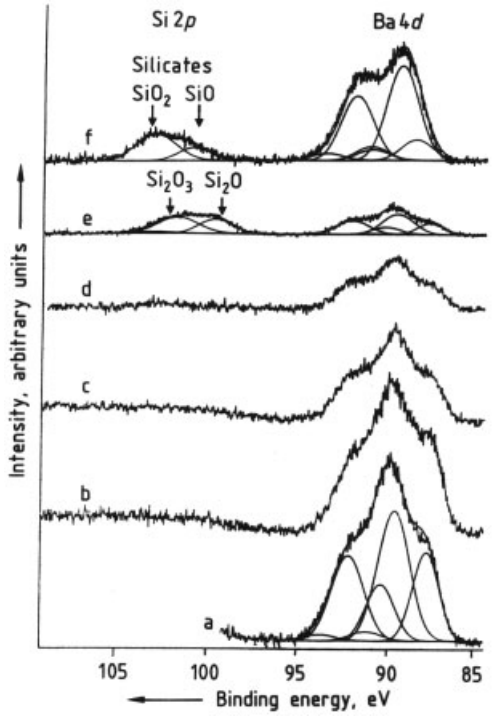

Figure 16. The Si 2p and Ba 4d XPS spectra [87] from the a) Clean surface of the superconducting oxide $YBa_2Cu_3O_{7-x}$; b) – e) Same surface after increasing deposition of b) 0.8 nm Si; c) 2.8 nm Si; d) 4.4 nm Si; e) 6.0 nm Si; f) Deposited surface after heating to 250 °C in oxygen (8×10^{-3} mbar)

2.2. Ultraviolet Photoelectron Spectroscopy (UPS)

In its principle of operation, UPS is similar to XPS, in that a surface is irradiated with photons and the energies of the ejected photoelectrons are analyzed. The physical relationship involved is the same as that of Equation (2.1) for

XPS, but in UPS the energy $h\nu$ of the exciting photons is much lower since the photons are derived from a gaseous discharge that produces hard UV radiation. The gas normally used is helium, which, depending on pressure conditions in the discharge, will provide line sources of energy 21.21 eV (He I) or 40.82 eV (He II) with very narrow linewidths (≈20 meV). Because of these low exciting energies, the binding energies E_B that appear in Equation (2.1) refer not to core levels, as in XPS, but to shallow valence band levels and to other shallow levels, such as those of adsorbates, near the valence band. For energy analysis, the CHA is again used, but the pass energies employed for UPS are much lower than in XPS since better energy resolution is required.

UPS is not an analytical technique in the sense that it can provide quantified elemental concentrations on a surface as XPS and AES can, but it is a powerful tool in basic research and is another of the group of most frequently used techniques. It is used in two ways. In its *angle-integrated* form it has been employed very extensively in studying the interactions of gaseous molecules with surfaces, usually in combination with other techniques such as XPS, AES, LEED, and TDS. The other way in which it has been used is in *angle-resolved UPS* (ARUPS). In this form the energy analyzer, a CHA of small radius, is mounted in such a way mechanically that it can traverse the space around the sample in both polar and azimuthal directions, while the angle of photon incidence is kept constant. The angle of incidence is then changed by rotation of the sample. If the sample is a single crystal with a clean surface, the dispersion of spectral features (i.e., their movement along the energy axis and their increase or decrease in magnitude) can be used theoretically to plot the band structure in the surface region. The changes in surface-associated band structure during reaction with gaseous molecules can then be followed, leading to additional information about the nature of the reaction.

2.3. Auger Electron Spectroscopy (AES)

After XPS, AES is the next most widely used surface analytical technique. As an accepted surface technique, AES actually predates XPS by two to three years, since the potential of XPS as a surface-specific technique was not recognized immediately by the surface science community. Pioneering work was carried out by Harris [135] and by Weber and Peria [136], but the technique as it is known today is basically the same as that established by Palmberg et al. [137].

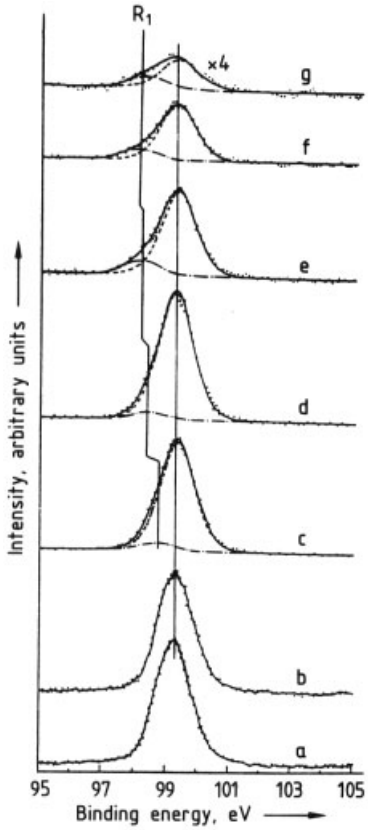

Figure 17. The Si 2*p* XPS spectra with Mg K_α excitation during deposition of thulium to a thickness of 12.5 nm at room temperature [112]
a) Clean silicon; b) 0.2 nm Tm; c) 0.5 nm Tm; d) 2 nm Tm; e) 4 nm Tm; f) 6.5 nm Tm; g) 12.5 nm Tm

2.3.1. Principles

The surface to be analyzed is irradiated with a beam of electrons of sufficient energy, typically in the range of 2 – 10 keV, to ionize one or more core levels in surface atoms. After ionization the atom can then relax by either of the two processes described in Section 2.1 for XPS, that is, ejection of a characteristic X-ray photon (fluorescence) or ejection of an Auger electron. Al-

though these are competing processes, for shallow core levels ($E_B < 2$ keV) the probability of the Auger process is far higher. The Auger process is described schematically in Figure 1 (left), which points out that the final state of the atom is doubly ionized.

Equation (2.2) shows that the Auger energy is a function only of atomic energy levels. Since no two elements have the same set of atomic binding energies, analysis of Auger energies provides elemental identification. Even if levels L_1 and L_{23} are in the valence band of the solid, analysis is still possible because the dominant term in Equation (2.2) is always the binding energy of K, the initially ionized level.

Since heavy elements have more levels than just K and L, Equation (2.2) also indicates that the heavier the element, the more numerous are the possible Auger transitions. Fortunately, large differences exist in the probabilities between different Auger transitions, so that even for the heaviest elements, only a few intense transitions occur, and analysis is still possible.

In principle, chemical as well as elemental information should be available in AES, since the binding energies appearing in Equation (2.2) are subject to the same chemical shifts as measured in XPS. In practice, since the binding energies of three levels are involved, extracting chemical information from Auger spectra is, in most cases, difficult. However, significant progress has been made in data-processing methods (see Section 2.1.4.2) yielding valuable chemical information.

The reasons why AES is one of the surface-specific techniques are given in Section 2.1.1, with reference to Figure 2. The normal range of kinetic energies recorded in an AES spectrum would typically be from 20 to 1000 eV, corresponding to inelastic mean free path values of 2 – 6 monolayers.

The nomenclature used in AES is also mentioned in Section 2.1.1. The Auger transition in which initial ionization occurs in level X, followed by the filling of X by an electron from Y and ejection of an electron from Z, would therefore be labeled XYZ. In this rather restricted scheme, one would thus find in the KLL series the six possible transitions KL_1L_1, KL_1L_2, KL_1L_3, KL_2L_2, KL_2L_3, and KL_3L_3. Other combinations could be written for other series such as the LMM, MNN, etc.

2.3.2. Instrumentation

2.3.2.1. Vacuum Requirements

The same considerations discussed in Section 2.1.2.1 for XPS apply to AES.

2.3.2.2. Electron Sources

The energy of the exciting electrons does not enter into Equation (2.2) for the Auger energy, unlike that of the exciting X-ray photon in XPS; hence the energy spread in the electron beam is irrelevant. What is relevant is the actual energy of the primary electrons because of the dependence on that energy of the cross section for ionization by electron impact. For an electron in a core level of binding energy E_B, this dependence increases steeply with primary energy above E_B, passes through a maximum about 3 – 5 times E_B, and then decreases to a fairly constant plateau. It is illustrated in Figure 18, in which various theoretical and experimental cross sections are shown for ionization of the Ni K shell, from Casnati et al. [138]. The implication of this dependence is that for efficient ionization of a particular core level, the primary energy should be about five times the binding energy of an electron in that level. Since in most samples for analysis, several elements are found in the surface region with core-level binding energies extending over a large range, choice of a primary energy that is sufficiently high to ionize all the core levels efficiently is advisable. Hence for *conventional AES*, primary energies are typically in the range 3 – 10 keV. In *scanning Auger microscopy* (SAM), much higher primary energies are used, in the range of 25 – 50 keV, because in this version of AES as small an electron spot size must be achieved on the sample as possible, and the required focusing can be effected only at such energies.

For the primary energy range 3 – 10 keV used in conventional AES, the electron emitter is thermionic, usually an LaB_6 or a tungsten cathode, and focusing of the electron beam is carried out electrostatically. Typically, such an electron source would be able to provide a spot size on the specimen of about 0.5 µm at 10 keV and a beam

current of about 10^{-8} A. The beam can normally be rastered over the specimen surface, but such a source would not be regarded as adequate for SAM. Sources for SAM may be of either the thermionic or the field-emission type, the latter being particularly advantageous in achieving minimum spot sizes since it is a high-brightness source. Focusing is performed electromagnetically, and at optimum performance a modern electron gun for SAM would be capable of a spot size of 15 nm at 30 keV and a beam current of about 10^{-10} A. The beam is rastered over the surfaces at scan rates variable up to TV rates, with the dimensions of the scanned area variable as well.

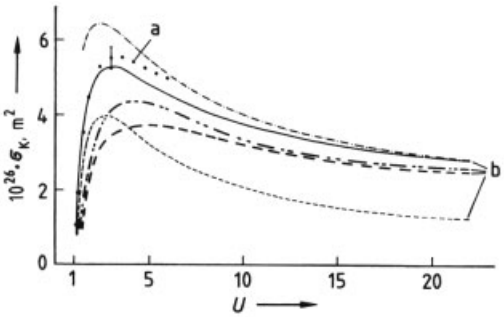

Figure 18. Electron impact ionization cross section for the Ni K shell, as a function of reduced electron energy U [138] $U = E_P/E_K$, E_P being the primary electron energy and E_K the binding energy of the K shell
a) Experimental points; b) Semiempirical or theoretical curves

2.3.2.3. Electron Energy Analyzers

An electron energy analyzer still often applied in AES is the cylindrical mirror analyzer (CMA). In the CMA shown schematically in Figure 19, two cylinders of radii r_1 (inner) and r_2 (outer) are accurately positioned coaxially. Annular entrance and exit apertures are cut in the inner cylinder, and a deflection potential $-V$ is applied between the cylinders. Electrons leaving the sample surface at the source S with a particular energy E on passing into the CMA via the entrance aperture are then deflected back through the exit aperture to the focus F. For the special case in which the acceptance angle $\alpha = 42.3°$, the first-order aberrations vanish, and the CMA becomes a second-order focusing device. The relationship between the electron energy E and the deflection potential is then

$$E/\mathrm{eV} = 1.31/\ln(r_2/r_1) \tag{2.11}$$

As in the CHA, therefore, scanning the deflection potential $-V$ and recording the signal as a function of electron energy provide the distribution in energy of electrons leaving the sample surface.

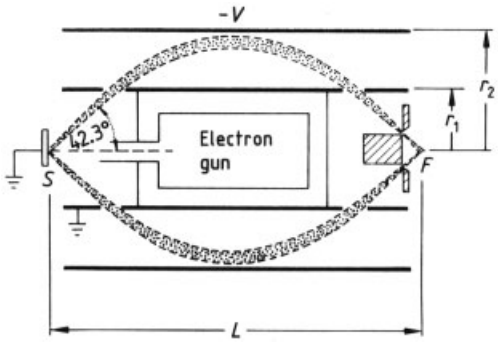

Figure 19. Schematic of a cylindrical mirror analyser (CMA)

If the angular spread of the acceptance angle is $\delta\alpha$, then the relative energy resolution of the CMA is

$$\Delta E/E = 0.18w/r_1 + 1.39\,(\delta\alpha)^3 \tag{2.12}$$

where w is the effective slit width. For a typical angular spread $\delta\alpha$ of ca. 6°, w can be replaced by the source size (i.e., the area on the sample from which electrons are accepted). The source size is very small, because of the focused primary beam, and the important property of the CMA is therefore its very high transmission, arising from the large solid angle of acceptance; for a $\delta\alpha$ value of ca. 6°, the transmission is about 14 %.

Conventionally, Auger spectra are presented as the differentiated energy distribution $\mathrm{d}N(E)/\mathrm{d}E$, rather than in the undifferentiated form $N(E)$ used in XPS. By differentiating, the visibility of Auger features is enhanced, as demonstrated in Figure 20, in which the differentiated spectrum from boron is shown above the corresponding $N(E)$ spectrum. The $N(E)$ distribution is nowadays recorded digitally, and the differentiation is carried out by computer. However, with the rapid improvement in the signal-to-noise characteristics of detection equipment, the trend in presenting Auger spectra is increasingly toward the undifferentiated $N(E)$ distribution.

Note from Figure 20 that although the true position of the boron KLL Auger peak in the $N(E)$ spectrum is at 167 eV, the position in the $dN(E)/dE$ spectrum is taken for purely reasons of convention to be that of the negative minimum, i.e., at 175 eV.

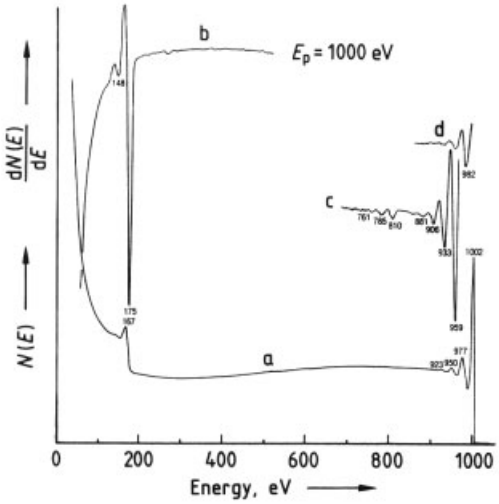

Figure 20. Secondary electron distribution from boron a) In the $N(E)$ mode; b) – d) In the $dN(E)/dE$ mode, whereby the differentiation reveals the plasmon satellites

2.3.3. Spectral Information

As stated above, the range of primary energies used typically in conventional AES is 3 – 10 keV, and the ionization cross section passes through a maximum at 3 – 5 times the binding energy of an electron in the ionized core level. Therefore the core levels that can be ionized efficiently are limited. K-shell ionization efficiency is adequate up to $Z \approx 14$ (silicon), L_3 up to $Z \approx 38$ (strontium), M_5 up to $Z \approx 76$ (osmium), and so on (for the much higher primary energies used in SAM, the ranges are correspondingly extended). This means that in various regions of the periodic table, characteristic Auger transitions occur that are most prominent under typical operating conditions.

Figures 21, 22, 23 show some of these prominent Auger features. They have all been taken from [139], and recorded using a CMA, in the differential distribution. Examples of the KLL Auger series are shown in Figure 21, of the LMM series in Figure 22, and of the MNN series in Figure 23. In the KLL series the most prominent feature is a single peak arising from the $KL_{2,3}L_{2,3}$ transition, with minor features from other Auger transitions to lower energy. The LMM series is characterized by a triplet arising from $L_3M_{2,3}M_{2,3}$, $L_3M_{2,3}M_{4,5}$, and $L_3M_{4,5}M_{4,5}$ transitions in ascending order of kinetic energy, with another intense peak, the $M_{2,3}M_{4,5}M_{4,5}$, appearing at very low energy (e.g., 50 eV for iron). Note that AES transitions involving electrons in valence band levels are often written with a V rather than the full symbol of the level. Thus $L_3M_{2,3}M_{4,5}$ and $L_3M_{4,5}M_{4,5}$ would often appear as $L_3M_{2,3}V$ and L_3VV, respectively, and similarly $M_{2,3}M_{4,5}M_{4,5}$ as $M_{2,3}VV$. In Figure 22 the increase in intensity of the L_3VV peak relative to the other two upon going from chromium to iron is due to the progressive increase in the electron density in the valence band. The characteristic doublet seen in the MNN series arises from the $M_{4,5}N_{4,5}N_{4,5}$ transitions, in which the doublet separation is that of the core levels M_4 and M_5. The characteristic doublet seen in the MNN series arises from the $M_{4,5}N_{4,5}N_{4,5}$ transitions, in which the doublet separation is that of the core levels M_4 and M_5.

Chemical effects in Auger spectra are quite commonly observed, but they are difficult to interpret compared to those in XPS since additional core levels are involved in the Auger process. Some examples of the changes to be seen in the KLL spectrum of carbon in different chemical environments are given in Figure 24 [140–142]. Such spectra are typical components of data matrices (see Section 2.1.4.2) derived from AES depth profiles (see below).

Associated with prominent features in an Auger spectrum are plasmon loss features, as are also associated with photoelectron peaks in an XPS spectrum. Plasmon energy losses arise from excitation of modes of collective oscillation of the conduction electrons by outgoing secondary electrons of sufficient energy. Successive plasmon losses suffered by the backscattered primary electrons can be seen in both the $N(E)$ and the $dN(E)/dE$ spectra of Figure 20, from boron. The same magnitude of plasmon loss, $\approx$27 eV, is also associated with the KLL Auger peak of boron. Plasmon losses are also present in Figures 21, 22, 23, 24, but they are difficult to disentangle from minor Auger features.

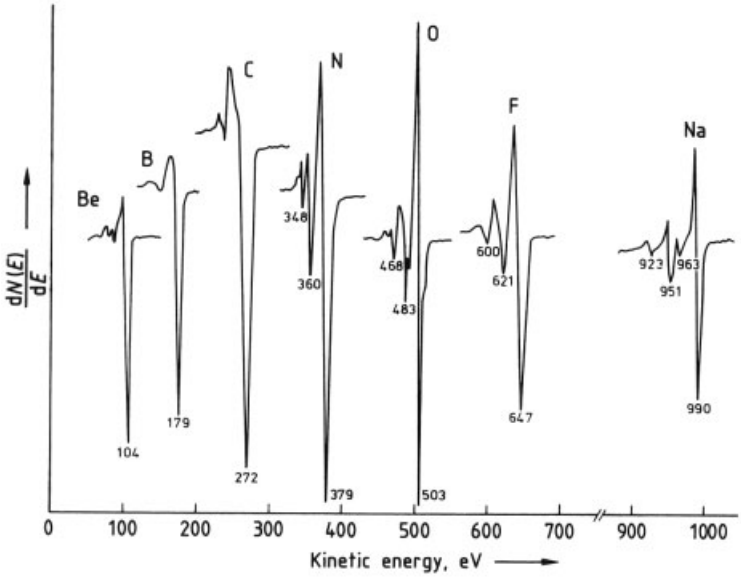

Figure 21. KLL Auger series characteristic of the light elements [139]

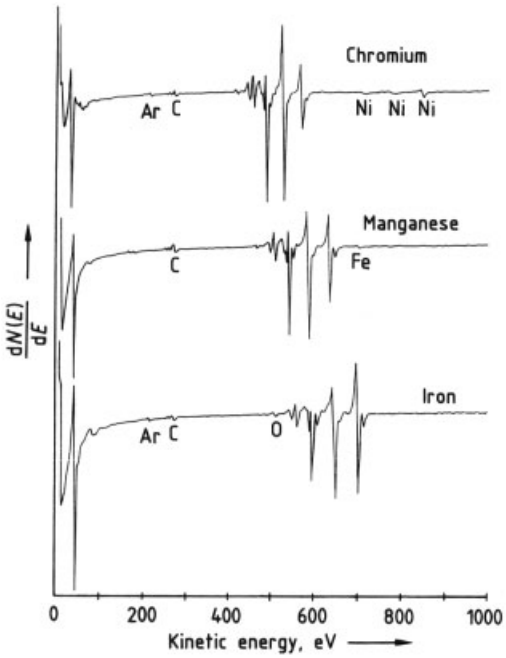

Figure 22. LMM and MMM Auger series in the middle of the first series of transition elements [139]

Another type of loss feature not often recorded in Auger spectra can also be seen in Figure 20, in the d$N(E)$/dE spectrum at 810 eV. This type arises from core-level ionization and forms the basis for the technique of core electron energy loss spectroscopy (CEELS). A primary electron interacting with an electron in a core level can cause excitation either to an unoccupied continuum state, leading to complete ionization, or to localized or partly localized final states, if available. Since the primary electron need not give up all its energy in the interaction, the loss feature appears as a step in the secondary electron spectrum, with a tail decreasing progressively to lower energies. Differentiation, as in Figure 20, accentuates the visibility. In that figure, the loss at 810 eV corresponds to the K-shell ionization edge of boron at about 190 eV.

2.3.4. Quantification and Depth Profiling

2.3.4.1. Quantification

If ionization of a core level X in an atom A in a solid matrix M by a primary electron of energy E_{p} gives rise to the current $I_{\mathrm{A}}(XYZ)$ of electrons produced by the Auger transition XYZ, then the Auger current from A is

$$I_{\mathrm{A}}(XYZ) = K\sigma(E_{\mathrm{p}},X)\,[1+r_{\mathrm{M}}(X,\alpha)]\,\bar{N}_{\mathrm{A}}\,\lambda_{\mathrm{M}}(E_{\mathrm{A}})\cos\theta \qquad (2.13)$$

where $\sigma(E_{\mathrm{p}}, X)$ is the cross section for ionization of level X by electrons of energy E_{p}; r_{M} the

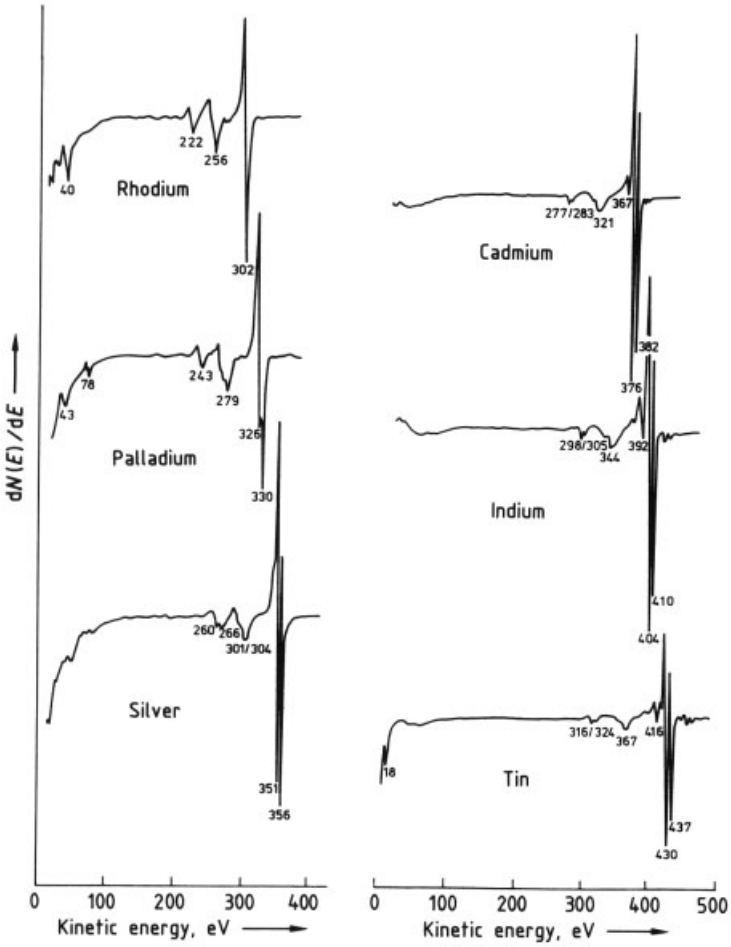

Figure 23. MNN Auger series in the second series of transition elements [139]

backscattering factor that takes account of the additional ionization of X (binding energy E_{B}) by inelastic electrons with energies between E_{p} and E_{B}; α the angle of incident electron beam to surface normal; $\bar{N}_{\mathrm{A}}$ the atomic density of A averaged over depth of analysis; $\lambda_{\mathrm{M}}(E_{\mathrm{A}})$ the inelastic mean free path in matrix M containing A, at the Auger kinetic energy E_{A}; and θ the angle of Auger electron emission to surface normal.

Similarly to Equation (2.6), K is a constant of proportionality containing fixed operational parameters such as incident electron current density, transmission of the analyzer at the kinetic energy E_{A}, efficiency of the detector at the kinetic energy E_{A}, and the probability of the Auger transition XYZ.

The cross section for electron impact ionization was already mentioned in Section 2.3.2.2 in connection with electron sources, and various experimental and theoretical cross sections are shown in Figure 18 for the particular case of the K shell of nickel. The expression for the cross section derived by Casnati et al. [138] gives reasonably good agreement with experiment, with the earlier expression of Gryzinski [143] also being useful.

For the inelastic mean free path, Equation (2.10) given for XPS also applies.

The new factor in Equation (2.13) compared to Equation (2.6) for XPS is the backscattering factor r_{M}. Ichimura and Shimizu [144] have carried out extensive Monte Carlo calculations of the backscattering factor as functions of pri-

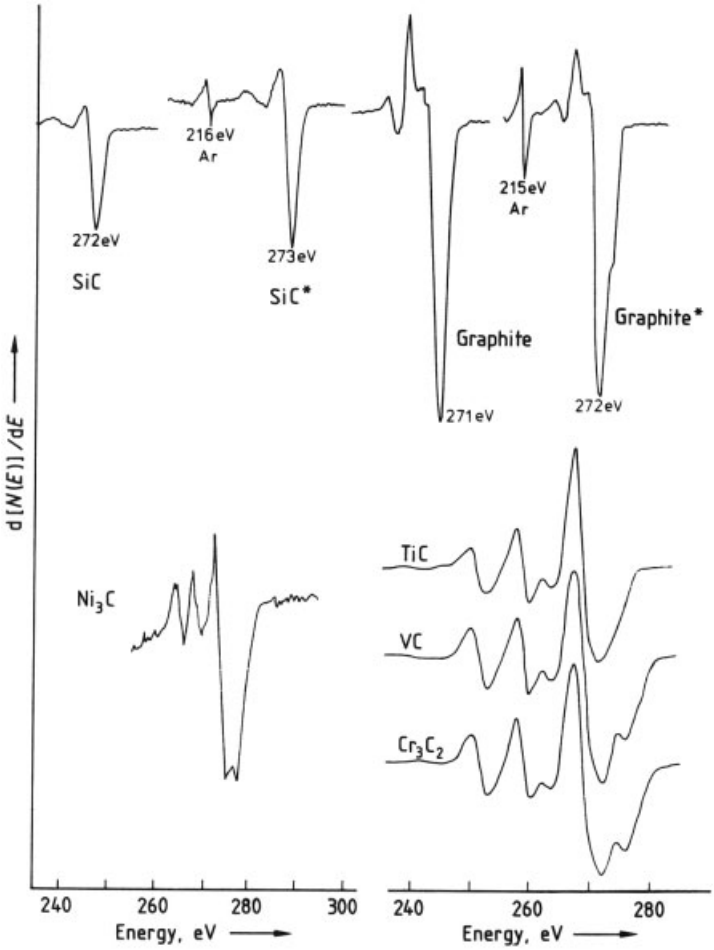

Figure 24. Examples of the effect of different chemical states on the KLL Auger spectrum of carbon [140–142] (SiC∗ and graphite∗ denote Ar^+-bombarded surfaces of SiC and graphite, respectively)

mary beam energy and angle of incidence, of atomic number and of binding energy, and a selection of their results for $E_p = 10$ keV at normal incidence is shown in Figure 25. The best fit of their results to experiment gives the following relationship:

$$r = \left(2.34 - 2.1Z^{0.14}\right) U^{-0.35} + \left(2.58Z^{0.14} - 2.98\right) \quad \text{for} \quad \alpha = 0^\circ$$

$$r = \left(0.462 - 0.777Z^{0.20}\right) U^{-0.32} + \left(1.5Z^{0.20} - 1.05\right) \quad \text{for} \quad \alpha = 30^\circ$$

$$r = \left(1.21 - 1.39Z^{0.13}\right) U^{-0.33} + \left(1.94Z^{0.13} - 1.88\right) \quad \text{for} \quad \alpha = 45^\circ \tag{2.14}$$

where Z is the atomic number and $U = E_p/E_B$.

Thus, as for XPS, in principle the average surface concentration $\bar{N}_A$ can be calculated from the measurement of the Auger current, according to Equation (2.13). Again, as in XPS, relative sensitivity factors are generally used. The Auger current I_A^∞ for the same transition *XYZ* in a standard of pure A is measured under the same experimental conditions as in the analysis of A in M, whereupon the ratio of the atomic concentrations is

$$\bar{N}_A / N_A^\infty = \frac{I_A\,(1+r_A)\,\lambda_A R_A^\infty}{I_A^\infty\,(1+r_M)\,\lambda_M R_M} \tag{2.15}$$

since the ionization cross section σ is the same for both standard and sample. Then, as in XPS, assuming nearly equal roughness factors $R_A^\infty \approx R_M$, Equation (2.15) can be reduced to

$$\bar{N}_A = S_A I_A \tag{2.16}$$

where

$$S_A = \frac{\bar{N}_A^{\infty}\,(1+r_A)\,\lambda_A}{I_A^{\infty}\,(1+r_M)\,\lambda_M} \tag{2.17}$$

where S_A is then the relative sensitivity factor. Normally, values of S_A are derived empirically or semiempirically. Tables of such sensitivity factors have been published by PAYLING [145] and by MROCZKOWSKI and LICHTMAN [146] for differential Auger spectra, and for the direct, undifferentiated spectra by SATO et al. [147]. A comprehensive discussion of the assumptions and simplifications involved and the corrections that should be applied is given in [18], [148]. Although modern Auger spectrometers detect the spectra in the undifferentiated mode, many workers still evaluate the spectra in the electronically differentiated mode, that is, using the peak-to-peak heights, so that changes in peak shapes due to changes in chemical bondings are not taken into account and quantification in AES is still not as accurate as that in XPS.

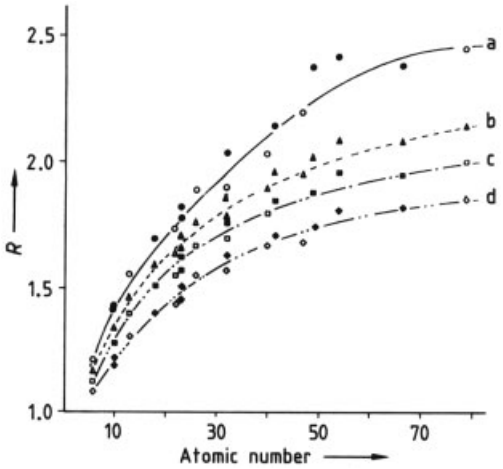

Figure 25. Values of the backscattering factor R (i.e., $1+r$) calculated as a function of atomic number for an electron beam of 10 keV at normal incidence by ICHIMURA et al. [144]
a) $E_B = 0.1$ keV; b) $E_B = 0.5$ keV; c) $E_B = 1.0$ keV; d) $E_B = 2.0$ keV

2.3.4.2. Depth Profiling

As in XPS, elemental distributions near the surface, but to depths greater than the analytical depth resolutions, are frequently required when using AES. Again, the universally employed technique is that of progressive removal of the surface by ion sputtering; again one has to be aware that the electron spectroscopic sampling depth corresponds roughly to the depth of the layer modified by preferential sputtering effects, and again the edges of the sputter crater should not contribute to the signal.

A variation on depth profiling that can be performed by modern scanning Auger instruments (see Section 2.4) is to program the incident electron beam to jump from one preselected position on a surface to each of many others in turn, with multiplexing at each position. This is called multiple-point analysis. Sets of elemental maps acquired after each sputtering step or each period of continuous sputtering can be related to each other in a computer frame-store system to derive a three-dimensional analysis of a selected micro volume.

2.3.5. Applications

Like XPS, the application of AES has been very widespread, particularly in the earlier years of its existence; more recently, the technique has been applied increasingly to those problem areas that need the high spatial resolution that AES can provide and XPS as yet cannot. Since data acquisition in AES is faster than in XPS, it is also employed widely in routine quality control by surface analysis of random samples from production lines of, for example, integrated circuits. Particularly in semiconductor industry, SIMS is a competing method. Note that AES and XPS on the one hand, and SIMS/SNMS on the other, both in depth-profiling mode, are complementary, the former gaining signals from the sputter-modified surface, and the latter from the flux of sputtered particles.

2.3.5.1. Grain Boundary Segregation

One of the original applications of AES, and still one of the most important, is the analysis of grain boundaries in metals and ceramics. Very small amounts of impurity or dopant elements in the bulk material can migrate under appropriate temperature conditions to the boundaries of the

grain structure and accumulate there. In that way the concentration of minor elements at the grain boundaries can become much higher than in the bulk, and the cohesive energy of the boundaries can be so altered that the material becomes brittle. Knowledge of the nature of the segregating elements and their grain boundary concentrations as a function of temperature can be used to modify fabrication conditions and thereby improve the strength performance of the material in service.

GORETZKI [149] has discussed the importance and the analysis of internal surfaces such as grain boundaries and phase boundaries, and given several examples. He emphasizes that to avoid ambiguity in interpretation of the analysis, the internal surfaces must be exposed by fracturing the material under UHV conditions inside the electron spectrometer. If this is not done, atmospheric contamination would immediately change the surface condition irrevocably. One of the examples from GORETZKI's paper is given in Figure 26, in which Auger spectra in the differential mode from fracture surfaces in the embrittled (Fig. 26 A) and unembrittled (Fig. 26 B) states of a 12 % Cr steel containing phosphorus as an impurity are compared. Embrittlement occurs when the steel is held at 400 – 600 °C for a long time. Note in the spectrum from the embrittled material not only the greatly enhanced phosphorus concentration at the grain boundary, but also increased amounts of chromium and nickel with respect to iron. By setting the analyzer energy first to that of the phosphorus peak, and then to that of the principal chromium peak, Auger maps could be recorded showing the distribution of phosphorus and chromium over the fracture surface. The two maps showed high correlation of the two elements, indicating association of phosphorus and chromium at the grain boundary, possibly as a compound.

More recent work [150], [151] on the above 12 % Cr steel deal with the enhanced appearance of phosphorus at the grain boundaries of embrittled material. Other high-chromium steels have also been investigated [152], [153]. REICHL et al. [154] investigated the temperature dependence of surface segregation of S and N in α-Fe. In an initial phase, S diffusion took place at the grain boundaries and displaced N at the surface at higher temperatures. Other applications in the study of grain boundary segregation in metals have included the effects of boron, zirconium, and aluminum on Ni_3Al intermetallics [155], the segregation of phosphorus and molybdenum to grain boundaries in the superalloy Nimonic PE16 as a function of aging treatment [156], and the enrichment of chromium carbide and chromium iron carbide at the grain boundaries in steel [157].

Application of AES to zirconia ceramics is described by MOSER et al. [158]. Elemental maps of Al and Si demonstrate the grain boundary segregation of small amounts of silica and alumina impurities in these ceramics.

2.3.5.2. Semiconductor Technology

With ever-increasing miniaturization of integrated circuits goes a need for the ability to analyze ever smaller areas so that the integrity of fabrication at any point on a circuit can be checked. This checking process includes not only microdetails such as continuity of components in the form of thin films and establishment of the correct elemental proportions in contacts, diffusion barriers, and Schottky barriers, but also the effectiveness of more macrotreatments such as surface cleaning. In all these types of analyses, depth profiling (see Section 2.3.4.2) is used extensively because, in general, elemental compositional information is required, not information about chemical state. As mentioned in Section 2.1.4.2, chemical-state information tends to become blurred or destroyed by ion bombardment.

Various materials have been prescribed as *diffusion barriers* to prevent one circuit material from diffusing into another (often Si) at the temperatures needed for preparation of certain components such as oxide films in integrated circuits. ZALAR et al. [159] studied the interdiffusion at the Al_2O_3/Ti interface. They covered Si(111) with TiN thin film as diffusion barrier, then with a 45 nm of polycrystalline Ti, and finally with 55 nm of amorphous Al_2O_3. The same samples were heated to different temperatures at a rate of 40 °C/min, and the interdiffusion processes were investigated by AES depth profiling. Figure 27 shows that oxygen from Al_2O_3 and, to a lesser extent, also Al migrate into the Ti layer. (The "N + Ti" intensity in the Ti layer is only due to Ti, while in the neighboring TiN layer the

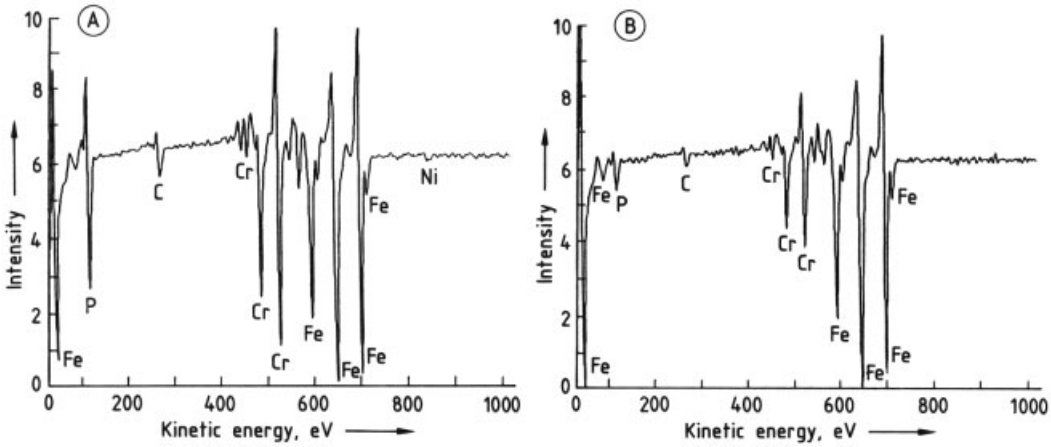

Figure 26. Auger spectra from fracture surfaces of a 12 % Cr steel [149]
A) Embrittled state; B) Unembrittled state

N + Ti intensity belongs to Ti and N because of peak overlapping.) Control measurements indicate that oxygen does not stem from the outer gas phase. In the interface region a new phase consisting of α-Ti_3Al is formed and detected by selected-area electron diffraction. Such AES investigations are normally carried out in combination with other surface-analytical techniques, in this case additionally with XPS, XRD, TEM and electron diffraction. Another work dealing with the analysis of CVD W_xN films as barriers for Cu metallization was published by Hua Li et al. [160], using AES, as well as RBS, AFM, XRD, and TEM.

The nature of metallic contacts to compound semiconductors has also been an interesting problem. Ohmic contacts must be made to semiconducting surfaces, otherwise circuits could not be fabricated. However, during fabrication, excessive interdiffusion must not occur, since otherwise the electrical characteristics of the semiconductor cannot be maintained. Among the many studies of the reactions of metallizing alloys with compound semiconductors is that of Roberts et al. [161], who investigated the W – TiN – Ti – Si contact structure using a multispectral Auger microscope (MULSAM) and TEM. Especially for the detection of small particles and defects which may be present in the ULSI regime, an Auger electron spectrometer with high spatial resolution imaging capability was developed which allows the inspection of wafers up to 200 mm in diameter [162].

AES has also been used to study preferential sputtering of $TiSi_x$ for forming low-resistivity conductor films in ULSI devices [163], and the electromigration behaviour of Au – Ag films on SiO_2 by using AES, XPS, and AFM [164].

AES has also been applied to a study of the oxidation of copper in printed circuit boards [524], of sulfur passivation of GaAs [525], of cleaning of germanium and GaAs with hydrogen dissociated by a noble-gas discharge [526], and of segregation and corrosion of silver layers electroplated onto light emitting diode (LED) lead frames as part of the final packaging [527].

2.3.5.3. Thin Films and Interfaces

The nature of the interface formed between very thin metallic films and substrates of various types has been studied extensively by AES, just as it has by XPS (see Section 2.1.6.6). The degree and extent of interaction and interdiffusion can be established by AES with depth profiling, although chemical information is normally absent, but the development of the interface is usually studied by continuous recording of Auger spectra during film deposition. Zhu *et al.* [165] deposited Ti or Cr layers on diamond samples. The Ti – diamond or Cr – diamond specimens were annealed at 600 °C for 4 h during metallization. Figures 28 A and B show the depth profiles before and after annealing of a Ti – diamond specimen, and Figures 28 C and D show the Ti

LM_1M_4 Auger line shapes in the $N(E)$ mode and the C KLL line shape in the differentiated $dN(E)/dE$ mode at various depths. The unstructured Ti signal at the surface can be attributed to TiO_2 due to a smooth oxidation of the surface. But the majority of the layer consists of TiC. The formation of TiC is also recognizable by the strongly structured carbon line, which changes its shape with depth and glides into that of diamond. All these features are consistent with the existence of TiC, indicating complete metallization.

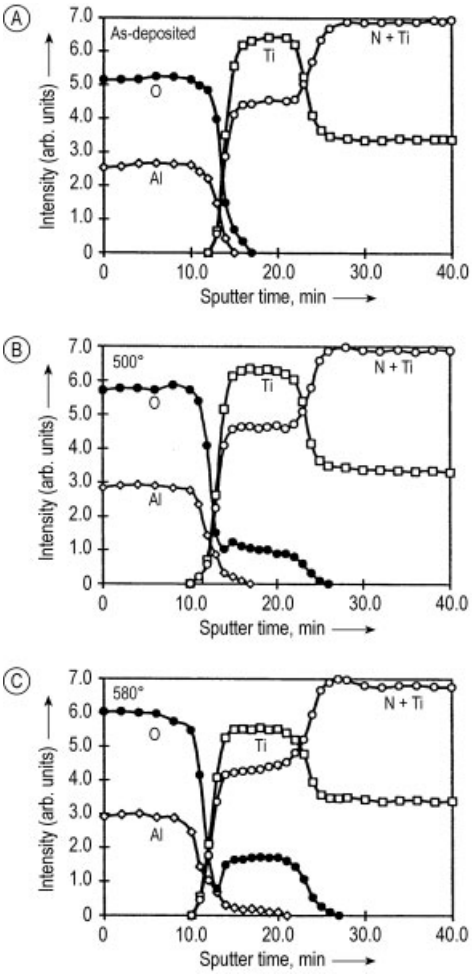

Figure 27. AES sputter depth profiles of the α-Al_2O_3 – Ti thin-film structure on a smooth Si substrate covered with a TiN thin-film diffusion barrier
A) As-deposited; B) After heating to 500 °C; C) After heating to 580 °C [159].

Other interesting thin-film studies using AES have included the growth of platinum on TiO_2- and SrO-terminated $SrTiO_3$ (100) single-crystal substrates [166], of epitaxial niobium films on TiO_2 (110) [167], the interaction of copper with rhenium (0001) surface [168], and the characterization of radio-frequency sputtered TiN films on stainless steel [169].

A multilayer-type structure, probably due to cords in the molten zone between single arc sprayed (0.25 MPa) Ni droplets and a steel substrate were found in AES point depth profiles [170]. This particular arc-spraying condition proved to yield the best adhesion. Plasma-sprayed Al_2O_3 layers separated from pre-oxidized Ni Substrate exhibited a micrometer-thick NiO layer on the substrate-sided face and micrometer-deep oxide interdiffusion [171]. Also in this case, AES point depth profiling substantiated technological assumptions on adhesion mechanisms.

2.3.5.4. Surface Segregation

The surface composition of alloys, compounds, and intermetallics may change profoundly during heat treatment as a result of segregation of either or both constituent and impurity elements. Surface enhancement may also occur through chemical forces during oxidation or other reaction. The mechanism of segregation is similar to that of grain boundary segregation already discussed, but of course at the free surface material is much more likely to be lost through reaction, evaporation, etc., so that excessive continued segregation can change the bulk properties as well as those of the surface. Nearly all such studies have been carried out on metals and alloys, and AES has been applied to the problem almost from its inception.

Sometimes the segregation of metallic and nonmetallic elements together to surfaces can result in the formation of surface compounds, as demonstrated by Uebing [172], [173]. He heated a series of Fe – 15 % Cr alloys containing either 30 ppm nitrogen, 20 ppm carbon, or 20 ppm sulfur to 630 – 800 °C and obtained the typical Auger spectra shown in Figure 29. In each case an excess of both chromium and the nonmetal occurs in the spectrum, over that expected if the bulk concentration were maintained. The fine structure of the nitrogen and carbon KLL spectra is characteristic of nitride and

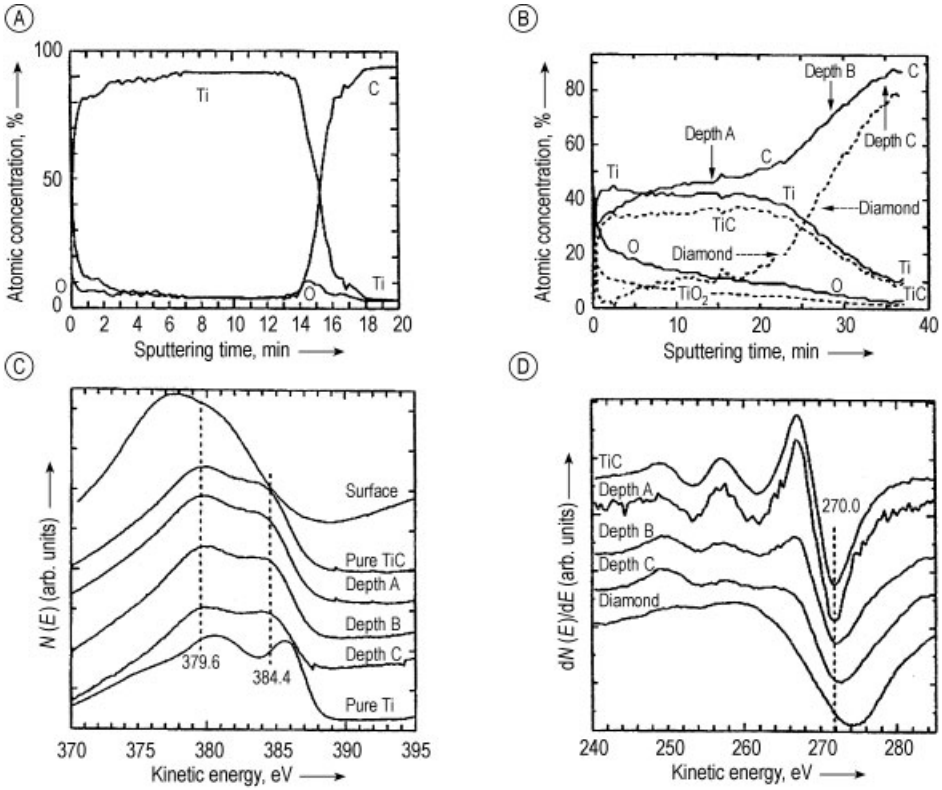

Figure 28. Depth profiles of a Ti – diamond layer before (A) and after annealing (B) at 600 °C for 4 h [165]. Auger line shapes of Ti LM_1M_4 (C) and C KLL (D) at various depths.

carbide, respectively. Hence, the co-segregation of chromium with nitrogen and with carbon has led to the formation of surface nitrides and carbides.

Song et al. [174] investigated surface segregation in $Ni_{0.65}Nb_{0.35}$ up to 500 °C and 0.1 MPa O_2 atmosphere. Below 100 °C, Nb segregated to the surface with formation of Nb_2O_5. At 300 °C and above, Ni segregated to the surface as Ni and NiO. This inverse segregation can be explained by different diffusion speeds of O and Ni in the Nb_2O_5 layer formed at the surface.

Other investigations involving AES, often with depth profiling, deal with the surface segregation of Ag in Al – Ag (4.2 %) [175], of Sn in Cu and formation of superficial Sn – Cu alloy [176], of Mg in Al – Mg alloys [177], and of Sb in Fe – Sb (4 %) alloy [178]. Note the necessity to disriminate segregation, that is, original sample properties, from the artifact of preferential sputtering.

2.4. Scanning Auger Microscopy (SAM)

If an incident electron beam of sufficient energy for AES is rastered over a surface in a manner similar to that in a scanning electron microscope (SEM), and if the analyzer is set to accept electrons of Auger energies characte ristic of a particular element, then an *elemental map or image* is again obtained, similar to XPS with the Quantum 2000 (see Section 2.1.2.5).

An example is shown in Figure 30 [179]. The surface was that of a fractured compact of SiC to which boron and carbon had been added to aid the sintering process. The aim of the analysis was to establish the uniformity of distribution of the additives and the presence or absence of impurities. The Auger maps show not only very nonuniform distribution of boron (Fig. 30 A) but also a strong correlation of boron with sodium (Fig. 30 C), and a weaker correlation of boron with potassium (Fig. 30 B). Point analyses at A

and B marked on the images reveal the presence of sulfur and calcium in some areas as well. In this case the sintering process had not been optimized.

Dedicated *scanning auger microscopes* with ultimate spatial resolution have been developed towards the smallest possible spot size and still acceptable Auger signal-to-noise ratio obtained within tolerable acquisition time. Electron guns used for SAM employ electromagnetic focusing. Since focusing electrons of low energy into a small spot is difficult, energies used in SAM are in the range of 25 – 50 keV, with beam currents on the order of 1 nA or less and a minimum spot size of about 15 nm.

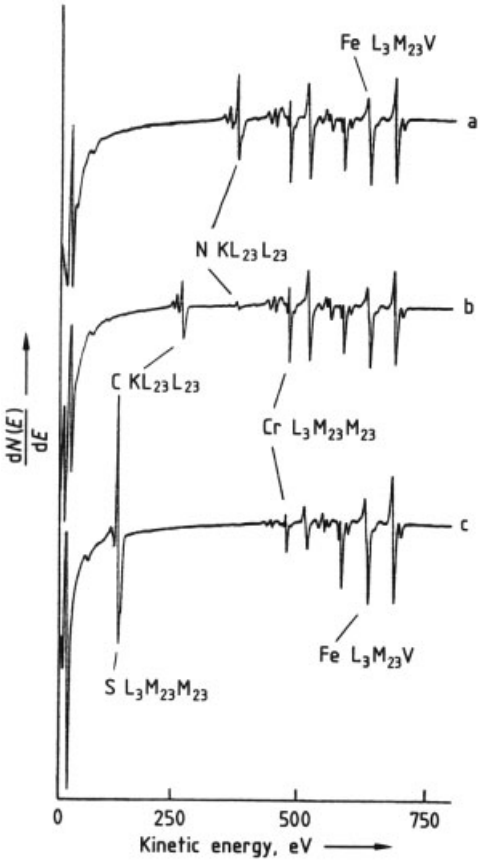

Figure 29. Formation of surface compounds on Fe – 15 % Cr alloys by co-segregation of chromium and a nonmetallic element [172]
a) Nitride; b) Carbide; c) Sulfide

2.5. Other Electron-Detecting Techniques

2.5.1. Auger Electron Appearance Potential Spectroscopy (AEAPS)

Due to the emission of an Auger electron as an alternative to soft X-ray photons, the total secondary electron yield will show an increase as an ionization threshold is crossed. It is the total secondary yield that is monitored in AEAPS, in a rather simple experimental arrangement. The secondary current arises not just from the Auger electrons themselves, but also from inelastic scattering of Auger electrons produced at greater depths below the surface. Since the yield change is basically a measure of the probability of excitation of a core-level electron to an empty state above the Fermi level, the fine structure above the threshold will be similar to that seen in SXAPS. One of the problems in AEAPS is that of a poorly behaved background, which means that SXAPS is preferred.

2.5.2. Electron Energy Loss Methods

2.5.2.1. Electron Energy Loss Spectroscopy (EELS) and Core-Electron Energy Loss Spectroscopy (CEELS)

The techniques of EELS and CEELS can be discussed together because they use the same experimental arrangement, basically that of conventional AES. A beam of primary electrons interacts with electrons in surface atoms, and the energy losses suffered by the primary electrons are measured by recording the appropriate part of the secondary electron spectrum in an energy analyzer. In *EELS* the losses are relatively small, being due to the energies required to stimulate electronic transitions within the valence band of the solid, and collective oscillations of the conduction electrons (plasmon oscillations). For each solid the plasmons have energies characteristic of the bulk structure and of the surface structure, the surface plasmon having theoretically an energy lower than that of the bulk by a factor of $2^{1/2}$. In *CEELS* the primary electron interacts with a core level, usually but not always leading to ionization, and the losses are in general much greater than in ELS. Thus in EELS the observed losses are normally within 30 – 50 eV

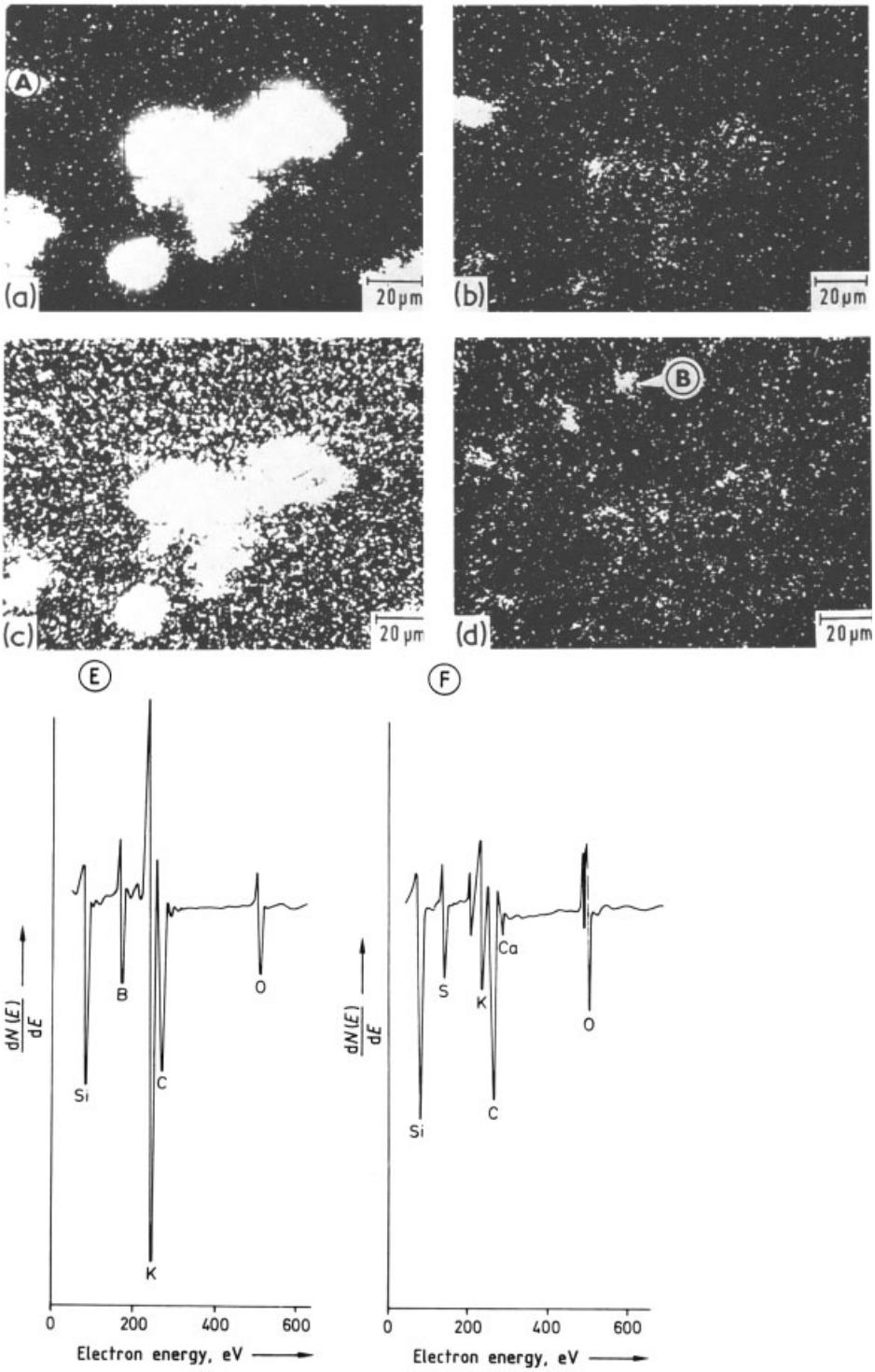

Figure 30. SAM map of fractured SiC after sintering with B addition [179]
a) – d) Elemental maps in boron, potassium, sodium, and oxygen, respectively; E), F) Point analyses at points ○A and ○B, respectively

of the primary energy, whereas in CEELS they may be several hundred electronvolts away.

Since all loss features are by definition associated with the primary energy, features characteristic of the surface can be compared to those of the bulk by variation of the primary energy (i.e., by effectively varying the electron escape depth). Such variation is standard practice in ELS, in which the range of primary energies used is typically 100 – 1000 eV. In CEELS, in contrast, although the relation between primary energy and escape depth still applies, the primary energy must be high enough to ionize the core level with adequate efficiency, as discussed in Section 2.3.2.2 for AES. Thus for CEELS the primary energy range would be the same as for conventional AES (i.e., 2 – 5 keV).

CEELS has been applied more to the study of core-electron excitation to localized final atomic states than to excitation to the continuum (i.e., complete ionization). Core-level ionization thresholds can in principle be used to obtain an elemental composition, but the spectral features are weak, and the analysis is performed far better by XPS. Nevertheless, where shallow, highly localized final states exist, CEELS has been able to provide useful information.

2.5.2.2. High-Resolution Electron Energy Loss Spectroscopy (HREELS)

In this technique, electrons of very low primary energy, typically 3 – 5 eV, are directed at a surface carrying adsorbed molecules. As the electrons approach closely enough, they interact with the molecular field and are reflected (i.e., scattered) having lost discrete amounts of energy corresponding to characteristic vibrational frequencies, either within the molecules or between the molecules and the surface. Analysis of the scattered electron energy distribution then reveals the energy losses, and hence the vibrational frequencies. If the molecules are attached to the surface only loosely (*physisorption*), many of the characteristic vibrations within a molecule are similar to those observable in the gas phase. These consist of various stretching and deformation modes of vibration. In addition, other modes appear as a result of attachment to the surface. In the gas phase a molecule can translate and rotate freely, but if one end is fixed to a surface it can no longer do so, and therefore certain translational and rotational modes of vibration become observable that cannot be observed in the free molecule. Should the molecule be attached strongly to the surface (*chemisorption*), perhaps with the loss of reactive groups, and certainly with a transfer of electronic charge, then the characteristic free-molecule stretching vibrations can become so distorted that they no longer resemble those seen in the gas phase. Thus the HREELS spectrum from a complex polyatomic molecule adsorbed on a surface can itself be very complex and contain a great deal of information concerning the way in which the molecule has reacted with the surface and also with neighboring molecules. The orientation of the molecule can often be deduced as well, particularly when HREELS is used with other techniques such as LEED, UPS, and TDS.

Because the energy widths of the characteristic vibrational energies are small, < 0.1 meV, for optimum energy resolution one must achieve as narrow an energy spread in the incident electron beam as possible. In all HREELS equipment, therefore, extreme measures are taken to reduce the spread in a variety of ways. Most importantly, the incident electrons are energy selected in a monochromator that is basically an electron energy analyzer used in reverse. Monochromators properly constructed can reduce the energy spread in the incident beam to 2 – 3 meV, but to maintain that width in the beam at the sample surface requires the complete elimination of stray magnetic and electrostatic fields around the sample. Thus very careful screening of the entire HREELS analytical arrangement is necessary. Many experiments are also carried out at low temperature to reduce the thermal broadening of characteristic vibrations.

2.5.3. Diffraction Methods

2.5.3.1. Low-Energy Electron Diffraction (LEED)

Both LEED and RHEED are probes of surface structure that depend on the existence of long-range order of periodic structure in the surface. They do not provide compositional information and are always used along with other techniques such as AES that can analyze for composition.

Low-Energy Electron Diffraction. In LEED a collimated and monoenergetic beam of electrons of variable energies between about 20 and 1000 eV strikes the surface of a single crystal at normal incidence. An energy selector, almost universally consisting of a set of spherical grids centered on the specimen, repels all secondary electrons except those elastically scattered (i.e., those at the primary energy). The elastically scattered electrons are then accelerated to a spherical fluorescent screen, on which the diffraction pattern can be seen. Measurement of diffraction spot intensity can be made either with a photometer or with a movable Faraday cup.

Diffraction occurs due to the wave nature of electrons. The waves represented by the incident electron beam interfere as a result of reflection by a periodic crystal structure in such a way that the reflected beams are anisotropic in direction and are in fact concentrated in certain directions related directly to the crystal structure. Where the reflected beams are intercepted by the fluorescent screen, a spot pattern results. The purely geometric relationship between the observed pattern and the surface structure is governed by the well-known Bragg equation, which connects the directions of the interference maxima with the surface lattice spacing and the primary energy. The observed diffraction pattern is a direct representation of the reciprocal lattice of the surface. Since the electron wavelength is of the same order of magnitude as the surface lattice spacing, only a few diffraction spots appear at low primary energies (<100 eV), but as the energy is increased, further spots appear and move with energy toward the normally deflected (i.e., undiffracted) beam. At the same time, the existing spots vary in intensity, appearing and disappearing on the fluorescent screen as they move with energy inward along radial directions.

The simple geometric, or kinematic, representation of the LEED pattern is valid only for scattering from those atoms actually in the surface (i.e., exposed to the incident electron beam). However, the incident beam has a finite depth of penetration, which of course increases with increasing energy; therefore the periodicity of the crystal normal to the surface must also be taken into account. This can be explored by recording the intensity of a particular diffraction spot as a function of primary electron energy (i.e., wavelength). The normally reflected beam will have intensity maxima due to the vertical periodicity, which are called primary Bragg peaks if the maxima are predicted by the Bragg equation, and secondary Bragg peaks if they are not. The latter are caused by multiple scattering effects, which can be accounted for only in a dynamic theory of scattering. The latter is complicated and can be applied only by using computer simulation, in which a feasible structure for the surface is guessed, from which diffraction intensities are calculated and compared with experiment. In this way a surface structure can be solved completely, but ambiguities cannot be eliminated in most cases.

2.5.3.2. Reflection High-Energy Electron Diffraction (RHEED)

The same diffraction mechanism is used in RHEED, but the experimental arrangement is quite different, and the observed pattern differs as well. Much higher primary electron energies are used than in LEED, in the range of 15 – 30 keV, and the primary beam is directed onto the crystal surface at near grazing incidence, at angles $< 5°$. As a result, the diffraction pattern is a mixture of spots and streaks. For such low angles of incidence the surface must be very smooth and flat, which is a limitation of the technique, but the requirement also means that any growth feature appearing above the original surface is easily identified since it produces its own characteristic diffraction pattern in transmission. In a variation of RHEED, the angle of the primary beam is varied backward and forward through a few degrees, which is called *rocking*, by analogy with the rocking patterns recorded in X-ray diffraction. Analysis of surface structure, with the assistance of sophisticated data processing, is less ambiguous when the rocking variation is used.

RHEED has also been used in basic studies, although to a far lesser extent because of its inherent restrictions, but in recent years it has found technological application as well. Semiconductor surfaces are usually flat, and epitaxial films grown on them are often flat as well. In those circumstances RHEED has proved to be a useful tool in the quality control of the structural condition of a semiconductor substrate and

then of the perfection or otherwise of any films or layers grown or deposited on it. Thus each of the large UHV systems built for molecular beam epitaxy (MBE) contains a RHEED diagnostic system as one of the standard inspection techniques of the epilayer during and after growth.

2.5.4. Ion-Excitation Method

2.5.4.1. Ion (Excited) Auger Electron Spectroscopy (IAES)

Auger emission following creation of a core hole by electron or photon irradiation has been described under AES and XPS. Incident ions can also create core holes, so that Auger emission as a result of ion irradiation can also occur, giving rise to the technique of IAES. The ions used are normally noble-gas ions, but protons and α particles have occasionally been used as well. In addition to the normal Auger features found in AES and XPS spectra, peaks are found in IAES spectra arising from Auger transitions apparently taking place in atoms or clusters sputtered from the surface. These peaks do not always coincide with those found in gas-phase excitation, and others are found that are not present in gas-phase measurements. Because of the complexity of the spectra, IAES cannot be used directly as an analytical technique, but it is very useful in basic physics experiments that study Auger processes occurring in excited atoms.

2.5.4.2. Ion-Neutralization Spectroscopy (INS)

The technique of INS is probably the least used of all those described here, because of experimental difficulties, but it is also one of the physically most interesting. Ions of He^+ of a chosen low energy in the range 5 – 10 eV approach a metal surface and within an interaction distance of a fraction of a nanometer form ion – atom pairs with the nearest surface atoms. The excited quasi molecule so formed can deexcite by Auger neutralization. If unfilled levels in the ion fall outside the range of filled levels of the solid, as is the case for He^+, then an Auger process can take place in which an electron from the valence band of the solid fills the core hole in the ion and the excess energy is given up to another valence electron, which is then ejected. Since either of the valence electrons can come from anywhere within the valence band, the observed energy spectrum reflects the local density of states at the solid surface, but is a self-convolution of the LDOS and of transition probabilities across the valence band. The technique of INS requires a complex mathematical treatment for deconvolution of the spectra. Because the technique of STS is also capable of extracting the LDOS, the results from INS and STS should be comparable, but such a comparison does not seem to have been attempted yet.

2.5.4.3. Metastable Quenching Spectroscopy (MQS)

MQS is in a sense an extension of INS. Instead of He^+ ions, helium atoms in metastable states are used in the incident beam, at the same low energies. The excited singlet state $He*(2^1S_1)$ has an energy of 20.62 eV and a lifetime of 2×10^{-2} s, and so is suitable. It is produced by expanding helium gas at high pressure through a nozzle into a cold cathode discharge sustained by combined electrostatic and magnetic fields. The high fields prevent ions and fast neutrals from leaving the discharge, and the beam is then nearly all $He*(2^1S_1)$. As the metastable ion approaches the surface, either of two mechanisms can lead to deexcitation, both resulting in Auger emission similar to INS. If the excited level in the atom can resonate with empty states at the Fermi level of the surface, electron transfer from the atom to the surface can occur, leading to resonance ionization of the helium atom. Auger neutralization then takes place as in INS, and the resultant spectrum is again a self-convolution of the LDOS. If, however, the excited level cannot resonate with empty surface states, then direct Auger deexcitation can occur, in which the hole in the inner shell of the metastable helium atom is filled from a surface state of the sample, followed by ejection of the excited electron from the helium atom. The process is also called Penning ionization. In the latter process, only one electron is ejected, and the resultant spectrum is thus an unconvoluted reflection of the LDOS.

2.5.5. Inelastic Electron Tunneling Spectroscopy (IETS)

IETS is unique in that it is entirely surface specific, but does not require a vacuum environment and has almost never been performed in UHV. Its principle is relatively simple. If two metals are separated by a thin (≈3 nm) insulating layer, and a voltage is applied across them, electrons can tunnel from one to the other through the insulator, and if no energy is lost by the electrons, the process is known as elastic tunneling. However, should discrete impurity states occur in the interface between either metal and the insulator, or molecules exist in such an interface that possess characteristic vibrational energies, then the tunneling electron can give up some of this energy either to the state or to the vibrational mode, before reaching the other metal. This is called inelastic tunneling. Obviously, the applied voltage must be greater than the state or vibrational energies. If the current across the metal – insulator – metal sandwich is recorded as a function of applied voltage, the current increases as the threshold for each state or vibrational mode is crossed. The increases in current are in fact very small, and for improved detectability, the current is double differentiated with respect to voltage, thereby providing in effect a vibrational spectrum that can be compared directly with free-molecule IR and Raman spectra.

The metal – insulator – metal sandwich is known as a tunnel junction, and its preparation is all-important. The standard junction consists of an aluminum strip ca. 60 – 80 nm thick and 0.5 – 1.0 mm wide deposited in a very good vacuum onto scrupulously clean glass or ceramic, the surface of the strip then being oxidized either thermally or by glow discharge. The resultant oxide layer is extremely uniform and about 3 nm thick. Introduction of adsorbed molecules onto the oxide layer, or "doping" as it is called, is then effected either by immersion in a solution followed by spinning to remove excess fluid, or by coating from the gas phase. The final stage in preparation is deposition of the second metal (invariably lead) of the sandwich; this deposition is carried out in a second vacuum system (i.e., not the one used for aluminum deposition), the final thickness of lead being about 300 nm, and the width of the lead strip being the same as that of the aluminum strip. The reason for using lead is that all IETS measurements are carried out at liquid-helium temperature, 4.2 K, to optimize the energy resolution by reducing the contribution of thermal broadening to the linewidth, and of course lead is superconducting at that temperature.

Normally not one, but several, junctions are fabricated simultaneously, since the likelihood of junction failure always exists. After fabrication the electrical connections are made to the metal electrodes, the junctions are dropped into liquid helium, and measurements are commenced at once.

Although insulators other than aluminum oxide have been tried, aluminum is still used almost universally because it is easy to evaporate and forms a limiting oxide layer of high uniformity. To be restricted therefore to adsorption of molecules on aluminum oxide may seem like a disadvantage of the technique, but aluminum oxide is very important in many technical fields. Many catalysts are supported on alumina in various forms, as are sensors, and in addition the properties of the oxide film on aluminum metal are of the greatest interest in adhesion and protection.

3. Ion Detection

3.1. Secondary Ion Mass Spectrometry

3.1.1. Static Secondary Ion Mass Spectrometry (SSIMS)

SSIMS emerged as a technique of potential importance for surface analysis owing to the work of Benninghoven et al. in Münster [180] during the late 1960s and early 1970s (i.e., at roughly the same time as AES). The prefix "static" was added to distinguish the technique from "dynamic" SIMS, the difference between the two lying in the incident current densities used: on the order of < 1 nA/cm^2 for SSIMS but much higher for dynamic SIMS. With this low primary ion dose, spectral data can be generated on a timescale that is very short compared to the lifetime of the surface layer (typically a surface destruction of < 1 %).

Together with XPS and AES, SSIMS ranks as one of the principal surface analytical techniques. Since its sensitivity for elements greatly

exceeds that of the other two techniques and a large amount of chemical information is available, its use is rapidly expanding in many application fields.

3.1.1.1. Principles

A beam of positive ions bombards a surface, and the resulting interactions cause the emission of a variety of types of secondary particles, including secondary electrons, Auger electrons, photons, neutrals, excited neutrals, positive secondary ions, and negative secondary ions. SSIMS is concerned with the last two of these, *positive* and *negative secondary ions*. The emitted ions are analyzed in a mass spectrometer, which results in positive or negative mass spectra that consist of parent and fragment peaks characteristic of the surface. The peaks may be identifiable as arising from the substrate material itself, from contamination and impurities on the surface, or from deliberately introduced species adsorbed on the surface.

When a heavy energetic particle such as an argon ion (typically 1 – 15 keV) strikes a surface, it will not be stopped short by the first layer of atoms, but continues into the surface until it comes to a halt as a result of energy lost in atomic and electronic scattering. Along its way, the ion displaces some atoms from their normal positions in the solid structure, which as they recoil displace others, which displace additional atoms, and so on, resulting in a complex sequence of collisions. Depending on the energy absorbed in an individual collision, some atoms are permanently displaced from their normal positions, whereas others return elastically after temporary displacement. This sequence is called a collision cascade and is illustrated schematically, along with some other processes, in Figure 31. As the cascade spreads out from the path of the primary ion, its effect may eventually reach atoms in the surface layer, and with enough kinetic energy left, bonds may be broken so that material leaves the surface in form of neutral or ionized atoms or clusters. The resulting secondary particle emission is of low energy (peak of energy distribution between 5 and 10 eV) with over 95 % of the secondary particles originating from the two top monolayers of the solid. Therefore, detecting these ionized particles directly with a mass spectrometer results in very high surface sensitivity.

The theoretical aspect of the collision cascade is given by SIGMUND's collision cascade model [181] which explains much about the parameters involved in the removal of material from a surface by sputtering, although it cannot predict the degree of positive or negative ionization of the material. A unified theory of secondary ion formation in SSIMS does not yet exist, although many models have been proposed for the process. Since the secondary ion yield (i.e., the probabilities of ion formation) can vary by several orders of magnitude for the same element in different matrices, or for different elements across the periodic table, lack of a means of predicting ion yield in a given situation limits both interpretation and quantification. Some theoretical and semitheoretical models have successfully made predictions for a very restricted range of experimental data, they fail when extended further. The most important theoretical models are briefly addressed in Section 3.1.2.1; for further information see, e.g., [182].

3.1.1.2. Instrumentation

The instrumentation for SSIMS can be divided into two parts: (1) the primary ion source in which the primary ions are generated, transported, and focused towards the sample, and (2) the mass analyzer, in which sputtered secondary ions are extracted, mass-separated, and detected.

3.1.1.2.1. Ion Sources

Two types of ion source are particularly suited for SSIMS. The first produces *positive noble gas ions* (usually argon) either by electron impact (EI) or in a plasma created by a discharge (see Fig. 48). The ions are then extracted from the source region, accelerated to the chosen energy, and focused in an electrostatic ion optical column. More recently, it was shown that the use of primary polyatomic ions such as SF_5^+, created in EI sources, could enhance the molecular secondary ion yield by several magnitudes [183], [184]. Cesium ions also are sometimes used to enhance the secondary ion yield of negative elemental ions, as well as that of some polymer

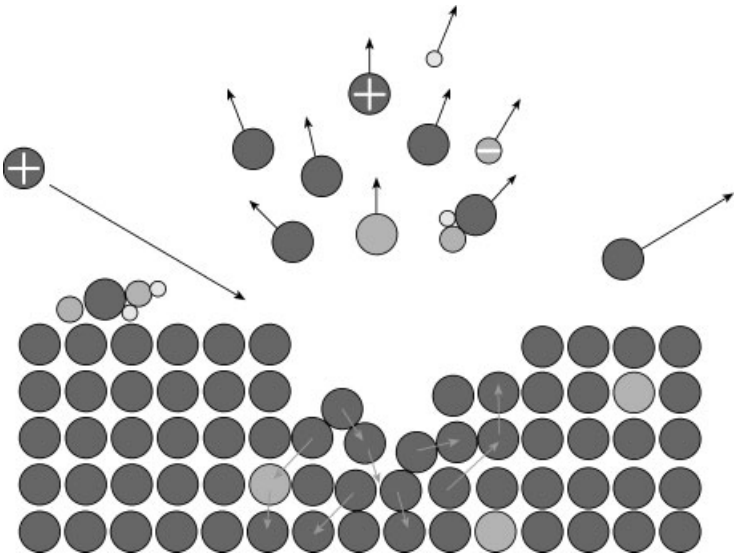

Figure 31. Schematic of the sputtering process

fragments [185]. They are produced by surface ionization with an extraction technique similar to that of EI sources.

Argon ion guns of the electron-impact type operate in the energy range of 0.5 – 10 keV with a 1 – 10 µA dc current, providing optimum spot sizes on the surface of about 10 µm. Discharge-type argon ion guns, often using the duoplasmatron arrangement, operate at up to 15 keV, with currents variable up to 20 µA and, depending on the particular focusing lens system used, providing spot sizes down to about 2 µm.

The second type of ion gun produces *positive ions from a liquid metal* (almost always gallium) in the manner shown schematically in Figure 32 [186]. A fine needle (f; tip radius ≈ 5 µm) of refractory metal passes through a capillary tube (d) into a reservoir of liquid metal (e). The liquid is drawn up through the tube over the needle tip by capillary action. Application of 10 – 30 keV between the needle and a nearby extractor electrode (b) draws the liquid up into a cusp (Taylor cone) through the combined forces of surface tension and electrostatic stress. Ions are formed in the region just above the cusp, which are then accelerated through an aperture in the extractor and focused through an ion optical column onto the sample surface. Gallium is mainly used as the liquid metal because its melting point of 35 °C allows it to remain liquid at room temperature; other metals such as In and Au have been tried but require heating of the reservoir. Because ion production occurs in a very small volume, *gallium liquid metal ion sources* (LMIS) have a very high brightness (ca. 10^{10} A m^{-2} Sr^{-1}). As a result, the ion beam may be focused to a fine spot, resulting in a spot size of 0.2 µm at 8 – 10 keV, which is reduced to about 20 nm at 30 keV.

All ion gun optical columns provide deflection plates for scanning the ion beam over areas adjustable from many square millimeters to a few square micrometers. They have been adapted for pulsing by the introduction of deflection plates which rapidly sweep the beam across an aperture. By applying an ion beam bunching technique, ion pulses of less than 1 ns width can be produced.

3.1.1.2.2. Mass Analyzers

To minimize surface damage, SSIMS mass spectrometers should be as efficient as possible in detecting the total yield of secondary ions from a surface. Also, to be able to separate elemental from molecular species, and molecular species from each other, the mass-resolving power, usually expressed as mass m divided by separable mass Δm should be very high. With this in mind, two types of mass spectrometers have been used: in early work mainly quadrupole mass filters

and, more recently, time-of-flight mass spectrometers.

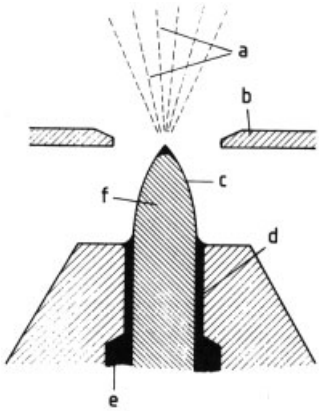

Figure 32. Schematic of the design and operation of a liquid-metal ion source (LMIS) [186]
a) Metal ions; b) Extractor; c) Liquid-metal film; d) Capillary tube; e) Liquid metal; f) Needle

Quadrupole Mass Spectrometer. The quadrupole mass spectrometer has been in use for many years as a residual gas analyzer (with an ionizing hot filament), as well as in desorption studies and SSIMS. It consists of four circular rods, or poles, arranged so that they are equally spaced in a rectangular array and exactly coaxial (Fig. 33 [187]). Two voltages are applied to the rods, a dc voltage U_{dc} and an rf voltage ($U_{rf} \sim U_0 \cos\omega t$). When an ion with a certain mass-to-charge ratio m/q enters the space between the rods, it is accelerated by the electrostatic field, and for a particular combination of dc and rf voltages, the ion has a stable trajectory and passes to a detector. For other combinations of voltages, the trajectory diverges rapidly and the ion is lost either by hitting one of the poles or by passing between them to another part of the system. The mass resolution is governed by the dimensions of the mass spectrometer, the accuracy of its construction, and the stability and reproducibility of the ramped voltage. The quadrupole mass spectrometer is compact, does not require magnets, and is entirely ultrahigh vacuum compatible – hence its popularity. It does have disadvantages, however, in that its transmission (typically $< 1\,\%$) is very low and decreases with increasing mass number. Furthermore, it is a scanning instrument that allows only the sequential transmission of ions, all other ions being discarded. Hence, the information loss is very high.

Time-of-Flight Mass Spectrometer (TOFMS). Because the quadrupole mass spectrometer is unsuitable for the analysis of large molecules on surfaces or of heavy metals and alloys, the TOF mass spectrometer was developed for use in SSIMS by Benninghoven et al. [188]. In a TOF mass analyzer (Fig. 34), all sputtered ions are accelerated to a given potential V in the range of 2 – 8 keV, so that all ions possess the same kinetic energy. The ions are then allowed to drift through a field-free drift path of a given length L before striking the detector. According to the equation $(mL^2)/(2t^2) = qU_o$, light ions travel the fixed distance through the flight tube more rapidly than identically charged heavy ions. Thus, the measurement of the time t of ions with mass-to-charge ratio m/q provides a simple means of mass analysis with $t^2 = (mL^2)/(2qU_o) \propto m/q$. Because a very well defined start time is required for the flight time measurement, the primary ion gun must be operated in a pulsed mode to deliver discrete primary-ion packages [189]. Electric fields (e.g., ion mirrors [189], [190] or electrical sectors [191], [192]) are used in the drift path to compensate for different incident energies and angular distributions of the secondary ions. For good mass resolution ($m/\Delta m \approx 10\,000$), the flight path must be sufficiently long (1 – 1.5 m), and sophisticated high-frequency pulsing and counting systems must be employed to time the flight of the ion to within a nanosecond. A malor advantage of TOFMS is its ability to simultaneously detect of all masses of the same polarity. Raw data acquisition [193] allows the reconstruction of TOF-spectra for any ion species as a function of depth and lateral position.

The original TOF design [188] used pulsed beams of argon ions, but commercial development of LMIS has significantly extended the capabilities of the TOF system. The principle of LMIS operation allows the beam of $^{69}Ga^+$ and $^{71}Ga^+$ ions to be focused to a probe of 50 nm minimum diameter, while being pulsed at frequencies up to 50 kHz and rastered at the same time. Mass filtering selects either of the isotopes before the beam strikes the sample, unless isotopically enriched gallium is used. Operating an LMIS at beam currents of 10 – 100 pA,

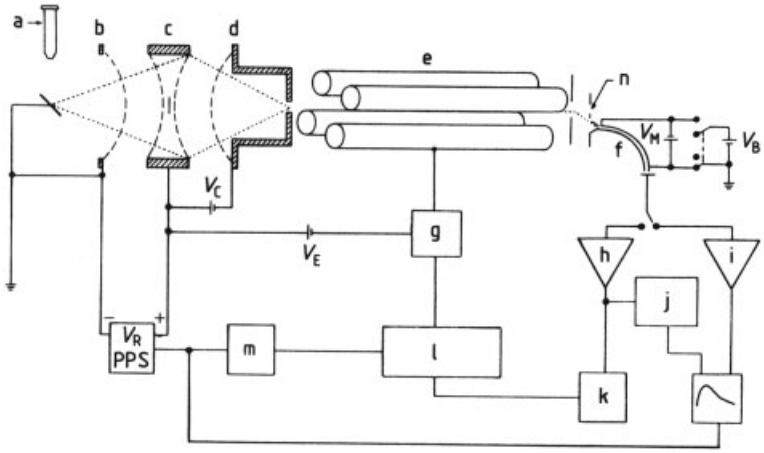

Figure 33. Experimental arrangement for SSIMS used by Krauss and Gruen [187] employing a quadrupole mass spectrometer for mass analysis and a retarding-field analyzer for prior energy selection
a) Ion gun; b) – d) Lenses 1 – 3; e) Quadrupole mass spectrometer; f) Charge detecting electron multiplier; g) Quadrupole power supply; h) Pulse amplifier; i) Iso amplifier; j) Rate meter; k) Multichannel analyzer; l) Mass programmer; m) Sawtooth generator; n) Exit hole for neutrals

opens up the possibility of performing scanning SSIMS and of producing secondary-ion images in selected masses. Pulsing of the ion source is achieved by rapid deflection of the ion beam across a small aperture, the pulse length being variable between 1 and 50 ns. The flight path length is 1.5 m, part of which includes an energy compensator (ion mirror) to ensure that all ions of the same mass, but of different energies, arrive at the same time. Secondary electron detectors (SED) allow topographical images produced by ion-induced secondary electrons to be generated; a low-energy electron flood gun (> 20 eV) can be used for charge compensation when an insulating sample is analyzed; a temperature-controlled sample holder permits SSIMS data to be obtained as a function of surface temperature (temperature-programmed SIMS, TP-SIMS).

With such a TOF-imaging SSIMS instrument, the useful mass range is extended beyond 10 000 amu; the mass resolution $m/\Delta m$ is ca. 10 000 with simultaneous detection of all masses; and within each image, all masses can be detected. The amount of data generated in a short time is enormous, and sophisticated data-acquisition systems are required to handle and process the data.

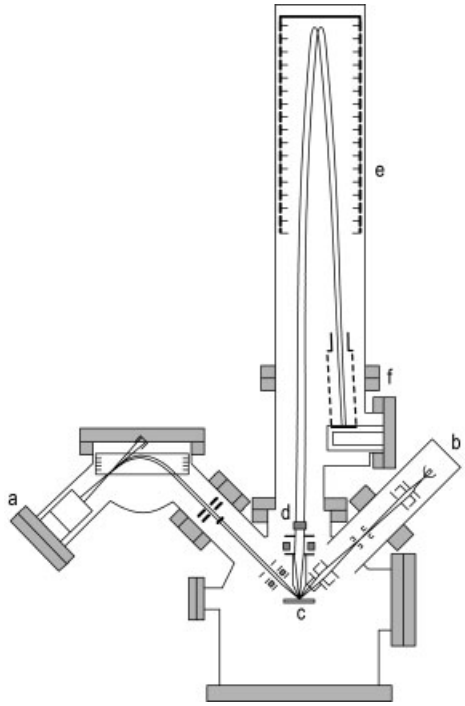

Figure 34. Schematic of an imaging time-of-flight SSIMS system
a) Electron-impact ion source (Ar^+); b) Liquid-metal ion source (Ga^+); c) Coolable and heatable target holder; d)Secondary-ion optics; e) Gridless reflectron; f) Detector

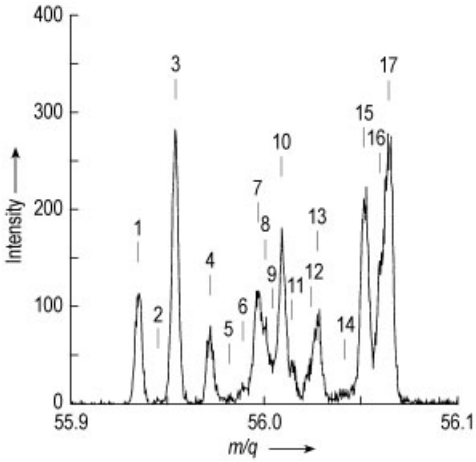

Peak	Ion	m/q
1	Fe	55.9349
2	MnH	55.9459
3	Si_2	55.9539
4	SiCO	55.9718
5	SiN_2	55.9831
6	C_2O_2	55.9898
7	$SiCH_2N$	55.9957
8	$^{29}SiC_2H_3$	56.0000
9	$Si^{13}CCH_3$	56.0038
10	SiC_2H_4	56.0082
11	C_2H_2NO	56.0136
12	$^{13}CC_2H_3O$	56.0217
13	C_3H_4O	56.0262
14	$C_2H_4N_2$	56.0374
15	C_3H_6N	56.0500
16	$^{13}CC_3H_7$	56.0581
17	C_4H_8	56.0626

Figure 35. High-mass-resolution TOF-SIMS spectrum of an Fe-contaminated Si wafer; Mass resolution R ≈ 15 000

3.1.1.3. Quantification

By bombarding a surface consisting of species A with primary ions, the surface coverage of A is reduced. The A particle can be removed by desorption, by fragmentation in the case of molecular species, or by being driven into a deeper layer. The ratio of the number of sputtered particles to the number of primary ions is given by the disappearance yield $Y_{\mathrm{D}}(\mathrm{A})$:

$$Y_{\mathrm{D}}\,(\mathrm{A}) = \frac{\text{No. of sputtered } A}{\text{No. of primary ions}} \tag{3.1}$$

A quantity related to the disappearance yield for a particle of a species covering a solid surface with a surface density $\vartheta(\mathrm{A})$ is the disappearance cross section $\sigma_{\mathrm{D}}(\mathrm{A})$, which corresponds to the average area damaged by one primary ion impact:

$$\sigma_{\mathrm{D}}\,(\mathrm{A}) = \frac{Y_{\mathrm{D}}\,(\mathrm{A})}{\vartheta\,(\mathrm{A})} \tag{3.2}$$

The change in the surface density $\vartheta(\mathrm{A})$ with increasing primary ion fluence F_{PI} is given by:

$$\vartheta\left(\mathrm{A}, F_{\mathrm{PI}} = \vartheta\,(\mathrm{A})_0 \cdot \exp\left(-\sigma_{\mathrm{D}}\,(\mathrm{A})\,F_{\mathrm{PI}}\right)\right) \tag{3.3}$$

Some particles sputtered from the surface are neutral, while others are charged. Molecular particles can be emitted as intact molecules and as fragments. The probability for the desorption of A into the emission channel $\mathrm{X}_i^{\mathrm{q}}$ is given by the transformation probability $P(\mathrm{A} \rightarrow \mathrm{X}_i^{\mathrm{q}})$:

$$P\left(\mathrm{A} \rightarrow \mathrm{X}_i^{\mathrm{q}}\right) = \frac{\text{No. of emitted particles } \mathrm{X}_i^{\mathrm{q}}}{\text{No. of sputtered } A} \tag{3.4}$$

where i is an index distinguishing between the different emission channels, and q is the charge of the particle X_i, where $q = 0, \pm 1, \pm 2, \pm 3, \ldots$ The transformation probability for ionization cannot be easily measured because it strongly depends on matrix and concentration. In fact, it represents the principal obstacle in achieving proper quantification in SSIMS.

The average number of emitted particles $\mathrm{X}_i{}^{\mathrm{q}}$ per incident primary ion is given by the secondary ion yield $Y^{(\mathrm{q})}(\mathrm{X}_i^{\mathrm{q}})$:

$$Y^{(q)}\left(\mathrm{X}_i^{\mathrm{q}}\right) = \frac{\text{No. of emitted particles } \mathrm{X}_i^{\mathrm{q}}}{\text{No. of primary ions}} \tag{3.5}$$

For the detection of sputtered ions, the transmission $T(\mathrm{X}_i^{\mathrm{q}})$ of the mass spectrometer and the detection probability $D(\mathrm{X}_i^{\mathrm{q}})$ must be taken into account. Their product gives the probability of detecting one sputtered particle $\mathrm{X}_i^{\mathrm{q}}$. For a TOF mass spectrometer, $T(\mathrm{X}_i^{\mathrm{q}})$ typically is 10 – 100 %, and $D(\mathrm{X}_i^{\mathrm{q}})$ 10 – 50 %.

The number of detected particles per incident primary ion is given by the detected yield $Y(\mathrm{X}_i^{\mathrm{q}})$:

$$Y\left(\mathrm{X}_i^{\mathrm{q}}\right) = \frac{\text{No. of detected } \mathrm{X}_i^{\mathrm{q}}}{\text{No. of primary ions}} \tag{3.6}$$

An important quantity is the so-called useful yield $Y_{\mathrm{u}}(\mathrm{A}, \mathrm{X}_i^{\mathrm{q}})$:

$$Y_{\mathrm{u}}\left(\mathrm{A}, \mathrm{X}_i^{\mathrm{q}}\right) = \frac{\text{No. of detected } \mathrm{X}_i^{\mathrm{q}}}{\text{No. of sputtered } A} \tag{3.7}$$

The useful yield provides an overall measure of the degree to which the sputtered material is used for analysis. It is a quantity employed to estimate the sensitivity of the mass spectrometric method. Values of $Y_{\mathrm{u}}(\mathrm{A}, \mathrm{X}_i^{\mathrm{q}})$ for elements typically range from 10^{-6} to 10^{-2} in the case of TOF-SIMS. The number of sputtered particles A per incident primary ion (*sputtering yield*) can be measured from elemental and multielemental standards under various operational conditions and can therefore be estimated to reasonable accuracy for the material being analyzed by judicious interpolation between standards.

Since measuring A can be problematic, quantification is normally carried out by relative sensitivity factor (RSF) methods. If a species A on the surface is detected by means of the ion $\mathrm{X}_i^{\mathrm{q}}$, the ratio of the detected ion current $I_{\mathrm{A}}\mathrm{X}_i^{\mathrm{q}})$ to the primary ion current I_{PI} and the surface density $\vartheta(\mathrm{A})$ is called the practical sensitivity factor $S_{\mathrm{p}}[\mathrm{X}_i^{\mathrm{q}}(\mathrm{A})]$:

$$S_{\mathrm{p}}\left[\mathrm{X}_i^{\mathrm{q}}(\mathrm{A})\right] = \frac{I_{\mathrm{A}}\left(\mathrm{X}_i^{\mathrm{q}}\right)}{\vartheta(\mathrm{A})\, I_{\mathrm{PI}}} \tag{3.8}$$

The ratio of two practical sensitivity factors is called the RSF $S[\mathrm{X}_i^{\mathrm{q}}(\mathrm{A}), \mathrm{X}_i^{\mathrm{q}}(\mathrm{B})]$, which is independent of the primary ion current I_{PI}:

$$S\left[\mathrm{X}_i^{\mathrm{q}}(\mathrm{A}), \mathrm{X}_{i'}^{\mathrm{q}'}(\mathrm{B})\right] = \frac{S\left[\mathrm{X}_i^{\mathrm{q}}(\mathrm{A})\right]}{S\left[\mathrm{X}_{i'}^{\mathrm{q}'}(\mathrm{B})\right]} = \frac{I_{\mathrm{A}}\left(\mathrm{X}_i^{\mathrm{q}}\right)}{I_{\mathrm{B}}\left(\mathrm{X}_{i'}^{\mathrm{q}'}\right)} \frac{\vartheta(\mathrm{B})}{\vartheta(\mathrm{A})} \tag{3.9}$$

By using RSFs, it is possible to determine surface densities of other species, if the surface density of the reference species $\vartheta(\mathrm{B})$ is known.

Although the RSFs contain matrix-dependent quantities, their variations are damped to some extent by virtue of taking ratios, and in practice the RSF is assumed constant for low concentrations of A (e.g., < 1 atom %). It can be evaluated from measurements on a well-characterized set of standards containing A in known dilute concentrations. The accuracy of the method, however, is not as high as in Laser-SNMS and XPS.

3.1.1.4. Spectral Information

An SSIMS spectrum, like any other mass spectrum, consists of a series of peaks of varying intensity (i.e., ion current) occurring at certain mass numbers. The masses can be allocated on the basis of atomic or molecular mass-to-charge ratio. In metal and semiconductor surfaces, many of the more prominent secondary ions happen to be singly charged atomic ions, which makes allocation of mass numbers slightly easier. Masses can be identified as arising either from the substrate material itself, from deliberately introduced molecular or other species on the surface, or from contamination and impurities on the surface. Complications in allocation often arise from isotopic effects. Although some elements have only one principal isotope, in many others the natural isotopic abundance can cause difficulties in identification.

Figure 35 shows the positive SSIMS spectrum from a silicon wafer, illustrating both the allocation of peaks and the potential problems with isobars. SSIMS reveals many impurities on the surface, particularly hydrocarbons, for which it is especially sensitive. The spectrum also demonstrates the reduction in isobaric interference through high mass resolution. For reasons discussed in Section 3.1.1.2, the peak heights cannot be taken to be directly proportional to the concentrations on the surface, and standards must be used to quantify trace elements.

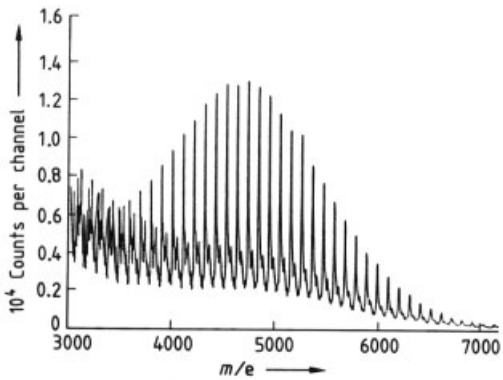

Figure 36. Positive SSIMS spectrum of polystyrene (M_{n} = 5100, $M_{\mathrm{n,\,calc.}}$ = 4500, $M_{\mathrm{n,\,corr.}}$ = 5600) from time-of-flight mass spectrometer [194]

The relationship between what is recorded in an SSIMS spectrum and the chemical state of the surface is not as straightforward as in XPS and AES (see Chap. 2). Because of the large number of cluster ions that appear in any SSIMS spectrum from a multicomponent surface (e.g., while studying a surface reaction), a great deal

of chemical information is available in SSIMS, potentially more than in XPS. The problem in using the information from a cluster ion lies in the uncertainty of knowing whether or not the cluster is representative of a group of the same atoms situated next to each other in the surface. For some materials, such as polymers, the clusters observed are definitely characteristic of the material, as seen in Figure 36 from Bletsos et al. [194], in which the TOF-SSIMS spectrum of polystyrene reveals peaks spaced at regular 104 mass unit intervals, corresponding to the polymer repeat unit. In the characterization of additives in plastic materials, SSIMS measurements have also proved to be a very powerful tool. However, in oxidation studies, for example, many metal – oxygen clusters are generally found in both positive and negative secondary-ion spectra, whose relationship to the actual chemical composition is problematic. Figure 37 from McBreen et al. [195] shows the SSIMS spectra recorded after interaction of oxygen with half a monolayer of potassium adsorbed on a silver substrate. Both the positive and the negative spectrum show evidence of K_iO_j clusters which the authors were unable to relate to the surface condition, although quite clearly much chemical information is available. Sometimes, the use of pure and well-characterized standards can be of help in the interpretation of complex spectra.

In conclusion, SSIMS spectra provide not only evidence of all the elements present, but also detailed insight into the molecular composition. Quasimolecular ions can be desorbed intact up to 15 000 amu, depending on the particular molecule [196] and on whether an effective ionization mechanism is present. Larger molecules can be identified from fragment peak patterns which are characteristic of the particular molecules. If the material being analyzed is completely unknown, spectral interpretation can be accomplished by comparing the major peaks in the spectrum with a library of standard spectra.

3.1.1.5. Applications

In general, SSIMS tends not to be used for quantifications of surface composition because of all the uncertainties described above. Its application has been more qualitative in nature, with emphasis on its advantages of high surface specificity, very high sensitivity for certain elements and molecules, multiplicity of chemical information, and high spatial resolution in the TOF imaging mode. It can be applied to determine the surface composition and existing surface contaminants as a function of lateral position for all kinds of materials such as metals, alloys, semiconductors, oxides, carbides, ceramics, glasses, organic molecular species, organometallics, biomolecules, polymers, etc. It can also be applied to study surface phenomena such as adhesion, chemi- and physisorption, wettability, wear and tear, lubrication, chemical reactivity, corrosion, surface diffusion, segregation, etc. As a common analyzing tool, it is used for numerous applications such as coatings, composite materials, micro- and nanoelectronics, heterogeneous catalysis, biosensors, combinatorial chemistry, pharmaceuticals, medical implants, chromatographies, sensors, painting, environment control and protection, failure analysis, etc.

3.1.1.5.1. Oxide Films

Compared to XPS and AES, the higher surface specificity of SSIMS (1 – 2 monolayers versus 2 – 8 monolayers) can be useful in more precisely establishing the chemistry of an outer surface. Although from details of the O 1s spectrum, XPS could give the information that OH and oxide were present on a surface, and from the C 1s spectrum that hydrocarbons and carbides were present, only SSIMS could say which hydroxide and which hydrocarbons. In the growth of oxide films for various purposes (e.g., passivation or anodization), such information is valuable since it provides a guide to the quality of the film and the nature of the growth process.

One important area in which the properties of very thin films play a crucial role is the adhesion of aluminum to aluminum in the aircraft industry. For bonds of the required strength to form, the aluminum surfaces must be pretreated by anodization, which creates not only the appropriate chemical conditions for adhesion, but also the correct structure in the resultant oxide film as well. Films formed by anodizing aluminum in phosphoric acid under ac conditions were studied by Treverton et al. [197] using

a combination of SSIMS and XPS. Knowledge of both the chemistry of the outer monolayers in terms of aluminum oxides and hydroxides, and of the nature and degree of any contamination introduced by the anodizing process, was important. Figure 38 compares the positive and negative SSIMS spectra after anodizing for 3 and 5 s at 20 and 50 Hz. Both sets of spectra show that the longer anodizing treatment caused significant reduction in hydrocarbon contamination (peaks at 39, 41, 43, 45 amu in the positive spectrum, and 25, 41, 45 amu in the negative), and marked increases in oxide- and hydroxide-associated cluster ions at 44, 61, 70, and 87 amu (positive) and 43 and 59 amu (negative), respectively. At higher masses, peaks allocated to $Al_2O_4H^-$, $Al_2O_3^-$, and $Al_3O_5^-$ were also identified. Residual PO_2^- and PO_3^- from the anodizing bath were reduced as a result of the longer anodizing time. The low level of contaminants is important, since the adhesive coating can then wet the surface more effectively, leading to improved adhesion. In addition, the change from AlO^+ (mass 43) to $AlOH^+$ (mass 44) in aluminum chemistry as a result of going from 3- to 5-s anodizing is a significant effect that could not have been established by XPS.

3.1.1.5.2. Interfaces

The protection of steel surfaces by paint depends significantly on the chemistry of the paint – metal interface. The system has many variables, since the metal surface is usually pretreated in a variety of ways, including galvanizing and phosphating. Indeed, there are likely several interfaces of importance, and corrosion protection may be a function of conditions at all of them.

As one of the final steps in car production, visual checks are performed on the quality of the paint layers, and if defects are observed, paint repair is necessary. Figures 39 and 40 describe the analysis of such a defect (crater in a red paint coat) [198]. The upper sequence of Figure 39 depicts a series of optical images of the defect for different fields of view (the right-hand image is about 150 × 150 µm). Paint remnants can be observed in the center of the crater. The second row of Figure 39 shows ion-induced secondary electron images taken from the crater with a highly focused liquid metal ion beam. The images clearly depict the crater topography.

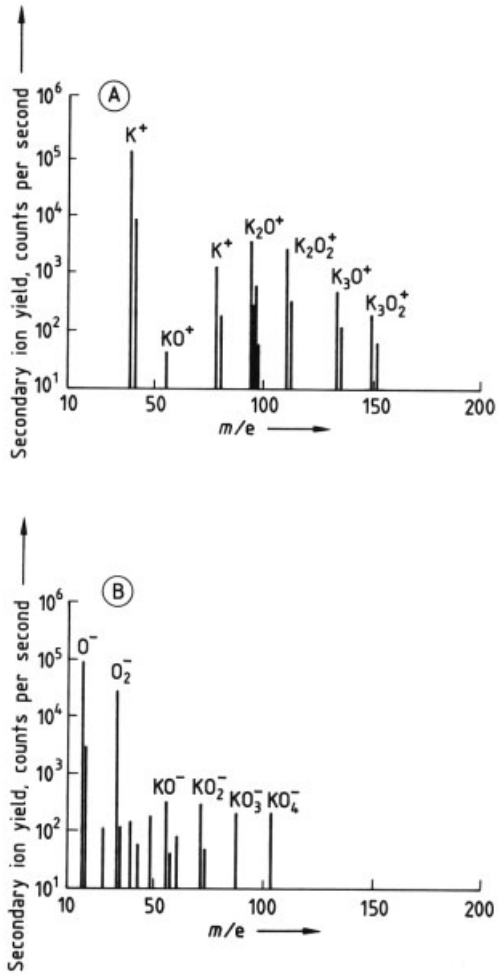

Figure 37. SSIMS spectra following interaction of oxygen with half a monolayer of potassium adsorbed on silver, at 300 K [195]
A) Positive SSIMS spectrum; B) Negative SSIMS spectrum

Figure 40 shows the secondary ion spectra recorded from within the defect and from its surroundings. The spectrum from the defective area clearly shows the presence of a polluting perfluorinated polyether structure (peaks due to C_xFy^+ and the peak at mass 47 amu (CFO^+) are diagnostic of the polyether structure).

The third sequence in Figure 39 shows mass-resolved secondary ion images recorded from the crater. On the far left, the lateral distribution of the uncharacteristic hydrocarbon $C_3H_3^+$ can be seen. The signal is enhanced in the crater

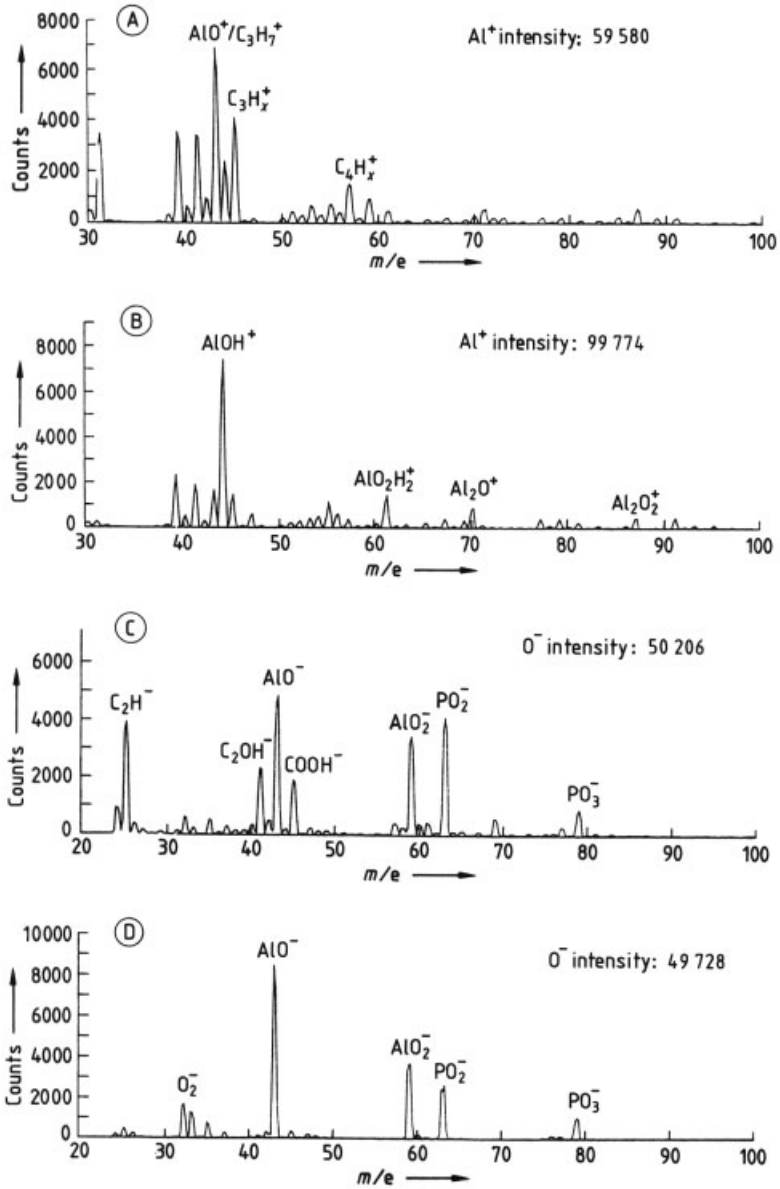

Figure 38. SSIMS spectra from oxide films formed on aluminum after anodization treatment in phosphoric acid for 3 or 5 s [197]
A) Positive SSIMS spectrum, 3 *s*; B) Positive SSIMS spectrum, 5 *s*; C) Negative SSIMS spectrum, 3 *s*; D) Negative SSIMS spectrum, 5 *s*

due to an increased sputter yield. Topographical information can be obtained from hydrocarbon images. The next image shows the lateral distribution of the silicone oil which is incorporated into the paint as a smoothing agent. The silicone oil is only detected from the paint itself and not in the crater bottom. Therefore silicone oil can be excluded as the cause of the defect. The last three images show the lateral distributions of several $C_xF_y^+$ fragments and reveal unambiguously that the perfluorinated polyether is found exclusively in the crater.

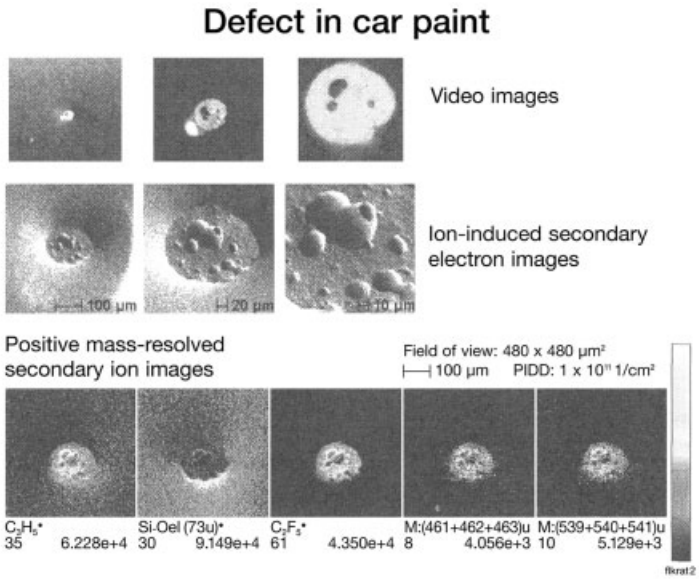

Figure 39. Video images, ion-induced secondary electron images, and mass-resolved ion images of a defect in car paint; white signal corresponds to high intensity, black to low intensity [198]

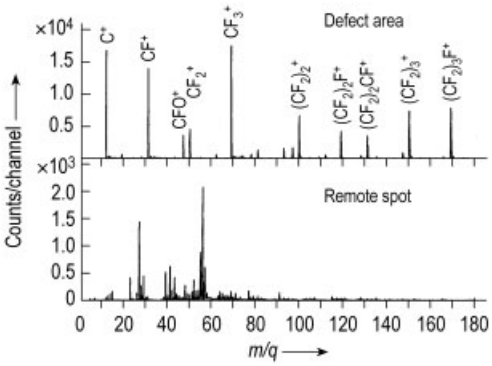

Figure 40. Positive TOF-SIMS spectra from the defective area and from a spot remote from the defect [198]

Later it was found that the polluting lubricant droplets originated from the transport belts used in production; they had fallen into the paint bath and prevented adhesion of the paint to the metal. It can be concluded that especially the high sensitivity of SSIMS in detecting submonolayer coverages of organic species makes it an extremely powerful tool for solving such interface problems.

3.1.1.5.3. Polymers

Treatment of polymer surfaces to improve their properties with respect to wetting or water repulsion and to adhesion is a standard procedure. The treatment is designed to change the chemical properties of the outermost groups in the polymer chain without affecting bulk polymer properties. Any study of the effects of treatment therefore requires a technique that is specific mostly to the outer atomic layers, and hence SSIMS is extensively used in this area.

One of the favored forms of surface treatment is plasma etching. Depending on the conditions of the plasma discharge (i.e., nature of discharge gas, gas pressure, rate of gas flow, discharge power and frequency, and substrate polarity), plasma etching can alter the chemical characteristics of a surface over a wide range. A plasma discharge is by nature highly controllable and reproducible. The effects of plasma treatment on the surface of polytetrafluoroethylene (PTFE) were studied by Morra et al. [199] using combined SSIMS and XPS. They found, as evidenced in Figure 41, that not only did the plasma change the surface chemistry, but the surface emission fragments themselves led to a temporary change in the plasma-discharge conditions. The untreated (Fig. 41 A) and 15-min oxygen-plasma-treated (Fig. 41 C) positive-ion spectra are virtually identical, whereas the 0.5-min oxygen-plasma-treated spectrum (Fig. 41 B) is

quite different. The initial breakup of the PTFE surface produced hydrocarbon fragments which reacted with the oxygen to produce water vapor, thus changing the local nature of the discharge gas. The important deduction to be made is that one of the vital parameters in plasma treatment is the length of contact time; in this case, too long a contact time resulted in no surface changes.

Other SSIMS studies of polymer surfaces have included perfluorinated polyether [200], low-density polyethylene [201], poly(ethylene terephthalate) [202], and the oxidation of polyetheretherketone [203].

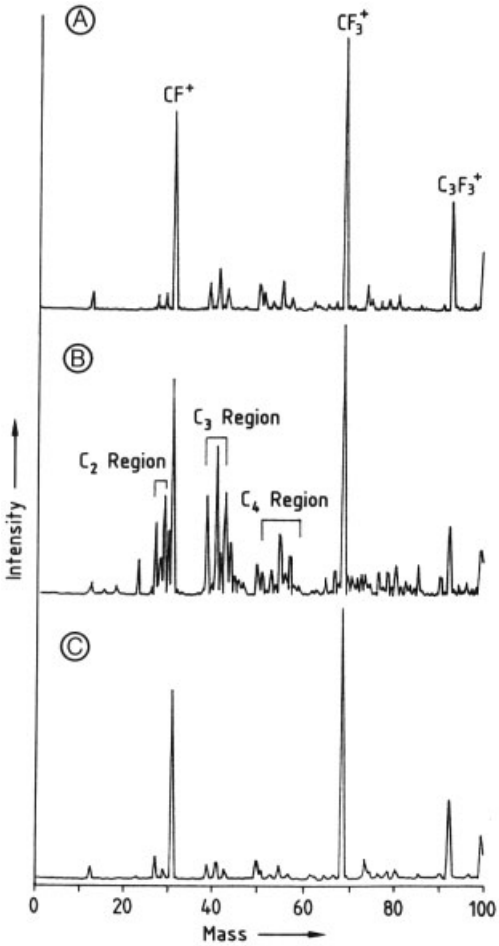

Figure 41. Positive SSIMS spectra from PTFE [199]
A) Untreated; B) 0.5-min oxygen plasma treatment; C) 15 min oxygen plasma treatment

3.1.1.5.4. Biosensors

Biosensor chip technology is a subject of growing interest for a variety of applications. In particular, DNA-sequencing chips based on the method of sequencing by hybridization are growing in importance. Unknown DNA sequences, which are typically labeled with radioactive or fluorescent markers, are hybridized to known complementary short DNA sequences called oligodeoxynucleotides (ODN) that have been immobilized on a solid surface [204], [206]. This method has become valuable for clinical diagnostics, the sequencing of cDNAs or the partial sequencing of clones, DNA and RNA sequencing, gene polymorphism studies, and identification of expressed genes. Using peptide nucleic acid (PNA) hybridization biosensor chips [205], [206] would even allow the sequencing of unlabeled DNA. PNA is a synthetic DNA analog in which both the phosphate and the deoxyribose of the DNA backbone are replaced by polyamides, resulting in a phosphate-free backbone of PNA. This allows the detection of the presence of DNA by a phosphate signal.

Figure 42 depicts some TOF-SIMS spectra obtained from ODN and PNA immobilized on silanized silicon wafers. The spectra clearly demonstrate that the masses corresponding to PO_2^- and PO_3^- provide the best correlation to the presence of ODN, and they can be used to precisely distinguish between ODN and PNA. The CF_3^- and $C_2O_2F_3^-$ peaks in the PNA spectra are due to trifluoroacetic acid, which was part of the PNA solution. Signals for the deprotonated forms $(Cyt - H)^-$ and $(Thy - H)^-$ of the bases cytosine and thymine, respectively are observed for both immobilized PNA and ODN sequences and can be used to detect the presence of these bases.

Figure 43 shows the thermal stability of immobilized ODN and PNA. The signal for the Thy and Cyt bases obtained with temperature-programmed SIMS starts to decrease at about 150 °C for ODN, and at about 200 °C for PNA. This variance is caused by the differences in binding strength between the bases and the sugar – phosphate backbone or the peptide backbone, respectively.

The data show that SSIMS can be used as a tool for characterizing the various steps in the production of biosensors or even for sequencing.

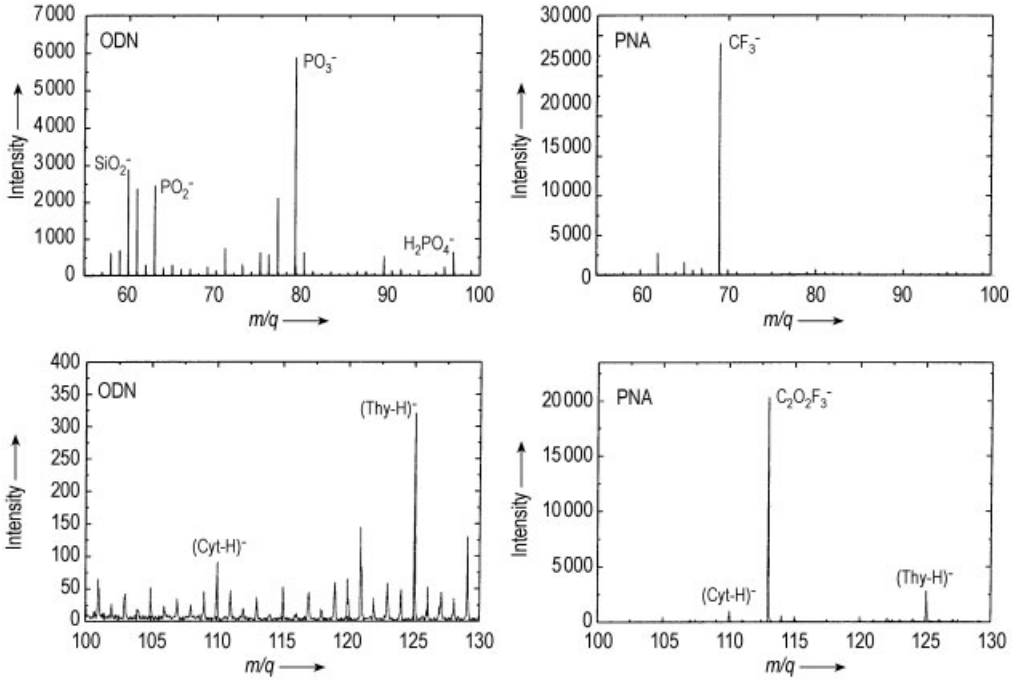

Figure 42. Negative TOF-SIMS spectra obtained from oligodeoxynucleotides ODN (left) and PNA (right) [205]

In a similar way, SSIMS can be applied to solve a variety of problems in bioanalytical chemistry, e.g., screening of combinatorial libraries, characterizing Langmuir – Blodgett layers, etc.

3.1.1.5.5. Surface Reactions

Because of the inherently destructive nature of ion bombardment, the use of SSIMS alone in the study of the reactions of surfaces with gases and vapors must be viewed with caution, but in combination with other surface techniques it can provide valuable additional information. The parallel techniques are most often XPS, TDS, and LEED, and the complementary information required from SSIMS normally refers to the nature of molecules on surfaces and with which, if any, other atoms they are combined.

A typical SSIMS spectrum of an organic molecule adsorbed on a surface is that of thiophene on ruthenium at 95 K, shown in Figure 44, from the study of Cocco and Tatarchuk [207]. The exposure was only 0.5 Langmuir (i.e., 5×10^{-7} torr s = 37 Pa s), and the principal positive ion peaks are those from ruthenium, consisting of a series of seven isotopic peaks around 102 amu. However, ruthenium – thiophene complex fragments are found at ca. 186 and 160 amu, each showing the same complicated isotopic pattern, indicating that even at 95 K interaction between the metal and the thiophene hadruthenium – thiophene complex fragments are found at ca. 186 and 160 amu, each showing the same complicated isotopic pattern, and this indicates that even at 95 K, interaction between the metal and the thiophene occurred. In addition, peaks for thiophene and protonated thiophene are observed at 84 and 85 amu, respectively, with the implication that no dissociation of the thiophene had taken place. The smaller masses are those of hydrocarbon fragments of different chain lengths.

SSIMS has also been used in the study of the adsorption of propene on ruthenium [208], the decomposition of ammonia on silicon [209], and the decomposition of methanethiol on nickel [210].

3.1.1.5.6. Imaging

SSIMS has been applied in the TOF-SSIMS imaging mode to study very thin layers of or-

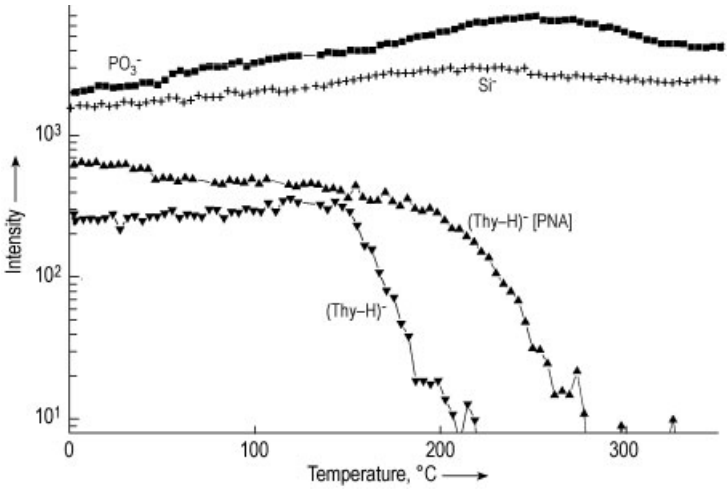

Figure 43. TP-SIMS measurements of an ODN (c = 0.1 mM) and PNA (c = 0.1 mM, curve labeled as PNA) TTTTCCCTCTCTC-sequence [205].

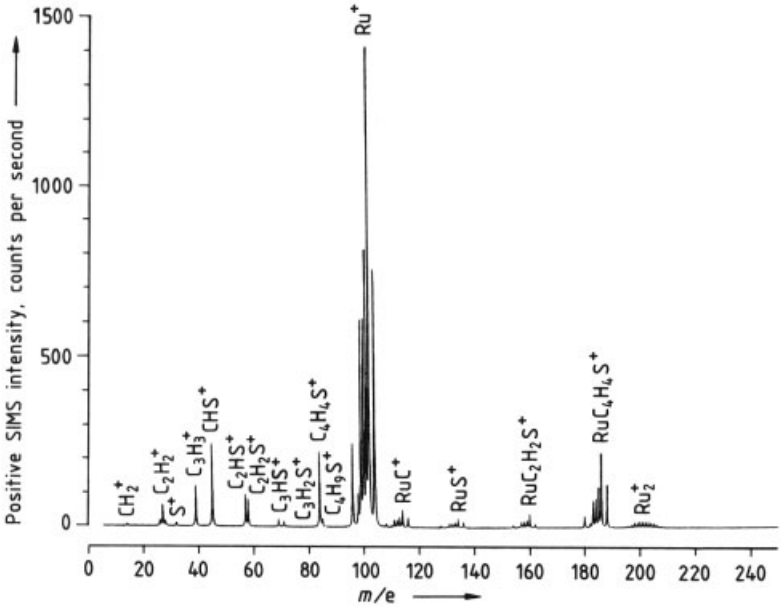

Figure 44. Positive SSIMS spectrum from Ru (0001) after exposure to 0.5 L thiophene at 95 K [207]

ganic materials [211–215], polymeric insulating materials [216], and carbon fiber and composite fracture surfaces [217]. In these studies, a spatial resolution of ca. 80 nm in mass-resolved images was achieved.

Figure 45 shows several mass-resolved images of polymer blends of Nylon 6 (Ny6) and Bisphenol A polycarbonate (PC), mixed for 20 and 40 min respectively [218]. Reactive blending of polymers and exchange reactions that may occur during the melt-mixing process have attracted considerable interest among many researchers. Reactive mixing of immiscible polymers can lead to partial compatibilization of the two components with properties advantageous for technological applications. It is evident that the 20-min sample is quite inhomogeneous, with large distinct domains of different composition,

while the 40-min sample shows a more homogeneous distribution. Detailed spectral analysis of the various imaging regions supported the picture that reactive mixing and compatibilization of Ny6 – PC blends occurs via the incorporation of PC in the nylon phase.

TOF-SIMS imaging can also be used to detect extremely small particles on various substrates such as silicon wafers or photographic films [219]. Figure 46 depicts images of silver bromide and silver chloride crystals, demonstrating the high spatial resolution (50 nm) possible with TOF-SIMS. Current development is aiming to reach a spatial resolution for chemical mapping of about 20 nm.

3.1.1.5.7. Ultrashallow Depth Profiling

TOF-SIMS is well established for surface analysis under static conditions. More recently, TOF-SIMS has also proved to be a very powerful tool for ultrashallow depth profiling, having the advantage of simultaneously detecting all elements of interest. Especially the dual-beam mode [220] (see Section 3.1.2.2.1) enables optimal depth resolution, since sputtering conditions can be independently optimized. Figure 47 depicts an ultrashallow TOF-SIMS depth profile of a 100 eV B implant in Si, capped with 17.3 nm of Si. The measurement was performed with 600 eV SF_5^+ sputtering and with O_2 flooding. The original wafer surface into which the B was implanted is indicated by the maxima of the alkali metal and C signals. Due to these contaminants, the ^{30}Si signal exhibits a minimum. The B profile shows a dynamic range of more than 3.5 decades and a depth resolution of ≤ 0.5 nm.

3.1.2. Dynamic SIMS

Dynamic SIMS is a standard technique for the measurement of trace elements in semiconductors, high-performance materials, coatings, and minerals. The main advantages of the method are the excellent detection limit of less than 1 ppm for all elements, the isotopic sensitivity, the inherent possibility of measuring depth profiles, and the capability for fast direct imaging and 3D detection of species distribution.

The first SIMS instrument was built by Herzog and Viehboeck [221] in Vienna in the 1940s. In the early 1960s, Herzog and Liebl [222] built the first sophisticated SIMS instrument, nearly at the same time as Castaing and Slodzian in Paris [223]. In 1970, Benninghoven was the first to use the acronym SIMS [224].

3.1.2.1. Principles

The bombardment of a sample with a dose of high-energy primary ions (1 – 20 keV) results in the destruction of the initial surface and near-surface regions (see Section 3.1.1.1). If the primary ion dose is higher than 10^{11} ions/mm^2 the assumption of an initial, intact surface is no longer true. A sputter equilibrium is reached at a depth greater than the implantation depth of the primary ions. The permanent bombardment of the sample with primary ions leads to the following effects.

Compensation of Preferential Sputtering. The species with the lower sputter yield is enriched at the surface. This effect is called preferential sputtering and complicates, for example, Auger measurements. The enrichment compensates for the different sputter yields of the compound or alloy elements; thus in dynamic SIMS the flux of sputtered atoms has the same composition as the sample.

Atomic Mixing. Depending on their mass, energy, and impact angle the primary ions reach a mean depth until they are finally stopped by many collisions with sample atoms. Also, sample atoms are moved from their initial locations (see Section 3.1.1.1). This results in atomic mixing of the near-surface region and limits the depth resolution. Small primary ion energies and oblique bombardment angles reduce this effect.

Implantation of Primary Ions. The primary ions are implanted in the sample and thereby influence the chemical constitution. For energies on the order of 20 keV, the implantation depth is about 30 nm. The sputter yield, i.e., the ratio of secondary to primary particles (not only ions), is energy dependent and has a maximum around 10 keV.

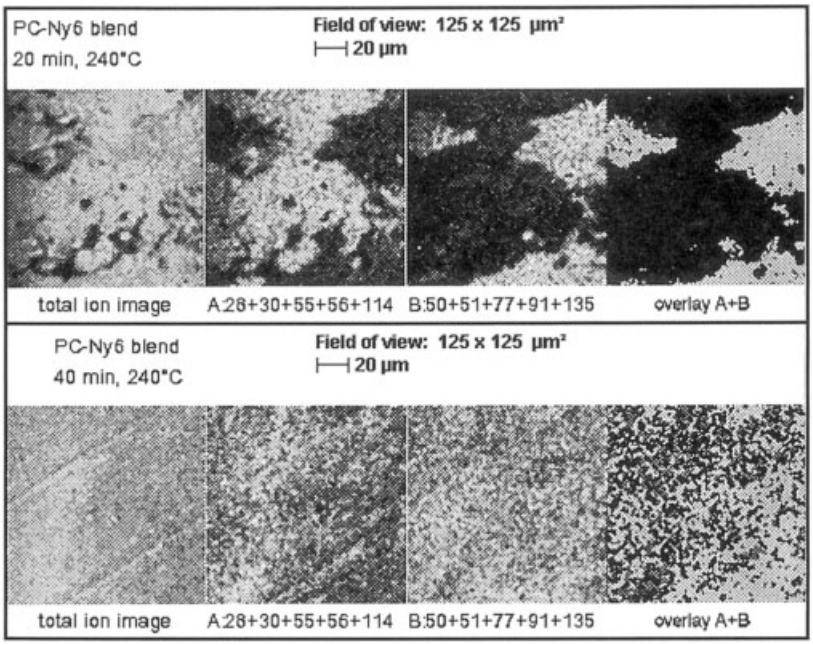

Figure 45. Secondary ion images of PC – Ny6 blends melt-mixed for 20 min (top) and 40 min (bottom). Images A and B were obtained by adding the mass-resolved images pertaining to the masses listed below each image. Also shown are the overlays of A (black) and B (white) [218].

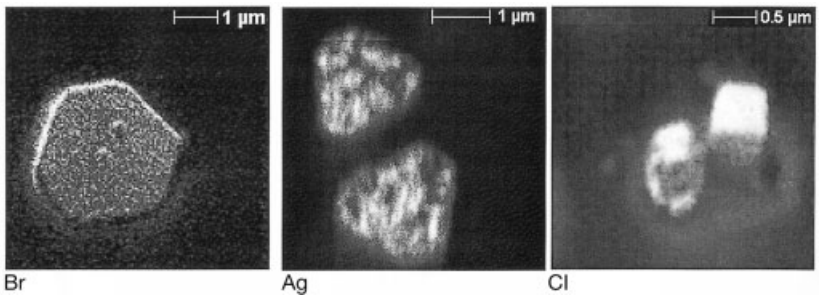

Figure 46. High-resolution TOF-SIMS images of silver bromide and silver chloride crystals

Crater-Bottom Roughening. Depth resolution is also limited by roughening of the crater bottom under ion bombardment. On polycrystalline samples this can be due to different sputter yields of different crystal orientations, since the sputter yields of single crystals can vary by a factor of two depending on their orientation. Due to this type of roughening, depth resolution deteriorates with increasing depth.

Sputter-Induced Roughness. Sputtering single crystal semiconductors, in particular II-VI semiconductors, with Cs^+ ions results in surface roughening. Until now there is no satisfactory explanation of this effect. This effect and the corresponding reduction of the depth resolution can be avoided by rotating the sample during the measurement.

Charging Effects. The electrostatic potential of nonconducting samples and layers is changed due to the bombardment of charged atoms and the emission of secondary electrons and ions. Usually this results in a positive charging of the sample and a corresponding drift of the secondary ions' energy. On good insulators this

charging reaches such high values that the primary ions are repelled by the surface, and analysis becomes impossible. To compensate for this effect, low-energy electrons can be directed to the surface ("electron shower").

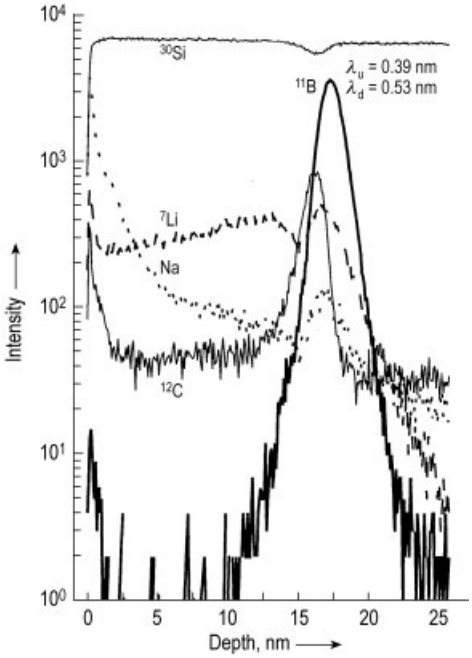

Figure 47. Multi-element depth profile of a B layer in Si performed with 600 eV SF_5^+ sputtering [220]

Secondary Ion Formation. Due to the complex situation on the surface, a satisfactory theoretical description of the ionization process is not yet possible. Different ionization mechanisms have been proposed.

Electron-Tunneling Model. Several models based on a quantum mechanical concept have been introduced. One idea is that one electron of the conducting band tunnels to the leaving atom or vice versa. The tunneling probability depends on the ionization potential of the sputtered element, the velocity of the atom (available time for the tunneling process), and the work function of the metal (adiabatic surface ionization, Schroeer model [225]).

Broken-bond models were especially developed to describe the ionization process of ionic compounds, especially under primary oxygen bombardment or gas admission ("O_2 shower"). The model assumes the building of an oxide layer on the surface. The binding electrons remain at the oxygen atom, and therefore positively ionized atoms are emitted [226].

Local Thermodynamic Equilibrium (LTE). The LTE model has only historical importance. The idea was that under ion bombardment a near-surface plasma is generated in which the sputtered atoms are ionized [227]. The plasma should be under local equilibrium, so that the Eggert – Saha equation for determination of the ionization probability can be used. The important parameter was the plasma temperature; given the concentration of one element, this temperature was determined. The theoretical background of the model is not applicable. The reason why it gives semi-quantitative results is that the exponential term of the Eggert – Saha equation also fits to quantum mechanical expressions.

3.1.2.2. Instrumentation

The basic instrumental setup for dynamic SIMS is the same as for SSIMS (Section 3.1.1). Depending on the required intensity, beam diameter, and ion species, different ion sources are used. Several mass analyzers with different characteristics enable a broad range of applications.

3.1.2.2.1. Ion Sources

Bombardment with reactive species increases the ionization yield and therefore SIMS sensitivity. Electropositive elements are more sensitive under oxygen bombardment, and electronegative elements under cesium bombardment (see Section 3.1.2.3). Therefore in dynamic SIMS usually Cs^+ or O_2^+ ions are used for sample sputtering. Because of the lower melting point of gallium and hence the higher achievable brightness of a Ga^+ source, this type is also used for fine-focus imaging applications, despite its lower sensitivity. The liquid-metal ion source (LMIS) described in Section 3.1.1.1.1 can also be run with Cs^+; O_2^+ beams are produced in electron impact/extraction or duoplasmatron sources.

Duoplasmatron (Figure 48). In the Duoplasmatron, gas-discharge ion sources are used for bombardment with oxygen or argon. In dynamic SIMS, especially the use of O_2^+ ions is common because of the chemical enhancement effect. With a Duoplasmatron, ion beam currents

of several microampere can be generated. The diameter of the beam can be focused down to 0.5 µm. The attainable particle current density of the beam (ions per square micrometer per second)) is an inherent property of the ion source, the so-called brightness. Hence a decrease in the beam diameter is associated with a reduction of the beam current.

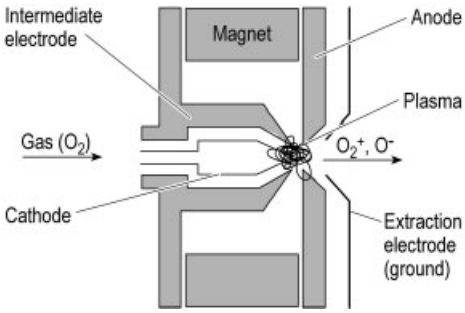

Figure 48. Schematic of a Duoplasmatron ion source

Dual-Beam Technique for TOF Instruments. Because of their low duty cycle, TOF instruments are not recommended for common depth profiles. If only one fine-focus pulsed ion gun is used, the sample is bombarded by primary ions for only 0.1 % of the time; therefore, it would take a long time to reach deeper regions. However, by using a second ion gun working anticyclic to the "measuring" beam without an applied accelerating potential, substantial amounts of sample material can be removed. One advantage of this dual-beam technique is that sputtering for measurement and sputtering for material removal are not coupled, and so beam diameter, energy, mass, and angle of the "removal" gun can be optimized independently of the "measuring" gun. A disadvantage is that atoms sputtered during the material-removal phase are not analyzed. This reduces the detection power. Simultaneously detecting all masses can only partially compensate for this disadvantage, especially in implantation profiles where only one mass is of interest.

3.1.2.2.2. Mass Analyzers

Apart from the quadrupole and TOF analyzers described in Section 3.1.1.1.2, the most important types of mass analyzers used in common dynamic SIMS instruments employ a magnetic sector field.

Magnetic Sector Field. In a magnetic field $\boldsymbol{B}$ an ion with velocity $\boldsymbol{v}$ and charge q experiences a centripetal force $\boldsymbol{F}$, known as the Lorentz force:

$$\boldsymbol{F} = q \cdot (\boldsymbol{v} \times \boldsymbol{B}) \tag{3.10}$$

This force must be balanced by the centrifugal force of the ion mv^2/r, where r is the radius of curvature of the trajectory. Hence

$$\frac{mv^2}{r} = q \cdot (\boldsymbol{v} \times \boldsymbol{B}) \tag{3.11}$$

Due to this force, particles with an energy $E = mv^2/2$ are held on circular paths of radius r:

$$r = \sqrt{\frac{2E\,(m/q)}{\boldsymbol{B}}} \tag{3.12}$$

Secondary ions are accelerated from the surface to a certain energy. After passing an entrance slit, different masses are forced to follow different radial paths in the homogeneous magnetic field. Only mass m ($\pm\,\Delta m$), defined by

$$\frac{m}{q} = \frac{\boldsymbol{B}^2 r^2}{2E} \tag{3.13}$$

is able to pass the exit slit and reach the detector. Since the formation probability of singly ionized atoms is considerably higher than that of more highly charged particles, q in Equation (3.4) can normally be set to unity. By scanning the magnetic field, all masses are detected sequentially.

Two types of commercial magnetic sector field instruments will be described in the following:

The *direct-imaging magnetic sector mass analyzer* (Fig. 49) has the unique property that all parts (lenses, ESA, and magnetic sector field) of the secondary ion optic are stigmatic (comparable to light microscopes). This means that all points of the surface are simultaneously projected into the analyzer.

The *Mattauch–Herzog geometry* (Fig. 50) allows the detection of several masses simultaneously and is therefore ideal for scanning instruments [228]. Up to five detectors are adjusted mechanically to locations in the detection plane, that is, to masses of interest. Thus, it is possible to detect, for example, all isotopes of an element simultaneously in a certain mass range. Also fast, sensitive, and precise measurements

of distributions of different isotopes are feasible, and this allows of isotope ratios of small particles visible in the image to be calculated. The only commercial instrument of this type (Cameca Nanosims 50) uses a coaxial optical design of the ion gun and the secondary ion extraction and provide a lateral resolution of 50 nm.

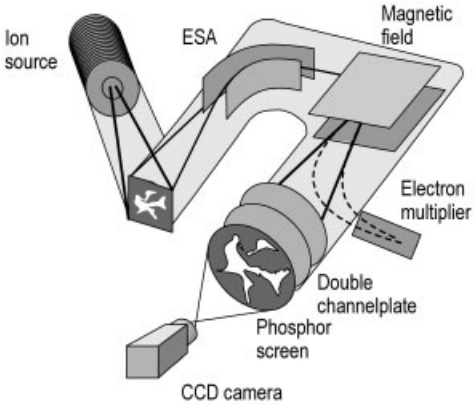

Figure 49. Basic set-up of a direct-imaging magnetic sector instrument. The stigmatic secondary ion optics consists of an electrostatic analyzer (ESA) and a magnetic sector field.

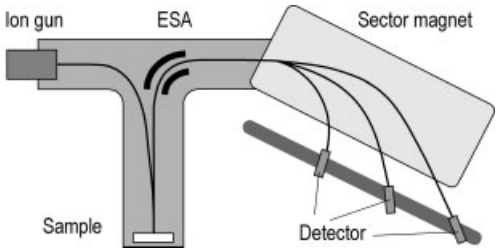

Figure 50. Schematic drawing of a Mattauch – Herzog magnetic sector instrument with simultaneous mass detection

3.1.2.2.3. Detectors

Ion intensities up to a count rate of two million are measured with an electron multiplier. Above this value a Faraday cup is used. Its ion – current amplification must be adjusted to fit the response of the electron multiplier.

For the measurement of local ion intensities in the direct-imaging mode (see Fig. 49), amplification ensuring laterally uniform single-ion detection is necessary. Depending on the sensitivity of the detector, single or double channel plates are used. Two laterally resolving detectors are in use: ccd cameras and resistive anode encoders (RAEs).

For standard *CCD cameras*, a double channel plate (amplification $> 10^6$) is necessary. For high-sensitivity cameras (sensitivity $> 10^{-3}$ lux, cooled or with internal amplification) a single channel plate suffices. By controlling the channel plate voltage (i.e., amplification) one can achieve a high dynamic range with this system [229].

The *resistive anode encoder* (RAE) has the advantage that the single ion events are detected digitally; therefore, it delivers quantitative results, independent of local differences in the amplification of the channel plate. One disadvantage is that the count rate is limited to 200 000.

3.1.2.3. Spectral Information

The element sensitivity is determined by the ionization probability of the sputtered atoms. This probability is influenced by the chemical state of the surface. As mentioned above, Cs^+ or O_2^+ ions are used for sample bombardment in dynamic SIMS, because they increase the ionization probability – the so-called chemical enhancement effect.

The near-surface region is partially oxidized in the case of O_2^+ bombardment. During the sputtering process the chemical bonding of the oxides is broken. The binding electrons have a higher affinity to oxygen; therefore, the (metallic) binding partner leaves the surface positively charged.

Bombardment with Cs^+ results in coverage of the surface with Cs. This reduces the work function and results in a higher probability that an electron from the conducting band tunnels from the surface to the leaving atom during the sputtering process. This effect results in a higher rate of formation of negatively charged atoms, especially of electronegative elements.

Both models have been described theoretically and verified by experiments. Due to the complex surface structures of real samples, there is no quantification algorithm based on physical models yet.

3.1.2.4. Quantification

Since the ionization probability is strongly dependent on the matrix, it is necessary to use standards.

Relative Sensitivity Factor. The most accurate and therefore most popular quantitative method to reduce these influences is to analyze the unknown sample together with a similar sample of known content. Generally, the relationship between measured intensity and content of each sample is defined via the *relative sensitivity factor* (RSF):

$$\frac{I_{\mathrm{el}}}{I_{\mathrm{ref}}} = \mathrm{RSF} \cdot \frac{c_{\mathrm{el}}}{c_{\mathrm{ref}}} \tag{3.14}$$

where ref specifies a matrix element, and el indicates the analyzed element. By using a standard sample with known concentration, the RSF can be made adjustable. By using standards relative precision up to 1 % is possible.

Implantion Standards. The production of homogeneous solid-state standards is costly. Dynamic SIMS has the advantage that nonhomogeneous *ion-implantation standards* can also be used. Given the implantation dose, the RSF can be calculated by summing the signal during a depth profile. Therefore, calculations must be performed with the mean intensity ratio indicated by the sum in Equation (3.15) to obtain the respective RSF.

$$\mathrm{RSF} = \frac{f_{\mathrm{ref}} c_{\mathrm{ref}} \varrho N_{\mathrm{A}} d}{f_{\mathrm{el}} Q_{\mathrm{T}} M C} \cdot \sum_{i=1}^{C} \frac{I_{\mathrm{el}}(i)}{I_{\mathrm{ref}}(i)} \tag{3.15}$$

where f_{ref} is the isotope abundance of the reference, c_{ref} the concentration of the reference in atoms/atom, ϱ the matrix density in g/cm^3, N_{A} the Avogadro number (6.022×10^{23} mol^{-1}), d the depth of the crater in cm, f_{el} the isotope abundance of the element, Q_{T} the implanted dose of the element in atoms/cm^2, M the molecular mass of the matrix, C the number of measured cycles, $I_{\mathrm{el}}(i)$ the signal intensity of the element in counts/s of cycle i, and $I_{\mathrm{ref}}(i)$ the signal intensity of the reference in counts/s of cycle i.

For quantification of multilayers, each layer must be calculated separately. Especially for oxide layers, quantification is not possible due to the variation in the ionization probability in the different layers. Modern materials have a complex three-dimensional inner structure with many different phases. For these samples quantification is not possible, but technologists are often only interested in relative differences between several samples, or they already know the bulk concentrations and are only interested in the elemental distribution.

MCs$^+$ Ions. With Cs bombardment, the matrix effect can be significantly reduced by using the *MCs$^+$ ion signals* for quantification. The detection limit is reduced by two or more orders of magnitude, and in some cases even standard-free quantification is reported [230]. One disadvantage is the high masses of the MCs$^+$ ions, because many mass interferences occur in this mass range.

3.1.2.5. Mass Spectra

Magnetic sector field instruments have mass resolutions up to $m/\Delta m = 20\,000$, quadrupole instruments are limited to a mass resolution of approximately $m/\Delta m = 500$. Both types of instruments provide a mass range up to 500. High mass resolution is needed to separate mass interferences of molecular and atomic ions. Due to the mass defect of the binding energy of the nucleus, atomic ions have a slightly smaller mass than corresponding molecular ions. Typically, mass resolutions between 5000 and 10 000 are necessary.

There is a second method to avoid mass interferences. Molecular ions have a steeper energy distribution than atomic ions, that is, lower abundance at higher energies. By energy filtering, that is gating of low-energy ions, the molecular ions are suppressed. This method also results in a reduction of the detection limit. Figure 51 shows that energy filtering hardly reduces the signal intensities of the atomic ions but strongly suppresses molecular ions. The ^{63}Cu signal is not visible without offset because there are many interferences, e.g., ^{51}V^{12}C.

3.1.2.6. Depth Profiles

The measurement of depth profiles is based on the detection of the masses of interest during sputter removal of the sample material. The evaluation of such an experiment has the following limitations:

– The depth resolution is limited by atomic mixing, roughness of the initial sample surface, and roughening processes during sputtering.

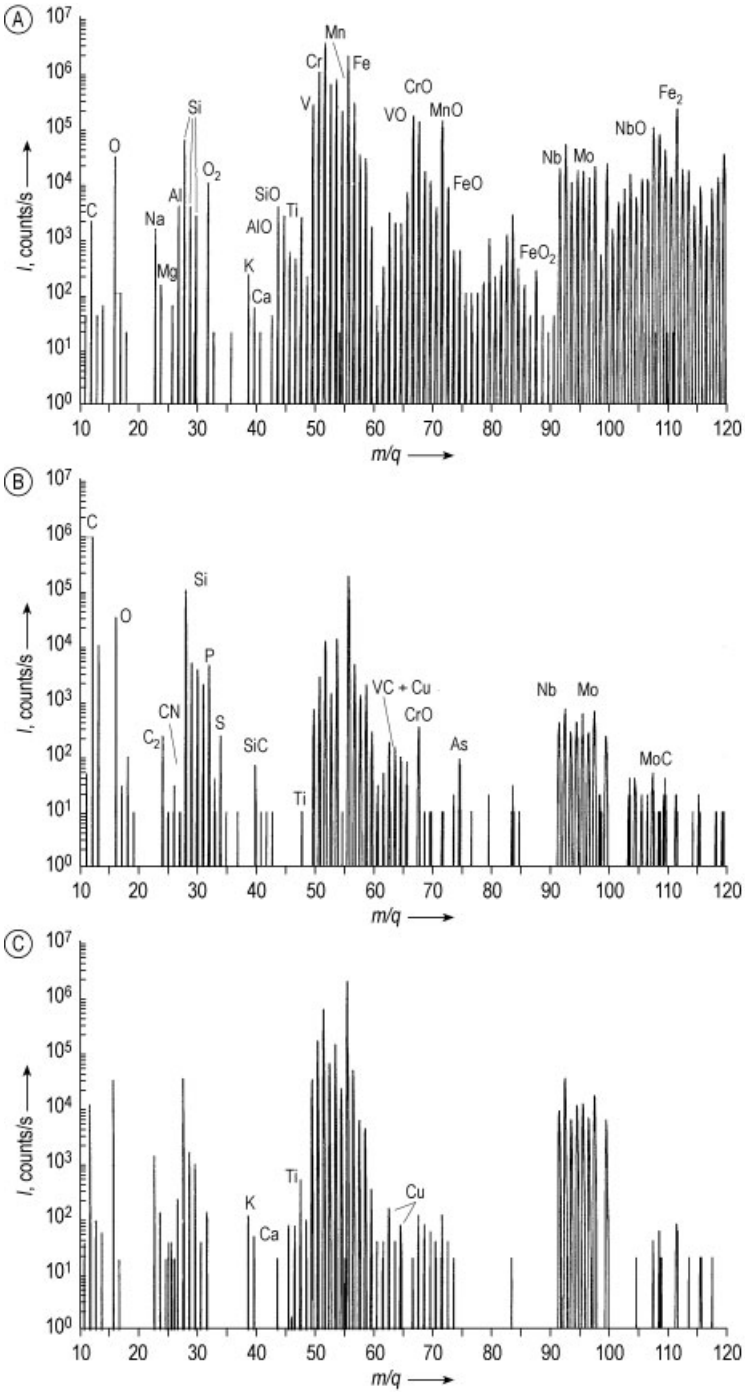

Figure 51. SIMS spectra of a high-speed steel. A) Primary ions: O_2^+, no secondary ion energy limitation, electropositive elements are sensitive, many molecule ions are visible. B) Same parameters, but 300 V offset was used; the molecular ion intensities are reduced significantly. C) Primary ions: Cs^+, 300 V offset was used; therefore, electronegative elements are more sensitive.

– The determination of the depth scale by the measurement of the crater depth may be hampered by the fact that multilayer systems consist of layers with different sputter yields.
– Furthermore, the different chemical constitution of the layers may cause different RSFs for a given element in different layers. These effects may make the determination of the depth scale and the RSF for each layer necessary.

SIMS depth profiles have no significant background; only the elements H, C, and O have a background due to residual gas. This background is only dependent on the gas pressure in the sample chamber, and not on the primary-ion beam intensity. To test the influence of the signal background due to the residual gas, the ion beam current can be reduced for several cycles. All signals should be reduced in the same way; if one signal is not reduced this indicates that it is caused by residual gas.

To illustrate the use of the isotope sensitivity of SIMS, Figure 52 shows a depth profile of a passivation layer on high-purity chromium. The layer was produced by a two-step process: first the sample was oxidized for 30 min in air with natural isotope concentrations. In the second step, the gas was changed to a mixture of ^{18}O and ^{15}N. The profile shows that the ^{18}O layer is located on the top; this indicates that the chromium diffuses to the surface and reacts there with the oxygen. The high dynamic range of SIMS forces the use of logarithmic scales over 6 or 7 orders of magnitude. This has to be kept in mind when interpreting depth profiles; for example, diffusion profiles appear as straight lines. The diffusion of oxygen through the layer to the interface with the metallic chromium is slower by a factor of 50. The enrichment of carbon at the interfaces is due to gas change.

3.1.2.7. Imaging

Scanning SIMS. In the scanning (or microprobe) mode the image is measured sequentially point-by-point. The lateral resolution of the element mapping in scanning SIMS is only dependent on the primary beam diameter. Therefore, mostly LMISs are used. Beam diameters down to 50 nm with high total intensities of 1 nA can be achieved. The beam diameter of oxygen is limited to 0.5 µm by the used Duoplasmatron; for mapping electropositive elements, this necessary because of the chemical enhancement effect.

The counting time per pixel and the number of pixels determine the measurement time of an image. In the case of low concentration of the elements of interest the Poisson statistics of the small integer values of the secondary ions determine the counting time.

A reduction in the measurement time of element distributions is made possible by the simultaneous detection of several masses. This can be achieved only by the use of a magnetic sector field spectrometer with Mattauch – Herzog geometry [228] (Fig. 50) and parallel detection of up to five masses by mechanically adjusted electron multipliers.

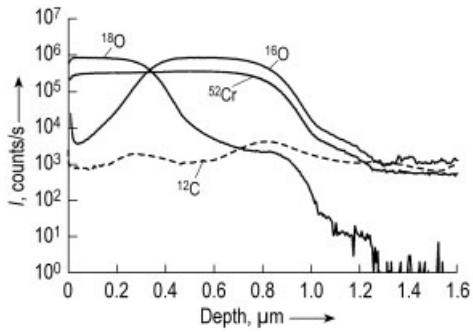

Figure 52. Depth profile of a passivation layer on high-purity chromium. The ^{18}O layer is on the top, the ^{16}O layer at the interface with the metal.

Direct-Imaging Mode. In the direct-imaging (or microscope) mode (see Figs. 53 and 54) all pixels of the element distribution are detected simultaneously by using stigmatic secondary ion optics. The lateral resolution of the image is not influenced by the diameter of the primary ion beam. The resolution is determined by the energy distribution of the secondary ions. The energy and angle distributions result in a chromatic aberration of the first electrostatic lens, the emission lens [185]. The resolution can only be increased by limiting the energy (or angle) acceptance of the instrument; this, of course, reduces the transmission and therefore the detection limit. In practice this effect limits the lateral resolution to about 1 µm; only for main components is 0.5 µm possible.

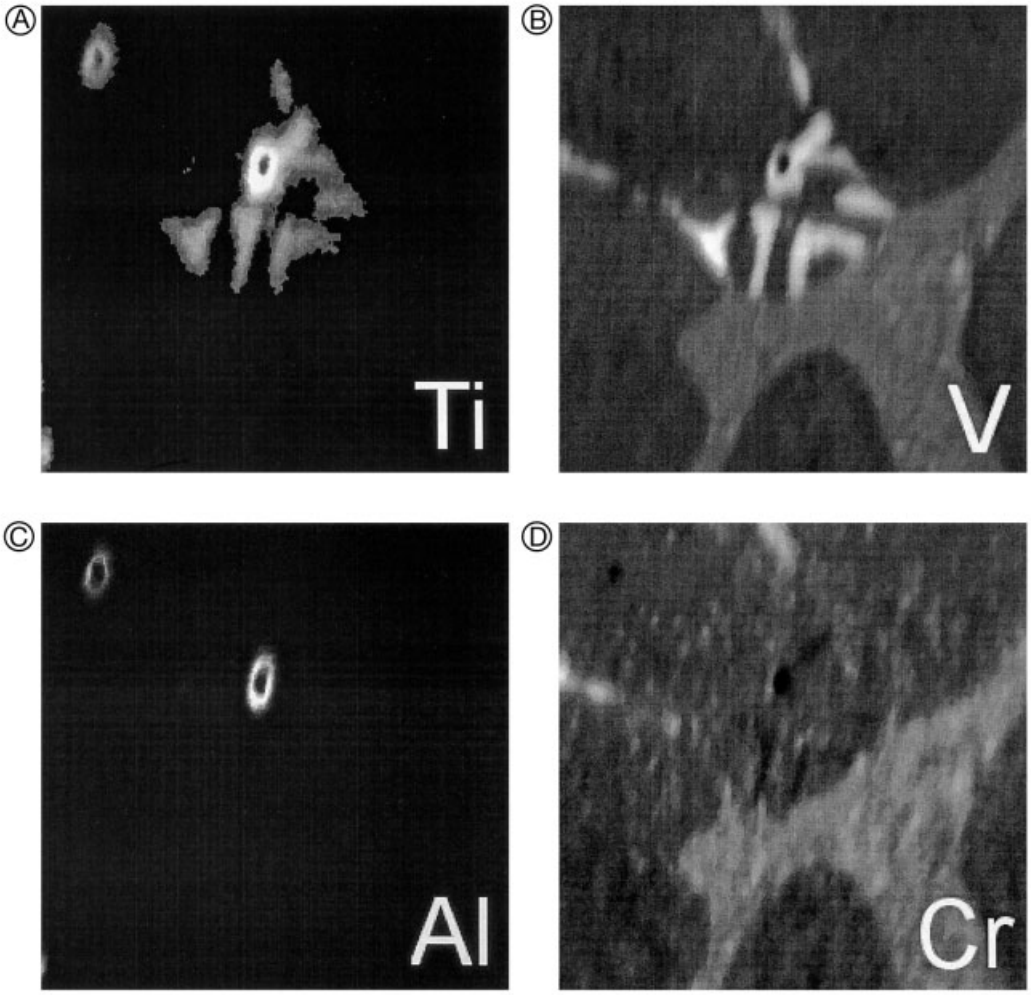

Figure 53. Scanning SIMS image of the carbide structure of a high-speed steel [231]. In the V distribution different phases (MC and M_2C) are visible. The Al distribution shows the shell structure of the NMI (non-metallic impurities) acting as condensation nucleus. Measurement time: 10 min, Image size: 64 x 64 µm. Primary ions: O_2^+, Primary energy: 5.5 keV.

In practice the quality of the image is also reduced by the use of high mass resolution and energy offset, and in many cases it is therefore not possible to avoid mass interferences. By the use of image processing tools for classification of mappings of different masses, determination of element distributions is possible [232].

For secondary-ion detection a laterally resolving detector is necessary. In the first step a channel plate for amplification is used, and the secondary electrons of the output of this device are accelerated either to a fluorescent screen or to a resistive anode. In the case of a fluorescent screen, the image is picked up by a CCD camera and summed frame by frame by a computer. This system has the advantage of unlimited secondary-ion intensities, but compared to the digital detection of the resistive anode decoder, the lateral and intensity linearity is not so well defined.

The advantage of the imaging mode is fast data acquisition. Because all pixels are projected and detected simultaneously, the measurement time of one distribution is extremely low.

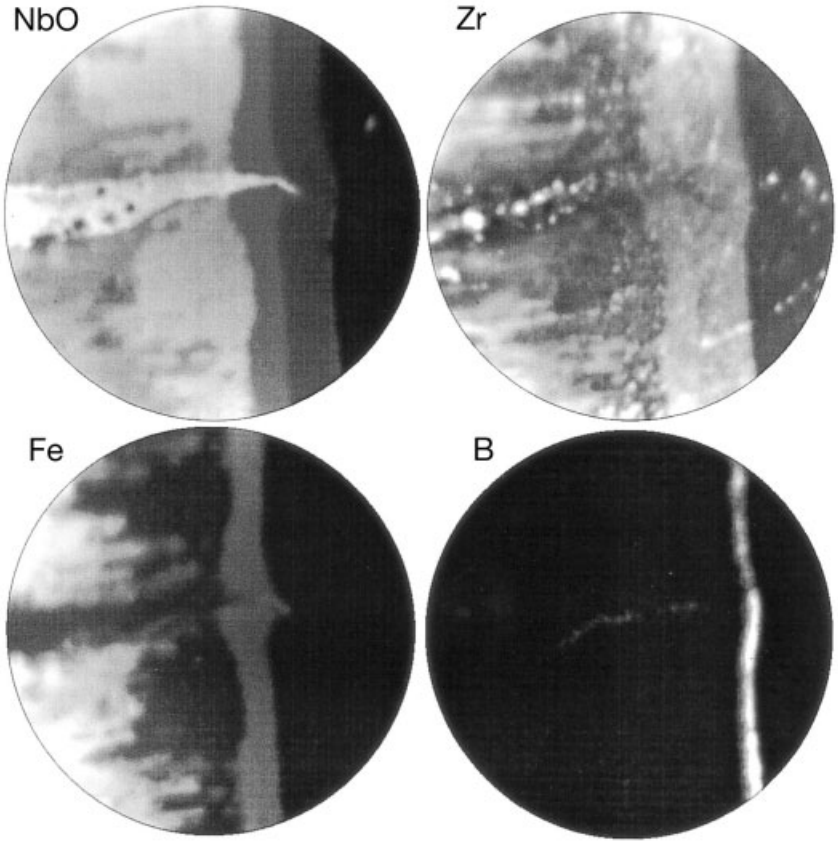

Figure 54. Direct-imaging mode SIMS image of a passivation layer on a niobium alloy [233]. The boron enrichment at the interface is not visible by EPMA. Measurement time: 10 s, image diameter: 150 µm, primary ions: O_2^+, primary energy: 5.5 keV.

3D-SIMS (Fig. 55) is a further development combining depth analysis and imaging analysis. In the imaging mode the summing time of single ion distributions is low (1 – 5 s). This enables the possibility of cyclic image acquisition during a normal depth profile. Measurements of large and therefore representative volumes (150 × 150 × 10 µm) in a short time (1 h) are possible [234]. In the scanning mode, the sequential detection of the single pixels and voxels results in long measuring times, and in practice only small volumes (10 × 10 × 1 µm) can be measured [235].

3.1.2.8. Applications

3.1.2.8.1. Implantation Profiles

Ion implantation is a common method for doping semiconductors. Since the concentrations of the dopants (mostly B and P) are very low, a dynamic range of more than five orders of magnitude is often necessary. The measurement of P is more difficult than that of B due to the mass interference of the ^{30}SiH and the ^{31}P ions. A high mass resolution of $m/\Delta m = 5000$ or an energy offset of 300 V is necessary.

To obtain high implant concentrations at large depths and with low damage, channeling is used.

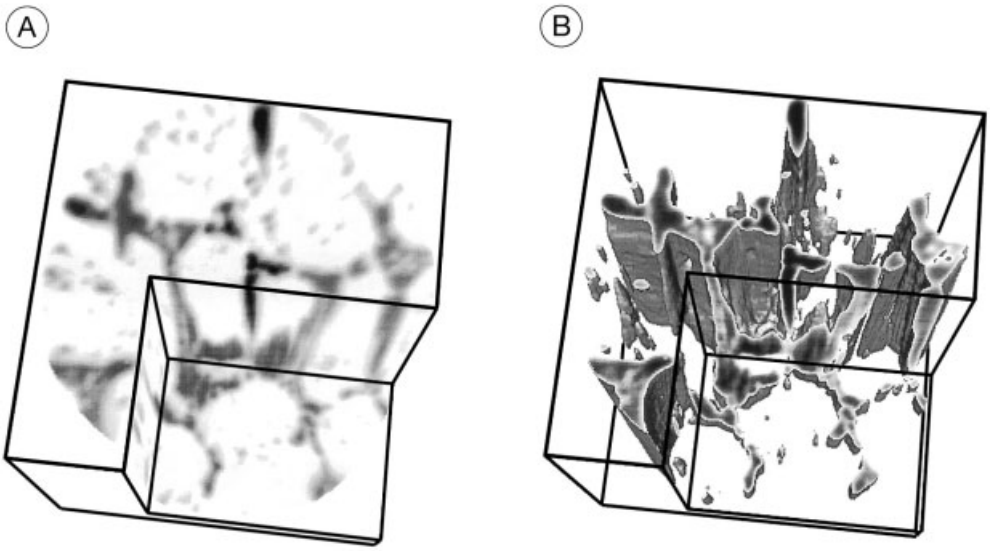

Figure 55. 3D-SIMS measurement of the V distribution in high-speed steel. Dimensions: 150 × 150 × 10 µm, primary ions: oxygen, primary energy: 5.5 keV, primary intensity: 2 µA. A) Block view; B) Isosurface view.

In this case the wafer is bombarded in a high-index crystal direction so that the implant ions can penetrate the solid mostly along the "channels" between the occupied lattice positions. However, the formation of a small number of defects restricts the use of the channeling effect to the upper limit of implantation concentrations.

The depth profiles in Figure 56 show that the typical flat channeling implantation profile is generated with low doses only. Increasing the dose superimposes the normal implantation profile. Undertaking such experiments with homogenous wafers allows the calibration of models for semiconductor production.

3.1.2.8.2. Layer Analysis

The benefits provided by surface-coating techniques have the common goal to increase productivity and reduce costs [236]. CrN provides excellent material properties to meet these requirements. It has high hardness, superior oxidation stability, high corrosion resistance, low friction coefficient, and good adhesion. Because of its outstanding properties CrN is often applied in the machining of iron-based materials and wear parts. Although CrN already has extraordinary material properties, further improvement is essential. For this reason it is very important to have not only exact but also complete information about the presence of desired or undesired trace elements within these layers. Figure 57 shows the depth profile of such a layer; an enrichment of Ti and Al at the interface to the substrate is visible. The Si signal in the layer increases with depth. It is not possible to determine from the depth profile alone whether diffusion of Si into the CrN layer occurs (see below).

Figure 58 shows 3D-SIMS distributions of the elements Si, Al, Ti, and Cr. The Cr distribution is shown from the bottom side to illustrate the rough interface. It is evident that the interfaces are not smooth. This is the reason for the slowly decreasing Cr signal. As can be seen in the 3D distribution the different depth profiles of Si, Ti, and Al in the layer are due to respective particulate inclusions.

From the depth profile alone it is not distinguishable whether the profile shape is a result of diffusion of elements or due to geometric effects. This example demonstrates the ability of 3D-SIMS to improve the information content of depth profiles.

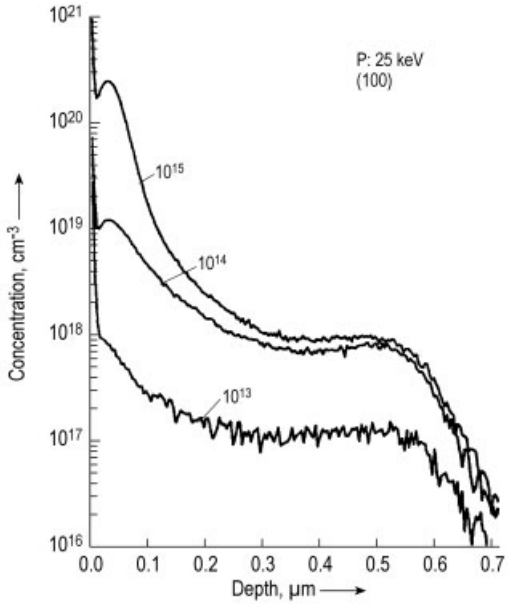

Figure 56. Depth profiles of P implantations in Si. Primary ions: Cs^+, primary energy: 14.5 keV.

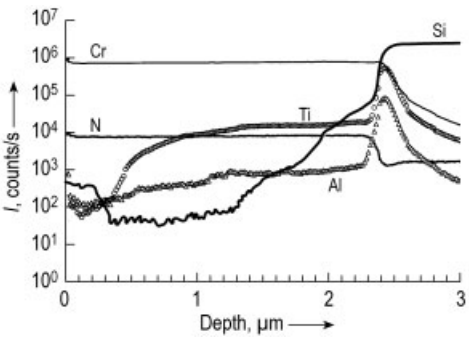

Figure 57. Depth profile of a CrN layer deposited on Si. Primary ions: O_2^+, primary energy: 5.5 keV, primary intensity 2 µA, measuring time: 1 h.

3.1.2.8.3. 3D Bulk Element Distribution

Steel alloyed with ca. 1 % Al shows a higher resistance to oxidation at high temperatures. Powder metallurgical, hot isostatically pressed steels have a better macro- and microscopic homogeneity of the distribution of all elements. Combining these advantages should result in higher performance of high-speed steels.

The steel particles are covered with an Al_2O_3 layer before hot isostatic pressing [237]. 3D SIMS measurements (Fig. 59) shows that this layer still exists in the compacted material. Carbonitrides are precipitated as small particles with a diameter of 30 nm inside the particles. As an example for a trace element with concentration in the ppm range, the distribution of Ca^+ ions is shown.

3.2. Secondary Neutral Mass Spectrometry (SNMS)

3.2.1. General Principles

The limitations of SIMS – some inherent in secondary ion formation, some due to the physics of ion beams, and some due to the nature of sputtering – have been mentioned in Section 3.1. Sputtering produces predominantly neutral atoms; for most of the elements in the periodic table the typical secondary ion yield is between 10^{-2} to 10^{-5}. This leads to a serious sensitivity limitation when extremely small volumes must be probed, or when high lateral and depth resolution analyses are needed. Another problem stems from the fact that the secondary ion yield can vary by many orders of magnitude as a function of surface contamination and matrix composition making quantification very difficult. Quantification can also be jeopardized by interferences from molecules, molecular fragments, and isotopes of other elements having the same mass as the analyte. Very high mass resolution can reject such interferences but only at the expense of detection sensitivity.

To improve sensitivity and quantification, several post-ionization methods have been developed. These techniques decouple the sputtering and ionization processes by ionizing the *sputtered neutrals* (SNs) after emission from the sample surface. While the *secondary ion* (SI) yield can vary widely with the surface composition, this is not true of the SN yield. For example, a variation in the SI yield over three orders of magnitude (10^{-2} to 10^{-5}) produces less than a 1 % change in the SN yield. As a result, SNMS has much smaller matrix effects than SIMS. Still, changes in sputter yield (as a function of matrix) can cause quantification errors. These can be corrected by using standards or by determining the various sputter yields. Recombination ef-

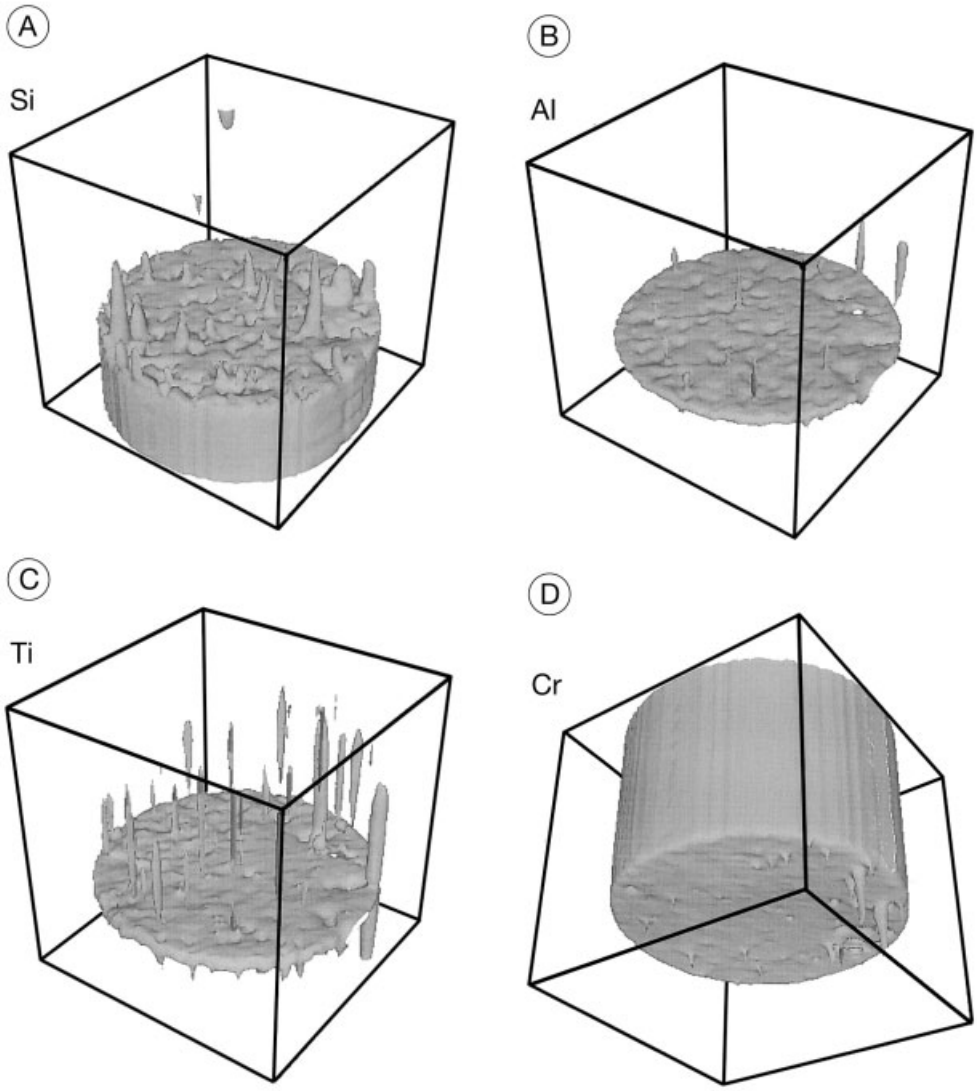

Figure 58. 3D depth profile of the CrN layer on Si substrate; same measurement as Figure 40. The CrN layer thickness is 2.3 µm, diameter 150 µm. The Cr distribution is shown from the bottom side.

fects and incomplete atomization are other possible sources of matrix effects in SNMS.

Compared to the other methods of surface and thin film analysis, the main advantages of SNMS are:

- Compared to XPS and AES sputter depth profiling: After attaining sputter equilibrium, and until a layer with different sputtering behavior is reached [238], the SN flux represents stoichiometry and not the altered layer concentrations evolving due to preferential sputtering effects.
- Compared to GD-OES (and -MS, if used for depth profiling): SNMS provides better depth resolution (1 nm instead of 10 nm range) and hardly any formation of molecules with the plasma gas (Ar), unlike the GD techniques, where argides are formed due to the comparatively high pressure.

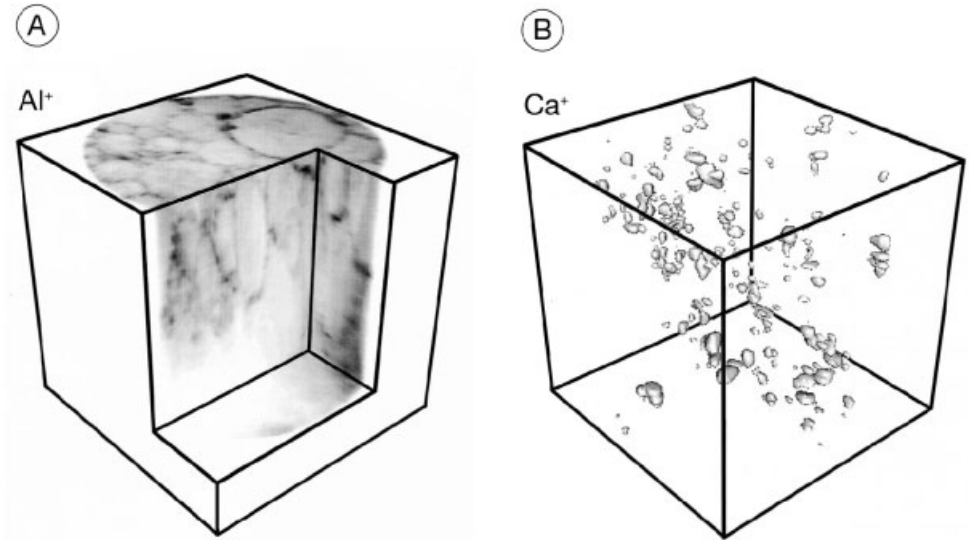

Figure 59. 3D element distribution of powder metallurgically produced steel. Edge of the cube: 150 µm. Primary ions: 0.1 µA Cs^+ (top), 2 µA O_2^+ (bottom), depth: 1 µm, (top), 11 µm (bottom).

Depending on the matrix and the post-ionization technique, SN^+ spectra can be dominated by atomic or by molecular signals. Compounds with high mass differences between the constituent elements and strong chemical bonds, e.g., result in significant amounts (10 – 100 % of the metal monomer) of molecules, mostly dimers, even if post-ionized by electrons. Then, isobaric interferences from molecular signals have to be considered.

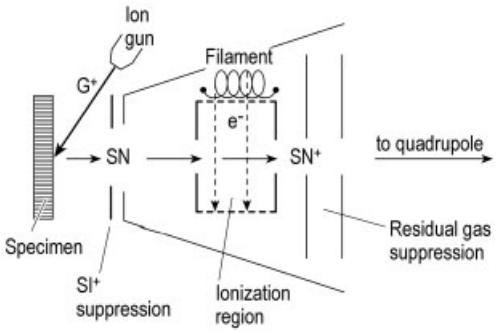

Figure 60. Schematic of an electron-beam SNMS setup (after [240])

SNMS sensitivity depends on the efficiency of the ionization process. SNs are post-ionized by electrons from a broad electron beam or a high-frequency (HF) plasma (i.e., an electron gas), or, most efficiently, by photons from a laser. Particularly the photoionization process allows the fragmentation rate of sputtered molecules to be adjusted by varying the laser intensity, pulse width, and/or wavelength.

SNMS is suitable for quantitative element depth profiles of metallic and electrically insulating samples. Laser SNMS enables the additional acquisition of 2D element distributions; with HF-plasma SNMS bulk analysis is also feasible.

3.2.2. Electron-Beam and HF-Plasma SNMS

3.2.2.1. Principles

The basic principle of electron-beam SNMS, as introduced by Lipinsky et al. in 1985 [239], is simple (Fig. 60): As in SIMS, the sample is sputtered with a focused ion beam. SN post-ionization is accomplished with an electron beam accelerated between a filament and an anode. The applied electron energy ($E_e \approx 50 \pm 20$ eV) is higher than the range of first ionization potentials (IP) of the elements (4 – 24 eV; see Fig. 61). Typical ionization probabilities are in the 0.01 % range. SI^+ and residual gas suppression is achieved with electrostatic lenses prior to SN post-ionization and energy filtering, respectively.

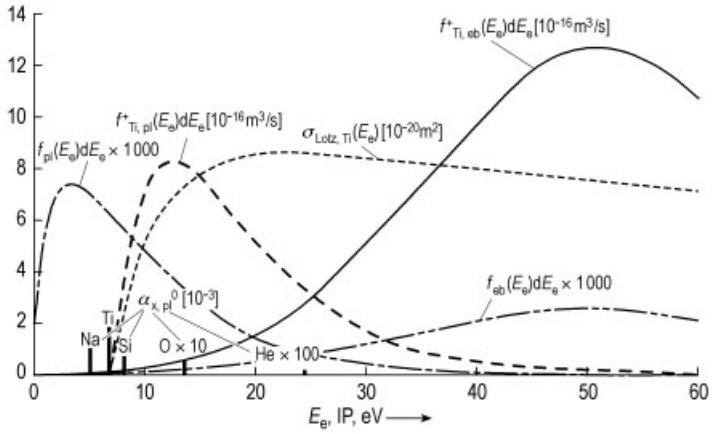

Figure 61. Distributions $f(E_e)dE_e$ of electron energy (E_e) for a low-pressure HF-plasma (suffix pl, Maxwellian, with temperature $T_e = 80\,000$ K) and an electron beam (suffix eb, simplified to Gaussian shape with 40-eV half-width); $\sigma_X(E_e)$ of the E_e-dependent electron-impact ionization cross section for X = Ti calculated according to LOTZ [241]; $f^+_X(E_e)dE_e$ (X = Ti) of the differential ionization rate constant (see Section 3.2.2.4); and respective calculated post-ionization probabilities α^0_X for X = Na, Ti, Si, O, and He, in the order of their first ionization potentials IP, for the HF-plasma parameters given in Section 3.2.2.4

OECHSNER's basic diagram of HF-plasma (or electron-gas) SNMS (Fig. 62) [242] was first published in 1966 [243]. An inductively coupled, low-pressure, HF noble gas plasma (mostly used: Ar, density $n_{pl} \approx 10^{10}\,cm^{-3}$) serves for sputtering by means of the ion component, and for postionization by the Maxwellian electron gas. Its temperature $T_e \approx 10^4 - 10^5$ K fits the given IP range well; depending mostly on IP and on the plasma (i.e., electron) density n_e, ionization probabilities are mostly in the 0.1 % range. Conducting samples introduced into the plasma are set to a negative potential U_{DBM} (DBM: direct bombardment mode) of a few hundred volts. Accelerated by the difference between U_{DBM} and plasma potential (ca. $+30 \pm 5$ V), Ar^+ ions are able to induce comparatively shallow subsurface collision cascades with sputter yields Y on the order of 0.1 – 1.

Positive secondary ions (SI^+) are repelled by an electrode in electron-beam SNMS or completely (in DBM) or widely (in HFM; see below) re-attracted by the sample surface in HF-plasma SNMS. Ionized contaminating species are suppressed by energy filtering.

3.2.2.2. Instrumentation

The electron-beam SNMS setup is shown schematically in Figure 60. The SN flux through the ionization region, which is a few centimeters in length, is crossed by an electron current in the lower milliampere range. Instead of the simple lens depicted in Figure 60, an electrostatic analyzer can serve for energy discrimination of SN^+ against ionized residual gas (1 – 10 eV vs. 10 meV). In contrast to the other SNMS techniques, electron-beam SNMS modules can easily be an added onto many existing UHV-based surface analysis instruments. Compared to a plasma, the lack of positive-charge carriers in an electron beam gives rise to space charge limitations.

A scheme of the commercially available INA-3 type spectrometer (SPECS, Berlin, Germany) is given in Figure 63. A constant Ar pressure of 0.3 ± 0.1 Pa is maintained by a piezo-valve between the high-pressure supply and the cylindrical plasma chamber (volume: 1.26 L; typically used values are given). The HF power of 150 ± 30 W is supplied from an HF generator by a single-turn coil. Electron cyclotron wave resonance [244] is effected, and skin effects of the electron gas are suppressed by a static magnetic field in the millitesla range, maintained by a constant current of 5 ± 2 A flowing through two parallel, rectangular Helmholtz coils with 184 turns each, 25 cm apart. Delivering primary ion current densities of 0.1 – 1 mA/cm^2, the plasma can effect sputter rates in the range of 0.1 – 1 nm/s. If U_{DBM} is fitted appropriately to n_e and T_e with a given distance between sample surface and

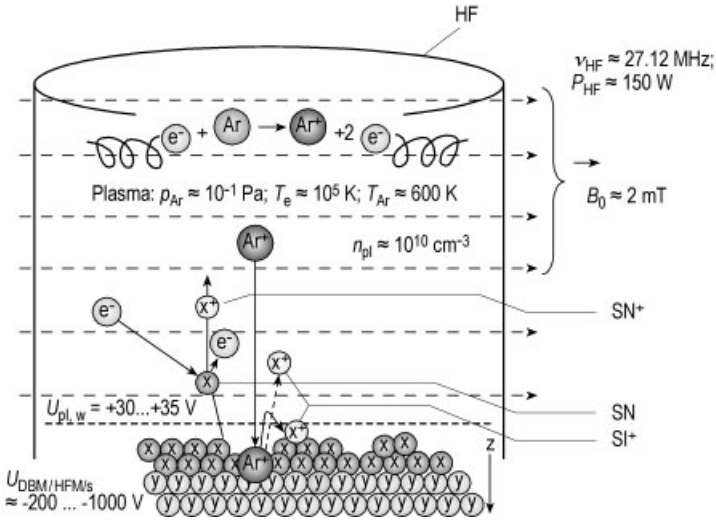

Figure 62. Basic principles of HF-plasma SNMS.
ν_{HF}, P_{HF}: HF generator frequency and power, respectively; p_{Ar}: plasma gas (Ar) pressure; T_e, T_{Ar}: electron and plasma gas temperatures, respectively; $n_{pl} = n_e$: plasma (e^-, Ar^+) density; B_0: magnetic flux; U_{DBM}, U_{HFM}: DC or HFM bombardment voltage, respectively (see text); U_s: surface potential vs. ground (HFM); $U_{pl,w}$: plasma sheath edge potential; $SN^{(+)}$: (post-ionized) secondary neutrals; SI^+: positive secondary ions.

plasma sheath edge, sputter erosion occurs with perfect lateral homogeneity, that is, nanometer depth resolution, on a spot 2 – 5 mm in diameter [245]. This compares to removing sand grains from a soccer field layer by layer; rough surfaces, of course, yield less ideal depth resolutions. The large area sputtered with high current density is one of the reasons why typical signal intensities are an order of magnitude higher than in electron-beam SNMS.

The gas temperature of the Ar plasma in the range of 600 ± 100 K results in temperatures of water- and liquid-nitrogen-cooled samples of 350 ± 30 and 200 ± 30 K, respectively. In addition, C, N, and O species desorbed from the chamber walls cause a plasma contamination level in the range 0.01 – 0.1 %.

After post-ionization in the cylindrical plasma space, 3 cm long, between the sample surface and the opposite wall, the SN^+ enter a 90° electrostatic ion energy analyzer (ion optics) that suppresses ionized plasma gas particles by a factor of $10^8 - 10^9$; noise levels are correspondingly low (1 Hz). Optical transmission is on the order of a few per cent.

In both electron post-ionization techniques, mass analysis is performed by a quadrupole mass analyzer (see Section 3.1.1.1.2), and pulse counting by a dynode multiplier. In contrast to a magnetic sector field, a quadrupole mass analyzer enables swift switches between mass settings, thus enabling continuous data acquisition for many elements even at high sputter rates within thin layers.

With useful yields typically around 10^{-10} – a value that also holds for electron-beam SNMS – typical measured intensities are in the range $10^4 - 10^6$ for HF-plasma SNMS, depending on the material and U_{DBM}. A typical plasma and a low U_{DBM} of around 300 V gives ultimate depth resolution and low intensities, whereas a U_{DBM} of 800 V enables bulk analysis in the parts-per-million range (apart from C, N, and O that are implanted from contamination) but poor depth resolution.

Electrically insulating materials can be analyzed in HF-plasma SNMS by applying a square-wave HF in the range of 100 kHz to the sample (Fig. 64). Dielectric charge transfer at the start of a period shifts the surface potential to the applied amplitude U_{HFM}. Ar^+ ions are attracted from the plasma and sputter the surface until the end of Δt^-. Then the potential increase $\Delta U_s = 10 - 100$ V caused by their charge is converted to a positive absolute ΔU_s, which is diminished to less than 1 V within < 0.1 µs by

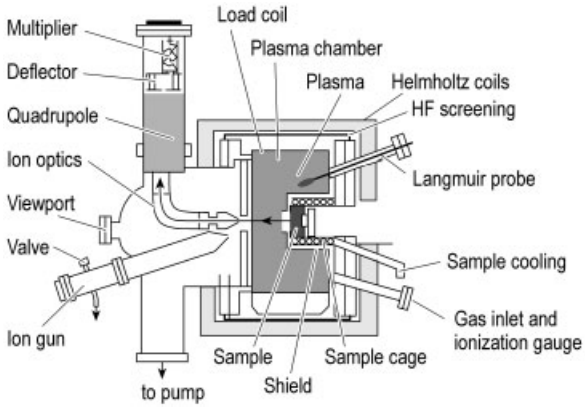

Figure 63. Schematic of the commercial INA-3 HF-plasma SNMS apparatus (SPECS, Berlin, Germany)

the plasma electron current. The missing sputtering time of Δt^+ and, more important, the often low sputter yield of dielectrics, especially oxides, give rise to lower intensities than given above.

A novel HF-plasma SNMS instrument which can be combined with XPS has been developed [246]. Detection limits in the parts-per-billion range have been attained on a dedicated HF-plasma instrument attached to a double-focusing mass spectrometer [247].

3.2.2.3. Spectral Information

Spectral information in electron-beam and HF-plasma SNMS, both being dynamic sputtering methods, is similar to dynamic SIMS. Since molecules from a heavily sputtered surface survive a collision cascade and an electron impact to a comparatively low extent, emphasis lies on elemental analysis based on mostly atomic spectra. Binary molecules formed from main constituents may add restricted speciation information. Multiple SN ionization occurs rarely because of the low probability of a second electron impact. As a consequence, electron-beam and HF-plasma SNMS spectra show a low degree of interference and are easy to interpret. Figure 65 shows a typical HF-plasma SN^+ mass spectrum, obtained from a Ti – Al alloy. The quadrupole mass detctor is the reason for Δm being unity throughout the spectrum. As discussed above, molecular intensities are in the per cent range of the parent peaks, doubly ionized species appear in the per mil range, i.e., for main constituents only, and adsorbed or implanted plasma gas contamination gives rise to a few additional low-intensity peaks.

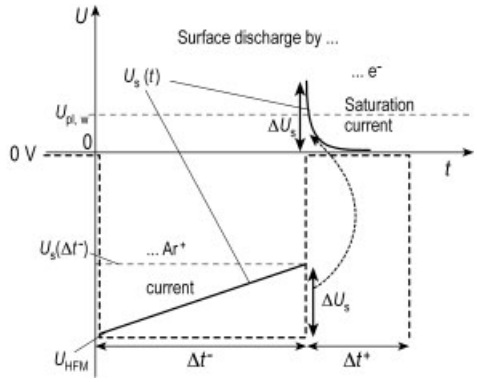

Figure 64. HFM principle
––– Applied voltage; —— Variation of surface potential U_s with time t

3.2.2.4. Quantification

The measured SN^+ intensity $I_X(E_{X,0})$ of element X, its SN being sputter emitted with energy E_0, is given by

$$I_x\left(E_{X,0}\right) = \frac{I_{G+}}{e_0}\Delta Y_x\left(E_{X,0}\right) W_X n_e t_X\left(E_{X,0}\right) F_X^+ T_{X+} \qquad (3.16)$$

where I_{G+}/e_0 is the primary ion particle current (e_0 = elementary charge); $\Delta Y_X(E_{X,0})$ the sputter yield of X valid for the emission energy range

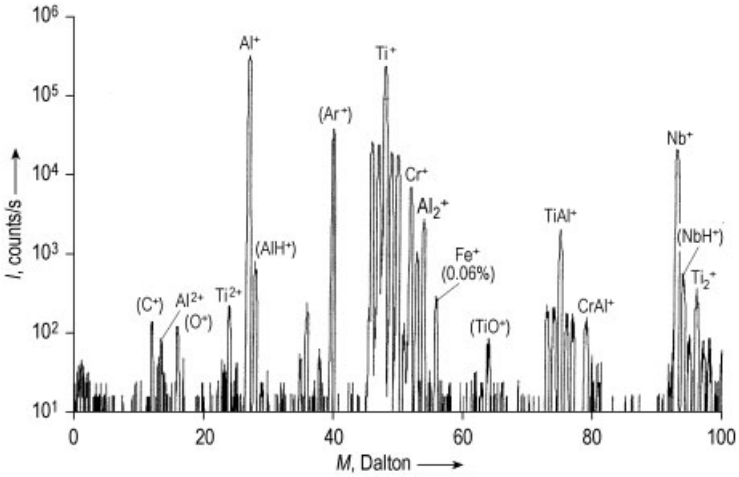

Figure 65. HF-plasma SNMS survey spectrum of a powder metallurgically produced Ti – Al alloy sample. In parentheses: signals coming from adsorbed and re-sputtered plasma gas contamination. Sample: Ti-48Al-2Cr-2Nb (mol-%)

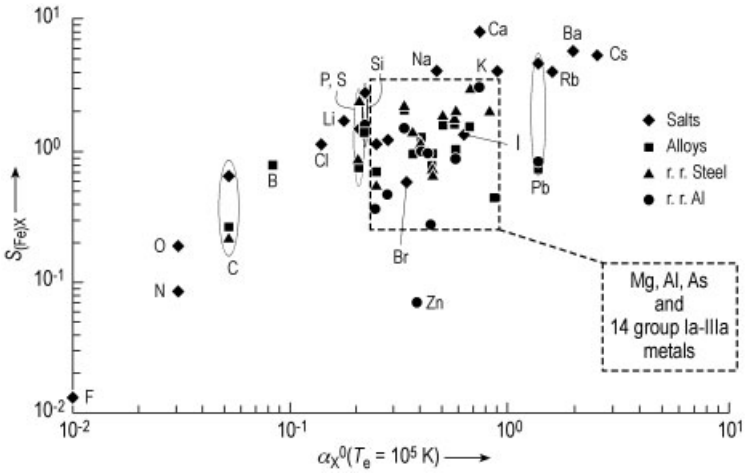

Figure 66. Experimental, Fe-related HF-plasma SNMS sensitivity factors $S_{(\mathrm{Fe})\mathrm{X}}$ with elements X ordered according to their post-ionization probabilities α_{X}^{0} calculated according to [253]; from [250] (salts); [251] (alloys); [252] (round robins, r.r.).

$E_{\mathrm{X},0} \pm \Delta E_{\mathrm{f}}/2$, ($\Delta E_{\mathrm{f}}$ = width of the energy window of the ion optics); W_{X} the probability for an SN emission into the ionization volume (electron beam, plasma) within the acceptance angle of the detection system (10^{-3} range); n_{e} the electron density; $t_{\mathrm{X}}(E_{\mathrm{X},0})$ the dwell time of an X particle with $E_{\mathrm{X},0}$ in the ionization region; T_{X^+} the combined spectrometer transmission and detector response probability for X^+ ions; and F_{X}^{+} is the ionization rate constant [248] (see also Fig. 61):

$$F_{\mathrm{X}}^{+} = \int_{E_{\mathrm{I,X}}}^{\infty} \sigma_{\mathrm{X}}(E_{\mathrm{e}}) \sqrt{\frac{2E_{\mathrm{e}}}{m_{\mathrm{e}}}} f(E_{\mathrm{e}})\,\mathrm{d}E_{\mathrm{e}} = \int_{E_{\mathrm{I,X}}}^{\infty} f_{\mathrm{X}}^{+}(E_{\mathrm{e}})\,\mathrm{d}E_{\mathrm{e}} \tag{3.17}$$

where $E_{\mathrm{I,X}} = \mathrm{IP}_{\mathrm{X}}$ is the ionization potential of X, σ_{X} the total ionization cross-section, $f(E_{\mathrm{e}})$ the distribution of the electron energy E_{e}, and m_{e} the mass of the electron at rest.

The product $n_e t_X(E_{X,0})F_X^+$ is equal to the post-ionization probability α_X^0:

$$\alpha_X^0 = n_e t_X\left(E_{X,0}\right) F_X^+ \tag{3.18}$$

The values $\alpha_{X,pl}^0$ in Figure 61 were calculated with $n_{pl} = n_e = 6 \times 10^{15}\,cm^3$, $E_{X,0} = 30\,eV$ (optics setting), and $l_{pl} = 3\,cm$ long path through the plasma. With a mean electron density $n_{eb} = 8 \times 10^{14}\,m^{-3}$, $E_{X,0} = 10\,eV$, and $l_{eb} = 3\,mm$, the corresponding $\alpha_{X,eb}^0$ amount to 6, 9, 13, 45, and 180 % of the given $\alpha_{X,pl}^0$ values (X = Na, Ti, Si, O, He).

The absolute detection (or sensitivity) factor is

$$S_X = W_X \alpha_X^0 T_{X+} \tag{3.19}$$

and can be related to a reference value S_{ref} according to

$$S_{(ref)X} \equiv \frac{S_X}{S_{ref}} = \frac{I_X}{I_{ref}} \frac{c_{ref}}{c_X} \tag{3.20}$$

since for a multicomponent material emitting only atomic SN after attaining sputter equilibrium, the total sputter yield Y is

$$Y_X = c_X Y \tag{3.21a}$$

$$Y = \sum_{i=1}^{n} Y_i = \frac{e_0}{I_{G^+}} \sum_{i=1}^{n} \frac{I_i}{S_i} \tag{3.21b}$$

$i = 1, 2, ..., X$; n = sample elements.

The differences in post-ionization probability α_M^0 of M = Ti and Al on the one hand, and of Cr and Nb on the other hand, both pairs being present in the same concentrations, are the major reason that the respective signals in Figure 65 are similar but not equal. Figure 66 [249] shows that experimental relative detection factors vary roughly with theoretical α_X^0. This variation by three orders of magnitude is mostly due to F, O, N, C, and Zn (as yet unexplained), with more than one order of magnitude, as well as some alkali and alkaline earth metals with another half order of magnitude. The $S_{(Fe)X}$ of all other elements represented in Figure 66 vary between 0.3 and 3, irrespective of matrix and measurement conditions; depending on the latter, identical $S_{(Fe)X}$ vary by factors of 2 5. It is this narrow range of $S_{(ref)X}$ variation compared to SIMS which has made SNMS famous for its ease of quantification.

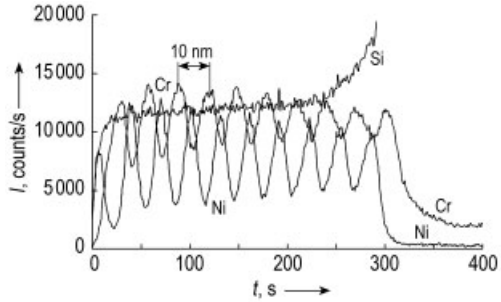

Figure 67. HF-plasma SNMS sputter time profile (DBM) of a multilayer system consisting of 10 double layers, 5 nm Cr + 5 nm Ni each, on a Si wafer.

3.2.2.5. Element Depth Profiling

Figure 67 gives an example for the depth resolution routinely achieved with both electron-beam and HF-plasma SNMS on suitable samples. If Equation (3.21) is applicable, that is, all sputtered material was recorded with ΣI_i, then Ye_0/I_{Ar^+} is equal to $\Sigma I_i/D_i$ [254], and depths z can be calculated according to:

$$z = \frac{I_{G^+} tY}{e_0 A_s n_u} \tag{3.22a}$$

$$n_u = \frac{\varrho_s N_A N_u}{M_u} \tag{3.22b}$$

where t is the sputter time, A_s the sputtered sample area, n_u the number density of formula units u (e.g., Ta_2O_5, $Ti_{0.48}O_{0.5}Cr_{0.02}$, ...), ϱ_s the sample density, N_A the Avogadro number, N_u the number of atoms per formula unit, and M_u the relative mass (in Daltons) of u.

Absolute detection factors D_X must be determined for this procedure by integrating intensities over time while sputtering suitable pure element samples and determining the crater volume; in the case of HF-plasma SNMS, the weight loss can also be measured.

Figure 68 shows that good depth resolution can also be maintained on sputtering through an insulating micrometer-thick layer.

3.2.2.6. Applications

Figure 69 exemplifies the use of HF-plasma SNMS in materials research [255]. The intermetallic compound γ-TiAl to which 2 mol % Cr

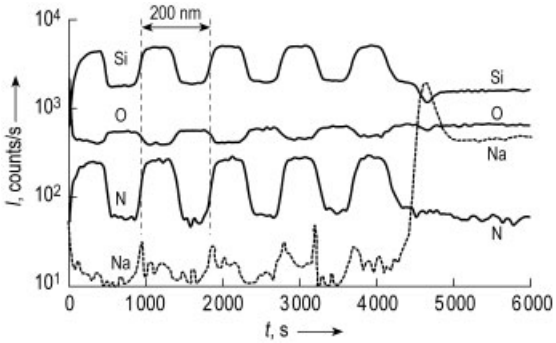

Figure 68. HF-plasma SNMS sputter time profile (HFM) of a multilayer system consisting of five double layers, 100 nm SiO_2 + 100 nm Si_3N_4 each, on a glass substrate (courtesy of V.-D. Hodoroaba, BAM Berlin, and Schott Glas, Mainz, Germany)

was added to increase the ductility was heated to 1070 K for less than 4 s only and cooled down (A) and or kept at this high temperature for 8 h. In the first case, an oxide layer about 150 nm thick formed with Al_2O_3 enriched at the top. Figure 69 B shows that after this initial oxide layer formation, nitrogen starts to penetrate to the interface between oxide and alloy and oxidizes the latter by forming nitrides. The N enrichment at the interface and tracer experiments [255] furthermore suggest that nitride has continuously been converted to oxide between the oxide layer and the nitride-enriched interface layer. Figure 69 B also shows that Ti diffused outward to form an upper TiO_2-rich layer.

Applications of electron-beam and HF-plasma SNMS have been reported in the following areas:

- Aerosol particles [256]
- X-ray mirrors [257], [258]
- Ceramics and hard coatings [259–263]
- Glasses [264]
- Interface reactions [265]
- Ion implantations [266]
- MBE layers [267]
- Multilayer systems [268]
- Ohmic contacts [269]
- Organic additives [270]
- Perovskite-type and superconducting layers [271]
- Steel [272], [273]
- Surface deposition [274]
- Sub-surface diffusion [275]
- Sensors [276–278]
- Soil [279]
- Thermal barrier coatings [280]

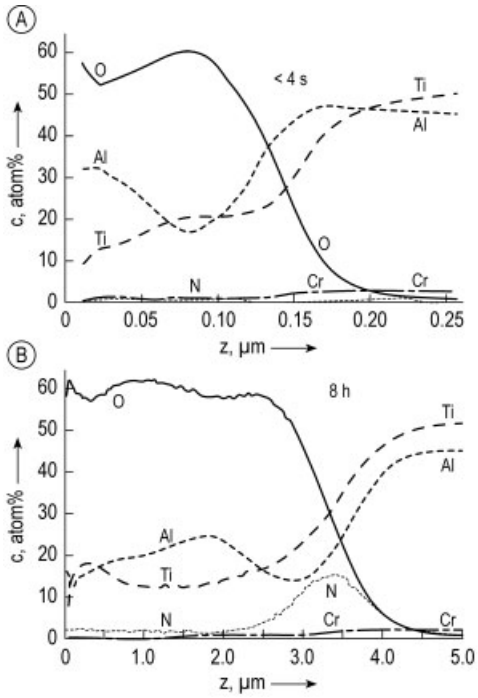

Figure 69. HF-plasma SNMS sputter depth profiles (DBM) of Ti-48Al-2Cr (atom %) samples oxidized at 800 °C in air for less than 4 s (A) and 8 h (B)

3.2.3. Laser-SNMS

3.2.3.1. Principles

3.2.3.1.1. Nonresonant Laser-SNMS

Surface analysis by nonresonant (NR) Laser-SNMS [281–285] has been used to improve the

ionization efficiency while retaining the advantages of probing the neutral component. In NR-Laser-SNMS, an intense laser beam is used to nonselectively ionize all elements and molecules within the volume intersected by the laser beam (see Fig. 70). With sufficient laser power densities, it is possible to saturate the ionization process. For NR-Laser-SNMS, adequate power densities are typically achieved only in a small volume at the focus of the laser beam. This limits sensitivity and leads to problems with quantification due to the differences in effective ionization volume for different elements. The nonresonant post ionization technique provides rapid, multielement and molecular survey measurements with significantly improved ionization efficiency over SIMS, but it still suffers from isobaric interferences.

3.2.3.1.2. Resonant Laser-SNMS

Resonant (R) Laser-SNMS [286–291] has almost all the advantages of SIMS, electron-SNMS, and NR-Laser-SNMS, but it has the additional advantage of using a resonance laser ionization process which selectively and efficiently ionizes the desired elemental species over a relatively large volume (see Fig. 70). For over 80 % of the elements, R-Laser-SNMS has an ionization efficiency of alnost unity over a large volume, so the overall efficiency is greater than that of NR-Laser-SNMS. Quantification is also made simpler because the unsaturated volume (where ionization is incomplete) is insignificant compared to the saturated volume. This leads to consistently high sensitivity and great accuracy in the R-Laser-SNMS measurements. Particularly the extremely high selectivity prevents almost all isobaric and molecular interferences. This improves the accuracy of the data and simplifies its interpretation without requiring very high resolution mass spectrometry. The main drawbacks accompanying this immunity to interference are that only one element is analyzed at a time and that low-repetition-rate pulsed lasers impose a low duty cycle. The use of multiple lasers permits the simultaneous ionization of multiple elements, but this is practically limited to two to three elements. Advances in diode-pumped Nd:YAG lasers and optical parametric oscillators (OPOs) will allow increased repetition rates of over 1 kHz as opposed to 30 – 100 Hz.

3.2.3.1.3. Experimental Setup

In Laser-SNMS experiments, the sample is bombarded with a finely focused pulsed primary ion beam (see Section 3.1.1.1.1). While the primary ion pulse is striking the target, voltages on the extraction electrodes are set so that electric fields retard positive secondary ions. After the primary ion pulse, the voltages are switched so that ions formed by laser post-ionization of the neutral particles in the sputtered cloud have significantly different energy than the secondary ions formed in the sputtering process. Ions are analyzed by a TOF mass spectrometer with an electrostatic energy analyzer or a reflectron which discriminates against the secondary ions. The TOF allows all photoions to be simultaneously measured in each laser shot.

Laser-SNMS imaging is performed by either scanning the ion beam over the sample or by translating the sample under a fixed ion beam. If an LMIS is used for sputtering, nanoanalysis down to 50 nm can be achieved. Charge compensation for analyses of insulators is possible by using pulsed low energy electrons, which are introduced during the time interval between sputtering pulses. Laser-SNMS depth profiles are obtained by sputtering the sample with a continuous ion beam and taking data with a pulsed ion beam in the center of the crater (dual-beam mode, see Section 3.1.2.2.1). In Laser-SNMS, as in other microprobe techniques (see Section 3.1.2.8), the combination of 2D imaging with depth-profile capabilities allows the visualization of the three-dimensional distribution of elements. However, the extremely high overall efficiency of R-Laser-SNMS (typically 3 – 8 % useful yield for most elements) permits the measurement of low concentrations of analytes in much smaller volumes, thus enabling higher lateral and depth resolution in these 3D visualizations. This capability could greatly contribute to the study of dopant and contaminant distributions in semiconductor devices.

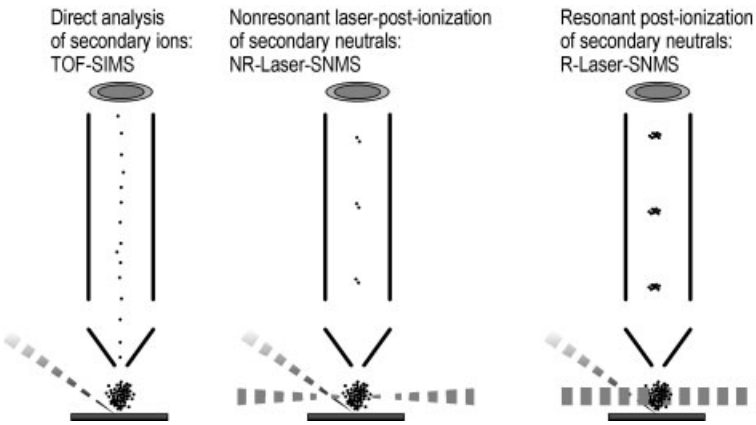

Figure 70. Comparison between TOF-SIMS and Laser-SNMS

3.2.3.1.4. Ionization Schemes

As illustrated in Figure 71, several laser schemes can be used to ionize elements and molecules. Scheme (a) represents nonresonant ionization. Since the ionization cross section is very low, a very high laser intensity is required to saturate the ionization process. Scheme (b) shows a simple single-resonance scheme. This is the simplest but not necessarily the most desirable scheme for resonant post-ionization. Cross sections for (photo)ionization are much lower than those for resonant excitation steps; therefore, the photoionization laser intensity must be proportionally higher in order to efficiently ionize the atoms. High-intensity visible or UV laser beams can cause nonlinear ionization of interfering species, which is rarely a problem with high-intensity IR beams. Therefore, whenever selectivity is an issue, IR wavelengths for the photoionization step are preferable. Scheme (c) is the most versatile of the schemes shown, and it gives the best performance for metals: H, He, C, N, O, F, Ne, P, S, Cl, Ar, Br, Kr, I, Xe, and Rn cannot be analyzed using Scheme (c) with commercial laser systems and frequency upconversion in nonlinear crystals. Resonance ionization Scheme (d) can be used when an autoionizing structure exists in the ionization continuum. The cross section for autoionization is significantly larger than for normal photoionization; therefore, the intensity of the ionizing laser can be reduced. This allows the use of visible wavelengths without extensive nonlinear ionization of interfering species. Scheme (e) is used when even the lowest energy excited states of the analyte element require vacuum-UV (VUV) wavelengths. It differs from Scheme (c) only in the experimental method used to generate the wavelength for the first transition. For wavelengths below about 195 nm, either four-wave mixing or anti-Stokes Raman must be used. This makes Scheme (e) the most experimentally complex of the resonance ionization schemes. With this technique it is possible to measure the amount of ^{81}Kr (a radioactive isotope with approximately 1 part in 10^{18} natural abundance in the atmosphere) in a water sample after several isotope-enrichment steps [292]. Scheme (f) is generally used when the energy of the excited state structure is too high for Schemes (b), (c), or (d) and the experimental complexity of scheme (e) is to be avoided. Examples are C, N, O, P, S, Cl, Br, Kr, I, and Xe. The two-photon transition used in scheme (f) is a second-order process, which is why its cross section is quite low compared to the one-photon transitions of the other schemes. This means that high-intensity light is needed, and focusing of the laser beam is generally required. In fact, the intensity of the light needed for efficient two-photon excitation is usually sufficient to ionize the excited state, and so Scheme (f) often requires only a single wavelength. This experimental simplicity is counterbalanced by the fact that the high-intensity light can cause nonlinear processes, including multiphoton ionization, to occur in many other species, leading to interferences.

Isotope shifts for most elements are small in comparison to the bandwidth of the pulsed lasers

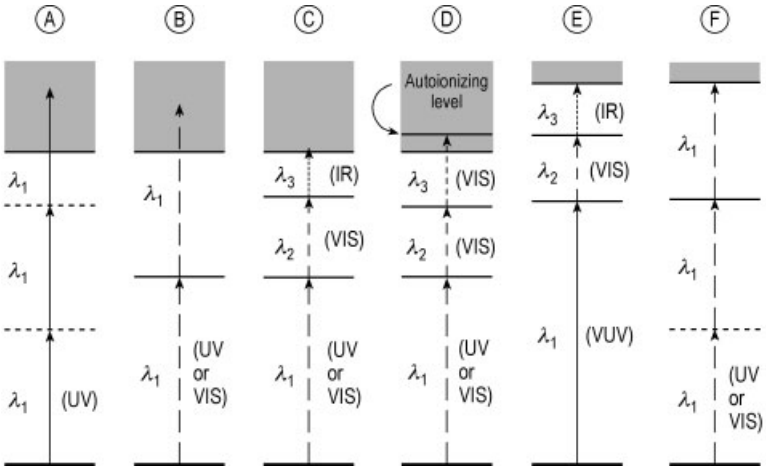

Figure 71. Schemes of laser post-ionization steps

used in resonance ionization experiments, and thus all the isotopes of the analyte will be essentially resonant with the laser. In this case, isotopic analysis is achieved with a mass spectrometer. Time-of-flight mass spectrometers are especially well-suited for isotopic analysis of ions produced by pulsed resonance ionization lasers since all the ions are detected on each pulse.

3.2.3.2. Instrumentation

A versatile Laser-SNMS instrument consists of a microfocus ion gun, a sputtering ion gun, a liquid-metal ion gun, a pulsed flood electron gun, a resonant laser system consisting of a pulsed Nd:YAG laser pumping two dye lasers, a non-resonant laser system consisting of a high-power excimer or Nd:YAG laser, a computer-controlled high-resolution sample manipulator on which samples can be cooled or heated, a video and electron imaging system, a vacuum lock for sample introduction, and a TOF mass spectrometer.

3.2.3.3. Spectral Information

Compared to electron impact on molecules, laser photoionization can be made to induce fragmentation of molecules to a much lesser extent by optimizing the laser parameters: intensity, pulse width and wavelength. This results in a higher degree of chemical information from Laser-SNMS spectra. On the other hand, significant multiple ionization of atoms typically occurs when using high laser powers to achieve high post-ionization rates.

3.2.3.4. Quantification

Only a portion of the neutral particles which is in the ionization volume (IV) of the laser beam is in principle available for the ionization. For this description the geometrical yield $Y_{\mathrm{IV}}(\mathrm{X}_i^0)$ is introduced (for notation, see Section 3.1.1.2):

$$Y_{\mathrm{IV}}\left(\mathrm{X}_i^0\right) = \frac{\text{No. of } \mathrm{X}_i^0 \text{ in the ionization volume}}{\text{No. of } \mathrm{X}_i^0} \tag{3.23}$$

Post-ionization occurs with probability $\alpha(\mathrm{X}_i^0 \rightarrow \mathrm{X}_j^{\oplus})$, where $\oplus$ denotes a photoion:

$$\alpha\left(\mathrm{X}_i^0 \rightarrow \mathrm{X}_j^{\oplus}\right) = \frac{\text{No. of generated } \mathrm{X}_j^{\oplus}}{\text{No. of } \mathrm{X}_i^0 \text{ in the interaction volume}} \tag{3.24}$$

Molecular particles can dissociate during the ionization process. This is taken into account by the different indices i and j.

For post-ionized particles, it is possible to define a generalized transformation probability $P^{\oplus}(\mathrm{A} \rightarrow \mathrm{X}_j^{\oplus})$:

$$P^{\oplus}\left(\mathrm{A} \rightarrow \mathrm{X}_j^{\oplus}\right) = \frac{\text{No. of generated } \mathrm{X}_j^{\oplus}}{\text{No. of sputtered } A} \tag{3.25}$$

$P^{\oplus}(\mathrm{A} \rightarrow \mathrm{X}_j^{\oplus})$ is given by Equation (3.26):

$$P^{\oplus}\left(\mathrm{A} \rightarrow \mathrm{X}_j^{\oplus}\right) = \sum_i P\left(\mathrm{A} \rightarrow \mathrm{X}_i^0\right) Y_{\mathrm{IV}}\left(\mathrm{X}_i^0\right) \alpha \left(\mathrm{X}_i^0 \rightarrow \mathrm{X}_j^{\oplus}\right) \quad (3.26)$$

The summation over i is necessary, because different precursors X_i^0 can be transformed into the same photoion $\mathrm{X}_j^{\oplus}$ by fragmentation.

For the detection of the generated ions particles $\mathrm{X}_i^{\mathrm{q}}$, the transmission $T(\mathrm{X}_i^{\mathrm{q}})$ of the mass spectrometer and the detection probability $D(\mathrm{X}_i^{\mathrm{q}})$ are considered in the same way as discussed in Section 3.1.1.2.

An important quantity is, again, the useful yield $Y_{\mathrm{u}}(\mathrm{A}, \mathrm{X}_i^{\oplus})$:

$$Y_{\mathrm{u}}\left(\mathrm{A}, \mathrm{X}_i^{\oplus}\right) = \frac{\text{No. of detected } \mathrm{X}_i^{\oplus}}{\text{No. of sputtered } A} \quad (3.27)$$

Values of $Y_{\mathrm{u}}(\mathrm{A}, \mathrm{X}_i^{\oplus})$ for elements typically range from 10^{-4} to 10^{-2} in the case of NR-Laser-SNMS, and from 10^{-2} to 10^{-1} in the case of R-Laser-SNMS. If the experimental parameters are not well known, the concentration of A can also be quantified by using the relative sensitivity factor (RSF) method [see Equations (3.8) and (3.9) in Section 3.1.1.2].

3.2.3.5. Applications

3.2.3.5.1. Nonresonant Laser-SNMS

Figure 72 [284] shows a typical NR-Laser-SNMS spectrum of a contaminated Si wafer under low dose sputtering conditions (primary ion fluence $< 10^{14}\ \mathrm{cm}^{-2}$). For post-ionization, the radiation of an excimer laser operating at a wavelength of 248 nm was focused into the sputtered plume (intensity $10^{10}\ \mathrm{W\,cm}^{-2}$). Since the ionization cross-section of Si at 248 nm is relatively low, the intensity of the matrix signals (Si, SiO) is lower than expected. Due to the use of a high-resolution TOF mass spectrometer (mass resolution $m/\Delta m = 3000$), the metal contaminants (Mn: 20 ppm, Ni: 130 ppm, Cu: 500 ppm, and Zn: 400 ppm) are clearly separated from the molecular signals at the same integral mass. The ultimate detection limits of NR-Laser-SNMS are given by the amount of available sample material and by the useful yield. If the outer monolayer is completely consumed within an area of $10^{-4}\ \mathrm{cm}^2$ and if the useful yield is $10^{-2} - 10^{-3}$, a surface coverage of about 10^7 atoms/cm^2 can be detected by at least 10 ions. This limit can only be reached if the corresponding ion signal is free of background and free of mass interferences. Table 6 presents the relative sensitivity factors $S(\mathrm{M}^+, \mathrm{Si}^+ + \mathrm{SiO}^+)$ with respect to the sum of the matrix signals Si^+ and SiO^+ and the detection limits (DL) determined for several metals on Si wafers by nonresonant postionization at 248 nm and 193 nm [284]. Due to the low matrix dependence of Laser-SNMS, the same sensitivity can be achieved for the detection of trace metals in matrices like polymers or large biomolecules as for semiconductor matrices.

Element mapping with nonresonant Laser-SNMS can be used to investigate the structure of electronic devices and to localize defects and microcontaminants [293]. Typical SNMS maps for a GaAs test pattern are shown in Figure 73. In the subscript of each map the number of counts per pixel is given. The images were acquired by using a 25 keV Ga^+ liquid-metal ion source with a spot size of about 150 – 200 nm. For the images shown only 1.5 % of a monolayer was consumed ("static SNMS").

Table 6. NR-Laser-SNMS: Relative sensitivity factors $S(\mathrm{M}, \Sigma\mathrm{Si})$ and detection limits DL for metals on Si wafer surfaces

Element	$S(\mathrm{M}, \Sigma\mathrm{Si})$		DL, atoms/cm^2	
	193 nm	248 nm	193 nm	248 nm
Al	0.9	0.9	8×10^9	3×10^{10}
Ti	1.9	9	5×10^9	5×10^9
V	2.1	42	3×10^8	3×10^8
Cr	2.9	28	3×10^9	3×10^9
Fe	1.6	31	5×10^9	5×10^9
Co	1.1	6.9	5×10^8	5×10^8
Ni	0.9	14	1×10^9	1×10^9
Cu	1.2	24	5×10^9	5×10^9
Ga	4.4	0.4	5×10^8	8×10^9
As	1.9	10	3×10^8	3×10^8
Mo	1.8	32	3×10^8	3×10^8
W	5.1	39	5×10^8	5×10^8

3.2.3.5.2. Resonant Laser-SNMS

Figure 74 shows Cu depth profiles obtained from silicon and 0.8-µm silicon oxide on silicon samples. For this experiment, half a silicon wafer without an SiO_2 layer and half a SiO_2-covered

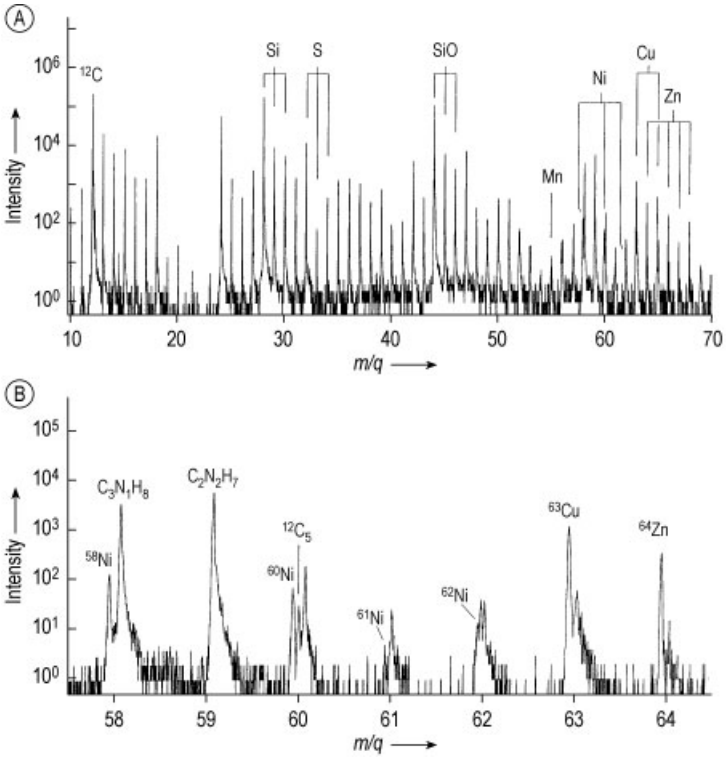

Figure 72. A) Nonresonant Laser-SNMS spectrum of a Si wafer contaminated with various transition metals and hydrocarbons [284]. B) Detail of spectrum A)

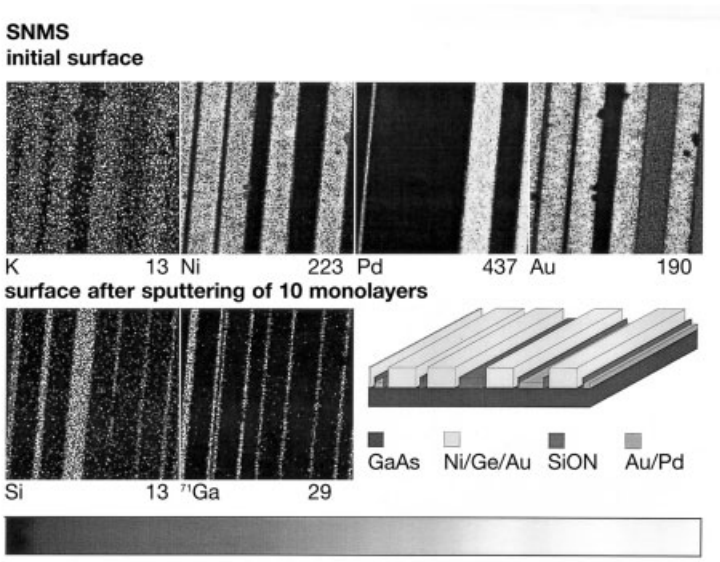

Figure 73. Nonresonant Laser-SNMS mapping of a contact test structure on GaAs. Field of view $40 \times 40\,\mu m$ [293]

Si wafer were simultaneously implanted with 5×10^{14} ^{63}Cu atoms per square centimeter at 150 keV. For analysis, both samples were cooled to 107 K to reduce migration effects, and depth profiles were taken under the same experimental conditions. The integrated Cu signal for both samples, that is, the Cu sensitivity factor for Cu in silicon and for Cu in silicon oxide, was almost the same (a difference of less than 3 %, which is due to the differences in sputter yield), demonstrating the matrix independence and the quantification accuracy possible with R-Laser-SNMS. The dynamic range for ^{63}Cu is better than six orders of magnitude. Due to the limited mass discrimination of the implanter, 2 % ^{65}Cu is present. At the SiO_2 – Si interface, a higher Cu concentration was observed. The isotopic ratio at this pile-up near the interface and below is closer to that of natural copper, and this indicates that the sample was contaminated with Cu when the oxide was grown. The Cu profile in Si was broadened due to channeling effects which occur in crystalline silicon but not in amorphous silicon oxide [294]. The lowest measured signal corresponds to a concentration of 400 ppt, which is still not the detection limit.

Figure 75 shows R-Laser-SNMS images of copper concentration in the region of Te and Cd inclusions in a CdZnTe (CZT) film. It is clear from the images that the copper concentration is approximately 10 times higher in the Te inclusion (peak concentration approximately 50 ppb) than in the surrounding CZT matrix and approximately 500 times higher in the Cd inclusion. These data support the theory that the copper preferentially migrates to Te – Cd second-phase regions in the CZT matrix [295]. Further experiments must be performed to verify whether the solubility of Cu is much higher in Cd than in Te. The same imaging technique can be used to locate and quantify pharmaceutical products in tissue with subcellular resolution and sub-ppm detection limits [296].

R-Laser-SNMS can also be applied to detect Sn-labeled DNA at positively hybridized and unhybridized sites on a DNA biosensor chip, which can be used, for example, for genetic-disease and cancer diagnostics [297], [298]. DNA diagnostics by sequencing by hybridization (SBH) involve binding a small (typically 18 – 20-mer) oligonucleotide to a chip. The chip may be glass, silicon, gold, or platinum. The polymerase chain reaction is used to produce fragments of the genomic DNA with an attached label. The size of the fragment can vary from a few dozen to several hundred or several thousand nucleotides.

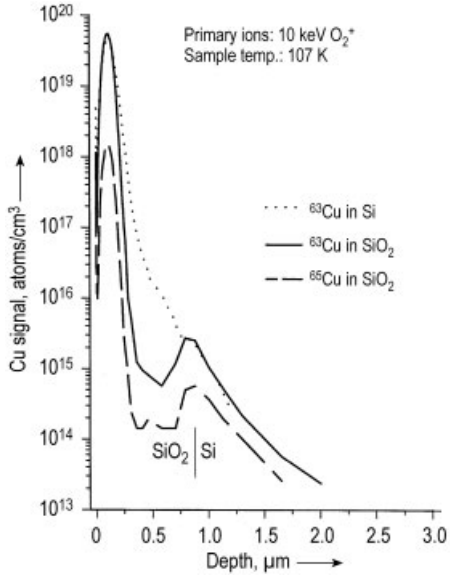

Figure 74. Depth profile of ^{63}Cu in Si and of ^{63}Cu and ^{65}Cu in SiO_2 [294]. Implantation parameter: 5×10^{14} ^{63}Cu atoms/cm^2 at 150 keV

Figure 76 shows a R-Laser-SNMS image of ^{118}Sn obtained from a DNA biosensor chip. For this experiment, two different 17-mer oligonucleotides were bound to platinum circles on a silicon wafer (three locations each). Enriched ^{118}Sn-labeled DNA which was completely complementary to one of these oligonucleotides and noncomplementary to the other was hybridized to this chip. After the hybridization, the sample was washed to remove unhybridized probes. The image shows three peaks at the complementary DNA sites. The discrimination between hybridized and unhybridized sites is better than 100. A single point on the top center spot was analyzed twice before the full image was taken, leading to a slight signal reduction. The same technique can be readily applied to a variety of problems in bioanalytical chemistry, such as screening of combinatorial libraries.

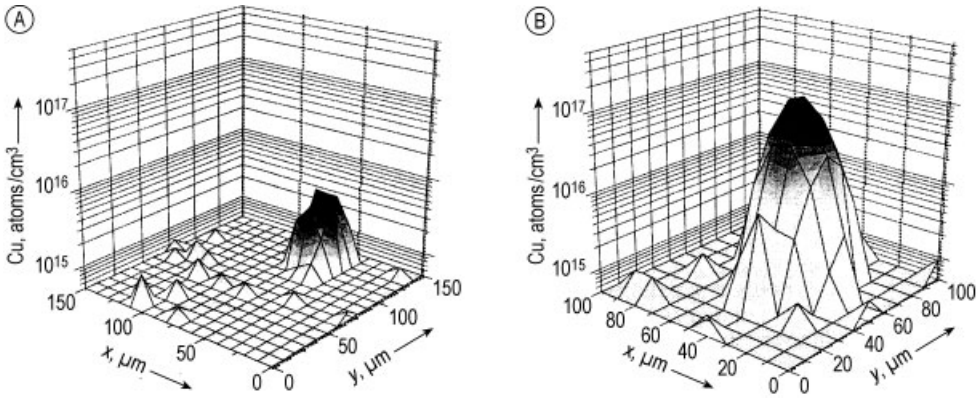

Figure 75. R-Laser-SNMS images of copper around a tellurium (A) and a cadmium (B) inclusion [295]

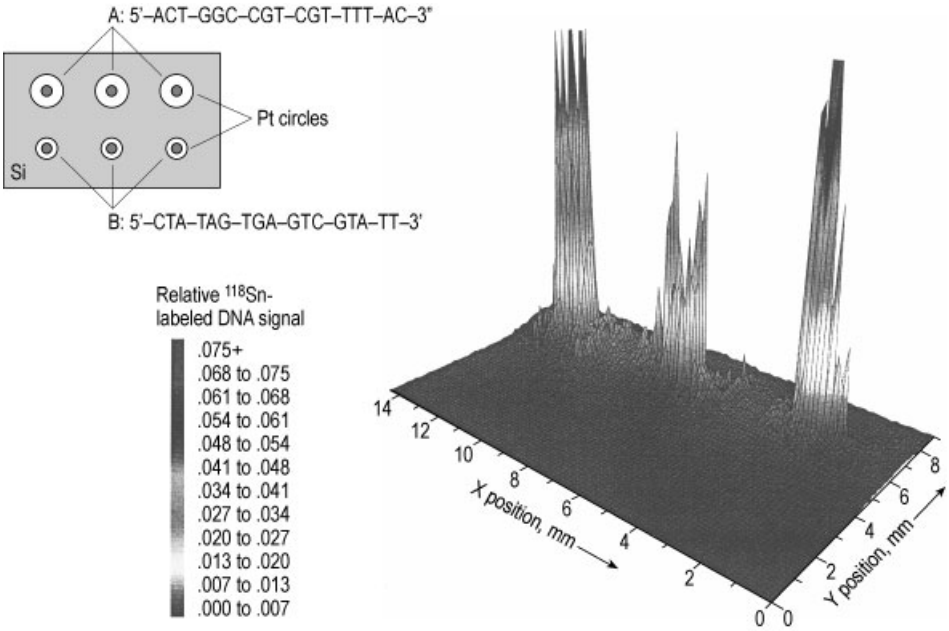

Figure 76. ^{118}Sn image of Sn-labeled DNA hybridized to six oligonucleotide spots, three of which were complimentary, and three noncomplementary

3.3. Ion-Scattering Techniques

3.3.1. Rutherford Backscattering Spectroscopy (RBS)

Rutherford backscattering spectroscopy is one of the most frequently used techniques for quantitative analysis of composition, thickness, and depth profiles of thin solid films or solid samples near the surface region. It has been in use since the 1960s and has evolved since then into a major materials-characterization technique. The number and range of applications are enormous. Because of its quantitative feature, RBS often serves as a standard for other techniques.

3.3.1.1. Principles

In RBS, a beam of monoenergetic ions, usually H^+ or He^+ of typical energy 0.5 – 2.5 MeV, is directed at a target, and the energies of the ions which are scattered backwards are analyzed. In the backscattering collision, energy is

transferred from the impinging particle to the stationary target atom. The energy ratio between the projectile energy E_1 after collision and the energy E_0 before collision, derived from binary collision theory, is [299], [300]:

$$\frac{E_1}{E_0} = K = \left[\frac{\left(M_2^2 - M_1^2 \sin^2\theta\right)^{1/2} + M_1 \cos\theta}{M_2 + M_1} \right]^2 \quad (3.28)$$

The energy ratio E_1/E_0, known as the kinematic factor K, shows that the energy after scattering depends only on the mass M_1 of the projectile, the mass M_2 of the target atom, and the scattering angle θ (i.e., the angle between incident and scattered beam). If the parameters M_1, E_0, and θ are known, M_2 may be determined and the target element identified.

For direct backscattering through 180°, the energy ratio has its lowest value given by:

$$\frac{E_1}{E_0} = \left(\frac{M_2 - M_1}{M_2 + M_1}\right)^2 \quad (3.29)$$

In the case $M_1 = M_2$, the incident particle is at rest after a central collision, and all the energy is transferred to the target atom. For target atoms with $M_2 < M_1$ no backscattering occurs.

The energy of the backscattered ion is given by Equation (3.28) only for scattering by an atom at the surface of the target. However, in RBS the ion beam penetrates the target, and an ion may be backscattered by target atoms at any point along its path. In the energy region used for RBS, the ion trajectory is a straight line (apart from the backscattering collision) along which the ions lose energy primarily through excitation and ionization of atomic electrons (electronic energy loss). The energy loss per unit path length dE/dx is called the stopping power. These additional energy losses broaden the peak observed in an RBS spectrum of a thin film.

The situation is illustrated in Figure 77. The upper part shows a thin film of Ni deposited on a Si substrate. Only particles scattered from the front surface of the Ni film have an energy given by the kinematic equation (Eq. 3.28), $E_1 = K_{Ni}E_0$. As particles traverse the solid, they lose energy along the incident path. Therefore, particles scattered from a Ni atom at the interface Si – Ni have an energy smaller than $K_{Ni}E_0$. On the outward path the particles lose again energy. On emerging from the surface, the particles scattered at the interface have a total energy difference ΔE relative to particles scattered at the surface, causing a broad peak in the backscattering spectrum. The peak width ΔE is related to the thickness of the Ni film (see below).

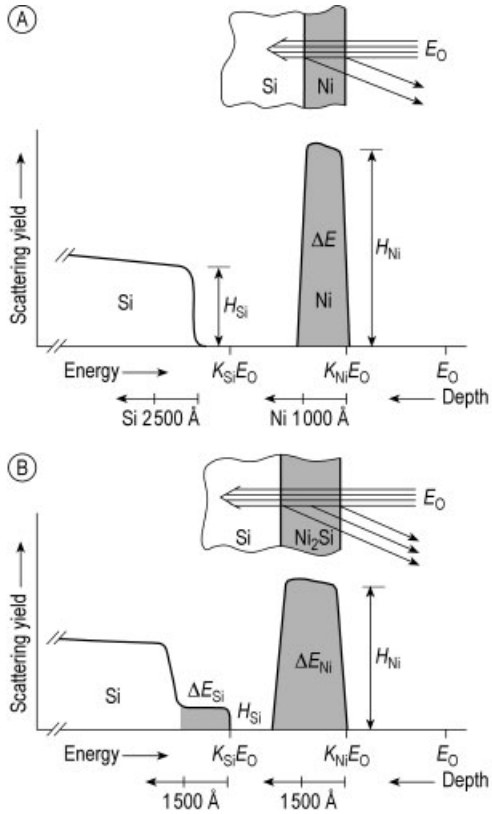

Figure 77. Schematic backscattering spectra for MeV He^+ ions incident on a 100 nm Ni film on Si (A) and after reaction to form Ni_2Si (B). Depth scales are indicated below the energy axes [299].

For Si atoms the kinematic factor is smaller ($K_{Si} < K_{Ni}$) due to their lower mass. Backscattered particles from Si atoms appear at lower energies in the spectrum. Because no Si atoms are at the surface, the energy spectrum produced by scattering from Si starts at an energy lower than $K_{Si}E_0$ and then extends to zero energy, since the Si atoms form a substrate of effectively infinite depth.

For a surface layer consisting of atoms with lower mass than the substrate atoms, the peak due to the surface layer merges with the broad continuum and appears as a small feature on top of it. For surface analysis by RBS, conditions

must be such that the mass of surface atoms is considerably higher than the mass of substrate atoms in order to resolve the peak due to surface atoms completely. Only under these circumstances is the sensitivity of RBS comparable to those AES and XPS, that is, up to ca. 10^{-3} atomic layers. No chemical information is available.

So far we have assumed that all the target atoms are equally accessible to the projectiles. In a single crystal, atomic rows and planes can guide energetic ions along the channels between rows and planes, so that the ion beam penetrates deeply into the crystal, an effect known as channeling [299–301]. Channeling occurs when the ion beam is carefully aligned with a major symmetry direction of the single crystal. Figure 78 shows a side view of this process in which most of the ion beam is steered through the channels formed by the strings of atoms. Channeled particles cannot get close enough to the atomic nuclei to undergo large-angle Rutherford scattering, and hence the scattering is drastically reduced by a factor of about 100. Atoms at the surface produce a surface peak in the backscattering spectrum because the ion beam is scattered from the surface atoms with the same intensity as from a random array of atoms. The second atom in each string of atoms is completely shadowed in a perfectly rigid lattice due to the "shadow cone" formed by the ion trajectories. Thus, channeling improves enormously the sensitivity of RBS to atoms at the surface.

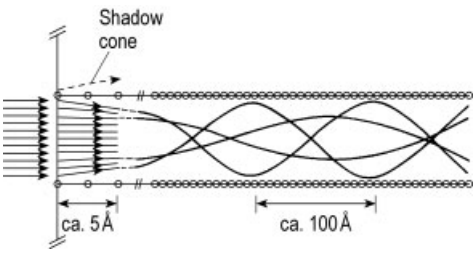

Figure 78. Schematic of particle trajectories undergoing scattering at the surface and channeling within the crystal. The depth scale is compressed relative to the width of the channel in order to display the trajectories [299]

3.3.1.2. Instrumentation

Ion beams suitable for RBS are produced in particle accelerators. By far the most widely used device for the production of MeV ion beams is the Van de Graaff electrostatic accelerator, either single-ended or double-ended (tandem Van de Graaff). The reader is referred to [302] for detailed information about accelerator technology. Accelerators normally produce ions in several charge states, and even a multiplicity of species. To extract a beam suitable for materials analysis, the beam is passed into the field of an analyzing magnet. A mass and charge selected beam then enters an UHV environment via differentially pumped apertures and is steered to the target by electrostatic or magnetic lenses. The beam size at the target is typically 1 mm^2. Small beam spots, if desired, can be obtained by suitable lens systems. Selected area analysis using small spots is usually carried out by specimen manipulation. For general analysis the ions strike the specimen at normal incidence. When structural information is sought (RBS + channeling), the specimen is mounted on a multiple-axis manipulator and may be rotated about the point of ion impact to vary the angle of incidence about channeling directions.

The scattering angle for optimum mass resolution would be 180° (a certain ΔM_2 gives the largest change in K when $\theta = 180°$, Eq. 3.28). Because of the detector size, in practice $\theta \approx 170°$ is chosen. The backscattered ions are usually detected by a solid-state detector, either a silicon barrier detector or a passivated implanted planar silicon (PIPS) detector. For conventional RBS with He^+ or H^+ ions at 1 – 2 MeV the PIPS detector has the better energy resolution (about 10 keV compared to 15 keV for surface barrier detectors [303]). The detector signals, which exhibit good proportionality to the energy of the incident particles, are amplified and sorted in energy in a multichannel analyzer.

Increasing use has been made of medium-energy ions and of heavier ions to optimize certain features of backscattering analyses. *Medium energy ion scattering* (MEIS) employs ions with energies of 100 – 500 keV together with electrostatic analyzers or time-of-flight (TOF) detectors. MEIS offers very shallow analysis depth with resolution up to monolayers [304]. *Heavy ion backscattering spectroscopy* (HIBS) uses ions such as ^{12}C, ^{16}O, or ^{35}Cl together with TOF detectors to obtain extremely high sensitivities (ca. 10^9 atoms/cm^2) for trace analysis [305].

3.3.1.3. Spectral Information

An RBS spectrum contains information on the mass of the scattering atoms, the composition of a surface layer, the depth of scattering atoms, and thickness of a surface layer.

The evaluation of the mass of the target atoms from the energy of backscattered particles is described in Section 3.3.1.1. An example of the nature of the compositional information obtainable from an RBS spectrum is based on Figure 77. The lower part of the figure shows schematically aa Ni film on Si that has undergone reaction to form Ni_2Si. After reaction, the Ni signal ΔE_{Ni} has spread slightly due to the presence of Si atoms contributing to the energy loss. The Si signal shows a step at the energy $K_{Si}E_0$ corresponding to Si in the Ni_2Si. The hatched peak areas in the figure are a measure of the number of particles scattered by Si or Ni atoms. Since the probability of a collision between the projectile and a target atom is easily obtained (see Eq. 3.32), the number of Si and Ni atoms in the target and hence the composition N_{Ni}/N_{Si} may be calculated.

Given the composition of a layer, it is possible to establish a depth scale for an element distribution or to measure the layer thickness from the energy of the scattered particles. This depends on the energy loss of the projectile on its inward and outward paths, as described in Section 3.3.1.1. The difference in energy ΔE of a particle scattered at the surface and a particle scattered at a depth x is given by

$$\Delta E = [S]\,x \tag{3.30}$$

where $[S]$ is the energy loss factor, which depends on the stopping power on the inward and outward paths, the kinematic factor K, and the orientation of the sample to both the incident beam and the detector direction [300]. In Figure 77 the depth scales are indicated.

RBS together with channeling can be used to detect and measure the thickness of an amorphous layer on an otherwise crystalline substrate. The method is commonly used to obtain structural information on the damaged surface layer of ion-implanted semiconductors and to study the damage removal upon annealing. An example is shown in Figure 79 for GaAs [306]. If the incident ion beam is directed onto the GaAs crystal in a ⟨110⟩ channeling direction, the backscattering yield is drastically reduced (d) compared to the yield when the beam is incident in a nonchanneling or "random" direction (a). Note the small surface peak in the channeled spectrum. Ion implantation with 120 keV Si^+ ions to a dose of $5 \times 10^{15}\,cm^{-2}$ produced an amorphous surface layer in GaAs. Because no channeling is possible in this layer, the intensity of the backscattered signal under channeling conditions increases and is as high as in the random case. Ions passing through the amorphous layer experience the crystalline nature of GaAs, which results in a reduced backscattering yield (b). From the width of the broad peak in curve b the thickness of the amorphous layer was determined to be about 140 nm. After annealing at 950 °C the intensity of the backscattering yield under channeling conditions dropped substantially (c). The spectrum resembles closely that of crystalline GaAs, which indicates that the implantation damage is annealed out almost completely.

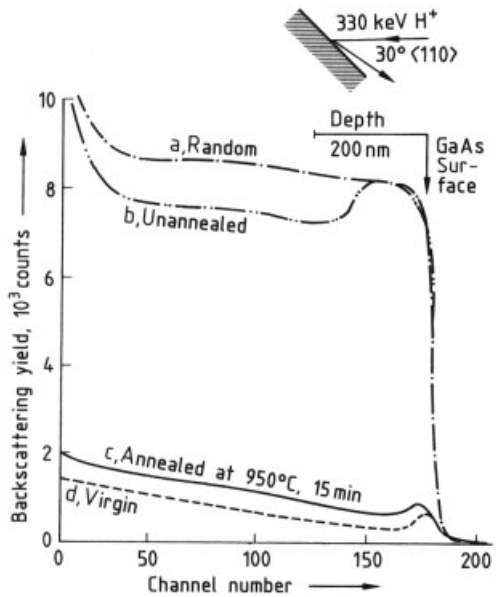

Figure 79. RBS spectra from GaAs implanted with silicon before and after annealing at 950 °C (120 keV Si^+ – GaAs, dose: $5 \times 10^{15}\,cm^{-2}$) [306]
Uppermost spectrum is taken in a nonchanneling direction; the others are in channeling directions

One of the most fascinating applications of channeling RBS is the study of lattice locations of impurity atoms. By measuring the angular de-

pendence of the backscattering yield of the impurity and host atoms around three independent channeling axes it is possible to calculate the position of the impurity. Details can be found in [301].

3.3.1.4. Quantification

For an ion beam with the total number Q of ions impinging on a thin film, the number Q_A of particles backscattered from atoms of type A and registered in the detector (also called yield Y_A), is given by:

$$Q_A = Y_A = Q N_A \sigma_A \Delta\Omega \tag{3.31}$$

where N_A is the areal density of atoms A in the film (atoms/cm^2), σ_A the differential scattering cross section (cm^2/sr), and $\Delta\Omega$ the solid angle of the detector. The differential scattering cross section describes the probability that a projectile is scattered by a target atom through an angle θ into a solid angle dΩ centered about θ. If the interaction potential between the particle (M_1, Z_1) and the target atom (M_2, Z_2) during scattering is given by the Coulomb potential, the cross section is given by the Rutherford formula:

$$\sigma(E,\theta) = \left(\frac{Z_1 Z_2 e^2}{4E}\right)^2 \cdot \frac{4\left[\left(M_2^2 - M_1^2 \sin^2\theta\right)^{1/2} + M_2 \cos\theta\right]^2}{M_2 \sin^4\theta \left(M_2^2 - M_1^2 \sin^2\theta\right)^{1/2}} \tag{3.32}$$

Equation (3.32) is given in cgs units. For practical calculations, the value $e^2 \approx 1.44 \times 10^{-13}$ MeV cm is useful. For standard RBS with 1 – 2-MeV He$^+$ ions, the use of the Rutherford cross section is justified (giving the technique its name). Deviations occur at both higher and lower energies [300].

For a particular primary ion and fixed experimental conditions, the scattering cross section is a function only of the mass and atomic number of the scattering atom, and can be calculated. Thus, in principle, all the terms in Equation (3.31), except for the required N_A, are known. Therefore, RBS is an absolute method that does not require the use of standards.

For a compound film $A_m B_n$, the composition may be calculated from Equation (3.31) to be:

$$\frac{n}{m} = \frac{N_B}{N_A} = \frac{Q_B \sigma_A(E,\theta)}{Q_A \sigma_B(E,\theta)} \tag{3.33}$$

Note that this ratio depends only on the ratio of measured yields Q_A/Q_B and knowledge of the cross section ratio σ_A/σ_B. The hard-to-measure quantities Q and $\Delta\Omega$ have cancelled.

For targets containing several elements which might produce overlapping peaks, RBS spectra are analyzed by computer simulations. The energy spectrum of the backscattered particles is calculated for the actual experimental parameters and an assumed target composition. Then the target composition is altered (either manually or by using a least-squares fitting procedure) until the calculated and measured spectra are closely matched. A widely used RBS analysis program is the RUMP code [307], [308]. Commercial software packages are also available [300].

The accuracy in RBS results is ca. 3 % for areal densities and better than 1 % for stoichiometric ratios. These high accuracies are only obtained when all relevant quantities are measured or evaluated carefully. Pitfalls which often prevent RBS from achieving its full accuracy are described in [309]. Calibration can be achieved by measuring standards obtained by either implanting into or depositing on a light element (silicon) a known amount of a much heavier element (e.g., Ta or Sb).

3.3.1.5. Applications

Since RBS is a major technique for the analysis of thin solid films and surface layers, and the number of published applications in these areas is enormous. The number of publications in which RBS has been used, either alone or together with other analytical techniques, is about 1000 per annum. The majority of applications are in the field of semiconductor technology. Ion-implantation damage and damage annealing, implantation profiles, multilayer systems grown by molecular beam epitaxy or chemical vapor deposition, silicide formation, and diffusion barriers for contacts are among the topics routinely investigated by RBS.

A typical example is the formation of a buried β-$FeSi_2$ layer in Si [310]. The system is a favorable one for RBS, since Fe is much heavier than

Si. The $FeSi_2$ was synthesized by implanting Fe to a high dose (2.7×10^{17} cm^{-2}) into Si and performing a high temperature annealing. $FeSi_2$ exists in different phases, which complicates the formation of the low-temperature semiconducting β-phase. After a short annealing (10 s) at 1150 °C, α-$FeSi_2$ is formed. The transformation to β-$FeSi_2$ is performed by prolonged annealing (17 h) at 800 °C, a temperature well below the $\alpha-\beta$ transition. Figure 80 shows the RBS spectra. The Si yield starts at the surface energy $K_{Si}E_0$ and shows a large dip at lower energies, whereas the Fe peak starts at energies below $K_{Fe}E_0$, which indicates that the $FeSi_2$ layer is buried below the Si surface. The difference in Fe signal heights for the α- and β-phases indicates a change in density during the phase transformation. It is estimated that the α-phase contains ca. 17 % structural vacancies. The channeling spectrum of the β-phase shows that the buried layer is crystallographically aligned with the Si [(001) planes of β-$FeSi_2$ are parallel or slightly off-oriented to (111) Si].

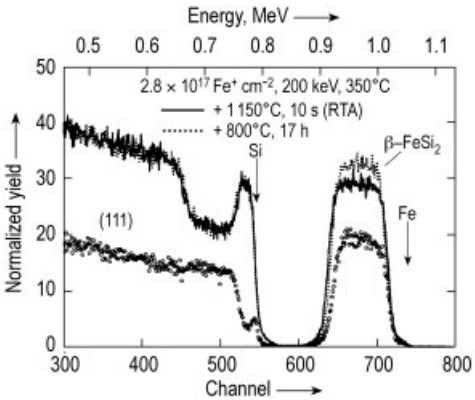

Figure 80. RBS spectra of buried α-$FeSi_2$ (full line) and β-$FeSi_2$ (dotted line) layers taken in random direction. The $\langle 111 \rangle$ channeling spectrum (circles) refers to the β-$FeSi_2$ layer [310].

RBS and channeling are extremely useful for the characterization of epitaxial layers. An example is the analysis of a $Si_{1-x}Ge_x - Si$ strained layer superlattice [311]. Four pairs of layers, each about 40 nm thick, were grown by molecular beam epitaxy on a (100) Si substrate. Because of the lattice mismatch between $Si_{1-x}Ge_x$ ($x \approx 0.2$) and Si, the $Si_{1-x}Ge_x$ layers are strained. Figure 81 shows RBS spectra in random and channeling directions. The four pairs of layers are clearly seen both in the Ge and Si yield. The spectrum for channeling in the $\langle 100 \rangle$ direction is typical, whereas a large dechanneling is observed in the $\langle 110 \rangle$ direction, due to the lattice distortion at the interface $Si_{1-x}Ge_x - Si$. (The strained $Si_{1-x}Ge_x$ layer grows such that the in-plane lattice constant is the same as that of Si, whereas the perpendicular lattice constant is larger. Therefore, $\langle 100 \rangle$ channels are rather undistorted and diagonal $\langle 110 \rangle$ channels are distorted by a small angle.) The large dechanneling yield is a direct consequence of the lattice strain. The distortion can be measured as a small angular shift in angular scans at energies corresponding to specific Si and $Si_{1-x}Ge_x$ layers. By applying high-resolution RBS, even the intermixing of Si and Ge during epitaxial growth has been studied [312].

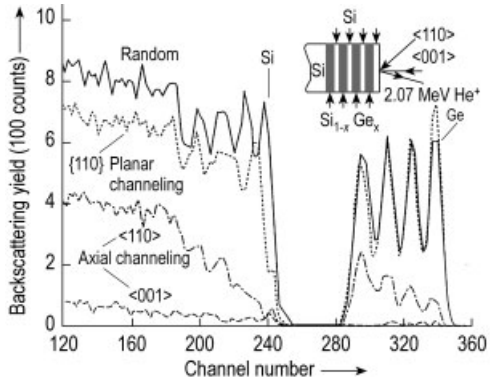

Figure 81. RBS spectra of 2.07 MeV He^+ ions backscattered from a $Si_{1-x}Ge_x - Si$ strained layer superlattice. The full line refers to a random direction; the other spectra are taken along various channeling directions [311].

Other fields for RBS analysis include optical and dielectric materials, hard and protective coatings, superconductors, and magnetic materials.

As an extremely useful method for applied materials science RBS has also contributed to basic science. A well-known example is surface melting [313]. Figure 82 shows normalized backscattering yields of H^+ from Pb near the melting point in the $\langle 101 \rangle$ direction (double alignment). The surface peak increases with increasing temperature and develops into a broad peak at 600.5 K. At 600.8 K, just above the melting temperature (600.7 K), the random spectrum is obtained, as is expected for a bulk liquid.

The broad peak below the melting temperature (curve d) can be fitted by assuming a disordered liquid surface layer (M) plus a partially ordered transition layer (I) between the liquid and the crystalline substrate. These were the first direct measurements of surface melting. Later experiments showed that the surface melting is strongly dependent on the orientation of the surface.

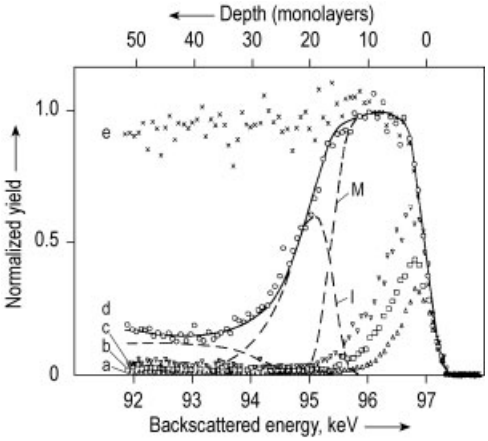

Figure 82. Normalized backscattering yields of H^+ ions from Pb near the melting point, with the incident beam and scattered beam directed along ⟨101⟩ crystal axes (double alignment)
a) 295 K; b) 506 K; c) 561 K; d) 600.5 K; e) 600.8 K
Spectrum d was fitted by a sum of contributions M from a liquid surface layer, and I from a partially ordered transition layer [313]

3.3.2. Low-Energy Ion Scattering (LEIS)

3.3.2.1. Principles

In principle, a low-energy ion scattering (LEIS) experiment is like RBS at low energies: an ion beam is sent to a target at a large angle with respect to the surface plane, and backscattered projectiles are detected at a large scattering angle [314], [315]. In contrast to RBS, in a typical LEIS experiment only positive ions are detected. Consequently, the fraction of singly charged ions P^+, i.e., the yield of singly charged ions divided by the total backscattered yield, has to be included when calculating the ion yield Q_A^+ backscattered from an atom of species A:

$$Q_A^+ = Q_0 N_A \frac{d\sigma_A}{d\Omega} P^+ S T \Delta\Omega \tag{3.34}$$

where Q_0 is the number of incident projectiles, N_A the areal density of atoms A in the surface (atoms/cm^2), $d\sigma_A/d\Omega$ the differential scattering cross section, S a steric factor describing the influence of plural scattering, and T and $\Delta\Omega$ the transmission function and the solid angle of the spectrometer, respectively.

The scattering cross-section is considerably different from the Rutherford cross-section, since the distance of closest approach R_{min} is rather large at low energies. Thus, electronic screening of the interaction between the nuclei is important. The screened scattering potential $V(r)$ reads:

$$V(r) = V_C(r) + V_e(r) = V_C(r)\,\Phi(r/a) \tag{3.35}$$

where V_C and V_e are the potentials of the target nucleus and target electrons, respectively; $\Phi(r/a)$ is the screening function, which describes how the Coulomb potential is weakened by electronic screening; and Φ is a function of the reduced distance r/a, where the screening length a is characteristic of the ion – target combination. Within the Thomas – Fermi – Molière model [299], Φ and a are given by Equations (3.36):

$$\Phi(r/a) = \sum_{i=1}^{3} \delta_i \exp\left(\frac{-c_i r}{a}\right) \tag{3.36a}$$

$$a = \frac{0.8852 \cdot a_0}{\left(\sqrt{Z_1} + \sqrt{Z_2}\right)^{2/3}} \tag{3.36b}$$

where $a_0 = 0.529$ Å (Bohr radius), and the constants $\delta_1 = 0.35$, $\delta_2 = 0.55$, $\delta_3 = 0.1$, $c_1 = 0.3$, $c_2 = 1.2$, and $c_3 = 6$. Thus, the influence of screening is negligible for small distance ($\Phi \to 1$ for $r \to 0$) and limits the interaction to distances comparable to the screening length ($\Phi \to 0$ for $r >> a$).

From the screened potential, the scattering angle $\theta(b)$ that corresponds to a given impact parameter b is obtained by solving the scattering integral, that is, the general relationship $\theta(b)$ for an arbitrary central potential [316]. From $\theta(b)$, the differential scattering cross section is readily obtained:

$$\frac{d\sigma}{d\Omega} = \left|\frac{b}{\sin\theta}\frac{db}{d\theta}\right| \tag{3.37}$$

In Equation (3.34), the major unknown is the charge fraction P^+, which still is a subject of theoretical investigation [317]. The state-of-the-art knowledge is as follows: When noble gas ions are used as projectiles, the detection of backscattered positive ions leads to the extreme surface sensitivity that is typical for LEIS, since $P^+ > 0$ is observed only when the projectile is scattered by a surface atom (due to their large binding energy, the mean charge state of low energy noble gas ions in any matter is neutral).

When only two charge states (0, +1) are of relevance, there are two contributions to P^+:

1) Projectile ions that have escaped neutralization along their trajectory
2) Those which were neutralized on their incoming path, re-ionized in the surface collision, and remained in the charged state when they left the surface:

$$P^+ = P^+_{\mathrm{surv,in}} P^+_{\mathrm{surv,out}} + \left(1 - P^+_{\mathrm{surv,in}}\right) \cdot P_{\mathrm{ReI}} P^+_{\mathrm{surv,out}} \tag{3.38}$$

where $P^+_{\mathrm{surv,in}}$ and $P^+_{\mathrm{surv,out}}$ are the probabilities that the ion survives neutralization on the incoming and the outgoing paths, respectively, and P_{ReI} is the probability for reionization in a surface collision. P_{ReI} depends strongly on the distance of closest approach in the surface collision and on the detailed electronic structure of projectile and target atom, and it is more important at higher energies [318]. Therefore, at low energies the survivals usually dominate the charge fraction. As long as only one neutralization process (Auger or resonance process) dominates in the energy range of interest, the probability that the ion escapes neutralization P^+_{surv} is obtained from the transition rate $R(t)$ in transitions per second:

$$P^+_{\mathrm{surv}} = \exp\left[-\int \mathrm{d}t R(t)\right] \tag{3.39}$$

Taking into account that neutralization means tunneling of a target conduction electron to the ion, the time integral can easily be replaced by integration over the distance s from the surface by use of the identity $\mathrm{d}t = \mathrm{d}s/v_\perp$, where $v_\perp$ is the component of the ion velocity perpendicular to the surface. From this, the velocity dependence of the survival probability $P^+_{\mathrm{surv},l}$ is obtained:

$$P^+_{\mathrm{surv},l} = \exp\left[-\frac{1}{v_{l,\perp}} \int \mathrm{d}s R(t)\right] = \exp\left[-\frac{v_{\mathrm{c}}}{v_{l,\perp}}\right] \tag{3.40}$$

In Equation (3.40), l can mean in or out, taking into account that the ion velocity after scattering is reduced due to the recoil energy transferred to the target atom. v_{c} is obtained by integration of the transition rate along the trajectory and is a characteristic velocity, representative of the neutralization efficiency: large values of v_{c} correspond to strong neutralization.

From Equation (3.40) we can draw the following conclusions:

1) At high energies (small values of $1/v_\perp$) the ion fraction is high, since there is only a short time for neutralization
2) For a given geometry, the ion fraction decreases exponentially with the reciprocal velocity

These theoretical predictions were experimentally verified for numerous target materials (see Fig. 83 [319]). Note the pronounced difference in neutralization between carbon atoms in a carbide and in graphite. This is one of the rare examples where matrix effects are observed.

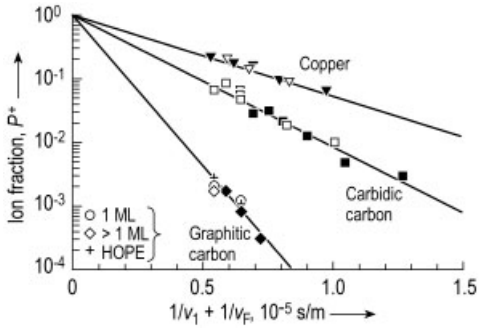

Figure 83. Ion fraction of ^{3}He and ^{4}He ions (open and full symbols, respectively) in LEIS from Cu and different types of C as a function of the sum of the reciprocal velocities of the incoming and the scattered ion [319]

3.3.2.2. Instrumentation

An RBS and a LEIS setup are compared in Table 7.

Table 7. Comparison of experimental setups for RBS and LEIS

RBS	LEIS
Accelerator	Ion source (< 5 keV)
Deflection magnet	Wien filter
PIPS detector	Electrostatic spectrometer or TOF spectrometer plus microchannel plate
H^+, He^+ ions	Noble gas ions (He^+, Ne^+)

Due to its lower beam energy, an LEIS ion source is much more compact than an accelerator for MeV ions, but it has the same purpose: to provide a beam of ions with well-defined energy and mass. In LEIS, the latter demand is fulfilled by a Wien filter that selects ions of one specific mass, like the deflection magnet in RBS. In LEIS, the energies of the ions are too low for PIPS detectors and would lead to a signal to noise ratio of less than unity. Therefore, a stack of microchannel plates is used to detect the ions in LEIS. The detection efficiency of the microchannel plates is included in the experimental transmission factor T in Equation (3.34).

The energy selection is performed either by an electrostatic analyzer or by a TOF spectrometer. Both types of spectrometer have advantages and disadvantages. In a TOF measurement, ions and neutrals are separated by a post-acceleration of the ions. Without post-acceleration, neutrals are also detected, thereby increasing the intensity by orders of magnitude; this usually leads, however, to the loss of surface sensitivity. For analysis of surface structures, however, this is not a problem, since there the shadow cones may be used to regain surface sensitivity (see below).

A conventional electrostatic spectrometer works only with ions and is therefore surface sensitive, but it has low efficiency, since only ions in a narrow energy window around a well-defined pass energy are detected simultaneously; the pass energy has to be scanned to acquire the spectrum. The detection efficiency can be increased by orders of magnitude by using a double-toroidal energy analyzer in combination with a position-sensitive detector, since such a system measures simultaneously an energy spectrum of backscattered ions in a larger energy window. Such a set-up, known as ERISS, has been developed at Eindhoven University [320] (see Fig. 84). ERISS allows static LEIS measurements (without noticeable damage), even at polymer surfaces (see below) for which the sputter rate even under LEIS conditions is very high.

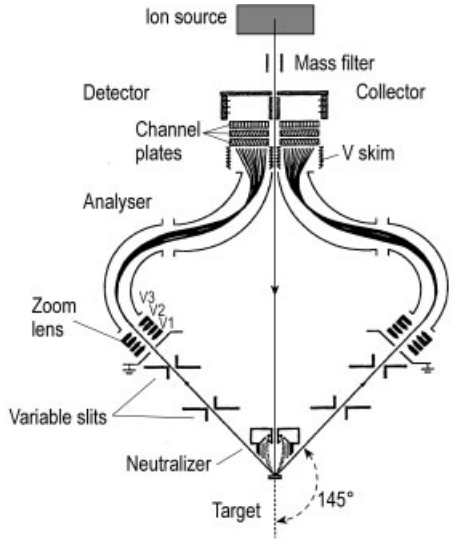

Figure 84. Schematic diagram of the ERISS setup [320]

3.3.2.3. Information

The energy spectrum in LEIS contains two types of information: energy and ion yield.

Energy Information. In the discussion up to now we have only considered the typical LEIS experiment, that is large angles of incidence and of exit with respect to the surface plane. Under these conditions, quantitative composition analysis is generally possible, since the ion – target interaction can be considered as binary collision due to the absence of matrix effects (see below).

Surface composition analysis by LEIS is based on the use of noble gas ions as projectiles, making use of the superb surface sensitivity of LEIS in this case. A consequence of this surface sensitivity is that the LEIS energy spectrum consists of lines, one per element, if the masses differ sufficiently. The lines are narrow, since inelastic energy losses play a minor role here. Thus, the information on the atomic species present is deduced from the energy of the backscattered ions, which may be converted to the mass of the scattering center. (Fig. 85 [321]). Figure 85 shows that the mass range in which LEIS is sensitive depends on the projectile mass.

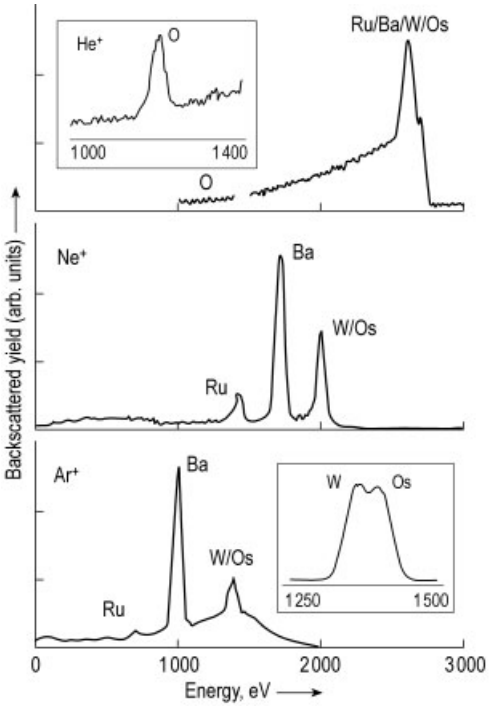

Figure 85. LEIS spectra obtained from an Os – Ru top layer dispenser cathode with 3-keV He, Ne, and Ar projectiles, respectively. The He spectrum demonstrates the absence of O on the undamaged cathode surface, the inset of the He spectrum was obtained for a cathode exposed to 20 Langmuir oxygen at room temperature [321].

Yield Information. As can be seen from Equation (3.34), the backscattered ion yield in LEIS is a measure of the concentration of a certain element in the surface, if the other quantities are known. This is generally the case: the scattering cross section can be calculated along the lines of Equations (3.36) and (3.37) or found in tabulations; S is less problematic for large angles of incidence and exit (and may be considered constant); the transmission probability T is proportional to E for any electrostatic spectrometer that varies the pass energy (as long as the detection efficiency is energy-independent); the solid angle $\delta\Omega$ of the spectrometer can be assumed constant. Hence, for fixed experimental conditions (constant primary energy, constant geometry etc.), Equation (3.34) may be simplified to:

$$Q_A^+ = \eta_A c_A \tag{3.41}$$

where η_A is the sensitivity factor which contains all element specific quantities, and c_A is the concentration of element A in the surface.

In Figure 86, LEIS spectra are shown for two completely different types of Al_2O_3 samples, that is, α-alumina (sapphire) and γ-alumina powder with high specific surface area. Very similar results were obtained in both cases after thermal treatment at 400 °C [322]. The lowering of the Al signal in γ-alumina was ascribed to shielding by hydroxyl groups, which are typical adsorbates on γ-alumina.

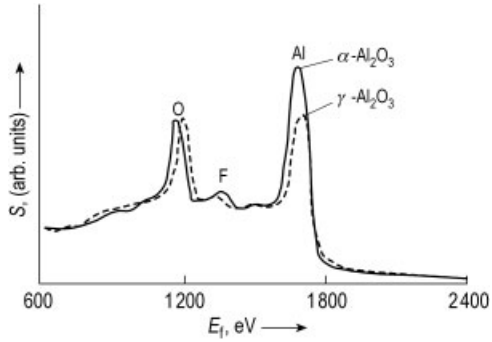

Figure 86. LEIS obtained from different types of alumina with 3-keV He ions [322]

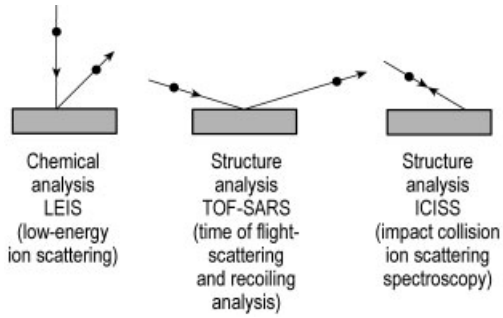

Figure 87. Different experimental geometries for low-energy ion spectroscopies

When measuring LEIS at small angles to the surface, both, for the incoming and for the outgoing particles, information on atomic structure may be obtained for a given surface (see Fig. 87). Both TOF-SARS and ICISS make use of the shadow cone and of focusing the enhanced ion intensity at the shadow onto the neighboring atom (see below): at glancing incidence, all surface atoms are hidden in the shadow cone of the first atom in a plane or terrace, and the scattered intensity is low; when the angle of incidence

is increased (with respect to the surface), at a critical angle the neighboring atoms leave the shadow cone and are hit by an enhanced flux of projectiles, which leads to a peak in the scattered intensity as a function of the angle of incidence (see below).

3.3.2.4. Quantification

The usefulness of Equation (3.41) depends crucially on whether or not the sensitivity factor η_A depends on the presence of other elements in the surface ("matrix effects"). It is an experimental finding that neutralization generally depends only on the atomic number of the scattering center, and matrix effects rarely occur. An instructive example is the neutralization of He by Al in the pure metal and in alumina. The slopes of the neutralization curves turn out to be the same in these two cases, that is, matrix effects are absent [323]. This is a strong indication that in the neutralization process not only the valence/conduction electrons, but also atomic levels below the valence/conduction band are involved.

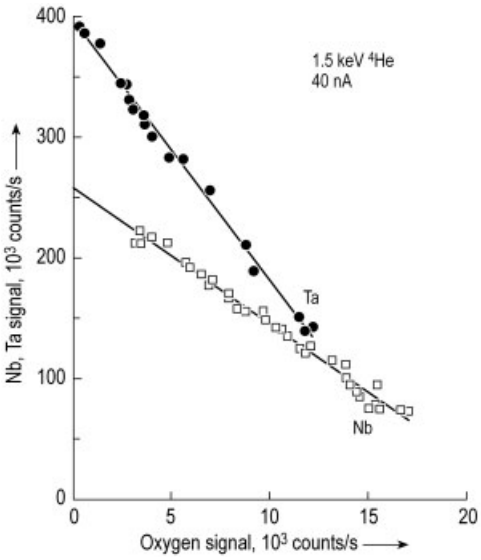

Figure 88. Peak intensities of Nb and Ta versus the oxygen peak intensity for adsorption of oxygen on the pure metals [324]

From a practical point of view, it is more convenient to measure intensity ratios instead of absolute intensities. Thus, for example, Cu can serve as a reference material, relative to which the ion intensities backscattered from the atoms of the surface under consideration are measured:

$$\frac{Q_A^+}{Q_{Cu}^+} = \frac{\eta_A c_A}{\eta_{Cu}} \tag{3.42}$$

A further example where quantitative surface composition analysis is possible for a nontrivial surface is shown in Figure 88, where for the systems Ta + O and Nb + O adsorption, the ion signal from the metal is shown as a function of the ion signal from O. In this binary case the following concentration equations hold:

$$c_{Ta} + c_O = 1 \tag{3.43a}$$

$$c_{Nb} + c_O = 1 \tag{3.43b}$$

Inserting Equations (3.43) into Equation (3.42) yields:

$$Q_O^+ = \eta_O - Q_{Ta}^+ \frac{\eta_O}{\eta_{Ta}} \tag{3.44}$$

From Equation (3.44), one expects a linear relationship between Q_O^+ and Q_{Ta}^+, Q_{Nb}^+. Indeed, this is experimentally observed (see Fig. 88 [324]).

3.3.2.5. Applications

Applications of LEIS are widespread, going far beyond the needs of a typical surface-science laboratory, since LEIS is also capable of yielding information on insulating samples and on very rough structures (e.g., catalysts). The main problem with rough insulating surfaces is that they need charge neutralization (flooding by thermal electrons which neutralize the charging of the surface by incoming ions and emitted energetic electrons).

An example is a LEIS study on a specific spinel, namely, $ZnAl_2O_4$, for which cations (Zn) in tetrahedral sites are expected [325] to be less stable and therefore move to sites below the surface where they are better shielded, yielding a lower LEIS signal. This was confirmed by Brongersma et al [326] (Fig. 89. This figure shows that LEIS is very sensitive to Zn, as shown by LEIS from ZnO, but for the spinel no Zn is visible in the surface.)

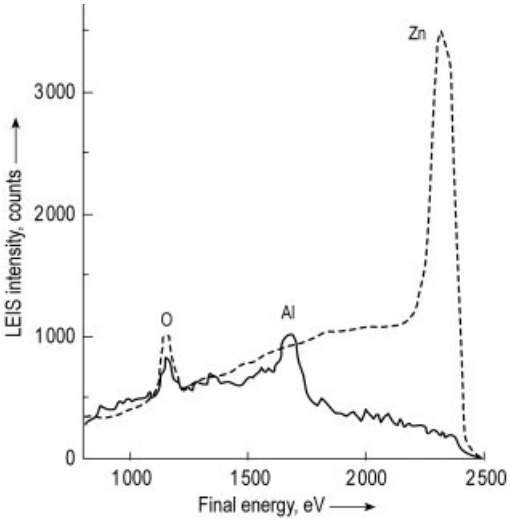

Figure 89. LEIS spectra of 3 keV He^+ scattering from $ZnAl_2O_4$ (solid line) and from ZnO (dashed line) [326]

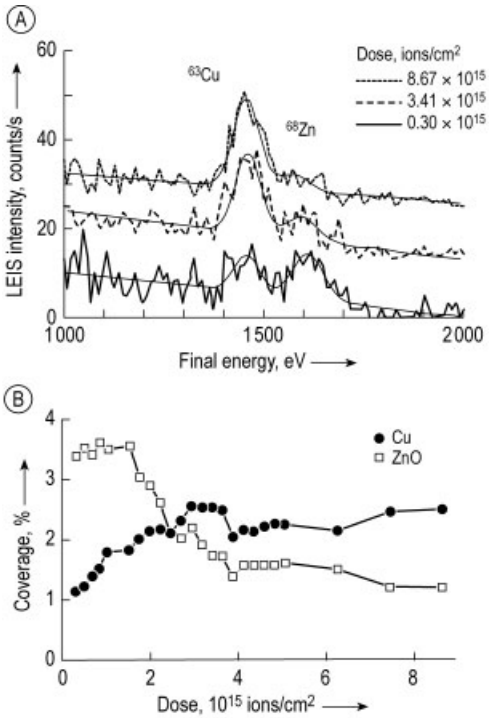

Figure 90. A) LEIS spectra of the $^{63}Cu-^{68}ZnO-SiO_2$ catalyst obtained with 5-keV Ne^+ ions for different doses (see inset, spectra are shifted vertically for clarity). Catalyst reduction temperature 700 K. Solid lines: fitted Gauss peaks [327]
B) The relative coverage of Cu and ZnO on the silica-supported catalyst, reduced at 700 K, as a function of the ion dose [327]

An even more ambitious goal is to characterize an unsupported catalyst, since in this case the surface is extremely rough, and the target rapidly deteriorates under bombardment: energy deposition leads to an enormous erosion, since the substrate cannot dispose of the energy deposited due to the the low thermal conductivity. As a consequence, static LEIS conditions have to be applied to obtain information on the surface alone. Figure 90 A shows a series of LEIS spectra obtained with 5-keV Ne^+ ions on a $^{63}Cu-^{68}ZnO-SiO_2$ catalyst, reduced at 700 K [327]. These LEIS spectra were obtained at ion doses of 3×10^{14}, 3.41×10^{15}, and $8.67 \times 10^{15}\ cm^{-2}$, respectively. Due to the use of isotopically enriched Cu and Zn, and of Ne^+ ions as projectiles, Cu and Zn can clearly be separated in the LEIS spectrum. A strong dose dependence is seen. Figure 90 B shows the dose-dependent surface concentrations of Cu and Zn. At low doses ($< 1.5 \times 10^{14}\ cm^{-2}$) the Zn concentration remains constant, while the Cu concentration increases. At these low doses, a hydroxyl layer on top of the catalyst is sputtered. The Zn signal stays constant despite removal of the adsorbate, indicating that at the virgin surface the Zn concentration was even higher.

As a final application, the analysis of a surface structure by TOF-SARS is discussed (for a general discussion, see [328]). Figure 91 shows that the scattered intensity is increased when projectiles hit a surface at the critical angle of incidence α_c due to focusing of collisions onto one of the neighboring atoms in the surface (A) or below the surface (B). Note that also recoiling atoms can be emitted in the same direction. Figure 92 a TOF spectrum is shown, measured for bombardment of CdS(0001) by 4 keV Kr^+ ions [329]. Also a spectrum of Ne projectiles scattered by Cd in CdS is shown as a function of the angle of incidence α. The peak around $\alpha \approx 20°$ is due to focusing of the projectiles by first-layer Cd atoms onto their first layer Cd neighbors, by which they are scattered into the detector. The peak at $\alpha \approx 65°$ is due to focusing of the projectiles by first-layer Cd atoms onto their second-layer Cd neighbors, by which they are scattered into the detector. Analysis of these critical angles permits the lateral first-layer interatomic spacings and the first-second interlayer spacing to be determined (see Fig. 91).

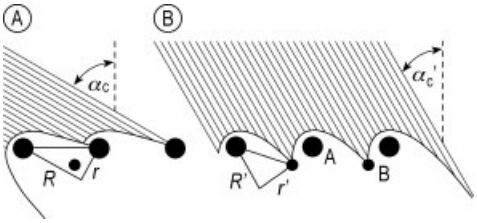

Figure 91. Schematic illustration of projectile trajectories, showing focusing collisions when the projectiles impinge at a critical angle α_c [330]

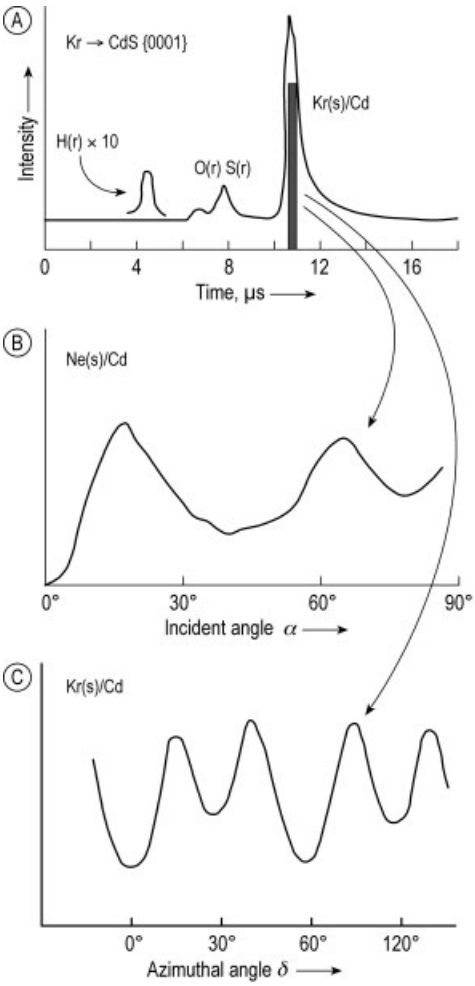

Figure 92. Schematic of a TOF-SARS experiment and examples of TOF-SARS spectra for Kr scattering from CdS: TOF spectrum; scans of the angle of incidence α and the azimuthal angle δ [329]

Finally, Figure 92 shows an azimuthal scan, again for CdS(0001) and Kr^+ ions incident at a grazing angle. This allows one to determine the surface periodicity, since the scattered intensity is minimal for incidence along the crystallographic axes, again because of shadowing. The intensity rises when the direction of incidence is tilted with respect to the crystallographic axis by a critical angle. The widths of the minima are related to the interatomic spacings along the particular directions. Wide, deep minima are expected for short interatomic spacings, because of the large azimuthal tilt needed to move the neighboring atom out of the shadow cone.

3.4. Other Ion-Detecting Techniques

3.4.1. Desorption Methods

3.4.1.1. Electron-Stimulated Desorption (ESD) and Electron-Stimulated Desorption Ion Angular Distribution (ESDIAD)

Electron irradiation of a surface, particularly one covered with one or more adsorbed species, can give rise to many types of secondary particles, including positive and negative ions. In ESD and ESDIAD, the surface is irradiated with electrons of energies in the range of 100 – 1000 eV, and the ejected positive-ion currents of selected species are measured in a mass spectrometer. If the angular distribution of the secondary ions is also measured, either by display on a screen or by using position-sensitive detection, then electron-stimulated desorption (ESD) becomes electron-stimulated desorption ion angular distribution (ESDIAD).

The electron desorption techniques are not used, and probably cannot be used, for compositional analysis, but they provide valuable information on the nature of electronic interactions leading to the breaking of bonds and, in the angle-resolved form, on the geometry of surface molecules and the orientation of broken bonds. The primary electrons do not, at the energies employed, succeed in breaking molecular or surface-to-molecule bonds or in knocking ions out of the surface directly, but the process is one of initial electronic excitation. An electron is absorbed by a surface – adsorbate complex or an adsorbed molecule itself, leading to excitation to an excited state by a Franck – Condon process. If the excited state is antibonding and the molecule or radical is already far enough

from the surface, desorption can occur. Since return to the ground state after excitation is a much more probable process, the cross sections for ion desorption are low ($10^{-20} - 10^{-23}$ cm^2). If core-level ionization is involved in the initial interaction with the incident electron, rather than valence levels as in the Franck – Condon-type excitation, then a desorption mechanism based on Auger decay has been proposed. The core level left behind after interatomic Auger decay creates a positive ion, which is then expelled by the repulsive Madelung potential.

The most common ions observed as a result of electron-stimulated desorption are atomic (e.g., H^+, O^+, F^+), but molecular ions such as OH^+, CO^+, H_2O^+, and CO_2^+ can also be found in significant quantities following adsorption of H_2O, CO, CO_2, etc. Substrate metal ions have never been observed, and this means that ESD is not applicable for surface compositional analysis. The most important application of ESD in the angularly resolved form ESDIAD is in determining the structure and mode of adsorption of adsorbed species. This derives from the fact that the ejection of positive ions in ESD is not isotropic. Instead the ions desorb along certain directions only, characterized by the orientation of the molecular bonds that are broken by electron excitation.

3.4.1.2. Thermal Desorption Spectroscopy (TDS)

Thermal desorption spectroscopy (TDS), sometimes called temperature-programmed desorption (TPD), is simple in principle. A gas or mixture of gases is allowed to be adsorbed on a clean metal foil for a chosen length of time; then, after the gas has been pumped away, the foil is heated at a strictly linear rate to a high temperature, during which the current of a particular ion or group of ions is monitored as a function of temperature. The ion masses are selected in a quadrupole mass spectrometer. As the binding energy thresholds of the adsorbed species on the surface are crossed, peaks in the desorbed ion current appear at characteristic temperatures. From the characteristic temperatures and the shape of the desorption peak above the threshold, the activation energies for desorption can be obtained, along with information about the nature of the desorption process. The mass spectrum from the mass spectrometer provides information about the species that actually exist on the surface after adsorption.

Although simple in principle, experimental artifacts that are possible in TDS must be avoided. Thus, ions accepted by the mass spectrometer must originate only from the surface of the foil, and in addition, the temperature distribution across the foil should be uniform to avoid the overlapping of desorption processes occurring at different temperatures. To ensure that these experimental requirements are met, the angle of acceptance into the spectrometer is restricted by placing a drift tube with an aperture between the foil and the spectrometer, and the foil itself is usually in the form of a long thin ribbon, from whose center section only are ions accepted. In addition, the heating rate must be sufficiently fast that the desorbed species accepted into the mass spectrometer is characteristic of the desorption process, but not so fast that the pressure rise around the foil is too great.

With correct experimental procedure, TDS is straightforward to use and has been applied extensively in basic experiments concerned with the nature of reactions between pure gases and clean solid surfaces. The majority of these applications have been catalysis related (i.e., performed on surfaces acting as models for catalysts), and in every case TDS has been used with other techniques such as UPS, ELS, AES, and LEED. To a certain extent it is quantifiable, in that the area under a desorption peak is proportional to the number of ions of that species desorbed in that temperature range, but measurement of the area is not always easy if several processes overlap.

3.4.2. Glow Discharge Mass Spectroscopy (GDMS)

In GDMS the surface to be analyzed is sputtered by a glow discharge in which the surface is the cathode, and the contribution to the sputtered flux of secondary neutrals is measured. On entering the discharge region, many of the sputtered atoms are ionized. The ions are then extracted from the discharge and passed via differential pumping stages into a region at much lower pressure where they can be mass analyzed.

Either magnetic or quadrupole mass spectrometers are used for mass analysis.

Like SSIMS and GDOS, GDMS has been developed commercially and has also had wide industrial application. Most of the applications have, like GDOS, consisted of rapid profiling of surface films many micrometers thick, with special emphasis on the depth distribution of trace elements, and would not be regarded as surface specific. Nevertheless, again like GDOS, GDMS is in principle a surface-specific technique and has occasionally been used in an ultrahigh-vacuum system and at very low erosion rates.

One of the attractions of GDMS is the relative ease and accuracy of quantification by using standard materials and relative sensitivity factors. For most elements, particularly the metallic and semiconducting elements, the spread of sensitivity factors is less than an order of magnitude, comparable to the situation in XPS. In addition, the actual sensitivities are significantly higher than in XPS or AES, with analysis at the part-per-million level being routine and at the part-per-billion level being quite possible.

3.4.3. Fast Atom Bombardment Mass Spectroscopy (FABMS)

Fast-atom bombardment mass spectrometry is very similar to SSIMS in practice, the only difference being that instead of using positively charged ions as the primary probe, a beam of energetic neutral atoms is used. Secondary ions are emitted as in SSIMS and analyzed in a mass spectrometer, usually of the quadrupole type. The beam of fast atoms is produced by passing a beam of ions through a charge-transfer cell, which consists of a small volume filled with argon to a pressure of about 10^2 Pa. Charge transfer occurs through resonance between fast argon ions and argon atoms with thermal energy, with efficiencies of 15 – 20 % if geometric and pressure conditions are optimized. Residual ions are removed by electrostatic deflection.

FABMS has two advantages over SSIMS, both arising from the use of neutral rather than charged particles. Firstly, little or no surface charging of insulating materials occurs, so that organics such as polymers can be analyzed without the need to employ auxiliary electron irradiation to neutralize surface charge. Secondly, the extent of beam damage to a surface, for the same particle flux, is much lower using FABMS than SSIMS, thus allowing materials such as inorganic compounds, glasses, and polymers to be analyzed with less worry about damage introducing ambiguity into the analysis.

3.4.4. Atom Probe Microscopy

The atom *probe field-ion microscope* (APFIM) and its subsequent developments, the *position-sensitive atom probe* (POSAP) and the *pulsed laser atom probe* (PLAP), have the ultimate sensitivity in compositional analysis (i.e., single atoms). FIM is purely an imaging technique in which the specimen in the form of a needle with a very fine point (radius 10 – 100 nm) is at low temperature (liquid nitrogen or helium) and surrounded by a noble gas (He, Ne, or Ar) at $10^{-2} - 10^{-3}$ Pa. A fluorescent screen or a microchannel plate is situated a few centimeters from the needle. A high positive voltage is applied to the needle, which causes noble-gas atoms approaching the needle to be ionized over points of local field enhancement (i.e., over prominent atoms at the surface); the ions are then repelled from these points and travel in straight lines to the screen or plate, where an atomic image of the needle tip is formed. This is called a field-ion image, and the voltage at which the image is optimized is called the *best imaging voltage.* The latter is in the range 5 – 20 kV.

3.4.4.1. Atom Probe Field-Ion Microscopy (APFIM)

If the voltage is raised further above the imaging voltage, the cohesive energy that binds atoms to the surface can be exceeded, and atoms can be removed by what is termed *field evaporation.* The evaporation field required is a function of the sample material and the crystallographic orientation of the needle. The removal of atoms is the basis of APFIM and its daughter techniques. If the evaporation field is pulsed with very short ($\leq$ 10 ns) voltage pulses of extremely sharp rise times ($\approx$1 ns), and the time taken for a field-evaporated atom to travel from the tip to a detector is measured, then the mass-to-charge ratio

of the ionized atom can be established; that is, elemental identification of the individual atom is possible. The region on the tip from which atoms are removed is selected by rotating or tilting the tip until the desired region, as viewed in the FIM image, falls over an aperture of about 2 mm diameter in the fluorescent screen. Most field-evaporated atoms will strike the screen, but those from the selected area will pass through the aperture into the time-of-flight mass spectrometer. For an aperture of 2 mm, the area analyzed on the tip is ca. 2-nm in diameter.

3.4.4.2. Position-Sensitive Atom Probe (POSAP)

Pulsed operation of the APFIM leads to analysis of all atoms within a volume consisting of a cylinder of ca. 2 nm diameter along the axis of the tip and aperture, with single atomic layer depth resolution but no indication of just where within that volume any particular atom originated. The advent of position-sensitive detectors has allowed APFIM to be extended so that three-dimensional compositional variations within the analyzed volume can be determined. This development is called position-sensitive atom probe (POSAP). The aperture and single-ion detector of APFIM are replaced by a wide-angle double channel plate, with a position-sensitive anode just behind the plate. Field- or laser-evaporated ions strike the channel plate, releasing an electron cascade, which is accelerated toward the anode. The impact position is located by the division of electric charge between three wedge-and-strip electrodes. From this position the point of origin of the ion on the tip surface can be determined, since the ion trajectories are radial. Thus after many evaporation pulses, leading to removal of a volume of material from the tip, both the identities and the positions of all atoms within that volume can be mapped in three dimensions. Because an evaporated volume may contain many thousands of atoms, the data collection and handling capabilities must be particularly sophisticated. POSAP is the only technique available for identifying and locating precipitates, second phases, particles, and interfaces on an atomic scale, and has therefore had considerable application in metallurgical and semiconductor problems.

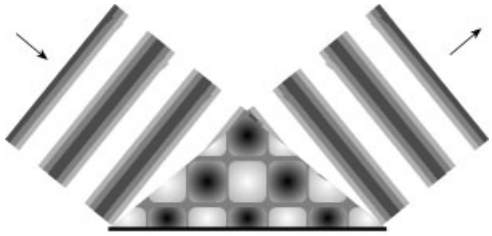

Figure 93. Interference of the incoming and the reflected X-ray waves in the triangular region above a flat and thick reflecting substrate. The strength of the electromagnetic field is represented on the gray scale with instantaneous crests (white) and troughs (black). In the course of time, the pattern moves from left to right [352].

4. Photon Detection

4.1. Total-Reflection X-ray Fluorescence Analysis (TXRF)

4.1.1. Principles

Less than two decades after the first practical application of X-ray fluorescence [331], Moseley established the relation between the reciprocal wavelength of emitted characteristic X-ray fluorescence (XRF) and the atomic number Z of the elements [332], [333]. However, the XRF arrangement was not sensitive enough for trace or ultratrace analyses. Compton reported in 1923 that the reflectivity of a flat target increases below a critical angle of 0.1° under conditions of total X-ray reflection. The high reflectivity of the sample support reduced the spectral background of the support and improved the detection limit (DL) down to picogram levels in the early 1970s, when Yoneda and Horiuchi applied the principle of TXRF mainly to ultratrace elemental microanalyses of biological samples [334]. For more details on the history of TXRF, see [335–337].

Since then, TXRF has become a standard tool of surface and subsurface microanalysis [338–342]. In 1983, Becker reported on the angular dependence of X-ray fluorescence intensities in the range of total reflection [343]. Further development in the field of TXRF are improved detection limits [344] in combination with subtle surface-preparation techniques [345], [346], analyte concentrations extended even to ultratraces

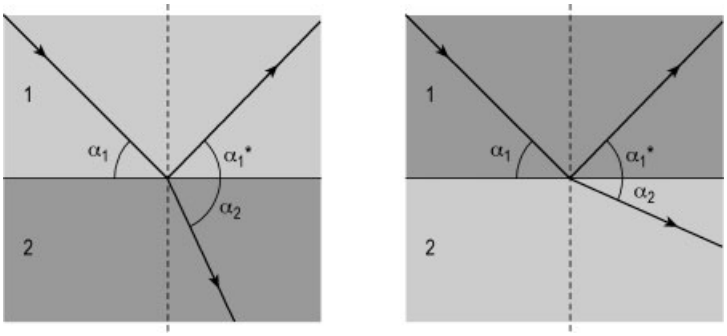

Figure 94. Incident, reflected, and refracted beam at the interface of two media 1 and 2. On the left, medium 2 is optically denser than medium 1 ($n_2 > n_1$); on the right, it is vice versa ($n_1 > n_2$) [352].

(pg) of light elements such as Al [347], speciation of different chemical states [348] and novel optical arrangements [349] and X-ray sources [350], [351].

For information on both the history and principles of and future trends in XRF, see → X-Ray Fluorescence Spectrometry. For TXRF, see the handbook by KLOCKENKÄMPER [352] and current reviews [353–355].

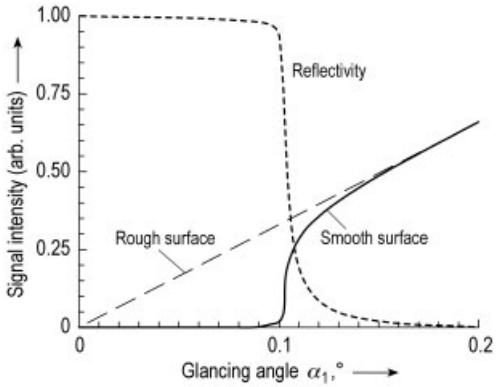

Figure 95. Signal intensity of a thick, flat, and smooth Si substrate (—-), calculated for an impinging $Mo_{K\alpha}$ beam. In addition, the reflectivity R (·····) is shown as function of the glancing angle α. Below $\alpha_{crit} = 0.102°$, total reflection occurs with a stepwise increase in reflectivity and a stepwise decrease in signal intensity. The oblique dashed line represents the intensity from a rough Si substrate [352].

The basis of XRF analysis is the photoelectric absorption and the subsequent emission of X-ray photons characteristic of the fingerprints of analyte atoms in the sample. Element composition can be quantified by the relative intensities of the individual element fingerprints. In a TXRF arrangement, however, the primary beam interacts with the surface analyte atoms as a standing wave field (Fig. 93) and excites surface atoms rather than the bulk, in contrast to XRF. The emitted intensity is directly proportional to the intensity of the standing wave field. In the subsurface, the primary beam generates an evanescent wave field. Hence, the characteristic fluorescence of analyte atoms is a response to these varying excitations by the standing and evanescent wave fields, and its dependence upon the glancing angle can be exploited for stratigraphy of parallel-layered structures and for distinguishing particulate from film-type surface contamination.

In X-ray spectrometry, attenuation, deflection, and interference must be considered. Attenuation is described by the well-known Lambert – Beer law and the mass attenuation coefficient as given for conventional XRF.

Practical X-ray energies do not exceed 100 keV. The primary beam is mainly attenuated by the photoelectric effect. Scattering, both elastic (Rayleigh) and inelastic (Compton), represents a minor contribution to attenuation at energies below 100 keV.

Reflection and refraction of X-rays follow the laws of optics (Fig. 94): The glancing angles of incidence (α_1) and reflection (α_1^*) are equal. The respective refraction angles follow SNELL's law for different phase velocities v in media 1 and 2:

$$v_2 \cos\alpha_1 = v_1 \cos\alpha_2 \tag{4.1}$$

When the glancing angle α_2 of the refractive beam becomes zero, the refractive beam runs tangential to the boundary surface. Below the

critical glancing angle of incidence, reflectivity increases to 100 % and the penetration, and hence information, depth is limited to a few nanometers. Reflection below the critical angle is defined as total reflection. The critical angle α_{crit} depends on the photon energy E and the atomic mass M of the medium, its atomic number Z, and its density ϱ (Eq. 4.2). The signal intensity is also a function of surface roughness.

$$\alpha_{\mathrm{crit}} \approx \frac{1.65}{E} \sqrt{\frac{Z\varrho}{M}} \tag{4.2}$$

At α_{crit} the penetration depth approaches a minimum, particularly in the case of reflective surfaces, such as chemo-mechanically polished Si. Total reflection disappears on rough surfaces (Fig. 95). Below α_{crit} penetration depth is in the range of a few nanometers (Fig. 96).

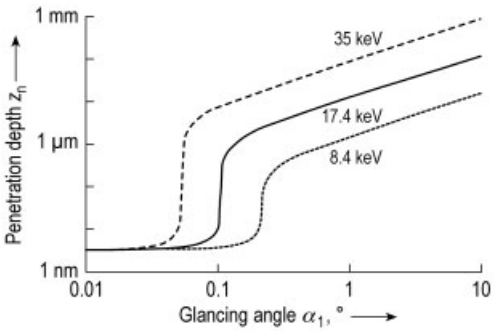

Figure 96. Penetration depth z_{n} of X-rays impinging on silicon at a variable glancing angles α_1. The curves were calculated for three different photon energies. The dashed vertical line represents the respective critical angle [352].

4.1.2. Instrumentation

The construction of a TXRF system, including X-ray source, energy-dispersive detector, and pulse-processing electronics, is similar to that of conventional XRF. In addition, the geometric arrangement must provide for total reflection of a monochromatic primary beam. The totally reflected beam interferes with the incident primary beam. This interference causes the formation of standing waves above the surface of a homogeneous sample, as depicted in Figure 93, or within a multiple-layered sample. A part of the primary beam fades away in an evanescent wave field in the bulk or substrate [359].

A TXRF instrument (Fig. 97) consists of an X-ray source, a low-pass filter, collimator, monochromator, sample holder, detector and an electronic registration unit. Various target materials (e.g., Mo, W, or Mo – W) alloys are used in fine-focus, sealed-anode X-ray tubes powered by a water-cooled 3.5 kW generator. Maximum permissible power for high-Z anodes is about 2 – 3 kW. During their lifetime of 2000 – 4000 h, they gradually lose intensity. Rotating anodes are warranted for 2000 h of operation at higher operating powers of up to 30 kW and more intense water cooling (15 L/min instead of 5 L/min).

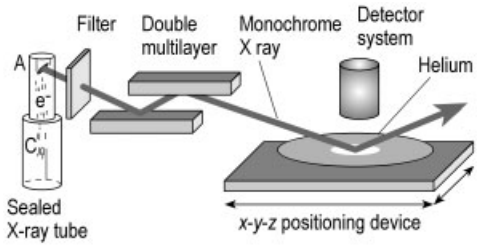

Figure 97. Schematic of a TXRF system equipped with a fine-focused anode

A double reflector is used as low-pass filter [340]. Tunable monochromators in combination with alloy anodes are also commercially available. For sensitive surface analyses monochromators are necessary. They are made of LiF, highly oriented pyrolytic graphite (HOPG), or multilayers of W – Si, W – C, or Mo – B_4C, which provide for intense and monochromatic radiation, preferentially $W_{L\beta}$ or $Mo_{K\alpha}$. The sample positioning device must operate at high geometrical reproducibility in all three dimensions, with automated sample load and unload stations. The sample must be in an evacuated or He-flushed chamber to avoid disturbing absorption by air.

The energy-dispersive (EDX) solid-state detector (SSD, Figs. 98 and 99) is made of lithium-drifted Si crystal [Si(Li)]. Between thin p- and n-type layer lies a high-resistivity Si crystal of centimeter dimensions. The front and end planes of the crystal are coated with Au and serve as electrodes. The crystal, cooled to 77 K by liquid

nitrogen, represents a $p-i-n$ diode (Fig. 99). An incident X-ray photon with an energy above 2 keV causes a cascade of photoelectron – hole pairs into which the initial photon energy is completely converted. Due to the reverse bias, electrons rapidly drift to the positively polarized n-type electrode, and holes to the negative p-type electrode. The number of electron – hole pairs is proportional to the energy of the impinging photon; hence, the charge pulse is a measure of the characteristic photon energy. Finally, the X-ray spectrum is established by a multichannel analyzer which distributes the pulses according to their height into a series of channels, each of which is reserved for a small range of pulse heights (i.e., energies), and adds them up to give the measured intensity.

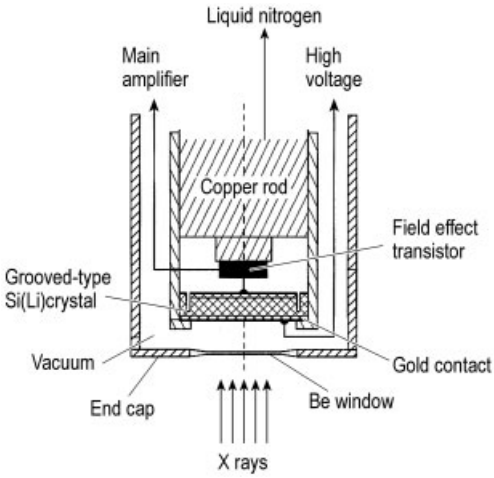

Figure 98. Section drawing of the front end of an SSD with a grooved Si(Li) crystal. Crystal and preamplifier are connected to a cooled copper rod and shielded by a case with an end cap and Be window [352], [360].

Optimal resolution, that is,low full width at half-maximum (FWHM), is a trade-off between high count rate, i.e., low dead-time, and good spectral resolution.

Silicon drift detectors (SDDs, Figs. 100 and 101) now also provide sufficient resolution (FWHM = 0.175 keV) for sample spots of size 2×2 to 100×100 mm, and enable high-speed operation ($> 10^5$ counts/s). SDD can be combined with microelectronics and applied in portable TXRF models for microanalytical applications [361]. They must be cooled by a Peltier element.

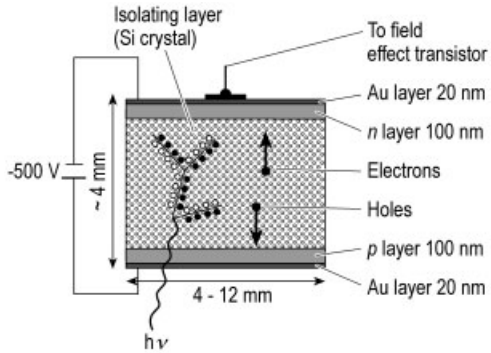

Figure 99. Semiconductor detector operated as a pin diode with a reverse voltage or bias. An incident X-ray photon ultimately produces a series of electron – hole pairs. They are swept out by the bias field of − 500 V: electrons in the direction of the n-layer, and holes in the direction of the p-layer. Thus, a small charge pulse is produced (after [361]).

4.1.3. Spectral Information

Spectral information from TXRF spectra is similar to conventional XRF-EDX spectra. The analyst must be always aware of potential spectral interferences that cannot be resolved with standard EDX detectors [347]. Spurious peaks can be assigned to escape peaks in energy dispersive spectra. An escape peak emerges when intense fluorescence of sufficiently high energy (giving rise to the *mother peak*) impinges on the SSD, expelling, as usual, photoelectrons from Si core shells. Excited Si atoms emit X-ray photons that are usually reabsorbed by the crystal. Photoelectrons generated by these photons initiate additional electron – hole cascades (i.e., charge pulses) in the same way as the sample fluorescence. However, photons near the SSD crystal surface also escape without producing a pulse, thus carrying off the difference between the energy of the mother peak and the generated core photoelectron. This "escaped" energy gives rise to a separate, low-energy *daughter peak*. Some elements whose Kα peaks interfere with escape peaks of other elements in SSDs or SDDs are listed in Table 8.

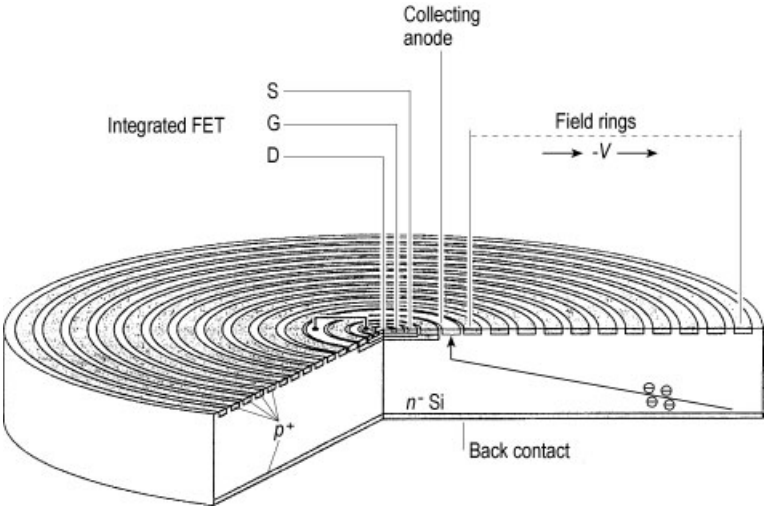

Figure 100. Schematic view of a cylindrical silicon drift detector (SDD) [361]

Table 8. Some elements whose Kα peaks interfere with escape peaks of other elements in SSDs or SDDs [347]

Elements	Escape peaks
V	$Co_{K\alpha}$, $Mn_{K\beta}$, $Dy_{L\alpha}$, $Ho_{L\alpha}$, $Er_{L\alpha}$, $Gd_{L\beta}$
Cr	$Co_{K\alpha}$, $Fe_{K\beta}$, $Er_{L\alpha}$, $Tm_{L\alpha}$, $Tb_{L\beta}$, $Dy_{L\beta}$
Mn	$Ni_{K\alpha}$, $Co_{K\beta}$, $Yb_{L\alpha}$, $Lu_{L\alpha}$, $Hf_{L\alpha}$, $Ho_{L\beta}$
Fe	$Cu_{K\alpha}$, $W_{L\alpha}$, $Ta_{L\alpha}$, $Hf_{L\alpha}$, $Ni_{K\beta}$, $Tm_{L\beta}$, $Er_{L\beta}$
Co	$Zn_{K\alpha}$, $W_{L\alpha}$, $Cu_{K\beta}$, $Yb_{L\beta}$, $Lu_{L\beta}$, $Re_{L\alpha}$, $Os_{L\alpha}$
Ni	$Ga_{K\alpha}$, $Pt_{L\alpha}$, $Ir_{L\alpha}$, $Ta_{L\beta}$, $Hf_{L\beta}$
Cu	$Zn_{K\beta}$, $Au_{L\alpha}$, $W_{L\beta}$, $Hg_{L\alpha}$, $Ge_{K\alpha}$

Spurious peaks can also appear by energy doubling if the charge collection induced by the first fluorescence photon is not complete when a second photon from the same fluorescence source impinges. Then, the detector registers twice the number of charges, and a sum-up peak results. Sum-up peaks can be reduced by a pulse pile-up rejector (PUR). Contamination along the beam path, e.g., Fe in the detector window (Be), may also result in spurious peaks and limit the detection capability for certain elements [362].

4.1.4. Quantification

Reliable quantification is based on peak-search software that combines peak localization, peak identification, and element deduction. Element deduction means that, for unambiguous detection, at least two of the principal peaks be detected for each analyte of interest. In trace analysis, only the strongest α peaks can be detected, and special attention must be paid to interfering satellites and spurious peaks.

By using an optically flat plate of a pure element such as Ni, TXRF systems can be calibrated absolutely [363]:

$$I_X = B_X c_X \tag{4.3}$$

where I_X is the background-corrected net intensity of the principal peak of analyte X, B_X a proportionality factor relative to the absolute sensitivity of the standard reference (e.g., an Ni plate), and c_X the concentration of X. Multielement analyses are based on known relative sensitivities S:

$$S_j = \frac{I_j / c_j}{I_{ref} / c_{ref}} S_{ref} \tag{4.4}$$

where the suffix ref denotes reference values. Values of relative sensitivities can be determined by theoretical calculations including, for example, mass absorption [364]. The fluorescence yield is angle-dependent. Because matrix effects are absent in the case of minute sample amounts ($< 1\,\mu g/cm^2$), quantification by internal standardization is possible. Equation (4.4) holds correspondingly.

The bulk-type response curve also depends on surface roughness [365]. Therefore, reference materials must be carefully investigated by angle scan prior to use. Angle-scan characteris-

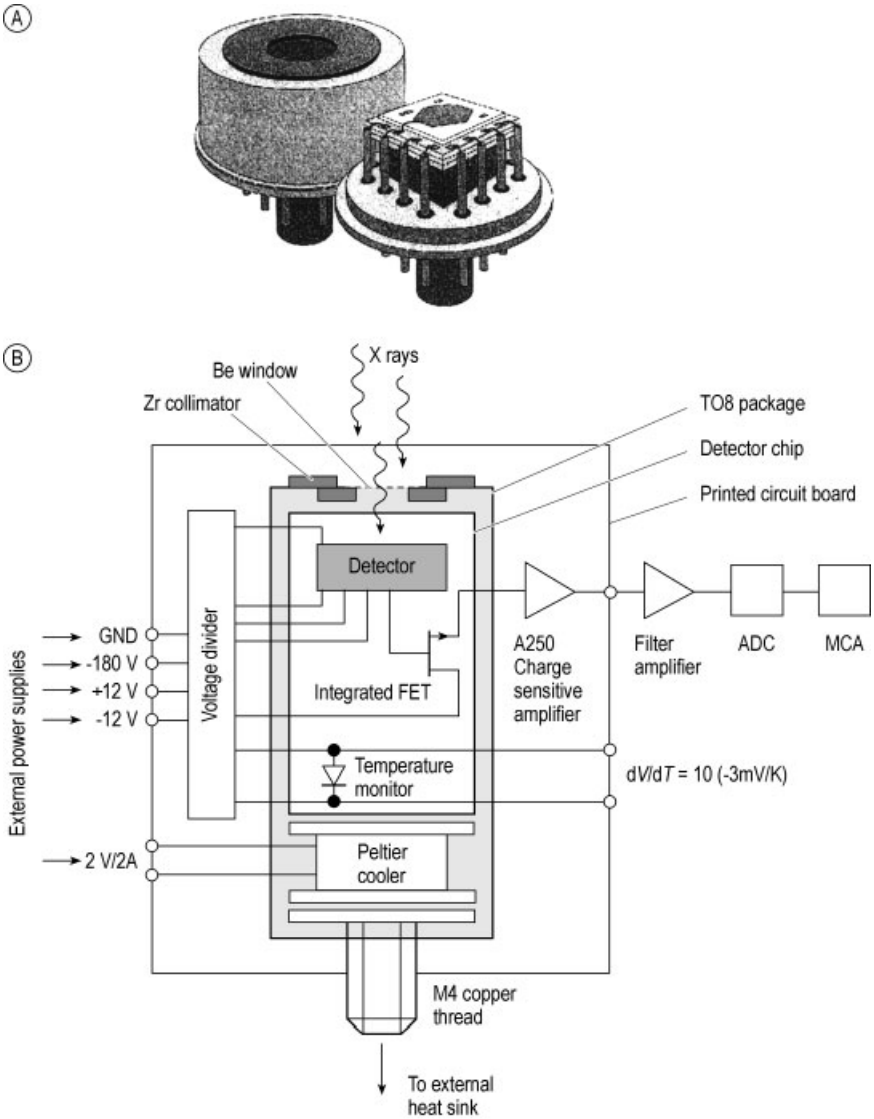

Figure 101. Block diagram (A) and assembly image (B, also with cap removed) of a silicon drift detector module, including electronics.

tics of the sample, that is, recording the fluorescence intensity at more than one glancing angle near α_{crit}, should not deviate from the reference. The measurement must be carried out under similar optical conditions.

With monochromatic $\mathrm{Au}_{\mathrm{L}\beta}$ (11.44 keV) excitation, the variation of Cu depth distribution in monocrystalline Si with time was studied by means of TXRF [366]. Such and similar effects must be carefully investigated when using reference samples of unknown behavior. Complementary analytical techniques may be helpful. However, the propagation of errors must be taken into consideration; and, moreover, the calibration must cover the concentration range of interest [367] because the accuracy of reference concentrations strongly affects the accuracy of the results [368], [369]. External calibration with reference materials can yield deviations of 3 – 20 % from the true value [370].

However, reliable preparation of controlled spiking has been reported [371]. For general issues of external calibration of TXRF systems, see [372] and [373].

4.1.5. Applications

TXRF is an ideal tool for microanalysis [352], since it has a broad range of linearity (10^{13} – 10^{8} atoms/cm^2) and is extremely surface sensitive and matrix-independent. TXRF can be applied to a great variety of different organic and inorganic samples, such as water, pure chemicals, oils, bodily fluids and tissues, and suspended matter, down to the picogram range.

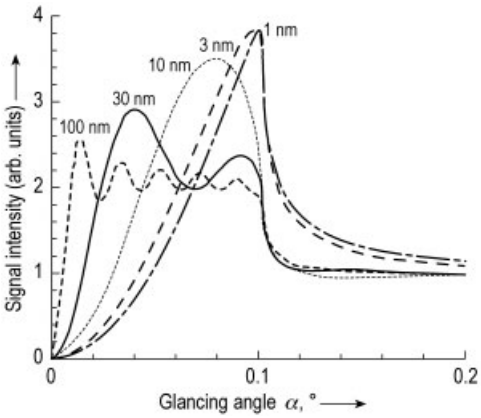

Figure 102. Fluorescence signal of small particles or thin films deposited on a silicon substrate as sample carrier. The intensity was calculated for particles, thin films, or sections of different thickness but equal mass of the analyte, and plotted against the glancing angle α. A $\mathrm{Mo}_{\mathrm{K}\alpha}$ beam was used for excitation. Particles or films more than 100 nm thick show double intensity below the critical angle of 0.1° [352].

4.1.5.1. Particulate and Film-Type Surface Contamination

The penetration depth of impinging X-rays is very limited under the condition of total reflection. For light substrates such as Si, quartz, and poly(methyl methacrylate), the spectral background of TXRF is six orders of magnitude smaller than that of XRF. The fluorescence signal arises from the uppermost layer; therefore, the scattered intensity, that is, background, is lower than in a conventional XRF arrangement. If metallic smear, organic tissues, or thin (< 100 nm), dried residues of a solution, suspension or dispersion are present on a substrate within the standing wave front, strong oscillation of signal intensity results (Fig. 102). Such oscillations vanish when grain size or film thickness increase to about 1000 nm [374].

One can clearly distinguish thick-film from particulate analytes by means of angle characteristics (cf. curves b and c in Fig. 103). The particulate and thin-layer angle characteristics do not differ at large angles [375], [376]. However, surface roughness crucially modifies peak height [365] and may simulate a stratified structure [377].

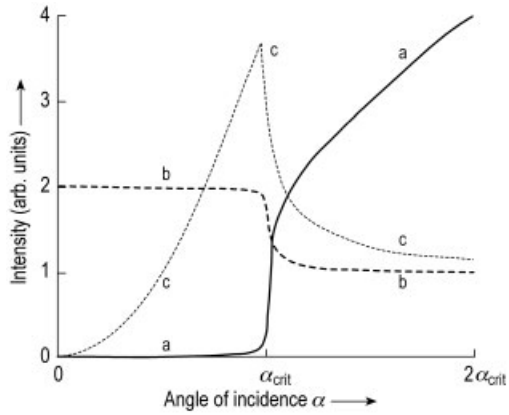

Figure 103. Characteristic intensity profiles for three different kinds of concentrations
a) Bulk; b) Particulate; c) Thin-layer
The critical angle α_{crit} is determined by total reflection at the substrate ([352], after [375])

4.1.5.2. Semiconductors

Metrology and contamination analysis in particular have been decisive factors for profitable semiconductor production [378]. Semiconductor applications of TXRF go back to the late 1980s and were introduced by Eichinger et al. [379], [380]. Due to its high sensitivity, wide linear range, facile spectrum deconvolution, and ease of calibration, TXRF has rapidly become the economic workhorse of monitoring cleaning efficiency and front-end wafer cleanliness

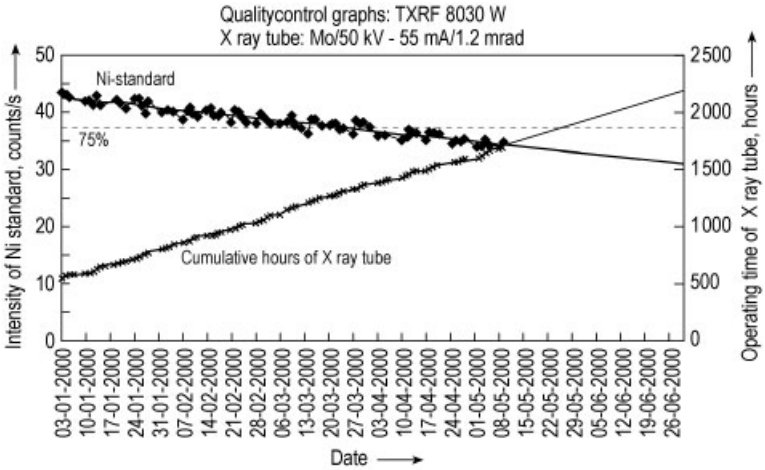

Figure 104. Statistical process control chart for TXRF measurement systems. The sensitivity is monitored by daily calibrations with a Ni standard reference sample. When the X-ray intensity sinks below 75 % of the original intensity, the tube must be replaced with a new one.

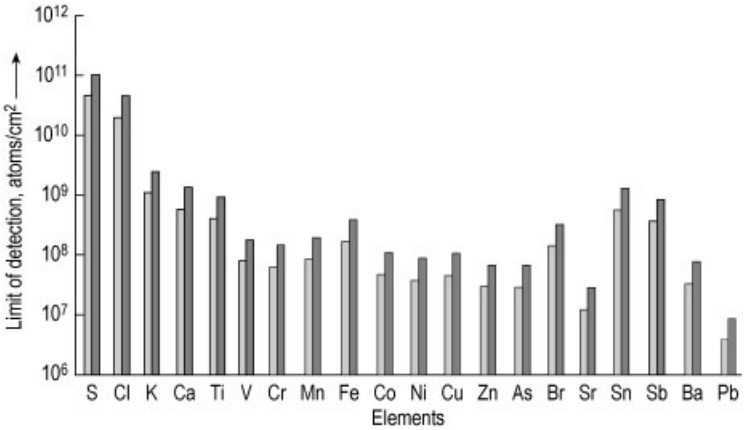

Figure 105. Limits of detection (3 σ) for VPD-TXRF measurements on 200- and 300-mm Si wafers

[381], [382]. TXRF provides reliable quality-control data that can be easily integrated into the data-management system [383], [384]. Automatic operation with repeated one-point calibration every 8 h, as well as automatic data management and transfer, has led to abundant applications of TXRF (Fig. 104).

Detection limits for various elements by TXRF on Si wafers are shown in Figure 105. Synchrotron radiation (SR) provides a bright and horizontally polarized X-ray excitation that is tightly collimated and diminishes the Compton scattering of silicon. Developments in the field of SR-TXRF and extreme-ultraviolet (EUV) lithography promise improved sensitivity down to the range of less than 10^7 atoms/cm^2 [344], [385–388]. At present, practical applications of SR-TXRF focus on fast mapping of wafers up to a diameter of 300 mm and reference measurement capabilities in the range of $10^5 - 10^7$ atoms/cm^2 [389].

4.1.5.2.1. Depth Profiling by TXRF and Multilayer Structures

Buried layers are important components of microcircuits, and TXRF is a sensitive microanalytical tool for inspection of buried layers

(Fig. 106). On a thick substrate, the angle characteristics is a function of layer thickness. [390], [391]. First-principle calculations have been reported on model-free TXRF of stratified structures [392].

Shallow doping profiles, particularly those of As, require nanoscale information on dopant distribution. Though SIMS can be reliably applied for layers below 5 nm depth, the technique of choice for the upper subsurface is angle-dependent TXRF [393]. Using angle-scan TXRF allows stratified microstructures to be accurately analyzed for element composition, layer thickness, and density (Fig. 107) [372]. A combination of TXRF with layer-by-layer chemical etching provides reproducible results on the stratigraphy of metallic implants [346], [394–396].

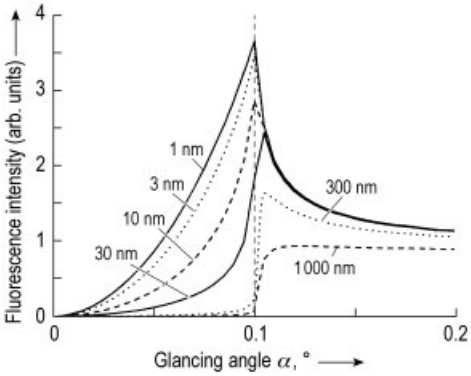

Figure 106. Fluorescence intensity from layers buried in a thick substrate. The intensity as a function of the glancing angle was calculated for layers of different thickness but with a constant area density of the analyte. Silicon was assumed as substrate and $Mo_{K\alpha}$ X-rays as primary beam. Total reflection occurs below 0.1°. Without total reflection, the dashed horizontal line would be valid in all cases [352].

4.1.5.2.2. Vapor Phase Decomposition (VPD) and Droplet Collection

Vapor-phase decomposition and collection (see Figs. 108, 109, 110) is a standardized method of silicon wafer surface analysis [342]. The native oxide on wafer surfaces readily reacts with isothermally distilled HF vapor and forms small droplets on the hydrophobic wafer surface at room temperature [397]. These small droplets can be collected with a scanning droplet. The scanned, accumulated droplets finally contain all dissolved contamination in the scanning droplet. It is dried on a concentrated spot (diameter ca. 150 µm) and measured against the blank droplet residue of the scanning solution [398–400]. VPD-TXRF has been carefully evaluated against standardized surface analytical methods. Care must be taken to use reliable reference materials [401–403].

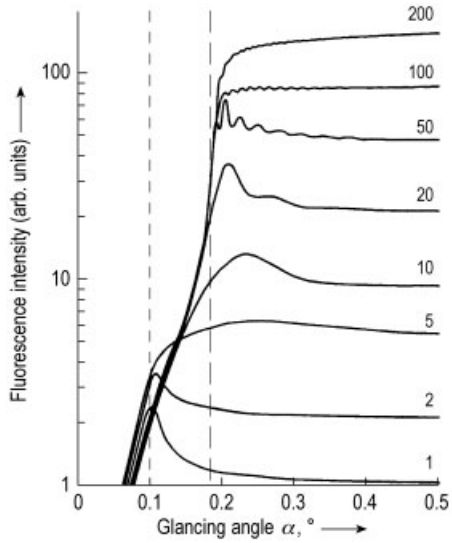

Figure 107. Angular dependence of the fluorescence radiation emitted from a Co-layered Si substrate. The $Co_{K\alpha}$ intensity is plotted semilogarithmically for layers of different thickness (curve parameter in mm). The maxima for the ultrathin Co layers are localized at the critical angle of Si (dashed vertical line). They are shifted to the critical angle of Co (dotted vertical line) if the layer is more than 10 mm thick ([352], after [372]).

The accuracy of VPD-TXRF depends largely on the reliability of the reference material and on the reproducibility of the droplet-drying procedure (Fig. 109) [406], [407]. Mass absorption sets the upper limit of reliable analyses at 10^{14} atoms/cm^2 [408].

VPD-TXRF is also a facile technique of interface analysis [409], [410]. Automated VPD equipment (Fig. 108) improves both the detection limit (upper range of 10^7 atoms/cm^2) and the reliability (by $> 50\,\%$) of the VPD-TXRF

Figure 108. Wafer surface preparation system (WSPS): automated VPD system for wafer diameters from 100 to 300 mm

measurement [345]. Current research focuses on issues of sample holders [411], [412] and light-element detection capability [413–415].

4.2. Glow Discharge Optical Emission Spectroscopy (GD-OES)

Glow discharge optical emission spectroscopy (GD-OES) is becoming one of the most important techniques for the direct analysis of solids at their surface, through their depth and in their bulk. The ease and speed of operation, high sensitivity to all elements, accuracy of analysis, and breadth of application make GD-OES a very versatile and powerful technique. GD-OES has proven its capability, in particular, for the rapid and reliable analysis of metallic coatings, oxide scales, and diffusion and alloy layers, and can thus be used for quality control of technical products during different stages of heat treatment, such as nitriding or nitrocarburizing steps, and for hard coating and PVD processes.

4.2.1. Principles

Glow discharges are low-pressure electrical discharges in a noble gas at pressures of $10^2 - 10^3$ Pa [416]. The glow discharge in the discharge chamber consists of several alternating dark and luminous zones. The complicated structure may be altered by varying the geometry of the chamber, especially the distance between the two electrodes. In general, when the electrode distance is relatively long, the positive column (the bright zone near the anode) is the most prominent of the zones in the discharge. When the space between the electrodes is reduced, the positive column shrinks and finally disappears, while the negative glow (the bright zone near the cathode) and the cathode dark space are hardly affected. Both negative glow and cathode dark space are essential to maintain the discharge and cannot be dispensed with. In this case, the discharge is called an obstructed glow and can be realized when the interelectrode separation is just a few times the thickness of the cathode dark space.

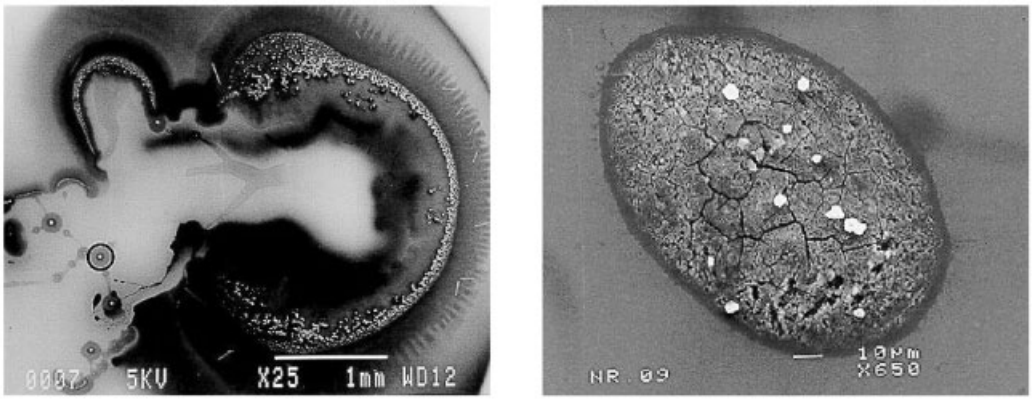

Figure 109. Dried VPD droplets. Left: worst case, the VPD solution exploded under fast drying with an infrared lamp, droplet size of a few mm; Right: best case (WSPS), VPD solution dried under controlled conditions with vacuum and carrier gas [367].

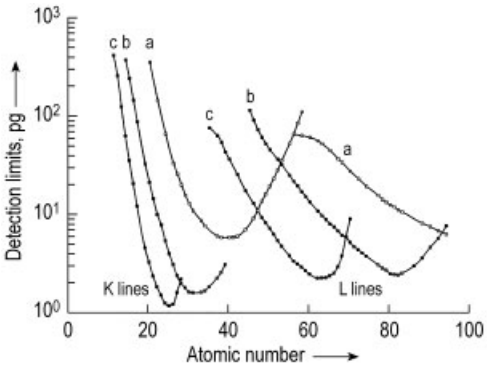

Figure 110. Detection limit of TXRF for the residues of aqueous solutions as a function of the atomic number of the analyte element. Three excitations modes were used: a) W tube; 50 kV; Ni filter, cutoff 35 keV; b) Mo tube; 50 kV; Mo filter, cutoff 20 keV; (c) W tube, 25 kV, Cu filter. The three curves to the left were determined by the detection of K peaks; the three curves to the right, by that of L peaks ([352], after [404], [405]).

Discharge tubes working under obstructed-glow conditions are employed extensively as excitation sources for optical emission spectrometry. For this purpose the analytical sample of interest serves as the cathode. During the discharge, the cathode is bombarded by positive ions formed from the noble gas. This bombardment produces a surface erosion called cathodic sputtering: atoms, electrons, and ions are removed from the surface of the material constituting the cathode. This sputtered material then participates in different phenomena in the discharge. The result is an emission of photons corresponding to the characteristic spectral lines of the elements in the plasma. Thus the glow discharge provides information on the elemental composition of the analytical material being removed layer-by-layer from the cathode. The sputter rates vary from 0.1 nm/s up to more than 100 nm/s, depending on the cathode material and the excitation conditions used. Therefore, the glow discharge provides a rapid and detailed characterization of near-surface layers, interfaces, and bulk composition when the spectral line intensities of the elements of interest are recorded simultaneously.

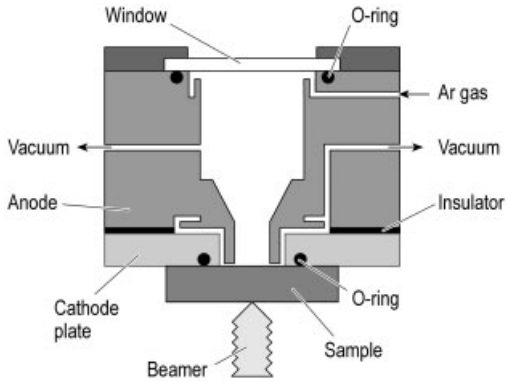

Figure 111. Diagram of a typical glow discharge device used for GD-OES depth profiling [422]

4.2.2. Instrumentation

Glow Discharge Sources. If the potential applied across the cathode and the anode is constant (dc) then conductive materials can be analyzed; if the potential is varied at radio frequency (rf) then both conductive and nonconductive materials can be analyzed.

The most important dc-discharge source used to date in GD-OES is the *Grimm-type source* [417], the principle of which is illustrated in Figure 111. In the Grimm source the anode is no longer a flat plate but a hollow tube, the annular face of which approaches to within 0.1 – 0.2 mm of the surface of the sample. This restricts the discharge to a sample area equal to the aperture of the anode tube (typically 4 – 8 mm). The sample is mounted against the discharge chamber and acts as cathode. When the discharge current is sufficiently high, in the so-called *abnormal glow regime*, the whole area is covered by the discharge, very efficient cathodic sputtering results, and the material of the sample is easily removed. To maintain the required low pressure, the sample must be vacuum tight and sealed to the discharge chamber by an O-ring. Normally, flat samples with sufficient low surface roughness are used, although samples with other regular shapes (rods or wires) can also be used with special sample holders and a modified front plate of the discharge chamber. The chamber is evacuated, usually by a rotary pump, and flushed continuously by a noble gas (generally argon). The pressure inside the source during sputtering is typically 300 – 1300 Pa. The optical radiation of the excited sample atoms is observed end-on through a window with the required transmittance; MgF_2 is used most commonly as window material.

Other well-known discharge sources include the boosted sources, which use an additional electrode to achieve a secondary discharge, a magnetic field or microwave power to enhance the efficiency of excitation and thus the analytical capabilities, but none of these sources has been applied so far to surface or depth profile analysis.

The more recent Marcus rf source [418] is operated at pressures similar to those of dc discharge sources, but the essential characteristic is the conventional 13.56 MHz radio-frequency power which is applied to the back of the sample by an impedance matching device. For proper operation the sample must be electrically isolated from all metallic parts of the source and ancillary components to reduce rf radiation and conduction losses. In this way the sample no longer acts as cathode, and the second electrode or counterelectrode (corresponding to the anode in the dc device) is no longer a tube, but is represented by the metallic walls of the discharge chamber. The sample area exposed to the discharge is confined by an orifice disk and is much smaller than the counterelectrode, so that most of the drop in rf potential occurs at the sample and no significant sputtering of the counterelectrode occurs.

The most important difference between dc and rf discharge source is the ability of the rf source to sputter both electrically conductive and nonconducting materials. This enables the elemental analysis of insulating samples directly in the solid state.

Continuous discharge and sputtering of the sample are accomplished by the negative dc self-bias voltage, or offset, which is acquired at the surface of the insulator without any applied external dc potential. This phenomenon is explained by the different mobility of positive ions and electrons, which reverse their direction of movement toward the sample with every half-cycle of the bipolar rf potential. The positive charge of the insulator surface during bombardment with ions is readily neutralized and shifted to negative charge by electrons in the next half-cycle, and this results in the observed steady-state negative offset voltage and hence the effective sputtering of the sample by positive ions.

Spectrometer. Apart from the glow discharge source with its associated gas and power supplies, vacuum pumps and controls, the optical emission spectrometer is the most important part of a GD-OES instrument. To make use of the analytical capability of the discharge source, the spectrometer must have a sufficient spectral resolving power and an adequate spectral range. Depending on the application, a commercial spectrometer can cover the wavelength range from 110 to 800 nm, which contains the most sensitive lines of all elements, including those of the light elements (Li, Be, B, C etc.)

and gases (H, O, N, Cl). The high spectral resolution is required to avoid spectral interference with lines of other elements, in particular of the discharge gas argon.

The most widely used spectrometer designs are the Paschen – Runge polychromator configuration for simultaneous spectrometers and the Czerny – Turner monochromator for sequential analyzers. In the Paschen – Runge mount the concave grating focuses the spectral lines of preselected elements onto fixed exit slits, which are positioned on a Rowland circle. Up to 64 elements can be detected simultaneously in this way. In the Czerny – Turner mount quasi-monochromatic radiation is observed at the exit slit. By turning the plane grating around its axis it is possible to scan a spectral range and to profile an emission line, including background radiation and neighboring spectral lines.

Signal Acquisition. The optical radiation passing the exit slit(s) of the spectrometer is recorded by a detector system. The best known detector is still the photomultiplier tube (PMT) which is implemented in most commercial spectrometers. The rapid development in the performance of CCD and similar detection techniques makes this detector system very attractive in the future, especially when combined with the well known Échelle spectrometer. This configuration enables a small-sized spectrometer with high spectral resolving power and high flexibility in line selection.

In the common polychromator system, each detection channel is equipped with a PMT and the corresponding high-voltage supply. Each supply can be adjusted individually to give the appropriate sensitivity for each element required to match the analytical conditions. Due to the linear characteristic of the PMT, the output signal is proportional to the radiation power of the spectral line recorded by each detection channel. The analogue output signals are amplified, converted into digital information and transferred to a PC for further data processing. In case of depth profiling and continuous sample erosion it is necessary to use fast A/D converter, since the acquisition rate must be high (ca. 1 kHz) to ensure a sufficient depth resolution.

4.2.3. Spectral Information

The emission process in a glow discharge source depends on several factors: the number of atoms in the plasma, the availability of energetic particles such as high-energy electrons and metastable argon atoms which provide the excitation energy by numerous collision processes, and the excitation cross sections of the competing energy levels of the elements. It can be estimated that the number of argon atoms in the discharge exceeds the number of sputtered sample atoms by a factor of about 10^3. Since only a small fraction of sputtered atoms is charged and energetically excited, the discharge plasma for a given sample material is dominated mainly by argon atoms. This fact explains why calibration curves obtained with GD-OES are mostly linear over a wide range of elemental concentration. The operating conditions of the discharge, i.e., low pressure and low power input, result in low kinetic temperatures of about 500 – 1000 K and small physical line widths of emission lines. Therefore, the spectral information obtained from the glow discharge source is highly specific and selective. With constant excitation conditions, the spectral line intensities are proportional to the number densities of sputtered atoms in the plasma and hence the element concentration in the sample used. To maintain reproducible excitation conditions in the glow discharge source, the working parameters (e.g., argon pressure, dc current or rf power) are carefully controlled.

4.2.4. Quantification

Due to the complex nature of the discharge conditions, GD-OES is a comparative analytical method, and standard reference materials must be used to establish a unique function relating the measured line intensities and the elemental concentration. In quantitative bulk analysis, which has been developed to very high standards over decades, calibration is carried out with a set of calibration samples of compositions similar to the unknown samples. Normally, a major element is used as reference, and the internal-standard method is applied. This approach is not generally applicable to depth-profile analysis, since the different layers encountered in

a depth profile often represent widely different types of material, which means that a common reference element is not available.

The quantification algorithm most commonly used in dc GD-OES depth profiling is based on the concept of *matrix-independent emission yield* R_{ik} [419]. In this approach the observed line intensity I_{ik} is described by the concentration c_i of element i in the sample j and by the sputtering rate q_j:

$$I_{ik} = c_i q_j R_{ik} \tag{4.5}$$

The sputtered mass δm_i of element i during time increment δt is given by:

$$\delta m_i = I_{ik} \delta t / R_{ik} \tag{4.6}$$

The emission yield R_{ik} is defined as the radiation of the spectral line k of an element i emitted per unit sputtered mass and must be determined independently for each spectral line. The quantities q_j R_{ik} are derived from a variety of different standard bulk samples with different sputtering rates. In practice, both sputtering rates and excitation probability are influenced by the working conditions of the discharge. The systematic variation of discharge voltage U_g and current I leads to the empirical intensity expression [420]:

$$I_{ik} = K_{ik} c_i C_Q I^{A_k} f_k (U_g) \tag{4.7}$$

where K_{ik} is an atom- and instrument-dependent constant characteristic of the spectral line k of element i; A_k is a matrix-independent constant characteristic of the spectral line k; $f_k(U_g)$ is a first- to third-degree polynomial, also characteristic of the spectral line k; and C_Q is a constant related to the probability of a sample atom being ejected during the sputtering process. The emission-yield happroach as proven to be extremely successful for a large and increasing number of applications and has been implemented in commercial GD-OES depth profiling instruments.

4.2.5. Depth Profiling

The primary information obtained in GD-OES depth profile measurements is the relative intensity from the elemental detection channels as a function of sputtering time. The intensity versus time curves obtained for different elements can be converted into concentration versus depth curves by applying Equations (4.5) – (4.7) and the sum normalization of all concentrations to 100 %. The depth is determined from the sputtered mass, that is, the quantity obtained by the emission-yield technique. The density of composite material is calculated by an approach which is based on a weighted average of the density of pure elements. This method gives very accurate results for all types of metal alloys, but tends to be less accurate in the case of compounds which contain light and gaseous elements (oxides, nitrides, carbides, etc.). In general, accuracy, precision, and repeatability of quantitative GD-OES depth-profile analyses are assured on the basis of interlaboratory round-robin tests [421], by layered certified reference materials [422], and by comparison with other surface techniques [423–425].

The information depth achieved by GD technique is determined, in principle, by the depth of penetration of the incident ions, which is in the range of a few nanometers at the relatively low energies employed in the discharge. The practical depth resolution, however, is larger and is determined by several effects which are due to the sputtering process (preferential sputtering, atomic mixing), the sample properties (surface roughness, polycrystalline structure), and most seriously by the nonuniform erosion of the sample material. Most of these effects are also inherent in other techniques using sputtering for surface and depth-profile analysis. In practice, the depth resolution of GD-OES depth profile measurements is roughly proportional to the sputtered depth, and usually deteriorates from the nanometer range obtained at near-surface layers to micrometer range at depths of several micrometers [426], [427]. The lateral resolution of the GD-OES technique, on the other hand, is restricted by the size of the sputtered area of the sample surface (usually 4 – 8 mm in diameter) and is much larger than with other surface techniques. Both lateral and depth resolution of GD-OES are usually adequate, however, for the chemical characterization of technical products.

4.2.6. Applications

In comparison to other surface analytical techniques, GD-OES has the advantage of easy use,

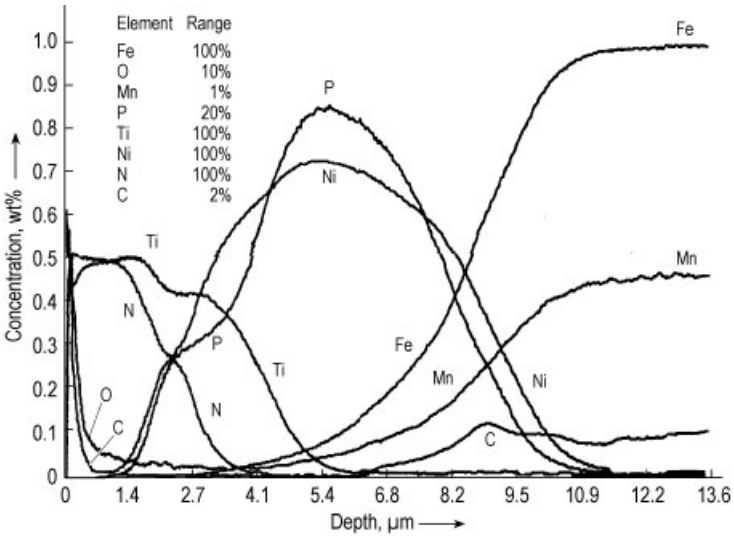

Figure 112. Quantitative depth profile of a TiN – NiP coating on steel [431]

rapid in-depth analysis, wide depth range, and sensitivity, for all elements. Therefore, GD-OES depth profiling is of primary interest for industrial use, especially for quality control of large-scale technical products. The numerous applications cover the field of different coating technologies of metallic surfaces, the investigation of corrosion effects, and reactive diffusion processes in layer – bulk interfaces [428].

In recent years the applications of GD-OES concentrated mainly on steel analysis and the qualitative characterization of surfaces of cold- and hot-rolled steels in order to control enrichment or depletion of minor and trace elements during the manufacturing process or the formation of oxide scales [423], [429]. Zinc-based coatings are of great industrial importance in materials for buildings and for car manufacturing. Because of the wide industrial application of hot-dip galvanizing and electroplating technologies, the rapid and reliable determination of the thickness of the zinc coatings (typically 5 – 30 µm) and the composition of major elements in these layers [420], [430] is a great challenge to surface- and depth-profiling methods. GD-OES has proven its capability in this field (Fig. 112 [431]) and is accepted as a standard method [432]. Other typical applications of GD-OES include the analysis of surface hardened steel after the process of nitriding or carburizing [433], and after different processes of hard coatings, i.e. physical vapor deposition (PVD), of TiN layers [434].

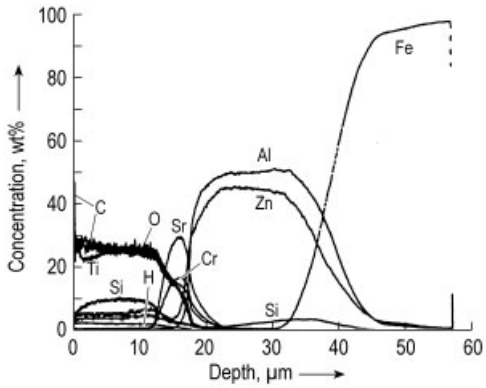

Figure 113. Quantitative depth profile of a commercial pre-painted metal-coated steel sample [435].

Besides the analysis of relatively thick metallic coatings and diffusion profiles, GD-OES is also applied successfully to the analysis of thin protective layers (thickness typically < 100 nm) such as phosphate and chromate layers on steel [420] or surface films on aluminum [435].

The examples given so far represent basic applications of GD-OES depth profile analysis with the dc glow discharge source. The rf source has unsurpassed capability when composite materials such as ceramics, coated glass surfaces, and polymer coatings are analyzed. A quantitative depth profile on pre-painted metal-coated steel is shown as an example in Figure 113 [436].

4.3. Surface-Sensitive IR and Raman Spectroscopy; Ellipsometry

4.3.1. Reflection – Absorption IR Spectroscopy (RAIRS)

Reflection – absorption infrared spectroscopy is a vibrational spectroscopic method that is frequently applied in the research areas of chemical, physical, and biological research [437–440]. The technique is well established for the identification and characterization of chemical state and structure of molecules and thin films adsorbed on metallic and non-metallic surfaces. Vibrational spectra are used as characteristic fingerprints for adsorbate molecules, adsorption configurations, and structures. The method is known under various acronyms: RAIRS, IRAS, FT-IRAS, IRRAS, and ERIRS (see Chap. 1), whereby the different names refer to the same principle of measurement using infrared radiation and a reflection – absorption geometry. This technique involves a single external reflection at a substrate covered by the sample, whereby the molecules of the sample absorb radiation at their vibrational frequencies and corresponding absorption bands occur in the spectrum of the reflected radiation. Substrates are chosen which do not show disturbing absorption bands in the spectral range of interest.

4.3.1.1. Principles

Conventionally, RAIRS has been applied to qualitative and quantitative characterization of adsorbed molecules or films on mirrorlike (metallic) substrates [437]. In the 1990s its applicability for quantitative analysis of adsorbates on nonmetallic surfaces (e.g., semiconductors, glasses [439], and water [441]) has been proven, too. However, the dielectric properties of the substrate and the adsorbate determine the optimum condition under which the RAIRS experiment should be carried out. The excitation of vibrational adsorbate modes on metals is governed by a *surface-selection rule*. The amplitude and phase of the reflected radiation depend on the direction of the electric field vector, which is composed of one component parallel to the reflection plane (p-polarized) and another normal to the reflection plane (s-polarized). Considering Fresnel equations, only the p-polarized radiation shows a significant resulting electromagnetic field near the metal surface for high incident angles of radiation near grazing incidence (at ca. 80°). In contrast; the mean-square electric field of radiation polarized in the surface plane is negligible because of the 180° phase shift after reflection at the metal. Therefore, only vibrations with dipole components perpendicular to the surface can be excited efficiently. With p-polarized radiation and incident angles near grazing incidence an increase in signal intensities by a factor of about 17 can be achieved in comparison to transmission experiments [437].

RAIRS spectra show structures of absorption bands, related to vibrations of the bulk or adsorbed molecules. From reflectance spectra the absorbance is commonly determined by calculation of $-\log(R_s/R_o)$, where R_s is the reflectance from the adsorbate-covered substrate, and R_0 the reflectance from the bare substrate. For thin films with strong dipole oscillators the Berreman effect, which leads to an additional spectral feature in the reflectance, must be considered (see Section 4.3.3). The frequencies, intensities, full widths at half maximum, and lineshapes of bands in the absorption spectrum yield information about adsorption states, chemical environment, ordering effects, and vibrational coupling. For example the structure, orientation, and reaction kinetics of alkoxides on Cu(111) surfaces have been determined [442].

In case of films on nonmetallic substrates (semiconductors, dielectrics) the situation is much more complex. In contrast to metal surfaces, both parallel and perpendicular vibrational components of the adsorbate can be detected. It has been shown that sign and intensity of RAIRS bands strongly depend on the angle of incidence, the polarization of the radiation, and the orientation of vibrational transition moments. Figure 114 shows IR reflection spectra

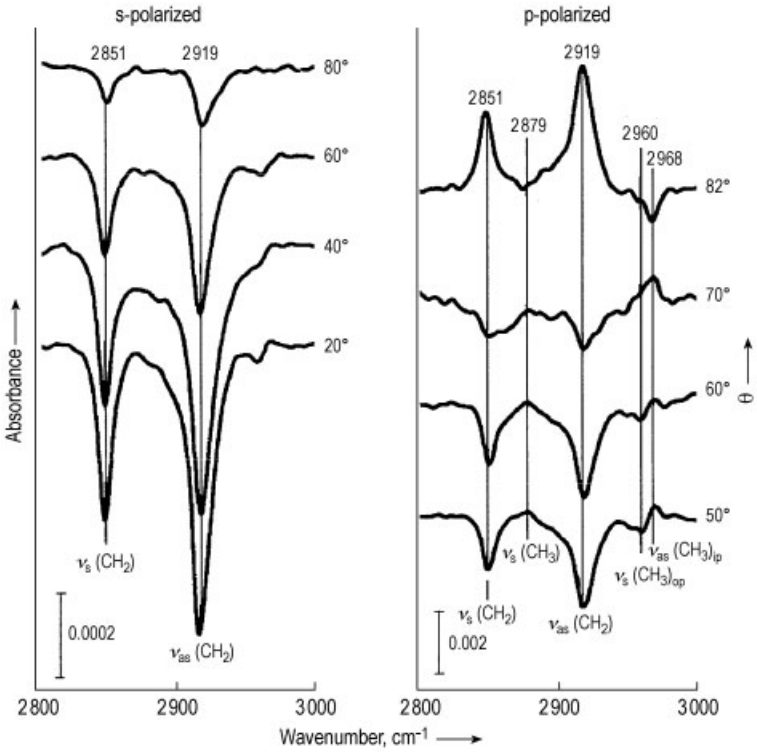

Figure 114. IR reflection spectra of an ODS monolayer on silicon for s- and p-polarised radiation at different incident angles θ. Symmetric (s), asymmetric (as), in plane (ip) and out-off plane (op) CH_2/CH_3 vibrations are marked [439].

of an octadecylsiloxane (ODS) monolayer on a silicon substrate [439] for s- and p-polarized radiation at different incident angles. Absorption bands are assigned to fundamental symmetric and asymmetric CH stretching vibrations. The spectra show significant changes as a function of radiation incidence angle and direction of polarization. In the case of p-polarization the sign of RAIRS bands changes by crossing the Brewster angle at 73°. In addition the spectra provide information concerning the bonding between adsorbate and substrate. By comparison with calculations of relative intensities of absorption bands [439], an average tilt angle of the hydrocarbon chains of about 10° towards the surface normal was concluded.

4.3.1.2. Instrumentation and Applications

In the IR spectral range Fourier transform (FT) interferometers are generally utilized. In comparison to dispersive spectrometers, FTIR offers higher optical throughput and the multiplex advantage at equivalent high spectral resolutions. In the near-IR region, a tungsten-filament lamp, in the mid-IR region a silicon carbide rod (glowbar), and in the far-IR region a mercury-arc lamp are employed as radiation sources. A special radiation source is a synchrotron, which offers high brilliance over the whole IR region. For example, it was applied for in-situ measurements on the anodic oxidation of copper in aqueous solution [443]. Often RAIRS is used as a complementary method to HREELS (see Section 2.5.2.2), which is also a standard method for investigation of dipole-active vibrations, but has a low spectral resolution (ca. 20 cm^{-1} on semiconductor surfaces).

RAIRS is a nondestructive IR technique and also has especially versatile applications, because it does not require vacuum conditions, which are strictly necessary for electron-

spectroscopic methods. Therefore, RAIRS is in principle applicable to growth processes [444]. By applying a polarization – modulation technique, surfaces in can be investigated a gas phase. Higher surface sensitivity is achieved by modulation of the polarization between s and p. Furthermore, this method allows anisotropic near-surface absorption to be distinguished from the isotropic absorption in a gas phase [445].

RAIRS is routinely applied in the area of chemically modified surfaces. A variety of surface systems in electrochemistry [449], polymers research [438], [450], catalysis [437], [445], self-assembling monolayers [439], [440] and protein adsorption [440], [447] have been investigated.

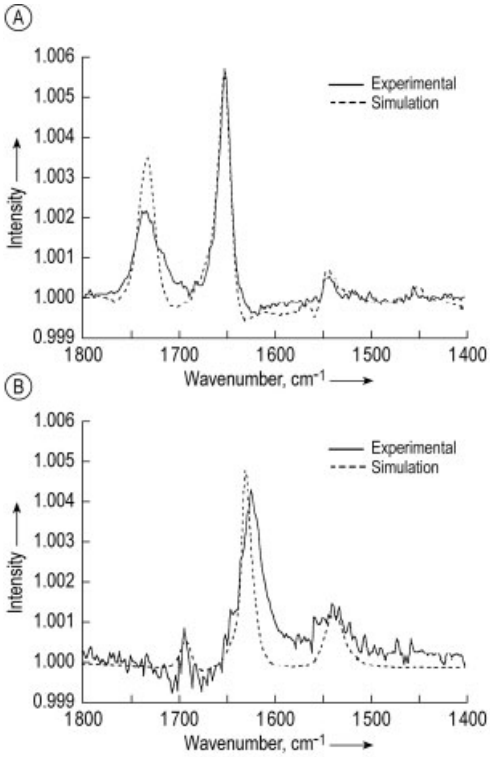

Figure 115. Experimental and simulated PM- (polarisation-modulated) RAIRS spectra of a single monolayer of PGB (A) and $K(LK)_7$ (B) at the air water interface. The surface tension was 20 mN/m [448].

Besides studies of adsorbates on metal and semiconductor surfaces, considerable interest has been aroused by the structural characterization of monolayers at air – water interfaces [441], [446]. Experiments at the gas – water interface have been carried out in biochemical research on the characterization of membrane proteins [447] and structures of biological molecules [448]. Figure 115 presents experimental and simulated RAIRS spectra of a single monolayer of PGB (poly-γ-benzyl-L-glutamate) and a synthetic peptide $K(LK)_7$ at the air – water interface [448]. In the simulations the different secondary structures of both molecules and anisotropic optic constants were taken into account. The PGB spectra show significant features resulting from amide I, amide II ,and ester bands in an α-helix structure. For $K(LK)_7$ the simulations reveal molecules lying in antiparallel β-sheets on the water surface. The wavenumber of the amide I mode ($1622\,cm^{-1}$) provides information on interactions between the antiparallel β-sheets and the substrate. The knowledge of such optical data is of great interest for determination of the main secondary structures (α-helix, β-sheets) found in proteins and polypeptides [448].

4.3.2. Surface Raman Spectroscopy

4.3.2.1. Principles

Raman scattering [451] is one of the phenomena that occur when electromagnetic radiation interacts with a molecule or a crystal. By the inelastic scattering process, photons lose energy for exciting molecular or lattice vibrations (Stokes scattering) or even gain energy (anti-Stokes scattering) when the transition starts from an excited vibrational state with a final state of lower energy (e.g., ground state). When a sample is irradiated by monochromatic light, the scattered light exhibits a spectrum of wavelengths in which the intensity peaks are shifted from the excitation wavelength by an amount of energy equivalent to that for the excitation of molecular vibration modes or crystal phonons. These vibrational spectra are highly specific for the material studied and can be used for an unambiguous identification of substances.

The Raman cross section of molecules is generally low. Therefore, intense light sources and low-noise detectors must be used, and the high sensitivities required for surface analyses

are difficult to achieve. Another issue in Raman spectroscopy is the superposition of the stronger fluorescence of the substances or their impurities on the weak Raman spectrum. This problem can be overcome by the use of longer excitation wavelengths, which cannot excite electronic transitions. This concept works well for numerous samples. However, near-infrared excitation wavelengths result in lower sensitivity in comparison with visible excitation because of the noise of detectors working in this range.

Raman spectroscopy does not need vacuum and is, in principle, applicable to all optically accessible samples. Even inclusions in glasses or minerals, substances in aqueous solutions, or substances packaged in transparent glass or plastics can be analyzed. Small particles down to 1 μm can also be characterized by proper focusing of the exciting laser beam.

4.3.2.2. Surface-Enhanced Raman Scattering (SERS)

The application of Raman spectroscopy in surface analysis is limited by the low scattering cross section of the molecules and the correspondingly low intensities of their Raman bands. This situation can be clearly improved by surface-enhanced Raman scattering (SERS) [452], [453]: The Raman scattering of molecules adsorbed on rough metal surfaces or small metal particles (Fig. 116) can be increased by six or even more orders of magnitude by electromagnetic and chemical effects. This large enhancement enables easy analysis of molecular monolayers on appropriate surfaces. Even 1 % of a monolayer has been observed, and the detection of single molecules on optimum metal particles has been reported [454].

The first SERS experiments were performed with electrochemically roughened electrodes and metal colloids. Meanwhile, numerous further types of suitable SERS substrates are known; for example, metal island films, metal films over nanoparticles or rough substrates, gratings, and sputter-deposited metal particles. Colloidal and sputtered particles can be deposited on many different materials, thus making SERS a widely applicable method.

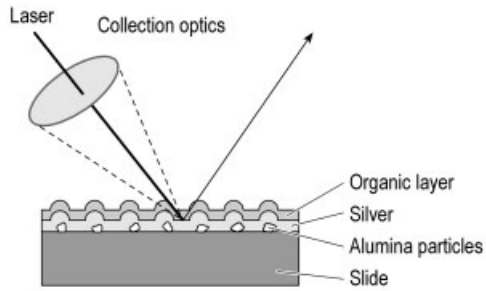

Figure 116. Scheme of an SERS-active substrate and the measuring arrangement. Alumina nanoparticles are deposited on a glass surface and produce the required roughness. A thin silver layer is evaporated onto the nanoparticles and serves for enhancement. Organic molecules adsorbed on the silver surface can be detected by irradiation with a laser and collecting the Raman-scattered light.

4.3.2.2.1. Instrumentation

The scattering effect was first observed by C. V. Raman in 1928 by using sunlight and complementary filters [451]. Later mercury vapor lamps were used for illumination of samples. Current Raman spectrometers use lasers as monochromatic light sources. Different types of lasers can be used, depending on the required wavelength, sensitivity, and spectral resolution. Gas lasers such as Ar^+ and HeNe lasers are commonly used because of their narrow, well-defined spectral lines, stability, and beam quality. High intensity Ar^+ lasers require high maintenance costs and consume a lot of electrical power and cooling water. Diode lasers are more convenient. Powerful diode lasers for Raman spectroscopy generally emit red wavelengths which are suitable for avoiding fluorescence and fit to the sensitivity range of silicon detectors. Diode lasers must be frequency-stabilized since their emission frequency is not defined by narrow transitions. Furthermore, spontaneous emission background and low-intensity sidebands must be filtered out when applying diode lasers for high-sensitivity Raman measurements. Solid-state lasers emitting in the near-infrared such as Nd:YAG lasers are used in combination with Fourier transform spectrometers.

Sample optics focus the laser onto the sample and collect the Raman-scattered light. In the widespread backscattering or 180° arrangement, a single objective serves to both focus the laser

and collect the scattered light. This makes measurements easy, since only the distance between sample and objective need be adjusted. Transparency of the whole sample is not needed with a backscattering arrangement. Therefore, Raman spectroscopy is well suited for in situ measurements at interfaces between gases, liquids, and solids with at least one transparent phase.

Elastically scattered laser light must be removed in front of the spectrograph because it is several orders of magnitude more intense than the Raman-scattered radiation and would otherwise produce stray radiation at the detector. Nowadays, holographic notch filters are mostly used for this purpose because of their high transmission of desired and strong suppression of undesired wavelengths. Premonochromators are preferable for measurements at low wavenumbers and for varying laser wavelengths.

The spectral analysis of the Raman-scattered light is performed by dispersive or Fourier transform (FT) spectrometers. Dispersive spectrometers decompose the radiation according to its wavelength by a grating monochromator. FT spectrometers calculate the spectrum from an interference pattern (interferogram) produced by an interferometer with a variable path length within one of the interferometer arms. Monochromators are preferable in the visible spectral range. Compact and robust monochromators without moving parts can be used in rough environments, whereas larger high-resolution devices are mainly used in research. In combination with array detectors, imaging monochromators can use the multichannel advantage and thus produce large signal-to-noise (S/N) ratios during short measuring times. FT spectrometers also produce considerable S/N ratios by virtue of their throughput advantage. However, the noise of their near-IR detectors is considerably larger than that of detectors operating in the visible region. The main advantages of FT spectrometers are strong suppression of fluorescence by using long wavelength excitation and variable spectral resolution.

Charge-coupled device (CCD) detectors are preferred in the visible spectral range because of their low dark current, high quantum efficiency, and the multichannel capability. They must be cooled thermoelectrically or by liquid nitrogen for high-sensitivity applications such as surface spectroscopy. The sensitivity of CCD detectors decreases considerably for wavelengths larger than 1 μm due to the band gap of the silicon material. Therefore, other semiconductor detectors with larger dark currents, such as germanium or InGaAs, must be used for near-IR Raman spectroscopy.

Different accessories are available for Raman spectroscopy. Confocal microscopes provide a lateral resolution of about 1 μm and a depth resolution down to about 2 μm. The depth resolution is especially important for surface and interface spectroscopy since it helps to eliminate Raman signals from the bulk phases. Fiber optics can be used for guiding the exciting laser light to the sample and the Raman-scattered light to the spectrometer, thus making sample changing easier and avoiding hazards due to freely propagating laser light. Special sample cells were constructed for measurements at high or low temperatures, high pressures, and for laser multipass through gaseous samples.

4.3.2.2.2. Spectral Information

At least for large molecules, Raman spectra show numerous bands which cannot always completely be assigned to certain vibrational modes. However, the many bands, when measured with appropriate spectral resolution, can serve for an unambiguous identification of substances by comparing the spectral pattern (fingerprint) to those of references if suitable reference spectra are available.

Several bands in characteristic spectral regions can be definitely assigned to molecular or crystal vibrations, especially if some information on the chemical composition of the sample is available. For instance, these bands allow the identification of aromatic or aliphatic hydrogen atoms, carbonyls, thiols, double or triple bonds, heterocycles, and other groups or skeletal components. Vibrations of weakly polar and even symmetrically bonded atoms generally result in intense Raman bands because Raman activity is connected with changes in the molecular polarizability during the vibration. This is a major difference to IR spectroscopy, where bands from vibrations causing changes in the dipole moment dominate the spectrum. Raman bands of carbon – carbon bonds are readily observable.

Strong vibrational coupling along carbon chains and symmetry selection rules make these bands sensitive markers for substitution at aromatic rings and the conformation of the carbon backbone. Structural isomers usually possess very different spectra and can therefore be easily distinguished.

Polarization effects are an additional feature of Raman spectroscopy that improves the assignment of bands and allows the determination of molecular orientations. For isotropic phases, the analysis of polarized and nonpolarized bands allows the determination of the symmetry of the respective vibrations. For aligned molecules in crystals or at surfaces, up to six independent Raman spectra can be measured in dependence on polarization and propagation direction of incident and scattered light with respect to the molecular or crystal axes.

4.3.2.2.3. Applications

Developments in Raman equipment have led to a considerable increase in sensitivity. This allowed reactions of organic monolayers at carbon surfaces to be monitored [455] and the structure of Langmuir – Blodgett monolayers to be analyzed without any enhancement effects. This unenhanced surface Raman spectroscopy is expected to be applied to various technically or scientifically important surfaces and interfaces. Nevertheless, it requires a careful optimization of the apparatus, data treatment, and sample preparation. The surfaces of porous materials such as catalysts, molecular sieves, and adsorbents are much more readily accessible to Raman spectroscopy, since larger amounts of adsorbed substance can be placed within the laser focus, thus contributing to the scattering process.

SERS exhibits excellent sensitivity but is restricted to specially prepared "active" surfaces. Silver, gold, and copper electrodes are easily activated by roughening with reduction – oxidation cycles. Therefore, SERS was broadly applied in electrochemistry for monitoring adsorption, orientation, and reactions of molecules on these electrodes in situ. Further electrode and catalyst materials are accessible to SERS by deposition of such thin layers on SERS-active silver supports.

Coating of metal SERS substrates with organic receptor layers makes SERS promising for highly sensitive chemo-optical sensors for organic trace analysis in water and air [456]. The combination of selective adsorption with the measurement of the molecule-specific SERS spectra can reduce cross-sensitivities considerably. SERS substrates with bare metal surfaces irreversibly adsorb thioorganics (Fig. 117) and other compounds and they thus can serve for the detection and identification of very low gas or solution concentrations of these substances [457].

SERS was also applied as sensitive, molecule-specific detection method in chromatography, e.g. thin layer, liquid and gas chromatography.

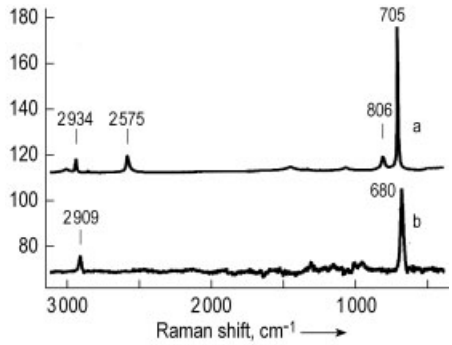

Figure 117. Raman spectrum of methyl mercaptan (a) and SERS spectrum of methyl mercaptide (b) formed by adsorption of the mercaptan on a silver surface. The surface reaction is proven by the disappearance of the S–H stretching and bending bands at 2575 and 806 cm^{-1}, respectively. The Raman shift of the C–S stretching band around 700 cm^{-1} is decreased during adsorption by withdrawal of electron density from the C–S bond due to bonding to the silver. The symmetric methyl stretching band appears above 2900 cm^{-1} [457].

4.3.2.3. Nonlinear Optical Spectroscopy

Unlike linear optical effects such as absorption, reflection, and scattering, second-order nonlinear optical effects are inherently specific for surfaces and interfaces. These effects, namely, second harmonic generation (SHG) and sum frequency generation (SFG), are dipole forbidden in the bulk of centrosymmetric media. Especially in the investigation of isotropic phases such as liquids, gases, and amorphous solids,

signals arise exclusively from the surface or interface region, where broken symmetry is found. Nonlinear optics are applicable in situ without the need for vacuum, and in addition they show a fast time response.

Surface SHG [458] produces frequency-doubled radiation from a single pulsed laser beam. Intensity, polarization dependence, and rotational anisotropy of the SHG provide information on the surface concentration and orientation of adsorbed molecules, as well as on the symmetry of surface structures. SHG was successfully applied for the analysis of adsorption kinetics and ordering effects on surfaces and interfaces, reconstruction of solid surfaces and other surface phase transitions, as well as potential-induced phenomena on electrode surfaces.

SFG [459], [460] uses two different laser frequencies for the generation of the sum frequency. Tuning one laser in the infrared spectral range allows molecular vibrations of adsorbed molecules to be monitored. SFG includes the capabilities of SHG and can, in addition, identify molecules and their structure on the surface by analyzing the vibration modes. It has been applied to the observation of surfactants at liquid surfaces and interfaces, as well as the ordering of interfacial water molecules.

4.3.3. UV – VIS – IR Ellipsometry (ELL)

4.3.3.1. Principles

Ellipsometry is a method of measuring film thickness, refractive index, and extinction coefficient of single films, layer stacks, and substrate materials to very high sensitivity. Rough surfaces, interfaces, material gradients, as well as mixtures of different materials can be analyzed.

Film thicknesses between 0.1 nm and 100 µm can be measured, depending on the spectral range used for the analysis and the homogeneity of the thicker films. Thicknesses smaller than 1 µm can be determined with a sensitivity better than 0.01 nm. Thicknesses in the micrometer range can be analyzed with a sensitivity typically better than 1 nm.

The refractive index of a film or a substrate material can be measured with a sensitivity better than 5×10^{-4}, which is the highest available value for noninvasive optical measuring methods, especially for thin films. The extinction coefficient can be measured with nearly the same sensitivity, which corresponds to a lower limit of $10 - 100\,\mathrm{cm}^{-1}$ for the absorption coefficient of the material.

The number of measurable layers of a stack is only limited by the optical contrast between the different layers. In practice, stacks of ten layers and more can be analyzed by ellipsometry. Further advantages of ellipsometry compared to other metrological methods are the noninvasive and nondestructive character of the optical method, the low energy input into the sample, the direct measurement of the dielectric function of materials, and the ability to take measurements in any kind of optically transparent environment.

The principal of measurement is shown schematically in Figure 118. Linear polarized light is reflected from a sample surface, which must be flat and sufficiently reflecting. The polarization state of the incident light is changed by the reflection into elliptically polarized light. The parameters measured by ellipsometry are $\tan\Psi$, the amplitude ratio of the resolved components of the electric field vector of the reflected light parallel to and perpendicular to the plane of incidence, and $\cos\Delta$, the phase difference of the two components. An unambiguous determination of the phase angle Δ in the whole data range between 0 and 360° is possible by using a retarder to give circularly polarized light as incident light in a second measurement in which the parameters $\tan\Psi$ and $\sin\Delta$ are determined.

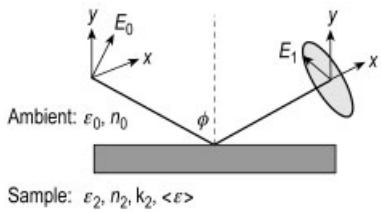

Figure 118. Principle of ellipsometry measurements

The information about the properties of the sample are contained in the complex ratio ρ of the Fresnel reflection coefficients parallel to (r_p) and perpendicular to (r_s) the incident plane-polarized electric field vector:

$$r = \frac{r_p}{r_s} \tag{4.8}$$

The fundamental equation of ellipsometry [461]

$$r = \tan\Psi \cdot e^{i\Delta} \tag{4.9}$$

describes the connection between the measured quantities Ψ, Δ and the sample properties contained in the reflection coefficients, and hence in ρ. In principal this complex equation can be solved analytically only for pure substrates. Each more complex structure requires an optical modeling of the sample and a fit of the calculated ratio ρ to the measured quantities Ψ and Δ. Ψ and Δ are always correct quantities provided the ellipsometer is working correctly. Film thickness, refractive index, extinction coefficient, and other parameters are calculated on the basis of a model. The choice of a correct optical model of the sample is a fundamental assumption for correct values of all calculated parameters.

The thickness of a film influences the interference of light waves reflected from the front and back side of the film, and hence the reflectance. Therefore, the thickness of an absorbing film can only be measured as long as there is still a contribution of back-side reflection to the reflectance of the sample. Typical measurable thicknesses of metallic layers are less than 50 nm.

Ellipsometric measurements depend on the incident angle ϕ due to the angle dependence of the reflection coefficients r_s and r_p. Large differences between both quantities are found for angles above 50°, whereby r_p has a local minimum at the so-called Brewster angle. This angle is between 50 and 60° for transparent materials and greater than 70° for most absorbing materials. Ellipsometric measurements close to the Brewster angle of the substrate are especially sensitive for very thin layers on top of the substrate. Multiple-angle measurements are suitable for confirming optical models of a sample structure.

Two measured ellipsometric angles Ψ and Δ at a fixed wavelength and a fixed incident angle allow the calculation of at most two other parameters, for example, film thickness and refractive index of a transparent layer. Multiple-angle measurements increase the number of measured quantities and hence the number of parameters which can be determined for a specific sample. But in this case, too, the number of measurable parameters will typically not larger than four or five. Wavelength-dependent ellipsometric measurements (spectroscopic ellipsometry) further increase the number of measurable parameters, and hence the complexity of the analyzable sample structure. Spectroscopic ellipsometry allows the calculation of the dispersion of the refractive index and the extinction coefficient of materials, as well as the analysis more complex sample structures (e.g. multilayer stacks, interfaces, material gradients, material compositions) on the basis of parameterized dielectric functions which drastically reduce the number of unknown parameters compared to the number of measured ellipsometric angles.

Spectroscopic ellipsometry is sensitive to the dielectric functions of the different materials used in a layer stack. But it is not a compositional analytical technique. The combination with one of the compositional techniques such as AES or XPS and with XTEM for information about the vertical structure can provide valuable additional information for creating a suitable optical model of an unknown complex sample structure.

4.3.3.2. Instrumentation

The basic configuration of an ellipsometer is very simple: light source – polarizer – sample – polarizer – detector. Advanced configurations use a phase shift device (retarder) before or after the sample.

There are manys in which a specific ellipsometer can be realized. The main differences consist in the way in which the light is polarized and the detection of the polarization state of the reflected light. The light source can be monochromatic (lasers), white light (Xe arc lamp, deuterium lamp, halogen lamp) with a monochromator, or silicon carbide rods (glowbars; mainly for IR ellipsometry).

Photomultiplier tubes (vacuum UV), Si photodiodes (UV – VIS), Ge or InGaAs photodiodes (NIR), and MCT (HgCdTe) or DTGS detectors (MIR) are used as broadband detectors. They must be operated with a monochromator if a white-light source is used. Fast spectroscopic ellipsometers use photometers and diode arrays as detectors. This allows simultaneous detection of

multiple wavelengths rather than sequential detection. State-of-the-art ellipsometry in the IR spectral region (ca. 0.8 – 100 µm wavelength) uses modulated white light from an FT-IR instrument and broadband detectors to measure the reflected light. Intensity spectra are generated by Fourier transformation of the detected signal. Ellipsometric measurements are performed at a fixed angle of incidence. Each ellipsometer configuration can be operated in a discrete-wavelength or spectroscopic, that is, variable angle of incidence mode.

Null ellipsometers are one of the oldest configurations, in which a revolving polarizer and a revolving compensator are used to change the polarization state of the incident light, and a revolving polarizer (analyzer) is used to analyze the reflected light. The angular position of the optical elements is changed until the light intensity on the detector is zero. This configuration is one of the most accurate but also the slowest, even if automated.

Rotating-element ellipsometers have a continuously rotating element (10 – 40 Hz frequency of revolution), which allows automated acquisition and faster measurement. Most of the instruments used nowadays have this configuration. Polarizer (rotating polarizer ellipsometer, RPE) or analyzer (rotating analyzer ellipsometer, RAE) can be used as rotating element. The detector signal is measured as function of time and is then Fourier analyzed to obtain the ellipsometric angles Ψ and Δ. The RPE configuration has the advantage that any polarization sensitivity of the detector (photomultiplier tube, monochromator) does not influence the measurement. The polarization state of the detected light is fixed by the fixed analyzer position. The disadvantage is the sensitivity to residual polarized light of the light beam incident on the rotating polarizer. The RAE configuration has the advantage of being insensitive to any residual polarization of the light source (laser, white-light source with monochromator), but errors can occur due to polarization sensitivity of the detector (photomultiplier tube). With polarization-insensitive Si photodiodes as detectors, this configuration is most often used for discrete wavelength ellipsometers. Typically, single-wavelength ellipsometers are equipped with a quarter-wave plate (compensator) which can be moved into the light beam. The linear polarized light is changed into circularly polarized (in general, elliptically polarized) light after passing through the quarter-wave plate. Thereby, the measured ellipsometric angle Δ is shifted by 90°, and it can be determined over the whole data range of 0 to 360°, and the large uncertainty in Δ when its value is close to 0 or 180° is removed.

Spectroscopic ellipsometers that measure Δ over the whole data range require an achromatic compensator with a phase shift close to 90° over a large spectral range. The compensator can be located between polarizer and sample, or between sample and analyzer.

Step-scan polarizer/analyzer (SSP/SSA) and continuously rotating element configurations are used for diode-array or FT-IR-based ellipsometers. In the SSP or SSA modus, complete intensity spectra are measured at a number of angle positions of the revolving polarizer/analyzer. The acquisition time varies between some seconds (UV – VIS) and several minutes (IR), depending on the intensity of the incident beam and the reflectance of the sample. The measurements are fast compared to the monochromator-based configurations and very accurate regarding the exact positioning of the revolving element. The continuously rotating element configuration is used for very fast spectroscopic ellipsometers measuring complete ellipsometric spectra in less than 100 ms. For such instruments, the intensity at each wavelength (pixel of the diode array) is accumulated over a 45° revolution angle of the rotating element.

A few instruments have a continuously rotating compensator and fixed polarizer and analyzer. The advantage of this configuration is the insensitivity to residual polarization of the light source or a polarization sensitivity of the detector. The disadvantage is the large influence of a small misalignment of the rotating achromatic compensator on the performance of the instrument.

Polarization modulation ellipsometers use a photo-elastic modulator to modulate the polarization state of the incident beam. Polarizer and analyzer are fixed during the measurement. A Fourier analysis of the time-dependent signal gives the ellipsometric parameters Ψ and Δ. The advantages of this configuration is the very high measuring speed (typical modulation frequency of 50 kHz) for discrete wavelengths and the lack of moving parts. The ellipsometric angle Δ can

be measured over the whole range of 0° to 360°. Disadvantages are the frequent calibration of the instrument owing to the high temperature dependence of the polarization modulator and the limited measurement speed for complete spectra.

The infrared ellipsometer consists of a combination of a Fourier transform spectrometer (FTS) and a photometric ellipsometer. One of the two polarizers (the analyzer) is moved step by step in four or more azimuths, because the spectrum must be constant during the scan of the FTS. From these spectra, the $\tan\Psi$ and $\cos\Delta$ spectra are calculated. In this case, only Δ is determined in the range of 0 – 180° with a strongly reduced accuracy in the neighborhood of 0 and 180°. This problem can be overcome by using a retarder (compensator) with a phase shift of approximately 90° in a second measurement. Thereby, $\cos\Delta$ and $\sin\Delta$ are measured independently with full Δ information [465].

In a second kind of IR ellipsometer, a photoelastic modulator (PEM) as a dynamic retarder replaces the static one. The PEM produces a sinusoidal phase shift of approximately 40 kHz and supplies the detector exit with signals of the ground frequency and the second harmonic. From these two frequencies and two settings of the polarizer and PEM the ellipsometric spectra are determined [466]. This ellipsometer system is mainly used for fast and relative measurements.

4.3.3.3. Applications

UV – VIS – NIR Spectral Region. Spectroscopic ellipsometry in the UV – VIS – NIR region is mainly used for the characterization of new materials (e.g., photoresists, polymers, low-*k* dielectrics, semiconductors, composite materials), the analysis of multilayer stacks (e.g., low-*k* and antireflection (AR) layer stacks on architectural glass, flat-panel applications, silicon IC technology, optoelectronic devices, photonic devices), the qualification of a large number of deposition processes for production and R&D applications, as well as in-situ monitoring of growth and deposition processes. Spectral range and speed of the measurement are key features which recommend the use of spectroscopic ellipsometers for R&D and production surveillance. The existence of spectroscopic ellipsometers in the spectral range of 140 nm to 2.3 µm allows for many applications in metrology.

Ellipsometry in the vacuum-UV (< 190 nm) allows the analysis of materials for next-generation lithography (photoresists, AR coatings) at the latest exposure wavelengths (157 nm and 193 nm). The short wavelengths increase the sensitivity of ellipsometric measurements of ultrathin films (< 10 nm). New prospects are expected for the analysis of thin metallic and dielectric layers.

In the UV most of the interesting materials like Si, polysilicon, SiGe, GaAs, and other semiconductor materials are strongly absorbent; this permits surface-sensitive measurements. Surface roughness, native oxide covering, material composition, and structural properties can be analyzed.

The VIS – NIR spectral range is mainly used to measure the thickness of single films and layer stacks. Many materials (dielectrics, semiconductors, polymers) are transparent in this spectral range. The analysis of complex layer stacks often requires the knowledge of the dielectric function of each material in the stack or parametrized dispersion models to reduce the number of unknown parameters. The Cauchy formula is often used for transparent layers, the Drude – Lorentz formula works well for metals and IR-active materials, the LENG formula [462] can be used for semiconductor materials, and the Tauc – Lorentz formula [463] works best for amorphous materials. The general Gaussian-broadened polynomial superposition (GBPS) parametric dispersion model [464] can be broadly applied to most materials, including crystalline and amorphous semiconductors, metals, and organics. Figure 119 shows measured ellipsometric spectra of an a-Si/Si_xN_y multilayer deposited on GaAs. Four layers of Si_xN_y and three layers of a-Si were alternately deposited. The dispersion of each layer material was measured on single films. Film thicknesses of 78 nm for a-Si and 145 nm for Si_xN_y were determined for the different layers of the stack by using appropriate parametrized dispersion models for each material. Other typical multilayer applications are dielectric stacks on silicon, spectral selective mirror stacks, and Bragg reflectors for laser devices.

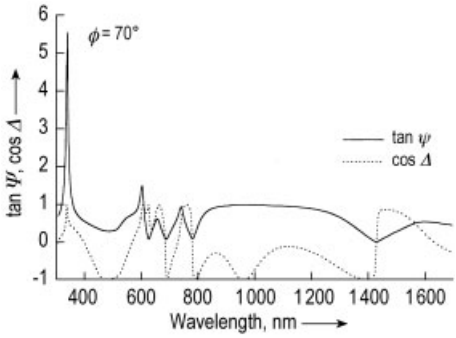

Figure 119. Measured ellipsometric spectra of an a-Si/SiN multilayer stack (7 layers) on GaAs.

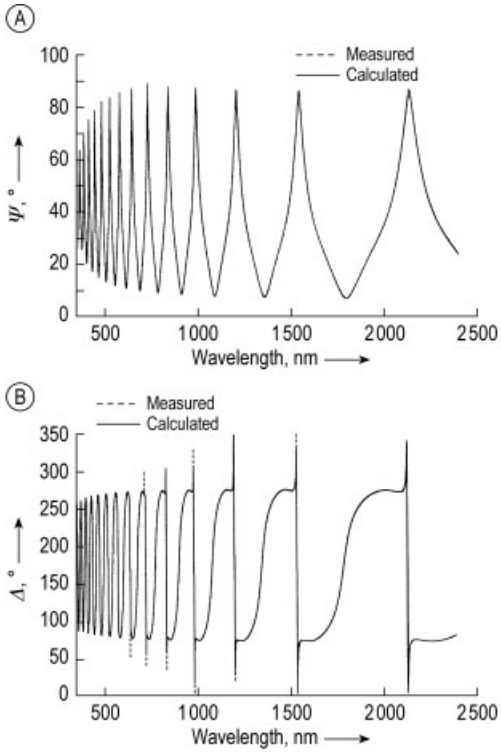

Figure 120. Measured and calculated ellipsometric spectra of a 2.5 µm thick SiO_2 layer on Si

Ellipsometric measurements in the NIR are often used for the determination of thickness and refractive index of thicker dielectric films and semiconductor stacks based on InP and GaAs, which are transparent in this spectral region. The wavelength and film thickness should not differ much, so that a highly sensitive ellipsometric measurement of film thickness and refractive index is obtained. Figure 120 shows ellipsometric spectra of an 2.5 µm thick SiO_2 layer on silicon measured between 350 nm and 2.3 µm wavelength. The higher sensitivity in the NIR spectral range is clear. The growing number of applications of 1.55-µm semiconductor lasers forces the application of NIR ellipsometers to control and to monitor AR coatings on laser facets.

Laterally inhomogeneous films and patterned structures of microelectronic and optoelectronic applications require small measuring spots. Measurements in 50×50 µm areas are standard for microspot spectroscopic ellipsometers used in fabrication lines. Areas more than ten times smaller can be analyzed by using discrete-wavelength ellipsometers equipped with laser light sources.

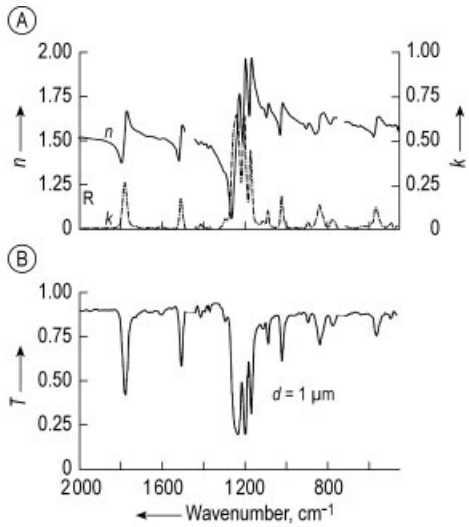

Figure 121. Bulk n, k spectrum of polycarbonate (A) and transmission T spectrum (B) calculated from n, k for a thickness of 1 µm

Infrared Ellipsometry. Infrared ellipsometry typically is carried out in the MIR range of 400 to 5000 cm^{-1}, but also in the NIR and FIR. The resonances of the molecule vibrations or phonons in the solid states provide typical features in the tan Ψ and Δ spectra in the form of relative minima or maxima and dispersionlike structures. For an isotropic bulk, the calculation

of the optical constants – refractive index n and extinction index k is straightforward. In all other cases (thin films and anisotropic materials) iteration procedures are applied. In ellipsometry only angles are measured, and therefore the results are absolute values without using any standard.

Two types of resonances have to be distinguished in the infrared: the strong and the weak oscillator. The strong oscillator is characterized by a region in which the refractive index is lower than unity for wavenumbers higher than the oscillator frequency and is mostly found in inorganic compounds with reststrahlen bands (n, k spectrum of quartz glass, Fig. 124). The weak oscillators always have a refractive index greater than one and are typical of organic compounds (Fig. 121, top). The optical constants of these spectra are calculated by using relations of the metal optics [467]. From the determined ellipsometric optical constants, other experimental configurations can be calculated. In Figure 121 (bottom) the calculated transmission spectrum of the spectra in Figure 121 (top) is shown, which now is suitable for the direct comparison with library spectra.

The spectra of a strong oscillator have some special features which differ from those of a metallic or a dielectric substrate. Different structures in tan Ψ and Δ are observed on a metallic substrate, dependent on the thickness of the film (Fig. 122). For a very thin film up to approximately 100 nm, the Berreman effect is found near the position of $n = k$ and $n < 1$ with a shift to higher wavenumbers relative to the oscillator frequency. This effect decreases with increasing thickness ($d >$ ca. 100 nm) and is replaced by the excitation of a surface wave at the boundary of the dielectric film and metal. Now the oscillator frequency (TO mode) can also be observed. On metallic substrates for thin films ($d <$ ca. 2 µm) only the z component of the electric field is relevant. Thin films on a dielectric substrate always show the oscillator frequency and the Berreman effect simultaneously, because in this case all three components of the electric field are possible (Fig. 123). The relative slope of tan Ψ (max. or min.), Δ, and reflectivity for the parallel component depends on the relative position with respect to the Brewster angle (greater or lower). Greater thickness of the films on both types of substrates leads to dominant interferences.

Double and buried layers can be detected well by infrared ellipsometry if the film material is characterized by reststrahlen bands. The response is observed at the position of the Berreman effect (Fig. 124 with SiO_2, Si_3N_4 layer, and the SiO_2 spectrum for comparison) and also by the TO mode for greater thickness. The thinnest film was observed for the H-terminated silicon surface with a thickness of 0.2 nm [468].

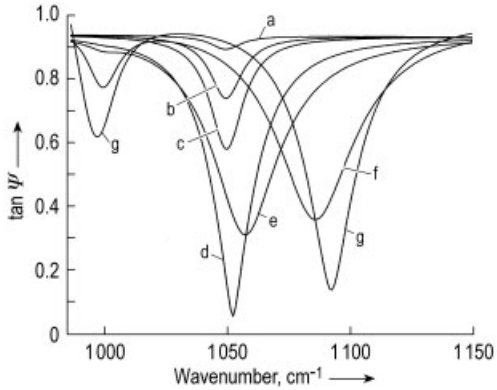

Figure 122. Different spectral features of tan Ψ for a strong model oscillator at 1000 cm^{-1} on a metal substrate. The TO mode (1000 cm^{-1}), Berreman effect (1050 cm^{-1}), and excitation of a surface wave (1090 cm^{-1}) are seen for different thickness: (a) 1, (b) 5, (c) 10, (d) 50, (e) 100, (f) 500, and (g) 1000 nm

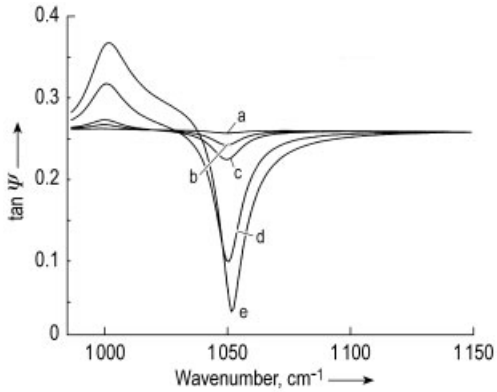

Figure 123. Tan Ψ spectra for the same oscillator as in Fig. 122 for a silicon substrate, thicknesses: (a) 1, (b) 5, (c) 10, (d) 50, and (e) 100 nm.

The determination of the optical constants and the thickness is complicated by the problem of calculating three results from two ellipsometric values. This problem can be solved by the oscillator fit in a suitable wavenumber range or

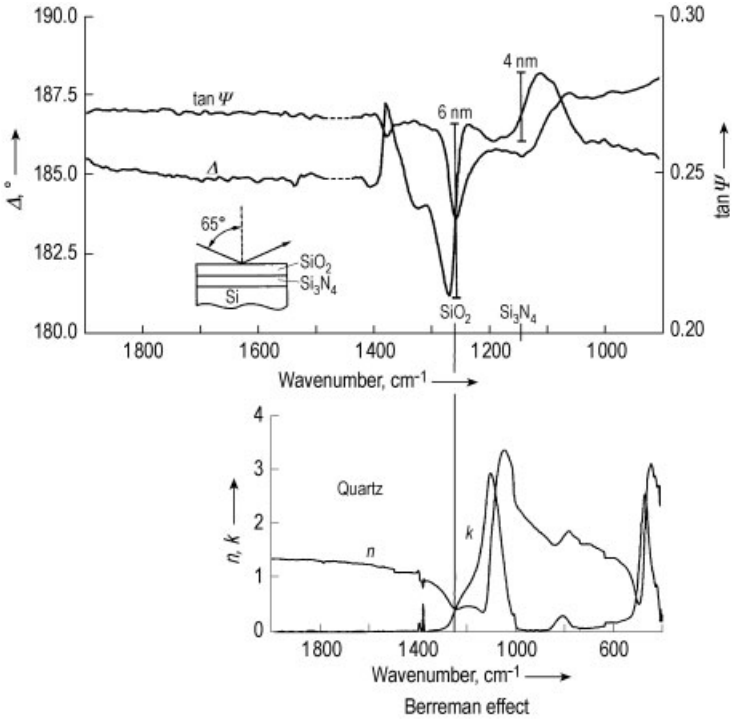

Figure 124. Tan Ψ and Δ spectrum of a double layer of SiO_2 (6 nm) and Si_3N_4 (4 nm) on a silicon substrate (top) and SiO_2 spectrum (bottom) for comparison.

by exploiting the fact that in the infrared ranges without any absorption always exist. In this case the thickness and the refractive index outside of the resonances can be determined for instance by the algorithm of Reinberg [467].

In general it is possible to investigate very thin films up to the subnanometer region by infrared wavelengths which are much greater than the thickness of the film with a ratio of 1:10 000, because of the interference optics of the strong oscillator (Berreman effect).

4.4. Other Photon-Detecting Techniques

4.4.1. Appearance-Potential Methods

Appearance-potential methods all depend on detecting the threshold for ionization of a shallow core level and the fine structure near the threshold, differing only in the way in which detection is performed. In all of them the primary electron energy is ramped upward from near zero to whatever is appropriate to the sample material, while the primary current to the sample is kept constant. As the incident energy is increased, it passes through successive thresholds for the ionization of core levels of atoms in the surface. An ionized core level, as discussed earlier, can recombine by the emission either of a characteristic X-ray photon or of an Auger electron.

4.4.1.1. Soft X-Ray Appearance-Potential Spectroscopy (SXAPS)

In SXAPS the X-ray photons emitted by the sample are detected, normally by letting them strike a photosensitive surface from which photoelectrons are collected, but also – with the advent of X-ray detectors of increased sensitivity – by direct detection. Above the X-ray emission threshold from a particular core level the excitation probability is a function of the densities of unoccupied electronic states. Since two electrons

are involved – the incident and the excited – the shape of the spectral structure is proportional to the self-convolution of the unoccupied state densities.

4.4.1.2. Disappearance-Potential Spectroscopy (DAPS)

Crossing an ionization threshold means that electrons are lost from the primary beam as a result of ionization of a core hole. Thus, if the reflected current of electrons at the primary energy, more usually termed the elastically reflected current, is monitored as a function of energy, it should show a sharp decrease as a threshold is crossed. This is the principle of operation of DAPS. It is in a sense the inverse of AEAPS, and indeed if spectra from the two techniques from the same surface are compared, they can be seen to be mirror images of each other. Background problems occur also in DAPS. The principal advantages of AEAPS and DAPS over SXAPS is that they can be operated at much lower primary electron currents, thus causing less disturbance to any adsorbed species.

4.4.2. Inverse Photoemission Spectroscopy (IPES) and Bremsstrahlung Isochromat Spectroscopy (BIS)

Irradiation of a surface with electrons leads to the emission not only of X-ray photons of energies characteristic of the material, but also of a continuous background of photons called *bremsstrahlung radiation.* If a detector is set to detect only background photons of a particular energy, and the primary electron energy is ramped upward from zero, then the variations in photon flux should mirror the variations in the densities of unoccupied electron states. The process is called inverse photoemission because it is clearly the opposite of ordinary photoemission as in XPS and UPS. Although the name *inverse photoemission spectroscopy* (IPES) should apply to all forms of the inverse photoemission technique, in fact it is confined by usage to the form in which the energy at which the photons are detected is in the UV region, typically with $h\nu = 9.7$ eV. If, on the other hand, a crystal monochromator normally used to provide monochromatized X-rays for XPS is used in reverse, with a detector placed where the electron source is in Figure 3, then detection is at the X-ray energy $h\nu = 1486.6$ eV, and the technique is called *bremsstrahlung isochromat spectroscopy* (BIS).

Since IPES maps the densities of unoccupied states, it is related to other techniques that do the same (i.e., STS and SXAPS). When used in conjunction with a technique that maps the densities of occupied surface states, such as UPS or ELS, a continuous spectrum of state density from occupied to unoccupied can be obtained. Just as in UPS, in which angular resolution allows elucidation of the three-dimensional occupied band structure, so in IPES angular resolution allows mapping of the three-dimensional unoccupied band structure. This version is called *KRIPES* (i.e., K-resolved IPES).

4.4.3. Ion-Beam Spectrochemical Analysis (IBSCA)

Ion irradiation of a surface produces visible-light photons as well as many other types of particle. The light emission arises from the deexcitation of ions and excited neutrals, and is characterized by a spectrum consisting of a series of sharp lines at wavelengths associated with the type of ionized or excited atom. Spectral analysis thus provides elemental identification by reference to standard emission wavelength tables. The technique is called *ion-beam spectrochemical analysis* (IBSCA) but has several other names as well. Typical operating conditions include bombardment by Ar^+ ions at energies in the range 2 – 20 keV, with current densities of 10 – 100 $\mu A/cm^2$; the light emitted just above the target surface is then focused onto the entrance slit of a scanning optical monochromator, which analyzes the light spectrum. Spatial resolution depends on the type of ion source being used, and can be on the order of 2 – 5 μm with a suitable type such as a duoplasmatron. Because emission takes place from-particles ejected from the surface layer, the surface specificity is similar to that of SSIMS (i.e., one atom layer). Quantification can in principle be carried out theoretically but is normally performed by relative intensity measurements, that is, by measuring the intensity of emission from a pure elemental standard under

the same instrumental conditions of analysis as used for an unknown.

5. Scanning Probe Microscopy

The invention of the scanning tunneling microscope (STM) by Binnig and Rohrer in 1981 [470], [471], has triggered a tremendous development of scanning probe microscopy (SPM) techniques [472]. The basic principle all scanning probe microscopes have in common is shown in Figure 125. A sharp probe (e.g., tip, optical fiber, pipette) is rastered across a sample surface by means of piezoelectric translators, while a certain signal is recorded by the probe for each single image point. Because the size of the probe's apex and the distance between the probe and sample surface can be smaller than the wavelength of the analytical signal (e.g., electrons, electromagnetic radiation) information can be obtained in the so-called near-field and thus the resolution is no longer diffraction-limited as given by the wavelength of the signal. Since the measured signal constitutes local information on the surface below the probe, the term local-probe technique is frequently used. An important feature of this concept is to use the local signal for monitoring the distance between probe and surface. Thus, topographical information, in principle down to atomic resolution, can be obtained in real space. Moreover, the possibility of positioning sharp probes above a surface with angstrom and sub-angstrom precision opens up the exciting possibility of accessing local spectroscopic information, even confined to spots as small as a single atom.

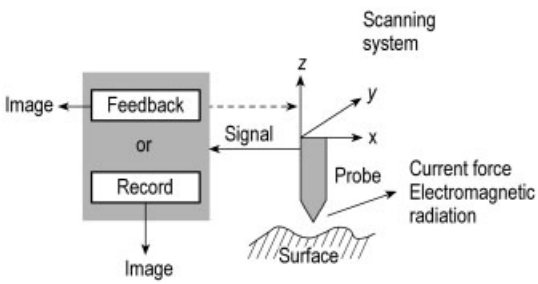

Figure 125. General scheme of scanning probe microscopy

Although a large number of scanning probe microscopy techniques utilizing various signal generation mechanisms has emerged, atomic force microscopy and scanning tunneling microscopy are the most important techniques with a wide range of applications in science and technology.

5.1. Atomic Force Microscopy (AFM)

5.1.1. Principles

Atomic force microscopy (AFM), invented in 1985 by Binnig, Quate, and Gerber [473], [474], belongs to the group of scanning force microscopy (SFM) techniques, which are based on the measurement of different forces (e.g. attractive, repulsive, magnetic, electrostatic, van der Waals) between a sharp tip and the sample surface. Imaging is accomplished by measuring the interaction force by means of the deflection of a soft cantilever while raster-scanning the tip across the surface (Fig. 126). Although various types of forces are encountered when a tip approaches the sample surface, signal generation in AFM is essentially based on interatomic repulsive forces, which are of extreme short range nature (Fig. 127). In the ideal case the tip can be considered to be terminated by a single atom, which means that direct contact of the tip with the sample surface is confined to an extremely small area. As a consequence there is always an interatomic repulsive force at this small contact area due to interpenetration of the electronic shells of the tip and substrate atoms. Since the interatomic repulsive force is influenced by the total electron density around an atom, and not just by particular states of energy as in scanning tunneling microscopy, this force can be used to map the topography of the surface down to atomic dimensions. In addition to this short-range force, long-range forces (e.g., Coulomb forces between charges, dipole – dipole interactions, polarization forces, van der Waals dispersion forces, capillary forces due to adsorbate films between tip and substrate), which can be attractive or repulsive, are also encountered (Fig. 127). Although both types of forces contribute to the total force acting on the probing tip, only the varying repulsive interatomic force

allows high-resolution imaging of surfaces, because of its extremely short range nature. Since long-range attractive forces pull the tip towards the sample surface, they give rise to an increase in the local repulsive force, which can deteriorate the experimental conditions. This can lead, for instance, to artefacts due to sample deformation or, in the worst case, the sample can be destroyed by the tip's scratching across the surface. Hence, it is important to minimize these long range forces (e.g., by imaging under liquids) to achieve very low repulsive forces (nanonewton and less) in the contact area between tip and sample. This is especially important when looking at soft materials, which can be deformed or destroyed easily by the load of the tip.

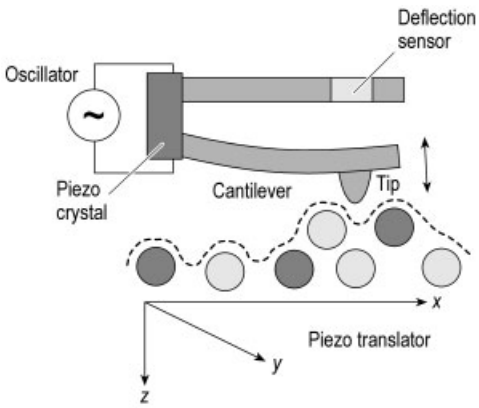

Figure 126. Principle of AFM. The sample, symbolized by the circles, is scanned by means of a piezoelectric translator. The piezo crystal and the oscillator are only needed for tapping-mode operation.

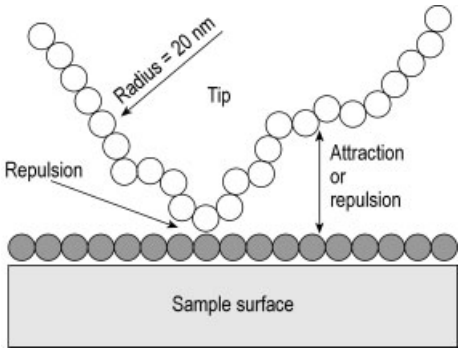

Figure 127. Schematic view of forces encountered when the tip touches the sample surface. Light circles symbolize tip atoms, and dark circles sample atoms.

The interaction between tip and sample can be described by force – distance curves. The force – distance curve in Figure 128 shows how the force changes when the sample surface approaches the tip. At large separations there is no interaction, and the observed force is zero (straight line between 1 and 2, if we assume that there are no electrostatic charging forces). At position 2 the tip jumps into contact due to attractive van der Waals interaction. As the sample is further moved towards the tip, the total force acting on the cantilever becomes repulsive. When the sample is retracted again, the force is reduced along the line from position 3 to 4. Below the zero-force line in the diagram the net force acting on the cantilever becomes attractive, because the tip is held at the surface due to adhesion. At 4 the adhesion force and the cantilever load are just balanced, and the tip flips off the surface when the sample is further retracted. For AFM measurements, the force can be set along the curve between positions 3 and 4, preferably close to 4 to minimize the contact force. The value of the pull-off force can be reduced significantly by imaging under liquids due to elimination of capillary condensation forces, which pull the tip towards the sample.

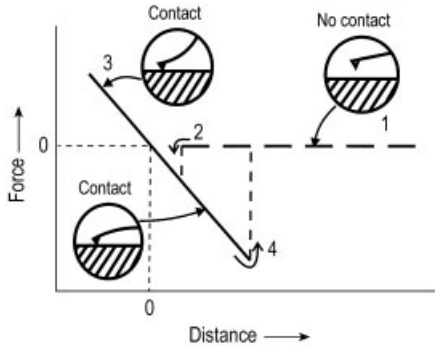

Figure 128. Force – distance curve depicting the interaction of the AFM tip with the sample surface. The operating range of the AFM is between positions 3 and 4. For explanation refer to the text.

One option for imaging surfaces under gentler conditions is tapping-mode AFM [475], [476]. In this mode of operation, the cantilever with the tip is driven near its resonance frequency by means of a piezo oscillator (Fig. 126), so that only intermittent contact between tip and sample occurs. In tapping-mode AFM to-

pographic information is retrieved from the amplitude signal of the oscillating cantilever. The advantage of this technique is that, due to the intermittent contact, lateral shear forces can be eliminated. Thus, scratching of soft samples or removal of loosely bound surface features (e.g., particles, thin films) can be avoided.

AFMs can be operated in two different modes: the constant-force mode and the constant-height mode. In the *constant-force mode*, the cantilever deflection, and hence the force, is kept constant by readjusting the sample in the vertical direction to follow the topographic features of the surface. In this mode comparatively large and rough sample areas can be imaged without destroying the tip, the sample surface, or both ("tip crash"), but the scan rates must be kept comparatively low to allow the feedback system to respond to height changes. In the *constant-height mode* the vertical position of the sample is kept constant, and the varying deflection of the cantilever is recorded. In this mode higher scan rates can be achieved, which is advantageous for eliminating thermal drifts in high-resolution imaging, but large scan sizes should be avoided, because tip crashes are possible.

5.1.2. Instrumentation

The following basic components of an AFM can be identified:

- A sharp tip mounted on a soft cantilever
- A detection system for measuring the deflection of the cantilever
- A piezoelectric translator to move the probe relative to the sample
- A feedback system to keep the deflection constant by height readjustment of the probe
- An imaging system to convert the single data points into an image

The tips and the cantilevers have to fulfill a number of requirements. The tip should be very sharp with a small radius of curvature and a high aspect ratio so that it can trace fine details on the surface. The cantilever must be softer than the bonds between the atoms of the sample to achieve a measurable deflection without destructive displacement of surface atoms. Moreover, the resonance frequency should be high to minimize noise due to coupling with low-frequency noise (e.g., vibrations of buildings) from the environment. Low spring constants and high resonance frequencies can only be achieved by making the cantilevers very small. Cantilevers with force constants of less than 0.1 N/m and resonance frequencies up to several hundreds of kilohertz can be produced by the multistep lithographic processes common in microelectronics technology. Such microfabricated cantilevers [477], [478] allow imaging with forces typically in the nanonewton and sub-nanonewton range. These forces are about 10 000 to 100 000 times lower than the force of gravity introduced by a fly (1 mg) sitting on a surface. Commercially available cantilevers are usually made of silicon nitride or single-crystal silicon. The length of such cantilevers is typically 100 to 200 μm, the total size of the integrated tip is on the order of 10 μm, and the radius of curvature at the apex can be as small as 5 – 10 nm.

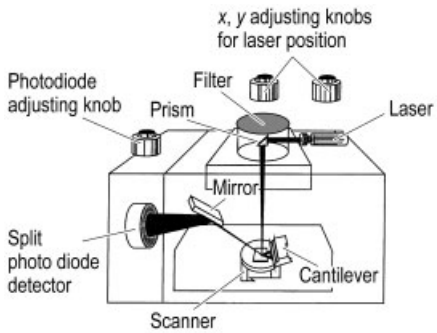

Figure 129. Schematic view of the deflection sensing system as used in the NanoScope III AFM (Digital Instruments, Santa Barbara, CA). The deflection of the cantilever is amplified by a laser beam focused on the rear of the cantilever and reflected towards a split-photodiode detector.

The force acting on the tip is sensed by measuring the deflection of the cantilever. A large number of options (e.g. tunneling current, interferometry, capacitance, laser-diode feedback) [474] for measuring this deflection have been described, but the most important one is the optical-lever technique [479] depicted in Figure 129. In this technique a laser beam is focused and positioned on the rear of the cantilever. The reflected laser beam is directed towards a double-segment photodiode. As the cantilever deflection, and thus the mirror plane for the laser beam,

changes, the position of the laser beam on the photodiode, and hence the difference signal between the two segments, changes, too. Therefore, the difference signal is a sensitive measure of the cantilever deflection, with a resolution of 0.01 nm.

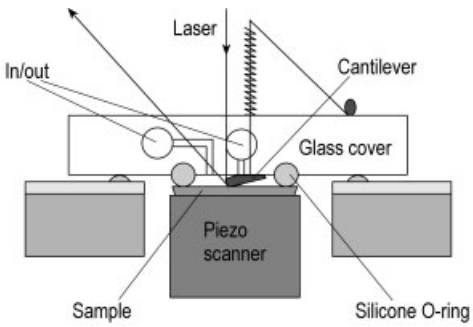

Figure 130. Cross-section view of a liquid cell for in situ AFM measurements of surface processes

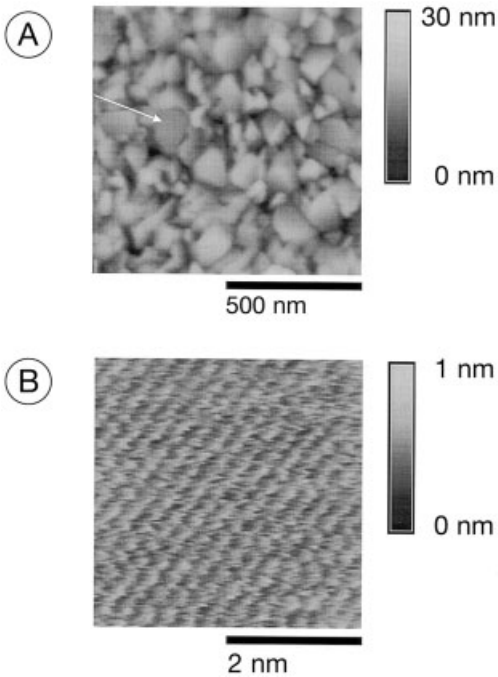

Figure 131. A) AFM image of a PVD gold film on silicon; image size: 1 × 1 µm, depth scale: 30 nm from black to white; B) Atomic-resolution image of the crystallite marked by the arrow in A; image size: 4 × 4 nm; depth scale: 1 nm from black to white

Usually, in AFM the position of the tip is fixed and the sample is raster-scanned. After manual coarse approach with fine-thread screws, motion of the sample is performed with a piezo translator made of piezo ceramics such as. lead zirconate titanate (PZT), which can be either a piezo tripod or a single-tube scanner. Single-tube scanners are more difficult to calibrate, but they can be built more rigidly and are therefore less sensitive towards vibrational perturbations.

In contrast to many other surface analytical techniques, such as scanning electron microscopy, AFM does not require a vacuum. Therefore, it can be operated under ambient conditions and allows processes at solid – gas and solid – liquid interfaces to be directly observed. The latter can be accomplished by means of a liquid cell, shown schematically in Figure 130. The cell is formed by the sample at the bottom, a glass cover holding the cantilever at the top, and a silicone O-ring seal between. Studies with such a liquid cell can also be performed under potential control, which opens up valuable opportunities for electrochemistry [480], [481]. Moreover, imaging under liquids opens up the possibility to protect sensitive surfaces by in situ preparation and imaging under an inert fluid [482].

5.1.3. Applications

The great analytical potential of AFM is based on a number of properties summarized in the following: Atomic resolution can be achieved, but imaging can also be performed on areas larger than 100 µm square. This capability allows overview images to be obtained, which can be used to zoom in on details with high resolution without changing the sample or the instrumental setup, which would make it impossible to find the identical sample position again. For example, Figure 131 A shows an image of a PVD gold film on silicon, which gives a good overview on the size and distribution of the single crystallites, which are typically 100 nm in diameter [483]. The root-mean-square (rms) roughness determined from this image is 3 nm. From XRD measurements it is known that the surface of the gold film is mainly a (111) surface. Atomic-resolution images of the position marked with the arrow in Figure 131 A revealed a gold (111) surface for that specific crystallite

(Fig. 131 B). Although the corrugation in one lattice direction is more prominent, which can be caused by convolution with an asymmetric tip, a hexagonal symmetry can be observed, and the spacings are in good agreement with the expected values (0.29 nm).

A further important advantage of AFM is the fact that images contain direct depth information, which makes the technique a valuable metrological tool. Figure 132 A shows, for example, an AFM image of a structured photoresist on a silicon surface. The pattern observed was produced by ion lithography. Figure 132 B depicts a cross-section profile through the image shown in Figure 132 A. This example nicely demonstrates that the AFM data can be conveniently analyzed by the software to directly measure the height of surface features.

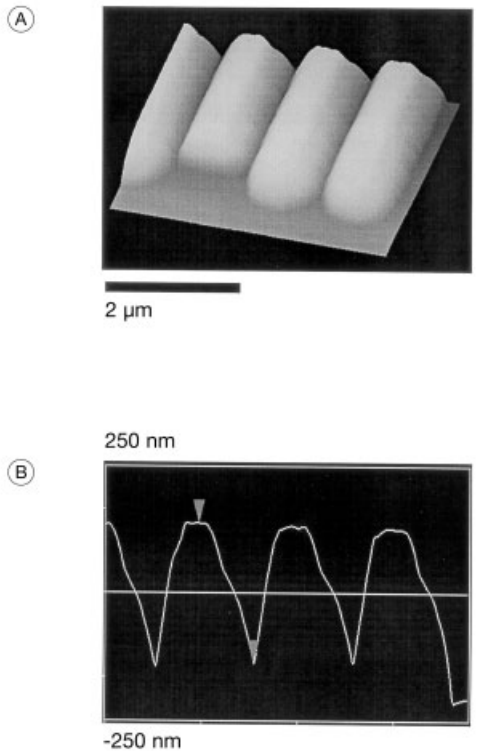

Figure 132. A) AFM image of a photoresist layer on silicon, generated by ion lithography; B) Cross-section profile through the AFM image shown in A

Besides conducting samples insulators can also be imaged readily with AFM without the need to coat them with a conductive film. This means that sample preparation is simple in many cases and that artefacts introduced by the coating process can be avoided. Therefore, a wide variety of organic [484], [485], biological [486], [487], and inorganic insulators [488] can be studied in a straightforward manner. Figure 133 shows as an example the topographic image of an organic filtration membrane. The pores in this Lavsan (polyethylene terephthalate) membrane were produced by α-particle bombardment, and their diameter is approximately 200 nm. Between the pores is a uniform surface with an rms roughness of 7 nm.

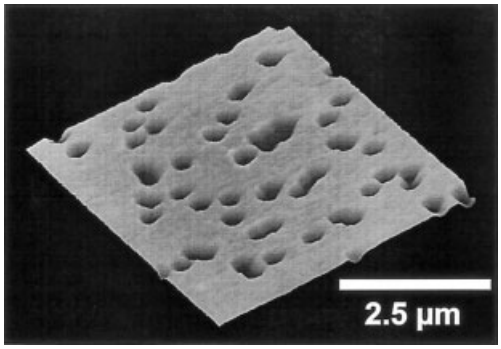

Figure 133. AFM image of a Lavsan filtration membrane; image size: $5 \times 5\,\mu m$, depth scale: 500 nm from black to white

An important potential of AFM for surface chemistry is its capability to perform in situ measurements under liquids and in air. This opens up the opportunity to directly investigate surface processes like corrosion, crystal growth, thin-film deposition, and electrochemical processes at electrode surfaces [480–482], [489], [490]. As an example of in situ AFM imaging, Figure 134 shows a series of AFM images depicting the deposition of an octadecylsiloxane (ODS) monolayer on silicon. Such self-assembled monolayers (SAMs) can be obtained by deposition of octadecyltrichlorosilane from toluene [491]. It can be seen that large islands with diameters of typically 5 µm are deposited on the surface right from the beginning of the growth experiment. This indicates that, under the given conditions,

such ordered structures already exist in the deposition solution. Once the uncovered surface area no longer offers sufficient space for deposition of further larger islands, continuing film formation can only proceed by growth of already deposited islands by attachment of smaller units, a process which is comparatively slow. During course of time the growing islands coalesce and it can take up to 1 h or more until full coverage of the surface is reached. From such AFM images surface coverages can also also be evaluated on a quantitative basis. Since AFM is able to reveal the structure of submonolayer islands in such deposition experiments with submicrometer resolution, it has significantly contributed to the understanding of growth mechanisms, which is important for controlling film growth by variation of deposition parameters such as the water content of the precursor solution or the temperature [491].

A shortcoming of AFM is that chemical information or material specific properties cannot be obtained from AFM images directly. However, a number of AFM-related techniques which allow information to be accessed in addition to the topography have been introduced. Friction force microscopy (FFM) or lateral force microscopy (LFM) measures the friction between tip and sample surface. This is of particular interest for research in the field of tribology [492], [493]. The principle of the FFM has also been used in a development called the chemical force microscope (CFM) [494]. In this method the tip is chemically modified to make it sensitive towards different chemical surface compositions. In this way image contrast could be achieved due to variations in adhesion (and thus also in friction) between tip and sample. In Young's modulus microscopy (YMM) [495], differences in the surface elasticity can be accessed by measuring of the response of the tip to the vertical modulation of the sample while the tip is in contact with the surface. The adhesion between tip and sample surface can also be visualized by force – distance curve measurements (Fig. 128) [496–498]. The electric force microscope (EFM) utilizes electrostatic forces or force gradients between a conductive tip and a sample for signal generation. In this way information on surface charges [499], [500], topography, capacitance (dielectric constant), and potential (see [474], and references cited therein) can be obtained. In magnetic force microscopy (MFM) [501], magnetostatic forces are measured by interaction of magnetic domains on a surface with a magnetized tip. MFM is mainly interesting as a tool for applications in manufacturing magnetic storage device [502] and allows magnetic patterns to be imaged with a resolution of ca. 50 nm.

5.2. Scanning Tunneling Microscopy (STM)

5.2.1. Principles

In scanning tunneling microscopy (STM) [470], [471] the probe is a sharp metal tip that is scanned across a conducting surface at distances on the order of typically 1 nm (Fig. 135). A bias voltage of typically a few millivolts is applied between the tip and the sample and leads to a tunneling current on the order of a few nanoamperes. Figure 136 shows what happens when a voltage is applied between two metal electrodes separated by a very small distance. If a voltage V is applied then the Fermi levels E_F are shifted relative each other by an energy eV, where e is the electrostatic charge of an electron. Due to the energy barrier between the two metals, which is on the order of the metal's work function ϕ, at small voltages a conventional current cannot flow. However, because the electrons can be described by wave functions decaying exponentially across the energy barrier, the probability of finding an electron on the opposite side of the barrier is not zero. This phenomenon is known as the quantum mechanical tunneling effect. As a consequence, electrons can flow from occupied states on the negative electrode to unoccupied states on the positive electrode, whereby at small voltages they do not change their energy, which is described as elastic tunneling. The dependence of the tunneling current I_T on the applied voltage V_T and the tip-surface separation s is given by

$$I_T \approx V_T \exp\left(-A\phi^{1/2}s\right) \tag{5.1}$$

where A is a constant given by $A = 2[(2m_e)^{1/2}]/h$ (m_e = mass of an electron). Due to this exponential relation, the tunneling current is an extremely sensitive measure of the separation. For

Table 9. Typical data for the more important techniques of surface and thin-film analysis

Technique	Primary probe	Elemental range	Type of information	Depth of information	Lateral resolution	Sensitivity, atom %	Ease of quantification	Insulator analysis	Destructive	UHV environment
XPS	X-ray photons (Al or $Mg_{K\alpha}$)	all except H, He	elemental, chemical	4–8 monolayers	100 µm (standard); 3 µm (imaging)	0.1	good	yes	sometimes	yes
UPS	UV photons (He I, He II)	(adsorbed molecules)	chemical (valence band)	2–10 monolayers	1 mm	1	n.a.***	yes	sometimes	yes
AES	e^- (3–10 keV)	all except H, He	elemental, (chemical)	2–6 monolayers	20 nm	0.2	moderate	no	often	yes
SAM	e^- (20-50 keV)	all except H, He	elemental, (chemical)	2–6 monolayers	5 nm	0.2	moderate	no	often	yes
SSIMS	Ar^+, Ga^+	all	chemical	1–2 monolayers	50 nm	10^{-6}–10^{-2}	poor	(yes)**	(yes)	yes
DSIMS	Ar^+, Ga^+	all	chemical	1–2 monolayers	20 nm (imaging); 50 µm (depth profile)	10^{-7}–10^{-1}	poor*	(yes)**	yes	yes
SNMS, e^- beam	Ar^+, Cs^+	all	elemental	1–2 monolayers	5 µm	10^{-2}–1	moderate*	yes (with rf)	yes	yes
SNMS, plasma	Ar^+, Kr^+	all	elemental	1–2 monolayers	1 mm	10^{-3}–10^{-1}	moderate*	yes	yes	yes
SNMS, laser	Ar^+, Ga^+	all	chemical, elemental	1–2 monolayers	50 nm	10^{-7}–1	moderate*	yes	yes	yes
RBS	H^+, He^+	high Z on low Z	elemental, depth distribution	1 monolayer to 2 µm	5 µm to 1 mm	10^{-6}	good	yes	no	no
ISS	He^+, Ne^+	all except H, He	elemental	1–2 monolayers	ca. 1 mm	10^{-4}	good	yes	not with He^+	yes
TXRF	X-ray photons		elemental	1 monolayer	(none)	10^{-4}–10^{-1}	yes	yes	no	no
GDOS	A^+ (Ar^+)	all	elemental, depth profile	10 nm	none	10^{-6}	moderate	yes	yes	no
RAIRS	IR photons	(adsorbed molecules)	chemical (vibrational)	1-2 monolayers	ca. 1 mm	10^{-4}	n.a.	no	no	no
ELL	UV/VIS/IR photons	index of reflection	chemical	1 nm	ca. 1 mm	1 of adsorbant	poor	(yes)	no	no
SERS	laser photons	restricted	chemical (vibrational)	10-100 nm	ca. 0.5 mm	10^{-6}	moderate	yes	no	yes
AFM	mechanical force	all	topographic	1 monolayer	1 atom	(1 atom)	n.a.	yes	no	no
STM, STS	high field	all	topographic, chemical	1 monolayer	1 atom	(1 atom)	n.a.	no	possibly	yes

* Good with near-matrix standards. ** With electron charge compensation. *** n.a. = not applicable.

Table 10. Suppliers of equipment for surface and thin-film analysis

Supplier	Electron spectroscopy	Mass spectroscopy	Photon spectroscopy	Scanning probe microscopy
Gammadata Scienta P.O. Box 15120 750 15 Uppsala Sweden www.gammadata.se	XPS			
Jeol USA, Inc. 11 Dearborn Road Peabody MA 01960 USA www.jeol.com	XPS, AES	SIMS		
Kratos Analytical Wharfside Trafford Wharf Road Manchester, M17 1GP United Kingdom www.kratos.com	XPS, AES			
Omicron Vakuumphysik GmbH Idsteiner Straße 78 65232 Taunusstein Germany www.omicron-instruments.com	XPS, AES, SAM, UPS, LEED	ISS		STM, AFM
Physical Electronics, Inc. 6509 Flying Cloud Drive Eden Prairie MN 55344 USA www.phi.com	XPS, AES, SAM, UPS	SIMS		
Specs GmbH Voltastraße 5 13355 Berlin Germany www.specs.de	XPS, AES, UPS, LEED	SIMS, SNMS, ISS		
Staib Instruments GmbH Obere Hauptstraße 45 85354 Freising Germany www.staib-instruments.com	AES, RHEED			
tectra GmbH Reuterweg 65 60323 Frankfurt a.M. Germany www.tectra.de	LEED			
VG Scientific The Birches Industrial Estate Imberhorne Lane East Grinstead West Sussex RH19 1UB	XPS, AES, SAM, UPS, HREELS	SIMS, SNMS, GDMS, APFIM, ISS	BIS	STM, AFM

Table 10. (Continued)

Supplier	Electron spectroscopy	Mass spectroscopy	Photon spectroscopy	Scanning probe microscopy
United Kingdom www.vgscientific.com				
VSW Ltd. Unit 4, Heather Close Lyme Green Business Park Macclesfield, Cheshire SK11 0LR www.vsw.co.uk	XPS, AES, UPS, EELS	ISS	BIS	
Atomika Instruments Bruckmannring 40 85764 Oberschleißheim Germany www.atomika.com		SIMS	TXRF	
Cameca FRANCE 103 Boulevard Saint Denis 92403 Courbevoie Cedex France www.cameca.fr		SIMS		
Hiden Analytial Ltd. 420 Europa Boulevard Warrington WA5 5UN United Kingdom www.hiden.co.uk		SIMS		
National Electrostatics Corp. 7540 Graber Road P.O. Box 620310 Middleton WI 53562-0310 USA www.pelletron.com		RBS		
Charles Evans & Associates 810, Kifer Road Sunnyvale CA 94086 USA www.cea.com		RBS		
Ital Structures Via M. Miscone 11/d - Z.I. Baltera 38066 Riva del Garda Italy www.italstructures.com			TXRF	
Philips Analytical Lelyweg 1 7602 EA Almelo The Netherlands www.analytical.philips.com			TXRF, ELL	
Rigaku International Corporation 3-9-12 Matsubara-cho Akishima-shi, Tokyo 196 Japan			TXRF	

Table 10. (Continued)

Supplier	Electron spectroscopy	Mass spectroscopy	Photon spectroscopy	Scanning probe microscopy
www.rigaku.com				
Jobin Yvon S.A. 16-18, rue du Canal 91165 Longjumeau Cedex France www.jyinc.com			GD-OES, RAIRS, ELL	
LECO Instrumente GmbH Benzstraße 5b 85551 Kirchheim bei München Germany www.leco.com			GD-OES	
SPECTRUMA Spectro Analytical Instruments GmbH Boschstraße 10 47533 Kleve Germany www.spectruma.de			GD-OES	
ABB Bomem inc. 585 Charest East, Suite 300 Québec, QC G1K 9H4 Canada www.bomem.com			RAIRS	
Beaglehole Instruments Ltd. 127 Kelburn Parade Wellington New Zealand www.beaglehole.com			ELL	
J.A. Woollam Co. 650 J Street Suite 39 Lincoln NE 68508 USA www.jawoollam.com			ELL	
NFT-Nanofilm Technologie GmbH Anna-Vandenhoeck-Ring 5 37081 Goettingen Germany www.nanofilm.de			ELL	
Sentech Gesellschaft für Sensortechnik GmbH Bahnstraße 3 82131 Stockdorf Germany www.sentech.com			ELL	
Sopra, Inc. 26, rue Pierre Joigneaux 92270 Bois-Colombes			ELL	

Table 10. (Continued)

Supplier	Electron spectroscopy	Mass spectroscopy	Photon spectroscopy	Scanning probe microscopy
France www.sopra-sa.com				
Waterloo Digital Electronics Division of WDE Inc. 279 Weber St. N. Waterloo, Ontario N2J 3H8 Canada www.onramp.tuscaloosa.al.us			ELL	
Bruker Optics Inc. 19 Fortune Dr. Manning Park Billerica MA 01821-3991 USA www.bruker.com			RAIRS, Raman	
ChemIcon Inc. 7301 Penn Avenue Pittsburgh Pennsylvania 15208 USA www.chemimage.com			RAIRS, Raman	
Chromex 2705-B Pan American Fwy NE Albuquerque NM 87107 USA www.chromexinc.com			Raman	
Kaiser Optical Systems Inc. Box 983 371 Parkland Plaza Ann Arbor MI 48106-0983 USA www.kosi.com			Raman	
Nicolet Instrument Corporation 5225 Verona RD Bldg 5 Madison WI 53711-4495 USA www.nicolet.com			RAIRS, Raman	
Ocean Optics Europe Soerense Zand 4a 6961 LL Eerbeek The Netherlands www.topsensors.nl			Raman	
Perkin Elmer Instruments 761 Main Avenue Norwalk CT 06859 USA instruments.perkinelmer.com			RAIRS, Raman	
Renishaw plc			Raman	

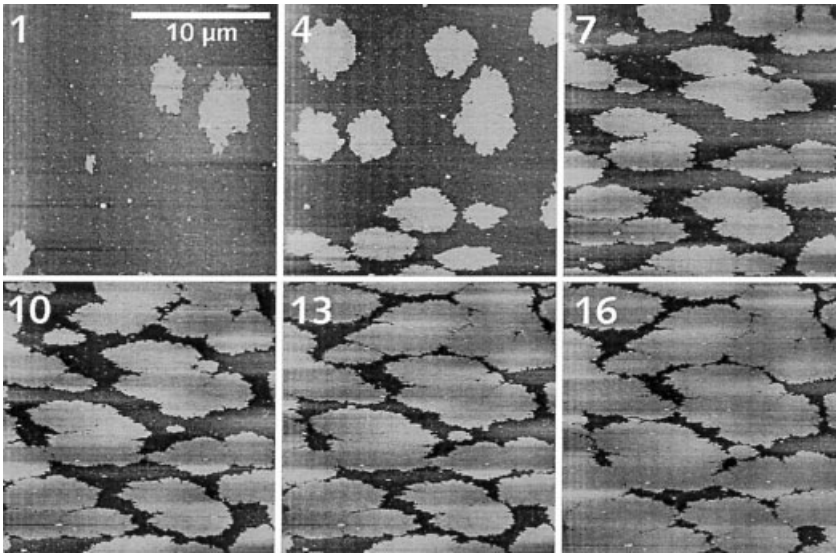

Figure 134. Series of sequential in situ AFM images of the growth of ODS film on silicon. The numbers indicate the adsorption time in minutes.

Table 10. (Continued)

Supplier	Electron spectroscopy	Mass spectroscopy	Photon spectroscopy	Scanning probe microscopy
New Mills Wotton-under-Edge Gloucestershire GL12 8JR United Kingdom www.renishaw.com				
Burleigh Instruments Inc. Burleigh Park Fishers NY 14453 USA www.burleigh.com				STM, AFM
Digital Instruments GmbH Janderstraße 9 68199 Mannheim Germany www.digital-instruments.com				STM, AFM
LK Technologies Inc. 3910 Roll Avenue Bloomington IN 47403 USA www.lktech.com				STM
ThermoMicroscopes 1171 Borregas Avenue Sunnyvale CA 94089 USA www.thermomicro.com				AFM, STM

example, a variation of the separation by 0.1 nm changes the tunneling current by a factor of approximately 10. This means that in principle variations in the separation of 0.001 nm can be monitored by keeping the current constant within a few percent. However, the tunneling current depends not only on the tip – surface separation, but also on the electronic structure of the surface. This is important when interpreting STM images on the atomic scale, because in that case, in contrast to AFM, in general the image

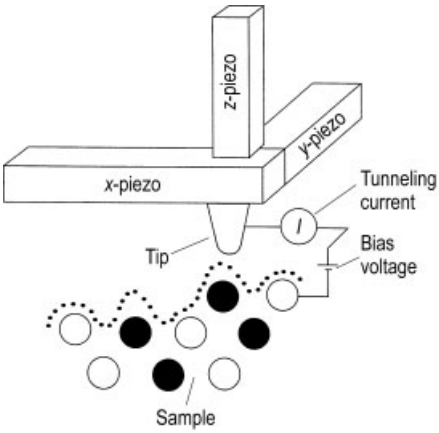

Figure 135. Principle of STM

no longer simply shows the topography of the surface, but rather a contour map of the local density of states near the Fermi level ($LDOS_{E_F}$) if small voltages in the millivolt range are applied. For larger voltages up to a few volts the density of other electronic states can be imaged. Note that the relations given above are only an approximation under the assumption that the tip and the applied voltage do not influence the electronic structure of the surface. Furthermore, for an exact description, three-dimensional tunneling must be considered to take into account the complex spatial structure of the electronic states (wavefunctions) involved [504].

There are two ways of operating an STM: the constant-current mode and the constant-height mode. In the *constant-current mode*, the tip is scanned across the sample surface at a fixed bias voltage, and the tunneling current is kept constant by readjusting the tip in the vertical direction to follow the topographic features (or, on the atomic level, the local density of states) on the surface. In this mode, comparatively large and rough sample areas can be imaged without destroying the tip, the sample surface, or both ("tip crash"), but the scan rates must be kept comparatively low to allow the feedback system to respond to height changes. For images taken in constant-current mode at a certain bias voltage, the term constant-current topograph (CCT) is sometimes used. In the *constant-height mode* the vertical position of the tip is kept constant and the varying tunneling current is recorded. In this mode higher scan rates can be achieved, which is advantageous for eliminating thermal drifts in high-resolution imaging, however, large scan areas should be avoided, because tip crashes are possible.

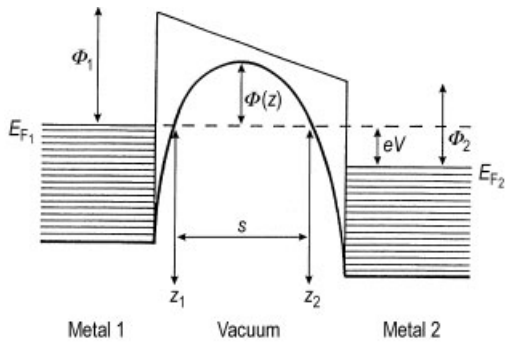

Figure 136. Energy level diagram of a tunneling junction [503]

5.2.2. Instrumentation

Figure 137 shows a schematic overview of an STM system consisting of the following basic components:

- A sharp metal tip
- A piezoelectric translator to move the tip relative to the sample
- Control electronics for applying the bias voltage and measuring the tunneling current
- A feedback system to keep the tunneling current constant by height adjustment of the tip
- An imaging system to convert the single data points into an image

The tips used for STM experiments should be sharp and stable. Chemical stability can be

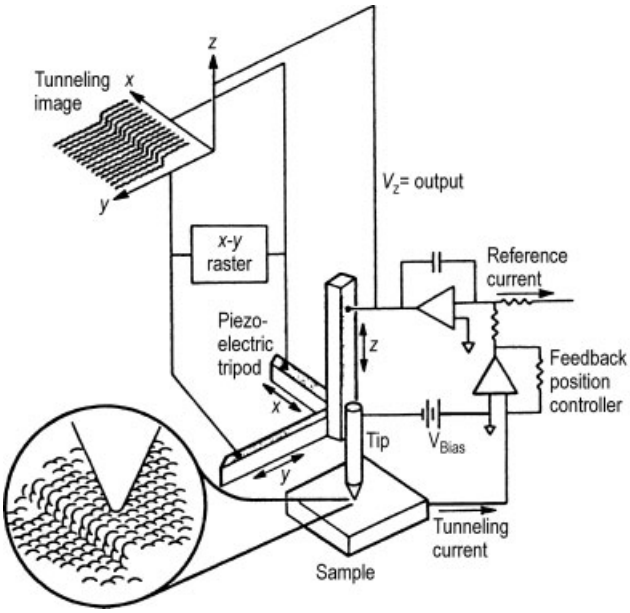

Figure 137. Scheme of an STM system [505]

achieved by using a noble metal. Mechanical rigidity can be attained by using short wires. Alloys of Pt and Ir are frequently used for fabrication of STM tips. They can be produced in a surprisingly simple way just by cutting a metal wire with conventional cutting tools. Due to their high chemical stability, such Pt – Ir tips are well suited for atomic-resolution experiments on flat samples. However, due to their low aspect ratio near the apex they fail to trace steep features and narrow trenches. Therefore, electrolytically etched tungsten tips with higher aspect ratios are sometimes used as an alternative, although they are less stable against oxidation.

Usually, in STM the position of the sample is fixed and the tip is raster-scanned. Like in AFM, after manual coarse approach with fine-thread screws, motion of the tip is performed with a piezo translator made of piezo ceramics like lead zirconate titanate (PZT), which can be either a piezo tripod or a single-tube scanner.

Although physical studies of the electronic structure of surfaces have to be performed under UHV conditions to guarantee clean uncontaminated samples, the technique does not require vacuum for its operation. Thus, in-situ observation of processes at solid – gas and solid – liquid interfaces is possible. This has been utilized, for instance, to directly observe corrosion and electrode processes with atomic resolution [471], [506].

5.2.3. Lateral and Spectroscopic Information

STM images can only be interpreted in terms of surface topography for surface structures with dimensions well above the atomic scale. In general, images taken in constant-current mode deliver contour maps of the local density of states. If the polarity of the sample is negative, then states in the valence band are imaged. For a positive polarity of the sample the distribution of electronic states in the conduction band can be recorded. If different chemical species are present on the surface, the image contrast is further influenced by the varying effective barrier height (work function) at different positions. As an example Figure 138 shows an STM image of a silicon (111) surface with one-third of a silver monolayer on top of it [507]. The high contrast between the silicon surface and the silver islands does not represent the real height of the

islands, which appears exaggerated due to the lower work function in the silver regions. Such local differences in the effective barrier height can be directly imaged by modulating the tip in the vertical direction and recording dI/ds, which according to Equation (5.1) is proportional to $\phi^{1/2}xI$. STM can also yield spectroscopic information by recording dI/dU curves at fixed positions, which is called scanning tunneling spectroscopy (STS). This opens up the opportunity to perform electron spectroscopy with a resolution down to a single atom. An other option for obtaining spectroscopic information is simultaneous recording of STM images at various bias voltages. This can be accomplished by performing point spectroscopy at every image point or by modulating the bias voltage while scanning. In the literature simultaneous recording of images at various bias voltages has been called current imaging tunneling spectroscopy (CITS). An example is the discrimination between Ga and As atoms on a GaAs (110) surface in an atomically resolved STM image by simultaneously recording of two images at reversed bias voltage by Feenstra et al. [508].

5.2.4. Applications

Similar to AFM, an advantage of STM is the fact that a wide range of scan sizes from more than 100 µm down to the atomic level can be covered in one experiment. Because the information is obtained in real space, local defects (e.g. monoatomic defects, steps, dislocations) can be investigated. This clearly is a great advantage compared to diffraction methods, which rely on extended periodic structures and thus show averaged information. Moreover, the possibility to obtain spectroscopic information makes the STM a very valuable tool for studying surface processes on the atomic scale. An instructive example highlighting several figures of merit of STM is given in [509–511] and described below. Figure 139 shows an STM image of the unoccupied states of a Si (111) 7×7 surface exposed to 0.2 L of O_2 at 300 K. Besides the characteristic pattern of the 7×7 structure of the Si (111) surface, pronounced dark and bright spots can be observed as a result of exposure to oxygen. To clarify the mechanistic origin of these spots, tunneling spectra at various atomic sites in the 7×7 unit cell (Fig. 140) were recorded and compared with UPS spectra and theoretical results from tight-binding slab calculations for various adsorption models of oxygen. First of all, Figure 140 shows that the different sites can be clearly distinguished by the respective tunneling spectra, whereas photoemission spectra only deliver averaged information on the electronic structure of the surface. The spectra labeled A, B, and C were obtained on sites that were unchanged by adsorption of oxygen. The spectra labeled D, E, and F were obtained on oxygen-induced dark, bright, and perturbed (gray) adatom sites, respectively. Comparison of these spectra with theoretical calculations and a thorough evaluation, including UPS spectra for different temperatures and oxygen pressures, led to the conclusion that the bright spots (E) are silicon atoms which have maintained their dangling bond but are perturbed by a neighboring oxygen atom inserted in the back bond of the adatom. The dark spots were interpreted as oxygen atoms saturating the dangling bond of a silicon atom which has another oxygen inserted in its back bond. Although this is a rather early work in the field of scanning tunneling spectroscopy, it is still an excellent example for atom-resolved investigation of surface chemistry with STM.

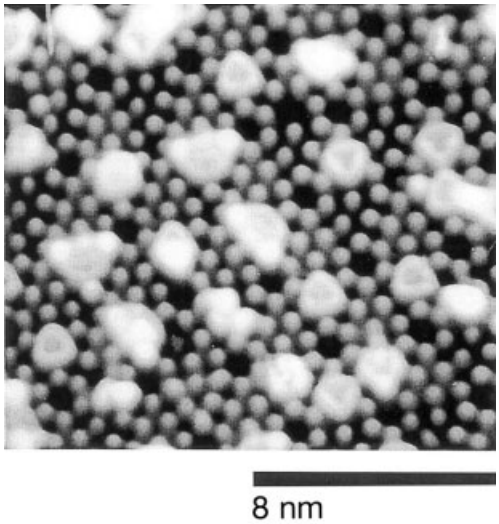

Figure 138. STM image of silicon covered with one-third of a monolayer of silver [507]

STM has, however, not only been applied to metals and semiconductors, but also to a wide variety of organic and biological systems, such as thin films on conducting substrates [484], [485] as well as protein and DNA molecules [512–515].

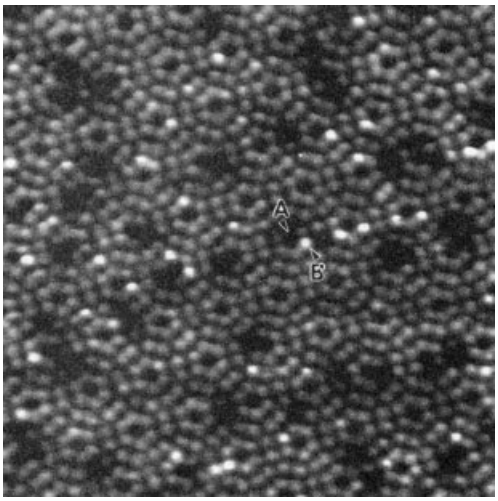

Figure 139. STM image of a Si (111) 7×7 surface exposed to 0.2 L of O_2 at 300 K. The sample bias voltage was + 2 V. Dark and bright sites generated by oxygen exposure are marked with A and B, respectively [510].

A variant of STM is *spin-polarized STM* (SP-STM), which works with a ferromagnetic tip and can be used to probe surface magnetism with high resolution [516], [517]. Other modifications of STM involve electromagnetic radiation, whereby two basic concepts can be distinguished: measuring of photons generated in the tunneling gap by inelastic tunneling effects [518], [519], which is called *photon-emission scanning tunneling microscopy* (PESTM), and generation of electrical currents by irradiation of the tunneling gap [520–523], which is know as *photon-assisted scanning tunneling microscopy* (PASTM).

6. Summary and Comparison of Techniques

Important surface analytical techniques are summarized and compared in Table 9.

7. Surface Analytical Equipment Suppliers

A large number of suppliers of surface analytical equipment exist; a few provide for several techniques; the majority specialize in just one or two. Only suppliers of equipment for techniques that are discussed in more detail are listed in Table 10. The authors cannot warrant for completeness.

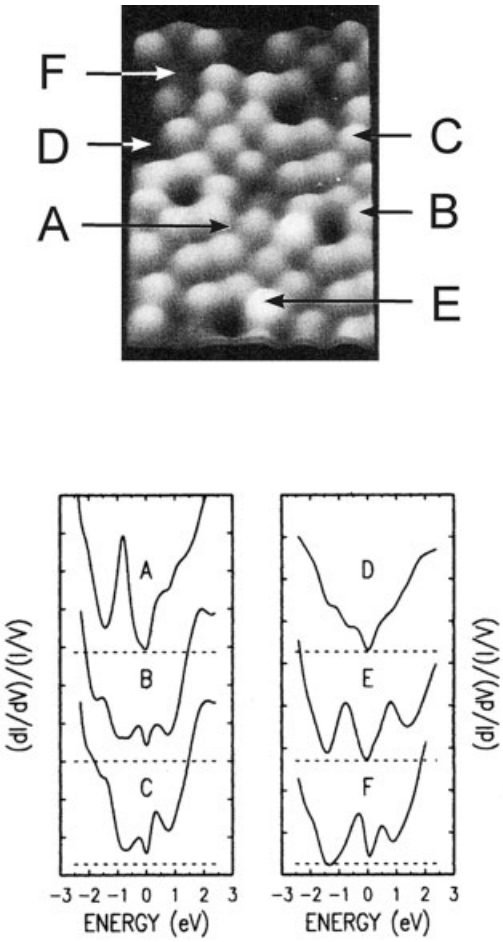

Figure 140. Top: STM image of a region of an O_2-exposed Si (111) 7×7 surface. The sites labeled A, B, and C have not been changed by exposure to oxygen; D, E, and F correspond to oxygen-induced dark, bright, and perturbed (gray) sites, respectively. Bottom: tunneling spectra corresponding to the different sites [510].

8. References

1. J. C. Rivière: *Surface Analytical Techniques,* Oxford University Press, Oxford 1990.
2. M. P. Seah, D. Briggs in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 1,** Wiley, Chichester 1990, pp. 1 – 18.
3. D. M. Hercules, S. H. Hercules, *J. Chem. Educ.* **61** (1984) 402 – 409.
4. M. J. Higatsberger, *Adv. Electron. Electron. Phys.* **56** (1981) 291 – 358.
5. H. W. Werner, *Mikrochim. Acta Suppl.* **8** (1979) 25 – 50.
6. A. E. Morgan, H. W. Werner, *Phys. Scr.* **18** (1978) 451 – 463.
7. J. Tousset, *Le Vide* **33** (1978) 201 – 211.
8. G. B. Larrabee, *Scanning Electron Microscopy*, part I, 1977, 639 – 650.
9. A. Benninghoven, *Appl. Phys.* **1** (1973) 3 – 16.
10. K. Siegbahn, C. Nordling, A. Fahlman, R. Nordberg, K. Hamrin, J. Hedman, G. Johansson, T. Bergmark, S.-E. Karlsson, I. Lindgreen, B. Lindberg: *ESCA: Atomic, Molecular, and Solid State Structure Studied by Means of Electron Spectroscopy,* Almqvist and Wiksells, Uppsala 1967.
11. P. Auger, *J. Phys. Radium* **6** (1925) 205 – 209.
12. M. P. Seah, W. A. Dench, *Surf. Interface Anal.* **1** (1979) 2 – 11.
13. D. Briggs, J. C. Rivière in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 1,** Wiley, Chichester 1990, pp. 85 – 141.
14. M. P. Seah in J. M. Walls (ed.): *Methods of Surface Analysis,* Cambridge University Press, Cambridge 1989.
15. B. P. Tonner, T. Droubay, J. Denlinger, W. Meyer-Ilse, T. Warwick, J. Rothe, E. Kneedler, K. Pecher, K. Nealson, T. Grundl, *Surf. Interface Anal.* **27** (1999) 247 – 258.
16. J. H. Scofield, *J. Electron. Spectrosc. Relat. Phenom.* **8** (1976) 129 – 137.
17. R. F. Reilman, A. Msezane, S. T. Manson, *J. Electron Spectrosc. Relat. Phenom.* **8** (1976) 389 – 394.
18. M. P. Seah in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 1,** Wiley, Chichester 1990, pp. 201 – 255.
19. P. Brüsch, K. Müller, A. Atrens, H. Neff, *Appl. Phys. A* **38** (1985) 1 – 18.
20. S. Hofmann in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 1,** Wiley, Chichester 1990, pp. 143 – 199.
21. S. W. Gaarenstroom, *J. Vac. Sci. Technol.* **16** (1979) 600 – 604.
22. S. Hofmann, J. Steffen, *Surf. Interface Anal.* **14** (1989) 59 – 65.
23. H. Bubert, H. Jenett, *Fresenius' Z. Anal. Chem.* **335** (1992) 643 – 647.
24. C. Gatts, A. Zalar, S. Hofmann, M. Rühle, *Surf. Interface Anal.* **23** (1995) 809 – 814.
25. J. Card, A. L. Testoni, L. A. Le Tarte, *Surf. Interface Anal.* **23** (1995) 495 – 505.
26. T. Do, N. S. McIntyre, *Surf. Interface Anal.* **27** (1999) 1037 – 1045.
27. R. Reiche, S. Oswald, K. Wetzig, M. Dobler, H. Reuther, M. Walterfang, *Nucl. Instrum. Methods Phys. Res. B* **160** (2000) 397 – 407.
28. T. Morohashi, T. Hoshi, H. Nikaido, M. Kudo, *Appl. Surf. Sci.* **100/101** (1996) 84 – 88.
29. M. J. G. Wenham, I. R. Barkshire, M. Prutton, R. H. Roberts, D. K. Wilkinson, *Surf. Interface Anal.* **23** (1995) 858 – 872.
30. M. N. Souza, C. Gatts, M. A. Figueira, *Surf. Interface Anal.* **20** (1993) 1047 – 1050.
31. H. Bubert, N. Niebuhr, *J. Microsc. Soc. Am.* **2** (1996) 35 – 41.
32. T. Watanabe, S. Kishida, K. Ishihara, Y. Yamauchi, H. Tokutaka, *Hyomen Kagaku* **19** (1998) 98 – 105.
33. P. Grunow, H. Paes, C. Gatts, W. Losch, *Thin Solid Films* **275** (1996) 191 – 194.
34. H. Tokutaka, K. Yoshihara, K. Fujimura, K. Iwamoto, K. Obu-Cann, *Surf. Interface Anal.* **27** (1999) 783 – 788.
35. B. Lesiak, J. Zemek, A. Jozwik, *Appl. Surf. Sci.* **135** (1998) 318 – 330.
36. H. Bubert, H. Hillig, *Mikrochim. Acta* **133** (1999) 95 – 103.
37. A. E. Munter, B. J. Heuser, K. M. Skulina, *Physica B (Amsterdam)* **221** (1996) 500 – 506.
38. P. Coxon, J. Krizek, M. Humpherson, I. R. M. Wardell, *J. Electron Spectrosc. Relat. Phenom.* **52** (1990) 821 – 836.
39. C. D. Wagner, *Faraday Discuss. Chem. Soc.* **60** (1975) 291 – 300.
40. C. D. Wagner, L. H. Gale, R. H. Raymond, *Anal. Chem.* **51** (1979) 466 – 482.
41. C. D. Wagner in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 1,** Wiley, Chichester 1990, pp. 595 – 634.
42. I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997.
43. D. Briggs, J. P. Espinós, J. M. Sanz (eds.): *ECASIA 99,* Wiley Interscience, New York 2000.
44. J. M. Walls (ed.): *Methods of Surface Analysis,* Cambridge University Press, Cambridge 1989.

45. D. P. Woodruff, T. A. Delchar: *Modern Techniques of Surface Analysis,* Cambridge University Press, Cambridge 1988.
46. J. F. Watts: *An Introduction to Surface Analysis by Electron Spectroscopy,* Oxford University Press, Oxford 1990.
47. G. Ertl, J. Küppers: *Low Energy Electrons and Surface Chemistry,* 2nd ed., VCH, Weinheim 1985.
48. D. Briggs (ed.): *Handbook of X-Ray and Ultraviolet Photoelectron Spectroscopy,* Heyden, London 1977.
49. J. Wolstenholme, *Mater. World.* **7** (1999) 412 – 414.
50. J. M. López Nieto, V. Cortés Corberán, J. L. G. Fierro, *Surf. Interface Anal.* **17** (1991) 940 – 946.
51. T. A. Patterson, J. C. Carver, D. E. Leyden, D. M. Hercules, *J. Phys. Chem.* **80** (1976) 1700 – 1708.
52. A. Paul Pijpers, R. J. Meier, *Chem. Soc. Rev.* **28** (1999) 233 – 238.
53. A. Boix, J. L. G. Fierro, *Surf. Interface Anal.* **27** (1999) 1107 – 1113.
54. T. Fujitani, I. Nakamura, S. Ueno, T. Uchijiama, J. Nakamura, *Appl. Surf. Sci.* **121/122** (1997) 583 – 586.
55. T. Miyazaki, T. Doi, M. Kato, T. Miyake, I. Matsuura, *Appl. Surf. Sci.* **121/122** (1997) 492 – 495.
56. D. Robba, C. Casale, M. Notaro, D. M. Ori in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 169 – 72.
57. A. Galtayries, J. Grimblot, J.-P. Bonnelle, *Surf. Interface Anal.* **24** (1996) 345 – 354.
58. G. Beamson, G. Briggs, *High Resolution XPS of Organic Polymers,* Wiley, Chicester 1992.
59. D. Briggs, *Surface Analysis of Polymers by XPS and Static SIMS,* Cambridge University Press, Cambridge 1998.
60. A. A. Galuska, *Surf. Interface Anal.* **27** (1999) 889 – 896.
61. D. Wolany, T. Fladung, L. Duda, J. W. Lee, T. Gantenfort, L. Wiedmann, A. Benninghoven, *Surf. Interface Anal.* **27** (1999) 609 – 617.
62. J. Marchand-Brynaert, *Surfactant Sci. Ser.* **79** (1999) 91 – 124.
63. L. Dai, A. W. H. Albert, *J. Phys. Chem. B* **104** (2000) 1891 – 1915.
64. J. E. Fulghum, *J. Electron Spectrosc. Relat. Phenom.* **100** (1999) 331 – 355.
65. K. Asami, K. Hashimoto, *Langmuir* **3** (1987) 897 – 904.
66. R. Ramanauskas, *Appl. Surf. Sci.* **153** (1999) 53 – 64.
67. G. Pajonk, H. Bubert, *Fresenius' J. Anal. Chem.* **365** (1999) 236 – 243.
68. F. Bentiss, M. Traisnel, L. Gengembre, M. Lagrenee, *Appl. Surf. Sci.* **152** (1999) 237 – 249.
69. J. H. Qiu, P. H. Chua, *Surf. Interface Anal.* **28** (1999) 119 – 122.
70. J. F. Marco, A. C. Agudelo, J. R. Gancedo, D. Hanzel, *Surf. Interface Anal.* **26** (1998) 667 – 673.
71. J. F. Marco, A. C. Agudelo, J. R. Gancedo, D. Hanzel, *Surf. Interface Anal.* **27** (1999) 71 – 75.
72. L.-S. Johansson, T. Saastamoinen, *Appl. Surf. Sci.* **144/145** (1999) 244 – 248.
73. G. Chen, C. R. Clayton, *Surf. Interface Anal.* **27** (1999) 230 – 235.
74. D. Sprenger, H. Bach, W. Meisel, P. Guetlich, *Surf. Interface Anal.* **20** (1993) 796 – 802.
75. K. Idla, A. Talo, H. E.-M. Niemi, O. Forsen, S. Ylasaari, *Surf. Interface Anal.* **25** (1997) 837 – 854.
76. C. Ziegler, B. Reusch, J. Geis-Gerstorfer, *Fresenius' J. Anal. Chem.* **361** (1998) 547 – 553.
77. E. McCafferty, *Proc. - Electrochem. Soc.* **98 – 17** (1999) 42 – 55.
78. Chin-An Chang, Yong-Kil Kim, A. G. Schrott, *J. Vac. Sci. Technol. A* **8** (1990) 3304 – 3309.
79. M. F. Fitzpatrick, J. F. Watts, *Surf. Interface Anal.* **27** (1999) 705 – 715.
80. J. F. Watts, M. F. Fitzpatrick, T. J. Carney in *ECASIA97,* Wiley, Chichester 1997, pp. 113 – 116.
81. D. Wolany, T. Fladung, L. Duda, J. W. Lee, T. Gantenfort, L. Wiedmann, A. Benninghoven, *Surf. Interface Anal.* **27** (1999) 609 – 617.
82. C. V. Bechtolsheim, V. Zaporojtchenko, F. Faupel, *Appl. Surf. Sci.* **151** (1999) 119 – 128.
83. J. Zhang, C. Q. Cui, T. B. Lim, E.-T. Kang, K. G. Neoh, *Surf. Interface Anal.* **28** (1999) 235 – 239.
84. H. Horn, S. Beil, D. A. Wesner, R. Weichenhain, E. W. Kreutz, *Nucl. Instrum. Methods Phys. Res., Sect. B* **151** (1999) 279 – 284.
85. E. Capelli, S. Orlando, F. Pinzari, A. Napoli, S. Kaciulis, *Appl. Surf. Sci.* **138/139** (1999) 376 – 382.
86. P. Steiner, V. Kinsinger, I. Sander, B. Siegwart, S. Huefner, C. Politis, R. Hoppe, H. P. Mueller, *Z. Phys. B* **67** (1987) 497 – 502.
87. C. Ziegler, F. Baudenbacher, H. Karl, H. Kinder, W. Göpel, *Fresenius' J. Anal. Chem.* **341** (1991) 308 – 313.

88. Y. Hwu, L. Lozzi, S. La Rosa, M. Onellion, H. Berger, F. Gozzo, F. Levy, G. Margaritondo, *Appl. Phys. Lett.* **59** (1991) 979 – 981.
89. Y. Hwu, M. Marsi, C. Hwang, J. Seutjens, D. C. Larbalestier, M. Onellion, G. Margaritondo, *Appl. Phys. Lett.* **57** (1990) 2139 – 2141.
90. Q. Y. Ma, M. T. Schmidt, L. S. Weinman, E. S. Yang, S. M. Sampere, S. W. Chan, *J. Vac. Sci. Technol. A* **9** (1991) 390 – 393.
91. R. P. Vasquez, M. C. Foote, B. D. Hunt, L. Bajuk, *J. Vac. Sci. Technol. A* **9** (1991) 570 – 573.
92. I. N. Shabanova, O. V. Popova, V. I. Kormilets, N. M. Nebogaticov, V. I. Kukuev in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 917 – 920.
93. H. Behner, G. Gieres, B. Sipos, *Fresenius' J. Anal. Chem.* **341** (1991) 301 -307.
94. W. A. M. Aarnink, J. Gao, H. Rogalla, A. van Silfhout, *J. Less Common Met.* **164/165** (1990) 321 – 328.
95. C. C. Chang, M. S. Hedge, X. D. Wu, B. Dutta, A. Inam, T. Venkatesan, B. J. Wilkens, J. B. Wachtman Jr., *J. Appl. Phys.* **67** (1990) 7483 – 7487.
96. G. Frank, C. Ziegler, W. Göpel, *Phys. Rev. B. Condens. Matter* **43** (1991) 2828 – 2834.
97. S. Tanaka, T. Nakamura, M. Iiyama, N. Yoshida, S. Takano, *Jpn. J. Appl. Phys.* **30** (1991) L 1458 – L 1461.
98. S. Kohiki, S. Hatta, T. Kamada, A. Enokihara, T. Satoh, K. Setsune, K. Wasa, Y. Higashi, S. Fukushima, Y. Gohshi, *Appl. Phys. A* **50** (1990) 509 – 514.
99. S. Kohiki, J. Kawai, S. Hayashi, H. Adachi, S. Hatta, K. Setsune, K. Wasa, *J. Appl. Phys.* **68** (1990) 1229 – 1232.
100. M. A. Sobolewski, S. Semancik, E. S. Hellman, E. H. Hartford, *J. Vac. Sci. Technol. A* **9** (1991) 2716 – 2720.
101. D. R. C. Hoad, W. R. Flavell, S. E. Male, I. W. Fletcher, G. Beamson, *Surf. Interface Anal.* **21** (1994) 764 – 770.
102. G. Subramayam, F. Radpour, V. J. Kapoor, G. H. Lemon, *J. Appl. Phys.* **68** (1990) 1157 – 1163.
103. K. Tanaka, H. Takaki, K. Koyama, S. Noguchi, *Jpn. J. Appl. Phys.* **29** (1990) 1658 – 1663.
104. H. Yamanaka, H. Enomoto, J. S. Shin, T. Kishimoto, Y. Takano, N. Mori, H. Ozaki, *Jpn. J. Appl. Phys.* **30** (1991) 645 – 649.
105. T. Suzuki, M. Nagoshi, Y. Fukuda, K. Ohishi, Y. Syono, M. Tachiki, *Phys. Rev. B Condens. Matter* **42** (1990) 4263 – 4271.
106. A. Fujimori, Y. Tokura, H. Eisaki, H. Takagi, S. Uchida, E. Takayama-Muromachi, *Phys. Rev. B Condens. Matter* **42** (1990) 325 – 328.
107. M. Klauda, J. P. Stroebel, J. Schloetterer, A. Grassmann, J. Markl, G. Saemann-Ischenko, *Physica C* **173** (1991) 109 – 116.
108. K. Okada, Y. Seino, A. Kotani, *J. Phys. Soc. Jpn.* **60** (1991) 1040 – 1050.
109. A. Hartmann, G. J. Russel, D. N. Matthews, J. W. Cochrane, *Surf. Interface Anal.* **24** (1996) 657 – 661.
110. P. Adler, H. Buchkremer-Hermanns, A. Simon, *Z. Phys. B Condens. Matter* **81** (1990) 355 – 363.
111. K. H. Young, E. J. Smith, M. M. Eddy, T. W. James, *Appl. Surf. Sci.* **52** (1991) 85 – 89.
112. S. Gokhale, S. Mahamuni, K. Joshi, A. S. Nigavekar, S. K. Kulkarni, *Surf. Sci.* **257** (1991) 157 – 166.
113. S.-I. Fujimori, Y. Saito, K.-I. Yamaki, T. Okane, N. Sato, T. Komatsubara, S. Suzuki, S. Sato, *Surf. Sci.* **444** (2000) 180 – 186.
114. F. Esaka, K. Furuya, H. Shimada, M. Imamura, N. Matsubayashi, T. Kikuchi, H. Ichimura, A. Kawana, *Surf. Interface Anal.* **27** (1999) 1098 – 1106.
115. L. Ren You, N. K. Huang, H. L. Zhan, B. Yang, D. Z. Wang, *Appl. Surf. Sci.* **150** (1999) 39 – 42.
116. L. Ottaviano, L. Lozzi, A. Montefusco, S. Santucci, *Surf. Sci.* **443** (1999) 227 – 237.
117. G. Iucci, G. Polzonetti, P. Altamura, G. Paolucci, A. Goldoni, R. D'Amato, M. V. Russo, *Appl. Surf. Sci.* **153** (1999) 10 – 18.
118. L. L. Cao, Y. M. Sun, L. Q. Zheng, *Wear* **140** (1990) 345 – 357.
119. S. Noël, L. Boyer, C. Bodin, *J. Vac. Sci. Technol. A* **9** (1991) 32 – 38.
120. A. Osaka, Y.-H. Wang, Y. Miura, T. Tsugaru, *J. Mater. Sci.* **26** (1991) 2778 – 2782.
121. S. M. Mukhopadhyay, S. H. Garofalini, *J. Non Cryst. Solids* **126** (1990) 202 – 208.
122. E.-T. Kang, D. E. Day, *J. Non Cryst. Solids* **126** (1990) 141 – 150.
123. Z. Hussain, E. E. Khawaja, *Phys. Scr.* **41** (1990) 939 – 943.
124. R. K. Brow, Y. B. Peng, D. E. Day, *J. Non Cryst. Solids* **126** (1990) 231 – 238.
125. B. A. van Hassel, A. J. Burggraaf, *Appl. Phys. A* **52** (1991) 410 – 417.
126. P. Prieto, L. Galán, J. M. Sanz, *Surf. Sci.* **251/252** (1991) 701 – 705.
127. K. Oyoshi, T. Tagami, S. Tanaka, *J. Appl. Phys.* **68** (1990) 3653 – 3660.

128. Z.-H. Lu, A. Yelon, *Phys. Rev. B Condens. Matter* **41** (1990) 3284 – 3286.
129. A. Carnera, P. Mazzoldi, A. Boscolo-Boscoletto, F. Caccavale, R. Bertoncello, G. Granozzi, I. Spagnol, G. Battaglin, *J. Non Cryst. Solids* **125** (1990) 293 – 301.
130. R. W. Bernstein, J. K. Grepstad, *J. Appl. Phys.* **68** (1990) 4811 – 4815.
131. A. Ermolieff, F. Martin, A. Amouroux, S. Marthon, J. F. M. Westendorp, *Semicond. Sci. Tech.* **6** (1991) 98 – 102.
132. S. V. Hattangady, R. A. Rudder, M. J. Mantini, G. G. Fountain, J. B. Posthill, R. J. Markunas, *J. Appl. Phys.* **68** (1990) 1233 – 1236.
133. Maria Faur, Mircea Faur, D. T. Jayne, M. Goradia, C. Goradia, *Surf. Interface Anal.* **15** (1990) 641 – 650.
134. H. Hasegawa, M. Akazawa, H. Ishii, A. Uraie, H. Iwadate, E. Ohue, *J. Vac. Sci. Technol. B* **8** (1990) 867 – 873.
135. L. A. Harris, *J. Appl. Phys.* **39** (1968) 1419 – 1427, 1428 – 1431
136. R. E. Weber, W. T. Peria, *J. Appl. Phys.* **38** (1967) 4355 – 4358.
137. P. W. Palmberg, G. K. Bohm, J. C. Tracy, *Appl. Phys. Lett.* **15** (1969) 254 – 255.
138. E. Casnati, A. Tartari, C. Baraldi, *J. Phys. B* **15** (1982) 155 – 167.
139. L. E. Davis, N. C. MacDonald, P. W. Palmberg, G. E. Riach, R. E. Weber: *Handbook of Auger Electron Spectroscopy,* Physical Electronic Ind., Minnesota 1976.
140. E. Kny, *J. Vac. Sci. Technol.* **17** (1980) 658 – 660.
141. M. A. Smith, L. L. Levinson, *Phys. Rev. B Solid State* **16** (1977) 1365 – 1369.
142. M. A. Smith, L. L. Levinson, *Phys. Rev. B Solid State* **16** (1977) 1365 – 1369.
143. M. Gryzinski, *Phys. Rev.* **138** (1965) A 305 – 321.
144. S. Ichimura, R. Shimizu. *Surf. Sci.* **112** (1981) 386 – 408.
145. R. Payling, *J. Electron Spectros. Relat. Phenom.* **36** (1985) 99 – 104.
146. S. Mroczkowski, D. Lichtman, *J. Vac. Sci. Technol. A* **3** (1985) 1860 – 1865.
147. T. Sato, Y. Nagasawa, T. Sekine, Y. Sakai, *Surf. Interface Anal.* **14** (1989) 787 – 793.
148. M. P. Seah, *Surf. Interface Anal.* **9** (1986) 85 – 98.
149. H. Goretzki, *Fresenius' Z. Anal. Chem.* **329** (1987) 180 – 189.
150. R. K. Wild, J. Hickey in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 643 – 646.
151. D. Saidi, N. Souami, M. Oulladj, M. Negache, S. Lebaili in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 683 – 686.
152. I. A. Vatter, J. M. Titchmarsh, *Surf. Interface Anal.* **25** (1997) 760 – 776.
153. M. Norell, L. Nyborg in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 687 – 690.
154. B. M. Reichl, M. M. Eisl, T. Weis, H. Hutter, H. Stoeri, *Fresenius' Z. Anal. Chem.* **353** (1995) 762 – 765.
155. T. H. Chuang, *Mat. Sci. Ens.* **A141** (1991) 169 – 178.
156. D. J. Nettleship, R. K. Wild, *Surf. Interface Anal.* **16** (1990) 552 – 558.
157. A. Bukaluk, *Appl. Surf. Sci.* **144/145** (1999) 395 – 398.
158. E. M. Moser, M. Metzger, L. J. Gauckler, *Fresenius' Z. Anal. Chem.* **353** (1995) 684 – 689.
159. A. Zalar, B. Baretzky, F. Dettenwanger, M. Rühle, P. Panjan, *Surf. Interface Anal.* **26** (1998) 861 – 867.
160. H. Li, S. Jin, H. Bender, F. Lanckmans, I. Heyvaert, K. Maex, L. Froyen, *J. Vac. Sci. Technol. B* **18** (2000) 242 – 251.
161. R. H. Roberts, M. Prutton, D. K. Wilkinson, I. R. Barkshire, C. J. Hill, P. J. Pearson, P. D. Augustus, D. K. Skinner, K. Stribley, *Surf. Interface Anal.* **26** (1998) 461 – 470.
162. D. K. Fillmore, H. A. Krasinski, *Surf. Interface Anal.* **26** (1998) 109 – 112.
163. C. Palacio, *Surf. Interface Anal.* **27** (1999) 1092 – 1097.
164. L. Cao, F. Shi, W. Song, Y. Zhu, *Surf. Interface Anal.* **28** (1999) 258 – 263.
165. Y. Zhu, W. Yao, B. Zheng, L. Cao, *Surf. Interface Anal.* **28** (1999) 254 – 257.
166. A. D. Polli, T. Wagner, T. Gemming, M. Rühle, *Surf. Sci.* **448** (2000) 279 – 289.
167. J. Marien, T. Wagner, G. Duscher, A. Koch, M. Rühle, *Surf. Sci.* **446** (2000) 219 – 228.
168. R. Wagner, D. Schlatterbeck, K. Christmann, *Surf. Sci.* **440** (1999) 231 – 251.
169. T. W. Kang, T. W. Kim, *Appl. Surf. Sci.* **150** (1991) 190 – 194.
170. S. Bredendiek, H. Jenett, *Mikrochim. Acta* **107** (1992) 219 – 225.
171. S. Bredendiek, H. Jenett, *Fresenius' J. Anal. Chem.* **346** (1993) 315 – 317.
172. C. Uebing, *Surf. Sci.* **225** (1990) 97 – 106.

173. C. Uebing in I. Olefjord, L. Nyborg, D. Briggs (eds.): *ECASIA 97,* Wiley, Chichester 1997, pp. 58 – 67.
174. Z. Song, D. Tan, F. He, X. Bao, *Appl. Surf. Sci.* **137** (1999) 142 – 149.
175. H. K. Lee, R. W. Hyland Jr., H. I. Aaronson, P. P. Wynblatt, *Surf. Sci.* **408** (1998) 288 – 299.
176. G. Contini, V. Di Castro, N. Motta, A. Sgarlata, *Surf. Sci.* **405** (1998) L 509 – L 513.
177. D. T. L. van Agterveld, G. Palasantzas, J. T. M. De Hosson, *Appl. Surf. Sci.* **152** (1999) 250 – 258.
178. E. Clauberg, C. Uebing, H. J. Grabke, *Appl. Surf. Sci.* **143** (1999) 206 – 214.
179. R. Hamminger, G. Grathwohl, F. Thümmler, *J. Mater. Sci.* **18** (1983) 353 – 364.
180. A. Benninghoven, *Z. Phys.* **220** (1969) 159 – 180.
181. P. Sigmund in: N. H. Tolk, J. C. Tully, W. Heiland, C. W. White (eds.): *Inelastic Ion-Surface Collisions,* Academic Press, New York 1977.
182. J. C. Rivière: *Surface Analytical Techniques,* Oxford University Press, Oxford 1990.
183. D. Stapel, O. Brox, A. Benninghoven, *Appl. Surf. Sci.* **140** (1999) 156 – 67.
184. K. Kötter, A. Benninghoven, *Appl. Surf. Sci.* **133** (1998) 47 – 57.
185. A. Benninghoven, F. G. Rüdennauer, H. W. Werner: *Secondary Ion Mass Spectrometry,* Wiley, New York 1987.
186. P. D. Prewett, D. K. Jefferies, *J. Phys. D* **13** (1980) 1747 – 1755.
187. A. R. Krauss, D. M. Gruen, *Appl. Phys.* **14** (1977) 89 – 97.
188. P. Steffens, E. Niehuis, T. Friese, D. Greifendorf, A. Benninghoven, *J. Vac. Sci. Technol. A* **3** (1985) 1322 – 1325.
189. E. Niehuis, T. Heller, H. Feld, A. Benninghoven, *J. Vac. Sci. Technol. A* **5** (1987) 1243.
190. V. I. Karataev, B. A. Mamyrin, D. V. Shmikk, *Sov. Phys. Techn. Phys.* **16** (1972) 1177.
191. B. W. Schueler, *Microsc. Microanal. Microstruct.* **3** (1992) 119.
192. T. Sakurai, T. Matsuo, H. Matsuda, *Int. J. Mass. Spectrom. Ion Phys.* **63** (1985) 273.
193. K. Iltgen, C. Bendel, E. Niehuis, A. Benninghoven, *J. Vac. Sci. Technol. A* **15** (1997) 460.
194. I. V. Bletsos, D. M. Hercules, A. Benninghoven, D. Greifendorf in A. Benninghoven, R. J. Colton, D. S. Simons, H. W. Werner (eds.): *SIMS V,* Springer, Berlin 1986.
195. P. H. McBreen, S. Moore, A. Adnot, D. Roy, *Surf. Sci.* **194** (1988) L112 – L118.
196. D. van Leyen, B. Hagenhoff, E. Niehuis, A. Benninghoven, I. V. Bletsos, D. M. Hercules, *J. Vac. Sci. Technol. A* **7** (1989) 1790.
197. J. A. Treverton, J. Ball, D. Johnson, J. C. Vickerman, R. H. West, *Surf. Interface Anal.* **15** (1990) 369 – 376.
198. B. Hagenhoff, D. Rading in J. C. Rivière, S. Myhra (eds): *Handbook of Surface and Interface Analysis,* Marcel Dekker, New York 1998, pp. 209 – 253.
199. M. Morra, E. Occhiello, F. Garbassi, *Surf. Interface Anal.* **16** (1990) 412 – 417.
200. D. E. Fowler, R. D. Johnson, D. van Leyen, A. Benninghoven, *Surf. Interface Anal.* **17** (1991) 125 – 136.
201. G. J. Leggett, D. Briggs, J. C. Vickermnan, *Surf. Interface Anal.* **17** (1991) 737 – 744.
202. W. D. Ramsden, *Surf. Interface Anal.* **17** (1991) 793 – 802.
203. D. J. Pawson, A. P. Ameen, R. D. Short, *Surf. Interface Anal.* **18** (1992) 13 – 22.
204. H. F. Arlinghaus, M. N. Kwoka, K. B. Jacobson, *Anal. Chem.* **69** (1997) 3747 – 3753.
205. H. F. Arlinghaus, C. Höppener, J. Drexler in: *Proc. SIMS XII*, Elsevier Science, Amsterdam 2000, p. 951.
206. C. Höppener, J. Drexler, M. Ostrop, H. F. Arlinghaus in: *Proc. SIMS XII*, Elsevier Science, Amsterdam 2000, p. 915.
207. R. A. Cocco, B. J. Tatarchuk, *Surf. Sci.* **218** (1989) 127 – 146.
208. B. H. Sakakini, I. A. Ransley, C. F. Oduoza, J. C. Vickerman, M. A. Chesters, *Surf. Sci.* **271** (1992) 227 – 236.
209. X.-L. Zhou, C. R. Flores, J. M. White, *Surf. Sci. Lett.* **268** (1992) L267 – L273.
210. M. E. Castro, J. M. White, *Surf. Sci.* **257** (1991) 22 – 32.
211. D. Briggs, M. J. Hearn, *Surf. Interface Anal.* **13** (1988) 181 – 185.
212. A. Benninghoven, *Angew. Chem. Int. Ed. Engl.* **33** (1994) 1023 – 1043.
213. A. Benninghoven, B. Hagenhoff, E. Niehuis, *Anal. Chem.* **65** (1993) 630A-640A.
214. N. Bourdos, F. Kolmer, A. Benninghoven, M. Sieber, H.-J. Galla, *Langmuir* **16** (2000) 1481 – 1484.
215. M. Deimel, H. Rulle, V. Liebing, A. Benninghoven, *Appl. Surf. Sci.* **134** (1998) 271 – 274.

216. D. Briggs, M. J. Hearn, I. W. Fletcher, A. R. Waugh, B. J. McIntosh, *Surf. Interface Anal.* **15** (1990) 62 – 65.
217. M. J. Hearn, D. Briggs, *Surf. Interface Anal.* **17** (1991) 421 – 429.
218. A. Licciardello, C. Puglisi, D. Lipinsky, E. Niehuis, A. Benninghoven in A. Benninghoven, B. Hagenhoff, H. W. Werner (eds.): *SIMS X,* Wiley, New York 1997.
219. F. Kollmer, R. Kamischke, R. Ostendorf, A. Schnieders, C. Y. Kim, J. W. Lee, A. Benninghoven in Proceedings SIMS XII, Elsevier Science, Amsterdam 2000, p. 329.
220. K. Iltgen, A. Benninghoven, E. Niehuis in G. Gillen, R. Lareau, J. Bennett, F. Stevie (eds.): *SIMS XI,* Wiley, New York, 1998, pp. 367 – 370.
221. R. F. K. Herzog, F. P. Viehböck, *Phys. Rev.* **76** (1949) 855L.
222. R. F. K. Herzog, H. Liebl, *J. Phys.* **34** (1963) 2893.
223. R. Castaing, G. Slodzian, *J. de Microscopie* **1** (1962) 395.
224. A. Benninghoven, *Z. Phys.* **239** (1970) 403.
225. J. M. Schroeer, T. Rodin, R. Bradley, *Surf. Sci.* **34** (1973) 511.
226. Z. Sroubek, *Phys. Rev. B* **25** (1983) 604.
227. C. A. Anderson, J. R. Hinthorne, *Anal. Chem.* **45** (1973) 1421.
228. J. Mattauch, R. Herzog, *Z. Phys.* **89** (1934) 786.
229. H. Hutter, M. Grasserbauer, *Mikrochim. Acta* **107** (1992) 137.
230. J. M. Schroeer, H. Gnaser, H. Oechsner in *Proceedings SIMS IX,* Wiley, New York 1994, p. 394.
231. Ch. Bruner, H. Hutter, K. Piplits, M. Gritsch, G. Pöckl, M. Grasserbauer, *Fresenius' J. Anal. Chem.* **361** (1998) 667 – 671.
232. T. Stubbings, M. Wolkenstein, H. Hutter, *J. Trace and Microprobe Techniques* **17** (1999) 1 – 16.
233. M. Gritsch, K. Piplits, R. Barbist, P. Wilhartitz, *Microchim. Acta* **133** (2000) 89 – 93.
234. M. Grasserbauer, H. Hutter in *Proceedings SIMS IX,* Wiley, New York 1994, p. 545.
235. F. G. Ruedenauer, F. G. Steiger in *Proceedings SIMS VI,* Wiley, New York 1987, p. 361.
236. C. Jarms, H.-R. Stock, P. Mayr, *Surf. Coat. Technol.* **108 – 109** (1998) 206 – 210.
237. C. Pollak, B. Kriszt, H. Hutter, *Mikrochim. Acta* **133** (2000) 261 – 266.
238. Th. Albers, M. Neumann, D. Lipinsky, L. Wiedmann, A. Benninghoven, *Surf. Interface Anal.* **22** (1994) 9.
239. D. Lipinsky, R. Jede, O. Ganschow, A. Benninghoven, *J. Vac. Sci. Technol. A* **3** (1985) 2007 – 2017.
240. A. R. Bayly, J. Wolstenholme, C. R. Petts, *Surf. Interface Anal.* **21** (1993) 414 – 417.
241. W. Lotz, *Z. Phys.* **232** (1970) 101.
242. R. Jede, O. Ganschow, U. Kaiser in D. Briggs, M. P. Seah (eds.): *Practical Surface Analysis,* 2nd ed., **vol. 2,** Wiley, Chichester 1992.
243. H. Oechsner, L. Reichert, *Phys. Lett.* **23** (1966) 90 – 92.
244. H. Oechsner, *Plasma Phys.* **835** (1974) 16.
245. E. Stumpe, H. Oechsner, H. Schoof, *Appl. Phys.* **20** (1979) 55.
246. H. Oechsner, M. Müller, *J. Vac. Sci. Technol. A* **17** (1999) 3401 – 3405.
247. H. Gnaser, H.-J. Schneider, H. Oechsner, *Nucl. Instrum. Methods Phys. Res. B* **136 – 138** (1998) 1023 – 1027.
248. A. Wucher, *J. Vac. Sci. Technol. A* **6** (1988) 2287.
249. H. Jenett in H. Günzler, A. M. Bahadir, R. Borsdorf, K. Danzer, W. Fresenius, R. Galensa, W. Huber, M. Linscheid, I. Lüderwald, G. Schwedt, G. Tölg, H. Wisser (eds.): *Analytiker-Taschenbuch* **vol. 16,** Springer, Berlin 1997, pp. 43 – 117.
250. M. Fichtner, J. Goschnick, U. C. Schmidt, A. Schweiker, H. J. Ache, *J. Vac. Sci. Technol. A* **10** (1992) 362.
251. R. Jede, H. Peters, G. Dünnebier, O. Ganschow, U. Kaiser, K. Seifert, *J. Vac. Sci. Technol. A* **6** (1988) 2271.
252. J. Bartella, R. Fuchs, J. Goschnick, D. Grunenberg, *Fresenius' J. Anal. Chem.* **346** (1993) 131.
253. A. Wucher, F. Novak, W. Reuter, *J. Vac. Sci. Technol. A* **6** (1988) 2265.
254. H. Oechsner in M. Grasserbauer, H. W. Werner (eds.): *Analysis of Microelectronic Materials and Devices,* Wiley, Chichester 1991.
255. H. Jenett, J. D. Sunderkötter, M. F. Stroosnijder, *Fresenius' J. Anal. Chem.* **358** (1997) 225 – 229.
256. J. Goschnick, C. Natzeck, M. Sommer, *Appl. Surf. Sci.* **144 – 145** (1999) 201 – 207.
257. S. Hopfe, N. Kallis, H. Mai, W. Pompe, R. Scholz, S. Völlmar, B. Wehner, P. Weißbrot, *Fres. J. Anal. Chem.* **346** (1993) 14 – 22.
258. R. W. Martin, Teak Kim, D. Burns, I. M. Watson, M. D. Dawson, T. F. Krauss, J. H. Marsh, R. M. De la Rue, S. Romani, H. Kheyrandish, *Phys. Status Solidi A* **176** (1999) 67 – 71.

259. S. Dreer, *Fresenius' J. Anal. Chem.* **365** (1999) 85 – 95.
260. I. Galesic, B. O. Kolbesen, *Fresenius' J. Anal. Chem.* **365** (1999) 199 – 202.
261. T. Tharigen, G. Lippold, V. Riede, M. Lorenz, K. J. Koivusaari, D. Lorenz, S. Mosch, P. Grau, R. Hesse, P. Streubel, R. Szargan, *Thin Solid Films* **348** (1999) 103 – 113.
262. M. Witthaut, R. Cremer, A. von Richthofen, D. Neuschütz, *Fresenius' J. Anal. Chem.* **361** (1998) 639 – 641.
263. H. Boerner, H. Jenett, V.-D. Hodoroaba, *Fresenius' J. Anal. Chem.* **361** (1998) 590 – 591.
264. B. Hueber, G. H. Frischat, A. Maldener, O. Dersch, F. Rauch, *J. Non-Cryst. Solids* **256 & 257** (1999) 130 – 134.
265. R. Getto, J. Freytag, M. Kopnarski, H. Oechsner, *Mater. Sci. Eng. B* **61 – 62** (1999) 270 – 274.
266. S. Kruijer, O. Nikolov, W. Keune, H. Reuther, S. Weber, S. Scherrer, *J. Appl. Phys.* **84** (1998) 6570 – 6581.
267. G. A. Seryogin, S. A. Nikishin, H. Temkin, R. Schlaf, L. I. Sharp, Y. C. Wen, B. Parkinson, V. A. Elyukhin, Yu. A. Kudriavtsev, A. M. Mintairov, N. N. Faleev, M. V. Baidakova, *J. Vac. Sci. Technol. B* **16** (1998) 1456 – 1458.
268. M. C. Simmonds, A. Savan, H. van Swygenhoven, E. Pfluger, *Thin Solid Films* **354** (1999) 59 – 65.
269. A. Vogt, A. Simon, H. L. Hartnagel, J. Schikora, V. Buschmann, M. Rodewald, H. Fuess, S. Fascko, C. Koerdt, H. Kurz, *J. Appl. Phys.* **83** (1998) 7715 – 7719.
270. J. Goschnick, U. Maeder, M. Sommer, *Fresenius' J. Anal. Chem.* **361** (1998) 707 – 709.
271. J. A. Rebane, O. Yu. Gorbenko, S. G. Suslov, N. V. Yakovlev, I. E. Korsakov, V. A. Amelichev, Yu. D. Tretyakov, *Thin Solid Films* **302** (1997) 140 – 146.
272. C. Degueldre, D. Buckley, J. C. Dran, E. Schenker, *J. Nucl. Mater.* **252** (1998) 22 – 27.
273. D. Sommer, A. Essing, *Rep. Eur. Comm.* **EUR 18864** (1999) 1 – 126.
274. T. A. Dang, T. A. Frisk, M. W. Grossman, C. H. Peters, *J. Electrochem. Soc.* **146** (1999) 3896 – 3902.
275. I. Baumann, F. Cusso, B. Herreros, H. Holzbrecher, H. Paulus, K. Schaefer, W. Sohler, *Appl. Phys. A* **68** (1999) 321 – 324.
276. M. Miloshova, E. Bychkov, V. Tsegelnik, V. Strykanov, H. Klewe-Nebenius, M. Bruns, W. Hoffmann, P. Papet, J. Sarradin, A. Pradel, M. Ribes, *Sens. Actuators* **B7** (1999) 171 – 178.
277. R. B. Vasiliev, M. N. Rumyantseva, S. E. Podguzova, A. S. Ryzhikov, L. I. Ryabova, A. M. Gaskov, *Mater. Sci. Eng. B* **57** (1999) 241 – 246.
278. J. Goschnick, C. Natzeck, M. Sommer, F. Zudock, *Thin Solid Films* **332** (1998) 215 – 219.
279. J. Goschnick, M. Sommer, *Fresenius' J. Anal. Chem.* **361** (1998) 704 – 707.
280. H. Nickel, D. Clemens, W. J. Quadakkers, L. Singheiser, *PVP (Am. Soc. Mech. Eng.)* **374** (1998) 357 – 361.
281. C. H. Becker, K. T. Gillen, *Anal. Chem.* **56** (1984) 1671.
282. C. H. Becker, *J. Vac. Sci. Technol. A* **5** (1987) 1181.
283. J. B. Pallix, C. H. Becker, N. Newman, *J. Vac. Sci. Technol. A* **6** (1988) 1049.
284. M. Terhorst, R. Möllers, A. Schnieders, E. Niehuis, A. Benninghoven in A. Benninghoven, Y. Nihei, R. Shimizu, H. W. Werner (eds.): *SIMS IX,* Wiley, Chichester 1994, pp. 434 – 437.
285. A. Schnieders, R. Möllers, M. Terhorst, H.-G. Cramer, E. Niehuis, A. Benninghoven, *J. Vac. Sci. Technol. B* **14** (1996) 2712.
286. N. Thonnard, J. E. Parks, R. D. Willis, L. J. Moore, H. F. Arlinghaus, *Surf. Interface Anal.* **14** (1989) 751.
287. H. F. Arlinghaus, M. T. Spaar, N. Thonnard, *J. Vac. Sci. Technol. A* **8** (1990) 2318.
288. H. F. Arlinghaus, M. T. Spaar, N. Thonnard, A. W. McMahon, T. Tanigaki, H. Shichi, P. H. Holloway, *J. Vac. Sci. Technol. A* **11** (1993) 2317.
289. D. L. Pappas, D. M. Hrubowchak, M. H. Ervin, N. Winograd, *Science* **243** (1989) 64.
290. M. J. Pellin, C. E. Young, W. F. Calaway, J. E. Whitten, D. M. Gruen, J. B. Blum, I. D. Hutcheon, G. J. Wasserburg, *Phil. Trans. R. Soc. Lond. A* **333** (1990) 133.
291. S. W. Downey, A. B. Emerson, R. F. Kopf, *Inst. Phys. Conf. Ser.* **114** (Resonance Ionisation Spectroscopy) (1991) 401.
292. N. Thonnard, R. D. Willis, M. C. Wright, W. A. Davis, B. E. Lehman, *Nucl. Instr. and Meth. Phys. Res. B* **29** (1987) 398 – 406.
293. J. Zehnpfenning, E. Niehuis, H.-G. Cramer, T. Heller, R. Möllers, M. Terhorst, A. Schnieders, Zhiyuan Zhang, A. Benninghoven in A. Benninghoven, Y. Nihei, R. Shimizu, H. W. Werner (eds.): *SIMS IX,* Wiley, Chichester 1994, pp. 561 – 564.

294. H. F. Arlinghaus, C. F. Joyner, *J. Vac. Sci. Technol. B* **14** (1996) 294 – 300.
295. H. F. Arlinghaus, C. F. Joyner, J. Tower, S. Sen in A. Benninghoven, B. Hagenhoff, H. W. Werner (eds): *SIMS X,* Wiley, New York 1997, pp. 463 – 466.
296. H. F. Arlinghaus, M. T. Spaar, R. C. Switzer, G. W. Kablaka, *Anal. Chem.* **69** (1997) 3169 – 3176.
297. H. F. Arlinghaus, M. P. Kwoka, X. Q. Guo, K. B. Jacobson, *Anal. Chem.* **69** (1997) 1510 – 1517.
298. H. F. Arlinghaus, M. P. Kwoka, K. B. Jacobson, *Anal. Chem.* **69** (1997) 3747 – 3753.
299. L. C. Feldman, J. W. Mayer: *Fundamentals of Surface and Thin Film Analysis,* North Holland, New York 1986.
300. J. A. Leavitt, L. C. McIntyre, M. R. Weller in J. R. Tesmer, M. Nastasi (eds.): *Handbook of Modern Ion Beam Materials Analysis,* Materials Research Society, Pittsburgh 1995, pp. 37 – 81.
301. M. L. Swanson in J. R. Tesmer, M. Nastasi (eds.): *Handbook of Modern Ion Beam Materials Analysis,* Materials Research Society, Pittsburgh 1995, pp. 231 – 300.
302. S. Humphries: *Principles of Charged Particle Acceleration,* Wiley, New York 1986.
303. E. Steinbauer, P. Bauer, M. Geretschläger, G. Bortels, J. P. Biersack, P. Burger, *Nucl. Instr. Meth. B* **85** (1994) 642 – 649.
304. J. Vrijmoeth, P. M. Zagwin, J. W. M. Frenken, J. F. van der Veen, *Phys. Rev. Lett.* **67** (1991) 1134 – 1137.
305. J. C. Banks, B. L. Doyle, J. A. Knapp, D. Werho, R. B. Gregory, M. Anthony, T. Q. Hurd, A. C. Diebold, *Nucl. Instr. Meth. B* **136 – 138** (1998) 1223 – 1228.
306. R. S. Bhattacharya, P. P. Pronko, *Appl. Phys. Lett.* **40** (1982) 890 – 892.
307. L. R. Doolittle, *Nucl. Instr. Meth. B* **9** (1985) 344 – 351.
308. L. R. Doolittle, *Nucl. Instr. Meth. B* **15** (1986) 227 – 231.
309. J. A. Davies, W. N. Lennard, I. V. Mitchell in J. R. Tesmer, M. Nastasi (eds.): *Handbook of Modern Ion Beam Materials Analysis,* Materials Research Society, Pittsburgh 1985, pp. 343 – 363.
310. S. Mantl, *Nucl. Instr. Meth. B* **80/81** (1993) 895 – 900.
311. N. R. Parikh, G. S. Sandhu, N. Yu, W. K. Chu, T. E. Jackman, J. M. Baribeau, D. C. Houghton, *Thin Solid Films* **163** (1988) 455 – 460.
312. K. Nakajima, A. Konishi, K. Kimura, *Phys. Rev. Lett.* **83** (1999) 1802 – 1805.
313. J. W. M. Frenken, J. F. van der Veen, *Phys. Rev. Lett.* **54** (1985) 134 – 137.
314. E. Taglauer in J. C. Vickerman (ed.): *Surface Analysis – The Principle Techniques,* Wiley, New York 1997, p. 215.
315. H. Niehus, W. Heiland, E. Taglauer, *Surf. Sci. Rep.* **17** (1993) 213 – 303.
316. E. S. Parilis, L. M. Kishinevsky, N. Yu. Turnaev, B. E. Baklitzky, F. F. Umarov in S. I. Nizhnaya, I. S. Bitensky (eds.): *Atomic Collisions on Atomic Surfaces,* North-Holland, Amsterdam 1993.
317. E. C. Goldberg, R. Monreal, F. Flores, H. H. Brongersma, P. Bauer, *Surface Science* **440** (1999) L875 –L880.
318. S. Tsuneyuki, M. Tsukada, *Phys. Rev. B* **34** (1986) 5758 – 5768.
319. L. C. A. van den Oetelaar, S. N. Mikhailov, H. H. Brongersma, *Nucl. Instr. Meth. B* **93** (1994) 210 – 214.
320. H. H. Brongersma, P. A. C. Groenen, J.-P. Jacobs, *Mater. Sci. Monogr.* **81** (1994) 113 – 182.
321. A. W. Denier van der Gon, R. Cortenraad, W. P. A. Jansen, M. A. Reijme, H. H. Brongersma, *Nucl. Instr. Meth. B* **161 – 163** (2000) 56 – 64.
322. G. C. van Leerdam, PhD Thesis, University of Technology, Eindhoven, The Netherlands 1991.
323. J.-P. Jacobs, PhD Thesis, University of Technology, Eindhoven, The Netherlands 1995.
324. L. C. A. van den Oetelaar, J.-P. Jacobs, M. J. Mietus, H. H. Brongersma, V. N. Semenov, V. G. Glebovsky, *Appl. Surf. Sci.* **70/71** (1993) 79 – 84.
325. H. C. Yao, M. Shelef, *J. Phys. Chem.* **78** (1974) 2490.
326. H. H. Brongersma, J.-P. Jacobs, *Appl. Surf. Sci.* **75** (1994) 133 – 138.
327. M. M. Viitanen, W. P. A. Jansen, R. van Welzenis, H. H. Brongersma, *J. Phys. Chem. B* **103** (1999) 6025 – 6029.
328. H. Niehus, R. Spitzl, *Surf. Interface Anal.* **17** (1991) 287 – 307.
329. J. W. Rabalais, *Surf. Interface Anal.* **27** (1999) 171 – 178.
330. A. Krauss, O. Auciello, J. A. Schultz, *MRS Bull.* **20** (1995) 18 – 23.
331. M. I. Pupin, *Science* **3** (1896) 538 – 544.
332. H. G. J. Moseley, *Philos. Mag.* **26** (1913) 1024.

333. H. G. J. Moseley, *Philos. Mag.* **27** (1914) 703.
334. Y. Yoneda, T. Horiuchi *Rev. Sci. Instrum.* **42** (1971) 1069.
335. P. Wobrauschek, P. Kregsamer, C. Streli, H. Aiginger, *X-Ray Spectrom.* **20** (1991) 23 – 8.
336. P. Kregsamer, *Spectrochim. Acta B* **46** (1991) 1333 – 40.
337. H. Aiginger, *Spectrochim. Acta B* **46** (1991) 1313 – 21.
338. H. Aiginger, P. Wobrauschek, *Nucl. Instrum. Methods* **114** (1974) 157.
339. J. Knoth, H. Schwenke, *Fresenius' Z. Anal. Chem.* **291** (1978) 200.
340. H. Schwenke, J. Knoth, *Nucl. Instrum. Methods* **193** (1982) 239.
341. Semiconductor Equipment and Materials International, SEMI Standards E 45.
342. Semiconductor Equipment and Materials International, SEMI Standards M 33.
343. R. S. Becker, J. A. Golovchenko, J. R. Patel, *Phys. Rev. Lett.* **50** (1983) 153.
344. P. Wobrauschek, R. Görgl, P. Kregsamer, C. Streli, S. Pahlke, L. Fabry, M. Haller, A. Knöchel, M. Radtke, *Spectrochim. Acta Part B* **52** (1997) 901 – 6.
345. L. Fabry, S. Pahlke, L. Kotz, P. Wobrauschek, Ch. Streli, P. Eichinger, *Fresenius' J. Anal. Chem.* **363** (1999) 98 – 102.
346. R. Klockenkämper, A. von Bohlen, *Spectrochim. Acta B* **54** (1999) 1385 – 92.
347. Ch. Streli, P. Kregsamer, P. Wobrauschek, H. Gatterbauer, P. Pianetta, S. Pahlke, L. Fabry, L. Palmetshofer, M. Schmeling, *Spectrochim. Acta B* **54** (1999) 1433 – 41.
348. X. Xiao, S. Hayakawa, Y. Goshi, M. Oshima, *Anal. Sci. (Jpn)* **14** (1998) 1139 – 44.
349. P. K. de Bokx, S. J. Kidd, G. Wiener, H. P. Urbach, S. de Gendt, P. W. Mertens, M. M. Heyns, *Proc. - Electrochem. Soc.* (1998), PV 98-1 (Silicon Materials Science and Technology, **Vol. 2**) pp. 1511 – 1523.
350. H. Schwenke, GKSS, Geesthacht, personal communication, 1999; *Proc. GMM Workshop*, Erlangen Feb. 23-24, 2000, Gesellschaft für Mikroelektronik VDE/VDI.
351. P. Wobrauschek, C. Streli: *Advanced Light Sources, ADLIS,* TU Wien, Vienna 1999, pp. 284 – 300.
352. R. Klockenkämper: *Total Reflection X-ray Analysis,* Wiley Interscience, New York 1997.
353. P. J. Potts, A. T. Ellis, P. Kregsamer, C. Streli, M. West, P. Wobrauschek, *J. Anal. At. Spectrom.* **14** (1999) 1773 – 99.
354. P. Wobrauschek, C. Streli in *Encyclopedia of Analytical Chemistry*, in press.
355. K. N. Stoev, K. Sakurai, *Spectrochim. Acta B* **54** (1999) 41 – 82.
356. K. Yakushiji, S. Ohkawa, A. Yoshinaga, J. Harada, *Jpn. J. Appl. Phys.* **32** (1993) 1191 – 6.
357. K. Yakushiji, S. Ohkawa, A. Yoshinaga, J. Harada, *Jpn. J. Appl. Phys.* **33** (1994) 1130 – 5.
358. B. W. Liou, C. L. Lee, *Chin. J. Phys.* **37** (1999) 623 – 30.
359. K. Yakushiji, S. Ohkawa, A. Yoshinaga, J. Harada, *Jpn. J. Appl. Phys.* **32** (1993) 4750 – 1.
360. K. L. Williams: *An Introduction to X-Ray Spectrometry,* Allen & Unwin, London 1987.
361. L. Strüder, P. Lechner, P. Leutenegger, *Naturwissenschaften* **85** (1998) 539 – 543.
362. K. Yakushiji, S. Ohkawa, A. Yoshinaga, J. Harada, *Anal. Sci. (Jpn.)* **11** (1995) 505 – 11.
363. J. Knoth, H. Schwenke, P. Eichinger in M. Heyns, M. Heuris, P. Mertens (eds.): *Ultra-Clean Processing (UCPSS'94, IMEC),* ACCO, Leuven 1994, p. 107.
364. G. Y. Tao, S. J. Zhuo, A. Ji, K. Norrish, P. Fazey, U. E. Senff, *X-Ray Spectrom.* **27** (1998) 357 – 66.
365. K. Tsuji, T. Yamada, T. Utaka, K. Hirokawa, *J. Appl. Phys.* **78** (1995) 969 – 73.
366. Y. Mori, K. Shimanoe, *Anal. Sci. (Jpn)* **12** (1996) 277 – 9.
367. L. Fabry, S. Pahlke, L. Kotz, *Fresenius' J. Anal. Chem.* **354** (1996) 266 – 70.
368. P. Mertens, S. DeGendt, K. Kenis: *Calibration Accuracy of Different Atomika TXRF Instruments,* IMEC, Leuven 1996.
369. I. Rink, H. Thewissen, *Spectrochim. Acta B* **54** (1999) 1427 – 1431.
370. Y. Mori, K. Shimanoe, T. Sakon, *Anal. Sci. (Jpn)* **11** (1995) 499.
371. R. Hölzl, K. J. Range, L. Fabry, D. Huber, *J. Electrochem. Soc.* **146** (1999) 2245.
372. A. C. Diebold, *J. Vac. Sci. Technol. A* **14** (1996) 1919.
373. R. S. Hockett, *Anal. Sci. (Jpn)* **11** (1995) 511.
374. H. Schwenke, J. Knoth, U. Weisbrod, *X-Ray Spectrom.* **20** (1991) 277.
375. U. Weisbrod, R. Gutschke, J. Knoth, H. Schwenke, *Fresenius' J. Anal. Chem.* **341** (1991) 83.
376. W. Berneike, *Spectrochim. Acta B* **48** (1993) 269 – 75.
377. H. Schwenke, R. Gutschke, J. Knoth, M. Kock, *Appl. Phys. A* **54** (1992) 460

378. L. Fabry, Y. Matsushita, *Proc. Electrochem. Soc*. (1998), PV 98-1(Silicon Materials Science and Technology, **Vol. 2**) p. 1459.
379. P. Eichinger, H.-J. Rath, H. Schwenke in D. C. Gupta (ed.): *Semiconductor Fabrication: Technology and Metrology, ASTM STP 990,* ASTM, 1989, p. 305.
380. R. S. Hockett, *Adv. in X-Ray Anal.* **37** (1994) 565 – 565.
381. V. Penka, W. Hub, *Fresenius' Z. Anal. Chem.* **333** (1989) 586 – 589.
382. J. A. Sees, L. H. Hall, *J. Electrochem. Soc.* **142** (1995) 1238 – 1241.
383. L. Fabry, *Accred. Qual. Assur.* **1** (1996) 99.
384. L. Fabry, L. Köster, S. Pahlke, L. Kotz, J. Hage, *IEEE Trans. Semicond. Manuf.* **9** (1996) 428.
385. C. Streli, P. Wobrauschek, V. Bauer, P. Kregsamer, R. Görgl, P. Pianetta, R. Ryon, S. Pahlke, L. Fabry, *Spectrochim. Acta B* **52** (1997) 861 – 872.
386. P. Wobrauschek, P. Kregsamer, W. Ladisich, C. Streli, S. Pahlke, L. Fabry, S. Garbe, M. Haller, A. Knöchel, M. Radtke, *Nucl. Instr. Meth. A* **363** (1995) 619 – 620.
387. F. Comin, M. Navizet, P. Mangiagalli, G. Aposolo, *Nucl. Instr. Meth. B* **150** (1999) 538 – 542.
388. H. Binder (M + W Zander, Stuttgart) personal communication, 2000.
389. S. A. McHugo, A. C. Thompson, C. Flink, E. R. Weber, G. Lambie, B. Gunion, A. MacDowell, R. Celestre, H. A. Padmore, Z. Hussain, *J. Cryst. Growth* **210** (2000) 395 – 400.
390. H. Schwenke, W. Berneike, J. Knoth, U. Weisbrod, *Adv. X-Ray Anal.* **32** (1989) 105.
391. H. Schwenke, J. Knoth in R. van Grieken, A. Markowicz (eds.): *Handbook on X-Ray Spectrometry,* Dekker, New York 1993, p. 453.
392. A. D. Dane, H. A. van Sprang, L. M. C. Buydens, *Anal. Chem.* **71** (1999) 4580 – 4586.
393. H. Schwenke, J. Knoth, L. Fabry, S. Pahlke, R. Scholz, L. Frey, *J. Electrochem. Soc.* **144** (1997) 3979 – 83.
394. R. Klockenkämper, A. von Bohlen, H. W. Becker, L. Palmetshofer, *Surf. Interface Anal.* **27** (1999) 1003 – 1008.
395. R. Klockenkämper, A. von Bohlen, *J. Anal. At. Spectrom.* **14** (1999) 571 – 576.
396. R. Klockenkämper, A. von Bohlen, *Anal. Commun.* **36** (1999) 27 – 29.
397. T. Shiraiwa, N. Fujino, S. Sumita, Y. Tanizoe in D. C. Gupta (ed.): *Semiconductor Fabrication: Technology and Metrology, ASTM STP 990,* ASTM, 1989, p. 314.
398. A. Huber, H. J. Rath, P. Eichinger, T. Bauer, L. Kotz, R. Staudigl, *Proc. - Electrochem. Soc.* (1988), PV 88-20 (Diagnostic Techniques for Semicond. Materials and Devices) p. 109 – 112.
399. C. Neumann, P. Eichinger, *Spectrochim. Acta B* **46** (1991) 1369 – 1377.
400. G. Buhrer, *Spectrochim. Acta B* **54** (1999) 1399.
401. A. C. Diebold, P. Maillot, M. Gordon, J. Baylis, J. Chacon, R. Witowski, H. F. Arlinghaus, J. A. Knapp, B. L. Doyle, *J. Vac. Sci. Technol. A* **10** (1992) 2945 – 2952.
402. S. Metz, G. Kilian, G. Mainka, C. Angelkort, A. Fester, B. O. Kolbesen, *Proc. - Electrochem. Soc.* (1997) PV 97-22 (Crystalline Defects and Contamination: Their Impact and Control in Device Manufacturing II) p. 458 – 467.
403. L. Fabry, S. Pahlke, L. Kotz, E. Schemmel, W. Berneike, *Proc. - Electrochem. Soc.* (1993), PV 93-15 (Crystalline Defects and Contamination) pp. 232 – 239.
404. U. Reus, K. Freitag, A. Haase, J. F. Alexandre, *Spectr. 2000* **143** (1989) 42.
405. A. Prange, H. Schwenke, *Adv. X-Ray Anal.* **35B** (1992) 899.
406. L. Fabry, S. Pahlke, L. Kotz, *Adv. in X-Ray Chem. Anal. Jpn.* **27** (1995) 345.
407. H. Kondo, J. Ryuta, E. Morita, T. Yoshimi, Y. Shimanuki, *Jpn. J. Appl. Phys. Part 2 No. 1A/B* **31** (1992) L11.
408. R. Klockenkämper, A. von Bohlen, *Spectrochim. Acta B* **48** (1989) 461.
409. I. Rink, H. Thewissen, *Spectrochim. Acta B* **54** (1999) 1427 – 1431.
410. M. Matsumura, T. Sakoda, Y. Nishioka, *Jpn. J. Appl. Phys.* **37** (1998) 5963 – 5964.
411. H. Yamaguchi, S. Itoh, S. Igarashi, K. Naitoh, R. Hasegawa, *Anal. Sci. (Jpn)* **14** (1998) 909.
412. H. Yamaguchi, S. Itoh, S. Igarashi, K. Naitoh, R. Hasegawa, *Bunseki Kagaku* **48** (1999) 855L.
413. M. Yamagami, M. Nonoguchi, T. Yamada, T. Shoji, T. Utaka, S. Nomura, K. Taniguchi, H. Wakita, S. Ikeda, *X-Ray Spectrom.* **26** (1999) 451.
414. M. Yamagami, M. Nonoguchi, T. Yamada, T. Shoji, T. Utaka, S. Nomura, K. Taniguchi, H. Wakita, S. Ikeda, *Bunseki Kagaku* **48** (1999) 1005.
415. G. Vereecke, M. Schaekers, K. Verstraete, S. Arnauts, M. M. Heyns, W. Plante, *J. Electrochem. Soc.* **147** (2000) 1499.

416. B. Chapman: *Glow Discharge Processes,* Wiley, New York 1980, pp. 77 – 138.
417. W. Grimm, *Spectrochim. Acta B* **23** (1968) 443.
418. M. R. Winchester, C. Lazik, R. K. Marcus, *Spectrochim. Acta B* **46** (1991) 483.
419. A. Bengtson, A. Eklund, M. Lundholm, A. Saric, *J. Anal. At. Spectrom.* **5** (1990) 563.
420. A. Bengtson, *Spectrochim. Acta B* **49** (1994) 411.
421. A. Bengtson, S. Hänström, E. Lo Piccolo, N. Zacchetti, R. Meilland, H. Hocquaux, *Surf. Interface Anal.* **27** (1999) 743.
422. M. R. Winchester, U. Beck, *Surf. Interface Anal.* **27** (1999) 930.
423. E. Jansen, *Materials Sci. Eng.* **42** (1980) 309.
424. M. Ives, D. B. Lewis, C. Lehmberg, *Surf. Interface Anal.* **25** (1999) 191.
425. K. Shimuzu, H. Habazaki, P. Skeldon, G. E. Thompson, G. C. Wood, *Surf. Interface Anal.* **27** (1999) 1046.
426. A. Quentmeier, *J. Anal. At. Spectrom.* **9** (1994) 355.
427. K. Shimuzu, H. Habazaki, P. Skeldon, G. E. Thompson, G. C. Wood, *Surf. Interface Anal.* **27** (1999) 950.
428. R. Payling, D. Jones, A. Bengtson (eds.): *Glow Discharge Optical Emission Spectrometry,* Wiley, Chichester 1997.
429. R. Berneron, J. C. Charbonnier, *Surf. Interface Anal.* **3** (1981) 134.
430. R. Payling, D. G. Jones, *Surf. Interface Anal.* **20** (1993) 787.
431. H. Böhm, *Oberflächen Werkstoffe* **4** (1994) 8.
432. A. Bengtson, S. Hänström, E. Lo Piccolo, N. Zacchetti, R. Meilland, H. Hocquaux, *Surf. Interface Anal.* **27** (1999) 743.
433. E. Rose, P. Mayr, *Mikrochim. Acta* **I** (1989) 197.
434. H. Böhm in R. Payling, D. G. Jones, A. Bengtson (eds.): *Glow Discharge Optical Emission Spectrometry,* Wiley, Chichester 1997, p. 676.
435. K. Shimuzu, H. Habazaki, P. Skeldon, G. E. Thompson, G. C. Wood, *Surf. Interface Anal.* **29** (1999) 151.
436. R. Payling, D. G. Jones, S. A. Gower, *Surf. Interface Anal.* **20** (1993) 959.
437. B. E. Hayden in J. T. Yates, T. E. Madey (eds.): *Vibrational Spectroscopy of Molecules on Surfaces,* Plenum Press, New York, 1987, pp. 267 – 344.
438. M. D. Porter, *Anal. Chem.* **60** (1988) 1143A-1155A.
439. H. Brunner, U. Mayer, H. Hoffmann, *Appl. Spectr.* **51** (1997) 209 – 217.
440. P. Tengvall, I. Lundström, B. Lindberg, *Biomaterials* **19** (1998) 407 – 442.
441. C. R. Flach, A. Gericke, R. Mendelsohn, *J. Phys. Chem. B* **101** (1997) 58 – 65.
442. S. C. Street, A. J. Gellman, *J. Phys. Chem.* **100** (1996) 8338 – 8348.
443. C. A. Melendres, G. A. Bowmaker, J. M. Leger, B. Beden, *J. of Electroanlytical Chem.* **499** (1998) 215 – 218.
444. W. Richter, D. R. T. Zahn in G. Bauer, W. Richter (eds.): *Optical Characterization of Epitaxial Semiconductor Layers,* Springer, Berlin 1996, pp. 60 – 62.
445. G. A. Beitel, A. Laskov, H. Oosterbeek, E. W. Kuipers, *J. Phys. Chem.* **100** (1996) 12494 – 12502.
446. Y. Urai, C. Ohe, K. Itoh, M. Yoshida, Ken-ichi Iimura, T. Kato, *Langmuir* **16** (2000) 3920 – 3926.
447. H. Lavoie, J. Gallant, M. Grandbois, D. Blaudez, B. Desbat, F. Boucher, C. Salesse, *Mat. Sci. Eng. C* **10** (1999) 147 – 154.
448. T. Buffeteau, E. Le Calvez, S. Castano, B. Desbat, D. Blaudez, J. Dufourcq, *J. Phys. Chem. B* **104** (2000) 4537 – 4544.
449. P. Christensen, A. Hamnett, *Electrochim. Acta* **45** (2000) 2443 – 2459.
450. C. Sammon, J. Yarwood, N. Everall, *Polymer* **41** (1999) 2521 – 2534.
451. C. V. Raman, K. S. Krishnan, *Nature* **121** (1928) 501.
452. M. Fleischman, J. P. Hendra, A. J. McQuillan, *Chem. Phys. Lett.* **26** (1974) 163.
453. D. J. Jeanmaire, R. P. Van Duyne, *J. Electroanal. Chem.* **84** (1977) 1.
454. S. Nie, S. R. Emroy, *Science* **275** (1997) 1102.
455. Y.-C. Liu, R. L. McCreery, *J. Am. Chem. Soc.* **117** (1995) 11254.
456. W. Hill, V. Fallourd, D. Klockow, *J. Phys. Chem. B* **103** (1999) 4707.
457. W. Hill, B. Wehling, D. Klockow, *Appl. Spectrosc.* **53** (1999) 547.
458. N. Bloembergen, P. S. Pershan, *Phys. Rev.* **128** (1962) 606.
459. Q. Du, E. Freysz, Y. R. Shen, *Science* **264** (1994) 826.
460. G. L. Richmond, *Anal. Chem.* **69** (1997) 536A.
461. R. M. A. Azzam, N. M. Bashara: *Ellipsometry and Polarized Light,* North Holland, Amsterdam 1977, p. 287.
462. J. Leng, J. Opsal, H. Chu, M. Senko, D. E. Aspnes, *Thin Solid Films* **313 – 314** (1998) 132.

463. G. E. Jellison Jr., F. A. Modine, *Appl. Phys. Lett.* **69** (1996) 371.
464. B. Johs, J. A. Woollam, C. M. Herzinger, J. Hilfiker, R. Synowicki, C. L. Bungay, *Crit. Rev. Opt. Sci. Technol.* **CR72** (1999) 29 – 58.
465. A. Röseler, *Thin Solid Films* **234** (1993) 307 – 313.
466. A. Canillas, E. Pascual, B. Drevillon, *Rev. Sci. Instr.* **64** (1993) 2153 – 2159.
467. M. Born, E. Wolf: *Principles of Optics,* Pergamon Press, Oxford 1964.
468. H. Angermann, W. Henrion, A. Röseler, M. Rebien, *Mater. Sci. Eng. B* **73** (2000) 178 – 183.
469. A. R. Reinberg, *Appl. Opt.* **11** (1972) 1273.
470. G. Binnig, H. Rohrer, Ch. Gerber, E. Weibel, *Phys. Rev. Lett.* **49** (1982) 57.
471. D. A. Bonnell (ed.): *Scanning Tunneling Microscopy and spectroscopy. Theory, Techniques and Applications,* VCH Publishers, New York 1993.
472. R. Wiesendanger: *Scanning Probe Microscopy and Spectroscopy. Methods and Applications,* Cambridge University Press, Cambridge 1994.
473. G. Binnig, C. F. Quate, Ch. Gerber, *Phys. Rev. Lett.* **56** (1986) 930.
474. D. Sarid: *Scanning Force Microscopy,* Oxford University Press, New York 1991.
475. Q. Zhong, D. Inniss, K. Kjoller, V. B. Elings, *Surf. Sci. Lett.* **290** (1993) L688.
476. P. K. Hansma, J. P. Cleveland, M. Radmacher, D. A. Walters, P. E. Hillner, M. Bezanilla, M. Fritz, D. Vie, H. G. Hansma, C. B. Prater, J. Massie, L. Fukunaga, J. Gurley, V. Elings, *Appl. Phys. Lett.* **64** (1994) 1738.
477. T. R. Albrecht, S. Akamine, T. E. Carver, C. F. Quate, *J. Vac. Sci. Technol. A* **8** (1990) 3386.
478. O. Wolter, Th. Bayer, J. Greschner, *J. Vac. Sci. Technol. B* **9** (1991) 1353.
479. G. Meyer, N. M. Amer, *Appl. Phys. Lett.* **53** (1988) 2400.
480. S. Manne, P. K. Hansma, J. Massie, V. B. Elings, A. A. Gewirth, *Science* **251** (1991) 183.
481. A. A. Gewirth, *AIP Conference Proceedings* **241** (1992) 253.
482. T. Prohaska, G. Friedbacher, M. Grasserbauer, H. Nickel, R. Lösch, W. Schlapp, *Anal. Chem.* **67** (1995) 1530.
483. G. Friedbacher, T. Prohaska, M. Grasserbauer, *Mikrochim. Acta* **113** (1994) 179.
484. J. Frommer, *Angew. Chem.* **104** (1992) 1325.
485. H. Fuchs, *J. Mol. Struct.* **292** (1993) 29.
486. H. G. Hansma, *J. Vac. Sci. Technol. B* **14** (1996) 1930.
487. S. Kasas, N. H. Thomson, B. L. Smith, P. K. Hansma, J. Miklossy, H. G. Hansma, *Int. J. Imaging Systems and Technology* **8** (1997) 151.
488. S. N. Magonov, *Appl. Spectroc. Rev.* **28** (1993) 1.
489. P. E. Hillner, A. J. Gratz, S. Manne, P. K. Hansma, *Geology* **20** (1992) 359.
490. I. Schmitz, T. Prohaska, G. Friedbacher, M. Schreiner, M. Grasserbauer, *Fresenius' J. Anal. Chem.* **353** (1995) 666.
491. T. Leitner, G. Friedbacher, T. Vallant, H. Brunner, U. Mayer, H. Hoffmann, *Mikrochim. Acta* **133** (2000) 331.
492. C. M. Mate, G. M. McClelland, R. Erlandsson, S. Chiang, *Phys. Rev. Lett.* **59** (1987) 1942.
493. R. M. Overney, T. Bonner, E. Meyer, M. Rüetschi, R. Lüthi, L. Howald, J. Frommer, H.-J. Güntherodt, M. Fujihira, H. Takano, *J. Vac. Sci. Technol. B* **12** (1994) 1973.
494. C. D. Frisbie, L. F. Rozsnyai, A. Noy, M. S. Wrighton, C. S. Lieber, *Science* **265** (1994) 2071.
495. P. Maivald, H.-J. Butt, S. A. C. Gould, C. B. Prater, B. Drake, J. A. Gurley, V. B. Elings, P. K. Hansma, *Nanotechnology* **2** (1991) 103.
496. H.-J. Butt, *Biophys. J.* **60** (1991) 777.
497. A. L. Weisenhorn, P. Maivald, H.-J. Butt, P. K. Hansma, *Phys. Rev. B* **45** (1992) 11226.
498. J. H. Hoh, J. P. Cleveland, C. B. Prater, J.-P. Revel, P. K. Hansma, *J. Am. Chem. Soc.* **114** (1992) 4917.
499. J. E. Stern, B. D. Terris, H. J. Manin, D. Rugar, *Appl. Phys. Lett.* **53** (1988) 2717.
500. B. D. Terris, J. E. Stern, D. Rugar, H. J. Manin, *J. Vac. Sci. Technol. A* **8** (1990) 374.
501. Y. Martin, H. K. Wickramasinghe, *Appl. Phys. Lett.* **50** (1987) 1455.
502. G. Persch, H. Strecker, *Ultramicroscopy* **42 – 44** (1992) 1269.
503. G. F. A. Van der Walle, E. J. Van Loenen in M. Grasserbauer, H. W. Werner (eds.): *Analysis of Microelectronic Materials and Devices,* Wiley, Chichester 1991.
504. P. K. Hansma, J. Tersoff, *J. Appl. Phys.* **61** (1987) R1.
505. J. A. Golovchenko, *Science* **232** (1986) 48.
506. O. M. Magnussen, J. Hotlos, G. Beitel, D. M. Kolb, R. J. Behm, *J. Vac. Sci. Technol. B* **9** (1991) 969.
507. H. Neddermeyer, *Trends Analyt. Chem.* **8** (1989) 230.
508. R. M. Feenstra, J. A. Stroscio, J. Tersoff, A. P. Fein, *Phys. Rev. Lett.* **58** (1987) 1192.

509. Ph. Avouris, *J. Phys. Chem.* **94** (1990) 2246.
510. Ph. Avouris, I. W. Lyo, F. Bozso, *J. Vac. Sci. Technol. B* **9** (1991) 424.
511. Ph. Avouris, I.-W. Lyo, *AIP Conference Proceedings* **241** (1992) 283.
512. S. M. Lindsay, L. A. Nagahara, T. Thundat, U. Knipping, R. L. Rill, B. Drake, C. B. Prater, A. L. Weisenhorn, S. A. C. Gould, P. K. Hansma, *J. Biomol. Struct. Dyn.* **7** (1989) 279.
513. R. J. Driscoll, M. G. Youngquist, J. D. Baldeschwieler, *Nature* **346** (1990) 294.
514. P. G. Arscott, V. A. Bloomfield, *STM SFM Biol.* (1993) 259 – 272.
515. M. Amrein, H. Gross, R. Guckenberger, *STM SFM Biol.* (1993) 127 – 175.
516. S. F. Alvarado, *J. Appl. Phys.* **64** (1988) 5931.
517. R. Wiesendanger, H.-J. Güntherodt, G. Güntherodt, R. J. Gambino, R. Ruf, *Phys. Rev. Lett.* **65** (1990) 247.
518. J. K. Gimzewski, B. Reihl, J. H. Coombs, R. R. Schlittler, *Z. Phys. B* **72** (1988) 497.
519. R. Berndt, R. R. Schlittler, J. K. Gimzewski, *J. Vac. Sci. Technol. B* **9** (1991) 573.
520. M. Völcker, W. Krieger, T. Suzuki, H. Walther, *J. Vac. Sci. Technol. B* **9** (1991) 541.
521. Y. Kuk, R. S. Becker, P. J. Silverman, G. P. Kochanski, *J. Vac. Sci. Technol. B* **9** (1991) 545.
522. S. Akari, M. Ch. Lux-Steiner, M. Vögt, M. Stachel, K. Dransfeld, *J. Vac. Sci. Technol. B* **9** (1991) 561.
523. D. G. Cahill, R. J. Hamers, *J. Vac. Sci. Technol. B* **9** (1991) 564.
524. A. Manara, V. Sistori, *Surf. Interface Anal.* **15** (1990) 457 – 462.
525. J. Shin, K. N. Geib, C. W. Wilmsen, Z. Lillienthal-Weber, *J. Vac. Sci. Technol. A* **8** (1990) 1894 – 1898.
526. S. V. Hattangady, R. A. Rudder, M. J. Mantini, G. G. Fountain, J. B. Posthill, R. J. Markunas, *Mater. Res. Soc. Symp. Proc.* **165** (1990) 221 – 226.
527. A. Hupfer, J. Albrecht, D. Dietrich, *Fresenius' J. Anal. Chem.* **341** (1991) 439 – 444.

Chemical and Biochemical Sensors

Karl Cammann, Institut für Chemo- und Biosensorik, Münster, Germany (Chap. 1, 4, 5, Sections 2.2, 2.3.1.3, 2.3.1.4, 2.3.4, 2.4, 2.5)

Bernd Ross, Institut für Chemo- und Biosensorik, Münster, Germany (Sections 2.3, 2.3.1, 2.3.1.2)

Andreas Katerkamp, Institut für Chemo- und Biosensorik, Münster, Germany (Sections 2.3.2, 3.3.2)

Jörg Reinbold, Institut für Chemo- und Biosensorik, Münster, Germany (Sections 2.2.2, 2.3.3)

Bernd Gründig, SensLab GmbH - Bioelektrochemische Sensoren, Leipzig, Germany (Sections 3.1 – 3.2.3, 3.5)

Reinhard Renneberg, Hong Kong University of Science & Technology - Department of Chemistry, Hong Kong, PR China (Sections 3.1, 3.2.4 – 3.4)

1. Introduction to the Field of Sensors and Actuators 2
2. Chemical Sensors 4
2.1. Introduction 4
2.2. Molecular Recognition Processes and Corresponding Selectivities . 10
2.2.1. Catalytic Processes in Calorimetric Devices 10
2.2.2. Reactions at Semiconductor Surfaces and Interfaces Influencing Surface or Bulk Conductivities 11
2.2.3. Selective Ion Conductivities in Solid-State Materials 18
2.2.4. Selective Adsorption – Distribution and Supramolecular Chemistry at Interfaces 19
2.2.5. Selective Charge-Transfer Processes at Ion-Selective Electrodes (Potentiometry) 19
2.2.6. Selective Electrochemical Reactions at Working Electrodes (Voltammetry and Amperometry) 20
2.2.7. Molecular Recognition Processes Based on Molecular Biological Principles 21
2.3. Transducers for Molecular Recognition: Processes and Sensitivities 21
2.3.1. Electrochemical Sensors 21
2.3.1.1. Self-Indicating Potentiometric Electrodes 22
2.3.1.2. Voltammetric and Amperometric Cells 31
2.3.1.3. Conductance Devices 38
2.3.1.4. Ion-Selective Field-Effect Transistors (ISFETs) 48
2.3.2. Optical Sensors 51
2.3.2.1. Fiber-Optical Sensors 51
2.3.2.2. Integrated Optical Chemical and Biochemical Sensors 54
2.3.2.3. Surface Plasmon Resonance 56
2.3.2.4. Reflectometric Interference Spectroscopy 57
2.3.3. Mass-Sensitive Devices 58
2.3.3.1. Introduction 58
2.3.3.2. Fundamental Principles and Basic Types of Transducers 59
2.3.3.3. Theoretical Background 67
2.3.3.4. Technical Considerations 74
2.3.3.5. Specific Applications 77
2.3.3.6. Conclusions and Outlook 80
2.3.4. Calorimetric Devices 81
2.4. Problems Associated with Chemical Sensors 84
2.5. Multisensor Arrays, Electronic Noses, and Tongues 85
3. Biochemical Sensors (Biosensors) 87
3.1. Definitions, General Construction, and Classification . 87
3.2. Biocatalytic (Metabolic) Sensors . 88
3.2.1. Monoenzyme Sensors 89
3.2.2. Multienzyme Sensors 91
3.2.3. Enzyme Sensors for Inhibitors – Toxic Effect Sensors 93
3.2.4. Biosensors Utilizing Intact Biological Receptors 93
3.3. Affinity Sensors – Immuno-Probes 95
3.3.1. Direct-Sensing Immuno-Probes without Marker Molecules 96
3.3.2. Indirect-Sensing Immuno-Probes using Marker Molecules 98
3.4. Whole-Cell Biosensors 103
3.5. Problems and Future Prospects . 105
4. Actuators and Instrumentation . 106
5. Future Trends and Outlook 107
6. References 108

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_121

1. Introduction to the Field of Sensors and Actuators

When a person suffers the loss of one of the five senses it is recognized as a serious obstacle to carrying out everyday activities. In the case of blindness and deafness the expression "disabled" is even used as an indication of the severity of the loss, just as with those whose limbs (legs or arms) or other crucial body components are incapable of functioning properly. In the animate world generally, sensing and then acting or reacting is closely linked with the very essence of life, where input from some type of sensory system always precedes intelligent action. From the earliest days of mankind it was the availability of a powerful "coping mechanism" that constituted the driving force leading first to intelligent action and later to the unbiased accumulation of knowledge.

Any intellectual information structure capable of producing theories is dependent upon a reliable input of information. Imperfect observations, or perceptual processes subject to frequent errors or some bias, will never provide a satisfactory basis for working theories subject to verification by others. In this respect analytical chemistry can be regarded as one of the cognitive means available to the chemist interested in ascertaining the composition of the material world and the distribution and arrangement of atoms and molecules within a particular sample. Since both the nature and the three-dimensional arrangement of atoms in space determine the characteristics of matter, reliable information with respect to both is crucial for understanding the whole of the material world. This observation is equally valid for the inorganic, the organic, and also the living world. A high level of structured information is essential to the work of most natural scientists. The ability to describe in detail a complex chemical process (e.g., the combustion of fuel in an engine or waste in a municipal incinerator, the synthesis of minerals deep within the earth, or the course of biochemical reactions taking place in living systems) is crucially dependent on the quality of the information available. High technology is impossible without detailed information on the state of the system in question, as has been amply demonstrated in semiconductor technology, where knowledge of the purity of the materials and the state of cleanliness in the production plant is absolutely essential.

Complex and especially dynamic systems involving chemical processes are often described in terms of simplified models. The more chemical information is available, the better will be the model – and consequently the more complete the understanding and control over the system. Chemical analysis constitutes the primary source of information for characterizing the various states of such a system.

As a result of a tremendous increase in the demand for analytical chemical information, the number of analyses conducted for acquiring information about particular samples is increasing dramatically. It is thought that worldwide more than 10^9 analyses are now performed every day – with a double-digit growth rate. The latter because quality control has become a worldwide topic, agreed upon in many international trade contracts. Many complex systems are governed by time-dependent parameters, which makes it essential that as much information as possible be acquired within a particular period of time so that the various reactions taking place within the system can be followed, and to provide insight into the dynamics of the mechanisms involved. Devices designed to permit continuous measurement of a quantity are called *sensors.* Various types of sensors have been developed for monitoring many important *physical* parameters, including temperature, pressure, speed, flow, etc. The number of potential interferences is rather limited in these cases, so it is not surprising that the corresponding sensors offer a level of reliability considerably higher than that for sensors that measure *chemical* quantities (e.g., the concentration of a certain species, the activity of an ion, or the partial pressure of a gaseous material).

There are various major fields of application for chemical sensors: environmental studies, quality control of chemically produced compounds and processed food and biomedical analysis, especially for medical diagnostics. In the field of environmental analysis there is a great demand for the type of continuous monitoring that only a sensor can provide, since the relevant parameter for toxicological risk assessment is always the *dose* (i.e., a concentration multiplied by an exposure time). The study of environmental chemistry also depends upon a continuous data output that provides not only baseline lev-

els but also a reliable record of concentration outbursts. *Data networks* can be used to accumulate multiple determinations of a particular analyte regardless of whether these are acquired simultaneously or at different times within a three-dimensional sampling space (which may have dimensions on the order of miles), thereby providing information about concentration gradients and other factors essential for pinpointing the location of an emission source. The inevitable exponential increase in the number of chemical analyses associated with this need for maximum information threatens to create serious financial and also ecological problems, since most classical analytical chemical techniques themselves produce a certain amount of chemical waste.

Our society is characterized by a population with an average age that is continuously increasing. This has stimulated the widespread biomedical application of chemical sensors for monitoring the health of the elderly without confining them to hospitals. Sensors for determining the concentrations of such key body substances as glucose, creatinine, or cholesterol could prove vitally important in controlling and maintaining personal health, at the same time enhancing the quality of the patient's life. The same principle underlies sensor applications in intensive care wards. Sensors also offer the potential for significant cost reductions relative to conventional laboratory tests, and they can even improve the quality of the surveillance because continuous monitoring can lower the risk of complications. For example, the monitoring of blood potassium levels can give early warning of the steady increase that often precedes an embolism, providing sufficient time for clinical countermeasures.

It should be noted at the outset, however, that the development costs associated with a chemical sensor are typically many orders of magnitude higher than those for a sensor measuring a physical quantity. A satisfactory return on investment therefore presupposes either mass production in the case of low- cost sensors or special applications that would justify an expensive sensor.

Another important field of application for chemical sensors is *process control.* Here the sensor is expected to deliver some crucial signal to the actuators (valves, pumps, etc.) that control the actual process. Fully automated process control is feasible with certain types of feedback circuits. Since key chemical parameters in many chemical or biochemical processes are not subject to sufficiently accurate direct determination by the human senses, or even via the detour of a physical parameter, sensor technology becomes the key to automatic process and quality control. Often the performance of an entire industrial process depends on the quality and reliability of the sensors employed. Successful adaptation of a batch process to *flow-through reactor production technology* is nearly impossible without chemical sensors for monitoring the input, the product quality, and also – an increasingly important consideration – the waste. *Municipal incinerators* would certainly be more acceptable if they offered sensor-based control of SO_2, NO_x, and, even more important, the most hazardous organic compounds, such as dioxines. Compared with the high construction costs for a plant complete with effective exhaust management, an extra investment in sensor research can certainly be justified.

These few examples illustrate the increasing importance of chemical sensor development as a key technology in various closed-looped sensor/actuator-based process control applications. Rapid, comprehensive, and reliable information regarding the chemical state of a system is indispensable in many high technology fields; for this reason there is also an increasing need for intelligent or "smart" sensors capable of controlling their own performance automatically.

Finally, there is an urgent and growing demand for analytical chemists and engineers skilled in producing reliable chemical sensors delivering results that are not only reproducible but also accurate (free of systematic errors attributable to interferences from the various sample matrices).

2. Chemical Sensors [1–9]

2.1. Introduction

Until the introduction in the 1980s of the lambda-(λ) probe [10–12] for oxygen measurement in the exhaust systems of cars employing catalytic convertors (see Section 2.2.3), the prototype for the most successfully applied chemical sensor was certainly the electrochemical

pH electrode, introduced over 50 years ago and discussed in detail in Section 2.3.1.1. This device can be used to illustrate the identification and/or further optimization of all the critical and essential features characterizing sensor performance, including reversibility, selectivity, sensitivity, linearity, dynamic working range, response time, and fouling conditions.

Electrochemical electrodes for chemical analysis have actually been known for over 110 years (the Nernst equation was published as early as 1889), and they can therefore be considered as the forerunners of chemical sensors. In general, a chemical sensor is a small device placed directly on or in a sample and designed to produce an electrical signal that can be correlated in a specific way with the concentration of the *analyte,* the substance to be determined. If the analyte concentration varies in the sample under investigation, the sensor should faithfully follow the variation in both directions; i.e., it should function in a reversible way, and it should deliver analytical data at a rate greater than the rate of change of the system (quasi-reversible case). From the standpoint of practical applications, complete and fast reversibility is relatively unimportant so long as measurements (even readings of such as from a dosimeter) are acquired with sufficient rapidity compared with changes in the analyte concentration that are to be followed. Another restriction arises

from the demand for a reagent-free measurement: an extremely selective or specific cooperation between the analyte atom, ion, or molecule and the surface of a receptor-like sensor element must then be the sole basis for gaining information about the analyte concentration.

The general construction principle of a chemical sensor is illustrated in Figure 1. The most important part is the *sensing element* (receptor) at which the molecular or ionic recognition process takes place, since this defines the overall selectivity of the entire sensor. The sensing-element surface is sometimes covered by an additional surface layer acting as a protective device to improve the lifetime and/or dynamic range of the sensor, or to prevent interfering substances from reaching the sensing-element surface. The analyte recognition process takes place either at the surface of the sensing element or in the bulk of the material, leading to a concentration-dependent change in some physical property that can be transformed into an electrical signal by the appropriate transducer. The transducer cannot itself improve the selectivity (exceptions are transducer like electrochemical electrodes which intrinsically deliver an analyte proportional signal), but it is responsible for the sensitivity of the sensor, and it functions together with the sensing element to establish a dynamic concentration range. The electrical signal of the transducer is usually amplified by a device positioned close to the transducer or even integrated into it. If extensive electrical wiring is required between the sensor and the readout – control unit, the electrical signal might also be digitized as one way of minimizing noise pick-up during signal transmission. Several chemical signals produced by multiple sensing elements, each connected to its own specific transducer, can be transmitted via modern multiplexing techniques, which require only a two-wire electrical connection to the control unit also delivering the electrical power ("Bus"-technique in German).

If a biochemical mechanism (mostly enzymatic catalysis, immuno-chemical reaction or complementary DNA hybridization) is used in the molecular recognition step, the sensor is called a *biosensor.* Systems of this type are dealt with in Chapter 3.

Key features of chemical sensors subject to specification are, in order of decreasing importance:

1) Selectivity
2) Limit of detection (LOD) for the analyte
3) Accuracy (trueness – a new English term – more important than reproducibility)
4) Sensitivity (slope of the calibration curve)
5) Dynamic response range
6) Stability
7) Response time
8) Reliability (maintenance-free working time)
9) Lifetime

Selectivity. The *selectivity* of a chemical sensor toward the analyte is often expressed in terms of a dimension that compares the concentration of the corresponding interfering substance with an analyte concentration that produces the same sensor signal. This factor is obtained by dividing the sensitivity of the sensing device toward the corresponding interfering substance

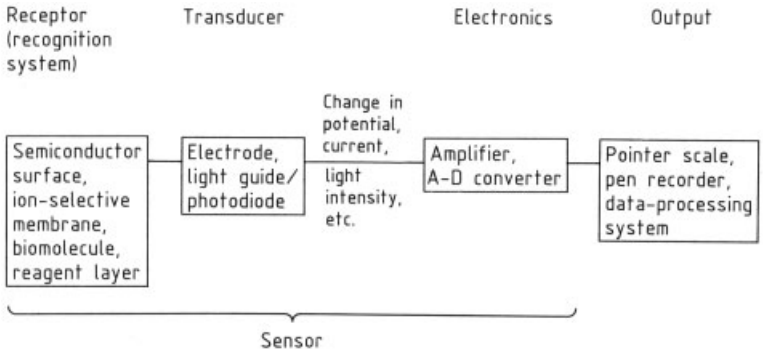

Figure 1. Schematic arrangement of a typical chemical or biochemical sensor [13]

(= slope of the calibration curve; see below) by the sensitivity for the analyte. Typical selectivities range from 10^{-3} to $<10^{-12}$ the last extremely small value stands for the sodium error of classical pH-glass-electrodes. In order to estimate the error attributable to limited selectivity the concentration of the interferent must be multiplied by the selectivity coefficient. If, for example, a sodium-selective glass electrode shows a typical selectivity coefficient toward potassium ions on the order of 10^{-3}, a thousand-fold excess of potassium- over sodium-ion concentration would produce an equivalent sensor signal, therefore doubling the reading and leading to an error of 100 %. It should be noted, however, that such selectivities are themselves concentration- and matrix-dependent and often depends on the way they are determined. This means that near the limit of detection of a particular sensor an interfering species will disrupt the measurement of an analyte to a greater extent than it will at higher concentrations. Therefore, response factors can rarely be used for correction purposes.

The selectivity is the most important parameter associated with a chemical sensor since it largely determines the trueness of the analytical method. Trueness (or *accuracy*) is more than simply precision, repeatability, or reproducibility: it represents agreement with the true content of the sample. This is the concentration found independent of the method, the time, the analyst, the laboratory, the country, etc. and is demanded in international trade agreements as comparability. All the other parameters mentioned are subject to systematic error, which can only be recognized if an analyte is determined by more than one method. Since selectivity is always limited, all chemical sensors are prone to report higher concentrations than a sample actually contains. In the case of environmental analysis this positive error can be regarded as a safety margin if relevant interferents are also present. In such a situation the sensor acts as a probe for establishing when a sample should be taken to be analyzed in the laboratory. The reproducibility of chemical sensor measurements is often well below 5 % relative.

However, it should also be noted that sensing elements are subject to poisoning, resulting in a diminished response for the analyte. Sensor poisons are substances that either cover the sensing surfaces, thereby influencing the analyte recognition process or, in the case of catalytic sensors, decrease the activity of the catalyst. Since these fouling effects are not linear, predictable, or mathematically additive, great care is necessary whenever data from a sensor array are treated by the methods of mathematical data management (e.g., pattern recognition, neural networks, etc.). The influence of sensor poisoning can be completely different and unpredictable for every single sensing element. There is no simple theory that permits the transfer of poisoning effects with respect to one sensing element to other parts of a sensing array as a way of avoiding recalibration. Thus, any attempt to overcome insufficient selectivity by the use of a sensor array with multiple sensor elements displaying slightly differing selectivities for the analyte and the interferents, followed by the application of chemometrics to correct for selectivity errors, is very dangerous in the context of unknown and variable sample matrices. Unfortunately, this is usually the situation in environ-

mental analysis. In the learning and/or calibration phase of sensor-array implementation it is essential that the sample matrix be understood. In the case of a known matrix and a closed system, and if the number of possible interferents is > 5, the number of calibration steps (involving different mixtures of the interfering compounds, selected to cover all possible concentration and mixture variations) required to provide reliable information on the 5-dimensional response space of the sensor array for a $5 + N$ array is extremely large (usually several hundred). With real samples the time necessary for calibrating the sensors might exceed the period of stability. Since with open environmental systems the type and number of interferents is not known, any attempt to correct for inadequate sensor selectivity by use of a sensor array increases the risk of systematic errors. This assertion is valid for every type of mathematical data treatment (pattern recognition, neural networks, etc.) for samples of unknown and/or variable matrices. However, for trace analysis in the sub-ppm range all analytical methods (with the exception of isotope dilution) are subject to systematic errors, as has been demonstrated with interlaboratory comparison tests using even the most sophisticated techniques [14]. Therefore one should not expect an accuracy of better than 1 – 5 % relative from sensor measurements.

Limit of Detection (LOD) for the Analyte. In analytical chemistry the *limit of detection* (LOD) is exceeded when the signal of the analyte reaches at least three times the general noise level for the reading. No quantitative measurement is possible at this low concentration; only the presence of the analyte can be assumed with a high probability (> 99 %) in this concentration range. The range of quantification ends at ten times the LOD. Therefore, the LOD of an analytical instrument must always be ten times smaller than the lowest concentration to be quantified. It is worth mentioning in this context that LOD values reported by most instrument manufacturers disregard the influence of interfering compounds. Furthermore, the actual LOD is often worse than published values with simple mono-analyte solutions when an analyte must be determined within its characteristic matrix (e.g., a pollution site, urine, or blood).

In some cases with extremely sensitive sensors a zero-point calibration might be difficult to perform, because traces of the analyte would always be present or might easily be carried into the calibration process by solvents or reagents. The lowest *measurable* level is then called the *blank value.* The LOD is then defined as three times the standard deviation of the blank value, expressed in concentration units (not in the units of the signal, as is sometimes improperly suggested).

Normally the LOD becomes worse as a sensor ages.

Sensitivity. As mentioned above, the *sensitivity* of a sensor is defined by the signal it generates, expressed in the concentration units of the substance measured. This corresponds to the slope of the corresponding calibration curve when the substance is the analyte, or to the so- called response curve for interferents. With some sensors the sensitivity rises to a maximum during the device's lifetime. A check of the sensitivity is therefore a valuable quality-assessment step. Intelligent sensors are expected to carry out such checks automatically in the course of routine performance tests. Since in most cases the sensitivity depends on such other parameters as the sample matrix, temperature, pressure, and humidity, certain precautionary measurements are necessary to ensure that all these parameters remain constant both during calibration and in the analysis of real samples.

Dynamic Response Range. The *dynamic response range* is the concentration range over which a calibration curve can be described by a single mathematical equation. A potentiometric sensing device follows a logarithmic relationship, while amperometric and most other electrochemical sensors display linear relationships. Both types of signal-to- concentration relationship are possible with optical sensors; in absorption measurements it is the absorbance with its logarithmic base that is the determining factor, whereas a fluorescence measurement can be described by a linearized function. The broader the concentration range subject to measurement with a given sensing system, the less important are dilution or enrichment steps during sample preparation. The measurement range is limited by the LOD at the low concentration end, and by

saturation effects at the highest levels. Modern computer facilities make it possible also to use the nonlinear portion of a calibration curve as a way of saving sample preparation time, but analytical chemists are very reluctant to follow this course because experience has shown that the initial and final parts of a calibration curve are generally most subject to influence by disturbances (electronic and chemical). A good chemical sensor should function over at least one or two concentration decades. Sensors with excellent performance characteristics include the lambda oxygen probe and the glass pH electrode, both of which cover analyte concentration ranges exceeding twelve decades!

Stability. Several types of signal variation are associated with sensors. If the signal is found to vary slowly in two directions it is unlikely to be regarded as having acceptable stability and reproducibility, especially since it would probably not be subject to electronic correction.

Output variation in a single direction is called *drift.* A steadily drifting signal can be caused by a drifting zero point (if no analyte is present) and/or by changes in sensor sensitivity (i.e., changes in the slope of the calibration curve). Drift in the sensor zero-concentration signal (*zero-point drift*) can be corrected by comparison with a signal produced by a sensor from the same production batch that has been immersed in an analyte-free solution.

When drift observed in a sensor signal is attributable exclusively to zero-point drift, and there has been no change in the sensitivity, a one-point calibration is permissible: that is, either the instrumental zero point must itself be adjusted prior to the measurement, or else the sensor signal for a particular calibration mixture must be recorded. Assuming there is no change in the slope of the calibration curve, this procedure is equivalent to effecting a parallel shift in the calibration curve. If the sensor is routinely stored in a calibrating environment between measurements then such a one-point calibration becomes particularly easy to perform.

Special attention must be directed toward the physical and chemical mechanisms that cause a sensor signal to drift. Differential techniques are satisfactory only if the effect to be compensated has the same absolute magnitude and the same sign with respect to both devices. A consideration often neglected in potentiometry is proper functioning of the reference electrode. Several authors [15–17] have suggested that in ion-selective potentiometry a "blank membrane" electrode (one with the same ion-selective membrane composition as the measuring electrode but without the selectivity-inducing electroactive compound) would create the same phase-boundary potential for all interfering compounds as the measuring electrode itself. This is definitely not the case, however. At the measuring electrode, analyte ions establish a relatively high exchange-current density, which fixes the boundary potential. The effect of a particular interfering ion will depend on its unique current–voltage characteristics at such an interface. Therefore, interfering ions may establish different phase-boundary potentials depending on the particular mixed potential situations present at the membranes in question, which means that a differential measuring technique will almost certainly fail to provide proper compensation.

If both the zero point of a sensor and its sensitivity change with time, a two-point calibration is necessary. The combination of a changing slope for the calibration curve and a change in the intercept is generally unfavorable, but this is often the situation with catalytic gas sensors. If the slope of the calibration curve also depends on the sample matrix, only standard addition techniques will help.

The stability of a chemical sensor is usually subject to a significant aging process. In the course of aging most sensors lose some of their selectivity, sensitivity, and stability. Some sensors, like the glass pH electrode, can be rejuvenated, while others must be replaced if certain specifications are no longer fulfilled.

Response Time. The *response time* is not defined in an exact way. Some manufacturers prefer to specify the time interval over which a signal reaches 90 % of its final value after a tenfold concentration increase, while authors sometimes prefer the 95 % or even 99 % level. The particular percentage value chosen represents a pragmatic decision, since most signal–time curves follow an exponential increase of the form:

$$\text{signal} \times \left(1 - \mathrm{e}^{-kt}\right)$$

where the true final value is unknown and/or will never be reached in a mathematical sense. Therefore, specification in terms of the time constant k is clearer. This corresponds to the time required for a signal to reach about 67 % of its final value, which can be ascertained without waiting for a final reading: as the slope of the curve ln (signal) vs. time. If the response is constant and independent of the sample matrix, this equation can be used to calculate a final reading immediately after the sensor has been introduced into the sample.

Typical response times for chemical sensors are in the range of seconds, but some biosensors require several minutes to reach a final reading. At the other extreme, certain thin-film strontium titanate oxygen sensors show response times in the millisecond range.

The response time for a sensor is generally greater for a decreasing analyte concentration than for an increasing concentration. This effect is more pronounced in liquids than in gaseous samples. Both the surface roughness of the sensor and/or the dead volume of the measuring cell have some influence on the response time. Small cracks in the walls of a measuring cell can function as analyte reservoirs and diminish the rate of analyte dilution.

With certain sensors the response time can also depend on the sample matrix. In the presence of strongly interfering substances the response time for a chemical sensor might increase as a result of an increase in the time required to reach final equilibrium.

Reliability. There are various kinds of reliabilities. One involves the degree of trust that can be placed in an analytical result delivered by a particular chemical sensor. Another is a function of the real time during which the sensor actually performs satisfactorily without a breakdown and/or need for repair.

There is no way to judge fairly the analytical reliability of a chemical sensor, since this depends strongly upon the expert ability of the analyst to choose a suitable sensor and a suitable sample preparation routine in order to circumvent predictable problems.

With respect to the first generation of sensors, the ion-selective electrodes, the experience of many users has led to the following "reliability hit list", with the most reliable devices cited first: glass pH electrodes, followed by fluoride, sodium glass, valinomycin-based potassium, sulfide, and iodide electrodes, and culminating in the electrodes for divalent ions. In a general way, sometimes very personal judgements regarding reliability have always been linked to a particular sensor's specificity toward the analyte in question.

Reliability can be improved considerably if analytical conclusions are based on measurements obtained by different methods. For example, it is possible to determine sodium either with an appropriate glass-membrane electrode or with a neutral carrier-based membrane electrode. If both give the same analytical results, the probability of a systematic error is very low. Likewise, potassium might be measured using selective membrane electrodes with different carrier molecules. This is of course different from analysis based on a sensor array, in which the selectivities of the individual sensors differ only slightly, and different also in the sense that no correction procedure is invoked in case two results are found to disagree. Use of the Nernst–Nikolsky equation (see Section 2.3.1.1), together with the introduction of additional electrodes for determining the main interferents, has been shown to be an effective way of correcting errors in the laboratory with synthetic samples. Since selectivity numbers (i.e., parameters like the selectivity coefficient) are influenced by the sample matrix through its ionic strength, content of surface-active or lipophilic compounds, or interfering ions, it is dubious whether data from real environmental samples should be corrected by this method.

The length of time over which a chemical sensor can be expected to function reliably can be remarkably great (in the range of years), as in the case of glass-membrane or solid-state membrane electrodes, the lambda probe (Section 2.2.3), and the Taguchi gas sensor (see Section 2.2.2). On the other hand, a biosensor that depends upon a cascade of enzymes to produce an analyte signal will usually have a short span of proper functioning (a few days only). Potentiometric ion-selective membrane electrodes and optical chemosensors have lifetimes of several months. In the case of biosensors it should be noted that anything that changes the quaternary space-orientation of the recognition biomolecules will destroy the proper function-

ing of the sensor. Enzymes can be influenced by such factors as pH, certain heavy-metal ions, certain inhibitors, and high temperature (resulting in denaturation).

Lifetime. As already mentioned above, some rugged solid-state sensors have *lifetimes* of several years. Certain sensors can also be regenerated when their function begins to deteriorate. The shortest lifetimes are exhibited by biosensors. Ion-selective electrodes and optical sensors based on membrane-bound recognition molecules often lose their ability to function by a leaching-out effect. In optical sensors the photobleaching effect may also reduce the lifetime to less than a year. On the other hand, amperometric cells work well for many years, albeit with restricted selectivities.

Problems associated with inadequate lifetimes are best overcome with mass-produced miniaturized replacement sensors based on inexpensive materials. Minimizing replacement costs may well represent the future of biosensors. Installation of a new sensor to replace one that is worn-out can also circumvent surface fouling, interfering layers of proteins, and certain drift and poisoning problems. The integrated optical system IOS developed at the ICB Muenster and described in Section 3.3.2 has shown for the first time how to work with pre-calibrated low-cost immuno-chips with long storage capability.

Comparison of Sensor Data with that Obtained by Traditional Analytical Methods. It is extremely difficult to compare sensor data with traditional data; indeed, generalizations of this type are rarely possible anywhere in the field of analytical chemistry. Sensor developers are often confronted with the customer's tendency to consider use of a sensor only if all else has failed. This means that the most adverse conditions imaginable are sometimes proposed for the application of a chemical sensor. It is also not fair to compare a device costing a few dollars with the most expensive and sophisticated instrumentation available, nor is it appropriate to compare the performance of a chemical sensor with techniques involving time- consuming separations. In this case only the corresponding detector should be compared with the sensor. Sometimes the very simple combination of a selective chemical sensor with an appropriate separation technique is the most effective way to obtain the redundant data that offer the highest reliability.

Chemical sensors are generally superior to simple photometric devices because they are more selective or faster (as in the case of optical sensors based on the photometric method), more flexible, more economical, and better adapted to continuous sensing. The latter advantage can also be achieved by traditional means via flow-through measuring cells, but this leads to a waste of material and sample, and also to the production of chemical waste. Reagent-free chemical sensors show their greatest advantages in continuous-monitoring applications, in some of which they are called upon to fulfill a control function without necessarily reporting an analytical result, as in the case of the lambda probe described below (Section 2.2.3).

2.2. Molecular Recognition Processes and Corresponding Selectivities

In any sensing element the functions to be fulfilled include sampling, sample preparation, separation, identification, and detection. Therefore, successful performance in these tasks largely determines the quality of the chemical sensor as a whole. Selective recognition of an analyte ion or molecule is not an easy task, since more than five million compounds are known, and a real sample may contain hundreds of potentially interfering compounds. Recognition can be accomplished only on the basis of unique characteristics of the analyte in question. Several different recognition processes are relevant to the field of chemical sensors, ranging from energy differences (in spectroscopy) to thermodynamically determined variables (in electrochemistry), including kinetic parameters (in catalytic processes). The most specific interactions are those in which the form and the spatial arrangement of the various atoms in a molecule play an important role. This is especially true with biosensors based on the complementary (lock-and-key) principle. Here the analyte molecule and its counterpart have exactly complementary geometrical shapes and come so close together that they interact on the basis electrostatic interaction or with weak van der Waals interaction

forces. Apart from biomolecules, supramolecular chemistry has been given increased attention [18], [19].

2.2.1. Catalytic Processes in Calorimetric Devices

Pellistors are chemical sensors for detecting gaseous compounds that can be oxidized by oxygen. A catalyst is required in this case because the activation energy for splitting a doubly-bonded oxygen molecule into more reactive atoms is too high for the instantaneous "burning" of oxidizable molecules. In most cases platinum is the catalyst of first choice because of its inertness. The principle underlying oxidizable-gas sensors involves catalytic burning of the gaseous analyte, which leads to the production of heat that can in turn be sensed by various temperature-sensitive transducers. Often what is actually measured is the increase in electrical resistance of a metal wire heated to an elevated temperature (ca. 300 – 400 °C) by the current flowing through it. However, it is also possible to use more sensitive semiconductor devices or even thermopiles in order to register temperature changes of $< 10^{-4}$ °C.

In order to understand the selectivity displayed by a pellistor toward various flammable compounds it is necessary to consider the elementary steps in the corresponding catalytic oxidation of analyte at the catalyst surface. Since this is in fact a surface reaction, various adsorption processes play a dominant role. First, oxygen must be adsorbed and chemisorbed, permitting the oxygen double bond to be weakened by the catalyst. Then the species to be oxidized must also be adsorbed onto the same catalyst surface where it can subsequently react with the activated oxygen atoms. The process of adsorption may follow one of two known types of adsorption isotherms, as reflected in the calibration curves for these devices. An equilibrium consisting of adsorption, catalyzed reaction, and desorption of the oxidized product leads to a constant signal at a constant analyte concentration. The sensor response function is influenced by any change in the type or number of active surface sites, since this in turn affects both the adsorption processes and the catalytic efficiency. Likewise, compounds that have an influence on any of the relevant equilibria will also alter the calibration parameters. Especially problematic are strongly adsorbed oxidation products, which lower the turnover rate of analyte molecules and thus the sensitivity of the gas sensor.

Given the sequence of events that must occur when an organic molecule is oxidized by atmospheric oxygen, thereby delivering the heat that is actually to be measured, one can readily understand the importance of changes in the adsorption and desorption equilibria. The selectivity observed with such a calorimetric device arises not because some gas reactions are associated with larger enthalpy changes than others, but rather because those gas molecules that exhibit the most rapid adsorption and desorption kinetics are associated with the highest turnover rates. The latter of course depend on molecular-specific heats of ad- and desorption, which have a major influence on the overall reaction kinetics. A change in these specific heats always results in a corresponding change in sensor selectivity. Consequently, any change in the catalyst material, its physical form, or its distribution within the mostly ceramic pellet will in turn alter the gas selectivity and sensitivity. The same is true for variations in the working temperature. On the other hand, nonreacting compounds can also influence a sensor's response and thereby the calibration function if they in some way affect one of the relevant equilibria and/or the catalytic power of the catalyst. Catalyst poisons such as hydrogen sulfide or organic silicon compounds show a strongly detrimental effect on sensor response.

Summarizing the selectivity characteristics of these calorimetric gas sensors, any gas will be subject to detection if it can be catalytically oxidized with a high turnover rate by atmospheric oxygen at elevated temperature. From a kinetic point of view, smaller molecules like carbon monoxide or methane are favored. In contrast to biosensors there is here no precise molecular "tight-fit" recognition of the geometrical form of the analyte. Thus, the selectivity of this type of gas sensor is rather limited. Differences in the often reaction-rate controlled adsorption and desorption processes for different oxidizable gas molecules are not sufficiently large to allow selective detection of only a single compound. However, this is not necessarily a disadvantage in certain sensor applications, such as detecting

the absence of explosive gases (especially important in the mining industry) or carbon monoxide in an automobile garage. In the latter case any positive error resulting from gasoline interference could in fact be regarded as providing a safety margin, since unburned gasoline should be absent from such locations as well.

2.2.2. Reactions at Semiconductor Surfaces and Interfaces Influencing Surface or Bulk Conductivities

Introduction. Since the early 1960s it has been known that the electrical conductance of certain semiconductor materials such as binary and ternary metal oxides (e.g., SnO_2 [20], ZnO [21], Fe_2O_3 [22] – all of which are *n*-semiconductor materials – and CuO or NiO [23] – *p*-semiconductors) depends on the adsorption of gases on their solid surfaces. The underlying principle here involves a transfer of electrons between the semiconductor surface and adsorbed gas molecules, together with charge transduction in the interior of the material. Typical gases detected by semiconducting devices include oxidizable substances such as hydrogen, hydrogen sulfide, carbon monoxide, and alkanes (SnO_2, ZnO, etc.), as well as reducible gases like chlorine, oxygen, and ozone (NiO, CuO).

In 1967 both Shaver [24] and Loh [25] described effects achievable with oxide semiconductors modified by the addition of noble metals (e.g., Pt, Pd, Ir, Rh), and since that time the sensitivity and selectivity of semiconductor sensing devices has been significantly enhanced. Intense efforts in this direction, coupled with the further addition of metal oxides [26–28], resulted in widespread application of semiconductor gas sensors beginning in the 1970s.

One of the earliest SnO_2 sensors, designed by N. Taguchi, is referred to as the "Figaro sensor" (see below). Sensors of this type make it possible to detect as little as 0.2 ppm of an oxidizable compound such as carbon monoxide or methane [29]. Nevertheless, certain details of the associated sensing mechanism are still not fully understood theoretically. An important aim is to overcome limitations of the present generation of sensors, especially instability, irreproducibility, and nonselectivity [30].

Construction and Characterization of Semiconductor Sensors. Semiconductor gas sensors are characterized by their simple construction. A schematic overview of the construction principle of a homogeneous semiconducting gas sensor is provided in Figure 2 A [31]. Sensor operation is based on a change in the surface resistance (or conductance) of an oxidic microcrystalline semiconductor in the presence of interacting gases. A time-dependent record reflecting transient exposure of a sensor to a gas leading to an increase in conductance is illustrated in Figure 2 B. The sensitivity of a semiconductor sensor is strongly affected by its operating temperature, which is normally in the range 200 – 400 °C. Chemical regeneration of the oxide surface is possible by a reheating process.

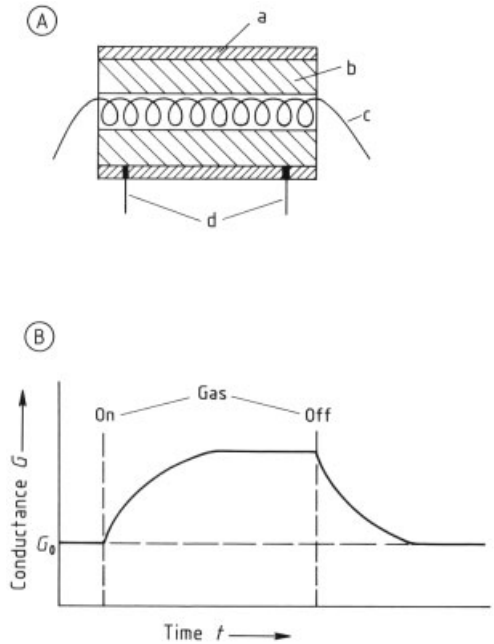

Figure 2. A) Schematic drawing of a homogeneous semiconducting gas sensor; B) Result of transient exposure of the sensor to a gas that increases the conductance [31]
a) Semiconductor; b) Ceramic; c) Heater; d) Contacts
(with permission from Elsevier, Amsterdam)

A commercial sensor of the "Figaro" type (TGS 813) is shown in Figure 3 A; 400 000 such sensors had already been sold in 1988 [32].

Miniaturization leads to a more modern version of the SnO_2 sensor, normally prepared by thick-film techniques (Fig. 3 B) in which thin

SnO_2 films, insulator layers (SiO_2), and integrated heating films are sputtered onto silicon substrates. This approach is compatible with high rates of heating and low-cost production [13].

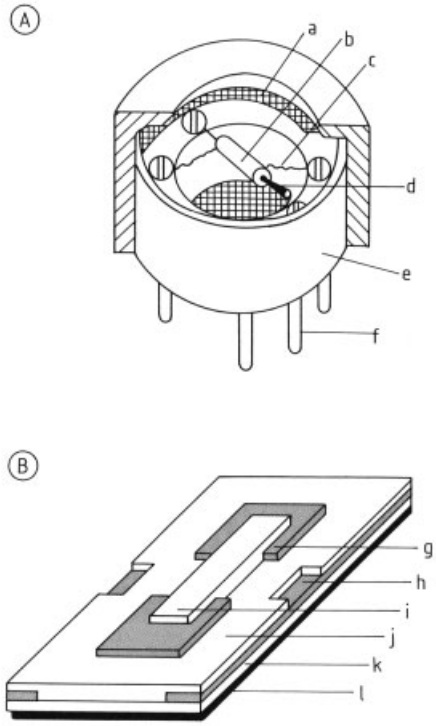

Figure 3. Design characteristics of gas sensors (with permission from VDE-Verlag, Berlin)
A) Figaro Type TGS 813 [32]; B) SnO_2 thick-film type [13]
a) Stainless steel screen; b) Sensor element; c) Leads; d) Heater; e) Epoxy resin; f) Nickel contact; g) Gold; h) Heat contact; i) SnO_2; j) Insulating layer; k) Heater; l) Substrate

Working Principles and Theory. The mechanisms responsible for semiconductor gas-sensor operation can be divided into two classes. The first class involves changes in bulk conductance (transducer function), while the second relies on changes in surface conductance (receptor function). The physical phenomena associated with these two mechanisms are shown schematically in Figure 4 A [33]; Figure 4 B addresses the same problem at the microstructural level [34].

The description of functional principles that follows relates directly to *n*-semiconductors, but its application to *p*-semiconductors is straightforward. In a first step, oxygen molecules from the air form a layer of more or less strongly adsorbed (chemisorbed/ionisorbed) oxygen molecules at the surface, resulting in a local excess of electrons. In other words, oxygen acts as a *surface acceptor,* binding electrons from the surface space – charge layer. With respect to the principal energy states (levels) of the electrons in the surface space – charge layer, ionisorption results in a decrease in the electron concentration and an increase in the electronic energy (Fig. 5) [32].

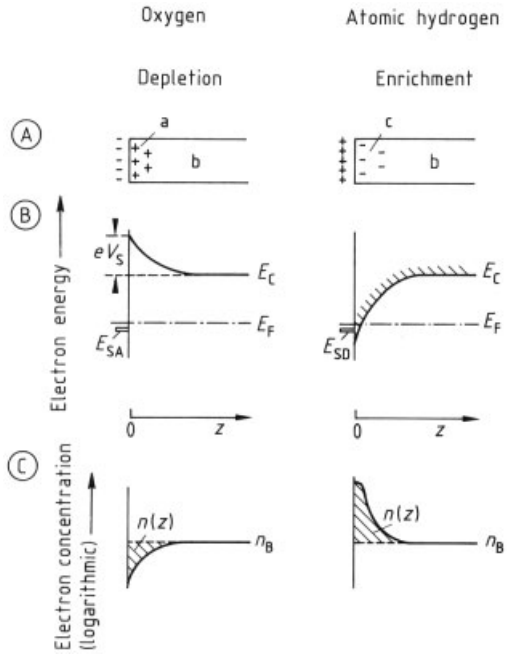

Figure 5. Space – charge layer at the surface of ZnO after exposure to oxygen and hydrogen (with permission from VDE-Verlag, Berlin)
A) Vacuum; B) Band structure in the vicinity of the conducting edge; C) Concentration n (z) of electrons in the conductance band [32]
a) Ionized donors; b) Crystal; c) Electrons
E_{SA} = Acceptor level; E_{SD} = Donor level; n_B = Bulk electron density; z = Distance from surface; eV_S = Surface energy barrier height; E_C = Conductance band edge; E_F = Fermi level

Subsequent reaction with reducing gases (e.g., CH_4 or CO) leads to an increase in charge density (and therefore an increase in conductivity), associated with three possible mechanisms [23], [31], [35]:

1) Adsorption of the reducing molecules as donors, causing electrons to be shifted into the conductance band of the oxide

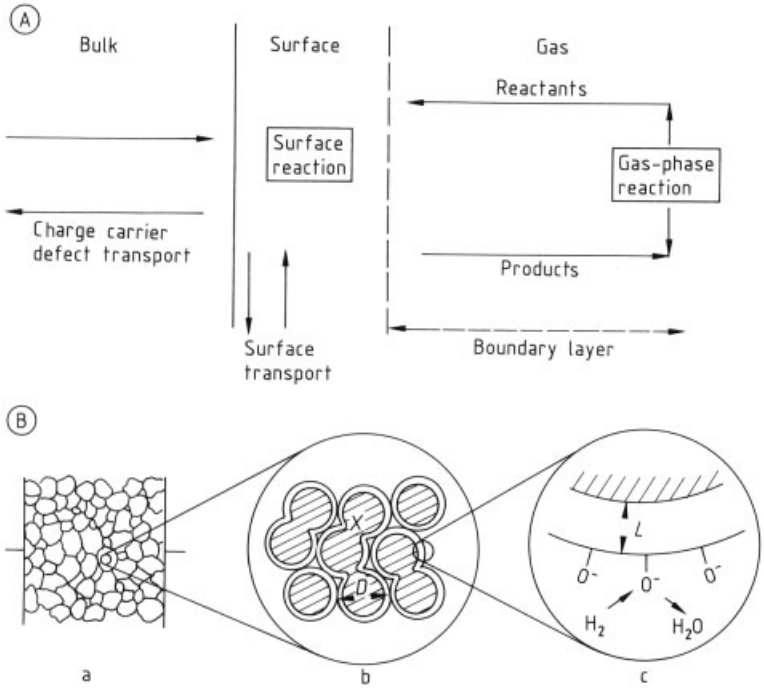

Figure 4. A) Illustration of the physical and chemical phenomena involved in the transduction of a change in the gaseous atmosphere into a change in conductance of a ceramic oxide (with permission from Adam Hilger, Bristol); B) Surface reactions (receptor function) may occur between adsorbed species and defects in the solid (transducer function), between different adsorbed species, or between gas molecules and adsorbed species (or any or all of these together), and the surface reaction may involve a catalyst [33], [34] (with permission from Elsevier, Amsterdam)
a) Element (output resistance change); b) Microstructure (transducer function); c) Surface (receptor function)
D = Particle size; X = Neck size; L = Thickness of space – charge layer

2) Reaction of the reducing molecules with ionisorbed oxygen under conditions leading to the production of bound electrons
3) Reduction of oxidic oxygen by the reducing molecules, resulting in oxygen vacancies which act as donors, thereby increasing the conductance

Cases 1) and 2) alter the amount of charge stored in the *surface states*, and therefore the amount of charge of opposite sign in deeper parts of the region. For a theoretical derivation of the relationship between the conductance of a semiconducting oxide layer and the composition of the gaseous surroundings, the following facts must be considered [30], [36]:

Oxygen must be present; that is, these sensors respond only to nonequilibrium gas mixtures containing both combustible gases (CO, hydrocarbons, H_2, etc.) and oxygen.

There exists a temperature of maximum response; that is, the relative change in conductivity upon introduction of a combustible gas increases with increasing temperature, but falls to zero at sufficiently high temperatures.

With respect to the relationship between conductivity (σ) and gas partial pressure, Equation (1) has been found to apply:

$$\sigma \sim p^{\beta} \tag{1}$$

where p is the partial pressure of the combustible gas and β is generally in the range 0.5 – 1.0 depending on the mechanism.

This microscopic model is limited to idealized thin films (10 – 100 nm of, for example, SnO_2). Here the gas – solid interactions can be described in terms of the electronic surface states, which are related to energy levels in the band gap model (Fig. 5). Other surface phenomena involved are lattice defect points, trace impurities, and material segregations near the surface. Crystal dislocations and adsorbed atoms or molecules also play an important role in states

at the gas – solid interface. As far as *n*-semiconductors are concerned, both adsorbed hydrogen and oxygen vacancies can function as surface donors, whereas ionisorbed oxygen is a surface acceptor. A surface potential difference develops as a consequence of negatively charged adsorbed oxygen (O_2^- and O^{2-}) and positively charged oxygen vacancies within the space – charge region below the surface layer. Increasing the amount of negatively charged oxygen at the surface causes the surface potential to increase up to the Fermi level (the highest level occupied by electrons). This defines the surface potential and represents the surface state, but the precise magnitude of the potential is a function of the oxygen partial pressure. The distribution of the various adsorbed oxygen species depends on the temperature, and is influenced by the presence of hydrogen and other gaseous compounds. With respect to the two possible charged oxygen species, O_2^- and O^-, it can be assumed that only O^- is reactive, and that the rate of interconversion of the species is low compared to the rate of the surface combustion reaction, consistent with the following kinetic scheme [33], [37]:

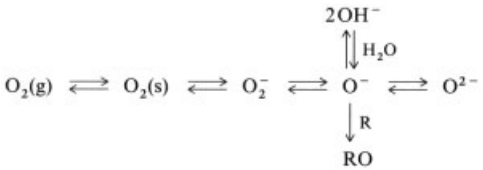

At low temperature the adsorbed species is mainly O_2^-, which is converted into O^- when the temperature is increased above 450 K.

If a combustible gas R reacts with the adsorbed oxygen species, a steady state occupancy Θ of the surface state is established, which is less than the equilibrium occupancy in air. The following mechanism can be assumed:

$$n' + 1/2\,O_2(g) \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} O^-(s)$$

$$R(g) + O^-(s) \xrightarrow[k_2]{} RO(g) + n'$$

where n' denotes a conductance electron.

Necessary conditions for sensitivity with respect to the partial pressure p of the combustible gas are then:

$$k_{-1} << k_2 p\,(R) \text{ and } k_1 p\,(O_2)^{1/2} << k_2 p\,(R) \quad (2)$$

in which case

$$\Theta = \frac{k_1 p\,(O_2)^{1/2}}{k_{-1} + k_1\,(O_2)^{1/2} + k_2 p\,(R)} \quad (3)$$

The third case described above is somewhat different: here the observed effect is related to a change in bulk conductance as a function of stoichiometric changes in oxygen activity in the interior of the crystal lattice. The change in conductivity can be described by the relationship [33]:

$$\sigma = \sigma_0 \exp\,(E_A/kT)\,p\,(O_2)^{\frac{1}{n}} \quad (4)$$

where σ_0 denotes the electrical conductivity in air, k denotes the Boltzmann constant, T is the temperature in Kelvin, E_A is an activation energy, and the sign and value of n depend on the nature of the point defects arising when oxygen is removed from the lattice. At low temperature (< 500 °C) the rate-determining step is interfacial combustion, whereas oxygen vacancies within the lattice dominate at higher temperatures below ca. 1000 °C.

Thus, the observed overall conductance represents a combination of surface effects (both electronic and ionic) together with grain-boundary and volume-lattice effects. The principles elaborated above have been applied widely in the control of oxygen in combustion processes at high temperatures, as in automobile engines.

Parameters. Under a constant set of conditions (temperature, air, pressure, humidity, etc.) the ratio of the conductance in sample air to that in pure air (G/G_0) is proportional to the analyte concentration (Fig. 6). Apart from the concentration of adsorbed gases, other parameters, both physical and chemical, may also have an influence on the conductance and the optimum working conditions. Considering both components of the conductance change described above, the most important factors influencing conductance are:

1) The microstructure of the semiconducting particles and the surface composition, both of which are characterized by the crystallite size D, the grain-size distribution, and the coagulation structure, all subject to some control through the inclusion of additives.
2) The working temperature of the material, which has a unique optimum (maximum)

value for a particular material composition, and must therefore be determined separately for every analyte gas and every substrate composition (Fig. 7).

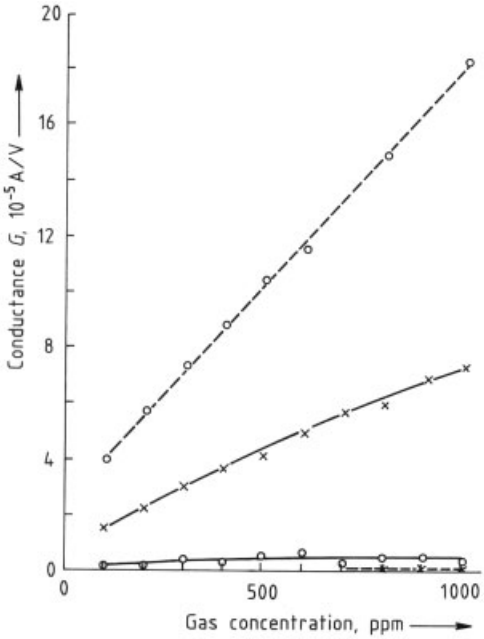

Figure 6. Sensitivity data from commercial SnO_2-based sensors, with sintered layers containing stabilizing additives like SiO_2 (conductance as a function of CO and CH_4 concentration in air); the Type 711 sensor probably operates at a lower temperature and incorporates different catalytically active additives than the Type 813 sensor [31], [38] (with permission from Elsevier, Amsterdam)
– – – – – Sensor Type 711; ——— Sensor Type 813; × = CH_4; o = CO

Tin dioxide has become the favorite material for sensor applications of this type because of its simple preparation and high sensitivity at low operating temperature. Commercially available sensors are usually sintered $\geq 700\,°C$ to ensure sufficient mechanical strength; the cross-sectional diameter of the SnO_2 particles is typically ≥ 20 nm. Unfortunately there is a sharp decrease in gas sensitivity (corresponding to a decrease in the resistance of the SnO_2) when the particle size D increases beyond about 6 nm (Fig. 8 A) [34]. Additives such as barium, phosphorus, or lanthanum oxides can be used to restrict thermal growth of the SnO_2 particles to a limit of ca. 10 nm (Fig. 8 B) [34].

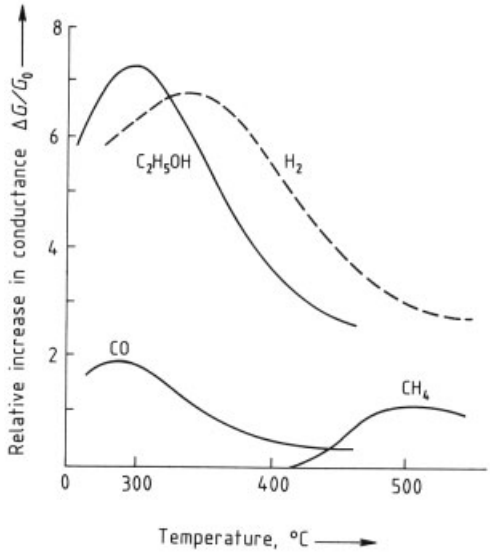

Figure 7. Sensor response to reducing gases (1000 ppm in air) as a function of temperature (sintered layer of SnO_2, thickness 0.05 mm, 0.05 wt % Sb; based on [39], [31]) (with permission from Elsevier, Amsterdam)

Various approaches have been applied to interpreting the parameters influencing gas sensitivity. Models that relate gas sensitivity to grain size (D) and the geometry of contacts between the SnO_2 particles include the intergrain model and the neck model. An extended approach taking into account both models has been developed by Yamazoe and Miura [34], in which gas sensitivity is related to electron charge transfer within the microstructure of the polycrystalline elements (cf. Fig. 4 B).

Selectivity. The selectivity of all semiconductor gas sensors is very limited, but it can be optimized within narrow limits by choosing the best operating temperature and the most appropriate dopant. As in the case of pellistors, overall selectivity is controlled by the ultimate rate-determining step; i.e., the compound that

shows the highest oxidation rate determines the selectivity. As is generally the case, the slowest step controls the overall reaction rate. In order to construct a sensor that selectively transforms only the analyte into another product, side reactions must be prevented between the compound that is actually detected (here oxygen) and compounds other than the analyte. This is nearly impossible to achieve in the case of a considerable excess of interfering compounds. Since the physicochemical phenomenon underlying heterogeneous catalysis is based on adsorption–desorption steps and subject to the kinetics of the heterogeneous chemical reaction itself, no known mechanism displays exceptional selectivity. It is not generally possible for a solid-state surface to recognize a single class of molecules because the forces involved in the ad- and desorption steps are rather strong, and are influenced only by overall molecular design and not by details of molecular structure, in contrast to the situation with biocatalysts. It must also be kept in mind that every compound can be considered as a potential interferant, since it might influence the heterogeneous rate constant for catalysis even if it does not itself react with oxygen. This also means that every catalyst poison, such as silicones or even dust blocking the active surface, is likely to interfere in a measurement, which in turn limits the possibility of using chemometric methods to correct for errors caused by lack of selectivity.

Humidity effect. Humidity often influences both the zero point of a Taguchi sensor and the sensitivity, which makes corrections difficult as shown in Figure 9.

Table 1 demonstrates that this is the general behavior of semiconductor gas sensors. The slope of the calibration curve can vary by up to a factor of 5. In this table, three different Taguchi sensors (TGS 800, TGS 813, and TGS 822) were carefully compared [257] with respect to their humidity interference using ethanol as the analyte.

This behavior has to be taken into account if these devices are used in electronic noses!

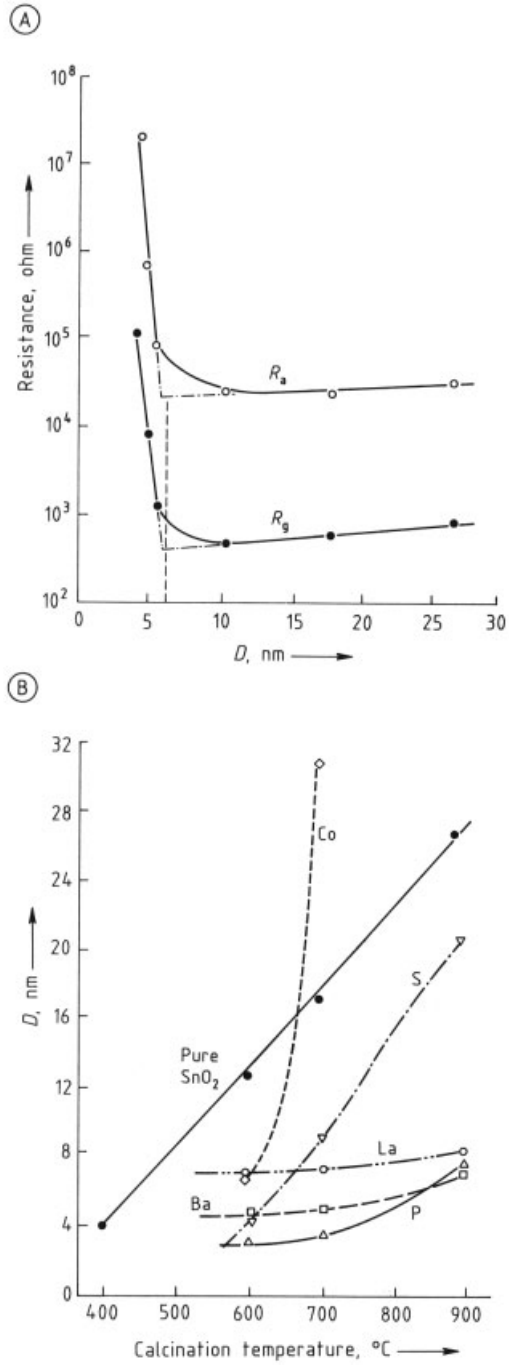

Figure 8. A) Influence of crystallite size (D) on electrical resistance for an SnO_2-based sensor in dry air (R_a) and in air containing 800 ppm H_2 (R_g) at 300 °C (elements sintered at 400 °C); B) Changes in SnO_2 crystallite size as a function of calcination temperature for pure SnO_2 samples and SnO_2 impregnated (5 %) with oxides of various foreign metals [34]
(with permission from Elsevier, Amsterdam)

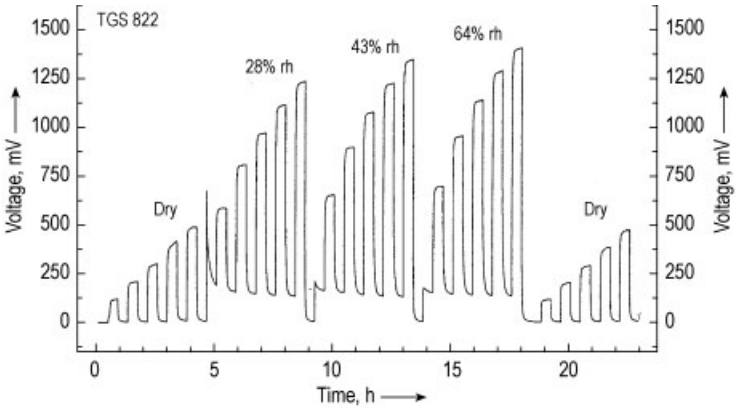

Figure 9. Effect of the relative air humidity on the response of an SnO_2 semiconductor sensor (Taguchi Type TGS 822) on the calibration cycle with ethanol as analyte. The ethanol concentrations were varied from 13, 26, 65 ppm and the relative humidity between dry, 23 %, 43 %, and 64 % rh.

Table 1. Humidity interference of SnO_2-based gas sensors. Analyte, ethanol; relative humidity rh; S = sensitivity /slope of calibration curve at 12.5 ppm ethanol; norm = factor of sensitivity increase in corresponding humid air (values in parenthesis)

Ethanol, ppm	rh, %	*S*, mV/ppm		
		TGS 800	TGS 813	TGS 822
13	0.00 ± 0.00	14.368 (1.00)	3.568 (1.00)	9.571 (1.00)
26	22.23 ± 0.42	57.752 (4.02)	13.850 (3.88)	47.423 (4.95)
39	42.13 ± 1.18	61.776 (4.30)	15.488 (4.34)	52.851 (5.52)
52	64.43 ± 0.93	63.932 (4.45)	16.801 (4.71)	54.275 (5.67)
65	0.00 ± 0.00	13.978 (0.97)	3.330 (0.93)	9.296 (0.97)

2.2.3. Selective Ion Conductivities in Solid-State Materials

The lambda probe, based on selective oxygen ion diffusion through solid yttrium-doped ZrO_2 above 400 °C, is an extremely selective sensor for gaseous oxygen. Frequently used for measuring the residual oxygen content in exhaust gases from internal combustion engines, the lambda probe is primarily a potentiometric device that functions as a concentration cell in which gaseous oxygen is in equilibrium with lattice oxygen in the solid electrolyte (Fig. 10). A higher concentration of oxygen on one side of the membrane leads to a potential difference which, according to the Nernst equation, is proportional to the ratio of the two oxygen concentrations. Likewise the fluoride ion-selective electrode based on selective fluoride-ion diffusion through solid-state (crystalline) LaF_3 is an extremely selective sensor for fluoride ions in solution.

Selectivity in both of these cases is controlled by the size and the charge of the diffusing ion. The diffusion process takes place via discrete ion jumps to nonoccupied lattice positions and/or from and to interstitial positions. Ion selection here can be regarded as a filtering process, with the additional requirement for an appropriate ionic charge to prevent overall electroneutrality. Because the analyte molecules and ions involved are rather small, they can be separated quite readily from larger molecules or ions. Thus, the solid-state material functions like an ideal permselective membrane. Any solid state material characterized by exceptionally high ionic conductivity, such as Ag_2S (which has extremely high Ag^+-ion conductivity), will be a very good sensor for that ion. In the case of the ion-selective membrane electrodes used as Nernstian potentiometric sensors in solution, however, more detailed consideration is necessary (see Section 2.3.1.1).

Since the analyte recognition process is here rather straightforward, and because the filtering effect is accompanied by additional restrictions (appropriate charge and multiple straining), the selectivities obtained through this approach to

measuring are extremely high. Many of the compounds that interfere do so not through direct participation in the solid-state diffusion process but rather by affecting the analyte concentration on the sensor surface by other means (e.g., reactions with the analyte, etc.). Thus, with the help of appropriate catalysts and high temperatures the λ-probe can also be used for the detection of other analytes that are oxidized at the surface and thereby reduce the amount of adsorbed oxygen.

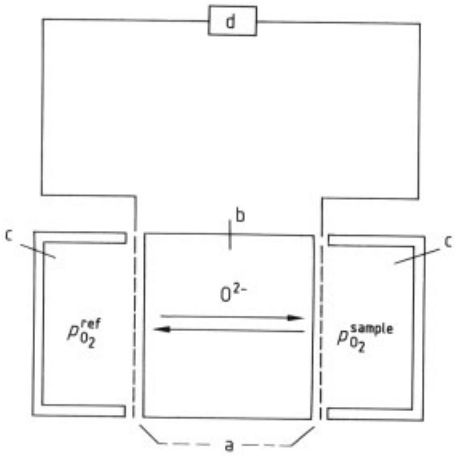

Figure 10. Schematic representation of a lambda probe operated as a potentiometric sensor
a) Porous electrode; b) ZrO_2, doped; c) Oxygen; d) Voltage measurement system

2.2.4. Selective Adsorption – Distribution and Supramolecular Chemistry at Interfaces

Supramolecular chemistry has been a source of fascinating new molecular recognition processes. At the active center of a biocatalyst (enzyme) and/or in the binding region of a large antibody macromolecule, the recognition process is attributable mainly to a perfect bimolecular approach of complementary structures. Association or binding occurs only if the relatively weak intermolecular forces involved (dipole – dipole and van der Waals forces) are further supported by a uniquely favorable bimolecular approach that encourages interatomic interactions. A selective reaction of this type requires that the host molecule be carefully tailored so that it matches the guest molecule perfectly. Molecular modeling with attention to the activation energies for closest approach is an especially valuable tool in the design of synthetic host molecules for particular analytes. The analyte can be either a charged or an uncharged species. In the case of an uncharged molecule, however, selective binding of the analyte to the host molecule must be detected with the aid of a mass-sensitive transducer or by following the changes of certain optical parameters.

Further research in the field of supramolecular chemistry for sensors seems warranted given the prospect for improved selectivity based on custom-tailored synthetic guest molecules for the selective binding of particular host molecules (analytes). Even differentiation between enantiomers may be achievable. The latter could be demonstrated by employing known stationary phases used in gas chromatography for the separation of enantiomers also on the surface of mass sensitive transducers. It has been claimed that the technique of molecular imprinting polymers (MIP) is able to perform a similiar high selectivity. While this may be achievable in chromatographic separations due to the high repetition of multiple interface equilibria, it has not yet been convincingly shown for a chemical sensor with a single "hit and recognize" event. The rigidity of the polymer backbone of the imprint makes it highly unlikely that the same recognition power as flexible biomolecules with their "induced-fit" approach can be achieved.

2.2.5. Selective Charge-Transfer Processes at Ion-Selective Electrodes (Potentiometry)

An important aim of quantitative analytical chemistry is the selective detection of specific types of ions in aqueous solution. An electrode capable of selectively detecting one type of ion is called an *ion-selective electrode* (ISE). The operating principle of a particular ISE depends upon the nature of the interaction between ions in solution and the surface of that electrode. Consistent with thermodynamic distribution, the ion in question crosses the phase boundary and interacts with the electrode phase after first discarding its hydrate shell (new studies have shown that this may not always be the case in a quantitative sense, since water can be detected in ion-selective PVC membranes), which results in a net flow of ions across the boundary. Ideally, the counterion would not take part in this flow, so

a charge separation would develop, causing the counterion to remain in the neighboring border area. This is almost the situation when there will be no net energy gain by this transfer due to some supramolecular binding, and a distribution because of the lipophilicity of counterions should be less pronounced. The resulting electrode potential (difference) between the interior of the electrode and the bulk of the sample solution can be described thermodynamically and kinetically – again, in the ideal case – by the Nernst equation.

The selectivity of such a potentiometric device is that of an ideal Nernstian sensor, the potential of which remains stable irrespective of the fact that some current may be flowing through the sensor interface. An ideal Nernstian sensor is characterized by a current – voltage curve that is very steep and nearly parallel to the current axis.

The selectivity of a potentiometric ion-selective interface is controlled largely by interfacial charge-transfer kinetics and not by the stability constant for reaction of the analyte ion with the electroactive compound (e.g., valinomycin in the case of potassium ion). The exchange-current density for the analyte ion relative to interfering ions determines the extent of potentiometric selectivity. Any factor that increases the former (conditioning of the electrode with the measuring ion, ion-pair formation by lipophilic counterions, etc.) leads to a better and more selective Nernstian sensor. The ideal electroactive compound apparently behaves like a selective charge-transfer catalyst for the analyte ion. Advantages that might perhaps be gained from tailored host molecules for accelerating the transfer of analyte ions into the surface of the electrode have not yet been fully investigated. Recognition of an analyte ion depends upon size and charge as well as the speed with which the ion loses its solvent sphere prior to entering the electrode phase. Therefore, the electroactive compounds most suitable for embedding in the membrane of an ion-selective electrode (usually consisting of PVC containing a plasticizer) are those that play an active role in successive replacement of the often firmly held hydrate shells of the analyte ions. Results clearly demonstrate that the thermodynamic complex stability constants determined in homogeneous solution (e.g. via NMR) do not mirror the potentiometric selectivity! A newly synthesized compound for sodium ions, as an example, turned out to yield a very selective PVC membrane for $Pb(OH)^+$ ions [40]. While in theory supramolecular recognition worked perfectly well in homogeneous solutions, at the interface a different mechanism obviously plays a dominant role.

2.2.6. Selective Electrochemical Reactions at Working Electrodes (Voltammetry and Amperometry)

Depending on the potential range at the working electrode, as many as about ten different electrode processes can be separated voltammetrically based on the typical resolution of a polarographic curve, which requires a difference voltage of about 200 mV for the half-wave potential and a concentration-dependent limiting current. Selectivity in the process is therefore rather limited, and depends on such thermodynamic parameters as the reaction enthalpy for reduction or oxidation at the electrode surface. The potential of the working electrode determines whether a particular compound (analyte) can be oxidized or reduced, and compounds with similar redox characteristics might interfere.

The problem of interference is of course greatest when extremely positive or negative working electrode potentials are applied at the working electrode, since at high potential the small differences between electrochemical reaction rates become negligible. The "resolution" of a typical current – voltage curve is diminished if the electrochemical reaction is kinetically hindered, resulting in so- called irreversible current steps that are less steep. Irreversible electroactive species require a greater overpotential before the diffusion- controlled limiting current plateau is reached, therefore increasing the chance that an interfering compound will become electrochemically active. The limited resolution power of voltammetry can be improved by the use of chemically modified electrode surfaces. If a surface layer or a thin membrane is introduced between the phase boundary this additional layer will increase the selectivity if only the analyte can pass through this layer.

In certain biosensors an irreversible heterogeneous reaction is transformed into a homogeneous redox reaction via a very reversible redox

system. The latter can be transformed back at the working electrode by a much lower overpotential, thereby reducing the chance of co-oxidation or co-reduction of interfering compounds. Since the limiting current is strictly proportional to the analyte concentration and also to the concentrations of interfering compounds, chemometric data treatment and/or differential measurements may be used to correct for errors.

2.2.7. Molecular Recognition Processes Based on Molecular Biological Principles

Metabolic biosensors are based on special enzyme- catalyzed reactions of the analyte. The recognition process, also known as the lock-and-key principle, is extremely selective, permitting differentiation even between chiral isomers of a molecule. Recognition in this case demands a perfect geometrical fit as well as the appropriate dipole and/or charge distribution to permit binding of the analyte molecule inside the generally much larger biomolecule (host). Since fitting is associated with all three dimensions, a very large biomolecule with its stabilized structure is more effective than the smaller host molecules commonly encountered in supramolecular chemistry. The selectivity of a specific biocatalytic reaction can be further improved by ensuring detection of only the transformed analyte or a stoichiometric partner molecule through a chemical transducer located behind the recognition layer. In a certain sense, this type of biosensor represents the highest possible level of selectivity, and comparable results would be difficult to achieve by other means.

Immunosensors based on mono- or polyclonal antibodies constitute the second best choice with respect to high selectivity. Depending on the amount of effort expended in screening and choosing the desired antibodies from among the millions of different antibody molecules available, the selectivity can be remarkably high. On the other hand, it is also possible to choose antibody molecules that bind only to a certain region (epitope) of the analyte molecule, leading to biomolecules capable of recognizing all members of some class of compounds sharing similar molecular structures. This type of sensor, though it cannot be calibrated, is very valuable for screening purposes used to detect the presence of an only certain class of substance . The application of antibody molecules has been considerably simplified by the large-scale preparation of monoclonal versions with precisely identical features. It has even become possible to construct immunosensor arrays based on monoclonal antibody molecules with exactly matched selectivities. Results obtained from such arrays must be evaluated by modern methods of pattern recognition.

In order to overcome the limited stability of large protein molecules, research aims isolate only the binding region of the F_{ab} part (antibody-binding fragment) of an antibody molecule. Smaller molecules are of course likely to display somewhat more limited recognition ability, however, with selectivities approaching those of supramolecular host molecules synthesized in the traditional way. Research has also focussed on recombinant antibodies produced from large protein libraries without the need for any animal immunization.

2.3. Transducers for Molecular Recognition: Processes and Sensitivities

2.3.1. Electrochemical Sensors

Electrochemical sensors constitute the largest and oldest group of chemical sensors. Although many such devices have reached commercial maturity, others remain in various stages of development. Electrochemical sensors will be discussed here within the broadest possible framework, with electrochemistry interpreted simply as any interaction involving both electricity and chemistry. Sensors as diverse as enzyme electrodes, high-temperature oxide sensors, fuel cells, and surface- conductivity sensors will be included. Such sensors can be subdivided based on their mode of operation into three categories: *potentiometric* (measurement of voltage), *amperometric* (measurement of current), and *conductometric* (measurement of conductivity) [13], [41].

Electrochemistry implies the transfer of charge from an electrode to some other phase, which may be a solid or a liquid. During this process chemical changes take place at the electrodes, and charge is conducted through the

bulk of the sample phase. Both the electrode reactions themselves and/or the charge transport phenomenon can be modulated chemically to serve as the basis for a sensing process. Certain basic rules apply to all electrochemical sensors, the cardinal one being the requirement of a closed electrical circuit. This means that an electrochemical cell must consist of at least two electrodes, which can be regarded from a purely electrical point of view as a sensor electrode and a signal return.

Another important general characteristic of electrochemical sensors is that charge transport within the transducer portion of the sensor and/or inside the supporting instrumentation (which constitutes part of the overall circuit) is always electronic. On the other hand, charge transport in the sample under investigation may be electronic, ionic, or mixed. In the latter two cases electron transfer takes place at the electrode – sample interface, perhaps accompanied by electrolysis, and the corresponding mechanism becomes one of the most critical aspects of sensor performance.

The overall current – voltage relationship is complex, and it can vary as the conditions change. The relationship is affected primarily by the nature and concentration of the electroactive species, the electrode material, and the mode of mass transport. A total observed current can be analyzed in terms of its cathodic and anodic components. If the two currents are equal in magnitude but opposite in sign, there must be an *exchange current* passing through the electrode-surface/sample interface.

Both the charge-transfer resistance and the exchange-current density are critically important parameters in the operation of most electrochemical sensors, since they reflect the kinetics of an electrode reaction. These parameters are directly proportional to the electrode area, so the smaller the electrode the higher will be the resistance, all other parameters remaining unchanged. Therefore, it is the electrochemical process taking place at the smaller electrode that determines the absolute value of the current flowing through the entire circuit. The auxiliary electrode will begin to interfere only if its charge-transfer resistance becomes comparable to that of the working electrode. Microelectrodes represent one approach to avoiding such interferences (see Section 2.3.1.2).

On the other hand, in zero-current potentiometry the relative sizes of the two electrodes is immaterial. Acquiring useful information in this case requires only that the potential of the working electrode be measured against a well-defined and stable potential from a reference electrode. Any foreign potential inadvertently present within the measuring circuit can contaminate the information, so it is mandatory that the reference electrode be placed as close to the working electrode as is practically possible. Thus, in an amperometric (or conductometric) measurement the source of information can be localized by choosing a small working electrode, whereas it cannot be localized with zero-current potentiometric measurements.

2.3.1.1. Self-Indicating Potentiometric Electrodes

Fundamental Considerations. Within the context of this chapter "potentiometry" is understood to mean the measurement of potential differences across an indicator electrode and a reference electrode under conditions of zero net electrical current. Such a measurement can be used either for determining an analyte ion directly (direct potentiometry) or for monitoring a titration (see below).

Potentiometry has proven to be well suited to the routine analysis of a great number and variety of analytes [5], [40], [42–60]. Ion-selective electrodes (ISEs) are commercially available for many anions and cations, as indicated in Tables 4, 5, 6, 7. Other analytes can be determined using ISEs in an indirect way. Chapter 3 deals with various types of potentiometric biosensors. A special advantage of ion-selective potentiometry is the possibility of carrying out measurements even in microliter volumes without any loss of analyte.

At present, most ISEs unfortunately do not provide absolute selectivity for a single ion. This fact requires that one have access to very detailed information regarding the nature of samples to be analyzed so that interferences can be either eliminated or otherwise taken into account.

The theory of (ion-selective) potentiometry has already been introduced in Section 2.2.5. Here it is necessary only to remind the reader of the Nernst equation and the Nernst – Nikolsky equation, both of which are

used in the evaluation of potentiometric measurements.

The *Nernst equation* is valid only under ideal conditions (with no interfering ions, etc.):

$$E = E^0 + (RT/zF)\ \ln a_{\mathrm{M}} \tag{5a}$$

$$= E^0 + (0.059/z)\ \log a_{\mathrm{M}} \tag{5b}$$

The *Nernst–Nikolsky equation* on the other hand takes into consideration the influence of interfering ions on the potential of an ISE:

$$E = E^0 + (RT/z_{\mathrm{M}}F)\ \ln\left[a_{\mathrm{M}} + \sum K_{\mathrm{MI}}\,(a_{\mathrm{I}})^{z_{\mathrm{M}}/z_{\mathrm{I}}}\right] \tag{6}$$

where

E = potential of the ISE, measured with zero net current
E^0 = standard potential of the ISE
R = gas constant (8.314 J K^{-1} mol^{-1})
T = absolute temperature (K)
F = faraday constant (96 485 C/mol)
a_{M} = activity of the ion to be measured
a_{I} = activity of an interfering ion
$z_{\mathrm{M}}, z_{\mathrm{I}}$ = electrical charges of the measured and interfering ions
K_{MI} = selectivity coefficient

The value of K_{MI} depends both on the activity a_{M} and on the particular combination of analyte ion and interfering ion. Bakker [48] suggested another equation which described the EMF versus concentration behavior of an ISE with a higher accuracy. The selectivity coefficients are determined here without conditioning the electrode with the measured ion, which yields constant selectivity coefficients. The new equation describes the situation when a monovalent analyte ion I^+ is interfered by a divalent ion J^{2+} especially well, and *vice versa*:

$$E = E_{\mathrm{I}}^0 + \frac{RT}{F}\ \ln\left(\frac{a_{\mathrm{I}}\,(\mathrm{IJ})}{2} + \frac{1}{2}\sqrt{a_{\mathrm{I}}\,(\mathrm{IJ})^2 + 4a_{\mathrm{J}}\,(\mathrm{IJ})\left(K_{\mathrm{IJ}}^{\mathrm{pot}}\right)^2}\right) \tag{6a}$$

for $z_{\mathrm{I}} = 2$ and $z_{\mathrm{J}} = 1$

$$E = E_{\mathrm{I}}^0 + \frac{RT}{F}\ \ln\left(\sqrt{a_{\mathrm{I}}\,(\mathrm{IJ}) + \frac{1}{4}K_{\mathrm{IJ}}^{\mathrm{pot}}a_{\mathrm{I}}\,(\mathrm{IJ})^2} + \sqrt{\frac{1}{4}K_{\mathrm{IJ}}^{\mathrm{pot}}a_{\mathrm{I}}\,(\mathrm{IJ})^2}\right) \tag{6b}$$

Horvai [49] developed another more general equation for ion-selective ion-exchange membranes, which also considers the concentration dependency of the selectivity coefficient:

$$E \cong E_{\mathrm{I}}^0 + \frac{RT}{z_1F}\ln a_1\,(\mathrm{IJ}) + \frac{RT}{z_1F}K_{\mathrm{IJ}}a_{\mathrm{I}}\,(\mathrm{IJ})^{z_{\mathrm{J}}/z_{\mathrm{I}}}\,a_{\mathrm{J}}\,(\mathrm{IJ}) \tag{6c}$$

One can see here, separated from each other, the response of the ISE towards the pure analyte ion I and (as the last term) the EMF deviation caused by the interfering Ion J. Note that there is no logarithm in the last term and also that the last term depends both on the interfering and primary ion concentrations.

Types of Ion-Selective Electrodes. Simple metal-ion electrodes such as those based on silver, gold, platinum, etc., will not be dealt with in this discussion since they are used mainly for indicating purposes in potentiometric titrations. Apart from these, the following types of ISEs can be distinguished:

1) Glass-membrane electrodes
2) Solid-membrane electrodes
3) Liquid-membrane electrodes (including PVC membrane electrodes)
4) ISEs based on semiconductors (ion-selective field-effect transistors, etc.)

These four types will be the subject of more detailed consideration.

Glass-Membrane Electrodes [50]. The best-known glass-membrane electrode is the pH electrode. Its most important component is a thin glass membrane made from a sodium-rich type of glass. Depending on the intended application the membrane may take one of several geometric forms: spherical, conical, or flat. For applications in process streams, special pH electrodes have been devised that are resistant to high pressures, and sterilizable pH electrodes are available for biotechnological applications.

Examples of the different types of pH electrodes are illustrated in Figure 11. The *spherical* type is most frequently used for the direct measurement of pH or for acid–base monitoring. It is robust and appropriate for most routine applications. *Conical* electrodes can be used as "stick-in electrodes" for pH measurements in

meat, bread, cheese, etc. A conical membrane can easily be cleaned, which is important if measurements are to be made in highly viscous or turbid media. A *flat* membrane facilitates pH measurements on surfaces, such as on human skin.

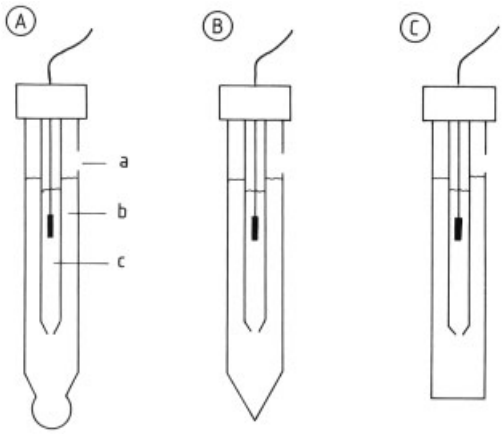

Figure 11. Various types of glass-membrane pH electrodes (with permission from Springer-Verlag, Heidelberg)
A) Spherical membrane; B) Conical membrane; C) Flat membrane
a) Fill hole; b) Inner filling solution (buffer); c) Reference system

The pH electrode must be combined with a reference electrode [51] that provides a constant potential for completing the electrical circuit. A suitable reference system is frequently integrated into the body of the pH electrode itself. Combined electrodes of this type are particularly easy to handle. pH Measurements at > pH 10 by means of glass electrodes are subject to interference by alkalimetal ions, especially Na^+. The selectivity coefficient K_{MI} for interference by the sodium ion is 10^{-13} in special lithium containing glasses with a reduced alkali error. This means that a 10^{13}-fold concentration of sodium is required to produce the same potential as a given concentration of protons.

It is important that the surface of a glass membrane be swollen before measurements are taken. Prior to first use the electrode should therefore be allowed to swell overnight in a 3 mol/L KCl solution. For the same reason the electrodes should also be stored in a solution of this type between measurements and after measurements have been conducted in nonaqueous solvents. Here, the choice of a double junction reference electrode with the same solvent as in the measuring solution and added organic salt for sufficient conductivity is important. In the case of endpoint detection a platinum/ferrocene redox element without a liquid junction could also be used.

Apart from the pH-selective glass-membrane electrodes, other commercially available glass-membrane electrodes are responsive to sodium ions, although the selectivity of a pNa electrode is much lower than that of a pH electrode. Approximate selectivity coefficients K_{MI} in this case are 10^3 for interference by H^+ and 10^{-3} for interference by Li^+, K^+, and Cs^+. However, an interference by H^+ ions is seldom serious since by appropriate pH adjustment one can easily shift the H^+ ion concentration to a range where it does not influence the electrode potential.

Solid-Membrane Electrodes. The sensing element of a solid-membrane electrode consists of a material showing ionic conductivity for the particular ions that are to be determined. Such a device may be manufactured from a single crystal, as in the well-known fluoride-selective electrode, where the sensing element is a single crystal of lanthanum fluoride doped with europium(II) fluoride to lower the electrical resistance. Alternatively, a membrane can be produced by grinding the crystalline sensor material together with an appropriate additive (PTFE powder, silicone rubber), followed by solidification either by application of high pressure or by addition of a cross-linking agent. It is thus essential that the crystal particles stay in contact which each other. In addition to these so-called heterogeneous membranes, the homogeneous membranes based on silver sulfide are better known. They consist of Ag_2S which is freshly precipitated, washed, dried, and pressed into a pellet with a traditional KBr-IR press. Since Ag_2S is rarely stoichiometric it shows a high Ag^+ ion conductivity leading to a perfect Nernstian-sensor for that ion. Because the standard exchange current density at its suface is much higher than on a silver metal surface (overpotential due to crystallization polarization) it senses lower Ag^+ ion concentrations much better. By incorporating about 30 weight % of a certain silver halide into this pellet one obtains the corresponding halide sensitive membrane. Note, the same result would be obtained if the appro-

priate silver halide were to be added as a powder to the measuring solution. That means that it is still the Ag^+ ion which determines the potential via the corresponding solubility product. If certain metal sulfides such as CuS, PbS, or CdS are added the membrane senses those metal ions via two solubility products and an intermediate S^{2-} ion-concentration. Thus, Ag^+ ions must be totally absent as well as compounds oxidize S^{2-}.

The membrane is then fixed to the electrode body. Electrical contact is established either by a conducting adhesive or by an internal reference system. Figure 12 illustrates schematically the construction of two types of solid-membrane electrodes. A detailed compilation of commercially available solid-state electrodes is provided in Tables 4, 5, 6, 7 at the end of this section.

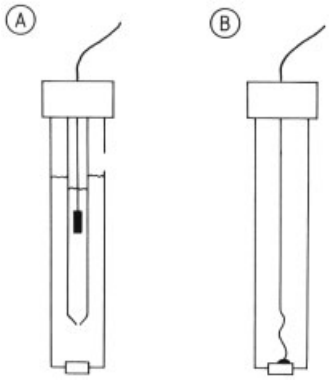

Figure 12. Two types of solid-membrane electrodes (with permission from Springer-Verlag, Heidelberg)
A) With internal filling solution and reference system; B) With solid contact

Generally speaking, interference must be anticipated from any ion that tends to react with the membrane material to form compounds less soluble than the membrane itself. Thus, the chloride electrode, which consists of a $AgCl/Ag_2S$ membrane, is poisoned by mercury(II) ions. Due to the solid state of the membrane the reaction rate is rather low, so a poisoned membrane can be repaired by carefully polishing it with an appropriate polishing powder (as described in instruction manuals furnished with the electrodes).

Liquid-Membrane Electrodes. The electroactive phase in this type of electrode consists of a liquid that has been stabilized mechanically. For this purpose the electroactive components may be dissolved in a highly viscous and apolar organic solvent such as one of the higher alcohols or paraffin. A porous material (e.g., commercial membrane filters such as Sartorius cellulose nitrate with pore diameters of about 0.45 µm), are initially made hydrophobic with, e.g., hexamethyl disilazane and then soaked with this solution and introduced in place of the solid membrane in an electrode of the type described above. More frequently, however, the liquid is not fixed in a porous material, but rather embedded into a high molecular mass polymer like PVC. In this case the plasticizer (normally present at a concentration of about 66 % by weight) acts as a solvent for the electroactive component (the concentration, which seldom exceeds 1 %!). Simplified construction schemes for these two types of electrodes are illustrated in Figure 13.

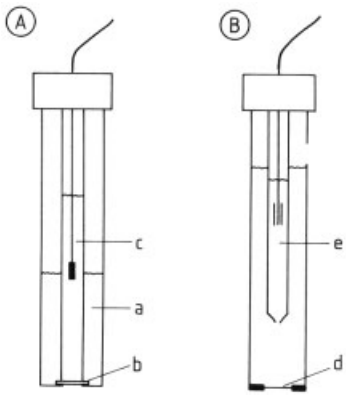

Figure 13. Two types of liquid-membrane electrodes (with permission from Springer-Verlag, Heidelberg)
A) Porous-membrane electrode; B) PVC-membrane electrode
a) Reservoir for liquid ion carrier; b) Porous membrane; c) Reference system; d) PVC membrane; e) Reference system

For applications like those discussed in Sections 2.2.4 and 2.2.5 the electroactive component of a liquid-membrane electrode would consist of a (charged) ion exchanger or an (uncharged) ion carrier. In both cases the electroactive component must have a rather high affinity for the ion to be determined, but at the same time display the greatest possible selectivity with respect to interfering ions. Table 2 lists ion exchangers and ion carriers frequently used in liquid-membrane ISEs. For an excellent and almost exhaustive compilation of ISEs based

on this principle see the reviews of BAKKER, BÜHLMANN, and PRETSCH [59], [60].

Table 2. Example of ion exchangers and ion carriers frequently encountered in liquid-membrane electrodes

Ion	Electroactive compound
Cations	
K^+	valinomycine
NH_4^+	nonactine – monactine
Li^+	lithium carrier, e.g., ETH 1644 (Fluka)
Ca^{2+}	calcium salt of didecylphosphonic acid
Anions	
NO_3^-	Ni(*o*-phen)$_3^{2+}$
ClO_4^-	Fe(*o*-phen)$_3^{2+}$
BF_4^-	Fe(*o*-phen)$_3^{2+}$

The first part of these publications deals with the most recent theoretical treatment while the second part describes in detail nearly all known membrane compositions for nearly every ion of particular importance. It should be noted that PVC-based ISEs are especially easy to construct in the laboratory just before they are to be used. The Fluka company offers a large collection of electroactive compounds (including the most famous ones with their characteristic ETH number), the right PVC and plasticizer and also construction information. Recently reasons for the typical detection limits in the range of μM solutions have been found [61]. If the internal filling solution contains too high a concentration of the measured ion it will be distributed through the entire membrane and can be exchanged with interfering ions at the measuring surface and thereby produce a certain concentration in the sample solution. Thus, in case of trace determinations the inner concentration of the measured ion should be $<$ μM. However, care is needed not to loose the internal measured ion by adsorption processes. Thus, an appropriate ion-buffer leading to a constant low concentration is recommended.

For anion determinations, most PVC-membrane based ISEs follow the so-called Hofmeister series, which is: the more lipophilic anion is always more favored by the organic membrane and therefore sensed better. However, there are exceptions to this general rule which are also most interesting from a theoretical point of view, especially when a supramolecular carrier is used and more than the typical one lipophilic salt addition is performed. Because of this, owing to the incorporation of a lipophilic cation and anion, a carbonate selective ISE could be constructed with decreased interference from salicylate by five orders of magnitude [62]. This addition surprisingly increases the charge transfer resistance of the ISE so that it can be measured besides the high bulk resistance in series. However, the ratio of the transfer resistances for the analyte and the interfering ions changes too, in favor of the measured ion!

ISEs Based On Semiconductors (*ISFETs*). Devices based on ion-selective field-effect transistors (ISFETs) are sufficiently distinctive to warrant separate treatment in Section 2.3.1.4.

Instrumentation. The discussion here of instrumentation required for potentiometric measurements will be limited to a brief overview of suitable reference electrodes [51]. The conventional reference electrode consists of an electrode of the second kind, like the Ag/AgCl/KCl system. Much effort has been devoted to the development of miniaturized reference systems suited to direct integration into ISFETs, but most systems of this type fail to fulfill all the requirements for a "true" reference electrode with a well-defined and stable potential difference relative to the standard hydrogen electrode. The greatest problem is the short distances needed for the liquid junction, leading to an influx of unwanted compounds within too short a time period. Reference electrodes with gel-stabilized internal electrolytes are produced for applications in process analysis.

Table 3 provides an overview of the most frequently used reference electrodes.

Evaluation of Potentiometric Results. The aim of all methods for the evaluation of potentiometric data is the computation of substrate activities from voltage changes observed between an ISE and a reference electrode. Generally speaking, three different approaches are possible: (1) the direct potentiometric measurement of activities, (2) monitoring during the course of a titration, and (3) flow-injection analysis (FIA).

A determination by *direct potentiometric measurement* is accomplished either by calibrating the electrode with solutions of known concentration, or by using the techniques of standard addition or standard subtraction. Since calibration with standard solutions usually in-

Table 3. Common reference electrodes

Electrode	Potential vs. NHE	Useful temperature range, °C	Interferences
$Hg/Hg_2Cl_2/KCl$ (0.1 mol/L)	+0.334 V		
$Hg/Hg_2Cl_2/KCl$ (1.0 mol/L)	+0.280 V	15 – 70	complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
$Hg/Hg_2Cl_2/KCl$ (satd.)	+0.241 V		complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
Ag/AgCl/KCl (0.1 mol/L)	+0.290 V		complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
Ag/AgCl/KCl (1.0 mol/L)	+0.222 V	15 – 110	complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
Ag/AgCl/KCl (satd.)	+0.197 V		complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
$Hg/Hg_2SO_4/K_2SO_4$ (0.5 mol/L)	+0.682 V	15 – 70	complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
$Hg/Hg_2SO_4/K_2SO_4$ (satd.)	+0.650 V		complexing agents (CN^-), S^{2-}, strong oxidizing or reducing agents
Tl(Hg)/TlCl/KCl (3.5 mol/L)	− 0.575 V	15 – 120	S^{2-}, strong oxidizing agents; no interference from complexing agents

volves solutions of pure salts, thorough knowledge of the samples to be analyzed is a prerequisite to avoiding erroneous results due to interfering ions present in the sample matrix. Influences attributable to other substances present in the sample that do not interfere directly with the determination but produce changes in ionic strength can be minimized by adding a total ionic strength adjusting buffer ("TISAB solution") to both standard solutions and samples. Standard-addition or -subtraction methods should be invoked whenever it is not possible to make appropriate standard solutions (owing to a very complex matrix, for example). Figure 14 illustrates the evaluation of a standard-addition analysis.

Monitoring a titration by means of an ISE has the advantage over direct potentiometric measurement that the accuracy is determined largely by the titration reaction and not simply by the calibration function of the electrode. This is especially important for ISEs with changing or unknown slopes. For this reason the determination of a surfactant should always be carried out as a titration and not as a direct measurement.

Flow-Injection Analyses with ISEs are especially easy to carry out. If the ISE has a flat surface it can be connected to the liquid junction of a reference electrode via a small strip of filter paper. This paper strip is then held by a clamp and the carrier stream of the FIA system is applied just before the measuring ISE, passing over its surface and later over the reference electrode. New levels of ultra trace analysis can be reached – as already mentioned – by never conditioning the ISE membrane with higher concentrations of the measured ion either from the outside or from the inside! This automatically leads to a differential FIA operating mode in which the injected sample segment is first passed over one of the ISE surfaces and then with a matched retarding coil to the other side of the ISE membrane. Figure 15 shows results and the experimental set-up with two external reference electrodes, the ISE membrane and the redarding coil which should direct the sample segment to the second site of the ISE-membrane exactly at the moment it leaves the first side [63]. In ultra trace analysis the carrier solution should contain only traces of the measured ion in order to get a stable voltage reading. Note that the injection volume in Figure 15 is in the µL range. Thus, such a simple and low cost approach allows the determination of pg amounts as it is demonstrated in Figure 16.

Commercially Available ISEs. Tables 4, 5, 6, 7 provide an overview of ISEs commercially available at the time. ISEs described only in the literature, and which must therefore be selfconstructed, have not been included in the tables. Instructions for the preparation of PVC-membrane elecrodes can be found in the literature [4], [5], [15].

Table 4. Commercially available ISEs with selectivity for cations

ISE type	MT [a]	Concentration range, mol/L [b]	Useful pH range [b]	Temp. range, °C [b]	Membrane resistance, MΩ [b,c]	Suppliers [d]	Interfering ions [b,e] in order of decreasing interference; miscellaneous remarks
pNa^+	G	$2 \cdot 10^{-6} - 1$	pH = pNa + 4	0 – 60	100 – 200	CIMORWΩ	H^+, K^+, NH_4^+, Li^+; Ag^+ must be absent
pK^+	P	$10^{-6} - 1$	2 – 12	0 – 40	10 – 20	CIMORWΩ	Rb^+, Cs^+, NH_4^+, Tl^+, H^+, Ag^+, $Tris^+$, Li^+, Na^+; cationic surfactants must be absent
pCa^{2+}	P	$5 \cdot 10^{-7} - 1$	2.5 – 11	0 – 40	1 – 4	CIMORWΩ	Pb^{2+}, Hg^{2+}, H^+, Sr^{2+}, Fe^{2+}, Cu^{2+}, Ni^{2+}, NH_4^+, Na^+, $Tris^+$, Li^+, K^+, Ba^{2+}, Zn^{2+}, Mg^{2+}; cationic surfactants must be absent
pCd^{2+}	S	$10^{-7} - 10^{-1}$	2 – 12	0 – 100	≤ 1	MOWΩ	Cu^{2+}, Hg^{2+}, Ag^+ must be absent; Fe^{2+}/Fe^{3+} interfere if present in concentrations $> c\,[Cd^{2+}]$
pPb^{2+}	S	$10^{-6} - 10^{-1}$	4 – 7	0 – 100	≤ 1	MORWΩ	Cu^{2+}, Hg^{2+}, Ag^+ must be absent; Fe^{2+}/Fe^{3+} and Cd^{2+} interfere if present in concentrations $> c\,[Pb^{2+}]$
pCu^{2+}	S	$10^{-8} - 10^{-1}$	2 – 12	0 – 80	≤ 1	CMORW	Hg^{2+}, Ag^+, S^{2-} must be absent; Cl^-, Br^-, Fe^{3+}, Cd^{2+} interfere if $c > c\,[Cu^{2+}]$; Anions forming complexes with Cu^{2+} (Hal^-, OAc^-) interfere; this ISE is sensitive to light, so all measurements should be carried out in the dark
pAg^+	S	$10^{-7} - 10^{-1}$	2 – 12	0 – 80	≤ 1	CIMORWΩ	Hg_2^{2+}, Hg^{2+}, proteins
pNH_4^+	P	$< 10^{-6} - 1$	4 – 7	0 – 50	≤ 10	IR	cationic surfactants must be absent; interferences from K^+, Rb^+, H^+, Cs^+, Na^+/Li^+, Sr^{2+}, Ba^{2+}, Mg^{2+}

[a–h] See Table 7.

Table 5. Commercially available ISEs with selectivity for anions

ISE type	MT [a]	Concentration range, mol/L [b]	Useful pH range [b]	Temp. range, °C [b]	Membrane resistance MΩ [b,c]	Suppliers [d]	Interfering ions [b,e] in order of decreasing interference; miscellaneous remarks
pF^-	S	10^{-6} – sat. solu.	5 – 11	0 – 100	≈ 0.2/≤1	CIMORWΩ	pH < 5, pH > 11
pCl^-	S	$5 \cdot 10^{-5}$ – 1	2 – 12	0 – 50	≤ 0.1	CIMORWΩ	CN^-, I^-, Br^-, $S_2O_3^{2-}$, NH_3, OH^-, metal ions forming complexes with Cl^-; S^{2-} must be absent
pBr^-	S	$5 \cdot 10^{-6}$ – 1	2 – 12	0 – 50	≤ 0.1	CIMORWΩ	CN^-, I^-, $S_2O_3^{2-}$, NH_3, Cl^-, metal ions forming complexes with Br^-; S^{2-} must be absent
pI^-	S	10^{-7} – >1	2 – 12	0 – 80	≤ 0.1	CIMORWΩ	S^{2-}, CN^-, Br^-, Cl^-, $S_2O_3^{2-}$, metal ions forming complexes with I^-
pCN^-	S	$< 10^{-6}$ – 10^{-2}	11 – 13	0 – 80	≤ 30/≤0.5	CIMORWΩ	S^{2-} must be absent; interferences from I^-, CrO_4^{2-}, $S_2O_3^{2-}$, Br^-, Cl^-, metal ions forming complexes with CN^-
$pSCN^-$	S	$5 \cdot 10^{-6}$ – 1	2 – 12	0 – 50	≤ 0.1	MOW	S^{2-} ≈I^-, Br^-, CN^-, $S_2O_3^{2-}$, Cl^-, OH^-
$pClO_4^-$	P	$7 \cdot 10^{-6}$ – 1		0 – 40		O	not sorted: I^-, NO_3^-, Br^-, ClO_3^-, CN^-, NO_2^-, HCO_3^-, CO_3^{2-}, Cl^-, $H_xPO_4^{x+3-}$, OAc^-, F^-, SO_4^{2-}
pNO_3^-	P	10^{-5} – 1	3 – 12	0 – 50	1	CIMORWΩ	anionic surfactants must be absent; interference from SCN^- ≈MnO_4^-, NO_2^- ≈ClO_4^-, I^-, Br^-, HCO_3^-, F^- ≈SO_4^{2-}
pBF_4^-	P	$7 \cdot 10^{-6}$ – 1	2.5 – 11	0 – 40		MOW	ClO_4^-, I^-, ClO_3^-, CN^-, Br^- ≈NO_2^-, NO_3^-, Cl^-, PO_4^{3-}, OAc^-, F^-, SO_4^{2-}
pS^{2-}	S	10^{-6} – 10^{-1}	12 – 14	0 – 80		CIMORWΩ	Hg_2^{2+}, Hg^{2+}, proteins

[a–h] See Table 7.

Table 6. Commercially available ISEs with selectivity for gases

ISE type	MT[a]	Concentration range, mol/L[b]	Useful pH range[b]	Temp. range, °C[b]	Membrane resistance, MΩ[b,c]	Suppliers[d]	Interfering ions[b,e] in order of decreasing interference; miscellaneous remarks
pNH_3	M	$< 10^{-6} - 5 \cdot 10^{-2}$	alkaline soln.	0–50		IMOWΩ	all types of detergents and wetting agents must be absent; interferences from volatile amines
pCO_2	M	$7 \cdot 10^{-6} - 2 \cdot 10^{-2}$	acidic soln.	0–50		IOΩ	all types of detergents and wetting agents must be absent; interferences from SO_2, NO_x, H_2S
pNO_x	M	$10^{-6} - 5 \cdot 10^{-3}$	acidic soln.	0–50		IO	all types of detergents and wetting agents must be absent; interferences from SO_2, CO_2, volatile carbonic acids
pCl_2	M	$10^{-7} - 3 \cdot 10^{-4}$		0–50		O	oxidizing agents

[a–h] See Table 7.

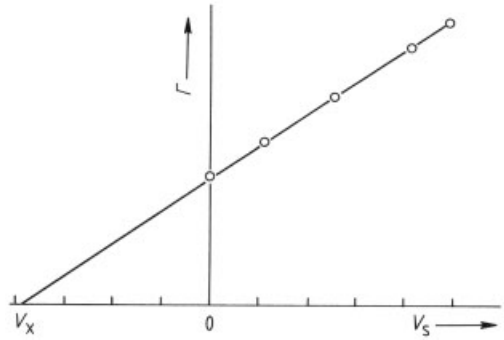

Figure 14. Plot for the evaluation of a standard-addition analysis
V_S = Volume of standard solution; V_x = Volume of analyte
$\Gamma = pE \cdot V_{total}$, with $pE = 10^{\Delta E.(z/S)}$ (ΔE = Potential difference; z = Charge of the ion, including sign; S = Slope of the electrode, including sign)
The unknown concentration of the analyte c_x is $c_x = (-c_s \cdot V_x)/V_{initial}$, where c_s = Concentration of the standard solution and $V_{initial}$ = Initial volume of the analyte

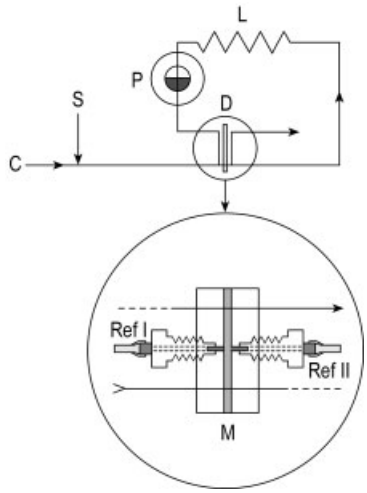

Figure 15. Differential flow-injection analysis with one single ISE membrane and two reference electrodes in an electrolytic closed loop.
An injector block used in liquid chromatography injects about 500 µL of sample into the carrier stream. The volume of the carrier stream loop from one side of the ISE membrane to the opposite side including the volume of both membrane sides is also approximately 500 µL. Thus, when the sample segment is fully washed out of the first membrane side the first front of it reaches the backside. Through this a positive and a negative analyte peak are obtained, which look like a second derivative of a peak.
C = carrier solution; S = sample (introduction with injector loop); P = pump; L = loop adapted to different sample volumes; D = detector; M = ISE membrane with two spacers with a flow channel on each side; Ref I and Ref II = commercial reference electrodes with threads

Table 7. Commercially available ISEs with selectivity for miscellaneous species

ISE type	MT [a]	Concentration range, mol/L [b]	Useful pH range [b]	Temp. range, °C [b]	Membrane resistance, MΩ [b,c]	Suppliers [d]	Interfering ions [b,e] in order of decreasing interference; miscellaneous remarks
Hard [f]	P	$7 \cdot 10^{-6}$–1		0–50		O	not sorted: Na^+, Cu^{2+}, Zn^{2+}, Fe^{2+}, Ni^{2+}, Sr^{2+}, Ba^{2+}, K^+ oppositely charged bulky ions interfere with the titration reaction; high concentrations of nonionic surfactants may interfere
Surf [g]	P	$\approx 10^{-6}$; [h]	1–13	0–40		MO	

[a] MT = Membrane type; G = Glass-membrane electrode; P = PVC-membrane electrode; S = Solid-state membrane electrode; M = gas-permeable membrane electrode. Glass and solid-state membranes are more resistant to certain organic solvents (acetone, methanol, benzene, dioxane, etc.). PVC-membrane electrodes must not be allowed to come into contact with any organic solvent, because this dramatically reduces the lifetime of the electrode.
[b] Mean values as provided by the suppliers.
[c] Only if different suppliers provide very different values for membrane electrical resistance are two values listed in the table.
[d] This list was compiled in Spring 1993 and should not be regarded as complete; C = Ciba Corning; I = Ingold Messtechnik; M = Metrohm; O = Orion; R = Radiometer; W = WTW; Ω = Omega.
[e] A very detailed compilation of selectivity coefficients is available in [23].
[f] Hard: water hardness; i.e., divalent cations.
[g] Surf = ionic surfactants.
[h] This type of ISE should be used only for the monitoring of titrations; the value cited refers to the miminum titratable concentration of an anionic surfactant (pure solution).

2.3.1.2. Voltammetric and Amperometric Cells

Information is obtained with this type of electrochemical sensor from either the combined current/potential – concentration relationship (voltammetry) or from the current – concentration relationship alone (amperometry).

A voltammetric measurement is accomplished by scanning the potential difference across an electrochemical cell containing a working electrode, a reference electrode and usually an auxiliary electrode separated from the sample solution with a diaphragma from one preset value to another and recording the cell current as a function of the applied potential. A curve so generated is known as a *voltammogram.* Amperometric measurements are made by recording the current flow in a cell at a single applied potential.

The essential operational function of a transducer for this purpose is the transfer of one or more electrons to or from the transducer surface (= working electrode). The flow of these electrons is what constitutes the output signal. Voltammetric and amperometric devices are also capable of conferring a degree of selectivity on the overall sensing process by the proper choice of the working electrode potential against a reference electrode.

Furthermore they are relatively simple: in the most elementary case, the transducer (electrochemical cell) consists of nothing more than two electrodes immersed in a suitable electrolyte. A more complex arrangement might involve the use of a three-electrode cell mentioned above, with the advantages that the reference half-cell is not disturbed (polarized) by the current flowing and that the *ixR* drop can be compensated by such a set-up.

In spite of the inherent diagnostic advantages of voltammetry, a transducer based on this technique represents a rather cumbersome approach to sensing, mainly because of the electronic circuitry required to scan the applied potential, the time needed for a single scan, and the evaluation algorithm identifying the correct current peak to be measured. Accordingly, most sensing applications involve cells operating in an amperometric (fixed potential) mode.

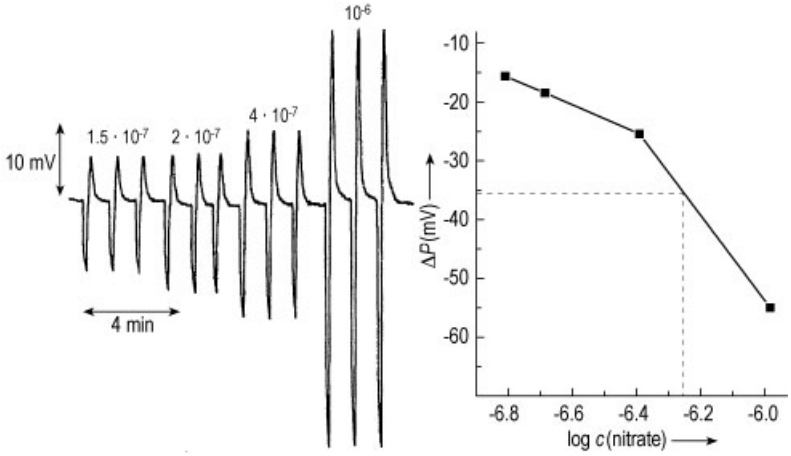

Figure 16. The calibration plot in the lower ppb range shown was constructed with peak-to-peak values resulting in a double Nernstian response. Because of the FIA detection the ISE membrane is never equilibrated with higher concentrations of the analyte ion.
Details: nitrate-selesctive ISE membrane: 1 % Tetradodecylammoniumbromide; 50 mol % of the first is potassium-tetrakis-(4-chlorophenyl)-borat; 49 % 2-.NPOE; 50 % PVC
Carrier solution: 10^{-3} M $Li_2SO_4 + 10^{-7}$ (!) M KNO_3

Voltammetry. When a slowly changing potential is applied to an electrode immersed in an electrolyte solution containing a redox species, a current will be observed to flow as soon as the applied potential reaches a certain value. This current arises from a heterogeneous electron transfer between the electrode and the redox couple, resulting in either oxidation or reduction of the electroactive species. At a sufficiently oxidizing or reducing overpotential the magnitude of the current may become a function of mass transfer of the redox species to the electrode. In a well-stirred solution so-called quasi-stationary current – voltage curves with typical current steps and plateaus are obtained. The half-step (wave) potential is characteristic of the type of species being electrochemically oxidized or reduced, the height of the current steps represents the concentration of it. Without stirring, current peaks are obtained, since the diffusion towards the working electrode surface is no longer sufficient to maintain a constant current. Redox couples that give rise to symmetrical current peaks separated by $58/z$ mV (at 25 °C) with cyclic voltage ramps in unstirred solutions are frequently referred to as *reversible couples.* If the rate of electron transfer between the redox couple and the electrode is high compared with the rate of mass transfer, then the electrode reaction is reversible. Under these circumstances the concentration ratio for oxidized and reduced forms of the couple at the electrode surface is described by the Nernst equation [Equations (5a) and (5b)].

If electron transfer between the redox species and electrode is very slow (kinetical controlled) relative to the mass transport of solution species to the electrode, then the observed current will not be a function of mass transport. In this case, the low rate of electron transfer results in a concentration ratio of the two forms of the redox couple at the electrode surface that no longer conforms to the Nernst equation. Current – voltage curves of these so- called irreversible processes show the corresponding current peaks more to be separated and of different heights and are of limited analytical utility.

An intermediate situation arises when the electron-transfer and mass-transport rates are comparable. Such quasi-reversible electrode reactions are quite common, and their analytical utility depends to a large extent on careful control of the mass-transfer rate in the electroanalytical method used for their study.

Instruments suitable for voltammetry and amperometry consist of three basic components: a wave-form generator, some form of potential control, and an electrochemical cell. Mod-

ern electroanalytical systems employ a three-electrode arrangement for the electrochemical cell. A device called a *potentiostat* is used to maintain a programmed or fixed potential difference between the two current- carrying electrodes (the working electrode and the auxiliary electrode) relative to a third electrode (reference electrode), the function of which is to provide a fixed potential reference in the cell [64], [65].

Both potentiostats and waveform generators have benefited substantially from the introduction of operational amplifiers, as well as from the availability of desktop computers. Electrochemical cell design is now wellestablished both for static systems and for detectors situated in flowing reagent streams.

Despite the fact that solid electrodes have been in use for electroanalytical purposes for some time and are well documented, selection of an appropriate electrode material and its preparation before use remains an area of considerable interest and great controversy among electroanalytical chemists [66]. Solid electrodes for analytical use should always be prepared in a rigorous and reproducible manner. This generally involves polishing the electrode physically with successively finer grades of carborundum or diamond paste. The polished electrode should then be rinsed thoroughly to remove as completely as possible all traces of the polishing materials.

Many different applied potential waveforms have been employed for analytical purposes, but only a few are of mainstream significance [67], [68]. These are presented in Table 8.

In *linear-sweep* voltammetry the current is recorded while the potential of the working electrode is swept from one selected limit to another at a rate between 1 mV/s and 1000 mV/s. In *cyclic* voltammetry, two linear sweeps are recorded, as shown in Table 8. The distinguishing feature of cyclic voltammetry is that electrogenerated species formed in the forward sweep are subject to the reverse electrochemical reaction in the return sweep [69]. Figure 17 shows a typical cyclic voltammetric curve together with the important parameters. In both methods, as the potential is swept to the electroactive region of the redox couple, the current response rises to a peak value before decaying. This decay is caused by depletion of the electroactive species in the zone close to the electrode surface, which means that the diffusion zone spreads out further into the bulk solution.

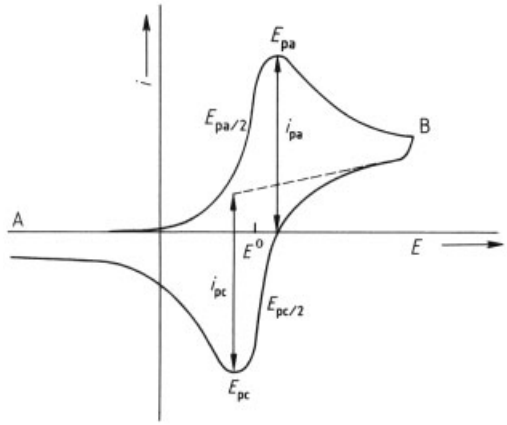

Figure 17. Typical potential vs. current (E vs. i) curve from a cyclic voltammetric experiment with a reversible redox couple
A) Starting potential; B) Reverse potential
E^0 = Standard potential; E_{pa} = Anodic peak potential; E_{pc} = Cathodic peak potential; $E_{pa/2}$ = Anodic half-wave potential; $E_{pc/2}$ = Cathodic half-wave potential; i_{pc} = Cathodic peak current; i_{pa} = Anodic peak current

Pulse and square-wave voltammetry [70] are much better candidates for incorporation into a chemical sensor than linear-sweep or cyclic voltammetry, primarily because of the possibility of discriminating between faradaic and capacitive currents. When a potential pulse is applied to an electrode, the capacitive current that flows is proportional to the magnitude of the pulse, and it decays exponentially with time. The faradaic current, on the other hand, decays according to the square root of time. Figure 18 illustrates the expected decreases in faradaic and capacitive currents, showing that the capacitive current decreases faster than the faradaic one. This gives the opportunity of sampling the current only after the capacitive one can be neglected.

Proper selection of the measuring time permits a signal-to-noise ratio to be improved dramatically. This characteristic is exploited in normal-pulse, differential-pulse, and square-wave voltammetry. In the first of these techniques it represents the only mechanism for decreasing the effect of capacitive current. Further elimination of the capacitive current can be achieved in differential-pulse voltammetry by limiting the duration of the applied pulse and

Table 8. Voltammetric techniques (adapted from [26]); $* \rightarrow$ = Point at which the current is measured; τ = Portion of the staircase during which pulses of width $\tau/2$ are superimposed; E_{sw} = Square-wave amplitude.

Technique	Applied potential*	Measured signal	Comments
Linear-sweep voltammetry	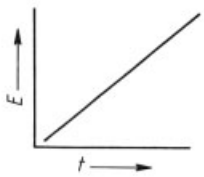	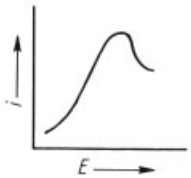	used mainly for diagnostic purposes; current measured throughout the scan
Cyclic voltammetry	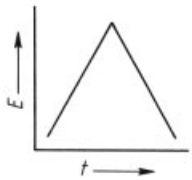	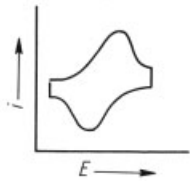	reversal of the potential at the end of the forward scan makes this a powerful diagnostic tool; current is measured throughout the scan
Differential-pulse voltammetry	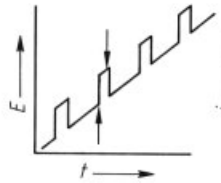	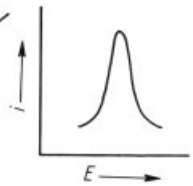	the displayed signal is the difference between the current just prior to and at the end of the applied pulse; low scan rates, but good limits of detection
Square-wave voltammetry	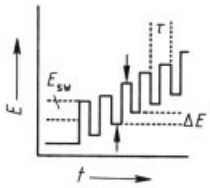	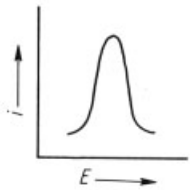	the displayed signal is the difference between currents measured on the forward and reverse pulses; rapid scans and good detection limits

by subtracting the current observed immediately prior to the imposition of the pulse. Increased rejection of the charging current leads to improved detection limits; indeed, this variant of voltammetry permits concentrations as low as $10^{-8} - 10^{-7}$ mol/L to be measured quite readily. In order to avoid problems associated with the electroactive species, the delay between application of successive pulses must be approximately one-half second. This in turn imposes a limit on the scan rate in differential-pulse voltammetry, thereby decreasing its usefulness in sensor applications.

The technique of square-wave voltammetry offers greater promise as a voltammetric method for probing selective chemistry because of the high rate at which the corresponding scan can be executed. The analytical signal in this technique constitutes a difference between the current for the forward pulse and the current for the reverse pulse. Because of the large amplitude of the square wave in a reversible reduction, a reduced electroactive species formed at the electrode during the forward pulse is reoxidized by the reverse pulse. Consequently, the sensitivity of the method is enhanced relative to differential-pulse voltammetry.

This element of speed is crucial to square-wave voltammetry, because, like all voltammetric techniques based on pulse waveforms, the measured current is proportional to $t^{-1/2}$. However, in contrast to the other pulsed voltammetric techniques, square-wave voltammetry causes very little of the depletion that gives rise to distortion of the current – voltage waveform. Accordingly, square-wave voltammetry is uniquely

capable of benefiting from high scan rates. A typical compromise frequency for the waveform is 200 Hz. Coupled with a dE value of 10 mV, this gives rise to a scan rate of 2 V/s. The scan rate is limited by a concomitant increase in the capacitive current. At a solid electrode, the square-wave method is no better than pulsed-voltammetric techniques with respect to rejection of the capacitive current.

Another significant advantage of square-wave technology is the possibility of rejecting a wide range of background currents. With respect to capacitive currents this is achieved by the subtraction of two currents in a manner analogous to differential-pulse voltammetry. Slowly varying capacitive currents that arise when surface groups reorganize on certain types of solid electrodes are also eliminated provided the rate of variance is sufficiently low.

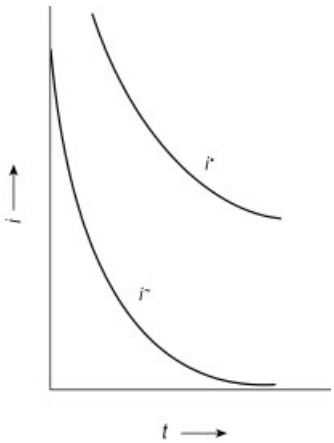

Figure 18. Decrease in faradaic and capacitive current with elapsed time from the imposition of a potential pulse $i^{\bullet}$ = Faradaic current; $i^{\sim}$ = Capacitive current

Amperometry. Amperometry has traditionally been concerned with maintaining a fixed potential between two electrodes, but pulsed techniques have attracted considerable attention as well. The applied potential at which current measurements are made is usually selected to correspond to the mass-transport-limited portion of the corresponding voltammetric scan. Table 9 summarizes the various amperometric methods.

Theoretically, the current obtained in a quiescent solution at a conventional set of electrodes gradually decays to zero according to the Cottrell equation [71]:

$$i_j = nFAC\,(D/\pi t)^{1/2} \tag{7}$$

where t is the elapsed time from application of the potential pulse.

The observed decrease in current is due to a slow spread of the diffusion layer out into the bulk solution, with a concomitant decrease in the concentration gradient. In practice, this process continues for ca. 100 s, after which random convection processes in the solution take over, putting an end to further movement of the diffusion layer. Waiting nearly two minutes to obtain a steady-state current is not a particularly attractive alternative. Accordingly, in amperometric measurements for sensor applications the spread of the diffusion layer is controlled by invoking one or more of the following mechanisms:

1) *Convective Diffusion.* From a practical point of view, convective diffusion can be achieved in two ways: by moving the solution relative to the electrode, or moving the electrode relative to the solution. Of the systems developed to move the electrode, only one merits consideration here, the *rotating disc electrode* (RDE), but there are few possibilities for realizing this electrode in a sensor.
2) *Imposition of a Physical Barrier in the Form of a Membrane.* There are four primary advantages to covering the electrodes of an amperometric device with a membrane permeable only to the analyte: (a) poisoning of the electrodes by electroactive or surface active species is limited; (b) the resolution of the system is enhanced, because extraneous electroactive species that might otherwise undergo electron transfer at the electrode are excluded; (c) the composition of the electrolyte occupying the space between the membrane and the electrodes remains constant; and (d) the membrane forms a physical barrier to prevent the diffusion layer from spreading into the bulk solution.
The limiting current obtained at a membrane-covered amperometric device is a function of time, reaching in due course a steady-state value. The current – time transient cannot be described by a single equation, because different transport mechanisms operate during this time interval. For a typical membrane-covered amperometric oxygen detector with an electrolyte thickness of

Table 9. Amperometric methods (adapted from [26])

Method	Applied potential	Measured signal	Comments
Chronoamperometry	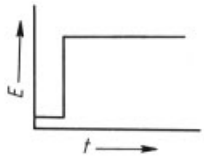	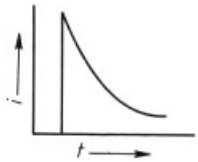	the electrode potential is pulsed to a region in which the analyte is electroactive; a decaying current reflects the growth of the diffusion layer
Pulsed amperometry	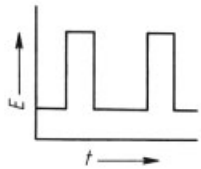	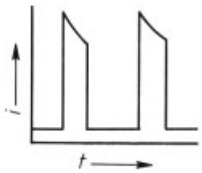	the electrode potential is pulsed briefly to a region in which the analyte is electroactive; between pulses the diffusion layer may be eliminated by forced or natural convection
Pulsed amperometric detection	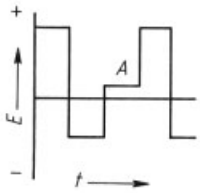	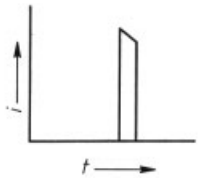	electrode conditioning, analyte sorption, and catalytic electrooxidation are all promoted by the use of this waveform; the current is measured only during the last part of the cycle (region *A*)

10 µm, a membrane thickness of 20 µm, an electrode radius > 2 mm, and a diffusion coefficient for the analyte two orders of magnitude lower in the membrane than in the electrolyte solution, it can be shown that the current – time transient can be described by three equations. For times > 20 s after the imposition of the potential pulse to the electroactive region, the limiting current is given by

$$i_j = nFAC\,(P_m/d) \tag{8}$$

where P_m is the permeability coefficient of the membrane and d is its thickness.

An important consequence of this equation is that the limiting current is a function of the permeability of the membrane. Unfortunately, membrane properties such as permeability are themselves a function of time and ambient conditions, particularly temperature. Accordingly, there are advantages to operating membrane- covered detectors away from this steady-state condition. For a typical detector, the limiting current < 100 ms after imposition of the measurement potential pulse is given by

$$i_j = nFAC\,(K_b/K_0)\,(D_e/\pi t)^{1/2} \tag{9}$$

where t is elapsed time since imposition of the measuring potential, D_e is the diffusion coefficient for the analyte in the electrolyte, and K_b is the distribution coefficient for the analyte at the internal electrolyte – membrane interface. The ratio K_b/K_0 expresses the salting-out effect of the electrolyte solution.

In effect this equation describes the diffusion-limited current arising within the electrolyte layer. In other words, the diffusion layer has not had sufficient time to spread to the membrane, and so the limiting current is completely independent of the membrane. This has the advantage that alterations in membrane characteristics, as well as alterations in ambient conditions that might affect membrane performance, have no effect whatsoever on the limiting current. One of the factors to which this applies is stirring. Under steady-state conditions, stirring of some type is necessary to prevent changes that would otherwise be caused by the diffusion layer moving out into the bulk solution,

although the rate of such stirring is not critical. When transient measurements are made, no stirring whatsoever is required. The true situation regarding transient measurements with amperometric devices is actually not quite so simple as depicted here; for example, there is also a need to take into account contributions from the capacitive current.

The classic amperometric membrane-covered device is the Clark oxygen sensor [72], which consists of a platinum working electrode, an Ag/AgCl reference electrode, a KCl electrolyte solution, and an oxygen-permeable membrane made of teflon or silicone rubber. Figure 19 is a schematic diagram of such a sensor. Dissolved oxygen passes through the permeable membrane, after which it is reduced at the working electrode. The current flow is proportional to the amount of oxygen present in the solution. The following reactions take place at the surface of the working electrode:

$$O_2 + 2\,H_2O + 2\,e \longrightarrow [H_2O_2] + 2\,OH^-$$

$$[H_2O_2] + 2\,e \longrightarrow 2\,OH^-$$

A separate polarographic wave corresponding to the reduction of hydrogen peroxide is observed only with a mercury cathode, which produces two waves for the reduction of oxygen. At other electrodes (Pt, Au, C) only one fourelectron wave is obtained, corresponding to complete reduction to four hydroxide ions [73]. The potential E of the working electrode (at the diffusion-limiting current plateau) usually lies between -0.6 and -0.9 V vs. Ag/AgCl.

Progress in semiconductor technology during recent years has made it possible to miniaturize the classical oxygen sensor [74–77] for use as a transducer in several chemical and biochemical sensors.

3) *Pulsing the Electrode from a Region of Electroinactivity to a Region of Transport-Controlled Electroactivity, after which the Potential is Restored to the Electroinactive Region.* For electrochemical sensors without membranes, applying a potential pulse from a region of electroinactivity to a region of electroactivity has several advantages. For example, the duty cycle of the pulse can be arranged in such a way that the electrode spends the majority of its time at the electroinactive potential. Under these circumstances electrode reactions give rise to products that poison the electrode, but these products are of minimal significance because they are produced at a very low rate. Moreover, natural processes of convection are able to redistribute the electroactive species during the time that the electrode is at the rest potential, thus limiting the spread of the diffusion layer into the bulk solution. The drawback to this technique is the appearance of a capacitive current. Limiting the size of the pulse decreases the contribution of this current, but a small pulse from a rest potential to the electroactive potential is achievable only for reversible electrode reactions. The alternative is to allow sufficient time for the capacitive current to decay away between imposition of the potential pulse and measurement of the current.

4) *Pulsed Amperometric Detection* [78]. This is a relatively new technique that has made possible the direct amperometric determination of many compounds formerly considered unsuitable for analysis in this way. The corresponding waveform is illustrated in Table 9. This rather complex waveform serves several purposes. On platinum, for example, and at the extreme anodic potential, surface-confined oxidation products that tend to poison the electrode are desorbed. This process occurs concurrently with the formation of PtO_2 on the surface. At the cathodic potential the surface oxide is subsequently reduced, and a clean platinum surface is regenerated. In some cases, sorption of the analyte occurs at this cathodic potential, so the analytical signal is recorded at an intermediate potential. Several different mechanisms can give rise to this signal. Electroactive species that are not adsorbed at the electrode may undergo conventional oxidative electrode reactions. Alternatively, electroactive species that are adsorbed at the electrode during the cathodic swing of the cleaning cycle may undergo either conventional oxidative desorption or oxidative desorption catalyzed by an oxide species on the surface of the electrode. For many compounds the latter reaction is much more favorable from a thermodynamic

point of view than oxidation in the absence of a catalyst. This type of reaction occurs most readily when both the analyte and the oxide are confined to the surface. Finally, adsorbed electroinactive analytes may also be detected by virtue of their ability to block the electrode sites at which the PtOH formation occurs. Consequently, any faradaic current arising from the oxidation of platinum is diminished in the presence of adsorbed electroinactive analyte.

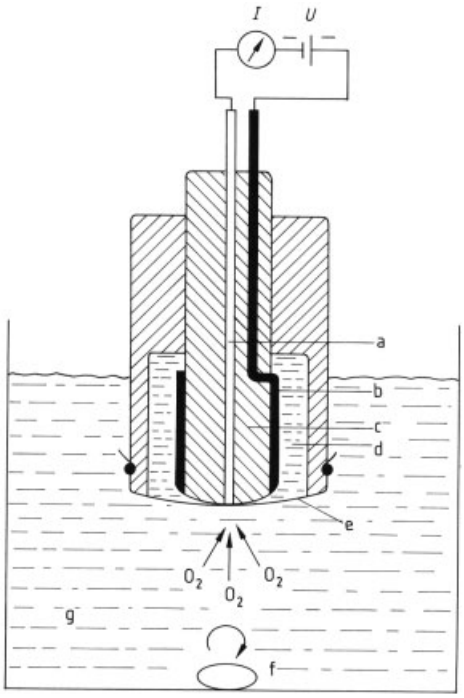

Figure 19. Schematic diagram of a Clark-type oxygen sensor
a) Cathode; b) Anode; c) Insulator; d) Electrolyte; e) Membrane; f) Stirrer; g) Sample for analysis

5) *Microelectrodes.* Electroanalytical applications using microelectrodes have become more common in recent years [79–83] as silicon technology has opened new possibilities for constructing electrodes with micrometer and submicrometer dimensions [84], [85]. The rate of electrolysis at such electrodes is approximately equivalent to the rate of diffusion. Linear-sweep voltammetry in this case gives rise to the same S-shaped current – voltage curves observed with rotating disc or dropping mercury electrodes. The time-dependent current arising at a microelectrode is given by the equation for the chronoamperometric current at a conventional electrode under spherical diffusion conditions:

$$i_t = nFACD\left[\left\{1/(tD\pi)^{1/2}\right\} + \{1/r\}\right] \quad (10)$$

where r is the electrode radius. Indeed, a sphere or hemisphere represents a good model for the diffusion zone that surrounds such an electrode, and because of enhanced mass transport to the electrode, a steady-state current is achieved very rapidly once a potential pulse has been applied to the electrode. Although the diffusion zone is large in this case relative to the electrode surface, it is small compared to the zone surrounding a conventional electrode. This small size means that the faradaic current associated with a microelectrode is relatively immune to the effects of convection in the bulk solution, and the current in a flowing stream is independent of the rate of flow. Decreased effects due to capacitance and resistance at microelectrodes, coupled with high mass-transport rates, make it possible for electrochemical measurements to be obtained in this way in cells containing highly resistive solutions. In fact, provided sufficient free carrier is available to charge the double layer, microelectrodes can be used for direct electrochemical measurements in the gas phase [86]. One drawback of employing working electrodes with diameters less than about 5 μm are the small currents which will flow through such small interfaces. Current measurements in the pA range are prone to electromagnetic interferences (e.g., AC pick-up) and thus, the electric noise level can be a limiting factor. This problem can be solved by employing an array of microelectrodes all at the same potential and separated from each other by about ten times their diameter. Such arrays have been constructed and have shown all the benefits of single microelectrodes [87], [88].

2.3.1.3. Conductance Devices

Conductance is used as a sensor signal in two different areas of sensor technology. On one hand it

is possible to determine the electrolytic conductance of electrolytes, which permits conclusions to be reached on a wide range of electrolyte properties. Moreover, partial pressures of gaseous compounds can be determined in the gas phase by examining their influence on the electronic conductance of such materials as semiconducting oxides [89–92].

Conductometric Sensors for Monitoring Electrolytic Conductance. *Introduction.* Electrolytic conductance serves as a useful signal of the electrical conductivity of mainly aqueous electrolyte solutions, and *conductometry* has been developed into an electrochemical analytical method based on measuring the conductance in electrolyte solutions. All ions in such a solution act as current carriers as a function of their mobility, charge, and concentration. For this reason conductivity measurements are inherently nonselective with respect to any system to which a changing voltage difference has been applied and the diffusion current is measured. Despite this disadvantage, conductivity measurements are indispensable sources of information, especially for monitoring the chemical purity of water samples. In pollution control, conductometric data provide a reliable measure of the total content of ionic pollutants, which is sometimes all that is required. Advantages of conductivity detectors include their universal applicability, simplicity, low price, and wide effective concentration range. Conductivity measurement is one of the most commonly used electrochemical techniques in the control of industrial chemical processes. Conductivity detectors are also employed in ion chromatography and HPLC as a way of detecting separated ions in the eluate from a column [93]. Furthermore, conductometric analyzers are used in the monitoring of atmospheric pollutants such as acidic or alkaline gases (e.g., HCl, CO_2, SO_2, SO_3, NH_3) after these have been transferred as ions into aqueous solution [94], [95]. Conductivity measurements are nondestructive with respect to the sample, and they are always advantageous when electrolytes must be determined in a medium of low self- conductivity.

Definitions. The electrolytic conductivity σ of a solution is defined as the conductance G of the electrolyte between two electrodes, each with an area A, separated by a distance d. The conductivity σ is then given by

$$\sigma = G \cdot d/A \tag{11}$$

The traditional unit of electrolytic conductivity, Siemens per centimeter (S/cm), has been replaced by S/m [96], but the new unit has not yet been accepted widely by the instrument-producing industry, nor by the users of conductivity measurement devices. The English-language literature has historically made use of the "mho" (reciprocal ohm), which is another way of expressing the conductance G. The quotient d/A (cm^{-1}) is defined as the cell constant k, and depends on the geometry of the cell. The electric field is never homogeneous over the cross section of a real cell, so it is not in fact possible to calculate a cell constant k directly from geometric dimensions; it must instead be determined with calibration solutions of known conductivity [97].

The temperature dependence of σ can be deduced from ionic mobilities. This dependence is expressed in terms of the temperature coefficient $\alpha = (d\sigma/dT)(1/\sigma)$, which generally has a positive value between 1 and 3 % per degree Celsius. Knowledge of α in turn makes it possible to calculate conductivities at temperatures other than the temperature of measurement. Whenever the need arises to determine the conductivity of a solution with a very low electrolyte concentration it is essential also to take into account the self-dissociation of pure water, which has a conductivity $\sigma_{25} = 0.05483$ µS/cm.

Measurement Principles. The equipment used for measuring conductance usually consists of a conductivity sensor, a measuring attachment, and a temperature- compensation unit. Conductance values are obtained by establishing either the current flow between a set of electrodes subjected to a constant AC voltage, or the current induced in the secondary coil of a transformer connected to the primary coil by the electrolyte of interest. Capacitive coupled cells have been developed as well, but because of their expensive construction they are no longer of interest [98–101].

Two-Electrode Cells. Two-electrode cells are equipped with a pair of electrodes across which both the voltage and the current can be measured. Use of an AC voltage source is a convenient way of avoiding polarization effects at the

electrode/electrolyte interface and electrolyte decomposition. Figure 20 illustrates a typical construction together with the corresponding equivalent circuit. The advantage of this design is the simplicity of both the cell and the measuring attachment.

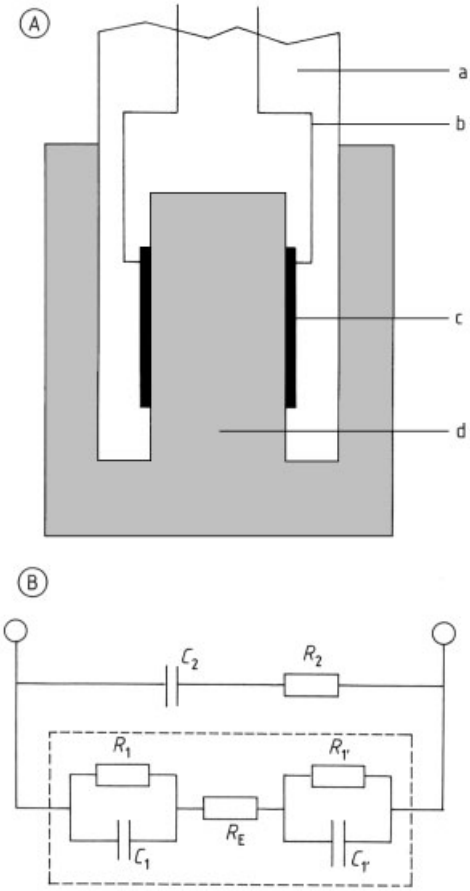

Figure 20. Depiction of a two-electrode cell (A) together with the corresponding equivalent circuit (B)
The dashed rectangle includes all network components involving the electrolyte in contact with the electrodes.
a) Cell body; b) Leads; c) Electrode; d) Sample
R_E = Electrolyte resistance; R_1, $R_{1'}$ = Charge-transfer resistances at the interface electrode – electrolyte; C_1, $C_{1'}$ = Capacitance of the electrochemical double layer; R_2, C_2 = Shunt circuit in the cell

Disadvantages are associated with the need for measuring both the current and the voltage at a single electrode, which may lead to electrode polarization and a measured conductivity lower than the true value. The resistance contributed by a cell in contact with a sample must also be taken into consideration in conjunction with other resistance and capacitance values. At high conductivities ($\sigma > 1$ mS/cm) the resistance of the connecting wires is another significant factor, and electrode reactions can cause polarization of the electrode – electrolyte interface. These problems can be overcome most effectively with an AC voltage oscillating at 1 – 5 kHz and electrode materials with minimum polarizability due to a large surface area, such as platinum covered with a layer of platinum black. At low conductivities ($\sigma < 10$ µS/cm) measured conductivity values may be too high because of capacitive – ohmic shunts caused by the cell body or the connecting wires ("Parker effect"). Contamination by adsorbed ions is also a problem with porous electrodes. The measurement frequency should therefore be kept as low as possible, and the electrodes should be smooth. Resistive precipitates at the electrode surface are sometimes a cause of artificially low conductivity.

For the reasons mentioned above, the "cell constant" is in fact constant only over a limited conductivity range. Suitable values for the cell constant depend on the anticipated level of conductivity:

Expected conductivity	Cell constant
< 0.1 µS/cm – 1 mS/cm	0.1 cm^{-1}
1 µS/cm – 100 mS/cm	1.0 cm^{-1}
10 µS/cm – 1 S/cm	10 cm^{-1}

Four-Electrode Cells. The four-electrode cell was developed as a way of overcoming the problems inherent in two-electrode cells. Two current electrodes in this case generate an electric field, and two separate voltage electrodes are provided for measuring the voltage drop with the aid of a high-impedance amplifier. This design is insensitive to polarization effects, and it also avoids interference from insulating films or precipitates because these are formed outside the zone of measurement. The result is a single system that can be used to measure any conductivity > 1 µS/cm.

Inductively Coupled Cells. The principle underlying inductively coupled conductivity cells is very simple. Two coils of a transformer are shielded from each other magnetically such that the only coupling loop is formed by an insulating tube filled with the sample. An AC voltage in the primary coil induces an AC current in the sample, the magnitude of which depends on the dimensions of the loop and the conductivity of the sample. The voltage induced in the second coil

depends only on the input voltage of the primary coil and the conductance of the sample. Typical working frequencies are 80 – 200 Hz. An advantage of these cells is that they permit contactless measurement. Very high conductivities (>1 S/cm) can be measured without polarization errors because of the contactless decoupling of the signal. Even thin resistive films within the tube fail to falsify the measured conductivity. Another advantage associated with cells of this type is the possibility of encapsulating the entire device in a polymer offering high chemical resistance. Inductively coupled conductivity cells are thus of special interest for measurements in highly conducting and aggressive media [102].

Ion-Selective Conductometric Microsensors. The greatest problem in miniaturizing potentiometry was the lack of a longer working reference electrode. The development of ion-selective conductometric microsensors mainly for cations such as H^+, Li^+, Na^+, K^+, NH^{4+}, Ca^{2+}, Cd^{2+}, Pb^{2+}, Hg^{2+}, and Ag^+ thus far in the ICB Muenster overcomes this problem [103]. In principle, they are partly based on the ion-selective optodes introduced by the late Prof. Simon who showed that the measured ion can also be selectively extracted into an ionophore-based membrane. In order to maintain electroneutrality in the bulk of the membrane a co-immobilized indicator dye looses a proton which passes into the sample solution and thereby changes the color. The influx of the measured cation, which will be complexed by the neutral ionophore and leads to an even larger cation complex in the PVC membrane, is electrically compensated by an outflux of protons that dissociate from a neutral lipohilic acid. This exchange of protons from the membrane against the measured cation will cause the membrane conductivity to change. This can easily be followed by a planar interdigital electrode array onto which a corresponding membrane is drop-coated. In order to keep the lipophilic acid in the membrane in its neutral state and in order to control the interfacial ion-exchange processes in a reproducible way, all measurements must be performed in a buffer of an appropriately chosen and constant pH. Figure 21 shows the measuring set-up and Figure 22 some results of a valinomycin based potassium-selective membrane. The length A of the interdigital electrode fingers is only about 1 mm and α the width of a finger and spacing lies in the region of about 10 µm. The membrane thickness is δ in the range 0.5 – 5 µm. The membrane cocktail – similar to the one in case of ISE membranes – can be brought onto the interdigital electrode array by dip- or drop-coating. U_{in} an alternating voltage in the range 10 – 100 mV and is applied via the corresponding output of a lock-in amplifier. U_{out} at the load resistor is fed to the input of the lock-in amplifier. The in-phase component of the complex impedance is measured and evaluated. In general, the characteristics of those bulk conductive microsensors resemble those of the corresponding ISE membranes with slight variations in the selectivity pattern. The advantage is that the accuracy of their measurements does not drop with increasing charge of the measured cation. With a membrane thickness in the region of 1 µm the response time is in the second range. A difference compared with ISE membranes lies in the increased concentration of ionophore that is added up to about 5 % by weight. These ion-selective conductometric microsensors do not need any reference electrode and can be scaled down to the µm range. The lifetime depends on the lipophilicity of the ionophore and plasticizer. Normally they work for up to several weeks and show a relatively small drift. The further advantages are that they are easily produced by CMOS technology and open also the way to the detection of higher charged analyte ions for which the Nernst equation only allows a small mV change per decade concentration change. Despite the fact that their selectivity pattern is more similar to the one of the corresponding optodes they lack the drawback of the latter with respect of chromo- or fluorophore bleaching. Furthermore, in case of light absorption measurements a compromise between sensitivity (depending on the thickness) and response time must be made.

Because of the excellent sensitivity, ease of fabrication, and storage capabilities these microsensors are ideally suited for medical bedside monitoring or "point of care" devices. Figure 23 shows a novel cartridge developed at the ICB Muenster for blood electrolyte measurements. In comparison with existing devices no open blood transfer with a certain risk of infection is needed. In another special version of the cartridge, a valve allows the continuous uptake of blood in the syringe needing only the first 100 µL for the electrolyte. The content of the syringe can

then be analyzed as usual. The evaluation is performed by putting the cartridge in a hand-held readout unit. Note, prior to blood sampling the same cartridge has to be checked by the same readout unit for quality control reasons and recalibration.

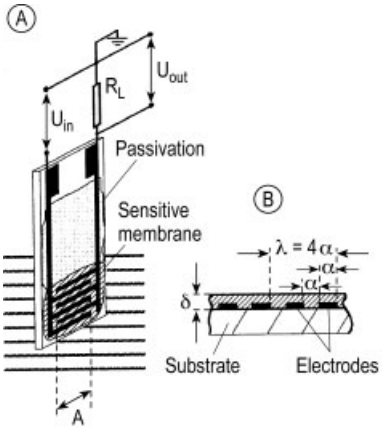

Figure 21. Schematic drawing of the construction and measuring circuit of ion-selective conductometric microsensors not requiring any reference electrodea) General view; b) Cross section

Conductometric Sensors for Monitoring Gaseous Compounds. *Introduction.* Gaseous compounds in the gas phase are detected with chemically sensitive devices that respond electrically to changes in conductance. BRATTAIN and BARDEEN (using germanium, in 1952) [104] and HEILAND (zinc oxide, 1954) [105] were the first to observe modulation of the electrical conductance of a solid-state device by gases from the surrounding atmosphere. The most commonly used material for conductometric solid-state gas sensors is now tin dioxide [106]. As noted previously (Section 2.2.2) commercial sensors based on polycrystalline SnO_2 ("Taguchi sensors") [107] are produced by companies such as Figaro Inc., for a wide range of industrial and domestic applications, including gas monitors, leak detectors, and alarm systems for toxic or inflammable gases. Apart from SnO_2, an *n*-semiconductor, many other inorganic semiconducting oxides can serve as sensitive materials in gas sensors, including TiO_2, ZnO, and Fe_2O_3 with various additives and dopants [13], [108]. Certain organic semiconductors such as phthalocyanines with appropriate central metal atoms, or polymers like polypyrrole or polysiloxane, also show sensitivity to gaseous compounds. These sensors are often referred to as homogeneous or heterogeneous semiconductors, or as dielectric sensors. Detailed lists of materials sensitive to detectable gaseous compounds together with the corresponding operating temperatures are provided in [89–92], [108]. The design principle upon which such sensors are based is quite simple: a sensitive semiconducting sample is placed between two metallic contacts, and the combination is equipped with a resistance heater capable of maintaining the appropriate working temperature, usually between 100 °C and 700 °C. Semiconductor sensors have dimensions in the range of millimeters, and are easily fabricated by either thick- or thin-film technology [109].

Heterogeneous Semiconductor Sensors. The response of a heterogeneous semiconductor sensor is based on changes in surface conductance as mentioned in the introduction, mainly due to changes in the free electron concentration within a thin surface layer of the semiconductor as a consequence of charge exchange with adsorbed species accompanying chemisorption or heterogeneous reactions [33] (see Section 2.2.2). The most commonly used semiconducting materials are the two very stable oxides SnO_2 [110], [111] and ZnO [112]. Because the conductance change is a surface phenomenon, the ratio of surface to bulk material should be made as large as possible (using thin layers) to maximize the effect. For sintered polycrystalline materials the grain size should also be as small as possible.

At a given temperature the number of gas molecules adsorbed at a semiconductor surface – hence the conductance – correlates exponentially with the partial pressure of the gas. Oxidizing gaseous compounds such as oxygen, nitrogen dioxide, and chlorine cause a resistance increase by extracting electrons from the conduction band of an *n*-type semiconductor, while reducing gases like hydrogen, carbon monoxide, methane, hydrocarbons, or alcohols lead to a resistance decrease since they are themselves oxidized by the adsorbed oxygen . Humidity decreases the resistance of an *n*-type semiconductor, causing considerable interference.

Response times for sensors of this type are usually in the range of seconds or minutes. For some gases it has been found that catalytic effects at the semiconductor surface can have a

significant impact on the response. Thus, sensor properties such as response and selectivity can be improved by adding catalytically active metals like palladium, rhodium, silver, or platinum. This doping of the semiconductor surface is one important way of modifying the selectivity of a sensor [113–115], but another possibility for adjusting sensor properties for a particular application is careful selection of the working temperature [116].

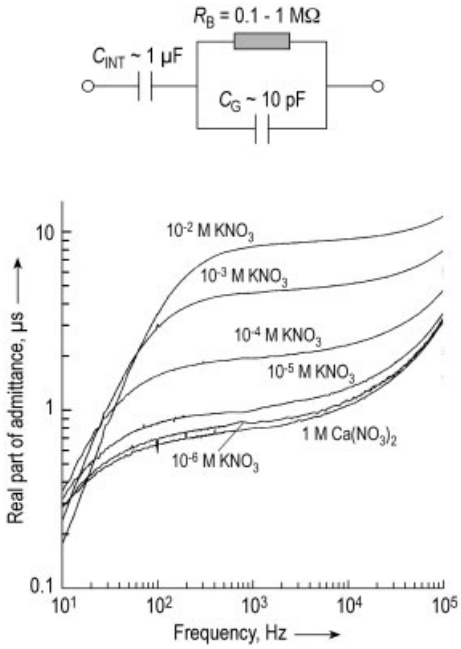

Figure 22. Simplified electronic equivalent circuit (measured is R_B = bulk membrane resistance) and impedance spectra of the real part (admittance) in a 1 M $Ca(NO_3)_2$ (!) supporting electrolyte and increasing K^+ concentrations with a valinomycin-based potassium conductometric microsensor. Note that even μM amounts of K^+ ions in a well conducting electrolyte solution can be sensed

The influence of temperature on the selectivity of a sensor depends on the fact that each gas has its own *characteristic temperature:* a temperature at which the reversible relationship between gas concentration and the concentration of electrons at the semiconductor surface (and therefore the conductance) reaches a maximum [110]. Working temperatures for SnO_2-based sensors usually lie between 200 °C and 450 °C depending on the application.

Various techniques are used for the fabrication of semiconductor sensors. Conductance sensors from structurized sintered polycrystalline ceramics can be produced by thick- or thin-film technology. Chemically sensitive materials in the form of single crystals or whiskers can be attached to electrodes by thin- or thick-film techniques as well. Mass production of sensors requires that the resulting devices be characterized by a defined level of conductance. For example, the conductance of polycrystalline SnO_2 can be adjusted by subsequent thermal treatment > 800 °C under a controlled partial pressure of oxygen. Another approach to defined conductance involves doping the semiconductor with antimony or fluorine. The reproducibility and stability of a sensor signal is generally diminished by irreversible changes in the defect structure of the semiconductor surface.

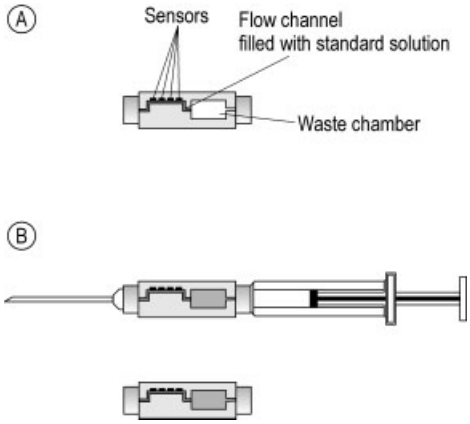

Figure 23. Measuring principle of a cartridge system using microsensors for the case of medical "point of care" blood electrolyte measurements
A) Storage of the cartridge with a standard solution; recall of data; measuring of the standard solution for plausibility and quality controls.
B) Drawing of the blood sample; displacement of standard; disconnection of the cartridge; insertion of the cartridge into the analyzer; measuring of the undiluted blood sample.

Homogeneous Semiconductor Sensors. The effectiveness of this type of semiconductor sensor depends on changes in bulk conductance. The stoichiometry and concentration of bulk point defects, and thus the conductance of such materials, is a function of the partial pressure of a gaseous analyte. Conductance changes arise through reversible changes in small deviations from ideal stoichiometry, which in turn influence the concentrations of bulk point defects and free electrons [33]. The deviation from ideal stoichiometric composition is usually very small; for example, in $Cu_{2-x}O$ the deficit x in the Cu

ratio is on the order of 0.001. Since conductance and deviation from ideal stoichiometry are proportional, a measured conductance can be used to determine a corresponding partial pressure. Regarding the kinetics of these sensors, an important factor is diffusion of a compound like oxygen or the corresponding lattice defects that compensate for nonideal stoichiometry within the bulk of the semiconductor. Any gaseous reaction involving point defects at the surface of the semiconductor must also be appropriately fast. Both requirements can be met through the combination of high temperature and a catalytic coating (e.g., Pt).

Most bulk-conductance sensors consist of semiconducting oxides the conductance of which changes with the partial pressure of oxygen. They are used primarily for monitoring oxygen partial pressures in the air or in combustion gases. Titanium dioxide (TiO_2) is one of the most commonly used homogeneous sensor materials for the determination of oxygen partial pressures [117]. At 900 °C and an oxygen partial pressure of 0.1 MPa, TiO_2 has precisely the stoichiometric composition. At lower oxygen partial pressures TiO_2 has an oxygen deficit x expressed by the formula TiO_{2-x}. This oxygen deficit specifies the number of unoccupied oxygen sites and the corresponding excess of trivalent titanium in the lattice. Both defects act as electron donors, thereby contributing to the concentration of free electrons in the conductance band of the semiconductor. The oxygen deficit, and thus the conductance, varies inversely with the partial pressure of oxygen in a characteristic way. Thus, the conductance σ_e of an n-type semiconducting oxide like TiO_{2-x}, CeO_{2-x}, or Nb_2O_{5-x} is given by the equation:

$$\sigma_e = \text{const}\,(T)\cdot p\,(O_2)^{-1/m} \tag{12}$$

where const(T) is a temperature-dependent constant and m usually varies between 4 and 8 depending on the nature of the disorder in the oxide structure.

On the other hand, p-type semiconducting oxides like $Cu_{2-x}O$, $Ni_{1-x}O$, and $Co_{1-x}O$ have a deficit x of metal relative to ideal stoichiometry. This deficit increases with increasing oxygen partial pressure. Ionic lattice defects occurring in oxides of this type act as electron acceptors. In this case the concentration of the defects, and therefore the conductance, increases with increasing oxygen partial pressure according to the expression:

$$\sigma_e = \text{const}\,(T)\cdot p\,(O_2)^{1/m} \tag{13}$$

Here again, m usually varies between 4 and 8.

The strong exponential dependence of const(T) on temperature is a great disadvantage with homogeneous semiconductor sensors, one that can lead to large measurement errors. Nevertheless, appropriate sensor design makes it possible to compensate for the temperature dependence. In the case of a TiO_2 combustion-gas sensor, for example, the conductance of an oxygen-sensitive TiO_2 sample can be compared with that of a second sample displaying the same response to temperature but coated in such a way as to eliminate its sensitivity to gas composition. The conductance difference between the two samples is then a function only of the gas composition, not of fluctuations in the temperature.

An important requirement for any metal oxide under consideration for use as a homogeneous semiconducting sensor is that it display a sufficiently high bulk diffusion coefficient $\tilde{D}$ for the conductance-determining species. Response times t for oxide layers with a thickness y can be estimated from the Einstein equation:

$$y^2 = 2\tilde{D}\cdot t \tag{14}$$

For example, a Cu_2O layer with a thickness of 1 µm reaches a response time of ca. 1 ms, corresponding to a diffusion coefficient of ca. $10^{-5}\,cm^2\,s^{-1}$, only upon heating to 700–800 °C. The diffusion coefficient can be altered by doping, but it also depends on the extent of deviation from ideal stoichiometry.

A number of oxides (e.g., TiO_2, CeO_2, Nb_2O_5, CoO) or titanates with a perovskite structure (Ba/Sr/$CaTiO_3$) have been successfully tested for application as extremely fast responding oxygen sensors [118–120] allowing individual cylinder control in automobile engines. Most of these investigations involved determination of the oxygen partial pressure in exhaust gases from combustion processes at partial pressures between 10 and 10^{-18} kPa. At low oxygen partial pressures and high temperature, certain oxides (e.g., Cu_2O, CoO, NiO) are unsuitable because of their reduction to pure metal.

Sensors based on TiO_2 ceramics are by far the most highly developed devices. These consist of porous sintered ceramic materials with a platinum surface doping, and are used to determine air – fuel ratios in the exhaust gases of combustion engines.

Apart from TiO_2, a number of other semiconducting oxides might be useful as sensor materials for special applications over limited ranges of oxygen partial pressure, as in monitoring the level of oxygen in inert gases. There is also a good possibility of optimizing various ternary and higher oxide combinations for sensor application by suitable choice of composition with respect to stability, diffusion, and surface properties [121–123]. Thus, barium and strontium ferrates show conductances with considerably less temperature dependence relative to TiO_2 or CoO.

Organic Conductometric Sensors. For several years, certain organic semiconducting polymers have attracted attention in the search for new chemical sensors [124]. The possibility of utilizing conductivity changes induced by gases adsorbed on organic materials is interesting for a number of reasons. For example, it might permit the detection of very low levels of atmospheric pollutants. Moreover, organic materials are much more easily modified than inorganic materials with respect to such characteristics as sensitivity, working temperature, and selectivity. The greatest disadvantage of organic materials in gas detection is that they are usually very poor conductors, and conductivity measurements would be correspondingly difficult. Also, organic materials are thermally unstable, so it is often impossible to use them at temperatures at which gas – solid interactions proceed rapidly and reversibly.

Phthalocyanines constitute one group of organic materials with thermal stability $> 400\,^{\circ}C$, and they are usually considered to have semiconducting rather than insulating properties. Thin layers of phthalocyanine complexes based on metals like lead [125] and copper [126] have been particularly carefully studied as potential gas sensors [127]. The conductance of these films is a very sensitive function of gases with electron-acceptor properties, including nitrogen dioxide, chlorine, and iodine. Films of phthalocyanines with various central metal atoms can be deposited on substrates with a planar interdigital electrode array by sublimation or screen printing. Changing the nature of the central metal atom makes it possible to adjust the sensitivity and selectivity of a phthalocyanine film in accordance with the proposed application [128–133].

Another organic semiconductor with gas-sensitive properties is polypyrrole. Chemiresistors with ambient-temperature gas sensitivity to both electron-donating and electron-accepting gases have been achieved using thin polypyrrole films prepared by electrochemical polymerization on electrodes. Sensitivities have been investigated toward such gases as nitrogen dioxide [134], ammonia [135], and methanol [136], [137]. A wide variety of polymers of this type is available, including substituted polypyrroles, polythiophenes, polyindoles, and polyanilines. This is leading the way to combining more of these sensors in sensor arrays and using chemometrics to evaluate unknown smells (electronic noses).

Humidity Sensors. Humidity sensors are devices that respond to changes in water-vapor pressure by a change in electrical resistance, capacitance, or a combination of the two as reflected in the impedance. The sensitive material might be an electrolyte, an organic polymer, or a metal oxide subject to a change in ionic – electronic conductivity or capacitance with respect to humidity as a consequence of physical adsorption or absorption of water molecules. In general these devices offer a very cost-effective continuous or spot- check approach to monitoring humidity in air and other gases [138], [139].

Electrolyte humidity sensors. Electrolyte humidity sensors are based on an electrolyte solution (e.g., the hygroscopic LiCl) the ion conductivity of which changes as a result of evaporation or condensation of water in response to changes in the relative humidity of the surrounding atmosphere. The fabrication of these sensors is quite straightforward. A porous substrate or an organic binder is impregnated with the electrolyte in such a way as to prevent electrolyte from flowing out of the sensor even at very high humidity. Contact is maintained with platinum electrodes on both sides of the substrate. Even thin films of potassium metaphosphate or barium fluorite are suitable for humidity measurements. Despite their slow response and limited linear

range these sensors have an outstanding reputation due to good reproducibility, long-term stability, and low cost.

Ceramic Humidity Sensors. The first widely-used ceramic humidity sensor was the aluminum oxide sensor, in which aluminum is oxidized anodically in an acidic solution to form a porous oxide film, sometimes less than 1 μm thick, that displays the necessary hygroscopicity. Adsorption of water molecules causes a decrease in resistance and an increase in capacitance, which is reversed by subsequent desorption. Other materials are also used in humidity sensors, including ZnO with added Li_2O and V_2O_5, colloidal Fe_3O_4, α-Fe_2O_3, mixed oxides of TiO_2 and SnO_2, $MgCr_2O_4-TiO_2$, $TiO_2-V_2O_5$, $ZnCr_2O_4-LiZnVO_4$, $MgAl_2O_3$, $MgFe_2O_4$, or $H_3PO_4-ZrSiO_2$, as well as $PbCrO_4$ with added alkali oxide [140]. Sometimes the signal of such a sensor is subject to drift as a result of changes in the surface structure or contamination from other adsorbed gases. Nevertheless, reproducibility can be restored by heat cleaning.

The most important considerations with respect to sensor characteristics are surface properties (hydrophilic – hydrophobic), pore-size distribution, and electrical resistance. To ensure adequate sensitivity and response a sensor of this type should consist of a very thin film of porous ceramic with a porosity $> 30\,\%$. The contacting electrodes may be interdigitated or porous sandwich-type structures made from noble metals (e.g., Pd, Pt, Au) and so constructed that they do not obstruct the pores of the oxide film. Humidity affects not only the resistance of a porous ceramic but also its capacitance by extending the surface area in contact with the electrodes. The high dielectric constant of adsorbed water molecules also plays an important role.

Polymeric Humidity Sensors. Certain polymers are capable of taking up water as a function of the relative humidity of the surrounding atmosphere, leading to changes in such mechanical and electrical properties as volume, resistance, and dielectric characteristics. A number of polymers have been used as sensitive materials in this regard, including ion-exchange resins, polymer electrolytes, hydrophilic polymers, and hygroscopic resins [141]. Hydrophilic polymers [e.g., poly(ethylene oxide) – sorbitol] swell by absorption of water. If such a polymer is filled (mixed) with carbon or metal powder of very small particle size, the resistance observed depends on the distance and the number of particles that are in contact with each other. Uptake of water leads to swelling of the polymer and disturbance of the ohmic contacts between conducting particles, which in turn increases the resistance. For this reason the resistance correlates with humidity. Useful materials include cellulose esters and polyalcohols.

The resistance of polymers with dissociable groups (e.g., sulfonates, amines, amides, or even hydroxides) also depends on humidity. The resistance of such polymers in fact decreases with absorption of water because of ionic dissociation and an increase in ionic conduction.

The absorption of water by polymers is especially suited to the development of a capacitive humidity sensor based on the high dielectric constant of water, since capacitance depends on the area, the thickness, and the dielectric constant of a dielectric. Uptake of water causes an increase in the dielectric constant followed by an increase in capacitance. Polyimide is a suitable sensitive material, but so are certain inorganic ceramics such as Al_2O_3 [142] and low-density Ta_2O_5, which change their capacitance with the formation of water dipoles. Low-density Ta_2O_5 layers can be formed by anodic oxidation of sputtered tantalum films, whereas polyimide films are deposited by spin coating and subsequent polymerization.

Conducting Polymer-Based Gas Sensors. The above mentioned ion-selective conductometric microsensors can easily be transformed into gas sensors by omitting the ionophore and adding an organic salt for a certain ground conductivity instead. These PVC/plasticizer membranes are sensitive to reaction with certain gases that can enter the membrane phase and alter the conductivity of the organic ions either by increasing their dissociation from an ion-pair situation or by increasing their diffusional property due to membrane swelling. Figure 24 gives an example of when the salt and the plasticizer are present simultaneously. The selectivity for a certain gas can be modulated by changing the type and concentration of the latter. Therefore, much more variations are possible compared with other polymer-based gas sensors based only on their known gas chromatographic distribu-

tion behavior. A typical membrane composition is: about 50 % PVC, 30 – 35 % plasticizer, the remainder being the organic salt such as tetraalkylammonium halides. The compounds are as usual dissolved in tetrahydrofurane and then drop-coated on the interdigital electrode array with similar dimensions as in the above mentioned case of the conductometric ion-selective microsensors.

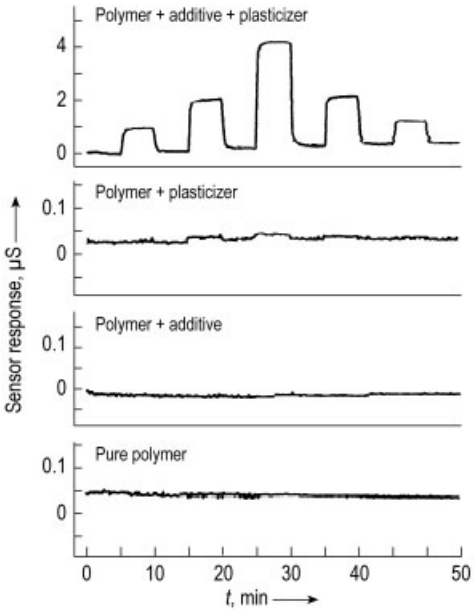

Figure 24. Variations of the sensor membrane composition. Sensor response against different concentrations of ethanol (500 ppm, 1000 ppm, 2000 ppm).
Polymer: polyisobutylene; plasticizer: ortho-nitrophenyloctylether; organic salt: tetraoctylammoniumbromide

Gases that can be detected in the lower ppm range with response time of a few seconds are so far: tetrachloroethene, ethanol, benzene, toluol, and xylene. The measurement technique is exactly the same as with the conductometric ion sensors for application in liquids (Fig. 21). Compared with other gas-sensing devices the measurements are performed at room temperature. Thus, the energy consumption is very low, which will open the way to hand-held instruments. Unfortunately the optimization of the membrane composition for one gaseous analyte has yet to be on a purely empirical base. First studies at the ICB Muenster showed that the sensitivity and selectivity could be changed by many orders of magnitude [143], [144]. One example of a membrane with a certain selectivity towards tetrachloroethene to be deployed in the dry cleaning business is shown in Figure 25. Other examples are shown together with electronic nose developments.

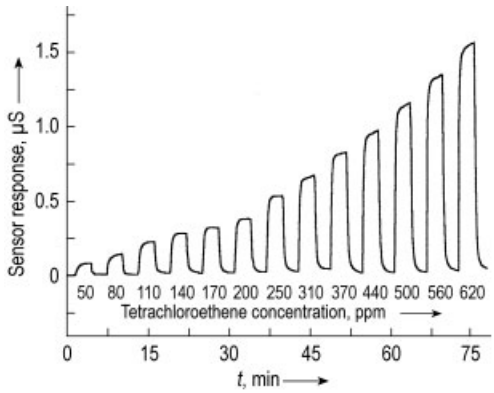

Figure 25. Polymer membrane sensor response against tetrachloroethene. The membrane contains polyethylenoxide, 2-fluorophenyl-2-nitrophenylether, tetraoctylammoniumbromide.

Here the relevant limits of 25 ppm room concentration and 270 ppm in the cleaner before door is opened can easily be detected. Compared with the semiconductor gas sensors, humidity only changes the zero point for the measurements not the slope of the calibration curve. Figure 26 shows the humidity effect on a conductometric ethanol microsensor. Thus any placement of the sensor in a location with similar humidity but zero analyte concentration will correct for that.

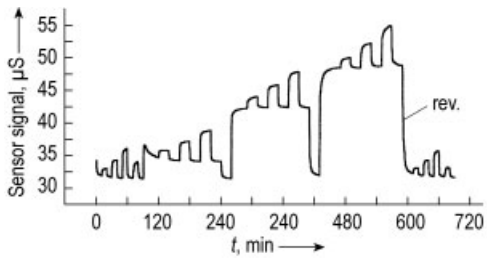

Figure 26. Humidity effect on a conductometric polymer gas sensor for ethanol.
Response toward: 350, 500, and 700 ppm ethanol at 10 %, 40 %, and 60 % relative humidity; membrane composition, 54 % polyepichlorhydrin, 36 % o-nitrophenyl-octyl-ether, 10 % tetradodecylammoniumbromide

In order to solve the humidity problem of most gas sensors the ICB Muenster developed and instrument called “Air-Check” in which a

short but efficient pre-sampling on Tenax material takes place. With this the sensitivity could be increased as needed, water no longer interferes and the selectivity could also be increased by analyzing the time course of the rapid thermal desorption step. The whole adsorption – desorption cycle can be performed automatically within less than 1 min!

2.3.1.4. Ion-Selective Field-Effect Transistors (ISFETs)

As described in Section 2.3.1.1, with conventional ion-selective electrodes (ISEs) the ion-selective membrane is placed between the analyte and the internal reference solution. An internal reference electrode makes electrical connection to the internal reference solution, and the measurement system is completed by a second reference electrode in contact with the analyte solution. In any modern pH meter the internal reference electrode is connected to the gate metal of a metal-oxide semiconductor field-effect transistor (MOSFET), which acts as an impedance- converting device (Fig. 27). Normally, a conventional ISE functions very well. However, in certain applications problems can arise due to large size, high cost, and sensitive construction.

These problems might be resolved by the ion-selective field-effect transistor (ISFET). In an ISFET the ion-selective membrane is placed directly on the gate insulator of the field-effect transistor; alternatively, the gate insulator itself, acting as a pH-selective membrane, may be exposed to the analyte solution (Fig. 28). If one compares an ISFET with the conventional ISE measurement system, the gate metal, connecting leads, and internal reference system have all been eliminated. An ISFET is a small, physically robust, fast potentiometric sensor, and it can be produced by microelectronic methods, with the future prospect of low- cost bulk production. More than one sensor can be placed within an area of a few square millimeters. The first ISFETs were described independently by Bergveld [142] and Matsuo, Esashi, and Inuma [145]. The Institute of Microtechnology of the University of Neuchatel is very active in this field [146].

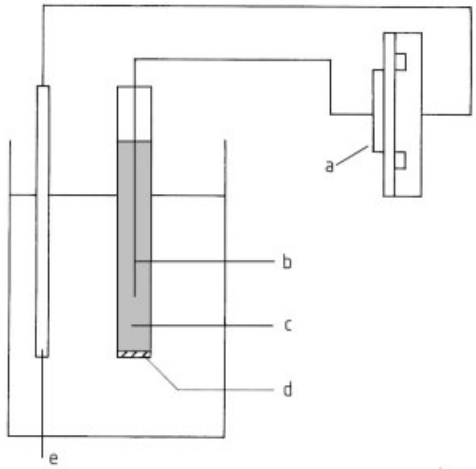

Figure 27. Schematic diagram of a membrane-electrode measuring circuit and cell assembly
a) MOSFET; b) Internal reference half cell; c) Internal filling solution; d) Ion-selective membrane; e) Reference electrode

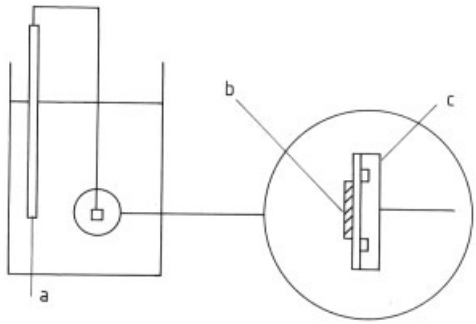

Figure 28. Schematic diagram of an ISFET measuring circuit and cell assembly
a) Reference electrode; b) Ion-selective membrane; c) ISFET

ISFET Operation. The ISFET can be regarded as a special type of MOSFET. To understand the operation of an ISFET it is therefore necessary first to understand the operation of a MOSFET.

An n- channel MOSFET (Fig. 29) consists of a *p*-type silicon substrate with two *n*-type diffusions, the source (c) and the drain (d). The structure is covered with an insulating layer (b), on top of which a gate electrode (a) is deposited over the area between the source and the drain. A voltage is applied to the gate that is positive with respect to both the source and the bulk material. This voltage produces a positive charge at the gate electrode and a negative charge at

the surface of the silicon under the gate. The negative charge is formed by electrons, which are the minority carriers in the substrate, and this in turn creates a conducting channel between the source and the drain. The density of electrons in the channel, and thus the conductivity of the channel, is modulated by the applied gate voltage (V_G). If a constant voltage (V_D) is applied between the source and the drain a change in the conductivity of the channel leads to a change in current between source and drain. The drain current I_D in the unsaturated region ($V_D < V_G - V_T$) of a MOSFET is equal to:

$$I_D = \mu C_{ox} \frac{W}{L} \left[(V_G - V_T) V_D - \frac{1}{2} V_D^2 \right] \quad (15)$$

where

μ = Electron mobility in the channel
C_{ox} = Oxide capacitance per unit area
W/L = Ratio of channel width to length
V_T = Threshold voltage

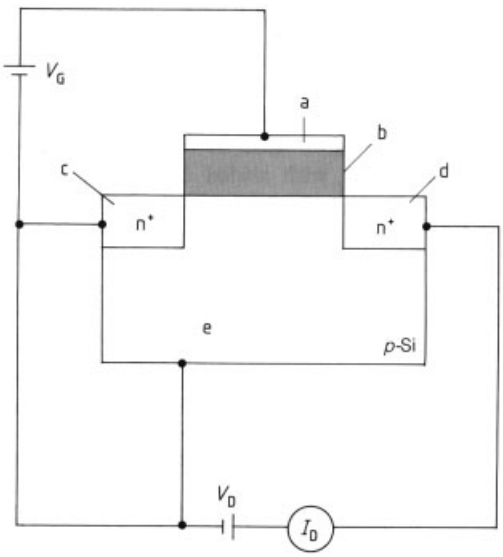

Figure 29. Schematic diagram of an n-channel MOSFET
a) Gate metal; b) Insulator; c) Source; d) Drain; e) Bulk

A detailed description of the MOSFET is provided in [147].

In an ISFET (Fig. 30) the gate metal is replaced by a selective membrane and an electrolyte solution, or simply by the electrolyte solution, which is in contact with a reference electrode. The potential of the gate with respect to the substrate is the sum of the potentials of the various electronic and electrochemical processes that occur in the external path between the silicon substrate and the gate; i.e., the reference electrode half-cell potential E_{ref}, the solution – membrane or solution – insulator interfacial potential φ_M, and the metal – silicon contact potential. These voltages are added to the externally applied voltages V_g, and the resulting voltage has the same significance and function as that defined in the theory of MOSFET operation. V_{T*}, the threshold voltage of the ISFET, can be described as:

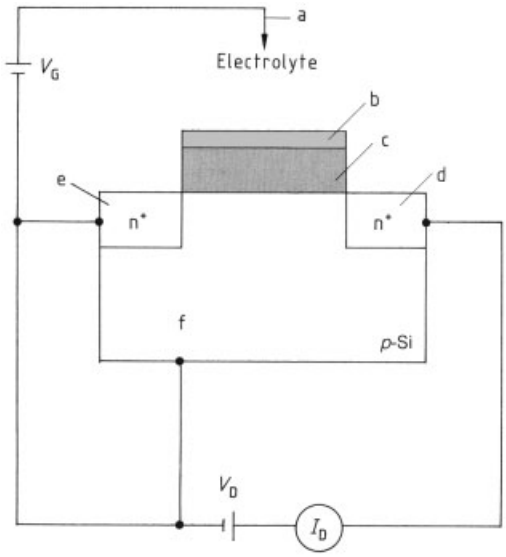

Figure 30. Schematic diagram of an ISFET
a) Reference electrode; b) Ion-selective membrane; c) Insulator; d) Drain; e) Source; f) Bulk

$$V_{T*} = V_T - E_{ref} - \varphi_M \quad (16)$$

From Equation (15) we obtain the following relationship:

$$I_D = \mu C_{ox} \frac{W}{L} \cdot \left[(V_G - V_T + E_{ref} - \varphi_M) V_D - \frac{1}{2} V_D^2 \right] \quad (17)$$

φ_M, the solution – membrane or solution – insulator interfacial potential, depends on the activity of the ions in the analyte solution according to the Nernst – Nikolsky equation [Eq. (6), Section 2.3.1.1] for ISFETs with an additional membrane, or according to a similar equation (see below) for ISFETs with a solution – insulator interface.

If all parameters in Equation (17) apart from φ_M are kept constant, changes in I_D are caused exclusively by changes in the activity of ions in the analyte solution. Thus, in an ISFET, the drain current is modulated by the ionic activity in the analyte solution. A detailed description of ISFET operation is presented in [148].

pH-Sensitive ISFET s. Gate insulators such as Si_3N_4, Al_2O_3, and Ta_2O_3 have been found to be pH sensitive, so FETs in which bare gate insulators of this type are exposed to a solution respond to changes in pH. The pH response can be explained by the "site-binding theory." This model assumes that there are ionizable binding sites present at the surface of an insulator, and that these can exhibit amphoteric behavior with respect to the potential-determining ion (i.e., H^+):

$$A-OH \rightleftharpoons A-O^- + H_S^+$$

$$A-OH + H_S^+ \rightleftharpoons A-OH_2^+$$

According to this model the pH dependence of φ_M is given by the following equation:

$$\varphi_M = 2.3\frac{kT}{q}\left[\frac{\beta}{\beta+1}\right](pH_{pzc} - pH) \quad (18)$$

under the condition that $\beta \gg q\varphi_M/kT$. pH_{pzc} is the pH value for which $\varphi_M = 0$; β is a parameter that reflects the chemical sensitivity of the outer gate insulator, with a value dependent on the density of hydroxyl groups and the surface reactivity. Surfaces with a high β value show Nernstian behavior. For a detailed description of the site-binding model, see [148].

ISFETs Sensitive to Other Ions. Deposition of an ion-selective membrane on top of the gate insulator opens the way to measurement of ions other than H^+. Most of the sensitive materials for ISE applications described in Section 2.3.1.1 have been used in conjunction with ISFETs, providing sensors covering a wide variety of species. Thus, ISFETs with solid membranes have been described with AgCl – AgBr membranes sensitive to Ag^+, Cl^-, and Br^- [149], for example, or with LaF_3 membranes sensitive to fluoride [150]. Ion-sensitive polymer-matrix membranes (liquid membranes with a polymer matrix) have also been used as sensitive membranes for ISFETs. The first experiments involved mainly PVC membranes [151], [152], but these membranes show poor adhesion and poor mechanical strength. To improve membrane adhesion, modified PVC was utilized as a matrix material [153]. The use of silicones as matrix materials has made it possible to prepare very durable ISFETs with polymer-matrix membranes [154]. Other reports describe the use of photopolymerized polymers as matrix materials [155], [156].

Problems. The commercial development of ISFETs has been slow despite intensive research. From the standpoint of broad commercialization the ISFET is in competition with conventional ISEs with respect to price, performance, and reliability, and it is essential that some means be devised for producing ISFETs by commercially acceptable methods. Reaching this stage requires that several problems be solved, however.

ISFETs have so far been used primarily in conjunction with conventional macro reference electrodes. This fact greatly limits the potential benefits to be gained from miniaturization of the sensor. An optimal reference electrode for ISFETs should be miniaturizable and display long-term stability, and it should be subject to fabrication with microelectronic techniques. Attempts have been made to miniaturize macro reference electrodes [157], or to use modified surfaces with extremely low surface-site densities as reference FETs [158], but to date the abovementioned requirements have not been met.

A further problem relates to encapsulation. The ion-sensitive gate area is exposed to the analyte solution, but all other parts of the sensor must be insulated from the solution. In most cases encapsulation is carried out by hand using epoxy resins, but it is impossible to produce ISFETs at low cost by this technique. Both photolithographic methods [159], [160] and an electrochemical method [161] have been proposed for the solution of this problem. One very promising development is the backside-contacted ISFET [162]. The electrical contacts in this case are protected from the analyte solution by means of O-rings, thereby circumventing the need for resins.

As noted previously, ISFETs selective to ions other than H^+ are prepared by deposi-

tion of an additional membrane on top of the gate insulator. If such ISFETs are to be commercialized this membrane must be deposited by a mass-production technique, which will be difficult to achieve with polymer membranes. One possible solution is the development of photolithographically patternable ion-sensitive membranes [163], but such membranes have so far been developed only for a few ions, and problems in the photolithographic deposition of more than one type of membrane on a single chip have not been resolved. The interface between an insulator and a sensitive membrane is also ill-defined in a thermodynamic sense, but this problem was solved by introducing a hydrogel between the sensing membrane and the insulator. Futhermore, ISFETs suffer from drift. The influence of this effect would be eliminated if the sensor were to be used in flow-injection analysis (FIA), so successful applications can be anticipated for ISFETs in the field of dynamic measurement [164].

Only if all the problems cited are successfully resolved will the ISFET live up to the promises that have been held out for it for more than two decades: low cost, multifunctionality, and a robust and reliable construction.

2.3.2. Optical Sensors

2.3.2.1. Fiber-Optical Sensors

The use of optical fibers has become increasingly important in the field of chemical and biochemical sensors. This is clearly due to the fact that fiber-optical sensors combine the strengths of well-known bulk optical analytical methods with the unique advantages of optical fibers. More specifically:

1) An enormous technology base already exists for familiar optical methods of chemical analysis. There is probably no class of chemical analyte that has not at some time been the subject of optical determination through absorption or fluorescence spectroscopy. Methods also exist for labeling a target analyte with an appropriate chromophore or a fluorophore, a common practice in immunoassays.
2) Fiber-optic techniques offer increased sensitivity relative to conventional bulk optic approaches. For example, evanescent wave (EW) spectroscopy is significantly more sensitive than bulk attenuated total reflection spectroscopy. Fiber-optical EW spectroscopy is the only technique suitable for use with highly absorbing or scattering media.
3) The interrogating light in a fiber-optics system remains guided, so no coupling optics are required in the sensor region, and an all-fiber approach is feasible. This represents a sharp contrast to bulk optics, where proper optical adjustment is extremely critical. Fiber optics also provides a good solution to the problem of wide separation between the sensing area and the detector, as in the in vivo monitoring of medical parameters.
4) Recent developments in optical communications and integrated optics have led to optical materials that permit significant miniaturization of the sensing and detecting devices.
5) Since the measurement signal is optical, electrical interference is avoided, and the absence of electrical connectors ensures safe operation in hazardous environments, such as those containing explosive vapors.

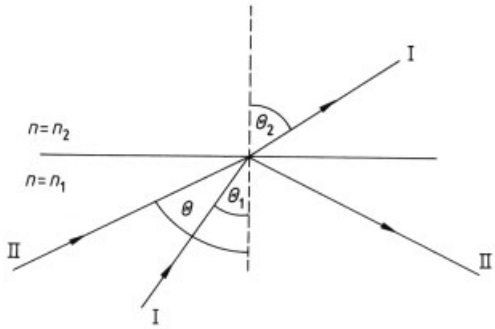

Figure 31. Refraction and total internal reflection
n = Refractive index

The light inside an optical fiber is guided by the principle of total internal reflection, as described by Snell's law and depicted in Figure 31. A ray of light striking the interface between two media with differing refractive indices n_1 and n_2 is refracted (ray I) according to the relationship $\sin(\Theta_1)\times n_1 = \sin(\Theta_2)\times n_2$. If the condition $n_1 > n_2$ is fulfilled, total internal reflection will be observed for several angles of

incidence. Thus, at an angle close to the horizontal the incident beam will no longer be refracted into the second medium, but will instead be totally reflected internally at the interface (ray II in Fig. 31). The *critical angle of incidence* $\Theta_c = \arcsin(n_2/n_1)$, which defines the onset of total internal reflection, is determined by the refractive indices of the two media, and whenever $\Theta > \Theta_c$ light is guided inside the medium characterized by n_1.

The general structure of an optical fiber is illustrated in Figure 32. Light entering from the left is confined within the core, which is surrounded by a cladding of lower refractive index (n_{clad}). The whole arrangement is surrounded by a nonoptical jacket, usually made from plastic and designed to provide both stabilization and protection. The fibers themselves can be constructed from a wide variety of transparent materials with differing refractive indices, including various glasses, fused silica, plastics, and sapphire. Most commercially available fibers have glass cores with n_{core} in the vicinity of 1.48 and diameters ranging from 5 – 200 µm; typical claddings have $n_{clad} < 1.48$ and thicknesses of 50 – 1000 µm. An optical fiber is usually characterized by its numerical aperture $NA = (n_{core}^2 - n_{clad}^2)^{1/2}$. The value of NA determines the cone angle Θ_c within which incident light will become entrapped by the fiber, where $\Theta_c = \arccos(NA)$; $\Theta_{max} = \arcsin(NA/n_{env})$ for the cone angle Θ_{max} with which light leaves the fiber at its end (see Fig. 32).

The simple ray-optical description of light-guiding presented above fails to account for several phenomena important with respect to fiber-optical sensors. The most important of these is the *evanescent wave*. The intensity of light that is totally reflected internally does not in fact fall abruptly to zero at the core – cladding interface. Instead, the electromagnetic field intensity decays exponentially as a function of distance from the interface, thereby extending into the medium of lower refractive index. This field is called the *evanescent field*, and its penetration depth d is defined by the distance from the interface within which it decays to a fraction 1/e of its value at the interface. The angle of incidence Θ strongly influences the penetration depth. Values for d vary from 50 – 200 nm, increasing as Θ decreases; d reaches its maximum when Θ is very close to the critical angle.

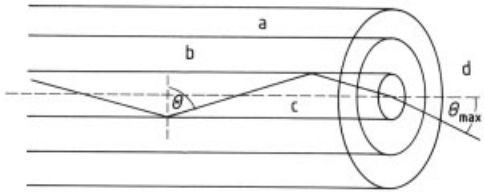

Figure 32. Structure of an optical fiber
a) Jacket; b) Cladding, $n = n_{clad}$; c) Core, $n = n_{core}$; d) Environment, $n = n_{env}$

Fiber-optical sensors can be constructed in either of two fundamentally different optical configurations with respect to the sensing area. The first is a wave-guide-binding configuration, in which recognition occurs at the surface of the fiber core via the evanescent field. This is achieved by removing the jacket and cladding from the fiber and causing the analyte molecules to bind to the core surface, as in the example described below. Such an arrangement is described as an *evanescent wave sensor* or *intrinsic fiber-optical sensor*. The resulting sensitivity depends greatly on the value of NA [165].

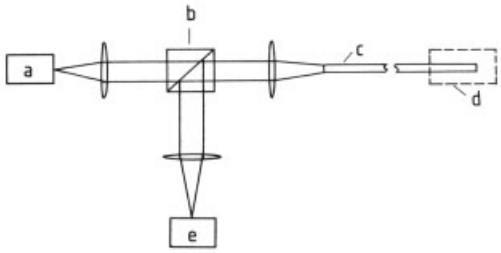

Figure 33. General configuration of an optical detection system for a fiber-optical sensor
a) Source; b) Beam splitter; c) Fiber; d) Sensing area; e) Detector
The excitation system consists of a xenon or tungsten lamp together with a monochromator or optical filter, an LED, or a laser, whereas the light detection system might be a monochromator or filter with a photomultiplier or photodiode.

Alternatively, the fiber-end configuration might be considered, in which analyte is released for recognition purposes into the illuminated space at the end of the fiber. In this case the fiber acts only as a light pipe, resulting in what is known as an *extrinsic fiber-optical sensor*.

Both configurations entail similar optical systems: a spectral light source for excitation (e.g., a xenon or tungsten lamp equipped with a monochromator or optical filter, an LED, or a laser) and a spectral light-detection device consisting of a monochromator or filter to-

gether with a photomultiplier or photodiode. A schematic diagram applicable to the optical systems for these two sensor configurations is presented in Figure 33.

More detailed information regarding the physical principles of fiber-optical chemical and biochemical sensors and the corresponding detection systems is available in [166] and [167].

Biochemical Sensors. Antibodies are widely used as recognition elements in fiber-optical biosensor applications. This is due to the high specificity of the binding reaction between an antibody and an analyte (antigen). To determine this reaction optically, labeling agents such as fluoresceine, rhodamine, or phycoerythrine are used, all of which show absorption and emission bands in the visible region of the spectrum. Linear amplification can be achieved by binding more than one label. Depending on the circumstances, either the antigen or the antibody might be tagged (see below).

The evanescent wave (intrinsic) configuration is most important for biosensors because it permits the antibody or antigen to be immobilized directly on the core surface, and it extends the path for possible interaction between the light and the recognition elements. Accordingly, only tagged molecules inside the evanescent field are subject to excitation. The amount of light energy absorbed, or the light emitted from these molecules and collected by the fiber, gives rise to the measurement signal. In contrast to the extrinsic sensor approach, the intrinsic arrangement therefore reduces noise that would otherwise result from free labeled molecules or intrinsic fluorescence within the environment [167]. Despite these advantages, however, extrinsic biosensors have also been constructed, with binding devices placed in front of the fiber as a way of fixing the recognition elements. These devices can be either glass plates or transparent membranes. This offers the advantage of easy removal and replacement of the recognition elements [168].

Figure 34 provides an example of a biosensor of the intrinsic type. Here the antigen has been immobilized at the fiber core surface where it has been allowed to bind with a labeled antibody. Energy from the evanescent wave is absorbed by the labels, generating a detectable fluorescence signal. Addition of unlabeled analyte to the surrounding solution leads to competition for the binding sites. This permits some of the labeled antibody molecules to diffuse away from the surface, causing the fluorescence signal to decrease. The same general approach is also applicable to immobilized antibodies and tagged antigens [169].

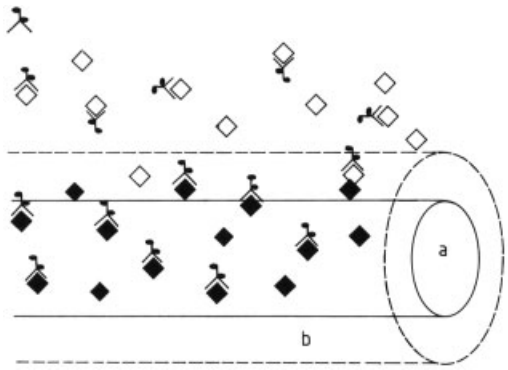

Figure 34. A fiber-optical immunosensor, in which labeled antibody molecules are gradually replaced by free analytes, causing the fluorescence signal to decrease
a) Core; b) Region of evanescent wave
♦ Fixed analyte; ◇ Free analyte; Tagged antibody

Competition between a labeled antigen and an unlabeled analyte can also be effected in an external column, where the labeled antigen has previously been bound to an antibody. Once the competition reaches equilibrium any labeled antigen that has been released can be detected with antibody immobilized on a fiber. In this case the fluorescence signal would be expected to increase with an increased concentration of analyte.

Enzymes are in some cases also useful for recognition and labeling purposes. For example, glucose oxidase has been used to determine glucose on the basis of free oxygen that is consumed, where the disappearance of oxygen is measured with a fiber-optical oxygen subsensor [170]. Labeling of antibodies or antigens with urease or alkaline phosphatase and subsequent addition of a fluoro- or chromogenic substrate leads to detectable production of fluorophores or chromophores. The signal in this case is subject to multiple enhancement since it is proportional to the concentrations of both labeled antigen and antibody as well as to incubation time.

Chemical Sensors. Recognition with a *chemical sensor* is achieved via an analyte-specific reaction involving chemical compounds located inside the sensing area. The reaction must be one that is accompanied by changes in absorbance or fluorescence characteristics. Several examples are described below. A common physical arrangement utilizes a fiber-end approach and two single fibers or two fiber bundles, one for guiding light into the sensing area and the other to carry it away. The chemical reagents are confined within a transparent membrane or a small segment of the fiber itself, near its end and shielded by a membrane.

A particularly wide variety of fiber-optical sensors has been developed for measuring pH values, and many of these sensors are also used as subsensors for other analytes. Sensors for pH determination can be divided into two general types. The first relies on colorimetric acid – base indicators to produce changes in absorbance. For example, the dye phenol red can be copolymerized with acrylamide and bisacrylamide to yield dyed polyacrylamide microspheres. These are in turn packed into a length of cellulose dialysis tubing placed at the end of a pair of fibers that can serve as a vehicle for monitoring changes in absorbance [171]. Bromothymol blue immobilized on styrene – divinylbenzene copolymer has been similarly employed [172]. Alternatively, fluorometric acid – base indicators can be used to detect changes in fluorescence intensity that accompany protonation. Thus, one pH sensor was based on fluorescence quenching by H^+ of the excited state of immobilized fluoresceinamine [173]. 8-Hydroxypyrene-1,3,6-trisulfonate (HPTS) fixed on ethylene vinyl acetate has also been invoked as a fluorescence pH sensor [174].

Chemical sensors for other ions have been reported as well, including the halides and pseudohalides [175]. A sensor of this type is based on the dynamic quenching of fluorescence emission from glass-immobilized, heterocyclic, acridinium and quinolinium indicators. The fluorometric process involved is known to obey the Stern – Volmer equation [176]. Al(III) ions have been determined with morin (3,5,7,2′,4′-pentahydroxyflavone) [177]. Thus, aluminum displaces the hydrogen ion when it binds to morin, leading to a change in fluorescence properties. Be(II) has also been determined using morin [178]. Furthermore, reports have appeared on sensors for Mg(II), Zn(II), and Cd(II) based on quinoline-8-ol sulfonate immobilized electrostatically on an anion-exchange resin [179].

Gas sensors for carbon dioxide [180] and ammonia [181] in aqueous solution have been constructed on the basis of a gas-permeable membrane and an optical pH subsensor. Gases diffusing through the membrane produce a change in the pH in the sensing region of a fiber-optical pH sensor. Sulfur dioxide [182] and oxygen [183] sensors take advantage of the quenching of the energy transfer (Förster transfer) between excited pyrene and perylene, both of which can be immobilized on a silicone matrix. Other O_2 sensors rely on the effect of fluorescence quenching as described by the Stern – Volmer equation. These generally involve metal complexes like tris(2,2′-bipyridine)ruthenium(II) [184]. In most cases fluorescence quenching also leads to a shortening of the fluorescence lifetime, which can be determined, for example, by phasesensitive detection [185]. The advantage of this method is that fluorescence lifetime is independent of the fluorophore concentration, and there is no signal drift due to bleaching.

2.3.2.2. Integrated Optical Chemical and Biochemical Sensors

Generally speaking, integrated optical devices can be constructed using techniques similar to those for manufacturing semiconductor devices. This offers great potential for cost reduction and facilitates the miniaturization often required in sensor applications. Furthermore, such sensors retain all the advantages of the well-known fiber-optical sensors.

The propagation of a light ray confined within a planar dielectric film (or *waveguide*) is depicted in Figure 35. Guidance of the light is possible only if total internal reflection occurs at the interfaces film – cover (I) and film – substrate (II). This condition is met if the refractive index of the film n_f is greater than that of both the cover n_c and the substrate n_s. Moreover, the angle of incidence Θ must be greater than the critical angles for total internal reflection at the interfaces I (Θ_{cI}) and II (Θ_{cII}), given by the expressions $\sin(\Theta_{cI}) = n_c/n_f$ and $\sin(\Theta_{cII}) = n_s/n_f$. To fulfill

the self- consistency condition within the waveguide, the phase shift of the light wave as it travels from point 1 to point 3 must be an integral multiple m of 2π, where m is called the *mode number*.

$$\left(\frac{2\pi}{\lambda}\right)\cdot L\,(n_{\mathrm{f}},h,\Theta)+\Phi_{\mathrm{I}}\,(n_{\mathrm{f}},n_{\mathrm{c}},\Theta)$$
$$+\Phi_{\mathrm{II}}\,(n_{\mathrm{f}},n_{\mathrm{s}},\Theta)=2\pi m \qquad (19)$$

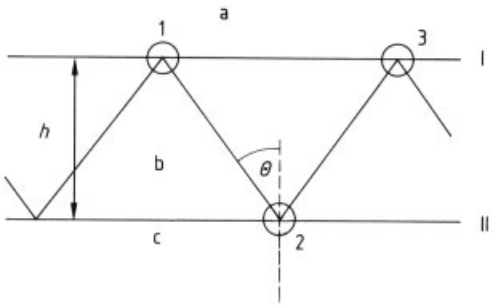

Figure 35. Propagation of a confined light ray within a planar waveguide mounted on a substrate
Most substrate materials are glasses or crystals with a refractive index of 1.4 – 2.4, and the cover material is air; the thickness of the film ranges from 0.1 to several micrometers.
a) Cover, $n=n_{\mathrm{c}}$; b) Film, $n=n_{\mathrm{f}}$; c) Substrate, $n=n_{\mathrm{s}}$

Φ_{I} and Φ_{II} are the phase shifts caused by reflection at the interfaces I and II, and $(2\pi/\lambda)\,L$ is the shift accompanying travel along the optical path $L=2\,h\cos(\Theta)$. The values of $n_{\mathrm{f}}, n_{\mathrm{c}}, n_{\mathrm{s}}$, and m are usually fixed by the fabrication process for the waveguide, and λ is established by the light source. The value of Θ can therefore be calculated from Equation (19). It is useful to define an *effective refractive index* n_{eff} for the guided mode, defined as $n_{\mathrm{eff}}=n_{\mathrm{f}}\sin(\Theta)=n_{\mathrm{eff}}\,(n_{\mathrm{f}},n_{\mathrm{c}},n_{\mathrm{s}},h,m,\lambda)$, which, apart from λ, depends only on the material and geometric parameters of the waveguide. The parameter n_{eff} can be used to treat all types of light-propagation effects within the waveguide.

Even though light is totally reflected internally at the interfaces, its intensity does not fall to zero outside the waveguide. Just as with optical fibers, the small part of the guided light located outside the waveguide is referred to as the evanescent field, with a penetration depth ranging from 100 – 300 nm depending on the waveguide performance. A more detailed mathematical treatment of planar and strip waveguides is presented in [186].

Waveguides intended as chemical or biochemical transducers can be operated in either of two modes. A comprehensive physical description of these operating modes and the corresponding efficiencies is provided in [187].

First, it is possible to use a waveguide for acquiring refractive index data. This can be accomplished by coating the waveguide with a reactive but transparent chemical layer of a thickness greater than the penetration depth of the evanescent field. In this case the layer can be treated as if it were a cover material. Any chemical reaction with the layer will result in a change in the refractive index n_{c}, which in turn leads to a change in the effective refractive index n_{eff}. This approach has been used successfully in the production of sensors for CO_2 and SO_2 [188].

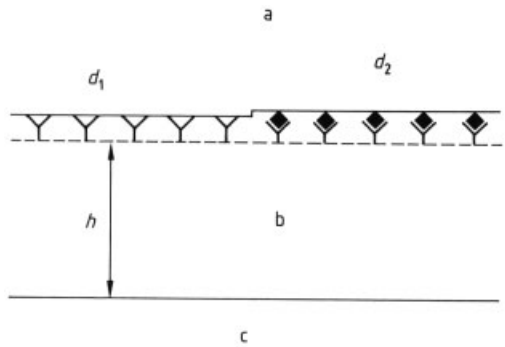

Figure 36. Specific antibody – antigen binding at the surface of a planar waveguide, causing the thickness of the adlayer to change
a) Cover, $n=n_{\mathrm{c}}$; b) Film, $n=n_{\mathrm{f}}$; c) Substrate, $n=n_{\mathrm{s}}d_1$, d_2 = Layer thicknesses before and after binding

The adsorption and desorption of molecules is depicted in Figure 36 as an example of the second operational mode. The waveguide in this case is coated with a thin chemical layer consisting, for example, of an antibody. When the antibody – antigen binding reaction occurs, the effective thickness [187] of the entire waveguide increases, which again causes a change in the effective refractive index of the guide. Direct immunosensors based on this method have often been reported in the literature [189], [190].

Detecting a change in n_{eff} is a problem that can be solved in various ways. An interferometer approach analogous to the Mach – Zehnder interferometer [191], [192], the Fabry – Perot interferometer [188], or the difference interferometer [193] has often been used. The general arrangement of a Mach – Zehnder interferometer is illustrated in Figure 37 A. The diameter of the strip waveguide is usually on the order

of 5 µm, with a distance between the two arms of < 1 mm, and a total length for the device of < 3 cm. Only the signal arm with its reactive chemical layer is exposed; the reference arm is covered by a dielectric film with a refractive index lower than that of the waveguide. The output power P_{out} of the device depends on the extent of the light phase shift $\Delta\Phi = (2\pi/\lambda)\, L\, n_{\mathrm{eff}}$ between the signal arm and the reference arm; e.g., $P_{\mathrm{out}} = P_{\mathrm{in}} \cos^2(\Delta\Phi/2)$.

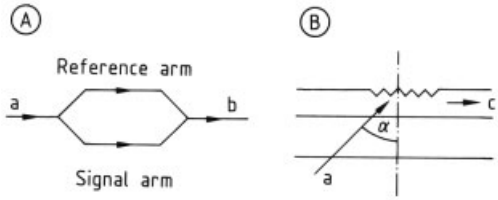

Figure 37. General arrangement of a Mach – Zehnder interferometer (A) and an input grating coupler (B)
In A only the signal arm is exposed to an active chemical layer while the reference arm is covered, whereas in B the grating is provided with an active layer and exposed to the sample.
a) Input beam; b) Output beam; c) Incoupled beam

Grating couplers can also be used to determine the value of n_{eff}, as depicted schematically in Figure 37 B. For efficient light *incoupling* the angle α must be consistent with the expression $n_{\mathrm{eff}} = n_0 \sin(\alpha) + l\,(\lambda/\Lambda)$, where l is the diffraction order and Λ the grating period. According to this relationship, a change in n_{eff} leads to a change in the angle α that will reinforce light guided within the waveguide. A grating coupler can also be used in an *outcoupling* arrangement. In this case it is the angle for the outcoupled light that would conform to the grating equation. A detailed description of grating- coupler methods is presented elsewhere [194].

2.3.2.3. Surface Plasmon Resonance

Transducers for optical sensing systems in chemical and biochemical sensors have also been developed based on the sensitivity of surface plasmon resonance (SPR) devices. Since first described in 1983 [195], SPR transducers have been used in many different applications, including immunoassays [195], [196] and sensors for gases [195], [197] and liquids [198].

Electron charges located at a metal – dielectric interface can be caused to undergo coherent fluctuations known as *surface plasma oscillations* or *surface plasmons* (SP). The frequency ω of such longitudinal oscillations is related to the wave vector $\boldsymbol{k}_{sp}$ *by the dispersion relationship.*

$$k_{\mathrm{sp}} = \left(\frac{\omega}{c}\right) \cdot \sqrt{\frac{e_{\mathrm{M}} \cdot e_{\mathrm{D}}}{e_{\mathrm{M}} + e_{\mathrm{D}}}} \tag{20}$$

where $\varepsilon_{\mathrm{M}}(\omega)$ is the permittivity of the metal and $\varepsilon_{\mathrm{D}}(\omega)$ that of the surrounding dielectric (cf. Fig. 38); c is the velocity of light in a vacuum. The permittivity of the metal ($\varepsilon_{\mathrm{M}} = \varepsilon_{\mathrm{M1}} + \mathrm{i} \cdot \varepsilon_{\mathrm{M2}}$) is complex and the value of $\varepsilon_{\mathrm{D}} = n_{\mathrm{D}}^2$ is real, where n_{D} is the refractive index of the dielectric. A surface plasmon can be excited only if the conditions $\varepsilon_{\mathrm{M1}} < 0$ and $-\varepsilon_{\mathrm{M1}} - > \varepsilon_{\mathrm{M2}}$ are fulfilled. For example, gold, silver, and palladium are potentially applicable metals with negative values of $\varepsilon_{\mathrm{M1}}$ for frequencies corresponding to visible light [199].

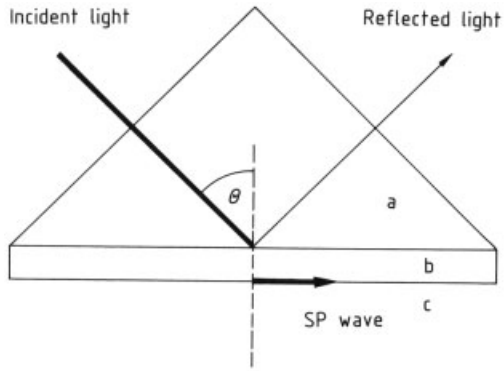

Figure 38. Attenuated total reflection coupler used to excite surface plasmons at the metal – dielectric interface; excitation is observed as a minimum in the intensity of the reflected light by scanning the angle of incidence
a) Prism; b) Metal; c) Dielectric

SP excitation can be accomplished in several ways. A very simple and often used approach is illustrated in Figure 38. RATHER [199] has described this as the *attenuated total reflection* (ATR) coupler. Here the base of a prism is coated with a thin metal film approximately 50 nm thick. Incident monochromatic light is reflected at this interface in such a way that a projection of the light vector $\boldsymbol{k}_0$ on the interface becomes $\boldsymbol{k}_{\mathrm{x}} = \boldsymbol{k}_0\, n_{\mathrm{p}} \sin(\Theta)$. If the resonance (Re) condition $\boldsymbol{k}_{\mathrm{x}} = Re\{\boldsymbol{k}_{\mathrm{sp}}\}$ is fulfilled, the evanescent field of the incident light excites an SP at the metal –dielectric interface, although this is only possible with light polarized parallel to the plane of incidence (p-polarized). The excitation

is observed as a sharp minimum in the intensity of the reflected light, and it can be detected by scanning the incident angle.

Using such an ATR device, small changes in $\varepsilon_D = n_D^2$ and the build-up of a thin layer at the metal – dielectric interface can easily be detected by measuring the shift in the *resonance angle* Θ.

For an immunosensing application the metal layer would be coated with appropriate antibodies, either immobilized covalently or embedded into a matrix layer. Upon exposure to a specific antigen the resonance angle should shift relative to that observed in an analyte-free environment due to the increase in layer thickness that accompanies binding of the analyte. For example, a thickness change of 0.1 nm results in a 0.01° change in the resonance angle, a shift that can be measured quite accurately (see also Section 3.3.1).

If an ATR device is to be used for chemical sensing, the metal surface must be coated with a thick dielectric layer of a reactive chemical. Chemical reaction within the layer then produces changes in n_D that can be detected by the corresponding shift in Θ.

2.3.2.4. Reflectometric Interference Spectroscopy

Gauglitz introduced reflectometric interference spectroscopy (RIFS) into the field of chemical and biochemical sensing [200], [201]. Figure 39 shows the underlying optical principle. The interference of the schematically shown partial light rays 2 and 3 leads to an intensity modulation of the overall reflected light spectra. This must be happening within the so-called coherence length of light after leaving the substrate. The coherence length of light using, e.g., a tungsten light bulb in the wavelength range of 400 – 1000 nm is about 20 µm. Therefore, the thickness of the interference causing layer has to be less than 20 µm and has to exceed the wavelength of the used light to give a measurable interference pattern. A layer thickness between of 0.5 and 10 µm with $n_2 \approx 1.4$ fits well this demand.

Light rays which undergo multiple reflections at the different interfaces contribute little because of their lower intensity. The scattered light should be as low as possible as with light from other sources. The equation for the interference pattern is:

$$I_{\text{all}}(\lambda) = I_1 + I_2(n_2) + I_3(n_2) + 2\cdot(I_2(n_2)\cdot I_3(n_2))^{1/2}\cdot \cos\{(4\pi/\lambda)\cdot(n_2\cdot d)\} \quad (20a)$$

Figure 40 shows a typical interference pattern after an alteration of the refractive index and the thickness of the interference layer. The shift $\Delta\lambda$ is caused by Δd and Δn_2 whereas the chance in the modulation depth Δg is caused by Δn_2 only.

Thus, the basic response of an RIFS based polymer or immunosensor is the alteration of the optical pathway. For the evaluation of the interference pattern several algorithms have been developed. After the determination of the minima and maxima the interference order k is determined using a series of calculations and curve

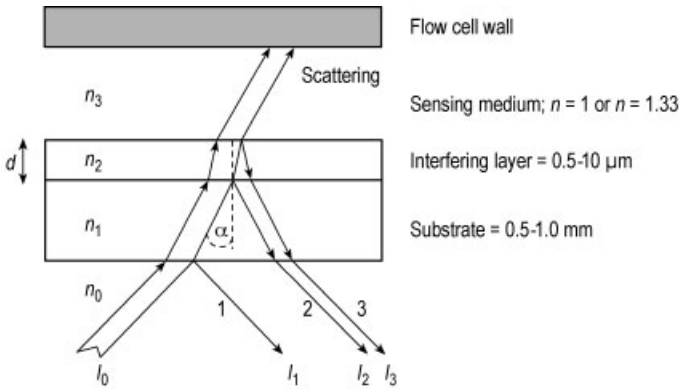

Figure 39. Schematic sketch of the principles of RIFS in a flow through set-up. Schematically shown light rays I_2 and I_3 can be re-collected with a fiber optic and lead to a monochromator with a CCD detector.

fitting techniques using newly developed software. Before a measurement is carried out a reference spectrum is obtained from an uncoated sensor including all spectral information except for the interference pattern. All further spectra obtained are divided by the reference spectrum to obtain the natural interference pattern.

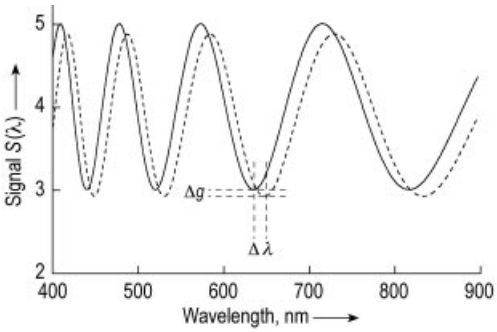

Figure 40. Observed interference pattern of RIFS normally recorded via a CCD array; comprehensive software [202] is needed to gain the needed information from this spectrum

2.3.3. Mass-Sensitive Devices

2.3.3.1. Introduction

Mass changes are usually detected with the aid of a gravimetric device, such as a balance. Ideally, balance sensitivity can be extended down into the microgram range, but the desire for further improvements in sensitivity and miniaturization of the experimental setup has led to the development of other mass-sensitive devices: the so-called quartz microbalance (QMB), which is the best-known system, and other, more recent devices such as the surface acoustic wave oscillator (SAW), the Lamb wave oscillator (LW), and the acoustic plate-mode oscillator (APM), which represent the present state of the science (Section 2.3.3.2). These devices are all based on the *piezoelectricity* of solids, (see Section 2.3.3.3), a phenomenon that permits the generation of vibrations or the propagation of waves. A piezoelectric substrate, usually a quartz plate, is stimulated to vibration by an oscillating electric circuit. This corresponds to an interconversion of electrical and mechanical vibrations (deformations of the crystal), resulting in a device that can be regarded as either an oscillator or a resonator.

Bulk waves, surface waves, and waves of other types (see Section 2.3.3.3) are associated with characteristic resonance frequencies corresponding to the most stable frequencies of vibration. Such a resonance frequency is itself highly dependent on the mass of the oscillating plate. A slight change in the mass of the plate will result in a considerable change in the resonance frequency, which can be measured very accurately. An oscillator of this type can therefore be used as a mass-sensitive transducer. Other factors that may influence the observed frequency shift include changes in the viscosity and density of surface layers as well as such physical parameters as forces, temperature, and pressure.

To construct a mass-sensitive sensor the surface of a piezoelectric material must be coated with an appropriate chemically or biochemically reactive layer capable of interacting with the proposed analyte and – ideally – recognizing it as well (see Section 2.3.3.4). Such an interaction causes the mass of the crystal to increase, resulting in a measurable shift in the resonance frequency. A relationship thus exists between the amount of analyte adsorbed or absorbed and the frequency shift of the oscillating system. Resonance frequency changes can be established very accurately, so very small changes in mass can be detected. A theoretical approach applicable to simple cases is described in [203].

To date, mass-sensitive devices have been used mainly for sensing analytes in the gas phase; they are also applicable in principle to liquids, but this presupposes a more sophisticated physical arrangement, a much more complicated signal control, and very careful interpretation. Many published papers have dealt with the potential use of QMB, SAW, and, more recently, APM and LW devices as mass-sensitive detectors in chemical and biochemical sensors, but no such device has yet established itself commercially – with the exception of QMBs employed for thin-layer thickness measuring, which cannot be regarded as chemosensors in the strictest sense of the term [204].

Historical Development. The first report concerning the use of piezoelectric resonance from a quartz crystal in a gravimetric microbalance is by J. Strutt (Lord Rayleigh), who in 1885 described a shift in resonance frequency accompanying an infinitesimal mass change in a mechanical oscillator [205]. Fundamental studies into the mass sensitivity of quartz resonators

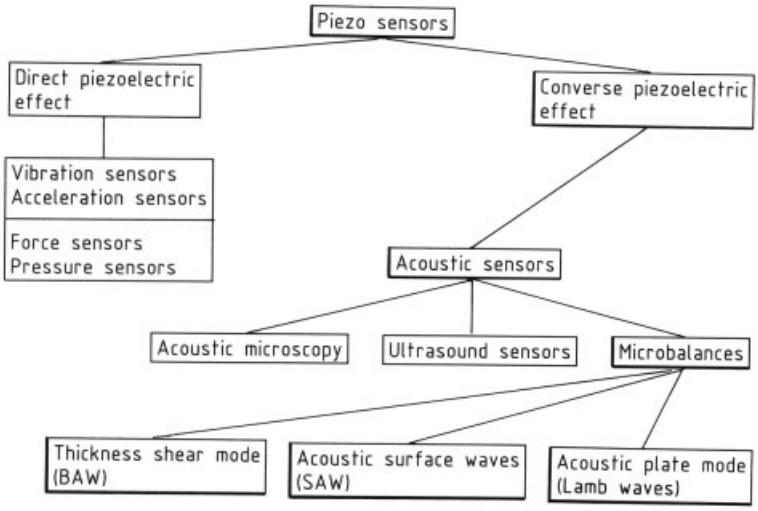

Figure 41. Selected applications of piezoelectric transducers with respect to both the direct piezoelectric effect (active transducer) and the converse piezoelectric effect (passive transducer) [222] (with permission from Springer-Verlag, Heidelberg)

and their potential application for measuring the thickness of thin layers were reported in 1959 by SAUERBREY [203], [206]. KING was the first to use a quartz microbalance "QMB" (in 1964) as a gas-phase detector in an analytical application [207], [208]. Numerous more recent papers discuss the fundamentals [209–216] and applicability [217–220] of bulk acoustic wave (BAW) piezoelectric resonators as chemical and biochemical sensors in the gas phase as well as in liquids, but physical devices for the measurement of pressure, temperature, or distance have attracted greater commercial interest [221], [222]. Nevertheless, quartz microbalances based on bulk acoustic waves do find frequent use in analytical research and development.

RAYLEIGH in an early paper [205] also noted the existence of waves that are propagated only in relatively thin surface layers, but it was not until 1979 that WOHLTJEN and DESSY adapted these so-called surface acoustic waves (SAWs) for incorporation into a chemical gas sensor [223–225]. The chief advantages of SAWs are increased sensitivity relative to BAWs of lower resonant frequencies and a greater potential for miniaturization. On the other hand, SAWs are associated with a variety of circuitry problems resulting from the relatively high operating frequencies required for adequate sensor sensitivity. A modification of this type of wave was introduced by CHANG and WHITE [226], [227], who adapted a SAW sensor by creating an exceptionally thin membrane-like piezoelectric region in the substrate directly beneath the acoustic path. This in turn led to the production of *Lamb waves*, which proved to be particularly suitable for chemical sensor applications, opening the way to the so-called plate-mode oscillator, a device representing the present state of the art.

2.3.3.2. Fundamental Principles and Basic Types of Transducers

Solids and other vibrationally active mechanical structures are capable of being stimulated by electrical, thermal, or optical means either to vibration or to the further propagation of waves. Piezoelectric solids, preferably quartz (SiO_2), are especially susceptible to emission and reception of mechanical vibrations and waves. The anisotropic nature of quartz supports various types of waves, each with its characteristic frequency, amplitude, and propagation properties.

Survey of the Different Types of Piezoelectric Substrates. A summary of the various types of transducers involving piezoelectric solids, including their chemical and physical applications, is presented in Figure 41. Acoustic gravimetric sensors can be categorized generally on the basis of their design, their wave mode (BAW, SAW, APM, LW), and their frequency

range. Figure 42 offers a schematic overview of the principles and design characteristics of various piezoelectric transducers. Criteria for establishing the usefulness of a particular vibrational system for the determination of mass include [209]:

1) The system should be readily excitable, preferably by electrical means
2) The system must lend itself to coupling with a frequency- (or period-) measuring device without significant disturbance
3) The system must be associated with sharply defined resonant frequencies so that the frequency (or period) of vibration can be determined precisely within a reasonably short period of time
4) In order to provide the required mass sensitivity, the mass-induced change in resonant frequency must be greater than the instability level of the resonant frequency and such uncertainties in the frequency (or period) as may accompany the measurement technique employed
5) Resonant frequency changes due to environmental disturbances (e.g., fluctuations in temperature or pressure, the presence of electrical or magnetic fields, or external mechanical stress) must be small relative to those caused by mass changes
6) An equation must be available in analytical form relating mass changes to the corresponding shifts in resonant frequency

Bulk Acoustic Wave Transducers. Waves of the bulk acoustic type (BAWs) can be invoked for microgravimetric applications through the use of either AT- or BT- cut plates derived from a single crystal of α-quartz, where "AT" and "BT" refer to cuts with specific orientations relative to the main axis of the crystal (see below and [229]). Figure 43 A represents a schematic diagram of such a quartz crystal with its associated electrodes and sensing layers. This diagram also indicates how a species to be detected might be absorbed into the sensing layer. Both sides of the quartz are in contact with electrodes. The corresponding electrodes are usually 300 – 1000 nm thick and 3 – 8 mm in diameter, and are made of gold, nickel, silver, or aluminum [217]. Introduction of oscillating electrical energy through an appropriate electrical circuit causes the mechanical structure to begin vibrating. In the case of AT- cut quartz the crystal vibrates in the thickness shear mode, as shown in Figure 43 B (see also Fig. 58).

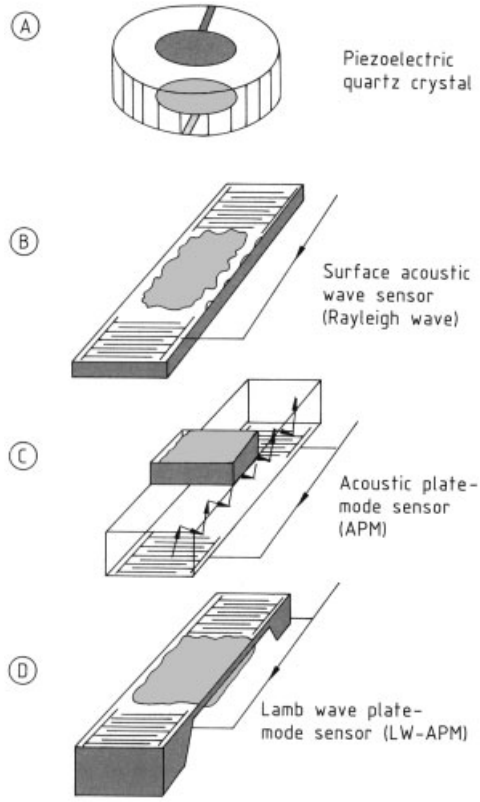

Figure 42. Schematic overview of acoustic gravimetric devices, illustrating the principles of wave propagation and design [228] (with permission from VDE-Verlag, Berlin) A) BAW; B) SAW; C) APM; D) LW substrates

Theory. Microgravimetric application of acoustic devices presupposes a quantitative relationship between an observed relative shift in the resonance frequency and an added mass. In the case of a quartz plate vibrating in a liquid an extended approach is required with respect to changes in the viscosity and density. The first useful relationship of this type was developed by SAUERBREY, who treated the added mass as an "added thickness" of the oscillator. The relationship itself was based on a quartz crystal vibrating in its thickness shear mode in the gas phase (Fig. 44). The observed oscillation frequency (f_q) is inversely proportional to the thickness t_q of the crystal; $f_q = N/t_q$, where N is a frequency constant ($N = 0.168$ MHz cm for AT- cut quartz at room temperature). The mass

of the crystal is $m = \varrho A t_q$, where ϱ is the crystal density and A its cross-sectional area. In this derivation the mass increment due to a foreign mass Δm is treated as an equivalent change in the mass of the crystal. Invoking the approximation $\varrho_q = \varrho_f$ leads to the equation:

$$\Delta f_q = -f_q \frac{\Delta m}{m} \tag{21}$$

which finally provides the following expression for an AT- cut crystal:

$$\Delta f_q = -2.3 \times 10^6 f_q^2 \frac{\Delta m}{A} \tag{22}$$

where f_q is expressed in megahertz, Δm in grams, and A in square centimeters. Since f_q for a quartz crystal is typically 10 MHz, and a frequency change of 0.1 Hz can be readily detected by modern electronics, it is possible in this way to detect mass changes of as little as about 10^{-10} g/cm^2 in the gas phase (!) [230].

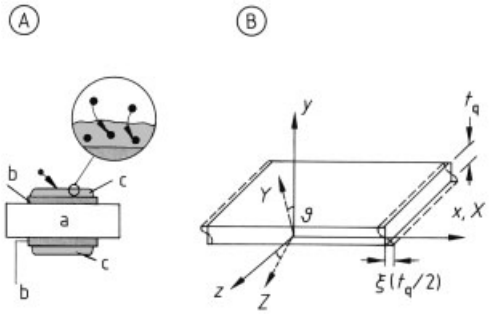

Figure 43. Schematic diagram of a quartz crystal augmented with electrodes and sensing layers, illustrating the way in which the species to be detected is absorbed into the sensing layer (A) and the ideal thickness shear mode of a quartz plate (B) [230], [203]
a) Quartz crystal; b) Electrodes; c) Sensing layers x, y, z = Coordinates of the quartz plate; X, Y, Z = (IEEE-standard) coordinates for a single crystal of α-quartz, where X is the polar axis and Z is the optical axis; t_q = Thickness of the quartz plate, where the wavelength $\lambda = 2\,t_q$; ξ = Shear amplitude; ϑ = Angle of rotation

Lu has developed a model involving propagation velocities for sound waves in the substrate (for quartz, v_q) and in the sensing layer (v_f) [216]. Reflection and refraction occur at the interface between the crystal (q) and the film (f) (Fig. 44) analogous to the optical reflection and refraction observed at a boundary between two materials with different optical densities. The shear velocities in the quartz crystal and the film coating, respectively, are:

$$\text{quartz } v = \sqrt{\frac{\mu_q}{\varrho_q}}\,\text{q} \tag{23}$$

$$\text{film } v = \sqrt{\frac{\mu_f}{\varrho_f}}\,\text{f}$$

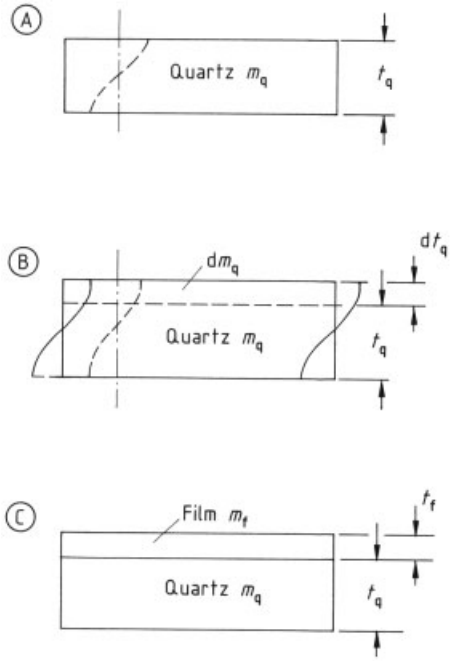

Figure 44. Shear wave (with permission from Plenum Publishing, New York)
A) In a quartz crystal; B) After an incremental increase in the crystal thickness; C) With the addition of a film m_q, t_q = Mass and thickness of the quartz; m_f, t_f = Mass and thickness of the film

Given a knowledge of the shear moduli μ_f and μ_q, Equation (23) can be used to calculate the shear velocity of the wave. Ideal wave-propagation behavior is observed with a crystal upon which such a film has been deposited provided there is a negligible amount of frictional loss. In this case the resonance condition is valid, which leads to Equation (24):

$$\tan\left(\frac{\pi \cdot f_c}{f_q}\right) = \left(\frac{\varrho_f \cdot v\text{f}}{\varrho_q \cdot v\text{q}}\right) \tan\left(\frac{\pi \cdot f_c}{f_f}\right) \tag{24}$$

where $f_c = \omega/2\pi$ is the resonance frequency of the crystal upon which material has been deposited. The terms $\varrho_f \cdot v_f$ and $\varrho_q \cdot v_q$ are the acoustic impedances Z_f and Z_q of the film and crystal, respectively. Their ratio, $Z = Z_f/Z_q$, is an important parameter with respect to acoustic

matching of the materials. For optimum resonance conditions Z should be as close as possible to unity, because its value affects the resilience of the whole assembly, and maximum stability corresponds to $Z \sim 1$ [231].

For example, Z has a value of 1.08 for a piezoelectric crystal covered with aluminum electrodes (shear-mode impedance of SiO_2: $Z_{SiO^2} = 8.27 \times 10^6\ \mathrm{kg\,s\,m^{-2}}$; that of Al: $Z_{Al} = 8.22 \times 10^6\ \mathrm{kg\,s\,m^{-2}}$), indicative of the good match associated with this material combination. Deposition of gold electrodes on an SiO_2 surface leads to a Z-value of 0.381, which is barely acceptable (shear-mode impedance of gold: $Z_{Au} = 23.2 \times 10^6\ \mathrm{kg\,s\,m^{-2}}$) [231].

From a practical point of view it is important to develop a relationship involving more experimentally common parameters influencing the frequency response. The frequency response of a piezoelectric sensor in the presence of a gas at a particular pressure can be related generally to three effects: the hydrostatic effect (p), the impedance effect (x), and the sorption effect (m):

$$-\frac{\Delta f}{f_0} = \left(\frac{\Delta f}{f}\right)_p + \left(\frac{\Delta f}{f}\right)_x + \left(\frac{\Delta f}{f}\right)_m \tag{25}$$

For gravimetric measurements and sensor applications, the latter effect (that is, to added mass) is the most important, and the other two effects can be viewed as nonspecific interferences. Because both of these terms are three orders of magnitude smaller than the mass term, the mass relationship alone is sufficient for characterizing measurements in the gas phase. However, situations do exist in which the other two effects cannot be ignored. For instance, if the pressure is changed from that of the surrounding air to a vacuum the first term is no longer negligible, and in the case of liquids any change in viscosity or density alters the impedance, so this term must be taken into consideration.

Other theoretical approaches relevant to applications in the gas phase and dealing with thickness-shear mode acoustic wave quartz sensors are presented in [213], [232–237].

Bulk Acoustic Waves in Liquids. When a quartz crystal is placed in a liquid, there is a significant change in both the density and the viscosity of the surrounding medium. Both become much larger than in the gas phase, and the consequences must be taken into account with respect to the second term in Equation (25). Thus, the Sauerbrey equation is no longer applicable if the viscoelastic properties of the contacting medium (liquid and/or film) change during the course of an experiment. The liquid now represents an additional mass load, which produces coupling between the substrate elastic shear wave and the liquid. Kanazawa derived the first theoretical relationship linking the properties of a liquid to the frequency response of a piezoelectric crystal. In a liquid medium, the shear motion of the crystal causes motion to occur in the adjacent layer of solvent molecules (usually water). Up to a layer thickness of approximately 1 µm the solvent molecules move with a certain degree of "slip," which corresponds to a phase shift in the third and subsequent layers. Therefore, vibrational energy from the oscillator is continuously dissipated into the liquid. If only one side of the crystal is coupled to the liquid, the result is a standing wave perpendicular to the substrate surface. For liquids with Newtonian behavior this wave can be described theoretically as a damped wave. As is usual for this type of wave, a decay constant k is used to describe the magnitude of the damping. In this case the constant k can be related to $(\omega/2\,\nu)^{1/2}$, where ν is the kinematic viscosity. The damping distance (limit of wave propagation) is on the order of micrometers, while the frequency shift can be expressed [238] as:

$$\Delta f = f_0^{3/2} \left(\frac{\nu}{\pi \mu_q \varrho_q}\right)^{1/2} \tag{26}$$

Figure 45 shows the effect of microscopic kinematic viscosity at the crystal – liquid interface, where the magnitude of the observed frequency shift has been plotted against the corresponding increase in density for a glucose solution.

The frequency shift for quartz is highly dependent on changes in the viscosity and density of a surrounding liquid. This situation can be discussed in a straightforward way on the basis of the *quartz equivalent circuit* [231], [239], shown in Figure 46, which consists of the following elements:

1) The $L' - C'$ combination, which determines the resonance frequency f_m of the motional arm
2) A parallel capacitance C_p, caused by the electrodes on the quartz and any stray capacitance

3) A resistive component R'

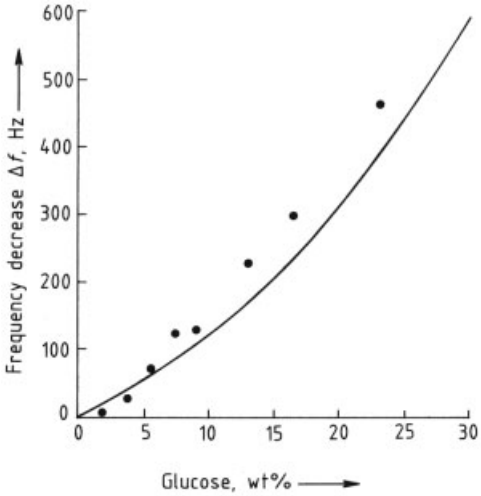

Figure 45. The effect of kinematic viscosity on resonance frequency as a function of glucose concentration (in water) for a 5-MHz quartz crystal
——= Calculated; • = Measured (with permission from Am. Chem. Soc., Washington)

The resistance R' is a measure of mechanical losses from the vibrating quartz, and it therefore reflects the viscoelastic properties of the contacting liquid [240], [241] or a deposited film [242]. Surface roughness also influences the value of R' [243]. This is often not considered in certain applications. It is highly unlikely that in the case of a directly sensed immuno-surface-reaction the binding events are uniformly smeared out. Likewise, immobilized peptide molecules may alter the tertiary shape by changing the ionic strength, resulting in a different surface roughness. Changes in the vibrating mass correspond to changes in the self-inductance L' [244], [245]. Thus, the resonance frequency of the motional arm (f_m; see Fig. 46) changes in proportion to the mass. In deriving Equation (26) KANAZAWA considered only those L' values for the motional arm that contribute to a frequency shift, but in a real experiment any change in $\Delta R'$ is normally accompanied by a change in L' (typical values for R' in air and in water are $7\,\Omega$ and $350\,\Omega$, respectively, for a 6 MHz quartz crystal). Thus, when a quartz crystal is immersed in liquids of different viscosities, both L' and R' can be expected to vary. Use of an impedance analyzer makes it possible to determine the increase in R' associated with operation of a quartz crystal in air versus a liquid. The frequency shift of the motional arm can be calculated from the values of C' and L' using the formula [239]

$$f_m = 1/[2\pi(L'\,C')^{1/2}].$$

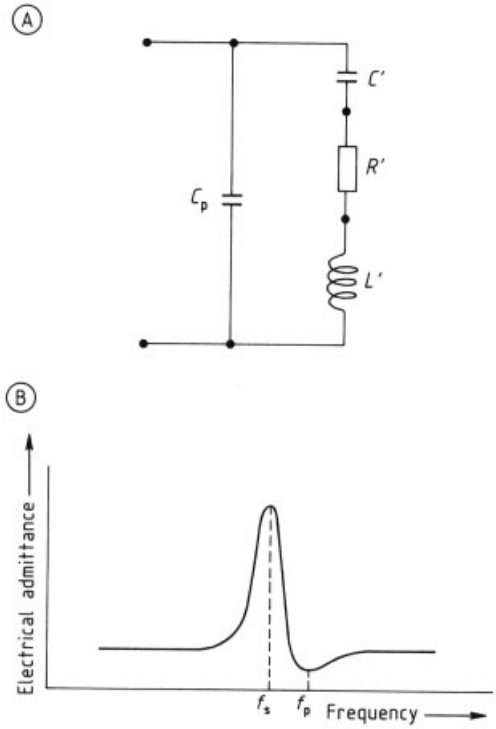

Figure 46. Equivalent electrical circuit (A) and resonance frequency of a mounted piezoelectric crystal oscillator (B) [231], [239]
C_p = Capacitance of the mounted crystal; C', L' = Rigidity and mass of the added material; R' = Equivalent to the mechanical loss, which also represents the acoustic (mechanical) load; f_s, f_p = Series and parallel resonance frequencies
(with permission from Plenum Publishing, New York)

SAW Transducers. Surface acoustic wave (SAW) sensors have been designed with sensitivity to many different physical quantities, including force, acceleration, hydrostatic pressure, electric field strength, dew point, and gas concentration [246]. The operating principle of an SAW for use with gases is conceptually quite simple. A surface acoustic wave – a periodic deformation perpendicular to the material surface – is transmitted across the surface of some appropriate solid that has been exposed to the atmosphere subject to analysis. With a homogeneous substrate, usually a piezoelectric solid

like quartz or $LiNbO_3$, the required SAW (sometimes also called a Rayleigh wave) is generated by means of an interdigital transducer (IDT).

In contrast to a quartz microbalance, the requisite IDT components are deposited only on one side of the crystal. These take the form of planar, interleaved, metal electrode structures, (with permission from Plenum Publishing, New York) where adjacent electrodes are supplied with equal but opposite potentials (Fig. 47). Application of a time-varying r. f. potential causes the crystal to undergo physical deformations, and if these are confined to the surface region of the crystal, the result is a surface acoustic wave (Fig. 47 B) [248]. Such vibrations will interfere constructively only if the distance $L/2$ between two adjacent "fingers" (see Fig. 37 A) is equal to one-half the elastic wavelength. The frequency $f = v_R/L$ that corresponds to such a cumulative effect is called the *synchronous frequency* or *resonance frequency* (f) [230].

The waves that are generated have their mechanical and electrical components in a single plane, the *sagittal* plane, which is perpendicular to the substrate surface. The phase velocity (v_R) and the amplitude (A) of an SAW are determined by elastic, piezoelectric, dielectric, and conductive properties of the substrate as well as by its mass. If one of these parameters can be modulated in an appropriate way by the quantity of interest, a sensing device can be constructed. Such a modulation could be induced in the transducer itself and/or in the transmission region (waveguide or *delay line*). With layered substrates the physical properties of each layer as well as the thicknesses of the layers determine the phase velocity and the amplitude of the resulting SAW [230]. Various energy-loss mechanisms also operate on the propagating SAW: scattering loss due to finite grain size, thermoelastic loss due to nonadiabatic behavior of the acoustic conductor, viscous loss caused by the dissipation of energy in a direction perpendicular to the transducer – environment interface, and hysteresis absorption due to irreversible coupling between SAW energy and an adsorbate [231]. Any addition of mass during the sensing step may change the magnitude of one or more of these phenomena, and may also lead to changes in v_R and A, especially a frequency shift (Δf) or a relative frequency shift ($\Delta f/f$). An exact description of the physical processes involved in the generation, propagation, and detection of SAWs is complex, and is available from such sources as [248], [249].

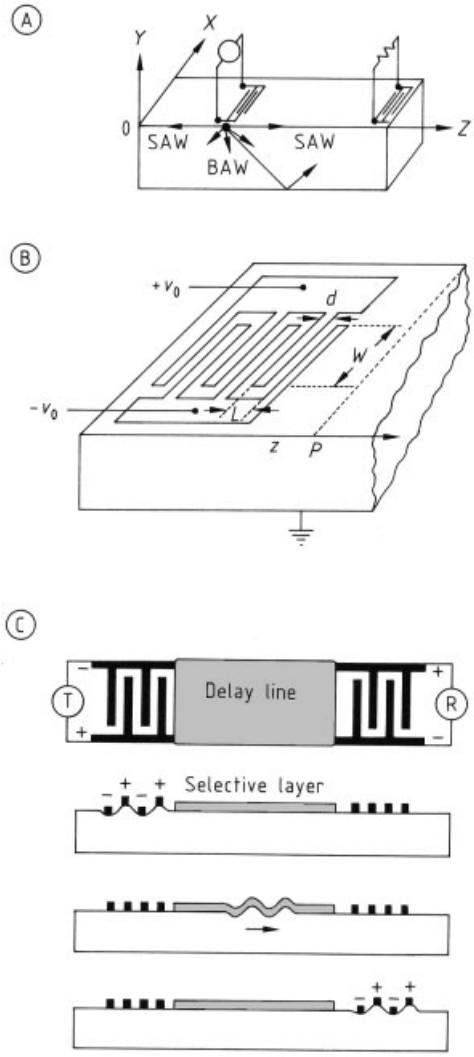

Figure 47. Interdigital metal pattern of a uniform transducer; the IDT behaves like a sequence of ultrasonic sources or receivers [247]
A) An SAW device with input and output transducers; B) A uniform transducer defined by half-spatial period L, electrode width d, and aperture width W; the observation line P used for derivation of the impulse response is at a distance $z + L$ from the first source; C) Schematic diagram of an SAW sensor with a transmitter T, a receiver R, and a chemically selective layer deposited in the form of a delay line [231] (with permission from Plenum Publishing, New York)

The considerations presented above provide the basis for using SAWs in gas-monitoring devices, because observed frequency shifts are pro-

portional to added mass. However, the characteristic frequency of an SAW is also sensitive to changes in temperature or pressure. The ideal substrate would be one with a zero temperature delay coefficient but a high piezoelectric coupling effect. Despite extensive materials research efforts, it is currently necessary to choose between these parameters; e.g., ST- cut quartz is the preferred substrate if a zero temperature coefficient is most important, and cuts of lithium niobate are selected for high piezoelectric coupling.

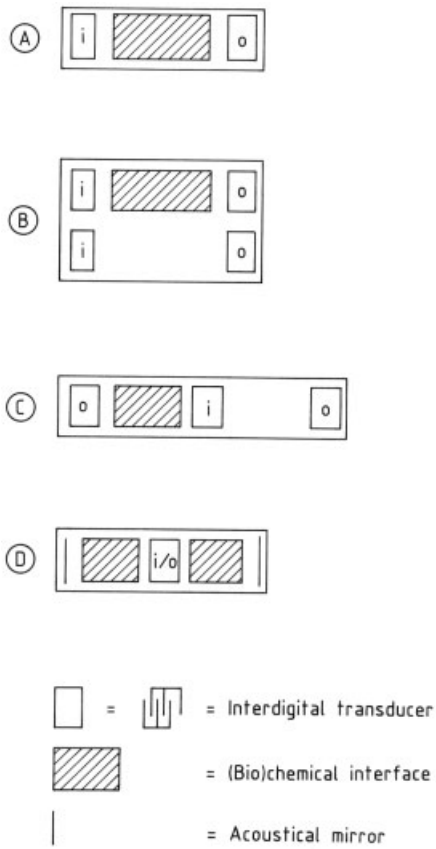

Figure 48. Various possible sensor configurations for SAW chemical sensors [230]
A) Single delay line; B) Dual delay line; C) Three-transducer device; D) Resonator
i = Input transducer; o = Output transducer

Sensor Configurations. A single SAW sensor may sometimes suffice for a gas sensing application (Fig. 48 A), but often the sensors display undesirable sensitivity to such effects as temperature, pressure, and ambient humidity. Overcoming these problems requires the establishment of a relative signal output. For example, VETELINO et al. have described a dual-line SAW gas sensor [230], [250], [251]. One delay line in this case is used for measuring, while the other acts as a reference (see Fig. 48 B). The advantage of a dual delay-line configuration is that it transforms the output signal into a relative change in oscillator frequency, which can be attributed only to the effects of the analyte. D'AMICO et al. have described a system based on a three-transducer delay line (see Fig. 48 C). This device also has two dual-delay lines, but the paths are established in this case by one input IDT and two output IDTs. Again, one path of the delay line is coated and the other remains uncoated as a way of minimizing the effects of both temperature and pressure [252].

The first use of an *SAW resonator* sensor was reported by MARTIN et al. [253]. The corresponding resonator incorporated a single IDT located between two acoustically reflecting mirrors. The distance between the mirrors was so adjusted as to ensure constructive interference between successive reflections, resulting in a maximum for one particular frequency (Fig. 48 D).

Theory. The SAW velocity v_{R} can be affected by many factors, each of which is associated with a potential sensor response:

$$\frac{\Delta v}{v_{\mathrm{R}}} = \frac{1}{v_{\mathrm{R}}}\Big(\frac{\partial v}{\partial m}\Delta m + \frac{\partial v}{\partial c}\Delta c + \frac{\partial v}{\partial \sigma}\Delta\sigma + \frac{\partial v}{\partial \varepsilon}\Delta\varepsilon + \frac{\partial v}{\partial T}\Delta T + \frac{\partial v}{\partial P}\Delta P + \dots\Big) \tag{27}$$

where v_{R} is the phase velocity unperturbed by such external factors as m, c, σ, and ε (the mass, elastic constant, electrical conductivity, and dielectric constant of the solid medium, respectively), as well as T and p, the environmental temperature and pressure [246]. Three types of films have been considered for use as reactive layers:

1) Nonconductive isotropic overlay film
2) Electrically conducting overlay film
3) Metal-oxide semiconducting film

A theoretical treatment can be developed for each of these film types with respect to its properties, involving, in the nonconductive case, the thickness t_{f}, the mass density ϱ, λ (the Lamé constant), and μ (the shear modulus).

Regarding the relationship $\Delta v/v = -\Delta f/f$ between SAW velocity and the variation in the SAW dual-delay line oscillator frequency, perturbation analysis leads to the following result for the relative frequency shift [249]:

$$\Delta f = (k_1 + k_2)\, f_0^2 t_f\, \varrho - k_2 f_0^2 t_f \left[\frac{4\mu\,(\lambda+\mu)}{v_r^2\,(\lambda+2\mu)}\right] \quad (28)$$

where k_1 and k_2 are material constants for the SAW substrate. Additional information with respect to these equations is available from [224], [246], [249].

Thus, the film mass per unit area is $t_f\,\varrho$. If chemical interaction does not alter the mechanical properties of the film, the second term in Equation (28) can be neglected, in which case the frequency shift Δf can be attributed exclusively to the added mass:

$$\Delta f = (k_1 + k_2)\, f_0^2 t_f \varrho \quad (29)$$

Coating a polymer film 1 µm thick onto a quartz SAW sensor operating at 31 MHz should therefore cause a frequency shift $\Delta f = -130$ kHz [231].

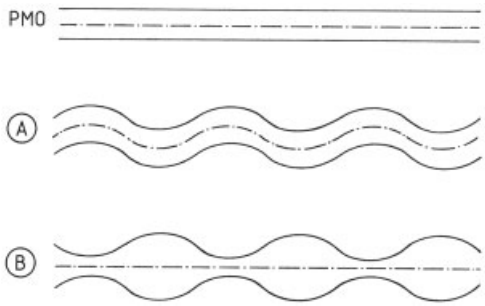

Figure 49. Lamb waves propagating in a thin plate-mode oscillator (PMO) [226], [231] (with permission from Plenum Publishing, New York)
A) Symmetric waves; B) Antisymmetric waves

Plate-Mode Oscillators (PMOs). *Lamb Wave Oscillators.* In all the SAW devices discussed above, the acoustic wave propagates in a slab of material whose thickness is infinitely larger than the wavelength λ of the propagating wave. When the thickness of the plate is reduced to such an extent that it becomes comparable to λ, the entire plate becomes involved in the periodic motion, producing a symmetric and antisymmetric *Lamb wave* (LW) (Fig. 49). This behavior is observed in plate-mode oscillators with thicknesses of a few micrometers. A cross-sectional view through a typical PMO is shown in Figure 50.

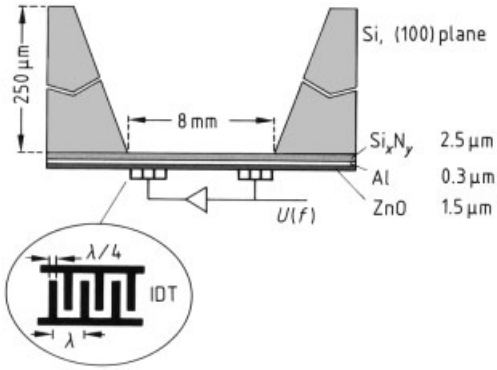

Figure 50. Cross section through a Lamb wave PMO, including the corresponding interdigital transducers [228]

Certain performance factors make the LW-PMO potentially attractive as a chemical sensor. Both surfaces contribute to the signal, so the observed sensitivity is greater than for a corresponding SAW device. However, the most important advantage follows from the fact that the velocity of the lowest-order wave in the antisymmetric mode is much lower than in the case of the corresponding SAW oscillator. This may be important for applications involving mass-sensing in liquids, which are problematic with high-frequency SAW devices. As the frequency of the Lamb wave decreases below the velocity of the compressional wave in the liquid, energy loss in the perpendicular direction decreases as well. A Lamb wave also has a greater frequency dispersion than the corresponding surface acoustic wave, resulting in increased resolution and sensitivity.

However, there are certain practical problems. The fragility of the thin plate and its consequent sensitivity to external pressure is particularly serious. Performance analyses as well as details regarding fabrication of the device can be found in the original paper [227].

Acoustic Plate-Mode Oscillators and Love Plate Devices [254]. These two types of acoustic PMOs, like BAW sensors, are characterized by surface-guided, shear-horizontally (SH) polarized waves, which have a horizontally polarized component of displacement (see also Fig. 58). In principle the performance of such APMs is similar to that of the Lamb wave sensors described above, but the thickness of the substrate

is greater, so waves propagate from the surface into the bulk of the material. Energy is confined mainly in the bulk of the plate as the wave propagates through multiple reflections, generating displacements in both the upper and lower surfaces. The Love plate sensor includes a waveguide structure, whereby the SH wave is confined within an elastic layer deposited on a SAW substrate capable of supporting SH waves. Figure 51 provides cross-sectional views of APM and Love plate devices.

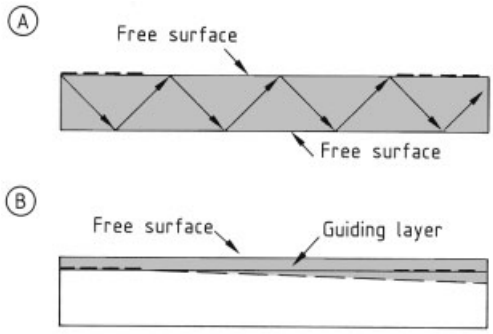

Figure 51. Cross-sectional view of (A) APM and (B) Love plate devices [254]

Acoustic sensors based on SH waves have become rather widespread due to their ability to function in liquid media.

Relative Sensitivities of BAW, SAW, LW, APM, and Love Plate Devices. Because of excitation and propagation differences among the various transducers described, the mass-deposition sensitivity of the several devices would be expected to vary considerably, with some geometries proving more attractive than others for particular sensing applications. Table 10 offers a performance comparison for these transducers. Table 10 also summarizes the calculated frequency-to-mass sensitivities for the various devices. Sensitivity variations can be explained on the basis of energy confinement on the sensing surface and a particular device's operating frequency. A decrease in the thickness of the substrate leads to a sensitivity increase, which can be realized with either an acoustic plate-mode or a Lamb wave device. These are therefore associated with the highest sensitivities, followed by the Love plate, which also has a very thin layer. Less sensitivity is available with BAW devices, for which the sensitivity depends on the thickness of the crystal.

The above statements have to be corrected because of very recent developments. With the commercial availability of a 155 MHz BAW quartz at reasonable prices the QMB should be more favored. According to the Sauerbrey equation the new BAW device shows a much higher sensitivity than the traditional 10 MHz QMB. Table 11 summarized recent data of the different noise levels published in the literature. By this it becomes evident that the new 155 MHz device will be the transducer of choice when an extremely sensitive gas sensor transducer is needed. This sensor also shows almost no drift. Thus drift compensation methods often used with SAW devices are no longer needed. The price of only a few Euro for this mass-produced unit will make it highly competitive in sensor applications.

2.3.3.3. Theoretical Background

Piezoelectricity. *The Piezoelectric Effect.* Because all acoustic gravimetric sensors are based on the phenomenon of piezoelectricity, it seems appropriate to discuss briefly the effect itself. Piezoelectricity was first observed by the Curie brothers (Jaques and Pierre) in 1880 [258]. It is a reversible phenomenon, consisting of linear electromechanical interactions between mechanical and electrical properties in certain crystals (Fig. 52). The effect is generated, as already mentioned, by application of an AC potential to the piezoelectric material via contact electrodes. One can distinguish between bulk- and surface-generated AW (BAW and SAW, respectively) when the electric field is applied across the substrate or only at its surface, respectively.

A piezoelectric effect occurs only with those ionic crystalline solids whose crystals contain a polar axis along which the physical properties are not constant. The *direct* piezoelectric effect is observed when an applied force (F) produces an electrical polarization (P), and a *converse* (reciprocal) piezoelectric effect results when an applied electrical field (E) induces a strain (S, deformation of the crystal) (Fig. 53).

A simplified view of piezoelectricity at the microscopic level is provided by the example of α-quartz (Fig. 54). Here a net dipole mo-

Table 10. Comparison of sensitivities and other characteristics of acoustic gravimetric sensors [228], [255], [256]; international symbols (short form) have been used to denote specific cuts of the various materials

Acoustic wave type	Sensitivity formula, $cm^2/\mu g$	Calculated sensitivity, $cm^2/\mu g$	Frequency, MHz	Thickness, µm	Material
Bulk wave resonator (BAW)	$S_m = -\frac{2}{\varrho\lambda} = -\frac{1}{\varrho t_f}$	− 23	10	165	AT- cut quartz
Surface (Rayleigh, SAW)	$S_m = -\frac{K(\sigma)}{\varrho\lambda}$	− 129	100	NA∗	ST- cut quartz, $LiNbO_3$
	$0.8 < K < 2.2$	− 516	400	NA∗	$K = 1.16$; $\sigma = 0.35$
	$0 < \sigma < 0.5$				
Lamb plate (LW), A_0 mode	$S_m = \frac{1}{\varrho t_f}$	− 450	4.7	3	ZnO on SiN
		− 951	2.6	(not given)	
Acoustic plate mode (APM, SH)		− 3000	NA∗	1	ST- cut quartz
		− 300	NA∗	10	
		− 30	NA∗	100	
Love plate (transverse surface)		− 37	100	NA∗	$LiNbO_3$
		− 182	500	NA∗	

∗ not available.

Table 11. Transducer noise of the new polymer-coated 155 MHz BAW in air in comparison with mass-sensitive an SAW-transducer of different operating frequencies from literature values. The integration times of the frequency counters were in the range between 1.0 s and 2.0 s; resonance frequency f_0

Transducer	f_0, MHz	Research group/[257]	Noise, Hz
AT-quartz			
QMW; self-build	155	Cammann, Reinbold [257]	0.2 – 1.0
SAW $LiNbO_3$			
Plessey Semiconductors	67	Wohltjen et. al.	9 – 10
SAW_{quartz} covered with different Me(II)-	39		11
phthalocyanines, leading to	52	Nieuwenhuizen et al.	7
different noise	78		17 – 29
Xensor Integration, Delft	80	Göpel et al.	1 – 5
Microsensor Systems, Inc.	158	Zellers, Patrash et al.	11 – 15
Microsensor Systems, Inc.	158	Grate, Mc Klusty et al.	5 – 7
Sawtek, Inc. Orlando, FL	200	Grate, Mc Klusty et al.	3 – 10
Microsensor Systems, Inc.	300	Grate, Mc Klusty et al.	26 – 55
Microsensor Systems, Inc.	400	Grate, Mc Klusty et al.	15 – 40
R2632 Siemens; self-build	433	Göpel et al.	1 – 5

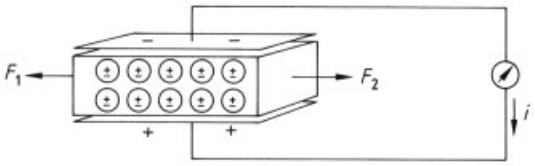

Figure 52. Schematic illustration of the way mechanical and electrical energy can be interconverted with the aid of an appropriate crystal [259]
F_1, F_2 = Mechanical forces; i = Electrical current

ment will arise if the hexagonal structure is stretched or compressed along a direction parallel (Fig. 54 B) or perpendicular (Fig. 54 C) to one of the three in-plane symmetry axes. When symmetry is achieved, the net dipole moment of the molecule vanishes.

Structural Aspects of Piezoelectric Solids. In order for a crystal to be piezoelectric it must be noncentrosymmetric; i.e., it cannot contain a center of symmetry. Solids are usually grouped for structural characterization purposes into seven crystal systems (from lowest to highest symmetry: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic), which can in turn be divided into 32 point groups, depending upon point symmetry. Of these, 11 classes are centrosymmetric and 20 are piezoelectric. An exceptional case is class

432 from the cubic crystal system, because it is neither centrosymmetric nor piezoelectric. In addition to the 21 noncentrosymmetric classes, 11 classes can be distinguished on the basis that they possess no plane of symmetry. This means that they are associated with both right- and left-handed forms, which cannot be interconverted by simple rotation, a phenomenon known as enantiomorphism [261].

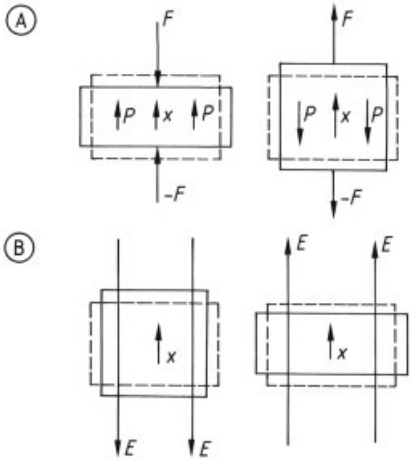

Figure 53. Schematic representation of the direct piezoelectric effect (A) and the converse (reciprocal) piezoelectric effect (B) for an X-quartz plate (right-handed quartz) [222]
F = Applied force; P = Polarization; E = Applied electrical field
(with permission from Springer-Verlag, Heidelberg)

Table 12 summarizes on the basis of international point-group symbols the structural distribution with respect to piezoelectric classes, centrosymmetric classes, and enantiomorphism. For example, $LiNbO_3$, which belongs to the class $3\,m$, is piezoelectric, but not enantiomorphic, whereas α-quartz, in class 32, is both piezoelectric and enantiomorphic. The latter characteristic is illustrated in Figure 55.

Not every solid material associated with a piezoelectric class is suitable for practical applications, because in some cases the piezoeffect is too weak. Large piezoelectric coefficients are a prerequisite for an electromechanical transducer, like those found, for example, with α-quartz or $LiNbO_3$. Therefore, these are the solids most often used as mass-sensitive plates or substrates.

Table 12. Distribution of crystal structures (point groups) with respect to centrosymmetric, piezoelectric, and enantiomorphic classes [261]

Crystal system	Centro-symmetric classes	Piezoelectric classes	Classes with enantiomor-phism
Triclinic	$\bar{1}$	1	1
Monoclinic	$2/m$	$2, m$	2
Orthorhombic	mmm	$222, mm2$	222
Tetragonal	$4/m, 4/mmm$	$4, \bar{4}, 422, 4mm, 42m$	4, 422
Trigonal	$\bar{3}, \bar{3}m$	$3, 32, 3m$	3, 32
Hexagonal	$6/m, 6/mmm$	$6, \bar{6}, 622, 6mm, 6m2$	6, 622
Cubic	$m3, m3m$	$23, \bar{4}3m$	23, 432

Theory and Physics of Piezoelectricity. The discussion that follows constitutes a very brief introduction to the theoretical formulation of the physical properties of crystals. If a solid is piezoelectric (and therefore also anisotropic), acoustic

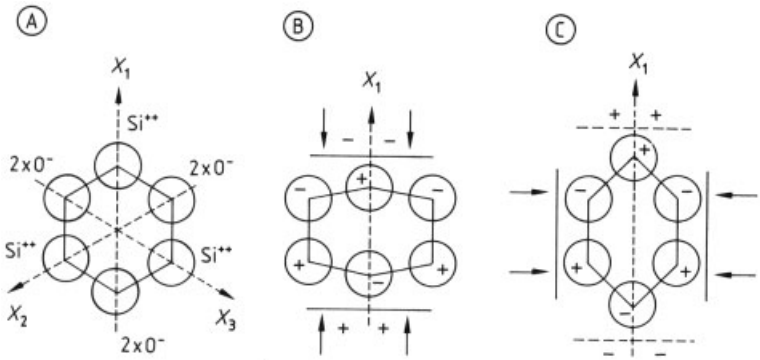

Figure 54. Physical interpretation of the piezoelectric effect [260]
A) Simplified microscopic model of a quartz crystal lattice; B) Longitudinal effect; C) Transverse effect
"Si^{++}" and "$2\,O^-$" refer to centers of gravity (circles) for charges associated with the two types of atoms, where the tetrahedral "SiO_4" structure has been projected onto a plane (as a hexagon)

displacement and strain will result in electrical polarization of the solid material along certain of its dimensions. The nature and extent of the changes are related to the relationships between the electric field ($\boldsymbol{E}$) and electric polarization ($\boldsymbol{P}$), which are treated as vectors, and such elastic factors as stress $\boldsymbol{T}$and strain ($\boldsymbol{S}$), which are treated as tensors. In piezoelectric crystals an applied stress produces an electric polarization. Assuming the dependence is linear, the direct piezoelectric effect can be described by the equation:

$$\boldsymbol{P}_{\mathrm{i}}=d_{\mathrm{ijk}}\boldsymbol{T}_{\mathrm{jk}} \text{ or } \boldsymbol{P}_{\mathrm{i}}=e_{\mathrm{ijk}}\boldsymbol{S}_{\mathrm{jk}} \tag{30}$$

The quantities d_{ijk} and e_{ijk} are known as the *piezoelectric strain* and *piezoelectric stress* coefficients, respectively.

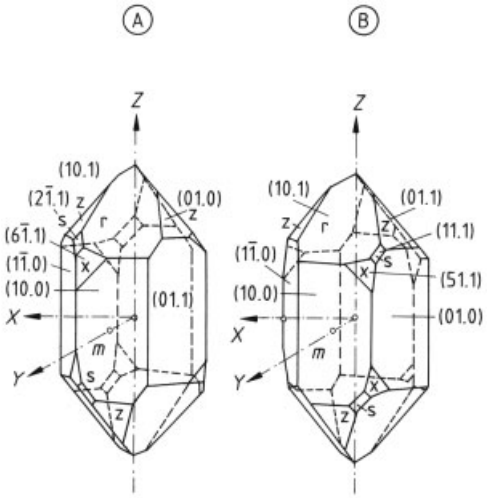

Figure 55. Major crystal surfaces of α-quartz together with their Bravais – Miller indices (hk.l) (with permission from The Institute of Electrical and Electronics Engineers, New York)
A) Left-handed form; B) Right-handed form; Cartesian coordinates are specified in accordance with the IEEE standard of 1987 [229]

Conversely, if such a crystal is placed in an electrical field it will become deformed, a phenomenon known as the converse or reciprocal piezoelectric effect (Fig. 53 B). The contributions of stress and strain in this case can be expressed:

$$\boldsymbol{S}_{\mathrm{jk}}=d_{\mathrm{ijk}}\boldsymbol{E}_{\mathrm{i}} \text{ or } \boldsymbol{T}_{\mathrm{ij}}=e_{\mathrm{ijk}}\boldsymbol{E}_{\mathrm{i}} \tag{31}$$

Since the stress and strain tensors are symmetric, the piezoelectric coefficients can be converted from tensor to matrix notation. Table 13 provides the piezoelectric matrices for α-quartz together with several values d_{ijk} and e_{ijk}, including the corresponding temperature coefficients [259], [261].

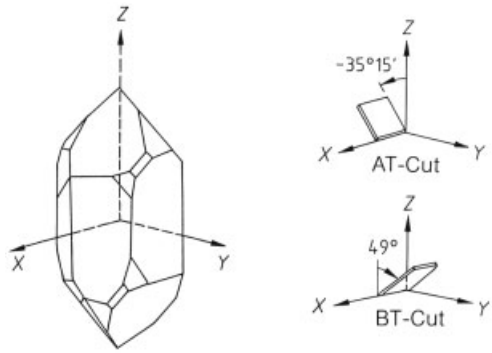

Figure 56. Oriented quartz crystal cuts in relation to the natural crystal [209] (with permission from Elsevier, Amsterdam)

For sensor applications, the magnitudes of d_{ijk} and e_{ijk} should be as great as possible, and the temperature dependence should be as small as possible. Although lithium niobate displays larger stress and strain constants than α-quartz, the temperature influence in this case is also great. Therefore, α-quartz is the material most often selected for piezoelectric sensor applications. Because of the anisotropic behavior of a piezoelectric material, properties like resonance frequency and temperature dependence can be optimized by cutting a plate from a single crystal in a particular way. Figure 56 shows two possible cuts with respect to a natural α-quartz crystal. For use as acoustic gravimetric sensors, only AT- and BT- cut quartz plates are useful. These *y*-rotated cuts provide two different high-frequency plates that vibrate in a shear mode along an axis parallel to the major surface, as indicated previously (see also Fig. 43 B). Minimizing temperature effects requires that the plate be cut at a very precise orientation. The temperature dependence of the relative frequency for various cuts of quartz crystal is shown as a function of the angle of cut in Figure 57 [262].

More detailed theoretical treatments of piezoelectric solids and their properties are provided by [222], [263–265]. Another more sophisticated approach based on dynamic quantum mechanics is presented in [266]. Interest has also developed in piezoelectric polymers, especially

Table 13. Piezoelectric matrices for α-quartz and $LiNbO_3$, together with the corresponding piezoelectric stress and strain coefficients [259], [261]

Classification	α-Quartz: Trigonal system, class 32			$LiNbO_3$: Trigonal system, class 3 m			
Matrix	$\begin{pmatrix} d_{11} & -d_{11} & 0 & d_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & -d_{14} & -2d_{11} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$			$\begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & -2d_{22} \\ -d_{22} & d_{22} & 0 & d_{15} & 0 & 0 \\ d_{31} & d_{31} & d_{33} & 0 & 0 & 0 \end{pmatrix}$			
Piezoelectric strain constants	d_{11}	d_{14}		d_{22}	d_{31}	d_{33}	d_{15}
	2.3	−0.67		21	−1	6	68
	2.31	−0.670		22.4	−1.2	18.8	78.0
	2.31	−0.727		20.8	−0.85	6.0	69.2
Temperature coefficients of piezoelectric strain constants ($\times 10^{-4}/°C$)	$T_{d_{11}}$	$T_{d_{14}}$		$T_{d_{22}}$	$T_{d_{31}}$	$T_{d_{33}}$	$T_{d_{15}}$
	−2.0	17.7		2.34	19.1	11.3	3.45
	−2.15	12.9					
Piezoelectric stress constants	e_{11}	e_{14}		e_{22}	e_{31}	e_{33}	e_{15}
	0.173	0.04		2.5	0.2	1.3	3.7
	0.171	0.0403		2.43	0.23	1.33	3.76
Temperature coefficients of piezoelectric stress constants ($\times 10^{-4}/°C$)	$T_{e_{11}}$	$T_{e_{14}}$		$T_{e_{22}}$	$T_{e_{31}}$	$T_{e_{33}}$	$T_{e_{15}}$
	−1.6	−14.4		0.79	2.21	8.87	1.47

as a result of their strongly piezoelectric properties [222].

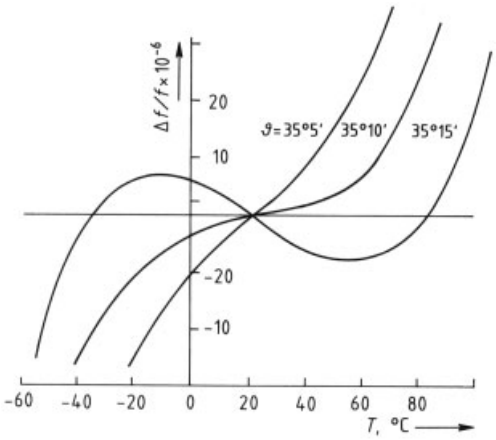

Figure 57. Temperature dependence of the relative vibrational frequency for different AT-cuts of a quartz crystal [262] (with permission from Elsevier, Amsterdam)

Vibrations and Waves. The theory of acoustic waves in solids is well understood, and many comprehensive descriptions exist [249], [267]. The description presented here should suffice to explain sensor function despite the fact that it is limited and qualitative.

Acoustic waves in solids consist of oscillatory motions of the material comprising the solid. The stress – strain relationship treated in the preceding section leads ultimately to two types of acoustic displacements: *longitudinal* or *tensile* motions that cause the dimensions of the solid body to change along a single direction without accompanying changes along the orthogonal directions, and *shear* motion, which produces changes in the shape of the solid body. Figure 58 illustrates these two types of motion [256].

In piezoelectric solids, an applied force generally creates both quasi-longitudinal and quasi-shear waves, where the acoustic displacement in a wave will be largely either longitudinal or shear. Pure longitudinal (L) and shear waves (SH = horizontally polarized shear waves; SV = vertically polarized shear waves) are generated if wave propagation occurs along certain crystallographic axes or with specific orientations. The generation of mixed wave types is also possible, but normally not desirable for sensor applications.

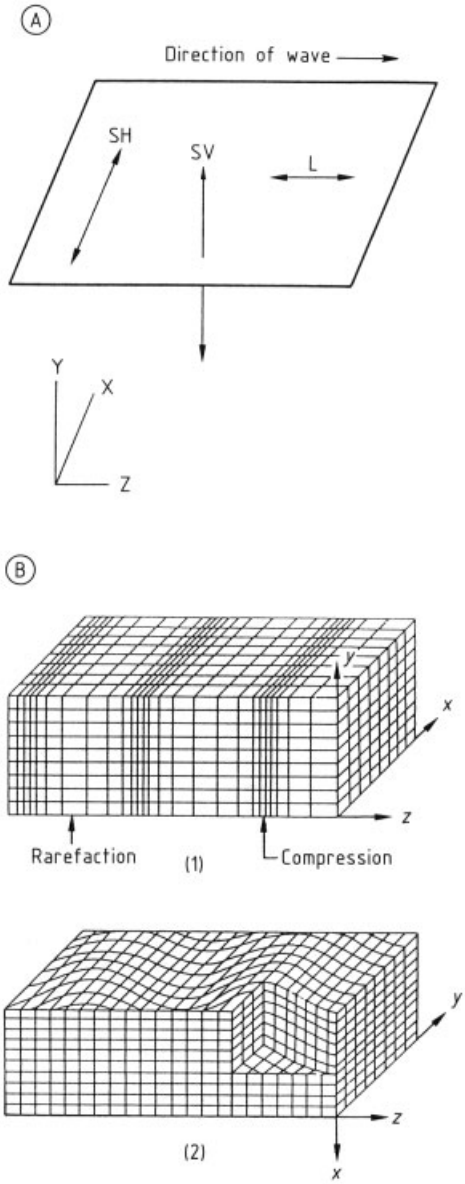

Figure 58. A) Schematic representation of longitudinal and shear displacement [256]; (with permission from Elsevier, Amsterdam) B) Grid diagrams for plane uniform waves propagating along the z axis in a material of infinite extent [248] (with permission from J. Wiley & Sons, New York): (1) Tensile motion (one dimension); (2) Shear motion (two-dimensional; y-polarized in this case)
L = Longitudinal (tensile) motion; SH = x-Polarized shear (horizontal); SV = y-Polarized shear (vertical)

As far as microgravimetric sensors are concerned, two types of wave-propagation devices can be distinguished: those characterized by wave propagation perpendicular to the plate surface (BAW) and those in which the waves propagate along the surface (SAW) or in the bulk of the substrate (LW, APM). Transducers of the former type are normally resonators. A given device might be designed to utilize either longitudinal or shear waves propagated between the faces, as in microbalances based on AT- or BT- cut quartz plates operating in their SH high-frequency fundamental mode. Only acoustic waves satisfying boundary conditions for the resonator surfaces will propagate efficiently across such a device. The waves can be visualized as standing sinusoidal waves with displacement nodes at the resonator surfaces (see 2.3.3.2).

Transducers in the second category include wave delay-line oscillators (see Figs. 42 and 47), which are based primarily on two types of waves. One is a coupled linear combination of two wave-equation solutions, an L-wave and an SV-wave. In this wave, also called a Rayleigh wave, particle displacement takes the form of an elliptical motion in the plane perpendicular to the surface (sagittal plane) and containing the direction of propagation. Some displacement also occurs outside the sagittal plane, but the deviations are minor (Fig. 59). Two of the most important cuts providing these wave types, and which are used in chemical sensors, have been illustrated in Figure 56. Waveguides that rely on Rayleigh wave devices for their operation are often referred to as surface acoustic wave (SAW) devices, and they have been extensively employed as sensors.

The second type of wave in this context is a horizontal shear (SH) wave, whose acoustic displacement is in the plane of the waveguide. An SH wave generally travels at a different velocity than a Rayleigh wave, and for piezoelectric waveguides the SH wave becomes a true surface wave provided the acoustic impedance at the surface differs from that in the substrate bulk. Such a difference can be caused by piezoelectric stiffening of the SH mode, deposition of a layer of dissimilar material on the surface, the presence of a surface grating, or liquid in contact with the surface. Depending on their origin, SH surface waves are referred to in the literature as Bleustein – Gulayev (BG) waves, Stoneley waves, SH-SAW, or surface transverse waves. Pure surface transverse waves (shear waves) appear to show promise for applications in liquids, especially biosensor applications, because the dissipation of energy across the phase bound-

Cut	Designation	Type of wave	Operating frequency
AT-Quartz	(YXl)-35°	Thickness shear mode	0.5-100 MHz

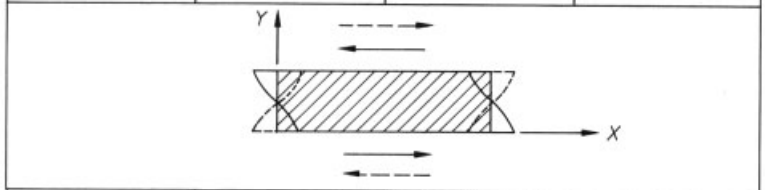

GT-Quartz	(YXlt)-51°/-45°	Longitudinal extension	50-550 kHz

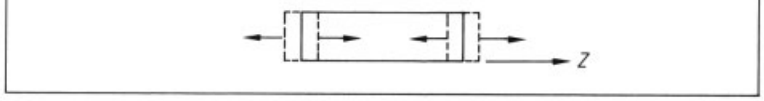

ST-Quartz	(YXl)-42.5°	Surface wave, vertically polarized	50-1500 MHz

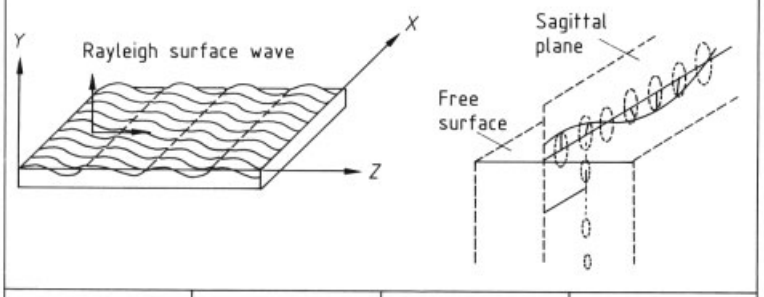

Various Y-rotated quartz cuts		Surface wave, horizontally polarized	e.g., 330 Mz

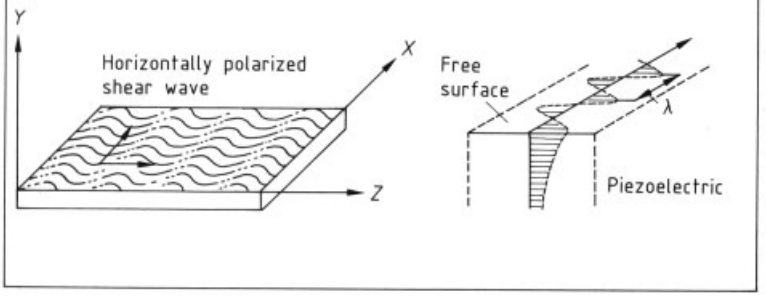

Figure 59. Wave properties associated with several different types of quartz cuts [260]; see also Figure 48

ary into the liquid is approximately zero, which greatly reduces the generation of longitudinal sound waves in the liquid itself.

Plate-mode waves are analogous in form to Rayleigh and SH waves in a semi-infinite waveguide. Two groups of plate modes can be distinguished: the first consists of Lamb waves or flexural waves, where the acoustic displacement is a combination of longitudinal and vertical shear motion; a second family involves SH acoustic displacement [256]. Because of their frequency range (100 kHz to the GHz region) these sound waves fall in the ultrasound category, with a propagation velocity of 1 – 10 km/s (cf. electrical signals, which travel at 100 000 – 300 000 km/s).

Interdigitated Transducer (IDT). The IDT was invented by White and Voltmer in 1965. The IDT is a planar, interleaved metal structure at the surface of a highly polished piezoelectric substrate (see Figure 60). The adjacent electrodes are given equal potentials of the opposite sign. The resulting spatially periodic electric field produces a corresponding periodic mechanical strain pattern employing the piezoelectric effect. This gives rise to generation of surface acoustic waves (SAW), provided that the surface is stress free. In general both SAW and BAW may be generated by IDT.

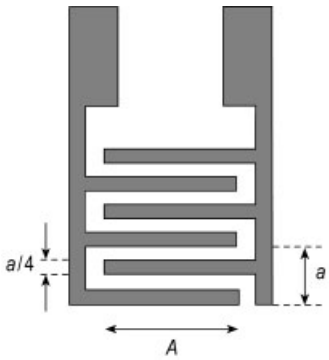

Figure 60. Layout of an Interdigitated Transducer (IDT) with period a and aperture A. The period determines the wavelength of the generated SAW $\lambda = a$. The aperture determines the width of the radiated beam and is typically 10 – 100 wavelengths in magnitude. The IDT length is $L = N\lambda$, when IDT contains N electrode pairs

The IDT behaves as a sequence of ultrasonic sources. For an applied sinusoidal voltage, all vibrations interfere constructively only if the distance $a/2$ between two adjacent fingers is equal to half the elastic wavelength. The frequency $F_0 = V_{AW}/\lambda = V_{AW}/a$ that corresponds to this cumulative effect is called the synchronous frequency or the resonance frequency. The bandwidth of an IDT is narrower, when there are more fingers. When IDT is used the AW velocity is determined by the plate material and orientation, while wavelength depends only on the ITD periodicity.

SAW are emitted in both opposite directions, which result in an inherent minimum of 3 dB transducer conversion loss at F_0. A minimum insertion loss of 6 dB is found for a delay line. The acoustic aperture A defines the effective region of transduction between two adjacent electrodes. The IDT is uniform if a constant aperture is obtained.

SAW component design is based on the application of an equivalent circuit model [29] using the values of the piezoelectric coupling coefficient of the material, F_0, and the static capacitance [30].

The frequency at which the AW device operates depends on:

- the acoustic wave velocity in the substrate material
- the IDT finger spacing (for SAW)
- the substrate (plate) thickness

Additionally, the type of AW generated by IDT depends on:

- the crystal cut
- the orientation of the IDT relative to the crystal cut
- the substrate (plate) thickness
- the wave guiding mechanism (BAW, SAW, APM)

Acoustic wave devices (AWD). Figure 61 shows an overview of the basic types of AWDs, indicating the polarization of the generated acoustic waves, and whether the devices can be operated in liquid.

2.3.3.4. Technical Considerations

The (Bio)Chemical Interface. A piezoelectric transducer is transformed into a mass-sensitive sensor by coating it with an appropriate (bio)chemical film or layer, which may

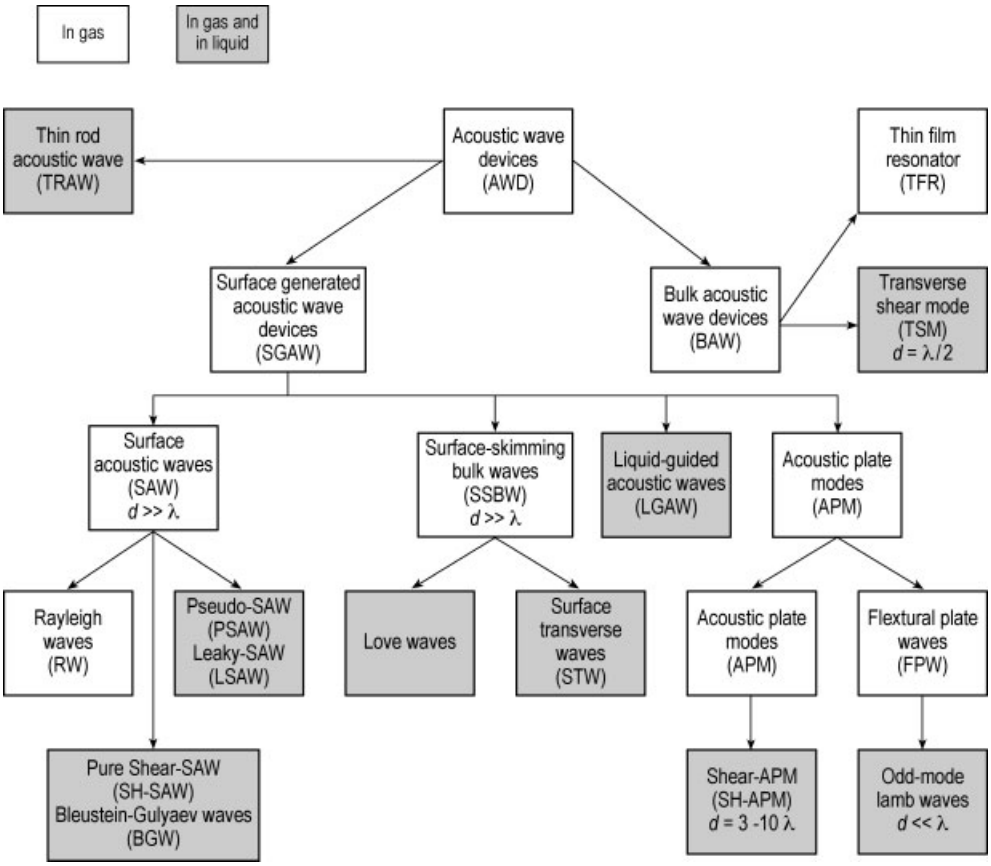

Figure 61. Flow diagram illustrating the basic types of acoustic wave devices. Here, *d* is the substrate (vibrating plate) thickness and l is the acoustic wave wavelength

be inorganic, organic, or even metallic in nature. Ideally the film should be strongly adherant and deposited in such a way as to obtain a plane surface. Concerning immobilized large biomolecules, certain doubts exist over the demand for a plane surface and the request for a strong adherence. It seems highly unlikely that an immobilized macromolecule will follow any surface movement of the transducer as a whole entity given its flexible structure. The thickness of such a layer is usually on the order of 1 – 100 µm depending upon the nature of the transducer. The chemically sensitive layer serves as an interface between the environment and a data-processing system, and it is responsible for both the sensitivity and the selectivity of the sensor. This layer should display the following properties:

1) Reversibility, sensitivity, and selectivity
2) Inertness with respect to chemical cross-influences, such as humidity or carbon dioxide
3) Ease of coating, strong adhesion, and long life

Most chemically sensitive layers are insulators, such as organic substances with special functionalities, but conducting materials such as metal films or conducting polymers are also useful. Examples include:

1) Donor – acceptor functionalities; e.g., carbocations [268], M(II)porphyrins [269], or betaines [270]
2) Host – guest binding systems (supramolecular compounds); e.g., molecular cavities [271], clathrate systems [272]
3) Pure metallic layers; e.g., palladium for the detection of hydrogen [273]

Interaction of the Chemical Layer with an Analyte. Interaction with an analyte can take many forms ranging from adsorption – absorption to chemisorption, including the compromise of coordination chemistry. The following types of chemical interactions can be distinguished [274]:

1) *Absorption.* Here the analyte is distributed between the chemical interface (a liquid film, amorphous solid, or polymer, analogous to the stationary phase in gas chromatography) and the surrounding medium. In the case of a gas interacting with the chemical interfacial layer, once equilibrium has been achieved the amount of analyte present in the layer is a function of the partition coefficient between the gas and the interface (as specified by Raoult's and Henry's laws) as well as the thickness and area of the interface. Because distribution tends to be determined mainly by the polarity of the chemical interface and the analyte, little selectivity is expected.
2) *Adsorption.* The analyte in this case interacts only at the surface of the interfacial layer, and no chemical bonds are formed or broken. Attraction energies are usually in the range 0 – 10 kJ/mol (i.e., attributable to van der Waals forces), extending in the case of hydrogen bonds to as much as 40 kJ/mol. Adsorption may result from physical attraction between a nonreactive gaseous analyte and a metal surface, for example, or formation of a donor – acceptor association (H^+ or e^-). Adsorption is a universal phenomenon, and it is quite nonselective. The basic principles of gas adsorption have been described by Langmuir [275] and de Boer [276].
3) *Chemisorption* [277]. Chemisorption involves very strong interaction – up to ca. 300 kJ/mol – developed at the surface of the chemical interface, including the formation and breaking of chemical bonds. High selectivities can therefore be anticipated, but there is a simultaneous loss of reversibility. Because of the partially irreversible nature of bond formation, such a sensor would actually behave more like a dosimeter, although this might be advantageous in some applications.
4) *Coordination chemistry* [274]. A compromise offering both selectivity and reversibility of analyte binding is presented by the area of coordination chemistry. A typical coordination compound consists of a central metal ion M surrounded by a neutral or charged (often organic) ligand. The extent of selectivity can often be influenced by the choice of the metal ion as well as by the choice of the ligand – taking into account both electronic and steric factors. Complexation with an analyte produces changes in the properties of the coordination compound, and these changes can be subject to detection. The selectivity is a function of structural, topological, and polarity parameters.

Other types of chemical interfaces have also been used for microgravimetric gas detection, including polymeric phthalocyanines [278], porphyrins [279], ferrocenes [280], metal clusters [281], crown ethers [282], and cyclodextrins [283].

In many cases it has proven to be very difficult to find a selective chemical interface for a particular analyte, so the use of sensor arrays has been proposed. Numerous papers have dealt with this approach, invoking both BAW and SAW sensors. Pattern-recognition techniques employing increasingly powerful microprocessors are required to generate the appropriate signals [284].

Coating Methods. The stability of a (bio)chemical interface depends heavily on the attachment of the material to the surface of the sensor device. One or more of the following coating methods is employed according to the nature of the sensitive layer:

Smearing [224]
Spin coating [285]
Drop coating [103], [257]
Solvent evaporation [286]
Spraying [287]
Langmuir – Blodgett film formation [288]
Physical vapor deposition [289]
Sputtering [290]
Chemical immobilization [291]

The first five methods listed involve simple coating procedures in which the interfacial material (usually nonpolar) is dissolved in an organic solvent and then deposited directly onto the substrate surface. Evaporation of the solvent leaves a film that adheres to the surface of the substrate by physical bonds. The

Langmuir – Blodgett technique makes it possible to achieve very thin films. This is a dipping procedure in which molecules are transferred with great accuracy from the surface of a liquid to the surface of the substrate. Even monolayers can be prepared in this way. More complicated methods include physical vapor deposition (PVD) and chemical vapor deposition (CVD). Very planar layers can be achieved in this way, but the reaction conditions are usually quite drastic, so the technique is essentially restricted to materials that are relatively inert, such as inorganic compounds or metals.

2.3.3.5. Specific Applications

Having reviewed the fundamental aspects of microgravimetric transducers and piezoelectric solids with respect to the generation and propagation of vibrations and waves, it is now appropriate to turn to the practical considerations of instrumentation and application, addressing first the BAW devices.

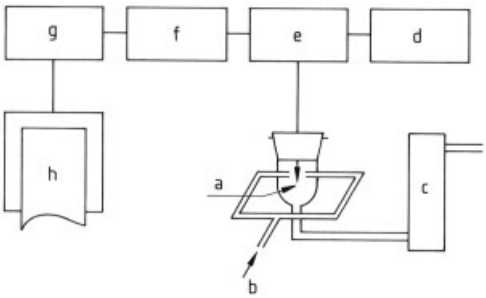

Figure 62. Typical experimental apparatus based on a piezoelectric quartz-crystal detector [217] (with permission from Elsevier, Amsterdam)
a) Piezoelectric quartz crystal; b) Sample; c) Flow meter; a) Piezoelectric quartz crystal; b) Sample; c) Flow meter; d) Power supply; e) Oscillator; f) Frequency counter; g) Digital-to-analog converter; h) Recorder

Bulk Acoustic Wave Sensors. *Experimental Arrangement.* A typical experimental setup involving a piezoelectric mass-sensitive quartz crystal detector is depicted schematically in Figure 62 [217]. The piezoelectric quartz crystal (a) is shown here inserted directly in a gas stream (b) with a flow velocity of ca. 10 – 100 mL/min, but a measuring cell containing a stationary atmosphere of the analyte gas is also useful. The oscillator (e) is usually powered by a regulated power supply (d; e.g., 5 – 15 V) that drives the quartz crystal. The frequency output from the oscillator is monitored with a frequency counter (f), which should be modified by a digital-to-analog converter (g) to permit the frequency data to be recorded.

Parameters that adversely affect the performance of a piezoelectric sensor include the mass of the chemical layer, built-in stress produced by the chemical layer, stress from the electrodes (clamping), and changes in temperature. A more sophisticated device such as a microprocessor-controlled dual- crystal instrument would be required to eliminate these undesirable effects [292].

The device described is designed for measurements in the gas phase; applications in a liquid medium demand an extended version, including a liquid-tight box, an oscilloscope or network analyzer, and provisions for more detailed signal analysis (impedance analysis). Improvements in the oscillating circuit are particularly important for obtaining satisfactory results with liquids [293–295]. A proper interpretation of any observed frequency change in liquid media still remains. Can the densitiy and the viscosity of the sample liquids be controlled so accurately (sometimes within the fifth digit after the decimal point) and does the surface roughness not change during the measurements increasing the viscoelastic effects?

Applications. This section will be limited to a brief consideration of the use of BAW sensors in gas-sensing devices and as sensors for liquid phases. Piezoelectric microbalances can also be used for monitoring such heterogeneous samples as aerosols and suspensions, as discussed in [262].

Gas-Phase Sensing. In principle, both inorganic and organic analytes are subject to detection. Figure 63 provides an example of results obtained from the novel 155 MHz piezoelectric quartz microbalance (QMB) resonator used to detect gaseous tetrachloroethylene (C_2Cl_4). The used absorption layer was polyepichlorhydrin (PECH). Observed response curves are illustrated as a function of time and for various analyte concentrations. [257].

A wide variety of solvent vapors can be detected with this new QMB with extremely high sensitivity (starting in the ppb range) and showing a dynamic working range of up to 4 decades

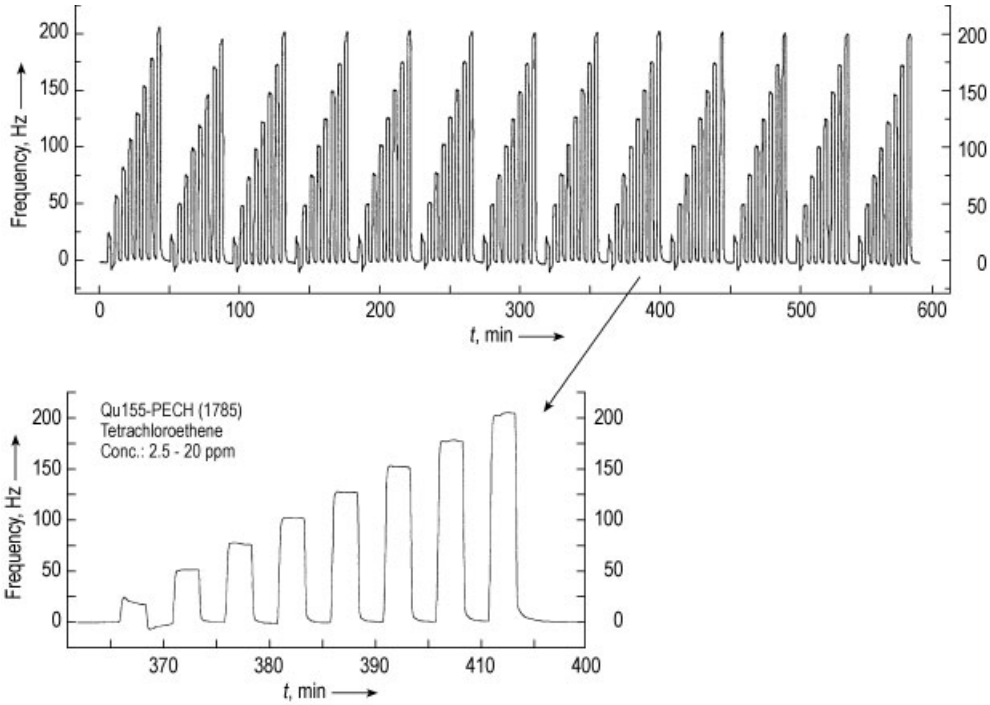

Figure 63. QMB-based gas sensing with polyepichlorohydrin (PECH) as polymer. The upper original recorder traces show 13 complete calibration cycles and demonstrates the very high reproducibility of this gas sensor in this low concentration range. The lower part shows the details of one cycle between 2.5 and 20 ppm tetrachloroethene. With a sensitivity of ≅ 10 Hz/ppm tetrachloroethene and an LOD of about 100 ppb this device is more sensitive than any Taguchi gas sensor or the FID considering its low noise, in the range of only about 0.5 Hz! Compared with a 25 MHz QMW the sensitivity could be increased by one order of magnitude [257]

with response times below 10 s, indicating the potential of this approach to gas-phase monitors and sensors for applications in environmental analysis or process control.

Table 14 provides an overview of the detection of selected gaseous analytes. Information regarding interfering gases can be found in the original literature. The effect of changes in the relative humidity of the environment tends to be a general problem.

Applications in Liquids. Apart from the classical sensors designed for detecting ions, piezoelectric crystal resonators are the preferred devices for conducting biochemical measurements in liquids. Figure 64 illustrates results for the detection of human immunodeficiency virus (HIV) antibodies by means of a synthetic HIV peptide (p 24) immobilized on a piezoelectric quartz sensor. The data refer to the frequency shift observed upon addition of authentic HIV antibody. The quartz device was operated in this case in its fun-damental thickness shear mode at 20 MHz, with an oscillator circuit stability of ca. 0.5 Hz. The quartz plate was coated with HIV peptide (p 24) dissolved in phosphate buffer at a dilution of 1 : 1000. Figure 65 shows the corresponding response as a function of different concentrations of monoclonal HIV antibodies. This device has not yet been brought onto the market despite complete automation and further improvements. Thus, some remarks concerning mass-sensitive transducers employed in liquids must be made, since it is still doubtful if the frequency change observed by passing a sample solution over such a sensor surface really reflects a mass change. Even if the user will manages to control the density and viscosity of the sample solutions within a narrow range determined by the required sensitivity (up to 5 digits after the point) a change in the surface roughness by non-homogeneously attached macromolecules during the immuno reaction may affect the vis-

Table 14. Survey of selected reports on BAW sensors for inorganic and organic analytes [230]

Analyte	Range, ppm	Chemical interface	References
Inorganics			
CO	1 – 50	HgO/Hg→Au	[297]
CO_2	≤ 100 000	7,10-dioxa-3,4-diaza-1,5,12,16-hexadecatrol	[298]
$COCl_2$	8 – 200	methyltrioctylphosphonium dimethylphosphonate	[299]
HCN	13 – 93	bis(pentane-2,4-dionato)nickel(II)	[300]
Organics			
Nitroaromatics	0.001 – 100	carbowax 100	[301], [302]
Dimethylhydrazine	NR∗	polybutadienes	[303]
Acetoin	0.008 – 0.120	tetrabutylphosphonium chloride	[304]
Formaldehyde	0.010 – 100	formaldehyde dehydrogenase	[305]
Vinylchloride	≤ 80	amine 220	[306]

∗ not reported.

coelastic effect. In addition, the complex structure of proteins/antibodies may contain a variable amount of occluded water and ions, which depends on the ionic strength and other factors that have to be carefully controlled to get the true analytical result. result.

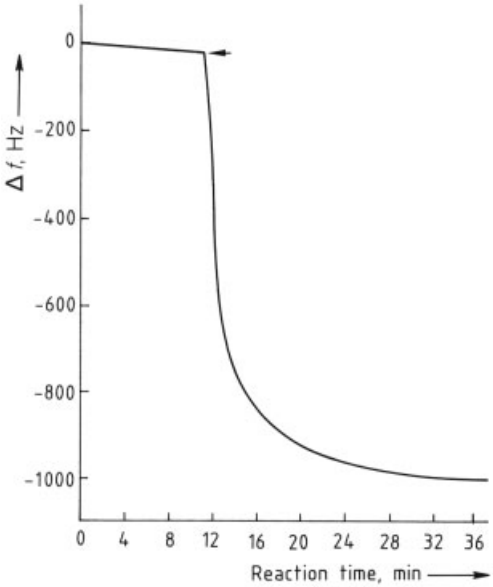

Figure 64. Response for the reaction of anti-HIV antibodies (dilution 1 : 1000) with adsorbed HIV peptide; the saturation value for the frequency shift is reached ca. 20 min after sample introduction (indicated by the arrow) [307] (with permission from Elsevier, Amsterdam)

Surface Acoustic Wave Sensors. Many reports have appeared regarding the use of SAWs as gas detectors. These reports also discuss problems associated with the choice of a coating, achieving adequate specificity, selecting a coating thickness, and depositing reproducible levels of polymer coating on the surface of the sensing element. The variety of coatings used is as diverse as the nature of the analytes, covering the range from simple polymers through metallocyanines to metal films [308]. The main drawback after the introduction of the 155 MHz QMB was the lack of availability and the price. Concerning the LOD, their inherent higher noise excludes, in general, the ppb range. The application of SAW sensors in liquid systems is still an active area of research, so this topic will not be treated. Information regarding liquid applications for mass-sensitive devices, especially SAW, APM, and LW devices, is provided in [309], [310].

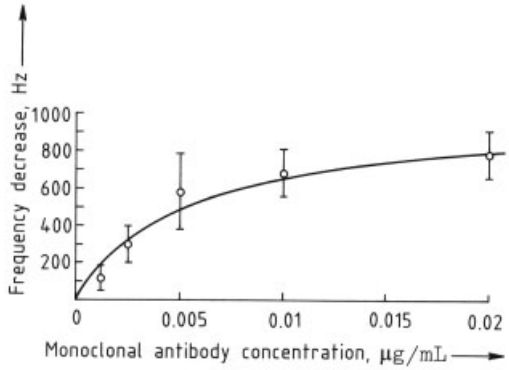

Figure 65. Frequency response for the reaction of HIV antibodies at several concentrations with adsorbed HIV peptide [307] (with permission from Elsevier, Amsterdam)

Experimental Arrangement. A schematic representation of a SAW sensing device is provided in Figure 66. This particular resonator was used by D'Amico et al. [311] for the detection of

hydrogen. It consists of a dual delay-line structure in which one of the propagation paths (l_S) is coated with a thin palladium film to act as the sensing channel, while the other (l_R) is uncoated and serves as a reference.

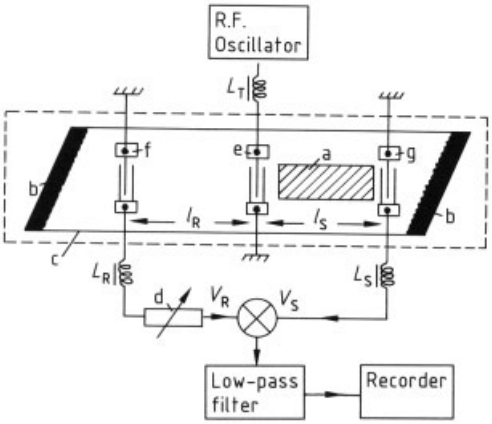

Figure 66. Schematic representation of a SAW hydrogen sensor [311] (with permission from Elsevier, Amsterdam) a) Palladium film; b) Acoustic absorbers; c) YZ-$LiNbO_3$; d) "Stretched line," an adjustable resistance for matching the frequency of V_R to that of V_S; e) Input transducer T; f) Output transducer T_R; g) Output transducer T_S; l_R, l_S = Propagation path (l_R is the distance between T_R and T; l_S between T_S and T); L_T, L_R, L_S = Series inductors (establishing a fixed e.g., 50 Ω electric input – output impedance); V_R, V_S = Output voltage of the dual delay line

Applications. The type of results to be expected with a variety of analytes in the gas phase is illustrated by the example of an NO_2-SAW sensor. The device in question was developed by Nieuwenhuizen and Venema [312], and had an operating frequency of 79 MHz with a metal-free phthalocyanine coating. A frequency change of ca. 670 Hz was observed upon exposure of the sensor to 88 ppm of nitrogen dioxide in air at 150 °C. Figure 67 A demonstrates the sensitivity of the device as a function of nitrogen dioxide concentration, while Figure 67 B shows the response of a typical sensor as a function of time with alternating exposure to nitrogen dioxide and air. Possible interferences were also investigated, as described in a similar report published earlier by the same authors [313].

A summary of selected papers describing the use of SAWs for gas sensing is provided in Table 15, subdivided according to the type of analyte.

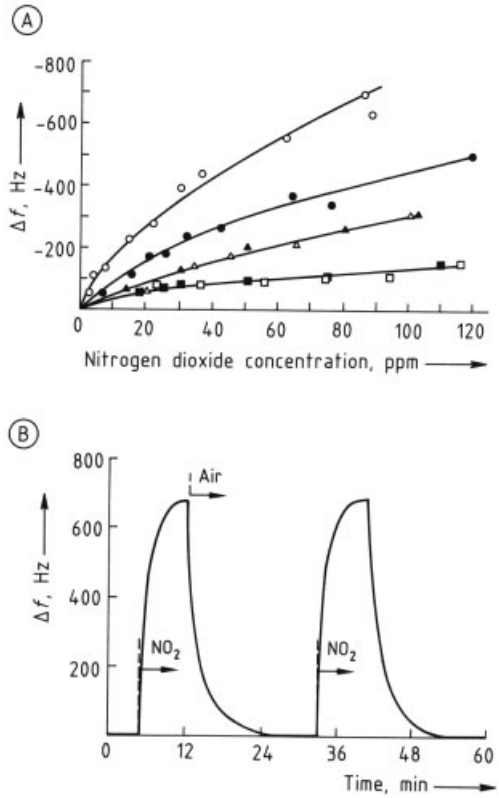

Figure 67. A) Sensitivity of an NO_2-SAW resonator at 120 °C (●, ■, △) and 150 °C (○, □, ▲) with three different metal-free pthalocyanine layers: PC 42 (●, ○), PC 46 (■, □) and PC 47 (△, ▲) [312]; B) Sensor output as a function of time at 150 °C (PC 42, 88 ppm of NO_2)
(with permission from Elsevier, Amsterdam)

2.3.3.6. Conclusions and Outlook

Mass-sensitive devices offer great promise as chemical sensors, particularly since they rely on the most basic physical effect that accompanies the interaction of one chemical substance with another: the mass effect.

Both BAWs and SAWs have already been successfully applied to the monitoring of many gases, vapors, and more recently, liquids. One of the most critical aspects of chemical sensor research is the search for a suitable chemical interface. The selectivity and sensitivity of mass-sensitive devices can be improved by the use of more carefully designed polymers and other special coating materials. Correction algorithms might also be developed to compensate for variations due to changes in humidity, temperature,

Table 15. Survey of selected reports on SAW gas sensors [308] (with permission from Adam Hilger, Bristol)

Analyte	Sensor material	Frequency, MHz	Temperature, °C	Chemical interface	References
Inorganics					
H_2	$LiNbO_3$	75		Pd	[311], [314]
H_2	quartz	23		Pd	[315]
Halogens	quartz	50 – 80	30 – 150	Cu – PC [a]	[316]
H_2O	$LiNbO_3$	75	0 – 100	polymer	[317], [318]
NO_2	$LiNbO_3$	150	70	Cu – PC [a]	[319]
Organics					
DMMP [c]	quartz	31 – 300	20 – 80	polymers	[320]
Styrene	quartz	30		Pt complex	[286]
Cyclopentadiene	quartz	31	25	PEM [b]	[321]
Various	ZnO	109	24	ZnO	[322–324]
Various	quartz	158	35	polymers	[325]

[a] PC = Phthalocyanine.
[b] PEM = Poly(ethylene maleate).
[c] DMMP = Dimethyl methylphosphonate.

pressure, etc. Humidity poses one of the greatest problems with respect to most sensing devices. Better design of the electronic oscillator circuits might also increase the stability of the signal in response to a gas or vapor, thereby improving the signal-to-noise ratio and raising the level of sensitivity.

Research on chemical sensors tends to be unusually multidisciplinary. Such new techniques as the use of Love plates and LW-APM make it possible to combine the advantages of microcompatibility, high sensitivity, and suitability for work in both the gas and liquid phases. In particular, the new 155 MHz BAW could open up new horizons in gas sensor developments and electronic noses. The higher sensitivity of this transducer now allows dynamic ranges larger than 4 orders of magnitude with response ranges in the seconds range. Taking the first derivative of the absorption or desorption signal this device can be used as a more or less selective GC detector [257]. A 5 ppm peak of tetrachlorethene can then be sensed with a signal-to-noise ratio of about 3000.

Nevertheless, anyone proposing to conduct investigations in this area should be aware of the complexity of interpreting data accumulated from multicomponent gas mixtures over a prolonged period of time, and of parameter identification and interpretation in the case of liquids [230], [308].

2.3.4. Calorimetric Devices

Introduction. Most chemical as well as enzyme-catalyzed reactions are accompanied by changes in enthalpy. For this reason, calorimetric transducers represent a universally applicable approach to chemical or biochemical sensors.

Thermal devices for detecting an enthalpimetric effect (either exo- or endothermic) are insensitive to changes in optical properties of the sample, so colored solutions and turbidity do not constitute interferences. There is also no need for a detectable reaction product, whereas amperometric sensors, for example, require the formation of some product subject to transformation at an electrode. Amplification can be achieved by immobilization of a catalyzing reagent that enhances the rate of enthalpy change. Thermal transducers are especially suitable for continuous measurements in flow systems. An important disadvantage, however, is the requirement for thermostatic control of the device. For this reason calorimetric systems tend to be rather large. In addition, calorimetric sensors are relatively complicated to use [326], [327]. Table 16 provides molar enthalpies for several important types of chemical reactions.

The measurement signal in this case, ΔT, is related to the change in enthalpy ΔH, but also to the heat capacity of the system c_s, according to the following equation:

$$\Delta T = n\Delta H / c_s$$
$$n = \text{Moles of product} \tag{37}$$

This means that the heat capacity of the sensor itself should be minimized to ensure sensitive measurements. A temperature change can easily be transduced into an electrical signal. The usual measuring device is based on a reference

thermal transducer incorporated into a Wheatstone bridge.

Table 16. Molar reaction enthalpies for selected processes [328], [329]

Reaction	Examples	Enthalpy $-\Delta H_R$, kJ/mol
Oxidation	methane; O_2	800
Neutralization	NaOH, HCl	55
Protonation	tris(hydroxymethyl) aminomethane; H^+	47
Enzyme catalysis	enzyme: catalase; substrate: H_2O_2	100
	enzyme: glucose oxidase; substrate: glucose	80

Thermal Transducers Commonly Utilized in Chemical and Biochemical Sensors. *Metal-Resistance Thermometers.* The electrical resistance of many metals rises sharply with increasing temperature. Resistors made of platinum (called "Pt 100" because they provide a resistance of 100 Ω at 0 °C) are often used for temperature measurement in the range from − 220 °C to +750 °C. Other metals like nickel or copper are also applicable for temperatures ≤ 150 °C. Self-heating, caused by a current flow, interferes with the measurement, and the response time of a metal resistor is relatively long (ca. 5 s). The advantages of such resistors are high sensitivity and long-term stability. Platinum resistance devices are especially common in catalytic chemical sensors. A reducing gas may be oxidized in the presence of a heated catalyst (e.g., Rh, Pd, or Pt) in order to increase the reaction rate. In 1962 Baker [330] described a device called a *pellistor* that effectively separates the catalyst from the platinum wire (see Fig. 68). The detection limit for methane with this device was found to be 20 ppm.

Thermistors. These are semiconductor resistances with temperature coefficients sufficiently high to make them suitable for use in temperature measurement. Many semiconductors have negative temperature coefficients (NTC), which means that their resistance decreases with increasing temperature. Thermistors with reproducible temperature coefficients are difficult to produce, and self-heating within the sensor is always a problem. On the other hand, thermistors are usually both inexpensive and sensitive, and they can be used over a wide temperature range (0 – 1000 °C).

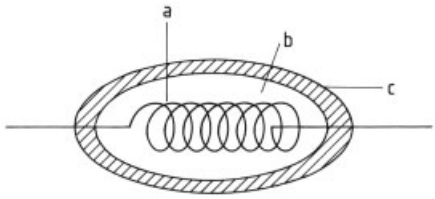

Figure 68. Pellistor device
a) Metal resistor; b) Inactive material; c) Catalyst

The thermistor most commonly used in chemical or biochemical sensors, known as an *enzyme thermistor* (ET), was designed in 1974 by Mosbach et al. [331]; it consists of an enzyme reactor specially adapted for use in a flow system. A split-flow arrangement was suggested in 1976 [332], whereby one portion of the analyte solution serves as a reference and flows through a blank column, while another passes through the enzyme reactor itself. The reference signal is subtracted from the signal produced by the enzyme reactor (Fig. 69) [333]. Samples are analyzed at a rate of ca. 20/h. Since the enzyme reaction and the detecting system are separated in space, an enzyme sensor of this type cannot be strictly interpreted as a "biosensor" according to the definition provided in Section 3.

Many other devices featuring lower detection limits have been developed, including a four-channel enzyme-based thermistor that permits the detection of four substrates simultaneously [334]. Another approach involves fabrication of a thermal microbiosensor on a silicon chip. In this case the sensor consists of a micromachined enzyme reactor, inlet and outlet flow channels, and a microthermistor [335].

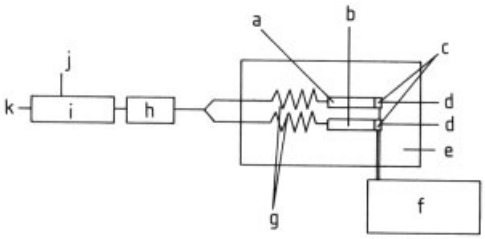

Figure 69. Enzyme thermistor system (with permission from Academic Press, Orlando)
a) Reference reactor; b) Enzyme reactor; c) Thermistors; d) Waste; e) Insulation; f) Wheatstone bridge; g) Heat exchangers; h) Pump; i) Injector; j) Sample; k) Buffer

Thermocouples/Thermopiles. Thermocouples operate on the basis of the Seebeck effect. If wires fabricated from two different metals or

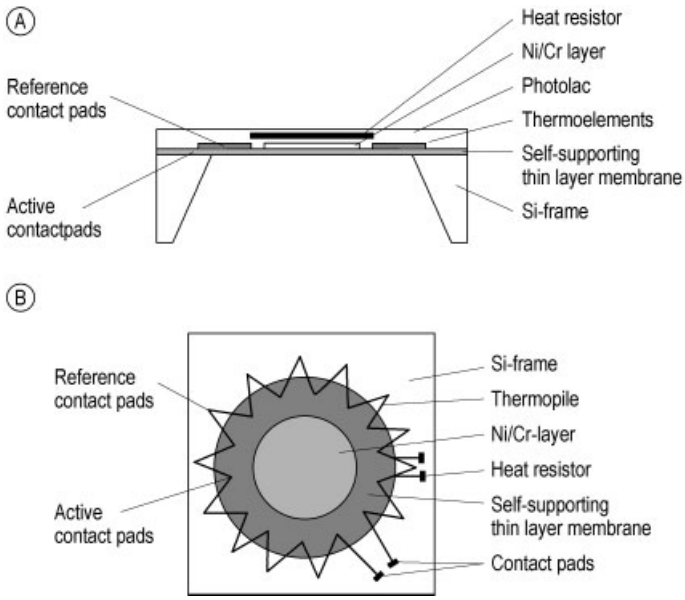

Figure 70. Construction details of a microsystem technology approach towards thermopiles. The extremely sensitive thermopile was produced in the "Institut für physikalische Hochtechnologie" IPHT - Jena

semiconductors are soldered together to form a circuit, any temperature difference that exists between the joined points leads to a measurable potential difference, the magnitude of which depends on the extent of the temperature difference and the materials involved. The use of several thermocouples connected in series (*thermopiles*) increases the sensitivity of the sensor, but if a single thermocouple is damaged the entire sensor is affected. It is more difficult to miniaturize thermoelectric sensors than resistance devices, and the long-term stability of thermocouples is often not good. A typical response time is less than one second, and sensors of this type can be used over a temperature range from $-200\,^{\circ}$C to $+1600\,^{\circ}$C.

Thermal devices based on thermocouples have been extensively investigated in recent years. Guilbeau et al. described in 1987 a thermoelectric sensor for the measurement of glucose [336]. This device incorporated an antimony – bismuth thin-film thermopile with enzyme immobilized on the side containing the active junctions but not on that of the reference junctions. Exposing the sensor to a solution containing the substrate causes the enzyme-catalyzed reaction to take place, resulting in heat exchange that is detected by the thermoelectric transducer. At the ICB Muenster, a thermopile produced with microsystem technology in the "Institut für physikalische Hochtechnologie" IPHT Jena (Germany) with the extreme sensitivity of about 10^{-4} K has been used for constructing a GOD-based glucose sensor which does not need any calibration. Figure 70 shows the construction of the thermopile with the microsystem technology.

Figure 71 shows the experimental set-up for absolute glucose sensing via the total heat of this specific GOD-based enzymatic reaction, which is strictly proportional to the number of moles reacting [337]. The heat measured in the diagram ΔH versus the glucose concentration follows the theoretical predictions. Thus, no calibration has to be performed, and no standard is needed!

Summary. Most chemical calorimetric sensors are based on the pellistor device, and are designed for monitoring gases; sensor systems of this type have been subjected to intensive investigation [338]. In general, however, sensor determination of the heat of a reaction is rare, primarily because of the availability of other, more sensitive transducers. On the other hand, thermometric biosensors are applicable to a wide range of analytes, including enzyme substrates, enzymes, vitamins, and antigens. Many applications have been reported in clinical analysis,

process and fermentation control, and environmental analysis [339]. Thermal biosensors are seldom used in industry because they are difficult to control and relatively expensive. Universal applicability and new techniques for fabricating microthermosensors suggest that these may find wider use in the future.

2.4. Problems Associated with Chemical Sensors

All nonbiochemical sensors suffer from a lack of that selectivity which would permit their wide application without detailed knowledge regarding the matrix composition of the samples under investigation. Application is therefore restricted to those areas in which the sample matrix is thoroughly understood and not subject to change. Compared to the selectivity problem, all other problems (e.g., stability, sensitivity, lifetime) are relatively unimportant, since they can be resolved by incorporating the sensor into some type of intelligent instrument capable of carrying out all necessary corrections, recalibrations, and performance tests. In some cases insufficient selectivity of a chemical sensor can be circumvented by integrating a selective chemical reaction into the instrumental design. In this way many interfering compounds are eliminated and/or the analyte of interest is selectively transformed into a species capable of being detected unambiguously. A simple approach to this end takes advantage of the flow-injection analysis (FIA) technique.

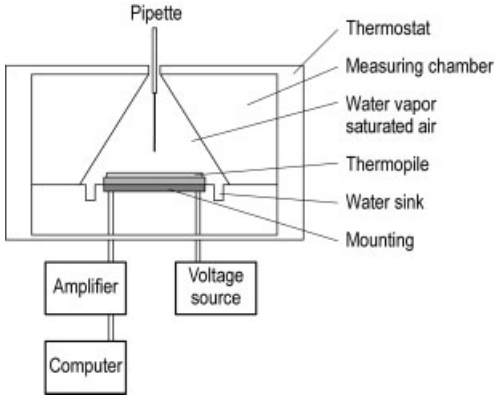

Figure 71. Measuring cell for absolute calorimetric GOD-based glucose determinations. The opening must be minimized because otherwise the water evaporates too fast and interferes with the measurement by its heat of evaporation. In order to accelerate the enzymatic reaction peroxidase was added simultaneously. As soon as the drop of blood was added the temperature – time course was recorded and integrated. The glucose concentration can then be directly calculated via the published value for ΔH for this reaction as Figure 72 demonstrates

Any chemometric method of error correction with respect to lack of selectivity must be

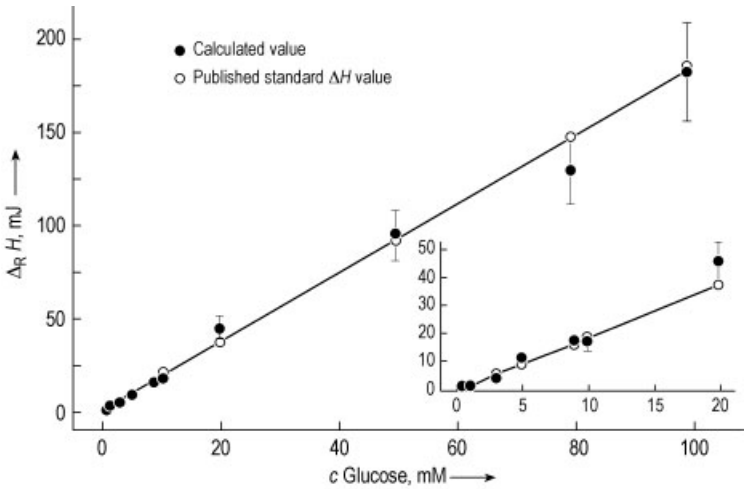

Figure 72. Several measurements with known glucose concentration are plotted on a diagram calculated from the "theoretical" ΔH values for this aliquot. Note the thermopile and the set-up was not optimized further. The only drawback is the fact that both enzymes are lost after each determination as they are part of the receptor buffer solution. This is also the case in traditional enzymatic analysis with photometric evaluation. Here, only a so-called spectral-line photometer can work without further calibration

checked very carefully, especially if it is to be applied to an open system with varying and/or unknown interfering compounds. In some cases a prior chemical separation and/or selection step is required.

Despite the superb selectivity or even specificity of biochemical sensors for recognizing particular analyte molecules, interfering compounds must be dealt with even here in real sample applications. The compounds in this case do not disturb the actual measurement of the analyte; instead they influence the recognition process. For this reason it is important to ensure the use of a stable and well-defined host molecule with an analyte-fitting geometry that is not subject to change. Anything that might alter the host would influence analyte recognition. Thus, any compound or measuring condition capable of changing the tertiary structure of the sensitive protein skeleton of some biomolecule used for recognition purposes would disturb the measurement: first by modifying the selectivity, but also by altering the sensitivity. It is well established that most selective biomolecules are very stable only in a physiological environment similar to the one in which they were produced; even changes in the ionic strength can influence the tertiary structure of a delicate protein macromolecule. Furthermore, certain metal ions (e.g., Hg^{2+}, Pb^{2+}, Ag^{+}, Cu^{2+}) are known to be bound to sulfur containing molecules (or destroy a stabilizing S – S bridge in a protein) and thus degrade recognition and thus strongly interfere. A similar effect is produced by surface-active molecules (detergents). Bearing these complications in mind, the superb selectivity of biochemical sensors must be evaluated in a broader perspective.

2.5. Multisensor Arrays, Electronic Noses, and Tongues

Theoretically, and under ideal conditions, a sensor array containing several sensors with slightly different selectivities and sensitivities towards the analyte and its main interferences should make it possible to compensate for errors caused by the limited selectivity of each individual sensor. The task resembles the solving of mathematical equations with several unkowns. There must be as many independent equations as unknowns. In sensorics this means: one additional but slightly different sensor is required for each interfering compound. The task then becomes one of solving n equations for n unknowns, which can be accomplished with an algorithm for matrix calculations and an electronic calculator. However, there are several restrictions that limit the use of this compensating technique.

First, the interfering components must be known in advance so that one can prepare a set of calibration mixtures for obtaining the individual analytical functions for each sensor under variable interference (matrix) concentrations. The approach is therefore limited to a closed sample location, and it presupposes complete information regarding potential interferents within the volume in question. This may prove applicable in certain production control situations, but certainly not in cases where the sample matrix is not sufficiently known a priori, as in the case of most environmental and some clinical analyses. Such precautions cannot be circumvented even with the use of the most sophisticated modern approaches to pattern-recognition analysis and/or so-called adaptive neural-network treatment. All potentially interfering compounds must be known in advance and introduced into the array under study during the so- called learning phase through a permutation of all possible types of mixtures. If for example one analyte is to be determined in a concentration range of 1 to 100 ppm and only five interfering compounds in a similar concentration range are present, and the concentration range should be checked for linearity by five concentrations between 0 and 100 ppm one has a permutation power of about a faculty of 10. Much time may be consumed in the preparation of the necessary mixtures and in performing all the measurements. In the end, the drift associated with most sensors and the limited lifetimes of biochemical sensors could render the whole procedure meaningless!

Second, the linearity and true independence of all effects cannot be guaranteed. In some cases of only one sensor poisoning (e.g. catalyst poisoning) all analytical functions (selectivities and sensitivities) obtained during prior calibration and/or learning are rendered invalid for every individual sensor. How is a sensor array to distinguish between irreversible poisoning effects and the normally reversible disturbances attributable

to ordinary interferents? How should it differentiate between the effects of a positive error (where a portion of the signal is due to interferents) and general matrix effects resulting in diminished sensitivity? Especially in environmental trace analysis (<10 ppm) each sample is very likely to contain more than ten interfering compounds. Intimate knowledge of each, as well as the availability of appropriate standards to permit extensive permutation at various concentrations presupposes the preparation of countless mixtures to be utilized in the training process for the sensor array. It must also be recognized that even trace amounts of certain compounds may totally disrupt the analyte recognition process.

Taking all this into consideration, as well as the fact that most chemists lack fundamental experience with sensor arrays in real open-environmental trace analysis, considerable caution is warranted with respect to this approach to increasing the selectivity of sensors and correcting for interferants. In general, application of a simple and rapid preliminary separation step will be more productive with respect to the goal of achieving a reliable analytical result. Literature examples of successful sensor-array applications related to the identification of beverages or the characterization of odors [340–342] should not be extrapolated to the quantitative analysis of a single analyte. Many of the sensor-array applications (just identification) described in the literature [343–351] involve problems that could have been solved by much simpler means, such as exact density measurement, possible now with extreme accuracy. Differentiation among various samples with similar compositions in the context of a closed system (i.e., without the variability associated with an open matrix) can usually be achieved more economically on the basis of simple physical measurements. Nevertheless, electronic noses and tongues on the market have found some application despite the humidity interference of the first category in certain food processing processes, in which they sense any deterioration in the process (e.g., a rotten tomato in ketchup processing or a bad coffee bean in the roasting process. For those applications in the gas phase the ICB Muenster recently introduced a novel electronic nose based on the new sensor class of conductometric gas microsensors (polymer + plasticizer + organic salt). It has already been mentioned that the permutation of polymer, plasticizer, and organic salt together with their concentration allows the construction of an indefinite number of different membranes with different selectivities. This is shown in Figure 73.

The identifying power of an array of only five of such sensors is shown in Figure 74 [144], [352]. The intended application area was the detection of BETX (benzene, ethylbenzene, toluene and the different xylenes) near gasoline stations for reasons of industrial hygiene. The advantages of this electronic nose using only one type of sensor principle is the low energy consumption (conductivity measurements at room temperature) making hand-held devices feasable and the fast exchange of the sensor array in the form of a cheap sensor card not larger than a business card [353].

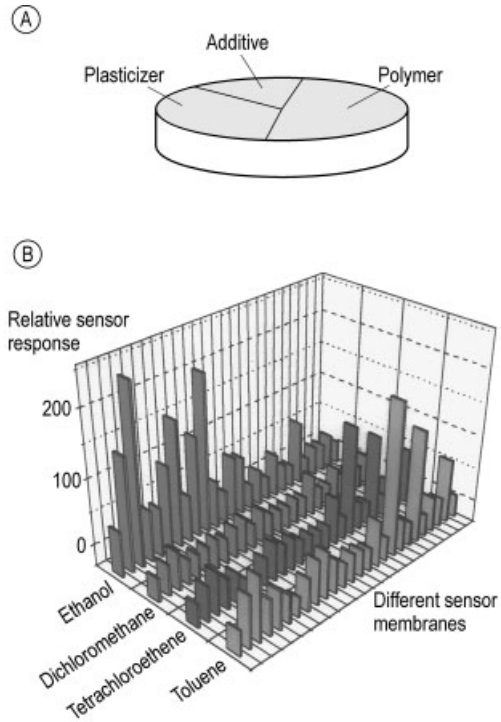

Figure 73. General composition of gas selective conductometric polymer membranes (A) and selectivity pattern against several organic solvents (B)

The above mentioned electronic nose was further improved by incorporating an extremely fast enrichment and water elimination step (in the seconds range) and increasing the resolving power following the time course of a fast ther-

mal desorption step. Since the air sampling time, the desorption time and the number of conductometric gas sensors can be individually set, a very versatile and extremely sensitive electronic nose (called: Air-Check) with sensitivities in the ppb range and almost no interference by humidity was developed in the ICB Muenster. It will be positioned in clean rooms for the purpose of surveillance.

3. Biochemical Sensors (Biosensors)

3.1. Definitions, General Construction, and Classification

Biosensors are in general small devices based on a direct spatial coupling between a biologically active compound and a signal transducer equipped with an electronic amplifier (Fig. 75). This definition of an I.U.P.A.C. working group is unfortunately very specific. The crucial thing here is the demand for a "direct spatial coupling" of biomolecules and transducers. This might have influenced the development of further commercial available biosensors since much time has been spent on biomolecule immobilization and stabilization oriented towards biosensors that could be transported and stored without a need for a refrigerator. The intimate contact of the less stable biological recognition element with a transducer surface requires either one shot devices and throwing both away after one deteriorates or a reproducible surface rejuvenation with the biological receptor by the user. The latter is the problem since the user is not an expert in controlled coverage of surfaces. Concerning the life and working time and ease of exchanging the more labile biological recognition element, devices which separate the recognition space from the detection space show certain advantages though they cannot be named a biosensor. It is, e.g., quite feasible to employ the labile biomolecules with or without the required cofactors or reagents in a freeze-dried state in a cartridge to be activated only shortly before its use. This cartridge is then to be used upstream of an optical, electrochemical or calorimetric flow through cell. Keeping this cell for longer use allows more sophisticated constructions. With modern microsystem technology the whole device could be rather small and difficult to visualize at a first glance in that the demanded intimate contact for being a biosensor is missing.

Biological systems at various levels of integration are used to recognize a particular substance that is to be determined in a specific way. Different types of biologically sensitive materials can be used for selective recognition, including enzymes, multienzyme systems, organelles, photosensitve membranes from plants and bacteria, membranes generally, protoplasts, whole intact cells, tissue slices, antibodies, lectins, DNA, and transport protein or receptor systems isolated from cell membranes. The macromolecular biological compound is ordinarily immobilized in close proximity to the transducer surface, thereby facilitating direct or mediated sig-

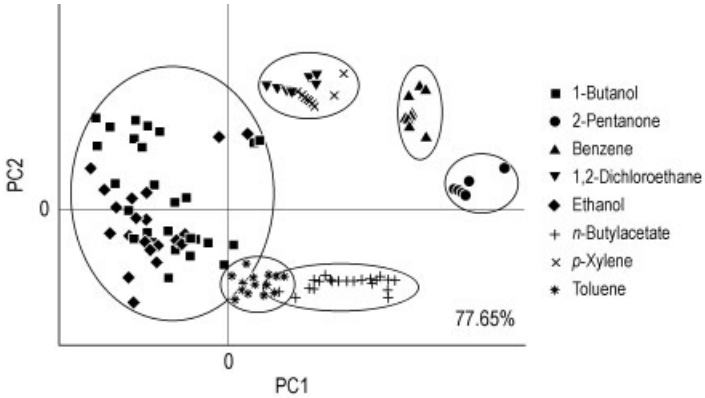

Figure 74. Identifying important BETX compound in the ppm range with only five conductometric gas sensors (ICB-Muenster patent) via principal component analysis (PCA), preliminary results [352]

nal transfer to the transducer. Besides fixing the receptor on the transducer surface this immobilization also serves the important function of stabilizing the biological material.

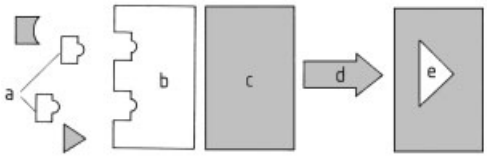

Figure 75. The basic principles of a biosensor
a) Analyte; b) Biological receptor; c) Transducer; d) Electrical signal; e) Signal processing

Table 17. Transducer systems used in the design of biosensors

Transducer	Indicating a change in:	Caused by :
Amperometric electrodes	electron transport	enzyme-catalyzed oxidation or reduction of an analyte
Potentiometric electrodes, ISFETs, MOSFETs, conductometric devices	ion concentration	enzyme-catalyzed oxidation, hydrolysis, or cleavage of C – C, C – O, C – N, or other bonds
Calorimetric devices (thermistors, thermopiles)	enthalpy	enzyme-catalyzed reactions
Piezoelectric crystals	mass	bioaffinity reaction (between Ab – Ag, Ab – hapten, an analyte and receptor protein, or analyte and lectin
Fluorometric fiber optics	fluorescing concentration of enzyme substrates/products or immunolabels	enzyme and bioaffinity reactions
Waveguides	optical path (adlayer)	bioaffinity reaction
Devices based on surface plasmon resonance	resonance (adlayer)	bioaffinity reaction

The first indicating step is specific complex formation between the immobilized biologically active compound (in a broader sense, the receptor R) and the analyte (A). This selective and usually reversible interaction produces secondary chemical or physical signals that are recognizable by a suitable electrochemical, optical, enthalpimetric, conductometric, or piezoelectric transducer. The transducer of choice depends on the nature of the molecular-level binding affinity with respect to the analyte, which may cause changes in such parameters as the redox state of the receptor compounds, the concentrations of electrochemically or optically active substances that are consumed or produced, or heat, mass, or electromagnetic radiation (see Table 17) [354–357].

Depending upon the mechanism of biochemical interaction between the receptor and the analyte, two basic types of biosensors can be distinguished: *biocatalytic* (metabolic) sensors, and *bioaffinity* sensors (Table 18).

Among the analytes detectable with biosensors are carbohydrates, organic acids, alcohols, phenols, amines, heterocyclic compounds, proteins, enzymes, cofactors, inorganic ions, hormones, vitamins, peptides, drugs, xenobiotics, microorganisms, gases, and pesticides. Analyte concentrations measurable with various types of biosensors range from $10^{-7}\,\mathrm{mol\,L^{-1}}$ (biocatalytic sensors) to as little as $10^{-15}\,\mathrm{mol\,L^{-1}}$ (affinity sensors). Based on their level of integration, biosensors can be subdivided into three generations:

1) *First Generation.* The receptor is entrapped between or bound to membranes, and the combination is fixed on the surface of an appropriate transducer or known chemical sensor .
2) *Second Generation*. The receptor is bound covalently to the transducer's surface, thereby eliminating the need for a semipermeable membrane.
3) *Third Generation*. The receptor is bound to an electronic device that transduces and amplifies the signal (e.g., the gate of a field-effect transistor, an array of amperometric ultramicroelectrodes, or interdigitated microelectrodes), utilizing the techniques of microelectronics and micromechanics.

It should be mentioned that this classification is not also meant to be an order of increasing performance. In general rugged systems have more advantages than complicated and expensive constructions.

3.2. Biocatalytic (Metabolic) Sensors

Metabolic biosensors are based on the specific recognition of some enzyme substrate and its subsequent chemical conversion to the corresponding product(s). The biocatalytic reaction itself accomplishes specific binding between the active site of the enzyme and the substrate

(recognition process), conversion of the substrate to product, and release of the product from the active center, which in turn leads to regeneration of the binding affinity and catalytic state of the enzyme.

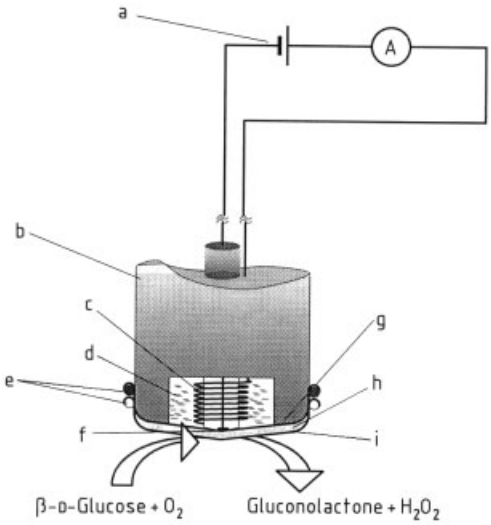

Figure 76. Configuration of an amperometric enzyme electrode for glucose determination
a) Polarization voltage (−600 mV or +600 mV vs. Ag/AgCl); b) Electrode body; c) Ag/AgCl pseudo-reference electrode; d) Electrolyte solution; e) O-Rings; f) Platinum working electrode; g) O_2/H_2O_2-permeable membrane; h) Glucose oxidase (immobilized layer); i) Semipermeable membrane

The linear measuring range for metabolic sensors depends on the "apparent" Michaelis–Menten constant (K_{Mapp}) of the immobilized enzyme (not free in solution and approachable from all directions, thus, "apparent"), which describes a particular substrate concentration associated with half the maximum rate of the enzyme reaction. The reaction rate of an immobilized receptor enzyme is controlled by the rate of substrate and product diffusion through both the semipermeable membrane covering the enzyme and the layer of immobilized enzyme itself, resulting in a higher K_{M} value than that anticipated under conditions of kinetic control. Enzymes, coupled enzyme sequences, organelles, microorganisms, or tissue slices can all be used as receptor materials for metabolic sensors. Under certain conditions, cosubstrates, effectors (inhibitors, activators), and enzyme activities can also be determined by metabolic sensors.

3.2.1. Monoenzyme Sensors

Enzyme sensors rely on an immobilized enzyme as the biologically active component. Many biocatalytic reactions, especially those of oxidoreductases, hydrolases, and lyases, are associated with the consumption or formation of electroactive substrates or products in concentrations proportional to the analyte concentration. For this reason most commercially available biosensors are enzyme sensors based on electrochemical transducers.

Typical electrochemically detectable (co-) substrates and products include oxygen, hydrogen peroxide, hydrogen ion, ammonia, carbon dioxide, reduced or oxidized cofactors, and redox-active (oxidized or reduced) prosthetic groups in oxidoreductases, all of which can easily be converted into electrical signals by suitable transducers (e.g., amperometric or potentiometric electrodes or conductometric sensors, Table 19).

Immobilization of an enzyme in or at an artificial matrix or a membrane usually stabilizes the biocompound, thereby increasing the half-life for activity. Techniques of enzyme immobilization include physical or chemical adsorption, ionic and covalent bonding (perhaps to functionalized transducer surfaces), cross-linking by bifunctional agents, and entrapment in such polymer matrices as natural or artificial gels and conducting polymers. The first biosensor was described by Clark and Lyons in 1962, who introduced the term "enzyme electrode" [358]. In this first enzyme sensor, an oxidase enzyme in a sandwich membrane was placed next to a platinum redox electrode. A platinum anode polarized at +0.6 V (vs. SCE) responded to peroxide produced by the resulting enzyme reaction with the substrate. Application of a polarization voltage of −0.6 V permitted the amperometric detection of oxygen consumption. The primary target substrate was β-D-glucose (Fig. 76). Stoichiometric turnover of the glucose oxidase-catalyzed reaction yields a sensor response proportional to the glucose concentration up to the apparent K_{M} value of the immobilized enzyme. This in turn led to the first glucose an-

Table 18. Biosensor specification according to the nature of the primary sensing reaction [356]

Biocatalytic (metabolic) sensors		Bioaffinity sensors	
Indicating reaction:*			
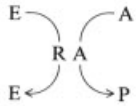			
Analytes	Receptors	Analytes	Receptors
Substrate	enzyme	protein	dye
Cosubstrate	enzyme system	glycoprotein	lectin
Cofactor	organelle	cells	antigen
Prosthetic group	microorganism	antigen/hapten	antibody
	tissue slice	antibody	RNA/DNA
		neurotransmitter	receptor protein
		pesticide	
		enzyme (system)	

*E = Enzyme; A = Analyte; P = Product; R = Receptor; RA = Receptor–analyte complex.

alyzer for measuring glucose in dilute blood on the basis of an amperometric enzyme sensor. A commercial analog, the Yellow Springs Instrument Model 23 YSI, appeared on the market in 1974. The same indicating technique has more recently been applied to many other oxygen-mediated oxidase systems, which constitute the most widely exploited enzyme classes for the construction of biosensors.

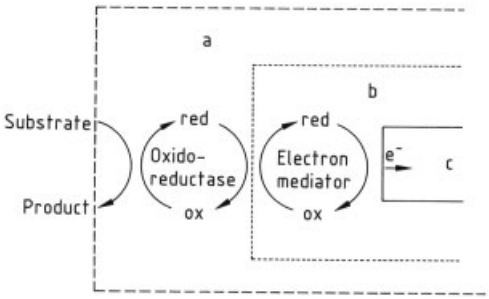

Figure 77. Operating principle of a mediator-modified amperometric biosensor, involving an oxidase or a dehydrogenase (in the case of a reductase, electron transfer would proceed from the electrode to the analyte substrate)
a) Amperometric enzyme sensor; b) Mediator-modified electrode; c) Redox electrode

Amperometric Biosensors Using Artificial Electron Mediators. Amperometric biosensors for monitoring oxidase-catalyzed oxygen consumption or the production of hydrogen peroxide are, under certain circumstances, restricted in their use because of a limited availability of dissolved oxygen. This may necessitate careful pretreatment of the sample with rigorous exclusion of oxygen-consuming compounds, as well as application of a high polarization voltage, which may cause interfering currents in complex matrices.

It should be possible to overcome such problems by direct electron transfer between the redox-active prosthetic group of an enzyme and the electrode surface. In a few cases, especially ones involving small redox proteins such as cytochrome c, efficient electron transfer has been achieved in this way with promoters that direct the active redox centers to the electrode surface. Macromolecular redox proteins exhibit rather low rates of heterogeneous electron transfer because according to the Marcus theory tunneling is not possible over the long distance from the active center to the electrode surface. Thus, the transfer is often accompanied by a large overpotential, but the use of redox mediators may increase the electron-transfer rate dramatically. These low-molecular-mass (in)organic redox species undergo facile reversible redox reactions with both the prosthetic groups of the enzyme and the electrode surface according to the general scheme depicted in Figure 77. The resulting electron shuttle reduces the overvoltage of the biological redox system to the formal potential of the mediator. Suitable redox mediators are redox systems which show a high exchange current density and include quinoid redox dyes,

quinones, organometallic compounds, and fulvalenes (see Table 20) [359–363].

Table 19. Common configurations for monoenzyme sensors

Enzyme class	Detectable compound involved in the enzyme reaction	Typical transducer
Oxidoreductases		
Oxidases	O_2, H_2O_2	pO_2 electrode, noble-metal – carbon redox electrode
Oxygenases	O_2, quinones, catechol	pO_2 electrode, (amperometric) carbon electrode
NAD(P)-dependent dehydrogenases	NAD(P)H/NAD(P)	(amperometric) mediator-modified redox electrode, fluorometer, photometer
PQQ-dependent dehydrogenases	PQQ/$PQQH_2$	
Reductases	NAD(P)H/NAD(P) Heme-, Cu-, Mo-, FeS-containing centers	
Hydrogenases	H_2, NAD(P)H, NAD(P), $FADH_2$, $FMNH_2$, FeS-, Ni-containing centers	
Hydrolases	NH_3/(NH_4)	(potentiometric) ammonia-sensitive electrode
	CO_2	(potentiometric) pCO_2 electrode
	H^+	pH electrode, ISFETs, coulometer, conductometric and enthalpimetric devices
	redox-active products (phenolic, thio-group-containing compounds)	(amperometric) redox electrodes

In constructing such a biosensor the redox mediators can be immobilized at the electrode surfaces in various ways, including chemisorption or covalent bonding, chemical or physical deposition of a polymer film containing the mediator molecules, or physical incorporation of the redox mediators (which are only slightly water soluble) directly into the electrode material. One ingenious approach involves covalent binding of the mediator to the electrode via an appropriate *spacer*, or electrical "wiring" of the enzyme through an electron-relaying redox polymer network. In both cases electron transfer from the redox centers to the electrode surface is thought to occur by electron tunneling [364–366]. The use of a redox mediator increases the dynamic range of a sensor, minimizes interferences, and provides an indicator reaction that is usually independent of the oxygen content of the matrix. Furthermore, such mediators undergo reversible redox reactions with redox-active coenzymes, dehydrogenases, reductases, and hydrogenases, as well as the redox-active products of enzyme- catalyzed reactions, which opens many new possibilities for incorporating these classes of oxidoreductases into new types of biosensors.

3.2.2. Multienzyme Sensors

Coupled enzyme reactions mimic certain metabolic situations in organelles and cells that have themselves already been exploited in biosensors [367], [368]. The use of enzyme sequences facilitates:

1) Broadening the scope of detectable analytes
2) Improving selectivity
3) Increasing both sensitivity and dynamic range
4) Eliminating interferences

Depending on the target analyte, sample matrix, required measuring range, etc., enzymes can be coupled to produce reaction sequences operating either in series or in parallel. An alternative approach to realizing multienzyme sensors is the use of cell complexes, whole cells, or tissue slices comprising physiologically integrated enzyme systems (see below).

Enzyme Sequences Operating in Series. The number of substances that can be determined directly by monoenzyme sensors is limited, because many enzyme- catalyzed reactions involve no physicochemically detectable substances. To form readily detectable compounds it is often necessary, therefore, to couple together several different enzyme reactions in a sequential way. The analyte of interest is first converted by some specifically catalyzing enzyme into a product that is itself convertable in the course of one or more subsequent enzyme reactions involving detectable secondary (co)substrates or products. Biosensors based on such enzyme sequences have been developed for the determination of many substrates, including disaccharides (sucrose, maltose, cellulose, lactose), cholesterol, creatinine, and nucleic acids [356].

With certain enzyme sequences it is possible to achieve chemical signal amplification

Table 20. Redox mediators leading to efficient electron transfer with redox proteins

Mediator type	Examples	Structures
Quinones	benzoquinone	
Quinodimethanes	tetracyanoquinodimethane (TCNQ)	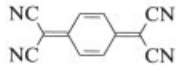
Tetrathiafulvalene (TTF) and its derivatives		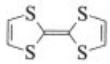
Metallocenes and their derivatives	ferrocene	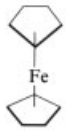
Phthalocyanines (Co, Fe, Mn, Cu)	cobalt phthalocyanine (Co-PC)	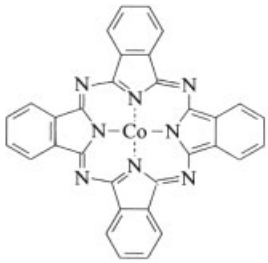
Quinoid redox dyes		
Phenazines	*N*-methylphenazonium (NMP^+) salts	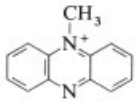
Phenoxazines	meldola blue	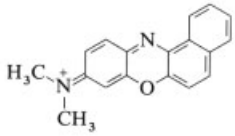
Phenothiazines	methylene blue	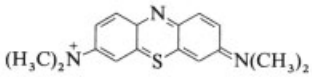
Indamines	toluylene blue	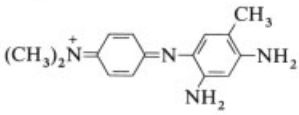
Viologens	methyl viologen	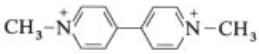
Complex compounds		
Ruthenium complexes		$[Ru(NH_3)_6]^{3+/2+}$, $[Ru(en)_3]^{2+/3+}$
Osmium complexes		$[Os(bpy)_2Cl]^{3+/2+}$
Cobalt complexes		$[Co(bpy)_3]^{3+/2+}$, $[Co(en)_3]^{2+/3+}$
Iron complexes		$[Fe(CN)_6]^{3+/4+}$, $[Fe(phen)_3]^{2+/3+}$

through the accumulation of a particular product by cycling the substrates or cofactors. The sensitivity of some biosensors can be enhanced in this way by as much as three orders of magnitude. Signal-amplifying enzyme systems have been based on oxidase – dehydrogenase, dehydrogenase – transferase, and phosphorylase – phosphatase – oxidase sequences, permitting the determination, for example, of lactate, ATP, and phosphate at the picomole per liter level (Fig. 78) [357–369].

Enzyme Sequences Operating in Parallel. Multienzyme sequences operating in parallel are used for the elimination of interfering agents, expansion of the dynamic measuring range, and (in principle) the indirect detection of an electrochemically inactive reaction catalyzed by one of the enzymes in a substrate competition. *Anti-interference* enzyme systems are the most important. Substances interfering with an indicating reaction can be eliminated as one way of improving the analytical quality of the measurement by covering the immobilized enzyme with an anti-interference enzyme layer, in which a second enzyme converts the disturbing substance into an inert product. This strategy has made it possible to remove such interfering agents as ascorbic acid and endogenous glucose [367].

Biosensors Based on Integrated Biological Enzyme Systems. Cellular organelles and membranes, as well as whole cells or tissue slices from animal or plant sources comprising physiologically coupled multienzyme systems are often employed as receptor materials in biosensors based on electrochemical detection. Native physiologically integrated enzyme systems display enhanced stability due to their in vivo environment. Biological material of this type offers the potential for low cost and simple preparation. Furthermore, this opens the way to analytes that are converted to electrochemically detectable products not by simple enzyme reactions, but rather by complex multienzyme reactions within an intact biological structure. Analytes subject to such receptor systems include chloroaromatics, caprolactam, naphthalene, tryptophan, several hormones, and acetic acid [354], [370].

3.2.3. Enzyme Sensors for Inhibitors – Toxic Effect Sensors

Sensors for inhibitors are based on competitive inhibition between a substrate and an inhibitor for an enzyme binding site, inhibition of an enzyme reaction by product accumulation, irreversible binding of an inhibitor to an active binding site, or redox-active heme or SH groups of enzymes that lead to the blocking of enzyme activity.

When thylakoid membranes, protoplasts, or whole cells are used as receptor materials, various inhibitors cause metabolic "short circuits," or inhibition of the (photosynthetic) electron-transport system, which results in a decrease in such metabolic parameters as oxygen consumption or production, photocurrent development, dehydrogenase activity, or fluorescence. In contrast to metabolic sensors, the first sensing step here is a result of binding of the inhibitor to an active binding site or redox center on the enzyme. Sensors of this type usually operate at low enzyme-loading factors and with substrate saturation in order to ensure a highly sensitive signal of limitation in the kinetically controlled reaction due to the presence of an inhibitor. Examples of such inhibitors (analytes) include toxic gases, pesticides, chloroorganic compounds, phenols, and heavy-metal ions [371], [372]. To detect toxic effects on vital enzymes in a reproducible manner, to include also synergistic or antagonistic effects, is only possible with biosensors or bio-tests. This advantage as a fast warning device is still not fully used. Too often biosensors are developed as if they should compete with instrumental analysis. Given the pace of development in this area, especially concerning miniaturization, speed, and versatility, the biosensor as a mono-analytical device will have problems in gaining a larger market share. However, as a toxic warning device triggering some screening and sampling it is without competition.

3.2.4. Biosensors Utilizing Intact Biological Receptors

Molecular receptors are cellular proteins (often membrane-bound) that bind specific chemicals

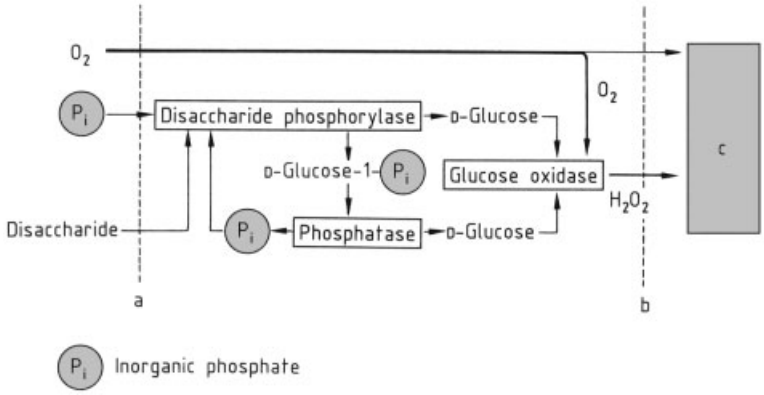

Figure 78. Enzyme sequence for the detection of inorganic phosphate, based on the principle of chemical signal amplification [369]
a) Dialysis membrane; b) H_2O_2/O_2 permeable membrane; d) Working electrode

in such a way as to produce a conformational change in the protein structure. This conformational change in turn triggers a cellular response: opening of an ion channel, for example, or secretion of an enzyme. In general, molecular receptors can be distinguished from other types of receptors (including larger multicomponent systems) on the basis of their composition, which is a single protein, though this protein may contain more than one subunit. Among the important receptors known to date are receptors for hormones (control agents for a wide range of cellular and body functions), amino acids, insulin, and neurochemical transmitters, in addition to receptors capable of binding synthetic bioactive chemicals, such as drugs [373], [374].

Molecular receptors have two important properties relevant to their possible incorporation into biosensor devices: intrinsic signal amplification and high specificity. The specificity is a consequence of a highly evolved binding region within the receptor that is fully optimized for binding one particular ligand by a variety of electrostatic, hydrogen bonding, van der Waals, and hydrophobic forces.

Limitations affecting the study and use of molecular receptors include:

1) Relatively difficult experimental procedures for the isolation of the receptor proteins (even prior to purification), which are both time- consuming and expensive, and often require animal sources
2) An absence of techniques for obtaining more than very small amounts of pure receptors
3) The rapid loss of biological function that usually follows isolation

Molecular biology has brought considerable progress in the isolation and purification of receptors, while advances in artificial membrane technology and reconstitution have reduced the problem of instability.

Comparatively few biosensors of this type have been reported [373], [374], primarily because of problems associated with the successful isolation of molecular receptors (in sufficient quantity and purity) and their subsequent stabilization in artificial environments in such a way that appreciable biological activity is retained.

A clear exception, however, is a fiber-optic sensor based on evanescent fluorescence developed by Eldefrawi et al. [374], [375] that contains nicotinic acetylcholine receptor (nAChR). This was the first neurotransmitter receptor to be purified and characterized in vitro, and it is involved in chemical signalling. The sensor in question consisted of nAChR immobilized on the surface of a quartz fiber. Determination of the concentration of acetylcholine and other natural neurotransmitters (e.g., nicotine) was accomplished by measuring the extent to which they inhibited the binding of a standard solution of a fluorescence-labeled toxin (fluorescein isothiocyanate-α-bungarotoxin, FITC-α-BGT). Binding determinations were based on measurements of fluorescence intensity in the

region of the evanescent wave (that is, at the fiber – solution interface, see Section 2.3.2.1), where the nAChR was immobilized. The sensor had an acceptable signal-to-noise ratio of 10^2 and an analysis time < 10 min (based on the initial rate of signal increase rather than a steady-state value), and it was capable of operating under conditions of high ionic strength. However, it suffered from the problem of non-regeneration of the sensing element because of the irreversibility of FITC-α-BGT binding to the nAChR.

A similar receptor-based biosensor using electrochemical detection was reported for the detection and measurement of riboflavin (vitamin B_2) [373], [374]. This device is based on the competition for aporiboflavin-binding protein (apoRBP) between riboflavin present in the sample subject to analysis and a riboflavin analogue bound in a membrane. In the absence of riboflavin all the apoRBP is bound to the analogue on the membrane. When riboflavin solution is added, the five-fold greater affinity of apoRBP for its natural cofactor (riboflavin) results in displacement of apoRBP from the membrane into the solution, causing a change in the electrical potential of the membrane. A riboflavin measurement range of $0.1 - 2 \times 10^{-6}$ mol/L was reported.

Further progress in molecular receptor-based sensors will be a function of advances in molecular biology with respect to enhancing the quantity, purity, and stability of receptors that can be used in vitro.

3.3. Affinity Sensors – Immuno-Probes

Affinity sensors take advantage of physicochemical changes that accompany complex formation (e.g., changes in layer thickness, refractive index, light absorption, or electrical charge). Once a measurement has been made the initial state must be regenerated by splitting off the complex. Because of this they cannot be used for continuous monitoring nor to follow declining concentrations of the analyte. Therefore, they cannot be named biosensors into the strict sense of the I.U.P.A.C. definition to be "reversible". Thus, they are named immuno-probes. Since the splitting of the complex can be very fast (e.g., with pH < 2.0 or chaotropic reagents) and with the process being automated (e.g., in FIA-systems) the response time of an immuno system can reach less than 1 min and, thus, becomes comparable to enzymatic biosensors working under diffusion control.

Examples include immunosensors utilizing antibodies or antigens for recognition purposes, lectins, and "true" biological receptors in matrix-bound form (see Section 3.2.4).

Immunosensors Based on Antibody Molecules. Biosensors based on the antigen – antibody interaction are called *immuno-probes.* The high specificity and high sensitivity that typifies an antigen – antibody (Ag – Ab) reaction has been used in a vast range of laboratory-based tests incorporating antibody components (immunoassays). The analytes detected and measured have included many medical diagnostic molecules, such as hormones (e.g., pregnancy-related steroids), clinical disease markers, drugs (therapeutic and abused), bacteria, and such environmental pollutants as pesticides. A crucial distinction must be made between an immunoassay and an immuno-probe, however, with respect to the technology relevant to the analysis. The ultimate aim is a low-cost, disposable immuno-probe, with the entire device being used only once. For example, it has now proven possible in practice to develop an instrument containing a disposable "sensor chip" that can be used for Ab or Ag tests.

Two different types of immuno-probes have been described:

1) *Direct immunosensors* register immunochemical complex formation at the transducer surface via electrochemical, mass or optical changes. The advantage of a direct immunosensor is that measurement of the antigen – antibody interaction can be accomplished immediately without any need for additional antibodies or markers (enzymes, fluorescent labels, etc.).
2) *Indirect immunosensors* are based (like homogeneous and heterogeneous immunoassays) on labeling of one of the partners in an immune interaction. One class of indirect immunosensors derived mainly from enzyme immunoassays (EIA) encompasses the enzyme immunosensors. These combine the

high selectivity of an immunoreaction with the high sensitivity characteristic of the enzymatic amplification effect. However, the enzyme can affect the molecular recognition step because of its large size and needs in general special attention (cooling) during storage. In the case of the widely used assay technique of Enzyme Linked Immuno Sorbent Assay (ELISA) the use of enzymes generates two further manipulations: a) the addition of a suitable sustrate and b) the addition of a reagent which stops the enzymatic reation. This is not necessary when using modern fluorescence markers.

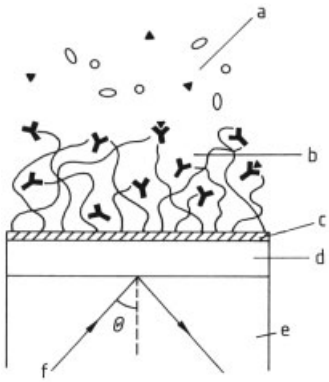

Figure 79. Operating principle of an SPR immunosensor a) Bulk solution; b) Carboxymethylated dextran matrix; c) Linker layer, $HO-(CH_2)_n-SH$; d) Gold film; e) Glass substrate; f) Light
(with permission from GEC Journal of Research, Essex)

3.3.1. Direct-Sensing Immuno-Probes without Marker Molecules

Among direct immunosensors the most favored are optical sensors, because they incorporate the advantages of fiber optics or optoelectronic transducers. The nonelectrical properties of optical systems offer important safety advantages for medical, industrial, and military applications. Remote sensing is clearly advantageous, for example, for explosives and toxins. Optical-fiber probes are also mechanically flexible, small, inexpensive, and disposable.

The detection of an evanescent field in an optical waveguide (see Section 2.3.2.1) is a highly sensitive technique especially suitable for monitoring the antibody – antigen interaction [373], [374]. One partner of the immuno pair (e.g., the antibody) is immobilized at the surface of the waveguide, and its reaction with the other partner in the test sample is monitored. An evanescent wave generally penetrates a distance of several hundred nanometers extending the full length of an antibody – antigen complex, so in contrast to conventional techniques there is no need to carry out a prior separation of the nonspecific biological components. There is no need for washing steps.

The technique of surface plasmon resonance (SPR) represents a very sensitive way to follow an immuno-reaction on an interface , and its feasibility has been adequately demonstrated [376]. As noted previously (Section 2.3.2.3), SPR is a phenomenon produced by a beam of light directed onto a glass – metal interface, usually a glass prism in contact with a layer of gold or silver (Fig. 79). At some specific angle (the resonance angle) a component of the electromagnetic light wave propagates in the metal along the plane of the interface, taking the form of surface plasmons. The resonance angle is quite sensitive to changes in refractive index and dielectric constant at the interface up to a distance of several hundred nanometers from the actual metal surface, with the sensitivity decreasing exponentially as a function of distance from the surface. Immobilization of an antibody on the surface causes a measurable shift in the resonance angle, and binding of a high-molecular antigen to the immobilized antibody leads to a further change. For typical biological systems this binding-induced shift in resonance angle (expressed in resonance units) is approximately linearly proportional to the concentration of bound antigen (or antibody if antigen has been pre-immobilized).

The SPR-based biosensor systems BIAcore and BIAlite, developed by Pharmacia Biosensor AB and now available commercially, represent a breakthrough in immunosensor technology, particularly in the use of SPR techniques. Apart from high cost, the only limitation associated with this technique is that the sensitivity depends on the molecular mass of the adsorbed layer, which in turn controls the optical thickness, so low concentrations of small molecules (molecular mass < 250, including most haptens) are unlikely to be measurable directly without additional labeling. In

addition to SPR, Mach–Zehnder interferometers [377], RIFS [201], [202], and so-called grating couplers [378] are also used for optical immunosensing. An interesting step towards miniaturization was introduced by the development of practicable fiber optical SPR systems by Jorgenson and Yee [379] and Katerkamp et al. in the ICB Muenster [380], [381]. The latter device was so small and low-cost that application as a hand-held refractometer with a resolution in the fifth digit after der decimal point was feasible. The fiber optical SPR tips were rapidly exchangeable and bending of the fiber had no negative influence on the result.

All the above mentioned transducers measure changes in the refractive index and layer thickness together! With respect to quantitative assays one should bear in mind that those techniques only sense a mean layer thickness. Thickness extensions that occur because of large macromolecules are leveled out. Futhermore, the change of the refractive index inside the evalescent field (within the surface layer) may also depend on ions or water molecules entrapped in the proteins. The latter is dependent on the ionic strength. Thus, changes in the ionic strength of the sample solution may also produce a signal. In addition, if antibody molecules are immobilized on the transducer surface, how can anyone differentiate between an immunoreaction with the analyte or a change in the tertiary and quarternary structure of the antibody molecule caused by other effects giving rise to a change in the thickness of the layer?

Alternative transducer technologies that have been applied in immunosensor research include piezoelectric and electrochemical systems. *Piezoelectric* immunosensors [383], [384] (see Section 2.3.3) tend to suffer from significant levels of nonspecific binding to the piezoelectric substrate, which makes accurate analyte quantification difficult. However, this type of immunosensor may have an important role to play in the search for traces of volatile drugs in security monitoring, and it offers the advantages of relatively small size and low manufacturing cost. The main disadvantage from the standpoint of immunosensing is that reproducible measurements have so far been achieved only in the gas phase, whereas the Ab–Ag interaction is restricted to aqueous solution. As has already been mentioned above, working with mass-sensitive transducers in liquids gives rise to doubts about the correct interpretation of frequency shifts. Here, the same as in case of the optical transducers, they need to be differentiated from real mass changes. Any structural changes of large protein molecules which induce associated changes of entrapped water and ions will result in a measurable signal. Despite the fact that these frequency changes have nothing to do with specifically bound analyte molecules these signals can be very reproducible. Without a thorough and convincing validation with real samples, the analytical accuracies are questionable. This is not to say that qualitative binding studies or determination of association constants could not be measured. All transducers needing no label compound are ideal for screening purposes. They are mostly used for the detection of specific receptor functions.

Potentiometric immunosensors [356] (see Section 2.3.1.1) are unfortunately associated with a low signal-to-noise ratio as a result of the low charge density on most biomolecules relative to such background interferences as ions. These sensors also show a marked dependence of signal response on sample conditions, including pH and ionic strength. Similarly, *immunoFET* devices (see Section 2.3.1.4), which measure very small changes in an electric field when an Ab–Ag binding reaction occurs, tend to suffer from practical problems associated with membrane performance and various artifacts. The theoretical basis of the sometimes observed voltage change remains unclear. The only plausible explanation for the fact is that clearly the binding of a neutral antigen molecule to a large antibody molecule with multiple charges (depending on the pH and isoelectric point) may change the charge distribution at the interface or may lead to a different mixed potential. If the whole immuno-reaction producing an analyte proportional signal is not watched carefully and controlled by convincing validation studies, a certain bias can easily be overlooked by all so called direct sensing techniques.

Main Drawback of all Direct Sensing Techniques. It has become a well known fact that the sensitivity (or LOD) of all direct sensing techniques is not any signal-to-noise ratio nor any theoretical mass-loading sensitivity! It is the lack of control over non-specific binding processes, which are not caused by the very

specific molecular interaction under study. All direct sensing immuno-techniques cannot differentiate between competing surface processes leading to adsorbed particles with various heats of adsorption. Even if working with very hydrophilic surfaces to prevent this, the ultimate LOD of such an immuno-probe depends, according to the laws of Analytical Chemistry, on the standard deviation of such non-specific binding (LDO = 3 x σ_{blank})! This renders many extreme sensitive transducers useless for practical applications in real world samples.

Amperometric or fluorescence-based immunotechnology (immuno-electrodes or capillary fill devices with a fluorescence label, for example) has shown more promising results at the research prototype stage than other types of transducers. One preferred technique continues to be the indirect amperometric immunosensor approach described below, which uses enzymes as markers. The second emerging technique is based on fluorescence dyes which are now commercially available together with a simple covalently binding protocol (one to two steps only). Fluorescence can be as sensitive as enzyme-based amperometry especially if modern fluorophores are used with excitation wavelengths in the near IR and the emission shifted about 100 nm to higher wavelengths. Around 600 – 800 nm is an observation window for biological systems since no natural fluorescence (e.g., tryptophane) occurs in that region. Furthermore, cheap laser diodes of around 680 nm with high energy are available.

3.3.2. Indirect-Sensing Immuno-Probes using Marker Molecules

The most satisfactory marker molecules for use in this context have been fluorophores and enzymes. As regards the sensors themselves, several fundamental types exist.

Reactor immuno-systems are based on a cartridge or microcolumn packed with appropriate immobilized antibodies (which serve as so-called catching antibodies). As in the case of enzyme immunoassays, the competition, sandwich, and replacement principles are all subject to exploitation. The amount of bound (or replaced) antigen is determined via the extent of enzyme activity remaining after substrate addition, which can be established with an electrochemical or optical sensor. For many applications (on-line measurements, coupling with flow-injection analysis), repeated and automated measurements are required. Because of the extremely high loading of catching antibodies, reactor immunosensors have the advantages of high sensitivity, regenerability, and reproducibility. On average they can be reused 50 – 100 times without loss of activity. The US Naval Research Laboratories were very successful with respect to sensitivity, selectivity, reliability, and simplicity with several prototypes of reactor-based systems they developed under the guidance of Frances Ligler [382]. The developed immuno-systems allow field determinations and have been extensively tested and validated in the field of environmental analysis. In this respect, it is worthwhile to mention that the US EPA has adopted immuno-kits for certain important analytes in the environment, e.g. gasoline spills. If the fluorophore is carefully selected sensitivities in the ppb range with a FIA-like set-up and response times under 1 min are possible!

Membrane immuno-probes employ antibodies immobilized directly at the surface of a sensor, or at a membrane covering the sensor, as shown in Figure 80. An example of the use of this type of immunosensor is illustrated schematically in Figure 81. The analyte to be detected is an early myocardial infarction marker, the fatty acid binding protein (FABP). A modified Clark electrode is covered by a membrane bearing immobilized monoclonal catching antibodies directed against FABP. After binding of the analyte and washing, a secondary polyclonal antibody is added to the solution. This secondary antibody is conjugated with glucose oxidase (GOD). The next step is incubation followed by further washing, after which glucose is added to the measuring solution. Glucose is oxidized by GOD in an amount directly proportional to the amount of FABP, yielding H_2O_2, the concentration of which is monitored with an amperometric sensor, as shown in Figure 82. This system was developed at the ICB Muenster with the aim of developing a rapid test for a myocardial infarct to be performed actually in the ambulance with only one drop of blood. It works fully automatically. One drawback was, however, the use of a labile enzyme. Despite the fact that GOD was the most stable known enzyme, some pre-

cautions had to be taken and also the need for substrate addition and stopping of the reaction needed computer-controlled timing.

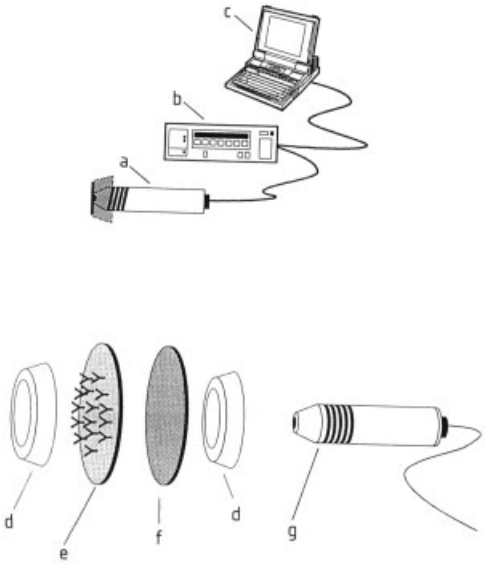

Figure 80. Design principle of an immunosensor system a) Immunosensor; b) Electronic unit; c) Computing unit; d) O-Rings; e) Catching antibody membrane; f) Dialysis membrane; g) Pt vs. Ag/AgCl electrode

One example of a nice commercial immunosensor is produced by Serono Diagnostics. This fluorescence-based evanescent-wave immunosensor [374], [375], [376] incorporates a novel capillary-fill design, and is shown in Figure 83.

The system consists of two glass plates separated by a narrow capillary gap of 100 mm. The lower plate acts as an optical waveguide and contains on its surface an immobilized layer of antibodies. The system benefits from its capillary-fill system, by means of which a fixed volume of sample is drawn into the space between the plates regardless of the volume of the bulk sample (a blood droplet for example). Delivery of a highly reproducible volume of sample eliminates a significant source of error in analyte measurement. The Serono immunosensing format constitutes a competitive assay, though a sandwich-type assay could also be used. The upper plate contains on its surface a layer of fluorescence-labeled analyte trapped in a water-soluble matrix. On addition of sample the labeled analyte is released into solution, where it competes for antibody sites on the lower plate with any unlabeled analyte present in the sample. After a fixed incubation period the fluorescence signal of the evanescent waves is measured and related to analyte concentration.

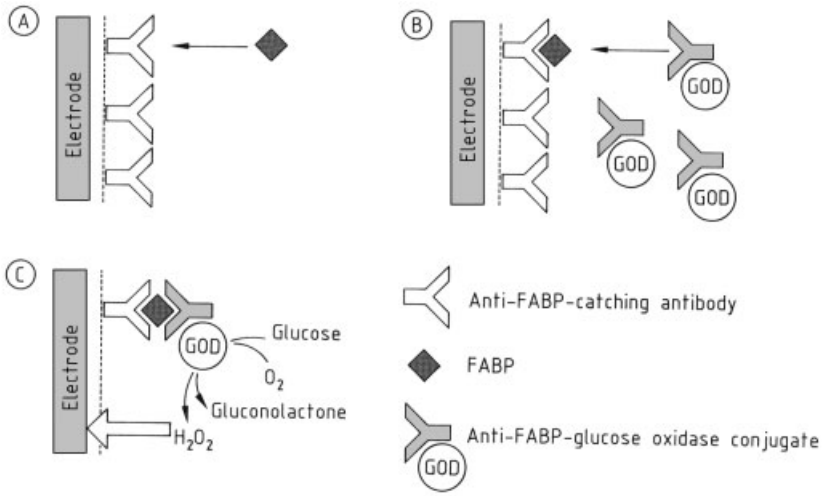

Figure 81. Schematic representation of the operation of an immunosensor for detecting the fatty acid binding protein (FABP) A) Analyte binding step; B) Introduction of a secondary polyclonal antibody conjugated with glucose oxidase (GOD); C) Determination of GOD activity after addition of glucose

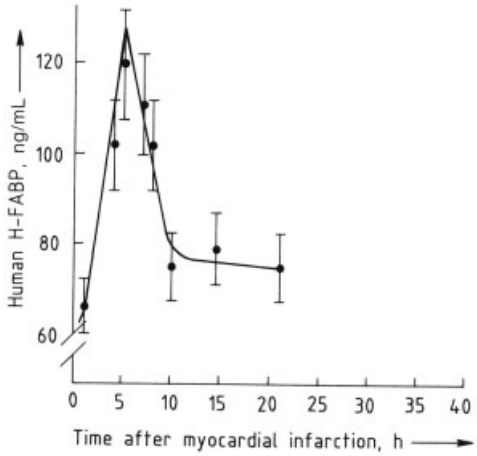

Figure 82. Analytical results obtained with the human heart FABP sensor described in Figure 81 as a function of time after onset of a myocardial infarction

This technology does have shortcomings, however, some of which are related to the practical problems associated with carrying out measurements on whole blood. These include poor capillary flow because of the "stickiness" of blood compared with water, increased incubation time resulting from slow dissolution of the labeled analyte matrix, and the presence of fluorescent and highly colored molecules in blood, such as hemoglobin. Drawbacks include the necessity for an incubation period of several minutes (due to Ag – Ab interaction) and increased manufacturing costs (each analyte to be tested requires a dedicated sensor, which involves labeling and immobilization prior to physical construction).

While immunological analysis with one of the different ELISA methods works very well and reliably in the laboratory, the fabrication of pre-calibrated small stick-like probes has usually failed until recently. The main problem was the lack of information one has about the surface density of active immunopartners and a way of reproducibly manufacturing those single-use devices with antibodies or analyte molecules

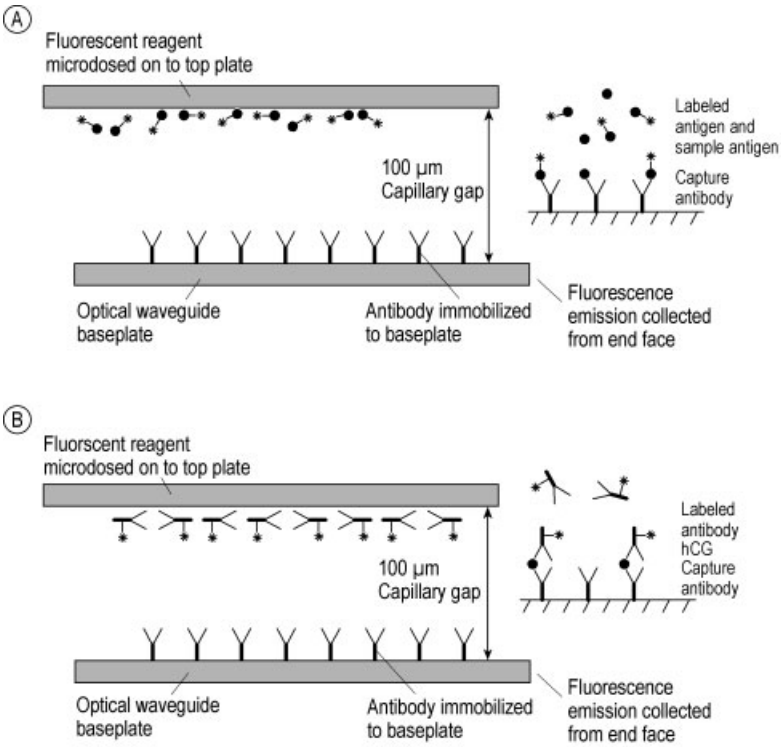

Figure 83. Principle of the patented Capillary-Fill Device. The upper part shows schematically a competitive immuno-assay; the lower part a sandwich assay. The optical waveguide is always the lower baseplate. Here a diode laser beam is used to produced an evanescent field in the sample inside the capillary.

bound to the surface. Thus, one or more calibration solutions to be run simultaneously on all surfaces with immobilized biomolecules were always necessary. Only if full control of the surface density and state (active or not, freely accessible or not, etc.) were be given pre-calibrated immuno-probes for such fields other than clinical diagnostic and environmental field analysis with quantitative results would be feasible.

The Integrated Optical System – IOS. The solution to overcoming the main drawbacks of all hitherto developed biosensors, with the exception of the glucose sensor (poor storage capability), is a single-use sensor chip on which the labile biomolecules could be kept active forever by freeze-drying them. The solution to overcoming the pre-calibration problems is to employ a kinetic data evaluation method. The solution to overcoming the mostly diffusion limited immuno-reaction at an interface is to reduce the Nernstian diffusion layer thickness significantly so that differences in the association kinetics become apparent. To overcome poor sensitivity the solution is to use stable fluorophores with high quantum yield and emission in the near IR. In order to overcome washing steps, the evanescent field method for excitation and back-coupling has to be employed used. The IOS-chip (developed at the ICB Muenster) consists of a plastic baseplate with a refracting index of about 1.50. The mostly aqueous samples lie in the range of about 1.33. Thus, if the interface plastic base sample solution is illuminated, total reflection takes place and the light is reflected back into the plastic chip material. However, the reflected light does not fall abruptly to zero in the sample phase but fades exponentially (evanescent field). The evanescent field reaches out into the sample medium only about 100 nm. The first 20 nm contain about 20 % of the optical power radiated onto the interface. The dimensions of most antibodies lie in a range between 5 and 7 nm. This means that only those fluorophore-labeled antibodies that are located directly on the surface are excited to fluorescence. Labeled but not specifically bound antibodies in the sample volume above the interface are not excited. Together with further improvements the IOS technology was patented wordwide by the ICB Muenster (prior art: DE 19628002 A2/DE 19711281 A2 inventors: Katerkamp, Meusel, Trau).

The plastic base chip, onto which a diode laser beam (λ near the absorption wavelength of the fluorophore) is located, is directed from underneath. The corresponding antibodies or antigens are immobilized on the chip (about 30×10 mm) surface in a simple and straightforward batch process of several thousands or more. Then, in order to reduce the Nernstian diffusion layer thickness a very thin double adhesive spacer (about 50 µm or lower) and a cut-out sample flowing channel is put onto the side with the immobilized molecules. A plastic cover with two openings for the sample in- and outlet is put on forming the necessary narrow detection channel for the hydrodynamics. The sample is brought into the measuring channel by suction forces applied at the outlet within the apparatus. However, on the sample inlet side of the chip, an easy to change container (dwell) is located. On its wall all necessary reagents, buffer, etc. are freeze-dried. The type of recognition molecule is determined by the method. In general all known ELISA methods are feasible with higher speed and sensitivity. Of particular interest are generic chip designs in which, e.g., a labelled antibody (e.g., with a Cy5 kit) reacts with the antigen of the sample in the small dwell at the inlet and is then filtered into the latter via a membrane filter with immobilized antigen. Thus, the excess antibodies are not passing through since they are captured on the filter surface. However, the antigen – antibody complex reaches the optical channel after application of a slight vacuum on the outlet side. Here a catching layer for general antibody recognition makes the whole device generic! The dynamic measurement of the slope of the increasing fluorescence signal falling onto a simple optical detector below the chip is important. In the described operational mode this slope strictly depends on the concentration of the antigen. This slope can be measured in a few seconds! In cases of extremely low levels, the measuring time can be increased. By this kinetic evalution and viewing of the whole active area the inactivation of receptor molecules on the surface is more or less evened-out. The slope depends much more on the concentration of the partners and the constant hydrodynamics.

This IOS technology has recently been licensed to several companies in the field of diagnostics and will change the future of biosensing because of its inherent advantages:

- no washing and separation steps are needed for performing solid phase immunoassays in this arrangement
- reduction in measuring time from several hours to a few seconds
- dramatic increase in sensitivity (< ppb)
- automatic computer control of the whole process:
  1) insert the analyte specific sensor chip in the read-out unit
  2) fill the sample container (sample volume needed for slope determination < 200 µL)
  3) start the measurement process (chip moves into read-out unit and the outlet is punctured (closed previously to prevent uncontrolled intake by capillary forces) and the constant metering pump forces the sample through the measuring channel.

The low-cost chip manufactured in mass-produced plastic injection molding can also contain several functional layers for optimum sample treatment:

- adjustment of pH and/or ionic strength
- filtration of sample (with or without analyte coating)
- separation of cellular blood
- components
- release of further immuno reagents or hydrodynamic puffers.

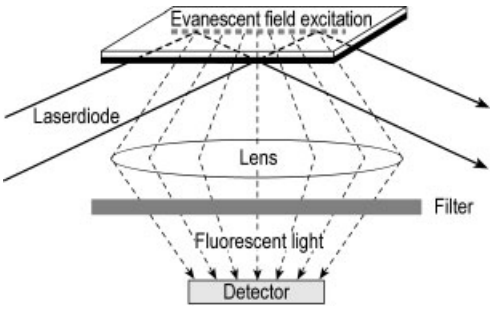

Figure 84. Optical set-up and construction of the IOS-chip. The laser light from a low-cost laser diode is directed towards the lower plastic baseplate of the immuno-probe chip. The evanescent field is produced within the flow channel produced by the spacer and the cover with sample entrance and outlet holes. The fluorescence of labeled molecules bound to the surface of the baseplate is collected, after passing an interference filter corresponding to the fluorescence wavelength, by a low-cost detector.

The Figure 84 shows the IOS chip design and Figure 85 the simple optical set-up. In Figure 86 a typical calibration plot for the quantification of the important pregnancy hormone hCG is presented. Note: This calibration curve was obtained by using single-use sensors at the corresponding concentration. The error brackets show the standard deviation obtained in each case with five different IOS sensors from one production batch!

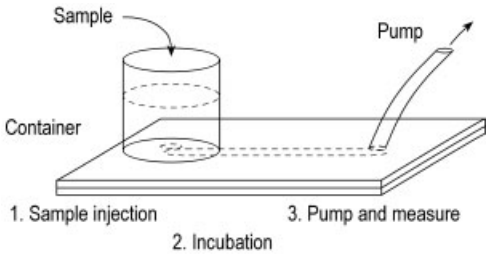

Figure 85. Simplicity of the whole measuring process. In the sample container (small dwell) all necessary reagents are present. After sample injection the reagents (with appropriate antibodies) dissolve in the sample and after a variable incubation for the immunological reaction the IOS-chip is automatically transported into the readout instrument where the outlet is punctured and a pump starts to suck the sample through the narrow measuring channel.

What is evident from that calibration curve is its extremely high dynamic range, not typical in the ELISA techniques! The reproducibility can be brought well below 5 % relative, which is outstanding in the trace analytical range. The matrix effect between an aqueous buffer solution and urine can be about 10 % (less) and between buffer and serum about 20 % (less) relative. Both can be compensated for with sufficient accuracy.

Outlook for the IOS technology-DNA-Chip. The IOS technology developed at the ICB Muenster is a base technology with high market orientation, with only as much sophisticated high-tech as needed, optimized for ease of production and quality control, increased production units by generic approach (one base chip for antibody capturing can be used for a wide number of different analytes depending only on the sample container at the inlet). All these advantages can even be topped by the fact that with another optical system delivering a focussed image of sensing spots onto a CCD camera, multi-analyte detection is feasible! Even further, fluorescence-based hybridization events are also sensed with the multi-analyte IOS technology. The upper limit is only given by patent violation reasons and lies at 400 detection spots. This applica-

tion can further trigger mass-production and has the great advantage compared with other DNA-arrays that it quantifies automatically in real time!

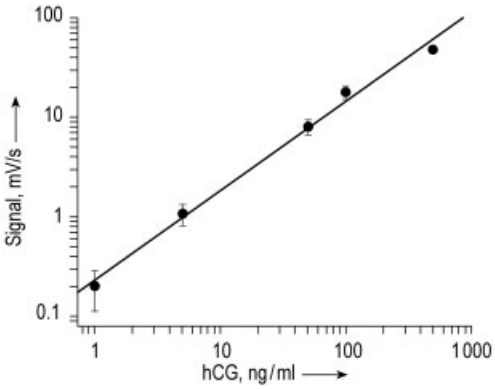

Figure 86. Calibration curve for the pregnancy hormone hCG. Each measurement was made with a new IOS sensor. The change of the fluorescence signal is plotted vs. the concentration.

3.4. Whole-Cell Biosensors

Biological cells can be regarded as small membrane-bound structures containing a high concentration of chemicals, including enzymes, nucleic acids, ions, many types of protein, small organic molecules, and a host of others. A major challenge in exploiting whole (intact) cells in biosensor research is the problem of achieving selective transduction of a specific biochemical event or process within the context of this complex cellular chemistry [357], [370], [373], [374], [386] (Fig. 87).

The sensitivity and specificity of a particular cell with respect to a particular chemical varies enormously depending on the type, source, and environment of the cell and the precise nature of the chemical itself. It is not possible to provide any universal guidelines beyond such generalizations as the observation that most mammalian cells respond to a range of highly toxic chemicals, including cyanides. Even in this case it is known that hydrogen cyanide can interact with a number of different cellular components (such as electron transport proteins and enzymes) because of its small molecular size (which facilitates diffusion through biological membranes) and the ability of the cyanide anion to bind strongly to a wide variety of biomolecules.

The stability of biological cells is of critical importance in biosensor applications. In general, cells derived from higher organisms are rather difficult to isolate and maintain in a viable form for the extended periods of time required in most sensor formulations. The most biologically robust cells are derived from less complex organisms, particularly yeast, bacteria, and algae. Cells from these organisms can generally be stabilized in an in vitro environment more easily than complex cells; indeed, many of the organisms themselves are unicellular (yeast, bacteria, algae), leading to biosensors incorporating truly "living" biological components.

Natural cell differentiation and the growth characteristics of such whole-cell biosensors sometimes cause problems associated with signal drift over time. Nevertheless, there are potential advantages to be gained by exploiting whole cells in a sensor format, including:

1) Versatility (cells respond to a wide range of chemicals)
2) The presence in situ of biological structures supporting multicomponent or multistep biological reaction sequences useful for detection purposes and ensuring optimum enzyme activity (many enzymes lose activity when isolated in vitro)
3) Low preparation cost; i.e., growth in culture

For the reasons outlined above, nearly all whole-cell biosensors reported to date have employed relatively simple bacterial or algal cells [372], [387–390]. Existing whole-cell biosensor technology has also been based exclusively on electrochemical transduction of cellular responses.

The increasing interest in whole-cell biosensors since the first reports by Divies [387] and by Rechnnitz [388] has led to the development of a large number of analyte-specific amperometric and potentiometric sensors. In these devices an immoblized layer of whole cells (entrapped in a gel, for example, or membrane-supported) is placed at the surface of an electrode. The cells then function as an enzyme source, where the enzyme of interest is present at a high relative concentration and is effectively able to function in its natural optimized environment. Bacterial cells in particular may contain very large amounts of particular enzymes relative to all others, espe-

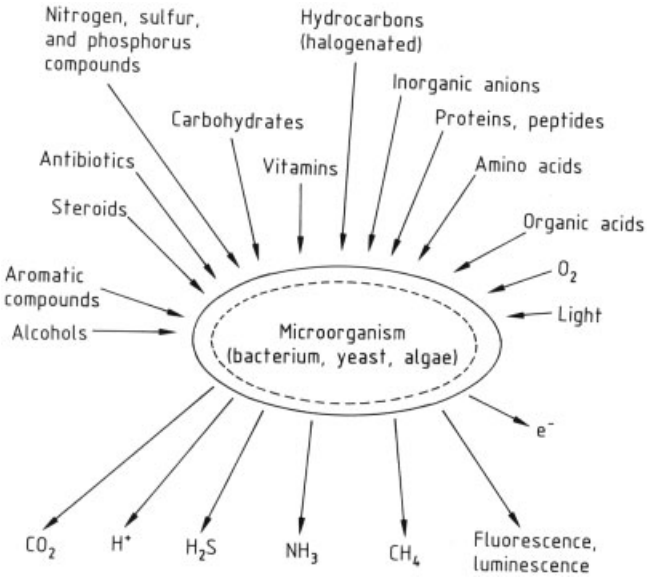

Figure 87. Microorganisms as receptor systems in biosensors useful for detecting specific groups of compounds

cially if they are grown on a single source of carbon or nitrogen.

Karube [389] has pioneered in the development of many bacterial electrodes based on this principle using various bacteria depending on the target analyte. Ion-selective electrodes for NH_3, O_2, CO_2, H_2S and H^+ have all been used in conjunction with immobilized whole cells. Riedel et al. [390] have demonstrated that preincubation of certain bacterial electrodes with the desired analyte (substrate) can enhance the sensitivity of a sensor toward that chemical by a factor of as much as 25. This induction approach may prove to be widely applicable. The shelf-life of a whole-cell electrochemical sensor can extend up to several weeks with fully optimized storage conditions (low temperature, for example). Microbe thermistors (sensors that respond to the heat evolved during bacterial metabolism of a substrate) have also been developed, but these present problems once again with respect to analyte specificity.

A significant advance in whole-cell biosensor technology accompanied the development by Rawson et al. [372] of a mediated electrochemical sensor employing a layer of blue-green algae (cyanobacteria). These authors demonstrated that a low molecular mass redox-active mediator, such as $Fe(CN)_6^{3-}$ or benzoquinone, could accept electrons from the photosynthetic energy chain of a cyanobacterium and relay them, via diffusion, to an electrode. The sensor proved capable of detecting aqueous herbicides on the basis of their inhibition of electron transport in the cells. This resulted in a dramatic fall in current, since fewer electrons reached the mediator and hence the electrode. The advantage of this approach to the use of whole cells is that it permits reasonably specific detection of a group of chemicals (herbicides, in this case) rather than limited-range detection of a single substance.

The principal application at the present time for cell sensors is in the measurement of biochemical oxygen demand (BOD) in wastewater. For this purpose yeast and/or bacterial cells (mainly the yeasts *Trichosporon* sp. or *Issatchenkia* and the bacterium *Rhodococcus erythropolis* [370], [386]) are immobilized on top of a Clark oxygen electrode and the oxygen consumption is measured. The great advantage of BOD sensors is a short measurement time (20 s to several minutes) compared with the classical BOD test in wastewater treatment, which requires five days. The main disadvantage is that a BOD sensor reflects only low molecular mass and easily assimilable substrates. High molecular mass substrates such as proteins and starch cannot penetrate the protective dialysis mem-

brane of the sensor. Pretreatment with hydrolytic enzymes could solve this problem. In general, however, BOD sensors, represent a promising approach to the control of wastewater facilities.

Table 21. Potential fields of application for biosensors

Field of application	Biosensor opportunities
Medicine	patient self-control, home health care (glucose, lactate, creatinine, phenylalanine, histamine)
	in vivo analysis
	long-term in vivo control of metabolites and drugs
	control element (biotic sensor) for artifical prosthesis and organs
	rapid analysis at intensive-care units
	surface imaging of organs during implantation
	bedside monitoring
Clinical chemistry	diagnostics for metabolites, drugs, enzymes, vitamins, hormones, allergies, infectious diseases, cancer markers, pregnancy, etc.
	laboratory safety
Environmental protection	pollution control
	monitoring/screening of toxic compounds in water supplies, solid and liquid wastes, soil and air (e.g., pesticides, inorganic ions, explosives, oils, PAHs, PCBs, microorganisms, volatile vapors and gases)
	self control of industrial companies and farms
	alarm systems for signaling hazardous conditions
	determination of organic load (BOD)
Chemical, pharmaceutical, food industry	monitoring and control of fermentation processes and cell cultivation (substrates, metabolites, products)
	food quality control (screening/detection of microbial contaminations; estimation of freshness, shelf life; olfactory qualities and flavor; rancidity; analysis of fats, proteins, carbohydrates in food)
	study of efficiency of drugs
	detection of leakage and hazardous concentrations of liquids and gases in buildings and mine shafts
	indoor air quality checks
	location of oil deposits
Agriculture	quality control of soils
	estimation of degradation/rottage (e.g., of biodegradable waste, or in wood or plant storage)
	rapid determination of quality parameters of milk
Military	detection of chemical and biological warfare agents (e.g., nerve gases, pathogenic bacteria, viruses)

3.5. Problems and Future Prospects

Compared with traditional sensors and classical methods of analysis, biosensors offer an attractive alternative for the rapid, selective, and highly sensitive assay of a wide variety of analytes. Further advantages of this cheap and economical tool emerge with the possibility of direct measurements within complex matrices, demanding only a rudimentary sample pretreatment, and the prospect of an uncomplicated measuring procedure that avoids the use of toxic reagents and organic solvents. Potential fields of application for biosensors are suggested in Table 21.

Until recently, biosensors were available for only a few analytes (roughly two dozen). This is mainly a consequence of insufficient reliability associated with poor stability of the biomaterial, including a dependence on physical and chemical parameters in the environment and interferences within the transducers. In the case of implantable biosensors for in vivo monitoring of metabolites or feedback control of artificial organs (e.g., the pancreas), a lack of biocompatibility with the body environment, organic interferences, and inadequate stability of the biological receptor have so far prevented long-term implantation. Nevertheless, with respect to rapid, sensitive, and selective low- cost determination of a great variety of relevant compounds no suitable alternative to the principle of biosensors is apparent at this time.

Constructing biosensors with greater reliability requires a better understanding of the basic principles of enzyme catalysis and immunochemical reactions, protein structures, pathways for receptor-based signal amplification, and the interfacial behavior of biocompounds at the artificial transducer surface. Biosensor research must therefore be directed toward the following points:

1) Protein engineering, with the aim of enhancing the stability of enzymes, catalytic antibodies, and bifunctional enzymes
2) Screening stable enzymes from microorganisms (extremophiles)
3) Utilization of antibody fragments (Fab, Fab′), bispecific antibodies, and single-chain and single-domain antibodies (tailored through genetic engineering and its expression in bacteria)
4) Recombinant antibodies
5) Modification of the efficiency, specificity, and modulation of antibodies by site-directed mutagenesis
6) Use of DNA libraries for the production of monoclonal antibodies to meet specific demands of the desired transducer

7) Selection and specification of biological receptor systems applicable to the quantification of olfaction and taste
8) Manipulating enzymes and cofactors by electrically wiring redox-active centers to transducer surfaces, utilizing the principles of both molecular electronics and biological electron transfer
9) Investigation of interfacial effects and interactions between biocompounds and transducer surfaces at the molecular and atomic levels by exploiting new methods of surface analysis (e.g., scanning microscopy techniques, fast-atomic-bombardment mass spectrometry, laser-assisted mass spectrometry, time-of-flight secondary ion mass spectrometry, Fourier-transform infrared spectrometry, ellipsometry, X-ray photoelectron spectroscopy, and electron microscopy)
10) Materials research, including biocompatible membranes (based on compounds found within the body) and redox-active electrically conducting (metal)organics and conjugated polymers
11) Development of (multifunctional) silicon-based ultramicroelectrodes covalently linked to biomaterials
12) Preventing the fouling of the biological compounds through microbiological degradation and attacks from proteases

Concerning methodological applications of biosensors, most developments relate to enzyme electrodes integrated into laboratory analyzers for the determination of low molecular mass compounds in the medical and biotechnological fields. A further important line of development is directed toward disposable biosensors for screening and semiquantitative analysis in environmental control, patient home diagnostics, and food-quality control. Direct testing would avoid the time- consuming procedures of sample storage and transfer as well as expensive laboratory analyses. A medium-term application is expected to be the use of biosensors in (quasi)continuously operating random-process analyzers based on micromachined fluidics elements, together with low-maintenance systems for the control and monitoring of water supplies and clinical bedside monitoring. Finally, a long-term target is the development of highly reliable implantable biosensors for long-term implantation. Biosensor applications in the field of instrumentation should also contribute to the solution of many important problems in conventional measuring techniques, opening major new areas to modern analysis.

4. Actuators and Instrumentation

The instrumentation associated with chemical sensors is rather simple and inexpensive, especially in the case of electrochemical transducers and the new IOS technology developed at the ICB Muenster . All that is normally required is a high-ohmic voltmeter (pH meter), a conductometer, and/or polarographic equipment. Fiber-optical devices, on the other hand, can be quite expensive depending on quality requirements with respect to the monochromatic light source. Spectrophotometers with fiber-optical inputs and outputs are already commercially available. A laser source is required in certain cases of sensitive fluorescence measurements with small-volume samples. Exploiting the modern techniques of acoustic wave modulation also requires more than a simple frequency counter. For example, a network analyzer can be of great help in following the shift in a resonance curve caused by interaction of an analyte with the transducer surface.

The highest costs are usually related to development of a selective analyte recognition layer, perhaps the most important part of a complete chemical sensor. Transducer technology is much more highly developed; indeed, there seems to be little need here for further improvement without equivalent improvements in the selectivity of sensor elements.

An important area of sensor research is microsystem technology (mst). A microsystem is defined as one too small to be fabricated by traditional miniaturized mechanical techniques, especially one that is dependent on modern mass-production techniques involving silicon or LIGA technology. A *system* must, by definition, consist of at least two interactive processes. A typical example is the antiblocking brake system (ABS) in an automobile, in which acceleration sensors are capable of detecting a blocked wheel and selectively controlling the brake power to that wheel via actuators.

Examples of analogous complete systems in chemistry are still rare. The best example is perhaps the "artificial pancreas," in which a glucose sensor measures the actual blood-glucose concentrations in order to control an appropriate actuator, the insulin pump. Bedside devices for this purpose have operated successfully for many years, but all attempts at implantation have so far failed, mainly because of the limited lifetime of the glucose biosensor and biofouling of the biosensor membrane by macrophages and other body constituents.

A complete system providing both a sensor and an actuator would be ideal in the field of process control, but because of a lack of truly reliable chemical sensors on the market the concept has not been widely implemented. One exception relates to the analytical method of coulometry, a technique that offers great potential for delivering chemical compounds to a controlled reaction. Especially attractive in this context is the method of constant- current coulometry, which can be carried out with an end-point sensor and a coulometric actuator for maintaining a generator current until the end-point has been reached. In this case both of the required devices can be miniaturized and constructed with the same technology.

5. Future Trends and Outlook

Surveys that have been conducted on the subject of chemical sensors all predict a growing market during the late 1990s and the early years of the 21st century [391–393]. Anticipated developments in automation, environmental protection, health care, and high technology ensure a strong demand for sensors. However, devices of this type must become increasingly reliable, must survive a thorough validation, and become even simpler to use. Japanese researchers see the first area for mass application of sensors to be the health care industry, especially with respect to the elderly. The demographic prospect of an increasing number of retirees and a lack of sufficient care centers may trigger the use of many sensors for monitoring the health of people living for as long as possible in their own homes. It is now clear that the application area of health care will dominate the application of biosensors. The "point of care" idea will replace large central laboratories to which the labile biological sample has to be transported. The reliabilty, built-in validation checks, and ease of use will allow their use by the normal physician who will have then the result within a few minutes. While the numbers of DNA arrays used in the genom research area will become saturated shortly, the use of simple arrays for looking for specific protein expression patterns will increase. Combining bio-informatics with the latter in finding super-marker molecules for extremely early tumor diagnostics will be possible. This leads to the corresponding mass-production of these diagnostic chips.

With respect to environmental chemistry, controls over the emission of organic compounds will undoubtedly be strengthened in the future. For example, municipal incinerators without provisions for the effective measurement of all hazardous organic compounds will soon be strictly forbidden everywhere. It seems likely that a large fraction of the moneys previously spent for defense purposes will be redirected toward improvements in continuous measurement technology. Sensors will be required to compete in this respect with such fast separation methods as capillary electrophoresis or flash chromatography, which permit the separation of most compounds within only a few minutes.

One critical remark is in order regarding papers published in the field of chemical sensors. Many such reports reflect a lack of input from qualified analytical chemists during the development and testing stages. Insufficient understanding is often apparent with respect to the analytical process itself and the difficulties involved in obtaining a truly reliable result; indeed, many sensor researchers seem to be unaware of such basic problems as sample – matrix effects. Anyone who claims to have developed a new sensor should be required to demonstrate its effectiveness with a real sample under real conditions. The latter is required and called validation by international quality standards (ISO 25). Publications that fail to reveal this most important feature of a sensor are also essentially worthless. Chemical sensors are analytical devices, so they should be subject to the same standards of characterization as other instrumental tools. Quality-control considerations mandate critical evaluation of any new molecular sensor presented to the scientific community. Every report concern-

ing a new sensor development should clearly delineate the crucial features, with none omitted.

The most important parameters for characterizing or validating a chemical sensor or a biosensor include:

1) *Selectivity.* The selectivity with respect to important known interferents for the particular application in question should always be expressed numerically; expressions like "relative selectivity" toward the analyte, or "not affected by typical concentrations of interfering compounds" are inappropriate.
2) *Sensitivity and Limit of Detection (LOD).* The slope of a calibration curve expressed in the appropriate units (e.g. microamps, millivolts, or hertz) per millimole or micromole should always be provided both for a matrix-free solution and for a sample in its matrix. The corresponding LOD should also be reported in both cases. In the absence of a blank the signal 3 × larger than the noise determines the LOD. The limit of quantification is typical 10 × LOD.
3) *Stability of the Analytical Function (Drift).* A zero-point drift value should be provided for a sensor in a matrix-free solution and also with respect to real samples involving a matrix. The same applies to the stability of the slope (varying sensor sensitivity).
4) *Response Time.* Since response time is also a function of the sample matrix it should be measured both in pure calibration solutions and with real samples. Generally two different response times are observed, one for a concentration rise and one for decreasing analyte concentration, with the latter usually slightly greater. Since IUPAC has not yet recommended a specific procedure for measuring this parameter, one should express the response time in the form of a time constant (=1/e of the final value), since this is the approach taken with electronic circuits. Only in the case of a sensor that never reaches its final value should other approaches be considered for characterizing the rate of response to a changing analyte concentration.
5) *Sensor Lifetime.* Especially in the field of biosensors the lifetime of a sensor in contact with a calibration solution or a real sample, or during storage in a refrigerator with an optimal conditioning solution, is of considerable interest, and it should always be mentioned. Conditions leading to rapid sensor fouling should also be identified in order to permit appropriate precautions to be taken. Examples include the fouling of certain enzymes by inhibitors or heavy metals, or microbiological attack. Any use of conditioning solutions containing metal complexation reagents, inhibitor scavengers, sodium azide, etc., should be clearly described.

Other important information might include total power consumption, auxiliary equipment costs, total elapsed time between failures, temperature compensation measures, or maintenance requirements.

Finally, sensor developers must remain aware of progress within instrumental analytical chemistry generally, especially in the area of fast high-resolution chromatography. Indeed, progress has already been so great in the latter field that more than 20 anions (including phosphate) can now be separated completely in less time than would be required for a response from a multienzyme phosphate biosensor! The serious challenge this poses to sensor advocates will be obvious.

6. References

1. W. Göpel, J. Hesse, J. N. Zemel (eds.): *Sensors – A Comprehensive Survey,* **vols. 1 – 8 + updates,** VCH Verlagsgesellschaft mbH, Weinheim 1989 – 1995.
2. T. Seiyama, S. Yamauchi (eds.): *Chemical Sensor Technology,* **vols. 1 + 4,** Elsevier Science Publisher B, Amsterdam 1988.
3. R. W. Murray et al.: "Chemical Sensors and Microinstrumentation," *ACS Symp. Ser.* **403** (1989).
4. J. Janata: *Principles of Chemical Sensors,* Plenum Press, New York 1989.
5. K. Cammann, H. Galster: *Das Arbeiten mit ionenselektiven Elektroden,* 3rd ed., Springer Verlag, Berlin 1996.
6. J. Janata, R. Huber: *Solid State Chemical Sensors,* Academic Press, New York 1985.
7. K. Cammann et al.: *Analytiker Taschenbuch,* **vol. 15,** Springer-Verlag, Berlin – Heidelberg 1997, pp. 3 – 40.
8. J. Reinbold, K. Cammann: *Analytiker Taschenbuch,* **vol. 16,** Springer-Verlag, Berlin – Heidelberg 1997, pp. 3 – 42.

9. K. Cammann (ed.): *Instrumentelle Analytische Chemie,* Spektrum-Verlag, Heidelberg 2000.
10. T. Takeuchi, *Sens. Actuators* **14** (1988) 109.
11. G. Hötzel, H. M. Wiedemann, *Sens. Rep.* **4** (1989) 32.
12. J. P. Pohl, *GIT Fachz. Lab.* **5** (1987) 379.
13. K. Cammann et al., *Angew. Chem.* **103** (1991) 519; *Angew. Chem. Int. Ed. Engl.* **30** (1991) 516 – 539.
14. W. Masing: *Handbuch der Qualitätssicherung,* Hanser Verlag, München 1988.
15. K. Cammann: *Working with Ion-Selective Electrodes,* Springer Verlag, Berlin 1979.
16. A. K. Covington: *Ion Selective Electrode Methodology,* vols. I and II, CRC Press, Boca Raton, Fla., 1979.
17. H. Freiser (ed.): *Ion Selective Electrodes in Analytical Chemistry,* Plenum Press, New York 1979.
18. K. Schierbaum, *Sens. Actuators B* **18 – 19** (1994) 71 – 76.
19. J. Reinbold et al., *Sens. Actuators B* **18 – 19** (1994) 77 – 81.
20. N. Taguchi, JP 4 5 38 200, 1962.
21. T. Seiyama, A. Kato, K. Fujiishi, M. Nagatani, *Anal. Chem.* **34** (1962) 1502.
22. Y. Nakatani, M. Sakai, M. Matsuoka in T. Seyama et al. (eds.): *Proc. Int. Meet. Chem. Sensors (Fukuoka, Japan),* Elsevier, Amsterdam, Sept. 1983, pp. 147 – 152.
23. S. R. Morrison: "Semiconductor Gas Sensors," *Sens. Actuators* **2** (1982) 329.
24. P. J. Shaver, *Appl. Phys. Lett.* **11** (1967) 255.
25. J. C. Loh, JP 4 3 28 560; FR 1 545 292, 1967.
26. N. Yamazoe, Y. Krakowa, T. Seiyama: "Effects of Additives on Semiconductor Gas Sensors," *Sens. Actuators* **4** (1983) 283.
27. S. Matsushima, Y. Teraoka, N. Miura, N. Yamazoe, *Jpn. J. Appl. Phys.* **27** (1988) 1798.
28. M. Nitta, S. Kanefusa, S. Ohtani, M. Haradome, *J. Electr. Mater.* **13** (1984) 15.
29. A. Chiba: "Development of the TGS Gas Sensor," in S. Yamauchi (ed.): *Chemical Sensor Technology,* vol. 4, Elsevier, Amsterdam 1992, pp. 1 – 18.
30. T. A. Jones: "Characterisation of Semiconductor Gas Sensors," in P. T. Moseley, B. C. Tofield (eds.): *Solid State Gas Sensors,* Adam Hilger, Bristol 1987.
31. G. Heiland, D. Kohl, *Sens. Actuators* **8** (1985) 227.
32. W. Weppner: "Halbleiter-Sensoren" in R. Grabowski (ed.): *Sensoren und Aktoren,* VDE-Verlag, Berlin 1991, pp. 167 – 183.
33. D. E. Williams: "Characterization of Semiconductor Gas Sensors," in P. T. Moseley, B. C. Tofield (eds.): *Solid State Gas Sensors,* Adam Hilger, Bristol 1987.
34. N. Yamazoe, N. Miura: "Some Basic Aspects of Semiconductor Gas Sensors," in S. Yamauchi (ed.): *Chemical Sensor Technology,* vol. 4, Elsevier, Amsterdam 1992, pp. 19 – 41.
35. G. Heiland: "Homogeneous Semiconducting Gas Sensors," *Sens. Actuators* **2** (1982) 343.
36. H. Windischmann, P. Mark, *J. Electrochem. Soc.* **126** (1979) 627.
37. S. Chang, *J. Vac. Sci. Technol.* **17** (1980) 366.
38. J. Watson: "The Tin Dioxide Gas Sensor and Its Applications," *Sens. Actuators* **5** (1984) 29.
39. N. Komori, S. Sakai, K. Konatsu in T. Seyama et al. (eds.): *Proc. Int. Meet. Chem. Sensors (Fukuoka, Japan),* Elsevier, Amsterdam, Sept. 1983, p. 57.
40. B. Ahlers: *Dissertation*, University of Muenster 1996.
41. M. L. Hitchmann, H. A. O. Hill, *Chem. Br.* **22** (1986) 1177.
42. D. Ammann: *Ion Selective Microelectrodes – Principles, Design and Application,* Springer Verlag, Berlin 1986.
43. P. L. Bailey: *Analysis with Ion-Selective Electrodes,* Heyden & Sons, London 1976.
44. R. Bock: "Nachweis- und Bestimmungsmethoden," in *Methoden der Analytischen Chemie,* vol. 2, part 2, Verlag Chemie, Weinheim 1984, pp. 33 – 83.
45. K. Cammann: "Fehlerquellen bei Messungen mit ionenselektiven Elektroden," in *Analytiker-Taschenbuch,* vol. 1, Springer Verlag, Berlin 1980.
46. R. A. Durst: "Ion-Selective Electrodes," *NBS Special Publication No. 314,* 1969.
47. S. Ebel, W. Parzefall: *Experimentelle Einführung in die Potentiometrie,* Verlag Chemie, Weinheim 1975.
48. E. Bakker, *Trends Anal. Chem.* **16** (1997) 252 – 260.
49. G. Horvai, *Trends Anal. Chem.* **15** (1997) 260 – 266.
50. G. Eisenman: *Glass Electrodes for Hydrogen and Other Cations, Principles and Practice,* Marcel Dekker, New York 1967.
51. D. J. G. Ives, J. Janz: *Reference Electrodes,* Academic Press, New York 1961.
52. R. Koryta: *Ion Selective Electrodes,* Cambridge University Press, Cambridge 1975.
53. T. S. Ma, S. S. M. Hassan: *Organic Analysis Using Ion-Selective Electrodes,* 2 vols., Academic Press, New York 1982.

54. G. J. Moody, J. D. R. Thomas: *Selective Ion Electrodes,* Merrow Publishing, Watford 1977.
55. F. Oehme: *Ionenselektive Elektroden: CHEMFETs – ISFETs – pH-FETs, Grundlagen, Bauformen und Anwendungen,* Hüthig Verlag, Heidelberg 1991.
56. E. Pungor: *Ion-Selective Electrodes,* Akadémiai Kiadó, Budapest 1973.
57. F. Honold, B. Honold: *Ionenselective Elektroden,* Birkhäuser Verlag, Basel 1991.
58. K. Umezawa, Y. Umezawa: *Selectivity Coefficients for Ion Selective Electrodes,* University of Tokyo Press, Tokyo 1983.
59. E. Bakker, P. Bühlmann, E. Pretsch: "Carrier-Based Ion-Selective Electrodes and Bulk Optodes. 1. General Characteristics," *Chem. Rev.* **97** (1997) 3083 – 3132.
60. P. Bühlmann, E. Pretsch, E. Bakker: "Carrier-Based Ion-Selective Electrodes and Bulk Optodes. 2. Ionophores for Potentiometric and Optical Sensors," *Chem. Rev.* **98** (1998) 1593 – 1687.
61. Y. Mi, S. Mathison, R. Goines, A. Logue, A. Bakker, *Anal. Chim. Acta* **397** (1999) 103 – 111.
62. A. Schwake: *Dissertation*, University of Muenster 1999.
63. F. Zuther: *Dissertation*, University of Muenster 1997.
64. A. J. Bard, L. R. Faulkner: *Electrochemical Methods,* J. Wiley, New York 1980.
65. A. M. Bond, H. B. Greenhill, I. D. Heritage, J. B. Reust, *Anal. Chim. Acta* **165** (1984) 209.
66. A. J. Bard: *Electroanalytical Chemistry,* vol. 9, Marcel Dekker, New York 1976.
67. G. Henze, R. Neeb: *Elektrochemische Analytik,* Springer Verlag, Berlin 1985.
68. K. Cammann: "Elektrochemische Verfahren," in H. Naumer, W. Heller (eds.): *Untersuchungsmethoden in der Chemie,* Thieme Verlag, Stuttgart 1986.
69. P. T. Kissinger, W. R. Heinemann, *J. Chem. Educ.* **60** (1983) 702.
70. J. Osteryoung, J. J. O'Dea in A. J. Bard (ed.): *Electroanalytical Chemistry,* **vol. 14,** Marcel Dekker, New York 1986, pp. 209 – 308.
71. P. Vadgama, G. Davis, *Med. Lab. Sci.* **42** (1985) 333.
72. L. C. Clark, US 2 913 386, 1958.
73. H. Suzuki et al., *Sens. Actuators B* **2** (1990) 297.
74. S. C. Cha, M. J. Shao, C. C. Liu, *Sens. Actuators B* **2** (1990) 239.
75. W. Sansen et al., *Sens. Actuators B* **1** (1990) 298.
76. H. Suzuki et al., *Sens. Actuators B* **1** (1990) 528.
77. H. Suzuki et al., *Sens. Actuators B* **1** (1990) 275.
78. D. S. Austin et al. *J. Electroanal. Chem. Interfacial Electrochem.* **168** (1984) 227.
79. R. M. Wightman, *Anal. Chem.* **53** (1981) 1125.
80. T. Hepel, J. Osteryoung, *J. Electrochem. Soc.* **133** (1986) 752.
81. L. Zhaohui et al., *J. Electroanal. Chem. Interfacial Electrochem.* **259** (1989) 39.
82. C. L. Colyer, J. C. Myland, K. B. Oldham, *J. Electroanal. Chem. Interfacial Electrochem.* **263** (1989) 1.
83. J. Osteryoung, S. T. Singleton, *Anal. Chem.* **61** (1989) 1211.
84. R. E. Howard, E. Hu: *Science and Technology of Microfabrication,* Materials Research Society, Pittsburgh 1987.
85. C. D. Baer, N. J. Stone, D. A. Sweigart, *Anal. Chem.* **58** (1988) 78.
86. J. Ghoroghchiana et al., *Anal. Chem.* **58** (1988) 2278.
87. H. Meyer et al., *Anal. Chem.* **67** (1995) 1164 – 1170.
88. M. Wittkamp, G. Chemnitius, K. Cammann, M. Rospert, M. Mokwa, *Sens. Actuators B* **43** (1997) 87 – 93.
89. W. Göpel, J. Hesse, J. N. Zemel: *Sensors,* vol. 2, VCH Verlagsgesellschaft, Weinheim 1991.
90. W. Göpel, J. Hesse, J. N. Zemel: *Sensors,* vol. 3, VCH Verlagsgesellschaft, Weinheim 1991.
91. P. T. Moseley, B. C. Tofield: *Solid State Gas Sensors,* Adam Hilger, Bristol 1987.
92. P. T. Moseley, J. O. W. Norris, D. E. Williams: *Techniques and Mechanisms in Gas Sensing,* Adam Hilger, Bristol 1991.
93. H. Emons, G. Jokuszies, *Z. Chem.* **28** (1988) 197.
94. R. Kalvoda, *Electroanalysis* **2** (1990) 341.
95. F. Opekar, A. Trojanek, *Anal. Chim. Acta* **203** (1987) 1.
96. ISO/TC 147 (draft):Electrolytic Conductivity, Beuth Verlag, Berlin.
97. DIN 38 404:Deutsche Verfahren zur Wasser-, Abwasser- und Schlammuntersuchung, VCH Verlagsgesellschaft, Weinheim 1993.
98. F. Oehme in W. Göpel, J. Hesse, J. N. Zemel (eds.): *Sensors,* vol. 2, VCH Verlagsgesellschaft, Weinheim 1991.

99. F. Oehme, R. Bänninger: *ABC der Konduktometrie,* Polymetron AG, Mönchaltorf (Switzerland).
100. K. Rommel, *VGB Kraftwerkstech.* **65** (1985) 417.
101. K. Rommel: *Kleine Leitfähigkeitsfibel,* Wissenschaftlich-Technische Werke, Weilheim 1988.
102. F. Oehme, M. Jola, *Betriebsmeßtechnik,* Hüthig Verlag, Heidelberg 1982.
103. M. Niggemann: *Dissertation*, University of Muenster 1999.
104. W. H. Brattain, J. Bardeen, *Bell Syst. Tech. J.* **32** (1952) 1.
105. G. Heiland, *Z. Phys.* **138** (1954) 459.
106. J. Watson, *Sens. Actuators B* **8** (1992) 173.
107. N. Taguchi, US 3 644 795, 1972.
108. D.-D. Lee, D.-H. Choi, *Sens. Actuators B* **1** (1990) 231.
109. B. C. Tofield in P. T. Moseley, B. C. Tofield (eds.): *Solid State Gas Sensors,* Adam Hilger, Bristol 1987.
110. R. Lalauze, C. Pijolat, S. Vincent, L. Bruno, *Sens. Actuators B* **8** (1992) 237.
111. J. Watson, *Sens. Actuators* **5** (1984) 29.
112. R. M. Geatches, A. V. Chadwick, J. D. Wright, *Sens. Actuators B* **4** (1991) 467.
113. S. R. Morrison, *Sens. Actuators* **12** (1987) 425.
114. S. Matsushima et al., *Sens. Actuators B* **9** (1992) 71.
115. T. Maekawa et al., *Sens. Actuators B* **9** (1992) 63.
116. S. Nakata, H. Nakamura, K. Yoshikawa, *Sens. Actuators B* **8** (1992) 187.
117. U. Kirner et al., *Sens. Actuators B* **1** (1990) 103.
118. N. Li, T.-C. Tan, *Sens. Actuators B* **9** (1992) 91.
119. U. Lampe, M. Fleischer, H. Meixner, *Sens. Actuators B* **17** (1994) 187 – 196.
120. N. Koshizaki, K. Yasumoto, K. Suga, *Sens. Actuators B* **9** (1992) 17.
121. M. Egashira, Y. Shimizu, Y. Takao, *Sens. Actuators B* **1** (1990) 108.
122. P. T. Moseley, D. E. Williams, *Sens. Actuators B* **1** (1990) 113.
123. E. Ye. Gutman, I. A. Myasnikov, *Sens. Actuators B* **1** (1990) 210.
124. R. S. Tieman, W. R. Heineman, J. Johnson, R. Seguin, *Sens. Actuators B* **8** (1992) 199.
125. Y. Sadaoka, T. A. Jones, W. Göpel, *Sens. Actuators B* **1** (1990) 148.
126. S. Dogo, J. P. Germain, C. Maleysson, A. Pauly, *Sens. Actuators B* **8** (1992) 257.
127. J. W. Gardner, M. Z. Iskandarani, B. Bott, *Sens. Actuators B* **9** (1992) 133.
128. M. Trometer et al., *Sens. Actuators B* **8** (1992) 129.
129. J. D. Wright, *Prog. Surf. Sci.* **31** (1989) 1.
130. T. A. Temofonte, K. F. Schoch, *J. Appl. Phys.* **65** (1989) 1350.
131. S. Kanefusa, M. Nitta, *Sens. Actuators B* **9** (1992) 85.
132. W. Göpel, K. D. Schierbaum, D. Schmeisser, D. Wiemhöfer, *Sens. Actuators* **17** (1989) 377.
133. B. Bott, S. C. Thorpe in P. T. Moseley, J. O. W. Norris, D. E. Williams (eds.): *Techniques and Mechanisms in Gas Sensing,* Adam Hilger, Bristol 1991.
134. T. Hanawa, S. Kuwabata, H. Yoneyama, *J. Chem. Soc. Faraday Trans. 1* **84** (1988) 1587.
135. C. Nylander, M. Armgarth, I. Lundström, *Anal. Chem. Symp. Ser.* **17** (1983) 203.
136. P. N. Bartlett, P. B. M. Archer, S. K. Ling-Chung, *Sens. Actuators* **19** (1989) 125.
137. P. N. Bartlett, P. B. M. Archer, S. K. Ling-Chung, *Sens. Actuators* **19** (1989) 141.
138. H. Arai, T. Seiyama: "Humidity Control," in W. Gö-pel, J. Hesse, J. N. Zemel (eds.): *Sensors,* vol. 2, VCH Verlagsgesellschaft, Weinheim 1991.
139. A. K. Michell in P. T. Moseley, J. O. W. Norris, D. E. Williams (eds.): *Techniques and Mechanisms in Gas Sensing,* Adam Hilger, Bristol 1991.
140. T. Nenov, S. Yordanov, *Sens. Actuators B* **8** (1992) 117.
141. Y. Sakai, Y. Sadaoka, H. Fukumoto, *Sens. Actuators* **13** (1988) 243.
142. P. Bergveld, *IEEE Trans. Biomed. Eng.* **BME-17** (1970) 70.
143. K. Buhlmann: *Dissertation*, University of Muenster 1997.
144. K. Buhlmann, B. Schlatt, K. Cammann, A. Shulga, *Sens. Actuators B* **49** (1998) 156 – 165.
145. T. Matsuo, M. Esashi, K. Inuma, *Proc. Dig. Joint Meeting,* Tohoku Sect. IEEEJ, Oct. 1971.
146. J. E. Moneyron, A. De Roy, J. P. Besse, *Solid State Ionics* **46** (1991) 175.
147. S. M. Sze: *Physics of Semiconductor Devices,* Wiley, New York 1981.
148. P. Bergveld, A. Sibbald: *Analytical and Biomedical Applications of Ion Sensitive Field Effect Transistors,* Elsevier, Amsterdam 1988.
149. Y. G. Vlasov, D. E. Hackleman, R. P. Buck, *Anal. Chem.* **51** (1979) 1579.
150. W. Moritz, I. Meierhöffer, L. Müller, *Sens. Actuators* **15** (1988) 211.

151. S. D. Moss, J. Janata, C. C. Johnson, *Anal. Chem.* **47** (1975) 2238.
152. A. Sibbald, P. D. Whalley, A. K. Covington, *Anal. Chim. Acta* **159** (1984) 47.
153. D. J. Harrison et al., *J. Electrochem. Soc.* **135** (1988) 2473.
154. P. D. van der Wal et al., *Anal. Chim. Acta* **231** (1990) 41.
155. A. van den Berg, A. Griesel, E. Verney-Norberg, *Sens. Actuators B* **4** (1991) 235.
156. R. W. Catrall, P. I. Iles, J. C. Hamilton, *Anal. Chim. Acta* **169** (1985) 403.
157. R. Smith, D. C. Scott, *IEEE Trans. Biomed. Eng.* **BME 33** (1986) 83.
158. A. van den Berg, P. Bergveld, D. N. Reinhoudt, E. J. R. Sudhölter, *Sens. Actuators* **8** (1985) 129.
159. N. J. Ho, J. Kratochvil, G. F. Blackburn, J. Janata, *Sens. Actuators* **4** (1983) 413.
160. C. Dumschat et al., *Sens. Actuators B* **2** (1990) 271.
161. K. Domansky, J. Janata, M. Josowicz, D. Petelenz, *Analyst (London)* **118** (1993) 335.
162. H. H. van den Vlekkert et al., *Sens. Actuators* **14** (1988) 165.
163. C. Dumschat et al., *Anal. Chim. Acta* **243** (1991) 179.
164. P. Bergveld, *Sens. Actuators B* **4** (1991) 125.
165. W. F. Love, L. J. Button, R. E. Slovacek: "Optical Characteristics of Fiberoptic Evanescent Wave Sensors," in D. L. Wise, L. B. Wingard (eds.): *Biosensors with Fiberoptics,* Humana Press, Clifton, N.J., 1991.
166. S. J. Lackie, T. R. Glass, M. J. Block: "Instrumentation for Cylindrical Waveguide Evanescent Fluorosensors," in D. L. Wise, L. B. Wingard (eds.): *Biosensors with Fiberoptics,* Humana Press, Clifton, N.J., 1991.
167. I. M. Walczak, W. F. Love, T. A. Cook, R. E. Slovacek: "The Application of Evanescent Wave Sensing to a High-sensitivity Fluoroimmunoassay," *Biosensors Bioelectron.* **7** (1992) 39 – 47.
168. B. J. Tromberg, M. J. Sepaniak, T. Vo-Dinh, G. D. Griffin: "Fiberoptic Probe for Competitive Binding Fluoroimmunoassay," *Anal. Chem.* **59** (1987) 1226 –1230.
169. R. B. Thompson, F. S. Ligler: "Chemistry and Technology of Evanescent Wave Biosensors," in D. L. Wise, L. B. Wingard (eds.): *Biosensors with Fiberoptics,* Humana Press, Clifton, N.J., 1991.
170. W. Trettnak, P. J. M. Leiner, O. S. Wolfbeis: "Fiberoptic Glucose Biosensor with an Oxygen Optrode as the Transducer," *Analyst (London)* **113** (1988) 1519 –1523.
171. I. J. Peterson, S. R. Goldstein, R. V. Fitzgerald, D. K. Buckhold: "Fiberoptic pH Probe for Physiological Use," *Anal. Chem.* **52** (1980) 864 – 869.
172. G. F. Kirkbright, R. Narayanaswamy, N. A. Welti: "Studies with Immobilized Chemical Reagents Using a Flow- cell for the Development of Chemically Sensitive Fiberoptic Devices," *Analyst (London)* **109** (1984) 12 – 19.
173. L. A. Saari, W. R. Seitz: "pH Sensor Based on Immobilized Fluoresceinamine," *Anal. Chem.* **54** (1982) 821 – 823.
174. Z. Zhujun, W. R. Seitz: "A Fluorescence Sensor for Quantifying pH in the Range from 6.5 to 8.5," *Anal. Chim. Acta* **160** (1984) 47 – 55.
175. E. Urbano, M. Offenbacher, O. S. Wolfbeis: "Optical Sensor for Continuous Determination of Halides," *Anal. Chem.* **56** (1984) 427 – 430.
176. O. S. Wolfbeis, E. Urbano: "Fluorescence Quenching Method for Determination of Two or Three Components in Solution," *Anal. Chem.* **55** (1983) 1904 – 1907.
177. L. A. Saari, W. R. Seitz: "Immobilized Morin as Fluorescense Sensor for Determination of Aluminum(III)," *Anal. Chem.* **55** (1983) 667 – 670.
178. L. A. Saari, W. R. Seitz: "Optical Sensor for Beryllium Based on Immobilized Morin Fluorescence," *Analyst* (London) 109 (1984) 655 – 657.
179. Z. Zhujun, W. R. Seitz: "A Fluorescent Sensor for Aluminum(III), Magnesium(II), Zinc(II) and Cadmium(II) Based on Electrostatically Immobilized Quinolin-8-ol-sulfonate," *Anal. Chim. Acta* **171** (1985) 251 – 258.
180. Z. Zhujun, W. R. Seitz: "A Carbon Dioxide Sensor Based on Fluorescence," *Anal. Chim. Acta* **170** (1984) 209 – 216.
181. M. A. Arnold, T. J. Ostler: "Fiberoptic Ammonia Gas Sensing Probe," *Anal. Chem.* **58** (1986) 1137 –1140.
182. A. Sharma, O. S. Wolfbeis: "Fiberoptic Fluorosensor for Sulfur Dioxide Based on Energy Transfer and Exciplex Quenching," *Proc. SPIE Int. Soc. Opt. Eng.* **990** (1989) 116 – 120.
183. A. Sharma, O. S. Wolfbeis: "Fiberoptic Oxygen Sensor Based on Fluorescence

Quenching and Energy Transfer," *Appl. Spectrosc.* **42** (1988) 1009 – 1011.
184. O. S. Wolfbeis, M. J. P. Leiner, H. E. Posch: "A New Sensing Material for Optical Oxygen Measurement, with Indicator Embedded in an Aqueous Phase," *Mikrochim. Acta* 1986, no. 3, 359 – 366.
185. K. W. Berndt, J. R. Lakowiz: "Electroluminescent Lamp-Based Phase Fluorometer and Oxygen Sensor," *Anal. Biochem.* **201** (1992) 319 – 325.
186. "Integrated Optics" in T. Tamir (ed.): *Topics in Applied Physics,* Springer Verlag, Berlin 1975.
187. K. Tiefenthaler, W. Lukosz: "Sensitivity of Grating Couplers as Integrated-optical Chemical Sensors," *J. Opt. Soc. Am. B Opt. Phys.* **B 6** (1975) 209 – 220.
188. A. Brandenburg, R. Edelhauser, F. Hutter: "Integrated Optical Gas Sensors Using Organically Modified Silicates as Sensitive Films," *Sens. Actuators B* **11** (1993) 361 – 374.
189. D. Schlatter et al.: "The Difference Interferometer: Application as a Direct Immunosensor," in: *The Second World Congress on Biosensors,* Oxford, Elsevier Advanced Technology, 1992, pp. 347 – 355.
190. Ph. Nellen, K. Tiefenthaler, W. Lukosz: "Integrated Optical Input Grating Couplers As Biochemical Sensors," *Sens. Actuators B* **15** (1988) 285 – 295.
191. R. G. Heidemann, R. P. H. Kooyman, J. Greve: "Performance of a Highly Sensitive Optical Waveguide Mach – Zehnder Interferometer Immunosensor," *Sens. Actuators B* **10** (1993) 209 – 217.
192. N. Fabricius, G. Gaulitz, J. Ingenhoff: "A Gas Sensor Based on an Integrated Optical Mach – Zehnder Interferometer," *Sens. Actuators B* **7** (1992) 672 – 676.
193. Ch. Fattinger, H. Koller, P. Wehrli, W. Lukosz: "The Difference Interferometer: A Highly Sensitive Optical Probe for Molecular Surface- Coverage Detection," in: *The Second World Congress on Biosensors,* Oxford, Elsevier Advanced Technology, 1992, pp. 339 – 346.
194. K. Tiefenthaler: "Integrated Optical Couplers As Chemical Waveguide Sensors," *Adv. Biosensors* **2** (1992) 261 – 289.
195. B. Liedberg, C. Nylander, I. Lundström: "Surface Plasmon Resonance for Gas Detection and Biosensing," *Sens. Actuators* **4** (1983) 299 – 304.
196. P. B. Daniels, J. K. Deacon, M. J. Eddowes, D. Pedley: "Surface Plasmon Resonance Applied to Immunosensing," *Sens. Actuators* **14** (1988) 11 – 17.
197. J. Gent et al.: "Optimization of a Chemooptical Surface Plasmon Resonance Based Sensor," *Appl. Opt.* **29** (1990) 2343 – 2849.
198. K. Matsubaru, S. Kawata, S. Minami: "Optical Chemical Sensor Based on Surface Plasmon Measurement," *Appl. Opt.* **27** (1988) 1160 – 1163.
199. H. Rather: *Surface Plasmons,* Springer Verlag, Berlin 1988.
200. G. Gauglitz et al.: DE 4 200 088 A1, 1993.
201. A. Brecht, J. Ingenhoff, G. Gauglitz, *Sens. Actuators B* **6** (1992) 96 – 100.
202. G. Kraus, G. Gauglitz, *Fresenius' J. Anal. Chem.* **349** (1994) no. 5, 399 – 402.
203. G. Sauerbrey, *Z. Phys.* **155** (1959) 206 – 222.
204. H. K. Pulker, J. P. Decostered in C. Lu, A. W. Czanderna (eds.): "Applications of Piezoelectric Quartz Crystal Microbalances," *Methods and Phenomena,* vol. 7, Elsevier, Amsterdam 1984, Chap. 3.
205. Lord Rayleigh, *Proc. London Math. Soc.* **17** (1885) 4 – 11.
206. G. Sauerbrey, *Z. Phys.* **178** (1964) 457.
207. W. H. King, Jr., *Anal. Chem.* **36** (1964) 1735 – 1739.
208. W. H. King, Jr., US 3 164 004, 1965.
209. A. W. Czanderna, C. Lu: "Applications of Piezoelectric Quartz Crystal Microbalances," in C. Lu, A. W. Czanderna (eds.): *Methods and Phenomena,* vol. 7, Elsevier, Amsterdam 1984. chap. I.
210. M. Thompson et al. *Analyst (London)* **116** (1991) 881 – 890.
211. L. Wimmer, S. Hertl, J. Hemetsberger, E. Benes, *Rev. Sci. Instrum.* **55** (1984) 605 – 609.
212. C.-S. Lu, O. Lewis, *J. Appl. Phys.* **43** (1972) 4385 –4390.
213. J. G. Miller, D. I. Bolef, *J. Appl. Phys.* **39** (1968) 5815 – 5816.
214. E. Benes, *J. Appl. Phys.* **56** (1984) 608 – 626.
215. D. R. Denison, *J. Vac. Sci. Technol.* **10** (1973) 126 –129.
216. C.-S. Lu, *J. Vac. Sci. Technol.* **12** (1975) 578 – 583.
217. G. G. Guilbault: "Applications of Piezoelectric Quartz Crystal Microbalances," in C. Lu, A. W. Czanderna (eds.): *Methods and Phenomena,* vol. 7, Elsevier, Amsterdam 1984, chap. 8.
218. H. Beitnes, K. Schrøder, *Anal. Chim. Acta* **158** (1984) 57 – 65.
219. A. Kindlund, H. Sundgren, I. Lundström, *Sens. Actuators* **6** (1984) 1 – 17.

220. G. G. Guilbault: *Ion-Selective Electrode Review,* **vol. 2,** Pergamon Press, Oxford 1980, pp. 3 – 16.
221. R. M. Langdon: "Resonator Sensors – a Review," *J. Phys. E* **18** (1985) 103 – 115.
222. J. Tichy, G. Gautschi: *Piezoelektrische Meßtechnik,* Springer Verlag, Berlin 1980.
223. H. Wohltjen, R. Dessy, *Anal. Chem.* **51** (1979) 1458 –1464.
224. H. Wohltjen, R. Dessy, *Anal. Chem.* **51** (1979) 1465 –1470.
225. H. Wohltjen, R. Dessy, *Anal. Chem.* **51** (1979) 1470 –1478.
226. C. T. Chang, R. M. White: "Excitation and Propagation of Plate Mode in an Acoustically Thin Membrane," *Proc. IEEE Ultrasonics Symp.* 1982, 295 – 298.
227. R. M. White, P. W. Wicher, S. W. Wenzel, E. T. Zellers: "Plate-Mode Ultrasonic Oscillator Sensors," *IEEE Trans. Ultrason. Dev. Ferroelectr. Freq. Contr.* **UFFC-34** (1987) 162 – 171.
228. T. G. Giesler, J.-U. Meyer: *Fachbeilage Mikroperipherik,* vol. 6 (1992) XLVI.
229. IEEE Standard on Piezoelectricity (ANSI/IEEE Std. 176-1987), The Institute of Electrical and Electronics Engineers, New York 1987.
230. M. S. Nieuwenhuizen, A. Venema: *Sensors* – A Comprehensive Survey, vol. II, VCH Verlagsgesellschaft, Weinheim 1991, p. 652.
231. J. Janata: *Principles of Chemical Sensors,* Plenum Press, New York 1989, pp. 55 – 80.
232. J. G. Miller, D. I. Bolef, *J. Appl. Phys.* **39** (1968) 45.
233. K. H. Behrndt, *J. Vac. Sci. Technol.* **8** (1971) 622.
234. A. P. M. Glassford in A. M. Smith (ed.): *Progress in Astronautics and Aeronautics,* **vol. 56,** American Institute of Aeronautics and Astronautics, New York 1977, p. 175.
235. A. P. M. Glassford, *J. Vac. Sci. Technol.* **15** (1978) 1836.
236. V. Mecea, R. V. Bucur, *Thin Solid Films* **60** (1979) 73.
237. R. A. Crane, G. Fischer, *J. Phys. D* **12** (1979) 2019.
238. K. K. Kanazawa, J. G. Gordon, *Anal. Chem.* **57** (1985) 1770.
239. C. Fruböse, K. Doblhofer, C. Soares, *Ber. Bunsen-Ges. Phys. Chem.* **97** (1993) 475 – 478.
240. R. Beck, U. Pittermann, K. G. Weil, *Ber. Bunsen-Ges. Phys. Chem.* **92** (1988) 1363 – 1368.
241. F. Eggers, Th. Funck, *J. Phys. E* **20** (1987) 523 – 530.
242. A. Glidle, R. Hillman, S. Bruckenstein, *J. Electroanal. Chem. Interfacial Electrochem.* **318** (1991) 411 – 420.
243. R. Beck, U. Pittermann, K. G. Weil, *J. Electrochem. Soc.* **139** (1992) no. 2, 453 – 461.
244. R. Beck, U. Pittermann, K. G. Weil, *Ber. Bunsen-Ges. Phys. Chem.* **92** (1988) 1363 – 1368.
245. D. Salt: *Handbook of Quartz Crystal Devices,* van Nostrand Reinhold, New York 1987.
246. Z. P. Khlebarov, A. I. Stoyanova, D. I. Topalova, *Sens. Actuators B* **8** (1992) 33 – 40.
247. V. M. Ristic: *Principles of Acoustic Devices,* J. Wiley and Sons, New York 1983.
248. P. Hauptmann, *Sens. Actuators A* **25** (1991) 371 – 377.
249. B. A. Auld: *Acoustic Fields and Waves in Solids,* vol. II, J. Wiley and Sons, New York 1973.
250. J. F. Vetelino, D. L. Lee, WO 83 1511, 1983.
251. A. Bryant, D. L. Lee, J. F. Vetelino, *Ultrason. Symp. Proc.* 1981, 171 – 174.
252. A. D'Amico, *Sens. Actuators* **3** (1982/83) 31 – 39.
253. S. J. Martin, S. S. Schwartz, R. L. Gunshor, R. F. Pierret, *J. Appl. Phys.* **54** (1983) 561 – 569.
254. E. Gizeli, A. C. Stevenson, N. J. Goddard, C. R. Lowe: *Chemical Sensors,* Waseda University Int. Conference Center, Tokyo, Sept. 13 – 17, 1992.
255. P. Hauptmann: *Sensoren, Prinzipien und Anwendungen,* Hanser Verlag, München 1990.
256. G. J. Bastiaans in S. Yamauchi (ed.): *Chemical Sensor Technology,* Elsevier, Amsterdam 1992, pp. 181 – 204.
257. J. Reinbold: *Dissertation*, University of Muenster 2000.
258. J. Curie, P. Curie, *Bull. Soc. Min. Paris* **3** (1880) 90.
259. R. A. Heising: *Quartz Crystal for Electrical Circuits – Their Design and Manufacture,* van Nostrand, New York 1947, p. 16.
260. S. Büttgenbach: *Mikromechanik,* Teubner, Stuttgart 1991, p. 39.
261. L. E. Halliburton, J. J. Martin in E. A. Gerber, A. Ballato (eds.): *Precision Frequency Control,* vol. 1, Acoustic Resonators and Filters, Academic Press, New York 1985.
262. M. H. Ho: "Applications of Quartz Crystal Microbalances in Aerosol Measurements," in C. Lu, A. W. Czanderna (eds.): "Applications of Piezoelectric Quartz Crystal

Microbalances," *Methods and Phenomena,* vol. 7, Elsevier, Amsterdam 1984.
263. W. G. Cady: *Piezoelectricity,* McGraw Hill, New York 1946.
264. J. F. Nye: *Physical Properties of Crystals,* Oxford University Press, London 1957.
265. J. C. Brice: "Crystals for Quartz Resonators," *Rev. Mod. Phys.* **57** (1985) 105.
266. M. Born, K. Huang: *Dynamical Theory of Crystal Lattices,* Clarendon Press, Oxford 1954.
267. E. K. Sittig: "Acoustic Wave Devices," in *Encyclopedia of Physical Science and Technology,* vol. 1, Academic Press, New York 1987, pp. 83 – 109.
268. F. L. Dickert, M. Vonend, H. Kimmel, G. Mages, *Fresenius Z. Anal. Chem.* **333** (1989) 615 – 618.
269. A. W. Snow, J. R. Griffith, N. P. Marullo, *Macromolecules* **17** (1984) 1614.
270. F. L. Dickert, *Chem. Unserer Zeit* **26** (1992) 138 – 143.
271. B. A. Cavic-Vlasak, L. J. Rajakovic, *Fresenius J. Anal. Chem.* **343** (1992) 339 – 347.
272. E. Weber, A. Ehlen, C. Wimmer, J. Bargon, *Angew. Chem.* **105** (1993) 116 – 117.
273. I. Lundström, M. S. Shivaraman, C. M. Stevenson, *J. Appl. Phys.* **46** (1975) 3876 – 3881.
274. M. S. Nieuwenhuizen, W. Barendsz, *Sens. Actuators* **11** (1987) 45 – 62.
275. I. Langmuir: "The Adsorption of Gases on Plane Surfaces of Glass, Mica and Pt," *J. Am. Chem. Soc.* **40** (1918) 1361.
276. J. H. de Boer: *The Dynamical Aspects of Adsorption,* Academic Press, Oxford 1953.
277. A. Clark: *The Chemisorptive Bond,* Academic Press, New York 1974.
278. F. M. Moser, A. L. Thomas: *Phthalocyanine Compounds,* Reinhold Publ. Co., New York 1963, chap. 8.
279. L. F. Barringer, D. P. Rillema, J. H. Ham IV, *J. Inorg. Biochem.* **21** (1984) 195.
280. J. R. Lenhard, R. W. Murvay: "Chemically Modified Electrodes," *J. Am. Chem. Soc.* **100** (1978) 7880.
281. J. S. Miller (ed.), *ACS Symp. Ser.* **182** (1981) part 15.
282. M. T. Reed et al.: "Thermodynamic and Kinetic Data for Cation Macrocyclic Interactions," *Chem. Rev.* **85** (1985) 271.
283. D. W. Armstrong, *J. Liq. Chromatogr.* **7, Suppl. 2** (1984) 353 – 376.
284. M. S. Nieuwenhuizen, A. Venema: "Mass-Sensitive Devices," in W. Göpel et al. (eds.): *Sensors,* vol. II, chap. 13, VCH Verlagsgesellschaft, Weinheim 1991, pp. 657 – 658.
285. J. G. Brace, T. S. Sanfelippo, S. Joshi, *Sens. Actuators* **14** (1988) 47 – 68.
286. E. T. Zellers, R. M. White, S. M. Rappaport, S. W. Wenzel, *Proc. Int. Conf. Sens. Actuators 4th* 1987, Tokyo: IEE Japan, pp. 459 – 461.
287. M. S. Nieuwenhuizen, A. J. Nederlof, *Sens. Actuators B* **2** (1990) 97 – 101.
288. A. W. Snow et al., *Langmuir* **2** (1986) 513 – 519.
289. A. J. Ricco, S. J. Martin, T. E. Zipperian, *Sens. Actuators* **8** (1985) 105 – 111.
290. A. Byrant et al., *Sens. Actuators* **4** (1983) 105 – 111.
291. M. Janghorbani, H. Freund, *Anal. Chem.* **45** (1973) 325 – 332.
292. A. Kindlund, H. Sundgren, I. Lundström, *Sens. Actuators* **6** (1984) 1.
293. D. Soares, *J. Phys. E,* in press.
294. C. Barnes, *Sens. Actuators A* **29** (1991) 59 – 69.
295. S. Bruckenstein, M. Shay, *Electrochim. Acta* **30** (1985) 1295 – 1300.
296. F. L. Dickert, A. Haunschild, V. Maune, *Sens. Actuators B* **12** (1993) 169 – 173.
297. M. Ho, G. G. Guilbault, E. P. Scheide, *Anal. Chem.* **52** (1982) 1998 – 2002.
298. J. M. Jordan, Dissertation, University of New Orleans 1985.
299. A. Suleiman, G. G. Guilbault, *Anal. Chim. Acta* **162** (1984) 97 – 102.
300. J. F. Alder, A. E. Bentley, P. K. P. Drew, *Anal. Chim. Acta* **182** (1986) 123 – 131.
301. Y. Tomita, M. H. Ho, G. G. Guilbault, *Anal. Chem.* **51** (1979) 1475 – 1478.
302. J. A. O. Sanchez-Pedreno, P. K. P. Drew, J. F. Alder, *Anal. Chim. Acta* **182** (1986) 285 – 291.
303. G. M. Varga, Jr., US NTIS AD Report 780171/5 GA, 1974.
304. E. C. Hahn, A. Suleiman, G. G. Guilbault, J. R. Cananaugh, *Anal. Chim. Acta* **197** (1987) 195 – 202.
305. G. G. Guilbault, *Anal. Chem.* **55** (1983) 1682 – 1684.
306. M. J. van Sant, Dissertation, University of New Orleans 1985.
307. C. Kößlinger et al., *Biosens. Bioelectron.* **7** (1992) 397 – 404.
308. C. G. Fox, J. F. Alder: "Surface Acoustic Wave Sensors for Atmospheric Gas Monitoring," in P. T. Moseley, J. Norris, D. E. Williams (eds.):

Techniques and Mechanism in Gas Sensing, Adam Hilger, Bristol 1991, chap. 13.
309. J. Auge, P. Hauptmann, F. Eichelbaum, S. Rösler, *Sens. Actuators,* in press.
310. J. W. Grate, S. J. Martin, R. M. White, *Anal. Chem.* **65** (1993) 987 – 996.
311. A. D'Amico, A. Palma, E. Verona, *Sens. Actuators* **3** (1982/83) 31 – 39.
312. A. Venema et al., *Sens. Actuators* **10** (1986) 47.
313. A. Venema et al., *Electron. Lett.* **22** (1986) 184.
314. A. D'Amico, A. Palma, E. Verona, *Appl. Phys. Lett.* **41** (1982) 300 – 301.
315. A. D'Amico, M. Gentilli, P. Veradi, E. Verona, *Proc. Int. Meet. Chem. Sens. 2nd* 1986, 743 – 746.
316. A. W. Snow et al., *Langmuir* **2** (1986) 513 – 519.
317. J. G. Brace, T. S. Sanfelippo, S. Joshi, *Sens. Actuators* **14** (1988) 47 – 68.
318. J. G. Brace, T. S. Sanfelippo, *Proc. 4th Int. Conf. Sens. Actuators* 1987, 467 – 470.
319. M. S. Nieuwenhuizen, A. J. Nederlof, M. J. Vellekoop, A. Venema, *Sens. Actuators* **19** (1989) 385 –392.
320. H. Wohltjen, *Sens. Actuators* **5** (1984) 307 – 325.
321. A. W. Snow, H. Wohltjen, *Anal. Chem.* **56** (1984) 1411 – 1416.
322. S. J. Martin, S. S. Schwartz, R. L. Gunshor, R. F. Pierret, *J. Appl. Phys.* **54** (1983) 561 – 569.
323. S. J. Martin, K. S. Schweizer, S. S. Schwartz, R. L. Gunshor, *Ultrason. Symp. Proc.* 1984, 207 – 212.
324. A. J. Ricco, S. J. Martin, T. E. Zipperian, *Sens. Actuators* **8** (1985) 319 – 333.
325. J. W. Grate et al., *Anal. Chem.* **60** (1988) 869 – 875.
326. K. Mosbach, *Biosens. Bioelectron.* **6** (1991) 179.
327. P. T. Walsh, T. A. Jones in W. Göpel et al. (eds.): *Sensors: A Comprehensive Survey,* **vol. 2,** VCH Verlagsgesellschaft, Weinheim 1991, p. 529.
328. A. F. Hollemann, E. Wiberg: *Lehrbuch der anorganischen Chemie,* De Gruyter, Berlin 1985.
329. B. Danielsson, F. Winquist in A. E. Cass (ed.): *Biosensors: A Practical Approach,* Oxford University Press, Oxford 1989, p. 191.
330. A. R. Baker, GB 892 530, 1962.
331. K. Mosbach et al., *Biochim. Biophys. Acta* **364** (1974) 140.
332. B. Mattiasson et al., *Anal. Lett.* **9** (1976) 217.
333. B. Danielsson et al., *Appl. Biochem. Bioeng.* **3** (1981) 103.
334. H. G. Hundeck et al., *GBF Monogr. Ser.* **17** (1992) 321.
335. B. Xie et al., *Sens. Actuators B* **6** (1992) 127.
336. E. J. Guilbeau et al., *Trans. Am. Soc. Artif. Intern. Organs* **33** (1987) 329.
337. G. Steinhage: *Dissertation*, University of Muenster 1996.
338. M. G. Jones, T. G. Nevell, *Sens. Actuators* **16** (1989) 215.
339. B. Danielsson et al., *Methods Enzymol.* **137** (1988) 181.
340. S. Wold et al.: *Food Research and Data Analysis,* Applied Sciences, Barking 1983, pp. 147 – 188.
341. T. Nakamoto, K. Fukunishi, T. Moriizumi, *Sens. Actuators B* **1** (1989) 473.
342. K. Ema et al., *Sens. Actuators B* **1** (1989) 291.
343. J. R. Stetter, P. C. Jurs, S. L. Rose, *Anal. Chem.* **58** (1986) 860.
344. U. Weimar et al., *Sens. Actuators B* **1** (1990) 93.
345. W. P. Carey et al., *Sens. Actuators* **9** (1986) 223.
346. D. Wienke, K. Danzer, *Anal. Chim. Acta* **184** (1986) 107.
347. B. R. Kowalski, *Anal. Chem.* **47** (1975) 1152.
348. B. R. Kowalski, *Anal. Chem.* **52** (1980) 112 R.
349. E. Frank, B. R. Kowalski, *Anal. Chem.* **54** (1982) 232 R.
350. K. R. Beebe, B. R. Kowalski, *Anal. Chem.* **59** (1987) 1007 A.
351. D. W. Osten, B. R. Kowalski, *Anal. Chem.* **57** (1985) 908.
352. B. Schlatt: *Diplomarbeit*, University of Muenster 1996.
353. A. Schulga, K. Cammann in G. Henze, M. Köhler, J. P. Lay (eds.): *Umweltdiagnostik mit Mikrosystemen,* Wiley-VCH, Weinheim 1999, pp. 14 – 32.
354. A. F. P. Turner, I. Karube, G. S. Wilson (eds.): *Biosensors – Fundamentals and Applications,* Oxford University Press, Oxford 1987.
355. H.-L. Schmidt, R. Kittsteiner-Eberle, *Naturwissenschaften* **73** (1986) 314 – 321.
356. F. Scheller, F. Schubert: *Biosensors, Techniques and Instrumentation in Analytical Chemistry,* vol. 11, Elsevier, Amsterdam 1992.
357. E. A. H. Hall: *Biosensors,* Open University Press, Buckingham 1990.
358. L. C. Clark, Jr., C. Lyons: *Ann. N. Y. Acad. Sci.* **102** (1962) 29 – 45.

359. J. J. Kulys, A. S. Samalius, G. J. S. Sviermikas, *FEBS Lett.* **114** (1980) 7 – 10.
360. M. J. Eddows, H. A. O. Hill, *Biosci. Rep.* **1** (1981) 521 – 532.
361. M. L. Fultz, R. A. Durst, *Anal. Chim. Acta* **140** (1982) 1 – 18.
362. L. Gorton, *J. Chem. Soc. Faraday Trans. 1* **82** (1986) 1245 – 1258.
363. B. Gründig et al., in F. Scheller, R. D. Schmid (eds.): "Biosensors: Fundamentals, Technologies and Applications," *GBF Monogr.* **17** (1992) 275 – 285.
364. Y. Degani, A. Heller, *J. Phys. Chem.* **91** (1987) no. 6, 1285 – 1289.
365. Y. Degani, A. Heller, *J. Am. Chem. Soc.* **111** (1989) 2357 – 2358.
366. W. Schuhmann, T. J. Ohara, H.-L. Schmidt, A. Heller, *J. Am. Chem. Soc.* **113** (1991) no. 4, 1394 –1397.
367. F. W. Scheller, R. Renneberg, F. Schubert, in S. Colowick, N. O. Kaplan, K. Mosbach (eds.): *Methods in Enzymology,* **vol. 137,** Academic Press, San Diego 1988, pp. 29 – 44.
368. F. Schubert, D. Kirstein, K. L. Schröder, F. W. Scheller, *Anal. Chim. Acta* **169** (1985) 391 – 406.
369. A. Warsinke, B. Gründig, DE 4 227 569, 1993.
370. K. Riedel, B. Neumann, F. Scheller, *Chem. Ing. Tech.* **64** (1992) no. 6, 518 – 528.
371. M. H. Smit, A. E. G. Cass, *Anal. Chem.* **62** (1990) 2429 – 2436.
372. D. M. Rawson, A. J. Willmer, *Biosensors* **4** (1989) 299 – 311.
373. R. S. Sethi: *Biosensors & Bioelectronics* **9** (1994) 243 – 264.
374. M. P. Byfield, R. A. Abuknesha: "Biochemical Aspects of Biosensors," *GEC J. Res.* **9** (1991) 97 – 117.
375. M. E. Eldefrawi et al., *Anal. Lett.* **21** (1988) 1665 –1680.
376. S. Lofas, B. Johnsson, *J. Chem. Soc. Chem. Commun.* 1990, 1526 – 1528.
377. F. Brosinger et al., *Sens. Actuators B* **44** (1997) 350 – 355.
378. D. Kröger, A. Katerkamp, R. Renneberg, K. Cammann, *Biosensors & Bioelektronics* **13** (1998) 1141 – 1147.
379. WO 9416312 A1, 1994 (R. C. Jorgenson, S. S. Yee).
380. A. Katerkamp et al., *Mikrochim. Acta* **119** (1995) (1 – 2) 63 – 72.
381. M. Niggemann et al., *Proc. SPIE-Int. Soc. Opt. Eng.* (1995) 2508. Chemical, Biochemical, and Environmental Fiber Sensors VII, 303 – 311.
382. U. Narang, P. R. Gauger, F. S. Ligler, *Anal. Chem.* **69** (1997) no. 14, 2779 – 2785.
383. J. E. Roederer, G. J. Bastiaans, *Anal. Chem.* **55** (1983) 2333 – 2336.
384. G. G. Guilbault, J. H. Luong, *J. Biotechnol.* **9** (1988) 1 – 10.
385. M. P. Byfield, R. A. Abuknesha, *GEC J. of Res.* **9** (1991) 108.
386. M. Hikuma et al. *Biotechnol. Bioeng.* **21** (1979) 1845 – 1850.
387. C. Diviès, *Ann. Microbiol. (Paris)* **126A** (1975) 175 – 186.
388. G. A. Rechnitz, *Science* **214** (1981) 287 – 291.
389. I. Karube, S. Suzuki, *Ion Sel. Electrode Rev.* **6** (1984) 15 – 58.
390. K. Riedel, R. Renneberg, F. Scheller, *Anal. Lett.* **53** (1990) 757 – 770.
391. Frost & Sullivan: *The European Market for Industrial Gas Sensors* and The European Market for Portable and Transportable Analytical Instruments, Frost & Sullivan Eigenverlag, Frankfurt 1992.
392. INFRATEST: *Chemische und biochemische Sensoren – Marktübersicht,* Infratest Industria, 1992.
393. MIRC: *The European Market for Process Control Equipment and Instrumentation,* Frost & Sullivan Eigenverlag, Frankfurt 1992.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Microscopy

Standard Article

Andres Kriete[1], Heinz Gundlach[2], Severin Amelinckx[3], Ludwig Reimer[4]

[1]School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, USA

[2]Carl Zeiss, Jena, Federal Republic of Germany

[3]Universiteit Antwerpen (RUCA), Antwerpen, Belgium

[4]Physikalisches Insitut Universität Münster, Münster, Federal Republic of Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2005 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.

DOI: 10.1002/14356007.b06_213.pub2
Article Online Posting Date: July 15, 2005

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Modern Optical Microscopy
1.1. Introduction
1.2. Basic Principles of Light Microscopy
1.2.1. Optical Ray Path
1.2.2. Imaging Performance and Resolution
1.2.3. Characteristics and Classification of Lenses
1.2.4. Eyepieces and Condensers
1.3. Illumination and Contrast Generation
1.3.1. Optical Contrast Generation
1.3.2. Fluorescence Microscopy
1.4. Inverted Microscopy
1.5. Optoelectronic Imaging
1.6. Confocal Laser Scanning Microscopy
1.6.1. Basic Principles
1.6.2. Imaging Performance
1.6.3. Instrumentation
1.6.4. Imaging Modalities and Biomedical Applications
1.7. Computer Applications in Digital Microscopy
1.7.1. Image Analysis
1.7.2. 3D Visualization
1.8. High-Throughput Screening for Histopathology and Drug Development
2. Electron Microscopy
2.1. Introductory Considerations
2.2. Conventional Transmission Electron Microscopy (CTEM)
2.2.1. Introduction
2.2.2. Scattering by Atoms: Atomic Scattering Factor
2.2.3. Kinematic Diffraction by Crystals
2.2.3.1. Lattice, Reciprocal Lattice
2.2.3.2. Geometry of Diffraction
2.2.4. Dynamic Diffraction by Crystals
2.2.4.1. General Considerations
2.2.4.2. Basic Equations
2.2.4.3. Dynamic Rocking Curve
2.2.4.4. Anomalous Absorption, Bormann Effect
2.2.4.5. Lattice Fringes

2.2.4.6. Faulted Crystals
2.2.4.7. Moiré Patterns
2.2.5. Operating Modes of the Electron Microscope
2.2.5.1. Microscope Optics
2.2.5.2. High-Resolution, High-Magnification Mode
2.2.5.3. Diffraction Mode
2.2.6. Selected-Area Electron Diffraction (SAED)
2.2.7. Diffraction Contrast Images
2.2.7.1. Imaging Modes
2.2.7.2. Dislocation Contrast
2.2.7.3. Extinction Conditions for Defects
2.2.7.4. Domain Textures
2.2.7.5. Interface Contrast and Domain Contrast
2.2.7.6. Strain Field Contrast
2.2.8. Convergent Beam Diffraction
2.2.8.1. Geometry of Convergent Beam Patterns
2.2.8.2. Point Group Determination
2.2.8.3. Space Group Determination
2.2.8.4. Foil Thickness Determination
2.2.9. High-Resolution Electron Microscopy
2.2.9.1. Image Formation in an Ideal Microscope
2.2.9.2. Image Formation in a Real Microscope
2.2.9.3. Resolution Limiting Factors
2.2.9.4. Image Formation Models
2.2.9.5. Image Interpretation
2.2.10. Scanning Transmission Microscopy
2.2.11. Z-Contrast Images
2.2.12. Analytical Methods
2.2.12.1. X-Ray Microanalysis
2.2.12.2. Electron Energy Loss Spectrometry (EELS)
2.2.13. Specimen Preparation
2.2.13.1. Diffraction Contrast Specimens
2.2.13.2. High-Resolution Specimens
2.2.14. Applications to Specific Materials and Problems
2.2.14.1. Crystal Structures
2.2.14.2. Defects
2.2.14.3. Small Particles
2.2.14.4. Surface Studies
2.2.14.5. Thin Epitaxial Layers
2.2.14.6. “In situ” Studies
2.3. Scanning Electron Microscopy
2.3.1. Introduction
2.3.2. Instrumentation
2.3.2.1. Electron Guns
2.3.2.2. Electron Probe Formation
2.3.2.3. Detectors
2.3.3. Electron–Specimen Interactions
2.3.3.1. Elastic and Inelastic Scattering
2.3.3.2. Electron Diffusion
2.3.3.3. Emission of Secondary and Backscattered Electrons
2.3.3.4. Specimen Charging and Damage
2.3.4. Image Formation and Analysis
2.3.4.1. Topographic and Material Contrast
2.3.4.2. Electron Channeling Effects
2.3.4.3. Imaging and Measurement of Surface Potentials
2.3.4.4. Imaging of Magnetic Fields
2.3.4.5. Electron-Beam Induced Current
2.3.4.6. Cathodoluminescence
2.3.4.7. Special Imaging Methods

2.3.5. Elemental Analysis
2.3.5.1. X-Ray and Auger Electron Emission
2.3.5.2. X-Ray Spectrometers
2.3.5.3. X-Ray Microanalysis
2.3.5.4. Special X-Ray Techniques

[Top of Page]

1. Modern Optical Microscopy

Andres Kriete, Heinz Gundlach, Severin Amelinckx

1.1. Introduction

Optical microscopy goes back as far as the 16th century, when magnifying glasses and optical lenses became available. ANTONI VAN LEEUWENHOEK (1623 – 1723) observed structures lying beyond the resolution limit of the eye, such as bacteria, with a 270-fold magnifying, single lens. The exact design of the microscope was established from a theory of image formation developed by ABBE in 1873 [1]. With the introduction of apochromatic lenses (⟶ Optical Materials – Effect of Partial Dispersion on Optical Imaging) and oil immersion, a theoretical resolution limit of 0.2 µm became available at this early stage. Light microscopy advanced further by new lens designs, optical contrast enhancement methods, and application of fluorescent and immunofluorescent dyes, also in combination with inverted microscope designs. Optoelectronic methods, such as video techniques, have further broadened the spectrum of light microscopic applications. Finally, laser scanning techniques and confocal imaging allowed the investigation of microscopic structures three-dimensionally. Compared with a standard microscope, the confocal microscope has enhanced lateral and axial resolution, as well as it has improved contrast, and in particular it can remove out-of-focus blur. Thus, an optical, noninvasive sectioning capability is achieved. The breakthrough of confocal microscopy in biomedicine and material science boosted the development of a variety of commercial confocal laser scanning microscopes. Highly developed optoelectronic components such as laser light sources and sensitive detectors, as well as the application of specific fluorescent dyes, enabled wide acceptance of this new imaging technique together with effective use of powerful computers to digitally store, visualize, and analyze three-dimensional data. The combination of these developments has also led to high-throughput applications of modern light microscopy for drug development.

1.2. Basic Principles of Light Microscopy

1.2.1. *Optical Ray Path*

The rigid *stand* of the microscope is the “carrying” element, with mechanical and optical parts attached. The *objective* is the central optical unit, generating an intermediate image that is magnified by the *eyepiece* for visual observation [2-4]. *Condenser* and *collector* optimize specimen illumination. Lens designs corrected for a *finite length* form a mirrored, magnified, aerial (intermediate) image, which is projected infinity by the eyepiece (Fig. 1). The distance between the back focal plane of the objective and the primary image plane (which is the front focal plane of the eyepiece) is defined as the *optical tube length*. The *mechanical tube length*, measured from the mounting plane of the objective to the mounting plane of the eyepiece, is 160 mm for currently used microscopes. Any aberrations present in the intermediate image, such as the lateral chromatic aberration (LCA), are compensated by the eyepiece. Optical – mechanical components such as the Bertrand lens, or epi-illuminating reflectors mounted within the optical path require a so-called telan system. Objectives of *infinity-corrected designs* generate an infinite image projected into a real image by the tube lens (Fig. 2). The infinity color-corrected system (ICS) optic is an example of how the tube lens together with the objective can correct aberrations, particularly chromatic ones [5]. The intermediate image is monochromatic; consequently, photographic and video recording do not require compensating lenses. Moreover, the path between lens and tube lens may be used by all optical – mechanical components for various illuminating and contrast generating methods without telan optics (Fig. 2).

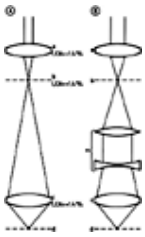

Figure 1. Schematic optical path for objectives of finite length; compensating eyepieces correct for chromatic aberrations

A) Conventional microscope with finite optics; B) Microscope with finite optics and telan optics for insertion of optical components

a) Eyepiece; b) Primary intermediate image; c) Objective; d) Object plane; e) Telan system (incident light fluorescence, Optovar, Bertrand lens)

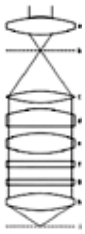

Figure 2. Schematic optical path for objectives corrected for infinity with color- corrected intermediate image

a) Eyepiece; b) Plane of intermediate image; c) Tube lens; d) Bertrand lens slider; e) Optovar 1 X, 1.25 X, 1.6 X, reflector for incident light; f) λ/4 plate, plate, compensators; g) Differential interference contrast (DIC) slider; h) Objective; i) Specimen plane

1.2.2. *Imaging Performance and Resolution*

The image generation in microscopy cannot be explained sufficiently by means of geometric optics alone; wave optics

must also be considered [1, 6]. Imaging a point source generates a diffraction pattern in the focal plane. The central parts of this Airy disk contain most of the intensity; therefore, the diameter of this kernel can be defined as the smallest resolvable image element (Fig. 3 A – D). The diameter of the kernel d_k is given by

$$d_k = 1.22\lambda / n \cdot \sin \alpha \tag{1.1}$$

with λ being the wavelength, *n* the refractive index of the material, α the angle between the optical axis and the marginal beam of the light cone that enters the objective, and $n \cdot \sin \alpha$ the numerical aperture (*NA*).

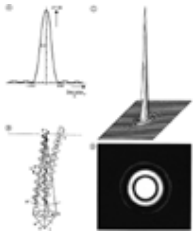

Figure 3. A) and B) Intensity distributions in the image of a luminous point; C) Computed intensity distribution; D) Airy disk

a) Image plane; b) Lens; c) Objects plane

The resolution limit for two points is reached when the maximum of the first diffraction spot coincides with the first minimum of the second one. ABBE developed this theory and proved it experimentally in his historical papers [1]. The resolution is calculated from the available numerical aperture and the wavelength λ by

$$d_r = 1.22\lambda / NA_{obj} + NA_{cond} \tag{1.2}$$

As an example, a lens and condenser of *NA* = 1.4 and λ = 530 nm give a smallest separation of *d* = 200 nm. Resolution can be improved either by a higher *NA*, which is difficult to obtain in lens design, or by using shorter wavelength, as in ultraviolet microscopy [7].

Both the size of the diffraction spot and the resolution of the eye define the utilizable magnification. Since the resolution of the eye is ca. 2 – 4 arcminutes, total magnification should not exceed 500 – 1000 times the numerical aperture. Magnification above this limit does not deliver any additional information, but may be used for automatic measuring procedures or counting.

1.2.3. *Characteristics and Classification of Lenses*

The imperfections of optical lenses give rise to various distortions, either monochromatic or chromatic, (see also ⟶ Optical Materials – Effect of Optical Material on Imaging) that are visible in the diffraction pattern. Monochromatic distortions include aberration, curvature, astigmatism, and coma. Chromatic aberrations are caused by differences in the refractive index of glasses. These types of errors are minimized by combining several pieces of glass with different form and refractive index, or by using diaphragms. The three main categories of lenses are

1. Achromatic objectives
2. Fluorite or semiapochromatic objectives
3. Apochromatic objectives

Fluorite and apochromatic lenses are mostly designed as plane (i.e., corrected for a flat field of view). Computers allow calculation of the diffraction pattern under various imaging conditions to optimize imaging performance for a wide range of applications. The present advantage in modern lens design and production is indicated by the image quality of the ICS Plan Neofluar class. At a field-of-view number of 25, color photography, as well as all optical contrast enhancement methods, are supported with achromatic correction comparable to apochromatic lenses with a good transmission down to UV at 340 nm. These objectives are generally applicable for all fluorescent methods (see Sections Fluorescence Microscopy, Imaging Modalities and Biomedical Applications). Plan apochromates feature a still higher aperture: chromatic rendition is improved, as well as correction at the field-of-view boundary; therefore these objectives are very suitable for color microphotography. In addition, a complete correction not only at the focus level but also for a limited range above and below the focal plane is achieved. For specific applications, the following lenses are also available:

1. Multi-immersion objectives of type Plan Neofluar, corrected for use with oil, glycerol, or water, with or without a cover glass
2. Objectives for specimen without cover slide
3. Objectives with a long working distance
4. Objectives with iris diaphragm
5. Objectives for phase contrast investigation
6. Objectives, strain free for polarized light microscopy
7. Ultrafluars, corrected between 230 and 700 nm

1.2.4. *Eyepieces and Condensers*

The main characteristics of *eyepieces* are magnification and type of correction, such as color compensation and correction of viewing angle. The correction of the *condensers* must match the properties of the lenses. Special or combined condensers are required for certain methods of contrast generation or for longer working distance with inverted microscopes.

1.3. Illumination and Contrast Generation

Illumination and contrast are central points in the discussion of imaging performance. Without sufficient contrast, neither the required magnification nor a maximal resolution can be obtained. A. KÖHLER developed optical illumination for microscopy in 1893 [8]. The main advantage of Köhler illumination is that in spite of the small filament of the light bulb, the entire area of the lamp field stop, as well as the illuminated part of the specimen, have a uniform luminance (Fig. 4). The twofold ray path of diaphragms and lenses suggests that adjustment of the condenser iris, being the illumination aperture, controls resolution, contrast, and resolution depth. This principle is also important for correct setup of the confocal microscope as well as electron microscopy.

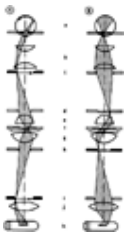

Figure 4. Schematic of Köhler illumination

A) Image-forming ray path; B) Illuminating ray path

a) Eye; b) Eyepiece; c) Eyepiece field stop; d) Exit pupil of objective; e) Objective; f) Specimen; g) Condenser; h) Aperture diaphragm; i) Lamp field stop; j) Light collector; k) Light source

1.3.1. *Optical Contrast Generation*

For the following considerations the principles of wave optics are used, which describe image formation as a result of diffraction and interference.

Dark-Field Illumination. In dark-field illumination the illuminating aperture is always set higher than the lens aperture. Only light waves being diffracted by the microstructures are captured by the lens. Objects appear as bright diffraction patterns on a dark background.

Phase Contrast. Based on ABBE's theory the Dutch physicist FRITS ZERNIKE (Nobel prize in 1953) developed phase contrast imaging in 1935, to convert the invisible phase shift within the specimen into recognizable contrast [9]. A phase plate located in the lens modifies the phase and amplitude of the direct (illuminating) wave in such a way that a phase shift of $\lambda/2$ is obtained (Fig. 5). The amplitudes of the direct diffracted waves are equalized, and by destructive interference, dark-phase objects appear on a bright background. The application of this technique ranges from routine diagnosis to cell biology research [6, 10]. Today this method is also used advantageously in combination with epifluorescence. The specimen is adjusted under phase contrast before switching to fluorescence light in order to avoid bleaching, or both imaging modes are combined to match the fluorescent signal with the morphological image.

Figure 5. Image path of the phase contrast microscope, converting phase shifts into a noticeable contrast

a) Illuminating diaphragm ring; b) Condenser; c) Specimen generating a phase shift; d) Objective plate; e) Phase plate; f) Intermediate image

Nomarski Differential Interference Contrast (DIC). In 1955 the physicist GEORGE NOMARSKI simplified the two-beam interference microscope in a way that it became available for routine microscopy [11] (Fig. 6). DIC uses modified Wollaston prisms (Fig. 6 B), lying outside the focal areas of condenser and objective [11-13]. Normal lenses are used; the contrast can be adjusted by shifting the lens-sided prisms. As in phase contrast, differences in the optical path length are visualized. Objects appear as a relief, and thick specimens lack the halo-typical pattern of phase contrast imaging. Full aperture of both the objective and the condenser allow optical sectioning (Fig. 7). Combined with inverted microscopes (see Section Inverted Microscopy) and contrast-enhanced video microscopy (see Section Optoelectronic Imaging), ultrastructures in cells and tissues can be imaged. Certain applications combine DIC with phase contrast. For routine applications, phase contrast lenses may give better results at lower magnification. The DIC image is well suited to be combined with a fluorescent image as well.

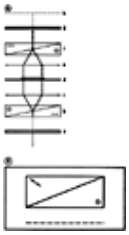

Figure 6. A) Image path of a two-beam interference microscope; B) Nomarski prism with direction of optical axes and position of interference plane indicated

a) Polarizer (45°); b) First Wollaston prism; c) Condenser; d) Specimen; e) Objective; f) Second Wollaston prism; g) Analyzer (135°); h) Intermediate image

Contrast is generated by a phase shift of the polarized and prism-split illuminating beams, which are combined by the second prism

Figure 7. A) – C) Optical sectioning of salivary gland of drosophila by DIC Nomarski (Specimen courtesy of L. Rensing, Universität Bremen, Germany)

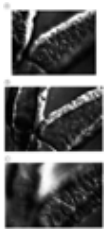

Polarization and Reflection Contrast Microscopy. Polarization microscopy is used preferentially for certain applications in cell research, to enhance double diffraction (anisotropic) structures such as muscle fibers or the mitotic apparatus. Quantitative studies require strain-free polarization lenses for the polarized light used. With the help of reflection contrast microscopy [14, 15], visible interference is generated at epi-illumination, for example, to study adhesive properties of living cells on glass surfaces [16]. Moreover, this method may be combined with phase contrast or Nomarski DIC contrast. For the evaluation of immunogoldmarked specimens, immunogold staining (IGS), reflection as well as transmission contrast is feasible.

1.3.2. *Fluorescence Microscopy*

Basic Principles. The development of immunofluorescence greatly influenced medical practice and basic biomedical research [17-20]. Many biomedical specimens emit *autofluorescence* if illuminated properly. Today a wide range of fluorescent microstructures can be excited by application of fluorochromes (*secondary fluorescence*). Immunofluorescence is based on antigen – antibody connectivity [21]. This couples the high specificity of the immune reaction with the extreme sensitivity of fluorescence microscopy. Various markers for antibodies exist and derivates of fluorescein, mainly fluorescein – isothiocyanate (FITC), are the preferred fluorochromes. For this kind of microscopy, special lamps and filters are required. Light sources used include high-pressure bulbs as well as lasers. Fluorescence microscopy is used either in transmission or in epi-illumination; the latter combines phase contrast and Nomarski DIC [22, 23]. Fluorescence microscopic objectives with high aperture and a good transmission between 340 and 700 nm are preferred. Multi-immersion lenses of plan-type Neofluar have been designed specifically for such applications, as well as low-magnification lenses featuring high apertures (primary magnification 40, *NA* 1.3 oil). Plan apochromates are more suitable for illumination in visible light. The brightness *b* of fluorescent images increases proportionally to the square of the numerical aperture of lens L and condenser C, and decreases inversely with the square of the total magnification *M*

$$b \sim \frac{NA_L^2 + NA_C^2}{M^2} \tag{1.3}$$

This makes ocular lenses with low primary magnification an ideal choice. At epi-illumination the brightness depends on the power of four of the aperture, since the lens acts as the condenser as well:

$$b \sim \frac{NA_L^4}{M^2} \tag{1.4}$$

In the 1980s, fluorescence methods became a major tool for sensitive and highly specific detection in biomedical research and medical diagnosis.

Fluorescent Dyes and Markers. The development of reagents based on fluorescence includes the use of biological molecules: antibodies, RNA, DNA, proteins, and other macromolecules and physiologic indicators [17-19]. These methods have enabled new findings in cell biology, molecular biology, developmental biology, and cancer research. The cytoskeleton (Fig. 8) is responsible for the cell shape, cell division, capacity of the cell to move, and the intracellular transport of cell organelles [24, 25]. Multiparameter fluorescence microscopy uses a broad selection of fluorescent dye molecules to examine several properties or components of a single cell simultaneously (Fig. 9).

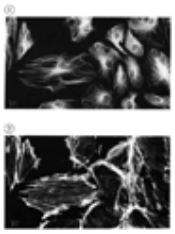

Figure 8. Double fluorescence micrograph of epithelial kidney cells

A) Microtubules, fluorescent dye FITC; B) Actin filament, fluorescent dye tetramethylrhodamine-B-isothiocyanate (TRITC)

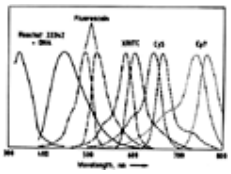

Figure 9. Excitation and emission curves of various fluorochromes

Filter curves depict dyes Hoechst 33342 (suitable for DNA), FITC (for actin), rhodamine (endosomes), and cyanine 5 (mitochondria)

(Courtesy of L. TAYLOR, University of Pittsburgh, United States)

A series of new probes suitable for this purpose has been developed by extending the usable spectrum to the near

infrared (Fig. 10). *Fluorescence-based in situ hybridization* (FISH) is rapidly being recognized as a powerful tool for detecting and quantifying genetic sequences in cells and isolated chromosomes [26]. FISH is based on the highly specific binding of pieces of DNA, labeled with a fluorescent probe, to unique sequences of DNA or RNA in interphase nucleus or metaphase chromosomes. Recent developments enable the simultaneous visualization of three or more DNA regions (chromosome painting) [27].

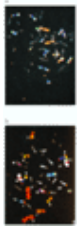

Figure 10. A) Micrograph by triple exposure of color slide film using conventional fluorescence microscopy; B) Pseudocolored digital image recorded with the CCD camera for chromosome 1, 2, 7, 8, 9, 12 and the human X chromosome

Green = FITC; Red = TRITC; Blue = DAPI

(Courtesy of T. CREMER, Heidelberg, Germany and Oxford University Press)

Comprehensive reviews of available fluorochromes have been published elsewhere [28-31]. UV lasers broaden the spectrum of applicable fluorescence dyes. The green fluorescence protein (GFP) is a strongly fluorescent reporter molecule that stems from the jellyfish *Aquorea victoria* [32]. It has a fluorescein-like characteristic, it requires no cofactors or substrates, and is species-independent. GFP engineering has produced different color mutants which have been used for gene expression, tracing of cell lineage, and as fusion tags to monitor protein localization [33].

Fluorescent semiconductor nanoparticles (quantum dots, Q-dots) have the benefit of bright fluorescence without any photobleaching, thus allowing for extended-time observation of a biological system [34]. Simultaneous tracking of Q-dots of different colors Q-dots is possible due to their broad absorption and narrow emission spectrum [35]. They are excited by wavelengths below 480 nm, and emit at various wavelengths from 490 to 700 nm with sharp, well-defined absorption spectra [36]. Full width at half maximum intensity ranges from 20 to 40 nm. Different types of coating chemistries have been employed to render the Q-dots hydrophilic and suitable for further conjugation with antigens and antibodies of interest (see Fig. 19) [37].

Time-Resolved Methods. Special methods and dyes have been developed to observe functional phenomena in living cells. Fluorescence lifetime imaging microscopy (FLIM) monitors fluorescence lifetime as a function of environmental parameters of cellular components. These include Ca^{2+}/pH-sensitive indicators, also called biosensors [19]. Typically a fluorescent dye is chemically linked to a macromolecule, creating an analogue of the natural molecule. The fluorescent analogue is specifically designed to measure chemical properties of the macromolecule, instead of tracking its location within the cell. Time-gated multichannel plates (MCPs) are used for time- and spatially resolved measurements.

Fluorescence recovery after photobleaching (FRAP) and the measurement of the fluorescence resonance energy (FRET), which allow the diffusional mobility of cellular components and their interaction at the molecular level to be monitored are other varieties of time-resolved methods [38, 39]. A novel method is fluorescence correlation spectroscopy (FCS), which measures the statistical fluctuations of fluorescence intensity within a confocally illuminated volume. Correlation analysis allows the concentration of particles and their diffusion to be determined [40, 41]. FCS has proved be useful in the study of enzyme kinetics, DNA/protein interactions, receptor/ligand interactions, and cellular measurements [42]. Molecular interactions can be studied without interfering with the natural binding conditions, since only the reference substances in competition experiments need to be fluorescently labeled.

1.4. Inverted Microscopy

Inverted microscopes were originally designed for the evaluation and counting of plankton and sediments [43]. The first inverted research microscope which included photomicrography, microphotometry, cinemicrography, and video imaging was the Axiomat [4, 44-46]. Inverted microscopes in U-design (Fig. 11) with a rigid stage are of great advantage for investigations in cell and developmental biology, as well as in cell physiology experiments that are performed in optical microscopes [47-49]. Identification and selective micropreparation on live nuclear components in oocytes of *Xenopus* in the DIC Nomarski technique facilitate high-resolution optical microscopy in different planes of the specimen (optical sectioning), which can be isolated by selective micropreparation for subsequent analysis in the electron microscope [50].

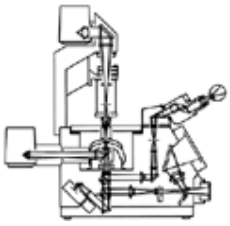

Figure 11. Light path of an inverted microscope

In the life cycle of the green alga *Acetabularia*, high-resolution phase and DIC microscopy have made part of the endoplasmic reticulum and lampbrush chromosomes visible [51]. The distinct advantage of inverted compared to upright microscopes is the unhindered application of cell and tissue culture chambers in conjunction with micromanipulators, micropipettes, microelectrodes (patch clamp method), and injection systems for experimental cell research. Computer-controlled microinjection can be performed automatically [52]. Manipulation of cells, organelles, and genomes is also possible by means of laser microbeam tweezing and optical trapping [53].

1.5. Optoelectronic Imaging

Electronic imaging methods such as video microscopy and laser scanning microscopy can extend resolution slightly, but the ability to extend the detectability of subresolved structures and subresolution movements is more important. A video camera can generally be mounted to the same port employed for normal photomicrography. In *video-enhanced contrast microscopy* (VEC) (Fig. 12), high-resolution video cameras based on a vidicon tube are used. This principle of adjusting gain, and offset and digital image processing is also used in laser scanning microscopes (see Section Confocal Laser Scanning Microscopy). Digital imaging and image processing make it possible to see structures that are smaller than the resolution limit [12, 54-57]. By means of this method, single microtubules can be observed [58, 59]. The size of each microtubule is only one-tenth of the visible-light resolution limit. In the 1980s, solid-state cameras were introduced into microscopy [60, 61]. Most of them are based on a *charge coupled device* (CCD), which produces digital data directly for computer processing and image analysis. For low light level (LLL) or video intensification microscopy (VIM), i.e., weak fluorescence or small FISH signals (see Fig. 10 B), cooled, sensitive CCD cameras for black and white or color are available now and have replaced tube cameras [62]. Digital imaging techniques have gained increasing attention as supplements to conventional epifluorescence and photomicrography [63]. *Microphotometry* permits spectrophotometric analysis of substances, as well as statistical and kinetic measurement and compilation of two-dimensional intensity profiles [49].

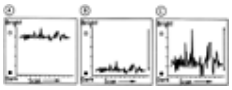

Figure 12. Principle of video-enhanced contrast microscopy

A) Original image; B) Offset (brightness); C) Contrast gain

1.6. Confocal Laser Scanning Microscopy

1.6.1. *Basic Principles*

The fundamentals of confocal microscopy go back to 1959, when MINSKY described a new kind of microscopic apparatus [64]. The historical development of confocal microscopy is reviewed elsewhere [65, 66]. Two different kinds of confocal instruments are available: the tandem scanning microscope and the laser scanning microscope. The *tandem scanning microscope* uses normal light chopped by a rotating Nipkow disk [67]. The light beams reflected by the specimen cross the rotating disk on its opposite side (thus tandem), and due to the symmetric pattern of holes, out-of-focus blur is removed. This system allows real-time and real- color imaging, but is limited in sensitivity because of the fixed size of the pinholes.

The more commonly applied *confocal laser scanning microscope* (CLSM) features a variable pinhole and allows adjustment of magnification by modifying scan angles. A schematic example of the principle of confocal laser scanning is given in Figure 13. The laser light is focused via the scanner (b) through the tube lens (c) and the objective (d), and illuminates a small spot in the specimen (e). Emitted light emanating from the focal plane and the planes above and below (dotted and dashed lines) is directed via the scanner to the dichroic beam splitter (a) where it is decoupled and directed onto a photomultiplier (i). A pinhole (h) in front of the photomultiplier is positioned at the crossover of the light beams emerging from the focal point. This plane corresponds to the intermediate image of the Köhler illumination described in Section Illumination and Contrast Generation. Light emanating from above and below the focal point has its crossover behind and before the pinhole plane so that the pinhole acts as a spatial filter. Numerous papers elucidate the basic aspects of confocal image formation and application [68-72], including a handbook [73].

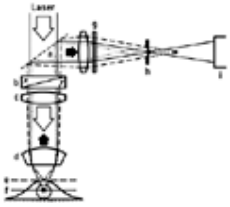

Figure 13. Ray path in the confocal microscope

a) Dichroic beam splitter; b) Scanner; c) Tube lens; d) Objective; e) Specimen; f) Focal plane; g) Filter; h) Pinhole; i) Detection system (photomultiplier)

Since only a small specimen point is recorded, the entire image of a specimen is generated by either moving the specimen (*stage scanning*) or moving the beam of light across the stationary specimen (*beam scanning*). The latter is utilized in all commercial instruments (Fig. 14). Theoretical considerations have shown that both scanning forms give identical resolution. The optical sectioning capability at lateral scanning opens a way to image a three-dimensional structure with a slight axial stepping of the stage (see Fig. 15).

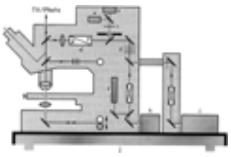

Figure 14. Setup of the confocal microscope, based on a research light microscope (Zeiss)

a), b) Photomultiplier tubes; c) Confocal variable pinhole; d) Scanners; e) Barrier filter; f) He – Ne laser (543 nm, 633 nm); g) Stage; h) Ar laser (488 nm, 514 nm); i) UV laser (350 – 360 nm); j) Antivibration table

Light of internal (He – Ne) and external (Ar ion or UV) lasers passes the scanner unit; switchable mirrors and barrier filters permit multiwavelength operation. Additional barrier filters are located between the confocal variable pinhole and the beam splitter adjacent to the photomultipliers.

Figure 15. A) Conventional versus confocal imaging (inserted region of interest) of a lung structure in fluorescence;

B) Multislice image sequence obtained by optical sectioning

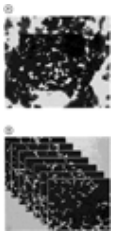

1.6.2. *Imaging Performance*

The imaging characteristic in confocal microscopy is given by the properties of the specimen weighted by the applicable spatial confocal response function. An explanation based on wave optics suggests that the point spread functions of the illuminating path and the detection path (resembling the photon detection possibility) are convolved, consequently resulting in a narrow point resolution [74].

Describing image formation of an reflecting object and the corresponding coherent imaging properties must be distinguished from a fluorescent object viewed in incoherent imaging because in the latter the illuminating wavelength and the detected wavelength are different. In terms of resolution, the Rayleigh criterion determines the minimum resolvable distance between two points of equal brightness (see Eq. 1.2). For *incoherent imaging* using a lens with a certain numerical aperture *NA* at a wavelength λ, the *lateral resolution* d_r obtainable in confocal imaging is [71]:

$$d_r\,(\text{conf.}) = 0.46\lambda/NA \tag{1.5}$$

With a wavelength of 514 nm (green argon laser line) and a numerical aperture of 1.4, this equation gives a resolution of d_r = 157 nm. Compared to conventional imaging (see Eq. 1.2), the resolution is ca. 32 % higher [75].

An approximation of the *axial resolution* d_z is given by

$$d_z\,(\text{conf.}) = 1.4\lambda/NA^2 \tag{1.6}$$

The values (514 nm and *NA* = 1.4) give a d_z of 367 nm.

One way to measure the axial response of a confocal microscope is to record light intensity from a thin layer with infinite lateral extension. In a conventional microscope, no difference — and therefore no depth discrimination — are obtained. However, in confocal microscopy, a strong falloff occurs if the object is out of focus. The free width at half maximum (FWHM) of such a function is proportional to the optical section thickness [73, 76]. The optical section thickness also depends on the diameter of the pinhole. The optimum pinhole size *D* with a maximum deterioration of 10 % in reflection is given by [72]

$$D = 0.95\lambda M/NA \tag{1.7}$$

with magnification *M*, λ, and *D* given in micrometers. In fluorescence microscopy the size has to be reduced 10 – 30 %, depending on the ratio of excitation to emission wavelengths. Another way to describe the imaging performance is by the modulation transfer function (MTF) (see also ⟶ Photography – Modulation Transfer Function.); such measured and theoretically derived three-dimensional functions have been published [77, 78].

1.6.3. *Instrumentation*

Confocal microscopic equipment is available as an add-on to a standard microscope or in the form of more fully integrated devices, including computer control equipment. Both upright and inverted types of microscopes are available. The inverted confocal microscope allows bigger specimens to be studied and offers easy handling of micromanipulating or injecting devices (see also Section Inverted Microscopy). Beam scanning is obtained by rotating galvanometric mirrors. Axial control is realized by a stepping motor drive or a piezoelectric translator. Most instruments are equipped with several detectors to record multifluorescence events simultaneously. Some also offer transmission detectors. Signals from different channels can then be mixed electronically (see Fig. 16). Confocal transmission microscopy has also been studied. Improvement of contrast and resolution depending on the pinhole size of a thin specimen has been documented in a confocal transmission arrangement using differential interference contrast [79].

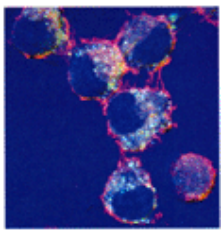

Figure 16. Multichannel confocal image of macrophages (J774, mouse cell line)

Extrathyroidal release of thyroid hormones from thyroglobulin is marked by fluorescence dye (first channel, green); actin cytoskeleton is labeled with rhodamine-phallodin (second channel, red). Morphological information (third channel), providing a dark background in the cells, is detected in transmission.

(Permission of K. BRIX, V. HERZOG, Institut für Zellbiologie, Bonn, Germany)

Variations of the confocal design have been developed, such as bilateral scanning in real time with a scanning slit [80]. Because of the fast scanning, TV detectors such as a sensitive CCD camera are best suited. Slit-scan confocal microscopes offer a better signal-to-noise ratio and real-time imaging capabilities but a slightly reduced resolution

compared to a conventional CLSM.

Various types of lasers can be linked to the microscope [81], but up to now only continous-wave gas lasers have been used routinely. This includes argon-ion lasers having excitation lines at 488 and 514 nm and helium – neon lasers with a major line at 633 nm and a weaker line at 543 nm. The use of UV lasers requires achromatic lenses and specifically designed scanning mirrors. Lasers used in the UV range are argon-ion UV or argon-ion tunable lasers, as well as krypton-ion and helium – cadmium lasers. Depending on the emission lines of the fluorescent dyes, various filters are used in front of the photomultiplier, characterized by quantum efficiency and spectral response. The total sensitivity of a system is given by the transfer efficiency of the various optical components [82] and, in addition, the quantum efficiency of the detector and its associated electronics [83].

Simplified types of laser scanning microscopes, such as the *fiber scanning optical microscope* have also been proposed [77]. Here the fiber can act both as a light delivering source and as a pinhole. A totally different physical approach to reduce out-of-focus blur is based on a physical phenomenon called *double-photon (2 p) excitation*, whereby fluorescence is generated by simultaneous absorption of two photons of long wavelength at a single molecule. The necessary energy emitted by pulse lasers is present only at the focal plane; thus a “confocal” effect is produced. At turbid tissue media, the signal level under 2 p-excitation drops much faster than that under single-photon excitation, although image resolution is higher in the former case [84, 85].

The diffraction pattern of the illuminating beam is elongated axially compared to the lateral direction (see Eq. 1.5 versus 1.6). This anisotropy is caused basically by the limited numerical aperture of the microscopic lens. An optimal diffraction pattern would be achieved by an unlimited, 360° aperture. This concept is known as the *4-pi microscope*. Prototypes use two facing lenses or two lenses oriented at an angle (theta microscopy). The object in focus is illuminated from two sides, and fluorescence emanating from the object is detected through both lenses [86]. Such concepts have been recognized as entry-points for fluorescence nanoscopy [87].

Special techniques for 3D imaging of particular thick specimen several microns in depth have also been suggested, including optical coherence tomography [88] and selective plane illumination microscopy [89].

1.6.4. *Imaging Modalities and Biomedical Applications*

Imaging Modalities. Confocal contrast generation include reflectance, rescattering, and fluorescence. The first results in confocal imaging were obtained on unstained tissue in *reflectance* [90]. Reflectance is also the preferred imaging modality used in industrial inspection of semiconductors, often combined with methods such as optical beam induced conductivity (OBIC).

In biomedical science, the study of unstained tissue is limited due to reflectance. The reflectance must be intensified by metal impregnation (e.g., Golgi stain); one such example is given in Figure 17. Other examples include peroxidase – labeling with nickel intensification in neurobiology [91] or silver-stained nucleolar organizer region-associated proteins (AgNORs). Gold immunolabeling has also been used [92]. In *rescattering*, variations in the refractive index can give rise to reasonable contrast. One particular example is in ophthalmology [93] where structures in the transparent cornea such as cell nuclei, keratocytes, neurons, or the fibers of the lens can be visualized (see Fig. 18).

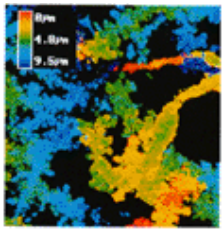

Figure 17. Golgi-stained neuron: three-dimensional display of a confocal data set acquired in reflection

Color depth coding visualizes the three-dimensional neuron; various spines of different size give a brushlike impression.

(Specimen courtesy of B. FORTMANN, Institut für Ernährungswissenschaften, Gießen, Germany)

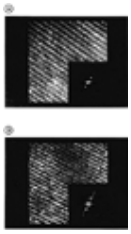

Figure 18. Confocal images of the eye's lens fibers, resulting from rescattering at microstructures with different indices of refraction

A) Imaging through the 400 µm transparent cornea;

B) Imaging with cornea removed

The inserted Fourier power spectrum documents the loss in spatial frequency content.

(Specimen courtesy of B. R. MASTERS, USUHS, Bethesda, United States)

Fluorescence is the most commonly used imaging mode in confocal microscopy, such as, applications in cell and neurobiology [93-100], which include the use of autofluorescence, specific dyes in combination with antibodies, and in situ hybridization. For specifics on fluorescent dyes and markers see Section Fluorescence Microscopy.

Biomedical Applications. Optical sectioning devices, being noninvasive and having low radiation damage, are ideal for studying living specimen, organs, and tissues. With confocal microscopes, structures can be observed in vivo at full three-dimensional microscopic resolution [94]. Small organisms or organisms in an early stage of development can be observed, as in embryogenesis [97, 98]. Concerning individual structures, kinetics of labeled receptors (see Fig. 19), morphological changes, such as the movement of cell compartments (see Fig. 20), growth of neurons [99], or synaptic plasticity in the retina, have also been studied [100]. Confocal microscopy is also used in clinical imaging, which includes the observation of wound healing, flow processes in veins, and ophthalmology [94, 96] (see Fig. 18).

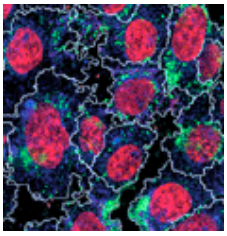

Figure 19. A431 cells stained to study EGF-receptor internalization. 605 streptavidin Q-dots linked to EGF-R were in contact with the cells for 20 min. Nucleus is stained with Dapi, and cell membranes with DiO.

Green = Q-dots; Red = DAPI; Blue = DiO

Image courtesy E. PAPAZOGLU, Drexel University, Philadelphia

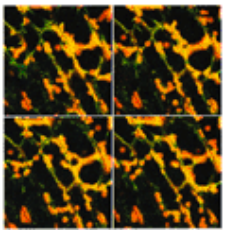

Figure 20. Application of confocal microscopy to plant cells (*Egeria densa Planchon*), in vivo

Cell boundaries and the movement of chloroplasts are visualized by autofluorescence (green) and FITC labeling (red).

(Permission of I. DAHSE, Institute für Biophysik, Jena, Germany)

1.7. Computer Applications in Digital Microscopy

Computer applications in digital microscopy include image storage in databases, compression, enhancement, analysis, and visualization. Digital analysis applied to microscopical data sets has the aim of automatically extracting structures of interest out of digitized images and then quantifying them. This includes field-specific or stereological, geometrical or object-specific, and intensity-related parameters. In the case of 3D images, typically generated by a confocal imaging technique, a computer graphical 3D visualization of the data stack is performed, which allows inspection of the original data and subsequent image-processing operations.

Different digital image formats are used in microscopy, for example, the TIFF format. In 1999, a microscopic imaging standard was established by the ARC-NEMA Institute as part of the DICOM standard (Digital Imaging and Communication in Medicine). This format, DICOM-VL (Visible Light), regulates storage of instrumental and specimen parameters besides the image information. This fosters the use of microscopic images in telemicroscopic applications and in medical networks. The vast amount of data in larger studies, and in particular multidimensional data, has led to the development of new concepts in computing and visualization that utilize special data compression techniques [101], as well as compressions schemes for regulated environments [102].

1.7.1. *Image Analysis*

Image quality assurance is an important control mechanism for image data to be analyzed quantitatively. Appropriate techniques include sampling [103], focus setting techniques [104], image quality measurement [105], and calibration routines [106]. Some of the undesired phenomena which influence or prevent successful application of image analysis require preprocessing. This includes attenuation of the laser beam in thick specimens, photobleaching or saturation of the fluorescence, autofluorescence, and background noise, and the distortions caused by the embedding media [92, 94, 107]. The loss of resolution in thick specimens is shown in Figure 18. The correction of such phenomena is often limited by the underlying theoretical models selected, since the effects are highly object dependent.

Undesired effects in fluorescence confocal microscopy can be corrected by ratio imaging. Ratioing (i.e., the compensation of two channels by digital division) can substantially reduce cross-talk effects in double detectors [72]. Environmental markers indicating the pH value or the level of free calcium ions (Ca^{2+}) also require ratio imaging to compensate for structural densities or section thickness [19].

Due to the anisotropic resolution of the data sets, various ways of improving axial resolution are available; *deconvolving methods* use the point spread function of the optics, which has to be measured first [108]. The algorithms can be differentiated further by their ability to incorporate noise. The classical filter is the Wiener filter; new developments include the iterative *maximum likelihood estimation* [109]. This method was extended with a blind deconvolution algorithm, using certain constraints to model the point spread function, which is a promising approach for biomedical users (see Fig. 21). However, calculation times are still in the range of 1 h for some hundred iterations. A different approach is the *nearest-neighbor deconvolution* developed for bright-field imaging [110], by taking into account the blur present in the planes above and below the focal plane. More recently, this method has been contrasted with a no-neighbor deconvolution method [111]. Besides directly improving data, isotropy can also be obtained by interpolating the lateral dimension [92]. Transformation of data from a cubic format into other lattices has also been proposed, based on mathematical morphology [112].

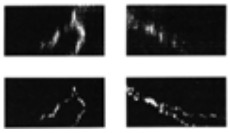

Figure 21. Projectional side views of rat hippocampal dendrites stained with Lucifer yellow

Top: Original confocal data set; Bottom: Restored data following 100 iterations of the maximum likelihood blind deconvolution algorithm

(Permission of T. HOLMES, Rensselaer Insitute, Troy, United States)

Sets in digital image processing to enhance, filter, segment, and identify structures are frequently applied to three-

dimensional microscopic data sets. This can be done either section by section or by using the corresponding three-dimensional extension of digital filters, such as the Gauss, Laplace, Sobel, and Median filter or noise removal algorithms [113-116]. Segmentation often requires interaction of the user, since automatic segmentation algorithms generally rely on the intensities of the voxels only. Once the binary images have been obtained, binary morphological operations using structuring elements for erosion and dilatation can be performed [117].

Digital measurements address the size and volume of structures, quantities, or forms and relation between volumetric subtleties. To measure volumes, all voxels that identify a certain structure are summed up. It is particularly difficult to measure the surface of structures, and most methods offer an estimation based only on stereological methods. Object counting in volumes usually requires the definition of guard boxes, which are moved throughout the volume. Objects falling into such boxes are evaluated by following certain stereological rules to avoid any bias. Stereological considerations are also used to estimate spatial statistics, such as numerical densities or nearest-neighbor distances; such methods have been used extensively in three-dimensional cytometry [132].

1.7.2. *3D Visualization*

Visualization of serial sections is the most frequently used computer technique in confocal microscopy, independent of specific applications [118-120]. Most commercial instruments have three-dimensional software implementations ready for the user. Three-dimensional objects are displayed on the computer screen by staking up the individual sections. Such techniques have been reviewed in several articles [121-124]. Besides surface rendering based on contours, volumetric representations that rely on the intensities of the voxels are the preferred techniques in confocal microscopy. A comparative assessment of both techniques is given in [125]. Reconstructions are presented three-dimensionally on a computer graphics screen by using either visual cues, animation, of stereoscopic displays.

One method used in *volume rendering* is to define imaginary rays through the volume (*ray-casting techniques*) [126, 127]. The way of casting is accomplished by the definition of certain projection geometry. Given a central projection, the angles of the rays are defined by the matrix of the screen and the center of projection. A common implementation traverses the data in a front-to-back order, which has the advantage that the algorithm can stop when a predefined property of the accumulation process has been reached. This class of methods is also referred to as *image order rendering*, since the matrix of the screen is used as the starting point for the accumulation process. Because of the projection geometry, such a ray-firing technique does not necessarily pierce the center of a voxel, and an interpolation has to be taken into account. To improve the output of ray- casting renderings, preprocessing of the voxel scenes may be performed as, for example, in the renditions of complex chromatin arrangements [116]. Modalities available in image order rendering include maximum projection (Fig. 22), integration, and surface mode (Fig. 18).

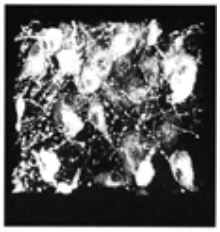

Figure 22. Volume rendering of a confocal sequence

The neural network of immunohistochemically stained cells, together with connecting fibers in the neural optical tract of hamsters is visualized by this projection

(Specimen courtesy of H. KORF, Zentrum für Morphologie, Frankfurt, Germany)

A much simpler geometry is realized in a parallel projection, such as in *object order rendering*. Here the volume is rotated according to the current viewing angle and then traversed in a back-to-front order [128]. During traversal, each voxel is projected to the appropriate screen pixel where the contributions are summed or blended, a situation that is referred to as compositing. During the accumulation process other properties of voxels besides the intensities may be considered, such as opacity, color, gradient values [129, 130], or the simulation of fluorescence processes (SFP) [131]. However, since volume rendering relies entirely on voxel attributes, it has some disadvantages if analytical representations are required. Therefore, elements of computer graphics are combined with volumetric renditions. Examples can be found in vector descriptions of volumetric data sets [132] or in geometric attributes expressing local properties and topologies that can be embedded in the three-dimensional reconstruction [122, 133].

1.8. High-Throughput Screening for Histopathology and Drug Development

The prevailing tissue diagnostic methods use histological staining protocols, which are the basis of most pathological assessments. Specimens prepared from tissue blocks or biopsies are embedded in paraffin, cut into 5–7 µm thick sections, placed on glass slides sized 20× 50 mm, and are stained with Hematoxilyn-Eosin (H&E). Complete imaging of such a glass slide at 20× microscopic magnification generates several hundred individual digital color images. Specialized robotic microscopes, some developed out of telepathology [134], allow scanning of a complete slide within minutes. These solutions have implemented a continuous stage movement. The information about focus levels to guide the auto-focusing mechanism is generated either prior to the scanning process by probing and interpolating randomly selected points on the slide, or by a parallel detection path that looks ahead and continuously determines focal-position information. These systems have been automated further by robotic slide–loading mechanisms that allow the scanning of hundreds of slides without any user interaction. The images are digitized, automatically assembled into montages, and saved to disk for automated analysis. These automated solutions can be used in conjunction with special preparation techniques such as tissue microarrays (TMA). The TMA is composed of hundreds of tissue cores on one slide. Such preparations can substantially increase the throughput of many investigations, specifically in those areas where tissues of many donors have to be compared, such as the testing of antibodies or identification of biomarkers [135, 136].

In drug discovery, advances in combinatorial chemistry and genomics have led to the development of high–throughput screening (HTS) for drug target development [137]. Conventional cell-based HTS assays rely on measuring single parameters from lysates of cell populations grown in multi-well high-density plates (whole-well average measurement), at

single time points (end-point format) with very limited temporal resolution and no spatial resolution. By using standard or more advanced plate readers it is possible to reach the speed requirement of HT primary campaigns for drug discovery. New screening approaches are referred to as “high–content screening (HCS)”. HCS relies on more sophisticated assays where the higher quality of the data being generated is believed both to increase the likelihood of discovery and to accelerate the profiling of successful drugs [138, 139]. A number of automated imaging platforms have been introduced to the market, and each instrument offers different solutions toward imaging and specimen handling, fluidics/kinetics capabilities, sample throughput, and data analysis to accomplish HCS. Some of these imaging workstations are confocal (based either on laser line or spinning disk) while some are not. Confocal imaging platforms support kinetics, such as calcium and sodium fluxes, mitochondria, and changes in plasma membrane potential. Also supported are end-point cell assays , e.g., cytotoxicity, apoptosis, cell cycle, translocations, using common commercially available fluorescent probes [140].

[Top of Page]

2. Electron Microscopy

Severin Amelinckx and Ludwig Reimer

2.1. Introductory Considerations [154]

There are two main classes of electron microscopy (EM) techniques. In the first class, the electron probe is a stationary beam incident along a fixed direction. This incident beam can be parallel [conventional transmission electron microscopy (CTEM), high-resolution transmission electron microscopy (HRTEM), high-voltage transmission electron microscopy (HVTEM), selected-area electron diffraction (SAED)] or convergent [convergent-beam electron diffraction (CBED), convergent-beam electron microscopy (CBEM)]. The resolution is determined by the quality of the imaging optics behind the specimen. Instruments implementing these techniques are conceptually related to classical light microscopes.

In the second class of methods, a fine electron probe is scanned across the specimen, and transmitted electron (TE) [scanning transmission electron microscopy (STEM)] or the desired excited signal such as secondary electrons (SE) or backscattered electrons (BSE) [scanning electron microscopy (SEM)], and/or AE [Auger Electron spectroscopy (AES)/scanning Auger microscopy (SAM)] is selected, detected, and the signal amplified and used to modulate the intensity of another electron beam which is scanned synchronously eith the first over the screen of a TV monitor. The stationary beam methods are based on image formation processes, whereas the scanning methods are essentially “mapping” techniques. Their resolution is mainly determined by the probe size, i.e., by its electron probe formation optics. The magnification is geometric; it is determined by the ratio of the areas scanned by the electron beam on the screen and synchronously by the electron probe on the specimen.

Analytical microscopes (AEM) have been developed which make it possible to implement a number of imaging modes, diffraction modes, and analytical modes in a single instrument and often also to apply them to a single specimen. The synergism of these tools has the potential to provide detailed structural and chemical information on the nanoscale. There is a trend of extending stationary beam probe analytical tools such as X-ray microanalysis, Auger electron spectroscopy (AES), and electron energy loss spectroscopy (EELS) into mapping methods which make it possible to image spatial distributions of chemical elements.

2.2. Conventional Transmission Electron Microscopy (CTEM)

2.2.1. *Introduction*

The images produced in transmission electron microscopy are essentially due to local diffraction phenomena; absorption contrast plays only a minor role. Not only is electron diffraction responsible for the image formation; it also allows establishment of the orientation relationship between direct space, as observed in the image, and reciprocal space as imaged in the diffraction pattern (see ⟶). Images and the corresponding diffraction patterns are equally important for a detailed interpretation; both should always be produced from the same selected area, with the specimen orientation unchanged. For this reason, most commercial microscopes have the ability to switch easily from the diffraction mode to the imaging mode and vice versa (see Section Operating Modes of the Electron Microscope).

2.2.2. *Scattering by Atoms: Atomic Scattering Factor*

Electrons are scattered by atoms as a result of Coulomb interaction with the nucleus and the electron cloud. The atomic scattering factor $f_e(\theta)$ thus contains two terms of opposite sign

$$f_e(\theta) = \frac{m e^2 \lambda}{2h^2}[z - f_x(\theta)] / \sin^2\theta \qquad (2.1)$$

where z is the atomic number; $f_x(\theta)$ the scattering factor for X-rays; m the electron mass; e the electron charge; λ the wavelength of the electrons; and h Planck's constant [155].

The first term clearly relates to the nucleus, whereas the second term is due to the electron cloud. The interaction with matter is stronger ($\times 10^4$) for electrons than for X-rays or neutrons, which interact only with the electron cloud or with the nucleus, respectively. As a result multiple scattering will not be negligible in electron diffraction experiments. Moreover, electron scattering is oriented mainly in the forward direction. Tables of $f_e(\theta)$ for different atoms are given in [156].

2.2.3. *Kinematic Diffraction by Crystals*

2.2.3.1. *Lattice, Reciprocal Lattice*

The amplitude diffracted by an assembly of atoms results from interference of the waves scattered by the atoms (i.e., the scattered amplitude in any given direction is obtained by summing the amplitudes scattered in that direction by the individual atoms and taking into account the phase differences due to path differences resulting from the geometry of the

assembly). For a physical and mathematical description of the important terms lattice and reciprocal lattice, see [157] and ⟶ .

2.2.3.2. *Geometry of Diffraction*

Bragg's Law [158]. Like other diffraction phenomena in periodic structures, electron diffraction can be described in either direct or reciprocal space. Bragg's diffraction condition describes diffraction in direct space. A diffracted beam will be formed whenever the incident beam encloses with the set of lattice planes an angle $\theta_{\mathbf{H}}$, such that the path difference 2 $d_{\mathbf{H}}$ sin $\theta_{\mathbf{H}}$ between waves diffracted by successive lattice planes is an integer *n* of wavelengths λ of the radiation used:

$$2d_{\mathbf{H}} \sin \theta_{\mathbf{H}} = n\lambda \tag{2.2}$$

where **H** denotes the Miller indices of the plane. In Figure 23 this path difference is indicated by a thicker line. For 100-kV electrons, λ = 4 pm, and as a result, the Bragg angles $\theta_{\mathbf{H}}$ are very small — a few degrees at most.

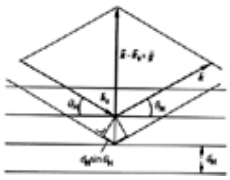

Figure 23. The derivation of Bragg's diffraction law in direct space

Ewald Construction [159]. Diffraction conditions can also be formulated in terms of reciprocal space as

$$\bar{\mathbf{k}} = \bar{\mathbf{k}}_0 + \bar{\mathbf{g}} \tag{2.3}$$

where $\bar{\mathbf{k}}_0$ is the wave vector of the incident electron beam and $\bar{\mathbf{k}}$ the wave vector of the diffracted beam, where $|\bar{\mathbf{k}}| = |\bar{\mathbf{k}}_0| = 1/\lambda$; $\bar{\mathbf{g}}$ is a reciprocal lattice vector. Equations (2.2) and (2.3) have the same physical content. They can be obtained by expressing the conservation of energy and of momentum of the incident electrons on scattering. More details on Ewald's construction are given in [160] and in ⟶ .

Diffraction by a Thin Foil. Since electrons are strongly “absorbed” in solids the specimen must be a thin foil (< 300 nm); otherwise no diffracted beams will be transmitted. The number of unit cells along the normal to the foil is thus finite, whereas along directions parallel to the foil plane, the number of unit cells can be considered infinite. In such specimens the diffraction conditions are relaxed, and diffraction also occurs for angles of incidence deviating somewhat from the Bragg angle. In reciprocal space this relaxation results in a transformation of the sharp reciprocal lattice nodes into thin rods (so-called relrods) perpendicular to the foil plane. The Ewald sphere intersects such a rod to produce a diffracted beam. The direction of this beam is obtained by joining the center of Ewald's sphere with this intersection point. The resulting diffraction pattern can to a good approximation be considered to be a planar section of the reciprocal lattice. Indexing of the diffraction spots is simple and unambiguous.

The distance $\bar{\mathbf{s}}_{\bar{\mathbf{g}}}$ by which the lattice node *G* is missed by Ewald's sphere is called the *excitation error.* It is a vector along the direction of the foil normal joining the reciprocal lattice node *G* to the intersection point with Ewald's sphere. It is positive when pointing in the sense of the incident beam and negative in the opposite case.

Column Approximation [161]. The small magnitude of the Bragg angles causes the electrons to propagate along narrow columns parallel to the beam direction. The intensity observed in a point at the exit face of the specimen is thus determined by the amplitude scattered by the material present in the column located at that point. In a perfect foil the intensity does not depend on the considered column, but when defects are present, it does.

Kinematic Rocking Curve. The amplitude of the beam scattered by a foil of thickness z_0 is obtained by summing the amplitudes scattered by the volume elements along a column perpendicular to the foil surface and taking into account the phase differences due to their depth position in the column. For a perfect foil the intensity of the scattered beam, I_s for reflection **H** is given by

$$I_s(s,\ z_0) = F_{\mathbf{H}}^2 \sin^2 \pi s z_0/(\pi s)^2 \tag{2.4}$$

where $F_{\mathbf{H}}$ is the structure factor [143] with indices **H**.

The dependence of I_s on *s* (i.e., on the direction of the incident beam and on the foil thickness z_0) is called the *rocking curve* (see Fig. 24). The separation of the zeros is $1/z_0$, and the full width of the central peak is $2/z_0$.

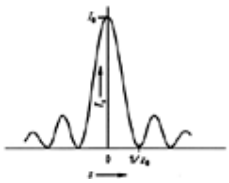

Figure 24. Plot of scattered intensity I_s versus excitation error, i.e., the “rocking curve” according to the kinematic approximation

Bent Contours [162]. For a curved foil, *s* varies along the specimen; the loci of constant *s*, imaged as bright and dark fringes, are called bent extinction contours. They form as shown schematically in Figure 25 for a cylindrically bent foil.

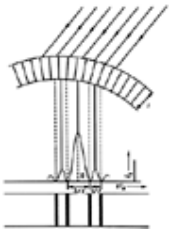

Figure 25. Formation of bent extinction contours by a cylindrically bent foil

Thickness Contours. The thickness dependence of I_s is periodic with period $1/s_{\bar{g}}$: the period thus diverges for $s_{\bar{g}} = 0$ (i.e., for the exact Bragg condition). The loci of equal thickness are also imaged as lines of equal intensity; they are called thickness extinction contours or wedge fringes.

Kikuchi Lines [163]. In rather thick specimens a second type of diffraction pattern becomes prominent, which consists mainly of lines rather than spots. These lines occur in parallel pairs; the line closest to the center of the diffraction pattern is darker, whereas the one further from the center is brighter than the background (Fig. 26). These lines are the intersections of Ewald's sphere with rather flat double cones having semiapex angles equal to 90° – θ_H, where θ_H is the Bragg angle associated with the pair of lines ($\theta_H \approx 1 – 2°$). The angular separation of the bright – dark line pair is 2 θ_H.

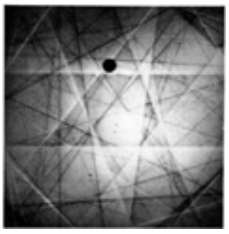

Figure 26. Kikuchi line pattern produced by a silicon foil

Note the parallel pairs of bright (excess) and dark (deficiency) lines. The deficiency lines are close to the center. The excess lines are separated in angle by 2 θ_H from the corresponding deficiency lines

If the Bragg condition for a set of lattice planes is satisfied exactly, the bright Kikuchi line associated with that set of planes will pass through the corresponding Bragg spot, whereas the dark one will pass through the origin. The separation of a Bragg spot from its corresponding Kikuchi line is a measure of the excitation error *s* of that reflection. As a result, the Kikuchi line pattern provides a means to measure *s* [164].

2.2.4. *Dynamic Diffraction by Crystals*

2.2.4.1. *General Considerations*

According to kinematic theory the intensity of diffraction spots is proportional to the square of the structure factor (Eq. 2.4). This simple proportionality is lost in reality because of multiple diffraction effects, which are taken into account adequately in dynamic theory [165].

Although kinematic theory does not predict the intensities of diffraction spots correctly, it is useful because the geometric features of diffraction are well described. For large *s* or very small foil thicknesses the relative spot intensities are qualitatively reproduced. For *s* = 0, kinematic theory breaks down since a diffracted beam with diffraction vector $\bar{g}_1$ may then acquire an amplitude comparable to that of the incident beam, and thus act as an incident beam and give rise to diffraction in the reverse sense (i.e., with diffraction vector $-\bar{g}_1$).

For most orientations of the incident beam, many diffracted beams are excited simultaneously and a section of reciprocal space is imaged. However by carefully orienting a sufficiently thick specimen the intensity of a single diffracted beam can be maximized: this is called a *two-beam case*. Such specimen orientations should be used, whenever possible, in the quantitative study of defects.

2.2.4.2. *Basic Equations*

The two-beam dynamic theory describes the interplay between transmitted and diffracted beams by means of a set of coupled differential equations of the form [144, 166].

$$\mathrm{d}T/\mathrm{d}z = (\pi i/t_{-g})S \exp 2\pi i s z \tag{2.5a}$$

$$\mathrm{d}S/\mathrm{d}z = (\pi i/t_g)T \exp -2\pi i s z \tag{2.5b}$$

where *T* and *S* are the complex amplitudes of the transmitted and scattered beams, respectively; *z* measures the depth in the foil. The exponential factors take into account the growing phase shift due to the excitation error, with increasing distance *z* behind the entrance face. The factor *i* takes into account the phase jump on scattering; t_g and t_{-g} are called the extinction distances. The extinction distance, which has the dimension of length, is a measure of the strength of the reflection $\bar{g}$. For low-order reflections of elemental metals it is of the order of a few tens of nanometers, but for weak superstructure reflections it may be several hundred nanometers. Whether a foil is “thick” or “thin” depends on the number of extinction distances in the foil thickness. Kinematic theory is valid only for foils with a thickness that is a fraction of an

extinction distance.

Absorption can phenomenologically be accounted for by assuming the extinction distances to be complex [167]. Formally $1/t_{\bar{g}}$ is replaced by $1/t_{\bar{g}} + i/\tau_g$, where τ_g is called the absorption length. Empirically, it is found that τ_g ranges from $5\,t_g$ to $15\,t_g$. The substitution can be made, either in the Equation set (2.5) or directly in the solution of this set.

2.2.4.3. *Dynamic Rocking Curve*

By ignoring absorption and taking into account the fact that $T = 1$ and $S = 0$ at $z = 0$ (entrance face), integration of Equation set (2.5) gives

$$T(s,z) = [\cos \pi\sigma z - i(s/\sigma)\sin \pi\sigma z] \exp \pi i s z \tag{2.6a}$$

$$S(s,z) = (i/\sigma t_g) \sin \pi\sigma z \exp \pi i s z \tag{2.6b}$$

and thus

$$I_s = SS^* = (1/\sigma t_g)^2 \sin^2 \pi\sigma z \tag{2.7}$$

$$I_T = 1 - I_s \tag{2.8}$$

where

$$1/\sigma = t_g/(1 + s^2 t_g^2)^{1/2} \tag{2.9}$$

For $s = 0$ this reduces to $\sigma = 1/t_g$ (i.e., the depth period is now $t_{\bar{g}}$ and for large s: $\sigma = s$). The divergence of kinematic theory for $s = 0$ is now removed. The depth variation of the transmitted and scattered beams, leading to thickness extinction contours, is represented schematically in Figure 27 for $s = 0$ as well as for $s \neq 0$.

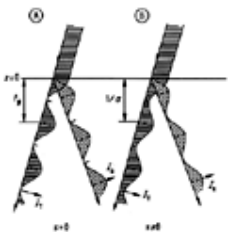

Figure 27. Interplay between incident and diffracted beam according to the dynamic theory

A) Bragg's condition is satisfied exactly, the depth period is t_g; I_S and I_T pass periodically through zero; B) $s \neq 0$ the depth period is now $1/\sigma_g$; I_T varies periodically but does not pass through zero; I_S and I_T are complementary since absorption was neglected

2.2.4.4. *Anomalous Absorption, Bormann Effect [167, 168]*

Taking absorption into account gives more complex rocking curves of the type shown in Figure 28 for a foil of thickness $t = 3\,t_g$ and $\tau_g = 10\,t_g$. The curve for the diffracted beam (b) is symmetric, whereas the curve representing the transmitted beam (a) is asymmetric in s. For the same absolute value of s, the transmitted intensity is larger for $s > 0$ than for $s < 0$. This asymmetry is known as the Bormann effect; it was demonstrated initially for X-ray diffraction under dynamic conditions [168]. The steep change of I_T with s in the vicinity of $s = 0$ causes a large sensitivity of I_T to small changes in orientation.

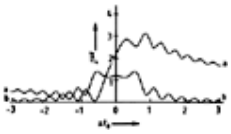

Figure 28. Rocking curves for transmitted (a) and scattered (b) beams according to the two-beam dynamic theory, with anomalous absorption taken into account

2.2.4.5. *Lattice Fringes [169]*

The Electron Wave Function at the Exit Face. The amplitude of the transmitted wave for a unit incident wave can be represented by $T \exp 2\pi i \bar{K} \cdot \bar{r}$, where $\bar{K}$ is the wave vector of the incident wave inside the crystal (i.e., corrected for refraction at the vacuum – crystal interface). Similarly the diffracted wave can be represented by $S \exp 2\pi i(\bar{K} + \bar{g}) \cdot \bar{r}$, where $\bar{g}$ is the diffraction vector (i.e., the reciprocal lattice vector corresponding to the excited reflection). The wave function of the electrons emerging at the exit face of the foil is then

$$\psi = (T + S \exp 2\pi i \bar{g} \cdot \bar{r}) \exp 2\pi i \bar{K} \cdot \bar{r} \tag{2.10}$$

The observed intensity

$$I_\psi \equiv \psi\psi^* = I_T + I_S + 2\sqrt{I_T I_S} \sin (2\pi g x + \varphi) \tag{2.11}$$

with

$$\tan \varphi = (s/\sigma) \tan \pi\sigma t; \tag{2.12}$$

x is measured along the exit surface in the direction of $\bar{g}$: further, the following abbreviations are used: $I_T \equiv T\,T^*$; $I_S \equiv S\,S^*$. This expression represents sinusoidal fringes with a period $1/|\bar{g}|$; they can be considered as forming an image of the lattice planes $\bar{g}$.

2.2.4.6. *Faulted Crystals*

Planar Interfaces. Planar interfaces are imaged as fringes parallel to the foil surface; their characteristics depend on their geometric features. *Translation interfaces* (stacking faults, out-of-phase boundaries, etc.) separate two crystal parts related by a parallel translation $\bar{R}$ (Fig. 29 A), and *domain boundaries* separate domains that differ slightly in orientation (i.e., for which the excitation errors are different $\Delta s = s_1 - s_2$; (Fig. 29 B) [170].

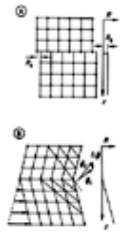

Figure 29. Schematic of two different types of interfaces

A) Translation interfaces with constant displacement vector $\bar{R}$; B) Domain boundary between two regions for which the excitation error is slightly different; the displacement R increases with increasing distance from the interface (the homologous diffraction vectors $\bar{g}_1$ and $\bar{g}_2$ are slightly different: $\Delta\bar{g} = \bar{g}_2 - \bar{g}_1$)

Diffraction Equations for Faulted Crystals. The displacement over R_0 (R_0 = constant) of the exit part of a column with respect to the entrance part (Fig. 29 A), the two parts being separated by a fault plane, can be accounted for by noting that a supplementary phase change $\alpha = 2\pi\bar{g}\cdot\bar{R}$ occurs on diffraction but not on transmission (see Fig. 30). The path difference Δ between diffracted waves 1 and 2 is $\Delta = 2\,R_0 \sin\theta_H$, where θ_H is the Bragg angle and hence $2\,d_H \sin\theta_H = \lambda$ with $d_H = 1/|\bar{g}|$. This gives for the phase change $\alpha = (2\pi/\lambda)\,\Delta = 2\pi\,\boldsymbol{g}\,R_0$, where R_0 is in general the component of $\bar{R}$ along $\bar{g}$ (i.e., $gR_0 = \bar{g}\cdot\bar{R}$).

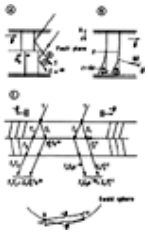

Figure 30. Diffraction by a faulted crystal

A) Effect of a translation interface with vector $\bar{R}_0$; B) Effect of a displacement field $\bar{R}(\bar{r})$; C) Schematic of the expressions for transmitted and scattered beam amplitude for a foil containing a general interface parallel to the foil surfaces

The equations describing diffraction by the displaced exit part are obtained by the substitution $S \longrightarrow S \exp -i\alpha$ in Equations (2.5). The amplitudes T and S emerging from a column intersecting such an interface are then obtained by integrating the Equation (2.5) set along a column down to the level of the interface and subsequently further down to the exit face, using the adapted form of Equation (2.5)(Fig. 30).

The effect on T and S of the presence of a defect described by a displacement field $\bar{R}(\bar{r})$ (for instance due to a dislocation) (Fig. 30) can be taken into account in Equations (2.5) by replacing the constant s by an effective local value s_{eff} which now depends on z, i.e., on the depth along the columns, and which is given by $s_{eff} = s + \bar{g}(d\bar{R}/dz)$. This can be shown rigorously but it can also be understood intuitively by the following geometrical considerations (Fig. 30 B).

Let the diffraction vector of the considered set of lattice planes be $\bar{g}$ at the level z below the entrance face. At level $z + dz$ the displacement $\bar{R}$ has increased by $d\bar{R}$ and as a result the local vector $\bar{g}$ is rotated over a small angle $\delta\theta \simeq \tan\theta = dR/dz$. The change in s corresponding to this rotation of $\bar{g}$ over $\delta\theta$ is given by $\delta s = g\,\delta\theta$, i.e., by $\boldsymbol{g}\,(dR/dz)$ or somewhat more general by $d(\bar{g}\cdot\bar{R})/dz$, expressing that only displacements which change the orientation of the vector $\bar{g}$ are operative.

In the particularly simple case of a domain boundary as modeled in Figure 29 B, $R = kz$ (k = constant) in the exit part. Hence, at the level of the domain boundary, s changes abruptly into $s + k$ along the integration column.

Diffraction by a Crystal Containing a Planar Interface [171]. The amplitudes of the transmitted and scattered beams for a foil containing a planar interface, with displacement vector $\bar{R}$ (Fig. 30 B), can be formulated as

$$T(s_1, s_2, z_1, z_2, \alpha) = T_1(s_1, z_1)\,T_2(s_2, z_2) + S_1(s_1, z_1)\,S_2(-s_2, z_2)\,\exp\,i\alpha \tag{2.13}$$

$$S(s_1, s_2, z_1, z_2, \alpha) = T_1(s_1, z_1)\,S_2(s_2, z_2)\cdot\exp\,(-i\alpha) + S_1(s_1, z_1)\,T_2(-s_2, z_2) \tag{2.14}$$

with $\alpha = 2\pi\bar{g}\cdot\bar{R}$. These relations can also be derived from the adapted set of Equation (2.5).

The first equation states that the amplitude of the transmitted beam T results from the interference between the doubly transmitted beam (i.e., the beam transmitted by the first part T_1 and subsequently transmitted by the second part: $T_1 \times T_2$)

and the doubly scattered beam $S_1 \times S_2^{(-)} \exp i\ \alpha$. The factor $\exp i\ \alpha$ takes into account the phase shift on the amplitude S_2 scattered by the second part, resulting from the translation of the second part of the foil.

A similar interpretation can be given to the second expression. The amplitudes $T_1(s_1, z_1)$, $T_2^{(-)}$, $S_1^{(-)}$, $S_2^{(-)}$ are given by Equations (2.6 a) and (2.6 b). The indices 1 and 2 refer to the entrance and exit parts, respectively: the minus superscript means that in the corresponding Equations (2.6 a) and (2.6 b), s must be replaced by $-s$, because diffraction occurs from the negative side of the set of lattice planes (i.e., the diffraction vector is $-\bar{g}$).

If $s_1 = s_2$ but $\alpha \neq 0$, Equations (2.13) and (2.14) refer to a pure translation interface. If $s_1 \neq s_2$ and $\alpha = 0$, the expressions are applicable to a domain boundary. For mixed interfaces, $s_1 \neq s_2$ as well as $\alpha \neq 0$. If the interfaces are parallel to the foil surfaces, transmitted and scattered intensities are constant over the specimen area, but they vary in a pseudoperiodic fashion with the thicknesses z_1 and z_2 of entrance and exit parts, with $z_1 + z_2 = z_0$ (total thickness). The expressions describing this variation explicitly are obtained by substituting Equations (2.6 a) and (2.6 b) into (2.13) and (2.14).

If the interfaces are inclined with respect to the foil surfaces this depth variation is displayed as a pattern of fringes parallel to the foil surfaces in the area where the two wedge-shaped crystal parts overlap.

The fringe profiles of α-fringes ($s_1 = s_2$; $\alpha \neq 0$) and δ-fringes ($\Delta s = s_1 - s_2 \neq 0$; $\alpha = 0$) in bright field (BF) and dark field (DF), exhibit different properties. The symmetry properties allow the fringe patterns due to the two kinds of interfaces to be distinguished. The properties of α-fringes with $\alpha = \pi$ are singular [172].

2.2.4.7. *Moiré Patterns [173-175].*

The superposition of two identical crystal foils having a small orientation difference θ (rotation moiré) or of two parallel crystal foils with a small difference in lattice parameters (parallel moiré) gives rise to a fringe pattern, which can to a good approximation be considered as the coincidence pattern of the lattice fringes corresponding to the diffraction vectors $\bar{g}_1$ and $\bar{g}_2$ active in the two crystal foils.

The fringes are parallel to the average direction $\bar{g}$ of the two diffraction vectors $\bar{g}_1$ and $\bar{g}_2$ in the case of rotation moiré patterns and perpendicular to the common direction of the two diffraction vectors in the case of a parallel moiré. From the moiré spacing Δ, small differences can be deduced in the lattice parameters d_1 and d_2 of the two components of the sandwich since $\Delta = d_1 d_2/(d_1 - d_2)$, or the small orientation difference $\theta = |\Delta\bar{g}|/|\bar{g}|$ in the case of a rotation moiré. Figure 31 shows an optical analogue for the two types of moiré fringes produced by the superposition of two line patterns.

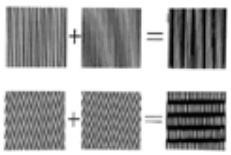

Figure 31. Optical analogue for the formation of moiré fringes in a sandwich crystal

(Upper) The two crystals are parallel, but the lattice parameters are slightly different; (Lower) The two crystal parts exhibit a slight orientation difference

2.2.5. ***Operating Modes of the Electron Microscope [144, 146, 147]***

2.2.5.1. *Microscope Optics*

The ray paths in an electron microscope are shown in Figure 32, according to the geometrical optics approximation, for the two main operating modes: high-magnification, high-resolution imaging, and selected area diffraction.

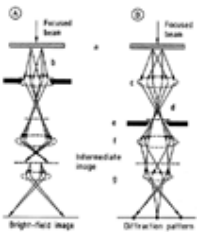

Figure 32. Schematic of the beam path in an electron microscope according to the geometrical optics approximation

A) High-magnification, high-resolution mode; B) Selected area diffraction mode

a) Specimen; b) Objective aperture; c) Objective lens; d) Field-limiting aperture; e) Gaussian image plane; f) Intermediate lens; g) Projector lens

The microscope is essentially a three-lens system: an objective lens (c), an intermediate lens (f), and a projector lens (g). Each lens may be a composite lens. A movable selector aperture (b) is present in the image plane of the objective lens; a second aperture (d) is placed at the objective lens, close to the back focal plane. With the first aperture a small area (≤1 µm) of the image (i.e., of the specimen) is selected, with the second aperture a single beam is selected or a number of the image-forming diffracted beams.

The characteristics of the objective lens are crucial because they determine to a large extent the image resolution and the contrast. The other two lenses mainly provide the desired magnification.

Although magnetic lenses normally rotate the image about the optical axis, in microscopes designed at the end of the 1980s, these rotations are compensated by a suitable device, and the image and the diffraction pattern have the same orientation.

2.2.5.2. *High-Resolution, High-Magnification Mode*

In the first (Fig. 32) imaging ray path the beam produced by a source (tungsten filament, LaB_6, field emission gun; see also Section Electron Guns) and collimated by a condenser lens system, is scattered by the object, and an image is formed in the image plane of the objective lens. By means of the selector aperture, an area of the specimen is selected and magnified by the intermediate lens, which is focused on the image plane of the objective lens and provides a magnified image in the image plane of the intermediate lens. This image serves as the object for the projector lens, which forms the final image either on a fluorescent screen, on a photographic plate, or on an image intensifier followed by a TV camera. In the latter case — which is virtually a requirement for successful high-resolution work — the image can be viewed on a TV monitor and the final adjustments performed before the image is recorded photographically. Electronic recording on tape is of course possible directly via the TV signal; this mode is used mainly for so-called in situ dynamic studies.

2.2.5.3. *Diffraction Mode (Figure 32 B)*

When operating in the diffraction mode the intermediate lens (f) is weakened (i.e., its focal length is enlarged) so as to cause the back focal plane of the objective lens (c) to coincide with the object plane of the projector lens (g). This produces a magnified image of the diffraction pattern. In doing so, the selected area is not changed since only the intermediate lens current is changed; the diffraction pattern is thus representative of the selected area. The area selected in the diffraction mode is much larger than the field of view under high-resolution conditions.

As the electron beam passes through the specimen, various interactions occur next to diffraction: characteristic X-rays are produced, the electrons suffer energy losses, etc. These effects can be used for analytical applications (see Section Analytical Methods).

2.2.6. ***Selected-Area Electron Diffraction (SAED)***

ED methods (SAED, CBED) are a useful complement to X-ray diffraction, especially when the material is available only as small microcrystals. Then only powder diffractometry can be applied when using X-rays. However, even fine powders usually contain single crystals of a sufficient size to produce a single-crystal ED pattern that allows the approximate (but unambiguous) determination of the lattice parameters. These approximate lattice parameters in turn allow unambiguous indexing of the powder diffraction pattern and subsequently make possible a precise lattice parameter measurement. In addition, ED patterns often exhibit weak reflections due to superstructures, which are not visible in X-ray powder diffraction.

The spatial geometry of diffuse scattering can be reconstructed more easily from ED patterns than from X-ray diffraction patterns, since ED patterns are planar sections of reciprocal space, which is not the case for X-rays.

2.2.7. ***Diffraction Contrast Images [144, 164]***

2.2.7.1. *Imaging Modes*

Diffraction contrast images are usually obtained under two-beam conditions. By means of the aperture placed close to the back focal plane of the objective lens, either the transmitted or the diffracted beam is selected and the corresponding diffraction spot is highly magnified (Fig. 33). The image obtained in this way is a map of the intensity distribution in the selected diffraction spot. The image made in the transmitted beam is a *bright-field image*, whereas the image made in the diffracted beam is a *dark-field image*. Since the electrons are confined to narrow columns, the latter can be considered the picture elements of this image. The amplitude of the transmitted (or scattered beam) associated with a point at the back surface of the specimen is obtained by summing the contributions of the volume elements along a column parallel to the incident beam centered on that point. The amplitude depends on how the excitation error *s* changes along the column, which in turn varies with changes in the orientation of the diffracting planes along the column. Strain fields cause such orientation changes and can thus be imaged in diffraction contrast. Diffraction contrast images do not reveal the crystal structure but are very sensitive to the presence of strain fields caused by defects such as dislocations.

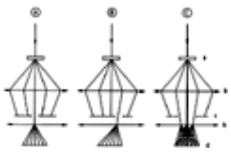

Figure 33. Schematic representation of different imaging modes

A) Bright-field diffraction contrast mode; B) Dark-field diffraction contrast mode; C) High-resolution imaging mode

a) Specimen; b) Lens; c) Aperture; d) Screen

2.2.7.2. *Dislocation Contrast [141, 144]*

Consider a foil (Fig. 34) containing an edge dislocation in E, with an orientation such that in the perfect parts of the foil the amplitude of the incident beam is approximately equally divided between the transmitted and the diffracted beams. To the left of the dislocation the considered lattice planes are locally inclined in such a way as to better satisfy Bragg's condition (i.e., *s* is smaller) than in the perfect parts; more electrons are diffracted than in the perfect part. A lack of intensity will then be noted left of the dislocation in the transmitted beam, leading to a dark line in the bright-field image. At the right of the dislocation, Bragg's law is less well satisfied than in the perfect part. This leads to a lack of intensity in the scattered beam and thus to a dark line in the dark-field image. The image is on one side of the dislocation: the side on which the dark line occurs in the bright-field image is called the *image side*. According to this model the image side is where the lattice planes are tilted into the Bragg orientation.

Figure 34. Schematic representation of the image formation at an edge dislocation in a thin foil

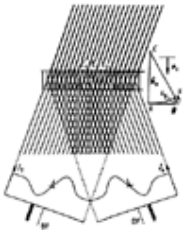

The line thickness is a measure of the beam intensity. The image profiles for bright-field (BF) and dark-field (DF) images are represented schematically along with diffraction conditions in the perfect part of the foil

The conditions for kinematic diffraction [176] are best approximated in the *weak-beam* method, which consists of making a dark-field image in a weakly excited diffraction spot. The dislocation image then consists of a narrow bright line on a darker background.

2.2.7.3. *Extinction Conditions for Defects*

Diffraction by a family of lattice planes that remain undeformed by the presence of the dislocation will not produce an image of the dislocation [144, 177].

Quantitative images based on dynamic theory can be computer simulated by using the methods developed in [178]. A comparison of simulated and observed images allows determination of all the relevant parameters of various dislocation types (Fig. 35). More details are given in [179].

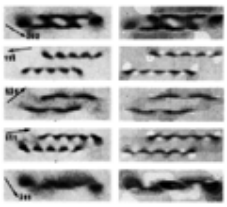

Figure 35. Examples of computed images (right) and the corresponding observed images (left) of the same dislocation by using different diffraction vectors [178]

2.2.7.4. *Domain Textures*

Phase transitions in crystals are usually accompanied by changes in symmetry, the space group of the low-temperature phase being a subgroup of the space group of the high-temperature phase. On cooling, the crystal breaks up into *orientation variants* (or twins) related by the lost symmetry elements of the high-temperature phase and *translation variants* (out-of-phase boundaries) related by the lattice translations lost on forming the superlattice [180]. The composite diffraction pattern of a domain texture is the superposition of the diffraction patterns due to individual domains. The diffraction pattern taken across a *reflection twin interface* contains one row of unsplit spots, which is perpendicular to the mirror plane. All other diffraction spots are split along a direction parallel to the unsplit row of spots [143].

In diffraction contrast images made in a cluster of *split diffraction spots*, orientation domains exhibit differences in brightness (i.e., domain contrast), especially for conditions close to $s = 0$ where the intensity variation with s is steep (Fig. 36). The presence of translation variants is not reflected in the diffraction pattern, except by some diffuse scattering.

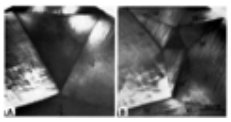

Figure 36. Domain texture in β-lead orthovanadate resulting from the $\gamma \longrightarrow \beta$ transformation

A) The central triangular area still consists of the high-temperature γ-phase, whereas the other domains are in the low temperature β-phase; B) The same area as A) after additional cooling. The unstable triangular γ-area has been reduced further (Courtesy of C. Manolikas)

Images made in *unsplit reflections* do not reveal brightness differences due to orientation differences of the lattice, but they may exhibit domain contrast due to differences in the structure factor. This is the case for Dauphiné twins in quartz, which are related by a 180° rotation about the threefold axis (Fig. 37). The two domains produce the same diffraction pattern since their lattices are common, but certain reflections have structure factors that differ in magnitude [181]. A dark-field image made in such a reflection will exhibit a different brightness in the two domains, which is called the structure factor contrast [144].

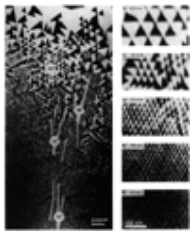

Figure 37. Dauphiné twin lattice in α-quartz close to the phase transition into the high-temperature β-phase viewed along the threefold axis

A temperature gradient is established across the specimen area. The α-phase is broken up in regular arrays of columnar domains whose size decreases with increasing temperature [181]

2.2.7.5. *Interface Contrast and Domain Contrast*

Interface contrast (i.e., an α-fringe pattern) is produced in the projected area of the interface even if the structure factors in the two domains have the same magnitude, but differ in phase by α; no domain contrast is produced under these conditions. Interface contrast is also produced at *orientation domain boundaries*; δ-fringes are produced in the projected area, since the *s*-values on both sides of the boundary are different in general [182]. Extinction of the δ-fringe pattern occurs if the *s*-values are the same for the two domains.

The presence or absence of domain contrast across a fringe pattern helps to distinguish the two types of interface. If the two domains on either side of the interface exhibit the same brightness for all reflections, the fringes must be due to a translation interface; they must exhibit α-character. Domains exhibiting a difference in brightness are separated by a domain boundary and the fringes are of the δ-type.

2.2.7.6. *Strain Field Contrast*

The strain field associated with precipitate particles in a matrix whose lattice parameters differ slightly from those of the precipitate can also be revealed as regions of different brightness [183]. For example, a spherical particle in an elastically isotropic matrix produces a spherically symmetric strain field. In certain parts of this strain field the vector $\bar{R}$ is locally perpendicular to $\bar{g}$; such areas will show up with the background brightness. In other areas, $\bar{g}$ is parallel to $\bar{R}$; such areas will show up with a brightness that is different from the background, either lighter or darker.

2.2.8. *Convergent Beam Diffraction [148, 184]*

2.2.8.1. *Geometry of Convergent Beam Patterns*

In the early 1980s, a class of applications based on the use of a convergent beam of electrons (convergent beam diffraction, CBD) was developed. Whereas parallel beams allow high spatial resolution, *high angular resolution* is more easily achievable with convergent beams. For these applications the electrons are incident along directions within a cone of revolution or along the surface of a cone (hollow cone method) having its apex in (or close to) the sample plane. Under these conditions, electrons are incident on the specimen under all possible directions within a certain angular range, which should be of the order of the Bragg angles of the lowest-order reflections of the material under study. In practice, the semiapex angle α varies from 2 to 10 mrad. Also, the corresponding diffracted beams for each reciprocal lattice node form cones. The intersections of these cones with the photographic plate are circular disks whose radii are proportional to the semiapex angle of the cone of incident directions (Fig. 38). If the specimen is perfectly flat and defect free, these disks are images of the angular intensity distribution in each particular reflection.

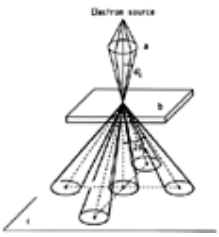

Figure 38. Production of a convergent beam electron diffraction pattern

a) Probe forming optics; b) Specimen; c) Film

The incident beam is convergent. Diffraction disks are centered on each diffraction spot.

2.2.8.2. *Point Group Determination [185, 186]*

When the incident beam is parallel to a zone axis, the symmetry of the intensity distribution in the complete disk pattern, including the fine lines in the 000 disk, reflects the point symmetry of the crystal along this zone axis (Fig. 39).

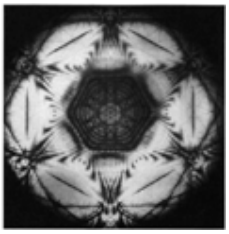

Figure 39. Zone axis convergent beam electron diffraction pattern revealing the presence of a sixfold rotation axis [184]

Several techniques can be used to determine the point group [185]. The method developed by BUXTON et al. [186] is based on the use of zone axis patterns and of dark-field diffraction patterns obtained under exact two-beam conditions. By choosing the appropriate optical conditions, these disks do not overlap and the exact Bragg spots remain in the centers of the corresponding disks. The optics required for the application of these methods are described in [187]. With increasing cone angle the disks start to overlap and the nature of the pattern changes, but the symmetry elements are conserved.

A disk in the CBED pattern usually consists of broad fringes due to the dynamic interaction between beams belonging to the zero-order Laue zone (ZOLZ). Finer lines are formed by the interaction of high-order Laue zone reflections (HOLZ) with the zero-order beams (Fig. 40). The geometry of the HOLZ lines is determined by the accelerating potential and the lattice parameters; comparison of computer-simulated and observed patterns allows determination of one parameter if the other is known [187, 188].

Figure 40. Deficiency HOLZ (high-order Laue zone) lines in the central disk of a convergent beam diffraction pattern of

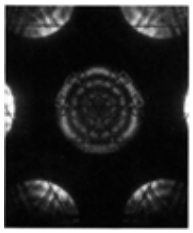

silicon [184]

2.2.8.3. *Space Group Determination*

Space group determination is based on the observation of the Gjønnes – Moodie lines [189] in certain “forbidden” diffraction disks. Reflections that are kinematically forbidden may appear as a result of double diffraction under multibeam dynamic diffraction conditions. If such forbidden spots are produced along pairs of different symmetry-related diffraction paths that are equally excited, the interfering beams may be exactly in antiphase for certain angles of incidence if the structure factors have opposite signs. Since the convergent beam disks are formed by beams with continuously varying directions of incidence, this condition will always be satisfied somewhere in the disk. The locus of points for which the phase difference is exactly π is a dark fringe along a diameter of the disk for which the Bragg condition is satisfied exactly (Gjønnes – Moodie lines, Fig. 41).

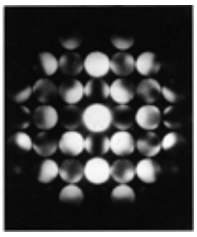

Figure 41. Convergent beam electron diffraction pattern exhibiting Gjønnes – Moodie lines due to the presence of glide mirror planes and twofold screw axis [184]

2.2.8.4. *Foil Thickness Determination*

From the dark-field two-beam rocking curve (Section Geometry of Convergent Beam Patterns) observed in the disk corresponding to the diffracted beam, the specimen thickness can be deduced with high accuracy provided the accelerating voltage and the interplanar spacing of the excited reflection are known [190].

2.2.9. ***High-Resolution Electron Microscopy***

In recent years, high-resolution electron microscopy has become an important tool in the study of complicated crystal structures and their defects. Whereas detailed considerations of electron optics are of only marginal importance in the case of diffraction contrast images, they become essential for a discussion of the atomic resolution structure images produced by crystals.

2.2.9.1. *Image Formation in an Ideal Microscope (Fig. 42)*

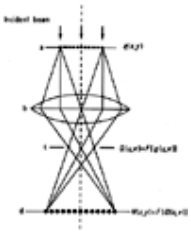

Figure 42. Image formation in an ideal microscope

a) Specimen; b) Objective lens; c) Back-focal plane, objective aperture; d) Image plane

The diffraction pattern Q (u, v) is the Fourier transform of the object q (x, y); the image is the inverse Fourier transform Ψ (x, y) of the diffraction pattern

Let the crystalline object be a thin foil characterized by a two-dimensional transmission function q (x, y) that describes at each point of the exit surface of the specimen the amplitude and phase of the electron beams emerging from the column situated at (x, y) after dynamic diffraction in the foil. The diffraction pattern can be described to a good approximation as the Fourier transform Q (u, v) of the object function q (x, y). This diffraction pattern acts in turn as a source of Huyghens wavelets, which interfere to form the image, after linear magnification by the optical lens systems; the image is, in turn, the Fourier transform Ψ (x, y) of the diffraction pattern.

2.2.9.2. *Image Formation in a Real Microscope*

In real microscopes the situation is complicated by the presence of finite apertures and lens aberrations. The apertures truncate the Fourier transform by admitting only a finite number of beams to the image formation process. Lens aberrations depend on the angle β that a diffracted beam encloses with the optical axis of the microscope, hence they introduce angle-dependent phase shifts between diffracted beams. Moreover, they cause “blurring” of the image, a point of the object plane being represented as a disk in the image plane and vice versa.

Spherical Aberration. Spherical aberration causes electron beams enclosing an angle β with the optical axis to suffer a phase shift χ_S with respect to the central beam:

$$\chi_S = \frac{1}{2}\pi\, C_S \beta^4/\lambda \tag{2.15}$$

where C_S is the spherical aberration constant.

Defocus. Under exact Gaussian focusing conditions the contrast produced by an object that changes only the phase of the incident electron wave (i.e., a so-called phase grating) is minimal. This is the case for thin crystalline foils. Visual contrast improves when working in underfocused conditions. The phenomenon is in a sense similar to phase contrast in optical microscopy of phase objects. Defocusing causes a relative phase shift of the diffracted beams enclosing different angles with the central beam. The phase shift caused by a defocus distance ε with respect to the Gaussian focus is given by

$$\chi_D = \pi\, e\, \beta^2/\lambda \tag{2.16}$$

The total phase shift caused by spherical aberration and defocusing is thus:

$$\chi(\beta) = \left(\frac{1}{2}\pi\, C_S \beta^4 + \pi\, e\, \beta^2\right)/\lambda \tag{2.17}$$

$\chi(\beta)$ depends on β in the way shown in Figure 43, i.e., there is a flat minimum at $\beta_{min} = (-\varepsilon/C_S)^{1/2}$ and a zero at $\beta = (-2\ \varepsilon/C_S)^{1/2}$.

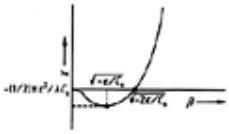

Figure 43. Dependence on β of the phase shift $\chi(\beta)$ of a beam enclosing an angle β with the optical axis of the microscope

Phase Grating Approximation [191]. The effect of this phase shift on the image parameter can be understood most easily by discussing the case of a pure phase grating, which is a reasonable model for a very thin crystalline foil. Such a specimen causes a phase shift χ of the electron wave on passing through the foil because the wavelength of the electron is different in vacuum and in the crystal. This phase shift in the point (x, y) can be written as $\sigma\, \Phi(x, y)$ where Φ is the projected potential integrated over the sample thickness along the propagation direction of the electrons, and $\sigma = \pi/\lambda E$ (λ = wavelength of electron in vacuum; E = accelerating voltage). Ignoring absorption, the object function for a thin foil can be written as

$$q(x, y) = \exp\, i\sigma\varphi(x, y) \simeq 1 + i\sigma\varphi(x, y) \tag{2.18}$$

Taking also into account the phase shift (Eq. 2.17) caused by the lens system on the diffracted beams, and neglecting absorption it can be shown by Fourier transformation [192] that the image intensity is finally given by

$$I(x, y) \simeq 1 - 2\sigma\varphi(x, y) \tag{2.19}$$

whereby it was assumed that $\sin\chi = -1$ ($\cos\chi = 0$). The image contrast, defined as $(I - I_0)/I_0 = -2\ \sigma\ \Phi(x, y)$, where I_0 is the incident intensity, is thus directly proportional to the projected potential $\Phi(x, y)$ provided $\sin\chi = -1$.

Provided $\sin\chi = -1$, columns of large projected potential are imaged as dark dots. If $\sin\chi = +1$, the same columns would be imaged as bright dots.

Optimum Imaging Conditions. The most stringent requirement is $\sin\chi = -1$; this condition can only be met approximately and only in a limited β interval. The condition $\sin\chi = +1$ can be satisfied only for a limited number of discrete β-values.

The situation is somewhat comparable to positive and negative phase contrast in optical microscopy. The lens imperfections have been used to introduce a phase shift of $\pi/2$ in the same way as the quarter wavelength ring in the optical microscope. Only beams passing through the "window" in which the condition $\sin\chi = -1$ is met, interfere with the required phase relationship that causes the image to represent maxima in projected potential as dark areas.

From Equation (2.17) it is clear that $\chi(\beta)$ depends on the parameters C_S, ε, and λ. The condition $\sin\chi = -1$ can thus be met by adjusting one or several of these parameters. For a given instrument C_S and λ are fixed and the observer can meet this condition by optimizing the defocus ϵ. From Equation (2.17) it follows that χ adopts the essentially negative value $\chi = -\pi\ \epsilon^2/(2\ \lambda\ C_S)$ for $\beta = (-2\ \epsilon/C_S)^{1/2}$. $\sin\chi$ will be -1 if $\pi\ \epsilon^2/(2\ \lambda C_S) = \pi/2$, i.e., for

$$e_S = -(\lambda\, C_S)^{1/2} \tag{2.20}$$

For the defocus $\epsilon = \epsilon_S$ the curve $\sin\chi$ versus β will exhibit a rather flat part in the region around

$$\beta = (-e/C_S)^{1/2} \tag{2.21}$$

where $\sin\chi = -1$. The underfocus value ϵ_S corresponding to this optimum is called Scherzer defocus [193]. Most high

resolution images are made under this defocus condition and in the thinnest part of the sample.

2.2.9.3. *Resolution Limiting Factors*

Aperture. The presence of an aperture admitting only beams that enclose an angle with the optical axis not exceeding β_A imposes a resolution limit, called the Abbe limit [194]. A point in object space is imaged as a circle with a radius

$$\varrho_A = 0.61\lambda/\beta_A$$

Chromatic Aberration. Instabilities ΔE in the high voltage E of the microscope cause a spread in the wavelength of the incident electron, which blurs the image. Variation ΔI in the lens current I similarly causes a spread in the focal length of the lenses, which also contributes to such blurring. Furthermore, the inelastic scattering of electrons in the specimen causes slight changes in the wavelength of electrons emerging from the specimen. The net result of these different phenomena can be described as an effective spread in the focal length of the objective lens:

$$\Delta f = C_C\left[(\Delta E/E)^2 + 4(\Delta I/I)^2\right]^{1/2} \tag{2.22}$$

where C_C is the chromatic aberration constant; the corresponding disk of least confusion in object space is $\varrho_C = \beta\ \Delta f$. In concrete cases, $\Delta E/E \leq 10^{-6}$; $\Delta F \approx 10$ mm, $C_C \approx 10$ mm.

Astigmatism. Ideal lenses should have cylindrical symmetry. However, in reality, slight deviations may occur, leading to a dependence of the focal length on the azimuth of the ray path considered. Astigmatism can occur for various reasons, including inhomogeneities in the polepiece material, asymmetry in the windings, or a dirty aperture. As a result of astigmatism a point source is imaged as two different line foci at right angles to each other. At exact focus the image deformation disappears but the image is still blurred. Astigmatism can be corrected by a stigmator — a device (usually an octopole) that makes the lens appear perfectly cylindrical — by applying weak additional magnetic fields. Careful correction of astigmatism is essential for the production of high-quality, high-resolution images.

Beam Divergence. The intense illumination required for high-resolution imaging is obtained by a condenser lens system that produces a slightly conical beam with a typical apex angle of the order of 10^{-3} rad. This also leads to some image blurring since the final image is the superposition of images corresponding to the different directions of incidence.

Mechanical Instability. The obtainable resolution depends not only on the lens characteristics but also to a large extent on the mechanical stability of the microscope, which should be installed on a vibration-free heavy concrete block. In many cases the mechanical stability is actually the resolution-determining factor. The specimen holder should moreover be creep free, which is a problem for hot and cold stages. High-resolution work is therefore performed almost exclusively at room temperature.

Ultimate Resolution. The final disk of confusion due to the different nonmechanical blurring effects is given approximately by

$$\varrho = (\varrho_A^2 + \varrho_S^2 + \varrho_C^2)^{1/2} \tag{2.23}$$

which depends on β. At low angles the aperture effect is usually dominant, whereas at high angles the spherical aberration is the limiting factor (for 100-kV microscopes, $C_S \approx 8.2$ mm and $C_C \approx 3.9$ mm). The optimum radius of least confusion is achieved for a certain β-value and the radius of confusion corresponding to that optimum is

$$\varrho_0 = 0.9\,\lambda^{3/4} C_S^{1/2} \tag{2.24}$$

Therefore, a gain in resolution can be achieved by reducing C_S and by reducing λ (i.e., using a higher voltage). At present, many microscopes used in material science and in solid-state chemistry operate at 200 – 400 kV.

At higher voltages, radiation damage to the specimen severely limits observation time. Advanced designs aim at reducing C_S to allow a decrease of the high voltage for a given resolution. Practical resolution limits are at present ca. 0.16 nm, but new developments tend to peak this value to below 0.10 nm.

Directly interpretable high-resolution images (i.e., having a direct relationship to either the projected lattice potential or the electron density) are formed by the interference between beams passing through the “window” in the image transfer function (ITF). As a result, only low-order reflections (i.e., low-order Fourier components of the lattice potential) generally contribute to the image, which limits the detail that can be resolved (irrespective of the point resolution) since fine details are carried by the high-order Fourier components. However, information transmitted beyond the window (i.e., by the rapidly oscillating part of the image transfer function) can also be used. This information can be extracted by computational techniques, for instance, by using as input a set of images made at a series of closely spaced defocus values [195].

2.2.9.4. *Image Formation Models [196]*

Fourier Model. High-resolution image formation is considered most conveniently as occurring in two steps. In the first step, electrons propagate through the crystal and produce a two-dimensional periodic electron distribution at the exit face, which images the projected lattice potential or the projected electron density. In a second step, this exit face acts as a planar distribution of point sources of spherical electron waves, which interfere behind the foil and form the diffracted beams that move in the lens system of the microscope. Interference of these beams produces the final image, which will be directly interpretable only if the beams interfere with the correct phases.

The resolution of the image depends on the spatial frequencies of the highest-order beams admitted by the selector aperture that still contribute to the image with the correct phase. Only beams passing through the window in the ITF contribute to a directly interpretable image.

The image will exhibit more detail as the order of included reflections increases, but it will give a faithful representation of the structure only if the Fourier components have correct relative phases; this puts a practical limit on the number of useful reflections.

Complete software packages for the simulation of high-resolution images are available commercially (see, e.g., [197]).

Channeling Model. An alternative model that emphasizes the particle nature of electrons provides a simple intuitive picture [198]. An atomic column parallel to the incident beam is a cylindrically symmetrical potential well, which acts on the moving incident electrons as a succession of alternating convergent and divergent lenses. At the entrance face of the foil the incident electron distribution is uniform, but in the first part of the foil the electrons are attracted toward the core of the atom columns and focusing occurs, transforming part of the potential energy of the electrons into kinetic energy. Subsequently the electrons repel one another and defocusing occurs. The focusing – defocusing motion occurs periodically as a function of the depth in the foil, the depth period being a function of the atomic number of the atoms in the column. The heavier the atoms are, the smaller is the depth period. As a result, for a foil of given thickness, one type of atom column will give rise to a relative maximum in the electron distribution, whereas another type may give rise to a relative minimum. These extremes are not necessarily imaged as extremes of the same nature. A maximum may be imaged either as a bright dot or as a dark dot, depending on the defocus and on the foil thickness (see Fig. 44). This model explains why different atom columns produce dots of different brightness. The focusing and defocusing behavior was confirmed by computer simulations, which also showed that the depth periods are in the 4 – 10-nm range depending on the atomic number.

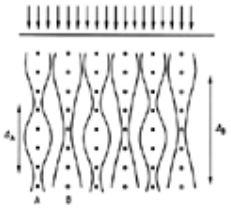

Figure 44. Channeling of electrons along columns of atoms parallel to the incident beam

Alternating focusing and defocusing occurs. The focusing distances Δ_A and Δ_B are different for A and B columns

2.2.9.5. *Image Interpretation [196]*

Trial and Error. Digital simulation programs for HRTEM are usually based on the Fourier approach. They allow one to compute the image of a given structure projected along a given zone axis as a function of the foil thickness and of the imaging conditions (defocus) for a microscope with known characteristics (spherical aberration coefficient, accelerating voltage, beam convergence) (Fig. 45).

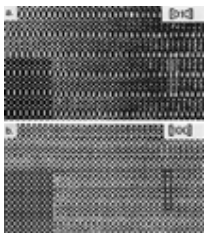

Figure 45. High-resolution image of the high-temperature superconductor $Y_2Ba_4Cu_7O_{15-\delta}$. The bright dots reveal the heavy atom columns as projected along two different zones. Simulated images are reproduced as insets. The projected unit cell is outlined (Courtesy G. Van Tendeloo). The images were made at optimum defocus. The two images refer to different specimens, with comparable thickness.

Identification of a structure or of a defect in a structure proceeds by “trial and error”. A model is proposed, and a matrix of images that varies the two main independent variables (foil thickness and defocus) is computed.

These theoretical images are then compared with the observed images at the different thicknesses along a wedge-shaped part of the foil. A solution is considered acceptable when the correspondence between observed and computed images is judged visually to be “good enough” for all available thicknesses at a constant defocus. This comparison allows one at the same time to estimate the foil thickness and the defocus; these quantities are usually not known a priori. Numerical criteria to quantify the goodness of fit have been proposed but have seldom been applied as yet. In recent developments, high-resolution images have been used to determine the chemical composition along individual columns in rather special circumstances, such as along the interfaces in synthetic layer structures of semiconductors grown by molecular beam epitaxy. The composition profile across an interface can be obtained at an atomic scale. The application of these methods requires a certain amount of a priori knowledge concerning the structure.

Direct Retrieval. “Direct retrieval” use as the input a series of images taken at closely spaced defocus values (focus variation method). If the microscope parameters describing the transfer function and its inverse are known, the projected wavefunction at the exit plane of the foil can be reconstructed. From this, the projected structure (the object) can be retrieved by using an analytical formulation of the channeling model. Knowing the projected structure along more than one zone axis allows one to reconstruct the three-dimensional structure. Less a priori knowledge is required than for the methods based on “trial an error”.

2.2.10. ***Scanning Transmission Microscopy [199]***

Transmission electron microscopy can also be performed with a scanning incident beam (STEM). In a STEM instrument a

fine convergent electron probe, formed by demagnifying a small, but brilliant, electron source, is scanned over the specimen area of interest. The incident beam is focused on the specimen plane and a convergent electron diffraction pattern is formed in a plane behind the specimen. Parts of this diffraction pattern (CBED pattern) can be selected by an appropriate aperture, and the signal detected hereby gives rise to an electronic signal which is displayed on a TV monitor, the scan of which is operated synchronously with the probe scan. A bright field (BF) image is obtained when the directly transmitted beam is selected. If one or several beams outside the central beam are selected, a dark field (DF) image is produced.

High resolution is achieved by making the effective probe size as small as possible. For this purpose use is made of a field emission gun with an effective source diameter on the order of 5 nm; demagnification then allows a probe size on the order of 1 nm in diameter to be obtained. The beam current is of the order of 0.5 nA.

The essential electron optics of STEM and CTEM instruments are in principle not very different, even though the electrons travel in opposite senses in the two instruments. A comparison of the principles of STEM and CTEM is represented schematically in Figure 46. In CTEM the image signal is produced in parallel (i.e., simultaneously) in all parts of the image plane (e.g., a photographic plate), whereas in STEM the signal is generated in serial form, as a time-dependent electric signal, which makes it easy to apply on-line image processing. Like in CTEM various signals, excited by the passage of the electron beam through the specimen, can be displayed by using the appropriate detectors.

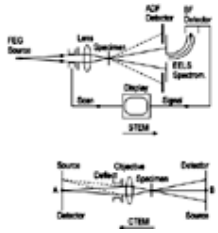

Figure 46. Scanning transmission electron microscopy

Top: Schematic representation of the essential components of a scanning transmission electron microscope (ADF: annular dark field; BF: bright field; FEG: field image gun)

Bottom: Illustrating the reciprocity relationship between the beampaths in STEM and CTEM microscopes; the electrons move in opposite senses (courtesy of J. Cowley)

2.2.11. *Z-Contrast Images [200]*

The commonly used imaging modes (CTEM, HRTEM) are based on interference and diffraction and rely strongly on coherently scattered electrons. However, simultaneously with the coherent scattering, incoherently scattered electrons also are produced by thermal diffuse scattering and in particular by Rutherford scattering. In the STEM mode these incoherently scattered electrons can be used to image atom columns, provided that the proper electron optics are available. A fine electron probe is obtained by focusing the incident convergent electron beam in the sample and scanning over the foil. Thereby, a significant fraction of the incoherently scattered electrons, emerging from the sample at relatively large scattering angles, is captured in an annular aperture and detected.

in the case of incoherent imaging conditions, the contrast transfer function is a monotonously decreasing function of the spatial frequency, whereas under coherent imaging conditions it is a rapidly oscillating function. This has important consequences. In the coherent case the brightness of a dot imaging a given atom column may change from bright to dark and vice versa as a function of defocus. In contrast, in the incoherent case, the relative brightness of a dot remains consistently the same, independent of defocus, i.e., there is no contrast reversal. Moreover, the dot brightness increases monotonously with increasing average *Z* value of the atoms in the column as a consequence of the contribution of the Rutherford scattering to the incoherently scattered electrons. Incoherent images can therefore be interpreted on an intuitive basis even for relatively large thickness. It can be shown that structure retrieval is in principle possible and simpler than in the case of coherent imaging. The method is therefore well suited to the study of geometric defects in crystals.

2.2.12. *Analytical Methods*

2.2.12.1. *X-Ray Microanalysis [150]*

Principle. An important feature of electron microscopy is the possibility of performing in the same instrument both a qualitative and a quantitative elemental analysis of the same small crystal fragment that produced the electron diffraction pattern and the image. The most commonly used method consists in principle of detecting the characteristic X-rays excited by the high-energy electrons passing through the specimen.

For this purpose, an X-ray spectrometer, mounted sideways on the microscope column at the level of the specimen holder, captures the X-rays emitted in all directions by the specimen within a certain solid angle centered around the chosen takeoff direction and transmitted through an X-ray transparent window (Fig. 47). The spectrum consists of the characteristic X-ray lines superposed on a continuous background of bremsstrahlung (see also ⟶ Surface and Thin-Film Analysis – Ultraviolet Photoelectron Spectroscopy (UPS), ⟶ Surface and Thin-Film Analysis – Auger Electron Spectroscopy (AES)).

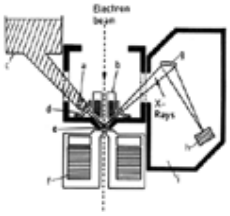

Figure 47. Specimen – detector configuration used in X-ray microanalysis

Left: Energy-dispersive X-ray system (EDX) with Si(Li) detector; Right: Wavelength-dispersive spectrometer (WDS) with bent crystal

a) Si(Li) detector; b) Mini lens; c) Liquid-nitrogen reservoir; d) X-ray window; e) Specimen; f) Objective lens; g) Crystal; h) Detector; i) Crystal spectrometer

The physical basis of X-ray microanalysis is Moseley's law [201], which establishes a direct and systematic relation

between the atomic number Z and the characteristic X-ray wavelength λ in a family of X-ray lines (e.g., associated with similar transitions in different elements):

$$\lambda = A/(Z - C)^2 \tag{2.25}$$

A and C depend on the considered family; K_α, K_β, . . . , L_α, . . . , X-ray lines are produced when an inner shell (e.g., K-shell) electron has been ejected from the atom by absorbing an amount of energy from an incident electron, exceeding the critical ionization energy associated with that shell. On deexcitation, this vacant level is filled by an electron originating from a higher-lying shell or subshell of the same atom. The transition is accompanied either by the emission of X-rays, which is the phenomenon of interest here, or by the ejection of an Auger electron (⟶ Surface and Thin-Film Analysis – X-Ray Photoelectron Spectroscopy (XPS)). Equation (2.25) shows that the atomic number can in principle be determined by measuring the characteristic X-ray wavelength.

Wavelength-Dispersive Spectrometry (WDS). Two types of X-ray spectrometers are in use: wavelength-dispersive spectrometers and energy-dispersive spectrometers.

In wavelength-dispersive spectrometers, dispersion is achieved by an analyzing crystal. Since the Bragg angle for a given set of lattice planes depends on the wavelength of the incident radiation, X-rays of different wavelengths can be separated spatially by the crystal, and a spectrum can thus be analyzed by changing the orientation of the analyzing crystal by θ and detecting the amount of radiation passing through a slit whose angular position is rotated simultaneously over 2θ. Often the analyzing crystal, which acts as a grating, is cylindrically curved so that focusing along a line occurs. The detector, usually a gas proportional counter, is placed in the focus and connected to a single channel analyzer, followed by standard counting electronics.

Depending on the wavelength range to be detected, different analyzing crystals are necessary; provision is usually made on the equipment to allow switching in several crystals from a built-in carousel. Extension of this method to long-wavelength X-rays is hampered by the absence of suitable crystals with large interplanar spacings. Synthetic layered structures, prepared by molecular beam epitaxy, and Langmuir – Blodgett layers have been used for this purpose since the end of the 1980s.

Energy-Dispersive Spectrometer (EDS). In energy-dispersive spectrometers, spectral analysis is based on measuring the X-ray energy rather than the wavelength. As a detector, a solid-state ionization chamber is used (e.g., a $p - i - n$ junction in silicon, an intrinsic germanium crystal, or a mercuric iodide crystal). The energy resolution is of the order of 150 eV or less for X-ray energies in the range 1 – 12 keV (Fig. 48). The silicon $p - i - n$ junction is produced by diffusing lithium (at elevated temperature) into a p-type silicon slab under a reverse bias, creating an intrinsic region via the exact compensation of fixed acceptors by mobile lithium donors, as a result of electromigration of the lithium ions. Under reverse bias an electric field is created in the broad intrinsic region, which acts as the sensitive volume. Such detectors must be kept permanently at liquid-nitrogen temperature since lithium ions migrate at room temperature. When an X-ray photon passes through the intrinsic region, its energy is dissipated almost entirely by the formation of electron – hole pairs, which are swept away by the reverse bias field and produce a voltage pulse across the detector crystal that is proportional to the energy of the incoming X-ray photon. The X-ray spectrum is then established by analysis of the pulse heights with a multichannel analyzer. For qualitative analysis, only the line positions are important, but for quantitative analysis the line shape and the peak heights are also used. Fundamental for the interpretation of EDS spectra is the assumption that all the energy of the X-ray photon is deposited in the detector (silicon escape peaks are ignored). The number of charges n created by a photon with energy E is then given by $n = E/E_i$, where $E_i = 3.6$ eV for silicon. Since the charge pulses to be detected are small, reducing the noise of the detector is essential; this is a second reason for cooling the detector. EDS has many advantages over WDS, the main ones being the much shorter acquisition time and the smaller minimum useful probe size. On the other hand, complications also occur that do not arise in WDS instruments; the major ones are escape peaks, pulse pileups, electron beam scattering, peak overlap, and window absorption effects. In practice, the latter prevents detection of light elements with $Z \leq 10$ (beryllium window) unless a windowless detector is used.

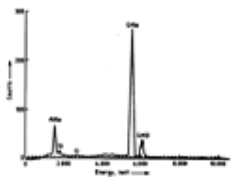

Figure 48. Example of X-ray spectrum obtained by means of an energy dispersive spectrometer

The material is a chromium aluminum alloy.

Quantitative Analysis. Quantitative analysis is based on the assumption that the intensity of the Lorentzian X-ray peak is proportional to the number of atoms of the species responsible for the peak. Since the theoretical relations are complicated and involve a large number of parameters, some of which are difficult to evaluate, well-characterized homogeneous standards are generally used for calibration, the composition of which has been determined by classical chemical macroscopic methods. These standards, which should also have similar shapes to the sample, are then measured in EDS equipment under the same well-defined and reproducible conditions as the unknown sample, and the spectra are compared. This calibration leads to the so-called k_i factors (coefficients of proportionality) for the different elements [202]:

$$C_{i_{\text{unknown}}}/C_{i_{\text{standard}}} = I_{i_{\text{unknown}}}/I_{i_{\text{standard}}} = k_i \tag{2.26}$$

where C_i are weight fractions.

Since k_i may depend on the particular matrices, Equation (2.26) must be corrected as follows [150, 203]:

$$C_{i_{\text{unknown}}} / C_{i_{\text{standard}}} = Z_i \times A_i \times F_i \qquad (2.27)$$

where the matrix correction factors Z_i>, A_i, and F_i still depend on the element under consideration; Z_i is the factor due to atomic number; A_i is due to X-ray absorption; and F_i to X-ray fluorescence. Other correction factors are due to the specimen geometry because X-ray production is depth dependent since it is proportional to the local electron energy. In the thin foils used in TEM the *ZAF* correction factors can often be neglected. X-ray microanalysis allows quantities of material of the order of 10^{-11} g to be analyzed; the concentration can be as low as 0.01 %, and the precision is of the order of 1 %.

For a more detailed discussion of the method and its technicalities, see [202-204].

Special Effects [205]. For foil orientations close to an exact Bragg orientation, the transmitted and scattered wave fields both propagate along the lattice planes. In a model crystal with a primitive structure, one wavefield has its maximum amplitude along planes coinciding with the atomic planes; the other has its maximum along the planes midway between atomic planes. For $s > 0$, the latter wavefield is excited, and therefore easy transmission occurs since the electrons in this wavefield avoid the atom cores (Bormann effect). For $s < 0$, the former wavefield is excited, but strongly absorbed, since the electrons pass close to the atom cores and can excite X-ray emission. X-ray emission is thus expected to be enhanced on the $s < 0$ side of a bent contour ($s = 0$).

In a crystal whose superstructure consists of alternating A and B atomic layers parallel to the considered lattice planes, which are close to the exact Bragg position, one type of wavefield is peaked along the planes of A atoms, whereas the other is peaked along the B-atom planes. Depending on the sign of *s*, one of the wavefields is enhanced relative to the other; hence the A and B atomic layers will be excited unequally. Under such conditions, X-ray microanalysis may produce erroneous results for the A/B ratio. Therefore, foil orientations should be avoided that cause an extinction contour to pass through the selected area due to anomalous transmission. On the other hand, this effect has been exploited to locate the site occupied by a third atomic species, in an AB superstructure, using a technique called ALCHEMI (atom location by channeling-enhanced microanalysis) [205].

2.2.12.2. *Electron Energy Loss Spectrometry (EELS) (* ⟶ Surface and Thin-Film Analysis – Electron Energy Loss Methods*) [206]*

Principles [150, 207]. Apart from elastic scattering, which is responsible for electron diffraction, inelastic scattering events also occur as electrons pass through the foil. Inelastic processes can be caused by (1) single electron excitations, such as X-ray and Auger electron production; and (2) collective excitations, such as volume and surface plasma oscillations (plasmons) and phonons. Collective excitation can be revealed indirectly as characteristic energy losses of the incident electrons. The energy of elementary excitations (plasmons and phonons) is low compared to the incident electron energy (100 – 140 kV).

Plasmons have energies in the range of 10 – 20 eV, whereas phonon energies are of the order of 10 meV. Individual quantized plasmons are thus much easier to detect than phonons. Since the positions of the plasmon loss peaks are characteristic of the materials, they can be used as analytical tools for aluminum and magnesium, for example.

However, for chemical analysis, the absorption edges in EELS curves are more important (Fig. 49); they reflect the absorption phenomena leading to X-ray production and exhibit a fine structure on the high-energy side. This fine structure is referred to as EXELFS (extended energy loss fine structure) [208], which is the analogue of EXAFS (extended X-ray absorption fine structure). The steep rise on the low-energy side of the absorption edge is due to the excitation of inner-shell electrons and characterizes the element.

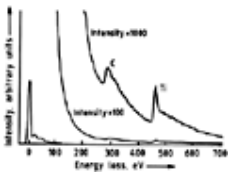

Figure 49. Example of EELS (electron energy loss spectrometer) spectrum of TiC

On the low-energy side the plasmon peak is visible. The fine structure on the high-energy end is caused by interference of waves reflected by the neighboring atoms

The fine structure is produced by the electron wave originating from an inner shell that is partially back-reflected by the surrounding atoms, which leads to a modulation of the excitation probability of inner-shell electrons. This fine structure can therefore provide information not only on the chemical nature of the absorbing atom, but also on the number of nearest neighbors and their distance.

As an analytical tool, EELS is in a sense complementary to X-ray microanalysis and makes use of the same type of phenomena. EELS works well for light elements, whereas X-ray microanalysis is much more difficult to apply to elements with $Z \leq 11$.

Spectrometers. The simplest system uses a magnetic prism to produce spatial separation of electrons with different energies. The spectrum is scanned by means of a slit, and a phosphor coupled to a photomultiplier is used as the detection system.

One of the most advanced types of equipment is the ω electron spectrometer [209]. It consists of four magnetic prisms that bend the incident beam back to its original direction; it can be mounted on an electron microscope column behind the specimen where inelastic scattering is produced. Slits allow a certain energy range to be selected from the spectrum. The spectrometer can thus also act as a filter and can produce, e.g., zero-loss images in which the blurring due to inelastic

scattering is eliminated.

Since the 1980s *X-ray microanalysis* based on energy-dispersive spectrometers has become a routine technique. Equipment as well as data processing hardware and software are available commercially as optional accessories for most microscopes. In the late 1990s the range of applications of EELS spectroscopy was considerably expanded and extended to nanoscale samples. The commercial availability of user-friendly equipment compatible with most transmission electron microscopes has greatly stimulated its use. Also the improvements in the interpretation methods and the development of the corresponding computer software greatly contributed to the expansion of its use. It is now possible to map the spatial distribution of selected features of the information contained in the EELS spectrum. In this way spatial distribution of different elements (also light elements) can be mapped with nanometer resolution, but for instance also the spatial distribution of different ionization states of the same element can be visualized. These advances were made possible by the use of arrays of diodes that allow parallel detection instead of the initially used serial detection.

2.2.13. *Specimen Preparation*

The objective of all specimen preparation methods is to obtain a crystal fragment that is sufficiently thin to transmit electrons with energies of 100 keV or more, and has a sufficiently large lateral size. The numerous methods available depend strongly on the type of material and observation method.

2.2.13.1. *Diffraction Contrast Specimens*

Specimens should have a thickness of a few hundred nanometers. If too thick, the foil is not transparent; if too thin, achieving a two-beam orientation is difficult; moreover, the sample is not representative of bulk material for a number of applications.

Electrically conducting bulk samples (metals and alloys) can generally be thinned by electropolishing; *bulk ceramic samples* are often chemically thinned and polished. The resulting sample is usually perforated in the center; the edges of the hole are then wedge shaped and have transparent edges.

Insulating thin foils may become charged and deflect the electron beam; this can be avoided by coating the specimen with a thin layer of amorphous carbon, which is electrically conducting.

Thin films of uniform thickness can be obtained by *vapor deposition* under vacuum or by *chemical deposition* techniques; such samples usually have microstructures that are different from those of samples prepared from bulk material.

Layered crystals exhibiting a pronounced cleavage plane, such as graphite or MoS_2, can be thinned by repeated cleavage to produce rather uniformly thin specimens.

A truly universal thinning method does not yet exist. However, *ion beam bombardment under grazing incidence* is applicable to many materials, in particular to brittle ones such as most semiconductors. In the first step, these materials are mechanically polished to produce slabs with dimples; subsequent ion beam thinning under grazing incidence then produces a hole whose edges are wedge shaped. Thinning can be achieved either from one side only or from both sides of the slabs; this is important, for instance, for quantum wells grown by molecular beam epitaxy. Surface damage left after ion bombardment can often be removed by a final chemical polish. Figure 50 schematically represents thinning methods used for semiconductor specimens.

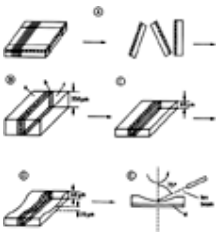

Figure 50. Preparation of a cross section of integrated circuits

A) Specimen is cut into thin slabs; B) Slabs are glued on a glass slide; C) Composite specimen is thinned mechanically; D) Thin specimen is dimpled; E) Final thinning is done by ion bombardment from one side while the specimen is rotated

a) Integrated circuit; b) Glue line; c) and d) Glass slide

(Courtesy of IMEC)

2.2.13.2. *High-Resolution Specimens*

The various methods described above can also of course be used to prepare specimens for use in high-resolution imaging. However much thinner specimens (1 – 10 nm) are required than for diffraction contrast. Nevertheless, more rudimentary methods can often be used successfully, since much smaller lateral dimensions of the specimen can be tolerated. Specimens of brittle materials (e.g., oxide superconductors) are in many cases obtained simply by crushing the material in a quartz mortar under a protecting organic liquid. In such a way the crystal fragments are dispersed in a suitable liquid, and a drop of the suspension is deposited on a grid covered with a holey carbon film and allowed to dry. The edges of fragments protruding over the holes are potentially suitable specimens, but actually only a small fraction of the crystal fragments are sufficiently thin to produce high-quality images. The thickness of the specimen is often the resolution-limiting parameter. Materials that deform plastically when crushed at room temperature often behave as brittle when crushed while immersed in liquid nitrogen.

2.2.14. *Applications to Specific Materials and Problems*

The number and diversity of applications of different forms of electron microscopy have become so large that an exhaustive survey, even when restricted to chemical aspects, is almost impossible. In particular, its application to structural problems in solid-state chemistry has evolved explosively in recent years. The examples given in what follows attempt to show as large a diversity as possible.

2.2.14.1. *Crystal Structures*

Only in simple cases can an “ab initio” structural determination be made by electron microscopy and diffraction, essentially because in electron diffraction the intensity of the reflections is not simply proportional to the square of the structure amplitude, but also depends in a complicated manner on the orientation and the thickness of the sample.

However, whereas X-ray and neutron diffraction provide only average structures (⟶), electron diffraction and microscopy can provide detailed information on local structures, especially when the average structure is known from other techniques. Also planar defects (i.e., local deviations from the average structure) can be studied to suggest possible new structures in which such defects occur periodically.

Since electron microscopy requires minute samples, the crystallography of materials that are available only as small particles (e.g., catalysts) can be studied in direct and reciprocal space. The early crystallographic studies on C_{60} and C_{70} crystals were for instance, performed by electron diffraction [210]. Even fine powders or fine-grained ceramics contain single crystal particles large enough to produce a single crystal electron diffraction pattern; therefore, the lattice parameters and the unit cell can be determined easily, albeit with moderate precision. Subsequently, X-ray powder diffraction patterns can be indexed unambiguously and the lattice parameters determined with high precision. Electron diffraction and especially high-resolution electron microscopy are complementary techniques to X-ray and neutron diffraction.

Imaging Modes for Various Materials.

Polytypes, Mixed-Layer Polytypes. The structures of polytypes are usually based on the various stacking modes of closely packed layers of spheres. The stacking sequence can best be revealed by imaging the structure along the zone parallel to the direction of the close-packed rows of atoms. Figure 51 shows a foil of SiC – 15 R imaged along the close-packed rows by using three different modes; it illustrates clearly the effect on the image detail of increasing the number of beams participating in the image-forming process. In Figure 47 A, only two successive spots out of the densely populated row of 000l spots are included; only the superspacing consisting of five-layered SiC lamellae is revealed as a line image. The isolated stacking faults are imaged as anomalously wide line spacings.

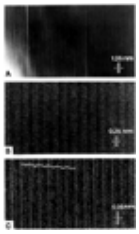

Figure 51. Imaging of the 15 R polytype of SiC along the closely packed rows of atoms

A) Two successive spots of the 0001 row contribute to the image; only the five-layer lamellae are imaged as lines; singular line spacings represent stacking faults; B) Cluster of five successive spots of the 0001 row contributes to the image; single SiC layers are revealed; each lamellae of A) is seen to contain five layers; C) Spots from parallel rows are allowed to contribute to the image; the stacking sequence ABCAC can be deduced from the two-dimensional image since close-packed atom columns are imaged as dots

Including in the aperture, a cluster of at least five successive spots of the central row between the 0000 and 00015 reflections, but excluding parallel rows, reveals the individual SiC layers and the superperiod due to the five-layered lamellae as lines; it shows in particular that the superperiod contains five layers (Fig. 47).

Finally, when spots from neighboring dense rows are also collected, a two-dimensional dot pattern is obtained and the close-packed rows are imaged as bright dots; the stacking sequence ABCAC . . . is directly revealed (Fig. 47).

The same imaging mode is used for the study of mixed-layer compounds based on the epitaxy of close-packed layers [211].

Ordered Alloys. The diffraction patterns of ordered binary alloys, derived from the face-centered cubic (fcc) structure, consist of strong basic spots that would also be produced by the disordered fcc lattice, and of weaker superlattice spots characteristic of the superstructure. The first-order superlattice spots correspond to larger interplanar spacings and are thus closer to the origin than the basic spots. Therefore images can be produced by selecting superstructure spots next to the direct beam only and excluding basic spots. The image so obtained (*bright-field superlattice image*) reveals only the configuration of the minority atoms as bright dots, since these atoms determine the superlattice. This information is sufficient to determine completely the projection of the superstructure, provided the basic lattice is known.

An image can also be formed exclusively by means of the superstructure spots situated within a reciprocal unit cell of the basic lattice (*dark-field superlattice image*) [212]. This last mode produces the best-contrasted images of the minority atom sublattice (Fig. 52).

Figure 52. Dark-field superlattice image of Au_4Mn

The manganese atoms are represented as bright dots. Note the presence of two orientation variants and several translation variants [212, 214]

If basic and superlattice spots are included in the image-forming process the basic lattice is also revealed; minority and majority atoms are imaged as dots with a different brightness. Such images contain information that is redundant if only the superstructure is sought. Examples of superlattice images of ordered alloys obtained by using the bright-field superlattice mode are reproduced in Figure 53.

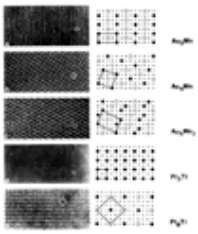

Figure 53. Various superstructures in ordered binary alloys imaged using the bright-field superlattice mode and compared to structural models

Long-Period Interface Modulated Structures. Many long-period structures are derived from simpler structures by the periodic introduction of planar translation interfaces. Typical examples are the long-period antiphase boundary structures in many ordering alloy systems such as Au – Mn, Cu – Au, and the shear structures derived from a simple transition-metal oxide structure such as WO_{3-x}, TiO_{2-x}. Such compounds produce characteristic diffraction patterns. In combination with their high-resolution images, their structures can usually be elucidated completely, including local deviations from the average structure, which often occur in such phases.

In a model [213], the superstructure is assumed to consist of identical slabs of basic structure with thickness *D* limited by planes normal to $\bar{e}_n$. Usually the slab thickness is equal to an integer number *n* of unit cell parameters *a* of the basic structure i.e., *D* = *na*, but this need not be the case. Successive slabs are separated by planar interfaces (stacking faults, anti-phase boundaries, discommensuration walls, etc.) with a displacement vector $\bar{R}$ and a unit normal $\bar{e}_n$.

The diffraction pattern of such a structure consists of clusters of equally spaced superstructure spots (spacing: 1/*D*), called satellites, located around the positions of the basic spots. The intensity of the satellites decreases rapidly with separation of the satellite from the basic spot to which it belongs, which also determines its position. The positions $\bar{H}$ of satellites are given by the relation [213]

$$\bar{H} = \bar{h} + (m + \bar{h} \cdot \bar{R})\, \bar{q} \qquad (2.28)$$

(*m* = integer; order of the satellite; $\bar{q} = (1/D)\bar{e}_n$. The diffraction pattern thus exhibits main (or basic) spots at $\bar{h}$. With each basic spot $\bar{h}$ a linear sequence of equidistant satellites $m\,\bar{q}$ is associated. These sequences are perpendicular to the interfaces and they are shifted with respect to the positions of the basic spot over a fraction $\bar{h} \cdot \bar{R}$ of the intersatellite spacing; $\bar{q}$ is the wave vector of the modulation. Provided the basic structure is known, the long-period structure can be determined from the geometry of the diffraction pattern.

If the displacement vector $\bar{R}$ is parallel to the interfaces, the latter are termed conservative since the long-period structure has the same chemical composition as the basic structure.

If $\bar{R}$ is not parallel to the interfaces the superstructure may have a composition different from that of the basic structure; this is the case for shear structures, in which the interfaces perpendicular to $\bar{e}_n$ are generated by removing slabs of material with thickness $\bar{R} \cdot \bar{e}_n$, followed by closing the gaps so created by the displacement $\bar{R}$. If the removed slab has a composition different from that of the basic structure, the overall composition changes in the process.

The alloy, based on a face-centered cubic lattice, with ideal composition $Au_{22}Mn_6$, produces the diffraction pattern shown in Figure 54 A and represented schematically in Figure 54 B. It consists of sequences of superstructure spots, marked as full dots in Figure 54 B, associated with the basic spots marked as small open disks in Figure 54 B, which are the spots that would be produced by the Au_4Mn structure. Both types of spots are present simultaneously in Figure 54 A because the selected sample contained a small area of Au_4Mn as well as an area exhibiting the $Au_{22}Mn_6$ structure. From the direction of the rows of superstructure spots, their spacing, and the fractional shifts, the model of Figure 54 D can be derived [214]. In this model, only the manganese columns are represented. The high-resolution image (Fig. 54 C) made in the bright-field superlattice mode can be compared directly with the model of Figure 54 D; only the minority atom columns are revealed as bright dots.

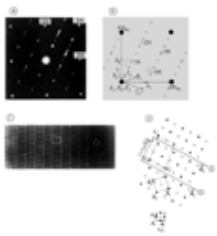

Figure 54. One-dimensional nonconservative long-period superstructure of the Au_4Mn structure, with theoretical composition $Au_{22}Mn_6$ [214]

A) Composite diffraction pattern along the [001] zone of Au_4Mn; B) Schematic representation of composite diffraction pattern (open dots indicated locations of basic spots, full dots are superstructure spots); C) High-resolution image of the superstructure (on the right, a small area of basic Au_4Mn structure is also visible); D) Model of the one-dimensional superstructure; only manganese columns are represented

Molecular Crystals; Fullerites. Organic crystals are usually prone to ionization damage and decompose very rapidly under electron irradiation; they can thus be studied for only a short time (a few seconds) and only with a very low electron beam intensity. Transmission electron microscopy has, therefore, seldom been applied to organic crystals. However, the all-carbon molecules C_{60}, C_{70}, etc., (fullerenes) discovered at the end of the 1980s resist electron radiation fairly well. Early structural studies on the crystalline phases of fullerenes (fullerites) were performed mainly by electron microscopy because only small quantities of sufficiently pure material were available. At room temperature, C_{60} crystals were found to have the

fcc structure, often containing intrinsic stacking faults, twins, and other defects characteristic of low stacking fault energy fcc alloys [215]. Figure 55 shows, for instance, various faults in a crystal of C_{60}: an intrinsic fault containing a dipole of stair rod dislocations in S, and a Frank partial dislocation and its associated stacking fault in A. The diffraction effects and the microstructure caused by the orientational phase transition in C_{60} at 255 K, from the room temperature fcc phase to the simple cubic low-temperature phase, were studied by means of electron microscopy [216] (Fig. 56).

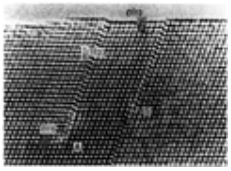

Figure 55. High-resolution image of a fragment of a C_{60} molecular crystal: bright dots represent rows of C_{60} molecules

Note the presence of a stepped intrinsic stacking fault (in S). Also a Frank-type partial dislocation and its associated stacking fault are present (in A) [215]

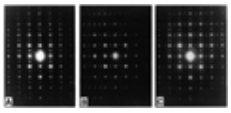

Figure 56. Diffraction patterns of the simple cubic phase of C_{60} along a cube zone [216]

A) One variant; B) Perpendicular variant to A); C) Both variants present

Vapor-grown C_{70} crystals were found to belong to two different structural types. About 1 % of the vapor-grown crystals are hexagonally close packed; the rest are face-centered cubic. Many crystal fragments exhibit both crystal structures. The hexagonally close-packed crystals undergo several orientational phase transitions on cooling, leading to a monoclinic orientationally ordered superstructure at low temperature [217]. The fcc crystals undergo a single sluggish phase transition into a rhombohedral phase.

Quasi Crystals. Until 1984, one of the most firmly established fundaments of crystallography was the postulate that crystals have three-dimensional translation symmetry, and that only symmetry rotation axes with multiplicities 1, 2, 3, 4, and 6 can occur because they are the only ones that are consistent with homogeneous space filling. However electron microscopy and electron diffraction have shown that 5-, 8-, 10-, and 12-fold symmetry axes occur in the structure of certain, usually rapidly cooled, alloy phases with complicated compositions [218, 219]. Careful tilting experiments have demonstrated the presence of icosahedral and dodecahedral symmetry groups in many of these phases. Their electron diffraction patterns consist of sharp spots, whose geometry reflects the above-mentioned symmetry elements, but they do not exhibit translation symmetry. The same conclusions follow from high-resolution images [220].

At present, numerous alloy systems are known to exhibit noncrystallographic symmetry elements and to lack three-dimensional periodicity; they are called quasi crystals. Figure 57 shows the diffraction pattern of a quasi crystal with composition Al – Mn exhibiting a fivefold axis, whereas Figure 58 shows the corresponding high-resolution image; pentagonal arrangements of bright dots are visible in many places.

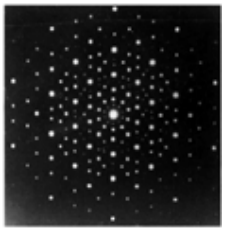

Figure 57. Diffraction pattern of the icosahedral phase of quenched Al – Mn alloy exhibiting fivefold symmetry (Courtesy of G. VAN TENDELOO)

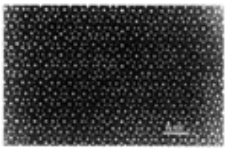

Figure 58. High-resolution electron micrograph of rapidly cooled Al – Mn alloy showing the absence of a periodic structure; pentagonal arrangements of dots can be observed (Courtesy of G. VAN TENDELOO)

Artificial Layer Structures and Semiconductor Devices. In recent years the growth of artificial layer structures, which can be used as quantum wells, has become an important research activity. The properties of such layered structures are determined to a large extent by the perfection of the interfaces, as well as the regularity of their spacing. Both parameters can be studied by means of high-resolution, cross-sectional samples. The steps in the preparation of cross-sectional samples are discussed in Section Specimen Preparation. The method is clearly destructive and cannot be used for fabrication control. However, it is the most direct way to calibrate other nondestructive methods.

Figure 59 shows a high-resolution image of a quantum well consisting of the alternation of GaAs and (GaAl)As. The positions of the interfaces are indicated by arrows. The interfaces are flat to within two or three atomic layers; no interfacial dislocations are observed. The composition profiles across such interfaces can in principle be studied by measuring the brightness of the dots representing the atom columns [221].

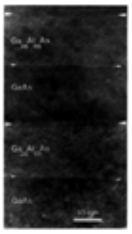

Figure 59. High-resolution image of quantum well consisting of alternation of pure GaAs and (GaAl)As

For the study of semiconductor devices, diffraction contrast images made in a high-voltage electron microscope (≈ 1000 kV) are often useful since thicker specimens can be tolerated than in a conventional microscope (≈ 200 kV). An example is shown in Figure 60, which is the cross section of a field effect device. The different layers have been indicated along with the orientation of the silicon substrate. The black dots numbered 1 – 5 are images of dislocations seen end on, which were nucleated by the stresses set up as a result of the oxidation process. From the spatial distribution of the dislocations the magnitude of the stress can be estimated.

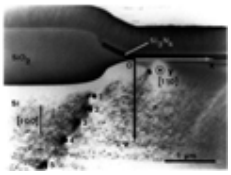

Figure 60. Cross section of field effect device

The procession of dislocation 1 . . . 5 results from stresses set up by the oxidation process (Courtesy of J. VAN HELLEMONT)

During crystal growth, small oxide particles are often formed in the interior of silicon single crystals. Such particles cause a compressive stress on the surrounding matrix. This may lead to prismatic punching, whereby disks of self-interstitials limited by loops of perfect dislocations are emitted, thereby relieving the stresses. These loops can glide along cylindrical surfaces whose cross section is determined by the size and shape of the particle, and whose generators are parallel to the different glide vectors of silicon, which are of type 1/2 [110]. The geometry of the process can be established with diffraction contrast images (see Fig. 61).

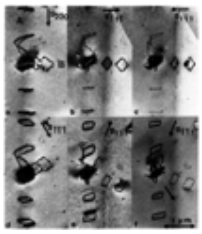

Figure 61. Oxide particle in the interior of a silicon single crystal

Stresses have been relieved by prismatic punching. The dislocation loops are imaged using different reflections (Courtesy of H. BENDER)

Carbon Nanotubes, Onions. The preparation of fullerene-containing soot by the vaporization of graphite electrodes in an electric arc leads simultaneously to the formation of very fine *hollow carbon needles* or “tubules” with diameters as low as 1 nm. Electron microscopy and electron diffraction have shown that these needles consist of concentric seamless graphene tubes (2 – 15 tubes) [222]. The most remarkable feature is the helical character of some of the tubes within a tubule, as deduced from fiber diffraction patterns (Fig. 62). The helical structure is a consequence of the stepwise increase in circumference of successive concentric tubes by πc (c = lattice parameter of 2 H graphite). Since πc is not commensurate with the a-parameter of graphite the $c/2$ spacing between successive tubes and their seamless character can be reconciled only with the changing diameter if some of the tubes at least become helical. However, a detailed analysis of the fiber pattern has shown that in most tubules the majority of tubes are nonhelical [223].

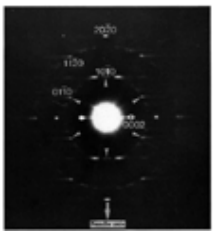

Figure 62. Electron diffraction (fiber) pattern produced by a graphite nanotube showing that certain graphene tubes are helically wound, whereas others remain normal

Note the outward streaking of the diffraction spots

(Courtesy of X. F. ZHANG)

Intense “in situ” electron irradiation of particles of carbon soot for tenths of minutes in the electron microscope produces spherical particles consisting of *onion-like* concentric spherical shells of graphene-like layers [224]. The microscope allows the concentric shell structure to be imaged during its formation.

High-T_c Superconductors. High-resolution electron microscopy and electron diffraction have been applied extensively in structural studies of high-T_c oxide superconductors. In particular, $YBa_2Cu_3O_{7-\delta}$ and the related 1 – 2 – 3 family of compounds have been studied in great detail. Numerous studies of superconducting compounds can be found in *Physica*

C from 1989 onward. The generic 1 – 2 – 3 compound ($YBa_2Cu_3O_{7-\delta}$) has a structure derived from the triperovskite structure in which the succession of layers along the *c*-direction is CuO – BaO – CuO_2 – Y – CuO_2 – BaO, etc. The CuO layer consists of Cu – O – Cu – O – chains parallel to the b_0-direction; this feature breaks the tetragonal symmetry and reduces the symmetry to orthorhombic. The average structure was determined by X-ray and neutron diffraction, but several important structural features were discovered by electron microscopy. The oxygen deficiency was found to be accommodated by the absence of oxygen in a fraction of the chains in the CuO layer [225]. The structure in which one of two chains is alternately free of oxygen, the so-called 2 a_0 (or ortho II) superstructure, has an ideal composition with $\delta = 0.5$. It occurs in two orientation variants, corresponding to the (110) reflection twin texture of the 1 – 2 – 3 compound, which was also discovered by electron microscopy. The presence of the 2 a_0 structure was shown to be responsible for the 60 K plateau in the T_c versus composition curve [225]. Locally, structures with a period of 3 a_0 occur [226].

Much attention has been devoted to various substitutions in CuO layers, replacing copper ions by other metallic ions such as Fe, Co, etc., but also substituting by complex anions such as CO_3^{2-}, SO_4^{2-}, PO_4^{3-}, or NO_3^-. High-resolution electron microscopy has demonstrated that the substitution occurs in the CuO layers, and has allowed the resulting superstructures to be visualized. In the case of SO_4^{2-} substitution, for instance, the superstructure has been shown not to be commensurate with the basic 1 – 2 – 3 lattice [227]. Figure 63 shows a view along the b_0-direction of the 1 – 2 – 3 matrix. The prominent, bright dot squares reveal the SO_4^{2-}-containing rows that replace CuO chains. Their arrangement is not periodic and can best be described as resulting from a concentration wave with a wave vector that is not commensurate with the 1 – 2 – 3 lattice along the *a*-direction.

Figure 63. High-resolution image along the $[010]_0$ direction of the SO_4^{2-}-substituted 1 – 2 – $(YBa_2[Cu_{3-x}(SO_4)_x]O_{7-\delta})$ compound

The SO_4^{2-}-containing rows parallel to [010] are marked as small crosses of prominent bright dots. (Courtesy of T. KREKELS) [227]

Replacing copper by cobalt or gallium in the CuO layer results in the formation of corner-sharing chains of CoO_4 (GaO_4), tetrahedra along the [110] and [1$\bar{1}$0] directions of the 1 – 2 – 3 matrix lattice [227]. The resulting structure is still orthorhombic, but with a diagonal unit cell ($a_0 \approx b_0 \approx a_p\sqrt{2}$), where a_p is the lattice parameter of the basic perovskite. Electron microscopy [228] showed that a superstructure is formed, which is localized in the Co – O (Ga – O) layers, and in which alternating parallel CoO_4 (GaO_4) chains have mirror-related configurations, leading to doubling of the lattice parameter perpendicular to the chain direction. Figure 64 is a view of the cobalt-containing material along a $[120]_p$ direction, which allows the period doubling along a_0, as well as the possibility of polytypism along c_0, to be observed.

Figure 64. High-resolution image along the $[120]_p$ zone of the cobalt-substituted 1 – 2 – 3 compound $YBa_2Cu_2CoO_7$

Dark dots reveal indirectly the arrangement of chains of corner-sharing CoO_4 tetrahedra that replace the CuO layers. (Courtesy of T. KREKELS) [228]

Superionic Conductors. Superionic conductors often consist of a stable rigid framework forming large channels along which ions can move easily. The channels in such structures can be made visible in high-resolution images made along a zone parallel to the channel direction. Figure 65 shows such a high-resolution image of the natural mineral hollandite with idealized composition $Ba_xMn_8O_{16}$ [229]. Octagonal channels, formed by interconnected strings of edge and corner-sharing MnO_6 octahedra, are imaged as octagons of black dots. These channels are occupied by barium ions, which are also visible as black dots in the centers of the octagons.

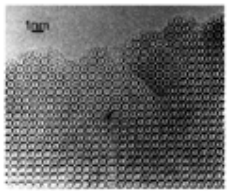

Figure 65. Structure of natural monoclinic hollandite $Ba_xMn_8O_{16}$ as viewed along the [010] zone

The image is compared with a simulated image in the inset [229].

Polymers. Although most organic materials usually deteriorate rapidly from ionization damage in the electron beam, meaningful observations can be made on certain polymers. In Figure 66, fiber patterns of a poly(*p*-phenylene) stretched six to seven times are reproduced (PPV, an alternating copolymer of *p*-phenylene and acetylene) at two different photographic exposures to reveal the details of the pattern close to the origin [230]. The direction of stretching is indicated by arrows. The patterns of the left column (A) and (C) refer to the pristine material. On doping with $FeCl_3$ this material becomes a good electrical conductor. The effect of $FeCl_3$ doping on the diffraction pattern is shown in the right column (B and D). Whereas in the undoped material the 001 diffraction vector encloses an obtuse angle with the row of hoo

reflections, this angle becomes 90° in the doped material, which suggests that the symmetry changes from monoclinic to orthorhombic on doping.

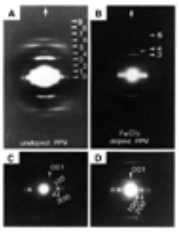

Figure 66. Fiber texture pattern of a stretched PPV polymer

A) and C) Pristine polymer; B) and D) After $FeCl_3$ doping

(Courtesy of X. F. ZHANG)

2.2.14.2. *Defects*

Dislocations. Dislocation configurations have been studied mainly by diffraction contrast bright-field imaging since this method is very sensitive to lattice strain fields. Under these conditions, dislocations appear as dark lines when two-beam diffraction conditions close to $s \approx 0$ are used. A network of dislocations in the basal plane (0001) of zinc is visible in Figure 67.

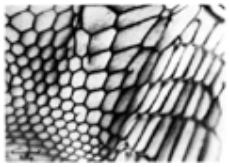

Figure 67. Network of undissociated dislocations in the (0001) plane of zinc

Images of dissociated dislocations (i.e., dislocation ribbons) have been used extensively as a means to deduce the stacking fault energy [141].

The separation between partial dislocations is usually measured on a weak beam image (see Section Dislocation Contrast). The image, which is now a very narrow bright line on a darker background, is essentially kinematic. In the high-resolution image of Figure 68, dissociated dislocations in silicon are viewed along the [110] zone. Between the two partial dislocations an intrinsic stacking fault is present. Such configurations have been used to deduce stacking fault energies [141, 192].

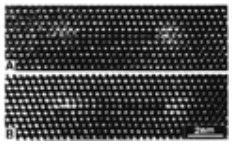

Figure 68. Dissociated dislocations in silicon viewed along the [110] zone in the high-resolution mode (Courtesy of H. BENDER)

Planar Defects. The diffraction contrast images of planar interfaces that intersect the foil surfaces consist of a set of fringes parallel to the closest surface with a depth period given by $1/\sigma_g$ (which is equal to t_g for s = 0; see Section Faulted Crystals). If the planar interface is parallel to the foil surfaces, which is often the case in cleaved foils, the faulted area exhibits only a brightness difference.

High-resolution images can reveal stacking faults directly, especially in close-packed structures. In such structures the image made along a zone parallel to the close-packed rows reveals the stacking directly. A stacking fault in a wurtzite crystal is reproduced in Figure 69.

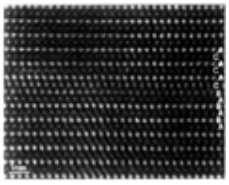

Figure 69. High-resolution image of a stacking fault in 2H-wurtzite (Courtesy of H. BENDER)

Domain boundaries and twins in planes that intersect the foil surfaces are imaged as δ-fringes (Section Faulted Crystals) [171]. The extinction condition is now $\Delta s = 0$ (or $\Delta \bar{g} = 0$!) which is satisfied for the family of planes common to the two domains. When the interfaces are nearly perpendicular to the foil plane the main effect is domain contrast (i.e., a brightness difference in the domains on either side of the interface) as shown in Figure 70, where the same twin boundaries in the high-temperature superconductor $YBa_2Cu_3O_{7-\delta}$ have been imaged in three different modes.

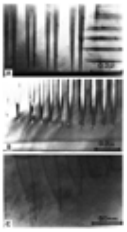

Figure 70. Twin boundaries in $YBa_2Cu_3O_{7-\delta}$

(A) Domain contrast; (B) Interface contrast; (C) High-resolution imaging (Courtesy of H. W. ZANDBERGEN)

2.2.14.3. *Small Particles*

Like most other electron microscope techniques, transmission electron microscopy allows the particle-size distributions and particle shapes of very fine powders to be studied. Such studies are important in several areas of research: including catalysis, magnetic recording, and photography. Transmission electron microscopy provides, in addition, the possibility of obtaining electron diffraction patterns of single particles as well as high-resolution images. In Figure 71 single particles of silver vapor deposited on an amorphous carbon substrate is shown. They are clearly not single crystals but aggregates of multiply twinned fcc crystallites having an overall icosahedral shape.

Figure 71. Small silver particles consisting of an icosahedral aggregate of multiply twinned face-centered cubic crystals (Courtesy of C. GOESSENS)

The topotactic dehydration reaction [231] transforming goethite into hematite was studied “in situ” at room temperature and shown to produce small twinned hematite crystals (Fig. 72). In the initial stages of the dehydration process, satellite spots appear in the diffraction pattern around the hematite spots; they were found to be due to a texture consisting of a periodic alternation of voids and hematite crystals, rather than to a long-period superstructure.

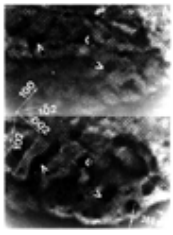

Figure 72. Small particles of twinned hematite formed in the electron microscope as a decomposition product of goethite [231]

The crystal habit of the silver halide particles used in photographic emulsions is important since it is one of the parameters determining the photosensitivity of the film. The hexagonal and triangular (111) tabular crystals of AgCl were shown to be in fact multiply twinned on planes parallel to the habit plane. Depending on the parity of the number of twin lamellae the particle grows into either hexagonal or a triangular habit.

When silver halide crystals with a cube habit are exposed to electrons, photolytic metallic silver specks are formed “in situ” in an epitaxial relation to the silver halide substrate. This gives rise to moiré fringes parallel to the cube edges due to the difference in lattice spacing between AgCl and Ag (Fig. 73).

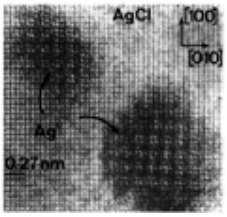

Figure 73. Silver chloride crystal covered by silver specks viewed along the cube zone

2.2.14.4. *Surface Studies [232]*

Crystal surfaces can be studied in a transmission electron microscope by different modes. It is possible to use an electron beam, in reflection under *grazing incidence*, magnifying the Bragg reflected beam due to planes roughly parallel to the surface. The surface of the sample must be almost parallel to the beam. The resulting image is highly distorted in the sense that the magnification in a direction perpendicular to the beam is much lower than the magnification along the projection of the electron beam. This Bragg reflection mode can be used to image surface steps and to detect surface superstructures due to reconstruction or to adsorbed layers by diffraction. A very clean vacuum is required.

A second mode consists of observing the surface profile in transmission with the *electron beam parallel to the surface*. In this way, surface relaxations and surface reconstruction can be observed together with atomic resolution of the substrate. An example of the cleavage surface of a high-T_c superconductor is reproduced in Figure 74. It allows the layer within the structure to be located along which cleavage takes place [233].

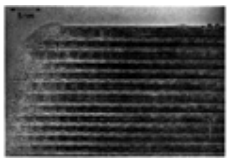

Figure 74. High-resolution image of the free cleavage surface of a superconductor ($Bi_2Sr_2Ca_1Cu_2O_x$)

Cleavage occurs between the two BiO layers [233]

2.2.14.5. *Thin Epitaxial Layers*

The development of modern miniaturized electronic devices relies heavily on the use of thin epitaxial films of which the relevant physical properties sensitively depend on their thickness and on their microstructure. In recent years high-resolution electron microscopy has extensively been applied for the primary characterization of such films. In particular various modes of accommodating the misfit between a single-crystal substrate and the epitaxial films have been discovered by this method.

Routine characterization, which necessarily has to be based on nondestructive methods can be “calibrated” by means of HRTEM, which is a destructive, but highly informative method. Examples are shown in Figure 75.

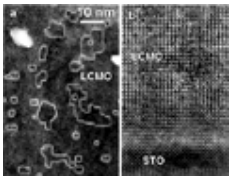

Figure 75. Thin film of $La_{1-x}Ca_xMnO_3$ on a $SrTiO_3$ (STO) single-crystal substrate

a) Plane view, low-resolution image revealing antiphase boundaries

b) High-resolution image of a cross section specimen (courtesy O. LEBEDEV)

2.2.14.6. *“In situ” Studies [234]*

By using different types of stages the specimen chamber of the microscope can be transformed into a small laboratory. Heating and cooling stages make it possible to study “in situ” the changes in crystal structure and in microtexture accompanying phase transitions, by observing (1) the appearance or disappearance of superstructure spots in the diffraction pattern, and (2) the fragmentation with translation and/or orientation domains in the direct space image.

Environmental cells can be used to study in situ solid – gas reactions, such as oxidation processes or decomposition reactions.

Plastic deformation can be studied and the dislocation propagation observed in real time in a straining stage.

The creation of radiation damage by ionization or by electron – atom collisions can be studied in medium- and high-voltage microscopes. The formation of agglomerates of point defects, such as dislocation loops and stacking fault tetrahedra, can be observed during irradiation with the image-forming electrons. (Light atoms, such as oxygen, are readily displaced by electrons of 400 kV.)

Radiation ordering and disordering in alloy systems have successfully been studied in high-voltage microscopes at different temperatures.

2.3. Scanning Electron Microscopy

2.3.1. *Introduction*

In a scanning electron microscope (SEM) [235-237], a small electron probe 1 – 10 nm in diameter scans in a raster across the surface of the specimen (Fig. 76). The incident electrons are elastically and inelastically scattered by the specimen. *Elastic scattering* results in large scattering angles and zigzag electron trajectories. Therefore, a fraction of electrons can leave the specimen as backscattered electrons (BSE) (Fig. 77). The slowing down of electrons by *inelastic scattering* results in an electron range *R*. Electrons from the specimen atoms that are excited by inelastic scattering can leave the specimen as *secondary electrons* (SE) from a thin surface layer Λ_{SE} of ca. 1 – 10 nm. By convention, electrons in the energy spectrum (Fig. 78) with $E \leq 50$ eV are called SE. The secondary electrons consist of (1) SE1 excited by the primary electrons; (2) SE2 excited by BSE on their path through the surface; (3) SE3 are excited when BSE strike the lower polepiece; or flow (4) as SE4 through the polepiece bore (see Fig. 77). The ionization of inner atomic shells results in the emission either of characteristic X-ray quanta (X) or of Auger electrons (AE).

Figure 76. Schematic cross section of a scanning electron microscope (SEM)

a) Electron gun; b) Condenser lenses; c) Scan coils; d) Objective; e) Photomultiplier; f) Amplifier; g) Scan generator; h) Cathode-ray tube

BSE = Backscattered electrons; SE = Secondary electrons; SC = Specimen current; EBIC = Electron-beam-induced current; CL = Cathodoluminescence; X = X-rays

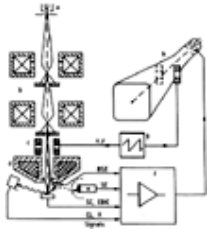

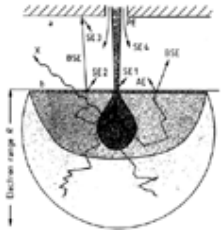

Figure 77. Excitation of backscattered electrons (BSE), Auger electrons (AE), and different groups of secondary electrons (SE) by primary electrons (PE)

Volumes of electron diffusion, BSE trajectories, and emitted X-rays are shown with increasing density of gray levels

a) Polepiece; b) Specimen

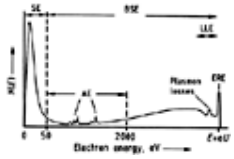

Figure 78. Schematic energy spectrum of emitted electrons with secondary (SE), backscattered (BSE), elastically reflected (ERE), low-loss (LLE), and Auger electrons (AE)

The *image* is formed by the signal of emitted secondary electrons, backscattered electrons, Auger electrons, absorbed specimen current (SC), or X-ray quanta, which modulate the intensity of a cathode-ray tube rastered in synchronism (Fig. 76). Conventional SEMs work with electron acceleration voltages of 5 – 30 kV, whereas a *low-voltage scanning electron microscope* (LVSEM) [238, 239] uses 0.5 – 5 kV.

2.3.2. *Instrumentation*

2.3.2.1. *Electron Guns*

The following types of electron emitters are used in SEM electron guns.

Thermionic cathodes consist of a directly heated tungsten hairpin cathode at T_c = 2500 – 3000 K, or an indirectly heated pointed rod of lanthanum or cerium hexaboride (LaB_6, CeB_6) at 1400 – 2000 K. The electrons must overcome the work function Φ of 4.5 eV (W) or 2.7 eV (LaB_6) by thermal activation (Fig. 79, curve a). Between the cathode at the potential $-U$ and the grounded anode, a negatively biased Wehnelt electrode forms a crossover of diameter 20 – 50 µm (W) or 10 – 20 µm (LaB_6) as an effective electron source. The emitted electrons show an energy spread ΔE = 1 – 2 eV (W) or 0.5 – 1 eV (LaB_6). A measure of the quality of an electron gun is the axial gun brightness β:

$$\beta = j/\pi\alpha^2 \approx j_c E/\pi k T_c \qquad (2.29)$$

where j is the current density, α the aperture angle, and E the electron energy. The axial gun brightness is constant for all points on the axis regardless of lenses and aperture diaphragms (j_c and T_c are current density and temperature, respectively, at the cathode). It increases proportionally to the electron energy E, with $\beta \approx 10^5$ A cm^{-2} sr^{-1} (where sr denotes steradian) for tungsten at E = 20 keV and about ten times higher values for LaB_6.

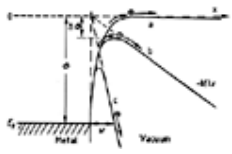

Figure 79. Potential at the cathode – vacuum interface for W and LaB_6 thermionic emission, ZrO – W Schottky emission, and field emission guns

a) Thermionic emission; b) Schottky emission; c) Field emission

Φ = Work function; E_F = Fermi energy

Schottky emission cathodes consist of zirconium-doped tungsten tips, with a radius of about 0.5 – 1 µm, coated with a ZrO layer. This layer decreases the work function from 4.5 to 2.7 eV. A Schottky emission cathode works with a higher electric field strength at the tip that decreases the work function by $\Delta\Phi$ (Schottky effect, Fig. 79, curve b) and concentrates the emission at the tip with a virtual electron source having a diameter of 15 –20 nm. However, the electrons still have to overcome the work function by thermal activation. The energy spread is 0.3 – 1 eV.

Field-emission guns (FEG) consist of tungsten tips with a radius of ca. 0.01 – 0.1 µm. An intermediate anode at ca. 1 – 2 kV extracts electrons by tunneling through the barrier of the work function, which is w = 1 – 10 nm in width (Fig. 79, curve c). Two types of FEG exist: (1) *cold FEGs* that work with the tip at room temperature, and the tip is flashed once a

day, (2) *heated FEGs* work at 1800 K. The energy spread is only 0.2 – 0.4 eV for cold and 0.5 – 0.7 eV for heated FEGs. The gun brightness of FEGs $\beta \approx 10^7 - 10^8$ A cm^{-2} sr^{-1}.

2.3.2.2. *Electron Probe Formation*

The crossover of a thermionic gun or the virtual point sources of Schottky or field-emission guns are demagnified by two to three electromagnetic lenses (Fig. 76), so that a geometric probe size d_g is formed at the specimen with a diameter of 1 – 10 nm. A diaphragm of ca. 50 – 200 µm diameter in front of the last lens limits the electron-probe aperture α_p. The electron – probe current I_p, d_g, and α_p cannot be changed independently because of the conservation of gun brightness β in Equation (2.29):

$$I_p \approx \frac{\pi}{4} d_g^2 j_p = \frac{\pi^2}{4} \beta d_g^2 \alpha_p^2 \tag{2.30}$$

The geometric probe size d_g is superposed by aberration disks of spherical (d_s) and chromatic (d_c) aberration and diffraction (d_d):

$$d_s = 0.5 C_s \alpha_p^4; \quad d_c = C_c(\Delta E/E)\alpha_p; \quad d_d = 0.6\lambda/\alpha_p \tag{2.31}$$

with the aberration constants C_s and C_c, respectively; ΔE is the energy spread of the electron gun, and λ is the deBroglie wavelength

$$\lambda = h/mv = 1.23/\sqrt{E} \tag{2.32}$$

with λ in nanometers and E in kiloelectronvolts. An astigmatism of the last probe-forming lens can be compensated using a stigmator, which consists of magnetic quadrupoles.

The (quadratic) superposition of d_g, d_d, d_s, and d_c results in an effective probe diameter d_p (Fig. 80), where d_g and d_s dominate for high electron energies and thermionic guns, and d_d and d_c dominate for low-voltage SEM and field-emission guns. This superposition results in a minimum probe size between 1 and 10 nm at optimum apertures of tens of milliradians. An anticipated increase in resolution (decrease of electron probe diameter) requires a compensation of the chromatic aberration by using a combination of electrostatic and magnetic multipoles.

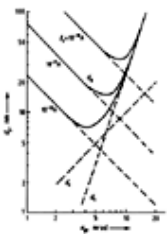

Figure 80. Double-logarithmic superposition of geometric probe size d_g for different probe currents I_p and aberration disks d_c and d_s versus the electron probe aperture α_p for a thermionic tungsten cathode

E = 20 keV; β = 7×10^4 A cm^{-2} sr^{-1}; C_s= 50 mm; C_c = 20 mm; ΔE = 1 eV

The low optimum probe aperture α_p of the SEM results in a large depth of field D_f:

$$D_f = \delta/\alpha_p = \Delta/\alpha_p M \tag{2.33}$$

where $\Delta = \delta M \approx 0.1$ mm is the resolution and M the magnification on the cathode-ray tube (CRT) screen. As a consequence, an SEM has a two-order-of-magnitude greater depth of field than a light microscope even at low magnification.

To decrease the aberration constants C_s and C_c, the focal length of the last probe-forming lens can be reduced by a stronger excitation of the magnetic lens, which means a short working distance between specimen and polepiece for high resolution. The aberration constants can be decreased further if the specimen is positioned in the lens field; this makes through-lens detection of secondary electrons necessary.

2.3.2.3. *Detectors*

Secondary electrons are most frequently used for image formation and can be detected by a scintillator – photomultiplier combination (*Everhart – Thornley detector*, Fig. 81). A large fraction of slow SE is attracted by a positively biased grid (b) and accelerated to a yttrium – aluminum – garnet (YAG) scintillator or a P47 powder scintillator (c) biased at +10 kV. The photons generated are guided in a light pipe (d) to a photomultiplier (f). This SE detector shows a high signal-to-noise ratio and a large bandwidth up to 10 MHz. The collection field of an Everhart – Thornley detector mounted on one side of the specimen can disturb the electron probe in LVSEM. In this case, beam deflection and distortion can be avoided either by a combination of electrostatic and magnetic quadrupoles forming a Wien filter, by a retarding electrostatic lens, or by through-lens detection of SE.

Figure 81. Everhart – Thornley detector for SE

a) Specimen; b) Collector grid and screen; c) Scintillator; d) Light pipe; e) Optical contact; f) Photomultiplier;

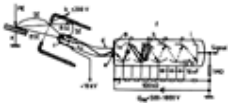

g) Photocathode; h) Dynodes; i) Anode

Backscattered electrons have enough energy for direct production of a greater number of photons in a scintillator coupled via light pipe to a photomultiplier. However, the Everhart – Thornley detector cannot be used effectively for faster BSE because only a small fraction hits the scintillator area. Therefore, *annular or semiannular top detectors* collecting BSE with a high takeoff angle or *ring detectors* with low takeoff, are applied to make the best use of the angular characteristics of BSE. Alternatively, *semiconductor detectors* can be used. The BSE produce a large number $n = E/E_i$ of electron – hole pairs, where E_i = 3.6 eV is the mean energy per excitation in silicon. These charge carriers can be separated in a *p – n* junction and form an electron beam-induced current (EBIC). BSE can also be detected by the *conversion of BSE to SE3* at the polepiece and other parts of the specimen chamber, when a negatively biased electrode around the specimen retards SE1 and SE2 from the specimen.

Detectors for LVSEM. Both scintillator and semiconductor detectors show a signal decreasing linearly with decreasing electron energy, and a threshold energy of ca. 1 – 5 keV. Therefore, their application in LVSEM requires postacceleration. As a recent alternative, *microchannel plates* (MCPS) can be used for the detection of BSE. These consist of a slice from a boule of tightly packed, fused tubes of lead-doped glass with a 10 – 20 µm inner diameter and a resistance of 10^8 – 10^9 Ω over their length. Annular disks 3 – 4 mm thick can be mounted below the polepiece. The incident electrons produce SE at the inner tube wall, which are accelerated by a continuous voltage drop along the tube with a bias of 1 keV and are multiplied inside the MCP tubes by a multiplier-like action. The anode plate and a preamplifier at a potential of 1 kV are electrically insulated by an optical decoupler or by transmitting a modulated 30-MHz signal.

2.3.3. *Electron – Specimen Interactions*

2.3.3.1. *Elastic and Inelastic Scattering*

Elastic scattering of incident electrons results from the attractive Coulomb force of the nucleus screened by the atomic electron cloud with a differential Mott cross section of

$$\frac{d\sigma_M}{d\Omega} = r(\theta)\frac{e\,Z}{4\,(4\,\pi\,\varepsilon_0)^2\,m^2\,v^4}\frac{1}{\sin^4\,(\theta/2)} \tag{2.34}$$

where $r(\theta)$ is the ratio between the Mott and Rutherford cross sections, Ω is the solid angle, Z the atomic number of the nucleus, ϵ_0 the vacuum permittivity, and m the mass of the electron [235]. The strong differences between Mott and Rutherford cross sections result from taking account of the spin – orbit coupling of electrons and solving the relativistic Dirac equation, whereas the Rutherford cross section is only a solution of the Schrödinger equation not containing the spin.

Inelastic scattering results in an excitation of electrons of the solid and a corresponding energy loss ΔE of the incident electron. Information about the differential inelastic cross section $d^2\sigma/d\Omega\, d(\Delta E)$ can be obtained from dielectric theory or experimental electron energy loss spectra (EELS; see Section Electron Energy Loss Spectrometry (EELS)) of high-energy electrons [240, 241].

A series of inelastic scattering processes with statistical energy losses results in a slowing down of electrons, which can be described by a mean energy loss per unit path length (Bethe stopping power S)

$$S = \left|\frac{dE_m}{ds}\right| = \frac{2\,\pi\,e^4\,N_A\,r\,Z}{(4\,\pi\,\varepsilon_0)^2\,A\,E}\ln\,(1.166E/J) \tag{2.35}$$

with the mean ionization potential $J \approx 12.5\,Z$. As energy decreases, fewer subshells are ionized, which can be described by a parabolic decrease of $1/S$ below $E/J = 6.3$.

2.3.3.2. *Electron Diffusion*

The decrease of mean electron energy E_m with increasing path length s of electron trajectories due to Equation (2.35) results in a mean total path (Bethe range R_B) that increases with increasing Z of the material. In low-Z material with less frequent large-angle scattering, R_B and the practical range R are equal, whereas in high-Z material with more frequent scattering, the trajectories are strongly curled and $R < R_B$, as demonstrated by Monte Carlo simulations in Figure 82. The range

$$R \approx 6.6E^{5/3} \tag{2.36}$$

is to a large extent independent of material, when R is measured in units of mass thickness (µg/cm^2) and E in keV.

Figure 82. Monte Carlo simulations of 10- and 1-keV electron trajectories in carbon and gold

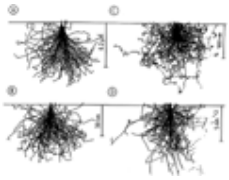

A) C, 10 keV, ϱ = 1 g/cm^3; B) C, 1 keV, ϱ = 1 g/cm^3; C) Au, 10 keV; D) Au, 1 keV

2.3.3.3. *Emission of Secondary and Backscattered Electrons*

The *backscattering coefficient* η describes the fraction of *primary electrons* leaving the specimen with an energy reduced by energy losses. Emitted electrons with energies of ≤50 eV are called secondary electrons (see Section Introduction) and are described by the *secondary electron yield* δ. Both quantities depend on electron energy *E*, atomic number *Z* of the specimen, and surface tilt angle ϕ (ϕ = 0: normal incidence), and show characteristic energy and angular distributions that are important to the discussion of image formation in using BSE and SE signals.

Figure 83 shows the dependence of the *backscattering coefficient* η on atomic number *Z* for different electron energies *E*. For *E* > 5 keV, η increases monotonically with increasing atomic number. For multicomponent targets

$$\eta = \sum c_i \eta_i \tag{2.37}$$

shows a best fit to experiments, where c_i represents the mass fractions. This dependence of η on *Z* is responsible for the atomic number (compositional) contrast of the BSE signal for *E* = 5 – 30 keV (see Section Topographic and Material Contrast). The backscattering coefficient η is approximately independent of *E* in the range 10 – 100 keV. Below 5 keV, η decreases for *Z* > 30 and increases for *Z* < 30 with decreasing *E*. The reason for this is the Mott elastic cross section (Eq. 2.34). An increasing tilt angle ϕ of the specimen (ϕ = 0: normal incidence) results in an increase of η.

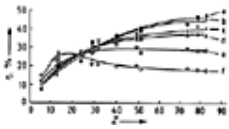

Figure 83. Backscattering coefficient η as a function of atomic number *Z* for different electron energies *E*

a) *E* = 5 keV; b) *E* = 4 keV; c) *E* = 3 keV; d) *E* = 2 keV; e) *E* = 1 keV; f) *E* = 0.5 keV

The *secondary electron yield* δ shows a maximum for primary energies *E* of a few hundred electron volts and decreases ~$E^{-0.8}$ for higher energies (Fig. 84). For *E* > 5 keV, the SE yield contains the contributions of SE1 excited by the primary electrons (PE) (Fig. 76) and of SE2 excited by the BSE. For *E* < 5 keV SE1 and SE2 can hardly be distinguished because the exit depths of SE and BSE have the same order of magnitude. The dependence of δ on the surface tilt angle ϕ can be approximated by $\delta \sim \sec^n \phi$, where the exponent *n* decreases from 1.3 (Be) to 1.1 (Al) and 0.65 (Au).

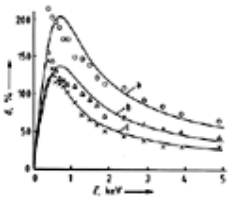

Figure 84. Decrease of SE yield δ with increasing electron energy *E* at normal incidence for

a) Au; b) Cu; c) C

2.3.3.4. *Specimen Charging and Damage*

Insulating specimens show a *negative charging* by adsorbed electrons, when the total electron yield $\sigma = \eta + \delta$ is less than unity beyond a critical energy E_2 where σ changes from values larger than to values smaller than unity. Therefore, a conductive coating is necessary for working at high electron energies. With decreasing energy, σ can become greater than unity below E_2. In this case, more electrons leave the specimen as SE or BSE than are absorbed. The specimens becomes *positively charged* by a few volts only because SE of low energies will be retarded. Therefore, insulating specimens can be observed at low energies without coating, although rough specimens can still show negative charging when, for example, SE cannot escape from holes.

Temperature-sensitive specimens can be damaged by *heat generation*. The increase of surface temperature at the electron probe decreases with decreasing energy, and such specimens can be observed better by LVSEM. Biological specimens are damaged by ionization, which results in a loss of mass and finally a polymerized carbon-enriched conglomerate within a layer of the order of the electron range *R*.

2.3.4. ***Image Formation and Analysis***

2.3.4.1. *Topographic and Material Contrast*

Surface topography can be imaged with the SE signal, where the contrast is generated by the dependence of δ on the tilt angle ϕ of a surface element, and by shadowing effects caused by reduced SE collection from surfaces with normals opposite to the direction of SE collection or inside holes and trenches. This results in a contrast nearly equivalent to a light illumination from the detector, which, however, is disturbed by the diffusion contrast caused by SE2 generated by BSE at a

greater distance from the electron impact. This typically results in bright zones near edges with a width of about the range *R*. Therefore, a low-voltage SEM at 1 – 5 keV decreases the diffusion effect and shows a better topography, as demonstrated in Figure 85. With in-lens operation and through-lens collection of SE by an Everhart – Thornley detector (ETD) inside the last lens, the SE signal becomes independent of azimuth, which is an advantage for the metrology of integrated circuits, but a disadvantage for recognizing the topography and distinguishing elevations and indentations, for example.

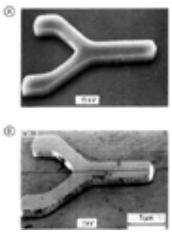

Figure 85. SE image of an etched structure (Y) on silicon with

A) 10-keV; B) 1-keV electrons

The dependence of δ and especially of η on atomic number *Z* (Fig. 83) leads to an atomic number or compositional contrast that can be used for the discrimination of phases with different mean atomic numbers, as demonstrated by an eutectic alloy shown in Figure 86.

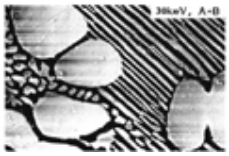

Figure 86. Example of material contrast of an Al – Ag eutectic alloy recorded with 30-keV electrons

2.3.4.2. *Electron Channeling Effects*

Electron waves entering a single crystal propagate as Bloch waves, which results in channeling effects and an orientation dependence of all electron – specimen interactions that are concentrated at the nuclei, such as large-angle scattering for backscattering or inner-shell ionization for X-ray and Auger electron emission.

When rocking an incident electron beam, the backscattering coefficient is modulated by a few percent. This results in an electron channeling pattern (ECP) (Fig. 87) containing excess (bright) or defect (dark) Kikuchi lines (see Section Geometry of Diffraction) that are the intersections of the Kossel cones of apex 90° – θ_B and an axis normal to the lattice planes of distance *d*. Therefore, such a pattern contains the crystal symmetry, and the distances of opposite Kikuchi lines include an angle 2 θ_B, which fulfills the Bragg condition

$$2\,d\,\sin\theta_B = \lambda \tag{2.38}$$

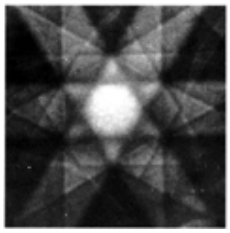

Figure 87. Electron channeling pattern of a (111) silicon surface recorded with 20-keV electrons

Scanning a polycrystalline specimen results in changes of brightness forming the crystal orientation or channeling contrast (Fig. 88).

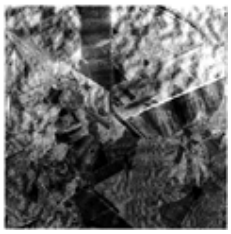

Figure 88. Superposed crystal orientation and magnetic contrast type I on polycrystalline cobalt

Another type of channeling pattern is the electron backscattering pattern (EBSP) in which the intensity modulation can be observed on a fluorescent screen as the dependence of backscattering on takeoff direction. An EBSP has the advantage of covering a large angular range of about ± 30°, whereas the rocking for ECP can be realized only with maximum ± 2 – 3°.

2.3.4.3. *Imaging and Measurement of Surface Potentials*

A positively biased or charged part of the specimen retards low-energy SE and appears darker than the surrounding parts at ground potential. Negatively biased or charged parts appear brighter because more SE are repelled and can reach the collection field of the ETD. This results in the voltage contrast. However, no unique relation exists between brightness and bias because the SE signal depends on the surrounding potentials (near-field effect) and the collection efficiency of an ETD (far-field effect).

For quantitative measurement of surface potentials U_s on integrated circuits, a high extraction field strength (400 – 1000 V/cm) is needed at the surface, which decreases the influence of near fields by neighboring potentials. The potential is measured by the shift in the SE spectrum after passage through a spectrometer of the retarding or deflection type [242].

2.3.4.4. *Imaging of Magnetic Fields*

Magnetic fields of the specimen can act on primary, secondary, and backscattered electrons by the Lorentz force $F = e\, v \times B$. The *magnetic contrast type 1* is caused by the deflection of SE by external magnetic fields, which can be observed for magnetic recording media and uniaxial ferromagnetic materials. The best contrast is observed when only about half of the SE are collected. Then the signal intensity can change by 1 – 10 % for opposite directions of the stray field. Figure 88 shows the superposition of magnetic contrast type 1 and the channeling contrast on a polycrystalline cobalt specimen.

Magnetic contrast type 2 is caused by the deflection of BSE in internal magnetic fields. However, sufficient contrast can be observed only for tilt angles ϕ of 40 – 50°, and the contrast increases with increasing energy, but is only a few per thousand for opposite directions of *B*. Another possibility for measuring external stray fields is the deflection of primary electrons, which pass at a short distance parallel to the surface.

2.3.4.5. *Electron-Beam Induced Current*

The mean number E/E_i of electron – hole pairs generated in semiconductors — with a mean formation energy of E_i = 3.6 eV in silicon, for example — normally recombine. The electric field inside depletion layers separates the charge carriers, and minority carriers can diffuse to the depletion layer and contribute to the charge collection I_{cc} or electron-beam induced current. Depletion layers can be formed by *p* – *n* junctions parallel or perpendicular to the surface or by Schottky barriers formed by a nonohmic evaporated metal contact. Therefore, a scanning electron probe becomes a useful tool for qualitative and quantitative analysis of junctions and semiconductor parameters [243], which is demonstrated by the following examples:

1. Imaging the position and depth of depletion layers below conductive pads and passivation layers, by utilizing the increasing penetration depth (range) with increasing electron energy
2. Measuring the width of depletion layers and their increase with increasing reverse bias
3. Imaging sites of avalanche breakdown in depletion layers
4. Imaging lattice defects (dislocations, stacking faults) and dopant striations that actively influence the diffusion length, for example, by a Cottrell atmosphere of dopant atoms
5. Measuring the diffusion length *L* from a semilogarithmic plot of the EBIC signal [~ exp (–*x*/*L*)] versus the distance *x* of the electron probe from a perpendicular *p* – *n* junction, and measuring the lifetime τ with a signal [~ exp (–*t*/ τ)] when chopping the electron beam
6. Measuring the surface recombination rate S from the variation of EBIC with increasing electron energy and depth of carrier formation

2.3.4.6. *Cathodoluminescence*

The excitation of light from materials stroked by electrons, known from fluorescent screens of TV tubes for example, is called cathodoluminescence (CL). In semiconductors with a direct band gap, electrons that are excited from the valence to the conduction band can recombine with the holes by emission of radiation, whereas semiconductors with an indirect band gap have a reduced probability of radiative recombination. In semiconductors and most inorganic materials, CL depends strongly on the concentration of dopants, which can either enhance CL by forming luminescence centers for radiative transitions or quench CL by forming centers of nonradiative transitions.

Organic specimens such as anthracene and plastic scintillator material, for example, can also show CL. Fluorescent dyes are used in biology for selective staining. However, all organic material is damaged by electron irradiation, and CL is quenched at incident charge densities a few orders of magnitude lower than those necessary for radiation damage of the crystal structure and loss of mass by ionization processes.

The CL signal can be detected by a photomultiplier or dispersively by using a spectrometer between specimen and detector. As in color TV, the signals from three detectors with color filters can be used for a real-color image [244].

2.3.4.7. *Special Imaging Methods*

The *specimen current mode* uses the current

$$I_s = I_p[1 - (\eta + \delta)] \tag{2.39}$$

to earth, which is complementary to the SE and BSE emission and can be used for imaging the compositional contrast of BSE or the magnetic contrast type 2.

In *environmental* SEM, the partial pressure of water or other gases can be increased near the specimen by differentially pumped diaphragms. This offers the possibility of observing wet specimens without drying, for example [245]. The production of ions in the gas also reduces the negative charging of the specimen by absorbed electrons in nonconductive

material.

In the *thermal-wave acoustical mode*, the electron beam is chopped at frequencies in the 100-kHz to 5-MHz range and produces periodic specimen heating. The periodic changes in thermal expansion excite an acoustic wave that can be picked up by a piezoelectric crystal transducer of lead zirconate titanate.

2.3.5. *Elemental Analysis*

2.3.5.1. *X-Ray and Auger Electron Emission*

Either ionization of an inner shell and the subsequent filling of the vacancy by an electron from higher energy states result in the emission of an X-ray quantum, or the energy is transferred to another electron in the higher shell, which leaves the atom as an Auger electron (see also ⟶ Surface and Thin-Film Analysis – X-Ray Photoelectron Spectroscopy (XPS)). The sum of the X-ray fluorescence yield ω and the Auger electron yield is unity, and ω decreases strongly from near unity to low values with decreasing Z and increasing X-ray series n = K, L, M. The X-ray quantum energies of a series increases approximately $\sim \sqrt{Z}$(Moseley's law), which is basic for elemental analysis by X-rays.

2.3.5.2. *X-Ray Spectrometers*

Two different types of X-ray spectrometer exist: wavelength- and energy-dispersive spectrometers (WDS and EDS, see Section X-Ray Microanalysis), which can be used in either an X-ray microanalyzer (XRMA) or an SEM.

A *wavelength-dispersive spectrometer* can analyze the characteristic lines of all elements down to beryllium when special analyzing crystals with a large lattice-plane spacing are used for Bragg reflection. XRMA with electron energies between 10 and 60 keV can work with WDS for recording three different wavelengths simultaneously, whereas only one WDS is used in an SEM.

The electron – hole pairs produced in the *energy-dispersive spectrometer* are separated by a voltage drop of ca. 1 kV at the crystal and result in a charge pulse proportional to the X-ray quantum energy. This pulse is amplified by a charge-sensitive preamplifier and runs through a unit for pulse shaping so that it can be recorded in a multichannel analyzer (MCA). The number of counts in a channel that is proportional to the quantum energy is increased by one unit. This allows all quantum energies to be recorded simultaneously, and the growth of a spectrum can be followed on the screen of the MCA (Fig. 89). Therefore, EDS are often also used in XRMA to survey the emitted X-ray spectrum. The vacuum inside an EDS can be separated from the vacuum of the microscope by a 8 – 10-µm-thick beryllium foil, which absorbs the characteristic radiation of elements $Z \leq 11$ below E_x = 1 keV. An organic foil or a windowless detector can also be used to record the lines of carbon (E_x = 250 eV), oxygen, and nitrogen, for example.

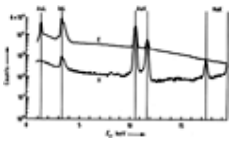

Figure 89. Energy-dispersive X-ray spectrum of InAs (logarithmic scale of counts per channel) recorded by excitation with 30-keV electrons (E) and X-ray fluorescence (X) using the molybdenum K radiation excited in a molybdenum foil target in front of the specimen

WDS and EDS show the following characteristic differences. *WDS* requires very accurate orientation of the electron impact on the Rowland circle within an area a few micrometers in diameter. The specimen must be planar and is shifted mechanically to record a line scan or elemental map. The electron probe can, by contrast, be scanned over an area of a few millimeters without any loss of efficiency when an EDS is used. WDS analyze only one line with a resolution of about 10 eV. The proportional counter allows a high counting rate of about 10 000 counts per second (cps). Therefore, high probe currents and a large number of counts per unit time can be recorded. The energy resolution of an *EDS* is only about 100 – 200 eV, and a ten times greater fraction of the continuum below the characteristic X-ray peaks is counted compared to WDS. This decreases the signal-to-background ratio in EDS. The poorer resolution of an EDS can also result in an overlap of neighboring X-ray lines, whereas all lines can normally be separated by a WDS. An EDS must work with lower probe currents because all quanta are counted and only a count rate of ca. 10 000 cps guarantees no overlap of sequential pulses. Pulse-rejection electronics are used when the time between two pulses is shorter than the pulse-shaping time necessary for the MCA.

2.3.5.3. *X-Ray Microanalysis*

For a quantitative XRMA [246], the number n_a of counted characteristic X-ray quanta of an element a of concentration c_a in an alloy or compound is compared with the number n_s from a pure elemental standard (c_s = 1) or a compound of known composition. Only an approximate value for c_a can be obtained from the k ratio

$$k = n_a/n_s \approx c_a/c_s \tag{2.40}$$

The exact measurement of concentration needs a *ZAF correction method* (see Quantitative Analysis.). This considers the differences in the stopping power and the backscattering of electrons between specimen and standard (*Z correction*), in the depth distribution of X-ray generation and *absorption* (*A*) of X-rays recorded under a takeoff angle of 40 – 60°, and in the excitation of X-ray quanta of the element of interest by characteristic X-rays and continuum of the matrix (*fluorescence F*).

Special correction programs must be used, for example, for tilted specimens, thin film coatings, small particles, and biological specimens.

2.3.5.4. *Special X-Ray Techniques*

When the specimen is scanned in a raster or along a line, the pulses of selected characteristic X-ray lines can produce an elemental distribution profile or map, respectively.

In a *single crystal*, the excited X-rays are Bragg diffracted at the lattice plane, and their isotropic angular characteristics show defect and excess Kossel lines of apex angle 90° – θ_B, which can be used for accurate measurement of lattice parameters and strains when the Kossel pattern is recorded on a photographic emulsion.

X-ray fluorescence analysis can be applied in an SEM when generating X-rays in a thin foil, which stops the electrons, but the transmitted X-rays can excite X-ray fluorescence in the specimen, which is recorded by an energy-dispersive spectrometer. The background of the spectrum is much lower than that from direct electron excitation (Fig. 89, curve X versus curve E), which results in a better signal-to-background ratio for trace elements. However, only larger areas 0.1 – 1 mm in diameter can be analyzed.

The small source of X-rays generated by an electron probe on a bulk target can also be used for X-ray projection microscopy.

[Top of Page]

References

1. E. Abbe: “Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung,” *Arch. Mikrosk. Anat.* **9** (1873) 413.
2. D. Gerlach: *Das Lichtmikroskop*, 2nd ed., Thieme Verlag, Stuttgart 1985.
3. H. Gundlach: *Mikroskope. Handbuch Biomedizinische Technik*, vol. **4**, Springer Verlag, TÜV Rheinland, (1991) pp. 1 – 20.
4. K. Michel: “Axiomat von Zeiss, ein Mikroskop mit neurem Konzept,” *Zeiss. Inf.* **20** (1973).
5. F. Muchel: “ ‘ICS’-A New Principle in Optics,” *Zeiss Inf.* **30** (1988) 20 – 27.
6. K. Michel: *Die Grundzüge der Theorie des Mikroskops*, 3rd ed., Wissenschaftliche Verlagsgesellschaft, Stuttgart 1981.
7. A. Köhler: “Mikrophotographische Untersuchungen mit ultraviolettem Licht,” *Z. Wiss. Mikrosk. Mikrosk. Tech.* **21** (1904) 129 – 165, 273 – 404.
8. A. Köhler: “Ein neues Beleuchtungsverfahren für mikrophotographische Zwecke,” *Z. Wiss. Mikrosk. Mikrosk. Tech.* **10** (1893) 433 – 440.
9. F. Zernike: “Das Phasenkontrastverfahren bei der mikroskopischen Beobachtung,” *Phys. Z.* **36** (1935) 848.
10. H. Gundlach: “Neuere Entwicklungen und Anwendungen in der Lichtmikroskopie,” *Gynäkologie* **23** (1990) 328 – 335.
11. G. Nomarski: “Microinterferometre Differentielle a Ondes Polarisees,” *J. Phys. Radium* **16** (1955) 9.
12. R. D. Allen *et al.*: “The Zeiss Nomarski Differential Interference Equipment for Transmitted-Light Microscopy,” *Z. Wiss. Mikrosk. Mikrosk. Tech.* **69** (1969) 193 – 221.
13. W. Lang: “Nomarski Differential Interference Contrast Microscopy,” *Zeiss Inf.* **16** (1968) no. 70, 114 –120.
14. M. Francon: *Einführung in die neuen Methoden der Lichtmikroskopie*, Verlag Braun, Karlsruhe 1967.
15. W. I. Patzelt: “Reflexionskontrast. Eine neue lichtmikroskopische Technik,” *Mikrokosmos* **3** (1977) 78 – 81.
16. S. Pentz *et al.*: “Darstellung des Adhäsionsverhaltens kultivierter Leberzellen an Glasoberflächen während der Mitose durch Reflexionskontrast-Mikroskopie,” *Zeiss Inf.* **25** (1980) 91, 41 – 43.
17. H. Gundlach *et al.*: “Immunfluorescance Microscopy with the New Zeiss Photomicroscope Axiophot,” *Zeiss Inf.* **29** (1997) 36 – 39.
18. A. J. Lacey: *Light Microscopy in Biology*, IRL Press, Oxford University, 1989.
19. D. L. Taylor, M. Nederlof, F. Lanni, A. S. Waggoner: “The New Vision of Light Microscopy,” *Am. Sci.* **80** (1992) 322 – 335.
20. D. L. Taylor *et al.*: *Fluorescence Microscopy of Living Cells in Culture*, Academic Press, San Diego 1989.
21. A. H. Coons *et al.*: “Immunological Properties of an Antibody Containing a Fluorescent Group,” *Proc. Soc. Exp. Biol. Med.* **47** (1941) 200 – 202.
22. J. S. Ploem: “The Use of a Vertical Illuminator with Interchangeable Dichroic Mirrors for Fluorescence Microscopy with Incident Light,” *Z. Wiss. Mikrosk. Mikrosk. Tech.* **68** (1967) 129 – 143.
23. L. Trapp: “Über Lichtquellen und Filter für die Fluoreszenzmikroskopie und über die Auflichtfluoreszenzmethode bei Durchlichtpräparaten,” *Acta Histochemica* **7** (1965), pp. 327 – 338.
24. W. W. Franke: “Different Intermediate-Sized Filaments Distinguished by Immunofluorescence Microscopy,” *Proc. Nat. Acad. Sci. USA* **75** (1978) no. 10, 5034 – 5038.
25. O. Greulich *et al.*: “The Light Microscope on its Way from an Analytical to a Preparative Tool,” *J. Microsc. (Oxford)* **167** (1992) 127 – 151.
26. P. Lichter *et al.*: “Analysis of Genes and Chromosomes by Nonisotopic in Situ Hybridization,” *Gata* **8** (1991) no. 1, 24 – 35.
27. Chr. Lengauer et al., “Chromosomal Barcodes produced by Multifluorescence in situ Hybridization with Multiple YAG Clones and Whole Chromosomes painting probes,” in *Human Molecular Genetics,* 5th ed., vol. 2, Oxford University Press, pp. 505 – 512.
28. D. M. Shotton: “Confocal Scanning Optical Microscopy and its Applications for Biological Specimens,” *J. Cell. Sci.* **94** (1989) 175 – 206.

29. J. N. Turner: “Confocal Light Microscopy. Biological Applications (Special Issue),” *Electron Microsc. Tech.* **18** (1991) 1.
30. R. P. Haugland: *Molecular Probes: Handbook of Fluorescent Probes and Research Chemicals*, Eugene Molecular Probes Inc., 1989.
31. R. Y. Tsien, A. Waggoner: “Fluorophores for Confocal Microscopy: Photophysics and Photochemistry,” in [73], pp. 169 – 178.
32. M. Chalfie, Y. Tu, D. C. Prasher, “Green Fluorescent Protein as a Marker for Gene Expression,” *Science* **263** (1994) 802 – 805. Links
33. R. Heim, R. Y. Tsien, “Engineering Green Fluorescent Protein for Improved Brightness, Longer Wavelength and Fluorescence Resonance Energy Transfer,” *Curr. Biol.* **6** (1996) 178 – 182. Links
34. T. M. Jovin: “Quantum Dots Finally Come of Age,” *Nat Biotechnol.* **21** (2003 no. 1):, 32 – 33. Links
35. B. R. Goldman: “Multiplexed Toxin Analysis using Four Colors of QuantumDot Fluoroagents,” *Anal. Chem.* **76** (2004) 684 – 688. Links
36. B. Ballou, et al.:, “Noninvasive imaging of quantum dots in mice,” *Bioconj. Chem.* **15** (2004), vol. **15**, pp. 79 – 86.
37. W. J. Parak, et al.:, “Conjugation of DNA to Silanized ColloidalSemiconductor Nanocrystalline Quantum Dots,” *Chem. Mater.* **14** (2002) 113 – 2119.
38. B. Herman, *Fluorescence Microscopy*, BIOS Scientific Publisher, Oxford, 1998.
39. E. A. Jares-Erijman, T. M. Jovin:, “FRET Imaging”. *Nat. Biotechnol.* **21** (2003) no. 11, 1387 – 1395. Links
40. E. L.Elson, D.Magde, “Fluorescence Correlation Spectroscopy (I). Conceptual Basis and Theory,” *Biopolymers* **13** (1974) 1 – 27. Links
41. R. Rigler, J. Widengreen, Ü. Mets, “Interactions and Kinetics of Single Molecules as Observed by Fluorescence Correlation Spectroscopy,” *Fluorescence Spectroscopy — New Methods and Applications*, Springer, Heidelberg, 1992, 13 – 24.
42. M. Eigen, R. Rigler, “Sorting Single Molecules: Application to Diagnostics and Evolutionary Biotechnology,” *Proc. Natl. Acad. Sci. USA* **91** (1994) 5740 – 5747. Links
43. H. Utermöhl: “Neue Wege in der quantitativen Erfassung des Planktons,” *Verh. intern. Verein Limnol.* **5** (1931) 567 – 596.
44. H. Gundlach et al.: “Mikrokinematographie mit dem Mikroskop “Axiomat,” *Res. Film* **9** (1976) no. 1, 30 – 36.
45. H. Gundlach: *Die Anwendung von inversen Mikroskopen in der Zell- und Entwicklungsbiologie*, Suppl. GIT, Verlag E. Giebler, Darmstadt 1981, pp. 9 – 15.
46. H. A. Tritthart *et al.*: “The Zeiss Axiomat Applied to the Examination of the Morphology and Physiology of Myocardial Cells in Culture,” *Zeiss Inf.* **25** 1980 no. 90, 10 – 13.
47. M. Horster *et al.*: “Application of Differential Interference Contrast with Inverted Microscopes to the In-vitro Perfused Nephron,” *J. Microsc. (Oxford)* **117** (1979) 375 – 379.
48. F. K. Möllring: “Inverse Mikroskopie IM 35 und ICN 405 für Biologie, Medizin und Metallographie,” *Zeiss Inf.* **23** (1977) 18 – 19.
49. H. Piller: *Microscope Photometry*, Springer, Berlin 1977.
50. H. Gundlach *et al.*: “Identification and Selective Micropreparation of Live Nuclear Components with the Zeiss IM 35 Inverted Microscope,” *Zeiss Inf.* **25** (1980) no. 91, 36 – 40.
51. H. Spring *et al.*: “DNA Contents and Numbers of Nucleoli and Pre-rRNA-Genes in Nuclei of Gemetes and Vegetative Cells of Acetabularia Mediterranea,” *Exp. Cell. Res.* **114** (1978) 203 – 215. Links
52. W. Ansorge *et al.*: “Performance of an Automated System for Capillary Microinjection into Living Cells,” *J. Biochem. Biophys. Methods* **16** (1988) 283 –292. Links
53. “Single Molecule Biochemistry Using Optical Tweezers.” *FEBS Lett.* **430** (1998) no. 1–2:, 23 – 27. Links
54. R. D. Allen *et al.*: “Video-Enhanced-Contrast Polarization (AVEC-Pol) Microscopy: A New Method Applied to the Detection of Birefringence in the Motile Reticulopodial Network of Allogromia Laticollaris,” *Cell Motil.* **1** (1981) 275 – 289.
55. R. D. Allen *et al.*: “Video-Enhanced-Contrast, Differential Interference Contrast (AVEC-DIC) Microscopy: A New Method Capable of Analyzing Microtubule-Related Motility in the Reticulopodial Network of Allogromia Laticollaris,” *Cell Motil.* **1** (1981) 291 – 302.
56. S. Inoue: “Foundations of Confocal Scanning Imaging in Confocal Microscopy” in [73], pp. 1 – 14.
57. S. Inoue *et al.*: “The Acrosonal Reaction of Thyone Sperm Head Visualized by High Resolution Video Microscopy,” *J. Cell Biol.* **93** (1982) 812 –819. Links
58. J. H. Hayden *et al.*: “Detection of Single Microtubules in Living Cells: Particle Transport Can Occur in Both Directions Along the Same Microtubule,” *J. Cell Biol.* **99** (1984) 1785 – 1793. Links
59. T. Salmon *et al.*: “Video-Enhanced Differential Interference Contrast Light Microscopy,” *Bio Techniques* **7** (1989) 624 – 633.
60. D. J. Arndt-Jovin *et al.*: “Fluorescence Digital Imaging Microscopy in Cell Biology,” *Science (Washington D.C.)* **230** (1985) 247 – 256.
61. Y. Hiraoka *et al.*: “The Use of a Charge-Coupled Device for Quantitative Optical Microscopy of Biological Structures,” *Science (Washington D.C.)* **238** (1987) 36 – 41.
62. H. H. Sedlacek *et al.*: “The Use of Television Cameras Equipped with an Image Intensifier in the Immunofluorescence Microscopy,” *Behring Inst. Mitt.* **59** (1976) 64 – 70.
63. H. Gundlach *et al.*: *Electronic Photography-Technology, Systems and Applications in Microbiology. Proc. of Int. Symposium on Electronic Photography*, Cologne The Society for Imaging Science and Technology, Springfield, Virginia USA, 1992, pp. 60 –65.

64. M. Minsky: US 3013467, 1961.
65. G. Cox: “Photons Under the Microscope. Marvin Minsky – the Forgotten Pioneer?,” *Australian EM Newsletter* **38 no. 4**, (1993) 4 – 10.
66. C. J. R. Sheppard: “15 Years of Scanning Optical Microscopy at Oxford,” *Proc. Roy. Mic. Soc.* **25** (1990) 319 – 321.
67. M. Petran, M. Hadravsky, M. D. Egger, R. Galambos: “Tandem-Scanning Reflected Light Microscope,” *J. Opt. Soc. Am.* **58** (1968) 661 – 664.
68. G. J. Brakenhoff, P. Blom, P. Barends: “Confocal Scanning Light Microscopy with High Aperture Immersion Lens,” *J. Micros. (Oxford)* **117** (1979) 219 –232.
69. T. Wilson: “Optical Sectioning in Confocal Fluorescence Microscopy,” *J. Microsc. (Oxford)* **154** (1989) 143 – 156.
70. C. J. R. Sheppard, A. Choundhury: “Image Formation in the Scanning Microscope,” *Opt. Acta* **24** (1977) 1051 – 1073.
71. C. J. R. Sheppard, T. Wilson: “Depth of Field in the Scanning Microscope,” *Opt. Lett.* **3** (1978) 115 – 117.
72. P. Wallen, K. Carlsson, K. Mossberg: “CLSM as a Tool for Studying the 3-D Morphology of Nerve Cells,” in [119], pp. 110 – 143.
73. J. Pawley: *Handbook of Biological Confocal Microscopy*, 2nd ed. Plenum Press, New York 1995.
74. S. Hell, E. Lehtonen, E. H. K Stelzer: “Confocal Fluorescence Microscopy: Wave Optics and Applications to Cell Biology,” in [32], pp. 145 – 160.
75. I. J. Cox, C. J. R. Sheppard, T. Wilson: “Superresolution in Confocal Fluorescent Microscopy,” *Optik (Stuttgart)* **60** (1982) 391 – 396.
76. T. Wilson (ed.): *Confocal Microscopy*, Academic Press, London 1990.
77. C. J. R. Sheppard, M. Gu: “3-D Transfer Functions in Confocal Scanning Microscopy,” in [119], pp. 251 – 280.
78. M. Gu, *Principles of 3D Imaging in Confocal Microscopes*, World Scientific, Singapore, 1996.
79. C. J. Cogswell, J. W. O'Bryan: “A High Resolution Confocal Transmission Microscope,” *SPIE Proceedings* **1660** (1992) 503 – 511.
80. G. J. Brakenhoff, K. Visscher: “Bilateral Scanning and Array Detectors,” *J. Micros. (Oxford)* **165** (1990) 139 – 146.
81. E. Gratten, M. J. van der Veen: “Laser Sources for Confocal Microscopy,” in [73], pp. 69 – 98.
82. K. S. Wells, D. R. Sandison, J. Strickler, W. W. Webb: “Quantitative Fluorescence Confocal Laser Scanning Microscopy,” in [73], pp. 39 – 53.
83. J. Art: “Photon Detectors for Confocal Microscopy,” in [73], pp. 127 – 139.
84. W. Denk, J. H. Strickler, W. W. Webb: “Two-photon laser scanning fluorescence microscopy,” *Science (Washington D.C.)* **248** (1990) 73 – 76.
85. M. Gu, X. Gan, A. Kisteman, M. G. Xu, *Appl. Phys. Lett.* **77** (2000) no. 10, 1551 – 1553. Links
86. S. Lindek, E. H. K. Stelzer, S. W. Hell: “Two New High-Resolution Confocal Fluorescence Microscopies (4pi, theta) with One- and Two-Photon Excitation,” in [73], pp. 445 – 458.
87. S. W. Hell: “Toward Fluorescence Nanoscopy”. *Nat. Biotechnol.* **21** (2003) no. 11, 1347 – 1355. Links
88. J. G. Fujimoto: “Optical Coherence Tomography for Ultrahigh Resolution in vivo Imaging”. *Nat. Biotechnol.* (2003) no. 11, 1361 – 1367. Links
89. J. Huisken *et al.*, “Optical Sectioning Deep Inside Live Embryos by Selective Plane Illumination Microscopy”. *Science*. **305** (2004) 1007 – 1009. Links
90. M. D. Egger, M. Petran: “New Reflected Light Microscope for Viewing Unstained Brain and Ganglian Cells,” *Science (Washington D.C.)* **157** (1987) 305 –307.
91. J. S. Deitch, K. L. Smith, J. W. Swann, J. N. Turner: “Parameters Affecting Imaging of HRP Reaction Product in the Confocal Scanning Laser Microscope,” *J. Microsc. (Oxford)* **160** (1990) 265 – 278.
92. J. P. Rigaut, S. Caravajal-Gonzalez, J. Vassy: “3-D Image Cytometry,” in [119], pp. 205 – 237.
93. B. R. Masters, A. Kriete, J. Kukulies: “Ultraviolet Confocal Fluorescence Microscopy of the In-Vitro Cornea: Redox Metabolic Imaging,” *Applied Optics* **32** (1993) no. 4, 592 – 596.
94. A. Villringer, U. Dirnagl, K. Einhäupl: “Microscopical Visualization of the Brain in Vivo,” in [119], pp. 161 – 181.
95. R. Yuste, F. Lanni, A. Konnerth (eds.): *Imaging Neurons: A Laboratory Manual*, CSHL Press, Cold Spring Harbor, 2000.
96. B. R. Masters: “Confocal Ocular Microscopy – a New Paradigm for Ocular Visualization,” in [68], pp. 183 – 203.
97. W. A. Mohler, J. G. White, “Stereo-4D Reconstruction and Animation from Living Fluorescent Specimens,” *Biotechniques* **24** (1998) 1006 – 1012. Links
98. R. G. Summer, P. C. Cheng: “Analysis of Embryonic Cell Division Patterns Using Laser Scanning Confocal Microscopy,” in G. W. Bailey (ed.): *Proc. Annual meeting Electron Microsc. Soc. Am.* **47** (1989) 140 – 141.
99. N. O'Rourke, S. E. Fraser: “Dynamic Changes in Optic Fiber Terminal Arbors Lead to Retinotopic Map Formation: An In-Vivo Confocal Microscopic Study,” *Neuron* **5** (1990) 159 – 171. Links
100. A. Kriete, H.-J. Wagner, “Computerized Spatio-temporal (4D) Representation in Confocal Microscopy: Application to Neuroanatomical Plasticity,” *J. Micros. (Oxford)* **169** (1993) 27 – 31.
101. A. Kriete, N. Klein, L. C. Berger, “Data Compression in Microscopy: a Comparative Study,” *Proc. SPIE.* **3605** (1999) 158 – 168.
102. M. W. Tengowski: “Image Compression in Morphometry Studies Requiring 21 CFR Part 11 Compliance: Procedure is Key with TIFFs and Various JPEG Compression Strengths,” *Toxicol. Pathol.* **32** (2004), 258 – 263.
103. I. T. Young, “Characterizing the Imaging Transfer Function,” in D. L. Taylor, Y. L. Wang (eds.): *Methods in Cell Biology*, Academic Press, San Diego, 1989, pp. 1 – 45.

104. R. A. Jarvis, “Focus Optimization Criteria for Computer Image Processing,” *The Microscope* **24** (1976) 163 – 180.
105. A. Kriete, “Image Quality Considerations in Computerized 2D and 3D Microscopies,” in P. C. Cheng, T. H. Lin, W. H. Wu, J. L. Wu (eds.): *Multidimensional Microscopy*, Springer, New York, 1990, pp. 141 – 150.
106. I. T. Young, Quantitative Microscopy, IEEE Eng. in Medicine and Biology, Jan./Feb. 1996, pp. 59 – 66.
107. T. Visser, J. L. Oud, G. T. Brakenhoff: “Refractive Index and Axial Distance Measurements in 3-D Microscopy,” *Optik (Stuttgart)* **90** (1991) 17 – 19.
108. J.-A. Conchello, E. W. Hanssen: “Enhanced 3-D Reconstruction from Confocal Scanning Microscope Images. Deterministic and Maximum Likelihood Reconstrauctions,” *App. Opt.* **29** (1990) no. 26, 3795 – 3804.
109. T. J. Holmes, Y.-H. Liu: “Image Restoration for 2-D and 3-D Fluorescence Microscopy,” in [119], pp. 283 – 323.
110. D. A. Agard, J. W. Sedat: “Three-Dimensional Architecture of Polytene Nucleus,” *Nature (London)* **302** (1983) 676 – 681.
111. G. Wang, W. S. Liou, T. H. Lin, P. C. Cheng: “Image Restoration in Light Microscopy,” in T. H. Lin, W. L. Wu, J. L. Wu (eds.): *Multidimensional Microscopy,* Springer Verlag, New York 1993, 191 –208.
112. F. Meyer: “Mathematical Morphology: 2-D to 3-D,” *J. Microsc.* **165** (1992) 5 – 28.
113. P. C. Cheng et al.: “3-D Image Analysis and Visualization in Light Microscopy and X-Ray Micro-Tomography,” in [119], pp. 361 – 398.
114. K. Mossberg, U. Arvidsson, B. Ulfhake: “Computerized Quantification of Immunoflourescence Labelled Axon Terminals and Analysis of Co-Localization of Neurochemicals in Axon Terminals with A Confocal Scanning Laser Microscope,” *J. Histochem Cyctochem.* **38** (1990) 179 – 190.
115. B. Roysam *et al.*: “Unsupervised Noise Removal Algorithms for 3-D Confocal Fluorescence Microscopy,” *Micron and Micros. Acta* **23** (1992) in press.
116. M. Montag *et al.*: “Methodical Aspects of 3-D Reconstruction of Chromatin Architecture in Mouse Trophoblast Giant Nuclei,” *J. Microsc. (Oxford)* **158** (1990) 225 – 233.
117. V. Conan *et al.*: “Geostatistical and Morphological Methods Applied to 3-D Microscopy,” *J. Microsc. (Oxford)* **166** (1992) 169 – 184.
118. N. S. White: “Visualization Systems for Multidimensional CLSM Systems,” in [73], pp. 211 – 254.
119. A. Kriete: *Visualization in Biomedical Microscopies 3-D Imaging and Computer Applications*, VCH-Verlagsgesellschaft, Weinheim, Germany 1992.
120. R. Eils, C. Athale:, “Computational Imaging in Cell Biology,” *J. Cell Biol.* **161** (2003) 477 – 481. Links
121. H. Chen, J. W. Sedat, J. A. Adard: “Manipulation, Display and Analysis of Three-Dimensional Biological Images,” in J. Pawley (ed.): *Handbook of Biological Confocal Microscopy,* Plenum Press, New York 1990, pp. 141 – 150.
122. D. P. Huijsmanns, W. H. Lamers, J. A. Los, J. Strackee: “Toward Computerized Morphometric Facilities: a Review of 58 Software Packages for Computer-Aided 3-D Reconstruction, Quantification, and Picture Generation from Parallel Serial Sections,” *Anat. Rec.* **216** (1986) 449 – 470. Links
123. A. Kriete, P. C. Cheng (eds.): “3-D Microscopy (Special Issue),” *Computerized Medical Imaging and Graphics*, vol. 17, Plenum Press, N.Y. 1993, p. 8.
124. R. Ware, V. LoPresti: “Three-Dimensional Reconstruction from Serial Sections,” *Comp. Graph.* **2** (1975) 325 – 440.
125. J. K. Udupa, H. M. Hung: “Surface Versus Volume Rendering: a Comperative Assessment,” *VBC '90*, Proceedings IEEE, Atlanta 1990, pp. 83 – 91.
126. S. D. Roth: “Ray-Casting for Solid Modeling,” *Comput. Graphics Image Processing* **18** (1982) 109 –144.
127. H. K. Tuy, L. T. Tuy: “Direct 2-D Display of 3-D Objects,” *IEEE CG & A* **4** (1984) 29 – 33.
128. G. Frieder, D. Gordon, R. A. Reynolds: “Back-to-Front Display of Voxel-Based Objects,” *IEEE CG & A* **5** (1985) no. 1, 52 – 60.
129. R. A. Debrin, L. Carpenter, P. Hanrahan: “Volume Rendering,” *Computer Graphics* **22** (1988) no. 4, 65 – 74.
130. M. Levoy: “A Hybrid Ray-Trace for Rendering Polygon and Volume Data,” *IEEE CG & A* **3** (1990) 33 – 40.
131. H. T. M. van der Voort, G. J. Brakenhoff, M. W. Baarslag: “Three-Dimensional Visualization Methods,” *J. Microsc. (Oxford)* **153** (1989) no. 2, 123 – 132.
132. P.-O. Forsgren: “Visualization and Coding in Three-Dimensional Image Processing,” *J. Microsc. (Oxford)* **159** (1990) no. 2, 195 – 202.
133. M. Levoy *et al.*: “Volume Rendering in Radiation Treatment Planning,” *VBC '90*, Proceedings IEEE, Atlanta 1990, pp. 4 – 10.
134. R. S. Weinstein *et al.*: “Telepathology Overview: From Concept to Implementation.” *Hum. Pathol.* **32** (2001) 1283 – 1299 Links
135. J. Kononen *et al.*: “Tissue Microarrays for High-Throughput Molecular Profiling of Tumor Specimens.” *Nat. Med.* **4** (1998) 844 – 847. Links
136. T. Braunschweig, J.-Y. Chung, S. M. Hewitt “Perspectives in Tissue Microarrays.” *Combinatorial Chem. High Throughput Screening* **7** (2004) 575 – 586.
137. R. Srinivasan, S. Q. Yao, D. S. Yeo “Chemical Approaches for Live Cell Bioimaging,” *Combinatorial Chem. High Throughput Screening* (2004) 597 – 604.
138. D. L. Taylor, E. S. Woo, K. A. Giuliano: “Real-Time Molecular and Cellular Analysis: The New Frontier of Drug Discovery,” *Curr. Opin. Biotechnol.* **12** (2001) 75 – 81. Links
139. A. Dove:, “Screening for Content – The Evolution of High Throughput,” *Nature Biotech.* **21** (2003) 859 – 864.
140. Z. E. Perlman “Multidimensional Drug Profiling by Automated Microscopy.” *Science* **306** (2004), **306**: 1194 – 1198. Links

General References

141. S. Amelinckx: “The Direct Observation of Dislocations,” Suppl. 6 in F. Seitz, D. Turnbull (eds.): *Solid State Physics,* Academic Press, London 1964.
142. S. Amelinckx, R. Gevers, J. Van Landuyt (eds): *Diffraction and Imaging Techniques in Material Science*, North-Holland Publishing Company, Amsterdam 1970, 1978.
143. F. R. N. Nabarro (ed.): *Dislocation in Solids*, North-Holland, Amsterdam 1979.
144. P. B. Hirsch, R. B. Nicholson, A. Howie, D. W. Pashley, M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths, London 1965.
145. H. Bethge, J. Heydenreich (eds.): *Electronenmikroskopie in der Festkörperphysik*, Springer Verlag, Berlin 1982.
146. J. C. H. Spence: “Experimental High Resolution Electron Microscopy,” *Monographs on the Physics and Chemistry of Materials*, Oxford Science Publications, Clarendon Press, Oxford 1981.
147. G. Thomas: *Transmission Electron Microscopy of Metals*, John Wiley and Sons, New York 1962.
148. J. C. H. Spence, J. M. Zuo: *Electron Microdiffraction*, Plenum Press, New York 1992.
149. R. W. Cahn, P. Haasen, E. J. Kramer (eds.): *Materials Science and Technology,* vol. 2 A, VCH Verlagsgesellschaft, Weinheim 1992.
150. D. C. Joy, A. D. Romig, Jr., J. I. Goldstein (eds.): *Principles of Analytical Electron Microscopy*, Plenum Press, New York 1986.
151. L. Reimer: “Transmission Electron Microscopy,” *Springer Series in Optical Sciences*, 4th ed., Springer Verlag, Berlin 1997.
152. J. M. Cowley (ed.): “Electron Diffraction Techniques,” vols. 1 and 2, *Monographs on Crystallography 3,* International Union of Crystallography, Oxford University Press, Oxford 1992.
153. P. G. Merli, M. Vittori Antisari (eds.): *Electron Microscopy in Materials Science*, World Scientific, Singapore 1992.

Specific References

154. S. Amelinckx, D. Van Dyck, J. Van Landuyt, G. Van Tendeloo (eds.): *Handbook of Microscopy*, VCH, Weinheim, Germany 1997.
155. N. F. Mott, H. S. W. Massey: *The Theory of Atomic Collisions*, Clarendon Press, Oxford 1949.
156. J. A. Ibers, B. K. Vainshtein: “Scattering Amplitudes for Electrons,” in K. Londsdale (ed.): *International Tables for X-Ray Crystallography*, vol. **3**, Kynoch, Birmingham 1962.
157. in [149], p. 250.
158. W. L. Bragg, *Nature (London)* **124** (1929) 125.
159. P. P. Ewald, *Ann. Phys. (Leipzig)* **54** (1917) 519.
160. in [149], p. 251.
161. S. Takagi, *Acta Crystallogr*. **15** (1962) 1311. Links
162. R. D. Heidenreich, *J. Appl. Phys.* **20** (1949) 993. Links
163. S. Kikuchi, *Jpn. J. Phys.* **5** (1928) 23.
164. in [141], p. 125.
165. H. A. Bethe, *Ann. Phys. (Leipzig)* **87** (1928) 55. C. H. MacGillavry, *Physica (Amsterdam)* **7** (1940) 329.
166. A. Howie, M. J. Whelan, *Proc. Roy. Soc. London A*, **263** (1961) 217. A. Howie, M. J. Whelan, *Proc. Roy. Soc. London A*, **267** (1962) 206.
167. H. Yoshioka, *J. Phys. Soc. Jpn.* **12** (1957) 628. H. Hashimoto, A. Howie, M. J. Whelan, *Proc. Roy. Soc. London A* **269** (1962) 80.
168. G. Bormann, *Z. Phys.* **42** (1941) 157. G. Bormann, *Z. Phys.* **127** (1950) 297.
169. H. Hashimoto, M. Mannami, T. Naiki, *Philos. Trans. R. Soc. London A* **253** (1961) 459. J. W. Menter, *Proc. Roy. Soc. London A* **236** (1956) 119.
170. G. Van Tendeloo, S. Amelinckx, *Acta Crystallogr. Sect. A: Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **A 30** (1974) 431.
171. S. Amelinckx, J. Van Landuyt, in [142], p. 107. H. Hashimoto, M. J. Whelan, *J. Phys. Soc. Jpn.* **18** (1963) 1706. P. B. Hirsch, A. Howie, M. J. Whelan, *Philos. Trans. R. Soc. London A* **252** (1960) 499. P. B. Hirsch, A. Howie, M. J. Whelan, *Philos. Mag.* **7** (1962) 2095.
172. C. M. Drum, M. J. Whelan, *Philos. Mag.* **11** (1965) 205. J. Van Landuyt, R. Gevers, S. Amelinckx, *Phys. Status Solidi* **7** (1964) 519.
173. G. A. Bassett, J. W. Menter, D. W. Pashley, *Proc. Roy. Soc. London A* **246** (1958) 345. D. W. Pashley, J. W. Menter, G. A. Bassett, *Nature (London)* **179** (1957) 752.
174. H. Hashimoto, R. Uyeda, *Acta Crystallogr.* **10** (1957) 143.
175. R. Gevers, *Philos. Mag.* **7** (1963) 769.
176. D. J. H. Cockayne, I. L. E. Ray, M. J. Whelan, *Philos. Mag.* **20** (1969) 1265. D. J. H. Cockayne, M. J. Jenkins, I. L. E. Ray, *Philos. Mag.* **24** (1971) 1383. R. de Ridder, S. Amelinckx, *Phys. Status Solidi B* **43** (1971) 541.
177. S. Amelinckx, P. Delavignette, *J. Appl. Phys.* **33** (1962) 1458.
178. P. Humble, in [142], p. 315. P. Humble, *Aust. J. Phys.* **21** (1968) 325. A. K. Head, *Aust. J. Phys.* **20** (1967) 557. A. K. Head, P. Humble, L. M. Clarebrough, A. T. Morton, G. T. Forwood: “Computed Electron Micrographs and Defect Identification,” in: S. Amelinckx, P. Gevers, J. Nihoul (eds.): *Defects in Crystalline Solids*, vol. **7** , North-Holland, Amsterdam 1973.
179. [149], p. 58.
180. C. Boulesteix, J. Van Landuyt, S. Amelinckx, *Phys. Status Solidi A* **33** (1976) 595. Links
181. G. Van Tendeloo, J. Van Landuyt, S. Amelinckx, *Phys. Status Solidi A* **33** (1976) 723. Links M. Snijkers, R.

Serneels, P. Delavignette, R. Gevers, S. Amelinckx, *Chryst. Lattice Defects* **3** (1972) 99.

182. R. Gevers, J. Van Landuyt, S. Amelinckx, *Phys. Status Solidi* **11** (1965) 689. M. J. Goringe, U. Valdré, *Proc. R. Soc. London A* **295** (1966) 192.
183. M. F. Ashby, L. M. Brown, *Philos. Mag.* **8** (1963) 1083, 1649.
184. J. W. Steeds, E. Carlino, in [153], p. 279.
185. P. Goodman, *Acta Crystallogr. Sect. A: Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **31** (1975) 793; *Acta Crystallogr. Sect. A: Cryst. Phys. Diffr. Theor. Gen. Crystallogr.* **31** (1975) 804. M. Tanaka, *J. Electron Microsc. Tech.* **13** (1989) 27. Links J. W. Steeds, R. Vincent, *J. Appl. Crystallogr.* **16** (1983) 317. Links
186. B. F. Buxton, J. A. Eades, J. W. Steeds, G. M. Rackham, *Philos. Trans. R. Soc. London* **A 281**, 171. J. Eades, M. Shannon, B. Buxton in O. Johari (ed.): *Scanning Electron Microscopy*, I.I.I.R. Institute, Chicago 1983, p. 83.
187. M. Tanaka, M. Terauchi, T. Kaneyama: *Convergent Beam Electron Diffraction I and II*, Jeol Ltd., Tokyo.
188. N. S. Blom, E. W. Schapink, *J. Appl. Crystallogr.* **18** (1985) 126. Links
189. J. Gjønnes, A. F. Moodie, *Acta Crystallogr.* **19** (1965) 6567. P. Goodman, G. Lempfuhl, *Z. Naturforsch. A* **19** (1964) 818.
190. W. Kossel, G. Möllestedt, *Ann. Phys. Leipzig* **36** (1939) 113. C. H. MacGillavray, *Physica (Amsterdam)* **7** (1940) 329.
191. G. R. Grinton, J. M. Cowley, *Optik (Stuttgart)* **34** (1971) 221. J. M. Cowley, S. Iijima: "The Direct Observation of Crystal Structures," in H. R. Wenk (ed.): Electron Microscopy in Mineralogy, Springer Verlag, Berlin 1976, p. 123.
192. S. Amelinckx in [149], p. 5.
193. O. Scherzer, *J. Appl. Phys.* **20** (1949) 20. Links
194. E. Abbe, *Arch. Mikrosk. Anat.* **9** (1873) 413.
195. D. van Dyck, M. Op de Beeck, *Proc. Int. Conf. Electron Microsc.* **12th** (1990) vol. 1, 64. D. van Dyck, W. Coene, *Ultramicroscopy* **15** (1989) 29. D. van Dyck, M. Op de Beeck, *Proc. Int. Conf. Electron Microsc.* **12th** (1990) 26.
196. D. van Dyck: "High Resolution Electron Microscopy" in [154] vol. I, Chap. 1.1.2, p. 353.
197. J. M. Cowley, A. F. Moodie, *Acta Crystallogr.* **10** (1957) 609. Links
198. D. van Dyck *et al.*in W. Krakow, M. O'Keefe (eds.): *Computer Simulation of Electron Microscope Diffraction and Images*, The Minerals, Metals and Materials Society, 1989, p. 107.
199. J. M. Cowley: "Scanning Transmission Electron Microscopy" in [154] vol. II, Chap. 2.2, p. 563.
200. S. J. Pennycook, D. E. Tesson, P. D. Nellist, M. F. Chisholm, N. D. Browning: "Scanning Transmission Electron Microscopy: Z-Contrast" in [154] vol. II, Chap. 2 – 3, p. 595.
201. H. G. J. Moseley, *Philos. Mag.* **26** (1913) 1024; *Philos. Mag.* **27** (1914) 703.
202. E. L. Hall, in [149], p. 147.
203. M. H. Jacobs, J. Baborovska: *Electron Microscopy*, The Institute of Physics, London 1972, p. 136. K. F. J. Heinrich: *Electron Beam X-Ray Microanalysis*, Van Rostrand, New York 1981.
204. S. J. B. Reed: *Electron Microprobe Analysis*, Cambridge University Press, London 1975.
205. J. C. Spence, H. Tafto, *J. Microsc. (Oxford)* **130** (1983) 147.
206. C. Colliex: "Electron Energy Loss Spectrometry Imaging" in [154], vol. I, Chap. 1.3, p. 425.
207. B. Jouffrey, in [153], p. 363.
208. R. D. Leapman, L. A. Grunes, P. L. Fejes, J. Silcox, in B. K. Teo, D. C. Joy (eds.): *EXALFS Spectroscopy*, Plenum Press, New York 1981, p. 217.
209. G. Zanchi, J. P. Perez, J. Sevely, *Optik* **43** (1945) 495. W. Pejas, H. Rose, *Electron Microscopy*, Microscopic Society of Canada, Toronto 1978, p. 44. H. T. Pearce-Percy, D. Krahl, J. Jeager in D. G. Brandon (ed.): *Electron Microscopy*, vol. **1**, Tal International, Jerusalem 1976, p. 348.
210. G. Van Tendeloo, M. Op de Beeck, S. Amelinckx, J. Bohr, W. Krätschmer, *Europhys. Lett*. **15** (1991) no. 3, 295.
211. S. Kuypers *et al.*, *J. Solid State Chem*. **73** (1988) 192. Links
212. S. Amelinckx, *Chim. Scr*. **14** (1978/1979) 197.
213. J. van Landuyt, R. de Ridder, R. Gevers, S. Amelinckx, *Mater. Res. Bull.* **5** (1970) 353. Links
214. G. van Tendeloo, S. Amelinckx, *Phys. Status Solidi A* **43** (1977) 553; Links *Phys. Status Solidi A* **49** (1978) 337; Links *Phys. Status Solidi A* **65** (1981) 431. Links
215. S. Muto, G. Van Tendeloo, S. Amelinckx, *Philos. Mag.*, in press.
216. G. van Tendeloo *et al.*, *J. Phys. Chem.* **96** (1992) 7424.
217. G. Van Tendeloo *et al.*, *Europhys. Lett*. **21** (1993) 329. M. A. Verheijen *et al.*, *Chem. Phys.* **166** (1992) 287. Links
218. D. Schechtman, I. Blech, D. Gratias, J. W. Cahn, *Phys. Rev. Lett.* **53** (1984) 1951.
219. C. Janot: "Quasicrystals, A Primer," *Monographs on the Physics and Chemistry of Materials*, Oxford Science Publishers, Oxford 1992.
220. R. Penrose, *Bull. Inst. Math. Appl.* **10** (1974) 266.
221. A. Ourmazd, D. W. Taylor, J. Cunningham, C. W. Tu, *Phys. Rev. Lett.* **62** (1989) no. 8, 933. Links A. Ourmazd, R. H. Baumann, M. Bode, Y. Kim, *Ultramicroscopy* **34** (1990) 237.
222. S. Iijima, *Nature (London)* **354** (1991) 56. S. Iijima, T. Ichihashi, Y. Ando, *Nature (London)* **356** (1992) 776.
223. X. F. Zhang *et al.*, *J. Cryst. Growth* **130**, (1993) p. 36 ff. (1993).
224. D. Ugarte, *Nature (London)* **359** (1992) 707.
225. G. Van Tendeloo, H. W. Zandbergen, S. Amelinckx, *Solid State Commun.* **63** (1987) no. 5, 389 – 393; Links *Solid State Commun.* **63** (1987) no. 7, 603 – 606. Links H. W. Zandbergen, G. Van Tendeloo, T. Okabe, S. Amelinckx, *Phys. Status Solidi A* **103** (1987) 45 – 72. Links

226. T. Krekels *et al.*, *Appl. Phys. Lett.* **59** (1991) 3048. Links T. Krekels *et al.*, *Solid State Commun.* **79** (1991) 607 – 614. Links
227. T. Krekels *et al.*, *Physica C* **210** (1993), 439 – 446. Links
228. T. Krekels *et al.*, *J. Solid State Chem*. (1993), in press.
229. L. C. Nistor, G. Van Tendeloo, S. Amelinckx, *J. Solid State Chem.* **105** (1993), 313 – 335. Links
230. X. F. Zhang, personal communication.
231. F. Watari, J. van Landuyt, P. Delavignette, S. Amelinckx, *J. Solid State Chem.* **29** (1979) 137 – 150. Links F. Watari, P. Delavignette, S. Amelinckx, *J. Solid State Chem.* **29** (1979) 417 – 427. Links
232. K. Yagi, in [152], vol. 2, p. 261. K. Yagi, *J. Appl. Crystallogr.* **20** (1987) 147. Links T. Hasegawa *et al.*, *The Structure of Surfaces II*, Springer Verlag, Berlin 1987, p. 43.
233. H. W. Zandbergen *et al.*, *Physica C* **158** (1989) 155. Links H. W. Zandbergen, W. A. Groen, F. C. Mijlhoff, G. Van Tendeloo, S. Amelinckx, *Physica C* **151** (1988) 325.
234. H. Fujita (ed.): *In-situ Experiments with High Voltage Microscope*, Research Center for High Voltage Electron Microscopy, Osaka 1985.
235. L. Reimer: “Scanning Electron Microscopy. Physics of Image Formation and Microanalysis,” *Springer Ser. in Opt. Sciences*, vol. 45, 2nd ed., Springer Verlag, Berlin 1998.
236. O. C. Wells: *Scanning Electron Microscopy*, McGraw-Hill, New York 1974.
237. D. B. Holt, M. D. Muir, P. R. Grant, I. M. Boswarva: *Quantitative Scanning Electron Microscopy*, Academic Press, London 1974.
238. J. B. Pawley: “LVSEM for High Resolution Topographic and Density Contrast Imaging,” *Adv. Electron. Electron. Phys.* **83** (1992) 203 – 274.
239. L. Reimer: *Image Formation in Low-Voltage Scanning Electron Microscopy*, SPIE Press, Bellingham 1993.
240. H. Raether: “Excitation of Plasmons and Interband Transitions by Electrons,” *Springer Tracts Mod. Phys*. **88**, Springer, Berlin 1980.
241. R. F. Egerton: *Electron Energy-Loss Spectroscopy in the Electron Microscope*, Plenum Publishing, New York 1986.
242. E. Menzel, E. Kubalek, *Scanning* **5** (1983) 151 –171.
243. D. B. Holt, D. C. Joy: *SEM Microcharacterization of Semiconductors*, Academic Press, London 1989.
244. G. V. Saparin: “Cathodoluminescence” in P. W. Hawkes, U. Valdrè (eds.): *Biophysical Electron Microscopy,* Academic Press, London 1990, pp. 451 – 478.
245. G. D. Danilatos: “Foundations of Environmental Scanning Electron Microscopy,” *Adv. Electron. Electron. Phys.* **78** (1988) 1 – 102.
246. K. F. J. Heinrich: *Electron-Beam X-Ray Microanalysis*, Van Nostrand-Reinhold, New York 1981.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Rheometry

Trevor A. Strivens, ICI Paints Plc., Slough, Berkshire, United Kingdom

1. The Science of Rheology 1
1.1. Introduction 1
1.2. Terminology 2
1.2.1. Rheological Measurement Terms . 2
1.2.2. Rheological Classification of Materials 3
1.2.3. Classification of Rheological Experiments and Measuring Systems . . . 5
1.3. Examples of Rheological Behavior 6
1.3.1. Newtonian Liquids 6
1.3.2. Elastic Solids 8
1.3.3. Simple Viscoelastic Liquids 8
1.3.4. Complex Viscoelastic Liquids . . . 10
1.3.5. Thixotropic Liquids 11
2. Rheological Measurements 12
2.1. Basic Deformation Modes for Measurement 12
2.1.1. Shear Deformation 12
2.1.2. Extensional Deformation 12
2.2. Torsional Shear Deformation . . . 13
2.2.1. Rheometer Types 13
2.2.2. Measurement Geometries 14
2.2.2.1. Parallel Plates 14
2.2.2.2. Cone and Plate 15
2.2.2.3. Concentric Cylinder (Cup and Bob, Couette) 15
2.2.2.4. Other Geometries 16
2.3. Capillary (Poiseuille) Shear Flow 16
2.4. Extensional Flow 18
2.5. Measurement Procedures 19
2.5.1. Steady-State Measurements 19
2.5.2. Dynamic (Oscillatory) Measurements 19
2.5.3. Transient Methods 21
2.5.4. Combined Methods 22
2.5.5. Normal Stress Measurements 22
3. Results of Rheological Measurements 23
3.1. Steady-State Measurements 23
3.2. Dynamic (Oscillatory) Measurements 25
3.3. Transient Measurements 26
4. References 28

1. The Science of Rheology

For basics of rheology, see also →Fluid Mechanics; →Plastics, General Survey; →Plastics, Properties and Testing.

1.1. Introduction

Rheology is defined as the science of deformation and flow of matter in response to the application of mechanical force. The etymology of the word emphasizes flow (Greek, $\varrho\varepsilon\omega$ and $\lambda o\gamma o\sigma$ mean "to flow" and "word," respectively). Although rheology overlaps classical (solid) mechanics at one end of its range and fluid mechanics at the other, it generally includes materials that show both deformation and flow under the influence of mechanical force. This excludes chemically cross-linked materials such as rubber, which may exhibit flow because of chemical degradation resulting from heat generated by the application of excessive mechanical force (thermomechanical degradation). However, it includes materials such as solid emulsion paints or thermoplastic polymers, which become liquid under the influence of stress and increased temperature, respectively. Many of these materials exhibit viscoelasticity; i.e., both irreversible (energy dissipative processes, viscous flow) and reversible (energy storage processes, elasticity) changes. Energy is dissipated mainly in the form of heat or sound, and stored as potential energy by conversion from kinetic energy. The exact balance between irreversible and reversible processes when force is applied to the substance and deformation results, depends on factors such as temperature and the magnitude of the force or deformation applied. It also depends on the time scale of the laboratory experiment relative to the characteristic times of internal material processes (e.g., molecular or particle diffusion). Some materials exhibit plastic flow, which is different in character from viscous flow but still irreversible (energy dissipative).

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_279

The science of rheology can be said to have been founded with the meeting of an international scientific committee on April 29, 1929 to draw up a constitution for the first society of rheology. Although the systematic study of rheology dates from then, SCOTT-BLAIR [15] has pointed out a number of rheological observations dating back to very ancient times (ca. 1650 B.C.) with the use of water clocks by ancient Egyptians (a forerunner of modern flow cups for measuring viscosity).

In 1660, ROBERT HOOKE discovered the law governing (linear) elastic deformation. In 1687, NEWTON proposed his law relating stress to velocity gradients in fluids. With these two laws, the modern era of rheology began, with substantial contributions over the next 200 years from (among others) EULER, CAUCHY, COULOMB, POISEUILLE, HAGEN, COUETTE, REYNOLDS, and BINGHAM, before the foundation of the first society of rheology in 1929 in the United States.

Because of the wide range of materials and physical concepts involved, rheology is a truly interdisciplinary science. Thus it unites contributions from mathematics, physics, and chemistry, and from specialists in colloid science, materials science, polymers, chemical engineering, and computer modeling.

1.2. Terminology

1.2.1. Rheological Measurement Terms

The two main modes of deformation, or ways in which materials can be changed in shape by the application of force, are stretching (or *extensional deformation*) and *shearing deformation*. These modes can be divided further. For example, extensional deformation can be *uniaxial*, (i.e., stretching in one direction only, as in stretching a thread or filament of material from end to end) or *biaxial* (i.e., stretching simultaneously in two directions, as in inflating a rubber balloon). Shear deformation can be *simple* or *torsional*.

In *torsional deformation*, the shear is rotational around a central axis. *Simple shear deformation* is shown in Figure 1. The upper plate can move parallel to the lower fixed plate under the influence of an applied force F. A thin slice of material of constant thickness h is trapped between the two plates and adheres to both (i.e., no slip occurs as the upper plate moves). The area of the upper plate is A. The *shear stress* σ (sometimes written as τ, which usually denotes relaxation time) is defined as force per unit area (i.e., F/A, measured in pascals). If the material is a *perfect* (or ideal) *elastic solid*, it obeys Hooke's law and is referred to as a *Hookean solid* or body. When force is applied, the upper plate moves a certain distance b, then stops in the new position for as long as the force is applied. The amount of deformation is defined as b/h (providing the angle Θ is small) and is known as the *shear strain* γ. Since the material thickness h remains constant, then according to Hooke's law, the shear strain is proportional to the shear stress applied. The constant of proportionality is the *shear modulus* of *elasticity* G. The unit of G is again pascal because the shear strain is dimensionless. Thus $G = \sigma/\gamma$.

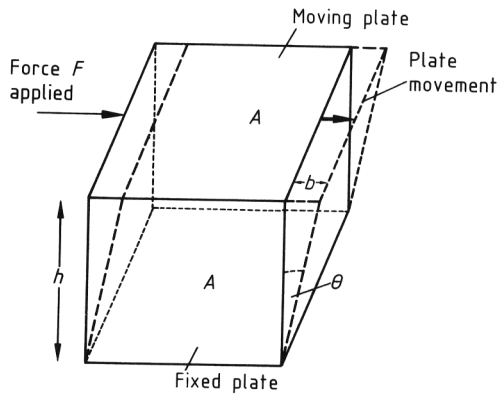

Figure 1. Diagram of simple shear to show basic rheological parameters

If the stress is removed, the upper plate returns immediately to its original position and the mechanical energy stored as potential energy in the stretched material during application of the force is recovered in the kinetic energy of the return movement.

If the slice of the material is a simple *ideal liquid*, it obeys Newton's law and is referred to as a *Newtonian liquid*. When force is applied, the upper plate moves at a constant velocity v for as long as the force is applied. In an analogous fashion, a *shear strain rate*, usually known as the *shear rate*, is defined as v/h (denoted by $\dot{\gamma}$, i.e., $d\gamma/dt$). [Shear rate is sometimes symbolized as

D, but $\dot{\gamma}$ is used to avoid confusion with the Deborah number (see Section 1.2.2) or the differential operator d/dt, which are sometimes written as D.] In Figure 1, the liquid velocity is assumed to decrease linearly across the gap between the moving upper and the fixed lower plates. The shear rate is proportional to the stress, and the constant of proportionality is the *shear viscosity* η in units of pascal seconds (Pa · s). If the force is removed, the plate stops moving (unless the inertia due to its mass continues its motion for a short time) but does not return to its initial position. There are no stretched elements in the material that, by relaxing, can pull the plate back toward its original position, and the potential energy of the plate is zero. The mechanical energy applied has been dissipated as heat [16].

Analogously, an *elongation strain* ε and *strain rate* $\dot{\varepsilon}$, as well as an elongation modulus (*Youngs's modulus*, E) and an *elongation viscosity* $\bar{\eta}$, may be defined. Under certain well-defined material conditions, the elongation modulus is three times the shear modulus, and the elongation viscosity three times the shear viscosity (Trouton's law or ratio [17] →Fluid Mechanics, Chap. 4.3.4.).

1.2.2. Rheological Classification of Materials

Many materials show characteristics of both ideal solids and viscous liquids: they are *viscoelastic solids* or *liquids* depending on which material characteristic predominates. If the modulus or viscosity is independent of strain or strain rate, respectively (i.e., the materials obey Hooke's or Newton's laws), they are referred to as *linear* materials because their response to deformation is linear. In practice, most materials obey these laws over limited ranges of stress, strain, or strain rate. When the modulus becomes a function of strain and the viscosity a function of strain rate (usually, in both cases a decreasing function) the material response is *nonlinear*. At high shear rates, nonlinear responses can lead to so-called normal forces. These arise because the force exerted by the moving liquids becomes more and more out of alignment with the flow as the shear flow increases (see Section 2.5.5).

Most materials, even if linear, do not respond instantaneously to the application of a force; even if these materials are linear, they show a *time-dependent* response. This time dependence becomes much more obvious if the applied stress or strain itself varies with time, for example, in *oscillatory or dynamic* experiments where the variation is sinusoidal, or in transient experiments where the time scale of the material response is long compared to those of experimental artifacts. Whether a predominantly elastic or viscous response is obtained is often dictated by the relationship between the material response time (material characteristic time) and the time scale of the experimental observation. This time ratio is the Deborah number D, named by M. Reiner after the Hebrew prophetess Deborah who sang "the mountains flowed before the Lord" (Judg. 5) [18]. Low Deborah numbers indicate viscous, and high numbers (>1) elastic response [19]. Thus, even simple liquids can exhibit elasticity at very high frequencies [20].

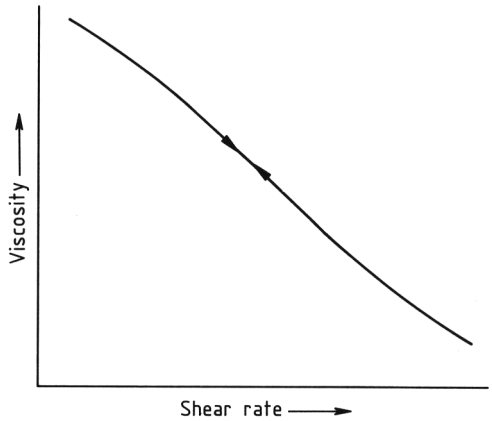

Figure 2. Viscosity – shear rate profile (pseudoplastic)

A number of qualitative, classificatory terms are used in rheology to describe various types of material response. If the shear rate is increased linearly with time between zero and a maximum value and then decreased over the same range the resulting viscosity – shear rate plots for the increasing and decreasing shear rate halves of the experiment may either be superimposed on each other (*pseudoplastic behavior*, Fig. 2), or the plot of increasing shear rate may be above the plot of decreasing shear rate, to form a hysteresis loop (*thixotropic behavior*, Fig. 3) (see also →Fluid Mechanics, Chap. 4.1.2., →Fluid Mechanics, Chap. 4.1.3.). Instead of continuous shear rate variation, with modern rheometers a series of measurements can often be pro-

grammed at discrete shear rate values (ascending and descending as before), with a predetermined time at each value of shear rate before the viscosity is measured. Whether pseudoplastic or thixotropic behavior is observed depends on the ratio of the time the material requires to reach a steady state to the experimental time at each shear rate.

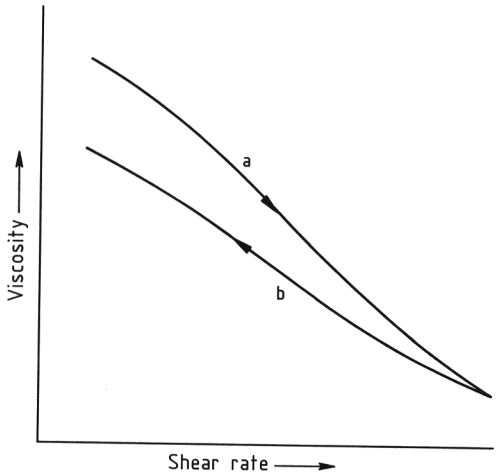

Figure 3. Viscosity – shear rate profile (thixotropy)
a) Increasing shear rate; b) Decreasing shear rate

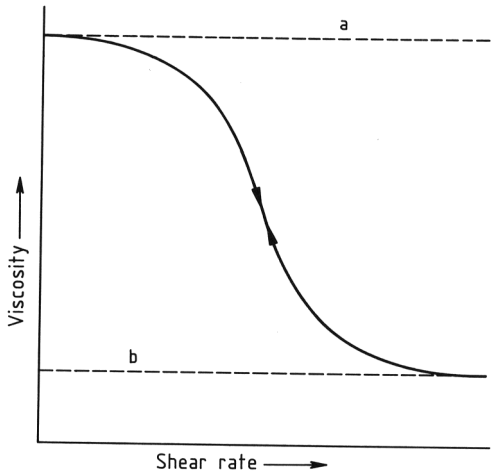

Figure 4. Viscosity – shear rate profile with Newtonian plateaus
a) Lower Newtonian plateau; b) Upper Newtonian plateau

At both low and high shear rates, materials may show regions of Newtonian behavior, with pseudoplastic behavior at intermediate shear rate values (Fig. 4). These regions are the *lower Newtonian* and *upper Newtonian plateau regions*, respectively. The terms lower and upper refer to the shear rate values and not to the Newtonian viscosity values in each region.

Many familiar materials, such as emulsion paints, toothpaste, and cream, exhibit structure at rest (i.e., they have a more or less solid appearance). However, when sheared by stirring they become liquid-like. Although they may not show thixotropy at moderate to high shear rates, they nearly always do at low shear rates. In addition, the viscosity appears to increase more and more rapidly as lower and lower shear rates are used, with no sign of the approach of a lower Newtonian plateau region (Fig. 5). A linear plot of shear stress versus shear rate appears to extrapolate not to the origin but to a positive value of shear stress, the so-called *yield stress* [21]. Experimentally, determination of the yield stress is difficult because of the long times most of these materials require to reach a steady shear stress value at very low shear rates, problems of data extrapolation, and instrumental difficulties, etc. These materials nearly always show an upper Newtonian region because shearing breaks down the structure formed at rest and the material flows like a "normal" liquid. Other materials have different behavior at high shear rates: in particular, the viscosity increases with increasing shear rate. This behavior is referred to as (→ Fluid Mechanics, Chap. 4.1.3.):

1) *Shear thickening*, when the increase is due to enhanced flocculation of particles in dispersion (orthokinetic flocculation) [22]
2) *Rheopexy*, when the shear flow orients anisotropic particles in dispersion into favorable positions for flocculation [23]; it is usually time dependent

These phenomena have also been observed in polymer solutions and other materials, but they are associated predominantly with the rheological behavior of dispersions. Flocculation refers to the formation of clusters, chains, or networks of particles due to net attractive forces existing between the particles of dispersions. Such flocculation may also lead to the shear thinning phenomena already discussed [24], [25].

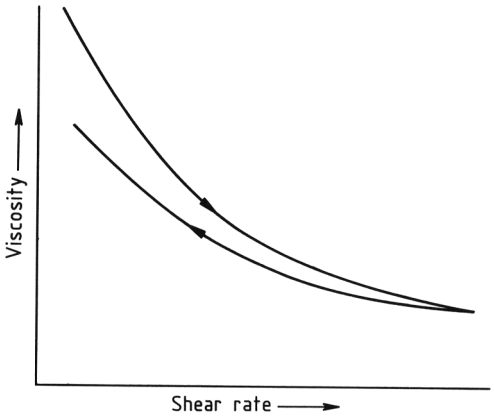

Figure 5. Viscosity – shear rate profile of a structured liquid

Other terms refer to the approximate or exact conformation of the rheological data to a particular mathematical equation, empirical equation, or theoretical model. Thus, if rheological parameters exhibit a linear dependence on frequency or shear rate on double logarithmic plots, the rheological behavior is referred to as *power law behavior*. If the material shows a yield stress and the square root of the measured viscosity (the so-called apparent viscosity) is proportional to the square root of the shear rate, the material is described as having Casson flow behavior, after the theoretical model of structure breakdown developed by Casson (see also →Fluid Mechanics, Chap. 4.1.) [26].

Finally, *reduced viscosity values* are often used. Two kinds of reduced viscosities exist. Firstly, Newtonian plateau viscosity values for *dispersions or polymer solutions* are often divided by the viscosity of the dispersion medium or the solvent, respectively, before the viscosity is related to disperse phase volume or polymer concentration. Secondly, in describing *shear rate dependence*, certain equations such as the Cross equation use a reduced viscosity of the kind

$$\frac{\eta_a - \eta_x}{\eta_o - \eta_x}$$

where η_a is the measured or apparent viscosity, and η_o, and η_x are, respectively, the lower and upper Newtonian plateau values of the viscosity.

A full glossary of rheological terms based on the 1975 British Standard [27] and the U.S. Society of Rheology recommendations [28] is found in [29].

1.2.3. Classification of Rheological Experiments and Measuring Systems

In addition to the simple type of shear deformation described (Fig. 1), two other types are also of considerable importance. These are (1) *pipe, capillary,* or *Poiseuille flow* and (2) *torsional flow*. Capillary flow in so-called *U-tube viscometers* forms the ultimate measurement standard for Newtonian liquids. The flow is characterized by a nonlinear velocity distribution across the tube radius. Further, this distribution depends on the rheology of the test liquid. Since the velocity distribution can be predicted accurately for Newtonian and for a few (mathematically) simple non-Newtonian liquids (e.g., power law fluids), U-tube viscometers are used primarily for Newtonian liquids and for very high shear rate measurements (where liquids tend to be Newtonian, anyway). Since fresh unsheared liquid is passing through the tube continuously, fewer problems occur because of errors due to mechanical energy dissipation as heat at high shear rates than in the case of the small static sample volumes used in other types of shearing deformation.

Simple deformation is really not very practicable for steady shearing measurements, because of the limitation in total strain (or the product of shear rate and time) that may be applied (one plate sliding over another with the material confined between) and the difficulties of confining all but the most viscous liquids in the gap between the plates. A simple way of reducing these difficulties is to use circular plates with a rotational movement around a central axis, instead of a sliding motion between two parallel plates; this is referred to as *torsional shear deformation* or flow. Again, one plate is fixed and the other moves with the gap between them kept constant. This measuring geometry is known as *parallel plate* geometry. However, the angular velocity of the moving plate now varies across the plate radius (zero at the center and maximum at the edge of the plate); consequently, the shear rate does also. By substituting, for one plate, a cone whose tip touches the other plate, the shear rate becomes independent of the radial distance from the center of the moving

plate. This occurs because the angular velocity increases linearly with distance from the center of the moving plate, but so does the gap between the plates (Fig. 6). This type of measuring geometry is known as *cone and plate geometry*. In practice, a small part of the cone tip is cut off or truncated (see Section 2.2.2.2). Another torsional measuring geometry is the so-called *Couette*, *concentric cylinder,* or *cup and bob*, where the material to be measured is confined in the space between the cup and a cylindrical bob inside it. By keeping the radii of the cups large and the gap between the cylinder inner wall and the bob wall small, the difference in radial velocities at the two walls is small. Consequently, the shear rate across the gap is fairly uniform (see Fig. 7), although not as uniform as in the cone and plate geometry. However, large errors may arise from the contribution of the flow between the bob end and the cup wall. Cone angles, as well as the radius of cone or plate and the radii of Couette components, can be used to control the shear range within given limits of rotation speeds.

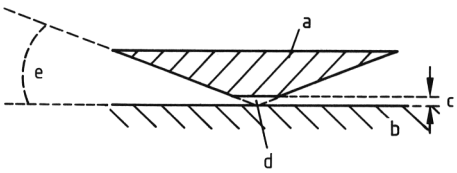

Figure 6. Cone and plate geometry
a) Cone; b) Plate; c) Gap; d) Truncated cone tip; e) Cone angle

By rotating one member of the measurement geometry at a given speed, viscosity – shear rate measurements can be obtained. If the time scales are correct (see Section 1.2.2), *steady-state* measurements are obtained. By oscillating one member between limits at a controlled amplitude and frequency, particularly with a sinusoidal relationship to time, *oscillatory* or *dynamic* measurements are performed. Other time dependences, such as square wave or triangular, can also be applied. They are generally more difficult to use, from the point of view of both instrumental capability and mathematical analysis. An *impulsive* or *transient measurement* is made by applying a single square wave input or a linearly or logarithmically increasing input of known level for a certain period of time and measuring the material response as a function of time (see Section 2.5.3).

In *extensional flow*, the measuring geometry and test method depend on the shear viscosity and cohesive strength of the liquid. Thus, for liquids with high shear viscosity and high cohesive strength such as polymer melts, threads can be formed that can be stretched or wound on reels under well-defined conditions. More solid-like melts can be made into sheets for biaxial deformation and into bars or rods for uniaxial deformation. As the shear viscosity is reduced, a thread of liquid can be drawn into a tube from the surface of a liquid sample, and a stable liquid profile can be maintained in this manner. The thread or column can then be stretched by moving the tube vertically away from the liquid surface. However, the presence of a disperse phase, such as solid particles of filler or pigment, or gas bubbles (e.g., air), reduces the cohesive strength of the liquid and leads to instability or breakdown of the liquid column. Further reduction of the shear viscosity will render the formation of liquid columms or threads impossible, because of the very low cohesive strength of liquid threads. However, many low-concentration, high molar mass polymer solutions have extensional viscosities many orders of magnitude greater than their very low shear viscosity of a few millipascal-seconds [30]. Such liquids can be measured by the opposing jet apparatus (commercially available), by the four-roller technique (if the liquid is transparent), or in a semiquantitative manner by pressurized flow cups or thread formation as drops break away from a syringe needle tip (see Section 2.4).

1.3. Examples of Rheological Behavior

(→Fluid Mechanics)

1.3.1. Newtonian Liquids

The results of oscillatory and creep measurements for a Newtonian oil are given in Figures 8 and 9. In Figure 8, the real and imaginary components (G' and G'') of the complex modulus $G*$, as well as the dynamic viscosity η' are plotted as a function of frequency. At low frequencies the dynamic viscosity is equal in value to the (Newtonian) steady rate shear viscosity

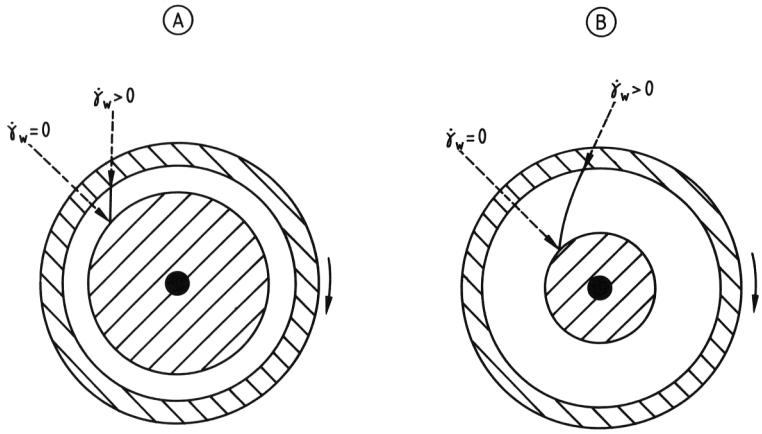

Figure 7. Shear rate distribution in Couette geometry
A) Narrow gap, linear variation of shear rate across gap; B) Wide gap, nonlinear variation of shear rate across gap

value. Correspondingly, the imaginary component of the modulus or *loss modulus* G'' is proportional to frequency f and is much greater in value than the real component or *dynamic elasticity modulus* G'. This is typical of the dynamic response of a liquid. The loss modulus represents the portion of the total mechanical energy, supplied during the applied strain cycle, which is dissipated as heat during the cycle. The portion stored as potential energy by elastic deformation of the material is represented by the modulus of elasticity G'. The ratio of the imaginary to the real component of the (complex) modulus of elasticity yields the value of tan δ, where δ is the phase angle difference between the applied stress or strain cycle and the corresponding response (measured) cycle. The modulus of elasticity is very small at low frequencies but increases sharply (to the second power) with increasing frequency. This means that at higher frequencies, the two plots cross and the apparently Newtonian liquid becomes elastic at ca. 40 Hz. Correspondingly, the dynamic viscosity η' decreases as the high-frequency end is approached. Since simple liquids exhibit elasticity only at very high frequencies [20], many orders of magnitude above the ones shown in Figure 8, clearly another source of energy storage exists. This additional source is the measuring system itself, particularly the contribution to the overall response signal from its inertia and elasticity. This contribution becomes a maximum at the so-called natural frequency, which is close to 40 Hz for the particular measuring systems used here. The inertia term increases with the second power of the frequency. If this inertia contribution is subtracted, the elasticity modulus G' is low and virtually independent of frequency. Thus, measurement conditions should be chosen so as to eliminate or minimize the contribution of the measurement system either by appropriate corrections or by replicate measurements using different measurement systems or conditions. It is most important to have access to raw measured data values so that such corrections can be made by the operator and to understand exactly how printed rheological parameters are derived from such data in the computer software of commercial rheometers.

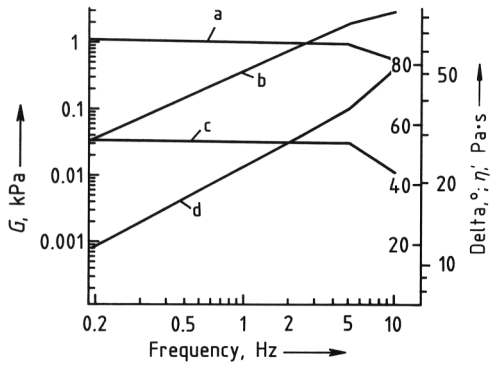

Figure 8. Oscillatory data (Newtonian liquids)
a) Delta; b) Loss modulus G''; c) Dynamic viscosity η'; d) Dynamic modulus of elasticity G'

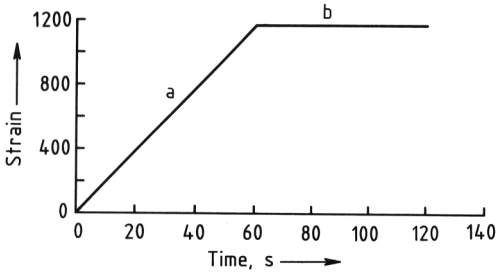

Figure 9. Creep data (Newtonian liquids)
a) Retardation period; b) Relaxation period

Results for creep measurements on the same oil are shown in Figure 9. The creep experiment is a typical transient experiment in which a constant stress is applied virtually instantaneously to the samples and the corresponding strain induced in the sample is measured as a function of time (*retardation period*). The stress is then removed instantaneously, and the sample strain is again measured as a function of time during its recovery from stress application (*relaxation period*). In the case of a Newtonian oil, the strain increases linearly with time during the retardation period. However, during the relaxation period, the strain remains constant.

When the stress is removed, all the mechanical potential energy supplied during the retardation period has dissipated. Consequently, the sample remains at the maximum strain value attained during the retardation period, and no change occurs during the relaxation period measurements.

1.3.2. Elastic Solids

As an example of an elastic solid, Figure 10 and Figure 11 show the results of oscillatory (dynamic) and transient (creep) experiments on a 6 % (food) gelatin gel in water at 25 °C. The dynamic data show two striking features: first, the dynamic modulus of elasticity G' (curve a) is independent of frequency, and second, it is much greater than the loss modulus G'' (curve b). The loss modulus G'' exhibits a minimum but very little frequency dependence. The creep experiment (Fig. 11) shows an almost constant strain value during the retardation period, with almost complete and instantaneous recovery to zero strain during the relaxation period. In the figure compliance (measured strain divided by applied stress) is plotted against time. In creep experiments, it is more convenient to use compliance instead of modulus of elasticity. Compliance is denoted by J and related reciprocally to G, i.e., the unit of compliance is Pa^{-1}. Deviations from ideal Hookean behavior can be attributed to either slight irreversible deformation or, more likely, some failure in instrumental or measurement conditions.

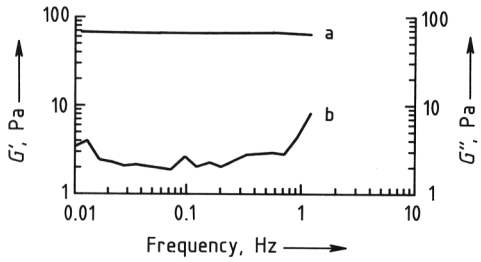

Figure 10. Oscillatory data (6 % gelatin solution in water)
a) Dynamic modulus of elasticity; b) Loss modulus

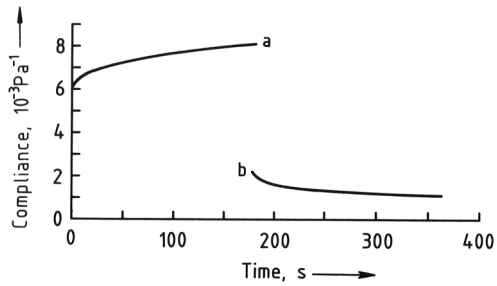

Figure 11. Creep data (6 % gelatin solution in water)
a) Retardation; b) Relaxation

1.3.3. Simple Viscoelastic Liquids

The 0.1 mol/L cetyltrimethylammonium bromide (1 : 1 molar ratio) – salicylic acid solution (CTAS) system in water forms long rod-like micelles that like many other such materials, exhibit strong elasticity while flowing easily like a liquid [31]. The 0.1 mol/L CTAS system is fairly unique because it exhibits virtually ideal behavior, in that it has a single relaxation time and can be modeled by the simple Maxwell model (→Plastics, Properties and Testing) [32].

As can be shown from viscosity – shear rate measurements, the viscosity is Newtonian at low shear rate values (i.e., it has little or no de-

pendence on shear rate). At higher shear rates, the viscosity decreases rapidly with increasing shear rate. At even higher shear rates the viscosity once again approaches Newtonian behavior (see Fig. 4). To display the full rheological profile of a given sample, data from more than one measurement system must often be combined, as has been done for the concentrated poly(methyl methacrylate) dispersion in the hydrocarbon system shown in Figure 12. Misalignment of data, as shown in this example, can arise from small calibration errors, difficulties in obtaining repeatable samples, or sample handling.

Figure 13 shows oscillatory data for the 0.1 mol/L CTAS system. Now the features of both the Newtonian liquid and the Hookean solid are combined, with a narrow transition region between them in the middle of the frequency range. At low frequency, the dynamic viscosity η' is constant and the loss modulus G'' is greater than the dynamic modulus of elasticity G' (predominantly viscous response). The loss modulus and the modulus of elasticity increase approximately linearly, and to the second power, respectively, with increasing frequency. At the high-frequency end, the response is predominantly elastic. The dynamic modulus of elasticity is approximately constant, and the loss modulus lies below it and decreases rapidly (as does dynamic viscosity). In the narrow frequency region in between, the loss modulus passes through a sharp maximum, at which frequency (the *crossover frequency*) the modulus of elasticity equals the loss modulus and has a value approximately half that of its maximum at higher frequencies. This crossover frequency corresponds to the *relaxation time* (the reciprocal of the frequency divided by 2π). In Figure 13, the relaxation time is several seconds. This finite relaxation time can be demonstrated even more clearly by the (transient) stress relaxation experiment. Here, a sudden strain is applied to the sample, and the resultant torque (stress) exerted by the sample is measured as a function of time. After electronic smoothing, the stress – time curve shows an initial time-independent value (corresponding to the high-frequency modulus G' value of Fig. 13) and then decreases fairly rapidly toward zero (Fig. 14). The inflection point of the plot G' versus time (curve b) corresponds to the relaxation time. This can be seen more clearly if the first differential of the curve is plotted against time (curve c, relaxation spectrum in Fig. 14). The time corresponding to the peak of this curve is the relaxation time. With identical samples and test conditions (not the case in Figs. 13 and 14), the two methods should agree fairly closely for the relaxation times.

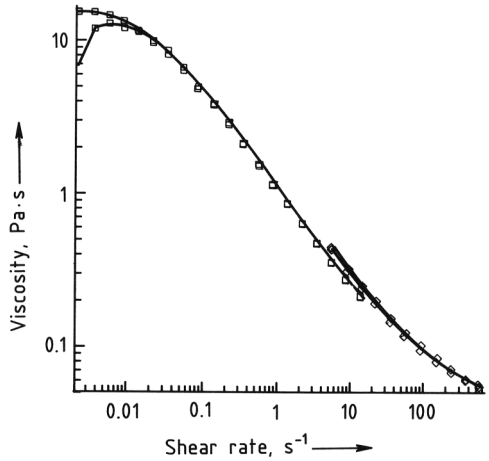

Figure 12. Viscosity – shear rate data [concentrated poly(methyl methacrylate) dispersion]
□ and ◇ denote data obtained with different equipment

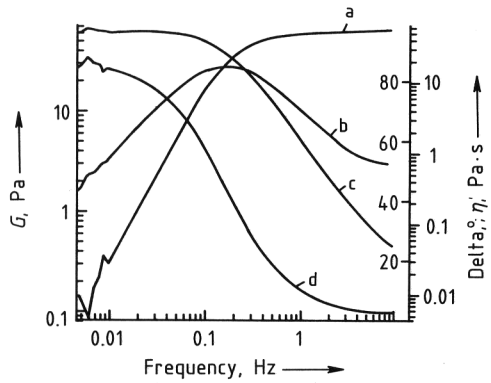

Figure 13. Oscillatory data (0.1 mol/L CTAS)
a) Dynamic modulus of elasticity G'; b) Loss modulus G''; c) Dynamic viscosity η'; d) Delta

The reason for the necessary electronic smoothing of the relaxation data in Figure 14 is that the elasticity of the sample causes a sudden twisting force (torque) to be applied to the calibrated bar of the measuring system. The system then oscillates at its natural frequency in the same way a pendulum does when moved from its position and released. Like the pendulum, this

initial oscillation decays at a rate dependent on the velocity of the dissipation of the kinetic energy of movement by the surrounding medium. In the case of the measuring system it decays rapidly because of the sample viscosity. The initial peak in the spectrum (curve c) at ca. 20 ms represents this initial decay due to the measuring system.

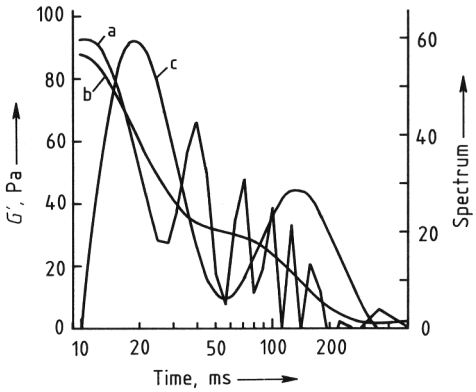

Figure 14. Stress relaxation data (0.1 mol/L CTAS)
a) Dynamic modulus of elasticity G'; b) G' after electronic smoothing; c) First differential of b

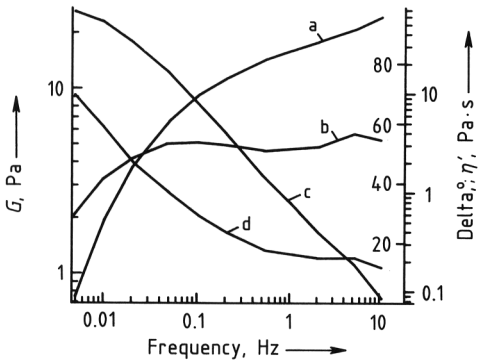

Figure 15. Oscillatory data (concentrated dispersion)
a) Dynamic modulus of elasticity G'; b) Loss modulus G''; c) Dynamic viscosity η'; d) Delta

1.3.4. Complex Viscoelastic Liquids

In Figure 15, the results of dynamic (oscillating) measurements on a more complex system are depicted. The system measured here is a concentrated dispersion of poly(methyl methacrylate) (PMMA) in a paraffin. The dispersion is sterically stabilized with poly(hydroxystearic acid) (PHS) – PMMA graft copolymers (for preparation details, see [33]). Now a much broader transition between liquid-like viscous response and solid-like elastic response is obtained. In particular, the peak in the loss modulus G'' plot becomes very shallow, and the crossover of the loss and elasticity modulus plots occurs lower than the peak frequency. The crossover frequency now corresponds to the longest relaxation time of the system, and the broader transition is due to a range of relaxation times in the system. This contrasts with the simple system in Figure 13, where only one relaxation time occurs. By superimposing many simple systems like that of Figure 13 – each with its own relaxation time and appropriately weighted low-frequency plateau viscosity and high-frequency plateau elasticity values – the broader, less definite transition behavior shown in Figure 15 is obtained. However, as Figure 12 shows, the system is still strongly pseudoplastic under continuous shear, with a well-developed lower Newtonian viscosity and the approach of an upper Newtonian region clearly visible at higher shear rate. Although the particle-size distribution is not particularly narrow ($\pm 10\,\%$ of the mean particle diameter), the range of transition times is apparently more dependent on concentration than on particle-size variation. If the dispersion is diluted gradually, the range of relaxation times decreases and the transition region becomes narrower. At any given moment, because of their concentration and lack of free volume, the particles are arranged in random groupings of touching particles, with each group having a different size and, consequently, its own characteristic relaxation time. Thus a range of relaxation times is created. As the system is diluted, Brownian motion and more free volume lead to the breakdown of the largest groups into smaller groups. Thus the range of group sizes becomes smaller, and the range of relaxation times is reduced. In addition, the plateau value of the modulus of elasticity high-frequency decreases sharply. Since this value reflects the total number of particle – particle contacts per unit volume, it supports the particle contact group explanation of the relaxation time range observed [34].

1.3.5. Thixotropic Liquids

In Figure 16, steady shear viscosity measurements are presented for a typical industrial thixotropic material. Here the shear rate is varied continuously through an increasing and a decreasing cycle. Immediately after completion of the first cycle, a second identical cycle of shearing is applied to the sample. Two factors are striking: (1) the first cycle shows a thixotropic loop (i.e., the viscosity – shear rate plot for the descending half of the cycle lies at viscosity values below those for the ascending – increasing shear rate – half of the shear cycle); and (2) in the second cycle, the plots are superimposed on each other. In other words, the sample has now become pseudoplastic in its continuous shear rheology.

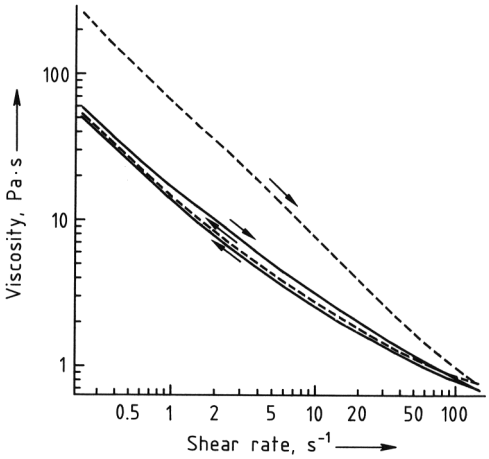

Figure 16. Two shear rate cycles (thixotropic material) – – – – = first cycle; —·— = second cycle

These data show two of the characteristics of thixotropic materials. First, an interaction exists between experimental time and material response time. Although not shown here for reason of clarity, increasing the cycle time from the 5 min used here decreases the loop width or area enclosed by the ascending and descending plots, whereas decreasing the cycle time increases the loop width. Furthermore, the existence of a previous shear history, as exemplified by the first cycle of shearing, leads to a change in the sample rheology observed.

Second, although an upper Newtonian limit is approached at the high shear rate of the cycle, no corresponding limiting value of the viscosity appears to exist at low shear rates. Instead, if shear stress is plotted versus shear rate and the plot is extrapolated to zero shear rate, it does not go through the origin but through a positive value of the shear stress, the so-called yield stress (see Section 1.2.2). Although experimental problems occur in obtaining reliable yield stress values, and the physical reality of the yield stress has been debated [21], no doubt exists about the practical reality of such an effect. This is evident in the difficulty of producing flow for these materials, particularly in start-up of operations such as pumping or mixing efficiency in large vessels with paddle stirrers. Such practical difficulties are magnified if elasticity is present in the material at rest.

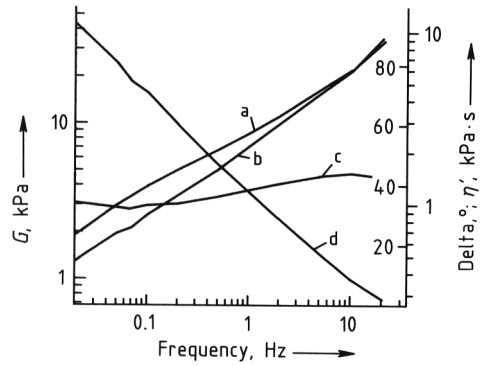

Figure 17. Oscillatory data for thixotropic material
a) Dynamic modulus of elasticity G'; b) Loss modulus G''; c) Delta; d) Dynamic viscosity η'

The elasticity effects in thixotropic materials can be demonstrated by dynamic (oscillatory) measurements. Data for the same industrial material are shown in Figure 17. Plots of the modulus of elasticity G' and the loss modulus G'' on double logarithmic axes are parallel straight lines (i.e., both moduli exhibit a power law dependence on frequency). This sort of dependence is characteristic of a wide range of relaxation times, with the maximum possible time being infinite. Such a range of times arises from the tenuous structure in terms of both size and degree of branching, and extent of distribution across the volume of the sample. Such structures are characteristic of associating (flocculating) particles in dispersion, associating polymer chains in solution, or mixtures of the two. Since such associations frequently result from physical rather than chemical bond-

ing, they are susceptible to breakage by temperature increase or small mechanical deformation, thus giving rise to other characteristic features of thixotropic materials. These physically bonded structures may extend throughout the sample, giving rise to modulus of elasticity values greater than the loss modulus. However, because of their tenuous network and readily labile nature, they rarely exhibit the predominantly Hookean-type elasticity behavior characteristic of chemically cross-linked polymer gels or very regularly and highly (physically) cross-linked gels, such as the gelatin system discussed [35].

2. Rheological Measurements

2.1. Basic Deformation Modes for Measurement

Two basic modes of deformation are described in detail, namely, shear and extensional deformation. Other modes exist (e.g., bulk or compressional deformation) that are not discussed here.

2.1.1. Shear Deformation

Three submodes of shear deformation exist – simple, torsional, and capillary, or Poiseuille deformation. *Simple shear flow* has conceptual, but little practical, value. Simple shear flow is limited to highly viscous or stiff materials, such as printing ink or asphalt, and can be used in very limited shear rate or total shear strain ranges. Most measurements are made with torsional shear or capillary flow equipment.

Capillary flow is applied to low-viscosity fluids, and gravitational force is used to drive the flow. However, if air or gas pressure is used to drive the flow, fluids with a much higher viscosity can be measured at much higher shear rates. In this way, polymer solutions or melts can be measured. The combination of high viscosity and high shear rate leads to high rates of energy dissipation as heat. Because of the continuous flow of fresh sample into the capillary and the low residence time of a particular volume element of the sample in the capillary, the high rate of energy dissipation usually does not lead to problems of sample temperature control and consequent errors in viscosity measurements.

Torsional shear measurements, in contrast, use the same aliquot of sample confined between the elements of the measuring geometry. Because of the small gap between the elements of the measuring geometry necessary to achieve high shear rates and the reduced sample size, the energy dissipation rate becomes a problem in terms of temperature control. The narrow gap also presents some severe mechanical design problems. However, the narrow gap combined with the relatively large radius of the cup and cylinder (*Couette geometry*) gives a shear rate that is virtually constant across the gap. Although the *cone and plate geometry* is more suitable for achieving uniform shear rate across the sample, the high rotational speed and consequently strong centrifugal forces tend to force the sample from the gap, which limits the maximum shear rate obtainable.

Higher shear rate values can be obtained with the *Couette geometry,* due to containment of the sample within the cup. By contrast, capillary flow leads to a gradient of shear rate across the radius of the capillary. Whereas the gradient can be predicted for Newtonian liquids, this is more difficult if the viscosity – shear rate relationship is unknown. However, because many liquids show near-Newtonian behavior at very low and very high shear rates, capillary flow is more widely applicable than merely as a standard reference measurement for viscosity, despite the difficulty in calculating the shear rate gradient.

2.1.2. Extensional Deformation

Although a large number of commercial rheometers are available to perform torsional flow measurements, relatively few exist for measurement of extensional flow. This reflects the much greater practical difficulty of making such measurements, particularly at high rates of extension or with very mobile liquids. In addition, the problem of generating flows approximating steady extension is by no means trivial. For sim-ple liquids, Trouton's rule states that the

extensional viscosity is three times the shear viscosity value. However, the extensional viscosity of several high molar mass dilute polymer solutions is many orders of magnitude higher than the shear viscosity. For example, METZNER and METZNER [36] measured very dilute aqueous solutions of polyacrylamide (100 mg/L) and found shear viscosities of 1.4 mPa · s and extensional viscosities of 9 Pa · s (Trouton ratio 6500 instead of 3). Other systems showed Trouton ratios as high as 30 000 [37]. Measurements using a different technique substantiate these findings [38]. However, a disparity exists between results obtained with different techniques on the same sample [39]. Practically, only steady-state measurements can be performed because of experimental difficulties. However, stress transients are observable before steady-state flow is attained. Some of the disparities between measurement methods may relate to the failure to obtain true steady-state values or the onset of molecular degradation in high stretching rate flows (dilute polymer solutions).

Two main types of flow in extension have been distinguished by WALTERS [40]. These are (1) uniaxial, and (2) biaxial and two-dimensional flow. Most methodical studies use *uniaxial methods.* Some studies of *biaxial and two-dimensional flow* have been made by DENSON and coworkers [41], in which the formation of bubbles by inflation of small areas of plastic sheet has been studied. Even if the flow may be steady (Eulerian sense), each liquid element must be subjected to a rate of deformation that is independent of time: i.e., transients have died away, and the flow is steady in this sense also (Lagrangian sense). In this latter sense, particularly with low-viscosity liquids at high extension rate, achieving steady extensional flow under practical laboratory conditions is difficult.

Extensional flow has many obvious practical applications in industries based on polymers, for example, fiber spinning, extrusion, vacuum forming, and blow molding. Extensional flow can also be applied in studies, e.g., on surface coatings, atomization during paint application, or the properties of coatings applied by roller coating. Other important areas are lubrication, drag reduction, and flow through porous media (of interest to the oil industry).

2.2. Torsional Shear Deformation

2.2.1. Rheometer Types

Two basic types of rheometer are available for rheological measurement in torsional shear deformation: controlled shear strain – shear (strain) rate and controlled stress instruments. Historically, the former type is older, the first one having been described by COUETTE in 1890.

In controlled *shear strain – shear rate instruments*, the lower member of the measuring geometry is rotated through a small angle (shear strain) or rotated at a controlled speed (shear rate). The sample, confined by the system geometry, exerts a force on the fixed upper member as a result of movement of the lower. This force or torque is measured by means of a sensitive transducer (e.g., a piezoelectric device) or by the (relatively) small deflection of a calibrated torque bar, which couples the upper member to the body of the instrument. The angular deflection or rotation speed is then converted into the applied shear strain or shear rate, and the measured force or torque into the required stress. The applied shear strain or shear rate can be varied in time to increase or decrease linearly, or it can be applied discontinuously with varying rest times between applications.

Controlled stress instruments operate according to a different principle. The lower member of the measuring geometry is rigidly attached to the body of the instrument. By means of a controlled electric current applied to an induction motor, a rotary torque is applied to the measuring head to which the upper member of the geometry is rigidly attached. Movement of the head can be measured by, for example, an optical diffraction grating system, in terms of both angular deflection (shear strain) and angular rotation speed (shear rate). By varying the current, the stress can be controlled as required.

The following applies to both types of instrument:

1) Because of increasing requirements of sensitivity, accuracy, etc., the measuring heads are nearly always supported by air bearings that minimize errors due to friction in bearings.
2) Most modern rheometers incorporate computers that, besides converting measured

(raw) data into the required rheological parameters, also control the experimental procedure. This allows considerable flexibility of control and permits a sequence of inputs to be designed to simulate complex practical situations.

3) When measuring transient properties, a potentially important factor is the input time of the two types of instrument. Thus the imposition of a steady torque by electrical means is extremely rapid. Instantaneous (Hookean) elasticity can be measured since measurement can start 1 ms after commencement of stress application with a measurement rate of $120\,s^{-1}$ in a typical commercial instrument. By contrast, a purely mechanical drive system requires a longer time (rise time) to reach a preset strain level (typically 20 ms to reach the preset strain zero). This results in modification of the short-time stress relaxation behavior observed, particularly if fast relaxation processes occur in the sample. With torque bar measurements, rapid application of the strain leads to an initial oscillating stress (the torque bar behaves like a suddenly struck pendulum). The extent of oscillation and rapidity of decay depend not only on the properties of the measuring head and geometry, but also on the energy dissipative capability of the sample. Although electronic filtration can be used to remove the initial oscillation, it may not be entirely satisfactory. With electrically driven instruments, such instrumental transients are small and of very short duration and do not disturb the measurements.

Some constant-stress instruments are equipped with software that, by using feedback control of stress level, can provide controlled shear rate measurements. Despite the apparent duplication of steady-state, transient, and oscillation measurements on both stress and shear rate instruments, the two types of instrument provide very different and complementary kinds of information. The selection of instrument must depend on both the type of flow application to be studied and the sample to be measured. For example, when studying the flow-out of a wet paint film after application, a stress-controlled rheometer is more appropriate because the flow of the wet paint is caused and driven by stresses or forces originating from gravitational forces and from surface tension gradients on the film surface [42]. In contrast, when studying stirring or mixing problems, shear rate controlled rheometers are more appropriate because the rotation speed of a stirrer blade or a mixing head will determine the shear rates applied to the sample.

2.2.2. Measurement Geometries

The most common geometries of commercial rheometers are parallel plate, cone and plate, and concentric cylinder. Others are available, or can be made, but they have rather more specialized uses. A detailed description of all geometries, together with theoretical considerations, can be found in [9], [43], [44].

2.2.2.1. Parallel Plates

The parallel plate rheometer is a logical development of the parallel plate simple shear geometry (see Section 1.2.3). Instead of a sliding motion, however, the plates are replaced by disks a fixed distance apart, which can rotate about a central axis. This means that both the strain and the shear rate vary across the radius of the disk, from zero at the center to a maximum at the disk edge. This renders parallel plates unsuitable for viscosity – shear rate measurement (unless the liquid is Newtonian). However, the geometry is highly suitable for dynamic (oscillatory) measurements. Because the linear response region is generally used, the strain amplitude variation across the radius is immaterial. The advantages of parallel plates are:

1) Correction for fluid inertia effects is relatively easy (this is important for combinations of low-viscosity liquids and high-frequency measurement)
2) The gap can be varied freely

Fluid inertia effects can be significant, since they depend on the magnitude of the product αh, where h is the plate separation and α is given by the equation

$$\alpha^2 = -\frac{i\omega\varrho}{\eta^*} \tag{1}$$

where ω is the angular frequency (2π times the frequency in Hertz), ϱ the liquid density, $\eta*$ the complex dynamic shear viscosity, and $i^2 = -1$. Clearly αh can be reduced by decreasing the plate distance h, but this procedure has limitations. First, when working with dispersion systems, a value of h more than 50 times the average disperse particle diameter should be used to avoid errors due to wall effects [45]. Second, the narrower the gap, the greater is the likelihood of error due to edge effects, which result from the sample drying at the edge. Examples showing the magnitude of fluid inertia effects are found in [46].

Errors due to *wall effects* arise from the presence of a concentration gradient in particles in dispersion systems (sometimes referred to as the depletion layer). This occurs because the wall exerts a repulsive force (hydrodynamic, not colloidal in origin) on the particle as it approaches. *Loss of solvent* by evaporation during the test is always a problem, because it leads to the buildup with time of concentrated viscous or solid (skin) layers at the sample – air interface. This buildup leads to enhanced torque readings and hence measurement errors. Evaporation can be slowed by coating the sample surface with a layer of immiscible and impermeable liquid; however, the best solution is to saturate the atmosphere around the sample surface with solvent vapor.

2.2.2.2. Cone and Plate (Fig. 6)

To eliminate variation of shear strain or strain rate across the radius of the parallel plate, one of the disks is replaced with a cone (see Section 1.2.3), thus ensuring that the shear rate remains constant. In addition, a small cone angle is required to achieve this [47]. A small part of the cone center is truncated (see Fig. 6), so that it has no direct contact with the plate. Direct contact would result in the risk of friction and, hence, errors in torque measurement and mechanical wear. The smaller the cone angle (measurement at the rim), the greater is the shear rate for a given rotational speed. Cone angles normally vary between 0.5 and 5°.

In addition to the source of error already mentioned under parallel plates, another is due to *errors in gap setting.* To achieve a uniform shear rate, the gap between the cone and the plate must be adjusted carefully so that the cone tip (afterwards removed by truncation) just touches the center of the plate. Since the gap settings are typically between 15 and 60 µm and many commercial rheometers are equipped with a micrometer, dial gauge, etc., that cannot be set to better than 1 µm, errors can be of the order of 2 – 6 %. If programmed temperature tests are carried out on the same samples, errors due to gap setting variation from thermal expansion can be considerable. *Wall effects* can cause serious errors, particularly with concentrated dispersions [48]. *Fluid inertia effects* are particularly difficult to correct, because the α term now appears in the solution of a partial differential equation instead of the ordinary differential equation required for describing responses in parallel plate or concentric cyclinder geometry. A series solution for this equation have been given with Bessel functions and associated Legendre functions in the product αr, where α is as defined above and r is the cone radius [49]. These equations are so complicated that apparently no one has utilized them, even with the aid of computers.

The cone and plate geometry is the obvious choice for viscosity – shear rate measurements. It is not suitable for oscillatory measurements unless fluid inertia corrections can be neglected.

2.2.2.3. Concentric Cylinder (Cup and Bob, Couette)

The concentric cylinder geometry possesses two main advantages over the cone and plate and the parallel plate geometries: (1) a larger shearing surface, and therefore, greater sensitivity when measuring low-viscosity materials; (2) minimization of errors due to drying or skinning at the sample surface – air interface. This last point is not immediately obvious. In normal sample filling, the sample level should be above the top of the bob but below that of the cup. This ensures that most of the torque produced by the sample originates in its bulk as the sample is sheared in the narrow gap between the outer wall of the bob and the inner wall of the cup. Since the shaft of the bob normally has a much smaller diameter than the bob itself, the torque contribution from any thickening with time on the sample surface is minimal. Evaporation from the sample surface tends to be into a stagnant air layer

if the surface level is below the top of the cup, and under such conditions a layer of immiscible, impermeable liquid can easily be placed on the sample surface to prevent evaporation loss. By arranging the system so that the cup and bob radii are large, and the gap between the bob and cup inner wall is small, the shear situation for any small volume element of the sample becomes approximately that of simple shear with uniform shear rate across the gap. In addition, the generally larger sample volumes reduce errors due to heating at high shear rates, and confinement of the sample at high shear rates helps prevent sample loss by centrifugal force effects.

However, the concentric geometry has a major drawback: the corrections required because of end effects (i.e., the fact that some of the torque contribution from the sample is due to the end of the bob, where the shear regime is ill-defined). For example, errors as high as 25 % were found with Newtonian standard oils, using a conventional narrow gap cone ended bob [50]. Other options commercially available are bobs with hollow ends – the idea being to trap an air bubble under the bob. Thus the air pocket reduces the torque contribution from the bob end. However, due to problems from the wetting effects, of shear on the bubble and the rim on the end of the bob, the end effect may still be finite but also variable. Other means of minimizing or eliminating end effects are discussed in [51].

Alternatively, end effects may be *calibrated* out by measuring the torque – angular rotation speed ratio for various heights of sample in the cup. By extrapolating the torque to zero, the extra bob height equivalent to the end effect can be determined and allowed for in viscosity calculations. Within experimental error, these end effect corrections are independent of rotation speed. Corrections must be done for each sample; the calibration for a Newtonian liquid cannot be used [51].

Although end effects appear not to be significant for linear viscoelastic materials in dynamic measurements [52], the effects of *fluid inertia* may be significant. Again, the mathematical treatment of these effects, although easier than in cone and plate geometry, is still more difficult than with parallel plate geometry. A method of approximation using a series to obtain the modulus value is given in [53] and the whole topic is discussed in [54].

2.2.2.4. Other Geometries

Other geometries include double cone systems (one cone fitting into another wider-angle cone containing the sample), various squeezing or sliding flow instruments, instruments with noncoaxial or spherical geometries, and various penetrometers. These are all described in [2], [3], [9].

Whereas most of these have limited general applicability, some are commercially available. For example, the *double cone system* is claimed to be useful for elastic samples, which may fracture when sheared in conventional geometries. *Squeezing and sliding flow instruments* are of interest for very viscous materials (e.g., printing inks), while many types of *penetrometer* (instruments, in which depth – time profiles are measured for a suddenly released loaded needle or other probe) are used in industry for measuring solid and semisolid materials, such as lubricating greases, soaps, gels, foods, or paints.

The *noncoaxial geometries* were potentially of great interest in that they allowed dynamic data to be derived from continuous rather than oscillatory motion. However, they do not provide any significant advantage over conventional rheometers in terms of either the accessible frequency range or the amelioration of sources of error [55].

2.3. Capillary (Poiseuille) Shear Flow

Two main type of materials are tested by capillary rheometers: low-viscosity liquids, where the driving pressure is provided by gravitational force (hydrostatic head type), and high-viscosity liquids such as polymer melts, where the driving pressure is provided by gas or hydraulic pressure. All capillary rheometers are easily temperature controlled. Often when using metal capillaries, good control at elevated temperature is easier than in torsional rheometers. Glass U-tube-type viscometers used for capillary measurements on low-viscosity liquids can be immersed in a temperature-controlled water bath for good control of temperature. Besides being more easily thermostated, capillary viscometers

are generally simpler, cheaper, and more accurate than torsional viscometers.

During the flow of a liquid through a capillary, three regions of flow can be distinguished: entry flow, viscometric flow, and exit flow. The *entry flow region* represents the "funneling" effect of the liquid as it flows from a relatively wide-diameter reservoir into the capillary, and this effect takes some time to relax to its limiting value. Estimates of the length of the entry region vary, it is generally accepted as about 40 times the capillary diameter. To achieve consistent results, the reservoir diameter should be at least 15 times the capillary diameter [56]. In the *viscometric region*, the stresses produced by flow along the tube axis are constant. Clearly, pressure measurements should be made in this region. The *exit region* is induced by the exit of the free liquid stream from the capillary end. It is usually shorter than the entry region. The free stream may have a larger diameter than that of the capillary (die swell) and is often associated with elasticity in the liquid. Experimentally, the length of the viscometric region is maximized by using a sufficiently long capillary tube for the experiment.

The *experimental procedure* consists of measuring the volume flow rate of the liquid and the total pressure drop between the ends of the capillary tube. If the sample is an unknown material, a number of capillaries of different length and interior diameter may be required to establish the flow curve of the liquid and to make suitable correction for end effects (both entry and exit) and other sources of error.

The greatest problem with *interpreting the results* of capillary flow measurements is calculation of the velocity profile across the diameter of the tube. For a Newtonian fluid, the profile can be predicted to be parabolic; for non-Newtonian fluids the profiles can be calculated only if the viscosity – shear rate relationship is known. This can be done for simple rheological models, such as power law or Bingham fluids (→Fluid Mechanics, Chap. 4.1.1., →Fluid Mechanics, Chap. 4.1.2.). However, other models give intractable or complicated mathematical formulas. Although capillary measurement methods are commonly used for non-Newtonian liquids, the results are often presented as flow curves (i.e., pressure drop – flow volume rate plots). Methods do exist for deriving viscosity – shear rate values from flow curve data, but these are valid, in general, only for fluids in which the departure from Newtonian behavior is small. Nevertheless, despite these problems, capillary flow measurements are attractive not only because of their simplicity and cheapness in combination with accuracy, but also because of their ease of application to very high shear rate measurements. With respect to errors and repeatability, the results of capillary measurements on Newtonian liquids are an order of magnitude better than corresponding results on the best commercial torsional instruments [50]. Furthermore, most torsional rheometers are not capable of measuring at shear rates much above $1000\,s^{-1}$ because of the various errors involved. Although the precise limit depends on the geometry and the sample involved, the essential problem is the low sample volume with small gaps and high rate of energy dissipation at high shear rates, with consequent temperature control problems.

In the case of *hydrostatic-type viscometers* such as the Ostwald U-tube viscometer, the driving pressure is set by filling a reservoir with sample and timing, between two graduation lines, the emptying of the reservoir through the capillary tube. Such capillary instruments, usually made of glass, are low-stress instruments (1 – 15 Pa) and have maximum shear rate values of ca. $1 - 20\,000\,s^{-1}$ based on efflux times of 200 – 800 s [57]. Their main use is as secondary standards for viscosity measurements [58], [59], but they are important in establishing the molar mass of polymers or polymer fractions (from gel permeation chromatography separations, for example) by measuring the low shear rate (Newtonian) viscosity of dilute solutions (→Plastics, Analysis, Chap. 5.2.). Another important industrial application is the calculation of liquid flow in pipes.

Finally, *efflux* or *flow cups* are widely used in industry. Basically, these are hydrostatic head instruments, consisting of cylinders with a conical bottom, pierced at the center by a short capillary whose length does not usually exceed the cup wall thickness. They may be operated at elevated (e.g., the Saybolt and Redwood cups used in the petroleum industry) or ambient (e.g., the Ford or BS cups used in the paint industry) temperature. The cup is filled to a mark with the capillary plugged. The plug is then removed, and the time required to fill a graduated flask placed

underneath to a specified volume, or until the fluid stream first breaks into drops is measured. Although these methods give reasonable results for quality control or diluting to a given viscosity (e.g., for spray application of paint), they are not very sound in concept. First, they are totally unsuitable for non-Newtonian liquids because the gravitational driving stress varies considerably due to the large drop in liquid level (ca. 0.06 m) and also the capillary length is so short that viscometric flow is never established.

2.4. Extensional Flow

Extensional flow instruments can be classified according to Nazem and Hill [60]. Controllable experiments refer to those in which the stress or extensional strain rate is kept constant and the steady value of the other parameters is measured. In practice, controllable experiments are applicable only to high-viscosity materials, such as polymer melts. Noncontrollable techniques must be applied to low-viscosity liquids.

Polymers. Polymer specimens are commonly *stretched between two blocks*, one fixed and one movable. This places a constraint on the maximum strain, and hence strain rate, that can be attained because of the finite length of the equipment and the requirement allowing transient stresses to die away to steady levels. By suitable electronic or mechanical means, a programmed force or length increase can be applied so as to maintain constant stress [61] or constant extensional strain rate [62–64]. An improved strain rate controlled instrument has been described that has two sets of rollers at either end of the specimen [65]. This allows a maximum rate of ca. 1 s^{-1}, which is nearly an order of magnitude higher than the other instruments available. Examples of these instruments are commercialy available.

In *biaxial extension,* sheets of polymeric material can be tested either by bubble inflation or by the apparatus developed by Meissner [66] in which an octagonal clamping system is driven by a complex servomotor system.

Mobile Liquids. For mobile liquids, steady-state measurements in both the Eulerian and the Lagrangian senses (Section 2.1.2) are impossible with the present state of knowledge [67]. Therefore, noncontrollable methods are applied. Of these, the laboratory equivalent of fiber spinning equipment is commercially available and widely used [68–73]. The liquid being tested is extruded through a die and collected on a turning spool. Both the diameter of the threadline as a function of position and the flow rate are measured, and the extensional stress is calculated from the force on either the die or the turning spool. The limitation of this method is probably the cohesive energy of the liquid and whether, in practice, a coherent stable thread can be drawn from the liquid.

Alternative methods applicable to more mobile liquids are the open siphon method [74–77], the converging flow method [36], [78–84], the triple jet method [38], and the more recent (commercially available) opposing jets method. In the *open siphon* or *Fano flow method*, the sample is sucked into a siphon until a stable column is formed [75]. This has an advantage over methods involving dies, because the liquid has not suffered any (perhaps unknown) deformation before being stretched [73]. *Converging flow techniques* are simple to use in that liquid flows from a reservoir through an orifice into a capillary. However, analysis of the measurements is difficult, and measurement requires that the liquid be transparent. In the *triple jet technique*, two high-velocity jets converge on a central lower-velocity jet of the same liquid, thus causing the central jet to be stretched. Finally, in the *opposing jets technique,* two jets are mounted horizontally: one on a fixed vertical arm and one on a pivot equipped with a transducer. The jets are immersed in the liquid under test, and liquid is sucked in or blown out through the jets at a controlled flow rate by two syringes also filled with test liquid. The flow rate is controlled by a constant-speed stepping motor, which drives a moving block that pulls or pushes the plungers of the two syringes equally. By measuring the deflection of the pivoted jet and applying a measured but variable force to maintain its position, the extensional stress is obtained. Despite the limitation imposed by the maximum travel distance of the plungers in the syringes, extensional viscosities can be measured at strain rates up to 50 000 s^{-1} or more.

2.5. Measurement Procedures

Three basic measurement procedures are described – steady state, dynamic (oscillatory), and transient – together with a statement of the equation for calculating stress, strain, or strain rate for each of the main measurement geometrics (see Table 1 and Table 2). Because both extensional deformation and capillary shear deformation are carried out only under steady-state conditions, the bulk of this section is devoted to torsional measurements. Extensional flow equations tend to depend on the apparatus and technique used. Detailed discussion of the equations along with their theoretical derivation, is given in [2] and [7].

2.5.1. Steady-State Measurements

Steady-state conditions are achieved by choosing a value of stress or shear rate to apply to the sample and waiting sufficiently long for the measured shear rate or stress value to become constant before proceeding to the next value. Generally, one starts with low applied values and proceeds stepwise to the maximum required value. The sequence is then reversed to check that shearing has not brought about any irreverisble change in the sample. This is necessary for torsional shear when the same sample volume is used for all measurements. When fresh sample is supplied continuously to the measuring region (e.g., in capillary flow), it is not necessary unless the bulk sample is suspected of changing with time, due to sedimentation of the dense dispersion phase, slow phase separation, etc.

In practice, the more solid-like the material, or the more viscous the liquid, or the closer the material is to phase separation or changes, then the longer are the times required to attain steady state. For example, some structured emulsion paints may require 5 min or more to reach a measured value constant to within 10 % over 30 s at a moderate applied stress (100 Pa) [85]. In these cases, thixotropic behavior may have to be accepted for reasons of expediency, i.e., the time required to complete the entire measurement procedure or irreversible changes in the sample over time (e.g., evaporation losses, sedimentation), and the accuracy requirements made less stringent.

The more fluid the material, the more likely is the occurrence of pseudoplastic or Newtonian behavior. In this case, the time to complete the entire measurement may be considerably reduced by using a continuous sweep up and down instead of the stepwise procedure.

2.5.2. Dynamic (Oscillatory) Measurements

Whereas the objective with steady shear measurement is to measure the viscosity over a range of imposed stresses or shear rates, the objective of dynamic measurements is to study the sample as near rest as possible. Shear moduli values are therefore usually measured as a function of applied shear stress or strain at a fixed frequency to define the linear response region (i.e., the range of amplitude values in which the moduli are constant). In practice, the linearity-limiting amplitude does not vary very much with frequency. However, some structured materials have a linear response region at very low applied stress or strain amplitude values, so compromises must sometimes be made between the demands of linearity and the ability of the rheometer to make accurate measurements at low amplitudes.

Few workers have studied the *nonlinear dynamic viscoelastic* properties of materials. Not only are the measurements more complicated and the data more difficult to analyze, theoretical development is still rudimentary, at least in its application to rheological properties. In this respect the work of Onogi and his group should be mentioned [86].

The formulas in Table 2 are applicable to controlled-strain amplitude rheometers only. Those for controlled-stress instruments are similar, but differ in detail.

The equations show the relationship between the measured parameters and the complex dynamic shear viscosity $\eta*$. The measured parameters are the phase angle δ and the stress amplitude θ_0. The ratio of the stress amplitude θ_0 to the strain amplitude input θ_1 is denoted by θ [87].

In Table 2, K and I are the torque constant and the moment of inertia about the axis of the torque-measuring element of the rheometer, respectively; $\omega = 2\pi f$ is the angular frequency (radians per second); and f denotes the measuring frequency (hertz). In the fluid inertia correction

Table 1. Equations for flow shear stress, and shear rate for different measuring geometries

Measuring geometry	Shear stress		Shear rate	
Parallel plate (plate radius R; separation h; distance from center of plate r; angular velocity Ω)	total torque G $G=\int_0^R 2\pi r^2 \sigma \mathrm{d}r$	(2)	$\dot{\gamma} = \frac{\Omega r}{h}$	(3)
	for Newtonian liquid $\eta=\frac{2hG}{\pi\Omega R^4}$	(4)		
Cone and plate [symbols as for parallel plate: θ is angle between cone and plate surface (radian)]	$\sigma=\frac{G}{2\pi R^3}$	(5)	$\dot{\gamma} = \Omega/\theta$	(6)
Concentric cylinder (cup radius R; bob radius r; bob height h)	$\sigma=\frac{G}{2\pi R^2 h}$ where $R = R$ or $r*$	(7)	$\dot{\gamma} = \frac{r\Omega}{R-r}$	(8)
Capillary flow (tube radius a; length l; pressure difference between ends P; volume flow per second V)	$\sigma_\mathrm{w} = \frac{aP}{2l}$	(9)	$\dot{\gamma}_\mathrm{w}=\frac{4V}{\pi a^3}$	(10)
subscript w indicates stress at tube wall	Newtonian liquids only in general, $V=-\frac{\pi a^3}{\sigma^3}\int_0^{\sigma_\mathrm{w}} \sigma^2 f(\sigma)\,\mathrm{d}\sigma$ and $\eta=\frac{\sigma_\mathrm{w}}{\dot{\gamma}_\mathrm{w}}$			(11)

∗ If $r/R = 0.98$, the two stress values agree within 4 % [85].

Table 2. Equations for oscillatory flow∗

Measuring geometry	Without fluid inertia corrections		With fluid inertia corrections	
Parallel plate	$\frac{e^{i\delta}}{\theta}= 1-\frac{i}{\eta^*}S$ with $S=\frac{2h\left(K-I\omega^2\right)}{\pi a^4\omega}$	(12)	$\frac{e^{i\delta}}{\theta}= \cos\alpha h+S'\alpha h\sin\alpha h$ with $S' = \frac{2\left(K-I\omega^2\right)}{\pi\varrho a^4 h\omega^2}$	(13)
Concentric cylinder	as parallel plate, with $S=\frac{\left(R^2-r^2\right)\left(K-I\omega^2\right)}{4\pi L R^2 r^2\omega}$	(14)	$\frac{e^{i\delta}}{\theta}=\frac{\pi r^2}{2R}X$ $\{\alpha[J_2(\alpha r)Y_1(\alpha R)-Y_2(\alpha r)J_1(\alpha R)]+\alpha^2 S''X$ $[J_1(\alpha r)Y_1(\alpha R)-Y_1(\alpha r)J_1(\alpha R)]$ with $S''=\frac{K-I\omega^2}{2\pi\varrho r^3 L\omega}$	(15)
Cone and plate	as parallel plate, with $S=\frac{3\alpha\left(K-I\omega^2\right)}{2\pi r^3\omega}$	(16)	see [49]	

∗ For explanation of symbols, see Section 2.5.2.

term $\alpha^2 = i\omega\varrho/\eta*$, where ϱ denotes the density of the sample and $i^2 = -1$. For parallel plates, h denotes plate separation and a the plate radius; for concentric cylinders, R_0 and R_I denote the radii of the outside (internal cup radius) and the inside cylinders (bob), respectively; L is the bob length; and J and Y are Bessel functions of the first and second kind. Finally, for the cone and plate geometry, α_0 and r denote, respectively, the angle between the cone and the plate and cone radius.

Resolving Equation (12)

$$\frac{e^{i\delta}}{\theta} = 1 - \frac{i}{\eta^*}S \tag{12}$$

for the elasticity modulus components yields

$$G' = \frac{\omega S\theta\,(\cos\delta - \theta)}{\theta^2 - 2\theta\cos\delta + 1} \tag{17a}$$

$$G'' = \frac{-\omega S\theta\sin\delta}{\theta^2 - 2\theta\cos\delta + 1} \tag{17b}$$

The corresponding equations for the controlled stress rheometer are

$$G' = C\,(\theta\cos\delta - 1) + \frac{CI\omega^2}{K} \tag{18a}$$

$$G'' = C\theta\sin\delta \tag{18b}$$

$$C = \frac{\pi a^4}{2hK}$$

The equation taking account of fluid inertia effects is also simpler for controlled-stress instruments [85]

$$\frac{e^{i\delta}}{\theta} = \frac{K - I\omega^2}{K} + iC\alpha h\eta^*\omega\cot(\alpha h) \tag{19}$$

Besides fluid inertia effects, other sources of error can occur in dynamic measurements due to natural frequency effects and sample stiffness. The sample stiffness is, however, rarely important. Correction methods have been described [88], but the situation should be avoided by ensuring that θ is always small compared to unity.

The *natural frequency* ω_0 of the measuring system is given by $\omega_0^2 = K/I$. At this frequency, $\theta = 1$ and $\delta = 0°$ for controlled-strain experiments and the values of G' and G'' become indeterminate (Eqs. 17, 18). However, for controlled-stress instruments, $\delta = 90°$ and, again, the G' and G'' values are indeterminate. Although accurate values can be measured on either side of the natural frequency, at higher frequencies the accuracy of measurement becomes more limited as frequency increases beyond the natural frequency value. This is due to the increasing dominance of the inertia term ($I\omega^2$) on the right-hand side of Equations (17) and (18) [89]. A better procedure is to shift the value of the natural frequency by changing either K (by using, for example, a stiffer torque bar) or I (by altering the mass of the measuring geometry – e.g., substituting a glass or plastic cone for a steel cone or using a smaller-diameter cone). By this means, the natural frequency can be shifted out of the required frequency range.

In addition to the forced oscillation (i.e., where an oscillation frequency is imposed on the sample – measuring system combination), the system may be in so-called *free oscillation.* However, this technique should properly be classified as a transient method (see Section 2.5.3).

2.5.3. Transient Methods

In the free oscillation method, the sample – measuring system is deflected (usually in torsion), then released and allowed to oscillate until damping by the sample stops the oscillation (→Plastics, Properties and Testing). The oscillation frequency and the amplitude decrease (decrement) per cycle are measured. Because the oscillation frequency can often be arranged by judicious choice of mass, etc., to be in the 0.1 – 2-Hz range, measurements can be made rapidly. Its speed of measurement and inherent simplicity make this an attractive technique for studying rapid changes in rheological properties during, for example, resin curing or polymer cross-linking. Using a woven braid to support the initially low-viscosity fluid, GILLAM and coworkers have made extensive use of this technique [90]. The other primary transient methods are creep and stress relaxation (→Plastics, Properties and Testing). In *creep measurement,* the sample undergoes a stepwise increase in stress, and the resulting strain is measured as a function of time. The stress is then removed, and the decay of strain with time is followed. In the *stress relaxation experiment,* a stepwise strain is applied to the sample, and the resulting stress is measured as a function of time until the stress either decays to zero or becomes constant.

When using, for example, parallel plates in a torsional shear stress relaxation test, the modulus of elasticity $G(t)$, defined as the stress $\sigma(t)$ at time t divided by the applied strain γ, is given by

$$G(t) = \frac{2hT}{\pi R^4\varphi} \tag{20}$$

where T is the measured torque at time t and φ is the angular deflection (radians). In general, the equations in Table 1 may be used to calculate stress, and by substituting γ for $\dot{\gamma}$ and φ for Ω the strains can be found. In creep measurements the compliance $J(t)$ is often calculated.

The same considerations still apply in choosing measuring geometries, and cone and plate geometry is probably best wherever possible. In addition, the rise time (i.e., the time to reach the chosen stress or strain value from an initial zero value) must be taken into account. Ideally, the sample's relaxation times should be longer than the rise time. The application of sudden deflections to a torque bar can lead to initial free oscillations that can swamp sample response in the initial few seconds of a stress relaxation experiment. The magnitude of these oscillations depends on the brevity of the rise time, the strain value applied, and the stiffness of the sample. This last factor can be reduced by increasing the gap between parallel plates, reducing the cone diameter, etc.

2.5.4. Combined Methods

With the aid of computer software, many commercial torsional shear rheometers can provide any required combination or sequence of basic measurement procedures just described. This is particularly useful in two ways. First, many structured materials are very sensitive to stress or strain and have very long *recovery times.* Consequently, the procedure of sampling from the bulk and loading the sample in the rheometer may impose sufficient stress or strain to change its properties significantly (i.e., the sample may suffer unknown deformations before measurement). The long recovery times may mean that even leaving the sample for a long time before measurement is not sufficient for it to return to its original state. This can be circumvented by imposing an initial deformation history, for example, by applying a shear rate until a constant stress value is obtained. The measurement can then be carried out immediately afterwards, and the effect of varying the initial history on the measurements can be studied. The second possibility is that by varying the time interval between the completion of the initial history and the start of measurement, the recovery process can be determined. Particularly useful in measuring recovery processes are repeated oscillatory measurements over a period of time at single frequencies or (better still) superimposed multifrequency waveforms, which are becoming available in many software packages. Such techniques are also useful if the sample is changing with time because of some other effects than relaxation, e.g., chemical cross-linking.

The second way in which such combined methods are useful is to *simulate complex flow problems.* For example, paint applied to a vertical wall is spread by the high stress exerted by a roller or brush; the wet liquid film should then remain unchanged under the influence of low stresses due to gravity. This unchanged state should remain for up to 5 min, after which viscosity increases or the appearance of elasticity – due to the loss of solvent by evaporation (drying) or the onset of curing – prevents further change. If flow does occur under the influence of gravity, it sometimes does not show immediately, but appears after a few tens of seconds as drips, runs from edges, and wavy irregular patterns of uneven thickness. These, of course, spoil the appearance and protective power of the final paint coating. This process may be simulated by applying a high stress for a short time (brush application), followed by a creep measurement made at small applied stress (0.2 – 5 Pa), which can be varied to simulate variations in wet-film thickness and specific gravity. The creep measurement will then show whether significant flow occurs at a specific stress (film thickness) and within a certain specified time (drying time).

Other examples stem from the tendency of many disperse systems to form *structures or become solid-like* when at rest for long periods. Whereas such structure usually breaks down when a stress or strain is applied to it and the material liquefies, in the design and operation of pumping and mixing equipment, the pressure (stress) or mixing speed (shear rate) required to liquefy the sample and the length of time for which it should be applied, must be known. Rheological measurement procedures can be designed to simulate these conditions, whereas dynamic measurements can be used to probe the nature and strength of the structure at rest.

2.5.5. Normal Stress Measurements

When an elastic liquid is sheared (e.g., a polymer solution or melt) a significant component of the total force is often developed in the liquid, which is in a direction normal to the direction of

shear. Thus, in addition to the shear force usually measured, a significant force may be measured by mounting a suitable transducer in the face of the cone or plate of the measurement geometry. These so-called normal forces lead to rod climbing or Weissenberg effects (→Fluid Mechanics, Chap. 4.3.) in which such liquids form a convex meniscus around a rotating stirrer rod or shaft [91], [92].

Normal force effects occur in all the basic measurement techniques, provided the material possesses sufficient elasticity and is fluid enough to be sheared without sample fracture. Dispersion systems, particularly well-stabilized ones, do not usually possess sufficient elasticity, even when very concentrated, to exhibit measurable normal force values. Methods for measuring normal force in continuous shear have been described in [93]. In dynamic measurements the measured normal force oscillates at twice the applied frequency [94].

3. Results of Rheological Measurements

Reviews of rheological results and theory include [10], [12], [14].

3.1. Steady-State Measurements

At low concentrations, both polymer solutions and dispersions follow Einstein's formula and are Newtonian liquids (viscosity independent of shear rate)

$$\eta_r = \frac{\eta_o}{\eta_s} = 1 + 2.5\varphi \tag{21}$$

where η_o is the viscosity at low shear rate, η_s is the solvent or dispersion medium viscosity, and φ is the total phase volume occupied by the polymer molecules or particles.

The number of interactions increases with increasing particle or molecular concentration, either by interaction through collision or by modification of the flow fields around the particles and the mutual interaction of these fields. These interactions increase the amount of energy dissipation per unit time during shearing, and consequently, the overall viscosity of the dispersion or solution also increases. Finally, at ca. 20 % volume concentration, the number of interactions becomes sufficiently large that the effect of strong shearing is to reduce their number and intensity by changing a random, three-dimensional fluctuating network of interactions into a less random, two-dimensional fluctuating network. Such reduction in interactions results in less energy dissipation and, consequently, a lower viscosity for the whole system. This leads to the appearance of two Newtonian viscosities: a high value corresponding to the three-dimensional random situation at low shear rate (lower Newtonian plateau) and a lower value corresponding to the more ordered situation at high shear rate (upper Newtonian plateau). At intermediate shear rates, the viscosity is shear rate dependent. At these intermediate values, a balance is struck between the shear rate trying to order the system and the thermal motion of the solvent molecules trying to disorder the particles or solute molecules. This balance may be achieved rapidly. If the interactions are relatively strong, the rate at which the balance is reached may be of the same order or greater than the rate of change of shear rate (if the shear rate is constant for a certain experimental time, then the time to reach balance may exceed the experimental time). This has two consequences:

1) The flow behavior shows time dependence (i.e., it is thixotropic rather than pseudoplastic)
2) Because the associative interactions lead to a higher effective concentration or a higher phase volume than that of the actual concentration, the onset of shear rate dependence of the viscosity is observed at much lower concentration than 20 % phase volume

Shear rate dependences are shown diagrammatically for dispersions in Figure 12 and for polymer solutions in Figure 18. Well-stabilized dispersions usually show a stronger dependence on shear rate of viscosity than polymer solutions, as indicated.

As the concentration increases, the increasing number of multiple interactions leads to greater increases of viscosity for each equal increment of concentration. Both the upper and the lower Newtonian values increases sharply, but the difference between them also increases (i.e., the shear rate dependence becomes stronger). Finally, a critical volume fraction is reached at

which the particles of the dispersion or the solute molecules have so little free volume that they are in contact with each other constantly, thus forming a network throughout the volume of the liquid. At this point, the viscosity of the liquid becomes infinite and the liquid exhibits solid-like behavior. Finally, if the volume fraction becomes higher, even these rearrangements under continuous shearing will be impossible, and the material becomes entirely solid-like. These changes normally take place over an effective phase volume fraction of 0.50 – 0.72.

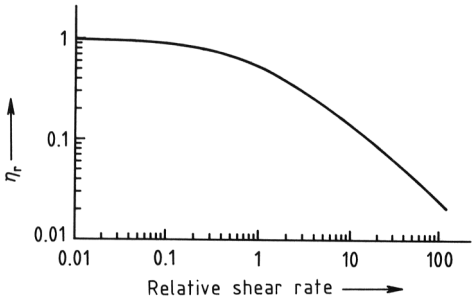

Figure 18. Viscosity – shear rate profile for polymer solution (diagrammatic)

Differences occur between different systems in this critical volume fraction range. Thus, well-stabilized *hard particle dispersions* exhibit very abrupt changes in viscosity over a narrow concentration range, whereas the changes in *polymer solutions* are far more gradual, due to changes in the polymer molecular volume as the concentration increases. In dilute polymer solutions in a thermodynamically good solvent, the macromolecular coils are fully expanded. As the polymer concentration increases, polymer – polymer interactions becomes stronger than polymer – solvent forces. As a consequence, less solvent is "trapped" within the polymer coil, and the coil contracts in volume until it reaches so-called unperturbed dimensions [95]. Beyond this minimum volume, coils can still overlap or entangle as concentration increases further. Thus the volume shrinkage leads to an initial small change with concentration (dilute solution), then a more rapid change over a fairly wide concentration region where the swollen coils come closer to each other and contract to the unperturbed dimensions (semidilute solution), and then finally a steep viscosity increase with concentration due to overlap (concentrated solution). This last increase is not usually as strong as in concentrated dispersions, due to coil flexibility, and continuous measurements are possible up to 100 % polymer content. Because hard particles are not flexible, continuous shear measurements on concentrated dispersions are not possible beyond a certain critical maximum volume packing fraction, which seems to depend on the nature of the stabilization often around that corresponding to random packing (0.639) for sterically stabilized [96], and higher (0.74) for charge-stabilized dispersion [97], [98].

Some *bimodal particle dispersions* can be fluid up to volume fractions of 0.85 or slightly more, if the relative particle sizes and the concentration of each size are chosen carefully [99]. Apparently, when the large particles reach the critical packing fraction, the smaller particles can fit in the interstices between them. Emulsions can also reach similar phase volumes before the overall viscosity becomes very high [100], because the droplets can undergo distortion (due to the flexibility of the interface and the fluidity of their interiors) and pack in hexagonal cross-sectional structures. The continuous phase is then confined to the interfacial surfaces and small pockets where more than two interfacial surfaces meet. The maximum phase volume that can be reached usually depends on the onset of phase inversion. Similar considerations apply to foams (gas-in-liquid dispersions), but in practice, because of the much greater overall fragility of the structure compared to emulsions, they are usually unable to sustain much shear.

Polymer melts exhibit a similar transition in viscosity behavior with (average) molar mass. Thus at low molar mass, the macromolecular coils are not long enough to overlap or entangle and are in their unperturbed dimensions. As molar mass increases, the (low shear rate) viscosity becomes dependent linearly on the molar mass. Beyond a certain critical molar mass, overlap or entanglement commences and the viscosity becomes dependent on molar mass to the (average) power of 3.4. A more or less pronounced transition region may occur, due to the effects of molar mass distribution [101].

Additional effects may occur in *charge-stabilized dispersions,* due to the presence of charged ions of opposite sign to those on the particle that are attached to the particle surface, by

either chemical bonds or adsorption of charged stabilizers (surfactants or polymer molecules). Thus a charge double layer is formed at the surface of the particle. The thickness of this layer depends on the electrolyte concentration in the dispersion medium; indeed, if the medium is carefully deionized by dialysis or similar techniques, the double layers can extend so far that even dispersions as dilute as 2 % phase volume can show solid-like characteristics and, if the particle size is uniform, quasi-crystalline ordering [102], [103].

Anisotropic particles often exhibit effective phase volumes different from their geometric volume. Such particles may also show specific effects due to interactions when oriented in velocity gradients, for example (rheopectic flow). *Micellar systems* often exhibit effects intermediate between emulsions and dispersions as concentration increases. *Liquid-crystal systems* share some of these properties, but their extreme anisotropy leads to many interesting and complex rheological phenomena [104]. The rheological properties of emulsions, micellar, and liquid-crystal systems are often extremely temperature sensitive.

Finally, net attractive forces between particles in dispersion or between particles and polymers dissolved in the medium can lead to the formation of complex space-filling structures (*flocculates*), whose effective phase volume may vastly exceed the phase volume of the particles alone. If such structures form a network throughout the volume of the liquid, it behaves like a solid at rest or low shear rates (infinite viscosity). If such structures are separated by a reasonable volume of medium, flow will occur but the viscosity will be much higher than that expected on the basis of the dispersion phase volume alone. However, unless the attractive forces involved are very strong or covalent bonding is involved (e.g., in a polymer gel), these structures will be broken down readily by shearing, and continuous shear rate measurements will indicate pseudoplastic or (more likely) thixotropic flow.

3.2. Dynamic (Oscillatory) Measurements

For a Newtonian liquid, the loss component G'' of the complex dynamic modulus of elasticity $G*$ shows a plot of slope of unity against frequency on double logarithmic axes, whereas the elasticity component G' is much lower [61], often impossible to measure. This is equivalent to a constant real component of the dynamic shear viscosity η' that is independent of frequency (i.e., Newtonian).

Few liquids exhibit viscoelastic responses that are characterized by single relaxation times. Exceptions are some micellar solutions, particularly a number of cationic surfactants, which in the presence of certain organic anions, form long stiff micelles. These solutions possess very definite viscoelasticity, probably due to the formation of transient networks resulting from random interlocking of the very long stiff micellar rods. A typical example is the 0.1 mol/L 1 : 1 molar ratio cetyltrimethyammonium bromide – salicylic acid solution (see Fig. 13, Section 1.3.3). Apart from these systems, neither simple supercooled liquids (supercooled to bring the transition region within the frequency range of the dynamic mechanical equipment; dielectric measurements do not have this restriction, but even at ambient temperature, nonassociating liquids usually exhibit a range of relaxation times [105], [20]) nor dispersion systems exhibit such behavior, although some dispersions may have very weak viscoelasticity of this kind over a limited concentration range [34]. Most simple liquids, micellar solutions, colloidal solutions, emulsions, and dispersions have a wider transition region from predominantly viscous to elastic properties. Thus, crossover ($G' = G''$) occurs at a lower frequency than the G'' peak, with the crossover frequency corresponding to the longest characteristic time of the system. Such a range of characteristic times reflects a restricted range of free volume and interactions available to each disperse phase particle or to each liquid molecule in the case of supercooled liquids (see Section 1.3.4).

The range of characteristic times is limited by comparison with a *polymer macromolecular chain,* where the chain can be modeled as a string of beads connected by springs, with each bead dissipating energy by frictional resistance with the surrounding medium (whether it be a liquid, as in polymer solutions, or other macromolecules, as in polymer melts). For flexible chains or high molar masses, the number of beads can be large, and the range of char-

acteristic times and the transition region correspondingly wide. The theories developed on this general model by BUECHE [106], [107], ZIMM [108], ROUSE [109], SLONIMSKY [110], and others show a particular distribution of characteristic times, and predictions based on these theories agree well with experimental data. A typical data set for a concentrated dispersion is shown in Figure 15, and for a polymer solution in Figure 19.

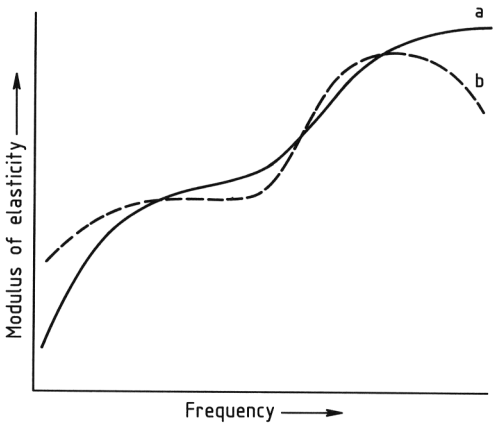

Figure 19. Oscillatory profile for polymer solution (diagrammatic)
a) Dynamic modulus of elasticity G'; b) Loss modulus G''

If the particles in a dispersion are *flocculated,* the floccules tend to be linear, highly branched, or fractal structures and then the possibility of wide distributions of characteristic times exists, with the maximum times being very large if structures filling the liquid volume are formed. This process becomes analogous to the changes in dynamic mechanical behavior of a gelling polymer system with reaction time (→Polymerization Processes, Chap. 2.1.3.). Thus, initially, the dispersion is viscous or viscoelastic with a wider transition region than Maxwell liquids. As flocculation or gelation proceeds, the transition region widens close to the gelation point. At this point, the region is so wide that both components of the complex modulus (G', G'') show power law dependence on frequency with exponents less than unity. Beyond this point, the strength and number of links in the polymer gel network or particle network increase. Because of this, the exponent of the modulus of elasticity G' – frequency power law relationship decreases toward zero, and G' becomes progressively greater than the loss modulus G'' over the entire frequency range (normal frequency ranges, 0.001 – 40 Hz, supplied by typical commercial rheometers are assumed). The changes and the viscoelastic properties characteristic of each stage of this process are shown in Figure 20.

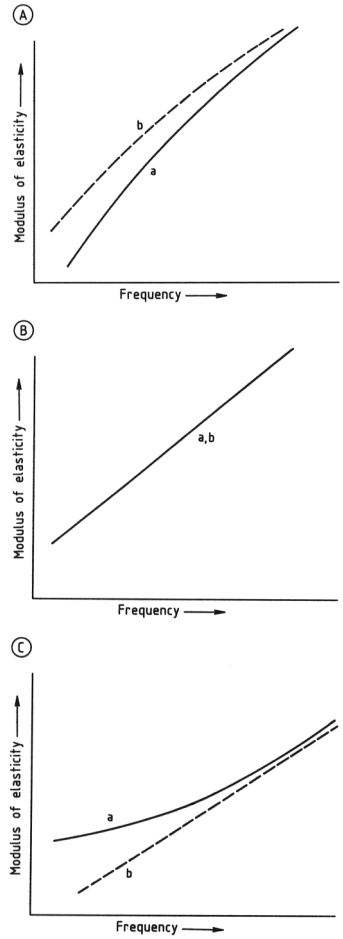

Figure 20. Oscillatory data for gelation process (diagrammatic)
A) Just before transition; B) Transition liquid→gel; C) Just beyond transition
a) Dynamic modulus of elasticity G'; b) Loss modulus G''

3.3. Transient Measurements

The two major types of measurements used are constant stress (creep) and constant strain (stress relaxation). Other methods – for example, measurements of stress transients on application of a constant shear rate or of stress relaxation on

removal of constant rate of shear – are less frequently employed because of measurement errors and limited applicability. They are used quite often to confirm the predictions of constitutive equations. Of the two main methods, creep often gives more obvious indications of structural change.

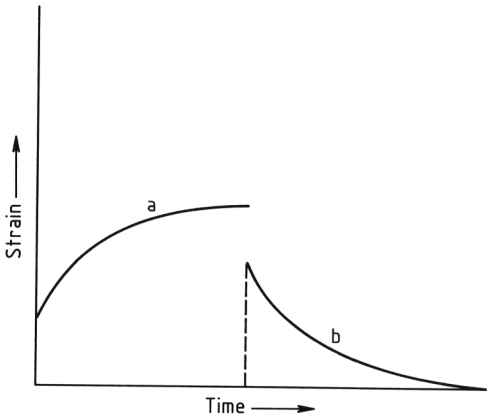

Figure 21. Creep (anelastic behavior)
a) Retardation (stress on); b) Relaxation (stress off)

Even allowing for instrumental errors, few *solids* exhibit perfect Hookean behavior, except at low stress or strain or short times. Many solids exhibit elastic behavior, as shown in Figure 21 for a creep experiment. This is typical of many metals (→Mechanical Properties and Testing of Metallic Materials, Chap. 2.7., →Mechanical Properties and Testing of Metallic Materials, Chap. 2.8.), polymers (→Plastics, Properties and Testing), etc. An initial instantaneous deformation is followed by a gradually decreasing deformation, which reaches a constant strain value after sufficient time. When the applied constant stress is removed, an instantaneous reduction in strain value occurs, followed by a slower time-dependent reduction of strain to the initial zero value. No irreversible deformation occurs, and thus no dissipation in the deformation process. Essentially, anelasticity can be defined as Hooke's law behavior extended to include the time dependence of deformation, as first done by Zener [111]. Two additional constraints are added: (1) at any value of time, the deformation (strain) is proportional to stress; and (2) given sufficient time, the strain will always reach an equilibrium value. Thus, although the modulus or compliance is time dependent, it is not strain or stress dependent.

Structured liquids and *weak solids,* such as paints and swollen gels, exhibit true viscoelastic properties. These are shown in Figure 22, again for creep experiments. Compared to elastic behavior, two essential differences can be seen. First, after a sufficiently long time for stress application, the strain becomes a linear function of time (i.e., some true viscous flow occurs, usually at avery low shear rate, and no equilibrium strain value is reached). Second, when the stress is removed, the strain value tends to an asymptote (i.e., a finite nonzero strain value at long times). So no recovery to zero occurs. However, when these plots are converted to compliance – time plots for different stress values, they may still be identical (i.e., the viscoelasticity is linear in nature). Apparently, in such materials, numerous strong-bonded volumes exist, whose boundaries have weakly bonded interfaces. At low stress, the weak bonds of the interface break, allowing some relative and irreversible movement of the volumes. Many of these interfacial bonds can reform in new positions due to the slow relative motion of the volumes. This ensures some recovery from the deformed state when the stress is removed [111], [112]. However, at higher stress, the response is still viscoelastic, but no longer linear. At still higher stress, total breakdown of structure and rapid flow takes place, as shown by the sharp increase in strain value with time and the slight or zero strain recovery when the stress is removed (Fig. 23).

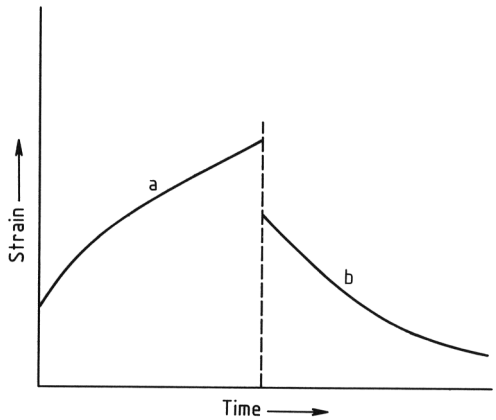

Figure 22. Creep (viscoelastic behavior)
a) Retardation (stress on); b) Relaxation (stress off)

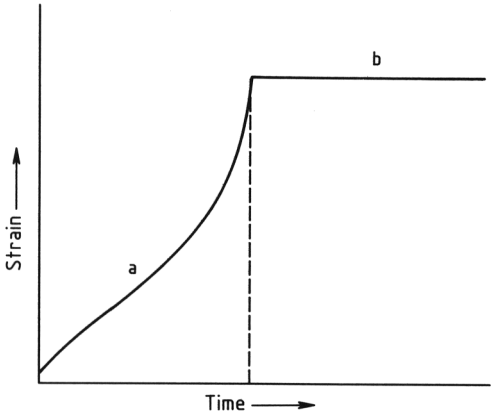

Figure 23. Creep (breakdown to flow)
a) Retardation (stress on); b) Relaxation (stress off)

Similar but less distinct changes can be seen in *stress relaxation plots* as the strain value increases. Thus an initial constant stress value is observed, which then declines either toward zero (flow) or to another constant value (gel). Progressive increase in strain may give modulus – time plots that are identical (linear response), and then the initial stress plateau may decrease or last for a shorter time before disappearing entirely. Finally, a rapid decline to zero stress indicates flow.

4. References

General References

1. H. A. Barnes, J. F. Hutton, K. Walters: *An Introduction to Rheology,* Elsevier, Amsterdam 1989.
2. K. Walters: *Rheometry,* Chapman & Hall, London 1975.
3. R. W. Whorlow: *Rheological Techniques,* Ellis Horwood, Chichester 1980.
4. J. R. van Wazer, J. W. Lyons, K. Y. Kim, R. E. Colwell: *Viscosity and Flow Measurement,* Interscience, New York 1963.
5. J. D. Ferry: *Viscoelastic Properties of Polymers,* 3rd ed., Wiley-Interscience, New York 1970.
6. M. Doi, S. F. Edwards: *The Theory of Polymer Dynamics,* Clarendon Press, Oxford 1986.
7. G. V. Vinogradov, A. Ya. Malkin: *Rheology of Polymers,* Mir Publishers, Springer Verlag, English version 1980.
8. G. W. Scott-Blair: *A Survey of General & Applied Rheology,* 2nd ed., Pitman, London 1949.
9. S. Oka in F. R. Eirich (ed.) *Rheology: Theory & Applications,* vol. 3, Academic Press, New York 1960, chap. 2.
10. J. M. Krieger, W. B. Russel, R. L. Hoffman, Th. F. Tadros in M. S. El-Aasser, R. M. Fitch (eds.): Future Directions in Polymer Colloids, *NATO ASI Ser. E,* no. 138, Martinus Nijhoff, Dordrecht 1987, chap. 6 – 10.
11. H. L. Goldsmith, S. G. Mason in F. R. Eirich (ed.): *Rheology: Theory and Applications,* **vol. 4,** Academic Press, New York 1967, chap. 2.
12. J. W. Goodwin in: *Specialist Periodical Reports of the Chemical Society (London) Colloid Sci.* **2** (1975) no. 7; *Adv. Colloid Interface Sci.* **42** (1992) 303.
13. J. Mewis, *J. Non Newtonian Fluid Mech.* **6** (1979) 1.
14. W. W. Grassley: "The Entanglement Concept in Polymer Rheology," *Adv. Polym. Sci.* **16** (1974);*Comprehensive Polymer Science,* vol. 2, Pergamon Press, Oxford 1989.

Specific References

15. G. D. Scott-Blair in [8], pp. 5 ff.
16. D. Halliday, R. Reswick: *Physics,* 3rd ed., Wiley-Interscience, New York 1978, pp. 482 – 483.
17. K. Walters in [2], p. 232.
18. M. Reiner, *Phys. Today* **17** (1964) 62.
19. A. B. Metzner, J. L. White, M. M. Denn, *Chem. Eng. Prog.* **62** (1966) 81.
20. J. Lamb, *Rheol. Acta* **12** (1973) 438.
21. D. C.-H. Cheng, *Rheol. Acta* **25** (1986) 542.
22. In [1], p. 166.
23. In [1], p. 159.
24. D. J. Shaw: *Introduction to Colloid and Surface Chemistry,* 4th ed. Butterworth/Heinemann, Oxford 1992, pp. 210, 220 – 221, 238 – 240.
25. H. L. Goldsmith, S. G. Mason in [11], sect. V, pp. 161 – 185.
26. N. Casson in C. C. Mill (ed.): *Rheology of Disperse Systems,* Pergamon Press, Oxford 1959, pp. 84 – 104.
27. British Standard, Glossary of Rheological Terms, BS 5168 : 1975.
28. J. M. Dealy, *J. Rheology (N.Y.)* **28** (1984) 181.
29. In [1], pp. 159 – 169.
30. K. Walters in [2], p. 233.
31. T. A. Strivens, *Colloid Polym. Sci.* **267** (1989) 269.

32. T. Shikata, H. Hirata, T. Kotaka, *Langmuir* **3** (1987) 1081.
33. K. E. J. Barrett (ed.): *Dispersion Polymerisation in Organic Media,* Wiley-Interscience, London 1975, p. 235.
34. T. A. Strivens, in press.
35. F. Mainardi, E. Bonetti, *Prog. Trends in Rheology (Suppl. Rheol. Acta)* **II** (1988) pp. 64 – 67.
36. A. P. Metzner, A. B. Metzner, *Rheol. Acta* **9** (1970) 174.
37. A. B. Metzner, *Rheol. Acta* **10** (1971) 434.
38. D. R. Oliver, R. Bragg, *Rheol. Acta* **13** (1974) 830.
39. K. Walters in P. Moldenaers, R. Kennings (eds.): "Theoretical and Applied Rheology," *Proc. Int. Congr. Rheol., 11th* 1992, vol. 1, p. 16.
40. K. Walters in [2], pp. 210 – 212.
41. C. D. Denson, D. L. Crady, *J. Appl. Polym. Sci.* **18** (1974) 1611.
42. T. A. Strivens in R. Lambourne (ed.): *Paint Technology,* Ellis Horwood, Chichester, 1987, chap. 15.
43. K. Walters in [2], pp. 133 – 160.
44. R. W. Whorlow in [3], chap. 3, pp. 113 ff.
45. T. A. Strivens, *J. Colloid Interface Sci.* **57** (1976) 476.
46. K. Walters in [2], Fig. 6.9, 6.10, p. 138.
47. K. Walters in [2], p. 47.
48. T. A. Strivens in [45].
49. M. C. Nally, *Br. J. Appl. Phys.* **16** (1965) 1023.
50. T. A. Strivens, unpublished results.
51. R. W. Whorlow in [3], pp. 138 – 142.
52. R. W. Whorlow in [3], p. 255.
53. H. Markowitz, *J. Appl. Phys.* **23** (1952) 1070.
54. K. Walters in [2], pp. 134 – 136.
55. K. Walters in [2], pp. 197 – 198.
56. K. Walters in [2], p. 95.
57. *Kirk-Othmer,* 3rd ed., **20,** 284.
58. Methods for the Determination of the Viscosity of Liquids, BS 188 : 1977 (1988).
59. ASTM D 445.
60. F. Nazem, C. T. Hill, *Trans. Soc. Rheol.* **18** (1974) 87.
61. F. N. Cogswell, *Plast. Polym.* **36** (1968) 109.
62. R. L. Ballman, *Rheol. Acta* **4** (1965) 137.
63. G. V. Vinogradov, A. I. Leonov, A. N. Prokunin, *Rheol. Acta* **8** (1969) 482.
64. T. J. Peng, *Jet Propulsions Laboratories Quart, Rev.* **2** (1972) 40.
65. J. Meissner, *Rheol. Acta* **10** (1971) 230; **8** (1969) 78.
66. A. Demarmels, J. Meissner, *Rheol. Acta* **24** (1985) 253.
67. K. Walters in [2], p. 222.
68. J. Mewis, A. B. Metzer, *J. Fluid. Mech.* **62** (1974) 593.
69. J. A. Spearof, A. B. Metzner, *Trans. Soc. Rheol.* **16** (1972) 495.
70. C. A. Moore, J. R. A. Pearson, *Rheol. Acta* **14** (1975) 436.
71. C. D. Han, L. Segal, *J. Appl. Polym. Sci.* **14** (1970) 2973.
72. D. Acierno, J. N. Dalton, J. M. Rodriquez, J. L. Little, *J. Appl. Polym. Sci.* **15** (1971) 2395.
73. N. E. Hudson, J. Ferguson, P. Mackie, *Trans. Soc. Rheol.* **18** (1974) 541.
74. F. A. Kanel: Ph. D. Thesis, Univ. of Delware, 1972.
75. G. Astarita, L. Nicodemo, *Chem. Eng. J. (Lausanne)* **1** (1970) 57.
76. G. Fano, *Arch. Fisiol.* **5** (1908) 365.
77. J. R. A. Pearson, J. T. F. Pickup, *Polymer* **14** (1973) 209.
78. T. Kizior, F. A. Sayer, *Trans. Soc. Rheol.* **18** (1974) 271.
79. D. R. Oliver, R. Bragg, *Can. J. Chem. Eng.* **51** (1973) 287.
80. D. R. Oliver, *Chem. Eng. J. (Lausanne)* **6** (1973) 265.
81. H. Gisekus, *Rheol. Acta* **7** (1968) 127; **8** (1969) 411.
82. F. N. Cogswell, *Polym. Eng. Sci.* **12** (1972) 64.
83. A. B. Metzner, E. A. Uebler, C. F. Chan Man Fong, *AIChE J* **15** (1969) 750.
84. G. Marucci, R. E. Murch, *Ind. Eng. Chem. Fundam.* **9** (1970) 498.
85. K. Walters in [2], p. 56; T. A. Strivens, unpublished results.
86. S. Onogi, T. Matsumoto, *Polym. Eng. Rev.* **1** (1981) 45.
87. K. Walters in [2], chap. 6.
88. K. Walters in [2], pp. 155 – 160.
89. T. E. R. Jones, K. Walters, *Rheol. Acta* **10** (1971) 365.
90. J. B. Enns, J. K. Gilham, *J. Appl. Polym. Sci.* **28** (1983) 2831.
91. H. A. Barnes, J. F. Hutton, K. Walters in [1], pp. 60 – 64.
92. G. V. Vinogradov, A. Ya. Malkin in [7], pp. 303 –321.
93. R. W. Whorlow in [3], pp. 177 – 187.
94. G. V. Vinogradov, A. Ya. Malkin in [7], p. 317.
95. E. F. Casassa, G. C. Berry in: *Comprehensive Polymer Science,* vol. 2, Pergamon Press, Oxford 1989, chap. 3.
96. T. A. Strivens, *Colloids Surf.* **18** (1986) 395; *Colloid Polym. Sci.* **261** (1983) 74.

97. R. Ottewill, *Prog. Colloid Polym. Sci.* **67** (1980) 71.
98. J. D. Ferry in [5], pp. 519 ff.
99. C. Parkinson, S. Matsumato, P. Sherman, *J. Colloid Interfac Sci.* **33** (1970) 150.
100. B. W. Barry, *Adv. Colloid Interfac Sci.* **5** (1975) 37.
101. T. G. Fox, S. Gratch, S. Loshack in F. R. Eirich (ed.): *Rheology: Theory and Applications,* vol. 1, Academic Press, New York 1956.
102. N. Ise, H. Matsuoka, K. Ito, H. Yoshida, J. Yamanaka, *Langmuir* **6** (1990) 296.
103. M. Tomita, K. Takano, T. G. M. van de Ven, *J. Colloid Interfac Sci.* **82** (1983) 367.
104. V. G. Kulichikhin, V. S. Volkov, N. A. Plate in *Comprehensive Polymer Science,* **1st suppl,** Pergamon Press, Oxford 1991, chap. 15.
105. G. Glarum, *J. Chem. Phys.* **33** (1960) 679.
106. F. Bueche: *Physical Properties of Polymers,* Interscience, New York 1962.
107. F. Bueche, *J. Chem. Phys.* **24** (1956) 269.
108. B. H. Zimm, *J. Chem. Phys.* **24** (1956), 269.
109. P. Rouse, *J. Chem. Phys.* **21** (1953) 1272.
110. G. L. Slonimsky, *J. Polym. Sci. Part C* **16** (1967) 1667.
111. C. Zener: *Elasticity and Anelasticity of Metals* Chicago University Press, Chicago 1948.
112. J. S. Kochler: *Imperfections in Almost Perfect Crystals,* Wiley-Interscience, New York 1952, p. 197.

Rhodanides → Thiocyanates, Inorganic
Rhodium → Platinum Group Metals and Compounds
Riboflavin → Vitamins
Risk Assessment → Chemical Products: Safety Regulations
Road Construction Materials → Asphalt and Bitumen
Rocket Propellants → Propellants

Laboratory Information and Management Systems (LIMS)

KLAUS SCHUCHARDT, Veba OEL AG, Gelsenkirchen, Federal Republic of Germany

REINHARD SCHYMIK, Unternehmensberatung Triestram & Partner GmbH, Bochum, Federal Republic of Germany

1. Introduction 1
2. Definition 2
2.1. Historical Development 2
2.2. Outline Solutions 2
2.3. User Expectations 2
2.4. Task Areas 2
2.5. Functionality 3
3. Objectives 3
4. LIMS in an Information Network . . 3
5. Nature of a LIMS 4
5.1. Data Areas 4
5.2. Functional Principle 4
5.3. Analysis – Sample 4
6. Forms of Order 5
7. Basic Functionality 5
7.1. Registration 5
7.2. Sample Distribution – Dissemination of Information 6
7.3. Work Planning 6
7.4. Recording of Results 6
7.5. Validation 6
7.6. Publication 7
7.7. Archiving 7
7.8. Reserve – Waste Disposal of Samples 7
7.9. Addition to Basic Functions 7
8. Expanded Functionality 7
8.1. Integration of Extraneous Programs . 7
8.2. Cost Accounting 8
8.3. Inventory Management 8
8.4. Internal Quality System 8
9. Adaptation of LIMS to Laboratory Tasks 8
10. LIMS as a Component of Quality Assurance 9
11. Development Prospects 9
12. References 10

1. Introduction

Laboratory information and management systems (LIMS) support both the administrative tasks of order processing and also the acquisition and assessment of results in the laboratory. LIMS "informs" about results and "manages" order processing.

The laboratory as a production center for discoveries – and so subject to the rules of production planning – is a new and unfamiliar idea. The laboratory is still being recognized as a creative (innovative) work center with a problem-oriented working procedure for nonrecurrent problems. Work of this kind is generally considered to be unpredictable and therefore only plannable to a limited degree. If this was true, however laboratory work could not be combined with the systematic arrangement that supports an information system, and data processing support in the laboratory would only be of minor importance.

In an analytical laboratory, orders and work flow can effectively be dealt with according to a plan and guidelines. For every laboratory there is a limited, yet established array of services depending on the instrumental equipment, the qualifications of personnel, and the objects to be analyzed.

Orders, samples, and analytical methods can be systematically described by a suitable data model.

Laboratory orders are plannable and can be carried out according to the rules of operations scheduling. The uncertainties of time for laboratory orders are greater for material and methodical reasons compared with production planning. In the laboratory operations scheduling is usually controlled with less attention to particulars. Nevertheless the preconditions exist for a time, capacity, and personnel plan together with those for an order processing procedure supported by data processing. Laboratory information and management systems undertake these tasks.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_307

2. Definition

2.1. Historical Development

The handling of large data streams such that arise in spectroscopy and chromatography, e.g., autosamplers in chromatography, led to the development of laboratory data systems. Initially these systems took over control functions for instruments, but were later developed into laboratory-wide management systems with the functional areas of data acquisition, evaluation, and order management. LIMS has developed from initially central solutions (host) with a comprehensive problem definition to administrative systems at departmental computer and networking level. Contributing to this development were the decreasing prices for hardware and the promising development of instrumental analysis (on-site information). The availability and use of relational data bank management systems represented a further innovative step in the development of LIMS. It also introduced a transition to the more extensive use of standard software, and solution by an individual program was increasingly supplanted. Most of LIMS development is definitely in the technical field, with little innovation in functionality, i.e., in the rules by which the data are processed. With respect to functionality, in addition to the sample-oriented character that is still preferred today, an extension to order- or project-related applications becomes apparent. At the same time a development fostered by quality assurance can be recognized: the LIMS application is no longer regarded as an isolated solution for the laboratory, but rather is understood to be a component or an element of quality management and assurance for the entire company.

2.2. Outline Solutions

In outline solutions, the goal is to start out from commercial standard software. This is supported by the rapid development of relational data bank management systems. One of the great advantages of such systems lies in a flexible adaptability of the functionality to (future) requirements. However, the use of relational data bank management systems simultaneously makes new and greater demands on the efficiency (speed, security of access) of the hardware.

The data processing implementation of LIMS was pushed further and further into the background in favor of the specific application.

2.3. User Expectations

The user expectations of data processing systems have developed with the available possibilities. The demands for user friendliness and operating convenience have increased. This development is taken into account by graphical user interfaces (GUI) and interactive auxiliary functions, but the user also expects an improved and expanded system functionality. Systems can be adapted and configured to different tasks by the user without programming operations. One of the areas of LIMS development is the integration of software extraneous to LIMS. The specialized software packages are so integrated into the functionality of LIMS that a continuous form of processing is obtained. Globally, the task of LIMS as the exchange of data between LIMS and other data processing applications becomes an absolute necessity.

2.4. Task Areas (see Fig. 1)

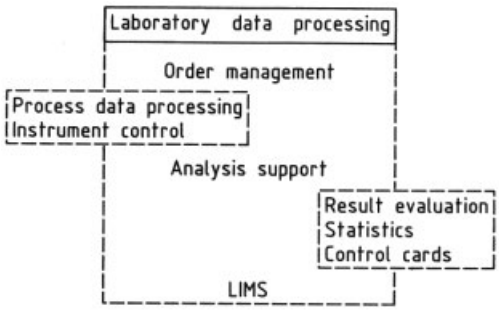

Figure 1. Laboratory data processing using LIMS

LIMS mainly covers the administrative part of support in the laboratory. The main objectives are order management and the collection of results together with the resulting archiving functions. Traditionally, LIMS has been a sample- or object-oriented system for an analytical laboratory. LIMS is being expanded to form a support for an order-oriented or (more

comprehensively) a project-oriented operating procedure of research and development laboratories. The *sample-oriented* form can be understood as a subgroup of the project-oriented form. By this definition, the *order-oriented form* is the connecting link between these two types. The *project-oriented form* is based on the processing and examination of samples, but in addition it contains features of a system of project management with respect to time and content and a system of project documentation.

2.5. Functionality

The LIMS functionality is directed to the task of providing a service. No recognized LIMS standard exists, but there is substantial agreement on the functionality of LIMS applications. The user expects and assumes large functional differences in the individual LIMS applications. However, since the laboratory problems are in fact not that different, the real functional differences between systems result essentially from the limitation of central LIMS functions to individual tasks specific to the company. These include, for example, terminologies, reporting, and forms of data evaluation. The central functionality, i.e., the actual laboratory procedure, shows no fundamental differences. In terms of the LIMS data model, this means that the assumed differences lie mostly in the information that describes tasks and samples and not in that necessary for the operational sequence.

3. Objectives

Compared with other investments in the laboratory, the introduction of a LIMS is exceptional by virtue of the costs and the project duration. In addition to the external costs for hardware, software, and services, further requirements are the preparation of the master data and introduction of the system. This investment requires an economic justification. Measurable and quantifiable benefits are *rationalization effects*, performance increases, and task extensions, which occur both in the laboratory and in other areas of the company. Further positive but not immediately quantifiable consequences result from the higher standards of work expected.

In addition to rationalization, the second aspect, an *improvement in quality* of the laboratory work, is equally important together with the creation of *centralized information* on quality data. Aspects such as *quality assurance* and *data exchange* have also become more significant. Without a LIMS, developments as initiated by computer-integrated manufacturing (CIM) or total quality management (TQM) are scarcely conceivable and not practical.

The systematic approach leads to working rules for the laboratory that are indispensable for a quality assurance system. In addition, the laboratory work becomes both internally and externally comprehensible.

Following the systematic approach required by LIMS, a data base also is created that allows meaningful archive evaluation. Of the secondary reasons for introduction of LIMS, the continual availability of information is especially important.

As a result of appropriate data processing support as provided by a LIMS, conditions are created for evaluating data at a complexity appropriate to a specific problem. Manual data processing, which also is possible in principle, is generally not practiced due to the cost of reliable data gathering, the time required for processing, or the lack of technical competence.

4. LIMS in an Information Network

A laboratory data system must not be considered solely from the perspective of the laboratory; the global company environment also must be taken into account. This results in additional demands for extending communication beyond the laboratory.

Data that are valid throughout the company are thus incorporated in the master data base of a LIMS. In principle, therefore, such general data could be accessed by the LIMS (i.e., not filed in the master data base) to avoid a redundant holding of data (and consequently problems with the data files developing in separate directions). However, the utilization of general data files by LIMS instead of using its own master data base is not yet possible, since the continual availability of data for the LIMS cannot yet be guaranteed. In the same way, organizational difficulties

such as systematic arrangement, competence, or responsibility are obstacles to the consistent application of distributed data banks.

Alongside the LIMS computer system in the laboratory, there are usually other computing systems (GC, HPLC, spectroscopic, and workstations). These subsystems are to be subsidiary to the LIMS. At the same time the distributed intelligence at work stations and analytical instruments is used so as to prevent too much interference with analytical processing as a result of malfunctions. Outside the laboratory area, computer applications exist in which laboratory or laboratory-relevant data are processed or managed (e.g., commercial applications, process control systems).

The demarcation of LIMS to systems of the latter kind is relatively easy, because these processing and managing systems do not directly influence current laboratory procedures. In these external systems, data are filled once the processing of an order is complete. Data from these systems are retrieved by LIMS for updating its own data files. This data transfer is merely a technical problem (and a problem of data format conversion). The inclusion of analytical subsystems in LIMS, on the other hand, which lead to different forms of data exchange requires individual, complex solutions. The data flow resulting from this must be fully integrated into the LIMS functionality.

5. Nature of a LIMS

A LIMS is described by its data model and the functionality. Within this framework a LIMS is designed by the user. The description of the analysis and the samples is particularly important since the samples constitute the object of processing, the analysis the form of processing in the laboratory work.

5.1. Data Areas

In principle, three forms of logical data area exist in a LIMS:

1) Master data
2) Dynamic data or order data
3) Archive data

The *master data base* contains the "knowledge" of the laboratory, and thus all information necessary for carrying out the work there. It consists of general, analytical, and task-describing tables. The information contained in the master data base includes among other items the description of the organization, the authorities of personnel, and the descriptions of equipment, analytical methods, and samples. In addition, processing of different orders is described in the master data base. This includes sampling plans, analysis plans, monitoring of testing agents, and processing strategies according to experience. When an order is received, this master data base is always re-accessed and with the information filed there a task is set up in the *dynamic data area*. The raw data are assembled in this temporary data area and, with validation, transferred to the *archive area*. The archive thus originates in the master data.

5.2. Functional Principle

A LIMS is an information system built on experience. Flexibility and adaptability of the system therefore depend on the collection of master data which reflect the user's experience and his "view of the laboratory world". The master data are updated in a dynamic process by new experiences. Due to this continually expanding basic knowledge, LIMS are influenced by experience and is adaptive. If a master data base is of poor quality, therefore, only limited benefit can be expected from the use of even functionally very powerful systems.

5.3. Analysis – Sample

The analytical method is a measuring process that is applied to a sample (object) and whose outcome can comprise one or more results. The sample itself is not characterized by the analytical method, but is described by its own characteristics. The processing of samples (analyses) generates the processing order. The project is made up of various orders (individual steps of a project). The existing logical distinctions between project, order, sample, and analytical method must be reproduced in the data base of the LIMS.

Structure of an Analytical Method. An analytical method is modeled in LIMS in the form of its algorithm (procedure). This procedure must not process nonanalytical information, otherwise the necessary validity is lost.

Test Values – Statistical Characteristics. In addition to the procedural models, method-related statistical evaluations of analytical results must be defined in LIMS. The individual results are characterized by range of application, significance, and precision (e.g., standard deviation, repeatability, reproducibility), which are all filed together with the analytical method.

Calibration. For the calculation of analytical values, prior calibration or access to calibration values for the instruments or methods used is necessary in many cases. These calibration values and the calibration planning are assigned to the method, with consideration taken for the instrumental equipment.

6. Forms of Order

In principle, three forms of order are possible:

1) Frequency-controlled orders
2) Event-controlled orders
3) Random orders

For frequency- and event-controlled (routine) orders, the order data are known within the LIMS, but the *random order* must first be set up on the basis of the master data. It is therefore advisable to provide different functions for recording routine and additional orders.

Production monitoring is usually carried out in a frequency- or event-controlled manner. LIMS uses the corresponding work schedules for several purposes:

1) Sampling schedule
2) Automatic prerecording of orders
3) Order recording
4) Capacity planning
5) Cost accounting

In addition to the routine and medium-term processing of individual samples, all-encompassing mechanisms that enable long-term and project-oriented processing are desirable. These forms of order require individual activities to be planned with regard to time and content. The project orders filed as a network are monitored and controlled by LIMS by means of statements of internal costs and status reports.

From the recorded project schedules, the project activities are described in detail by means of a sample-oriented order-processing schedule. With this, a new sample-oriented order which describes the processing of actual samples, is created for each individual step of the whole project.

In the description of an order the LIMS treats separation operations (phase separations, distillations, preliminary separations, etc.) as conventional sample-preparation techniques. In addition, the individual subsamples are described and the corresponding analysis is allocated to them. Only in this way is an unequivocal allocation of results possible in sample calculations and evaluation methods (and project).

7. Basic Functionality

The functionality of a LIMS is marked by five key areas:

1) Maintenance of master data
2) Order creation
3) Analytical processing
4) Management functions
5) Archive management

7.1. Registration (see Fig. 2)

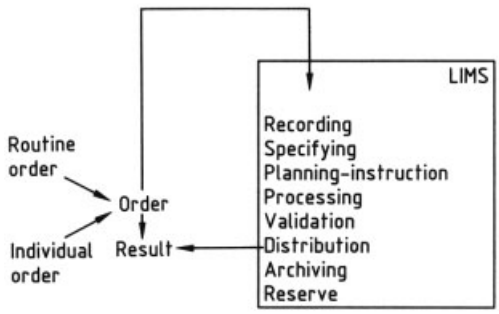

Figure 2. Basic functionality of LIMS

The work flow in a laboratory begins with the *recording* of an order to examine an object. This

means a description of the object to be examined, by which it can be identified in the laboratory and used for the subsequent report to the orderer or for subsequent evaluation. The order (i.e., the required analyses, tests, etc.) is allocated to the object identified. Thus with the recording of an order a processing objective must be known. By means of the work order recorded, LIMS draws up the individual orders for the laboratories and work groups involved. The actual implementation of an order is adapted to the nature of the order so that the necessary operating sequence can be efficiently arranged:

Routine Record. With the aid of a routine sample plan, a task list is drawn up automatically from the master data. The registration can, therefore, be carried out by LIMS independently or by the user in confirmation of receipt of the sample (order). For sampling that occurs before the recording of the order, LIMS can draw up task lists for the samplers in advance of the routine sample plans.

Individual Record. Individual orders are recorded by the user, and at the same time master data are re-accessed and the order adapted to the requirement on the basis of the system's proposals.

All data of an order must be modifiable until processing is complete. The modifications must be performed by an authorized person and documented so that they can be understood later on.

7.2. Sample Distribution – Dissemination of Information

Labels, sample waybills, raw data forms, and assorted task lists for work planning can be printed with the aid of LIMS within the laboratory. However, the objective is to ensure the availability of information even without printed lists.

Information about the order, processing situation, and whereabouts of the samples of an order must be available through the LIMS to all those entitled to it. The samples are distributed to the working groups and LIMS upholds the checking of the samples' whereabouts.

7.3. Work Planning

The user decides on the processing sequence with the aid of the task lists drawn up by the LIMS. Since in most LIMS the master data base does not contain information on the capacity of the individual working groups and orders cannot be prepared by LIMS according to their actual processing (network plan), LIMS are not able to carry out independent operations scheduling. If analytical subsystems are attached, generally the person in charge must coordinate the loading of an autosampler between LIMS and subsystem. In principle though, methods exist for automatic sample identification by reading devices.

7.4. Recording of Results

The measurements are carried out and the raw data are recorded in the LIMS. In the next processing step, the LIMS calculates the results from the raw data. If several series of measurements exist, LIMS combines the results to give mean values and calculates the actual precision. The results are examined for their compliance to the range of application and precision as a function of the analytical method. In a further testing stage, the results are compared with expected sample-dependent values (e.g., specifications). Finally, after the completion of all testing stages, the results are released by the person in charge of the analytical method.

7.5. Validation

The results are subjected to a stepwise validation in the LIMS. Already during the first recording, the range of application, precision, and the desired object-related values were controlled. After the release of the individual results, a further checking step follows.

For this purpose, LIMS compiles the results produced at various processing units and tests these results for evaluations. The LIMS constructs the analytical report of the order from the combination of individual validated results.

At the general report stage a final validation occurs with inclusion of possible plausibility tests. In this stage, the LIMS notes the authority and responsibility of the user so as to eliminate unauthorized release.

7.6. Publication

The printout of completed (validated) analytical reports is automatically checked by LIMS. If a printout has to be distributed, then a centralized distribution is preferable to improve the handling efficiency. As distribution is manual the LIMS assists in sorting and copying reports according to requirements. The objective is also, in this case, to make results available to other receiving computer systems (process computers, administration computers, etc.) in order to avoid re-inputs of data.

Although LIMS allows the creation of the individual report as an exception, generally LIMS should produce reports in systematized style which can be reduced to four basic forms:

1) Summary
2) Standard report (e.g., according to EN45001)
3) Chronological survey
4) Survey by object (plant, product, etc.)

7.7. Archiving

After validation and release, the activity data from the order are converted into archive data and, therefore, withdrawn from any further editing. Herewith the analytical reports acquire document status. Direct amendment of these archived data (reports) is disallowed for documentation reasons. However, well-founded invalidity declarations for individual results and the nomination of new findings with appropriate authorization and justification are allowed. Recipients on the original distribution list of the report are informed automatically by LIMS of any correction (even authorized) of archive data.

7.8. Reserve – Waste Disposal of Samples

As specified in the order record, samples from processing are disposed of as waste or transferred to a reserve, which is also managed by the LIMS.

7.9. Addition to Basic Functions

The basic functionality of a LIMS has to be expanded for actual laboratory requirements. The processing paths described here are certainly adequate for most operations, but further requirements of the functionality are intrinsic to the management of samples during and after processing. An *expanded basic functionality* of order processing includes the correcting and annotating orders as well as assembling order and task lists according to various criteria. Further functions make it possible for the LIMS to produce compilations selected from the archive data and process them graphically. Through the use of interfaces, the selected data can be made available to other data-processing applications. The LIMS provides the user with means of creating reports and defining the reports' contents.

8. Expanded Functionality

8.1. Integration of Extraneous Programs

The compilation of results is followed by the evaluation phase. This means that the LIMS must be able to compile data which can then be processed in other programs. Such an interface is not restricted to the programs mentioned here (e.g., word processing, statistics, and graphical presentation), but is also necessary for integrating the LIMS into a data-processing environment extending throughout the company. Although the same basic approach is followed when integrating LIMS into a global data processing environment, complicated individual solutions also have to be expected.

The integration of a *word processing system* presents a special problem, since the analytical results are not subject to the same protection rules within the word processing system as within LIMS; i.e., they can be modified rather easily. This difference can be desirable, but it raises the question of the reliability of data taken over from LIMS in the outgoing reports. The recipient must therefore be able to distinguish between the different forms of report.

8.2. Cost Accounting

In the cost documentation on laboratory expenses, a LIMS can distinguish between personnel and material costs. The costs per "customer" are assigned through collection at cost centers which makes invoicing possible. Capacity utilization statistics for internal laboratory purposes can be generated from this accounting information. These statistics concerning costs, work load, etc., can provide a basis for medium-term investment decisions.

8.3. Inventory Management

Because all the necessary items of information are available from the order processing, LIMS can accept them into an inventory management system. LIMS continually offers an exact survey of the stock of reserve samples, of stock location, and the storage time.

8.4. Internal Quality System

The internal quality system controls the recording of reference materials and the proper use of analytical instrumentation, maintenance, calibration measurements, reference measurements, repairs, and purchases, etc. In this way LIMS ensures that only sucessfully checked analytical systems and methods are used for the processing of orders.

In addition, the stocks of reference materials and standards are maintained and managed within LIMS.

9. Adaptation of LIMS to Laboratory Tasks

If an electronic data processing system is to be introduced into an existing (laboratory) environment then it must be adapted to it. But since electronic data processing offers the possibility to perform new tasks and modify job processing, changes must also be made in the laboratory field for best utilization of LIMS.

Since the laboratory sequences are similar in all areas of industry, LIMS have developed as broard-based solutions. With similar central functionality, the differences when applied to different laboratory tasks lie particularly in the screen display layout, the logical data structures, and the functional complexity.

Therefore most of the laboratory requirements can be met by a ready-made system if its level of technical development is sufficiently high. A small proportion of the requirements can only be met with limitations; and an even smaller fraction of tasks requires actual changes or extensions. Slight adjustments within the screen display layout and functionality will now and then become necessary after operating experience.

Outline Solutions. After drawing up the performance specifications and examining the profitability, there are two alternatives for the introduction of LIMS:

1) Purchase of a ready-made system (and possibly then adapt)
2) In-house implementation

The purchase of a system must be preferred since, although in-house implementation certainly enables better adaptation to the requirements, it also entails a considerably greater risk. Besides, an individual solution is not subject to general development as a commercially available system. The standard solution demands compromises of the user. These are also necessary in the case of an individual solution, for reasons of cost and efficiency.

With the introduction of LIMS, the desired final state of organization has to be defined and compared with the present state in order to achieve the highest possible degree of systematization. Existing complexities in a laboratory environment are generally determined by past experience and resources and can, in all probability, be simplified or abandoned.

On average, two years are required to implement a LIMS for a medium-size application. Because of the complexity of the system, it is advisable to introduce LIMS stepwise in order to have a consolidation phase after each stage. Implementation can be done in three steps:

1) General LIMS introduction
2) Integration of analytical subsystems
3) Integration into the company's communication network

After the first phase, the LIMS functionality is available to the laboratory. In the second phase, operations are optimized by on-line recording of results. This concludes the adaptation of LIMS within the laboratory. In the third phase, areas outside the laboratory are integrated into the scope of the LIMS.

Adaptations. In principle, LIMS can be extensively adapted to the user's requirements. However, the adaption possibilities available to the user must be clearly distinguished from those available to the manufacturer of the system. In general the *user* can only adapt the system through arrangement of the master data. For the *system manufacturer* it is possible to make both adaptations conforming to the system (arrangement of the display surface) and changes of data structures and functionalities deviating from the system. A change of data structure and functionality can be equated to an individual approach. A LIMS with modified data structure and functionality is not subjected to general further development and maintenance by the manufacturer.

Systems originating from the use of powerful development tools are to be regarded similarly. The advantage here is the more probable minimization of the associated high risk compared with a new individual solution. With any individualistic approach, the future user and the system manufacturer need to be highly competent.

10. LIMS as a Component of Quality Assurance

Extensive responsibilities for documentation are imposed on the operator of a laboratory through the requirements of Good Laboratory Practice, an accreditation according to EN 45001, or in the framework of certification according to ISO 9001-3. A LIMS is basically not a quality assurance system, but rather an object to be itself safeguarded within such a system. That is, the LIMS must be arranged so as to guarantee the requirements of a testable mode of operation. However, within a laboratory a LIMS satisfies many quality requirements through the retention of master data necessary for the system and through the work flows specified by the system. Consequently, the LIMS considerably simplifies the observance of the quality requirements mentioned (and even makes them practically realizable from the expenditure perspective.

A version management of the entries can thus be carried out within the maintenance of the master data. With that, a methods and product archive, among other things, builds up. The handling of testing agents, standards, and reference materials is controlled and documented by LIMS. With proper application of the LIMS a task can be reproduced at any time from the recorded and stored data by reviewing the acquisition of results for an order and the associated decisions.

At the same time the LIMS ensures that in carrying out their duties all users always have the up-to-date information available.

11. Development Prospects

The main emphasis in the development of LIMS will be its further functional integration, into the laboratory itself and within the company. Project-related operating principles and control of the work flows are the main emphases of development in the laboratory. In production control, further possibilities for the laboratory are appearing in the automation of measuring procedures. The penetration of information and function throughout the company will involve the potentialities of a LIMS particularly in quality assurance, production control, and information appraisal. Further technical developments (hardware, standard software) are of secondary importance for the LIMS, since their result is not to develop the application per se of the system but rather primarily to improve its potential uses.

In addition to these lines of development, which mean an updating of the known system functionality, an innovative development step is conceivable. The LIMS, hitherto only passively recording and archiving data, can actively interpret data relationships on a structured knowledge base and consequently influence analytical procedures. Economic considerations will determine which of the directions of development will be favored. Of these, project LIMS, integration within the company, and aspects of quality assurance are emerging as the main points of concern.

12. References

1. S. Ebel: "Anforderungen an Labordatensysteme," *Nachr. Chem. Tech. Lab.* **36** (1988) M 1 – M 14. K. Schuchardt, G. Fischer, A. Golloch: "Labor-Informations- und Management-Systeme," *Nachr. Chem. Tech. Lab.* **40** (1992) M 1 – M 20. R. R. Mahaffey: *LIMS Applied Information Technology for Laboratory,* Van Nostrand Reinhold, New York 1990. R. D. Mc Dowall: *Laboratory Information Management Systems,* Sigma Press, Wilmslow, UK, 1987. V. Neitzel: *Labordaten-Verarbeitung,* VCH, Weinheim 1992.

Lacquers → Paints and Coatings

Process Control Engineering

Martin Polke, Lehrstuhl für Prozeßleittechnik, RWTH Aachen, Federal Republic of Germany (Chaps. 1, 2, Sections 5.1 – 5.3, 6.1, 10.7, 11.1 – 11.3, Chap. 14)

Wolfgang Ahrens, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 2, Section 13.1)

Ulrich Pallaske, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 3.1, 3.3, 3.4)

Rudolf Metz, Bayer AG, Leverkusen, Federal Republic of Germany (Section 3.2)

Norbert Kuschnerus, Bayer Ltd., Tokyo, Japan (Section 3.5)

Ulrich Epple, GfP, Gesellschaft für Prozeßtechnik, Aßlar, Federal Republic of Germany (Sections 4.1, 4.5, Chaps. 7, 14)

Ernst-Dieter Gilles, Institut für Systemdynamik und Regelungstechnik, Universität Stuttgart, Federal Republic of Germany (Section 4.2)

Jörg Raisch, University of Toronto, Department of Electrical Engineering, Ontario, Canada (Section 4.2)

Herbert Wehlan, Institut für Systemdynamik und Regelungstechnik, Universität Stuttgart, Federal Republic of Germany (Sections 4.3, 4.4)

Michael Heim, Lehrstuhl für Prozeßleittechnik, RWTH Aachen, Federal Republic of Germany (Sections 5.1 – 5.4, 6.1, 11.2, 11.3)

Reiner Hotop, Bayer AG, Leverkusen, Federal Republic of Germany (Section 5.1)

Eberhard Nicklaus, Bayer AG, Leverkusen, Federal Republic of Germany (Section 5.1)

Wolfgang Noerpel, Bayer AG, Leverkusen, Federal Republic of Germany (Section 5.1)

Michael Freytag, Bayer Antwerpen N.V., Antwerpen, The Netherlands (Sections 5.1, 5.2)

Klaus Hartmann, Miles Inc., Pittsburgh, PA 15205-9741, United States (Sections 5.1, 5.2)

Günter Pinkowski, Krohne Meßtechnik, Duisburg, Federal Republic of Germany (Section 5.5)

Hans-Josef Schneider, Bayer AG, Leverkusen, Federal Republic of Germany (Section 6.2)

Michael Björkmann, ABB Ströberg Drives, Vaasa, Finland (Section 6.3)

Wolfgang Mutz, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 6.4, 10.7)

Hartwig Steusloff, Fraunhofer Institut für Informations- und Datenverarbeitung, Karlsruhe, Federal Republic of Germany (Chap. 8)

Gerd-Ulrich Spohr, Siemens AG, Köln, Federal Republic of Germany (Chap. 9)

Roland Vogelsgesang, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 10.1 – 10.6)

Norbert Ingendahl, Lehrstuhl für Prozeßleittechnik, RWTH Aachen, Federal Republic of Germany (Section 11.2)

Karl-Heinz Schmitt, Bayer AG, Krefeld, Federal Republic of Germany (Section 11.4)

Hasso Drathen, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 12)

Dietrich Balzer, Elpro AG, Berlin, Federal Republic of Germany (Sections 13.2, 13.3)

1. Introduction 3

2. Information Structures in Process Control Engineering 5

2.1. Principles 5

2.2. Architectural Principles for Information Structuring 9

2.3. Applications in Process Control Engineering 24

3. Knowledge about the Process 39

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b06_317

3.1. Principles 39
3.2. Analysis Methods for Process Quantities 44
3.3. Process Models 49
3.4. Modeling 58
3.5. Management and Utilization of Information 60
4. From Process Knowledge to Process Control 72
4.1. Principles 72
4.2. Feedback Control 74
4.3. Optimal Control 87
4.4. Binary Control 90
4.5. Operational Control of Process Plants 94
5. The Process Control System and Its Elements: Process Sensor Systems . 112
5.1. Principles 112
5.2. Process Sensor System Technology . 126
5.3. Sensor Systems for Special Applications 141
5.4. The Market for Sensors and Sensor Systems 170
5.5. Field Installation and Cable Routing 174
6. The Process Control System and Its Elements: Process Actuator Systems 184
6.1. Principles 184
6.2. Actuator Systems for Material and Energy Streams 189
6.3. Electrical Drives in the Chemical Industry 202
6.4. Electric Power Supply Systems . . . 210
7. The Process Control System and Its Elements: Distributed Control Systems 215
7.1. Principles 215
7.2. System and Component Structure . 226
7.3. Process Control Operating System . 241
7.4. General System Services 248
7.5. Design and Commissioning 257
8. The Process Control System and Its Elements: Information Logistics . . . 262
8.1. Principles 262
8.2. Functional Structures and Information Flow in Production Companies 262
8.3. Computer Communications Between and Within Control Levels 263
8.4. Computer Communications in Industrial Production; Standards . . . 265
8.5. MAP/TOP: Protocol Standards for Information Integration in Production Companies 266
8.6. Field Bus Systems 269
8.7. Quality Assurance: Conformance and Interoperability Tests 276
8.8. Methods and Tools for Protocol Specification 277
8.9. Steps toward Computer-Integrated Production 280
9. Computer-Aided Methods 280
9.1. Principles 280
9.2. System Analysis 281
9.3. CAE System for Process Control Engineering 282
9.4. Structure of a CAE System 283
9.5. Aids for Hardware Design 284
9.6. Aids for Software Design 286
9.7. Outlook 289
10. Design and Construction of Process Control Systems 289
10.1. Principles 289
10.2. Organizational Requirements 291
10.3. Decision Phase 301
10.4. Specifications 302
10.5. Execution Phase 313
10.6. Quality Assurance 320
10.7. Process Control Rooms 325
11. Operation 330
11.1. Principles 330
11.2. Human – Process Communications . 335
11.3. Process Analysis and Process Optimization 354
11.4. Maintenance Strategies 366
12. Standards, Committees, and Associations 371
12.1. Principles 371
12.2. Standardization Bodies and Other Organizations Involved in Standardization 372
12.3. Technical and Scientific Bodies 378
12.4. Shows and Fairs 384
13. Integration of Knowledge-Based Systems in Process Control Engineering 387
13.1. Principles 387
13.2. Knowledge-Based Approach 390
13.3. Knowledge Engineering 401
14. Appendix 410
14.1. Glossary 410
14.2. Abbreviations 424
15. References 426

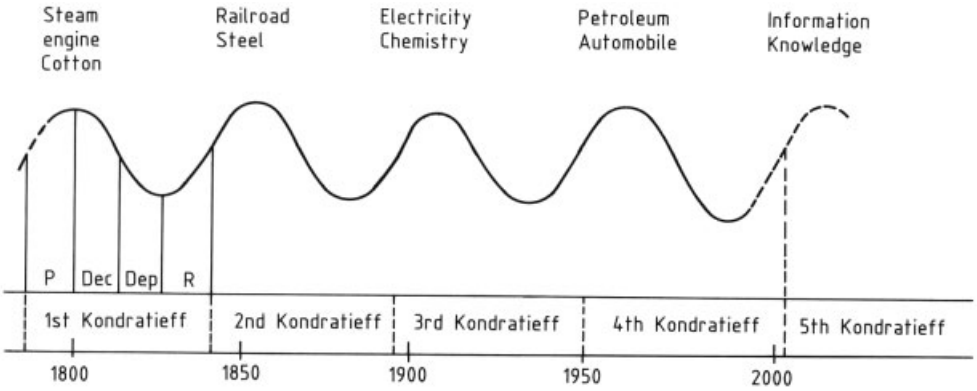

Figure 1. The Kondratieff waves [116]
P = prosperity; Dec = decline; Dep = depression; R = recovery

1. Introduction

Since the invention of the steam engine in the 1700s, the world has gone through four long economic cycles, which were described by the Russian economist Nikolai D. Kondratieff (1892 – 1930) and are known as Kondratieff waves or cycles [115].

Each upward phase (identified as "prosperity" in Fig. 1) was borne by inventions such as the steam engine (Watt, 1769), the mechanized weaving loom (Cartwright, 1784 – 1786), the steam locomotive (Stephenson, 1814), the electrodynamic principle (Siemens, 1866), the diesel and Otto motors (1893 – 1897, 1876), and fundamental innovations in chemical process engineering such as the manufacture of basic chemicals.

The world economy entered another long upward phase at the beginning of the 1980s.

The development of the new basic innovation called "information technology", characterized by the invention of the transistor [117] and the subsequent rapid development of miniaturization and laser technology, but also by results from the field of informatics such as object orientation [118] and Petri networks [119], is described in depth in [120]. The economic boom identified with the new resource "information" [121] is a global phenomenon with heretofore unknown dynamic qualities. International business, national economies, and individual companies find in it growth and innovation potential that could scarcely have been imagined. Porter describes the importance of information as an essential resource for a successful corporate strategy [122]: *"The question today is not whether information technology has significant effects on a company's competitive situation, but merely when and how these effects will become apparent. . . . Anyone who does not react today will be compelled to follow a course that others have set."*

The fifth Kondratieff cycle, however, is not only an imposing technical and economic upswing but carries along with it profound social, ecological, cultural, and intellectual changes and challenges. These connections, often revolutionary in scope, were pointed out by Schumpeter [123].

The fifth Kondratieff cycle will render permanent changes in the way the individual thinks and acts and also into the strategies that corporations and national and international institutions will pursue. A technological advance will become a structural change [124].

Along with the classical production factors–energy, labor, raw materials, and proven production technology–information has become not just an additional factor but probably the most important factor.

CIM (computer-integrated manufacturing) and CIP (computer-integrated processing) are not merely slogans today [125]. Demands made by society, such as increased environmental awareness, better and more consistent product quality, improved delivery, and more econom-

ical management, necessitate production facilities and processes with greater flexibility and ease of operation.

Strategy, that is, the management of the company as a whole, means identifying goals and achieving them. Because a company's definition of goals must always be dynamic, it is important to inquire into what impact changes in the company's environment will have. As trends in sociocultural, ecological, technological, domestic economic, international economic, and politicallegislative settings become stronger and change more rapidly, optimal access to information becomes a central requirement for a successful strategy (e.g., in marketing, sales, and production).

But strategy also means making available the means for business success. An essential success factor is thus the exploitation of new information technologies based on engineering, physics, and information science.

Information is an unlimited resource that can be managed only if it can be structured. In the economic, administrative, and technical spheres, those responsible for information management must optimize the utilization of information by coordinating all parties involved and by creating a technical infrastructure.

The German term *Prozessleittechnik* (process control engineering) was coined at Bayer AG in 1980 as a working title covering the measurement, control, and electrical engineering groups [126]. The expression was chosen to denote the integral view of all those concerned with the management or control of production programs based on engineering processes.

Since that time, a number of events have made it clear that classical or signal-oriented measurement and control technology is being supplanted by information-oriented process control engineering.

Long discussions among experts regarding the terms "control" and "automation" came to an end when Lauber stated a valuable dichotomy, placing humans at the center of the analysis [127]: The role of humans can be (1) to specify the framework and the sequence of the process, which then runs automatically, or (2) to gain the maximum information about the process and intervene in it directly, becoming the agent that ultimately controls it. This question is discussed in more depth in Section 4.3. The authors of this article have a clear preference for the more universal term "control" and make use of various levels or degrees of automation (see also [128]).

"Essential impetus for development today is coming from information technology, which has become a significant tool in measurement and automation technology. Only information-oriented technology enables the industrial user to operate processes and plants in compliance with the criteria of flexibility, productivity, safety, and environmental protection. The integration of humans into the production process is of great importance if the operator must, for example, perform control functions at the interfaces with the automation system and with the process" [129].

The national and international standardization organizations of greatest relevance to process control engineering, such as the International Electrotechnical Commission (IEC) with its Technical Committee TC 65 and the German Electrotechnical Commission (DKE) have also replaced the term "measurement and control technology" with the new complex of "control engineering" (see Chap. 9). The Standardization Working Group on Measurement and Control in the Chemical Industry (NAMUR) had already announced such a new orientation in September 1987 [125].

The strong influence of information technology on automation technology and process control engineering is visible not just in the area of devices and systems (especially process control systems; see Section 5.2), but in all aspects of design, construction, operation, maintenance, and so forth (procedures) [130].

BASF has made their training programs in operational practice available to the public [131]. Detailed procedural instructions can be found in Strohrmann's [132] course on process control engineering (serialized in *atp* since 1984). Gruhn et al. [133] and Lauber [127] have published accounts of structuring practices for automation with an emphasis on process engineering and process computing automation technology, respectively. Töpfer and Besch have also described principles [134].

The time now seems ripe for attempting to present an integrated, information-oriented exposition of this subject.

The information structures employed in control engineering and the architectural principles

that generate them are therefore discussed and explained with practical examples (Chap. 2).

Next, the computational and mathematical methods used to acquire, reduce, model, and document process information are presented (Chap. 3).

Information about the process is used for process control (Chap. 4) by applying isolated and integrated control concepts and object-oriented process-control strategies (e.g., recipe management).

The control components of production plants, that is, the electrotechnical devices used for process control (sensor and actuator systems, process control systems including communication technology, process power supply, design of control and power distribution rooms and cable runs), are next analyzed from the information-technology point of view (Chaps. 5, 6, 7).

Computer-aided techniques (C-methods) have proved an especially useful information-technological support for the engineering of process control systems. General aspects of such systems and detailed structures of data-modeloriented design aids now in use or under development are described in Chapter 9.

The procedures used for the design, construction, and start up of process control facilities are set forth from a systems viewpoint and on a structural and sequential basis. Decision phases with the most important design document, (requirements and specifications) are clearly distinguished from execution phases. Quality assurance for the design, construction, and start up of process control systems is described in Chapter 10.

In the operation of production plants with process control systems, a change is taking place from signal-oriented man – machine communication to information-oriented, state-based man – process communication. This enables the plant operator to utilize knowledge about the process, formulated in an ergonomically correct way, to control the process. Continuous process analysis and process optimization take on new importance in this context. Strategies for maintaining or restoring the availability of process control functions are described in Chapter 11.

The standardization of process control techniques and systems in national and international bodies, societies, and institutions is presented in Chapter 12.

Finally, the integration of knowledge-based systems and modern concepts of fuzzy logic and neural networks are discussed in Chapter 13.

2. Information Structures in Process Control Engineering

2.1. Principles

Information as a Corporate Resource. In principle, information is an unlimited resource that can be managed only if it is structured. It is the task of information management to optimize the utilization of information by insuring the efficient interplay of all parties involved and providing the technical infrastructure.

The objective of information structuring is to reduce the degree of complexity by appropriate modeling (formation of data models and functional models) and thus create the prerequisites for information processing (input/output, storage, and analysis) by means of information systems. What has to be overcome, above all, is not the material but rather the inadequacy of the human intellect to deal with complexity.

The importance of adopting suitable structuring aids is demonstrated by the way in which technical, administrative, management, and logistical data "circles"–none, as a rule, consistent with the others–frequently come together.

Because they continue to be developed, application systems arise whose functionalities often overlap (Fig. 2, see next page) [135].

What is more, the demand for information depends very heavily on the level of automation present in individual production facilities. Production engineering systems often feature non-communicating automation islands. In fact, all such information circles are necessary, from the technical and economic standpoint, in order for the corporation to accomplish its task unhindered by departmentalization.

If information as a corporate resource is to be optimally deployed, it is essential to integrate information technology, and this can be done only when information concepts have been harmonized.

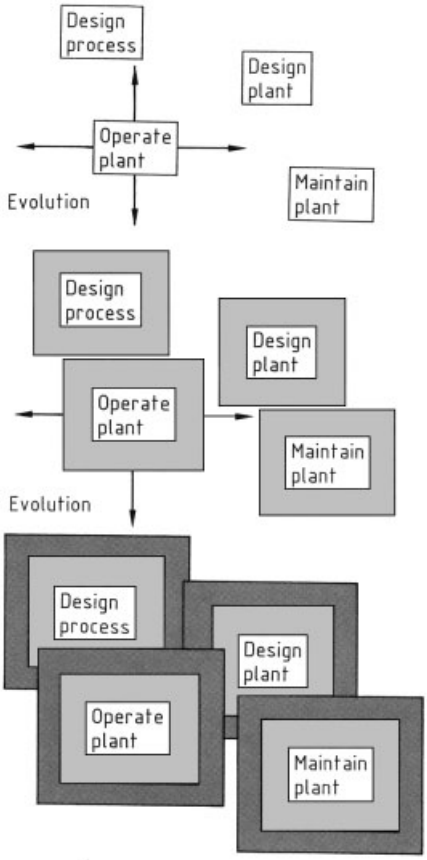

Figure 2. Convergence of isolated data "circles" and automation islands [135]

Integration of Information/Information Models and of Data/Data Models. Information integration is achieved by consolidating information across projects, leading to a mandatory corporate information model [136–139].

Such information models [140] are said to be "semantic" or "conceptual" because, at the user level, they reflect the definitions and interrelationships of concepts but not the realization of information in data-processing (DP) terms or the selection of information for specific applications. In contrast to earlier management information system (MIS) implementations, the result is an extremely stable information model. Appropriate representation techniques yield data from information and data models from information models.

Since 1970, theoretical publications on the form of such data models have been concerned chiefly with how the organizational problem of large data-base systems can be solved. For example, the mathematically grounded relational data model with its normal forms [141] was created first; the entity – relationship (ER) model, which stresses the semantic aspect of the data, came later [142].

In 1975, the ANSI/X3/SPARC DBMS Study Group introduced the "conceptual scheme", a comprehensive, implementation-neutral structure description of corporate information. This conceptual scheme is of central importance for the understanding and management of information systems [136, 143]; see also the SPARTEN information and control systems (ISS-$\alpha\alpha$) developed at Bayer AG [144].

Further detailed exploration and discussion of the conceptual scheme has taken place in the ISO Working Group with the same title. A detailed report is presented in [145].

Work on these aspects has continued; some results have been summarized and various methods, procedures, and tools to aid in the creation of data models have been set forth [146–149].

In the development of information systems, data properties relating to content, application logic, and organizational aspects are first examined in a "data structure analysis." The results must be described in a semantic or conceptual information model before the data model (which depends on the form of presentation) and thence the data-base design can be derived. The data-base design must take account of requirements and restrictions having to do with data-processing hardware and software. Restrictions based on processing considerations are brought in at another level.

The data-base design can thus be represented in a three-level model:

- Conceptual level: The result is the semantic or conceptual information model
- Logical level: The result is the logical data model, which contains restrictions having to do with data-base hardware and software
- Physical level: The result is the physical data model, which makes allowance for processing restrictions

A conceptual information model is constructed with an eye not to isolated information-processing functions, but to the objects of corporate information processing that are common

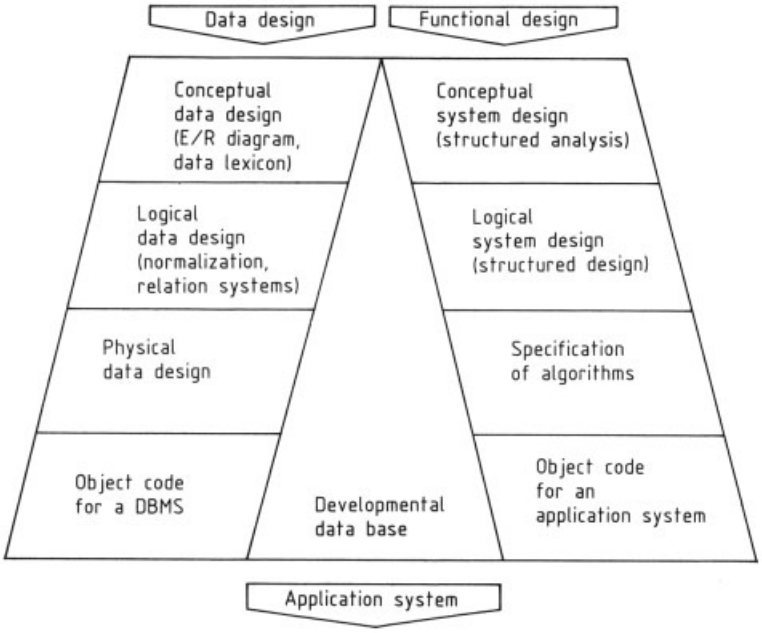

Figure 3. Methodological concept [151]

to all such functions. This is the strategic dimension of data modeling.

Such conceptual corporate data models are now being created in many companies. As a corporation-wide, application-neutral, implementation-neutral structural description of corporate information, such a scheme also constitutes the common linguistic basis for communications between persons involved in DP management. Because it is independent of the allocation of data among the company's computers, independent of physical data storage, and independent of the data-base systems employed, such a scheme can unify information systems based on a wide variety of hardware and software.

Integration of Methods. In the first period of software engineering, the software life-cycle model was in the forefront of discussion. Software development environments were designed in accordance with this philosophy. Alternative philosophies include rapid prototyping and participative system development [150]. Both aim at the early involvement of end users, whose acceptance is the single decisive criterion for the success or failure of an application system.

The appearance of data bases gave new impetus to the development of data modeling techniques. Because these are based on a variety of implementations, they are largely incompatible. For example, an elevator control system does not require a data-base design; an information system has no time-critical concurrent processes; and an expert system uses different knowledge representations from a procedural program.

If the data model has a hierarchical, network, or relational structure, different design methods are used.

An integrated methodology must generate both the data design and the functional design (Fig. 3).

It has often been overlooked that every design level of this methodological concept deals with a different aspect of the problem.

Thus, data and functional models are built up at the uppermost, conceptual level and then, in the further specification, broken down into a representation that can run on certain hardware.

Integration of Representation Means. The basic prerequisite for the common storage of design data and the transfer of information between methods is an appropriate representation form.

Today, graphical and textual representation means can be thought of as being diametrically opposed.

Each, however, has its justification. Graphical means of representation (flowchart, function diagram, entity-relationship diagram, polling hierarchies, etc.) are a suitable idiom for use in the human – computer interface (Table 1).

Textual means of representation (pseudocode, structured text, computer languages of various generations) yield expressions that are understandable chiefly to the computer.

Table 1. Methods and their points of emphasis [152]

Method	Points of emphasis
Structured Analysis (SA)	Representation of functions, data regions, interfaces, and data flows connecting them, in the form of hierarchically organizable nets
Structured Design (SD)	Representation of modules and their interfaces in "Structured Charts" and module level model
SADT	Representation of functions, as well as data and control flows connecting them, in the form of hierarchically organizable nets
Entity-Relationship Model	Representation of objects and their quantifiable relations (1: 1, 1: n, etc.) in the form of nets
Decision Tables	Representation showing under what combination of conditions certain actions can be performed, in the form of tables
Finite-State Automata	Representation of states and state transitions, and also of inputs initiating changes of state as well as outputs initiated upon changes of state, in the form of nets
Petri Nets	Representation of states and events in the form of nets whose behavior can be dynamically investigated with the aid of the Petri net rules
ISAC	Representation of information sets, as well as information flows and functional sequences, in the form of seven types of nets
Jackson Structured Programming (JSP)	Representation of the data structure of a problem, derivation of the program structure from this data structure
Jackson System Development (JSD)	Representation of objects and their actions in the form of trees, and of the connections between objects in the form of nets
MASCOT	Representation of processes (and devices) and their communication regions in nets
Functional nets	Representation of processes and devices as well as their communication, in the form of nets
HIPO	Tree-type breakdown of a system and definition of every part in terms of input and output data and required operations (steps)

There must therefore be a way to translate between graphical and textual means of representation.

The task remains to harmonize these means of representation with one another, obtaining a canonical set of description means that can be used to describe data structures and functional structures with regard to both static and dynamic properties.

Integration of Tools. The most important tools for data integration are data-base systems and data dictionaries [138, 153].

In order that the data in a data base will represent a consistent system of facts, the permissible and necessary relationships between data are specified by means of integrity conditions and monitored and controlled in real time by a data-base management system. Data-base systems offer a range of concepts that support "information management."

In particular, these include concepts for modeling and manipulating data in an application-oriented fashion (data models), for integrating various applications over the same data base without redundancy (data integration), for making an application neutral to changes in the logical and physical organization of the data (data independence), for insuring data consistency, multiuser synchronization, and data security, and for safeguarding the stored data against unauthorized access.

A data-base system is said to be of hierarchical, network, or relational type, depending on the underlying data model.

Hierarchical and network data-base systems have become established in commercial, administrative applications. Relational data-base systems are gaining ground in engineering applications, such as the support of design processes (in the design of mechanical parts, LSI circuits, process monitoring and control equipment, and software development) and process control. Distributed and object-oriented data-base systems have not yet become widely available from commercial sources.

In contrast to data-base systems, a data dictionary is a system for storing and processing data definitions (data descriptions) for one or more software systems. In its simplest form, a data dictionary can be regarded as a well-organized notebook containing basic information about all the data elements stored in a computer for an application area [154–156].

These support not only a one-time, uniform, consistent documentation of information objects, but also the management of other objects used in the DP environment (such as jobs, projects, users, etc.) as well as their relationships. Data dictionary systems are themselves data-base systems that use a central data base to manage all information needed to describe, pro-

cess, and use the objects of an application area, regardless of their internal organizational structure. Data dictionary systems also support data management for application systems and thus contribute to the efficient development, implementation, and description of software systems.

A number of other tools have been devised in the course of the software engineering discussion. Each of these tools is suited to a certain class of problems; each either extends over multiple phases or is specific to one phase [140, 157, 158].

For example, editors are for the creation of graphics and texts for use in a variety of methods; compilers, linkers, debuggers, and so forth, for the implementation phase; test tools, for the testing phase.

Tools for retrospective documentation are becoming more and more important at present.

Decades of software development have produced software systems that can scarcely be maintained any more (the "software crisis"). The documentation of these systems is often in a wretched state.

Because of changes in methodology, it is no longer sufficient to document such systems in the context of their own (usually obsolete) approach. What should be aimed for is to create new documentation for old systems, with an eye to future trends in software engineering.

Retrospective documentation is a more and more pressing need in a management context, where decades' worth of investment in software must be safeguarded. This does not necessarily mean that obsolete software systems must continue to be used in their existing form. The minimum requirement is that the information and technical knowledge contained in the system must be revealed and made accessible for new developments.

The integration of tools means that documents can be passed along from phase to phase and from tool to tool.

Integration of tools can succeed only if information, data, and methods have been integrated. This applies not just to all engineering activities during the lifetime of a technical device (see Chap. 10), but also to the design of monitoring and control equipment itself, especially of process control engineering systems (see Chap. 7).

2.2. Architectural Principles for Information Structuring

Systems and Architectural Principles. Real-world systems and elements have internal structures. Along with Aristotle's doctrine of categories [159], the Linnaean system–still a valid descriptive model employed in botany–provides an example [160]. Darwin's ideas on the origin of species [161] also illustrate this structuring process.

In the 1700s, "system" at first meant a collection of linked truths. The links had to be methodologically correct and completely deducible from an underlying principle.

KANT's concept of system stresses the theory of systematics. In his "Critique of Pure Reason," he emphasizes the significance of architecture as the art of systems.

Thus a system, traditionally, is a closed whole whose parts are interrelated, interact, and satisfy certain constraints.

The parts, regarded as elementary, are called elements of the system. A system consists of elements with properties; the elements are structured and linked together by the relationships in accordance with a system principle. Systematization involves specifying system principles and devising system structures. Systems are characterized by the fact that they have properties that cannot always be accounted for by the properties of their elements. Thus simply collecting elements does not form a system. Instead, each system has a structure that is generated by the relationships among its elements.

Systems can be subdivided into parts on as many levels as desired, whereby all parts that are not elements of the system are referred to as subsystems. System elements are the parts of a system whose further subdivision is impossible or undesired.

The sections that follow state principles for the structuring of information. They are called architectural principles because they can be identified at various points in the design process and in various representations of the information system, thus marking the architecture of the information system. The first principles discussed are general ones in the phases of design and implementation.

Principles in the "definition of requirements" phase, quality assurance in parallel with other

activities, and operation of a facility might be added here. These are also referred to as con-struction principles, since they are applied mainly in the construction of DP systems.

This is demonstrated by the example of modern operator interfaces, where structuring of information is used to support human – process communication (see Section 11.2).

Higher concepts of functional and data structuring also make use of the design principles discussed above but represent a more application-oriented aspect of structuring.

The focus is on the object-oriented formulation, which can be applied to both functions and substances [162] in an integrated methodological concept. It is the tool for the decomposition principle, with which complicated objects are broken down into their constituent parts.

The next most important structuring principle is abstraction, which is a fundamental approach to concept formation. Adopted from expert-systems technology, abstraction is introduced as a construction principle of application systems through class-forming, concept-forming, complex-forming, and functional abstraction. Closely related to the principle of abstraction is that of inheritance [163–165].

The last section deals with transformation principles. Here the various models used for problem-solving are described, starting with procedurally oriented formulations (with their counterpart in functional design) and passing through logic-oriented formulations (as now employed by many expert systems) to state-oriented or event-oriented concepts (which come much closer to process-related DP problems).

State-oriented and event-oriented formulations in turn are closely related to object-oriented ones. Objects are in certain states and communicate with one another via established protocols (see Chap. 8).

The complexity of present-day systems increasingly calls for a holistic, network style of thinking [166, 167] if the basic laws governing systems, their dependences, and their interac-

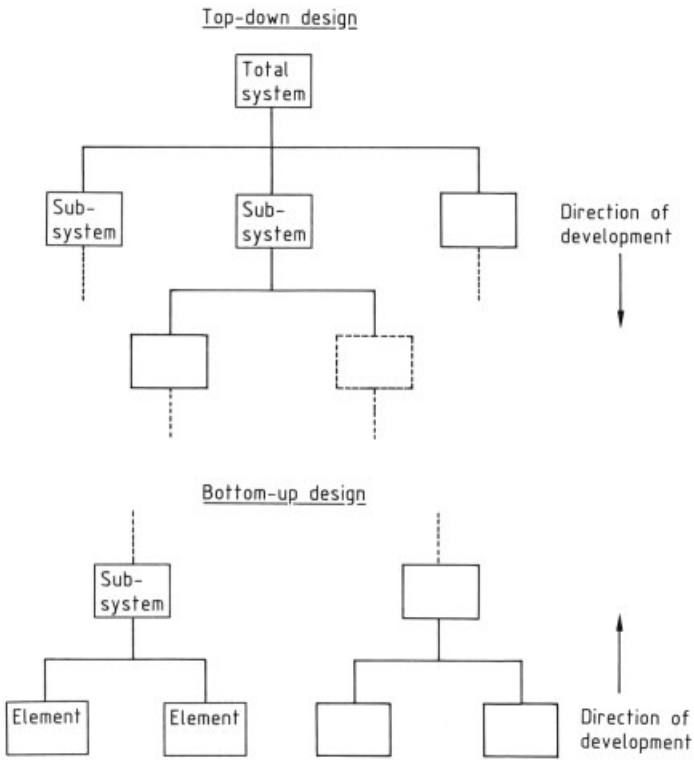

Figure 4. Design principles

tions are to be understood. It will take a new kind of thinking to comprehend the dynamics of a continuous process, plan a monitoring and control system for it, erect and operate the system, employ higher-level information systems, and in the process not overlook human beings [168].

The Decomposition Principle.

Design Principles. Structuring principles occur in all phases of system development. In the design phase, for example, the top-down and bottom-up principles are often encountered. In the top-down approach, a whole is broken down into parts step by step, whereby the complexity of the parts under consideration is systematically decreased (decomposition principle).

In bottom-up design, single parts are assembled step by step and made into a whole.

In practice, mixed forms are often encountered (Fig. 4, see previous page).

Bottom-up design comes into play when prefabricated elements are available from which the overall systems can be assembled. This principle is often used in mathematical and statistical applications, where the elementary mathematical and statistical functions are known and can be made available in program libraries. New applications can be analyzed only by the top-down approach, though there is a trend toward reusable elements here too (see, e.g., Section 4.5).

Implementation Principles. The implementation phase also exhibits principles that ultimately determine the structure and thus the understandability and maintainability of software systems.

The more ordered the structures are, the better they lend themselves to verification, modification, and maintenance. The simplest ordering is linear. Iterations, recursions, and branches generate more complex structures (Fig. 5).

Recursive data and functional structures are preferred today because of their mathematical compactness. It can be shown, however, that iterative and recursive structures are equally powerful and can be transformed to one another.

Another type of structure is the directed tree with strict hierarchy or polyhierarchy, which ultimately becomes the general network if the rigorous requirement that every node has only a single input is dropped (Fig. 6, see next page). The last is especially important in so-called master–slave structures. The slave should have only one master, since otherwise there can be conflicts in execution. Ring and star structures form a separate class.

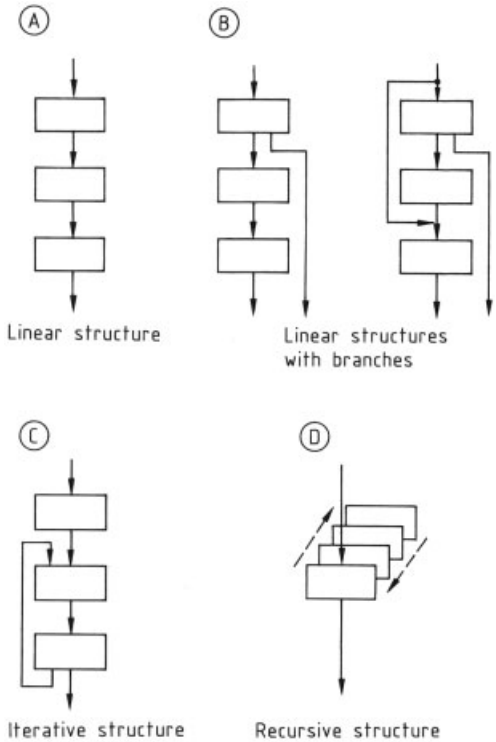

Figure 5. Types of structure

Principles for User Interface Design. The user gains access to the system through the user interface. A number of structuring principles apply to this interface, not all of which can be discussed here.

Design principles such as the law of like form, the law of color, or the law of proximity (Fig. 7) should be employed in the layout of masks and printouts [169, 170].

Principles such as conformity to expectation in dialog with the system (also called "principle of minimum surprise"), consistency of dialog elements, and visualization of the application context also play an important part.

Modern software ergonomics focuses on job design for people, a process that is heavily influenced by software [171, 172].

An important goal in software writing must be the design of complete activities. A complete activity has been defined as an activity that, to-

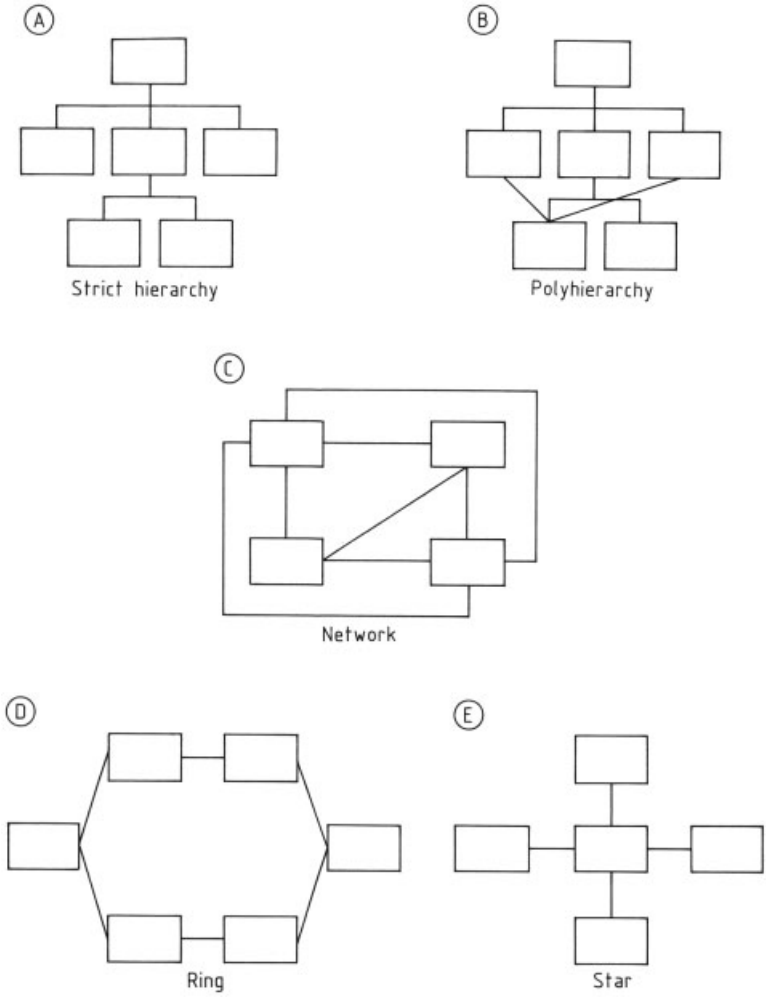

Figure 6. Types of tree and network structure

gether with simple execution functions, includes the following:

- Preparatory functions: identification of goals, development of procedures
- Organizational functions: agreement with other workers as to tasks
- Monitoring functions: feedback to the worker on attainment of goals

Incomplete activities are activities where there is essentially no possibility of independent goal identification and decisionmaking, no scope for individual ways of performing tasks, or no adequate feedback. Shortcomings in job design impair the motivational and learning potentials of the work process and affect the worker's welfare and job satisfaction. Intimately linked with the completeness of activities is the degree to which the worker has opportunities to make decisions as to time and substance, that is, freedom of action.

This freedom of action must permit creative users of an information system to devise an individualized procedure in accordance with their experience and working style. At the very least, defining freedom of action should not be part of the program author's responsibility; it is a function of corporate management and personnel practices. Observed acceptance problems often stem from misjudgments in this area.

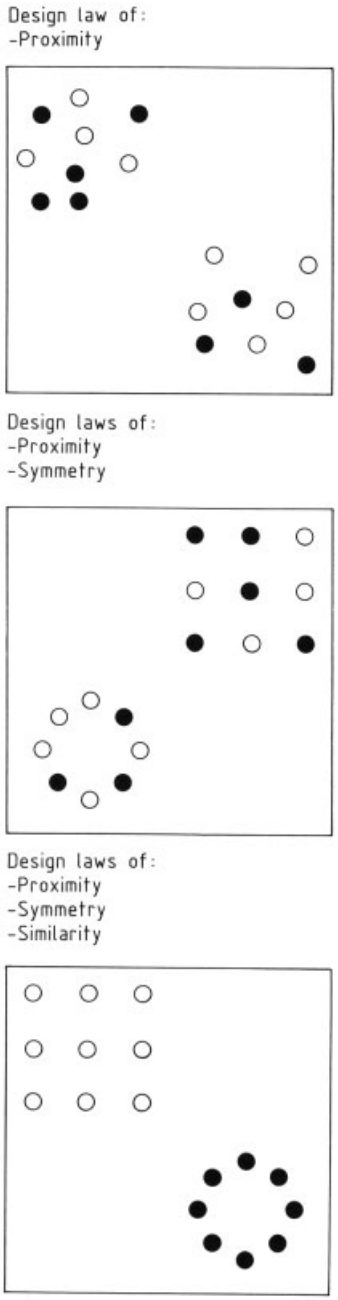

Figure 7. Simple design laws [169]

A new type of human – computer interaction has come into being with the new class of workstations. This user interface has the new feature that it presents the user with graphical images of customary objects belonging to the office world. Such objects can be selected and then manipulated with the aid of a pointing device (mouse, light pen).

SHNEIDERMAN [173] introduces "direct manipulation" as a collective term for several new principles in the user interface. The most important principles are the following [174, 175]:

- Permanent visibility of the object as an icon or pictogram
- Quick, reversible, one-step user actions with immediate feedback
- Replacement of complicated commands with physical actions (see also Section 11.2)

Higher Concepts of Functional and Data Structuring. Historically, the functional structuring of application systems was emphasized in the early period of data processing. A complex function was successively broken down into partial functions until elementary functions realizable in DP hardware and software were arrived at (flowcharts).

Data were fitted into the functional structure but remained meager, as they had to be given in the programming languages known at the time.

The structuring principle was oriented to the procedure employed in problem-solving. As a result, functional structuring led to process-oriented or program-oriented structures (Fig. 8, see next page).

Even today, functional structuring is still the classical approach, because it leads directly to realizable, procedurally-oriented solutions. At Bayer, for example, all engineering tasks were modeled by the SADT method, somewhat as shown in Figure 304. [177, 178].

Thus the modeling concept (functional structuring) and the implementation concept (procedural programming language) correspond.

Data structuring is of secondary importance: What is interesting is the processing of the data, and data-base systems were not yet in use at this time.

However, functional structuring has several weaknesses:

- There are scarcely any objective criteria for breaking the function down into partial functions. Decomposition is often a matter of the developer's discretion.
- The resulting data structures unavoidably grow out of the functional structure and are monolithically bound to it. Any change in the

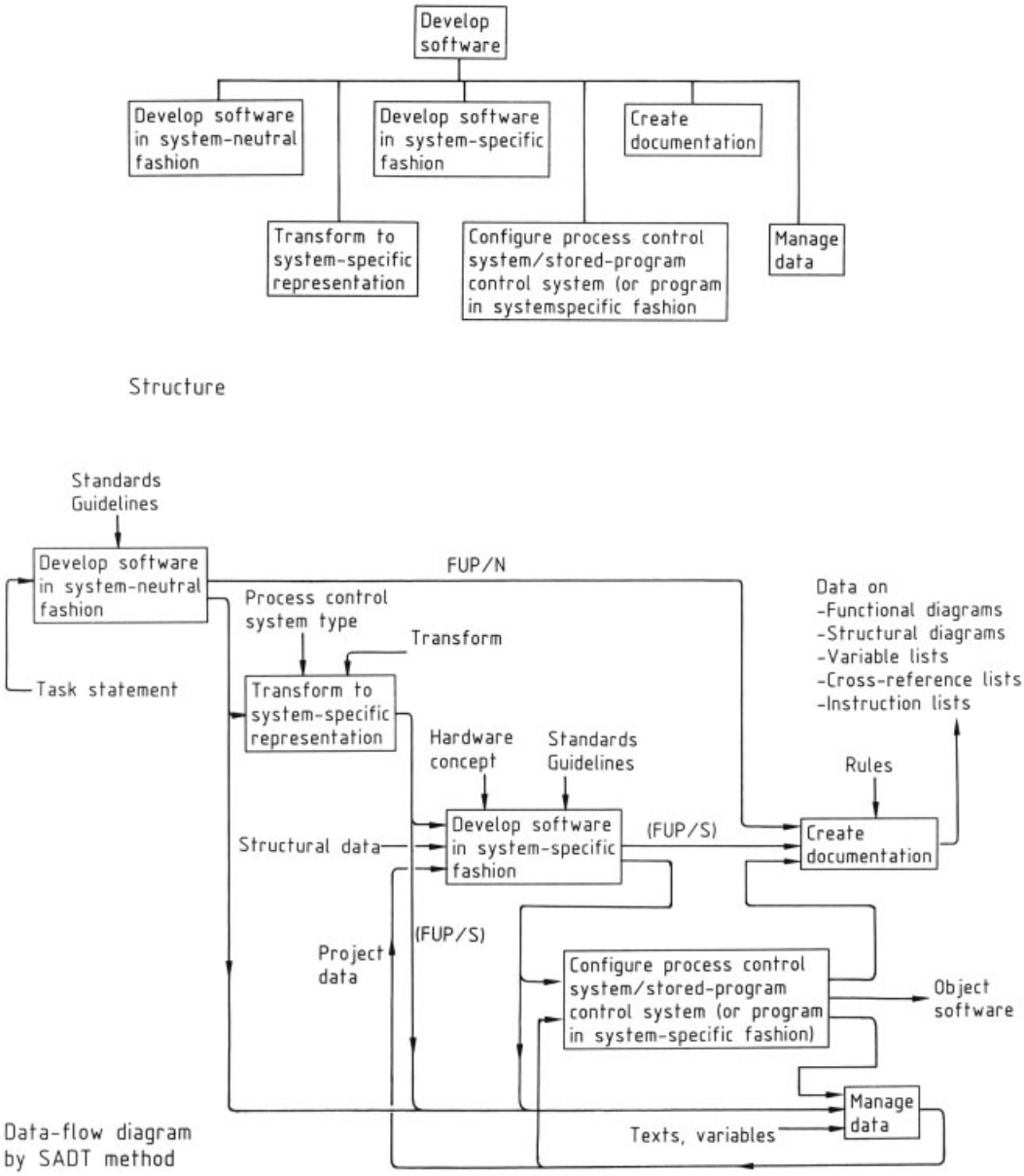

Figure 8. Functional structuring for the example "develop software" [176]

data structure impacts on the functional structure, and vice versa.

– Frequently, insufficient stress is laid on reusability, since it cannot be known how and in what connection an arbitrarily defined function or an arbitrarily defined module may be used again–in contrast to mathematical and statistical program libraries, where reuse is straightforward. The complexity of such a DP system increases more than proportionally to the functionality.

In the early 1970s, greater value was accorded to data structuring. This was the period of developments such as the PASCAL programming language and the first hierarchical data-base systems. Data structuring became more important.

Subsequent development is marked by attempts to shift the meaning of the data (the infor-

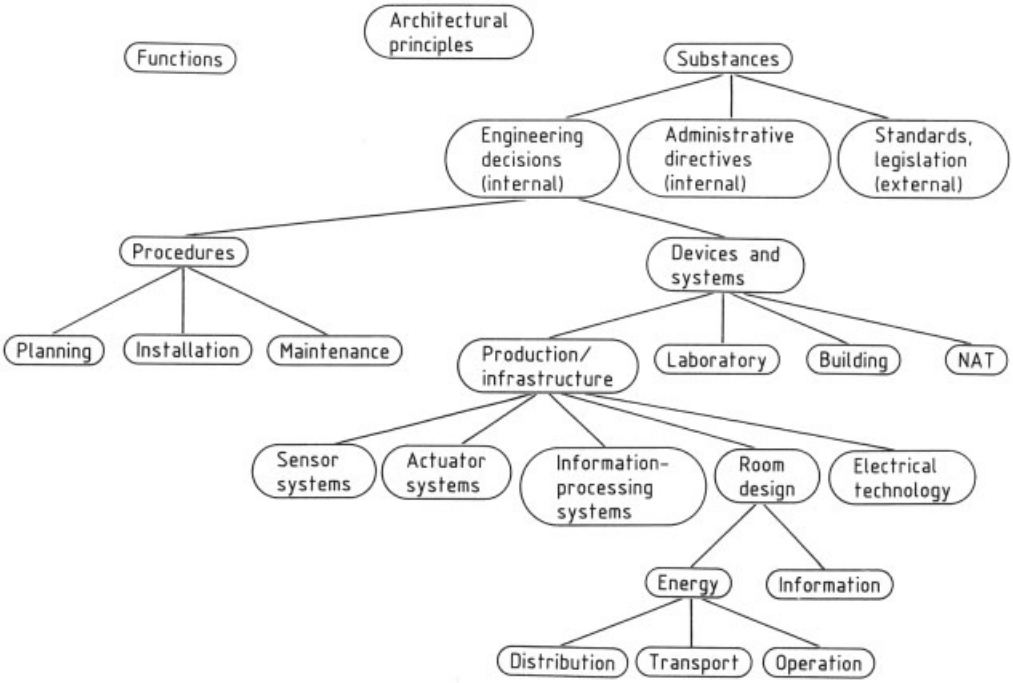

Figure 9. Structure of PCE propositions

mation) out of the programs and into the stored data, and to generate program systems in a manner directly opposed to the functional approach. Now the first step in system development is to design the data model on which the later program systems will work.

One speaks of data-oriented or object-oriented structuring, since the analysis is focused on information objects (data structures) as representations of real or abstract objects.

Objects are described and related to one another through their properties, also called attributes. Typical relationships are abstraction and mapping (correspondence) relationships [137].

The result of this approach to structuring is the semantic data model, which is designed independently of the functions defined on it [146, 149].

In his epistemological discussions, Wendt [147] classifies objects as concrete and abstract. Concrete objects include things and processes; abstract objects include concepts (see [162] and Fig. 9).

After the existence of objects and the properties of objects, the relationships of objects are taken up third. A first type of object relationship is described by ordering, which is arrived at through a comparison of some property.

Other important object relationships are part-whole relationships and instrumentality relationships. An instrumentality relationship is said to exist between a thing (e.g., a process facility, a process monitoring and control system) and a process (e.g., a continuous process, a control process) if the thing plays a part in the process.

Inseparably connected to processes are states. States include the initial and final times of processes or subprocesses.

If countability is associated with the concept of object, and measurability with the concept of property, then the concept associated with the concept of relationship is decidability [147]. Objects are counted, properties measured or described (see "Scales" in Chapter 3), and relationships are decided.

Thus the three fundamental concepts – object, properties, and relationship – have been introduced. These make it possible to structure experience, and thus they govern the world of human thought [147, 179, 180].

Concrete objects are the plant, the apparatus, the process monitoring and control system, the production process, the reaction process, the sensor – actuator process, the data-processing process, and so forth. This group of objects is also called "devices"; devices must be designed, built operated, maintained, and optimized, these activities being known as "procedures." The outcomes of these procedures are represented as flowcharts, construction drawings, process monitoring and control station diagrams, and so forth. In the Cassirer scheme, both groups of objects are called substances. Functions that describe a task or activity (e.g., plant/apparatus coding; see Section 2.3) also have a representation in the object-oriented formulation (see Figs. 23, 25, and 65).

Functions (i.e., tasks) are implemented through procedures (abstract objects) from devices (concrete objects). Figure 9 gives a large-scale survey of "object worlds." For clarity in terminology, relationships are described not with the function concept usual in informatics, but with the epistemologically based relation concept (cf. entity relationship) [162].

The abstract representations are the objects that can be manipulated in the computer. They always relate to an object, which they describe from a certain abstract point of view.

The object-oriented approach thus represents a "natural" modularization concept, since the objects can be extracted directly from the application. If the functions are likewise structured in an object-oriented manner, this constitutes a principle of functional structuring that is far less sensitive to modifications.

The Principle of Abstraction. The (provisionally) last step in development combines relational and object-oriented structuring. In the "abstract" data type, data structures and relational structures are merged into a whole. The relations enclose the data as if in a shell; the data can be manipulated only through the relations "guarding" them. The high modularity of these constructs suggests that there may be significant gains in the reusability and maintenance of software. This principle is referred to as data encapsulation or "information hiding" [165, 181–183].

The structuring principle is thus concerned with the structuring of data and relations, their (static) structural and (dynamic) sequential structures, and their interactions.

Usually, the sensory or empirical view of things is said to be "concrete," while any description in terms of concepts is said to be "abstract" (see also [147]).

Abstraction means reducing complexity by neglecting all inessential properties or features. One also speaks of models.

Models are always simplified pictures of the real objects (abstractions) as viewed from a certain aspect (see also Sections 3.1 and 3.3).

In process engineering, flowsheets are models of a real facility in graphical form. They are standardized and represent the medium of communication for persons involved in the design, construction, and operation of plants (Fig. 10).

Mathematical models in the form of systems of algebraic and differential equations depict partial aspects of the physicochemical and apparatus-specific parts of processes. Such models are used, for example, in model-aided measurement techniques for the control of processes, the detection of faults, and simulation (see Chaps. 3 and 4).

A plant model is a representational picture of the plant that is much favored as an object of study by plant designers and operators.

The three-dimensional representation of plants in CAD systems is a way of getting the advantages of such models while using only DP resources.

Along with mathematics, the field of informatics now offers methods with which complex engineering systems can be modeled, simulated, and optimized.

It must not be overlooked that data models are always reductions from real systems. Models, no matter what kind or in what medium, represent a given system from various points of view. They emphasize essential properties and suppress inessential ones. Within these limitations, however, models are an indispensable part of any modern information technology.

Class-Forming Abstraction. Object orientation means that once the objects have been identified, they are then described in terms of their properties or attributes. Each such property consists of a category and a value; these can be scaled either metrically or nonmetrically (see Section 3.1 and Fig. 40). The number of properties describing an object is not intrinsically re-

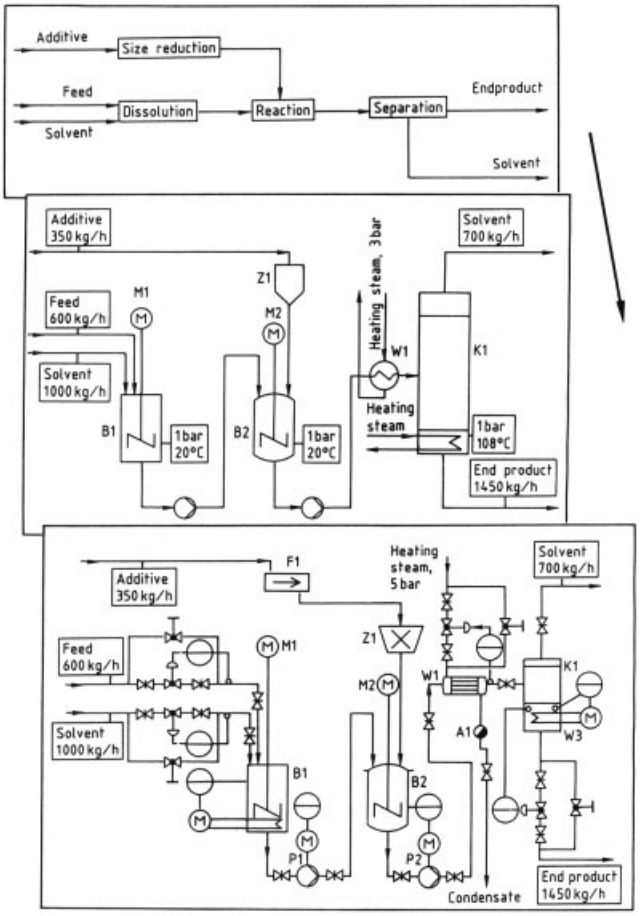

Figure 10. Design steps in process engineering [184]

stricted. The treatment is generally limited to the attributes relevant to a mode of examination or an application, or to characteristic attributes; this is equivalent to an abstraction (Fig. 11). Such attributes can be listed in a table (Fig. 12, see next page).

One column in the table describes one characteristic property or attribute of the object. The table represents a "class of objects"; that is, a collection of objects having at least one attribute category in common.

The attributes describing a class simultaneously identify the type of an object. If the same attribute categories are employed for the description of more than one object, these categories are said to be "generic."

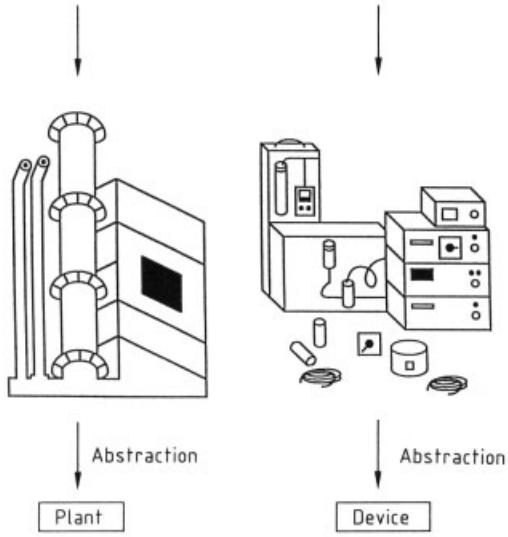

Figure 11. From things in the real world to abstract objects

Table

Name of plant	Identification	Plant complex	Battery or works	Capacity	Utilization

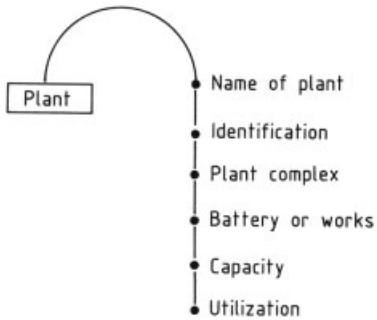

E/R diagram

Figure 12. The table as a form for describing object properties

Name of plant	Building	Plant complex	Battery or works	Capacity	Utilization
Sulfuric acid unit A	4711	Sulfuric acid	Brazil	100000 t/a	85%

Instances of object class "plant"

Figure 13. Instances of a class of objects

A given object in the real world is called an instance of the class (Fig. 13). The concept of table, also called relation, leads to the relational data base, which utilizes the relation (= correspondence, dependence, etc.) as its organizing principle. The table as a data structure is purposely not given any further structure, so that it is possible to create a "relational algebra" that is mathematically well-defined. Any manipulation of the data base is thus broken down into set-theory operations.

Apart from class-forming abstraction, the identification of objects is important. Each instance of a class must be identified by a unique name and/or a unique identification number.

Without exception, every one of the classification schemes devised earlier for identification purposes has failed, because objects cannot be unambiguously classified (see Concept-Forming Abstraction) and an excessively narrow decimal classification system often proved too restrictive. Classification systems were weak forms of data modeling at a time when data bases and semantic data models were not yet known.

What is recommended today is an unsuggestive, system-free, numeric or alphanumeric code and the description of properties–including the membership of the objects in one or more classes (polyhierarchy)–in a semantic data model.

Another form of identification has been discussed in connection with direct manipulation. An object is identified by "pointing" at it. This principle of identification is extremely important for the ergonomics of user interfaces (cf. the icons first introduced by Apple in the Macintosh). A third option is to describe objects so that they can be found again. The relational data base draws its strength from this approach. Structured query language (SQL) represents an industry standard and makes the results independent of the data-base system.

Concept-Forming Abstraction. Concept formation is an act of abstraction that leads to concepts and hierarchies of concepts. It is a generalizing process, in which particulars are disregarded. Unimportant properties of the object are left out; those remaining govern the class of the objects. Each abstraction rests on a certain objective of thought or speech, for example the objective of being able to think or speak about [179]:

- Things, regardless of engineering details
- Procedures, regardless of what principles or concepts apply to them
- – Factual statements, regardless of meaning-preserving transformations in the form of circumstances

According to Frege [185], things that are already linguistically available are grouped, from the viewpoint of something they have in common, into a "concept," which offers an invariant way of speaking or thinking about things already linguistically available.

The formation of "abstract" concepts cannot take place in isolation from the intended purpose or objective. This may be a disadvantage if one wishes to do modeling in the most application-independent way possible at an early stage of system design. For example, a biologist might group dogs and cats together as domestic animals, place these in turn under mammals, and so forth, while a jurist might distinguish movable and immovable objects simply because movable objects can be stolen and such taking is a statutory offense. Darwin, in "The Origin of Species," had already discussed the difficulties involved in devising consistent taxonomies and hierarchies of concepts [161].

The concept arising from concept formation is not specified in advance but lies in the discretion of the data modeler. It is therefore quite possible for multiple taxonomies to exist in one and the same realm of objects. This can be illustrated by a simple example: If one abstracts from the manufacturer of certain bimetallic sensors, the concept or class used is that of manufacturer, while the concept introduced is that of bimetallic sensor (Fig. 14 A; see also [186]). If one wishes to introduce the concept of expansion sensor along with the concept of bimetallic sensor, there is a further abstraction process using the principle of measurement as class (Fig. 14 B) and leading to the more abstract concept of temperature sensor.

Now it might occur to one to set forth the sensor program in closed form by manufacturer; now the taxonomy looks different again (Fig. 14 C; see also Chap. 5).

Complex-Forming Abstraction, Aggregation. In complex-forming abstraction, unlike objects are combined into a new object-concept. The process is also termed aggregation, because the new concept denotes either the static structure of one object or the relationship between objects. The object newly created by aggregation can also have new attributes that are made possible only by the new whole. A distinction can be made between this operation and agglomeration, a simple collection of objects in which the attributes present merely change in value.

Aggregation can also be represented in graphical form. Figure 15 A shows the static structure of a sensor system. The concept "customer order" (Fig. 15 B), on the other hand, is constructed from the concepts of customer, article, and time. The new customer order concept contains part of the information from each of the original concepts [148].

Quite early DP programs were able to handle aggregation in the form of parts lists.

Functional Abstraction. A system can be modeled in greater or lesser detail and in a more or less abstract fashion, depending on the objective. Different persons, with different tasks and areas of responsibility, take different views of the system. Systems analysts and system modelers always describe a system as an excerpt of reality viewed from one or several of these standpoints. According to Polke [168] (see Fig. 29), Rasmussen introduced systematic order into the more-or-less intuitively grounded models [187], describing a system (Fig. 16) as functional ab-

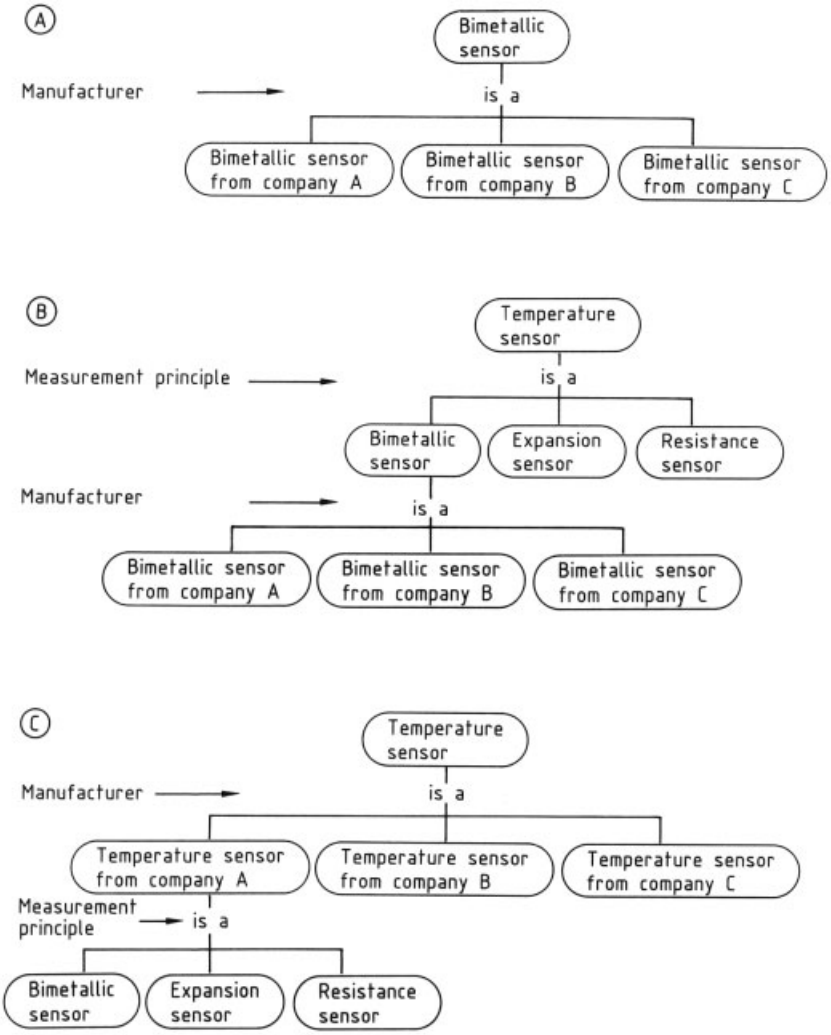

Figure 14. Taxonomies of sensor systems
A) Abstraction according to manufacturer; B) Abstraction according to measurement principle; C) Alternative taxonomy

stractions over multiple levels. The number of levels can differ from one system to another, depending on the type of system and the objective. Nor do all the levels have to be implemented. The design levels of this model extend from intensional models at the uppermost level to extensional ones at the lowermost.

Extensional models relate to the specific physical world, while higher-level models represent the purpose and objectives that are to be achieved with the system.

Scales for the desired behavior must to be developed from the purpose of the system; in other words, criteria for errors and malfunctions can be formulated only against the background of the process objective.

In certain operating states, for example, alarms may be meaningless; measurement signals may have different nominal and limiting values, depending on the operating mode; and so forth. The causes of errors or malfunctions arise from the physical world (bottom-up), while the criteria for defining errors can only be derived from the process goals (top-down).

This approach to modeling brings order to the great variety of representations of a system.

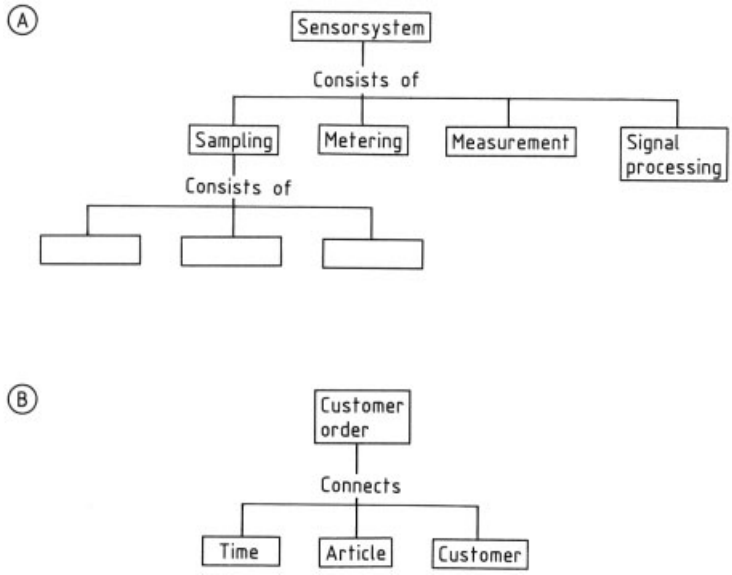

Figure 15. Examples of aggregation
A) Structure of a sensor system [232]; B) Complex object: customer order

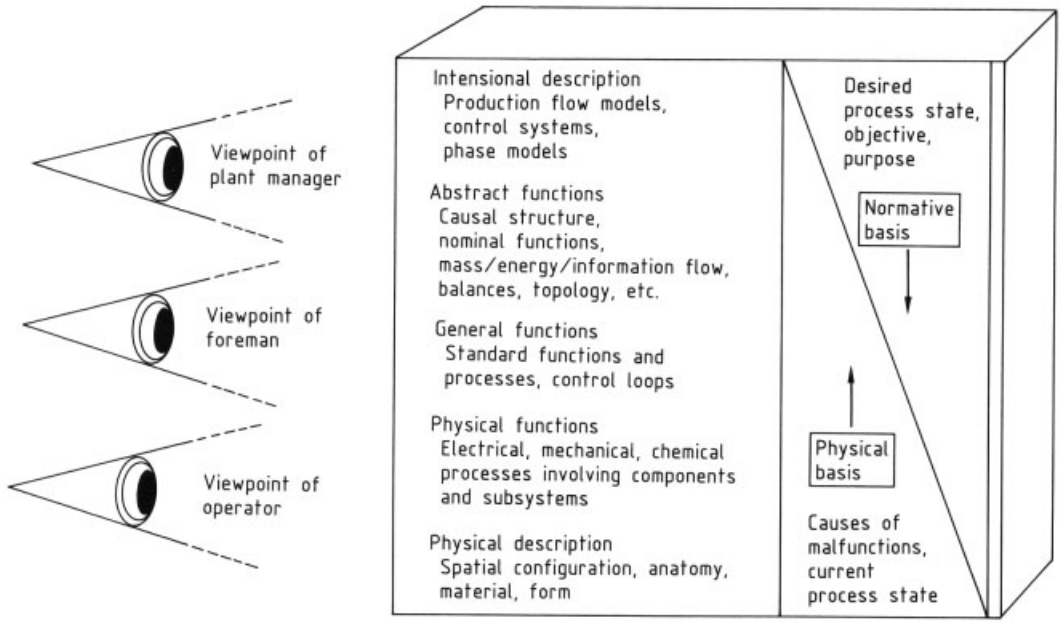

Figure 16. RASMUSSEN's scheme of viewpoints [187]

Above all, it demonstrates that no single representation is adequate; a different view of the system may make a different representation useful.

Many technical and economic systems have been created under this architectural principle. Typical examples are computer architectures; architectures of communications equipment, the latter known as the ISO-OSI reference model (e.g., [188]); and the level model of a corporation discussed in Section 2.3 [168, 189]. The structures of monitoring and control systems men-

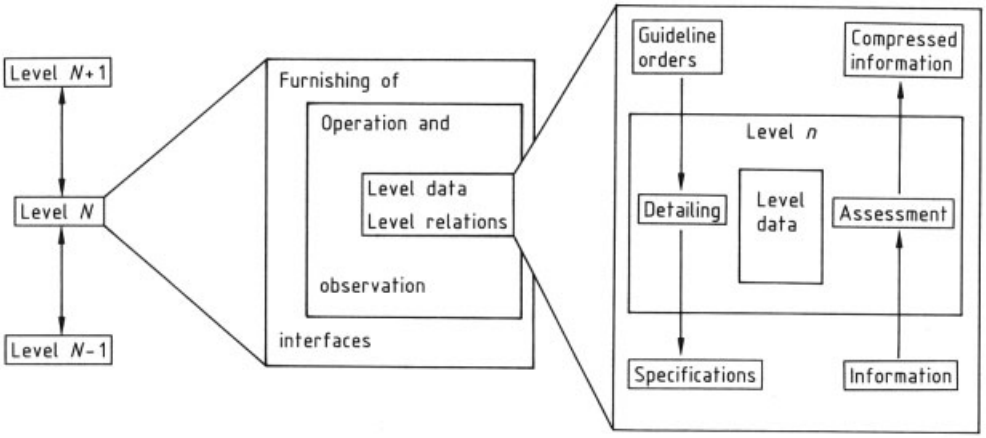

Figure 17. Hierarchy of functions (level model)

tioned in DIN 19 222 also fall under this principle [190].

The principle of functional abstraction is based on the following rules:

- The levels must offer generally usable functions and relations; that is, they must represent a good abstraction.
- The levels must have clear interfaces and minimal upward and downward connection; they must obey the secrecy principle.
- They must completely implement the required functions; that is, they must be self-contained, with the properties of locality and internal strength.
- A level "uses" the services of the level immediately below it. Jumping over levels is not permitted.

It follows that all functions requiring intensive data exchange with one another should be grouped on the same level (Fig. 17).

The design process for engineering systems is also oriented to such a level model. For example, mechanical design begins with the functional structure, proceeds through the operative principle of design synthesis, and finishes with dimensional synthesis.

The Transformation Principle. In chip design, the following levels are recognized: system design, logic design, circuit design, and layout design. The process makes use of detailing algorithms.

Configurable detailing algorithms are not known for the corresponding steps in plant design based on flowsheets (Fig. 10).

If the design process within a process monitoring, control, and automation system is to be completely or partly automated, it must be clear above all how the representation form of one design level is to be transformed into the representation form of the next one. For the developer and designer of an automation system, such a concept supports creative design activities, design verification, project planning, and documentation.

This transformation process in informatics is fully automated by problem-oriented languages, assemblers, and on to machine code. Programs that do these transformations are called compilers, assemblers, and so forth.

Sequence-oriented description concepts and representations are required in order to describe the functions of information-processing operations. One speaks of a transformation principle because input data are transformed to output data and initial states to final states.

The following description types are distinguished on the basis of how the operator is implemented:

- Sequence-oriented descriptions
- Logic-oriented descriptions
- State-oriented descriptions

Sequence-Oriented Descriptions. Classically, information-manipulating processes are described in terms of individual, application-

neutral elements. It can be shown that any routine or any procedure can be described in terms of the linguistic elements sequence, alternative, and iteration. In a procedural programming language, it would therefore be sufficient to make these three constructs available.

Typical graphical descriptive media for procedural concepts include

- Flowcharts
- Nassi – Shneiderman diagrams
- SADT diagrams
- Michael Jackson diagrams

Textual media include, among others,

- Pseudocode
- Structured text
- Programming languages such as FORTRAN and PASCAL

The analogy between operators in mathematical physics and their importance in the description of state changes (including time-dependent changes) can only be mentioned here [191].

Logic-Oriented Descriptions. Logic-oriented concepts relate not so much to the "how" of the process as to the "what." Graphical media used for this purpose include

- Combinational circuits
- Logic diagrams
- Decision tables

Textual media include

- Fourth-generation languages
- Algebraic specifications
- Logic-oriented languages

State-Oriented Descriptions. The automation of technical systems often requires the control of processes that take place in a predominantly step-by-step (i.e., discrete) fashion. The complexity of a discretely controlled system becomes very high when there are processes parallel in the time domain (i.e., concurrent processes). If these processes use shared resources, there is always the risk of "deadlock." The processes block one another in such a way that they cannot continue to operate without outside intervention. Such deadlocks cannot easily be foreseen. The need for an analysis and, possibly, a simulation of these systems is thus manifest (compare the typical situation of management in a multi-project environment [192]).

Such problems can be described well in terms of mechanisms or automata.

A finite-state automaton is a system having a specified number n of possible internal states $z_1, \ldots, z_n$. For each state z_i there is a well-defined set of inputs for which the automaton changes to another state. An output can occur when such a transition takes place. Automata can be used to describe the time order of state transitions in technical systems.

State graphs (Fig. 18), with circles for states of the system and arcs for transitions between states, are used to visualize these systems.

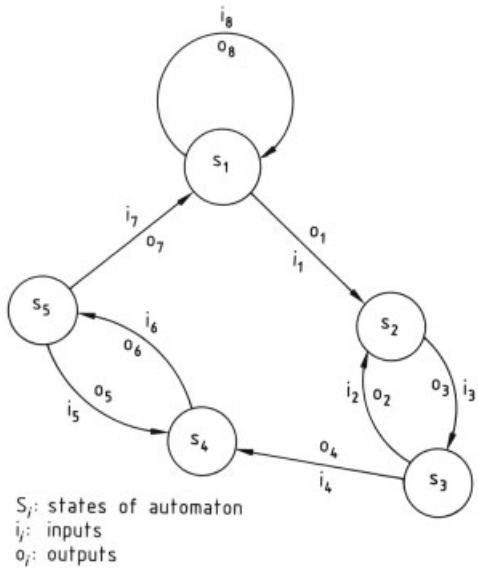

Figure 18. State graph illustrating an automaton

In practical cases, however, the number of global states increases rapidly, so that the representation of the time order by state graphs becomes unwieldy.

The Petri net, an extension of the state graph concept, offers a more elegant way to describe such systems.

A Petri net is a directed graph whose nodes are of two types, places and transitions. Arcs, which can connect two nodes only if they are of different types, represent logical or causal relationships between nodes [172, 193–201].

Using a mode of analysis different from that of state graphs, the Petri net technique allows the description of systems that represent more extensive automata. Thus, the circles (places)

represent partial states in the system and, therefore, conditions for the occurrence of certain events that can occur in the system and lead to state transitions.

The partial state represented by a place is in effect if the place holds one (or more) marks or "tokens."

A transition is enabled, that is, it can occur or "fire," when all its input positions (arcs leading to the transition) have tokens while all its output positions (arcs leading away from the transition) have none (Fig. 19 A).

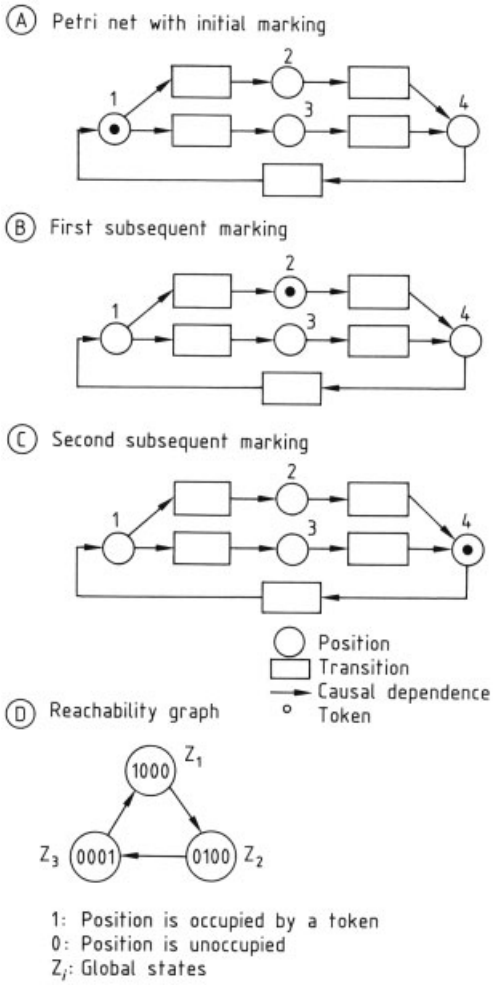

Figure 19. Petri net with initial and subsequent markings and the associated reachability graph

When an enabled transition fires, tokens are removed from its input places and placed on its output places, which now may enable transitions in their turn. In larger nets, this rule means that several transitions may be enabled at once; this is known as concurrent execution (Fig. 19 B and C).

A Petri net lends itself not just to graphical representation but also to mathematical description. On the basis of such a description, the dynamic behavior of the net can be studied by using computer-aided analytical tools.

Interesting properties of the net include, for example, reachability and the presence of deadlocks: If each possible occupation of the Petri net with tokens (each "marking") is interpreted as a possible global state of an automaton, the associated state graph is also called a reachability graph. It immediately shows whether there are markings of the Petri net that can be reached in just one way or cannot be reached, whether there are deadlocks, or whether the initial marking can be reached again (reversibility; see Fig. 19 D).

Petri nets can be broken down in hierarchical fashion and so support top-down and bottom-up design.

In the area of process control, Petri nets are usually represented by functional charts. The technique of functional charts makes it possible to portray directly the structure of a Petri net and offers linguistic elements for the specification of control commands to be issued at the individual steps (nodes of the net).

One graph that is especially easy to comprehend is the "phase model" of production, which is discussed in more detail in Section 2.3.

2.3. Applications in Process Control Engineering

Object Worlds of Continuous Production Processes. If the system to be designed or operated (whether it is an apparatus, a plant, a control system, a software system, or some other system) with all its subsystems and components is placed at the focus of computer support, the monitoring and control of continuous processes can be broken down into plant engineering, process engineering, and process control engineering worlds.

In order to describe the actual operation, these are supplemented by the world of process communications, logistics, and maintenance (Fig. 20) [202].

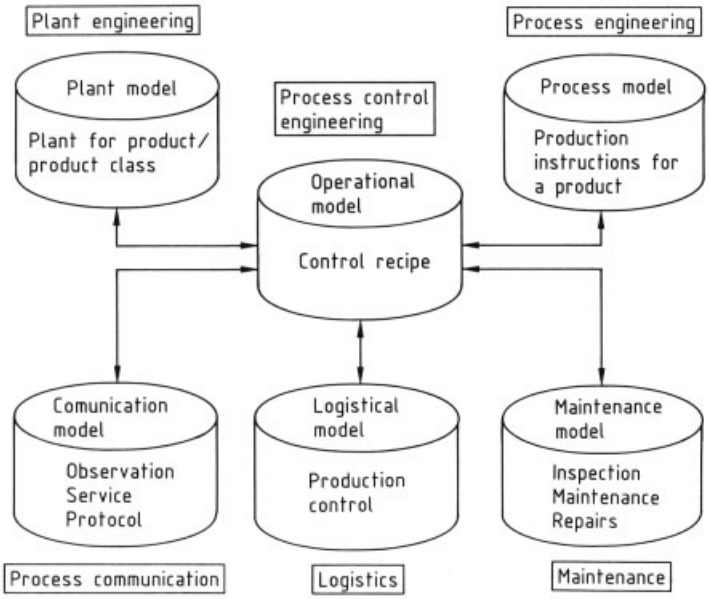

Figure 20. Object worlds of continuous production processes [202]

In an object-oriented approach, substances in these worlds can be defined as functions. They represent the uppermost level of a corporate data model within process control engineering [198].

Plant Engineering. Typical descriptions in the field of plant engineering take the form of data sheets (Fig. 21, see next page).

The abstraction principles discussed in Section 2.2 can be seen in this data sheet: The glass-lined stirred tank, as represented in Figure 22 (page 27), consists of drive, clutch, stirrer, tank, stirrer shaft, shaft seal, motor, and gearbox (complex-forming abstraction, aggregation).

Some of the objects are further specified; for example, the stirrer is identified as an impeller, dual-impeller, or anchor type, and the stirrer shaft seal is specified as a stuffing box or a rotating mechanical seal (concept-forming abstraction).

All objects listed on the data sheet are described by a series of attributes and thus stand for a class of objects (class-forming abstraction).

The "plant and apparatus coding" (PAC) scheme can be viewed as a first formulation and crystallization point for a corporate data model [203]. This coding system describes the static structure of a plant in terms of the required functionality (Fig. 23, page 28). It has analogies with the "power plant coding system" [204].

While most description systems do not separate the functions (tasks) of a plant (abstract) from the real technical plant (device), the PAC gives a rigorous description only of the function, up to the technical device and its functional components, and is therefore invariant to actual technical implementations (Figs. 23 and 65).

Earlier DIN representations [205] could not employ the object orientation logically utilized in the implementation of the PAC [203].

For this reason, stress must be laid on the worldwide harmonization of concepts; this comment applies equally to all three of the "object worlds" under discussion here.

Process Engineering. The following example of process engineering illustrates a taxonomy of unit operations (Fig. 24, page 28) [206].

The classification scheme first looks at the most important operations, such as separation, combination, and so forth. The next criterion applied is the state of aggregation of the starting material (see also [207]).

The general production instructions for the making of a given product (Fig. 25, page 28) also arise from process engineering.

While the first attempts to formalize production instructions were based on the classical structures of thinking in process engineering [208], more recent approaches start with the phase model of production derived from the

	Dept.	Process/Plant _____			Section
		Works/battery _____	GB/ZB _____		Sheet No. of
		Building _____	Project No. _____		
1	**Enameled Stirred Tank, Stirrer with Drive for Stirred Tank,**				Code
2	____ m^3				
3	Customer	File No.		Date	Class/Rev. No.
4	Department	Comm. No.	Inquiry No.		
5	Drawing No.		Order No.		Inventory No.
6	Manufacturer	Stock No.	Bid No.		
7			Contract No.		Code No.
8	Previous Order/Contract Nos.				

9	Drive	Top drive ○	Bottom drive ○	Inf. variable from to	Drawing No.
10	Requirements	Speed rpm		In steps	
11	DIN 28161-A ○	Sliding-ring seal removable at side ○			
12	Gearing	Not variable ○		Pole-reversing ○	Drawing No.
13		Speed variable ○		Gear-shift ○	
14		Continuously	Hydraulic ○	Change gears ○	
15			Mechanical ○	V-belt pulleys ○	
16			Electrical ○	Torque transmitted Nm	
17		Manufacturer		Model, size	
18	Skirt	Companion dimensions DIN 28162 Part 1 ○			
19	Clutch	Motor-gearbox		Gearbox-stirrer shaft	
20					
21	Motor	Power kW		Enclosure	Remark:
22		Speed rpm		Explosion protection	
23		Voltage V		Pole-changeable ○	
24		Frequency Hz		OK for variable frequency ○	
25		Design		Size	
26	Stirrer	Impeller ○	DIN 28157 ○	One-piece ○	Drawing No.
27		Dual impeller ○		Split ○	
28		Anchor ○	DIN 28158 ○	Total stirrer length mm	
29				Stirrer diam. mm	Enamel grade
30					
31	Stirrer shaft	Diam. in seal mm		Shaft enameled in seal ○	Drawing No.
32		Shaft split ○		Shaft bushing metallic ○	
33		Shaft integral ○		Material	
34		Shaft end DIN 28159 ○			Enamel grade
35					
36	Stirrer shaft seal: Inst. dim.	DIN 28138 Part 2 ○			
37	Stuffing box	Oil buffer chamber ring ○		External cooling ○	Drawing No.
38		Leak catching cup ○			
39	Packing rings	Number required		Material	
40		Manufacturer		Model, size	
41	Rot.mech. seal	Manufacturer		Model, size	Remark:
42	WN 1560 Pt 3 ○	GL No.			
43	Face ring matl	Product side		Atmosphere side	
44		Flushing medium			
45	Flush pressure	Manufacturer		Model	
46	Recirculation	Thermosiphon ○		Forced circulation ○	
47		Cooler ○	No cooler ○	Flush pressure bar	
48				Circulation rate L/min	
49	Inspections	Factory Standard 6345			

50	**Remarks:** Spare part(s) to fit Stock No.:	Whse. No.
51		
52		
53		
54		Recd. AMV:
55		
56		
57		
58		

59	Name		Change	1	2	3	4
60	Date						

Figure 21. Data sheet for an enameled stirred tank

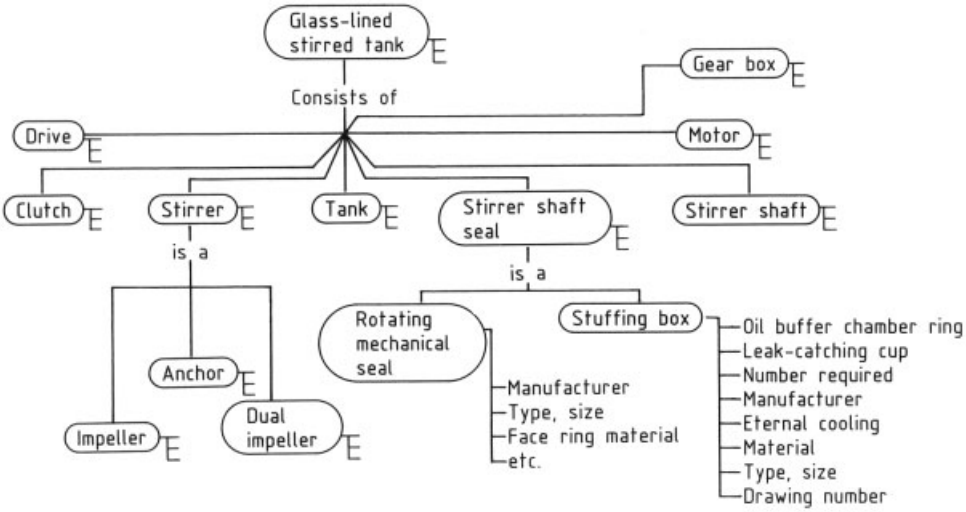

Figure 22. Data model of a stirred tank in the entity – relationship (ER) representation

transformation principle [209]. The configuration principle of modern process control systems builds on knowledge from object-oriented programming and, more recently, experience in modularization [210–212]. Figure 26 (page 29) illustrates the correspondence of plant model and process model (see also Fig. 65).

Process Control Engineering. To illustrate the relationship between process engineering and process control engineering, consider what happens in the transition from a generally valid recipe for making product A to a recipe for making product A in plant X. In a taxonomy of recipes, this is a concretization of the "plant-independent" basic recipe into a plant-dependent control recipe. New research by the ISA, particularly on the topic of batch control systems [213], and by NAMUR-AK 2.3 on requirements for recipe-mode operation [214] must be taken seriously as a basis for discussion. Figure 27 (page 29) illustrates the principal recipe types [214].

If one assumes a self-contained recipe for the individual operating modes (see Fig. 123), the taxonomy of recipes shown in Figure 28 results (see page 29).

Level Model of a Corporation. Structuring by hierarchical levels (functional abstraction) is encountered in a variety of forms (e.g., for the description of computer networks or data bases). A corporation can also be represented in this way.

According to a suggestion made earlier [189, 215], if the abstraction principles stated above are applied to a corporation, four levels can be identified. By way of example, Figure 29 shows schematically a model of a company involved in the process industry. The four levels are the corporate management level, the production management level, the process control level, and the field level. The dispositive quality of the functions decreases from top to bottom, while the operative quality increases.

In what follows, the corporate, production, and process control levels are discussed for the production division of a chemical industry corporation [168, 216].

Corporate Management Level. In a world where increasing competition demands rapid and often massive reactions in many markets, the corporate management level, where goal-setting and long-term planning are done, is becoming more and more important.

The key condition for the support of the corporate management level, once systems-engineering problems have been solved, is effective control of the lower levels, in particular the establishment of conceptual information structures.

For this purpose, the information available on the subordinate levels must be compressed in a way appropriate to its use. A strongly

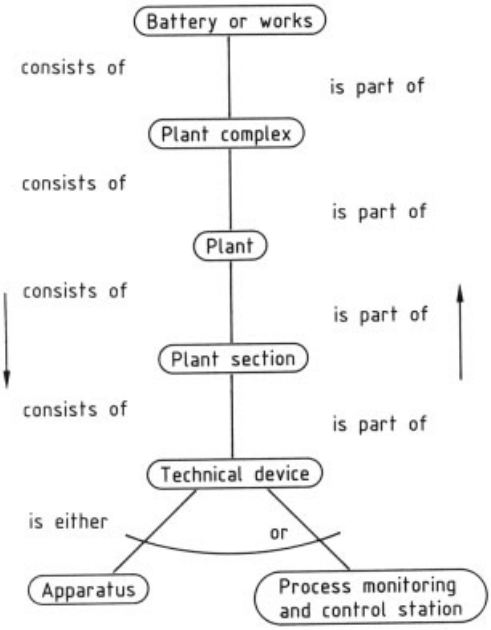

Figure 23. Plant model [204]

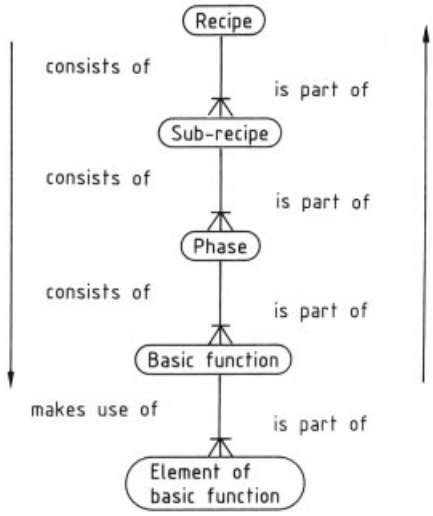

Figure 25. Structure of production instructions (process model, recipe) [172, 198]

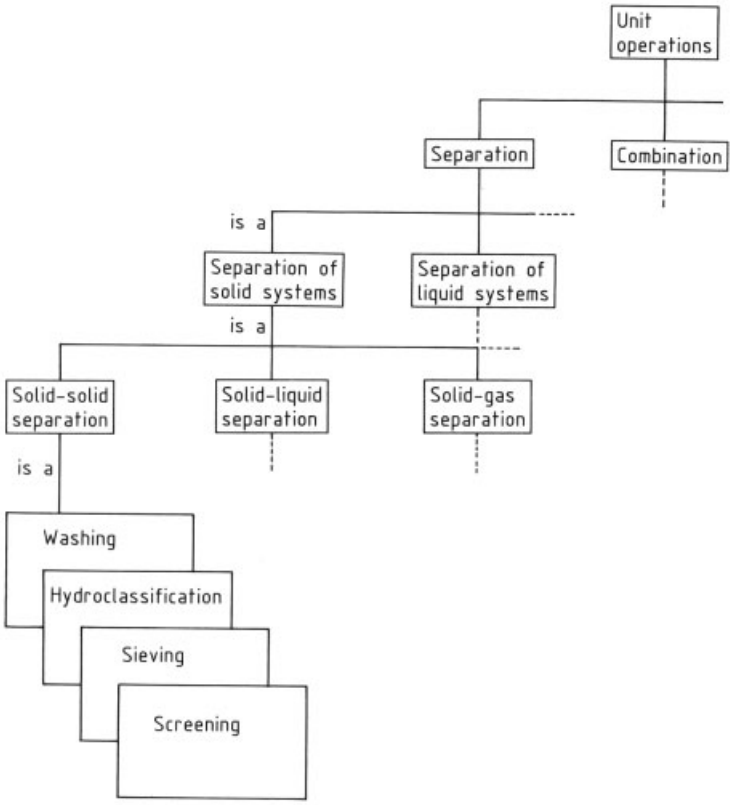

Figure 24. Taxonomy of unit operations (process engineering) (GDR standard [206])

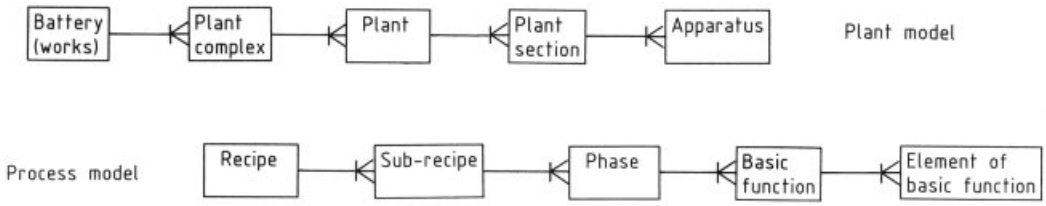

Figure 26. Correspondence of process model and plant model

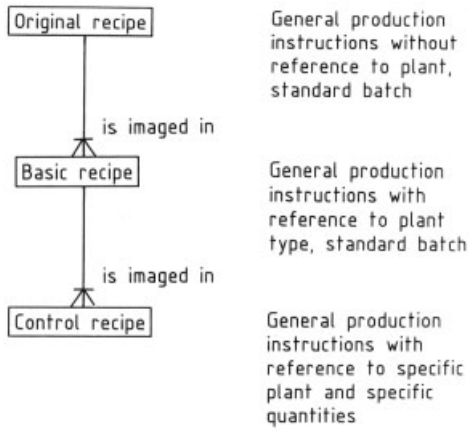

Figure 27. Types of recipe [214]

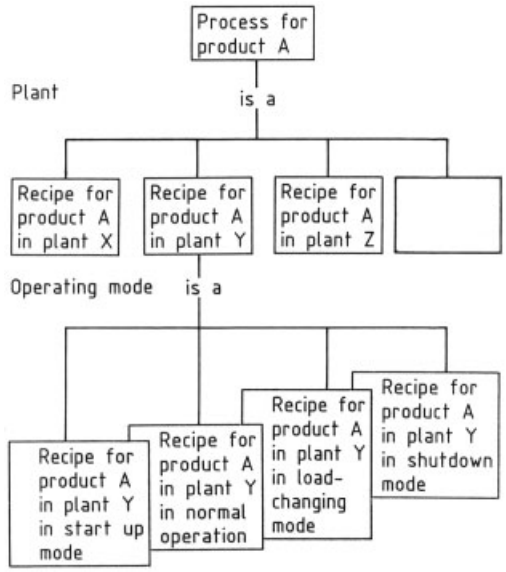

Figure 28. Taxonomy of production instructions

department-oriented view is supplanted by a supra-departmental one; a country-specific view by an increasingly international and global one; and a focus on the individual industry or sector is supplanted by a view that takes in many industries.

The increasing importance of effective product development management, for example, makes clear the need for a supra-departmental view. In order to avoid costs for dead-end development and delays in development, concrete requirement profiles should be derived from consumer needs at an early stage, the existence of competing products should be investigated, and the requirements of downstream processing industries should be taken into consideration (see also Section 11.3 and Figs. 397 ff.). This information should be derived with internal development capabilities. Corporate management should be able to examine the development process at any time. Product quality and environmental protection are other examples where new information aggregates, cutting across classical departmental lines and extending all the way up to the corporate management level, are gaining in importance and must therefore be made available (see also Fig. 6. in [217]).

Any company that wants to make intelligent decisions on a worldwide basis regarding, for example, research and production sites and markets to be entered, while optimizing its decisions in a supra-departmental fashion, must have consistent, up-to-date information covering all countries of interest. The result will be increased demands on internal accounting in the sense of comprehensive control, that is, a control instrument for corporate management. Highly diverse, historically evolved structures cannot meet these requirements.

Particularly in the chemical industry, the product orientation that comes about because of manufacturing conditions must be harmonized with the interests of the consumer market, which often have to do with more than one product. An integrated way of looking at problems can result in synergistic effects in purchasing management

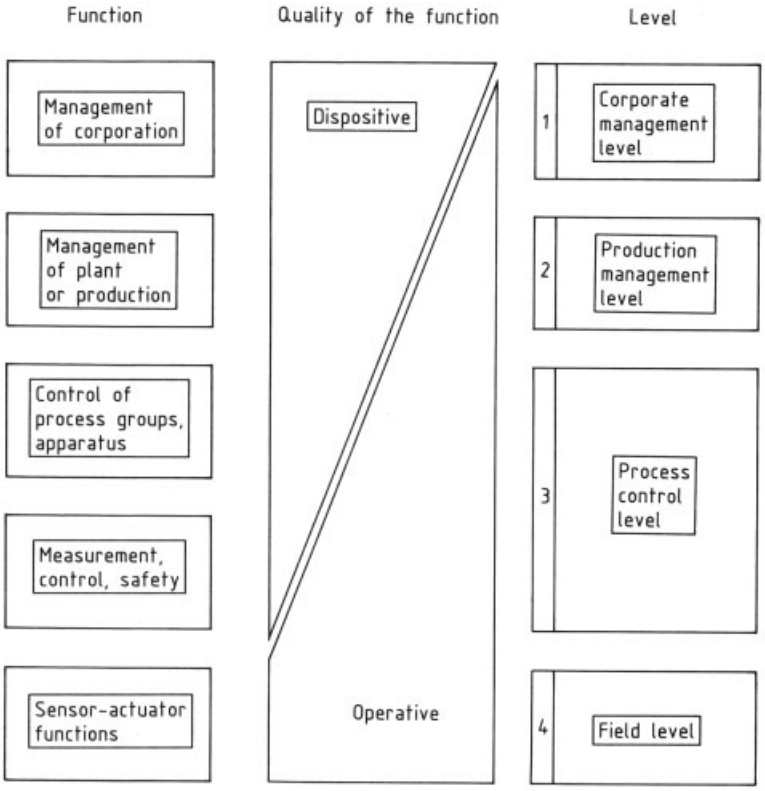

Figure 29. Level model of a corporation [168]

as well. A high degree of information transparency at a level higher than the industry or sector is therefore required.

The contribution of information technology at the corporate management level was very limited in the past, because it was restricted to supporting activities directly related to performance or order filling. The technology was marked by an orientation toward single tasks. Data structures were devised in a task-specific manner, in the narrowest sense; data bases were hierarchically structured (see also [144]). Data communications between programs also featured a mainly hierarchical structure. The task-oriented application systems based on this technology naturally could make little contribution to supra-departmental views. The core requirements of integration and flexibility could not be met. Personal computers were often made available to staff personnel on the management floors in an attempt to close this gap; most such computers were manually supplied with data from the functional systems and–not surprisingly–produced a wealth of incompatible data.

More recently, however, information technology has reached a stage in which the logical requirements of integration and flexibility can increasingly be realized. This applies both to information storage and to information retrieval and interpretation. The first comprehensive concept was developed for all engineering divisions of Bayer in 1989 – 1992 [178, 218].

Requirements imposed on the lower levels are illustrated for the case of production in what follows. A supra-departmental view is selected; that is, the interfaces to other departments are examined (see also Chapter 8).

Production Management Level. The discussion concerning the "factory of the future" has provided an outline of how the industrial production of the future will look. Although each expert's picture of this factory is different in detail, there is a consensus as to the need for increasingly extensive automation (see Fig. 124)

of production processes for reasons of quality assurance and optimal utilization of raw materials [216, 219].

Mass production is no longer a prerequisite for automation. The contemplated decentralized automation world, with the self-contained manufacturing islands already in existence, can be realized only if it is connected by unified information and communications systems. The objectives are to improve material flow, shorten storage times, increase the speed of transportation, and cut setup times when the product mix changes. These objectives are essentially the same no matter whether production is a manufacturing or a chemical process. Even so, there is a great difference between the implementations of these two types of systems.

Typical for the manufacturing industry are situations in which one or a few end products are assembled from many individual parts. The assembly process is complicated but lends itself to a virtually complete computer analysis [220].

In process industries, such as the chemical industry, the situation is reversed: There is one starting material from which one product or several related products are made (see diagram F8 on p. 379 in [216]). This simple, branching flowsheet allows stage-to-stage planning; any change will affect the next stage downstream, but can still be made at any time. The reason is that the process and product properties in each process element are interrelated (see Fig. 34). The interrelations, in general, can be described only in a phenomenological way, so that theoretical relations cannot be applied directly.

However, there are also many multiply branched flowsheets in which many end products are produced via several intermediates. These multi-stage batch or continuous production processes have the feature that the central event, the chemical reaction itself, is essentially a probabilistic occurrence and thus calculable only to a limited extent [220]. Such processes cannot be handled by production planning methods that are oriented toward parts lists.

Thus the systems integrated in CIM cannot be immediately and directly extended to continuous-process production.

The general structure shown in Figure 17 is now implemented in a manner specific to the production management level in Figure 30 [221, 222]. Production specifications supplied by the corporate management level are the following:

- Type of product
- Quantity of product
- Delivery schedule
- Special contract information (e.g., special packaging or formulation)

Detailing procedures are carried out with the aid of level-specific algorithms. Such procedures include adoption of production specifications, selection of basic recipes, comparison of material and apparatus requirements with inventories and free capacities, allotment to production units, assignment of apparatus-specific control recipes, and preparation of schedules. This detailing is often an iterative process that becomes necessary when the nominal/actual comparison in the evaluation algorithm signals disturbances in the production process or when preconditions for detailing are changed during the detailing process itself.

This block of functions has a strongly dispositive character. Once the detailing and disposition process has been completed, the production control function communicates it to the lower process control level in the form of a control recipe (when, what, where, how). The process control level in turn reports the relevant process and product data upward for further compression, either while the process is running or after it has been completed.

These downwardly and upwardly oriented algorithms include specific level data, which take the form of static inventories of apparatus and recipes. The updated execution plans, inventories, and actual product and process data are kept accessible in dynamic data bases.

This type of data management naturally includes a suitable documentation function, which can generate both routine reports and notices of special requirements. Another algorithm, not shown in Figure 30, deals with the optimization function. Points such as batch size, tooling time, startup and shutdown procedures, and stockpiles should be optimized. It is also important to modify the basic recipe and control recipe after the interpretation of process and product data (see Section 3.2) if product consequences reveal incompatibilities or if certain operating points turn out to be sensitive.

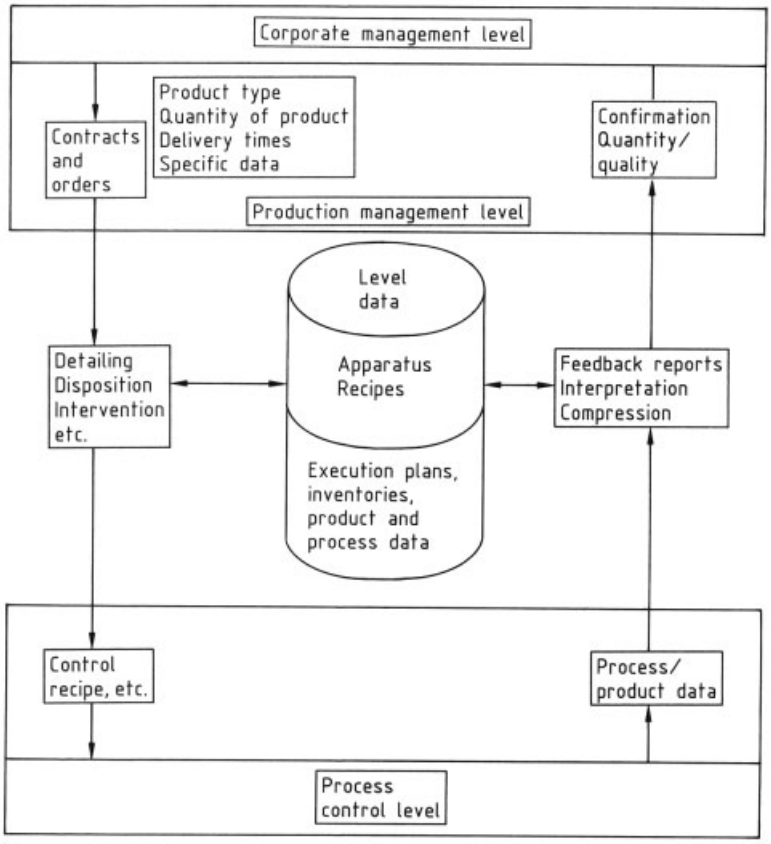

Figure 30. Production management level: structure and relationships

In the near future, such a production management system will become a permanent component of all production operations in the process industries. These functions will be implemented in a decentralized manner in individual operations.

Production management systems must be available in a configurable, off-the-shelf form, if only for economic reasons. The individual functions must therefore be implemented in the form of self-contained, comprehensible subsystems (Fig. 31).

According to Riemer, users can do the following on the basis of information that is kept accessible and up-to-date [223] (see also [221, 222]):

- On the basis of sales and operating quantity plans and using inventories and orders, they can determine in advance the net demand for products and feedstocks over all production stages (production planning).
- For this planned demand, on the basis of the instantaneous inventory, purchasing, and orderbook situation, they can ensure the timely procurement of all materials needed for fulfillment of the plan and monitor execution (demand disposition).
- They can keep their own production operations on target as to schedule, quantity, quality, and cost; allot materials and capacities; prepare working documents; continuously monitor production progress; and take prompt action when actual values deviate from nominal values (production control).
- They can accept specifications for production process control; support and monitor the process; and, by direct intervention as well as aggregation and structuring, prepare the continuously acquired production data for production control (process control; see Section 4.5).

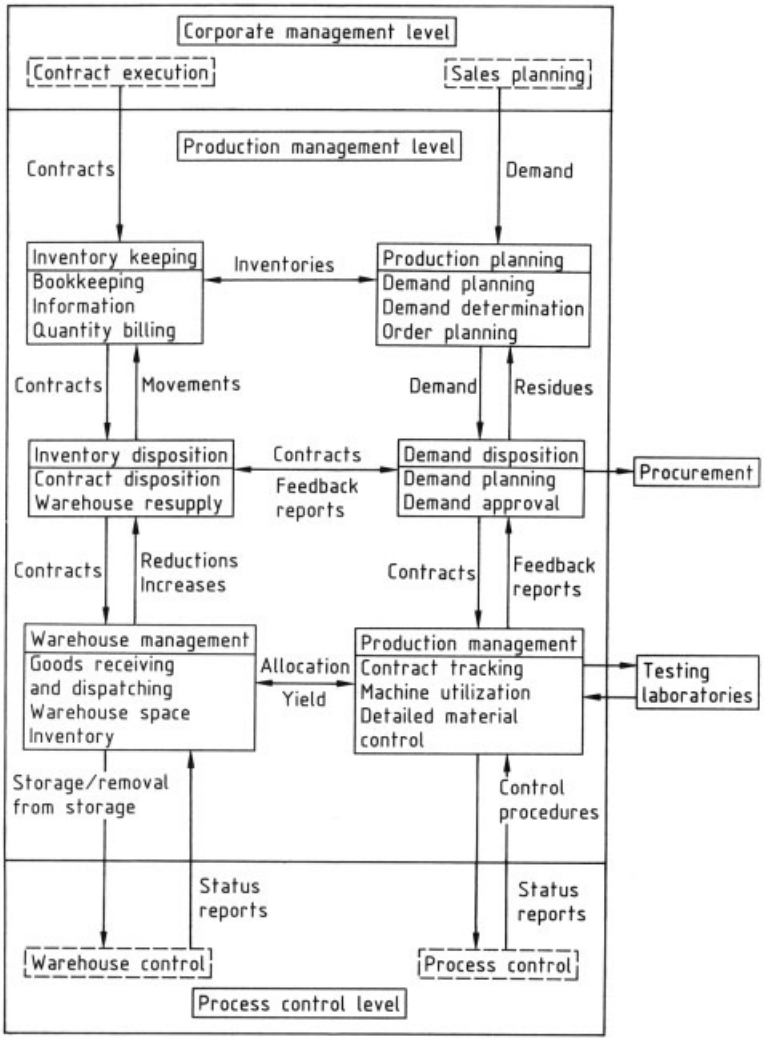

Figure 31. Functions of a production management system [223]

- They can better support the management of object (products, packaging, etc.) inventories (inventory keeping).
- They can automatically or manually check and plan inventories on the basis of customer, internal, and production orders (inventory disposition).
- They can have the materials at their disposal stored, removed from storage, commissioned, packaged, sealed, and prepared for shipping (warehouse management).
- They can store materials and remove them automatically and with optimal routing of warehouse vehicles (warehouse control).
- They can monitor what is happening in the process, using the minimum number of clear signals and standard information, whether by ad hoc queries on the screen or as printout or graphical output (information and report management).
- They can acquire current data with the required level of detail, store the data, and make them available at various locations for the various functional divisions (corporate planning, controlling, accounting, procurement, duties and taxes, etc.) whether for information or for processing (interfaces).

The example of the production management level shows how the analog of CIM in manufacturing operations is computer integrated processing (CIP) [224] in continuous-process operations. Clearly, information has now become a factor of production.

Process Control Level. In the process control level, production orders (i.e., requirements on product, quantity, and delivery time) are transformed to realization processes. This change requires that a processing facility be available, the control recipe (i.e., the recipe adapted to the actual facility being used) has been determined, and all necessary resources are available (see Section 4.5).

The production process is generally subdivided into a sequence of process elements each self-contained in time. The structure of the process monitoring and control system must correspond to this horizontal breakdown of the process. Process control functions for plant sections must be performed in decentralized process control stations and must be suited to handling the information flow in parallel with the process [168, 225]. The performance of the production order then involves the allotment of partial control recipes to plant sections and the phase-byphase execution of the recipes through the basic functions of process control engineering (see Figs. 25 and 38).

Corresponding to the hierarchy of processes and subprocesses, the following functional complexes can be distinguished on the process control level (Fig. 32; see also Chaps. 7 and 8) [226]:

- Central functions for controlling the process itself (process operating computers, process control computers, process AI computers) and for controlling the process monitoring and control system (process engineering computers for design and diagnostic service). The complex operations involved in diagnosing possible error states of a process control system are merely indicated here (see Chap. 7 and Sections 11.2 and 11.3). The analogy with medical diagnostics should also be mentioned [227].
- Individual functions and coordinaton functions: the grouping of all process-level functions in aplant section(process section or process element) in a decentralized process control station (process-level components, PLC).
- System communications functions thought of as interface functions (system bus, field bus).

Phase Model of Production. An important aid for systems analysis of production processes is the phase model of production, which was adopted from software engineering [228] and belongs to the category of product nets. Other models lack any representation of product properties in between process elements, which is a prerequisite for constructable quality assurance and process reliability; the phase model supplies this representation. As suggested by Büsing [229], the phase model is readily derived from the basic flowsheet by attaching a label for the products between the individual process elements (Fig. 33). Each rectangular box stands for a process element, and each circle for a product. The individual process elements can be regarded as operators that transform the product properties (see "Transformation Principle" in Section 2.2). This method facilitates, among other things, the representation of information flows, which are needed to control the entire process in such a way that it is always held in its instantaneous nominal state. The method also requires a clear picture of what information is (or should be) available in and downstream of each individual process element for process monitoring and quality assurance (see Section 11.3 and [230, 231]).

Monitoring and control functions relating to certain product and/or process properties must be carried out in each process element. Product and process properties are the object classes of information about the process [232]. In Figure 34 product and process properties are listed for the example of one portion of a phase model. Every product or process property consists of a category and a value.

Special attention should be paid to scaling (see Section 3.1, Fig. 40, and [233]).

Process properties are the following:

- State variables such as pressure, temperature, and concentration. The process can be described in terms of these or with derived state functions.
- Process parameters such as heat-transfer coefficient and catalyst activity. These identify the constraints under which the process runs. Process parameters are stationary or at least quasi-stationary quantities.

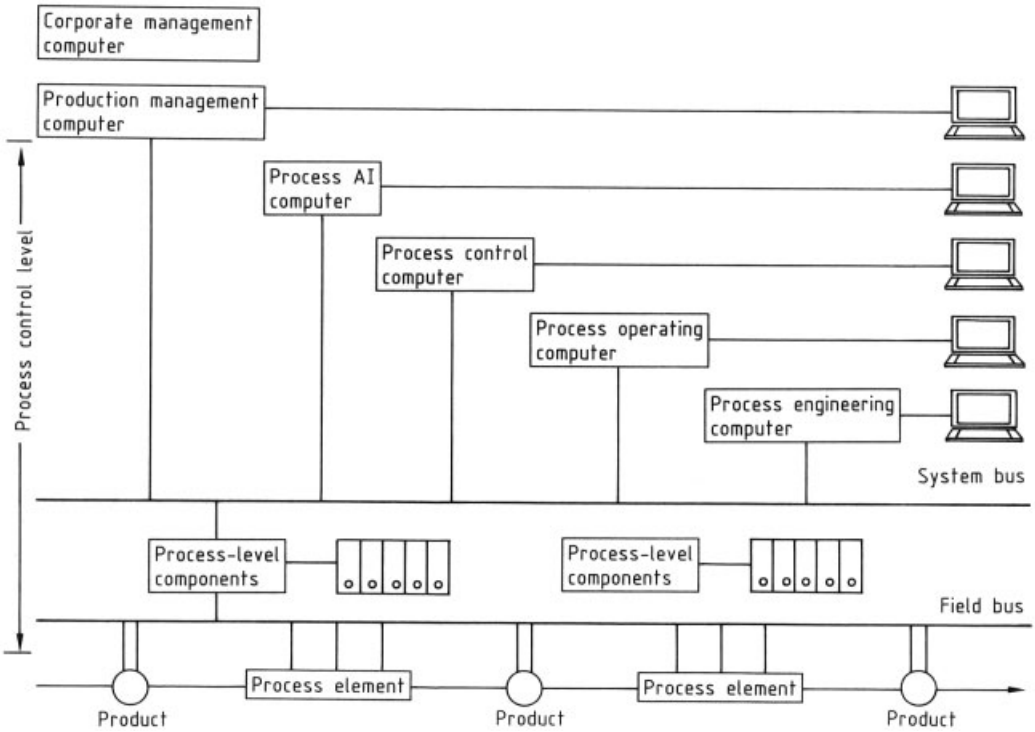

Figure 32. Ideal structure of a process/production control system

– Control variables. These characterize intervention in the process.

In theory, the information budget of an entire process can be completely described with an appropriate set of these quantities. In practice, however, it often happens that not all the desired state variables and control variables can be measured or specified. Other information is often employed as a substitute:

– Setpoints are correlated with the desired control variables.
– Process indicators are empirically correlated with one or more state variables (these may also be, for example, product properties).

Product properties are the following:

– Physical quantities
– Chemical quantities
– Technological properties
– Product indicators

The last two items, again, are stand-in values. Technological properties are determined in special tests when relevant product properties depend on physical quantities that are difficult to measure or when the relationship to physical quantities is not known. Product indicators are correlated with physical or chemical quantities. Process properties can be employed as product indicators. The importance of this interplay between product properties and process properties was pointed out as early as 1967 when Rumpf introduced the concept of "property functions" [234].

Process indicators [235] and product indicators are the forerunners of the results that are now obtained by model-aided measurement techniques [236].

Product properties and process properties can be represented as property profiles information stream accompanying the material stream in the phase-model [168] (see Fig. 35). As nominal values, they are an essential part of the production instructions (Section 4.5). As actual values, they are used in process monitoring (Section 11.2), process control (Section 4.5), process analysis (Sections 3.2 and 11.3), and quality assurance (see also [216,237,238]). Note that it

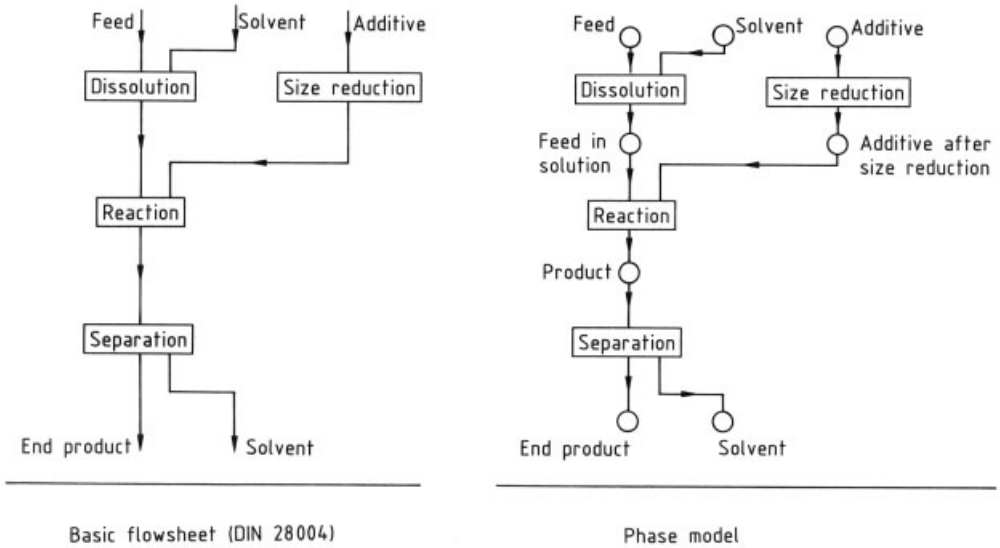

Figure 33. Development of the phase model from the basic flowsheet

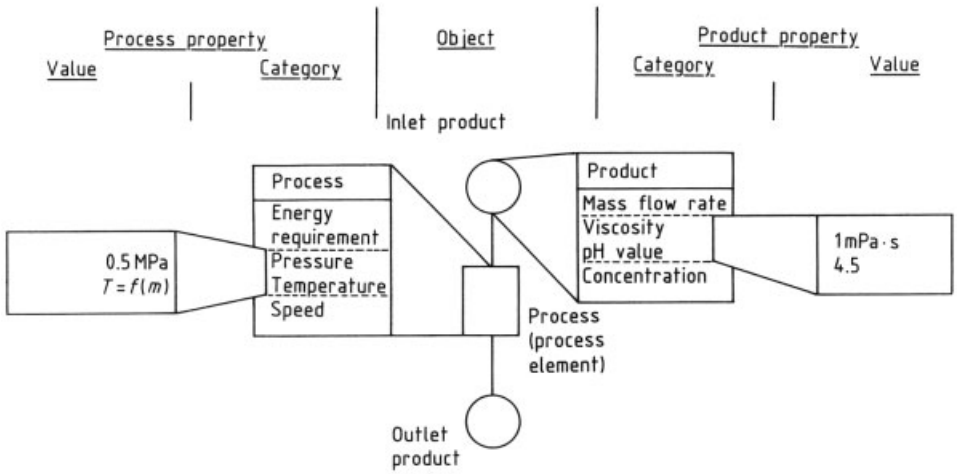

Figure 34. Excerpt of the phase model with product and process properties

is not just the values of the product properties that change from stage to stage; categories may also disappear and other new ones may appear. An understanding of this point is necessary in order to follow, for example, the suggestion of GILLES that quality be introduced as a state variable [230].

Figure 36 shows the property profile "quality" for an intermediate product, with its ranges of validity [239]. Figure 37 uses the same method to illustrate reliability.

Process engineering is customarily concerned with the material flow (and the energy flow, although this is not illustrated here) in a production operation, whereas the task of process control engineering is to deal with the information flow. The cross-link between the material flow and the information flow is provided by sensors and actuators. Sensor technology (Chap. 5) acquires the information required for process monitoring and control. The function of actuator technology is to derive actions on the process from information extracted from the information stream (see Chap. 6).

A phase model can be hierarchically refined in accordance with structuring principles. This

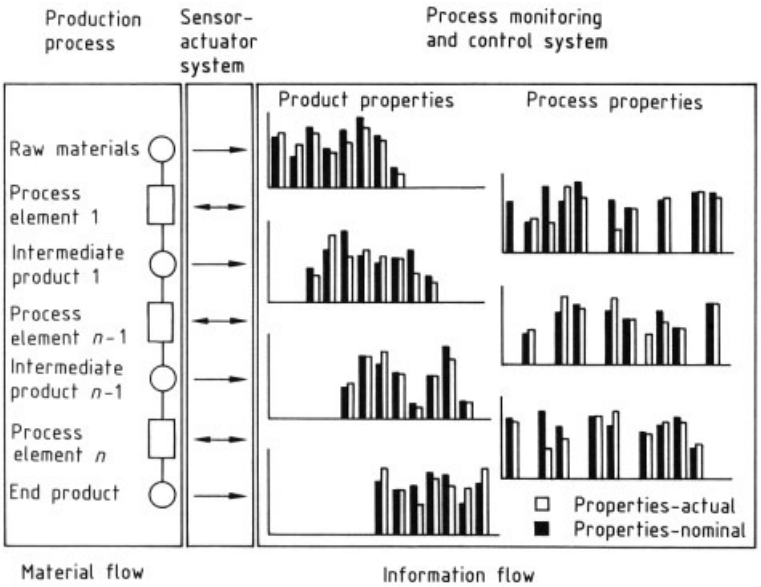

Figure 35. Profiles for process and product properties [135]

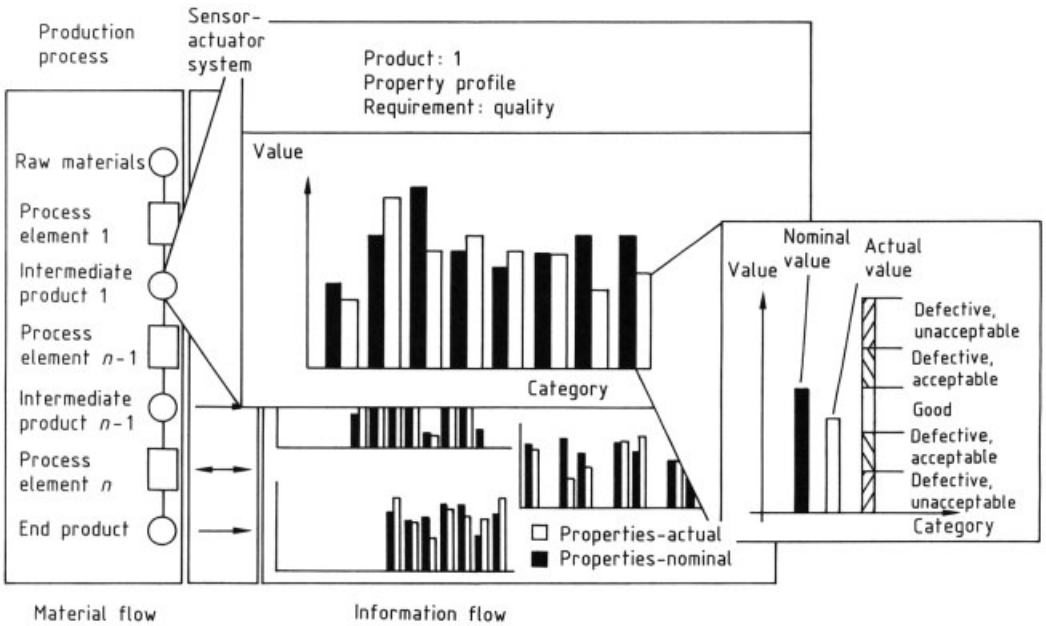

Figure 36. Material and information flows in the phase model of quality [238]

has been done for the example of automating a wastewater treatment plant [240]. Figure 38 shows the detailing process for chlor-alkali electrolysis [209].

Once the process analysis has been refined to the point that products cannot be broken down any further, a transition should be made from a product net to a causal model, either time-based or condition-based.

This transition is accomplished by transforming the interpretation "product with property profile" to the logical proposition, "product present in sufficient quantity and adequate quality at time t_0."

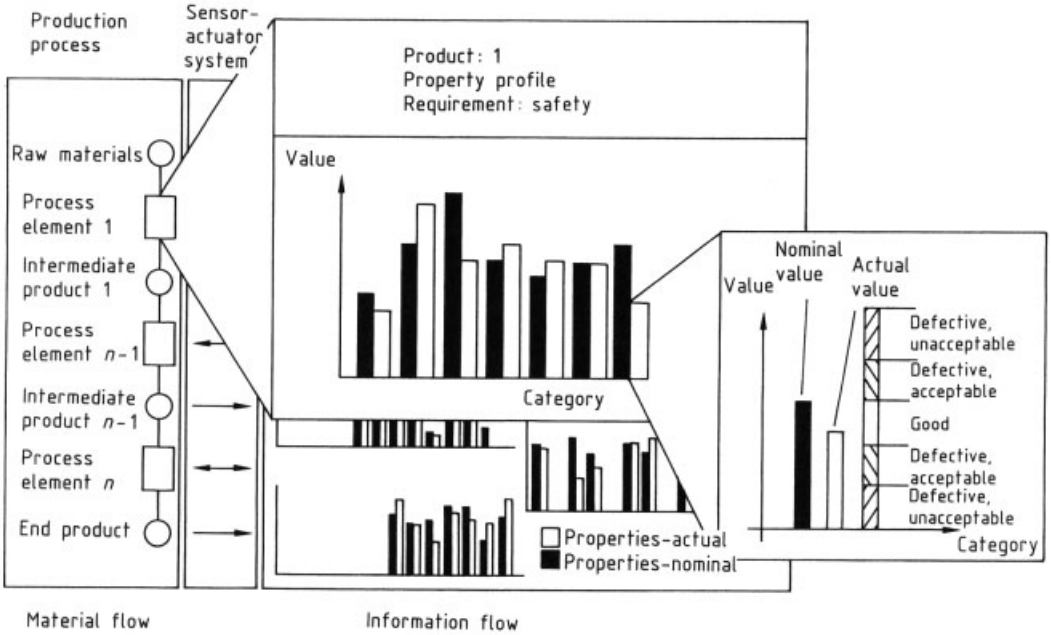

Figure 37. Material and information flows in the phase model of safety [238]

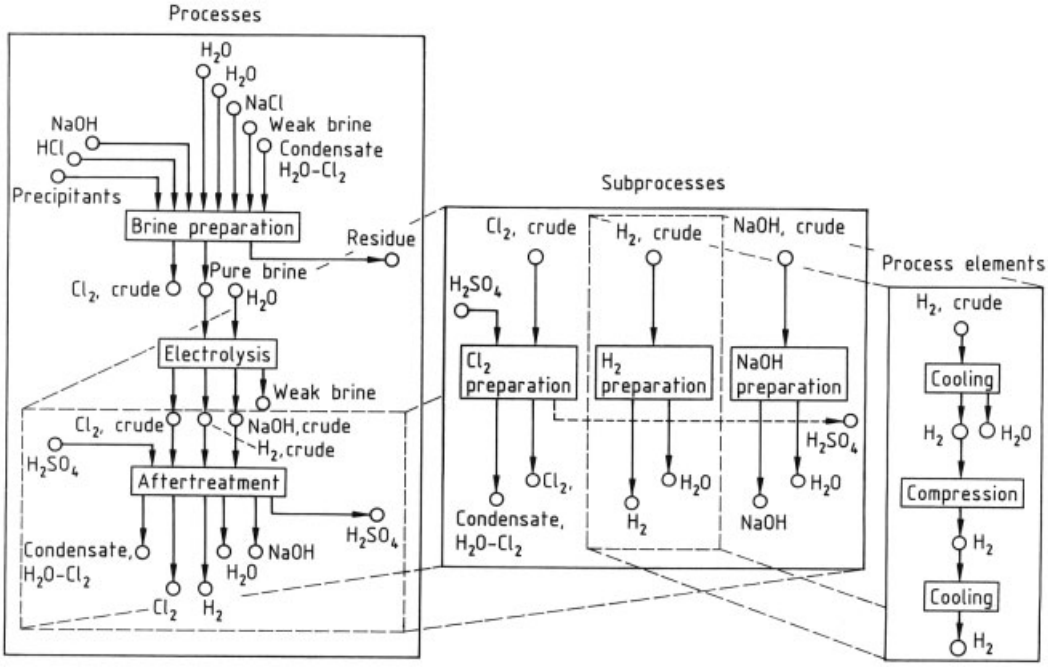

Figure 38. Detailing of the phase model for the example of chlor-alkali electrolysis

This logical proposition can be supplemented by further ones, for example:

- All resources needed for carrying out the process (plants, production specifications, monitoring and control system) are available and serviceable at time t_0
- All required energy sources are available in the proper quantity and quality at time t_0
- All safety devices are operable at time t_0
- Approval for startup has been issued at time t_0

With such a system of conditions applying to each minimal production unit, the way from product net to causal net is evident. In Figure 39, the transition is illustrated, starting with the phase model and ending with Petri nets, which in turn make it possible to present a control process in function charts complying with DIN 40 719 [241].

The left-hand part of Figure 39 shows the phase model of a simple production process. Each box represents a process element and corresponds to a unit operation. On this level of the hierarchy, changes in product properties are in the foreground; the way in which the unit operations are implemented is not of interest yet. If the top-down design process is continued and the technical implementations of the unit operations are specified (e.g., as shown in the diagram of Fig. 25), the middle part of Figure 39 shows the resulting detailed description. Each unit operation is now represented by its own Petri net relating to the setting of process properties. Basic functions are employed to establish the process properties. The production specification (or recipe) sets forth under what conditions which process property must be set to what value. Basic function elements are utilized in order to execute the basic functions. The resulting function chart describes the detailed solution in process control engineering terms (see also Section 4.5).

The phase model of production is thus the appropriate tool for describing (Chapter 10) and analyzing (Section 11.3) production processes with regard to quality assurance (Figs. 34, 35, 36), process reliability requirements (Fig. 37), static structuring (Fig. 38), and dynamic structuring (Fig. 39).

3. Knowledge about the Process

3.1. Principles

An essential feature of modern process control engineering is that knowledge about the production process is utilized to the fullest. "Knowledge about the process" does not refer just to knowledge of the static and dynamic attributes or properties of the object classes of the production process (e.g., process or product properties [242]), as now represented in entity–relationship models (Chap. 2) and often taking the form of observations and measurements. Instead, this phrase denotes knowledge about how these attributes depend on one another. Because they pertain to the system "production process," they will also be called "system quantities" in what follows.

This knowledge about the interdependence of system quantities can exist in various forms (e.g., characteristic curves, correspondence tables, mathematical process models). The power of such records depends heavily on the type of system quantity.

The type of quantity is established by the kind of scale used to measure it [243]. There are four types of scales: nominal, ordinal, interval, and proportional (Fig. 40). They differ in information content and interpretability. For example, the nominal scale permits only the relationship like/unlike (or equal/unequal) between two fixed values of a quantity; this applies to quantities such as tank numbers, production line numbers, and so forth. The ordinal scale contains the further information that an order relation exists, as in sensory evaluations or quality levels. With these scales, items can be identified not only by a wide variety of possible symbols but also by numbers, but the intervals between such numbers have no empirical meaning: The scales represent only topological information. Metric scales, where the differences between numerical values are interpretable, are the domain of physical quantities. An example is temperature measured in degrees Celsius, which is an interval scale. If the ratios between the numerical values of a quantity have a physical meaning, as is the case for masses, absolute temperatures, and so forth, the scale is said to be a proportional or ratio scale.

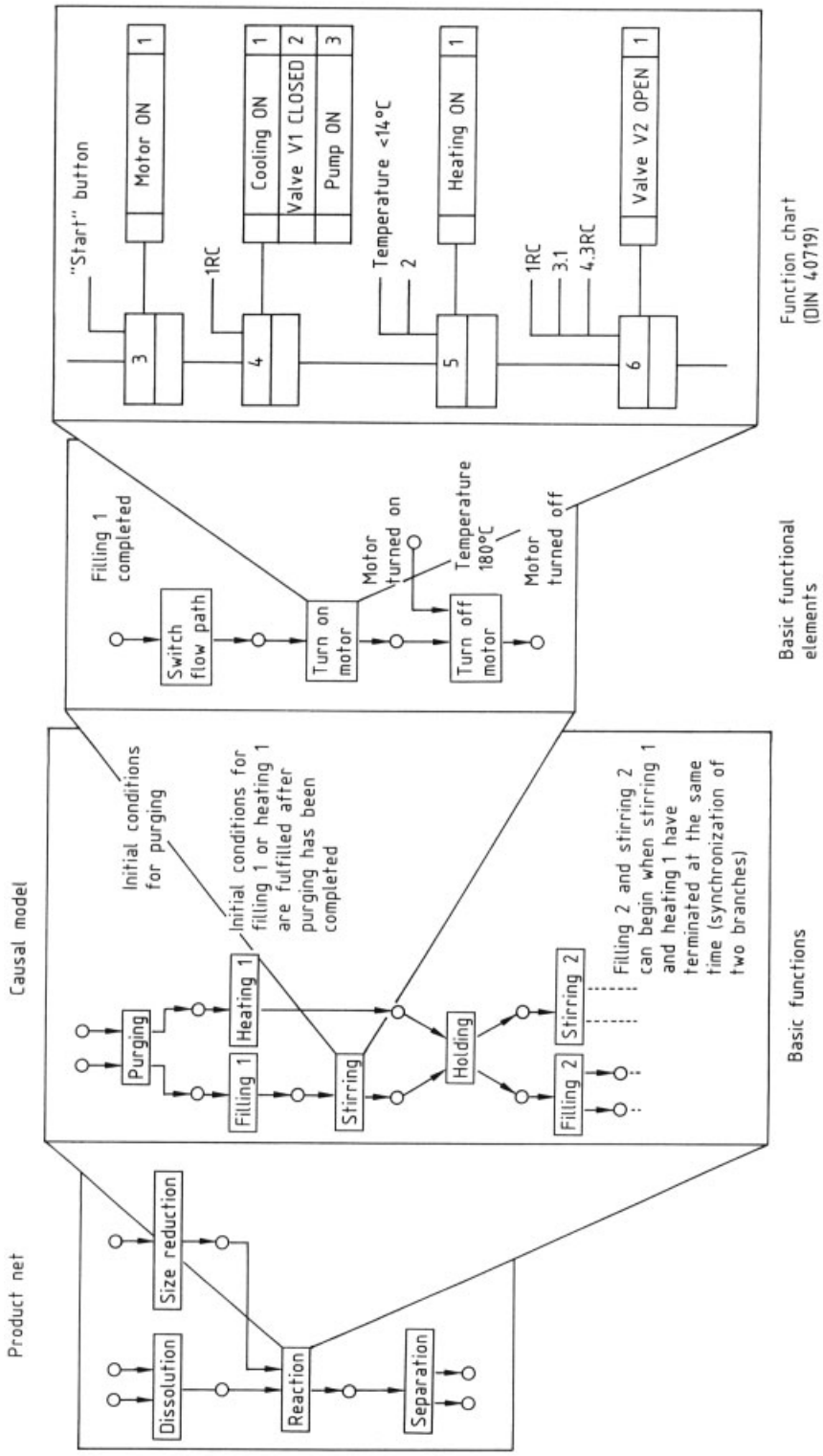

Figure 39. Transformation from the phase model via Petri nets to the function chart

Scale type	Interpretable relations				Examples
	Like/unlike	Ordering	Difference	Ratio	
Nominal	X				Tank no. Production no.
Ordinal	X	X			Quality classes Sensory evaluations
Interval	X	X	X		Temperature (°C) Time (h)
Proportional	X	X	X	X	Temperature (K) mass, energy

Figure 40. Types of scales

It is evident that the ways of representing knowledge depend on the type of scale applicable to the system quantities. Mathematical process models expressed in systems of differential equations can be employed only when the quantities lend themselves to a metric type of scale.

Before dealing with ordering methods for system quantities and the structure of mathematical process models, first some key concepts of systems theory will be elucidated that are significant in a process control engineering (PCE) context. The notion of "system" presented here contains a deliberate balance between needed generality (to afford sufficient degrees of freedom for separating the overall system into subsystems) and certain properties that are absolutely necessary for arriving at a general system structure. This is a point of great importance in the writing and application of software based on modular principles. The system concept is closely related to the phase model of production processes (Fig. 41) [244, 245], in which each process step can be interpreted as a transformation (see Sections 2.2 and 2.3). Certain product properties that exist before the inlet to the process stage, often represented as an unordered set $\{E_i\}_n$ or as the endpoint of a property vector in product space [246], are mapped onto the property vector existing after the process operation n with coordinates $\{E_\mathrm{i}\}_{n+1}$.

In simplified form, property vectors are written as x_in (inlet to the process stage) and x_out (outlet from the stage), as shown in Figure 41.

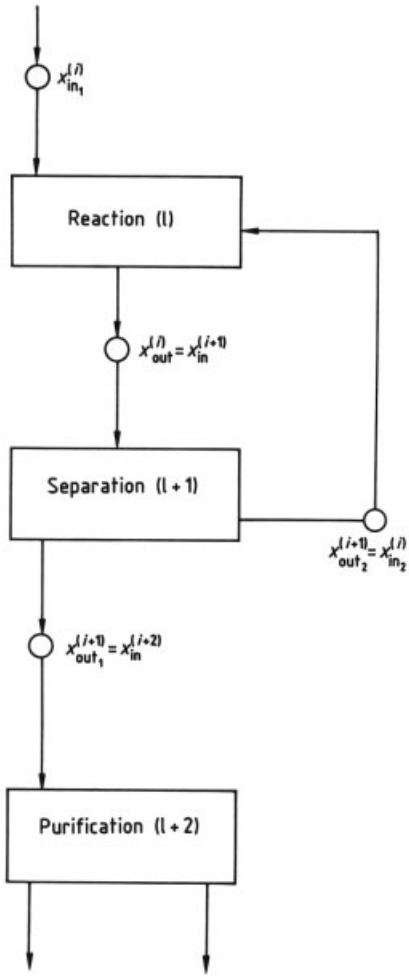

Figure 41. Phase model of a production process (excerpt)

The transformation can then be written as:

$$x_{\text{out}} = f\,(x_{\text{in}}) \tag{3.1.0}$$

In recent decades, the specialist literature has included many contributions attempting to justify a general theory of systems [247–254]. The axiom-based treatments, like all axiomatizations, represent a formal abstraction of concrete experience. The basic systems-theoretical concepts will be discussed here only in so far as they are essential in a PCE context.

In what follows, all system quantities are assumed to be "measurable," that is, representable on metrical scales.

The extension of this analysis to topological scaling, which has its own descriptive forms, leads to indefinite, conditional statements closely related to "fuzzy logic" and the methods of artificial intelligence (AI). They are not within the scope of this article (see also [255]). The scaling must, however, always be considered in the synthesis and analysis of production processes.

If the behavior of a reactor, such as a cascade of stirred tanks, a tubular reactor, or the like, is examined the system quantities can be classified in several categories.

For example, the following classes can be distinguished:

Inputs u_i: Mass flow rate in, concentrations and temperature in the inlet stream, cooling/heating temperature, etc.

Outputs y_i: Mass flow rate out, concentrations and temperature in the outlet stream, density of the outlet stream, etc.

System parameters p_i: Heat-transfer areas, heat-transfer coefficients, specific heats, enthalpies of reaction, kinetic constants, geometric quantities (pipe diameters and lengths), etc.

In this classification, the inputs play the role of independent (environmental) variables acting on the system. They can be thought of as perturbations or, when purposely manipulated, as control variables.

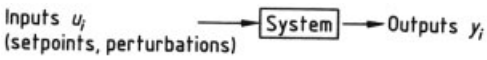

Figure 42. Schematic of a system

The outputs, on the other hand, are the means by which (at least in principle) the system acts on its environment (see Fig. 42).

It may seem surprising that state variables have not been mentioned up to this point. In fact, there is an important class of systems in which only steady-state processes take place, and in principle these can be described without the use of state variables. For example, in a cascade of stirred tanks operating in the steady state, the outputs are uniquely determined by the inputs. The introduction of state variables x_i is necessary only when investigating the time variations of systems. If, for example, the startup behavior of a reactor is to be studied it is not sufficient to know the time dependence $u\,(t\,)$ of the inputs; the "internal state" of the system must be known at least at a certain "initial time" t_0 before the outputs $y\,(t\,)$ can be predicted. This situation is reflected in the precise mathematical description when the system is formulated as a system of differential equations

$$\dot{x} = f\,(x, u, t) \tag{3.1.1}$$

which can be solved uniquely only by specifying an initial state x_0 at an initial time t_0.

Formulation (3.1.1) suggests that we generally attempt to describe a system by finding a vector x that has a minimum number of components. In the description of a chemical reactor, for example, one often arrives at a state vector x having the "reaction coordinates" ξ_i and the temperature T as components. All other quantities, such as concentrations, density, etc., are grouped in a vector z and can be stated as functions of x, u, and time t:

$$z = g\,(x, u, t) \tag{3.1.2}$$

The output vector y can be defined in a similar way:

$$y = C\,(x, u, t) \tag{3.1.3}$$

If a state vector x with the minimum number n of components is selected, n is said to be the number of degrees of freedom, or the order of the system.

Systems with finite order are also referred to as having lumped (concentrated) parameters. Systems with infinite order are then said to have distributed state variables. The state variables in such systems are generally continuous functions of the position coordinates.

The terms "system" and "process" are frequently used as synonyms. Here, however, "process" is defined as a one-parameter family of system states with the time t as the parameter. Accordingly, processes are generally dynamic; that is, their states are truly time-dependent unless special precautions are taken. Even when the inputs of a system are held constant over time, the system state will be a function of time at least during a startup phase. A special case of extraordinary importance for many problems in practical process control is that of stationary processes, where the system states $x(t)$ persist in a fixed state x_s and thus become independent of time. The state x_s is also said to be steady or stationary. The maintenance of specified steady states is one of the key tasks of process control. Steady states x_s should be considered only in time-invariant systems, where the time does not appear explicitly on the right-hand side of system equations (3.1.1) to (3.1.3). Clearly, the steady states in systems of this kind depend on the (constant) inlet variables u selected, and since $\dot{x}_s = 0$, the stationary system relating to (3.1.1) – (3.1.3) is immediately obtained if (3.1.1) is replaced by

$$0 = f(x_s, u) \tag{3.1.4}$$

For practical purposes, it is important to have a knowledge of the system stability near a stationary state x_s: If a process trajectory begins near state x_s, does the trajectory move farther away from x_s or return to x_s? This raises the question of how the system reacts to small disturbances away from its stationary point x_s. In order to handle this as well as some other questions, it is useful to have a concept of distance in relation to vectors. As is customary, the vector norm $\| x \|$ is used here; the distance between two vectors x_1 and x_2 is taken to be the norm of the difference: $\| x_1 - x_2 \|$. In what follows, the commonly encountered Euclidean norm

$$\|x\| = \|x\|_2 = \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2} \tag{3.1.5}$$

is used, where the ξ_i $(i = 1, 2, \ldots, n)$ are the components of the vector x.

Reconsidering the stability in the vicinity of a stationary state x_s. We first recognize that only the state equations $\dot{x} = f(x, u)$ have any role to play in this respect. LYAPUNOV defines the stability of a stationary solution x_s of $\dot{x} = f(x, u)$ in the following way:

1) x_s is stable if, for any sphere K_r with arbitrary radius r about x_s, there exists a number $\varrho > 0$ such that the following statement holds for a process trajectory with initial value $x(t_0)$: $x(t)$ lies in K_r for all $t \geq t_0$ provided that $\| x(t_0) - x_s \| < \varrho$ (Fig. 43).
2) x_s is asymptotically stable if x_s is stable and there exists a number ϱ_0 such that $\| x(t_0) - x_s \| < \varrho_0$ implies the convergence of $x(t)$ to x_s; that is, $\| x(t) - x_s \| \to 0$ as $t \to \infty$.
3) x_s is said to be unstable if x_s is not stable.

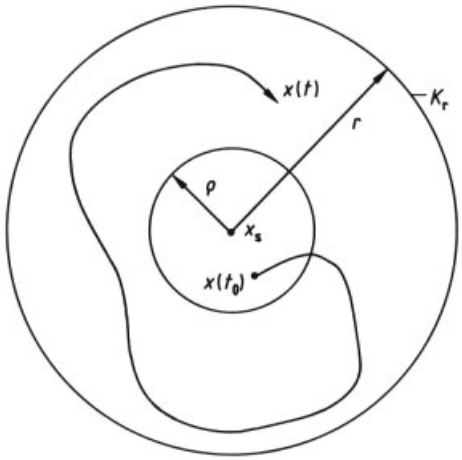

Figure 43. Lyapunov stability

More on stability and related problems can be found in [256–258].

The property of asymptotic stability says nothing about the rate at which process trajectories beginning near x_s converge to x_s. The notion of the relaxation time T_R serves this purpose: If the estimate

$$\|x(t) - x_s\| \leq \gamma e^{-\frac{t - t_0}{T_R}} \tag{3.1.6}$$

holds for all trajectories beginning sufficiently close to x_s, then x_s is exponentially asymptotically stable with relaxation time T_R.

Systems with a globally valid relaxation time T_R (i.e., T_R independent of specially chosen inputs u) have a remarkable property:

For sufficiently slow variations in the inputs u, they behave in a quasi-stationary manner. More precisely: If the rate of change with time $\dot{u}(t)$ of the input vector satisfies $T_R \cdot \| \dot{u}(t) \| \ll \| u(t) \|$, the time-variant state $x(t)$ can generally be calculated from the formally stationary equation $0 = f(x(t), u(t))$. The quasi-stationary system is purely algebraic (it no longer contains differential equations), and its dynamics depend solely on the time dependence of the inputs $u(t)$. The system thus has no intrinsic dynamics but responds to the time variations of u without inertia (time lag).

The transition to a quasi-stationary system is a radical reduction of the system: The initial system (3.1.1) has n degrees of freedom, but the new system has no internal degrees of freedom. Consider, for example, an ideally mixed stirred tank containing an indicator substance whose time-dependent concentration $u(t)$ in the inlet fulfils $\tau \cdot | \dot{u}(t) | \ll | u(t) |$ (where τ is the mean residence time in the tank). The outlet concentration $y(t)$ then satisfies $y(t) = u(t)$. The intrinsic dynamics of the tank, represented, for example, by its residence-time spectrum, is no longer identifiable.

The quasi-stationary property of subsystems within a specified overall system is not just important for numerical treatment: It is one of the key concepts for explaining "synergetic" effects in nonlinear systems [259].

Finally, two terms whose understanding is basic to the mathematical solution of most control engineering problems: controllability and observability [254, 260–264].

If at time t_0 a system is in a certain state x_0 and the inputs exhibit a time dependence $u(t)$ such that the system is in a specified state x_1 at a later time t_1, the system is said to be controllable from x_0 to x_1. If any x_0 is controllable to any x_1, the system is said to be completely controllable. Consider, for example, a stirred reactor in which the reaction $A \rightarrow B$ takes place, whereby the reaction is dependent on the temperature Θ. Let the state vector be $x = (\Theta, [A], [B])^T$, and the inputs u_1 = heating/cooling temperature Θ_K and u_2 = inlet concentration $[A]_{in}$ of species A. Then any state x_0 is controllable to any state x_1 provided the time interval from t_0 to t_1 is sufficiently long. However, if u_2 is not permitted as an input, then clearly x_0 is not controllable to x_1 if $[A]_0$ and $[A]_{in}$ are smaller than $[A]_1$. The controllability concept is of crucial importance for the problem of feedback control (returning the state to the setpoints) and the related problem of pole specification.

In the same way that the controllability links inputs u with state variables x, the observability relates the outputs y to the state variables. The state x is said to be observable at time t_1 if $x(t_1)$ can be determined from a knowledge of $y(t)$ in the finite time interval $t_0 < t < t_1$. In the reaction system discussed above, for example, suppose the state variable Θ is simultaneously taken as a measured quantity y. Then the remaining state variables [A] and [B] can be calculated from the time variation of the temperature Θ, and are hence observable, provided the enthalpy of reaction is nonzero. The observability concept is basic to the concept of the Luenberger observer in control engineering [254, 260, 265, 266].

3.2. Analysis Methods for Process Quantities

Introduction. Suppose k process quantities (i.e., state variables, inputs, or outputs) are measured at a certain time. These observations can be expressed as a k-dimensional vector $V(1, \ldots, k)$. If there are many observations, they form a cluster of points in a k-dimensional vector space, which will be called observation space or property space in what follows. Mathematical methods of linear algebra and statistics now become applicable [246]. Important uses of these methods include data reduction and transformation. A data reduction can be performed if the point cluster lies, at last approximately, in an m-dimensional subspace of the k-space. A transformation changes the coordinate system; that is, it generates a more easily handled representation by replacing the measured process quantities with "parameters" calculated from them.

Statistical models are valid only in the region studied; extrapolation beyond this region is not permitted. Furthermore, such models hold only under the assumption that all independent variables influencing the dependent variables have been identified.

Preparation of Process Properties. A statistical analysis can be broken down into the functions of acquisition, archiving, preparation,

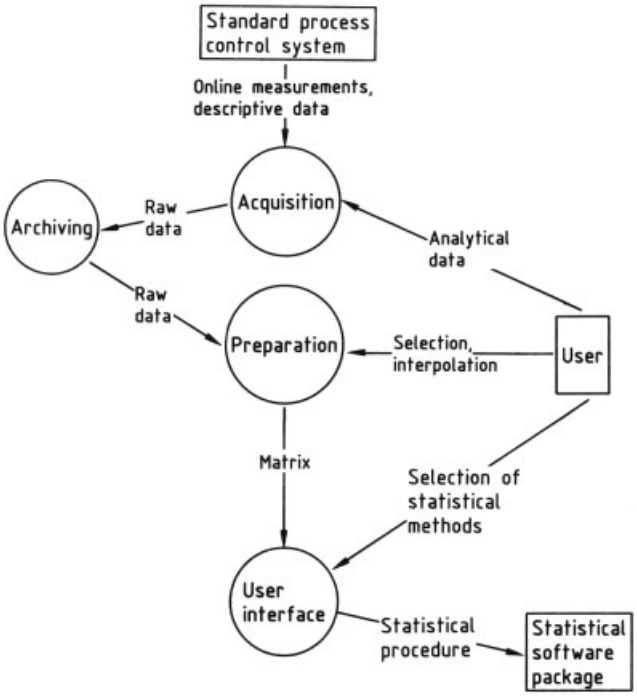

Figure 44. Acquisition, archiving, preparation, and interpretation of product and process data

and interpretation (Fig. 44). Statistics proper is part of the interpretation function and will be discussed in depth in the next section. Acquisition and preparation, however, are vital for the practical use of statistical methods and should therefore be explained in further detail.

Acquisition. Statistical research often entails a laborious data collection process. In plants equipped with process monitoring and control systems, the system acquires a large amount of process data for observation and for use in plant operation. Thus it is obviously desirable to use process monitoring and control systems to acquire data for use in statistical studies. When this is done, archiving must be added to the functions of a typical system so that data collected at a given time can be accessed at a later time. What is more, certain data on the configuration of the process control system must also be archived so that the archived process data can be properly interpreted even after a reconfiguration. Statistical software packages (e.g., SAS) can be used for the actual interpretation step; such packages for use on PCs offer good functionality at reasonable prices.

Preparation (Formatting). The starting point for many statistical procedures is a matrix W in which the rows represent observations and the columns represent the observed features (Fig. 45).

A typical requirement for a process monitoring and control system is that k process quantities be selected in a specified time interval from t_0 to t_m at a specified sampling rate. This would yield an $m \times k$ matrix (i.e., a matrix having m rows and k columns), which can be input directly to statistical analysis.

However, commercial software for process control systems does not to have this functionality. Nor does the simple selection of an $m \times k$ matrix satisfy all practical requirements, since professional presentation requires that descriptive data (name of measurement point, unit of measurement, etc.) be included. The extracted data must be output in a format that can be interpreted by commercial PC software. Because the result of data preparation is essentially a matrix, it makes sense to adopt a file format used in spreadsheet-type analysis.

Software. The preparation (formatting) and statistical analysis of process properties is not a

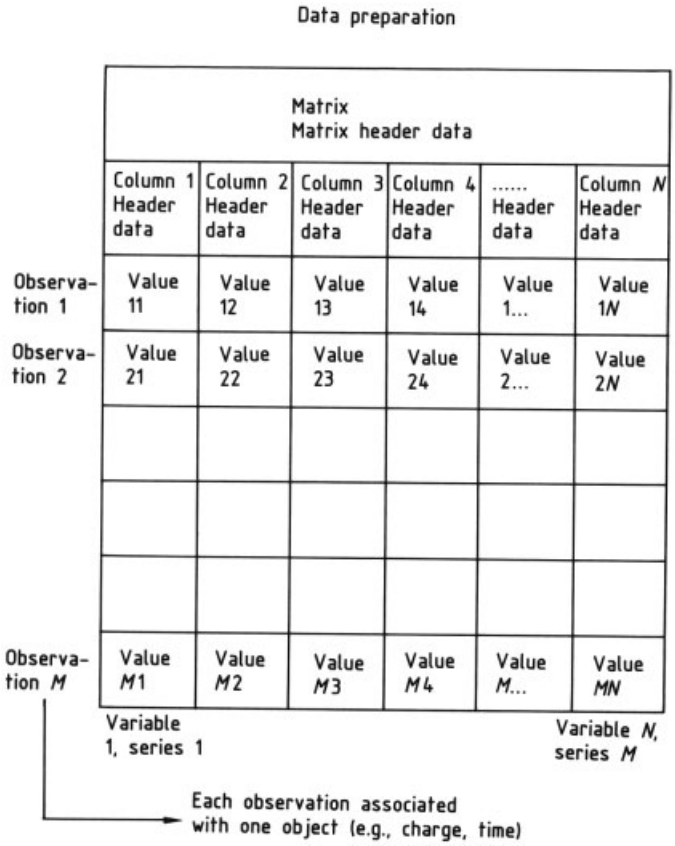

Figure 45. Data matrix for observations

routine activity for the person who does it. This makes it all the more important for the software to have a simple user interface. The format of the measurement station list must offer an easy way of marking the stations chosen for data acquisition (Fig. 46) [267]. Since all the selected points are sampled at the same time intervals, which as a rule do not coincide with the rhythm of data acquisition, an interpolation type must be specified for each individual station. The interpolation types include linear interpolation; constant incrementing of the measured value up to the next measurement; and mean, minimum, or maximum value (when several acquisition events fall into one data-preparation interval).

Once the process information has been prepared, the data can be transferred to a statistical software package via a suitable interface. As already mentioned, the user expects that not only the measurements, but also the descriptive data, will be extracted from the process control system. The statistical package must then use these descriptive data as captions, axis labels, etc., without any intervention by the user.

Many applications allow the user to define a catalog of statistical methods that can be employed. In such cases, a user interface with the following properties can then be implemented, either as part of the statistical software package or as an add-in:

- Spontaneous selection of a variable as a dependent variable
- Selection of statistical methods by marking a list
- The possibility of depositing rules for conditioned or excluded statistical methods
- Dynamic specification of numbers of variables and observations and identification of descriptive data for statistical procedure

If the statistical package employed does not have this functionality, the user interface can be implemented with additional software. A statis-

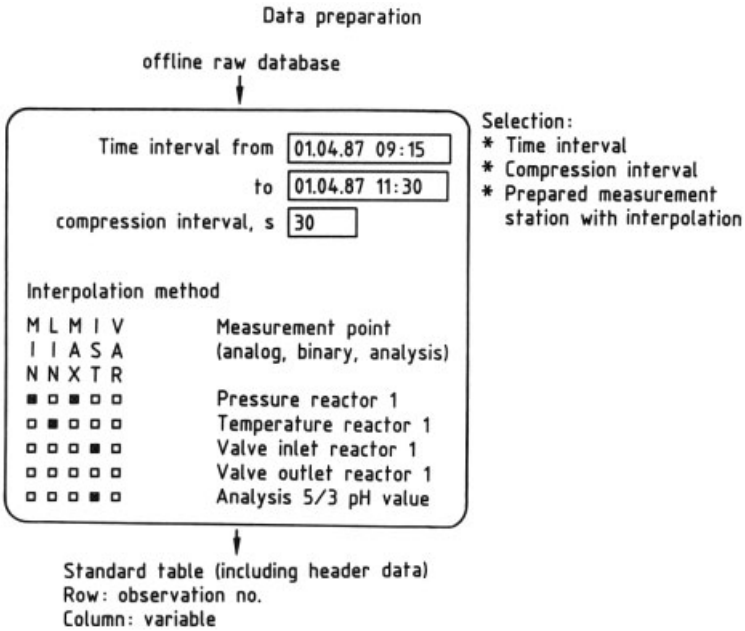

Figure 46. Interpolation of raw data to produce a standard table for subsequent analysis

tical procedure will then be written in which the numbers of variables and observations, names of dependent variables, and descriptive data are represented by placeholders. The steps in the procedure are defined as statistical methods. The add-on software uses selection rules to interpret user inputs, descriptive data, and files, replaces the placeholders, and ignores methods not selected. The procedure that results can now be handled directly by the statistical package.

Statistical Methods. This survey briefly describes individual statistical methods and discusses their applicability to the interpretation of process information.

Elementary Statistics. The examination of position parameters, such as mean, median, and mode, is meaningful especially when the process is expected to be stationary or pseudostationary (e.g., a constant flow rate in a continuous process). If the usual variability of the process has already been determined on the basis of large data sets, statistical quality control with control charts is possible. The purpose of control charts is to discriminate as accurately as possible between normal fluctuations of a value and unusual variations.

An important point when using a statistical test is to follow a definite methodology. This includes exact formulation of the hypothesis, so that the probabilities of Type I errors (rejection of a correct hypothesis) and Type II errors (acceptance of a false hypothesis) can be precisely studied and evaluated.

In many cases it is useful to determine whether a classical distribution function can be assumed. The important tests used in practice–the F test (comparison of two variances) and the t test (comparison of two means)–assume a normal distribution of the population.

Experimental Design. The concern of experimental design is to determine which of k potential factors $x_1, \ldots, x_k$ actually influence a dependent variable y, and to do so in the most efficient way by means of planned experiments. Two-level factorial experiments are important. First, a realistic minimum value (− level) and a realistic maximum value (+ level) are selected for each factor. Not only quantitative factors but also qualitative ones are possible; for example, + denotes product from plant A and − denotes product not from plant A.

The complete design then consists of 2^k experiments. It is obtained by taking all combina-

tions of the − and + levels for the w_i. The effects are the actions of the factors themselves, as well as the interactions of two or more factors. The level of an interaction is obtained by multiplying the levels of the individual factors. For example, if the factors w_1, w_2, w_3 in an experiment have levels +, −, −, then in this test the effect $w_{1,2,3}$ has level +.

After performing the experiments, the effects are calculated by taking the mean of all results from trials in which this effect had level + and subtracting the mean of all results from trials in which the effect was −. The literature gives formulas for deciding, on the basis of these results, whether an effect is significant.

If the assumption is justified that interactions of more than two factors are not to be expected, then the complete factorial experimental design can be replaced by a fractional factorial design calling for fewer tests.

Methods for generating partial factorial designs can be found in, for example, [268]. Software for creating and evaluating designs is available.

The interpretation of two-level factorial experimental designs is a special case of analysis of variance, a technique that can be used with any number of levels. The trick is to use two-level factors first to investigate which potential factors truly affect the experimental result y. The nature of this influence (e.g., whether it is linear) is investigated only after the merely notional factors have been identified and omitted.

Principal Components. On economic grounds, it is often not possible to carry out a designed experiment, and one must be content with random process measurements from the plant. It may be difficult to gain any knowledge from such a collection of data. Principal-component analysis is a method of orienting and, in some cases, reducing high-dimensional observations.

If the observations exist in k-dimensional space, principal-component analysis answers the following question: For any i, $0 < i < k$, how well can the k-dimensional observations be described in a subspace U of dimension i (Fig. 47) ? While a regression model minimizes only the squared distances in the direction of the dependent variable y, principal-component analysis minimizes the sum of the squared Euclidean distances between U and the individual observations. In this way, all the k process variables are equivalent; none is identified as dependent. Scaling now becomes a problem, especially when a variety of physical units are present. A change of scaling means a numerical change in the Euclidean distance in the k-dimensional observation space, and thus changes the result of the principal-component analysis. In practice, all the process variables are usually scaled so as to have mean 0 and variance 1. When this is done, however, the scaling becomes experiment-dependent. It is also possible to determine a technical minimum and maximum for each process variable, then define the scaling to give these the values −1 and +1.

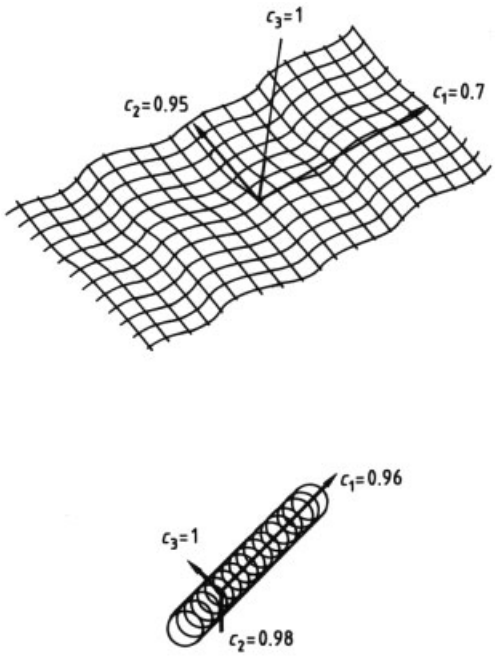

Figure 47. Effect of the shape of a point cluster on the principal components

The principal-component analysis determines, for every i, $0 < i < k$, the i-dimensional subspace U_i for which the sum of the squared Euclidean distances to the observations is a minimum. The procedure has the convenient mathematical property that these subspaces are all nested, that is, $U_1 \subseteq U_2 \subseteq \ldots \subseteq U_{k-1}$. They can thus be stated as a sequence of k basis vectors; these are the principal components. The mathematical procedure yields an eigenvalue for each principal component; the sum of all the eigenvalues is k. If, for some i, the sum of the first i

eigenvalues $l_1 + l_2 + \ldots l_i$ is close to k, the point cluster of the observations can be regarded as i-dimensional to satisfactory accuracy. Often, $c_i = (l_1 + l_2 + \ldots + l_i)/k$ is used as a parameter, known as the variance contribution of the first i principal components; this has a maximum value of 1.

Time-Series Analysis. A typical problem leading to a statistical study of process variables is the estimation of dead times. The central concept is cross-correlation: The correlation coefficients between the time series $x(t)$ and all time series $y(t+nT)$ are calculated, where T is the sampling rate. If N denotes the value of n that has the largest correlation coefficient, then NT is the estimated dead time.

This method has been extensively developed. Nonlinear relations between $x(t)$ and $y(t)$ can now be included, and processes for which the dead time is not a constant but depends on the nature of $x(t)$ can be investigated.

Systems Engineering. A number of software packages for the statistical analysis and graphical presentation of data have become commercially established. The methods employed are complex and exist in a wide range of variations, and no practically useful functionality is left out. The publishers of most of this software offer their products for a variety of system platforms (MS-DOS, UNIX, etc.), and data can be shared with other widely used software products (database management programs, spreadsheets, etc.). Examples of established statistical packages are SAS and RS/1. These products are continually being upgraded as standards develop (e.g., X/Windows). While market conditions are generally good for statistical applications software, there are some problems having to do with inadequate management of process data (Fig. 48). In practice, statistical analysis includes the functions of data acquisition, archiving, and preparation; these have already been described. But no software standards have yet been adopted in this area, and there are no dominant products in the market. One step in the right direction is The Fix, a process data acquisition system published by Intellution for MS-DOS. The system supports the preparation of historical data in Data Interchange Format (DIF), which many spreadsheet programs, as well as the SAS and RS/1 statistics packages, can handle.

Process monitoring and control systems published by the established firms, however, do not support the generation of a matrix (Fig. 49) including descriptive headers in place of a series of measured values.

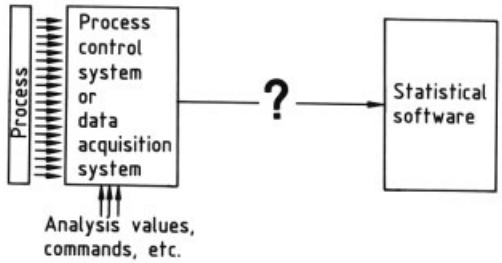

Figure 48. Problems in transferring inadequately organized data to statistical software

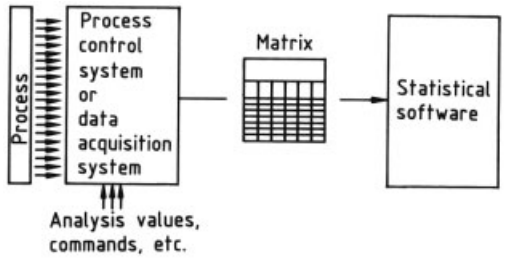

Figure 49. The data matrix as interface between process control or data acquisition system and statistical software

3.3. Process Models

Introduction. The general concept "model" is very broad. Any means of analysis and synthesis that is used in learning about reality or in actively designing and fabricating human artifacts can be called a model (see also Section 2.2). Typically, models are simplified pictures from a particular viewpoint of the properties and functions of the real object on which they are based. As for the chemical plant models in Figure 50 (see next page), it can be stated generally that a model reflects only a certain portion of the whole, from a certain aspect and for certain purposes.

The example clearly shows the three main features of all models [269]:

1) Mapping.
2) Abbreviation: The model generally lacks many of the attributes of the original. In fact,

this information loss is often an advantage because incidentals can be ignored in creating the model. One discards intellectual ballast so as to be able to focus better on the interesting aspect of the whole. These comments are closely related to the third model characteristic.

3) Pragmatism: The model represents the original, aiding in the acquisition of knowledge and permitting a limited range of mental or actual operations.

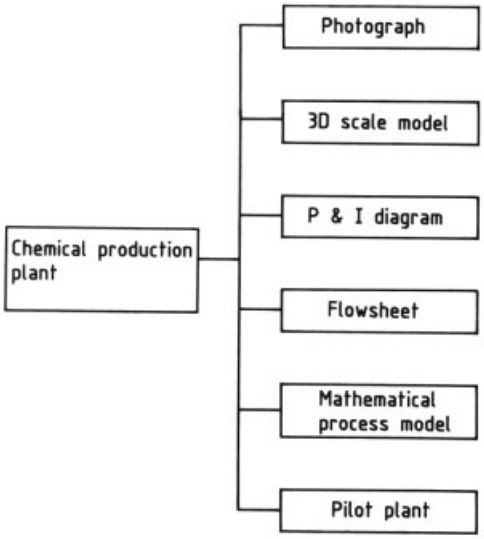

Figure 50. Model and reality

In contrast to models that represent physical images or simulations of particular originals, mathematical models are mental constructs that generally apply to an entire class of objects and situations. They thus fulfill a key prerequisite for being "scientific" ("Scientific theories are general laws" [270]). Mathematical models are formulated (as differential equations, etc.) in terms of variables that cannot enter into a concrete calculation until boundary or initial values and other model parameters have been established. It is in setting these values that the user applies the mathematical model to the "special case."

The mathematical model, as the image of an entire class of processes, constructed with the aid of the basic disciplines (chemistry, physics, process engineering, etc.), starts out with only a theoretical or prospective relationship to actual calculations (Fig. 51). Its use in computation depends on the existence of a solution algorithm; in general, a computer program is still needed. The usefulness of a mathematical model is thus determined not only by its consistency with the underlying real object, but also by the bounds set by numerical mathematics and computing capacity set. These limits are still relevant in practice today, because many of the processes found in the chemical industry are highly complicated from the standpoint of mathematical modeling (e.g., emulsion polymerization with its many associated phenomena). Advances in the calculability of mathematical models are essentially governed by advances in computer hardware. An example is the method of finite elements, used for the solution of partial differential equations. Although the fundamental idea was formulated by Courant in the 1940s [271], this publication remained largely unknown, and finite elements were "rediscovered" by engineers in the 1950s. Only then did rapid progress occur in its numerical principles, and these advances accompanied the development of computer technology.

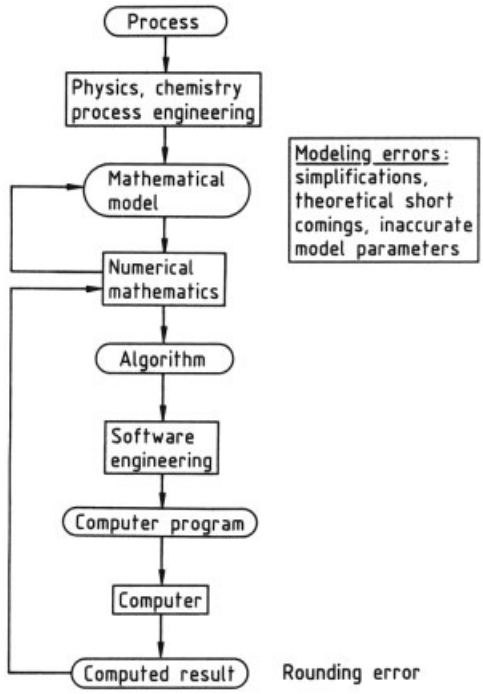

Figure 51. From process to computed result

Another example is adaptive control. The first publications on this method appeared in

the 1950s [272, 273], but the impetus for further study came only when the development of fast process control computers made it practically feasible.

The creation of a mathematical model thus means keeping track of numerical problems, the capacity of the computers to be used, and (especially when writing large programs) software engineering considerations. In particular, the three main sources of error (modeling errors, procedural errors, and rounding errors) must be treated as a compromise to minimize the total error. This topic has also been dealt with in [274, 275].

Mathematical Models: Survey and Structural Properties. This section treats mathematical models that describe the space – time behavior of processes whose model variables and parameters are measured on at least an interval scale (see Section 3.1). The general mathematical description of such processes consists of coupled systems of partial differential equations. The models under consideration are those in which the system is described in terms of model variables classified as inputs u_i, state variables x_i, outputs y_i, and model parameters p_i (see Section 3.1).

In creating and using mathematical models of process apparatus, it is particularly important that the model nearly always consists of two parts, the physicochemical part and the apparatus-specific part (Fig. 52). The physicochemical component of the model includes relationships whose validity is totally independent of the particular apparatus in which the modeled operations take place. This self-consistency of the physical component insures that the model can be applied to a variety of apparatus; this feature is naturally a great advantage in the architecture of large applications software packages (material property databases, etc.). The fact that apparatus and facilities can be simulated with off-the-shelf engineering software, without regard to the wide range of variation in the details, is a consequence of the strong structural similarity of the essential apparatus-specific process equations. All space – time processes taking place in continuous-flow equipment, for example, obey balance equations written for conserved quantities (e.g., energy and chemical species), in which one side always contains terms known from the continuity equation while the other side contains diffusive terms, sources, and sinks, which represent only the pure, non-apparatus-specific physics. As special cases, the continuity terms contain both purely spatial variability (steady continuous process) and purely temporal variability (batch process).

1. Physicochemical part (state relations)

Enthalpy	$J = J(x, T)$
Density	$\varrho = \varrho(x, T)$
Ideal gas law	$pV = vRT$

2. Apparatus-specific part (process equations)

Balance equations $\frac{\delta x}{\delta T} + \mathrm{div}\,(\vec{v}\,x)$

= diffusion term + sources − sinks

"Engineering formulas"

Initial and boundary conditions

Figure 52. Structure of mathematical process models

"Stage models" are often used in simulation. Such models describe apparatus or parts of apparatus in terms of fictitious or real stages that have no spatial variability of the state quantities in a finite volume. The differential operator $\mathrm{div}\left(\vec{v}\cdot x\right)$ is then replaced by appropriate difference operators.

In addition to the balance equations, the typical information includes initial and boundary conditions for the apparatus-specific equations. A number of software products now available make extensive use of these structural properties, both in their design and in the user interface. The conserved quantities are momentum (Navier – Stokes equations), mass (continuity equation), and enthalpy (heat balance). Other conserved quantities, as well as sources and sinks, can generally be programmed by the user.

The divergence term $\mathrm{div}\left(\vec{v}\cdot x\right)$ has uses that go far beyond describing convection in space. The balance equation can be applied to other pseudospatial propagation phenomena without change.

Two examples:

1) Particle-size distribution
   "Space coordinate" = particle diameter
   Velocity v = rate of growth of particles
   Conserved quantity x = number of particles
2) Particle age distribution
   "Space coordinate" = particle age τ

Velocity $v = d\tau/dt = 1$
Conserved quantity x = number of particles

As a concrete example to show how the model equations are split, Figure 53 shows the mathematical model of a distillation column.

1. Physicochemical part (state relations)

Equilibrium relations	$y^* = y(x, T)$
Boiling curve	$p = p(x, T)$
Ideal gas law	$pV = vRT$
Liquid density	$\varrho = \varrho(x, T)$
Enthalpy of liquid	$J = J(x, T)$
Heats of vaporization	$r = r(T)$
Reaction rates	$R = R(c, T)$

2. Apparatus-specific part (process equations)

Material balances	$(n_i \cdot x_i) = \Sigma$ inputs $- \Sigma$ outputs
Heat balances	$W_i = \Sigma$ inputs $- \Sigma$ outputs
Separation effect of trays	$\dfrac{y_i - y_i + 1}{y_i^* - y_{i+1}} = E$
Pressure balance	$\Delta p_i = g_1$ (flow rates, densities, x, T, geometry)
Weir relation	Liquid flow rate = g_2 (liquid volume, geometry)

Figure 53. Mathematical model of distillation columns

For the physicochemical part, equilibrium relations are needed between the mole fractions in gas and liquid, the relation for the boiling point line (which interelates the pressure with the molar fraction of liquid and the temperature), the ideal gas law, and relations for the liquid density, liquid enthalpy, heat of vaporization, and (if appropriate) reaction rates. All these are purely physical state equations. Given this list of relationships, only an expert might suppose that someone is creating a mathematical model for a distillation column. The apparatus component of the model is required to reveal which apparatus is to be modeled. In the case of a distillation column, this would comprise the material and heat balances for condenser, trays, bottom, and vaporizer; the separating effect of the trays (giving the departure from ideal equilibrium, described by, for example, the Murphree efficiency E); the pressure balances for determining the pressure drop across individual trays (these use basic geometrical data such as perforation diameter, tray thickness, and active tray surface area); and finally a weir relation to give the downflowing liquid flow rate (again based on geometrical data such as weir height).

Spatial discretization with stage models is a special case of a widespread practice (e.g., difference methods, method of finite domains, Galerkin method, with the important special case of finite elements), which meets practical requirements because partial differential equations can generally be solved only by discretizing to finite-dimensional substitute models. These are represented as systems of differential algebraic equations (DAE systems) of the form

$$A(x, z, u)\,\dot{x} = f(x, z, u) \tag{3.3.1}$$
$$0 = g(x, z, u) \tag{3.3.2}$$

The output equation

$$y = C(x, u) \tag{3.3.3}$$

is also needed.

The special structure of the left-hand side of (3.3.1), where the matrix A depends on x, z, and u, expresses the fact that nearly all process models of practical relevance are linear in the time derivatives $\dot{x}$ *of the state variables* x. For the numerical solution of DAE systems, see [276, 277].

In Section 3.1, the state vector x with minimum number n of components was introduced to describe the system, where n is equal to the number of degres of freedom in the system. Unfortunately, the mathematical modeling of a system based directly on the physical phenomena often leads to DAE systems (3.3.1), (3.3.2) in which the number of degrees of freedom is smaller than the number of components of vector x (DAE systems with index greater than 1).

For example, in mathematical models of distillation columns, the boiling-line equation couples the pressure with the differential equation variables temperature and mole fractions. If the pressure is assumed to be constant, the number of degrees of freedom is thus reduced.

The numercial handling of such systems with index greater than 1 is problematical because the solution contains time derivatives of the inputs $u(t)$ up to order $q - 1$, where q is the index of the DAE system [278–280]. Thus DAE systems with index greater than 1 are mixed problems involving both integration and differentiation.

Example: The equation $\dot{x}_1 = x_0 - x_1$ (stirred tank with mean residence time $\tau = 1$) with $x_1 = u(t)$ [with $u(t)$ a specified function] has solution $x_0 = u(t) + \dot{u}(t)$. Here the number of degrees of freedom is actually zero; the index is 2.

Design problems such as this example, where certain state or initial values are specified and certain inputs are to be calculated, generally have a high index.

To transform a DAE system with index $q \geq 2$ to an equivalent $q = 1$ system, index reduction methods have been devised [278, 280]. For large models, however, these are often time consuming since the transformations must be performed with closed formulas, and an expert system that will let the computer do the transformations is not available.

If the system is discretized not only in space but also in time, by considering the system quantities only at specified (generally equally spaced) times $t_i = t_0 + i\Delta t$ $(i = 0, 1, 2, \ldots)$, the time-continuous models (3.3.1) – (3.3.3) become time-discrete models of the form

$$A(x_i, z_i, u_i)\, x_{i+1} = F(x_i, z_i, u_i) \tag{3.3.4}$$

$$0 = g(x_i, z_i, u_i) \tag{3.3.5}$$

$$y_i = C(x_i, u_i) \tag{3.3.6}$$

Time-discrete systems offer a number of advantages, including simpler numerical solution. Because of their inaccuracy (especially when the system behaves in a strongly nonlinear manner), their use is generally recommended only for on-line adaptation of the system parameters. The structural properties in most cases are invariant to the change from time-continuous to time-discrete models, but there are changes in, for example, the concepts of observability and controllability [260].

Up to now we have discussed dynamic models that include both spatial and temporal variations of the system. These models are naturally of interest as aids where the time dependence is an essential feature of the problem. These situations include studies of system behavior by means of:

1) Stability analysis
2) Step response of the system
3) Controller design and tuning of controller parameters; simulation of genuinely dynamic processes such as batch operations
4) Startup and shutdown strategies in continuous equipment
5) Failure analyses

Finally, dynamic models are increasingly used in identifying optimal process control (e.g., in time-optimized control of batch processes when safety and product quality are competing considerations). Besides these advanced simulation techniques, dynamic process models are used on-line in the form of observers, such as the Luenberger observer [265, 266] or the Kalman – Bucy filter [281, 282].

As discussed in Section 3.1, the notions of steady-state and quasi-steady-state process models come directly from Equations (3.3.1) – (3.3.3) if $\dot{x}$ is set equal to zero. Naturally, steady-state process models find use chiefly with steady continuous processes and in connection with problems such as

1) The design and optimization of plants and apparatus
2) The analysis of plant characteristics

The state variables can in principle be eliminated from these models, so that a direct relationship appears between the outputs y and the inputs u:

$$y = h(u) \tag{3.3.7}$$

Equation (3.3.7) is the simplest case of a transfer model, in which the value of the output vector y is obtained directly from that of the input vector u by means of a calculation specification. Such a simple link between y and u cannot in principle hold in dynamic process models.

For example, the transfer function between the concentrations u (input) and y (output) for an ideally mixed stirred tank with no reaction is given by

$$y(t) = \frac{1}{\theta} \int_{-\infty}^{t} \mathrm{e}^{-\frac{t-\tau}{\theta}} \cdot u(\tau)\, \mathrm{d}\tau \tag{3.3.8}$$

where θ is the mean residence time.

This formula represents the "steady-state" solution of the dynamic model

$$\dot{x} = \frac{1}{\theta}(u - x) \tag{3.3.9}$$

$$y = x \tag{3.3.10}$$

It is important, however, that in Equation (3.3.8) the value of y at time t does not, as in (3.3.7), depend solely on the value of the input u at time t, but on the time variation of u throughout the past.

Transfer models for dynamic systems are functionals of the form

$$y(t) = h_1\left(t, u(\tau)_{\tau \leq t}\right) + h_2(t, u(t)) \quad (3.3.11)$$

The second term $h_2(t, u(t))$ in Equation (3.3.11) takes into account the possibility that the effect on y of time variations in u may be present not just in smoothed form, as in h_1 (discontinuous functions in u become continuous functions y; continuous u become continuously differentiable y), but in unsmoothed form. Then h_2 describes the "feedthrough" of u to y.

The importance of transfer-function models (Eq. 3.3.7 and 3.3.11) is that they make it possible to examine how the inputs u affect the outputs y without knowing the internal structure of the system (represented by the state variables x_i and their dependences on u and on one another). This may be the only approach in, for example, complex systems (e.g., dynamics of drugs in human beings and animals) or in design calculations where only the input/output behavior of the apparatus is known at first and the detailed engineering has not yet been done (e.g., pre-simulations based on linear quantitative balances).

Transfer-function models thus belong to a class of models whose structure is only slightly or not at all influenced by the "nature" of the system being modeled. These purely mathematical (statistical or regression) models therefore offer great freedom in the choice of model structure. The model parameters must always be obtained by fitting to measurements made in the underlying real system. Because such regression models are based on measurements (see Section 3.2) and no attempt is made to derive a predictive theory from them, the model cannot be applied or extrapolated beyond the range of the measured values and the specific apparatus on which the measurements were performed, in contrast to theoretical models. While pure regression models cannot be employed in the design phase, they have an important application in on-line simulation (see Section 11.3). Measurements on the running process make it possible to treat and adjust the model parameters as time-variant quantities, so this accurate knowledge of the model parameters permits the use of regression models with simple structure. When these "adaptive" process models are used, the control parameters can also be modified on-line to fit the current process situation. This technique is called "adaptive control."

Particularly important here are linear process models. In the state – space representation, they have the form

$$\dot{x} = A(t) \cdot x + \mathrm{B}(t) \cdot u(t) \quad (3.3.12)$$

$$y = C(t) \cdot x + \mathrm{D}(t) \cdot u(t) \quad (3.3.13)$$

The system parameters are thus the totality of components in the matrices A (system matrix), B (input or control matrix), C (observation matrix), and D (pass-through matrix), which completely characterize the model of Equations (3.3.12) and (3.3.13). For this model, the block-diagram visualization customary in publications on control engineering takes the form shown in Figure 54.

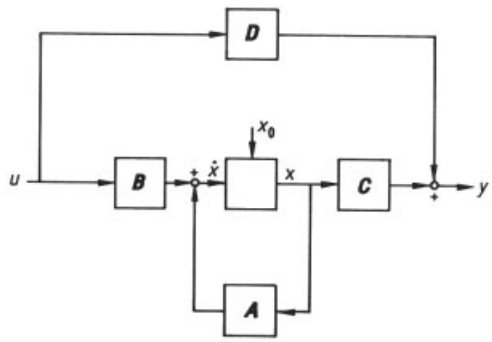

Figure 54. Block diagram of a linear process model

The input x_0 to the integrator block symbolizes the effect of an initial condition $x(t_0) = x_0$ on the solution behavior. The transfer-function model associated with Equations (3.3.12) and (3.3.13) has the form

$$y(t) = \int_{-\infty}^{t} K(t,\tau) \cdot u(\tau)\, \mathrm{d}\tau + D(t) \cdot u(t) \quad (3.3.14)$$

where the matrix $K(t, \tau)$ depends on the time variables t and τ. Equation (3.3.14) gives a mapping $y = F(u)$ from the set of input functions u into the set of outputs y; this mapping satisfies the superposition principle

$$F(u_1 + u_2) = F(u_1) + F(u_2) \quad (3.3.15)$$

(image of sum = sum of images) and

$$F(\alpha \cdot \mathrm{u}) = \alpha \cdot F(u) \quad (3.3.16)$$

(image of α times $u = \alpha$ times image of u).

Linear Process Models. The linear process models introduced above require some more discussion, since they are still very important in both systems analysis and the design and on-line modeling of control systems (despite advances in the calculation of nonlinear systems). Their importance undoubtedly stems from the fact that their simple structure can be completely understood from fundamentals and that many structural properties of nonlinear process models are conserved in linearization (approximation of nonlinear models by linear ones).

What is more, the calculation of linear process models involves far fewer problems than that of nonlinear models. The use of linear models should always be considered when it is certain that the process trajectories in state space do not depart from the region in which linearization is valid. An example is control to maintain a stationary operating point by using an on-line linearized process model.

Special importance attaches to the time-invariant linear models

$$\dot{x} = A \cdot x + B \cdot u\,(t) \tag{3.3.17}$$

$$y = C \cdot x + D \cdot u\,(t) \tag{3.3.18}$$

which are set by the constant (time-invariant) matrices *A, B, C,* and *D*.

As an example of linearization about a stationary operating point, let us analyze the reaction A → B taking place in a stirred tank, using the heat balance. The nonlinear process model is described by the equations

$$\dot{a} = \frac{1}{\theta} \cdot a_{\mathrm{E}} - \frac{1}{\theta} \cdot a - S \cdot e^{-T_{\mathrm{A}}/T} \cdot a \tag{3.3.19}$$

$$\dot{T} = \alpha\,(T_{\mathrm{c}} - T) + \beta \cdot S \cdot e^{-T_{\mathrm{n}}/T} \cdot a + \gamma \cdot (T_{\mathrm{E}} - T) \tag{3.3.20}$$

where a and a_{E} are the concentrations of A in the tank and in the inlet stream; T is the reaction temperature, T_{E} the inlet temperature, and T_{C} the coolant temperature; θ is the mean residence time and S the collision factor; T_{A} is the activation temperature of the reaction; and α, β, and γ are further model parameters.

Consider given values of the inputs $\bar{a}_E$, $\bar{T}_{\mathrm{E}}$, and $\bar{T}_{\mathrm{K}}$, and the associated stationary values $\bar{a}$ *and* $\bar{T}$ of the state variable. If the deviations are defined as $a = \bar{a} + x_1$, $T = \bar{T} + x_2$, $a_{\mathrm{E}} = \bar{a}_E + u_1$, $\bar{T}_{\mathrm{E}} = T_{\mathrm{E}} + u_2$, and $T_{\mathrm{K}} = \bar{T}_{\mathrm{K}} + u_3$. Linearization then yields the relations

$$\dot{x}_1 = \frac{1}{\theta} u_1 - \frac{1}{\theta} x_1 - S \cdot \mathrm{e}^{-T_{\mathrm{A}}/T} \cdot x_1 - \frac{S \cdot T_{\mathrm{A}}}{\bar{T}^2} \mathrm{e}^{-T_{\mathrm{A}}/\bar{T}} \cdot \bar{a} \cdot x_2 \tag{3.3.21}$$

$$\dot{x}_2 = \alpha \cdot u_3 - \alpha \cdot x_2 + \beta \cdot S \cdot \mathrm{e}^{-T_{\mathrm{A}}/\bar{T}} \cdot x_1 + \frac{\beta \cdot S \cdot T_{\mathrm{A}}}{\bar{T}_2^2} \mathrm{e}^{-T_{\mathrm{A}}/\bar{T}} \cdot \bar{a} \cdot x_2 + \gamma \cdot u_2 - \gamma \cdot x_2 \tag{3.3.22}$$

If the temperature is known as an output (measurement) and we accordingly set $y = x_2$, the linearized system has the system matrix

$$A = \begin{bmatrix} -\frac{1}{\theta} - S \cdot \mathrm{e}^{-T_{\mathrm{A}}/\bar{T}} - \frac{S \cdot T_{\mathrm{A}}}{T^2} \mathrm{e}^{-T_{\mathrm{n}}/\bar{T}} \cdot \bar{a} \\ \beta S e^{-T_{\mathrm{A}}/\bar{T}} - \alpha - \gamma + \frac{\beta \cdot S \cdot T_{\mathrm{A}}}{T_2} \mathrm{e}^{-T_A/\bar{T}} \cdot \bar{a} \end{bmatrix} \tag{3.3.23}$$

the input matrix

$$B = \begin{bmatrix} \frac{1}{\theta} & 0 & 0 \\ 0 & \gamma & \alpha \end{bmatrix} \tag{3.3.24}$$

observation matrix $C = (0\ 1)$, and feed-through matrix $D = (0\ 0\ 0)$.

The stability (also of nonlinear systems) at a stationary state x_{s} can be investigated with the system matrix A obtained by linearization about this state. Thus, for example, the system is asymptotically stable at x_{s} if and only if all the eigenvalues λ_{i} of A have negative real parts.

In the example described, the eigenvalues of the system matrix A have negative real parts if the conditions

$$\psi = \frac{1}{\theta} + \alpha + \gamma + S \left(1 - \beta \frac{T_{\mathrm{A}}}{\bar{T}^2} \bar{a}\right) \mathrm{e}^{-T_{\mathrm{A}}/\bar{T}} > 0 \tag{3.3.25}$$

and

$$\varphi = \alpha + \gamma + S \left([\theta\,(\alpha + \gamma)] - p \frac{T_{\mathrm{A}}}{\bar{T}^2} \bar{a}\right) \mathrm{e}^{-T/\bar{T}} > 0 \tag{3.3.26}$$

are fulfilled. In this case the stationariness $\bar{a}$, $\bar{T}$ is asymptotic. Criteria of this type for the stability of linear systems can be generalized to arbitrary n-row system matrices A (e.g., the Hurwitz criterion) [257].

The transfer-function model associated with a time-invariant linear model is

$$y\,(t) = \int_{-\infty}^{t} K\,(t - \tau) \cdot u\,(\tau)\,\mathrm{d}\tau + D \cdot u\,(t) \tag{3.3.27}$$

The matrix $K(t)$ in Equation (3.3.27), known as the impulse response matrix, is given by

$$K(t) = C \cdot \mathrm{e}^{At} \cdot B \tag{3.3.28}$$

(For the definition of the matrix exponential function appearing in this formula, see [260].) If the system has no feed-through (i.e., $D = 0$), and the idealized impulse function $\delta(t)$ (delta function) is taken for $u(t)$, the result is

$$y(t) = \int_{-\infty}^{t} K(t-\tau) \cdot \delta(\tau)\,\mathrm{d}\tau = K(t) \tag{3.3.29}$$

$K(t)$ itself is thus the response when a unit impulse $\delta(t)$ is input to the system.

Because transfer-function models are often obtained from measurements alone, Equation (3.3.28) cannot be used to determine $K(t)$, since the matrices A, B, and C are unknown in these cases.

Example: The residence-time distribution of a continuous-flow apparatus is $w(t)$, and $w(t)\Delta t$ is the probability that a particle entering the apparatus at time $t_0 = 0$ leaves it in the interval $(t, t + \Delta t)$. By the superposition principle, the following relation is then obtained between the inlet flow rate u and the outlet flow rate y of an indicator substance:

$$y(t) \int_{-\infty}^{t} w(t-\tau) \cdot u(\tau)\,\mathrm{d}\tau \tag{3.3.30}$$

which is of the type of Equation (3.3.27) with $K(t) = w(t)$.

Accordingly, $w(t)$ can be found as the response to a unit impulse; this is a widely used method for residence-time distributions.

When time-invariant linear process models are investigated, the use of integral transformations is often very productive. The Laplace transformation, which is particularly important, is discussed only briefly here. This transformation, defined by

$$F(s) = \int_{0}^{\infty} \mathrm{e}^{-s \cdot t} \cdot f(t)\,\mathrm{d}t \tag{3.3.31}$$

assigns to the function $f(t)$, $t \geq 0$, a function $F(s)$ of the independent variable s. Here s can be taken as a complex variable; $F(s)$ is then differentiable (holomorphic) at least in the convergence region of the improper integral in Equation (3.3.31). The convergence region is a half-plane $Re(s) > \sigma$, with boundary $Re(s) = \sigma$ dependent on the function $f(t)$. To make the mapping property of Equation (3.3.31) especially clear, the notation

$$F(s) = L(f(t)) \tag{3.3.32}$$

is also used; $f(t)$ is then called the "original function" and the Laplace transform $F(s)$ is also called the "image function." Of the characteristics of the Laplace transformation, here only three that are particularly important for application to linear process models are mentioned:

1) The transformation is linear. From Equation (3.3.31) it follows that

$$\begin{aligned} &L(a_1 \cdot f_1(t) + a_2 \cdot f_2(t)) \\ &\quad = a_1 \cdot L(f_1(t)) + a_2 \cdot L(f_2(t)) \end{aligned} \tag{3.3.33}$$

2) The application of L to linear models thus preserves linearity.

$$L \cdot \frac{\mathrm{d}f(t)}{\mathrm{d}t} = s \cdot L(f(t)) - f(0) \tag{3.3.34}$$

3) If the convolution integral

$$\int_{0}^{t} f_1(t-\tau) \cdot f_2(\tau)\,\mathrm{d}\tau \tag{3.3.35}$$

is abbreviated as $f_1(t) * f_2(t)$, then

$$L(f_1(t) * f_2(t)) = L(f_1) \cdot L(f_2) \tag{3.3.36}$$

Equations (3.3.34) and (3.3.36), in particular, illustrate the simplifications that result from the use of the Laplace transform. Relatively complicated operations such as differentiation and convolution in the original domain turn into simple algebraic operations.

Computations with the Laplace transform can be schematized as shown in Figure 55. Correspondence tables, showing the original and transformed functions corresponding to a number of elementary functions, are employed. Consider, for example, the linear model

$$\dot{x} = a \cdot \mathrm{x} + b \cdot u \tag{3.3.37}$$

with initial condition $x(0) = x_0$. By using Equation (3.3.34):

$$s \cdot X(s) - x_0 = a \cdot X(s) + b \cdot U(s) \tag{3.3.38}$$

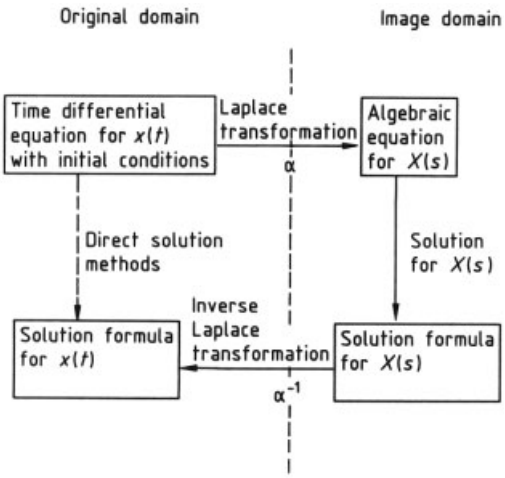

Figure 55. Application of the Laplace transformation

From this the following purely algebraic relation between the transforms X (s) *and* $U(s)$ is obtained

$$X(s) = \frac{x_0}{s-a} + \frac{b \cdot U(s)}{s-a} \tag{3.3.39}$$

Together with the correspondence $L(\mathrm{e}^{\mathrm{at}}) = 1/(s-a)$ and the property (Eq. 3.3.36), transforming (3.3.39) to the original domain directly gives the solution formula

$$x(t) = x_0 \cdot \mathrm{e}^{at} + \int_0^t \mathrm{e}^{a(t-\tau)} \cdot b \cdot u(\tau)\,\mathrm{d}\tau \tag{3.3.40}$$

For further details on the Laplace transformation, the reader should consult the specialist literature [283, 284].

The Laplace transformation can also be applied to vector-valued functions such as those appearing in the general model of Equations (3.3.17) and (3.3.18). Consider the case with no feed-through ($D = 0$). Application of the Laplace transformation gives

$$Y(s) = G(s) \cdot U(s) \tag{3.3.41}$$

which offers a simple link between the transforms $U(s)$ of the input and $Y(s)$ of the output in terms of the transfer matrix

$$G(s) = C \cdot (s \cdot I - A)^{-1}\,\mathrm{B}$$
$$(I = \text{identitymatrix}) \tag{3.3.42}$$

Note that, as usual, we are considering the state that depends only on the temporal behavior of n; that is, $x(0) = 0$ and Equation (3.3.41) holds only for inputs $u(t) = 0$ for $t < 0$. Equation (3.3.41) in the transform domain corresponds to the relation

$$y(t) = \int_0^t K(t-\tau) \cdot u(\tau)\,\mathrm{d}\tau \tag{3.3.43}$$

in the original domain, which follows from Equation (3.3.27) under the stated conditions. Thus the role of the impulse function response matrix $K(t)$ in the original domain is played by the transfer matrix in the transform domain.

A function of great importance in practice is the frequency response $G(j\omega)$, which is obtained by letting s take on imaginary values $j\omega$ (ω real) in Equation (3.3.42). The frequency response satisfies the fundamental relation

$$y(t) = G(j\omega) \cdot \mathrm{e}^{j\omega t} \cdot u_0 \tag{3.3.44}$$

which in the original domain gives a simple description of the response to periodic input functions $u(t) = e^{j\omega t} u_0$. Since the frequency response $G(j\omega)$ contains all the information about the process, an obvious way to extract this information is to measure the response functions $y(t)$ over an entire spectrum of input frequencies ω and use Equation (3.3.44) to determine $G(j\omega)$ from these. This "frequency response method" is also a highly sensitive procedure for stability analysis that does not require a knowledge of the system matrix A and its eigenvalues. The determination of the frequency response from measurements with the aid of periodic input functions also makes it possible to obtain the impulse response $K(t)$ more accurately than with the impulse method, even though the amount of measurement required is greater, since $K(t)$ is calculated from the frequency response $G(j\omega)$ through the inverse Fourier transformation:

$$K(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(j\omega)\,\mathrm{e}^{j\omega t} d\omega \tag{3.3.45}$$

This expression can be inverted in accordance with the Fourier integral formula:

$$G(j\omega) = \int_{-\infty}^{\infty} K(t)\, e^{j\omega t} dt \tag{3.3.46}$$

If the time dependence of the transfer function is replaced by a position dependence and, therefore, the temporal frequencies by spatial frequencies, the analogous operations lead to the contrast transfer function (optical transfer function) known from image transmission and photography. This function verifies the transformation principle described in Section 2.2 as well as the phase model derived from that principle in Section 2.3 [285].

3.4. Modeling

The derivation of a mathematical model for a given purpose involves several problems. The objective governs

1) The type of model (dynamic, time-continuous, time-discrete, stationary, time-variant, etc.)
2) The structure of the model (theoretical, statistical, transfer-function, linear, etc.)

The restrictions mentioned in Section 3.3, regarding the computability of the model and computing time problems, must also be taken into account. It is often not possible to give an a priori answer to the question of which physical processes should be incorporated into the model and which can be neglected. The addition of a further physical phenomenon to the model may render the former numerical method of solving the model equations useless and may lead to a large increase in total computing time. In such cases, attempts are made to use a simplified version of the model. Ultimately, however, the power of the model depends on the comparison of computed results with experimental data. The modeling process often becomes an iterative procedure in which calculation and experimentation alternate. It is not always the case that successive, improved versions of the model, when compared with the experimental data, will converge to the final model. Occasionally, new measurements are needed for deciding whether to accept or reject certain model assumptions.

These problems suggest a generalization of the problem statement: Given some competing models $M_1, M_2, \ldots$, describing a certain system, to find quantitative decision criteria for identifying the "best" model at minimal testing cost. A number of publications on this model discrimination problem have appeared [286–288]; in the well-known Box – Hill method [289], an entropy value is assigned to each model based on its agreement with previous experimental results, and the next test point is chosen so as to maximize the entropy difference between models.

Section 3.3 identified discretization as an essential step in obtaining finite-dimensional equivalent models of the type defined in Equations (3.3.1) – (3.3.3). This implies, in general, the problem of choosing the correct state variables in the vector x. Usually there is a wide range of options, and this transformational freedom should be utilized in mathematical models in order to exhaust certain optimization possibilities. The following main objectives can be stated:

- Models as small as possible
- Short computing times
- Model structures as simple as possible

These goals are interrelated and they also compete with the requirement of a minimum model accuracy. Known examples in which suitable coordinate transformations are used to give simpler model structures include the diagonalization of linear systems, with a diagonal matrix $\boldsymbol{A}$ as system matrix, and the change to dimensionless notation, which indeed often works wonders.

The requirement of small but efficient computational models must often be invoked, especially for on-line use. The model reduction or order reduction technique described here involves the replacement of a mathematical model by a smaller model. There are essentially two approaches

- The use of special physicotechnical properties of the system under consideration yields a reduced model that is often highly efficient
- The second approach consists of purely mathematically formal methods for obtaining reduced process models

In both procedures, a key role is played by the coordinate transformation concept. An attempt is made to position the new coordinate system so that the fewest possible coordinates

(state variables) are needed and so that these reflect the truly relevant degrees of freedom in the system. The technique involves projecting onto the state – space subspace that is essential for the processes taking place in the system. The idea is closely related to the principal-component analysis method used in Section 3.2 to spot relationships in existing data.

A well-known example of such a projection is the expansion in specified functions in the Ritz – Galerkin method. In this technique, eigenfunctions of the operator or so-called functions with compact support are employed in a finite-elements procedure. Other examples are Fourier series expansion; the method of moments used to solve systems of equations arising in, for example polymerization kinetics; and the use of symmetry properties generally. Mention should also be made of similarity theory [290–292], in which quantities relevant to the problem are determined, a dimensional analysis is performed, and the resulting set of dimensionless quantities is smaller in number than the dimensioned starting quantities.

Another example of order reduction is statistical mechanics, in which the ca. 10^{26} state variables that describe the ensemble of molecules in the subject system are projected onto a few new quantities, such as temperature, pressure, and enthalpy.

In the examples examined up to this point, either the transformations were physically motivated or plausible trial functions, series expansions, or the like were applied, and the coefficients in the expansions were taken as the new state variables. A number of methods have been devised [293–301] which attempt to eliminate this element of arbitrariness. In a certain class of transformations, these techniques identify the one for which a certain figure-of-merit functional is minimized.

Figure 56 illustrates the method schematically. Any reducible system has the following property: Every solution trajectory, after quickly passing through an initial branch, merges into a one-dimensional curve. This system therefore has just one relevant degree of freedom, even though it is two-dimensional. The discrepancy between the number of degrees of freedom n and the number of relevant degrees of freedom r is large for most large systems and represents the key to their reducibility. The appropriate coordinate transformation must be found so that z_1 will be the relevant state variable in the new coordinate system (as shown in Fig. 56) while z_2 plays a secondary role and can be neglected. If a certain set M of representative solutions $x(t)$ and associated solutions $x*(t)$ of the reduced system is chosen and the merit function

$$J = \sum_M \int_0^\infty \|x(t) - x^*(t)\|^2 \mathrm{d}t \tag{3.4.1}$$

[299] is examined, then for specified r the reduced system is defined such that J takes on a minimum value $J_{\min}$, whereby

$$e_{\min} = 100 J_{\min} / \sum \int_0^\infty \| \mathrm{x}(\mathrm{t}) \|^2 \mathrm{d}t \tag{3.4.2}$$

is a measure of the mean relative error of the solutions $x*(t)$ of the reduced system. Figure 57 shows $\varepsilon_{\min}$ as a function of the dimension of the reduced system r, where the original mathematical model was that of a distillation column with 33 trays and three components. The number n of differential equations is 68. The set of all response functions to step inputs was selected as the representative ensemble of solutions. The error $\varepsilon_{\min}$ decreases with increasing r, becoming zero for $r = 68$, where the unreduced original model is reached. Even for r as small as 4 the error is only ca. 0.5 %. Thus the problem requires a reduced model with only four differential equations instead of 68.

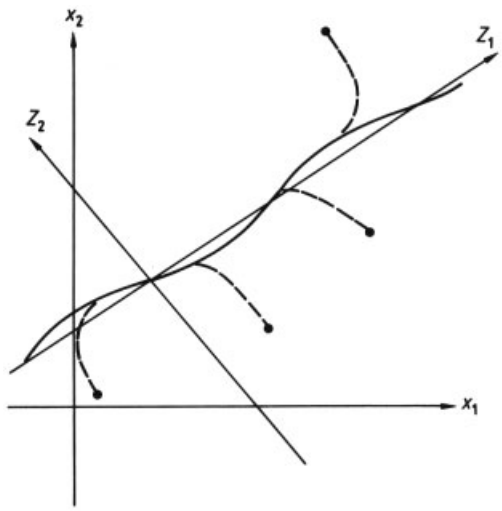

Figure 56. Schematic diagram of model reduction

r	ε_{min}, %
4	0.532
5	0.337
6	0.195
7	0.124
8	0.0730
9	0.0527
.	.
.	.
.	.
68	0

r = number of degrees of freedom of reduced system

ε_{min} = mean relative error of solutions of reduced system

Figure 57. Model reduction for a distillation tower

3.5. Management and Utilization of Information

Introduction. As early as 1986, the Association of Large Power Plant Operators (Verband der Grosskraftwerksbetreiber) carried out a thorough discussion of changes in process control engineering and the demands imposed on the management of engineering documentation [302].

With increasing complexity of plants and processes, the number of monitoring and control devices has increased out of proportion. The main reason has been the growing level of process automation. Rationalization has had little to do with this development; a far more important point is that stringent requirements on product quality not only call for increased information about the process but also demand more powerful process modeling aimed at better reproducibility and direct control of product quality. In addition, a keener awareness of plant safety has led to extensive safety analyses, which in turn have given rise to practices that frequently involve monitoring and control equipment.

These growing demands can easily bring about a documentation conflict between designers and builders of process plants, on the one hand, and operators and maintenance personnel, on the other. The management of data on control devices must enable the use of the documentation right through from concept, through design, execution planning, construction, and operation of the plant. The benefits of such an information

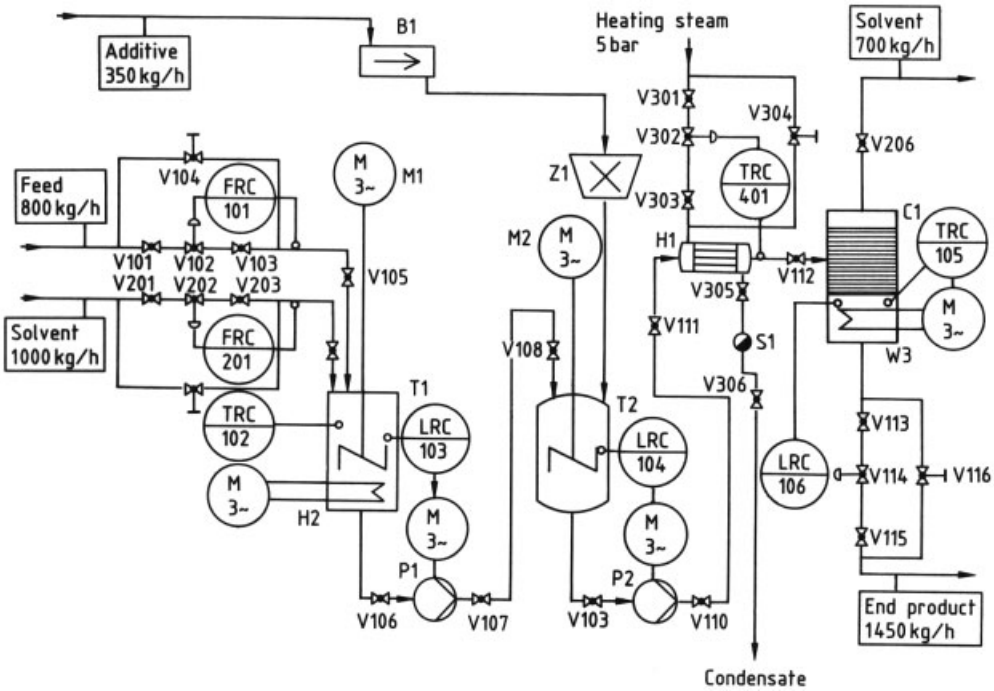

Figure 58. Example of a piping and instrumentation diagram (DIN 28 004, Part 1) [303]
B = belt conveyor; C = column; H = heat exchanger; M = motor; P = pump; S = separator; T = vessel, tank; V = valves; Z = size reduction; FRC = flow registration control; LRC = level registration control; TRC = temperature registration control

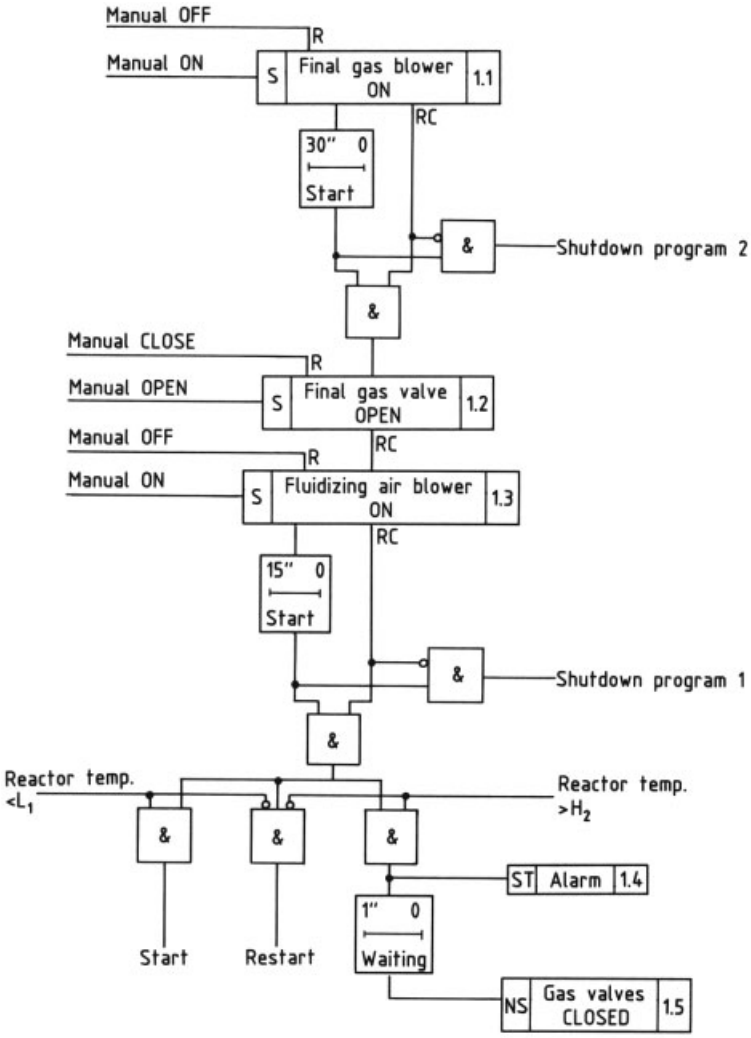

Figure 59. Example of a control system function diagram (DIN 40 719, Part 6)

system are not limited to rationalization; consistent documentation also simplifies the exchange of experience between operators and builders of plants.

Many discussions have contributed to the new edition of DIN 40 719, Part 2, ("Circuit Documentation: Labeling of Apparatus, Signals, and Documents"). The basic ideas can be extended in large part to the documentation of plant monitoring and control systems.

Forms of Organization. The classical organization of control engineering documents is functional (see Chap. 2). Documents are created in order to accomplish a certain task. The instrumentation of "something" is set forth in, for example, the piping and instrumentation diagram (Fig. 58). "Any" control system can be described by functional diagrams (Fig. 59) and schematic diagrams (Fig. 60) or program flowcharts. Documents for the installation of "any" process control device include schematics as well as terminal diagrams.

This type of documentation does fulfill its intended purpose, but it leaves the user virtually unable to form a comprehensive picture of the properties of the control system. Indeed, the user can hardly get an idea of the system from any other viewpoint than that of documents prepared in the past (see Section 2.2). As a consequence,

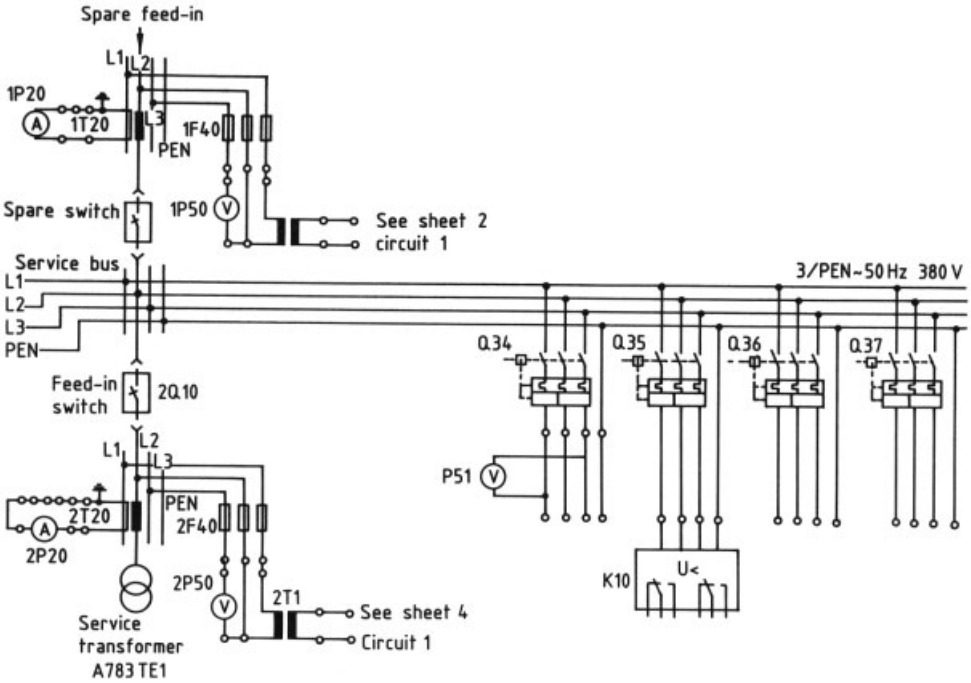

Figure 60. Example of an electrical schematic (DIN 40 719, Part 3)

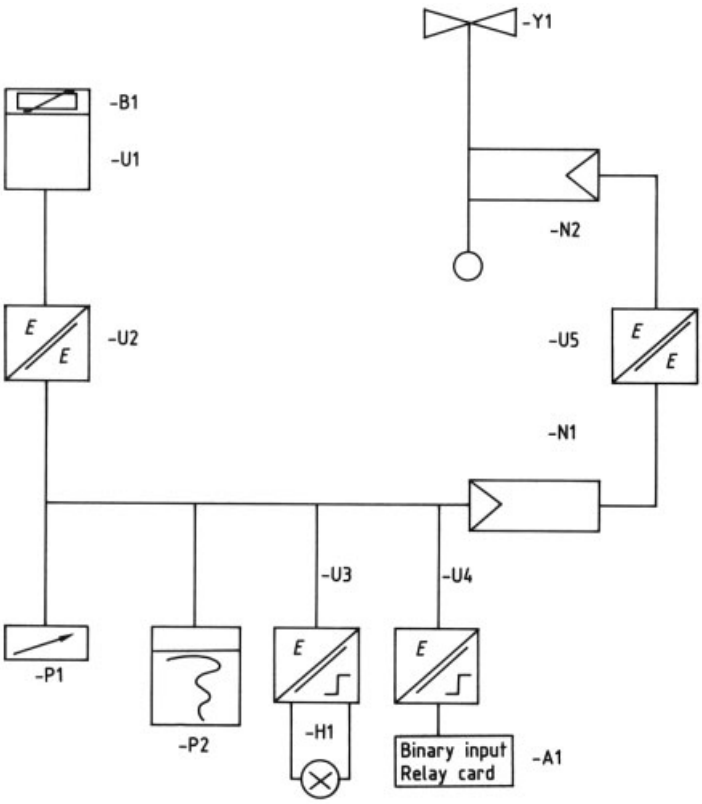

Figure 61. A process control device (control station) from the design viewpoint: station (loop) diagram

No.	Control station KACP	Functions checked	Reason for check	Doc.reqd.?	Reqd. condition
1	V105 BA20 P2	I S + A +	Plant safety	Y	Shut down

Object-related instructions:
Written permission, station setup data sheet, station diagram required

Inspection: Check pressure measurement with I S + A + — Report

1. Perform visual inspection

   Are the elements of the process control station fully labeled, in good mechanical condition, and clean? — OK / OK after repair / Not OK

2. Connect mA test generator in place of transducer.

3. Check measuring range of readout at three points; note actual values.

Nominal values:	Test generator	Readout	Tolerance	Actual values
First value	4.0 mA	0.2 bar	± 0.02 bar	_____ bar
Second value	12.0 mA	0.6 bar	± 0.02 bar	_____ bar
Third value	20.0 mA	1.0 bar	± 0.02 bar	_____ bar

   Are the actual readout values within tolerance? — OK / OK after repair / Not OK

4. Check maximum limit indication.

   Limit: 0.90 bar — Tolerance: ± 0.02 bar

   When limit is exceeded, following indications must occur:

   Does visual indication "P02 max" occur in A 639 W 1 panel 7? — OK / OK after repair / Not OK

   Does acoustic indication occur in A 639 W 1? — OK / OK after repair / Not OK

5. Check maximum limit response.

   Limit: 0.95 bar — Tolerance: ± 0.02 bar

   When limit is exceeded, following responses must occur:

   Does valve P2 – Y1 open? — OK / OK after repair / Not OK

   Does pump P27 shut off? — OK / OK after repair / Not OK

6. Restore to initial condition.

Remarks: _____

Date: _____ Inspector's name: _____ ☐ Not inspected

Figure 62. A process control device (control station) from the maintenance viewpoint: partial job plan

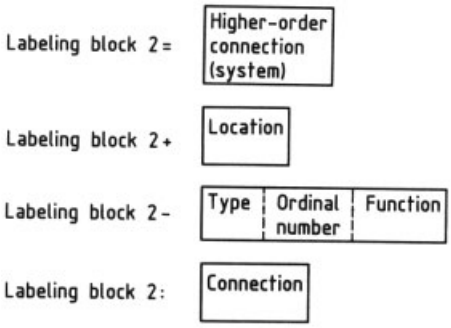

Figure 63. Devices (DIN 40 719, Part 2)

whenever a new question arises, it is often necessary to create new supplementary documents even though these may contain scarcely any new information.

One remedy for this documentation bottleneck is object-oriented management of data on monitoring and control devices. Documents for a given task are then representations of object-related information, that is, projections from certain viewpoints (Figs. 61 and 62). Implementation of this concept requires, above all, the devising of a consistent data model for hardware

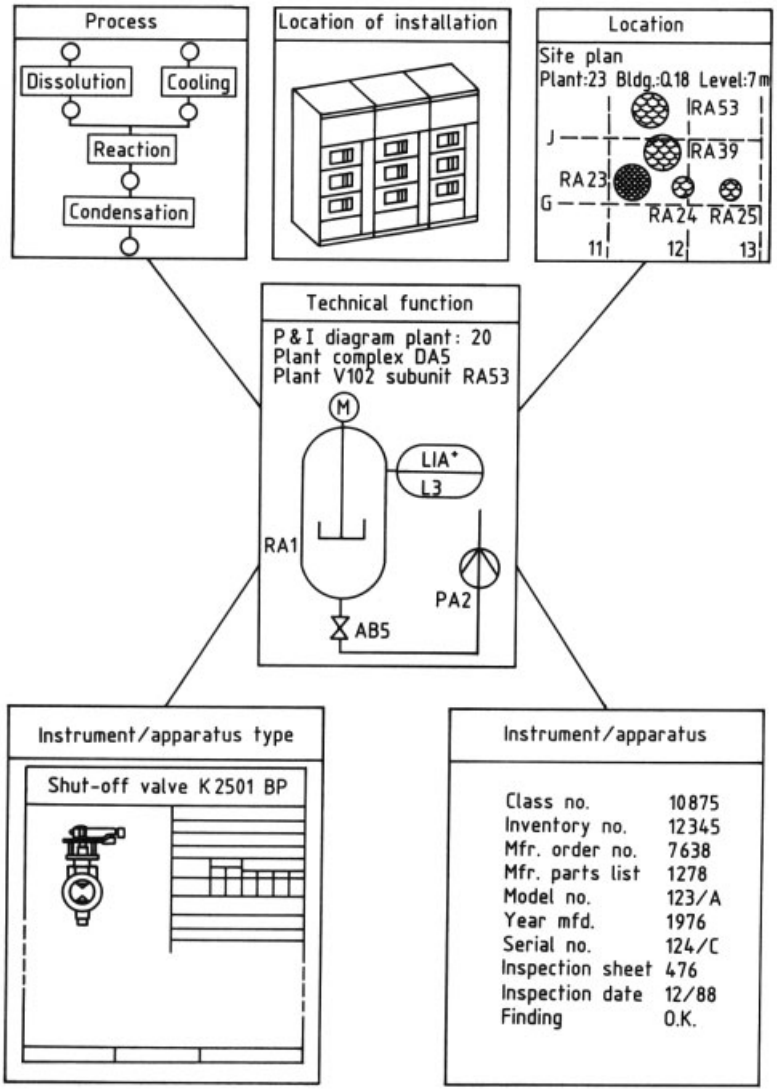

Figure 64. Information blocks associated with process plants

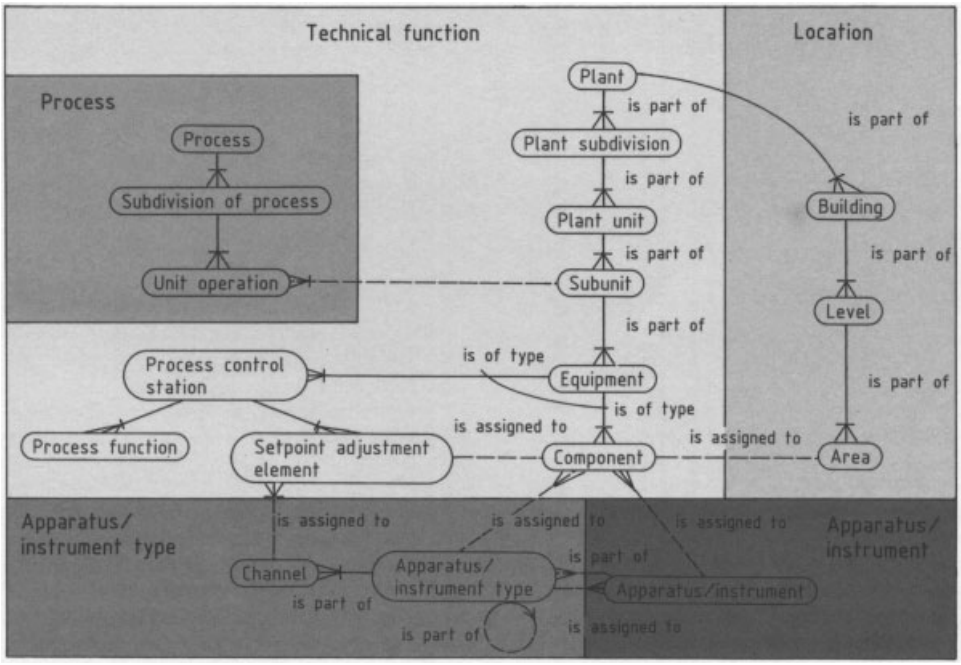

Figure 65. Simplified data model for information blocks relating to process plants

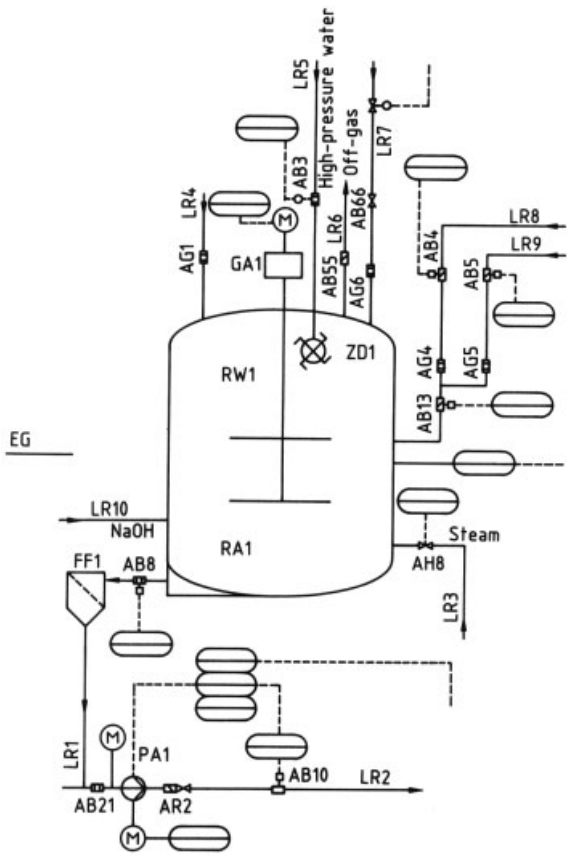

Figure 66. Piping and instrumentation diagram of a plant subunit

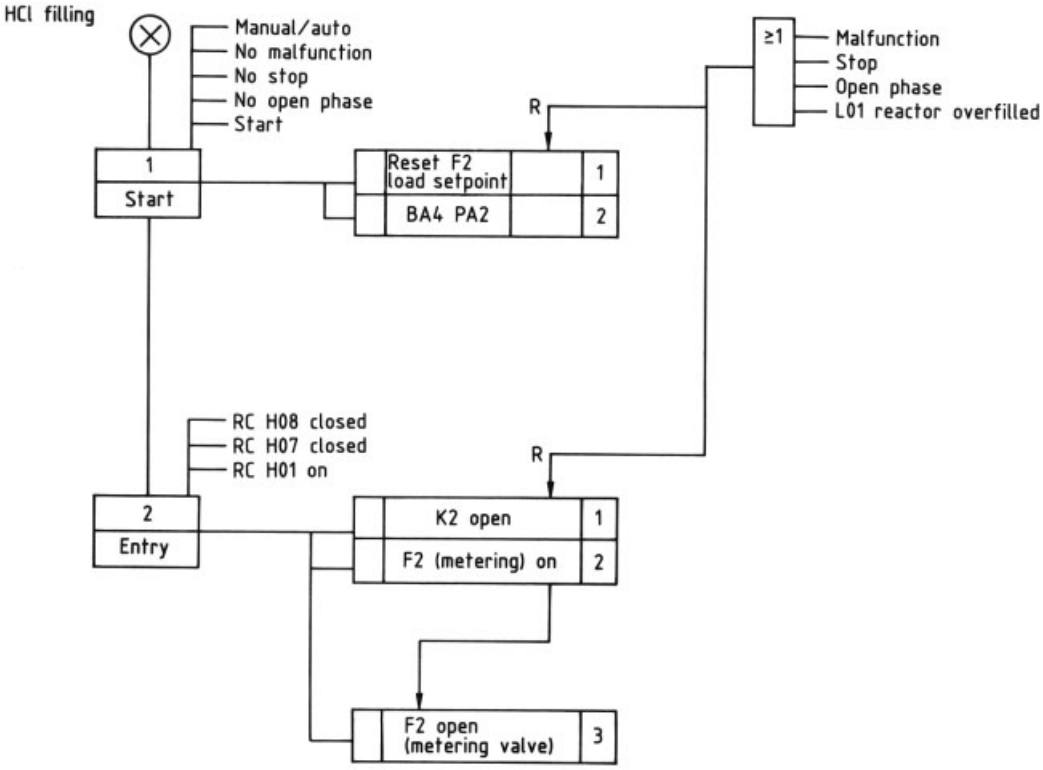

Figure 67. Function chart of a plant subunit

objects (see Section 2.2). It has proved desirable not to model the process control system in isolation but to see it as a component of the process plant.

The initial step must be to understand which data blocks are the fundamental ones. DIN 40 719, Part 2 defines the most important data blocks for electrical apparatus (Fig. 63).

This is not adequate for process plants, since it does not include the purpose of the facility: the process. Incorporating the process into the information structure becomes more important for process control engineering, the more use is made of higher process control functions such as recipe procedures and process models for process control (see Section 4.5).

It has proved desirable to separate detailed description of the engineering properties of a single instrument (substance) from information about its technical function (see also Section 2.3). This makes it possible to manage hardware improvements by replacing devices in the plant with others that perform the same function. All these aspects lead to an information structure for process plants and their process control systems as shown in Figure 64.

When an object-oriented model is created, these information blocks must be described interms of interlinked individual objects. Figure 65 presents a simplified example in which the "entity-relationship model" described in Chapter 2 was used.

When such a model has been set up, it is possible to file all information by object. For example, data on a process control station is filed only under that station. Data on a higher-order plant subunit are filed only under that subunit. They are automatically relevant to the lower-order control stations. Data on an apparatus or instrument type are filed only under that apparatus or instrument type. They are automatically relevant to the associated process control stations.

Application. After such a data model has been created (provided it is complete), documents can be generated at any time and tailored

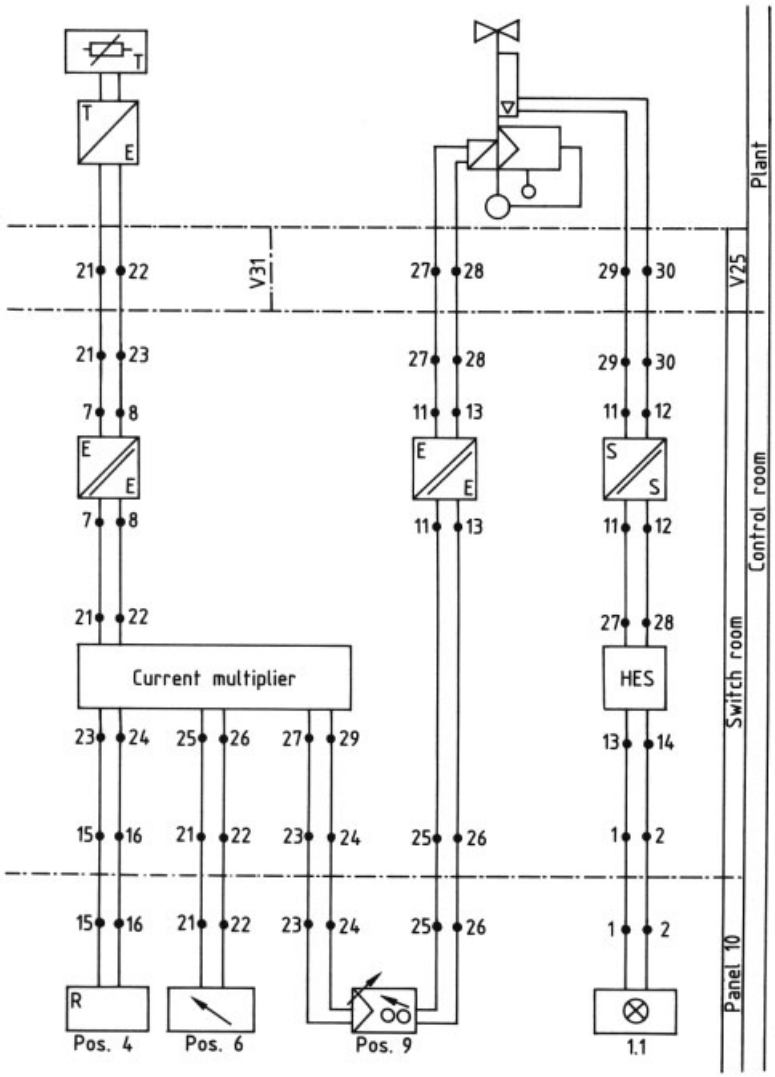

Figure 68. Station (loop) diagram

to any objective. It is at least equally important that given information is managed by object, not in various documents. A computerized revision service can then be set up easily, so that information retrieved at any time will be internally consistent.

Flowsheets are now prepared for a plant (plant unit) or subunit in accordance with this data model (Fig. 66).

A functional diagram may refer to a certain subunit (Fig. 67) or to a certain control station (it is then known as a station or loop diagram; Fig. 68).

This object-oriented management of engineering data makes it possible to work with the same documentation at every stage of plant activity and to update the documentation step by step. A glance at the phases of plant activity (Fig. 69) shows the feasible and necessary level of detail in the documentation at each stage (see also Chap. 7).

Structuring of the processes in terms of unit operations (DIN 28 004) is possible and necessary even at the basic research stage. Plants can already be subdivided into units as documented in the basic flowsheet (Fig. 70). The assignment

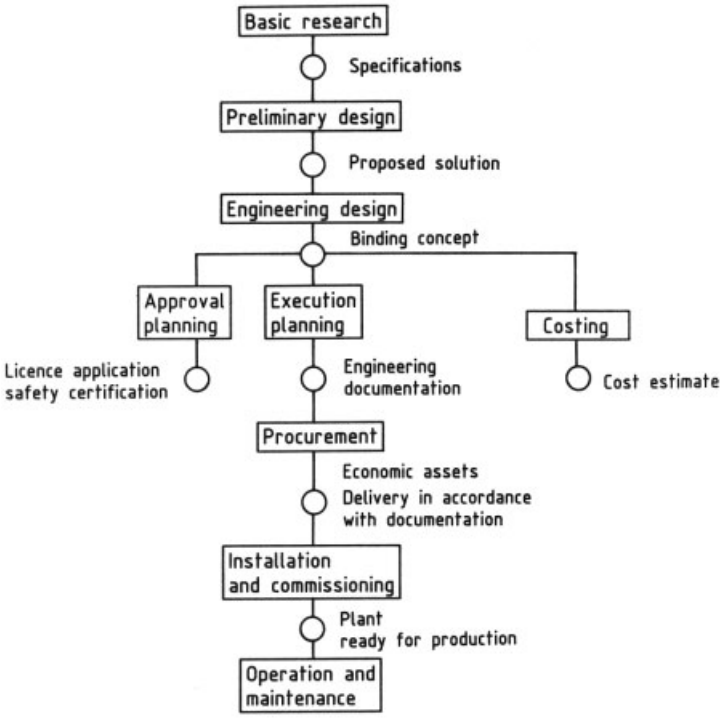

Figure 69. Phases of plant activity

of unit operations to plant subunits, which is so important for higher-level automation tasks, can take place at this stage.

The phase model derived from experience in software engineering (discussed in detail in Section 2.3) is suitable for this purpose.

Quality assurance, in particular, calls for a systematic representation of product and process requirement profiles, since the needed data acquisition measures as well as sensor and actor systems must be established from this standpoint (Fig. 33). In this connection, the reader is referred to Japanese Industrial Standard JIS Z 8206-1982, "Graphical Symbols for Process Charts." This goes beyond DIN 28 004 by specifying which quantity and quality checks must be performed on product streams. More recent practice is to generate process control flowsheets and checklists as tools for quality assurance; these can also serve as evidence that the quality assurance system itself (Fig. 71) conforms to the standard.

The most important process control equipment is selected in the engineering design stage. Documentation relating to each control device, such as a control station, can be created here.

Of special importance in the approval planning (regulatory engineering) phase is the safety analysis, which establishes particular safety objectives and practices. It is important that all information used in this analysis refer to definite, documented engineering hardware (Fig. 72), since only then can the engineering of protective measures be assessed before commissioning and during plant operation.

At the stage of execution planning, all engineering details are fixed. The number of implementation options is unlimited in principle, but in any real enterprise the factors of economics and practicability call for the selection of procedures as well as instruments and systems with which, or from which, the requisite functions can be developed. These choices then lead to design and installation standards, but also to equipment and system standards. As discussed in Chapter 2 and at the beginning of this section, these decisions are also documented in the object-oriented framework (Fig. 73).

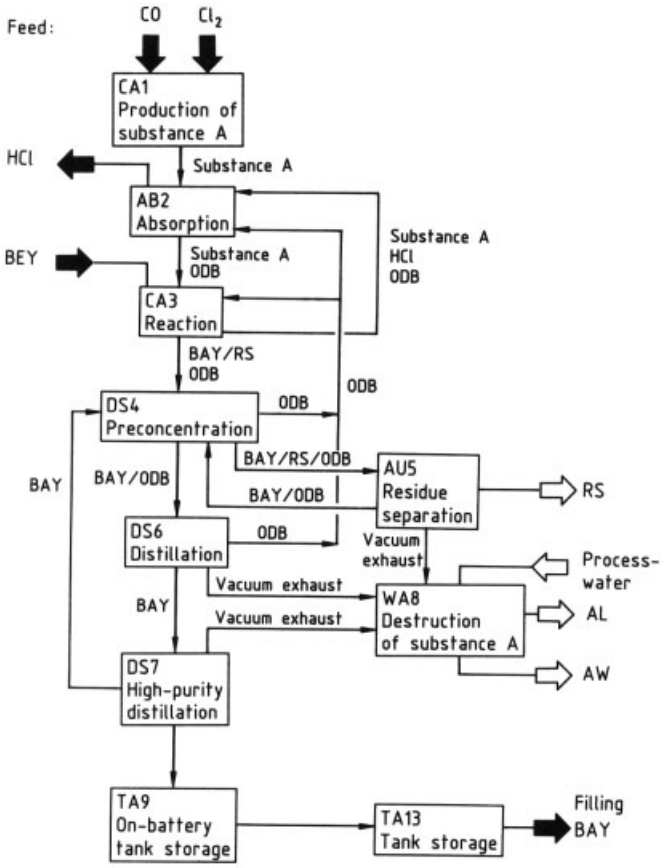

Figure 70. Basic flowsheet of a plant

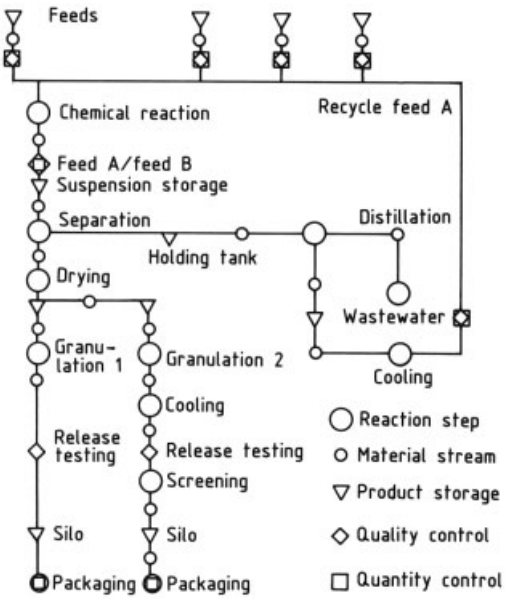

Figure 71. Process monitoring flowsheet of a vulcanization accelerator (cf. JIS Z 8206)

Examples of activities during execution planning are the preparation of control station diagrams for process control devices and the specification of requirements on engineering properties. This information is filed and managed in the process control station. Realizable hardware units, for example, are handled as instrument types derived from the instrument standard, while the functional elements are assigned to control devices (Fig. 74). This has the great advantage that a distinction can be made between requirement and implementation and that, when instruments differing in type are interchanged, only the assignment must be changed. Manual updating of engineering details becomes unnecessary. Similarly, the basic functions of the process monitoring and control systems are assigned to the control devices. In addition to a more rational configuration, this offers the

Safety analysis
Hazard discussion Plant: _____ Date: 10/27/88 V ____ -A ____ -G Version: 0 Page: _____

Facility identification (Internal Standard 9060)

Plant	Plant subdivision	Plant unit	Subunit	Bldg.	Piping and instrumentation diagram	Date
2 2	P P G 6	V 6 9	R A 9 1			

VA No. : _____
AT No. : _____

Process section: _____
Plant component: _____

Deviation from nominal condition:

DEV.NO.	Description
1	Flow rate (feed) too low

Possible causes

Cause No.	Description
1	Pump malfunction (PA2 or PA3)
2	Fittings closed (AB7 or AB12 or AH17)
3	Flow ratio control defective (F2 and F3)

Figure 72. Example of a form used for plant safety analysis
The information from this form must be correlated with the corresponding device (shaded blocks)

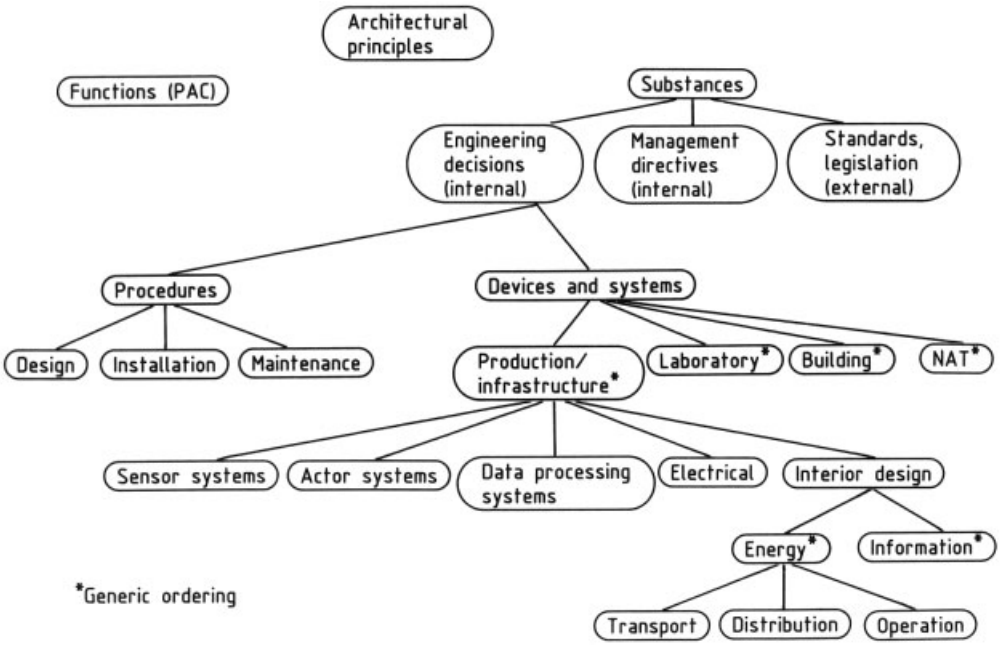

Figure 73. Structure of decisions relating to process control stations

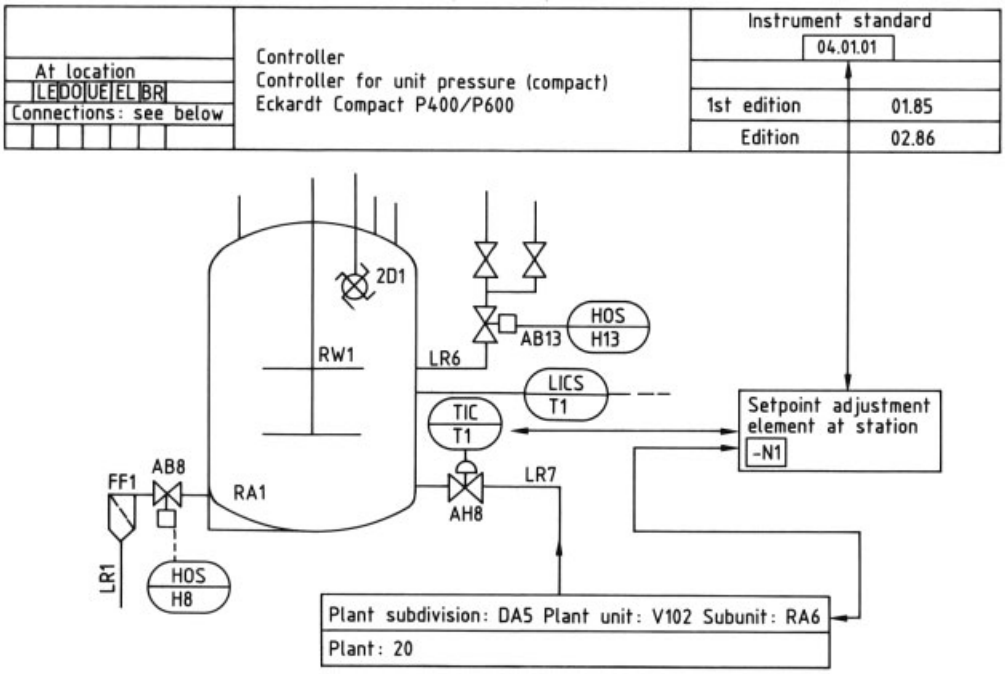

Figure 74. Assignment of instrument types to technical function for the example of a process control station

FS-IT Maintenance

Plant	Plant subdiv.	Plant unit	Subunit
22	FSDL	V40	WS33

Inspection plan
Job specification
Mechanical

Job number 15	Group number 432	As of: 1.89	AKZ 622001	Area: Dralon	Operation: Spinning
Building: E4	Eng.-No.: 075350	Order no.:	Test no.: 3	Registration no. 24	Location no.: 108
Planning date: 15.04.89	Function group: 432	Inconvenience bonus:	Carried out by:		

Type of work:

Component	Operations:
PA 1	Replace GLRD, adjust coupling
PA 4	Check tightness of motor can, adjust coupling
RM 1	Remove NaOH feed and clean nozzles
VE 1	Inspect/replace V-belts
F8 1	Clean mist collector
WS 31	Use pressure cleaner to remove biological and other deposits; perform visual leak check

Figure 75. Example of a maintenance and inspection plan for a component
Each instruction relates to the functioning of one device

advantage of greater reliability with regard to the consistency of engineering data (see Section 4.3).

Similarly, individual devices can be assigned by technical function if necessary. This would be the case, for example, if custom instruments were used instead of off-the-shelf devices, if the instruments were subject to individual inspection, or ifmaintenance cost records were to be kept for an individual unit.

Such assignments are important items of information for installation and maintenance. When necessary, they must be indicated locally. Maintenance and inspection work is classified by technical function (Fig. 75). The performance of such work and the results of inspections are documented by technical function. The analysis and repair of defects are done chiefly in reference to the technical function. In the course of examination, the engineering details on the assigned information can be considered; technical defects can be localized through the assignment of installation locations.

Such an information model makes possible an integral analysis of process facilities and, ultimately, of an entire company. The details of the model must be tailored to the technical circumstances. The storage and management of information in relation to engineering objects is crucial for its consistent use throughout the economic life of the plant. Only in this way is it possible to generate documents, at any time and at comparatively low cost, permitting instant analysis from any desired viewpoint.

4. From Process Knowledge to Process Control

4.1. Principles

Industrial processes, in contrast to natural processes, are artificially designed, consciously initiated, and purposely controlled so as to serve a certain end. It is not enough to understand an industrial process and be able to describe its internal dynamics in terms of a model; if the process is to take place at all, some form of external control is needed. This external control includes all functions required for coordinating and controlling the process. The structure, form, and implementation of these process control functions are the subject of the chapter that follows.

The sections on feedback control and optimal control set forth the most important design methods from control theory. Section 4.4 explains the concept of binary control in a variety of applications. In Section 4.5 it is shown how the operational control hierarchy can be structured for a plant as a whole. Process control functions play a central role in operational process control. For better understanding, these can be explained briefly before going on to the rest of the chapter.

Figure 76 shows schematically how process control functions work. The same diagram applies to all control functions. Each control function, on the basis of its own logic, seeks to bring the actual properties of a system into accord with the specified nominal properties. Each control function thus embodies certain assumptions about the behavior of the system. Using this knowledge, a control function can issue control commands in such a way as to influence the system in the desired fashion.

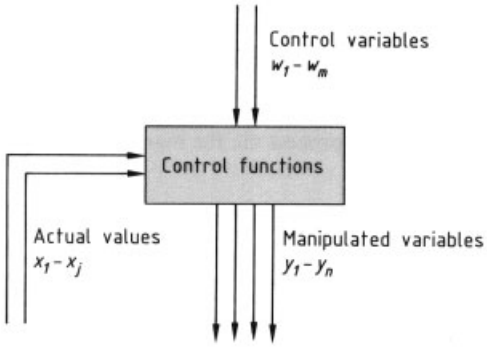

Figure 76. Schematic diagram of process control functions

In process control, the control variables and manipulated variables correspond to variable process or product properties. Manipulated variables either impinge directly on actuators or play the role of control variables, acting on lower-level control functions. In this way, process control can be built up in a hierarchical manner.

If we consider, for example, the adjustment of the pH in a reactor by the addition of acid, we obtain the hierarchy illustrated in Figure 77.

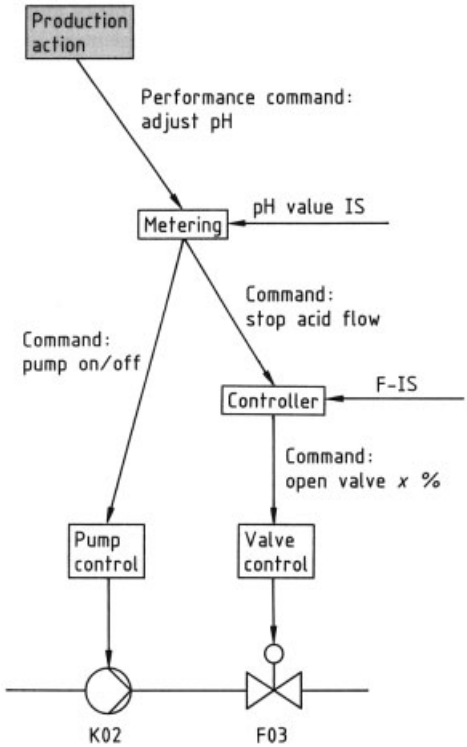

Figure 77. Schematic example of the hierarchy for a pH adjustment

The task of the production action is to make a product having certain properties at the discharge from the process section. For this purpose, one of the manipulated variables available to it for controlling the process is the pH in the vessel.

The metering device seeks to establish the specified time dependence of pH in the vessel; it therefore passes on the acid flow rate to the lower-level flow controller as a setpoint.

The flow controller, in turn, sets the required flow rate by specifying a certain valve position. The valve position controller exerts a corresponding control pressure to insure that the valve comes into the required position.

In most cases, a process control function has only one control variable; in general, however, m control variables and n manipulated variables must be allowed for.

The control concept can be customized for any process control function. The methods of control engineering are therefore applicable to process control functions at all levels in the hierarchy.

Process control functions can be classified in four groups by the concept employed:

1) Timed process control functions
2) Event-controlled feedback functions
3) Continuously controlled functions
4) Process control functions (optimized with respect to some quality criterion)

The selection of a concept depends less on the type of process control function and its position in the hierarchy than on the quality requirements, the anticipated severity of disturbances, the possibility of directly or indirectly measuring certain process properties, and (for the fourth group) the availability of suitable process models.

All the concepts are equally capable of reproducing a specified process if it is assumed that there are no disturbances.

Only when malfunctions or variations affect the course of the process will the four groups differ in their capabilities.

In practice, the concepts are combined as needed.

In what follows, the various concepts are illustrated for the case of a pH control. The task for the process control function is to regulate the addition of acid so as to keep the pH in a reactor at the proper value for the process.

Timed process control function: Acid is admitted to the reactor at a fixed, specified flow rate $q(t)$. If there are no disturbances, the desired pH value is established in the reactor. The actual pH value is not fed back.

Event-controlled feedback function: Acid is admitted to the reactor at a fixed, specified flow rate $q(t)$ until a certain specified pH has been reached; the flow is then cut off. If the process takes place without disturbances, the pH in the reactor follows the desired trajectory. If there are disturbances, the actual pH agrees with the desired pH, at least at the end point. The pH must be measured at least once.

Continuously controlled function: The admission of acid is controlled so that the pH in the reactor follows a fixed, specified control trajectory. Control actions compensate for upsets. Continuous pH measurement is required.

Optimal process control function: The control trajectory pH(t) is not specified in fixed form but is determined on the basis of the instantaneous process state and a quality criterion. If the fixed trajectory of concept 3 is assumed optimal in the sense of the quality criterion, functions 3 and 4 initially result in the same control trajectory. No difference is seen until a disturbance causes the actual value to depart from the nominal trajectory. Under concept 3, the control system attempts to force the process back onto the old trajectory; under concept 4, the process is guided along a new trajectory that is optimal when the actual situation is taken into account.

4.2. Feedback Control

Introduction. This section explains the concepts of feedback control and feedforward control. It discusses the distinction between the two as well as what process properties and what process requirements make a feedback structure necessary. The drawbacks to a feedback structure are also elucidated.

Any process is characterized by the action of the inputs on the outputs (Fig. 78). Inputs are classified as manipulated variables and disturbances. The former make it possible to intervene in the process, while the latter are undesired environmental effects that act on the process (see Chap. 3). Process outputs are quantities that are to be influenced in an appropriate way. Desired values for the outputs are called setpoints. Time-varying setpoints are called command or reference variables.

The objective of any control system is to interact with the process via the manipulated variables so as to insure stability and good performance.

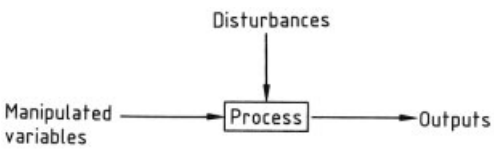

Figure 78. Schematic diagram of a process

The second point means that the outputs should react in a desired fashion to a change in the setpoints (reference variables) or the disturbances. In particular, they should react sufficiently fast, and the steady-state difference between setpoints and outputs should be sufficiently small.

Feedforward control can be diagrammed as an open loop (Fig. 79). On the basis of externally specified setpoints and measured disturbances, the controller generates manipulated variables as process inputs. No check is made to determine whether the selected manipulated variables actually have the desired effect on the outputs. With this structure the objectives cited can be achieved only if

- All disturbances to the process can be measured
- The effets of control actions on the outputs are known with sufficient precision, so that the mathematical process model underlying the controller design reflects the main characteristics of the real process
- The uncontrolled process is stable (an unstable process cannot be stabilized by feedforward control)

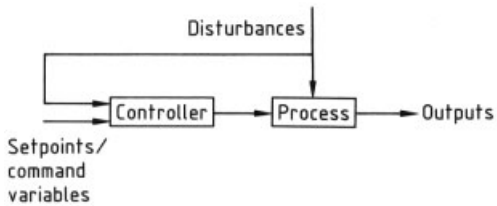

Figure 79. Feedforward control

In contrast, good performance can often be achieved with feedback control (Fig. 80), even if the above requirements are not met. Some of the outputs to be influenced are measured and compared with their setpoint values. The differences between these values are transmitted to the controller, which uses them in calculating the required control actions. The drawbacks of a feedback control structure are that feeding back the outputs

- Introduces additional measurement errors
- May (if the controller is inappropriately designed) make a stable process unstable
- Fails to react to disturbances until after they have already influenced the outputs

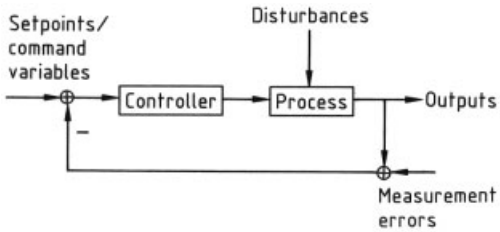

Figure 80. Feedback control

In view of the last point, feedback and feedforward control concepts are often combined. An example is a feedback structure with an additional disturbance input to the controller (Fig. 81), in which measurable disturbances are partly compensated by feedforward control but in general they cannot be completely suppressed because of shortcomings in the mathematical process model. Feedback control is used in an attempt to diminish the "residual effects" of these disturbances, as well as the effects of unmeasurable disturbances on the outputs. The difficult aspect of such "mixed" concepts is generally the design of the feedback component. Therefore, the description in what follows will be restricted to this component. Detailed discussions of feedforward and feedback control concepts, along with analyses of their advantages and disadvantages, can be found in [304, 305].

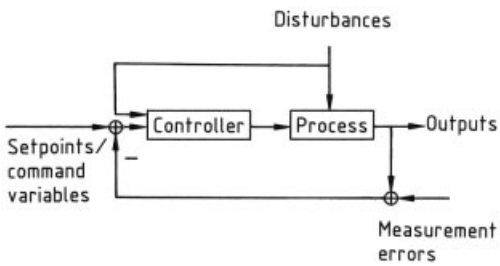

Figure 81. Feedback configuration with additional feedforward control of disturbances

Characterization of Desired Control Loop Properties. Before discussing methods of control system design, it is necessary to explain some concepts required for characterizing properties of a control loop.

This section is limited to linear process models (see Section 3.3) and controllers. This restriction is justified in most cases because engineering processes often behave in an approximately linear fashion near the operating point and therefore can be handled by a linear controller. The deviation of the process behavior from the linearity of the model equation can then be treated as part of the model uncertainty, which must be taken into account in controller design anyway. It is also assumed that process model and controller are time invariant. The frequency response is then an ideal tool for investigating control loop properties.

Frequency Response. A linear, time-invariant system can be described by its frequency response $G(j\omega)$ (see Chap. 3). For a single-variable system (i.e., a system having one input and one output), $G(j\omega)$ is a complex-valued scalar, which can be graphically represented in either of two ways. First, it can be plotted in the complex plane for all (positive real) frequencies ω; this gives the Nyquist locus or polar plot of the frequency response. Second, the complex variable $G(j\omega)$ can be split into modulus (magnitude) $|G(j\omega)|$ and phase arg $[G(j\omega)]$ by polar decomposition,

$$G(j\omega) = |G(j\omega)|\, e^{j\arg[G(j\omega)]} \tag{4.2.1}$$

and each of these real quantities plotted separately as a function of ω. If logarithmic scales are used for frequency and modulus and a linear scale for phase, the result is the Bode plot of the frequency response. The modulus scale is usually multiplied by 20 and the unit of modulus is then the decibel: $|G(j\omega)|_{\mathrm{dB}} = 20 \log |G(j\omega)|$. Figures 82 and 83 show the polar plot and the Bode plot of the frequency response for a second-order lag element.

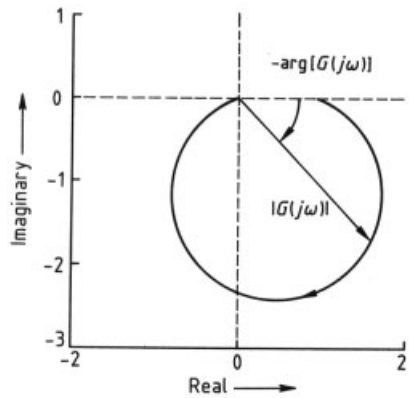

Figure 82. Nyquist locus or polar plot of frequency response

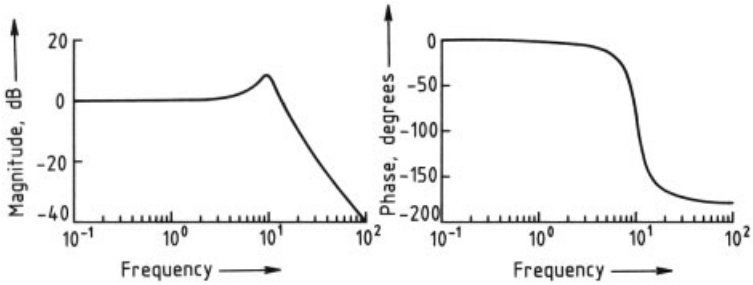

Figure 83. Bode plot

The frequency response has a straightforward physical explanation: If an asymptotically stable system characterized by $G(j\omega)$ is excited by a harmonic oscillation with frequency ω and amplitude $\hat{u}$ and the transients are allowed to decay, the output of the system oscillates harmonically at the same frequency, but with a phase shift φ and a different amplitude $\hat{y}$ (Fig. 84). The ratio of the output and input signal amplitudes (gain) and the phase shift between the two signals are described by the modulus and phase of $G(j\omega)$, respectively, and can thus be read directly from the Bode plot:

$$|G(j\omega)| = \hat{y}/\hat{u} \tag{4.2.2}$$

$$arg\,[G(j\omega)] = \varphi \tag{4.2.3}$$

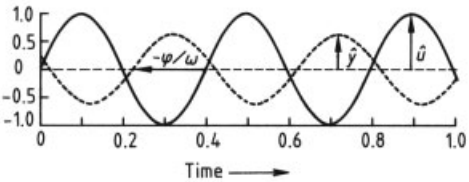

Figure 84. Harmonic oscillations at the input and output of a linear time-invariant system

The frequency response is analyzed in detail in control engineering textbooks, for example, [304–306].

The requirements imposed on the closed control loop can be quantified with the aid of the frequency responses $G(j\omega)$ of the process and $K(j\omega)$ of the controller.

Stability. The Nyquist criterion represents a necessary and sufficient condition for the stability of a closed control loop [304, 307]. It requires that the Nyquist locus of the open-loop frequency response

$$Q(j\omega) := G(j\omega)\,K(j\omega) \tag{4.2.4}$$

encircles the "critical point" $(-1{,}0)$ $m/2$ times counterclockwise as the frequency ω goes from 0 to ∞, where m is the number of unstable poles of the open loop. Because the Nyquist locus of the example in Figure 85 encircles the critical point 1/2 times, the associated closed loop is stable if and only if either $G(j\omega)$ or $K(j\omega)$ (but not both) has one unstable pole.

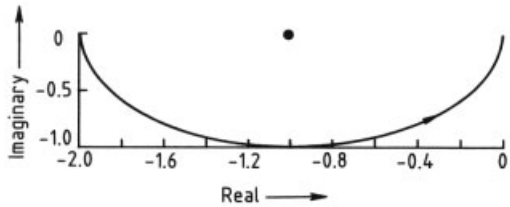

Figure 85. Example of Nyquist stability criterion

Response to Reference Inputs and Disturbances; Noise Suppression. To determine whether a control loop exhibits good response to reference inputs and disturbances and has the ability to suppress measurements noise, it is necessary to examine how all quantities acting on the control loop from the outside (Fig. 86) influence the control error.

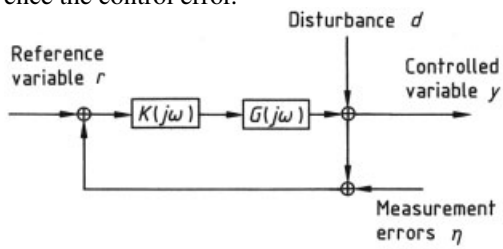

Figure 86. Linear feedback loop

The error $e(j\omega)$ is defined by

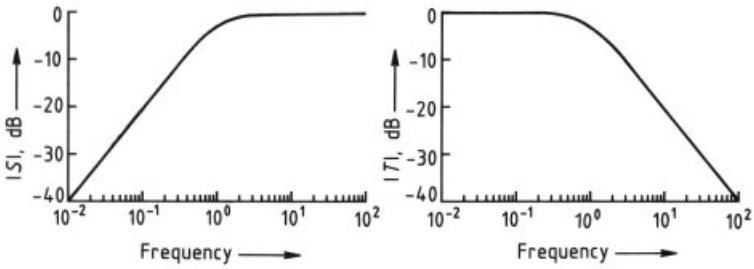

Figure 87. Typical shapes of $|S(j\omega)|$ and $|T(j\omega)|$

$$e(j\omega) := r(j\omega) - y(j\omega) \tag{4.2.5}$$

Hence

$$e(j\omega) = \underbrace{\frac{1}{1 + G(j\omega)\,K(j\omega)}}_{S(j\omega)}[r(j\omega) - d(j\omega)] + \underbrace{\frac{G(j\omega)\,K(j\omega)}{1 + G(j\omega)\,K(j\omega)}}_{T(j\omega)}\eta(j\omega)$$

The quantity $S(j\omega)$ is called the sensitivity function, because it gives the sensitivity of the control error to a disturbance acting on the output and to the reference variable. $T(j\omega)$ is known as the complementary sensitivity function. The objective is to keep $|S(j\omega)|$ and $|T(j\omega)|$ small, and thus to achieve a small gain between quantities acting on the control loop from outside and the control error. Unfortunately, this is generally impossible, since for any ω

$$S(j\omega) + T(j\omega) = 1 \tag{4.2.7}$$

This "fundamental dilemma" of control engineering, embodied in Equation (4.2.7), is often alleviated because in most casses reference variables and disturbances contain mainly low-frequency components and hence vary comparatively slowly with time, while measurement noise commonly occurs at high frequencies. It is then sufficient to make $|S(j\omega)|$ small at low frequencies and $|T(j\omega)|$ small at high frequencies. Figure 87 shows typical plots of the sensitivity and complementary sensitivity functions. The bandwidth ω_b of the closed loop is the frequency above which

$$|T(j\omega)| < 0.708 \text{or} |T(j\omega)|_{dB} < -3 \tag{4.2.8}$$

It "separates" the frequency range in which the closed loop has good response to reference inputs and disturbances from the frequency range in which measurement noise is suppressed. Establishing the desired bandwidth is thus one of the most important decisions in any controller design problem.

Robustness. The discussion up to this point has had to do with how to insure desired properties for the "nominal" control loop (i.e., the system consisting of process model and controller). The objective is to transfer these properties to the real control loop (the system consisting of the real process and the controller). In this context, robustness means that the properties of the nominal control loop should be robust with respect to errors in the model.

Robustness can be investigated only if an upper boundary can be given for the size of the model error. Often an upper bond for the relative model error is estimated by assuming that the real process behaves in a linear and time-invariant fashion near the operating point under consideration. The (unknown) frequency response of the process is represented by $G_r(j\omega)$. If the known frequency response of the model is $G(j\omega)$, the relative model error is then defined by

$$\Delta(j\omega) = \frac{G_r(j\omega) - G(j\omega)}{G(j\omega)} \tag{4.2.9}$$

Suppose the modulus of the relative model error is bounded by a known real-valued and positive function $p(j\omega)$:

$$|\Delta(j\omega)| < p(j\omega) \tag{4.2.10}$$

Figure 88 shows a typical curve of $p(j\omega)$, which grows rapidly with increasing frequency.

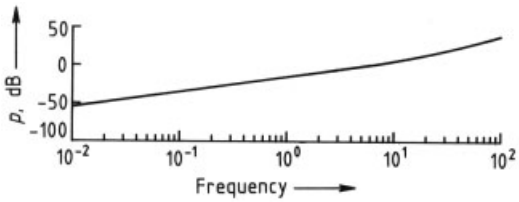

Figure 88. Typical error bound

If G_r and G have the same number of unstable poles, it is easy to show [308] that the stability of the control loop is robust to all model errors permitted by Equations (4.2.9) and (4.2.10) if and only if the following holds for the nominal complementary sensitivity function

$$|T(j\omega)| < \frac{1}{p(j\omega)} \tag{4.2.11}$$

To achieve good performance of the real loop the nominal sensitivity function must usually be made sufficiently small at low frequencies while the nominal complementary sensitivity function must be sufficiently small in the frequency range where the model uncertainty is large (see Fig. 88) [309].

Open-Loop Shaping. The Nyquist criterion makes it possible to guarantee closed-loop stability by appropriate manipulation of the open-loop frequency response $Q(j\omega)$. The sensitivity function and the complementary sensitivity function are used to guarantee the other control-loop properties stated. Since both quantities depend on Q in a simple way (see Eq. 4.1.6), all the desired properties of the control loop can be achieved by suitable "shaping" of the open-loop frequency response $Q(j\omega)$. The term "open-loop shaping" is often used in this context. Good response to reference inputs and disturbances thus requires that

$$|Q(j\omega)| >> 1 \tag{4.2.12}$$

at low frequencies, whereas suppression of measurement noise requires that

$$|Q(j\omega)| << 1 \tag{4.2.13}$$

at high frequencies. In the region near the gain crossover frequency ω_d, however, $|Q(j\omega)|$ must not roll off too rapidly. (The gain crossover frequency is the frequency at which $|Q(j\omega)|$ intersects the 0 dB line.) Otherwise, disturbances, reference variables, and measurement noise can produce large control errors in this frequency range. Moreover, rapid gain roll-off near the crossover frequency threatens closed-loop stability. This is a direct consequence of the Bode gain – phase relation [304, 310]. The ideal would be a 20 dB per decade roll-off for $|Q(j\omega)|$ in this frequency range. If this is the case, the gain crossover frequency ω_d is a good approximation to the bandwidth ω_b of the closed loop.

Multivariable Systems. The discussion to this point has related only to systems with one input and one output. Most industrial processes, however, have multiple manipulated variables and multiple outputs to be suitably influenced. Each of the manipulated variables will, in general, affect several or even all of the controlled variables. The process is then said to be a multivariable (multi-input, multi-output) system, and its dynamics (assuming linearity and time-invariance) are described by the frequency response matrix

$$G(j\omega) = \begin{bmatrix} G_{11}(j\omega) \cdots G_{1q}(j\omega) \\ \vdots \qquad\qquad \vdots \\ G_{p1}(j\omega) \cdots G_{pq}(j\omega) \end{bmatrix} \tag{4.2.14}$$

Most of what has been discussed up to now can be extended to multivariable systems with no difficulty.

The multivariable Nyquist criterion is a necessary and sufficient condition for the stability of the closed loop. It requires that the frequency-dependent eigenvalues of the matrix $Q(j\omega)=G(j\omega)K(j\omega)$ (the"characteristic loci"), taken together, encircle the "critical point" $m/2$ times [311, 312].

To assess the other control-loop properties, the concept of gain must be generalized [313]. In the single-variable case, the gain of an asymptotically stable system corresponds to the modulus of the frequency response. The situation is somewhat more complicated in multivariable systems. Such a system has a gain that depends on the direction of the input vector. The gains of most interest are the minimum and maximum (frequency-dependent) gains. These are characterized by the minimal and maximal singular values of the frequency response matrix:

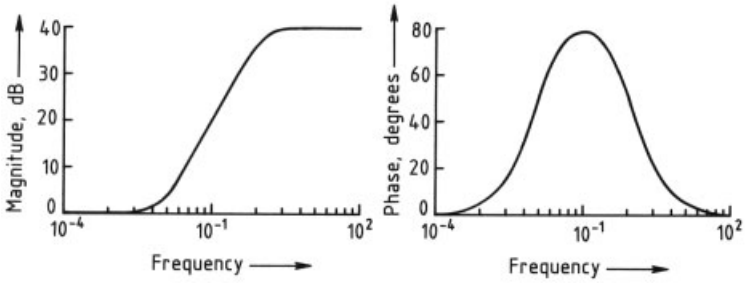

Figure 89. Bode plot of a lead compensator

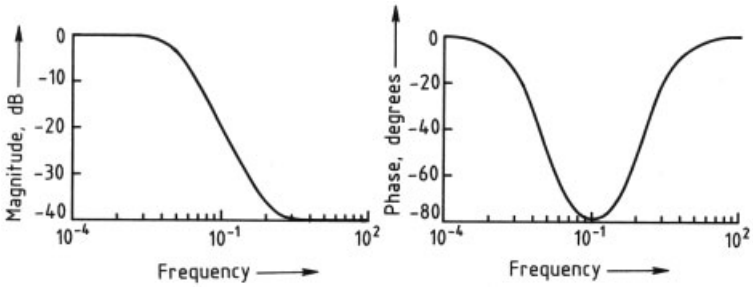

Figure 90. Bode plot of a lag compensator

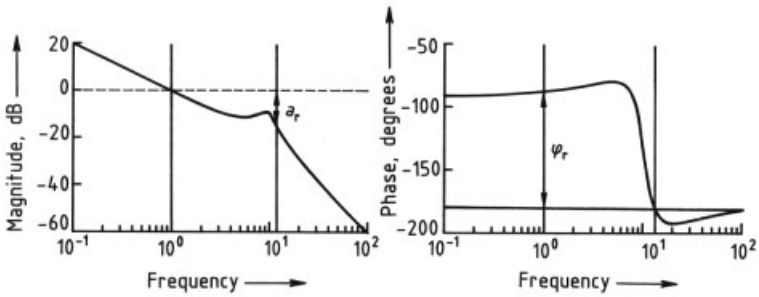

Figure 91. Bode plot for $Q(j\omega)$ in the example

$$\min_{u\neq 0}\frac{\|y(j\omega)\|}{\|u(j\omega)\|}=\sigma[G(j\omega)]$$
$$=\sqrt{\lambda_{\min}[G^T(-j\omega)G(j\omega)]} \tag{4.2.15}$$

$$\max_{u\neq 0}\frac{\|y(j\omega)\|}{\|u(j\omega)\|}=\bar{\sigma}[G(j\omega)]$$
$$=\sqrt{\lambda_{\max}[G^T(-j\omega)G(j\omega)]} \tag{4.2.16}$$

Here $\|\ldots\|$ denotes the Euclidean vector norm (see Chap. 3), while $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues.

With these gains, the requirements derived for single-variable systems can readily be extended to the multivariable case [308]:

$$\bar{\sigma}[S(j\omega)]<<1 \text{ or } \sigma[Q(j\omega)]>>1 \quad \text{for} \quad \omega<<\omega_{\mathrm{b}} \tag{4.2.17}$$

$$\bar{\sigma}[T(j\omega)]<<1 \text{ or } \bar{\sigma}[Q(j\omega)]<<1 \quad \text{for} \quad \omega>>\omega_{\mathrm{b}} \tag{4.2.18}$$

The sensitivity matrix S and the complementary sensitivity matrix T are defined similarly to the scalar variables in the single-variable case: $S:=(I+GK)^{-1}$ and $T:=(I+GK)^{-1}GK$.

If the requirements on the various controlled variables differ sharply, frequency-dependent weighting matrices for S, T, and Q can be used to handle the situation.

Controller Design Methods. This section surveys the most important methods of controller design.

Classical Control Engineering. The methods of classical control engineering continue to be dominant in many areas. In general, these methods assume a linear, time-invariant, single-variable process model. If the requirements on the controlled system are not too stringent, however, these methods can still be used in many cases even where such models give a highly erroneous description of the real process.

The most important design method in classical control engineering is the frequency response method [304,306,314]. It has the objective of shaping the open-loop frequency response $Q(j\omega)$ in accordance with what was said in the preceding section. For this purpose compensators

$$K(j\omega) = k\frac{1 + j\omega T_1}{1 + j\omega T_2} \tag{4.2.19}$$

are connected in series with the process. The compensator increases the phase if $T_1 > T_2$ (lead compensator) and decreases it if $T_2 > T_1$ (lag compensator); see Figures 89 and 90. A special case of the lag compensator for $k = \tilde{k}T_2$ and $T_2 \to \infty$ is the proportional-plus-integral (PI) element. Because of the particular axis scaling, the Bode diagram of a series system can be obtained by simply adding the amplitude and phase curves of the individual elements. It is thus easy to see what kind of compensating network is required to modify the curves in the desired way. For example, Figure 83 shows the frequency response of a process model; the bandwidth of the closed loop is required to be $\omega_b = 1$. In this case a PI element with frequency response $K(j\omega) = (1 + 0.1j\omega)/j\omega$ can be used as the compensator. The resulting amplitude and phase plots for the open loop are shown in Figure 91. Clearly, $Q(j\omega)$ satisfies requirements of Equation (4.2.12) at low frequencies and of Equation (4.2.13) at high frequencies. The polar plot of $Q(j\omega)$ can be used to check whether the Nyquist criterion is also satisfied. Figure 92 shows the curve near the origin. The quantities ϕ_r ("phase margin") and a_r ("gain margin") give the "distance" at which the polar plot of the frequency response passes the critical point (i.e., how "close" the loop is to instability). The gain and phase margins can also be read directly from the Bode plot in Figure 91. The following can serve as a rule of thumb for the use of the frequency-response method: $a_r > 2.5$ or $|a_r|_{dB} > 8$ and $\phi_r > 30°$. Because the gain and phase margins in this example are sufficiently large, a further compensation of $Q(j\omega)$ can be dispensed with.

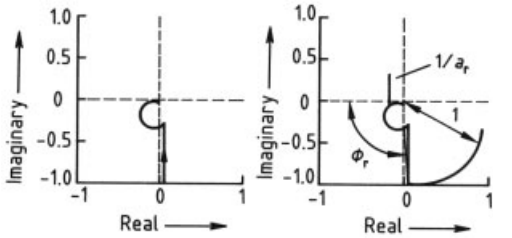

Figure 92. Polar plot for $Q(j\omega)$ in the example

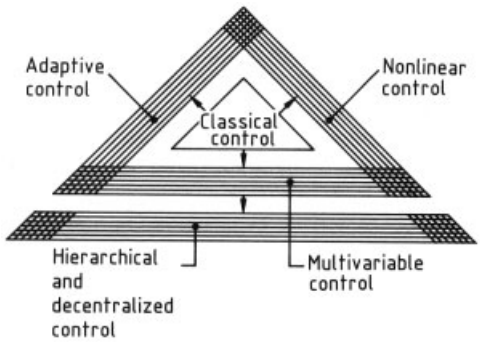

Figure 93. Classification of methods in control theory

Another widely used technique is the root-locus method [304,307,314]. This design procedure starts not with the frequency response, but with the transfer function of the plant (process) model. The location of poles and zeroes of this transfer function indicates how the poles, and thus the dynamics of the closed loop, vary as a function of the controller parameters and structure. This method is accordingly suitable for insuring the stability of the control loop. On the other hand, it generally proves quite difficult

to insure other desired properties of the control loop. Another disadvantage of this design technique it that it cannot be applied to process modes with time delays.

In process engineering, a controller structure (PID) is often specified and the parameters are obtained by experimenting on the process and evaluating the Ziegler – Nichols rules [307, 314]. When such a procedure is used in practice, there is no need for an explicit process model; the theoretical justification, however, presumes the existence of linear, time-invariant single-variable models.

Discarding one or more of the three model assumptions in classical control engineering leads to the other three (overlapping) "classes" of control engineering (Fig. 93).

Table 2. Multivariable design method

Time-domain	Frequency domain	
	Direct Nyquist array method	
	Inverse Nyquist array method	
	Characteristic locus method	Open-loop shaping
	RFN method	
	LQG/LTR method	
LQG method	H_2 Minimization	
Pole assignment	H_∞ Minimization	Closed-loop shaping
	IMC method	

Multivariable Control. Of the three model assumptions in classical control theory, the one furthest from reality is usually that the process can be described in terms of several independent single-variable models. Ignoring the links between various manipulated and controlled variables can lead to major problems: input controllers that work against one another (wasting energy) and instability of the controlled process may result. Design methods that take account of the links and thus the multivariable structure of the control loop exist in both the frequency and the time domains. Frequency domain methods begin with a transfer or frequency response matrix as the plant model; time-domain methods are based on state-space models (see Chap. 3). The frequency-domain methods can be classified in two groups: generalizations of the classical frequency-response method, in which an attempt is made to manipulate the open-loop frequency response matrix $Q(j\omega)$ in an appropriate way (open-loop shaping), and approaches that seek to influence directly, or optimize, the closed-loop frequency response matrices (closed-loop shaping). Table 2 lists the most commonly used linear, time-invariant multivariable design met ods. Some of these techniques are fully described in textbooks of multivariable control engineering [312, 315–318].

The LQG (linear quadratic Gaussian) method is based on a linear, time-invariant state-space model (see Chap. 3) of the control loop:

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t) \tag{4.2.20}$$

$$y(t) = Cx(t) + v(t) \tag{4.2.21}$$

where $w(t)$ and $v(t)$ are error terms representing, respectively, disturbances and measurement noise acting on the plant. They are modeled as white (Gaussian) random processes with a mean of zero [319, 320]. Their spectral densities P and R provide a measure of the "strength" of these random processes. They tell how dependable the deterministic components of the models are. If the goal is to keep the plant as close as possible to the operating point $x = 0$, subject to a restriction on the use of control energy, then it is desirable to seek a control strategy that will minimize the cost function

$$J = E\left[\int_0^\infty \left(x^T L x + u^T M u\right) \mathrm{d}t\right] \tag{4.2.22}$$

The (positive semidefinite) matrix L penalizes deviations of the state variables from zero, and the (positive definite) matrix M penalizes the control energy. The symbol E represents the expected value. This problem can be solved straightforwardly. The separation principle states that the problem can be approached in two independent steps:

– First, ignore the fact that the state is not measurable and suppose it is exactly known. The optimal control strategy is then a constant state feedback:

$$u(t) = -K_c x(t) \tag{4.2.23}$$

Here K_c can be uniquely calculated from the model matrices A and B and the weight matrices L and M.

– In the second step, determine an optimal (in the sense of minimum error variance) estimate $\hat{x}$ for the unmeasurable state vector x. This is done by the use of a Kalman – Bucy filter [319,321] (Fig. 94). The Kalman – Bucy filter includes the deterministic part of the plant model and thus attempts to simulate the behavior of the state variables. What is more, however, the model also receives measurement information y, which is used for correcting the estimate (unless it agrees with the measurement $\hat{y}$ predicted by the model internal to the filter). The filter matrix K_f that specifies this correction is uniquely determined by the matrices A and C of the deterministic model component and the spectral densities P and R of the stochastic component. The optimal control strategy is now to replace the true state x in Equation (4.2.23) by the estimate $\hat{x}$.

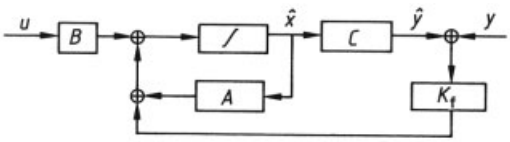

Figure 94. Kalman – Bucy filter

The derivation and details can be found in, for example, [322,323]. Despite its theoretical elegance, the use of the LQG method in practice commonly presents problems. In chemical engineering, state – space models rarely have a simple physical interpretation. Generally, one has to be content with simple frequency-domain models (transfer function matrix, frequency response matrix) derived from the input/output behavior of the real process. While these state – space models can be determined from such input – output models their state variables then often fail to correspond to physically relevant quantities. This fact makes it difficult to determine a suitable weighting matrix L. Finding the spectral densities (especially for the "system noise" w) may well prove to be a matter of luck as well. In such cases, the elements of P and L offer a large number of design parameters whose effects on the desired control-loop behavior are difficult to assess.

One way of avoiding this problem is the LQG/LTR (loop transfer recovery) design method [318,324], in which the design parameters P, R, L and M of the LQG design are divorced from any physical meaning and selected such that the open-loop frequency response matrix $Q(j\omega)$ satisfies the requirements stated in the preceding section. For this reason, the LQG/LTR design technique is often classified as a frequency-domain method.

A further time-domain method is pole assignment [307,315]. It has the objective of specifying the poles and thus the dynamics of the closed loop by means of appropriate state feedback. If the state is not measurable, it is estimated by a Luenberger observer. The Luenberger observer resembles the Kalman – Bucy filter in that it is an algorithm for reconstructing the system state and in that it contains a model of the plant. It differs from the filter only in the way the correction matrix K_f is computed. In the case of the observer, this is done by specifying the dynamics of the equation for the estimation error. There is again a kind of separation principle in pole assignment: the observer matrix K_f and the state-estimate matrix K_c can be designed independently. The use of this technique may also create problems if the state variables have no direct physical meaning. Furthermore, establishment of the control-loop dynamics does not guarantee satisfactory loop behavior.

The characteristic locus curve method [325] seeks to meet the requirements stated in the preceding section by manipulating the eigenvalues of the open-loop frequency response matrix $Q(j\omega)$. These eigenvalues $\lambda_{Qi}(j\omega)$ are called "characteristic frequency responses." If the corresponding eigenvectors are collected in the matrix $V(j\omega)$, the following holds

$$Q(j\omega) = V(j\omega)\,\mathrm{diag}\left\{\lambda_{Qi}(j\omega)\right\} V^{-1}(j\omega) \tag{4.2.24}$$

Spectral decompositions for the sensitivity matrix and the complementary sensitivity matrix are then given by

$$S(j\omega) = V(j\omega)\,\mathrm{diag}\left\{\frac{1}{1+\lambda_{Qi}(j\omega)}\right\} V^{-1}(j\omega) \tag{4.2.25}$$

$$T(j\omega) = V(j\omega)\,\mathrm{diag}\left\{\frac{\lambda_{Qi}(j\omega)}{1+\lambda_{Qi}(j\omega)}\right\} V^{-1}(j\omega) \tag{4.2.26}$$

The eigenvalues of S and T thus depend in a very simple way on the $\lambda_{Qi}(j\omega)$. Although the gains of a matrix cannot be deduced from the absolute values of its eigenvalues, these do serve as a good indication of the gains in many cases. Therefore, small $|\lambda_{Qi}(j\omega)|$ at high frequencies and large $|\lambda_{Qi}(j\omega)|$ at low frequencies are required. The crossover frequencies for the characteristic frequency responses are selected to correspond to the desired closed-loop bandwidth. What is more, the multivariable Nyquist criterion makes it possible to say how often the locus curves of $\lambda_{Qi}(j\omega)$ (the "characteristic locus curves") must enclose the critical point $(-1,0)$ for the (nominal) control loop to be stable. Once the characteristic frequency responses have been chosen, the behavior of $\bar{\sigma}[S(j\omega)]$ and $\bar{\sigma}[T(j\omega)]$ must be examined to ensure that Equations (4.2.17) and (4.2.18) are satisfied. Now a way must be found to modify the characteristic loci in the desired way. The ideal method is to use a controller frequency response matrix having the same eigenvectors as the plant frequency response matrix. The following would then hold:

$$Q(j\omega) = \underbrace{V \text{ diag}\{\lambda_{Gi}(j\omega)\} V^{-1}}_{G(j\omega)} \underbrace{V \text{ diag}\{\lambda_{Ki}(j\omega)\} V^{-1}}_{K(j\omega)} = V(j\omega) \text{ diag} \underbrace{\{\lambda_{Gi}(j\omega)\lambda_{Ki}(j\omega)\}}_{\approx \lambda_{Qi}(j\omega)} V^{-1}(j\omega) \tag{4.2.27}$$

In this case, the $\lambda_{Gi}(j\omega)$ could be treated as single-variable systems, and suitable single-variable controllers $\lambda_{Ki}(j\omega)$ could be designed by the classical frequency-response method. Unfortunately, this is rarely possible, because the elements of V are generally nonrational in $j\omega$, and the resulting controller matrix thus cannot be realized. One therefore has to be content with approximating the ideal controller by approximating the complex matrices $V(j\omega_i)$ with real matrices for selected frequencies ω_i.

A shortcoming of the characteristic locus method is that the $|\lambda_{Qi}(j\omega)|$ provide only an indication of the actual gains of $Q(j\omega)$. In the RFN (reversed-frame normalizing) method [326], attempts are made to bring the absolute values of the eigenvalues and the gains (singular values) of $Q(j\omega)$ into coincidence. If this is successful (approximately), all the conditions stated in the preceding section can be met through manipulation of these quantities. A drawback of the characteristic locus and RFN methods is that the characteristic frequency responses cannot be associated with individual plant input and output variables. There is accordingly no systematic way of making different controlled variables display different transient responses or different steady state accuracy. Such difficulties are less pronounced (or nonexistent) in the methods discussed below.

The direct Nyquist array method [311] breaks the design process into two steps:

1) First, an attempt is made to approximately diagonalize the plant model by means of a simple – and, if possible, constant – compensator K_c. This means that off-diagonal elements of the frequency-response matrix should be small in magnitude compared to the diagonal elements. In particular, if the elements of the compensated $p \times p$ matrix $Q_c = GK_c$ satisfy

$$\sum_{k=1,k\neq i}^{p} |Q_{C_{ki}}(j\omega)| < |Q_{C_{ii}}(j\omega)| \tag{4.2.28}$$

for $i = 1, \ldots, p$ and $0 \leq \omega < \infty$, then $Q_c(j\omega)$ is said to be (column-)diagonal-dominant. The diagonal-dominance of a frequency-response matrix can easily be checked graphically. The polar plots of the diagonal elements $Q_{c^{ii}}(j\omega)$ are overlaid with circles whose radii are given by the left-hand side of Equation (4.2.28). The bands formed by these circles are called Gershgorin bands. If Equation (4.2.28) is satisfied, these bands do not cover the origin of the complex plane. A number of methods have been advanced for obtaining narrow Gershgorin bands [311, 315, 318].

2) In the second step, the small off-diagonal elements of Q_c are ignored. Single-variable controllers $K_i(j\omega)$ are designed for the principal diagonal elements of the compensated plant

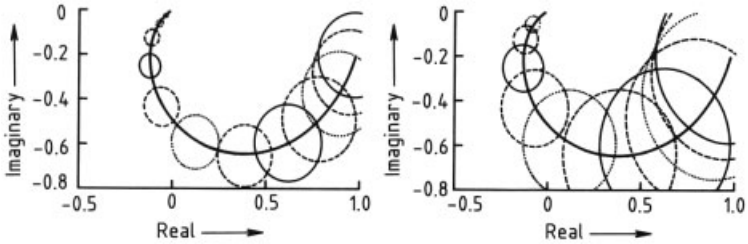

Figure 95. Gershgorin bands for a 2×2 system

model Q_c; thus the open-loop frequency-response matrix is obtained:

$$Q(j\omega) = \underbrace{G(j\omega)K_c(j\omega)}_{Q_c(j\omega)} \text{ diag}\{K_i(j\omega)\} \quad (4.2.29)$$

This operation alters the form of the Gershgorin bands but leaves their width unchanged, since both the center and the radius of each Gershgorin circle are multiplied by the same factor. The Gershgorin bands of $Q(j\omega)$ make it possible to assess the stability of the closed loop, since they contain the characteristic loci curves. If the bands do not cover the critical point (−1,0), they determine how many times the characteristic locus curves encircle this point. Accordingly, the $K_i(j\omega)$ must be chosen such that the Gershgorin bands of Q do not cover the point (−1,0) and, collectively, they encircle this point counterclockwise $m/2$ times. Figure 95 shows an example for $m = 0$. To perform a classical single-variable design for the i-th controller component, the corresponding frequency-response element of the compensated plant model also must be known for the case where the remaining control loops are already closed. Because small off-diagonal elements are still present, this frequency response will not coincide with $Q_{c^{ii}}$. Its polar plot, however, lies in the interior of the Gershgorin band associated with $Q_{c^{ii}}$. This is a further reason for the importance of narrow Gershgorin bands.

Experienced users generally prefer the inverse Nyquist array method [311], which has the advantage that the relationship between $y_i(j\omega)$ and $u_i(j\omega)$ becomes sharper when the other control loops are closed. This relationship is localized by Ostrowski bands, which lie in the interior of the Gershgorin bands. The method has the disadvantage that it entails the use of inverse frequency responses. Furthermore, the technique can be applied only to systems in which the number of manipulated variables equals the number of controlled variables.

A design method that has been gaining popularity in the process engineering field is the IMC (internal model control) technique [327]. In its simplest form, this method assumes a stable plant. The controller then contains a plant model in parallel with the actual plant (Fig. 96). The requirements on the closed control loop are easily fulfilled with this structure, because

1) The closed loop is stable if and only if the controller component $\Gamma(j\omega)$ is stable
2) The sensitivity matrix and the complementary sensitivity matrix depend on Γ in a very simple way:

$$S = I - G\Gamma$$
$$T = G\Gamma \quad (4.2.30)$$

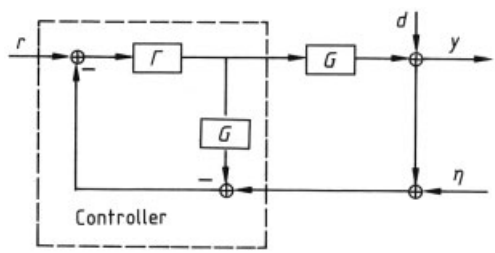

Figure 96. IMC structure

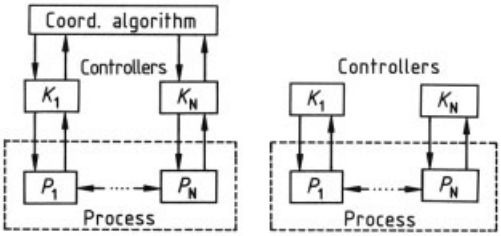

Figure 97. Hierarchical and decentralized control structures

A promising approach to the design of multivariable controllers is H_∞ minimization [328, 329]. The discussion in the preceding section showed what kind of behavior should be sought for the maximum singular values of the sensitivity matrix and the complementary sensitivity matrix. In addition, the matrix $K(j\omega)S(j\omega)$ can be included in the analysis. This matrix indicates how disturbances in the control loop affect the manipulated variables (see Fig. 86). Typically, $\bar{\sigma}[K(j\omega)S(j\omega)]$ is required to be bounded for all frequencies. This is intended to prevent the manipulated variables being saturated when disturbances occur. To avoid high-frequency oscillation of actuators, a further requirement is that $\bar{\sigma}[K(j\omega)S(j\omega)]$ roll off at high ω. For S, T, and KS to have small gains, each in its own frequency range, distinct frequency-dependent weighting matrices W_S, W_T, and W_KS are assigned to them. A controller is then sought that stabilizes the control loop and minimizes

$$\sup_\omega \bar{\sigma} \begin{bmatrix} W_\mathrm{S}(j\omega)S(j\omega) \\ W_\mathrm{T}(j\omega)T(j\omega) \\ W_\mathrm{KS}(j\omega)KS(j\omega) \end{bmatrix} := \left| \begin{matrix} W_\mathrm{S}S \\ W_T T \\ W_\mathrm{KS}KS \end{matrix} \right|_\infty \quad (4.2.31)$$

$\|\ldots\|_\infty$ is known as the H_∞ norm. Algorithms for solving this problem are complex and computationally intensive. For the user, however, this point is usually of as little importance as the nonuniqueness of the solution. The design engineer need only select the weighting matrices. As a rule, they are chosen to be diagonal matrices with stable, minimum-phase elements. Low-pass elements are commonly used for weighting S, so that small $\bar{\sigma}[S]$ is obtained at low frequencies. Similarly, high-pass elements are used for weighting T and KS. This method has the considerable advantage that the elements of the weighting matrices can be directly assigned to distinct manipulated and controlled variables. Their effect on closed-loop behavior is thus easy to assess. The selection of weights is illustrated in page 86.

Decentralized and Hierarchical Control. When the known multivariable methods are used, difficulties arise if the process to be controlled has many coupled inputs and outputs. Such large-scale systems should be broken down into subproblems. This can be done heuristically (and the subproblems defined, for example, as physical parts of the plant) or by means of formal techniques (e.g., the block relative-gain array [330]). A (multivariable) controller is assigned to each subsystem; in the design of this controller, however, links between the subsystems cannot be completely neglected. If the controllers operate independently of one another, this is known as decentralized control; if they are coordinated at a higher level, the control scheme is said to be hierarchical (Fig. 97) [331–335]. Most known hierarchical and decentralized methods belong to the time domain. With regard to their use in practice, essentially the same problems arise as in the LQG or pole assignment methods. The development of frequency-domain decentralized and hierarchical methods promises to be helpful.

Adaptive and Nonlinear Control. Linear control methods reach their limits in processes having no operating point or a continuously variable operating point (batch processes).

Conceptually, the simplest way of dealing with problems involving an operating point that varies slowly over time is to use adaptive control [336, 337]. Adaptive controllers continuously modify themselves to match the process being controlled. In general, only the simplest class of adaptive methods are employed in practice: self-tuning controllers. An indirect variant of such a controller is made up of a three-level algorithm: on-line estimation of parameters of a process model with specified (usually linear) structure; calculation of parameters for a controller structure, also specified; and action of the resulting controller on the process (Fig. 98). In this way, the entire (nonlinear) control problem is broken down into an estimation problem and a straightforward, generally linear, control problem. Adaptive control structures can therefore be treated as special nonlinear control structures.

If the operating point of the process changes very rapidly or a linear description is altogether impossible, explicitly nonlinear methods must be considered. However, these methods, such as exact linearizaton [338], are restricted to a very small class of nonlinear systems. Exceptions are semi-heuristic methods such as nonlinear versions of MPC (model predictive control) [339].

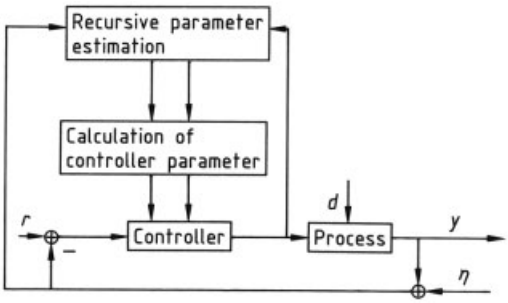

Figure 98. Self-tuning controller (indirect)

Design Aids. Most modern design techniques are so complex that they could not be employed without the help of suitable CACSD (computer-aided control systems design) packages. A continuously updated survey of available software is the ELCS (Extended List of Control Software) [340]. A further summary can be found in [341].

Design Example. This section shows how the classical frequency-response method and H_∞ minimization can be used in the design of controllers for a pilot-plant distillation column with 40 trays (Fig. 99). The device separates a mixture of methanol and propanol. The control objective is to keep the concentrations constant in the bottoms and distillate as the feed rate or feed concentrations change. As is common in binary separations, the concentrations and temperature on an individual tray are related in a unique way. Instead of the concentrations, which are difficultto measure, the temperatures can therefore be controlled. Simulation studies based on a detailed nonlinear model (using the DIVA simulation system [342]) have shown that the temperatures on the 14th and 28th trays are most sensitive to disturbances in the feed. These two temperatures are therefore chosen as controlled variables. The manipulated variables are the heat input Q_E and the reflux ratio $\varepsilon = L/V_K$. The controller design is based on a simple linear model obtained by evaluating step responses:

$$\begin{bmatrix} \Delta T_{14} \\ \Delta T_{28} \end{bmatrix} = \underbrace{\begin{bmatrix} -\dfrac{1220}{6.1j\omega+1} & \dfrac{100}{(1.2j\omega+1)(1.5j\omega+1)} \\ -\dfrac{1500}{2.7j\omega+1} & \dfrac{324}{1.6j\omega+1} \end{bmatrix}}_{G(j\omega)} \begin{bmatrix} \Delta\varepsilon \\ \Delta Q_E \end{bmatrix} \tag{4.2.32}$$

Here Δ denotes deviations of the indicated quantities from the desired stationary operating condition. The time constants of the model are given in hours. The most important specification for linear controller design is the closed-loop bandwidth, which should be close to $40\,\mathrm{h}^{-1}$.

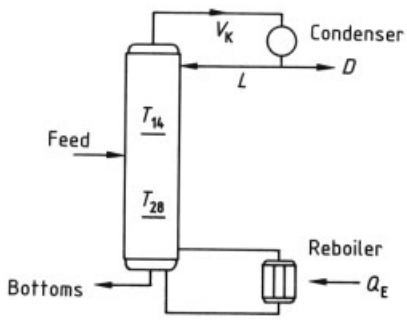

Figure 99. Distillation column
V_K = vapor stream; L = reflux stream; D = distillate; Q_E = heat input rate

In the classical frequency-response method, the off-diagonal elements are neglected and only the diagonal elements of G are considered. Figure 100 shows the magnitudes for these elements. To suppress steady state errors PI controllers are used. Their time constants are selected such that the magnitude of both elements rolls off at a rate of ca. 20 dB per decade over the whole frequency range. The controller gains are set so that the two magnitude curves coincide over a broad frequency range and the bandwidth

of the closed multivariable control loop takes on the desired value. The last point is verified by computing the singular values of the complementary sensitivity matrix (Fig. 101).

In the H_∞ minimization method, weight matrices are selected for *S, T*, and *KS*. Diagonal matrices with minimum-phase elements are employed. Figure 102 shows the amplitude characteristics of the elements of W_S (low-pass) and W_T (high-pass). All elements are chosen such that their gains intersect the 0 dB line at 40 h^{-1}, because closed-loop bandwidth should lie close to this frequency. Small constants (0.01) are chosen as diagonal elements of W_{KS}, since in this example no problems with actuator saturation or high-frequency actuator oscillation are expected. The H_∞ cost function is completely defined in this way. It is minimized by a controller of order eight, but this can be reduced to order 5 by model reduction techniques. Figure 103 gives plots of the singular values of *S* and *T*.

Both controllers were tested in a pilot plant. It was found that the controller obtained by minimization of H_∞ met all practical requirements, while the classical PI controller required on-line tuning of the gains.

Details and comparisons with other multivariable design methods can be found in [343].

4.3. Optimal Control

Introduction. Before something can be optimized, the term "optimal" must be defined. In quantitative form, this is done by establishing a performance index (performance criterion, optimization criterion).

The optimization of a dynamic system means formulating the dynamic system behavior in an optimal way under the selected criterion. The system is characterized by the time-dependent processes by which it goes from a given initial state $x(t_0)$ to a final state $x(t_e)$.

The optimization of a controlled dynamic system generally begins with the system already specified; only the control scheme is varied to achieve the optimal system behavior. If the control structure is fixed, so that only its parameters can be varied, we speak of parametric op-

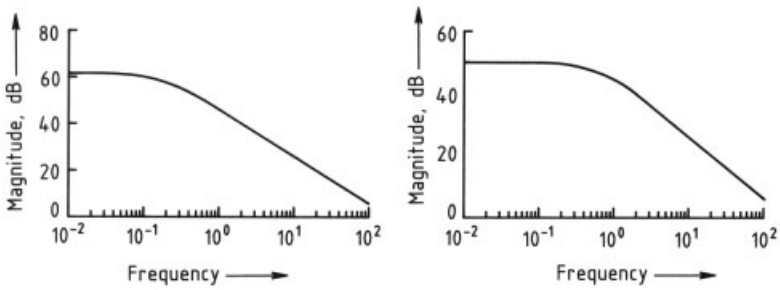

Figure 100. Magnitudes of the diagonal elements

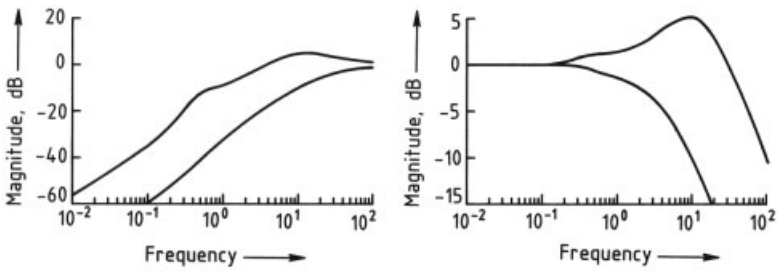

Figure 101. Singular values of the sensitivity matrix and complementary sensitivity matrix

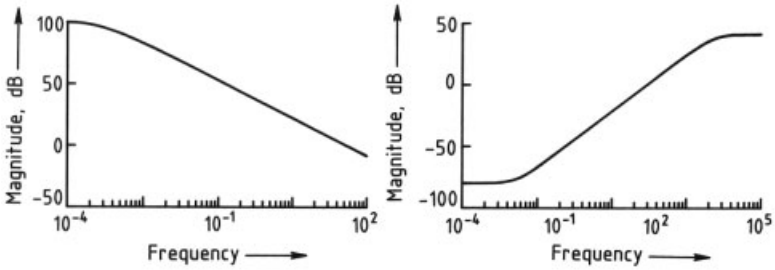

Figure 102. Magnitudes of the weight elements

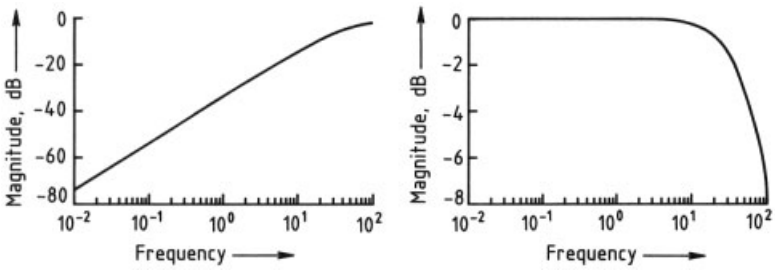

Figure 103. Singular values of the sensitivity matrix and complementary sensitivity matrix

timization. For example, the Ziegler – Nichols rules (see Section 4.1) yield optimal controller parameters for PI or PID controllers and a class of systems. If, on the other hand, the control scheme – including its structure – can be freely chosen, this is referred to as structural optimization.

In the case of parametric optimization, the numerical values of the variable parameters are assigned a numerical rating based on the performance index; this rating is to be maximized or minimized through the variation and is thus a function of the parameters. In structural optimization, the entire time dependence of the control function $u(t)$ in the time interval $t_0 \leq t \leq t_e$ is varied. The performance index assigns a numerical rating to this function and is therefore a functional (quality functional, cost functional) [344].

Usually, the term "optimal control" is understood as referring to the results of a structural optimization, and this meaning will apply in the discussion that follows.

Optimization Problem. The starting point is a mathematical description of the system in state space (see Section 3.3) in the form of the system equation:

$$\dot{x}(t) = f(x(t), u(t), t) \tag{4.3.1}$$

What is now sought is the control function $u(t)$ that yields the minimum value of the performance index J for the process by which the system goes from the initial state $x(t_0) = x_0$ to a final state $x(t_e) = x_e$. This is the optimal control function $u*(t)$. Solving the system equation with given x_0 and $u*(t)$ gives the optimal state – space trajectory $x*(t)$ associated with $u*(t)$.

The final time t_e may be fixed or subject to choice, and similarly for the final state x_e or its components. Quantities that can be freely selected are available for variation; that is, they can be incorporated in the performance index.

Along with specifications of this kind, there are often constraints: State or control variables cannot, or must not, go outside certain value ranges at any time during the transition. For example, a value can be at most completely open or completely closed, the concentration of a species cannot take on values outside the range from zero to unity, and a reactor temperature must not go above a certain upper boundary or below a certain lower boundary.

The problem thus comprises

1) The system equation

2) The initial state x_0
3) The performance index J
4) Possible specifications of t_e and x_e
5) Possible constraints on x and u

Open- and Closed-Loop Optimal Control. For a given initial state x_0, the solution of the optimization problem yields an optimal control function $u*(t)$. It does not, however, say anything about how the control function can be technically implemented. There are two approaches to doing so:

In open-loop optimal control, the system state x is measured once at time t_0 and $u*(t)$ is determined from the result. For the remainder of the time interval $t_0 < t \leq t_e$, $u*(t)$ is presented at the control inputs of the system; no further account is taken of the system state. The control function is thus a function of time only from time t_0 onwards; it can be generated, for example, by a computer, which merely plays the role of a function generator.

Such an open-loop structure has, however, the disadvantage that the system follows the optimal trajectory $x*(t)$ only if no disturbances occur (see Section 4.1). For this reason, open-loop optimal control is often replaced by closed-loop optimal control. The optimal controller continuously generates the control function based on the current system state. If there are no disturbances, the controller holds the system on the optimal trajectory, thus behaving in the same way as the open-loop optimal controller. If disturbances occur, resulting in departures of the system from the optimal trajectory, the controller will initiate corrections so as to offset these departures while the transition process is still under way.

Performance Indices. A general form of performance index, which includes many special cases of importance in engineering [307], is the functional

$$J = h\left(x\left(t_e\right), t_e\right) + \int_{t_0}^{t_e} f_0\left(x\left(t\right), u\left(t\right), t\right) \mathrm{d}t \qquad (4.3.2)$$

The first term in this Bolza performance index evaluates the system behavior at the end of the transition process, while the integral term evaluates the entire process. This is known as a Mayer performance index if $f_0 = 0$, and as a Lagrange index for $h = 0$.

The following cases occur under the Lagrange performance index:

- $f_0 = u^{\mathrm{T}}(t)Mu(t)$, the energy-optimal performance index
- $f_0 = x^{\mathrm{T}}(t)Lx(t)$, the ISE performance index (performance index of the integral squared error type)
- $f_0 = 1$, the time-optimal performance index

The diagonal elements of the matrices M and L evaluate the squares of the control and state variables, respectively. All the other matrix elements are often simply set equal to zero.

Methods. This section lists the most important methods and concepts, with references to the literature.

The classical method of minimizing functionals is calculus of variations [307, 345, 346]. This technique is suitable for solving optimization problems without constraints on the control and state variables. An important special case, the closed-loop optimal control of linear systems [322, 323, 347] with quadratic performance index

$$J = x^T\left(t_e\right) Sx\left(t_e\right) + \int_{t_0}^{t_e} \left(x^T\left(t\right) Lx\left(t\right) + u^T\left(t\right) M\, u\left(t\right)\right) \mathrm{d}t$$

leads to the so-called Riccati controller.

Systems with constraints on the control variables can be optimized with the Pontryagin maximum principle [348–350]. Such constraints arise, in particular, in time-optimal control problems where the solution is expected to yield control functions that alternate between the upper and lower extreme values of the control variables. Feldbaum's theorem states, for a given class of systems, how many such alternations are required at a maximum.

Dynamic programming [351–353] is a general optimization method that can be used for, among other purposes, the optimization of dynamic systems. It offers the possibility of utilizing the constraints not only on the control variables but also on the state variables. An essential foundation of dynamic programming is Bellman's principle of optimality, which says that any final portion of an optimal trajectory is itself an optimal trajectory. This principle implies the basic procedure: The optimal trajectory and the associated open-loop optimal control scheme are determined in successive backward steps, beginning with the final state x_e.

A thematically organized literature survey of the optimization topics discussed here, as well as others, can be found in [307]; see also [354–359].

4.4. Binary Control

Principles. Classical control theory is concerned chiefly with "analog" signals, which are discrete as to time and value. The mathematical description of the process ("plant") and of the automation device ("controller") employs differential equations as well as transformations (Fourier, Laplace).

Industrial control practice, on the other hand, deals with all devices used to influence plants and machinery in a purposeful way. Traditionally, however, the emphasis is on working with binary signals (e.g., in relay and contactor controls) in open-loop structures [360]. Analog control loops occur only as low-level elements. Boolean algebra and its adaptation to hardware (switching algebra) are employed for the mathematical description of binary control [361].

These contrasting approaches have historical roots: Analog thinking relates to early work on the automation of continuous operations such as those practiced in the process industries, while binary thinking grows out of attempts to automate piecewise operations such as those used in the production industries.

Process automation as now practiced in industry involves the convergence of both approaches. This applies to processes being automated and to the automation equipment itself, but not to the mathematical apparatus.

Operations. Operations that have long been automated by open-loop control methods are now being operated with closed-loop methods as well; an example is the motion programs of industrial robots.

Process engineering, a classical domain of closed-loop control, now makes increasing use of open-loop control, both for the automated startup and shutdown of continuous processes and for automated batch production.

At the same time, a trend toward higher-level optimization of the entire production process, with the goals of improved efficiency and flexibility, is apparent in all areas of industry. The result is generally coupling of subprocesses in the sense of material streams, energy, and information. Process automation thus calls for a holistic examination of continuous and discrete subprocesses, analog and binary process variables, and continuous and sequential operations. (This section does not deal with a third class, object-related operations [362], in which individual, identifiable objects pass through the process. As far as process engineering is concerned, this concept figures chiefly in high-level production automation; examples of "objects" are charges and production campaigns. See also Chap. 2 and Section 4.5.)

Automation Equipment. Just as continuous and discrete subprocesses have been linked into "hybrid" processes, automation equipment has also undergone a merger with classical measurement and control devices (Fig. 104 A) becoming components of integrated automation systems (Fig. 104 B). This development has, however, been driven less by user demand for integrated process automation than by technological advances in instrumentation and control. Mechanical, pneumatic, hydraulic, and electromechanical devices have gradually had to give way to microprocessor-based electronics. Hence the step from consistent instrumentation and control devices to integrated process automation systems was obvious.

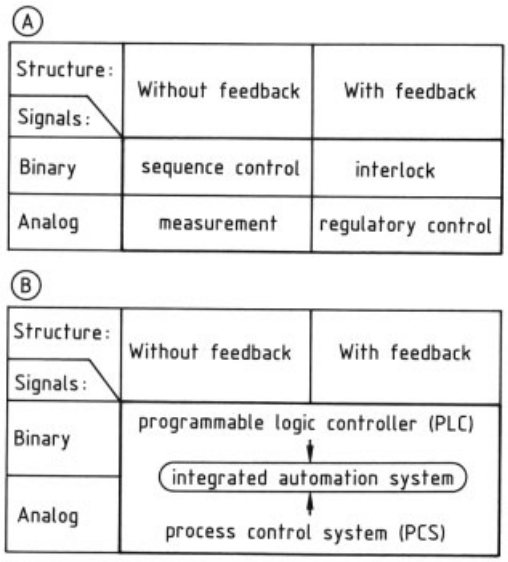

Figure 104. Application areas of automation systems
A) Classical automation devices; B) Modern automation systems

The adoption of these systems by users has been aided by the appearance of two variants that use software to mimic the operator interface of classical instruments and controls (Fig. 104 B):

1) The programmable logic controller (PLC) for users oriented toward open-loop control in the production industries. Programming may be performed graphically with ladder diagrams, which are similar to the schematic diagrams of relay-based control systems.
2) The process control system (PCS) for users oriented toward closed-loop control in the process industries. The PCS interface can, for example, imitate conventional control stations with live flow diagrams and controller front panels (see Section 5.2).

The combination of first-generation PCS and PLC into integrated "hybrid" process automation systems has led to a second generation of process control systems [363].

Mathematical Methods. A hybrid system consisting of a hybrid process and a hybrid automation system can be simplified by splitting it into a continuous, "analog" part and a discrete, "binary" part (Fig. 105). These subsystems have common structural properties: They contain internal feedback paths and must be treated as dynamic systems with memory elements.

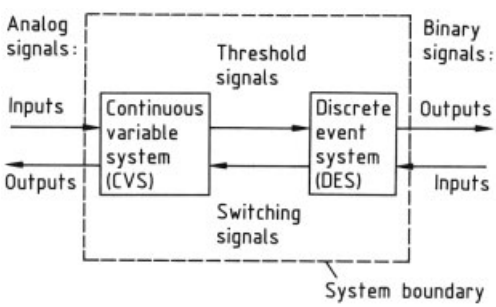

Figure 105. Hybrid system
Input and output signals connect the system to surrounding systems (e.g., systems on higher and lower levels; operator interfaces)

The mathematical methods applied to the two classes of systems have not yet converged [364], mainly because the theoretical techniques are at far different stages of development.

For continuous systems (and also for the closely related sampling systems), control theory offers a polished, complete, and proven mathematical toolkit. For the modeling, analysis, and synthesis of dynamic discrete event systems, in contrast, there is only a rudimentary selection of diverse methods, including switching algebra [361], Petri nets [365], Boolean differential calculus [366], and theory of automata [367].

Terminology. The foundation of binary control is switching algebra [361], which works on two-valued or logical variables. Its axioms establish the elementary logical operations conjunction (AND), disjunction (OR), and negation. Figure 106 shows schematic symbols for the hardware implementations of these operations. Other basic operations are derived from the elementary ones: NAND, NOR, equivalence, antivalence (exclusive OR or EXOR operation), and implication. Combining the basic operations generates switching functions, the implementations of which are called combinatorial circuits or logic controllers. The rules of switching algebra permit the mathematical analysis of networks. The Quine – McCluskey method can be used for the minimization of given logic functions [368].

If a combinatorial circuit contains internal feedback paths, memory elements (flipflops) can arise (Fig. 107). In this way, a sequential circuit or sequential controller is obtained. There are two classes of flipflops: level-operated and edge-triggered. In level-operated devices, the output can change its state as soon as one input state changes; in edge-triggered circuits, the output can change its state only when a certain (e.g., positive) edge is presented at a certain input, called the clock input. If, along with the logic functions, a sequential circuit contains only edge-triggered flipflops whose clock inputs are all driven by a central system clock, it is said to be a synchronous sequential circuit (since all the memory states can change only in synchrony with the edge of the clock pulse). Otherwise, the sequential circuit is said to be asynchronous.

A systematic design procedure exists for synchronous sequential circuits. The starting point is a graphical description of the problem, called the state diagram [369]. The technique does not, however, insure a minimal realization of the sequential circuit. State diagrams are also unsuitable for the analysis of complex circuits, since a

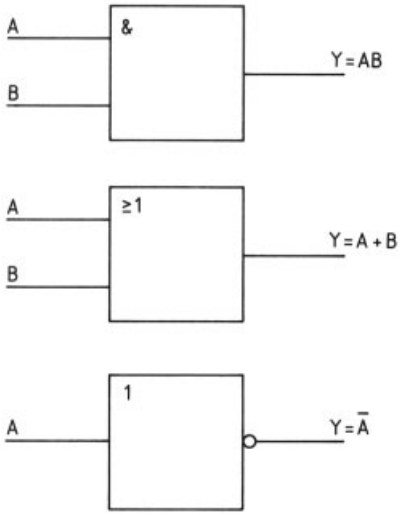

Figure 106. Schematic symbols for the elementary logical operations AND, OR, and negation, according to IEC 117-15
A, B inputs; Y output

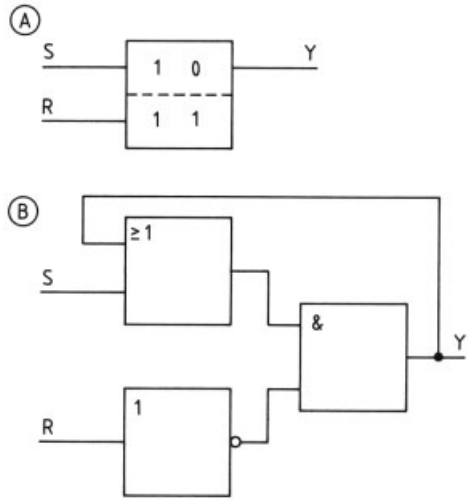

Figure 107. RS flipflop
A) Schematic symbol; B) Construction from elementary logical operations
S set input; R reset input (with priority over S)

sequential circuit with n memory elements can have up to 2^n states.

Besides the combinatorial and memory elements, there is a third type of basic functional element: timers, which shortens, lengthens, or delays binary impulse signals. They are of very minor importance in the theoretical analysis of binary control systems but are frequently used in practice.

The program of a control system means the establishment of all signal-processing functions. This can take either of two forms. In a wired-program controller, the functions are established by hardware connection of the individual functional elements. If the connections can be altered in a simple way, the controller is said to be reprogrammable; otherwise it is called fixed-programmed.

In a stored-program controller (or programmable logic controller, PLC), both the functional elements and their connections are simulated in software on a digital computer. The control program is thus part of the computer program. The program for a RAM-programmed controller is stored in read – write memory (random-access memory, RAM); that for a programmable controller with exchangeable memory is stored in read-only memory (ROM).

A PLC includes a digital computer and is thus essentially a synchronous processor. It can, however, be used to simulate combinatorial circuits, asynchronous and synchronous sequential circuits, timers, and also analog signal-processing elements such as controllers, or any combination of these, to a very good approximation. It accordingly represents the most flexible and advanced way of solving all control problems, and it is the most cost-effective technique for all complex control problems.

A number of text-oriented and graphics-oriented languages are available for the specification of control tasks and the programming of PLCs [361]:

- Structured text (ST; IEC 1131/3) is particularly suitable for writing implementation-neutral specifications.
- The instruction list (IL; DIN 19 239, VDI 2880) is a low-level assembler-like PLC programming language (i.e., similar to machine language).
- With the ladder diagram (LD; DIN 19 239, VDI 2880), simulations of relay-based circuit diagrams are employed for PLC programming. This is a graphical language.
- The function block diagram (FBD; DIN 19 239, DIN 40 719, IEC 117-15) technique, another graphical language, uses elements similar to the circuit symbols for logic gates and memory elements in digi-

tal schematics (see Fig. 108). If individual logic and memory elements in the logic diagram are combined into macros ("actions" and "steps"; see Fig. 108), the result is a sequential function chart (SFC).

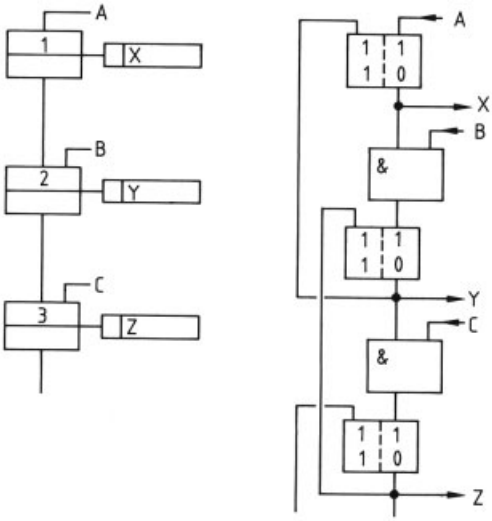

Figure 108. Linear step sequence
Macro representation: sequential function chart (left); Explicit representation (right) 1, 2, 3 steps; A, B, C transitions; X, Y, Z actions

– The sequential function chart (DIN 40 719/6, VDI 2880) is especially suitable for sequential controls where the individual control goes through states or steps not in an arbitrary sequence but (at least section-wise) in a fixed, linear sequence. This condition does usually hold in practice.

Besides these standardized languages, there are many manufacturer-specific languages, partially constructed from elements of the standardized languages (e.g., Siemens STEP-5).

Sequential function charts include not only sections of linear step sequences, but also branches. A distinction is made between OR (alternative) branches and AND (parallel) branches. In the latter, concurrent control sequences take place in the controller; that is, sequences that proceed in parallel and independently of one another until they are again synchronized to one another at the end of the AND branch. Such controls cannot be usefully diagrammed and analyzed with state diagrams; a much better tool for this purpose is offered by Petri nets [365]. The draft of DIN 40 719/60 for sequential function charts, in contrast to the old version of DIN 40 719/6, includes some important restrictions on branching options to permit Petri net modeling of concurrent controls.

Both Petri nets and sequential function charts are suitable for the specification, simulation, analysis, programming, operator interface, and documentation of controllers. They are particularly useful as support for hierarchical control concepts, which are often important in practice.

Classification of Control Systems. The common language of automation technology includes many control terms, most of which are defined in DIN 19 237 and many of which were mentioned in the preceding section. In what follows, these terms are listed under different classification criteria.

– Signal-oriented classification: analog, digital, binary control
– Function-oriented classification: combinatorial, sequential control
– Structure-oriented classification: asynchronous, synchronous control
– Technology-oriented classification: wired-program control, stored-program (programmable) control; (wired-program control:) fixed-programmed, reprogrammable; (stored-program control:) RAM-programmable, exchangeable memory; relay, contactor, electronic control; mechanical, hydraulic, pneumatic, fluidic control, etc.
– Hierarchy-oriented classification: individual, group, coordinating control
– Application-oriented classification: interlock control, sequence control; burner, scale, transport, recipe control, etc.

Another concept is mode control, which refers to a control level, higher than the signal-processing part of a sequential control, in which an operating mode is selected. VDI/VDE 3683 and DIN 19 237 define the following modes (among others): automatic, semiautomatic, single step, step setting, manual. In hierarchical control concepts, the operating modes available may differ from one level of the hierarchy to another and must be coordinated in a suitable way.

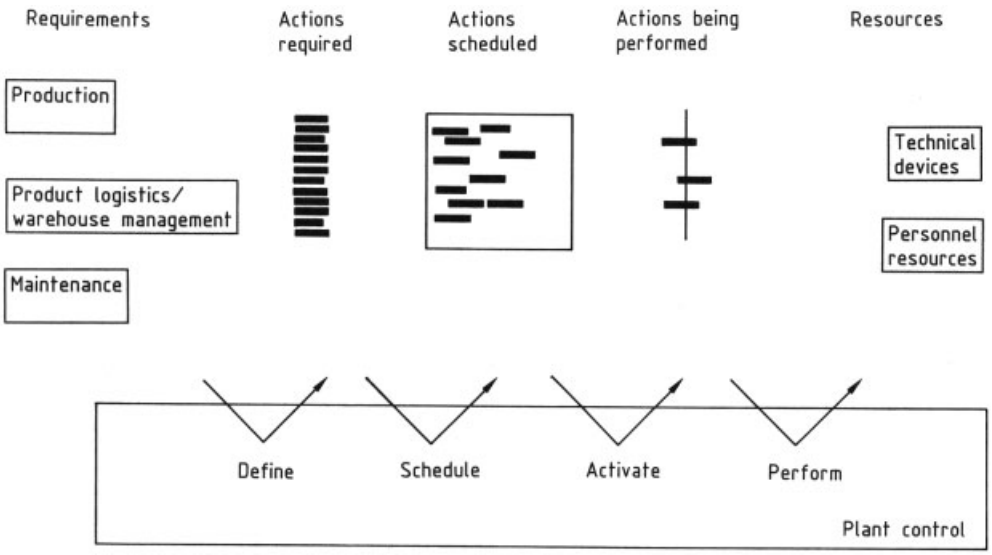

Figure 109. Actions as a means of plant control

4.5. Operational Control of Process Plants

Requirements on the operational control of process plants have grown much more stringent. The reasons include increasing flexibility in production, shortages in personnel and plant resources, shorter product throughput times, production structures with cross-links due to energy and product coupling, quality assurance standards, and the greater and greater number of maintenance operations that have to be performed at regular intervals. Optimized production programs and maintenance practices must be integrated into a plant control concept against the background of environmental and occupational safety and health requirements and the reliability standards that they imply.

A number of approaches have been devised and used with success for specific problems.

Brombacher [370] and Niese [371,372] used the term "process logistics" to describe proposals they developed for managing product streams given the coupling of process elements in normal operation, but also in transient and exceptional situations.

Under the leadership of NAMUR, "recipe" control concepts were developed for batch processes; these have been implemented as standard products by control system manufacturers [373].

Riemer has described concepts for production control, which have been turned into tools [374].

The same holds for special functions at the plant complex control level [375].

A closer analysis reveals that these concepts do accomplish their specific goals but are not consistent with one another in the sense of operational process control.

In future, accordingly, a broader approach must be taken to the operational control of process plants.

The action and resource model discussed here is one such expanded approach to describing the operational control of a process plant. The model is constructed in an object-oriented way. A wide variety of requirements, relating to recipe-based process control, logistical product management, maintenance, and other areas, can be reflected in the model.

The model includes a general formalism for resource allocation and control. Automated tools support the application of the formalism. In the context of the level model, the action model represents a logical standard interface between the production or plant complex level and the process control level.

Introduction. The fundamental task of operational process control is to coordinate and, with the aid of plant resources, carry out various

tasks stemming from production requirements, in-plant product logistics, or technical support and plant maintenance.

The action and resource model provides a formalism for systematically finding and implementing solutions.

The action model is based on the idea that every task arising in operational plant control can be completely accomplished by performing individual actions. These actions are performed through orders on the available resources.

The process control problem is thus broken down into three parts:

1) Mapping of tasks into individual actions
2) Planning and scheduling of these actions
3) Control and monitoring of the performance of the actions

Figure 109 illustrates the basic structure of the model.

Some important concepts figuring in the model are explained first. These should be understood as working concepts and are used here only in a very restricted sense.

Functional Units. A functional unit is, quite generally, a unit aggregated from various functional elements and, as a whole, accomplishing a self-contained task. Functional elements are selected such that they can be directly implemented with the help of existing plant equipment. Thus they are not fictitious functional objects but concretely represent the capability of a group of process devices. A plant complex includes a variety of functional units, such as plant sections, maintenance foreman groups, auto transport services, energy sources, and so forth.

In general, functional units are organized in a hierarchical manner (see Chap. 2).

Maintenance groups, for example, are organized as follows:

Maintenance technician
Maintenance gang
Maintenance foreman group
Maintenance support group

Process-oriented functional units might be organized as follows:

Individual functional unit
Group functional unit
Plant section unit
Plant unit
Plant complex

The internal structure of these units is discussed in depth below.

Resources. In the concept discussed here "resources" means "that which can perform services in the plant." Resources correspond to functional units that do not functionally overlap and which together represent the capability of the plant. Resources so defined are "intelligent" units to which the plant can pass on orders either directly or through plant actions. Resources are part of the permanent equipment of a plant complex; they are present whether they are processing an order or not. The sum of the resources reflects the functional capability of a plant complex.

Several special rules apply to resources:

– Self-sufficiency:
  Each resource has the capability of managing itself and independently carrying out orders. For the real performance of its task, each resource has its own apparatus and resources to which it has unrestricted access. Each resource can thus be planned, scheduled, and worked independently of all other resources.
– Control interface:
  Each resource has a control interface and an allocation mechanism associated with the interface. A resource may be needed for a variety of plant actions. At any time, however, only one action is allocated to a resource by means of orders. The planning and scheduling of resource utilization is accomplished by higher-level functions. Independently of such planning and scheduling, however, each resource has an allocation mechanism, which locally protects it against improperly being assigned more than one order.
– No parallel processing of orders:
  A specific feature of this model is that each resource can process only one order at any time. Accordingly, there is no concurrent working of orders within a resource. This provision is crucial in determining the overall structure of operational process control.

This "flatness" and the ban on parallel processing of orders are an important advantage

in making operational control easy to understand and simple to implement, but they represent a severe restriction on structuring. Functional units, as substantial objects, can be aggregated in accordance with the structural principle. The rules for resource formation now require that one level of aggregation be selected and permanently identified as the resource level used for operational process control. Only the functional units on this level are competent to perform operational process control services. If, for example, the maintenance foreman group is identified as the resource for the maintenance hierarchy, the operational process control system can then issue orders only to the foreman group. The issuance of an order to a gang or to a support group would not be permitted. In this example, the rules would demand that a foreman group be processing only one order at any time. This is, however, generally not so in practice. Accordingly, a foreman group is not the proper resource for operational process control. The resources identified should be those on a lower level of aggregation. Gangs, in general, process only one order at any time; hence gangs meet the criteria for a resource and would be a suitable choice as schedulable units for operational control of the maintenance area.

It is important to stress that this restriction to one level of aggregation relates solely to operational process control. All the levels of aggregation are still of interest from the standpoint of other organizational concerns.

Actions. The plant complex can accomplish its task by issuing orders directly to its resources. Such a manual procedure is, however, expensive for the plant complex and also hard to track. The action model offers a more convenient solution. An action is a self-contained control unit that permits the automatic performance of an entire functionally coherent block of tasks. Typical actions in this sense might be

- "Pump X liters from tank A to tank B,"
- "Do the scheduled maintenance on emergency trip system X,"
- or
- "Produce X tons of A in accordance with recipe B."

For each action there is a performance specification, which sets forth how the action is to be carried out. Any action may require more than one resource for its performance.

Actions can be combined and grouped at the plant complex control level. At the interface to the lower-order process control system, however, each action is essentially a self-sufficient, independent, self-contained object. The model says nothing about the complexity of an individual action. Both extremely complex actions and elementary actions can be defined and carried out simultaneously in a system. Actions need not be free of redundancy as to their functionality. For example, an action such as "transfer X liters from rail transport to tank A" may well have the same elements as the action "Pump X liters from tank B to tank A." Regardless of the nature of the task or the type of resources employed to accomplish it, a common scheme is used for performing and managing actions. In this way, they form an object class with well-defined properties. In the hierarchy, actions form a standard interface between requirements on plant management and operational performance. This formal decoupling is an essential goal of the action concept, but it should not be taken to imply that task and execution can be analyzed and organized independently of one another. If the structures of resources and those of tasks are not consistent, the result will be a poorly organized package of actions that cannot be efficiently carried out.

Thus it makes no sense, for example, to define special plant engineering resources simply for purposes of process control if such resources cannot be utilized to perform maintenance actions or contribute to in-plant product logistics.

Central importance attaches to the development of a resource structure that is matched (in substance and form) to the internal concerns of various task areas.

Requirements on Operational Plant Control. Operational plant control must meet requirements arising from many areas. These are broken down by the concerns of the areas where they originate. The next section presents examples of "Requirements from Production Planning," "Requirements from In-Plant Product Logistics," and "Requirements from Plant Support" in order to show how actions can be derived from requirements.

Requirements from Production Planning. From production planning, the plant complex re-

ceives orders to produce certain products, along with product-related specifications such as product name, nominal quantity, nominal quality, and nominal deadline. In order to accomplish this production, a number of actions must be initiated and carried to completion. These actions can be analyzed in two ways:

- An operational breakdown of the production process into steps corresponding to the subdivisions of the process
- A purely quantitative breakdown of the production process into lots making up the desired (nominal) quantity
- Operational breakdown of the production process

For structuring purposes, a production process can be broken down by using the phase model (see Chap. 2, Figs. 33, 34, 35, 36, 37, 38). Figure 110 shows an example for a process A. At the left, the entire process is shown as a closed block with inlet products IP1 to IP3 and outlet products OP1 and OP2. The guiding principle in breaking down the process is that the operations belonging to a process can be separated by simply dividing up the material system by product and dividing up each operation into phases by time.

If this is possible, individual, self-contained subprocesses corresponding to the phases of the phase model are obtained. Each subprocess is characterized by the performance of certain operations. In the example shown, these operations are reaction A1, separation A2, reaction A3, mixing A4, and reaction A5. The products at the beginning and end of a subprocess have well-defined properties. In the example, these products include the inlet and outlet products of the overall process plus intermediate products ZP1 to ZP4. The subprocesses are coupled to one another via these products.

Depending on the complexity of the overall process, the same scheme can be re-applied to break down the subprocesses.

This method reaches its limit when operations cannot be further separated either by a product-related subdivision of the material system or by a time subdivision into phases.

Production specifications are instructions for the execution of a production in a plant. In other words, they are performance specifications for production actions. They extend beyond general process descriptions, stating precise conditions to be maintained and states to be established in order to carry out a certain process in a plant.

These conditions include, for example, properties of products that enter the process. The starting material present in the plant at the beginning of a process should also be regarded as a product occurring in the process.

An important initial condition is the state of the plant at the beginning of the process (see also Section 2.3). Initial condition refers not only to state variables in the narrow sense, such as temperature or control state, but also to general properties such as "degree of contamination," "functional reserve," "modification effected," and so on. These conditions govern the need for tooling or cleaning orders.

The fundamental part of the production specification, however, describes how certain process properties must be controlled in the course of the process so that the desired overall process can be carried out. This also includes the establishment of certain intermediate and end product properties [376]. Each production specification is based on a model of a reproducible process.

By analogy to the breakdown of a process, a production specification can also be divided into parts (see Fig. 25). It should be noted that every portion of the specification, by virtue of the analysis, contains supplementary conditions on the inlet product (feed) properties and on the states of plant sections at the beginning of the subprocess.

When such a breakdown is carried out, a production specification becomes a number of subspecifications, each of which describes the execution of a subprocess in a plant section (see Fig. 26).

It is evident that the analysis of processes and production specifications has a direct effect on the structure of action objects and the way they are managed. This can be illustrated by two extreme situations as shown in Figure 111.

In Example 1, the process has not been broken into parts. In other words, the action takes in all partial production steps including intermediate product transport. The overall production process, including the time sequence, is described by a single production specification and, in its entirety, corresponds to a single action.

In Example 2, the overall process is divided into subprocesses, for each of which there is

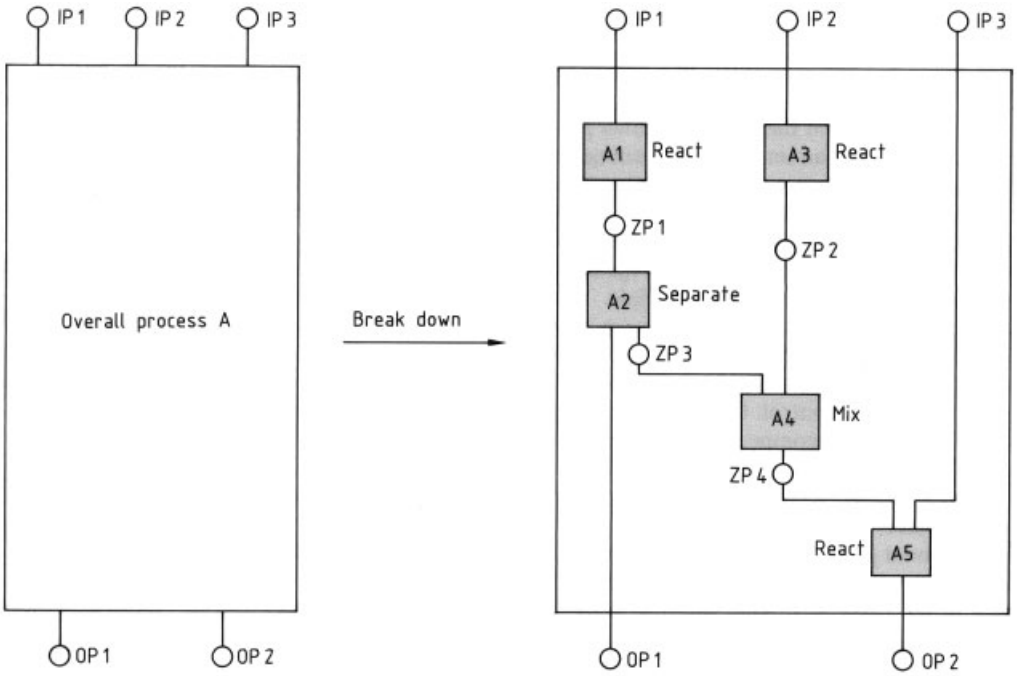

Figure 110. Breakdown of a process in the phase model

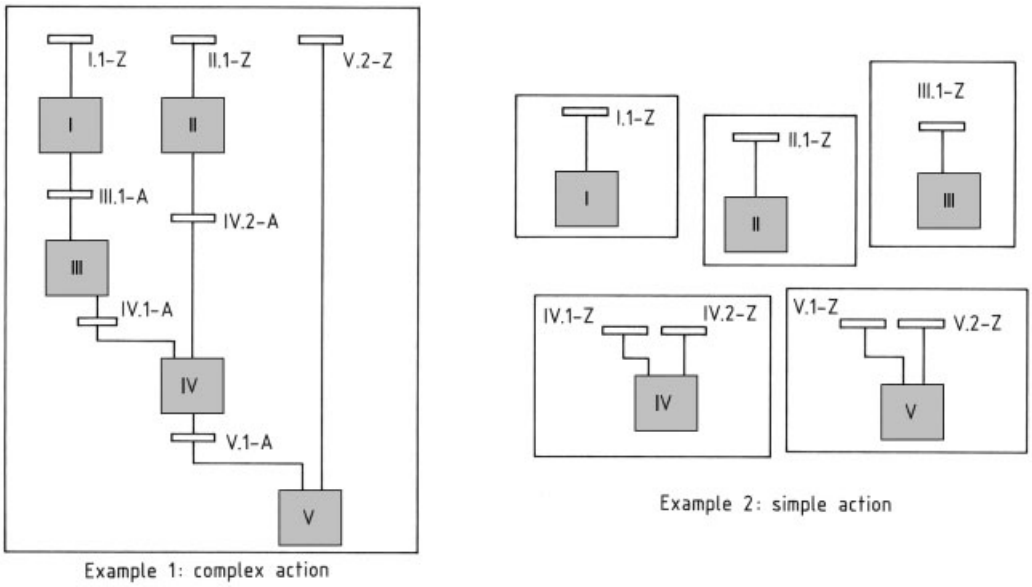

Figure 111. Complex and simple actions
Z = prerequisites strictly state-oriented; A = prerequisites also sequence-oriented

a partial production specification independent of the other subprocesses. An independently schedulable action is assigned to each subprocess. Temporal coordination is achieved through the availability of intermediate products and the initial conditions for the individual subprocesses. Coordination comes under the planning

and scheduling functions of the production or plant complex control level.

What level of "granularity" is appropriate for a given production process will depend on the requirements and constraints imposed. The action concept permits arbitrarily broad or granular actions and does not place any restriction on the depth of the breakdown.

– Quantitative breakdown of the production process

If the nominal product quantity is too large to be produced in one step in a closed material system, the overall production can be subdivided into individual batches. A number of batches can be combined into a lot, and a number of lots into a campaign. The NAMUR definition states that a lot is "a delimitable and identifiable quantity of product having a uniform quality. In the case of a batch production process, a lot may include one or more batches" [373].

Again, the action concept allows any degree of aggregation. In general, actions will relate to the production of one batch; the combination into lots is a planning and scheduling task. In the extreme case, however, a complete lot can be defined as a single action.

Requirements from In-Plant Product Logistics. In-plant product logistics refers to the control of product (material) flow within the plant complex. This activity supplies inlet products (feeds) to the "producing" plant sections and removes the end products and byproducts. The tasks of in-plant product logistics also include the control of internal buffers, tank storage, bunkers, high-bay warehouses, and devices used for storage and reclaiming. This area of responsibility may also extend to the control of tank cars, rail cars, and other mobile storage located on-battery.

Depending on the application, product-logistics activities may take place in the context of production actions or may be independently planned and scheduled tasks executed by separate actions. Such actions might then be:

– "Transfer by pumping,"
– "Convey from tank X to tank Y,"
– "Transfer from tank car unloading station A to tank X,"
– or the like.

These actions are required to be highly flexible as to scheduling. Often the orders are not issued until shortly before performance, or rescheduling may be necessary at the last minute (e.g., tank cars are in the wrong sequence on the track, operators at the tank car terminal wish to carry out order A before order B, etc.). This required flexibility in the planning of logistic actions and the requirement that noncentral (local) entities must have access to scheduling (e.g., from the receiving station) must be taken into account in the design of the scheduling system.

For higher-level product management, the granular logistic elements are combined into more complex units. For example, starting products may be delivered to the plant complex, whereby each delivery may consist of a number of tank cars. Within the plant complex, not the receipt of the delivery as a whole but rather the discharge of an individual tank car is generally regarded as a single action. Again, however, the action concept is open and permits modification to suit the organizational setup of the plant. (For example, a complete coal train is unloaded in a different way from a delivery made up of separately marshalled cars.)

Requirements from Plant Support. Plant support, in order to accomplish its tasks, must periodically carry out plant inspection, maintenance, and repair operations. These can also be defined and performed as actions, for example:

– "Calibrate probe X,"
– "Drain separator sump,"
– "Do scheduled maintenance on pump A,"
– "Inspect overcharging protection B,"
– and so forth.

If a maintenance action renders plant equipment unavailable for production while the action is being performed, the maintenance action must be included in the planning and scheduling system like any production action. The action concept permits a systematic treatment of maintenance actions in resource allocation. What is more, it allows plant-wide monitoring of the progress of planned maintenance. Every maintenance operation defined as an action in the system monitors its own correct performance. If unplanned events occur, the progress status of the action should reflect them. If necessary, an explicit notification should be provided to plant management or the planning and scheduling system (see Fig. 20).

Maintenance operations can be broken down in two ways: (1) by the functions or functional units maintained and (2) by expected frequency, for example, on a time or wear basis. Ultimately, these analyses lead to individual activity elements. Job planning combines these elements into packages meaningful from the engineering, functional, and management standpoints, and the packages are treated as maintenance actions for planning, scheduling, and administrative purposes. How this grouping of elements takes place in a particular case will depend in the maintenance concept practiced at the plant complex (see Section 11.4).

Analysis of Plant Equipment and Definition of Resource Structure. The plant complex has technical equipment and personnel resources at its disposal to perform its tasks. This section describes how the apparatus and personnel resources can be broken down and how they can be combined into technologically meaningful functional units. On the basis of the system of functional units created in this way, a suitable resource structure can then be established.

This description is for the example of plant engineering functional units but can be extended to other areas as well.

Breakdown of Plant Equipment. The purpose of analyzing equipment and apparatus available in the plant complex is to identify clearly delimited, nonoverlapping elements. Elements belonging to various devices can then be assembled into meaningful functional units. The process is illustrated for the combination of plant equipment elements and control system elements into plant engineering functional units.

– Breakdown of Process Equipment
  Production takes place in process equipment and facilities where the product is conveyed, stored, and transformed. A process plant or facility ultimately consists of valves, pumps, pipes, and so on. Between the plant as a whole and these single elements, however, a number of useful analytical levels can also be found. An important level is the plant section. From some points of view, the plant section is actually more important than the plant. Plant sections are typically delimited in such a way that further subdivision of the material system present in them is impossible or needless. Thus exactly one subprocess can run in a self-sufficient way within every plant section. Here "exactly" means that one, but only one, subprocess can take place in any plant section at a given time.
  A plant section has inlets where products enter it and outlets where products leave it. These inlets and outlets serve to connect plant sections to one another. The topological structure implies certain rules for product streams between plant sections. Figure 112 shows the most important basic ways of connecting plant sections. In the train, a product moves through one plant section after another (stage by stage). If there are several trains, a product moves through only one of the trains. In present-day chemical plants, especially plants where several different products are made, these structures are no longer found in pure form. Instead, the plant sections are connected by a complicated network of piping. Figure 113 gives an example of such a complex structure. A plant section consists of the components exclusively assigned to it, such as valves, pumps, sensors, piping connections, and so on.
  Within a plant section these components can be assembled into groups. In general, a group has the property that it effects a certain flow of matter or energy in the plant section. Figure 114 shows a reactor with its groups.
  Each group consists of the components exclusively assigned to it.
  Groups are coupled to one another by transfer points for matter or energy, called connections.
– Breakdown of Process Control
  The goal of process control is to bring about certain process properties. A certain control logic is required for this purpose. The basic structure of this control logic is described by a control model.
  The basis for the control model employed here is a hierarchical control model. As Figure 115 shows, a control unit receives orders, commands, and setpoints from a higher-level unit. It executes the order by, in turn, specifying orders, control values, control parameters, and commands to lower-order control units. How it does this is an aspect of its internal logic. The methods of feedforward and feedback control can be used to derive a suitable internal logic.

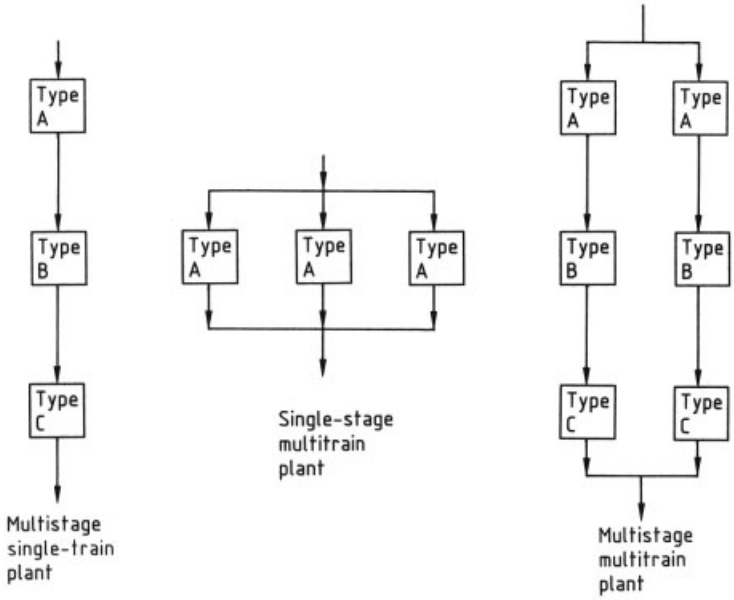

Figure 112. Basic connection patterns of plant sections

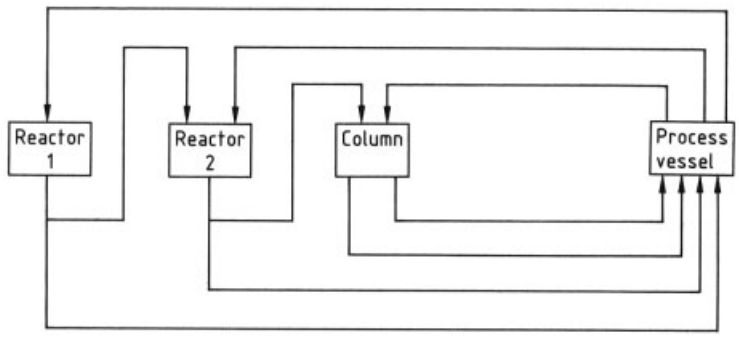

Figure 113. Real network of connections between plant sections

Grouping of Functional Units. The purpose of forming functional groups is to combine plant equipment in various areas in such a way that it together accomplishes a partial task of the plant. Workable modules for accomplishing the overall task of the plant are obtained in this manner. This is illustrated by an example of the grouping of plant units and control units.

Plant (equipment) units and control units can be combined into functional units. The model provides that one control unit is assigned to each plant unit.

At the lowest level, for example, individual control elements such as valves, pumps, electrical heating devices, and so forth, are grouped with individual control units to form "individual functional units." Such an individual functional unit for the blocking and release of a product stream is illustrated in Figure 116.

Each individual functional unit permits a certain process property to be set (see also Section 6.2).

The second level comprises group functional units. Figure 117 shows schematically how a group functional unit is built up. A group func-

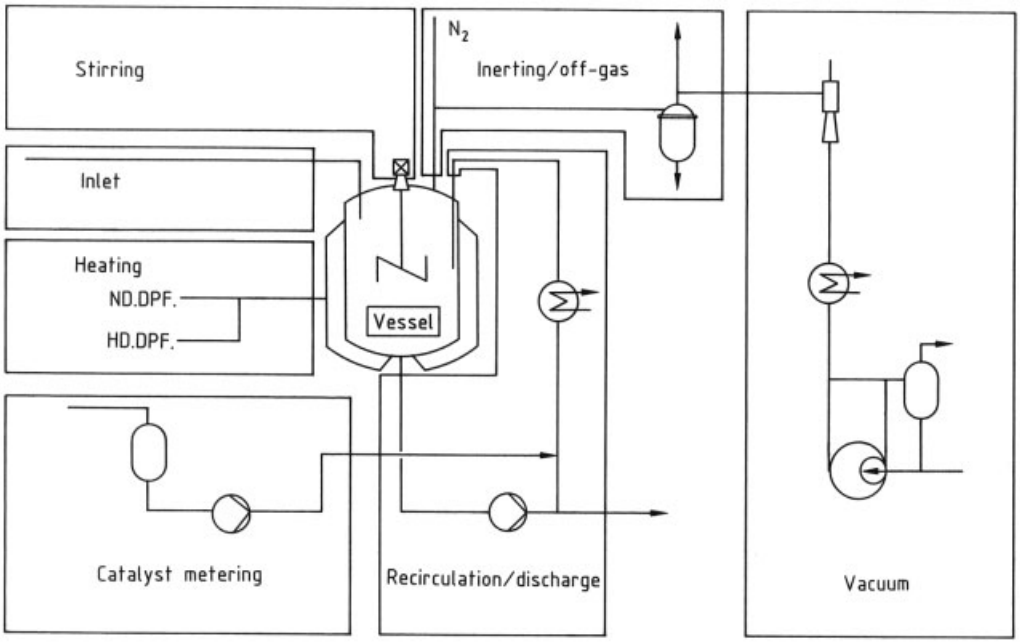

Figure 114. Groups of a reactor

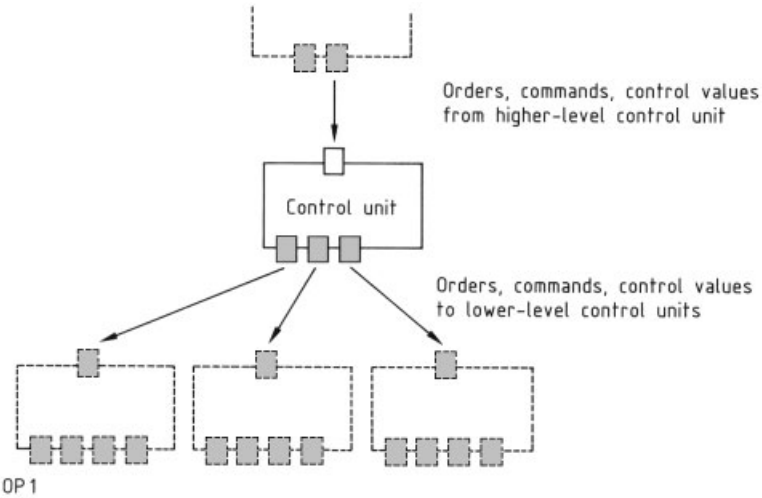

Figure 115. Hierarchical control model

tional unit consists of a group of individual functional units exclusively assigned to it, together with a group control unit, which controls and coordinates the activities of the individual functional units.

The overall goal of a group functional unit is to combine a group of individual functional units from the control standpoint in such a way that together they act on the process in a manner that is meaningful from the process engineering viewpoint. Individual effects that do not add up

to an effect meaningful in terms of process engineering are excluded by the group functional unit.

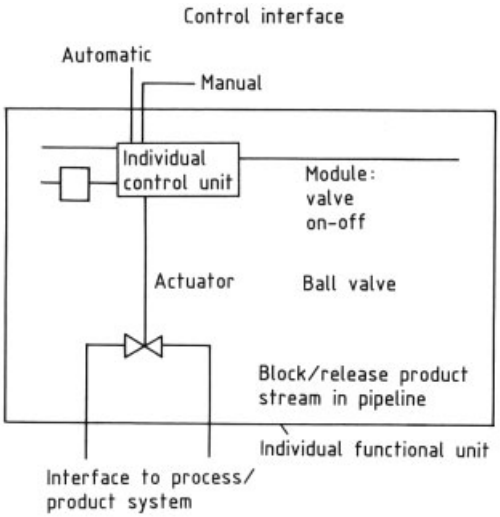

Figure 116. Structure of an individual functional unit

If several effects useful from the process engineering standpoint can be brought about with a group of individual functional units, the group control unit supports each of these effects with its own control structure. The group functional unit is said to have alternative or selectable "modes."

The functionality of the group functional unit is implemented when the group control unit issues specifications to its assigned individual control units in accordance with the control model.

Figure 118 illustrates the structure of a group control unit.

Note that the group functional units discussed here have also been described under the name "basic functions" [377, 378]. Because this term leads to repeated misunderstandings, however, and has been under discussion in NAMUR [373], it is not employed here.

The third level is that of plant section functional units. A plant section functional unit consists of the group functional units of the plant section and a plant section control unit. The plant section control unit reflects all the process control capabilities of the plant section. Typically, group functional units permit an arbitrary number of meaningful program options. Which option is needed in a particular case is set forth in the performance order for the individual process. In this way, when issuing an order to the plant

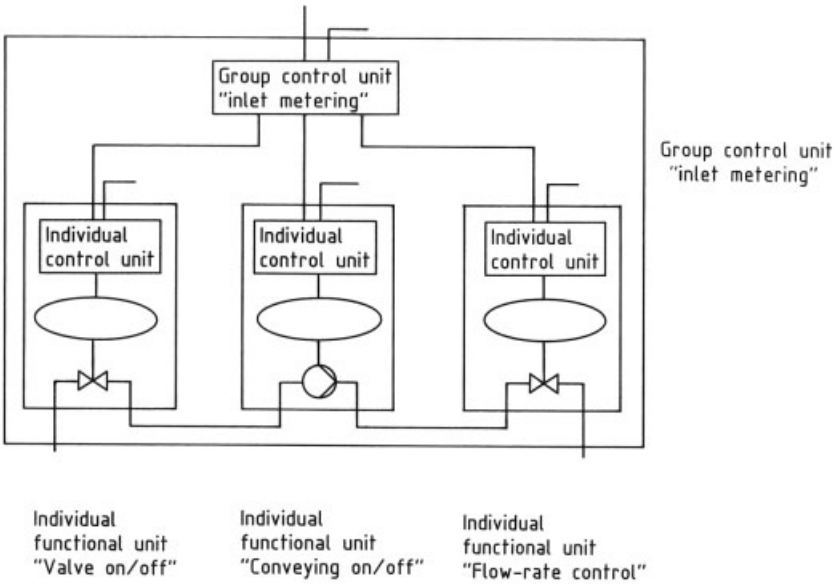

Figure 117. Structure of a group functional unit

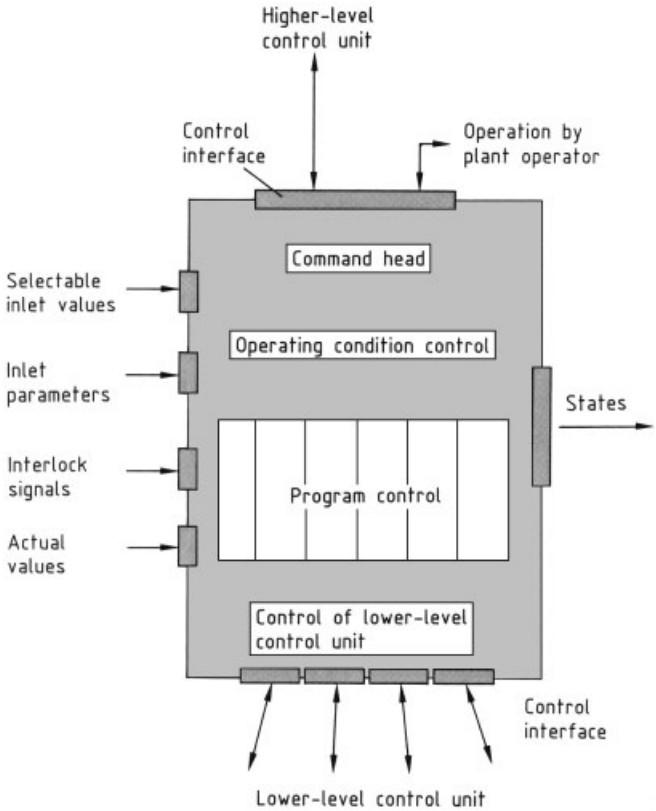

Figure 118. Structure of a group control unit

section control unit, it is necessary to provide not only control values, control parameters, and commands, but also portions of the performance specification.

To sum up:

Each level offers certain possibilities for functional intervention in the process; that is, each level can carry out certain process control orders. Furthermore, each level completely reflects the process control capability of the plant.

The higher the level, the more complex the task. In the simplest case, the order is to open or close a valve or to set a certain flow rate.

Every functional unit contains control units that carry out specific tasks. A control unit receives orders and/or control values from a control unit on the next higher level and, in turn, issues orders and/or control values as setpoints to control units on the next lower level. The number of levels in the hierarchy can be selected on technological grounds. Criteria might be, for example, the size and complexity of the process plant or facility. For the reference model under consideration here, a two-level structure below the plant section is chosen.

Determination of a Suitable Resource Structure. Establishing a suitable resource structure means selecting an appropriate set from the many functional units for operational plant control. The following criteria must be taken into account:

- The set chosen must fully cover the plant properties relevant to operational plant control.
- The functional units in the chosen set must not overlap. No resource may contain devices belonging to another resource.
- The structure of the resources must have a modularity suitable to the requirements.

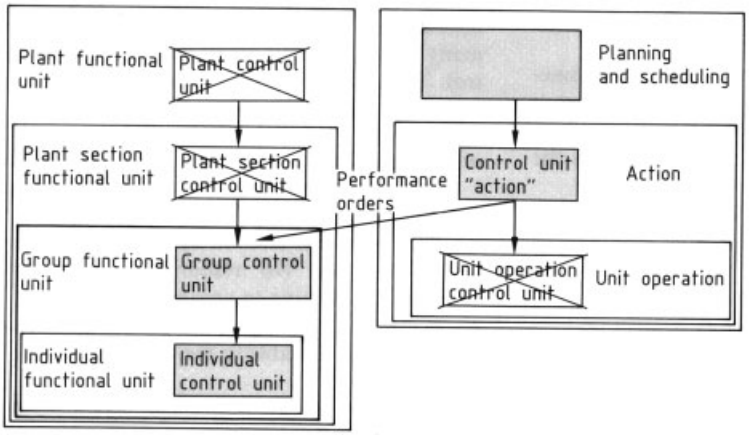

Figure 119. Control hierarchy

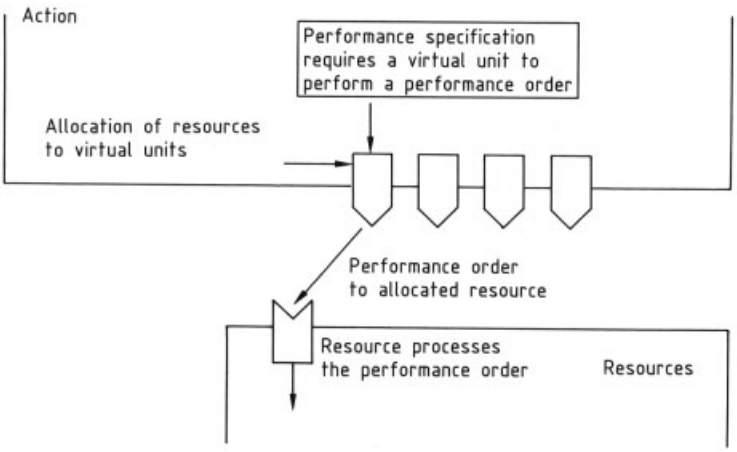

Figure 120. From performance specification to implementation

In what follows, possible resource structures are discussed for the example of plant functional units.

The criteria of completeness and freedom from overlaps are satisfied at the levels of individual functional units, group functional units, and plant section functional units.

Accordingly, the level most suitable for analysis can be selected.

Maximum flexibility is gained by selecting individual functional units as resources, though at the cost of involved performance specifications. The fact that there are group control functions, which can coordinate the individual functional units independently of the action being carried out, is not included in the treatment. The coordination of individual functional units must be explicitly described in each concrete case by means of instructions contained in the performance specifications for the action.

If group functional units are chosen as resources, virtually all the order-neutral conditions and program steps are contained in the internal logic of the resources. Simple and powerful performance specifications result. Group functional units are an appropriate choice. In the model

under consideration, group functional units are identified as plant engineering resources. In this way, every order involved in operational process control references one group functional unit. Orders to individual functional groups are no longer possible. Plant sections are not known to the operational plant control system, and there are no plant section control units.

Figure 119 shows the control hierarchy that results from identifying group functional units as resources.

In some process control concepts in use at present, plant section functional units instead of group functional units are the resources for allocation, planning, and scheduling. This notion is suggested by the rule: "At any time, only one subprocess can take place in a plant section." In practice, however, this approach results in some difficulties and disadvantages. Such problems arise, for example, in managing the interconnection of plant sections. Other difficulties come about when actions in the areas of in-plant product logistics or plant support are considered; here it is a hindrance to assume that one, and only one, action can be performed in a plant section at any time. For example, given certain constraints it is quite possible that production, calibration, and maintenance actions can be carried out simultaneously in a plant section.

A component-oriented control structure leads to further difficulties as well. In multifunction facilities, plant sections often carry out many different subprocesses. The needed control specifications can no longer all be established as "modes" embodying all the capabilities of the plant. They must be loaded into the plant section control systems before the subprocesses are performed. This procedure of loading program sections into existing control units entails serious conceptual problems.

On the whole, there is no advantage in the functional coupling of group functional units to form plant section functional units that can be allocated. The reference model accordingly looks directly to group functional units as the allocatable plant engineering resources.

Regardless of the resource structure, the plant section naturally plays a major, if not the central, role as the smallest unit of the plant engineering world.

Consequences of the Choice of Resources. The selection of a functional unit as resource not only implies additional formal requirements on the functional unit itself, but also has a strong effect on the semantics of the performance specifications for actions and on the overall structure of the plant control system. These consequences must be considered when making the choice.

Additional Formal Requirements. Resources in this model are self-sufficient, largely self-managing, and self-protecting objects. All resources must exhibit certain standard properties. In this way, the issuance of an order to a resource is standardized and is formally independent of the type of resource. An order to a maintenance gang does not differ formally from the order "heat" given to a group functional unit.

The formal provisions include the following standard properties of a resource:

Standard order interface. Each resource must have an interface through which it can receive orders, setpoint values, control parameters, and so forth. The interface is standardized and is, formally, the same for all resources. It is dedicated to communications with the allocating customer (e.g., action X, action Y, manual intervention).

Standard status interface. Each resource must have an interface at which it makes certain (generally standardized) significant states available as information for other units. This interface provides information to all interest units and is held available by the resource as a read-only interface. It is formally standardized and is the same for all resources. The value ranges of the state variables "availability state," "type of allocation," and "operating state" are consistent for all types of resources.

Allocation logic. Each resource protects itself against multiple allocations. The allocation logic uses a simple principle: The first to allocate an available resource is the customer. A customer must release the resource in order to make it available to other potential customers. Prioritization and the identification of optimal allocation sequences are in the planning and scheduling domain; these operations are not supported by the resource itself. This simple structure appears to be especially appropriate for process engineering, where briefly interrupting one production process to make the resource used for it available for another task is generally not desirable.

Self-management and self-organized activity. Each resource is a functional unit that manages itself and operates in a manner internally independent of an allocating customer. Even unallocated resources carry out their control tasks and are continuously active and operational. Resources are responsible for the correct functioning of the equipment subordinate to them and the management of their resources. They are intelligent alarm and report generators.
Self-sufficient protective logic. Each resource has protective logic to prevent its creating a forbidden condition. This protective logic operates under its own rules, independently of the order system, and insures priority local protection.

Semantics of Performance Specifications. Once the resources have been chosen, the reference elements are identified for the formulation of actions. The selection of resources thus influences the way of thinking and communicating during operation. The importance of this fact is commonly underestimated when specialists analyze their individual areas.

The linguistic consequences can be illustrated for plant engineering resources.

Three examples of plant engineering resources are considered.

If the input/output channel is chosen as the resource, the performance specification will contain elements such as "set output BA43," "read input AE65," and so forth.

But then both the EMR/process control system engineer who writes an automation function and the chemist who formulates a recipe must ultimately do so on this linguistic level.

It is obvious that this "channel level" is unsuitable as a medium of description.

If individual functional units are selected as resources, the requirements can be formulated in engineering language. The performance specifications now contain elements such as "pump 34 on," "valve 56 closed," "setpoint for controller 36 = 50 °C," and the like. Accordingly, this level permits an engineering formulation and includes device-specific details. But the formulation depends on the specific instrumentation present in the plant.

Finally, if the group functional units are chosen as resources, the instrumentation-related details become internal to the resources. Group functional units represent the functionality of plant (equipment) units. Now the performance specifications will contain elements such as "heat to 40 °C" (vessel heating), "meter in 100 L at 1 L/s" (inlet line), "mix vessel contents" (stirrer), and "transfer up to a maximum of 300 L" (inlet line). In order to avoid terminological confusion, Figure 120 shows the situation schematically. The focus is on the "performance order," which corresponds to a concrete instruction to a group functional unit to perform a certain procedure. Every group functional unit has the capability of doing certain procedures. On the other side are the possible performance orders, that is, the linguistic elements from which the performance specifications are assembled.

Group functional units are a convenient, meaningful, and workable level. Choosing this level permits both the formulation of performance specifications for automatic actions and the manual control of actions by issuing single performance orders at a technologically understandable, not excessively complicated level.

Actions. Each action serves to accomplish a self-contained partial task and, within the plant complex, is a separate, identifiable object. How an action is to be carried out is set forth in its performance specification. The system responsible for the performance of an action is identified in planning and scheduling. This performing system must know the performance specification or must be given it. For example, the processing of a production action can be transferred to the plant operator or the process control system. If the action is assigned to the process control system, this system independently performs the action in accordance with the performance specification stored in the system. If the action is assigned to the plant operator, the operator handles the action by manually issuing performance orders in accordance with the performance specification contained in the operating manual. Processing includes issuing performance orders, controlling the allocated resource, collecting archival data, generating the proper reports, giving the ready signal, and monitoring correct operation.

Each action is generated, planned, and scheduled by the plant planning entity and is transferred to a performance system where it is actively processed. It is then deleted from the performance system and the planning and scheduling system, remaining only in the archive sys-

tem. Figure 121 shows the states of an action and the possible transitions.

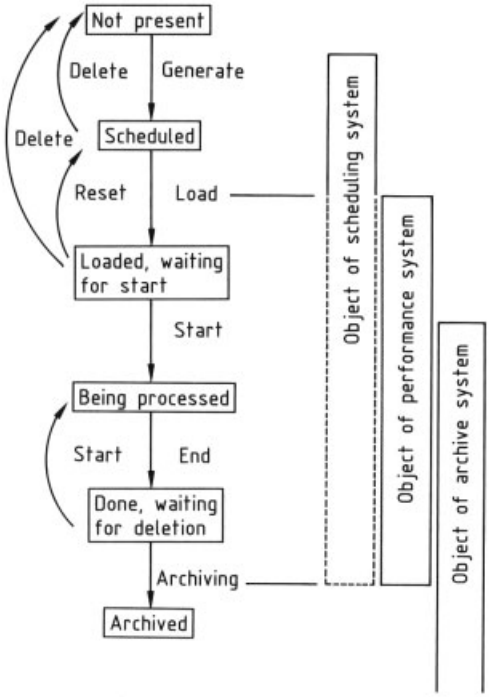

Figure 121. Operating states of an action

Lot planning, tank storage planning, maintenance planning, and other activities generate requirements for the performance of certain actions in certain spans of time. In order to schedule the performance of these actions, all actions planned up to a certain time "horizon" are reported to the in-plant planning and scheduling entity. An action thus becomes an object in the planning and scheduling system, which maintains a list of all actions generated. It makes available an operator interface, with whose aid further actions can be generated and existing actions can be rescheduled, modified, and deleted if necessary. The planning and scheduling system supports planning by means of prospective calculations, then uses the list of actions to be performed together with the current status to arrive at resource allocations, determine the contents of product storage units, and predict the future states of the actions.

For this purpose, there is a simulation specification for each type of action. With the help of this specification, which treats each action as a planning and scheduling unit, it is possible to forecase when which plant resource will be needed, and to say when and where substances are to be received and discharged (and in what quantities).

On the basis of such a forecast, the action plan can easily be checked for critical resource allocations or critical states of product storage units.

During production, this forecast must take account of the actions now running and those loaded. In other words, the planning and scheduling system must always know the present status of all actions being processed.

Scheduled actions are initially pure scheduling objects in the planning and scheduling system. At this point, they have no access to resources. To be performed, actions must first be loaded. When an action is loaded, a performance system is ordered to perform it; this system receives all the required functions. If there is no instruction to the contrary, the performance system then automatically carries out the action.

An important property of the loading operation is that pre-loading is possible, meaning that an action can be loaded into the performance system long before its scheduled performance. In this way, the requirements on the availability of the planning and scheduling system are eased. An adequate reserve of actions can always be kept loaded in the performance systems. For example, all production actions to be performed automatically in the next 24 h can be loaded into the process control system; the maintenance group is given a list of all maintenance actions to be carried out for the week; the plant operator receives a list of all production actions to be performed manually during the weekend.

A performance system of special interest for process control engineering is the process control system. The loading and automatic execution of actions might be thought of in this way: When an action is loaded, a control unit is created in the process control system and initialized.

The control unit is a self-contained object with the ability to verify conditions and to perform its task automatically. It is no longer dependent on the intervention of higher levels but still can, if necessary, be influenced by higher levels. The loading of an action can be initiated either manually by the plant operator or automatically by the planning and scheduling system.

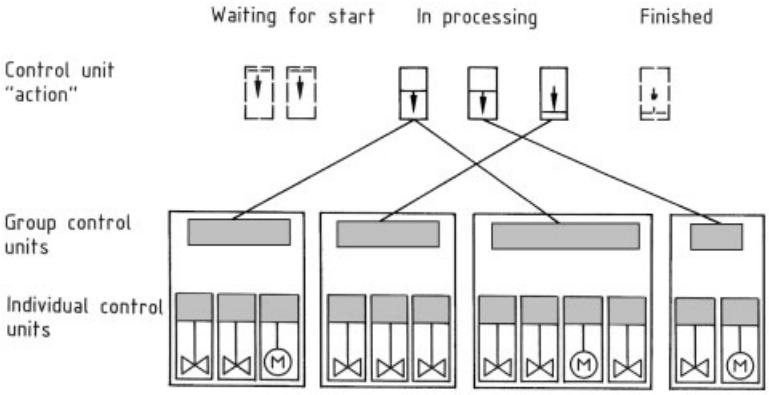

Figure 122. The action as control unit

Once loaded, the action considered as a control unit goes into an initialized condition: "loaded, waiting for start." It has no effect on process control and no access to resources, and it can be deleted or reset at any time without having exerted any influence on the system. The processing of the action begins when the control unit is started; during this operation, the control unit is given access to the needed resources. It gains such access through allocation of group functional units and the construction of the control links required for this purpose. Two concepts should be distinguished here. In the simpler one, all links that the action will need at any time during processing are made at the outset. In a somewhat more complicated concept, only those links called for during the initial phase of processing are present at any time; links that are no longer needed are dynamically broken while the action is still being processed.

Once an action is complete, all links created for it are abolished.

While an action is being processed, the action control unit acts like a normal, permanent control unit (Fig. 122).

Once an action has left its initialized state and has been started, it cannot return to the initialized state. The action has been launched, cannot be called back, and must be processed all the way to a correct or incorrect termination. Every action, once started, is stored in an archival system; it cannot be cancelled without entry in the archive. All stored information about the progress of the action and all stored information about group and individual functional units working under its control can be accessed by virtue of this entry in the archival system.

The form of storage depends on the implementation.

It is evident that the action model described here is especially suitable for batch processes, where a one-to-one mapping of actions to production actions is possible. The steps in an action (see the state model in Fig. 121) largely correspond to the steps in a segment of a batch production process.

This is not the case in continuous processes, which are characterized by the continuous operation of process steps rather than by steps with definite terminations. The stress is on managing and controlling the modes and transitions between modes as shown in Figure 123.

Nonetheless, continuous processes can be described in terms of the action model. One simple option is to "discretize" the continuous process into small increments delimited in time and having definite production objectives.

Automatic Performance of Actions. The automatic performance of actions means implementing the functions required to complete the action by using automatically functioning software or hardware modules. This can be achieved, for example, by implementation in a process monitoring and control system, a maintenance system, or another system for carrying out actions.

The reference model has a structure appropriate for this purpose: Resources and the actions themselves are built up in object-oriented fashion, in two senses: first, that each resource or ac-

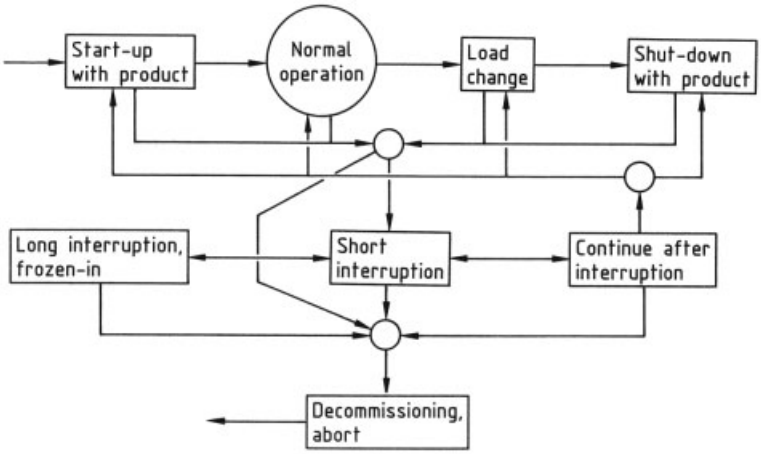

Figure 123. Operating modes of a segment of a continuous process [370]

tion contains its own control unit and thus acts as a self-contained unit in the system; second, that the resource or action has properties that can be derived from a type concept. Control units occupy places in a control hierarchy. Each control unit is connected to the next higher-level control unit and the next lower-level control units through well-defined control interfaces. In addition, each control unit has interfaces through which it can exchange information with all other units.

In process control systems, control units can be implemented as modules. This is a common and a standardized practice in process control engineering; for example, the IEC standardized languages (WG65A) are based on an object-oriented modular approach.

Example of a Modular Solution. The following section shows how the action and resource model can be implemented in a modular form in a process control system. The modular model has the following properties:

– *Support of the type concept*
  The type concept is a special manifestation of a one-level abstraction hierarchy. Each module is of a certain type. The type defines all methods, the description of connectors (input/output interface module), and the description of internal variables. Further, the type includes certain type-specific variable values. The entity, as a management unit, contains not only its name but also its specific data set and the reference to its type. Accordingly, the entity contains no methods or variable descriptions, only parameters.
  In contrast to the individual device approach, in which each compact controller contains the type information (e.g., control algorithm), in process control systems the type in-formation can be contained only once, in one component. Entities then contain only the module state information and module-specific parameters. Under this concept, the type information must already be present in the target system before the creation of an entity. Loading of type information is a separate step from instantiation of the individual module.
– *Data exchange through connectors and links*
  According to this concept [379], data exchange between modules must obey the following principle: Every module offers certain connectors to serve as interfaces to the external world. Two matching connectors of two modules are connected to each other via a link. Once the link is in existence, data can be exchanged. The construction of the link itself is a separate process from data exchange. Links are independent management objects in the system and can be created and again destroyed while production is under way.

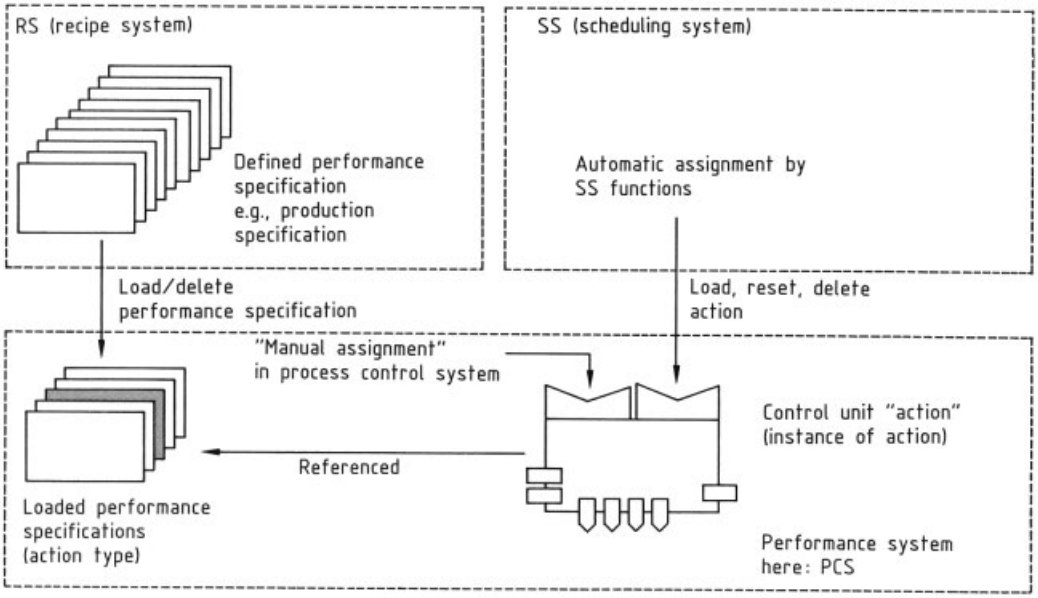

Figure 124. Separate loading of performance specification and action (performance system here = process control system)

– *Closure of modules*
  Each module as a whole is a unit that can be manipulated. Modules are created by simple instantiation commands. The prerequisite is that the proper type must be present. Once modules have been created, they can be incorporated in the processing by an activation command. They bring along all the prerequisites for their function. Modules can be created at any time, and also deleted again, while production is under way. Once modules have been deleted, no fragments remain in the system.

A particularly simple implementation of the action concept is obtained if all functions required for the performance of the action are contained in the performance specifications (and thus in the type description). The individual action contains not specific functions but only parameters for the control of the performance specification. In this way, there is a strict division between the creation and loading of performance specifications, on the one hand, and the creation and processing of individual actions, on the other (Fig. 124). This strict separation is not a practical drawback, since in general the creation of performance specifications takes place separately from the scheduling and operational performance of the actions. The various tasks are often accomplished on different levels: Performance specifications are created and managed in a "recipe" system, while action management and initiation take place in a scheduling system.

An important performance feature of a process control system is how convenient it is to create and load a new performance specification, modify an existing performance specification, or generate the associated simulation specifications, original data sets, and parameter masks for actions being created.

Another important point, however, is that it is possible to attribute all required manipulations in the performance system to operations formulated in the modular concept. Thus no loading, writing, or coupling steps on the modular system are necessary. In this way, the action concept can be logically implemented in a modular system.

Automation. A very important property of the model is that automatically performed actions and actions manually performed by the plant operator can be handled concurrently in a process control system. It is thus possible to build up to a higher level of automation in a step-by-step fashion.

When actions can be automatically completed, a system acquires the degree of automation that Polke has called "logistical mode" (see Section 11.1) [376].

Summary and Outlook. The action and resource model attempts to describe a general structure of operational plant control. Important features of the model are the resource structure, the standardization of the order interface, and the introduction of "actions" as self-contained control and management objects. This contribution has taken the example of plant engineering resources in order to show how such a structure can be created and implemented in a suitable way.

The concept suggested is, however, a technology-free general mechanism. It is strictly limited to describing the operational performance of actions; thus it contains no specifications on how the planning and scheduling task is to be accomplished or on how recipes are to be generated, and it provides no rules as to how resources are to be selected and organized.

Special models are required for these purposes. By way of example, a resource model for plant engineering resources is described and its consequences explored.

A plant complex must carry out its tasks while fulfilling a variety of requirements and satisfying a number of constraints. "Plant-wide" decisions are needed on a daily basis. The action and resource model enables the plant to formulate its operational decisions in a consistent way and implement them in a systematic manner. The results are not just a simplification of scheduling, but also an improvement in the quality of planning and performance of in-plant actions. The documentation of plant activities is part of the concept and can be synthesized in a systematic way.

The action and resource model is logically object-oriented. It can be extended directly to a corresponding modular system.

With regard to usability, this model has the great advantage that technical devices and personnel can both be used in constructing resources and performance systems. Plants can thus be controlled in a consistent way, regardless of their instantaneous degree of automation.

A special task for the future is to integrate existing process control concepts for production control, plant management, recipe-based process control, and production logistics into the general methodology of the action and resource model, thus arriving at a comprehensive solution for the operational control of process plants.

5. The Process Control System and Its Elements: Process Sensor Systems

5.1. Principles

Sensor, Sensor Technology, and Sensor System: Classification and Standards. Modern process control engineering, with a distributed structure of information processing based on microelectronic devices, has greatly expanded the potential of the information economy for the control, monitoring, optimization, and reliable conduct of processes [380].

If this increasing potential of information processing is to be utilized fully, the amount of information acquired directly or indirectly as to the status of the process must also be augmented.

A relative bottleneck exists here, because the development of sensor technology lags behind that of information processing [380].

There are a number of reasons for this. Technical functional units in the immediate vicinity of the process are always much more expensive and complicated than those in a protected environment such as the control room. Furthermore, the continued use of conventional, internationally standardized signals together with proven transducer technology stands is hindering further development. Finally, information processing is often regarded as being much more important than information acquisition.

Decentralized process automation is beginning to extend distributed "intelligence" based on the use of microelectronics to the point where measurements are acquired. Therefore, the meanings of "sensor element," "sensor," "sensor technology," "sensor system," and "direct signal processing" have become blurred.

For example, consider the "critical" link in the action chain associated with measurement. In the current language of process and production engineering, the functional element that embodies the interaction between the material system and the instrumentation is referred to as a "sensor" [381].

Here, "sensor" no longer means the same thing as it commonly does in the electronics industry, where it is restricted to the proportional transformation of nonelectrical input quantities to electrical output signals [382, 383]. The appli-

cable standard (DIN 19 226) is therefore in need of revision. A useful starting point for a meaningful definition of the term "sensor" is offered by Sheet 3, "Apparatus Terms," of VDI/VDE Guideline 2600, "Metrology" [384]. This document employs "pickup" or "primary element" for the (entire) device that accepts the quantity for measurement at its input and emits a corresponding measurement signal at its output. The "sensor" proper, on the other hand, is the part of the pickup that directly acquires the quantity for measurement and reacts in some sensitive fashion to the quantity. Even this definition, however, is appropriate only for a relatively small fraction of the sensors currently utilized for measuring material quantities. The definition does apply, for example, to semiconductors and piezoelectric sensors; it is difficult to apply to IR photometers and paramagnetic oxygen sensors, because here the interaction between the material system and the instrument does not take place in the sensor, as VDI/VDE 2600 presumes, but via an energy field serving as an auxiliary quantity [385].

The present terminology and standards, such as the definition in VDI/VDE 2600, are thus insufficient and require further discussion. An unambiguous resolution of terminology in this area is essential in the long term.

Suggestions for refinement of the terminology have been made in [381, 383, 385, 386].

Here, it is proposed to separate information acquisition in and on the process from information acquisition in the laboratory and, as applicable, in the pilot plant. Both classical "process measurement" [387, 388] and "process analytical techniques" [389] must fulfill fundamentally different requirements from the comparable instruments in the laboratory. The shift from information acquisition in the plant laboratory and "process analysis rooms" to acquisition in the production process itself justifies the information-oriented term "process sensor system technology" as opposed to "laboratory measurement technology."

Process Models and Sensor Systems. The concept of "process model" has been known theoretically since the 1960s; it encompasses "black box," mathematical, physical, and other types of models. Especially important in connection with measurement technology are adaptive process models, also known as observer models.

Hundreds of books and tens of thousands of other publications have appeared on "process models" (see also Chapter 3), "process identification," and "process simulation." There are academic departments concerned with these topics, and many congresses and symposia on them are held every year. Although the theory is advanced, there is a wide gap between theory and practice [380, 390].

According to Perne and Polke [391], the process model is ultimately a mapping between product properties and process properties.

Problems of product variation or product development could thus be attributed to a process variation if a suitable process model were available. Identifying such a variation would provide a starting point for "optimization" under a wide range of criteria (see Section 11.3).

This problem can be handled systematically and – most importantly – optimized only if a representative mapping or transformation (i.e., the process model) can be found between the process space and the product space (see Fig. 408).

But why use a model instead of the real production process? Simply because the real process is generally much too complex and opaque to permit direct operations based on physical and chemical relationships and algorithmic descriptions of these (see Chapter 3).

However, this chapter is less concerned with the representability of a model in relation to the real process and the desired product qualitites, and more interested in the reaction on sensor elements and sensor system technology.

It is important, above all, that questions of sensor technology not be separated from the topics of process models, process properties, and product properties.

Material Flow, Energy Flow, Information Flow, and Sensor System. As Figure 125 shows, there is a prallelism and correlation between material – energy flow and information flow. The link between an event or state in the process and the information system is effected by the sensor system.

Accordingly, the sensor system has to carry out two basic tasks. First, the input quantity resulting from an event or state must be sensed and transformed into a signal quantity suitable

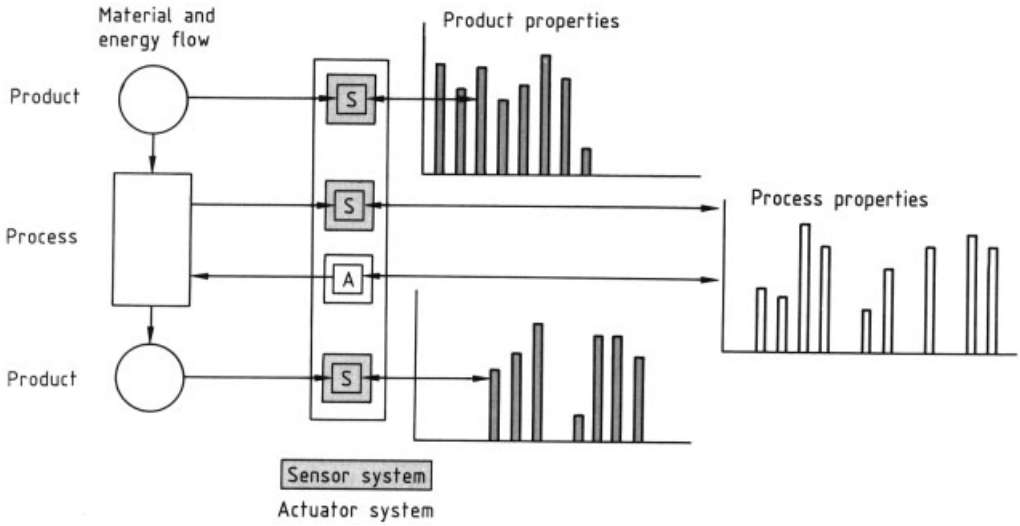

Figure 125. Position of sensor systems relative to material, energy, and information streams

for processing. Second, the signal quantity must be set in relation to its cause so that interpretation can extract usable information (see also Section 5.2).

The sensor itself must perform the first task, while the second is in the realm of information processing. Some thought should also be given to the possibility of augmenting the information by serial, parallel, or associative linking of single items, whereby the result may be of far greater relevance than the sum of the individual pieces of information. Typical examples are the temperature profile, the stress tensor, the density field, and the concentration gradient.

This multi-sensor approach is widely used in the field of biological sensors. A typical example for the capabilities of an array of identical sensors is image recognition by the eye. A further advantage of arrays becomes apparent in biological sensors but has been little utilized in industrial sensors: While the individucal biological sensor is markedly inferior in performance to many industrial sensors, a system with intelligent preprocessing of signals achieves an amazing degree of flexibility in dealing with unforeseen events. Naturally, this flexibility is more vital to biological systems than to industrial ones, where the service loads can be established more precisely. Nonetheless, the gain in flexibility through multi-sensor technology is of interest for industrial systems as well [392].

What holds for any economic system also applies to the information economy [393]: One can use only what one has. The first step, accordingly, must be information acquisition. It is reasonable to limit the information acquired to that which is important and actually necessary. There are two reasons for such a restriction: to avoid an information flood and to hold the cost of instrumentation within acceptable bounds. Both experience and mathematical and physical methods can be used as guides for limiting the information flow.

Because it is often difficult to identify the relevant properties, that is, the properties that adequately describe the process or the product, more easily measurable properties (indicators) are often resorted to. Process optimization requires better penetration into the process, and this in turn means finding ways to make relevant properties – i.e., those that can serve as significant process indicators – more measurable.

Hesse, in his editorial in the "Sensors 1992" special issue [394], used the following terms to describe the shift in emphasis in the development of sensor systems:

"At the beginning of the 1980s, the problems were to systemize technical terminology, to discover effects and technologies that could be exploited in practice, and to find out how medium-sized companies could establish themselves in this field. Later, in the mid-1980s, publications on practical sensors utilizing a wide

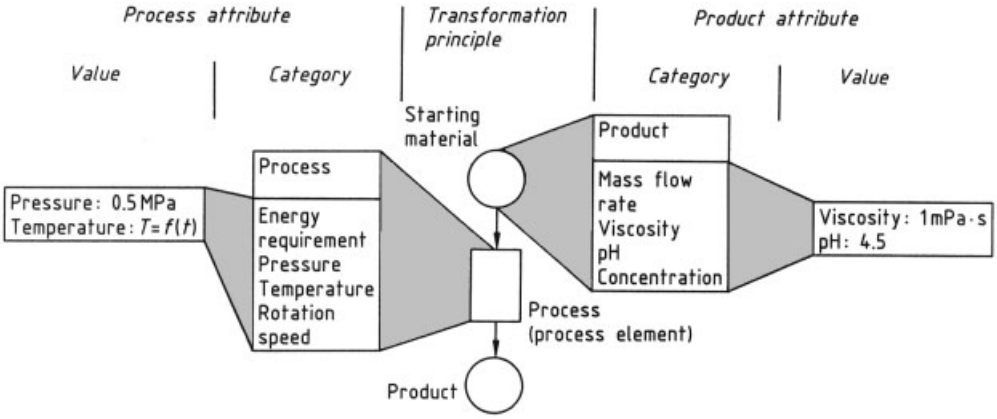

Figure 126. Assignment of attributes in the phase model of production [395]

range of technical principles became dominant. These were still in the foreground in the early 1990s, but systems aspects are now being examined along with applications and on an equal footing.

In fact, systems concepts should have been considered right from the start. These aspects aid in the development of strategic insights and objects, since they describe the overall high-level solution. The introduction of process control engineering in the chemical industry can serve as a model: Sensors are dictated by the system itself and can thus be developed under explicit constraints."

For these reasons, some concepts having to do with the use of sensor (system) technology in process control engineering are discussed first.

From Phase Model of Production to Process Information Model (see Chapter 2). A key aid in describing and analyzing processes is the phase model of production. This object-oriented semantic tool originated in software engineering. The outcome, a directed bipartite graph, belongs first to the category of product nets (see Chapter 2).

The phase model of production consists of two object classes, "product" and "process element." Individual products and process elements are described in terms of attributes. The attributes are the object classes of information about a process; they are defined by category and value (Fig. 126 [395]).

The value is made up of a numerical value and a unit. Whenever metric scaling is possible (see Chapter 3), the units employed should comply with U.I.P. 20 (1978) [396] and ISO 31/0 ("General Principles Concerning Qualities, Units and Symbols") and DIN 1313 [397].

Higher-level requirements such as "product quality," "process safety," and so forth are imposed on the production process.

Objective: Product Quality. Product quality refers to a subset of the actual properties of a product, which in terms of category and value form an intersection with the nominal properties demanded by the customer (Fig. 127). The development objective is to maximize this intersection within reasonable tolerances.

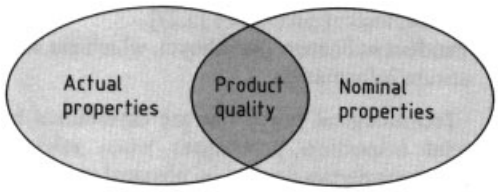

Figure 127. Product quality [398]

Ojective: Process Safety and Environmental Safety. Safety tasks such as process safety and environmental safety can be accomplished with the resources of process engineering and process control engineering. This can be achieved by means of technical or organizational actions [399] or a combination thereof (see DIN V 19 250 [400]). This procedure makes it possible to find alternative solutions that are equivalent from the safety standpoint, so that the most economical solution can be applied in imple-

mentation [401]. In the new NAMUR Recommendation, the required process control safeguards are matched to the risk arising from a process facility or the absence of such process control safeguards. This risk assessment is necessarily qualitative and must be carried out case by case. Other procedures that lead to a comparable level of safety should, however, not be ruled out. The protective objective in the context of this recommendation is to prevent injury to persons as well as major environmental harm and damage to property. In the case of major damage to property, damage assessed as serious danger (e.g., according to [402]) should be distinguished from damage evaluated as affecting the company's interests.

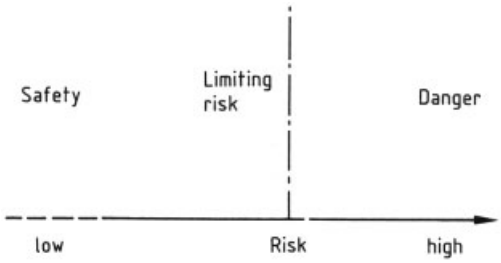

Figure 128. Safety, risk, and danger [403]

Some important concepts should be defined here (see Fig. 128), without an in-depth examination of the extensive safety literature [401, 403–405] or a workshop on Safety in Chemical Production:

- Danger: A danger is a situation in which the risk exceeds the limiting risk.
- Limiting risk: The limiting risk is the highest still-acceptable risk of a certain technical process or state. In general, the limiting risk cannot be quantified; it is described indirectly or in terms of safety considerations.
- Risk: The risk associated with a particular technical process or state is described in terms of probabilities. Two factors are included: the expected frequency of occurrence of an event leading to damage and the extent of damage expected when the event does occur.
- Safeguards are actions taken to reduce risk by lowering the frequency of occurrence, the extent of damage, or both.
- Safety is a situation in which the risk is lower than the limiting risk.
- Availability: The availability of a technical device is the probability that the proper functionality is present.

It follows that, with regard to the safety-related availability of process control devices, solutions for holding the risk less than the limiting risk should be evaluated from the economic standpoint only if they are equivalent from the safety standpoint.

In contrast, it is useful and necessary to analyze other availability characteristics to determine the economic consequences of malfunctions. This statement applies to the logistical availability of equipment that would impair the ability to deliver (just-in-time, just-in-case) and also to the process availability of production equipment that would impact directly on product quality or plant condition (e.g., production stoppages, market losses, repair costs).

The unambiguous description of proper functioning, that is, the complete set of all nominal attributes needed for such a description, leads to the information model of quality (Fig. 36) or the information model of safety (Fig. 37). The histograms illustrated show, for each product object or process object, all attributes necessary for its description, along with the nominal values and their tolerances ("good," "defective, acceptable," and "defective, unacceptable" ranges), which may be scaled either metrically or topologically (see Chapter 3).

This information must be statically and dynamically defined at the maximum level of detail [395]. What is more, the spatial and temporal representativeness of the attribute values must be taken into account, especially for ex-line sensor systems (Figs. 144 and 145).

Product and Process Properties. The attributes used to describe the objects of the production process are process properties and product properties.

The term process properties includes the following:

- State variables, such as pressure, temperature, and concentration, with which the process can be described either directly or via derived state functions.

– Process parameters, such as heat-transfer coefficient and catalyst activity, which characterize the constraints under which the process takes place. Process parameters are steady or at least quasi-steady quantities.
– Controlled variables, such as motor speed and setpoints, which characterize actions on the process.

In theory, the process can be described completely if a suitable set of these quantities is available. In practice, however, it often happens that not all the desired state variables and controlled variables can be measured or specified. Substitute information often must be resorted to:

Setpoints, such as stirrer speed, are correlated with the controlled variables or represent residual categories.

Process indicators [406] are empirically correlated with one or more state variables (and may also be product properties; see Section 11.3). These represent substitute information.

Controlled variables and setpoints characterize the part of the process state that is imposed from outside. It is the task of the actuator system to convert the specification, provided by the information flow, to an action on the process. The remaining process properties characterize information that is to be acquired from the process with the aid of sensor technology.

Product properties are the following:

– Physical quantities
– Chemical quantities
– Technological properties [406]
– Product indicators (see above), which are substitute information

Technological properties are determined by special inspection procedures when relevant product properties depend on physical quantities that can be determined only with great difficulty, or if the relationship with the physical quantities is currently unknown. Product indicators are empirically correlated with physical and chemical quantities, but process properties can also be employed as product indicators. All product properties characterize information that can be acquired by using sensors.

"Technological" and "Physical" Test Procedures. Some fundamental problems in establishing material properties are illustrated for the case of rubbers [406].

Technological Test Methods. The testing and characerization of chemical substances is generally limited to assessment based on technological tests or application-oriented model tests, which are single-point measurements which have limited predictive power and do not take account of the functional relationships actually present. Technological tests do not directly lead to conclusions about the physical and chemical properties of chemical substances. This means that such data alone are insufficient as a basis for the optimization of chemical production processes.

The essence of technological testing can be illustrated by a simple diagram (Fig. 129). A certain material M_j under a certain load X_i exhibits a technological behavior Z_{ij}.

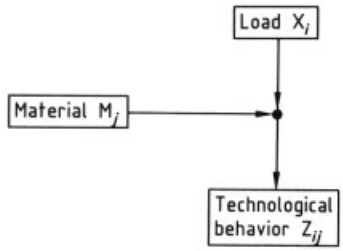

Figure 129. Technological inspection [406]

When the test load (generally complex) is specified in terms of the arbitrary design of a testing machine, the actual loading often cannot be expressed physically. For example, the Mooney tester [407–409] has an inconsistent shear rate and inhomogeneous stress and temperature distributions in the steady state. Furthermore, the variation of temperature with time during the measurement varies in an uncontrolled way as a function of the specimen used. The single-point measurement performed, namely the torque after 4 min, describes only an integral response of the material relative to the test-chamber geometry as the specimen is subjected to a load that is neither homogeneous nor steady.

The predictive power of this measurement is thus limited to comparison of materials under these conditions. In general, no certain conclusion can be drawn as to the processing qualities of the rubber tested in this way, since the shear loads that occur in industrial processing machinery are several orders of magnitude higher.

Physical Test Methods. A material can be better characterized in terms of its physical properties. These quantities are obtained by applying simple loads and measuring the relationships between load and response. Such simple loads include linear, uniaxial stress, laminar flow at constant shear rate, and a steady homogeneous temperature distribution in the specimen. Complicated loading modes such as occur in practice can generally be synthesized by superimposing simple loading states.

The material properties M_{ij} are material-dependent quantities described a functional relation:

$$Z_{\mathrm{ij}} = f\left(M_{\mathrm{ij}}, X_{\mathrm{ij}}\right) \tag{5.1.1}$$

For example, Hooke's law

$$\tau = G\gamma \tag{5.1.2}$$

describes the behavior of many elastic solids. Here, the stress τ is the load ($X_j = \tau$), while the deformation or strain γ represents the material behavior or response ($Z_{ij} = \gamma$). A material constant, the shear modulus $G = \tau/\gamma$, interrelates these two quantities and characterizes the mechanical properties of the material ($M_{ij} = G$).

Figure 129 shows that the technological behavior is governed by the material and the loading. The task of a phenomenological theory is to define the formalism that describes the observed relations (Eq. 5.1.1) and the properties M_{ij} of the material (Fig. 130).

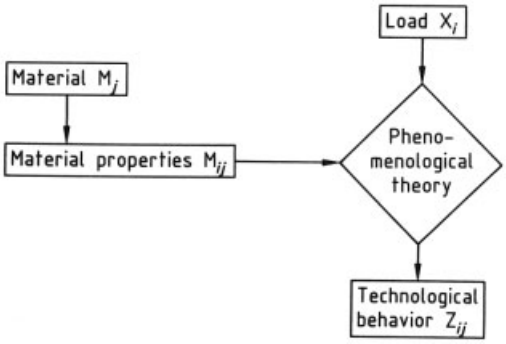

Figure 130. Physical inspection [406]

Idealized models of substances and bodies play a major role here, for example the ideally elastic (Hookean) solid, the Newtonian fluid, and the Maxwell and Voigt bodies (Fig. 131). Deviations from the behavior of the idealized body are handled by restricting the range of validity (Hooke range) or by introducing correcting relations. For example, the viscosity of a general non-Newtonian fluid is described by

$$\tau = \eta \mathrm{d}\gamma/\mathrm{d}t \tag{5.1.3}$$

where τ is the shear stress, $\mathrm{d}\gamma/\mathrm{d}t$ is the shear rate, and the viscosity is treated as a function of the shear rate $\mathrm{d}\gamma/\mathrm{d}t$,

$$\eta = \eta\left(\mathrm{d}\gamma/\mathrm{d}t\right) \tag{5.1.4}$$

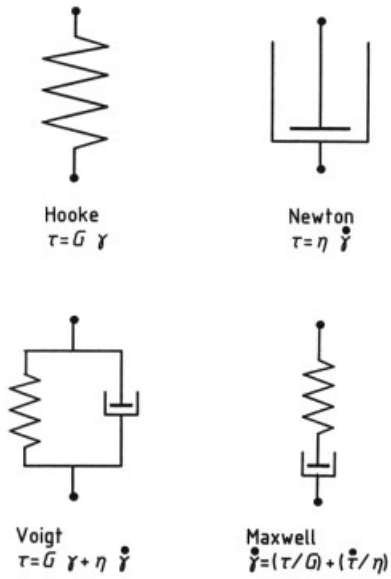

Figure 131. Rheological model [406]

The phenomenological theory describes the physical properties of a material by interrelating the load state and the behavior of the material, whereas polymer physics seeks to devise a theory that correlates the physical properties of a material M_{ij} with the chemical composition and structure of the material or a structure model (e.g., Kuhn coil model [410], Pechold link theory [411]); see Figure 132. If this were achieved, there would be a basis for the planned development of particular polymers.

Among the most important properties of elastomers is the viscoelastic behavior. There is still no self-contained phenomenological theory to describe the entire mechanical deformation behavior, including fracture. For small deformations, relaxation spectroscopy [411] can

be used for complete determination of the viscoelastic behavior of polymeric materials. Under harmonic loads, the complex shear modulus $G*(\omega,T)$ is measured as a function of frequency ω and temperature T. The results can be used to calculate the relaxation time spectrum $H(\tau_R,T)$ (τ_R = relaxation time), which is closely related to the structure. Using the Boltzmann superposition principle allows the material behavior under complicated, nonharmonic loads to be calculated.

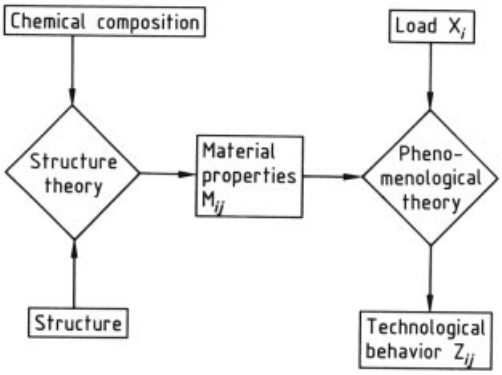

Figure 132. Physical inspection and structural theory [406]

These methods fail at large deformations, where the relation between stress and strain becomes nonlinear. The Boltzmann principle, which states that the effect of a composite load can be treated as the sum of the effects of its composite simple loads, does not apply in such cases. Fundamental findings and results on the material properties of multiphase chemical substances, including these application-relevant ranges, can be found in [412–414].

The question arises to what extent the mechanical behavior of elastomers can be completely characterized despite the lack of a unified nonlinear theory of viscoelasticity. NINOMIYA and FURUTA [415] proposed that in general four functions are sufficient to characterize the deformation behavior:

- The relaxation time spectrum $H(\tau_R)$, which characterizes linear viscoelastic behavior at small strains
- The stress – strain curve $\tau(\gamma)$, which characterizes the nonlinearity
- The fracture strain $\varepsilon_B(T)$, which characterizes fracture behavior
- The shift factor $a_T(T)$, which characterizes the temperature dependence [416]

Such a description of a polymer, however, requires substantial efforts. To determine the relaxation time spectrum and shift factor, the complex modulus must be measured over a wide range of temperatures and frequencies. Moreover, extensive calculations must be carried out.

The difference between technological and physical test methods can be summed up as follows: Technological tests investigate the behavior of the material in an application-oriented model test, generally involving a single-point measurement under complex loading conditions. In contrast, physical tests are concerned with the fundamental behavior of the material. Simple and easily quantifiable loading modes are selected and then varied such that the fields of properties are determined in multidimensional product spaces (see Fig. 404).

Despite the obvious advantages of the physical analysis, technological tests will continue to be important because of their lower cost.

A crucial point in selecting a method is the purpose of testing. Three aims must be considered:

1) Product characterization in general, in the sense of quality assurance (e.g., DIN ISO 9001 [417])
2) Acceptance testing and production control
3) Prediction of the practical behavior of a material

The objective of product characterization is to achieve at the most complete possible description of product properties in terms of instrument-neutral physical quantities, to determine the key parameters that affect the product properties, and to quantify functional relationships.

Accordingly, product characterization encompasses the area of physical testing. The demands on instrumentation and the amount of work required for interpretation are so great that these methods are generally only justified for new or as yet unknown products. This is the only way to carry out a reliable, exact comparison between materials.

As a technique for routine inspection, for example in acceptance testing and production control, sampling inspection of material properties is generally adequate. Sampling techniques (re-

presentativeness over space and time) are applied to both the quantity of material and the selection of loading mode combinations. The principles of statistical quality control should be observed [418].

The creation of correct, optimal experimental designs, however, presupposes a comprehensive knowledge of material properties.

The requirement on acceptance testing procedures is that they can be performed by the vendor at one location and the buyer at another, using different instruments. The parties must agree on the test conditions (establishing the controlling parameters); this is one of the key tasks of international standardization (e.g., ISO). In contrast to product characterization, single-point measurements may be relevant here as well.

In production control, various batches of a product are subjected to spot checks to verify that they have identical physical properties. The important prerequisite is good reproducibility of measurements on a given instrument. It may be desirable to choose complex test conditions in the interest of quick, effective monitoring [419].

There are two possible approaches to forecasting the mechanical and technological behavior of a material in practice. First, tests can be performed under the process conditions and loading modes encountered in practice; this step should be preceded by a thorough analysis of the process (see Section 11.3 with Figs. 395 and 396). Second, in principle it should be possible, given suitable results from the determination of the physical material properties, to calculate in advance how the material will behave under the required loading conditions. The second method, however, often runs into the difficulty already mentioned: the lack of applicable theories. For this reason, the first approach (simulation of practical conditions) is becoming the more important.

By analogy with this example of mechanical product properties, the electromagnetic product properties defined by Maxwell's equations

$$\boldsymbol{D} = e\boldsymbol{E} \tag{5.1.5}$$

$$\boldsymbol{L} = \mu\boldsymbol{H} \tag{5.1.6}$$

$$\boldsymbol{I} = \sigma\boldsymbol{E} \tag{5.1.7}$$

can be examined. The first product properties determined here (generic class "application features") are due to molecular motions and are measured as collective phenomena. For example, suitable spectroscopic or chromatographic methods make it possible to resolve the required product reactions over a variable frequency and temperature field:

$$e = e\,(T, \omega, ...) \tag{5.1.8}$$

Again, the limit of the linear relations is found by considering the variation of the field quantity:

$$e = e\,(T, \omega, E) \tag{5.1.9}$$

The product properties are in general complex:

$$e^* = e_\mathrm{r} + ie_i \tag{5.1.10}$$

Model-Aided Measurement Procedures for Determining Product and Process Properties. From the connection between process model and information demand, it is possible to identify ways and means of overcoming the limitations of sensor technology by utilizing the present capabilities of information processing (see also [380]). The broad potential of model-aided methods, which has been under-exploited in process engineering, should be mentioned first (see Fig. 133).

In the context of intelligent sensor systems, this procedure makes it possible to reconstruct unmeasurable process and product properties from measurable quantities by interpreting a priori knowledge about the process in the form of a mathematical model. Ideally, the entire process state can be derived from the available measurements in this way. By integrating a priori knowledge into the measurement process, the predictive power of measured quantities is markedly enhanced in both scope and reliability.

The prerequisite, however, is a sufficiently accurate mathematical model of the process under consideration [420–424]. Especially in chemical process engineering, however, mathematical modeling presents some difficulties. Many processes, because of their complex nature, do not lend themselves to accurate mathematical description. Process facilities often include subprocesses that are not subject to the classical mathematical model-building procedure.

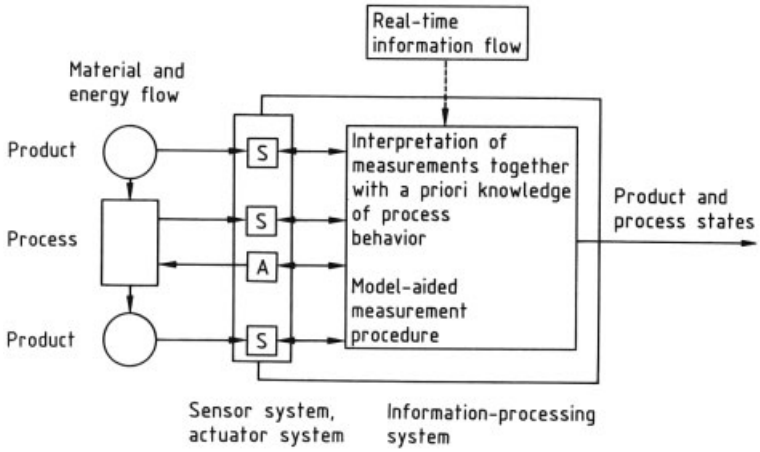

Figure 133. Model-aided measurement procedure

In such cases, for example where the product properties and their relationships can be described only in qualitative terms, new ways of incorporating a priori knowledge about the process into the measurement must be sought. The methods of applied information processing can be expected to reveal new solution approaches. Under some circumstances, verbally formulated algorithms and heuristic methods can lead on to a qualitative description of the process. The use of knowledge-based methods is more and more attractive, for example the inference methods employed in expert systems [425–427] and fuzzy-logic methods [428–430].

Interesting tasks arise here for future research. An initial approach was the project "Technical Expert Systems for Data Interpretation, Diagnostics, and Process Control (Tex-I)," in which several participants (Siemens, Interatom, Bayer, and others) attempted to adapt expert systems techniques to process control. Other projects in this context are SENROB and NERES [431].

It is to be hoped, however, that the practical benefits of model theories will be studied and utilized more intensively than in the past. However, casual use of modern or simply fashionable modeling methods to make up for shortage of process knowledge must be avoided [380].

Figure 134 outlines a model-aided measurement technique for the case where a mathematical process model is available. The mathematical model is set up in parallel with the process and run (simulated) in real time (for example on a mi-

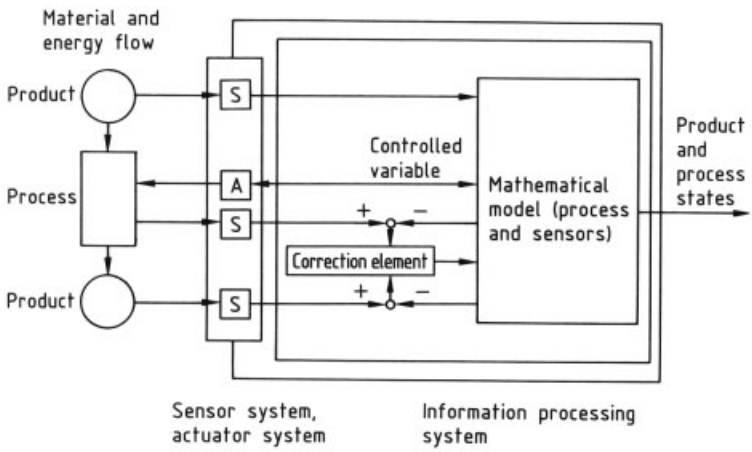

Figure 134. Schematic diagram of model-aided measurement procedure

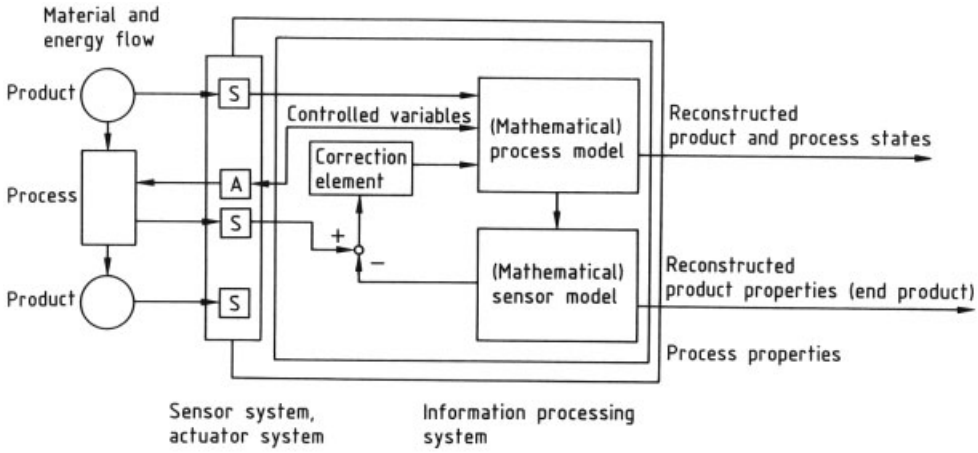

Figure 135. Structure of a model-aided measurement procedure

crocomputer). The known controlled variables acting on the process are supplied to the model as inputs. The class of model-aided techniques includes the Kalman filter and the Luenberger observer. These are similar in their fundamental structure, differing in that the Kalman filter is designed on a stochastic basis while Luenberger observers are designed by deterministic methods. The structure cited here applies to both continuous and batch processes.

In general, the properties of starting materials are interpreted as inputs, since they fix the starting conditions. They are therefore fed to the model as inputs, together with the controlled variables acting on the process actuators.

It is tempting to use the model to reconstruct the unmeasurable part of the process state in view of all knowledge about the inputs. A variety of problems arise, however. Even if an exact model exists, the simulated process state may coincide with reality only when the initial conditions selected for the model are identical with those of the process. But the initial conditions of the unmeasurable state components are generally also unknown. Further, it must be assumed that unavoidable modeling errors and perturbations acting on the process will lead to increasing divergence between the actual and simulated state variables. Forcing convergence between the model and the process, regardless of the selected initial conditions, perturbations, and error effects, means taking corrective action in the model. This can be done by comparing the actual measured process and end-product properties with the same properties calculated from the simulated process state with sensor models. If the resultant difference signals are then amplified in a suitably designed correction element, the correction can be made in the process model itself. The chief problem is to design the correction element in such a way that the behavior of the model will converge to that of the real process as rapidly as possible.

Figure 135 shows the structure of a model-aided measurement method in a simpler form. Here, the product properties and the controlled variables of the process have been combined into a single multidimensional input vector. Similarly, the output vector includes all the measured properties of the process and of the end products. The block diagram clearly shows that the process model, via the measured output variables, is made to track the real process in the manner of a servo follower. The correction element, acting as a controller (closed-loop controller), compensates for differences in behavior between the real process and the model resulting from perturbations, modeling errors, and incorrect initial conditions. In the case of the Kalman filter, it is generally assumed that both process noise and measurement noise act on the plant [420, 432]. Under the assumption of certain properties for this stochastic perturbing signal, the correction element is designed such as to minimize the expected value of a quadratic function of the difference between the actual and estimated process states. In the design of a Kalman filter, the a priori knowledge thus includes not only the pro-

cess and sensor models but also models of the two stochastic processes. In the case of a Luenberger observer, in contrast, the design process is purely deterministic [421,423].

For this concept of estimating or reconstructing the process state to work in a satisfactory manner, it must be insured that the process can be observed by means of the measurable quantities. In other words, all components of the process state manifest themselves in various ways in the time variation of the measured process and product properties. Criteria are available for investigating the observability of a process [420, 421]. Examples of applications in process engineering can be found in, for example, [433].

The discussion to this point has related to the determination of process and product properties in general; measurement and test methods employed in the laboratory, pilot plant, and industrial plant must meet some additional requirements. This chapter is exclusively concerned with measurement in the plant itself, or in an environment close to the process or similar to it. In the past, this subject was referred to as process metrology, or, when analytical questions were being emphasized, process analysis.

The dividing line between these two fields is fluid, and often they are inseparable. It is therefore proposed that only the distinction between process-level (local) and remote sensor system technology be preserved, so that process sensor system technology becomes an appendage of the well-known field of laboratory metrology [434] in which special requirements have to be met.

The computer processing of information from the sensor system changes the profile of requirements. In manual operation, measured values are commonly interpreted and processed by an expert plant operator. In this form of information processing the fault tolerance and error rate depend on the plant operator's level of experience and knowledge.

Computer processing imposes the following new requirements on the sensor system:

- Reliability at an acceptable maintenance cost
- Functionality even in extraordinary process states (startup, shutdown, disturbances)
- Accuracy [435] and dependability

These requirements result from advances in microelectronics that make computerized information processing a possibility. Microelectronics is also an important tool for the realization of these requirements. After optimization of the instrumentation and installation, the design of intelligent sensors with the ability to report faults and initiate maintenance procedures automatically is in many cases the most economical way to satisfy the requirement of reliability.

The requirement of functionality even in extraordinary process states implies not only robustness but also a thorough understanding of what happens on in the process. Especially in sensor systems used to measure quantities that depend on the composition of the measured stream, many independent variables must be handled, and up to now it was only rarely the case that these were known for all possible process states.

Measurement (from the physicist's standpoint) becomes metrology (from the eingineer's). It is not enough that a method works in principle; it is essential that it always functions effectively [436].

The use of computers for information processing, however, makes it possible to acquire information more economically than before. In signal-oriented process control, the motto was: "Acquire information item by item, in parallel, and always," whereas information-oriented process control engineering follows the principle: "Acquire information in the smallest possible amount, but as much as necessary, and only then when it is to be used" [437].

The advantage, especially in batch processes, is obvious. Large savings are possible in the wear-out reserve (see Section 11.4 and Figs. 420, 421) and thus in maintenance costs if, for example, a pH probe in a reaction vessel is brought into contact with the product only when the pH must be measured and otherwise is held in a protected standby condition (see Section 5.3 and [438]).

The new requirement profile and the new principle of information acquisition have influenced not only instrumentation but also the interface between the sensor and information-processing systems (Fig. 125). Information coupling, for example to a process control system, is no longer limited to the transmission of a measured value. Sensor status messages can also be included (e.g., reports that a unit has failed or malfunctioned or that a maintenance procedure is under way), and it should also be possible for

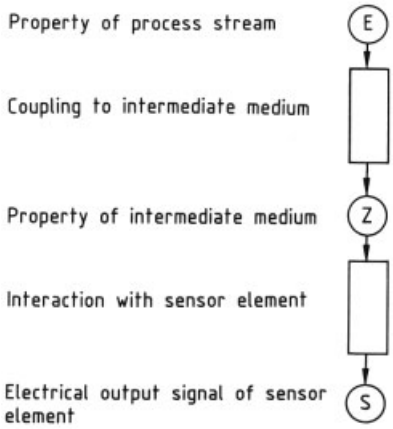

Figure 136. Information flow in sensor systems: Temperature measurement with a jacketed thermocouple

the sensor system to receive information from the process control system (e.g., an instruction to hold itself available; see Figs. 136, 137, 138, 139, 140, 141, 144, 145, 146, 147). The same holds for actuator systems (see Section 6.1).

The definition of the extended interface between the sensor and information-processing systems should also include the signal levels and types and the signal contents.

Besides the change in emphasis as a result of computerized information processing, new measurement functions are being assigned to sensor technology (as they have been throughout the development of this technology). When existing processes, such as polymerizations, are automated and optimized, the need for information is reconsidered and new constraints applied, with

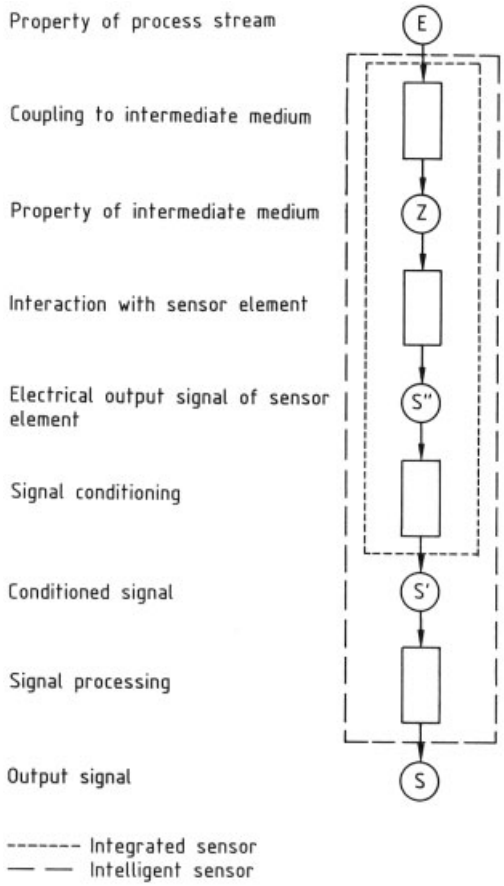

Figure 138. Information flow in sensor systems: temperature measurement with jacketed thermocouple and signal processing

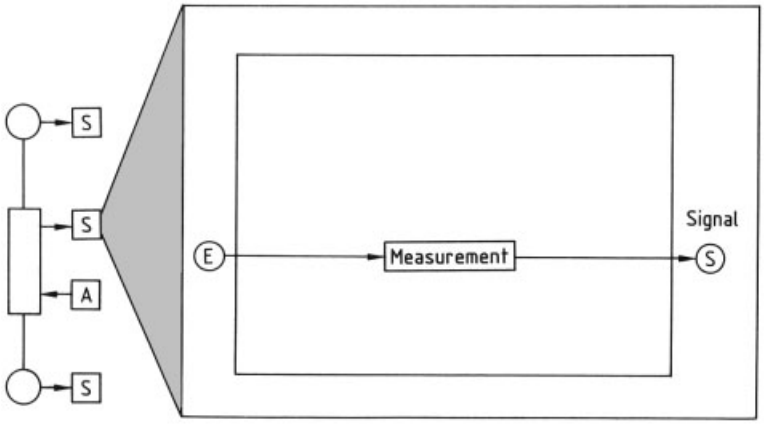

Figure 137. Information flow in a simple sensor system [389]

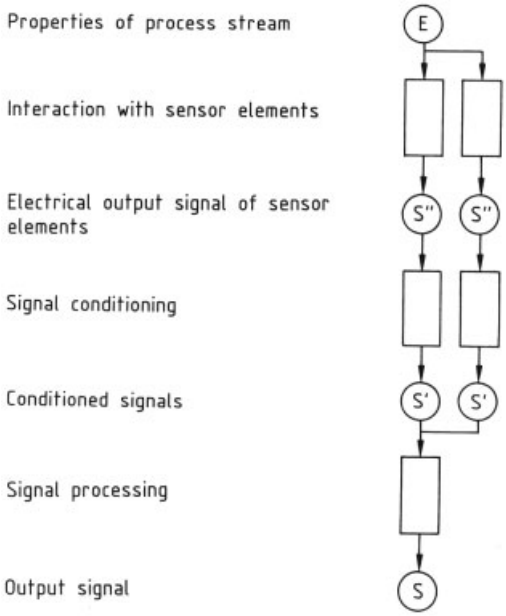

Figure 139. Information flow in sensor systems: sensor system with compensation (e.g., pH measurement with temperature compensation)

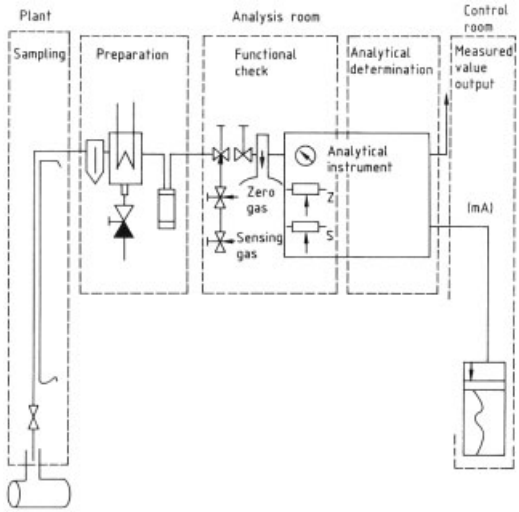

Figure 141. Functional groups of a process analyzer (gas analyzer) [389]

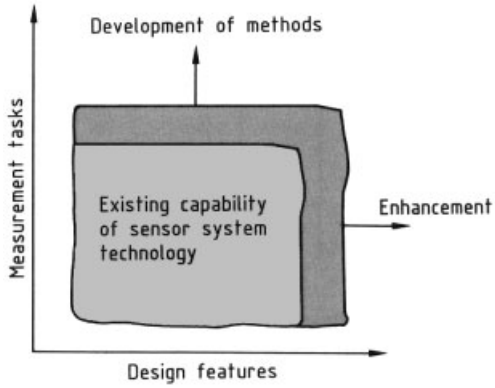

Figure 142. Continuing development of sensor system technology

the consequence that new measurement tasks are created or an old task is accomplished. The introduction of novel processes (e.g., biotechnology) involves new problems for sensor technology [439,440] (see Section 5.3).

The requirements that must be met in the further development of sensor technology for production processes fall into two independent classes (Fig. 142):

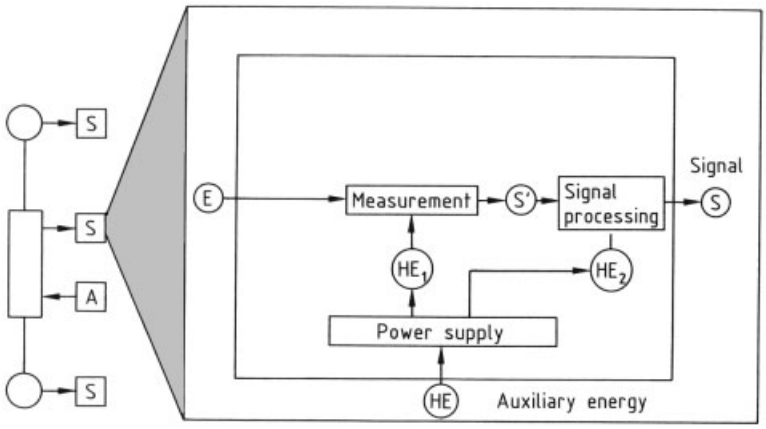

Figure 140. Information flow in a sensor system with auxiliary energy and signal processing [389]

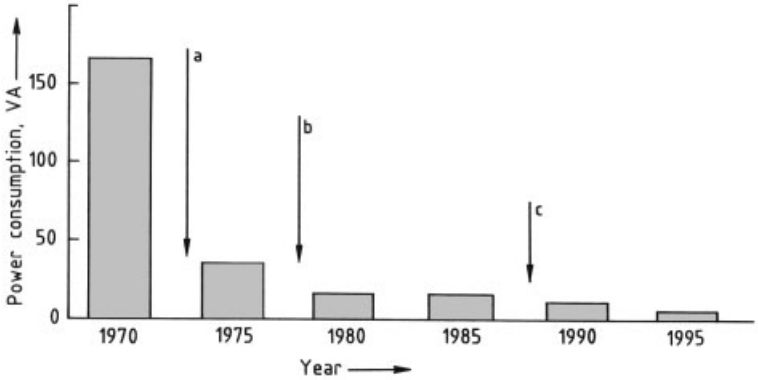

Figure 143. Development of the magnetic induction flowmeter (MID) illustrated by power consumption over time [441]
a) Introduction of switched d.c. field; b) Introduction of compact devices; c) Introduction of microprocessor control

The changed profile of requirements and the need for higher availability in computerized information processing mean that existing sensor systems must be enhanced. An example is the development of the magnetic induction flowmeter. In 20 years, the power consumption – a key factor in internal heat generation, service life, overall dimensions, etc. – has been reduced by a factor of 20 (Fig. 143) [441].

New problems require the development of new methods. Again, the realization of technical systems, often based on well-known methods of physics, is facilitated by the new capabilities of microelectronics and modern materials. Examples are systems using the Coriolis force to determine mass flow rate and others using the radar principle for level measurement.

5.2. Process Sensor System Technology

Information Flow in Sensor Systems. Section 3.5 showed that it is desirable to separate the precise description of the technical properties of an individual instrument from information about its technical function (Fig. 65). This makes it possible to maintain required functionality even when instruments are replaced.

Thus it is advantageous to apply the same formalism to information transformation (from the measured property of the material stream to the output signal) that was employed in structuring the material flow in production processes [386]. This leads to the idea of information flow in the sensor system, whereby a sharp distinction is made between the form taken by the desired information after each transformation step (corresponding to the product elements in the production process) and the metrological process steps required for information transformation (corresponding to the process elements of production).

As an example of a simple sensor system, Figure 136 illustrates the measurement of temperature with a jacketed thermocouple.

The measured property E of the stream is coupled to an intermediate medium and thus transferred to the latter.

The intermediate medium with property Z now interacts with the sensor element, which transforms Z into an electrical output signal S.

The information flow between the process and a simple sensor system is illustrated, in abstract form, by Figure 137 [389, 442].

It is usually necessary to amplify and condition the electrical output signal of the sensor element (Fig. 138):

The electrical output signal of the sensor element S'' is amplified to give a signal S'.

This amplified signal S' is linearized, mapped onto a unit interval, and made available as a signal S at the sensor system output.

The term *integrated sensor* refers to a system in which the functions of coupling, interaction with sensor element, and signal conditioning are integrated in a single component. If signal processing is included, the term *intelligent sensor* is used, especially when signal processing involves complex calculations and range matching steps [443].

In a simple sensor system with compensation (Fig. 139), an additional sensor is needed to de-

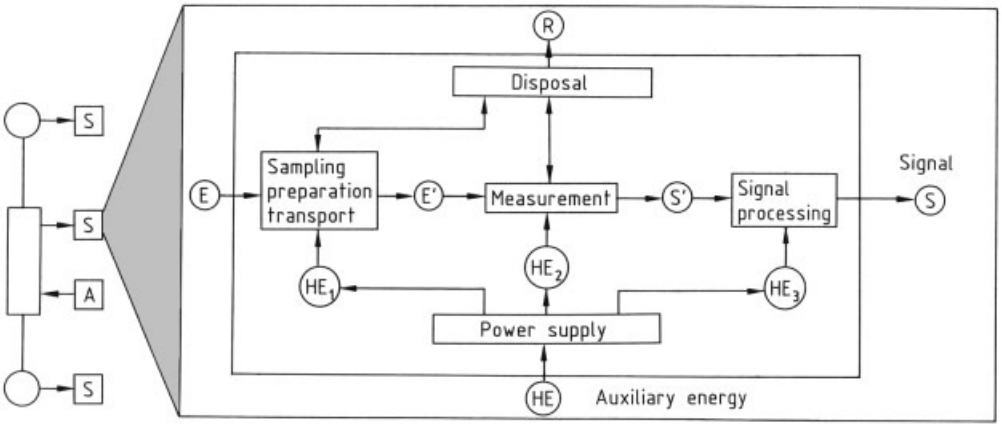

Figure 144. Information flow in a sensor system with sampling and other auxiliary functions (disposal, delivery, etc.) [444]

termine the parameter whose conditioned signal is incorporated into the signal processing. Auxiliary energy is generally required for this purpose; in abstract form, Figure 137 is elaborated to Figure 140. The energy needed for signal conditioning and signal processing is supplied to the sensor system by a power supply device that must be analyzed independently as far back as the design stage.

Sensor systems in production processes are often highly complex in structure. Figure 141 shows the conventional representation, for the example of a gas analyzer [389].

Here, the measurement proper is preceded by the functions "sampling," "sample transport," and "sample preparation." Sampling has the task of acquiring a representative sample of the stream to be measured. The sample must be transported without alteration or falsification to the analyzer, which performs the actual measurement. This stage of sample transport often involves conveying the sample over a long distance. The "sample preparation" function involves adapting the properties of the process stream at the sampling point to conditions compatible with the measuring instrument itself. The measured property of the stream – for example, the concentrations of certain components in the stream – must naturally be preserved in a well-defined way. The planning and execution of sampling, sample transport, and sample preparation are strongly dependent on the properties of the stream material, the analytical objective, and the measuring method employed. This strong application dependence is the reason that sensor systems for determining composition-dependent quantities are complex in structure and must be serviced by specialists.

Along with the sensor system modules just described, Figure 144 also shows the functions "delivery" and "disposal." The disposal step matches the properties of the sample stream at the analyzer output to the permissible conditions at the point of return to the stream. The delivery function modifies the auxiliary energies and substances such that they can be utilized in the sensor system. Both these functions must be included in the diagram of the entire system because the correctness of the measurement also depends on them and because they are potential sources of error for the sensor system as a whole.

On grounds of economy, it may be desirable to withdraw samples at various points in the process and measure them with a single analyzer. The functions of sampling, preparation, and transport are then preceded by a measuring point selection step (Fig. 145). In combination with signal processing for each sampling point, this function correlates the analyzer output signals (S_1 to S_n) with the properties to be determined (E_1 to E_n). Conversely, several analyzers can determine a variety of values on a single sample by accessing the same sample preparation.

As in the case of the production process, monitoring and control systems must also be available for complex sensor systems. All complex sensor systems include for this purpose auxiliary devices such as displays, pressure gauges, flowmeters, or valves for the admission of cal-

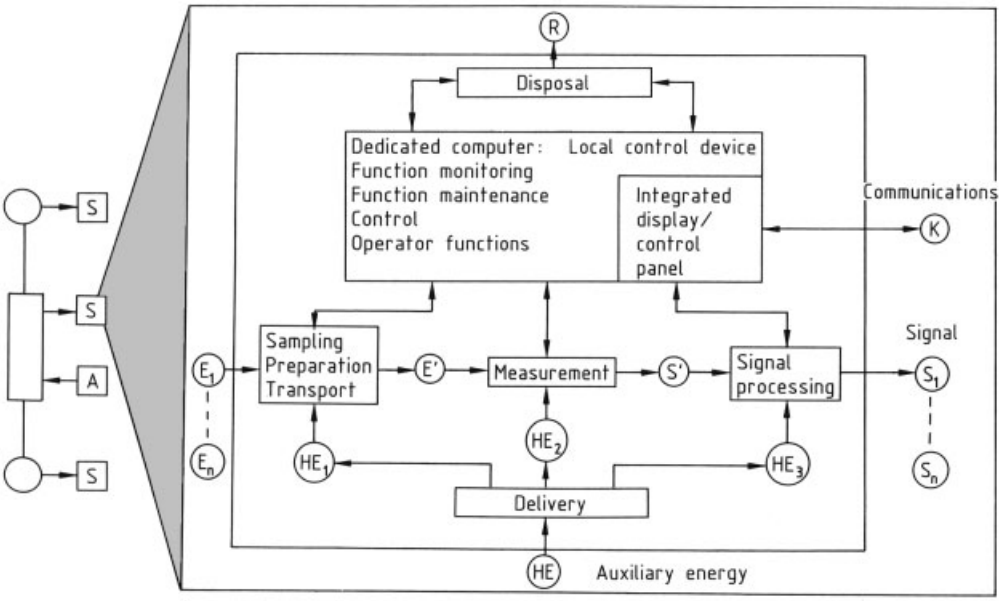

Figure 145. Information flow in a sensor system with sampling, auxiliary functions, and integrated dedicated computer [444]

ibration substances. In more abstract terms: the background functions grouped together as "functional check, control of maintenance procedures, and setting of parameters" must be supported as a minimum [442].

These functions are performed manually in conventional sensor systems. At intervals specified by the maintenance strategy, the system is inspected and serviced (e.g., calibrated) by maintenance personnel. However, analogous to the production process, this routine activity is increasingly being automated (Fig. 145) and turned over to a dedicated computer system that monitors all the basic functions of the system, controls procedures, and executes corrective actions. For this, the dedicated computer requires system-internal sensors and actuators. The computer furnishes status information about the sensor system and can execute incoming instructions, for example, to come into action in an intermittent mode (on-request or on-event). The outcome is a general structure for complex sensor systems. The information blocks K and S are the basis of the protocols for the future field bus.

"Intelligence" and Sensor Systems. In an intelligent sensor system, the functions discussed above are integrated or are supported in such a way that it can be regarded as a sensor system with communication capabilities. In Figure 146, these links are provided by a field bus, whose function could equally well be carried out by the use of smart technology based on the HART protocol.

In what follows, the functions of sensor systems are described as classical or intelligent.

Classical functions include:

- Sensing functions, i.e., transformation of physical and chemical effects into a signal suitable for downstream processing
- Signal conditioning (e.g., amplification)
- Signal output (part of the signal processing, such as A/D conversion, takes place in the local or process-level components)

Intelligent functions include:

- Signal processing
- Interconnecting measurements to expand information
- Auxiliary functions for on-line metrology
- Automatic execution of maintenance functions
- Failure strategies (see Fig. 138)

Figure 146 illustrates classical and intelligent sensor systems and their incorporation in higher-order information-processing systems. The use of a field multiplexer for the connection to the process monitoring and control system bus,

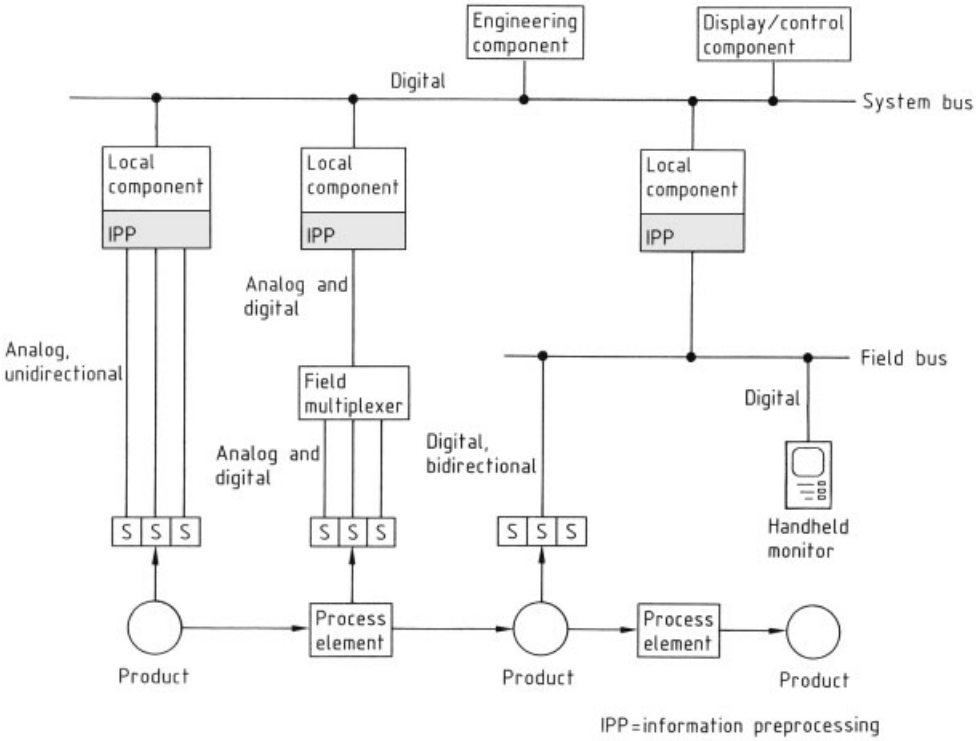

Figure 146. Movement of information processing nearer the process level

which is already a practicable transition approach (even in explosive environments); see Section 5.5.

The intelligent functions of a sensor system can be employed in a five-point program:

- Making all the needed information available
- In a form suited to the application
- At the proper time
- Ready for downstream processing
- Even in extraordinary process states

The goal of this development is to devise autonomous field-level devices that will automatically perform functions such as the maintenance functions of DIN 31 051 [445] discussed above. Intelligent process pH sensor systems can already carry out functions such as electrode cleaning, zeroing, calibration, flushing, and so forth, either "on-event" or "on-request."

Communication with Process Sensor Systems. Consideration of the functional scope of complex sensor systems leads to the question of how the sensor system can appropriately communicate with information-processing systems. The output signal categories M (messages), B (instructions), and S (signals) are the information blocks involved in communications (K) with process control and automation systems. A straight point-to-point connection with communication in one direction does not meet modern requirements. The need for bidirectional, open communications structures, and messaging services is essentially derived from the functions discussed above. More detailed information can be obtained from the literature on field buses [446, 447].

The connection of a classical sensor system to a local component (LC) is a unidirectional link employing a uniform signal such as 0 – 20 or 4 – 20 mA (Fig. 126-9). All that is transmitted is the measurement signal or a report of a device failure. Signal processing, that is, conversion to a value with a physical unit, takes place in the LC. Maintenance tasks must be performed locally in the sensor system. According to DIN 31 051 [445], these include inspection to evaluate the actual status, maintenance to preserve the nominal status, and repairs to restore

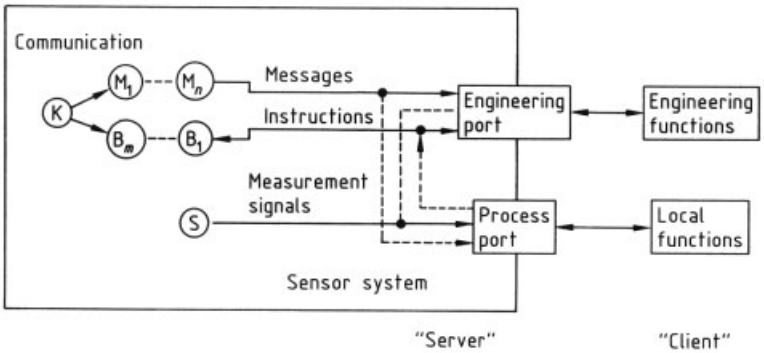

Figure 147. Functional breakdown of the sensor interface K (communications; see Fig. 145) into "process port" and "engineering port" [389]

the sensor system to nominal status. General discussion of these points can be found in Section 11.4.

If a dedicated computer performs the function monitoring of the intelligent measuring system and carries out simple maintenance procedures, using messages to report on both activities, the process sensor system acquires, functionally, a second interface (in addition to information delivery to the control room via the "process port"), which mediates information exchange with the engineering team via the "engineering port." This functional distinction is illustrated in Figure 147.

Given the wide range of methods employed in process sensor system technology and the broad spectrum of application-specific hardware configurations, the only way to achieve system-wide reliability and effective maintenance is by unifying the two interfaces, process port and engineering port. A flexible dedicated computer system, matched to the structure of complex process sensor systems, can organize the required information exchange from and to the sensor system. In this way, expensive custom solutions can be avoided and replaced by acceptable, standard solutions.

Information Exchange with Local Functional Units. According to SCHMITT the following holds for information exchange between the field level and the process monitoring and control system: "As little as possible – as much as necessary" [437]. This requirement governs the functional sensor interfaces shown in Figure 147 and is discussed further in Figure 148 in application to the process port.

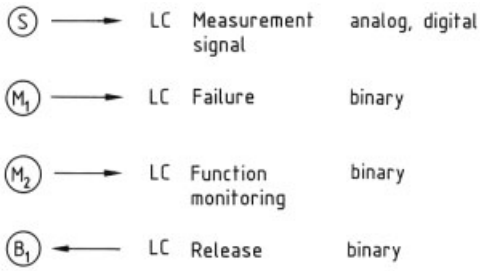

Figure 148. Information exchange between sensor system and local process control system component (LC) [389]

In relation to local (process-level) functions, the sensor system has only one task: making the measured value available in proper form. If the sensor system ceases to respond as specified because of an upset of malfunction, this fact should be reported in a group message "Failure" (M_1), which is to be forwarded directly to the display and control component of the control system (see Chapter 7). Any still-pending measurements are then rejected as unusable. This assessment of sensor information presupposes that the sensor system monitors its own internal technical functions.

If internal sensor function monitoring is initiated automatically or manually, measurement signals may appear momentarily that are correct in content but have nothing to do with the process and product properties being measured. This state is indicated by the status message "function monitoring" (M_2). Because of the absolute priority of the information measurement signal, it may be necessary to permit such nonavailability due to internal programs in the sensor system only when they can be tolerated by

the control system. An instruction "release" (B_1) from the process control system to the sensor system is a suitable way of handling this situation; for example, the instruction can inhibit or enable the calibration procedure.

The power and low cost of modern microprocessor boards make it possible to use them locally to optimize internal operations in sensor systems without having to shift such functions to central systems. In this way, straightforward system structures can be created. For the sake of manageability, information exchange with process-level functional units of the process monitoring system should be restricted to the measurement signal plus a few status messages and instructions. The handling instructions and procedures associated with each type of status message or instruction should be defined as a strategy in the process control system (for status messages) or in the sensor system (for instructions).

Information Exchange with Engineering Functional Units. The maintenance functions have the task of keeping up reliability and insuring availability. Routine inspection and function monitoring jobs should be automated and turned over to the sensor system itself whenever technically possible at reasonable cost. When service is needed or when a malfunction occurs, the engineering group is requested. From a functional standpoint, access to the sensor system and necessary information exchange take place over a separate interface, labeled "engineering port" in Figure 147.

If a process sensor system is fully equipped with self-monitoring functions and devices to maintain the functionality of both the instrument proper and the sample preparation part, the volume of information produced is difficult to manage without structuring. Engineering becomes inefficient and personnel find themselves overloaded if the separate functional concept and distinct operator interface of each instru-ment type and each manufacturer have to be kept in view.

The capabilities of future field-level communications are discussed here, based on analyses of information structuring for process analyzers [389].

For efficiency in engineering, it is proposed that information exchange (see Fig. 149) be accomplished in three functionally distinct information "planes" differing in complexity. A status summary is needed for the instrument proper; this includes a group message "failure" when the measuring device is no longer reliably available; a status message "function monitoring" when the sensor system is occupied with itself as a result of a manual intervention or automatic program; and the display "measurement" so that the engineer can make an on-site assessment. If the sensor system is equipped with self-monitoring functions that can detect a failure (for example due to drift in the analyzer or exhaustion of the test gas supply), the status message "maintenance required" should also be displayed.

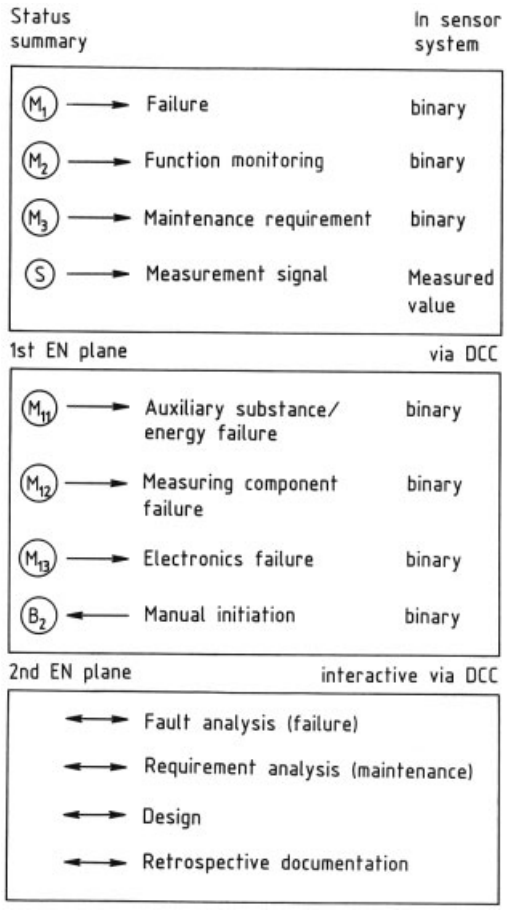

Figure 149. Structuring of the interface to engineering [389] DCC = display and control component; EN = Engineering

The inspection, maintenance, and repair of process sensor systems need not always involve specialists. It appears useful to be able to distinguish whether a given malfunction can be

remedied by simply replenishing auxiliary substances or replacing an electronic component, or whether the device can be put in service again only after a specialist works on the sample handling system or the physical part of the instrument.

For this purpose, the first engineering (EN) plane includes simple, binary diagnostic indicators. The "failure" message may, for example, be broken down into just three categories: M_{11} if there are problems with the delivery of auxiliary supplies or energy; M_{12} if there are problems in the measuring section, including sample preparation; and M_{13} if a failure occurs in the electronics. For general maintenance, simple instructions such as manual initiation of a function check should be integrated into this EN plane.

This initial diagnostic feature requires communications with the sensor system above and beyond a simple status indication. These communications are carried on through a display and control component near the sensor level. The display and control component can take the form of a keypad with display integrated into the sensor system; an engineering computer some way off; or a portable unit such as the handheld monitor shown in Figure 146, which is plugged into the sensor. As to what technical and functional capabilities exist and how they can be exploited, see [446,448].

Detailed diagnosis of the causes of malfunctions, which can only be performed by process sensor specialists, should be carried out in a second engineering plane, via one of the functional units. This kind of diagnostics includes single-fault analysis upon a system failure, as well as detailed need analysis upon a maintenance request. The design and commissioning of complex process sensor systems often include configuration and parametrization. These tasks are also performed by specialist personnel and thus form part of the second EN plane. For the manageability of process sensor systems, it is important that the system data be kept updated and that retrospective documentation be maintained.

When such a ternary structure is used in the EN interface, standardization across vendors and instruments appears possible at least for the status summary and the first EN plane [449]. For efficient maintenance, it would also be desirable if information exchange through the second plane could be structured in a consistent way and mediated by operator interfaces as similar as possible.

The "sensor-level display and control component" function is increasingly being implemented with a portable computer (laptop PC). If a standardized operating system and graphics environment are employed, vendor- and device-specific parametrization and diagnostic routines can be loaded and run in software. For process analyzer technology at least, such a consistent engineering tool offers substantially greater benefits than a variety of individual solutions with their associated library of manuals (see also [450]).

The communication ports of field-level devices should generally be implemented on the basis of the object-oriented model [451] defined by the MMS (Manufacturing Message Specification; see Chapter 8 and [452,453]). The field device plays the role of server and performs the services requested by the client; see Figure 150.

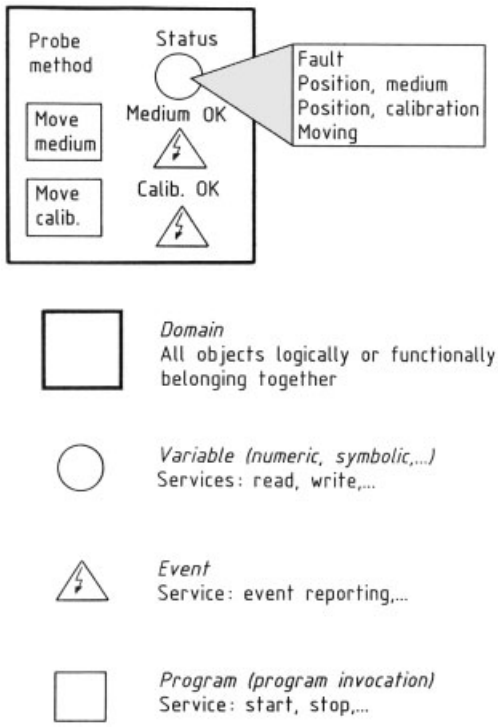

Figure 150. MMS communications objects

Linked with the cited background functions and their use for process control, maintenance,

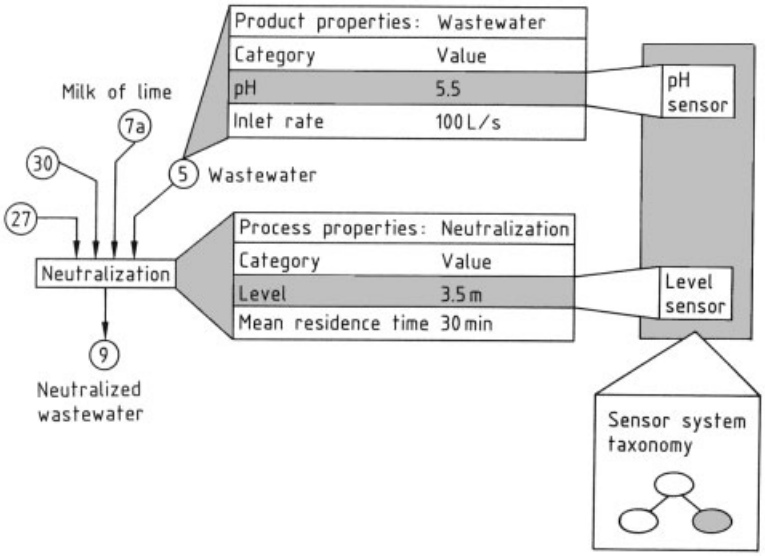

Figure 151. Determination of instrumentation costs

and design (configuration) tasks is the introduction of general, standardized descriptive languages for field-level devices; these are called Device Description Languages (DDLs). The objective of manufacturer-neutral field-level functions is that the field device to be operated should be addressable without any special knowledge of the physical device, even if devices are modified or replaced by hitherto unknown devices. Standardization efforts toward a universal device description are focused on creating a proposal for the ISA SP 50 international field bus standard [446, 454].

Furthermore, it is not sufficient to solve the problems of field-level communications without taking up questions of what user interfaces are appropriate for sensor systems and how these interfaces can be standardized [450, 388]. The application connected above the communications services must be dealt with at the same time. On practical grounds, a variety of devices having the same function must not have unlike operator interfaces [446, 455].

Moreover, the rapid growth in information demand and the need for auxiliary energy in process sensor systems – which may well exceed the permissible limits in intrinsically safe systems – have made it essential to consider separating the information and energy channels (see Section 5.5, Chapter 8, and [447]).

Determination of Instrumentation Cost. Given the information economy of a process modeled with the phase model (see Figs. 125 and 36), the instrumentation outlay can be determined at any level of detail (see Chapter 2).

Consider, for example, the process step of neutralization in a wastewater treatment process. Among the quantities to be determined are the pH and the inlet flow rate of the wastewater (Fig. 151). From the categories and attributes of the process property (here the wastewater pH), the property can be methodically assigned to a suitable sensor system. A CAE tool for this purpose, accessing a source database for sensor – actuator systems, is discussed below.

Figure 152 shows an example of how to structure a taxonomy of sensors for process engineering. The user is guided by generic attributes: application features and design features. For the wastewater pH example mentioned above, a sensor system can thus be defined that has the application features on-line measurement, determination of a collective property, measurement in the product stream, and so forth. In general, the requisite instrumentation cost can thus be determined from the information demand obtained from the phase model. Sensor or actuator systems are selected on the basis of the category of process and product properties, with the aid of a taxonomy of such systems (navigation, hypermedia, etc.). Manufacturers still offer

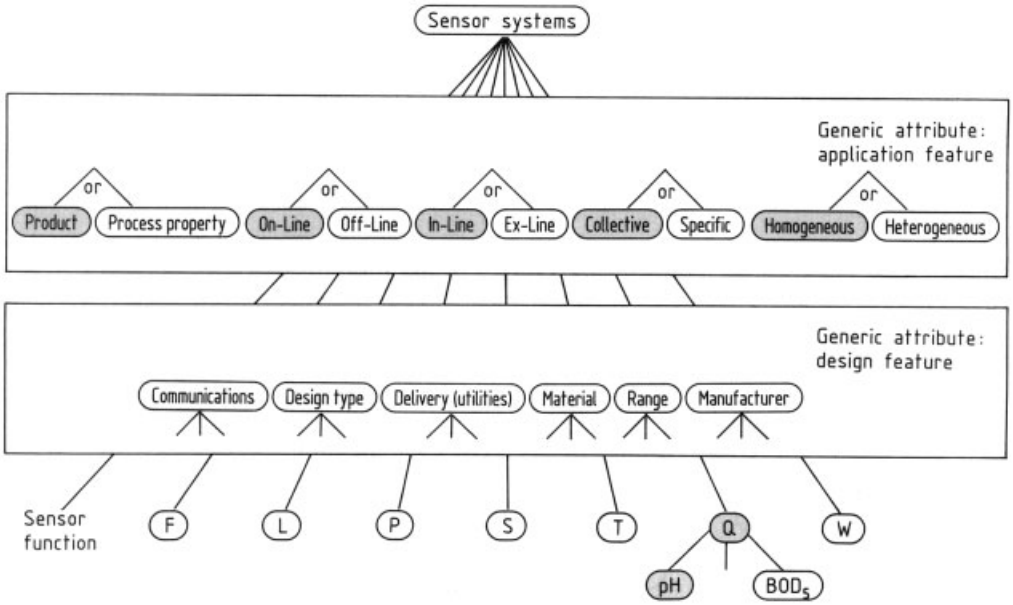

Figure 152. Taxonomy of process sensor systems: System properties required for determination of pH and BOD_5

monogram-like selection aids [456], which will certainly be replaced soon by expert-systems tools.

Taxonomy of Sensor Systems. The classical listing in book form, arranged by apparatus features, is no longer adequate for presenting an up-to-date survey of application options for modern sensor systems.

Nevertheless, such proven classics as Profos [457], Strohrmann [458], Hengstenberg [459], or the Kohlrausch [460] are indispensable aids. The journals TM, atp, and MPA have presented thorough surveys of problems and solutions (see also [388, 461–463]). However, in general, encyclopedic treatment must be replaced by structuring.

A priority task for structuring available sensor systems in a way relevant to process control engineering is to emphasize those features that actually bear on potential applications. As discussed previously, these features cannot be identified without a knowledge of process conditions met with in practice; that is, an apparatus-oriented analysis alone is insufficient. Instead, the conceptualizations introduced to structure information characterizing a production process lead to a powerful technique. In future, this approach may serve as a basis for developing powerful CAE tools for the systematic instrumentation of new plants.

Figure 9 shows that, if the distinction between function and substance – or function, function-generating procedures, and devices (Fig. 65) – is taken into account, then the world of sensor systems can first be represented in terms of the tasks meaningful in certain industries; see Table 3 [464] and the lower row of Figure 153.

Table 3. Letter symbols for process control point functions [464]

Letter symbol	Group 1: measured values and other input quantities		Group 2: processing (as trailing letter)
	As first letter	As additional letter	
A	alarm		
C			controlling
D	density	difference	
E	all electric variables		sensing element
F	flow rate	ratio	
I			indicating
L	level		low
P	pressure or vacuum		test point connection
Q	quality, e.g., analysis, concentration, conductivity	integrate or totalize	integrating or summating
R	nuclear radiation		recording
T	temperature		transmitting

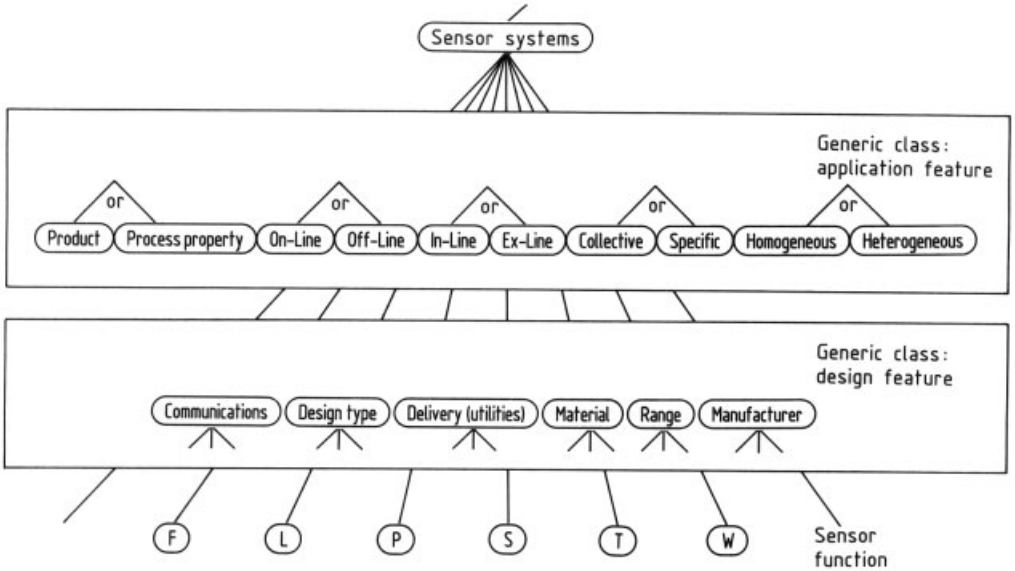

Figure 153. Taxonomy of process sensor systems: Breakdown of sensor functions

The E – R (entity-relationship) modeling technique described in Chapter 2 is employed, at various levels of hierarchical detail, for this purpose. Such an ordering rule, as explained in Chapter 2, is known as a "taxonomy" (Greek: taxis = order).

Generic Class: Application Features. The generic class of application features or task features, defined from the functional viewpoint, becomes the uppermost level in this context.

The following are employed as classification criteria:

- The attributes of the object classes of technical processes
- The degree of indirectness with which the measurement signal is determined
- The ratio of relaxation times between the acquisition of the measurement signal and the variations of the property of the stream being measured (see Section 5.1)
- Specification of the components of the substance measured

Under these criteria, sensor systems are now classified according to whether they determine product properties or process properties, whether they operate on-line or off-line, whether they are employed in-line or ex-line, whether they can determine collective or specific properties, and whether they are usable in homogeneous or heterogeneous phases.

These generic features are explained briefly below:

- On-line: A continuous correlation is possible between information provided and properties of the process or product. The essential condition for on-line measurement is that the relaxation time of changes in the process or product properties must be markedly longer than the time required to acquire, convert, and deliver the information $[\tau_{\mathrm{R(property)}} \gg \tau_{\mathrm{R(sensor)}}]$ (see also Section 5.1).
- Off-line: Discontinuous measurement; determination of sampling intervals requires knowledge of process characteristics (e.g., kinetics).

This pair of attributes describes the problems the (process) sensor system must handle in terms of communications.

- In-line: The systems used to prepare information about the process or product properties interact directly with the product stream (or a bypass stream).
- Ex-line: The required sampling operation (extractive) makes it necessary to take into account additional changes in the properties of the stream substance.

Questions of positional and temporal representativeness are extremely important here. This

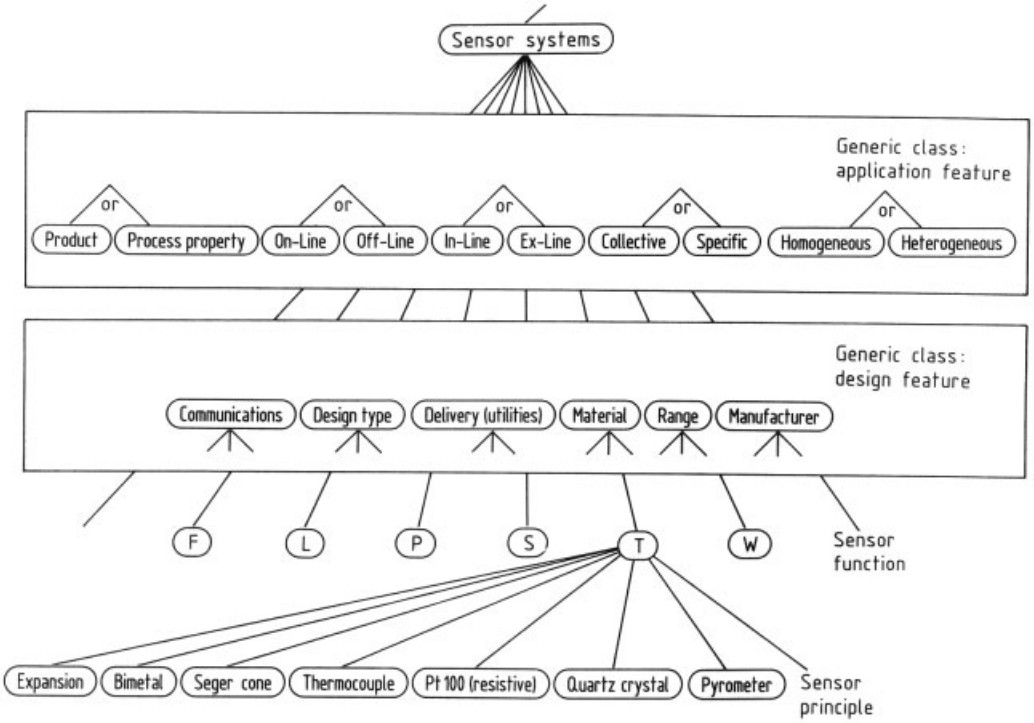

Figure 154. Taxonomy of process sensor systems: breakdown of sensor principles for the example of temperature measurement

pair describes the problems the system must deal with in terms of metrology.

- Collective: The quantity measured can provide information about the quality of the stream as a whole (e.g., density or flow rate), its reactions (e.g., pH, conductivity), and the relationships between its properties (e.g., relative concentration).
- Specific: The "manifestation of a property" determined makes it possible to gain qualitative and quantitative information about individual components of the stream; such techniques include analytical methods such as spectroscopy and chromatography.

This pair thus states clearly whether the chosen sensor system yields information about atomic/molecular, mechanical, and electromagnetic mechanisms.

- Homogeneous: The sensor system can acquire information only in homogeneous phases (e.g., flow-rate measurement with a mass-flow anemometer).
- Heterogeneous: The sensor system permits the determination of the measured quantity even in heterogeneous states of aggregation, such as dispersions (e.g., flow-rate measurement by the Coriolis principle).

This pair accordingly requires an unambiguous description of the sensor system functionality with regard to its possible use in multi-phase environments.

Generic Class: Design Features. The next level consists of a further class of features which basically summarize information about the design of the sensor systems under consideration.

Sensor systems are classified by communications capability, design type, utilities required, material of construction, range of measurement, and manufacturer. Individual features can be further structured. The dependent lines in Figures 153 and 154 illustrate this point. In principle, the depth of structuring is unlimited.

To illustrate characteristic features of importance in field-level devices, the safety requirements on process sensor systems are discussed at this point. If "safety" is defined as a risk less than the limiting risk [403], as in Section 5.1, safety-relevant properties of a process sensor system

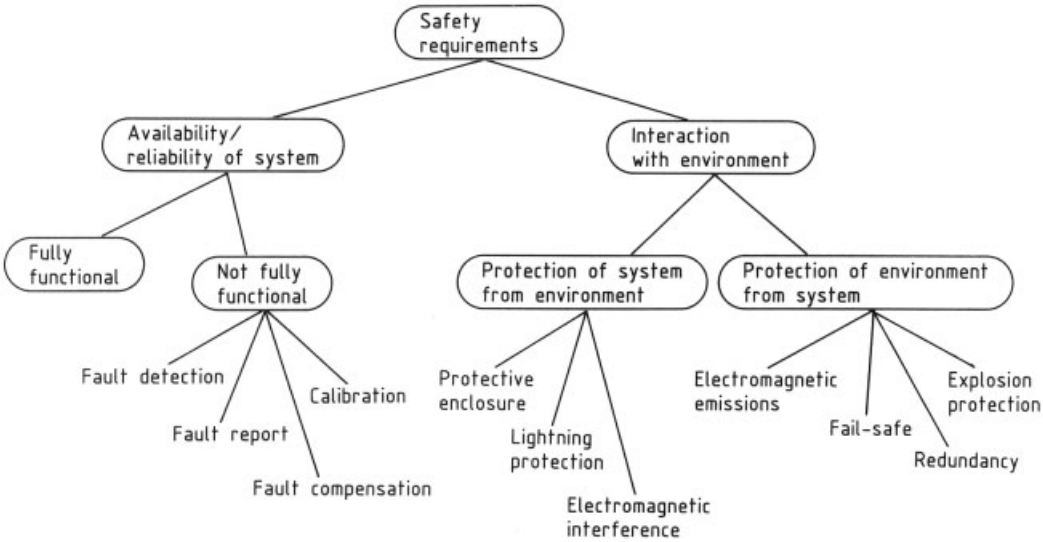

Figure 155. Safety requirements on process sensor systems

can be divided into two classes with regard to possible sources of danger:

- Properties related to insuring the availability and reliability of the sensor system itself
- Properties relating to reducing or preventing danger interactions of the sensor system with the system environment

These contributions to safety are correlated with possible technical safeguards in Figure 155. The potential dangers due to interactions with the environment are divided into those that require protection of the sensor system against the environment, and those requiring (possibly indirect) protection of the environment against the effects of the system.

As Figure 155 shows, present-day principles, guidelines, and standards on insuring the safety of process control hardware can be classified in the same way:

From a safety standpoint, enhanced availability and reliability gained through fault detection, fault reporting, fault compensation, and automatic calibration figure in the tendency toward increased functionality in field systems through integration of information-processing components, as discussed in this section. An initial and crucial requirement on the safety features of a process sensor system will be the ability of the system to verify and, if appropriate, report its complete functionality.

The protection of live or moving parts of an electrical device against foreign solid bodies or liquids is required by all national and international standards, for example, [465]. Because they are represented by the letter code IP (International Protection), these provisions are also referred to as IP protection types.

A topic still under vigorous discussion, in contrast, is how to insure electromagnetic (EM) compatibility. In German and European standards, electromagnetic compatibility is defined as the ability of an electrical device to function satisfactorily in its environment without unduly influencing this environment [466,467]. In Figure 155, measures to insure electromagnetic compatibility are divided into safeguards against EM interference induced in the device (static discharges, electromagnetic fields, and fast transients) [468] and safeguards against EM emissions into the environment, in particular human exposure, due to fields generated by the device or system [469].

Another type of protection incorporated into the field-level device is explosion protection [470]. The avoidance of flammable materials and explosive mixtures are usually the primary safeguards against explosions [458,471]. With regard to requirements on field-level systems, these practices are of relevance to design and material selection. Primary explosion safeguards must also be considered when analyzing the use

of sensor systems, especially the use of process control resources for plant security [401]. This area of applications includes all process control hardware employed for monitoring, protection, and damage limitation [472]; see also Section 10.4.

Secondary explosion protection in field-level process cotnrol systems includes all measures taken to prevent ignition of explosive substances by preventing energy transfer. The regulation governing electrical devices in explosion hazard areas distinguishes seven types of protection. Among these, intrinsic safety [473] sets rigorous criteria on integrated circuits in modern sensor systems (see Section 5.5).

For redundant and fail-safe design, see Section 10.4. Lightning protection practices are generally based on suppressor circuits [474].

The lowermost level is concerned with the physical, chemical, and other principles on which sensors operate. The commonly used nomenclature is set forth in DIN 19 227. Figures 153 and 154 show a selection of possible sensor principles; all others are indicated by the dependent lines. As in the case of the design features, further structuring is possible.

The foregoing application-oriented description can serve as a basis for a conceptual, object-oriented classification of sensor systems, differing sharply from the device-oriented and method-oriented structures, as listed below.

Methods of measuring "classical" quantities:
Temperature
– Resistance thermometer
– Thermocouple
– Radiative pyrometer
Pressure
– Elastic elements with capacitive, inductive, piezoresistive or piezoelectric signal conversion
Flow rate
– Magnetic-inductive
– Differential Level
– Hydrostatic
– Electrical and electronic
– Radiometric pressure

Nonspecific methods:
Density
– Radiometric
– Other
Viscometry
Thermal conductivity
Optical scattering and extinction
Acoustic

Group-specific methods:
Electrical conductivity
Flame ionization
Photoionization
Methods based on group-specific auxiliary reaction
Calorimetry

Specific methods:
Paramagnetism
Photometry
– Absorption (UV, VIS, NIR, IR)
– Diffuse reflectance
Gas chromatography
Liquid chromatography
Mass spectrometry
Electrochemical methods
– Potentiometry
– Amperometry
Titration
Analytical methods based on auxiliary reaction
– Chemoluminescence
– Absorption in aerosols or liquids
Special moisture measuring techniques

The concept presented here makes it possible to create flexible CAE tools, which will gain increasing importance in the design of process monitoring and control systems. For example, interactive and iterative equipment selection by the design engineer is aided by computer-generated querying of the features introduced here [475]. There is no need of once-for-all solutions or a priori optimization criteria. Hypermedia concepts can serve as navigation or explanatory aids [476].

Determination of Communications Costs. Field-level communications systems (i.e., field bus systems) will be essential components of a networked system. The design of future field bus systems will be based on the information economy derived from the phase model (Fig. 156). A suitable field bus system will be chosen on the basis of a taxonomy, here the taxonomy of field bus systems.

Typical message lengths for sensors and actuators, as well as maximum cycle times derived from the process and product properties,

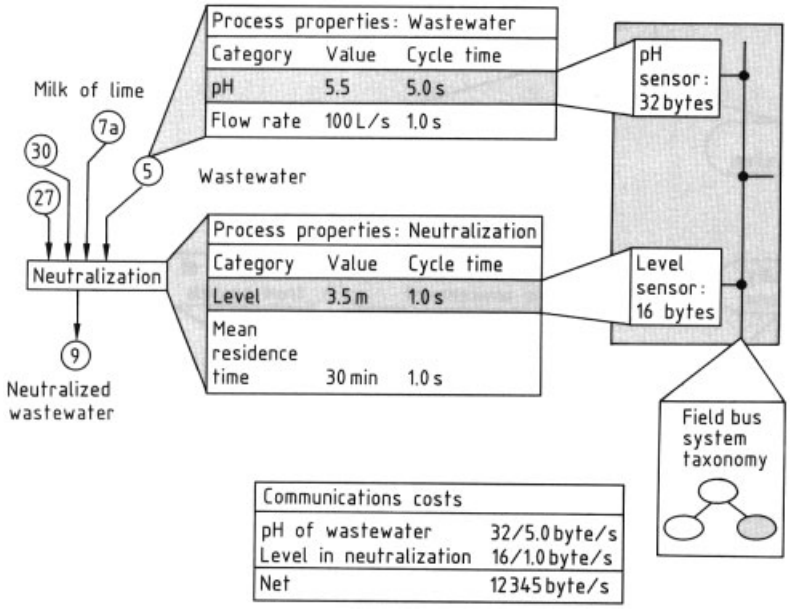

Figure 156. Determination of communications costs

are balanced in order to arrive at the communications costs for the system as a whole. The number of field bus systems and the number of segments in these systems are set on this basis. The result is a picture of a CAE tool for the information-oriented design of communications systems [477]. The advantages of information modeling dictated by the process are again manifest here.

A proposal for a taxonomy of field bus systems is shown in Figure 157 [477]. The generic attributes are architectural features and the communications structure. Monomaster systems have a pure master/slave structure with strictly hierarchical vertical communications between planes. Multimaster or peer-to-peer systems, in contrast, permit horizontal communications within a plane. For the wastewater treatment example of Figure 151, both monomaster structures within a plant section and multimaster structures for communications between plant sections are possible. Communications between local components and sensors (e.g., pH and flow sensors) within a plant section might be based on the PROFIBUS or the DIN measuring bus, while communications between plant sections and with higher-level central functions might be implemented on the basis of PROFIBUS.

General Requirements on Process Sensor Systems. A complete introduction to process sensor system technology would also cover selectivities, detection thresholds, time dependences, sensitivities, stabilities, and so forth, all qualities that have to be defined for each sensor system [442, 478]. These are technical specifications that differ from one measurement task to another because of differences in service conditions, products, perturbations, and so forth.

The profile of requirements on sensor systems is strongly influenced by the downstream computer processing of the information:

- In-line and ex-line sensor systems must not disturb the process.
- It is essential to separate the energy channel from the information channel.
- Communications capabilities and system capabilities must be insured.
- Electromagnetic compatibility is also required; in some cases, external constraints dictate “intrinsic safety” as well [386]. See the generic class “Design Features.”
- Special conditions in chemical production (corrosion, abrasion, etc.) must be taken into account when choosing the sensor principle and construction material.

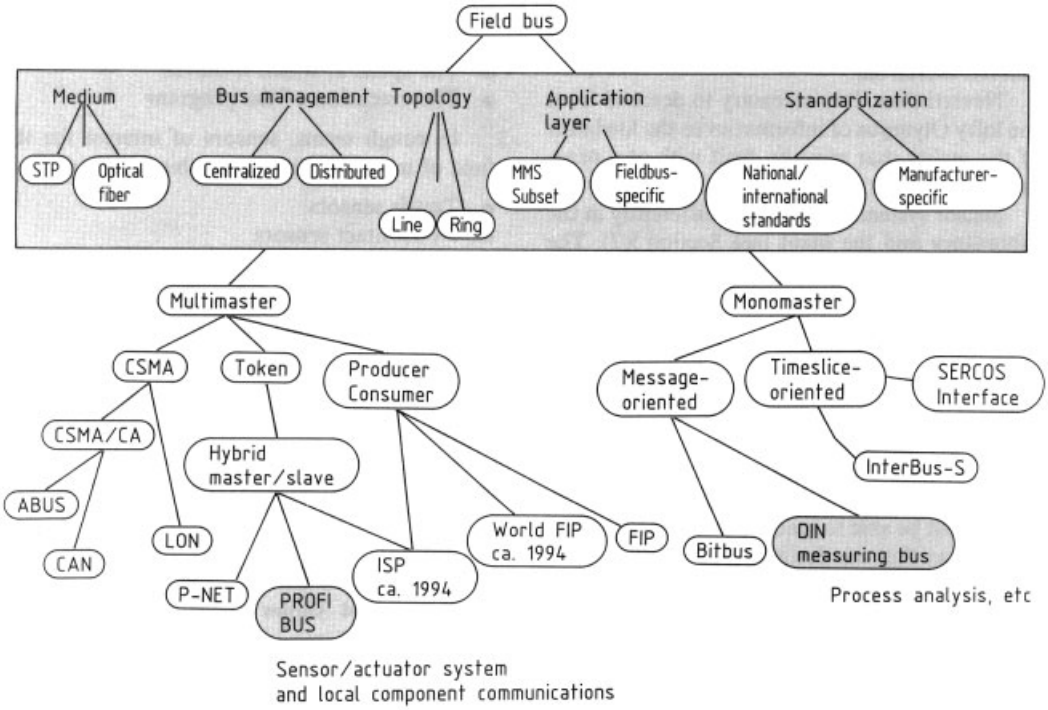

Figure 157. Taxonomy of field bus systems [477]

– One special requirement for in-line sensor systems is that they must remain functional even under extraordinary process conditions (e.g.,upon a change in the state of aggregation).

Economic Aspects. From the standpoint of the process control engineer who designs, commissions, and maintains a sensor system, meeting the stated requirements is closely linked to the lifetime costs of the sensor system. The term "cost of ownership" expresses the fact that the key criteria include not only the technical specifications and the resulting investment costs, but also longer-term aspects of sensor system use, such as training and maintenance costs. As discussed in Section 2.1, a holistic approach [479] must be taken from the outset.

In general, however, the main stress should be on the need for a process control station and a process sensor system. Total costs for a process monitoring and control system are largely determined by this sum. What is decisive is not the initial cost of devices and systems; the field installation costs make up a respectable fraction of total costs (see Section 10.2 and Fig. 320). Finally, the costs of control system components depends in part on the number of process control stations.

Often a fear of safety aspects and regulations leads to an exessive number of process control stations being set without a careful and critical analysis of the necessary equipment availability (Section 10.4).

Inadequate knowledge about the process at the time when costs are determined can, on the other hand, result in too few process control stations or incorrect equipment selection. This problem often cannot be corrected until commissioning, by which time costs are very high.

The departures from the original design may give nonspecialists the impression of "overengineering," especially if process control engineering resources must be used to correct solutions that are nonoptimal to poor from the process engineering standpoint. Overengineering due to ignorance is the most common cause of this phenomenon.

It is therefore essential that some tool such as the phase model of Figures 36 and 37 be employed at the bidding specification stage, in order

to take full account of all information needed for the production process.

5.3. Sensor Systems for Special Applications

In accordance with HESSE's suggestion [394] to put systems concepts at the beginning, Section 5.1 deals with systems integration of sensors or sensor systems, while Section 5.2 discusses process sensor systems from the standpoint of internal information flows and the taxonomies derived therefrom.

Nevertheless, it is necessary to descend from the lofty Olympus of informatics to the lowlands of the reality that must be dealt with at a pragmatic level.

Sensor systems behave very differently in the laboratory and the plant (see Section 5.2). The tasks they must perform, and therefore the requirements they have to meet, vary from one application to another.

Sensor Systems in Manufacturing [480]. For example, a sensor system in a manufacturing industry must provide an industrial robot with information about its surroundings that are relevant to the operation but not predictable. The robot must be able to react to this information in such a way that the desired result is achieved or maintained automatically. This is why a variety of sensors are employed to determine physical quantities and states, transform them into (generally) electrical signals, process them, and interpret them.

As a rule, information transmitted to the robot influences

- The position and orientation of the robot's manipulator or tool
- The speed at which it moves
- The execution of the program

In rough terms, sensors of interest for the field of industrial robots can be classified into

- Tactile sensors
- Noncontact sensors

Another principle of classification is the complexity of the information acquired, allowing sensors to be grouped into

- Sensors that transmit binary signals
- Sensors that transmit one-dimensional measurements (forces in a preferred direction, distances)
- Sensors that transmit multidimensional measurements (forces of arbitrary orientation)
- Sensors capable of pattern recognition (workpiece identification)

Figure 158 shows a taxonomy based on these criteria.

Under each of the categories cited, a number of functional principles can be employed, depending on the prevailing boundary conditions (Fig. 159).

The linking of robots to image-processing, pattern recognizing sensor systems and to tactile, multidimensional force and moment sensors is an especially urgent problem for future robotics applications.

Important future applications that can be exploited only with the aid of powerful sensors for industrial robots include

- The machining of workpieces
- Assembly
- Emptying of containers
- Continuous joining operations such as welding and seam cementing

The emptying of containers is a typical application for imaging sensors, which have the task of identifying workpieces in the container, measuring their orientations and positions, and providing the industrial robot with the information required to grasp the workpieces correctly.

The areas of machining and assembly are important applications for tactile sensors.

One task that illustrates significant problems and requirements in connection with industrial robots is machining with tolerances specific to the tool, the workpiece, and the robot.

If the tolerances and deviations exceed a certain value, it is no longer possible to achieve satisfactory machining results within a series of workpieces. An optimally designed, and therefore workpiece-variable, control of the machining process is only possible if the most important toleranced quantities are determined by sensors and the process is automatically optimized.

SCHWEIZER and WARNECKE distinguish two groups of variables: technological variables oriented to the machining process, such as power

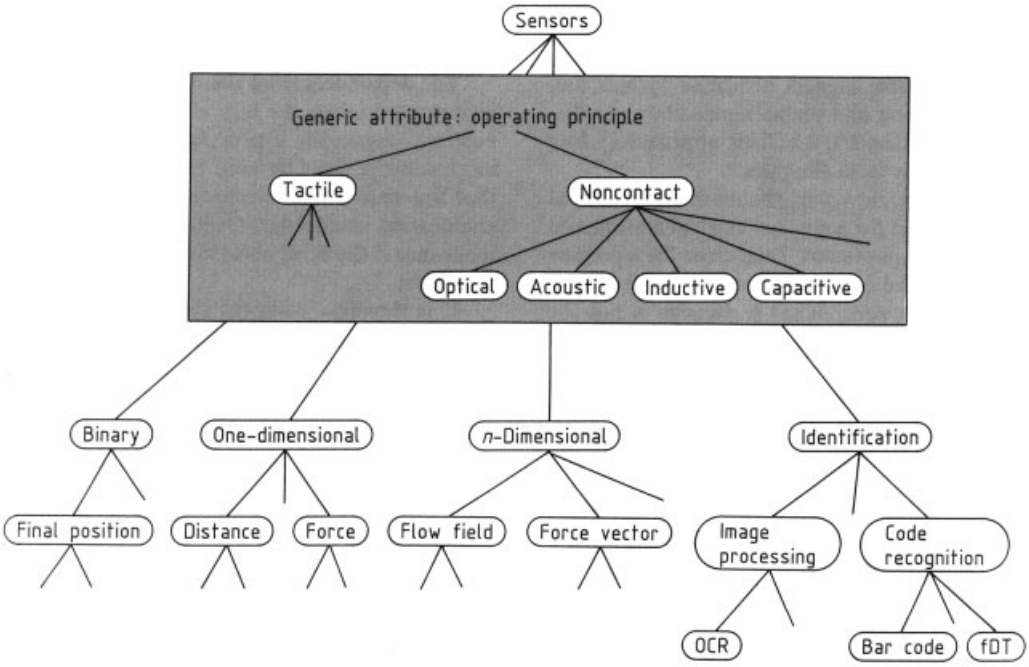

Figure 158. Taxonomy of sensors for manufacturing

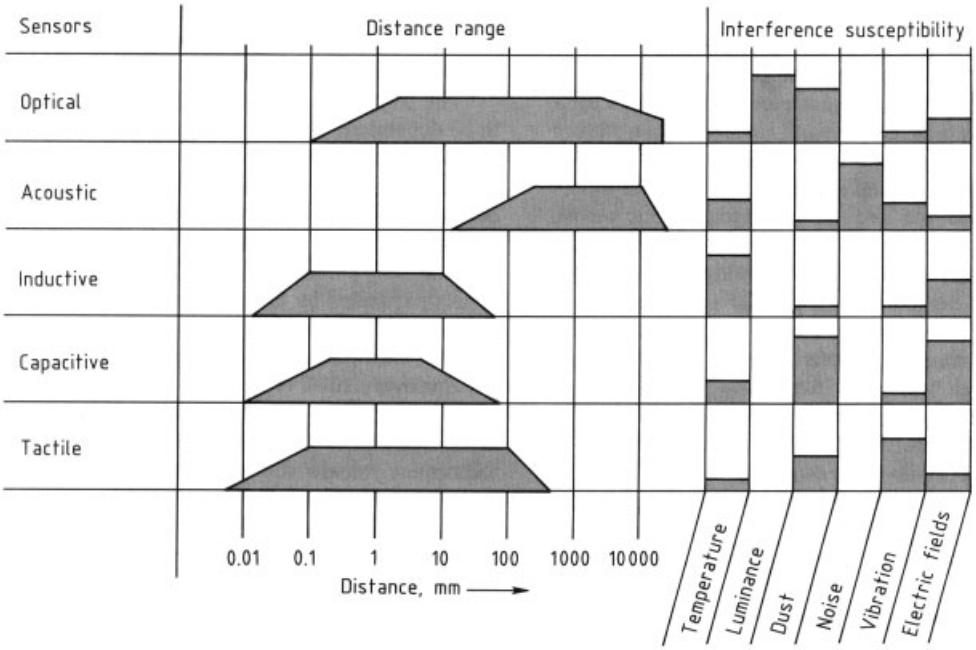

Figure 159. Distance measuring systems for industrial robots [480]

consumption and cutting forces, and geometrical variables such as tool wear, geometrical and clamping tolerances, and surface quality. The fundamental problem in such tasks lies in the high data rate and the mathematically demanding simulation process. Adaptation to a variety of tasks also complicates this application, which most commonly involves highly dedicated sensor systems.

Artificial sensors cannot offer the capabilities and flexibility of human senses, but they can reliably perform single functions in manufacturing. A survey of approaches in the field of quality assurance through production metrology can be found in [481].

In assessing future sensor technology, it is useful to compare biological and technical sensor systems [392].

In many areas technical sensors can be made substantially more powerful than biological ones. Furthermore, the spectrum of technical sensors is broader; that is, there are technical sensors for a wider variety of physical variables than nature has provided for among biological sensors. For example, technical sensors exist that can detect high- and low-frequency electromagnetic waves, electrical currents, and electric field strengths.

With regard to preprocessing by the sensor system, however, nature is far superior to artifice. Technical sensor systems at present are designed chiefly to detect events in preplanned technical processes. Accordingly, they can be more highly specialized for the expected events. Thus the knowledge that a foreseeable process event has taken place often comes solely from the fact that the sensor has responded. Scarcely any preprocessing is needed; the sensor "speaks for itself." By the same token, however, the system loses flexibility with respect to unexpected events. And this flexibility is particularly important for biological sensor systems, under some circumstances more important than high performance of individual sensors. Biological systems are designed to be more universal, while technical systems are more specialized. This notion is expressed in the term dedicated systems.

Most biological sensor systems are more "intelligently" constructed than artificial systems: Preprocessing takes place in amazingly powerful processing systems, with the result that vastly many individual items of information acquired by many identical sensors are reduced to summary information in a process of "supersign" formation. It is hardly possible to imagine the demand imposed on the processing system, not just by "concept-forming" logic but also by the coordination of the many sensors in arrays. In the eye, for example, the sensor array of the retina comprises millions of cone and rod cells. Nature again and again offers a very impressive demonstration to technology, showing how flexibly such array systems can react to unforeseen processes of many kinds. The array principle is now being increasingly used in technical systems, specifically where sensors and information-processing systems, given the proper degree of miniaturization and economy, can be integrated to form "intelligent" systems. This incipient symbiosis between electronic sensor technology and the microcomputer is opening up new dimensions for an innovative potential that has scarcely been tapped before [392].

Sensor Systems in Medical Technology [482]. Sensor systems used in medical diagnosis and monitoring must meet different requirements.

The physician, in examining a patient, used to be dependent on what the patient and his relatives said, the physician's own experience, and above all his senses. Today, medical technology, developing in parallel with general technology, offers valuable help to the physician. The range of his sensory perceptions, in particular, has been hugely expanded by technical resources, sensors (i.e., technical sensory organs) playing an essential role. Milestones in this development include the discovery of X-rays by Roentgen in 1896 and the recording of a electrocardiogram by Einthoven in 1902. Procedures based on these techniques remain the cornerstones of clinical and practical diagnosis today.

The principal objective in the use of medical instruments is to support the physician in administering the optimal treatment to the patient. Important narrower objectives in diagnosis and monitoring are to acquire reliable diagnostic information, place the lightest possible burden on the patient, and hold the user's costs and efforts within reasonable limits.

Three areas of medical technology are especially dependent on sensors: imaging techniques [483], physiological measurements [484,485],

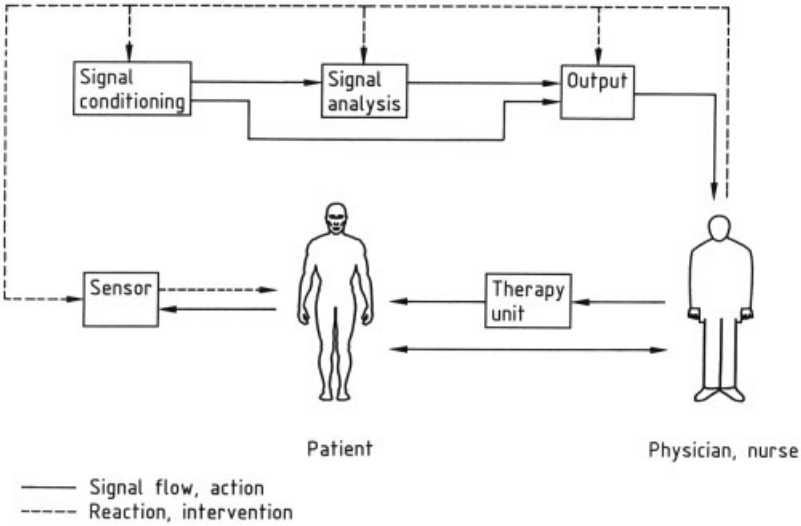

Figure 160. Diagnosis and therapy with medical devices [482]

and laboratory analytical techniques. In imaging and physiological measurement, the sensor is closely linked to the patient (Fig. 160); application aids or transmission media are also employed in many cases.

Only sensors and sensor systems with electrical output signals offer the possibility of instrumental processing, representation, and storage of information. A wide range of methods, drawn from a wide range of sciences, are available for the conditioning and interpretation of these signals; see, for example, the approach to dynamic entropy analysis in [486,487]. Nonelectrical sensors that are still of major practical importance include X-ray film and mercury thermometers and manometers.

The quality of the signals delivered by the sensor has a substantial influence on the value of the technical system as a whole. Often, no additional useful information can be derived from a poor output signal, no matter how extensive the post-processing. This consideration alone means that the effort of developing and maintaining sensors must be continued along with the work of applying them.

Imaging Techniques. Medical imaging techniques can reveal information that is not directly accessible to the human eye, either because of the nature and frequency of the radiation or because of the geometry of the body (Table 4). Except for thermography, these are active procedures; that is, the measurement is based on energy introduced into the body in the form of radiation or radiation-emitting substances. Additional contrast-enhancing substances are often applied as well. Crucial advances with regard to exposure time and image quality in some techniques have been achieved by using arrays (with many detectors working in parallel) instead of discrete detectors. Nuclear medicine employs as many as 60 photomultipliers; X-ray computer tomography (CT), over 2000 detectors; and ultrasonic examination methods, up to 200 piezoelements.

More recently, nuclear magnetic resonance (NMR) tomography has become an established technique. It has the advantage of using strong magnetic and high-frequency fields instead of ionizing radiation, so that the patient avoids the risks associated with radiation exposure.

Significant results can be anticipated from positron emission tomography (PET), which can be regarded as a symbiosis of nuclear medicine and computer tomography. This technique permits not only the three-dimensional determination of blood circulation in the brain, but also quantitative studies of metabolism. Together with single-photon emission computer tomography (SPECT), this procedure raises the hope that diseases in the future may be detected long before anatomical lesions or other symptoms develop [488,489].

Table 4. Sensors for imaging procedures in medical technology [482]

Imaging procedure	Sensor	Radiation
Nuclear medicine	scintillation crystal (NaI) + photomultiplier	γ-rays (radioisotope sources)
Classical X-ray (transmission image)	film + intensifier film ($CaWO_3$)	X-rays
	image intensifier	
	storage film (Ba halides)	
X-ray computerized tomography,	discrete scintillators (CsI)	X-rays
NMR tomography,	gas detectors (Xe)	
PET, SPECT	semiconductors (Ge, CdTe)	
Endoscopy	fiber optics	visible light
Thermography	thermofilm (liquid crystal)	infrared light (passive)
	bolometer	
	IR vidicon	
Nuclear spin tomography	HF antenna	high-frequency pulses, magnetic fields
Ultrasound:	piezoceramic	acoustic waves (MHz)
B display (realtime, compound)	piezofilm	
Transmission camera		
Transmission CT		
Doppler		

The images produced by nonelectrical sensor systems, such as X-ray film (with intensifying film) or fiber-optic endoscopy, can be electronically scanned for further processing.

Ultrasonic imaging presents a difficulty not encountered with other imaging techniques: the coupling of the sensor to the body. The contact must be made through a couplant having acoustic properties (impedance) similar to those of water or tissue; serious disturbances result from air inclusions.

Physiological Measurements. The task of physiological metrology is to determine time-dependent states and functions of individual organs and organ systems in the body.

In contrast to imaging techniques, the procedures in physiological metrology are mostly passive; that is, the signals acquired are chiefly mechanical and electrical signals inherent in the body. Quantitative results are needed in most cases. As a consequence, instruments used for temperature and pressure measurement in a medical setting must be calibrated against standards.

Parallels to process sensor system technology, with regard to nearly all application-oriented generic attributes (see Fig. 154), abound in this field.

In Table 5, important and characteristic applications are listed by groups of organ and the quantities measured. In addition, the relevant sensors and sensor principles are indicated. Further, the table indicates in what setting the techniques are used: D stands for clinical and practical diagnosis, M for monitoring (chiefly intensive monitoring), L for long-term diagnosis or monitoring under everyday conditions [490]. If a procedure can also be invasive, that is, if the sensor or parts of the sensor system are introduced into the body for the duration of the measurement or the body must be opened for the measurement, this is indicated by the letter I. Often information can be obtained only by invasive measurement; the decision is then based not only on the balance of benefits and risks for the patient, but also the cost of the sensor, which is influenced by more stringent requirements on quality and miniaturization.

In general, the following points are taken into consideration in the development and selection of sensors:

- Essential precondition: a satisfactory functional principle
- Reliable application with the least possible expenditure of time and materials for routine uses
- Compliance with national and international standards of technical safety
- Compatibility of materials, especially for internal and long-term use
- Resistance to a wide range of mechanical loads and chemical attack (contact agents, body substances, cleaning, disinfection, sterilization)
- Transportability and adequate shelf life under variable ambient conditions
- Economic mass production with reproducible properties

Robust sensors of simple design, which require no involved preparation or post-treatment, are advantageous; the increasing adoption of one-time articles is a direct consequence.

• Electrodes for Surface Potentials: One of the most frequently performed physiological measurements is studying the electrical activity of

Table 5. Physiological quantities and sensors listed by organ system [482]

Quantity	Sensor/ principle	Heart, circulatory	Respiratory	Gastro-intestinal, urogenital	Brain, nervous system, sense organs	Skin, connective
Force, acceleration	piezoresistive, piezoelectric, optical	D M ballistic cardiogram	D M respiratory mechanics, monitoring of apnea		D tremor	D M L motoric, weight
Pressure		D M L I blood pressure	D M respiratory pressure, resistance	D M I bladder pressure, uterine contraction pressure	D M I intracranial pressure, intraocular pressure	
Flow rate	optical, thermal, electromagnetic, ultrasound	D M I blood flow, pulse, stroke volume	D M respiratory flow, volume	D I urine flow, volume		
Sound	microphone	D heart sounds, Korotkoff's sounds	D auscultation	D intestinal sounds, fetal heartbeat	D hearing	
Temperature	thermal (NTC)	D M L integral quantity: body-core, skin temperature				
Electric potential, impedance	electrodes	D M L I EKG, rheogram	D M rheogram	D M fetal EKG, EMG	D M I EEG	D EMG
Magnetic properties	SQUID	D myocardiogram	D silicosis	D iron in the liver	D MEG	
Optical properties	photodetector			D amniotic fluid	D eye	D
Gases	electrochemical (polarography, potentiometry), spectroscopic	D M I O_2, CO_2, pH oximetry	D M O_2, CO_2	D I pH		D M I O_2, CO_2; tissue, transcu-taneous
Ions				D dialysis		D I
Organic substances		D I glucose enzymes	D alcohol acetone	D L		D I

* D = diagnostic use; M = monitoring; L = long-term diagnosis or monitoring; I = invasive.

organs by measuring electric potential differences at the surface of the body. The sensor system used to measure these low voltages (EKG amplitude ca. 1 mV; EEG amplitude ca. 50 μV) includes the electrodes and the electrolyte (each electrode plus electrolyte constitutes a half-cell) together with attachment agents, power cords, and connectors.

The criterion for choosing the electrodes and electrolyte is a well-defined, stable electrode potential with a low noise level. The time required to reach equilibrium after mechanical or electrical disturbances should be in the range of seconds, and the effect of imposed d.c. or a.c. currents should be slight (time-varying overvoltages, corrosion).

Electrolytes and adhesives should be skin-compatible; the electrolyte, as contact agent, must also lower the skin's electrical impedance at the point of the measurement.

For applications in which the required EKG quality is not too high, robust metal electrodes are frequently used. Where the requirements are more stringent or the conditions are more difficult (stress and long-term EKG, intensive care), the Ag – AgCl electrode is the only type employed. It is also the preferred electrode in ready-to-use one-time electrodes precharged with electrolyte.

• Sensors in Intensive Care: The principal intensive care sensor applications are in surgery and neurosurgery, cardiology, internal medicine, obstetrics/gynecology, and pediatrics. Typically the patient is connected to various diagnostic and therapeutic systems (e.g., infusion, respiration, dialysis) and also continuously attended. The sensors or probes attached to the body must not interfere with care and therapeutic measures and, conversely, must be affected as little as possible by these measures; sensors, instruments, and accessories must be as simple and reliable as possible.

A variety of designs are available for EKG, blood pressure, and temperature measurement in all fields:

- Ag – AgCl adhesive electrodes for adults and children, as well as for operations
- Blood pressure measurement by intermittent noninvasive means (cuff) or by continuous invasive means (catheter) with external pressure transducer
- Temperature sensors for body core or skin temperature

In addition, specific procedures and transducers are used in certain special areas, for example:

- Photoelectric monitoring of the peripheral pulse
- Monitoring of respiration in adults (thermistor) and in infants (rheographic, noncontact methods)
- Cerebral pressure
- Fetal EKG and contractions in obstetrics
- Monitoring of oxygen supply: CO_2 in respired air (by infrared absorption); transcutaneous measurement of O_2 and CO_2, chiefly in newborns

Sensors and their application must meet higher standards when used with infants, since all the bodily dimensions are much smaller than in adults and both risks and interactions must be more strictly avoided.

Invasive Measurement of Chemical Quantities: Except for CO_2 measurement in the respiration gas and transcutaneous measurement, the continuous measurement of substances involved in metabolism and other substances in the body has not yet become routine in intensive care medicine. Development of sensors for these purposes, especially in combination with catheters or for use as implants, has been going on for a long time; techniques known from the analytical laboratory, such as spectroscopy, polarography, and potentiometry, are in the foreground.

The use of specific membranes, electrolytes, and enzymes makes it possible to modify electrochemical sensors (oxygen and pH electrodes) for the measurement of other substances, such as glucose, CO_2, or ions.

The miniaturization of sensors and the mechanical and electrical connection of sensors to a catheter are tasks that have already been accomplished or are at least feasible in principle. Applications involving contact with blood have still not been satisfactorily solved, because in its properties and its interaction with materials, blood is probably more complex than any other fluid encountered in technology. The deposition of fibrin and other proteins in the sensor pocket has two consequences: an increased risk of thrombosis for the patient, and disruption of the primary contact surface for the measurement.

The routine adoption of new catheter sensors is also difficult because the station personnel must possess additional knowledge and capabilities, for example, in the operation and calibration of the instruments and in the assessment of the measurements.

Requirements as to the reliability and accuracy of catheter sensors are far surpassed by the requirements imposed on implantable sensors and sensors for controlled therapeutic procedures. Therapeutic systems controlled with chemical sensors are a topic of major interest for the future.

Promising applications for intensive care are respiration, dialysis, and balancing, for example the regulation of the water, electrolyte, and en-

ergy economy of the burn victims. Implantable devices for the delivery of medication, such as insulin delivery controlled by a glucose sensor, are a key area for long-term therapy.

The transition from the laboratory analyzer, still routinely used for this purpose, to catheter sensors has not yet been completed, even though the use of integrated circuits has already led to major advances in the direction of miniaturized and multiple sensors. The next step – to implantable sensors with long-term stability – may well be equally large. A contribution toward solving this problem may be anticipated from biosensors, in which integrated electronic structures are built and operated in direct contact with organic materials.

Sensor applications range from imaging systems to the invasive measurement of metabolic variables, with cardiological diagnostics and intensive care as points of emphasis. Metrological methods and application techniques merit special attention, as does the management of the patient – sensor interface.

A comprehensive literature survey can be found in [481]. Information on research projects in the methods and applications of microsystems technology can be found in [491], which contains a detailed bibliography, and [492].

Sensor Systems in Automobiles [493]. Automobiles today must satisfy increasingly stringent requirements for pollution control, economy, safety, and comfort. These standards can be met only with the help of electronics. Sensors are the essential link between physical, chemical, and technical processes in the automobile on the one hand and electronics on the other.

Sensors in an automobile must offer high reliability and accuracy in a harsh environment, together with small size and low cost. Costs for transmitters as used in process technology, where quantities are correspondingly small, range from a few hundred to a few thousand €(Euro); in contrast, the maximum long-term cost for many types of automobile sensors must be < € 5.

Classification of Sensors in Automobiles. The sensors needed in automobiles can be classified by use, characteristic curve, output signal type, and interpretation.

By Use:

- Functional sensors in drive-train control systems and in subsystems including the steering, braking, and body electronics.
- Sensors for monitoring and diagnostics, for example in the cooling, braking, lighting, fuel, and other systems.
- Sensors for safety and security purposes, for example in passenger protection, locking, and anti-theft systems.
- Sensors for information, for example in the measurement of fuel consumption and outside temperature, or in travel route monitoring [494].

By Characteristic Curve: Which of the following characteristics is preferred will depend on the sensor application:

- Continuous, linear: Preferred for systems that perform continuous measurements with a wide control or display range. Uniform sensitivity is desired.
- Continuous, nonlinear: Preferred for systems that perform continuous measurements with a narrow control or display range. High sensitivity over a limited range is desired.
- Discontinuous, limiting value: Used for state-dependent "program variation" in complex systems, for example to enrich the mixture when the throttle valve is full open and to monitor the critical limits in the automobile.
- Discontinuous, stepped: Suitable for monitoring components subject to wear where replacement must be planned ahead of time, such as brake lining replacement before a vacation trip.

By Output Signal and Interpretation:

- Voltage or current analog
- Amplitude analog
- Frequency analog
- Period analog
- Pulse-length analog
- Digital

Voltage and amplitude analog signals must be submitted to analog/digital (A/D) conversion before they can be processed in digital systems. Recently, however, inexpensive A/D converter circuits have appeared on the market in integrated form or partially integrated on a chip with the microprocessor. Integrated commuting circuits, known as multiplexers, make it possible to use these converters in a quasi-simultaneous manner for several pickups.

Time-varying signals such as frequency, period, and pulse-length analog signals have the advantage that they are not very sensitive to noise on transmission paths and they are easily digitized at the point of reception. The prerequisite, however, is that an appropriate converter circuit is located at the transmitter, because measured variables are not generally in this signal form. What is more, clocks and counters must be provided at the point where the information is to be processed.

Sensors with digital output signals have not been considered because of costs. The downward trend in price for integrated logic circuits, in particular microprocessors, has led to reconsideration of such new sensor configurations, in which microprocessors are included in the sensors. The resulting "smart sensors" offer marked advantages with regard to compensation of undesired variables as well as equalization; in some cases, A/D conversion can be carried out by the sensor itself. Code security procedures can insure reliable signal transmission, even in a noisy environment.

Requirements on Sensors

– Environmental Protection: Sensors used in automobiles operate in a hostile environment, being exposed to heat, cold, moisture, dirt, water spray, salt, electrical interference, mechanical vibrations, and mechanical shock. These factors make the measurement tasks immensely more difficult and represent a serious threat to sensor reliability and service life (Table 6).
  The most severe conditions are experienced by sensors mounted directly on the engine. A temperature swing from $-40\,^{\circ}$C to $+150\,^{\circ}$C and even around $800\,^{\circ}$C in the exhaust gas stream, as well as $100\,g$ accelerations, are common environmental conditions here, but components installed in the engine compartment also experience harsh conditions (Table 7).
  Whenever possible, sensors are therefore removed from the zones of most severe stress and installed at more favorable locations. For example, pressure sensors used to obtain the motor load signal are usually connected to the intake pipe by means of a hose.
– Reliability and Service Life: Reliability requirements for sensors on board an automobile are in all cases extremely high, but can be ranked as follows: safety functions, mobility functions, diagnostic and monitoring functions, and information. In addition to the needed reliability, there is often a further requirement that a sensor malfunction must also be detected so that the subsystem can switch over to emergency mode and permit at least a limited "limp home" function. Frequently this point has to be considered early in the sensor development process.
  As to service life, sensors must also fit in with the trend toward longer-lasting automobiles. Expected values are more than ten years or > 3000 h service.
– Accuracy: Although accuracy requirements in many instances are as high as those in

Table 6. Environmental influences on sensors and typical effects [493]

Environmental influence	Typical effects
Temperature/temperature change	aging, embrittlement, fatigue, cracking
Atmospheric humidity (with dew and fog)	stray currents, surface and contact corrosion
Precipitation (rain, snow, ice, frost)	electrolysis, swelling
Pressure (air, water)	function
Sand, dust	stray currents
Sunlight	ageing, cracking
Aggressive substances (NaCl, SO_2, H_2S)	surface and contact corrosion
Mildew	stray currents
Salt water and sea mist (road deicing salt, coastal areas)	electrolysis, contact and surface corrosion
Sulfur dioxide (industrial regions)	surface corrosion
Ammonia, perspiration	creep due to stray currents, loss on insulating resistance
Fuel and fuel vapors	surface corrosion, swelling
Brake fluid	surface corrosion, swelling
Battery acid	surface and contact corrosion
Motor oil	swelling

Table 7. Service conditions of sensors installed in the engine compartment [493]

Mechanical	
Vibrational acceleration	(potholed road)
	chiefly stochastic
	chiefly low-frequency (max. 10 – 30 Hz)
	effective (time-average) value: 5 – 25 m/s^2
	max. peak acceleration: 50 – 80 m/s^2
Climatic	
Temperature	normal 120°C; extreme 140°C
Max. humidity	38°C/95 % R.H.
Salt fog	24 – 96 h (DIN 50 021), depending on packing density
Water	splash to flooding
Dust and sand	density up to max. 5 g/m^3
Chemical	gasoline/oil vapors, alkaline cleansers
Biological	mildews and molds

process plants, this accuracy has to be attained at much lower cost. This can only be achieved by mass production in a highly automated manufacturing process. Because the required accuracy is generally the main factor governing selection of sensor principle as well as cost, it is important to take great care in determining the acceptable error. Excessive requirements have a superproportional influence on cost. Accuracies of 2 – 3 % are needed for many applications.

– Space Requirement and Weight: The trend toward smaller and lighter automobiles, dictated chiefly by fuel economy, means that components have to be made small; accordingly, there is often little space available for the installation of sensors. Lightweight, space-saving sensor designs are therefore preferred. Small sensors, furthermore, have the advantage that they experience smaller accelerations due to vibration and impact and thus are less prone to malfunction.

Quantities Measured in an Automobile. Despite the wide range of electrical and electronic systems used in automobiles, the number of physical quantities measured by sensors is limited. The most frequently used sensors are those measuring distance, angle, rotation speed, pressure, and temperature.

Less common but still important are sensors for flow rate, acceleration, force and torque, partial pressure, and humidity.

Even though the list of quantities is not long, a considerable variety of sensor types are employed because of the diversity of media (streams) measured, measurement ranges, required accuracy, and environmental standards. Two examples are given below:

– Engine Electronics for Gasoline Engines: Fuel injection and ignition must be controlled jointly and matched to each other. In other words, every operating condition must be identified and serve as input to the process of fuel metering and ignition timing. This results in following improvements: increased fuel economy, increased engine power per unit of displacement, and reduction of pollutant levels in the exhaust.
  The motor-electronics system needs sensors for engine speed, crankshaft position, air flow rate, air and engine temperatures, and intake pipe pressure (load signal). To insure compliance with stringent exhaust quality regulations, the system is augmented by a λ control with a λ probe. Other possible functions of the electronics include automatic transmission control, idle speed control, knock prevention, and exhaust recycling.
– Electronic Control of Diesel Engines: Diesel engines have traditionally featured mechanical control, but stricter air pollution and noise emission standards as well as demand have led to greater use of electronics. Figure 161 shows the process properties involved in optimizing a diesel engine along with the required sensors. Apart from the parameters listed above (e.g., rotation speed, temperature, pressure, and flow rate), quantities such as the beginning of injection at the nozzle and accelerator pedal position are measured.
  The example of the sensor detecting the start of injection illustrates how harsh some of the sensor service conditions are. This pickup is directly integrated into the nozzle holder and is thus in the immediate vicinity of the combustion chamber.

Measuring Principles of Automotive Sensors. The technical implementation of sensors for automotive service can be based on a wide range of physical effects (Table 8).

Table 8. Technologies for automotive sensors [493]

Quantity measured	Sensor principle and sensors
Distance/angle	short-circuit ring or disk potentiometer
Rotation speed	galvanomagnetic effects (Hall, magnetoresistive)
Pressure	piezoresistive semiconductor pressure pickups
	diaphragm unit with Hall, thin-film strain gage, or capacitive pickoff
	thick-film pressure pickup
Temperature	thin-film metallic sensors
	semiconductor (Si) sensors
Flow rate	heated wire or film
	propeller (fuel)
	circulating-ball sensor
Acceleration	spring/mass system with strain-gage (cast film, thin-film) pickoff
	piezoelectric sensor
	piezoresistive semiconductor sensor
Torque	magnetoelastic sensor (torductor)
	eddy-current sensor
	optical sensor
Oxygen concentration	O_2 concentration probe with ZrO_2 solid electrolyte

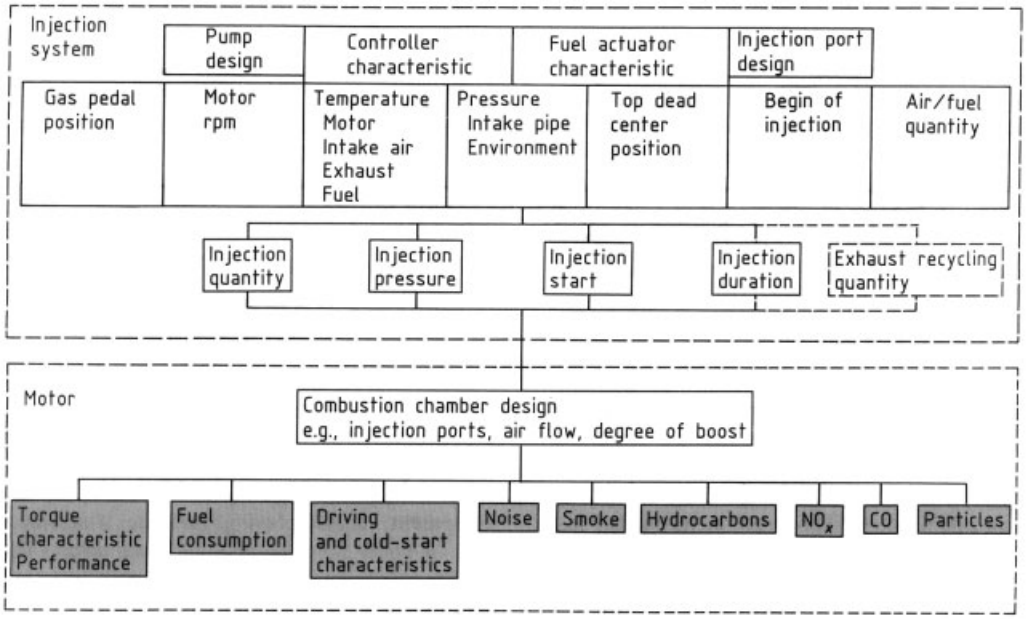

Figure 161. Process properties for optimization of a diesel engine [493]

Coating techniques are used in temperature sensors (e.g., Ni, Pt), strain gages (e.g., Ta), and rotation speed sensors (e.g., Permalloy).

The constant-resistance heated-wire sensor is used to measure air flow rates. The requisite long-term stability is imparted to the platinum wire by special surface treatment plus regular "refining."

Torque sensors would be useful especially in generating a load signal for electronically controlled automatic transmissions. The torque on the main shaft might be measured, for example. Because the shaft cannot be split, its elastic deformation can be exploited for this purpose. The eddy-current damping principle has proved suitable here. Two metal cylinders with lengthwise slots are slipped one over the other in the axial direction, with a clearance of about 10 cm, and rigidly attached to the shaft. When a torque is applied to the shaft, the cylinders rotate relative to one another. A high-frequency coil placed axially over the cylinder housings experiences unequal degrees of damping as a result.

Continuous acceleration sensors (accelerometers) are used for prompt collision detection so that passive passenger-restraint systems such as an airbag or belt tightener can be triggered. Spring-and-mass systems are a proven type of accelerometer; they are mechanically damped by being operated in silicone oil. A triangular leaf spring, preloaded in one direction, bears a mass at its free end. The inertia of the mass causes flexural loading of the spring in a roughly uniform manner, so that the spring is deflected accordingly. The flexural stress is detected by strain gages, which can also be thin-film devices.

In order to meet exhaust regulations and minimize pollutant emissions, the engine should be supplied with an air/fuel mixture as close as possible to stoichiometric ($\lambda = 1$), so as to insure proper functioning of the downstream catalysts.

Exhaust probes or λ probes permit accurate measurement of the oxygen content and regulation of the λ ratio. These probes have a characteristic curve such that even small departures from $\lambda = 1$ results in a very marked change in the voltage signal. For engines operated over a wider λ range, partial-pressure probes with continuous characteristics are in development.

The situation in the automotive sensor field can be summed up by saying that much has been achieved already but much development is still needed to reach further goals with regard to economy, safety, and exhaust cleanness with the aid of complex system concepts and electronics.

Above all, ways must be found of resolving conflicts between the objectives of performance and cost. One approach involves sensor technologies that offer major economies when put into mass production. The clear favorites at present are semiconductor and thin-film technologies. However, functional packaging, contacts, electronic interpretation, etc. account for a large fraction of sensor costs and largely determine performance and reliability. For these reasons, the industry must not simply further develop but must also invest in the development of these incidental technologies.

Sensor technologies of the same type, strongly oriented to specific requirements, are encountered in rail-mounted vehicles, ships, and aircraft. These application areas cannot be discussed further here.

Sensor Systems in Solid-State Process Engineering [495, 496]. Solid-state processes offer a further example of the relationships between product properties and process properties (as described in Chapter 3 and Section 5.1), especially with regard to quality assurance standards. The properties of disperse solids are influenced by the chemical composition and the physical state of the product (property function). For development, manufacturing, and use, it is necessary to know how the product state affects the quality attributes. Because quality attributes are determined on the end product, the characterization of the disperse state must be utilized in controlling the process. An important principle of quality assurance is that quality must be definable and thus inspectable. Metrology thus occupies a central place in quality assurance. The practices involved in quality assurance are oriented to the following steps, each of which contains specific problems for disperse products:

- All desired quality attributes are set forth in the profile of requirements
- The process must be structured
- Process control variables must be determined; property functions and modeling are particularly important here

Solid-State Processes. A solid-state process is a sequence of process elements put together in various combinations so that a solid with certain properties is created as the end product. A typical process sequence [497, 498] is shown in Figure 162 in terms of the phase model. As shown in Chapter 2, the phase model contains not only process elements but also the respective product states, and thus takes account of the way the product state changes in the course of the process.

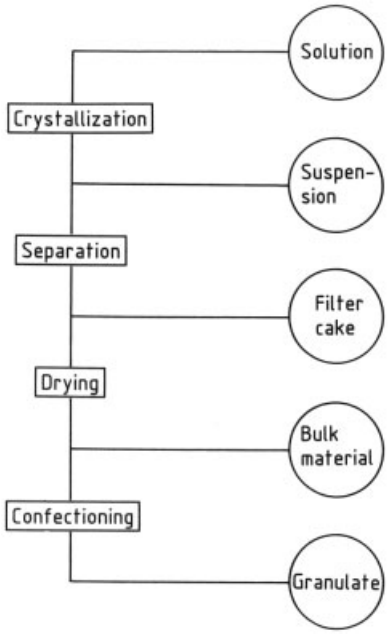

Figure 162. A typical solid-state process: process elements and product states [495]

After a chemical reaction, the solid is formed by crystallization or precipitation. It is separated from the liquid phase in a mechanical separation stage, such as filtration, then washed, and finally dried in a thermal operation. A confectioning step may follow, in which the product is put into the form offered for sale. Confectioning may include the unit operations of size reduction, agglomeration, mixing, or fractionation. The end product must ultimately be conveyed, stockpiled, packaged, or metered. The product state changes after each process element: from solution via suspension and filter cake to powder or granular form (depending on the drying or confectioning step). An important feature of the solid-state process is that the product properties do not depend just on the chemical composition but also depend heavily on the disperse properties of the product.

RUMPF treats this dependence of the technological properties of a product on the physical properties by introducing the property function [499]:

Technological property =	f (physical property)
	$Q_i = f\ (E_{\mathrm{p}\,i})\mathrm{E}_{\mathrm{i\ chem}} = \mathrm{const.}$
Technological properties Q_i	Physical properties $E_{\mathrm{p}\,i}$
Processing $Q_{i\,1}$	Dispersity properties $\mathrm{E}_{\mathrm{p}\,i}$
Filterability	Size
Formability	Shap
Agglomeration behavior	Porosity
Application $\mathrm{Q}_{\mathrm{i}2}$	
Activity	
Action	
Instantaneous property	

Property functions apply to the sequence of process steps (processing properties) and also to the end properties, that is, the quality of the product (application properties). In filtration, for example, the fineness of the product is crucial; in the case of an automobile, the color properties depend in part on how finely divided the pigments are.

The product state is not only crucial to the end quality but must also be considered for each process element in solid-state processes. Not only does the product state change in each process element, the product state itself significantly influences the process element. A knowledge of how the quality attributes depend on the product state is essential for the development, manufacture, and use of products. Structuring of the process into units (unit operations) or meaningful areas is necessary if hold points with intermediate goal values are to be defined. These intermediate targets must be checked with appropriate methods of measurement, as a means of early process monitoring. The state of the end product is not simply established in the last process element; every process element throughout the process contributes to it. For this reason, close tolerances must be maintained even in the intermediate stages if the subsequent stages are to remain in the correctable range. The following steps must be considered from a quality assurance standpoint:

- Establishing the quality levels in the profile of requirements
- Structuring the process and identifying the process control variables
- Insuring the reproducibility of the process
- Optimizing the process

The various types of product quality target values are listed below for the example of pigment manufacturing:

Application properties
- Color strength
- Color coordinates
- Purity

⋮

Processing properties
- Pourability
- Freeness from dust
- Dispersion hardness
- Gloss
- Interfacial properties
- Rheological properties

⋮

Safety
- Low content of side products
- Easy to dispose of
- Nontoxic
- Incapable of dust explosion

⋮

Efforts must be made to achieve a number of quality attributes, and these target values relate not just to the final quality (characterized by the application properties) but also to processing properties (see Fig. 397), environmental aspects, and safety considerations. All these quality attributes must be monitored (i.e., measured). Two examples indicate how difficult this is.

Suppose that a paint, for example an auto body paint, is to be manufactured from a combination of pigments. The first substantial problems arise in the testing of tinting strength, since this measurement cannot be performed on a dry pigment. In practice, an entire paint factory must be "simulated" before this test can be performed; in other words, a procedure similar to that in the paint factory must be followed on a laboratory scale (Fig. 163). The test preparation must be done in the same medium, in the same technical facility, and with the same energy expenditure as would be used by the paint manufacturer; the paint must be applied by the same technique; and the same instrument must be employed for color measurement.

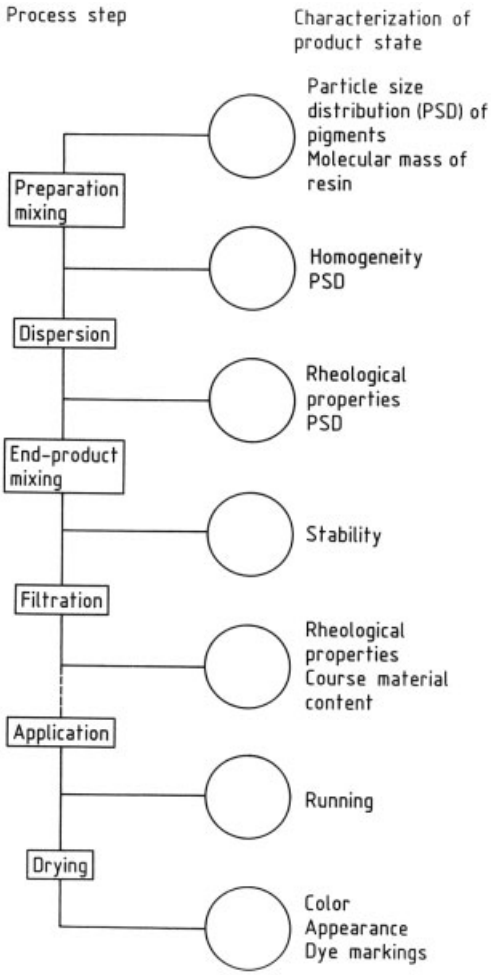

Figure 163. Process for manufacturing a paint, from the pigment to the painted surface [495]

All these requirements for pigment quality inspection cannot, of course, be satisfied when the pigment is to be used in paint for several users. Instead, from the test in one system, in one apparatus, and so forth (representative testing), deductions must be made about the behavior in other types of apparatus. The development of tinting strength in two unlike dispersion machines illustrate the potential problems of such deductions. In Figure 164, the tinting strength is plotted as a function of dispersion time. Unequal energies are applied in different dispersion machines, for example in a laboratory apparatus and the full-scale system. As a rule, tinting strength is determined after one or at most two dispersion times. The result is that the quality attribute "tinting strength" varies widely when the dispersion times and energies cannot be matched, or unambiguously correlated, with those of the user.

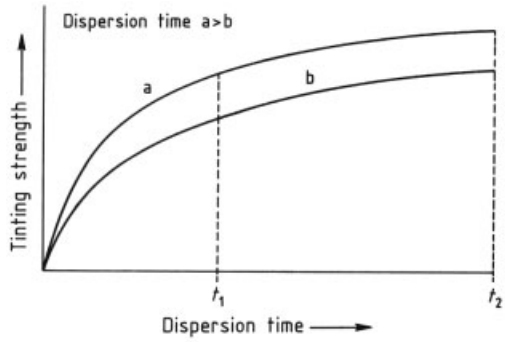

Figure 164. Tinting strength development for two dispersion energies [495]

This example shows what difficulties arise in quality inspection even when there is just one final target value. A process is designed and optimized for the final target values of a product, and the aim of quality assurance is to maintain these final target values. A standardized testing method leaves open questions of product quality for any individual case. Every user, however, wants a suitable pigment for its own formulation, where "formulation" or "recipe" covers both the paint system (including additives) and the apparatus and operating conditions.

Property Function. An important, necessary, but not sufficient step in quality assurance is quality control. As a rule, quality attributes are measured only on the end product. Thus quality can be controlled but not assured. Assurance, in the sense of "controlling," requires some way of influencing the process, and doing so at the earliest possible process stages. This in turn means that the process must be structured into single steps or meaningful subdivisions and that intermediate goals must be defined or established. These intermediate goals are suitable for early process control. The phase model (Fig. 162) and the property functions become especially important for quality assurance. Only when the relationships between quality attributes and physical quantities (property functions) are known is it possible to identify an appropriate process control variable. If the disperse state can be measured and controlled at the point where it is gen-

erated or changed, intervention in the process steps can be implemented at the proper time and thus quality can be controlled. In other words, a knowledge of the property functions is a prerequisite for the assurance of the quality of disperse products.

In the example of colored pigments discussed above, some of the property functions are well known. Figure 165 shows the target value tinting strength as a function of particle size. For two or more distinct target quantities, a compromise is necessary. The desired particle size should already be established at the crystallization stage.

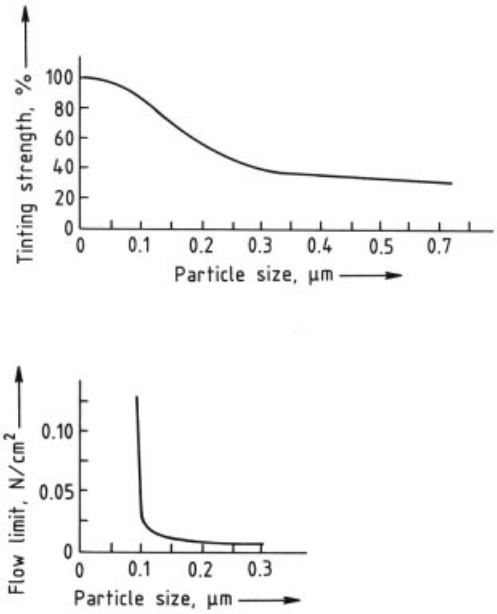

Figure 165. Property functions for a colored pigment [495]

The property functions must be known if crystallization is to be optimized with respect to the application properties. Each subsequent process step changes the particle size and the state of agglomeration [500]; that is, the generation of particles by crystallization should be controlled and changes later in the process should be avoided or monitored.

The dispersity values based on the property functions are important control variables and thus make it possible to implement early process control. To determine the property functions, both quantities must be known and measurable, namely the physical quantity (dispersiness) and the technological quantity (application or processing quantity). The problems that arise in the measurement of the latter are discussed in Section 5.1. The measurement of dispersity is the focus of particle-size measurement techniques. Some aspects of particle-size measurement will be discussed briefly below.

Examples of particle properties with corresponding test methods are listed below:

Particle size (distribution), agglomerate size (distribution)	diverse methods
Specific surface area	BET method
Flow behavior	shear test
Bulk density	determination of volume and mass
Compressibility	determination of volume and mass, force – distance diagram on compression
Redispersibility	transmission measurements PSD measuremens for various degrees of dispersion
Wetting, porosity	liquid absorption, capillary pressure

Generally, the particle size (primary particles up to agglomerates) and the particle size distribution are of fundamental importance for solid-state process engineering. A number of methods are available for their measurement. These have been described in many publications and analyzed with respect to time, costs, and problems [501,502]. The trend is toward measurement techniques and instruments that utilize the optical properties of disperse systems as a tool for characterization.

Instead of going further into individual measurement methods, the problems linked with the use of these methods is discussed here [503]; see Figure 140. As a rule, a sample must be taken from the process. (The special features of sampling of powders and granular materials have been thoroughly treated [504].) The sample must be transported to the measurement location and prepared so that the constraints imposed by the measurement technique can be complied with. Preparation may mean filtering or diluting the sample, cooling or heating it, dispersing it, or adding other substances to it. These measures, which can bring about a change in the product, transform the process state into the measurement state. Moreover, the time between sampling and result is often much too long. A measurement that does not become available until hours or even days after sampling is not useful for process control.

Thus the objective can only be on-line measurement [444] to obtain information on the product or process state quickly. Sample preparation must be automated so that the measurement state can be correlated with the process state in the most reproducible manner possible. Ultimately, however, in-line measurement (see "Generic Class: Application Features," page 135) must be employed to generate the required information about the process state. Two examples of new in-line methods of importance to solid-state processes are presented briefly.

- Laser scanners [505]: The particles are scanned with a laser and a chord – length distribution is obtained. The probe can be immersed in any vessel whose contents in are in sufficient motion ($0.01 \leq v \leq 1$ m/s). The particle-size measurement range is 1 – 1000 µm.
- Photon correlation spectroscopy [506]: In this laser technique, particle mobility is measured on the basis of Brownian motion. The Einstein relation links the stochastic motion with the particle size via the diffusion coefficient. This method is in successful use, for example for process characterization in vitamin production.

A comprehensive exposition of particle and particle-system characterization can be found in [496], which treats this topic within the framework of sensor system technology. Particle measurement, which extends far beyond the boundaries of particle-size analysis, is completely discussed as to status, trend, and demand in [507], which builds on ideas similar to those that underlie quality assurance in solid-state process engineering [508].

Sensor Systems for Environmental Protection. The prevention or reduction of environmental pollution caused by industry, commerce, agriculture, households, etc. requires not only technical measures but often also measurements, whereby process sensor systems (see Section 5.2) are used to provide the required information about the properties of the relevant product or process element, about water and air pollution caused, and also about noise pollution.

To enable quick reaction in the sense of environmental responsibility, and also as mandated by environmental legislation, the required information must be available on-line (see Section 5.2).

For this purpose it is especially important that information about the product properties be obtained in undistorted form. This can be achieved by measurements on the relevant product itself, performed in-line, that is, in the actual product stream. This is not a simple matter with many present-day sensor systems.

The special features of sampling (in comparison with solids sampling as discussed elsewhere in this section), with the aim of insuring spatial and temporal representativeness, are discussed further on.

Figure 166 shows, in highly simplified form, the tasks of process sensor systems in environmental protection. These systems acquire information about the product and process properties, measure workplace pollutant levels for occupational safety and health reasons, and monitor wastewater, solid and gaseous emissions in stack gases, and noise emissions for environmental protection purposes [509].

Flue-Gas Monitoring Sensors. By way of example, some of the requirements for flue-gas monitoring are examined for a coal-fired plant.

• Power-Plant Flue-Gas Monitoring: The German regulations on power station emissions require continuous measurements of the mass concentrations of particulates, CO, halogen compounds, NO_x, and SO_2, and also of the volume concentration of O_2 (Fig. 167). All instruments used must be tested for suitability; proper installation of measuring devices must be verified by services designated by the authorities; calibrations must be carried out every five years (plants with thermal power under 300 MW) or every three years (plants over 300 MW). In addition, unannounced functional tests are performed yearly.

The measured values (instantaneous values) are interpreted by

- Taking half-hour averages
- Taking daily averages
- Converting the averages to reference oxygen contents
- Classification of O_2-referred averages and storage in the form of frequency distributions

Pursuant to the regulation, the emission limits are deemed to be complied with if

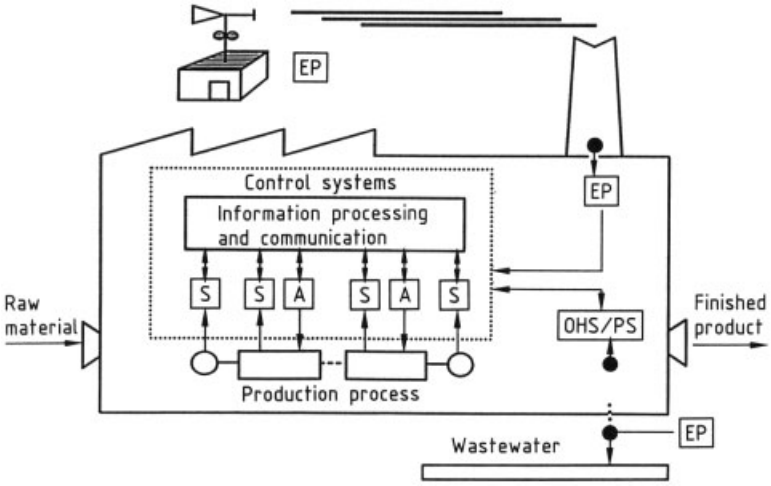

Figure 166. Process sensor systems (S) in environmental protection [509]
A = actuator system; EP = environmental protection; OHS = occupational health and safety; PS = plant safety

- All daily averages are $\leq$1.0 EL (EL = emission limit)
- 97 % of all half-hour averages are $\leq$ 1.2 EL
- 100 % of all half-hour averages are $\leq$ 2.0 EL

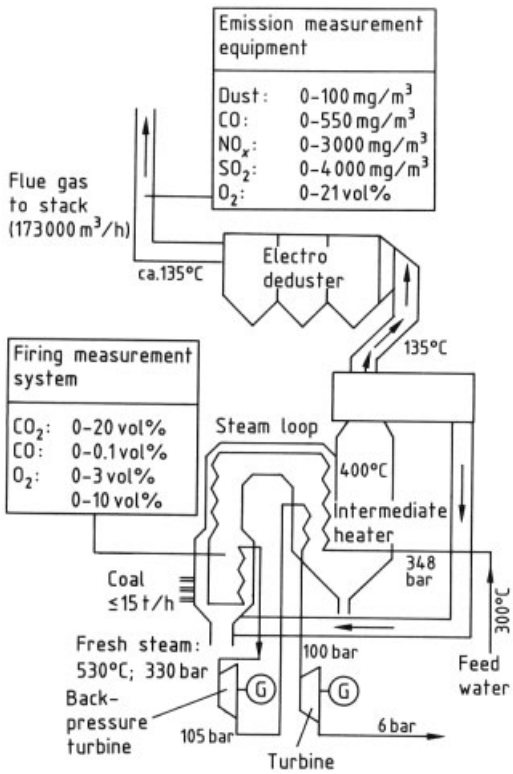

Figure 167. Example of a fired facility [509]

A cross-reference to the concepts of human – process communication (Section 11.2) is relevant here. If compliance with the emission limits is also to be monitored by the operating personnel on behalf of the operator, the levels should be represented by means of suitable presentation and interaction forms. An example of such a form is a cumulative frequency chart updated on-line (see Fig. 168) [510].

Sludge Incinerator: Sludge from biological wastewater treatment is dewatered in filter presses and is then forwarded to a furnace for combustion. Such an incineration system includes a hot ("dry") stack gas scrub, which is subject to HCl and SO_2 controls (Fig. 169).

The emission measurement hardware is designed to comply with TA Luft. Emissions (concentrations) are referred to wet gas in the measurement stream, so that the content of water vapor must be determined each time. This is done directly on the basis of the measured O_2 levels in the wet gas (O_{2W}) and the dry gas (O_{2D}). The measurements are analyzed by averaging, referring to O_2, and classification. Instruments are calibrated by a service holding an official permit.

General Emission and Impact Monitoring of Gaseous Pollutants: Gaseous pollutants represent a special danger from the environmental and occupational safety and health standpoints.

Formerly, this problem could be handled only with expensive suction-type gas collecting sys-

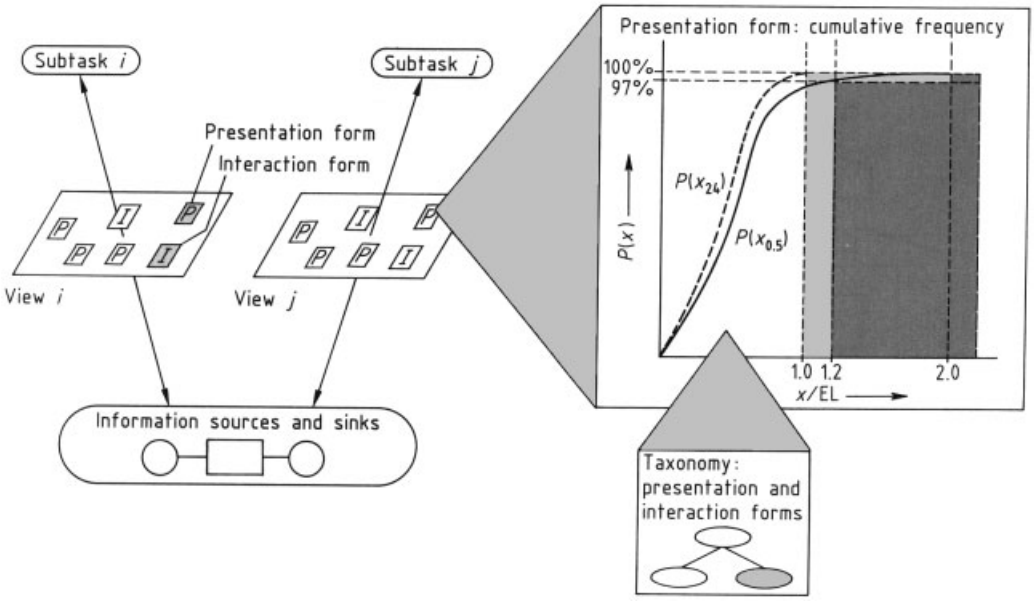

Figure 168. Viewpoints, presentation, and interaction [510]

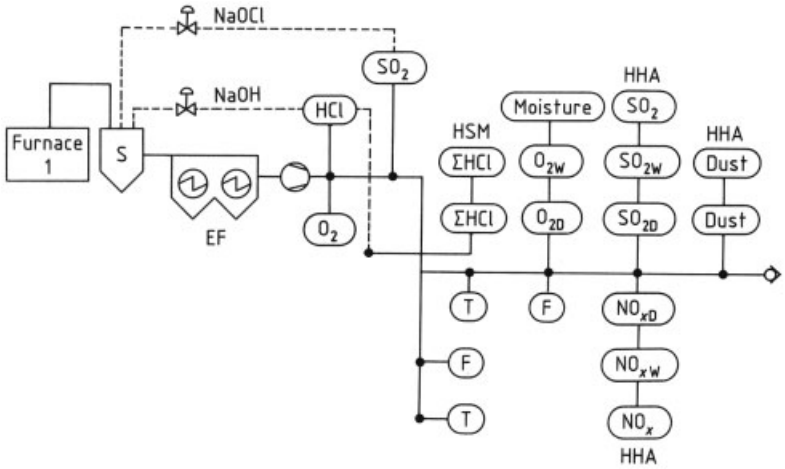

Figure 169. Diagram of analytical instrumentation installed in a sludge incinerator plant [509]
EF, electrostatic filtration; S, stack-gas scrubbing; HHA, half-hour average

tems, which delivered the mixture of pollutants to product-identifying sensor systems. Examples of such systems were spectroscopes, chromatographs, surface-sensitive metal oxides [511] and semiconductors, as well as physicochemical flame ionization detectors [512] and ZGSM gas detectors [513].

As described in Section 5.2 [386, 389], the sensor system proper then includes sample preparation, measurement, and application.

As early as 1985, WARNCKE [382] reported on optically integrating absorption measurements over long lines in the open atmosphere. The measurement of pollutant components in the infrared by this remote analysis technique is best performed with lasers. Rotating sensor systems place a "detection cone" around the object where emissions are expected to occur (Fig. 170).

By way of example, Figure 171 shows emission monitoring at a chlorine plant with an "an-

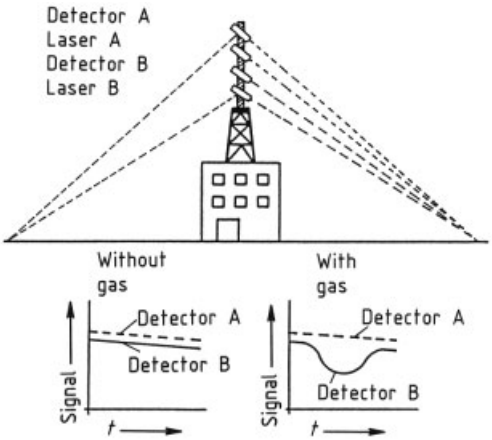

Figure 170. Continuous plant emission monitoring with rotating laser backscatter photometers [382]

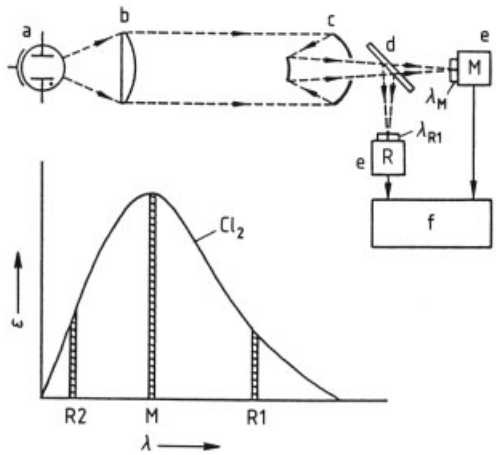

Figure 171. Optical monitoring of chlorine emissions [382] a) Flash lamp; b) Lens; c) Telescope; d) Beam splitter; e) Detector; f) Evaluation

alytical eye," an ultraviolet instrument using a conventional light source. The special innovation here is that the interference due to sunlight is compensated by using a pulsed light source. The source intensity during the pulse is much greater than that of the interfering light. Accordingly, in the detection range the aperture can be made small enough to prevent presaturation of the detector by the interfering light. If necessary, measurements can be performed at two reference frequencies to compensate for light scattering by particles.

This procedure amounts to setting up an optical enclosure around the whole production plant. Despite the use of modern lock-in amplifiers, the cross-sensitivity and relatively high cost of the apparatus has kept these developments from gaining wide adoption.

The current status of this process sensor system technology is surveyed by [514].

Wastewater Monitoring

- Monitoring of Combined Sewer Water: Much of the energy of reaction generated in a process is typically transferred to water (usually withdrawn from a stream) in heat exchangers. Along with storm sewer water, this cooling water is discharged directly into the receiving water through a combined sewer. Contamination can occur when rain falls on roofs, streets, and other areas containing pollutants; in the ordinary drainage lines; or through leaks in the heat exchangers.
- Process Wastewater Monitoring: All other water used in the plant, including scrub liquor from final cleanup stages, must be conveyed through the now-standard system (including mechanical, chemical, and biological treatment processes) in order to comply with legislative requirements.

Without going into detail on national and international laws and regulations governing wastewater, or on the legal consequences of a violation [515] (see Chapter 12), Figure 172 illustrates some aspects of the situation in Germany.

The new philosophy of international environmental protection declares the protection of natural waters from pollution to be a high good, regardless of whether the polluting wastewaters originate in households, public facilities, or industry. Accordingly, the availability of treatment functions in treatment plants is declared to be beyond the scope of economic discussion.

Raab and Wachter-Büchner state in particular [516]: "In order to comply with the legislative mandate, precautions commensurate with specific potential hazards must be taken. These include, first, technical precautions that reliably prevent contamination of the wastewater or the occurrence of severe pollutant burdens in the first place. An effective complement to these primary technical measures is continuous analytical monitoring of wastewater streams for substances foreign to water. In the chemical in-

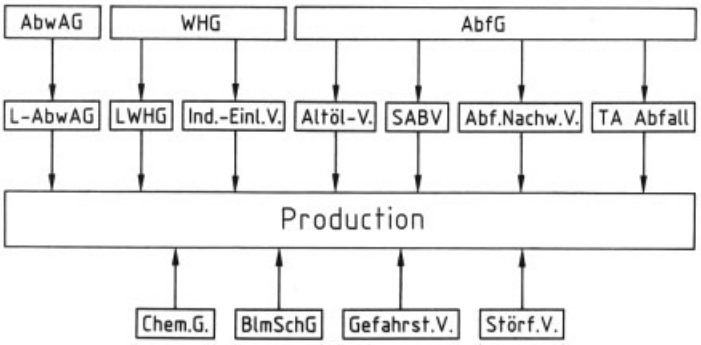

Figure 172. German wastewater legislation and regulations [515]
AbwAG: Wastewater Levy Act; WHG: Water Resources Management Act; AbfG: Waste Disposal Act; L-AbwAG: State Wastewater Levy Legislation; LWHG: State Water Resources Policy Legislation; Ind.-Einl.V.: Regulation on Indirect Dischargers; Altöl-V.: Waste Oil Regulation; SABV: Regulation on the Definition of Hazardous Waste; Abf.Nachw.V.: Waste and Residual Materials Monitoring Regulation; TA Abfall: Technical Directive on the Disposal of Hazardous Waste; Chem.G.: Chemicals Act; BImSchG: Federal Pollution Control Act; Gefahrst.V.: Hazardous Materials Regulation; Störf.V.: Regulation on Disruptions of Operations

dustry, many such substances are carbon compounds, which often means organic contaminants. The parameter TOC (total organic carbon), which states the total content of organic carbon in mg C per liter of water, has proved to be a highly suitable quantity to measure for this purpose.

Water monitoring is not, however, limited to the detection of undesired organic contaminants in wastewaters. Increasingly, it is also employed in process monitoring. Along with environmental protection applications, TOC measurement can also be found in production monitoring and control. A company has control of its production only if it also knows the makeup of its wastewater. A topic of recent discussion is the continuous measurement of TOC burdens in wastewater streams (balancing), which is a prerequisite for source apportionment of wastewater streams within a production complex.

Environmental protection regulations and production monitoring needs dictate the measurement mission. If it is taken seriously, this means the continuous determination of TOC. Only if wastewater streams are subject to continuous, on-line monitoring as near as possible to a potential point of entry of water pollutants can inadvertent contaminants be detected promptly and countermeasures put quickly in train. The monitoring of process wastes also requires continuous surveillance, whether to prevent overloading of the treatment plant or to promptly detect improper states in the plant. There is a definite trend toward process TOC measurement, that is, continuous 24-h determination of TOC in the wastewater. From the metrological point of view, this is not a simple task; from the instrumentation standpoint, it requires substantial expenditure.

According to DIN 38 409 Part 3 [517], TOC is the sum of organically fixed carbon in dissolved and undissolved organic compounds. The DOC (dissolved organic carbon), is the sum of carbon contained in dissolved organic compounds. Other terms in common use are TC and TIC. TC (total carbon) equals the sum of organically and inorganically fixed carbon in dissolved and undissolved compounds. The inorganic component of the TC is called TIC (total inorganic carbon). Each of these is expressed as a concentration in milligrams of carbon per liter (Fig. 173 [516]).

The basis for all methods used to determine organically fixed carbon in water is the oxidation of carbon to carbon dioxide. This can be accomplished in a thermal process (combustion), by wet chemistry (with suitable oxidizing agents), and/or by irradiation with ultraviolet light. The last method is restricted to solids-free samples.

The carbon dioxide resulting from oxidation is either determined directly or reduced with hydrogen and determined indirectly as methane. The determination can be carried out in the liquid or gas phase. Methods suitable for use in the liquid phase are photometry and conductometry. The most common technique, however, is

	mg C per liter in			
	organic compounds		inorganic compounds	
	dissolved	undissolved	dissolved	undissolved
TC (total carbon)	○	○	○	○
TIC (total inorganic carbon)			○	○
TOC (total organic carbon)	○	○		
DOC (dissolved organic carbon)	○			

Figure 173. Definitions [516]

to determine carbon dioxide in the gas phase by IR spectroscopy. For the TOC measurement methods in ordinary use, the individual process steps are summarized in Table 9 [516], together with the starting and end products of each phase transformation, up to the acquisition of the measurement.

Depending on the physical locations of the sensor systems and the way they are tied into the sewerage (process discharge, sewer junction, plant discharge, treatment plant), the requirements on the following performance features vary widely:

- Selectivity
- Detection threshold
- Time behavior
- Availability

Besides the carbon contents TC, TIC, TOC, and DOC, water quality is also determined by other product properties [515]:

Physical and chemical product properties
  Temperature
  Conductivity
  pH, redox potential
  Turbidity
  Contaminants from process
  COD
  pO_2
  Salt
  Ammonium
  Nitrate
  Phosphate
  Heavy metals
Biological parameters
  BOD_5
  Saprobic index
  Toxicity

The methodology of wastewater monitoring is shown in Figure 174 [515].

On-line instrumentation is available for the determination of most product properties. Such

Table 9. Steps in conventional TOC measurement

Process step	Starting/intermediate product	Phase transition	Intermediate/end product
I	organically/inorganically fixed carbon (in water)	separation inorganic/organic species	CO_2 (inorganic) + volatile organic carbon compounds (in gas phase) organic carbon (in liquid phase)
II	organic carbon (in liquid + gas phases)	oxidation thermal chemical photochemical (UV)	$CO_2 + H_2O$ + reaction components (in gas phase)
III	$CO_2 + H_2O$ + reaction components (gas phase)	measurement gas preparation (e.g., drying) absorption of CO_2 (in liquid reagent) reaction (hydrogenation of CO_2)	CO_2 (in gas phase) CO_2 (in liquid phase) CH_4 (in gas phase)
IV Analysis	CO_2 (in gas phase) CO_2 (in liquid phase) CH_4 (in gas phase)	selective IR absorption (photometry), color reaction/conductivity change (photometry/conductometry), combustion (flame-ionization detector)	TOC (mg C/L)

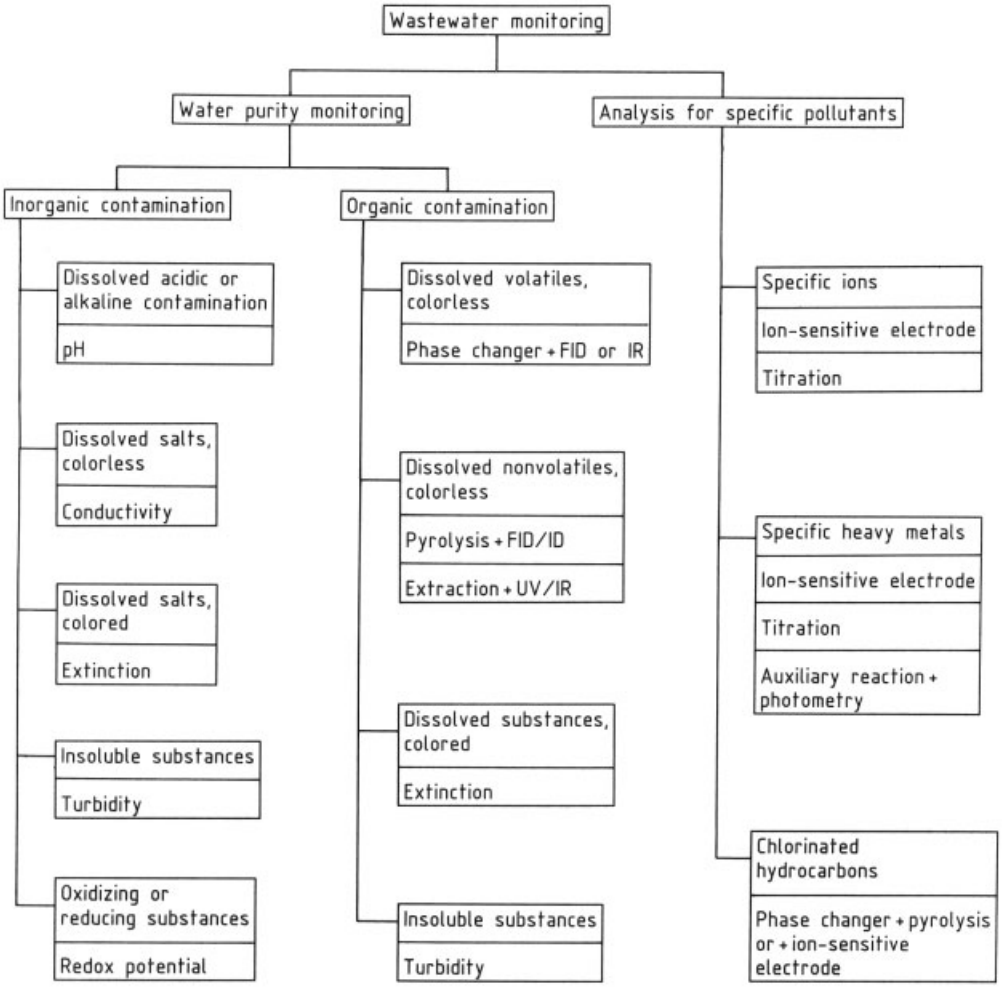

Figure 174. Methodology of wastewater quality monitoring [515]

fully automated devices take water samples, analyze them, and output measurements in computer-readable form in a continuous manner (or quasi-continuously in the case of cyclical analytical methods); see, for example, [518]. The spectrum of on-line measuring instruments that come under consideration ranges from robust devices proven over many years in process monitoring service to complicated and expensive special-purpose apparatus such as TOC analyzers (see above) and toximeters. Similarly broad is the range of performance features and costs for installation and maintenance of the measurement systems.

Representativeness in space and time merits special attention, since the products present in a plant wastewater system generally change from time to time, they are poorly mixed (in laminar flow), and if the pipes do not run full they do not reflect the actual product qualities. Furthermore, it must be ensured that no chemical reactions take place downstream of the sensor system that might lead to other pollutants that then remain undetected. The use of analytical instruments for wastewater analysis must always be examined from the standpoint of explosion hazard.

If there is no usable foreknowledge of the potential pollutants, measurement techniques of universal applicability must be employed; examples are TOC determination (see above) and flame – ionization detectors with phase changers. The apparatus is more expensive and cannot

always be installed in an explosion-proof manner; the selection of physical location may also be severely limited.

If instrumentation is present at the process discharge, expansion is possible by relocation aimed at bringing internal substreams under continuous monitoring. In this way the source of wastewater upsets may be localized [509].

Sensor Systems for Noise Pollution Control. The well-known instrumental methods of noise measurement cannot be treated in detail here, but for the sake of completeness a few remarks are offered from an applications standpoint.

The metrological tasks of noise pollution control in the environmental protection context include the following:

Noise level measurements in the vicinity for the determination and monitoring of noise contributed by the plant and by the surroundings, in particular by traffic.

Emission measurements to determine the level of noise radiated from individual pieces of equipment (acoustic power) and prepare an emission register (acoustic power levels of individual sources).

Calculation of impacts from emissions (noise radiation), with allowance for propagation damping between source and test point, to prepare an impact register (acoustic pressure level in the vicinity).

Comparison of measured and calculated impacts (levels) to validate the level mesurement technique or the propagation damping law used in the calculation [509].

In the framework of the occupational safety and health effort, these tasks are extended to the workplace proper. The large volume of interpretations require sensor and computer systems that are matched from an information-technology standpoint.

Sensor Systems in Polymer Production. In principle, the same problems arise in establishing the product properties throughout polymer production as are discussed in Section 5.1 [406], in Section 11.3 (process analysis and optimization aspects), and specifically in this section [495, 507].

Like all production processes, polymerization can be represented in terms of the phase model (see Chapter 2 and Fig. 175).

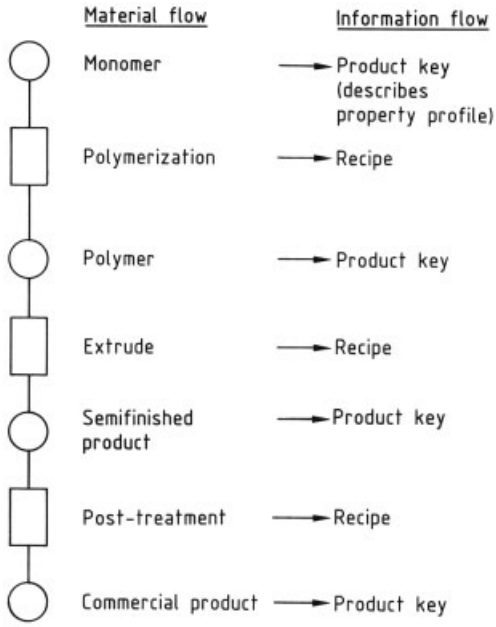

Figure 175. Material and information flows for production of a chemical material [391]

Determination of Relevant Product Properties throughout the Process. The first central problem in determining the product properties is that the properties established or required in the end product (elements of the requirement profile) cannot generally be represented after all the prior steps in the process (see Fig. 35); for example, even a change in state of aggregation makes it necessary to examine more than just one category of these product properties.

Determination of Relationships between Process and Product Properties. The second group of problems arises because the relationships between the properties required in the end product or technological application (requirement profile), the actual product properties (qualification profile), and the properties that exist in the prior stages of the production process are not directly known. Rumpf first discussed this question from the standpoint of experience with solid-state processes (see above), coining the term "property function" for these relations [499]. Later, Ecker et al. applied terms such as "process indicators" to similar solution approaches [406]. Here, as discussed in Section 5.1, the relationship between "technological" [386, 391, 406] and "physical" product properties was formulated. In acry-

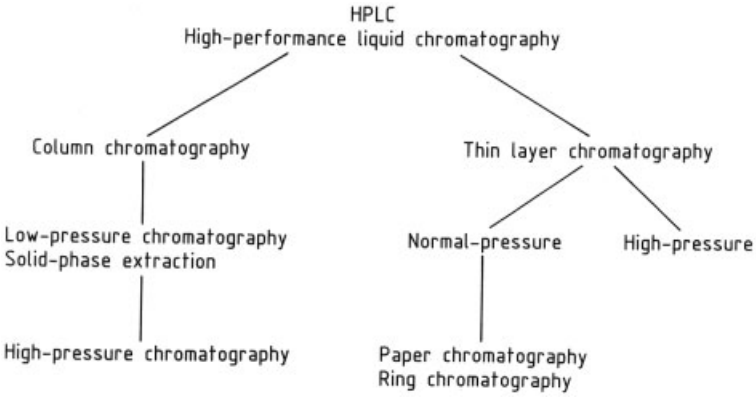

Figure 176. Scheme of HPLC methods [523]

lonitrile polymerization, for example, a viscosity quantity is often used as a control variable (K value) [519–521].

Relaxation Time of Product Properties. The third set of problems is due to the fact that important product properties cannot be determined online (see Eq. 5.2-1) because the mea-surement process has a relaxation time τ_{Rsens} (required for information acquisition) that is too long or fails to satisfy Equation 5.2-1.

In-Line Measurements. Another important aspect is that information acquisition, in its sensor-linked stage (see Fig. 145), often involves not the product itself (in-line, see Section 5.2) but rather an extract, usually prepared after being obtained by appropriate extraction of the product from the production events (ex-line).

Hotop and König have readdressed this problem with application to polymer processes [522].

Complex Shear Modulus. The product properties (chiefly the mechanical-technological ones) can be described by the complex shear modulus G^* (see Section 5.1). This "phenomenological" quantity, which statically and dynamically describes the reaction of a product when a mechanical field acts on it, affords a key with which the mechanical behavior can be traced back to elementary molecular motions, provided there is an adequate theoretical framework. Kuhn [410] and Pechold [411] carried out useful preliminary work, which Morbitzer et al. [412–414] has developed to the stage of practical application.

Molecular-Mass Distribution. The macroscopic mechanical behavior of polymers is essentially governed by the chain-length distribution or molecular-mass distribution and by the degree of cross-linking. A variety of HPLC methods are used here (Fig. 176 [523]).

Only when, as discussed above, there is a stable relationship between the molecular properties of the "process product" (in-line) and extracts rendered "measurable" through a series of differential-solubility operations (ex-line) can the conclusion stated above be justified (see also [524]).

In any case, a specific relation and a specific model must be sought for each application, so that process control consistent with quality assurance can be implemented. From a modern information-oriented standpoint, all such efforts lead to measurement techniques associated with process indicators [406] or supported by models [432]. A comprehensive bibliography can be found in [390].

Purity Guarantee. Similar aspects must be considered for the second collection of tasks, the guarantee of chemical purity for intermediate and end products.

It is beyond the scope and aim of this article to describe material-identifying spectroscopic and chromatographic methods (see [434] for more details).

An illustration of these points is shown in Table 10 [522].

Sensor Systems in Biotechnical Process Engineering. In the preface to [439] Schügerl states that the most important difference between biotechnology and its predecessor, fermentation technology, is that fermentation tech-

nology is an assemblage of hard-to-quantify empirical relations, while biotechnology is an assemblage of relations that are still empirical but can be represented in quantitative form. The foundation has been laid for biotechnology to arrive at a point where it can make quantitative predictions with the aid of general laws. The same holds for both microorganisms and cell lines.

Table 10. On-line measurement tasks in polymer processes [522]

Measurement	On-line analytical methods	
task	Conventional	New
Mixture, purity, and trace analysis	density, refractometry, photometry, titrimetry, GC, photometry after color reaction	spectrometry, liquid chromatography
Additives		
Side components		
Residual monomer		
Polymer		
End groups		
Molecular mass determination	viscosity, light scattering	gel chromatography
Average molecular masses		
Molecular-mass distribution		
Chemical constitution as function of molecular mass		
Characterization of polymer melt/commercial product	rheometry, photometry	spectrometry
Additive content		
Rheological properties		
Branching, cross-linking		
Chemical composition		
Residual solvent content		

Previously, production strains were always selected by tedious screening and then laboriously improved by induced mutation and screening. Today, these strains are increasingly being modified to order by using genetic engineering. Protein design makes possible the purposeful local modification of proteins through appropriate genetic modification of the microorganisms or cells. Metabolite design allows the metabolic pathways detrimental to production to be eliminated and desired pathways to be reinforced.

Analytical techniques play a key role in the industrial exploitation of the full biological potential of the microorganisms and cells obtained in this way. These techniques make possible a precise study of the cell physiology and allow quantitative determination of the regulation of the metabolism. Furthermore, the optimal process conditions for growth and for product formation can be brought about and maintained by means of process analysis.

The same rules apply here as for the other fields discussed in this section, though at a high level of abstraction; see the discussions of solid-state process engineering and polymer production engineering above, as well as Figures 125 and 126.

Sensor systems commonly used in biotechnical process engineering [440] are as follows:

On-line and in situ fluorimetry
- NMR spectroscopy
- ESR spectroscopy
- X-ray structure analysis

Chromatography
Mass spectrometry
Electrophoresis
Optical and acoustic microscopy
Flow cytometry
Flow injection analysis
Bioelectrodes
Biosensors
- Biocalorimeter
- Bio-FETs
- Piezoelectric biosensors
- Optodes

Antibody/antigen systems

Along with the classical methods, which are known at least in general principle in chemical process engineering, a new class of sensors are now appearing: biosensors, which are used especially to mesure specific biological product properties (see Fig. 152). In these sensors, only specific biological reactions yield a sensory effect.

As a rule, however, parallel, serial, or associative linking of information is used [382]. A typical example is given by fermenter analysis (see Fig. 177): The following specific items of information are available: pressure, flow rate, and oxygen content of inlet air; oxygen, carbon dioxide, and nitrogen content of outlet air; pH and turbidity of the solution; and contents of oxygen and carbon dioxide dissolved in the solution. A calculation relating the oxygen and carbon dioxide concentrations to the air through-

put (Fig. 178) gives the respiratory quotient as the ratio of oxygen uptake rate to carbon dioxide evolution rate. Together with a knowledge of the mass-transfer coefficients from the gas phase via the liquid phase to the cell, this makes possible an optimization of the air throughput at a higher level of assessment [382].

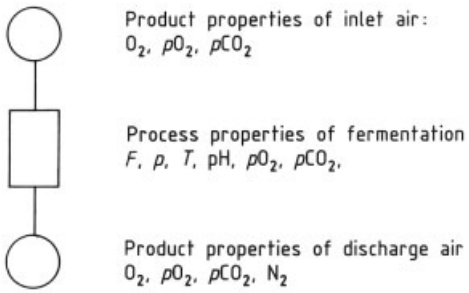

Figure 177. Measurements in fermentation [382]

As shown by Warncke [382] and Hotop and König [522], the greatest difficulty in biological product property determination lies in sampling and sample preparation, which must be done case by case. Without the aid of robotics in sampling and sample preparation and without information-oriented model support, sensor systems in biotechnical process engineering would be inadequate.

Additional constraints include foreign influences and the disposal of sample material after the measurement (see Fig. 144).

Biology itself, however, must increasingly become the guide in the development of biosensors. Multisensors, sensor arrays, and the processing of many simple signals in a manner similar to the neuronal structure of organisms should provide the solution to many problems [392].

Interesting results in the identification of process trends have been obtained by Adler [525], who used fuzzy clustering methods, which Winterstein has tested in practice by applying a multivariate sensor system to wastewater monitoring [526].

Outlook. In order to control a production process, it is necessary to look at information on various levels (Fig. 179). Sensor and actuator systems link the process with the information processing. Process monitoring and control systems take on the function of information processing and human – process communication.

The information orientation of process control engineering means that proven measurement techniques must be developed toward a sensor system technology.

The necessity of providing interpretable information to the process monitoring and control system at any time means a change in the requirement profile for sensor systems of the future:

- Reliability at reasonable maintenance cost
- Functionality even in extraordinary process states (e.g., startup, shutdown, upsets)
- Adequate accuracy and predictive power
- Easy adaptation to production variation [527]

Self-diagnosis and self-maintenance are important qualitites of future sensor systems. The information structure of a universal sensor system is illustrated in Figure 145.

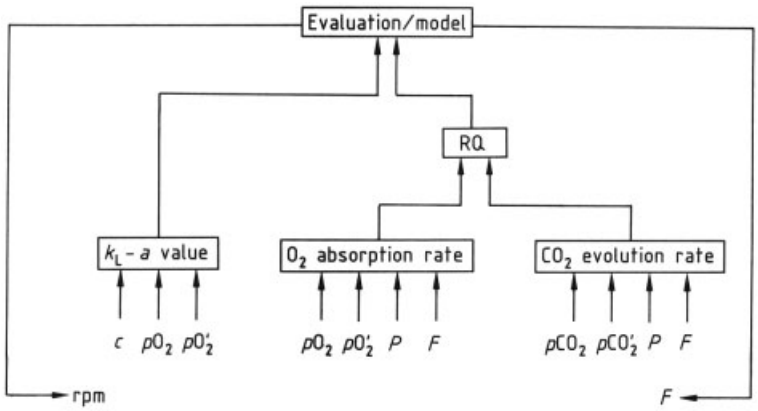

Figure 178. Gain in information due to coupling [382]

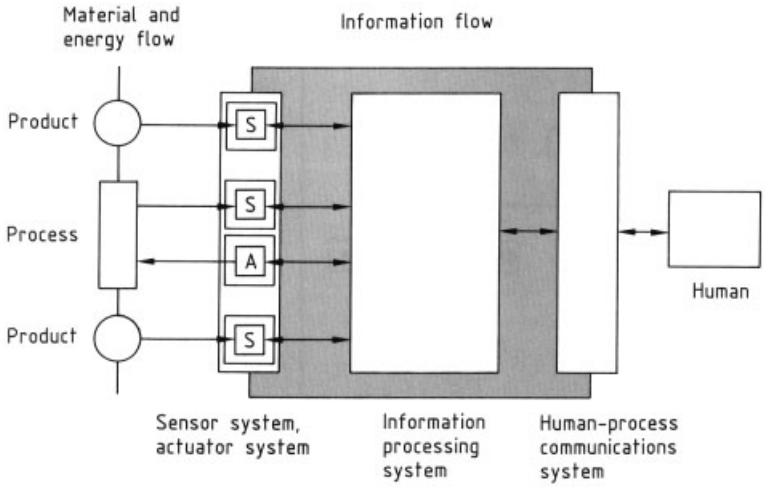

Figure 179. Sensor systems in the information chain from material/energy flow to human – process communication

Increasingly, manufacturers face the challenge of designing modular, engineered sensor systems. Design, together with easy installation and maintenance, represent untapped potentials for improved efficiency.

Of course, there are not yet any adequately tested sensor systems for polymer and biotechnical production, for the solid and suspension states of aggregation, and for many areas of environmental technology. Essentially, however, it is more important to bring the classical (physical, chemical, and technological) measurement techniques up to the new standard.

Hauptmann calls for the following among other points [528]:

"In order to develop sensors and bring them to market at low prices, some fundamental requirements must be met:

- Precise knowledge of the market and of international development projects as a basis for innovation
- Concepts, strategies, and visions of future sensor markets and technologies as a basis for long-term success
- Strategic partnership and cooperation with research institutes, universities, users, and manufacturers"

There is a clear trend away from the laboratory towards the process environment [382, 514].

Already, under the auspices of ISA, IFAC, VDI/VDE-GMA, and other organizations, numerous forums are available worldwide for the discussion of sensors, sensor systems, technologies, and applications.

These discussions, often academically conducted, are brought down to earth by applications-oriented conferences such as GVC, the Eisenhüttentag (Foundry Conference), and NAMUR, which deal with the harsh conditions of everyday service (see above).

With increasing success, attempts are being made to bring structure into this unbounded variety by using new methods of applied informatics [529] and to find routes for effective application through hypermedia (see Fig. 154).

Nevertheless, the need to set down in-depth knowledge in encyclopedias [530] will still exist, so that this knowledge can be kept up-to-date (with a view to the taxonomies cited above) through supplementation or modification of the assignment of object attributes.

Special sensor technologies are continually being reported on by technical and scientific journals such as TM, IFAC, ISA, and atp. These include:

- Optical sensors (including fiber optics)
- Chemical sensors, such as
  - Liquid electrolyte sensors
  - Solid electrolyte sensors
  - Mass-sensitive sensors
  - Optochemical sensors
  - Calorimetric sensors [528]
- Surface-active metal oxide and semiconductor sensors
- Biological sensors

Four general trends can be identified:

Advances in Existing Sensors. The development of existing sensors and the creation of sensor systems continues. Multisensor systems based on commercial sensors, coupled with new signal recognition and signal processing techniques, will open up new application areas. An important aspect is the modeling of the sensor with regard to the effect measured as well as parasitic effects. Pattern recognition in forms ranging from classical to fuzzy logic will promote this development.

Digital Sensors. The resonance sensor is an important step toward digital sensors. This class includes micromechanical resonant sensors as well as sensors based on piezoelectric materials. Analog sensors can be replaced by resonant detection mechanisms with no change in sensor configuration. The coupling of resonant structures with fiber-optic excitation, detection, and signal transmission will increase. The development of entirely novel sensor principles will largely remain an exception.

New or Modified Sensor Materials. Material development is currently a major theme in sensor development. An important role is played by polymers, ceramics, composites, alloys, biomaterials, modified glass-fiber materials, III – V semiconductors, and so forth. Silicon, however, will remain dominant and combinations of silicon with other materials will be increasingly used.

Microsystems. Microsystems technology is having a lasting effect on the field of sensors. A microsystem is a spatial integration of sensor, signal processing, data processing, and actuator. This technology combines methods from microelectronics, micromechanics, microoptics, and information technology. It even makes use of techniques from acoustics, hydraulics, chemistry, and biology. It includes the design, fabrication, and application of technical systems that integrate a variety of functions. The functional units of a microsystem interact closely with one another, and their geometrical dimensions, dictated by functional considerations, are typically in the range from millimeters down to micrometers. An essential criterion for the design of microsystems is optimization of the overall function. This is advantageous from the standpoint of technological effort, because it is not always necessary to optimize the subfunctions per se. But because the interaction between the parameters has the nature of a function itself, the function of a microsystem is more than the sum of the functions of the subcomponents (see also [491, 531]).

The principle that functionality can be enormously increased by a greater integration density while system reliability and economy are maintained provides the basis and motivation for future microsystems technology. The integration of electrical and nonelectrical subfunctions on a chip or substrate enhances the performance of microsystems and is therefore particularly interesting. The requirements that must be met when integrating sensor and electronics, as well as the actuator whenever possible, are, however, in part contradictory.

The manufacture of future microsystems will require not only suitable principles but also high-performance materials, technologies, and design systems. New assembly and connection techniques will play a crucial role, as will quality assurance. Present-day solutions, technological approaches, and principles of microelectronics, materials research, integrated optics, process engineering, production engineering, chemistry, biology, and medicine must all be examined with regard to their systems capability and integratability [532].

The development of sensor systems will, however, be influenced by two other trends:

Transition from Batch to Continuous Processes. As production is changed over from batch to continuous processes, information acquisition cycle times become shorter. Direct coupling or integration of self-contained, autonomous sensor-actuator systems (e.g., for metering pumps, variable-speed and variable-torque motors, and so on) will probably be a necessity (see Chapter 6).

Automation. The problems cited in the preceding sections show how important robot-aided sensor systems will be for sampling, sample preparation, disposal of auxiliary substances, and maintenance actions. For example, Figure 180 shows a fully automated pH sensor system with remotely calibrated probe [438]. Figure 181 shows the sequence of automatic functions in a phase model.

Information utilization leads development in a completely different direction.

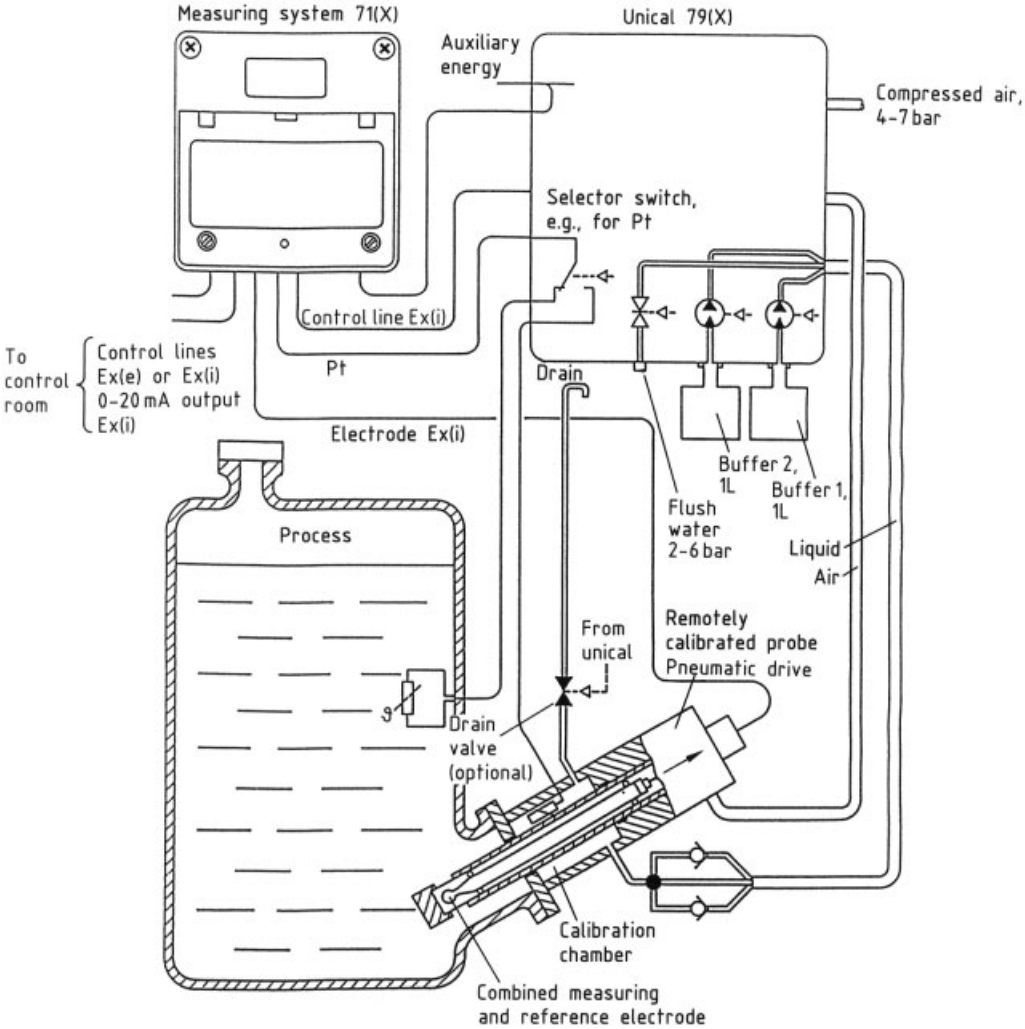

Figure 180. Example of an automated pH sensor system with remotely calibrated probe [438]

"Model-supported measurement methods," which have now become well established, should be mentioned first [390, 424].

The connection between process model and sensor system implies the ways and means for remedying fundamental shortcomings in measured-value acquisition with the present capabilities of information processing. Given these procedures, unmeasurable process and product properties can be reconstructed from easily measurable quantities through interpretation of a priori knowledge, represented in a mathematical model, about what happens in the process.

Furthermore, experience in petroleum exploration with properly positioned sensors has led to the use of tomographic methods so that, ultimately, spatial representativeness can be achieved on the basis of a one-point (local) measurement.

Modern methods of human – process communication (see Section 11.2) will make it easier to develop temporal single-point measurements by using characteristic curves in product and process configuration spaces (Fig. 182).

Sensor and actuator systems, as the link between material flow, energy flow, and information flow, will be designed in future such that they relieve the load on the central processing units of the process control system or actually become subsystems of it.

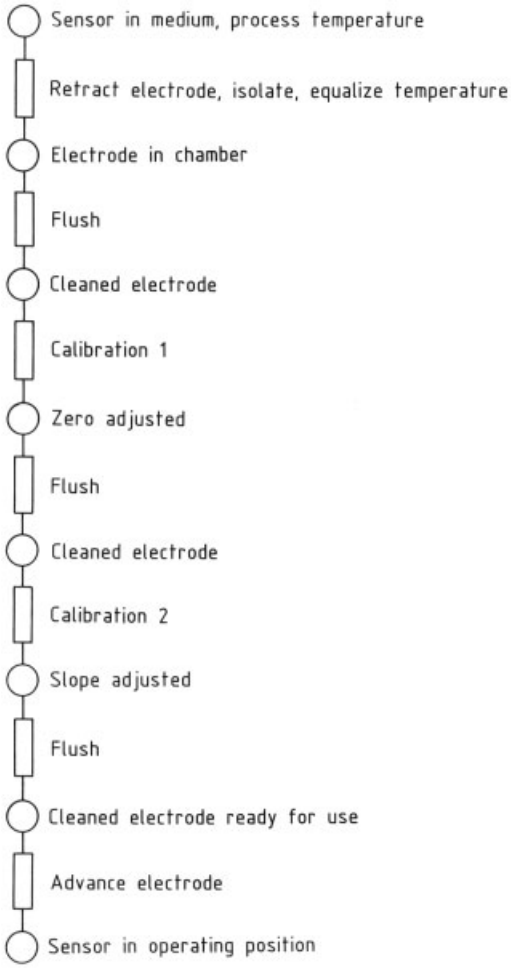

Figure 181. Subprocesses in calibration of a pH sensor

Developers and users, particularly in the area of sensor – actuator systems, will need to grow beyond the historic borders of their field and apply a keen and critical sense of analogy to the problem of transferring insights from other experiential worlds into process control engineering.

5.4. The Market for Sensors and Sensor Systems

Extensive market surveys [533, 534], shows and congresses [532, 535, 536] and seminars held by technical societies and manufacturers all come to one conclusion: The already large sensor market (world volume in 1988: ca. € 12×10^9) wil expand even further, reaching ca. € 23×10^9 in 1995 and > € 33 $\times10^9$ by the year 2000 [537]. Figure 183 shows a regional breakdown of the expected boom.

According to the Prognos figures, process engineering and plant construction are the largest customers of sensor and senosr system manufacturers at present (€ 4.6×10^9 in 1988). It will be these users that largely motivate and support the market growth; the absolute increase in sales to these customers will be exceeded only by applications in automaking, the relative increase by applications in building and safety engineering as well as in medical technology (Fig. 184).

If the forecasts are accurate, process engineering and plant construction will account for roughly one-third of sensor sales in 1995 (€ 7.5×10^9). This development will be motivated by requirements for increased product quality and plant availability, and also by more stringent environmental and occupational safety standards.

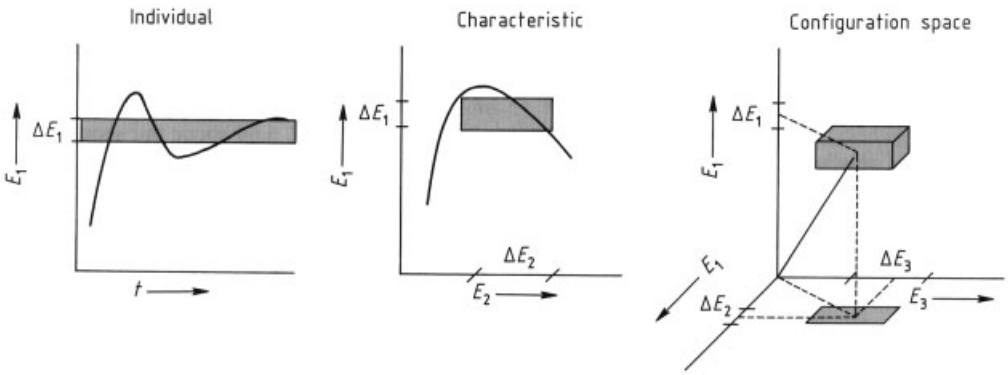

Figure 182. From process signal to configuration space (process and product property space)

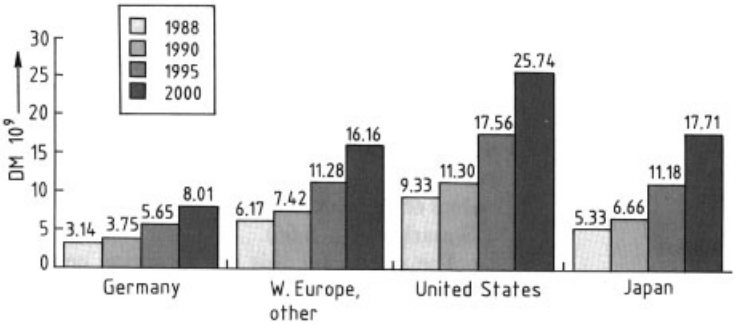

Figure 183. Sensor market broken down by region and year up to 2000 (1 DM ≙ 1.95583 €) [537]

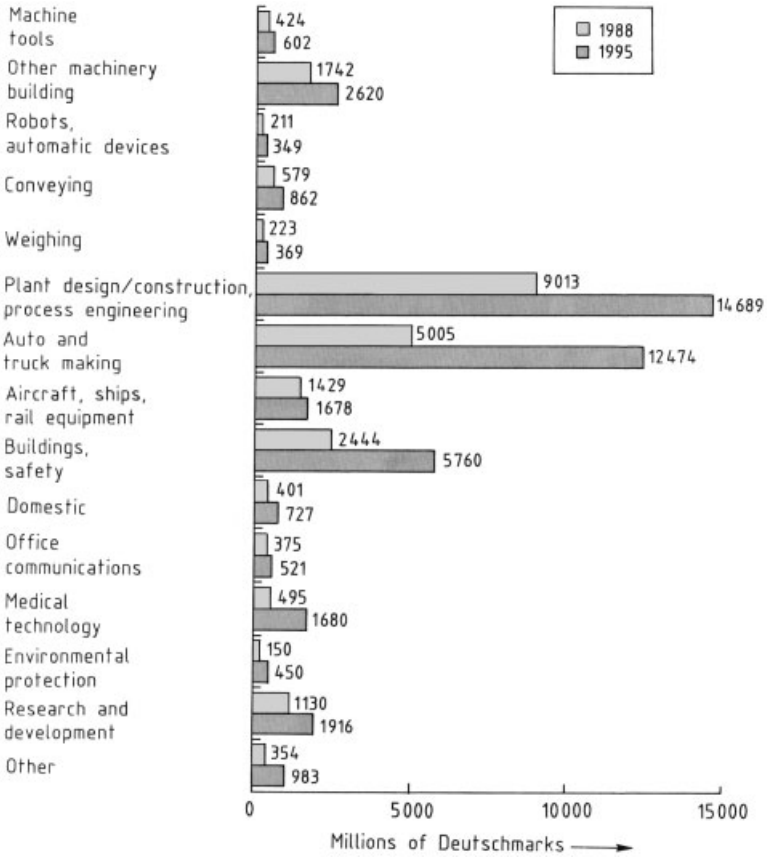

Figure 184. World sensor use broken down by application (1 DM ≙ 1.95583 €) [537]

The Prognos study cited above also shows that plant and process engineering are the dominant users of sensor technology on a region-by-region basis too (Western Europe, United States, Japan) and will probably remain so.

For the sensors used in plant design and construction, Figure 185 presents a summmary of application areas, while Figure 186 gives a similar breakdown for those used by process plant operators. Notable is the dominance of the petrochemical industry, which is accounted for by the high degree of automation in plants now onstream.

Figures 187 and 188 show the sensor market for plant construction and process engineering divided according to the physical quantities

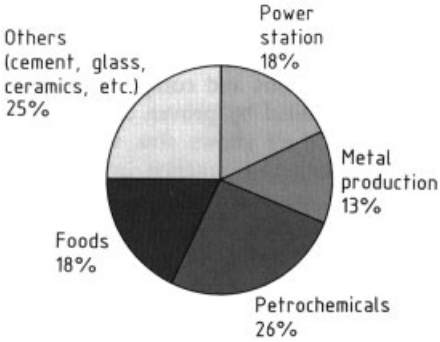

Figure 185. Use of sensor systems in plant design and construction, broken down by industry [537]

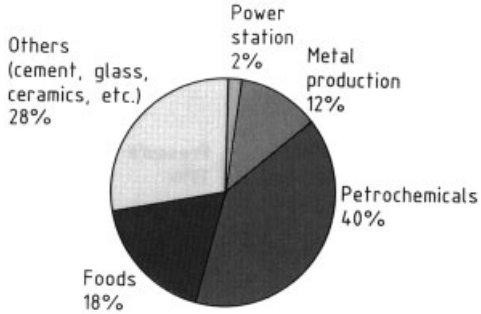

Figure 186. Use of sensor systems in process engineering, broken down by industry [537]

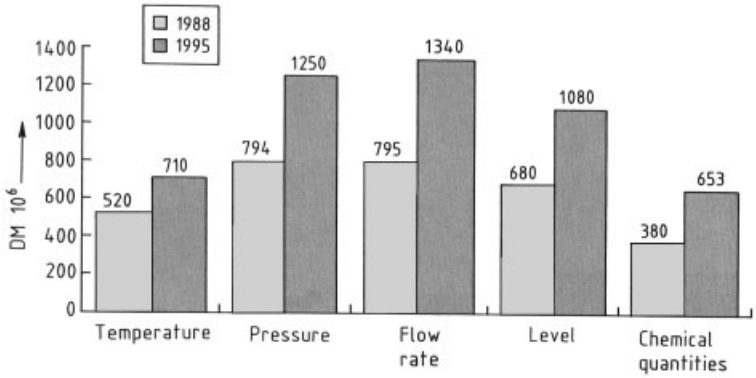

Figure 187. Western European sensor market in plant and process engineering, broken down by quantity measured [537]

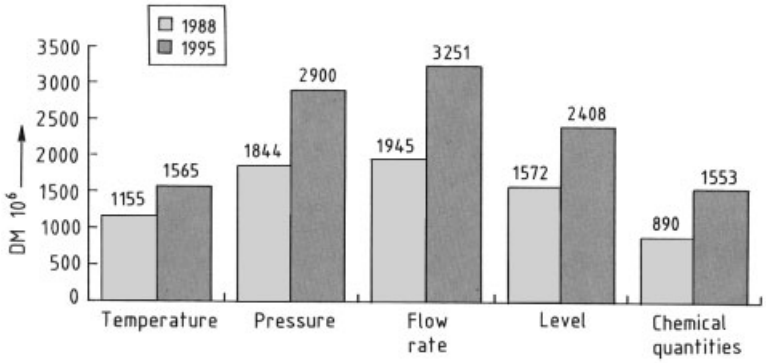

Figure 188. World sensor market in plant and process engineering, broken down by quantity measured (1 DM $\triangleq$ 1.95583 €) [537]

measured. Clearly the process properties identified as "classical" in Section 5.2 – temperature, pressure, flow rate, and level – account for the highest expenditures, while the properties grouped under "chemical quantities" represent a less important market for the process analysis technology discussed in Section 5.2.

It is interesting to do a cost analysis by the quantities themselves. A comparison of Figures 188 and 189, the latter showing the distribution of process control stations by quantity measured in medium-sized chemical process plants [448], suggests that temperature and level can be measured much more cheaply than pressure and flow rate. This is supported by a market sur-

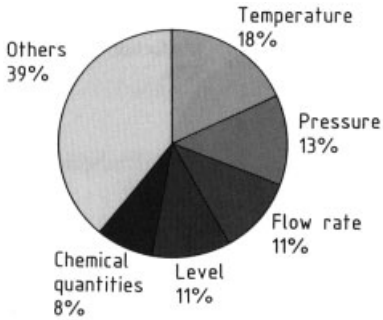

Figure 189. Number of process monitoring and control stations in chemical process plants, broken down by quantity measured [448]

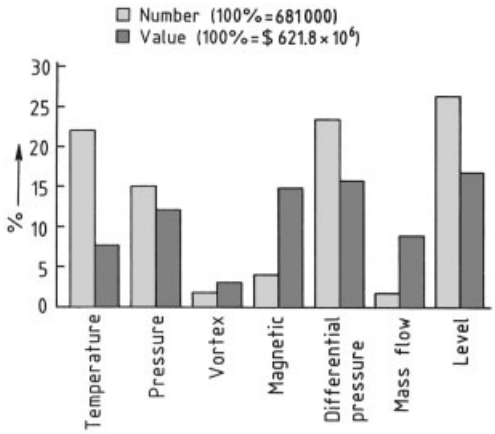

Figure 190. Numbers and values of transmitters sold in the U.S. market, 1990 [538]

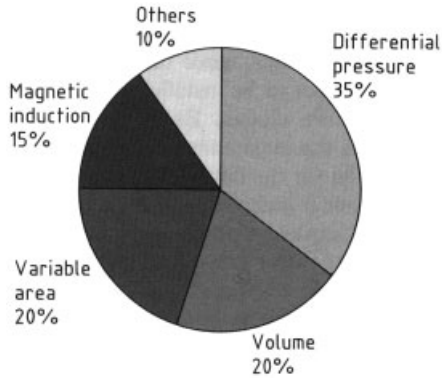

Figure 191. Flow rate and mass flow rate sensors broken down by type (estimated 1988 [537])

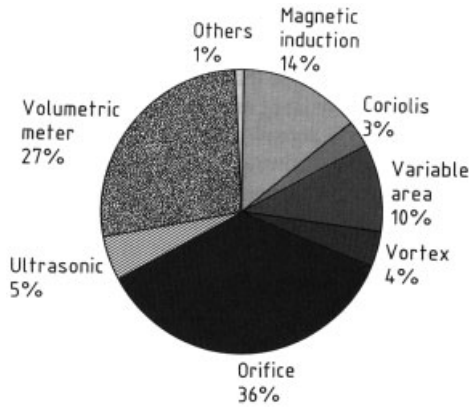

Figure 192. Western European sales of flow sensors in 1990, broken down by sensor type [441]

vey [538] that compares the numbers and values of standard sensor transmitters sold in the US market (Fig. 190).

The reason for this disproportion certainly lies in the use of many relatively simple and robust principles for measuring temperature (for example, with the Pt 100 resistance sensor) and level; with regard to the markedly higher figures for pressure and flow, a detailed comparison with other conventional "classical" technology (e.g., orifice plates) makes it clear that manufacture prices here are motivated by market policies. An intimate knowledge of the market reveals that modern "contactless" methods of temperature and flow measurement, which require costly signal processing arrangements, are being offered at dumping prices in comparison with conventional "mechanical" methods.

A more detailed analysis can be done for sensor systems used in flow rate measurement. According to present figures, these represent cost-intensive process control stations (see also Figs. 188, 189, and 190, in which the third to sixth transmitter categories can be employed in flow measurement).

Newer developments in this field make clear the advantages of intelligent field systems with remote diagnostic and maintenance capabilities. The additional functionalities of these systems, which in conventional technology are rather demanding in maintenance terms, should pay off in the long run by lowering maintenance costs and reducing energy losses (e.g., diminished pressure losses). There is an increasingly pronounced trend toward methods such as vortex shedding, magnetic induction, and Coriolis methods (direct determination of mass flow rate)

and away from the older methods of differential pressure and rotameter-type instruments (see Figs. 191 and 192).

Modern trends in flow measurement are reported in [441,539,540].

All forecasts on applications in plant and process engineering agree on the expected development thrusts and market opportunities: biosensors and microsensor technology [531,541].

5.5. Field Installation and Cable Routing

Introduction. Field installation and cable routing denotes the connections and structural work that must be performed to connect the field instrument to the data-processing system in the switch room and thus insure the flow of information between the field level and the process control level.

Modern installation technology is characterized by the use of an individual line from the sensor system to the switch room where the information-processing system (e.g., process control system) is located. Here, "sensor system" means the sensor with sensing element plus the metrology and instrumentation required for data acquisition [383]. It is not important whether the sensor system is made up of one unit (compact device) or several parts (e.g., the sensor plus a 19″ rack-mount transducer in the switch room). In the compact unit, sensing element and transducer are conbined in a single instrument. Compact devices have some major advantages.

As to functionality, there is a clear division between sensor system and data-processing unit. The sensor system generally performs the measurement, while all information processing takes place in the control system or stored-program controllers. The incorporation of microprocessor technology into sensor systems has turned them into "intelligent" systems, even if they show no outward sign of this new capability. The potential of intelligent sensor systems can be realized only through bidirectional communication comprising the transmission of all commands and information. This is also the only way to take advantage of the wide range of possible functional expansions of sensor systems. Field bus technology will not only allow bidirectional communication but also simplify field installation and reduce costs.

The cables needed for field installation are placed along the cable routes provided for them. Cable routing must comply with structural and safety requirements; important points include fire and mechanical protection of the cables. Other practical requirements can be found in the section on cable routing (see page 183).

Field Installation. *Types of Installation.* Even though the prospects for field installations with field bus and field multiplexer continue to be discussed, new systems and components today are generally installed by proven conventional methods. Figure 193 shows five variants of present-day installation practice in Germany [446].

In variants (a), (b), and (c) the transducer is located in the switch room, whereas in (d) and (e) it is in the field. On going from (a) to (e), the amount of functionality incorporated in the field-level device increases:

- Variant (a): The sensor element (e.g., a resistance thermometer) contains no electronics. The transducer is located in the switch room and is connected to the sensor by a two- to four-wire cable. If the cable has three or four conductors, the effect of the line resistance can be compensated. Because there are no electronics in the field and no auxiliary energy is supplied to the sensor, the signals depend only on the sensor and are said to be sensorspecific. The transducer in this variant is normally installed in a 19″ rack.
- Variant (b): The field-level sensor element (e.g., a pressure sensor) has a two-wire power supply cable and uses a second two-wire cable to transmit a signal (0 – 20 mV or 4 – 20 mV) back to the transducer, which is usually located in a 19″ rack. The signals to the transducer are manufacturer-specific, since the sensor signals are altered, but no intelligence resides in the field instrument.
- Variant (c): The sensor element in this variant has electrical preprocessing. In physical terms, the transducer is now present in the field. The manufacturer-specific signal, modulated onto the two-wire cable, is not yet finally processed; for example, linearization and damping have not been performed, and

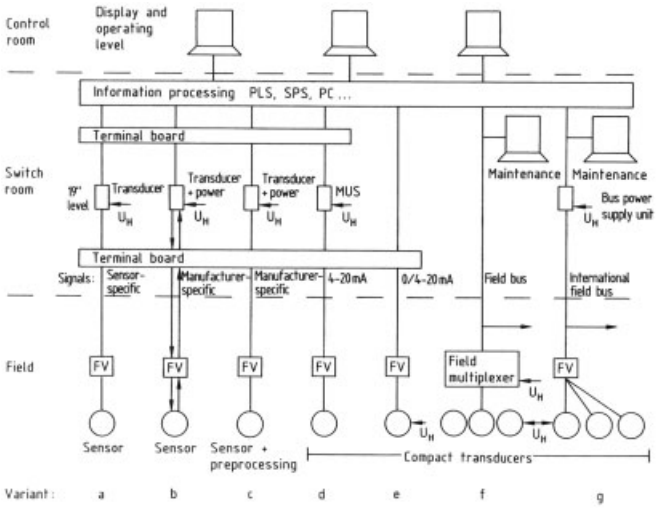

Figure 193. Development of field installation
PLS = Process control system; SPS = Stored-program controller; PC = Personal computer; U_H = Auxiliary energy; MUS = Transducer power supply; FV = Field distributor

the output has not been converted to a 0 – 20 mA or 4 – 20 mA signal. These functions take place in the 19″ rack-mount transducer, whose power supply also supplies the sensor element through the same two-wire cable.

– Variant (d): The transducer is located at the measurement station and is combined with the sensor element to form an overall system (compact instrument). A 19″ rack-mount transducer power supply powers the compact instrument, furnishing 4 mA. The transducer sends the measurement information to the power supply unit as a 0 – 20 mA or 4 – 20 mA signal, over the same two-wire cable. The transducer generates this information by varying its internal resistance.
– Variant (e): This variant also employs a compact device. Because of the physical principle used, instruments in this variant are supplied with a relatively large amount of auxiliary energy, so that a separate power supply line is necessary. The output signal corresponding to the measured value lies in the range 0 – 20 mA or 4 – 20 mA and can be input directly to post-processing or used for displays in either the switch room or the control room. Hence no other instrument is needed in the switch room.

All these variants have in common that the information-bearing signals from the field instruments are led to a field distributor or distributor box, to which other field-level instruments are also connected. The cable employed for connecting the field instrument to the field distributor is generally a twisted pair or four wires in two twisted pairs, with 0.5 mm^2 shielded conductors. Data lines and cables are described in DIN VDE 0800-0899.

Consider, for example, the 32-pole field distributor. From it, a 32-wire master cable usually runs to the terminal board in the switch room. Partial allocation is also possible, provided this is allowed by the allocation of master cable signal loops to the interface plane and no future change in the signal loops is permitted. The master-cable specifications could be as follows: 32-wire ($16 \times 2 \times 0.5\ mm^2$); i.e., 16 twisted pairs, not shielded pairwise but having a common shield.

In addition to the 19″ interface plane, when required, a connection to a second terminal board is also possible. The second terminal board provides the connection to, for example, the process-level components of the process control system. This connection can be made with a master cable having, for example, the following specifications: 32-wire ($16 \times 2 \times 0.5\ mm^2$); i.e., 16 twisted pairs, not shielded pairwise but having a common shield.

Field Installation in Explosion Hazard Areas. Special regulations apply to the installation and operation of equipment in explosion

hazard areas. These requirements have not been harmonized on a worldwide basis, so that some discrepancies arise. The discussion that follows applies to Europe.

- Definitions and Terms (VDE 0165): The basis for assessing the scope of the requirements to be imposed is a division of explosion hazard areas into zones by the probability of occurrence of hazardous atmospheres. The gases and vapors present in the plant are classified into explosion groups, while the instruments and other devices to be installed are divided into temperature classes. Each temperature class specifies the maximum surface temperature permitted for the device.
  - Zones: Zone 0 includes regions in which a hazardous explosive atmosphere is present at all times or over prolonged periods. Zone 1 includes regions in which hazardous explosive atmosphere is expected to form only rarely and for short periods. Zone 2 includes regions in which hazardous explosive atmospheres are expected to occur occasionally.
  - Explosion Groups: Explosive gases and vapors can be classified into explosion groups according to the minimum ignition temperature of their mixtures with air. The groups are designated IIA, IIB, and IIC; gases and vapors in group IIC are most easily ignited. Instruments and devices must be approved for the appropriate explosion group.
  - Temperature Classes: Instruments and devices for installation in explosion hazard areas are classified into temperature classes corresponding to the maximum surface temperature. The classes range from T1 (max. 450°C) to T6 (max. 85°C).
- Cables: Cables and wiring for circuits that are not intrinsically safe must be selected so that they can withstand the anticipated mechanical, chemical, and thermal stresses. The selection must comply with one of the following standards: DIN VDE 0298 Part 1, 3, or DIN VDE 0891 Part 1, 5, 6.
  On mechanical grounds, wire cross sections should not be smaller than certain values (VDE 0165). Conductors should generally be copper or, under certain conditions, aluminum.
- Placement of Cables and Wires: Cable and wiring penetrations to areas free of explosion hazard must be sealed adequately tightly; possible methods include sand seals, approved elastomers, and mortar seals. Cables and wires should have appropriate protection at points where there are especially severe threats due to thermal, mechanical, or chemical attack. One such form of protection is to place the cables and wires inside pipes. This should be done only with approved piping systems; case connections and transitions to cables and wires must also be approved. Once the installation for the explosion hazard area has been completed, it must be inspected by electrical specialists for compliance with ElexV § 12. A further inspection for compliance with the accident prevention regulations of UVV 4 § 5 must also be performed by an appropriately licensed person.
- Equipotential Bonding: In explosion hazard areas, equipotential bonding is required to prevent static discharges that might cause ignition. A good foundation earth is incorporated when the plant is first constructed, and the structural steel of the building and, for example, the switch room floor structure are grounded to this in star fashion. The interface cabinets and terminal board cabinets, as well as all other units, are grounded to this floor structure, along the shortest path. The interface cabinets contain the transducer power suppiles and other devices. Between two active devices, such as motors, located not more than 2.5 m apart, the measured resistance should not be greater than 0.2 Ω. The measured resistance between an active device and a passive object such as piping, located a maximum of 2.5 m apart, must not exceed 1 Ω.

An electrical specialist must verify compliance with the equipotential bonding requirements (UVV 4 § 5) visually and by measurement. These measurements must be recorded in an inspection record.

Systems with Intrinsically Safe Circuits. In the "intrinsic safety" type of explosion protection, electrical energy and surface temperature are limited such that sparking and thermal effects that could lead to ignition of certain atmospheres cannot occur. In other words, not even

a short circuit can cause sparking. This type of protection exists in all countries that accept the IEC standards, and also in the United States and Canada, which have not yet adopted them.

Only insulated wires (insulation test voltage 500 V a.c.) may be used in intrinsically safe circuits. Several intrinsically safe circuits may be combined in one cable. Cables and wires of intrinsically safe circuits must be labeled or color coded. Sheaths or jackets are color-coded in light blue.

Intrinsically safe circuits and circuits without intrinsic safety must not be laid together in cables, wiring, conduits, or conductor bundles. In cable ducts, insulated spacers must be inserted or wires with certain kinds of sheath or jacket must be selected.

When installing an intrinsically safe circuit with just one apparatus, care should be taken that the maximum permissible inductance and capacitance for the circuit, including cables and wiring, are not exceeded. These permissible values are found on the nameplate of the apparatus. Wires must be treated as lumped capacitances.

Intrinsically safe circuits from the field are generally led to interface planes located in the switch room, which provide the separation from circuits without intrinsic safety. Possible implementations include approved safety barriers and transducer power supplies. The transducer power supply limits the maximum power available on the intrinsically safe side and also supplies a maximum of 4 mA and, say, 15 V to the transducer. The sensor system connected to this device thus has ca. 60 mW available to supply all the electronics.

Comparison of Circuits with and without Intrinsic Safety. The preceding section said that a transducer drawing its power via the 0 – 20 mA or 4 – 20 mA signal has ca. 60 mW at its disposal. This implies that only relatively low-power sensor systems can be supplied in this way; however, such sensors make up well over half of all sensor systems. There are, however, other sensor systems that cannot now be operated at such low power levels. The trend is for marginal systems to be supplied with low power at the expense of reliability, accuracy, and/or stability over time. It should be considered whether these drawbacks outweigh the advantages. Such a decision must be made for each application.

The advantages of intrinsically safe sensor systems and their circuits are easy handling and low cable costs, since only two wires per unit are needed. Intrinsically safe instruments and circuits can also be worked on while in service without having to cut off the power.

When nonintrinsically safe apparatus is used, there must be a separate power supply for the sensor system. That is, separate signal and energy channels must be present (see also Sections 5.1 and 5.2), so that at least four wires are required. If the power supply voltage is, for example, 24 V d.c., the supply and signal lines can be put in one cable. This arrangement has the further advantage that the terminal board plane and interface plane can be omitted.

If the two approaches are compared, the intrinsically safe variant has the clear advantage as to handling, while the other offers significant cost benefits. Futhermore, when using intrinsically safe sensor systems that are barely able to function with the power available, some attention must be paid to reliability, accuracy, and time behavior. *Field Installation in the United States.* Installation techniques in the United States are quite different from those in Europe. In most cases, the cables are placed in conduits.

- Definitions and Terms: Classes are defined as follows: Class I relates to gases and vapors; Class II, to dusts; and Class III, to fibers or fiber dusts. Each class is further subdivided into Division 1 and Division 2. The discussion that follows concerns only Class I areas.
  - Division 1 is characterized by the permanent, occasional, or periodic occurrence of hazardous concentrations under normal conditions.
  - Division 2 is characterized by the fact that hazardous concentrations can occur only under accident or malfunction conditions.
- Cabling in Class I, Division 1: The only types of cabling permitted in these areas are cables in conduits and mineral-powder-insulated cables of type MI with the end closures in stress-relieving grips. Only materials approved for these areas may be used; for example, conduits and fittings must satisfy U.S. codes, some of which differ from the DIN standards. Conduits are inserted directly into the pressure-tight housing. The conduit system is completely "flameproof," so that a

spark originating in the conduit cannot escape to the outside.

– Cabling in Class I, Division 2: Apart from conduits, NEC Article 501-4 also permits sealed track channels or cables or wires of types MI, MC, ALS, CS, TC, or SNM with approved end fittings to be installed in stress-relieving grips.
  Normal wiring is also permitted for circuits that, under normal circumstances, do not liberate enough energy on a wire break, short circuit, or ground fault to ignite an explosive atmosphere.
  Housings, fittings, and connectors generally must not be explosion-protected.
– Installation Practice in the United States: The classification of wiring in the United States is similar to variants (d) + (e) in Figure 195, that is, with a field distributor and master cable to the switch room.

Even when cabling without conduit is permitted, common practice is to place all cables in conduits. This has historical reasons: When installation practice was codified in the United States, intrinsic safety did not yet exist, so all cables were run in conduits; in Division 1, the conduits also had to be pressure-tight. Sensor systems, as well as field distributors, were used in pressure-tight designs (as they still are). This type of installation, designed by piping engineers, was reflected in the standards. Changes, even when they appear to make sense, take place very slowly. By now, intrinsic safety does exist in the United States, but its availability has had little influence on installation practice.

The insurance industry actually promotes the retention of this practice. Plants are usually covered by "electrical insurance," written by companies that approve the installation method and generally expect conduits.

Comparison of European and American Field Installation.

Disadvantages of Conduit Use in the United States: The use of conduits increases the space requirement for the installation, since the routing has to allow for bend radii and offsets.
The cost of installation is much higher, not only because of the placement of the conduits themselves but also because of the difficulty of pulling the cables through the conduits, their inflexibility, and the clumsy and quite heavy pressure-tight distributor boxes.
Additional gas barriers must be placed in the conduits to prevent the movement of gases and vapors and the propagation of flames from one part of the electrical system to another.
In the long term, problems arise from condensation and the pitting that it causes, as well as from the higher transfer resistances due to corrosion. This is important because the conduit system is also used for equipotential bonding. Condensation is controlled by drains through which the liquid can be removed.
Advantages of Conduit Use: The conduits afford mechanical protection to the cables.
As electromagnetic compatibility becomes more and more a factor in plant design, conduits offer much better shielding. Especially for digital data transmission, it is very difficult to separate signals from noise, since they occur in part in the same band of frequencies.

Trends in Field Installation. *Cost Analysis for Present and Future Installations.* An important question in the evaluation of new installation techniques is that of cost. New methods must not only provide technical benefits but should also be, at least, no more expensive. One study has found that future installation practice can make a significant contribution to cost reduction [542].

The comparative cost analysis is based on four cabling concepts in a representative plant.

The plant consisted of 400 process control stations. For the first calculation, these were assumed to consist of 240 analog and 160 binary stations. The second calculation was based on 300 binary and 100 analog stations. For the variants with field bus, it was designated that fewer than half the process control stations could be connected to one field bus.

All design, installation, shop, and material costs were included, from the individual line to the field instrument up to and including the input/output cards of the process-level components.

The results are shown in Figure 194.

Variant I: conventional, parallel cabling. Cost = 100 % (basis for comparison).

Variant II: field multiplexer hardware used at the field level, so that only a few cables (field buses) had to be run from the field to the switch room. The result shows that there are marked

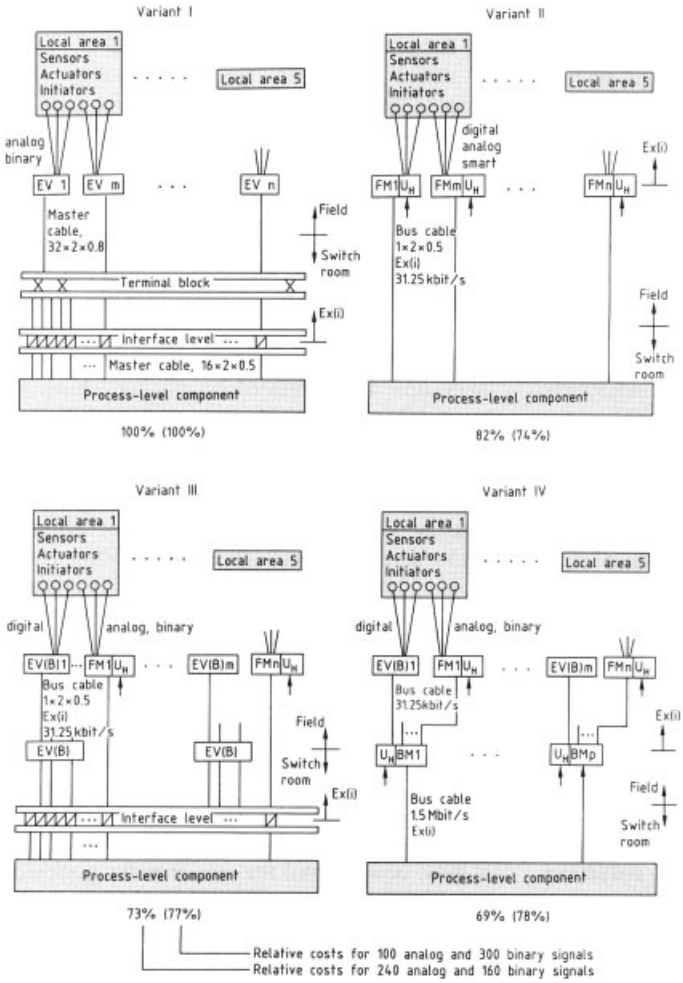

Figure 194. Cost comparison for field installation variants [542]
EV = Final distributor; FM = Field multiplexer; U_H = Auxiliary energy; BM = Bus multiplexer

savings. Cost = 82 % (74 % for the plant with the larger fraction of binary stations).

Variant III: field bus (intrinsically safe; transmission rate 31.25 kbaud) for the field instruments that can be directly connected to a field bus. The remainder of the field instruments are connected to a field multiplexer. Cost = 73 % (77 %).

Variant IV: field buses from variant III compressed into one field bus with higher data rate (not intrinsically safe, 1.5 Mbaud). Cost = 69 % (78 %).

When the analysis of "present-day" instrumentation was extended by distinguishing between separate versions and compact devices, it appeared that future trends in instrumentation would see only the compact instruments still in use, with the intelligent transducer generally forming a single unit with the sensor element. Under severe ambient conditions, it often makes sense to install the transducer with its complement of electronics near the sensor element. Because it is not desirable to connect a 19″ transducer on a field bus, this alternative will largely disappear in future.

Field Installations with a Field Multiplexer. The field multiplexer solution is illustrated in Figure 193, variant (f), and Figure 194. A mixed field installation with field multiplexer and field bus is also possible. The field multiplexer must be approved for service in the explosion hazard area (zone 1). It provides data compression so

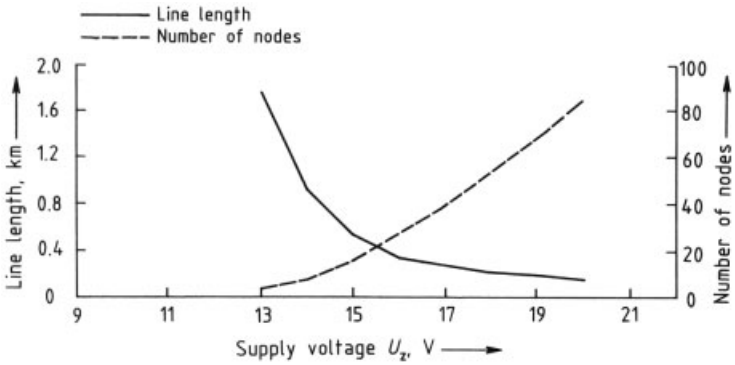

Figure 195. Results of the PTB field bus study (category ib, group IIB)

that the total cost of cabling to the switch room, the required terminal board, the 19″ rack for the transducers and their power supplies, and the necessary plant documentation are all substantially reduced in comparison with conventional installation practice.

Such a field multiplexer can be obtained in a range of sizes. The user must decide the number of channels connected to it above which partial or complete redundancy is to be employed. In the case of complete redundancy, a second field multiplexer is connected in parallel. The line from the field bus directly to the input/output cards of the field-level components in the switch room should not be run together with the existing field bus. The I/O interface module can also be designed with redundancy.

One possibility for the individual lines from field devices to the field multiplexer is an unshielded $2\times0.5\,\text{mm}^2$ twisted pair. The type of field bus cable depends on the run length and the data transmission rate. A shielded $2\times0.5\,\text{mm}^2$ twisted pair may be needed for simple applications, a coax cable for more difficult applications. The field multiplexer itself requires external power (e.g., 220 V a.c.).

Field Installation with a Field Bus (see Chapter 8).

– Principles: Figure 195, variant (g), shows a possible field installation with field bus. This is a simple example for the direct connection of intelligent field instruments to the field bus. For use in process situations, especially in explosion hazard areas, the field bus must be of the intrinsic-safety type, so that the disconnection and connection of a bus subscriber is possible even while the system is operating. The bus can accommodate a combination of transducers powered from the bus and transducers with external power, as shown in variants (d) + (e). Important is that the bus interface of the sensor system can always be connected to the Ex“i” (explosion protected, intrinsically safe) bus.

The field bus cabling concepts illustrated in Figures 193 and 194 all relate to a field bus employing copper wire. The field bus can also be realized in the form of an optical fiber cable or a radio link. Radio transmission plays a role in widely dispersed or poorly accessible plants and is not further considered here. Although the “Physical Layer” is now included in the IEC Draft International Standard, optical fiber has not yet been adopted as a physical data transmission medium. Nevertheless, optical fiber technology offers a serious alternative to copper wire.

– Comparison of Copper Wire and Optical Fiber: A field bus installation that seems appropriate for process engineering is that shown in Figure 194. The design in question has parallel individual lines from field instrument to field distributor. The several lines are combined into a two-wire line in the field distributor, and the two-wire line is then led to the process-level components. The field distributor is a purely mechanical component. In contrast, the optical fiber system has as its distributor an active element that forwards the signals after appropriate conditioning.

The advantage of optical fiber technology is that the information is unaffected by electromagnetic interference. The separate power supply to the instruments must, however, satisfy

the requirements of electromagnetic compatibility. For data transmission via copper wires, it can be expected that proper shielding or the use of special cables will still make it possible to comply with electromagnetic compatibility standards such as those of NAMUR. The price of such a cable is, however, much higher than that of a normal bus line.

Very high data rates can be realized with optical waveguide techniques (5 Mbaud). However, such a rate is not generally needed for the connection of sensor and actuator systems through the field bus.

The number of field instruments that can be connected to the international field bus (now in the definition phase) now appears to be 32. For a copper bus line intended for service in the explosion hazard area and also designed to power the instruments connected to it, the number of instruments is smaller, just nine at present. With optical fiber technology, 32 instruments can be connected to a field bus, even in explosion hazard areas. It depends on the application whether it is desirable to connect this many devices to one I/O card in the process-level component for the sake of redundancy. Present-day I/O cards can accomodate eight channels, roughly the same as the number of instruments that can be connected in explosion hazard areas when copper wire is employed.

One advantage of the copper bus line – that the field instruments can also be powered from the bus line – becomes a drawback in many cases. An attempt is sometimes made to use the bus to supply instruments that could yield more accurate, more reliable, and/or quicker results if more power were available. This dilemma of the instrument designer is often resolved with optical fiber technology, since each field instrument then requires a separate power supply.

Consider, for example, a plant with 200 field devices to be connected to the field bus. A calculation gives the following numbers of lines to be run:

Field bus with copper cable:
Number of buses: 23 (9 instruments per field bus)
Number of separate power supply lines: ca. 20 (10 % of 200)
Field bus with optical fiber cable:
Number of buses: 7 (32 instruments per field bus)
Number of separate power supply lines: 200 (possibly implemented with a field distributor)

- Bus Cable: The definition of a single bus cable for all conceivable applications is unlikely, since it would be too expensive for simple applications. The Draft International Standard for the Physical Layer of the international field bus proposal identifies four cable types for the slow (31.25 kbaud) field bus (ISA/SP50-1992-236P):
  - Type A cable: Shielded twisted pair with following specifications:
    $Z_o = 100\ \Omega \pm 20\ \%$
    Maximum damping = 3 dB/km
    Maximum capacitance = 2 nF/km (wire-shield)
    Maximum resistance = 22 Ω/km (per wire)
  - Type B cable: Twisted pair in a master cable with a common shield and certain specifications
  - Type C cable: Twisted pair, possibly in a master cable, but not shielded, with certain specifications
  - Type D cable: Simple (not twisted) pair in a master cable, again with established specifications; the worst of the variants
- Field Installation of Safety Circuits: How safety circuits will be handled in future installation practice can be examined on an application-specific basis. Simple binary operations, such as "if limit switch trips, cut off feed," should be distinguished from more complex conditions for triggering a safety circuit. Such a condition might be a mathematical operation on multiple binary and analog inputs.

The simple binary operation will continue to be wired separately and realized individually by appropriate modules. It may be that an "autonomous" unit in the field multiplexer, powered only by the system, can perform this binary operation, so that the cable need not be run to the switch room. For a more comprehensive safety assessment, it is conceivable that either a hardwired unit or a stored-program controller could be used.

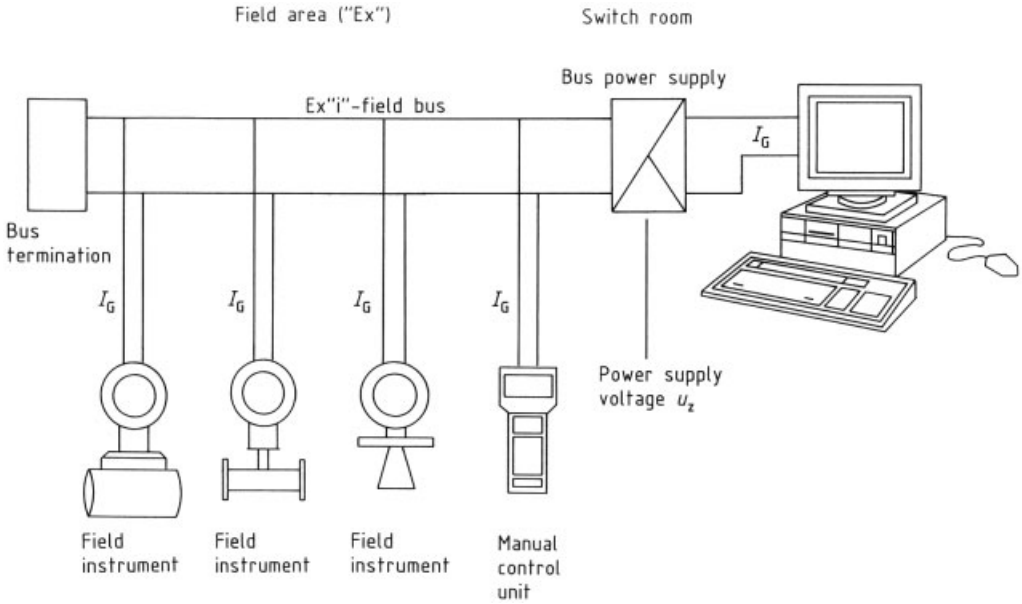

Figure 196. Base concept of the PTB study on field bus use in explosion hazard areas
I_G = base current

Field Bus in Explosion Hazard Areas. The situation is analyzed for copper wire as the data transmission medium. A field bus for service in explosion hazard areas must comply with the explosion protection class "intrinsic safety," so that an instrument can be connected or disconnected while the system is in operation. For low-power sensor systems, such as those usually employed for pressure and temperature, it should be possible to use the bus line for both information and power supply to the instruments.

Even externally powered sensor systems must at least carry out the bus connection in the "intrinsic safety" class, so that these can be connected to the Ex"i" field bus together with instruments not requiring external power.

– PTB Tests: A study by PTB has shown that simple descriptive features make it possible to connect instruments made by different manufacturers to an Ex"i" field bus complying with EN 50020 (Design Standard for Intrinsic Safety) with no special system certification. A variety of cable types were tested for a range of cable lengths and bus supply voltages [446].

Figure 196 illustrates the concept underlying the PTB tests. Only "one active device," the bus power supply, is connected to the field bus. The other devices are "passive" in that they cannot inject power into the bus. In this way, only one device can put power on the line, even in case of a malfunction. The passive devices have low maximum external capacitances and inductances ($C_i < 5$ nF and $L_i < 10$ µH) and minimum input voltages of 9 V. It was further required that each instrument connected must draw a base current of at least 10 mA. The cable length and number of nodes can be optimized by varying the power supply voltage and line length. The concept is applicable to a variety of field buses (Profibus, international field bus, etc.).

The most important finding was that over the wide range of parameters tested, and for various cable types and lengths and various bus power supply voltages, the probability of ignition was not increased by the length of the field bus, the number of bus connections, or the number of instruments connected to the bus. These results are illustrated in Figures 197 (category ib, group IIC, $R'_L = 44\ \Omega$/km) and 5.73 (category ib, group IIB).

For a realistic bus power supply voltage of 15 V, the maximum number of nodes is nine for group IIC and 27 for group IIB. The maximum cable length is 787 m for group IIC, and 310 m for group IIB. The bus power supply voltage or

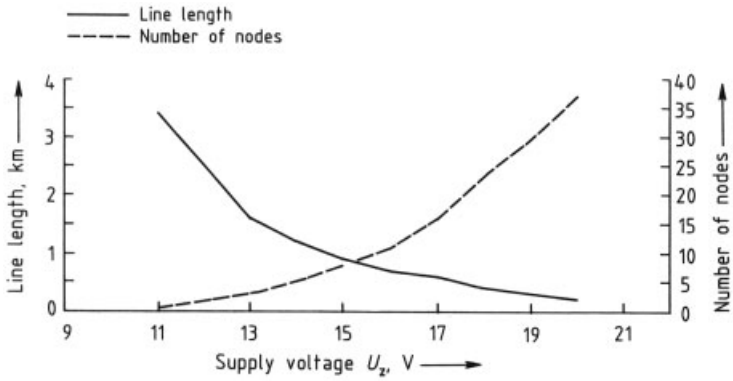

Figure 197. Results of the PTB field bus study (category ib, group IIC)

conductor cross section can be varied in order to adapt to a wide range of applications.

Important observations from the user's standpoint can be found in [447].

Cable Routing. The sensors, actuators, and so forth contained in a plant are usually connected to the process control system by cables. Cable routes are the "streets" along which the cables are laid.

The project manager, in collaboration with fire services and operators, must identify zones of heightened fire hazard and danger zones as defined in Elex V and VbF at the stage of site and building layout [543]. The overall concept for the main cable and piping routes must be defined at an early point in time and incorporated into the construction plans. Whenever possible, cables should be located only in areas of slight danger. Accessibility and ease of maintenance must always be kept in mind; this includes later expansions. The spread of fire through lining penetrations into fire hazard areas must be restricted by suitable covers.

When cables and pipes are routed together, the possibility of a harmful interaction through thermal, chemical, and mechanical attack must be borne in mind. Cable and pipe routes should therefore be kept separate when possible. When this is not possible and an adequate standoff to hot plant apparatus cannot be maintained, heat shielding must also be provided. Where areas of severe fire hazard cannot be avoided and the functioning of the process control system must be insured under all conditions, a fire-resistant lining or cladding or the flooding or spraying of cable routes may be necessary.

The hanger design must be suitable to the load. Lining may necessitate cooling of the affected section (forced ventilation). Ventilation systems must not extend over the boundaries of fire compartments, so that the spread of fires into other compartments will be hindered. Reliable extraction of cooling air (fumes and smoke in the case of a cable fire) must be insured. Chemical attack must be handled by appropriate selection of construction materials, coating, or jacketing.

The number and location of process control rooms must be selected to yield shorter and more secure cable routes (see Section 10.7). These rooms should be in areas of little danger. Process control rooms and principal cable routes should always be configured so that fires in the rooms and localized fires in plant areas will affect only a portion of the plant (decentralization). The insertion of cables into process control rooms should take place outside the explosion hazard area whenever possible. Massing of cable routes near such insertions should be avoided (subdivision).

If power, control, and data cables are run together, the thermal load on the cable route is diminished; there is no objection to this arrangement provided electromagnetic compatibility is insured. High-voltage cables (>1 kV), however, should be run by themselves. Dedicated cable routes may also be needed for emergency power supply and safety system cables, whose functioning must be guaranteed for as long as possible during upsets.

Bringing cables from outside to inside to supply a floor or platform is generally safer than bringing them from a central cable tray inside the building. When cables are run on exterior

walls or in open recesses indoors, any potential threat from their surroundings should be considered. Working platforms connected by fixed ladders should be built on exterior walls around the height of wall penetrations. In outdoor installations, cables can also be routed on the outside and, if necessary, should be provided with protection against thermal effects.

The cable route must be accessible to firefighters in case of fire. "Cable walls" must permit access to areas located behind them.

Exposed cables must not be run in staircases or on walkways that serve as escape routes.

Placement of cables in underfloor ducts should always be avoided. It creates the danger that aggressive fluids will penetrate the cables or that explosive gas-air mixtures will form and spread without hindrance.

In cases where the use of underfloor ducts cannot be avoided, possible sources of danger should be eliminated by appropriate precautions, for example by sand filling or ventilation. When sand filling is chosen, cables approved for underground placement and having adequate thermal ratings should be selected.

If fire protection considerations mandate the placement of cables in a cable pit or vault, this must be of fire-resistant design. Suitable access must be provided for both fire fighting and the operation of vent and exhaust valves. The chimney effect must be taken into account where cable fires can occur in cable pits. If necessary, the proper design precautions should be taken, such as fire partitions between floors and section-by-section removal of smoke. Compartmentalization should also be considered in horizontal cable routes.

Cable Routing inside Process Control Rooms. The design of cabling and wiring inside process control rooms should allow for some key constraints (see also Section 10.7):

- The penetration to the production facility must be gastight and safe with respect to fire hazards.
- For cables and wires that are to be led into cabinets or cubicles from overhead, or brought from overhead to terminal boards or mounting racks, cable trays should be provided below the ceiling.
- For cables and wires that are to be led into cabinets, cubicles, or racks from beneath, a double floor at least 20 to 30 cm high must be provided (Fig. 357). Cables and wires can be laid free on the floor. For larger cable cross sections, in particular, care must be taken that sufficiently large bending radii are allowed.
- On grounds of electromagnetic interference, cables and wires for power and those for data should be spatially separated.
- Massing of cables and lines should be avoided.
- Design must make adequate allowance for retrofit capabilities.

For high power density, continuous air cooling may be needed. The placement of product and energy piping (except process control system piping) in cable pits is always to be avoided.

In areas where there is a danger of dust explosions, design or organizational practices must insure that accumulations of dust are avoided.

Cable Penetration to the Switch Room. Cables running between the field and the switch room must be led into the switch room through a cable bushing in the wall. On grounds of fire protection, a bushing that meets the required fire protection class, such as F90, is selected. The fire department establishes what fire protection class is required. Not only the cable bushing but also the wall and the door to the switch room must comply with this fire protection class. In a low-cost method of penetration, holes of the desired diameter (e.g., 150 mm) are bored at appropriate points in the wall. The holes are half filled with the cables to be led through them. A fireproofing compound is then injected into the spaces between cables. The remainder of the opening is sealed with adequately close-fitting, flexible plugs inserted from both sides (the plugs should foam up in case of fire).

In future, 3-D CAD tools will be important aids in three-dimensional cable routing and piping design.

6. The Process Control System and Its Elements: Process Actuator Systems

6.1. Principles

Chapters 2 and 5 have pointed out the necessity of adopting suitable structuring practices to

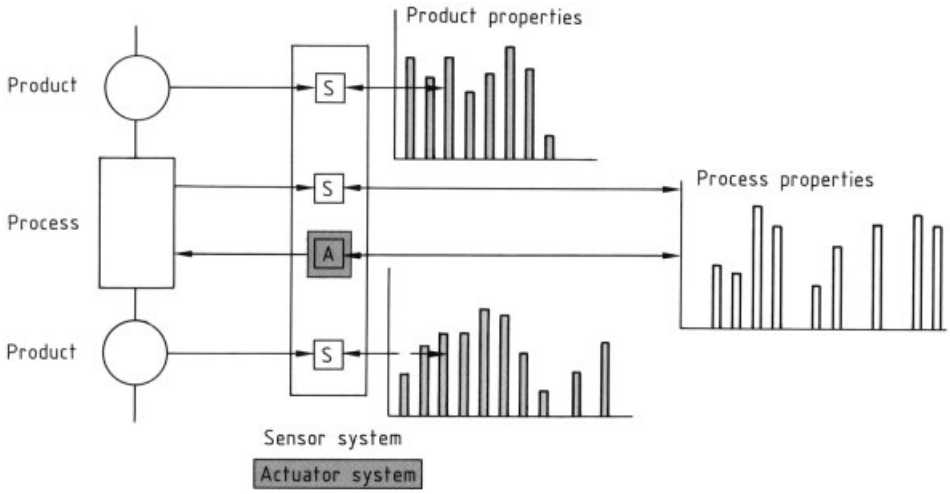

Figure 198. Actuator systems: analogy with the phase model of production

gain an overview of the wide variety of hardware implementations. The need for currency and completeness of description is complicated by the restriction to essentials (for the sake of clarity) and by the fact that technical progress continues after the manuscript of any publication is submitted. For these reasons, an attempt is made here (as in Chap. 5) to achieve some consistency in description, from the information point of view. This can easily be done, in object-oriented terms, by appropriate assignment of attributes. The viewpoint of the modeler is shaped by the objective, as in any modeling process, and if the objective changes then this viewpoint may (or must) change.

Information-oriented process control engineering uses the term "actuator systems" for all technical systems that enable intervention in the process properties, and ultimately the product properties (see Fig. 198). Regardless of the hardware implementation of the system under consideration, this choice of terminology has the advantage that the analysis focuses on the way in which the technical process is influenced or acted on, without necessarily being tied to implementation-dependent nomenclature such as controller, positioner, servo drive, and so forth. Furthermore, the degree of automation, the availability of energy, and the complexity of the actuator system remain initially unaffected by the choice of terms.

If an actuator system is analyzed from this viewpoint, a minimum of three components can always be distinguished (Fig. 199). First there is a functional unit that receives and interprets a control command (actuating command), represented by the control signal (actuating signal) from a higher level. The signal may originate in local components (Chap. 7) of a process control system or in programmable controllers. This functional unit of the actuator system is labeled as "control signal processing" in Figure 199. An energy-furnishing component, controlled by the output signal from control signal processing, can then be identified. In its simplest form, this component (called "servo drive" in Fig. 199) must act as an energy store or else utilize energy supplied via the control signal, passing on this energy in accordance with the control signal. Finally, the third functional unit is the actuator proper, often known as the "final control element." This component is in physical contact with the process properties [544]. The components listed here are found in related forms in the classical control devices (control valve, shutoff valve, ball cock, etc.).

The classical terms control element, control drive, and control signal processing indicate the unambiguously defined direction of action; they do not imply feedback (confirmation of execution or of effect achieved) as to the intervention in the process.

This type of analysis further reveals the problem of separating information flow from energy flow in such simple actuator systems. For example, the energy transmitted by the servo drive to

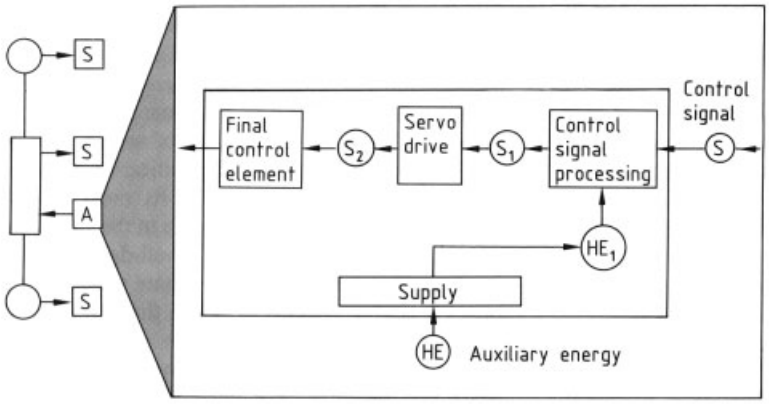

Figure 199. Structure of a simple actuator system

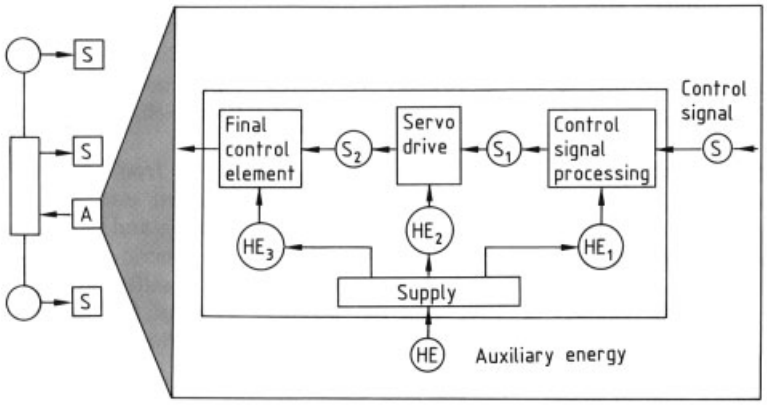

Figure 200. Structure of an actuator system with auxiliary energy

the final control element can simultaneously be treated as a control signal, but one whose action on the process properties is determined by quantities specific to the system and apparatus. It is clear that the resulting dependence of process interventions on unknown effects, such as turbulence, aging, and the like, does not promote goal-oriented process control.

Figure 200 shows an actuator system in which an auxiliary energy supply has been added. This configuration makes it possible – at least in principle – to separate information flow from energy flow, to decouple the actuator and the drive from the signal processing, and to achieve flexibility in design.

The trend toward "intelligent" actuator systems is illustrated in Figure 201. Here an integrated (dedicated) computer takes care of function monitoring, function maintenance as required, control of the process intervention, and other services relevant to design or maintenance purposes [545]. Communication with the dedicated computer takes place via a display and control panel, a handheld terminal, a PC, or a display and control component [546]. In the long term, communications requirements stemming from additional functionalities will make a field bus solution essential.

The shift of intelligence into the field level permits additional functionalities in the field (local) components. Because the real effect of an actuator operation on the process or product properties is of far greater interest for process control than information such as the actuator position, there is a trend toward coupling of information directly at the local level. The speed control of a pump as a function of the actual flow rate, as determined with an integrated sensor, is an example of fast control loops on the local level.

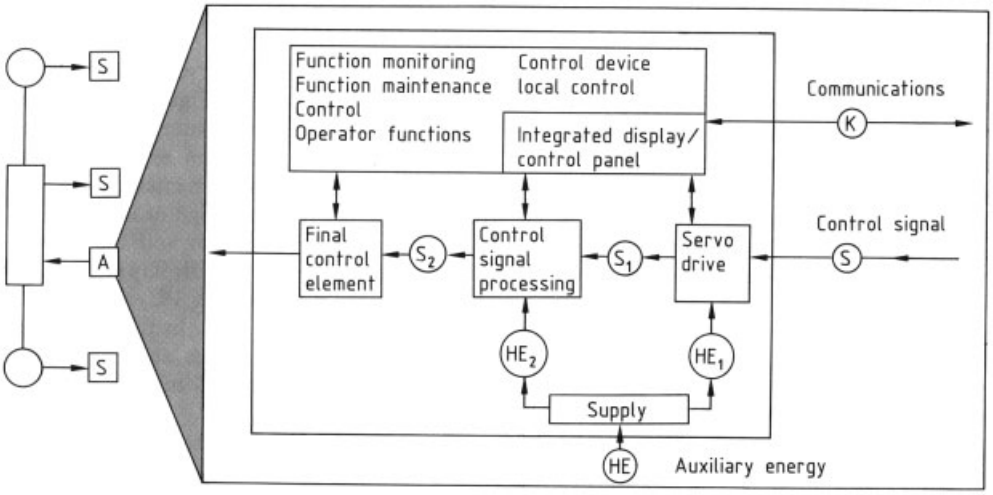

Figure 201. Structure of an "intelligent" actuator system

In order to survey and classify actuator systems, it is necessary first to define an appropriate viewpoint. In what follows, some criteria are described, leading to the classification presented here.

A breakdown of actuator systems meaningful in the automation context might be based on the degree of automation of process interventions (manual operation, automatic operation), the type of auxiliary energy (electrical, pneumatic, hydraulic, etc.), or even hardware criteria. With regard to information-oriented process control engineering, however, a different type of classification appears more desirable. Actuator systems are analyzed from the standpoint of possible actions on material and energy streams. An attempt to set up such a classification is described here.

The action of an actuator system can be characterized in terms of the possible change in process properties. The manifestation of a property category can be described as changing "continuously" or "discontinuously." The taxonomy that follows therefore uses the criterion of a continuous versus a discontinuous control range.

For example, a mass stream can be controlled in continuous fashion by a control valve or discontinuously by a shutoff valve.

In continuous actuator systems, a further division into proportional and integrating types is possible. Similarly, discontinuous systems can bring about defined absolute changes or incremental changes in the process properties.

An additional feature of actuator systems is important from the standpoint of safety. A system may have an undefined position on failure, where the actuator remains in its most recent position in case of an interruption in the signal or energy flow, or it may have a well-defined safe position (fail-safe). An intermediate group comprises actuator systems in which the drive has a memory characteristic that takes the undefined failure position of the actuator over to a well-defined rest position. For certain processes, this represents an ideal solution from the safety standpoint.

Figure 202 shows a taxonomy of actuator systems based on the criteria discussed above as well as the energy flows to be supplied. Because the application and design features here are generic properties, they apply to all the actuator systems discussed here.

Design features – ranging from physical form, material of construction, and control range to communications, safety, and manufacturer – are listed in a second generic class.

Actuator systems are first classified by objective, that is, actuator systems for material streams are distinguished from those for electrical energy streams.

For each group, systems that effect a quantitative change in the streams are next separated from those that essentially bring about a qualitative change (most often defined in energetic terms). The first type, acting quantitatively on the streams, are described as "controlling" sys-

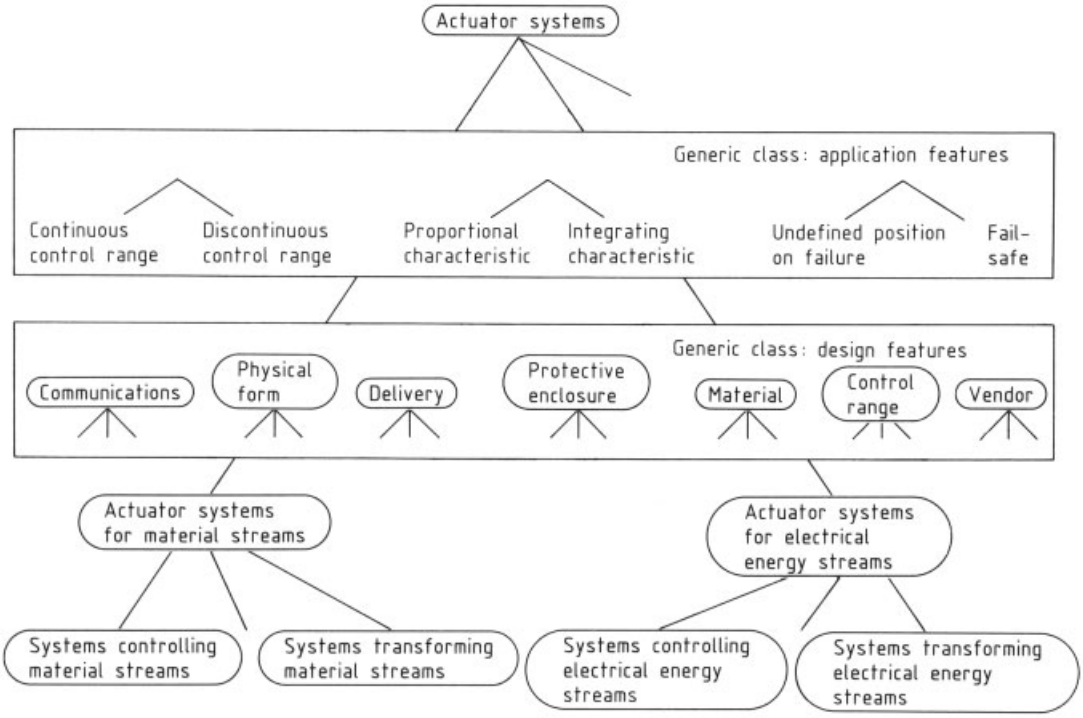

Figure 202. Taxonomy of actuator systems for material and energy streams

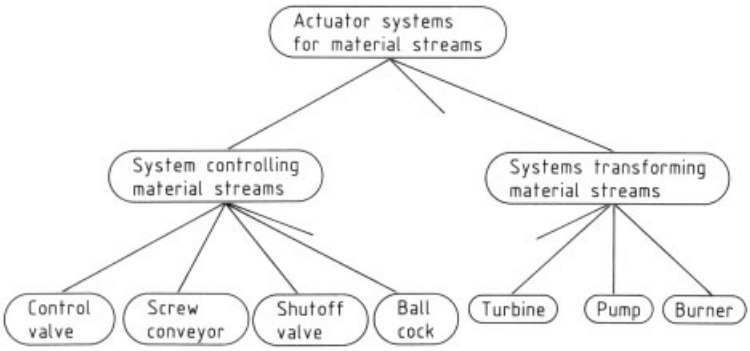

Figure 203. Taxonomy of actuator systems for material streams

tems in Figures 202, 203, 204, while the second type is described as "transforming" systems.

If we consider first actuator systems for material streams (Fig. 203), the group of controlling systems include "distribution" systems such as screw conveyors, metering apparatus, etc., and "throttling" systems such as control valves, shutoff valves and gates, cocks, and so forth. In manufacturing and process engineering these two classes are often regarded as independent groups.

The category of transforming systems includes not only the turbine and the burner, but also the pump. This is admissible since the primary use of a pump is to increase the pressure of a material stream; thus the pump, as an actuator system, alters a qualitative property of the material stream.

Other examples of material transforming systems can easily be classified; for example, a stirrer plus drive in a stirred tank reactor, or of a roller, a press, or a tubular reactor.

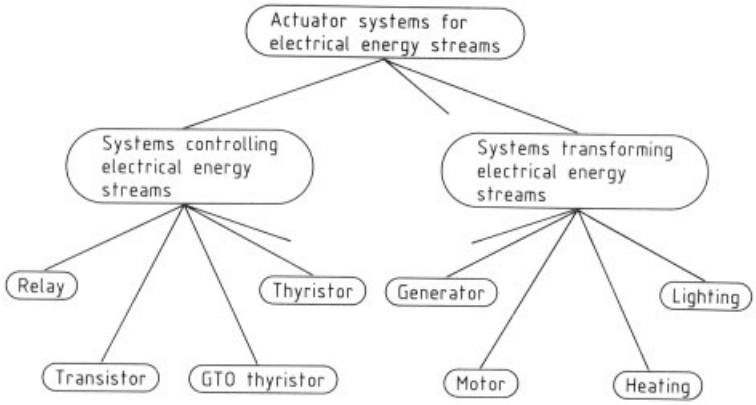

Figure 204. Taxonomy of actuator systems for electrical energy streams

Figure 204, based on the same classification principle, shows a taxonomy for actuator systems working on electrical energy streams. The quantity controlling systems include relays, thyristors, transistors, and gate-turnoff (GTO) devices (see also Section 6.3); the energy transforming systems include generators, lighting, motors, heating, etc.

A more detailed breakdown of energy controlling systems can be found in Section 6.2, which also distinguishes between switch and positioner on the basis of whether the control range is continuous or discontinuous.

Development trends in actuator systems can be briefly described as follows:

- Development from the actuator to the actuator system with integrated intelligence and the incorporation of lower-level control functions
- Expanded communications capability with field multiplexer or field bus [547]
- Improved service life, especially for actuators, due to novel materials, designs [548], and automatic maintenance techniques [549]
- Increased use of electronic frequency converters for motor speed control[550]

The trend toward intelligent actuator systems is possible because of advances in microelectronics. For example, a valve position controller with an integrated 16-bit microprocessor is now available [551]. Self-parametrization makes it much easier to adapt the device to the drive and the valve; the internal control loops of the actuator are more powerful; and the adaptability to a range of service conditions has led to a reduction in the number of position controller types needed.

The augmented functionality of actuator systems cannot be fully utilized unless there is a suitable facility for communication with higher levels in the process control system. Parametrization, function monitoring, and maintenance require special information, which will be made available by future field bus and communications concepts (see also [552] and Chap. 8).

Section 6.3 deals with the trend toward variable-speed drives. It is seen that improved control behavior (linear control characteristic), reduced energy consumption (especially under part-load conditions), motor power and motor losses restricted to the design values, and optimal matching to the process machinery used offer substantial potential for economic and technical optimization.

6.2. Actuator Systems for Material and Energy Streams

Introduction. Together with sensor systems, actuator systems link material, energy, and information streams in processing and manufacturing systems. The control path with sensor system, controller, and actuator system forms a closed control loop. Changes, whether in perturbing values, setpoints, controlled variables, or other quantities, manifest themselves in the system and also in the signals transmitted to and by these devices. With the aid of actuator systems, the material or energy stream is set in the con-

trol loop as a function of the system deviation. The action also depends on the characteristics of the control path, the sensor system, and the controller with its control parameters influencing its dynamics. With regard to the dynamics, the control element can be included in the control pathway and the servo drive (actuator drive) can be included in the control device. In the open loop case, the intervention in the process is effected through a control command, which takes the form of a control action or a manual action. The control pathway need not carry feedback.

In actuator systems, the control command (the output of the controller) must be tranformed to a control signal (Fig. 199). Servo drives in the form of power amplifiers are needed for this purpose. These may be pneumatic, electric, or hydraulic, but hydraulic drives are rarely used in process engineering. Because of their well-known advantages in manufacturing and robotics, however, hydraulic power amplifiers are preferred in these fields. They are robust drives whose high power density results in good control of complex motions, accurate positioning under load, and wide speed ranges. Problems, however, such as poor damping, marked nonlinearity, and load-dependent eigenfrequencies, should be mentioned.

Continuous and switching control elements are distinguished, just as control elements with memory must be distinguished from those with a well-defined rest position. Single action modes or combinations may be required, depending on process or safety requirements (see also Fig. 202).

Continuous control elements effect a steady or gradual change in mass or energy flow within the design range, while switching elements act in a step-by-step manner. In process control engineering, there is often only a minor difference between the two types. Control elements of continuous or switching type can be designed by modification of globe and plug valves, cocks, shutoff valves, and the like. The required characteristic determines whether the required accuracy of positioning or control can be achieved under given pressure conditions. The characteristic associated with the system (inherent characteristic) can be matched to the operating characteristic. With regard to accuracy of positioning, a corresponding resolution of the control signal must be ensured. Control elements such as slide valves or solenoid-actuated valves have a design that makes them unsuitable for continuous control and positioning.

Control elements with electric motor drives have memory and remain in the last position adopted if the auxiliary energy goes off or the control signal fails. Ordinarily, control elements take up a well-defined position on a failure of auxiliary energy or control signals. For material streams, this is generally full open or full closed, and for energy streams, full on or full off. In safety devices, this position is called the "safety position." Control elements such as gate valves or stop cocks with double-acting pneumatic drives can be made to exhibit memory by the addition of blocking valves; the use of return springs or pneumatic buffers can result in their exhibiting a well-defined rest position. The size of the pneumatic buffer depends on the number of switching operations and the quality of sealing. The advantage over energy storage in springs is that several switching operations can be performed after the occurrence of a fault.

The term "control" is explained in DIN 19 222 [553]. Control is the changing of mass, energy, or information flows by means of control elements. Standardized definitions of the terms actuator, actuator system, control element, controller, and control machine have not yet been adopted. They are used at present even though the demarcations between them are often not clearly defined.

Examples of Actuator Systems; Delimitation. Controllers or final control elements are also referred to as actuators or actuator systems, in analogy to sensors or sensor systems. Ordinarily, "sensor" means an element that performs a detection or measurement function. If this element is equipped with signal processing and other modules that permit further options, the result is a "sensor system." The analogy to "actuator" and "actuator system" can be made complete only if the "actuator" is defined exclusively as the part that comes in contact with the product.

Accordingly, the actuator is the final control element or control machine itself, that is, the throttling organ, the pump, the compressor, the blower, the resistance coil (for heating), or the fluorescent tube or incandescent lamp (for lighting). The corresponding actuator drives would

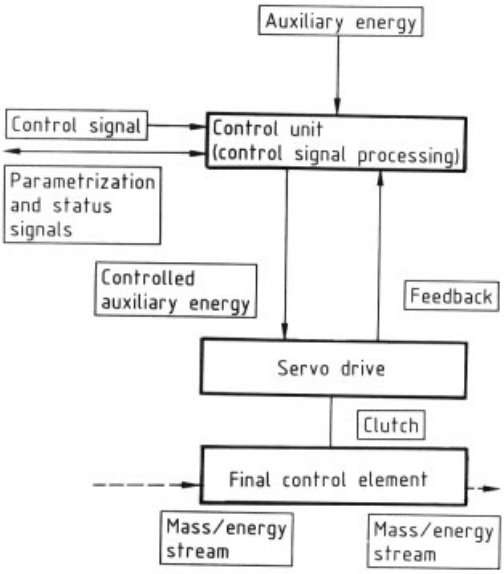

Figure 205. Structure of an actuator system

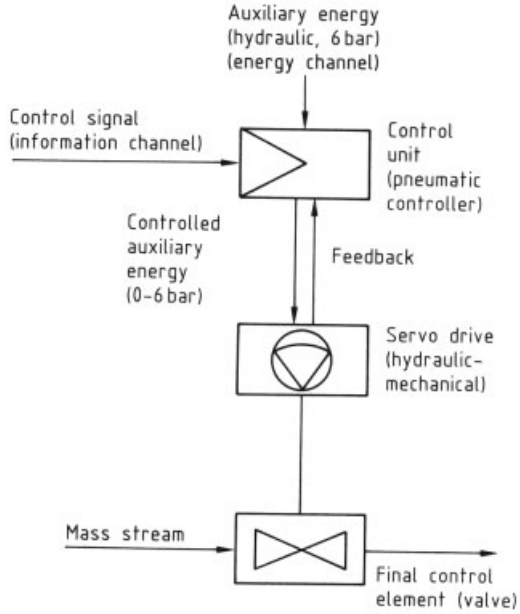

Figure 206. Actuator system for mass streams: structure of a control valve

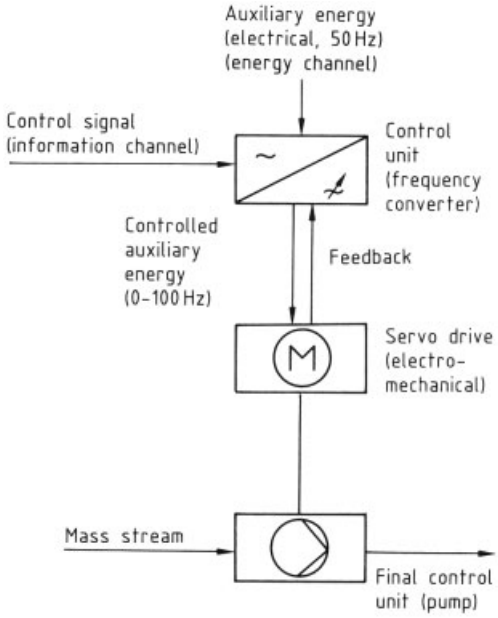

Figure 207. Actuator system for mass streams: structure of a variable-speed pump drive

be (for controllers of material streams) the diaphragm drive, the piston drive, the rotary drive, the motor, the motor with converter when speed control is required, or the switch or power controller for heating for lighting. The last two items apply to the case of controllers for energy streams. In common usage, however, the terms actuator, actuator system, controller, and final control element are often employed as synonyms, generally for the complete device.

Generally, the actuator system (Fig. 199) consists of a final control element (control machine), a servo drive, and a control device. Figure 205 shows the setup. The final control element is moved by the servo drive in order to change the flow of material and energy. The servo drive is driven externally with the control signal from a control device. With further devices, parametrization, configuration, status, feedback, and other signals can also be transmitted.

If these terms are applied to two well-known cases – the control valve and the variable-speed pump drive – the consistent representations in Figures 206 and 207 are obtained. Figure 206 shows a controller made up of a control element in the form of a valve (fitting) with a throttling organ, a pneumatic drive, and a control device. The control device is a positioner, which processes the control signal (as setpoint) to the position of the drive (as actual value) in the course of a closed-loop control operation. Along with the control section, the positioner includes a power section, where the control signal is amplified and which is to provide the desired positioning of the

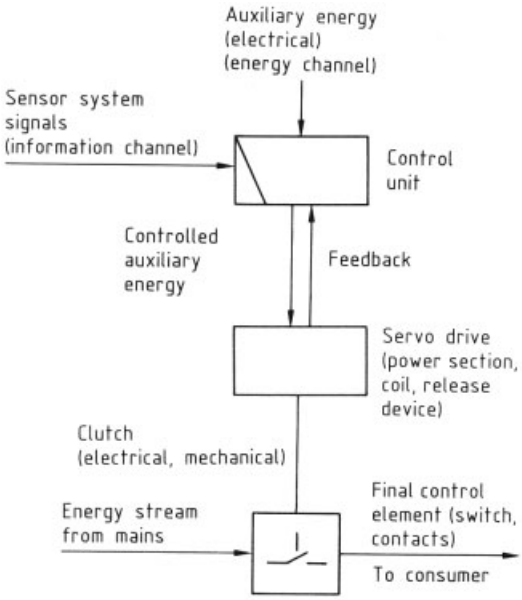

Figure 208. Actuator system for energy streams: structure of a switch for control or safety purposes

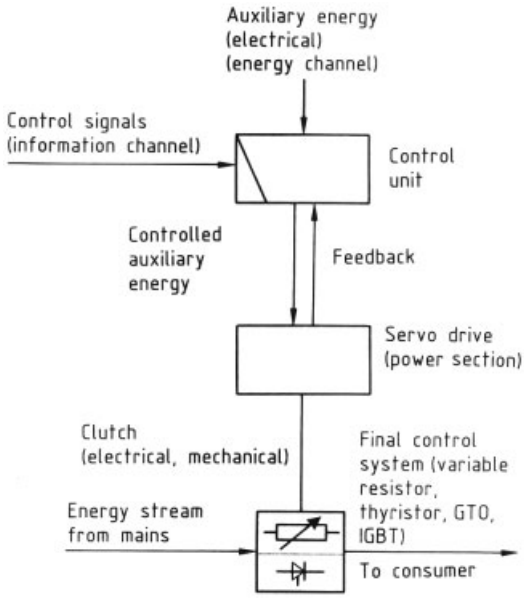

Figure 209. Actuator system for mass flow: structure of a positioner for a heating system

control element with the required quality and dynamics. The terms control element, actuator, and fitting are often used synonymously.

The variable-speed pump drive is a typical interconnection of a control element in the form of a pump, a control drive in the form of a motor with power section integrated into the frequency converter, and a control device in the form of an information-processing component. The control signal is fed to the control device, which monitors, reports, limits, diagnoses, tests, generates error messages, archives, establishes a connection as needed, communicates, and so forth; it can also be parametrized and configured (see also Fig. 201). Section 6.3 goes into more detail on variable-speed drives with d.c. and a.c. motors.

The simplest case of control of an energy stream is the making and breaking of a circuit to a consumer. Figure 208, which is based on Figure 205 and analogous to Figures 206 and 207, shows an actuator system whose control device receives the sensor system signals of interest to the operator, and appropriately switches auxiliary energy to the control drive (coil) in order to activate the control element (switch).

The terms can also be extended to a positioner in a heating system (Fig. 209). This is a continuously adjustable heating device, whereby the control signal is generally a unified current signal (0/4 – 20 mA). The control device, with monitoring, testing, reporting, communications, and display functions among others, drives the control drive or power section. The energy stream to the control element or to the heater coil is adjustable from the control room or locally. It can be set automatically or by hand – depending on what devices are connected in circuit with it – anywhere in a design range specified by the designer.

Points to consider in the design of a controller for a material stream include dynamics, operating point, positioning forces, operating range, characteristic, safety position, operability, service under explosion hazard conditions, and sealing. Design steps include the determination of the nominal diameter (DN), the rated pressure (PN), the construction materials for components in contact with the product, the type of sealing, the positioning forces, the noise emission, the possibility of cavitation, and the matching of the inherent characteristic to the operating and control drive characteristic. In the case of the variable-speed pump drive or a control machine, special consideration should be given to the motor startup characteristics, the by-

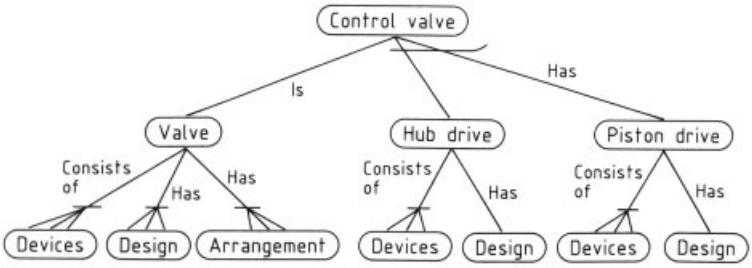

Figure 210. Descriptive structure of a control valve for design and installation

pass circuit, the power rating of the motor and the converter, the motor heating, the torque ratios, the proper matching of the actuator (in the form of a pump) to the motor, the cooling of the converter, the use of tested and certified devices for service in Zone 1, and so forth. In this connection, a number of regulations, specifications, technical guidelines, and so on must be complied with. A process control station data sheet is filled out in which the product, device, process, and design data are entered along with data on safety, availability, quality assurance, etc.

With the station data sheet and the requirements from the piping and instrumentation diagrams or the bid specifications, the appropriate design is prepared and the process control system station plan is generated. The procedure is illustrated for a control valve (Fig. 210); see also the generic classes in Figure 202.

The control valve consists of the valve itself plus a lever or plunger drive, depending on the motion of the throttling organ. The valve can be realized in the form of various devices, such as a globe, peristaltic, or plug valve. The design specification differs from one type of throttling organ to another. The configuration or the way the devices are installed in a process plant can also vary. The best installation position depends on the product, the heating medium, the sealing system, the flange design, and so forth. The design of the lever drive can be based on equipment offered by several manufacturers. As a rule, however, only a single design specification is considered. The same holds for the case of a plunger drive. The devices with their designs, the layouts, the selection, the installation of these devices have been fully described elsewhere. There are also many DIN, IEC, and VDMA standards. In many cases, calculation and design programs are offered with the respective data sheets. Examples of commercially available programs can be found in [553].

The path of the control signal via the control signal processing section of the control device, the control drive, and the final control element itself, where it is transformed to a position value, is shown in the middle part of Fig. 201. The control element alters a process value in a process element. Generally, one measured quantity is provided to the control device from the process. This quantity is needed as a control or monitoring quantity for the controller or actuator system. Feedback to the process control system or to local components takes the form of messages. In the control device, the actuator computer, together with its communications devices, is shown in Figure 201. This device performs function monitoring in the form of self-tests or externally initiated diagnostic tests. The result of these tests is directly displayed locally in error messages, error displays, status messages, and so forth, or (in future technology) will be transmitted over a field bus to a remote location. An integrated display/control panel is always included in complex actuator systems for local parametrization and configuration. Human – machine communication takes place via this interface.

Taxonomies of Actuator Systems for Material and Energy Streams (see Figs. 202, 203, 204). In what follows, an attempt is made to devise taxonomies of various actuator systems under various criteria [554]. The entity – relationship model is employed, converted directly to file structures (see Chap. 2 and [555]).

Controllers (actuator systems) are first classified in terms of their tasks. Controllers for material streams and those for energy streams are grouped separately (Fig. 211). At the next level

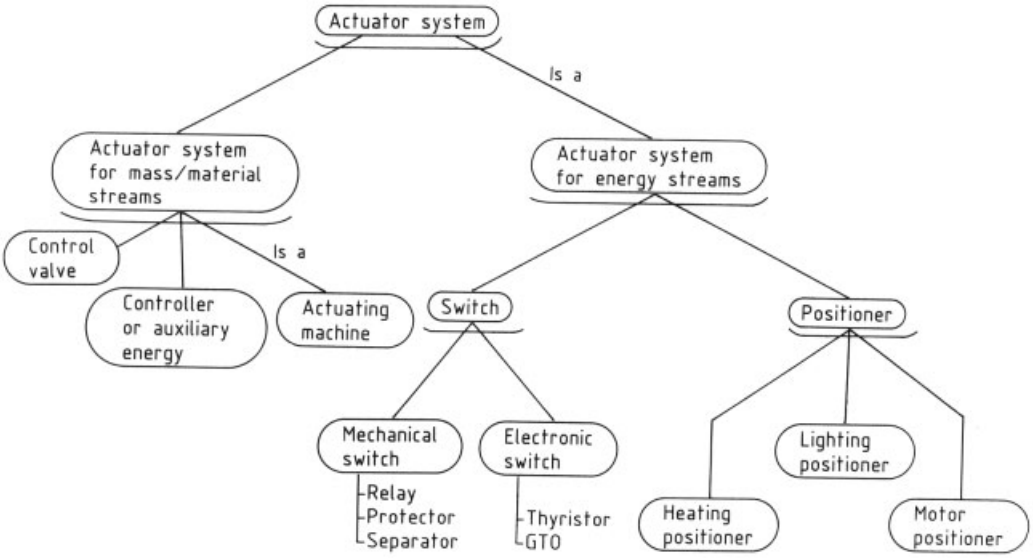

Figure 211. Taxonomy of actuator systems for mass and energy streams

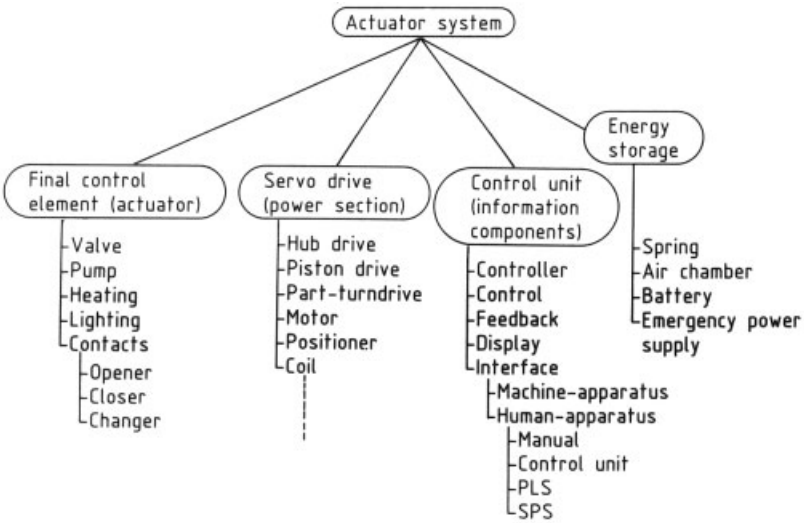

Figure 212. Taxonomy of actuator systems for mass streams according to functional criteria

of detail, controllers for energy streams are broken down into switches (e.g., Fig. 201) and positioners (Fig. 202).

Figure 212 shows the classification by technical and functional criteria. Each actuator system consists of a control element, a drive, a control device, and energy storage, if required.

The control element (actuator) may consist, for example, of a valve with a throttling organ, a pump with case and impeller, the inductive coils of an induction furnace, a halogen lamp, the mechanical contacts of a power contactor, or the power thyristors or transistors of a positioner. The drives (power section) are listed in a similar way. In more complex actuator systems, the control element and drive are based on a control device (information-processing component). This component, in the form of an information or signal-processing section, processes, conditions, and transmits the signals to the actuator system and also from the actuator system to and from higher-level devices.

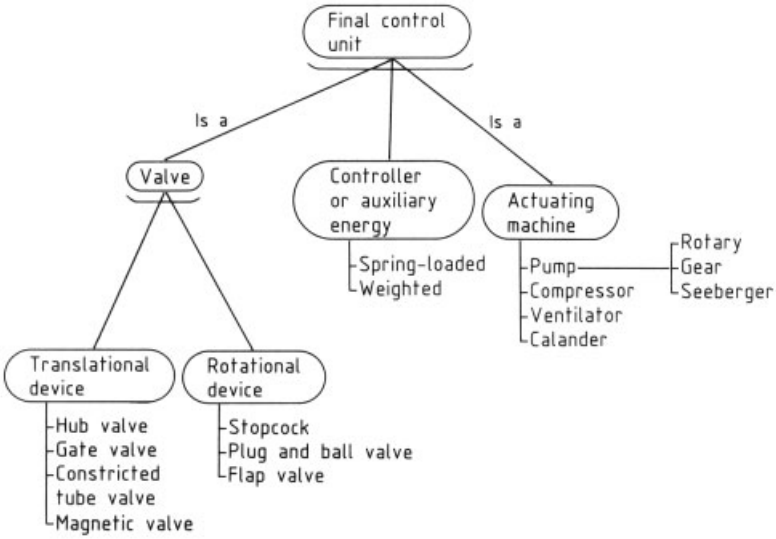

Figure 213. Taxonomy of final control elements for mass streams

When safety and availability requirements are more stringent, energy storage is required. If the auxiliary energy fails, this stored energy brings the control element to a desired safe position via the drive. Examples are compression and tension springs, which release energy on demand and test the "emergency case" upon every switching operation. A pneumatic buffer can be used as energy store; from which energy stored in compressed air can be withdrawn for a certain time and with which repeated switching operations can be carried out. In other applications, batteries and backup power supplies maintain safe operation. Battery-backed systems can be kept operable without interruption; with emergency generators, on the other hand, interruptions must be reckoned with from the outset. It is often sufficient that the motors continue to run after a short interruption.

Final control elements for material streams are classified as shown in Figure 213.

The control element may be a control valve, whereby the function of the throttling organ is carried out by translational or rotational motion. The throttle bodies are said to be straight-line (translational) or rotational.

Another possible classification, in accordance with the terms of DIN IEC 534, Part 1, is illustrated in Figure 214.

Mention should be made of the many and varied process connections, sealing systems, fitting types, throttling organ configurations, noise suppression options, drive design possibilities, and control device types.

The control element can also be a controller without auxiliary energy but employing spring or weight loading to insure the admission of a certain quantity (setpoint value) of material per unit time. Large numbers of such devices are used not only in process engineering, but also in heating and air conditioning as well as other branches of industry.

Control machines are becoming increasingly important with the development of modern power electronics, because they can be operated much more efficiently in variable speed mode than controllers with throttling organs. In systems with throttling devices, energy must first be introduced into the material stream, and then part of the energy dissipated by throttling [556]. Cavitation can result, with outgassing, phase-change effects, and noise as the consequences. For this reason, variable-speed motors are increasingly used to drive control machines (see Section 6.3 and [557]).

Drives. Each actuator system has a power section or drive system (Fig. 215), which must be matched to the control element or actuator. In devices with throttling organs, translational and rotational control drives are distinguished. Motors can be classified into d.c., multiphase (Fig. 216), compressed-air, and other types.

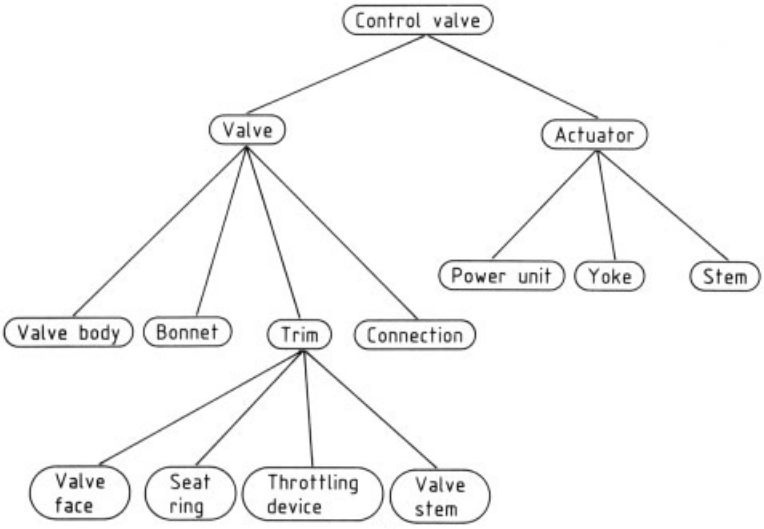

Figure 214. Taxonomy of a control valve [DIN IEC, Part 1]

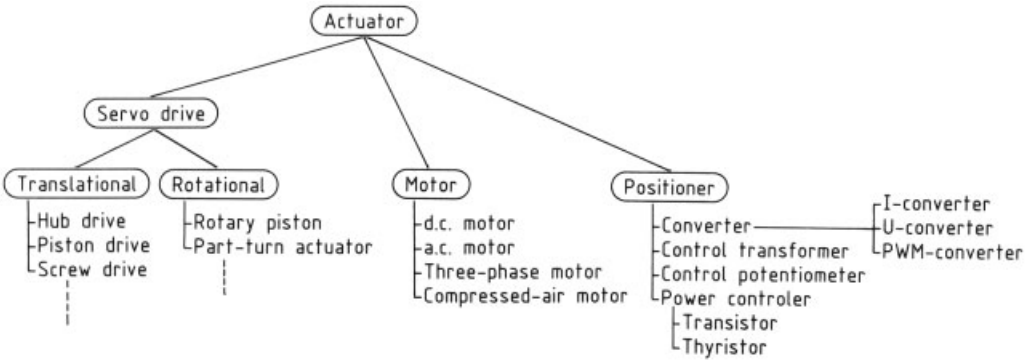

Figure 215. Descriptive structure of an actuator or power section

Depending on the network form, power, required dynamics and quality of control, field and later maintenance conditions, and so forth, the designer selects an asynchronous, a.c., synchronous, or d.c. motor. Attributes of these objects include power, rated voltage, speed, number of pole pairs, motor and explosion protection, case shape and size, type of protective enclosure, weight, ventilation, position of terminal box, manufacturer, supplier, model number, and stock number.

Devices for controlling energy streams can be broken down into switching devices and devices continuously controlling the energy stream (Fig. 211). Typical representatives of switching devices or switches are relays, contactors, disconnectors, power circuit breakers, and current-limiting circuit breakers. Each of these consists of a drive (operating mechanism) and a mechanical system, which in the simplest case (e.g., in a low-voltage, high-breaking-capacity breaker) is operated by hand. More complex devices employ a magnet coil or, in the medium- or high-voltage range, a motor drive. Switching functional elements are normally-open, normally-closed, and changeover switches. Figure 217 illustrates this taxonomy, but only for the mechanical switch side; electronically operated devices with corresponding functions can be classified similarly.

In the continuously operating devices, the energy stream or power can be set within a certain range, so that matching to a certain operating point within a stated range is possible (see also

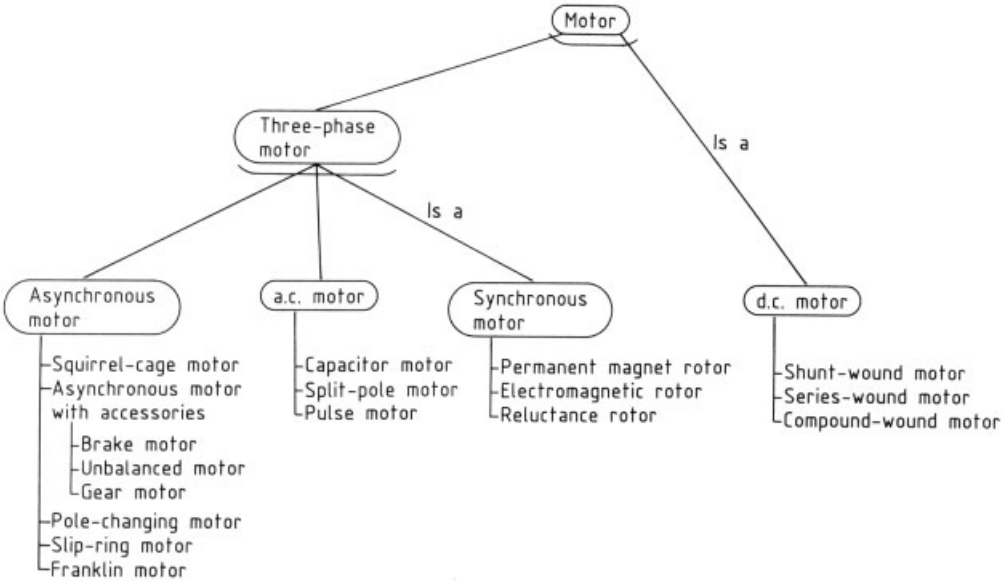

Figure 216. Taxonomy of servo drives for the example "motor"

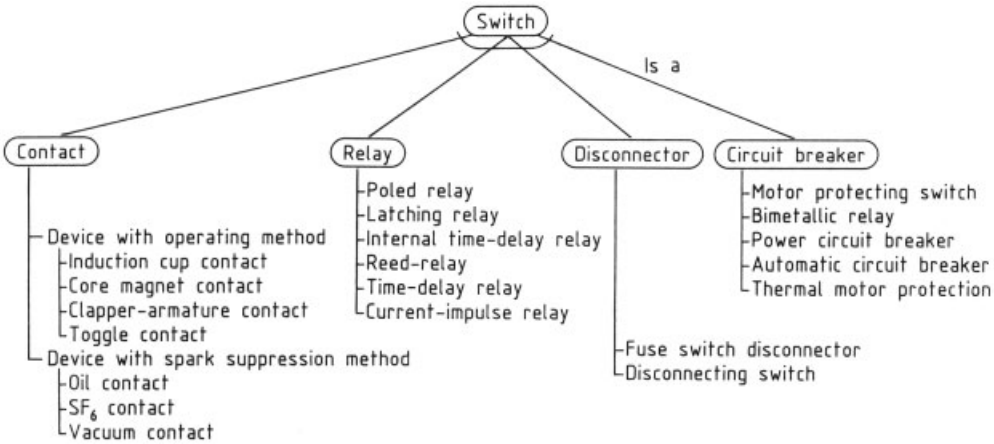

Figure 217. Taxonomy of discontinuous actuator systems for energy streams

the schematic diagram in Fig. 209). Positioners or position controllers can be used to set the voltage, current, frequency, and other values [556, 558]. Rectifier technology includes devices to transform or control electrical energy by the use of rectifier valves. The fraction of electrical energy switched, controlled, and transformed in the power electronics will increase steadily with growing requirements on the controllability and transformation of electrical energy. Applications for power electronics include industrial drives, electric heating, and energy generation and distribution. The basic functions of positioners can be classified as shown in Figure 218 (see also Section 6.3).

The basic functions are rectification, inversion, a.c. conversion, and d.c. conversion. In rectifiers and inverters, which are not regarded as actuator systems, the direction of energy flow is specified. In a.c. converters and d.c. converters, the direction of energy flow can generally be reversed. Variable rectifiers make it possible to control the output voltage and frequency and thus the energy flow. A possible classification of

rectifiers by direction of energy flow, into unidirectional and bidireactional rectifiers, would also be conceivable where the energy flow can be reversed. The a.c. and d.c. controls shown in Figure 218 should be emphasized; these can be used to control heating and lighting and also in power systems. Further, the various link converters used for motor speed control should be mentioned. There are various pulse-amplitude-modulated frequency converters, pulse-width-modulated frequency converters, and link converters. These devices are commonly referred to as motor controls, heater controls, and lighting controls (see Figs. 212 and 219). Their attributes as individual objects are power, voltage, current, driving signal, manufacturer, accessories, limiting, torque limiting, type of starting torque enhancement, power and torque transducers, and braking unit.

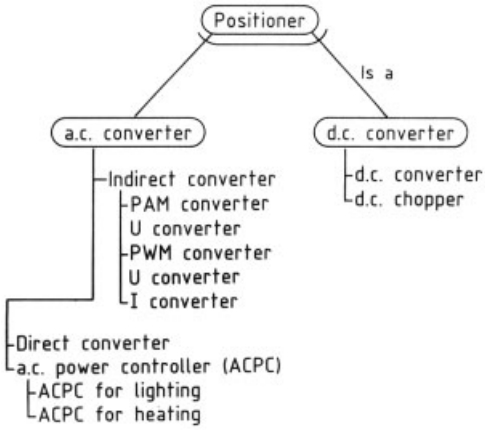

Figure 218. Taxonomy of continuous actuator systems for energy streams

In general, without going into the individual technical differences between frequency converters, a classification such as that shown in Figure 220 can be employed. As shown above, the devices can be broken down into final control element, control device, and power section (control drive). The entire equipment is also called a control machine in accordance with IEC 514. All centrifugal, turbine, gear, and plunger pumps, as well as special-purpose pumps, are included in the category of machines for handling gases, liquefied gases, and liquids. The control drive consists of a power section and the motor. The power section and the control device are commonly integrated, except for the transformer of the rectifier.

Control Device. The control device has two functions: to match the actuator system to various equipment connected before it and to the mode of operation, and to match it to the final control element and the drive itself. Figure 221 shows a possible classification.

In more complex actuator systems, control devices are made up of devices such as controllers and monitors with their respective memories, actuator buses, display/communications/feedback devices, accessories, integrated display and control panels, and so on. It can be expected that the control device will go through major changes in the future. The control device should make possible matching to the connector in conventional or smart technology (see Section 5.3 and Chap. 8) or in bidirectional field bus technology. The type of central electronic controls common up to now, with expensive cabling between the sensor and actuator systems and extensive setup operations during commissioning, will be modified for the sake of more easily understood, less expensive, distributed signal processing. For the user, such a device will become a "plug-and-play" unit in spite of its complexity. Not yet fully achieved are networking capabilities and a consistent, manufacturer-neutral, menu-based operator interface as desired by both operator and designer. Futhermore, major efforts are needed in order to bring about a unified field-level communications technology based on internationally accepted standards. Such technology will permit interchangeability of hardware across manufacturers. These interfaces must satisfy increasingly stringent requirements, partly because of the trend to "rationalization" and the desire for a central service. In the near future, this area will become a focus of the activities of manufacturers, users, and developers (see also Section 5.3 and Chap. 8).

The control section, as a module of the control device described above is needed for the human – device interface (Fig. 222). In the simplest case, the actuator system can be set with a handwheel, or via a manual control signal from the supervisory device of the controller or from the process control system. In automatic mode, positioning takes place after signal processing

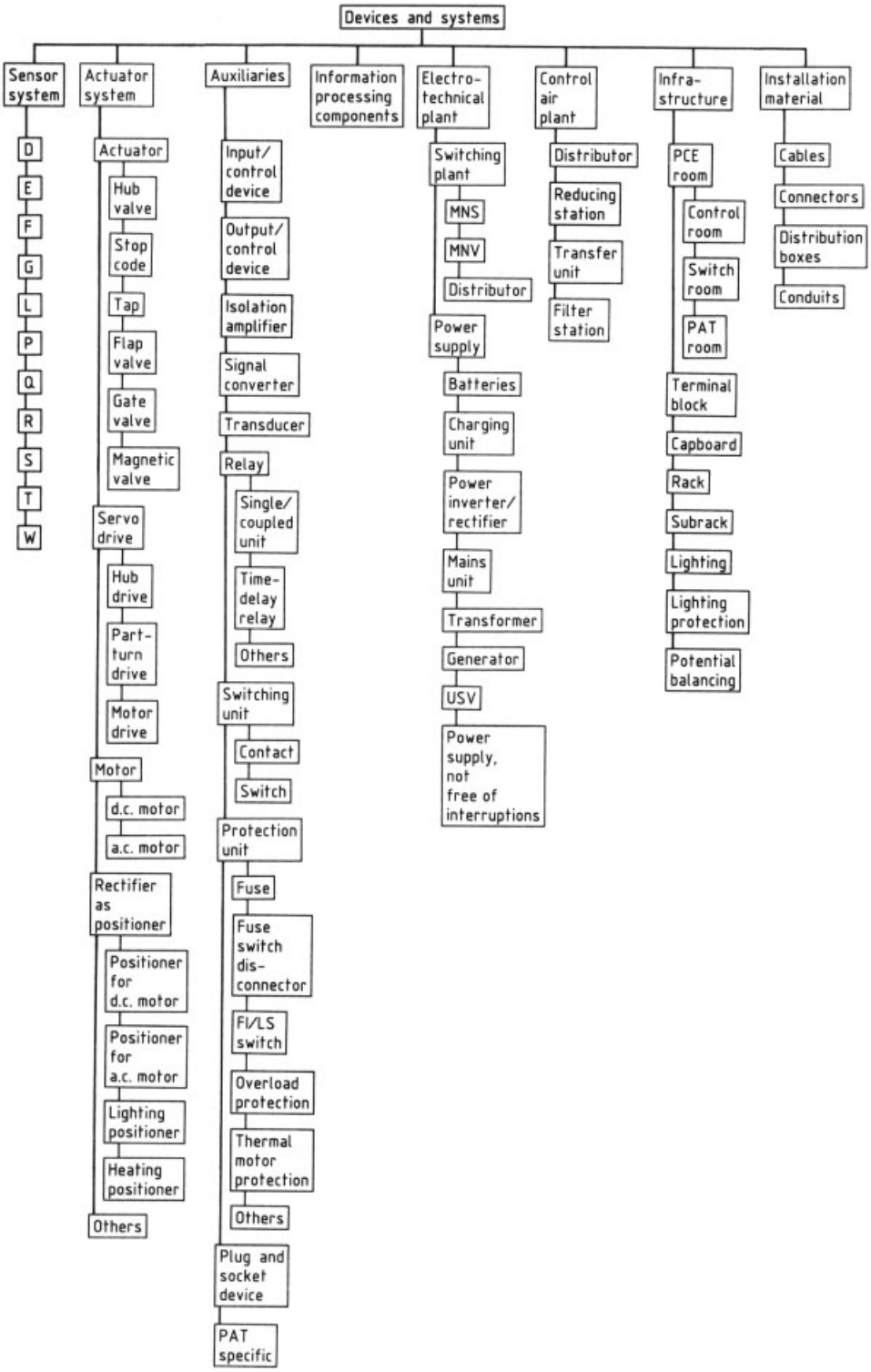

Figure 219. Device object tree for process control systems at the field level

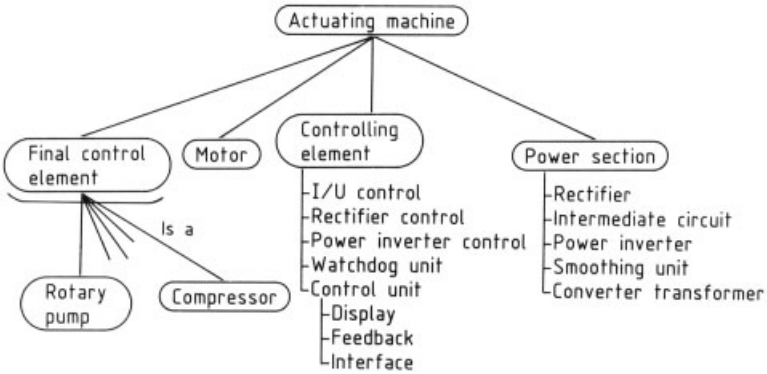

Figure 220. Descriptive structure of an actuating machine

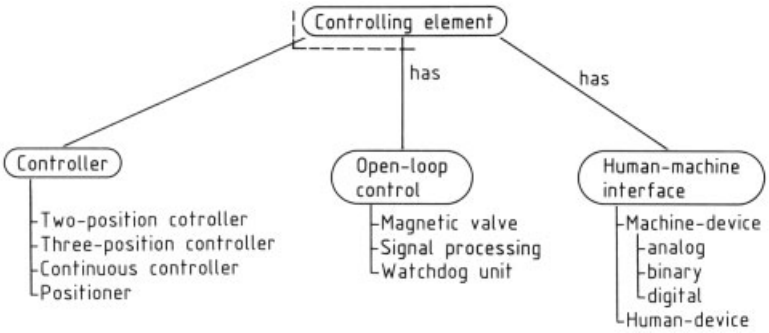

Figure 221. Structure of a controlling element

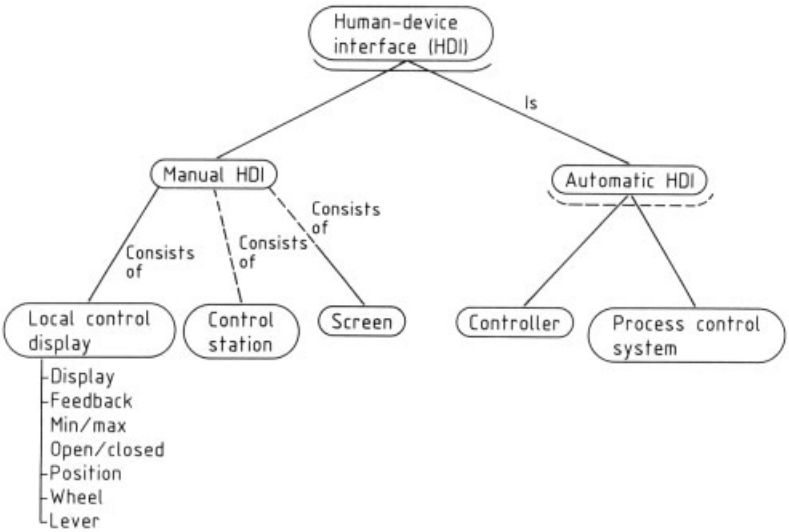

Figure 222. Structure of a human – device interface

with specified setpoint values; this function is performed automatically.

Figure 222 illustrates the possibilities of manual and automatic operation. Thus, there are two interfaces: a human – device in manual mode, and machine – device in automatic mode.

In manual operation, interventions can be accomplished by

- A local control station, in the simplest case a handwheel, lever, or the like
- A display and control panel, integrated into the actuator system, on which the system can be set
- A control panel at the supervisory device, for manual and automatic operation
- A screen mask on the display and control component where manual control signals can be set or nominal values and tolerances can be specified

In automatic mode, interventions take place

- Between the final control element and the field controller
- Between the control element and the individual controller in the control room
- Between the control element and the controller component in the process control system, stored-program controller, or PC

Outlook. The object-oriented representation in data processing has the following advantages for the user: Objects are presented in an intuitive, structured, transparent form, and searches (e.g., for a certain device) are carried out in prompted mode.

Because devices in this representation are merely sketched, further detailing is necessary at the lower levels if installation, shop, and maintenance workers are to find individual repair or replacement parts with given specifications and given relationships to other devices. Similar requirements arise in the design of a process monitoring and control system, where the specification document dictates the selection of devices; their identification by, say, a CAE plot symbol in a process monitoring and control functional chart; or their representation in a positional world with lower- and higher-level devices.

On the basis of Figure 219, this point is illustrated for the example of a device object tree [559].

In the object-oriented representation, the devices required for a process control system are described by groups. For example, they can be broken down into structural elements of equal rank: sensor system, actuator system, auxiliary device, information-processing components, electrical, control air system, infrastructure, and installation materials. Sensor systems are classified by physical quantities; actuator systems, by the criteria discussed above; auxiliary equipment, into input and output devices, switches, and protective device; information-processing components, into controllers, supervisory devices, software modules, stored-program controllers, and process control systems with their components, such as central unit, local components, display and control components, and so forth. Note that a "pure" or textbook-type classification is not possible in many cases, and for practical reasons a compromise must be found, especially in the case of actuator systems. Here, continuous control devices are listed under the structural element "actuator system," except for solenoid valves and switching-type control valves, while switching devices for energy streams are classified exclusively as "auxiliary equipment." Naturally, a practical worker seeking a controller without auxiliary energy will look among the other controllers under "information-processing components," and if in need of a reducing station will look under "control air system." Under this object-oriented approach, once a certain device has been found, design specifications, device configuration and installation standards, and even manufacturing standards (if needed) can be found in a further level having the same structure and the same nomenclature of structural elements. After a device has been selected, designed, and configured, the respective activities can be carried out in CAE applications modules, such as process control station data sheet, process control station plan, procurement, and so forth.

Such a system has the further advantages that innovations in hardware and installation can be adopted better and more quickly and that workers in these fields obtain design and manufacturing standards in easily understood form. What is more, changes in standards and directives can be complied with in a better and more economical way (see Chap. 9).

Finally, two more developments should be mentioned:

- Workers in automation, information, and microsystems engineering are energetically pursuing:
  - Piezoelectric actuators [560]
  - Shape-memory actuators
  - Magnetoelastic actuators
  - Micromechanical actuators
  - Actuators using electrorheologic and magnetorheologic liquids
  - Precision positioning with electromagnetic actuators

Information on these developments can be found in [561].

- The trend in the chemical industry toward smaller-sized batch processes is generating a demand for much faster metering systems, which can only be realized by direct coupling between metering element and flowmeter (see Chap. 5) [562].
- In addition to a number of other advantages (energy savings, decrease shear gradient at almost closed position of the control valve – important for biotechnological and polymeric materials – smooth start-up, etc. [563]), by means of r.p.m controlled drives, pumps with variable performance can adopt the functions of classical control valves [564, 565].
- The coupling of sensor and actuator systems in self-sufficient control loops at the field level allows the amount of communication with and the computing load on the process control system to be decreased. Expecially in the area of building automation, interesting application are already available at reasonable cost.
- In the field of sensor/actuator systems in particular it will be necessary for developers and users to move away from the traditionally narrow viewpoint and, by means of analogy, apply insights from other areas to process control engineering [564].

6.3. Electrical Drives in the Chemical Industry

Development of Electric Drives. Electric drives have long been used in large quantities in various applications in industry. The majority of the drives are not speed controlled in any way. The first major application for variable speed drives was in traction, applications, especially in locomotives and trams, and later also in metallurgy. Variable speed drives subsequently found their way into a large number of applications in general industry. The general trend is from fixed-speed to variable-speed drives. Typical applications for modern variable speed drives are, for example, in the field of traction, the main drives for TGV and ICE trains, main propulsion drives for ships, and a large number of process-oriented applications in the pulp and paper, chemical, and metallurgical industries. Other typical application areas cover manufacturing industry, energy production, and water treatment.

Most motors in use today are not speed controlled. About 95 % of all motors sold today are asynchronous motors, of which 95 % run at fixed speed, connected directly to the mains. In future, the number of asynchronous motors driven by frequency converters is expected to grow rapidly, at ca. 15 %/a, while the growth rate for all variable speed drives is ca. 7 %/a.

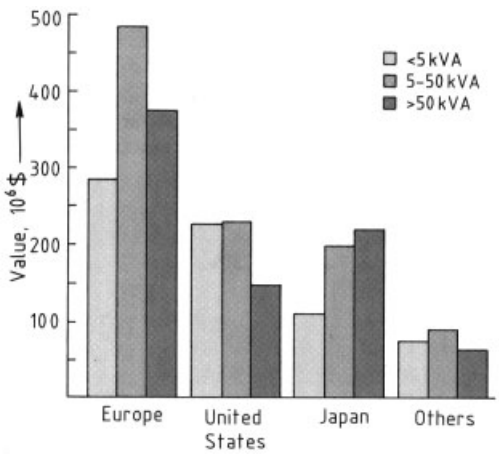

Figure 223. World market for variable-speed a.c. drives in 1991

The estimated world market for variable speed drives was ca. $ 3.3×10^9 in 1991, of which 55 % is a.c. and 45 % d.c., a trend that will intensify in the future. The a.c. market (Fig. 223) is concentrated to Europe, the United States, and Japan. The rest of the world accounts for ca. 10 % of the total market. The market split into the

various power ranges reflects the differing structures of the industry and the price of energy in the different areas. In terms of numbers, the majority of drives (ca. 90 %) are small (<5 kVA), ca. 8 % are in the range of 5 – 50 kVA, and only 2 % exceed 50 kVA. The number of motors controlled by variable speed drives in Europe is about 5 % for 5 kVA, 9 % in the range 5 – 50 kVA, and 16 % for >50 kVA. The larger the motor is, the more likely it will be driven by a frequency converter. Process demands at higher powers often require variable speed drives, and the realized energy savings are economically more important.

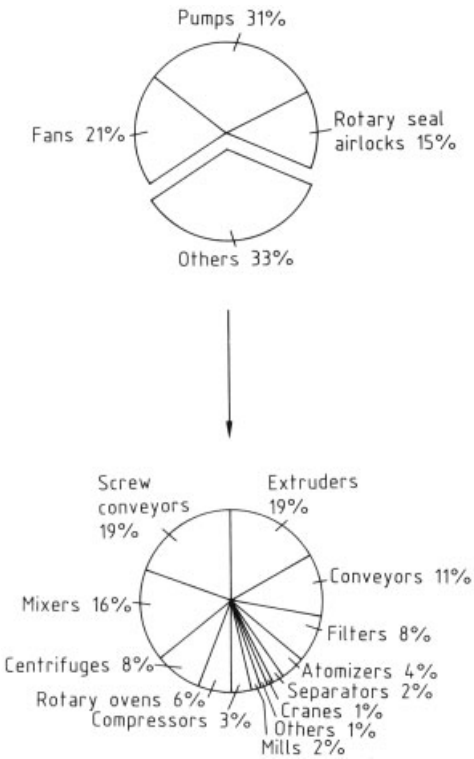

Figure 224. Areas of use of variable-speed drives in the German chemical industry

The basis for this development is the a.c. motor, which has a simple, robust, cost-effective, reliable, and rugged design that is relatively easy to modify for use in Ex-areas. Developments in power and control electronics have made drives more technically advanced and cost effective. The price of an a.c. drive has decreased by about 10 %/a during the last ten years, while the price for d.c. drives has shown a decrease of ca. 4 %/a. The total cost of a drive (drive electronics and motor) is about equal for a.c. and d.c. although a.c. can be slightly more expensive in the middle power ranges. The total life cycle costs for an a.c. drive are in general lower than those of a d.c. drive.

The German chemical industry has a large number of motors in use (Fig. 224), one estimate [561] is ca. 500 000 units. The average power is 7 – 13 kW, the minimum power is 120 W, and the maximum power is 6800 kW. About 80 % of all drives are smaller than 10 kW. Only about 10 % of all motors have variable speed; in two-thirds of the cases, this is achieved by pole changing. The large proportion of fixed speed motors indicates that other control methods than variable speed are used, usually control valves, bypass control, or mechanical variators.

Most of the drives in the chemical industry are used in the main process, although many drives are used in auxiliary processes in energy production and raw and wastewater treatment. The use of variable speed drives has increased markedly during the last few years. Most of the applications are pumps and fans, but several other applications are also quite common [566].

Variable speed drives are used mainly to improve process control [566]. By controlling the speed of the prime mover, the material flow can be better controlled; variable-speed pumps can handle logistics and the dosing of materials with low losses and low machine strain, thus improving process control and reducing mechanical wear. Variable speed drives are also used to replace control valves in polymer and biotechnical processes. The soft start feature of variable speed drives eliminates pressure shocks in pipes when motors are started and they also significantly reduce the thermal and mechanical strain on the motors at start, which usually permits more economic dimensioning of the total system. The supply dimensioning is eased by the lack of high starting currents and compensating capacitors. The direct drive energy savings, especially in centrifugal machines, are also of importance when deciding the type of drive required. A 40 % saving for variable-speed a.c. drives compared to fixed speed solutions is not uncommon. Variable speed drives also make less noise, as machines are used at optimum speed, rather than at full speed at all times.

Table 11. Main electrical data of various power semiconductors

	Thyristor	GTO	Bipolar transistor	IGBT	FET	MCT	FETh
Voltage, V	6 000	4 500	1 600	1 600	1 000	3 500	4 500
Current, A	5 000	4 000	1 200	400	10	50	50
Frequency, Hz	400	4 00	2 000	20 000	100 000	5 000	10 000
Power, kW	10^4	10^4	10^3	10^3	10^1	10^2	10^2
Chip size, cm^2	80	80	> 10	> 10	5	2	5
Year	1960	1977	1985	1991	1980/90	1993	1995
Processor/		8080/85	8086/186	8096/486		80586	
Memory	256 bit	4 kbit	64 kbit	1 Mbit	1 Mbit	16 Mbit	64 Mbit

Development of A.C. Drive Technology. *Power Semiconductors.* The development of power semiconductors is summarized in Table 11. The first commercially successful power semiconductor was the thyristor, developed in the early 1960s. The thyristor is a four-layer device that can be turned on (changed into a conducting state from a nonconducting one) by the application of a gate pulse. Because it works by positive feedback, to turn it off, the current through it must be decreased to zero, after which it regains its blocking, nonconducting state. If the applied voltage across the thyristor does not naturally reverse sign (as happens for components connected to the mains supply), a commutation (turn-off) circuit is required. This increases the size and the complexity of a thyristor drive and also limits the maximum switching frequency achievable. Thyristor PWM drives normally switch at ca. 400 – 600 Hz and the motor is thus fairly loud with a motor loadability of ca. 85 % at 50 Hz. The dU/dt is ca. 500 V/s. Thyristors are now used only in high power drives.

The next major power semiconductor is the GTO (gate turn-off thyristor). As the name implies, it is a thyristor that can be turned off as well as on. This is achieved by careful design of the semiconductor layers and device geometry. As the device can be turned off, no additional commutation circuitry is needed and the size of the drive thus decreases. The GTO is fairly sensitive to a correct handling at turn-off and is today used in high-power drives (200 kW) and at high voltages (>1 kV). An advantage of GTOs is that they are latching devices, like thyristors, and thus require no control power when they are in a stable state (on or off). In the case of short-circuits in the output, they pose problems, as they do not inherently limit the current and are difficult to turn off in this state. The switching frequency and dU/dt are comparable to thyristors.

In lower power drives today, the main power component used is the bipolar transistor. With this component powers of up to 200 kW can be achieved, and up to 600 kW with several transistors in parallel. The device consists of two Darlington-connected transistors. By controlling its base drive the device can be made to turn on and off. Transistors require a continuous base drive in order to conduct, i.e., they do not latch. In the case of shorts they are more forgiving than GTOs but they require a stiff base drive to turn them off. The simpler turn-off requirements simplify the control circuitry and allows drives to be smaller and more efficient. With Darlingtons, switching frequencies of up to 4 kHz are achieved with a dU/dt of ca. 1 – 2 kV/s. The higher switching frequency increases the motor loadability to ca. 90 % at 50 Hz and reduces motor noise.

In newer ranges of frequency converters, the power semiconductors are IGBTs (insulated gate base transistors). These devices are a combination of voltage-controlled field effect transistors (FETs) and bipolar transistors, connected so that the control is like the voltage control of a FET and the power side is of the low-drop bipolar type. These components inherently limit the current and they are thus more forgiving and easier to handle in short-circuit situations than transistors. As they are voltage controlled, not current controlled like bipolar Darlingtons, the control circuit is simplified and reduced in volume. There is also a significant reduction in the rating of the auxiliary power supply, as IGBTs require less control power than transistors. The IGBT allows switching frequencies of up to 20 kHz which increases the motor loadability to 100 % at 50 Hz and gives an attainable dU/dt of 5 kV/s.

New components are being developed by power semiconductor manufacturers, for example, MCTs (MOS controlled thyristor) and FETh

(field effect controlled thyristor), hybrids between conventional thyristors and FETs, with voltage-controlled switching and a very low state loss, which will reduce the size of future drives.

Frequency Converters. Frequency converters are divided into two main types, depending on their construction. One is the current source inverter (CSI) and the other the voltage source (PWM). Current source inverters contain a controlled thyristor bridge connected to the supply and a second one connected to the motor. The two bridges are connected by a large inductance. The current to the motor is controlled by the input bridge, and the frequency is controlled by the output bridge. The motor voltage then adjusts itself to the actual current and frequency. In older types, the output bridge consists of thyristors, which means that the motor is a part of the commutating (turn-off) circuit of the thyristors and thus every inverter and motor forms a pair. In newer drives the thyristors are replaced by GTOs or bipolar transistors, which eliminates this coupling. Due to the basic construction with a controlled input bridge, CSIs are suitable for single, stand-alone four-quadrant operation. As they contain a large choke, however, they are usually bulky and the drive may be noisy.

The most common type of drive today is the voltage source inverter. It consists of a diode input rectifier, which generates a constant intermediate circuit voltage. This intermediate circuit consists of a choke and a large capacitor. At lower powers the choke is often omitted. This causes the harmonic content of the supply current in the main to increase, so for installations requiring smooth mains current drives with an intermediate circuit, chokes are recommended.

The rectified and smoothed voltage is supplied to the motor via an output stage, which controls the voltage and the frequency of the waveform applied to the motor. The most commonly used components in this stage today are bipolar transistors at lower powers and GTOs at higher powers. The bipolar transistors are being replaced by IGBTs. The waveform applied to the motor consists of a pulse-train with a sine-weighted duty cycle (pulse-width modulated, PWM). The inductance of the motor filters the current resulting from this pulse-train to near-sinusoidal shape. The effective voltage applied is normally increased in direct proportion to the output frequency to generate a constant flux in the motor. This is possible until the mains voltage is reached (i.e., up to 50 Hz) and with the latest modulation techniques up to about 60 Hz. At higher frequencies the voltage remains constant, and the flux diminishes in inversely proportion to the output frequency. The maximum speeds that can be reached with standard squirrel cage motors are of the order of 40 – 120 Hz; higher speeds require special motors. The normal PWM inverter described above is used in about 90 % of all cases. It is a simple, straight-forward approach to variable speed control, requiring no special motors or any sort of feedback. As it is an open loop drive, its dynamics and precision are not of the highest degree. It is, however, sufficient for pumps, fans, extruders, conveyors, and simple machines.

For higher performance, vector-controlled drives are used. In these drives, a feedback signal is required from the rotor position. This information enables the controller to precisely calculate the torque on the motor shaft and also indicates the exact position of the shaft. The information is then fed to a PWM output stage, which controls the current and frequency in the desired manner. With this drive, full load torque at zero speeds and precise position control can be achieved. The drive always requires an encoder on the motor. Typical applications are synchronized lines, paper-machine drives, rolling-mill drives, and all applications where the torque must be precisely controlled.

Power semiconductor development has had a major impact on the size of drives. The average volume per power has decreased to 1/6 in ten years. Today the decisive factor determining the size of a drive is its thermal design. Modern power semiconductors allow a junction temperature of up to 150 °C, corresponding to a case temperature of ca. 100 °C. The case temperature is limited to about 80 °C, sometimes up to 100 °C.

Modern control electronics use 16 bitprocessors with designs for 32 bit processors or for signal processors in development. This trend of incorporating more computational power in the drive will continue, so that the possibilities for control offered by modern drives will increase. The requisite discrete logic is commonly realized in an ASIC (application specific IC). These are monolithic ICs built us-

ing standardized logic cells, so that a defined, application-specific function can be realized. One ASIC can replace several hundred conventional TTL circuits with large savings in space and powering. Lately, technologies have become available that allow a similar approach to analog circuit design.

These developments have also increased the reliability of drives. The MTBF (mean time between failure) for drives in 1983 was ca. 40 000 h, while in 1993 it was ca. 250 000 h. The number of parts necessary to build an 11 kW frequency converter has decreased from 2500 to 500, a trend that will surely continue. The basis for this lies in modern component technology, both for logic components and for power components, and in the better possibilities for simulating and measuring complicated and fast processes.

For critical applications, redundant drives are still recommended, especially in cases where there is a risk of injury to humans, where the process flow is critical and must be controlled at all times.

An aid to achieving high process reliability is the use of switchgear converters, which allow easy changing of drives and of system configuration. Combined with a fully digital commissioning and down-loading of various drive parameters, this allows an extremely rapid change of drives, simply by changing a cassette.

An external sign of quality control and an internal tool is a certification according to ISO 9000, Parts 1 – 3. This standard defines the requirements regarding standardized working methods and documentation for a company or profit center. Quality is also achieved by standardization of methods in development, production, and testing. Standardization allows a high repetition rate when using various components, either software or hardware. This allows a short development time for new products without endangering product quality.

As an example of standardization, consider solving EMC problems (electromagnetic compatibility), that is the problems occurring when devices have to work in an electrically noisy environment. The main standards in use are the various DIN/VDE and CISPR standards. Generally speaking, two types of interference occur:

- Low frequency: frequencies up to ca. 5 kHz and earth currents
- High frequency: frequencies from ca. 50 kHz to 30 MHz

Low-frequency interference can be divided into periodic interference (caused by nonlinear components such as diodes and thyristors) and aperiodic interference (caused by supply disturbances). High-frequency disturbance is caused by switching phenomena in power switches. The various interferences can influence other devices nearby, either through radiation or conductance. Generally, conducted interference is more problematical.

A standard solution to high- and low-frequency interference is adding filters to the input of the device. At high powers this is usually difficult, bulky, and expensive, especially as the high- and low-frequency ranges require different features. It is also possible to separate the power supply to interfering devices from all other supplies, or to use a d.c. distribution system, in which 12-pulse circuits for supplying the d.c. power distribution minimizes the 5th and 7th harmonic current in the supply. This d.c. power is then supplied to independent output stages in order to drive motors.

Aperiodic disturbances include supply failures, which can last from milliseconds to several seconds. Drives usually use the kinetic energy in the driven load as a supply when the supply fails and then accelerate to the set speed when the supply returns. The possible time the drive can survive depends on the type of load and its kinetic energy. Usually there are also other questions of synchronism: can the drives run when the total process is otherwise down?

To eliminate undesirable influences from the power stage on the control electronics, fiber optics are used. This allows a clean separation of the control circuits, which can be standardized irrespective of rated power, and the power stage, which can be separately optimized and standardized. This technique also allows customized interfaces to be developed for standardized, mass-produced drives.

In future, it will be possible to eliminate a significant number of sources of high- and low-frequency interference by using so-called resonance circuits. For example, with single-phase inputs, a completely sinusoidal supply current can be achieved. A further development allows

switching of the power semiconductors in a zero-power state, which reduces disturbances.

Drives as System Elements. A drive is regarded as an actuator from the systems engineering standpoint. It is connected directly to the process and to a field bus. Normally, only commands and status information are exchanged through the bus system, but in special cases the entire commissioning and the parameter setting must be performed associated through it. This can be done either from a remote panel connected to the field bus or from the engineering workstation functioning as bus master. The goal is to move the detailed process control to the sensors and actuators exchanging information through the field bus. To reach this goal intensive work is being done to standardize the field-bus protocols.

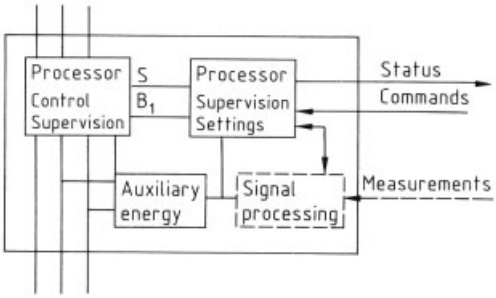

Figure 225. Information structure of a modern drive

A modern a.c. drive (Fig. 225) is functionally divided into the actual motor control and the interface with the external world, with control and setting possibilities. This concept uses at least two microprocessors (one for each function). The drive is self-contained, and no external power is necessary. It is in many cases possible to process data locally to control the drive. Such a drive will form a yield unit (Fig. 226), decreasing the load on the process controllers and on the control system. This is especially important in fast processes (for example, in dosing applications), so that the central control is not burdened with all details of the process. The drive is then closely connected to the process through local measurement and feedback, making it possible to implement local control of, for example, volume and heat flow. The bus system transmits only commands and status information from the various field units; details are handled locally.

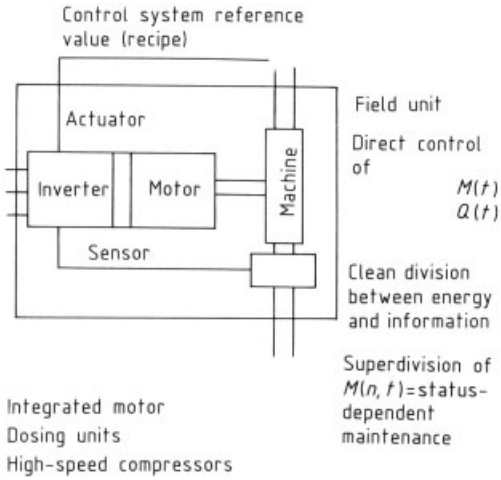

Figure 226. Field unit

The planning of a drive system is identical to the general planning of control systems. The basis for the planning is a requirement specification drawn up by the user, describing the function of the drive and its behavior in special cases such as mains failures. This must be carefully thought through and defined, as experience shows that most of the subsequent problems can be avoided here. Based on the requirement, an engineering specification is written, describing with which devices and with which technologies the desired functions will be realized. Here, the economic restrictions of the various projects must be taken into account. This work is usually facilitated by standardization within a plant or company. It is also recommended that a test specification be written for the possible delivery tests, preferably in cooperation with the chosen supplier.

When dimensioning and specifying a drive, the following features must be taken into account: voltage, power, torque requirement as a function of speed, speed range, control signals, braking capability, cabling recommendations, and EMC aspects. The behavior in special cases (short circuits, overloads, underloads, earth faults, etc.) must also be considered. The desired motor features must also be defined (construction, EX class, loadability, etc.).

Drives are considered actuators in modern CAE systems and they are treated as standardized objects with certain, defined features [567].

When a drive has been specified and built, an acceptance or delivery test is often required,

especially in the case of high-power drives. The goal of a delivery test is to establish that the drive fulfils all the requirements defined in the requirement specification. It is important to specify in detail how the various features must be measured to ensure that the specifications are fulfilled. At lower powers, a type test certificate is usually sufficient. The supplier must be able to test the behavior of the drive at high break-away torques or at mains failures. A test proves that the drive works as specified and allows a first, coarse setting of parameters to be done. It is usually not possible to optimize the drive; this must be done under true working conditions on site.

This optimization is made easier by digital technology, which also allows data transfer and trouble shooting over telephone lines. A multichannel recorder integrated into the drive also facilitates trouble shooting. For trend analysis, the drive itself stores and displays certain key data as a function of time. This can be done by a process control system, but it is simpler if the drive makes it itself and transmits only the results to the control system. The analysis of the trends makes it easier to judge when maintenance is needed, and avoids being bound to a fixed time interval. One example of this is the calculation of when bearings need to be lubricated. In this case, the drive calculates the lubrication interval and indicates when it is necessary to lubricate the bearings. This time depends on the speed at which the motor is running and can be appreciably longer than the time calculated from the nominal speed of the motor.

Trends. Two directions can be distinguished in the development of drives. On the one hand motors, and on the other hand power electronics and information processing electronics are continually being developed. In addition, a development in the mechanical construction of electric drives can be traced.

New motor ranges better suited to the use of converters are being developed. Standard motors are designed for mains use, as 95 % of motors are connected directly to the mains. Use with a converter makes different demands on the motor. FEM methods help when new, converter optimized designs are being simulated and built. They also aid in developing new high-speed motor concepts by optimizing the rotor slots for inverter use. The new slot shape reduces the losses for a 100 kW motor from 4 kW to 0.8 kW. This motor is designed for a compressor drive with a speed of 10 000 rpm.

The main problems of high-speed drives are mechanical and concern bearings, lubrication, and life time. Conventional lubrication is sufficient up to about 16 000 rpm above which it is necessary to use oil fog lubrication, which increases cost.

The high speeds also allow better material utilization. The high speed motor allows a higher power for the same motor size, or a smaller motor at the same power rating. In high-speed applications, gears are completely eliminated leading to lower investments, less maintenance, and better total efficiency.

The development of new magnets makes it possible to manufacture motors with permanent magnet rotos. The materials development has been rapid, making it possible to obtain realistic performance at a reasonable cost. The cost for neodynium magnets is today about 300 $/kg, allowing the development and sale of special motors. These motors have high efficiency, a power factor equal to unity, and a simple positioning drive without feedback can be realized, as the motor has no slip. Powers up to 100 kW have been reached under laboratory conditions. Research in this sector is concentrated on production methods, as it is difficult to work permanent magnets.

A motor of this type is also well suited to integral motor applications. In these motors, the motor and the drive are mechanically and thermally integrated, making it possible to talk about integrated field units consisting of motor, power electronics, and information processing electronics.

In the inverter, several areas are being developed. The development of microprocessors for the consumption and automotive industries makes powerful and cost efficient microprocessors available for use in drives as well. Semiconductor manufacturing technology also allows the design and manufacture of large ASIC circuits to replace discrete logic. The high processing power available makes it possible to realize complex algorithms, for example, the so-called flux control of a.c. motors. The magnetic vectors in the motor are calculated based on the position of the switches in the output power stage and on measurements of the currents and voltages.

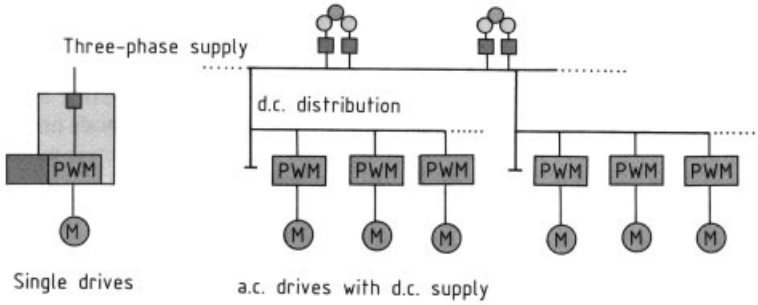

Figure 227. d.c. distribution

This requires a very powerful processor if it is to be performed in real time. This control method gives a standard a.c. motor dynamics superior to those of a d.c. motor.

The powerful processors make it possible to use adaptive control in motor control as well as in a possible controller built into a field unit. It is possible to optimize the motor control as a function of load. These adaptive methods make commissioning easier and are also a step in making drives more user-friendly.

Other trends are control algorithms that take into account the mechanical features of the shaft between motor and the driven machine. With an elastic or possibly an electroviscous clutch between motor and machine it is possible to reduce the impact of torque shocks in the load on the motor, but control of the load is more difficult. An algorithm of this type allows the use of smaller motors. Applications can be found in rolling mills and in long and heavy conveyor drives.

To reduce the size of the inverter and to modularize it, several approaches are being pursued. One is the so-called matrix inverter [569], which connects a three-phase motor directly to a three-phase supply and allows the motor speed to be controlled. The motor voltage does not correspond to the full mains voltage and thus a special motor is needed. One difficulty with this concept is that switches able to conduct and block in both directions are needed, and such components are not yet available in monolithic form. Semiconductor manufacturers are developing these components. As the matrix inverter does not contain any bulky components, intermediate circuit chokes, or and capacitors, it is a good candidate for integrated solutions.

Another solution for standardizing drives is d.c. distribution (Fig. 227) [568]. Instead of having one rectifier bridge, choke, and capacitor bank in each drive, the separate drives are supplied with a d.c. voltage from a common source. This larger unit contains one rectifier, choke, and capacitor bank. It can be smaller than the combined power of the individual units, as it is often possible to use the regenerative power from braking drives to supply driving ones. Typically, considerable savings are achieved, for example, in centrifuge applications, where by braking one centrifuge the energy can be used to accelerate others. The mains rectifier must only supply the energy lost. At higher powers it is also possible to use multiple-pulse connections to reduce the strain on the supply. Galvanic, periodic, and aperiodic disturbances are also easier to control.

A further step in modularization is the combination of power switches and their controlling circuits in one package. Modules are today available containing power switches and their drivers and protection circuits [569]. Such modules reduce the size of the inverter and its reliability increases as the number of separate components decreases.

This trend towards higher integration has also reached the chip level. Chips will soon be available with several functions integrated. This is called SMART power. Further in the future lies chips integrating power switches and total control of the drive. There are no theoretical difficulties in integrating the matrix inverter onto one chip but many practical difficulties. The limits of what can be done today lie at one power switch and its immediate drivers and protection circuits at a voltage of 600 V. These chips are very expensive, as volumes are small, the production process complex, and yields low.

The developments in the fields of power electronics, information electronics, silicon technology, and mechanics will combine in the form of integrated devices, where most of the necessary functions are integrated. The devices will integrate all necessary sensors, actuators, controls, and protections in a single mechanical unit.

The energy savings made possible by using variable speed drives, not only in the drive itself but through the total energy supply chain can be quite substantial. When a control valve is replaced by a variable speed drive, the efficiency of the total system is significantly increased, especially at low volume flows. The total efficiency can increase from 11.5 % to 23 %, an increase of 100 %. These savings reduce environmental pollution and require smaller investments, as the available energy is used more efficiently.

6.4. Electric Power Supply Systems

Principles. A useful way to visualize the structure of electric power supply systems in industrial plants is provided by the level model (Fig. 228, see next page). This scheme assigns a central role to the process power supply system, which connects actuators and sensors on the "consumer" level with power plants via medium-voltage distribution units on the "transport" level.

The process power supply system is tailored to the needs of each plant and must satisfy requirements under the following three headings [570]:

- Adequate capacity under normal conditions. The specific requirements are as follows:
  - Design for operation under rated conditions
  - Availability of electric power at any desired location
  - Flexible delivery of electric power under various service conditions
  - Voltages suitable for various consumer interfaces
- Availability requirements fall into three groups, each with its own set of objectives:
  - Availability from the safety standpoint (occupational safety and health, environmental protection)
  - Availability from the process engineering standpoint (product quality, safety of equipment)
  - Availability from the logistics standpoint (capacity of vendors, production costs)

  In engineering terms, this means:
  - Limited effects on the process when power disruptions occur
  - No interference of one electrical consumer with another
  - Limited effects on the power supply system in case of operation upsets (e.g., fires).
- Safety of electrical equipment. The specific requirements are as follows:
  - No danger to personnel
  - No environmental hazards

These three sets of requirements completely describe the profile of requirements for the process power supply system.

This section describes the essential concepts employed to meet the stated requirements [570, 571]. By way of illustration, the process power supply system for a wastewater treatment plant is described in [572].

Implementation of Normal Power Supply. The basis for supplying electric power for the nominal operation of a plant is established in the design phase. Here, the three-phase switchgear and control gear is dimensioned in view of the predicted requirements of consumers; utilization and demand coefficients are employed for this purpose. The numbers and sizes of power transformers needed are also determined at this point. Direct-current power supply systems are designed by a similar method, with the closed-circuit current and estimated component peak loads as starting points.

Under normal operating conditions, the entire power supply system of a plant is secured by splitting up the power demand among a number of sources, using in-plant generating capacity and imported power.

The relationship between in-plant and imported power depends on a number of factors, including the size of the plant and seasonal variations in steam demand.

In the chemical industry, the principle of cogeneration (combined heat and power generation) is used in creating the optimal power generation system.

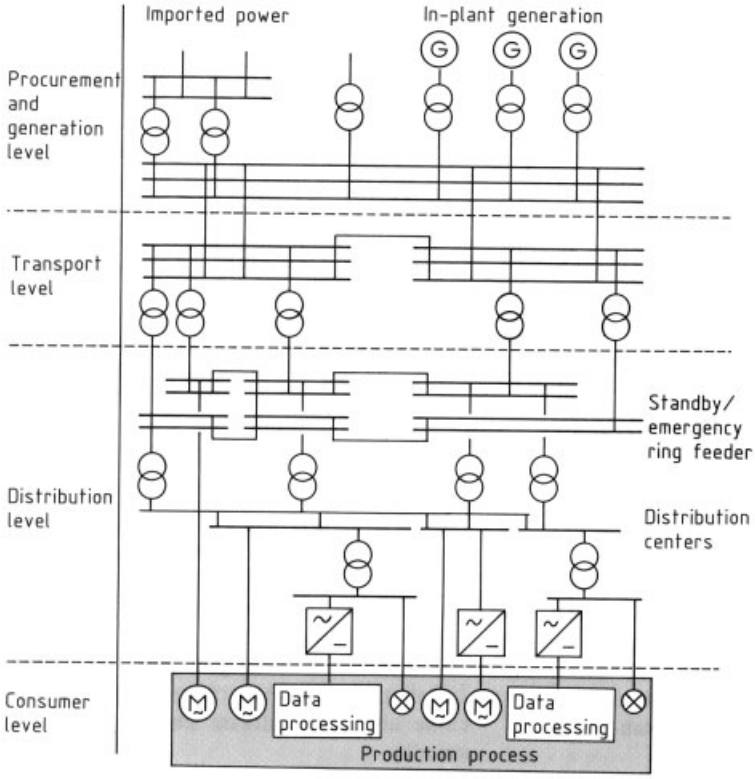

Figure 228. Level model of power supply

The fact that power to run the plant comes from two sources does not affect the design of electrical equipment inside the plant. The production process is embedded in a complex, two-dimensional network of cables that covers the entire battery area, connects supply to consumpxtion, protects the process systems, and insures any redundancy that may be required.

Cable networks in industrial plants may be radial or meshed. The radial design is more important today. The meshed network, which offers the advantage of two-way supply, is declining in importance, in part because problems can arise in dealing with the short-circuit power when existing facilities are expanded.

For special cases, distributed generation with diesel-generator sets or battery-backed converters serves to deliver reliable power. This approach should be used only when the solution for the normal case is not technically feasible or when the economics are very unfavorable.

Direct current systems are generally designed in distributed form within the plant unit or section. In future, the increasing number of variable-speed drives may also lead to other structures, such as d.c. networks.

Cable and installation costs can be controlled if the location of switch rooms and battery rooms in the central plant buildings is tailored to the process and/or such facilities are distributed in accordance with demand concentration.

Modern switchgear technology can react quickly to service changes by virtue of

- Modular switchgear design
- Busbar distributors, which are especially suitable for laboratories and workshops
- Worksite distribution boards for installation jobs

Process power supply systems in normal service are designed for standard voltages. Commonly used voltages are as follows:

Data acquisition and processing	24 V d.c.
Lighting	230/400 V a.c.
Wall receptacles	230/400 V a.c.
Low-voltage, low-power motors	230/400 V a.c.

Low-voltage motors, general	500/690 V a.c.
High-voltage motors	5 kV, 6 kV, or 10 kV a.c.

Special situations may require special voltages, such as 25 kV or 30 kV for electrolysis processes.

Implementation of High Availability (Conceptual Solution). Before the adoption of measures to enhance process power supply systems, two points must be considered:

- The availability of power supply systems in general is normally higher than 99 %. Outage time per year, as a rule, is in the range of minutes. This holds for every point in the network and for each consumer.
- The cost of upgrading, with the aim of approaching even closer to 100 % availability, increases with rising power level, increasing number of hierarchical levels in the system, and decreasing outage length. It is not, however, possible to achieve 100 % availability.

The profile of requirements for the process power supply system, as set forth in the bidding specifications, is met by standard and special practices.

Standard practices include the following:

- Use of network architecture typical of industrial plants, with multiple pathways at the supply and generation levels and complex network structure at the transport and distribution level (Fig. 228)
- Use of the IT network
- Use of modular switchgear
- Installation of A – B machinery, with one machine always on stream and the other on standby

Special practices include:

- Inclusion of a standby power supply system
- Upgrading of d.c. power supply system to a protected d.c. power supply system
- Splitting up both switchgear and cable installations into two subsystems
- Compliance with interference immunity requirements

In that follows, practices of a special nature are discussed in more depth; among this group of practices, only those involved in standby power supply and protected d.c. power supply are covered.

Two types of standby power are important for process power supply systems in industrial plants (Fig. 229):

- Standby power from a second, independent network
- Standby power from a standby generating system (diversity of power sources)

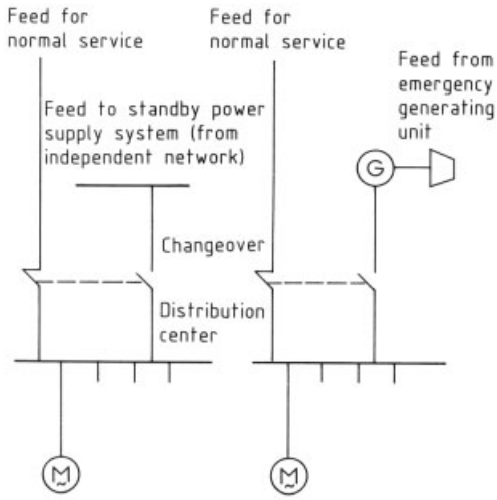

Figure 229. Standby power supply system

Standby power supply from a second, independent network is implemented by means of a reserve, backup, standby, or emergency ring feeder (Fig. 228), continuously supplied via automatic voltage detectors.

In general, when standby power supply systems are employed, the bidding specifications must list consumers by permissible outage time and assign them to classes of low-voltage switchgear. Figure 230 shows a classifiction that has proved acceptable as a first rough scheme.

Figures 231 and 232 identify types of low-voltage switchgear and the classes to which they are assigned.

Class 1 consumers must be further broken down for in-depth analysis.

Protected d.c. power supply systems (Fig. 233) feature 100 % reserve capacity because of

- The dual feeder arrangement
- Additional backup of each feeder in case of line outage

Classes of low-voltage switchgear	Possible outage times to consumers on failure of normal power supply	Details of low-voltage switchgear	
		Number of feeders	Devices actuated on failure of normal power supply
Class 1	0–15 s	2	Backup system continuously on line and ready
Class 2	>15 s		Backup system automatically switched on line
Class 3	At least until restoration of normal power	1	

Figure 230. Classes of low-voltage switchgear, possible outage times, and details of implementation

This redundancy must be extended to the data transmission and processing system, that is, to the cards in the subracks if possible.

Implementation of Safety. The terms safety, protection, and danger are interlinked by the concept of risk. Safety practices are intended to insure that the limiting risk is not exceeded. Such practices are mandated by legislation, regulation, engineering codes, and plant-specific direc-tives issued by operators.

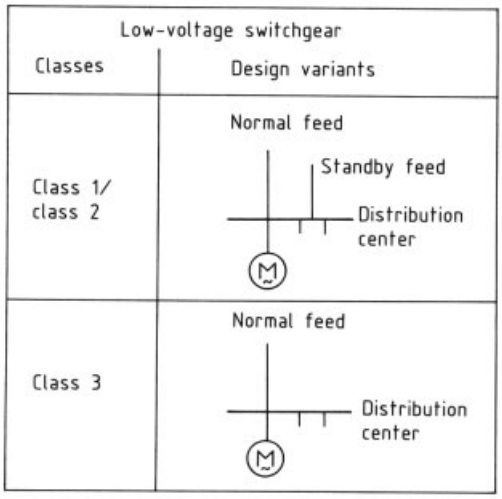

Figure 231. Classes and variants of low-voltage switchgear

In the case of power supply equipment or power supply systems in general, "safety" refers to two independent objectives:

- Protection against electric shock (personnel protection).
- Protection against failure of the electric power supply system, which is ultimately a question of "reliability" (safety of equipment, protection of product and plant). This

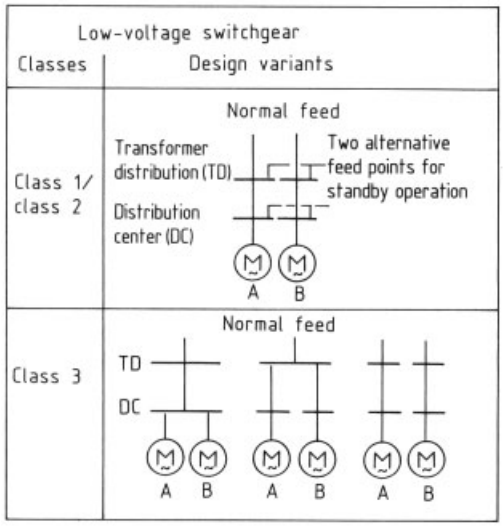

Figure 232. Classes and design variants of low-voltage switchgear in plant with A and B machines

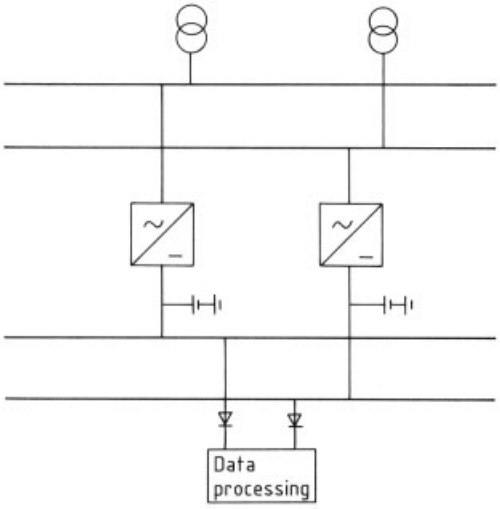

Figure 233. Protected d.c. power supply system

kind of protection is achieved by means of the measures employed for high availability.

Apart from functionality, protection against electric shock has been the second dominant theme in power supplies since the start of electrical engineering [573]. Notable among the many standards now in existence is the extensive VDE code, which is continually updated.

The practices listed below serve as examples of how personnel protection is achieved in process power supply systems for the chemical industry:

- Increased use of 30 mA ground-fault circuit interrupters. This switch cuts off the flow of fault currents that could cause fibrillation.
- Installation of devices safe against touch by the hand.
- Application of the accident prevention specification VBG 4, "Electrical Installations and Equipment."
- Wider use of safety switches on drives.
- Application of power supply options originally intended for escape route lighting to overall production power supply systems (Fig. 234) [574].

Environmental considerations apply chiefly to process engineering. In general, these requirements do no impact on the process power supply system, which comes into the picture only when "reliable" power is required for devices affording protection or limiting damage to process control systems [575]. It must be verified whether a standby power supply system or a proteted d.c. system is adequate. One type of such a "secure" power supply system is a high-availability power supply system.

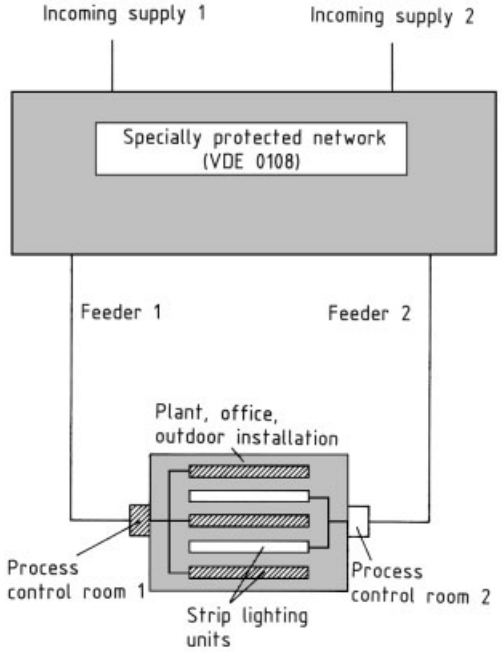

Figure 234. Lighting system fed from the specially protected network

High-Availability Process Power Supply System for a Wastewater Treatment Plant (Example). If a plant unit does not have its own waste disposal facilities, a central plant with biological treatment stages is of fundamental importance for maintaining production in a chemical plant. Such a treatment plant requires a high-availability power supply system, which must comprise all the levels of the power supply scheme and extend all the way down to the process control components in the field as well as control rooms [572].

The quantity structure of the plant as built is as follows:

- Distributed process control with dual host computer system
- 1800 process control stations
- 530 uncontrolled 3-phase drives with total power consumption of 10 MW
- 70 controlled electrical drives with total power consumption of 3.5 MW

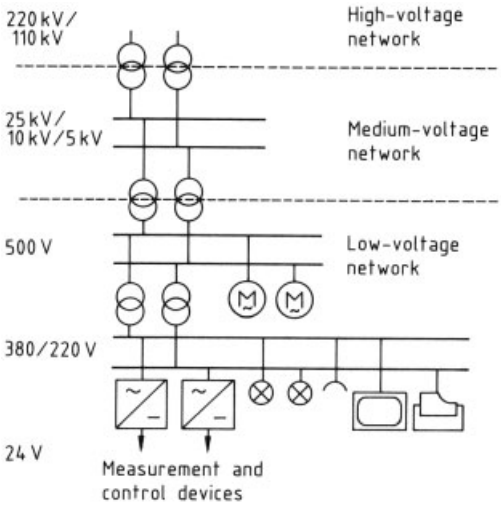

Figure 235. Voltage levels in a chemical plant

The following voltages are available (Fig. 235):

High voltage: 220 and 110 kV
Medium voltage: 25, 10, and 5 kV
Low voltage: 500 V for motors;
380/220 V for lighting, wall receptacles, computers, and hand tools; 24 V for measurement and control apparatus

Comparison with the level model (Fig. 228) shows the relationships to be as follows: The high-voltage network corresponds to the import and generation level; the medium-voltage network, to the transport level; the low-voltage network, to the distribution level.

The dual-bus structure illustrated in Figure 235, with the consumers separated into two subsystems, insures that the failure of one unit (e.g., a transformer or feeder) affects only some of the consumers at levels lower in the hierarchy. The failure of both system would, however, lead to shutdown of all consumers. However, this must be avoided for particularly important consumers. In such cases, further standby power supply practices come into play:

- Use of diesel generators
- Use of battery-backed uninterruptible power supplies
- Use of protected d.c. power supplies

Figures 236, 237, 238 show these backup arrangements for the low-voltage (500, 220, and 24 V) systems in the wastewater treatment plant.

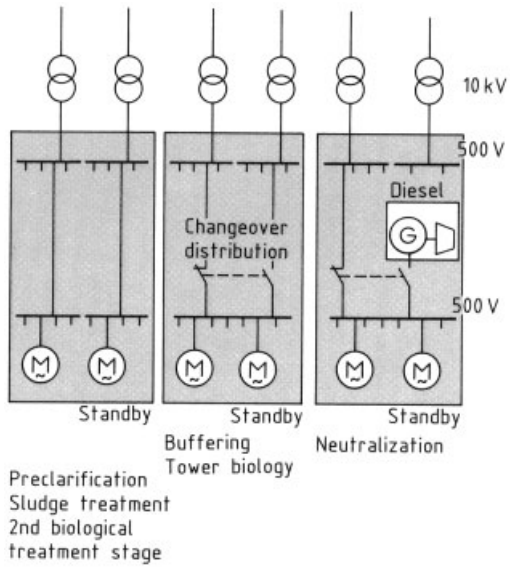

Figure 236. 500 V low-voltage supply system in a wastewater treatment plant

7. The Process Control System and Its Elements: Distributed Control Systems

7.1. Principles

The task of the process instrumentation and control hardware is to assist plant management in the control of its plants and in the control and monitoring of its production processes [576].

This chapter discusses the structure and function of process instrumentation and control hardware. The structure is discussed on the basis of a reference model.

In a general sense, process instrumentation and control hardware comprise all devices between the sensors and actuators and the operator controlling the process (see Fig. 239 and 363).

The process instrumentation and control hardware must perform a number of distinct functions. Figure 240, for example, shows a metering device. The required functions can be classified as follows:

- Controlling the process
  Automation functions that permit the self-contained control of plant units and process steps. Example:

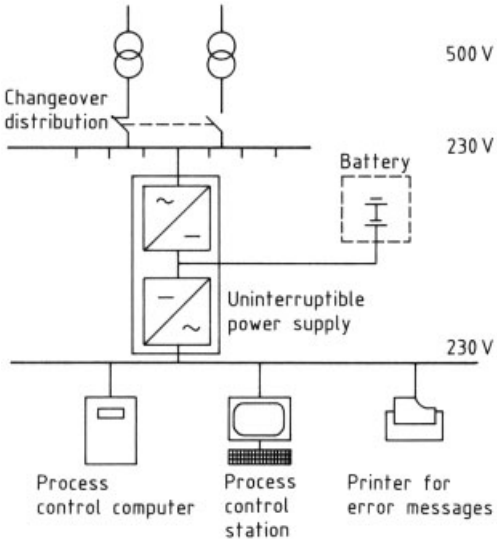

Figure 237. 230 V low-voltage supply system in a wastewater treatment plant

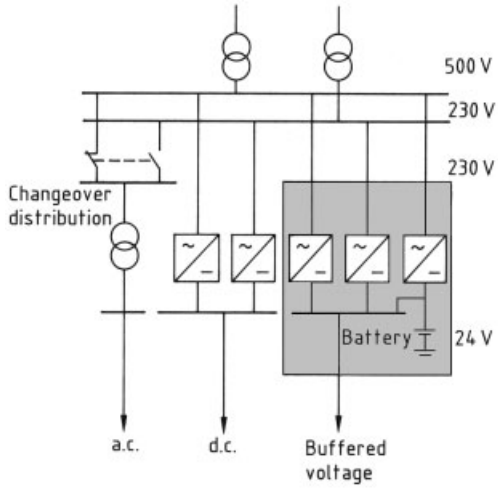

Figure 238. 24 V low-voltage supply system in a wastewater treatment plant

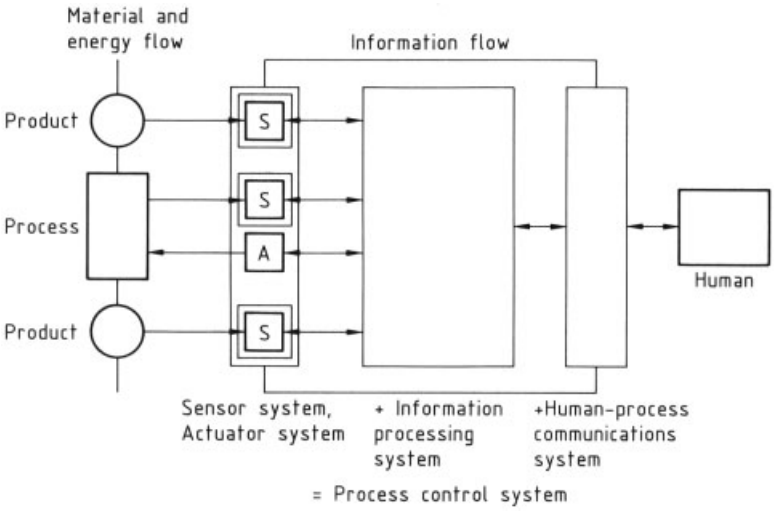

Figure 239. Process monitoring and control system

If action "metering on," then meter from B 1 to B 2 x liters at q L/s.

– Enhancing the safety and reliability of the process
  Functions prior in time, independent of control function specifications, that intervene in the process in order to enhance its safety and reliability. Because of high required availabilities and special performance provisions, these functions are often implemented in a separate "protective level." Example:
  If L 2 > L 2 max, pump K 1 is disabled.
– Providing displays
  Display functions that report plant and process status to the operator in a concentrated and easily understood form. Example:
  Display level L 1.
– Generating messages
  Message and alarm functions that explicitly notify the operator of certain events and provide information to the central message and alarm processing system. Example:

If L 2 exceeds L 2 max, transmit a message. (Class: "Alarm"; text: "L 2 max alarm")
- Archiving
  Archiving functions to store records of events and status logs. Examples:
  Archive all reports.
  Archive level L 1 every 10 s.
- Generating and interpreting reports Recording and evaluating (interpreting) functions that retrieve, evaluate, and print out archival and current data when a certain condition or command occurs. Example:
  Log the last metering operation performed (starting time, ending time, quantity, events).

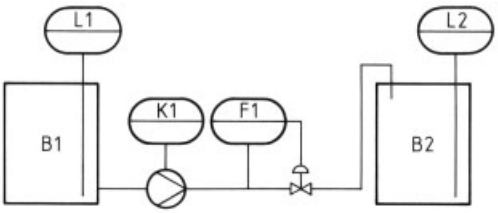

Figure 240. Example: metering system

If all functions to be performed by the instrumentation and control hardware are arranged in accord with the functionally abstract "level" model of the process control functions (see Section 2.2 and Fig. 17), some basic conclusions can be drawn as to the automation of this system.

The lower levels of the hierarchy contain many parallel functions with high intervention frequencies. These functions are chiefly suitable for automatic processing by process control devices. At higher levels, the intervention frequency generally decreases, but certain functions (e.g., load recipes) require the transfer of large amounts of data.

The automation of these functions generally proceeds from bottom to top, in keeping with the hierarchical breakdown of process control functions.

Different plant sections can differ in degree of automation. The degree of automation can be divided into four levels [577] (Table 14). Each degree includes not only the automation of the respective process control functions but also the automation of the associated safety/reliability functions. A safety- or reliability-enhancing function is assigned to the level at which it intervenes in the instruction branch. A safety function "if pump temperature > permissible maximum → turn pump off" is assigned to the individual function level; a safety function "if coolant temperature > permissible maximum → turn column off" is assigned to the plant section level.

The functions that a specific process instrumentation and control system must perform can be set down in a requirements list.

In principle, a project specification consists of a list of such required individual functions. Some of these functions, as shown in Figure 241 [578], are carried out by the plant operator, the rest by the process control system. The term "automatic" refers to a function that is performed by the process control system without external action. The degree of automation is the fraction of all functions that can be represented by automatic functions. Nevertheless, the overall scope of the control task must be considered regardless of the degree of automation. Each function that does not take place automatically must be performed by the plant operator. By changing the operating mode, the instantaneous degree of automation can be made lower than the maximum degree of plant automation provided for in the design (see Fig. 242).

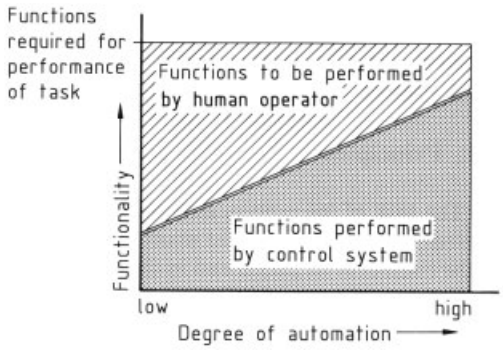

Figure 241. Degree of automation

When examining the properties of a process monitoring and control system, a distinction must be made between the external functional viewpoint and the internal structural viewpoint. The external functional view regards the finally configured process instrumentation and control system as a "black box" having certain functional qualities. The internal structural viewpoint regards how the system is structured and how the individual functions are implemented.

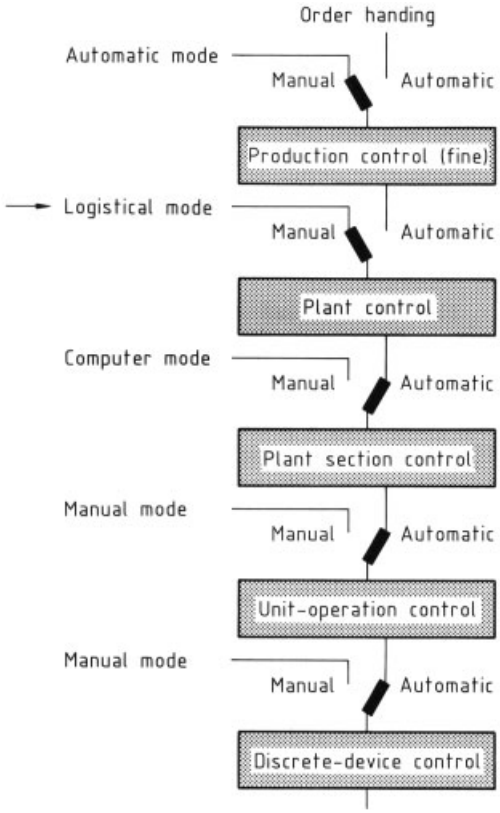

Figure 242. Auto/manual selection

From the process control standpoint, the internal structure of the process instrumentation and control hardware is often of secondary interest. The required automation functions, safety and reliability functions, archiving functions, operator and monitoring interfaces, and so forth can be implemented under a wide variety of system concepts. The differences, however, manifest themselves in important supplemental properties. The plant operator, for example, might see differences in system reaction times, the way in which data surges are controlled, the failure characteristics of components, or the central coordination of workstations. Important from the plant management viewpoint, then, are additional properties that cannot be implemented simply as individual functions but result from the structure of the process instrumentation and control system as a whole. The same applies even more so to the planning and operation of a monitoring and control system. Such additional properties of importance include the following:

- Transparency of structure
- Ease of operation
- Structure with few defects
- Capability of modular expansion
- High availability of individual functions

These can be realized only if they are incorporated in the control system design, i.e., in the internal structure, right from the start.

Historical Development. Today, the phrase "modern" process control systems refers more precisely to distributed or decentralized systems. The world of process control engineering before the development of the type of system discussed here (i.e., before ca. 1975) featured two extreme structures, both of which still exist.

In the *totally parallel structure*, each control task (control action, display, recording, etc.) is performed by a discrete device or combination of discrete devices. An example is the compact controller used in the implementation of a control loop. The parallel system terminates in the control room with a great number of instruments, buttons, and lights. It is left to the designer to ensure that the operator can scan the panels without becoming overstressed in critical situations (see Section 8.2).

The decentralized structure means that a plant so equipped has a fairly high availability, since the failure of a discrete device generally affects only the control task directly assigned to it. The solution of complex control problems, however, requires considerable interlinking of the individual process control stations (see Fig. 65). With discrete devices, this networking can be implemented only through explicit interconnection of devices. Subsequent changes of task are therefore expensive because the system must be rewired, and the flexibility of such a plant is low.

For more complex applications, there was the *totally centralized structure* implemented by the process control computer. This structure, which features the maximum degree of centralization on a single device, has all the process monitoring and control functions concentrated in one device, which is intended to alleviate the disadvantages of discrete-device technology.

All control tasks are carried out by a central computer with interfaces to the process. Individual process control stations are linked together by software, so that the scheme can easily be modified, for example, if the recipe is changed.

The resulting flexibility is partly offset by the lower availability of the plant. A failure of the central computer affects the entire process. Therefore, this structure has generally been used only in conjunction with special practices such as hardware redundancy or subordinate backup systems, which have lead to higher costs for such plants. Nevertheless, the centralized structure in this form still plays a role in the landscape of system structures [579].

Modern decentralized process control systems now fall between these two extreme structures (Fig. 243). The search for an optimal structure has led the manufacturers of these systems to solutions that differ in the degree of decentralization.

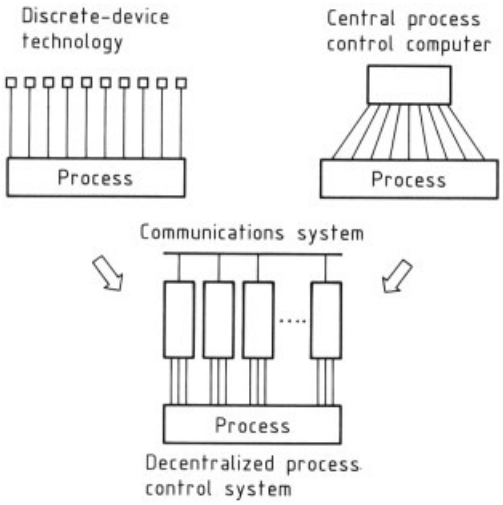

Figure 243. Decentralized process monitoring and control system compared with central process control computer and discrete-device structures

In simplified terms, a decentralized process control system arises through the logical subdivision of a centralized system. This approach is useful because only the use of a process control computer provides the conditions for implementing process monitoring and control functions through digital computer technology.

The development of microprocessors has made the desired or necessary computing power economically viable. Thus, another important prerequisite for the creation of decentralized process monitoring and control systems was fulfilled [580]. Such a system consists of individual computer components, each of which exhibits the properties of a process control computer to a limited extent. These components are linked together by a communication system. As a whole, they have the same capacity as a large central process control computer. However, the decentralized structure leads to higher availability than a centralized design.

The development of process instrumentation and control systems has followed three distinct lines:

- Decentralized stored-program controller systems
- Integrated decentralized process monitoring and control systems
- Intelligent central systems with dumb peripherals

Most process monitoring and control systems can be placed fairly unambiguously in one of these classes. Differences in historical development of the three groups have led to variations in characteristic features such as size of decentralized components, structure of system communication, type of internal processing, integration of operator control and monitoring into the overall concept, and type of design.

The Decentralized Stored-Program Controller (SPC) System. The starting point for SPC systems was the individual SPC (or programmable controller), which is a self-contained unit performing a particular automation task. The emphasis was on the solution of binary networking and sequencing problems. SPCs were conceived as an alternative to relay controllers, which were becoming increasingly complex, costly, and inflexible (as to wiring). For this purpose, the SPCs were for optimized fast execution of binary logic functions, the integration of control and monitoring functions initially playing a very minor role. Elements such as buttons, switches, displays, and lights were employed to operate the SPC and observe its functions. These were connected to the SPC via conventional field-level inputs and outputs. Outwardly, the SPC thus behaved as a wired discrete control device.

The development of SPC systems has led to three major new capabilities:

- Binary-word processing for handling integer and real variables
- Connection to serial bus systems for data exchange with other SPC components

– Support of display-aided operator control and monitoring systems

Compatibility requirements have dictated consistency of internal structure in the development of SPC-aided systems. Present-day process monitoring and control systems based on SPCs thus show traces of their original functions.

SPCs have gained wide acceptance. Other apparatus frequently contains self-sufficient automation components based on them and is sold in pre-assembled package form. Software creation and design are performed by the apparatus manufacturer, so that special in-house control system know-how can be exploited and protected.

From the plant engineering standpoint, manufacturing automation has been and still is the domain of SPCs. An early development was the linking of SPCs and bus systems to form large networks. The purpose of these communications systems was to permit the sharing of data, especially control data and status messages, between SPCs and between the central control computers and the SPCs. A typical feature is simple variable sharing, in which the communication event and all shared variables must be individually designed. An adequate integrated operating and monitoring structure with appropriate system-supported accesses has not yet become available (see Section 11.2).

SPCs are also being used increasingly for plant automation in the field of process engineering. The starting points are often logistical side areas such as formulating lines, packaging lines, and high-bay warehouses. Starting from these applications, SPCs are spreading into chemistry-related areas. Especially in small and medium-sized plants, SPCs are often preferred to decentralized process monitoring and control systems because of their lower hardware costs.

Although such SPC-based solutions are often more expensive than complete distributed control systems when a total cost analysis including engineering costs for design, maintenance, and technical support is carried out, the low cost of adopting them may make them more attractive to plant management when investment funds are being sought.

CRT-aided operator control and monitoring features are available for SPC systems (see Section 11.2). Generally intended for use at individual workstations with limited quantity structure, these devices offer good functionality. Problems arise in larger systems with coordinated multistation operation, increased availability requirements, and central system services (e.g., network clock synchronization).

The separate design of SPC automation functions, communications functions, and operator control and monitoring functions remains an unsolved problem. The lack of certain central functions that would be present in a process control operating system has proved especially serious in SPC-based systems.

Integrated Decentralized Process Control System. The development of integrated decentralized control systems began with the control of large continuous process plants. In contrast to SPCs, the initial motivation was not so much the need to increase the degree of automation as the wish to improve functionality at the individual-function level by making design and installation simpler and cheaper and by making operation and monitoring more straightforward.

The starting point came when discrete-device technology ran into its quantitative limits. In this approach, each individual function is implemented in a separate, individual device; every motor has its own start button, its own run light, and its own contactor control. Figure 244 (see next page) shows a control panel with discrete devices, which are wired with electrical signal lines on the back.

The increasing number of individual functions and the presence of up to 50 control panels in central control rooms meant that an immense collection of controls and indicators had to be put together, creating space problems and also difficulties in comprehensibility. The primary task of the plant operator was to maintain a steady watch over the continuous process. To read the instruments, the operator had to walk along meters of panel bays.

The main goal of the decentralized process control system was to replace this discrete technology with a system combining the advantages of distributed structure and easy operation with the advantages of centralized data management and easily understood, display-supported operation and monitoring.

Figure 245 (see next pages) illustrates the component structure employed in decentralized process control systems. The distributed com-

ponents contain the automation functions for the individual plant areas. They are organized in parallel with the process. The central components contain operator control, monitoring, archiving, evaluation, and central coordination functions [581–583].

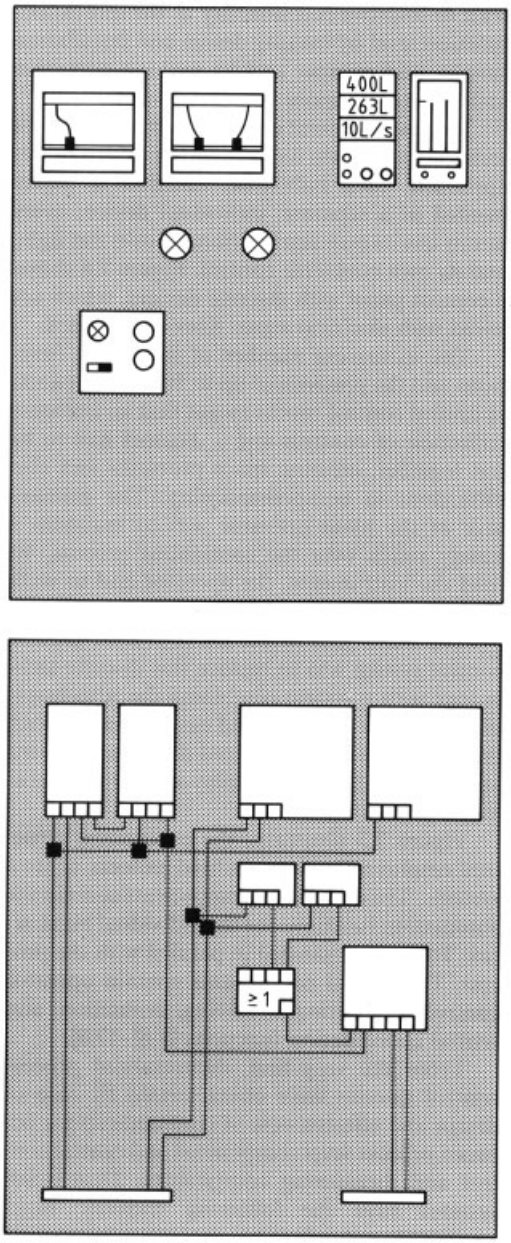

Figure 244. Discrete-device technology
Front and back of a control panel

From the very first, decentralized process control systems were characterized by ease of design and effectiveness of communications between process-level components and central control and monitoring components. Distributed systems have the following features: high-availability redundant bus systems, components that function reliably, low error rate in component design, and good technical support at the component level.

An important property of process control systems is the functional-module approach, in which software modules are handled in the same way as discrete devices. Functional modules can be inserted, switched on, removed, and connected to other functional modules while the plant is on stream. These properties result in simpler and hence cheaper design as well as high system reliability.

The operator interface of functional modules is designed to correspond to the discrete-device operator interface. Because of the similarity in operating properties, operating and maintenance personnel quickly learned to work with the new technology. This was an important reason for the successful introduction of decentralized process control systems over a broad front.

The successful introduction of distributed process control systems led to a demand for a higher degree of process automation. Concepts that have come to the fore in recent years have led to the integration of higher-order control functions, up to the level of recipe control and plant management, into the structures of distributed process control systems (see Chap. 2).

Intelligent Central Systems with Unintelligent Peripherals. The basis for this line of system development is the technology of process control computers, which became available for industrial service in the 1960s, that is, before stored-program systems and decentralized process control systems. Their introduction was greated with a certain technological euphoria, and due to the lack of alternatives this approach was adopted for the full range of plant automation, operation, and monitoring tasks. However, this concept was unsuccessful because of availability problems (due to the centralized structure), the fault susceptibility of the computers then in use, and the fact that both the hardware and the applications software could be supported and serviced only by specialists.

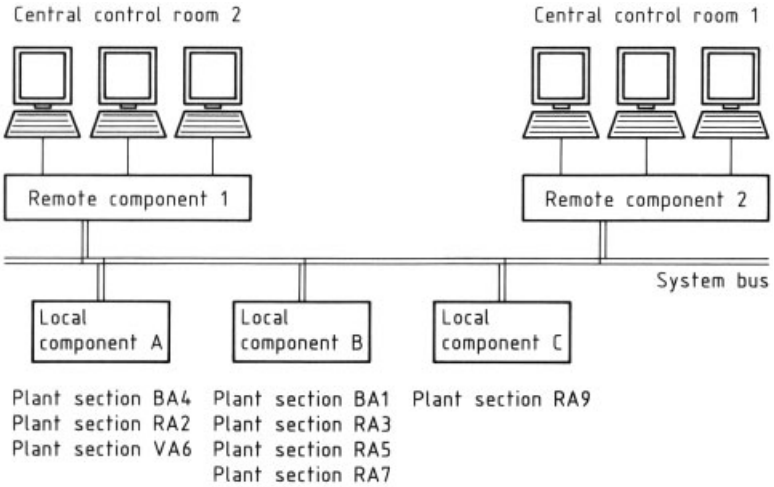

Figure 245. Component structure of a distributed process control system

Today, the only situations in which this concept is used are those where high-quality functions are required sporadically and access to installed monitoring and control systems is not possible. An example is the case of mini-plants and pilot plants, where availability requirements may not be high but it is necessary to test higher-performance control algorithms, implement model-based measuring techniques, or evaluate particular data. Furthermore, in research-level activities, personnel are often able to operate a process control computer themselves.

On the whole, however, this technology is declining in importance. It is increasingly common to use small, decentralized process monitoring and control systems or SPC-based systems for the control and operation of pilot plants. Special "research" functions can be implemented and tested on networks of flexible process control computers.

Structural Models. A process monitoring and control system is a complex pattern made up of many object classes; it can be structured in a wide variety of ways. As a result, a system description is difficult to understand and is always incomplete. Individual features can be explained only from a certain viewpoint, with other aspects being ignored. Nonetheless–and this is precisely the art of structural analysis–all aspects must be equally kept in view.

To provide a framework for the properties briefly discussed in the sections that follow, the most important structural models and the most important ways of looking at a process control system are described briefly.

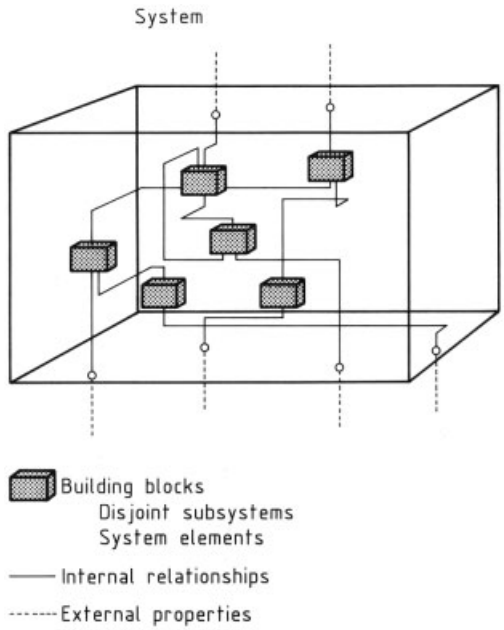

Figure 246. The building-block model

Building-Block Model. The building-block model represents the structure of a system unit in terms of disjoint elements and connections (Fig. 246). This model is generally applicable;

it can be employed, for example, to describe the structure of mechanical, electrical, and functional system units. The building-block model plays an important role in the structuring of process control systems.

The building-block model has some properties that set it apart from general system decomposition and aggregation schemes (see Section 2.2, Fig. 2):

- Nesting principle: Each element used in building up a system unit is completely and exclusively allocated to this system unit. It is, so to speak, "built into" the system unit and is no longer available outside the system unit.
- External interfaces: Each element and each system unit exhibits its properties at certain well-defined external interfaces (e.g., electrical terminals, functional module inputs/outputs).

External connections are possible only through these interfaces, which can in turn relate it to various worlds. A compact controller, for example, has terminals as its electrical interfaces, displays and buttons as its human–machine interface, and bore holes for mechanical mounting.

- Implementation background: Building-block models describe realizable system structures. Hence it is not meaningful to define fictitious (dummy) elements with no realistic background and thus to built up fictitious systems. Instead, the objective must be to identify elements such that they are realizable but also represent manipulable units in the context of the overall structure.

Building-block models are an effective aid to describing the structure of system units of all types. In process control systems, they are well-suited to describing the structure of the hardware, describing the structure of the logical components, and describing the structure of the system and applications software.

Types, Roles, and Functional Units. When one speaks of a system unit or a building block, one must bear in mind that the unit referred to is a self-contained object that can be examined from various standpoints: role (function), functional unit (device; see Fig. 247), type of role, and type of functional unit.

When a control system manufacturer refers to its control system, it means the functional types offered. The design of a control structure means interconnecting functional units. These units correspond to roles that are selected from a list of well-defined types of roles. Ultimately, what is installed are functional units. This distinction is just as important for design and commissioning as it is for the overall understanding of the system. The relationships are shown schematically in Figure 247.

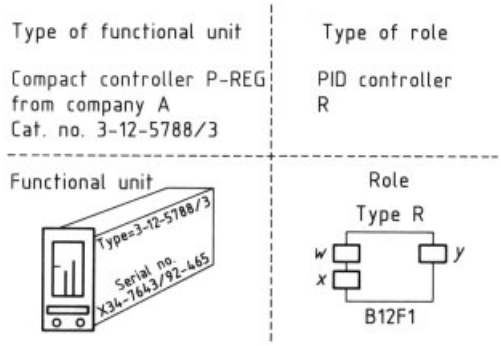

Figure 247. Roles and functional units

Functional units are the concrete devices, components, or software modules that actually perform a function.

The type of functional unit describes the essential qualities of a functional unit. Manufacturers' catalogs list, for example, the types of functional units available. If a catalog lists a P-REG compact controller, then one can order x units of this type. The manufacturer will then deliver x functional units of this type.

The type description of functional units goes beyond the purely functional description, taking in electrical properties, overall dimensions, weigths, prices, etc. The functional properties are a subset of the type properties of the real functional units.

Roles are fictitious functional units. An example of a role is flow controller F 1 in plant section B 12. This role can be defined–and used in building up the role system–independent of which functional unit (i.e., which device or software module) will later fulfill the role.

The model employed here specifies that each role is of a certain type. The type defines all the essential properties of a role. Obviously, any

number of roles of the same type can be defined in a plant.

Role systems are systems built up from discrete roles. In structure, role systems correspond to the building-block model. A functional role system is illustrated in Figure 248. This role system describes exclusively functional properties. It can be implemented with a wide variety of technologies involving a wide range of functional unit types, for example with software modules or in discrete devices. There are also more concrete formulations of role systems. For example, the control panel shown in Figure 244 can be thougth of as a concrete instance of the functional role system illustrated in Figure 248. In this role system, the roles are established much more fully. The description of each role includes not just functional properties but also electrical and constructional properties. These definitions can be made concrete if only functional units of a certain type are considered for its implementation.

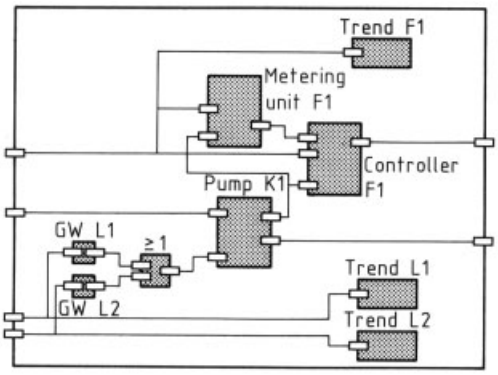

Figure 248. The system of roles

Section 7.5. gives a detailed account of how role systems are built up, concretized, and implemented.

Hierarchical Structural Models. The object worlds of a process monitoring and control system can be arranged in hierarchically related functional levels. Each functional level represents a level of abstraction on which certain functional elements are available and certain tasks (or functions) can be carried out (see Section 2.2).

The modeling of a process control system involves not just one hierarchical structure, but hierarchical structures based on various criteria.

- Operational levels: The hierarchical structure of the operational levels includes interfaces for
  - The control system manufacturer who develops the system
  - The project engineer who configures the process control system for the specific application
  - The plant operator who ultimately guides the process

Figure 249 illustrates such a structure. The lowest level is the base system level, which contains units and functionalities provided by the control system manufacturer, that is, not specially developed for the control system. These might include PCs; workstations; operating systems such as DOS, UNIX, or VMS; database management systems such as dBase, Informix, or Oracle; and all the hardware not specially designed for the control system.

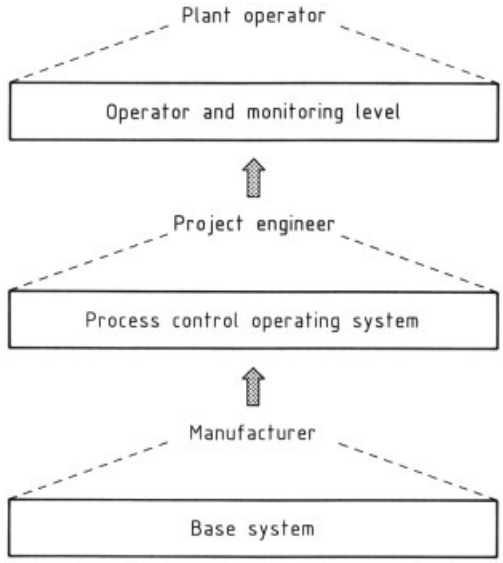

Figure 249. Operational levels

The second level contains the process control operating system. This level makes up the functionality of a manufacturer's control system. The user receives its units and functionalities when the control system is delivered.

The third level is the process control level. It is custom-designed and is available to plant management for active control of the process. It

satisfies all requirements on the control system as set forth in the project specifications.

In the model just described, the various levels are regarded as dense. This means, for example, that the designer does not need to access any level other than the process control operating system level (such as the base system). The operating system can include any required elements of the base system in its functional scope.

Thus the levels of functional abstraction correspond to self-contained working platforms. An important criterion for assessing a process control system is the functionality of the process control operating system level. Functional elements not available on this level must be programmed by the designer as required (see Figs. 16 and 17).

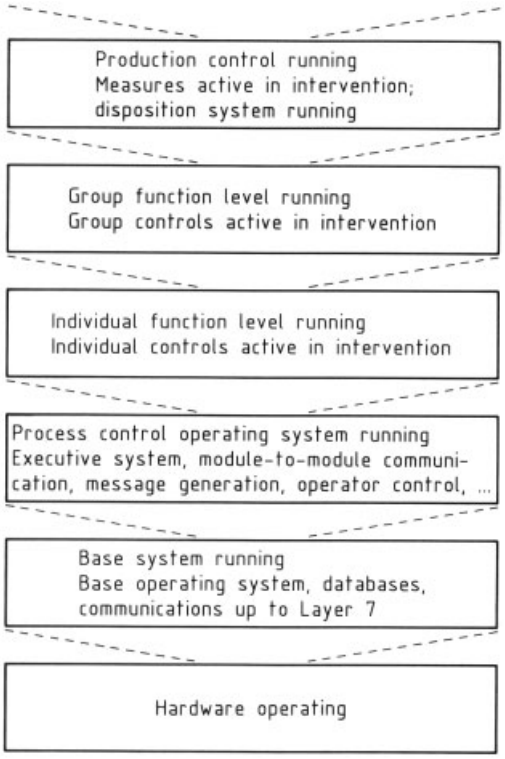

Figure 250. Virtual machines in a process control system

- Virtual machines: The execution functions necessary for the accomplishment of the tasks imposed can also be arranged in a hierarchical model by analogy to the operational levels. During its execution a function makes use of functions of the lower levels. The result is a system of virtual machines, each of which (as shown in Fig. 250) offers the higher-level functions an interface with certain services [584].

The hierarchical structure of execution functions is crucial for the internal implementation of program structures as well as the startup/restart behavior of the system.

At this point it becomes clear that the hierarchical structures illustrated are not to be viewed simply as fictitious structures but as concrete conceptual and real structure-generating programs. This is true of all the organizational schemes presented here.

According to a bottom-up principle, the functions on one level can be implemented with functional elements made available by the next lower level. In this way, more complex functional elements can be created and, in turn, made available for use by higher-level functions.

- Communications protocols: Communication between two applications can also be represented by a level model. Each protocol level corresponds to a "virtual medium" having certain properties, which the next higher protocol level can utilize in implementing its functions. A well-known scheme is the OSI reference model [585], which consits of seven layers and the overlying user functions and user data (see Fig. 296):
  - 7 Application layer
  - 6 Presentation layer
  - 5 Session layer
  - 4 Transport layer
  - 3 Network layer
  - 2 Data link layer
  - 1 Physical layer

  The structure of the communications layers (levels) is discussed in Section 7.2 for the example of systems communications and then described in general and comprehensive terms in Section 8.4.
- Process control: Process control tasks can be arranged into a hierarchical control model. In broad terms, the control model can be broken down into the following hierarchical levels: process control level, production management level, and corporate management level (see Section 2.3, Figs. 31 and 32).

The process control level can be divided into the following sublevels in accordance with the model of Section 4.5:
- Action level
- Group control level
- Individual control level
- Channel/signal level

Viewpoints. In the structural analysis of a process control system, the diverse viewpoints of the teams involved must be taken into account. For example, the following aspects are among those that must be considered:

- Development
- Sales
- Plant management
- Design
- Plant technical support and maintenance
- Device management

Each group represented by a viewpoint has an interest only in certain objects with certain properties. It does not make sense, however, to develop an isolated model from each individual viewpoint, covering only the objects of interest. Instead, it should be possible to relate all viewpoints to a single building-block model of a process monitoring and control system. What is needed are balanced model structures that will yield a useful mode of organization for all groups involved.

7.2. System and Component Structure

Overview. A decentralized process control system consists of individual components connected to one another through a system bus. The division of labor among the components varies in organization from system to system. One general distinction is between process-level or local components and remote components. Any component that has an I/O interface to the field is regarded as local, while any component that has no I/O interface to the field is considered remote. Figure 245 illustrates a typical functional structure.

Process-level components are generally organized in the same way as the plant. Each plant section is permanently assigned to a local component which controls and monitors the plant sections assigned to it. The component is autarkic (self-sufficient) and viable in the sense that it can perform its control and monitoring tasks independently. The result is a typical decentralized structure for process control systems as shown in Figure 251.

Remote components embody functions that are not directly a part of process-level control and monitoring. Examples of such functions are central operator functions, archiving functions, evaluation and interpretation functions, logging functions, production control functions, recipe management functions, and also engineering functions and diagnostic functions [581].

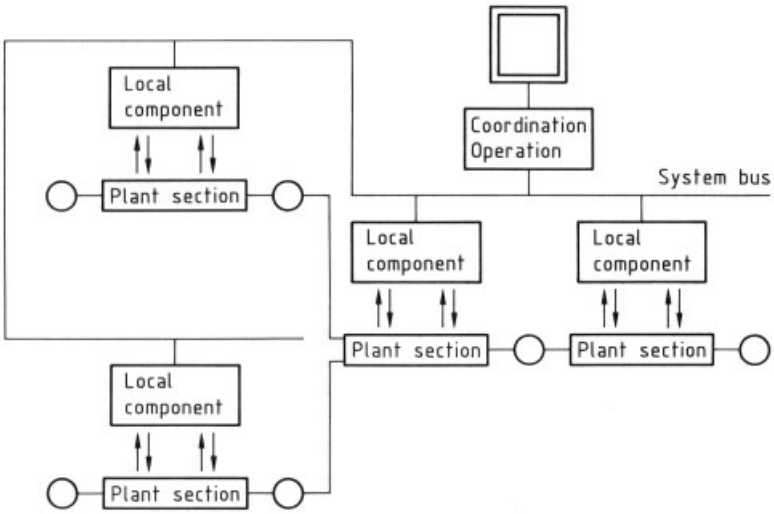

Figure 251. Organization of local components in parallel with plant organization

This division of labor can be modified from case to case. Many systems still have "entry-level" versions in which design functions, operator and monitoring functions, and archiving functions, for example, are integrated into the local components. Such a local component can be used alone, with no remote components. The introduction of economical remote components based on PC solutions has led to a perceptible decline in the use of such discrete component systems. In future, local components will perform only control and monitoring tasks, and even entry-level versions will feature a separate, small remote component.

Local components. A local component consists of a central unit and a peripheral unit, which in turn consists of peripheral or signal modules (Fig. 252)

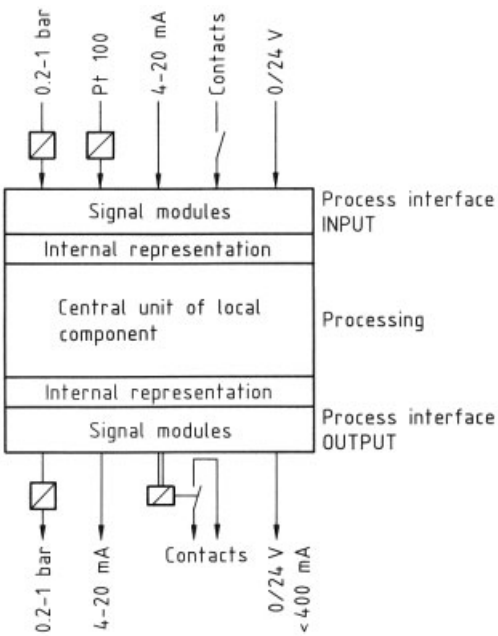

Figure 252. Coupling of field signals by signal modules

Peripheral (signal) modules differ in functional form and can accordingly be placed in four categories:

1) Intelligent discrete control modules, which completely match the functionality of an individual control unit
2) Straight interface modules, which serve only as signal converters
3) Subsystem connection modules, with which subsystems can be linked
4) Field bus modules, which connect intelligent actuators and sensors through a field bus

Intelligent peripheral modules include, for example, controller modules, metering modules, and control modules specific to motors and other devices. The functionality of the individual control level as a whole can thus be realized by separate, independent modules. This added decentrality does, however, have tradeoffs both in cost and in space requirements. If intelligent modules are not absolutely necessary for reasons having to do with the scanning rate or special availability requirements, they are not used at present; instead, their functions are implemented in software in the central unit of the local component. The function of the peripheral unit is thus reduced to simply the input and output of signals, which can be achieved in compact fashion by using multichannel interface modules. For example, input modules are needed in order to feed and read two-wire transducers; to process the signals from NAMUR initiators, strain gauges, and mechanical contacts; to acquire and linearize thermocouple and Pt 100 signals; to handle multirange frequency inputs; and to process 0 (4) – 20 mA, 0 – 10 V, 0 – 24 V, BCD, and other input signals from units. The corresponding output modules provide the following functions: output of current/voltage signals, possibly into standard loads (up to 24 V); power outputs for solenoid-actuated valves; floating outputs; pulsed outputs for integrating positioning drives; and so forth.

In process control engineering, it is important that terminals always be floating or isolated. Inputs for intrinsically safe measuring circuits are a desirable option.

The conversion of measurements and control information to electrical signals can be relocated into local hardware by using field multiplexers. These devices still provide conversions to and from standard electrical signals, but in this case the operation takes place in the field rather than in the control room.

Subsystem coupling modules allow the connection of intelligent subsystems such as scale controls, converters, process analyzers, press

controls, filter controls, and other machine-specific controls. In most cases, subsystem coupling functions support serial point-to-point connections, but there are also bus-type structures with the process control system as permanent master.

Field buses provide logical connections of actuators and sensors with no conversion to static electrical signals. Numerous proposals have been made for defining and standardizing field buses and marketing them as standard systems. Up to the present, it has not been resolved which solutions will finally prevail in the marketplace. Figure 253 illustrates some important field bus systems [586]. In addition to national criteria and corporate tactical decisions, differences in functional requirements also contribute to difficulties in the standardization of field buses.

The *central unit* of a local (process-level) component generally consists of a special microprocessor, ca. 1 – 4 MB RAM, and a connection to the system communications. Additional modules are available in many systems for implementing a local operator control and monitoring function or to provide a background memory for archiving purposes. It is important that the local component should have a process control operating system and be integrated–with its object and communications structure–into the overall system. This means that the system functions of the local components permit, for example, the support of central design and the provision of server services for the operator control, monitoring, archiving, alarm, and message generation functions.

Remote Components. Figure 254 illustrates the tasks (functions) to be performed by a process control system independent of how the system is implemented [587]. Two basic structures exist at present:

- First-generation control systems have no information model that applies throughout the system. The individual components are specially designed and optimized for the tasks they implement. Thus, operation, computer, and engineering components generally have different system architectures and can neither be configured and maintained in a uniform manner, nor controlled through a consistent user interface.
- Second-generation process control systems have an information model consistent throughout the system. Data processing management (operating system, data handling, communications, operator interface) for all remote components is implemented in the most uniform way possible (Fig. 255).

The advantage of this concept is that each function can, in principle, be performed in any component. The allotment of required functions to components in a specific case is carried out with regard to the following criteria: decentrality, uniform processor loading, minimal data exchange, and realization of certain redundancies. The operator interface is uniform across the system. Any function in any remote component can be accessed from any terminal. Thus, from a given terminal, one can configure a component x in the system, use an operator function to perform a certain action in a component y, view the archive stored in component z, and effect an input to the production control system located in component z. Which interventions are possible in a particular case is governed solely by access authorization. The extent to which the human – machine interface is consistent between packages will depend on the state of development of the individual system packages. The need for a consistent concept, however, requires a consistent systems approach from the outset. In small plants, this concept has the important advantage that a modest hardware platform can make available the entire system functionality; thus the hardware governs primarily the realizable quantity structures (quantitative) rather than the functions that can be performed (qualitative). If the requirements are increased, the power of the system can be adapted to the new tasks by the modular addition of hardware. The presence of a consistent information or data model is, however, an essential prerequisite (see Chap. 2).

System Bus (see also Chap. 8). The system bus is the nerve center of the process control system. The following services, among others, are performed via the system bus:

- Continuous delivery of data from local and remote components to the monitoring functions

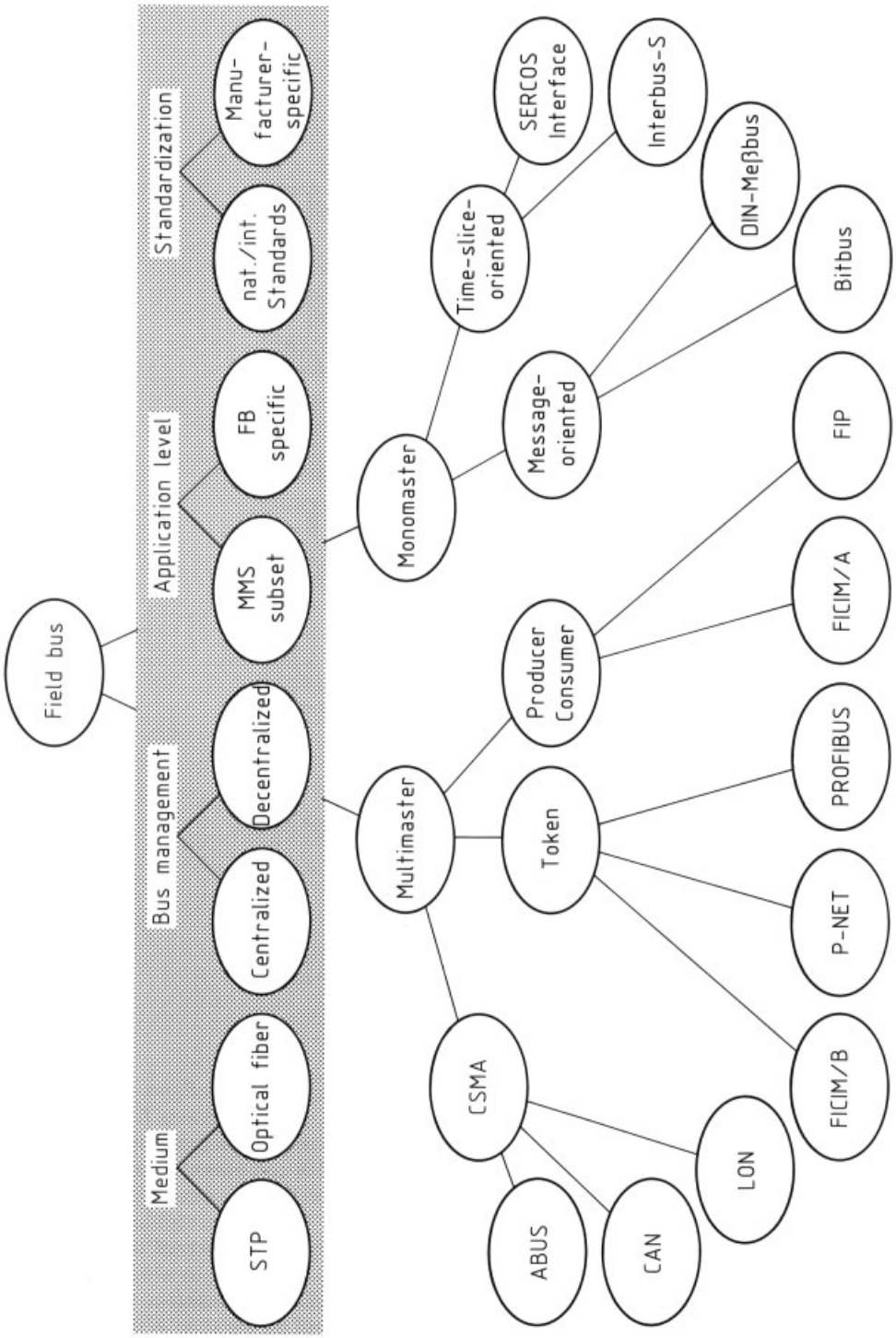

Figure 253. Classification of field bus systems

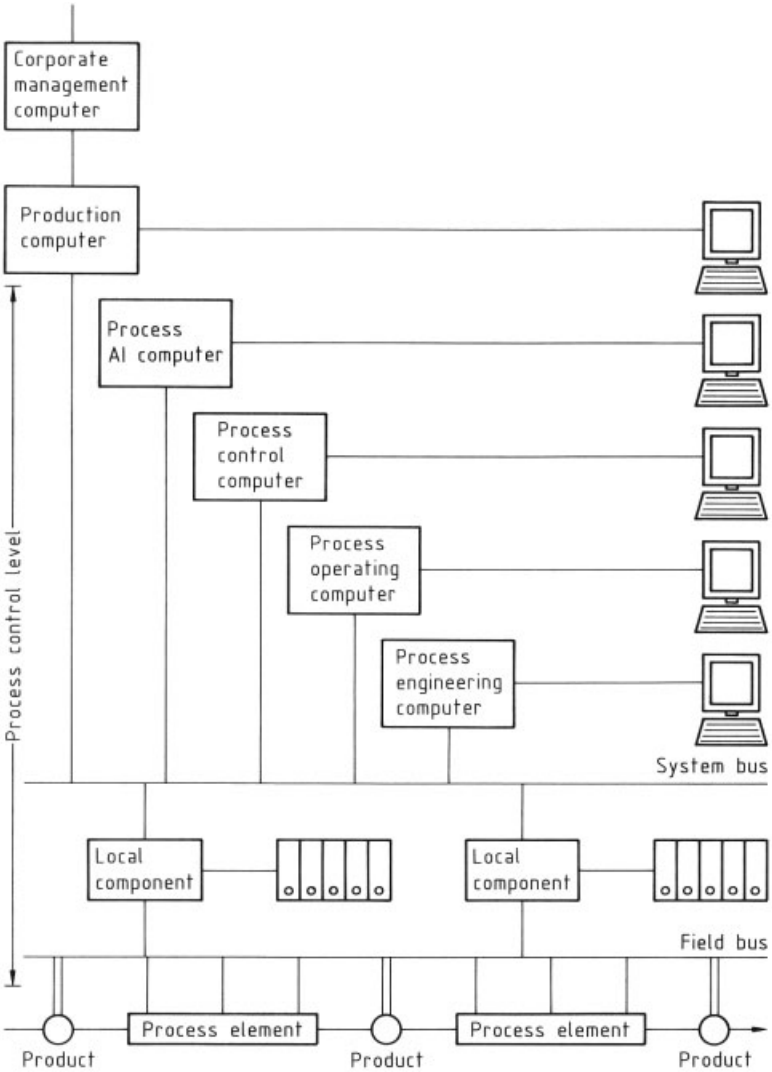

Figure 254. Ideal structure of a process/production control system

- Transmission of operator commands, acknowledgments, and instructions from central process and production control functions
- Provision of required data to the central archive
- Delivery of messages and alarms to the message receiving services
- Object manipulation and object information to assist central engineering
- Loading of actions (production specifications, execution specifications)
- Internal update functions when individual components go out and are restarted
- Temporal synchronization

The system bus of a process control must have the following technical qualities:

- High availability
- Insensitivity to electrical and electromagnetic interference
- Deterministic behavior, even when under load
- Open architecture (compliance with international and de facto standards)
- Hardware and software capable of adding or deleting subscribers while bus is in operation

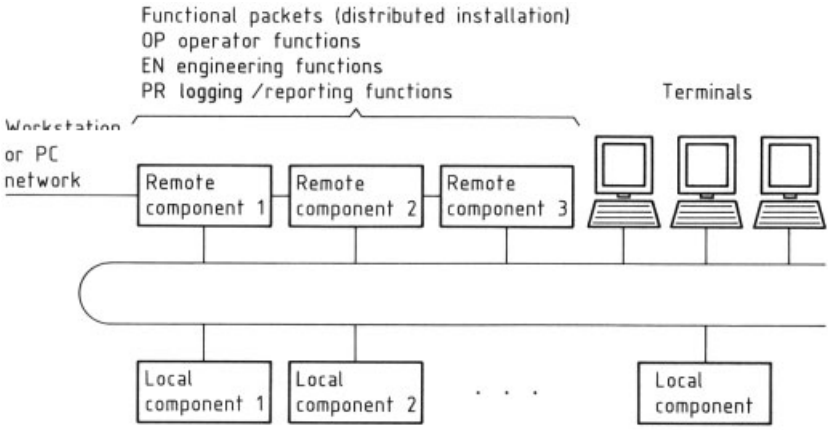

Figure 255. Process control system with continuous operating system platform for distributed use

The following features are possible ways to meet these requirements:

- Use of optical fibers for data transmission. Optical-fiber technology permits high transfer rates even in large-scale networks. Electromagnetic interference and potential problems in the bus system are physically excluded.
- Transparent redundancy of medium. The required high availability is possible only if the transport medium is redundant in design. Its redundancy must be transparent; that is, higher communication levels must not be affected by a switch from a pathway to the redundant pathway.
- Use of token-passing protocols. The use of the token method for transfers means that every subscriber can send important messages within a fixed length of time, even when the bus is heavily loaded.

Currently there are no standard system buses having these implementation features in the middle- (ca. 1 Mbit/s) and high-performance (ca. 10 Mbit/s) ranges. It is therefore no surprise that all system buses now available, while they do use standards for certain levels, still represent manufacturer-specific solutions on the whole.

Unfortunately this commercially important range of medium and high transfer rates is not currently experiencing any development that could lead to a standard solution with the properties needed in a process control system. It appears that there is still some time to wait for an "open" system bus for use in process control engineering.

The only hope at the moment is FDDI in the ultrahigh-performance field. FDDI has some prospect of acceptance as a general standard and has all the technical qualities expected in a system bus for process control systems. It is currently handicapped by its very high cost and by the space and power requirements of the interface groups to handle the high transfer rate (100 Mbit/s). Nonetheless, this approach is going to be a focus of attention in process control engineering circles.

The following comment should, however, be made on the evaluation of data-transfer rates.

If the transfer requirements imposed on a system bus in a process control system for controlling the process are considered, it can be seen that a gross transfer rate of 1 Mbit/s is sufficient for such tasks, even in larger networks. The net transfer rate strongly depends on a choice of individual transmission services and the network data distribution model.

Bottlenecks can occur, for example, when archives are updating one another, when static image information is being transmitted over the bus, or if an unsuitable data model has been chosen. A typical example of a transfer-intensive data model is the centralized display of all local component status information at remote component locations.

In the assessment of a bus system, therefore, the gross transfer rate is less important than a well-thought-out model of data management and data sharing on the network.

Component Structure of Commercial Process Control Systems. Figures are intended to explain the component structures of several process control systems. The selection is meant to provide examples and does not reflect either market position or technical ranking.

In Figures 256 – 269, the notation used for the components complies with the manufacturers' product designations.

The following symbols are used for functional packages:

AR Archiving functions
CO Control functions
EN Engineering functions
GA Gateway functions
IO Process input/output functions
OP Operator functions
PR Protocol (reporting/logging) functions
CD Coordination functions

Note that some functions relate to not a single component but several. In such cases, the packages are separated into client and server portions and a letter (C, S, or L) is appended in parentheses:

(C) Acts in the system as client of the function
(S) Acts in the system as server of the function
(L) Function is implemented locally with no involvementof system communications

The codes C and S are used only for components that support the function on the system side, that is, when communications need not be separately programmed.

Examples:

Engineering Functions. The engineering component in which design activities are carried out acts as a client in the system. The target system designed by the central engineering function acts as a server in the system. The engineering station would thus have (C) appended to its symbol, while the target systems designed by the central engineering functions via system communications would have (S) appended. Locally configurable components would have an (L).

Operator Functions. Functions available troughout the system provide for operator control, monitoring, and the transmission of messages and alarms. With these functions, information is shared between individual applications in the target system and the operator (via the human – machine interface). The display and operator components act as clients, while the target component in which the operation takes place or in which a status is interrogated is regarded as the server.

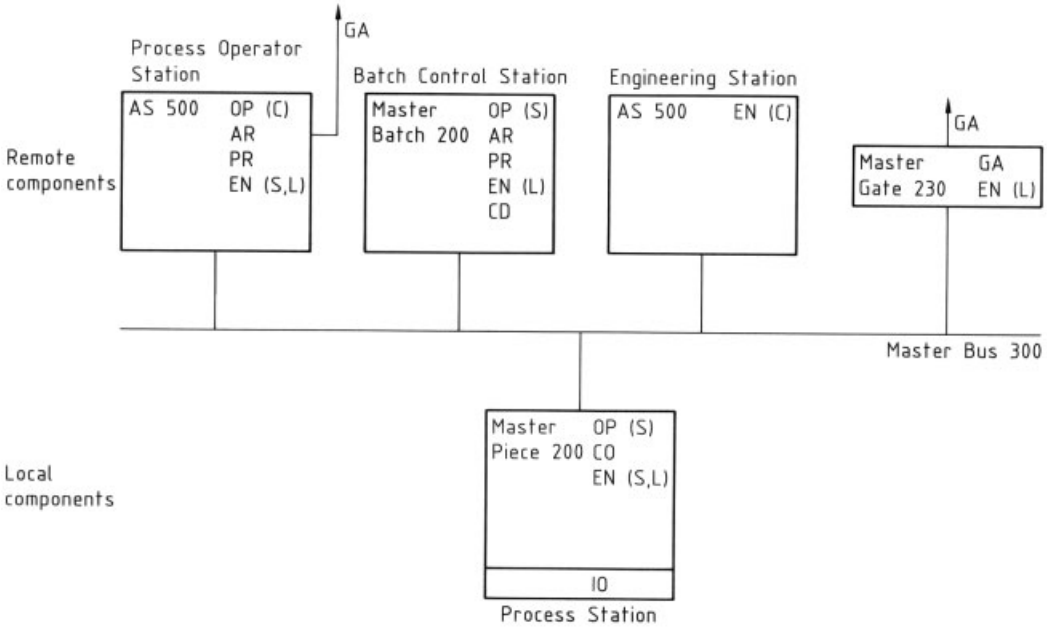

Figure 256. Functional structure, ABB Master

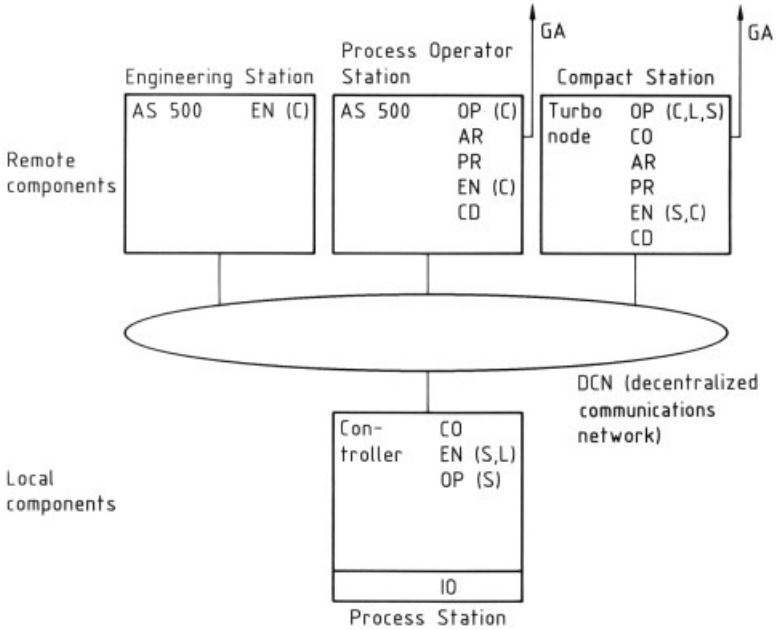

Figure 257. Functional structure, ABB MOD 300

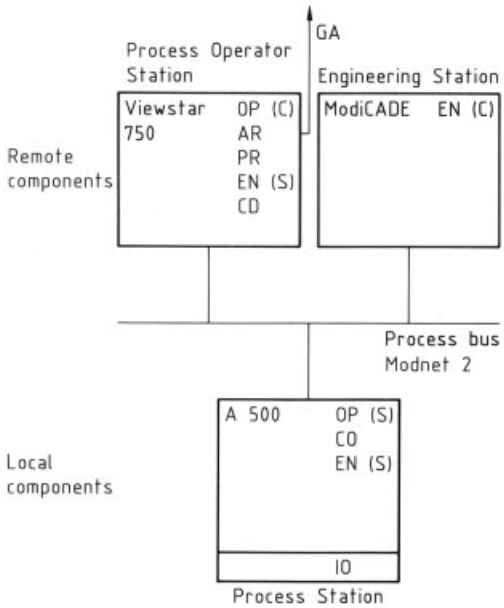

Figure 258. Functional structure, AEG Geamatics P

If one examines the structures in detail (see Figs. 256 – 269), one sees that most of the process control systems illustrated offer extensive systems solutions for the handling of local components. The major shortcomings are in communications between remote components. Server properties are largely lacking in these remote components. In other words, remote components often cannot be centrally designed (configured), and the normal operator control and monitoring functions do not make it possible to operate and monitor applications in other remote components.

Redundancy. The following concepts can be applied to redundant design:

Necessity of redundancy (see also Section 10.4)

The availability of the process control system depends on the frequency and duration of failures of its components. In general, when the process control system or a portion of it fails, the production plant or the respective plant section also fails.

If an especially high availability is required of a plant section, or if the failure probability of this plant section must be particularly low, the local components controlling it must be designed with redundancy. However, this is an exceptional case. In many situations, it is acceptable for individual plant sections to fail. Production can be briefly interrupted, or plant sections redundant from the production engineering standpoint can be switched over. The decentralized structure of

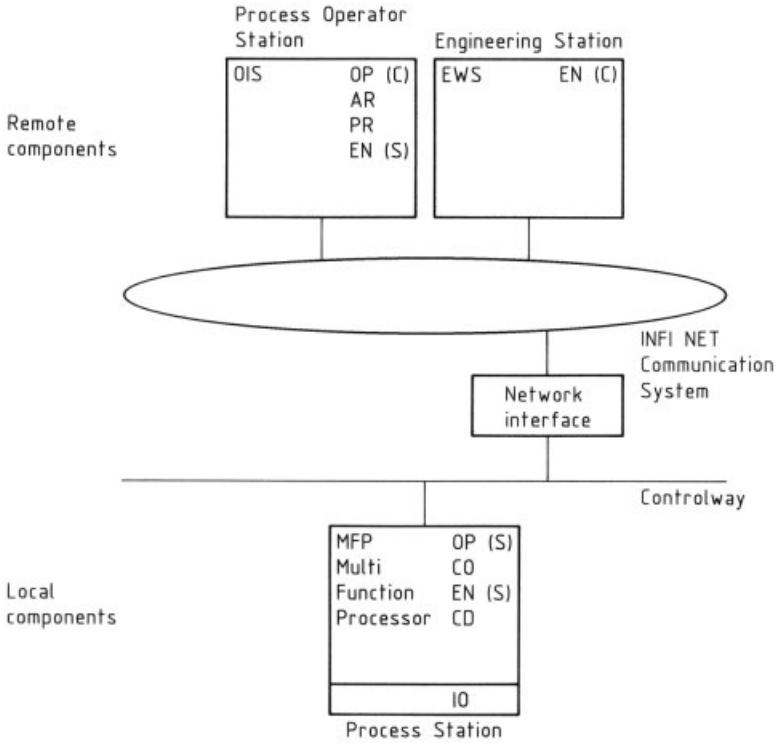

Figure 259. Functional structure, Bailey INFI 90

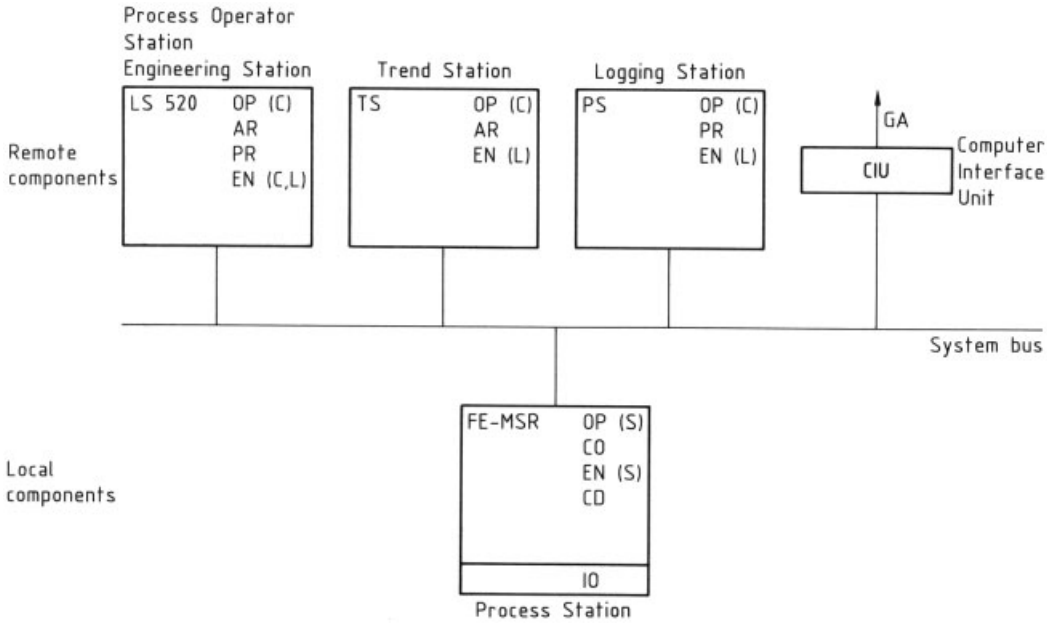

Figure 260. Functional structure, Eckardt PLS 80 E

the control system means that the failure of one local component affects only one of the redundant plant sections, provided the allocation has been chosen correctly. Here, the process control system need not exhibit additional redundancy.

In the case of continuous processes that run nonstop for 3 – 5 years, as is common in some industries, scheduled maintenance of the monitoring and control hardware becomes a problem. The performance of scheduled maintenance generally involves removing modules, while inspection entails removing entire system portions from active service. Redundancy is essential if this is to be done.

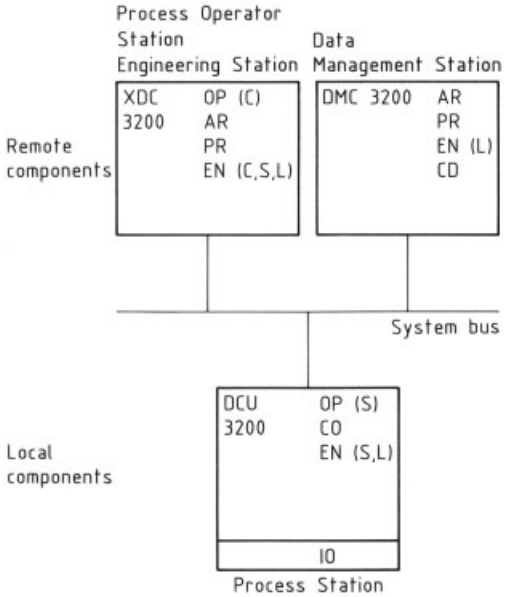

Figure 261. Functional structure, Fischer & Porter DCI System SIX

Dispensing with integrated operator and monitoring functions in the decentralized local components makes it necessary to have a redundant bus system with redundant implementation of operator functions in the remote components. Back-up operation with local intervention at the local components is generally no longer possible.

Modular redundancy concept

In the modular concept, the decision whether to make a unit redundant or not is made separately for each unit in a system. Three points favor modular redundancy:

- Redundancy is only desirable for well-defined, self-contained units
- Modularization creates redundancy nodes, which separate the individual redundant units from one another and increase the availability of the system
- Redundancy is expensive and it is therefore sensible to employ redundant design only where necessary for a specific application

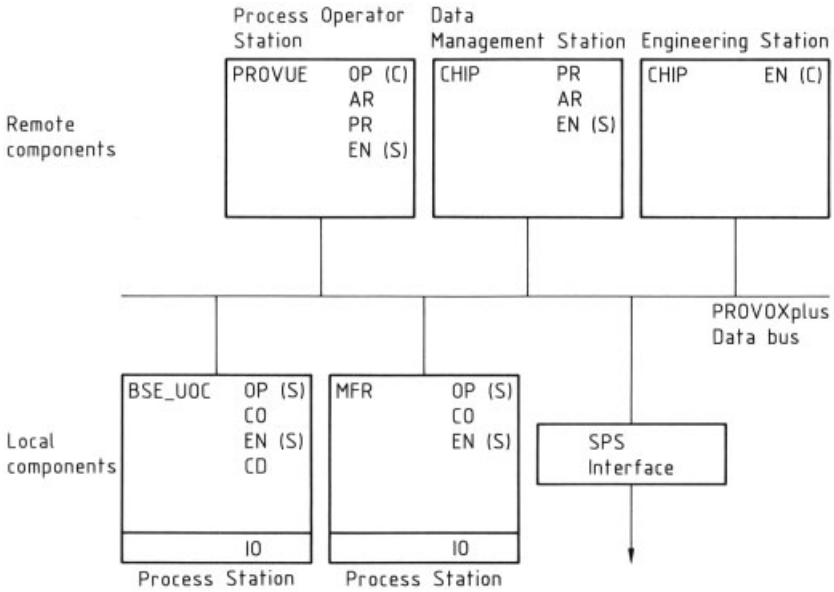

Figure 262. Functional structure, Fisher Controls PROVOXplus

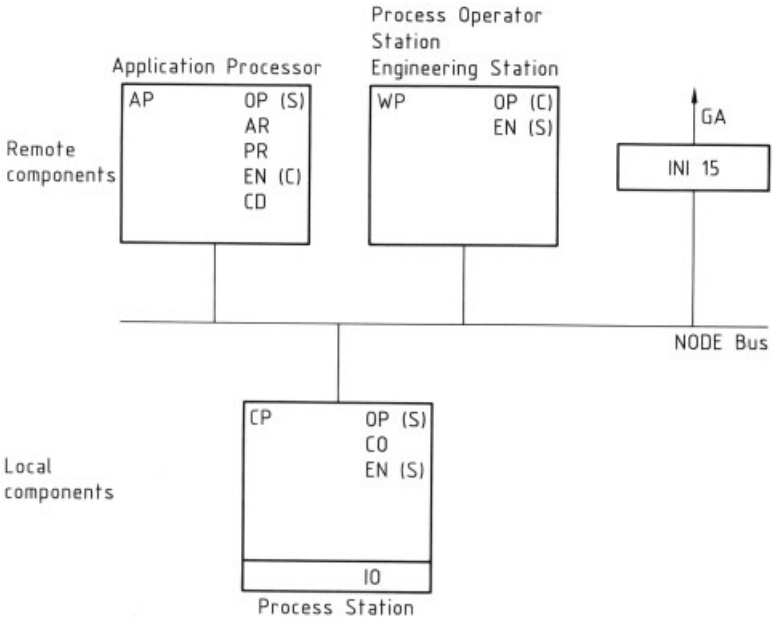

Figure 263. Functional structure, Foxboro IAS

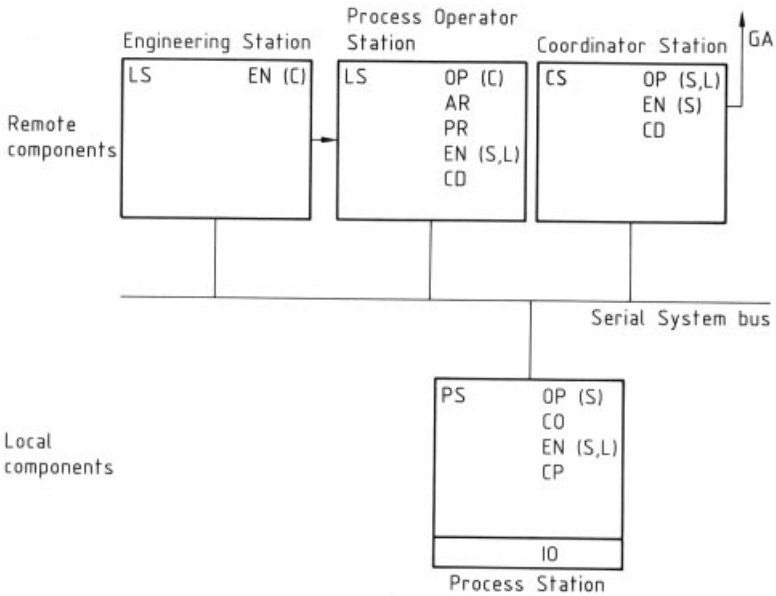

Figure 264. Functional structure, H & B Contronic P

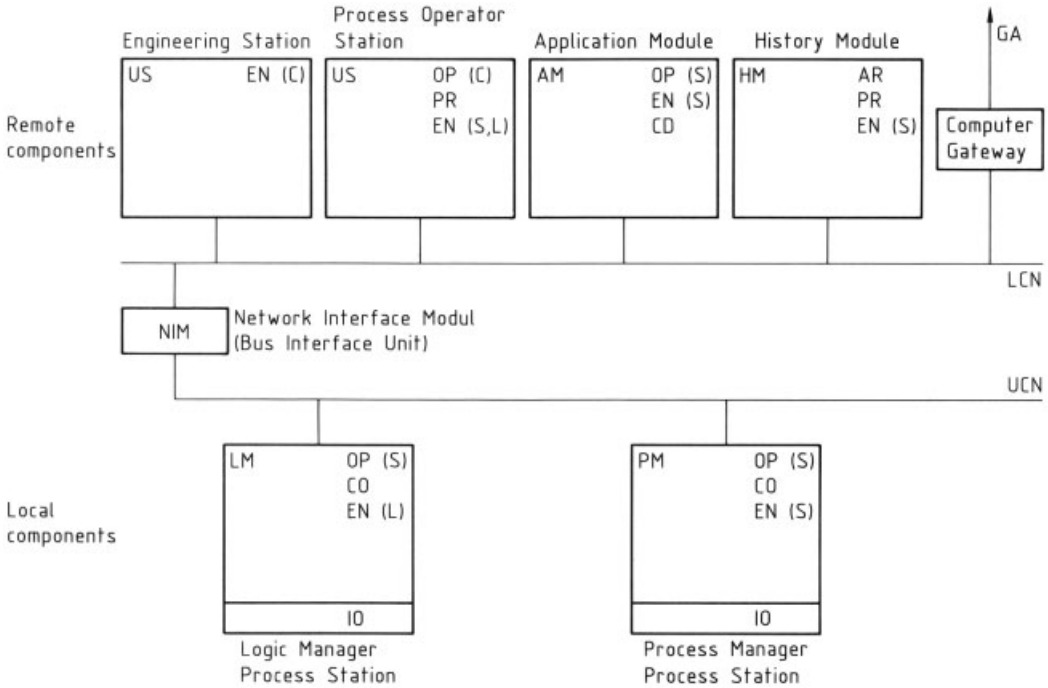

Figure 265. Functional structure, Honeywell TDC 3000

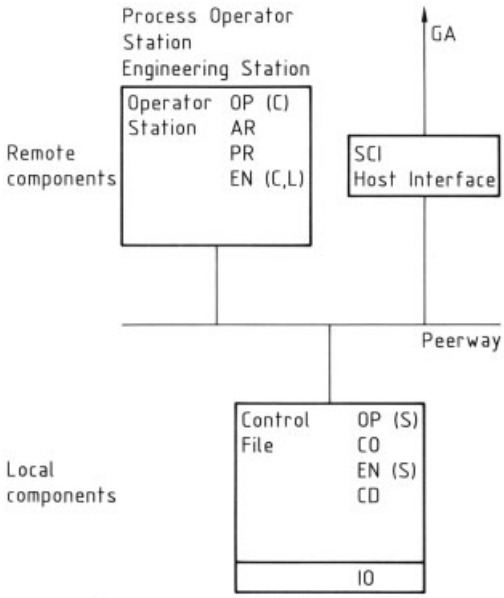

Figure 266. Functional structure, Rosemount System 3

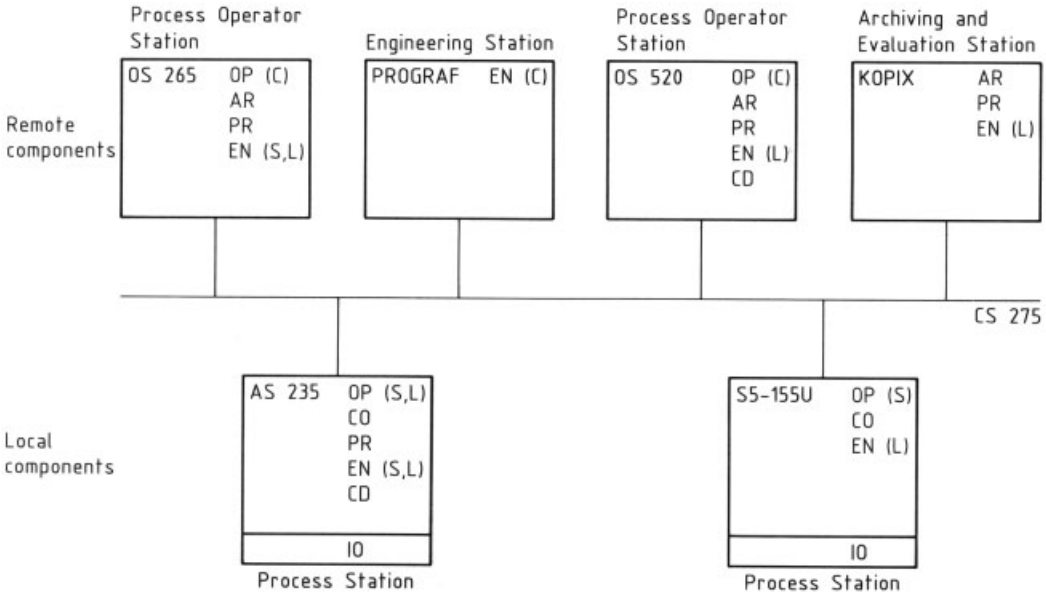

Figure 267. Functional structure, Siemens Teleperm M

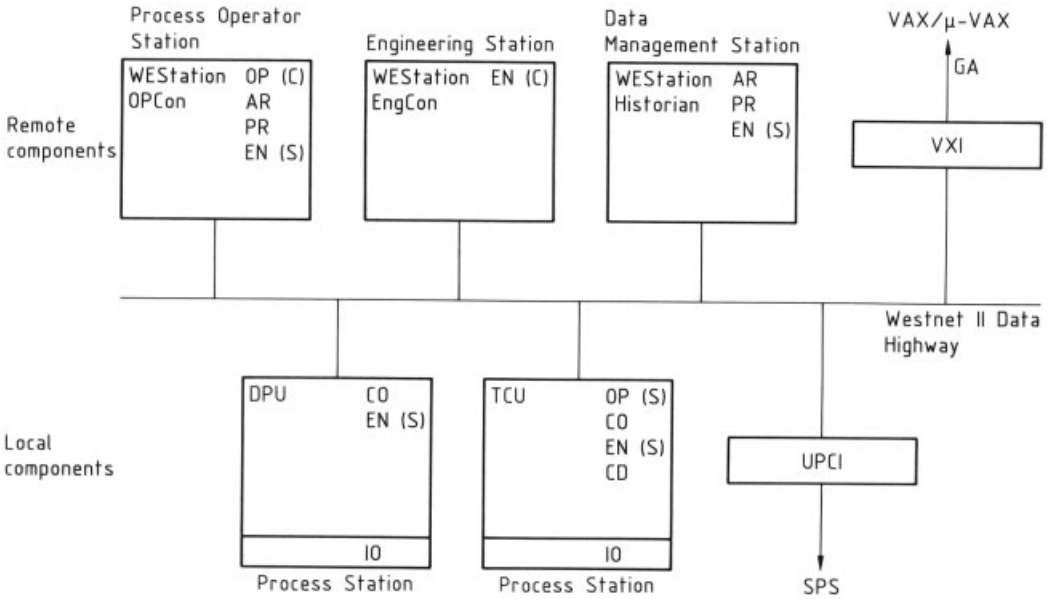

Figure 268. Functional structure, Westinghouse Controlmatic WDPF II

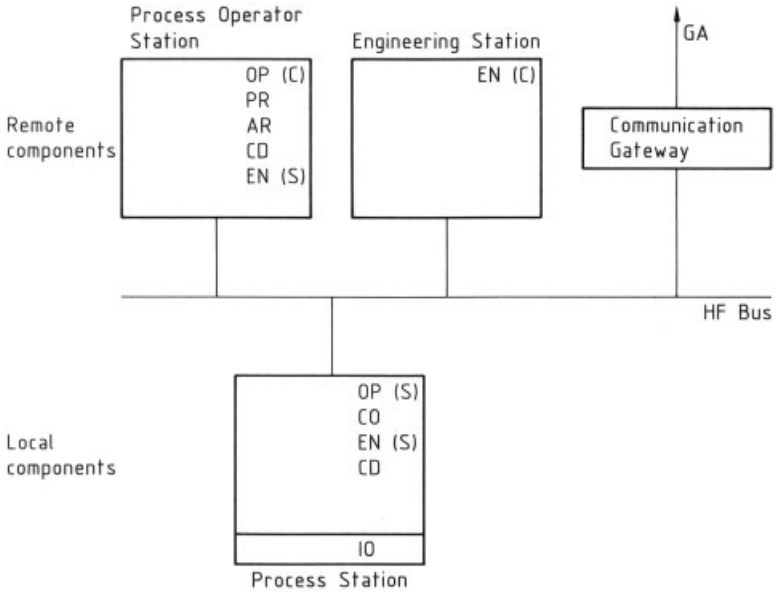

Figure 269. Functional structure, Yokogawa Centum XL

Figure 270 illustrates a modular redundancy concept for the system bus and local (process-level) components. This is an instance of transparent redundancy of the medium, which means that the individual data-transfer and processing units implement redundancy in such a way that higher-level system and user functions do not see it logically.

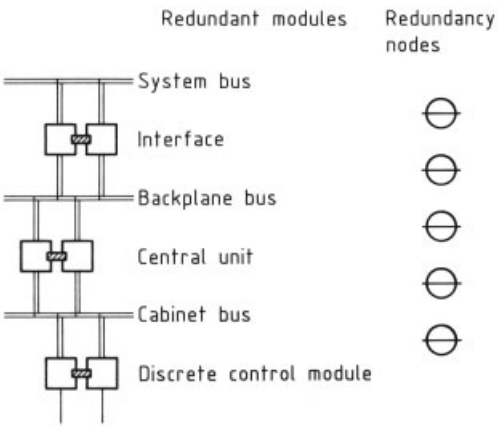

Figure 270. Modular redundancy concept

The situation is different in the area of remote components (i.e., on the "client side"), where standard components (workstations, industrial PCs) are employed in a nonredundant system structure. Redundant design is aided by special functions in the process control operating system and general system services and is ulti-mately realized in the structure of the user functions.

Quality of changeover to redundant units

The redundant layout of units means that if a unit A fails, it is possible to change over to another unit B which continues to implement the required functions.

An ideal changeover is as follows:

- A malfunction in A is correctly detected as soon as it occurs, and the changeover takes place before a single datum in A is altered
- Prior to the changeover, B is in exactly the same state as A
- The changeover takes place with no time delay
- The changeover is "bumpless"

This is, of course, an idealization. In real systems, a number of problems arise:

- Detection of the Malfunction. A malfunction can be detected by self-diagnosis in A itself or by external monitoring. External monitoring systems are commonly based on the comparator principle. In a 1-out-of-2 redundant setup, however, localizing the malfunction is a problem. If the two units do not agree, which is wrong, A or B? Another problem

is the time between the occurrence of a malfunction and its detection. Has A improperly changed its state in this interval as a result of the malfunction?

- Synchronization of B. B must begin working at the point where A quits. The question is how B, at an acceptable cost, can track A so as to hit the starting point as exactly as possible. The designer must then ask: How is the synchronization to be done? If A and B are clock-synchronized, or if there are synchronization points in the software, what deviations in the instantaneous states are tolerated and what deviations are unacceptable? Are data transfers in progress terminated and restarted, or is control passed "on the fly"? And so forth.
- Changeover. When the changeover takes place, all communications links of unit A are broken and replaced by links of unit B. The following problems arise at this point: How is the logical changeover accomplished? How long does B take to come into sync? How long will it take before all partners of A/B have dropped the old connections and made the new ones? Is the changeover automatic or manual? Does unit B remain on permanent standby, or must it be turned on or even fetched from storage?
- Restart. Once the malfunction has been corrected, how is the automatic restart of A (up to hot standby) supported without impairing the function of B?

The implementation of a modular redundant process control system is essential if the stated requirements are to be met. A good redundancy concept is the mark of distinction in a process control system. The questions stated under "Quality of Changeover to Redundant Unit" are not intended to lead the designer to a particular solution but rather to indicate the complexity of the problem and afford some warning against obvious but half-baked solutions.

Electrical and Mechanical Design. Process monitoring and control systems must function reliably, feature robust construction, and be adaptable and modifiable even while in service. Special importance therefore attaches to the mechanical and electrical design of these systems.

Mechanical Design. Local components are generally rack-mounted inside steel cabinets. Central modules are located in one plug-in unit; peripheral modules, either along with the central modules or in additional plug-in units. A module can be a simple card or an enclosed rack-mount device. The technique used to connect signal lines leading to the field varies from system to system; plugs on the front and backside are common, as is direct wiring to terminal posts or screw terminals in a special terminal bay.

An important mechanical design question is whether the cabinet must be accessible from one side or both sides. Practice has shown that access to the back of the plug-in units is necessary. This can be achieved by leaving a walkway at the back or, when this is not possible, by using a swing-out cage.

Remote components must also be considered in the course of mechanical design. Because these are generally off-the-shelf hardware, they must be mounted in an appropriately stable and secure way, either in control room panels or in properly equipped cabinets in the process control rooms (see also Sections 5.3 and 10.7).

Electrical design includes the overall electrical behavior of the process control system components. Important aspects of electrical design are power supply, isolation and earthing, protective enclosures and safety requirements, electromagnetic compatibility, and response to external conditions such as climatic, seismic, and chemical effects (see Section 6.4).

- Power supply.
  The process control system is responsible not only for controlling the plant but also for monitoring it. It is desirable to provide a highly reliable uninterruptible power supply for the process control system; this makes it possible at least to acquire and record information about the process and plant status even when power to the processing facility is interrupted. It is precisely these exceptional situations that make the process control system so vital as an information system (see Section 6.4).
- Isolation; earthing concept.
  A process control system communicates with a number of partners, some of them from different manufacturers. This is particularly true of local (process-level) components as well

as the subsystem and field areas. It must be possible to adapt the control system to the local grounding concept. The goal is thus a control system "open" from the grounding standpoint. This requirement will be fulfilled if each input and output in the process monitoring and control system is electrically isolated. The earthing concept determines which potentials are connected in a specific case (see Section 5.3).

- Protective enclosures and safety requirements.
  Process control systems are electrical systems and must satisfy regulations on shock-hazard protection and protection against the ingress of solid bodies and water. These regulations are based on IEC 529 (protection against foreign bodies and water) and IEC 536 (shock-hazard protection). In process control systems it is particularly important that the following sections be spatially separated from one another: low voltage; functional extra-low voltage with safe separation (PELV as specified in DIN 100, Part 410); and intrinsically safe circuits.
- Electromagnetic compatibility.
  Process control systems are operated in an electrically harsh environment. In many situations today, these systems are located next to static converters and other power electronics, in the same room or even cabinet to cabinet. Operating personnel use handheld radios when the cabinets are open, and the rebranching field cables can pick up interference. Therefore, electromagnetic compatibility is a special concern in process control systems. Electromagnetic compatibility is a measure of resistance to noise resulting from
  - Static discharges (IEC 801-2)
  - Electromagnetic fields (IEC 801-3)
  - Bursts [IEC 65(Co)39]
  - Hybrid voltage transients [IEC 801-5 (Draft)]

Other guidelines, recommendations, and standards have been issued by NAMUR and DIN; see, for example, [588–590].

For each device, the severity of the effect can be measured and classified. Four classes can be defined:

1) No impairment of operation or function
2) Momentary impairment of operation or function, self-correcting
3) Momentary impairment of operation or function in which restoration of service requires resetting or operator intervention
4) Permanent loss of function resulting from damage to the device (or its components)

In general, it is expected that the maximum effects under moderately severe testing (IEC Severity 3) will fall into Class 2.

- Other conditions.

This heading takes in all other environmental effects that might lead to impairment of the functioning of the process control system. They include, among others:

- Temperature and humidity
- Seismic stress
- Chemical attack (e.g., acidic gases)

Environmental conditions are defined and classified in IEC 721; test methods are set forth in IEC 68.

7.3. Process Control Operating System

A process monitoring and control system is a distributed system in which modular control structures are managed and executed in the most efficient way possible. This applies in particular to the local (process-level) area. The process control system should make it easy for the user to work with a modular system (as illustrated in Fig. 248) in a similar way to a system based on discrete devices.

For example, it should be possible to install or remove a module without interfering with the operation of other modules, and to make or break a connection or modify a parameter value while the system is in operation. Another important point is reliability. For example, the making of a connection should not inadvertently cause internal interference with a modular algorithm, while the manipulation of a module should not result in the accidental change of a parameter value in another module.

These requirements can be satisfied only if the process control system has an operating system with appropriate object manipulation and object-oriented programming capabilities; the operating system should also support the creation of modular software structures.

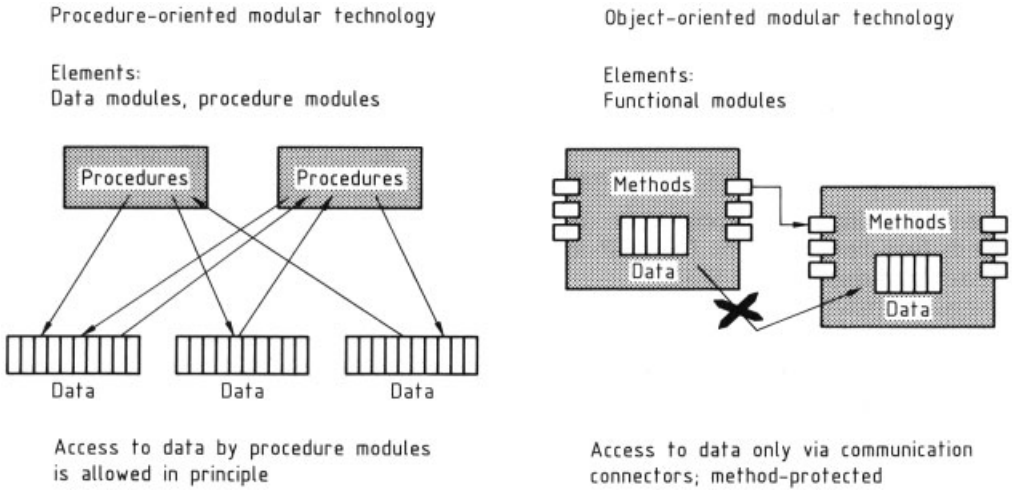

Figure 271. Procedure-oriented and object-oriented modular technologies

Three approaches have been taken to implementing these requirements.

In the first approach, the entire "modular intelligence" is concentrated in the design system. The local components responsible for control system operation do not have a modular structure. The design system makes a translator run, generating a transparent code from the objects managed, and loads this code into the local components for execution. The advantage of this technique is that the requirements on local components are modest; the local components can be structured easily and thus economically. On the whole, however, this concept has not yielded any satisfactory solutions. The main reason is that changing a single object requires retranslation and reloading of the entire program. Such a system is extremely resistant to changes.

When partial translation or partial loading is desirable, the design system must have a detailed knowledge of, and partial authority over, the data organization of the local components. The cost of developing such a design system is high and cannot be controlled, especially when the local components are further developed.

For these reasons, all successful SPC (stored-program controller) systems are decentralized process control systems that include management structures for objects in the local components themselves.

It is necessary to distinguish between the "procedure-oriented" modular approach used in SPC systems and the "object-oriented" modular approach used in distributed process control systems.

Figure 271 illustrates the difference. In the procedure-oriented modular technology, data modules and procedure modules are the objects. Distinct procedure modules communicate through data modules. The procedures describe certain data elements of the data modules, which in turn are read by their procedures. The connection does not exist as an explicit object. In this method, the system does not afford protection of access to the data areas; that is, all data can be read by any procedure, and also written if the procedure is write-enabled.

The right-hand part of Figure 271 shows the object-oriented modular technology, where data and procedures (here also called methods) are combined in a single module that can be handled only as a whole. Data sharing takes place via connections between the module connectors. Direct writing and reading of internal module data is not possible at all. The only way to access these internal data is locally, through the methods inherent to each module.

It is important to maintain the distinction between these two technologies. In practice, many misunderstandings arise when the term "modular technology" is used. In the SPC world, procedures are also referred to as "functional modules" and data as "data modules."

The forms of office organization shown in Figure 272 illustrate the properties and applications of the procedure-oriented and object-

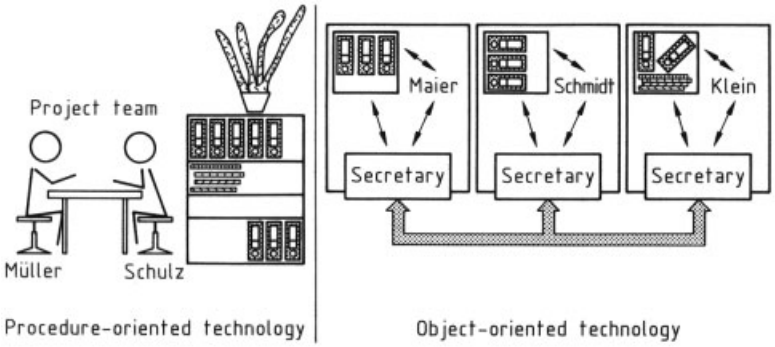

Figure 272. Examples of procedure-oriented and object-oriented setups

oriented approaches. On the left, Müller and Schulz work together in a project team. They sit across the table from each other, in a room containing a cabinet with a shared set of files. This arrangement corresponds to a procedure orientation. Müller and Schulz correspond to the procedures; the set of files, to the jointly available and processed data. This form of organization has the advantage that every worker has quick, direct access to the common data. The setup does, however, require intensive communication and agreement between the two team members: each must know exactly what the other is doing and what data are being changed. While this organizational form is feasible for small, easily overseen units, in larger and more complex structures it is no longer possible.

More complex structures require local data safeguards. Such a form of organization is shown on the right in Figure 272. Maier has his own set of documents, to which only he has access. No other team member can enter his office to view a file or enter anything in a file; anyone wishing to obtain such information or make such changes must ask Maier to provide information or make an entry. Maier himself decides whether and how to respond to this request. Maier and his data form a unit that, as a whole, performs certain tasks. This form of organization corresponds to an object orientation.

If the parallel is drawn to process control systems, it becomes clear that the traditional approach in SPC-based systems–which historically arose from small automation systems with fast access requirements–is procedure orientation. Decentralized process control systems, on the other hand, were designed for large, complex networks from the very beginning, and in these systems the object-oriented modular approach predominates.

The remainder of this section will describe the basic properties of a process control operating system supporting an object-oriented modular system.

Software Modules. One characteristic of a modular system is that a module–with all its parts, data records, connectors, and methods–is treated as a unit. In production service, from the communications standpoint, the module can be identified by a single reference.

Such an object-oriented modular concept can be implemented in either of two ways: by realizing the modules as flat, completely autonomous objects, or by realizing them as entities belonging to a class hierarchy.

Methods using discrete devices correspond to the first approach. Each compact controller has its own data storage, its own interfaces, and its own logic for implementing the PID algorithm. In a software-type concept, which follows the second approach, the modules are realized as instances of module types; this corresponds to a special, one-level abstraction scheme.

When it is realized in the functional system, each module is therefore split into two objects, a type object and an instance object. The type object contains the processing methods, a description of the data structure, and a description of the connector structure. The instance object (entity) contains the module name, the individual status information, and a reference to the type object.

This functional structure implies two properties that are significant for module handling in process control systems:

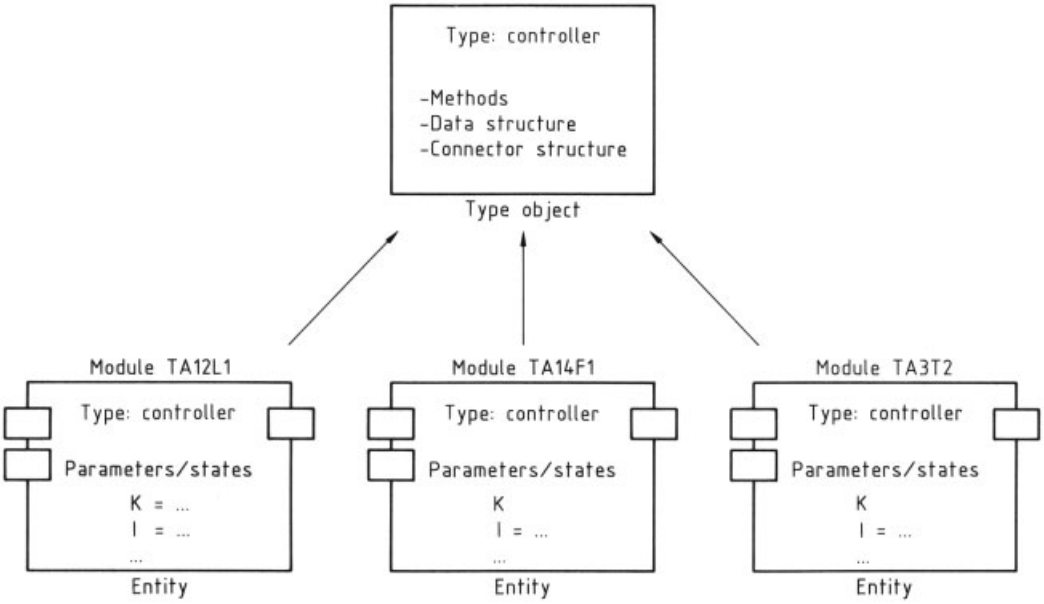

Figure 273. Software implementation of three modules of the type "controller"

- Type objects and entities can be created and loaded separately
- More than one entity can be defined for one type object

Figure 273 shows realizations of three controller modules. All three use the common type object "PID controller"; modules TA 12 L 1, TA 14 F 1, and TA 3 T 2 are three entities.

The creation and loading of types is an operation that generally calls for profound system knowledge and can be performed only by certain persons. In many process control systems, the user cannot create new module types with their own data structure, connector structure, and methods. The user must make do with the set of types provided by the manufacturer.

In contrast, it is very easy to work with instances under the software-type concept. Because all data structures, connector structures, and methods of a module are defined when the type is created, the instances need only be generated as data records, parametrized, and inserted in the processing sequence. These are simple operations, and in virtually every process control system they are supported by the operating system of the local components, even during production operation.

The software-type concept has proved effective for the handling, management, and sequential organization of the modular system in process control systems. In a variety of components, the concept has been embodied and optimized for the requirements of a real-time system with many objects to be processed concurrently.

Certain constraints relative to a general class hierarchy are unavoidable. They include the following:

- The functional system supports only one level of abstraction (entity – type).
- Type data are present in multiple copies and are thus not free of redundancy. The type data for a module are available locally, as loaded type objects, with respect to every processor.
- Each instance has a direct reference to the type object of its logical environment and knows what information it will find where.

As far as the external function of a module in the modular network is concerned, it makes no difference whether the module is realized as an autonomous object or as an instance of a type. In the software-type concept, the instances provide the external representation of the module with all of its properties.

Communications Network. A major task of the process control operating system is to manage and dynamically perform communications in the network of modules. Each module

has connectors serving as external interfaces, through which it exchanges information with its surroundings. All communications in the modular network can be depicted in terms of individual communications links, each between one module input and one module output.

From the applications standpoint, there are two basic forms of information exchange: the "get" relation and the "send" relation.

Figure 274 illustrates a *get* relation. B is interested in the present state of A and reads this state from A without feedback. Such a relation corresponds, in discrete technology, to acquiring a signal from an output terminal. A given output A can be read by various inputs.

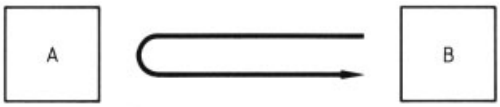

Figure 274. Get relationship

In this kind of relation, only the recipient is interested in exchanging information. The recipient needs a state that another unit is furnishing as an output, and needs it as quickly as possible. Regardless of the concrete form of this transfer–a network may well contain sending and event-driven system services–the relation as seen from the user's viewpoint corresponds to a get relation.

Relations of this type lead to low-error-rate structures that are easy to reconstruct in dynamic fashion. The fact of access without feedback means that get connections can be made without further safeguards.

Within the component, a get connection corresponds to a read access of an output variable and, accordingly, can be implemented efficiently.

Many process control systems permit the treatment of get connections as objects in the local components, or at least get connections internal to a component. Their management and general implementation are taken care of by the operating system.

Figure 275 shows a *send* relation. Here one application has a functional interest in communicating something to another application. This would be the case, for example, when higher-level units transmit orders to lower-level units; where monitoring units have the task of automatically generating messages and alarms; or where operator control units wish to send a new control value to an application.

Figure 275. Send relationship

The attribute send here does not refer to the type of transmission; it represents the fact that the initiative and interest in data exchange is primarily on the part of the sender.

A send relation is characterized by the following features:

- The application behind connector C knows under what conditions C is to send what information to whom and activates connector C accordingly.
- The message from C leads to a write access in D; that is, the state of D changes.
- The associated service in D must therefore have a protective mechanism that receives the message, checks it, and performs the write only after successful checking. The protective mechanism checks both sender and content.
- To avoid unpredictable response times, the server service responsible for checking the message and processing it on receipt must be initiated as soon as the message arrives.

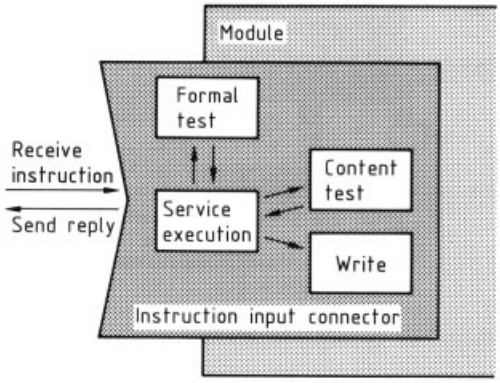

Figure 276. Instruction input connector

By way of example, Figure 276 shows an instruction input connector. When a message containing a new instruction arrives, the availability of the connector is first checked. If the connec-

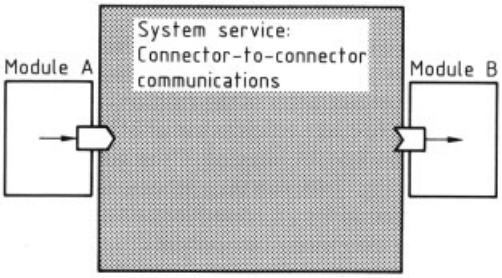

Figure 277. Communications as a system service

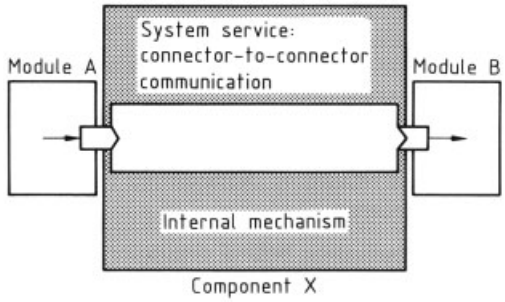

Figure 279. Communications internal to a component

tor is available, it is then occupied. The message is next checked for formal correctness (type of message, variable types, etc.). If the result is positive, the message is correct. This does not mean that the instruction can also be accepted by the module in terms of its content. The receiving service therefore has a link to a part of the module method specially allotted to the connector. This checks whether the content of the instruction is permissible from the application viewpoint. If the result is positive, the instruction is finally accepted and written as a new effective instruction. The sender of the instruction is notified of the acceptance, or the reasons for nonacceptance, of the instruction.

Most process control operating systems also permit unprotected writing into effective module input areas. Some users make intensive use of this capability, transforming the process control system into a quasi-procedure-oriented system. This is not, however, in keeping with "good programming practice" and should be avoided whenever possible.

A process control operating system must have the goal, as illustrated in Figure 277, of managing all communications relations between connectors in the modular network as objects and embodying their implementation, as seen from the user's viewpoint, in a system service. Many process control systems already do this within a local component, at least for relations internal to the component. The critical relations, however, are those between connectors belonging to modules implemented in different components. Here the user must explicitly design (configure) the communication in many cases. For example, it may be necessary to integrate and wire certain LAN communications modules. The message orientation and the strict formalization mean, however, that the connector model will later enable the operating system to manage and execute connector-to-connector relations over the LAN.

As a consequence of the special two-level module → connector object structure and the client/server character of the relations defined, itis possible to make efficient use of MMI services (see Section 8.5) for imaging. Open communications (see Fig. 278) can thus become a basic feature integrated into a communications network extending over many components.

Connector-to-connector communications can thus be organized as a network-wide mes-

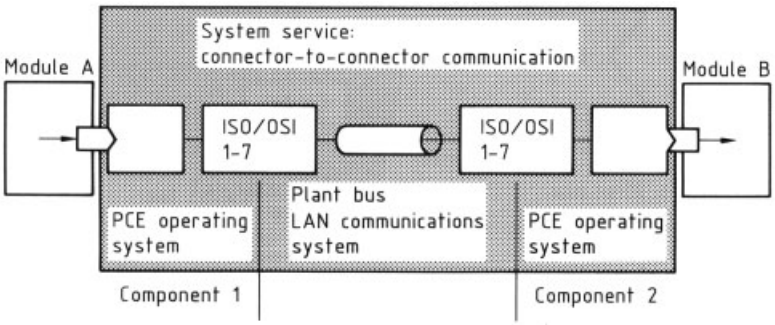

Figure 278. Connector-to-connector communications via the LAN

sage exchange system. A crucial point is that the simple local relations, with their efficient execution methods, are incorporated as an integral part of this overall system. If, as shown in Figure 279 (see next page), two applications are located in the same component, the process control operating system can implement connector-to-connector communications with optimized local access functions.

This technique can greatly increase the processing power of the execution system.

At the process (local) level, "get" relations within a component are the most frequent type of communications and are crucial for the loading of the communications network.

If all these relations were executed through a general message exchange system, the capacities of the functional units would be far too small. This level of performance can only be achieved by effective use of internal pointers and direct reading, both organized by the operating system.

Addressing. The objects and name spaces contained in a process control system depend on the organizational viewpoint and scheme. It is necessary, for example, to distinguish between the hardware viewpoint (component, module, unit), the access viewpoint (bus segment, subscriber, module, unit, connector, variable), the open communications viewpoint (VMD, domain, variable), the programming viewpoint (unit, connector), and the engineering viewpoints relating to the structure of the process plant, the breakdown of the process to be performed, or the structure of the production sequence.

Each way of analyzing the system has its own hierarchical name space, in terms of which the individual objects in the system so analyzed can be addressed. There are many mappings and assignments among the objects in various system breakdowns, and hence many object management structures and addressing options. These are not presented in detail here.

A question that is of interest, however, is addressing for LAN communications in the modular network. Most process control operating systems have not integrated communication via LAN into their module communications, and, therefore, either special communications functions must be installed (READ, WRITE, SEND, RECEIVE, . . .) or communications must be structured with direct, unprotected read or write access to the module variables.

Figure 280 shows one possible addressing scheme in the modular approach. The scheme is shaped by three aspects, which must be mapped into one another.

Addressing in the client is done from the programming viewpoint. The client connector wishes to make a connection to a certain server connector. The server connector is identified by its name and the name of the module (unit).

The process control operating system (OS) recognizes that this unit is located in another VMD and maps the requirements onto an MMI function. The unit is thus mapped into an MMI domain and the connector into an MMI variable in 1:1 fashion.

In the target VMD, the process control OS performs the MMI server functions and exchanges data with the server connector. The OS

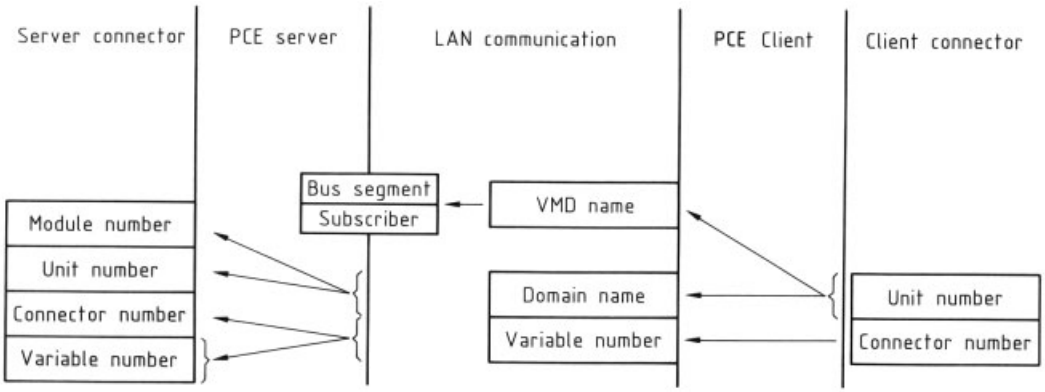

Figure 280. Example of connector addressing via LAN with names

uses its internal access addressing scheme, for example the module number, unit number, connector number, and individual variable numbers.

This is just one possibility. There are others, such as access through abbreviated addresses. The important point for effective open communications is that modules (units) and connectors can be mapped 1:1 into addressable communications objects.

Execution Properties. An important task of the process control operating system is to control the sequence of module processing. In general, a module consists of a main method and various connector-specific methods.

The main method is started by a call to the module by the executive system. In process control systems, calls typically come in cycles at fixed time intervals. Each module can be allotted such a fixed cycle for itself. Cyclic method calling is independent of whether or not changes of state have taken place at the input connectors. Cycles of between 1 and 4 seconds are typical in process engineering.

Connector-specific methods are permanently allocated to the input connectors. They are started in acyclic mode when messages arrive at the respective input connectors.

Asynchronous Operation. The use of fixed cycles generates deterministic sequences between certain modules. Even so, the modular network as a whole must be regarded as a network of modules in asynchronous operation. The processing of a module can be interrupted, for example, by the processing of a higher-priority cycle sequence or by connector-specific methods started in acyclic mode. Furthermore, if module processing is shared among several processors, cycles of equal priority also run asynchronously.

Asynchronous operation results in data consistency problems:

- While one module is running, the outputs of other modules may change
- While one module is running, processing of another module may be interrupted, so that its outputs represent results from different cycles

These problems are not generally relevant. The output states requested must be as current as possible. Inconsistencies are not a difficulty for the processing algorithms. Applications requirements, for example in module type creation, may call for appropriate measures to insure consistency in certain cases.

The efficiency-oriented concept of process control operating systems manifests itself again here: avoidance of unnecessary management overhead.

Synchronization methods are available to insure the consistency of output and input data. However, they impose a substantial burden on the executive system and markedly impair the performance of the local components. Most process control operating systems purposely dispense with system measures to impose consistency. Note that all the devices in a discrete control system also function in asynchronous mode without consistency-insuring measures.

7.4. General System Services

A control system is expected to offer, as an integral feature, certain general system services for production operations. These should be implemented as firmware in the process control operating system (OS).

The pertinent services include the following:

- Communications and organization services for operation, monitoring, the generation of messages and alarms, archiving, and report creation
- Time management and network synchronization services
- Monitoring and diagnostic services for process control hardware
- Services supporting the changeover between redundant units and the matching of their states

Integration of these services into the process control OS by the manufacturer is necessary because only in this way can they be implemented in a meaningful way and performed with the necessary effectiveness.

Unfortunately, there are no standards governing the functionalities of these system services; the available, manufacturer-specific solutions differ all the way back to their principles. The user must consider the structure of these manufacturer-specific system services and fit the applications software to it.

A manufacturer-neutral design procedure is therefore possible only within narrow limits. In principle, this design approach is restricted to the configuration of certain processing functions with standard languages such as FUP, module language, and so forth.

This part of the configuration job often makes up only 40 % of the whole. The remaining configuration work must deal with configuring and parametrizing manufacturer-specific system services.

This section will describe in detail the tasks and contents of the system services. Each service is treated as an individual functional module.

Monitoring (see also Section 11.2). Processes are generally monitored by using color graphics displays. The interface can be based on standard solutions from the PC or workstation realm. The monitoring services provide the monitor interface with the appropriate video and status information. One monitoring service includes transmission of the state value from its source to the monitoring system, maintaining an internal state map (state or status image) in the monitoring system, and displaying the state value at the interface. Figure 281 shows schematically a monitoring service.

The monitoring system always displays the current state values. If a state is dynamically changing, it is important to display its most recent value. Past values and timestamps are not of interest. On arrival, each new value overwrites the old value on the screen.

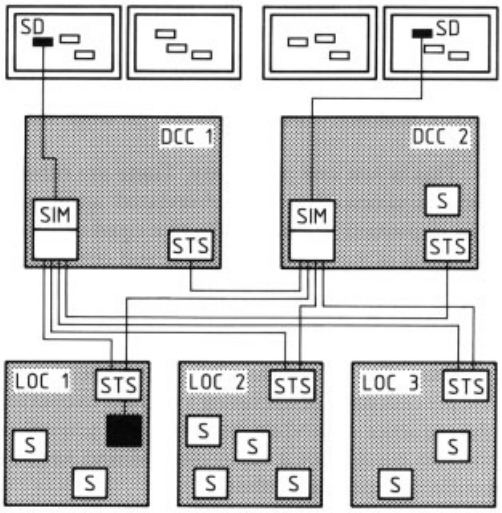

Figure 281. Schematic diagram of a monitoring service
S = state; SD = state display; SIM = state image; STS = state transmission server; DCC = display and control component; LOC = local component

All monitoring interfaces are implemented by the monitoring system. The same holds for the display of archival data, messages, records, and so forth.

When the message log is displayed, for example, the message system is the source of the states displayed. The timestamp of a message is a state of this system and is retrieved and displayed by the monitoring system in the same way as any other state.

State Displays. A state can be displayed in a variety of ways, and standard interfaces offer all the possibilities. It is therefore a fundamental question of human – machine or human – process communication design how a state is represented in each particular case.

Figure 282 presents examples of ways of depicting a binary state. Often, a binary state is represented by a color change in a text field or by the presence of a mark. In flowsheets, the process state that the binary state represents is indicated symbolically. For example, a curved arrow or a color applied to the valve symbol can be used to indicate whether a valve is open.

Figure 283 shows several representations for an integer or floating-point value. Here, standard systems offer a wide range of easily implemented display options (see Figs. 372 – 380).

Maintenance of State Image. The various ways of maintaining the state image in the monitoring system give rise to basic differences in process control system data models.

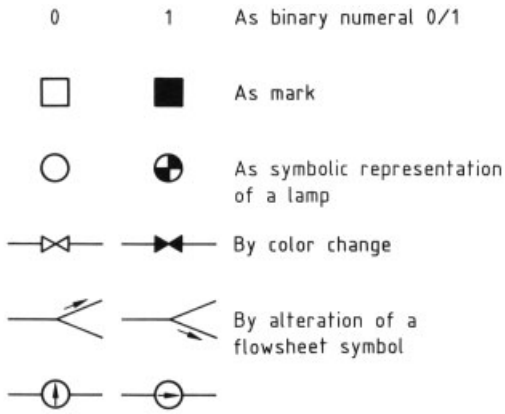

Figure 282. Representations of binary states

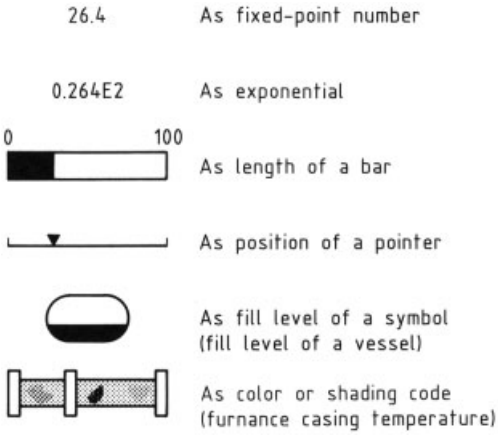

Figure 283. Representations of a floating-point value

Important goals in managing and maintaining the state image (map) in the monitoring system include the following:

- Keeping the displayed states updated: The current value of a source state should be displayed within one second if possible (or at least within two seconds in exceptional cases).
- Rapid preparation of a new image with the proper state values: When a new image is selected, the image together with all updated states should be constructed within 2 s (within 4 s in exceptional cases).
- Minimizing the data transfer cost: In most situations, the state sources are not located in the same display and control component as the monitoring function itself. The required states must then be transferred via the bus. In large systems in particular, or when the depth of detail is great, measures must be taken to reduce the amount of data transferred.
- Minimizing the cost of preparing the state information in the source system: Some effort must be expended in the source system to prepare the state data and organize the data transfer. This extra work must be considered from the outset when calculating the load on the source system. Problems arise, in particular, when several display and control components are used, each with its monitoring interface and all accessing states of a relatively small local component, or when a change-driven data transfer service is implemented.
- Modular organization: The state image maintenance procedure must be set up in such a way that new images can easily be inserted in a monitoring system and an additional display and control component, with its own monitoring system, can be connected to the system communications at any time.

The simplest way to organize and maintain the state image for monitoring is by using a 100 % image in the display and control components (DCCs). All states that might ever be required for monitoring are held in every DCC at all times. This solution, which is practicable in small one-station systems, leads to performance problems in large multistation setups. If the possibility of insetting detailed state information is not to be lost, data transfer must be optimized, for example by selective or change-driven transmission.

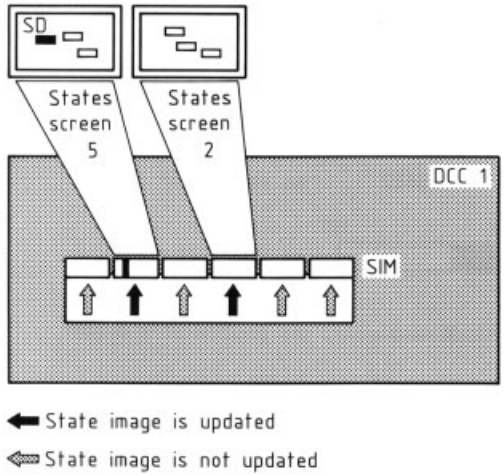

Figure 284. Selective state display
SD = state depiction; DCC = display and control component; SIM = state image

In selective transmission (Fig. 284), only those states that are currently displayed are updated in the state image. The data-transfer load thus depends on the scope of the images currently displayed, not on the scope of all images configured. The image can be carried to any depth, down to the smallest details, without any change in the preparation and transmission load. Selecting a new image means a change in the set of states to be updated. This must also be reported to the source systems and new arrangements made there. The process control operating systems of the components involved must there-

fore include an effective service for this purpose. Process control systems with the option of selective state images can generally be upgraded to large multi-workstation installations without difficulty.

The second possibility is change-driven transmission. In normal operation, many states undergo changes infrequently or not at all. This leads to the possibility of transmitting only when a state has changed significantly. The number of data to be transmitted under normal conditions is drastically decreased. The detection of state changes in the source system presents a problem. Any change of state in a binary value is significant, but in the case of an integer or floating-point value, a change is only significant if it exceeds a certain set distance from the old value. Detecting deviations leads to a perceptible load on the source system. A further drawback to change-driven transmission is that load fluctuations are difficult to calculate. Large numbers of switching operations take place in the plant precisely under critical conditions, so that there are simultaneous changes in nearly all process parameters. A change-driven system must be designed for these extreme peak loads.

Operator control refers to the specification of values by an operator via the system's own human – machine interface. The following units are involved:

- The operator, who communicates with the operator control system through the input device.
- The operator control system, which handles the selection and execution of instructions.
- The instruction acceptor, which checks the value entered. If the result is positive, the value is accepted and furnished as a new effective value to the device addressed.

Each operator console has its own operator control system, which functions independently of the others. The topology of the operator control system is shown in Figure 285.

An operator intervention is generally distinguished from a design intervention. The plant operator operates while the design engineer designs. The distinction is that an operator control action always communicates with an instruction acceptor (as an existing entity). Design (configuration) functions define, delete, and manipulate all possible types of objects. Linguistic usage aside, the actions of the designer can also be understood as "operating." Design functions are initiated and, on the basis of the operator instructions, cause appropriate object manipulations to be performed.

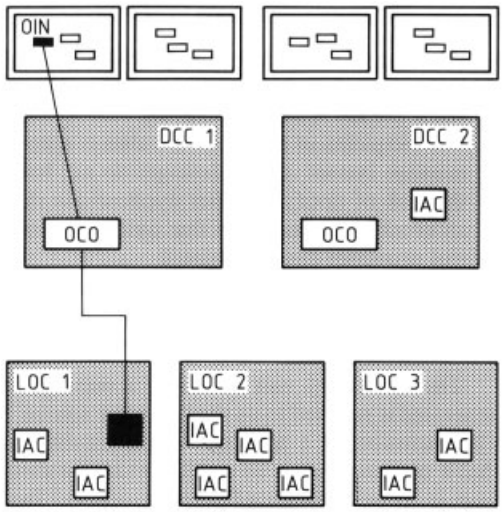

Figure 285. Structure of the operator control system
OIN = operator input; OCO = operator control; IAC = instruction acceptor; DCC = display and control component; LOC = local component

The sequence of events in a typical operator action is as follows:

Selection. The operator action is generally initiated from an operator window, which provides the operator with all information needed for operation, such as the current value of the variable to be manipulated, local enabling of operator control, control limits to be maintained, or process states relevant for operation. The structure of the window and the information furnished in it are governed by monitoring functions. An operator window should be openable regardless of operator authorization. The variables that can be manipulated by the operator are indicated in the window (e.g., shading as in a pull-down menu system).

The first step in the actual operator action takes place when the operator selects a variable or variable group for manipulation. The operator control system checks whether the selected variable is manipulable and whether the operator is permitted to manipulate it. Several options are

available for the nature and syntax of the selection. It can be effected, for example, by entering a command on the keyboard, or by clicking on a symbol marking the variable. Once a selection has been made, the operator control system is in instruction mode. Now the operator can identify and manipulate one of the selected variables or can terminate the control action. In general, concurrent selections of an instruction acceptor are permitted. The instruction acceptor does not know whether it is selected or not in an operator control system. The instruction acceptor is not occupied by being selected. The same variable may be concurrently selected for manipulation in several operator control systems.

Input. The control value is input in a specified input syntax. A wide variety of possibilities exist, depending on the type of input device, the type of value, and the control philosophy employed. Some possibilities are as follows:

- Input the value on the keyboard and press ENTER
- Click on the appropriate symbol to change a binary value
- Simultaneously press the command key and the release key (two-hand operation)
- Press an UP or DOWN arrow key to increase or decrease an analog value
- Move a slider

All these actions ultimately constitute a control instruction in the operator control system. Once the input is complete, the instruction is sent directly to the instruction acceptor concerned.

Control Instruction. The operator control system transmits the control instruction to the instruction acceptor, which may be located in the same component as the operator control system or in another component of the system. If the operator control system does not succeed in properly transmitting the instruction to the instruction acceptor, the manipulation is considered not possible or rejected. If the instruction is properly transmitted to the instruction acceptor, the acceptor begins to process its methods. First it checks whether the manipulation is permitted from the standpoint of the application. During the processing of these acceptor methods, no other control instruction can be transmitted to the acceptor. Although such a collision is extremely unlikely, the process control OS must exclude this possibility.

Processing of Instruction Acceptor Methods. Data security in a process monitoring and control system requires that the instruction acceptor have its own methods, which it uses to check whether it will accept a control instruction or not. System solutions in which the effective value in the instruction acceptor is directly overwritten by a control action should be avoided.

In the context of the connector model, an instruction acceptor is an input connector of a write connection. The events at the instruction acceptor can be described as follows in this model: The connector consists of two parts, a system unit and an application unit. The control instruction is first received by the system unit, which checks whether the connector is enabled and whether the instruction is formally correct. The system unit then starts the application-specific method component of the instruction acceptor. In this section, customized for the application, a check is made to determine whether the specified value is or is not permitted on the basis of the instantaneous state of the application. If the value is acceptable, it is taken over as the new effective value; if not, the old effective value is retained. The execution of these steps is recorded in a message stating that a manipulation was attempted and whether it succeeded or failed.

Reaction on the Part of the Process. The operator control function described here is strictly limited to the transmission of the control instruction. The consequences of such an instruction have nothing to do with the control function itself. Once the control instruction has been entered in the instruction acceptor, the control action is terminated. The process consequences of the control instruction can be tracked by, for example, the monitoring system.

An important point for the operation of a process control system is that the operator always be notified of the acceptance or nonacceptance of a manipulation. A control instruction must therefore be processed and passed on to the instruction acceptor immediately after its input. Processing in the component that houses the acceptor must be interrupted so that the checking methods of the instruction acceptor can be processed. The total time between the end of input and the report from the acceptor must be no more than 2 s, or 4 s in the worst case.

In order to maintain this timing, many systems simply acknowledge the input directly

from the operator control system, only supplying it to the instruction acceptor later. Such solutions should always be rejected. They give the operator a false picture of the facts, and if difficulties arise, they result in a loss of confidence in the correctness of the status display.

Message Handling. "Messages", in general terms, are explicit reports on important events. Not every change of state in the system leads to a message.

Among the purposes served by messages are the following:

- To inform the plant operator about events in the process
- To provide the maintenance personnel with information about malfunctions and changes of operating status in the plant
- To furnish the production manager with key data on the production process
- To document the process in the plant

The following restrictions apply:

- Messages do not serve to replace records of status, switching logs, and detailed event reports needed for event analysis.
- The recipients of messages are always special message system units. Messages do not have the task of notifying other applications that an event has occurred so that automatic control processes can be initiated.
- The message functions do not update the state image in the display and control components during startup or operation.

The functional unit in which a message is generated is referred to as the "message source". The inclusion of a message source establishes that a certain event is so significant that a message is to be transmitted when it occurs. Two distinct concepts are employed: decentralized (distributed) and central message generation.

Decentralized Message Generation. In the decentralized message generation, the message sources are built into the lowermost functional level at which the significance of a signal can be determined. Message sources can be integrated directly into the process functional units. Examples include the following:

- Malfunction of a motor interlock: Interlock malfunctions are generally classified and reported as significant events. A malfunction of the interlock is detected in the motor control and can be reported directly from this unit. In other words, the functional unit "motor" acts automatically and independently to generate an appropriate message.
- Channel failure in a module: Channel failures are always reported as control hardware malfunctions. An intelligent module can independently generate a message if such a malfunction occurs.
- Arrival at a mark in a recipe sequence: When a significant mark is reached in a recipe operation (e.g., "preliminary reaction completed"), the recipe control unit can automatically generate an operational message.

Under a concept of this type, messages are created in all local functional units as well as in various modules and components. The messages must first be conveyed securely to the respective message receivers. In the distributed concept–and this is an important advantage–message transportation can be effected by system services.

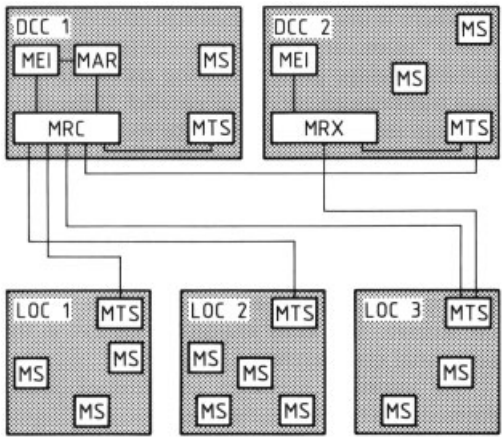

Figure 286. Structure of the message system in a system with decentralized message generation
DCC = display and control component; LOC = local component; MAR = message archive; MEI = message interpretation; MRC = message receiver (central unit); MRX = message receiver (auxiliary unit); MS = message source; MTS = message transmission receiver

Figure 286 illustrates one configuration. Messages are always generated in message sources and locally delivered to a message transmission server. Each logical unit has its own local message transmission server. The task of this

device is to collect the messages from a component, hold them in interim (buffer) storage, and transmit them by secure means to designated message receiving and archiving units.

Centralized Message Generation. A concept based on centralized message generation is shown in Figure 287. Messages are generated exclusively in the central component, where message generation, message archiving, and archive interpretation are generally combined.

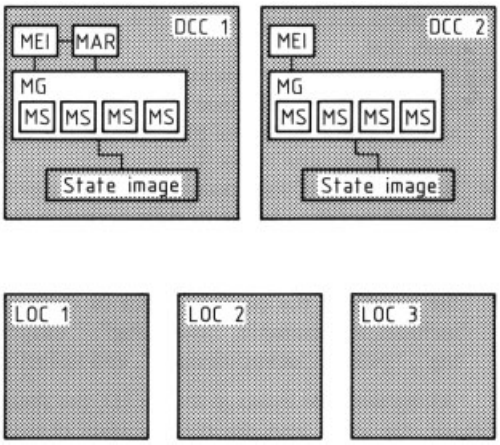

Figure 287. Structure of the message system in a system with centralized message generation
DCC = display and control component; LOC = local component; MAR = message archive; MEI = message interpretation; MG = message generation; MS = message source

The centralized concept is based on a different information model from the decentralized one. In the distributed concept, not all events must be reported to the central units via system communications. For example, a change of state of the control valve of a two-position controller is a normal operational event of secondary importance. This switching operation need not be reported and documented as an event. In the decentralized concept, reporting of the state (instantaneous valve position) for representation in an image is handled by services independent of message transmission.

The centralized concept links state and event reporting. The central component usually maintains a 100 % image of all required states for this purpose. When an event occurs in a central component, the state image is updated to reflect the state affected by the event. A check is run in parallel fashion to determine whether the event is one that must be reported. One problem in the centralized message concept is the management of information between multiple central stations. For every message, there must be no more than one message source in the entire system. The acknowledgement status is managed at this message source. Regardless of where it takes place, every acknowledgement must lead to resetting of the acknowledgement request, so that the acknowledgement is central for the entire system. In this respect, many control systems with centralized message generation (in particular SPC-based systems) do not have a clean structure. They do not support the sharing of message functions among central components and do not offer the possibility of central acknowledgement.

Content of a Message. The following information must be contained in a message:

- Who sent the message? Each message contains the identification of the message source and of the process unit in which the event being reported took place.
- When was the message sent? Every message contains a timestamp giving the time when the event reported took place.
- What was reported? Every message contains a precise description of its meaning. This description can be a plain text or a pointer to a plain text.
- How important is the message? Each message has a certain significance class, for example alarms, warnings, out of tolerance, sequential event report, and so forth.
- What change of state occurred? A message is sent whenever a change of state occurs in the message source. In general, a message contains the state in effect after the change and the nature of the change. Types of changes include, for example, “incoming”, “outgoing”, “pulse”, “acknowledged”, and so forth.

Along with the message source ID, the timestamp, the state, and the type of state change must be entered as instantaneous values at the moment when the message is created. Other information can be appended to the message later. The key here is the message source ID.

Message Transmission and Message Distribution. The task of the message transmission server is to receive, buffer, distribute, and transmit messages to the proper receiving and inter-

May 28, 1991 Sequential event display, Group V13RA4
16:54:05

Time	Unit	Text	Class	Type
9:41:23	V13RA4T3	temperature high	A	incoming
9:41:33	V13RA4T3	temperature high	B	acknowledge
9:43:22	V13RA4L2	level high	W	incoming
9:43:31	V13RA4L2	level high	B	acknowledge
9:44:38	V13RA4T3	temperature high	A	outgoing
9:44:46	V13RA4T3	temperature high	B	acknowledge
9:46:65	V13RA4L2	level high	W	outgoing
9:47:41	V13RA4F6	flow low	A	incoming
9:47:53	V13RA4F6	flow low	B	acknowledge
14:34:23	V13RA4K1	interlock malfunction	A	incoming
16:33:06	V13RA4F7	invalid measurement	A	incoming
16:53:53	V13RA4S1	please confirm recipe data	I	incoming

Figure 288. Sequential event display

pretation units (see Fig. 286). Message transmission servers must therefore insure that no message gets lost, even during a flurry of messages.

Message distribution can be described by the following general picture: The message receiving unit logs on to the message transmission servers every time a message is received.

There is one "main" message receiving unit. Other "secondary" message receiving units, for certain message groups, can log on to the message transmission servers in parallel. If the main message receiving unit fails, the message transmission server attempts to buffer messages until a replacement unit has logged on as the new main receiving unit.

Message Receiving and Archive Units. The task of a message receiving unit is to receive incoming messages and enter them in the logged-on archive. The most important archive is the sequential event archive.

All incoming messages are chronologically entered into the sequential event archive. Figure 288 shows a sequential event display, in this case an excerpt from the last messages entered in the archive. From the sequential event archive, messages can be grouped under a wide variety of criteria, for example sorted by charges, plant sections, significance classes, time intervals, and so forth, or by combinations of such criteria (see Fig. 65).

Besides the sequential event archive, there is usually a list containing all message sources that have generated current messages. Figure 289 illustrates such a list. A message source appears on this list (it is said to be "signaled") if either the event-generating bit is still set (value still out of range) or acknowledgement has not taken place yet.

Archiving. If processes are to be studied in retrospect, certain events and the time variations of certain states in the system must be archived. The resulting data can then be accessed by interpretive functions. The outcome of an interpretation can be output in the form of, for example, plots or lists. In modern systems, a strict distinction is observed between the data stored in the archive, the interpretive function per se, and the form and representation of the printout. While data storage in the archive is structured in keeping with considerations internal to the system, the user can define what the interpretive functions are to do and in what form the reports are to be output. Archiving presents a number of problems that can be solved only with the aid of special system services. Examples are:

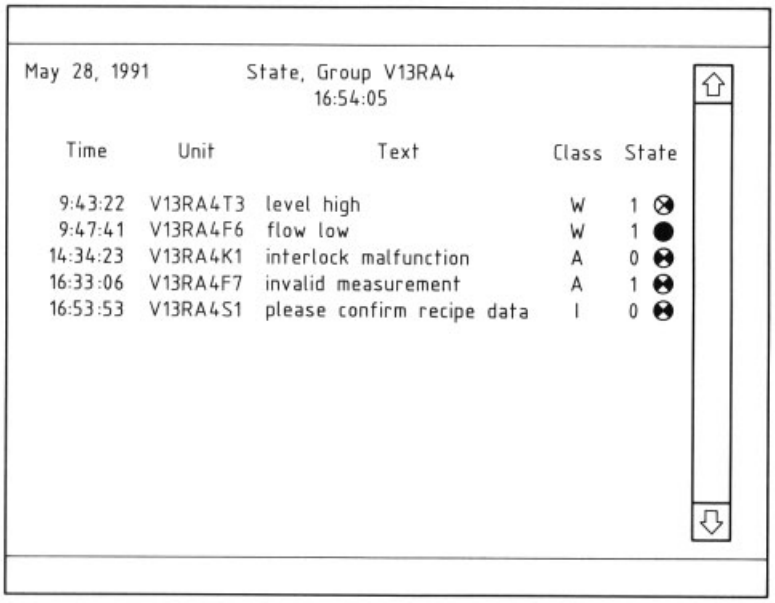

Figure 289. List of message-originating sources

- To append on-line process data quickly and reliably
- To carry out informational functions in parallel with the continuous addition of new data
- To safeguard the data automatically
- To implement the archiving functions in redundant fashion (see Section 3.2.)

Time Synchronization. Time (clock) synchronization is an essential system function in a distributed process control system. If events are to be listed in time order, the messages must be timestamped as near as possible to their place of origin. This may be in a local component or may be in a module where, for example, a control hardware malfunction is detected. A sequence of two events in two distinct system units can be correctly detected only if the clocks in the two units "tell the same time". Synchronization insures that the time discrepancy over all clocks in the system will fall within a narrow time interval. A typical value is 10 ms; specially equipped systems achieve discrepancies of < 2 ms.

The time is absolutely specified by a central system clock. Additionally, radio clocks are employed as system clocks, with the advantage that time changes (e.g., from standard to daylight-saving time) are performed automatically.

Process Control Hardware Diagnostics. The diagnostic functions monitor the operation of the control hardware and include tools for analyzing the state of the monitoring and control equipment. System diagnostic functions fall into three groups.

The first group includes *control hardware monitoring functions*, which are integral constituents of the individual control system modules and functional units. These functions run automatically and continuously. The result of each check is written into the disturbance status of the unit concerned and can be read from there. If the trouble status of a unit changes, the unit automatically sends a hardware message. Hardware messages get special treatment in many systems, although this is not really necessary. That a message is a hardware message can easily be determined, either directly by examining the message source ID or indirectly by means of a special flag in the message. In this way, control system hardware messages can be treated in the same way as process messages. Typical hardware messages include:

- Failure of an I/O channel (alarm)
- Failure of a redundant bus connection (warning)

The second group takes in *diagnostic functions* of interest to the *designer* or to the on-site

technician. These have to do with special system analyses that are initiated by the operator only as needed. The functions are supported, in many cases, by external diagnostic equipment. Examples from this group are functions for analyzing the following (and comparable functions):

- The current configuration
- Memory utilization and processor load
- Run-time measurements

The third group consists of *diagnostic functions* used only by specialists at the *manufacturing firm* which permit detailed analysis of system behavior.

7.5. Design and Commissioning

Before a plant is operated, it must be designed and commissioned. The engineering work necessary for this purpose is a key item in the total cost of a process control system. The distribution of the work varies widely. In a decentralized process control system with a well-designed OS and the proper design and commissioning support, pure system procurement costs are relatively high but design and commissioning costs are low. In a "cheap" system (most SPC-based systems fall into this category), the hardware is favorably priced but design and commissioning costs are substantial.

The discussion in this section is based on the structural model of design and commissioning shown in Figure 290.

Design. The design process begins with the technological task statement. In practice, however, all that is usually available is a piping and instrumentation diagram. Informative documentation on the functional relationships is either nonexistent or, for the project team at least, difficult to access. Functional data are often found scattered throughout the process development reports or in descriptions provided by the equipment fitter. Accordingly, one of the most important tasks for the project team is first to gather together all such information, develop the functions, and describe them in a formalized way in the project specification (see Chaps. 2 and 10). This formal summary may well take into account the structure of the later implementation. For process control, this means that the specification takes the form of not a list of requirements, but a role system with certain functional properties. This role system can be refined stepwise and ultimately can be employed as the basis for the implementation. Thus the process control specification can be expanded later, thus becoming the documentation of the process control system.

An important advantage of this approach is the systematic procedure. Each function discussed in project conferences can be directly incorporated into the role system as a property. As a result, efficient design work with few errors is possible, and the operational aspect of the procedure is implementation-oriented.

In what follows, the design steps required in such a procedure are presented in detail in terms of the structural model of Figure 290.

Definition of Technological Module Types Required. The definition of a suitable collection of module types generally precedes the design work. Industry and corporate standards apply in many cases. The GMA is now working to define a standard set of module types, and the IEC has recently taken up this project.

The following criteria must be considered in establishing the module types:

Manageable functional scope.

Functional modules should be neither too small nor too large in terms of functional scope.

Small modules permit very flexible structuring of functionality and thus can be used very generally. The disadvantage is the large number of small modules and connections required, resulting in increased design and object management costs. The error rate increases, and the efficiency of processing falls off drastically when the system is later mapped onto the functional system. Another drawback is the low degree of functional standardization that inevitably results from the element types defined.

Large modules permit fairly standardized structuring with a few modules. Large modules can be efficiently processed in the later mapping in the functional system. The disadvantages of large modules are the cost of defining them, their functional rigidity, and the possibility of an excessively broad spectrum of models. If large modules are to be used, a consensus must be found as to their functionality. Because these modules embody technological standards, this is a much costlier operation than for generic small modules. To cover all technological cases,

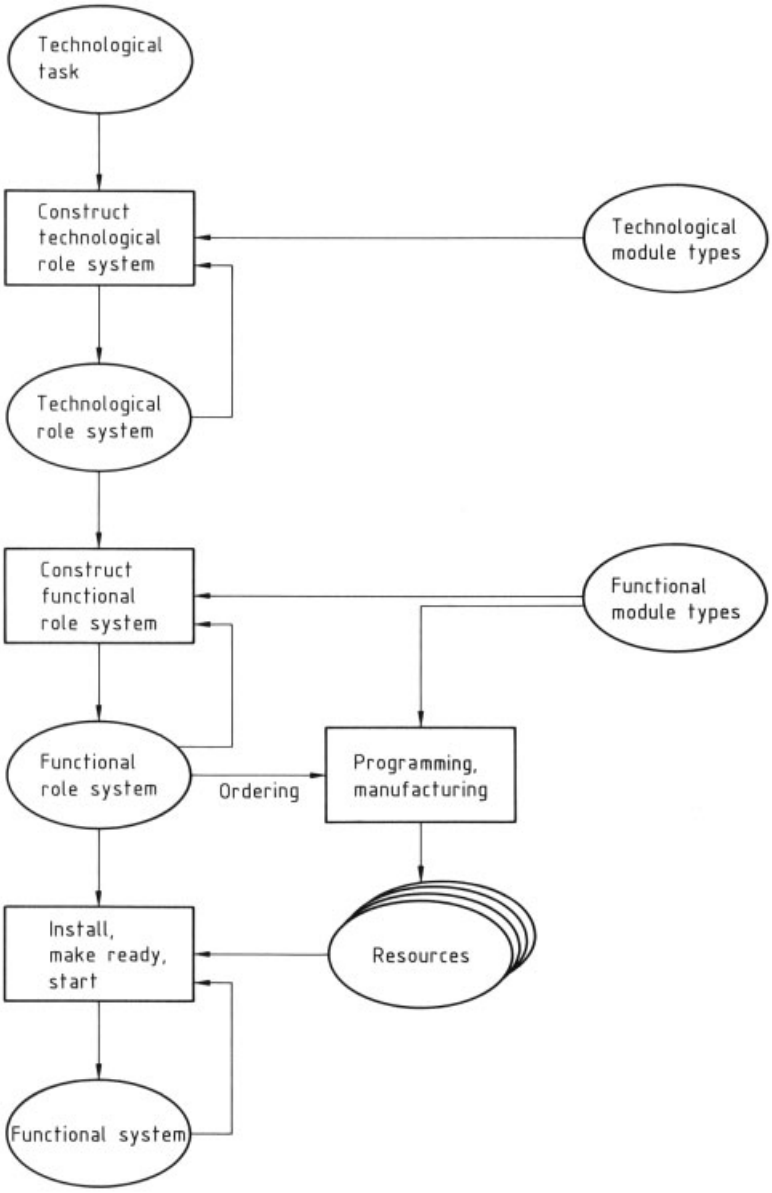

Figure 290. Structural model of system design and commissioning

either the modules must be made very extensive, with many selectable and parametrizable options, or a great number of types must be defined for fairly similar technological tasks.(What functions must a motor control module contain, and how many distinct motor control modules are desirable?)

Feasibility of using functional module types. It makes little sense to define technological module types such that their later mapping onto functional module types is difficult or impossible. It is therefore essential to check, while fixing the technological module types, how they can be realized in various systems.

Configuration of the Technological Role System. From the spectrum of types, the required

modules are identified, instantiated, connected, and parametrized. The result is a technological role system in modular form. In accordance with a structure such as the hierarchical structure of Section 4.5, this role system is broken down into individual function, group function, plant coordination, action, and other levels. This technological role system comprehensively and formally reflects all requirements derived from the technological task statement.

The documentation of the technological role system can be regarded as a formalized specification. In addition, it directly serves as the basic documentation for further design work specific to the process control system. The technological role system itself should be formulated in a manner largely independent of the control system (see also Chaps. 2 and 10).

Definition of Functional Module Types. Along with the firmware or special-purpose software, most manufacturers of process control systems supply a collection of functional module types. If the user remains with these prefabricated types, the complicated procedure of creating customer software module types can often be dispensed with. Indeed, many process control systems do not allow user programming of software modules.

Thus it is chiefly the manufacturer that defines the functional module types. The user must manage with these manufacturer-specific collections of types. Only in special cases can a competent user define and program new sets of types; such a user will attempt whenever possible to create software module types that exhibit a 1:1 correspondence with the functionalities of the technological module types used. Otherwise, the manufacturer must attempt to satisfy user requirements.

Configuration of the Functional Role System. The functions embodied in the technological role system are transformed into software structures specific to the control system. In this process, the structures of the technological role system are preserved in unchanged form as far as possible. A frequently used aid is the creation of"typicals."

Typicals are pre-assembled structures made up of functional module types. They are not types but structures having special naming conventions and which can be easily copied in the design interface. Typicals can readily be put together by the user. If typicals are created that functionally correspond to the technological module types, the functional role system can easily be configured with the aid of these typicals.

The configuration of the functional role system falls into two parts:

- Mapping the technological structure onto the functional structure specific to the control system
- Establishing the locations of individual functional units

"Location" here refers to the logical location of a unit. For a module, for example, this is the logical card slot; for a software module, it is the processor board in which the module is run.

Once the network of software modules has been defined, it is therefore necessary to establish which modules will be run on which processor boards. The component and module structure of the process monitoring and control system must be known, at least in outline, before this step can be done.

The result is the finally configured functional system. In a design system, functional units are handled as objects; they can be represented and documented in accordance with certain criteria and be loaded directly into the functional unit for execution.

Commissioning (see also Sections 10.5 and 10.6). Commissioning includes all actions required to bring a system or system unit from an initial state to an active working state. A single functional unit, a group of functional units, a whole system component, or a complete system can be commissioned. In defining what will be regarded as the initial state, one commonly takes an organizational (management) view, defining the initial state to be whatever state the commissioner finds when the operation begins. In a process control system, for example, the initial state is the system fully installed but not yet energized. When a compact device is purchased, commissioning starts with unpacking. Certain simple installation tasks may thus be included in commissioning.

From a systems engineering standpoint, commissioning is the last step involved in making the functional system available for service. The functional system includes all subsystems

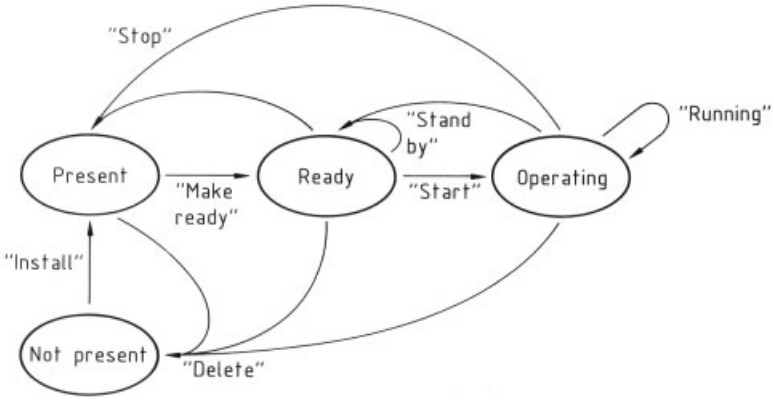

Figure 291. Availability states of a functional unit

needed for active implementation of production operation.

Availability of a Functional Unit. The functional system is the system that actively realizes production operation. Its units fulfill roles whose interplay results in the desired system functionality.

The separation of role systems and functional systems is of central importance for an understanding of commissioning.

A functional role system is set up in such a way that for every role there is a functional unit that plays this role. For a functional unit to carry out a role, it must be allocated to this role and must be made available at the defined location of the role. This part of the operation is called "installation".

Figure 291 illustrates the availability states of a functional unit along with the transitions between these states.

- Not present: There is no suitable functional unit at the location. For example, the card is not plugged in or the software module is not loaded.
- Install: Installation means making a functional unit available; for example, assembling a cabinet rack, plugging in a card, or loading a software module. There are two key events in installation: when the functional unit is physically present at the location, and when the functional unit has been configured for the task it is to perform. The order of these events is not significant here and can be determined case by case. For example, in installing a card, the jumpers are placed first and the card is then plugged in, whereas for installing a software module entity, the entity is created first and only then are values assigned for the parameters.
- Present: The functional unit is said to be present after installation; i.e., it is finally configured for its role and is available at the location designated for the role.
- Make ready: The state of the object after installation, for example of a newly loaded software module, is generally undefined. A special function called "make ready" must be carried out in order to bring the object to a well-defined "ready" state adapted to the instantaneous system state.
- Ready: The availability state "ready" means that the unit is not only present but also ready to start. In other words, at this instant all conditions exist for connecting the unit actively into the network. The "ready" state plays an important role specifically in distributed process control systems. Here the situation often arises where certain units have already been activated and begun carrying out their tasks. A further unit is now to be connected into this running structure. Naturally, such a connection must not be made except under certain conditions. The "ready" state indicates that these synchronization conditions exist at this moment and the unit can be connected into active operation.
- Start: From the "ready" state, a functional unit can be started. It then goes into the "operating" state.
- Operating: In this state, the unit is effective in the functional system as an active element. It is performing the role intended for it.

A typical situation in a distributed process control system is where the functional systems are operating in certain regions and not in others (see Fig. 292). The process control system is purposely designed so that functional units can be handled in a "granular" fashion. If a functional unit is not operating, no role that is performed by this unit can be realized. In the example shown, modules C and E are not operating; modules A, B, D, and F are operating. To the extent allowed by disrupted communications relations I, II, III, and IV, the operating modules are fulfilling the role intended for them.

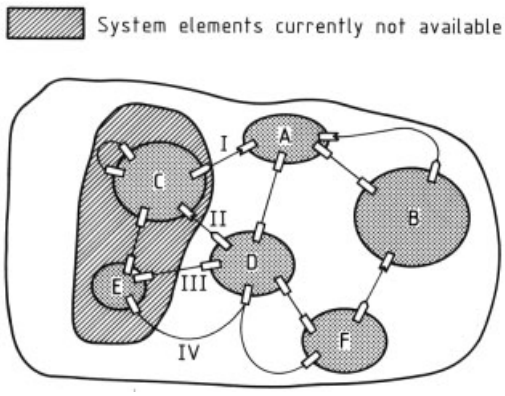

Figure 292. Functional system with several unavailable units

Hierarchy of the Functional System. The functional system of a control system should be regarded as having a multilevel structure. Each level satisfies the role requirements of the next higher level.

The structuring and implementation of a functional system are always carried from the bottom up, starting from the physical hardware units. This strict bottom-up principle underlies all commissioning and startup operations.

Figure 293 illustrates a typical functional system structure for control system hardware. Each functional unit can operate only when all functional units are operating in the lower levels on which it is built. If a functional unit fails, no functional unit in any higher level that requires the failed unit for the execution of its role will remain in the "operating" availability state.

- Hardware level: The hardware level furnishes devices, modules, displays, cables, and so forth as functional units. The lowest level of an industrial functional system is always the physical level. Without a physical basis, no industrial functional system can operate.
- Base system level: The base system level provides the base operating system, the required databases, and other standard tools (e.g., standard communications tools) as functional units. The control system functions are built up as actively available units from these standard tools.
- Process control level: The process control operating system level includes all functional units needed for the management and executive control of technological functional units. Examples from this class are system functions such as module processing, module-to-module communications, message generation, operator control, display, archiving, and so forth.

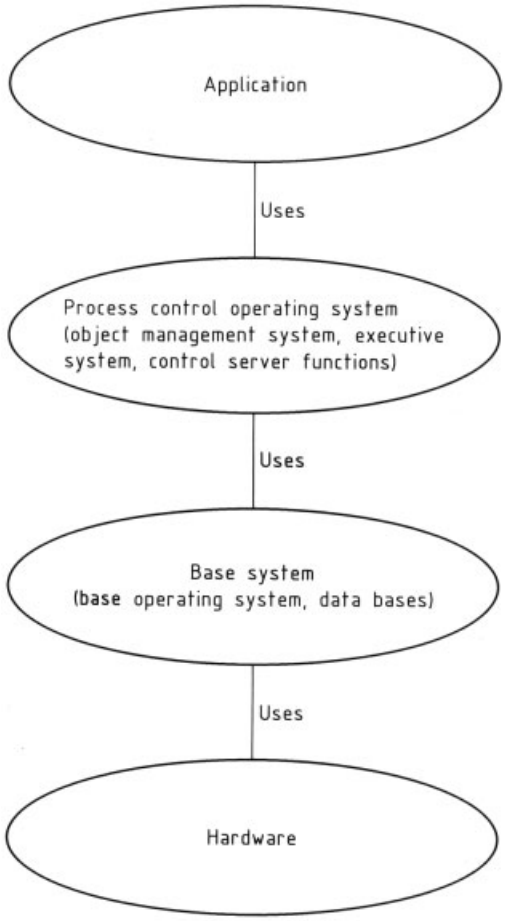

Figure 293. Hierarchical structure of the functional system

– Application level: The application level implements the process control functions on the basis of a functioning process control operating system. If process control is structured in a hierarchical fashion, the application level itself can be regarded as a hierarchical system of process control functions, one built on another. Commissioning of the process control system based on operating software modules is a special aspect that will not be discussed here. From the standpoint of the process monitoring and control hardware, the application level consists of a flat network of intercommunicating modules.

8. The Process Control System and Its Elements: Information Logistics

8.1. Principles

The need for information integration in production companies is generally acknowledged today, but this effort is subject to two kinds of limitations. The first are due to historical circumstances surrounding the introduction of computer-aided information processing ("island" solutions); the second have to do with the heterogeneity of the computer systems employed, which in turn stemmed from functional requirements. The starting point for solving the associated problems is the derivation of the requisite integration functions along with their scope; the process is aided by proper interface definitions and standards. Two worlds are involved here: the world of production engineering in the definition of what functions are needed, and the world of information and data processing in the implementation of these functions. Informatics is called on to create suitable methods and tools for the second range of problems. All experience to date shows that optimal solutions demand close cooperation between the two fields; this necessity represents a special challenge because of, among other factors, difficulties in understanding functional requirements and differences in terminology.

The new field of "information logistics" will contribute to an integrated view of methods and tools for information distribution and processing in production plants. By analogy with the logistics of materials, energy, or personnel, information logistics combines methods that are concerned with the modeling, distribution, storage, and retrieval of information. The new field must identify the interrelationships of these individual disciplines with regard to the tasks of information processing in production and, with the aid of tools, must put them to work in optimizing information-logistic systems. The methodological edifice of information logistics is not yet complete [591, 592], but separate disciplines do already exist in distinct form. One such is distribution – the "communication" of information within the company – which is treated in depth in this chapter.

8.2. Functional Structures and Information Flow in Production Companies

For a better understanding of what requirements manufacturing information systems must meet [593, 594], it is desirable to work out models of information flow in the company. If such information flows in production companies are examined more closely, an information hierarchy such as the one shown in Figure 294 is revealed [595]. The subdivision of production companies into various control and management levels arises from the way corporations are organized, and so the details will differ from one industry to another [596]. A consistent feature, however, is that the lowest level of the structure includes many single items of information about processes, while this information is more and more strongly condensed in the higher levels (all the way up to corporate management). Conversely, the objectives of the higher levels of management and control give rise to items of information about production steps or stages at the lower levels. Figure 294 shows that these processes can extend to various levels and take a variety of forms. Again, structural differences between industries and companies are a factor.

The subdivision into corporate, plant, and process levels (Fig. 294) is specific to the process industries, because here it is often the case that independent facilities on the process control level are supplied with production specifications (e.g., recipes) directly from the production

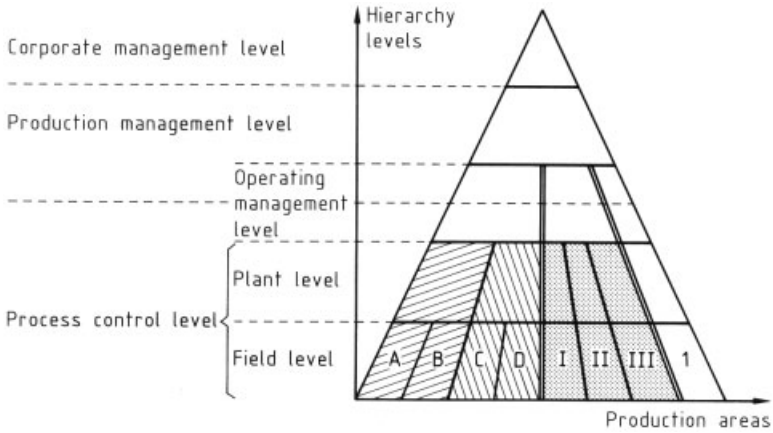

Figure 294. Levels of control in production

management level. In other industries, such as iron and steel, the products move in sequence through a number of plants or units where the raw material is upgraded, step by step, and ultimately converted to a salable product. This production structure is advantageously described by introducing an extra or intermediate level of management or control. This intermediate level breaks the general instructions from the production management level down into detailed instructions for the individual production units.

This structure implies that information flows both vertically (between control levels) and horizontally (within each control level). Both the transport and the storage of information reflect these information streams. Information logistics has to supply methods for the modeling and optimization of communications and information storage systems.

In general, the various levels employ disparate computer systems for information storage and processing. Information logistics, however, demands suitable high-capacity information pathways within and between the levels of control. It should be pointed out that information logistics is not just one of the tasks involved in information integration for production companies; it is a methodological aid to the planning and implementation of a corporate information utilization policy. Data communications and storage facilities do not become information communications and storage facilities until the meaning and utilization of the (digital) data have been established.

At first, information integration across the control levels of Figure 294 was thought of chiefly as a problem in manufacturing engineering. One indication of this is the wide adoption of the term CIM (computer-integrated manufacturing). Nevertheless, process engineering requires the same kind of information integration and has a functional structure wholly congruent with Figure 294. The term CIP (computer-integrated processing) is seen more and more often in this connection, showing that information integration is increasingly regarded as vital in both branches of production. It is therefore suggested that CIP be taken to mean computer-integrated production; this term expresses the fact that the same methodological aids will be applicable, and must be available, for information integration in all kinds of production operations. An analysis of the steps involved in manufacturing further shows that there are not only the discrete, piece-by-piece operations but also completely continuous "process-like" operations. If information integration becomes a way for methods of manufacturing engineering and process engineering to come together, this will be in accord with the situation in the plant. The integrative techniques of information logistics can then be put to optimal use.

8.3. Computer Communications Between and Within Control Levels

The need for information communication between the levels of Figure 294, as well as the

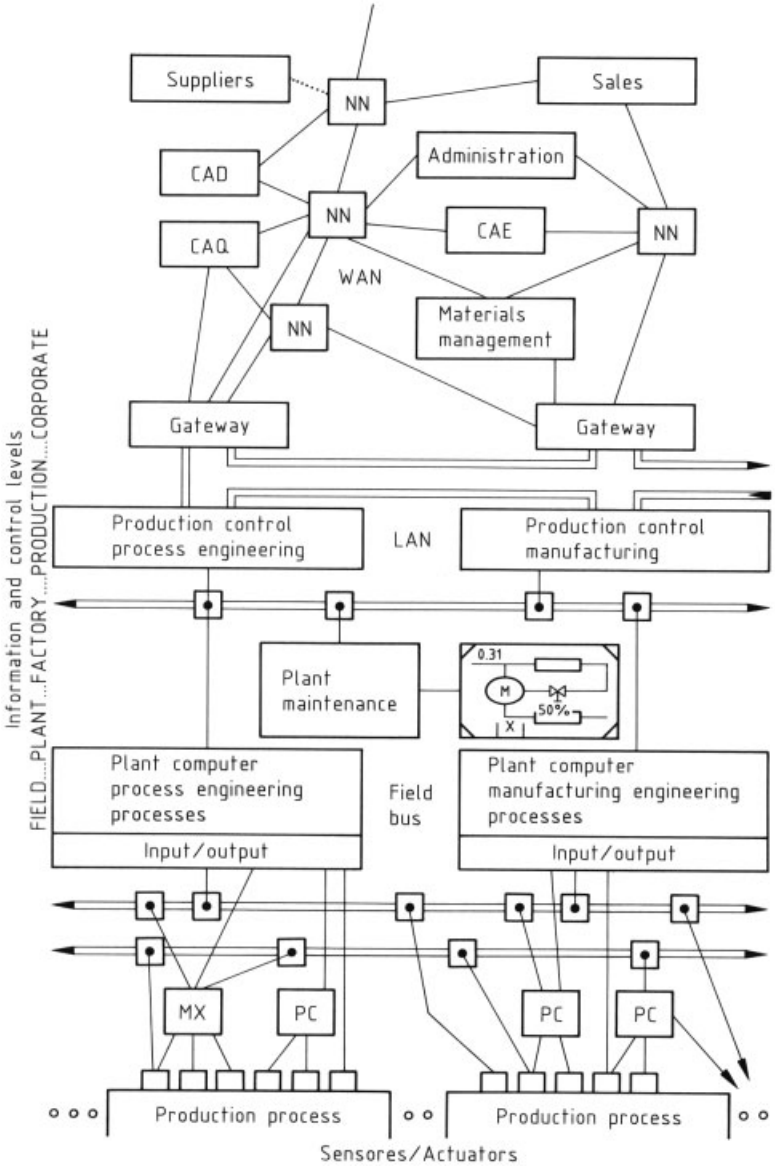

Figure 295. Networks for computer integration

requirements of information logistics, demand that appropriate aids be made available in the production company. As Figure 295 shows, the implementation of computer-aided information communication calls for a hierarchy of computer networks [595]. The information-handling facilities at each level of control must meet the following two requirements:

- There must be dedicated computers at each level. Information processing for more than one level can be combined in one computer, but the unavoidable increase in system complexity will result in a loss of flexibility, efficiency, and maintainability.
- The communications facilities at each level must be tailored to the particular needs of the level. The communications equipment

must also be able to connect heterogeneous computer systems together when a level uses computers of different generations, types, and capabilities because of special requirements, procurement strategies, or historical circumstances.

The networking concept illustrated in Figure 295 is one approach to meeting these demands. The following points may be important in individual cases:

The communications facilities must be designed and implemented with attention to the necessary time and consistency conditions. The effort of determining and modeling such conditions is justified by the optimal economics and flexibility of the communications system.

The communications system must be fault-tolerant as well as safe and secure with respect to intentional and inadvertent falsification of information. The redundancy and coding procedures employed for this purpose should be selected to fit each application [597].

Because production facilities are continually being changed, information logistics must be correspondingly flexible. Well-defined and, when possible, standardized interfaces and communications rules (protocols) are needed in order to meet this requirement economically, especially in heterogeneous systems.

Information integration across all levels of management and control is possible only if all information is in digital form.

However, information integration does not mean just data communications between one computer and another. Information logistics, in particular the transport and storage of digital data, must be supplemented by conventions on the meaning (semantics) of the digital data. The producer and consumer, who may be in different areas of activity, must interpret the digital data in the same way; only then do the data become information. Thus, when disparate computer systems are formed into an information network, it is particularly important that protocols be defined and standardized for semantic transition areas between applications systems. Formal methods of syntactical and semantic notation, based on experience and results from informatics, should be fully utilized for this purpose.

8.4. Computer Communications in Industrial Production; Standards

Standards already exist for the implementation of computer networks. These are discussed from a technical standpoint in what follows. The equally important economic aspects of various approaches call for an exact knowledge of each situation and its constraints. Computer-integrated production does not exist as an off-the-shelf solution, but the existence of standard protocols, however, makes it easier to handle specific applications problems.

Communications at the Process Control Level. Plant computers and the technical process often communicate (Fig. 295) through intelligent subsystems, for example stored-program control systems (SPS) or field-level multiplexers (MUX). Other direct methods of communication include digital input/digital output (DIDO) and standard serial and parallel buses (SCSI, IEEE 488, RS 232/422/485). Field buses, which permit serial data transfer under field conditions, are especially important and are treated in depth in Section 8.6.

Local Area Networks (LANs) at Higher Control Levels. Local networks have become vital in connecting computers with near real-time functions, such as plant unit, plant, and control-room computers, at the process and plant control levels. These setups allow both protected communication between two partners and unprotected simultaneous communication with more than one partner in the network (broadcasting). Physically, these networks are available based on either copper wire or optical fiber. The standard for levels 1 and 2 is IEEE 802, which has now been adopted by ISO as well. Line access protocols are CSMA/CD and token-passing. The CSMA/CD protocol (ISO IS 8802/3) has found wide acceptance in office systems, and a number of industrial products are on the market. This protocol is easy to implement, but the possibility of collisions means that transmission times are not well-defined. Its applicability under real-time conditions is therefore controversial.

In the token-ring protocol (ISO IS 8802/4), permission to transmit is linked with possession

of the "token," a special data structure sent in the form of a telegram. There is no possibility of colliding transmissions. Certain participants can be assigned priorities for permission to transmit through special token distribution mechanisms, so that even worst-case communications times are guaranteed. As a result, the token-ring protocol is particularly suitable for real-time applications. On the other hand, special fault-tolerance methods should be noted, such as the replacement of a lost token, which can markedly increase the administrative effort involved in running the token protocol compared with the CSMA/CD protocol.

Open Networks (WAN = Wide Area Networks). At the production and corporate levels of control, computers must communicate over distances longer than the few kilometers to which local area networks are restricted. Wide area networks (WAN) can link two communications partners over any distance, worldwide. In contrast to the broadcast mode of LANs, connections between nodes in open networks must be explicitly opened and closed. A variety of transmission media are available: all kinds of copper wires, wireless links (radio, satellite, infrared), and optical fibers. Open networks are characterized by the possibility of adding new participants while the network is in operation; therefore, communications standards are especially important in this area. ISO is making its contribution in the well-known OSI model [598]. Both LANs and WANs can be employed at all levels shown in Figure 294; special applications requirements may make the use of an open network desirable even at the process level of control.

Communication through Private Branch Exchanges (PBXs). The private branch exchange (PBX) has been a common installation in all kinds of companies for decades. For computer communications, the PBX has many of the properties of an open network [599]; for example, it sets up a point-to-point connection between two participants. A point to consider is the nonnegligible, and unpredictable, time needed to establish and drop connections, which are thus only partially suitable for real-time applications.

An important development is the transmission of digital data in parallel with telephone traffic, which is still partly analog but will be increasingly digital in future. With equipment now on the market, the Integrated Services Digital Network (ISDN) is limited to 64 kbit/s for the transmission of digital data. The PBX is therefore more suited to the transmission of digital data in parallel with speech at present, and isthus interesting chiefly at the corporate and production levels of control. The introduction of faster call-processing devices (with processing times in the millisecond range) will allow the PBX to penetrate local area network applications in the medium term.

8.5. MAP/TOP: Protocol Standards for Information Integration in Production Companies

Experience with the ISO standard protocols for data communications between computers [599] has shown that the standards are not adequate for true information communications. Up to the beginning of the 1980s, level 7 of the OSI model (the applications level) included only fairly general functions (file transfer, virtual terminal, etc.) with no reference to the meaning (semantics) of the transferred data. What was defined was thus tools with which the user could construct an information integration.

While this was a significant advance for computer communications applications, with their initially commercial orientation, the broad functionality and thus the semantic variety of the digital information that must be handled when computers are integrated into production meant that further standardization was needed on level 7 of the OSI model (see Fig. 296). This situation has led, since 1980, to proposals for MAP (Manufacturing Automation Protocols [600]) and TOP (Technical and Office Protocols). These proposals originated with users of automation and process control systems. MAP was developed by General Motors for purposes of manufacturing automation and, TOP, by Boeing as an approach to automating the engineering office.

MAP and TOP together provide for comprehensive information communications in industrial companies, from the administrative and resource allocation aspect through design and development (CAD) to computer-aided produc-

Application functions and data				
Assembly	CAE	Manufacturing	Quality control	Statistics
CAD	Office	Material management	Process control	Blue prints

Communications standards	CNMA		Testsystems
	MAP	TOP	
7 Application (Standard services)		Message handling system (MHS; CCITT X.400)	available
	ISO File transfer, access and management	(FTAM-subset; DIS 8571)	available
	ISO Manufacturing message service (MMS; DIS 9506)		available
	MAP Directory services (DS)		available
	MAP Network management (NM)		available
	ISO Application control service elements (ACSE; Subset: DIS 8649/1-2 and DIS 8650/1-2		
6 Presentation	ISO 8822, 8823		development
5 Session	ISO Session kernel	(IS 8326, IS 8327)	available
4 Transport	ISO Transport class 4	(IS 8072, IS 8073)	available
3 Network	ISO Connectionless internet	(DIS 8348, DIS 8473)	available
2 Data link	ISO Logical link control	(DIS 8802/2)	available
	Type 1	Type 1	
1 Physical	ISO Token passing bus (DIS 8802/4) (IEEE 802.4) Broadband / Carrierband	ISO CSMA/CD (DIS 8802/3) Baseband (IEEE 802.3)	

Figure 296. MAP/TOP protocol hierarchy

tion and quality control (CAQ). Both protocol environments constitute a key tool for CIP (Computer-Integrated Production).

The impetus for MAP and TOP came from users of industrial automation and process control systems. It is in the user's interest to select from the widest possible range of options the optimal solution to each problem while insuring the integrability of the new features. This is the alternative to the "single-source" solution often put together by one computer manufacturer. The interests of the suppliers and users of automation and process control systems are in conflict here: Complete standardization of interfaces for a wide variety of information-processing devices in a production company will breach the protective walls around computer networks, which have commonly been supplied by one vendor. On the other hand, this standardization trend expands the market for each vendor. The prerequisites are precise and strictly consistent standardization of protocols and a user with sufficient know-how and capacity to cope with the interoperability problems.

Objectives and States of MAP/TOP. The full definition and application of the MAP/TOP protocol hierarchy should guarantee direct communication between computers, control units, and their programming systems, even when these are supplied by different manufacturers. The objectives of MAP/TOP take in the following points:

The protocols should allow for the special properties and requirements of local plant and office networks. Further, they should be as compatible as possible with the respective open network protocols, so that communications between the two types of networks can be simplified (Fig. 295).

In contrast to the ISO protocols for open networks, with the few generally applicable functions of applications level 7, the MAP/TOP protocol hierarchy for level 7 includes special applications functions tailored for manufacturing and office automation (e.g., MMS = Manufacturing Message Specification and MHS = Message Handling Specification).

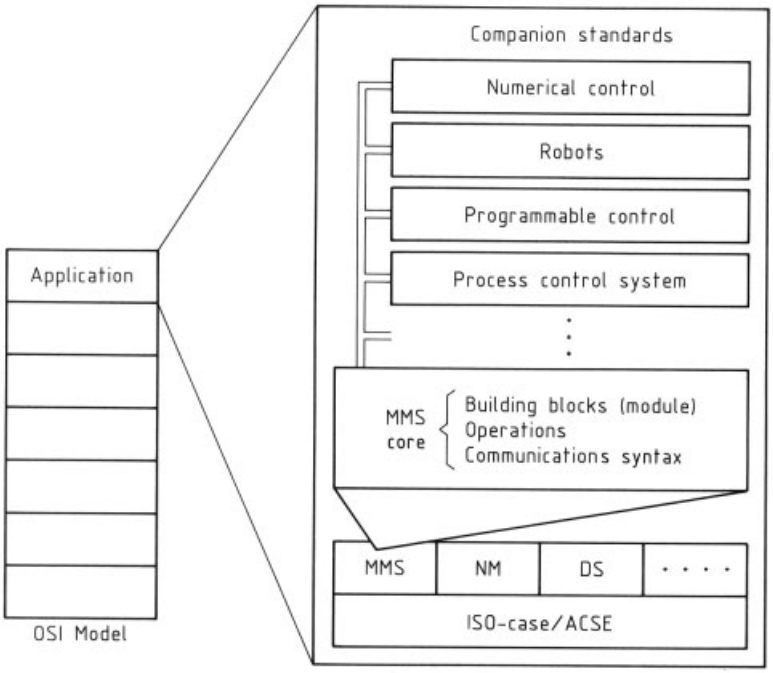

Figure 297. Application-oriented Companion Standards

In support of these ambitious objectives, the MAP/TOP definition includes quality-assurance measures such as conformity tests, which insure and certify the ability of protocol products from different manufacturers to communicate with one another.

Thus the objectives of MAP/TOP go beyond those of the ISO open network protocols, taking on the difficult task of identifying and standardizing applications functions in the manufacturing and office area at level 7. This makes it clear why the development of MAP/TOP, especially the applications protocols, is still in flux and will not be completed in the foreseeable future.

Figure 296 illustrates the MAP protocol environment as now available. At levels 1 and 2, the LAN protocols discussed in Section 8.3 are standardized. Levels 2 to 6 contain protocol definitions equivalent to WAN protocols (the same for MAP and TOP). The use of these protocol standards at levels 1 to 6 insures data communications between applications functions in production and office.

On applications level 7, three protocol types are distinguished. The first corresponds to the ISO protocols for WANs; these are essentially ISO standard applications protocols such as file transfer.

The second protocol type covers application-specific functions. Examples are the MMS protocol family in MAP and the MHS protocol system in TOP (Fig. 296). A third protocol type on level 7 deals with the functions of network administration; examples are Network Management (NM) and Directory Service (DS). These protocols also offer functions for handling errors as well as fault-tolerance practices.

Figure 296 shows the intention of the ESPRIT CNMA (Communication Network for Manufacturing Applications) project, a European undertaking that supports MAP and TOP in a similar way [601].

While the MAP specification focuses on manufacturing operations, a similar problem is evident in process operations. Proposals for a PMS (Processing Message Service) are therefore coming from a variety of sources. Its functions will correspond to the communications requirements of computers in the processing industries.

Generally, there is the problem of linking various functional environments to the basic MMS standard. Figure 297 shows the "Companion

Standards" concept that has now been chosen. A standardized MMS core includes generally usable concepts and structural specifications for applications standards that will serve as the basis for a variety of Companion Standards relating to technical specialties. This procedure allows the step-by-step construction and expansion of the applications function standards on the basis of user experience.

8.6. Field Bus Systems

Basis and Objectives of Field Bus Systems. Communication at the field level links field devices with control-level computers. The devices in question include sensor systems that locally acquire complex information about the production process; actuator systems that influence material and energy flows; and programmable (stored-program) controllers that perform local control and optimization tasks. The same holds, with suitable adaptations, for both processing and manufacturing operations.

Distribution, as a basic property of present-day control and instrumentation, extends to information processing at the field level and, to an increasing degree, blurs the distinction between sensors and actuators, on the one hand, and control computers on the other. This means that powerful communications systems meeting a variety of requirements are also needed at the field level. Examples include direct communication between intelligent sensors and actuators in local control loops, as well as communications links used to transmit product data between plant control systems and control computers.

Most signal transmission at the field level is by field devices individually wired to the control computers. The standard 0 – 20 mA and 4 – 20 mA signals are dominant in process engineering and have become standards in chemical engineering. In manufacturing, sensors and actuators on machines are also singly and directly linked to the cell computer or to programmable controllers.

The variety of information involved in modern production processes sets limits to this form of communication with regard to plannability, maintainability, immunity to electromagnetic interference, and economy. Field bus systems should overcome these limitations:

- The star configuration of individual lines is replaced by a digital bus known as the field bus. It is cabled to the various field devices and transmits digitized data via digital interfaces.
- Such a bus connects some tens of field devices among themselves and to a plant computer. If the devices are close together, the cost of cabling is substantially less than for individual cabling (Fig. 298). If field devices are assumed to be evenly distributed over a specified area, the savings in cabling costs can be estimated from Figure 299.
- The growing intelligence of field devices calls for the transmission not only of process measurements and control information but also of digital status and control data between field devices and control computers. The additional cabling that used to be required for this purpose can be omitted when the field bus is employed.
- The conversion of analog signals to digital locally, that is, at the connection to the field bus, prevents interference to the signal on route to the control computer. The redundancy of digital codes also makes it possible to reconstruct erroneous data.
- The standardization of field bus protocols permits standard information interfaces between field devices and a wide variety of control computers, so that open communication results at the field level. This is not possible at present with the wide range of information transmission media at the field level (Fig. 299).
- Diverse media can be used for digital data transmission (e.g., copper twisted pair, coaxial cable, optical fiber, radio waves), so that economic and technical optimization is possible (transmission rate, noise immunity, energy expenditure/intrinsic safety, electrical isolation, etc.).

These objectives and potential features make up the basis for field bus standards now being devised. The many degrees of freedom in the specification and implementation of field buses lead to a variety of approaches and have worked against the desired uniformity in field-level information interfaces.

Field Bus Systems in Development. Teams in several countries have made a start toward

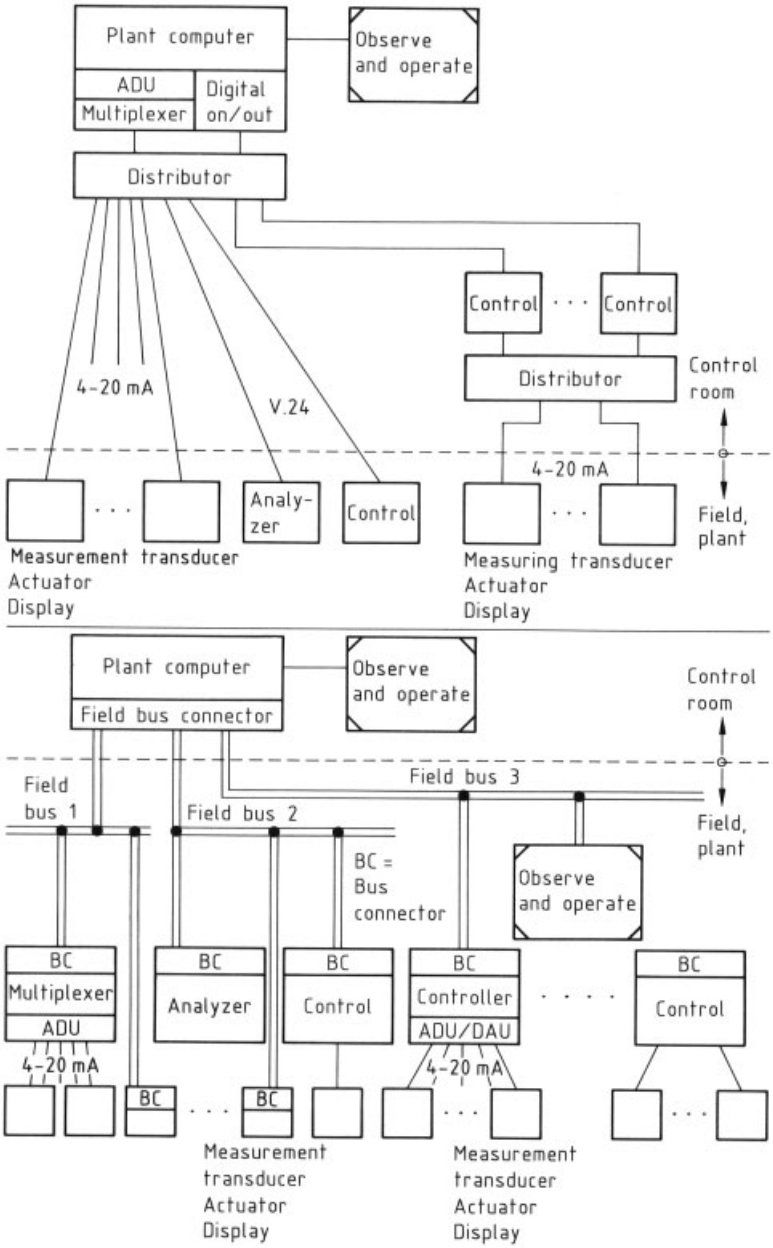

Figure 298. Transformation from star-configured individual cabling to field bus structures

specifying and implementing field bus systems. The properties of an architecture for all these projects is described first. The various approaches to implementing this architecture employed in the several projects are then discussed.

General Field Bus Architecture. A field bus system is a local network having a restricted number of nodes, which communicate exclusively among themselves or with a control computer. There is accordingly no need for the services of ISO levels 3 to 5. Presentation level services are also unnecessary in general form, since they can be performed in special form by the services of level 7 or 2. For open commu-

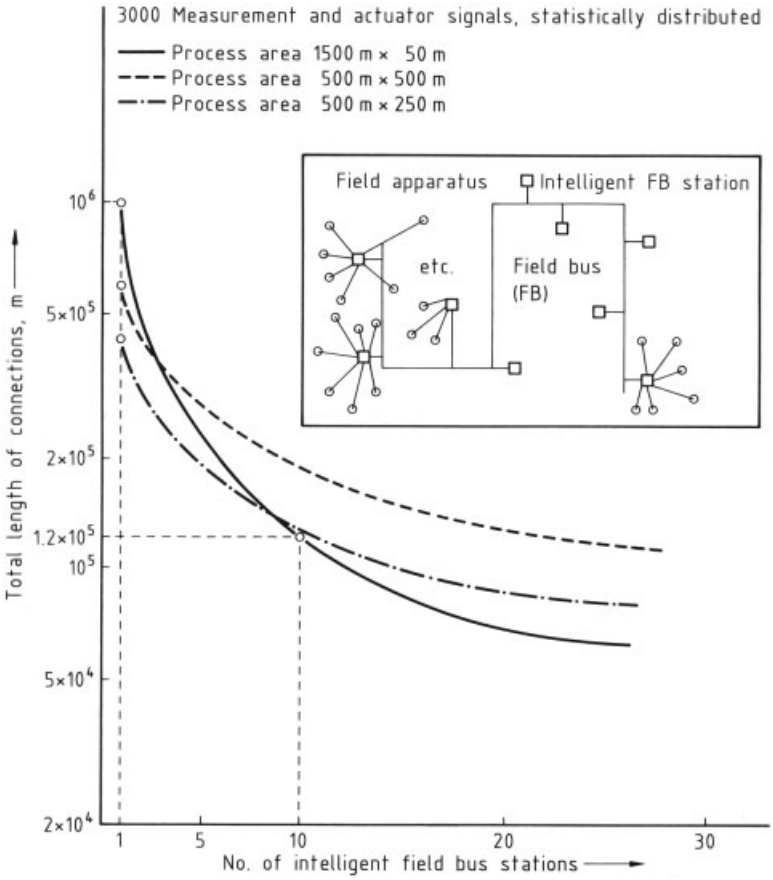

Figure 299. Estimation of cabling costs in field bus structures [615]

nication between field bus nodes, however, the services of level 7 must be specified. A general field bus architecture thus includes ISO levels 1 (physical layer), 2 (link layer), and 7 (application layer). This reduced architecture must be supplemented by network management services as well as application-specific services on level 7, such as those already introduced as Companion Standards [602] in MAP/TOP.

The American standards group ISA SP 50 has defined two performance classes (H1 and H2) for the *physical layer.* The scope of H1 is the replacement of existing analog standard interfaces (0 – 20 mA and 4 – 20 mA); the H1 specification must permit adequately fast and error-proof transmission of process, measurement, and control signals. Accordingly, the H1 specification is largely governed by process engineering requirements.

The H2 specification is to extend the H1 concept to applications with higher performance demand. This includes the fast transmission of large data sets, as is needed for the loading of machining programs in manufacturing engineering. H2 is therefore specified for manufacturing. IEC TC65C has created specifications for a field bus standard for use in industrial control and instrumentation. In addition to the physical layer performance classes already cited, it describes other properties: the topology (linear bus or tree), the transmission medium (twisted pair or, later, optical fiber or radio link), and finally the number of nodes on each field bus segment (32). For performance class H1, a data rate of 31.25 kbit/s and a maximum cable length of 1900 m are set forth. Performance class H2 has a data rate of 1 Mbit/s and cable lengths of up to 750 m. The intrinsic safety of the physical layer

is handled by limiting the electric power that can be carried on the bus. The H1 specification allows enough power to run at least five field devices, at least three of which can operate in a hazardous environment.

The H2 specification provides for at least three field devices to be supplied on the bus and permits AC transmission over the bus lines. Intrinsic safety is not explicitly considered in the H2 specification, since it is of no importance as a property of manufacturing systems. At present, power supply to field devices, even with auxiliary energy sources, is not a function of field bus systems; over the long term, however, separating the information channel from the (auxiliary) energy channel in process operations is a goal to be sought [603].

The specification of the *link layer* in field bus designs corresponds to the ISO/OSI Data Link Layer (DLL). In this way, the two sublayers LLC (Logic Link Control) and MAC (Medium Access Control) are established.

For the application area of field bus systems in general, the *link layer* must provide adequate efficiency (real-time environment) and also support access priorities as well as point-to-point connection or simultaneous connection of one node with several or all other nodes (multicast, broadcast). The essential novelty of the field bus concept as described, however, lies in the procedures by which nodes access the common data transmission medium (MAC). While the French FIP proposal [604] features a centrally organized mechanism in the sense of bus allocation, other proposals (PROFIBUS [605]) utilize a distributed access mechanism based on token passing as set forth in IEEE 802.4. These differences exist because FIP assumes better real-time properties for the central bus control. PROFIBUS uses standard microelectronics for the efficient performance of the more complex token-passing protocol, which allows more flexible parametrizing of the bus allocation for individual nodes [606].

The *application layer* must furnish application functions without regard to the protocols of the lower layers. These functions must insure the exchange of information in a distributed system and with adequate efficiency. Because applications in industry, whether processing or manufacturing, are subject to strong timing conditions, efficiency is particularly important. This holds for cyclic information transmission as well as the detection of spontaneous events. While the functionality of the protocol services on the application layer can be specified independently of the lower protocol layers, their performance under real-time conditions is crucially dependent on the efficiency of all protocol layers.

The application layer for field bus systems must offer cyclic and noncyclic services. Cyclic services are, for example, the reading and writing of process variables or parameters as well as access to information on the validity and consistency of such variables. Noncyclic services include the sending of reports based on an event, the sending of management information, and the initialization or configuration of subsystems during startup of technical processes.

A variety of approaches can be taken to the implementation of such services. The essential difference lies in the relationship between the information-generating and information-consuming processes. For example, client/server models as well as producer/consumer models can be used [607]. Requirements on the field bus for applications in process engineering have been thoroughly described by Pfleger [608] and Drathen [609].

Already in MAP and TOP it was seen that the specification of complete function sets for all conceivable applications in industrial production is not possible in a uniform standard. On the basis of an MMS core standard (Manufacturing Message Specification, ISO IS 9506), special function sets have therefore been standardized for various application areas; these are called *companion standards* [602]. A similar procedure is also provided for the standardization of field bus systems. Companion standards define function blocks, such as those for the control of processes (analog and digital input/output, PID controllers, etc.) or for the utilization of production facilities (configuration, object description, maintenance, startup, etc.). The companion standards for field bus systems will resemble those for MAP wherever appropriate in terms of functionality. Büsing has described the importance of MAP for process control engineering [610].

In the ISO/OSI model, the *network management* functions supplement the transmission protocol functions and communicate with them. As in any case where intelligent subsystems

are networked, network management is also required for a field bus system.

Of the five management areas of the ISO/OSI model, the following are pertinent to field bus systems:

- Performance management
- Configuration management
- Fault management

The management functions for a field bus system are implemented in a network management station. Its functions and services include system initialization, the loading of programs and parameters, the starting and stopping of the communications system, the detection and localization of faults, and the reconfiguration of the field bus system in reaction to faults identified. Furthermore, the management station monitors the performance and loading of individual communications channels, identifies bottlenecks, and generates planning data for network upgrading. Such a network management component appears in all the field bus system now in development.

Incorporation of Field Bus Systems into the Information Hierarchy. As shown in Figures 298 and 299, field bus systems connect intelligent field devices with one another and with plant computers. They must support information transport at the field level but are also incorporated, through the plant computers, into the overall information pyramid. Since field bus systems are used in both processing and manufacturing, their real-time performance and message formats are subject to a variety of requirements.

Within the field level, real-time requirements are critical when information transmission over a field bus is within a control loop. Depending on the type of process being controlled, guaranteed transmission times in the range of a few milliseconds are needed; the messages to be transmitted are, however, short. The situation is similar for field bus systems in manufacturing cells. Longer messages and less critical real-time conditions occur when information is exchanged between CAD/CAM systems and manufacturing cells. The guaranteeable response time over a field bus depends on the types of services implemented.

When field bus systems are integrated into the information pyramid, it must be kept in mind that levels 3 to 6 do not exist or are emulated only to the extent necessary for direct field bus operation. This means that a special field bus node must offer the necessary network functions and also must have the requisite information about the communications partners at higher levels of the information pyramid. As in reduced MAP architectures, there must be a gateway that, on the one hand, is fully networkable and, on the other, can execute the field bus protocols (Fig. 300).

Constraints Arising from Applications. The ESPRIT CNMA project (Communication Network for Manufacturing Applications) carried out a survey to find out what operational constraints would apply to a field bus standard in the fields of processing and manufacturing engineering:

Processing:
Field devices to be connected:
Programmable controllers, sensor and actuator systems, minicomputers, field multiplexers, and power supplies for simple sensors and actuators
Information to be transmitted:
Cyclic (length ≤ 4 bytes, transmission time between 100 ms and 2 s)
Alarms (length ≤ 2 bytes, transmission in a few ms)
Overall field bus transmission rate > 250 kbit/s
Process dimensions:
Tens of meters up to roughly 2 km; distance to control room between 50 m and 3 km; total bus cable length between 75 m and 10 km
Ambient conditions:
High temperatures ($> 80\,^{\circ}$C), humidity, dust, vibration, electromagnetic interference; the transmission medium must be resistant to oil, steam, and ultraviolet; insulation must not contain halogens
Safety:
Intrinsic safety is an essential property; safety classes IP 65, NEC Class 1 Division 2
Conformance to protocol specifications
Certification of conformance in order to insure connectability, interoperability, and interchangeability of field devices on the field bus is regarded as important

Manufacturing:
Field devices to be connected:
Programmable controllers, numerical controllers, robot controllers; intelligent sensor and actuator systems are of minor importance
Information:

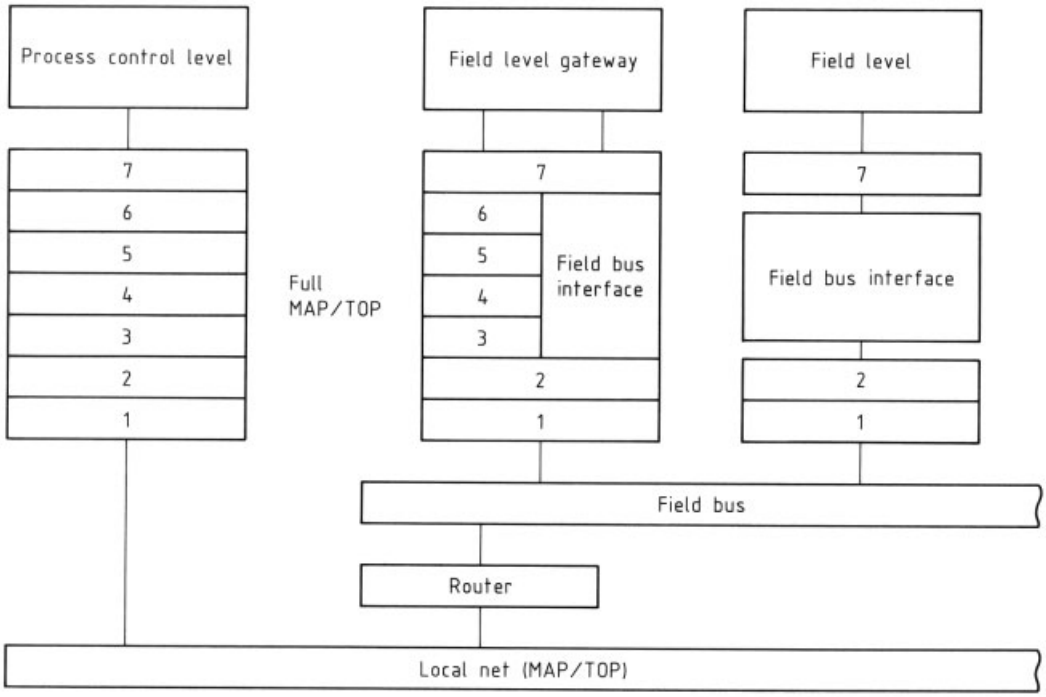

Figure 300. Linking of field bus and MAP LAN

Cyclic data (length 1 – 64 bytes, transmission time between 10 ms and 1 min)
Alarms (length 2 – 8 bytes, response time ca. 1 s)
Production data (length 2 – 256 bytes, time interval between 10 s and 8 h)
Process dimensions:
The diameter of manufacturing applications is ca. 10 – 80 m; usually the control room is nearby; bus cable length between 5 m and 50 m
Ambient conditions:
Severe electromagnetic interference, wide fluctuations in power supply voltage, static discharges
Safety:
No special requirements on safety or intrinsic safety of bus system
Protocol conformance:
Certification of field bus products for conformance to protocol is regarded as important

These survey results are reflected in different ways in existing field bus standards and those in preparation. The discussion that follows will deal with the most important properties of two field bus definitions. As early as 1989, Polke pointed out the needed separation into energy and information channels and questioned the requirement of "intrinsic safety" against the background of modern sensor and actuator systems [603].

PROFIBUS (Process Field Bus). PROFIBUS is a joint industry/science project in Germany, supported by the Federal Ministry of Research and Technology (BMFT), which prepared a field bus definition that is now a German standard (DIN 19 245). The PROFIBUS user organization (PNO) is charged to maintain and further develop the PROFIBUS standard.

PROFIBUS corresponds to the architectural principles discussed above and was conceived for use in both process and manufacturing industries. Its essential properties include the use of the token-passing access procedure on layer 2 and the Field Bus Message Specification (FMS) including network management over all layers. FMS [611] is derived from ISO IS 9506 (MMS = Manufacturing Message Specification) so that it can communicate with MAP applications in a simple manner and make use of experience gained in the development of MMS. A Lower Layer Interface (level 7 a) replaces the functions of ISO layers 3 – 6 needed for field bus

systems. A conformity test is included in the PROFIBUS definition.

At the physical layer, PROFIBUS defines a twisted pair cable with a maximum length of 1200 m (without repeaters) or 4800 m (with repeaters). If repeaters are used, the bus topology can be straight bus, tree, or star. The transmission rate can be between 9.6 and 500 kbit/s. The number of nodes can be up to 122, of which a maximum of 32 can be master stations. Redundancy, intrinsic safety, and transmission of auxiliary energy comply with the specifications of ISA SP 50.

Layer 2 of PROFIBUS (MAC) includes two protocols, the distributed token-passing protocol (IEEE 802.4 or IEC IS 88 204) and the centralized access protocol. In this way, communication among a number of master stations, or between one or more master stations and the substations (slave stations), is possible.

A special characteristic of the PROFIBUS definition is the use of standard processors for the protocol implementation, with a crucial effect on the efficiency of data transmission. On the basis of this efficiency, PROFIBUS at present is only conditionally suitable for service as a signal path between closed control loops with critical real-time conditions.

FIP (Factory Instrumentation Protocol). The FIP protocol definition has been in preparation since 1985, with support from the French ministry of higher education and research ministry. Its purpose is information transmission in a manufacturing setting. The FIP bus consortium includes more than 70 companies located in several European countries. There are at present three French standards for layers 1, 2, and 7. Other French standards for network management, MAP/MMS-compatible functions at layer 7, and optical fibers on the physical layer are in preparation [604].

The physical level allows for a maximum of 256 nodes. Three transmission rates are provided on the twisted-pair cable (31.25 kbit/s, 1 Mbit/s, 2.5 Mbit/s). The cable can be up to 3 km long at 1 Mbit/s. For optical fiber systems, the maximum length between two node connections is 1600 m (glass fiber) or 125 m (plastic fiber).

On layer 2, there is a centralized access control procedure (MAC), which supports both cyclic data transfer and the transmission of aperiodic data or longer messages.

The FIP definition, based on centralized bus control, assumes that data can also be transmitted within closed control loops. To achieve the required efficiency, protocols are implemented in specific ASICs differing in functionality. The FULLFIP ASICs include the standardized functions of layer 7, layer 2, and (in part) layer 1. This approach supports the connection of simple and inexpensive field devices (simple sensors) and minimizes the energy requirements of the interfaces. The decision to use special ASICs means, however, that they must be continuously modified to keep up with changing technology and standards. Given that international standardization of field buses has not yet stabilized, this point may be a problem.

The Route to Open Communication at the Field Level. Open communication at the field level means a consistent, internationally standardized protocol environment allowing field devices and plant computers made by a wide range of manufacturers to be linked to one another. This is particularly important at the field level, because a division of labor between manufacturers is natural and unavoidable in the area of field devices. Open communication has not yet penetrated into the production control and process control levels, but this fact cannot conceal the fundamentally different situation at the field level. Accordingly, manufacturers of field devices have not failed to come up with suggestions to remedy the lack of unity in the world of field bus definitions.

For example, the EUREKA field bus project has the goal of evaluating present field bus concepts and thus promoting the standardization of device interfaces to field buses. Another approach is embodied in the ISP project (Interoperable Systems Project), an international initiative by leading manufacturers of instrumentation and automation hardware wishing to create and standardize a field bus definition optimized for industrial applications on the basis of experience with existing field bus definitions.

The international standardization of field bus systems is proceeding in IEC TC65/SC65C, which is collaborating with the American standards body ISA. The ISA SP50 working group is creating a field bus specification for use in

international standardization. The H1 and H2 performance classes defined by ISA SP50 have already been mentioned. In 1989, ISA SP50 attempted to harmonize the discrepancies between the access control methods (layer 2, MAC) in various field bus definitions. No final result has been achieved, but there is a trend toward "distributed" access control as set forth in IEEE 802.4 or ISO IS 88 024 and used in PROFIBUS. The actual linking functions on layer 2 (LLC) have also been specified; the objective is to unify the most important mediation functions in all field bus proposals while simplifying these functions and thus improving their efficiency. On the whole, there is a trend toward the harmonization of field bus proposals, although the effect of existing national standards and international industrial cooperation on the likelihood of an international standard being accepted must not be underestimated.

The European Commission supported (under ESPRIT) the FICIM project (Fieldbus Integration into CIM) in order to promote the harmonization efforts discussed. From November 1990 to November 1992, the properties and interoperability of the two European field bus specifications (PROFIBUS and FIP) were tested in a pilot installation. The project had the participation of industrial companies as well as NAMUR, the Standards Working Group for Measurement and Control Engineering in the German Chemical Industry. The application functions and application standards were to support and insure the integration of information from these two field bus systems into a CIM system, independently of the lower field bus layers. The pilot phase of this project showed that such an integration goal is indeed achievable. The pilot facility is therefore being used for further studies of the interoperability of field bus systems.

The certification of field bus implementations for conformance to the standardized protocols is regarded as important in all field bus projects. Specifications and implementations of conformance test systems for all layers now exist for the FIP bus and the PROFIBUS. The Upper Tester, that is, the part of the test system above the protocol layer being tested, presents a special difficulty here. Because it is expected that field bus protocol implementations will be optimized for medium-sized and small intelligent field systems, the formerly usual (in MAP) configuration of the Upper Tester on the field bus device under test is no longer possible. The standardized conformance test for such devices thus includes an "Upper Tester Agent" (UTA) on the field bus device; through the field bus, this communicates with the actual Upper Tester on the separate test system.

The wide range of efforts to unify standardization work now in progress, and also to insure conformance of field bus components to the standards, suggests that open communication at the field level can be realized despite the variety of approaches now used. Manufacturers and users in the field of industrial technology should provide intensive support to this development work, since it is an important prerequisite for managing the complexity of future industrial control systems.

Again, a meaningful solution will be obtained only when the information model providing a semantic description of field-level problems has been adopted (see Chap. 2 and Section 5.1).

8.7. Quality Assurance: Conformance and Interoperability Tests

When computer communications equipment made by different manufacturers is used, the user must be able to rely on correct implementation of the protocols. The current state of the art does not allow a formal specification and thus an automatic verification of complex protocols, and so protocol products have to be tested with a standardized set of test cases. These "scenarios" define status data, parameters, and function calls, with whose aid all relevant protocol functions and protocol states can be tested as completely as possible or necessary. A protocol product that passes these tests is said to "conform" to the protocol specifications. If these specifications are taken to be complete and correct, a conformant protocol product will be able to communicate with any other corresponding protocol product.

In a full MAP/TOP or field bus protocol implementation, all protocols at the various levels must be tested. Figure 301 shows the general procedure for conformance tests [612]. The protocol being tested ("implementation under test") is enclosed by a test frame consisting of an upper tester (UT) and a lower tester (LT). While the LT responds to the test protocol in a well-

defined way and detects the reactions of the protocol, the UT interacts with the protocol as if it were a protocol product on the next higher level. The UT is assumed to function without errors; this supposition is justifiable only because the UT merely provides certain interface conditions (ASP = Abstract Service Primitive) and need not exhibit the full functionality of a protocol.

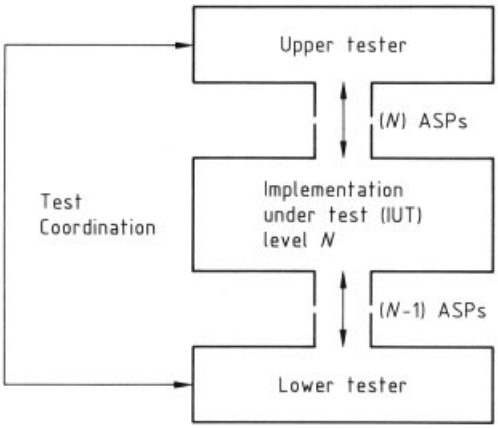

Figure 301. Principle of conformance testing [612]
ASP = Abstract Service Primitive

A test coordinator makes certain that the test scenarios set forth in the conformity test are run, causes the responses of the specimen to be stored, and interprets the test results. Interpretation consists of comparing the reactions of the specimen with a predefined set of expected reactions. The conformance of the protocol product or the presence of implementation errors can be deduced from this comparison.

The fact that protocol products have passed the conformance test does not necessarily guarantee their interoperability. Like any other test, the conformance test is incomplete; what is more, it cannot be more complete and correct than a possibly incomplete and inconsistent protocol specification. This uncertainty can be reduced only by additional interoperability tests.

An interoperability test studies the actual interoperability of protocol products, that is, their real ability to interact in open networks. Interoperability depends on conformance to protocols but also on the matching of parameters that can be adjusted when the protocols are run. In this context, the term profile means an application-specific selection of parameters and functions of a protocol in accordance with a standard. Depending on the ranges of variation of such parameters, different profiles may conform to the standard but not be interoperable. The problems associated with the range of options in profile definition are not discussed further here.

A final key aspect of the serviceability of protocol products is efficiency and performance. Even given correct and conformant protocol implementation, interoperability can be significantly impaired by, for example, wide differences in timing between protocol products. In future, accordingly, standardized interoperability and performance tests will be defined.

An interoperability test system must establish a connection between the protocol components being tested and monitor the resulting flow of data and information. The use of expanded lower testers (cooperating with a communications observer) is conceivable. The interoperability tester compiles the results from the lower tester and the communications observer.

While conformance tests are done against a (possibly partly formal) protocol specification, the interoperability test deals with communications aspects that cannot be so well specified. It is thus much harder to standardize this kind of test. The user of communications protocols, however, sees interoperability as the proper acceptance criterion for protocol products; this is why work is under way to standardize these tests.

8.8. Methods and Tools for Protocol Specification

The preceding sections have made it clear how crucial the degree of formalization of protocol specifications is for the correct implementation of protocol standards. This section describes some procedures and aids now being used in the specification and implementation of protocols and in test systems. The protocols in question are those in the context of the ISO – OSI reference model [598].

These protocols [598] define services that can be described by the functions invoked and the data exchanged. Services are composed of service elements such as establishing and dropping connections, transferring data, and exchanging acknowledgements. Such service elements are deemed indivisible sequences of events and must

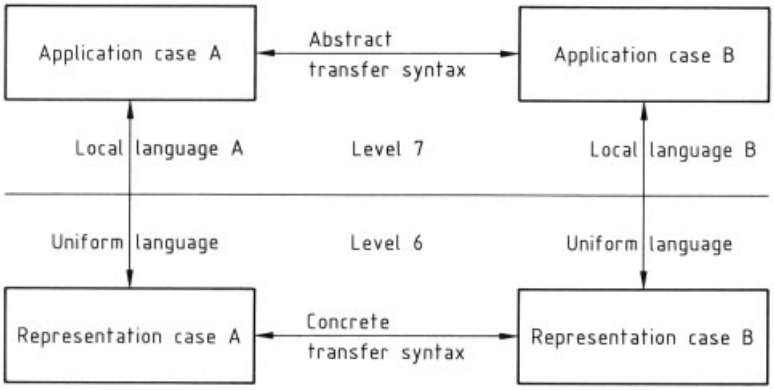

Figure 302. Representation context [616]

be implemented in a proper way with regard to fault tolerance and consistency requirements. They are neutral with respect to the data or function calls mediated by them.

Service elements are based on "service primitives," that is, functions such as "request," "response," "indicate," or "confirm." In presentday protocol specifications, these service primitives and thus the service elements built up from them are commonly described in natural language. In this way, a formal specification of the functional (procedural) part of protocols is not usually stated, notwithstanding the fact that these functional (procedural) parts of protocols represent finite automata. The form of description as a finite automaton is, however, used only in exceptional cases and for subfunctions. One important reason for this situation is the complexity of the automata, whose many and varied options make them hard and inconvenient to describe in the form of states and changes of state. Protocol Data Units (PDUs) exchanged with the aid of the service primitives, on the other hand, are now specified formally. A "language" of abstract syntactical notation is available for this purpose; ISO has standardized it under IS 8824, and it has become known under the name "ASN.1" (Abstract Syntax Notation One) [613]. Advantage has also been taken of experience gained in the informatics field with various higher-level programming aids and their data structure description methods.

At present, ASN.1 is used chiefly to define PDUs in protocols at the applications level of the ISO model [614]. As Figure 302 shows, applications programs or applications entities exchange information by means of PDUs. The ASN.1 language makes it possible to formulate these PDUs in the abstract manner standardized by IS 8824. It should be ensured that the meaning (semantics) of the PDUs in a given local applications syntax and in the abstract ASN.1 syntax remains unchanged. Thus, semantics-preserving transformations between the syntax of any local language and the abstract ASN.1 syntax must be defined.

The transmission of these PDUs via a communications system requires a concrete syntax, called the transfer syntax, that is understood by both communications partners. The transfer syntax must be defined at level 6 of the ISO model. Accordingly, ISO has standardized a "concrete transfer syntax" under IS 8825 [613], just as the abstract syntax is standardized under IS 8824. Thus, for PDUs actually to be transferrable between two applications entities, the abstract PDU definitions notated in ASN.1 must be translated from the real local syntax of each applications program into the IS 8825 transfer syntax. This assignment of a transfer syntax to an abstract syntax is called "representation context."

It is significant that the representation context standardized by ISO for ASN.1 is not a unique standard valid for all time. Other representation contexts especially suitable for certain applications areas can be defined later on; this is particularly important in connection with the companion standards discussed in Section 8.4. Thus, in order that one applications-specific representation context can be replaced by another for communications between applications program, every communications process begins with the

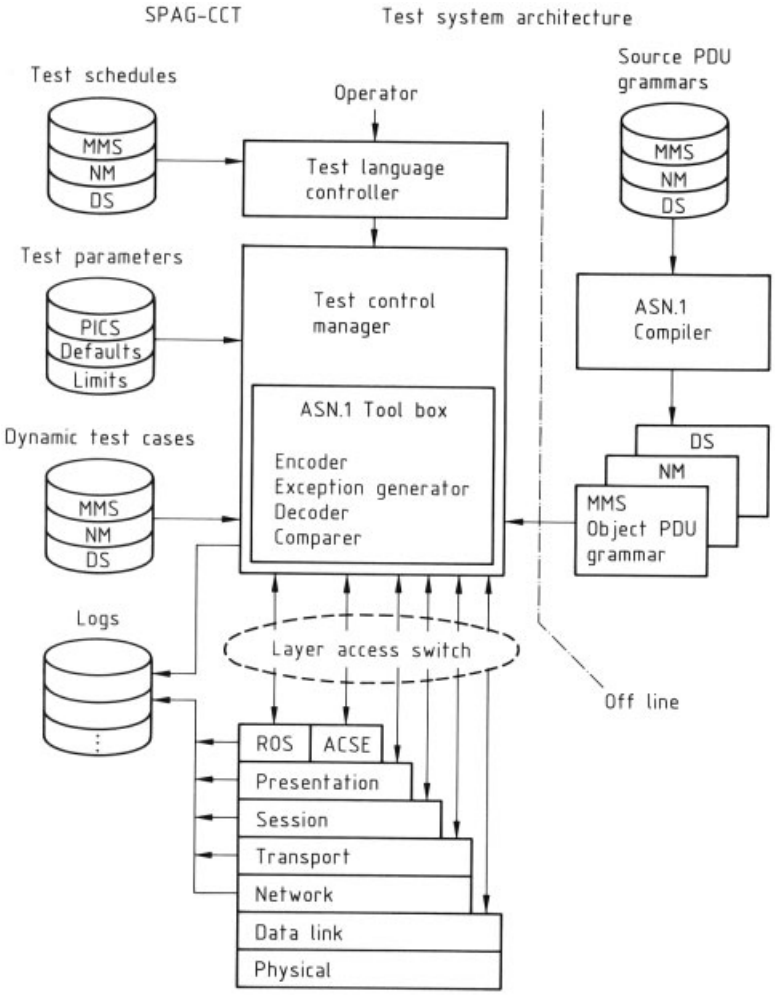

Figure 303. Implemented architecture of conformance testing

negotiation of the proper representation context. One example might be an exchange of communications of applications program systems between partners in the realm of process control engineering and partners in the realm of office communications. The difference in encoding conventions (e.g., measured data or CAD data) means that the representation context must be adapted.

There is already an alternative to the representation context ASN.1/Basic Encoding Rules, in the form of the "default context," which is used whenever no special representation context has been negotiated. This is the case in particular when a connection is being established. The default context, however, contains neither an abstract syntax notation nor a concrete transfer syntax. The unstructured bit stream between the two applications programs must be directly interpreted by the programs.

Because the PDUs of important MAP applications protocols, especially the MMS (Manufacturing Message Specification) services, have been specified in ASN.1, it is an obvious step to employ the same representation context for specifying the conformance tests for MMS. As Figure 303 shows, the conformance tests for the protocols MMS, Network Management (NM), and Directory Service (DS) are based on a set of ASN.1 tools with whose aid the test parameters and dynamic test cases for these three protocol groups have been formulated and processed. The

system of tools for conformity tests, as shown in Figure 303, were devised by the Institute for Information and Data Processing of the Fraunhofer Society. Though not completed at the time, they were used successfully in the conformity test at IENE '88 (International Enterprise Networking Event) in Baltimore, June 1988. It is expected that the representation context based on ASN.1 is sufficiently powerful to cover other Companion Standards as well. The standardization of PDUs, at least for the protocols named, is thus formalized to the extent that conformance tests can be derived from the actual protocol definitions in a direct manner, without substantial recourse to natural-language definitions.

8.9. Steps toward Computer-Integrated Production

Consistent information integration in production companies is the objective of the computer communication utilities derived and discussed in this chapter in the context of information logistics. While very important, they are by no means the only tools available for reaching this objective. Computer-Integrated Production (CIP) can be achieved only through collaboration between disciplines: production engineering (manufacturing engineering, control engineering, logistics, production planning, etc.), communications engineering (digital data transmission), and informatics (systems engineering, hardware and software technology). This integration of methods, in which information logistics is merely one part of the needed methodological edifice, is only just starting to gain support from products on the market and certainly cannot be achieved through standardized communications protocols alone. The custom CIP solution for a given plant often constitutes an analysis and modification of internal structures and modes of working.

The translation of such an internal reorientation to the use of computer communications techniques must take place in a continuously changing environment. Thus, MAP Versions 2.1 and 3.0 are not compatible; an early involvement in this protocol world gives rise to valuable experience, but also to a continuing need for new investment in more-advanced protocol products. Even within General Motors, initial experience with MAP was problematical. But it would be wrong to think that these unavoidable startup difficulties in a very complex field represent a failure of information integration. Outcomes of this experience have included more protocols on the applications level (or on all protocol levels at once), such as Network Management or Directory Service. In this way, gaps in the MAP/TOP definition are closed and the use of computer communications is facilitated.

The same applies to the present status of field bus standardization. Again, international differences in the view of applications requirements and product line definition must be ironed out in such a way that a single set of international standards can be devised in a usable form. The reaction of industrial firms to the very protracted standardization process shows once more that industrial users and manufacturer of communications systems are interested in open systems of information logistics.

All this development of tools will, however, be fruitless if users of CIP do not acquire the knowledge and training they will need to understand the total system. Along with education within the corporation, occupational training has a role to play. At the same time, there is a need for competent, neutral sources of advice. It is precisely small and medium-sized companies that lack the personnel resources to keep up with the rapid development of information integration and turn it to their own use, even though this assimilation is urgently needed to insure the future of the company's production and products. Regional sources of advice (chambers of commerce and industry) as well as research institutions cooperating with industry are increasingly available.

The path to computer-integrated production can be discerned today. It will lead, with much difficulty and many setbacks, to an improvement in the competitiveness and market position of those companies that have the commitment, creativitiy, and perseverance to follow this path.

9. Computer-Aided Methods

9.1. Principles

Coming years will see greater and greater importance attached to the use of computer-aided

methods and computerized tools for the design, erection, operation, and maintenance of processing facilities. The steadily increasing number and complexity of industrial plants and the requirement for complete, current documentation will generate a volume of work that could scarcely be handled without computer support. At present, however, the ratio of computer-aided tools to demand is still fairly small.

As the needed computer-aided engineering (CAE) systems are developed and selected, it is crucial that the engineering disciplines involved must be treated not as individual, independent working groups but as the intermeshed system they really make up, in which the results achieved in one discipline influence the work of the others. This does not mean that one "universal" tool has to be devised for all engineers. The model of a well-filled and organized "toolbox" appears to promise better success. Such a box contains a special tool for every job. The tools, however, are designed so that they all share consistent interfaces and thus can be combined with one another at will. Any problem that arises can thus be solved in optimal fashion; if the objective is changed, the "toolbox" can be adapted to the new situation through the modification of an existing tool or the addition of a new one.

This book does not deal with computeraided design tools for plant engineering (e.g., INTERGRAPH).

9.2. System Analysis

The background for system selection must include the formulation of overall objectives and possible interfaces.

The overall objective is broken down into partial objectives, which in turn are refined further into tasks. This procedure is continued until the individual tasks are manageable in their complexity and the functions needed for accomplishing them can be described along with the resulting data flow patterns.

To make this task analysis easier for the user and permit an unambiguous interpretation of the results, a variety of methods have been devised for displaying the process in graphical form with defined symbols. The best-known methods are structured analysis and design technique (SADT) [617] and structured analysis (SA) [618]. Either method can be carried out manually with pencil and paper, but increasingly they are embodied in software packages that make it possible to run these analytical tools on PCs and graphics workstations.

In what follows, the SADT method is briefly illustrated in a simplified example.

The essential elements of the SADT diagram are labeled boxes and arrows. Boxes symbolize activities (functions), while arrows stand for data and information and thus characterize the interfaces between boxes.

The following rules also apply:

- Input information always enters a box from the left, is transformed into output information in the box, and leaves the box on the right.
- Controlling information influences this transformation and may also be used per se in producing the output. Such information always approaches a box from above.
- The function of a box may be supported by mechanisms, which are optionally represented as arrows entering from below.

A model generated by SADT is a structured breakdown of a problem, represented as a hierarchically ordered sequence of diagrams. The topmost level of the hierarchy is a diagram with a single "black box" representing the total system, showing its connections to the outside. On the next level, this black box is broken down into a number of boxes; the same procedure is continued down the levels.

Figure 304 shows, in simplified form, the top two levels of an SADT model representing the engineering functions involved in the design, erection, and operation of a process plant.

The upper part of the figure shows the black box representing the overall system, with interfaces to the outside world. In the bottom part, the overall problem is broken down into the following subproblems:

- Design the process
- Design the plant
- Erect the plant
- Operate the plant

In the graphical presentation, the functions (boxes) are normally arranged on the diagonal from top left to bottom right. The order of functions is governed by the required flow of information and data, shown as arrows between the

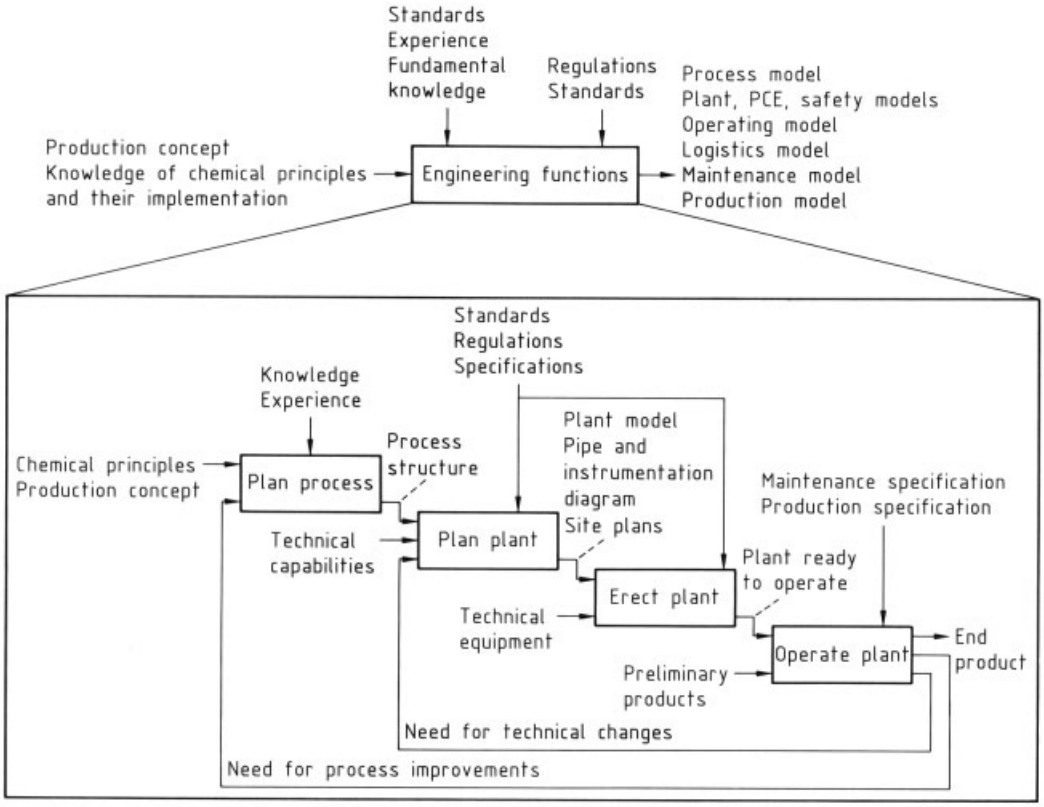

Figure 304. A SADT diagram (top two levels)

boxes. Feedback of information may also occur, in which case the procedure becomes interative. Interfaces to the outside world are also shown in greater detail.

Now each of the functions is further broken down in diagrams and documented. Obviously, complex problems yield many diagrams, all partially interdependent. Hence, if changes are made, these relationships must be taken into account and the appropriate corrections performed. The use of such a method thus calls for a considerable amount of administrative effort; on large projects, this is difficult to accomplish without suitable computer support. However, the implementation of a software project is virtually impossible without a prior system analysis.

In parallel to the design of the function model just described, it is necessary to define the overall system data model that applies to all functions. Only in this way can proper sharing of information among functions and consistent data management be insured. As a rule, the data model is set up as an entity-relationship diagram with data dictionary. Programs for this design task are also on the market; in many cases they are already combined with SA or SADT tools to make up software development systems (see also Chap. 2).

9.3. CAE System for Process Control Engineering

Within the overall system, the individual CAE tools used in the several engineering disciplines must have compatible interfaces. In addition each tool should have the optimal functional scope for the objectives of the field in which it is used. The special requirements on such a CAE system for use in process control engineering are therefore discussed in what follows.

A CAE system for process control engineering must support both the design and the later operation of the process measurement and control system, including maintenance (Section 11.4).

This latter aspect is particularly important here. Over a span of years, the design changes in an onstream plant due to detail improvements or process changes can easily loom as large as the original design. At the same time, the documentation must meet stringent standards of currency and consistency, since these changes generally must be performed while the plant is on stream or during brief shutdowns (see Section 3.5).

The CAE system must therefore be able to handle data having a variable and complex structure and to afford assistance in the processing of such data.

Process measurement and control functions are implemented in hardware and software. In practice, it has been found that a subdivision of the task into hardware design and software design is useful, and in fact these sub-tasks can even be carried out by different persons in the case of large projects. Both areas, of course, have access to a jointly defined interface, but are otherwise not temporally linked and can therefore work independently of one other.

The hardware data to be managed in the CAE system can be broken down into three classes:

- Station data: all information describing the task and the functional design of a process control station. As a rule, these are documented in station data sheets and station diagrams (see Figs. 339, 340, 341, 342).
- Device data contain the exact technical specifications of the devices employed. These serve as the basis for later procurement.
- Position data describe the configuration of the devices in the facility and contain the infrastructure created by the master cable. Typical forms of documentation are setup plans, terminal allocations, and cable plans.

The data describing the software have a different structure. Generally, modern process control systems have modular software. The modules are linked to one another and arranged into various levels of a hierarchy, depending on the task. The lowermost level is made up of modules that implement the functional design of a single process control station (Section 4.4). Higher-level modules coordinate the work of lower-level ones. Each module is composed of functional building blocks, which are the smallest software elements of a process control system. The most important documentation forms for software are function charts, logic diagrams, and element structural plans.

All these data groups are closely linked; together, they represent all the information about the process control equipment. In what follows, they are treated together under the term "project data." As a rule, these project data are managed by a database system (Chap. 10).

9.4. Structure of a CAE System

The ideal structure of a CAE system as shown in Figure 305 reflects the constraints discussed above.

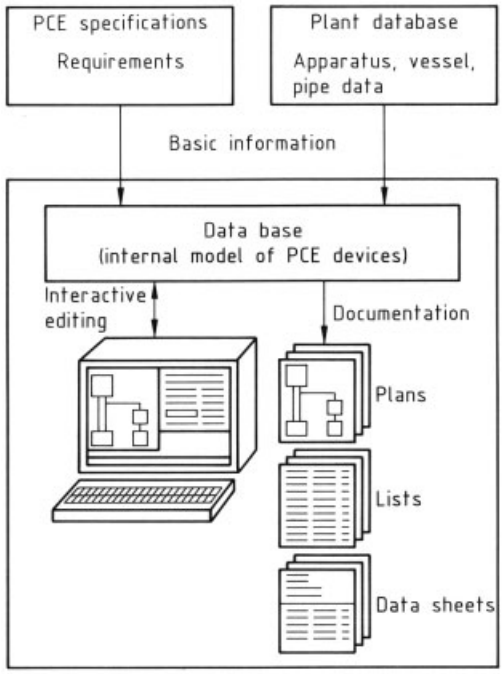

Figure 305. Structure of a CAE system for process control engineering

The heart of the system is a relational database, which contains the computer model of the planned equipment. All manipulations of the database take place interactively; depending on the task, graphical interfaces may be used.

At the start of the design work, basic information for the project is acquired from the other departments involved. The most important source is the process control tender specifications, which set forth requirements on the pro-

cess control devices to be implemented as well as their interconnections. Information is also required from the plant database, which supplies technical data on the plant sections affected as well as ambient conditions where the process control device is to be installed.

On this basis, the designer begins with the detail design of process control devices. This procedure is interactive and employs a graphical user interface. The system offers a number of tools that can, for example, prepare plans, enter specification data, define cables, or copy entire plans or cabinet terminal allocations. These manipulations gradually build up an internal model of the process control equipment in the database. Every interaction on the screen brings about a corresponding modification in the database.

This internal (computer) model also serves as the basis for all documentation. The various documents are merely the results of various inquiries formulated to the same database under various criteria. In this way it is insured that any subsequent changes will automatically be incorporated into the next round of documentation.

A CAE system with this level of complexity may consist of a number of largely independent modules, all of which communicate with one another through the common database. One possible subdivision into task areas is as follows:

- Hardware design
- Software design
- Procurement
- Installation

These are largely independent as far as processing is concerned and have access to one another only through a few well-defined interfaces. Accordingly, independent solutions are now on the market for certain of these task areas. When such products are employed, however, special attention should be focused on the possibility of later integration into an overall system.

Because process control devices cannot be designed and operated in isolation from the plant as a whole, smooth communication should be insured between this process and other departments, such as plant design, process engineering, purchasing, and maintenance [619]. The only way to be certain of this is to define a common data structure within the company: the company data model. Without such a data model, the danger exists that terms and information might be differently interpreted in the several systems, so that consistent data management would be impossible.

A crucial problem for the acceptance and efficiency of a CAE system is whether and to what extent the system user interface is adapted to the user's methods. Typical users impose varying, and in part inconsistent, requirements.

One user is the design engineer, who quickly becomes an "expert" through daily work with the system in a variety of modes, can operate it "in his sleep," and therefore wishes to get results quickly, with the fewest possible manipulations, and without having to read help messages.

In contrast, there is the operating engineer, who is confronted with the system only occasionally and briefly, for example during process upsets and when design changes are being worked out. The system should prompt this user in an intuitive way, with extra help functions, to avoid a prolonged learning process and the study of manuals. Because the drawing is the most comprehensible "language" in conventional design work too, it is an obvious step to design the entire user interface and system tools for largely graphical operation. In this way, the user's previous working methods are largely preserved, so that learning the system can be accelerated.

The use of hierarchically structured menu trees can drastically simplify the operation of choosing from a large number of devices or symbols. If the menus and the underlying database are cleverly designed, progressive narrowing of the field quickly leads to certain individuals without the use of lengthy catalogs and lists. This technique is widely used in modern systems [620].

9.5. Aids for Hardware Design

A number of CAE systems for process control hardware design have appeared on the market. They take two distinct approaches to implementing the task. One is to combine independent standard programs for database applications and 2-D graphics, using suitable auxiliary programs to link and supplement the commercial software. The other approach leads to an integrated system, in which the graphical tool is directly linked to the database and each is independent of the other. Both solution approaches will be

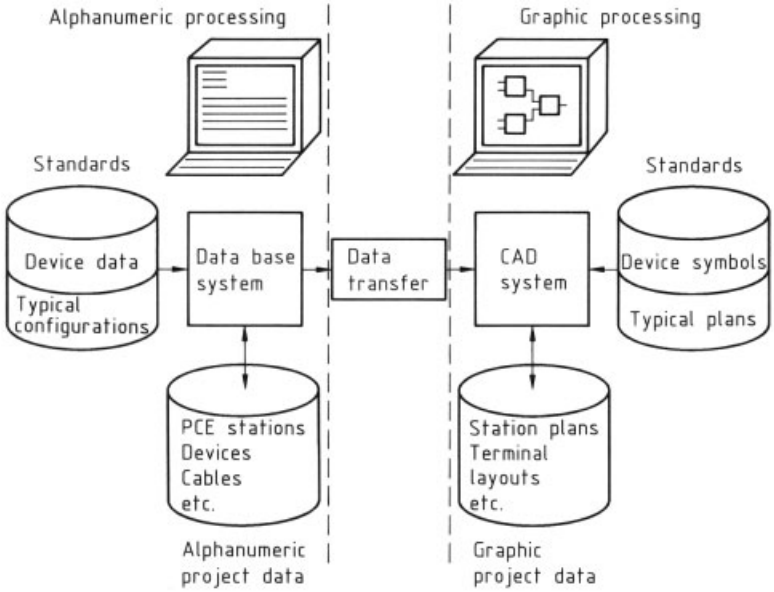

Figure 306. Structure of a combined CAE system

described in what follows, and their advantages and disadvantages will be discussed.

Combined CAE System. A combined system includes a free-standing, independent software package for alphanumeric tasks and one for graphical tasks, each package having its own data management features (Fig. 306).

Basic information contained in the alphanumeric package includes specifications on standard devices (process control equipment data; Figs. 10 and 339). Typical configurations, with information on combinations of devices for stated purposes, are also included as standards. In parallel, the graphical program package must have the respective symbols for the individual devices as well as typical plans for device combinations. It has to be explicitly verified that the typical configuration and the typical plan are logically consistent, since this forms the basis for subsequent automated processing (see also Figs. 340, 341, 342 and 61).

The first step in design is to input the process control stations, in the form of a station list, into the alphanumeric software and assign a typical configuration to each station on the basis of the task statement. These data are supplemented by structural information, such as the terminal allocations of cabinets and racks and the master cable connections. The devices used may also be specified more closely. Once the station and configuration data have been prepared, they can be exported to the graphical side with the aid on an auxiliary program. The respective plans are now prepared. Infrastructure information, such as terminal codes and the like, is entered into prepared text placeholders. The information managed in the alphanumeric module can thus be output as documentation along with the graphics.

There are a number of advantages to this approach. The use of common standard programs makes it possible to assemble such a system at relatively low cost. Because the programs do not have to run in parallel, the system can be and usually is executed on a PC with the MS-DOS operating system. The system as a whole is highly modular in structure, and the user can generally expand it with additional functions.

These advantages are partially offset by several drawbacks. The use of several software packages from different publishers means that the user must be familiar with the diverse operating concepts of the programs. Because configurations and plans are generated in different programs, there is a danger of input errors, which can lead to inconsistencies.

A combined CAE system thus has its strengths in the design areas, where a high degree of standardization is possible. If subse-

quent changes not complying with the standard are made, however, corrections must be made in both the alphanumeric and graphical components of the documentation.

Integrated CAE System. The integrated solution features a direct link between the graphics package and the database. Graphical manipulations thus influence the database contents immediately, while alphanumeric inputs influence the graphical representation on the basis of certain rules. This close connectedness both insures the consistency of the data and makes it possible to carry out plausibility checks. Interactive operation builds up an internal model of the process control equipment, which serves as the basis for all interpretations. All documentation, whether in graphical or alphanumeric form, is derived from this model (Fig. 307).

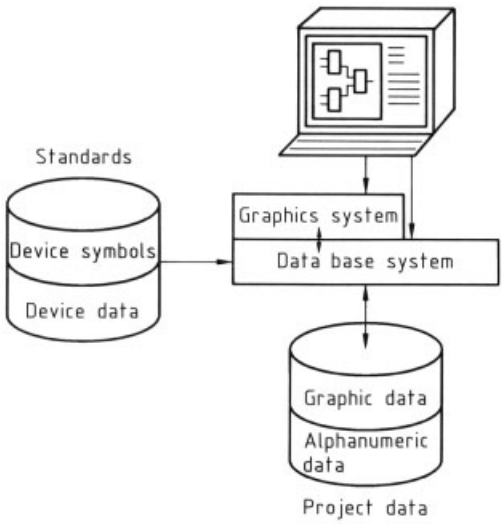

Figure 307. Structure of an integrated CAE system

The close coupling of graphics and database system usually requires the use of a multitasking operating system. The graphics packages employed are specially developed software with special provision for interaction with the database. UNIX graphics workstations serve as the hardware component.

The advantages of the integrated approach lie in the consistency of the data. In case of subsequent changes, corrections are essentially needed at just one place.

Because of their more complex structure, integrated systems are necessarily more expensive than combined ones. What is more, high-performance hardware is required if an acceptable response time is to be achieved [620–623].

Over the long term, the integrated CAE system is the better solution. Its present cost disadvantages are being offset by technical developments in the workstation market.

9.6. Aids for Software Design

Hierarchical Organization of Software. The principles of functional organization and progressive refinement have been tested in practice for the description of complex relationships (see Chap. 2). A good example of such a procedure are the design steps involved in the development of a complex integrated circuit such as a CPU. First, the required functions are represented in the form of modules at a fairly abstract level, and the requisite interactions between these modules are described. Next, the internal structures of the individual modules are established; the same principle is applied recursively in this step. The objective is thus narrowed down to the level of single transistor functions. In the subsequent fabrication operations, the complex functioning of the circuit is achieved through the combination and interconnection of many very simple subfunctions.

A similar procedure can be applied to the preparation of software for process control systems. The overall task is broken down into sub-tasks with well-defined interfaces (see Section 5.2). By narrowing down the description, one ultimately reaches the lowermost level of abstraction, which in the case of process control systems is formed by the components and logic functions available. The software for a system is finally written by appropriately combining these basic elements (Fig. 308).

If there is so much similarity in the procedures, the question arises whether parts of the highly advanced commercial CAE tools for circuit design might not be usable in modified form for process control system software development. In what follows, the minimum requirements on such a tool are stated briefly.

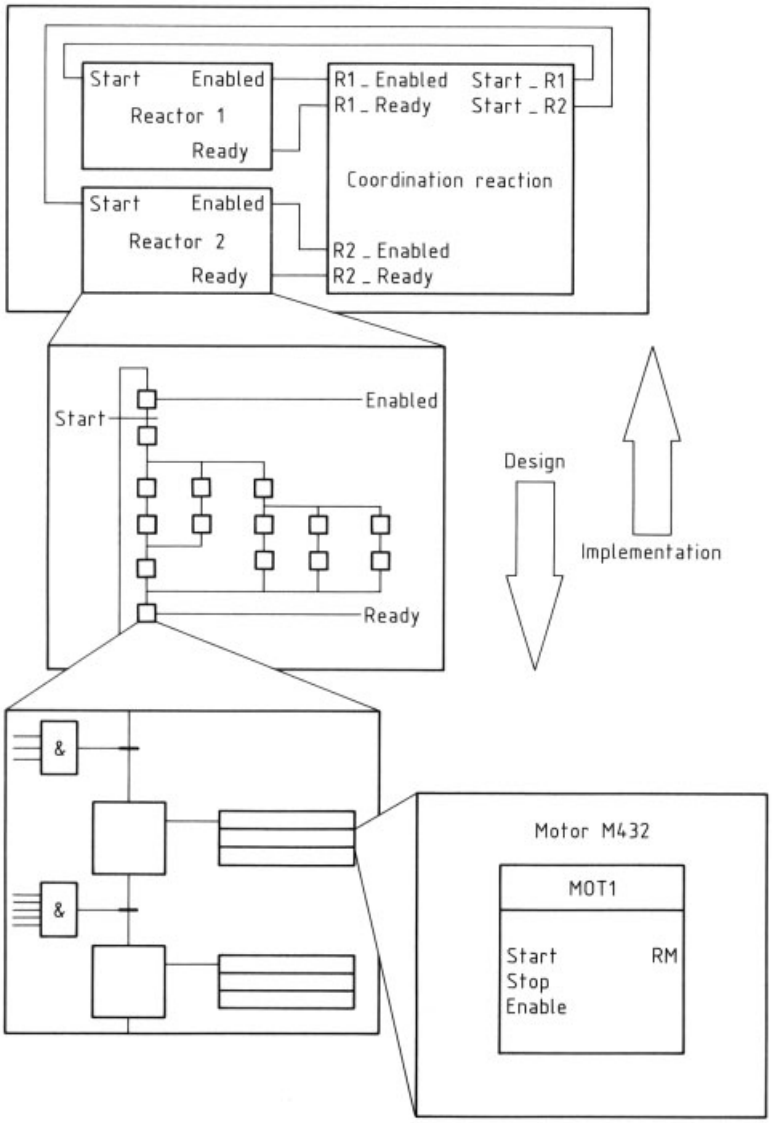

Figure 308. Progressive refinement of a software structure

Graphical Generation of Function Charts and Structural Diagrams. The building blocks or logic functions of the process control system, as furnished by the publishers, form the basis of the applications software. The stated task is accomplished by combining these, in suitable form, to make structures. The most important aid in design and subsequent maintenance is the function chart or structural diagram (Figs. 59 and 343). It shows graphically all the components used, with their interconnections and parameters. The creation of a graphical tool for generating these diagrams is crucial. The element symbols are selected from a library and placed on the screen interactively. The connections between elements are generated by constructing lines between the element terminals. Parameter values can also be interactively assigned to specific element terminals. The logical content of these diagrams must be determined by the system and automatically translated into a sequence of programming instructions for the process control system. Plausibility checks can be performed to eliminate programming errors

directly at the source. This procedure simultaneously insures consistency between the program and the documentation (see Section 10.6).

The CAE tool must support the method of successive refinement described above. The user must be able to generate abstract modules and their connections in a plan without needing to know their internal structures. Internal structure can be represented in a later diagram; the system automatically takes care of cross-referencing. In this way, complex structures can be detailed step by step, down to the component level. During code generation, the system must recognize and take into account information about the entire hierarchy of diagrams.

Another important point is the use of "typicals." These are multicomponent structures that can be used without alteration in many diagrams. Constructing a good library of typicals can drastically reduce the required design effort. At the same time, it leads to standardized software, since a fixed combination of elements can always be used for like functions. Such a capability must therefore be implemented in a CAE tool (see also Section 4.4).

Apart from building blocks defined specifically for each process control system, there are system-neutral descriptive forms for logic and sequential control systems. The appearance and interpretation of these are specified in ISO and DIN standards. It should therefore be possible to program controls with the aids described above, using the standard symbols. The translation to the language of a particular process control system is another function of the CAE tool.

Construction of Screen Diagrams. Like the preparation of component-oriented software, the programming of the user interface for a process control system is a particularly critical point (see Section 11.2). At present, this represents a major design effort, requiring support by CAE tools. Graphical plant diagrams used in interactive mode have come into widespread use in the user interfaces of process control systems. The graphical design of such images should be performed with tools whose quality and ease of use are consistent with those of commercial CAD systems. The definition of dynamic displays or controls might be done by marking the position and assigning the variable name. Diagrams constructed in this way should be usable for all components of the process control system without case-to-case modifications.

Documentation of Existing Facilities. A special problem arises when existing plants must subsequently be documented (see Section 3.5 and Chap. 10). Often, the current program no longer corresponds with the original documentation because changes have been carried out commissioning or maintenance. All process measurement and control systems can document the current program structure in the form of component and configuration lists. This form of documentation is, however, difficult to understand, and in practice it cannot be employed for troubleshooting. As a rule, therefore, there is no alternative to changing or regenerating function charts and structural diagrams on the basis of the current lists, but this procedure is time-consuming and expensive. Users are therefore demanding automatic graphical documentation of the current state of a process control system.

Synthetic diagram generation on the basis of information present in the control system constitutes a basic problem. Because the system contains no information about the original form of the diagram, the component and terminal positions needed for the representation must be determined by using algorithms such as "information flow top to bottom" and "minimize feedback paths." The consequence is that a synthetic diagram with given content may have a different graphical organization from a hand-drawn diagram. Furthermore, even small changes may alter the positions of some components relative to the old version because of differences in evaluation in the placement operation. This property may interfere with the immediate re-identification of a circuit. Nevertheless, retrospective graphical documentation can insure complete consistency between the documentation and the current program. For this reason, users are willing to tolerate the disadvantages of retrospective documentation until a better alternative becomes available.

System-Neutral Programming. Various manufacturers of process control systems have developed CAE tools, each tailored to a particular system. Although such a package provides effective support to the user, it also poses a new problem. Large firms and large engineering

offices, which usually prepare designs for a variety of control systems, would have to employ a large number of diverse, system-specific CAE systems in order to simplify software writing. Because these systems are mutually incompatible, porting existing application software to a new system ultimately means doing a complete redesign. Users have therefore welcomed the efforts of the IEC to devise a neutral, standardized programming language for process control systems. With such a language, system-independent and thus portable technology applications could be developed. Using such packages, the user could eventually perform tasks in a problem-oriented way regardless of which CAE system was in use.

9.7. Outlook

The development of CAE systems has been strongly influenced by rapid advances in computing hardware and software. If users are not to find themselves isolated in a few years and restricted to possibly obsolete computer systems, modern CAE systems must be made independent of both hardware and software. This objective can be achieved only by means of international standards.

In the operating systems area, UNIX Sys. V or OSF can serve as a standard. Most publishers have now adopted the MIT X-Windows System as the basis for the graphical user interface. A number of competing tools are available for the actual design of the user interface, but OSF MOTIF is emerging as the favorite. A CAE system developed with these tools can thus be regarded as having a good future.

One particularly critical point is the data management problem. Because plant design data generally must be kept for 20 to 30 years, stringent requirements apply to the dependability and development potential of the database program publisher. At the same time, exotic data structures should be avoided so that data can, at least in principle, be imported into a different database system. In the context of workstation networking, the potential of decentralized data storage should be investigated.

Graphical workstations, connected to one another and to service computers by local networks, will become the standard hardware. The simultaneous rapid rise of computer performance and decreasing hardware costs will allow use of high-performance solutions. The high requirement for storage space will be partly answered by advances in rewritable optomagnetic discs, which offer high data security and long-term stability.

10. Design and Construction of Process Control Systems

10.1. Principles

The previous chapters have described the functional and hardware elements of process control engineering. This chapter deals with the design and construction of process control facilities.

It is seldom that process control problems arise in isolation. As a rule, they are embedded in an overall problem. This may concern an investment project in the field of chemical plant construction; process control engineering then represents one component of the problem, while other disciplines such as machinery and apparatus, building construction, and piping are also involved.

In the design and construction of plants, a distinction is made between new construction and extension or reconstruction.

Extension and reconstruction activities are generally more difficult to plan than new construction projects, since interfaces to existing plants have to be taken into account and the existing infrastructure and engineering must be modified. The overall extent of the work usually only becomes apparent when planning is already under way. A complicating factor is that economics often dictates the incorporation of changes while the plant is on stream or during shutdowns. Because the scheduling and duration of shutdowns is governed chiefly by operational constraints, however, planning of schedules and resource utilization must be done with special care (Section 10.2).

New construction projects based on processes already in use within the company are generally simpler.

The handling of processes new to the company (i.e., created in process development or purchased from a licensor) is more difficult. Scaleup problems can involve substantial risks.

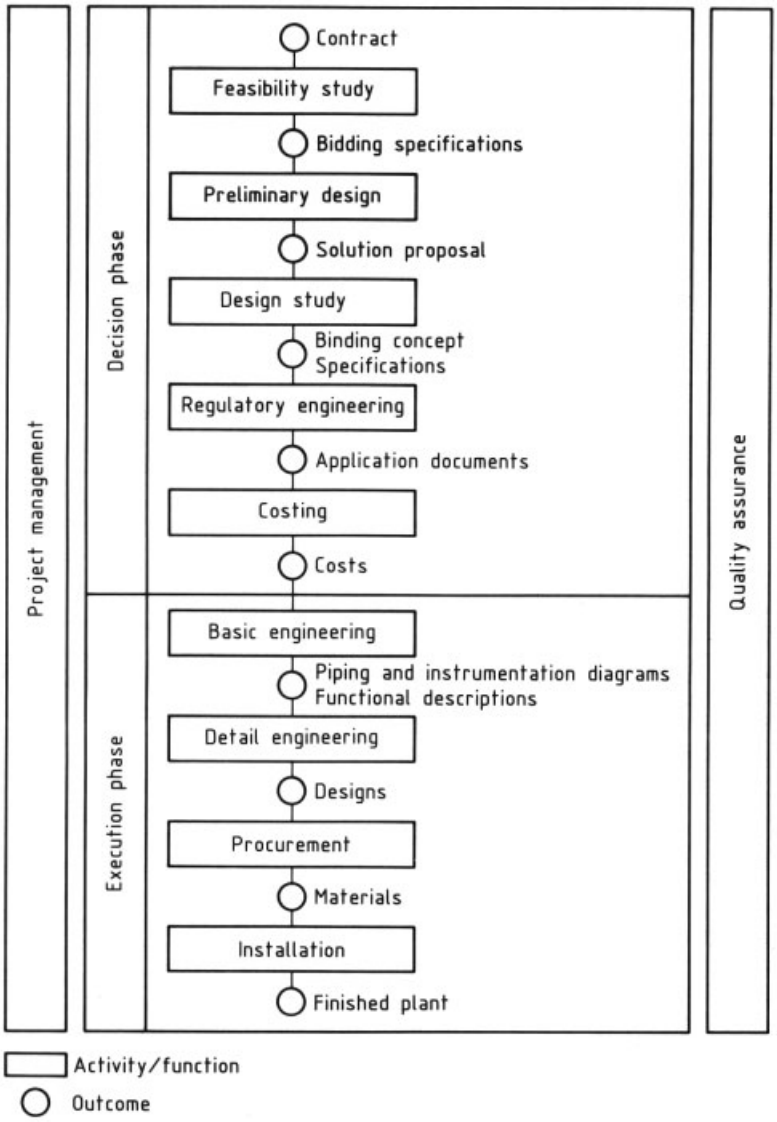

Figure 309. Phase model for chemical plant construction projects

Figure 309 gives a model for the implementation of chemical plant construction projects. The main blocks represent

- Project management
- Project planning and execution
- Quality assurance

The design and construction of plants is carried out in the framework of projects. Project management (Section 10.2) accompanies project activities, coordinates the work, and monitors compliance with directives.

In Germany, the usual approach project planning and execution is based on the regulation governing architects' and engineers' fees [624]. It is useful to illustrate the procedure in the form of a phase model, which breaks tasks or functions down into phases and steps and shows their outcomes (Fig. 309).

Note that the phase model represents an ideal. In actuality, the several phases merge continuously and may even overlap. Feedback and iteration between phases are also common. Inter-

faces between disciplines further complicate the process.

The phases can be grouped into two large phases on the basis of the problems to be solved; these are called the decision phase and the execution phase [625].

The decision phase (Section 10.3) involves the conceptual work. The task is formulated here; it can be thought of as the mission to be performed in the course of the project. The written description of the problem is called the performance specifications (Section 10.4); it forms an essential component of the decision phase.

The performance specifications give rise to the "binding concept," embodied in the specifications, which describes how and with what resources the stated task is to be performed (Section 10.4).

The concept has long-term consequences, since it influences the economics of the plant throughout its service life. Identifying the concept is therefore accorded special importance. This step requires great care and is supported by analyses of alternative solutions as well as an examination of the overall project.

One task may admit of several solutions, among which the optimal one for the company (from both economic and technical standpoints) is to be selected. The selection criteria may include

- Attainment of targets
- Fulfillment of economic criteria (Section 10.4)
- Compliance with regulations (Section 10.6)
- Availability of plant
- User-friendliness
- Use of proven techniques

The total analysis of all disciplines involved in an investment helps to prevent the overall project being harmed by suboptimization of individual disciplines.

The binding concept is the basis for corporate management's decision to proceed further.

The execution phase (Section 10.5) is oriented toward workmanship. Using results derived in the decision phase, it creates a functional plant.

When the chemical plant design and construction process is divided into decision and execution phases, with the stress on holistic analysis and conceptualization, the result is that the problem is handled in a fundamentally different way from what used to be common (see also Chaps. 2 and 8).

In the past, planning for certain functional specialties, including process control engineering, often caused delays in other parts of the process, such as plant design. The tasks of the "trailing" discipline were sometimes formulated from the standpoint of the "leading" one.

A holistic view of the task could scarcely come about in this way. If such a view was formed in the course of the project, any corrections arrived at through such an analysis, no matter how logical and necessary, were not carried out, simply because the planning processes in the several disciplines had reached such very discrepant stages by that time that revision of existing plans by the methods then in use could not be justified in terms of cost and lost time. The systematic use of data-model-oriented planning in plant construction, such as the Intergraph plant design system, makes it possible to implement changes such as those described in Chapter 9.

In the past, compromise solutions replaced logical, holistic solutions; suboptimization of individual disciplines was the rule.

The complexity and diversity of techniques now in use, together with the resulting interdependence, have led to continuous interaction between disciplines, virtually barring such an approach.

Only the analysis of the project in its totality, coupled with the conceptualizing approach and an intensive quality-assurance effort (Section 10.6) going along with the execution of the task, can satisfy the requirements imposed on the economics, safety, and environmental compatibility of engineering solutions today.

10.2. Organizational Requirements

The organizational concept in the execution of projects is called project management.

The term "project" refers to the performance of individual tasks bounded in time and having clear objectives; these tasks may further be characterized by features such as interorganizational components, many involved divisions, high level of complexity, and possibly heightened risk [626].

Various types of project exist, including

- Construction projects
- Plant projects
- Organizational (management) projects
- Research and development (R & D) projects
- Software projects

In general a project is the complete temporal framework of conditions of all substances for the realization of tasks represented by functions [627]. "Substances" are suitable procedures, systems, and elements (Figs. 10 and 65) and are also called "resources"; Figure 310 illustrates this relationship.

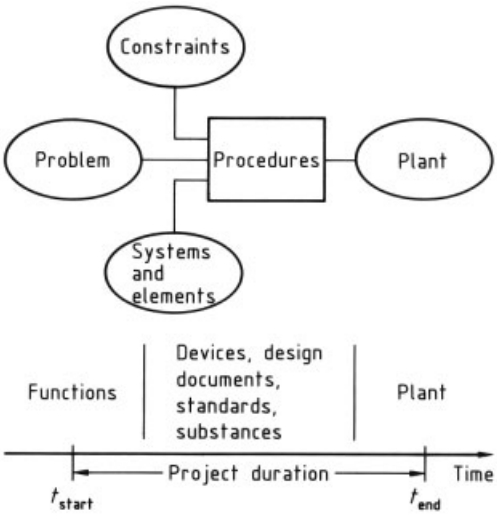

Figure 310. From problem to plant

In order to achieve the project objective, these resources must be procured, combined, coordinated, and utilized. Project management is responsible for these functions.

Organization. The organizational setup of project management differs from the "line" organization generally found in companies.

Organizational Forms. A variety of organizational forms are seen in project management. The three most important are briefly described below [628].

An across-lines form, also called "matrix" project organization, has proved especially successful (Fig. 311). The customary line organization is retained but is enlarged by a further dimension – project management. Certain aspects of the project, such as technical responsibility, remain under line jurisdiction; others, such as project responsibility, are transferred to project management.

This form of organization offers the advantage that personnel resources are optimally utilized, since employees remain in the project only as long as they are needed. As a result, specialist knowledge and special experience can be more easily brought in as well as transferred to subsequent projects. These advantages are partly offset by some drawbacks, in particular jurisdictional conflicts between line and project management and the relatively great effort required to demarcate responsibility.

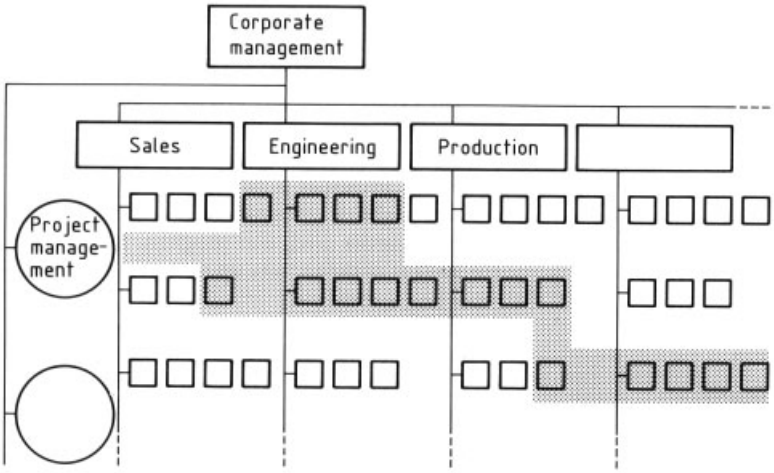

Figure 311. "Matrix" project organization [628]

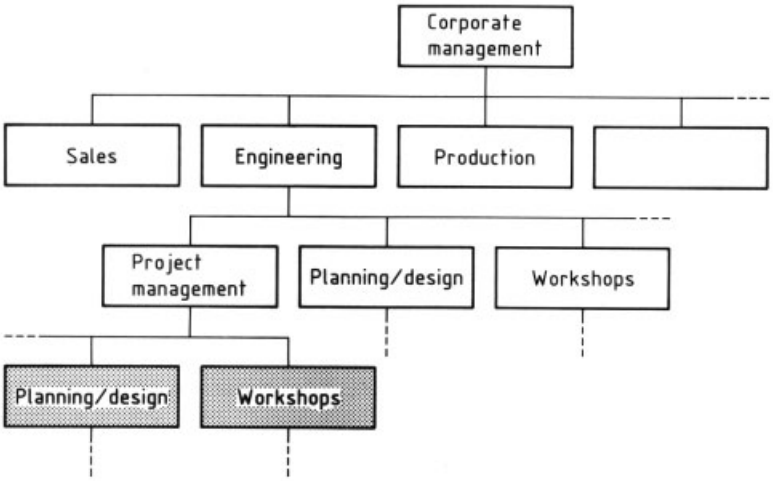

Figure 312. "Straight" project organization [628]

The advantages, however, outweigh the disadvantages. The matrix form of project organization has therefore found a wide variety of applications. It has been broadly accepted, in particular, by the chemical industry.

In "straight" project organization (Fig. 312), a separate, independent project line is created in addition to the main lines; project management has its own organization partly similar to those of the main lines. Project responsibility and technical jurisdiction are combined at a single point. Quick reactions are possible when difficulties arise in the planning and execution of the project. Against these advantages stand difficulties in recruiting personnel at the start of the project and in keeping these people occupied after its completion. Straight project organization is mainly used in large, long-term projects.

In "influence" project organization (Fig. 313), project management is simply a coordination instrument, without the authority to issue orders. It tracks progress with respect to schedules and costs and, if necessary, suggests actions to the respective line jurisdictions. Project management has no responsibility for the project; it is responsible only for providing timely information to the line jurisdictions. This arrangement does not favor successful cooperation across lines. When difficulties arise, reactions usually come slowly and delays result.

This form of project organization is usually practiced on smaller projects and projects involving several companies.

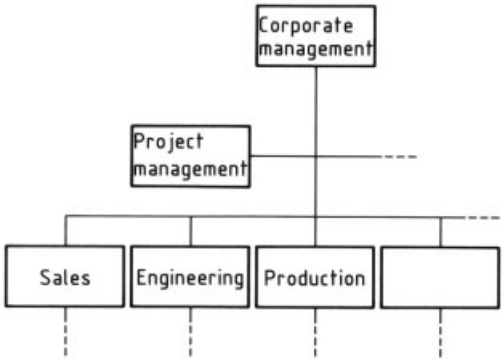

Figure 313. "Influence" project organization

Organizational Structure. For project management with matrix organization, especially in the chemical industry, a threefold division of responsibility has proved logical and useful (Fig. 314):

- Project management committee
- Project manager
- Project engineer

The project managment committee is the actual supervisory and decision-making body. It includes the decision makers of the crucial lines involved in the project, i.e., those issuing the order for it and those executing it. The project management committee formulates global mission targets. It monitors the complete and onschedule attainment of these targets, makes the needed decisions when targets come into conflict, and causes these decisions to be implemented in the

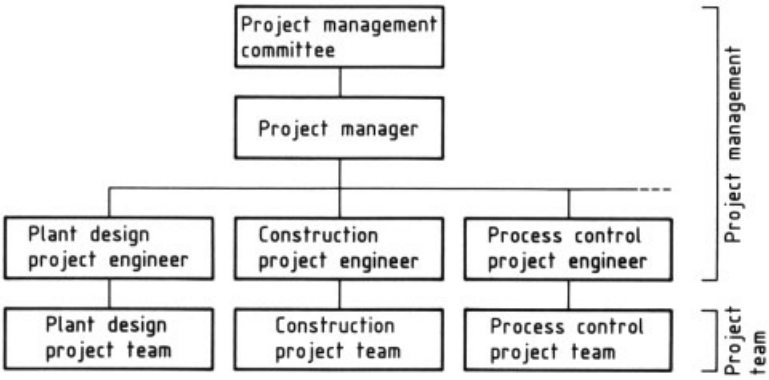

Figure 314. Organizational structure of project management

lines. The committee insures the availability of material and personnel resources.

The project manager is responsible for structuring the project. His duties include the definition of sub-tasks in accordance with the project structure. Principal activities are the planning and oversight of schedules, costs, materials and other resources used, and progress. In addition, the project manager sets up and implements the reporting system.

The project engineers support the project manager by providing organizational and technical guidance and oversight to the project teams, which are detailed by the lines to perform subtasks within the project.

The makeup of project management depends on the size and importance of the project (Fig. 315). The "size" of a project can be defined in a variety of ways; for investment projects, the level of investment is the key criterion.

In large projects, the project management committee is made up of delegated decision makers from top management. For small to medium sized projects, on the other hand, decision makers from lower strata of the hierarchy form the committee.

For large projects, a head-office project manager is designated; this person always performs management functions. Smaller projects do not require so much expense for the project management function, and a number of alternatives exist here. One manager can manage several projects at once, or one manager can focus on a single project but add specialist functions to management ones. The manager then becomes a participant in the project. The same holds for the

	Project management committee	Project manager	Project engineer
Key projects, projects of strategic importance	Decisionmakers from top levels in the hierarchy (top management)	o for one project: with pure management function	o for one project: with pure control and monitoring functions
Medium-sized projects	Decisionmakers from middle levels of the hierarchy	o for several projects: with pure management function or o for one project: with management and technical specialist function	o for several projects: with pure control and monitoring functions o for one project: with pure control and monitoring functions
Small projects	Decisionmakers from lower levels of the hierarchy		

Figure 315. Project management organization as a function of project size and importance

Figure 316. Tasks of project management

project engineers: They have 100 % of their time assigned to large projects, a lower percentage for smaller projects.

Tasks. Figure 316 summarizes the functions of the project management committee, project manager, and project engineer.

Major management activities include project structuring as well as cost, schedule, and resource management. Performing these functions requires close cooperation between all the activities involved in executing the project.

Project Structuring (see also Chap. 2 and Sections 3.5 and 4.3). The nature, scope, and complexity of a project dictate the manner and depth of structuring. This activity results in structural charts and can be based on a variety of approaches, with a corresponding variety of structural charts as outcomes [629].

The object-oriented structural chart breaks down the material and immaterial objects to be created by the project into components in hierarchical fashion. Figure 317 shows a portion of such a structural chart for the erection of a chemical plant.

The function-oriented structural chart shows the breakdown of project activities.

The structural chart of technical scope of performance shows what intermediate and final outcomes are to be generated by the project. All documents to be produced, such as reports, plans, drawings, lists, descriptions, certificates, contracts, and so forth are included, regardless of whether they will still be needed after the end of the project.

The last two types of structural chart are seldom made up by themselves; a combined form is usual.

The main goal of project structuring is to determine the total volume of work necessary for the success of the project. In this way, the preconditions for further project planning are fulfilled,

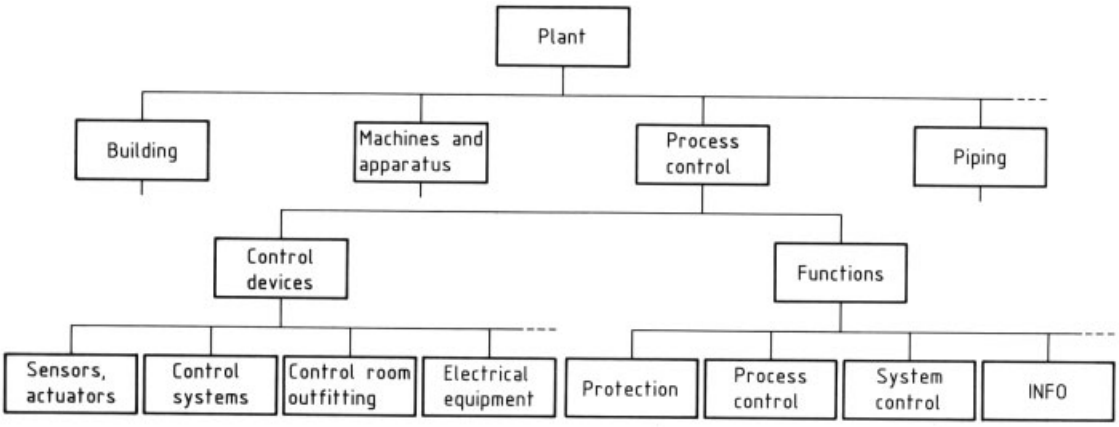

Figure 317. Example of an object-oriented structural chart

Total investment
- Apparatus, machinery
- Buildings
- Piping
- Process measurement and control
- Miscellaneous

Factor $\varphi_1 = f$(structure type, plant type, ...)
Factor $\varphi_2 = f$(complexity, materials of construction, ...)
Factor $\varphi_3 = f$(functional scope, required availability, safety aspects, ...)
Factor $\varphi_4 = f$(...)

$$\text{Total costs} = (1 + \sum_{i=1}^{n} \varphi_i) \cdot S_0$$

S_0 = total machinery and apparatus cost
S_i = total cost of category i
$\varphi_i = S_i / S_0$

Figure 318. Cost estimation by the multiplication factor method

since project structural charts influence, among other activities, the planning of schedules, costs, and resource utilization [629].

The transparency achieved through project structuring makes is possible to identify key tasks. It becomes easier to coordinate project teams and delegate their sub-tasks. Economic, technical, and also organizational risks, such as jurisdictional conflicts between project management and lines, can be delimited. Project structural charts are a prerequisite for effective project monitoring.

Cost management refers to all cost-related activities in the course of a project, such as costing, planning, monitoring, and billing.

Cost management is of central importance for a project. It must cover in detail

- Procurement costs for machinery, apparatus, buildings, process control equipment and systems, infrastructure, and other materials
- Erection and installation costs
- Engineering costs due to planning, design, and administrative activities
- Start-up costs

The specialist literature offers a number of methods for determining procurement costs in chemical plant construction [630]. The accuracy of the results depends on the method used, the depth of knowledge about the project, and other factors.

It is beyond the scope of this chapter to describe all methods in detail. Just two, widely used in plant construction, merit discussion.

The first can be called the "method of multiplication (or ratio) factors." The simplest application of the technique requires a rough, object-oriented breakdown of the overall costs into, say, the categories:

- Machinery and apparatus
- Buildings
- Piping
- Process control equipment and systems
- – Miscellaneous

The machinery and apparatus needed for the realization of a process are the first things to become known during an investment project. They can be calculated from offers or empirical figures and adjusted for inflation and market trends.

The principle of the method is that the costs of all other categories involved in the investment can be determined by using ratio factors based on the costs of apparatus and machinery (Fig. 318).

The person practicing the method needs experience to obtain usable results, since the factors include scope-dependent ranges. For example, office structures must be costed differently from industrial structures, and larger factors should be used for new "green-field" plant investments than for extensions based on existing buildings. The assessment of piping costs depends on, among other things, the complexity of the piping structure and the material of construction. Influences on the ratio factors for process control equipment include functional scope, required availability, and safety aspects.

Experience has shown that the method of ratio factors makes it possible to estimate costs with an accuracy of ca. $\pm 20\,\%$.

The second cost-estimation technique, is "unit-cost estimation." Like the ratio-factor method, it is based on an object structure, but a more detailed one.

The simplest case for cost estimation is the procurement of "substances" (see Fig. 10), provided the quantity structure is known accurately enough (see also Fig. 332).

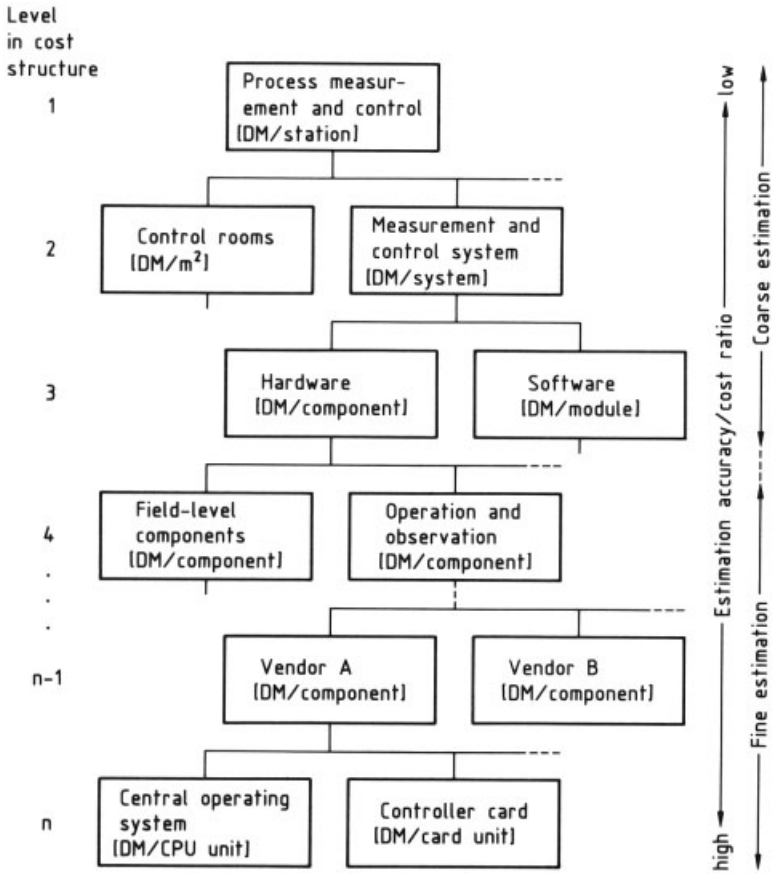

Figure 319. Example of a cost structure for unit-cost estimation

To every object at a structural level, the technique assigns a price referred to an object-specific unit (per item, unit area, etc.). The total price is obtained by multiplying this unit price by the quantity. For example, a possible cost structure for process measurement and control is shown in Figure 319.

This method is suited both to coarse and fine estimates, since cost estimates can be obtained for virtually any degree of subdivision. The accuracy of estimation increases with the level of detail. At the same time, however, more effort is required because the detailed breakdown calls for thorough knowledge, which can be acquired only gradually and after the project has been started.

The amount of effort can be reduced if the in-depth analysis is restricted to cost-relevant items while only a rough estimate is done for items of less relevance to the cost. This procedure does not result in any marked loss of accuracy. Cost-relevant items are those that make a large contribution to the total cost, for example by virtue of large quantities and/or high unit prices.

Incorrectly determined unit prices have less effect on the estimated cost than does an incorrect quantity structure. Experience indicates that this is where most errors occur.

Multiplication factors and unit prices are obtained by evaluating the invoices for a number of investment projects already carried out, extracting the cost structures, and then taking averages.

Today, installation is rarely carried out by in-house departments, and is mostly performed by contractors.

Installation can be paid for in any of three ways [631]: by expenses, by unit price, or by fixed price.

When charges are based on expenses, only the price per hour of work is established; the

fee is then calculated from the number of hours worked. This form of payment requires careful, continuous oversight.

When the unit price basis is used, a price is agreed on for each unit of performance. The total charge is the sum of these unit prices. This form of payment is suitable especially when changes and additions to the cost category are accepted. This billing system makes these particularly easy to handle. The amount of bookkeeping effort involved, however, is not inconsiderable.

The fixed price system calls for an exact description of the work to be done, largely completed plans, and installation conditions free of disruptions. The administrative cost is low. Changes in required installation may, however, lead to substantial cost overruns.

Engineering costs include all costs arising from

- Planning design activities
- Contract administration
- Supervision of installation work
- Start-up

In Germany, for example, engineering costs are largely set by the regulation governing architects' and engineers' fees [624].

Design costs are generally estimated on the basis of unit prices for stated functions. These activities can, however, vary in difficulty, and this fact is reflected in variable ratio factors. Some planning and design work, for example in the areas of feasibility study and project management, will be roughly equal in volume for any design project. This weighs more heavily on small projects than on large ones, so that small projects are more costly than large ones.

Contract administration and installation supervision are calculated on a percentage basis; commissioning is calculated in accordance with the man-hours required.

By way of example, Figure 320 (see next page) shows the cost structure of an investment for erecting a green-field chemical plant.

Cost estimates are made in parallel with the planning and design phase of a project. They become more accurate as the planning progresses.

The results of cost estimation form the basis for a profitability analysis and become part of the application for approval of the investment. Upon approval, the estimated costs become the binding cost framework for the project. Overruns are permitted only in a few justifiable (and exceptional) cases.

Adherence to the cost schedule requires continuous monitoring of costs over the project; an accounting of costs must be presented at the end of the project.

Schedule Management. In simple cases, scheduling can be carried out by using bar charts (Fig. 321, top), which show the beginning and duration of each activity as well as the earliest and latest occurrence of procedures along with buffer times. Such charts can be prepared at various levels of detail; if they are properly maintained, they provide a good overview of the schedule situation at any time. However, their critical disadvantage is that they do not show the interdependences between activities. For projects having a complex sequential structure, scheduling cannot be planned or overseen on this basis. Network charts can be of use in these cases [632].

In network charting, activities and relationships are stored in network files and graphically plotted, for example as shown in Figure 321, bottom. This technique makes it possible to determine not only the earliest possible but also the latest acceptable inception points for activities. Furthermore, if buffer times are analyzed, "critical paths" can be identified. A critical path is a path through the network that has no slack time.

Projects may contain more than one critical path, and these require especially close supervision during project execution. Schedule changes may open up new critical paths, but their influence on the overall schedule situation can be calculated in advance, allowing prompt reaction to such changes.

Network methods offer the project manager an aid in focusing on essentials in schedule supervision. The manager thus becomes free to perform other management tasks.

The charts used for this purpose must be continually updated. Because of the work involved in updating and the complexity that is sometimes associated with network charts, expecially in large projects, it is recommended that specially trained personnel be assigned to this job. A hierarchical procedure is also applied to network charts, where each decision maker is furnished with the requisite information structure.

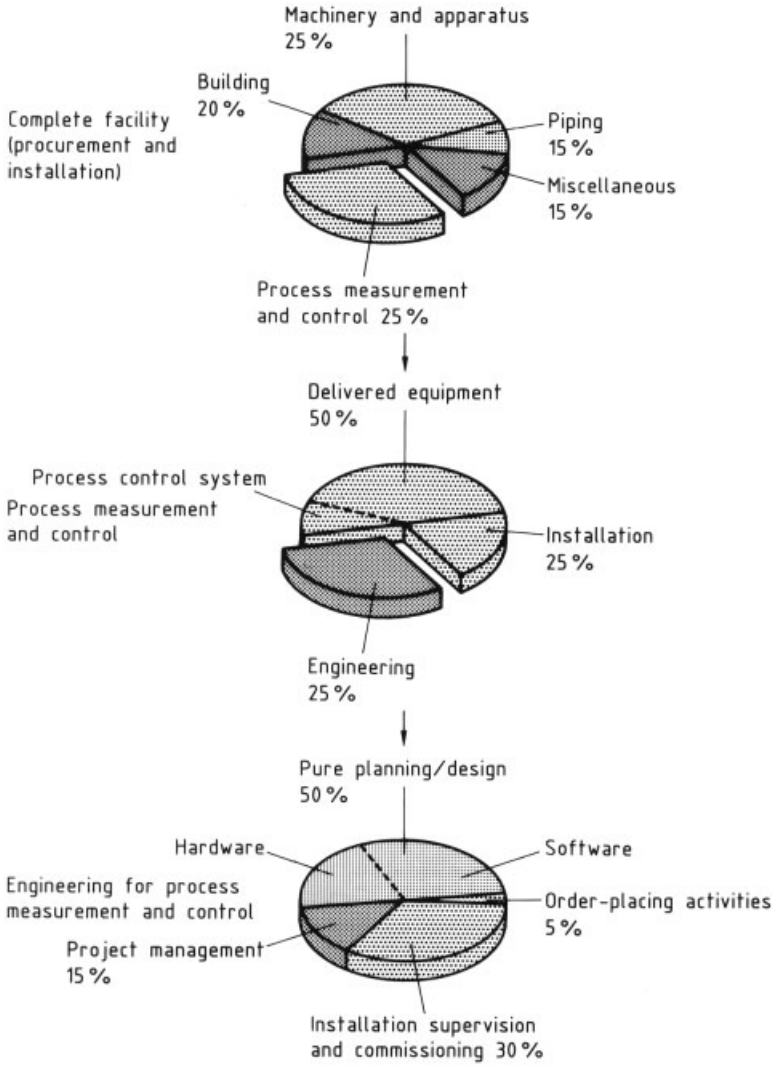

Figure 320. Cost structure of a chemical plant investment

Petri networks (see Chap. 2) are the basis for all types of network charts.

Resource Management. Resource management (resource allocation) is an extension of schedule management [629], whereby "resources" are personnel, machinery, funds, etc.

In general, schedules are drawn up as if the needed resources were available in adequate quantity at any time; however, only rarely is this actually so. Generally, in-house and outside resources are limited. Resource shortfalls or the complete unavailability of resources will result in delays if, for example, certain activities must be carried out later or not at all, or if activities that should be performed in parallel and have been so scheduled, have to be carried out in sequence instead.

Such imponderables in scheduling can be avoided if scheduling and resource planning are combined. There are two approaches: resource-constrained allocation and resource leveling.

Resource-constrained allocation is based on what resources are available. Deviations from the original schedule are allowed, but the project duration must be made as short as possible.

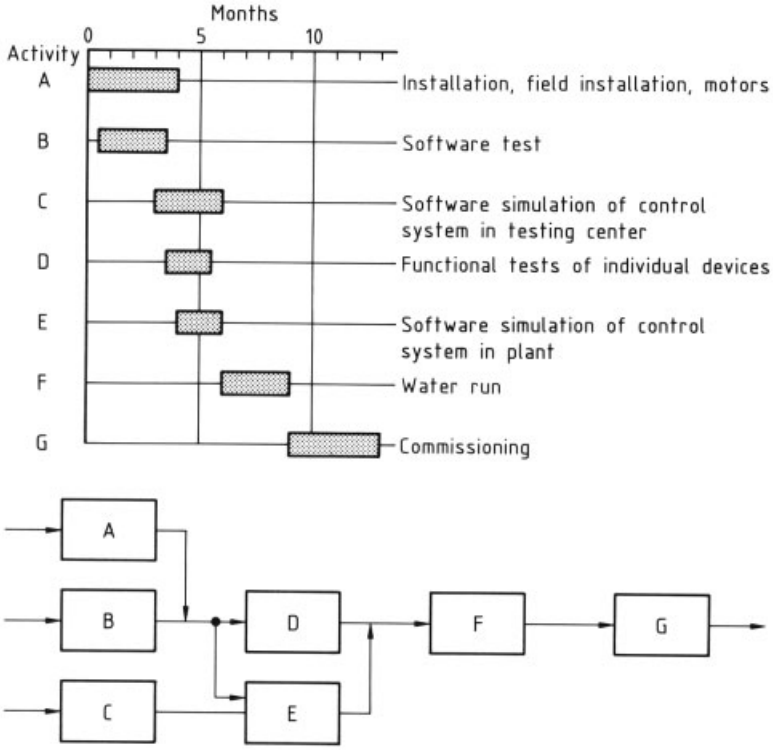

Figure 321. Bar chart (Gantt chart) and network chart

In resource leveling, the specified completion date is the top priority and additional resources are procured to the extent required to achieve it.

In practice, an intermediate course is generally followed.

The resource "personnel" is often a cause of special concern. In boom times, for example, a resource leveling approach is questionable because not enough of the specialists needed for execution of a project may be available. One way out of this conflict, though it is seldom considered, is to reduce the originally stated scope of work. The prerequisites for doing so are that the quality of the work will not suffer and that retrofitting will be possible later on at no significant extra cost.

The problem of resource management becomes more complicated if several projects differing in structure must be planned at the same time, as often occurs in the chemical industry. The problem can often be solved only by assigning priorities to the projects (Fig. 322).

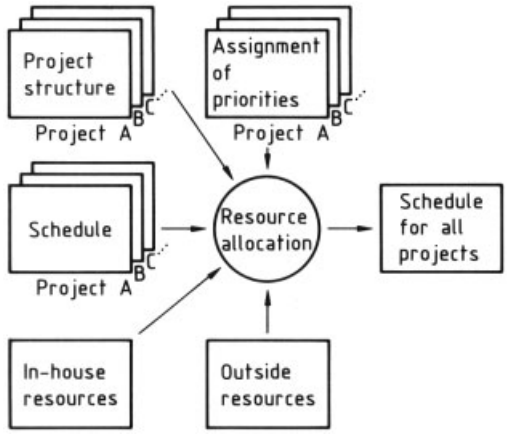

Figure 322. Resource management

Project activities are supported in a variety of ways by computer-aided management tools. These have been reviewed in [633] and [634].

The Microsoft project management system MS Project [635] appears to be finding acceptance in multiproject management situations.

- Economic aspects
  - Market interest
  - Contemplated production level
  - Nature of investment project
    - * New plant
    - * Expansion
  - Intended site
  - Intended capital investment
  - Economics or profitability
- Chemical principles
  - Reaction equations
  - Mole ratios
  - Reaction stages
  - Enthalpies of reaction
- Process principles
  - Structure
  - Product and heat streams
  - Process conditions
  - Mode of operation
    - * Batch/continuous
    - * Single-component/multicomponent
- Substance properties for feeds, intermediates, byproducts, and end products
  - Physical data
  - Toxicological data
  - Ecological data
  - Engineering data
  - Safety characteristics
- Constraints
  - Legislation and internal requirements
    - * Environmental protection
    - * Plant safety
    - * Occupational safety
  - Patents
  - Local conditions
- Description of plant
  - Structure
  - State of preservation
  - Utilization
  - Availability
  - Operation and monitoring
  - Engineering functions
  - Safety requirements
  - Infrastructure

Figure 323. Possible topics covered by feasibility study

10.3. Decision Phase

The fine structure of the decision phase can be seen in Figure 309. The purpose of the feasibility survey is to elucidate and develop the statement of tasks in close collaboration with the customer. Figure 323 shows possible points covered by a feasibility study for a chemical plant construction project.

The economic aspects are just as important as the chemical and process principles. These are supplemented by substance data and constraints. A detailed description of the plant completes the picture of the project. An additional analysis of weak points will reveal any flaws and thus contribute to making the task statement more concrete and precise (see Chapters 2, 3, 4).

This stage results in the performance specifications, which set forth the solution-invariant description of the problem to be solved (requirements). Its structure and content, in particular as it relates to plant safety, is discussed in more detail in Section 10.4.

In the preliminary design, a number of alternative solutions can be offered to the stated problem; from these, one is to be chosen as optimal for the company from both economic and technical standpoints (see Section 10.1), whereby a cost/benefit analysis may be useful.

Any optimization potential in the selected solution can then be revealed by sensitivity analysis, which investigates the influence of important, highly sensitive variables on the process by means of, for example:

- Mathematical models
- Application of empirical and theoretical knowledge
- Plausibility analysis

The optimization potential thus found can be realized in the design study. The result is a binding concept for both process and plant.

The binding concept is described in the specifications, whose structure and content are discussed further in Section 10.4, with particular regard to plant safety and economics.

The decision phase ends when the documents required for obtaining regulatory approval have been prepared and costing has been done (Section 10.2).

Errors in project planning and execution can arise from a wide variety of causes. The more serious are these problems and the later they are discoverd, they higher is the cost of correcting them.

10.4. Specifications

Bidding Specifications. The specifications describe the task in structured form, mainly from the user's viewpoint, and state what effort must be brought to bear [625, 636].

The bidding specifications can be regarded as the basis for a contract between the party designing and constructing the plant and the later plant operator. They are the basis of all further planning and design.

Structure and Content. The bidding specifications are broken down in parallel with the topics in the feasibility study (Fig. 323). They are broken down into a general section relating to all components and a specific section for each component (Fig. 324).

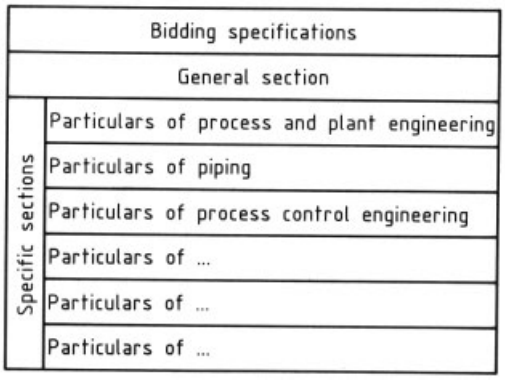

Figure 324. Breakdown of bidding specifications

The general section includes important information applicable to all components. Possible topics in the general section of the bidding specifications are:

1) General information on the project
2) Current position and shortcomings
3) Objective
4) Structure of the process and plant
5) Safety and availability requirements
6) Miscellaneous conditions to be considered
7) Framework concept

A vital part of the general section of the bidding specifications is the "framework concept," which describes the real environment of the plant as well as expected expansions. This section, in particular, describes interfaces to other plants (not just for process measurement and control).

The component sections of the specifications discuss these points only briefly, or may simply refer to the general section. Instead, these sections state the points of importance for the execution of the particular tasks. Possible contents of the section on process measurement and control are:

1) Typical requirements for process control station functions
2) Special control and monitoring functions
3) Process control functions
4) Human/process communications
5) Off-line interpretation functions
6) Infrastructure; room and building equipment
7) Communications hardware
8) Process analysis
9) Safety requirements
10) Availability requirements
11) Quality assurance
12) Miscellaneous

An important item in the bidding specifications is the description of functions used to control the process. The particulars of process measurement and control are described (Fig. 325, column 3) [637].

"Field-level" functions for the performance of control and safety tasks are distinguished from "higher-level" functions involved in the management of the process and the plant. The later include, for example, process operation, process logistics, and recipe-mode operation (Chaps. 2 and 4).

These functions in turn are interfaced to other higher-level functions involved in the management of a works/plant complex or corporate management (Fig. 325, columns 1 and 2) [637].

The breakdown of functions is based on the level model of production (Fig. 326) [638] (see also Chap. 2).

The totality of functions to be performed and their interfaces to one another defines the type and size of the process control systems. For example, Figure 327 shows a system that handles the functions of all levels of the production model (see also Section 5.2).

The holistic view and conceptual procedure require prompt acquisition of information about necessary and desired functions for the later operation of the plant. Particulars of various process management tasks are treated in depth in Section 4.3.

Corporate management	Plant management	Process management
■ Strategic planning ■ Logistical systems Contract performance Maintenance Operations planning Procurement ■ Information systems ● Sales and marketing Sales Partners Sales planning ● Production Production planning (overall) Capacity planning Investment planning ■ Quality assurance ■ Internal cost accounting ■ Personnel management	■ Administration and production logistics Production planning (detailed) Plant section operations planning Personnel planning Production flow monitoring Batch approval Schedule monitoring ■ Economic tasks Cost functions Cost analyses Statistics ■ Engineering tasks Recipe handling Assessment of process and product properties Machine changeovers Process engineering procedures Quality assurance	■ Higher-level functions: Management of processes and plants ● Recipe-mode operation ● Process operation Monitoring Intervention Documentation ● Process logistics Startup and shutdown Load and product changing Buffer strategies ● Disturbance handling ■ Field-level functions: Measurement, control, safety ● Determination of process properties ● Determination of product properties ● Setpoint (closed-loop) control ● Operational chart (open-loop) control ● Reporting of deviations ● Shutdown ● Operation and monitoring

Figure 325. Functions broken down by level of the production model [637]

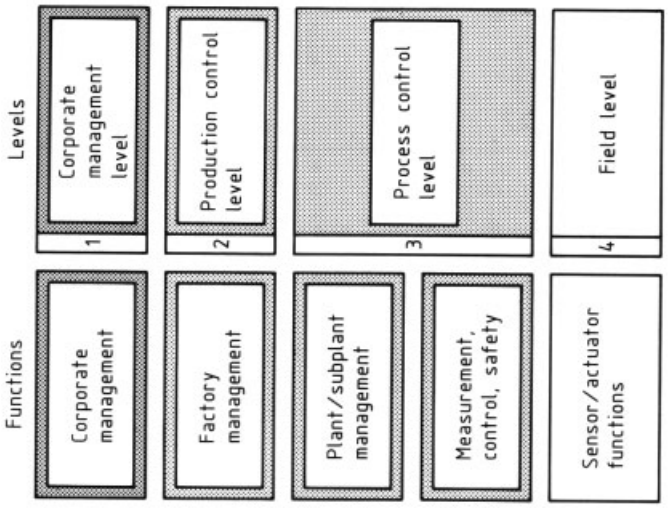

Figure 326. Level model of production [637]

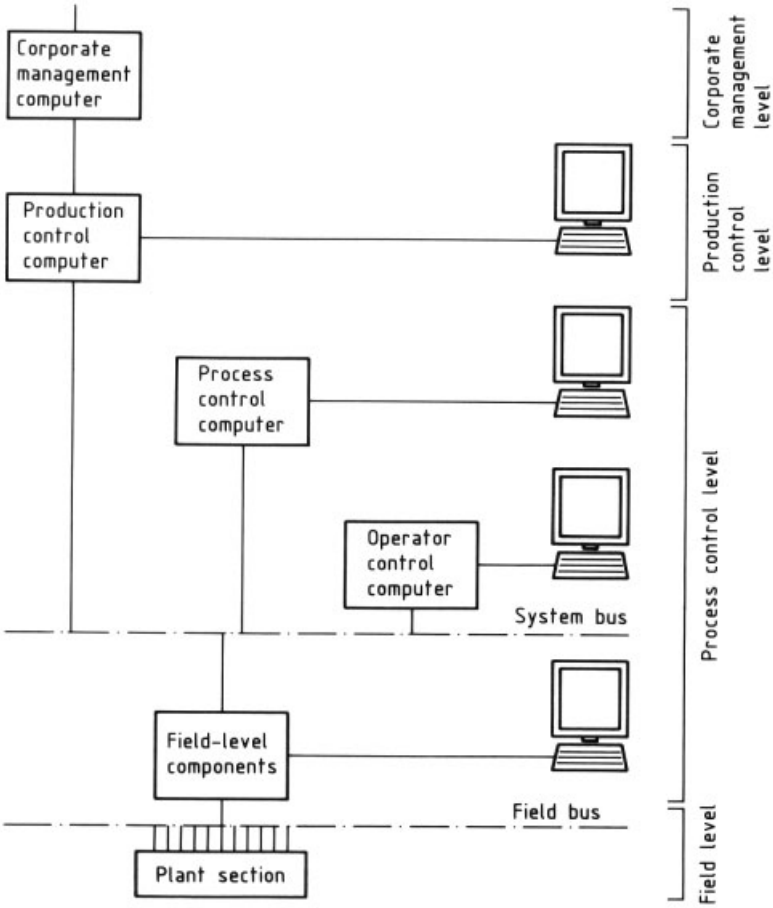

Figure 327. Schematic diagram of a control system [637]

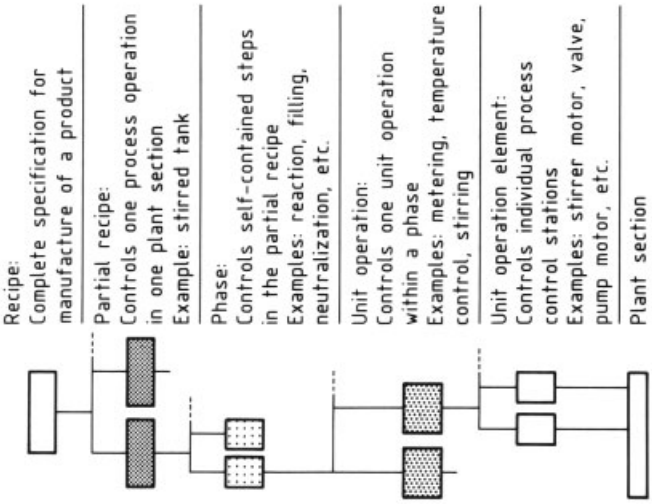

Figure 328. Example of recipe control [639]

Process/plant specifications	Process control specifications
o Process - Structure - Material and thermal balances - Unit operations - Product and process properties o Plant - Structure - Plant sections - Safety concept o Infrastructure o Miscellaneous - Construction concept - Provisional site designation	o Field-level and associated devices - Type of station and interconnection - Types of devices - Safety and availability measures o Systems hardware and software - Level of automation - System structure - Safety and availability measures - Operation and monitoring - Software specifications o Infrastructure and building equipment - Process control rooms - Power supply - Safety and availability measures
Rough cost estimation	

Figure 329. Structure and content of specifications

The rules stated in Chapter 2 apply in a special way to the description of recipe-mode operation [636, 639] (see Fig. 328), [640, 641].

As proposed earlier [636], the bidding specifications are now prepared in checklist form [642].

Process engineering functions may also include room and building functions (heating, climate control, ventilation, lighting, etc.) as well as communications (e.g., telephone).

Specifications. The specifications give a detailed description of the nominal condition, that is, how and with the requirements of the bidding specifications are to be realized. The specifications set forth the solution-variant realization. These documents are binding on the contractor (in-house planning and design department or outside firm). The specifications form the basis for the contract, especially when the planning and design work is contracted out [625].

Structure and Content. The specifications are organized similarly to the bidding specifications and consist of a general section and component-specific sections.

The process-engineering aspects governed by the specifications are illustrated in Figure 329. For the sake of comparison, provisions for the process and for the plant are listed in parallel [631, 643, 644].

Provisions on process and plant are first briefly explained (left column of Fig. 329).

The structuring of the process and of the plant is the same as set forth in the bidding specifications (see also "Phase Model of Production" in Chapter 2, as well as Figs. 330, 331, 332). The breakdown is taken over into the specifications unless changes have been made; otherwise, it is modified to accord with the new requirements.

The breakdown of the process is independent of whether it is already known or new within the company.

Even known processes must be restructured since a new investment always entails adaptation to the state of the art, and adaptation can strongly affect the process structure.

According to [645], the process can be broken down into process stages and unit operations. Some concepts in [645] must be modified, because there is no guarantee of completeness, uniqueness, or freedom from conflicts. "Process stages" are generally self-contained portions of a process.

Unit operations are the smallest steps in a process and correspond to process elements. Any unit operation converts inlet products with par-

ticular properties to outlet products with different product properties. As was discussed in Chapter 2, product properties are those qualities that characterize the product, such as physical and chemical values, engineering properties, and product indicators [236, 646].

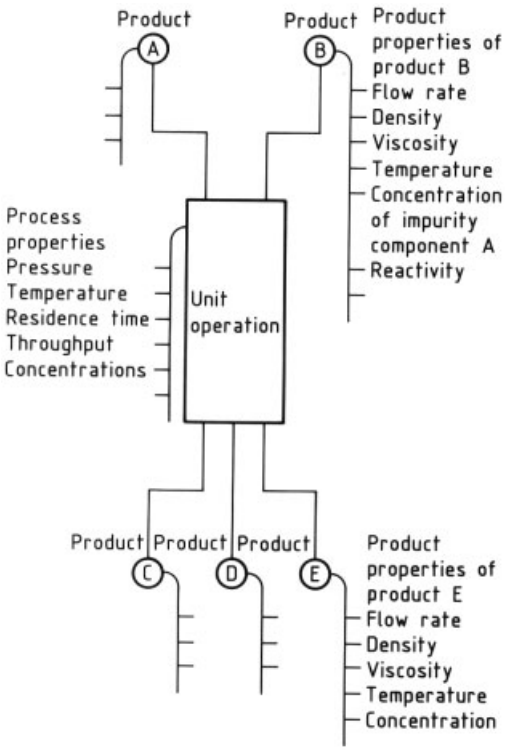

Figure 330. Extract from a phase model

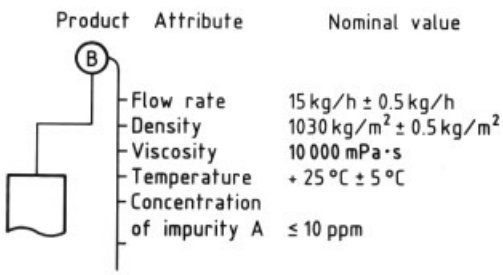

Figure 331. Example of a list of product properties

A unit operation can be specified as shown in, for example, Figure 24. It can be specified, for example, that a separation is to be carried out by distillation. Such a statement of the method of performance also establishes the basic form of the trajectories that describe the transformation of inlet products with their properties into outlet products with theirs. The properties that characterize this transformation process are the process properties.

Figure 330 shows a unit operation with the inlet and outlet products, their product properties, and the process properties. The product and process properties must be set forth in the specifications for the process. As Figure 331 shows, the acceptable ranges of variation (tolerances) should also be given if possible. In this way, a profile of requirements on the overall process is obtained; this should cover all operating modes (startup, shutdown, load change, etc.; see Fig. 332). This description, coupled with actual values, is the basis for a quality information system (QIS).

The profile of requirements is the essential condition for creating a reproducible operation [625]. The overall sensor and actuator system is later defined and quantified on the basis of this profile. Further more, possible ciritcal states of product properties, which might affect the availability of the sensor – actuator system, can be identified in this way (plugging due to solid/liquid phase separation, etc.).

Material and thermal balances indicate ways of insuring the supply of substances and energy to the process and the removal of substances and energy from it. They also give the consumption rates per unit amount of product, which are required for the profitability calculation at the end of the decision phase (Section 10.4).

Once the process structure has been completed, the plant structure is broken down by assigning unit operations to appropriate plant sections. A plant section is a functional unit of a plant that can be operated independently (corresponding to a unit operation). Examples of plant sections in chemical process plants are absorbers, chemical reactors, crystallizers, tank storage facilities, and so forth.

The totality of all plant sections required for the performance of a process is the plant.

Technical equipment in a plant section includes both devices involved in the process, such as vessels, machinery, apparatus, piping, and fittings, and process measurement and control units [647]. The detailed description of a plant section is based on these elements (see Fig. 23).

Safety and economic considerations result in a variety of requirements being imposed on the

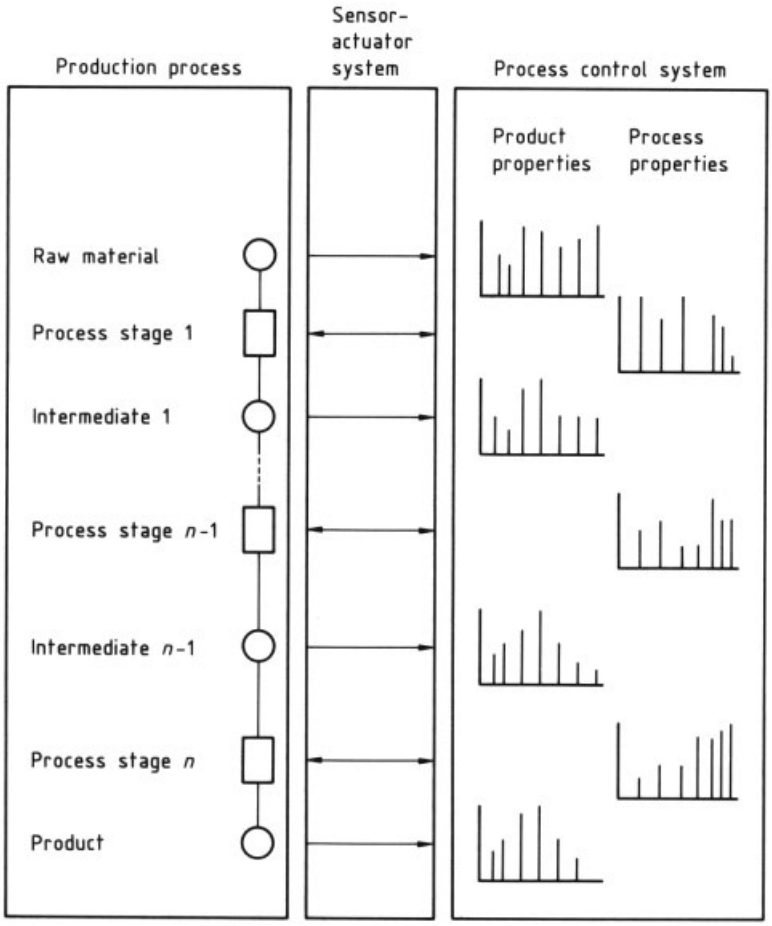

Figure 332. Property profile of an overall process [646]

objects of a plant, such as plant sections and technical equipment. Such requirements must be taken into account in planning and design, construction, and operation.

In the lifetime of a plant, a large quantity of data is generated, used, interchanged, and if necessary modified. If such data and other information is to be exchanged without losses, each item of information must be uniquely associated with the plant object to which it relates. This can only be achieved if the objects are uniquely labeled [648].

Plants and apparatus can be identified in various ways. For example, each can be assigned an unstructured designation.

In a complex system such as a chemical plant, it is better to structure the system in a hierarchical fashion, and then have the system of identification codes reflect this structure. In this way, an informative and easy-to-use identification of plant objects is obtained [649, 650].

An engineering object can be "named" in many ways; a few examples are: by purpose, by technical function, by location, or by device type.

If all such features were included at once in the object identification, an unwieldy code would result. For this reason, just a single attribute is employed. It is desirable to choose a feature that is completely and permanently identifies the object (e.g., the technical function [203]). This can be illustrated for a temperature measurement and control station.

Suppose the purpose of the station is to maintain a constant temperature in a certain unit operation, say mixing. The technical function is to

measure, record, indicate, and control (TIRC) at a certain point in the plant, a Pt 100 instrument being used as the measuring sensor.

If this sensor is now replaced by a thermocouple or if the plant is changed over to a different process with different unit operations (e.g., reaction instead of mixing), the station still retains its technical function.

The coding of technical objects is a planning/design task. It is carried out as early as possible, but at the latest when the specifications are prepared. This is done in entity – relationship (E/R) models, which are generated in a consistent way throughout the company.

On the basis of the safety provisions in the bidding specifications, the safety concept is derived from the operating conditions together with the potential hazard presented to persons and the environment by substances involved in the process.

A construction concept and a provisional identification of the site round off the process and plant concept.

The overall concept is documented in phase models plus the overall and process flowsheets discussed in Chapter 2, which contain fundamental and supplemental information.

The process measurement and control concept follows from the process and plant structure, the profile of requirements on the process, and the task statement contained in the bidding specifications.

The specifications set forth the types of process measurement and control stations and, roughly, the way in which they are interconnected.

The process and plant control functions performed by the control system determine the extent to which the plant is automated (see Section 4.3).

If functions on the plant control level are also to be implemented, a production control computer must be provided. The interfaces between functions at different levels must be described and specified as exactly as possible (see Chap. 2).

The number of field-level components and their assignment to plant sections is based on the process structure, the required functionalities, and the safety and availability requirements.

The operation and monitoring concept (see Section 11.2) essentially comprises the specification of operator displays, reports, and logs. The operator station concepts must also be worked out. These define the scope of tasks to be performed at the various control stations and the association of control stations with plants and plant sections [651].

The information structure should be hierarchical, with high-level diagrams containing few items of information, and station-level diagrams, extensive information.

Reports should be specified and classified as follows:

- Alarms
- Malfunctions (in the measurement and control system)
- Warnings (advance notice of alarms)
- Errors (in the peripheral hardware of the measurement and control system)

Safety and availability requirements govern how priorities are to be assigned to the messages and how they are to be presented in the control system (see also "Availability Aspects" below).

Logs and records are subdivided into, for example, batch records, shift logs, daily logs, and alarm logs.

If, for example, the bidding specifications call for recipe-mode operation, the software specifications (Fig. 329) should include the appropriate large-scale function charts, specify the selectable parameters of the recipes, and state how the recipes will be maintained and archived.

At this point, the task to be performed has been explained enough that a concrete solution proposal can be submitted with the aid of a process control system. This proposal must, however, take account of the measurement and control provisions described in the framework concept of the operation in which the process will be realized. For example, if a process control system from a certain manufacturer is already in use, then for maintenance reasons it is recommended that the planner stay with this manufacturer unless the adaptation to the task carries unacceptable risks with it.

Availability Aspects. *Basic Classification of Process Measurement and Control Equipment.* Availability requirements must be examined from several viewpoints:

- Safety
- Process engineering

– Logistics

Safety is mandated by legislation and regulations. The safety of persons in the plant (worker protection) and of the surroundings (environmental protection) must be guaranteed.

The safety of engineering processes is a key criterion. A high level of safety results from continual and intensive work on safety aspects. Once a high degree of safety has been achieved, it must be maintained and improved. The special handling of safety aspects in the bidding specifications and the specifications is meant to stress this requirement.

The term "process" is understood to mean the totality of interacting operations in a system by which matter, energy, or information is transformed, transported, or stored. In the case of technical processes, the physical quantities are determined and controlled by technical means [652].

The safety of a process is determined by the processes, plants, and process control equipment (Fig. 333). An essential contribution to process safety is due to process control equipment [653]. Such equipment can be classified as shown in Figure 333.

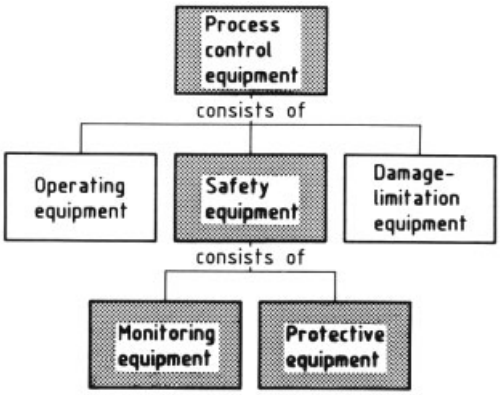

Figure 333. Summary of measurement and control equipment performing process safety functions

In assessing process safety, the following three types of operational conditions are distinguished [653]:

– Good range
– Acceptable range
– Unacceptable range

The good range is the intended range for the values of a process parameter. It is defined such that the plant does not sustain any damage beyond ordinary wear and tear, while the quality and quantity of goods produced satisfy the requirements imposed.

The acceptable range is adjacent to the good range. In it, the plant still sustains no damage beyond ordinary wear and tear; the quality and quantity of goods produced still satisfy the requirements to within extended tolerances.

In the unacceptable range, the value of a process parameter is such that the plant may sustain or cause greater damage, or the quality and quantity of goods produced no longer satisfy the requirements stated.

These parameter ranges can now be correlated with process measurement and control equipment. Devices used for measurement, control, reporting, and so forth serve to keep the plants operating as intended (in the good range). They continually intervene in the process. Only in the acceptable range do the process control safety devices first come into operation. On entry into this range, the monitoring devices respond, prompting the operator to take suitable action to bring the process back into the good range.

If these actions are not taken as expected, protective features in the process measurement and control system, such as safety valves, rupture disks, and the like, prevent entry into the unacceptable range.

Processes can further be provided with damage-limitation features such as spray-blocking walls, warning systems, and special traffic control systems that block off the site of the danger.

Figure 334 summarizes the correlation between the ranges of process parameters and the process measurement and control devices, showing the intervention levels and modes of action [653].

The safety requirements mandated by legislation for the overall process in a plant to be erected, and in particular for the process control equipment, must be precisely described in the bidding specifications, since they influence the process design as well as the design of the process measurement and control system.

The required availability from the process engineering standpoint affects both product quality and the protection of the facilities. The logistics-

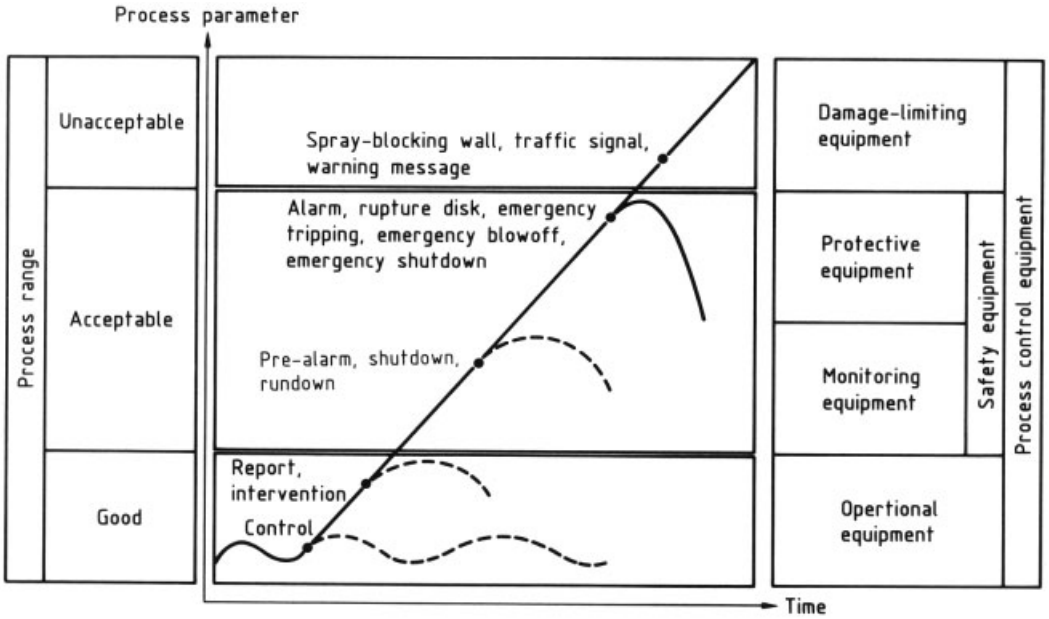

Figure 334. Process control devices and process safety [653]

based availability bears on the ability to deliver products and has direct effects on manufacturing costs. The availability requirements imposed by logistics and process engineering must be included in the economic picture, in contrast to the safety requirements.

The measures taken to comply with these requirements are set forth in the specifications.

Implementation. The safety and availability requirements are key factors affecting the structure and organization of field-level and system-level components. These lead to, among others, the need for the process sections and/or plant sections to be largely autonomous relative to one another and relative to malfunctions in higher-level operating and/or control computers.

To insure the protection of workers and the environment as mandated by law, a number of other technical measures set forth in codes and regulations must also be undertaken. A few of these are listed in Figure 345.

Allowance must always be made for defects that limit the proper functioning of technical equipment or even take it out of action. In the case of safety equipment, care must therefore be taken that no impairment or failure of the safety function occurs even when there are defects in the technical equipment performing such safety function.

This goal can be achieved through two distinct approaches to safety equipment design, called "fail-safe" and "redundant." In fail-safe design, all malfunctions deemed possible cause the safety devices to fail in a single well-defined sense, namely so as to trigger the safety or protective function. Redundant design means including more than one component for a certain task. If one component fails, a backup component takes over its function.

Redundancy in process control systems means that system components, such as field-level, operating and monitoring units, and/or bus systems, are present two or more times. The types of redundancy practiced in technical systems are as follows [654]:

Static redundancy. Instead of a single component, several components are installed in parallel. The system as a whole performs in the intended way as long as at least one unit is free of defects.

Dynamic backup redundancy. In case of malfunction, an additional component is switched in to replace the defective component. The additional component is not in operation when no malfunction is present.

Local dynamic backup redundnacy. Assigned to every component is a redundant compo-

nent, which is continuously reconfigured (hot standby).

Global dynamic backup redundancy. A redundant component is assigned to several components. In case of a malfunction, the redundant component is reconfigured once (warm standby).

Dynamic functional redundancy. When no malfunction is present, the redundant component performs its own tasks; if a malfunction occurs, it takes over the functions of the defective unit by restricting the capacity or shutting down inessential functions.

Backup. The general term "backup" refers to dynamic backup redundancy for a single process measurement and control function, such as control.

There are a variety of methods for acquiring redundant information about the process. Figure 335 presents some typical examples.

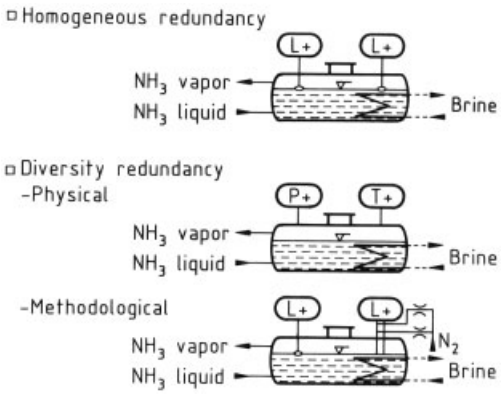

Figure 335. Redundancy in measurement [653]

The term homogeneous redundancy is used when the same physical quantities and the same measurement techniques are empolyed for data acquisition. Physical diversity uses distinct physical quantities. Methodological diversity means using the same physical quantity for data acquisition, but measuring it by different methods.

In process measurement and control systems, functions are implemented through hardware and software. Process control systems should be used to realize safety devices only when they can be validated, that is, when it can be demonstrated that the software involved in the safety function always performs properly and reliably under all circumstances. If this cannot be demonstrated, safety functions must be implemented by hardwired control units.

Redundancy implies a high-availability concept. Despite, or indeed precisely because of, the presence of multiple process components, such high availability is not always realized. Problems arising from the number, interconnection, and synchronization of components in a redundant system may work against it.

Safety concepts and availability requirements must also affect the security of the process power supply system. The number and type of incoming supply lines, multibus configurations, switching devices, and standby power systems depend on it [655]. Section 6.4 surveys the most important options.

The specifications contain additional provisions dealing with the infrastructure (Fig. 329), including process control areas, such as control and switching rooms, as well as overall building equipment.

The specifications end by giving cost figures and the profitability calculations based on them. These are crucial for the decision to continue with the investment project. If the economic criteria established by corporate management are satisfied, the project is continued; otherwise, either its scope is modified or its is terminated.

Economic Aspects. The purpose of any investment project is to generate profits. The specifications must therefore demonstrate that the proposed actions to carry out the tasks described in the bidding specifications will contribute to profits. The question of economics is vital, especially with regard to the desired level of automation (Section 4.3)

[625]. A separate economic calculation can be done for each automation level.

– Automation of field-level functions of the process control level (Fig. 325)

This level is mainly concerned with the functions of measurement, control, operation, and reporting. These can be implemented by conventional method or by process control systems. An analysis of several projects has shown that there is, in principle, no difference in financial outlay between these two options. Both investment costs and the breakdown of costs (Fig. 336)

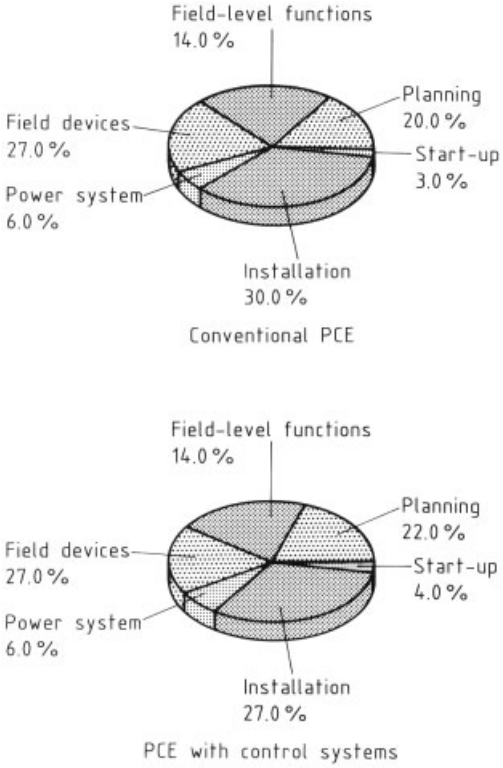

Figure 336. Comparison of cost breakdowns for the implementation of field-level functions of the process control levels with conventional process control technology and with process control systems [637]

are roughly comparable. On the basis of this knowledge, control tasks of the type discussed here should always be implemented with process control systems, since these have potential for expansion that is generally lacking with conventional methods.

– Automation of field-level and higher-level functions of the process control level

As Figure 325 shows, process control systems make it possible to carry out recipe-mode operation, tasks having to do with process logistics, and higher-level control functions. The automation of packaging lines or storage facilities can be included under this heading.

Performing these functions generally entails additional investment costs, which are, however, justified only if the annual profit gained through their use exceeds a certain value specified in advance by corporate management. The profit is calculated as the savings achieved through the investment minus fixed costs.

The profit expressed as a percentage of the total investment is called the return on investment (ROI).

$$\mathrm{ROI}\,[\%/\mathrm{a}] = \frac{\mathrm{savings}\,[\mathrm{DM/a}] - [\mathrm{fixedcosts}\,[\mathrm{DM/a}]]}{\mathrm{totalinvestment}\,[\mathrm{DM}]} \times 100$$

The reciprocal 100/ROI is the payback time, i.e., after how many years the total investment is paid back in profits:

Either figure can be used as a rough measure of the profitability of an action.

Table 12 shows a cost analysis done in this way. The benefits can be broken down in a variety of ways, but essentially into five categories,

Table 12. Cost analysis for the implementation of higher-level functions at the process control level using process measurement and control systems [656]

Task	No. of projects	Savings ∗, 10^3 €/a							Total investment, 10^3 €	Fixed costs, 10^3 €/a	ROI ∗∗, %/a
		Improved product quality	Reduced inspection	Energy savings	Personnel savings	Improved yield	Enhanced plant utilization	Sum			
Recipe-mode operation	2	–	+	–	232	–	197	429	665	146	40
		38	+	–	332	255	105	695	971	192	50
Process logistics	1	+	+	+	204	613	+	818	1278	304	40
Higher-level control technology	1	+	–	82	112	–	–	194	276	74	40
Automated packing/storage	2	+	–	–	+	–	+	245	232	59	80
								2086	1227	297	140

∗ + Beneficial but not quantifiable; – No benefits identifiable.
∗∗ Return on investment.

although not all of these can be quantified. If only the quantifiable items are included, satisfactorily high ROI values or satisfactorily short payback times (all < 2.5 years) are obtained.

These results favor the use of process measurement and control systems for the implementation of higher-level functions of the process control level.

– Automation of functions of the production control levels (Figs. 325 and 326)

The economic analysis is problematic because the simple investment return calculation methods described above are no longer adequate. The associated benefits of better information (i.e., comprehensive, up-to-the-minute, error-free, and perhaps condensed) being available as the basis for analyzing both engineering and management weak points and also for the real-time control of the operation cannot be quantified.

Ultimately, corporate objectives will dictate the implementation of a production control level. Failure to do so, however, may result in long-term competitive disadvantages.

10.5. Execution Phase

The fine structure of the execution phase is indicated in Figure 309

Basic Engineering. In the basic engineering phase, the individual tasks are formulated in their final form, in collaboration with the customer, and the functional solution is described. The work to be done here is listed below:

Planning and design for process and plant engineering:

– Design of plant sections (apparatus) in accordance with the process
– Decision on materials of contruction
– Preparation of site plans and provisional layouts
– Preparation of piping and instrumentation diagrams

Planning and design for process control engineering:

– Design of process control stations
  Establishment of structure
  Establishment of devices to be used
– Design of process measurement and control system
  Dimensioning of process interface
  Design of data processing software
  Dimensioning of the control system
  Dimensioning of control and other rooms
  Spatial disposition in the plant
– Preparation of piping and instrumentation diagrams

The process and plant engineering work includes plant design in accordance with the process, sizing of the plant sections (apparatus), and elucidation of materials questions. With regard to documentation, see in particular Section 3.5.

Preparation of Piping and Instrumentation Diagrams. A key task is the preparation of P & I (piping and instrumentation) flowsheets [645] (see Fig. 337). Provisional layouts and site plans are also drawn up.

The process control engineering organization is involved in the preparation of P & I diagrams, since its planning and design work will be based on these. Process control tasks are described in the station symbols (Fig. 338) of the P & I diagrams. This description should be sufficiently exact enough that it, together with the binding concept, makes possible the design of the PCE stations.

The process control station includes all measurement and control elements needed to perform the functions specified, such as sensors and/or actuators, signal transducers, power supplies, and processing functions, for example indicating (I), recording (R), controlling (C), etc. [657].

Process Control Station Data Sheet. The station design is set forth in a "data sheet," which must list all elements making up the station and also provide information on the substances with which the sensors and actuators come in contact and on the installation location (Fig. 339).

Given the measurement and control ranges of the sensors and actuators, suitable equipment is identified for the individual elements on the basis of instrument standards (see Chaps. 7 and 8 and Section 10.6).

Process Control Station Diagram. Station diagrams document how the equipment is wired together (Fig. 340). At this point, the planning and design work on stations utilizing conven-

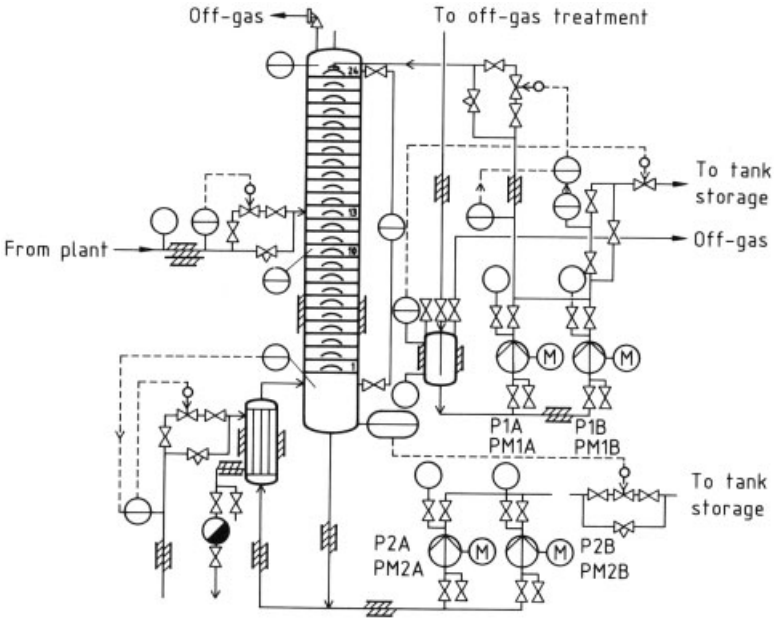

Figure 337. P & I flowsheet of a plant

tional technology is finished. If data processing functions are implemented in process measurement and control systems, additional design work is required for the hardware structure of the process interface as well as the software structure incorporated into the control system proper.

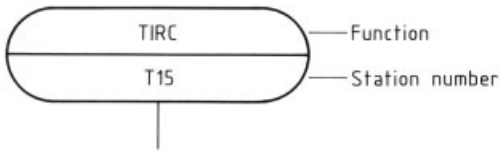

Figure 338. Representation of a process control station in the P & I flowsheet [647]

Analog and binary signals coming into the process control system must be transformed into information that this system can understand. The conversion takes place in the process interface and is performed by input/output components for the various types of signal. Backup controllers are also implemented with components. These components are shown in the station data sheet as well as the station hardware diagram (Fig. 341; see also Sections 7.2 and 10.2).

Station Software Diagram. The station hardware diagrams are supplemented by station software diagrams (Fig. 342; see also Section 9.6), which show how standard software modules, provided by the process control system manufacturer, are linked so as to build up the PCE station processing functions.

Function Chart. The data processing functions of a station must be performed in a coordinated manner; this requires a coordination function at a higher level in the hierarchy. The essence of this function is a task-specific sequence control that must be specially prepared for each field-level component of a measurement and control system. Sequence controls are formulated in a system-neutral way by using function charts (Fig. 343) [658]; the same holds for station-level interlocks.

The function chart is a process-oriented depiction of the control function, independent of the way in which it is implemented. It replaces or supplements the verbal description of the function. The function chart clearly and unambiguously describes the function and thus aids in understanding between manufacturer and user. Furthermore, it facilitates collaboration between specialists [659]. In future projects, Petri networks will be employed at an early stage, from which function charts can be automatically generated (see Chap. 2 and Section 4.3).

	Station data sheet		Prepared: 20.10.1989

Header

Facility		Process control station	*F1.2*
Plant complex	*AB-Reak.* Plant Plant section	Type of station	*FIC*
Building	*A123*	Interconnected with	*T1.1*
Project number	*1234567*	P&I diagram No.	*831227*

Description of substance

Station identification	*Steam inlet R01*		
Station is labeled	*Steam*		
Circuit type			
Installation standard			
Substance measured	R	Substance controlled	R
Designation of substance	*Steam 5 bar*	Name of substance	*Steam 5 bar*
Composition		Composition	
State of aggregation	*Vapor*	State of aggregation	*Vapor*
Corrosive constituents	*None*	Corrosive constituents	*None*
Suspended solids		Suspended solids	kg/m³
Density in operating state	kg/m³	Density in operating state	kg/m³
Density at standard conditions	kg/m³	Density at standard conditions	
Dynamic viscosity	mPa s	Dynamic viscosity	mPa s
Electrical conductivity	µS/cm		

	min.	norm.	max.	units		min.	norm.	max.	units
Pressure		*5*	*5.5*	*bar*	Pressure		*5*	*5.5*	*bar*
Temperature		*180*	*220*	*°C*	Temperature		*180*	*220*	*°C*
Flow rate		*500*			Flow rate				
Limits									

Substance measured/minus side			
Density/minus side			
Behavior in sampling line			

Location

Measuring location	R	Controll location/device	R
Installation location	*Steam R01*	Installation location	*Steam R01*
Piping/apparatus Nom. diam. DN *50*	Nom. press. PN *16*	Piping Nom. diam. DN *50*	Nom. press. PN *16*
Material	*GS-C25*	Material	*GS-C25*
Heating		Insulation thickness	mm
Insulation thickness	mm	Safety code	
Measuring span (M)	mm		
Floor clearance (HO)	mm		
Overall length (vessel) (HE)	mm	Control device, type	*241-1*
Flange, top/bottom DN/	DN	Control device Nom. diam. DN *50*	Nom. press. PN *16*
Ambient temp. from to	°C	Overall length/height *230* mm	*305* mm
Overall length (piping)	mm	Mode of operation	*Closed by spring*
Pressure drop	bar	Seat/critical volume flow rate	*31* mm /16
Safety code		Pressure drop	bar

List of station elements

No.	Code	Text	Input	Output	Location	Remarks	R
1	Own	Standard orifice			plant B1		
2	03.42.21	Transducer		4-20 mA	plant B1		
3	08.07.01	Diaphragm valve	0-6 bar		plant V5		
4	08.02.01	Solenoid valve	DC 24V	0-15 bar	plant Y1		
5	04.21.01	Position controller	4-20 mA	1.4-6 bar	plant N1		
6	14.22.05	Power supply	4-20 mA	4-20 mA	K1 E2 A2		
7	03.08.03	Signal transducer	4-20 mA	4-20 mA	K1 E3 A1	MUS/TVA	
8	3.8.1918	Signal transducer	2×0-20 mA	8×0-20 mA	K1 E4 A1	GUE	
9	Own	Control K1.1			K2 E3	IV 02	
10	45.45.45	Voltage distribution	DC 24V	DC 24V	K2 E4	SIV 60, F2	

Figure 339. Example of a station data sheet

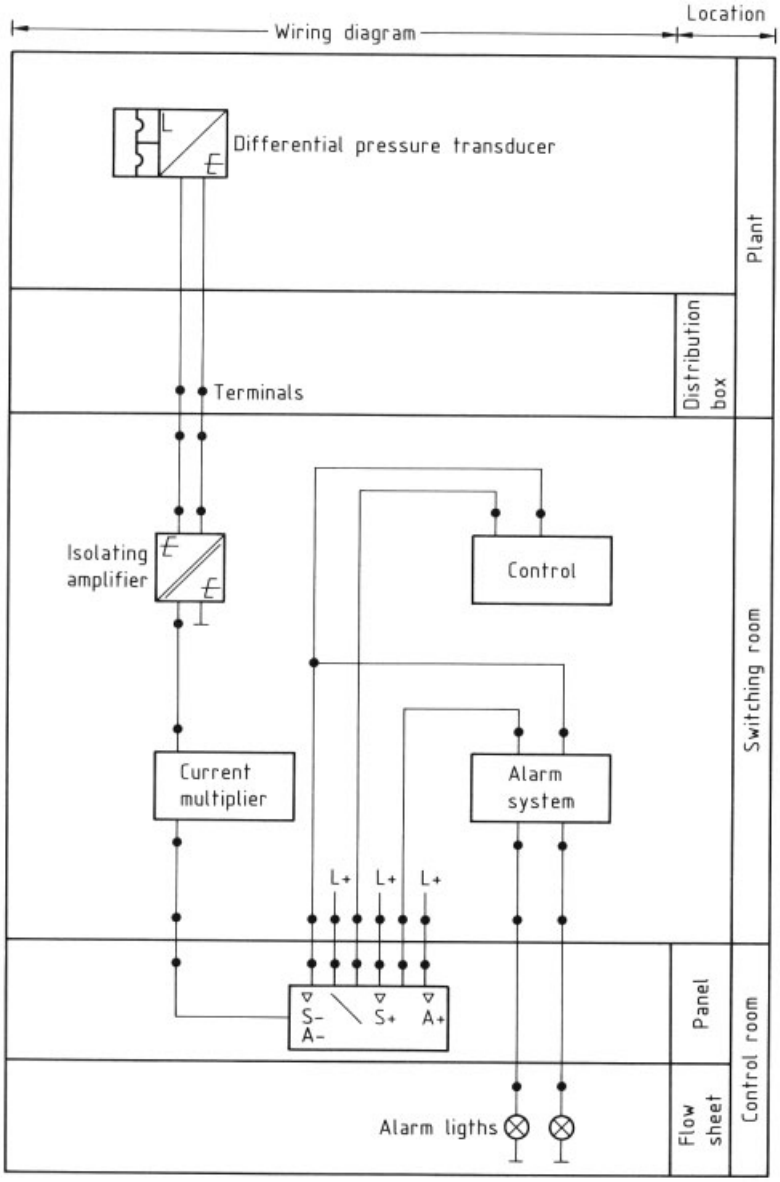

Figure 340. Example of a station plan for a conventional process control station

Dimensioning of Process Measurement and Control Components. Generally, a process measurement and control system consists of a number of field-level components. Like the station-level functions of field-level components, these must also be coordinated, but at a higher level, for example, in the context of recipe-mode operation. A process control computer is used for this coordination function.

The size and type of control computer will depend on the scope of the coordination function. The selection of the operating computer is based on the number of graphical operating displays (Section 11.2).

The number and size of the field-level components assigned to a process section is influenced by the number of stations to be implemented, the functional scope, and the requirements on availability and safety (Section 10.4).

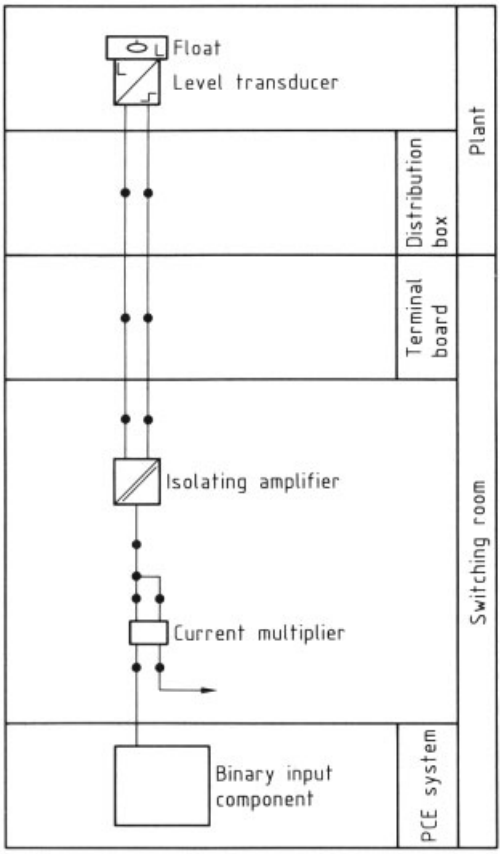

Figure 341. Example of a hardware station diagram for a station with signal processing (LS+)

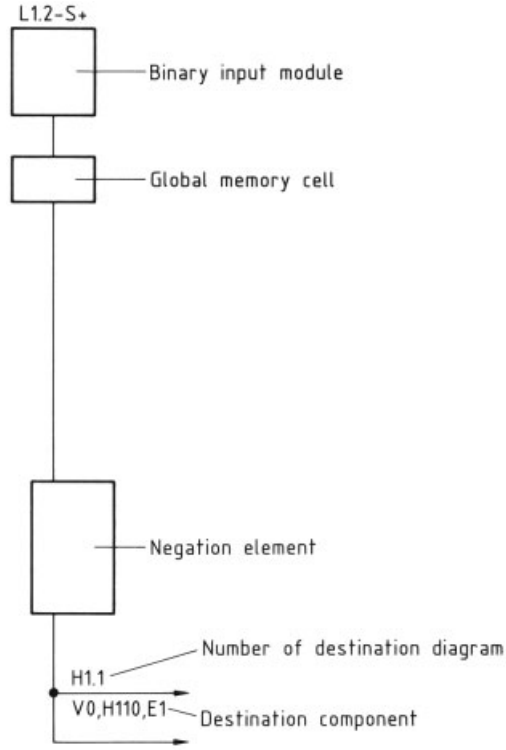

Figure 342. Example of a station software diagram for a station with signal processing (LS+) in a process measurement and control system

The station-level functions are built up from software modules. These differ in memory requirement and machine time, so that flow-through and pressure-control loops assembled from such modules may require 3 – 4 times as much memory and 4 – 6 times as much processing time as pure measurement loops.

If a large number of time-comsuming functions must be implemented in a single section of the process, field-level components whose time properties allow this must be used. Today, memory problems are of little importance.

The tasks of basic engineering also include dimensioning the rooms for the installation of process measurement and control equipment (see Section 10.7).

Detail Engineering. In the detail engineering phase, detailed and comprehensive documents are prepared for use in procurement, construction, operation, and maintenance of a plant. Consistent documentation is a vital consideration at this stage (see Section 3.5). The planning and design work can be broken down between process and plant engineering and process control as follows:

Planning and design for process and plant engineering
- Detailed layout of plant sections, apparatus
- Construction drawings
- Site plans, layouts
- Piping design

Planning and design for process control engineering
- Preparation of data sheets
- Completion of station diagrams and function charts
- Programming
- Working drawings of mimic diagrams for operation and monitoring
- Preparation of circuit and configuration diagrams
- Preparation of commissioning checklists

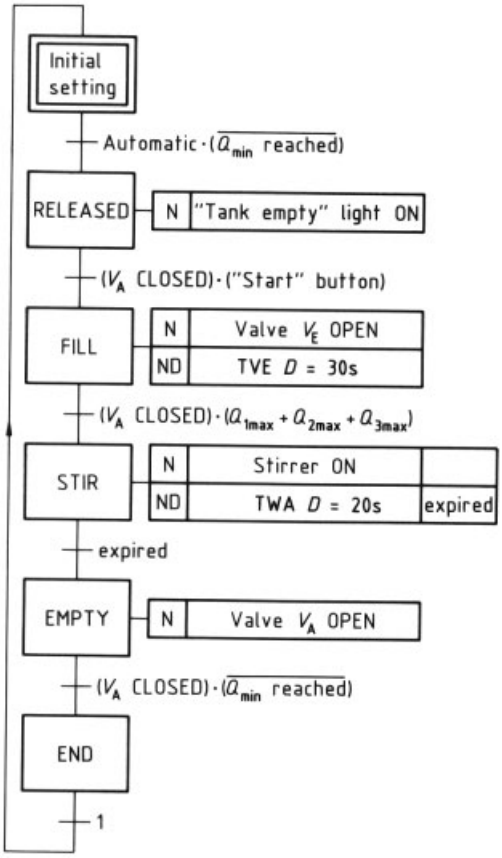

Figure 343. Example of a function chart for tank filling

Preparation of operating handbook
- Preparation for letting contract to outside firms
- Orders and contracts
- Scheduling and resource allocation for installation

The plant design activity prepares design drawings for machinery and apparatus. Construction design supplies building plans as well as calculations of foundations, floor loads, vibrational loads due to machinery, and so forth.

The piping activity supplements its design documents with piping isometrics and parts lists. Care should be taken here to achieve smooth cooperation with local field instrumentation and cable routing activities (e.g., with the aid of suitable 3 D CAE methods).

Planning and design of the process measurement and control system completes the documentation generated in the detail engineering phase.

The PCE station hardware diagrams contain additional information making it possible to localize the station elements by giving, for example, the switchbox identification, rack-mount module number, card slot, and also to identify the incoming and outgoing signal lines by referring to the terminal numbers. Figure 344 shows a detail of the station diagram of Figure 340 with this supplementary information.

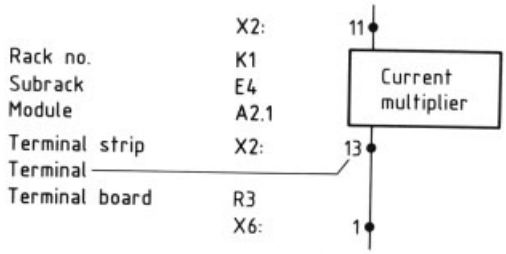

Figure 344. Excerpt from a supplemented station diagram

The station software diagrams, function charts, designs for recipe-mode control operations, and other higher functionalities are transformed into programs understandable by the process control system.

Because screen displays are associated with variables in the software modules, the preparation of screens for process monitoring and control is an expensive process. It can account for a substantial fraction of software costs. Before this step is performed, the screen must be designed. A useful tool for the process of agreement with the later operator of the plant is working drawings, usually prepared by CAD methods, which are easy to change and supplement. Special ergonomic requirements for state-oriented operation are discussed in Section 8.2.

An operating handbook must be written during detail engineering. It should assist in the early detection of design errors and the taking of proper corrective actions. Design errors can be remedied in a relatively cost-neutral way at this point. Corrections upon commissioning are expensive. Unfortunately, little use is made of this capability at present.

The following activities are carried out in collaboration with the purchasing departments:

- Requests for quotations to vendors and/or installation firms
- Orders

When contracting with outside firms, the nature and content of the contracts are agreed on with the legal and purchasing departments.

For installation, schedules should be prepared with allowance for delivery times (see also the network chart in Fig. 345).

It is also desirable to generate a commissioning concept, which should use diagrams and lists to document the following points (among others):

1) Personnel requirement
2) Schedule
3) Procedures
4) Type of functional tests, etc.

More and more computer-aided tools are used as planning and design aids. These support not so much the conceptual phase as the manual or material phase of design, where the plans needed for the erection and operation of the plant are drawn up (see Chap. 9).

Installation. Installation includes the erection of all classes of equipment involved in an investment project and the interconnection of all components into a completed, operational plant. The stages in installation are as follows [631]:

Preparation
Site preparation
Site organization

Execution
Steel construction
Installation of machinery, apparatus, tanks, etc.
Installation of piping
Installation of process measurement and control equipment
Installation in process control rooms
Installation of cable trays in the field
Field installation

Inspection
Machinery, apparatus, tanks
Operability
Tightness
⋮
Process measurement and control
Visual inspection
Wiring inspection
Functional testing

The preparations include the contruction of a site with site offices, storage areas, and installation areas, as well as the creation of a site organization to handle scheduling, resource allocation, costing, and the monitoring of performance relative to these plans.

Good installation planning reveals itself in that the installation of individual categories of assets is coordinated so as to insure smooth, reliable progress. Interferences between categories should not occur. What is more, all resources – such as installation personnel, plant sections, components, and other materials required for the installation of individual asset categories – and all working plans and related documents should be available on time and in adequate quantities and should meet stated quality standards.

The on-schedule installation of process measurement and control equipment requires special care, since as a rule it cannot be started until significant parts of the other categories have been installed. Detailed network charts (see above) can be helpful here.

The installation of process measurement and control equipment falls into three parts:

– Installation in control rooms
– Installation of cable trays in the field
– Installation of process measurement and control devices in the field

Installation "in the field" means all installation activities outside the process control rooms.

Installation work in the control rooms can begin at a very early time in the overall installation process, immediately after the buildings have been completed. Some 20 % of all process control installation work falls under this heading. Cable tray installation can proceed roughly in parallel with piping installation. But the installation of instruments cannot start until ca. 80 % of the piping installation has been done.

Installation ends with an inspection, which is part of the quality assurance (see Section 10.6).

The installation activity must be thoroughly cognizant of the "concurrence" of events in the sense of Petri network theory (see Chap. 2).

Installation offers a broad range of rationalization possibilities for later maintenance practices (see Section 11.4). The equipment manufacturers still have some way to go here (see Section 11.1).

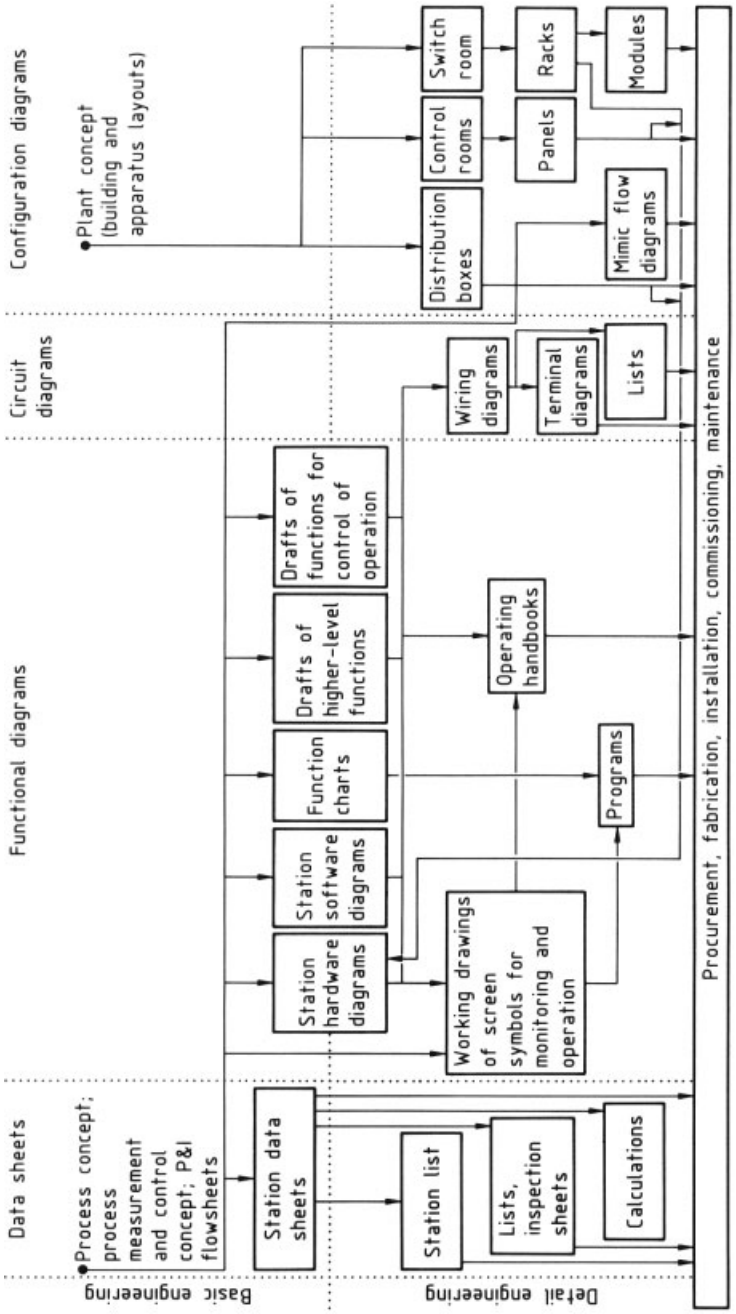

Figure 345. Information flow and process control engineering documentation in the basic and detail engineering phases [631]

10.6. Quality Assurance

The quality of the items prepared must be continuously assured during the project. These include plans, software specifications, programs, plant sections, installation, and so forth. Consistent documentation must be provided for all these activities (Section 3.5, [660,661]).

Major elements of quality assurance are

- Provisions governing planning, design, and installation
- Controls

– Inspection and testing

Today, inspection and testing includes data-model-oriented consistency checks, as known from software engineering, as well as final installation inspection and commissioning.

Governing provisions are, for example, legislation, standards, and internal corporate directives.

Legislative provisions describe the constraints under which a plant is to be erected and operated. They safeguard the environment, in the broadest sense, against possible threats (see Fig. 345).

In addition to legislation, a number of national, regional, and international standards must be considered. These standards essentially serve to ensure consistent planning, design, and execution (see also Chap. 12).

Internal corporate guidelines interpret and give concrete form to the provisions of laws and standards as well as corporate objectives (see Fig. 10).

Examples of provisions for process control engineering are given below:

General planning/design guidelines
– Documentation
– Nomenclature, identifications, abbreviations

⋮

Provisions governing project planning and execution
– Description of scope of work
– Costs; ordering
– Schedules
– ⋮

Substance data
Basis for calculations and design
Technical provisions
– Interconnection of devices
– Protective, safety, reporting systems
– Hardware and software structure of control systems
– Outfitting of process control rooms
– Equipment for building, labs, shops

⋮

Provisions for fabrication and installation
– Design of standard modules
– Design of protective enclosures, distribution boxes, instrument housings

⋮

"Design standard" is understood here to mean technical provisions for the function "planning and design." The design standard for process control engineering is supplemented by an equipment standard, which narrows the range of devices available on the market to a few choices, in particular those that have proved reliable in many years' service or through qualification tests (see Fig. 10).

The scope of work often goes through marked changes in the planning and design of a plant. The reason is inadequate preparation of the investment project. The nature and quantity of the technology to be implemented are not described exactly. Concepts change later on, and designs already adopted must be reworked. The resulting "patchwork" can hardly improve quality and, as a rule, actually degrades it.

Situations of this kind cannot be avoided, but proper preparation of the bidding specifications and specifications can significantly improve them. They must therefore be regarded as an important contribution to improving the quality of planning and design.

Design quality control is a major undertaking, because of the large volume of documents generated. It becomes more difficult if the design work is done by outside firms. For this reason, quality control is often reduced to just a spot-check of results. An object-oriented data model for the plant as a whole is thus all the more important (see Section 3.5 and Chap. 9).

It is shown below for the example of the hardware documentation of a process control engineering project that this procedure is quite adequate for practical use. A differentiated assessment of process control engineering documentation requires breaking it down into units for examination. In the case of hardware documentation, the logical unit is the documentation of a single process control station, comprising the station data sheet, the station diagram, and the corresponding information from the interconnection lists. These documents can be examined as soon as the execution planning has been completed; if the complete job (planning, design, installation supervision, and commissioning) is carried out by one contractor, they can be inspected after the process facility has been commissioned. In the case of a complete contract,

the contractor therefore has to examine the documentation only once.

For the documentation of a process control station to be deemed correct after commissioning, the documents included must meet all the following quality criteria:

- The station identification must be correct
- The documentation must correctly reflect the equipment installed
- Any cross-references required must be present
- Each terminal designation must be correct; that is, the information in the process control station diagram and in the associated interconnection lists must agree with the actual interconnection of the devices

The results of two sampling inspections of process control engineering documentation for a project comprising 324 PCE stations are shown in Figure 346. In the first examination, the sample size was increased in steps of 32 PCE stations (roughly 10 % of the total number of stations). Above a relative sample size of ca. 30 %, the relative error is nearly constant, and it appeared unnecessary to continue the inspection in this case.

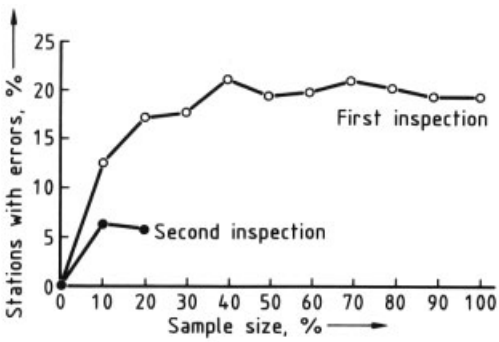

Figure 346. Quality inspection of process control engineering documentation of a project comprising 324 process control stations

In a second inspection by a different inspector, only those errors overlooked in the first inspection were noted. The result is around 5 % and must be regarded as the errors of the first inspector ("slip-through"). Similar slip-through values were found in studies of the manufacturing industries [662].

The results of this examination show that inspection of the process control engineering documentation by the contractor is essential [663]. In practice, the following procedure has proven suitable: If the customer's examination reveals an error lying above a predetermined threshold (e.g., 5 %), the customer returns the entire documentation to the contractor for complete revision. However, if the relative error is < 5 %, the contractor need only remedy the errors found [664].

On economic grounds, only spot checks can be carried out on the process control engineering documentation. Therefore, the risk exists that the error found in the spot check is greater (risk for contractor) or less (risk for customer) than the "true" error that would be found in a complete examination. To resolve this conflict, model calculations have been performed in order to determine the sample size for various project sizes (number of process control stations). The following assumptions were made:

- The error limit for acceptance of process control engineering documentation is 5 %
- The sample must be large enough that the probability of accepting the documentation (contractor's risk) when the documentation contains a relative error > 10 % (twice the acceptable error limit) is < 20 %
- The "hypergeometric distribution" is the appropriate mathematical description of the sampling inspection described [665]

The results of these model calculations are presented in Table 13. The most important result of these model calculations is that in projects with > 100 process control stations, given an acceptable limiting error rate of 5 %, sampling inspection of 50 stations is sufficient to decide on the acceptance of the process control engineering documentation or assurance measures. The advantages of sampling inspection in large projects are greatly reduced insepction costs.

Table 13. Sample size for quality inspection of process control engineering documentation

Number of	Sample size	
PCE stations	Absolute	Relative
20	14	70 %
50	35	66 %
100	50	50 %
500	50	10 %
1000	50	5 %

The monitoring of software design presents major problems. In principle, software design can be checked only during commissioning. A modular architecture is helpful if it permits step-by-step testing, from the single module up to the entire system. This type of examination cannot, however, eliminate software defects that lie outside the scope of the check. Software always hides certain imponderables; hence the recommendation that functions of immediate safety significance never be performed by software alone, but always in combination with a conventional safety level (Section 10.4 and [653]).

It can be expected that the quality of planning and design results will improve markedly if the computer methods still in development (see Chap. 9) are adopted.

Quality assurance also requires installation practices such as the following:

- Issuance of installation standards (technical provisions for the function "installation")
- Incoming inspection
- Supervision of installation activities
- Functional tests

An example for the contents of an installation standard for process measurement and control equipment is listed below:

Provisions on the mechanical construction of process measurement and control equipment in control rooms and other rooms
- Panel designs
- Cabinet and rack-mount designs
- Subracks
- Cooling air supply and distribution
- Cable basements
- Cable lead-throughs
- Cable trays and their routing on pipe bridges
- Distribution boxes
- Transducer boxes design and installation
- Terminal boxes

Provisions on electrical design
- Auxiliary power supply
- Grounding/equipotential bonding
- Cable routing
- Power supply and distribution
- Hookup methods

Provisions on the construction of process measurement and control stations

Installation material in configuration and interconnection

⋮

It contains, for example, provisions on
- The mechanical construction of process measurement and control equipment in control rooms and other rooms
- Electrical design
- The construction of individual stations

These standards should be supplemented by installation layout diagrams (Fig. 347), which use drawings and lists to set forth the installation materials needed for certain tasks as well as their configuration and interconnection.

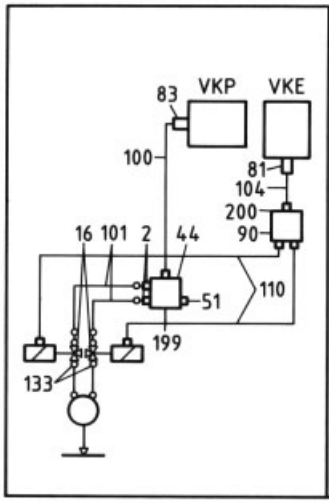

No.	Pos.	Name	Amount	Data sheet	Material
1	44	Hose-pipe coupling, triple	1	7.5 Bl. 1	Polymer
2	51	Cylinder-head screws R 1/4" × 12	1	7.5 Bl. 1	Polymer
3	81	Compression gland PG 13.5	1	6.9 Bl. 1	plastic
4	83	Compression gland PG 21	1	6.9 Bl. 1	plastic
5	100	Pressure transmission cable, 3 tubes, Type D2YBY	20 m	7.1 Bl. 1	
6	104	Wire NYM-I 5 × 1.5	20 m	6.3 Bl. 1	
7	110	Wire NYMHY-I 5 × 1.5	4 m	6.3 Bl. 1	
8	133	Seal R 3/8"	4		
9	2	GE 6-L-1/4" NPT	2	7.7 Bl. 1	steel/cd
10	16	GE 6-L-3/8"	4	7.7 Bl. 1	steel/cd
11	90	Branch box 8102/1 EXE G5 4 × 2.5	1	6.1 Bl. 1	
12	101	Pressure hose Nd4 with 6 mm pipe connection	4 m	7.1 Bl. 1	
13	199	Mount for pneumatic terminal box	1		steel
14	200	Mount for electrical terminal box	1		steel

Figure 347. Example of an installation configuration plan [631]

The quality of an installation activity is also enhanced if standardized, prefabricated, and pretested hardware components and other equipment are employed.

In process control engineering, the use of such standard elements in the form of panels, racks, cabinets, cable trays, subracks for alarm devices, power-supply equipment, and so forth has long been common.

Incoming inspections during installation relate not only to the work plans but also to the materials and equipment to be used. The plans must be complete and correct; materials and devices must comply with specifications.

It is desirable to have specialist personnel – for example, project engineers with knowledge of the process or installation foremen with competence in the work to be done – to supervise the installation activities. Such personnel should be designated by the customer of the project. Only in this way can high-quality execution be guaranteed.

An essential element of quality assurance is the final installation inspection (see "Installation" in Section 10.5, and [631]).

The process control engineering activity carries out three levels of inspections:

1) Visual inspection
2) Wiring checks
3) Functional testing

This procedure insures that the individual process measurement and control devices perform correctly and interact properly.

Visual inspection establishes correct mounting. Proper interconnection of the elements is determined by wiring checks.

The aim of functional testing is to ensure that the active operating sequences of the process measurement and control devices accomplish the functions specified for the process tasks. On the basis of the planning and design documents, the individual process control stations are systematically inspected from sensor to actuator. Process control functions are usually tested by first verifying the lower-level functions and then proceeding to successively higher levels.

A hierarchically organized process simulation would be desirable for functional testing, since only such a tool makes it possible to draw unambiguous conclusions as to whether the plant as a whole will also operate in the intended way. If such a simulation cannot be performed, as is still the case for most processes, this information cannot be acquired until commissioning (see Chaps. 3 and 11).

Commissioning is the initial operation of the entire plant, including process measurement and control equipment, by the operator. It is the last test showing that the process runs in the intended manner. Commissioning can thus be regarded as further quality assurance practice.

Commissioning is performed in three steps:

- Preparation
- Operation with "inert" media ("water run")
- Operation with chemicals

The first two steps are carried out by the departments involved in the planning, design, and erection of the plant. If the project has been contracted out, these steps are done by the outside firms together with the plant operators. The functions to be performed during commissioning are [653]:

Preparation
- Technical training of operating personnel (operating instructions)
- Technical training of maintenance personnel
- Safety training
- Stocking of spare parts and standby equipment
- Securing of feeds, auxiliary media, energy supplies
- Preparation of the plant (purging, drainage, venting, drying)
- Loading of the process measurement and control system

⋮

Operation with "inert" media ("water run")
- Commissioning of auxiliary systems (steam, cooling water, etc.); loops (heating water loops, etc.)
- Testing of safety features
- Simulated operation
- Testing in operating modes (startup, shutdown, load changing, etc.)
- Performance of needed amendments

⋮

Operation with chemicals
- Production test run
- Modification of process measurement and control equipment

– Guarantee run

⋮

After preparations are complete, the plant is started up with inert (i.e., unreactive) products, beginning with the auxiliary and service facilities. Next, circuits such as heating loops are commissioned. After the entire plant has been placed in operation, the safety devices are inspected.

The opportunity presents itself for a comprehensive test of all modes that will be encountered in later operation (startup, shutdown, load change). On the basis of the experience gathered, any amendments needed are now carried out. The documents are revised at the same time.

The next step is operation with chemicals, that is, a production test run. Responsibility for the plant ultimately passes over to the operator.

Ordinarily, a plant is started up at a low throughput and gradually brought up to rated output. The final modification of the process measurement and control equipment to the plant is carried out during this phase. Included in this modification are the adjustment of controller parameters, the input of limiting values, and so forth.

In particular, if the plant has been designed and built by an outside firm, evidence must be presented that the plant complies with the design as to capacity, yield, availability, and quality. This evidence takes the form of an extended test run. The plant documentation (see above) must be finalized at this time.

When commissioning has established quality, that is, shown that the process runs in the contemplated way, the project is over. A detailed final report is mandatory.

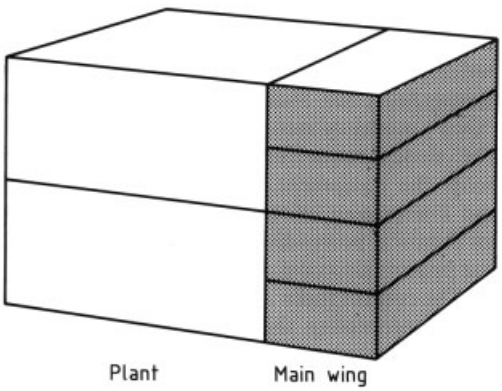

Figure 348. Relationship of plant and main wing

10.7. Process Control Rooms

Introduction. This section describes the requirements that process control rooms must meet, offers suggestions for their appropriate design, and reports operating experience [668].

In the context of optimal human – process communication, it is indicated what requirements, stemming from technical, ergonomic, and other knowledge, govern the architecture, construction, and outfitting of control rooms. The design suggestions are based on scientific publications, standards, guidelines, and operating experience. Details of execution can be found in vendors literature [669].

Architecture. *Relationship of Process Control Rooms to the Plant.* It is advantageous to locate process control rooms relative to the production facilities of the plant so as to create short paths for installation and for access to the plant. The control room should be on a level with those parts of the plant most often accessed. Control room location and structure must afford adequate protection [670].

The "main wing" layout (Fig. 348) offers an optimal way of meeting these requirements. Furthermore, direct connection between laboratory, superintendent's office, and social spaces promotes necessary communication. Protective functions are also taken care of, for example in relation to hygiene or safeguards against undesirable external influences.

In large-area combinations of facilities (Fig. 349), it may be useful to erect a freestanding control building to take over a variety of operational functions.

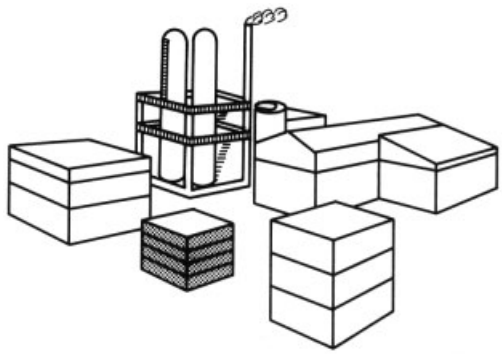

Figure 349. Layout of plant and control room buildings

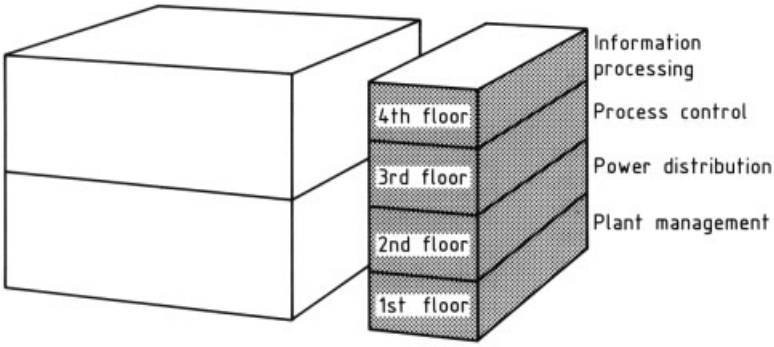

Figure 350. Breakdown of the main wing

Arrangement of Control Rooms. Process control rooms and other rooms that house, for example, plant management should be laid out in a functional way so as to permit an optimal flow of operations. This can be done whether these spaces are located in a main wing (Fig. 350) or in a separate building.

The control room is the central area for the performance of essential process control engineering functions. Such tasks are described in the "level" model of information (see Section 2.3). The switch room for information processing holds the equipment necessary for processing and interpreting information. The switch room for electrical power distribution is the site of central process electrical equipment.

The control room and switch room for information processing should be located close together. The electrical switch room should also be placed in the main wing. If distances are long, however, electrical distribution for the plant can be housed in more than one switch room (to minimize cabling costs).

The central communications equipment for a plant needs its own room. Rooms for process analytical instruments should be placed at crucial sampling points in the plant. Special requirements apply to these rooms [671].

The following points, in particular, must be considered when positioning control rooms relative to the plant and one another:

- Safe access and escape routes
- Optimal cable and utilities routing
- Fire safety
- Favorable location for fire fighting if necessary

Space Requirement for Process Control Rooms. Space requirements can be determined with graphs or ratios derived from operating experience, which offer guidelines (Fig. 351, see next page) for initial control room design. The increase in individual functions expected through the end of the project has to be estimated and reserve space must be provided.

Starting with some minimum size (ca. $25\,m^2$), the space required for a control room is worked out from the number of process control stations to be observed and controlled. The figure of $0.1\,m^2$ per process control station is a lower limit for conventional instrumentation. The use of process control systems and process control computers will reduce the area needed.

The space required for the information processing switch room is also governed by the number of process control stations. A minimum figure is again $25\,m^2$. With conventional instrumentation, the minimum additional area required is $0.2\,m^2$ per process control station. The use of process control systems will cut the space requirement.

The installed power is the essential factor determining the needed area for the power distribution switch room; the minimum size is $15\,m^2$. About one square meter is required for every $10-20\,kW$ of installed capacity. If there are many variable-speed drives, however, the required area per kilowatt may increase.

Ergonomics. Ergonomics is concerned with the interrelations of human beings and the world in which they work [666]. Due attention to ergonomic factors in the design and construction of control rooms is a precondition for good

human – process communication (see also Section 11.2).

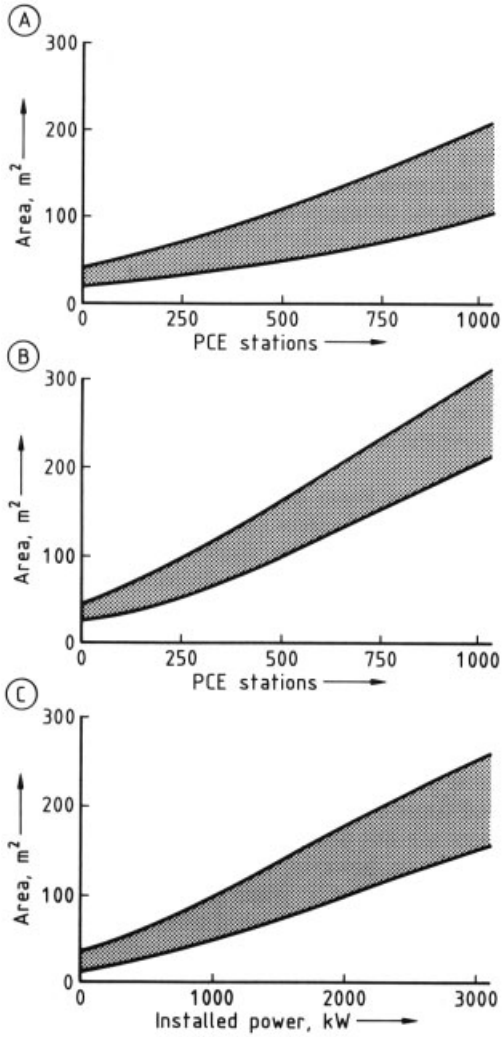

Figure 351. Space requirements for process control rooms

Anthropometry. Anthropometric considerations lead to guidelines for spacings, openings, and clearances in the placement of individual devices and equipment in control rooms.

Pushbutton controls, switches, and display elements should be arranged so as to allow optimal access and visibility (Fig. 352) [672, 673]. An important factor in determining the appearance and size of labels is the distance from which they will be viewed. When locating display and control panels, care should be taken that they are in easy view of the operator.

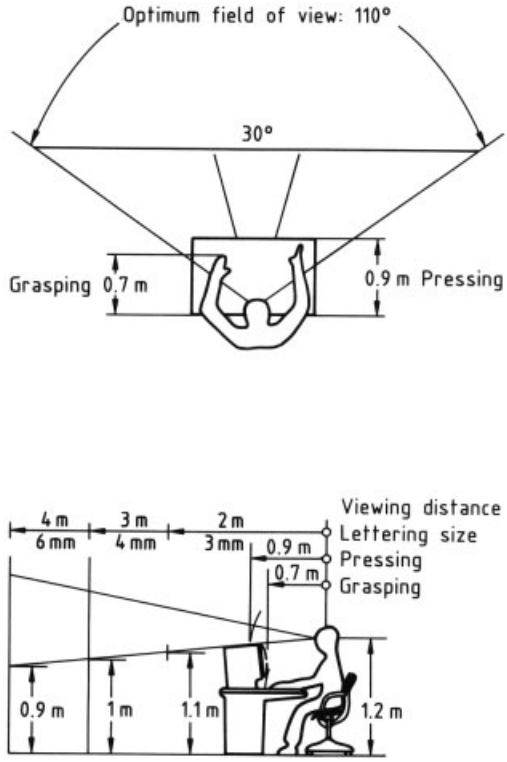

Figure 352. Optimal manipulation and viewing areas

Lighting. A variety of visual tasks must be performed in control rooms (Fig. 353). The type and intensity of illumination must be adapted to each such task [675]. Illuminances between 200 and 500 lux are recommended for error-free reading of meters and other readouts, illuminated symbols on panels, and characters on screens and displays. Lighting properly adapted to each visual task features separate, continuously variable brightness adjustments for the two zones. Fill-in lighting serves to enhance the local illuminance on printers, desks, and so forth. An appropriate configuration of lighting fixtures in combination with glare screens should help to prevent reflection from displays. Since control rooms are generally occupied day and night, the lamps should provide daylight-balanced lights in order to insure consistency in color and illumination.

Switchgear rooms should have an illuminance of 300 lux.

Backup lighting must be provided in control rooms and is recommended in switchgear rooms. Safety lighting must be provided for pur-

poses of accident prevention. Details of lighting design can be found in company literature [667].

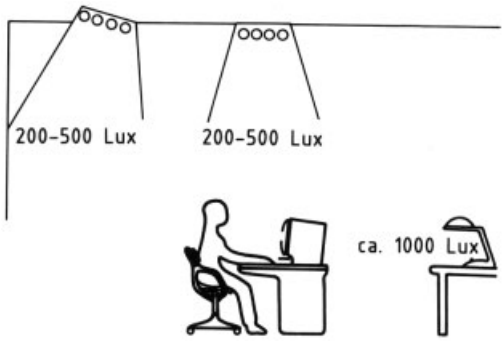

Figure 353. Control room lighting

Climate Control. Ambient conditions in the control room must be held constant within a fairly narrow range [674]. Furthermore, because personnel are present at all times, there must be a continuous, draft-free supply of fresh air (5 to 10 air changes per hour; Fig. 354).

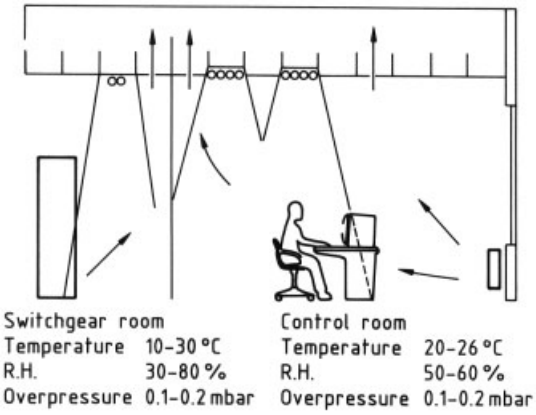

Figure 354. Ambient conditions

Switchgear rooms are not continuously occupied, and the tolerance range for the ambient conditions is therefore wider. In contrast to the control room, much heat is evolved here and must be removed by a waste-heat removal system. Computer components belonging to process monitoring and control systems may have stringent climatic requirements, which can best be satisfied by separate climate control devices.

Undesirable pollutants must be kept out of all rooms used for process control functions. A slight overpressure is therefore maintained in such areas; filtered fresh air from pollutant-free areas is delivered through special floor registers and extracted through ceiling grilles. The control room must have the capability of shutting down the ventilation of all rooms at any time.

Acoustics. Acoustic conditions in the control room should be such that personnel there experience no difficulty in understanding one another or in communicating with the outside world by telephone or intercom, and also such that acoustic signals can easily be detected by the human ear. Care should be taken that the noise level does not exceed the 55 dB (A) ambient noise, the reverberation time in the middle frequency range (500 to 3000 Hz) is ≤ 1 s, and the level of acoustical danger signals in an octave range lies a minimum of 10 dB (A) and a maximum of 20 dB (A) above the interference.

Control Room Design. *Doors and Windows.* On grounds of safety and health, pathways between the production facility and rooms that house process control functions may be equipped with airlocks. Airlock doors must be fireproof, self-closing, and gasketed, and they must open in the escape direction. Larger monitoring and control rooms must have a second emergency exit.

Because communication with the process takes place exclusively via process monitoring and control devices, windows opening into the plant are not needed for communication. Windows may, however, be required by workplace regulations [676]. When control rooms are located near explosion hazard areas, appropriately rated exterior windows must be provided.

Floors. Control rooms should be constructed with floating floors. Cables are led through floor penetrations or cable ducts placed in the floor (Fig. 355). A variety of special plastic floor tiles are available. Ceramic tile is used where contamination is severe. Double flooring in computer and switchgear rooms (Fig. 356; both figures see next page) permits free cable routing.

Ceilings and Walls. False ceilings are particularly suitable for control rooms. They offer a range of options both for concealing the unfinished ceiling and for the placement of functional elements such as fire alarm and fire-suppression systems, air conditioning, cable ducting, and lighting fixtures. Esthetic, acoustical, and cli-

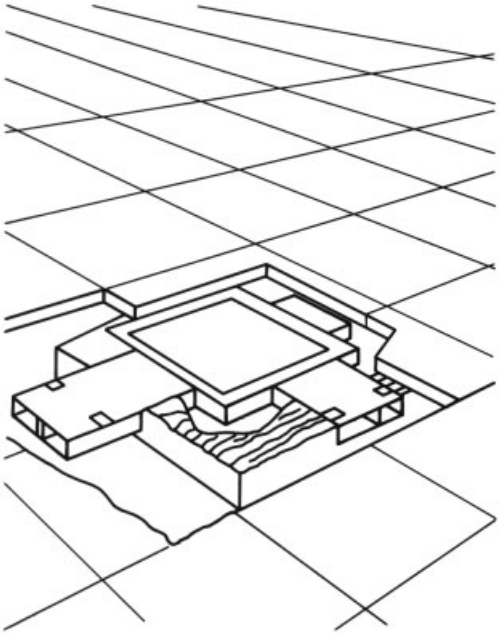

Figure 355. Cable duct installation

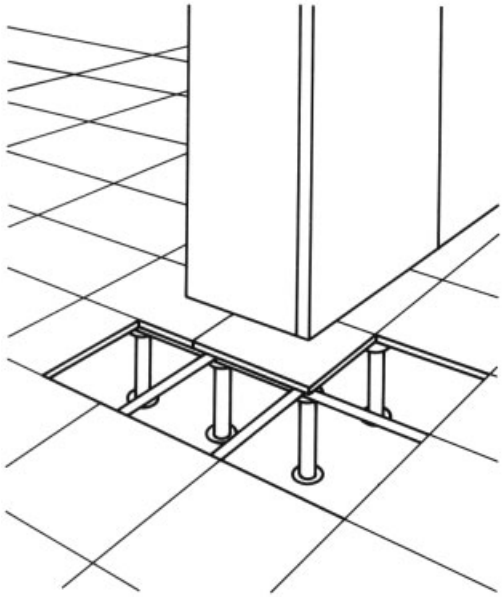

Figure 356. Double floor

matic considerations require attention to not only the ceiling but also the walls. All other rooms that house process monitoring and control functions have ceilings and walls painted in light colors.

The control room should have an unfinished height of 4 m (to bottom face of ceiling joists). The height of the false ceiling should be 3 to 3.50 m, depending on the size of the room. A height of 3 m (to bottom face of air ducts) is recommended for switchgear rooms.

Ceilings and walls must comply with pertinent fire-protection codes (F 30 or F 90); this applies particularly to cable penetrations.

Desks (Fig. 357) hold visual displays and controls such as keyboards, light pens, joysticks, mice, and touch-screen input devices. Horizontally and vertically pivoting display bases permit adjustment to suit individual preferences [673]. Additional free space is needed, for example, to accommodate a radio and telephones and to permit occasional writing.

Operator stations in control rooms are used for monitoring and controlling the process. They enable the operator to intervene in the process as required. An operator station (Fig. 358) generally consists of a color graphics monitor or monitors and controls. The controls usually consist of a keyboard (with movable or touch-sensitive keys) for the input of commands and/or a light pen with which the process can be controlled directly on the screen.

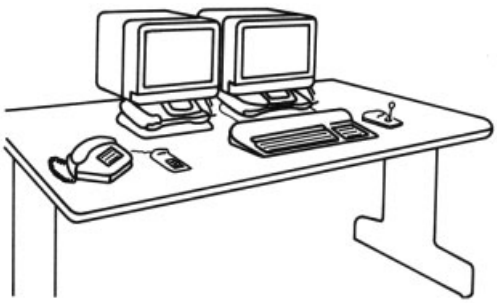

Figure 357. Operator desk

Figure 358. Operator station

The light pen has the advantage that input errors are largely eliminated and performance speed is roughly doubled. Fatigue-free operation with the light pen is, however, possible only with the elbow supported, so that the monitor should be located only ca. 40 cm away from the operator. Other control arrangements employ tracking devices such as joystick and mouse or permit operation by touching the front of the screen (touch-screen devices).

Hard-copy devices and printers are used to document the process in printed form. They should be installed in noiseproof, dustproof enclosures in a readily accessible location (unless they have to be placed in a separate room).

The *process communications wall* comprises panels holding extra monitors as well as individual displays and controls (Fig. 359). The preferred configurations are straight, L-shaped, and U-shaped. Mimic flowsheets can be laid out along the top. The various functional elements are modular devices. To make the spatial organization clear to the observer, they can be arranged in accord with the production flow.

Figure 359. Process communications wall, furniture, and accessories

Furniture and Accessories. Besides the monitoring and control devices, the control room should be equipped with the following (Fig. 359):

- Bookcases for documentation as well as storage of supplies such as printer paper
- Desk/worktable
- Bulletin board
- Flip chart
- Wall clock showing date, day of the week, and time
- Storage for protective clothing and safety devices (e.g., respiratory protection, fire extinguishers). This should be placed where the operating personnel can access it quickly and without hindrance on the way to the plant.

11. Operation

11.1. Principles

This chapter discusses some points relating to process control engineering that must be borne in mind when operating process plants.

Along with reliability (environmental, occupational safety, and health issues) [677–679], the most important quality of process operation is reproducibility. In other words, the required product properties [680] of the end products must lie within the respective tolerances (see also Fig. 34). The same applies to all relevant intermediate species, of course. Product properties should always be regarded as "primary nominal values" [681].

As shown in detail in Chapter 2, product properties are described in terms of individual process elements. These in turn are described in terms of process properties [680]. Process and product properties are linked by transformations expressed in "operators," and the product properties can thus be represented as functions of the process properties [682, 683]. Gilles [684] takes this a step further and, in a first formulation, allows the set of product properties in the production process to be categorized as a state variable.

Because the above properties or attributes are mostly multidimensional aggregations, it is meaningful to refer to "product spaces" and "process spaces."

Thus, to achieve product properties complying with the requirement profile, the respective process properties must be maintained in each process element. These in turn have their own requirement profiles with corresponding tolerances. These requirement profiles of the process properties in each process elements are to be understood as "secondary" nominal values or setpoints. They are constituents of the control

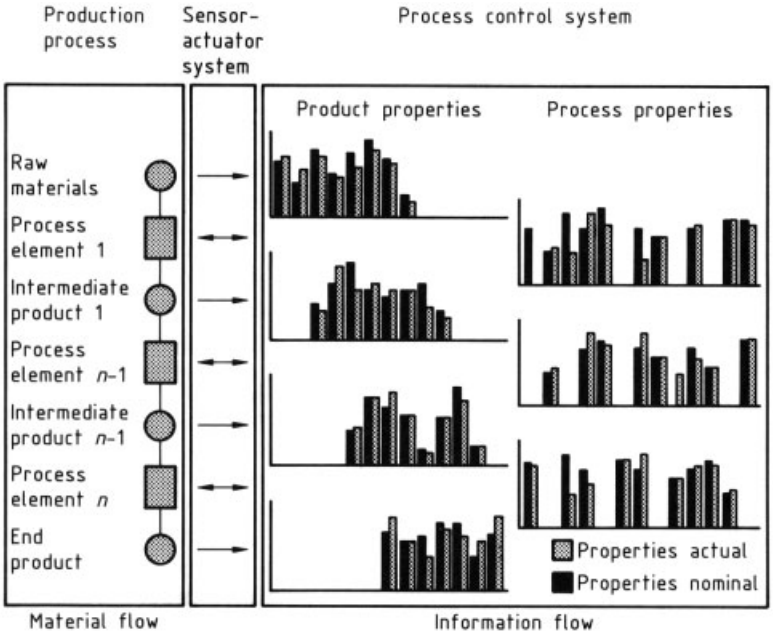

Figure 360. Nominal/actual comparison for a production process

loops (generally involving thermodynamic and reaction-kinetic quantities) in the process control system.

Figure 360 shows a nominal/actual comparison for a production process. This representation is the most important piece of information required for controlling the process. Detailed discussions can be found in Sections 2.3 and 11.2.

When production processes are to be optimized with respect to certain target quantities, process models must be created. These are math-ematical descriptions including the interdependences of process properties as well as the dependences of product properties on process properties. As shown in Chapter 3, models can be subdivided into purely empirical or statistical models and "theoretical" models.

Statistical models are derived from the actual (measured) product and process properties in the real process running in a pilot or full-scale facility (see Fig. 360 and Section 3.2). Its equations are purely mathematical in nature. It is difficult to assign physical interpretations to its variables by using techniques to reduce the dimensionality [685, 686].

Static and dynamic process models are essential for effective simulations.

In general, "simulation" refers to the use of information technology to forecast the behavior of the independent variables in a system when the independent variables or the initial and boundary conditions change [687, 688].

Provided the basic model (see Chap. 3) can be formulated in explicit mathematical form, the fundamental problem is to find suitable numerical procedures for the economical solution of the mathematical problem. As interactive communication between humans and the process becomes more important, the graphical presentation of the data model or functional model (see Chap. 2 and [689]) is gaining in significance [682].

Problem-oriented "object spaces," with the "attributes" of the objects as their coordinates, should be designed so as to promote human thought and learning (see Section 11.2). Information reduction is a key tool to doing this.

This holds both for training a user concerned with a technical, economic, or sociocultural system (see Chap. 2) and for devising solution aids for finding "optimal" operating points. Zeitz [687] is therefore justified in according the highest importance to such simulation. The framework for system modeling must be extended beyond the customary metric scaling into the realm of topology. In con-

trast to well-known solution approaches, simulation thus takes in predominantly algebraic and differential-equation-oriented structures as unified topologies of knowledge, as suggested by GILLES [690], in his discussion of knowledge-based process control engineering. Numerical and knowledge-based solution tools will have to be harmonized with one another in the future. For example, the NAMUR Status Reports include noteworthy articles offering many suggestions and solution approaches [691–697].

The modern model-aided procedure will also lead to new and expanded safety concepts. These include concrete approaches to the early detection of hazardous conditions based on model-aided measurement methods [680]. Experience in biotechnology also suggests that applications will appear in prognostic medical diagnostics.

One of the principal tasks of process control engineering is to increase the degree of computability in each process step and to represent it in suitable process models. However, only a purposeful collaboration with process engineering can lead to a fruitful symbiosis here. This holistic mode of analysis will be established in a new engineering discipline, process technology [698]; see also Section 11.3.

Representations that are closely related to the phase model of production are the Petri nets [699]. While the phase model represents the topology of product flow, Petri nets describe the system of conditions governing product flow. In the context of Petri net theory, any technical system can be represented in terms of two description categories:

- Existence of a state
- Occurrence of an event

Each state is terminated by at least one event and/or brought into being by at least one event [700]. States can also be interpreted as conditions that must be fulfilled for a transformation specification to be carried out. In order to describe the dynamic sequence, a further description element – the token – is introduced. The token moves through the static net as a traffic unit. The total population of the net by tokens characterizes a global state of the net.

The central modeling aspect of Petri nets is the concept of concurrency. Two events are said to be concurrent if they are causally independent of one another [701].

In processes, buffers are used in an attempt to decouple the flow of material and prevent disturbances propagating through the entire plant. Subsystems decoupled via buffers can operate independently of one another, that is, concurrently, within certain limits. In this way, Petri nets can be applied to production processes. The net theory offers the possibility of modeling such systems and determining their dynamic properties by computation or simulation. Local analysis and modeling has two effects: It reduces the complexity and it creates the possibility of using a process-logistical procedure along with decoupling via buffers [702].

In automatic mode (for which recipe-mode operation is a prerequisite), the process-logistical problem (Table 14) [703, 704] is that multiple intercoupled process elements operating in distinct modes must be managed in terms of their mass flow. The definition of operating modes is less well-known than the "recipe" concept discussed above. It describes all modes of operation within a process element, including exceptional and disturbance situations, in discrete form. It must be set in phase-specific fashion for the batch case and section-specific fashion for the continuous case. For a continuous process element and its transitions, the various operating modes are shown in Figure 361 [705]. Because of the wide range of possible mode specifications for various process elements, the process states desired in disturbance-free operation and the process states sought when disturbances are present are specified as functions of the instantaneous state, in well-defined steps, or are calculated in real time and set down in control tables. The contents of these mode-specific tables are derived from simulation calculations or experimental data.

Figure 362 shows schematically how the categories of importance for operation – reproduction, optimization, and automation – are interrelated.

WÜCHNER [706] has discussed which requirements apply to personnel when modern process control systems are used. KRAUSS [707] sees the acceptance of process control systems operating in information-oriented fashion as essentially insured by correct design of the display and control components of the control systems.

What applies to communication between humans and the process control system is equally

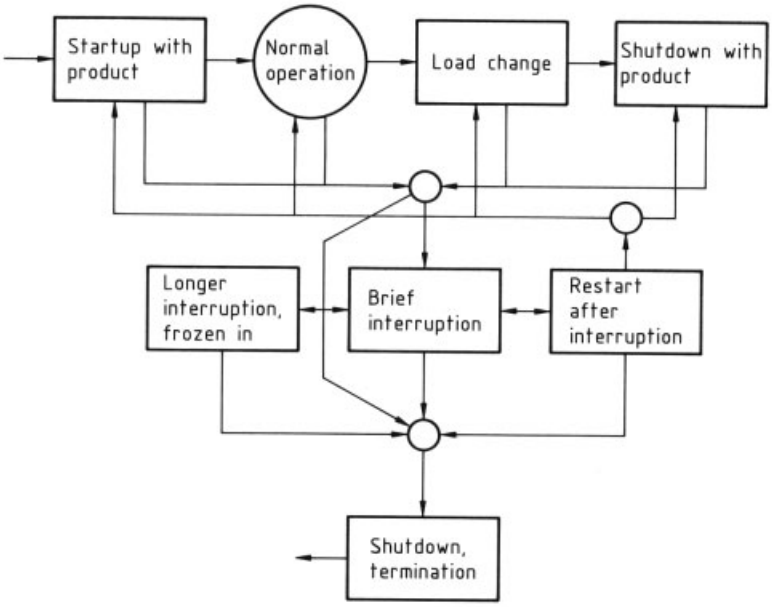

Figure 361. Operating modes of a continuous process section and their possible transitions [705]

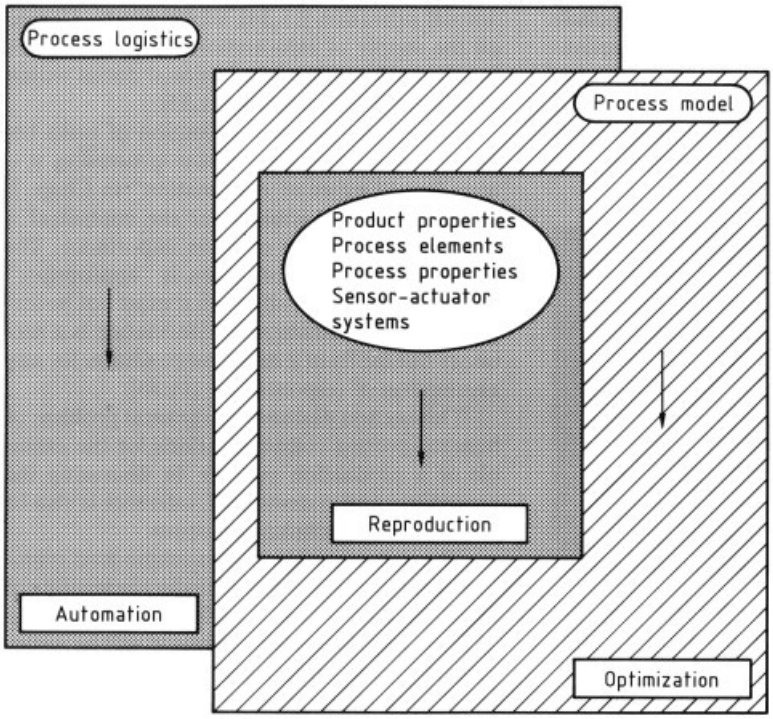

Figure 362. Reproduction, optimization, and automation

Table 14. Levels of automation

Process elements	Mode	Necessary information management	System hardware	Characteristics	Realization
Decoupled	1 manual	all flow paths and all process/product properties manually input as setpoints, managed, and documented	all sensors, all actuators, individual controls and control systems	reproducible production in all operating modes (process models)	$\geq$75 %
	2 computer	all flow paths and all process/product properties input to computer as setpoints, managed, and structured in documented form BF∗: switch net; set parameters; BF∗ provided with setpoints	as in 1, plus PCS∗ with BFE∗ and BF∗, PCC∗	reproducible production in all operating modes with recipe operation, process models, simulation, optimization	$\leq$20 %
Coupled	3 logistical	as in 2, plus buffer management, buffer strategy, source-point management	as in 2	buffer strategy-oriented recipe mode for all operating modes	$\leq$1 %
	4 automatic	as in 2 and 3, plus control of divergence of process elements through variation of process properties	as in 2, plus net orientation (Petri nets)	divergence-oriented recipe mode for all operating modes	1995

∗ BF = basic function; BFE = basic functional element; PCC = process control computer; PCS = process control system.

important with regard to the sensor systems, which serve as "eyes" on the process [708].

After this general discussion of operation, the following sections deal with human – process communications (Section 11.2), process analysis and process optimization (Section 11.3), and maintenance strategies (Section 11.4).

Human – process communication comprises the integration of information acquisition (sensor technology), information processing (control system), information and knowledge utilization (human component), and information feedback to the process (actuator systems). The totality of these processes (Fig. 363) insures proper operation of the process.

Process analysis and process optimization aimed at achieving permanent gains in reproducibility, safety, and economics is a single engineering task from the viewpoint of process control engineering.

Maintenance strategies aim to preserve and improve the availability of the production facility and its process control components.

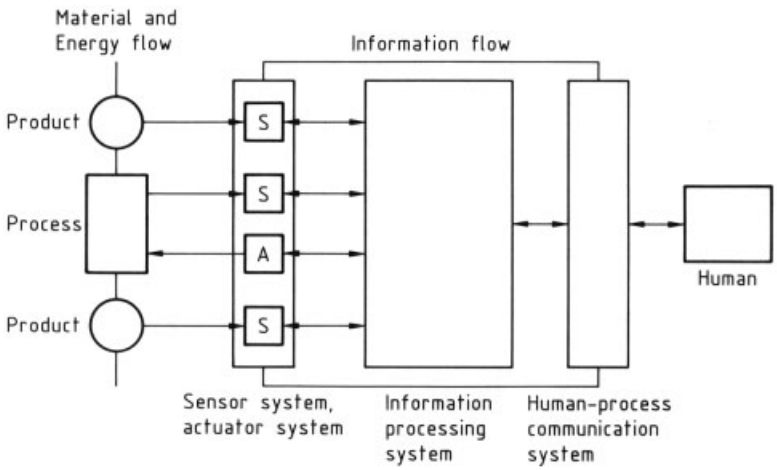

Figure 363. Human – process communication

11.2. Human – Process Communications

Introduction. Humans today are surrounded by a great number of technical processes. In production, new products are made from many elements by combining operations. Production arrives at new products through the reaction of feed materials. Energy generation and distribution processes are of major importance. Other important technical processes permit the transportation of people, goods, and information on land, by water, and in the air.

The human role in all these processes is to monitor and control. This can be carried out in a "planning and forethought" mode. The human optimizes technical processes as to structure and sequence, detects hazardous situations in the processes at an early stage, and can then intervene promptly to remedy them.

The complete interaction between human and process, which is gathered together under the phrase "human – process communication" in what follows, is governed by the operator, working at an operator station, who runs apparatus and devices, observes measurements, and alternately abstracts and concretizes in an attempt to comprehend what is happening in the process on the basis of individual-signal observations.

In a production process, information relating to an individual process apparatus is accessible at the individual measurement stations still to be seen everywhere (Fig. 364). There are also very large control rooms (Fig. 365) in which information from large plants is gathered together and displayed on consoles and panels. The color graphics display (Fig. 366) is a modern alternative offering an easily and quickly comprehensible representation, even for complicated processes.

Any production process has an a priori dependence on the combined functioning of physical, chemical, and other sensors, which provide the human operator with knowledge about the properties of products and processes by means of appropriate information processing. Automation of production processes, in contrast, leaves process events in a state that can ultimately be comprehended by human senses, except as required by systems-engineering and social-psychological exigencies [709].

The preceding chapters have taken up the problem of what information should be determined for process control purposes (see Chaps. 2, 3, and 7 and Section 4.5). Now to these is added the question of which presentation of the available information is best suited to human communication abilities. Modern resources of information technology hold much promise in this area. As yet, however, there has been no ex-

Figure 365. Large control room for monitoring operations

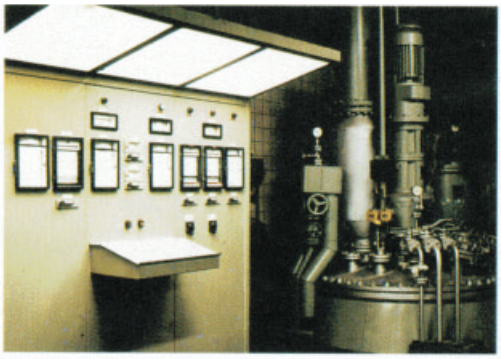

Figure 364. Individual measuring stand in process engineering

Figure 366. Color display with light pen and keyboard (Siemens Teleperm M [762])

act study of their capabilities in the context of information-oriented process control.

Shortcomings are often seen in the selection and presentation of information as commonly performed at present. For example, an intermittently severe cognitive stress is imposed on operators in "exceptional" situations, which include not only changes in operating values in continuous plants (startup, shutdown, load change), but also unforeseeable changes at the dispositive and operative levels. Ironically, plant automation often leads to a high degree of monotony during normal operation [710], which can be blamed on the supply of information being inadequate for the performance of additional duties within the sphere of competence [711].

Besides the proper distribution of tasks between human and information-processing systems (see Section 7.1, Fig. 241), the responsibility for the control of a process is a question of some importance. The question of what information is "appropriate" becomes significant, although it must not be forgotten that furnishing such information is a necessary but far from a sufficient condition for any assumption of responsibility [712].

This section discusses the general problem of human – process communications for the example of a production process. In discussing the procedures and systems used, reference is also be made to other technical processes.

The special problems of machine-level monitoring and control, grouped under the heading of MMI (Man – Machine Interface), is not the point of focus here, although at times it will be necessary to point out crucial differences between MMI and MPC (Man – Process Communication). A helpful introduction to human – machine communications can be found in [713]; a useful reference work is [714].

First, it will be beneficial to survey briefly the capabilities of humans in their interaction with large bodies of information today. In the commercial and administrative sphere, the high-speed printer serving as output medium in computing centers generates long columns of numbers and symbols on tall stacks of paper (Fig. 367); this device continues to play a central role. The term "business graphics" refers to a mode of presentation that is more suitable for summarizing complex information (Fig. 368).

Until recently, engineering devices first took form on the drawing board. Computer-aided design (CAD) marks the beginning of a new world in which design is supported by all the resources of modern computer technology.

Figure 367. High-speed printer with tables

Figure 368. Business graphics

Acquisition and Presentation of Information. In earlier centralized process control and more recent decentralized systems, all relevant information about the technical process (and today also about the distributed process control system) must be assembled at a central location so that process control can be optimized under holistic criteria. This central process control location must also have a high availability so that operating personnel can carry out monitoring and intervention at any time. High availability is also required because process condition documentation is often mandated by law.

Human – process communication began in the early 1900s with direct display and control by field-level devices (Fig. 369). The re-

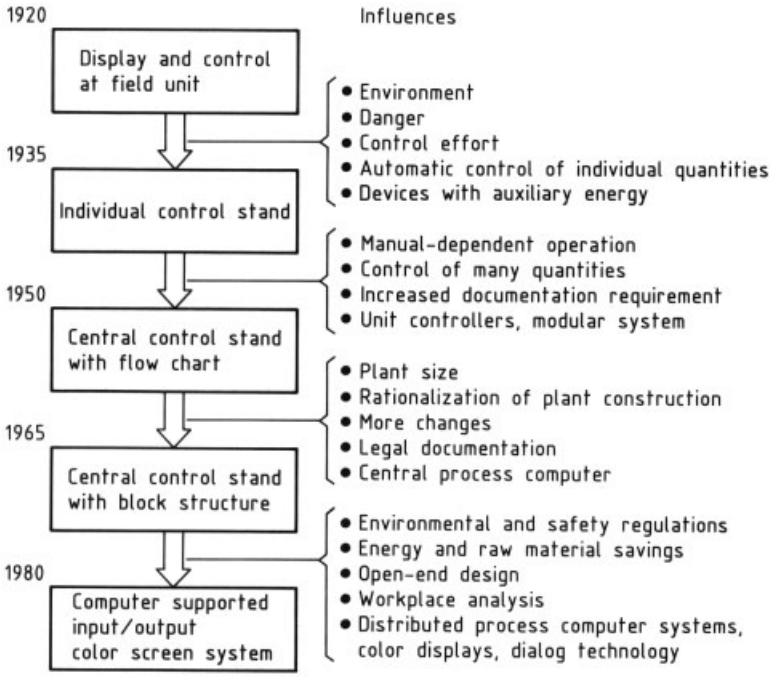

Figure 369. Development of process control room design [710]

sults gained, together with the development of auxiliary-energy field devices, led to a concentration of such devices in a local control stand. Increased plant size and more stringent documentation requirements, aided by unit controllers and modular control systems, led to the central control stand with a flowsheet-like representation of the process around 1950. This flowsheet technique has several advantages for human – process communications:

- The process structure (or "process topology") and the condition of the process are displayed together. An example is the combination track diagram and signal/control box used by the German Federal Railway (Fig. 370).
- The assignment of control and display elements to the process is implemented through a suitable configuration in an abstract, graphically represented process structure. As a result, control elements can be directly addressed and actuated while their place in the process structure is clearly shown.

With increasing process size and number of process states, the space requirements and costs of hardware process flowsheets have grown. In the case of extensive processes such as those in the petrochemical industry, process control rooms employing flowsheet displays are up to 100 m long and can be managed only by a correspondingly large and mobile staff (Fig. 365), hence the names "bicycle" and "rollerskate" contol rooms. High costs of creating such a display are followed by high costs of modifying it, so it is difficult to alter the flowsheet-type control room to keep up with changes in the technical process.

Figure 370. Combination track diagram and signal/control box used by the German Federal Railway

From 1965 on, increasing plant size and the growing number of display and control devices in the central process control room – along with the availability of central process control computers – led to the central control stand with block structure (Fig. 371), similar to the quick-scan display technique used in aircraft design. Control rooms employing the block-structure principle have the following features:

- Like instruments are placed in groups
- The relationship between operating and display elements and the technical process is identified by names (nameplates)
- The control and display elements are selected and actuated indirectly; there is no direct topological relationship to a graphical representation of the process.

Figure 371. Central control stand with block structure

The operating personnel must therefore remember the process structure. It takes several steps to select and actuate the control and display elements, and this operation calls for good memory and abstraction abilities. The purely technical capabilities of the block-structure control room as to modification are certainly superior to – and more user-friendly than – those of the flowsheet-type room, but it is more difficult for the operating personnel to keep up with frequent changes in the room.

With the availability of powerful and economical computer-supported color displays, the representation methods of the block-structure control room can be retained and adapted to the capabilities of the display screen.

The most important types of video display now in use are discussed on the basis of the control and monitoring devices employed in the process control systems TELEPERM M [715], MASTER [716], and Geamatics [717].

Figure 372 shows the general block-structure principle with color graphics displays, with details presented in Figures 373, 374, 375. These are standard or predefined displays and exist as prefabricated data records with the basic representation of images for plant areas, groups, and individual control loops. All the designer must do is adding plant-specific data to this data record (see Chap. 8).

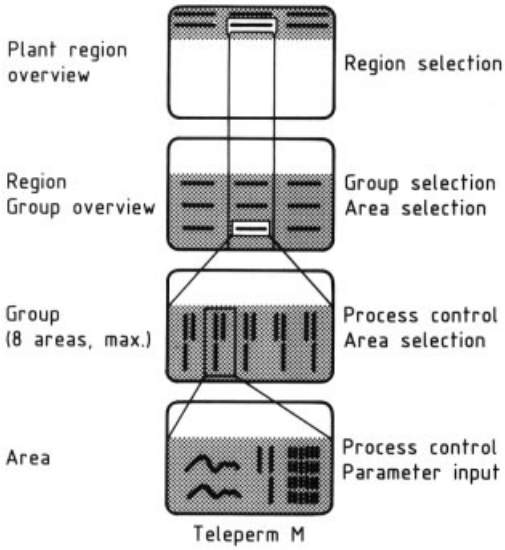

Figure 372. Block-structure principle with color displays [703]

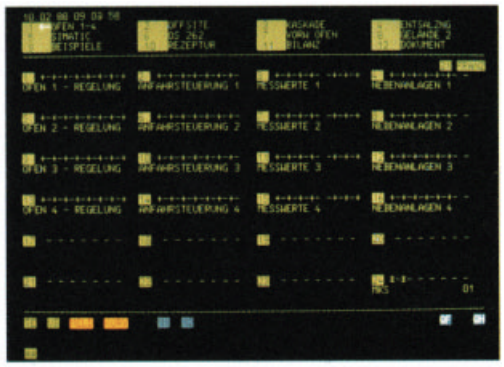

Figure 373. Summary diagram [715]

The so-called free display is the exact opposite of the standard displays. Here the designer has a totally free hand. Standard symbol sets offer an aid; elementary symbols can be com-

Figure 374. Group diagram [717]

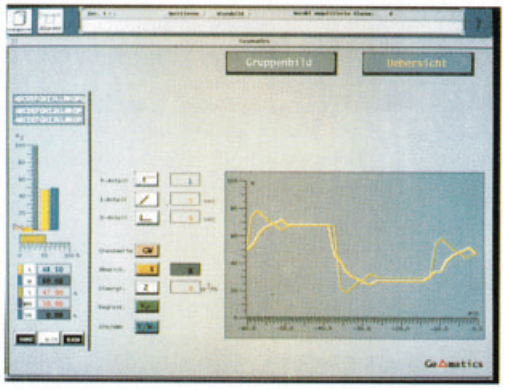

Figure 375. Loop diagram [716]

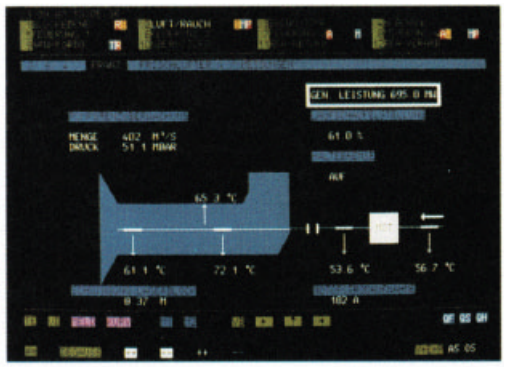

Figure 376. Free display [715]

bined to make new symbols or a toolbox with elementary graphics objects such as circle and rectangle. The designer can thus represent anything from a valve to a steam boiler. If certain symbols are not included, a new symbol set can be put up on the screen. Separately stored subimages can be inserted into various overall images, thus saving much time. State displays are also easy to create. In on-line operation, they appear on the screen with the true values, thus facilitating a nominal/actual comparison. Many such free displays can be stored in the operating computer as a data record and can be scrolled through sequentially (Figs. 376 and 377).

Standardized menus simplify the creation of curves and groups of curves, the assignment of curves to measurement stations, the assembly of curves into full plots, and the specification of storage form, display form, measurement range, and unit of measurement. The system may request all these details in interactive mode.

Curve plots can be presented in a standard or free display. Depending on the physical input device, the selection is made by pointing to a digital or a bar-graph display with a light pen (Fig. 378), touching a touch screen, or using a mouse or trackball to position a cursor. Curves are shown in windows with the required deeper level of detail.

Complete standard images can also be incorporated in these windows, for example, the standard representation of a plant group. Win-

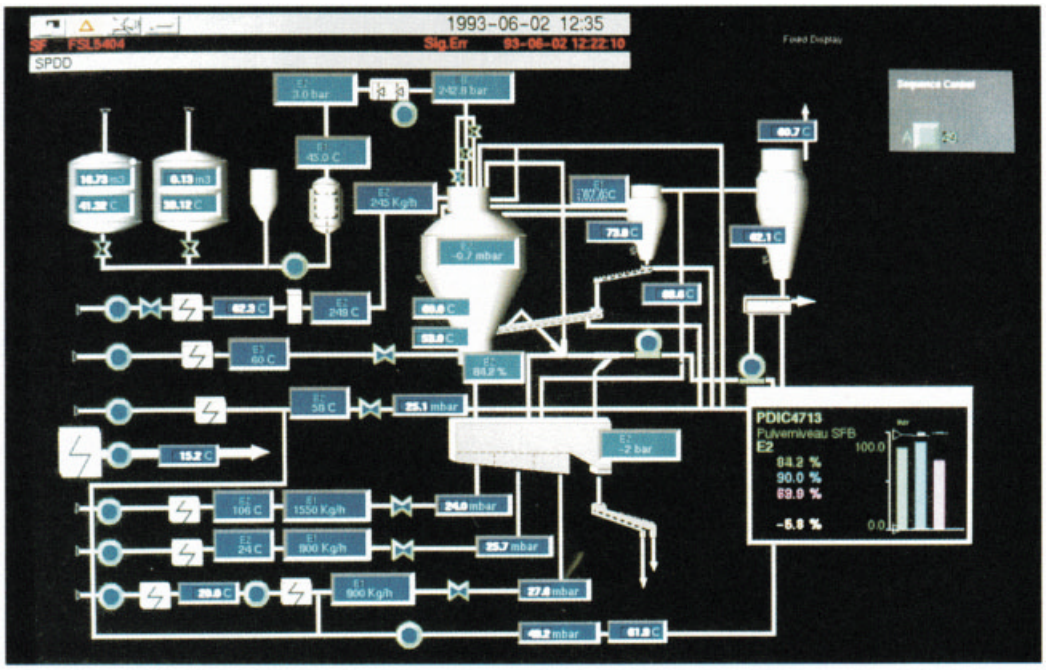

Figure 377. Free display [717]

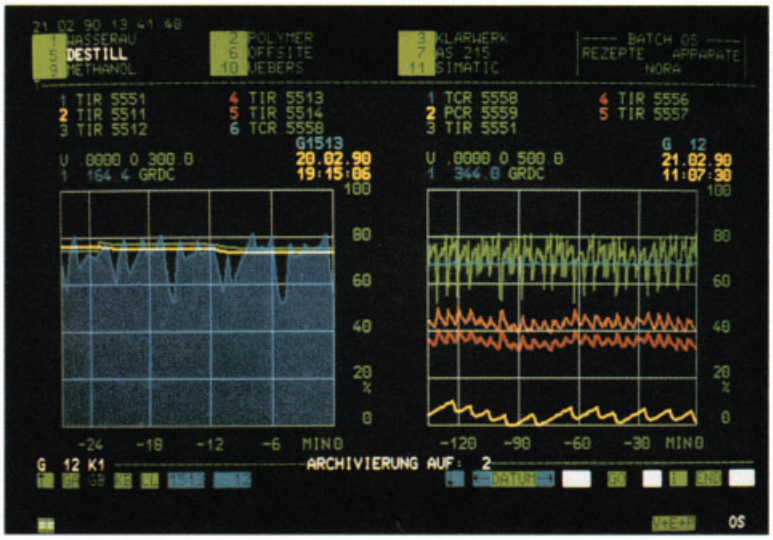

Figure 378. Curve plot [715]

dow support makes process displays far easier to understand, since only the permanently needed information is always visible; details appear in the window only when needed. The window size and position are determined by the designer; in modern systems, they can be modified on-line by the user (Figs. 379 and 380).

What is more, user procedures and reports can be freely designed. In doing this, the user determines everything: size and layout as well as dynamic information up to the way in which report creation is initiated. Reports can be generated by specifying a start time and a printing cycle (shift report) or certain events (charge re-

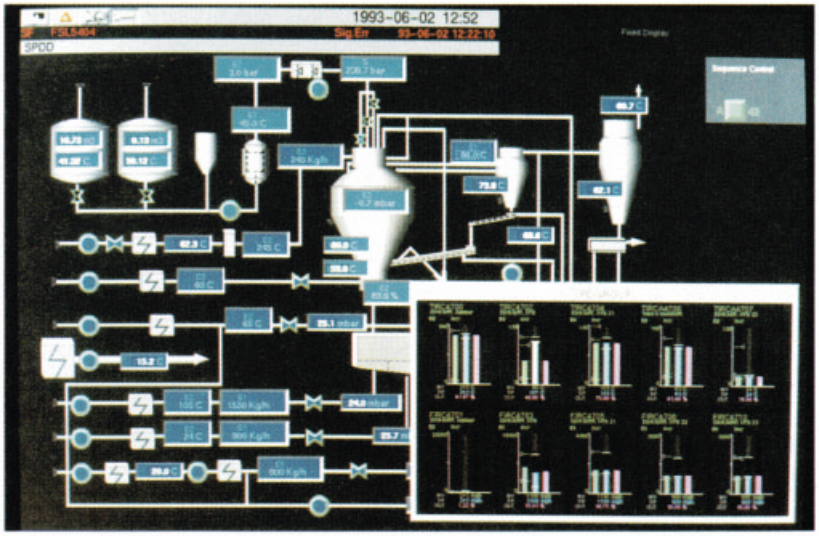

Figure 379. Windowing: standard displays in free display [712]

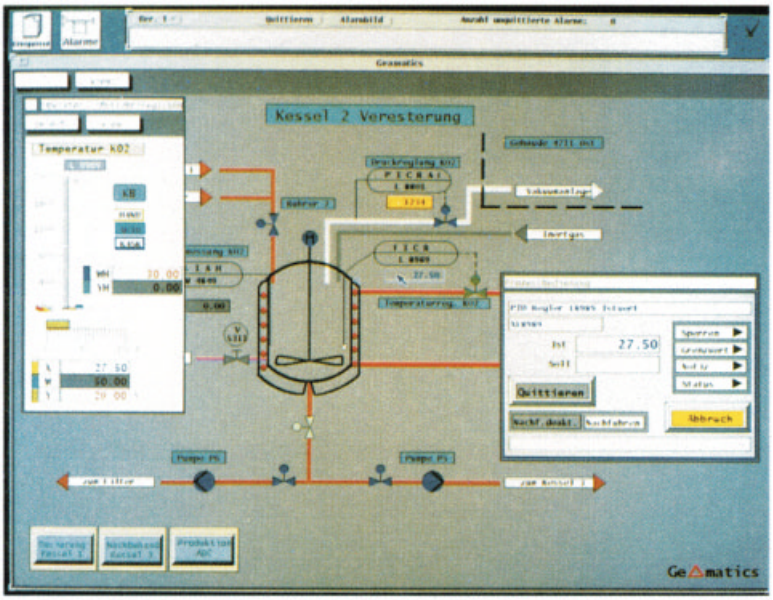

Figure 380. Windowing: standard displays and control panel in free display [716]

port). All reports can be retrieved at any time, for example, in order to reconstruct a certain plant state (Fig. 381).

Thus the windows technique pioneered by Xerox PARC, Apple Macintosh, and Microsoft has found use in color display components of process control systems.

The supplementing or replacement of the block-structure control room by the introduction of color video displays has also changed the mode of operation. In keeping with the concept of furnishing information about the process and product properties, apparatus, devices, and field-level (local) and process control systems to a central location at all times as needed for process control, additional analysis of the identification and selection of the desired information was required. The search for input techniques

that place an a priori limitation on the number of possible errors has led, for example, to the selection of well-defined screen regions by touching with a light pen, thus avoiding erroneous digital input. The touch screen is another approach to the same problem. Efforts to optimize "input technology" are related to the topics of accuracy, speed, and physiological and psychological burden on the operator. These points are discussed in [718, 719].

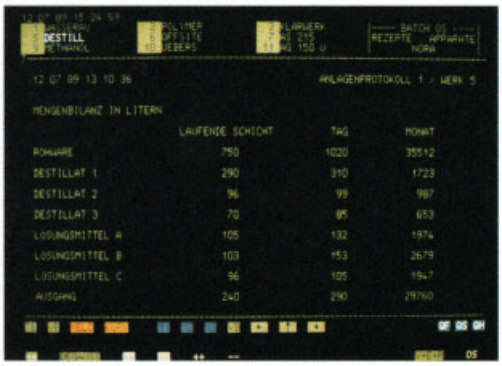

Figure 381. User report [711]

At this point, a fundamental difference between human – machine (man – machine) communication (MMC) and human – process (man – process) communication (MPC) should be emphasized. While MMC is based on a specified content and context of the information and goes on to formulate conditions such that the information flow through the human – machine interface will be suited to human capabilities, MPC is concerned with tasks arising in the course of process control and the information that must be furnished so that these tasks can be performed. From, this standpoint, the first questions raised in MPC relate to historically accreted information structures, which are largely oriented toward apparatus, devices, and control loops, and which supports the interpretation of information in precisely this context.

After a thorough analysis of the contributions made by the operator to process control (see Section 7.1, Fig. 241, and Section 11.1, Table 14) and the information that must be made available for this purpose, the design of human – process communication must take account of the "cognitive faculties" of humans, which also serve as the basis for the delivery of task-relevant information. Findings from psychology have a role to play here.

Interface design in software engineering and cognitive engineering employs an approach borrowed from cognitive psychology in which human beings and their communications capabilities are examined in terms of the "information processor" paradigm. An introduction to the principles and limitations of this approach can be found in [720].

According to the conceptual models of cognitive engineering, humans act on three planes (Fig. 382) [721]. First, from sensory afferents and associated arrays of features, signals are derived through training and directly transformed into actions (intuitive behavior). The second level of behavior is based on rules selected in accordance with a recognition procedure and the association of state and task. Finally, knowledge and experience are turned into goal-dependent decisions, which lead via a planning process to new rules of behavior. Clearly, the last two cognitive levels require a greater information-processing effort.

Any arrangement for human – process communication must therefore use an ingenious array of features and representation of the process state to provide especially good support to, for example, the "intuitive" level if the human operator is required to react quickly in the course of a monitoring task (such as to issue an alarm). In tasks having to do with quality, where links between process and product properties must be discovered and analyzed, possibly over wide spans of process sections, the cognitive faculties of the two higher levels should be supported in the model. According to ideas of cognitive engineering [722–724], this can be accomplished by providing information at various levels of detail or by using a rule-based advisory system (expert system).

How individual mental activities involved in operating and monitoring a (partially) automated technical process can be supported by appropriate software has been shown in [711]; see Figure 383.

In a paper titled "Mozart, or Our Dilemma When Dealing with Genius," Eigen [725] quotes from Eccles' work "The Human Brain" [726]. Here, a similar distinction is made between the dominant hemisphere, which controls conscious behavior in an analytical, logi-

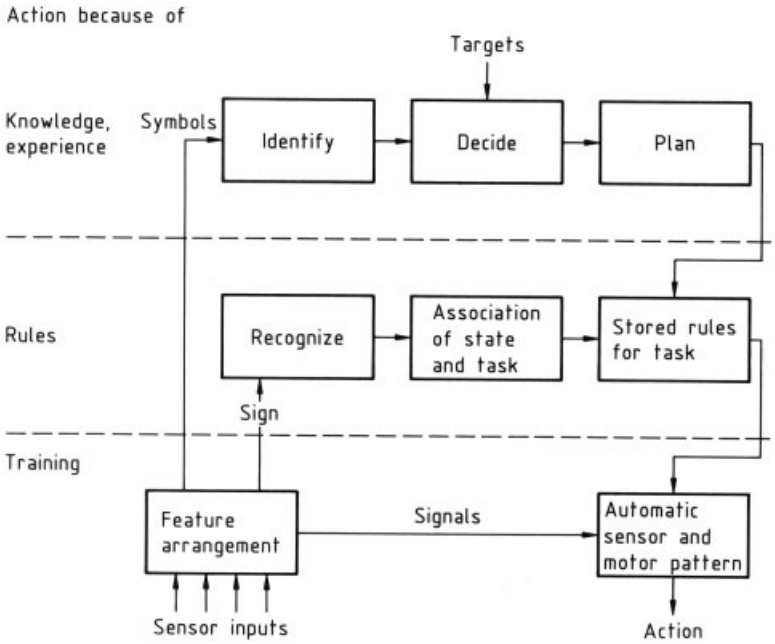

Figure 382. Levels of human behavior [721]

cal, comprehending way, and the subdominant hemisphere, which reacts in a musical, graphical, holistic way: This can be understood as implicitly relating to various human information-processing cababilities. Thus, pictorial representations, including graphics, make information comprehensible in a chiefly synthetic (holistic) manner, while texts appear to stimulate analytical, sequential thought.

Deepening the understanding of the perceptive, cognitive, and memory procedures of the human brain is the realm of psychologists. A guide to the abundant literature of psychology is given in [727].

Psychophysics deals with the quantification and scaling (see Section 2.3) of psychological meaningful phenomena. It studies the relationships between the "objective" features of the thing perceived, the underlying physiological processes, and the experience of perception. Gestalt psychology has also made contributions in this area [728]. Findings from these sciences find use today in software ergonomics among other fields [729]. Recommended intro-ductions to physiology are [730, 731]; an exposition of perceptual psychology is given in [732].

Now that magnetic-resonance and positron-emission imaging techniques have made it possible to localize functions of the human brain by functional activation of localizable glucose metabolism [733–736], neuropsychology may provide a bridge between physiology and psychology. Results from this discipline that may be significant for human – process communication are eagerly anticipated in [737].

A fundamental dispute between purely scientific-positivist and phenomenological-anthropological modes of analyzing psychological processes, in particular the concerns of perceptual psychology, was shown by Straus as early as 1936 [738]. Straus warns urgently against allowing an anthropomorphic interpretation of mechanical activities to be turned into mechanomorphic interpretations of human and animal behavior. Similar statements have been made by other members of the "holistic" school, such as Eccles, Popper, Lorenz, and Gibson [739]. They point to fundamental problems in the reductionist approach when regarded from the standpoints of neurology, philosophy, biology, and perceptual psychology (see, e.g., [726, pp. 34 ff]) and show how specific human abilities might be understood in terms of the integration of subfunctions and how the preconditions could thus be created for participation in

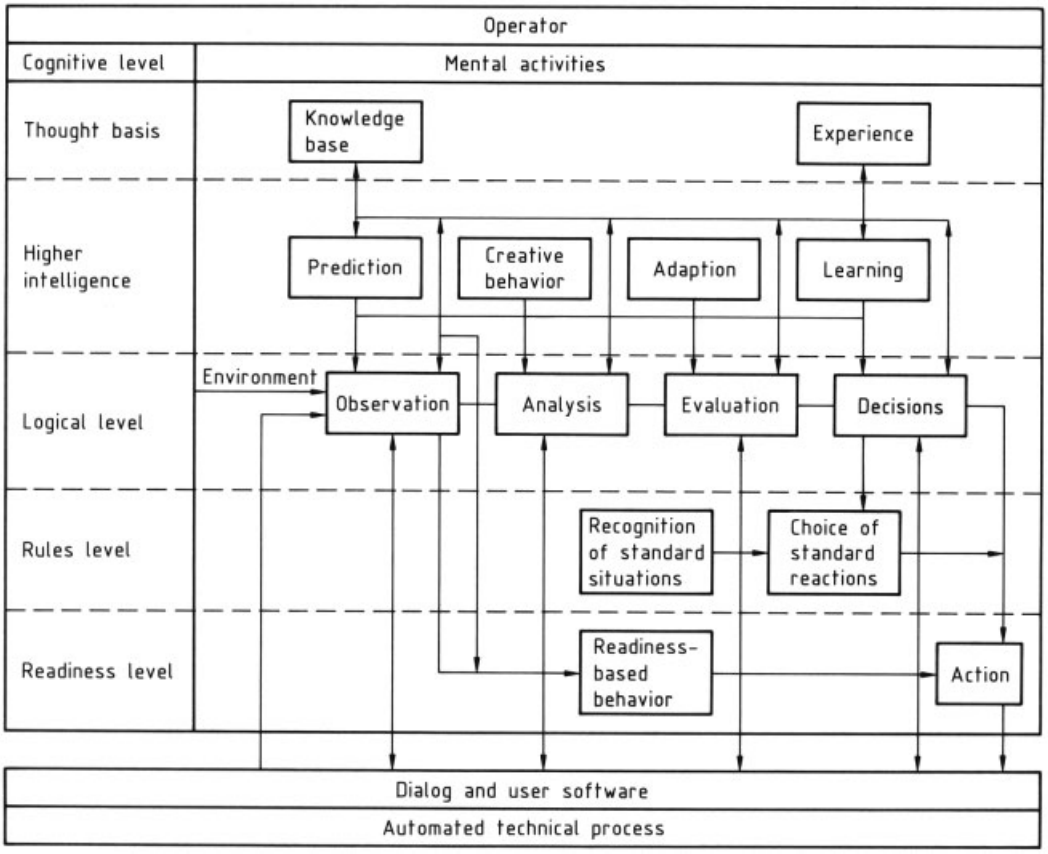

Figure 383. Support of operator's mental activities by interactive and custom software in the control and monitoring of an automated technological process [711]

supra-individual "worlds" such as culture and science (see [740, pp. 148 ff]).

The presentation of messages relating to the technical process must be oriented toward specific target groups of users. This does not simply refer to the experience, for example, that the message coding must correspond to the importance of the message itself – which is reflected in varying reaction times (Fig. 384) – or to the limited field of view of human vision (Fig. 385); such experiences must be taken into consideration when designing process control rooms and in particular when setting up the configuration of monitors [741] (see also Section 10.7).

The presentation of information must also take into account that process engineers, for example, need a different view of the process from day-shift or night-shift operating personnel, whose spectrum of tasks extends from process analysis and optimization to straight operation of the process in accordance with stated rules (Figs. 16 and 386).

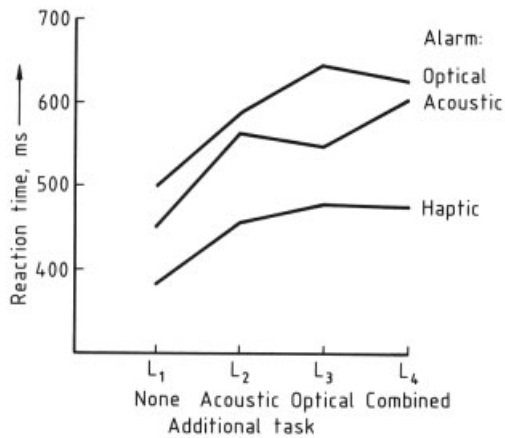

Figure 384. Reaction times to messages encoded in various ways [735]

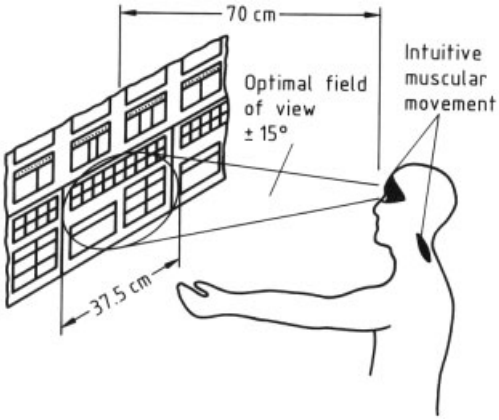

Figure 385. Field of human vision

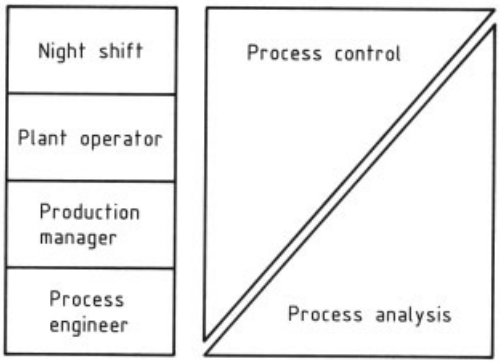

Figure 386. Process communication

Modern process control systems, with their display and control components, furnish a flexible tool with which, using the many individual process signals available, process states can be represented in an anthropotechnically and economically favorable fashion. As Figures 372–381 show, operating personnel build their own models of the technical process, based on the individual signals, so that they are then able to interpret these signals correctly and take appropriate action to control the process state.

The availability of high-performance computers suggests using the computer itself to compress individual signals into an overall view of the process – an image or, mathematically speaking, a model of the process – and conveying to the operating personnel this compressed information based on the individual signals. This would complete the transition to "information-oriented" process control engineering. In signal-oriented process control, the information philosophy was: "Acquire and present information item by item, in parallel, and always."

In information-oriented process control engineering, the philosophy can be reformulated as: "Acquire information in the smallest possible amount, but as much as necessary, and not until it is to be used, and in a form that is clearest for the operating personnel".

Thus a holistic point of view will make it possible to enhance the competence of the operating personnel with the objective of "state-oriented" human – process communication.

Information Reduction (see also Section 3.2). Complicated technical processes generate a large number of signals, which the human operator cannot take in and interpret in direct form. For example, if two dozen signals are plotted in bar-graph form, the task of deducing certain process states from certain patterns can generally not be performed by a human. Procedures of signal processing and pattern recognition must therefore be employed to reduce this complexity.

Several methods for reducing the dimensionality are known, for example, from speech processing [742]. Orthogonal transformations, the best-known of which is the Fourier transformation, play an important role in this procedure. The idea is to represent an original signal by a linear combination of sine and cosine signals differing in frequency and phase. The weighting factors in the linear combination are the new dimensions, the number of which can be reduced by leaving out, say, high or low frequencies, with no substantial effect on the signal form.

Another orthogonal transformation is the Karhunen-Loève transformation, in which sines and cosines are replaced by functions derived from the original signal. These are represented as vectors whose contributions to the total signal decrease with increasing ordinal number. By forming a linear combination of a few such vectors, the original signal can be reconstructed well; the weighting factors in the linear combination are the new transformed signals with smaller dimensionality [743, 744].

The method of factor analysis takes the large amount of possible information about the product and the process and generates a system of mutually independent new variables that can be ordered, so that it is possible to say how much

each of these new parameters contributes to the total information.

If such methods are used uncritically, the problem arises that the new dimensions may not lend themselves to ready interpretation by the user. For example, signal mean values and signal ratios are highly suitable and are also subjectively accepted as indicators.

This discussion has mentioned and briefly described just a few methods for reducing the dimension. Applying such a method also reduces the redundancy, since as a rule a redundant signal can be represented as a combination of other signals, so that recovering them does not supply any new knowledge about the process state.

This situation typically occurs in an overinstrumented system where there are, for examaple, more sensors than are required for an unambiguous determination of the state. However, overinstrumentation also makes it possible to carry out signal reconstruction for signal sources that may have failed because of some technical defect (redundant instrumentation).

Even though there may be substantial time shifts between signals, especially in extended technical processes, there are deterministic relationships between them (e.g., measurements upstream and downstream of a lag element). In many cases, such shifts can be detected by cross-correlation analysis, but in general the result for a signal source is that not only the measurement at a time T but also some number N of measurements obtained at earlier times must be taken into consideration (see Fig. 387).

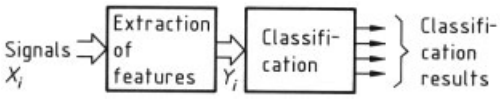

Figure 387. Schematic diagram of pattern recognition [735]

Pattern-recognition techniques afford still more support to the operator. For example, hazardous states in a process can be detected on the basis of a pattern in the signal. Figure 387 illustrates the principle of pattern-recognition systems. The original signals are first subjected to feature extraction. This is analogous to the dimension-reduction operation described above; the smallest possible number of features, and hence highly informative features, are derived from the original signals (see also Chap. 2).

This means that the subsequent classification can be done in a space of few dimensions. Classification involves forming regions in the N-dimensional feature space, under the assumption that patterns lying close together in feature space belong to the same class. In a two-dimensional feature space, such regions can be delimited by, for example, a chain of line segments.

If these lines are drawn automatically, possibly with operator interaction, the classifier is said to "learn." Under certain circumstances, depending on operator behavior, such a classifier may be able to adapt to changed process situations.

By using cluster analysis [745, 746], class formation can be carried out automatically without external supervision. Each new pattern is assigned to the region (cluster) to which it is closest. Cluster analysis can be used, for example, to find automatically and subsequently identify particularly frequent and typical process states. In future, this type of procedure will undoubtedly provide significant aid to the operator (see also Fig. 422).

These order-reduction procedures make it possible to represent process states in three-dimensional form, making it possible to visualize the product or process space (Fig. 388).

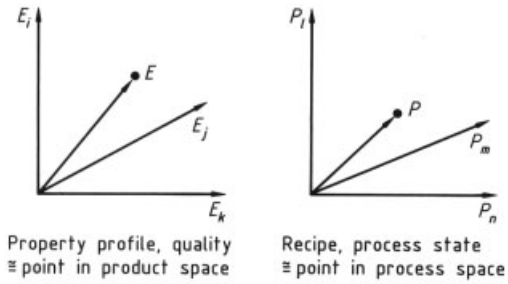

Figure 388. Analogy between product and process spaces [743, 744]

A similar mode of analysis is arrived at if the methods of similarity theory are applied to a signal pool [685, 686]. The direct presentation of the dimensioned X-quantities in X-space with X-relationships is replaced by the dimensionless presentation of π-quantities in π-space with π-relationships.

The presentation of state spaces, as a prerequisite for "state-oriented" human – process communication [711, 747, 748], requires procedures

by which relationships between the process and product properties can be obtained from the time dependences of these properties (as conventionally supplied by the instruments currently used for process monitoring).

Thus the problem is to derive the function $X_i(X_j)$ from the time dependences $X_i(t)$ and $X_j(t)$ of the measurements.

It is straightforward to plot the last N pairs of values (X_i, X_j) acquired to give a two-dimensional scatter diagram in coordinates X_j versus X_i. This representation may actually make it possible to derive the desired relationship. Under some circumstances, regression methods can also be used; for example, the mean and dispersion of X_i can be obtained for each X_j product. Utmost care is advisable, however, because no a priori prediction can be made as to the form of the curve.

Two further difficult time problems must be solved:

- In general, the function $X_i(X_j)$ is not stationary. Otherwise it could be determined once for all and would have no predictive power for the operator. The time constant that describes the behavior governs the size of the "window" over which the function $X_i(X_j)$ can be determined. The window may be fixed (a new measurement vanishes every time a new scan takes place) or decaying (e.g., exponential); the computing effort required is then less. The operator receives a picture of the function that shows a continuous slight variation.
- X_i and X_j may be very closely related but with a (possibly long) time shift or bias. If the X_i and X_j measured at the same time are plotted, the result may have no predictive power at all (Fig. 389).

The time shift can be identified by cross-correlation analysis, after which measured values $X_i(t)$ and $X_j(t+\Delta t)$ can be paired. It is also possible to create a three-dimensional representation with the time offset as the third dimension. The resulting "mountain range" would reveal where the dependence was most distinct. Changes in time offset, in any case, provide an important indicator of the process state (see also Section 3.2).

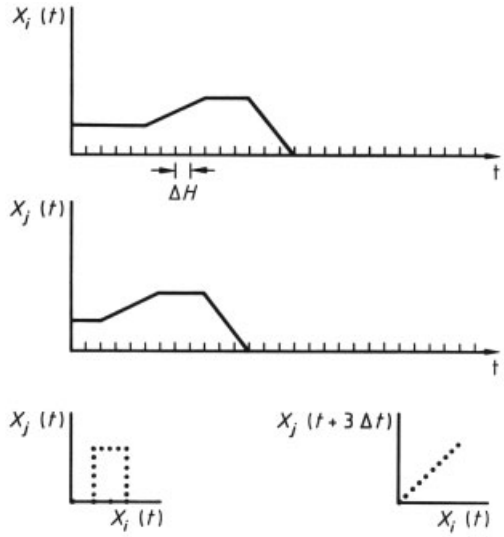

Figure 389. Effect of time delays

Information Storage under Real-Time Conditions. In future human – process communication systems, a major role will be played by the computer's internal data model [749–752]. This model must contain all data acquired and stored during the operation of the process control system (see Chaps. 2, 9, and particularly 8). Such a procedure will require data storage under real-time conditions, and this objective has given rise to the conception and development of real-time databases. The following typical features distinguish these from classical databases:

- They contain instantaneous data generated and required in real-time process operation. In addition, there are data to be archived for infrequent later interpretation. These archival data will replace the process logger charts common today; the large bodies of data acquired will typically be stored on video-disk memory devices. A real-time database, in the narrow sense, is concerned with the management of instantaneous data.
- These databases are very small in comparison with commercial databases.
- Reaction-time requirements are stringent, both for the storage of data (which must always be carried out such that a consistent data model is preserved) and for its retrieval (since such requests may be made by both the process and the operator). Reaction times in the millisecond range and very short time inter-

vals for insuring data consistency are needed here.

In the implementation of real-time databases, the result is extensive utilization of main memory.

The use of a real-time database in human–process communication systems will lead to major simplifications in programming. With the software tools belonging to a data base, individual queries can easily be handled; support of a query language with graphical output resources will simplify interaction with the operator.

Graphical Technology. EIGEN [685] has shown that pictorial representations (graphics) make it easier to comprehend information than does text.

The use of computer graphics in mathematics and the natural sciences is surveyed in [753]. Graphical kernel systems such as GKS/GKS-3D or PHIGS/PHIGS$^+$ are device-independent basic software systems for graphical programming; implemented on workstations, they increasingly form the basis for CAD tools [752,753]. Multidimensional state-space representation, in particular, requires images with almost photographic realism; today these can be generated with efficient algorithms and data structures or with ray-optics techniques from physics (ray tracing) [754].

Finally, computer graphics can handle demanding visualization tasks only with the aid of knowledge-based methods. Conversely, as discussed in Chapter 13, increasingly complex artificial-intelligence (AI) systems depend on graphical visualization aids to create viewable images of models and states. As a result, the disciplines of graphical data processing and artificial intelligence are becoming more and more interwoven, as is reported in [755].

What follows is a discussion of the graphics techniques available in current process monitoring and control systems. All graphical output devices now in use for human–process communication are color displays. Graphics systems in first-generation process monitoring and control systems feature medium resolution (roughly equivalent to the broadcast television standard) and permit the semigraphical representation of process states and process flowsheets. "Semigraphical" means that the screen is subdivided into predefined fields containing, for example, 8×8 pixels, in each of which alphanumeric characters or graphical image elements can be placed. A character generator driven in an 8-bit code contains the bit pattern for the characters and the selected graphical elements.

The refresh memories in modern systems are too large to be displayed on the screen. Thus the process image can be moved back and forth so that the screen appears to represent a segment of a large landscape (rolling image).

On the input side, the keyboard is complemented by a variety of pointing devices. These facilitate both graphical input and selection from menus shown on the screen. Typical devices for this purpose include trackballs; joysticks; touch screens, in which various physical methods are used to detect the location of the finger touching the screen; and light pens.

Each of these systems gives x/y coordinates on the screen, which are accepted as inputs to the operator software.

The trend in development of graphical devices for display and control components is toward pixel-oriented color displays. These offer much better resolution as well as physical parameters (19″ and 21″ screens with resolutions of 1280×1024 dots, displaying 256 colors simultaneously out of 16.7 million). There is a trend toward the standardization of hardware and software. The devices are based on PCs or workstations with RISC processors and run under operating systems such as DOS, UNIX, or their derivatives.

The principle embodied in these user interfaces is splitting into components that are relevant for the user interface and those that are specific to the processing tasks in an application [756]. The distinguishing feature from the user's standpoint is the structuring of the entire screen area in movable, variable-size, iconizable subregions or windows, which can accept user input and display representations independently of one another. Screen windowing and network-wide window management tasks are performed by basic windowing systems, which are included in the operating system or added on as utility libraries (e.g., Xd Windows [757]). "Network-wide" means that the user, from any network node, can interact with an application running on any other computer(s) in the network. One possible principle for such interaction is the client-server principle. Above the window manage-

ment level are graphical user interfaces (OSF-Motif, Open Look in the UNIX environment, MS Windows or Presentation Manager in the DOS environment).

The problems relating to control and monitoring components (see Chap. 7) have now impelled a number of manufacturers to create their own process visualization features for certain information-processing systems that have no display and control components of their own, especially systems based on stored-program controllers. One of the main components of these visualization devices, which can simultaneously carry out control functions, is a database system permanently linked to the process and hence running in real time. Because only a few of the user interfaces based on the well-known operating systems have been standardized (e.g., X-Motiv in the UNIX realm or Windows in the DOS environment), a number of very diverse "artistically" designed tools have appeared in this area. These require an extremely critical examination before they are put to use.

Several systems based on this architecture have been offered especially for process visualization (e.g., InTouch [758], Factory Link [759], APROL [760], Dynavis-X [761] COROS-LS-B [762]). Some process monitoring and control systems have integrated display and control components (e.g., SATTLINE [763], Viewstar 750 [764], Advant-Station [766]). The discussion that follows introduces special characteristics of these systems without describing individual products.

The functionality of these systems is outstanding because elementary, interactive, and scalable graphics objects are available. System resources, such as fonts, fill patterns, and color tables, can be utilized, and raster or bitmap graphics, either scanned in or generated by other graphics software, can be incorporated. With the aid of these design resources, predefined graphical input and output objects specific to process control engineering can be created and placed in libraries for later use. These include virtual keys, switch panels, slider controls, analog displays, and so forth. Another aspect of the graphical functionality is the "logical zoom" function, in which excerpts of images can be displayed in varying sizes and levels of detail and information is added as the zoom level increases [761].

Visualization systems based on graphical user interfaces are further characterized by expanded capabilities for defining the interaction (dialog) between user and process. These features include the interpretation of user inputs, the execution of appropriate commands, the creation of graphical objects when certain events take place, and so on. Along with well-known event-oriented and data-oriented concepts, complex control structures made up of command sequences, branches, and loops can be initiated, windows can be opened and closed, and graphical objects can be created and deleted. Flexible, state-dependent image structures can be built up in this way and need not be oriented to fixed image hierarchies.

Another characteristic is that these systems can be set up to utilize open system services present in any hardware and software platform so as to expand the process visualization functionality. Such mechanisms include system services for data exchange, command transmission, and simple integration of application-specific software upgrades. Examples are Dynamic Data Exchange, the Dynamic Link Library concept and multimedia add-ons under Windows and (in part) OS/2, database query languages such as SQL, and pipe mechanisms under UNIX. In this way, open visualization systems offer better integration capabilities than the display and control components in first-generation process control systems. One example is the connection of the visualization system to a spreadsheet analysis in which process data can be linked with mathematical functions on-line. Unified operator control concepts in a user interface come into play here because different programs then present themselves to the user with the same appearance mode of interaction (dialog). The utilization of system services also includes the possibility of creating help systems and defining macros.

For the designer of a process visualization utility, ease of use dictates that the capabilities of the graphical user interface be employed in design. "Toolkits" or interface modules are not adequate for this purpose. They merely provide development interfaces at the programming-language level and require much specialized knowledge. Abstraction levels suitable for the design of display and control components offer classes of tools, such as CAD-like editors for graphical interfaces in combination with editors

for formulating "interactive" control, so that interface objects can be selected, constructed, furnished with attributes, and linked to user actions and process data (interface builder and user interface management systems [756]). Resources taken over from the CAD world, such as traps in conjunction with underlaid rasters, further facilitate the work.

From the process control engineering standpoint, the high flexibility of visualization systems also holds some risks. In the design of process visualization, the assignment of appropriate parameters should be used to restrict the user's options in interactive mode. For example, it should not be possible to place other windows on top of alarm windows. Similarly, a basic subdivision of the screen into fixed-position, fixed-size summary and detail windows should be permitted. Application-specific software add-ons and process data manipulations in other software packages may prove far more successful if access rights questions are not handled as functions of user or process state. In "closed" process monitoring and control systems, such mechanisms are made available as system services.

Of particular importance in process control engineering is the seamless integration of visualization design into the engineering process as a whole. This includes, for example, the linking of process visualization and dynamic process data generated in the design of process-level (local) components. Capabilities should be provided that make available the engineering data which indicates which process data are used in which visualization functions (design database).

Anticipated developments in graphics hardware for human – process communication are discussed below. Modern visualization systems intended for use as display and control components in process control systems are based on 2D graphics. Three-dimensional images can also be displayed on a 2D screen. The representation of the Space Shuttle in Figure 390 shows how hidden lines and surfaces can be removed from the three-dimensional picture by using computer graphics. However, mapping a three-dimensional model onto a two-dimensional surface (well-defined portion of the image, well-defined angle of view) places a considerable load on the computer that constructs the individual points of the image. Accordingly, 3D graphics on a 2D screen is severely limited in speed. Graphical representation capabilities in process control systems with modern 2D graphics systems are illustrated in Figure 376.

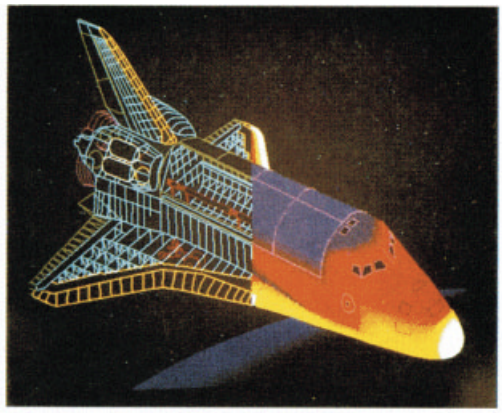

Figure 390. Three-dimensional representation of the Space Shuttle on a two-dimensional screen [765]

The performance of graphics systems is improving continuously. Graphics systems from the CAD realm and scientific data processing already hold an internal 3D representation, which they map onto the selected viewport. This implies a substantially larger graphics memory (a complex volume model, for example, requires 500 000 vectors, each represented by several bytes, so that the total memory demand is several megabytes). The enormously high performance required of graphics systems in these applications stem from close-to-reality graphical representation combined with animation capabilities. An example is the simulation of poor visibility in the training of aircraft pilots [767]. Graphics systems, usually implemented on a dedicated board or even as multiprocessor systems, can now plot up to 10^6 3D vectors per second. The workstation CPU and the graphics processors split the work up in a rendering pipeline [768]. Data transmission over an internal bus is a bottleneck, as in many other situations. Modern graphics workstations have main memory capacities of 16 – 128 MB and even more. The capabilities of modern high-performance graphics hold the promise of real-time visualization of complex data collections in n dimensions, a procedure that has now become the state of the art in scientific data processing. Isolines, textures, and other resources are used along with color to handle the n-th dimension or variant in graphical repre-

sentations. Photographs of glowing metal slabs or bars provide an example. From the color of a glowing piece of metal, it is easy to determine the temperatures of the individual zones. The color scale runs from white (highest temperature) to black (lowest temperature). In Figure 390, color is used as the fourth dimension in visualizing the temperature distribution of the Space Shuttle. Color can also be used to visualize other quantities such as height, material density, etc. The use of color as a supplementary element in graphics is advantageous chiefly in raster graphics, since colors can be gradated extremely finely in this kind of representation (high color resolution).

There have also been new developments in pointing devices. The mouse is now commonplace in office automation. In conjunction with the use of workstations as display and control components, it will be used in the process control room environment as well. Futuristic concepts of interactive devices go yet further: An accelerometer on the fingertip makes it possible to determine the direction in which the finger is pointing, and observation of the pupil tells where the operator is gazing. Electronic gloves equipped with sensors (datagloves) are already in use, chiefly for research and in virtual reality systems [769]. The first commercial applications have come about in the CAD field. In the n-dimensional data or feature space concepts discussed above, an operator using such an interactive tool can manipulate data and freely select a point for viewing images, thus wandering through the feature space.

Multimedia Techniques. Control and monitoring functions in process control systems have recently been expanded by techniques broadly referred to as multimedia. STEINMETZ has given a qualitative, application-neutral definition of this term [770]. The types of media supported and the options for media processing on a computer system are key features. Multimedia systems must be able to process discrete media (e.g., text and graphics) and continuous media (e.g., audio and video) independently of one another. The computer is extending its capacities to every medium by which information can be propagated and presented. The central points in multimedia are integration of media, range of media, distribution of multimedia information over a computer network, and user interaction with a multimedia system.

From the standpoint of information technology, crucial tasks for media integration are to create digital representations of the media, to develop suitable storage media, and to augment communications networks. The main problem is the high flow rate of time-critical multimedia data. For example, the digitization of a video sequence at normal resolution requires data rates of over 20 MB/s. Key technologies for implementation are digital signal processors for data compression, optical storage media, and high-data-rate networks.

Ultimately, multimedia systems make sense only if efficient system management insures the on-time processing of, in particular, continuous media. Requirements in this area apply to real-time processing in multimedia systems, specific service reservation, and the synchronization of linked information units or information streams relating to discrete and continuous media. Finally, the developers and users of a multimedia system should be able to operate in the way they have become accustomed to with discrete media. This means that multimedia functionality must be properly abstracted in development systems, user interfaces, and database systems. The object-oriented approach finds use here. Examples of the abstraction of multimedia capabilities can be found in multimedia add-ons to window-oriented graphical user interfaces running on personal computers and workstations. The desktop metaphor has been expanded to include, for example, a telephone, a dictating machine, and a video conference. These points will not be discussed at length here, nor will the associated problems, in particular the management of large bodies of data, diverse data formats, links between distinct media objects, and content-oriented searching. Reference should be made to the specialist literature [770–775]. In what follows, special requirements on the use of multimedia are presented, along with possible application areas in process control engineering.

Real-time requirements in process control mean that the correlation of data in a variety of media – especially continuous media – to objects, events, states, alarms, etc., is a central problem. Furthermore, because events in technical processes occur in stochastic fashion, it is impossible to predict at what time which multi-

media input/output devices need to be available. Conflicts, such as when several audio sequences must be output at the same time, must not occur, although it is entirely possible for the operator to be simultaneously occupied with several distinct tasks. These points can be traced back to the problem of multimedia resource management. A concept for the integration of multimedia in an information and diagnostic system for use in process control is proposed in [776]. A major role is played by the presentation system, which is responsible for synchronized input and output matched to system and user requirements.

Already common in the process industries is the use of cameras and microphones to monitor machinery, plant sections, and products. Generally, this involves self-sufficient analog monitoring systems with no connection to the process monitoring and control system, or else hybrid systems in which analog recording devices are controlled by computers through interfaces. In both cases, the data stream remains unprocessed and isolated from other information-processing functions. In digital multimedia systems of the future, integrated processing of all media will be possible. First, other information can be appended to audio and video signals (timestamp, links to process data, comments, etc.). The result will be new capabilities for messaging, reporting, and archiving systems. Second, digital image processing will offer new possibilities for human – process communication. Digital image processing is already employed in industrial robots (machine vision), in computer-aided quality inspection of finished parts, and in the monitoring of machinery [777, 778]. In human – process communication, digital image processing, perhaps in conjunction with CCD cameras, will provide a route to alternative methods by which product and process states can be determined and classified and the results presented in a visually informative way. An example is the assessment of temperature profiles in metal working by determining gray-scale distributions [779]. Finally, the operations of digital image processing have the goal of presenting the operator with visual information in an appropriate and easily remembered form. A closely related topic is pattern recognition, a set of techniques which permit the identification and grouping of logically like image contents and thus the identification of single objects [780].

In addition to the question of possible applications for multimedia systems, another point of interest for human – process communication is how these systems can be put to use. First of all, the modality – i.e., the way in which the operator's sensory organs take part in the interaction – of the operator – system interaction is expaned, for example by speech, but also by futuristic concepts such as the interpretation of body language [781, 782]. The objective is to make user interfaces more "natural" by letting them use a broader sensory spectrum. Information acquisition can be facilitated and supported through input and output in the original medium without transformation, as in voice mail, or through multimodal presentation of information in more than one medium. An application is the audio-visual notebook, in which video records with added comments provide information about special events.

A second system-related aspect is "active" user behavior. The pertinent concept here is hypertext or hypermedia [783], which is a concept for organizing and actively "exploring" information.

Essentially, a hypertext or hypermedia system has two components: networks consisting of elementary information units (nodes) and flexible, directed references (links), and tools for creating these networks, reading them, and so forth. Hypermedia differs from hypertext in that the nodes contain multimedia information. The user wanders through the network in survey fashion (browsing) or with a definite destination (navigation). The contents of nodes and links need to be static. The concept also includes executable code as well as the on-line determination of the destination of dynamic links as an aid to navigation. More complex structures can be generated by aggregating networks. For a more in-depth treatment and discussion of further aspects, such as programming and tool/system support, see [729, 784, 785].

The first applications of hypertext concepts came in the organization of documents in "information systems" (information kiosks) and in education. Hypermedia systems have found industrial use as technical information systems for commissioning, repair, and service documentation [786]. In technical processes, the operator of the future will obtain required knowledge, such as safety aspects, through information systems

of this kind. At present, work in this area is being carried on in a EUREKA project [787]. Hypertext concepts in human – process communication can also be put to work in on-line operator support or operator training systems, in particular systems that will guide the unpracticed operator to the correct control actions [788]. Similar considerations relating to ship steering are presented in [789]. Closely related to hypermedia concepts in process control is the linkage to knowledge-based methods; hypermedia systems are used as user interfaces for diagnostic systems, since operator behavior can be better supported with them [790] (see also Chap. 13).

Trend Forecasts. Instances of information-oriented process control engineering can be found today in aeronautics, the power-plant industry, and automaking.

Flying involves very complicated changes in attitude resulting from course changes. In modern aircraft, these trimming or dressing actions are performed for the pilot by a computer-supported control system. Thus the pilot can fully concentrate on determining the course, using a joystick to specify course changes to the autopilot, which in turn generates the control instructions needed to alter the attitude. In order to reduce further the cognitive burden on the pilot, aircraft such as the Airbus A310 include a display showing the instantaneous position with the planned course and a proposed course deviation. The justification for the deviation is incorporated into the display; for example, if it is an area of bad weather, a weather radar display is superimposed (Fig. 391) [791].

Similar developments are under way in automotive engineering. Navigation devices use graphics to display the position and the proposed course to the desired destination, taking account of current traffic conditions [792]. Required maintenance is determined by using models and reported to the driver; an example is the service-interval display developed by BMW [793].

Another example of information-oriented process communication is a display showing ignition angle plots for a digital motor ignition system [794].

In power-plant control engineering, the status of a boiler or turbine is displayed as an instantaneous operating point on a plot of characteristic curves. Any change in operating point can be effected by using a crosshair pointer to specify a new nominal operating point. Individual setpoints are then obtained from a process model [795].

New approaches to information-oriented process control engineering are also seen in the chemical industry. In general terms, product properties E can be represented as functions g of process properties P. This procedure is readily illustrated in process space (Fig. 392) [796]. For a certain product property E_0, there exists a function $g*$ of the process properties P_i and P_j, as shown by the curve. The thicker portion of the line characterizes the optimal re-

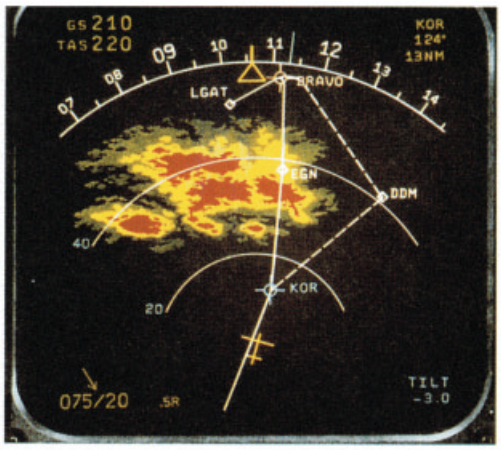

Figure 391. Airbus A310, navigational display [791]

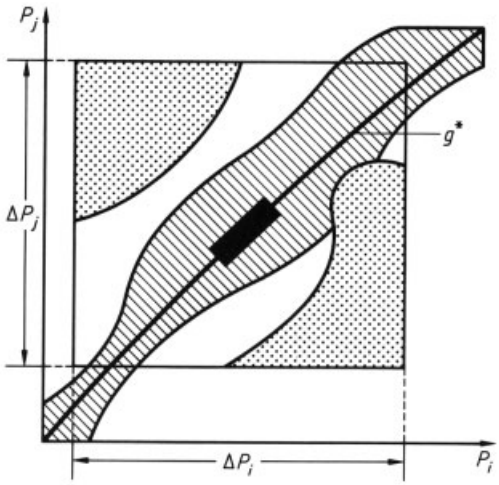

Figure 392. Process space (two-dimensional)

gion for the process, characterized by compliance with additional constraints. If certain tolerances ($E_0 \pm \Delta E_0$) are set when the values of the product properties are established, then in the process-space representation this corresponds to a certain variation range of the process properties (shaded region). In a specified technical apparatus, of course, the process properties can be varied only within certain ranges ($\Delta P_i, \Delta P_j$); the rectangle shown in the figure is obtained as a result. Safety and environmental restrictions narrow the acceptable region of the process space still further, leading to the areas enclosed by dashed lines in the figure.

Figure 393 shows another way of thinking about a real process. In process space (normally the coordinates are configuration parameters determined by order reduction), the hatched region shows the normal production condition for a certain product. The white rectangle located on this is the economically optimum operating region. The variation in the process properties allowed by a specified apparatus is indicated by the outline of a parallelepiped. The upper and lower gray areas show the restrictions imposed for safety and environmental reasons.

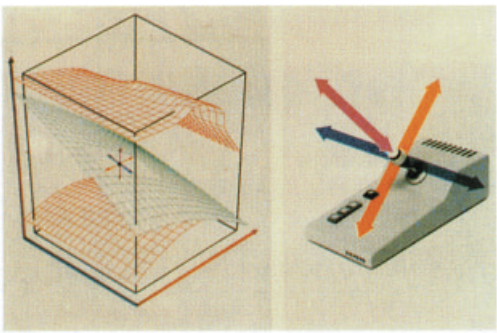

Figure 393. Process control with the joystick

Figure 393 also shows how the operating point on the screen can be moved into the target region, for example with a joystick. The change so specified is then automatically transformed to setpoint (control) signals [797].

Tracking the operating points in a process as just described indicates the distance from critical regions and, eventually, may also make it possible to initiate prompt computer-aided action to remedy disturbances. The analytical preconditions for doing this have long been known from fault-tree analysis.

Another type of process communication can be implemented with the help of Wolf diagrams [798].

These examples from a variety of technical application areas suffice to illustrate the present and foreseeable capabilities of information-oriented communication.

Information-oriented process control engineering will relieve operating personnel of some burden, help them avoid errors in disturbance situations, and promote the optimization of processes through improved modeling. This innovation will also make it easier to produce better and more consistent quality in accordance with specified nominal profiles. The operating personnel will be able to put more effort into studying the process as a basis for modeling. Furthermore, it will be possible to employ operating personnel with a broader spectrum of skills, and at the same time the monotony of continuously watching single signals will be broken up. Finally, the early detection of incipient transitions to undesirable operating states will lead to improved operational safety and reliability.

These are all good reasons for supporting information-oriented process control engineering. "Signal-oriented" process control engineering, which recognized only human – machine and instrument – machine modes of communication, will be replaced by "information-oriented" process control engineering, which includes the concept of human – process communication.

11.3. Process Analysis and Process Optimization

In 1984, Blass [799] proposed a method that requires a methodically structured procedure in the development of complex systems so that a clear, comprehensible framework can be obtained for the application of specialist knowledge and creativity and the range of possible approaches to finding a strong solution can be developed and exhausted. The method of choice should always be a systems approach, which furnishes methods, principles, and proven resources in a relatively abstract and easily interpreted fashion, in the form of the system concept, the systems procedure, and project management.

Nevertheless, there are still processes whose development has not had the benefit of the

systems approach. Thus, retrospective methods were sought for determining the validity of existing processes so that they can be optimized.

An example is the electrophotographic reproduction of images, for which BIXBY [800] and BICKMORE [801] found an approach that involved breaking complex processes down into steps and interpreting these in physical, chemical, and process-engineering terms.

Figure 394 shows how the transfer process can be analyzed into single, physically comprehensible steps and how information can be gained in this way about the overall process.

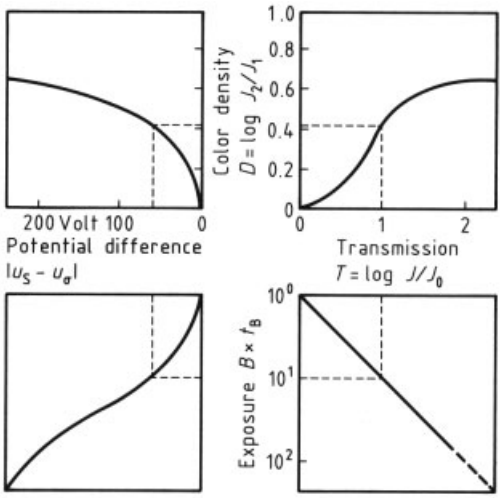

Figure 394. Analysis of the electrophotographic copying process into steps [800, 701]

This representation, often called a Jones diagram [802], is an effective way of taking individual transformation operations into account and converting object information into pictorial form [803].

Transfer Theory and Transformation Principle (see Chap. 2). Any optical copying process has the task of transforming the optical density distribution of the object $T(x)$ to an optical density distribution of the image $D(x)$. This process is described by transfer theory. Let the object $T(x)$ be expressed in terms of spatial frequencies R by the Fourier transform method:

$$T(x) = \int_R h(R)\, \mathrm{e}^{2iRx} \mathrm{d}R \qquad (11.3.1)$$

The frequency R denotes lattice periods per unit length. The image $D(x)$ is then represented by

$$D(x) = \int_R K(R)\, h(R)\, \mathrm{e}^{2iRx} \mathrm{d}R \qquad (11.3.2)$$

where $K(R)$ is the modulation transfer function (MTF), generally a complex-valued function. Its modulus describes the contrast with which the sublattice of frequency R is transferred from the object to the image.

The advantages of formulating the imaging process in this way is that the transfer fucntion collates all the properties of the transfer system in a single function; by multiplication of further transfer functions, processes can be represented in simple form and complicated processes can be broken down into steps.

At a high level of abstraction, profiles of product properties (Fig. 360) can be compared to optical density (blackening) distributions. Each process element then indicates (as discussed in Chap. 2) how each element of the product property profile is transferred to a new element, namely a property element of the end product, by a process step or process segment.

The analogy to the concepts advanced by GILLES [684, 804] and POLKE [805] is then logical.

In this relatively simple procedure, in which the modulation transfer function plays an effective modeling role, it was possible to describe and analyze the individual physical operations in the complex copying process. Later, the term "process analysis" was introduced [806] for quite different processes, such as the vulcanization of rubber. Existing processes were examined, from starting products through all process steps to end products, including the operational logistics, technical equipment, and process control engineering.

The entire process, the plant, and also the operational surroundings are thus examined with the objectives of separating dependent from independent variables; identifying possible improvements on the basis of the relations found; and devising and utilizing new – and possibly unconventional – solution approaches.

Figure 395 shows the material and data flows in a process, and Figure 396 shows the same relationships for the example of a vulcanization process. As early as 1972, a "process indicator," a

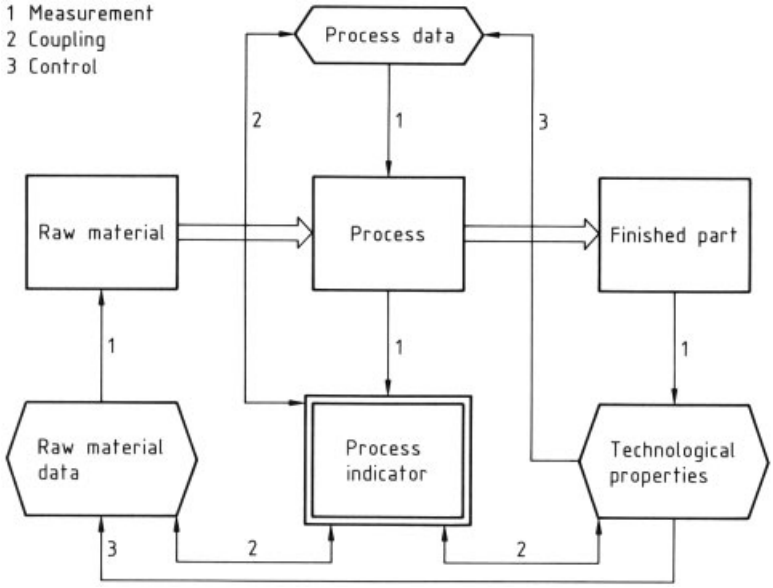

Figure 395. Process analysis (material and data flows)

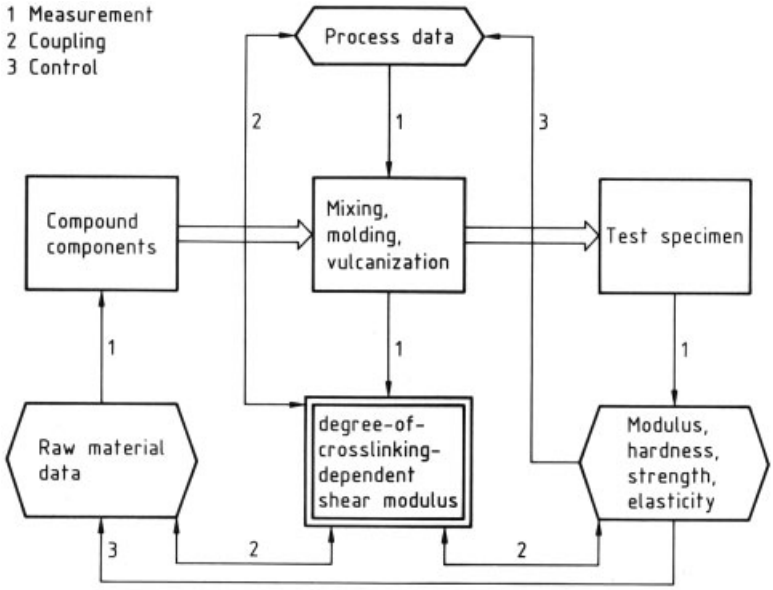

Figure 396. Process analysis (rubber vulcanization)

precursor of the model-aided measurement technique [807], was in use.

When there are obvious weak points in the operation, improvements are often carried out by taking single actions based on day-to-day experience. Such improvements have to do with the optimization of a unit operation, a plant section, or a single apparatus. Along with the advantage of solving the problem relatively quickly, this procedure has the drawback that the real cause of the problem remains undiscovered and substantial potential for improvement remains unexploited. As a rule, therefore, process analysis must always be performed from a holistic point of view so that improvements throughout the process and the plant can be effected. Finally,

process analysis is a tributary activity of process development (see Fig. 416).

The first step in a process analysis is to break down the overall process into individual process elements in such a way that the product property changes brought about by each element are covered by the phase model described in Chapter 2 (see Fig. 33).

As stated above, the transformation principle of Chapter 2 is used extensively (see also Fig. 417).

These relationships were also pointed out by Rumpf, though he used a different vocabulary, speaking of property functions in the course of the process [804].

This fundamental insight into the holistic analysis of a complete process is reinforced when process control systems are employed; very much more than previously, these presuppose a thorough knowledge of the process. This point was made in 1979 in the documentation of a correspondence course published by the Chamber of Engineering of the German Democratic Republic [808]. The form description of the dependent variables in a production process, of the product properties, as a consequence of the variation of the independent variables, of the process properties, has since come to be referred to as modeling (see also Chap. 3).

Here, the problems of process analysis are treated from the viewpoint of the process control engineer, who must acquire knowledge about the controlled process from the process engineer, the designer, the equipment fabricator, and others; that is, from the people who have the main responsibility for designing the apparatus, the facilities, and the processes to be run there, and whose work therefore must precede as well as supplement that of the process control engineer. This applies, in particular, to the analysis and description of processes in a form that is usable by the process control engineer and accessible to his methods and procedures.

The result of such a process analysis with a process control engineering orientation, however, not only provides the process control engineer with information about the process but also provides the process specialist with pointers on the active design of the process. This specialist is then better able to develop the process from the outset in such a way that it lends itself to the process control engineer's approach. This will greatly facilitate the design of the process control system. The outcome of such a process analysis can also be a considerable aid in the commissioning and maintenance of the plant [808].

The restriction to a process-control orientation of the process analysis has the result that the informational context becomes a key factor in the analysis. The work of Blass [799] has set the direction for the overall aspect of process analysis, which must always be kept in mind even when dealing with details. Together with presentations at the annual conferences of the VDI Association for Process Engineering and Chemical Engineering (GVC), Blass' publications provide an ever-fresh source of new methodological and technical information [809]. Blass also stresses the need for heuristic procedures, especially in engineering and scientific fields.

For this reason, what follows is a presentation of one such method that has been tested in practice, both for process analysis and for process development [683, 810].

Under certain conditions, product properties and process properties can be plotted as points in appropriately defined mathematical spaces. Some mappings between these spaces can be interpreted as production models that permit the formal mathematical description of product development in terms of process variation. There are tools, many of them tested in the field, for implementing this generally valid concept [683].

A company that wishes to succeed in the market must adapt the quality of its products to the requirements of the market. This adaptation can be interpreted as a measurement and decision problem.

The properties of the products produced by the company and required by the market should be quantifiable and their relative separation should be measurable. In principle, this objective can be achieved by reducing all "technological" properties to physical ones [806]. The physical properties assigned to a certain product are then interpretable as a point in a metric space. The nearness of the product quality to a required level can be objectivized in this way. The process of deciding whether product or process development should be carried out thus becomes transparent.

The question with what resources a quality improvement can be effected is first of all a question of operational capabilities: With what

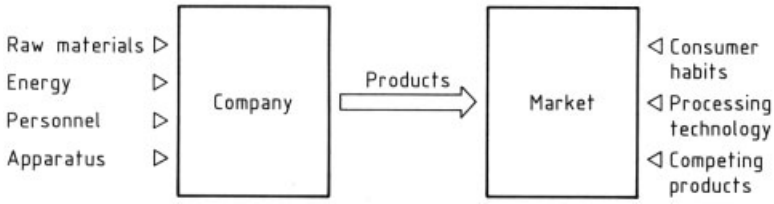

Figure 397. Factors affecting product quality (left) and market requirements (right)

processes are the current products made? Will changes of recipe in the context of existing processes result in the desired quality changes, or will process modifications or even new process development have to be carried out.

An analytical approach here affords some help in ordering the operational reaction possibilities. If a process is represented in terms of the set of assigned and independently variable process properties, then in analogy to the product space these span the process space, in which a metric is defined (provided the properties are measurable). A recipe is a certain expression of all the properties characterizing the process and can thus be interpreted as a point in this space (process state; see Fig. 388).

Now product development can be carried out in a controllable manner if it is possible to construct a mapping that assigns points in process space (recipes) to points in product space (product qualities). It is shown how the relationships between these spaces must be utilized in order to achieve product development through process variation.

It is further shown that process optimization, viewed from the standpoint just developed, generally reduces to a variational problem with constraints. The mapping from process to product space describes, in model terms, the interaction between plant and product during the production process. Accordingly, this mapping can also be used as an aid to solving control problems. The importance of such process models for the construction and utilization of technical information and control systems, as well as for the solution of automation problems with the aid of modern computer systems in the chemical industry, should be expressly reemphasized in this connection.

Product and Process Properties (see Section 2.3). The market exerts a variety of influences on product development. For example, consumer habits change over time, often because of changes in the priority of their needs [811]. New technologies in the converting industries impose other (usually more stringent) requirements on products. The market demand for products with novel or improved properties can also be triggered by the appearance of competing products (Fig. 397).

The company, for its part, offers to this market products whose properties are governed by the raw materials and energies used, the number and skills of the operating personnel, and by the technological condition of the apparatus available. Figure 398 illustrates the decision problem arising from the situation just described. If the existing properties of the products made by the company are no longer adequate to satisfy the needs of the market, the company must decide whether to risk the loss of competitiveness or to react to the new situation by undertaking product and/or process development. To render this decision process objective and intuitive, it is useful to define the "properties" of the products more exactly, on the one hand, and on the other to quantify the company's ability to react ("process") [810].

Product Properties. Beginning with the idealization that product properties can be measured directly and immediately, every product property can be plotted in a histogram, along with its tolerance range, and the full histogram can be interpreted as the property profile (Fig. 399). In a number of studies [812–819], this method has proved extremely helpful in the analysis of requirement and qualification profiles (see also Fig. 400).

Applying this idealization to the concept of product properties often reveals the practical difficulty of relating distinct characterizations of one and the same property as determined by subjective assessments. In principle, of course, even the "technological" properties can be reduced to

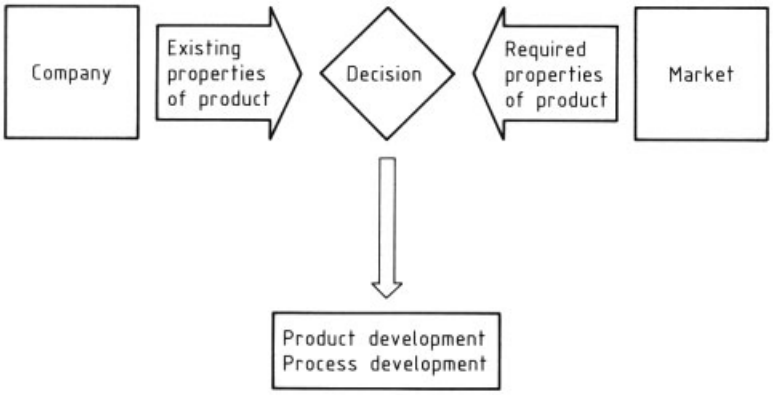

Figure 398. Decision problem for/against product development and/or process development

physical ones, measured, and mapped on metrical scales. For this to happen, however, the usually complex relationships must be elucidated, and only in a few cases can this be expected. The assessment problems arising here are not dealt with further; reference is made to the literature [820] and to Figure 40.

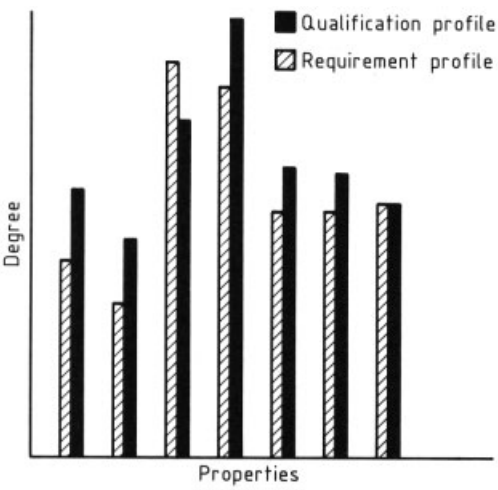

Figure 399. Plot showing qualification profile and requirement profile

If the treatment is initially limited to the case where all properties are measurable, at least one distance function is found that satisfies the conditions for a metric. Let E denote the set of product properties and d the distance function; then the ordered pair $\{E,d\}$ is a metric space. In accordance with standard mathematical practice, the distance measure is immediately dropped from the symbol and E is referred to as the product property space. The concept of "space" can be retained even when there are nonmetrical types of attributes, provided the distance concept is suitably generalized (e.g., topology and separation axiom → Hausdorff space).

As a rule, the property profile of a product includes redundancy. An analysis of the deterministic and/or statistical relationships of the profile leads to a reduction in the number of attributes needed to describe the object "product." This process is briefly described in Figure 401 and illustrated in Figures 402 and 403.

The result is that any product property profile can, with the aid of the concepts just developed, be treated as a point in a suitably selected product space. These concepts also make possible an exact definition of the hard-to-pin-down term "quality": Quality is a point in product space. If the components of the profile are also linearly independent and their realizations can be plotted on the real axis, then the property space is a vector space over the real numbers, and any profile can be interpreted as a vector in this space. Tolerances of the individual product properties allow the point to grow into an elementary volume (Fig. 404; see also Fig. 389).

Process Properties. A "process" is the operation of creating products having certain properties. Its operationalization leads to the elements of the process profile, the process properties P_j. A given realization of all the quantities characterizing the process (e.g., state variables, parameters, control variables, etc.) is called a process state. By construction, so to speak, all relevant

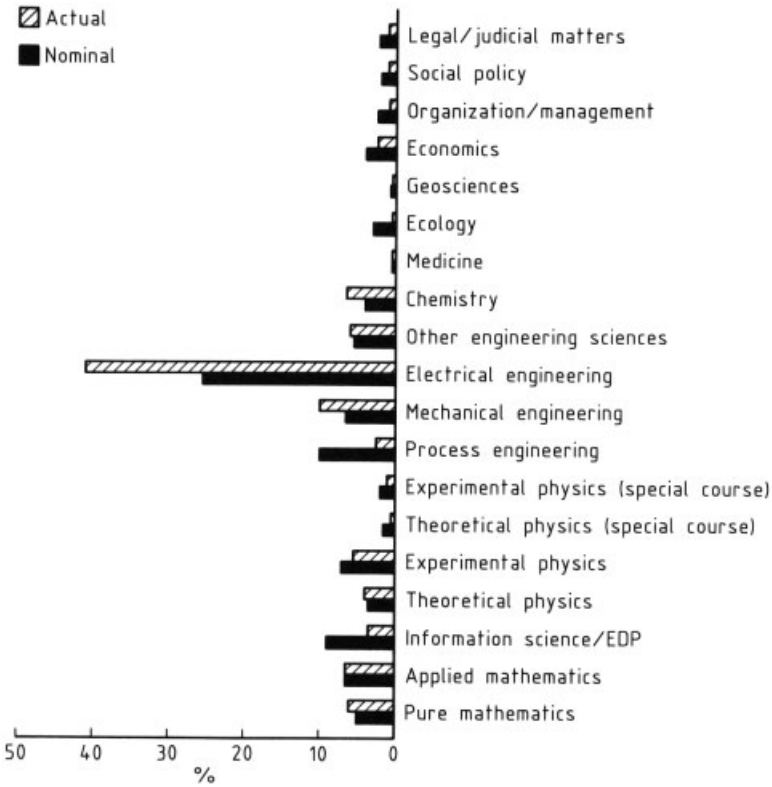

Figure 400. Qualification profiles of electrical engineers in process control engineering [817]

Reduction in number of independent components of profiles by identifying

a) deterministic relationships
b) statistical relationships

Examples	Relationship	Explanatory value
Electrodynamics	Physical theory	High
Diffusion equation	Semi-phenomenological model	↓
$J = \sigma E$	Empirical model	
$\gamma = \alpha + \beta x$	Regression model	Low

Figure 401. Analysis of product property profile (see also Section 3.4)

process properties involved in process development and plant design will be measurable (monitorable). In canonical fashion, the independently variable process properties span a vector space. By analogy with the product property profile, a recipe can now be interpreted as a point in the process space (Fig. 389).

Since the region of definition (range of variation) of each process property in real systems is an interval of finite length, in analyzing a particular plant the treatment can be restricted to a finite region in process space. All that is necessary is to replace E_j by P_j in Figure 404.

Product and Process Development. The entire production process is handled as a successive transformation of product property profiles (see Section 2.3). A process, specified by the process properties and represented by a point in

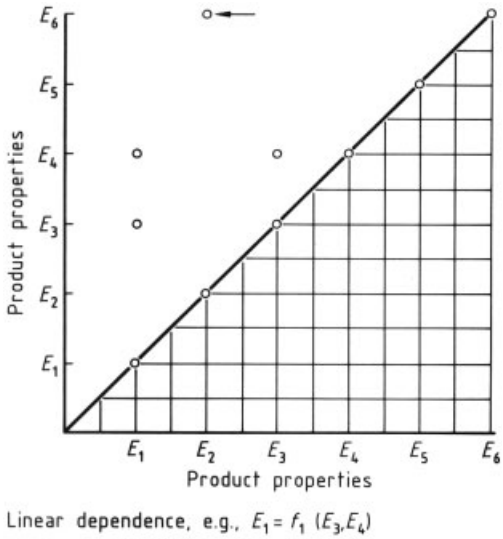

Figure 402. Correlation between components of the product property profile

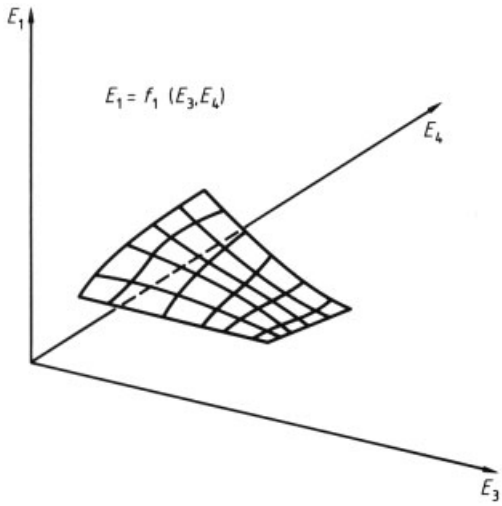

Figure 403. Functional dependences between components of the product property profile

process space, maps an initial property profile to a final property profile and thus links two points in property space. Now the procedure of "product development by process variation" can be formally represented as a function

$$F\,[P_1, \mathrm{P}_2, ..., P_j; E_1, E_2, ..., E_i; E_\mathrm{A}] = 0 \qquad (11.3.3)$$

where P_j are the process properties in process space, E_i are the product properties in product space, and E_A denotes the initial product properties upstream of the process (see also Figs. 360 and 417).

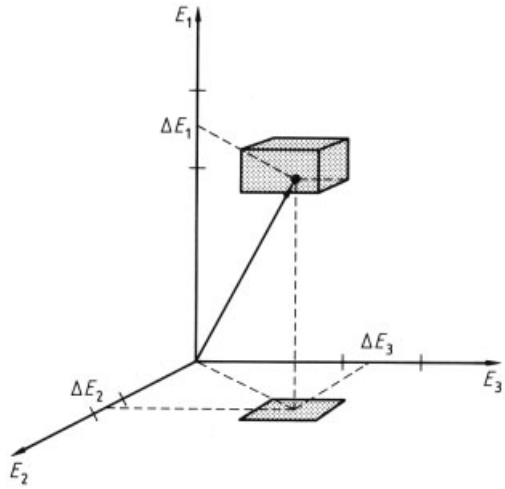

Figure 404. The product property profile as a multidimensional vector in product space

After these general preliminary comments, the particular question of concern in this discussion – the form of the product property profile – comes to the fore. The initial profile is taken as given and fixed. Equation (11.3.3) then depends on the variables P and E. If the market now demands from a product a certain nominal property profile E_S, this has its formal expression in the condition:

$$g\,(P_j) = 0 \qquad (11.3.4)$$

where g is specified by the initial property profile E_A and the nominal property profile E_S and by the interaction of plant, apparatus, and product taking place in the process step under examination.

Specifying one fixed property profile (a point in property space) reduces the number of degrees of freedom in process space by one. The set of permitted product states then lies on a hyperplane in process space (Figs. 405 and 406). After this analysis of the relationships between product and process spaces, an attempt is now made to answer the question of how a product property profile imposed by the market can be realized with the company's resources, taking both safety and economics into account.

Interpolation. In the first step, this objective is to be achieved and the possible experiential horizon of an existing plant is to be exhausted.

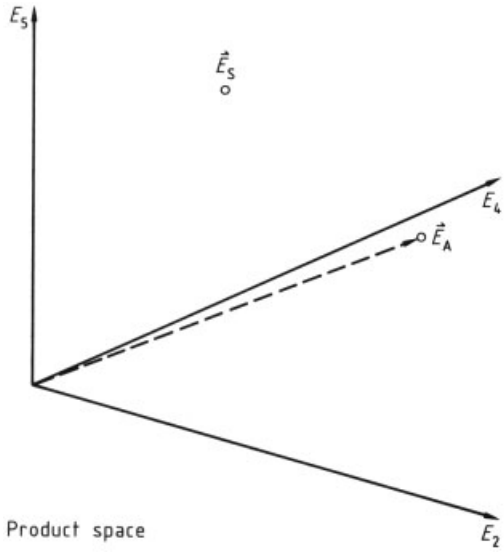

Figure 405. Initial product property profile E_A and nominal product property profile E_S as points in product space

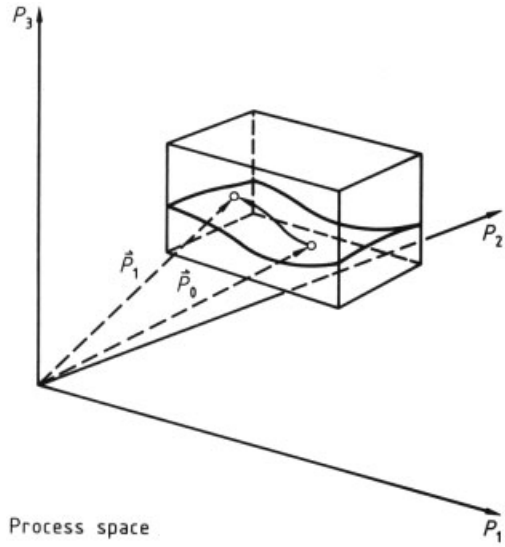

Figure 406. Hypersurface of property-equivalent process states in process space

A "recipe change of the first kind," which does not lead out of the original hyperplane in process space (Fig. 406), does not change the product properties (Fig. 405); such a change is therefore ruled out if the required profile E_S and the initial profile E_A are different. If, however, one moves off the hyperplane ("recipe change of the second kind"), products are now obtained with new property profiles (Figs. 407 and 408).

If one of these points lies sufficiently close to E_S, product development can be regarded as successful and no plant change is required.

If the requirements imposed cannot be satisfied in the context of the existing capabilities, the initial experiential horizon in process space must be enlarged.

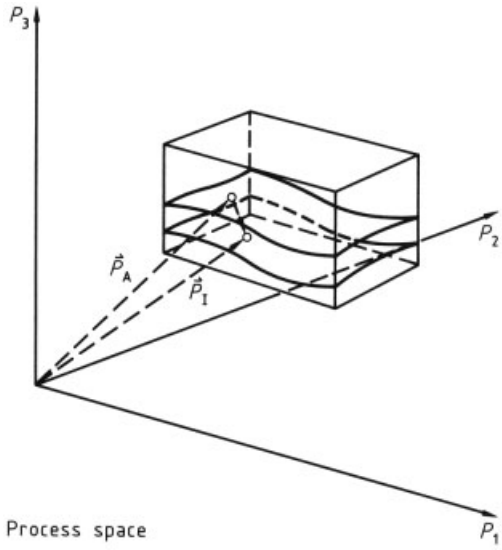

Figure 407. Property-changing process-state variations with no modification of plant

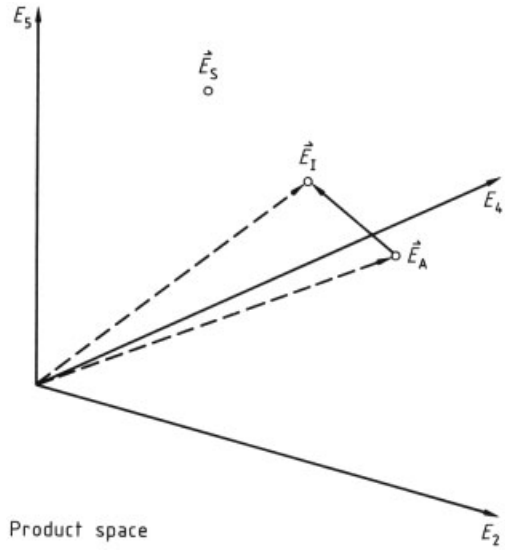

Figure 408. Possible property changes after the action described in Fig. 407

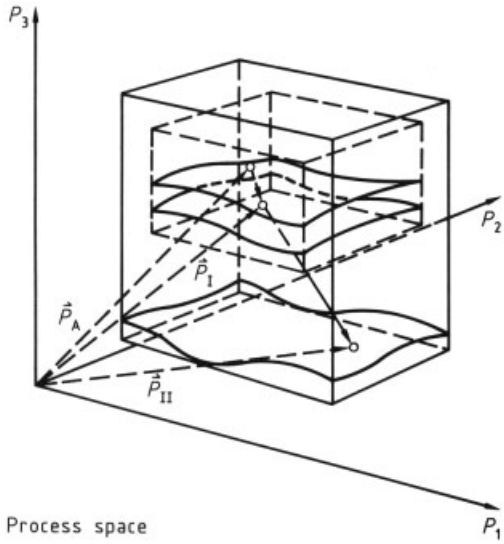

Figure 409. Property-changing process-state variations with modification of plant

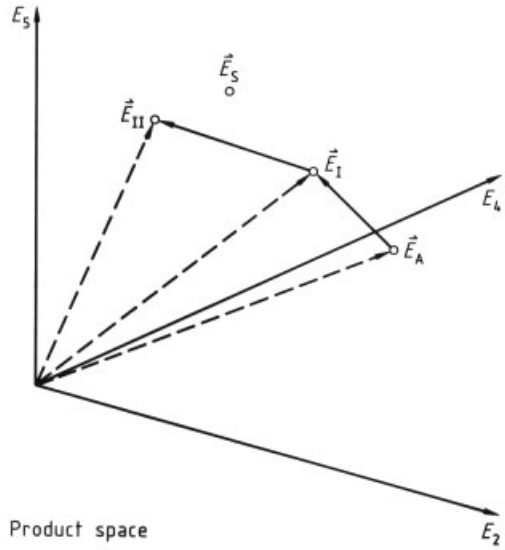

Figure 410. Possible property change after the action described in Figure 409

Extrapolation. In a second step, this is to happen while the dependences in process space are preserved. The same set of process properties are selected as above. If, in the development of products with new properties, the ability to produce the original assortment is to be preserved, then the range of variation of the individual process properties must be expanded by constructive actions. If the strategy requires the substitution of individual products in case of (partial) abandonment of the original assortment, subsequent process development comes down to generating a new range of variation of the process properties (upgrading, "reboring," debottlenecking).

Figures 409 and 410 illustrate this extrapolation step. If such an extrapolation does not lead

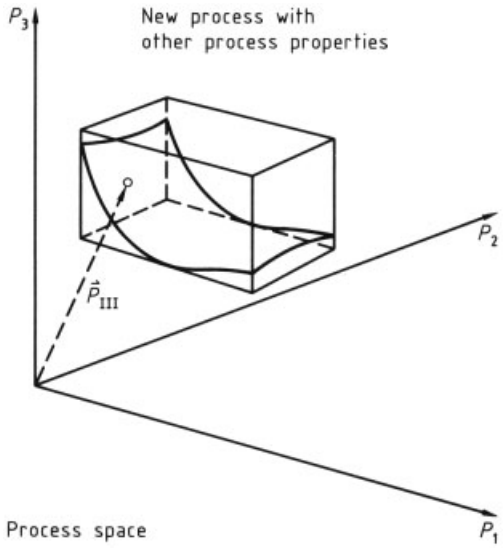

Figure 411. Hypersurface of property-equivalent process states in the new process space generated by process development

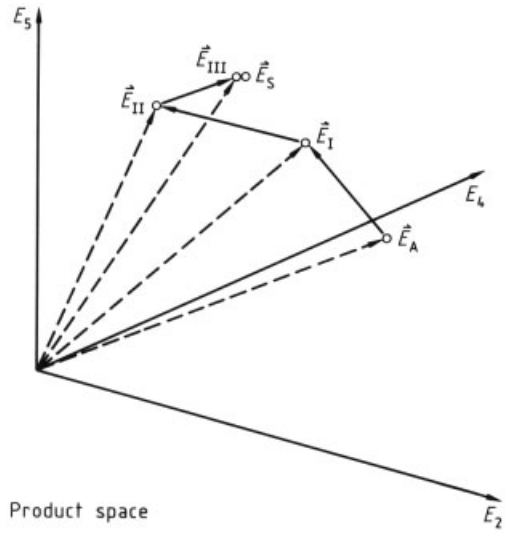

Figure 412. Possible property change after the new process development described in Fig. 411

to the target, the experiential horizon must be expanded in a third step by development of a new process or a new plant.

Under some circumstances, this leads to new relationships in process space (Figs. 411 and 412). It remains valid that a hypersurface (generally of a different nature) in process space corresponds to the (desired) property profile.

Optimization. Subject to the condition of generating equivalent property profiles, all process states on the respective process-space surface are equivalent. This fact can now be used in selecting a process state optimal with respect to the extremalization of a target quantity (cost, energy consumption, yield, etc.). From the mathematical standpoint, this commonly reduces to a variational problem with constraints (Fig. 413).

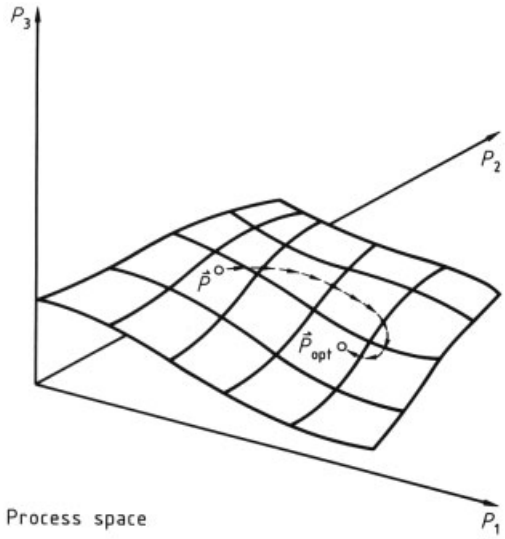

Figure 413. Optimization of the process with respect to a goal quantity

Noted that the concept developed above makes the actions to be taken in product development by process variation, as well as the way of evaluating their success, more intuitive, and also makes it possible to check them step by step (see also Figs. 414 and 415). However, a certain effort must be taken in order to prepare the required tools.

The question whether product development is necessary is thus answered by comparing the qualification profile E_A and the requirement profile E_S of the marketed product, that is, by deciding whether the existing properties of the product correspond to the properties required by the market.

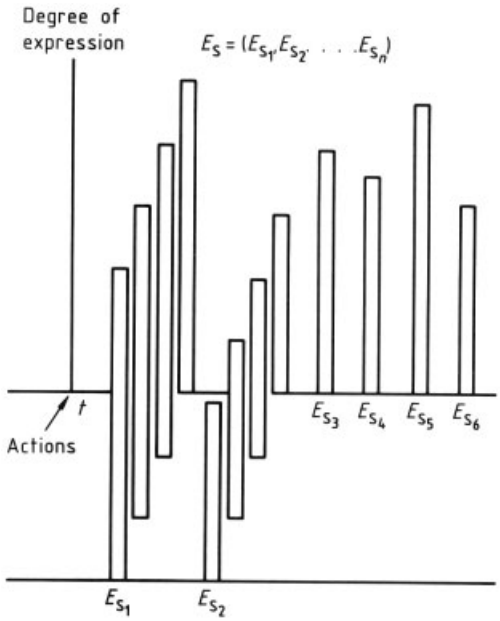

Figure 414. The product property profile as a function of time

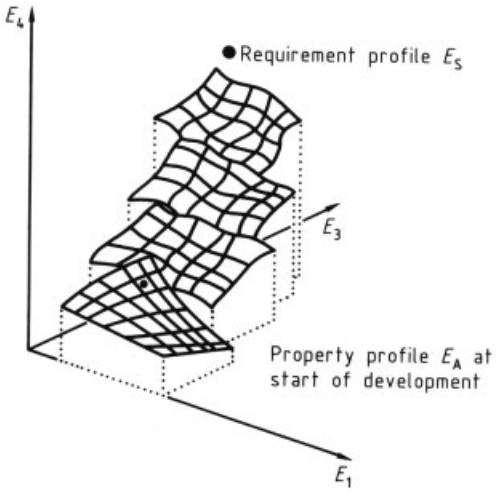

Figure 415. Dynamics of process development can be traced in product space: the portions of surfaces represent sub-goals of development

If the qualification profile and the requirement profile are in agreement to a great ex-

tent (that is, if the point in product space does not change), then a merely desirable improvement in the result does not demand any product change; if the process structure remains the same, however, itwill be useful to optimize the possible changes in process properties (top row in Fig. 416) without moving off the equiproduct surface (see also Fig. 413).

	Point in product space	Point in process space	Process structure
Process optimization	does not change	changes, stays on equiproduct surface	remains unchanged
Product modification	changes slightly	moves off equiproduct surface*	remains unchanged
New product development	changes greatly	moves off equiproduct surface	changes

*The equiproduct surface is the locus of all process property profiles that lead to the same product property profile.

Figure 416. Process/product variations

A product modification required by the market can be effected in either of two different ways:

- If the requirement profile is close to the qualification profile, then it is generally possible to continue working with the same process structure. For a product modification, the process properties are changed within the technically possible range of variation, but this necessitates moving off the equiproduct surface (middle row of Fig. 416).
- If the requirement profile and the qualification profile of the desired product are far apart, then as a rule it is also necessary to change the process structure, that is, carry out a new development (bottom row in Fig. 416).

Measures for Realization. *Production Model.* The material flow of a typical multistage production process in the chemical industry (phase model) is shown in Fig. 360). By abstraction with the aid of the concepts developed above, this material flow is mapped onto an information flow. The elements of the product space are described by the product properties and the points in the process space correspond to the process properties. Further details can be found in [807] and Figure 417 (see also Sections 2.3 and 3.5). Figure 418 shows how the information acquired can be utilized to discover the explicit relationships between plant, machine, apparatus, and product (i.e., between the process properties and the process) for each process step.

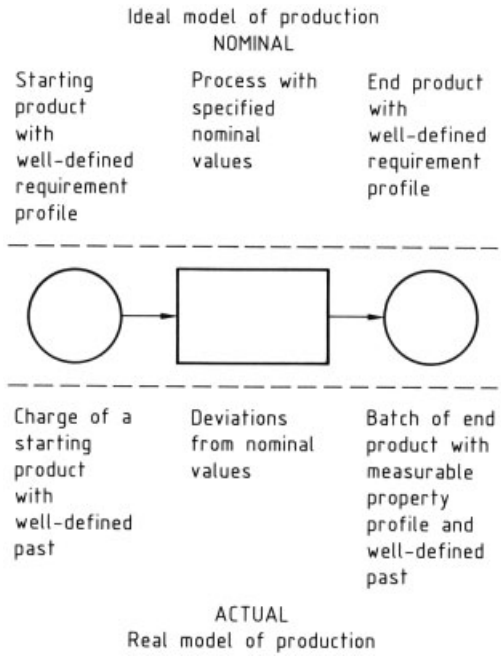

Figure 417. Ideal and real models of production (phase model)

In particular, methods of "system identification" make it possible to derive statistical process models from production data. Such models are needed for tasks in process control engineering (e.g., process indicators [806], observer model). Multivariate methods (such as factor analysis, cluster analysis, and analysis of variance) make it possible to establish functional relationships between process data and quality data (see Section 3.2).

In summary, it can be stated that the central objective of any process analysis and process

optimization is to achieve reproducible nominal product properties. Naturally, the first step is to describe with care the product properties, from the beginning up to the end of the entire production process, and of course also for each intermediate step. The creation of the phase model as shown in Figure 416 is therefore the central task of the process technologist. A holistic viewpoint is of fundamental importance (see also Chap. 10). This information-technical step is the most important task.

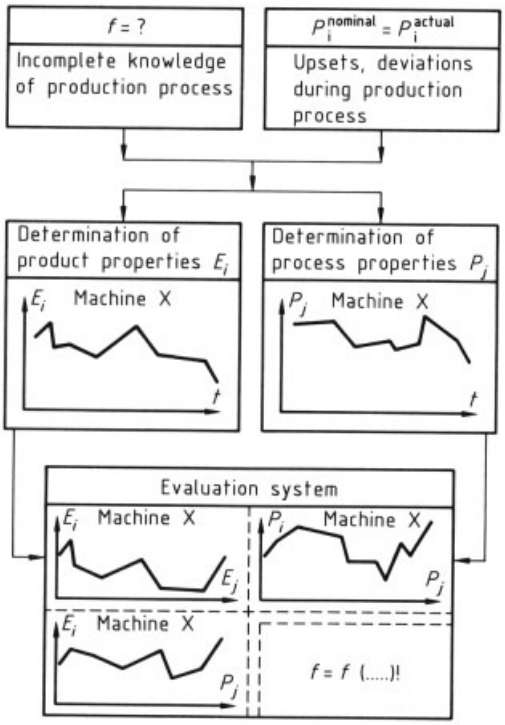

Figure 418. Evaluation system for quantifying the relationship between process and product properties (see Section 3.2)

Execution planning, execution, and quality assurance are described in depth in Chapter 10. They are to be applied, with appropriate modifications, to the problem and task discussed here.

Process analysis and thus process knowledge are also necessary conditions (as shown in Section 11.2) for state-oriented human – process communication so that the production process can be managed and controlled at all times and under all conditions.

11.4. Maintenance Strategies

Introduction. Before maintenance strategies are discussed, the key objectives of maintenance should be defined [821].

Production facilities are used for making consumer goods. During this production process, the facilities wear out in various ways. In order to remain functional despite wear, production facilities have a reserve of possible function fulfillments, called "wear-out reserve," which is gradually diminished during service (Fig. 419) and ultimately leads to breakdown.

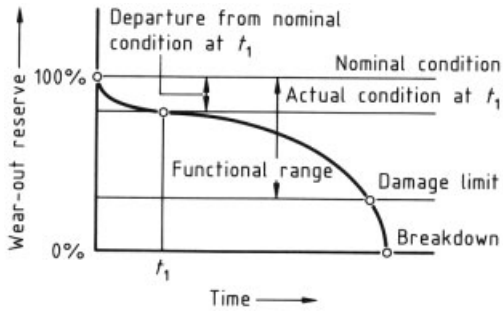

Figure 419. Wear-out reserve versus time

In the nominal condition, the production facility has a 100 % degree of function fulfillment. After a certain time, it loses its wear-out reserve. The production facility remains operable, however, over a broad range.

The objective of maintenance is to keep the wear-out reserve always above the "damage limit" in order to prevent breakdowns. Maintenance includes three activities [821]:

- Routine maintenance to preserve the nominal condition
- Inspection to measure and assess the actual condition
- Corrective maintenance to restore the nominal condition

It is recommended that corrective maintenance be subdivided into scheduled maintenance and forced repairs (unscheduled maintenance) (Fig. 420).

This distinction relates solely to the management of maintenance work, not to the quality of the work. In the case of unscheduled maintenance, including forced repairs, the schedule

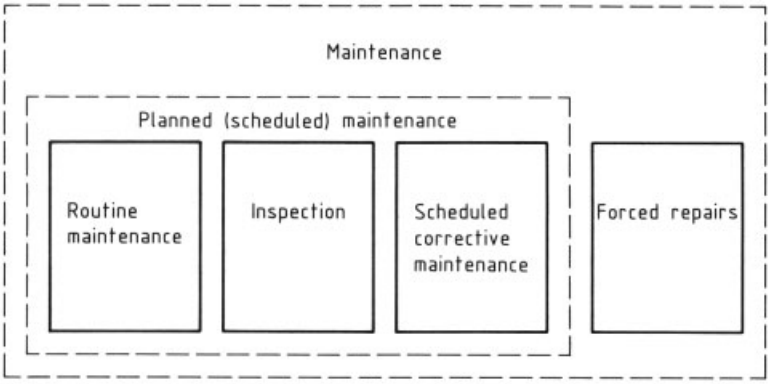

Figure 420. Scheduled and unscheduled maintenance

is dictated by the event, and the maintenance action is generally predetermined as well. The maintenance worker can thus do nothing but react. In the case of scheduled maintenance, the sequence of actions is determined in advance by the maintenance worker himself, who also influences the actions taken. In other words, the worker can take the initiative. Naturally, it is an important goal to put as much maintenance as possible into the "scheduled" category.

Maintenance will become even more important. The reasons are the increasing complexity of plants, the greater depth and breadth of automation, the growing size of production facilities, and the increasingly stringent requirements imposed on product quality, safety, and environmental protection.

Modern maintenance practices for process control and automation systems are discussed in [679].

As the tasks of maintenance change, its objectives also change. The goals of a maintenance organization have been stated as follows [822]: To guarantee the retention of value and functionality of production facilities while insuring compliance with requirements relating to

- Increasing safety and quality levels
- Reliability of on-time performance
- Economically optimal operation

Only when objectives have been clearly formulated can the "correct" strategies be identified and embodied in company policies so that economy measures in the maintenance area do not cause certain maintenance tasks to be neglected.

It is the increasingly high required safety and quality levels that mandate a change of priorities in cost-reduction strategies. Two common cost-reduction practices are

- Economical performance of maintenance activities
- Optimization of material and supply use in maintenance

But the potential of these has been largely exhausted. Now a new requirement has taken on top priority: reduction or even prevention of maintenance.

This practice, which involves the elimination of weak points, still holds a substantial potential for improvement. Production facilities needing fewer repairs have higher availability and can be operated both more economically and more safely. Finally, "maintenance" in a holistic sense has become a topic in the design process (Chap. 10) and must be considered at that stage.

Strategies for Performance of Maintenance Activities. Maintenance strategies serve to make maintenance a more systematic operation. They are oriented toward maintenance activities (Fig. 420). One speaks of a certain maintenance strategy, say an inspection-oriented strategy, if this activity (inspection) is purposely emphasized in the context of changing objectives and is not sacrificed when economy measures are introduced. Naturally, this does not mean that other maintenance activities are wholly de-emphasized.

Breakdown Strategy. The dominant role in the past was played by the breakdown strategy, which is based on forced repairs and, by definition, unscheduled maintenance. Despite the high

cost of devices used in process control engineering, this strategy is still fairly important because many of the sensors and actuators come into contact with products and hence experience contamination, deposits, plugging, and corrosion. It can be shown statistically that such breakdowns are not greatly reduced by more intensive scheduled maintenance. Some improvement is gained with measuring techniques, such as radiometry, that do not involve contact with the product.

Indeed, the breakdown strategy does have advantages: It entails low costs, the work can be done on site by foreman/craft teams, and downtime is generally short.

However, some significant drawbacks have also emerged:

- Costs increase with increasing degree of plant utilization and investment level
- Because of the "workshop" orientation, it may take much effort and expense to maintain operation at weak points such as those resulting from interaction with products; this strategy thus achieves the exact opposite of maintenance prevention
- More stringent requirements on safety, occupational health, and environmental protection, coupled with heightened public sensitivity, mean that a breakdown maintenance strategy cannot be practiced exclusively

Scheduled maintenance strategies have thus become more important.

Preventive Maintenance Strategy. Much was expected of preventive maintenance when scheduled maintenance strategies were first introduced. It is not just a matter of cost that has made this approach obsolete. The reliability of many devices (service lives of valves and motors) today is so great that preventive maintenance is justified only where an element wears out after a known time and can then be replaced as a unit, for example in the case of fluorescent tubes. Preventive maintenance is still necessary when there are special factors, such as plant safety or quality assurance considerations.

Inspection-Oriented Strategy. As safety and environmental requirements grow more stringent, the inspection-oriented strategy has come to play an important role. This applies to many devices significant for damage prevention or limitation, such as rupture disks, leakage retaining vessels, and protective and damage-limiting devices in process control systems. Legislation, such as the Störfallverordnung in Germany [Accident Regulation], mandates the continuous surveillance (inspection) of such devices in the form of periodic function checks. The performance and results of these actions must be documented.

These requirements have had dramatic effects. As a one-time engineering cost, they have given rise to extensive safety analyses with plant upgrades and danger aversion plans (e.g., functional inspection charts). More importantly, they have led to continuing additional labor costs for plant monitoring and safeguarding and for the documentation of actions carried out (see Section 3.5).

The specific cost of monitoring and documentation may decline over a span of years as new and improved inspection methods appear on the market. These might include instruments with automatic self-testing and auxiliary diagnostic devices in electronic systems (see Chap. 5), as well as model-aided early fault detection methods (e.g., [823]).

Another type of inspection-oriented strategy, visual inspection, is the basis for determining and assessing the conditaqtion of production facilities and process control equipment. The "preservative register" is a suitable technique for documenting the results [824].

Depending on the economic and political situation, another strategy may gain in importance: Maximization of plant availability, and thus safety, by application of the redundancy principle, as is already required in nuclear power plants [825]. The redundancy principle requires installation of spare devices, which automatically take up the function of broken-down units and report the malfunctions. This is not, however, a typical maintenance strategy but rather a practice that must be adopted at the design stage (see Chap. 10).

High availability, such as is achieved under the redundancy principle, improves the operating economics. The application of this principle leads, however, to higher costs, at least initially. Additional investment is required, which in turn gives rise to additional maintenance costs, since the redundant units must also be kept operable even when on standby status.

Scheduled Corrective Maintenance Strategy. The analysis of strategies would not be com-

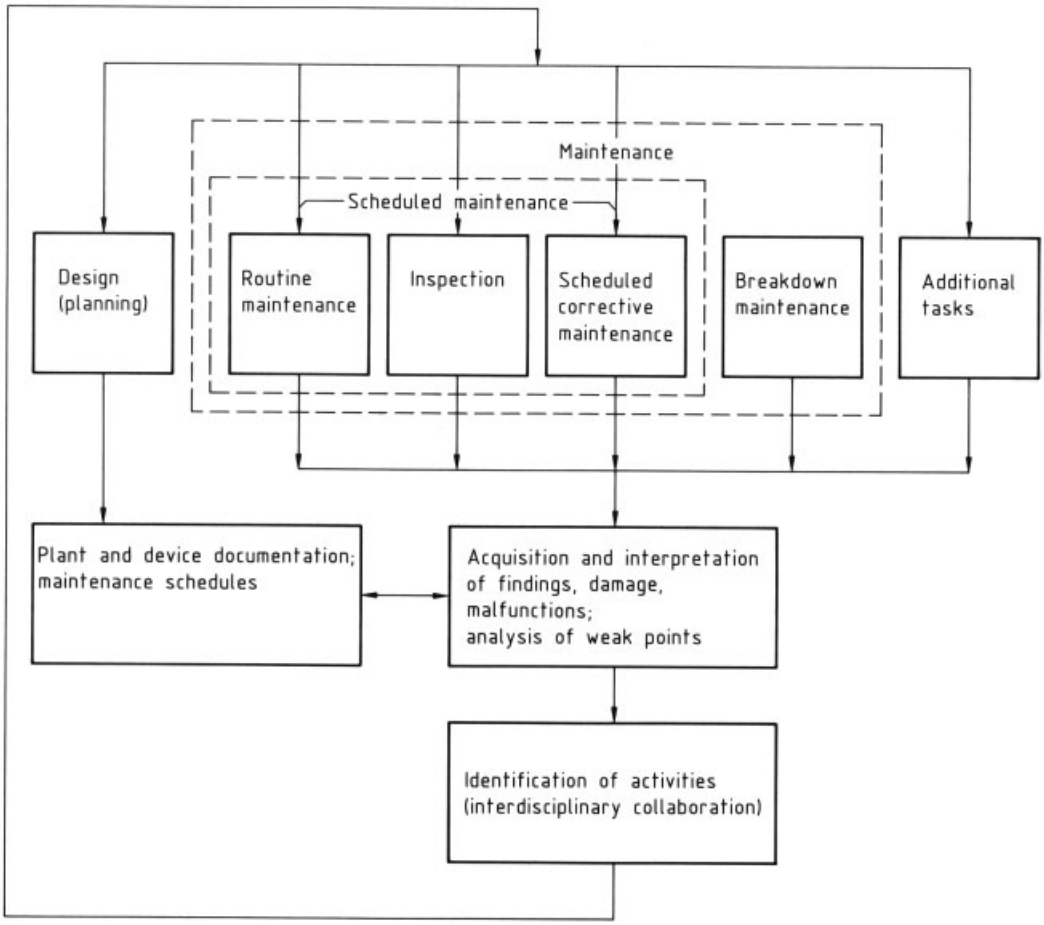

Figure 421. The maintenance control loop [822]

plete without the strategy of scheduled corrective maintenance. Scheduling corrective maintenance (repairs) does not have the aim of restoring the nominal condition; instead, alternative solutions must be sought, with an eye to economic, ecological, and process-engineering considerations. The following question must be examined: maintenance, investment, or both? This applies to plants, plant sections, apparatus, and process control equipment. The procedure, also referred to as "intelligent" maintenance, regards maintenance in the context of the plant economy, that is, holistically. The "plant economy" includes

- Design
- Fabrication, installation, setup
- Use (operation, maintenance, upgrading)
- Replacement

The centerpiece of the method is a decision model that provides a sort of value analysis to support decisionmaking about maintenance practices. On the basis of a situation analysis covering certain independent variables and constraints, the technique compares alternative actions and rates them through cost – benefit analysis. Decisionmaking is not, however, automatic; an educated interpretation of events is needed.

Maintenance Control Loop. Maintenance is concerned with the actual and nominal conditions of the facilities and equipment being maintained. Maintenance can thus be regarded as a control loop, as illustrated in Figure 421, which shows a breakdown of maintenance activities. Findings, damage, and malfunctions must be acquired and evaluated; weak points must be iden-

tified by a process of closing in. Finally, the loop must generate actions, which function as control actions. They react on the facilities being maintained (except for unschedulable breakdown maintenance).

In the maintenance control loop, planning (design) work is an essential component of the plant economy. It is at this stage that maintenance costs are set by the choice of process and plant design. Planning yields the plant and device documentation, which serve as controlling data, but the working plans for scheduled maintenance can also be generated in this stage. Maintenance-oriented installation is crucially important, along with the maintainable design of apparatus and equipment (see Section 3.5 and Chap. 10).

A holistic maintenance program requires the support of a powerful information system. Maintenance-relevant information stems from all maintenance activities. The documentation, however, must focus more on movement data generated on site by the operating engineers in the course of their work. This information goes into object-related history files, which can then provide the informational basis for documenting maintenance activities or analyzing weak points.

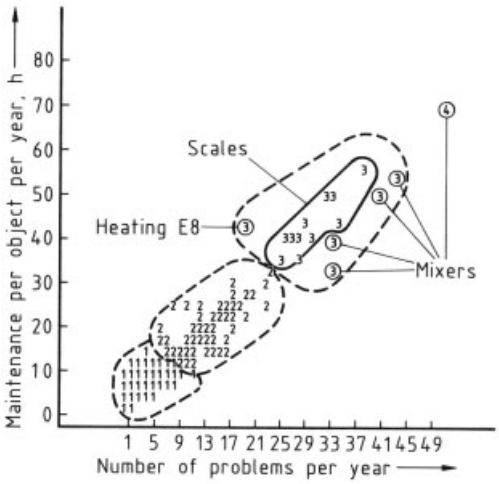

Figure 422. Use of cluster analysis to identify weak points in a process control system

Figure 422 illustrates the identification of weak points in process control systems. The ordinate is the total time spent in remedying problems, and the abscissa is the number of problems arising with individual objects in the system. Each entry characterizes one object. The weak points are located at top right. They stand out because of the large maintenance time requirement and the large number of problems. On the basis of the history files, the numbered groups can be formed by an computational technique called cluster analysis [826].

Outlook. Breakdown maintenance continues to play by far the major role in the chemical industry. What is lacking is the logical application of scheduled maintenance strategies. This has to do with a shortage of methods and tools with which a maintenance program oriented chiefly toward preserving and restoring produc-tive capacity can be conceptually and informationally integrated into the plant economy. There are, however, difficulties in achieving this kind of integration; the staggered and often uncoordinated adoption of information technology (hardware and software systems) has meant that bodies of engineering, administration, economic, and logistical data are mutually inconsistent. This underlines the importance of using the methods described above for structuring the plant economy, and only then implementing it technically (see Chap. 2).

In an integrated data processing system, maintenance should be a component of the plant economy with at least the following elements [823]:

1) A documentation system for the acquisition, structuring, editing, and storage of maintenance-relevant plant data
2) Planning and control systems for scheduled maintenance
3) A control system for generating action- and object-based cost and performance reports and calculating maintenance performance figures

A maintenance information system comprising these blocks should have compatible interfaces to other in-company data processing systems, especially to the production planning and production control, bookkeeping and cost accounting, materials records, and engineering subsystems (see also Fig. 20).

12. Standards, Committees, and Associations

12.1. Principles

The committees and associations that are either directly or indirectly concerned with process measurement and control are so many and diverse that it is impossible to present a complete list. This chapter merely indicates the points of concentration, functions, and objectives of the most important such standardization bodies.

What is Standardization? A clear answer to this question appears in DIN 820 Part 1; 01.86; Paragraph 2: "Standardization is the unification of material and immaterial objects for the benefit of the general public, carried out jointly and in a planned manner by the interested circles. It must not lead to a special economic advantage for individual parties. It promotes rationalization and quality assurance in business, technology, science, and administration. It contributes to the safety of persons and property as well as the improvement of quality in all areas of life. What is more, it contributes to good order and the dissemination of information in the specific area standardized. Standardization is practiced at the national, regional, and international levels."

Who May Create Standards? In Germany, only DIN and DKE are permitted to issue mandatory standards. Under the standardization agreement between DIN and the Federal Republic of Germany, May 5, 1975, DIN is to represent the standardization interests of the Federal Republic in regional and international standardization bodies. CEN and CENELEC European standards also become effective only after they have been published by DIN or DKE. IEC and ISO international standards can be adopted as national or European standards, but there is no obligation to do so.

German-speakers distinguish between the terms "norm" and "standard." Norms are always somewhat official, while standards are private in nature and represent agreement among certain groups. Both terms relate to technical specifications. In the case of a norm, such specifications are adopted by a governmental or internationally recognized standardization organization. Common usage, unfortunately, does not always maintain this precise distinction, and one often speaks of ISO standards. In the German sense, these are norms, since they are issued by ISO.

VDE and VDI standards (called "VDE-Bestimmungen" and "VDI-Richtlinien") do not have the status of norms in law. They are designated as the state of the art. The same applies to the recommendations and work sheets published from time to time by other bodies such as NAMUR, ZVEI, or VIK.

The purpose of these publications is to make results in science and technology known to a broad public as promptly as possible, since experience has shown that the issuance of norms is a very slow process. These papers generally become proposed norms for consideration by standardization bodies, are discussed among a broad public, and after some years appear in modified form as norms. The essential function of these associations is thus to create draft norms and evaluate draft norms.

Both norms and standards are described by the single word "standard" in American English. A "norm" in Germany (and many other countries) might be equated, quite roughly, to a mandatory standard or a standard embodied in a governmental regulation, while German "standard" might correspond to a "voluntary" standard. Standards issued by ANSI, ASTM, etc., are considered voluntary in the United States, but elsewhere they may be regarded as closer in legal status to "norms." The remainder of this chapter employs "standard" for both classes but provides clarification when necessary.

So-called industry standards are in essence neither standards nor norms, but successful products or specifications that have become established in the market.

Situation in the United States. The ISA occupies a dual position in the United States: As a technical-scientific organization (comparable with VDE in Germany), it not only takes part in the preparation of standards but also has the power to issue its own standards, having been granted this right, for the field of process control engineering, by the American National Standards Institute (ANSI; see Section 12.3).

Also deserving of mention is the Institute of Electrical and Electronics Engineers (IEEE),

which develops and generates international standards. Its publications, however, are not mandatory standards but recommendations.

Changes in the Standardization of Process Control Engineering. The 1980s saw a marked change in technology, from signal-oriented measurement and control to information-oriented process control engineering. Not only companies but also national and international standardization bodies have had to deal with this change.

While standardization bodies in the 1970s were concerned with embodying the state of the art in standards, in the early 1980s the emphasis changed to standardization in parallel with research. To lower research costs, a context for development was established in national and international standardization bodies.

The changes in technology and in standards emphasis also necessitated a new structure. This was acknowledged at the end of the 1980s when three new bodies were set up: IEC TC65 (Technical Committee 65 – Industrial Process Measurement and Control), DKE FB9 ("Special Area 9 – Measurement and Control"), and NAMUR ("Standards Working Group for Measurement and Control Engineering in the Chemical Industry").

12.2. Standardization Bodies and Other Organizations Involved in Standardization

The most important standards bodies for process control engineering, with their functions and objectives, are listed below:

CCITT. The Comité Consultatif Internationale Télégraphique et Téléphonique (International Consultative Committee on Telegraphy and Telephony) unites national postal and telecommunications organizations and has jurisdiction over telecommunications standards.

CEN. The Comité Européen de Normalisation (European Committee for Standardization) has the European national standardization bodies as its members. Together with CENELEC, it creates European standards and provisional standards.

CENELEC. The Comité Européen de Normalisation Electrique (European Committee for Electrical Standardization) includes European national standardization bodies concerned with electrical technology.

CEPT. The Comité Européen des Administrations des Postes et des Télécommunications (European Committee of Postal and Telecommunications Administrations) is the conference of European postal and telecommunications agencies.

DIN. The Deutsches Institut für Normung (German Standardization Institute) is the national institution that develops and adopts standards (mandatory standards).

DKE. The Deutsche elektrotechnische Kommission im DIN und VDE (German Electrotechnical Commission of DIN and VDE, VDE = German Electrical Engineers' Society) is the national organization that develops standards and safety provisions in the electrotechnical field.

ETSI. The European Telecommunications Standards Institute was established by European telecommunications agencies together with the EC Commission after substantial deficiencies had been revealed in the assimilation of standards in the telecommunications field. ETSI's task is European harmonization of standards for public-sector telecommunications services.

IEC. The International Electrotechnical Commission develops international standards in the electrotechnical field.

ISO. The International Organization for Standardization functions as an umbrella organization for nearly 90 national standardization bodies. In the interest of flexibility, ISO standards are formulated with many degrees of freedom and must be made more specific by lower-level institutions.

The *International Electrotechnical Commission (IEC)* was founded in 1906. Its membership includes more than 40 national committees, and the official languages are English, French, and Russian. The purpose of the IEC is to support international collaboration in all matters of standardization and certification in the electrical and electronics fields. It aims to achieve this by means of publications, recommendations, and international standards serving as models for incorporation into national legal codes. IEC standards serve manufacturers, consumers, research institutes, and governmental agencies concerned with international trade and electrical safety issues. Any country, without restriction, can be-

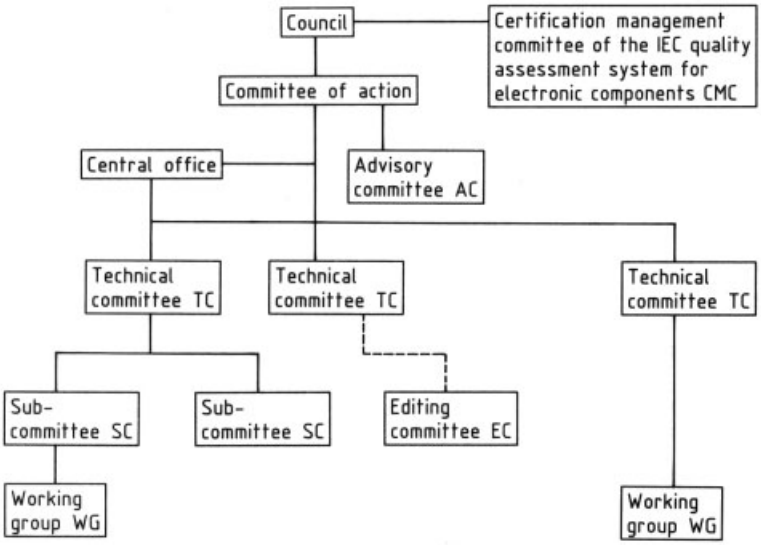

Figure 423. Structure of the IEC (simplified)

come an IEC member. Figure 423 summarizes the structure of the IEC.

The IEC Suborganization "Conformity Testing to Standards for Safety of Electrical Equipment" (IECEE) has the purpose of certifying products for conformity to IEC standards. The certification system applies to electrical devices used in the home, the office, and the workplace. For further information, see [828].

The IEC Technical Committee TC65, "Industrial Process Measurement and Control" (Fig. 424), has as its scope the creation of international standards for systems and elements of industrial process control, including batch and continuous processes, as well as the coordination of standards for measurement and control devices and systems. Subcommittee SC65A, "System Aspects," is concerned with standardizing generic aspects of systems, such as operating conditions, assessment of systems, functional reliability, and so forth. SC65B, "Devices," deals with standardization of devices and systems in the field, the switch room, and the control room. In SC65C, "Digital Communication," standards are developed for communications between subsystems.

The *European Committee for Standardization* (CEN) and the *European Committee for Electrical Standardization* (CENELEC) came into being in the early 1960s as regional standards bodies (Fig. 425). The membership of CEN/CENELEC includes the national standards institutes of all EC and EFTA countries. These institutes are also members of ISO (International Organization for Standardization) IEC (International Electrotechnical Commission).

CENELEC has the objective of harmonizing technical standards in Europe. In order to achieve these goals, the harmonization efforts of the Community are (under Article 100 of the EEC Treaty) to concentrate on solutions that establish binding criteria and make standards the basis for determining the technical details of products. Moreover, with regard to problems specific to new technologies, it is important that standardization at the trans-European level become effective at an early stage.

Publications created by the specialist activity of CEN/CENELEC and issued by CEN/CENELEC are called European Standards (EN), Harmonization Documents (HD), Provisional European Standards (ENV), or CEN/CENELEC Reports. European Standards and Harmonization Documents are referred to as "regional standards" in the sense of the ISO/IEC definition. The various publications differ in the following ways:

– The basic rule is to create European Standards, since it is important that the national standards of the member countries be identical wherever possible.

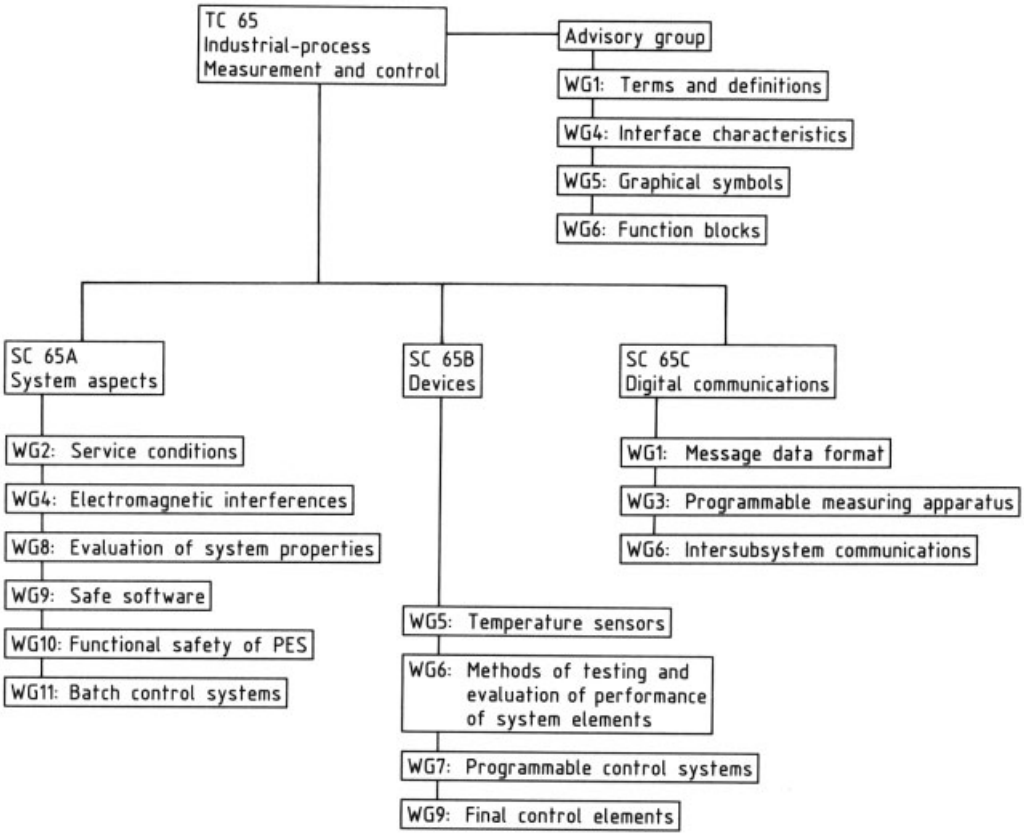

Figure 424. Structure of IEC TC 65

- Harmonization Documents are issued when the change to identical national standards is unnecessary or impractical, and especially when a unification can be achieved only by granting national variances.
- Provisional European Standards are planned standards for use on a preliminary basis.
- CEN/CENELEC Reports are intended to provide information.

The Vilamoura procedure has the goal of promoting collaboration between national committees on a regional level by decentralizing the earliest stages of work. All new intentions to standardize on the national level should be disclosed in detailed form, so that all members are acquainted with the content of the proposed project and can indicate their interest in active participation if they wish to do so.

"New intentions to standardize on the national level" means all intentions to standardize at the national level except for:

- Revision of existing national standards (see explanation below)
- Adoption of the results of CENELEC work
- Adoption of the results of IEC work

The results of national notifications lead to an EN, an HD, or an ENV. Notifications must be sent to all parties concerned (national committees, the CENELEC Secretariat, and the Technical Committee) for response. If no positive responses are received, the notifying national committee is free to develop a national standard.

The notification process for the revision of national standards is based on the assumption that the intention to revise is justified by the necessity of updating the standards. The revision of a national standard for the purpose of taking over an IEC standard, IEC provisional standard, or IEC publication, or for the purpose of bringing the national standard into conformity with an HD or an EN need not to be disclosed under this procedure.

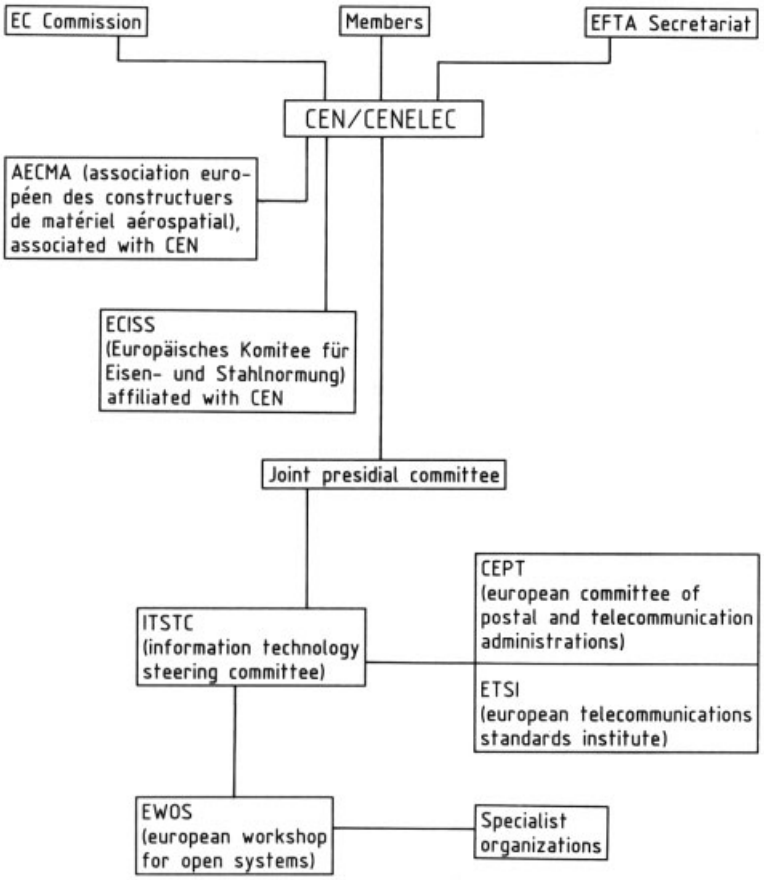

Figure 425. Organizational chart of CENELEC

The Technical Committees of CENELEC are numbered to correspond with those of the IEC. The member countries of CENELEC are listed below:

- Austria: Österreichisches Elektrotechnisches Komitee (ÖEK) beim Österreichischen Verband für Elektrotechnik (ÖVE) (Austrian Electrotechnical Committee of the Austrian Electrotechnical Society)
- Belgium: Comité Electrotechnique Belge (CEB), Belgisch Electrotechnisch Comité (BEC) (Belgian Electrotechnical Committee)
- Switzerland: Comité Electrotechnique Suisse (CES) (Swiss Electrotechnical Committee)
- Germany: Deutsche Elektrotechnische Kommission im DIN und VDE (DKE) (German Electrotechnical Commission of DIN and VDE)
- Denmark: Dansk Elektroteknisk Komite (DEK) (Danish Electrotechnical Committee)
- Spain: Asociación Española de Normalización y Certificación (AENOR) (Spanish Standardization and Certification Association)
- Finland: Finnish Electrotechnical Standard Association (SESKO)
- France: Union Technique de l'Electricité (UTE) (Technical Union of Electricity)
- United Kingdom: British Electrotechnical Committee (BEC), British Standards Institution (BSI)
- Greece: Hellenic Organization for Standardization (ELOT)
- Ireland: Electro-Technical Council of Ireland (ETCI)
- Iceland: The Icelandic Council for Standardization (STRI), Technological Institute of Iceland

- Italy: Comitato Elettrotecnico Italiano (CEI) (Italian Electrotechnical Committee)
- Luxembourg: Service de l'Energie de l'Etat (SEE) (State Energy Service)
- Netherlands: Nederlands Elektrotechnisch Comité (NEC) (Netherlands Electrotechnical Committee)
- Norway: Norsk Elektroteknisk Komite (NEK) (Norwegian Electrotechnical Committee)
- Portugal: Instituto Português da Qualidade (IPQ) (Portuguese Quality Institute)
- Sweden: Svenska Elektriska Kommissionen (SEK) (Swedish Electrical Commission)

The EC Commission has therefore devoted special attention to the development of standardized test methods for verifying compliance with the standards. The ultimate objective is a European certification framework that would award a publicly recognized certificate to products that conform to the standards. The ECITI (European Committee for Information Technology Testing and Certification) is a body created by the EC Commission for the purpose of promoting the testing of standard-conforming products in Europe.

CENELEC always follows the principle of using the results of IEC work whenever possible and also of carrying out new standardization projects within the IEC when there is a prospect of accomplishing this within a given time span. In this way, multiple treatments of the same topic are avoided. The extensive collection of IEC publications and the continuation of standardization in the IEC are important components of this effort. With the information agreement of November 1989 and the cooperation agreement that came into force on January 1, 1991, the essential tools for further collaboration between IEC and CENELEC and their members were created. The cooperation agreement aims, above all, to accelerate the standardization process by causing a draft prepared on one organization level to be submitted for agreement simultaneously and in parallel on the international and European levels.

DIN. The German standards system deals with all technical disciplines. DIN standards reflect actuality and document the level of technical experience. Each standard is investigated with regard to its economic impact, and only those that are absolutely necessary are issued. DIN standards contribute to rationalization, quality assurance, safety, and understanding in business, engineering, science, and administration, and also among the public. DIN and its agencies serve as the national representation in international and European standards institutions (ISO/IEC, CEN/CENELEC).

DKE. DKE is the German organization that develops standards and safety rules in the field of electrical technology. DKE is dedicated to the safety of persons using electricity on the job, at home, and during leisure time. DKE is further involved in solving problems of environmental protection, answering consumer questions, and creating nuclear safety regulations. DKE has jurisdiction over the harmonization of electrotechnical standards within the European Community.

In Germany, mandatory standards can be created only by DIN or DKE. Each European Standard must be incorporated by all CEN/CENELEC members into their own standards system. EFTA countries that have decided against the European Standards are free of this obligation. Adoption requirements such as those that apply to EN and HD documents do not exist for ISO and IEC international standards. On the European level, ISO/IEC standards are subjected to a questionnaire procedure; if they meet with agreement, they are uniformly adopted in all EC and EFTA countries.

Results of the work of DKE are issued primarily under the titles

- VDE-Bestimmung (VDE voluntary standard)
- DIN-Norm (DIN mandatory standard)
- DIN-EN-Norm (DIN European Standard), with DKE number if applicable
- DIN-IEC-Norm (DIN – IEC world standard), with DKE number if applicable

DKE's "Special Area" 9 (FB9) bore the name "Measurement and Control" up through 1990 and was then renamed "Process Control Engineering." FB9 (Fig. 426) creates basic standards for industrial process control as well as specialized standards for the following application areas: process engineering, power plants, networks, buildings, pipelines, ironmaking and iron founding, mining, shipbuilding, instruction and training, and laboratory practice.

It is not the job of standardization bodies to carry out preliminary technical and scientific work. This must be done by the societies that cover the various areas (e.g., NAMUR, GMA, ZVEI). This function includes the development of technical concepts, technical and scientific groundworks for standardization projects, and engineering studies and assessments. The procedure of standardization in parallel with development can only become an effective aid to both parties when close cooperation is achieved between these technical societies and the standards bodies.

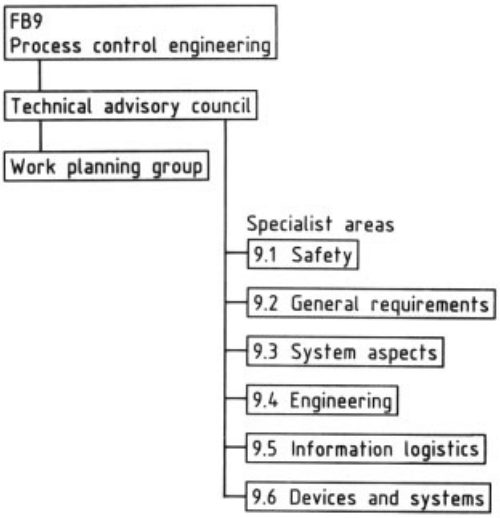

Figure 426. Areas of interest of DKE FB 9

VDE. The transfer of technical and scientific knowledge, as a service to all professional electrical and electronics engineers, is the first task of VDE. The spectrum ranges from historical aspects of electrical engineering, through occupa-tional and educational problems, to accident research, and the drafting of standards.

VDI. The transfer of technical and scientific knowledge is the primary objective of technical and scientific work performed by the Association of German Engineers (VDI). This is accomplished chiefly in the VDI specialist subdivisions.

An important emphasis in these subdivisions is the creation of standards documents known as *Richtlinien* (guidelines), which reflect the state of the art but do not have the legal status of mandatory standards.

VCI. Among the diverse committees and working teams of the Verband der Chemischen Industrie (VCI, Association of the Chemical Industry), the Engineering and Law Committee merits special mention. Through the "Measurement and Control" working team, VCI makes use of NAMUR specialist knowledge. This working team is identical with the governing body of *NAMUR* (see below).

NAMUR. The Normenarbeitsgemeinschaft für Mess- und Regelungstechnik in der Chemischen Industrie (NAMUR, Standards Working Group for Measurement and Control Engineering in the Chemical Industry; Fig. 427) takes in users of process control engineering (electrical, measurement, and control engineering) in the chemical and allied industries. It was founded in 1949 by BASF, Chemische Werke Hüls, DAG Troisdorf,Glanzstoff-Courtaulds,Henkel,Ruhrchemie, UK Wesseling, and Farbenfabriken Bayer. At present, NAMUR has 54 member companies, a third of them located outside Germany (in Switzerland, Austria, Belgium, and the Netherlands).

Activities of NAMUR relate chiefly to:

- Participation in the standards committees of DIN and DKE (in particular, DKE FB9, "Process Control Engineering")
- Participation and input in the creation of standards and specifications at the voluntary standards level (e.g., VDI/VDE-Richtlinien documents)
- Exchange of experience with process control devices, especially with newly developed instrumentation and other hardware
- Publication of NAMUR recommendations in all fields of process control engineering when no other standards exist
- Scientific support of technical journals
- Publication of NAMUR Status reports to document the state of the art in the field of process control engineering

Further details can be found in [224] and in [829].

VIK. The Vereinigung Industrielle Kraftwirtschaft (VIK, Industrial Energy Management Alliance) was founded in 1947.

The principal concerns of the Electrotechnical Committee are as follows:

- Exchange of experience

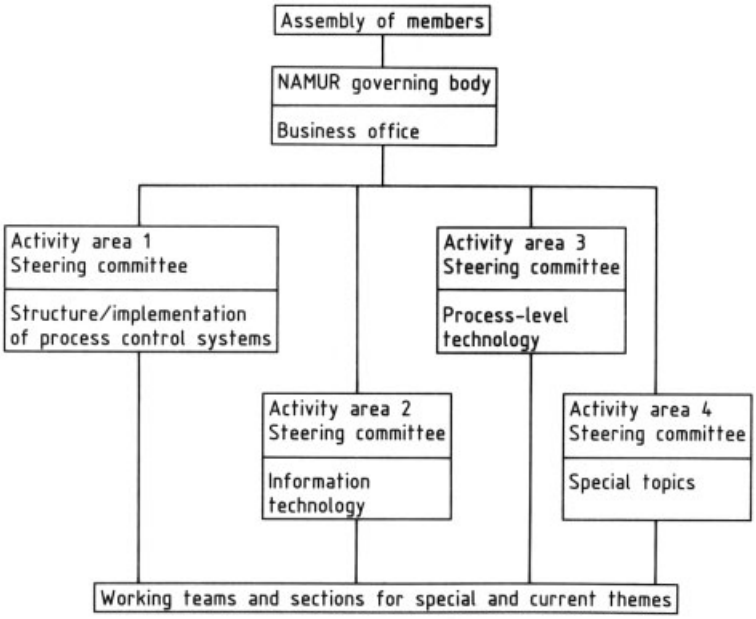

Figure 427. Organizational chart of NAMUR

- Influence on the preparation of devices suitable for consumer use
- Influence on the modification of VDE standards to conform with consumer requirements
- Influence on decreasing the number of designs
- Efforts to establish optimal conditions for accident prevention
- Attainment of economic benefit in the use of electrotechnical devices

This applies to electrical installations and systems in industrial power supply and production service.

ZVEI. The Zentralverband Elektrotechnik- und Elektronikindustrie e.V. (ZVEI, Central Federation of the Electrical and Electronics Industry) represents the interests of the German electrical and electronics industry in matters of economic and techology policy. ZVEI has the following tasks:

- To safeguard the common professional and business interests of its members before all agencies, in particular by cooperating with regulatory, economic, and other agencies and institutions and seeking to provide input to legislative measures
- To promote the general economic and technical development of the electrical and electronics industries

ISA. The Instrument Society of America (ISA) was established in 1945 in Pittsburgh, Pennsylvania.

It is a nonprofit organization active in the technology, science, and educational sector. Its concerns are with the theory, design, manufacture, and application of process control devices, systems, and computers. It issues the monthly International Journal of Instrumentation and Control (INTECH).

In the United States, the ISA made its name with an extensive training program and companion publications such as books, journals, software packages, videos, and CBT (computer-based training) packages. Moreover, ANSI (American National Standards Institute) has delegated to ISA the right to issue mandatory standards in the field of process control engineering. Valuable comments can be found in numerous ANSI/ISA standards sections.

12.3. Technical and Scientific Bodies

It is difficult and often impossible to distinguish between societies that are active in the standard-

ization field and bodies whose work is primarily technical and scientific. The emphasis for technical and scientific committees is on the sharing of experience in research and development, production and technology, training and continuing education, and safety. Such bodies can frequently be identified by the fact that they regularly organize high-level congresses. In this way they contribute to forming public opinion, and this in turn influences standardization.

In Germany, for example, the DVT unites a number of technical and scientific associations, such as DPG, DIN, GI, VDE, and VDI. Also important for process control engineering are GMA, Gesellschaft für Informatik (GI, Society for Information Processing), and Gesellschaft für Mathematik und Datenverarbeitung (GMD, Society for Mathematics and Data Processing), a public-sector research institution.

At the international level, there are a variety of national societies with special objectives. One such society oriented toward automation technology is IFAC; toward metrology, IMEKO; and toward the type testing of measurement and control equipment, the International Instrument Users Association.

DVT. The Deutsche Verband technisch-wissenschaftlicher Vereine (DVT, German Federation of Technical and Scientific Societies) includes about 100 technical and scientific organizations. It has the objective of analyzing high-level problems in science and technology and representing the interests of engineers and natural scientists in scientific, economic, social, political, governmental, and administrative circles. Its tasks include the promotion of applied sciences, the unification of common technical principles, the improvement of technical education, and cooperation in shaping legislation in the field of technology and technical management. The DVT regards itself as a link between members and also between the membership and business societies whenever technical and scientific problems are addressed.

AMA. The Arbeitsgemeinschaft Messwertaufnehmer e.V. (AMA, Data Sensors Working Group) is an association of manufacturing firms in the areas of sensor technology, sensor-based measurement and automation, and microsystems technology.

DPG. The Deutsche Physikalische Gesellschaft e.V. (DPG, German Physical Society) has the following objectives: to advance pure and applied physics, to bring its members and all physicists living in Germany closer to one another, to represent physicists as a body to other members of society, and to promote the exchange of experience among physicists in Germany and between them and their colleagues in other countries.

Notable special activities in the field of process control engineering are the conferences held by the Beratende Ausschuss der Industriephysiker (BAI, Advisory Committee of Physicists in Industry): the sensor technology conference in Münster and the automation conference in Bonn.

GMA. The Gesellschaft Mess- und Automatisierungstechnik im VDI und VDE (GMA, Metrology and Automation Society of the VDI and VDE) is active in measurement and control technology, automation, and process control engineering. The spectrum of related applications extends from the basic chemicals industry to the process industries, industrial and process engineering, energy generation and distribution, the automotive, transportation, and aerospace industries, construction engineering, medical technology, public utilities and waste disposal operations, and special applications in machinery, apparatus, and instruments.

Measurement and control, as a basic process-level technology, forms the basis for the functionally oriented, hierarchically structured field of automation and process control engineering. The principal points of emphasis in the action chain between process, automation system, and human being, as well as the higher-order, associative disciplines and engineering technologies, are communication, design, and simulation, together with information processing and, on this basis, the techniques of process operation and monitoring, automation equipment and systems, the technology of actuators and sensors, and also the dynamics of machinery and processes.

Areas of activity of the GMA:

- Exchange and evaluation of experience and information
- Holding conferences and lectures
- Cooperation with national and international bodies working in this field
- Stimulation and support of research

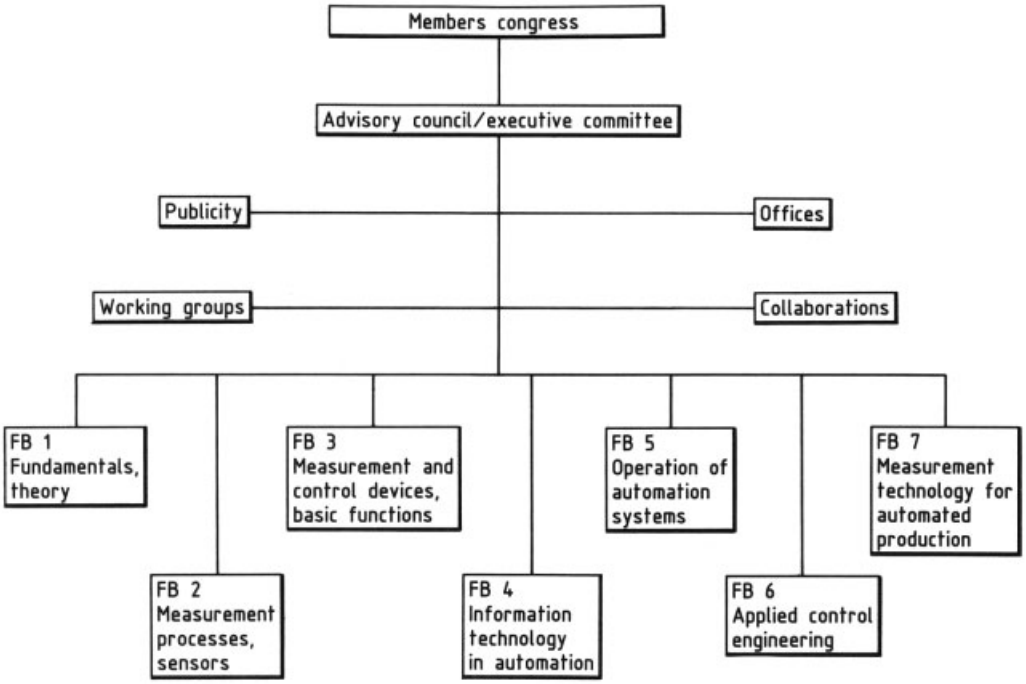

Figure 428. Organizational chart of GMA

- Publication and support of scientific journals
- Development of guidelines and other recommendations
- Sponsoring young engineers
- Involvement in education

National collaborations:

DGK	Deutsche Gesellschaft für Kybernetik
DGQ	Deutsche Gesellschaft für Qualität e. V.
DIN	Deutsches Institut für Normung
DKE	Deutsche Elektrotechnische Kommission im DIN und VDE
GI	Gesellschaft für Informatik e. V.
NAMUR	Normenarbeitsgemeinschaft für Meß-und Regelungstechnik in der Chemischen Industrie

International collaborations:

IMEKO	Internationale Meßtechnische Konföderation
IFAC	International Federation of Automatic Control

The organization of the GMA is shown in Figure 428.

GI. The Gesellschaft für Informatik (GI, Society for Information Processing) has as its objective the interchange of technical and scientific information in the field of "informatics" or information processing.

GMD. The Gesellschaft für Mathematik und Datenverarbeitung mbH (GMD, Society for Mathematics and Data Processing) is a public-sector research institution that uses basic research in the field of informatics to further the development of information technology and its applications. On the basis of wide-ranging basic research, conceptual work is performed and methods are devised that will lead to tools.

GVC. The VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen (GVC, the VDE Society for Process Engineering and Chemical Engineering) is the specialist organization of engineers active in process and chemical engineering as well as allied occupational groups. One of its central activities is the annual congress of process engineers. Its objectives are as follows:

- Sharing of technical and scientific experience
- Employment of specialist competence of the GVC technical committees
- Training and continuing education
- Collaboration with other institutions

– Social policy

IFAC. The International Federation of Automatic Control was founded in 1957 as a multinational organization. Its members are technical and scientific societies from 45 countries. Its objective is to support control science and technology in a holistic sense, that is, technically, physically, biologically, socially, and economically, both in theory and in practice. The task of IFAC is to support information exchange between national and international organizations active in the fields of automation and process control engineering. IFAC holds a world congress every three years; between congresses, it organizes symposia and workshops on special themes.

Figure 429 shows the organizational structure of IFAC.

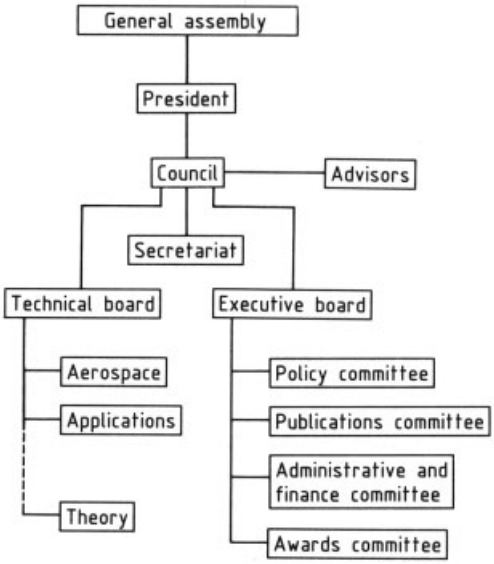

Figure 429. Organizational chart of IFAC

IFAC is a member of the "Five International Associations Coordinating Committee" (FIACC), which also includes the following bodies:

– IFIP International Federation for Information Processing
– IFORS International Federation of Operational Research Societies
– IMACS International Association for Mathematics and Computers in Simulation
– IMEKO International Measurement Confederation

Member societies of IFAC by country are as follows:

– Argentina: Asociacion Argentina de Control Automatico (AADECA)
– Australia: The Institution of Engineers
– Austria: Arbeitsgemeinschaft für Automatisierung
– Belgium: Federation IBRA/BIRA
– Brazil: Sociedade Brasileira de Automatica (SBA)
– Bulgaria: National Council of Automation
– Canada: Canadian National Committee for IFAC
– Chile: Asociacion Chilena de Control Automatico (ACCA)
– China: Chinese Association of Automation, Institute of Automation
– Czechoslovakia: CSFR National Committee of Automatic Control
– Denmark: El. Power Engineering Department, Denmark Technical University, Lyngby
– Egypt: Egyptian High Committee of Automatic Control
– Finland: Finnish Society of Automatic Control
– France: Association Française pour la Cybernetique, Economique et Technique – AFCET
– Germany: VDI/VDE Gesellschaft Mess- und Automatisierungstechnik (GMA)
– Greece: Technical Chamber of Greece
– Hong Kong: The Hong Kong Institution of Engineers
– Hungary: IFAC National Member Organization – Computer and Automation Institute
– India: India Institution of Engineers
– Israel: Israel Association of Automatic Control
– Italy: CNR Commissione IFAC
– Japan: Science Council of Japan
– Kuwait: Kuwait Society of Engineering
– Mexico: Asociacion de Mexico de Control Automatica – AMCA
– Morocco: Association Marocaine pour le Développement de l'Electronique, de l'Informatique et de L'Automatique – A.M.A.D.E.I.A.
– Netherlands: Royal Institution of Engineers
– New Zealand: Institution of Professional Engineers

- North Korea: Korean General Federation for Science and Technology, Korean Association of Electronics & Automation
- Norway: Norsk Forening for Automatisiering (NFA)
- Pakistan: The Institution of Engineers
- Poland: Polski Komitet Pomiarow i Automatyki i Robotyki
- Portugal: Associacao Portuguesa para o Desenvolvimento da Investigacao Operacional – APDIO, CESUR-IST
- Romania: Comisia de Automatizari, Inst. de Cercetari si Proiectari Automatizari – IPA
- Singapore: Instrumentation and Control Society
- South Africa: South African Council for Automation
- South Korea: Korean Association of Automatic Control
- Spain: Comite Español de la IFAC
- Sweden: Kommiten Svenska IFAC
- Switzerland: SGA, Schweizerische Gesellschaft für Automatik
- Turkey: Turkish National Member Organization – Otomatik Kontrol Türk Milli Comitesi (TOK)
- United Kingdom: United Kingdom Automatic Control Council – UKACC
- United States: American Automatic Control Council
- CIS: National Committee of Automatic Control
- Yugoslavia: Electrotechnic Faculty, Automatics and Systems Engineering Division

IMEKO. The Internationale Messtechnische Konföderation (IMEKO, International Measurement Confederation) was founded in 1958. It includes technical and scientific organizations in 31 countries. Its objectives are as follows:

- To support international sharing of experience in science and technology for the development of metrology, manufacturing and design of devices, and application of devices in research and industry
- To promote cooperation of scientists and engineers
- To hold a world congress every three years
- To establish Technical Committees
- To organize symposia, workshops, etc., on special themes and to publish the proceedings
- To cooperate with other international organizations having similar objectives(see IFAC)

A list of Technical Committees follows:

TC 1	Higher Education
TC 2	Photonic Measurement
TC 3	Measurement of Force and Mass
TC 4	Measurement of Electrical Quantities
TC 5	Hardness Measurement
TC 7	Measurement Theory
TC 8	Metrology
TC 9	Flow Measurement
TC 10	Technical Diagnostics
TC 11	Metrological Requirement for Developing Countries
TC 12	Temperature and Thermal Measurement
TC 13	Measurement in Biology and Medicine
TC 14	Measurement of Geometrical Quantities
TC 15	Experimental Mechanics
TC 16	Pressure Measurement
TC 17	Measurement in Robotics

Member societies of IMEKO by country are as follows:

- Australia: Institute of Instrumentation and Control
- Austria: ÖVE-ÖIAV Fachgruppe Messtechnik
- Belgium: Belgian Measurement Confederation
- Brazil: Brazilian Association of Mechanical Sciences
- Bulgaria: Committee for Measurement of the Association of Science
- China: Chinese Society for Measurement
- Czechoslovakia: Czechoslovak Scientific and Technical Society – CSVTS
- Denmark: The Danish Society for Engineering Metrology
- Egypt: Egyptian Organization for Standardization
- Finland: Finnish Society of Automatic Control
- France: Association Française pour la Cybernetique Economique et Technique – AFCET
- Germany: VDI/VDE Gesellschaft Mess- und Automatisierungstechnik (GMA)
- Hungary: Scientific Society of Measurement and Automation
- India: Institution of Instrumentation Scientists and Technologists
- Italy: National Research Council
- Japan: The Society of Instrument and Control Engineers
- Korea: Korea Standards Research Institute

- Morocco: Society for Automation and Control
- Netherlands: Royal Institution of Engineers Division for Automatic Control
- New Zealand: The Institute of Measurement and Control (Inc.)
- Norway: Norwegian Society of Automatic Control
- Philippines: Philippine Instrumentation and Control Society
- Poland: Committee for Measurement, Robotics and Automation – NOT
- Romania: Romanian Measurement Society
- Spain: Committee for Metrology, Spanish Association for Quality Control
- Sweden: Instrument Society of Sweden
- Switzerland: Schweizerische Gesellschaft für Automatik
- United Kingdom: The Institute of Measurement and Control
- United States: Instrument Society of America – ISA
- CIS: All-Union Scientific-Technical Society of the Instrument Industry
- Yugoslavia: JUREMA

International Instrument Users Association. The International Instrument Users Association unites three societies, "SIREP-WIB-EXERA" (EXERA=Association des Exploitants d'Equipment de Mesure, de Régulation et d'Automatisme, Association of Users of Equipment for Measurement, Control, and Automation), with the objective of performing suitability tests on measurement and control devices and systems. The concerns in this type testing are operability, safety, availability, accuracy, freedom from interference, ease of maintenance, etc. The results are communicated only to the member firms and the manufacturing firm. Membership is limited to user firms.

Addresses of Organizations Mentioned in Sections 12.2 and 12.3:

Arbeitsgemeinschaft Messwertaufnehmer e.V. (AMA)
Linprunstrasse 49
80335 München
Germany

CENELEC Central Secretariat
Rue de Strassart 35
B-1050 Bruxelles
Belgium
Telephone: + 32-2/519-6871
Fax: + 32-2/519-6919
Telex: 26257
Teletex: 206/2210097 = CENCEL

DIN Deutsches Institut für Normung e.V.
Burggrafenstrasse 6
10787 Berlin
Germany

DKE Deutsche Elektrotechnische Kommission im DIN und VDE
Stresemannallee 15
60596 Frankfurt/Main
Germany

Deutsche Physikalische Gesellschaft e.V. (DPG)
Hauptstrasse 5
53604 Bad Honnef
Germany

Deutscher Verband
technisch-wissenschaftlicher Vereine (DVT)
Graf-Recke-Strasse 84
Postfach 10 11 39
40239 Düsseldorf
Germany

Association des Exploitants d'Equipment de Mesure,
de Régulation et d'Automatisme (EXERA)
BP Cerchar No 2
60550 Verneuil en Halatte
France

Gesellschaft für Informatik e.V. (GI)
Godesberger Allee 99
53175 Bonn
Germany

VDI/VDE-Gesellschaft Mess- und Automatisierungstechnik (GMA)
Graf-Recke-Strasse 84
Postfach 10 11 39
40239 Düsseldorf
Germany

GMD-Birlinghoven
Postfach 12 40

Schloss Birlinghoven
53757 Sankt Augustin
Germany

VDI Gesellschaft Verfahrenstechnik und Chemieingenieurwesen (GVC)
Graf-Recke-Strasse 84
Postfach 10 11 39
40239 Düsseldorf
Germany

Internationale Elektrotechnische Kommission (IEC)
3, rue de Varembé
1211 Geneva 20
Switzerland

IFAC Secretariat
Schlossplatz 12
A-2361 Laxenburg
Austria

IMEKO Secretariat
H-1371 Budapest
P.O. Box 457
Hungary

Instrument Society of America (ISA)
67 Alexander Drive
P.O. Box 12277
Research Triangle Park NC 27709
U.S.A.

Contact address in Europe:
ISA International
Avenue Marcel Thiry 204
1200 Brussels
Belgium

NAMUR Normenarbeitsgemeinschaft für Mess- und Regelungstechnik in der chemischen Industrie
51368 Leverkusen, Bayerwerk
IN-PLT, Geb. H 1
Germany

SIREP International Instrument Users' Association
South Hill
Chislehurst, Kent BR 7 5 EH
England

Verband der Chemischen Industrie e.V. (VCI)
Karlstrasse 21
60329 Frankfurt
Germany

Verband Deutscher Elektrotechniker (VDE) e.V.
VDE-Haus, Stresemannallee 15
60596 Frankfurt/Main
Germany
Telephone: 069/6308-0

Verein Deutscher Ingenieure (VDI)
Graf-Recke-Strasse 84
Postfach 10 11 39
40239 Düsseldorf
Germany

Vereinigung Industrielle Kraftwirtschaft (VIK)
Ausschuss "Elektrotechnik"
Richard-Wagner-Strasse 41
45128 Essen
Germany

International Instrument Users' Association
WIB
Barronielaan 15
4818 PA Breda
Netherlands

Zentralverband Elektrotechnik- und Elektronikindustrie e.V. (ZVEI)
Stresemannallee 19
60596 Frankfurt/Main
Germany

12.4. Shows and Fairs

Shows and fairs in the field of process control engineering can be categorized into three major groups:

- Shows devoted to one specialty (e.g., process engineering, manufacturing engineering) and including process control engineering as a sideline. Significant here are ACHEMA for process engineering and the Hannover Messe Industrie for manufacturing engineering.
- Shows oriented toward process control engineering as a whole. In the forefront of this group are INTERKAMA in

Düsseldorf, BIAS in Milan, MESUCORA in Paris, HET INSTRUMENT in Utrecht, ILMAC/INELTEC in Basel, INSTRURAMA in Brussels, C & I in Birmingham, and the ISA Show in the United States.

– Shows that concentrate on a single area of process control engineering, such as Sensor in Nuremberg and Systec, Systems, Laser, and Analytica, all in Munich.

The alphabetical listing that follows is certainly not complete but does give more information on the shows mentioned above, citing the emphasis and a contact address for each.

ACHEMA – International Meeting for Chemical Technology and Biotechnology; includes a congress in the field of chemistry, with special emphasis on process control engineering

Areas of concentration: research and innovation; measurement and control; process control engineering; materials science and materials testing; laboratory and analytical technology; mechanical and thermal processes; packaging and storage technology; safety engineering; occupational safety and health; pumps, compressors, and fittings; biotechnology; environmental protection

Contact address:

DECHEMA
Deutsche Gesellschaft für Chemisches Apparatewesen, Chemische Technik und Biotechnologie
Postfach 15 01 04
60061 Frankfurt/Main
Germany

Analytica – International Technical Show for Biochemical and Instrumental Analytical Technique, with International Meeting

Technical show focusing on analytical laboratory techniques

Areas of concentration: analytical instruments and methods, laboratory technique and apparatus, laboratory data acquisition and processing, industrial chemistry, industrial analytic technique, materials testing, quality control, environmental analysis, biotechnology, medical diagnostics

Contact address:

Münchener Messe- und Ausstellungsgesellschaft mbH
Postfach 12 10 09
80034 München
Germany

BIAS – International Show and Congress in the Field of Automation, Instrumentation, and Microelectronics

Areas of concentration: process computers and process control systems, measuring instruments, controllers, valves, sensors, transducers, transmitters, components and accessories, control devices for processing and fabrication operations, logistics, communications, bus systems, remote control, robotic systems

Contact address:

E.I.O.M.
BIAS Secretariat
Viale Premuda, 2
I-20129 Milano
Italy

CeBIT – World Center for Office, Information, TelecommunicationsInternational technical show for computer techniques

Areas of concentration: screen communications, office communications, text processing, computer technology, data processing, information technology, software, telecommunications, communications technology, communications engineering

Contact address:

Deutsche Messe AG
Messegelände
30521 Hannover
Germany

Control & Instrumentation – National technical show and congress for process control engineering, with emphasis on the chemical industry; held at Birmingham, England

Areas of concentration: process control engineering, computer systems, control technology, actuators, sensors, warning systems, analytical technology, weighing, fiber optics, testing and inspection, CAE/CAD/CAM software, displays and operator interfaces

Contact address:

MGB Exhibitions Ltd.
C & I Secretariat
Marlowe House
109 Station Road
Sidcup, Kent DA 15 7 ET
England

Hannover Messe Industrie – The world's largest industrial show, focusing chiefly on control and process drive engineering but with process control engineering as a secondary theme

Areas of concentration: electrical and electronics engineering; microelectronics; energy, air pollution, and environmental engineering; industrial optics and lasers; materials of construction; surface technology; process drive engineering; control engineering; installation engineering; industrial robots; material-flow engineering

Contact address:

Deutsche Messe AG
Messegelände
30521 Hannover
Germany

Het Instrument – Technical Show for Industrial Instrumentation

Dutch national show in the field of process control engineering; technical show for health care, science, and industry; held in Utrecht

Areas of concentration: devices and systems (sensors, measurement, testing and inspection, weighing, and control engineering); medical technology; laboratory technology (analysis); electronics; research and innovation

Contact address:

Sekretariat Het Instrument
Postfach 1 52
Birkenstraat 108
3760 AD Soest
Netherlands

ILMAC – International chemical show for laboratory technology, chemical engineering, and measurement and automation

Areas of concentration: analytical laboratory techniques; isotope and radiation chemistry; process engineering; measurement, control, and automation engineering; biotechnology; laboratory animal management; environmental protection; health care

Contact address:

Schweizer Mustermesse
ILMAC Secretariat
Messeplatz
CH-4021 Basel
Switzerland

INELTEC – International electrical and electronics engineering show

Areas of concentration: power supply and distribution; energy supply systems; communications; control systems; components; measurement, testing, and inspection; power electronics

Contact address:

Schweizer Mustermesse
INELTEC Secretariat
Messeplatz
CH-4021 Basel
Switzerland

INSTRURAMA – Technical Show for Metrology, Instrumentation, Control, and Automation

National show for the field of process control engineering and laboratory technology

Areas of concentration: apparatus and laboratory equipment; devices, systems, and components for process control and management of industrial automation and informatics in the fields of energy, production, refining, communications, environmental protection

Contact address:

UDIAS
INSTRURAMA Secretariat
Avenue Slegers 203
B-1200 Brussels
Belgium

INTERKAMA – Innovation Market for Measurement and Automation

Largest international show in the field of process control engineering, taking in all aspects of the field but focusing on research institutions, professional qualifications, exhibitor seminars, and congress

Areas of concentration: process control and monitoring systems; measurement and control engineering; sensor technology; analysis and weighing; actuators; communications; data processing; displays and output devices; components and accessories; tools; testing and inspection; services

Contact address:

Düsseldorfer Messegesellschaft
bmH – NOWEA-INTER
KAMA Referat
Postfach 32 02 03
40417 Düsseldorf
Germany

ISA Show – American technical show for process control engineering, with emphasis on a congress; held at various sites

Areas of concentration: measurement and automation technology (see, e.g., INTECH/ISA 91)

Contact address:

Instrument Society of America (ISA)
67 Alexander Drive
P.O. Box 12277
Research Triangle Park NC 27709
United States

LASER – Innovative and Applied Optoelectronics International Technical Show and International Congress

Specialty show with stress on laser technolgy

Areas of concentration: laser and laser systems technology for manufacturing, communications, environment, medicine, science, optronics, components, optical measurement and inspection, holography, pattern recognition, image processing, fiber optics, sensor technology, microwave engineering

Contact address:

Münchener Messe- und
Ausstellungsgesellschaft mbH
Postfach 12 10 09
80034 München
Germany

MESUCORA – International technical show and congress for measurement and control engineering, automation, and industrial data processing, as well as engineering

Areas of concentration: dimensional and mechanical measurements, electrical measurements, process control and automation devices, inspection and test equipment, industrial and scientific data processing, telecommunications systems, engineering, laboratory apparatus, analytical instruments, biological and biomedical metrology, optics and electrooptics, nuclear measurements, vacuum and cryotechnology, instructional and learning aids

Contact address:

Association MESUCORA
11, rue Hamelin
F-75783 Paris Cedex 16
France

SENSOR, Nürnberg – Specialty show focusing on sensor technology

Areas of concentration: sensors, measurement and testing, analytical and weighing technology

Contact address:

Nürnberg Messe GmbH
Messezentrum
90471 Nürnberg
Germany

SYSTEC – International Technical Show and International Congress for Computer Integration in Industry

Technical show focusing on CIM

Areas of concentration: research and experimentation, development, and design of CAD, CAE, and CAM software; consulting; services; system components; networks; workstations

Contact address:

Münchener Messe-
undAusstellungsgesellschaft mbH
Postfach 12 10 09
80043 München
Germany

SYSTEMS – Computers and Communications – International Technical Show and International Congress

Specialty show for computer technology (from PC to the networked data processing scene)

Areas of concentration: hardware, software, communications technology, systems components, user-oriented DP problem solving

Contact address:

Münchener Messe- und
Ausstellungsgesellschaft mbH
Postfach 121009
80043 München
Germany

13. Integration of Knowledge-Based Systems in Process Control Engineering

13.1. Principles

Information technology now permeates the field of process control engineering, not just in the form of computers, storage media, and communications devices, but also via concepts, meth-

ods, and techniques for structuring a wide variety of information. These concepts, methods, and techniques now mark every facet of the field, from the structuring of knowledge about processes to be automated, to the use of increasingly powerful microprocessors in control facilities (sensor/actuator systems, process monitoring and control systems), the structuring of the process control engineer's work during the design, construction, and operation of control facilities, and the availability of computer-aided design tools. Automation is no longer conceivable without them. They are discussed in depth in preceding chapters.

The more advanced technologies of knowledge-based and adaptive systems are already becoming important.

Expert systems, or knowlegde-based systems, arose in the field of AI (artificial intelligence) research, chiefly at American universities and research institutions. The first expert systems appeared in the early 1970s, but their high demand for computer resources prevented their becoming established at that time.

Knowledge-based systems in the sense of a software technology are distinguished by the architectural principle that application-oriented knowledge and processing logic are separate. This can be referred to as a software technology because the field incorporates novel programming methods and techniques. It can also be said that a "paradigm shift" has taken place in informatics, since procedural programming (supported by programming languages such as FORTRAN, PASCAL, and PL/1) has now been joined, on an equal footing, by other paradigms: logic programming (PROLOG = programming in logic), rule programming (OPS 5), functional programming (LISP = list processing), and objectoriented programming (FLAVOR, SMALLTALK, C++, etc.).

Integrated in a hybrid programming system, all these paradigms can be utilized in parallel, depending on what representation or type of knowledge appears most suitable.

The aim of knowledge-based expert systems is to explicitly make available, in a knowledge base, the knowledge of experts as embodied in standards, regulations, and guidelines, but also the knowledge they have gained through experience.

Only later was it recognized that the simple, application-neutral processing logic of "inference machines" (or inference engines) is insufficient for handling certain types of problem.

In addition to the representation of knowledge, an ontological analysis of the problem type is needed. Thus another problem-solving strategy comes into play as used for example for seeking errors in a technical system [834] (diagnostic strategies) or for designing or planning of a technical system. The basis for the latter activity is "planning strategies," which are generally top-down, beginning with functional requirements and proceeding through design alternatives to engineering specifications.

Diagnostic strategies are relatively well understood today, and special expert-systems tools employing them are available. Planning strategies are not so well developed. A point of importance in the problem of planning is "nonmonotonic inference": During planning or design, the planner may find that an approach already adopted is not leading to success, so that design decisions must be withdrawn. Expert systems must be able to keep up with this nonmonotonic decision-making process.

The crucial question in expert systems is still how to acquire experience. In contrast to encoded technical knowledge, which is formalized, standardized, documented, and based on a consensus in the technical world, experience is not formalized, not standardized, not documented, and often not a matter of consensus. Nor can knowledge acquisition be approached by methods of classical systems analysis, since it is concerned with the understanding of mental processes.

For a certain class of problems, adaptive systems or neural networks offer a solution.

Neural networks are now being used successfully for pattern recognition and classification. Pattern recognition refers not just to the typical recognition of images, handwriting, and sounds, but also to the recognition of relationships between physical, chemical, and apparatus-specific quantities.

A neural network is a large, highly connected system of simple processors. Whereas the human brain consists of ca. 10^{12} neurons, artificial neural networks today have few neurons, generally no more than 100.

In such a network, a neuron receives inputs from several others. Only when the sum of the weighted input signals exceeds a threshold value does the neuron transmita signal to the neurons connected to it.

The "programming" of such a network, also called learning, consists of modifying the weights of the connections inside the network such that a pattern presented at the input causes the desired response at the output. One of many programming methods is to modify the internal weights on the basis of the difference between the actual output and the nominal output. When the learning phase is complete, the neural network can interpret new patterns correctly, albeit with a residual error.

New network designer's art consists of finding a design, a learning rule, and suitable representative examples so that an optimal neural network will result at the end of the learning phase.

Another approach to dealing with uncertainty and indistinctness is fuzzy logic. This technique, which has been brought to industrial maturity in Japan, is based on the mathematical theory of fuzzy sets.

In classical logic, a proposition can take only the values "true" or "false," 0 or 1. In fuzzy logic, a proposition or "linguistic variable" is said to have such-and-such percent of value *a* and such-and-such percent of value *b*.

For example, consider pressure as the variable. In an engineering system, a process control engineer might say that the pressure can take on qualitative values such as very low, low, moderte, high, and very high. Now a given pressure – whereby "pressure" is now understood as a linguistic variable in fuzzy logic – can be characterized by the degree to which it belongs to the set of "high" pressures. This is expressed by a numerical value between 0 and 1. Zero means that the pressure does not belong to the set of high pressures at all, while 1 means that it belongs to that set entirely.

A variable might be located, for example, in an intermediate range from high to very high. It no longer takes on a unique value but can be described by, for example, the expression "pressure = (very low = 0.0, low = 0.0, moderate = 0.0, high = 0.3, very high = 0.8)."

Astonishingly, fuzzy logic has found broad applications in control engineering. This should not, however, be surprising, because many quantities cannot be measured exactly and noise of various kinds is often present. A study performed by the Frauenhofer Gesellschaft gives a comprehensive view of the potential applications of fuzzy control, in which an extensive literature collection can be found [847].

Fuzzy-logic control systems in practice often show more stable behavior than conventional methods. Interesting developments have arisen through the combination of classical deterministic control algorithms with fuzzy logic. Combinations with neural networks make is possible to gain the advantages of both approaches. Another technique for finding and retrieving information involves "hypertext/hypermedia" systems. The basis for hypertext is that electronically stored knowledge bases are accessed not with formal query languages or commands but by associative searching in a network consisting of information nodes and arbitrary links between these nodes. Employing the principle of direct manipulation, the user follows a train of associations by clicking the mouse on interesting objects or interesting passages that appear in a document. Nonlinearity is the guiding principle of hypertext. The freedom to navigate within such a network can also lead to disorientation. Tools must therefore aid the user in moving around information space.

The technique of hypertext systems, often combined with object-oriented data bases, is used particularly for technical documentation, standards, and regulations.

Hypermedia systems result when the information nodes are not exclusively text nodes but can also offer graphics, audio, animation, and video sequences.

To summarize, information-oriented process control engineering is already fully operational with no end to its development in sight. This quick review of novel software technologies suggests that the immediate future will see remarkable new advances in information processing. This trend has led to many seminars and tutorials, with special mention being made here of the tutorials at the GMA Congress 1993, in which fuzzy control, neural nets modelling dynamic systems, ER/SADT, and information-oriented modelling concepts were reported in detail.

13.2. Knowledge-Based Approach

Introduction. The ideal structure of a process control system is presented in Chapter 7. In Figure 254, the "remote" components include an AI (artificial intelligence) computer, which forms the hardware basis for knowledge-based systems. In process control engineering, the latter term is understood as referring to a combination of conventional process automation approaches and AI methods.

The need for knowledge-based systems using AI computers in process control engineering arises from the development of the object being automated, that is, the processing system.

For example, the continuing trend toward more intensive process control – a consequence of technological progress in process, manufacturing, and automation engineering – has led to a number of qualitatively new and distinct features in modern production facilities. These include large product throughputs, extreme process parameters, operation up to the stability limits, high parametric sensitivity, direct coupling of some process units without intermediate storage, and a high degree of material, energetic, and informational interconnection between production areas and operational divisions. The objectives of process control have come to include process securing (enhancement of process reliability), plant availability, process efficiency, product quality, production flexibility, and environmental safety. For these reasons, complex control tasks at higher levels are steadily gaining in importance for process automation. Most of these tasks involve the solution of the following problems [848]:

- The complexity and high dimensionality of control tasks (more than five control variables that can be freely selected)
- The incomplete observability and the nonlinearity of the subsystems
- Compliance with complicated side conditions (including a number of logical conditions)

The classical methods of modeling and control often fail on grounds of expense or because of parameter identification problems or real-time requirements. Generally, higher-level control tasks can be accomplished only when their complexity and dimensionality are reduced by situation recognition techniques and when experiential knowledge (to delimit the solution domain) is combined with situational models (to define the solution point).

In the context of knowledge-based systems for process control engineering, the principal AI elements are expert systems, artificial neural networks, and fuzzy systems. Expert systems and fuzzy systems (both employing methods of mathematical logic) are similar in their information structure. They are used as rule-based systems in process automation, in both open loops (consultative systems) and in closed control loops. The sole difference is that expert systems employ "distinct" knowledge (precisely measured process signals, unambiguous and either/or rules) while fuzzy systems employ "indistinct" knowledge (membership functions, fuzzy rules). From the applications standpoint, artificial neural networks are very similar to statistical methods of process modeling.

This section deals mainly with the use of expert systems and neural networks. AI methods are just beginning to find wide use in process control engineering [848].

Rule-Based Systems. The main types of rule-based systems used in process control engineering are expert systems and fuzzy systems. These seek to imitate the process of human thought on a phenomenological level.

If these systems are to be used as control elements in process control engineering, it is necessary to analyze the process employed by the operator of a production plant in solving problems. Figure 430 shows the subtasks and the solution process for the example of a reaction to disturbances in the process plant.

An analysis of this problem-solving process shows the following:

- The solution process involves the steps of diagnosis, optimization, and planning. Each step is separated from the next by a simulation phase. In this way, solutions are validated on simulation models, which may be both heuristic models and, often, mathematical models "in the operator's head" (thought models). These simulation techniques, whether practiced consciously or not, form part of a continuous iterative mental process, which ultimately leads to optimal

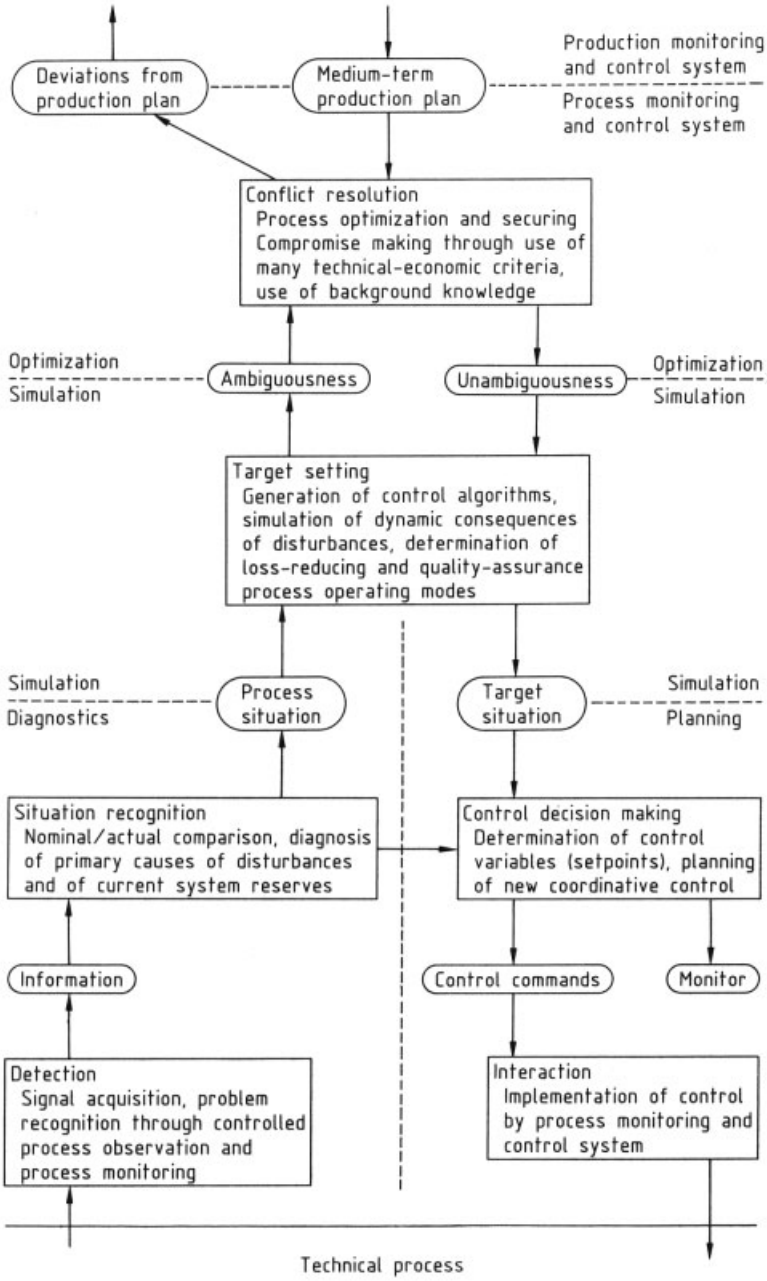

Figure 430. Tasks in knowledge-supported process control

conduct of the process. In other words, the suitability or optimality of the target situation is investigated by simulation and then corrected, leading to a corrected target situation, which in turn is simulated, and so on until the best result has been attained according to the operator's judgment.

– A nominal/actual comparison, the construction of control algorithms, and the determination of setpoint values, all as components of the diagnosis, simulation, and planning steps, can be performed with conventional control resources. The result is a higher degree of

automation and hence the accomplishment of process stabilization tasks.
- Conflict-resolution strategies in the context of optimization tasks can generally be implemented only by using artificial intelligence methods, namely the accomplishment of tasks relating to process optimization and securing. As a rule, knowledge-based systems here operate in consultative mode, with no increase in the degree of automation. If these tasks in all of the solution space are algebraically describable, however, it is possible to dispense with AI components.
- Knowledge-based systems cannot handle background knowledge (e.g., historical knowledge about the design process, knowledge of maintenance strategies or of the market situation). This part of the problem-solving process thus cannot be supported by knowledge-based process control engineering.

There are no fixed procedures for modeling the problem-solving process as a mental process. This is the fundamental distinction to the mathematical modeling of engineering processes, in which the application of balance equations (mass, enthalpy, and momentum balance) leads to an unambiguous and exact mathematical model.

Analyzing the problem-solving process from the standpoint of process control theory allows two classes of tasks to be identified: disturbance analysis and control synthesis.

Disturbance analysis encompasses solving the following tasks:

- Determination of primary casues of disturbances: Disturbances arise through slow parameter drift due to wear, contamination, and corrosion, or through sudden structural changes due to the failure of subsystems.
- Determination of secondary consequences of disturbances: Frequently occurring secondary effects are bottlenecks in feedstocks, energy-transfer media, intermediate products, or recirculating streams in adjacent process sections or in downstream process steps, which (depending on the process dynamics) most often manifest themselves immediately but, ultimately, further increase the resulting production losses.

On the basis of the primary causes of disturbances, a simulation of the dynamic consequences of disturbances thus provides important supplemental information about the objectives of measures taken against disturbances.

The following tasks must be carried out in control synthesis:

- Identification of primary repair measures means deriving the repair measures by which the primary causes of disturbances can be directly remedied.
- Determination of current system reserves: Along with the medium-term removal of primary causes of disturbances, in the short term it is necessary to make a control decision that will bridge over the failure of the malfunctioning subsystems for the duration of the repair measures and thus minimize the resulting production losses.
- Generation of a provisional operating regime: On the basis of the current system reserves, a new process regime is generated that will bridge over the system failure for the time required to remedy it and thus minimize the resulting production losses.
- Restoration of operation: Once the primary causes of disturbances have been remedied, the system must be restored to its normal state. This also requires for coordination of all subsystems.
- Process optimization: Current system reserves, as well as feedstock and energy resources, are deployed such that a technical-economic criterion takes on an optimum value while compliance with technical and economic constraints is maintained.

In what follows, some theoretical principles of rule-based systems are presented. These are essentially based on mathematical logic.

Neural Networks. In contrast to expert systems, artificial neural networks imitate the process of human thought at a biological level [848, 849]. The aim is a functional simulation of the human brain or of the human nervous system, which comprises some 25 billion nerve cells (neurons). Figure 431 shows a single neuron and the way in which neurons are interconnected.

A neuron receives its input information via the dendrites, which are connected by synapses

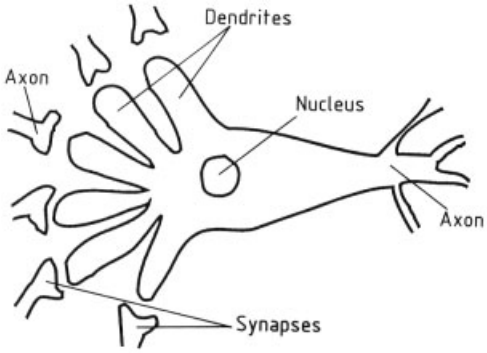

Figure 431. A biological neuron

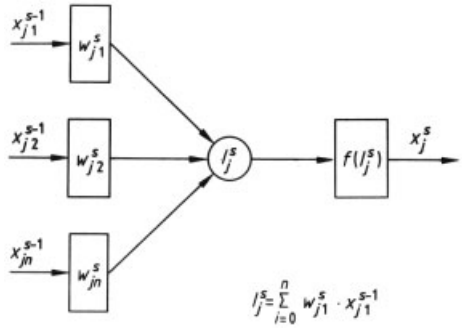

Figure 432. Functional diagram of an artificial neuron

to the information outputs of other neurons. Each neuron has only a single information output, the axon, which may be connected to as many as 10^4 "downstream" neurons through dendrites. From the sum of the inputs, an output dependent on the state of the neuron is generated and passed on to the downstream neurons by way of the axon. Biological information is thus transported throughout the network (nervous system), whereby it may also be modified. A significant feature of the process is the massive parallelism of information transport and processing. The above properties of neurons allow an abstract model of an artificial neuron (processing element) to be devised, which is described by the following rules (see Fig. 432):

- A neuron can adopt either of two states, the resting state and the action state
- A neuron j in layer s has multiple inputs $x_{ji}^{s-1}, \ldots, x_{jn}^{s-1}$ (synaptic connections) and one output (axon) x_j^s
- The output of the neuron is connected to the inputs of other neurons, and in this way a neural network is formed
- Some inputs and outputs of the neurons are connected to the outside world
- A neuron enters the action (excited) state when a sufficient number of its inputs have been excited

In this discussion,

x_{js}^{s} denotes the instantaneous output states of the j-th neuron of layer s,
w_{ji}^{s} the weightings of the connections of the i-th neuron in layer $s-1$ to the j-th neuron in layer s, and
I_j^s the weighted sum of the inputs of the j-th neuron in layer s.

The last rule, in particular, implies that the state of an artificial neural network depends only on the state of its elements. The individual neurons thus operate independently of one another.

For process control engineering the features of interest are the basic functions of neural networks (pattern association, pattern reconstruction). Here "patterns" are vectors (e.g., sets of measurements used to characterize a plant state or to assess product quality).

Three types of processing elements (units) are distinguished in an artificial neural network:

- Input units, which acquire their input information from the surroundings of the network (e.g., sensors) and are "visible" from outside the system
- Hidden units, which process incoming information and are "invisible" from outside the system
- Output units, which deliver output information to the surroundings of the network (e.g., actuators)

The individual units are grouped by function into layers, which are interconnected in a manner that depends on the problem. Figure 433 shows a typical multilayer network. Such networks are used principally for classification tasks.

Note that the behavior of the artificial neural network is strongly influenced by the network structure along with other factors.

A distinction is often made between the normal working phase (recall phase) and the learning or training phase of a neural network. Before the network becomes able to carry out given tasks, it must be trained for those tasks. Learning may be supervised or unsupervised. The first

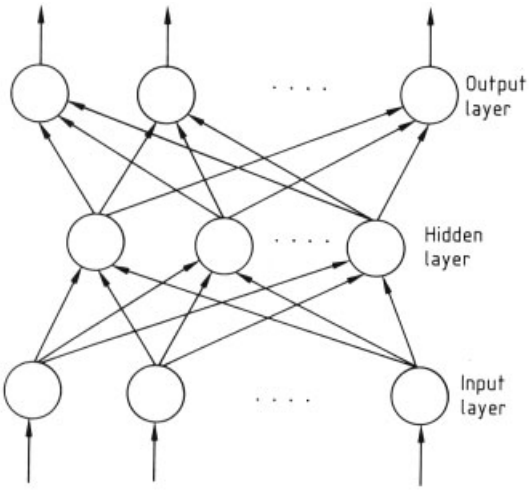

Figure 433. Multilayer network with one hidden layer

mode involves learning on the basis of specified examples. The back-propagation method is well known. In unsupervised learning, the network independently trains itself on the basis of rules in the input data (e.g., patterns).

In back-propagation, the output of the j-th element of layers s is determined by transformation with transfer function $f\left(I_j^s\right)$ (see Fig. 434 for an example).

$$x_j^s = f\left[\sum_{i=0}^{n}\left(w_{ji}^s \cdot x_i^{s-1}\right)\right] = f\left(I_j^s\right) \tag{13.2.1}$$

In back-propagation networks, the quality of the neural network is usually evaluated by a global error function E, which is needed in order to define local errors in the output layer and thus initiate the back-propagation of errors.

If it is assumed that vectors $\boldsymbol{i}$ and $\boldsymbol{d}$ are given to the network as the input and output patterns during the learning phase, while $\boldsymbol{o}$ represents the output vector calculated with the instantaneous weights, a global error function for the artificial neural network referred to the expected output $\boldsymbol{d}$ can be defined as

$$E = \frac{1}{2}\sum_{i=1}^{m}\left(\boldsymbol{d}_i - \boldsymbol{o}_i\right)^2 \tag{13.2.2}$$

where m is the number of components of vectors $\boldsymbol{d}$ and $\boldsymbol{o}$.

The aim of the learning process is to minimize the global error E of the network by selecting weights w_{ji}^s. Given the current set of weights

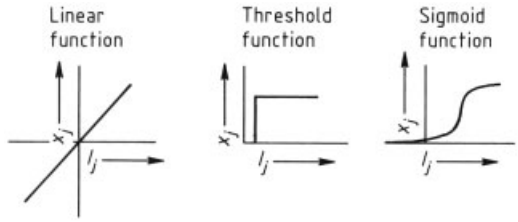

Figure 434. Possible transfer functions of a processing unit in a neural network

w_{ji}^s, the problem is to determine whether and by how much these weights must be changed. The formula

$$\Delta w_{ji}^s = -\alpha \cdot \frac{\delta E}{\delta w_{ji}^s} \tag{13.2.3}$$

is used for this purpose, where α is the learning coefficient.

Each weight is thus corrected in accordance with the magnitude and direction of the negative gradient on the surface of E. The partial derivative in Equation (13.2.3) is calculated as

$$\frac{\delta E}{\delta w_{ij}^s} = \frac{\delta E}{\delta I_j^s} \cdot \frac{\delta I^s}{\delta w_{ij}^s} \tag{13.2.4}$$

Setting

$$e_j^s = -\frac{\delta E}{\delta I_j^s} \tag{13.2.5}$$

and applying the chain rule twice gives the following relation between e^s and e^{s+1}, needed for the iterative computation:

$$e_j^s = f\left(I_j^s\right) \cdot \sum_{k=1}^{m}\left(e_k^{s+1} \cdot w_{kj}^{s+1}\right) \tag{13.2.6}$$

where m is the number of processing elements in layer $s+1$.

For the output layer

$$\begin{aligned} e_k^o &= -\frac{\delta E}{\delta I_k^o} = -\frac{\delta E}{\delta o_k} \cdot \frac{\delta o^k}{\delta I_k^o} \\ &= \left(d_k - o_k\right) \cdot f\left(I_k^o\right) \end{aligned} \tag{13.2.7}$$

Combining equations (13.2.3) and (13.2.4) and using Figure 432 gives

$$\Delta w_{ji}^s = x \cdot e_j^s \cdot x_i^{s-1} \tag{13.2.8}$$

Equation (13.2.8) represents the delta learning algorithm of the back-propagation method. Now the mechanism of back-propagation in the artificial neural network can be made explicit:

- Forward propagation of the inputs up to the output layer with the aid of Equation (13.2.1) and determination of the error at the output layer
- Successive correction of the weightings w_{ji} (with Eq. 13.2.8) during the process of back-propagation toward the input layer
- Recalculation of the output layer error with Equation (13.2.1) and recorrection of the w_{ji} with Equation (13.2.8), and so forth

Algorithmic Integration of Knowledge-Based Systems. Expert systems and neural networks have both advantages and disadvantages for process control (see Table 15).

Table 15. Comparison of expert systems and neural networks

System	Advantages	Disadvantages
Expert systems, fuzzy systems	use of surface knowledge	difficulties in knowledge acquisition (e.g., rules, membership functions)
Neural networks	ability to learn	difficulties in interpretation of solution (e.g., no possibility of explanation)

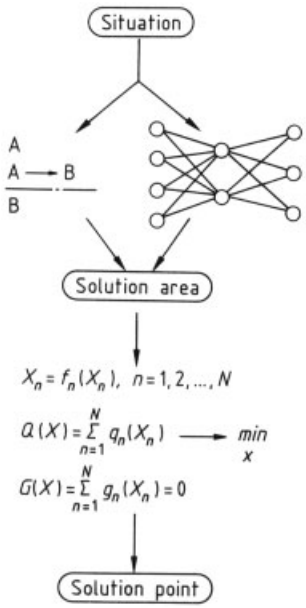

Figure 435. Algorithmic integration of knowledge-based systems

A useful way of compensating for the drawbacks is to employ both kinds of systems together. Figure 435 illustrates this point; the formulas presented there are explained in the text. This also enhances the software reliability. The artificial intelligence components determine the solution domain, while the conventional algebraic methods of process control give the solution point. From the standpoint of automation theory, the result is thus a localization of the process control task.

The mathematical models are significantly simplified in this way, because a mathematical description is needed only in the vicinity of the solution point. Linear models may well be adequate. Thus the control algorithms relating to process optimization, stabilization, and securing must be run under real-time-capable software, either on the process monitoring and control computer or on a special AI computer.

The definition of the solution domain by expert systems and neural networks (essentially in the context of classification tasks) is treated above. Now let us examine in more detail the application of algebraic methods to the determination of the solution point. It is assumed that a processing system can be broken down into subsystems (SS_n), interconnected by material and energy streams (see Figs. 436 and 437). This means simply that the phase model of production is employed in developing the software structure (see Chap. 2).

The following aspects are crucial in analyzing the system into subsystems:

- Topological (breakdown into technological elements, apparatus groups, etc.)
- Dynamic (breakdown into subsystems by dynamic properties, e.g., duration of dynamic transient processes, time lag)
- Combination of topological and dynamic aspects

The following notation is used:

$$\boldsymbol{X}_n = (U_n, Z_n, Y_n)$$

represents the vector of the process parameters of the n-th subsystem;

$$\dot{\boldsymbol{x}} = (X_1, X_2, ..., X_N)$$

the vector of the process parameters of the entire system;

$$\boldsymbol{X}_n = f_n(X_n)$$

the simplified mathematical model of the n-th subsystem;

$$G(\boldsymbol{X}) = 0$$

the coupling conditions of the subsystems; and

$$Q(\boldsymbol{X})$$

the goal function in process optimization, the Lyapunov function in process stabilization, and the negative first derivative of the Lyapunov function with respect to time in process reliability enhancement and quality assurance.

Thus the algorithmic structure of all process-control tasks is unified. As a consequence, marked simplifications have been achieved in the applications software (fewer software errors, better software maintenance).

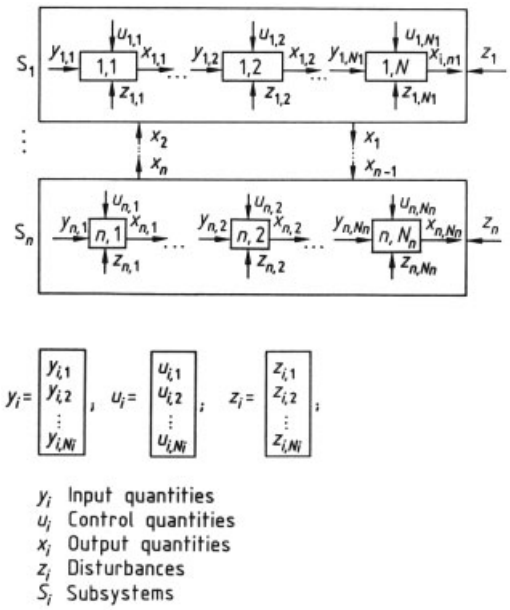

Figure 436. Decomposition of a technological system

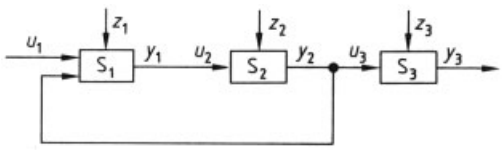

Figure 437. Example of a coupled total system

The use of the optimization condition $Q(\mathbf{X}) \rightarrow \min$ naturally requires some explanation.

With regard to the algorithmic integration of knowledge-based systems, conventional mathematical models are used as simulation systems for generating examples (in the case of neural networks) and of unmeasurable process information (in the case of expert systems). This point will be treated in more depth below.

Technical Integration of Knowledge-Based Systems. *General Principles of Integration.* Figure 438 shows the technical integration of expert systems and neural networks into the process monitoring and control plane. Knowledge-based process control can be represented in three levels:

- Elementary situation recognition with conventional methods of limit monitoring, deterministic or statistical class ordering, and trend analysis.
- Expert systems and neural networks combined with simulation systems based on mathematical models of the technological process. The objective is to identify the domain of the desired solution to the control problem.
- Conventional process control with the use of local or simplified mathematical models.

Integration of Expert Systems. The recommended procedure for technical integration of expert systems is to implement a real-time-capable expert system shell (expert system with empty knowledge base, i.e., without rule knowledge) on the control or AI computer, then fill it with concrete process-control knowledge in the course of knowledge acquisition (see Section 13.3). Such expert system shells must imitate the problem-solving process described here. A successful example is the PROCON expert system shell, which has been implemented in several cases [850–853]. This shell will be described in more detail in what follows.

System components: PROCON is subdivided into a development system and a runtime system with the following components (Fig. 439):

- Knowledge representation: object-oriented representation concept for explicit structuring of problem spaces

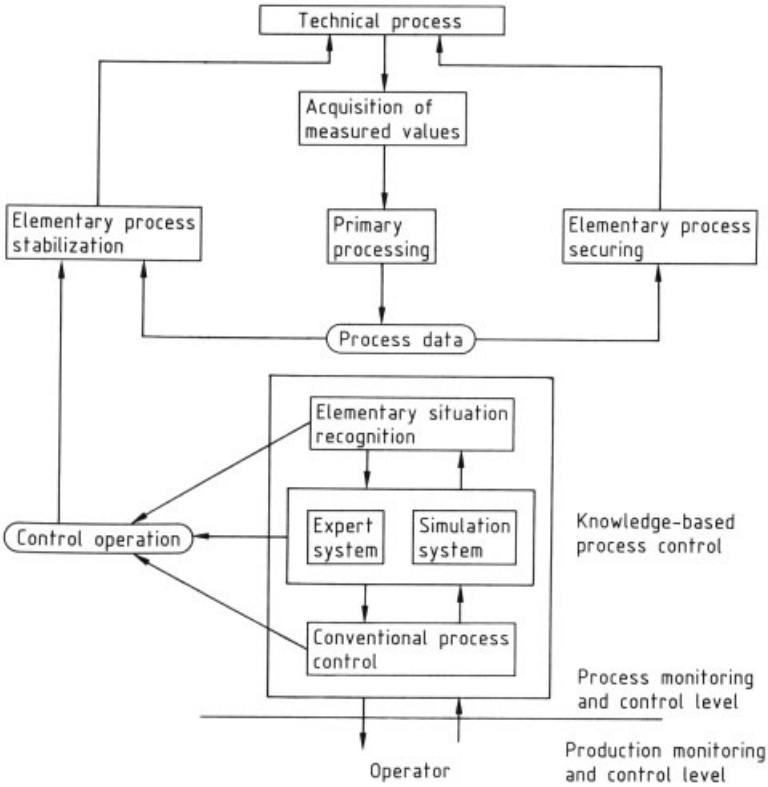

Figure 438. Technical integration of knowledge-based systems

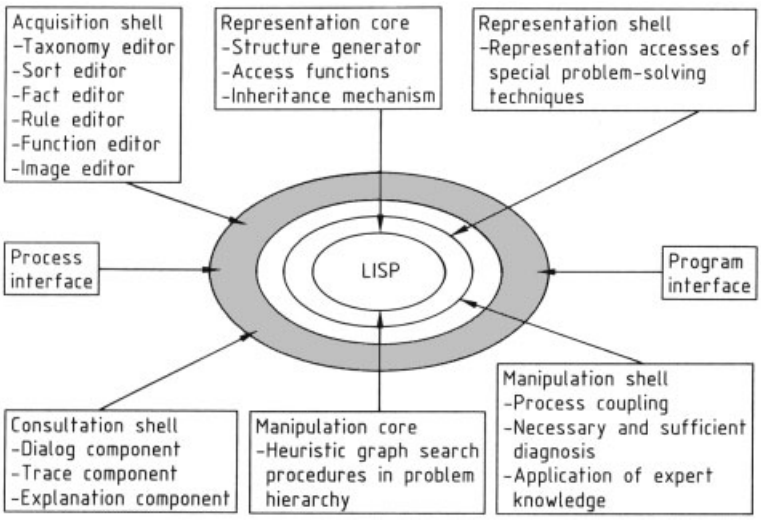

Figure 439. PROCON system components

- Knowledge manipulation: cooperative problem-solving processes for diagnosis and application of expert knowledge toward efficient delimitation of a solution
- Knowledge acquisition: problem-specific development environment for input, inspection, and transformation of knowledge into a runtime-efficient representation
- Knowledge consultation: user environment for verification and consultation of implemented knowledge bases

PROCON incorporates process, program, and dialog interfaces with the following features:

- Preparation of all needed process variables that can be acquired on-line, through coupling with the process control system
- Preparation of all needed process variables that can be acquired off-line, through interactive dialog with operating personnel
- Automatic interrogation of redundant information sources by the expert system whenever failures of measurement stations are reported
- Manual configuration of charge-dependent process parameters with parameter sets or in interactive dialog with operating personnel
- Rule-governed activation of external simulation and optimization programs via the program interface
- Manual, timed cyclical, or external consultation start

Requirements on the Knowledge Base. Applying expert systems to process control tasks means keeping pace with the process dynamics so that the integration of the knowledge base can be insured and critical reaction times met. Heavy modularization of expert knowledge into small knowledge units is dictated by both the need for adequate runtime efficiency of knowledge processing and the need for easy maintainability of the knowledge base as technological conditions change and operational experience is gathered. One useful modularization principle results from a division into disturbance and control classes. The set of all disturbances can be imaged on a set of disturbance classes that are qualitatively distinguished by primary causes and secondary consequences of disturbances. Similarly, all control actions can be placed in control classes on the basis of the disturbance class to be compensated and the available system reserves. If the two classifications are successively refined, a tree-like structure of disturbance and control classes is created. In the terminology of expert systems technology, this is referred to as a class hierarchy (taxonomy) of prototypical problem classes (objects). Descending through the problem hierarchy thus corresponds to localizing the current system disturbance and the resulting therapeutic control action. On a path from the root to a leaf of the taxonomy, each object encountered makes a contribution to the total solution of a problem. Each object contains a partial knowledge base made up of declarative (facts), functional (functions), and procedural (rules) knowledge specifications (Fig. 440).

Each object has access to the knowledge of all its predecessors in the problem hierarchy (inheritance):

- Declarative knowledge includes process information that can be automatically updated (values that can be acquired by on-line measurement, values that can be acquired by off-line laboratory work,etc.);user-configurable (technological) constants; interference variables derivable in the course of problem-solving; and representations parametrizable and external programs invokable by the rules of expert knowledge application (simulation and optimization models).
- Procedural knowledge. The following rule bases can be defined in each object:
- Rules for application of expert knowledge ("expertise" rules) are forward-chained production rules for the derivation of interferences from a given problem class (Fig. 440). They can derive interference variables, form solution texts to be displayed, parametrize representations for output, and invoke external programs for simulation and optimization.
- Diagnostic rules are classification rules for the derivation of so-called sympathy ratings for the objects in a problem hierarchy (Fig. 441). The expert system uses these ratings to decide what route to follow through the problem hierarchy, that is, what problem classes actually need to be examined in a concrete disturbance situation.

Problem Solving. The object-oriented structuring of the problem space permits efficient

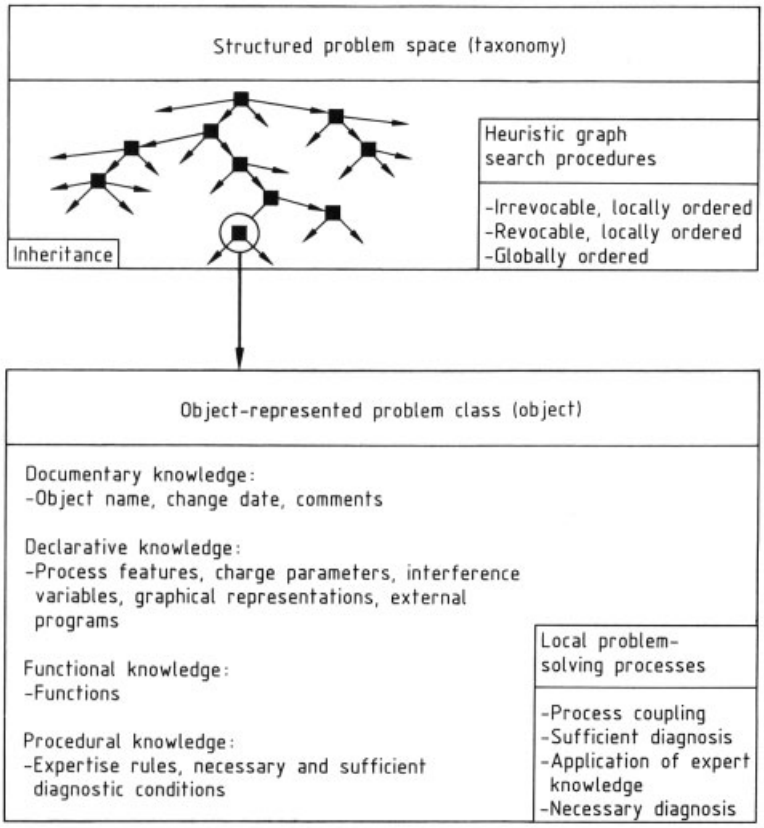

Figure 440. Concept of knowledge representation in PROCON

problem reduction through two problem-solving levels:

- Higher problem-solving level: Starting from the root of the problem hierarchy, a heuristic graph searching procedure determines a route through the structure of the problem space. In so doing, it establishes which problem classes actually need to be examined in a concrete disturbance situation, and in what order. The examination of a problem class is initiated by an activation message to its local problem-solving process.
- Lower problem-solving level: The local problem-solving of an activated object begins with a process coupling, then continues with an expertise and diagnostic process. Through the process coupling, all process features defined in the given object (with respective coupling parameters) are first updated. The expertise process (application of expert knowledge) that follows next evaluates the expertise rule base of the object in order to derive interferences from the problem class. Finally, the diagnostic process uses the diagnostic rule base of the object to rate its successors in the problem hierarchy.
  The remainder of the problem-solving process follows the rules of mathematical logic as described above.

Development Environment. The work of developing practically relevant knowledge bases is commonly divided among specialist experts (process engineers, automation engineers, expert operators) and, in the long term, makes possible the maintenance of the knowledge base as technological conditions change and operational experience is gathered. For this reason, special value attaches to interactive tools for the support of incremental, problem-specific knowledge acquisition. In PROCON, the acquisition component integrates the following aids in a self-contained development environment:

- Knowledge input: Graphically supported structure editor for taxonomy input. Mask editors for fact input, screen editors for rule and function definition.
- Knowledge inspection: Automatic search for inheritable knowledge specifications anywhere in the knowledge base; compressed explanation of defined knowledge contents of the active inheritance hierarchy during knowledge input.
- Knowledge transformation: Transformation of a problem-specific external knowledge representation into a runtime-efficient internal representation; formal-language or natural-language rule notation can be selected.
- Knowledge verification: Testing for syntactic closure and semantic consistency in the active inheritance hierarchy during rule translation; testing for adequacy and completeness in a comprehensive trace environment with case data management.

This development environment is the basis for knowledge acquisition as described in Section 13.3.

```
Rule D1

If
 the paste pump is not in operation &
 the level in the paste mixer is <5% &
 the level in VE1 is under 5% &
 the level in VE2 is under 5%

Then
 METER_DIST_1 90
 METER_DIST_2 50

Rule D2

If
 the paste valve is set to recycle &
 the level in the paste mixer is at Max-Alarm &
 the level in VE1 is over 5% &
 the level in VE2 is over 5%

Then
 METER_DIST_1 40
 METER_DIST_2 80
```

Figure 441. Examples of diagnostic rules (METER–DIST–1 and METER–DIST–2 are two lower-level objects of the current problem class)

Integration of Neural Networks [854]. In comparison with the conventional methods of process control engineering, neural information processing requires high-performance storage media as well as fast processing units (and possibly the use of special processors). Furthermore, the practical deployment of neural networks calls for ingenuity in network configuration as well as a training technique optimally suited to the application.

Hardware and Software Implementations. The recommended hardware support for neural network algorithms differs from the classical Von Neumann structure. The following solution approaches are of interest:

- The use of special neural processors to speed up serial operation. These units feature a separate floating-point processor, a supplemental cache memory, and complete primary storage capacity and employ the pipelining principle. Examples are the PC boards made by SAIC Corp.
- Multiple-instruction, multiple-data structures, which can be implemented with available "transputer" components; a number of software packages are already on the market for transputer boards.
- The use of vector processors. If these devices have a sufficient number of processor components, they make it possible to determine network activity layer by layer (i.e., to determine the activation vectors section by section with a high degree of concurrency).

At present, software simulators are used to convert the various network types to the computing and monitoring/control hardware currently used in process automation. Their implementation on general-purpose processors (e.g., Intel 8086) meets the requirements of flexibility, portability, and high integrability.

Only in exceptional cases can a dedicated implementation be recommended, especially since there are a number of powerful neural shells on the market. These systems generally have communications interfaces to the user and file levels and offer a wide range of distinct network types. Particularly important for integration into process control systems are runtime routines that enable customized, memory-efficient source code runtime modules to be integrated into the applications software. One example of such a software simulator is a development system called Neural Works Profes-

sional II, published by Neural Ware Inc. and intended for PC and workstation environments.

Note, however, that straight software implementations have the drawback of limited capacity, both in the training and in the implementation phase. A combination of dedicated hardware and dedicated software should therefore be preferred. The integration of parallel processing structures (e.g., in the form of transputers) into monitoring and control systems is anticipated.

13.3. Knowledge Engineering

Procedures and Methods. The creation of the knowledge base is the key step in the integration of knowledge-based systems into process control engineering. This procedure employs methods of knowledge engineering.

The term knowledge engineering denotes the totality of the processes involved in the design, implementation, and maintenance of expert systems (when expert system shells are used, this essentially means filling a knowledge base). Knowledge engineering thus continues throughout the life cycle of an expert system. The analogy to software engineering is clear. Knowledge engineering is the bottleneck in expert systems development (the same applies to automation engineering). Thus there is a parallel to the "software crisis" in the realm of conventional software.

Knowledge engineering, which can also be termed systems analysis, involves close collaboration among process control engineers, process engineers, and information specialists (Fig. 442).

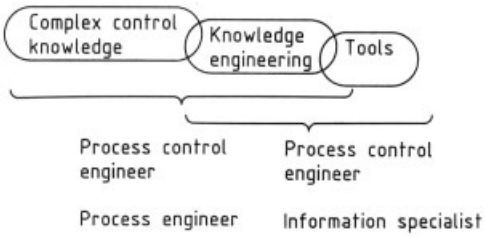

Figure 442. Interdisciplinary cooperation in knowledge engineering

In empirical social research, many years' experience and a broad methodological potential are available for the acquisition of knowledge and the traditional, nonformalized use of expert knowledge. Because these are also available to knowledge engineering, it is not surprising that the techniques currently in use for the acquisition of expert knowledge are largely derived from the methodological inventory of empirical social research. The techniques must, however, be modified to suit the needs of this special application area of process control engineering. Desirable, above all, is an appropriate combination of these techniques with the methodological potential built up over decades in theoretical and experimental process analysis. The knowledge to be incorporated in expert systems is generally found in the most diverse knowledge sources: technical manuals, operating instructions, case studies, and reports, but also data files and databases that may contain operational data for concrete process situations. In addition, personal experience and the knowledge of technical specialists (experts) is also sought. It is the last of these knowledge sources that is crucial in process automation. Moreover, a theoretical process analysis may also involve the creation of mathematical models that are then available as secondary knowledge sources.

Accordingly, a complex utilization of all knowledge sources – theoretical mathematical models, process data, and operator experience – is to be recommended as a way of reducing effort (see Fig. 443).

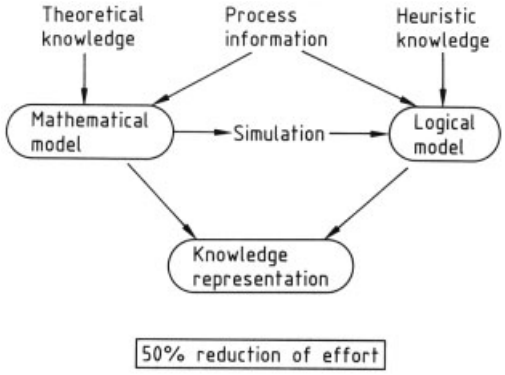

Figure 443. Reduction of effort by complex utilization of all knowledge sources

Formulated in general terms, the knowledge acquisition component of an expert system must support the transformation of this very heterogeneous knowledge to a homogeneous form, so

that it corresponds to the representation scheme defined by the structure of the knowledge base and by the interference component. Thus the knowledge acquisition component aids in the description of information or knowledge from the area of the user's tasks, building up a knowledge base from this information. The wide variety of knowledge sources makes this an exceedingly complicated problem, which entails a considerable effort in the case of an expert system for process automation. In present-day expert systems, various "manual" techniques are often employed for knowledge extraction; editors support only the final input to the knowledge base. When knowledge is prepared manually, however, relevant information is frequently lost or is not acquired at all. It is thus all the more important to develop knowledge acquisition components that will interactively support and assist the knowledge engineer (who is, as a rule, the process control engineer) in performing this task.

On the basis of these general aspects, a four-phase concept of knowledge engineering is recommended. This concept is oriented primarily toward process automation [855, 856]. The fundamental idea is to subdivide the iterative process of developing an expert system into the following phases (see Fig. 444):

- Definition
- Acquisition
- Operationalization
- Maintenance

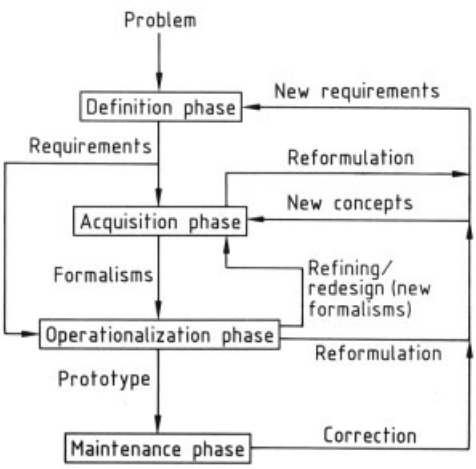

Figure 444. Four-phase concept

The definition phase provides a first global survey of the application domain. A specification of requirements must be created; that is, the problems for solution must be stated, their transformation into objectives and tasks of the expert system must be defined, and the feasibility of the expert system project must be investigated. In other words, the control task is established in this phase. The definition of the control task may be supported, for example, by comparing the disturbance amplitude D_z and frequency f_z for each control object. An analysis of this kind is done for the dominant disturbance in each control problem. Figure 445 shows that closed regions can be defined in the $D_z - f_z$ plane, in which process securing, optimization, and stabilization are to be performed.

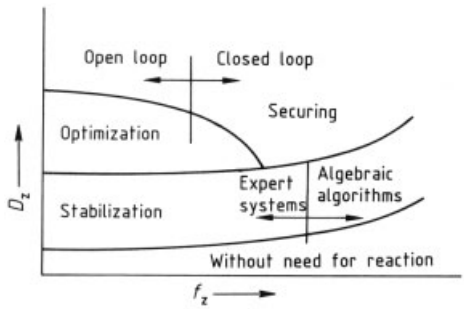

Figure 445. Relationship between automation tasks and properties of the disturbances

These process control components are defined as follows (see also Section 4.2):

- Process stabilization: Compensation of correctable disturbances with the aim of holding the process parameters constant at specified values
- Process optimization: Determination and setting of control variables so as to optimize a specified criterion while complying with specified restrictions
- Process securing: Compensation of uncorrectable disturbances with the aim of preventing unacceptable process states and product qualities or of minimizing their consequences

Figure 445 also shows, as a function of f_z, the trend in the use of expert systems and mathematical algorithms (conventional process control) as well as the use of consultative systems (open loop) and automatic systems (closed loop). In

this way, it is possible to identify a certain number of standard situations to which a particular control task applies. Expert systems are especially well-suited to the solution of process securing tasks, since as a rule there are no mathematical models for large disturbance amplitudes (emergencies), although a certain amount of empirical knowledge exists. The requirements worked out in the definition phase must be set forth in an appropriate form (specifications).

The acquisition phase includes analysis of disturbances, causal relationships, and possible control responses, as well as the aggregation of results into situation and control classes. The process analysis yields a formal model of the problem space in which the general questions to be answered in the definition phase should be detailed. The real difficulties in this phase lie in the design of knowledge bases. The knowledge should be structured to match the technology. It must directly mirror the operator's experience or mode of action and be determined chiefly by automation theory. In the operation phase, the knowledge is structured to match the system, as specified by the representation mechanisms that the expert system shell provides. This transformation of knowledge requires a knowledge engineer with experience and knowledge in many scientific disciplines.

Now let us take a closer look at "technology-like" structuring. The structure of the knowledge in this case is derived from the structure of the control object; a different form of structuring is generally meaningful in each disturbance class. Experience has shown that the technological system can be broken down into subsystems, each of which is controlled in autonomous fashion while all the subsystems are simultaneously coordinated. The decomposition of the entire technological system is based on the phase model of production (see Chap. 2). This principle can be illustrated for the example of process securing. The system is decomposed into a certain number of subsystems or levels, depending on the disturbance class; these subsystems show relevant differences in the dynamic properties relative to the dominant disturbances (Fig. 436). The upper (i-th) plane reacts with virtually no time lag to signals from the lower ($i+1$-th) plane. Each level contains a number of subsystems, which are connected to one another by energy, material, and information streams and exhibit similar dynamic properties.

On the basis of the structure of the control object, it is recommended that a hierarchical structure with multiple coordination levels be used for the process securing system. The subsystems should have the greatest possible autonomy (see Fig. 446). The subtasks of knowledge-based process control that are solved are the ones described above under "knowledge-based systems": determination of causes and consequences of disturbances, identification of control actions for the subsystems and for the system as a whole. One of the fundamental ideas in the technology-oriented structuring of knowledge is this: When an uncorrectable disturbance z_i takes place, the application of algorithm A_{1i} detects and compensates this disturbance (determination of primary causes of disturbance, derivation of primary repair measures). The "compensation" of a disturbance z_i means controlling a subsystem such that the given system is in a failure state for a time Δt shorter than some acceptable value Δt_{acc} ($\Delta t < \Delta t_{\mathrm{acc}}$). If this fails, coordination of two adjacent subsystems S_i and S_{i-1} is effected by using $A_{2(i-1)}$. First, the effect of the disturbed or failed subsystem S_i on S_{i-1} is identified (determination of secondary consequences of disturbance); second, the instantaneous dynamic reserves of S_{i-1} for compensating the failure of S_i are determined; third, a provisional regime (e.g., part-load or turndown regime) is generated. If even these steps do not compensate the disturbance, three subsystems are coordinated (for example, S_i, S_{i+1}, and S_{i+2} by using A_{3i}), then four subsystems, and so forth. This process is continued until the disturbance z_i has been compensated. Whenever a new subsystem is included, the above steps are repeated: identification of secondary consequences of disturbance, determination of instantaneous system reserves, and generation of a provisional regime. If successful compensation of the dangerous disturbance has not been achieved despite coordination of all subsystems, an emergency shutdown of the technological system is performed. In the context of knowledge-based process control, these coordinated shutdown modes can also be treated as provisional regimes. The normal regime is restored when the disturbance z_i is no longer present.

The operationalization phase includes the "system-like" structuring of the knowledge as well as steps to make the expert system functional.

Newly acquired knowledge is added during the maintenance phase (e.g., in the form of new "rules"). In addition, portions of the knowledge basis may have to be replaced altogether if changes take place in the technology. Expert systems are considerably more maintainable than conventional software systems, which is largely due to the fact that software errors do not necessitate complete reprogramming (software technology aspect). This leads to enhanced reliability of the process control system and lower maintenance costs.

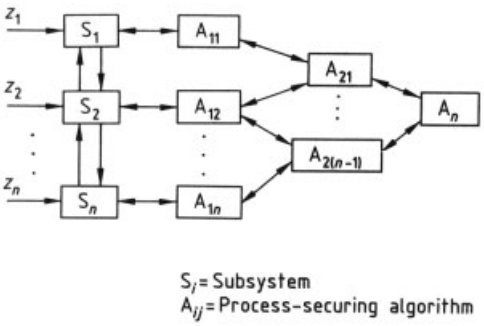

Figure 446. Structuring of a process securing system

Application Example. A chemical plant for the production of poly(ethylene terephthalate) fibers (trade name Grisuten) is equipped with a TDC 3000 process control system. The main tasks of the plant operator fall into the following categories:

- Process monitoring, the principal tasks being the detection and diagnosis of disturbances
- Process securing, i.e., the prevention of situations that would endanger the plant

In case of a disturbance, the operator must identify and implement a suitable reaction on the basis of available process information. Frequent operator interventions are needed in the following situations:

- Startup and shutdown
- Modifications of the operating regime
- Failure of plant sections
- Occurrence of critical process states that impair more than one plant section

Localized disturbances can generally be compensated by the existing automation system. In the above situations, the quality of operator decisions varies widely. Along with qualitatively good decisions by some operators, wrong decisions – some of them serious – are made by other, less experienced operators, especially when they are under psychological stress. Furthermore, the required specialist knowledge is commonly split up among several operators. The basis for an operator's decisions is essentially his own experience; theoretical background is less important. The main feedstocks are ethylene glycol and terephthalic acid (TPA). The process is continuous and can be broken down into the following phases (see also Fig. 447).

Metering/homogenizing. The feed materials are metered in fixed proportions into the paste mixer, where they are homogenized. Metering errors or inadequate homogenization of the feeds can lead to serious disturbances in the downstream system and to a major degradation of fiber quality.

Esterification. The reaction mixture is converted to diglycol terephthalate in a two-stage esterification reaction. The reaction is strongly endothermic, so heat has to be supplied. This process is carried out under vacuum, and the extracted vapors are distilled to recover ethylene glycol, which is recycled to the metering/homogenization stage. The crucial point here is to insure a minimum residence time of the reaction mixture in the reactors, to attain an appropriate degree of esterification.

Condensation. The diglycol terephthalate from the esterification stage is polycondensed to poly(ethylene terephthalate) in four condensation stages. The strongly endothermic reaction is carried out under vacuum with heating. The important quality criterion is that an appropriate degree of polycondensation must be attained after final condensation.

Spinning/finishing/extrusion. The polymer melt is spun and/or extruded in a melt-spinning stage. A number of treatments are performed on the spun filaments, after which they are ready for shipment. Every process step in the condensation and spinning/finishing/extrusion stages results in a significant increase in the viscosity of the reaction mixture.

The task of the expert system is to serve as a decision aid to the operator in carrying out

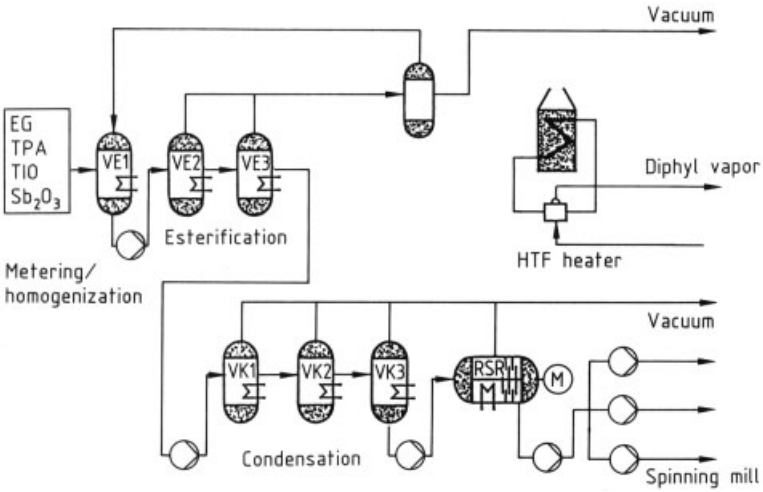

Figure 447. Process flowsheet
VE = esterification; VK = precondensation; RSR = ring-and-disk reactor; EG = ethylene glycol; TPA = terephthalic acid; HTF = heat-transfer fluid

process monitoring and securing tasks at the higher level. In particular, these tasks include:

Startup:
Cold startup of the chemical section of the plant (mean product temperature $< T_{\min}$)
Warm startup of the chemical section of the plant (mean product temperature $> T_{\min}$)
Failure of plant sections:
Failure of heating system for heat-transfer fluid (HTF)
Occurrence of critical process states:
Vacuum disturbance
Viscosity disturbance
Disturbance of mole ratio/COOH groups
Level and flow rate disturbances

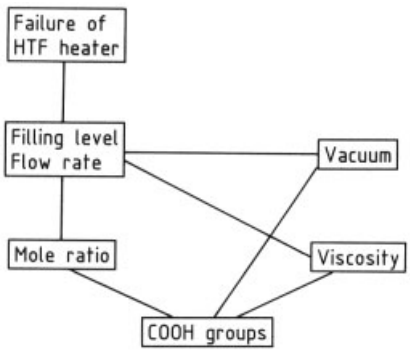

Figure 448. Disturbance classes and their interrelationships

The aim in reacting to disturbances is to insure high process reliability and product quality with the minimum reduction in plant capacity. Whenever possible, process data should be acquired from the process on-line, without placing extra demands on the operator. Interviews with experts have shown that all disturbance classes are interrelated in a complex manner (see Fig. 448). For this reason, in particular, they have not yet been completely brought under control.

Now let us examine the practical creation of a knowledge base more closely. The four-phase concept described above is applied to the disturbance class "HTF heater failure." The heat-transfer-fluid heating unit supplies heat to the strongly endothermic reactions in the VE and VK/RSR reactor groups (Fig. 447). A failure or loss of capacity in the HTF heater always results in severe disturbances (Fig. 448):

- Level and flow rate disturbances: Because of the rapidly cooling product lines, continuous flow of product between the individual reactors is not insured after just a short time. The consequences are unstable levels and residence-time disturbances.
- Vacuum disturbances: Unstable filling levels result in foaming in the reactors, which can result in product being carried into the vacuum system.
- Mole ratio disturbances: Residence time disturbances lead to an excessively high or low conversion of the reaction medium. The consequence is a severe decrease in fiber quality

and possible disturbances to the viscosity or COOH groups of the product.
- Viscosity disturbances: If the viscosity of the reaction process departs from the acceptable tolerance range. Both the processability and the quality of the product fall off drastically.
- COOH group disturbances: Because of excessive residence times, much of the feed ethylene glycol flashes off, leaving an excess of TPA and thus of COOH groups. As a result, major decreases in fiber quality occur.

For each of these disturbance classes, the proper responses to offset the disturbance (process securing) were ascertained by interviews with experts. The typical procedure for process securing (see above) was used as the basis for these interviews. Starting from the dynamic properties relevant in the case of an HTF heater failure, the technological system was broken down into subsystems S_1 (HTF heater) and S_2 (homogenizing, esterification, condensation, RSR reactor). The process securing system, in analogy to Figure 446, has the structure shown in Figure 449. This structure also describes the procedure used by the plant operator to compensate for the disturbance. The operator moves down a tree structure, level by level, until a solution to the process securing problem has been found.

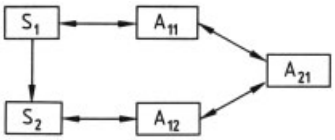

Figure 449. Structure of the process securing system upon a failure of the HTF heater

Figure 450 shows this procedure schematically. The respective decisions and actions are listed in accordance with the algorithms A_{ij} in Figure 438.

Figure 450 also demonstrates the structure of the technology-oriented knowledge base; reactions to possible downstream disturbances (level, vacuum, etc.) are merely indicated. Direct conversion to the expert system would have the consequence that many rules or knowledge items would have to be input more than once (e.g., 12 times for the cold and warm startup knowledge sub-base). This expresses the fact that several disturbance classes can be correlated with a control class. Large memory problems would result. Furthermore, searching problem space with such a tree would narrow down the problem very slowly, so that the solution process would not be efficient enough (real-time problem). For this reason, the knowledge base was structured in a situation-oriented fashion, as shown in Figure 451. The drawback to this structuring is that conclusions about possible consequent dis-turbances must be drawn at the very first level. Thus warm startup, cold startup, and other situations are shown not as consequent situations of an HTF heater failure but as general and equivalent disturbance classes. Technology-oriented structuring makes it possible to include the complete knowledge. Situation-oriented structuring, in contrast, improves the efficiency of knowledge processing for process securing problems in which the focus is on coordination of more than one processing unit.

In general, the final structure of the knowledge base (see Fig. 452) is governed by the procedure learned from the plant operator, by automation theory, and also by the structuring capabilities of the expert system. The reasons for this structure are as follows:

- The decision tree is simplified (fewer levels)
- Knowledge items are represented only once
- Memory requirements are reduced
- Reliability is improved
- The system gains in maintainability

As has been mentioned, each primary cause of disturbances leads to a different decomposition of the technological plant, which logically yields a different structure for the process securing system. For example, consider a vacuum disturbance, which can also be a secondary consequence of a disturbance. The coordinative control action to be generated (see Fig. 452) will depend on how widespread the disturbance is within the chemical section of the plant, which, dynamically and topologically, can be broken down into three subsystems: esterification (S_1), condensation (S_2), and ring-and-disk reactor (S_3). The method of knowledge acquisition is the same as described above for the case of an HTF heater failure.

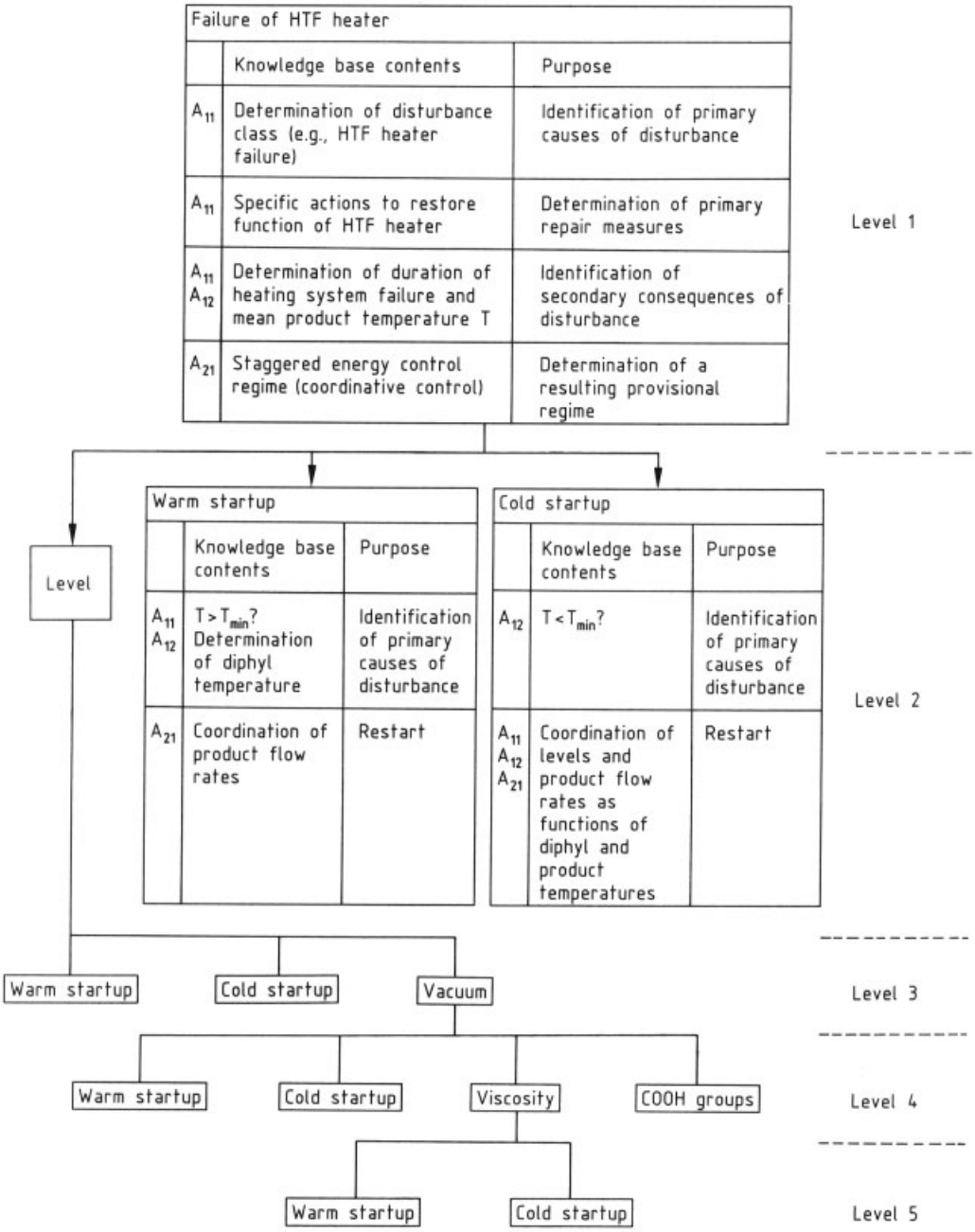

Figure 450. Technology-oriented knowledge base

The knowledge base for process securing (improvement of process reliability) in the chemical fiber plant includes, for one plant section, some 100 objects (problem classes) in six levels with about 220 process attributes, 40 interference variables as well as process constants, and 500 rules, of which 160 are diagnostic and 340 govern the application of expert knowledge. The expert system developed is coupled to the existing TDC 3000 process control system and

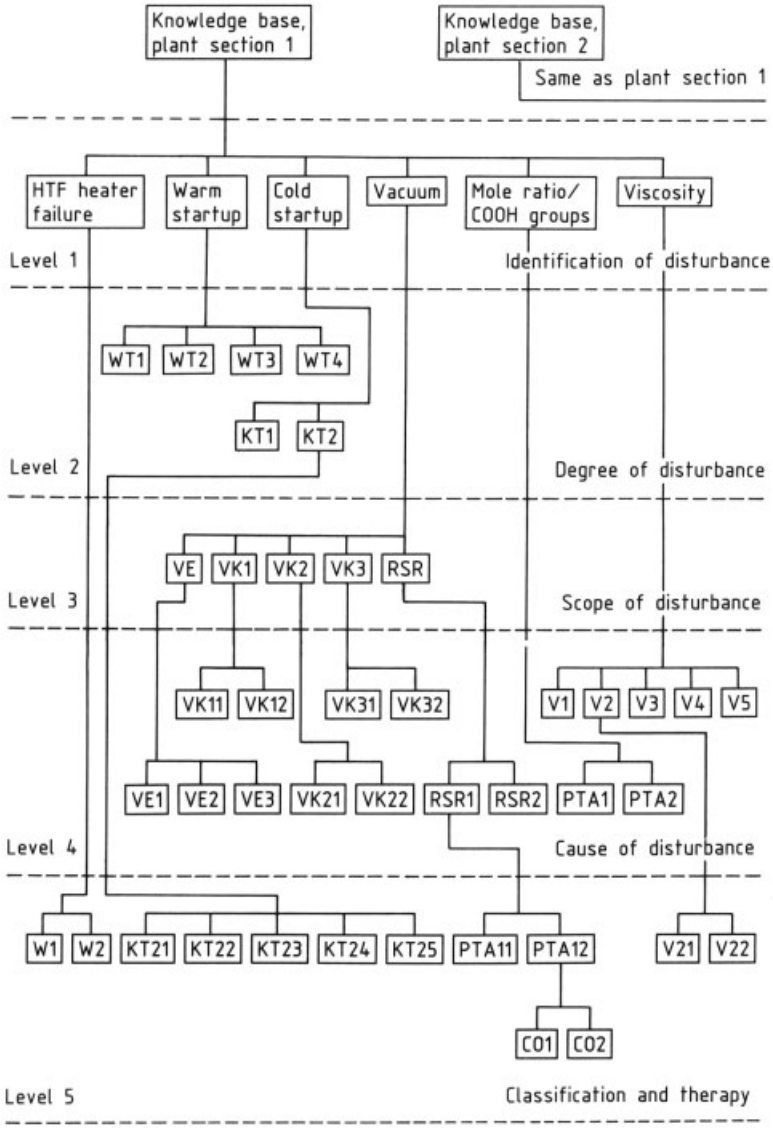

Figure 451. Structure of the knowledge base in the expert system

is in permanent service. Typical response times of the system are in the range of 4 – 10 s; roughly 70 –80 % of the time spent is for on-line acquisition of process data needed for expertise. The development extended over a period of about one year, during which time, 11 experts were interviewed. Figures 453, 454, 455 illustrate the consultation of the expert system by the plant operator.

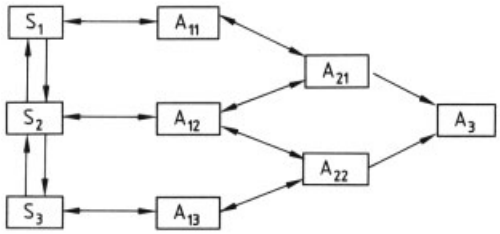

Figure 452. Structure of the process securing system upon a vacuum disturbance
S_1 = esterification; S_2 = condensation; S_3 = ring-and-disk reactor

Translate 2nd Editor Display

```
Rule E10

If
 the level in VK1<25% and
 the melt pump downstream of VE2 is not running

Then
 ACTION
 -Set precondensation startup parameters
  (activate explanation if desired) and

 ACTION-E
 -Startup parameters for precondensation
    Pressure        Level
  VK1 -A kPa        -A%
  VK2 -A kPa        -A%
  P_VK1, L_VK, P_VK2, L_VK
```

Representations
Units
Functions
Constants
Attributes

Bool
Flow rate
Integer
Number
Percent
Real
String
Temperature
Time

```
F12020     ...the primary diphyl flow rate
HB2081     ...paste pump 2081 is in order
HB2131     ...the melt pump downstream of VE2 is running
LRC2213    ...the level in VK1
LRC2302    ...the level in VK2
```

Figure 453. PROCON expert system shell, acquisition component, definition of a rule base for application of expert knowledge (expertise rule base)

Arguments Rule text State Results Secure Quit

```
Argument values   Evaluation of expertise rules in PROD_TEMP_4_TA1
Argument texts

Condition 1:
 LRC2213              =20%

Condition 2:
 HB2131               =FALSE

RULE E10
 Because
  the level in VK1<25% and
  the melt pump downstream of VE2 is not running
 RECOMMENDED ACTION=Set precondensation startup parameters
                    (activate explanation if desired)

 RECOMMENDED ACTION-E=Startup parameters for precondensation:

                        Pressure   Level
                 VK1    100 kPa    50.0%
                 VK2    120 kPa    50.0%
```

Knowledge base: GRISUTEN Test consultation

Figure 454. PROCON expert system shell, consultation component, expertise rule trace with inspection

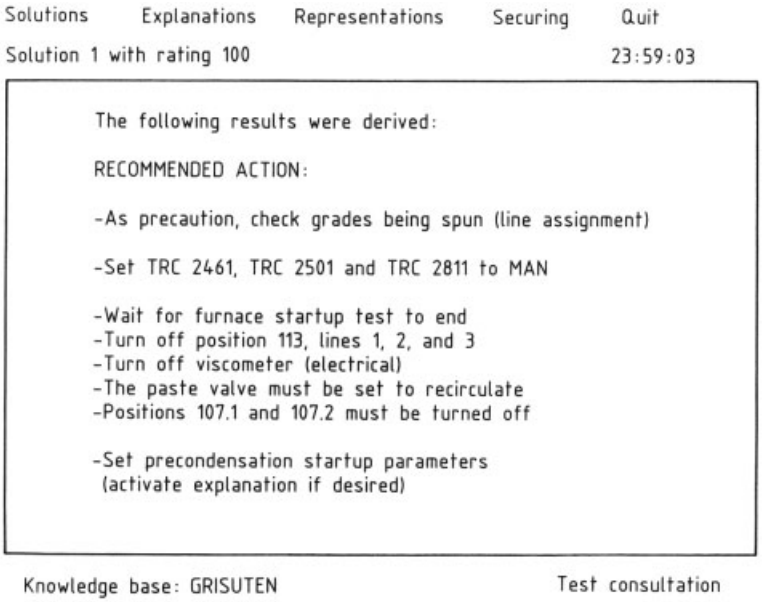

Figure 455. PROCON expert system shell, consultation component, presentation of results

14. Appendix

14.1. Glossary

Prefatory Note. During the period of transition to a holistic mode of analysis in process control engineering, the terminologies of individual specialties are not adapted to one another, and contradictions may even occur.

The development of an internally consistent, generally accepted terminological structure for process control engineering is a critical and ambitious task. It is the concern of the standardization bodies with jurisdiction over this field.

The following list of terms cannot and is not intended to anticipate such undertakings. The purpose here is to create a unified platform of concepts to serve as a basis for the treatment here. The list defines only a sampling of terms, the ones most important for understanding, and makes no claim of completeness.

The following abbreviations are used:

DF Definition
EP Explanation
LI Literature (references)
EX Example
SN Synonym
CL Classification

Abstraction

DF Abstraction denotes reducing complexity by neglecting all inessential properties or attributes. Which properties and attributes are essential and which can be regarded as inessential will depend primarily on the viewpoint adopted and on the question or problem stated. Both viewpoint and formulation of problem will determine the result of any abstraction.
There are several forms of abstraction:

DF 1 Class-forming abstraction
Class-forming abstraction takes at set of unrelated objects and orders the set in accordance with common properties and attributes. All objects that possess a common set of attributes can be assigned to one class. All objects that belong to this class are referred to as "instances" of the class.

EP 1 Abstraction can take place in several stages.
Associated with any abstraction hierarchy is a taxonomy of concepts or terms denoting the objects in the individual classes.

DF 2 Complex-forming abstraction
Related *objects* can be combined into a *system* with its own characteristic proper-

ties. When unlike objects are joined into a system, the diversity of properties of the individual objects is mapped onto a few external properties of the system. This procedure is called complex-forming abstraction.

SN 2 Special case: Aggregation

DF 3 Functional abstraction
When aggregation is perform, typically not just the complexity of the object but also the level (plane) of examination changes. With every aggregation step, the examination becomes more "abstract" [857]. The levels may be characterized, for example, as physical, operative, tactical, strategic, or intensional. Such a hierarchical structure is also referred to as functional abstraction.

Aggregation

DF Combination of related *objects* into a *system* having characteristic properties. Aggregation is a case of complex-forming abstraction. The objects examined are *classes* or *instances*.

EX By means of assembly, a collection of initially noncoherent objects such as wheels, body, engine, transmission, etc. becomes a system, in this case an automobile. The system exhibits its own characteristic properties, which can only be derived indirectly from the interaction of the element properties. For example, an engine is characterized by the following properties: number of cylinders, way in which valves are operated, number of valves, weight, displacement, etc. The transmission is characterized, among others, by the gear ratios. If the total system "automobile" is examined, the interaction of the properties of the individual elements generates a huge variety of overall system properties. From a given standpoint, however, only a certain subset of these will be of interest. For example, if an automobile is regarded as an element of a transportation system, its behavior can be characterized by a few properties, such as maximum number of passengers, maximum speed, fuel economy, and range. The properties of the individual objects from which the automobile was assembled are no longer of interest.

EP Aggregation forms complex objects by stepwise combination of simple objects.

SN Complex-forming abstraction

Attribute

DF A characteristic or feature of an object (see *Property*).

SN Property

Automatic

DF Performance of a task by a self-contained technical system without outside aid.

LI DIN 19 233

Basic Function

DF 1 Function of process control. Establishment of an independently set process property in a plant section.

DF 2 Functional unit with the capabilities of establishing certain process properties in a plant section. The functionality that a plant section possesses for intervention in the process is disjointly distributed over the basic functions. At a higher level of abstraction, the functionality of a plant section for process control is obtained by superposition of the functionalities of its basic functions.

DF 3 Software/hardware module with the capability of realizing the functionality of a basic function (DF 2) (see also *Functional Unit*).
The term "basic function" is also used with another meaning in process control engineering:

DF 4 In the sense of DIN 19 243, Part 1, basic functions are the smallest building blocks for the acquisition, processing, and output of analog and binary quantities. Here, basic functions include, for example, acquisition via analog input, BCD encoding, and the formation of a reference value.

DF 5 Synonymous with *group control*.

DF 6 Synonymous with *monitoring and control function*.

EP Here, "basic function" is exclusively used in the definitions given in DF 1 and DF 2.

Basic Functional Element

DF 1 Function, element of a *basic function* (DF 1).

DF 2 Functional unit or *functional element* (DF 2) from which a *basic function* (DF 2) is constructed.
In this sense, a basic functional element can be a process intervention, a discrete control device, a group control, or any *functional element* or *functional unit* that is used to construct a *basic function*.

DF 3 Software/hardware module with the capability of implementing a single basic functional element (DF 2) (see note under *Functional Unit*).

Breakdown (Analysis)

DF Elucidation of certain properties of an *object* through a system of simpler objects interacting in an ordered fashion.
The breakdown process is subject to the nesting principle.

EP The system is to be viewed as a model such that, for the question or problem stated, it can explain the external behavior on the basis of the interaction of system elements.
Quite different breakdowns may be obtained, depending on the question formulated and the model selected.

EX An automobile might be broken down
– from the assembly standpoint, into wheels, engine, body, etc.
– from the disposal standpoint, into plastic parts, metal parts, textiles, etc.

Call

DF One-time activation of an algorithm (better: the message that initiates the activation).

EX An analog controller operates continuously and need not be called; a discrete controller, on the other hand, operates under a cyclically (periodically) called algorithm.

Class

DF Group of objects characterized by a set of common attributes.

Command Mode

DF The command mode is a characteristic property of technical units of a command hierarchy. It indicates whether the unit in question receives its control values and commands from the plant operator ("manual" command mode) or from a higher-level technical unit ("automatic" command mode).

Connector

DF Connectors are subobjects of functional modules. Through them, the functional module mediates its message traffic with the surroundings. Each connector thus incorporates certain client and/or server properties of the module in a task-oriented fashion. A connector has an interface to the communications system, its own method for handling the message traffic, and an internal interface with access to the information held and used in the module.

EP Module inputs and outputs, as defined, for example, in the languages of IEC 1131, are special connector types. They serve primarily to handle "get" connections, in which the recipient, through an input, reads information from the output of another module. The IEC languages provide no possibility of handling more complex communications functions (e.g., semaphore-protected write) at the module surface. Separate "communications modules" must be defined for this purpose. With the introduction of connectors, these functions can be incorporated in the semantics of the module (component) language.

LI [858]

Control

The term "control" is used with five distinct meanings:

DF 1 A sequence of action: "Control" or "feedforward control" denotes an open sequence of action over a single transmission element. This meaning is contrasted with "feedback control," which denotes a closed sequence of action.

LI 1 DIN 19 226

DF 2 A binary operation: "Binary control" denotes binary processing of input signals yielding binary output signals.

LI 2 DIN 19 237

EP Here analog processing is contrasted with binary control. There are two forms of binary control: *logic control* and *sequence control*.

DF 3 A functional unit with the capability of performing all functions occurring on a control level (plane, layer). In the control system, it receives tasks from a functional unit located higher in the hierarchy, executes these tasks on the basis of a specified method or recipe, and concludes them by specifying appropriate instructions to the next lower level.

CL *Group control, plant section control, plant control, production control*, etc.

DF 4 Hardware/software modules that realize the functional unit of DF 3 (see note under *Functional Unit*).

DF 5 In a hardware context, a programmable logic controller (PLC).

Control Value

DF A control value is a time-varying setpoint of a controlled *process property*. The expression "controlled process property" signifies that there is a concrete *process control function* that (in normal operation) guides the actual value of the process property so as to track the nominal value in a specified fashion.

EP The control task can be carried out in various ways, for example

– by closed-loop control
– with event-controlled feedback
– by open-loop control

Decentralized Control System

DF 1 Totality of all technical apparatus that makes possible the function "monitoring and control".

DF 2 In a hardware sense, the monitoring and control system is part of the process monitoring and control hardware.

Degree of Automation

DF The degree of automation is the fraction of the process control tasks that is performed automatically by the monitoring and control system.

LI [859]

Device Control Function

DF *Function* for the control and monitoring of devices included in the monitoring and control system but not acting directly on the process. In the case of a sensor, for example, the following functions are carried out: select measuring range, initiate calibration, select place of measurement, etc. In the case of an actuator, the device control functions might include the selection of certain actuator characteristics (e.g., converter), the initiation of a calibration (e.g., metering scales), or the selection of a redundant resource (e.g., dual pump).

EP The functional unit *discrete control device* possesses the functionality to take over the device control function.

Discrete (Control) Device; Discrete Device Technology

DF *Functional unit* with the capability of controlling and monitoring an actuator or sensor. This functional unit generally contains both functional elements for controlling and monitoring the device itself and functional elements for the processing of useful information, for example, to output control commands or to accept and prepare measured values.In the actuator branch, the discrete device acts as a *group control*. Here, the task is to bring about certain process properties in accordance with the specified *control values*.

Element

DF *Object* that is in a well-defined relationship to other objects in the context of a *system* and that cannot or is not to be further decomposed in the context of the present analysis.

EP In contrast to an object in general, associated with an element per se are the relations (relationships, dependences) to the other elements in the system. Certain functions are transferred to an element in the context of the system.

EX A screw in a 100-pack is an object. In a certain system, the screw that holds two parts A and B together is an element of the system. In the system, the *function* "connect A and B" is transferred to this element.

SN System element

Entity

DF 1 A certain state of being of an existing thing; also the thing itself.

DF 2 In connection with object orientation, another name for an object that exists in reality or conceptually and is of essential importance to the problem under consideration. Entities thus form an excerpt of the real world that is necessary for the handling of the problem under consideration. Abstract concepts such as "function" can also be entities.

LI [860]

Entity – Relationship Model

DF The entity – relationship (E/R) model describes the *entities* of an analysis space together with the relations existing between these entities.

EP E/R models are suitable for describing static relations between entities (consists of . . . , is a . . . , is derived from . . .).

EP E/R models in general offer a graphical syntax for the intuitive representation of entities and their relations.

LI [861]

Execution Unit (see also Functional Unit)

DF Execution units are functional units that are capable of realizing a certain role. They are based on resources that possess the required functionality; spatially and logically, they are placed in a suitable layer; and they are configured so that they exhibit the special functionality needed for the role.

EP See *Role*.

SN 1 Device
Here the term "device" should, however, be understood as including units realized logically and in software.

SN 2 Here, execution units are referred to as functional units.

Framework Concept

DF Document describing the structure within which long-term development is to be accomplished.

Function

DF 1 Clearly defined task within a larger totality.
Functions are a separate category of objects.

EP 1 Note: A function is the task proper, and must be strictly distinguished from the capability (functionality) of a unit to perform a task.
In the context of the structure of a system, functions are mapped onto units (*functional units*) that possess the functionalities required for the functions. This transfer makes the units into elements of the technical system, which now implement the functions.
(See also *Functionality, Functional Unit, Element, Substance*).

EP 2 The term "function" is also used informally in relation to concepts that do not have to do with functions (see, e.g., *Functional Unit, Functional Element, Basic Function*; also the commonly used term "functional module" frequently refers to a program module).
Note: Although these expressions are inappropriate with regard to clear categorization of objects, they are not changed here but are used in accordance with common practice.

DF 2 In the process control engineering context, the only other meaning of "function" that is of any interest is the mathematical function. A function is a variable quantity whose value depends in an unambiguous way on the value of certain other quantities. (It is a mapping between algebraically structured sets.)

Functional Element
An ambivalent term denoting both the elements of functions and the elements of functional units.

DF 1 Element of a *function*.

EX 1 Display analog value, transform signal, link values, form limiting value, verify status signal, control valve, block motor, control, . . .

DF 2 Functional unit with elementary functionality.

EX 2 Functional elements, defined as the smallest units of functionality under consideration, include display elements, signal-transforming elements, linking elements, and limit pickups; or, specified from a technological standpoint, process monitoring and control station element, ba-

sic functional element, . . . , motor control, valve control, PID controller, . . .

EP Functional elements can be regarded as the bridge between the top-down analysis of functions and the bottom-up synthesis of functionality. A function must be broken down in such a way that it ultimately consists of functional elements (*Functions*, DF 1), each of which can be fully implemented by functional elements (*Functional Units*, DF 2).

Functionality

DF The capability of an object to perform certain tasks (*functions*).

EP Functionality denotes a fundamental capability of an object and in this sense is a resource. This resource can be utilized by employing the object as an element of a system. In the system context, certain tasks are assigned to the object (as an element of the system). In this way, a part of the functionality of the object is used.

EX A drive, as an object (*functional unit*), has the following capabilities: "can be at a stop," "can turn to the left," and "can turn to the right." The drive M 2 associated with roller R 4 of a roller conveyor has the tasks of turning to the right or of being at a stop. The capability of the object "drive M 2" expressed as "can turn to the left" is not utilized in this system environment.

CL Functionality denotes not an object but a set of properties of an object.

Functional Unit (see also *Execution Unit*)

DF An object with a certain required functionality; that is, an object having certain properties that are necessary for the implementation of the tasks to be performed.

CL Functional units are not functions but play the role of substances. Functional units have the character of resources and can be called on once and only once at any given time.

EX In the planning of a roller conveyor system (see the example under *Functionality*), the system is built up, say, from functional units such as drives, rollers, and so on. These are neither functions nor devices but postulated objects with well-defined functionality. In the planning of the roller conveyor system, these objects, viewed as available resources, are used as system elements; they are combined and linked in such a way that the roller conveyor system as a whole can fulfill the task allotted to it.

EP The definition of functional units is arbitrary. It would be possible to define any functional units with any functionality. It is, however, expedient to choose functional units in such a way that the set of properties (functionality) coincides with, or is a subset of, the property set of available substances. In the software realm, for example, this means that, in constructing a system, functional units may be used only if there are program modules corresponding to them.

Because of the correspondence between functional unit and object, synonymous expressions are used for the two even though this is not wholly correct (see *Function*).

Examples include the expressions "plant control," "discrete control device," "group control," and so forth.

An even more unfortunate situation arises when a synonymous term is used for the function, the functional unit, and the object. This happens with, for example, *basic functions* and *basic functional elements*. Metering is a function; there is a functional unit "metering device" for it; ultimately, there is a software module that implements the functionality of the metering device. In common usage, however, the expression "basic function" is applied to all three objects (see Section 4.5).

EP 1 Examples of the naming of functional units:

Classes of functional units include *production control, plant control, plant section control, group control, discrete device control, basic function*.

Types of functional units include metering device, motor, valve, level control, temperature measurement, reactor control, tower control, and so on.

Instances of functional units include (standard plant and apparatus coding)

motor V 13 R 4 K 12 and level control V 13 R 4 L 6.

EP 2 Here, the expression "functional unit" is also used for a subordinate concept, defined under *Execution Unit*.

Group Control

DF 1 *Functional unit* with the capability of establishing certain process properties in accordance with specified control values. The group control uses internal methods to calculate setpoint values, which it passes on to group controls or discrete control devices at the next lower level.

DF 2 Software/hardware module with the capability of implementing the *functional unit* "group control" (DF 1) (see note under *Functional Unit*).

DF 3 Functional unit with permanently defined, generally applicable control functionality.

LI [862]

Heuristics

DF Doctrine of the methodical acquisition of new knowledge that leads to acceptable results even though not guaranteeing a definite solution of a problem (since there is no proof of convergence). Heuristic procedures are frequently used in the engineering sciences and in operations research.

Inspection

DF Determination and evaluation of the actual state of a plant or apparatus.

LI DIN 31 051 (1985)

Instance

DF 1 All objects that belong to a class are referred to as instances of the class.

EP 1 This definition is general. In a hierarchy of classes, each class can be regarded as an instance of the next higher class. The definition makes no restriction as to the form of class formation.

EX 1 The electric motor is an instance of the class of motors. The asynchronous motor is an instance of the class of electric motors.

A certrain model of motor (manufacturer X, model Z) is an instance of the class of asynchronous motors. The motor with serial number XXX is an instance of a model of motor.

DF 2 Objects representing the lowermost level of a hierarchy of classes. An instance in this meaning cannot be further instantiated.

EX 2 The motor with serial number XXX is an instance in this sense. Every feature, at every time, has a certain expression (serial number, power, construction, voltage, explosion protection, instantaneous rotation angle and rotation speed, . . .).

EP 2 The instance is itself an *abstraction* of the concrete object and takes account only of the properties needed for classification and identification. Other properties of the real object do not become properties of the instance.

EP 2 Note: The instance of a *functional unit*, even though it has many of the same *properties* (*functionality*), is not identical with the instance of the corresponding physical object. They are correlated only temporarily, for the time of installation. If the physical object is removed (for example, taken into the workshop), the two objects may under some circumstances have different relationships.

Instruction Acceptor

DF The instruction acceptor is the connector through which a functional module receives a control instruction.

EP Each functional module can have more than one instruction acceptor.

The control instruction is a command specification with a write character. The instruction acceptor must accept the control instruction immediately after its arrival, independently of cyclic module processing. It checks the content of the control instruction from a formal and a semantic standpoint, sends a reply to the operating system; on acceptance of the command, it writes the new values to the module as effective values.

SN Operator control, control instruction.

Logic Control

DF Control using Boolean logical operations to assign certain output signal states to the input signal states.

LI DIN 19 237

Maintenance

DF 1 Measures taken to keep a plant in functioning condition.

DF 2 Measures to insure the nominal state of a plant.

LI DIN 31 051 (1985)

Model

DF Representation of the important relations between the properties of an object.

SN Model system.

Module

DF In general, a module is an object that forms a closed unit and can be manipulated as such.

EP The concept of "module" should not be further delimited. Subtle distinctions to other concepts, such as block, building block, unit, and so forth, based on the degree of complexity or other attributes, are not meaningful.

SN Apparatus module: device unit
Software module: functional module
Functional module: functional unit

Nesting Principle

DF Specification relating to the *aggregation* of *functional units*. *Instances* of *functional units* are resources that are available once. In the framework of an overall structure, they can be used only once and thus also aggregated only once.

EX 1 If gear Z 1 is installed in gearbox G 1, the gear is no longer available; the gearbox as a whole, however, is still available and can be installed in vehicle F 1. Once this has been done, again gearbox G 1 is no longer available; that is, it cannot simultaneously be installed in vehicle F 2.

EX 2 If a green box is packed inside a yellow box, the green box is no longer available and cannot also be packed inside a blue box. It is, however, possible to pack the yellow box (with the hidden green box inside it) in the blue box.

EX 3 In the case of physical objects, as in examples EX 1 and EX 2, the nesting principle is trivial. The important point is that it also applies to instances of functional units considered as "resources of functionality." If a functional unit A is an element of another functional unit B, functional unit A is no longer available for other functional units (in B, a certain task has been allotted to it that makes claims on its resources).

EP The nesting principle applies only to *instances* of *functional units*. It does not apply to the *aggregation* of *classes* or to the *aggregation* of *instances* of *functions*. A *functional element* can be simultaneously an *element* of more than one *function*.

Object

DF Subject of knowledge, perception, thinking, and action. In the context of object orientation: identifiable patterns that are describable in terms of properties. The internal structure of objects is not directly perceivable but may be deduced from the properties describing the object.

EP Both functions and substances can be treated as objects in this sense [863]. Object is not an antonym of function.

Operating Mode

This term is used with several meanings.

DF 1 Degree of automation of a plant.
Here, the operating mode is the plane (level) of operator control currently selected.

EX 1 A coarse classification includes the following modes: manual, computer, logistical, and automatic [859].

– Manual mode: the monitoring and control system takes over only certain process control tasks for the automatic execution of individual loops.

– Computer mode: the monitoring and control system takes over all process control tasks for the automatic performance of a process section in a plant section.

– Logistical mode: the monitoring and control system takes over all process control tasks for the automatic fulfillment of scheduled orders within a plant.

– Automatic mode: the monitoring and control system takes over all process control tasks including fine production control for the automatic scheduling and fulfillment of orders within a plant.

DF 2 One of the fundamentally different behavior patterns of functional units.

Alternatively, certain technological function units can carry out distinct (selectable) tasks.

EX 2 The functional unit "inert-gas blanketing" has the modes "pressurizing" and "purging." In the first, the task is to establish the pressure in accordance with the specified control value; in the second, to establish the flow of nitrogen through the system.

DF 3 Way in which a plant is run in dependence on the control of the operating value or throughput-to-capacity ratio (startup, normal running, load changing, emergency shutdown, etc.).

LI 3 [862]

Operating Value (Throughput-to-Capacity Ratio)

DF The ratio of product throughput through a plant to design throughput.

LI [862]

Parameter

DF 1 Parameter (of an equation). In functions and equations, an auxiliary quantity appearing alongside the variables proper and either left undefined or held constant.

DF 2 Parameter (of a path or trajectory). Special case of a parameter of an equation.

DF 3 Parameter (of a process). Variable that influences a technical, economic, or other process.

EP In the process control context, this term is often used in an imprecise and poorly understood way, and it should be largely avoided.

Partial Recipe

DF Complete description for making a certain end product or intermediate product in a particular type of plant section.
A partial recipe consists of:

– A specification as to the properties of the inlet products
– A specification as to the state of the plant section at the beginning of the process
– A manufacturing specification for the control of the process.

Phase

DF Phases subdivide *process elements* into a strict temporal order.

EP Each process element, at each point in time, is in precisely one phase. There is no parallelism of phases in a plant section.

EP Remark: In many cases, the phases are named after the unit operations that predominate in them. This often leads to misunderstanding. In a phase referred to as "meter starting substances," the unit operations "meter A," "meter B," "temper," and "mix" (agitate) can take place in parallel. The sequence of phases, however, is strictly serial.

EP The subdivisions of the *phase model* at the lowermost level of refinement correspond to the phases.

Phase Model

DF The phase model is a special network model with which a procedure can be systematically described. The phase model consists of two elements:

– Process elements in which n inlet products are transformed to m end products by means of operations.
– Product nodes with well-defined product properties, which connect the process elements to one another.

EP The phase model offers two substantial advantages:

– Logical network structure
The network structure permits systematic refinement (top-down method) from a coarse structure to a detailed description. The specification of nodes in terms of products makes it possible to describe and refine the individual process elements independently of one another. The phase model is not concerned with the form of the temporal and engineering realization of the coupling.
– The possibility of establishing conditions that must be observed, by specifying profiles of process and product properties in parallel with the design process.
The specification of profiles for the process properties to be brought about in the process elements and the product properties required in the nodes can be done in advance and refined with the network structure.
The phase model is the basis for the construction of quality assurance systems.

LI [859]

Plant

DF Combination of plant sections into a delimited technological unit.

EP The context in which the unit is delimited is not fixed. The important point is that each plant section can be assigned to only one plant.

EP An appropriate delimitation can be found if, for example, the plant is viewed as an analytical space for a fine production control system.

LI DIN 28 004

Plant Complex

DF Organizational combination of plants and other units into an overall heterogeneous unit.

Plant Control

DF Functional unit with the capability of initiating and coordinating the processing elements in the plant sections of a plant on the basis of a specified operating plan.

Plant Section

DF Self-contained unit for the realization of a technical task (see Fig. 26).

EP Among the points characterizing a plant section is that the product treated in it is to be regarded as substantially located in a balance space. This balance space may well be distributed (tower, fixed-bed reactor, etc.) but still forms one unit from the standpoint of process engineering.
Two stirred-tank reactors should be viewed as two separate plant sections because, in general, each is characterized by its own balance space.

LI DIN 28 004

Plant Section Control

DF 1 *Functional unit* with the capability of controlling the *group controls* or *basic functions* of a *plant section* and of automatically controlling a process section in accordance with a specified recipe.

DF 2 Hardware/software module for the realization of the "functional unit plant section control" (see note under *Function*).

Process (General)

DF A totality of interacting trains of events in a system by which matter, energy, or information is transformed, transported, or stored. The smallest unit for analysis is the process element.

EP Note: The term is used to refer to the actual train of events taking place in the real world, not to the system characterized by this process or the specification underlying the process.

EP In a processing system, "process" denotes the dynamic variation of product properties as a result of the action of unit operations or the variation of process properties due to the action of basic functions.

LI DIN 66 201

Process Control

DF Process control takes in all measures that effect the desired course of a process in the sense of established objectives. These measures either can be executed automatically by the process monitoring and control hardware itself, or may have to be performed by human operators.

LI DIN 19 222

Process Control Function

DF *Process monitoring and control function* with the task of influencing the course of a process in such a way that certain *process properties* will track specified *control values*. A process control function is realized by hierarchically structured process control units in the feedforward control branch and a network of process information processing units in the feedback branches (see also *Control Value*).

Process Control Station (Process Monitoring and Control Station; formerly Measurement and Control Station)
The concept of the process control station is a key one in process control engineering. Even so, it is difficult to define and delimit. In common usage, the term refers to a variety of concepts:

DF 1 A point of connection of the process.
In this sense, a process control station is the interface between the process monitoring and control functional unit and an engineering functional unit (used so, at

least in part, in the station data sheet or in the station detail of the piping and instrumentation diagram).

DF 2 A functional unit of process control. This functional unit is characterized by the fact that it possesses at least one process connection point. The task of the functional unit "process control station" is primarily to execute the assigned processing functions (O, I, C, S, . . .). Two remarks should be made, however:

– Only in a few cases is a processing function realized, as a whole, by a functional unit. In many cases (e.g., the S function, central A, I, O evaluating functional units), multiple functional units (and multiple process control stations, in the case of the S function) are involved in realizing a processing function.
– The differentiation between the functional unit "process control station" and more complex interconnections is indistinct (at what complexity does a functional unit with process connection points cease to be a process control station?).

DF 3 A combination of functions. In this sense, the process control station is given by the sum of the defined processing functions assigned to the process connection points (used so, at least in part, in the station detail of the piping and instrumentation diagram).

LI DIN 19 226

Process Control Station Element

DF Process control station elements are the smallest functional units into which a process control station (DF 2) is broken down. Process control station elements have functionality established such that they can be realized by available substances (devices, software modules, etc.). Each process control station element is needed for the performance of one or more processing functions.

Process Element

DF Smallest unit used in analysis of a process.

EP A process element can be realized by a process step or a process section, depending on whether the analysis is time-based or space-based.

Processing Function

DF Term denoting process monitoring and control functions allotted to a process control station. The processing functions are represented in the station detail.

EP Note that a processing function is generally realized through more than one functional unit and, in certain cases, through more than one process control station, despite the unambiguous allocation (see note under *Process Control Station*).

Process Intervention

DF Monitoring and control functional unit that combines all functional elements (process control station elements) of control hardware between discrete control device and process.

Process Monitoring and Control Function

DF *Function* realized with the aid of the monitoring and control hardware. In general, monitoring and control functions are closed lines of action between process elements, between operator and process, between operator input and operator output, or in general the response of the outputs of the process control system to certain input signals.

CL Monitoring and control functions in process control engineering can be specified by control action in the following ways (among others):

Process control functions

– After operation DOS_3_ON, meter in x liters in t minutes
– After operation T5_ON, turn on plant section 5

Process securing (stabilization) functions

– If temperature T 15 > 80 °C, turn off pump 3

Process information functions

– Display temperature T 15 in on-screen flowsheet

Device control functions

– Calibrate pH probe Q 14

System control functions

– Select screen page plant section ST 3

Monitoring and control functions also occur in other settings such as manufactur-

ing engineering or building management engineering.

Process Monitoring and Control Hardware

DF All devices employed for the task of process monitoring and control.

LI DIN 19 222

SN Process control system

Process Property

DF Process properties are the categories of information about a process. They include state variables, process parameters, control variables, setpoint parameters, and process indicators.

LI [864]

Process State

DF Expression of the process property vector at a given time (see Section 11.3).

EP See also *Quality*.

Production Control

DF 1 Functional unit with the task of scheduling production requirements, preparing a production schedule, allocating the needed resources, and monitoring the performance of the schedule.

DF 2 Software/hardware module with the capabilities of realizing precisely the function "production control" (DF 1) (see note under *Function*; see also Fig. 31).

Product Properties

DF Product properties are the categories of information about a product. They include physical quantities, chemical quantities, technological properties, and product indicators.

LI [864]

Project

DF Expressed generally in the terminology used here, a project is the complete temporal framework of conditions governing all *substances* for the realization of tasks represented by *functions*. The substances consist of procedures, systems, and elements (see Figs. 10 and 65) and are also referred to as resources.

Property

DF A characteristic or feature of an object. A property consists of the type of feature (category, size) and the expression of the feature (intensity, value). The type of expression (metrical, topological, etc.) depends on the type of feature.

EX A house might have the properties
Type of feature: Expression:
Address Hauptstraße 3
D-11111 Kleinhausen
Owner Mrs. Meier
Height 8.25 m
Building style Jugendstil (Art Nouveau)

EP Classes are marked by common feature types. The expression is an instance quantity.

SN *Attribute*

Quality

DF Expression of the product property vector at a given time (see Section 11.3).

EP See also *Process State*.

Recipe

DF A recipe consists of the sum of the partial recipes needed to make an end product and the explicitly stated relationships between the partial recipes.

EP The term "recipe" has been introduced but is still subject to various interpretations. "Production instructions" or "production specification" would be a more precise term and should be given preference in the future.

Redundancy (General)

DF 1 Concept in information theory, denoting that a message contains elements that can be omitted as not providing any additional information but merely supporting the desired basic information.

– Beneficial redundancy: constituents that can be omitted would back up the information if other constituents were omitted.

– Empty redundancy: constituents that can be omitted would not back up the information content if other constituents were omitted.

DF 2 Severalfold presence of a functionality in order to enhance the availability of a function.

Redundant Measurement

DF The measurement of a certain process property is said to be redundant when the actual value of this process property can be acquired from at least two disjoint groups of available measurements.

EP If a measurement becomes unavailable, the disjointness means that only one group is affected, and the actual value continues to be available to the system. This is an example of beneficial redundancy.

- Homogeneous redundancy: the actual value of the desired process property is determined by using the same measuring principle and by multiple measurements of the same auxiliary process properties.
- Diversity: the term "diversity" (diverse redundancy) is used when:
  - The actual value of the desired process property is determined from measurements of different auxiliary process properties (example: determination of flow rate via pressure drop and with magnetic flowmeter).
  - The actual (instantaneous) value of the desired process property is determined on the basis of the same auxiliary process properties, but the auxiliary process properties are measured by different methods (example: level measurement with float and ultrasonic instruments).

Repair(s)

DF The restoration of the nominal condition of a plant.

LI DIN 31 051 (1985)

Requirements (see also *Specifications*)

DF Document describing the tasks (*functions*) that are to be implemented in the framework of a project. The initial description is solution-invariant.

SN Bidding specifications

Role

DF Roles are unambiguously described tasks within a larger context. The realization and performance of a role requires a functional unit (execution unit), which realizes the role in a certain layer at a certain time.

EX The relationships between roles, resources, functional (execution) units, and layer can be illustrated with three examples:

– Theatrical performance
Consider the role of "Gessler" in the play "William Tell" (virtual system).
The realization of this role demands an actor. Any actor who knows the text and is capable of playing Gessler is a potential candidate and thus belongs to the pool of resources. Now suppose the play is to be performed on October 16, 1995, at venue X. The role system thus requires that an actor suitable for the representation of Gessler be available at X on October 16, 1995. For the theater manager, this means that an actor Y must be sought and signed for the performance; that is, a resource must be selected and allocated. This can take place ahead of time when the theater season is being planned. For the play to be actually performed, it is important that Mr. Y be available and prepared for the role at X on October 16, 1995. He is configured for his role through makeup and appropriate costuming. Mr. Y is then an available function (execution) unit for the role of "Gessler" at X on October 16, 1995.

– Hardware component
Consider the following role: discrete control component for reactor X. It belongs to the process monitoring and control system being planned (virtual system). Each component of type Y made by manufacturer A possesses the required functionality and is thus a potential resource. As a functional (execution) unit, it is present (available) if, as a hardware part, it has been delivered, configured (e.g., jumpers placed), and installed in the proper slot. (This job is performed by installation.) In this state, the component can take over the role envisioned for it in the process monitoring and control system.

– Motor control module
Consider the role of the control of motor V 12 TA 45 K 2. The process control being designed is the virtual system. For the realization of the role, a software module must be placed in the process monitoring and control system and made available as a functional (execution) unit. Any

instance of the motor control type MOT becomes a candidate. Thus, an instance MOT–V 12 TA 45 K 2 is defined, configured (here this may mean parameter assignment: number of feedback messages, startup time, etc.), and installed in the component provided (which here corresponds to the layer). Now this software module is capable of realizing the role of the motor control V 12 TA 45 K 2.

SN Function

Securing (Stabilization) Functions

DF *Functions* for preventing unacceptable deviations of the *process,* of the *process control,* or of the *process monitoring and control hardware* from the nominal state. Securing functions do not intervene in the sequence of events under normal operating conditions and permit the *process control functions* to perform their task. Before a deviation enters the unacceptable range, the securing functions are activated, interrupting the access of the process control functions.

EX Typical securing functions are interlocking of functional units and STOP conditions in sequence controls.

EP Other expressions of this concept refer to improving or enhancing the reliability or safety of the process.

Sequence Control

DF *Control* in which the program is constrained to progress step by step and the advance from one step to the next step in the program depends on advancing conditions.

LI DIN 19 237

Depending on the type of conditions for advance, the sequence control is said to be time-dependent or process-dependent. One also speaks simply of timed control or event control.

Specifications (see also *Requirements*)

DF 1 Document that contains all information required for the realization of a certain system.

EP 1 The specifications describe how the functionality of a system as set forth in the requirements can be realized with the help of given resources. "Given resources" mean functional units to be used, objects (hardware/software) to be used, rules and procedures to be applied, and constraints to be observed (e.g., limitations on time, money, material, space, and personnel resources). The specifications are linked to the solution adopted.

DF 2 Separate document describing the constraints, rules, and data that must be followed in the procedure.

Substances

In philosophy, the concept "substance" is employed with a variety of meanings. Here it is used in the sense of CASSIRER [863]:

DF Substances are objects in the real world that can be described and identified in terms of attributes.

Substances either are real objects themselves or correspond virtually to real objects.

Functions always require appropriate substances for their realization.

EP Substances include systems and their elements (devices), as well as skills and methods (procedures). With the aid of procedures, functions are realized from systems and their elements through the use of a part of their properties.

Of the great number of possible substances (see Fig. 9), "standards" (e.g., design, installation, and hardware standards) are derived by means of appropriate definitions. Substances, as real objects, can be either hardware objects or software objects.

Subsystem

DF Part of a *system*, having the following properties:

– Elements of a subsystem are always elements of the surrounding system as well.

– The relationship between elements in a subsystem always corresponds to the same relationship between the same elements in the surrounding system.

– Distinct subsystems of a system can overlap.

EP A special case is the complete subdivision of a system into nonoverlapping subsystems. This case corresponds to a *break*

*down*in accordance with the *nesting principle.*

LI [865]

System

DF A delimited, ordered totality of interacting subsystems or system elements. This configuration is separated, or regarded as being separated, from its surroundings by an envelope.

EP One characteristic of a system is that it simultaneously admits of an external and an internal view.
In the external view, the system is examined from outside its envelope, which is thought of as opaque. The system behaves as an aggregated object (black box).
In the internal view, the system is examined from inside its envelope, which is thought of as opaque. From this standpoint, the structure and mode of operation of the system can be seen.
The properties of a system cannot generally be elucidated by superposing or aggregating the properties of its elements.

EX A gearbox is a system of gears, locking mechanisms, bearings, and shafts. From the outside, in the technological view, all that is interesting about a gearbox is the transmission ratio between the driven shaft and the driving shaft. The internal geometry and the number of gears are not important. From the inside, on the other hand, it is precisely the geometrical structure of shafts and gears that is of interest. The only way to determine the transmission ratio effective at the envelope is through a complicated computation based on the individual gears of the active branch.

EP It is important that a system admits of both views at the same time. It can be said to represent an object at the moment of aggregation or disaggregation.

LI DIN 66 201, DIN 19 226, [859]

Taxonomy

DF A methodology used in science, consisting in the assignment of concepts or individual phenomena to a general ordering scheme.

Type

DF Class whose objects exhibit a high degree of interrelatedness.

Unit Operations

DF The smallest physical or chemical operation performed, which acts on a material system in a plant section.

EP By means of physical or chemical operations, the material system in a plant section is transformed from a state A to a state B. These operations can be broken down into unit operations. As a rule, unit operations are carried out one after another, but there are unit operations that take place in parallel. Examples: conveying and drying in the pneumatic conveying dryer, reaction and mixing in the stirred-tank reactor, etc.

CL Unit operations can be classified under conveying, mechanical processes, force-field processes, processes with molecular driving force, and chemical processes.

LI [862]

14.2. Abbreviations

AI	Analog Input, Artificial Intelligence
AMA	Arbeitsgemeinschaft Meßwertaufnehmer
ANSI	American National Standards Institute
AO	Analog Output
AR	Archival Functions
ASN. 1	Abstract Syntax Notation One
ASP	Abstract Service Primitive
C^{++}	Programming language (object oriented)
C Methods	Computer-supported methods
CAD	Computer Aided Drafting, Drawing, Design
CAE	Computer Aided Engineering
CASE	Computer Aided Software Engineering
CCIT	Comité Consultatif Internationale Télégraphique et Téléphonique
CEN	Comité Européen de Normalisation
CENELEC	Comité Européen de Normalisation Electrique

CEPT	Comité Européen des Administrations des Postes et des Télécommunications
CIM	Computer Integrated Manufacturing
CO	Control Functions
CIP	Computer Integrated Processing; Cleaning in Process; Cleaning in Place
CPU	Central Processing Unit
CSMA/CA	Carrier Sense, Multiple Access, Collision Avoidance
CSMA/CD	Carrier Sense, Multiple Access, Collision Detection
DBMS	Data Base Management System
DCS	Distributed Control System
DEC	Digital Equipment Corporation
DES	Discrete Event System
DI	Digital Input
DIF	Data Interchange Format
DGK	Deutsche Gesellschaft für Kybernetik
DGQ	Deutsche Gesellschaft für Qualität
DKE	Deutsche Elektrotechnische Kommission
DLL	Data Link Layer
DO	Digital Output
DOS	Disk Operating System
DPG	Deutsche Physikalische Gesellschaft
DS	Directory Service
DVT	Deutscher Verbund Technisch-wissenschaftlicher Vereine
EAPROM	Electrically Alterable Programmable Read Only Memory
ECITC	European Committee for Information Technology Testing and Certification
EEPROM	Electrically Erasable Programmable Read Only Memory
EIA	Electronics Industries Association
ELCS	Extended List of Control Software
ElexV	Verordnung über elektrische Anlagen in explosionsgefährdeten Räumen
EPROM	Erasable Programmable Read Only Memory
EPS	European Physical Society
E/R	Entity – Relationship
ESPRIT	European Strategic Program for Research in Information Technology
ETSI	European Telecommunications Standards Institute
EUREKA	European Research Coordination Agency
EXERA	Association des Exploitants d'Equipement de Mesure, de Régulation et d'Automatisme
FIP	Factory Instrumentation Protocol
FLAVOR	Programming language
GA	Gateway Functions
GI	Gesellschaft für Informatik
GMA	VDI/VDE-Gesellschaft für Meß- und Automatisierungstechnik
GMD	Gesellschaft für Mathematik und Datenverarbeitung
GMR	VDI/VDE-Gesellschaft für Meß- und Regelungstechnik, (now: GMA)
GVC	VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen
HART	Highway Addressable Remote Transducer
HD	Harmonization Documents
HDLC	High Level Data Link Control
HHT	Hand Held Terminal
HPC	Human – Process Communication
IEC	International Electrotechnical Commission
IECEE	IEC Sub-Committee: Conformity Testing to Standard for Safety of Electrical EquipmentI
IEEE	Institute of Electrical and Electronics Engineers
IFAC	International Federation of Automatic Control
IFC	International Fieldbus Consortium
IFIP	International Federation for Information Processing
IFORS	International Federation of Operational Research Societies
IMACS	International Association for Mathematics and Computers in Simulation
IMEKO	Internationale Meßtechnische Konföderation
IMC	Internal Model Control
IMS	Information Management System
IP	Input Product

ISA	Instrument Society of America
ISDN	Integrated Services Digital Network
ISO	International Organization for Standardization
ISP	International Systems Project
LAN	Local Area Network
LISP	List Processing Language
LLC	Logical Link Control
LTR	Loop Transfer Recovery
MAC	Medium Access Control
MAP	Manufacturing Automation Protocol
MIS	Management Information System
MMC	Man – Machine Communication
MMS	Manufacturing Message Specification
MODEM	Modulator and Demodulator
MPC	Man – Process Communication
MVS	Multi-Virtual System
NAMUR	Normenarbeitsgesellschaft für Meß- und Regeltechnik in der chemischen Industrie
NIR	Near Infrared
OOA	Object Oriented Analysis
OOD	Object Oriented Design
OP	Output Product
OS	Operating System or Open System
OSF	Open Software Foundation
OSI	Open Systems Interconnection
P	Proportional (controller)
PBX	Private Branch Exchange
PC	Personal Computer
PDS	Plant Design System
PDU	Protocol Data Unit
PI	Proportional Integral(controller)
PID	Proportional Integral Differential (controller)
PL/1	Programming Language Number 1
PLC	Programmable Logical Controller
RFN	Reversed Frame Normalizing Process
PROFIBUS	Process Field Bus
PROLOG	Programming in Logic (programming language)
PROM	Programmable Read Only Memory
PTB	Physikalisch Technische Bundesanstalt
RAM	Random Access Memory
ROM	Read Only Memory
ROI	Return on Investment
SA	Structured Analysis
SADT	Structured Analysis and Design Technique
SAP	Service Access Point
SAS	Statistical Analysis System, Statistikprogramm der Firma SAS Institute
SIREP	International Instrument User's Association
SNA	Systems Network Architecture (IBM)
SPARC	Standards Planning and Requirements Committee (RISC Architecture)
ST	Structured Text
STEP 5	Programming language
SQL	Standard Query Language
TC	Technical Committee (IEC)
TIRC	Temperature Indicating Recording Controlling (DIN 19 227)
UNIX	Multitasking and Multiuser operating system
UTE	Union Technique de l'Electricité
UVV	Unfallverhütungsvorschrift
VbF	Verordnung über brennbare Flüssigkeiten
VCI	Verband der Chemischen Industrie
VDI	Verein Deutscher Ingenieure
VDE	Verband Deutscher Elektrotechniker
VIK	Vereinigung Industrielle Kraftwirtschaft
VLSI	Very Large Scale Integration
VMD	Virtual Manufacturing Device
VMS	Operating system from DEC
VOB	Verdingungsordnung für Bauleistungen
WAN	Wide Area Network
XP	Expert System
ZVEI	Zentralverband Elektrotechnik- und Elektronikindustrie

15. References

General References

Automation and Process Control Engineering

1. D. Balzer: *Wissensbasierte Systeme in der Automatisierungstechnik,* Carl Hanser Verlag, München 1992.
2. K. Bender (ed.): *Profibus – Der Feldbus für die Automation,* Carl Hanser Verlag, München 1990.

3. G. Brack: *Automatisierungstechnik für Anwender,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig –Stuttgart 1993.
4. G. Brack, A. Helms: *Automatisierungstechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1985.
5. G. Färber: *Prozeßrechentechnik: Allgemeines, Hard- und Software, Planungshinweise,* 2nd ed., Springer Verlag, Berlin 1992.
6. E. L. Grant, W. G. Ireson, R. S. Leavenworth: *Principles of Engineering Economy,* John Wiley and Sons, New York 1982.
7. J. Hengstenberg, B. Sturm, O. Winkler, K. H. Schmitt (eds.): *Messen, Steuern und Regeln in der Chemischen Technik,* **vols. 1 – 5,** Springer Verlag, Berlin 1980.
8. K. Hiroi: *Basis and Application of Digital Control Systems,* Japan.
9. ISA: Human Engineering for Control Centers, ISA Recommended Practice RP 60.3, ISA 1985.
10. R. Lauber: *Prozeßautomatisierung,* 2nd ed., **vol. 1,** Springer Verlag, Berlin 1989.
11. R. Magin, W. Wüchner: *Digitale Prozeßleittechnik,* Verlag Vogel, Würzburg 1987.
12. P. Metzing, D. Balzer, H. Reinhardt: *Prozeßsteuerung,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1988.
13. I. Morishita: *Digital Control Systems,* Society of Instrumentation and Process Control, Japan.
14. W. Motsch: *Prozeßrechner – Strukturen,* Verlag Vieweg, Braunschweig 1993.
15. NAMUR (Hrsg.): NAMUR-Statusberichte 1985, 1987, 1990, 1993, R. Oldenbourg Verlag, München.
16. G. Platt: *Process Control, a Primer for the Nonspecialist and the Newcomer,* ISA 1988.
17. M. Polke: *Prozeßleittechnik,* 2nd ed., Oldenbourg, München 1984.
18. R. Rück, A. Stockert, F. O. Vogel: *CIM und LOGISTIK im Unternehmen,* Carl Hanser Verlag, München 1992.
19. E. Schrüfer (ed.): *VDI-Lexikon Meß- und Automatisierungstechnik,* VDI-Verlag, Düsseldorf 1991.
20. G. G. Seip (ed.): *Elektrische Installationstechnik,* **vols. 1 –3,** Siemens AG, Berlin 1971.
21. A. Shoten: *PID Control,* Systems Control and Information Society, Japan.
22. G. Strohrmann: *Automatisierungstechnik,* **vols. I and II,** R. Oldenbourg Verlag, München 1990.
23. M. Syrbe, M. Thoma (eds.): *Fachberichte Messen-Steuern-Regeln,* **vols. 1 – 10,** Springer Verlag, Berlin.
24. H. Töpfer, P. Besch: *Grundlagen der Automatisierungstechnik,* VEB Deutscher Verlag Technik, Berlin 1990.
25. H. Töpfer, W. Kriesel: *Funktionseinheiten der Automatisierungstechnik,* VEB Deutscher Verlag Technik, Berlin 1988.
26. Kompaktseminar Prozeßleittechnik: Themen in Einzeldarstellungen, R. Oldenbourg Verlag, München 1986 ff.
27. *Systems Engineering* M. Bruns: *Systemtechnik,* Springer Verlag, Heidelberg 1991.
28. W. F. Daenzer (ed.): *Systems Engineering,* 6nd ed., Verlag Industrielle Organisation, Zürich 1988.
29. J. Raasch : *Systementwicklung mit strukturierten Methoden,* Carl Hanser Verlag, München 1992.
30. R. Unbehauen: *Systemtheorie: Grundlagen für Ingenieure,* R. Oldenbourg Verlag, München 1993.

General References

Computer Science

31. D. Abel: *Petri-Netze für Ingenieure,* Springer Verlag, Berlin 1990.
32. L. Christov: *Objektorientierte Programmiersprache C+++: Vollständige Einführung anhand von Beispielen,* VDI-Verlag, Düsseldorf 1992.
33. J. Fiedler, K. F. Rix, H. Zöller: *Objekt-orientierte Programmierung in der Automatisierung,* VDI-Verlag, Düsseldorf 1991.
34. G. Glöe (ed.): *Software-Zuverlässigkeit,* VDI-Verlag, Düsseldorf 1992.
35. M. Maser: *Grundlagen der allgemeinen Kommunikationstheorie,* Verlag Berliner Union, Stuttgart and Kohlhammer, Stuttgart 1971.
36. *Expertensysteme: Einführung in Technik und Anwendung,* Siemens AG, Berlin 1987.
37. U. Rembold: *Einführung in die Informatik,* Carl Hanser Verlag, München 1991.
38. E. Schnieder (ed.): *Prozeßinformation: Einführung mit Petrinetzen,* 2nd ed., Verlag Vieweg, Braunschweig 1993.
39. S. Wendt: *Nichtphysikalische Grundlagen der Informationstechnik,* Springer Verlag, Berlin 1989.
40. Various editors: Reihe Informatik Fachberichte, Springer Verlag, Berlin.

General References
Process Engineering

41. E. Blaß: *Entwicklung verfahrenstechnischer Prozesse,* Salle und Sauerländer Verlag, Frankfurt 1989.
42. W. Büchner, R. Schliebs, G. Winter, K. H. Büchel: *Industrielle Anorganische Chemie,* 2nd ed., Verlag Chemie, Weinheim 1986.
43. M. Engshuber, R. Müller: *Grundlagen der Verfahrenstechnik für Automatisierungsingenieure,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig – Stuttgart 1993.
44. E. Ignatowitz (ed.): *Chemietechnik,* Europa-Lehrmittel Verlag, Wuppertal 1980.
45. H. Robel et al.: *Lehrbuch der chemischen Verfahrenstechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1983.
46. *Ullmann,* 4th ed., **5,** 891.
47. K. Weissermel, H.-J. Arpe: *Industrielle Organische Chemie,* 3rd ed., VCH Verlagsgesellschaft, Weinheim 1990.

General References
Control Engineering and Simulation

48. A. Auer: *Speicherprogrammierte Regelung,* Hüthig, Heidelberg 1990.
49. A. Böttiger: *Regelungstechnik: Eine Einführung für Ingenieure und Naturwissenschaftler,* 2nd ed., R. Oldenbourg Verlag, München 1991.
50. F. Dörrscheidt, W. Latzel: *Grundlagen der Regelungstechnik,* 2nd ed., B. G. Teubner Verlag, Stuttgart 1992.
51. T. Ebel: *Regelungstechnik,* 6th ed., B. G. Teubner Verlag (Studienskripten), Stuttgart 1991.
52. O. Föllinger: *Nichtlineare Regelungen,* **vols. I and II,** R. Oldenbourg Verlag, München, vol. I (5th ed.) 1989, vol. II (6th ed.) 1991.
53. O. Föllinger: *Regelungstechnik: Einführung in die Methoden und ihre Anwendung,* 7th ed., Hüthig, Heidelberg 1992.
54. J. C. Gille, M. Pelegrin, P. Decaulne: *Lehrgang der Regelungstechnik,* R. Oldenbourg Verlag, München 1963.
55. E.-D. Gilles: *Systeme mit verteilten Parametern,* R. Oldenbourg Verlag, München 1973.
56. H.-M. Hanisch: *Petri-Netze in der Verfahrenstechnik: Modellierung und Steuerung verfahrenstechnischer Systeme,* R. Oldenbourg Verlag, München 1992.
57. R. Isermann: *Identifikation dynamischer Systeme 1: Grundlegende Methoden,* 2nd ed., Springer Verlag, Berlin 1992.
58. R. Isermann: *Identifikation dynamischer Systeme 2: Besondere Methoden, Anwendungen,* 2nd ed., Springer Verlag, Berlin 1992.
59. D. P. F. Möller: *Modellbildung, Simulation und Identifikation dynamischer Systeme,* Springer Verlag, Berlin 1992.
60. W. Oppelt: *Kleines Handbuch technischer Regelvorgänge,* Verlag Chemie, Weinheim 1972.
61. M. Papageorgiou: *Optimierung: Statische, dynamische, stochastische Verfahren für die Anwendung,* R. Oldenbourg Verlag, München 1991.
62. E. Samal: *Grundriß der praktischen Regelungstechnik,* 17th ed., R. Oldenbourg Verlag, München 1991.
63. G. Schmidt: *Grundlagen der Regelungstechnik,* Springer Verlag, Berlin 1989.
64. G. Schmidt: *Simulationstechnik,* R. Oldenbourg Verlag, München 1980.
65. J. R. Schwarz, B. Friedland (eds.) Linear Systems, Mc-Graw-Hill, New York 1965. W. W. Solodownikow: *Grundlagen der selbständigen Regelung,* **vols. 1 – 2,** R. Oldenbourg Verlag, München 1959.
66. M. Thoma: *Theorie linearer Regelsysteme,* Verlag Vieweg, Braunschweig 1973.
67. J. G. Truxal: *Entwurf automatischer Regelsysteme,* R. Oldenbourg Verlag, München 1960.
68. G. Wunsch, H. Schreiber: *Stochastische Systeme,* 3rd ed., Springer Verlag, Berlin 1992.

General References
Control Technology

69. A. Auer: *SPS – Progammierung: Beispiele und Aufgaben,* 3rd ed., Hüthig, Heidelberg 1991.
70. A. Auer: *SPS – Aufbau und Programmierung,* 4th ed., Hüthig, Heidelberg 1991.
71. P. Conrads: *Vom Umgang mit SPS,* Hüthig, Heidelberg 1991.
72. R. Haug: *Pneumatische Steuerungstechnik,* 2nd ed., B. G. Teubner (Studienskripten), Stuttgart 1991.
73. J. Petry: *SPS – Projektierung und Programmierung,* 2nd ed., Hüthig, Heidelberg 1990.
74. G. Wellenreuther, D. Zastrow: *Steuerungstechnik mit SPS,* Verlag Vieweg, Braunschweig 1991.

75. G. Wellenreuther, D. Zastrow: *Lösungsbuch Steuerungstechnik mit SPS,* Verlag Vieweg, Braunschweig 1992.

General References
Metrology
76. R. Best: *Digitale Meßwertverarbeitung,* R. Oldenbourg Verlag, München 1991.
77. W. H. Cubberly (ed.): *Comprehensive Dictionary of Measurement and Control,* 2nd ed., ISA 1992.
78. O. Fiedler: *Strömungs- und Durchflußmeßtechnik,* R. Oldenbourg Verlag, München 1992.
79. H. Gatzmanga et al.: *Prozeßmeßtechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1989.
80. W. Göpel, J. Hesse, J. N. Zemel (eds.): *Sensors – A Comprehensive Survey,* **vols. 1 – 8,** VCH Verlagsgesellschaft, Weinheim 1992.
81. K.-H. Gonschorek, H. Singer (eds.): *Elektromagnetische Verträglichkeit: Grundlagen, Analysen, Maßnahmen,* B. G. Teubner Verlag, Stuttgart 1992.
82. E. Hering, K. Bressler, J. Gutekunst: *Elektronik für Ingenieure,* VDI-Verlag, Düsseldorf 1992.
83. J. W. Klein, P. Dullenkopf, A. Glasmachers: *Elektronische Meßtechnik: Meßsysteme und Schaltungen (Teubner Studienbücher),* B. G. Teubner Verlag, Stuttgart 1992.
84. H. Kronmüller: *Prinzipien der Prozeßmeßtechnik,* **vols. 1 and 2,** Schnäker Verlag, Karlsruhe 1986.
85. S. Miyazaki et al.: *Instumentation and Control for Process,* Nikkan Kohgyo, Shuppan-sha, Japan.
86. P. Profos, T. Pfeifer (eds.): *Grundlagen der Meßtechnik,* 3rd ed., R. Oldenbourg Verlag, München 1992.
87. P. Profos, T. Pfeifer (eds.): *Handbuch der industriellen Meßtechnik,* 5th ed., R. Oldenbourg Verlag, München 1992.
88. W. Richter: *Grundlagen der elektrischen Meßtechnik,* Verlag Technik, Berlin 1989.
89. G. W. Schanz: *Sensoren – Fühler der Meßtechnik,* 3rd ed., Hüthig, Heidelberg 1993.
90. H. Schaumburg: *Sensoren,* B. G. Teubner Verlag, Stuttgart 1992.
91. G. Schnell (ed.): *Sensoren in der Automatisierungstechnik,* 2nd ed., Verlag Vieweg, Braunschweig 1993.
92. G. Strohrmann: *Meßtechnik im Chemiebereich,* R. Oldenbourg Verlag, München 1991.
93. H.-R. Tränkler: *Taschenbuch der Meßtechnik mit Schwerpunkt Sensortechnik,* R. Oldenbourg Verlag, München 1992.
94. *Ullmann,* 4th ed., **5,** 797 – 890.
95. *Instrumentation Handbook,* Instrumentation and Process Control Engineer's Association, Japan.

General References
Conference Proceedings (see also Chap. 12)
96. Beratender Ausschuß der Industriephysiker der Deutschen Physikalischen Gesellschaft: Physikertagung, e.g., Proceedings of: Der Physiker in der Automatisierungstechnik (1989); Sensorphysik und Sensortechnik (1984).
97. IFAC: IFAC-World Congress: Automatic-Control, Selected Papers of the IFAC World Congress, Pergamon Press, Oxford.
98. INTERKAMA: INTERKAMA-proceedings as series: "Fachberichte Messen, Steuern, Regeln", P. Thoma (ed.), Springer Verlag, Berlin, since 1989: G. Schmidt, H. Steusloff (eds.), R. Oldenbourg Verlag, München.
99. GVC: Jahrestreffen der Verfahrensingenieure, Kurzfassung der Vorträge und Poster.
100. VDI: Automatisierungstechnik Tagungsband, VDI-Verlag, Düsseldorf.

General References
Journals
101. automatica
102. Automatisierungstechnik (at)
103. Automatisierungstechnische Praxis (atp)
104. Chemie Ingenieur Technik (CIT)
105. Computers & Chemical Engineering
106. Control Engineering Practice
107. Informatik Spektrum
108. INTECH (ISA)
109. Journal of Process Control
110. messen & prüfen
111. Process Control and Quality
112. Technisches Messen (tm)
113. VDI-Z
114. ZWFCIM
115. N. D. Kondratieff: "Die langen Wellen der Konjunktur," in: *Archiv für Sozialwissenschaft und Sozialpolitik,* vol. 56, Mohr Verlag, Tübingen 1926.
116. Leo A. Nefiodow: "Der fünfte Kondratieff," in: *Strategien zum Strukturwandel in Wirtschaft und Gesellschaft,* Gabler-Verlag, Wiesbaden 1990.

117. J. Bardeen, W. H. Brattain: "Transistor, A Semiconductor Triode," *Phys. Rev.* **74** (1948) 230 – 231.
118. D. P. Chen: "The Entity Relationship Model," *ACM Trans. Database Syst.* **1** (1976) 9 – 36.
119. C. A. Petri: "Kommunikation mit Automaten," Dissertation, TH Darmstadt 1961.
120. H. Nottebohm: "Entwicklungen der Informationstechnik," *Automatisierungstechnische Praxis (atp)* (1985) nos. 5, 6, 224 – 230, 267 – 273.
121. M. Polke: "Information als kritische Ressource," in H. A. Henzler (ed.): *Handbuch der strategischen Führung,* Gabler-Verlag, Wiesbaden 1988, pp. 353 – 378.
122. M. E. Porter: *Wettbewerbsvorteile,* Campus Verlag, New York – Frankfurt 1986.
123. J. Schumpeter: *Theorie der wirtschaftlichen Entwicklung,* 6th ed., Duncker und Humbolt Verlag, Berlin 1964.
124. M. Polke: *Technology and Structural Change,* presented at Scottish Council, International Forum, 1986.
125. B. Sturm, B. Will, M. Polke: "Die NAMUR – gestern, heute und morgen," *Automatisierungstechnische Praxis (atp)* **30** (1988) 11 – 29.
126. A. Lippert et al.: Bayerinterne Untersuchungen und Organisationsanweisungen, Leverkusen 1979 – 80.
127. R. Lauber: "Prozeßautomatisierung, Prozeßleittechnik oder Prozeßdatenverarbeitung," *Automatisierungstechnische Praxis (atp)* **28** (1986) 244 –245.
128. R. Lauber: *Prozeßautomatisierung als Hochschultext,* 2nd ed., Springer Verlag, Berlin – Heidelberg – New York 1989.
129. VDI/VDE Gesellschaft Meß- und Automatisierungstechnik (GMA): *Aufgaben – Struktur – Aktivitäten,* Düsseldorf 1991.
130. M. Polke: *Der Physiker in der Automatisierungstechnik,* DPG, Kongreß-, Ausstellungs- und Verwaltungs-GmbH, Bad Honnef 1989.
131. R. Magin, W. Wüchner: *Digitale Prozeßleittechnik,* Vogel-Buch Verlag, Würzburg 1987.
132. G. Strohrmann: "Prozeßleittechnik," *rtp-Seminar, Regulungstechnische Praxis (rtp)* **26** (1989) no. 10, 543 – 550.
133. G. Gruhn et al.: *Verfahrenstechnik – Prozeßmeßtechnik, Prozeßsteuerung, Automatisierungstechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1985.
134. H. Töpfer, P. Besch: *Grundlagen der Automatisierungstechnik,* 2nd ed., Verlag Technik, Berlin 1989.
135. M. Polke: "Erforderliche Informationsstrukturen für die Qualitätssicherung," *GVC-Fachtagung "Qualität",* Heidelberg, June 1992.
136. A.-W. Scheer: *Wirtschaftsinformatik,* Springer Verlag, Berlin 1988.
137. M. Vetter: *Aufbau betrieblicher Informationssysteme mittels konzeptioneller Datenmodellierung,* Teubner, Stuttgart 1989.
138. G. Schlageter, W. Stucky: *Datenbanksysteme: Konzepte und Modelle,* Teubner, Stuttgart 1983.
139. H. Wedeking, E. Ortner: *Systematisches Konstruieren von Datenbankanwendungen: Zur Methodologie der angewandten Informatik,* Hanser Verlag, München 1980.
140. P. Kruck: "Methodische Software-Entwicklung," Dissertation, Universität-Gesamthochschule Wuppertal, 1987.
141. E. F. Codd: "A Relational Model of Data for Large Shared Data Banks," *Commun. ACM* **13** (1970) 377 – 387.
142. P. P.-S. Chen: The Entity-Relationship-Model – Toward a Unified View of Data, *ACM Transactions on Database Systems,* **vol. 1,** Cambridge, Mass., March 1976, pp. 9 – 36.
143. A.-W. Scheer: *EDV-orientierte Betriebswirtschaftslehre,* Springer Verlag, Berlin 1990.
144. H. Vossberg et al.: "Die Informations- und Steuerungssysteme der Sparten," Bayer, Leverkusen, pp. 1970 ff.
145. R. Durchholz: "Konzeptionelles Schema," *Informatik Spektrum* **7** (1984) 245 – 247.
146. E. Ortner: "Semantische Modellierung – Datenbankentwurf auf der Ebene der Benutzer," *Informatik-Spektrum* **8** (1985) 20 – 28.
147. W. Wendt: *Nichtphysikalische Grundlage der Informationsstechnik,* Springer Verlag, Berlin 1989.
148. E. Ortner, B. Söllner: "Semantische Datenmodellierung nach der Objekttypenmethode," *Informatik-Spektrum* **12** (1989) 31 – 42.
149. F. Gebhart: "Semantisches Wissen in Datenbanken – ein Literaturbericht," *Informatik-Spektrum* **10** (1987) 79 – 88.

150. H. Peschke: "Partizipative Entwicklung und Einführung von Informationssystemen," in H. Balzert (ed.): *Einführung in die Software-Ergonomie,* De Gruyter, Berlin 1988, pp. 299 – 322.
151. D. B. Preßmar: "Strategien für den Einsatz von Methoden und Werkzeugen in der Softwaretechnologie," in T. Gutzwiller, H. Österle (eds.): *Anleitung zu einer praxisorientierten Software-Entwicklungsumgebung,* vol. 2, Angewandte Informationstechnik, Verlag, Hallbergmoos 1988, pp. 9 – 24.
152. S. Löffler, A. Warner: "Integration von Software-Entwicklungswerkzeugen," in H. Österle (ed.): *Anleitung zu einer praxisorientierten Software-Entwicklungsumgebung,* vol. 1, Angewandte Informationstechnik, Verlag, Hallbergmoos 1988, pp. 29 – 37.
153. U. Rembold (ed.): *Einführung in die Informatik,* Hanser Verlag, München 1987.
154. K. Heuer: "Die Rolle von Dictionary-Systemen in Software-Produktionsumgebungen," in H. Österle (ed.): *Anleitung zu einer praxisorientierten Software-Entwicklungsumgebung,* vol. 1, Angewandte Informatik, Verlag, Hallbergmoos 1988, pp. 117 –133.
155. S. Holloway: "Data Dictionary as an Administrative Tool – How to Succeed in the Nineties," in: Pergamon Infotech (ed.): *Information Update* **1** (1987) no. 1, 17 – 25.
156. E. Ortner, J. Rössner, B. Söllner: "Entwicklung und Verwaltung standardisierter Datenelemente," *Informatik-Spektrum* **13** (1990) 17 – 30.
157. H. Balzert: *Die Entwicklung von Software-Systemen; Prinzipien, Methoden, Sprachen, Werkzeuge,* Bibliographisches Institut, Mannheim 1990.
158. H. Balzert (ed.): *CASE-Systeme und Werkzeuge,* Wissenschaftsverlag, Mannheim 1990.
159. Aristoteles: *Kategorien,* Akademie Verlag, Berlin 1984.
160. C. v. Linné: *Systema naturae,* British Museum of Natural History Publishers, London 1958.
161. C. Darwin: *Die Entstehung der Arten* (In der Übersetzung von C. W. Neumann), Reclam, Stuttgart 1963.
162. E. Cassirer: *Substanzbegriff und Funktionsbegriff,* 6th ed., Wissenschaftliche Buchgesellschaft, Darmstadt 1990.
163. W. Ahrens: "Einsatz von Expertensystemen in der Prozeßleittechnik," *Automatisierungstechnische Praxis atp* **10** (1987) 475 – 485.
164. W. Gotthard: "Datenbanksysteme für Software-Produktionsumgebungen," *Informatik, Fachberichte 1993,* Springer Verlag, Berlin 1988.
165. H. Stoyan, H. Wedekind: *Objektorientierte Software- und Hardwarearchitekturen,* Teubner, Stuttgart 1983.
166. F. Vester: *Unsere Welt – ein vernetztes System,* Deutscher Taschenbuch Verlag, München 1983.
167. F. Vester: *Leitmotiv vernetztes Denken,* Heine, Verlag, München 1988.
168. M. Polke: "Informationshaushalt technischer Prozesse, in *Prozeßrechner 1984,* Springer Verlag, Berlin 1984.
169. H Zwerina, P. Haubner: "Gestaltung von Informationen auf Bildschirmen," in K. P. Fähnrich (ed.): *Software-Ergonomie,* R. Oldenbourg Verlag, München 1987.
170. H. Zwerina: "Masken und Formulare," in H. Balzert (ed.): *Einführung in die Software-Ergonomie,* De Gruyter, Berlin 1988, pp. 163 – 174.
171. W. Hacker: "Software-Ergonomie: Gestalten rechnergestützter geistiger Arbeit, in W. Schänpflug (ed.): *Software-Ergonomie 87,* Teubner, Stuttgart 1987, pp. 31 – 54.
172. W. Ahrens, M. Polke: "Netzmodelle als systemtechnische Informationsbasis für die Prozeßleittechnik," *Automatisierungstechnische Praxis atp* **37** (1989) no. 3, 94 – 103; **37** (1989) no. 4, 138 – 144.
173. B. Shneiderman: "Direct Manipulation: A Step beyond Programming Languages," *IEEE Computer* **16** (1983) Nov. 8, 57 – 69.
174. R. Ilg, J. Ziegler: "Direkte Manipulation, in H. Balzert (ed.): *Einführung in die Software-Ergonomie,* De Gruyter, Berlin 1988, pp. 135 – 194.
175. K.-P. Fähnrich, J. Ziegler: "Direkte Manipulation als Interaktionsform an Arbeitsplatzrechnern," in H. J. Bullinger (ed.): *Software-Ergonomie 85,* Teubner, Stuttgart 1985.
176. K. Klose: "Software-Entwicklung für Prozeßleitsysteme," *NAMUR-Statusbericht 87, Prozeßleittechnik für die Chemische Industrie,* R. Oldenbourg Verlag, München 1987, pp. 224 – 229.

177. CAE 2000: Bayerinterner Code für die Modellierung der Ingenieurfunktion, Leverkusen 1989 – 1992.
178. M. Chaineux: "Ausrichtung und Konception von DV-Anwendungssystemen auf der Basis von Unternehmungsmodellen, CAE 2000," *GMA Kongreß,* Dresden, Sept. 1993.
179. A. L. Luft: *Informatik als Technik-Wissenschaft. Eine Orientierungshilfe für das Informatik-Studium. BI,* Wissenschaftsverlag, Mannheim 1988.
180. R. Hagemüller, L. Max, J. Wolf: "Das Entwicklungssystem DOMINO von Siemens," in T. Gutzwiller, H. Österle (eds.): *Anleitung zu einer praxisorientierten Software-Entwicklungsumgebung,* vol. 2, Angewandte Informationstechnik Verlag, Hallbergmoos 1988, pp. 41 – 61.
181. W. Ahrens: *Objekt-orientierte Wissensrepräsentation in der Prozeßleittechnik. Prozeßrechensysteme '88,* Springer Verlag, Berlin 1989, pp. 498 – 513.
182. E. F. Cox: *Object Oriented Programming,* Addison Wesley, Reading, Mass., 1986.
183. H. Stoyan: *Programmiermethoden der Künstlichen Intelligenz,* **vol. 1 (1988),** vol. 2 (1991), Springer Verlag, Berlin.
184. E. Ignatowitz: *Chemietechnik,* Verlag Europa-Lehrmittel, Wuppertal 1982.
185. G. Frege in G. Patzing (ed.): *Funktion, Begriff, Bedeutung. Fünf logische Studien,* G. Patzing Verlag, Göttingen 1966.
186. DIN 32 705, Klassifikationssysteme, Beuth-Verlag, Berlin 1987.
187. I. Rasmussen: The Role of Hierarchical Knowledge Representation in Decision Making on System Management. IEEE Transactions on Systems, *Man and Cybernetics,* **SMC-15** (1985) no. 2, 234 – 243.
188. F.-J. Kaufels: *Rechnernetzwerk-Systemarchitekturen und Datenkommunikation,* Wissenschaftsverlag, Mannheim 1989.
189. R. Perne, M. Polke: "Produktentwicklung durch Verfahrensvariation," *Regelungstechnik* **30** (1982) no. 5, 147 – 156.
190. DIN 19 222, Leittechnik, Begriffe, Beuth-Verlag, Berlin 1985.
191. B. Polke, G. Rauprich: *Operatoren zur Beschreibung zeitabhängiger Zustandsänderungen,* in press.
192. T. Suckut: "Informationsmodell für Mehrprojekt-Management," Dissertation RWTH, Aachen 1993.
193. E. Schnieder: *Prozeßinformatik, Einführung mit Petri-Netzen,* Vieweg & Sons, Braunschweig 1986.
194. C. A. Petri: "Kommunikation mit Automaten," Dissertation, TH Darmstadt 1961.
195. W. Reisig: *Petri-Netze,* Springer Verlag, Berlin 1982.
196. E. Schnieder, H. Gückel: "Petri-Netze in der Automatisierungstechnik," *Automatisierungstechnik at* **37** (1989) no. 5, 173 – 181;no. 6, 234 – 241.
197. D. Abel: "Modellbildung und Analyse diskret gesteuerter Systeme mit Petri-Netzen," *Automatisierungstechnik at* **36** (1988) no. 12, 455 – 462.
198. W. Ahrens, M. Polke: "Informationsstrukturen in der Automatisierungstechnik," in G. Schmidt, H. Steusloff (eds.): "Mit vernetzten, intelligenten Komponenten zu leistungsfähigen Meß- und Automatisierungssystemen", *INTERKAMA-Kongreß 89,* R. Oldenbourg Verlag, München 1989, pp. 47 –67.
199. E. C. Yamalidou, J. C. Kantor: "Modeling and Analysis of Discrete-Event Processes Using Petri Net Theory" *ATCHE Annual Meeting,* San Francisco 1989.
200. P. Winkler: "Anforderungsbeschreibung mit Netzmodellen," *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 1, 32 – 39.
201. J. C. Kantor, E. P. Patsidou: "On the Sceduling and Control of Multipurpose Batch Plants," *AICHE Annual Meeting 1989,* Paper no. 1436, San Francisco 1989.
202. A. Hartramph: personal communication 1989.
203. Anlagen- und Apparatekennzeichen der Bayer-AG, Werknorm 9060, Leverkusen 1989.
204. *KKS-Kraftwerk-Kennzeichensystem,* 4th ed., VGB Kraftwerkstechnik GmbH, Essen 1991.
205. DIN 28 004, Fließbilder Verfahrenstechnischer Anlagen, Beuth Verlag, Berlin 1977.
206. TGL 25 000, März 1974 Verfahrenstechnik, Grundoperationen, Klassifikation.
207. *Ullmann,* 4th ed., **4,** 1.
208. R. J. Uhlig: "Erstellen von Ablaufsteuerungen für Chargenprozesse mit wechselnden Rezepturen, *Automatisierungstechnische Praxis (atp)* **29,** (1987) no. 1, 17 – 23.
209. J. Lauber, M. Polke: "Das Phasenmodell der Produktion: Ein semantisches CAE-Werkzeug zur Prozeßbeschreibung," *Regelungstechnischen Kolloquium,* Boppard, March 1993.
210. O. Graf: "Standardisierte Grundfunktionen verringern Softwareaufwand bei Teleperm M," *Eng. & Autom.* **3** (1991) 2 – 5.

211. U. Epple: "Funktionsorientierte Strukturierung prozeßleittechnischer Aufgaben." *Automatisierungstechnische Praxis atp* **31** (1989) no. 10, 468 – 474.
212. U. Epple, H. Kopec, R. Schmidt; "Strukturierung von Prozeßführungsaufgaben und Leitsystemsoftware," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 10, 468 – 474.
213. T. G. Fisher: "Batch Control Systems – Design, Application and Implementation, ISA, Paris 1990.
214. F. J. Kersting et al.: "Anforderungen an Systeme zur Rezeptfahrweise," NAMUR Empfehlung NE 25, Version vom 19.5.1992.
215. M. Polke, F. Portele: "Stand und Entwicklung der Prozeßleittechnik für die Textilindustrie," *Melliand Textilber.* **65** (1984) 291 – 297.
216. M. Polke: "Produktionsplanung in der Chemischen Industrie, in D. Hahn, G. Laßmann (eds.): *Produktionswirtschaft,* vol. 2, Physica-Verlag, Heidelberg 1989, pp. 369 – 413.
217. M. Polke: "Prozeßleittechnik: Trend zur Automatisierung in der Chemischen Industrie hält ungebrochen an Achema 91," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 5, 219 – 222.
218. Bayer AG: CAE 2000 Bayerinterne Dokumentation, Leverkusen 1992.
219. M. Polke: "Information als kritische Ressource," in A. Henzler (ed.): *Handbuch Strategische Führung,* Betriebswirtschaftlicher Verlag, Wiesbaden 1988, pp. 353 – 378.
220. G. Schmidt: "Was sind und wie entstehen komplexe Systeme und welche spezifischen Aufgaben stellen sie für die Regelungstechnik?", *Regelungstechnik* **30** (1982) no. 10, 331 – 339.
221. W. Hofmann: "Aufgabe der Produktionsleitebene in der Chemischen Industrie," *NAMUR-Statusbericht 85, Prozeßleittechnik für die Chemische Industrie,* R. Oldenbourg Verlag, München 1985, pp. 44 – 50.
222. R. W. Peters: "Informationshaushalte in Labor und Produktion," *NAMUR-Statusbericht 85, Prozeßleittechnik für die Chemische Industrie,* R. Oldenbourg Verlag, München 1985, 26 – 35.
223. S. Riemer: "Funktionen der Produktionslogistik in der Verfahrenstechnik," in O. Schmidt, H. Steusloff (eds.): "Mit vernetzten, intelligenten Komponenten zu leistungsfähigen Meß- und Automatisierungssystemen, *INTERKAMA-Kongreß 89,* R. Oldenbourg Verlag, München 1989, pp. 381 – 390.
224. B. Sturm, B. Will, M. Polke: "Die NAMUR – gestern, heute und morgen," *Automatisierungstechnische Praxis (atp)* **30** (1988) no. 1, 11 – 29.
225. K. F. Früh: "Optimalstruktur dezentraler Prozeß-leitsysteme," *RTP Regelungstech. Praxis* **21** (1979) 273 ff.
226. W. Körner, P. Moll, M. Polke: "Zur funktionalen Gliederung von Leitsystemen, *Automatisierungstechnische Praxis (atp)* **30** (1987) no. 9, 393 – 400.
227. W. Kaufmann (ed.): *Diagnostische Entscheidungsprozesse in der inneren Medizin,* Schattaur Verlag, Stuttgart 1986.
228. M. Wienand: "Mündliche und unveröffentlichte Hinweise auf das DV-Phasenmodell bei der Konzeption von Unternehmens-Informationsmodellen," Bayer AG, Leverkusen 1974 ff.
229. W. Büsing: *Ergebnis vieler Diskussionen zur Klärung des von Polke/Wienand intuitiv vorgeschlagenen Phasenmodells,* personal communication, 1983.
230. E. D. Gilles, M. Wurst: "Prozeßführung zur Sicherung reproduzierbarer Qualität in der verfahrenstechnischen Produktion," lecture held on *GVC-Jahrestreffen,* Tagungsband, Stuttgart 1990, pp. 138, 139.
231. E. D. Gilles: "Prozeßführung zur Sicherung reproduzierbarer Qualität in der verfahrenstechnischen Produktion," *Chem. Ing. Tech.* **63** (1991) no. 9, 910 – 918.
232. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie – Status und Trend," *Chem.-Ing.-Tech.* **58** (1986) 557 – 564.
233. K. H. Padberg, P.-Th. Wilrich: "Die Auswertung von Daten und ihre Abhängigkeit von der Merkmalsart, part 1, Skalentypen," *QZ* **26** (1981) no. 6, 179 ff.
234. H. Rumpf: "Über die Eigenschaften von Nutzstäuben," *Staub Reinhalt. Luft* **27** (1967) 3 – 13.
235. R. Ecker, H. Kramer, H. K. Müller, M. Polke: "Neue Aspekte bei der Prüfung von Kautschuk und Gummi," *Kautsch. Gummi Kunstst.* **25** (1972) no. 1, 5 – 10.
236. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie – Status und Trend, *Automatisierungstechnische Praxis atp* **28** (1986) no. 9, 423 – 437; no. 10, 479 – 484.
237. M. Polke: "CAE zwischen Betreibern und Anlagenbauern," *Partnerschaft im*

Anlagenbau, Haus der Technik, Essen, Nov. 1992.
238. M. Polke: "Prozeßleittechnik; Sicherheit und Qualität in der Prozeßindustrie," *Vortrag auf dem VDE-Kongreß 1993,* Berlin, Januar 1993.
239. J. König: "Sicherheitstechnische Aspekte der Prozeßleittechnik," *Chem. Ing. Tech.* **59** (1987) no. 3, 196 – 204.
240. B. Bieker, H. Rückert: *Entwicklung eines Expertensystems zur Prozeßführung von Bioreaktoren einer Kläranlage der Bayer AG,* Leverkusen, Aug.1989.
241. DIN 40 719, Part 6, Schaltungsunterlagen, Regeln und graphische Symbole für Funktionspläne, Beuth-Verlag, Berlin, Mai 1992.
242. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie – Status und Trend," *atp* **28** (1986) no. 6, 423 – 431.
243. K. H. Padberg, P.-Th. Wilrich: "Die Auswertung von Daten und ihre Abhängigkeit von der Merkmalsart," *Z. Qualität Sicherheit* **26** (1981) no. 179–183.
244. M. Polke: "Prozeßleittechnik für die Chemie – Status und Trend," *Automatisierungstechnische Praxis atp* **28** (1985) no. 5, 214 – 223.
245. E. D. Gilles: "Auf dem Wege zu einer modellgestützten Prozeßleittechnik," *Automatisierungstechnische Praxis atp* **31** (1888) 265 – 270, 441 – 449.
246. R. Perne, M. Polke: "Produktentwicklung durch Verfahrensvariation," *Regelungstechnik* **30** (1982) no. 5, 147 – 156.
247. I. L. Casti: *Dynamical Systems and their Applications,* Academic Press, New York 1977.
248. C. T. Chen: *Introduction to Linear System Theory,* Holt Rinehart and Winston, New York 1970.
249. L. A. Zadeh, C. A. Desoer: *Linear System Theory,* Mc Graw-Hill, New York 1963.
250. R. Unbehauen: *Systemtheorie,* R. Oldenbourg Verlag, München 1969.
251. R. E. Kalman, P. L. Falb, M. Arbib: *Topics in Mathematical Systems Theory,* Mc Graw-Hill, New York 1969.
252. L. Padulo, M. Arbib: *System Theory,* Hemisphere, Washington 1974.
253. F. Pichler: *Mathematische Systemtheorie: Dynamische Konstruktionen,* De Gruyter, Berlin 1975.
254. H. W. Knobloch, H. Kwakernaak: *Lineare Kontrolltheorie,* Springer Verlag, Berlin 1985.
255. S. Wendt: *Nichtphysikalische Grundlagen der Informationstechnik,* Springer Verlag, Berlin-Heidelberg-New York 1989.
256. W. Hahn: *Theorie und Anwendung der direkten Methode von Ljapunow,* Springer Verlag, Berlin 1959.
257. W. Hahn, *Stability of Motion,* Springer Verlag, Berlin 1967.
258. J. La Salle, S. Lefschetz: *Die Stabilitätstheorie von Ljapunow,* Bibliographisches Institut, Mannheim 1967.
259. H. Haken: *Synergetik,* Springer Verlag, Berlin 1983.
260. T. Kailath: *Linear Systems,* Prentice Hall, Engelwood, N. J., 1980.
261. E. Gilbert: "Controllability and Observability in Multivariable Control Systems," *J. Soc. Ind. Appl. Math. Ser. A* **1** (1963) 128 – 151.
262. R. E. Kalman: "On the General Theory of Control Systems," *Proc. 1st. Cong. Autom. Control,* 1960, 481 – 493.
263. R. E. Kalman: "Mathematical Description of Linear Systems," *J. Soc. Ind. Appl. Math. Ser. A* **1** (1963) 152 – 192.
264. R. E. Kalman: *Lectures on Controllability and Observability,* C.I.M.E., Bologna 1968.
265. D. G. Luenberger: "Observers for Multivariate Systems," *IEEE Trans. Autom. Control* **AC-11** (1966) 190 – 197.
266. D. G. Luenberger: "An Introduction to Observers," *IEEE Trans. Autom. Control* **AC-16** (1971) 596 –603.
267. R. Metz: *Auswertesysteme. Der Physiker in der Automatisierungstechnik,* Deutsche Physikalische Gesellschaft e.V., Bad Honnef, 1989, pp. 125 – 141.
268. W. Masing: *Handbuch der Qualitätssicherung,* Hanser Verlag, München – Wien 1988.
269. H. Stachowiak: *Allgemeine Modelltheorie,* Springer Verlag, Berlin 1973.
270. K. R. Popper: *Logik der Forschung,* J. C. B. Mohr (Paul Siebeck), Tübingen 1982.
271. R. Courant: "Variational Methods for the Solution of Problems of Equilibrium and Vibration," *Bull. Am. Math. Soc.* **49** (1943) 1 – 23.
272. F. A. Aseltine et al.: "A Survey of Adaptive Control Systems," *IRE Trans. Autom. Control* **6** (1958) 102 –108.
273. R. E. Kalman: "Design of a Self-Optimizing Control System," *Trans. ASME* **80** (1958) 468 – 478.
274. A. Rosenbluth, N. Wiener: "The Role of Model in Science," *Philos. Science* **10** (1945), 316 – 321.

275. P. Eykhoff: *System Identification,* J. Wiley, London 1974.
276. K. E. Brennan, S. L. Campbell, L. R. Petzold: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations,* North-Holland, Amsterdam 1989.
277. C. W. Gear: "Simultaneous Numerical Solution of Differential-Algebraic Equations," *IEEE Trans. Circuit Theory* **18** (1971) 89 – 95.
278. R. Bachmann, L. Brüll, U. Pallaske: "A Contribution of the Numerical Treatment of Differential Algebraic Equations Arising in Chemical Engineering," DECHEMA-*Monogr.* **116** (1989) 343 – 349.
279. L. R. Petzold: "Differential-Algebraic Equations are not ODE's," *SIAM J. Sci. Stat. Comput.* **3** (1982) 367 – 384.
280. C. W. Gear, L. R. Petzold: "ODE Methods for the Solution of Differential-Algebraic Systems," *SIAM J. Numer. Anal.* **21** (1984) 716 – 728.
281. R. E. Kalman, R. S. Bucy: "New Results in Linear Filtering and Prediction Theory," *J. Basic. Eng.* **83** (1961) 95 – 108.
282. G. O. Kallianpu: *Stochastic Filtering Theory,* Springer Verlag, Berlin 1980.
283. G. Doetsch: *Einführung in Theorie und Anwendung der Laplace-Transformation,* Birkhäuser-Verlag, Berlin 1976.
284. O. Föllinger: *Laplace- und Fourier-Transformation,* Elitera-Verlag, Berlin 1977.
285. M. Polke: "Zur Physik der Elektrophotographie," in: *Festkörperprobleme VIII,* Vieweg/Pergamon Press, Oxford 1968, pp. 131 – 174.
286. W. G. Hunter, A. M. Reiner: *Technometrics* **7** (1965) 307 – 323.
287. A. C. Atkinson, V. V. Fedorov: *Biometrika* **62** (1975) 57 – 70.
288. F. J. Dumez, G. F. Froment: *Ind. Eng. Chem. Process Des. Dev.* **15** (1976) 291 – 301.
289. G. E. P. Box, W. J. Hill: "Discrimination Among Mechanistic Models," *Technometrics* **9** (1967) 57 –71.
290. J. Pawlowski: *Die Ähnlichkeitstheorie in der physikalisch-technischen Forschung,* Springer Verlag, Berlin 1971.
291. H. Görtler: *Dimensionsanalyse,* Springer Verlag, Berlin 1975.
292. A. Betz: "Ähnlichkeitstheorie und Modelltechnik," in: *Physikhüttte,* vol. I: "Mechanik," Ernst u. Sohn, Berlin-München-Düsseldorf 1971.
293. E. Eitelberg: "Modellreduktion durch Minimieren des Gleichungsfehlers," *Regelungstechnik* **26** (1978) 320 – 322.
294. F. D. Galiana: "On the Approximation of Multiple Input – Multiple Output Constant Linear Systems," *Int. J. Control* **17** (1973) no. 6, 1313 – 1324.
295. K. Glover: "All Optimal Hankel-Norm Approximations of Linear Multivariable Systems and Their L_∞-Error Bounds," *Int. J. Control* **39** (1985) no. 6, 1115 – 1193.
296. D. C. Hyland, D. S. Bernstein: "The Optimal Projection Equations for Model Reduction and the Relationship among the Methods of Wilson, Skelton and Moore," *IEEE Trans. Autom. Control* **AC-30** (1985) no. 12, 1201 – 1211.
297. L. Litz: *Reduktion der Ordnung linearer Zustandsraummodelle mittels modaler Verfahren,* Hochschul-Verlag, Stuttgart 1979.
298. B. C. Moore: "Principal Component Analysis in Linear Systems: Controllability, Observability, and Model Reduction," *IEEE Trans. Autom. Control* **AC-26** (1981) 17 – 31.
299. U. Pallaske: "Ein Verfahren zur Ordnungsreduktion mathematischer Prozeßmodelle," *Chem. Ing. Tech.* **59** (1987) no. 7, 604 – 605.
300. R. E. Skelton: "Cost Decomposition of Linear Systems With Application to Model Reduction," *Int. J. Control* **32** (1980) 1031 – 1055.
301. D. A. Wilson: "Optimum Solution of Model-Reduction Problem," *Proc. IEEE* **58** (1970) 1161 –1165.
302. E. Welfonder: "Leittechnikdokumentation aus Betreibersicht," *ETG/VGB-Fachtagung Betriebsgerechte Dokumentation in Kraftwerken, Einfluß der modernen Leittechnik und Planungsmittel,* Baden-Baden 1986.
303. E. Ignatowitz: *Chemietechnik,* Verlag Europa-Lehrmittel, Wuppertal 1980.
304. I. M. Horowitz: *Synthesis of Feedback Systems,* Academic Press, New York 1963.
305. W. Oppelt: *Kleines Handbuch technischer Regelvorgänge,* Verlag Chemie, Weinheim 1972.
306. O. Föllinger: *Regelungstechnik,* Hüthig, Heidelberg 1985.
307. O. Föllinger: *Optimierung dynamischer Systeme. Eine Einführung für Ingenieure,* Oldenbourg Verlag, München 1985.
308. J. C. Doyle, G. Stein: "Multivariable Feedback Design: Concepts for a Classical/Modern

Synthesis," *IEEE Trans. Autom. Control* **26** (1981) 4 – 16.
309. M. Morari, J. C. Doyle: "A Unifying Framework for Control System Design under Uncertainty and its Implications for Chemical Process Control," *Proc. Chem. Proc. Control* **III** (1986) 5 – 51.
310. J. S. Freudenberg, D. P. Looze: *Frequency Domain Properties of Scalar and Multivariable Feedback Systems,* Springer Verlag, Berlin 1988.
311. H. H. Rosenbrock: *Computer-Aided Control System Design,* Academic Press, London 1974.
312. D. H. Owens: *Feedback and Multivariable Systems,* Peter Peregrinus, Stevenage 1978.
313. A. G. J. MacFarlane, D. F. A. Scott-Jones: "Vector Gain," *Int. J. Control* **29** (1979) 65 – 91.
314. Y. Takahashi, M. J. Rabins, D. M. Auslander: *Control and Dynamic Systems,* Addison-Wesley, Reading, Mass. 1970.
315. R. V. Patel, N. Munro: *Multivariable Systems Theory and Design,* Pergamon Press, Oxford 1982.
316. U. Korn, H. Wilfert: *Mehrgrößenregelungen,* VEB Verlag Technik, Berlin 1982.
317. H. Tolle: *Mehrgrößen-Regelkreissynthese,* Oldenbourg Verlag, München 1983.
318. J. M. Maciejowski: *Multivariable Feedback Design,* Addison-Wesley, Reading, Mass., 1989.
319. A. Gelb (ed.): *Applied Optimal Estimation,* M.I.T. Press, Cambridge, Mass., 1974.
320. K. Brammer, G. Siffling: *Stochastische Grundlagen des Kalman-Bucy-Filters: Wahrscheinlichkeitsrechnung und Zufallprozesse,* Oldenbourg Verlag, München 1975.
321. K. Brammer, G. Siffling: *Kalman-Bucy-Filters Deterministische Beobachtung und stochastische Filterung,* Oldenbourg Verlag, München 1975.
322. H. Kwakernaak, R. Siwan: *Linear Optimal Control Systems,* J. Wiley, New York 1972.
323. B. D. O. Anderson, J. B. Moore: *Linear Optimal Control,* Prentice-Hall, Englewood Cliffs, NJ, 1972.
324. G. Stein, M. Athans: "The LQG/LTR Procedure for Multivariable Feedback Control Design," *IEEE Trans. Autom. Control* **32** (1987) 105 – 114.
325. A. G. J. MacFarlane, B. Kouvaritakis: "A Design Technique for Linear Multivariable Feedback Systems," *Int. J. Control* **25** (1977) 837 – 874.
326. Y. S. Hung, A. G. J. MacFarlane: *Multivariable Fedback: A Quasi-Classical Approach,* Springer Verlag, Berlin 1982.
327. C. E. Garcia, M. Morari: "Inernal Model Control: 1. A Unifying Review and Some New Results," *Ind. Eng. Chem. Process Des. Dev.* **21** (1982) 308 –323.
328. B. A. Francis: *A Course in H_∞ Control Theory,* Springer Verlag, Berlin 1987.
329. J. Doyle, K. Glover, P. Khargonekar, B. Francis: "State-Space Solutions to Standard H_2 and H_∞ Control Problems," *IEEE Trans. Autom. Control* **34** (1989) 831 – 847.
330. V. Manousiouthakis, R. Savage, Y. Arkun: "Synthesis of Decentralized Control Structures Using the Concept of Block Relative Gain," *Am. Inst. Chem. Eng.* **32** (1986) 991 – 1003.
331. D. D. Siljak: *Large-Scale Dynamic Systems,* North-Holland, New York 1978.
332. M. G. Singh, T. Titli: *Systems: Decomposition, Optimisation and Control,* Pergamon Press, Oxford 1978.
333. M. G. Singh: *Dynamical Hierarchical Control,* North-Holland, Amsterdam 1980.
334. M. G. Singh: *Decentralised Control,* North-Holland, Amsterdam 1981.
335. L. Litz: *Dezentrale Regelung,* Oldenbourg Verlag, München 1983.
336. C. J. Harris, S. A. Billings (eds.): *Self-tuning and Adaptive Control: Theory and Applications,* Peter Peregrinus, Stevenage 1981.
337. M. M. Gupta, C. T. Chen (eds.): *Adaptive Methods for Control System Design,* IEEE Press, New York 1985.
338. A. Isidori: *Nonlinear Control Systems: An Introduction,* Springer Verlag, Berlin 1985.
339. C. E. Garcia, D. M. Prett, M. Morari: "Model Predictive Control: Theory and Practice – a Survey," *Automatica* **25** (1989) 335 – 347.
340. D. K. Frederick, C. J. Herget, R. Kool, M. Rimvall: "ELCS – The Extended List of Control Software," unpublished, WGS, Eindhoven University of Technology, Eindhoven 1987.
341. M. Jamshidi, C. Herget: *Computer-Aided Control Systems Engineering,* North-Holland, Amsterdam 1985.
342. P. Holl, W. Marquardt, E. D. Gilles: "Diva – a Powerful Tool for Dynamic Process Simulation," *Comput. Chem. Eng.* **12** (1988) 421 – 425.

343. F. Allgöwer, J. Raisch: "Multivariable Controller Design for an Industrial Distillation Column," *Proceedings of 5th IMA International Conference on Control Theory,* 1988.
344. A. Duschek: *Vorlesungen über höhere Mathematik,* vol. 3, Springer Verlag, Wien 1960.
345. P. Funk: *Variationsrechnung und ihre Anwendung in Physik und Technik,* Springer Verlag, Berlin 1962.
346. E. Klingbeil: *Variationsrechnung,* Bibliographisches Institut, Mannheim 1977.
347. H. Schwarz: *Optimale Regelung linearer Systeme,* Bibliographisches Institut, Mannheim 1976.
348. W. Boltjanski: *Mathematische Methoden der optimalen Steuerung,* Carl Hanser Verlag, München 1972.
349. J. Hsu, A. Meyer: *Modern Control Principles and Applications,* part III, McGraw-Hill, New York 1968.
350. L. Pontrjagin et al.: *Mathematische Theorie optimaler Prozesse,* Oldenbourg Verlag, München 1967.
351. R. Bellmann: *Dynamic Programming,* Princeton University Press, Princeton 1965.
352. R. Bellmann, S. Dreyfus: *Applied Dynamic Programming,* Princeton University Press, Princeton 1964.
353. G. Schneider, H. Mikolcic: *Einführung in die Methode der dynamischen Programmierung,* Oldenbourg Verlag, München 1972.
354. M. Athans, P. Falb: *Optimal Control,* McGraw-Hill, New York 1966.
355. G. Bieß et al.: *Optimale Prozesse und Systeme,* Verlag Harry Deutsch, Thun – Frankfurt 1980.
356. A. Bryson, Y. Ho: *Applied Optimal Control,* J. Wiley, New York 1975.
357. E. Hofer, R. Lunderstädt: *Numerische Methoden der Optimierung,* Oldenbourg Verlag, München 1975.
358. U. Hoffmann, H. Hofmann: *Einführung in die Optimierung,* Verlag Chemie, Weinheim 1971.
359. H. Tolle: *Optimierungsverfahren,* Springer Verlag, Berlin 1971.
360. K. Stahl: *Industrielle Steuerungstechnik in schaltalgebraischer Behandlung,* Oldenbourg Verlag, München 1965.
361. K. H. Fasol: *Binäre Steuerungstechnik,* Springer Verlag, Berlin 1988.
362. R. Lauber: *Prozeßautomatisierung,* vol. 1, Springer Verlag, Berlin 1989.
363. W. Körner, P. Moll, M. Polke: "Zur funktionalen Gliederung von Leitsystemen," *Automatisierungstechnische Praxis (atp)* **29** (1987) no. 9, 393 – 400.
364. A. H. Levis: "Challenges to Control: a Collective View," *IEEE Trans. Autom. Control* **AC-32** (1987) no. 4, 275 – 285.
365. B. Rosenstengel, U. Wienand: *Petri-Netze – Eine anwendungsorientierte Einführung,* Vieweg Verlag, Wiesbaden 1982.
366. D. Bochmann, C. Posthoff: *Binäre dynamische Systeme,* Oldenbourg Verlag, München 1981.
367. R. Herschel: *Einführung in die Theorie der Automaten, Sprachen und Algorithmen,* Oldenbourg Verlag, München 1974.
368. E. Leonhardt: *Grundlagen der Digitaltechnik. Eine systematische Einführung,* Carl Hanser Verlag, München 1982.
369. S. Wendt: *Entwurf komplexer Schaltwerke,* Springer Verlag, Berlin 1974.
370. M. Brombacher, M. Polke: "Perspektiven der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* 29 (1987) 501 – 509, 558 – 567, 1987, pp. 22 – 35.
371. M. Niese, EP 0 104 411, 1986.
372. M. Niese: "Struktureller Aufbau und dezentrale Führung in Produktionsanlagen," *INTERKAMA Kongreß,* München 1989, pp. 339 – 349.
373. M. Bruns, M. Eibl, F. J. Kersting, R. J. Uhlig: "Anforderungen an Systeme zur Rezeptfahrweise," *Automatisierungstechnische Praxis (atp)* **35** (1993) no. 1, 40 – 44;no. 2, 104 – 108.
374. S. Riemer: "Funktionen der Produktionslogistik in der Verfahrenstechnik," *INTERKAMA Kongreß,* München 1989, pp. 381 – 390.
375. K.-F. Geibig: "Funktionen der Betriebsleitebene," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 2, 68 – 72.
376. M. Polke: *Prozeßleittechnik,* Universität Stuttgart, 1984 ff.
377. L. Prins, J. Stäheli: "Zur Automatisierung von diskontinuierlichen Vielzweckanlagen in Chemiebetrieben," *Regelungstech. Prax.* **20** (1978) no. 9, 256 –261.
378. U. Epple, H. Kopec, R. Schmidt: "Strukturierung von Prozeßführungsaufgaben und Leitsystemsoftware," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 2, 59 – 67.
379. U. Epple: "Die Bausteintechnik – Grundlage einer objektorientierten Netzstruktur für die Prozeßleittechnik, GMA-Fachbericht," *VDE-Kongreß,* Berlin 20.– 23.1.1993.

380. E. Pavlik: "Sensoren als Quelle für den Informationshaushalt technischer Prozesse," *46. NAMUR Hauptsitzung,* Darmstadt, 7 – 9 Dec. 1983, 48. Physikertagung and Frühjahrstagung, 15 March 1984, Münster, pp. 1322 ff, Physik-Verlag GmbH, Weinheim 1984.
381. R. Weisbeck: "Sensoren für die Chemische Industrie – eine kritische Bewertung des Standes der Technik und derzeitiger Sensorentwicklung," *46. NAMUR Hauptsitzung,* Darmstadt, 7 – 9 Dec. 1983.
382. H. Warncke: "Sensortechnik in der Chemischen Industrie: Aufgaben und Möglichkeiten der Stoffgrößenmeßtechnik," *Tech. Mess.* 52 (1985) no. 4.
383. H. Raab: "Digitale Sensoren und Sensorsysteme – Welche Informationen sollen sie liefern?," *Automatisierungstechnische Praxis (atp)* **30** (1988) no. 11, 534 – 544.
384. VDI/VDE 2600, Blatt 3: Metrologie, Gerätetechnische Begriffe, Beuth Verlag, Berlin 1973.
385. H. Warncke: Vorteile, Probleme und Möglichkeiten der in-situ-Messung von Stoffgrößen, Bayer-interner Bericht, Leverkusen, Nov. 1982.
386. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie – Status und Trends," *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 9/10, 423 – 432, 479 – 484.
387. *Ullmann,* 4th ed., **5,** pp. 797 – 890.
388. H.-J. Schneider: "Sensorsysteme für die Betriebsmeßtechnik und Kommunikation im Feld," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 10, 511 – 528.
389. K. Hartmann: "On-line Analytik aus funktionaler Sicht, *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 11, 510 – 515.
390. H. Schuler, Ch.-U. Schmidt: "Modellgestütztes Messen in chemischen Reaktoren," *Chem. Ing. Tech.* **63** (1991) no. 3, 189 – 203.
391. R. Perne, M. Polke: "Produktentwicklung durch Verfahrensvariation," *Regelungstech. (rt)* **30** (1982) no. 5, 147 – 156.
392. K. Garbrecht, A. Prommer: *Mikroelektronik. Eine Technik macht Epoche,* Siemens, Berlin 1981.
393. M. Polke: "Informationshaushalt technischer Prozesse," *Automatisierungstechnische Praxis (atp)* **27** (1985) no. 4, 214 – 217.
394. J. Hesse: *Sensoren auf dem Weg zur Systemintegration,* special edition, Oldenbourg Verlag, München 1992.
395. J. Lauber: "Das Phasenmodell der Produktion: Ein semantisches CAE-Werkzeug zur Prozeßbeschreibung," *Regelungstechnischen Kolloquium,* Boppard, Germany, Feb. 1993.
396. International Union for Pure and Applied Physics (IUPAP): Symbols, Units and Nomenclature in Physics, Document U.I.P.20 (1978), Physik Verlag GmbH, Weinheim 1980.
397. DIN 1313, Physikalische Größen und Gleichungen. Begriffe und Schreibweisen, Beuth Verlag, Berlin 1978.
398. M. Polke: "Erforderliche Informationsstrukturen für die Qualitätssicherung," *Chem. Ing. Tech.* **65** (1993) no. 7, 791 – 796.
399. W. König: Sicherheitstechnische Aspekte zur Prozeßleittechnik, *12. Arbeitsschutz-Symposium der Bayer AG,* Leverkusen, 22 Nov. 1985.
400. DIN V 19 250, Messen – Steuern – Regeln. Grundlegende Sicherheitsbetrachtungen für MSR-Schutzeinrichtungen, Beuth Verlag, Berlin 1989.
401. NAMUR-Empfehlung 31: Anlagensicherung mit Mitteln der Prozeßleittechnik, NAMUR, Darmstadt 1992.
402. Verordnung zur Durchführung des Bundesimmissionsschutzgesetzes (Störfall-Verordnung) of 19 May 1988 and 5 July 1991.
403. DIN VDE 31 000: Allgemeine Leitsätze für das sicherheitsgerechte Gestalten technischer Erzeugnisse. Begriffe der Sicherheitstechnik. Grundbegriffe, part 2, Beuth Verlag, Berlin 1987.
404. 2nd IUPAC-Workshop on Safety in Chemical Production, Pacifico Yokohama, Japan, May 31 – June 4, 1993.
405. M. Polke: "Process Automation in Chemical Industry," *2nd IUPAC-Workshop on Safety in Chemical Production,* Yokohama, Japan, June 1993.
406. R. Ecker, H. Kramer, H. K. Müller, M. Polke: "Neue Aspekte bei der Prüfung von Kautschuk und Gummi," *Kautsch. Gummi Kunstst.* **25** (1972) no. 1, 5 – 10.
407. M. Mooney, H. Rolla, H. Taylor, J. H. Fielding: "Development and Standardization of Test for Evaluating Processibility of Rubber," *Symposium on Rubber Testing,* American Society for Testing Special Technical Materials, 1974.
408. DIN 53 523, Bestimmung der Plastizität nach Mooney im Warmscherversuch, Beuth Verlag, Berlin 1965.

409. ASTM D927/49 T, Method of Test for Viscosity of Rubber and Rubberlike Material by the Sharing Disc-viscometer, 1962.
410. W. Kuhn: "Beziehungen zwischen Molekülgröße, statistischer Molekülgestalt und elastischen Eigenschaften hochpolymerer Stoffe," *Kolloid-Z.* **76** (1936) 3.
411. W. Pechold, S. Blasenbrey, S. Woerner, *Kolloid-Z. Polym.* **189** (1965) 14.
412. L. Morbitzer, H. J. Kreß, C. Lindner, K. H. Ott: "Struktur und Eigenschaften von Mehrphasenkunststoffen," *Angew. Makromol. Chem.* **132** (1985) 19 – 42.
413. L. Morbitzer, D. Kranz, G. Humme, K. H. Ott: "Structure and Properties of ABS Polymers. X. Influence of Particle Size and Graft Structure on Loss Modulus Temperature Dependence and Deformation Behavior," *J. Appl. Polym. Sci.* **20** (1986) 2691 –2704.
414. L. Morbitzer, U. Grigo: "Die wichtigsten polymerphysikalischen Aspekte des Polycarbonats," *Angew. Makromol. Chem.* **162** (1988) 87 – 107.
415. K. Ninomiya, J. Furuta: "Interrelation Between Data from Different Physical Tests," *Annual Meeting of IISPR,* Mexico City, May 1970.
416. J. D. Ferry: *Viscoelastic Properties of Polymers,* John Wiley & Sons, New York 1961.
417. DIN ISO 9001, Qualitätssicherungssysteme – Modell zur Darlegung der Qualitätssicherung in Design/Entwicklung, Produktion, Montage und Kundendienst, Beuth Verlag, Berlin 1990.
418. K. H. Padberg, P.-Th. Wilrich: "Die Auswertung von Daten und ihre Abhängigkeit von der Merkmalsart," *QZ* **26** (1981) no. 6, 179 – 184, 210 –214.
419. DIN ISO 9000, Qualitätsmanagement- und Qualitätssicherungsnormen, Leitfaden zur Auswahl und Anwendung, Beuth Verlag, Berlin 1990.
420. R. E. Kalman: "A New Approach to Linear Filtering and Prediction Problems," *J. Basic. Eng.* **82D** (1960) 35 – 45.
421. D. G. Luenberger: "Observers for Multivariable Systems," *IEEE Trans. Autom. Control* **11** (1966) 190 – 197.
422. A. Gelb (ed.): *Applied Optimal Estimation,* The MIT Press, Cambridge, Mass., 1974.
423. W. Brammer, C. Siffling: *Kalman-Bucy-Filter. Deterministische Beobachtung und stochastische Filterung,* R. Oldenbourg Verlag, München 1975.
424. E. D. Gilles: "Modellgestützte Verfahren in der Automatisierungstechnik," *Spring Meeting of DPG, Verhandlungen der Deutschen Physikalischen Gesellschaft,* series VI, vol. 24 (1989).
425. N. R. Iudica: "Expertensysteme," *RTP Regelungstech. Prax.* **26** (1984) no. 9, 402 – 407.
426. D. Balzer: *Wissensbasierte Systeme in der Automatisierungstechnik,* Hanser Verlag, München 1991.
427. H. Fittler: "Anwendung von Expertensystemen in der Prozeßautomation," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 4, 199 – 205.
428. H.-P. Preuß: "Fuzzy Control – heuristische Regelung mittels unscharfer Logik," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 4, 176 – 184, 239 – 246.
429. M. Koch, T. Kuhn, J. Wernstedt: "Ein neues Entwurfskonzept für Fuzzy-Regelungen," *Automatisierungstech. (at)* **41** (1993) no. 5, 152 – 158.
430. H. Kiendl, K. D. Meyer-Gramann, B. Cuno, M. Fritsch, Chr. Frenck: "Fuzzy Control," Beiträge zur "Theorie für Anwender," *Automatisierungstechnik (at)* **41** (1993) no. 2. A 5.
431. W. Brauer, D. Hernàndes: "Verteilte künstliche Intelligenz und kooperatives Arbeiten," *4. Internationaler GI-Kongreß Wissensbasierte Systeme,* München, Oct. 1991. *Informatik Fachberichte,* Springer Verlag, Berlin 1991.
432. E. D. Gilles: "Modellbildung, ein Hilfsmittel der Meßtechnik," *Tech. Mess.* **46** (1979) 225 – 232, 271 –274.
433. E. D. Gilles: "Application of Model Based Measuring Techniques in Chemical Engineering," *Proceedings of International Conference on Industrial Process Modelling and Control,* Hangzhou, China, 6 – 9 June, 1985.
434. *Ullmann,* 4th ed., **5.**
435. T. Pfeifer, P. Profos (eds.): *Handbuch der Industriellen Meßtechnik,* R. Oldenbourg Verlag, München 1992.
436. K. H. Feltgen: personal communication, 1972.
437. K. H. Schmitt: "Die Automatisierung von Chemieanlagen mit Mitteln der Systemtechnik – Analysen und Trends," *Chem. Ing. Tech.* **53** (1981) no. 8, 620 –625.
438. Knick Elektronische Meßgeräte GmbH & Co.: "Meß-System 71 pH," company brochure no.717 959 150 00, Berlin 1993.
439. K. Schügerl (ed.): *Analytische Methoden in der Biotechnologie,* Vieweg Verlag, Braunschweig 1991.

440. J. E. Bailey, D. F. Ollis: *Biochemical Engineering Fundamentals,* McGraw-Hill Book Company, New York 1986.
441. U. Martens: "Sensor-Systeme für Durchfluß und Füllstand," *55. NAMUR* Lahnstein Nov. 1992.
442. E. Nicklaus: "Auswirkungen moderner Prozeß-leittechnik auf Sensorsysteme," *Chem. Ing. Tech.* **59** (1987) no. 8, 615 – 621.
443. D. Hesse, H. Küttner: "Entwicklungstendenzen der Sensortechnik," *Ind.-Elektronik + Elektronik* **28** (1983) no. 5, 36 – 44.
444. M. Polke: "On-line Betriebsmeßtechnik," Einführungsvortrag in den Themenkreis der gleichnamigen Postersession auf dem Jahrestreffen der GVC, Wien, Oct. 1, 1992.
445. DIN 31 051, Instandhalten: Begriffe und Maßnahmen, Beuth Verlag, Berlin 1985.
446. G. Pinkowski: "Instrumentierungstechnik im Wandel," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 5, 247 – 254, 325 – 327. G. Pinkowski: "Fortführung der Ex-Feldbus-Untersuchung," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 11, 617 – 622. G. Pinkowski: "Einheitliche, herstellerunabhängige Anzeige-Bedienoberfläche für kommunikationsfähige Meßumformer," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 2, 530 – 537.
447. H. Drahten: "Der Feldbus aus Sicht des Anwenders," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 2, 55-64.
448. H. Raab: "Die Rolle der Sensoren in der Verfahrenstechnik," *Chem. Ing. Tech.* **61** (1989) 199 –205.
449. W. Stieler: "Sensorprinzipien zur Messung physikalischer Produkteigenschaften," company report, Bayer AG, Leverkusen, Nov. 1982.
450. H. Raab: "Zur Struktur der Anzeige-Bedienoberfläche digitaler Sensorsysteme," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 2, 64 – 71.
451. ISO: MMS 9506, part 1, Service Definitions; MMS, part 2, Protocol Specification, 1988.
452. C. Nøkleby: *The MMS Implementation on the P-NET Fieldbus System, 2nd International Conference on P-Net Fieldbus Systems,* Deggendorf 1992.
453. A. Valenzano, C. Demartini, L. Climiniera: *MAP and TOP Communications, Standards and Applications,* Addison-Wesley Publishing Company, Workingham 1992.
454. W. Stadter: "Device-Description-Language. Ein wichtiger Beitrag zur herstellerunabhängigen Austauschbarkeit von Feld- und Bediengeräten," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 1, 21 ff.
455. VDI/VDE Richtlinie: Einheitliche Anzeige- und Bedienoberfläche auf Personalcomputern für digitale Feldgeräte, Gründruck 1992.
456. Krohne America Inc.: "Flowmeter Selection Guide," company brochure, Peabody, Mass. 1992.
457. P. Profos, T. Pfeifer: *Handbuch der Industriellen Meßtechnik,* R. Oldenbourg Verlag, München 1992.
458. G. Strohrmann: *Einführung in die Meßtechnik im Chemiebetrieb,* R. Oldenbourg Verlag, München 1990.
459. I. Hengstenberg, B. Sturm, O. Winkler: *Messen, Steuern und Regeln in der chemischen Technik,* Springer Verlag, Berlin 1980.
460. F. Kohlrausch: *Praktische Physik,* Teubner Verlagsgesellschaft, Leipzig 1950.
461. A. Mangold: "Stellgeräte für verfahrenstechnische Anlagen," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 11, 585 ff.
462. M. Huk: "Stellgeräte für verfahrenstechnische Anlagen," *Automatisierungstechnische Praxis (atp)* **3** (1993) no 3, 131 ff.
463. H.-J. Schneider: "Sensorsysteme und die Kommunikation im Feld," *Automatisierungstechnische Praxis (atp)* **35** (1993) no. 2, 71 ff.
464. DIN 19 227, Graphische Symbole und Kennbuchstaben für die Prozeßleittechnik. T1: Darstellung von Aufgaben, Beuth Verlag, Berlin 1990 (draft).
465. DIN 40 050, IP-Schutzarten. Berührungs-, Fremdkörper- und Wasserschutz für elektrische Betriebsmittel, Beuth Verlag, Berlin 1980.
466. DIN VDE 0838, Rückwirkungen in Stromversorgungsnetzen, die durch Haushaltsgeräte und ähnliche elektrische Einrichtungen versursacht werden, T1 – T3, Beuth Verlag, Berlin 1987.
467. NAMUR-Empfehlung 21, Störfestigkeitsanforderungen, Part 1: EMV von Betriebsmitteln der Prozeß- und Laborleittechnik. NAMUR, Darmstadt 1988.
468. DIN VDE 0843, Elektromagnetische Verträglichkeit von Meß-, Steuer- und Regeleinrichtungen in der industriellen

Prozeßtechnik, T1 – T4, Beuth Verlag, Berlin 1984 – 1988.
469. DIN VDE 0848, T1 – T3: Gefährdung durch elektromagnetische Felder and T4 – T4 A1: Sicherheit bei elektromagnetischen Feldern, Beuth Verlag, Berlin 1982 – 1990.
470. NAMUR-Empfehlung 12, Explosionsschutz in Analysengeräteräumen, NAMUR, Darmstadt 1983.
471. G. Strohrmann: *Anlagensicherung mit Mitteln der MSR-Technik,* Oldenbourg Verlag, München 1983.
472. P. Netter: "Anlagensicherung mit Mitteln der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **35** (1993) no. 1, 45 – 51.
473. NAMUR Empfehlung 24, Anforderungen an Meßeinsätze für Temperatursensoren (Widerstandsthermometer, Thermoelemente) zum Einsatz in eigensicheren Stromkreisen, NAMUR, Darmstadt 1991.
474. F. Pigler: "Blitzschutz für Leittechnische Anlagen," *Automatisierungstechnische Praxis (atp)* **28** (1987) no. 5, 200 – 205.
475. R.-W. Peters, D. Plieninger: "Bayer-interne Arbeiten zu CAE-gestützten Arbeitsmethoden," Leverkusen 1993.
476. G. Heeg: "Objektorientierte Systeme," company brochure, Georg Heeg Verlag, Dortmund 1993.
477. M. Arnold: "Analyse der Kommunikation zwischen Simulationsmodellen technischer Prozesse und den Prozeßnahen Komponenten von Prozeß-leitsystemen," Diplomarbeit, RWTH, Aachen 1993.
478. E. Nicklaus: "Sensortechnik in der Chemischen Produktion," *Mess. Prüf. Auto* (1987) 518 – 522.
479. M. Polke: "Prozeßtechnik," Symposium der RWTH Aachen and Bayer Leverkusen 1988.
480. M. Schweizer, H. J. Warnecke: "Sensoren für die Fertigungstechnik," *48. Physikertagung Münster 1984,* Physik-Verlag, Weinheim 1984, pp. 1331 ff.
481. T. Pfeifer: "Fertigungsmesstechnik und Qualitätssicherung," Tagungsband zum Fachkongreß Microtecnic '92, *VDI Ber.* **1006,** Oct. 1992.
482. H. Reichenberger: "Sensoren für Diagnose und Überwachung in der Medizintechnik," *48. Physikertagung Münster 1984,* Physik-Verlag GmbH, Weinheim 1984, pp. 1357 ff.
483. E. Krestel (ed): "Bildgebende Systeme für die medizinische Diagnostik," Siemens AG, Berlin 1980.
484. W. König (ed.): *Klinische-physiologische Untersuchungsmethoden,* Thieme Verlag, Stuttgart 1972.
485. H. Kresse (ed.): *Kompendium Elektromedizin – Grundlagen, Anwendungen, Geräte,* Siemens AG, Berlin 1982.
486. S. Kutscha: "Statistische Bewertungskriterien für die Entropieanalyse dynamischer Systeme," *Fortschrittsber. VDI Z. Reihe 8,* **178** (1989).
487. S. Kutscha, H. Haubrich, K. Henning: "Lokalisierung von Epilepsieherden durch Einsatz der dynamischen Entropieanalyse," *Automatisierungstech. (at)* **40** (1992) no. 2, 70 – 76.
488. H.-W. Delank: *Neurologie,* Enke Verlag, Stuttgart 1991.
489. K. Poeck: *Neurologie,* Springer Verlag, Berlin 1987.
490. F. D. Stott et al. (eds.): "ISAM-Gent-1981," *Proceedings of the 4th International Symposium on Ambulatory Monitoring and the 2nd Gent Workshop on Blood Pressure Variability,* Academic Press, London 1982.
491. Bundesministerium für Forschung und Technologie: Mikrosystemtechnik 1994 – 1999 – Programm im Rahmen des Zukunftkonzeptes Informationstechnik, Referat 414, Bonn 1994.
492. AMA Fachverband für Sensorik e.V.: Sensor '93 Kongreß, Technologie-Forum Mikrosystemtechnik, Nürnberg, Oct. 1993.
493. P. Egelhaaf, F. Heintz, E. Zabler: "Sensoren im Kraftfahrzeug," *48. Physikertagung Münster 1984,* Physik-Verlag GmbH, Weinheim 1984, pp. 1367 ff.
494. U. Kiencke: "Sensorik im Kraftfahrzeug – vom Sensor zum System," *VDI Ber.* **939** (1992).
495. R. Polke: "Qualitätssicherung in der Feststoffverfahrenstechnik," *Chem. Ing. Tech.* **62** (1990) no. 10, 813 – 821.
496. R. Polke: "The Relevance of Particle Measurement Technology to Process Engineering," *Particle and Particle Systems Characterization* **10** (1993) 86 – 91.
497. R. Polke, N. Scholz, *Part. Charact.* **2** (1985) 39 –40.
498. Ullmann's Encyclopedia of Industrial Chemistry, Vol. B 2, VCH, Weinheim-New York 1988, pp.**1**-1 – **1**-7.
499. H. Rumpf: *Staub Reinhalt. Luft* **27** (1967) 3 – 13.
500. R. Polke: *6. Fachtagung Granulometrie,* Dresden 1989, vol. 1, 39 – 48.
501. K. Leschonski, W. Alex, B. Koglin: *Chem. Ing. Tech.* **46** (1974) 23 – 26; **47** (1975) 97 – 100.

502. G. König, R. Polke, N. Scholz: *Verfahrenstechnik (Mainz)* **3** (1986) 10 – 17.
503. R. Polke, M. Schäfer, N. Scholz: *4. Europ. Symp. Partikelmeßtechnik PARTEC,* Nürnberg 1989, vol. I, pp. 47 – 61.
504. K. Sommer: *Probenahme von Pulvern und körnigen Massengütern: Grundlagen, Verfahren, Geräte,* Springer Verlag, Berlin 1979.
505. Laser Sensor Techn. Inc.: "Lasentec," company brochure, Bellevue, WA 98009.
506. D. Horn, *Angew. Makromol. Chem.* **166/167** (1989) 139 – 153.
507. R. Polke, M. Rädle, N. Scholz: "Partikelmeßtechnik Status, Trend – Bedarf," *Chem. Ing. Tech.* **65** (1993) no. 10, 1191 – 1199.
508. J. Krekel, R. Polke: "Qualitätssicherung bei der Verfahrensentwicklung," *Chem. Ing. Tech.* **64** (1992) no. 6, 528 – 535.
509. H. H. Meyer, W. Noerpel, P. J. Schröder, W. Stieler: "Meßtechnik im Umweltschutz," company report, Bayer AG, Leverkusen, July 1988.
510. M. Arnold, N. Ingendahl, M. Polke: "Prozeßautomatisierung in der Chemischen Industrie," *Innovationsdynamik prozeßanalytischer, automatisierungstechnischer und verfahrenstechnischer Methoden in der chemischen Industrie,* VDI-Hallescher Bezirksverein, Halle 14 – 15 Oct. 1993.
511. G. Heiland: "Sensoren auf Metalloxidbasis," *48. Physikertagung Münster 1984,* Physik-Verlag GmbH, Weinheim 1984, pp. 1393 ff.
512. H. Warncke, M. Kahl, H. H. Meyer, P. Schürmeyer, DE 2 932 436, 1984.
513. D. Kitzelmann: "Elektrochemische Sensoren zur Erfassung von Schadstoffen in Industrie und Umwelt," *48. Physikertagung Münster 1984,* Physik-Verlag GmbH, Weinheim 1984, 1463 ff.
514. R. Hotop: "Neue Methoden der Prozeßanalysetechnik für die verfahrenstechnische Produktion, *Chem. Ing. Tech.* **65** (1993) no. 8, 921 – 928.
515. W. Noerpel: "Abwasserüberwachung mittels on-line Analytik," company report, Bayer AG, Leverkusen 1990.
516. H. Raab, A. Wachter-Büchner: "Kontinuierliche TOC-Meßtechnik in der Abwasserüberwachung," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 3, 128 – 136.
517. DIN 38 409, Summarische Wartungs- und Stoffkenngrößen (Gruppe): Bestimmung des gesamten organisch gebundenen Kohlenstoffs (TOC), Beuth Verlag, Berlin 1983.
518. Endress + Hauser, International Holding AG (eds.): *Abwasserhandbuch,* Reinach 1992.
519. G. Fikentscher: *Cellulosechemie* **13** (1932) 58.
520. M. Hoffmann, H. Krömer, R. Kuhn: *Polymeranalytik I,* Thieme Verlag, Stuttgart 1977.
521. H.-G. Elias: *Makromoleküle,* Verlag Hüthig & Wepf, Heidelberg 1975.
522. R. Hotop, J. König: "On-line Analytik in Polymerprozessen: Derzeitige Situation und Zukunftsaufgaben," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 12, 610 – 616.
523. W. Günther: "Die HPLC im praktischen Vergleich," *Messen, Regeln, Automatisieren 1993,* Verlag Hoppenstedt & Co., Berlin 1993, pp. 17 – 24.
524. B. von Falkai (ed.): *Synthesefasern,* Verlag Chemie, Weinheim 1981.
525. B. Adler: "Erkennen von Prozeßtrends durch Methoden der Fuzzy Clusterung und Fuzzy Logic," lecture at VDI Tagung *Innovationsdynamik prozeßanalytischer, automatisierungstechnischer und verfahrenstechnischer Methoden in der chemischen Industrie,* Preprints GVC-VDI, Düsseldorf, Oct. 1993.
526. M. Winterstein, G. Brückner, B. Adler: "Einsatz eines multivarianten Sensorsystems zur Abwasser-überwachung," lecture held at the GVC-VDI Tagung *Innovationsdynamik prozeßanalytischer, automatisierungstechnischer und verfahrenstechnischer Methoden in der chemischen Industrie,* Preprints GVC-VDI, Düsseldorf, Oct. 1993.
527. M. Polke: "Trends der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 9, 408 – 415.
528. P. Hauptmann, R. Lucklum: "Sensoren für die Prozeßtechnik," lecture held at VDI Tagung *Innovationsdynamik prozeßanalytischer, automatisierungstechnischer und verfahrenstechnischer Methoden in der chemischen Industrie,* Preprints GVC-VDI, Düsseldorf, Oct. 1993.
529. R.-W. Peters: "Informationsstrukturen für das Gebiet der Prozeßleittechnik und ihre Präsentation in einem Hypermediasystem", Dissertation, RWTH, Aachen 1994.

530. W. Göpel, J. Hesse, J. N. Zemel: *Sensors,* vols. 1 – 8, VCH Verlagsgesellschaft, Weinheim 1989.
531. G. Tschulena: "Von Mikrosensoren zu Mikrosystemen – Stand und Perspektiven," *TM Tech. Mess.* **60** (1993) no. 9, 313 – 318.
532. AMA: Sensor 91 – Kongreßband Marketing, Fachverband für Sensorik e.V., Nürnberg 1991.
533. ISA Directory 1993: Comprehensive Guide for Measurement and Control Market, Instrument Society of America, North Carolina, 1993.
534. *Messen, Regeln, Automatisieren; Einkaufsführer der Meß-, Regelungs- und Automatisierungstechnik,* Verlag Hoppenstedt & Co., Darmstadt 1993.
535. *6. Internationale Fachmesse mit Kongreß für Sensorik & Systemtechnik,* Nürnberg, Oct. 1993.
536. *MessComp '93-7,* Kongreßmesse für industrielle Meßtechnik, Wiesbaden, Sept. 1993.
537. Prognos AG: *Sensortechnik 2000,* Prognos Weltreport, vol. 1, Basel, Sept. 1988.
538. DMCW Associates: "U.S. Process Control Marketplace, Opportunities and Trends in the 1990's," final report, ref. no. MC-10, Menlo Park, CA 1991.
539. K. Günzel: "Trends in der Prozeßmeßtechnik," *Tech. Mess.* **59** (1992) no. 9, 333 – 339.
540. A. Brucker, H. Umbach: "Tendenzen bei der Messung von Durchfluß," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 3, 123 – 127.
541. Batelle-Institut: *Sensors, Miniturisation and Integration,* Batelle-Institut e.V., Frankfurt 1988.
542. G. Mark, R. Heidel, K. P. Frehse: "Wirtschaftlichkeit des Feldbuseinsatzes: Analyse des Installationsaufwandes unterschiedlicher Feldkommunikationssysteme," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 9, 507 – 512.
543. Bayer AG: *Kabel- und Rohrwege,* company report, Bayer AG, Leverkusen, Oct. 1987.
544. H. Janocha (ed.): *Aktoren,* Springer Verlag, Berlin 1992.
545. H. Drahten: "Der Feldbus aus Sicht des Anwenders," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 2, 55 – 64.
546. G. Pinkowski: "Instrumentierungstechnik im Wandel. Von der Smart-Technik zum herstellerunabhängigen Bedienkonzept," *Automatisierungstechnische Praxis (atp)* **6** (1992) no. 5, 247 – 254.
547. H. Junginger, H. Wehlan: "Der Feldmultiplexer aus Anwendersicht," *Automatisierungstechnische Praxis (atp)* **31** (1989) 12.
548. A. Mangold: ACHEMA '91: "Stellgeräte für verfahrenstechnische Anlagen," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 11, 585 – 593.
549. DIN 31 051: Instandhaltung, Begriffe und Maß-nahmen, Beuth Verlag GmbH, Berlin 1985.
550. M. Björkmann: Elektrische Antriebstechnik in der Chemischen Industrie, NAMUR Hauptversammlung 1991, Lahnstein.
551. B. Koenig, O. Ohligschläger: "Ein intelligenter elektropneumatischer Stellungsregler," *Automatisierungstechnische Praxis (atp)* **31** (1989) 8.
552. M. Pandit, J. König, H. Hoffman: "Ein intelligenter elektropneumatischer Stellungsregler," *Automatisierungstechnische Praxis (atp)* **31** (1989) 367 –373.
553. DIN 19 222: Leittechnik, Begriffe, March 1985.
554. M. Polke: "On-line Betriebsmeßtechnik," lecture at the *GVC Jahrestreffen,* Wien 1992.
555. P. Kruck: "Methodische Software-Entwicklung (dargestellt am Beispiel eines CAD-Systems zur Symbolischen Konstruktion), Dissertation, Bergische Universität GH Wuppertal 1987.
556. H.-J. Dorn: "Drehzahlveränderliche Pumpenantriebe: Eine Alternative zu Stellventilen?" *Automatisierungstechnische Praxis (atp)* **29** (1987) no. 6.
557. H. Linnenbrink: "Drehzahlveränderbare Antriebe, ein neuer Arbeitskreis im Arbeitsfeld," *3. NAMUR Hauptbericht, Hauptsitzung 53,* Lahnstein 1991.
558. Antriebstechnik von AEG, Fachbereich Antriebssysteme und Leistungselektronik A 95, Gleichstrom-, Drehstrom- und Servoantriebe, Kurzdokumentation, Bestell-Nr. 029.129 951, 1991.
559. Bayer-interner Arbeitskreis PLT-Festlegungen: Einführung von PLT-Festlegungen nach Strukturierung auf Basis von E/R-Diagrammen, Leverkusen 1990.
560. Canon, various company brochures, EOS Cameras, 1989.
561. Veranstaltungsprogramm zu Internationale Konferenz zu neuen Aktoren – Actuator 1994, Bremen.
562. V. Wagner et al., Bayer Internal Report, Leverkusen 1990.

563. K. Hartmann, W. Schepers: Drehzahlgeregelte Pumpenanlagen, Bayer internal lecture, Leverkusen, Oct. 12, 1988.
564. M. Polke: "Trends in der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 9, 408 – 415.
565. N. Haby: "Drehzahlveränderbare Pumpen als Stellglieder im Regelkreisen," *Regelungstechnische Praxis (atp)* **22** (1986) 6.
566. H. Linnenbrink: "Drehzahlveränderbare Antriebe. Ein neuer Arbeitskreis im Arbeitsfeld," *3. NAMUR-Hauptbericht, Hauptsitzung 53,* 1990.
567. M. Polke: "Trend zur Automatisierung in der Chemischen Industrie hält ungebrochen an: Gedanken zur ACHEMA 1991," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 5, 219.
568. M. Harmoinen: "Säädetyt AC – käytöt tulevaisuudessa," (Variable speed AC drives in the future), ABB Strömberg Oy, Power Electronics, Vaasa, Finland 1991.
569. Sanyo Electric, Semiconductor Division, Sanyo MST Hybrid System, Japan, Oct. 1989.
570. W. Mutz: "Prozeßnahe Stromversorgung," *Automatisierungstechnische Praxis (atp)* Sonderheft NAMUR Statusbericht 87, 1987.
571. W. Mutz: "Niederspannungstechnik für die Betriebe," EN-PLT-Kolloquium "Prozeßelektrotechnik," Bayer AG, Leverkusen 1987.
572. M. Freytag: "Sichere Stromversorgung für PLT-Einrichtungen," PLT-Kolloquium "Erfahrungen mit dem Einsatz moderner Prozeßleittechnik," Bayer AG, Leverkusen 1984.
573. K. Kling: Stand und Entwicklung der Errichtungsbestimmungen VDE 0100.etz b.H.6, pp. 128 – 131, 1974.
574. Bayer AG: "Beleuchtungsanlagen," PLT-Festlegungen 06.53.01, Leverkusen 1984.
575. Bayer AG: "Die Bedeutung des Sicherheitsschalters bei Arbeiten in Gefahrenbereichen technischer Arbeitsmittel," Arbeitsschutzfestlegung LE2, Leverkusen 1984.
576. G. U. Spohr: "Prozeßleitsysteme," in: *Prozeßleittechnik,* 1st ed., Oldenbourg Verlag, München 1992, Chap. 5.2.
577. M. Polke: "Prozeßleittechnik," Skriptum zur Vorlesung WS84/85 Universität Stuttgart, Ausgabe 1984 und folgende.
578. U. Epple: "Prozeßführung," in: *Prozeßleittechnik,* 1st ed., Oldenbourg Verlag, München 1992, Chap. 4.3.
579. M. Brombacher, M. Polke: "Perspektiven der Prozeßleittechnik," *Chem. Ing. Tech.* **60** (1988) no. 7, 539 – 546.
580. K. F. Früh: "Optimalstruktur dezentraler Prozeß-leitsysteme," *RTP Regelungstech. Prax.* **21** (1979) 273 ff.
581. W. Körner, P. Moll, M. Polke: "Zur funktionalen Gliederung von Leitsystemen," *Automatisierungstechnische Praxis (atp)* **29** (1987) no. 9, 393 –400.
582. U. Epple: "Funktionsorientierte Strukturierung prozeßleittechnischer Aufgaben," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 10, 468 –474.
583. L. Litz: "Anforderungen an die Prozeßleitsysteme der nächsten Generation," *53. NAMUR Hauptsitzung,* Lahnstein 1990.
584. ISO: MMS, Part 1: Service Definitions, MMS Part 2: Protocol Specification, 1988.
585. ISO: ISO/DIS 74 498 Open Systems Interconnection Basic Reference Model, May 1983.
586. M. Arnold: "Analyse der Kommunikation zwischen Simulationsmodellen technischer Prozesse und den prozeßnahen Komponenten von Prozeßleitsystemen," Diplomarbeit, Lehrstuhl für Prozeßleittechnik, RWTH Aachen, 1993.
587. M. Polke: "Information als kritische Ressource," in H. A. Henzler (ed.): *Handbuch strategische Führung,* Gabler, Wiesbaden 1988.
588. NAMUR Empfehlung 21: Elektromagnetische Verträglichkeit von Betriebsmitteln der Prozeß- und Laborleittechnik, NAMUR Geschäftsstelle, Leverkusen Dec. 1990.
589. DIN VDE 848 T1 – T3: Gefährdung durch elektromagnetische Felder, T4: Sicherheit bei elektromagnetischen Feldern, Beuth Verlag, Berlin 1982.
590. DIN VDE 843 T1 – T4: Elektromagnetische Verträglichkeit von Meß-, Steuer- und Regeleinrichtungen in der industriellen Prozeßtechnik, Beuth Verlag, Berlin 1987.
591. J. König: "Datenkommunikation in der Leittechnik," *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 5, 213 – 214.
592. M. Freytag: "Datenkommunikation in der Leittechnik – Status und Trend," *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 5, 215 – 222.
593. M. Polke: "Informationshaushalt technischer Systeme," in Prozeßrechner 1984, Prozeßdatenverarbeitung im Wandel 4,

Prozeßrechnertagung 1984, Karlsruhe, Springer Verlag, Berlin-Heidelberg 1984, pp. 62 – 70.
594. M. Syrbe: "Informationstechnik, Chance und Verantwortung," *Vortrag Produktionstechnisches Kolloquium,* Berlin 1986, Fraunhofer-Gesellschaft, München.
595. H. Steusloff: "Kommunikation in technischen Systemen," *GI/NTG Tagung Kommunikation in verteilten Systemen,* Karlsruhe, 11. – 15.3. 1985, *Informatik-Fachberichte* **111,** Springer Verlag, Berlin 1985, pp. 33 – 55.
596. R. W. Peters: "Funktionale Gliederung der Informationssysteme für Produktion, Labor und Infrastruktur; Funktionen der Produktionsleitebene," *48. NAMUR Hauptsitzung,* Lahnstein 1985; Berichtsband pp. 210 – 232.
597. D. Powell et al.: "The Delta-4 Approach to Dependability in Open Distributed Computing Systems," *FTCS 18,* Tokio 1988.
598. ISO: Open Systems Interconnection Basic Reference Model, ISO/DIS 7498, May 1983.
599. O. Gihr, E.-H. Göldner, P. J. Kühn, K. Sauer: "Lokale Netze und ISDN-Nebenstellenanlagen, Stand und Entwicklungstendenzen," *PIK (Praxis der Informationsverarbeitung und Kommunikation),* **3,** Hanser Verlag, München 1986, pp. 26 –34.
600. General Motors (GM): "Manufacturing Automation Protocol (MAP) Document Version 2," GM EMCC-APMES A/MD-39, General Motors Technical Center, Warren 1984, MI 48090-9040.
601. ESPRIT Project 5104 (CNMA): CNMA Implementation Guide, Revision 5.1, Volume 1 – 3. University of Porto (ed.), Ref. 5104.92.35/F1.PD, 1992.
602. H. Steusloff: "Technische Kommunikation in der industriellen Automation. INFINA '89, Informatik für die industrielle Automation," *VDI-Berichte* **723,** VDI-Verlag, 1989, pp. 319 – 338.
603. M. Polke: "Trends der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **31** (1989) 408 – 415.
604. Factory Instrumentation Protocol C46-602 Application Layer (MPS), C 46-603 Data Link Layer, C46-604, Twisted Pair Physical Layer. Französische Normungsorganisation AFNOR, Paris 1988.
605. DIN 19 245: Messen, Steuern, Regeln, PROFIBUS, Process Field Bus. Beuth Verlag, Berlin.
606. D. Janetzky, K. Watson: Performance Evaluation of the MAP Token Bus in Real Time Applications. Advances in Local Area Networks (eds.: K. Kümmerle, F. Tobagi, J. Limb). IEEE, New York 1987, pp. 411 – 425.
607. Fieldbus; ESPRIT-CIM-Report: CNMA, Phase 4, Addendum 2, Chapter 4, in Vorbereitung; Kommission der Europäischen Gemeinschaften, DG XIII, 200 Rue de la Loi, Brüssel, Belgien.
608. J. A. H. Pfleger: "Kommunikationssystem Feldbus," *Automatisierungstechnische Praxis (atp)* **28** (1986) 223 – 227.
609. H. Drathen: "Neuorientierung der leittechnischen Normung auf nationaler und internationaler Ebene," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 11, 580 – 585.
610. W. Büsing: "MAP für die Prozeßleittechnik: Die Herausforderung bleibt," *NAMUR Statusbericht '87, Prozeßleittechnik in der Chemischen Industrie,* R. Oldenbourg Verlag, München 1987, pp. 101 –106.
611. G. Scheidenberger: "Fieldbus Message Specification (FMS)". *Automatisierungstechnische Praxis (atp)* **31** (1989) 152.
612. D. Heger: "Konformitätsprüfung offener Kommunikationsprotokolle für die Automatisierungstechnik. Fachbericht Messen, Steuern, Regeln. Fortschritte in der Meß- und Automatisierungstechnik durch Informationstechnik," vol. 14, *INTERKAMA-Kongreß,* Oldenbourg Verlag, München 1986.
613. ISO: "Specification of Abstract Syntax Notation One (ASN.1)," IS 8824, May 1987; "Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.1)," IS 8825, May 1987.
614. W. Gora, R. Speyerer: *ASN.1, Abstract Syntax Notation One,* DATACOM Buchverlag, Pulheim 1987.
615. M. Syrbe: "Basic Principles of Advanced Process Control System Structures and a Realization with Distributed Microcomputers," *IFAC World Congress,* June 1978, Helsinki, Pergamon Press, New York 1979, pp. 393 – 407.
616. M. Siefert: "Entwurf und Implementierung eines ASN. 1 – Compliers," Diplomarbeit, Fakultät für Informatik, Universität Karlsruhe/FrauenhoferInstitut für Informations- und Datenverarbeitung, Karlsruhe 1988.

617. D. T. Ross, K. E. Schomann, Jr.: "Structured Analysis for Requirements Definition," *IEEE Transactions on Software Engineering* **SE-3** (1977) no. 1, 6–15.
618. T. De Marco: *Structured Analysis and System Specification,* Yourdon Press, New York 1978.
619. M. Polke: "Prozeßleittechnik; Trend zur Automatisierung in der Chemischen Industrie hält ungebrochen an," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 5, 219–222.
620. mbp: Produktbeschreibung des CAE-Systems HW-PLAN, Dortmund 1990.
621. G.-U. Spohr: "CAE für die Prozeßleittechnik," *51. NAMUR-Hauptsitzung,* Lahnstein, Dec. 1988.
622. G.-U. Spohr, R. Schrieber: "Anforderungen an CAE-Systeme," *INTERKAMA-Kongreß,* Düsseldorf, Oct. 1989, in: G. Schmidt, H. Steusloff (eds.): Mit vernetzten intelligenten Komponenten zu leistungsfähigeren Meß- und Automatisierungssystemen, Oldenbourg Verlag, München 1989, pp. 575–584.
623. G.-U. Spohr: "CAE in der Prozeßleittechnik," Vorlesungsmanuskript, RWTH Aachen, Lehrstuhl für Prozeßleittechnik, 1991.
624. HOAI: Rechtsverordnung zum Gesetz zur Regelung von Ingenieur- und Architektenleistungen vom 04.11.1971.
625. M. Polke: "Einführung in die Prozeßleittechnik," Manuskript zur Vorlesung, Universität Stuttgart 1984.
626. DIN 69 901, Projektwirtschaft; Projektmanagement; Begriffe, Beuth Verlag, Berlin, August 1987.
627. E. Cassirer: *Substanzbegriff und Funktionsbegriff,* 6th ed., Wissenschaftliche Buchgesellschaft, Darmstadt 1990.
628. W. F. Daenzer: *Systems Engineering,* Verlag Industrielle Organisation, Zürich 1988.
629. D. Reschke, A. Schelle, K. Schnopp: *Handbuch "Projektmanagement,"* Verlag TÜV Rheinland, Köln 1989.
630. H. H. Schulze, A. Schembra: "Schätzung der Inve-stitionskosten und Produktionskosten bei der Verfahrensentwicklung und Anlagenplanung," lecture held at the *GVC-Tagung* über Entwicklung und Auslegung verfahrenstechnischer Prozesse, Baden-Baden, 12/13 June 1989.
631. E. Willems, H. Hennecke: "Planung, Montage, Inbetriebnahme und Ingenieurbetreuung von Einrichtungen zum Messen, Steuern, Regeln," in J. Hengstenberg, K. H. Schmitt, B. Sturm, D. Winkler (eds.): *Messen, Steuern, Regeln in der chemischen Technik,* 3rd ed., vol. 5, Springer Verlag, München 1985.
632. H. Groh, R. Gutsch: *Netzplantechnik,* VDI-Verlag, Düsseldorf 1982.
633. S. Dworatschek et al.: "Project Management Software," *The International GPM/INTERNET-Symposium,* Garmisch, 15–18 June 1986.
634. A. Tampier: "Dezentrale Einzelprojektplanung und zentrale Multiprojektplanung im automatischen Verbund VAX-PC-VAX mit IST-Kosten-Übernahme," in D. Reschke, A. Schelle (eds.): *Beiträge zum Projektmanagement-Forum 1989,* Tagung Garmisch, 4–6 Oct. 1989.
635. MS-Project, software product von Micro-Soft.
636. M. Brombacher: "Das Lastenheft als Grundlage der Automatisierung chemischer Verfahren und seine Darstellung als Expertensystem," Dissertation, Technische Universität, München 1985.
637. M. Polke: "Informationshaushalt technischer Prozesse," *Automatisierungstechnische Praxis (atp)* **27** (1985) 161–171.
638. M. Polke: "Information als kritische Ressource," in H. A. Henzler (ed.): *Handbuch "Strategische Führung,"* Gabler-Verlag, Wiesbaden 1988.
639. T H. Müller-Heinzerling: "Flexible Standardsoftware zur Automatisierung von Chargenprozessen mit Prozeßleitsystem TELEPERM M," *Automatisierungstechnische Praxis,* **30** (1988) no. 6.
640. R. Schmidt, Diverse Bayer-interne Mitteilungen, Leverkusen 1989–1990.
641. M. Niese, EP 0 104 411, 1986.
642. N. Becker: Diverse Bayer-interne Mitteilungen, Leverkusen 1989–1990.
643. *Ullmann,* 4th ed., **4,** 70.
644. G. Bernecker: *Planung und Bau verfahrenstechnischer Anlagen,* VDI-Verlag Düsseldorf 1984.
645. DIN 28 004, Fließbilder verfahrenstechnischer Anlagen; Begriffe, Fließbildarten, Informationsinhalt, part 1, Beuth Verlag, Berlin, May 1988.
646. M. Polke: "Prozeßleittechnik in der Produktion; Voraussetzung für Diversifikation von Chemiefasern," *Lenzinger Ber.* **67** (1989) 12–23.
647. DIN 19 227, Bildzeichen und Kennbuchstaben für Messen, Steuern, Regeln in der Verfahrenstechnik; Zeichen für die funktionelle Darstellung, part 1, Beuth Verlag, Berlin, Sept. 1973.

648. W. Ahrens, M. Polke: "Informationsstrukturen in der Automatisierungstechnik," lecture held at Interkama-Kongreß 1989, in G. Schmidt, H. Steusloff (eds.): *Mit vernetzten, intelligenten Komponenten zu leistungsfähigen Meß- und Automatisierungssystemen,* R. Oldenbourg Verlag, München 1989, pp. 47 – 68.
649. E. Welfonder: "Leittechnik – Dokumentation aus Betreibersicht," *ETG/VGB-Fachtagung Betriebsgerechte Dokumentation in Kraftwerken, Einfluß der modernen Leittechnik und Planungsmittel,* Baden-Baden 1986.
650. DIN 32 705, Klassifikationssysteme; Erstellung und Weiterentwicklung von Klassifikationssystemen, Beuth Verlag, Berlin 1987.
651. R. Kessler: "Prozeßorientierte Strukturierung von verfahrenstechnischen Anlagen," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 10, 461 –467.
652. DIN 66 201, Prozeßrechensysteme; Begriffe, part 1, Beuth Verlag, Berlin, May 1981.
653. J. König: "Sicherheitstechnische Aspekte der Prozeßleittechnik," *Chem.-Ing. Tech.* **49** (1987) no. 3, 196 – 204.
654. R. Magin, W. Wüchner: *Digitale Prozeßleittechnik,* Vogel Buchverlag, Würzburg 1987.
655. M. Freytag: "USV und Notstromversorgung," *VDI-Tagung,* Fellbach/Stuttgart, 25. – 26.10.1989.
656. R. Peters: Diverse Bayer-interne Mitteilungen, Leverkusen 1985.
657. E. Ignatowitz: *Chemietechnik,* Verlag Europa-Lehrmittel, Wuppertal 1982, p. 112.
658. DIN/IEC 3B (Sec)40, Ausführung von Funktionsplänen für Steuerungen, Identisch mit IEC 3B (Sec)40, Oct. 1985.
659. DIN 40 719, Kennzeichnungsregeln zur funktions- und ortsbezogenen Kennzeichnung von Systemen und Anlagen, Einrichtungen und Betriebsmitteln, Anschlüssen, Signalen und technischer Dokumentation, Teil 6, Entwurf, Beuth Verlag, March 1989.
660. D. Plieninger: "Qualitätssicherung in der leittechnischen Planung," NAMUR-Hauptsitzung, Lahnstein 1991.
661. K. H. Schmitt: Qualitätssicherung in der Leittechnik, NAMUR-Hauptsitzung, Lahnstein 1991.
662. W. Masing: *Qualitätslehre, DGQ,* 5th ed., Beuth Verlag, Berlin 1977.
663. A. Müller: Qualitätsprüfung von PLT-Dokumentationen als Vertragsbestandteil bei der Vergabe von Ingenieurleistungen an Fremdfirmen, Bayer-interne Festlegung, Leverkusen 1992.
664. D. Plieninger, A. Müller: "Ansätze zur Qualitätssicherung in der leittechnischen Planung," *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 10, 564 ff.
665. H. Rinne, H.-J. Mittag: *Statistische Methoden der Qualitätssicherung,* 2nd ed., Hanser Verlag, München 1991.
666. G. Färber, M. Polke, H. Steusloff: "Mensch-Prozeß-Kommunikation," *Chem. Ing. Tech.* **57** (1985) no. 4, 307 – 317.
667. Bayer AG, Beleuchtungsanlagen, PLT-Festlegungen, 06.53.01, Leverkusen 1984.
668. Bayer AG: Prozeßleitwarten, Leitstände, Nebenräume. Richtlinien für die konstruktive Gestaltung, PLT-Festlegungen, Leverkusen, 1989.
669. R. Beuchelt, U. Fierus, H. Hennecke, E. Willems: PLT-Raumdesign, interne Richtlinie der Bayer AG Leverkusen 1985.
670. VDI/VDE 3546: Konstruktive Gestaltung von Prozeßleitwarten. Blatt 1 (1987): Allgemeiner Teil; Blatt 2 (1981): Bautechnische Maßnahmen; Blatt 3 (1988): Ausführung des Leitstandes; Blatt 4 (1986): Nebenräume; Blatt 5 (1991): Anordnung von Monitoren, Beuth Verlag, Berlin.
671. NAMUR Empfehlung NE 12: "Explosionsschutz von Analysegeräteräumen". NAMUR Geschäftsstelle Leverkusen, Bayerwerk, Leverkusen 1983.
672. DIN 33 402: Körpermaße des Menschen. Beuth Verlag, Berlin 1986.
673. DIN 33 414: Ergonomische Gestaltung von Warten. Beuth Verlag, Berlin 1988.
674. DIN 33 403: Klima am Arbeitsplatz und in der Arbeitsumgebung, Beuth Verlag, Berlin 1988.
675. DIN 5035: Beleuchtung mit künstlichem Licht, Teil 2: Richtwerte für Arbeitsstätten in Innenräumen und in Freien, Beuth Verlag. Berlin, 1990.
676. Verordnung über Arbeitsstätten. Bundesministerium für Arbeit und Sozialordnung, Referat Öffentlichkeitsarbeit, Bonn 1975.
677. J. König: "Sicherheitstechnische Aspekte der Prozeßleittechnik," *Chem. Ing. Tech.* **59** (1987) no. 3 216 – 218.
678. R. Bachmann, J. König: "Towards Increased Safety – Some Future Aspects of Modern Process Control for Chemical Plants," in *Automatic Control World Congress 1987,* Pergamon Press, Oxford.

679. R. Isermann (ed.): *IFAC-Symposium: Safeprocess,* Baden-Baden 1991.
680. E. D. Gilles: "Von der MSR-Technik zur wissensbasierten Prozeßleittechnik," *Chem. Anlagen + Verfahren* **4** (1985) 62 – 64.
681. M. Polke: "Erforderliche Informationsstrukturen für die Qualitätssicherung," *Chem. Ing. Tech.* **65** (1993) no. 7, 791 – 796.
682. G. Färber, M. Polke, H. Steusloff: "Mensch-Prozeß-Kommunikation," *Chem. Ing. Tech.* **57** (1985) no. 4, 307 – 317.
683. R. Perne, M. Polke: "Produktentwicklung durch Verfahrensvariation," *Regelungstechnik* **30** (1982) no. 5, 147 – 156.
684. E. D. Gilles, M. Wurst: "Prozeßführung zur Sicherung reproduzierbarer Qualität in der verfahrenstechnischen Produktion," lecture at *GVC-Jahrestreffen,* Stuttgart 1990, pp. 138 – 139.
685. J. Pawlowski: *Die Ähnlichkeitstheorie in der physikalisch-technischen Forschung,* Springer Verlag, Berlin 1971.
686. J. Pawlowski: "Ähnlichkeitstheorie – ein wirksames Instrument der Verfahrenstechnik," *GVC-Tagung Ähnlichkeitstheoretische Methoden und ihr verstärkter Einsatz in der Praxis,* Mainz, June 15, 1984.
687. M. Zeitz: "Simulationstechnik," *Jahrestreffen der Verfahrensingenieure,* Straßburg 1986, Oldenbourg Verlag, München 1987, pp. 164 – 169.
688. H. Schuler: Methoden zur Prozeßführung mit Simulationsmodellen, *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 10, 475 – 481;no. 11, 519 – 523.
689. W. Ahrens, M. Polke: "Netzmodelle als systemtechnische Informationsbasis für die Prozeßleittechnik," *Automatisierungstechnik (at)* **37** (1989) no. 3, 94 – 103.
690. E. D. Gilles: "Auf dem Wege zu einer modellgestützten Prozeßleittechnik," *50. Jahrestagung NAMUR,* Lahnstein 1987, Oldenbourg Verlag, München 1987, pp. 35 – 45.
691. E. D. Gilles, W. Marquardt: "Dynamische Anlagensimulation in der chemischen Verfahrenstechnik," *Automatisierungstechnische Praxis (atp)* **34** (1986) no. 2, 49 – 58.
692. W. Ahrens: "Einsatz von Expertensystemen in der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **29** (1987) no. 10, 475 – 485.
693. M. Brombacher: "Expertensysteme zur Unterstützung der Planung (Lastenheft)," *Automatisierungstechnische Praxis (atp)* **30** (1988) no. 6, 280 –291.
694. W. Marquardt: "Rechnergestützte Erstellung verfahrenstechnischer Prozeßmodelle," *Chem. Ing. Tech.* **64** (1992) no. 1, 25 – 40.
695. H. Fittler: "Anwendung von Expertensysteme in der Prozeßautomation," *Automatisierungstechnische Praxis (atp)* **33** (1991) no. 4, 199 – 205.
696. E. D. Gilles et al.: "Ein Trainingssimulator zur Ausbildung von Betriebspersonal in der chemischen Industrie," *Automatisierungstechnische Praxis (atp)* **32** (1990) no. 7, 343 – 350.
697. J. Laube, M. Polke: "Das Phasenmodell der Produktion: Ein semantisches CAE-Werkzeug zur Prozessbeschreibung," Lecture at *Regelungstechnische Kolloquim,* Boppard, March 1993.
698. A. Lippert et al.: "Gedanken zur Neugestaltung der Verfahrenstechnik," company report, Bayer, Leverkusen, Nov. 1988.
699. C. A. Petri: "Kommunikation mit Automaten," Dissertation, Technische Hochschule Darmstadt 1961.
700. W. Reisig: *Systementwurf mit Netzen,* Springer Verlag, Berlin 1985.
701. G. Scheschong: "Eine auf Petrinetzen basierende Konstruktionsanalyse und (Teile-)verifikationsmethode zur Modellierungsunterstützung bei Entwicklung von Informationssystemen," Dissertation, TU Berlin 1984.
702. M. Brombacher, M. Polke: "Perspektiven der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **29** (1987) no. 11, 503 – 509;no. 12, 558 – 561.
703. U. Epple: "Funktionsorientierte Strukturierung Prozeßleittechnischer Aufgaben," *Automatisierungstechnische Praxis (atp)* **31** (1989) no. 10, 468 –474.
704. O. Graf: "Standartisierte Grundfunktionen verringern Softwareaufwand bei Teleperm M.," *eng & auto* **3** (1991) 2 – 5.
705. Bayer AG, EPO 104 411, 1986 (M. Niese).
706. W. Wüchner: "Anforderungen an das Personal beim Einsatz moderner Prozeßleitsysteme, *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 1 13 – 18.
707. G. Krauss: "Akzeptanz der Prozeßleitsysteme durch die Betriebe," *Automatisierungstechnische Praxis (atp)* **10** (1985) no. 10 470 – 476.

708. H. Raab: "Die Rolle der Sensoren in der Verfahrenstechnik, und: Digitale Sensoren und Sensorsysteme: Welche Informationen sollen sie liefern," *Chem. Ing. Tech.* **61** (1989) no. 3, 41 – 47.
709. Verein Deutscher Ingenieure: "Der Ingenieur in Beruf und Gesellschaft," Handlungsempfehlung: Sozialvertragliche Gestaltung von Automatisierungsvorhaben, Düsseldorf, Dec. 1989.
710. L. Bainbridge: "Ironies of Automation," *Automatica* **19** (1983) no. 6, 775 – 559.
711. H. Grauel, H. Bauer, P. Metzing: "Erhöhung der Kompetenz des Betreiberpersonals – Ziel der zustandsorientierten Mensch-Prozeß-Kommunikation in Chemieanlagen," *MSR (Berlin)* **30** (1987) no. 10, 448 – 455.
712. H. Lenk: "Können Informationssysteme moralisch verantwortlich sein?" *Informatik Spektrum* **12** (1989) 248 – 255.
713. H. J. Charwart: *Lexikon der Mensch – Maschine – Kommunikation,* Oldenbourg Verlag, München 1992.
714. B. Geiser: "Was Window-Manager in der Leittechnik leisten sollen," *Elektronik* **16** (1991) 52 – 59.
715. Siemens AG: "Teleperm M, company brochure, no. A 19100 E816-B677 100 045 PA 10895, 1991.
716. AEG: Schirmbilder für Viewstar 750, Frankfurt.
717. ABB Prozeßautomation GmbH: Schirmbilder für Advant Station.
718. K. R. Boff, J. E. Lincoln: *Engineering Data Compendium – Human Perception and Performance,* Harry G. Armstrong Aerospace Medical Research Lab., Ohio 1988.
719. H. Schmidtke: *Handbuch der Ergonomie,* 2nd ed., Carl Hanser Verlag, München 1989.
720. N. Heineken: "Aufnahme und Verarbeitung von Information," to be published in H. Schmidtke: *Handbuch der Ergonomie,* 2nd ed., Carl Hanser Verlag, München 1989.
721. H. Steusloff: "Mensch-Prozeß-Kommunikation," *DPG-Arbeitstagung,* Bad Honnef, Dec. 1983, pp. 1 – 23.
722. R. W. Bailey: *Human Performance Engineering, A Guide for Systems Designers,* Bell Telephone Laboratories, Inc., Prentice-Hall, Englewood Cliffs, N.J. 1982.
723. I. Rasmussen: "The Role of Hierarchical Knowledge Representation in Decision Making in System Management," IEEE Transations on Systems, Man and Cybernetics, vol., SMC -15(1985) 2, pp. 234 – 243.
724. G. Johannsen: "Towards a New Quality of Automation in Complex Man-Machine Systems," *Automatica* **28** (1992) no. 2, 355 – 373.
725. M. Eigen: "Mozart – oder unsere Verlegenheit im Umgang mit dem Genie," in *Perspektiven der Wissenschaft,* DVA, Stuttgart 1988, pp. 76 – 104.
726. J. C. Eccles: *Das Gehirn des Menschen,* 6th ed., R. Piper, München 1975.
727. H. and K. Benesch: *dtv-Atlas zur Psychologie,* dtv, München 1987, pp. 90 – 123.
728. W. Metzger (ed.): "1. Wahrnehmung und Bewußtsein," *Handbuch der Psychologie,* vol. I, Hogrete, Göttingen 1966.
729. N. A. Streitz, P. A. Gloor (eds.): "Hypertext und Hypermedia: Von theoretischen Konzepten zur praktischen Anwendung," *Informatik-Fachberichte* **249,** Springer Verlag, Heidelberg 1990.
730. A. Despopoulos, S. Silbernagel: *Taschenatlas der Physiologie,* Georg Thieme Verlag, Stuttgart 1988.
731. R. F. Schmidt, G. Thews: *Physiologie des Menschen,* Springer Verlag, Berlin 1983.
732. A. Hajos: *Einführung in die Wahrnehmungspsychologie,* Wissenschaftliche Buchgesellschaft, Darmstadt 1980.
733. W. D. Heiss: "Kernspintomographie und Positronen-Emissions-Tomographie zur Lokalisation menschlicher Hirnfunktion," *Ver. Ges. Dtsch. Naturforsch. Ärzte* **114** (1986) 241 – 269.
734. G. Deutsch, S. P. Springer: *Linkes und rechtes Gehirn: Funktionale Asymmetrien,* 2nd ed., Spektrum der Wissenschaft – Verlagsgesellschaft, Heidelberg 1988.
735. H. Ostertag: "Positronen-Emissions-Tomographie (PET)," *Phys. Bl.* **48** (1992) no. 2, 77.
736. K. Poeck (ed.): *Klinische Neuropsychologie,* Georg Thieme Verlag, Stuttgart 1989.
737. R. J. Trotter: "Das Jahrzehnt des Gehirns," *Psychologie Heute,* Jan. 1992, 57 – 59.
738. E. Straus: *Vom Sinn der Sinne,* Springer Verlag, Berlin 1978.
739. J. J. Gibson: *Wahrnehmung und Umwelt,* Urban & Schwarzenberg, München 1982.
740. K. Lorenz: *Die Rückseite des Spiegels. Versuch einer Naturgeschichte menschlichen Erkennens,* dtv, München 1977.

741. VDI/VDE Gesellschaft Meß- und Automatisierungstechnik (GMA), VDI/VDE Richtlinie 3536, Düsseldorf, Jan. 1990.
742. D. O'Shaughnessy: *Speech Communication,* Addison-Wesley, Reading, Mass. 1987.
743. D. Fukunada, W. L. G. Koontz, *Inf. Control* **16** (1970) 85 – 101.
744. E. Paulus, D. Längle: "Die Anwendung der Karhunen-Loève-Entwicklung für die digitale Sprachanalyse und -synthese," *NTG-Fachtagung Signalverarbeitung,* Erlangen 1973, pp. 362 – 369.
745. R. M. Anderberg: *Cluster Analysis for Applications,* Academic Press, New York 1973.
746. H. Späth: *Cluster-Analyse-Algorithmen,* 2nd ed., Oldenbourg Verlag, München 1977.
747. K. Stephan, P. Metzing: "Automatisierung komplexer Chargenprozesse," *MSR (Berlin)* **29** (1986) 511 – 515.
748. B. Böhme: "Anwendung der Situationssteuerung in komplexen verfahrenstechnischen Systemen," in Wissenschaftliche Berichte, vol. 15, TH Leipzig, 1985, pp. 48 – 81.
749. R. Müller, F. Müller, A. Parusel: "Planning Man-Machine-Communication in Process Control Systems," *Proceedings of the BIAS 1990,* Milan, pp. 407 – 417.
750. G. S. Axelby, A. P. Sage (eds.): Zusammenfassung der Tagungsbeiträge der IFAC/IFIP/IFORS/IEA Conference on Analysis, Design and Evaluation on Man-Machine-Systems, *Automatica* **19** (1982) no. 6, 593 – 779.
751. J. Richter: "Experimentelle Untersuchung zur aufgabenbezogenen Gestaltung von Informationssystemen in zentralen Überwachungseinheiten," Dissertation, Humbold-Universität, Berlin 1982.
752. A. De Meyer: "Software verfahrensgerecht strukturieren," company report, Bayer Antwerpen NV, Antwerpen 1991.
753. H. Jürgens, D. Saupe (eds.): *Visualisierung in Mathematik und Naturwissenschaften,* Springer Verlag, Berlin 1989.
754. H. Müller: *Realistische Computergraphik,* Springer Verlag, Berlin 1988.
755. K. Kansy, P. Wisskirchen (eds.): *Graphik und KI,* Springer Verlag, Berlin 1990.
756. K. P. Fähnrich, C. Janssen, G. Groh: "Entwicklungswerkzeuge für graphische Benutzerschnittstellen," *Computer Magazin* **2** (1992) 6 – 13.
757. D. A. Young: *X Window Systems,* Prentice Hall, Englewood Cliffs, N.J. 1989.
758. Eckelmann: Industrieautomation, InTouch, company brochure, Wiesbaden-Erbenheim 1991.
759. PROLINK AG, Factory Link, company brochure, Seengen 1992.
760. Prozeß-Leittechnik GmbH, APROL, company brochure, Hilden 1991.
761. G. Geiser: *Mensch – Maschine – Kommunikation,* Oldenbourg Verlag, München 1990.
762. Siemens AG: Teleperm M: Das Prozeßleitsystem, company brochure no. A 19100-E815-A2-V1, 1983.
763. SattControl GmbH, SATTLINE, company brochure, Wuppertal 1990.
764. AEG, GeAmatics, Company Brochure Frankfurt 1992.
765. Eckardt AG: LS 130 Leistation des PLS 80 E, Systemübersicht PLS 180 t, Druckschrift B 133/0692, Frankfurt 1992.
766. ABB Prozeßautomation GmbH: ABB Master, company brochure, Mannheim 1992.
767. J. Encarnacao: "Die perfekte Täuschung als hochgestecktes Ziel," forschung – Mitteilungen der DFG **2** (1992) 4 – 7.
768. M. Utermöhle, H. G. Pagendarm: "Sichtweisen – Visualisierung," *Grundlagen und Entscheidungskriterien* **IX** (1992) 92 – 98.
769. H. W. Franke: "Welten aus Bits and Bytes," *VDI Nachr.* **9** (1992) 13 – 17.
770. R. Steinmetz, R. G. Herrtwich: "Integrierte verteile Multimedia-Systeme," *Informatik-Spektrum* **14** (1991) 249 – 260.
771. R. G. Herrtwich: "Betriebsmittelvergabe unter Echtzeitgesichtspunkten," *Informatik-Spektrum* **14** (1991).
772. D. P. Anderson: "Resource Management for Digital Audio and Video," IEEE Workshop on Real-Time-Operating Systems and Software, Charlottesville, May 1990.
773. P. Hoepner: "Synchronizing the Presentation of Multimedia Objects – ODA Extensions," *Multimedia – Systems, Interaction and Applications, 1st Eurographics Workshop,* Stockholm, April 1991, pp. 87 – 100.
774. N. S. Borenstein: *Multimedia Application Development with the Andrew Toolkit,* Prentice Hall, Englewood Cliffs 1990.
775. K. Meyer-Wegener: *Multimedia-Datenbanken,* B. G. Teubner, Stuttgart 1991.
776. J. L. Alty, C. D. C. McCartney: "Managing Multimedia Resources in Process Control," *Multimedia – Systems, Interaction and*

Applications, 1st Eurographics Workshop, Stockholm, April 1991, pp. 293 – 306.
777. E. Golla: "Multimedia-Anwendung für die Industrie," *CIM-Praxis* **1** (1993) 46 – 47.
778. W. Michaeli, K. Borgschulte, M. Philip: "Neue Konzepte bei der Automation in der kunststoffverarbeitenden Industrie durch Anwendung von Bildverarbeitungssystemen," *Proceedings of the 9th International Conference on Automated Inspection and Product Control,* Stuttgart, May 1989, pp. 115 – 123.
779. F. J. Kleiner, M. Liewld: "Einsatz von CCD-Array-Kameras für optoelektronische Messungen in der Umformtechnik," in: *Umformtechnik – Ideen, Konzepte und Entwicklungen,* Teubner Verlag, Stuttgart 1992, pp. 243 – 260.
780. P. Haberäcker: *Digitale Bildverarbeitung,* Carl Hanser Verlag, München 1989.
781. D. Weimar: "A Synthetical Sicual Environment with Hand Gesturing and Voice Input," *Proceedings of CHI' 89,* May 1989, pp. 235 –240.
782. D. M. Frohlich: "The Design Space of Interface," in *Multimedia – Systems, Interaction and Applications,* 1st *Eurographics Workshop,* Stockholm, April 1991, pp. 53 – 69.
783. H. Marmolin: "Multimedia from the Perspectives of Psychology," in: *Multimedia – Systems, Interaction and Applications, 1st Eurographics Workshop,* Stockholm, April 1991, pp. 39 –52.
784. R. Bogaschewsky, "Hypertext-/Hypermedia-Systeme – Ein Überblick," *Informatik-Spektrum* **15** (1992) 127 – 143.
785. M. Mühlhäuser: "Hypermedia-Konzepte zur Verarbeitung multimedialer Information," *Informatik-Spektrum* **14** (1991) 281 – 290.
786. D. Roller: "Technical Information System for Assembly, Test and Service Documentation," *Proc. of the 3rd Int. Conf. on Human Aspects of Advanced Manufacturing and Hybrid Automation,* Gelsenkirchen, Aug. 1992, pp. 273 – 278.
787. A. Lieberam: "Expertensysteme in der Chemischen Technik," *Chem. Ing. Tech.* **65** (1993) no. 4, 362 – 363.
788. H. Kumano: "Learning with Interactive Multimedia by Plant Operators," *IFAC Workshop on Production in Process Industry,* Düsseldorf, March 1993, pp. 127 – 134.
789. M. Z. Conklin: "Operator Performance Support System (OPPS)," *Proceedings of Technology* **2002,** Feb. 1992.
790. J. A. Mullens, L. C. Williams, W. L. Zabriskie: "Integrated Diagnostic System for a Process Controller," *International Mashinery Monitoring and Diagnostic Conference,* Las Vegas, September 1989.
791. J. P. Hack, P. H. Heldt: "Das Cockpit des Airbus A 310," *Spektrum der Wissenschaft,* March 1984, 38 – 52.
792. Blaupunkt GmbH, ALI, company brochure, Frankfurt 1984.
793. BMW AG, company brochure no. VS-202 IX/83 311 500 410, München 1983.
794. BMW AG, company brochure no. 411 070 111-1/84 VM, München 1984.
795. Siemens AG, Teleperm M – Das Prozeßleitsystem für Kraftwerke, company brochure no. A 19100 E661-A2-V1, 1984.
796. M. Polke, M. Wienand: "Anforderungs- und Qualifikationsprofile," Festschrift zum 60. Geburtstag von Prof. Dr. Stöckmann, Karlsruhe, Oct. 1978.
797. G. Farber, M. Polke, H. Steusloff: "Mensch –Prozeß – Kommunikation," *Chem. Ing. Tech.* **57** (1985) no. 4 307 – 309.
798. P. I. Burton, *Chem. Eng. (Rugby, Engl.),* March 1981, 103.
799. E. Blass: "Methodische Entwicklung verfahrenstechnischer Prozesse," *Chem. Ing. Tech.* **57** (1985) no. 3, 201 – 210.
800. W. E. Bixby, P. G. Andrus, L. E. Walkup, *Photogr. Eng.* **5** (1954) 195 – 201.
801. I. T. Bickmore, R. E. Hayford, H. E. Clark, *Photogr. Sci. Eng.* **3** (1959) 120 – 214.
802. L. A. Jones, *J. Franklin-Inst.* **187** (1920) 29.
803. M. Polke: "Zur Physik der Elektrophotographie," in: *Festkörperprobleme,* vol. VIII, Vieweg/Pergamon Press, Oxford 1968, pp. 131 – 174.
804. H. Rumpf: "Über die Eigenschaften von Nutzstäuben," *Staub Reinh. Luft* **27** (1967) 3 – 13.
805. R. Polke, N. Scholz: "Quality Control in Solids Process Engineering," *Part. Charact.* **2** (1985) 39 –48.
806. R. Ecker, H. Kramer, H. K. Müller, M. Polke: "Neue Aspekte bei der Prüfung von Kautschuk und Gummi," *Kautsch. Gummi Kunstst.* **1** (1972) 5 – 10.
807. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie, Status und Trend," *Automatisierungstechnische Praxis (atp)* **28** (1986) no. 9, 423 – 431;no. 10, 479 – 484.

808. H. Töpfer, H. Bischoff: "Grundlagen der Prozeßanalyse," KTD. Wissenschaftlich-Technische Gesellschaft für Meß- und Automatisierungstechnik, Lehrbrief 1 bis 4, Dresden 1979.
809. J. Lauber, M. Polke: "Das Phasermodell der Produktion als methodisches Hilfsmittel für die Prozeßentwicklung," *GVC-Sitzung Prozeß- und Anlagentechnik,* Oct. 1992, Wiesbaden.
810. M. Breitenfelder, M. Polke, M. Wienand: "Der Weg eines Produktes vom Konzept zur Kontrolle," lecture held by M. Polke at the Arbeitstagung *Forschungsmanagement in der Physik* 9. – 12. June 1976, Universität Bielefeld.
811. M. Polke, P. Siara: "Chemiefasern – Chancen und Probleme in der Zukunft," Vortrag anläßlich des Zukunftssymposiums – Textil "Märkte morgen", 1975, Verband der Textilindustrie Westfalen, Schriftenreihe Heft 22.
812. M. Polke, M. Wienand: "Der Physiker in der Industrie – Ausbildung und Beruf," *Phys. Bl.* **6** (1975) and 1 (1976).
813. R. W. Peters, M. Polke: "Verfahren zur Objektivierung der Studienreform," VDI-Hauptgruppe Ingenieurausbildung, Düsseldorf 1972.
814. R. W. Peters, M. Polke: "Beruf und Ausbildung der Ingenieure – Eine empirische Untersuchung," VDI-Hauptgruppe Der Ingenieur in Beruf und Gesellschaft, Düsseldorf 1972.
815. H. Hillmer, R. W. Peters, M. Polke: *Studium, Beruf und Qualifikation der Ingenieure,* VDI-Verlag, Düsseldorf 1976.
816. M. Polke: "Die Bedeutung des Qualifikationsprofils des Ingenieurs für die Gestaltung der Studiengänge in der Ingenieur-Aus- und Weiterbildung," *FEANI-OECP-Seminar,* München 1973.
817. H. Drahten, M. Polke: "Anforderungsprofil und Qualifikationsprofil als Basis zur Ermittlung des Weiterbildungsbedarfs – Ergebnis einer Umfrage," in G. Schmidt, H. Steusloff (eds.): "Mit vernetzten, intelligenten Komponenten zu leistungsfähigen Meß- und Automatisierungssystemen," *Tagungsband Interkama-Kongreß 1989,* Oldenbourg Verlag, München, pp.643 – 651.
818. M. Polke: "Aus- und Weiterbildung in und an moderner Prozeßleittechnik," in W. Brückers, N. Meyer (eds.): *Arbeit, Technik und berufliche Bildung im Metallbereich,* vol. 3, Verlagsgesellschaft, Köln 1988, pp. 27 – 53.
819. M. Polke: "Prozeßleittechnik in der Produktion – Voraussetzung für Diversifikation von Chemiefasern," *Lenzinger Ber.* **67** (1989) no. 8, 12 –24.
820. K. H. Padberg, P.-Th. Wilrich: Die Auswertung von Daten und ihre Abhängigkeit von der Merkmalsart, part 1, *Qualität und Zuverlässigkeit* **26** (1981) 179 – 183.
821. DIN 31 051, Begriffe und Maßnahmen, Beuth Verlag, Berlin 1985.
822. F. Zutt, H.-P. Hubich: "Sinnvolles Agieren war früher kaum möglich; Praxisbeispiel. Instandhaltung aus der Sicht der chemischen Industrie. Instandhaltungsmanagement der 90er Jahre," in W. Schulte, G. Küffner (eds.): *Blick durch die Wirtschaft,* Frankfurter Allgemeine Zeitung GmbH, Frankfurt 1988, pp. 287 – 298.
823. R. Isermann: Model Based Fault Diagnosis of Machines via Parameter Estimation Methods and Knowledge Processing,"1991.
824. H. Strickling et al.: "Erstellung von Erhaltungskatastern," company report, Bayer, Leverkusen FS 1975, PLT 1987.
825. W. Aleite, D. Wach: "Supervision and fault diagnosis," *IFAC-Symposium,* Baden-Baden 1991.
826. N. Kuschnerus: "Die EDV-unterstützte Planung der Instandhaltung: Instandhaltung als Teil der Anlagenwirtschaft," company report, Bayer AG, Leverkusen 1988.
827. W. Männel: "Modernes Instandhaltungsmanagement: Anforderungsgerechte IH-Konzepte mehren den Unternehmenserfolg," *IH-Markt,* 1989, pp. 4 – 14.
828. IEC: Status and Rules of Procedure of the IEC, 01 (Central Office), 809, 809 A, July 1990.
829. M. Polke, B. Will: "Tradition und Fortschritt – 40 Jahre NAMUR," *Automatisierungstechnische Praxis (atp)* **32** (1990) no. 3, 105 – 109.

General References

830. W. Ahrens: "Einsatz von Expertensystemen in der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **29** (1987) 475 – 485.
831. W. Ahrens: "Expertensysteme für die Prozeßführung, Erfahrungen aus dem Verbundprojekt TEX-I," *Chem. Ing. Tech.* **62** (1990) no. 8, 635 – 644.
832. R. Bernhardt, C. Koppermann, U. Müller-Nebler, K.-H. Wietschorke:

"Expertensystem für betriebsnahe Problemstellungen," *Automatisierungstechnische Praxis (atp)* **31** (1989) 580 – 587.
833. H. Fittler: "Anwendung von Expertensystemen in der Prozeßautomation," *Automatisierungstechnische Praxis (atp)* **31** (1989) 74 – 80.
834. W. Kaufmann (ed.): *Diagnostische Entscheidungsprozesse in der inneren Medizin,* Schattauer, Stuttgart 1986.
835. R. Soltysiak: "Praktische Anwendung von Expertensystemen in der Prozeßleittechnik," *Automatisierungstechnische Praxis (atp)* **32** (1990) 247 –257.
836. R. Soltysiak: "HEPROX, ein Expertensystemshell für Prozeßführungsaufgaben," *Automatisierungstechnische Praxis (atp)* **33** (1991) 74 – 80.
837. T. Mc Cusker: "Neural Networks and Fuzzy Logic, Tools of Promise for Controls," *Control Eng.* **37** (1990) May, 84 – 85.
838. B. Kosko: *Neural Networks and Fuzzy Systems,* Prentice Hall, Englewood Cliffs 1990.
839. L. A. Zadeh: *Fuzzy Sets, Information and Control,* **8** (1965) 338 – 353.
840. H.-I. Zimmermann: *Fuzzy Set Theory – and Its Applications,* 2nd ed., Klumer-Nijhoff Publ., Boston 1990.
841. I. Conklin: Hypertext: A Introduction and Survey, *IEEE Comp. Mag.* **20** (1987) no. 9, 17 – 41.
842. P. A. Gloor, N. A. Streitz (eds.): "Hypertext und Hypermedia," *Informatik Fachberichte 249,* Springer Verlag, Berlin 1990.
843. P. A. Gloor: *Hypermedia-Anwendungsentwicklung,* Teubner, Stuttgart 1990.
844. T. Nelsen: "A File Structure of the Complex, the Changing and the Indeterminate," *Proc. 20th Nat. ACM-Conf.* 1965, 84 – 100.
845. R. Eckmiller, G. Hartmann, G. Hanske (eds.): *Parallel Processing in Neural Systems and Computers,* Elsevier, Amsterdam 1990.
846. H. Ritter: *Neuronale Netze,* Addison Wesley, Reading, Mass., 1989.
847. G. Bretthauer, G. Straube: "Einsatzmöglichkeiten für Fuzzy-Regler," Frauenhofer-Institut für Informations- und Datenverarbeitung, Dresden 1993.

Specific References

848. D. Balzer, V. Kirbach, V. May: "Knowledge based Process Control," in: H. Steusloff, M. Polke (eds.): *Integration of Design, Implementation and Application in Measurement, Automation and Control,* INTERKAMA Tagungsband, Oldenbourg Verlag, München 1992.
849. D. Balzer, V. May, U. Starke: "Wissensgestützte Prozeßführung," part 1: *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 1, 36 – 43; part 2: *Automatisierungstechnische Praxis (atp)* **34** (1992) no. 2, 85 – 92.
850. D. Balzer, V. Kirchbach, V. May: "A Real-Time Expert System for Process Control," *Proceedings of the 2nd IFAC Workshop on Artificial Intelligenze in Real-Time Control, AIRTC ′89,* Shenyang 1989.
851. B. Böhme, U. Starke: "Expert Systems in Process Control," *Proceedings IFAC Symposium on Large Scale Systems,* Berlin 1989.
852. V. May, R. Wieland, B. Böhme, D. Balzer: "PROCON-I ein Expertensystemshell zur Lösung höherer Automatisierungsaufgaben," *Proceedings of the 33rd IWK,* Ilmenau 1988.
853. V. May: "Grundlagen und Entwicklung von Rahmen-Expertensystemen für Aufgaben der Prozeß-führung -das Expertensystemshell PROCON," Dissertation, TH Leipzig 1989.
854. D. Balzer, V. Kirbach: "Neuronale Netze in der Prozeßautomatisierung," *atp* **34** (1992) no. 7, 395 –401.
855. N. Kunde: "Methoden und Werkzeuge zur Wissensakquisition in der Prozeßautomatisierung," Dissertation, TH Leipzig, 1991.
856. U. Starke: "Eine Methodologie für das Knowledge Engineering in der Prozeßsteuerung," Dissertation, TH Leipzig 1990.
857. I. Rasmussen: "The Role of Hierarchical Knowledge Representation in Decision Making on System Management," IEEE Transaction on Systems Man and Cybernetics, vol. SMC-15 (1985) no. 2, 234 –243.
858. U. Epple: "Modular Technology, Basis of an Object-Oriented Network Structure for Process Control Engineering," *GMA-Fachbericht 4,* VDE-Verlag, Berlin 1993.

859. M. Niese, EP 0 104 411, 1986.
860. A. Strickling: "CAE in der Prozeßleittechnik," *NAMUR-Statusbericht 1987,* pp. 211 – 217.
861. P. Chen: "The Entity-Relationship Model – Towards a Unified View of Data," *ACM Transactions on Database Systems* (1976) 1.
862. H. Robel et al.: *Lehrbuch der chemischen Verfahrenstechnik,* VEB Deutscher Verlag für Grundstoffindustrie, Leipzig 1983.
863. E. Cassirer: *Substanzbegriff und Funktionsbegriff,* Springer Verlag, Berlin 1910.
864. E. D. Gilles, E. Nicklaus, M. Polke: "Sensortechnik in der Chemie – Status und Trend," *Automatisierungstechnische Praxis (atp)* **9** (1986) 423 – 431; **10** (1986) 479 – 484.
865. M. Polke: Prozeßleittechnik, Manuskript der Vorlesung, Universität Stuttgart.

Sustainability and the Chemical Industry – Ethical Aspects and Approaches

Ute Müller-Eisen, Bayer AG, Leverkusen, Federal Republic of Germany

Herwig Hulpke, Bayer AG, Leverkusen, Federal Republic of Germany

1. Introduction 1
2. Structural Change and Change of Paradigms 2
2.1. Chemistry and Industry 2
2.2. Social System 3
2.3. Environmental Awareness 4
3. Ethical Aspects 5
4. Chances for the Future 6
4.1. Concept of Sustainable Development 6
4.2. Dialogue 7
4.3. Risk/Chance Assessment 8
4.4. Innovations 10
5. Approaches of the Chemical Industry 10
5.1. Responsible Care 11
5.1.1. Basic Idea 11
5.1.2. Management Systems 11
5.1.3. Production-Integrated Environmental Protection 12
5.1.4. Product Stewardship 12
5.1.5. Dialogue 13
5.2. Environmental Agreements 14
6. References 14

1. Introduction

With the growing concern in society as a whole about the form the future will take, ethical aspects have acquired increasing significance for the chemical industry. They center on the responsibility of employers and the principles governing their actions. This is because a wide variety of complex questions exist – not only in the interaction between companies and their social and economic environment, but also within the companies themselves – which have considerable bearing on the fields of ethics and morality. Some time ago, such topics of discussion generally received little attention. This changed at the latest by the 1980s due to increased awareness for the environment and for life in general. Questions of responsibility have been integrated into society and industry following calls to align corporate activities not only with the present generation, but also increasingly with future generations. The central question is therefore: How can we make sure that the possibilities for development available to future generations are not restricted by our present-day actions? This is an important requirement within the framework of the idea of "Sustainable Development", as resolved with worldwide consensus at the environmental conference in Rio de Janeiro in 1992 (see Section 4.1) [1]. This resolution contains an unambiguous appeal to our responsibility to conserve the environment, to use its resources sparingly, and provide society exclusively with production processes and products which stand up to an examination of this kind. Today, it is not only technology itself which is at the center of criticism, but also its products. One aspect focuses on weighing up chances and risks for humans and the environment. An important concept is that ecology, economy, and social matters have equal ranking.

The call for Sustainable Development is linked closely with ethical considerations, because it asks questions about the behavior and responsibility of people in the present, but with a view to the future. Its implementation in the chemical industry is largely in the hands of science. The special consideration of the chemical industry is not just a matter of chance; because of its experience and know-how, the chemical industry has the potential to influence production and products in the sense of the abovementioned sustainability. However, its contribution should only be regarded as a partial solution, because Sustainable Development is influenced by many factors that are closely interlinked. Apart from this, the call for equal consideration of economy, ecology, and social matters ne-

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b07_001

cessitates optimization of the overall structure. All sections of the human community (producers, customers, consumers, etc.) must contribute to this. Since the chemical industry, as part of the overall system, extends into all three areas, its contribution must be seen against the background of their historical development and the history of paradigms.

2. Structural Change and Change of Paradigms

2.1. Chemistry and Industry

Striving to understand the processes of nature and exploiting them to improve the standard of living has accompanied the history of mankind since its earliest beginnings. Chemistry as a natural science sprang from this endeavor in close connection with the dominant *Weltanschauung* of the respective epochs. In the beginning, much practical knowledge about the characteristics and effects of natural substances was acquired by simple experimentation. This led to the production of commodity articles and the ability to distinguish between poisonous substances, medicinal substances, and foodstuffs. Dyes and ceramics are regarded as the oldest products of humans. Technology was given a particular stimulus by the development of metallurgy, and mention should perhaps also be made of glass production, tanning, soapmaking, and of foodstuffs like beer, wine, and sugar. One of mankind's major challenges was, and still is, not only to use natural substances but to redesign and manufacture substances himself. With increasing know-how and experience, chemistry become a systematic and exact science which, step by step, created the basis for the achievements of modern times and the commercial application of scientific findings.

Modern-day chemistry and its industrial significance only really began about 200 years ago with, among others, the production of soda and sulfuric acid in the late 1700s and early 1800s (inorganic chemistry). Other key steps included the discovery of coal-tar dyes, the development of organic chemistry (from 1850 onwards), and polymers and plastics in the 1900s. Fundamental discoveries and their development into industrial processes broke fresh ground and provided the potential for major qualitative advances. The success created a conception of chemistry that was characterized by a belief in progress and a demand from the marketplace.

The number and quantity of products and their consumption rose dramatically. Thousands of materials are nowadays manufactured on an industrial scale. The products generated by the chemical industry are now found in all walks of life, and it is impossible to imagine everyday life without them: examples are pharmaceuticals, textiles, dyes, plastics, coatings and engineering materials, developed specifically for applications in the construction industry, the home, transport, food, sport, leisure, medicine, etc.

As regards chemistry's dynamism of development and its self-conception, the environmental aspect tended to take a back seat to what could best be described as a unidimensional aim to achieve progress until the second half of the twentieth century. It was only with the strong upswing in the chemical industry in the 1950s, which led to significant impact of production on the environment (air, water, soil), that the need arose to take countermeasures. Being evidently a technical problem, it was tackled with the aid of technology where the problems manifested themselves, that is, at the end of the production process, and this led to the birth of end-of-pipe environmental protection. This provided for the treatment of wastewater, solid waste, and waste air in large, mostly centralized facilities such as wastewater-treatment plants and incinerators. These end-of-pipe measures produced success in a relatively short time, and made an important contribution to improving the environmental situation in the 1970s, for example, in Germany. In subsequent years, both for technological and for economic reasons, a structural change has been taking place throughout the chemical industry from end-of-pipe environmental protection to preventive, integrated environmental protection (see Section 5.1.2).

The chemical industry today operates globally. As a supplier to many other branches of industry, it is part of the economic chain and intricate social inter-relationships. The degree of interlinking is very high indeed. Structural changes in the chemical industry have numerous implications for the structure as a whole. Globalization and the interlinking of

present-day markets necessitates a holistic approach, in which a partial optimization (ecological/economic/social) can have a negative effect in the long term.

2.2. Social System

Industrial development over the last two hundred years is marked not only by technical and economic innovations. It has also been accompanied by a wide range of social changes that have affected man's conception of himself. Increasing mass production and mounting consumption, a high standard of living, high division of labor, and specialization of skills and know-how, an increasing flow of information, greater mobility, and so on, have gone hand in hand with the opening up of new markets and the creation of new branches of industry, areas of work, and professions. The level of interaction and interlinking has increased enormously in all areas of life and work, especially in the highly industrialized societies. Above all, the far-reaching developments in electronics (e.g., computer and satellite technology) have brought profound, as yet, not entirely foreseeable changes, and have exerted an influence right the way down to our way of thinking. Many processes have been accelerated and expanded. For a long time, the maxim was "higher, further, faster". Traditions were frequently challenged or even broken. Fundamental rules which were embodied, for example, in religion, morals, and legal systems, lost their validity. Adjustment and rethinking processes were needed.

New technologies with a deep level of penetration were installed. As a result, the radius of action of the companies increased, but so did the possibilities of individuals to liberate themselves from the constraints of nature with the aid of technology. However, this was accompanied not only by benefits and greater freedom. New dependences came into being, which resulted from the way our present-day life is organized. For example, modern means of transport and traffic routes have made it possible to separate the place where we live from the place where we work. However, we have as a result made ourselves dependent on reliable, flexible, and inexpensive means of transport. Another example is the far-reaching division of labor: humankind is nowadays increasingly less in a position – either technically or from the organizational viewpoint – to produce its own clothes and feed itself. Life, also that of the individual, nowadays takes place more on the basis of interpersonal and social agreements. People are dependent on the specialist knowledge and skills of experts. An important factor here is trust in the work of another. This dependence, however, generates a subliminal insecurity, because man is no longer able to control many components of his life himself.

There are other dependences, too: the standard of living in the highly industrialized societies grew in the second half of the twentieth century to such an extent that the elemental needs of man (food and somewhere to live) were essentially satisfied and were no longer regarded as vital needs. This opened up the possibility of shifting human needs to a higher plane. In line with Maslow's hierarchy of needs, new qualities came along which had to be fulfilled, such as leisure, self-realization, craving for admiration, etc. This has progressed so far that, in a highly industrialized society, the benefits of many products of the chemical industry are taken for granted. The negative effects of industrialization and technology are no longer tolerated. Today, a high, perhaps exceedingly high, level of safety is demanded; risks are rejected entirely. The logical consequence of this approach is that chances are lost, because chances and risks are closely linked to one another.

The technical/economic differentiation in the highly industrialized countries has created a variety of life styles and criteria of values. With regard to the future alignment of society, this pluralism is a factor which should not be underestimated, and one which increases the level of difficulty of future solutions.

2.3. Environmental Awareness

There is no doubt that the chemical industry has played a major role in the appreciable improvement in the standard of living and life expectancy of man, and that the western world is today living at a very high standard. Practical reality has, however, shown that the achievements of the chemical industry cannot have only advantages,

but also disadvantages for mankind. These include negative effects of production facilities on the environment (solid wastes, emissions, accidents) and the interaction of products with the environment and, in some cases, undesirable effects on human health. Yet negative effects such as these are not a phenomenon of the modern age. Throughout history, man has intervened in many different ways in the natural balance, and in some cases caused considerable changes. An example from the ancient world is the salization of Mesopotamia, and from the pre-industrial age, the deforestation in the Mediterranean region and the contamination of waterways by tanneries. However, for long time, the effects of human actions were limited to local, or at most, regional areas and were often reversible. The depth of technical penetration was insufficient to have broader implications. Processes of change proceeded relatively slowly and their range of action was fairly small. Today, however, the scope for human activity has attained a global dimension, which means that the consequences are far more interlinked and, as a result, may also become irreversible. Apart from the space aspect, there is also the time-related long-range effect.

The first major wave of events that alerted the public to the impact of industrial developments – especially those in the chemical industry – on the environment came in the 1950s and 1960s, when the effects of large-scale chemical production became apparent. At that time, however, this did not provoke any general discussion. The desire for the consumption of products that improve the standard of living was still the priority for most people. The chemical industry was not generally challenged, nor was its technology or its products. Although there was little public discussion, specialists began debating the problem back in the 1950s, and this debate gradually intensified as scientific knowledge grew, particularly in the field of (environmental) analysis. In parallel with this, scientific findings were collected about substances and products, with the result that the substance or product-related aspect was added to the environmental-segment-related approach, which was formerly directed primarily at the effects of the production processes. Following on from this, a broad international exchange of views among experts relating to the production and products of the chemical industry was initiated, which expanded over the next few years to include society as a whole, and which has resulted in present-day discussion.

Ethical aspects also became increasingly involved, all the more so as negative effects of chemical production became apparent, such as the impact on the environment of solid waste, wastewater and waste air, harmful side-effects of products, and chemical accidents. Because of the high consumption of resources in the industrialized world, questions were also raised about limitations to growth [2]. This development was accompanied by the mounting interest of the media in environmental topics, and partly also by the formation in the late 1970s of environmental groups calling for a change in the ecological awareness of society. Consequently, since the beginning of the 1980s, people throughout the world have become increasingly sensitive to environmental matters and to the risks generated by the production processes and products of the chemical industry. On some occasions, this progressed to such an extent that some groups wanted to do away with production but were not willing to do without the products. The situation was accompanied by a mounting scepticism towards technology (ranging from a vague uneasiness to demands for abandonment) and a growing, sometimes utopian call for absolute safety. The belief in traditional progress, based solely on technology, has been significantly relativised as a result. The evaluation of technology became a programmatic concept.

Growing environmental awareness was connected with the call for a form of environmental protection which penetrated all areas of life, and which, in some cases, assumed the character of a moralistic ideal and, at its climax, even manifested itself as *Zukunftsangst* (fear of the future). Discussion in western societies is nowadays characterized by the question of what kind of progress man needs and what he can accept responsibility for.

3. Ethical Aspects

The chemical industry is embedded in the social currents with their ecological, social, cultural, and political aspects. It must take into account and evaluate not only these developments, but also its own experience and know-how to make decisions for the future. The resulting activities

in turn have consequences. Since ancient times, the question "How should man act?" has been at the center of ethics (Greek: *ethos*), the philosophical science of morality [3]. Ethics, as the theory of values and norms, is nowadays often regarded as being synonymous with moral philosophy. The term moral (Latin: *mos*), which is often equated with ethics, should be understood in its meaning of morals in their entirety, as the practical application. Generally, ethics involves the question: To which values and norms, goals and objectives should man orientate his activities? The main subject of the ethical approach are human activities, with the accent either on the character from which these arise ("Gesinnungsethik"), or on the effects generated by them ("Erfolgsethik"). Ethical thinking also deals with the question of whether the moral willpower and esteems are innate, and thus, to a certain extent, common to all mankind, or whether they are gained from experience, and thus change from people to people and from era to era. Over the course of time, depending on the starting point of the approach, many different variations in the direction of ethics have been created.

The social way of thinking has changed over the course of time. Ethical reflection began in ancient philosophy with Socrates and Plato. Through Aristotle, it became a science of its own. Conceptions and analyses of these ancient thinkers have been influencing philosophy and thinking right through to the present time, and have consequently also influenced the course taken by man's understanding for nature and science. Other famous personalities like Galileo, Descartes, Kant, Hegel, and Schweitzer also contributed new or opposing aspects and findings (ethics of discourse, duty, virtue, utility, reason, etc.). They made a major contribution to the change in the paradigm. An admiration for nature alternated with a contempt for it and its exploitation. Man's relationship with nature ran in parallel with this, and his relationship to technology and the development of western culture was also shaped by philosophers. The modern technology of the present industrial countries can be seen as the offspring of the Renaissance. Technical procedures and the basic approach and attitude were decisively influenced by the philosophers Bacon and Descartes. Bacon developed the inductive method in science. He had the vision of expanding scientific knowledge to such an extent through rationally planned empiricism that man would acquire control over nature. Descartes developed Bacon's approach further. He saw man as a subject who had at his disposal everything else as objects. This had a major influence on the philosophy of technology.

There is, as yet, no conclusive answer to the question of man's relationship with nature. The duality of the object to be exploited and the subject to be protected is still evident today in the way the western world sees nature. This question, together with the fact that modern technology has resulted in activities of a new dimension, with new consequences, has led to the formation of one of the "new ethics", namely, "environmental ethics". This states that ethical decisions should involve the whole of humankind, include the future, and take appropriate account of ecological matters. All in all, it is a question of developing suitable ways for humankind to handle nature. For this, it is useful and indeed essential to have moral principles for such behavior. This, in turn, leads to questions about man's role in all this. Depending on the standpoint, a distinction can be made between anthropocentric (centering on humans), pathocentric (humans and animals are equal because of their capacity to suffer), biocentric (all types of animate nature are equal), and holistic (regarding animate and inanimate nature as a complete structure) environmental ethics. Human dominance is up for discussion. Many articles have been published on environmental ethics. All thoughts on this subject probe the emergence of environmental problems and seek a strategy to overcome them.

The philosopher Jonas developed the challenges of modern technology for the ethical orientation of human action into a theory of extended responsibility [4]. He regards ethics as no longer being confined to the immediate vicinity, with regard to time and space, of activities, but feels that present-day ethics should take into account the global conditions of human life, and, at the same time, attend to the distant future. He takes this right through to the question of the continued existence of humans as a species. In other words, ethics has to deal with activities which extend further into the future than ever before. A central question here is the aspect of irreversibility. Jonas' principle of responsibil-

ity should be used to slow down human activities and thus put a brake on the "Promethean-utopian creative urge". Jonas used as his compass the preconceived risk and the fear associated therewith. The change in the range of responsibility is, for Jonas, the central point of his ethics of the future, and it includes not only the future of humankind but also the future of nature.

It is certainly undisputed that humans should not be allowed to do everything they are technically capable of doing. However, they should not be excessively restricted in creating something new, because technology offers the chance of realizing Sustainable Development. Dispensing completely with technology is therefore just as harmful as using it unrestrictedly. The aim should be to act responsibly with a preventive approach. Standards or yardsticks are required for this, but they cannot be set by scientific knowledge alone. Today, more than ever before, an interdisciplinary, holistic approach is needed. This includes eliminating the sometimes exaggerated separation of the humanities from the natural sciences. The existing conflicts must be broached. For example, natural science, which is characterized by experimentation, must, because of the inductive methodology, get to grips with the reproach that, although it produces knowledge, it does not produce the necessary cognition. Discussion of the two directions is particularly important with regard to sustainability's call to give ecological, economic, and social matters equal ranking. There is no doubt that we need a new dimension and quality of human sense of responsibility which leads to a modern ethics of responsibility.

4. Chances for the Future

4.1. Concept of Sustainable Development

Against the background of a dynamically growing world population, limited natural resources, the sensitivity of our ecosystems, and lack of food for many peoples of the world, discussion about the further development and responsibilities of humankind today is characterized by the term "sustainability". This term is not new. It was used back in the early 1700s in German forestry parlance. It referred to the principle that utilization of a renewable resource such as a forest must not occur faster than the rate of regeneration of that resource. The term sustainability was extended to cover other natural resources (e.g., fisheries) and is becoming established worldwide. The concept of forestry management, which was originally primarily an economic one, was supplemented over the course of time by ecological aspects, and has ended up today more as an ethical matter. The term "Sustainable Development" was mentioned for the first time in 1980 in the World Conservation Strategy of the International Union for Conservation of Nature and Natural Resources (IUCN). In 1983 it was defined as follows in the UN-appointed World Commission for Environment and Development, chaired by the Norwegian politician Brundtland: "Sustainable Development is development that meets the needs of the present without comprising the ability of future generations to meet their own needs."

The Brundtland report *Our Common Future*, published in 1987, encouraged worldwide discussion on Sustainable Development [5]. It reached its temporary climax in 1992 at the UN Conference on Environment and Development in Rio de Janeiro [1]. There, Sustainable Development was unanimously declared the common aim and the guiding principle for national and international environment and development policy. It was defined in more detail in the so-called Agenda 21 [1], [6]. This lists, among others, measures for its concretization and a call for its implementation. It gave as one of the key requirements for achieving the goal of Sustainable Development, nationally and internationally, that the measures take equal regard of ecological, economic, and social matters. These three pillars for continued existence in the future substantiate the call for ethical responsibility. This means that, apart from the economic questions concerning the efficient allocation of goods and resources, they should also include not only ecological problems such as the shortage of resources and environmental pollution, but also social problems such as hunger, poverty, and disease, in order, for example, to also find solutions to the questions of justice between the rich "north" and the poor "south" (intragenerative conflicts) or between present-day and future generations (intergenerative conflicts). All sec-

tions of the human community must face up to this challenge. The UN Commission on Sustainable Development (CSD) has the job of monitoring the implementation of Agenda 21.

Nevertheless, it must be said that a comprehensive sustainability concept of this kind with the triad of ecological, economic, and social aspects contains a contradiction and, in many cases, an almost insoluble conflict of objectives. In any case, all the different approaches make it necessary to take a very wide-reaching, interdisciplinary view.

The idea of Sustainable Development is thus a call to integrate [7]. Ecology, economy, and social matters should continue to develop together in such a way that, in a longer term, none of the three areas is irreversibly damaged. The concepts dealing with the ways of implementing Sustainable Development and its content and objectives differ considerably. This is due to the complexity of the questions and the plurality of the world's societies. Yet we should not expect any simple, uniform solutions. Sustainability can neither be centrally predetermined nor scientifically derived and quantified. This is not possible for one thing because the needs of future generations can only be defined to a limited extent. Ideas of what sustainability is are, to a very large degree, time, situation, culture, and knowledge-dependent, and are subject to constant change. For example, it is not possible to conclusively answer the question of which raw materials should now be conserved for the future, and in what quantities. The basic idea of Sustainable Development, on the other hand, is to avoid narrowing the chances for future generations. The demand that sustainability should be understood in a naive manner, in line with the early considerations on forestry management, would certainly not be adequate today. Instead, the call for sustainability must be seen as a regulative idea, comparable with other major concepts like freedom or justice. Ideas such as this prescribe directions and name goals. Searching and learning processes in the various sections of society must be oriented to it, and it must be understood as a communicative task which encourages motivation and the willingness to participate. These are protracted processes that will never be concluded, because new findings will repeatedly make course corrections necessary in a continuous process of improvement. In other words, Sustainable Development can only be the result of a permanent social searching and learning process. As part of the overall system, the chemical industry has adopted the idea. In its implementation, it focuses on self-responsible actions, coupled with self-imposed commitments and co-operation. Sustainable Development is not a self-contained concept for action, but a framework for designing the future.

4.2. Dialogue

From the open conception of the idea of Sustainable Development and its integration principle there arises the direct necessity for social communication processes. With regard to the equality of ecological, economic, and social matters, these processes must contain interinstitutional and interdisciplinary discourse. However, this is only possible if the habits of demarcation and self-imposed isolation of individual institutions and specialist disciplines (which have grown with industrialization and are based on different sets of logic for action and regles du jeu and their own semantics) are broken down. The implementation of this idea must be a central goal of the integration concept. Communication is the hinge that can bring opposing points of view together.

Target-oriented communication is based on knowledge and understanding. The concept of sustainability in particular is dependent on mutual insight, because sustainability must be lived, it cannot be prescribed. For the social process of communication, it is therefore not sufficient to provide information. The information must be evaluated and turned into reality. Man finds himself in the dilemma of being flooded with information. There is an optimum opportunity to gain knowledge, but it must be examined with regard to its content, and this is almost impossible for an individual because of the volume and complexity of the data and the lack of yardsticks for evaluation. Such evaluation yardsticks must be drawn up in a dialogue and then be integrated in decision-making processes.

The willingness for dialogue and for being involved in processes to establish targets in society must be encouraged. Environmental education could contribute to this. In particular, it must extend the knowledge of systematic relationships,

particularly in the fields of ecology, economics, and society. An environmental education contributes to the formation of an awareness, which is a precondition for strengthening personal responsibility and motivation. These, in turn, form the basis for changes of attitude. Because Sustainable Development is a continuous process, it is essential that society is constantly willing to learn.

One factor concerning communication within the society which must not be ignored are the emotional aspects. These arise increasingly in discussion about the safety or risks of technology. Decisions, particularly by laypersons, are frequently not made on the basis of rational, well-founded considerations but are influenced by emotional factors, such as fear. Media reporting plays an important role in this. In general, emotions are very difficult to overcome with purely technical or scientific arguments, because there are two different planes here, the rational and the emotional, which inter-relate. In many cases, these planes collide irreconcilably with one another. Finding a solution is out of the question if emotions are not acknowledged as being arguments and given the same ranking as facts. Consequently, suitable forms of dialogue must be found which can bring about an agreement of opinion on the identification, evaluation and management of risks. Part of this dialogue is to discuss chances and risks and to weigh them against each other.

4.3. Risk/Chance Assessment

During the last two decades, there has been an increasing demand in the highly industrialized countries for safety and a declining willingness to take risks. Many people believe they are increasingly exposed to risks as a result of technical developments. There are a number of very different reasons for this feeling, depending on the individual's or society's mental approach, knowledge, and yardsticks. Undoubtedly, that the high present-day standard of living, with its high level of safety, and the fact that people can look beyond simply satisfying their basic needs and openly take in the side effects, too, play an important role. Other factors which have contributed to the intensification of discussions on the subject of risk include the improved possibilities for scientific research, e.g., environmental analysis, and the influence of the media.

Risk is a complex term with manifold problems concerning its comprehension. It is based on the word *risco*, which was common in Italy in the 1200s, when Venetian merchants used it to appraise the risk of loss of commercial fleets. In subsequent years, various disciplines (e.g., insurance mathematics, decision theory) have become involved with the phenomenon of risk, with the result that we have different definitions. Common to all of them is the fact that risk can be described as a concept for appraising an expected loss.

The term "risk" is linked to a whole chain of other terms: risk description/risk evaluation/risk comparison/risk perception/risk communication/risk management/risk acceptance. This shows that very different worlds of thinking come together in this term. Formulated simply, it covers the technical and scientific world of thinking, and the social and political world of thinking. The scope of competence for decision-making decreases in this order for the technical and scientific segment, but increases for the social and political segment. Figure 1 illustrates these relationships in schematic form: Risk description is predominantly a technical and scientific matter, while risk acceptance is primarily a social and political domain. This view should help to clear up the misunderstanding that a high level of expertise for describing risks necessarily leads to a high expertise for accepting risks. This does not apply conversely either.

Figure 1. Scope and competence for decision making

Deeper analysis of the subject of risk shows that only a holistic approach can lead to correct decisions. For the three segments of Sustainable Development, this means that ecologi-

cal, economic, and social risks must be weighed up against each other in order to proceed along the path of Sustainable Development. An approach to risk such as this can also be helpful in the context of conflict management, in which it has proved useful to examine conflict of facts/conflict of interest/conflict of convictions separately, and not to mix them unnecessarily.

Another source of misunderstanding is very often the opinion that a risk can be completely eliminated if the minimization measures go deep enough. A mathematical-formalistic approach shows, however, that risks, however comprehensive the countermeasures, can never be reduced to zero. This relationship can be described by a curve which approaches the baseline asymptotically but never reaches it (Fig. 2). In fact, the curve shows even more than this: measures to reduce a risk can themselves involve risks. Accordingly, there is a certain optimum of measures at which the risk is at a minimum. In other words, an excessive number of measures is just as harmful as too few. Dispensing completely with the benefits of something must also be viewed critically, because this approach can harbor new risks.

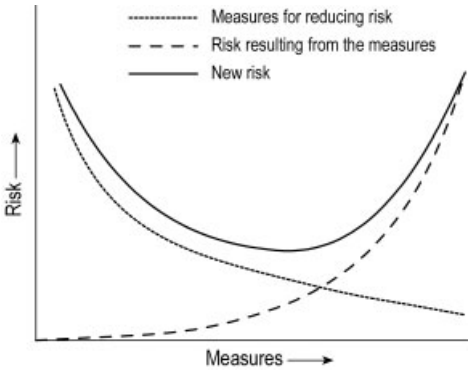

Figure 2. Relationship between risk and the measures taken to reduce it

There is an interesting analogy from the relationship between risk and the level of freedom in a society. The formal approach shows that plotting the level of freedom against the risk gives a curve with a distinct maximum (Fig. 3). This relationship must be interpreted such that only by accepting a certain level of risk can an optimum level of freedom be achieved. Too much risk is just as harmful for freedom as too little. The philosopher JASPERS described these often misunderstood circumstances as follows: "Human things do not permit any absolute safety. Freedom can only assert itself through freedom in risk. Anyone who wants absolute safety wants a lack of freedom and political death."

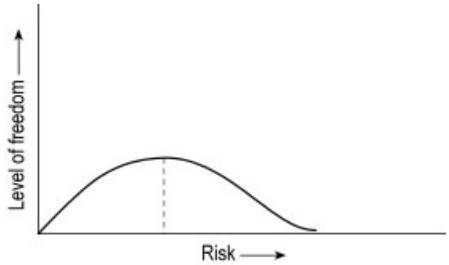

Figure 3. Relationship between risk and the level of freedom

The contemplation of freedom leads to a central, but frequently overlooked, aspect in the discussion of risk: risk infers not only an expectation of loss, but also an expectation of gain. Just as light and shade are connected with one another, so is risk connected with chance. Risk and chance cannot be separated from one another. Society, when considering questions of evaluation in technology, must be aware of this fact and its fate-related significance. For Sustainable Development, risk and chance provide the yardstick for measuring to what extent one of the three fields of ecology, economy, and social matters may be impaired. Risk/chance assessment is difficult, particularly because it poses the ethical question about criteria for whether a risk can be reasonably and responsibly accepted. The answer must be developed in social consensus with the aid of complex, cross-discipline discourse. One-sided approaches, which reveal only the chances or only the risks, are inappropriate. A differentiated view of the risks can also make an indispensable contribution to the concretization of the prevention principle.

4.4. Innovations

Concern about the future has also intensified the demand for appropriate, targeted innovations. They are seen as potential solvers of current problems, as a means to repair the past, and as the answer to ethical questions. It is also hoped

that they will have a long-term effect with regard to Sustainable Development.

Without innovations, human societies have no future. The emphasis on the importance of innovations is, however, not a modern idea. In the past, inventions were made that brought about profound changes in the social system and were connected not only with quantitative, but also with qualitative leaps. Printing, the railways , and fertilizers are just three examples.

In principle, the concept of innovation is very convincing. The actual meaning of the term innovation does, however, need evaluation. Innovation means renewal in the truest sense of the word, and, in the final analysis, creative destruction. In extreme cases, it can go as far as "revolution". In the technical world, a fundamental, far-reaching innovation is therefore often described as a "technical revolution". Innovation therefore means breaking with what currently exists and with traditions. It necessitates a willingness to make profound changes in structures and systems, and frequently unavoidably entails such changes. For humankind, this involves processes of rethinking and adjustment, which are not always desirable. They only receive approval if the change is accepted. This may be brought about either by insight or by suffering.

The call for innovations, however self-evident and logical it may be, must take several factors into account. First, innovations must not be confined to the field of research and development in the natural science segment. A demand for innovation also exists in other areas of science and in society and politics. Second, it is important to take a careful look at their interactions and consequences: Innovations are ambivalent, they conceal not only chances, but also risks. Third, creativity and hence innovation cannot be forced, it requires time and capital investment. Fourth, innovations must be accepted by society and be able to establish themselves in the marketplace. Fifth, innovations necessitate multidimensional interactions and a willingness of those involved to cooperate. Finally, they must not just be short-term solutions, they must take into account, in the long term and to an equal extent, the three fields of ecology, economy, and social matters.

It is important to look at this background when the chemical industry is called on to come up with innovations. That it should be involved is beyond doubt. Nevertheless, important preconditions for creativity include having sufficient space to move and adequate personal responsibility, which are oriented to the underlying social framework that has been established. Apart from the encouragement of innovative ideas, approaches already exist in the chemical industry for achieving Sustainable Development. They make no claim to finality but are parts of an iterative process.

5. Approaches of the Chemical Industry

Industry, and thus also the chemical industry, has adopted the idea of Sustainable Development. This was demonstrated, for example, by joining The Business Charter for Sustainable Development of the International Chamber of Commerce (ICC) [8], which was drawn up on the basis of the Brundtland Report in 1991 and contains 16 principles. Using natural resources sparingly and minimizing the impact of production and products on humans and the environment are two of the principles of the chemical industry. It has integrated Sustainable Development into its strategic targets and will contribute to the lengthy, continually advancing social process within the limits of its possibilities. It strives to intensify cooperation with politics and society.

5.1. Responsible Care

5.1.1. Basic Idea

The idea of Sustainable Development must be concretized by industry- and company-specific programmes and activities. An important contribution to this has come from the worldwide initiative of Responsible Care, which was developed by the chemical industry for the chemical industry (unilateral initiative) and is listed in Agenda 21 as a key step on the road to Sustainable Development. Responsible Care originated in the Canadian Chemical Producers Association. In 1984, this Association drew up guidelines containing the goal and the commitment to voluntarily and continuously improve services

in the fields of health care, safety, and environmental protection (HSE). Responsible Care thus goes beyond the legal requirements and is an initiative which relates to the future. It contains principles and guidelines for acting responsibly with regard to employees, neighbors, the public, and other branches of industry. In the meantime, the national associations of the chemical industry throughout the world have joined the initiative and formulated or adapted their guidelines. For the member companies of the associations, these guidelines are binding.

The concept of Responsible Care contains ten directives [9], which basically correspond to the demands laid down in The Business Charter for Sustainable Development (ICC) [8]. Among the key points in the initiative are the setting up of a suitable management system and the greater integration of individual employees. Their personal responsibility for environmental protection and safety is underlined, and they are called on to strengthen and sharpen their environmental awareness and personal initiative. They should act in line with corporate environmental protection and safety objectives, and supplement them with their own ideas. Apart from the guidelines, the Responsible Care concept also includes management practice codes for the fields of product stewardship, plant safety and hazard prevention, occupational safety, environmental protection, transport safety, and dialogue. The basic principles for these areas are described in Germany, for example, in guidelines issued by the national Chemical Industry Association (VCI) and many member companies [9], [10]. The task of environment management and of every individual is basically to put the guidelines into practice.

To effectively implement the basic aims of Responsible Care, a regular check must be made as to whether the self-imposed targets have been achieved by the measures taken. The aim of this target/actual comparison is to instigate any corrections which may be needed. It is, however, also the expression of a willingness to learn, which is necessary for a lasting goal (see Section 4.2). The workforce and the general public should be regularly informed about the goals and the results. For this purpose, the goals must as far as possible be measurable and understandable.

In summary, the three major elements of Responsible Care are proactive action, participation, and prevention. With Responsible Care, the ethical component has been incorporated into industrial programming. Responsible Care can be seen as a serious endeavour to take up Jonas' responsibility concept (see Chap. 3), and to assume responsibility in the sense of responsibility ethics with its own code of conduct.

The following section examine various examples from individual segments to demonstrate the practical implementation of the Responsible Care idea.

5.1.2. Management Systems

When setting up management systems, the aim is to find forms of organization which point the way to continuous improvement and set and perpetuate targets with personal responsibility. In clearly formal terms, it begins with taking stock of all the data of relevance to environmental protection and safety. From this, targets are set and announced, then measures are drawn up to implement them. After a fixed period of time, a comparison is made between the target and what has actually been achieved, for example in the form of an audit, either performed in-house or by external experts. This may result either in corrections to the original course or the setting of new targets. The formal procedures of the management system, information flows and responsibilities are compiled in a manual.

The preconditions for the success of management systems are the active involvement, motivation and creativity of the employees who are to achieve the improvements. Internal and external communication therefore plays a key role (see Section 5.1.4).

5.1.3. Production-Integrated Environmental Protection

The target aspects of Sustainable Development, namely to conserve resources and reduce the impact on the water, soil, and air, are identical to those of production-integrated environmental protection. Being a preventive form of environmental protection, it aims to use raw materials and energy sparingly, and to avoid, reduce, or recycle emissions of solid waste, wastewater, and waste air. Avoidance has the highest

priority, and disposal the lowest. Production-integrated environmental protection thus contrasts starkly with end-of-pipe environmental protection, which comes at the end of the production chain with subsequent waste disposal. Although this latter kind of industrial environmental protection was very successful in the 1960s and 1970s and made a major contribution to improving the environmental situation, it is nevertheless a kind of repair job, which, for its part, consumes energy and raw materials and generates considerable costs. Over the last few years, a structural change in environmental protection has taken place in the chemical industry away from disposal to preventive measures [11]. Here, the emphasis is quite clearly on production-integrated environmental protection.

Established production processes are being overhauled and new processes developed in accordance with the set targets of avoidance, reduction and recycling. The basic principle is nothing new in the chemical industry: the concept of coupled production – production in which byproducts from a process constitute the starting materials for new products – is extremely important for the chemical industry and has been practised for a long time. The environmental aspect has gradually been added in recent years. The search is for processes which utilize raw materials with a maximum efficiency and have as little impact as possible on the air, water, and soil from the very beginning. This can be achieved in various ways, including, for example, intelligent, innovative procedures in combination with optimum control techniques, the development of higher performance catalysts, and the setting up of closed material circuits.

Every modification to a process is a one-off situation which costs a lot of time and money and which must, in principle, answer questions concerning technical feasibility, economic viability, and ecological benefit. Interdisciplinary cooperation is needed to look at the relevant process from various points of view. Basically, modifications to processes should not be made merely under the aspect of ecology, but must also take other circumstances into account, for example, safety, material, energy, and economic factors. Wherever possible, these must be weighed up against each other. Semi-optimized, short-term solutions frequently create more problems in the long term and do not bring any real increase in efficiency.

Production-integrated environmental protection has already provided considerable impetus in the chemical industry in the past few years, and can look back on a number of successes [9], [11] (see also → Production-Integrated Environmental Protection). It offers chances not only for ecological improvements, but also for enhancing quality and saving costs. Because of these clear advantages over end-of-pipe environmental protection, production-integrated environmental protection will make an important contribution to safeguarding the future. In the context of Responsible Care, it will be consistently pursued with a high level of personal responsibility.

5.1.4. Product Stewardship

The integrated, preventive approach for protecting the environment and humankind also applies to the products of the chemical industry [9], [12], [13]. The aim is minimize the risks emanating from products with respect to their their production, packaging, marketing, handling, use, and disposal. Product stewardship means taking account of the health, safety, and environmental aspects of a product during its entire life cycle. It begins in the field of research and development, continues through production, storage, and marketing, but does not end when the product is used. It also considers its final destination, and includes recycling and disposal. It thus requires cooperation, not only from companies in the chemical industry, contractors and clients, but also from all other parties involved in the chain, starting with procurement of the raw materials and ending with the product's final disposal. This interindustry, intercompany communication allows a comprehensive examination and evaluation of the product, and provides an opportunity to search out and implement potential ecological improvements. Ultimately, it should lead to an improvement in product design.

Despite the overlap aspect, the individual companies and branches of industry must develop their own programmes and incorporate them into their management systems. Product stewardship is an important part of the Responsible Care initiative.

An efficient concept for product stewardship also encompasses risk management, which includes risk/chance assessment. Within this context, it is important to start with comprehensive knowledge of the ecological and toxicological properties and the effects of a substance in order to be able to describe its hazard potential. The actual hazard and the probability of its occurring, that is, the risk, can only be assessed after an analysis of its exposure. Here, an examination is made as to whether and to what extent ecosystems or living creatures are actually exposed to the substance. This is followed by an evaluation. This process must be carried out very carefully, and the chances and risks must be specifically weighed against each other. The process of actually passing judgement is best carried out at the interdisciplinary level. The result is a risk assessment which takes appropriate account of ecological and economic aspects. The next step after the evaluation phase is to deal with concepts and strategies to draw up measures through which risks can be compensated.

The range of tests and evaluations reveals very comprehensive, complex work. Scientific method development is being intensively pursued. Targeted communication between employees, customers, users, and contractors necessitates comprehensive information on the health, environment, and safety aspects of a product. The chemical industry is intensively involved not only in this, but also in the further development of the integrated concept of product stewardship.

5.1.5. Dialogue

To implement the idea of Sustainable Development, dialogue in society as a whole is absolutely essential (see Section 4.2). The chemical industry participates in this to a very high degree [9]. Firstly, it issues a wide variety of written and electronic information, for example, environmental reports, which provide information on recent achievements and developments and on the company's goals, and those of the chemical industry as a whole, in the context of Responsible Care [14].

Secondly, the chemical industry greatly encourages personal dialogue. This extends beyond the traditional exchange of know how and experience among experts, and now covers all sections of society. After all, only through communication between those affected and those involved is it possible to move closer together and, in particular, understand each other. This is an essential requirement in finding a consensus of opinion in situations of conflict. One important point here is that the doubts and misgivings of the population are taken seriously and that they are included in the planning. To improve communication, the chemical industry, for its part, stages a wide variety of events on a general or scientific basis (e.g., press conferences, open days, guided tours). Exhibition and communication centers offer on-site forums for discussion [15], and the Internet also encourages dialogue. Apart from this, the chemical industry actively participates in a wide range of discussions with the general public and with political leaders.

The chemical industry attaches major importance to internal dialogue. Staff motivation, which is so important for the Responsible Care initiative, is based on a consistent supply of information. For this purpose, the chemical industry has set up compulsory, target group-oriented training programmes for the workforce [16]. The aim is to further the knowledge of environmental protection and safety with relevance to the respective plant or department, and also to see it in a company-wide context. Another aim is to strengthen the feeling of personal responsibility. The system is supplemented by in-house newspapers and other written documentation.

5.2. Environmental Agreements

The chemical industry is prepared, together with politicians, to draw up environmental targets and solve recognized problems on its own responsibility and on the basis of agreements. A number of such agreements have been reached in the past few years in the fields of environment, safety, health, consumer protection and energy. These are either bilateral agreements between the industry and the authorities, or unilateral initiatives. Examples relating to the German chemical industry include the voluntary existing chemicals programme, the agreement to halt CFC production, the agreement with the city of Rotterdam to reduce contamination of harbor sludge, and the self-imposed commitment to lower CO_2

emissions. In practice, there are many possibilities for reaching voluntary agreements. They differ in their detail.

As a rule, the agreements are reached for the branches of industry by the relevant associations. In individual cases, the companies can also enter into agreements themselves and thus join in directly. In any case, each agreement has to be worked out individually and necessitates comprehensive dialogue. The actual route to achieving the target is largely left to the companies or branches of industry. Any scope for adjustment and cost benefits may be utilised. In the case of long-term targets, it is advisable to include quantified intermediate targets, in order to assess at a relatively early stage whether the measures taken so far are sufficient for reaching the target or whether corrections have to be made. The agreements include an obligation by the chemical industry to exert a vigorous influence to make sure the agreed aims are reached.

Complying with the agreements is important for the chemical industry. First of all, it offers the security of long-term planning and investment projects. Secondly, the agreements can lead both to ecological and to economic advantages and, in this respect, they also act as an instrument to stimulate the willingness and ability to innovate.

6. References

1. United Nations Conference on Environment and Development: Proposal of the Chairman of the Preparatory Committee for UNCED on the Declaration on Environment and Development, Rio de Janeiro 1992.
2. D. L. Meadows: *The Limits to Growth,* Universe Books, New York 1972.
3. *The New Encyclopedia Britannica*, vol. 6, 15th ed.
4. H. Jonas: *Das Prinzip Verantwortung,* Suhrkamp Taschenbuch 1085, Frankfurt 1984.
5. World Commission on Environment and Development (WCED): *Our Common Future (The Brundtland Report),* Oxford 1987.
6. Umweltpolitik, Konferenz der vereinten Nationen für Umwelt und Entwicklung: Agenda 21, Bundesministerium für Umwelt, Bonn 1992.
7. Enquete-Kommission "Schutz des Menschen und der Umwelt", Deutscher Bundestag: Konzept Nachhaltigkeit, Abschlußbericht, Bonn 1998.
8. International Chamber of Commerce: The Business Charter of Sustainable Development, Publication 210/ 356A, Paris 1991.
9. Verband der Chemischen Industrie: Responsible Care Report, Frankfurt 1997.
10. European Chemical Industry Council (CEFIC): Responsible Care Report, Brüssel 1997.
11. H. Hulpke, U. Müller-Eisen: "The Prevention Principle", *ESPR-Environ. Sci. & Pollut. Res.* **4** (1997) 146 – 153.
12. Verband der Chemischen Industrie, Leitfaden Produktverantwortung, Frankfurt 1996.
13. European Chemical Industry Council: Product Stewardship, Responsible Care for Business, Brüssel 1998.
14. Bayer: Environmental Report, 1997.
15. Bayer: A Bridge to Society, Communication Centre.
16. R. Barfknecht, H. Hulpke, U. Müller-Eisen: "Verantwortliche Mitarbeiter für Umweltschutz," in *Überzeugt vom Umweltschutz,* International Chamber of Commerce, Economica-Verlag, Bonn 1997.

Sutures → Surgical Materials

Ecology and Ecotoxicology

Otto Fränzle, Christian-Albrechts-Universität, Kiel, Federal Republic of Germany [(Sections1.1, 1.2, 2.1 (in part), 2.2, 2.3 (in part), 2.4, 3.1, 3.2 (in part)]

Milan Straškraba, Akademie věd České Republiky, Biomatematická laboratoř, Entomologický institut, České Budějovice, Czech Republic (Sections 2.1.4, 2.3.5, 3.2.3)

Sven Erik Jørgensen, Danmarks Farmaceutiske Højskole, Copenhagen, Denmark (Section 3.3)

1. Ecology and Environment 4
1.1. Humans and Environment 4
1.2. Scope of Ecology and Ecotoxicology 5
2. Principles of Ecology and Environmental Chemistry 7
2.1. Structure and Dynamics of Terrestrial Ecological Systems . . 7
2.1.1. Structure of Communities and Ecosystems 7
2.1.1.1. Description of Community Structure 8
2.1.1.2. Successions 13
2.1.1.3. Patch Dynamics 15
2.1.2. Water Balance of Ecosystems 18
2.1.2.1. Precipitation and Interception 18
2.1.2.2. Infiltration and Surface Runoff . . . 19
2.1.2.3. Deep Drainage 21
2.1.2.4. Soil Water 21
2.1.3. Energy Balance of Ecosystems . . . 21
2.1.3.1. Net Radiation 22
2.1.3.2. Soil Heat Flow 22
2.1.3.3. Sensible Heat Flow in the Lower Atmosphere 22
2.1.3.4. Latent Heat Flux 23
2.1.3.5. Efficiency of Photosynthesis 23
2.1.4. Nutrient Cycles and Productivity of Terrestrial and Aquatic Ecosystems 24
2.1.4.1. Nutrients 24
2.1.4.2. Nutrient Cycling in Ecosystems . . 27
2.1.4.3. Productivity of Ecosystems 30
2.2. Terrestrial, Lotic, Lentic, and Wetland Ecotones 38
2.2.1. Terrestrial and Lotic Ecotones . . . 38
2.2.2. Lentic and Wetland Ecotones 40
2.3. Sensitivity of Ecosystems and Ecotones 40
2.3.1. Different Notions of Community Stability 42
2.3.2. Biodiversity and Stability of Model Communities 42
2.3.3. Stability-Oriented Biodiversity Analyses of Real Communities . . . 45
2.3.4. Nondemographic Measures of Stability 48
2.3.5. Sensitivity of Aquatic Ecosystems . 50
2.3.5.1. Reactions of Aquatic Ecosystems to Stress 51
2.3.5.2. Sensitivity to Different Stressors . . 53
2.3.5.3. Sensitivity to "Global Change" . . . 55
2.3.6. Sensitivity of Ecotones 56
2.3.7. Stability and Resilience of Ecosystems in Evolution 57
2.3.8. Appraisal 59
2.4. Exposure and Effect Criteria of Chemicals 60
2.4.1. Equilibrium Constants for Chemical Distribution 60
2.4.2. Kinetic Constants for Environmental Processes 63
2.4.2.1. Hydrolysis 63
2.4.2.2. Photolysis 64
2.4.2.3. Oxidation and Reduction 65
2.4.2.4. Sorption and Ion Exchange 65
2.4.2.5. Biotransformation 67
2.4.2.6. Chemical Structure and Biodegradation 67
2.4.2.7. Bioconcentration, Bioaccumulation, and Ecological Magnification 69
3. Ecotoxicology 69
3.1. Exposure Assessment 70
3.1.1. Release Estimation 70
3.1.2. Elimination 72
3.1.3. Dispersion 72
3.1.4. Local and Regional Concentration Estimates 73
3.1.4.1. PEC Estimates in the Air Compartment 74
3.1.4.2. PEC Estimates in Soil and Subsoil . 74
3.1.4.3. Variogram Analysis 74
3.1.4.4. Modeling Chemical Distribution in Soil and Subsoil 78
3.1.4.5. PEC Estimates in Aquatic Compartments 79
3.2. Effects Assessment 83

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b07_019

3.2.1. Test Systems for Aquatic Ecosystems 84
3.2.1.1. Level 0 84
3.2.1.2. Level 1 88
3.2.1.3. Level 2 89
3.2.1.4. PNEC Values and Assessment Factors 91
3.2.2. Testing Strategies for Terrestrial Ecosystems 92
3.2.2.1. PNEC Estimates for Soil and Subsoil 92
3.2.2.2. Complex Assessment and Modeling of Vegetation Damage 96
3.2.2.3. Statistical Analysis of Forest Dieback 98
3.2.3. Impact of Xenobiotics on Aquatic Ecosystems 102
3.2.3.1. Approaches and Methods of Studying the Impact of Xenobiotics on Aquatic Life 102
3.2.3.2. Behavior of Xenobiotics in the Aquatic Environment and their Influence on Aquatic Organisms and Indirectly on Humans 104
3.2.3.3. Impact of Xenobiotics on Pelagic and Benthic Biocenoses 106
3.2.3.4. Ecological Risk Assessment of Xenobiotics 109
3.2.4. Testing Approaches to Atmospheric Effects 109
3.2.5. Appraisal 112
3.3. State of the Art of Modeling in Ecotoxicology 116
3.3.1. Characteristics of Ecotoxicological Models 116
3.3.2. Classification of Ecotoxicological Models 117
3.3.3. An Overview: Application of Models in Ecotoxicology 123
3.3.4. Ecotoxicological Processes 125
3.3.5. Parameter Estimation 128
3.3.6. Fugacity Models 130
4. Appendix 134
5. References 134

The entire topic was coordinated by Otto Fränzle

Abbreviations :

a species-specific maximum interception
A constant; evaporating surface
$ALT(t)$ leaf senescence function
b constant; density of vegetation cover; coefficient adjusting evaporation coefficient to other substances than oxygen
B natality
BCF bioconcentration factor
C connectance of the web
C_{dis} concentration at which long-lasting disturbances of essential soil functions occur
C_{eff} concentration at which low-level adverse effects first appear
CHAID chi-square automatic interaction detection
D disturbance; Simpson's diversity index; standardized damage level
2,4-D 2,4-(dichlorophenoxy)acetic acid
DOC dissolved organic carbon
e efficiency; vapor pressure
E evenness (equitability); rate
$E(N)$ photosynthetic efficiency of a particular needle age class
ECETOC European Chemical Industry Ecology and Toxicology Centre
eff specific effect constant
EFF average photosynthetic efficiency
EM ecological magnification (concentration trophic level n/concentration trophic level $n-1$)
$E_{\mathrm{o}}(\lambda)$ scalar irradiance
ES solar radiation
ET evapotranspiration
Ex excretion coefficient
f fugacity
F amount of food uptake per day
F_{m} site factors
$F*$ rate of infiltration
F_{k} correction factor
f_{oc} fraction of organic carbon in the solid being considered
F_{oc} distribution coefficient for organic matter
F_{s} distribution coefficient
GIS geographic information system
GW groundwater
GWP global warming potential
GWR groundwater run-off
h soil water suction; vector (Eq. 52)

H Shannon's diversity index; Henry's law constant
H_a sensible heat transfer through air
HC_p protective hazard concentration
H_s heat flow in soil
i cumulative infiltration at time t; flux
INT total intake of toxic substance per day
IRPTC International Register of Potentially Toxic Chemicals
J rate of irreversible phenomena (e.g., heat flux)
k von Karman's constant; rate constant
k_b $(=\mu_m/Y)$ biodegradation constant
K hydraulic conductivity; rate constant, buffer constant; sediment – water partition coefficient (Eq. 80)
K vertical eddy diffusion coefficient
$K_{a,\lambda}$ specific absorption rate
K_G gas film mass-transfer coefficient
K_L liquid film mass-transfer coefficient; evaporation coefficient (Eq. 89)
K_{OW} n-octanol – water partition coefficient
K_s saturated hydraulic conductivity, half-saturation constant (Fig. 10); concentration of substrate to support half-maximum specific growth rate (Eq. 50)
L specific coefficient (e.g., of thermal conductivity or diffusion)
$L(N)$ leaf mass of particular needle age class
LAI leaf area index
$LEAF$ total leaf mass
m expectation (Eq. 52); adsorbent mass
M mortality
n number of individuals per species; number of molecules adsorbed
n_m number of molecules required for a complete unimolecular surface film
N precipitation; rate of transfer; total number of individuals in the community
$N_i(R)$ net interception
NJ number of needle age classes
N_o above-canopy precipitation
NOEC no-observed-effect concentration
O_n organisms
ODP ozone-depletion potential
p concentration of chemical in solution
P precipitation; rate of absorption (photosynthesis); equilibrium vapor pressure
PCB polychlorinated biphenyl
PCP pentachlorophenol
PEC predicted environmental concentration
p_i proportion of individuals or biomass contributed to the total of the sample
PNEC predicted no-effect concentration
POCP photochemical ozone formation potential
p_o saturation vapor pressure
p_r amount of rain required to pond the soil
q instantaneous value of water vapor concentration; water vapor content at level
QSAR quantitative structure – activity relationship
r resistance factor; increased number of organisms per unit time and unit population (Eq. 81)
r_{ET} transfer rate of latent heat due to evapotranspiration
r_p rainfall intensity
R recovery; universal gas constant
R_c long-wave counterradiation of the lower atmosphere
R_n net radiant flux
R_s surface runoff
s sorptivity; specific leaf surface
S total number of species in the community; sorptivity; nutrient concentration; saturation concentration or solubility at fixed temperature; net source – sink term (Eq. 51); entropy; standardized damage level
$S*$ pondage capacity
SM soil moisture, volumetric soil water content
S_s IR absorption strength in the interval $800-1200\,cm^{-1}$
SW perched water
T safety factor (Eq. 65)
TCO_2 percentage of total CO_2
2,4,5-T 2,4,5-(trichlorophenoxy)acetic acid
TF throughflow
TJ season
To equilibrium concentration in water
TOD theoretical oxygen demand
T_r average residence time due to reactions
u velocity
U wind speed at level
v rate of infiltration (Eq. 11)
v_z vertical flow of water at depth z (Eq. 8)
Vd vapor density
w momentary vertical wind speed
W accumulation; body weight (dry or wet matter)
$W(t)$ accumulated pollutant at time t

x amount of substance adsorbed
X thermodynamic "force" (e.g., gradient of temperature or concentration); biomass per unit volume
Y biomass produced per unit amount of substrate consumed; random function (Eq. 52)
z depth, depth of lowest point of measurement (Eq. 8); elevation above surface (Eq. 21)
Z fugacity capacity
α measure of the effectiveness of Br in ozone depletion with respect to Cl; reflectivity (albedo)
β average interaction strength
β_{ij} effect of species j's density on species i's rate of increase
γ coefficient of absorption of photosynthetically active radiation
ε_λ molar absorptivity at wavelength λ
$\varepsilon\sigma T^4$ long-wave radiation from the earth's surface
μ specific growth rate
μ_m maximum specific growth rate
ϱ constant in Hammett – Taft equation; air density
σ Hammett constant; reflectivity
τ_s atmospheric lifetime
Φ_d quantum yield
Ψ soil water potential

1. Ecology and Environment

1.1. Humans and Environment

In the course of their evolution, humans have exerted an ever-increasing influence on the terrestrial and aquatic environments that ensured their livelihood. In early Paleolithic times, *Homo erectus* was dispersed over large geographical ranges of the Old World continents and had already attempted to circumvent environmental limitations, via culture, in expressing dietary preferences different from the youngest australopithecines and in colonizing both drier and colder environments. Because of their food preferences, these hunters and gatherers probably required larger territories and thus comparatively few Acheuleans could be supported in any one area; consequently, mobility and the periodicity of seasonal activities were considerable [1], [81]. The sites occupied indicate a preference for open, grassy environments with large herds of gregarious herbivores, which means that human influence on ecosystems was by no means different from that of hunting or gathering animals.

Things changed drastically during the tenth millennium B.C., when the first agricultural societies arose. The new subsistence system, based on crop planting, livestock raising, and permanent settlements, had been developed to the point that long-term population increases became possible. Significant ecosystem changes, often associated with disturbances of the hydrological cycle and soil erosion [82], resulted when native floras and faunas began to be replaced by domesticated stock and cultigens. Thus, not only were the largely prevailing agrarian and woodland ecosystems profoundly changed in species composition, but natural recycling processes also constantly lost importance.

In the present millennium, characterized by an unprecedented population increase in the last two centuries, the formerly "isolating" woodlands or uncultivated areas around the agrarian ecosystems and settlements have been constantly reduced, which means that the ecological buffering capacity of these nonagrarian areas, which had such an important role in the intricate ecosystem interplay of the past, no longer exists in densely populated regions of the world. Consequently, both aquatic and terrestrial types of ecosystems are now open to human impact on the regional and continental levels, and present-day analytical knowledge provides ample evidence that even the most remote areas on earth bear more or less marked imprints of human activity.

Among them, chemical imprints have attracted particular attention, because their importance has grown since the middle of the last century with ever-increasing speed, and the development is not likely to change in principle, albeit sectoral differentiations are evident. A comparison of high-volume chemicals illustrates for the United States that petroleum derivatives represent 10 % of the total number of entries in the inventories, but account for 55 % of the total production, whereas inorganics represent both 12 % of the materials and 12 % of the production. Another 6.9 % of the production is due to materials

Table 1. Estimated global release of organic reference chemicals in 10^3 t/a (after [84])

1,2,3-Trichlorobenzene	0.09	1,2-Dichloropropane	15
1,3,5-Trichlorobenzene	0.09	Diethyl phthalate	23.1
1,2,4,5-Tetrachlorobenzene	0.15	Pentachlorophenol	27
4-Chloroaniline	0.16	*p*-*p*′-DDT	28.5
2-Chloroaniline	0.17	Aniline	32.7
1,2-Dichloroethylene	0.3	Dichloroisopropyl ether	45
1,1,2,2-Tetrachloroethane	0.45	Nitrobenzene	46.5
4-Nitroaniline	0.55	Malathion	47.4
Hexachlorobutadiene	0.6	*o*-Xylene	63
N,N′-Dimethylaniline	0.72	1,4-Dichlorobenzene	66
3,4-Dichloroaniline	1.2	Hexachlorobenzene	80
4-Methylphenol	1.2	Tetrachloromethane	84
Dimethoate	1.8	*p*-Xylene	100
1,1,2-Trichloroethane	2.4	Atrazine	111
1,3-Dichlorobenzene	3	Alachlor	123
Dieldrin	3	Phenol	180
Tri-*N*-butyl phosphate	3	Styrene	210
1,2,4-Trichlorobenzene	3.3	Di-*N*-butyl phthalate	230
m-Xylene	3.6	Trichlorofluoromethane	291
Endrin	6	Trichloroethylene	326
Chlorobenzene	6.6	Di-2-ethylhexyl phthalate	402
Dimethyl phthalate	6.6	Tetrachloroethane	470
Aldrin	7.5	1,1,1-Trichloroethane	537
Heptachlor	8.1	1,2-Dichloroethane	547
1,2-Dibromoethane	9	Ethylbenzene	600
Parathion (methyl and ethyl)	9.6	Cyclohexanol	1 200
Chlorpyrifos	10	Toluene	1 800
1,1-Dichloroethane	12	Benzene	11 000
γ-HCH (lindane)	14.1		

that are residues from the processing of ferrous metals. Saturated hydrocarbons are responsible for 6.7 %. Well-defined organic substances are the most numerous, and cause major concern in testing, because they account for about 34 % of the inventory sample, but only 6 % of the total production. Polymers and plastics represent 24 % of the number of materials and 3 % of the total production [83]. Table 1 provides insight into the estimated global release of organic reference chemicals. The figures listed indicate that humans, owing to the accelerated growth of the chemical industry since the 1950s, produce and use many synthetic compounds that, in localized exposure at least, no longer appear to be negligible in comparison to natural fluxes and concentrations. Thus, the ecological role of humans has changed profoundly in little more than a century of their long cultural history, because they are now capable of and responsible for introducing new substances or increasing the concentration of natural ones in the environment to an extent unthought of in the past [2].

Because the products of the chemical industry are also a most essential part of both our economic and our social life, governments and industry alike have become increasingly concerned about the potential, unintended consequences that the use of man-made chemicals could have on both human health and the environment. Therefore, a number of key chemical-producing countries have passed, or are enacting, general substance control legislation (→ Legal Aspects, Chap. 2., → Legal Aspects, Chap. 6., → Legal Aspects, Chap. 7.). Common to all these legal instruments is the preventive aspect (i.e., the notification of chemicals prior to marketing), which entails the presentation of data derived from laboratory investigations together with additional information to permit the evaluation of potential hazards.

1.2. Scope of Ecology and Ecotoxicology

Environments are highly complex systems whose evolution is determined by complicated networks of positive and negative feedback loops. Human interference, particularly when inadvertent, is therefore reflected in an ever-

increasing number of environmental problems that cause major concern in today's societies. To better understand and resolve these problems, knowledge of the basic earth and life sciences unified under a comprehensive ecological perspective is essential, and a consideration of the bewildering number of interrelationships from a generalized system point of view appears indicated.

Ecological thinking began to emerge in the 17th and 18th centuries [3], while the very term "ecology" as derived from the Greek *oikos* (home) was coined and introduced into biological terminology by HAECKEL [4]. Paraphrasing HAECKEL, ecology can be defined as the scientific study of the interactions between organisms and their environment or, more informatively: "Ecology is the scientific study of the interactions that determine the distribution and abundance of organisms" [5]. The environment of an organism consists of all those physical and chemical (abiotic) or biotic (other organisms) factors and phenomena outside the organism that influence it.

At the *level of the organism*, ecology deals with how individuals interact with their biotic and abiotic environments (autecology). At the *population level*, ecology determines the presence or absence of particular species, their abundance or rarity, and trends and fluctuations in their numbers (demecology). Two approaches can be distinguished at this level: one deals first with the attributes of individual organisms and then considers the way in which they combine to define the characteristics of the population. The other deals directly with the characteristics of populations and relates these to relevant aspects of the environment. *Community ecology* deals with the structure of communities and with the pathways and fluxes of energy, water, nutrients, and other chemicals through them. This implies adequate consideration of human influence on nature (e.g., pollution and global warming).

Since ecology must deal explicitly with these three levels of the biological hierarchy and the overwhelming multitude of interactions between organisms and their physical, chemical, and biological environments, it is particularly subtle and complex. Yet ecology also has the distinction of being confronted with uniqueness in a very peculiar way: it must consider millions of different species, each composed of frequently countless numbers of genetically distinct individuals, all living and interacting in a highly varied and ever-changing world. Thus, "the beauty of ecology is that it challenges us to develop an understanding of very basic and apparent problems, in a way that recognizes the uniqueness and complexity of all aspects of nature but seeks patterns and predictions within this complexity rather than being swamped by it" [6].

Ecology of such a comprehensive scope cannot be an easy science but must be a complex interdisciplinary structure grouping several dozens of sciences on the basis of a hierarchically structured system of hypotheses. For instance, one of the major present-day ecological research schemes, the long-term ecosystem research project in the Bornhöved Lake District (Schleswig-Holstein) comprises no less than 25 disciplines, ranging from mathematics to social sciences, around a nucleus of physical, chemical, biological, and earth sciences [85].

Ecotoxicology may be defined as "a science which is concerned with the toxic effects of chemicals and physical agents on living organisms, especially on populations and communities within defined ecosystems; it includes the transfer pathways of those agents and their interactions with the environment" [7]. In formal analogy with the objectives of classical toxicology, which has evolved as an aut- or demecologically oriented discipline, ecotoxicology aims at defining toxic effects in terms of reactions of ecosystems or representative compartments of such systems. In view of the complicated structure of ecosystems and their tremendous variability in time and space, however, particular problems arise. While many experimental approaches exist to predict the hazards of drugs, food additives, and contaminants for humans (→ Toxicology, Chap. 3., → Toxicology, Chap. 5.), which are based on extrapolation from animal or other data, the analogous ecotoxicological procedures are much more complicated and hitherto far less satisfactory [8].

The very crux of the ecotoxicological extrapolation problem is the difficulty that an eventual decision-making process will lead to either "false negatives or false positives" [86]. In this connection, a false negative implies that a chemical that appears acceptable on the basis of laboratory data may cause considerable damage in

practical circumstances. If, for instance, the prediction of possible environmental exposure levels on the basis of standardized laboratory tests turns out to underestimate the actual levels occurring in "real-world" environments, populations of certain species will be much more at risk than expected. Thus, species with mono- or oligomaniacal feeding behaviors (i.e., food specialists), run a much higher risk of bioaccumulation effects if specific chemicals are applied against their favorite or only prey than less specialized predators.

In contrast, the false positive involves rejection of chemicals because laboratory tests indicate intolerable hazards for the environment, while in fact such undesirable effects are not likely to occur. Pesticides may be labeled toxic to fish on the basis of standardized laboratory trials, although this does not necessarily imply that they will always cause mortality in fish when applied in, or in the immediate neighborhood of, their natural aquatic habitat. The framework of tsetse fly control in West Africa amply demonstrated that the potentially fish-toxic insecticides endosulfan and deltamethrin could actually be used without having any noticeable effects on fish [87]. The effective dose rate for killing tsetse flies was minimized by careful selection of the insecticide formulation on the one hand and application methods on the other.

Thus, the hazards of pesticides usually depend on such factors as the amount applied, formulation, methods and times of application, and intensity of use, which are – in their extremely variable specific combinations – virtually impossible to simulate. Attempts must be made, therefore, to consider the ecological characteristics of target areas as well as possible nontarget effects in connection with the use pattern envisaged in order to avoid unrealistic predictions. A compound that appears unacceptable in a certain application in one place need not be so in another. Hence, basic physical – chemical data such as those described in the *Collection of Minimum Pre-Marketing Sets of Data* of the OECD Working Group on Exposure Analysis [88] have to be matched to data on the relevant properties of typified environments into which the compounds are ultimately released. From this particular perspective, the essential structural and dynamic characteristics of both chemicals and ecosystems that are relevant to ecotoxicology are described in the following chapters.

2. Principles of Ecology and Environmental Chemistry

2.1. Structure and Dynamics of Terrestrial Ecological Systems

2.1.1. Structure of Communities and Ecosystems

In a classic paper, TANSLEY [89] defined the term *ecosystem* as "... the whole *system* (in the sense of physics), including not only the organism-complex, but also the whole complex of physical factors forming what we call the environment of the biome – the habitat factors in the widest sense. Though the organisms may claim our primary interest, when we are trying to think fundamentally we cannot separate them from their special environment, with which they form one physical system. It is the systems so formed which, from the point of view of the ecologist, are the basic units of nature on the face of the earth .... These ecosystems, as we may call them, are of the most various kinds and sizes. They form one category of the multitudinous physical systems of the universe which range from the universe as a whole down to the atom."

From the definition of an ecosystem as an arbitrary unit with respect to both its spatial extent and the phenomena considered, ecosystems are often portrayed as food webs or graphs in which "all" the trophic relations between component species are given [9], [10], [90]. Species can also be grouped into a hierarchy of trophic levels or size spectra [11], [91], [92], which facilitates an understanding of the flows of energy and matter through a community and underscores biological differences in interactions within and between levels. Because of the impacts of pollution and environmental change on the living world [2], the concept of stability has also been germane to the argument of how best to depict ecosystems (Section 2.3.1).

Figure 1 illustrates that an ecosystem can be viewed as a composite of the lower levels of organization (i.e., communities, populations, and individuals). Coexisting individuals of a single species possess characteristics such as density,

sex ratio, age – class structure, rates of natality and mortality, and immigration or emigration that are unique to *populations*. The *community* is an assemblage of species populations that occur together in space and time.

In proceeding from the level of physiological and behavioral ecology – dealing primarily with individuals – to the ecology of populations and communities, collective properties such as species diversity, community biomass, and productivity can be identified. Organisms of the same or different species interact with each other in highly diversified processes of mutualism, parasitism, predation, and physical or chemical competition. Thus, the nature of a higher organizational level is obviously more than just the sum of its constituent parts; it is their sum plus the interactions among them. Thus, *emergent properties* appear only when increased complexity (i.e., different levels of organizational structure) is the focus of attention. A primary aim of community and ecosystem ecology is to determine whether repeating patterns in collective and emergent properties exist, even when the differences in the particular species that happen to be assembled together are great. Discussions need not be segregated as ecosystem rather than community topics, since these phenomena depend explicitly on fluxes between living and nonliving components of ecosystems. Thus, for instance, detailed knowledge of the role that communities play in biogeochemical cycling is essential for understanding and adequately combating the effects of acid rain or the increasing levels of atmospheric CO_2 on forests, fields, greenlands, streams, lakes, and other communities.

2.1.1.1. Description of Community Structure

The apparently simplest way to characterize a community is to list its species, but in practice this may become surprisingly difficult, because of taxonomic problems and because usually only a subsample of the organisms in an area can be counted. Thus, the *species richness* of different communities can properly be compared only if investigations are based on the same sample sizes in terms of the area of habitat studied, the time devoted to sampling, or the number of individuals or modules included in the samples. Statistics in general, and geostatistics in particular, are essential to ensure valid sampling (see, e.g., [12], [13], [93]).

The simplest measures that take into account both species richness and abundance or biomass patterns are diversity indices. As examples, the Simpson, McIntosh, and Shannon diversity indices are given:

The Simpson index [94] is

$$D=\left(\sum_{i=1}^{S} p_i\right)^{-1} \tag{1}$$

where D is the diversity index, S the total number of species in the community, and p_i is the proportion of individuals or biomass that it contributes to the total in the sample. An important further specification is expressed by the term evenness (or equitability, E), i.e., the proportion of the maximum possible value D would assume if individuals were completely evenly distributed among species:

$$E=D\cdot D_{\max}^{-1} \tag{2}$$

Evenness assumes a value between 0 and 1.

The McIntosh [95] index is given as

$$D=\left[N-\left(\sum_{i=1}^{n} n_i^2\right)^{\frac{1}{2}}\right]\cdot\left(N-N^{\frac{1}{2}}\right)^{-1} \tag{3}$$

where N is the total number of individuals in the community and n is the number of individuals per species. In Section 2.3.3, the McIntosh index is used in the framework of stability-oriented analyses of biodiversity.

Shannon's diversity index H [96] again depends on an array of p values; thus,

$$H=-\sum p_i \ln p_i \tag{4}$$

and

$$E=H\cdot H_{\max}^{-1}=H\,(\ln S)^{-1} \tag{5}$$

Since an infinite number of logarithmic systems exist, different authors use other logarithms, in particular those with bases 2 and 10, which should be specified when calculating H.

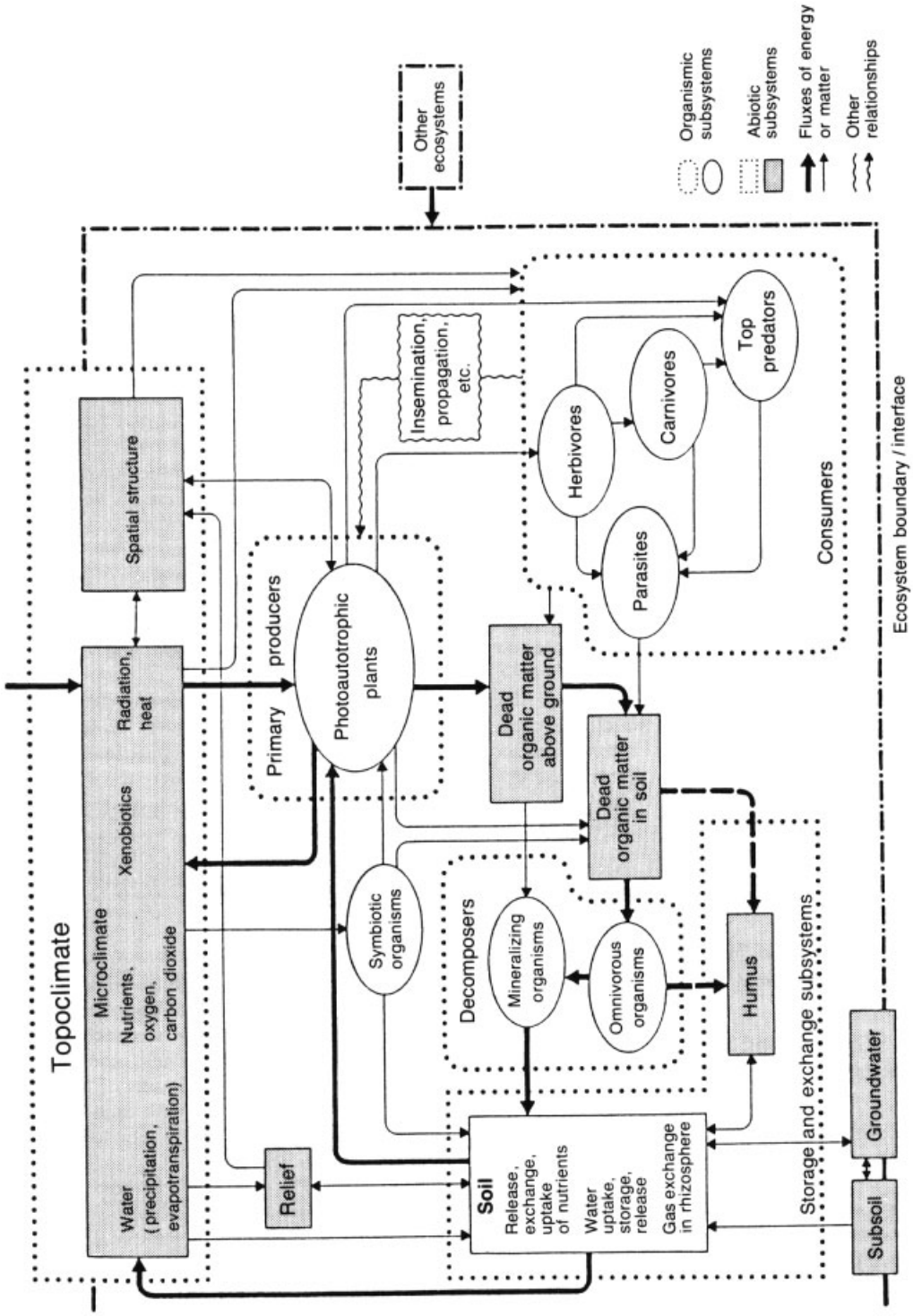

Figure 1. Graph model of an ecosystem (after [90], modified)

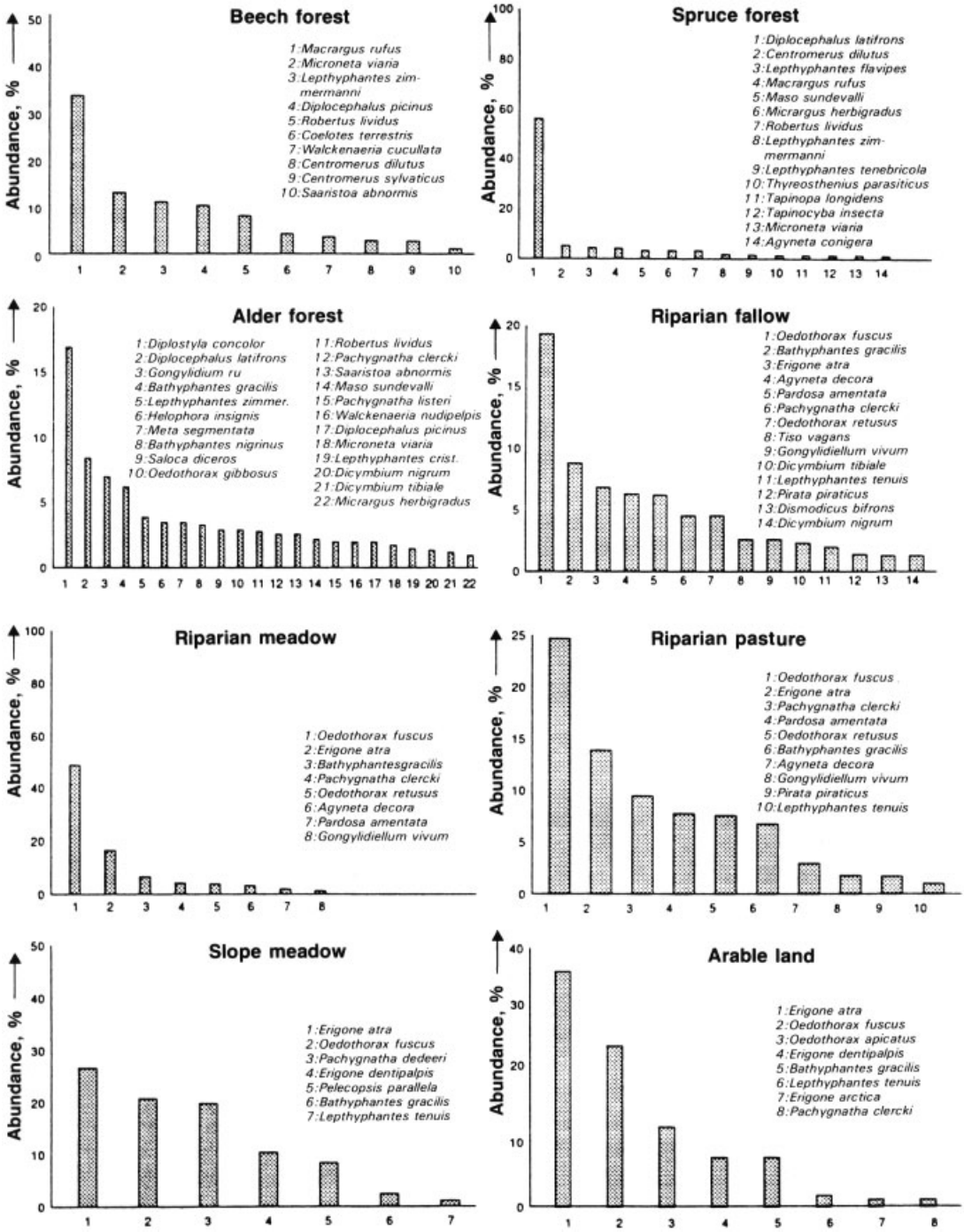

Figure 2. Rank – abundance histograms of epigeic spider communities from various ecosystems of the Bornhöved Lake District (Schleswig-Holstein) (after [85])

Rank – abundance diagrams provide a more complete picture of the distribution of species abundances in a community since they make use of the full array of p_i values by plotting p_i against rank. A rank – abundance diagram can be drawn for the number of individuals, for the area of ground covered by sessile species, or for the biomass contributed to a community by various species (see Fig. 2).

Other classical statistical techniques to reduce the subjectivity of community description are ordination and classification.

Ordination organizes communities on a graph so that those that are most similar in both species composition and relative abundance appear closest together. To this end, the axes of the graph are derived mathematically solely from the species composition of the various stands. Interpretation of the resultant patterns in terms of environmental variables is the second step, in which the scatter of points in the ordination is analyzed to see if the axes correspond to ecologically relevant gradients. Since the latter can be brought about by a host of interactions and a priori determination of the important ones is not possible, the same critical remarks relative to gradient analysis (see below) apply to the interpretation of ordination diagrams.

Classification, in a sense opposed to ordination, produces groups of related communities by a process conceptually related to taxonomic classification. Communities with similar species compositions are grouped together in subsets, and similar subsets may be further combined in a sequential manner by means of clustering algorithms [14], [97]. Figure 3 provides an illustration of such a community analysis (Fig. 3 A) and classification (Fig. 3 B), in which the primary characterization of the individual phytosociological units and the related Collembola faunas was effectuated by means of Renkonen indices.

Gradient analysis and canonical correspondence analysis are uni- and multivariate approaches to detect patterns in communities. In the first case, the investigator searches (sometimes with a considerable amount of subjectivity) for some feature of the environment that appears to matter to the organism. Then, data about the species concerned are organized along a gradient of that factor, which is not necessarily a priori the most appropriate factor. The fact that individual species from a community can be arranged in a sequence along a gradient may imply only that the factor chosen is loosely correlated with whatever really matters in the lives of the species considered. Thus, gradient analysis is a minor step on the way to the objective description of communities.

Canonical Correspondence Analysis. Much better suited to unravel the intricate interplay of environmental factors and to define their relative importance more precisely are multivariate methods (for a comparative analysis, see [98]). The canonical correspondence analysis [99] compares optimum and pessimum curves of species with Gaussian distributions on the basis of multivariate data obtained in the field. Figure 4 illustrates the results of such an analysis for the Collembola fauna of several stands of the Bornhöved Lake District (Schleswig-Holstein). In analogy to biplot analysis described below, in Figure 4 the orthogonal projection of the points marking species on the various vectors (environmental factors) defines the relative importance of these factors for the distribution of the Collembola species explored.

Biplot analysis [100], [101] permits analysis of the inherent structure of comprehensive geographic data matrices, representing, for instance, organisms (O_n) in relation to metrically defined site factors (F_m).

$$[\boldsymbol{M}\,(O_n, F_m)]_{1<n,m} \tag{6}$$

Such an $n \times m$ matrix is equivalent to a cluster in m-dimensional space, where m is the number of factors as related to n organisms considered. This equivalence means that any matrix with $n \times m$ elements may be represented by one vector for a row and another vector for each column such that the elements of the matrix are the inner products of the vectors. When the matrix is of rank 2, or if a higher-rank matrix can be closely approximated by a matrix of this rank by using singular-value decomposition procedures, the vectors may be plotted and the resulting matrix representation (i.e., the biplot), inspected visually. Thus, a biplot not only permits viewing the individual data and their differences, but

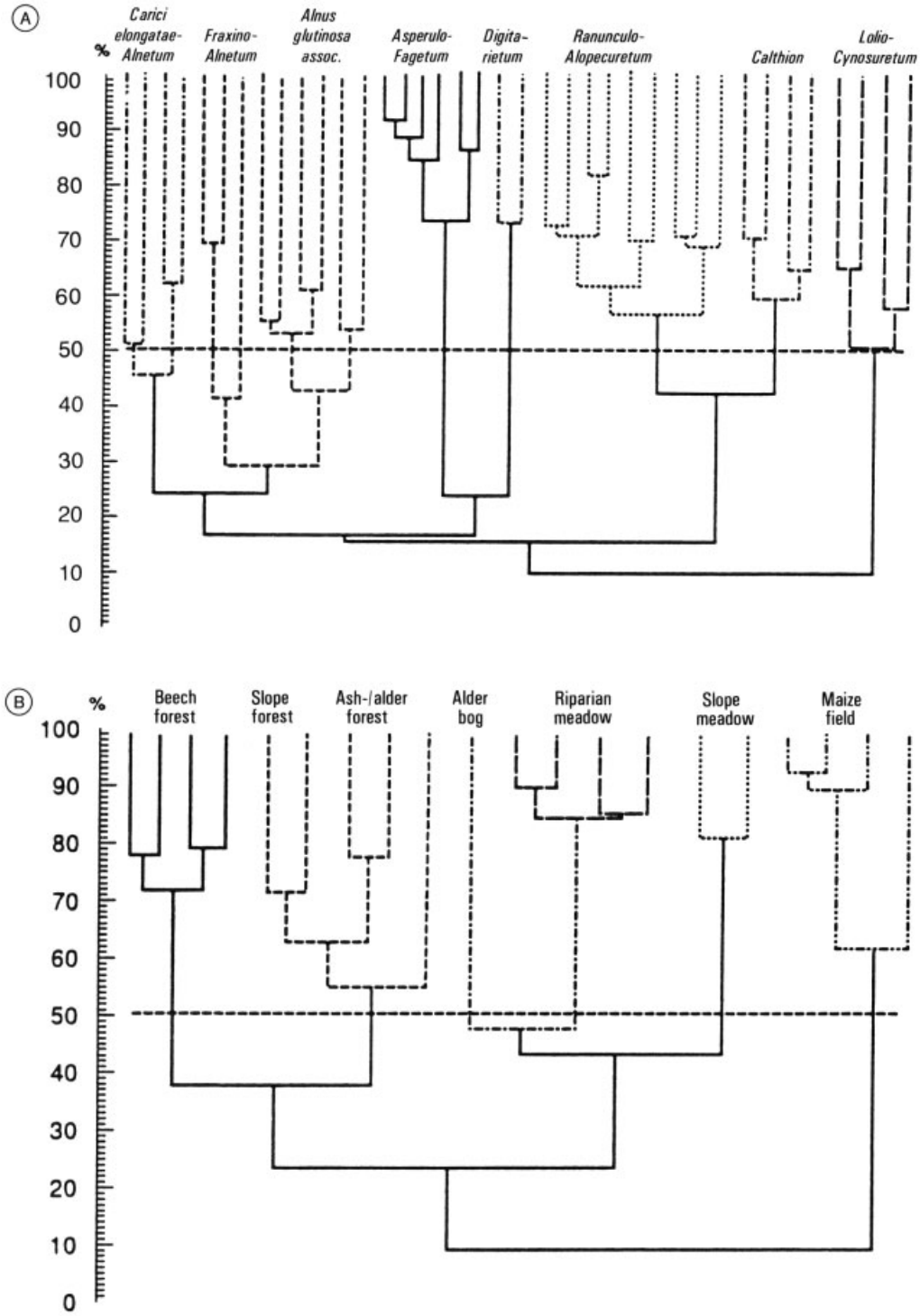

Figure 3. Cluster analysis of plant communities and related Collembola faunas in the Bornhöved Lake District (after [85]) A) Group analysis; B) Formation of subsets and clusters

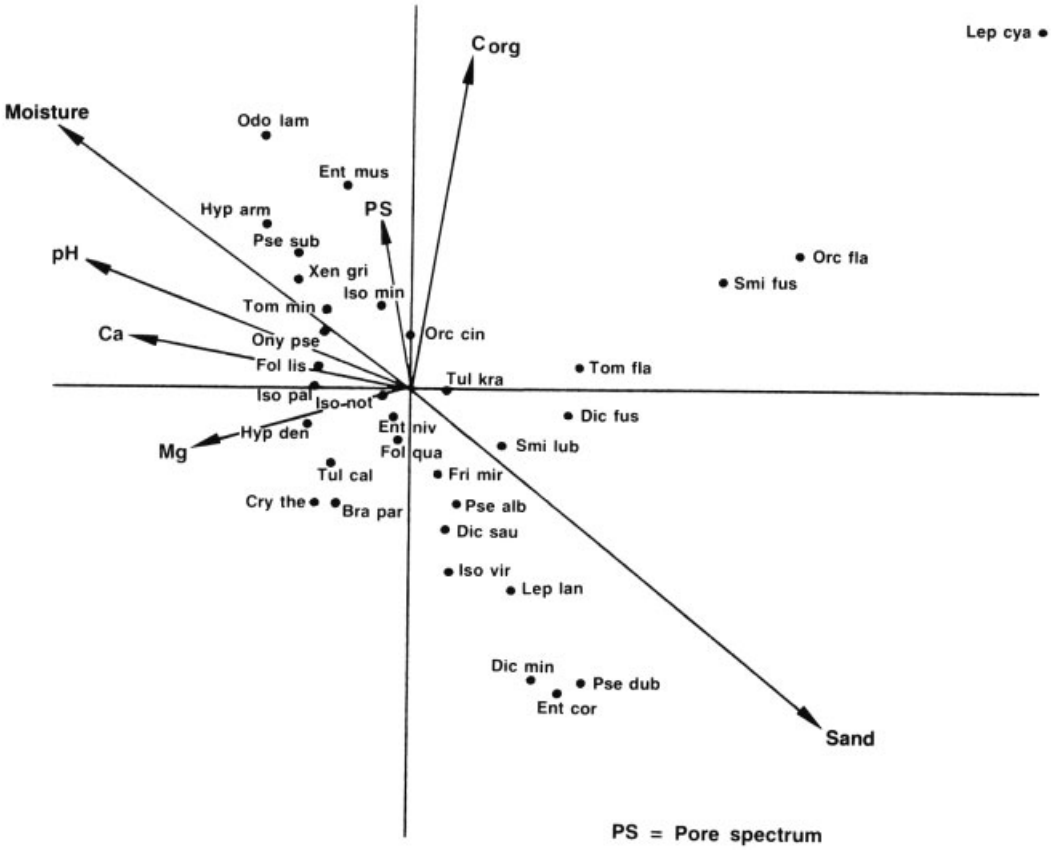

Figure 4. Canonical correspondence analysis of the Collembola faunas of different sites in the Bornhöved Lake District (after [85])
Species: *Brachystomella parvulu, Cryptopygus thermophilus, Dicyrtoma fusca, Dicyrtoma minuta, Dicyrtoma saundersi, Entomobrya corticalis, Entomobrya muscorum, Entomobrya nivalis, Folsomia listeri, Folsomia quadrioculata, Frisea mirabilis, Hypogastrura armata, Hypogastrura denticulata, Isotoma notabilis, Isotoma viridis, Isotomiella minor, Isotomurus palustris, Lepidocyrtes cyaneus, Lepidocyrtes lanuginosus, Odontella lamellata, Onychiurus pseudovanderdrifti, Orchesella cincta, Orchesella flavescens, Pseudachorutes dubius, Pseudachorutes subcrassus, Pseudosinella alba, Sminthurus fuscus, Sminthurus lubbocki, Tomocerus flavescens, Tomocerus minor, Tullbergia callipygos, Tullbergia krausbaueri, Xenylla grisea*

further allows scanning the standardized differences between units and inspecting variances, covariances, and correlations of the variables.

Chi-Square Automatic Interaction Detection. In addition to the above analytical procedures for metric data, chi-square automatic interaction detection (CHAID) is a hierarchically disaggregative classification procedure for nominal and ordinal data [98]. As a multivariable procedure, CHAID subdivides the base set of data automatically in relation to χ^2 values and significance levels into exhaustive subsets. Because of mathematical advantages and the easy interpretability of the resultant dendrogram presentation of data, it has found wide application in forest dieback studies in Germany [102], [103]. Figure 5 illustrates the results of a comprehensive CHAID analysis of environmental factors at 895 sites in North Rhine-Westphalia that proved to be of decisive importance for forest damage.

2.1.1.2. Successions

The relative importance in the community of many organisms (particularly those with relatively short lifetimes) changes with time of year because of seasonal change. In addition, interannual variations in the abundance of species oc-

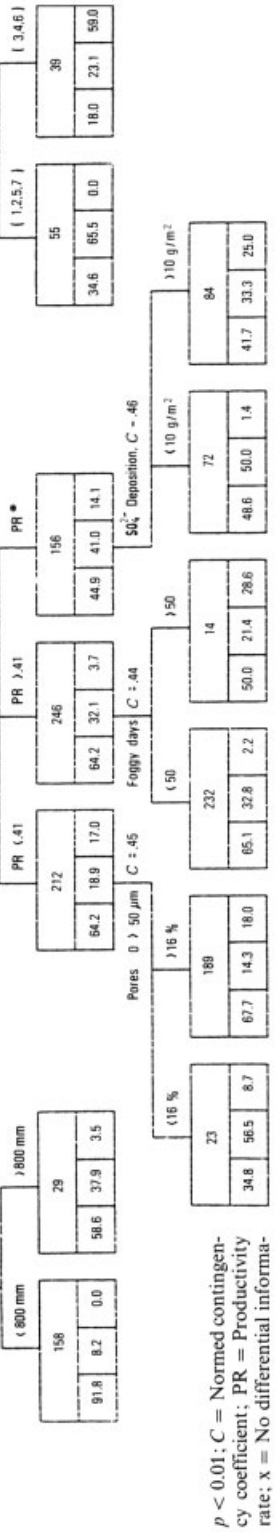

$p < 0.01$; C = Normed contingency coefficient; PR = Productivity rate; x = No differential information

Figure 5. CHAID dendrogram of site factors controlling forest damage in North Rhine-Westphalia (after [103])
$p < 0.01$; C = Normed contingency coefficient; PR = Productivity rate; x = No differential information

cur because individual populations respond to a multitude of environmental factors that influence their reproduction and survival rates. On still higher time scales, different processes of succession occur, which may be defined "as the nonseasonal, directional, and continuous pattern of colonization and extinction on a site by species populations" [6]. On any time scale, successions are of particular ecotoxicological interest since the varying communities normally have different stability, resilience, and consequently sensitivity properties.

The underlying mechanisms of such serial replacements may be quite different. *Degradative successions* usually occur over a relatively short time scale of months or years [104]. Any packet of dead organic matter – for example, litter, the body of an animal, or a fecal deposit – is exploited by microorganisms and detritivores. Normally different species invade and disappear in turn, as degradation of the organic matter uses up part of the resources and makes others available. Ultimately, degradative successions come to an end because the resource is completely metabolized and (partly) mineralized [15], [16].

In addition, autogenic and allogenic successions can be distinguished. *Allogenic successions* occur as a result of changing external environmental conditions (e.g., changes in habitat quality due to erosion or deposition, modifications of soil pH). *Autogenic successions* usually occur on newly exposed landforms, and one of the principal forces driving this type of serial replacement is the change in soil conditions caused by the early colonists. These may alter the availability of resources (e.g., nutrients) in a habitat, in such a way that the entry of new species is facilitated. The process of *facilitation* [105] may be particularly important in primary successions where conditions are initially severe [e.g., on till laid bare after the retreat of a glacier or inland ice (see Section 2.3.7)]. Another mechanism underlying autogenic succession is inhibition – specific modification of the environment by early species that makes it less suitable for recruitment of "late-successional" species.

Connell and Slatyer [105] proposed three models whose fundamental characteristics are summarized in Figure 6. Hemprich [106] modeled the changes in community composition on the basis of comparative phytosociological analyses in the Bornhöved Lakes area. Figure 7 is an example of successions in grassland communities that typically take from several decades to hundreds of years to run their course.

2.1.1.3. Patch Dynamics

Disturbances that open up gaps in woods may be caused by the death of a tree, high winds, lightning, animals, or lumberjacks. Perturbations in grassland communities may be due to frost, to grazing or burrowing animals, and to the feet and dung of grazers. The formation of such gaps of highly variable size is of importance to sensitive or sedentary species with a requirement for open space, whereas it matters comparatively little in the lives of mobile animals for whom space is not a limiting factor.

Viewing the evolution of communities or succession from such a "patch-dynamics" point of view leads to an important class of models consisting of a number of cells that are colonized at random by individuals of different species. Although species interactions within cells proceed according to Lotka – Volterra principles [107], the stochastically cellular structure of the community and the manifold opportunities in each generation for migration between cells allow species-rich communities to evolve and persist. Thus, patch-dynamics models are much more realistic than equilibrium models with resource partitioning and niche differentiation (e.g., [108–110]) or nonequilibrium models incorporating temporal variation in conditions (e.g., [111]), which both deal with closed systems. Patch-dynamics models including competitive interactions and dispersal are of particular interest in an ecotoxicological context. Dhôte and Hollier [112] considered the dynamics of forest stands from two complementary points of view, namely, their intrinsic biological behavior and their management. This involves explicit consideration of different spatial and temporal scales, levels of organization, and model aggregation. Yodzis [113] distinguished two different types of community organization depending on the type of competition between component species. The case in which some species are competitively superior is called "dominance controlled", while the other with

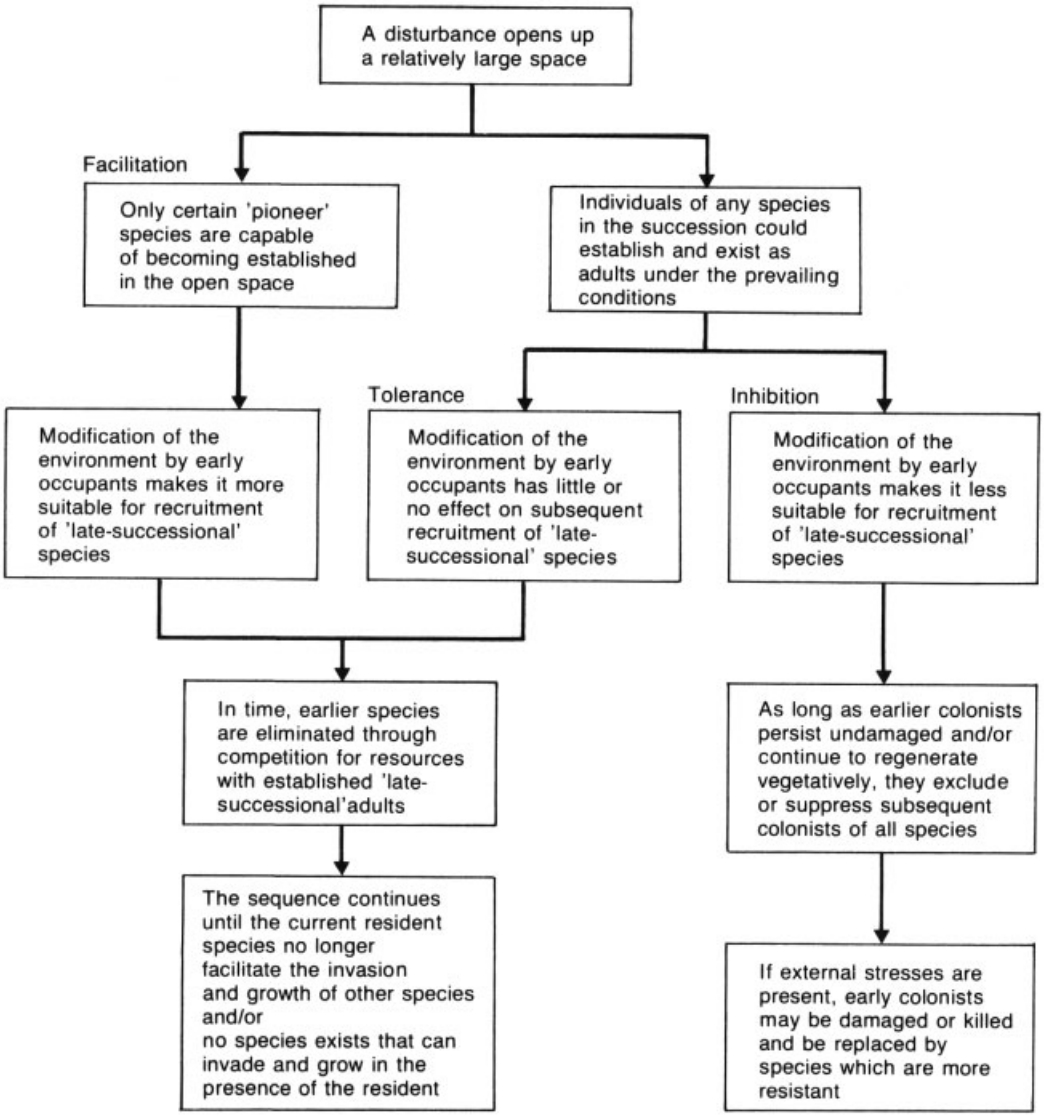

Figure 6. Essential mechanisms underlying terrestrial plant successions (after [105])

all species having similar competitive abilities is "founder controlled."

Dominance-Controlled Communities. In dominance-controlled communities, disturbances that open up gaps lead to relatively predictable species sequences, where early species are good colonizers and fast growers, whereas later species can tolerate lower resource levels and grow in the presence of early species, eventually outcompeting them. Thus, the community regains the climax stage when the most efficient competitors oust their neighbors. Disturbances may be synchronized, or phased, over extensive areas (e.g., bush or forest fires). Other disturbances are much smaller and thus produce a mosaic of patches, particularly if these disturbances are unphased. The resultant climactic macrosystem has a much higher species diversity than an extensive area undisturbed for a very long period and occupied by just a few dominant species.

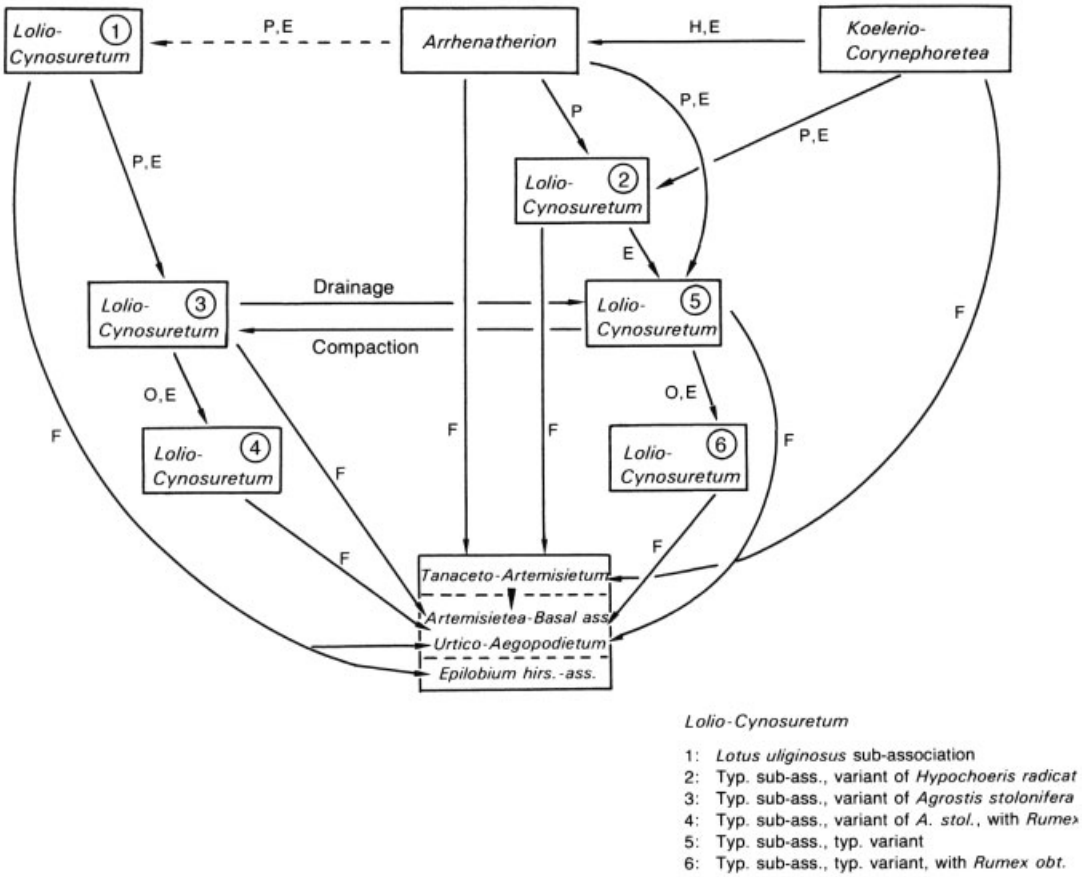

Figure 7. Successions in grassland communities of the Bornhöved Lake District (after [106], modified)
E = Eutrophication; P = Pasture; O = Overgrazing; H = Hay harvest; F = Fallow

Gap size and structure (defined in terms of relief, soil, climate, etc.) are important factors of community structure because they induce contrasting mechanisms of recolonization. The central parts of very large gaps are likely to be colonized by species propagules that can travel greater distances, whereas the smallest gaps may be filled by simple lateral movements of individuals around the periphery. Regrowth results from seeds, plants established prior to disturbance, and lateral growth of branches from surrounding trees. Nearer ground level, colonization is frequently due to saplings that were already present but suppressed [114]. Also, animals can be overrepresented in gaps, as LEVEY [115] has shown in a study of tropical rain forest in Costa Rica. Here, nectarivorous and frugivorous birds were much more abundant in treefall gaps where understory plants tend to produce distinctly more fruit over longer periods than conspecies fruiting under an undisturbed canopy.

Founder-Controlled Communities. Whereas in dominance-controlled communities colonizing ability and competitive status are inversely related, founder-controlled communities are characterized by species that are both good colonists and essentially equal competitors. If a large number of species nearly equivalent in their capacity to invade gaps, and equally tolerant of the abiotic conditions, can "defend" the gaps against all comers, then the probability of competitive exclusion is much reduced in communities where gaps open continually and randomly. On each occasion an

organism dies, a gap is reopened for invasion. Thus, all conceivable replacements are possible and biodiversity will be maintained at a high level. An additional factor is the unpredictability of neighboring individuals in space and time, which contributes to reducing the likelihood of competitive exclusion of one species by another in the cause of succession. In any case, however, numerous feedback relationships exist between local and regional successional dynamics and the energy and water balances of the ecosystems affected.

2.1.2. Water Balance of Ecosystems

A number of models for plant – environment interactions, and particularly for the utilization of water and energy by plants, have been developed since the 1950s. Several of these attempt to understand plant growth and water use as related to specific physiological and environmental parameters [116], [117]. These models, however, cannot be applied to situations for which few data are available unless a number of simplifying assumptions are used. On the regional and geographical levels, other models of a predominantly qualitative character have been suggested [118–120]. As a consequence of this dichotomous development, attempts have been made to unify these two approaches with a view to simplifying the comprehensive models, so as to make them applicable to regional use in areas with limited data, without introducing misleading oversimplifications.

Examples of soil – water balance studies that appear to partly fill in this middle ground are those of Fuchs and Stanhill [121], Slatyer [122], Benecke and Van Der Ploeg [123].

The soil – water balance equation is generally written in the form:

$$P - R_S - (GWR + TF) - ET + \Delta\,(SM + SW + GW) = 0 \tag{7}$$

where P denotes precipitation, R_S the surface runoff, GWR the groundwater runoff, TF the throughflow, ET evapotranspiration, SM soil moisture, SW the perched water (above a compact horizon), and GW groundwater. For the sake of greater convenience, $(GWR + TF)$ is termed deep drainage, and $\Delta\,(SM + SW + GW)$ is the change in soil water storage (initial minus final) during the period and for the depth of measurement. For a number of cases, deep drainage may be adequately defined as the amount of water passing beyond the root zone or, for experimental purposes, as the amount passing below the lowest point of measurement. All symbols have dimensions of length, for example, millimeters. Equation (7) can be written more specifically as

$$\int_{t_1}^{t_2} [(P - R_S) - (ET) - v_z]\,\mathrm{d}t = \int_{t_1}^{t_2}\int_{0}^{z} \frac{\partial\,(SM)}{\partial t}\,\mathrm{d}z\mathrm{d}t \tag{8}$$

where $(t_2 - t_1)$ is the time interval over which the measurements are made (s), z is the depth to the lowest point of measurement (cm), v_z is the net downward flux of water at depth z (cm/s), and SM is the soil moisture, i.e., the volumetric soil water content (cm^3 of water per cm^3 of soil). P, R_S, and ET are in units of mm/s or $g\,cm^{-2}\,s^{-1}$. In the absence of a water table near the surface, v_z is generally positive. Compared to aboveground measurements of estimates of the water vapor flux, the soil – water balance approach has the advantages of ease of data processing and integration, since the soil – water reservoir (SM) automatically integrates extraction rates between observations. The disadvantages are associated largely with a somewhat lower level of measurement accuracy and the difficulty of adequately assessing evapotranspiration during periods of rainy weather. Therefore, its applicability is, to a considerable degree, restricted to regions of relatively high potential evaporation rates and sufficiently well-defined alternations of rainy and dry weather.

2.1.2.1. Precipitation and Interception

The measurement of precipitation N at a site or in a region is generally considered simpler and more straightforward than that of the other terms in the water balance equation. The technical requirements are comprehensively described by Von Hoyningen-Huene and Nasdalack [124] and by Sevruk [125]. The identification

of unequivocal spatial structures should be controlled by means of variogram analysis, which is particularly important in vegetation stands (see Section 3.1.4.3).

Here, marked differences in the pattern of precipitation actually reaching the ground normally develop in many plant communities because of the gross interception of precipitation by the vegetation. Subsequently, precipitation is partly transferred to the soil by channeling down the main streams ("stemflow") and partly by dripping from branches, twigs, and foliage ("canopy leaching"), or it may be lost by evaporation from the wet surface. This latter proportion constitutes the term "net interception." Further differences in the amount of precipitation reaching the ground between plants ("throughfall") are due particularly to the disturbed wind structure and are most noticeable in the case of snow.

In view of the complicated physical nature of net interception, attempts to assess it by means of measurement and by indirect estimations are quite numerous. In the framework of a comprehensive agroclimatological model, BRADEN [126] developed the following interception estimate:

$$N_i\,(R) = a \cdot LAI \cdot \left(1 - \frac{1}{1 + N_0 b / aLAI}\right) \tag{9}$$

where (in the author's notation) $N_i(R)$ is the net interception; $a \cdot LAI$ the saturation parameter, dependent on leaf area index (LAI) and a species-specific maximum interception a; N_0 the above-canopy precipitation; and b the density of vegetation cover. The BRADEN approach has the double advantage of mathematical simplicity and physical foundation, and its validity has been widely tested [127].

With rain, drip or stemflow is usually observed after an area rain total of about 2 mm has been received. However, with freezing rain or snow under conditions favoring retention on the leaves, twigs, branches, and stems (i.e., low wind, temperatures a few degrees below freezing), several times this amount may be accumulated. Stemflow is enhanced by a smooth bark and by branches and leaves that are inclined upwards. Thus, in deciduous (beech) and evergreen (spruce) forests the amounts vary considerably, beech providing much more stemflow and much less interception loss, while oak would have an intermediate position (for greater details, see [2]).

2.1.2.2. Infiltration and Surface Runoff

Horton Overland Flow. Surficial runoff in the sense of Horton overland flow occurs whenever the rate of effective precipitation (i.e., precipitation net interception) exceeds the rate of infiltration ($F*$) and the resultant accumulation of surface water exceeds the pondage capacity ($S*$) at the point of measurement. The most important regulator is $F*$. It is useful in applied hydrology and in relation to pollutant transport to characterize the dynamics of infiltration by a small number of parameters. PHILIP developed a simple physical model of infiltration, which is, however, closely related to more precise diffusion descriptions of infiltration [128].

$$i = St^{0.5} + At \tag{10}$$

where i is the cumulative infiltration at time t, and the constants S ("sorptivity") and A have a physical meaning related to the diffusion analysis of infiltration. The first term on the right-hand side of Equation (10) describes the contribution to infiltration due to capillarity; the second term represents mainly the contribution due to gravity. The differential form of Equation (10) is Equation (11), where v is the rate of infiltration (cm/s):

$$v = \frac{1}{2} St^{-1/2} + A \tag{11}$$

Some values of the infiltration rate (mm/h) for particular soils in specified conditions are given below:

Sand, loess, silt	11 – 7
Sandy loam	7 – 4
Clayey loam, soils poor in organic matter	4 – 1
Clays, alkaline soils (solonets)	<1

These values are much lower than those determined by means of a field rain simulator in the Schönbuch Nature Reserve (Baden-Württemberg), where the use of buffered cylinder infiltrometers proved impossible because of the extremely high spatial variability of the infiltration-relevant soil characteristics. The experimental conditions were such that a constant surface runoff rate was produced by a

constant amount of rainfall when the soil had reached its saturation point. Since evaporation is negligible during the short duration of the experiment in comparison to the rainfall applied (100 – 250 mm/h) and the increment in soil and water storage is assumed to be zero, the difference between precipitation and surface runoff may be equated to infiltration [129]. For two clay soils tested, the minimum infiltration capacity amounted to 58 and 60 mm/h; for sandy soils, 79 mm/h was measured, while the infiltration rates of loamy soils varied between 63 and 76 mm/h. These results are, on the one hand, indicative of a marked enhancement of infiltration due to organic matter and, in particular, desiccation cracks and root voids; on the other hand, they point to high throughflow rates close to the surface (e.g., piping). As a consequence, no differences in runoff characteristics could generally be found on soils in coniferous, broadleaf, or mixed forests. Variation in the runoff rates was high when the soils were initially dry, but low when the soils were wet. A comparison of these minimum infiltration rates with the maximum net precipitation rates recorded within the last hundred years leads to the conclusion that surface runoff is an exceedingly rare phenomenon in temperate forests due to preferential seepage.

Horton overland flow appears to be a common process in semiarid and arid regions, where precipitation intensities are high and the infiltration capacity of the sparsely vegetated soil is low. It is further caused or intensified by the development of a crust on the soils because the surface layer becomes compacted and the pores blocked as a result of the redistribution of soil particles following raindrop impact. Crust formation due to lateral iron translocation is a particularly widespread phenomenon in ferric luvisols of the semihumid tropics, where it largely contributes to enhance pediplanation processes.

Saturated Overland Flow. In temperate environments with normally modest precipitation rates and well-structured soils, Horton overland flow is the exception rather than the rule, except under certain conditions of cultivation and when the ground is frozen. In temperate environments, all the pore spaces may become filled with water after a period of prolonged rainfall, thus saturating the soil. At this point the water table has risen to the surface and the effective infiltration capacity is consequently reduced to zero. Subsequent rainfall runs off directly across the surface of the slope as saturated (or saturation) overland flow. This situation is likely to come about toward the base of a slope or in microtopographical depressions on a slope where both local infiltrations and throughflow received from higher up the slope contribute to soil moisture. Smith and Parlange [130] describe simple relationships that enable saturating or ponding times to be estimated from values of saturated hydraulic conductivity K_s and sorptivity s. In Morocco Imeson [131] found that the amount of rain required to pond the soil p_r could be estimated reasonably well with one of these equations, namely,

$$\int_0^{t_{p_r}} p_r \, dt = \frac{A}{K_s} \ln \frac{r_p}{r_p - K_s} \tag{12}$$

where A is 0.5 s^2 and r_p is the rainfall intensity.

Because of the number of boundary conditions operative in runoff, the latter varies considerably with the amount, intensity, and duration of precipitation, as well as with slope configuration and soil fabric, which determine the degree and extent to which pondage can occur [129], [132], [133]. In natural situations, the slope is rarely constant and, while runoff tends to reduce soil water recharge at the top of the slope and increase it at the bottom, minor changes of slope generally modify the slope – runoff interrelation. Table 2 provides some runoff figures from two sites in the Federal Republic of Germany. Bochow is representative of loamy luvisols developed from Weichselian till, while Müncheberg is characterized by a podzolized cambisol on coversand overlying Saalian till.

Table 2. Hydrometeorological characteristics of two ploughland sites in northeastern Germany for 1961 – 1965 (after [2])

Characteristic	Müncheberg	Bochow
Number of erosion-inducing showers	30	39
Precipitation, mm		
Median value	14.8	11.8
Mean extreme value	23.6	18.4
Absolute extreme value	34.8	25.2
Runoff, mm		
Median value	1.2	2.1
Mean extreme value	3.3	5.1
Absolute extreme value	10.2	8.1

2.1.2.3. Deep Drainage

The deep-drainage term in balance Equation (7) comprises throughflow and groundwater flow and can be equated to a vertical flow that, in turn, may be calculated from hydraulic conductivity and soil water potential data. The normal equation for vertical flow of water v_z is

$$v_z = K + K\frac{\delta h}{\delta z} \tag{13}$$

where K is the hydraulic conductivity (cm/s) and $(\delta h/\delta z)$ is the rate of change of soil water suction (h in cm) with depth (z in cm). Soil water suction is derived from the soil water potential ψ (dyn/cm^2) by the relationship $h = -\psi/\varrho_W g$, where ϱ_W is the density of water and g the vertical acceleration due to gravity. Unless h is very small, $\delta h/\delta z$ is usually much greater than unity so the K term in Equation (13) is often negligible [122]. Under these circumstances, the deep-drainage term of the water balance equation is given by

$$GWR + TF = \int_{t_1}^{t_2} v_z \cdot \mathrm{d}t \tag{14}$$

where $(t_2 - t_1)$ is the time between observations [134], [135].

In other situations, a net upward flux of soil water into the root zones can occur from wetter underlying soil horizons or, in particular, from a water table closer to the surface. For comprehensive reviews of methods available for the determination of deep drainage with a particular emphasis on groundwater recharge, see [136–140].

Under certain soil conditions, diffuse water movement through the intergranular pore spaces and voids may be supplemented by concentrated turbulent throughflow in networks of pipes. These result from large voids that exist in many soils and are enlarged by soil fauna (e.g., mice, rats, hamsters, moles, weasels, ground squirrels) and the growth and decay of roots. Frequently, soil pipes develop at the interface between organic soil and the underlying mineral soil. Discharge in completely filled pipes varies depending on pressure and gravity potentials, and in partly filled pipes in response to the gradient of the water surface. Usually, therefore, pipe flow velocity is much more rapid than that of matrix flow. Table 3 quotes estimates of flow velocities. From these figures, pipe flow may be seen to attain considerable importance in chemical transport, although the distances covered are normally small in comparison to channel flow. Piping is perhaps more strongly associated with semiarid areas than with humid regions. Drainage and slope development in badlands all over the world are frequently dominated by piping.

Table 3. Flow velocities along different routes in a catchment (after [141])

Type of flow	Flow route	Velocity, m/h
Surface	channel flow	300 – 10 000
	overland flow	50 – 500
Soil flow	pipe flow	50 – 500
	matrix throughflow	0.005 – 0.3
Groundwater flow	limestone (jointed)	10 – 500
	sandstone	0.001 – 10
	shale	10^{-8} – 1

2.1.2.4. Soil Water

Measurement of changes in soil water storage is conducted most accurately by the use of weighing or hydraulic lysimeters, provided they are properly designed and sited [142]. Lysimeters cannot be used, however, when the nature of the species composition, the spatial structure of the vegetation cover, the depth and ramification of the root system, or other factors make it impossible to simulate the natural environment inside the lysimeter itself. In such cases, determinations of changes in soil water storage at different points in the plant community provide the only technique for evaluating $\Delta(SM + SW + GW)$. Soil water sensing equipment may still fall short of operator requirements, although marked advances have been made since the 1980s. Probably the most commonly used techniques, at the present time, are those of neutron moderation and tensiometers or tensiographs.

2.1.3. Energy Balance of Ecosystems

The disposition of radiant energy at the surface of the earth is of prime importance for

understanding soil – water balances and the related chemical transport and transformation processes. A formalized transcription of the energetic relationships yields the following form of the energy balance equation:

$$R_n + H_s + H_a + r_{ET} = 0 \tag{15}$$

In this equation, which summarizes the solar energy cascade with its numerous regulation and transformation components, any form of energy that is flowing toward the surface is considered positive and any form that is moving away from it is considered negative. R_n is the *net radiant flux*, H_S is the *heat flow* in soil at the surface, H_a is the *sensible heat* transfer through air at the surface, and r_{ET} is the transfer rate of *latent heat* due to evapotranspiration. Conventionally, all terms are expressed in joules per minute and square centimeters; $4.19\,\mathrm{J\,cm^{-2}\,min^{-1}}$ equals ca. 1 mm of water depth evaporated per hour.

2.1.3.1. Net Radiation

The earth's surface receives short-wave and long-wave radiation from the sun – some of which is lost through reflection – and long-wave counterradiation from the atmosphere, while itself emitting long-wave radiation. Whereas photochemical reactions are controlled mostly by short-wave radiation, the diurnal variability of surface temperatures and the related energy fluxes are very much under the influence of long-wave radiation. Thus, the net radiant flux R_n may be derived from the short- and long-wave radiation balances:

$$R_n = ES\,(1-\alpha) + R_c - \varepsilon\sigma T^4 \tag{16}$$

where the solar radiation ES is the sum of direct beam radiation ES_S and diffuse radiation ES_D, each term comprising short-wave and solar long-wave components; R_c is the long-wave counterradiation of the lower atmosphere; $\varepsilon\,\sigma\,T^4$ the long-wave radiation from the earth's surface (Stefan – Boltzmann radiation); and α the reflectivity. Regionally, the net radiant flux of the earth's surface varies considerably with a general trend toward a positive balance or surplus in low latitudes and a negative balance or deficit in high latitudes (for details, see [17]). Factors such as cloudiness and atmospheric humidity, which in conjunction with aerosols affect the transmission of both solar and terrestrial radiation, and the radiative properties of the soil and vegetation canopy create the spatial variation in radiation balance over the earth's surface.

2.1.3.2. Soil Heat Flow

The rate at which heat is transferred downwards into the soil and subsurface substrate (H_s) is directly related to the nature and efficiency of the distribution mechanisms. In solids, heat is redistributed by conduction, and the flow rate depends on *thermal conductivity*. Units are watts per meter per degree Celsius ($\mathrm{W\,m^{-1}\,{}^\circ C^{-1}}$) or joules per centimeter per second per degree Celsius ($\mathrm{J\,cm^{-1}\,s^{-1}\,{}^\circ C^{-1}}$). Under steady-state conditions, the flow of heat ($\partial Q/\partial t$) is related to the temperature gradient ($\partial\Theta/\partial x$) and the thermal conductivity λ by

$$H_s = \frac{\partial Q}{\partial t} = \lambda\frac{\partial\Theta}{\partial x} \tag{17}$$

A major problem in soils is that steady-state conditions are rarely achieved and thermal conductivity is a complicated function of granulometry, mineralogical composition, compaction, and water content. The conventional determination of thermal conductivity is consequently fairly difficult. An alternative parameter, *thermal diffusivity*, is used, which is given by $\lambda/(c\varrho)$, where c denotes the specific heat and ϱ the density. Units are square meters per second. For a homogeneous medium, thermal diffusivity defines the rate at which temperature changes ($\partial\Theta/\partial t$) take place:

$$\frac{\partial\Theta}{\partial t} = \frac{\lambda}{c\varrho}\cdot\frac{\partial^2\Theta}{\partial x^2} \tag{18}$$

2.1.3.3. Sensible Heat Flow in the Lower Atmosphere

The magnitude of the heat flow into the air at the surface H_a can be obtained by a difference when the other components of the energy balance in Equation (16) are measured, or it can be determined directly. Among the latter approaches, two methods, the aerodynamic method and the

Sverdrup – Albrecht method merit particular attention.

The *aerodynamic method* is based on very precise determinations of the vertical temperature and wind profiles, and involves a horizontal homogeneity of the surface over considerable distances in the luff of the measuring station.

The *Sverdrup – Albrecht method* determines either H_a or ET as components of the energy balance in Equation (19)

$$H_a = \frac{H_a}{ET}(R_n - H_s) \cdot \left(1 + \frac{H_a}{ET}\right)^{-1} \tag{19}$$

The ratio H_a/ET (Bowen ratio) can be estimated from measurements of the vertical gradients of temperature ($\Delta\Theta/\Delta z$) and vapor pressure ($\Delta e/\Delta z$) above the surface.

2.1.3.4. Latent Heat Flux

The latent heat flux can be estimated as a result of the vaporization or condensation of water. Consequently, five general methods are used to evaluate the water vapor flux caused by evapotranspiration.

Analogous to Equation (19), the *Sverdrup – Albrecht method* may be used to determine ET:

$$ET = (R_n - H_s)\left(1 + \frac{H_a}{ET}\right) \tag{20}$$

Thornthwaite and Holzman [143] suggested

$$ET = \frac{\varrho k (U_2 - U_1)(q_1 - q_2)}{[\ln(z_1/z_2)]^2} \tag{21}$$

where q is the average water vapor content, ϱ the air density (g/cm^3), U the average wind speed (cm/s), z the average elevation above surface (cm), and k von Karman's constant (0.4).

A *continuous record of water vapor convection by means of eddy-correlation techniques* as described by Dyer [144] yields estimates of ET by

$$ET = \frac{1}{t}\int_0^t \varrho w q \, dt \tag{22}$$

where ϱ denotes air density, w the momentary vertical wind speed, and q the instantaneous value of water vapor concentration. The problem of measuring ET as the time average of the vertical flux of water vapor thus becomes one of designing instrumentation capable of measuring the turbulent air motion and structure of q. This involves equipment whose response time is sufficiently short to take account of all frequencies in the turbulent spectrum contributing significantly to the flux. This requirement depends on both the height of measurement and the stability of the air, which means, in general terms, that the sensing elements should respond adequately to signals with a period of 1 s.

If all terms of the energy budget are known except the flux due to evaporation (or evapotranspiration), the latter can be obtained by a *difference*. Finally the latent heat flux can be found by *direct measurement*, i.e., by means of weighing lysimeters on the local scale or as the difference term of the water budget of a drainage system on the regional scale.

2.1.3.5. Efficiency of Photosynthesis

In comparison to the above components of the energy balance summarized in Equation (15), the absorption of photosynthetically active radiation in the plant cover plays a quantitatively minor role. Yet, it furnishes useful information on the amounts of solar energy available for photochemical reactions at the surface of soil and vegetation under natural conditions. A quantitative description of the absorption process is provided by

$$dP = \gamma (ES)_v \, s \, dz, \tag{23}$$

where P is the rate of absorption, γ is the coefficient of absorption of photosynthetically active radiation (i.e., wavelength ranges from 400 – 480 nm to 640 – 690 nm); $(ES)_v$ is the incident direct beam and diffuse short-wave radiation; s is the specific leaf surface (i.e., the surface of leaves in a unit volume); and z is the vertical coordinate counted upward [18].

According to Monteith and Szeicz [145], the daily energy rates required for gross assimilation are about 28.5, for respiration 6.7, and for net assimilation 21.3 J cm^{-2} d^{-1}. This means that normally the photosynthetic efficiency of the vegetation cover is in the order of magnitude of 1 – 2 % of the total incident radiation. These

findings were corroborated by numerous authors. Examples of studies from different parts of Central Europe are, for instance, [19] and [146]. From the purely quantitative budgetary point of view this may appear little, but with regard to bioproductivity and photochemical transformation reactions the situation is quite different. A number of compounds (e.g., various chlorinated aromatics) show distinctly higher conversion rates if adsorbed on particulate matter than if deposited as solids or on thin films [147]. This can be attributed to a bathochromic shift, changes in the relative extinction, the appearance of new absorption bands as a consequence of adsorption, and finally fixation of the dispersed molecules in the adsorbed phase resulting in an intensified compound – oxygen contact.

2.1.4. Nutrient Cycles and Productivity of Terrestrial and Aquatic Ecosystems

Speaking about nutrient cycles means considering how the same atoms of elements move through different components of nature or different biota, and participate in different kinds of reactions (Fig. 8). In principle, these are always chemical reactions; however, the conditions driving them vary highly if embedded in biochemical, biological, or ecological structures. A typical feature of the organic world is the hierarchy constructed of levels of increasing complexity (Fig. 9). In such a hierarchy, the control of processes is both bottom-up, from the low levels to high ones, and top-down, from the higher to lower levels. Biological reactions are controlled by several mechanical, chemical, and neurological mechanisms, which depend on the internal and external environment of biota. An example is the formation of urine in kidneys. In ecological reactions, the conditions for biological, biochemical, and chemical reactions are driven by population cycles of organisms, by mutual reactions of the environment and biota, and by interactions among biota themselves.

Biological production is based on the uptake of nutrients, but external conditions such as light, temperature, and oxygen have major effects. Production is the output of organic matter formation by plants, the primary producers, and of transformation by other groups of biota. Biomass is the organic mass of organisms, which can be expressed as fresh weight, as dry weight, as protein content, or in energy equivalents. Biomass is the standing stock, which is not necessarily identical with production. Deciduous forests in temperate regions, for example, produce yearly the biomass of leaves that then decay in fall. The next year, a new leaf biomass is formed.

In the following, nutrients and biological production are described separately for the sake of simplicity. However, they are shown to be inseparably interrelated. A distinction is made between terrestrial ecosystems (those inhabiting land) and aquatic ecosystems (those inhabiting water).

The effect of human activities, particularly of industry, on the environment is enormous. Humans are at present the dominant driving force on earth. They must take care that this force is not misused and is not driving humanity to extinction.

2.1.4.1. Nutrients

Nutrients are defined as those chemical elements that are essential for the construction of organisms. Depending on the proportion of the organism biomass that they represent, nutrients are usually divided into two classes: (1) macronutrients and (2) micronutrients or trace elements. *Macronutrients* represent major components of the biological material and are needed in large amounts. Usually, C, N, P, K, S, Ca, and Mg are considered macronutrients. *Micronutrients* are needed only in small amounts; micronutrients include Si, Mo, Cu, Zn, Mn, Fe, B, and Cl. The needs of organisms for elements belonging to the same class differ by a factor of ten. An indication of the ratio of the first three macronutrients in organisms is given by the stoichiometric ratio of the organisms' bodies. The average elemental composition (by weight) of aquatic autotrophs, for example, is about C : N : P = 106 : 16 : 1. In land plants the average ratio of C : N : P corresponds to 250 : 10 : 1 according to data in [21], and the entire set of macro-elements will give a C : N : P : K : Mg : S ratio of 250 : 10 : 1 : 5 : 5 : 1 : 0.5. Micronutrients are present in land plants in lower amounts (Fe 100, Mn 50, B 20, Zn 20, Co 6, and Mo 0.2 μg per gram dry weight). However, their ratios can vary

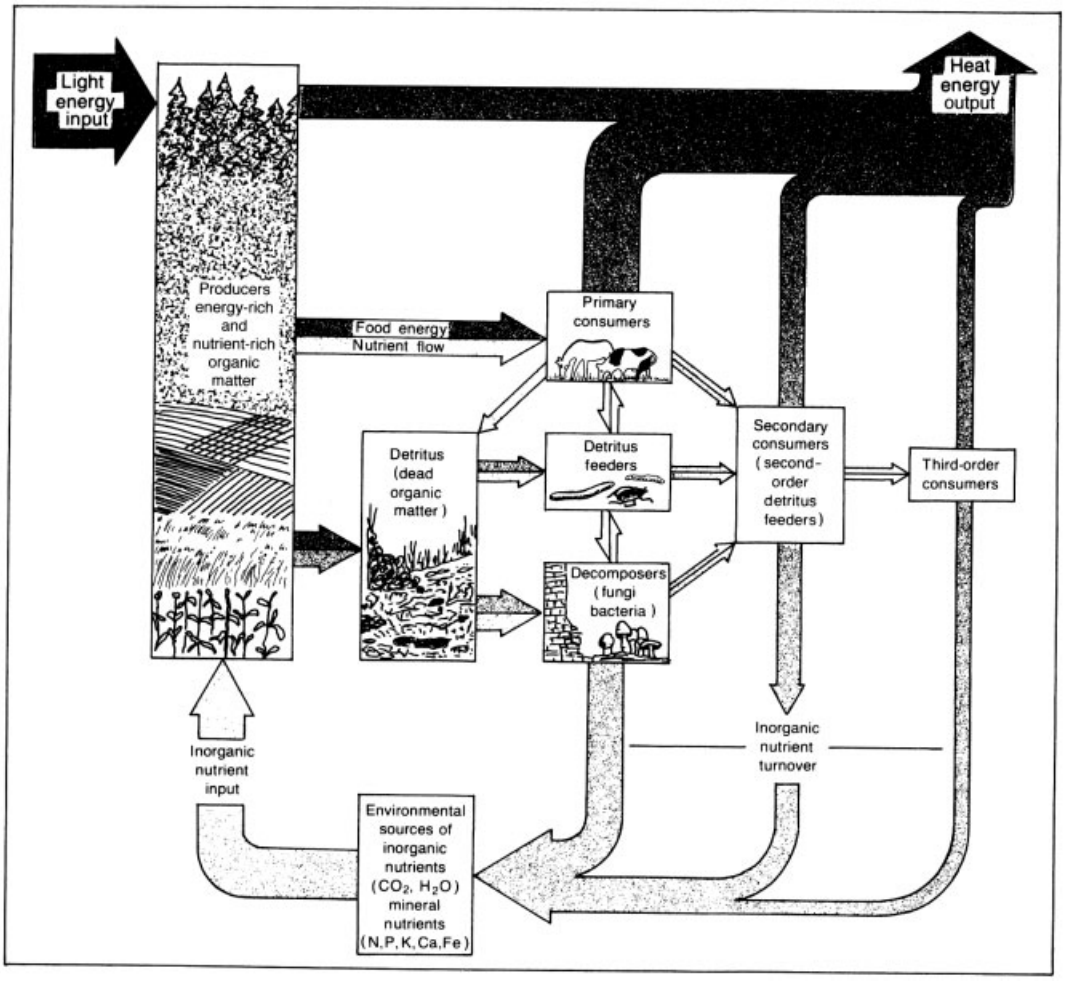

Figure 8. Nutrient cycling and productivity of ecosystems production decreases in relation to successive steps of the trophic chain (producers > herbivores > predators) (after [20], modified)

widely in individual plants (see, e.g., [148] for aquatic plants).

Chemical elements usually are not taken up as such by plants; rather, chemical compounds (e.g., C in the form of CO_2) are. Plants are now recognized to optimize their uptake. For example, in selection between nitrates and ammonia, the energetically more favorable ammonia is taken up first. Only some compounds of phosphorus are bioavailable [22].

A typical curve for the kinetics of nutrient uptake is given in Figure 10, which also shows that an element essential for the life of organisms may have a negative effect if it is present in excess quantities. In agriculture this is known as overfertilization. Plants need at least the first three macronutrients (C, N, P) simultaneously. The theory of limiting nutrient states that the nutrient in shortest supply represents the bottleneck. The uptake of all macronutrients is driven by the limiting one, the rest being taken up proportionally to the total needs.

Inorganic nutrients are taken up only by plants with photosynthetic pigments that use solar energy for the construction of highly structured organic matter from more simple inorganics. Such plants are called *autotrophs*; *heterotrophs* are organisms bound to organic matter formed primarily by autotrophs, but possibly transformed by other heterotrophs (e.g., carni-

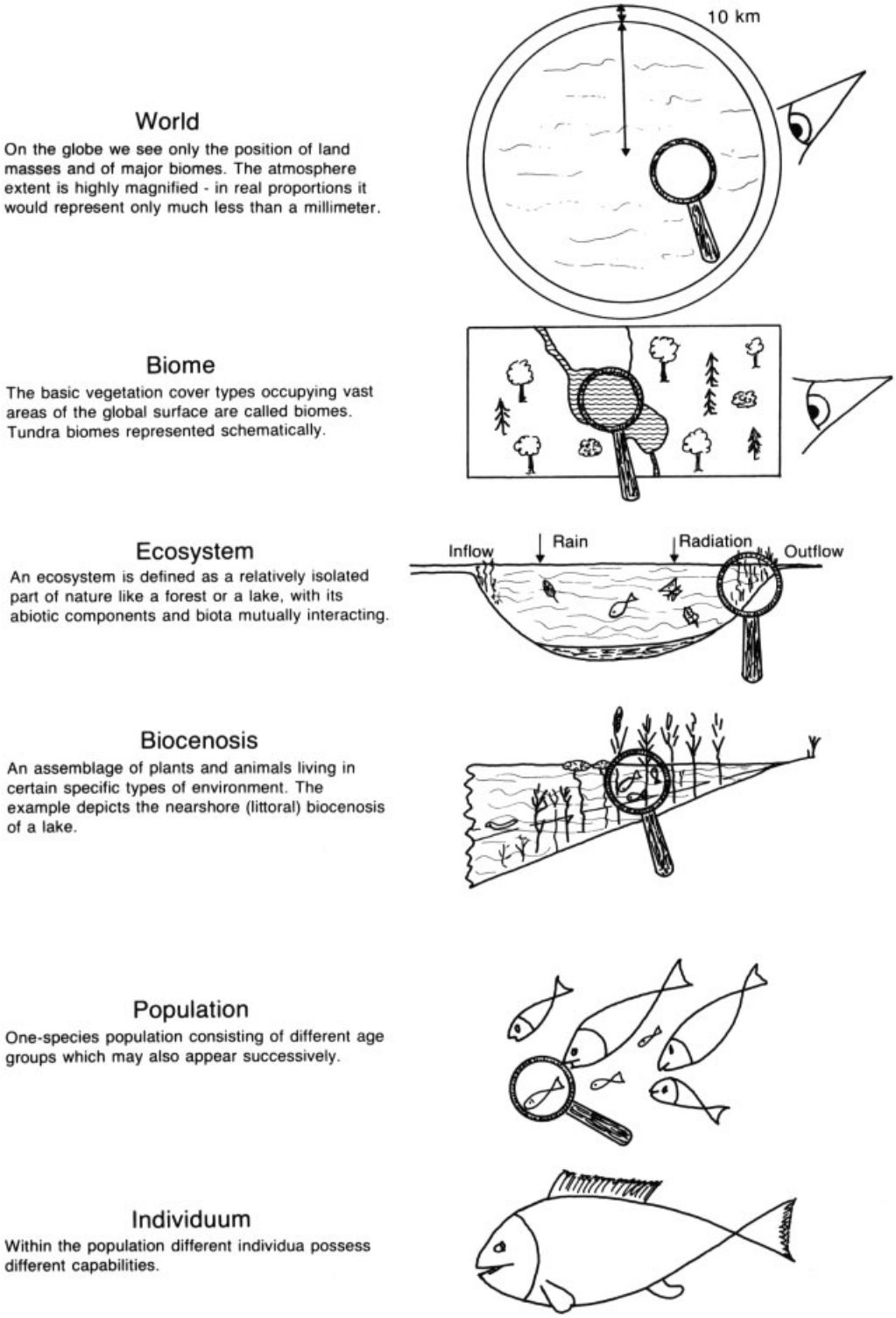

Figure 9. Hierarchy of biological organization; looking at selected objects of nature with successively larger or smaller magnification involves a simultaneous change of view

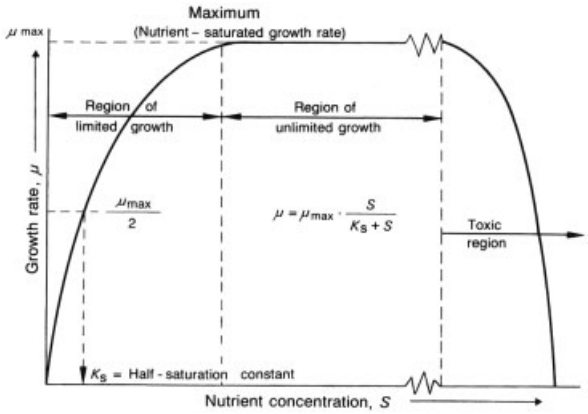

Figure 10. Kinetics of nutrient uptake by organisms
Organismic growth depends predictably on nutrient concentration, following a relationship known from enzyme kinetics.

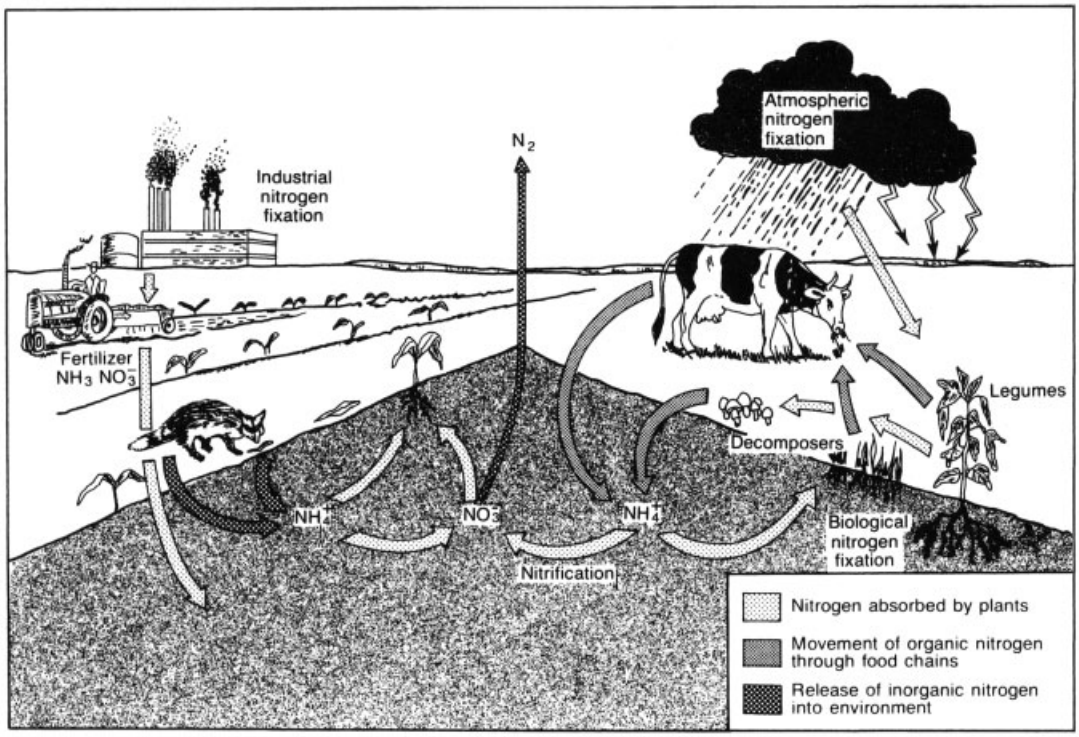

Figure 11. Nutrient (nitrogen) cycle in a man-made landscape (after [20])

vores). Bacteria, in this connection, are known as *decomposers*. They decompose matter to mineral components; bacteria belong to a larger group of destruents, degrading dead organic material, the so-called detritus.

2.1.4.2. Nutrient Cycling in Ecosystems

Two basic types of cycles can be distinguished: the short-circuit cycle in natural ecosys-tems (see Fig. 8) and the open cycle in agricultural ecosystems (Fig. 11). In natural ecosystems, nutrients circulate rapidly among the organisms,

soil, and groundwater. Soil constantly coveredby vegetation retains water with the dissolved nutrients. The vegetation cover also retains humidity. Evaporation creates backflow from the atmosphere as rain. Soil is rich in organic matter, which keeps the water capacity high. Therefore, nutrient loss from vegetated soils is low. Simultaneously, soil erosion and washout of nutrients are minimized.

In today's agricultural and inhabited landscapes, the situation is drastically different. The soil is barren or loosely covered, at least for a part of the year. Because of lack of rhizosphere, it is subject to high erosion and flushing out of nutrients to streams and rivers and, consequently, to lakes or sea. The soil is continuously deprived of organic matter, which decreases its water retention capacity. Instead of a short-circuit, an open cycle is observed, with considerable nonreversible losses. In the United States, soil erosion was estimated in 1968 and 1980 to represent 36×10^9 kg/a, about one-half of which originated from agriculture.

The opening of the cycle of organics and nutrients occurred in the Middle Ages when the increasing accumulation of people in towns caused diseases to spread. The construction of sewers disrupted the natural cycle, the original backflow of organic remnants and nutrients to the soil being diverted.

An experiment showing which major changes are produced by deforestation was obtained on the Hubbard Brook Ecosystem [4], [23]. Trees were logged within the entire watershed, and the outflow and its chemical composition were carefully measured. An increase in soil erosion and leaching of nitrogen, phosphorus, and sulfur to groundwater and streams was observed. Because water flux through the soil has increased, the soil loses its fertility. The organic forest floor rapidly decomposes and eventually disappears. The variability of flow in streams increases, with much more transport of the streambed produced during larger flows. However, the average flow decreases. Nitrogen released to the atmosphere during burning of remains has negative effects on the atmosphere. Reradiation from barren soil creates ground-level frosts on clear nights. Soil temperature is significantly increased on hot days.

Carbon Cycle. The global cycling of carbon is gaining in importance, not only because of plant nutrition but mainly because of its increasing concentrations in the atmosphere and the oceans. Because of the increase of several "greenhouse" gases – among which CO_2 plays the leading role – a global change in world weather conditions is expected. Carbon is essentially exchanged among the four reservoirs shown in Figure 12, where the past (in the atmosphere) and present reserves and flows are also given. The importance of forests and oceans as balancing elements in the CO_2 budget is enormous.

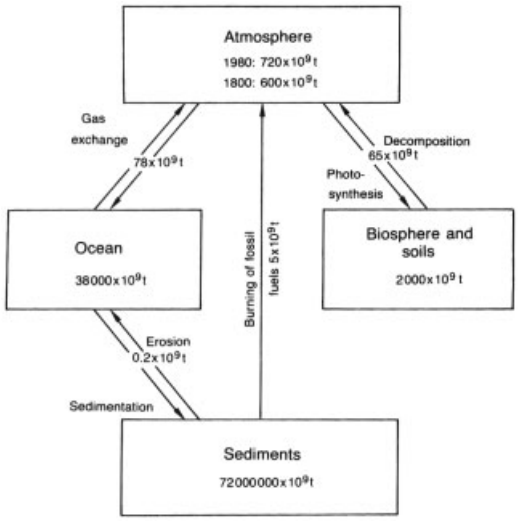

Figure 12. Major carbon reservoirs (after [24])

Nitrogen Cycle. Basic transformation processes in the terrestrial cycle of nitrogen are shown in Figure 13. Two pathways lead out of the short-circuiting between soil, crops, and animals: to the atmosphere and to water. Losses to the atmosphere have two dominant sources: industry, which releases NO_x; and agriculture, specifically animal husbandry, which produces NH_3. Both gases contribute to the greenhouse effects. Losses of nitrogen to waters also have negative consequences. Figure 14 shows the enormous rise in concentration of nitrates in water bodies in intensively cultivated landscapes of Europe and North America. Since the pristine state, the concentration has increased by a fac-

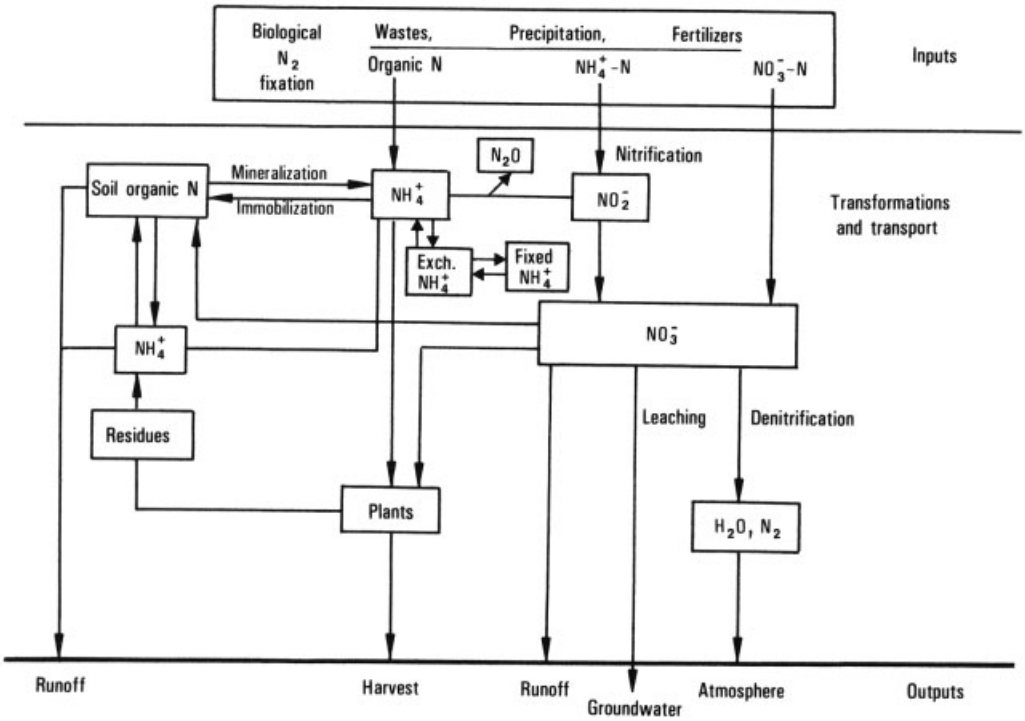

Figure 13. Soil nitrogen cycle (after [25], modified)

tor of ten and is occurring in both surface and groundwater. Most of this nitrogen increase is attributable to agriculture, for which it represents enourmous economic losses because of unutilized fertilizers. World fertilizer use increased from about 5 kg per capita in 1950 to 28 kg in 1985 [151]. The consequence of this fertilization of waters is an increased growth of aquatic vegetation, particularly of nuisance algae. In drinking water supply, excess nitrates result in hygienic problems, among them the deadly infant disease methemoglobinemia. For these reasons, every country has standards for the maximum permissible concentrations of nitrates in drinking water. The problem of cheap extraction of nitrogen from water is far from being solved on a global scale, although technological nitrogen-elimination processes are known. The nonsensicality of losing amounts of costly nutrients from agriculture on the one hand and eliminating them from waters at high costs on the other seems evident.

Phosphorus Cycling. A popular representation of the phosphorus cycle is given in Figure 15. Similarly to C and N, short-circuit cycling is characteristic of natural undisturbed vegetated landscapes. An open cycle with losses, mainly to waters, typifies agricultural and town environments. The difference from the previous cycles is that no phosphorus gas phase occurs. In organisms, phosphates are oxidized to produce energy in cellular respiration, and the phosphate is released in urine or similar waste back to the environment. When phosphorus compounds (particulate P, adsorbed P, dissolved orthophosphates, inorganic polyphosphates, organic phosphorus dissolved in the water phase) enter waterways, they enrich – indeed, overenrich – aquatic ecosystems. In waters, the processes of phosphorus uptake by aquatic flowering plants and microflora as well as processes of P immobilization and microbial synthesis are inadequate to utilize these amounts. Besides nuisance growth, considerable amounts of P-rich sediments accumulate in estuaries, lakes, and reservoirs. From water, little natural return to the land occurs. The sources of phosphorus for water are listed below:

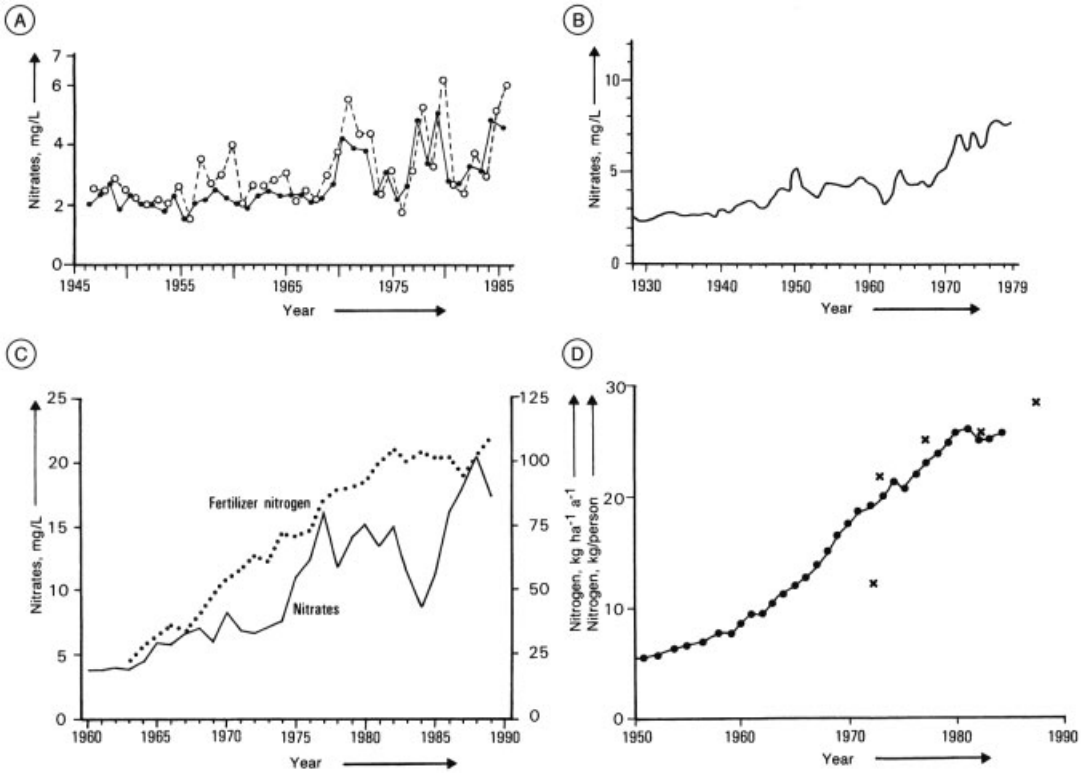

Figure 14. Trends of nitrate in freshwater and fertilizer use
A) Mean winter – spring concentrations of nitrate in Esthwaite Water and Blelham Tarn (Lake District, United Kingdom); B) Mean annual nitrate concentrations at Walton on Thames (United Kingdom), after [149]; C) Mean annual nitrate concentrations in the Slapy Reservoir (River Vltana, Czech. Republic), and related fertilizer use in the drainage basin [150]; D) Mean global fertilizer use per capita in 1950 – 1984 [151]; crosses represent newer data [26]

Natural surface load	0.07 – 0.12 kg/ha
Atmospheric load	0.4 – 1.2 kg/mm
From agricultural areas	up to 1.8 kg/ha
Livestock sewage	up to 13 kg per capita cattle
Sewers	0.85 kg per capita dweller
Septic tanks	0.14 kg per septic tank
Municipal sewage	1.1 kg per hectare of town area

Although great differentiation exists locally, two sources of phosphorus seem to dominate in the developed world: detergents and soil erosion. Phosphorus applied as fertilizer does not go into waterways in significant amounts unless some scattering from stores or spreading during aircraft application occurs.

2.1.4.3. Productivity of Ecosystems

Major biomes of the world are represented in Figure 16. The total biomass of plants on the globe is estimated to be 1.84×10^{15} kg [28]. The value is rapidly diminishing because of forest clearing, soil salinization, and desertification. In tropical countries of America, Africa, and Asia, the rate of loss represented 37, 48, and 40 %, respectively, between 1980 and 1985 [29]. Typical feedback between the nutrients and organisms is shown in Figure 17, which illustrates the dynamics of natural systems.

However, nutrients are not the only decisive agents for biological production. As evident from the summary photosynthetic equation

$$6\,CO_2 + 6\,H_2O \xrightarrow{\text{Energy}} C_6H_{12}O_6 + 6\,O_2 \qquad (24)$$

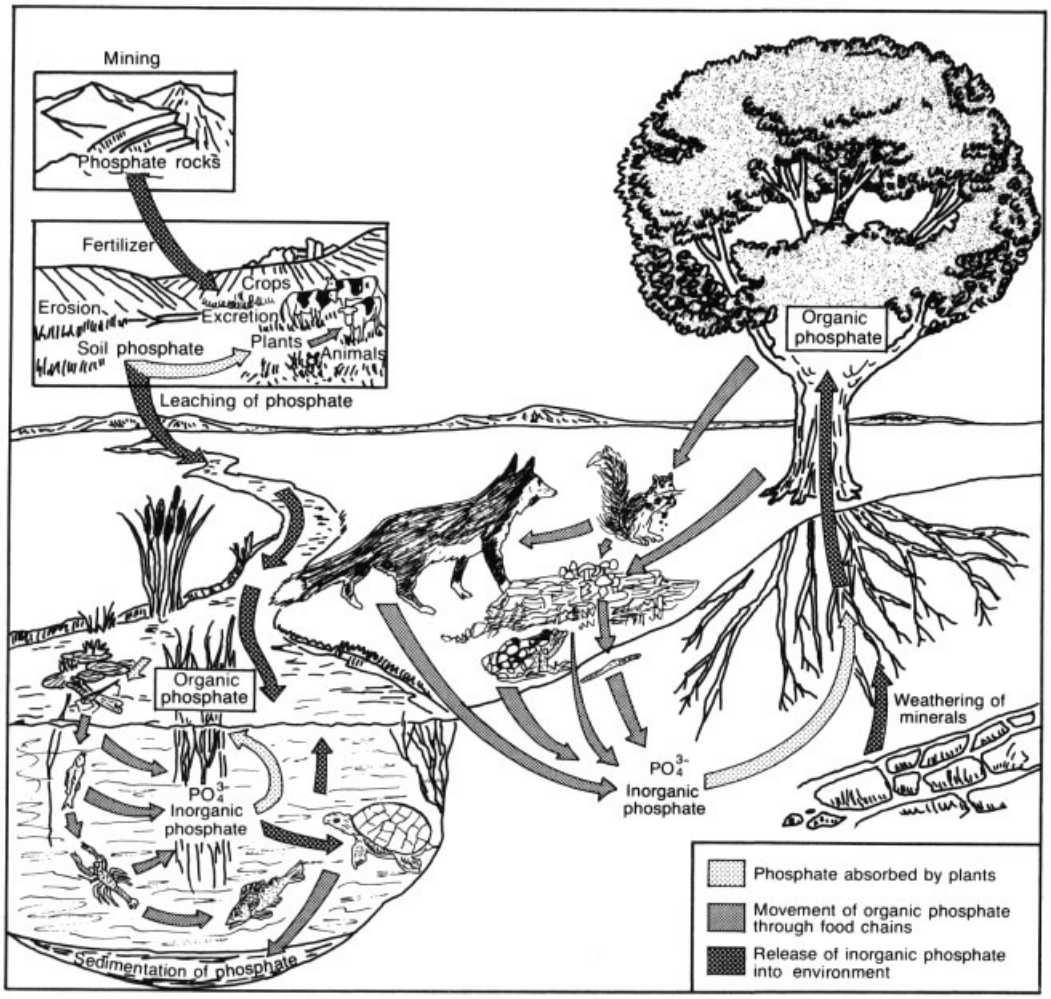

Figure 15. Phosphorus cycle (after [20])

energy (usually solar energy) and water are necessary. Like chemical reactions, biological reaction rates are also affected by temperature. The difference is that life is confined only to some −50 to +50 °C. Characteristic shapes of light and temperature dependence of organismic growth are given in Figure 18. For the above reasons the fate of nutrients is highly dependent on geographical position as well as seasonal and daily changes, as shown in Figure 19.

Productivity of Terrestrial Ecosystems. A distinction is made between natural and the agricultural production, although temperate forests are now also cultivated. Quantitative data on primary production around the world were elaborated in [28], [152]; data on sites where primary and secondary production was measured simultaneously are summarized in [30], [153]. The results of these studies are given in Table 4. Such estimates are methodologically very difficult, and some of them might be highly biased. Only the net aboveground production was measured; belowground production is known to represent up to five times as much as aboveground [31]. Net production is distinguished from gross production in that it does not include organic matter produced but used immediately for maintenance (respiration, excretion). The range of values for 104 sites was es-

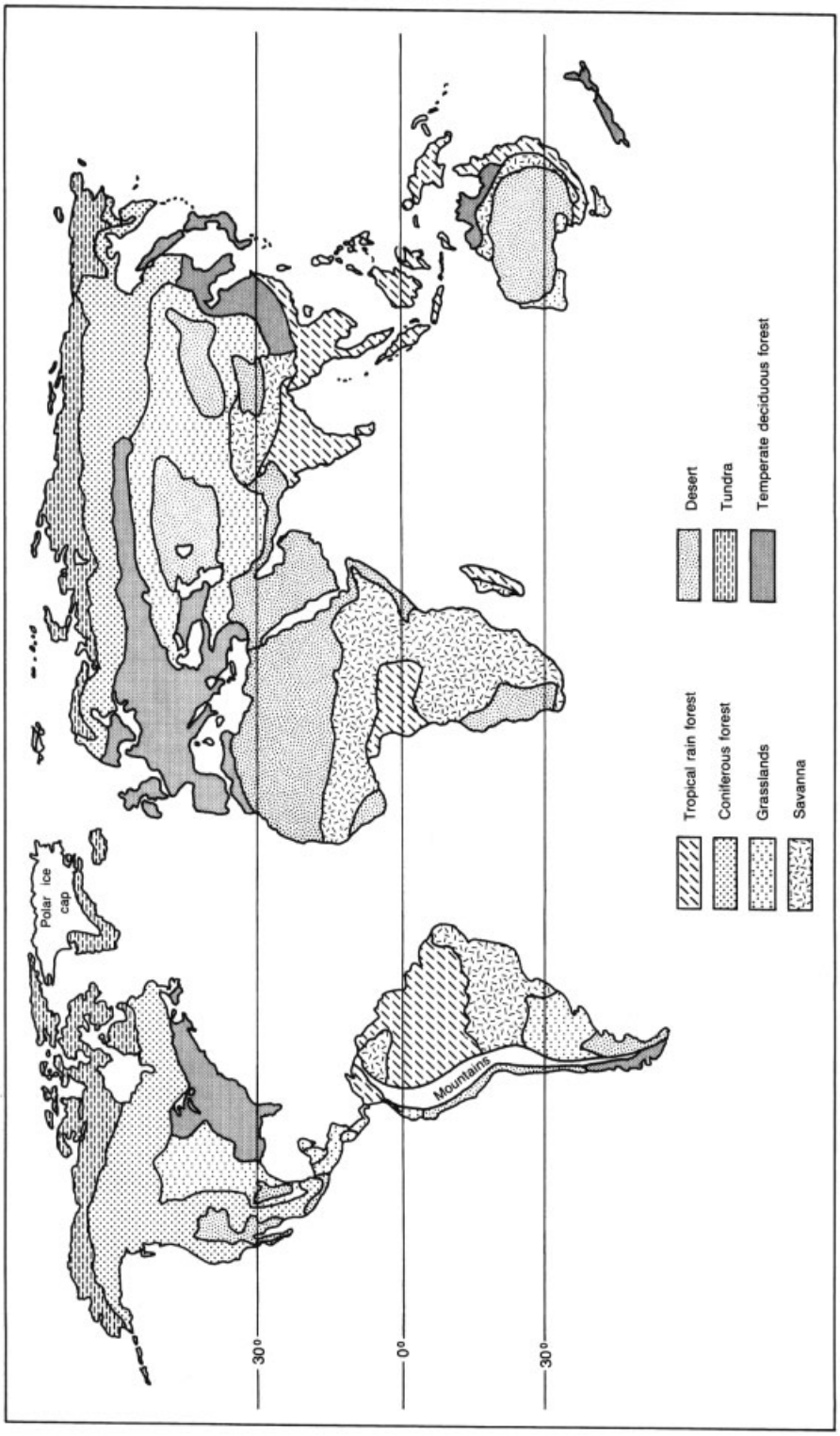

Figure 16. Major biomes of the world (after [27])

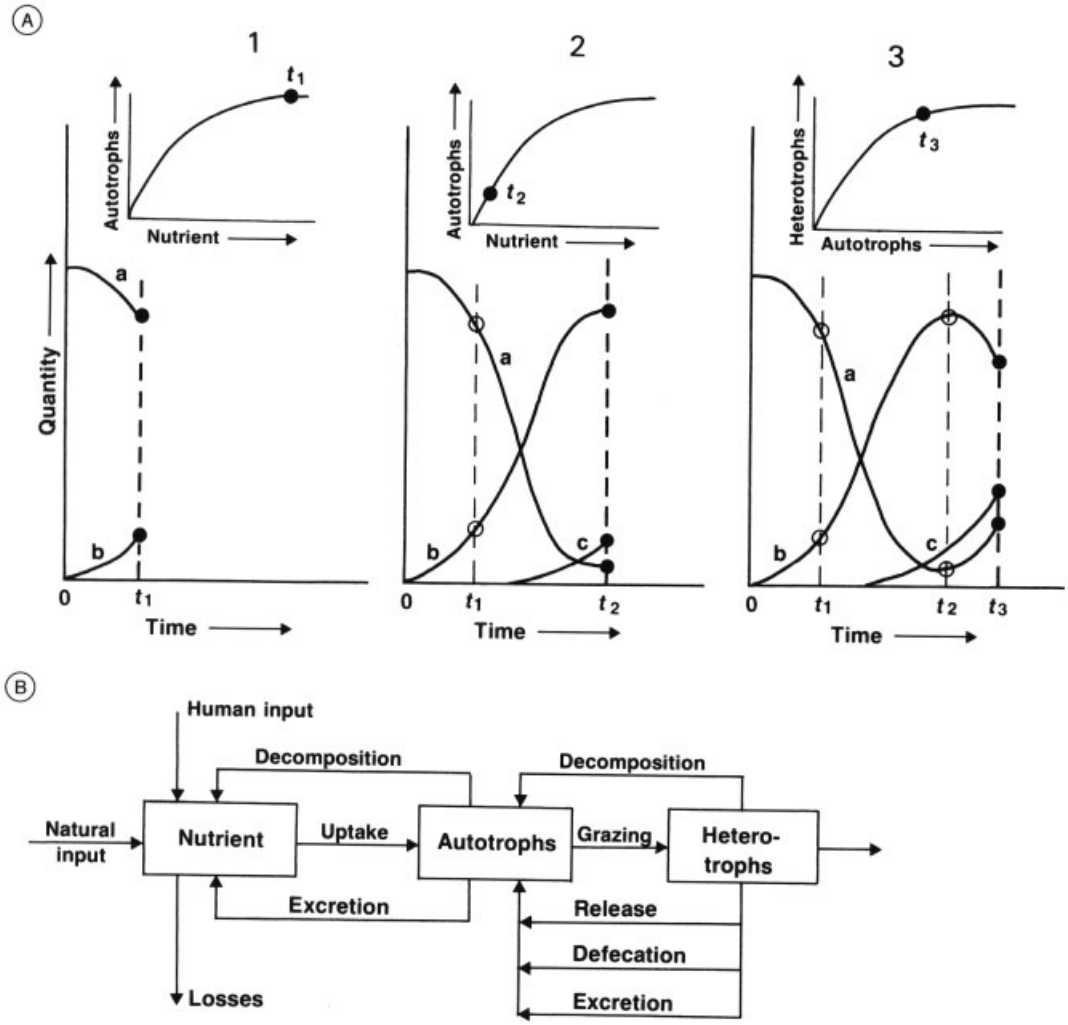

Figure 17. Essential processes of ecosystem dynamics
A) Time development of a population of autotrophs and heterotrophs in a nutrient-limited environment 1) At time t_1 (e.g., in winter) nutrients (curve a) are high and autotrophs (curve b) low, due to low temperatures. When autotrophs start to appear, they grow quickly and take up nutrients rapidly. 2) At time t_2, growth ceases, being severely limited by the decreasing quantity of nutrients (see inset). 3) Heterotrophs (curve c) appear with retardation, because their life cycles are long and at the beginning their abundance is limited by low food supply (i.e., low amounts of autotrophs on which they feed). At time t_3 autotrophs decrease, being heavily grazed on by herbivores (heterotrophs). Nutrients are regenerated both from senescent and decaying autotrophs and from excretion, defecation, and nutrient release (as in Fig. 17 B)
B) Main processes in the nutrient-productivity cycle. Losses are represented by flushing nutrients into groundwater and streams.

timated as 30 g fresh weight per square meter and year for a desert to 7050 $g\,m^{-2}\,a^{-1}$ for a tropical forest. The average estimates for various ecosystems range from 0 – 10 g dry weight per square meter and year for extreme desert through 10 – 250 for deserts and 400 – 2000 for boreal forests up to 1000 – 3500 for tropical rain forest and 800 – 6000 g dry weight per square meter and year for swamps and marshes. Table 4 shows that forests in general are the most productive ecosystems. Column 3 also shows that a high proportion of the primary production in nature is consumed by herbivores, be they insects, small rodents, or large ungulates. On average, 1.3 to 64 % of primary production is consumed in different biomes, with by far the highest pro-

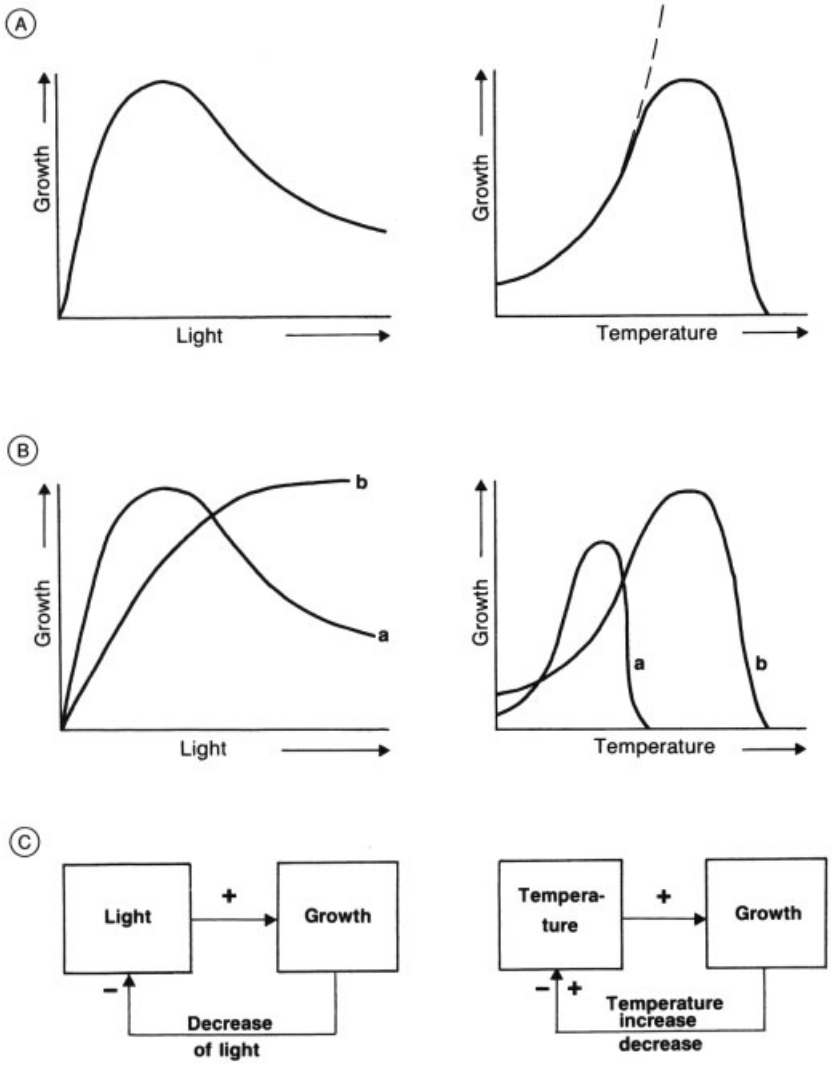

Figure 18. Light and temperature dependence of organismic growth and related feedback mechanisms
A) Typical shapes (equations not shown) of the dependence of plant growth on light and temperature. The dotted line shows the dependence typical for a chemical reaction (Arrhenius curve). B) In contrast to chemical reactions, which follow a fixed curve, the organisms are capable of adaptation to changing conditions. The two curves show changes for low (a) and high (b) light and low (a) and high (b) temperature conditions, respectively. Within the pertinent range of conditions the adapted organisms have the same high performance. However, full compensation is not always possible, as the temperature case shows. C) Feedback effect of growth on light (left) and temperature (right). In the case of light the feedback is negative; for temperature, either positive (increase of temperature, e.g., in water where more radiation is absorbed by microscopic algae) or negative feedback relationships (decrease of temperature by shading in forests) are possible

portion in tropical grassland. The primary production taken up is converted to herbivore production with an ecological efficiency (expressed as the amount of primary production converted to herbivore production) of about 0.35 %. The efficiency of herbivore utilization of primary production calculated from columns 3 and 4 results in an average value of 0.93 %, with a range between 0.09 for tropical grasslands and 13.3 % for salt marshes. Two types of ecological cycles in terrestrial ecosystems are distinguished in [31]: a short one, where most of the primary production

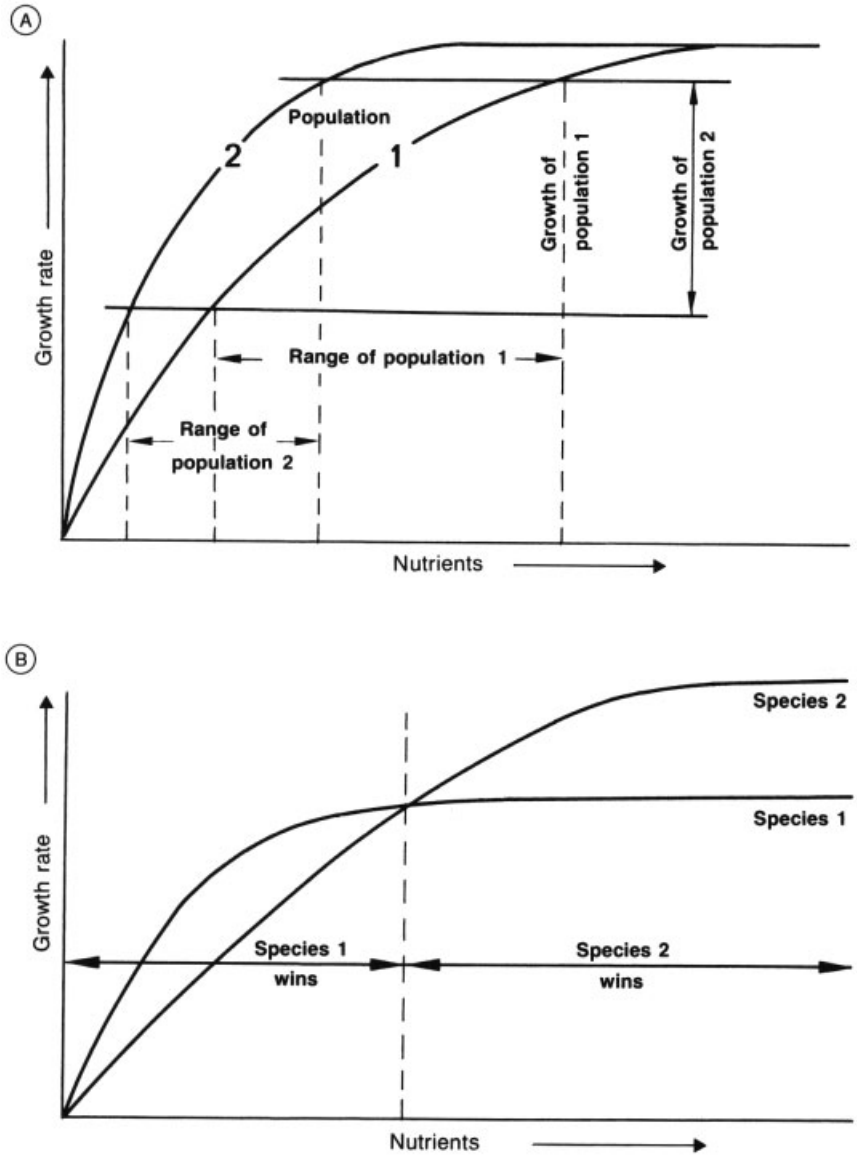

Figure 19. Adaptation of populations to nutrient supply and competition for nutrients
A) Populations adapt to low-nutrient conditions, as indicated by increasing slope of the curve. Population 1 living in high-nutrient conditions shows a lower initial slope of the curve, while population 2 living in low concentrations shows a high initial slope, enabling it to utilize low concentrations more efficiently
B) Species compete for nutrients. In a certain condition range, one or the other species dominates, depending on the range in which the curve reaches the highest values.

is immediately decomposed, while in the long one, herbivores not only consume a large part of primary production but also are preyed upon by carnivores and a complex food web originates. In Table 4, column 3, only temperate forests show the short cycle, with only about 1 % of primary production being consumed by herbivores. In all other instances the primary production consumed is of the order of magnitude of 10 %. For tropical grassland, seemingly the most complex system, more than half of the primary production is consumed by herbivores. Table 5 gives exam-

Table 4. Data on primary and secondary (herbivore) production of natural ecosystems in different biomes (after [153], modified) [a]

Ecosystem	Prim. prod. [b]	Biomass [c]	Consumption [d]	Herbivores [e]	Efficiency [f]
Desert	403	1	32 (7.9)	0.1	0.02
Tundra	257	0.5	16 (6.2)	0.2	0.07
Temperate grassland	1181	1	161 (13.6)	14.6	1.23
Temperate old field	1352	2.2	95 (7.0)	10.3	0.76
Tropical grassland	2694	8	1725 (64.0)	1.5	0.06
Temperate forest	5032	1.2	64 (1.3)	0.8	0.02
Tropical forest	5400	17	591 (10.9)		
Salt marsh	4328		345 (8.0)	46.0	1.06
Average	2276	8.1	861 (37.9)	8.0	0.35

[a] Values are given in grams of fresh weight per square meter and year.
[b] Prim. prod. = net aboveground annual plant production.
[c] Herbivore biomass in grams of fresh weight per square meter and year.
[d] Amount of primary production consumed by herbivores. The percentage of primary production being consumed by herbivores is given in brackets.
[e] Annual production by herbivores.
[f] Ecological efficiency, i.e., the percentage of primary production converted to herbivore production.

ples of net aboveground primary production of agrarian ecosystems and secondary production of domestic livestock compared to game.

Table 5. Examples of net aboveground primary production of agricultural eosystems and production of domestic livestock per area as compared to wild animals (data from [10], [23])

Primary production, kg dry weight per square meter and year		
Crop	World average	Maximum per country
Wheat	0.4	1.3
Corn	0.7	1.3
Rice	0.6	1.5
Potatoes	2.5	4.6
Soybean	0.35	0.6
Sugarcane	0.9	3.3
Sugar beet	2.0	
Secondary production		
Type of animals	Weight gain, kg fresh weight per day	Biomass, kg/km^2
Sheep	0.05	
Cattle	0.14	
Thomson's gazelle	0.06	
Topi	0.20	
Wildebeest	0.24	
Domestic livestock		10 000 – 27 000
Domestic livestock in East Africa		10 000 – 14 000
Wild ungulates in East Africa		61 000 – 87 000
Bison		12 000 – 17 000
Black tailed deer in California		9 200

Surprisingly, comparison of Table 5 with Table 4 shows that natural production and agricultural production are not much different, if one ignores the extreme energy needs to achieve high agricultural yields [154]. Often the fossil energy input is greater than the energy content of the food harvested [32]. According to [33], for an eightfold increase in yield of major crops, a twentyfold increase in energy applied is needed. The idea that much more output is obtained when artificially cultivating domestic livestock is not supported by these data. The use of natural vegetation by natural livestock is far from inefficient. That was recognized in tropical countries and became a political issue in opposition to what was exported from the developed countries as the "green revolution." The green revolution appears to lead to the continuing poverty of poor countries. They are made dependent on Western technology and on energy sources not available at home, without adequate rewards. The simultaneous trend for sustainable development leads to major revisions of prevailing ideas concerning agricultural production.

Productivity of Aquatic Ecosystems. The primary production of fresh waters may be divided into two groups: (1) production of flowering plants living in shallow water of marshes, swamps, and along the shores of lakes and slow-flowing streams; and (2) microflora inhabiting primarily the open lake and slow river central regions. Microflora consists of several groups of microscopic algae and species of cyanobacteria (also called blue-green algae). These free-floating organisms are called phytoplankton.

The transition between land and water belongs to the most productive plant environments: no water deficit exists; nutrients are obtained both from rich aquatic soil and from water. Under comparable geographic conditions, the highest production among all vegetation is obtained in aquatic macroflora. The dependence of aquatic primary production on geographical position is due to a combination of effects such as light and temperature. For phytoplankton, conditions of mixing in lakes dependent on geographical latitude and altitude are decisive. During the annual cycle, lakes undergo periods of intensive deep mixing to the bottom, as well as periods of stratification. During summer stratification, upper layers become warmer than deeper ones, and intensive mixing is confined to several top meters. The density difference between warm top and cold bottom layers is usually sharp. At a certain depth, a jump in temperature occurs between the two basic layers called the mixing layer (thermocline) and the hypolimnion. The depth of the mixing layer affects primary production. Algae moved within this layer periodically encounter high light intensity when taken to the water surface and reduced or no light when mixed to depths. Greater lakes are usually clearer and mixed to greater depths due to higher wind stress; their primary productivity is lower than that of shallow lakes with comparable nutrients.

The most important determinant of lake productivity is nutrient input. In the 1960s, controversy existed – mainly between representatives of the soap and detergent industry and limnologists studying lakes – whether carbon, nitrogen, or phosphorus is the nutrient most often limiting aquatic primary production. The situation is now clear: depending on the mutual relation among C, N, and P any of them can be limiting, as can several micronutrients. However, in most lakes at present, phosphorus is the limiting agent, since nitrogen is usually supplied from human activities in excess. Carbon can limit in instances of an excess of both nitrogen and phosphorus; carbon deficiency is due mainly to limited exchange of CO_2 between water and atmosphere. This is the case in highly fertilized fish ponds. A systematic difference exists between fresh water on the one hand and brackish and marine waters on the other: in the latter, nitrogen limitation prevails. Instances of limitation by micronutrients are reported rarely, particularly from regions without human activities, where the lack of some elements is geologically conditioned [34].

Table 6. Primary and secondary production in fresh waters, based on different sources and data obtained mainly during the International Biological Program [5] (for more detailed data see [31])

Type of water body	Macroflora Aboveground primary production, kg dry weight per square meter and year	Average biomass, kg dry weight per square meter and year
Lakes and streams	0.1 – 3.0	0.02
Swamps and marshes	0.8 – 8.5	1.5
	Microflora (phytoplankton)	
Phosphorus and nitrate concentrations, µg/L	Primary production, kg dry weight per square meter and year	Average biomass, g dry weight per square meter
Low P (<5) and N (<250)	0.015 – 0.09	0.4 – 2.0
Medium P (<30), N (<500)	0.075 – 0.3	2 – 5
High P (>30), N (>500)	>0.43	>5
	Type of water body fish production,∗ kg fresh weight per square meter and year	
Small lakes	0.002 – 0.18	
Great lakes	1 – 7	
Tropical lakes	2 – 250	
European fish ponds	500 – 850	
Chinese fish ponds	500 – 7500	

∗ Fish production is estimated from harvested production. According to recent estimates harvested production represents in lakes about 10 % of total production and in culture about 50 %.

Table 6 gives some ideas of the primary production of fresh water depending on phosphorus and nitrogen concentrations. In addition, there is a geographical differentiation of phytoplankton primary production at similar conditions of nutrients. In Arctic and Antarctic lakes values between 0.002 and 0.07 kg dry weight per square meter and year were observed. In alpine lakes the lowest values observed are similar to those in the Arctis but maximum values up to 0.2 are reached. For temperate and tropical lakes the ranges are 0.004 – 1.8 and 0.06 – 5.0, respectively. Some data on herbivorous production, represented by fish, are also indicated. The biomass of phytoplanktonic algae is about three factors of ten lower than in macrophytes, whereas production is similar. This is due to the short life cycle and high turnover rate of microscopic algae. Under intensive artificial cultiva-

tion, some algal species have produced up to 8 kg dry weight per square meter and year.

2.2. Terrestrial, Lotic, Lentic, and Wetland Ecotones

A boundary in nature may be created by discontinuities in the external influences that control the ecosystems or by discontinuities in the functional network of the systems themselves. Consequently, a wide spectrum of equilibrium system responses to values of environmental gradients exists.

In Figure 20 A the response of the ecosystem as characterized by one indicator variable is as smooth and linear as the changes in gradient; Figure 20 B depicts an abrupt break in response, probably due to a threshold response in one of the components of the ecosystem. In Figure 20 C the response has a twist, indicating that the system can have two stable equilibrium states. In these cases the ecosystem state and behavior with respect to driving variables depend on the history of the systems (hysteresis). Hysteresis plays a major role in catastrophe theory. In Figure 20 D the system has two stable states and can display multiple internal reactions to take on one or the other.

In more abstract or unifying terms the individual response curves of Figure 20 can be considered as a continuum of ecosystem responses to an environmental variable. Thus, the various responses grade into one another depending on internal reactions and autonomous features of the ecosystem. The environmental gradients can be continuous or discontinuous to form a variety of pattern that, together with the response of the ecosystems, control the nature and pattern of the transition zones between adjacent systems.

These transition zones or *ecotones* [156] have a set of characteristics uniquely defined by space and time scales and by the strength of interaction between adjacent ecosystems (see Section 2.3). Because ecotones usually represent relatively steep gradients in both abiotic and biotic variables, they facilitate the ecological analysis of mechanisms contributing to biodiversity and flow of energy or materials across the landscape [157].

2.2.1. Terrestrial and Lotic Ecotones

In view of the ecological importance of ecotones, they can be divided into three broad classes.

Terrestrial ecotones may be pragmatically defined as discontinuities between systems. Consequently, the most obvious ecotones in terrestrial ecosystem complexes involve transitions between plants of communities dominated by different life forms. Large-scale examples are transitions from tundra to taiga along thermal gradients or transitions from prairie to forest along continental moisture gradients; a frequent small-scale type is the transition from field to woodlot in agricultural landscapes. This pragmatic view of terrestrial ecotones depends highly on the contrast between the different elements contained in the definition of the landscape, and the degree of contrast is necessarily a function of the variables used by the observer to categorize the landscape elements [138], [158]. Thus, a more precise and generally applicable method of defining ecotones is based on areally valid measurements of ecologically relevant parameters and subsequent multivariate cluster analysis. The results are operationally homogeneous spatial units (patches) separated by distinct boundaries whose location and geometry depend on the variables used to characterize the primary measuring points or elementary spatial units, respectively; the fusion level of clustering adopted; and finally the scale selected [159], [160].

Lotic ecotones can be defined, in an analogous way, as fluvial boundaries. This definition indicates that resource patches are separated by both longitudinal (i.e., upstream – downstream) and lateral (landward) ecotones that operate over highly variable temporal and spatial scales. This type of ecotone appears to be highly sensitive to landscape changes caused by physical and biotic disturbances. Examples of direct influences of disturbances are removal of riparian vegetation by floods or landslides, and edaphic or vegetative modifications by animals and, in particular, humans. Indirect influences, such as changes in concentration gradients of dissolved chemical compounds and pathways of chemical reactions,

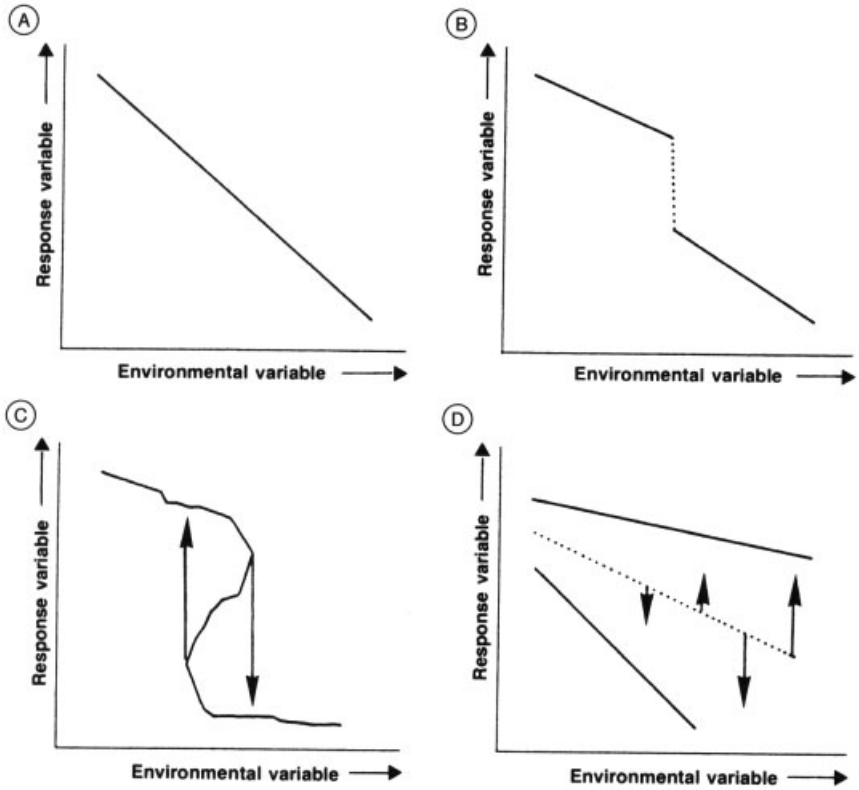

Figure 20. Ecosystem response to different types of environmental gradients (after [155], modified)
A) Case in which the response of the ecosystem is smooth with respect to the environmental variable; B) Case in which the response is discontinuous; C) Case in which the response is folded and results in hysteresis; D) Case in which the system can have multiple responses

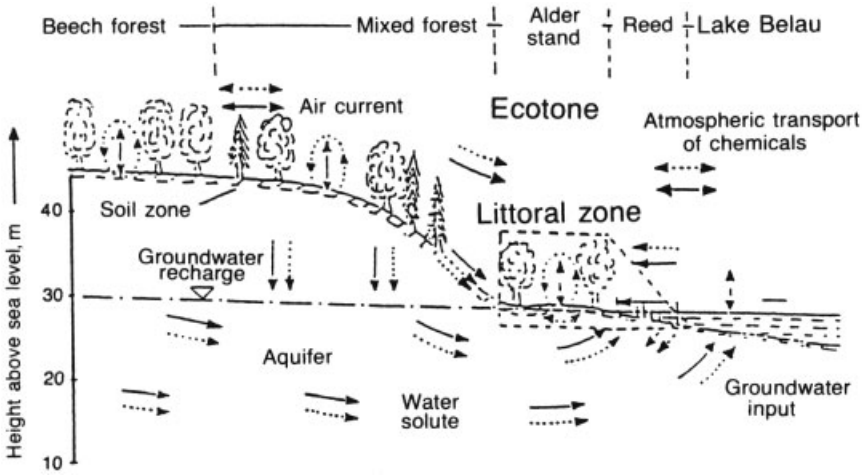

Figure 21. Exchange processes in ecotones of the Bornhöved Lake District, Schleswig-Holstein (after [157], modified)
a.s.l. = above sea level

may result from changes in edaphic and vegetative properties following direct impacts [2], [35], [161], [162].

2.2.2. Lentic and Wetland Ecotones

Lentic ecotones. (i.e., boundaries between water bodies such as lakes, ponds, or reservoirs and adjacent terrestrial patches) also play an important role in coupling ecosystems. With their steep gradients and typically great heterogeneity, lentic ecotones can contribute considerably to mitigating unfavorable changes in aquatic ecosystems. Furthermore, they regulate the landscape mosaic by controlling energy and nutrient flow between adjacent patches in the ecosystems [157], [163].

Wetlands are defined as lands transitional between terrestrial and aquatic ecosystems where the water table is at or near the surface or the land is seasonally covered by shallow water. Therefore they support, at least periodically, hydrophytes, and their substrate is classified predominantly as undrained hydric soil [164]. A clear-cut ecological distinction separates *tidal* and *inland wetlands*, which may be subdivided further [36]. Wetlands have internal and external boundaries separating distinct vegetation patches [165], [166]. They can be categorized into various types according to the flows that dominate in the tidal and inland environments. In both tidal and inland wetlands, vertical and horizontal transfers occur across a series of surficial and lateral boundaries. Flows across *surficial boundaries* (e.g., graded sediment beds or soil horizons) include transfer from aerobic to anaerobic soils, from soils to surficial vegetation and then to litter and vice versa, and from open water to the atmosphere.

Transfers across *lateral boundaries* include material flows from the upland to the wetland (upland – wetland ecotones) or from the wetland into adjacent open water (wetland – open-water ecotones), from groundwater aquifers into soils, or across vegetation patches with each community dominated by different species (wetland – wetland ecotones).

Figures 21 and 22 illustrate a combined lentic – wetland – upland ecotone situation in the Bornhöved Lake District (Schleswig-Holstein). The figures show that lentic ecotones form unusually differentiated habitats [136]. Various microenvironments are formed within the shore zone of the water body, and similar microenvironments occur in certain sections of water bodies contrasting in size, depth, productivity, and catchment area [167]. Lentic ecotones have rich plant and animal communities characterized by a shifting proportion of terrestrial and aquatic organisms.

Allochthonous and autochthonous material is the source of organic matter in ecotonal shore zones: matter originating from primary production by macrophytes and algae and accumulated material of terrestrial origin supplied by the drainage basin. A fast decomposition rate in several ecotonal habitats makes this material available to higher trophic levels much more quickly than in most land and water habitats [136], [163].

The majority of lentic ecotones in densely populated areas of the world is under visible human impact. Lake eutrophication, air pollution, sewage input, agricultural practices, and tourist activities often damage them. Disappearance of terrestrial and aquatic vegetation is the most common response to such unfavorable conditions.

2.3. Sensitivity of Ecosystems and Ecotones

In ecology, the sensitivity of biota or ecosystems to disturbances may be defined in terms of community stability. Questions of persistence and the probability of extinction are of particular importance in this connection; hence, these measures are also frequently defined as aspects of stability. Stability denotes the ability of a system to return to an equilibrium state after a limited disturbance, while persistence is an indication of resilience in the sense defined by Holling [168]: i.e., the capacity to absorb changes of state and driving variables and parameters. The more rapidly a system returns to equilibrium with the least fluctuation, the more stable it is, which means in terms of thermodynamics that stability is coupled with a (relative) minimum of entropy production [2], [169]. Further differentiations lead to various notions of community stability and nondemographic stability measures.

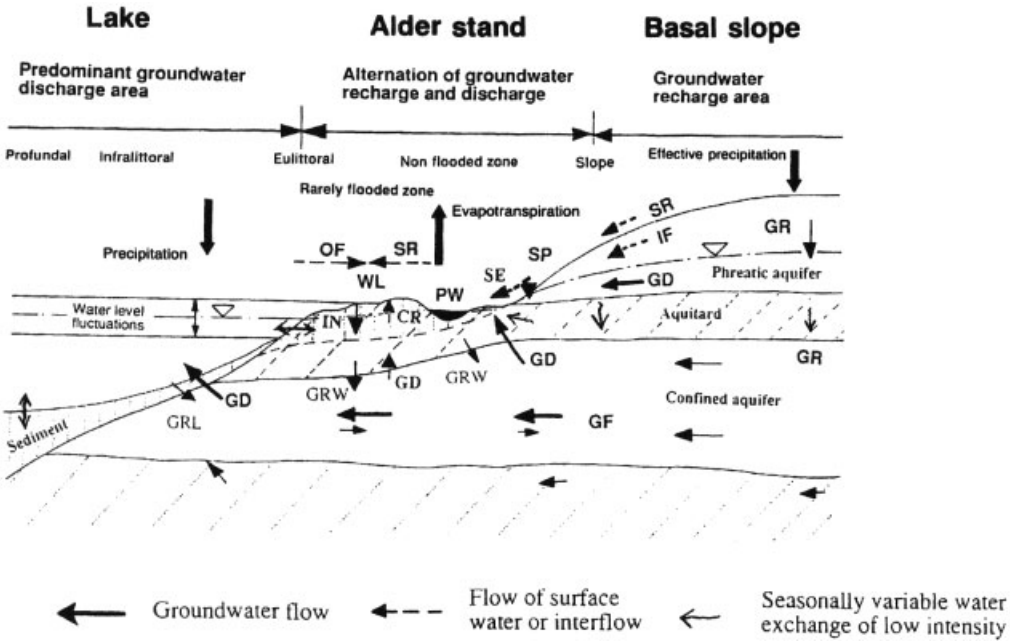

Figure 22. Water balance of ecotones adjacent to Lake Belau, Schleswig-Holstein (after [157], modified)
CR = Capillary rise; GD = Groundwater discharge; GF = Groundwater flow; GR = Groundwater recharge; GRL = Groundwater recharge by lake water; GRW = Groundwater recharge by wetland water; IF = Interflow; IN = Infiltration; OF = Overflow; PW = Ponding water; SE = Seepage on slope; SP = Spring at foot of slope; SR = Surface runoff; WL = Waterlogging

In the light of these definitions, a sensitive ecosystem is one that responds readily to a particular stress; an insensitive ecosystem may be oblivious to the stress. Sensitivity has also a temporal component: a system that responds more rapidly than another is considered more sensitive. Because of the complex character of system sensitivity to stress, various measurable aspects of sensitivity are recognized, which are summarized below:

– *Stability* how an (eco)system responds to perturbation and how readily it recovers
– *Resistance* rate at which (eco)system indicators change in response to a given stress – more resistant systems change more slowly
– *Recovery* how the system responds following removal of the stress. Recovery has two components: how rapidly and how effectively. (The temporal aspect is characterized as the system's resilience.)
– *Resilience* the inverse of the length of time required to return to near normal. According to WESTMAN [170], four components of resilience can be distinguished:
  – *Elasticity:* time required to restore the system to its initial steady state
  – *Amplitude:* zone of deformation from which the system will still return to its initial state
  – *Hysteresis:* extent to which the paths of degradation under chronic disturbance, and of recovery when disturbance ceases, are mirror images of each other
  – *Malleability:* degree to which the new steady state established following recovery differs from the original steady state

Most of the above definitions explicitly or implicitly assume that the system is in a steady state before a stress is applied. This is not necessarily true because ecosystems are generally continuously changing and developing. Therefore, the system need not return to a steady state after recovery, merely to a "near-normal" state.

In some instances the ecosystem can switch with relatively minor disturbances into another steady state. This is called bifurcation: rapid transition at a critical point into a new steady state. Examples of bifurcation of aquatic ecosystems are, for example, the rapid change of lakes from a clean-water state with low amounts of harmless algae to a highly polluted state with highly noxious blue-green algae, or the switching of prairie lakes with healthy fish populations to a state with fish-kills.

2.3.1. Different Notions of Community Stability

A first distinction is between local and global stability. *Local stability* or stability in the vicinity of an equilibrium point describes the tendency of a community to return to its original state when subjected to a minor perturbation. *Global stability* describes this tendency when the community is subjected to a major disturbance. The graphic visualization of solutions to the equations of population dynamics for an *m*-species community may be represented on some *m*-dimensional surface, where each point marks a set of populations. Then, equilibrium situations are in principle characterized by those points at which the surface is flat; by equating the configuration with that of a landscape, in customary geographical notation this means on hilltops and in valley bottoms. The hilltop equilibria, however, are obviously labile (i.e., unable to survive the smallest displacement); only the valley bottoms are stable configurations (see Fig. 23 A). In looking beyond the realm of linearized stability in the immediate neighborhood of equilibrium points for which straightforward mathematical tools exist the situation becomes more complicated. Appropriate representation of the *global stability* of a system implies recourse to nonlinear equations of population biology (see Fig. 23 B). Again in geographical terms, such a global analysis aims at comprehending the stability of an entire landscape, not just the immediate vicinity of equilibrium points. This involves appropriate recognition of the fact that real-world environments are uncertain and stochastic, which means that the corresponding environmental parameters in the model equations exhibit random fluctuations [6], [171], [172].

This leads to a yet more general meaning of stability, termed *structural stability*, which refers to the qualitative effects on solutions of the model equations that are due to gradual variations in the model parameters. Thus, a system may be considered structurally stable if these solutions change in a continuous manner. Conversely, a system is structurally unstable if gradual changes in the system parameters (e.g., alterations in site factors of a community) produce qualitatively discontinuous effects [2], [37], [172].

2.3.2. Biodiversity and Stability of Model Communities

An important issue of theoretical ecology is how biodiversity relates to community structure and which community-level properties emerge from the disparate interactions of organisms, populations, and site qualities. Possible examples of such "emergent properties" include trophic and guild structures, stability, resilience, and successional stages. A generalized offshoot of the concept of emergent properties is the notion that biological systems are hierarchically organized, with new properties at each level of the hierarchy [11], [38], [173]. According to this view, diversity is better understood if ecosystems or biota are decomposed hierarchically so that each process can be viewed as a stabilizing or disruptive factor at each level in a hierarchy of time and space scales.

During the 1950s and 1960s, the "conventional wisdom" in ecology [6] was that increased complexity within or diversity of a community lead to increased stability [39], [174]. May [175], however, came to contrasting conclusions by means of model food webs comprising a numberof interacting species. The term β_{ij} was used to measure the effect of species *j* 's density on species *i*'s rate of increase. Thus β_{ij} would be zero when there was no effect, whereas both β_{ij} and β_{ji} would be negative for two competing species, and β_{ij} would be positive and β_{ji} negative for a predator (*i*) and its prey (*j*). Setting all self-regulatory terms (β_{ii}, β_{jj}) in his randomly connected networks at -1, May distributed all other β-values at random, including a certain number of zeros. Thus, the cybernetic webs serving as ecosystem models could be described by

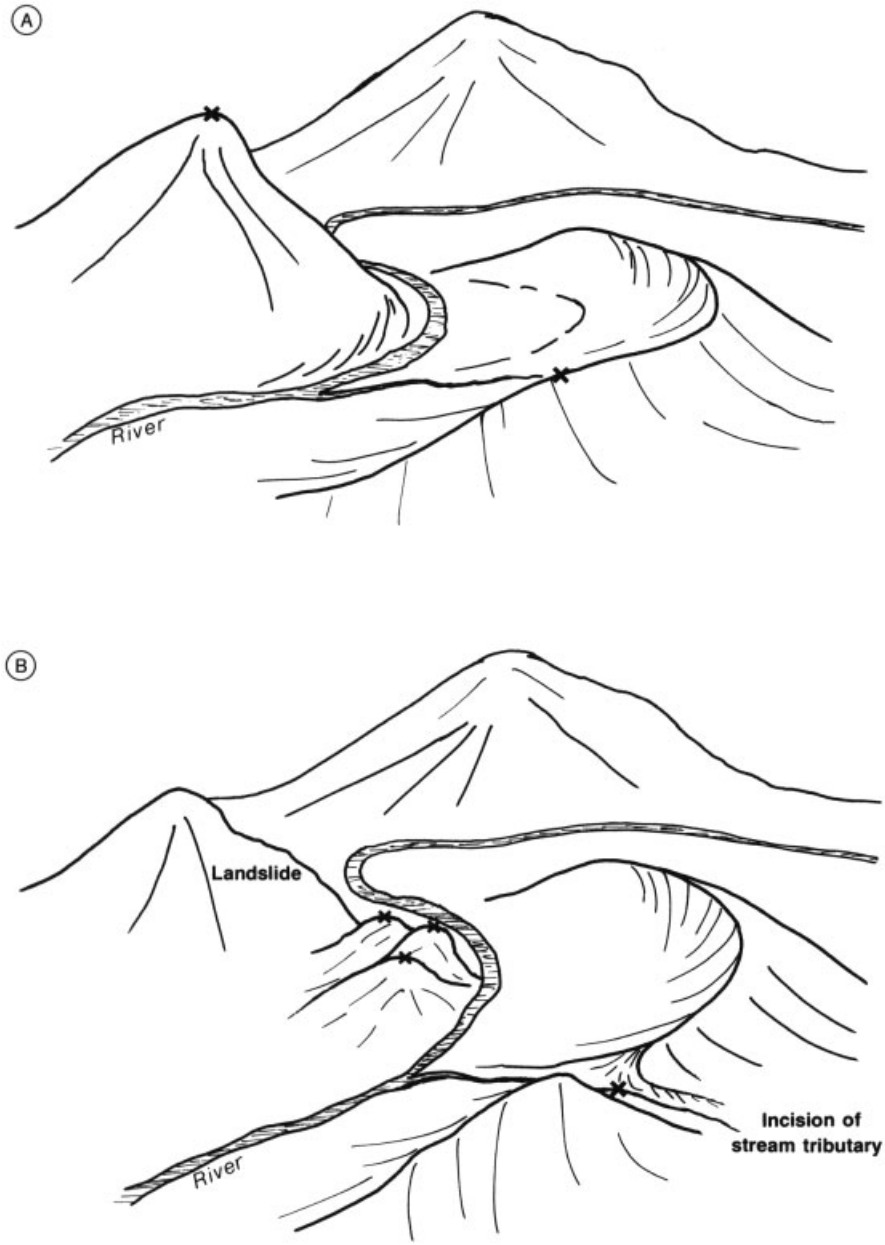

Figure 23. A) Visualization of system dynamics and stability conditions, equating phase – space characteristics with a landscape – labile equilibria (i.e., both low local and global stability); B) Illustration of global stability conditions by means of a landscape configuration, indicating different states of low local, but generally high stability

three parameters: S, the number of species; C, the connectance of the web, i.e., the fraction of all possible pairs of species interacting directly ($\beta_{ij} \neq 0$); and β, the average interaction strength, i.e., the average of nonzero β-values, regardless of sign.

Comparative analysis of these networks showed that they were likely to be stable (i.e., the populations represented would return to equilibrium after a small disturbance) only if

$$\beta\,(SC)^{1/2} < 1 \tag{25}$$

In other words, MAY's model suggests, like others, that "too rich a web connectance or too large an average interaction strength leads to instability. The larger the number of species, the more pronounced the effect." This inference is clearly in contradiction to the above conventional wisdom, but it indicates that no general, unavoidable connection exists between complexity or diversity, respectively, and community stability. Another question is, how much is MAY's result an artifact arising from particular characteristics of the model and the interpretative techniques applied [6], [37].

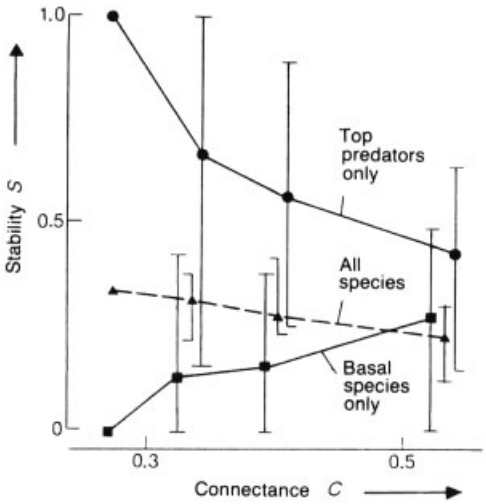

Figure 24. Species-deletion stability plotted against interspecific connectance (after [6])

The picture also alters if, instead of focusing on local stability and correspondingly minor disturbances, larger perturbations are considered or "species-deletion stability," which is particularly interesting from an ecotoxicological point of view. Under the assumption of the deletion of one species of a community owing to a persistent impact, the system is said to be species-deletion stable if all of the remaining species are retained at locally stable equilibria. In the simple six-species community of Figure 24 containing two top and two intermediate predators and two basal species (either plants or categories of dead organic matter), interaction strengths were defined at random, while connectance was varied systematically; thus, complexity could be equated with connectance. Again, stability generally decreased with increasing diversity, but this trend was reversed when basal species or elements were eliminated. However, within the simulated small subset of stable communities, resilience actually increased with complexity.

Other theoretical approaches to the stability problem are due to MCMURTRIE [176], PIMM [10], and ULANOWICZ [40], who established criteria for the stability of ecosystems under certain restrictive conditions, such as zero-sum games. In particular, they have shown that ecosystems can become unstable against weak perturbations when the complexity increases beyond a certain value, although the transition is not always sharp [177].

BACHAS and HUBERMAN [178] related the complexity of hierarchical structures, as measured by their diversity [179], to dynamical behavior. The result suggests that there might be a quantifiable relation between the diversity of a hierarchical structure and its stability. Furthermore, the results of DOREIAN [180] and of BRIAND and COHEN [181] suggest that the effective dimensionality of the space in which the interactions occur, and the overlaps in food webs, contribute to the nature of the structural representation of the community. Hence, it may be concluded that their stability will likewise be affected, which is in agreement with the results of a comparative model analysis of unstructured cooperative interactions between arbitrary species on the one hand and ecosystems with pyramidal organization on the other [182].

In returning to the original problem set out by MAY [175], this author derived a condition of stability for more general systems as a function of species diversity and the strength of interactions. The stability of such systems is determined by the behavior of the largest eigenvalue of matrices governing the response of the system to small perturbations. As a result, MAY and

coworkers showed in the case of nonhierarchical organizations how the removal of a zero-sum game condition can lead to further reduction of stability. Examining hierarchical ecologies, they demonstrated that these hierarchical structures are intrinsically more stable than unstructured ones [182].

The conflicting results among models developed so far must be considered in the light of model structure, particularly the number and composition of elements (i.e., species) simulated. In comparison to real communities with hundreds or thousands of species [136] and highly variable degrees of interaction, most models are extremely "impoverished" in species. Furthermore, the models quite often refer to randomly constructed communities, whereas real communities and ecosystems are far from being randomly constructed and normally display a complex hierarchical structure [2], [173]. Unstable communities are liable to collapse when they experience environmental conditions that reveal their instability, but the range and predictability of conditions may vary markedly from place to place. Under stable and predictable site conditions, a community will experience only limited fluctuations; thus, even a dynamically fragile one may still persist. By contrast, in a variable and largely unpredictable environment, only dynamically robust communities are likely to persist.

2.3.3. Stability-Oriented Biodiversity Analyses of Real Communities

Following MAY's approach, studies have been performed on the general importance of dynamic stability constraints by examining the relationship among S, C, and β in real communites. It follows from the inequality of Equation (25) that increases in S will lead to decreased stability unless compensatory decreases occur in C or β. Since data on interaction strengths for entire communities are unavailable, β is usually assumed to be constant. In such a case, communities with more species would retain stability only with a corresponding decline in average connectance, which means that the product SC should remain approximately constant.

REJMANEK and STARY [183] studied geographically distinct plant – aphid – parasitoid communities in Central Europe and found a decrease of C with an increase in S, with SC lying between 2 and 6. Evaluating the literature, BRIAND [184] analyzed 40 terrestrial, freshwater, and marine food webs and calculated a single value for connectance on the basis of both predator – prey and presumed competitive interactions (i.e., $C_{\max}$). Once again, connectance decreased in an approximately hyperbolic way with species number, and SC was about 7.

MCNAUGHTON's field test [185] of MAY's inference, carried out in 17 grassland stands in Tanzania's Serengeti National Park, showed that "both average interaction strength and connectance declined as species richness of the grassland increased. The correlation was somewhat stronger for interaction strength than for connectance." Hence, if the grassland communities analyzed are representative biocenoses, species-poor stands are likely to be characterized by strong interactions among species, while species interacting with many others do so relatively weakly. Diffuse competition obviously increases with the number of species; in addition, MCNAUGHTON's findings suggest that communities have a blocklike or patchy organization consisting of species interacting among themselves but only little with species in other blocks.

Table 7. Influence of nutrient addition on species richness, equitability (H/ln S), and diversity in two fields; and influence of grazing by African buffalo on species diversity of two vegetation plots (after [185])

	Control plots	Experimental plots	Statistical significance
Nutrient addition			
Species richness per $0.5\,m^2$ plot			
Species-poor plot	20.8	22.5	n.s.∗
Species-rich plot	31.0	30.8	n.s.∗
Equitability			
Species-poor plot	0.660	0.615	n.s.∗
Species-rich plot	0.793	0.740	$p < .05$
Diversity			
Species-poor plot	2.001	1.915	n.s.∗
Species-rich plot	2.722	2.532	$p < .05$
Grazing			
Species diversity			
Species-poor plot	1.069	1.357	n.s.∗
Species-rich plot	1.783	1.302	$p < .005$

∗ n.s.=not significant.

McNaughton [185] also tested the prediction that complex communities are less likely to return to their state prior to perturbation. In the first experiment, the perturbation resulted from addition of plant nutrients; in the second, it involved the action of grazing buffalo. The results summarized in Table 7 indicate that each perturbation significantly reduced the diversity of the species-rich but not the species-poor community.

According to these analyses and studies on the species diversity of tropical rain forest ecosystems in particular, a climax community is more likely to have both low fluctuations in composition (i.e., high stability) and low resilience as the homogeneity of its environment in space and time increases. Figure 25 summarizes the diversity and abundance spectra of numerous neo- and paleotropical rain forest communities as related to soil properties [2]. The diagram clearly shows that in most cases, rain forest stands on highly nutrient-depleted ferralsols, acrisols, and podzols of tropical lowlands attain diversity indices greater than 90 % of the theoretical maximum. However, a further differentiation according to local nutrient status and water budget is not possible with the data available. Nevertheless, stand 38 (Borneo) is of particular interest since it is representative of a forest that developed on a tropical bog; its exceptionally low index value allows the conjecture that its nutrient supply has fallen below a critical value, which is discussed in a broader thermodynamic context in Section 2.3.7.

In terms of May's stability criterion (Eq. 25), these findings would indicate a very low degree of connectance and average interaction strength between plants of the same species. With regard to nutrient uptake from the soil, this means a well-balanced low-grade intra- and interspecific competition involving a marked selection of K strategies (i.e., care for the offspring in the widest sense) and the formation of numerous ecological niches. Therefore, highly complex rain forest communities with their relatively constant environments appear to be more susceptible to outside, unnatural disturbances than the simpler, more robust communities of ectropical regions.

The example of Vietnam's inland forests is illuminating in this respect, since they still bear the scars of a ten-year herbicide spraying program by the U.S. Air Force during the Vietnam war. According to official U.S. figures, some 10.3 % of these forests, in addition to 36.1 % of mangrove forests, 3 % of cultivated land, and 5 % of other land, were sprayed. Thus, some 90 000 t of herbicides and antiplant chemicals was used between 1962 and 1971 [186]. Most of the spraying was with Agent Orange [i.e., a 1 : 1 mixture of (2,4-dichlorophenoxy)acetic acid (2,4-D) and (2,4,5-trichlorophenoxy)acetic acid (2,4,5-T)]; Agent White (2,4-D and picloram); and Agent Blue (cacodylic acid). Some 20 years after the spraying ended, little regrowth had occurred in areas sprayed three or four times. The fertility of the soils is rapidly reduced further if forests are replaced with grassland or bamboo, as has happened in a large part of the sprayed region. In particular, a loss of minerals and nitrogen has occurred, and in some cases a decrease in soil pH, which all inhibit diversified recolonization.

Defoliation has clearly affected the faunal assemblages, not only in the areas totally defoliated and subsequently converted into grassland but also in those regions less frequently sprayed where the plants in the upper reaches of the forest were damaged. In both areas, the number of animals and birds has dropped dramatically, and animals are also at considerable risk in some of the forest areas that have been isolated from the main body by herbicide application. Furthermore, the concern is that Agent Orange contained a high concentration of the highly toxic contaminant 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, which is both teratogenic and carcinogenic in animals, and toxic to humans as well.

In comparison to tropical lowland rain forests where species diversity and nutrient content of the soil are negatively correlated, the situation in temperate regions is different. The composition of 132 plant communities from all parts of the continental United States (except Alaska) evaluated in relation to climate and soils shows that the amount of water available for evapotranspiration is of decisive importance for species diversity. This is further corroborated by the highly significant correlation between diversity and the temporal distribution of rainfall. Conversely, the analysis of 194 plant communities in northern Germany led to the conclusion that, most frequently, high diversity coupled with medium to low proportional abundances occurs on sites

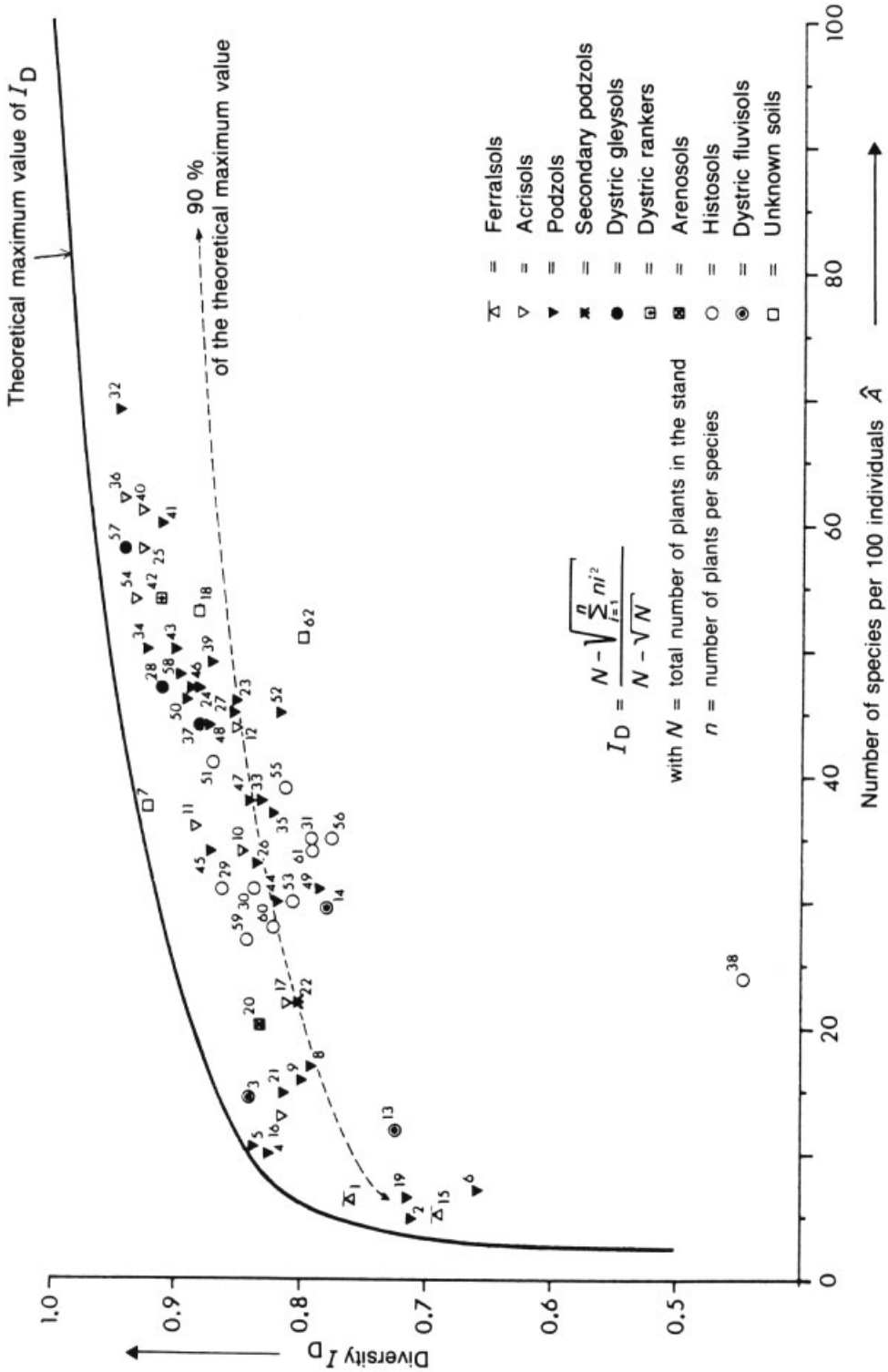

Figure 25. Diversity and abundance spectra of neo- and paleotropical rain forest communities

with high nutrient supply and medium to low soil moisture, whereas the combination of both high diversity and high abundance with high nutrient and medium to low moisture status is comparatively rare. These inverse diversity –nutrient relationships in tropical and ectropical environments, on the one hand, and the fact that soil moisture controls diversity in the United States but not in northern Germany, on the other, lead to the final question of whether a unifying interpretation of these apparently incoherent results is possible. This is given in Section 2.3.7 by considering ecological communities as dissipative structures.

2.3.4. Nondemographic Measures of Stability

The introductory remarks on stability and resilience indicate that community and ecosystem stability can be viewed from perspectives other than demographic ones. The highly different aspects of functioning (e.g., primary productivity, nutrient cycles, energy and water balance) and the composite aspects of structure such as standing-crop biomass appear to be particularly important [6].

McNaughtons's [185] studies on Serengeti grassland communities showed that species-rich associations responded to perturbations with a significant drop in species diversity, while species-poor communities did not. In terms of *primary productivity*, however, grazing as a natural perturbation reduced the standing-crop biomass in the species-poor grassland much more than that of species-rich communities. Similarly, several studies have revealed a marked influence of *food web* structure on community resilience in response to perturbations of energy and nutrient supplies. O'Neill [187] considered a three-compartment system consisting of active (photoautotrophic) plant tissue, heterotrophs, and dead organic matter, and found that the rate of change in standing crop depends on the energy transfer among these compartments. With regard to the autotrophic compartments, it depends on the input variable (net primary productivity) and two output variables (fraction consumed by heterotrophs and fractions lost to soil by litterfall) in addition to a great variety of translocation processes [85]. The rate of change of the heterotrophic compartment depends on two inputs (consumption of living plant biomass and dead organic matter) and two outputs (respiratory heat loss and defecation). Finally, as Figure 26 shows, the energetic exchange rate of soil organic matter depends on two inputs (litterfall and defecation) and two outputs (physical transport into neighboring ecosystems and consumption by heterotrophs).

On the basis of this simple configuration, and inserting real data from communities representing a pond, a spring and tundra, tropical forest, a temperate deciduous forest, and a salt marsh, O'Neill [187] subjected the models of these communities to a standard perturbation that consisted of a 10 % decrease in the initial standing-crop vegetation. The rates of recovery toward equilibrium thus monitored and plotted as a function of the *energy input* per unit standing crop of living tissue are presented in Figure 27. It shows that the flux of energy through the different subsystems has an important influence on community resilience; the higher the flux is per unit standing crop, the more quickly will the effects of perturbation be compensated. This, in turn, seems to depend in part on the relative importance of heterotrophs in the system considered. The pond as the most resilient system has a heterotrophic biomass 5.4 times that of autotrophs, reflecting the high turnover rate of phytoplankton; the least resilient tundra has a heterotroph : autotroph ratio of only 0.004. De Angelis [188] has come to analogous conclusions in terms of residence times of various nutrients in different compartments of woodland communities. Thus, a unit of nitrogen exists in soil for an average of 109 years, is taken into forest biomass for 88 years, remains in the litter for up to 5 years, and passes through the detrivore compartment in a few days. In contrast to nitrogen, calcium is a less tightly cycled element whose residence times in soil and biomass amount to only 32 and 8 years, respectively.

More detailed long-term investigations in the forest and agroecosystems of the Bornhöved study area (Schleswig-Holstein, Germany) corroborate these findings. Here, the nitrogen deposition amounted to 19.9 and 18.2 kg ha^{-1} a^{-1} in 1989 and 1990, respectively, with an ammonia fraction of 47 % (60 %) and 23 % (7 %) of organic nitrogen-compounds, while nitrate deposition yielded 6.1 and 5.9 kg nitrogen per hectare

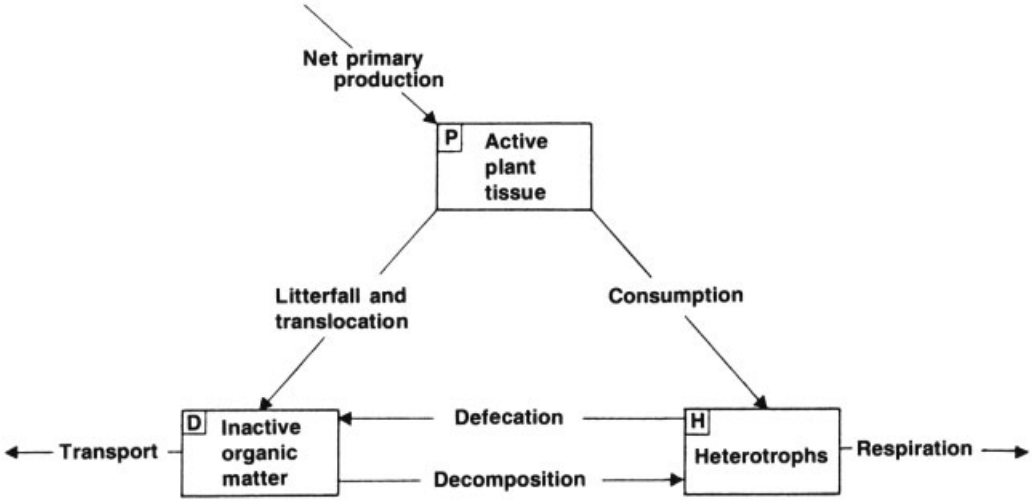

Figure 26. Three-compartment model of an ecosystem with typical energy transfers (after [187], modified)

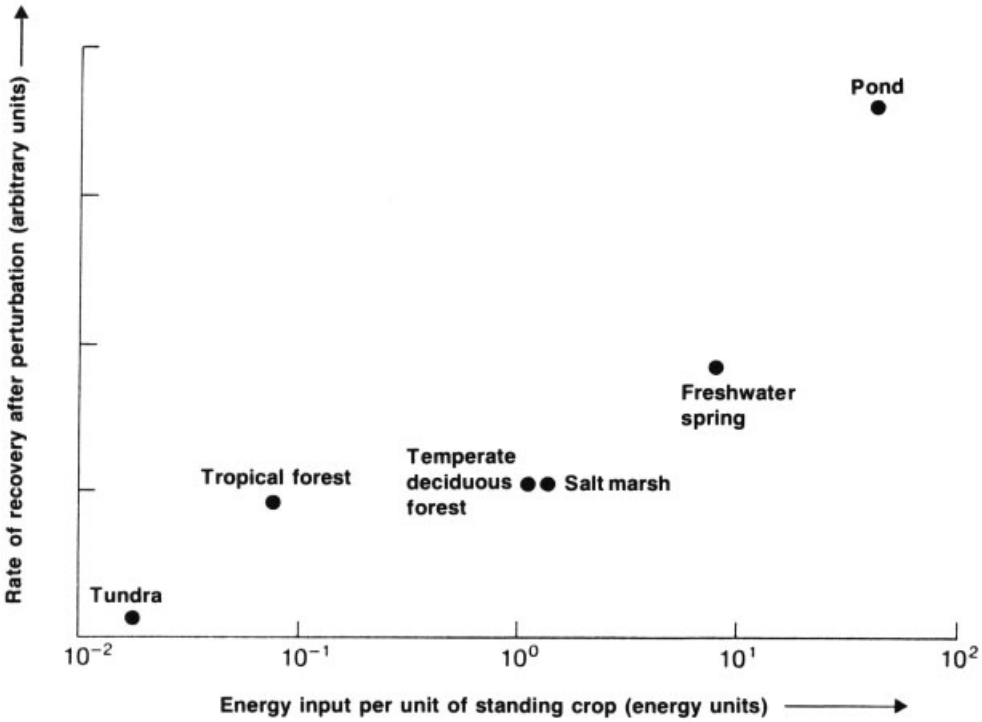

Figure 27. Recovery rate as an index of resilience after perturbation as a function of energy input per unit standing crop (after [187])

and year, respectively. The nitrogen balance of a field is summarized in Table 8.

Table 8. Nitrogen balance of a field in the Bornhöved Lake District

Process	Amount of nitrogen (measured, kg ha^{-1} a^{-1})	
	1989	1990
Initial N content	51 [a]	38
Total N input	384	222
NO_3-N	30	14
NH_4-N	118	64
N_{org}	236	144.5
Mineralization	221 [b]	215 [b]
Nitrification		
Plant uptake	206	121
N×leakage	110	129
Other N losses [c]	7	7
Residual N_{min} content	39	4.5
Balance	+59	+70

[a] Deduced value. [b] Net mineralization. [c] Denitrification and volatilization of ammonia.

The wet deposition of nitrates via canopy flow in a neighboring beech stand (*Asperulo–Fagetum*) of relatively uniform age structure (95 years) amounted to 30.3 kg nitrogen per hectare and year in 1989 and 37.5 kg in 1990, again with a major ammonia fraction. Comparably high were the deposition rates to the soils of a neighboring alder stand (*Carici elongatae–Alnetum*), where additional nitrogen fixation is due to the symbiotic activity of *Frankia*.

With regard to internal nitrogen fluxes the beech and alder stands differ considerably. In the first case, litterfall transported 78 kg nitrogen per hectare and year to the soil surface, where the chemical and microbial transformation processes induce seasonal changes in N_{min} content as illustrated by Figure 28. In the light of the site characteristics of this stand (Dystric Cambisol), the figures are indicative of rather limited nitrogen export from the community, which is further corroborated by measurements of the nitrogen fluxes in the related seepage water. In the distinctly more resilient alder stand, the nitrogen input due to litterfall amounted to 109.4 kg ha^{-1} a^{-1} in 1989 and 104.5 kg ha^{-1} a^{-1} in 1990, respectively. In comparison to the above beech stand, microbial degradation of the litter is very high in the *Alnetum*, and within 18 months almost all of the leaf tissue is degraded., Chap. 1.

2.3.5. Sensitivity of Aquatic Ecosystems

Aquatic ecosystems are natural collectors of materials from their surroundings, and water washes matter out from the land surface. Human activities increase this natural washout enormously by destroying the natural vegetation cover of the landscape due to agricultural, mining, construction, and forest-cutting activities. For organic matter of natural origin, with its contents of nutrients necessary for further growth of vegetation, a relatively closed cycle originally existed: the remains of organic matter consumed by humans and their domestic animals were returned to fields. In medieval times, increasing population concentrations in towns created local accumulations of decaying organic matter leading to the spread of infectious diseases. This problem consequently was solved by disruption of the feedback cycle. Organics with their content of elements essential for life were directed to streams, rivers, lakes, and ultimately the sea. The result was continuous deprivation of nutrients in soil and a need to subsidize them to maintain and enhance agricultural productivity. The open flow of nutrients from soil and deeper earth layers started to create unpleasant conditions by increasing aquatic production not only in streams, lakes, and ponds but also in the greatest lakes of the world and marginal seas such as the Adriatic. In addition, humans pollute water by applying various chemicals to the environment – both at home and in industry – and by extracting materials from the earth and transporting them over large distances, in part by means of waterways. Tanker catastrophes are becoming more and more common, and road transport is no less hazardous to waters.

Pollution sources are usually classified in two categories: point and nonpoint sources. Although *point sources* coming from a factory or town are technically much easier to handle, *nonpoint* sources are not only difficult to detect but still more difficult to eradicate. Typical nonpoint pollution sources originate from fertilizers and other chemicals spread on fields. Another source of water pollution is deposition of airborne pollutants, some of which reach water bodies in remote regions. The findings of elevated concentrations of DDT in Antarctica show the global scale of pollution problems.

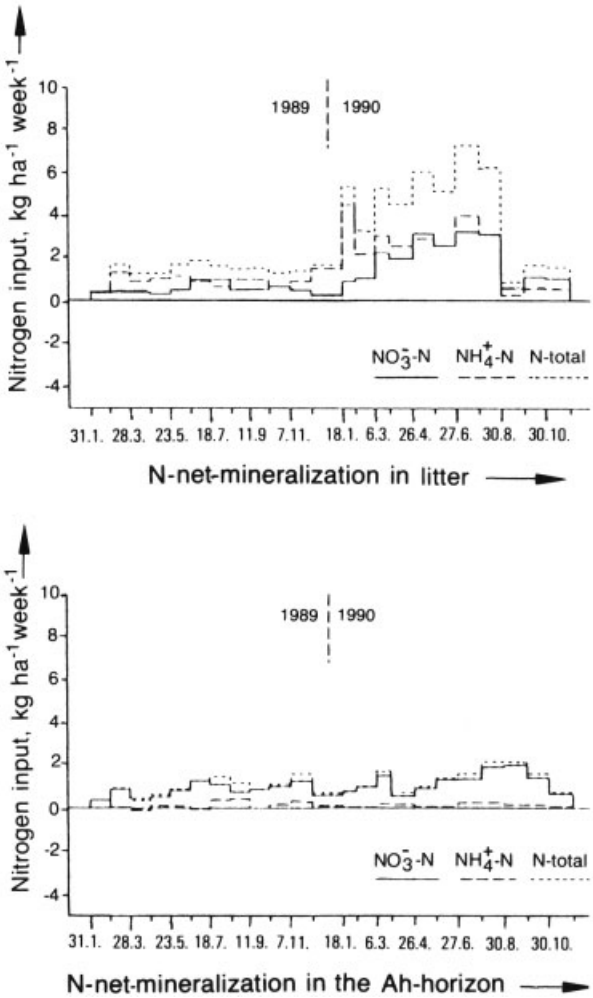

Figure 28. Seasonal variation of the N_{min} content of beech litter of a Schleswig-Holstein *Asperulo – Fagetum* stand in 1989 and 1990
Ah-horizon = part of the surface soil (→ Soil)

The sensitivity of aquatic ecosystems to disturbances differs depending on the type and size of the respective ecosystem and the nature of the disturbance. In scientific terminology, all disturbing factors are stressors and the term "sensitivity to stress" is used. Stress is usually defined as a measurable alteration of the state of organisms as induced by an environmental change that renders the individual, population, or community more vulnerable to further environmental change.

2.3.5.1. Reactions of Aquatic Ecosystems to Stress

Aquatic ecosystems under stress undergo both changes of structure and changes of function. *Changes of structure* are manifest by changes in the composition of populations of different species in the ecosystem. *Changes of function* are represented by differences in the organic matter productivity of the respective ecosystem and in the rates of release and utilization of different gases and minerals. Thus, both a sensitivity of ecosystem structure and a sensitivity of ecosystem function occur. No uniformity of

opinion exists as to whether the structure or function of ecosystems in general is more sensitive to various stressors. Several researchers feel that functional variables, especially those that are substrate limited, will always be less sensitive than structural measures because of functional redundance in the community. They suggest that any loss of functional capacity by one organism is immediately compensated by increased activity of another organism. Other scientists feel that functional capacity can be affected before compensatory mechanisms operate. This is especially likely when the compensatory mechanisms themselves are adversely affected by the stress or when they operate on a more lengthy time scale relative to the functional measure [189]. These authors have also summarized various functional reactions of ecosystems:

1) Community respiration increases
2) Productivity – respiration ratio becomes unbalanced
3) Productivity – biomass ratio increases as energy is diverted from growth and reproduction into acclimation and compensation
4) Importance of auxiliary energy increases (import becomes necessary)
5) Export of primary productivity increases
6) Nutrient turnover and nutrient loss increase
7) One-way transport increases and internal cycling decreases
8) Life span decreases; turnover of organisms increases
9) Trophic dynamics shift, food chains shorten, and functional diversity declines
10) Efficiency of resource use decreases
11) Condition declines

A well-documented example of greater sensitivity of functional variables of a terrestrial ecosystem is presented by Van Voris et al. [190], who measured structural and functional changes in small patches of a meadow subjected experimentally to cadmium stress. Functional changes such as spectral complexity and calcium export were affected at concentrations that did not affect structural measures such as the presence or absence of certain species.

In the direction of slowly and rapidly reacting ecosystems at opposite ends of the scale are pelagic lake and reservoir communities composed of very short-living (few hours to few days) species that react to any stress by their rapid appearance or disappearance, and forests with very long-living tree species that are incapable of rapid presence – absence types of reaction. Benthic communities of rivers and fish populations are intermediate: their life span is measured in months to years.

In aquatic ecosystems, many reports have shown greater relative sensitivity of structural than functional variables. For example, during his extensive comparative studies of Canadian Shield lakes, Schindler [191] found no significant changes in decomposition or nutrient cycling in whole lakes treated with acid, but species composition of phytoplankton was among the earliest indicators of change. Crossey et al. [192] found in impaired rivers that measures of production and respiration were more variable than the macroinvertebrate composition. Crumby et al. [193] studied the biological reaction of Roaring River in north-central Tennessee to stress caused by construction around the river and by inadequate agricultural practices in the watershed. Species composition changes were reflected in a general decline in numbers of intolerant species and simultaneous increase of tolerant species.

However, many reports of the greater sensitivity of functional variables of aquatic ecosystems also exist. Rodgers et al. [194] found that process rate changes were more sensitive than biomass or chlorophyll concentration in detecting the effects of diverse chemicals on periphyton in artificial streams. When dealing with enrichment, functional measures are often a good warning indicator. Uhlmann et al. [195] and Gnauck [196] summarized structural and functional changes in aquatic ecosystems and gave examples of experiments equivalent to those of Van Voris et al. [190]. Figure 29 compares the degree of variability (expressed by a relative index of instability) during a 50-d experiment with an experimental purification pond for several structural and a few functional variables. The greatest instability equivalent to the highest sensitivity is observed (with the exception of the alga *Scenedesmus obliquus*) for the biomass of individual species of organisms (index values between 0.85 and 1.45), while functional variables such as oxygen concentration, primary production, and turbidity and the summarizing biomass of functional groups of organisms such as zooplankton or phytoplankton or chlorophyll-

a levels have an index close to or below 0.5. The least sensitive appeared to be the organic carbon elimination capacity of the pond, which characterizes, from the human point of view, its most important function.

Once the stress has ceased, two possibilities exist:

1) Reversible changes are induced
2) The changes prove irreversible

The deforestation in the coastal hills of Venezuela is a drastic example of irreversible changes of an extensive ecosystem. The soil structure, seed sources, and local physical environment have changed to such a degree that forests have not returned even after the areas were abandoned by humans. Perhaps the example merely reflects an exceedingly long recovery period, and the system will eventually recover. Yet for practical purposes, these ecosystem changes are permanent. A dried lake and a watershed with homogeneously distributed amounts of toxic substances are examples of irreversible changes of aquatic ecosystems.

If the disturbance is sufficiently regular and of near-natural origin, components of the ecosystem may adapt and eventually require disturbance to maintain a normal, resilient system (tidal pools require daily exchange with the sea, pastures require mowing or grazing, many forests are adapted to periodic fires). In aquatic ecosystems, this system of reaction is seen in periodic pools, which are inhabited by animals adapted to desiccation. Their characteristic fauna will disappear when water is permanently present because they are inferior in competition with longer-lived species not needing desiccation.

For chemical stress, the recovery of aquatic ecosystems depends mainly on the degree of accumulation of the respective chemical in the environment and the rate of flow. This conclusion can be corroborated by comparing the findings in lakes with those in rivers. The recovery of eutrophic or highly polluted lakes after the sources of organic pollution and phosphorus causing lake eutrophication have been reduced is very slow. Recovery takes as much as ten to twelve years. The delayed reaction is due to the enormous quantities of decomposable organic matter and phosphorus stored in the bottom mud. Oxygen at the bottom is consumed during the decomposition of organic matter and, under such conditions, phosphorus is continuously released. This phenomenon is called internal P load, which indicates that when the external load ceases, an internal one plays a major role in continuing eutrophication. At the other end of the spectrum are rivers with high flushing. Yount et al. [197] made an overview of the recovery of rivers after a stress and found that almost full recovery may occur within about two years.

2.3.5.2. Sensitivity to Different Stressors

The sensitivity of marine and freshwater fish worldwide to recent changes in the environment was determined by means of changed species diversity [41]. The numbers of extinct, endangered, or threatened species and species of special concern represent 21 – 69 % of the total fauna in different regions of the world, with the exception of only 9 % in Latin America. The highest numbers were found in California and South Africa; low figures in Iran, Australia, Sri Lanka, and also Arkansas. Major causes of this situation are the extraction of water for various human uses, physical alteration of fish habitats (mainly by channelization, construction of weirs and reservoirs, siltation from mass erosion and degradation of wetlands), and pollution (mainly municipal and industrial point-source pollution, agricultural pollution, and acid rain). The sensitivity of aquatic ecosystems to different specific stressors depends on the type of ecosystem and the agent. For chemical agents the degree of interference with the structure and metabolism of organisms is important. However, even such components as structural macroelements of organisms may be harmful, once their concentrations in the environment exceed certain levels. This is the case, for example, with plant nutrients. When present in large excess, they decrease life functions (e.g., plant growth; see Section 2.1.4.3). Sensitivity to toxic chemicals is discussed in Section 2.4. Sensitivity to acidification as an environmental process is described in Section 2.3.5.3 together with other signs of global change.

The sensitivity of aquatic ecosystems to *ionizing radiation* is not well known. However, some reactions parallel to those observed for terrestrial ecosystems are expected to be valid.

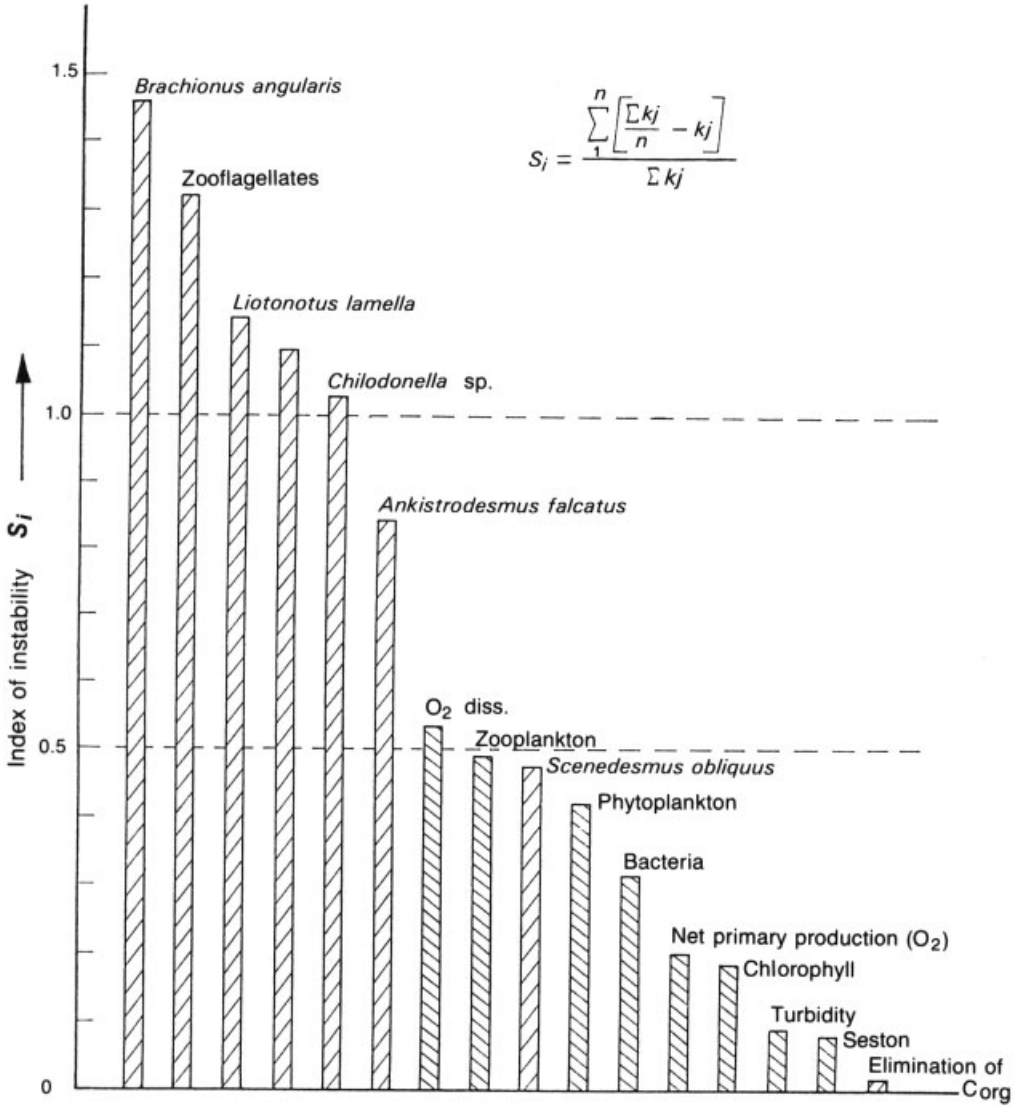

Figure 29. Comparison of the degree of variability, expressed in terms of an instability index, during a 50-d experiment with an artificial purification pond (after [195], modified)

The latter were studied in detail by WOODWELL [198]. He demonstrated that in a mixed forest – shrub – grassland exposed to different levels of ionizing radiation, the diversity of species is more sensitive than organic production. The sensitivity increases with the size of the organisms. The reason is that increased size puts greater demand on the photosynthetic mechanism for maintenance, leaving less for growth and repair. Sorting of the sensitivity was also observed on the basis of chromosome size: plants with large chromosomes were more sensitive.

Sensitivity to *communal organic pollution* is well known for both stream and lake ecosystems; this is the classical type of pollution studied by quantitative methods since the beginning of this century, the methods being continuously improved and new methods being continuously developed. From the pollution point of view, organic matter is not characterized as a chemical species but in a summarizing way, namely, the

total of easily decomposable or refractory matter. The difference between the two categories is arbitrary: easily degradable matter is decomposed by the natural population of microorganisms within five days' incubation. In streams, systematic changes of the composition of benthos are observed with certain amounts of organic pollution. This is the basis for the bioindicator approach to detect the degree of organic pollution. However, geographic differences in the benthic fauna of different regions; dependence on stream characteristics such as size, rate of flow, and bottom type; and combination of pollution types all complicate the evaluation.

2.3.5.3. Sensitivity to "Global Change"

One of the effects anticipated by part of the scientific community to cause major changes in world climate and consequently local weather and food production within the next decades is "global change." Since the 1940s, compounds such as CO_2, chlorofluorocarbons (CCl_2F_2 and CCl_3F), methane (CH_4), and dinitrogen monoxide (N_2O) have displayed notable and deleterious increases or decreases in the case of tropospheric ozone (O_3). These gases (except O_3) are classified as "greenhouse gases" because they have a greenhouse effect on earth's atmosphere.

About a hundred years ago, CO_2 values of 280 ppm (mL/m^3) were measured in the atmosphere; (in the 1990s), concentrations are about 25 % higher. Carbon is both fixed and released by oceans, biota, and human activities. Approximately 30 % of the global annual carbon fixation by plants occurs in the seas, which also form an integral part of the nitrogen flux of the earth. An estimate for 1980 gave the following figures for the fluxes of carbon among atmosphere, ocean, biosphere, and human activities [199]: 78×10^{12} kg C per year (flux ocean $\leftrightarrow$ atmosphere), 65×10^{12} kg C (flux biosphere $\leftrightarrow$ atmosphere), and 5×10^{12} kg C (flux due to human activities, mainly burning of fossil fuels and land clearance $\leftrightarrow$ atmosphere). Although the contribution by human activities is far less than the natural fluxes, it shifts the total balance significantly.

In the United States, the sources of CO_2 from human activities were estimated for 1987 as follows [42]: 11 % residential, 24 % industry, 30 % transportation, and 35 % electric and power generation. Industry is therefore an important CO_2 source, even when the significant portions of electric power generation and transport associated with industry are neglected.

The climate has changed throughout the last few thousand years due to natural causes; the influence of human activities superimposed on these variations has also contributed significantly. Scientific documentation demonstrates that the overall rise in CO_2 has already had marked effects on the biosphere. Comparison of different recent plants with herbarium specimens collected many years ago disclosed that the leaf area and the nitrogen content of leaves in 15 species of trees, shrubs, and herbs decreased over the last 240 years. Also, a systematic increase in the growth rate of bristlecone pines and a decrease in stomatal density of the leaves of plants over the last 200 years was observed [200]. Although no aquatic plants were included in the studies, they are expected to be subject to similar effects. The reported changes may seem unimportant; however, this may be only the easily observable tip of the iceberg. More relevant parameters cannot be compared with those of earlier times when no scientific measurements existed. The detection of changes increases the credibility of predictions by demonstrating that biological changes have already taken place and are not merely scientists' anticipation.

Despite many uncertainties in predictions, a certain consensus exists among part of the scientific community that the increase in CO_2 and the other greenhouse gases due to human activities will warm the climate by the middle of the next century between 1 and 5 °C ("global warming"). Greater temperature increases are expected at high latitudes, especially in winter, with consequences for ice melting and sea-level rise. Consensus also exists that changes of this magnitude in temperature will be accompanied by substantial regional anomalies in temperature, rainfall, evaporation, cloudiness, windiness, drought, and so forth that could have substantial impact on agriculture, water supply, human and animal health, biological diversity, energy demand, and sea level [199], [201].

The changes expected in aquatic environments due to these increasing quantities of greenhouse gases are of both a physical and a chemical nature. Physical changes mainly in-

volve global warming and an increased level of ultraviolet radiation at the earth's surface. Ultraviolet radiation in its present spectral composition is readily adsorbed in water, and only organisms living in the uppermost (few centimeters) water layers can be damaged. However, alterations in spectral composition of ultraviolet radiation expected as a consequence of global changes can markedly increase the layers affected.

One indirect effect of global warming on streams is expected in connection with increased variability in stream flow, associated with increased rates of sediment transport and sedimentation. Sediments are expected to fill in some pools and generally cover surfaces. However, streams may be flushed of material periodically. These conditions would favor invertebrates with shorter life cycles adapted to softer sediments, while larger invertebrates could be adversely affected by these conditions because they graze rock surfaces [43].

Indirect effects on aquatic ecosystems are also expected due to the high sensitivity of forest ecosystems to environmental variables. The sensitivity of forests to climate changes was estimated from changes observed during the geological past and from some geographical dependencies observed more recently [202]. The results indicate that during global change, shifts will occur in the geographical ranges of tree species, along with changes in forest productivity. Because of the interrelation of soil nutrients, soil moisture, groundwater, and tree growth, changes in forest ecosystems will have major consequences for aquatic ecosystems, primarily rivers. River flow will be changed and, with it, leaching of minerals and organics from soil will occur to varying extents, with consequences for river inhabitants [44]. Additional indirect effects will be caused by changes in agricultural production.

The probable effect of the expected climate warming on fish of the Great Lakes Basin was estimated by Meisner et al. [203]. They concluded that no full extirpation of species is to be expected, but the populations of the most valued fish, salmonids and coregonids, will decrease, whereas the species less valued by fishermen (cyprinids, esocids, centarchids, and ictalurids) will probably increase. Thus, a decrease of fishery yields of preferred species is expected.

Other global changes in aquatic ecosystems are already observed locally in acidified regions. They result from direct and indirect chemical effects caused by increasing CO_2 levels. Doubling of atmospheric CO_2 levels, which is expected to occur at the beginning of the next century, is supposed to decrease the pH of rainwater by about 0.4 unit, which means a pH drop in some localities to 4.0. The lower pH will increase the solubility of carbonate minerals and increase the calcium, magnesium, bicarbonate, and sulfate content of rivers draining carbonate terrain. The greatest secondary effects on organisms are caused by the increased solubility of aluminum at low pH. Aluminum would be released, predominantly in bedrock terrain. The speciation of aluminum in river water is important in determining the toxicity of elevated aluminum levels as well as the acidity of soil solutions. Inorganically complexed or free aluminum is more toxic than organically complexed aluminum. Lakes with present pH levels of 6.5 are expected to become acidified to pH <5, with considerable effects on inhabiting organisms. Major impoverishment of lakes due to acidification has already occurred in northern Europe and Canada, and in mountainous lakes of more southern localities [45].

2.3.6. Sensitivity of Ecotones

Like communities and ecosystems, ecotones also differ in physical stability or resistance to changes (resilience) caused by disturbances. In contrast to lotic ecotones, adjacent terrestrial ecosystems and ecotones may be viewed as occupying distinctly more stable positions in a landscape where natural disturbances due to geomorphic processes are less frequent than the time for patch adjustment to them (see Sections 2.3.5 and 2.3.7).

In distinguishing reactions, recovery and persistence phases, and recurrence intervals as characteristic parameters of system response to physical disturbance [46], the recovery and disturbance recurrence intervals are of particular ecological importance. The ratio of the recovery to disturbance recurrence intervals R/D suggests differences in recovery or sensitivity characteristics of both ecotones and ecosystems. For unstable systems, the R/D ratios exceed 1. Typically,

this is the case for predominantly transient systems such as many *lotic ecotones* [162]. Normally, deposition and erosion interact with riparian vegetation at varying intensities and frequencies so that communities are kept in the early stages of succession. A decrease in seasonal flooding of the riparian vegetation results in changes via several possible successional pathways, which may lead to the formation of purely terrestrial communities when the system is no longer influenced by flooding.

Examples of this type support the assertion of NAIMAN et al. [155] that a paradox of modeling interfaces is predicting when and where ecotone parameters become unpredictable relative to adjacent homogeneous sites. Chaos theory may be useful in dealing with this uncertainty because ecotones are not simple averages of adjacent systems; the spatiotemporal pattern of environmental factors and the interactions between patches are as important in determining ecotone dynamics on different scales as are the characteristics of adjacent patches.

2.3.7. Stability and Resilience of Ecosystems in Evolution

On a very general level, the apparently unusual – because inverse – relationship between the nutrient status of soil and the diversity of related tropical rain forest communities described in Section 2.3.3 may be interpreted in terms of biological thermodynamics [169]. Biotic communities are open systems in exchange of energy and matter with their environment; their entropy production dS thus comprises the terms $\mathrm{d_i}S$ and $\mathrm{d_e}S$, where $\mathrm{d_i}S$ denotes the entropy production within the system while $\mathrm{d_e}S$ is a flux term describing entropy "export" into the environment:

$$dS = \mathrm{d_i}S + \mathrm{d_e}S \tag{26}$$

The term $\mathrm{d_i}S$ can only be >0, but $\mathrm{d_e}S$ can also be negative. By identifying entropy with disorder, Equation (26) shows that an isolated system can evolve only toward greater disorder. For an open system, however, the "competition" between $\mathrm{d_e}S$ and $\mathrm{d_i}S$ permits the system, subject to certain boundary conditions, to adopt new states or structures. These are stationary if

$$dS = 0 \tag{27}$$

or

$$d_e S = -\mathrm{d_i}S < 0 \tag{28}$$

$\mathrm{d_i}S$ can be expressed in terms of thermodynamic "forces" X_i, and rates of irreversible phenomena J_i [165]; X_i may be gradients of temperature or concentration; the corresponding "rates" are then heat flux and chemical reaction rate. Hence,

$$\frac{d_i S}{\mathrm{d}t} = \sum_{i=1}^{n} J_i X_i \tag{29}$$

Around equilibrium, the relationship between fluxes and forces is linear

$$J_i = \sum_{j=1}^{m} L_{ij} X_j \tag{30}$$

where L_i represents specific coefficients (e.g., coefficients of thermal conductivity or diffusion).

Provided the reservoirs of energy and matter in the environment of the open system are sufficiently large to remain essentially unchanged, the system can tend to a nonequilibrium stationary state far beyond the domain of linear thermodynamics. This state may be associated with dissipative structures [204] (i.e., structures resulting from a dissipation of energy rather than from conservative molecular forces). In considering phytocenoses from the viewpoint of stationary dissipative structures, the relationship of $\mathrm{d_i}S$ and $\mathrm{d_e}S$ as expressed in Equation (28) and the specific boundary conditions controlling entropy production and flux rates appear to be particularly important. A consequence of Equation (28) is that, thermodynamically speaking, stability or the capacity to maintain a nonequilibrium steady state is coupled with a (relative) minimum of total entropy production dS. Clearly this can be accomplished by either minimizing $\mathrm{d_i}S$ or maximizing $\mathrm{d_e}S$ or by a combination of both strategies.

Concentration processes involved in the normal metabolic activities of living systems play an important role in this connection, as can be seen from

$$\Delta G^{\circ} = R \cdot T \cdot \ln \frac{C_2}{C_1} \tag{31}$$

where ΔG° is the difference in standard free energy, $R = 8.31\ \mathrm{J\,mol^{-1}\,K^{-1}}$, T = temperature in kelvin, and C_2, C_1 are higher and lower thermodynamic concentrations, respectively. Changes in concentration are a physical prerequisite for the production of a great many compounds, and an absolutely cogent one if substances are produced whose free energy is higher than that of the corresponding "raw materials."

On the nutrient-depleted soils of Amazonia (Fig. 25), for example, forest stands with their mosaic-like structure owing to the patchiness of rejuvenation processes (see Section 2.1.1) developed specific adaptations to facilitate these concentration processes. A dense root mat is formed over the soil in intimate contact with litter, with root tips growing upward into the fallen litter. Mycorrhizal associations, as the most common plant root – soil microorganism symbioses, benefit their host plants by effectively increasing the absorptive surface area of the root system, thus providing for increased uptake of nutrients. The dense fabric of predominantly fine roots also plays an important part in exchange and adsorption of nutrients from throughfall water. In addition, the structure of the foliage favors the use of nutrients during a long active life, the retransport of certain nutrients before leaf shedding, and a high polyphenol content and coriaceous nature, which both reduce herbivory. Another factor of great importance is the multilayered structure of the forest and the activities of epiphytes and microorganisms on much of the exposed surfaces, which together form highly efficient filtering systems scavenging nutrients from rainwater. Finally, biological nitrogen fixation in the root – humus – soil interface appears to play an important role in the nutrient budget [205]. Taking into account that even in the simplest cells the normal metabolic pathways imply several thousand complex chemical reactions that must be coordinated by means of an extremely sophisticated functional network, means that hierarchical order in both functional and spatiotemporal respects constitutes a further and most powerful negentropic factor. It characterizes every living system from the submicroscopic level to gigantic rain forest biomes such as the Amazonian Hylaea.

The effectiveness of these negentropic processes is further enhanced by most efficient entropy fluxes related to the transpiration and nocturnal respiration of plants. The molal entropy of H_2O increases from $63\ \mathrm{J\,mol^{-1}\,K^{-1}}$ (liquid) to $189\ \mathrm{J\,mol^{-1}\,K^{-1}}$ (gas) in the course of evaporation, and CO_2 has a molal entropy of $214\ \mathrm{J\,mol^{-1}\,K^{-1}}$. Consequently, the reverse process, the photosynthetic fixation of CO_2, is of comparable importance for the negentropy balance in the light of Equations (26) and (28).

As a consequence of these mechanisms, the results of the above comparative diversity analyses may be given a unified interpretation:

1) Species diversity is not a monotonous function of the nutrient status of soils. If soils form on basic or intermediate parent materials, pedogenic nutrient supply and species diversity may both increase during periods of $10^3 - 10^4$ (10^5) years, provided the water and energy factors are not limiting. In Figure 30, the period of ascending evolution is termed phase I. Shorter fluctuations of variable but generally decreasing amplitude are likely to be superimposed on this long-term trend. In phase II, which can last for several hundred thousand years, soils degrade in regard to nutrient supply, but species diversity keeps increasing for negentropic reasons (see the tropical rain forest stands in comparison to the ectoprical associations described in Section 2.3.3). Clearly such an evolution toward and including the following phase requires a sufficient degree of climatic and geomorphic stability (as has been the case in central Amazonia and considerable parts of Malesia), or else the continuity of pedogenesis will be interrupted by truncation processes that rejuvenate the soils (as happened in the major part of the Congo Basin). Phase III marks the eventual and comparatively acute decrease in diversity of phytocenoses once the nutrient status of soils has fallen below a critical level.
2) Spatial heterogeneity as a characteristic feature of the sites during phase I results in variability in the number of species and plant individuals. With this variability the associations can simultaneously retain genetic and behavioral types that can maintain their existence in low populations together with others capitalizing on opportunities for dramatic increase. The more homogeneous the environment in space and time (i.e., during phase II,

and probably also phase III), the more likely is the system to have low fluctuations (i.e., high stability). Tropical rain forests represent climatically buffered and (largely) self-contained systems with relatively low (natural) variability on highly impoverished soils of the ferralsol, acrisol, and podzol classes. Temporary external disturbances are consequently likely to affect a proportionally greater number of species than in phase I populations, where the interspecific differences in ecological potency are usually distinctly higher. Hence, resilience of mature biocenoses is low during phase II. In contrast, biocenoses may be highly resilient in phase III of evolution, although unstable.

The greatly reduced resilience of phase II stands is particularly well illustrated by the influence of clearing activities on tropical forests. The soil, with its low retention capacity, is irrevocably impoverished after a brief period of fertilization due to sudden release of nutrients in the biomass. The brief period of fertility is used by shifting cultivators to produce crops, but even this system, developed by small native populations "in close contact with the environment" [206], can induce severe losses if forced to fit a market-oriented economy or to feed larger populations [207], [208].

2.3.8. Appraisal

The analysis of real communities and the comparative assessment of ecological models [2], [37] show that real communities are not those homogeneous and temporally invariant systems described by simple Lotka – Volterra equations and exemplified by most laboratory microcosms. Population dynamics of real-world ecosystems are spatially distributed in general, associated with a marked temporal variation, providing a multitude of ways in which the probability of coexistence is enhanced and biodiversity increased. Coexistence under stochastic, nonequilibrium conditions as described by patch-dynamics models at different scales can be just as strong and stable as that occurring under a deterministic, niche-differentiation model [209].

In a single patch, species extinction can occur as a result of competitive exclusion, overexploitation, and other destabilizing interspecific interactions or be due to environmental instability (e.g., unpredictable disturbances or changes in conditions).

De Angelis and Waterhouse [210] and Remmert [211] have emphasized that the macroscale integration of unstable patches that are out of phase with each other can lead to persistent species-rich communities.

An important parallelism seems to exist between the properties of a community and the properties of its component populations, which will be subject to a (relatively) high degree of K selection in stable environments, while r selection (selection in the direction of high numbers of offspring) predominates in variable environments. The K-selected populations with their high competitive capacity, high inherent survivorship, and low reproductive rates are normally resistant to disturbances but, once perturbed, have little possibility of recovery (i.e., low resilience). The r-selected populations, by contrast, have less resistance but distinctly higher resilience. This inverse relationship between resistance and resilience is of particular importance in the ecotoxicological context (Section 3.2).

The above stability considerations mean that food chains should be shorter in "fluctuating" environments, since only the most resilient food webs would be expected to persist and short food chains are more resilient. An analysis of Briand's findings [184], however, shows that further critical tests of these logical ideas are required because the average maximal food chain lengths for 40 communities from both unstable and "constant" environments did not differ significantly. In conclusion, most communities are probably organized by a temporally and spatially varying mixture of " forces" – namely, competition, predation, and disturbances, with competition and predation being presumably less important in more disturbed environments [2], [212], [213]. Consequently, no single stability or sensitivity measure exists for a community. Stability varies with the aspect of the community or ecosystem under study and the nature of the disturbance. Without pertinent semantic specifications relating to defined temporal and spatial scales (duration and dimensionality of observations), indicator variables and the nature

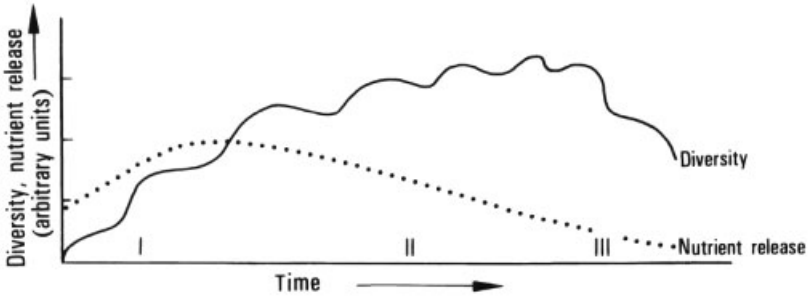

Figure 30. Generalized model of long-term ecosystem evolution in reaction to nutrient supply of soil

of perturbation statements about system stability would be meaningless in a general ecological and a more specific ecotoxicological context.

2.4. Exposure and Effect Criteria of Chemicals

Methodologies for estimating fluxes, transformations, and resultant environmental concentrations of chemicals depend on the knowledge of how rapidly the chemicals are discharged to a specific compartment in the environment and how rapidly they are transformed or removed by physical, chemical, and biological processes. Apart from the specific structure of the receiving compartment, particularly its physical and chemical characteristics, these processes are largely controlled by the inherent molecular properties of the chemicals released. These may be appropriately classified as sets of equilibrium and kinetic constants that define the exposure potential, and a complementary set of data that characterizes the effect potential, i.e., the chemical, biotic, and particularly, ecotoxicological effects of a substance.

In the following sections the essential exposure and effect criteria are described. Other chemical and physical properties such as flash points, flammability limits, and autoignition temperature are omitted because they are not of direct concern to the environmentalist. A comprehensive review of these and other dangerous properties of chemicals is given in [214]. With regard to the following quantitative specifications, however, note that chemicals are never 100 % pure and that the very nature and quantity of the impurities can have a significant impact on many environmental qualities. In particular, the following parameters are very sensitive to the presence of impurities: water solubility, odor characteristics and threshold values, biological oxygen demand, and toxicity [47].

2.4.1. Equilibrium Constants for Chemical Distribution

Density, Molar Mass, and Structure. The density of a substance, particularly in comparison to that of water or air, is of special environmental importance since it largely controls dispersion processes. Hydrocarbons are usually lighter than water. Figure 31 illustrates the relationships of molar mass and density.

Melting and Boiling Points. Normally, melting involves an increase in volume by about 10 %; exceptions are water, bismuth, and gallium. According to the sign of this change in volume, melting temperature rises or falls by 0.01 – 0.1 K/bar, which means that, under environmental conditions, pressure has little effect on melting point.

The melting and boiling points of the members of homologous series of hydrocarbons increase with increasing molar mass. When comparing organic compounds containing different halogens, however, the mass effect can be (partly) counterbalanced by polarization effects (e.g., in the case of CH_3F and CH_3Br). Substitution of another atom or group for hydrogen in an alkane results in an elevation of the boiling point. Thus, alkyl halides, alcohols, aldehydes, ketones, and acids boil at higher temperatures than ("elementary") hydrocarbons with the same carbon skeleton. The introduction of groups promoting association brings about a marked rise in boil-

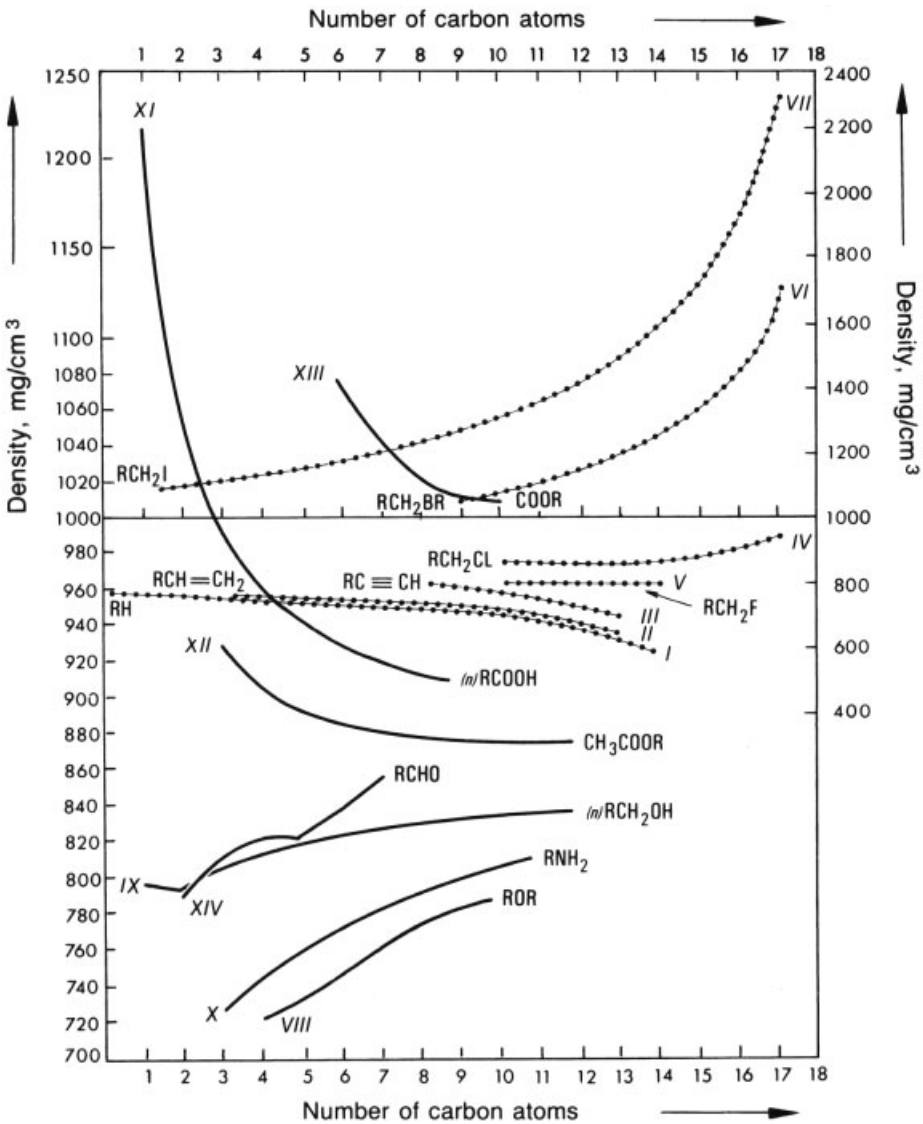

Figure 31. Relationships of molar mass and density (after [47], modified)
Dotted lines *I – VII* refer to right scale, lines *VIII – XIV* to left-hand scale

ing point, which is especially pronounced in alcohols and acids since hydrogen bonds can be formed [47].

Vapor Pressure, Vapor Density, and Volatility. The *vapor pressure* of a liquid or solid is the pressure exerted by the gas in equilibrium with the liquid or solid at a given temperature. The evaporative loss of a substance, volatilization, is a very important source of material for airborne transport and subsequent transformation and deposition processes. Vapor pressure data permit, in combination with solubility data, calculations of rates of evaporation of dissolved compounds from water by using Henry's law constant H, which is normally defined as

$$H = PS^{-1} \tag{32}$$

where P is the equilibrium vapor pressure of the pure chemical and S its saturation concentration or solubility at a fixed temperature.

Vapor density, compared to the density of air (=1), indicates whether a gas will be transported along the ground, possibly subjecting ecosystems to high exposure, or whether it will disperse rapidly. Vapor density is related to equilibrium vapor pressure through the fundamental state equation for a gas:

$$PV = nRT \tag{33}$$

If the mass of the substance and its molar mass are substituted for the number of moles n, this yields for the vapor density Vd [47]:

$$Vd = PM\,(RT)^{-1} \tag{34}$$

where P is the equilibrium vapor pressure (bar), $R = 8.31\ \mathrm{J\,mol^{-1}\,K^{-1}}$, M = molar mass (g/mol), and T = absolute temperature (K).

Volatilization of chemicals from water is an important transport process for a number of substances that have low water solubility and low polarity. Despite their very low vapor pressure, many compounds can volatilize rapidly because of their very high activity coefficients in solution, and by using the mathematical approach of Mackay and Paterson [215], [216] the rate constant for volatilization from water is given by

$$K_{\mathrm{vw}} = \frac{A}{V}\left(\frac{1}{K_{\mathrm{L}}} + \frac{RT}{HK_{\mathrm{G}}}\right)^{-1} \tag{35}$$

where A is the evaporating surface ($\mathrm{cm^2}$), V the liquid volume ($\mathrm{cm^3}$), H Henry's law constant (bar L/mol), K_{L} the liquid film mass-transfer coefficient (cm/h), K_{G} the gas film mass-transfer coefficient (cm/h), R the gas constant ($8.31\ \mathrm{J\,mol^{-1}\,K^{-1}}$), and T the temperature (K).

Volatilization processes from soil surfaces are complicated by variable contributions due to volatilization of the chemical from the water at the surface, evaporation of water itself, and the related capillary rise, which transports more water and dissolved chemical to the surface [217].

Water Solubility. Water solubility is an important parameter for the environmental assessment of all solid or liquid chemicals. In general, molecules with highly polar bonds will be relatively strongly attracted to one another; consequently, water and alcohol are potent solvents for ionic compounds such as salts, as well as polar nonionic molecules such as sugar. In contrast, nonpolar molecules such as tetrachloromethane, benzene, or gasoline are more weakly attracted to one another. They mix easily and are good solvents for nonpolar molecules, such as fats, grease, and paraffins, which are not soluble in water. Some solids are so firmly bonded that they are not significantly soluble in any common solvents except those with which they react chemically, whereby the nature of the molecule being dissolved is altered. Quartz is of this type: its solubility in pure water is of the order of 5 mg/L, but if the water contains a 0.01 M concentration of orthodiphenols, quartz is transformed into soluble esters and the solubility is increased to 40 mg/L [218].

The composition of natural waters varies considerably in space and time (see, e.g., [48]). Variables such as pH; water hardness; and concentration of cations, anions, and organic substances such as gelatin or fulvic and humic acids, all affect the solubility of chemicals in water.

Solubility of Mixtures. Mixtures of compounds behave differently from the individual compounds when brought in contact with water because each component partitions between the aqueous phase and the mixture. Compounds with a high water solubility tend to move into the aqueous phase, while the less soluble components have a greater tendency to remain in the mixture phase. Thus, the fractional composition of the water-soluble fraction differs from the original composition of the mixture, and the concentrations of the individual components of the water-soluble fraction are generally lower than their maximum solubilities [219].

***n*-Octanol – Water Partition Coefficient.** The n-octanol – water partition coefficient (K_{OW}) is defined as the ratio of the equilibrium concentration C of a dissolved substance in the two-phase system n-octanol and water (two solvents that are largely immiscible):

$$K_{\mathrm{OW}} = \frac{C_{\mathrm{octanol}}}{C_{\mathrm{water}}} \tag{36}$$

Usually the decadic logarithm of K_{OW} is given ($\log K_{\mathrm{OW}}$). Because K_{OW} is a dependent variable with regard to water solubility simultaneous

use of both in structure – activity relationships or multivariate classification procedures will give spurious correlations. Experimental determination of both the aqueous solubility and the *n*-octanol – water partition coefficient led to the following regression [220]:

$$\log K_{\mathrm{OW}} = 5.00 - 0.670 \log S \quad (37)$$

where S is the solubility in water (μmol/L). When solubility is expressed in mg/L, Equation (37) becomes

$$\log K_{\mathrm{OW}} = 4.5 - 0.75 \log S \quad (38)$$

The regression covers many classes of both liquid and solid compounds with highly different polarities, from hydrocarbons and organic halides to aromatic acids, pesticides, and polychlorinated biphenyls (PCBs). The corresponding correlation coefficient is 0.97, which normally allows an estimation within one order of magnitude of the partition coefficient of a given chemical from its solubility in water. However, the scatter increases considerably for solubilities greater than 100 mg/L, as the example of L-tyrosine shows, where the experimental value of $\log K_{\mathrm{OW}}$ is -2.26 while calculations using Equation (37) yield values of +2.7 or +2.5.

Furthermore, whether the above equations would also apply to chemicals with high dissociation constants, such as salts and strong acids or bases, is questionable since their activities in solution cannot be approximated by their concentration. Basically, the same restriction holds for aliphatic acids or bases whose partition coefficients can vary drastically with changes in pH [47], [221]. Nevertheless, the *n*-octanol – water partition coefficient has proved useful as a simple means of predicting biological uptake, lipophilic storage, and biomagnification factors (see Sections 2.4.2, 3.3.5 and Fig. 57), whereas extrapolations of the multifactorial soil adsorption processes are much more difficult.

2.4.2. Kinetic Constants for Environmental Processes

Every chemical reaction takes place at a definite rate depending on the specific conditions under which it occurs. These boundary conditions are the concentrations or pressure of the reacting substances, temperature, presence or absence of a catalytic agent, and radiation. Chemical kinetics deals with the rates and mechanisms of these reactions whose understanding is rendered difficult by the fact that both time and intermediate products are involved, compared to the subject of chemical equilibrium, which is concerned only with the initial and final states of reactants. Because of the complexity of these subjects, the following presentation is limited to determination of the essential rate constants for environmental processes from structure alone by using structure – activity relationships (see also → Toxicology, Chap. 2.).

2.4.2.1. Hydrolysis

Hydrolysis of compounds usually results in the introduction of a hydroxyl function and is most commonly associated with the elimination of a leaving group X:

$$\mathrm{RX} + \mathrm{H_2O} \rightarrow \mathrm{ROH} + \mathrm{HX} \quad (39.1)$$

In water, the reaction is catalyzed mainly by H^+ and OH^- ions; whereas in moist soil, loosely complexed metal ions such as copper or calcium may also constitute important catalysts for the hydrolysis of several types of chemical structures. In addition, sorption (→ Soil) of the chemical may increase its reactivity toward hydronium or hydroxyl ions [222], [223].

Mabey and Mill [224] reviewed kinetic data for hydrolysis of a variety of organic compounds in aquatic systems in relation to the chemical characteristics of most freshwater systems. Quantitative structure – activity relationships, used as predictive test methods to develop the essential kinetic data, generally take the form of linear free-energy relationships such as the Hammett and Taft

$$\log K_x / K_{\mathrm{o}} = \varrho\sigma \quad (39.2)$$

or Brønsted

$$\log K_x / K_{\mathrm{o}} = m\mathrm{p}K_{\mathrm{a}} + C \quad (39.3)$$

approaches, where ϱ is a constant for a given reaction under a given set of conditions, and σ is the Hammett constant, characteristic of a certain substituent; K_{o} and K_x are the rate constants for

the unsubstituted and substituted structures, respectively, while the dissociation constant pK_a refers specifically to the acidity of the leaving group.

Highly variable hydrolysis reactions have been observed in soil and sediment. This appears to be the result of unusual pH relationships at the surface of soil constituents, possible occurrence of metal-ion-catalyzed processes, and general acid- and base-catalyzed reactions, possibly involving phenolic, amine, or sulfide groups in soils [223]. In some cases (e.g., in the presence of clay minerals), hydrolysis rates of organic chemicals are accelerated considerably compared to bulk solution, while adsorption to humic components frequently impedes hydrolysis. A detailed understanding of the mechanisms involved is limited, however, and structure – activity relationships appear to be available for only a few compounds.

2.4.2.2. Photolysis

Direct Photolysis. Natural direct photolysis results from absorption of photons from the solar spectrum above 290 nm by the chemical. The specific absorption rate ($K_{a,\lambda}$) is a measure of the overlap between the solar spectrum and the electronic absorption spectrum of the compound (i.e., an inherent property of a photoreactive chemical) [225]:

$$K_{a,\lambda} = 2.303 E_0(\lambda)\,\varepsilon_\lambda \tag{40}$$

where $E_0(\lambda)$ is the scalar irradiance and ε_λ the molar absorptivity at wavelength λ.

Since pollutant concentrations are low under most environmental conditions, photolysis can be described by a first-order expression; i.e., the rate constant of direct photolysis k_d is directly proportional to pollutant concentration. From Equation (40), then

$$k_d = 2.303 \Phi_d \int E_0(\lambda)\,\varepsilon\, d\lambda \tag{41}$$

where integration is over the range of $Eo(\lambda)$, which is absorbed by the compound, and Φd denotes the quantum yield, which may attain a maximum value of 1. However, rapid direct photolysis can occur even with compounds that react with very low quantum efficiencies [225].

Since more than one primary reaction may be involved in a photochemical reaction, the primary quantum yield may be defined with respect to a specific reaction. Moreover, the sum of the quantum yields of all primary reactions, including deactivation, must be unity. The overall quantum yield, which is the quantity most often determined, on the other hand, does not distinguish between the effects of primary and secondary reactions and it may exceed unity.

Modeling photochemical reactions in natural waters is still in its infancy. No proven general model applicable to every kind of natural water is presently available for estimating the rates of consumption or production of chemical species through light-induced reaction pathways. A lot of fundamental and essential information for constructing a precise and realistic model is still not available. For example, the chemical composition of natural waters and their temporal and spatial variabilities are still inadequately known. Thus, the reactants that may participate directly and indirectly in photochemical reactions in a competitive, synergistic, or catalytic way cannot yet be fully identified. The rate at which a particular photochemical reaction may proceed in this poorly defined medium, and to which extend it will proceed cannot be quantified extensively by empirical data. Besides, as stated previously, even in a well-defined medium, the detailed reaction pathways of photochemical reactions can still rarely be identified definitively.

Indirect photolysis may occur if a natural material can absorb solar photons and transfer part of the energy to a chemical. Thus, compounds that do not absorb light directly can still undergo mediated photolysis, and chemicals that photolyze directly may photolyze much more rapidly in natural waters containing organic substances in dissolved or particulate form. Despite these difficulties, general rate expressions for sensitized photolysis can be defined. The rate of light absorption is

$$I^s_{a,\lambda} = 2.3 \varepsilon_{\lambda,s} E_o(\lambda)\,[S] \tag{42}$$

where $\varepsilon_{\lambda,s}$ is the molar absorptivity of the sensitizer and [S] is its concentration [225]. Thus, from Equation (42), the rate constants for indirect photolysis should increase with increasing [S] if $E_o(\lambda)$ is held constant. When the concentration of pollutant P is very dilute, the quantum

yield is directly proportional to this concentration. By denoting C_λ as the proportionality constant, the total rate expression becomes

$$(Rate)_\lambda = 2.3E_o(\lambda)\,\varepsilon_{\lambda,s}C_\lambda\,[S]\,[P] \quad (43)$$

Provided the sensitizer concentration is constant, which is usually the case in nature, the rate expression follows first-order kinetics (→ Air, Chap. 1.).

2.4.2.3. Oxidation and Reduction

Direct and indirect phototransformation as described in the preceding section are possible removal pathways only for those chemicals that, either by their very nature or due to fixation onto light-absorbing carrier substances, can absorb radiation in the range of the solar spectrum. However, the main transformations that lead to the removal of chemicals from the atmosphere and water involve reactions with photochemically generated *oxidants* such as the hydroxy radical (OH·), ozone (O_3), hydroperoxy radical ($HO\cdot_2$), organic peroxy radicals ($RO\cdot_2$), singlet oxygen ($O_2\,\Delta_g$), and nitrate radical ($NO\cdot_3$). Table 9 summarizes the reactivity of different classes of organic compounds toward some of the above oxidants.

Table 9. Rates of oxidation in air and water (after [222])

Type of oxidant	Class or structure	Half-life of reactant
HO· radical (in air)	*n*-, iso-, cycloalkanes	1 – 9 d
	alkenes	0.05 – 1 d
	halomethanes	0.2 – 100 a
	alcohols, ethers	1 – 3 d
	ketones	0.2 – 6 d
	aromatics	0.1 – 3 d
$RO\cdot_2$ radical (in water)	alkanes, alkenes	220 – 2000 a
	alkyl derivatives	220 – 2000 a
	phenols, arylamines	1 d
	hydroperoxides	150 min
	polyaromatics	10 d
$O_2\,\Delta_g$ (in water)	aliphatic compounds	100 a
	substituted or cyclic alkenes	8 – 40 d
	alkyl sulfides	1 d
	dienes	19 h
	enamines	15 min
	furans	1 h

The *reduction* of organic compounds requires an environment of low redox potential, i.e., eutrophic or deeper levels of aqueous systems and a soil depth at which oxygen is limited [226]. Reduction of organics can follow various mechanisms such as hydrogenation, halogenation, or coupled oxidation – reduction, many of which require catalysts. Like with many other chemical reductions, rates of reductive alteration of organics are temperature and pH sensitive [2]. Since the most common mechanism for reduction of organic molecules in the environment is via microbial attack, separating biologically induced from straight chemical reduction in situ is difficult. From the viewpoint of prospective hazard assessment, however, identifying those compounds that are potentially susceptible to reduction and the transformation products that might be deposited in the environment as a result is more important.

2.4.2.4. Sorption and Ion Exchange

Further processes of fundamental importance for the distribution of environmental chemicals in water and soil are sorption and ion exchange. Adsorption is the adhesion of ions or molecules to surfaces or to solid – liquid, liquid – gas, and liquid – liquid interfaces, causing an increase in the concentration of chemicals on the surface or interface over the concentration in solution. Adsorption occurs as a result of a variety of processes with a multitude of mechanisms when the concentration of chemicals lowers the free interfacial energy. The higher the free surface energy at an interface, the greater is the number of substances capable of lowering it.

Physical Adsorption. Molecules of a given material may be adsorbed from the gas or liquid phase to a solid by van der Waals forces. This process resembles a condensation that reaches equilibrium in a time limited chiefly by molecular diffusion rates. The weakness of van der Waals forces is reflected in the low adsorption enthalpy of physiosorptive (or unspecific) bonding; it amounts to hardly more than 40 kJ/mol [227]. Reversibility is a characteristic of physical adsorption that results from rapid equilibration; i.e., at a given temperature and pressure the same amount of material is adsorbed regardless of the direction from which equilibrium pressure and temperature have been approached. From the environmental point of view, however, the

fact that physical adsorption of gases by highly porous media such as silica gel, charcoal, and soil often exhibits "hysteresis," a form of irreversibility due to condensation of liquid in pores is very important. Analogous phenomena occur with chemicals bonded at solid – liquid interfaces.

Chemical Adsorption. Molecules of a given substance may also be adsorbed from the liquid or gas phase by a solid through the activation of valence forces. Consequently, they may be considered as reacting with atoms or molecules on the surface of the solid in a way analogous to any other chemical reaction, and the products of this reaction are usually conveniently considered to be species quite different from the original reacting molecules of the atoms (chemisorption or specific bonding in terms of soil chemistry). Temperatures at which chemisorption occurs are related not to the boiling point of the adsorbate (material being adsorbed), as is the case with physical adsorption, but to the stability of the surface compound and its rate of formation. Therefore, rates of chemisorption vary enormously with material and temperature, as do those of ordinary chemical reactions, and may be very much slower than rates corresponding to diffusion of gases or solutes to the surface. Furthermore, the rates are likely to vary strongly with the amount of material adsorbed. Because of these rate characteristics, the amount of adsorbate at a given pressure or concentration and temperature – when observed over practical time periods – may depend strongly on the way these boundary conditions have been reached. Therefore, irreversibility is extremely common in chemisorption, which is characterized by enthalpy values >100 kJ/mol [227].

Ion Exchange. The presence and extent of ionization also have a great influence on the chemical behavior of a substance, particularly on solubility, sorption, toxicity, and other biological characteristics. Ion exchange as an important sorption mechanism for substances prone to ionization is viewed as an exchange with some other ion that initially occupies an adsorption site on the solid (→ Ion Exchangers) [226]. For example,

$$\mathrm{H^+}\,\overbrace{\underbrace{\mathrm{Al^{3+}}}_{\mathrm{K^+\,Na^+}}}^{\mathrm{Ca^{2+}\,Mg^{2+}}} + 10\,\mathrm{NH_4^+} \rightleftharpoons \mathrm{NH_4^+}\,\overbrace{\underbrace{\quad}_{2\,\mathrm{NH_4^+}}}^{4\,\mathrm{NH_4^+}}\,3\,\mathrm{NH_4^+} \qquad (44)$$
$$+\,\mathrm{H^+} + \mathrm{K^+} + \mathrm{Na^+} + \mathrm{Ca^{2+}} + \mathrm{Mg^{2+}} + \mathrm{Al^{3+}}$$

The ion-exchange property of a soil results mostly from the clay and silt fractions, organic matter, and hydrous oxides of iron. Soil particles have an amphoteric character but normally carry a net negative charge. Positive charges, as indicated by the amphoteric nature of the clay fraction, may originate from hydrous oxides of iron (ferrihydrites), aluminum, and manganese, and from exposed octahedral groups that react as bases by attracting protons from the surrounding soil solution. The negative charge increases and the positive charge decreases with rising pH as a result of increasing ionization of the acid groups and decreasing protonation of the basic groups. For decreasing pH, the corresponding change occurs in the opposite direction. The carboxyl groups of humus ionize under acid conditions, the phenolic hydroxyl groups mainly above pH 6. Because of the highly irregular shape of many clay minerals (and clay – humus complexes, in particular) and the nonuniform distribution of charges in the particles, the surface charge density is quite variable.

To describe ion-exchange processes, several formulas have been proposed in the course of exchange studies. The first group is essentially empirical and intended mainly to provide mathematical formulations best fitting the experimental data. Among this group, the Freundlich [49], Langmuir, and BET (i.e., Brunauer – Emmett – Teller) are particularly important [228]. The second group of formulas is related to the law of mass action.

Empirical Adsorption Formulas. In the *Freundlich equation*, the concentration of a substance in solution is related to that adsorbed as follows [49]:

$$\frac{x}{m} = kc^{\frac{1}{n}} \qquad (45)$$

where x is the amount of substance adsorbed; m the mass of the adsorbent; c the concentration of substance remaining in solution; and k, n are specific constants.

The Freundlich formula is of a parabolic type and therefore cannot give a maximum adsorp-

tion value as does the *Langmuir theory.* According to the latter, adsorption is considered proportional to the pressure of gas (or concentration of a solute) and to the amount of surface not yet covered with adsorbed molecules. The rate of desorption (or evaporation) of adsorbed molecules is considered proportional to the number of such molecules. At equilibrium, these two rates must be equal; thus,

$$\frac{n}{n_{\mathrm{m}}}=\frac{bp}{1+bp} \tag{46}$$

This equation, in which n is the number of molecules adsorbed, n_{m} the number of molecules required for a completely unimolecular surface film, p the concentration of a chemical in solution, and b a constant at a given temperature, is the *Langmuir adsorption equation* (*Langmuir isotherm*).

A third theory of adsorption widely used for representing the adsorption of gases or vapors on solids was advanced by Brunauer, Emmet, and Teller [228]. In contrast to Langmuir's theory, the *BET theory* makes allowance for adsorption of multimolecular layers as is commonly observed when nonporous solids adsorb vapors at pressures approaching saturated vapor pressure. The BET equation is therefore

$$\frac{n}{n_{\mathrm{m}}}=\frac{b\frac{p}{p_0}}{\left(1-\frac{p}{p_0}\right)\left[1+(b-1)\frac{p}{p_0}\right]} \tag{47}$$

where p_0 denotes the saturation vapor pressure, while the other quantities have the same meaning as in Equation (46). The equation is designed for physical adsorption and predicts a very sharp increase in the amount of material adsorbed as the saturated vapor pressure is approached, and this is found to be the case [50].

2.4.2.5. Biotransformation

Biotransformation is a more accurate and general term than biodegradation, because many chemicals are transformed to products of comparable molecular complexity instead of being degraded. Chemically speaking, biotransformation in water and soil systems includes hydrolysis, oxidative cleavage, and reduction, even in aerobic systems. Kinetics of biotransformation are normally based on the Monod model [51], which couples the rate of loss of the chemical to the growth of organisms involved. In the natural environment, however, growth is usually controlled by nutrients present in comparatively constant amounts, and the kinetic expression for biotransformation processes simplifies to a second- or pseudo-first-order reaction:

$$d\,[C]\,/dt=k_{\mathrm{bt}}\,[\mathrm{C}]\,[\mathrm{B}]=k'_{\mathrm{bt}}\,[\mathrm{C}] \tag{48}$$

where d [C]/dt is the rate of transformation of C; [B] is the number of organisms per unit volume; k_{bt} the second-order rate constant for biotransformation; and k'_{bt} the pseudo-first-order rate constant. The half-life of the chemical under transformation ($t_{1/2}$ at a given cell concentration) can then be calculated as

$$t_{1/2}=\frac{\ln 2}{k_{\mathrm{bt}}\,[\mathrm{B}]} \tag{49}$$

In deriving this equation, the microbial community in water, sludge, or soil is assumed to already be acclimated to the chemical, so no lag time is involved in production of the necessary level of biotransforming organisms or appropriate mutants or enzymes.

Because of the remarkable versatility of microorganisms in their catabolic activity, many substances are subject to microbial mineralization or cometabolism in water and soil. If, however, microorganisms have no enzymes active in a necessary catabolic sequence or if the substrate is protected from microbial attack because of special properties of the environment, the compound tends to persist. In some instances, persistence may be attributed to absence of the requisite enzymes or to the lack of permeability of microorganisms to the substrate [229]. Other mechanisms of recalcitrance may be inherent in the molecule itself; in some cases, biotransformable compounds are degraded quickly in one environment but are partially or totally refractory in another.

2.4.2.6. Chemical Structure and Biodegradation

Biodegradation is the most important transformation mechanism for organic compounds in nature. It can occur under conditions in which

oxygen is present (aerobic) or absent (anaerobic). In aerobic processes, organic carbon is oxidized to CO_2, while in anaerobic processes it may ultimately be reduced to CH_4. In either case, the more important environmental variables affecting the rate and extent of biodegradation are temperature, pH, salinity, oxygen concentration, concentration of the chemical involved, ionic strength, composition of microbial associations and concentration of viable microorganisms, quantity and quality of nutrients (other than xenobiotic substances), trace metals, and vitamins.

In general, microbial degradation is impeded by ether bonds and by triple-substituted or condensed phenyl groups–particularly, if the latter have methoxyl, sulfone, nitro, or chlorine substituents in the *meta* position. Among the nitrophenols and chloroanilines, the *ortho*-isomers are relatively recalcitrant. Multiple substitution increases the recalcitrance of aromatic molecules considerably because their aerobic degradation is effected by relatively specific mono- and dioxygenases [229], [230]. Halogen substituents delay the electrophilic attack of these oxygenases considerably by reducing enzyme affinity. Therefore, multiple chlorine-substitued aromatics form relatively persistent compounds, the more so because the delayed cometabolic attack yields "dead-end" metabolites, which cannot be used for cell-building purposes or as an energy source.

Anaerobic Processes. As a consequence, reliable conclusions about biodegradation are not generally possible on the basis of chemical structure alone [47]; therefore, many laboratory methods have been developed for studying the aerobic biodegradative potentials of chemicals [231–234]. In comparison, tests for assessing anaerobic biodegradation have received less attention [84]. Nevertheless, anaerobic degradation is an important process for substances that are not aerobically biodegradable or whose physicochemical properties are such that their occurrence in an aerobic environment is restricted (e.g., chemicals that are strongly adsorbed or insoluble). For these substances, anaerobic biodegradation can be the major process responsible for breakdown in the environment.

Anaerobic biodegradation is complex and usually considered to occur in at least three consecutive main stages, the first of which involves hydrolysis of complex organic molecules such as carbohydrates, proteins, and lipids by the action of extracellular enzymes. In the subsequent stage (acidogenic step), the hydrolysis products are fermented to yield mainly short-chain fatty acids, alcohols, hydrogen, and carbon dioxide. Alcohols and acids are then converted in the acetogenic step. The last degradation step is effected by methanogenic bacteria (methanogens), which utilize acetate and hydrogen to form methane.

Anaerobic processes tend to be self-inhibitory because the hydrolysis of complex substrates such as fats, proteins, and carbohydrates yields volatile organic acids. By lowering the pH of the system, these may inhibit the growth of methanogens unless the system has sufficient buffer capacity [84]. Inhibition of gas production as a result of the toxicity of chemicals is also possible. Finally, some indications have been found that certain compounds (e.g., nitrilotriacetic acid) are degraded anaerobically only when aerobic organisms acclimatized to the compound are incorporated in the system [235].

When the organic compound liable to degradation or enzymatically induced metabolic transformations is utilized as a carbon source, the growth rate of the responsible microorganism depends on the concentration of the former. The rate of substrate utilization then becomes [52]

$$\begin{aligned} -dC/\mathrm{d}t &= \mu X/Y = (\mu_\mathrm{m}/Y)\cdot CX/(K_\mathrm{s}+C) \\ &= k_\mathrm{b}\cdot CX/(K_\mathrm{s}+C) \end{aligned} \tag{50}$$

where μ is the specific growth rate; X the biomass per unit volume; μ_m the maximum specific growth rate; K_s the concentration of the substrate to support half-maximum specific growth rate ($0.5\,\mu_\mathrm{m}$); $k_\mathrm{b} = \mu_\mathrm{m}/Y$, the biodegradation constant; and Y is the biomass produced from a unit amount of substrate consumed. The constants μ_m and K_s depend on temperature, pH, salinity, oxygen concentration, concentration of chemical, composition of microbial associations, and concentration of viable microorganisms, quantity and quality of nutrients (other than xenobiotic substances), trace metals, and vitamins.

2.4.2.7. Bioconcentration, Bioaccumulation, and Ecological Magnification

Bioconcentration is the concentration of a chemical in an organism due to direct uptake from the environment (e.g., ambient air or water), which does not include contaminated food. *Bioaccumulation* includes the latter pathway of concentration, while *biomagnification* considers only concentration processes via food uptake. *Ecological magnification* defines the increase in concentration of a substance in an ecosystem or a food web when passing from a lower trophic level to a higher one [53].

Bioconcentration or bioaccumulation is the result of both kinetic (diffusional transport and biodegradation) and equilibrium (partitioning) processes. It thus involves the following set of fundamental events [47]:

1) Partitioning of the xenobiotic between the environment and some surface of the organism
2) Diffusional transport of the molecules across cell membranes
3) Transport processes mediated by body fluids, such as exchange between blood vessels and serum lipoproteins
4) Concentration of the xenobiotic in various tissues depending on its affinity for specific biomolecules such as nerve lipids
5) Metabolism, cometabolism, or biodegradation of the xenobiotic

The relative importance of the above processes is likely to differ considerably as a result of both the physicochemical properties of the xenobiotic and the morphological and physiological structure of the target organisms. Therefore, an essential ecotoxicological requirement is that the determination of bioconcentration factors (*BCF*s) be complemented by an assessment of biomagnification values. Under real-world conditions, these would provide deeper insight into the complex bioaccumulation processes of xenobiotics as a result of transport and partitioning into higher levels of the food web, which may ultimately result in toxic concentrations [235].

On the basis of bioconcentration studies in different test organisms, correlations with both water solubility and *n*-octanol – water partition coefficients have been established. In some cases, the results proved excellent; in others, the correlations yielded only satisfactory structure – activity relationships (i.e., estimates of *BCF*s to within an order of magnitude for chemicals of $\log K_{ow}$ ranging from 1 to nearly 7).

3. Ecotoxicology

The production, use, and disposal of chemicals lead to their dispersion in the environment. The way in which substances are released depends on their physical and chemical properties (see Section 2.4.1), the production process (→ Production-Integrated Environmental Protection ; → Plant and Process Safety), including extraction of natural resources (→ Raw Materials and Energy) and manufacturing; and the use pattern and means of disposal (→ Waste).

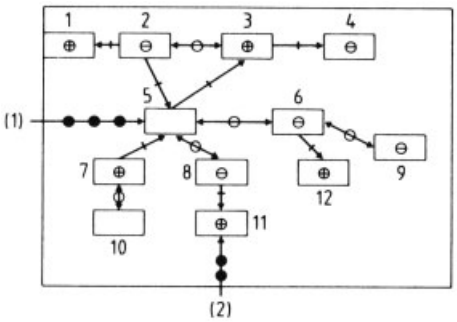

Figure 32. Component-specific influence of two xenobiotics on an ecosystem (after [159], modified)
Substance (1) affects a functionally central component, thus causing breakdown of the whole ecosystem. Substance (2) affects a peripheral component; even its deletion leaves the ecosystem practically unchanged.

Compartments of ecosystems have a limited capacity to assimilate chemicals by breakdown or dilution; if this capacity is exceeded, damage may ensue. Thus, the hazard posed by a single substance or a mixture of chemicals is a function of its inherent toxicity to organisms and the concentration attained. At the community and ecosystem levels, the sensitivity has to be considered specifically (Section 2.3), i.e., the functional position of the organisms affected with regard to the entire system. Figure 32 illustrates schematically the different influence of xenobiotics, affecting two elements of an ecosystem. In case (2), only a peripheral element (in the func-

tional sense of the term) is damaged or deleted, which leaves the system relatively intact. In case (1), however, a component of central importance is eliminated, which leads to breakdown of the entire system.

Thus, the assessment of whether a substance presents a hazard to organisms in the environment is based on a comparison of its predicted environmental concentration (PEC) with the predicted no-effect concentration (PNEC) for organisms in the ecosystem affected. Whenever the PEC : PNEC is <1.0, a substance may be concluded not to present an environmental hazard under the boundary conditions considered. In the opposite case, hazard assessment should be complemented by risk assessment approaches, which define the probability that a hazard will occur. Figure 33 illustrates the temporal and spatial scales that must be considered when analyzing the impact of environmental chemicals on biota or ecosystems.

PEC values can be measured in the field. The identification of unequivocal spatial structures related to emission, transport, and sedimentation normally involves the construction of isopleth maps by means of metric interpolation. A pertinent statistical method that permits testing the validity of interpolation procedures is variogram analysis, normally applied in conjunction with kriging methods (i.e. methods in geostatistics for valid derivation of isopleths) [236], [237] (Section 3.1.4). Alternatively, the predicted environmental concentration can be estimated, if the manner and quantity in which a chemical enters the environment, and its subsequent distribution and transformation, are known (see Section 2.4.2; [2], [238]). This normally involves the application of models that are available at different levels of complexity, ranging from empirical models for practical purposes to rather abstract ones aiming at generalized insights via simulation.

Effects assessment establishes a maximum concentration of a substance that does not cause adverse effects. The PNEC values may be derived from whatever ecotoxicological data are available, by using simple mono- and oligospecific test systems, or micro- and mesocosms, and ideally from toxicity measurements at the ecosystem level. The lower the level of assessment approach, the more appropriate are assessment factors to compensate for restrictions in the data.

3.1. Exposure Assessment

Exposure assessment involves the identification of point or diffuse emission rates and the subsequent distribution tendencies, based on predictive models. In the first assessment step, the initial concentration ($PEC_{initial}$) in each compartment of concern must be estimated. If necessary, a revision in light of relevant new information may lead to a revised $PEC_{initial}$ and a PEC_{local} or a $PEC_{regional}$. The PEC_{local} is the estimated concentration of a substance when released to a particular environmental compartment, with further fate processes, (e.g., volatilization, sorption, biodegradation) taken into account. The $PEC_{regional}$ denotes the predicted concentrations of a substance in all relevant compartments, when the further distribution and fate of the chemical are taken into account.

3.1.1. Release Estimation

For realistic exposure assessments, basically the entire life cycle of a chemical should be analyzed and PEC values calculated for each relevant process. To standardize the release estimation for screening purposes [239], the EEC and OECD have defined three different categories of use pattern: main, industrial, and use. The main category classifies substances in four groups (see Table 10).

Table 10. Release fractions in the screening phase of hazard assessment of chemicals related to the main category (after [239])

Main category	Release fraction of production volume	Examples
Closed system	0.01	chemical intermediates
Enclosed in a matrix	0.1	plastic additives
Nondispersive	0.2	photochemicals
Wide dispersive use	1.0	solvents, plant protection products, detergents

In addition to the main categories, there are 15 different “industrial” categories and 55 different “use” categories [240]. This classification scheme allows a stepwise approach in release

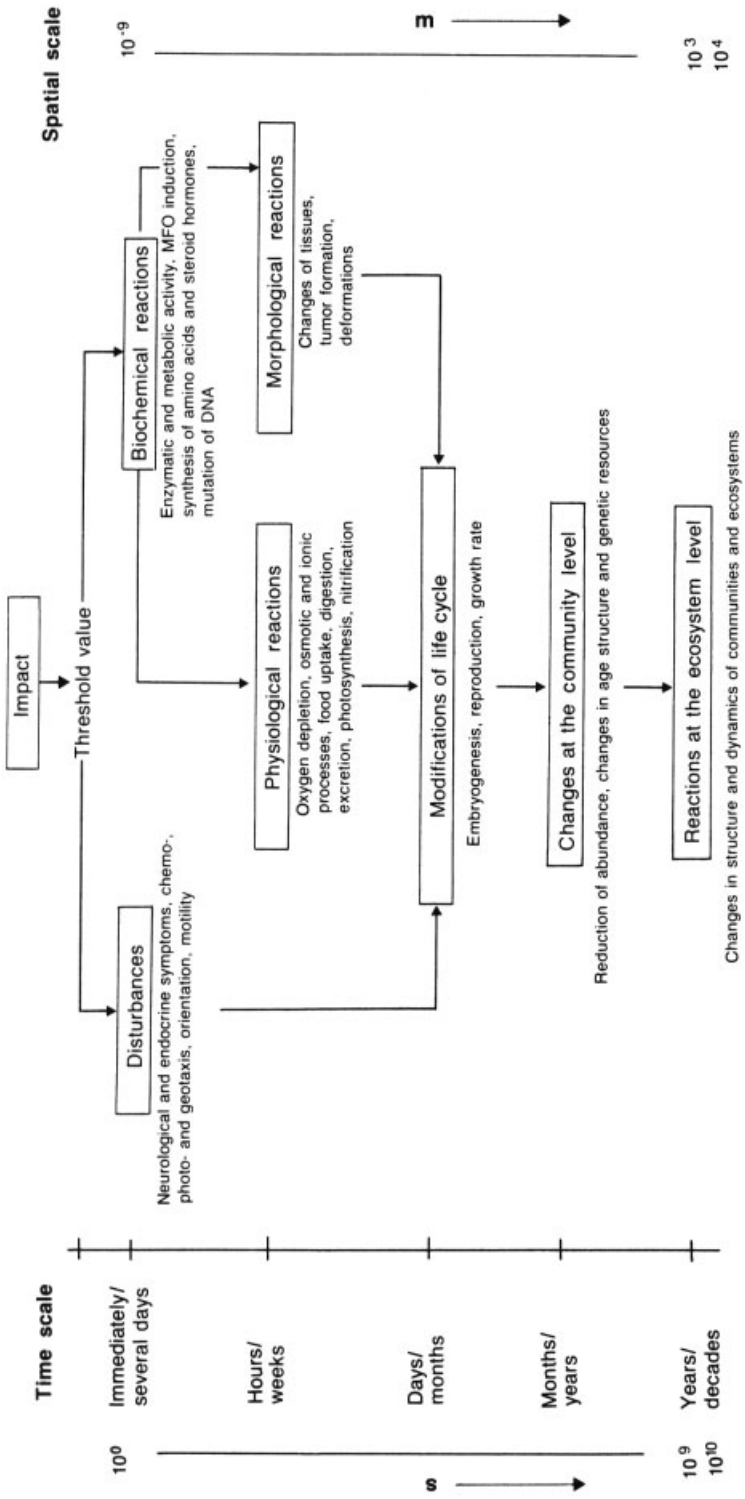

Figure 33. Temporal and spatial scales of the impact of chemicals on biota (after [53], modified)
MFO = monooxygenases

estimation. In the screening phase a rough estimate based on the main category may be sufficient, whereas in the confirmatory phase more refined release data (e.g., based on other categories) are needed. In the investigation phase, finally, detailed information on specific releases of the substance is required (→ Chemical Products: Safety Regulations, Chap. 3.).

3.1.2. Elimination

A reduction in the concentration of chemicals in gaseous or aqueous discharges prior to their release to the environment is termed elimination. Elimination may be due to physical, chemical, or biochemical processes. The most significant elimination mechanisms are biochemical processes, especially aerobic biodegradation by bacteria or an alternation of aerobic and anaerobic phases (see Section 2.4.2) [2], [35]. Physical processes transfer a substance from one phase to another. In the case of volatile substances, the aeration process normally enhances their removal by stripping them from the solid or liquid phases to the atmosphere [47]. Substances from exhaust gas streams may be eliminated by scrubbing, e.g., by adsorption on a suitable material or passage through a suitable aqueous solution.

3.1.3. Dispersion

Pollution transport and dispersion are directly related to wind speed or flow velocity of rivers, etc. The higher the velocity, the greater is the volume of air passing a point source per unit time, and the lower is the resultant concentration per unit volume. Higher wind speed also means greater turbulence, which involves fluctuations of the local wind vector on various scales, and these eddies act to diffuse a plume. In addition to this "stretching" phenomenon [241], wind speed is important for the transport distance. In strong winds, pollutants may be transported long distances, but the resultant dilution over 10 – 50 km is usually such that most individual plumes have lost their identity and contribute to a more general (background) contamination of the lower atmosphere. Consequently, weaker regional wind systems may exhibit much greater potential for pollution since both horizontal transport and eddy diffusion are reduced.

This simple fact becomes all the more important because, under these conditions, local wind systems tend to develop that are difficult to predict with any accuracy [54]. In so far as they are closed smaller-scale circulations (e.g., land and sea breezes, mountain and valley winds, city winds), they often produce recirculation of pollutants coupled with a progressive increase in pollutant loading with time. The city thermal wind system operating mainly in the urban canopy layer is particularly dangerous; it is produced by microscale processes in the streets [242–244]. Here, flow converges on the city (or its variable thermal subcenters) from all directions, rises, diverges, and then moves outward to subside on the urban – rural fringe, where it rejoins the inflow.

The worst conditions for dispersion, however, occur in the case of temperature inversion and consequent formation of a stable boundary layer. With the warmer air layer acting as a lid on the mass of cooler air below, pollutants cannot escape from the cooler zone and thus increase in concentration in the confined space. This situation can be brought about by cooling (usually radiative) from below, by warming (usually adiabatic) from above, or by advection of warmer or cooler air.

Since the 1970s, modeling of dispersion processes in air has become more and more sophisticated owing to the progressive consideration of pertinent boundary layer effects and concentration or removal processes ([245–248], e.g.). This involves not only a considerable increase in computer time but also, and often more prohibitive, much more detailed input data. In the light of these circumstances, a question of practical importance is whether simpler models are not likely to yield comparable results on the condition that their data base can be made proportionally more comprehensive.

Box models, including dry and wet deposition and chemical transformations, are likely to provide calculated concentrations that are in reasonable or good agreement with measured values on a local scale as well as with transport data over a distance of 1000 km. An exemplary prototype whose fundamentals have often been used was developed by [249], describing SO_2 and NO_x concentrations in the industrial areas

of North Rhine – Westphalia along the Rhine and Ruhr rivers.

Statistical models define immission concentrations by means of multiple regression relating immissions to emissions, meteorological parameters, topography, or other dispersion-controlling factors. Models of this type vary considerably with respect to both regression analytical approach and selection of data.

Observations of looping, coning, or (to a lesser degree) fanning plumes on an instantaneous basis reveal that the concentration is very peaked, with a marked maximum in the plume center. Averaged over growing time intervals, an increasingly smooth and wide plume envelope appears, which contains all of the short-period plume fluctuations. The wider spread results in a flatter concentration curve, with a correspondingly lower peak centered on the time mean axis of the plume (i.e., the mean wind direction). Analogous bell-shaped or Gaussian concentration profiles also characterize the vertical plane through the plume, although there may be a difference between the width of the horizontal and vertical profiles indicating that turbulence is greater in one or another direction.

Eulerian or Grid Model. The Eulerian or grid model is a mathematical representation of the pertinent transport processes and chemical reactions occurring in a grid of cells yielding the spatial and temporal variations of the ground cell concentrations of primary and secondary pollutants. It is characterized by fixed coordinate systems with respect to the ground and consists of a set of nonlinear coupled partial differential equations expressing the conservation of mass of each pollutant considered. Inputs to the model are numerical representations of the meteorological and emission conditions as a function of time of day, initial concentrations, and boundary conditions of inflow.

Lagrangian Model. In the trajectory or Lagrangian model, the time behavior of the concentration c of a pollutant is described by a differential equation of the type

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial z}\left(\underline{K}\cdot\frac{\partial c}{\partial z}\right) + S \qquad (51)$$

where c depends only on height above ground z and the time of travel t of an air parcel, while S refers to the net source – sink term representing the combined effects of emissions and chemical processes during transport, deposition, and flow across the upper boundary. $\underline{K}$ represents a vertical eddy diffusion coefficient whose values characteristically depend on time of day, height, and location. In comparison with the above Eulerian models, trajectory models have commendable mathematical advantages resulting from their moving coordinate systems with respect to a fictitious vertical air column moving horizontally with the advecting wind [250]. Considered as the present state of the art in operational long-range transport modeling, the Lagrangian approach has been used in numerous studies in the United States and Canada relating to transboundary air pollution.

3.1.4. Local and Regional Concentration Estimates

After emission, the environmental fate of a substance is governed by both transport and transformation processes that may be assessed by measurements or modeling approaches of which there are two basic types. The *local* or *monocompartment approach* is focused on the distribution of a substance within a particular environmental compartment, which leads to a $\mathrm{PEC}_{\mathrm{local}}$. The *regional* or *multicompartment approach* is directed toward the overall distribution, where transport occurs in all relevant environmental compartments (i.e., air, water, or some other medium) leading to $\mathrm{PEC}_{\mathrm{regional}}$ values.

The first approach is appropriate if the degradation rate is (distinctly) greater than the intercompartmental transfer rate; otherwise a multicompartment or multimedia approach is indicated (Section 3.3) [251]. Irrespective of the type of model to be used, the environmental parameters such as compartment volumes or compartment ratios must be defined; they are sensitive points for the final result. For substances that enter the environment from diffuse emission sources, a realistic PEC estimate cannot be calculated by means of the habitual point-source models. In such cases, the exposure assessment must be based on scale considerations and spatial modeling [252].

3.1.4.1. PEC Estimates in the Air Compartment

For the primary exposure assessment, models of advection and related dispersion processes may be used that compute concentrations of a substance at different receptor compartments. These require different minimum data sets consisting of discharge emission rates, effective height of source, and meteorological conditions, (stability classes of atmosphere, wind speed etc.). For regulatory purposes, the TA Luft model [253] is widely used (see also Section 3.1.3; → Air).

For subsequent revisions of the exposure assessment, information on the long-term fate of the chemical in the atmosphere is required. This involves the appropriate consideration of direct and indirect photodegradation and other chemical reactions [245], [254]. The reactivity of the hydroxy radical formed predominantly via photolysis of ozone by reaction of electronically excited oxygen atoms with water vapor is of fundamental importance in this connection.

A balance must be struck between simple models that do not represent the real-world situation and models that are too difficult to use. Thus, level 1 (equilibrium, conservative, steady state) and level 3 models (nonequilibrium, nonconservative, steady state) in the sense of the MACKAY et al. classification [251] are considered useful simple approaches. Level 2 models (equilibrium, nonconservative, and steady state) usually are of minor interest, whereas level 4 models (nonequilibrium, nonconservative, nonsteady state) should be used only for estimating disappearance of chemicals from the environment [238].

An important drawback of these "unit world" models is that they compute only one concentration value for each compartment, whereas measurements normally indicate that actual concentrations in the environment may range over several orders of magnitude. Thus, they are suited only to provide an indication of concentrations in places far from the source (i.e., background levels). Improvements are possible by performing a sensitivity analysis, by using statistical approaches, or by adapting the unit world to the region of interest [255], [256]. An alternative method is the application of dispersion models, provided that appropriate input data, representative dispersion parameters, and sufficient meteorological information are available (see Sections 3.1.3, 3.3).

3.1.4.2. PEC Estimates in Soil and Subsoil

Xenobiotics in soil and subsoil generally result from agricultural or industrial activities. The main problems in recent years were related to the application of manure and fertilizers, pesticides, sewage sludge, and to contaminated sites (see → Soil, Chap. 5., [2], [48]). Table 11 illustrates, on a comparative basis, heavy-metal concentrations in Schleswig-Holstein soils. In the light of the very high spatial variability of the soil cover, representative sampling is particularly important for concentration measurements. Regionalized variables should be applied to adequately characterize the chemical soil properties under consideration; consequently, variogram analysis in combination with kriging procedures is the appropriate method (see → Soil, Chap. 6.; [159], [236]).

3.1.4.3. Variogram Analysis

Mathematical Concept. A regionalized variable can be considered the realization of a random function Y. Usually, this function is assumed to be stationary; this implies the following:

1) The expectation (m) of Y at any point of x is constant and independent of x.

$$E\left[Y\left(x\right)\right] = \left(m\right) \tag{52}$$

2) The covariance function of any pair of points $\boldsymbol{x}$ and $\boldsymbol{x}+\boldsymbol{h}$ depends exclusively on the vector $\boldsymbol{h}$ and is independent of $\boldsymbol{x}$

$$E\left[Y\left(\boldsymbol{x}\right)Y\left(\boldsymbol{x}+\boldsymbol{h}\right)\right] - m^2 = K\left(\boldsymbol{h}\right) \tag{53}$$

In many cases, only the increments of the functions are supposed to be stationary. The intrinsic hypothesis for a vector $\boldsymbol{h}$ concerning expected value and variance is

$$E\left[Y\left(\boldsymbol{x}+\boldsymbol{h}\right) - Y\left(\boldsymbol{x}\right)\right] = 0 \tag{54}$$

$$Var\left[Y\left(\boldsymbol{x}+\boldsymbol{h}\right) - Y\left(\boldsymbol{x}\right)\right] = 2\gamma\left(\boldsymbol{h}\right) \tag{55}$$

Table 11. Average concentration of selected heavy metals in Schleswig-Holstein soils in mg/kg (after [257])

Substrate	As		Cd		Co		Cr		Cu		Hg		Ni		Pb		Zn	
	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C	A	C
S	2.61	**2.67**	**0.36**	**0.38**	1.96	**3.33**	2.55	**5.00**	**4.63**	**3.75**	0.082	**0.010**	8.97	**11.53**	**21.25**	**7.49**	**15.89**	**14.97**
Su2			0.46				5.67		7.70	5.61					23.91		25.95	16.41
Su3			0.31				18.83		13.26						34.72		43.26	
Su4																		
Su	7.13		0.37	0.42	3.30		11.45	16.03	**9.88**	4.75			11.02		29.00	6.29	**32.66**	15.61
Slu																		
Sl2			0.35				11.42		11.12				9.06		29.61		35.04	
Sl3	8.11		0.40	0.26	5.40		**22.73**	28.43	**9.21**	11.82			10.61	22.19	**30.67**	13.20	**43.91**	34.20
Sl4	7.21	**6.52**	0.51	**0.32**	6.73	**6.63**	**34.52**	**34.75**	**12.56**	**12.63**			**15.98**	19.65	**30.31**	**16.66**	**50.69**	**41.26**
Sl	**7.57**	**6.32**	0.43	**0.28**	**5.75**	**6.37**	**24.42**	**31.75**	**10.69**	**11.88**			**13.09**	20.46	**31.51**	**15.04**	**43.54**	**38.57**
Ls2																		
Ls3	7.05	**6.99**	0.70	**0.63**	7.84	**8.21**		**41.26**	13.83	**13.61**		0.022	19.61	**24.41**	31.99	**17.10**	63.59	**50.38**
Ls4				0.52				30.26		13.74				31.44		13.57		39.75
Ls	6.68	**6.68**	0.68	**0.60**	7.15	8.00	40.45	**39.07**	13.19	**13.63**	0.100	**0.022**	19.08	**25.44**	30.00	**16.49**	58.53	**48.38**
Lts		**6.45**		**1.15**		**11.62**				**15.35**		**0.023**		**31.09**		**15.58**		**52.92**
Lt2																		
Lt3																		
Ltu																		
Lt		**6.53**		**1.14**		**11.48**		59.71		**17.10**		**0.024**		**34.98**		**16.00**		**57.72**
Total	**6.78**	**5.46**	**0.45**	**0.55**	**5.34**	**7.24**	**21.04**	**30.20**	**10.05**	**10.70**	**0.085**	**0.019**	**13.63**	**22.90**	**29.04**	**13.34**	**39.64**	**36.75**

A = Average of A horizons (0 – 20 cm)
C = Average of C horizons
Total = Average heavy-metal content except marshland and peat
PNEC: As 20 mg, Cd 3 mg, Co 50 mg, Cr 100 mg, Cu 100 mg, Hg 2 mg, Ni 50 mg, Pb 100 mg, Zn 300 mg
0.00 for $n \geq 10$ and < 30
0.00 for $n \geq 30$

	Slu = Loamy-silty sand	S = Sand	Lts = Clayey – sandy loam
Su = Silty sand	Sl = Loamy sand	Sl = Loamy sand	Lt = Clayey loam
Su2 = Low silt content	Sl2 = Low loam content	Ls2 = Low sand content	Lt2 = Low clay content
Su3 = Medium silt content	Sl3 = Medium loam content	Ls3 = Medium sand content	Lt3 = Medium clay content
Su4 = High silt content	Sl4 = High loam content	Ls4 = High sand content	Ltu = Clayey – silty loam content

The so-called semivariogram (designated "variogram" further on), i.e., the function of the vector $\boldsymbol{h}$ is defined as

$$\gamma(\boldsymbol{h}) = 1/2Var\left[Y(\boldsymbol{x}+\boldsymbol{h})\,Y(\boldsymbol{x})\right] \tag{56}$$

From Equations (55) and (56),

$$\gamma(\boldsymbol{h}) = 1/2E\left[Y(\boldsymbol{x}_i+\boldsymbol{h}) - Y(\boldsymbol{x}_\mathrm{i})\right]^2 \tag{57}$$

In case of discontinuous data the expected value ($\boldsymbol{h}$) can be estimated by the formula

$$(h) = \frac{1}{2n}\sum_{i=1}^{n}\left[Y(\mathrm{x}_i+\mathrm{h}) - Y(\mathrm{x}_i)\right]^2 \tag{58}$$

where n is the number of pairs of points.

The points are situated in either one-, two- or three-dimensional space. In two-dimensional space as dealt with here, the coordinates h_1 and h_2 determine the vector $\boldsymbol{h}$. Hence, the variance of measured values depends on the distance and the direction of the difference vector $\boldsymbol{h}$.

In practice, the variogram is usually computed for four main axes (Fig. 34) to account for possible directional effects. Distinctions of the range in different directions (anisotropy) then enable a more detailed interpretation. A mean variogram that is independent of such directional effects is usually computed on the basis of the four directional variograms. A variogram is geometrically characterized by two boundary criteria called "sill" and "range" (Fig. 35). The latter denotes the maximum extent of influence, while the former is the analogous limiting value of influence as measured on the ordinate.

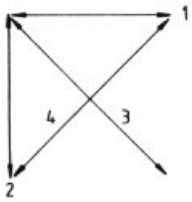

Figure 34. Principal axes of a variogram

Interpretation of the Variogram. *Zone of Influence*. The foregoing indicates that a variogram does not necessarily have a maximum or a level of stabilization. Because of this, two basic types of variograms are distinguished (Fig. 36). In Figure 36 A the maximum range is reached when the correlation between $Y(\boldsymbol{x})$ and $Y(\boldsymbol{x}+\boldsymbol{h})$ becomes nil

$$\gamma(\mathrm{h}) = 1/2Var\left[Y(\mathrm{x}+\mathrm{h}) - Y(\mathrm{x})\right] \tag{59.1}$$

$$= 1/2\left\{Var\left[Y(\mathrm{x}+\mathrm{h})\right] + Var\left[Y(\mathrm{x})\right]\right\} \tag{59.2}$$

$$= \frac{2\sigma^2}{2} = \sigma^2 \tag{59.3}$$

This means that a mathematically meaningful interpolation is possible if and only if the maximum distance between neighboring points of measurement is shorter than the distance indicated by the range value of the variogram. In Figure 36 B the zone of influence extends beyond the area examined.

Mathematical Models of the Curve. To best fit the curve of the variogram under construction to the sequence of points obtained primarily from computation of the variance function, various mathematical models are used, the most important of which are described briefly (Fig. 37). Power functions (Fig. 37 A):

$$\gamma(\boldsymbol{h}) = C\,|\boldsymbol{h}|^{\gamma} \quad \text{with}\, 0<\gamma<2 \tag{60}$$

$\gamma(\boldsymbol{h}) = C$ — $\boldsymbol{h}$ — is the special case of a linear model. Spherical model (Fig. 37 B):

$$\gamma(\mathrm{h}) = C\left[3/2\frac{\mathrm{h}}{a} - 1/2\frac{|\mathrm{h}|^3}{a^3}\right] \quad \text{for } |\mathrm{h}| \le a \tag{61}$$

$$\gamma(\mathrm{h}) = c \text{ for } |\mathrm{h}| > a$$

Exponential model (Fig. 37 C):

$$\gamma(\mathrm{h}) = C\left[1 - e^{\frac{-|\mathrm{h}|}{a}}\right];\ a \simeq 1/3 \text{ of range} \tag{62}$$

Gaussian model (Fig. 37 D):

$$\gamma(\mathrm{h}) = C\left[1 - e^{\frac{-|\mathrm{h}|^2}{a^2}}\right];\ a \simeq 0.58 \times \text{range} \tag{63}$$

Often, for a more exact fit the models must be combined. The variance of the sample is approximately the same as the value of the sill as far as bounded variograms are concerned. The model of the curve is important primarily for determination of the "nugget effect," a phenomenon that must be dealt with when the curve's behavior near the origin is considered.

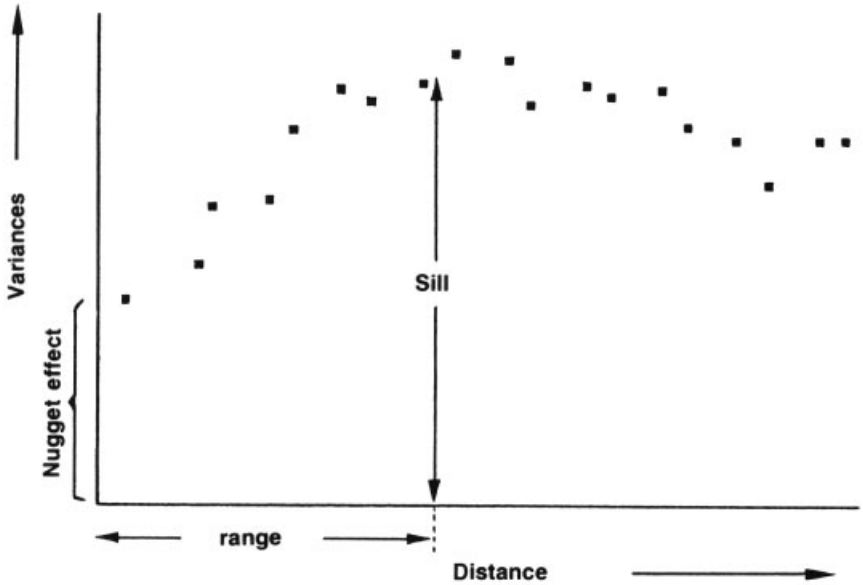

Figure 35. Sill and range of an experimental variogram

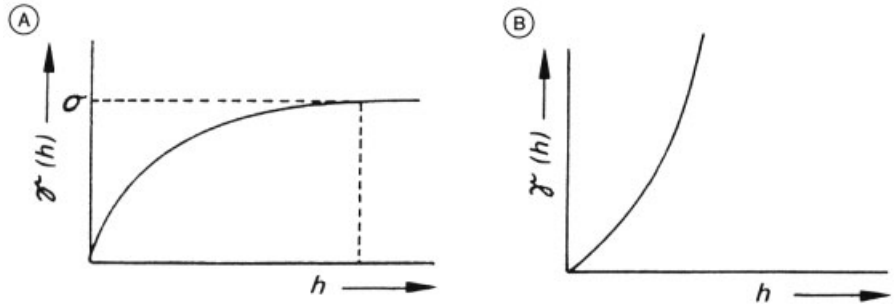

Figure 36. Bounded (A) and unbounded (B) variogram

Behavior of the Curve near the Origin.

1) A parabolic shape shows a high degree of continuity of the regionalized variable. It is differentiable.
2) A linear shape shows continuity "in average" according to MATHERON [236].
3) The curve does not intersect the abscissa at the origin. The nugget effect reveals great irregularity and can be caused fundamentally either by an extremely discontinuous distribution in the immediate neighborhood of the sample taken or by observational errors. Supplementary information is consequently needed.
4) A straight line parallel to the abscissa indicates that no correlation exists between any points $Y(\boldsymbol{x}+\boldsymbol{h})$ and $Y(\boldsymbol{x})$ whatever their distance might be. The sample does not show any spatial structure.

Anisotropies. Distributions characterized by different variabilities in different directions are reflected in the resulting variogram and called anisotropy.

1) Geometrical or affine anisotropy exists whenever elliptical zones of influence can be deduced in a two-dimensional space
2) In a three-dimensional space, variations might also appear in the vertical direction; this is the case of stratified anisotropy

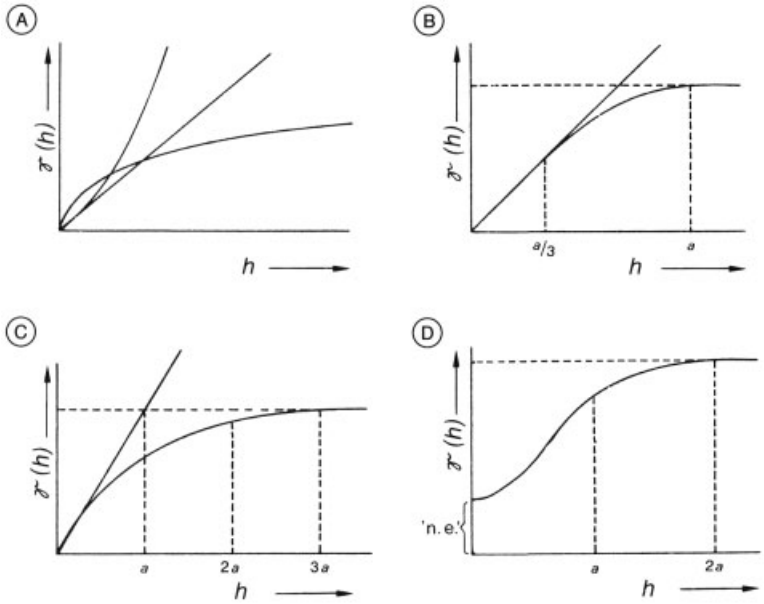

Figure 37. Mathematical models of a variogram
A) Power function model; B) Spherical model; C) Exponential model; D) Gaussian model with "nugget effect"

3.1.4.4. Modeling Chemical Distribution in Soil and Subsoil

Assessing chemical concentrations in soil by means of models normally involves two basically complementary approaches. One type of model simulates chemical movement in the unsaturated zone, (i.e., from the soil surface down to the water table); the other describes chemical transport in the saturated zone or aquifer. BONAZOUNTAS [258], FRÄNZLE [2], and MATTHESS [48] provide state-of-the-art reviews of these models.

Figures 38 and 39, representing one of the most comprehensive hierarchical models developed so far [252], illustrate that the two most important factors influencing chemical behavior in the unsaturated soil zone are sorption and biotransformation. The essential boundary conditions and processes involved (Section 2.4) are indicated in the model that is free from limiting meso- and macro-scale assumptions such as homogeneity (i.e., homogeneous compartments and uniform distribution of chemicals) and stationarity (i.e., large-scale equilibria in extensive compartments). Extended versions of this model serve specific purposes, e.g., modeling of carbon and nitrogen fluxes through ecosystems (Fig. 40).

Small-Scale Assessment of Chemical Distribution. Fate and behavior of chemicals can be described approximately by means of table functions, see Table 12 [259]. In combination with geographic information systems, they permit the production of smaller-scale single-species sensitivity or distribution maps. Six parameters, defined in score form on a 0 (practically none) to 5 (very much, very high) scale, are employed to describe chemical reactions in soil:

1) Solubility in mg/L (20°C) 0: <0.1, 1: 0.1 – 1, 2: 1 – 10, 3: 10 – 100, 4: 100 – 1000, 5: >1000
2) Adsorption, as defined by the Freundlich k values for organic matter (k_{oc}) and clay (k_c) at a soil – solution ratio of 1 : 2 0: <0.5, 1: 0.5 – 50, 2: 50 – 150, 3: 150 – 500, 4: 500 – 5000, 5: >5000 pH influence + or − (i.e., adsorption increases with increasing or decreasing pH)
3) Degradability, defined in terms of time required to produce an approximately 90 % transformation at a soil temperature ranging

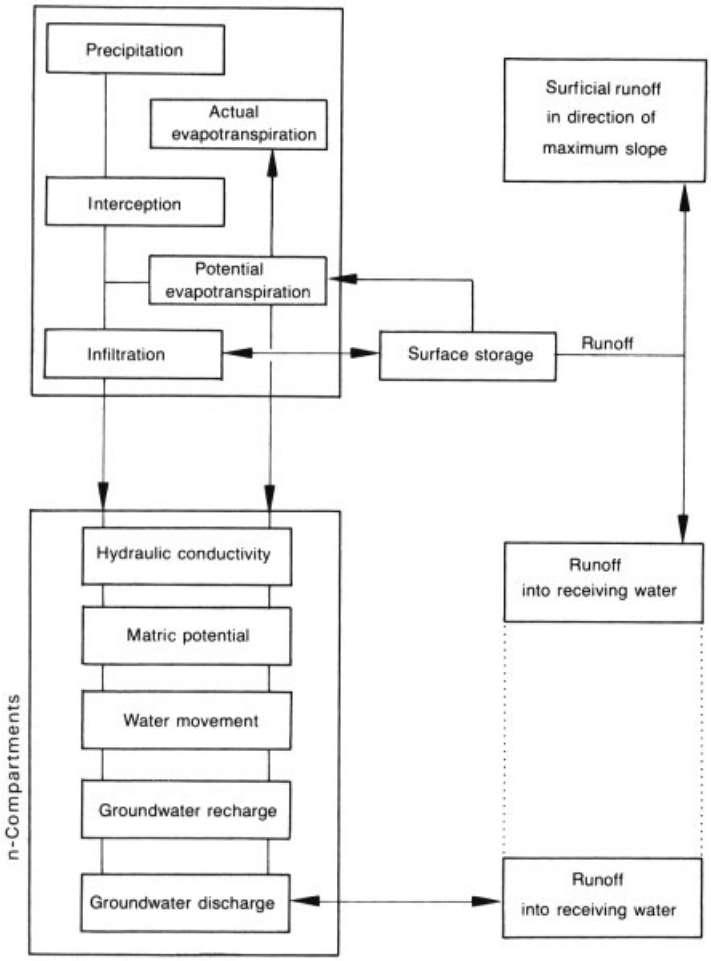

Figure 38. Soil moisture module of the integral WASMOD/STOMOD model (Wasserhaushaltsmodell/Stoffflußmodell, after [252], modified)

from 11 to 16 °C: 1: >3 years, 2: 1 – 3 years, 3: 18 weeks to 1 year, 4: <18 – 6 weeks, 5: <6 weeks

4) Volatility, defined by the vapor pressure of the pure substance in (hPa) at 20 – 25 °C: 1: $<10^{-3}$, 2: 10^{-3} – 100, 3: 100 – 500, 4: >500 or Henry constants 1: $<4\times10^{-6}$, 2: 4×10^{-4}, 3: 4×10^{-4} – 0.04, 4: >0.04

In considering the influence of organic matter content on the adsorption process, the average scores of Table 12 must be defined more precisely according to Table 13. Further refinements relate to texture, pH value, temperature, etc. (Tables 14, 15, 16).

3.1.4.5. PEC Estimates in Aquatic Compartments

Establishing valid monitoring networks or sampling grids poses distinctly lesser problems in surface waters than in soils, but the inherent difficulties should not be underestimated. For instance, determination of the surficial nitrate concentration pattern in a dimictic – holomictic lake of about 1-km^2 size requires 40 – 50 stations [167]. For other chemical species, network densities must be higher still in the light of variogram-analytical evaluations.

In the light of these technical difficulties, modeling approaches play a considerable role

Table 12. Behavior of organic chemicals in soil (after [259], for explanation of figures see)

Chemicals	Aqueous solubility	Adsorption to		pH influence	Degradation		Volatility	Concentrations∗, mg/kg		
		Organic matter	Clay		Aerobic	Anaerobic		A	B	C
Chlorinated alkanes and alkenes										
Dichloromethane	4	1	1	0	3 – 4	2	4	0.1	5	50
Trichloromethane	4	2	2	0	2 – 3	1 – 2	4	0.1	5	50
Tetrachloromethane	4	1 – 2	2	0	2 – 3	1 – 2	4	0.1	5	50
Dichloroethane	4	2 – 3	2	0	3	2	4	0.1	5	50
Monochloroethylene	4	1	1	0	4	4	3			
Trichloroethylene	4	2	1	0	3	2	4	0.1	5	50
Tetrachloroethylene	3	1 – 2	1	0	3	2	3	0.1	5	50
Carbocyclic compounds										
Benzene	4	1	1	0	3	2	4	0.01	0.5	5
1,2-Dichlorobenzene	3	3	2	0	2	2	3	0.05	1	10
1,2,4-Trichlorobenzene	2	3 – 4	1 – 2	0	2 – 3	2 – 4	3	0.05	1	10
Phenol	4	1 – 2	1 – 2	–	4	3	2	0.05	1	10
2,5-Dichlorophenol	4	3	2	–	3	4		0.01	0.5	5
2,4,5-Trichlorophenol	4	3	1	–	4	4		0.01	0.5	5
Pentachlorophenol	2	4	2	–	3 – 4	4	2	0.01	0.5	5
Aniline	4	1 – 2	1	+	4 – 5	4	2			
4-Chloroaniline	4	3	2 – 3	+	2	2	2			
Toluene	3	1	1	0	4	2 – 3	3	0.05	3	30
Toluidine	4	2 – 3	1	+	2	1				
Xylene	3	2 – 3	1 – 2	0	3		2	0.05	1	10

∗ A = reference level, can be regarded as unpolluted; B = indicative value for further investigation; C = clean-up necessary.

Table 13. Sorption capacity of soil related to organic matter content (after [259])

Organic matter content of soil, %	Relative sorption capacity (Table 12)				
	1	2	3	4	5
0.5 – 1	0	0	1	1.5	2
1 – 2	0	0.5	1.5	2	3
2 – 8	0.5	1	2	3	4
8 – 15	0.5	1.5	2.5	3.5	4.5
>15	1	2	3	4	5

∗ Average value of the uppermost 30 cm of soil (i.e., the Ah or Ap horizons).

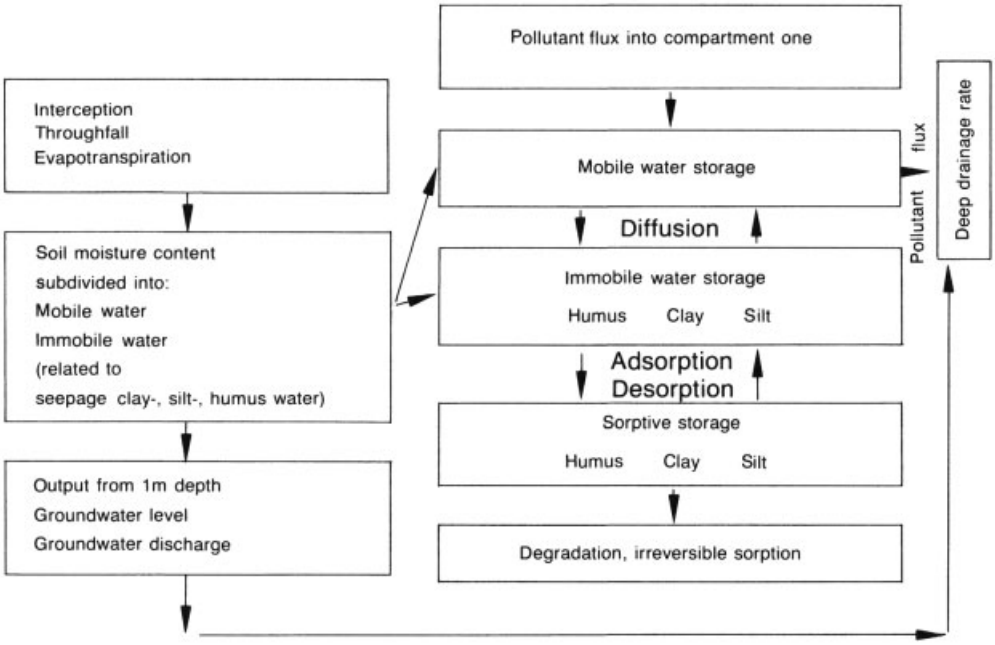

Figure 39. Solute module of WASMOD/STOMOD (Wasserhaushaltsmodell/Stoffflußmodell) (after [252], modified)

Table 14. Sorption capacity of soil related to texture (after [259])

Texture classes∗	Relative sorption capacity (Table 13)			
	1	2	3	4
1. S Ss′2∗	0	0.5	1	1.5
2. Sc2 Sl2, 3 Ss′ 3, 4 S′s S′	0.5	1	1.5	2
3. Sls′ Sl4 S′l2, 3 S′ls Ls Ls′ Sc3 Cs4 S′C2, 3	0.5	1	2	3
4. S′l4 Cl Cs2, 3 C′s Lcs Lc2, 3 Lcs′ S′c4	1	1.5	2.5	3.5
5. C	1	2	3	4

∗ Medium texture of the uppermost 30 cm of soil (i.e., the Ah or Ap horizons). S = Sand, S′ = Silt, L = Loam, C = Clay, s = Sandy, s′ = Silty, l = Loamy, c = Clayey; 2 = low, 3 = medium, 4 = high content.

Table 15. pH control of relative sorption capacity of soil (after [259])

pH influence	pH ($CaCl_2$)			
	>6.5	6.5 – 5.5	5.5 – 4	<4
+	+0.5	0	−0.5	−1
−	−0.5	0	+0.5	+1

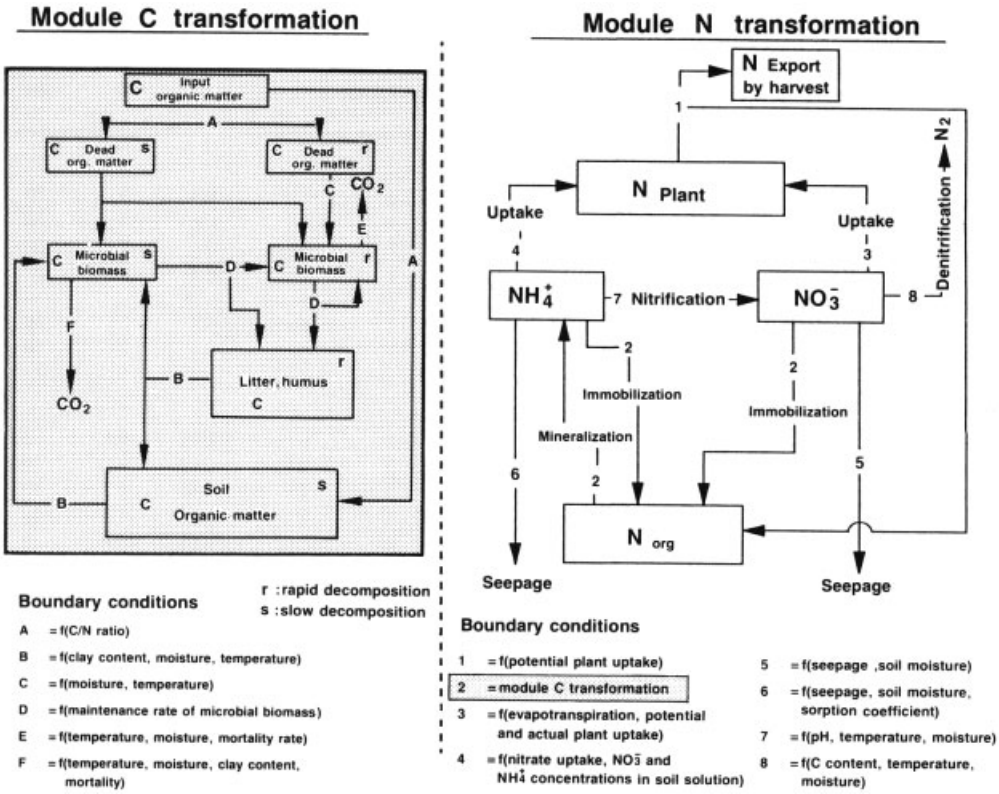

Figure 40. Carbon and nitrogen modules of the extended WASMOD/STOMOD (Wasserhaushaltsmodell/Stoffflußmodell) model (after REICHE, unpublished, modified)

Table 16. Degradability of organic chemicals related to mean air temperature (after [259])

Degrad-ability	Mean temperature of vegetation period, °C			Mean annual temperature, °C		
(Table 12)	21 – 16	16 – 11	11 – 6	12 – 9	9 – 6	6 – 3
1	1.5	1	0.5	0.5	0.5	0
2	2.5	2	1.5	1.5	1	0.5
3	3.5	3	2.5	2.5	2	1.5
4	4.5	4	3.5	3.5	3	2.5
4 – 5	5	4.5	4	4	3.5	3

[238]. Differentiation between large-scale dilution models operating at the regional level, (e.g., a whole river), and site-specific models requiring a more specific characterization of the projected site is indicated. Simple dilution models assume a homogeneous distribution of chemicals in a water body; i.e., for a given set of stream or river and sewage treatment (→ Wastewater, Chap. 1.) or discharge conditions, an average concentration is calculated. Site-specific models may yield concentration profiles of a chemical in the water body. Also concentrations of chemicals released into the river at many points can be estimated by dilution modeling. The simulation then comprises a relatively large region, typically a major catchment or an entire country. When a sufficient amount of data on dilution factors is available, chemical exposure can be expressed in the form of PEC frequencies in the rivers considered.

The data required to run a surface water model relate to the character of the chemical, the hydrological structure of the catchment, and

the discharge pattern [238]. The chemical data include molar mass, melting and boiling points, water solubility, ionization constant, and kinetic parameters relating to sorption, volatilization, photolysis, hydrolysis, oxidation, and biotransformation processes (see Section 2.4). Environmental data vary according to the complexity of the model, beginning with a stream dilution figure for a whole region and ending with more sophisticated approaches based on data on the number and discharge flows of sewage treatment plants and the flow regimes of rivers of a major catchment. Additional site-specific input data may include hydraulic geometry of channels, sediment load, terms of the catchment water balance, physical and chemical characteristics of the water body and its suspended and bottom sediments, etc. In addition, basic dilution models require data on the fraction of the chemicals released to the drain and its removal in different types of sewage treatment. Site-specific models require additional data on chemical loading for each simulated compartment. For spill models, volume and density of the spill are needed [260], [261].

For specific purposes, aquatic equilibrium modeling may become necessary to predict environmental concentrations more reliably. To this end, fast reactions of inorganic species dissolved in aqueous media can be classified under the following categories:

1) Reactions with solvent molecules, in particular dissociation
2) Substitution reactions with solvent or dissolved species
3) Redox reactions with dissolved gases or other ionic species

These reactions can lead to the formation of new complex ions with different ligands in the coordination sphere, ions with different oxidation state of the metal centers, or still other ions. The reactions are essential, since the medium in which the particular inorganic substance is dissolved largely determines the speciation of the particular metal ion. In addition, upon mixing of solutions of these metals with the environment, chemical modification of the species will occur initially by thermodynamically favored rapid reactions and subsequently by slower reactions [262–264]. The following slower reactions are important in aqueous media [258]:

1) Ligand substitution reactions of kinetically relatively inert ions
2) Electron-transfer reactions involving inner-sphere mechanisms for relatively inert ions and some outer-sphere reactions
3) Reactions with dissolved gaseous species or bacterially induced reduction
4) Precipitation of solids by formation of insoluble species through substitution reactions
5) Formation of metal hydroxides and sulfides by oxidation reactions

These slower reactions impact the speciation and concentration in the aqueous phase and determine both type and extent of further interactions in a way comparable to fast reactions. Therefore, the time frame for some of the above reactions may indicate that characterizing the speciation in the original aqueous phase will also define the chemical behavior of the compound in the environment, whereas in other cases, understanding speciation implies exact knowledge of environmental interactions. To this end, a larger number of speciation models have been developed to address a series of specific situations. They are based on chemical thermodynamic reactions and principles, and also account for additional processes such as redox reactions [265], adsorption [266], complexation [267–269], ligand exchange [270], and others [271].

3.2. Effects Assessment

The assessment of possible effects of chemicals on the environment involves the determination of no-effect concentrations (PNECs) for aquatic species, edaphic species, and for species that may be affected by ecological magnification, i.e., by accumulation of substances through the food web. In addition, assessment procedures are indicated for substances that may have a negative impact on natural physicochemical processes in the atmosphere (for definitions and terminology of risk, see Appendix, Chap. 4).

Hitherto the predictive assessment of dose – effect relationships between chemicals and biotic elements in the environment has usually started with laboratory tests. Their results need extrapolation to the complex real-world situation, which proves rather difficult, as a rule,

since the experiments are performed with defined amounts of chemicals in contact with a single species or restricted group of organisms. In contrast, the integral approach to assessing the effects of potential pollutants in the natural environment has to deal with the manifold interactions and complicating factors or boundary conditions, which are carefully avoided or controlled in the laboratory situation (see also → Toxicology, Chap. 3., → Toxicology, Chap. 4., → Toxicology, Chap. 5.).

The following are most prominent of these complicating factors in the natural environment [272]:

1) Uneven distribution of the pollutant in an ecosystem compartment because of gradients or currents, dilution, volatilization, sorption processes, etc. Bioaccumulation in successive elements of the food web may cause considerable variations in the individual load of different species.
2) Differences in specific sensitivity may lead to shifts in the abundance and diversity proportions of the biocenoses affected. Interspecific relationships, synergistic relations, and food web effects can be analyzed only in multispecies systems.
3) In a natural ecosystem, the effect of a pollutant on the biocenosis will be greatly influenced by weather, diurnal and seasonal variations, simultaneously occurring variable loads such as thermal and chemical loads, along with intermittent discharges.

To a certain extent, the influence of these complications can be accounted for by simulation techniques in the laboratory, but practical limitations necessarily exist. On the other hand, retrospective field experiments alone cannot be relied on since the doses are not known exactly, intentional pollution for experimental purposes is frequently prohibited, and biological observations are time-consuming and expensive. As a consequence, comparative tests should be carried out with selected model chemicals under both laboratory and natural conditions to determine the relative representativity of the test procedure. This validation of the tests should also pay appropriate attention to the important question of how far the test system or the experimental field plot is representative from a regional point of view [160].

In comparing this comprehensive, normative list of ecotoxicological models with the (still) predominant present-day practice, a fairly marked tendency exists to favor the bioindicator approach and aquarium or microcosm tests or technically comparable procedures [273], [274]. This has the appreciable advantage of relative simplicity and ensures a fairly high amount of reproducibility, but also poses the above problems as to rationally and reliably extrapolating to the complex real-world situation.

3.2.1. Test Systems for Aquatic Ecosystems

3.2.1.1. Level 0

Test procedures on the infraorganismic level (i.e., organells, cells, tissues) are omitted here and the hierarchical system suggested in the OECD Chemicals Testing Programme is followed [275]; consequently three successive levels of sophistication labeled 0, 1, and 2 are distinguished. The results of level 0 tests together with an increase in production of the chemical considered and its type of use pattern should indicate that subsequent tests on higher levels have to be carried out.

The following set of studies is considered relevant for the level 0 dossier:

1) Acute toxicity study with *Daphnia magna*
2) Acute toxicity study with fish
3) A static biodegradation test

According to the 7th Amendment to the Directive 67/548/EEC, Annex VII A, the minimal data basis (base set) for effects assessment purposes comprises acute tests for fish, *Daphnia*, and algae. In the OECD testing hierarchy, the latter forms part of level 1 enquiries.

Toxicity Test with *Daphnia magna*. The water flea *Daphnia magna* (Daphniidae, Cladocera) is a suitable test organism for static and semistatic tests and for flow-through systems. The results of these tests are relevant for initial hazard assessment because of the trophic importance of daphnids as primary consumers, their sensitivity to chemicals, and their relationship to various other ecologically important aquatic invertebrates. Also, automatic biomonitoring devices are feasible. The tests usually consist of two phases:

1) An acute phase that gives the 24-h EC_{50} value, the highest concentration that still causes no immobilization (EC_0), and the lowest concentration that causes 100 % immobilization (EC_{100})
2) A reproduction phase yielding the EC_{50} and LC_{50} values at 24, 48, and 96 h; 7 and 14 d; and the end of the test (further information is provided by time span until the emergence of the first brood, the number of live and dead young, their condition, etc.)

The slope of the time – response curve depends largely on the speed and degree of absorption, inherent detoxification, and excretion mechanisms. While a steep slope usually indicates rapid absorption and toxic effects appearing within a short period of exposure, a shallow slope may be due either to low absorption or to toxic effects delayed beyond the period of observation.

The 14-d effect level in relation to the 24-h LC_{50} is, to a certain extent, indicative of the potential for chronic toxicity of a compound. Since the test duration is limited to two weeks, however, it may provide misleading results if the chemicals considered are taken up slowly or have delayed toxic activity. Yet the *Daphnia* test is in any case useful for determining whether further ecotoxicity studies are necessary to indicate which parts of an ecosystem or which species are most sensitive. A low effect concentration may trigger further tests while, conversely, a relatively high *Daphnia* effect concentration may guide any further testing away from invertebrates [276], [277].

Acute Toxicity Studies with Fish. Fish are of particular importance in aquatic environments, and their survival in waters bears witness to the absence of toxic substances or adverse conditions for the species in question. Consequently, toxicity tests with fish have been carried out in laboratories for about a century, but not infrequently the results obtained have differed widely because of lack of standardization of test organisms and boundary conditions.

Therefore, considerable numbers of studies have been undertaken since the 1970s to define the parameters involved in performing toxicity tests: taxonomy, age, and size of fish; composition of environments, in particular pH value, buffering capacity, mineral and organic content, temperature, and oxygenation; duration of contact with toxic substances; etc. The result has been the perfection of numerous techniques that can be applied in fresh water or seawater; they yield satisfactory results, depending on their objectives, although they cannot always be applied readily to other cases in view of major differences in the experimental conditions and the sensitivity of species selected.

Species frequently used in such studies are *Brachydanio rerio*, *Carassius auratus*, *Leuciscus idus*, *Phoxinus phoxinus*, *Rasbora heteromorpha* (Cyprinidae); *Jordanella floridae*, *Oryzias latipes* (Cyprinodontidae); *Poecilia reticulata* (Poeciliidae); and *Oncorhynchus mykiss* (Salmonidae). They comprise species that occur in aquatic environments into which the chemicals under study are released and so-called aquarium fish, which (frequently) are not representative of natural environments but are more suitable for screening tests.

The acute toxicity test on fish [275] is specifically designed to estimate the LC_{50} with observations after 24, 48, 72, and 96 h. However, the longer durations are reported to yield possibly misleading results because test organisms are not fed during the experiment [278]. The resulting data are used to prepare concentration – mortality curves with LC_{50} values, which also permit estimation of the maximum concentration causing no mortality and the minimum concentration causing 100 % mortality. Where possible, LC_{50} values are also estimated for each of the recommended observation times, thus giving a more precise indication of the time – toxicity curve.

As for Daphniidae, comparison of the LC_{50} values with the potential environmental concentrations of a chemical will provide a first crude estimate of the potential hazard to fish. When the slope of the curve is steep and soon tends to be asymptotic to the time axis, longer exposure is not likely to result in a significant increase in mortality unless noxious effects that result from different mechanisms are delayed. Conversely, where the slope is shallow, cumulative toxicity and subsequent chronic effects may be indicated and an extension of the test period could result in a considerably lower LC_{50} estimate. For a more detailed review of the increasing range of types of acute toxicity tests on fish, see [47],

[275], [278], [279]. In view of the numerous tests suggested, however, one should note that environmentally acute lethal toxicity is not the major problem. Acute lethal poisoning is easily detectable and should be readily controllable as well. Thus, subchronic and chronic exposures are likely to be the sources of problems. Consequently, obtaining information on the approximate magnitude of acutely lethal concentrations over an exposure period of up to 96 h might appear sufficient: "The greater the range of water qualities in which the acute LC_{50} is measured and the greater the variety of fish species used, the more likely is it that sufficiently cautious estimates can be arrived at. In this field of toxicity testing it is more important to get some measure of the variability existing, and ever perhaps try to account for it, than it is to mask any real differences which exist by conveniently standardizing tests. (Standard tests with fish appear, in fact, not to be concerned with the assessment of toxicity as such but rather with reproducibly bioassaying the concentration of a given chemical in some particular quality of water)" [280].

Tests for Biodegradation. Assessment of the environmental hazard of a particular chemical involves a prediction of interactions that may lead to degradation of the chemical or its eventual accumulation in particular environmental compartments. In the framework of the OECD Chemicals Testing Programme [275], the following test procedures *defining ready biodegradibility* have been suggested:

1) Modified AFNOR test
2) Modified Sturm test
3) Closed bottle test
4) Modified OECD screening test

These tests provide limited opportunities for biodegradation and acclimatization. Hence, a chemical giving a positive result may be assumed to biodegrade in the environment rapidly and, therefore, may be classified as "readily biodegradable."

Modified AFNOR Test. The modified AFNOR test requires a test material that must at least be soluble at the concentration tested [40 mg dissolved organic carbon per liter (DOC/L)] have negligible vapor pressure, not be inhibitory to bacteria, and not significantly adsorb on glass surfaces. The test substance is degraded by chemicoorganotrophic microorganisms added in low concentrations with the products used as the sole source of carbon and energy. The organic carbon remaining in solution after 3, 7, 14, 28 (and 42) d is measured, and the corresponding level of biodegradation calculated. A detailed description of the test, which is sensitive to an order of magnitude of $\pm 10\,\%$, is given in AFNOR T-90-302 Test and the OECD Guideline 301 A [231].

Modified Sturm Test. The modified Sturm test has the same requirements as the preceding one, but uses the amount of CO_2 produced as indicator variable [281]. To this end, a chemically defined liquid medium, essentially free of other organic carbon sources, is spiked with the test substance and inoculated with a low concentration of sewage microorganisms. The CO_2 released is trapped as $BaCO_3$. By reference to suitable blank controls, the total amount of CO_2 produced by the test compound is determined and calculated as the percentage of total CO_2 (TCO_2) that the chemical tested could have theoretically produced in relation to its carbon content. In the absence of toxic effects on the inoculum, the reproducibility of the Sturm test is $\pm 5\,\%$. Unless the test is adapted to handle ^{14}C-labeled compounds, test chemical concentrations cannot be lower than 5 mg/L [231].

Closed Bottle Test. The purpose of the closed bottle test is measurement of the biodegradability of organic compounds in an aerobic, aqueous medium at a test concentration of 2 (i.e., standard concentration) to 10 mg/L of active material. The degradation or biotransformation is given as the biochemical oxygen demand (BOD) within 28 d as a percentage of either the theoretical oxygen demand (TOD) or the chemical oxygen demand (COD) [231]. Most data elaborated with this test pertain to water-soluble compounds, but volatile compounds and those of low water solubility may also be tested, at least in principle. In the closed bottle test, a predetermined amount of test substance is dissolved in a mineral nutrient solution to provide the above concentration of active substance per liter. The solution is inoculated with a small number of microorganisms from a mixed population (derived from unfertilized garden soil and activated sludge) and kept in closed bottles in the dark in a constant temperature bath at $20 \pm 1\,°C$. Degradation is followed by oxygen analyses over a

28-d period. A control with inoculum, but without test material, is run parallel for the determination of oxygen blanks. The reproducibility of the test is appropriate for screening purposes but not sufficient for a final decision regarding biodegradability. Although most experience has been gathered with water-soluble chemicals, the structure of the tests also permits, in principle at least, application to volatile and insoluble compounds.

Modified OECD Screening Test. The modified OECD screening test aims at measuring the ultimate biodegradability of water soluble, nonvolatile organic compounds in an aerobic, aqueous medium with a starting test concentration of 5 – 40 mg DOC per liter. Degradation is calculated as the percentage DOC removal within 28 d with respect to the test material. The compound to be tested is dissolved in an inorganic medium (i.e., a mineral nutrient solution, fortified with a trace element and essential vitamin solution), providing the above concentration. The solution is inoculated with a small number of microorganisms from a mixed population and aerated at 20 – 25 °C in the dark. Degradation is determined by DOC analysis over the 28-d period, which is checked by means of a standard. A control with inoculum, but without test material or standard, serves to determine DOC blanks [231]. Data obtained from these and similar tests are applicable only to aerobic biodegradation. Since a more comprehensive evaluation of the possible fate of a chemical must also take anaerobic situations into account, other test procedures may become necessary.

Other Tests for Ready Biodegradability. At least some 20 additional methods have been suggested to evaluate biodegradation in surface waters or sewage treatment plants [47]. On the basis of comparative degradation tests, Gerike [282] came to the conclusion that determination of DOC removal usually yields the most reliable and relevant results and could therefore constitute the major source of primary information on ready biodegradability.

Tests for Inherent Biodegradability. Substances that are not readily biodegradable in terms of the above or comparable tests may be examined with the aim of assessing their intrinsic potential for biodegradation (i.e., their liability to microbial breakdown under favorable conditions). At present, the modified Zahn – Wellens test and several methods for assessing the biodegradation potential of soil are available for determining inherent biodegradability [283].

The *Zahn – Wellens test* [284] has a number of advantages. As a batch die-away test, it employs activated sludge, mineral nutrients, and the test substance in solution as the only source of carbon. The mixture is agitated and aerated in a 1 – 4-L glass under diffuse illumination for up to 28 d. Degradation is monitored by determining the DOC (or COD) values in the (filtered) solution at regular intervals. The ratio of eliminated DOC (or COD) after each interval to the initial concentration is expressed as percentage biodegradation; plotted versus time it gives the biodegradation curve. This test provides a good indication of inherent biodegradability and permits acclimatization phenomena to be detected from the degradation – time relationships recorded. Its drawback lies in the high initial concentration of the test substance, which may exert toxic effects on the inoculum [284]. Like some others, the Zahn – Wellens test is suitable only for water-soluble, nonvolatile compounds.

The principle of the inherent biodegradability test in soil is described in OECD Test Guideline 304 A [234]. A small (50-g) sample of soil is treated with the ^{14}C-labeled test chemical in a biometer flask apparatus. The amount of $^{14}CO_2$ released from the test substance is absorbed in a 0.1 N KOH solution and determined by liquid scintillation counting. The test is one example of a class of indirect estimations of microbial degradation processes that generally rely on some unique characteristic of the compound in question, such as the release of $^{14}CO_2$ or of chloride from organics containing organically bound chlorine. Destruction of aromatic compounds in model systems can also be detected in the loss of ultraviolet light absorbance as the aromatic ring is cleaved [285].

Direct Measurement of Ecotoxic Substances. Among the direct measurement methods, gas chromatography provides a rapid, simple, and highly sensitive tool, especially in combination with mass spectrometry and in the form of multidimensional headspace capillary gas chromatography. The latter method increased the reliability of the qualitative and quantitative analysis of volatile compounds by the simultaneous chromatographic separation of samples in two

capillary columns of different polarity. It is not affected by disturbances due to salts and compounds with higher molar mass and avoids effects otherwise exerted by the extractant used. By virtue of its high degree of mechanization, headspace injection can be automated easily.

Bioassays. In bioassays for the degradation of toxic compounds, a plant, insect, microorganism, or other suitable indicator organism that is sensitive to the treated sample is exposed at regular intervals after initial modification of the soil. The substance or its inhibitory derivatives are assumed to remain as long as the susceptible species or strain fails to grow or develops atypically, and the length of incubation of the soil required before the indicator organism starts developing in the test sample as well as in untreated soil is considered a reflection of the time at which the toxic component has dissipated [286].

3.2.1.2. Level 1

Subject to positive results from the acute toxicity tests of level 0, an increase in production of the compound, and a risk estimate associated with the use pattern intended, the following test procedures are foreseen for the level 1 dossier:

1) Chronic toxicity study with *Daphnia magna*
2) Subacute toxicity study with fish
3) Algal test
4) Test for species accumulation
5) Additional biodegradation test

Chronic Toxicity Study with Daphnia Magna. The chronic toxicity study with *Daphnia magna* is thought expedient if the LC_{50} appeared lower than 100 mg/L, and it should also include determination of the "no-toxic-effect levels" for both reproduction and lethality. For higher LC_{50} values, such a test appears justified only on the condition that biodegradation has not been proved by the relevant test carried out on level 0.

Subacute Toxicity Study with Fish. The subacute toxicity study with fish is a medium-term test with an exposure time of at least 14 d. As such, it is generally dynamic and based on the substantiation of lethal or sublethal effects on growth, behavior, or reproduction, including the determination of "threshold levels." If the LC_{50} for fish has been assessed at values lower than 1000 mg/L, the same species is to be used for the subacute toxicity study. For higher LC_{50} values, such a test appears necessary only if the previous biodegradation study yielded inconclusive results [287].

Algal Test. An algal test [275] enables a determination of the effects of a substance on the growth of unicellular green algae under laboratory conditions over a period of several cell generations. Since algae, as primary producers, are at the very base of aquatic food webs, their increase or decrease may influence higher trophic levels in natural aquatic ecosystems. In this respect, the possible toxic influences of a chemical on algal growth are highly relevant for the assessment of potential effects in the natural environment. The value of the test is, however, necessarily limited by the inherent difficulty of extrapolating laboratory conditions to the natural situation with its involvement of many other biotic and abiotic factors.

To obtain reproducible results, the algae (frequently *Chlorella pyrenoidosa*, *Scenedesmus quadricauda*, or *Scenedesmus subspicatus*) are exposed to the test substance at defined concentrations for 96 h. The algal growth, measured by the increase in biomass or cell number, is determined at least 24, 48, 72, and 96 h after the start of the test, and the corresponding concentrations per milliliter are calculated. The EC_{50} value is calculated graphically or by computation.

For assessment purposes, both the intrinsic growth rate (i.e., slope of the curve during logarithmic growth) of the control and experimental algal population and the total yield have proved useful. Even if the 96-h EC_{50} value does not seem to reflect high toxicity, this may be due to the fact that the test chemical has degraded considerably in the course of the test.

Test for Species Bioaccumulation. The potential of a chemical to bioaccumulate in aquatic organisms is determined by means of a sequential static test for species bioaccumulation, for which fish are preferably the animals of choice. Bioaccumulation is expressed by the bioconcentration of the chemical in the test organism relative to that in water under steady-state conditions [234]. In the first step of the test procedure,

the time interval necessary to reach the steady-state concentration is determined by exposing one group of fish to one concentration of the test chemical without renewal of the test solution. In the second step, the dependence of bioaccumulation on various (i.e., at least three) concentrations of the test chemical is determined, and the result obtained is expressed in form of a sorption isotherm. In the third step of the procedure, the duration rate is measured. To this end, fish exposed to the maximum test compound concentrations in the preceding step are transferred to noncontaminated water from which the depurated compound is continually removed. The residual concentration in the fish is determined periodically until about 80 % depletion is reached or for a period of up to ten days, whichever comes first.

The test for species bioaccumulation was established with ^{14}C-labeled pesticides such as 2,4-D, atrazine, DDT, and others; consequently, such compounds can be appropriately used for standardization. The use of nonlabeled chemicals as standards requires highly sensitive (ppb-range) and compound-specific analytical methods. With respect to specificity, the test is suitable for nonionized organics that are not readily biodegradable, are relatively stable in the aquatic environment (i.e., half-life >4 d), are soluble in water at <2000 ppm (μg/g), and have an *n*-octanol – water partition coefficient (K_{ow}) >1000 and a volatility (i.e., water – air partition coefficient) >100.

With regard to direct bioaccumulation in individual species, it clearly correlates with the degree of lipophilicity, expressed in terms of the *n*-octanol – water partition coefficient (see Section 2.4.1). Exceptions normally relate to cases where K_{ow} overestimates the bioconcentration factor. The higher the K_{ow} of a chemical (i.e., the more lipophilic it is) the slower is it depurated (desorbed) from the organism. Since this process does not always follow first-order kinetics, however, apparent half-lifes should not be extrapolated beyond the ranges measured.

In addition, reactions of chemicals in living organisms may result in derivatives that are more lipophilic than the parent compound or in binding (parts of) the chemical to special cell constituents. The latter applies, in particular, to inorganic or organometallic chemicals. Clearly, in either case, the K_{ow} value of the parent compound as such is not an appropriate indicator of the bioaccumulation potential. For organic compounds ionizing under physiological conditions (i.e., in the pH range 3 – 9), K_{ow} is predictive only if the chemical has no specific reactivity to cell constituents.

If the studies carried out on level 0 did not yield a satisfactory amount of information about biodegradation, another *static biodegradation test* can be chosen. If its result is negative again, a dynamic test in a flow-through system on the basis of lower concentrations, together with another inoculum, should then be used.

3.2.1.3. Level 2

The results of studies carried out on test level 1 are preferably to indicate which tests are to be made on level 2; this level is characterized by

1) An additional test for bioaccumulation
2) A chronic toxicity study with fish
3) An acute and a subacute toxicity test with another organism

Additional Bioaccumulation Studies. In the framework of supplemental bioaccumulation studies, the water solubility and *n*-octanol – water partition coefficient of the chemical in question merit particular attention (Section 2.4.1). Depending on the results of the previous accumulation test on level 1 and the physical and chemical properties of the compound, a flow-through test may be appropriate. It provides an opportunity for more adequate regulation of test conditions, including actual exposure concentrations, and is not limited in duration, thus allowing more comprehensive assessment of the significance of exposure time in evaluating bioaccumulation. The test substances are dissolved, if necessary with surfactant, in water. The solution is introduced continuously into an aquarium with constantly flowing water where the test fish are kept. The concentrations of chemicals in both water and test fish are analyzed, and the BCF is calculated. Since the variability of bioconcentration factors in several fish species is fairly high, it may be useful to achieve a higher degree of normalization by relating BCFs to the lipid content of the fish.

A technical advantage of the easily applicable test [288] is the possibility of automatic control.

The dissolved oxygen concentration, pH, water temperature, etc. can be automatically and continuously measured and recorded.

Chronic Toxicity Study. A chronic toxicity study with fish is indicated if the "no-effect level" for reproduction is more than a factor 10 lower than the no-effect level for lethality of *Daphnia magna*. The same applies if the threshold level for fish as determined under level 1 is 10 times lower than the LC_{50} for fish. According to the relevant OECD Test Guideline [287], fish such as the aforementioned *Poecilia*, *Oryzias*, and *Jordanella* species should preferably be used as test organisms. On the basis of present limited experience, the tendency is to believe that the results of a toxicology study with a second species are not likely to change the estimated "no-toxic-effect" level to a great extent. Therefore, acute and subacute toxicity studies with another aquatic or terrestrial species, for instance the Japanese quail, should be performed only if the need is really indicated.

Representatives of the trophic levels described are listed below. Standard organisms are underlined. The assignment of an organism to a trophic level is based on the energy balance of the ecosystem considered and does not depend primarily on the species. Therefore, a population may represent more than one trophic level in dependence on the nutrition available for the species. In addition, earlier life stages often live on completely different nutrition compared to adults of the same species.

Primary Producers. Primary producers photo- or chemoautotrophically synthesize organic compounds using inorganic precursors. They include

1) Chlorophyll-containing species of vascular plants
2) Algae (e.g., green algae: *Selenastrum*, *Scenedesmus*, *Chlorella*; blue-green algae: *Microcystis*)
3) Purple sulfur bacteria, chlorobacteria
4) Chemoautotrophic bacteria (nitrifying bacteria, sulfur bacteria)

Primary Consumers. They live mainly on living or dead autotrophic organisms or on microorganisms. Representatives of this trophic level are especially plant-eating animals (i.e., species that are not carnivorous) of the following taxonomic groups:

1) Protozoa (e.g., *Uronema, Entosiphon, Tetrahymena*)
2) Annelids (e.g., *Tubifex, Enchytraeus*)
3) Crustacea (e.g., *Artemia*, *Daphnia sp.*, *Copepoda*, *Gammarus*, *Asellus*)
4) Mollusks (e.g., *Dreissensia, Mytilus*, *Ostrea*; several gastropods: *Patella, Viviparus*)
5) Insects (some insect larvae that are not carnivorous)
6) Nematodes (species that live in water)

Secondary Consumers. They live mainly on primary consumers. Among them are

1) Predatory insects and larvae of insects (e.g., *Chaoborus*)
2) Carnivore protozoa
3) Rotatoria
4) Coelenterates (e.g., *Hydra*)
5) Predatory copepods
6) Fish (Teleostei: e.g., *Cyprinus carpio, Brachydanio rerio, Poecilia reticulata, Oryzias latipes, Pimephales promelas, Lepomis macrochirus, Oncorhynchus mykiss* (former: *Salmo gairdneri*) *Leuciscus idus melanotus, Cyprinodon, Carassius*
7) Amphibians (e.g., *Rana*, *Xenopus*)

Decomposers. Organisms of this trophic level break down dead organic material to inorganic constituents. Representatives are

1) Bacteria (most species)
2) Fungi

Test organisms for benthic ecosystems

Biological group/test organism	Parameter tested
Benthic invertebrates	
Panagrellus redivivus	survival
Caenorhabditis elegans	survival
Tubifex tubifex	survival
Stylodrilus heringianus	survival, repellency, rate of transformation of the sediment, growth
Hyalella azteca	survival, growth, reproduction
Pontoporeia hyi (*Diporeia sp.*)	survival, repellency
Corbicula fluminea	survival, growth
Anodonata imbecilis	survival
Chironomus tentans	survival, growth, hatching
Chironomus riparius	survival, growth
Hexagenia limbata	survival, frequency of exuviation
Macrobenthic biocenoses	indices for biocenoses and populations
Macrophytes *Hydrilla verticilata*	length of shoots and roots, dehydrogenase activity, chlorophyll-a, peroxidases

3.2.1.4. PNEC Values and Assessment Factors

The function of hazard or risk characterization and assessment is the overall protection of the environment. With regard to aquatic ecosystems, two habitual assumptions are therefore made to extrapolate from the above single-species short-term toxicity data to ecosystem effects: (1) ecosystem sensitivity depends on the most sensitive species, and (2) protecting community structure protects community function. While the latter assumption is hardly questionable, the former needs further examination in light of the introductory statements to Chapter 3.

For many existing and most new substances, the pool of data on ecosystem effects is very limited; i.e., only short-term toxicity data are available as a rule. This implies the application of assessment factors that permit one "to predict a level at or above which the balance of probabilities suggests an environmental effect will occur. It is not intended to be a level below which the chemical is considered safe. However, again, the probability is that there will be no effects" [289]. When defining assessment factors, a number of uncertainties must be taken into account in attempting to extrapolate from mono- or oligospecific laboratory test systems to multispecies ecosystems: (1) intra- and interspecies variations, (2) acute to chronic toxicity extrapolation, and (3) extrapolation of laboratory data to field impact. Additional uncertainty results from the different testing methods themselves or interlaboratory variation in toxicity data.

When reviewing data from laboratory test systems the assessment factors of Table 17 are proposed. Instead of using assessment factors on the lowest-effect data, statistical extrapolation methods may be applied to a (relatively large) set of long-term data as an alternative approach. The underlying assumptions are [292]:

1) The distribution of species sensitivities follows a theoretical distribution function
2) The groups of species tested in the laboratory are a random sample of this distribution

This implies the use of no-observed-effect concentrations (NOECs) for identical parameters (e.g., reproduction) of the species considered [293], especially if only limited data from different species are available [294]. Normally long-term toxicity data are log-transformed and fitted according to the distribution function, and a prescribed percentile (e.g., 95 % [292]) of this distribution is used as a criterion. This means that the NOEC may be exceeded for 5 % of the species in the community. The validity of some of the above assumptions may be questioned (see also [294], [295]). First of all, it must be taken into account that for most compounds few longterm NOECs are available so that most tests on distribution functions have low power. Second, the NOECs available are not a random sample. Emans et al. [296] compared NOECs derived from multiple-species experiments for aquatic systems with extrapolated values [297], [298] and came to following conclusions:

1) Validation is rendered difficult by the lack of reliable data from multiple-species tests and long-term monospecific tests
2) Results from both statistical extrapolation approaches do not differ significantly
3) Results of the statistical extrapolation with 5 % confidence are in good agreement with results from (semi-) field tests

Table 17. Assessment factors for aquatic toxicity testing (after [289–291])

Toxicity data	Factor		
	EC	OECD	ECETOC∗
One short-term toxicity value: algae, daphnia, or fish		1000	
Short-term toxicity: algae, daphnia, and fish	1000 [lowest L(E)C_{50}]	100	200
Long-term toxicity: fish or daphnia	100 (NOEC)		
Long-term toxicity: algae, daphnia, and fish	10 (lowest NOEC)	10	5
Long-term toxicity: two species in two taxonomic groups	50 (lowest NOEC)		
Field or mesocosm toxicity	case-by-case evaluation		1

∗ ECETOC = European Chemical Industry Ecoloy and Toxicology Centre.

3.2.2. Testing Strategies for Terrestrial Ecosystems

Soil constitutes a central compartment of terrestrial ecosystems. Chemicals introduced into soil can be adsorbed on soil constituents or transformed by soil organisms; furthermore, they may be taken up by plants, washed out by rain or irrigation water, or evaporated in gaseous form. The extent of adsorption is controlled by various soil properties, including composition and content of organic matter, type and content of clay minerals, ion-exchange capacity, and acidity; it is also closely connected to the physical and chemical properties of the substance in solution (see Section 2.4.2). In view of this complex situation, the empirical knowledge of the behavior of environmental chemicals is still fairly limited, and the possible danger of excessive fertilizer or manure application, pesticides, and heavy-metal contamination of soils and crops has been recognized only in the last decades, mostly in relation to the agriculture use of sewage sludge or in connection with atmospheric deposition of organics or heavy metals.

Given the fundamental importance of soil for biological conservation, fertility, element fluxes, biodegradation of organic materials, and retention and filtration of groundwater, broad agreement exists as to the necessity of specific, use-oriented risk assessment procedures. As for risk characterization in aquatic ecosystems, the predicted no-effect concentration in soil ($PNEC_{soil}$) must be compared with the predicted environmental concentration in soil and subsoil (PEC_{soil}) as described in Section 3.1.4.

3.2.2.1. PNEC Estimates for Soil and Subsoil

Terrestrial Tests and Assessment Factors. A substance discharged into soil can have noxious effects on both the soil's physicochemical properties and its biocenoses. If the pore structure of a soil is affected by the chemical, the soil might silt up and be compacted; thus, the percolation rate is reduced considerably, the water balance is impaired (see Section 2.1.2), and the air supply is impeded. Since soil acts as a pH buffer system and a reservoir for bound substances, impairment of the soil adsorption capacity by xenobiotics can infer exhaustion of the buffer capacity. Impairment of microbial activity implies a reduction of the catabolic and metabolic capacity of the soil, which may lead to retarded liberation of nutrients and accumulation of mainly organic substances; this implies further effects on soil-dwelling organisms [35]. A chemical represents a risk to the soil compartment of an ecosystem if its concentration exceeds the degradation potential of the soil or if it changes the soil properties – in particular, the filtration, buffering, and transformation functions – in such a way that adverse effects result for the environment [299]. Thus, data on geoaccumulation as obtained in the framework of exposure assessment may give a first indication of possible changes in soil properties. The $PNEC_{soil}$ can be derived from the results of toxicity tests on soil and subsoil organisms by the application of assessment factors. If no ecotoxicological data are available for such organisms, the equilibrium partitioning method may be applied as a first approximation. It assumes that the bioavailability and therefore the toxicity of chemicals are determined by their concentration in the pore water of soil. The applicability of the method has been

tested less for soil- than for sediment-dwelling organisms; VAN GESTEL [300] has demonstrated its validity forearthworm toxicity for several chlorophenols, chlorobenzenes, and chloroanilines. Whether this also holds true for other compounds or species is questionable, however (e.g., lipophilic compounds or organisms that are exposed primarily via food).

For few existing chemicals, short-term terrestrial tests are available with microorganisms, earthworms, and higher plants; in addition, a long-term test is used with springtails. For new substances, two standardized short-term tests with higher plants and earthworms are required at level 1. If no acute tests are available, tests on primary producers, consumers, and decomposers should be applied for the preliminary stage of the concentration-effect assessment [301–304]. The Umweltbundesamt [290] suggests use of the same assessment factors for terrestrial as for aquatic ecosystems, depending on the type of investigations, (i.e., acute or long-term toxicity test) and the number of trophic levels studied (see Table 18).

Table 18. Assessment factors for terrestrial ecosystems (after [290])∗

Information available	Assessment factor
L(E)C_{50} for preliminary stage toxicity tests (e.g., acute toxicity to plants and earthworms or microorganisms)	1000
NOEC for one additional long-term toxicity test	100
NOEC for additional long-term toxicity tests for two species of two trophic levels	50
NOEC for additional long-term toxicity tests for three species of three trophic levels	10
Field data/data of model ecosystems	case by case

∗ The $\mathrm{PNEC_{soil}}$ is calculated on the basis of the lowest effect value measured.

Standardized long-term tests for earthworms [305] and for springtails [304] are available, both of which analyze effects on reproduction. The standardization of two long-term tests is close to completion: the test on staphylinids (Coleoptera) registers degree of parasitism, hatching rate, and reproduction [306]; the test on *Enchytraeus* (Annelida) can replace the reproduction test on earthworms [307]. VAN STRAALEN and VAN GESTEL [308], STAVOLA [309], and PEDERSEN and SAMSØE-PETERSEN [310] discuss test methods with various terrestrial species and different degrees of standardization.

Evaluation of Stress Tolerance of Soils as Ecosystem Compartments. To comply better with the exigencies of a provision-oriented environmental policy, the critical loads of soil with regard to nutrients, heavy metals, and xenobiotics require a generalized ecological evaluation, in addition to estimates of human exposure to contaminants in soil (see → Toxicology, Chap. 5.; → Soil, Chap. 6.2.). This implies the rational definition of soil standards that aim at sustaining appropriate physicochemical and biotic soil functions.

A recent approach of this type is based on the regionalized determination of geogenic and anthropogenic concentrations of potentially hazardous chemicals, and the assessment of different degrees of adverse chemical effects on soil and related biota [299] (see Table 19). Thus, the preventive threshold value C_{p} characterizes chemical concentration levels not affecting any essential ecological soil function. The C_{eff} value is indicative of first and low-level adverse effects and is defined by means of ecotoxicological test procedures and determination of the physicochemical behavior of xenobiotics in soil. By contrast, the C_{dis} value indicates long-lasting or permanent disturbance of essential biotic and abiotic soil functions. Among the extrapolation methods for determining ecologically relevant effect concentrations [293], [311–316], the last mentioned appears most appropriate. It defines a hazard concentration ($\mathrm{H}C_{\mathrm{p}}$) that ensures the protection of a certain number of soil-living animals. Thus, the approach is based explicitly on resilience considerations (see Section 2.1), using NOECs as reference values. The hazard assessment involves four steps:

1) Determination of at least five NOECs
2) Standardization of NOECs values with regard to a standard soil from at least three taxonomic groups according to the following equation

$$\mathrm{NOEC}' = \mathrm{NOEC}\,(L,H)\,\frac{R\,(25,10)}{R\,(L,H)} \tag{64}$$

where NOEC′ is the NOEC of standard soil; NOEC (L, H) ist the experimental NOEC

with given clay (L) and organic matter content (H); $R(25, 10)$ is the reference value of standard soil with 25 % clay and 10 % organic matter; and $R(L, H)$ is the reference value on the basis of the clay and organic matter content of the test soil.

3) Determination of a safety factor T in relation to the different sensitivity of indicator species

$$T = \exp\left[\frac{3S_{\mathrm{m}}d_{\mathrm{m}}}{\pi^2}\ln\left(\frac{1-\delta_1}{\delta_2}\right)\right] \tag{65}$$

where S_{m} is the standard deviation of ln NOECs; d_{m} is a factor dependent on the number of test species; δ_1 is the proportion of nonprotected species, related to defined hazard concentration (HC_{p}); δ_2 is the probability of overestimating HC_{p}; and T is a safety factor for NOEC average.

4) Based on this safety factor the hazard concentration HC_{p} is defined as

$$HC_{\mathrm{p}} = \frac{\exp(x_{\mathrm{m}})}{T} = \frac{\overline{\mathrm{NOEC}}}{T} \tag{66}$$

where m ist the number of test species; x_{m} the mean value of ln NOECs; and $\overline{\mathrm{NOEC}}$ the geometric mean value of NOECs.

The exemplary application of Equation (66) to cadmium and a test system of *Dendrobaena rubida, Lumbricus rubellus, Eisenia foetida, Helix aspera, Porcellio scaber, Platynothrus peltife*, and *Orchsella cincta* would yield an HC_5 (i.e., $\delta_1 = 0.05$) of 0.16 mg Cd per kilogram of soil. In comparison, the present A level of the Dutch "Leidraad Bodemsanering" is equivalent to 0.80 mg Cd per kilogram of standard soil.

As long as the C_{eff} value, defined by means of the above prodecure, is not exceeded, no essential reduction of ecological soil functions occurs and the long-term multifunctionality of soil is sustained. In contrast to this threshold, the C_{dis} value indicates a distinctly higher concentration of pollutants in soil that exerts long-lasting or irreversible effects on biotic and regulative soil functions. For assessment purposes, basically the same approach is foreseen as for the definition of C_{eff} (i.e., multispecies determination of HC_{p} with a p around 50). Unter these circumstances, neighboring ecosystems may also be affected due to export of primary pollutants or their metabolites. Thus, the C_{dis} value characterizes a threshold that necessitates protective or rehabilitation measures. Depending on both quality and intensity of land use, these may be quite different, including, e.g., limitations of crop production or reduction of fertilizer and pesticide application [299].

Complex Biotic Indicator Systems for PNEC Estimates. Toxic effects of chemicals in a biocenosis change according to the species composition of the community exposed to them. This is particularly important when a community is continually exposed to a concentration of a substance since its species are either least sensitive to this substance or adapted to it. Consequently, the effects of the substance on a selected laboratory test organism or a monospecific assemblage not previously exposed to that particular substance may be quite different. In consequence of these and other limitations of the aforementioned tests, artificial laboratory microecosystems (microcosms that simulate some of the characteristics of natural ecosystems) are more frequently used in toxicological studies [288], [317].

Model Ecosystem Approach. The model ecosystem approach uses physical models of natural ecosystems to simulate a selected subset of the relevant physical, chemical, and biotic characteristics of a potentially affected environment. These models may vary in complexity from a bacterial culture in a test tube to a net cage isolating a part of an ecosystem. Normally, the model ecosystem involves an aquarium or other small container in which soil, sediment, water, and several trophic levels of organisms are kept. Thus, METCALF and coworkers [318] studied the effects of chemicals in a terrestrial – aquatic interface to determine their potential to accumulate in food chain organisms and to degrade. Tests with the *Metcalf microcosm* have been standardized to permit comparison of different chemicals. Although this system does not allow for prediction of absolute concentrations of a substance expected or observed in a wide variety of actual environmental conditions, it has shown positive relationships among physical, chemical, and biotic properties, (e.g., water solubility and bioconcentration potential). However, it has not been used typically to determine rates of dominant processes, nor has it

Table 19. Complex ecological threshold values for soil quality characterization [299]

Natural background "clean soil"	Range of tolerable disturbance no further chemical stress	Range of permanent heavy disturbances changes in land use, hazard assessment, and restoration indicated
C_p	C_{eff}	C_{dis}
No reduction of ecological soil functions		
Multifunctionality: optimum range of ecological soil functions	first low-level adverse effects on ecological soil functions	long-lasting or permanent disturbances of essential ecological functions
No detrimental fluxes of chemicals	no detrimental fluxes of chemicals	detrimental fluxes of hazardous chemicals into neighboring ecosystems
Optimum conservation of species	almost complete conservation of species (HC_5)	species deletion (HC_{50})
Site-specific quasi-natural climax communities	$\longrightarrow$	possibilities of land use reduced
Any form of ecologically indicated land use possible	gradually increasing reduction of soil functions	biotic and abiotic soil functions reduced

sufficiently incorporated many of the dynamic properties found in other aquatic environments.

FREHE's *greenhouse approach* [319] is an example of microcosm research related to agriculture. Each microcosm, measuring $207 \times 193 \times 200\,cm^3$, was fixed on a lysimeter containing a soil mixture (loess loam, sand, and peat) with an earthworm (*Lumbricus terrestris*) population; an association of pea (*Pisum sativum*), vetch (*Vica sativa*), oat (*Avena sativa*), and mustard (*Sinapis alba*) was grown in it. The greenhouses were provided with ventilation and sprinkling systems permitting the application of di-(2-ethylhexyl) phthalate for testing purposes in aerosol form.

Studies in such small-scale laboratory systems are valuable tools to demonstrate some principles of concern in ecotoxicology and have to be included in the general knowledge of the fate of chemicals in natural large-scale ecosystems. They are indispensable additions to single-species experiments in order to more reliably evaluate the influence of interspecific interrelationships on the toxic potential of compounds. An example is zinc toxicity with respect to marine diatoms and dinoflagellates, which, in the majority of cases, proved to be definitely more pronounced in five-species tests than in monospecific cultures [320]. Hence, an abbreviated model ecosystem approach may prove useful in environmental concentration – fate studies by providing data on the rate constants of some of the processes that govern the transport and transformation of chemicals. Thus limited-scope model ecosystems contribute to determine whether a mathematical model accurately or approximately predicts concentrations in the various environmental compartments included in the model ecosystem. Expanded types of microcosms may be envisaged to include a substantially higher amount of that variety of physical, chemical, and biotic properties that a chemical is likely to encounter in the real environment. In terms of structural diversity, these comprehensive, and consequently more expensive, *mesocosms* lead over to biocenotic experiments under open-air conditions, which constitute the highest assessement level possible.

Biocenotic Experiments. Individual *populations* display marked changes in abundance as a direct consequence of the exposure pattern and an indirect consequence of the induced variations of intra- and interspecific competition and resource availability. Thus, almost all collembolan species exposed to cypermethrin displayed a longer increase in population density after an initial reduction, followed by a secondary decrease, probably due to excessive exploitation of food resources or an increase of predators [321]. In numerous experiments, complete annihilation of species was the exception; recedent taxa were mostly affected. Yet in a few cases, populations with dominant species were almost completely suppressed (e.g., the predatory mite *Rhodacarellus silesiacus* and the collembolan species *Isotomina bipunctata* on experimental plots treated with aldicarb [322]). Also, nontarget organisms experienced changes in their abundance spectra that lasted for several years [323].

At the *community level*, chemicals nearly always induce changes in the abundance spectra,

whereas changes in species composition occur less frequently, particularly as a result of persistent multifactorial stress [324], [325].

Ecosystem PNEC Assessment. Integral PNEC indicators at the *ecosystem level* are

1) Fluxes of macro- and micronutrients (K, Ca, Mg, N, P, S and Fe, Mn, Cu, Zn, Mo, B, Cl)
2) Duration of biochemical cycles
3) Energy flux rates and entropy production of phytocenoses
4) Changes in biodiversity and population dynamics
5) Structure of trophic networks or experimental food chains

Examples of the first groups of indicators are the influence of pentachlorophenol (PCP), 2,4-D, 2,4,5-T, and aldicarb on the nitrogen balance and litter degration of beech, pasture, and fallow ecosystems [326], [327]. In either case, degradation was impeded and the nitrogen content of the litter bags reduced. Shortly after application of the chemicals, the release of inorganic nitrogen compounds increased considerably at the expense of microflora and fauna. Furthermore, the retention capacity and the availability of many nutrients were diminished, and these effects lasted for more than four years depending on the chemical character and application rate of the compounds tested.

3.2.2.2. Complex Assessment and Modeling of Vegetation Damage

Pollutant gases such as sulfur dioxide and nitrogen oxides are frequently emitted in combination and normally with other pollutants such as soot and heavy-metal particles; in many places, photooxidants are emitted simultaneously. In these cases, the effects of combinations of pollutants are not simply the sum of the effects of individual components, but the relative concentrations of components or the specific sensitivity of the plants to one of them may determine the response. Thus, pollutants in combination may display antagonistic (reduced), additive (sum of individual pollutant effects), or synergistic (superadditive) effects.

In the past, most experiments on exposure to combinations of pollutants were carried out under laboratory conditions using higher concentrations of gases than are typical in the field. Newer fumigation techniques, for instance open-top chambers [328] or field fumigation systems [329], have now enabled the exposure of plants under field conditions with increasingly realistic concentration ranges. These experiments indicate that pollutant combination effects depend on pollutant concentrations; SO_2 and O_3 are often antagonistic at levels where the individual gases cause severe leaf injury, while superadditive effects predominate at lower concentrations (i.e., $SO_2 < 140\,\mu g/m^3$, $O_3 < 100\,\mu g/m^3$), particularly near the threshold dosage [330]. In addition, the injury symptoms for SO_2 and O_3 in combination are not a simple mixture; frequently, the symptoms of one pollutant occur more clearly or earlier than those of the other [331].

With regard to combinations of NO_2, SO_2, and O_3, effects were produced in many plants by lower concentrations of NO_2 and the other gases than if they acted alone, especially at the lower exposure levels [332], [333]. Synergistic effects of NO_2 in combination with SO_2 or O_3 apparently result from a decrease of the nitrate and nitrite reductase activities. While the latter tend to rise in response to NO_2 acting on its own, activation of these reducing enzymes is limited or even suppressed by the simultaneous presence of SO_2 or O_3 [334]. This adverse effect is further enhanced by the formation of free radicals. An inhibitory influence on the regulation of transpiration occurred at less than $20\,\mu L/m^3$ $SO_2 + 20\,\mu L/m^3$ NO_2 [335]. The combination of NO_2 with O_3 is likely to endanger vegetation more than $NO_2 + SO_2$ in combination and should receive greater attention since these pollutants form essential components of photochemical smog. The threat to vegetation may increase still further if all three pollutants interact simultaneously [332], [333], [336], [337].

Combination effects, however, also depend on the temporal order in which the pollutants affect the plants. If they are first exposed to NO_2 and thereafter to SO_2, the usual SO_2 effects are reduced, while the opposite sequence causes superadditive effects because SO_2 affects nitrite reductase [334], [338]. Attention must also be drawn to the combination of gaseous SO_2 with acid precipitation; the effects of this combina-

tion are discussed in connection with the new types of forest damage.

Critical Levels. In conclusion, and with some reservations resulting from the above remarks, critical levels for short-term and long-term exposures may be defined. In view of the fact that most horticultural and agricultural plants do not show adverse effects at *sulfur dioxide concentrations* $<30\,\mu g/m^3$, while particularly sensitive species of trees, moss, lichens, and bushy and grassland vegetation are adversely influenced by concentrations of ca. $20\,\mu g/m^3$, the critical level for SO_2 acting on its own is to be set at $20\,\mu g/m^3$ ($\approx 0.007\,mL/m^3$) as the annual mean value. Experimental fumigation trials that indicate first adverse effects from $70\,\mu g/m^3$ SO_2 (≈ 0.025 ppm) for cultivated plants form the basis for setting the short-term value at this concentration [331].

Leaf necroses were the predominant effects of short-term, mainly single exposures of plants to $>1\,mL/m^3$ *nitrogen dioxide*; lower concentrations caused alterations in photosynthesis and respiration as well as in organelle ultrastructures. The particular sensitivity of plants at night must also be taken into account, and the fact that high NO_2 concentrations are often associated with high NO levels. For long-term exposures, a seasonal differentiation of sensitivity is indicated. Thus, for summer, when the physiologically more active plants can rapidly metabolize NO_2, a half-year mean of $0.03\,mL/m^3$ appears tolerable. In winter, however, when plants are particularly sensitive, a half-year mean of $0.02\,mL/m^3$ must not be exceeded to protect both short-lived plants and long-lived arboreal flora.

Experimental investigations of *ozone damage* [337] permit definition of the following critical levels for the protection of vegetation against O_3 as a single pollutant. For longer-lasting exposures, $50\,\mu g/m^3$ ($\approx 0.025\,mL/m^3$) O_3 appears indicated as a vegetation period average (composed of 7-h daylight mean values).

For *NO_2 in combination with O_3 and SO_2*, the World Health Organization has adopted an Air Quality Guideline of $30\,\mu g/m^3$ NO_2 as the arithmetic annual mean for areas in which SO_2 and O_3 may occur in annual and vegetation period means of up to $30\,\mu g/m^3$ SO_2 and $60\,\mu g/m^3$ O_3 [333]. However, nitrogen oxides do not only have direct effects, they may also affect plants indirectly via the soil, with additional nitrogen input greatly endangering nutrient-poor plant formations such as heath and moor communities. An amount of $10\,\mu g$ NO_2 per cubic meter may suffice to damage the particularly sensitive ombrotrophic mires [339].

On the basis of exposure – effect relationships as determined for a variety of plants of widely differing sensitivity by various authors, Posthumus [340] deduced the following critical levels for *ammonia*:

$100\,\mu g/m^3$ ($\approx 0.14\,mL/m^3$) for monthly mean concentration
$600\,\mu g/m^3$ ($\approx 0.86\,mL/m^3$) for 24-h mean concentration
$10\,mg/m^3$ ($\approx 14.3\,mL/m^3$) for 1-h mean concentration

Impacts on Forest Productivity. Since the 1970s the indirect effects of acid precipitation on forest trees and crops has been a major area of both scientific and economic concern. In Germany, forest decline attributed to flue gases, particularly SO_2, was described in the neighborhood of industrial plants more than a century ago. The situation did not change essentially until the late 1960s, when a new type of forest injury became evident that was no longer related to local emitters but occurred to an increasing extent in presumably clean air areas far away from conurbation and industrial centers [341–345]. In the mid-1970s, severe dieback of silver fir (*Abies alba*) was observed in southern Germany. In the early 1980s, other conifers, especially Scotch pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*) showed similar symptoms over rapidly increasing areas, and deciduous species such as beech (*Fagus sylvatica*) and oak (*Quercus robur*) have recently been affected. According to the 1985 damage inventory, 52 % (i.e., 3.8×10^6 ha) of the forest areas in the 58 growth districts of the Federal Republic of Germany were injured. Level 1 of damage (reduction of vitality) characterizes about two-thirds of this portion, while medium to high damage (levels 2 – 4) occurs on 19 % of the forested area [346].

Comprehensive studies [347–350] show that the lack of nutrients, climatic factors [351], mistakes in forest management, infections, and pests

may play an important role in the present complex forest diseases; but a rapidly growing number of findings indicate that the above anthropogenic air pollutants, and possibly others such as triethyl lead or nitrophenols [352], [353], and related soil acidification are the essential direct and indirect causes of damage [328], [354–357].

At the cellular level, pollutants may lead to a disturbance of photosynthesis and respiration and the following primary and secondary metabolisms. In some cases, impairment of photosynthesis occurs before visible damage can be observed [358]. At the same time, photooxidants and acid rain cause weathering of cuticular structures, destruction of membrane systems, and leaching of nutrients from leaves. Quantitative and qualitative changes in photosynthesis [359] lead to reduced primary production, increased respiration, and a changed assimilate allocation to the root – mycorrhiza system, stem, branches and leaves, and also to the reproduction and phytogenous defense system. Such stressed trees emit unsaturated hydrocarbons such as ethylene, isoprene, and terpenes, which react with secondary pollutants (e.g., ozone) and highly phytotoxic tertiary pollutants in the immediate neighborhood of the plants. Weakening by reduced assimilate allocation to the roots leads to reduced uptake of nutrients from soil that has already been disturbed by acidification (i.e., impoverished in potassium and magnesium).

Murach [360] determined a calcium deficiency in the roots of declining spruce and fir, but the same levels of aluminum as in healthy trees; Rehfuess et al. [361] noted magnesium and possible calcium deficiencies by foliar analyses even on base-rich soils. Contrary to Ulrich's [362], [363] contention that a change in fine root biomass is paralleled by toxically high Al : Ca ratios, Rehfuess points out that the two phenomena are not necessarily synchronized, so that marked decreases in fine root biomass can precede the increase in soil solution aluminum.

Mineral deficiency due to leaching and soil acidification is frequently coupled with a relative surplus of nitrogen with respect to assimilated carbon as a consequence of enhanced deposition of nitrate and ammonia and the intensity of subsequent nitrification processes. Kriebitzsch [364], who studied nitrification in many types of acidic forest soils, divided them into groups A, B, C, and D. In group A, no nitrification occurred and ammonium was the only nitrogen source. In soil of groups B and C, partial nitrification occurred, while group D soils had complete nitrification. Former heathland soils in Germany now cultivated with various *Pinus* species and *Pseudotsuga* belong mainly to group A, where the nitrate level is low, whereas the ammonium concentration is high.

Quantitative and qualitative changes in the living biomass of forest trees will bring about changes in the food web of the forest ecosystem; the population of some consumer (e.g., insects) or decomposer organisms (e.g., fungi) may grow explosively while others are completely suppressed. The disturbance of soil biocenoses may then induce a reduced litter decomposition that leads to a further shortage of minimal nutrients. Trees subject to these stress factors are in general extremely sensitive to secondary damage by frost, drought, or insect infestation.

Visible *injury to crops* as a result of ambient acidic precipitation has not been reported from modern high-productivity agrarian ecosystems. Therefore, attempts to evaluate impacts of acidic deposition must be based on simulated precipitation in field experiments or controlled environments. Yet most of the crop cultivars studied in the field and the crop varieties studied in controlled environments exhibited no effect on growth or yield as a result of exposure to simulated rain usually up to ten times more acidic than ambient [365]. Likewise, out of five cultivars of soybeans studied, only one exhibited reduction in yield [366]. Genetic variations, possibly in combination with characteristics of the soil or simulated rain applied in these experiments, may account for the observed variability.

3.2.2.3. Statistical Analysis of Forest Dieback

Descriptive or static models of forest decline are analyses of the spatial correlation pattern of site qualities and damage levels [348], [367] that corroborate the importance of pedogenic nutrient supply and acidification processes for forest decline. Figure 41 illustrates, on the basis of comprehensive cross-tabulations of several hundred German stands, the extent of forest damage in terms of relevant site factors. The standardized damage level S (i.e., proportion of

injured : unaffected forest area) correlates best with the pedogenic nutrient supply, frequency of fog situations, and elevation, while the correlation with other site factors is also significant on the 99.9 % level although of little statistical relevance; $S = f(Sq)$ is a monotonous function, while $D = f(Nf)$ has a marked maximum. Thus, stands on members of the ranker, cambisol, and podzol soil groups with pH <4, marked nutrient deficiency, and concomitant reduction of buffering capacity down to the aluminum and iron ranges are most liable to dieback. Among airborne pollutants, SO_2 is particularly important. In combination with other pollutants, its effect on both vegetation and soil is enhanced by fog (contingency coefficient after Pearson $C = 0.53$). The amount of sulfur deposition is controlled mainly by precipitation ($C = 0.80$). Comparison of the pH and sulfur deposition curves ($C = 0.89$) summarizes the major influence of gaseous and particulate pollutants on forests as a result of the above mechanisms. Their inherent complexity implies the recognition that critical levels for tolerable pollutant concentrations can be defined adequately only with regard to the entire set of concomitant stress factors affecting a plant community.

A more sophisticated analysis of the data of forest decline monitoring programs can be brought about by the CHAID approach (see Section 2.1.1.1) [103], [368]. CHAID divides the total data set successively into subsets that are, at each level of partitioning, characterized by the very variable that best explains the distribution and intensity of damage. Thus, the explicative value of different combinations of decline-inducing factors can be tested on high levels of statistical significance. A further advantage is the easy interpretability of the CHAID dendrograms.

Thus, a dendrogram of forest decline of about 500 stands in North Rhine-Westphalia clearly indicates the predominant role of the factors age, crown density, and composition of stand as boundary conditions of forest decline, while climatic factors such as annual precipitation and anthropogenic pollutant gases play a distinctly lesser role. These results are corroborated by comparable findings in other parts of Germany [103].

Dynamic Models of Forest Dieback. Simulation models of forest decline have to deal with two types of change – (1) change in complexity through increasing or decreasing differentiation, and (2) adjustment to extrasystemic changes. Negative and positive feedback processes can be correlated with systemic continuity or discontinuity, respectively. Continuity results when deviation-correcting mechanisms operate to ensure that structure and function are kept viable within the given parameters of the system. Quantization occurs when deviation is amplified to the point at which no deviation-correcting mechanism can prevent rupturing of the basic systemic framework (i.e., when resilience mechanisms are no longer able to contain and canalize the energies and thrust that have been generated). The process alters the relationships of the system to its environment, creating new spatiotemporal, structural, and functional boundaries – in short, the system is transformed to a new level of internal organization and environmental integration. Therefore, and specifically with regard to forest decline, models must employ system concepts to explain the circumstances in which quantization can occur from one developmental stage to another. Hence, they have to account for both systematic levels of biocenotic organization and cybernetic processes that demonstrate (1) systemic self-stabilization within a given organizational level and (2) systemic transformation resulting in a biocenotic quantum jump across an environmental frontier.

Models with Nonlinear Dose – Effect Relationships. Following models of reduced photosynthesis and leaf formation, and increased leaf aging and shedding under air pollution stress [369], [370], Kohlmaier and Plöchl [371] developed a more comprehensive model concept for damage of Norwegian spruce. It includes the root system with reduced nutrient uptake and higher feeder root turnover, thus coupling the canopy and the soil root system. Reduced photosynthesis and changed assimilate allocation are simulated during the entire exposition to airborne pollutants, which may be a part of or the whole life span of a tree. Assimilate demand and supply have a hierachic structure with regard to, e.g., respiration, growth of foliage, roots, and living wood. The effect of an air pollutant on leaves and roots depends on both the

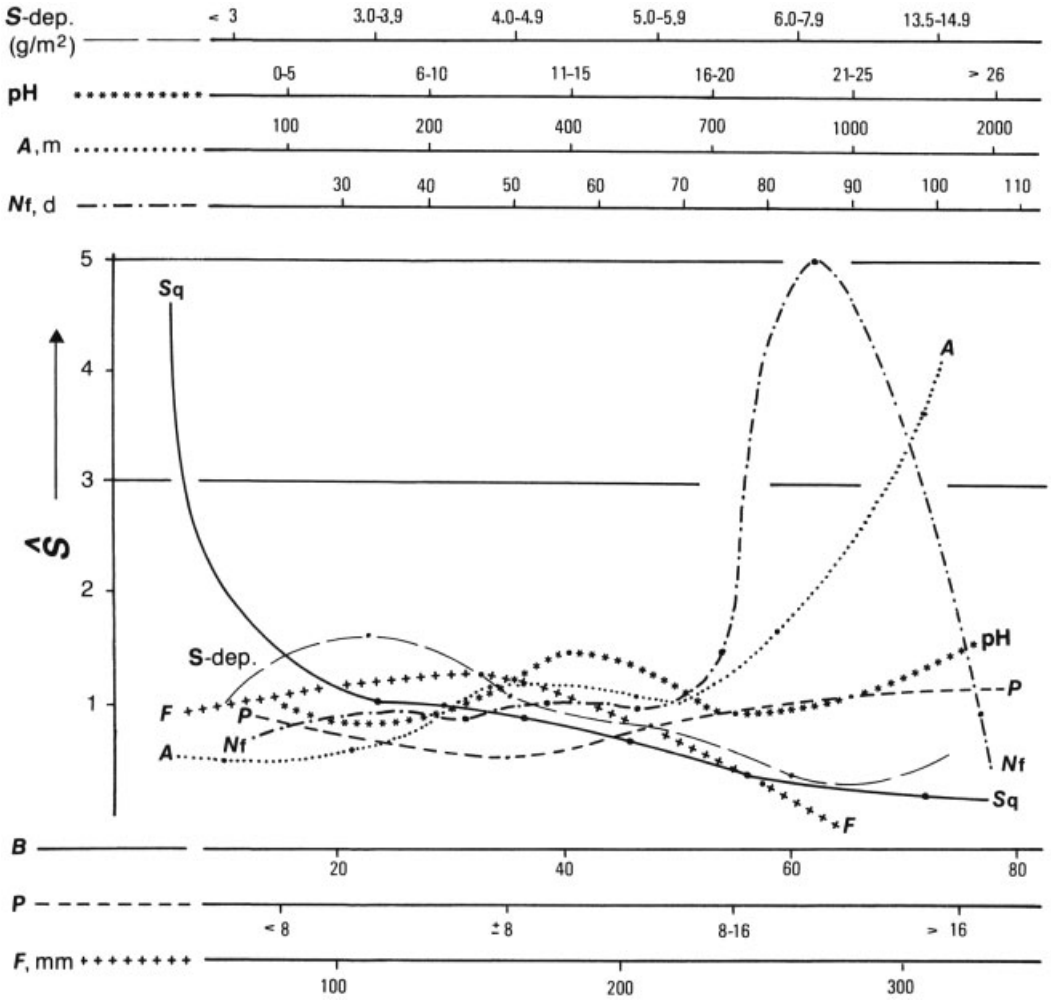

Figure 41. Forest decline as related to selected site factors
$\hat{S}$ = Ratio of damaged : undamaged forest area; A = Altitude (m); Nf = Number of foggy days; Sq = Soil qualitiy (nutrient supply); P = Porosity (percentage of pores >50 µm); F = Plant available field capacity (mm); pH = Percentage frequency of acid rain (i.e., pH <4.5); S-dep. = Sulfur deposition

specificity of instantaneous effects and the cumulative effects during the entire exposure time involving elimination and repair mechanisms of the plant. The complicated interaction pattern of these response mechanisms normally results in nonlinear dose – effect relationships with respect to concentration and accumulated pollutant and damage level of the plant. During one time step, the model first calculates the photosynthesis rate as dependent on external factors such as light intensity and air temperature, and internal parameters such as photosynthetic leaf efficiency; water uptake by roots is the limiting factor. Pollutants can enter the system by two different paths (i.e., leaves or needles and roots); in the first case, the photosynthetic efficiency of a particular needle age class is reduced; in the second, fine root turnover and consequently assimilate demand are increased. The momentary pollutant flux into the leaf is described in terms of gas exchange between the intercellular air and the ambient air by Fick's first law:

$$I_{\mathrm{in}}^{\mathrm{leaf}} = (c_{\mathrm{a}} - c_{\mathrm{i}}) / r_{\mathrm{ai}} \tag{67}$$

where the flux $I_{\mathrm{in}}^{\mathrm{leaf}}$ is proportional to the concentration difference $(c_{\mathrm{a}} - c_{\mathrm{i}})$ between ambient

and intercellular air, and r_{ai} is a specific resistance factor for a particular pollutant. The corresponding flux per unit canopy or unit ground area is then obtained by multiplication with the leaf area index LAI. For the boundary condition $c_{\mathrm{a}} \gg c_{\mathrm{i}}$ Equation (67) reduces to

$$I_{\mathrm{in}} = (r_{\mathrm{ai}}/LAI)\, c_{\mathrm{a}} = u_{\mathrm{dep}} c_{\mathrm{a}} \tag{68}$$

which is approximately equivalent to the dry deposition rate, with u_{dep} being the deposition velocity.

The pollutant load in the leaf or at its surface may be assumed to be eliminated either by physical processes (e.g., litterfall), by transport to other parts of the plant, or by metabolic processes. It appears more appropriate, however, to model the elimination process according to Michaelis – Menten kinetics of the type

$$d\bar{c}_{\mathrm{i}}/\mathrm{d}t = u_{\mathrm{dep}} c_{\mathrm{a}} - k_{\mathrm{out}}\, \bar{c}_{\mathrm{i}} / \left[1 + (\bar{c}_{\mathrm{i}}/K_{\mathrm{c}})\right] \tag{69}$$

where $\bar{c}_{\mathrm{i}}$ is the concentration per unit ground area and k_{out} the characteristic elimination constant, while the other symbols have the above meaning. The mode of interaction and the effect of a pollutant inside the plant are likely to be specific for the particular pollutant j. In the simplest case, the effect, above the specific no-effect level, will be proportional to the interior concentration $\bar{c}_i^{\mathrm{j}}$:

$$eff^{j}(t) = a_j\, \bar{c}_i^{j}(t) \tag{70}$$

where a_j is a specific effect constant relating to a particular process (e.g., photosynthesis). The effect variable *eff* (t) may be composed of linear and higher-order terms of n interacting species:

$$eff(t) = \sum_{j=1}^{n} a_j \bar{c}_i^{j}(t) + \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} \bar{c}_i^{j} \bar{c}_i^{k} \tag{71}$$

where synergetic effects are included in the second term with $j \neq k$.

According to Bossel et al. [370], the effect variable affects the leaf senescence function

$$ALT(t) = 0.1\,[1 + eff(t)] \tag{72}$$

with $0 \leq eff \leq 1.5$ and $0.1 \leq ALT \leq 0.25$, implying that the value 0.1 is the natural senescence rate without pollutant damage. The senescence rate influences the number of needle age classes:

$$NJ = 0.8/ALT(t) \tag{73}$$

and both determine the leaf's photosynthetic efficiency. Hence, the photosynthetic efficiency of a particular needle age class is

$$E(N) = 1 - ALT(t)\,[(N-1) + TJ] \tag{74}$$

where $0 < TJ < 1$ characterizes the season and $1 < N < J$ the age of a needle class.

The average photosynthetic efficiency is then expressed by

$$EFF = \sum E(N)\, L(N)/LEAF \tag{75}$$

where $L(N)$ is the leaf mass of a particular needle age class and $LEAF$ the total leaf mass. In unaffected systems, EFF ranges between 0.55 and 0.65 [371].

Acidification of soil and related leaching of mono- and bivalent cationic nutrients leads to higher feeder root turnover rates while a gradual loss of feeder root mass occurs. Following a titration curve of a buffered chemical system, this effect may be described by a logistic curve of the type

$$W_{\mathrm{to}}(t) = f[W(t)] = W_{\max} / \left[1 + F_k \left(3^{-Klt}\right)\right] \tag{76}$$

where $W(t)$ is the accumulated pollutant at time t; $W_{\max}$ the maximum accumulation; F_k the correction factor $(W_{\max} - 1)$; K the buffer constant; l the constant input; and $W_{\mathrm{to}}(t)$ the fine root turnover factor at time t [370].

This supply – demand model of photosynthetic efficiency and assimilate allocation appropriately describes the decline of a forest ecosystem under the influence of chronic air pollution stress. It clearly indicates that nonlinear effects may drastically accelerate the observed dieback phenomena, but more experimental data on both the biochemical and the physiological levels are required to better account for the combined effect of two or more pollutants [55], [372–374].

Time Series Analysis and Mapping. A common feature of the various methods of time series analysis of dynamic systems such as ecosystems is that they involve the observation of a complete set of state variables. This may constitute a major drawback in studies of empirical systems whose inherent dynamics are not known a priori. Under these circumstances, however, use can be made of the fact that in

dynamic systems with a finite number of state variables, information on the momentary value of all state variables can be replaced by information on the recent history of a part of the variables. GROSSMANN et al. [375] therefore used time series analysis in combination with geographic information systems (GISs) to develop a scenario that describes forest damage in sequential form by means of maps. A geographic information system is a computer hardware and software system designed to collect, manage, analyze, and display spatially referenced data. Geographic information systems are emerging as the major spatial data-handling tools for solving complex natural resource planning problems, and the use of GIS technology has revolutionary implications for conducting research and presenting research results.

The GROSSMANN et al. approach POLLAPSE is a combination of dynamic models writing information into the data bank of a GIS, which thus produces a time series of maps of forest growth or damage [375]. These maps can be compared with the actual development so that deviations become readily discernible. The hypotheses underlying this type of regionalized modeling are

1) Photochemical oxidants cause foliar damage, and subsequent attack of airborne pollutants leads to leaching, eventually resulting in nutrient deficiency [376]
2) Ulrich's soil acidification hypothesis [362]
3) A combination of the two preceding assumptions

The predictive quality of these hypotheses was tested in the Pfaffenhofen area (Bavaria) on the basis of 462 forest stands, and the POLLAPSE methodology showed that the combined hypothesis (3) yielded the relatively best interpretation of forest damage phenomena observed in terms of both spatial pattern and degree of damage. The above methodology of time series analysis and mapping also offers excellent possibilities for the management of forests or other biotic systems as a base for sustained viability.

3.2.3. Impact of Xenobiotics on Aquatic Ecosystems

Xenobiotics released to aquatic environments have a great impact on biocenoses inhabiting streams, wetlands, and lakes. This adversely affects not only the respective aquatic ecosystems, but also humans who are in contact with this water. Aquatic organisms show how far-reaching consequences can be caused by effects that are difficult to detect on a short-term scale and therefore easily overlooked. Similar to effects on aquatic biocenoses, in most instances the effects of xenobiotics on humans cannot be recognized, not only because the necessary analyses cannot be made, but also because small, difficultly detectable effects can grow into major and irreparable population damage. The following describes how difficult it is to foresee the effect of a chemical on an aquatic ecosystem merely from (easy to make) short-term tests. This is a warning that care for human health and life means that society cannot wait until some clearly seen negative effect is proved beyond any doubt but must react on early warnings obtained by sensitive methods. For more detailed information, see [56–64].

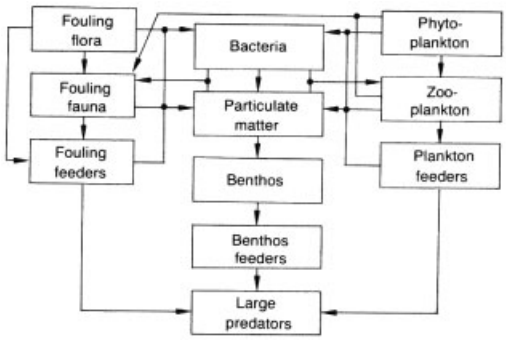

Figure 42. Food chains or food webs as the organization of organisms within an ecosystem according to food selection; arrows indicate directions of nutrient flow

3.2.3.1. Approaches and Methods of Studying the Impact of Xenobiotics on Aquatic Life

Aquatic life is organized into several hierarchic levels (Fig. 9 in Section 2.1.4):

1) Individual organisms
2) Monospecific populations
3) Communities of different species inhabiting the same environment, e.g., the pelagic communities freely swimming in water and benthic communities inhabiting the bottom of rivers and standing waters (ponds, lakes, reservoirs)

4) Ecosystems with their nonliving components interacting with organisms; within an ecosystem, organisms are organized into food chains (or more exactly, food webs) according to their selection of food items (Fig. 42)

Detection of the effects of xenobiotics is performed at all hierarchic levels of biological organization. In addition to this organizational hierarchy, a hierarchy of time scales of the effects can be distinguished, from very short ones measured in minutes up to very long ones measured in human generations [377], [378].

Methodology. The most common approaches to detect the impact of xenobiotics on aquatic life are listed below:

1) Classical toxicity tests on test individuals or cultivated populations of aquatic organisms to detect short-term lethal effects
2) Methods for detecting longer-term sublethal effects on test individuals and populations
3) Methods for detecting genotoxicity
4) Use of biomarkers, a relatively new approach for detecting stress effects of long-term exposure of individual organisms living in contaminated environments
5) Microcosm and mesocosm studies
6) Monitoring of long-term consequences of contaminants in nature by following the species diversity and loss of sensitive species in biocenoses and ecosystems

Other approaches are, e.g., predictions of toxicity based on molecular structure of the respective substances using quantitative structure – activity relationships (QSARs, see, e.g., [10], [220]), and tests for changes of morphologic or morphometric characteristics of organisms and detection of pathological individuals.

Classical Short-Term Toxicity Bioassays. Water organisms employed for these tests are usually some species of easily cultivated algae, water flies, and fish (see Section 3.2.1). Benthic, bottom-dwelling organisms commonly used are worms of the family Tubificidae living in the bottom mud of streams and lakes that are fairly polluted with organic matter of natural origin. In the simplest version, the tests consist of adding different concentrations of xenobiotics to jars (aquaria) containing the respective organisms and determining the time until half of the population dies (LC_{50} or LD_{50}). Recently, much care has been taken to standardize aquatic toxicity tests to maximize their major advantage – the rapid comparative evaluation of many xenobiotics [65].

Long-Term Sublethal Effects of Xenobiotics. In the aquatic environment, long-term sublethal effects are studied by measuring different reactions of test organisms. An example is the behavioral test with the freshwater bivalve mollusk *Dreissensia polymorpha* [379]. Its filtering rate, which is determined anemometrically in this test, represents a vital function, providing food and therefore vitality, growth, and reproduction of the organism. Detection limits of long-term sublethal tests are several orders of magnitude higher than those of short-term toxicity tests.

Genotoxicity. Compounds such as diethylstilbestrol that disrupt the endocrine systems of fish, wildlife, and humans were found to induce large-scale dysfunction at the human population level [66] by causing genetic and reproductive changes. Genotoxicity is studied by observations of the offspring of the affected population.

Determination of Biomarkers. Measurements of body fluids, cells, or tissues are carried out that indicate in biochemical or cellular terms the presence of contaminants or the magnitude of the host response [63]. Examples are, e.g., determination of marker enzyme activity in cells of target tissues, serum enzyme activity, appearance or loss of reaction products of an enzyme; major changes in the concentration of a cellular component such as glycogen, fat, mucus, or metallothionein; and measurements of damage to DNA.

Microcosm and Mesocosm Studies (see Section 3.2.2.1). Microcosms are laboratory model ecosystems consisting of mixed natural populations of organisms transferred from nature to a laboratory jar or aquarium, sometimes an arrangement that imitates a stream. Mesocosms are of the same origin, but of much larger volume and usually kept outdoors in their natural environment (e.g., [59], [380–382]). When xenobiotics are added, the changes in some common properties are measured (e.g., species composition, diversity, photosynthesis, or respiration).

Studies of Changes in Structure and Function of Communities and Ecosystems. Toxic chemicals can affect independently both the structure and the function of aquatic ecosystems. Low levels of perturbation may cause changes in community composition or structure without a change in function. Alternatively, a process, the rate of which is changed by stress (such as respiration or photosynthesis), may be reduced under sublethal stress, but the community structure is left intact. According to [383], the most predictable changes seem to be the disappearance of sensitive species from communities under environmental stress.

The integrated bioindicator approach currently gives the best results. It involves monitoring a suite of selected exposure and effects indicators at several levels of biological organization from the biomolecular to the community level [63], [384].

Critical Evaluation of the Above Approaches to Aquatic Ecosystems. For aquatic environments, *short-term lethal toxicity tests* have many advantages, as well as many disadvantages. Their major advantage is the ability to rapidly compare the relative toxicity of specific chemicals or specific mixtures. The limitations become particularly obvious when trying to use their results for higher levels of biological organization in the aquatic ecosystem. Thus, they do not account for the effects of

1) Chemical speciation in the environment
2) Kinetics and hysteresis in sorption of chemicals to sediments
3) Bioaccumulation through the food chain from the lowest level of algae and plants to the highest level of fish
4) Modes of toxic action on survival, growth, and reproduction of individual species populations

For aquatic sediments, an additional inadequacy is that sediment toxicity tests often rely on equilibration of chemicals in a sediment – water mixture or employ benthic organisms whose survival, growth, or reproductive success may depend not only on the presence of the toxicant but also on physical characteristics of the sediment.

The application of results of (laboratory) toxicity tests to natural populations is especially risky for the following reasons:

1) The way in which laboratory organisms are exposed to pollutants differs from exposure in the environment
2) Laboratory tests deal with single chemicals, and organisms are exposed to complex mixtures in the environment
3) Criteria for effects in the laboratory are not important functional endpoints in population and systems dynamics

The inadequacy of short-term tests for the aquatic environment may be shown by conclusions drawn from careful monthly bioassay results with the survival and reproduction of a water fly (*Ceriodaphnia*) and larvae of fathead minnow. These indicated that the effluent of an energy plant at Oak Ridge (Tennessee) causes no observed effect. However, community surveys demonstrated clear effects on diversity and species richness downstream of the power plant [63].

The use of *biomarkers* is more relevant to conditions in nature than acute toxicity methods. Biomarkers provide evidence of exposure to rapidly metabolized contaminants and in this way integrate pharmacodynamic and toxicological interactions. Correlations between biomarkers of exposure and ecological effects are expected to be better than those between populationmonitoring parameters and indicators of exposure.

Observations in nature can be performed only in rare instances when the concentration of the xenobiotic in nature is increased for a prolonged period. These observations cannot be used for testing new materials. Therefore an integrated system of xenobiotic hazard evaluation has been developed (see Sections 3.2.1, 3.2.3.4).

3.2.3.2. Behavior of Xenobiotics in the Aquatic Environment and their Influence on Aquatic Organisms and Indirectly on Humans

A chemical of potentially xenobiotic character follows a complicated flow from production site to the environment, as shown in Figure 43. The situation in fresh water depends not only on the direct emission of the respective chemical to the water body but also on its emissions and fates

in soil and groundwater (both associated with vegetation growth and decay), on the chemical and biological processes in the water body, and on the bioaccumulation in different organisms of a food chain. Environmental processes connected with local hydrometeorology, with precipitation and flow, and with temperature and light conditions modify the fate of xenobiotics. Figure 44 illustrates the behavior of hydrophobic xenobiotics in standing waters, particularly the exchange between the pelagic and benthic environment (i.e., between water and sediments). The participating major processes are shown: sedimentation, resuspension, and mud – water exchange. For hydrophilic xenobiotics, another important process is bioaccumulation (i.e., accumulation inside the organism's tissues and organs, mainly in fat and liver). In addition to these processes of an individual chemical compound, complex interactions occur between different chemical species in nature.

Many organic contaminants are readily taken up and *bioaccumulated* by aquatic organisms [61], [68] if the chemicals are dissolved in water. The bioavailability is much lower if the same compounds are sorbed to soil or sediment particles. Many toxic chemicals do not bioaccumulate but rather are metabolized. In many cases the metabolites are more toxic than the parent compounds present in the environment. Bioaccumulation, the process by which a substance is taken up by an aquatic organism from water and through food, is found to depend strongly on the size (weight) of the respective organisms (see, e.g., for pesticides [56], [387]). This is due to the dependence of most physiological processes on size (e.g., [59]). Biomagnification within aquatic food chains (i.e., the increasing concentration of toxic substances in higher organisms) was found important only for some chemicals (e.g., DDT [388]).

Humans are affected by the processes in water in different ways, as a result of their direct and indirect contact with water, but also through air, soil, and food. Large differences exist in the biological availability of contaminants associated with different environmental media, as well as individual and species-specific differences in pharmacodynamic disposition.

In most instances, one thinks of direct effects of the xenobiotic on organisms. However, recent ecological findings demonstrate that in nature the indirect effects are more important [389]. They are caused, e.g., by changes in food composition when some organisms become rare or disappear because of toxic effects and by changes in competition and other possible relations between organisms. An easily understandable example is the eradication of an aquatic weed species. Indirectly, this will cause a complete change of abiotic conditions not only for other species that depend on the weed directly, such as invertebrates living on it, but also for fish species finding their shelter.

Adaptation. Organisms are capable of adapting to the effects of xenobiotics to a certain degree, and this can have both positive and negative effects on individual organisms (including humans) and on the total environment, as well as on environmental management. Adaptive acquisition of degradative abilities for some organic compounds or of resistance to heavy metals has been demonstrated for bacteria in laboratory ecosystems [390]. Such an adaptive response may be caused by different molecular and biochemical processes: (1) induction of specific enzymes in members of the community, resulting in an increase in the observed degradative capacity of the total community; (2) growth of a specific subpopulation of a microbial community able to take up and metabolize the substrate; and (3) selection of mutants that acquire altered enzymatic specificity or novel metabolic activities at the onset of exposure of the community to the introduced compounds.

The *negative effects of adaptation* may be exemplified by pests (e.g., insects or weeds) in the framework of eradication programs [67]. In West Africa, the widespread parasitic disease filariosis is caused by the *Filaria.* This parasite is transmitted to humans by blackflies. To eradicate the flies, the organophosphate Temephos was applied on a large scale in about 18 000 km of rivers in the region during 1974 – 1980. However, resistance to this xenobiotic developed soon in one region and spread rapidly. Therefore, other means of eradication were sought that would act more persistently and not show the negative effects of the insecticide on other organisms in the rivers, since Temephos causes up to 100 % mortality of many insect species living on the river bottom. No toxicity to fish was observed. However, the insecticide accumulated in fish tissues,

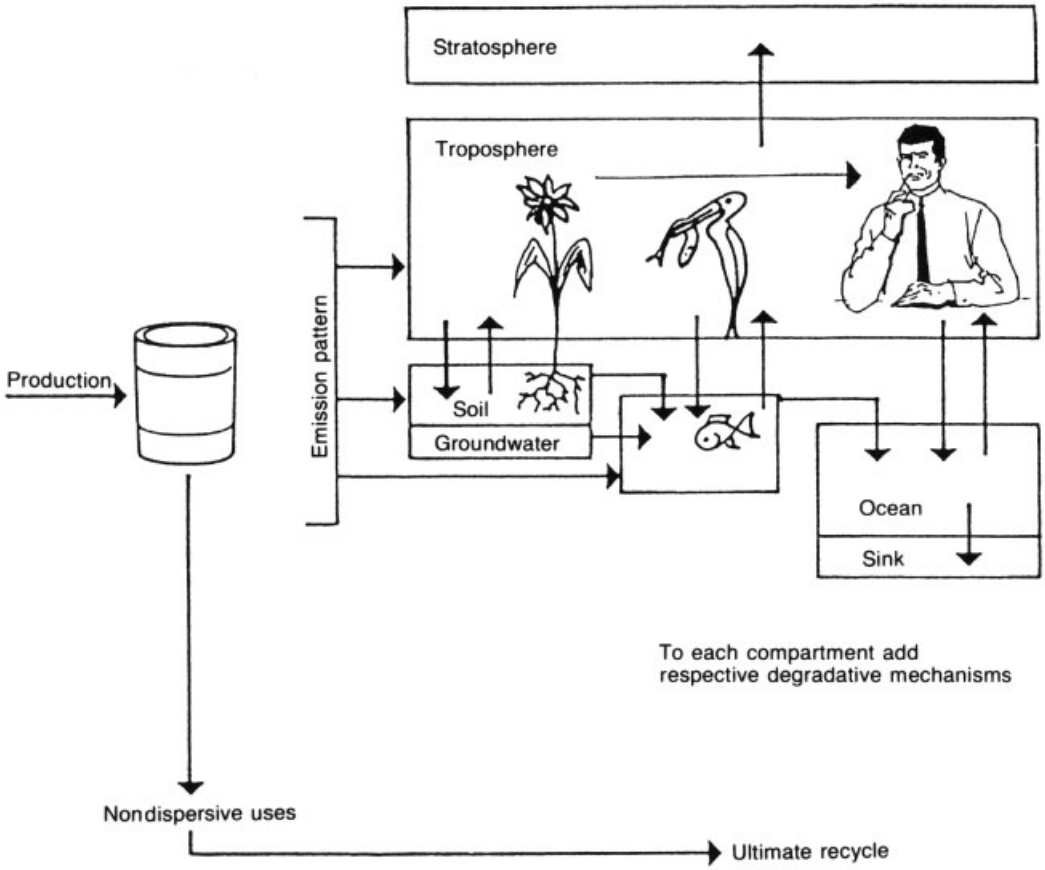

Figure 43. Schematic representation of the flow of a chemical from production to the environment (after [385])

which may have negative effects on humans that consume the fish.

In addition to the use of another more toxic organophosphate (chlorophoxime), a biological method of eradication was developed. This method demonstrated the general advantage of using biological agents. The bacterium *Bacillus thuringiensis* appeared to be very selective in killing only *Simulium* and few other species, and no adaptation of the fly populations was observed. Therefore, the application of a biological agent has the following advantages: (1) it is selective for the target goal; (2) no negative side effects are observed; (3) no rapid adaptation of pests takes place; and (4) the cost of application is lower.

Because of adaptation and other differences, the applicability of basic ecotoxicological principles derived in the temperate region to warm, arid, and cold climates is not without problems [67]. No simple straightforward relation exists for assessing regional differences of toxic effects due to differences in temperature and associated climatic factors. The general belief that metabolic rates are much higher in tropics has not been supported; nevertheless, an increased risk of toxic effects in tropics is to be expected.

3.2.3.3. Impact of Xenobiotics on Pelagic and Benthic Biocenoses

The data bank of the International Register of Potentially Toxic Chemicals (IRPTC) of UNESCO contains a record of some 8000 chemicals, and the number is growing rapidly. Therefore, only a few illustrative examples are described in the following, with an emphasis on related problems. The quantitative aspects of the

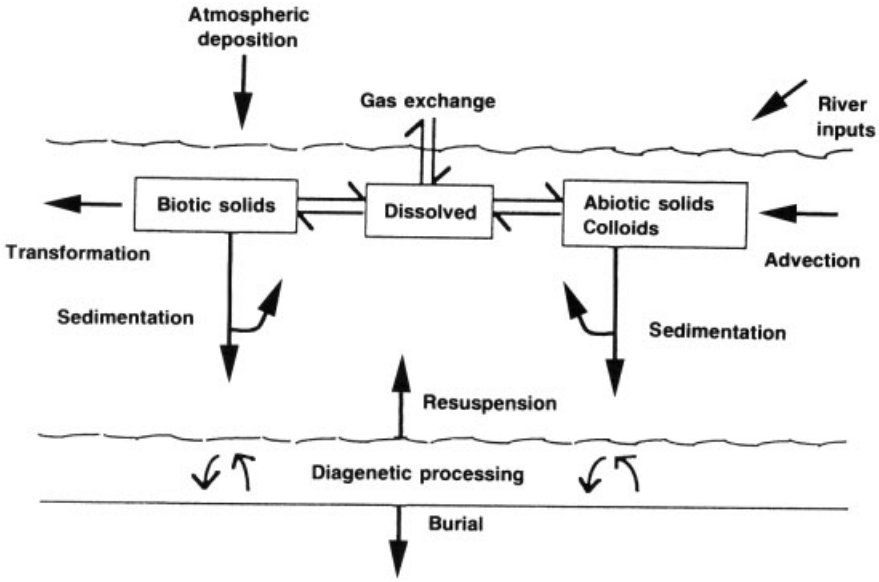

Figure 44. Behavior of hydrophobic xenobiotics in standing waters (after [386])

processes and predictions of the impacts of xenobiotics are described in Section 3.3, dealing with ecotoxicological modeling.

An example of the dependence of toxicity not just on one element, but on the ratio of several, can be demonstrated by the case of arsenic. REUTHER [391] found for Swedish lakes that, when added as arsenate, arsenic was readily accumulated in plankton at concentrations close to those found in nature. If the ratio P : As was low, arsenic proved toxic at much lower concentrations.

For the trace metals cadmium, lead, copper, and zinc, the effect of external variables on the amounts present in the environment was specific for each element [392]; especially important for Cd and Pb were sediment- and water-related abiotic variables. For the essential trace metals, the variance expressed by the effect of the common physical variables of the surrounding environment was very low: from 25 % for Zn up to 37 % for Cd and Cu. The quantities of the common xenobiotic cadmium in the aquatic environment can be demonstrated by the example of the river Ruhr in Germany (Fig. 45). A long-term effect of Cd on aquatic biocenoses was measured by using the microcosm method [380]. Figure 46 shows a steady decrease of the number of species in the microcosm with increasing cadmium concentration.

Some detergents greatly affect nature not only because of the amount of phosphates they contain (leading to eutrophication of water bodies), but also because of their toxicity. The detergent FIT was found to be highly toxic to frogs and their larvae [394].

Evidence exists of the sometimes close relation between the occurrence of some xenobiotics and the amounts produced in the respective region. Figure 47 shows how the quantities of PCBs, DDT, mirex, and chlorobenzene accumulated per year in layers of some U.S. lake sediments follow the sales or production rates of the respective years.

A systematic study of 50 chemical compounds demonstrated that their joint toxic effects can be considered simply as additive [395]. The authors proposed a quantitative structure–activity relationship (QSAR) technique for predicting the concentrations of the components of mixtures that would cause 50 % inhibition of the toxic effect by joint action. How far the effect of other compounds can act synergistically is not clear.

Range of cadmium concentrations of different waters in the Ruhr area

	Concentration, μg/L
Groundwater	0.05 – 0.15
Drinking water	0.05 – 0.2
Falling rain at Essen	0.2 – 1.1
Storm water flow from paved areas	1–10
Seepage water from sewage sludge deposits	<1
Seepage water from industrial sludge deposits	5–50

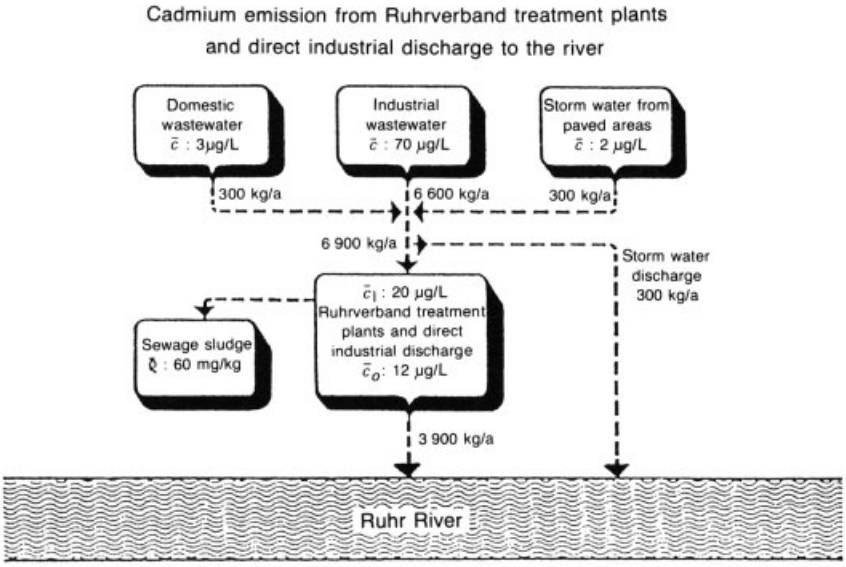

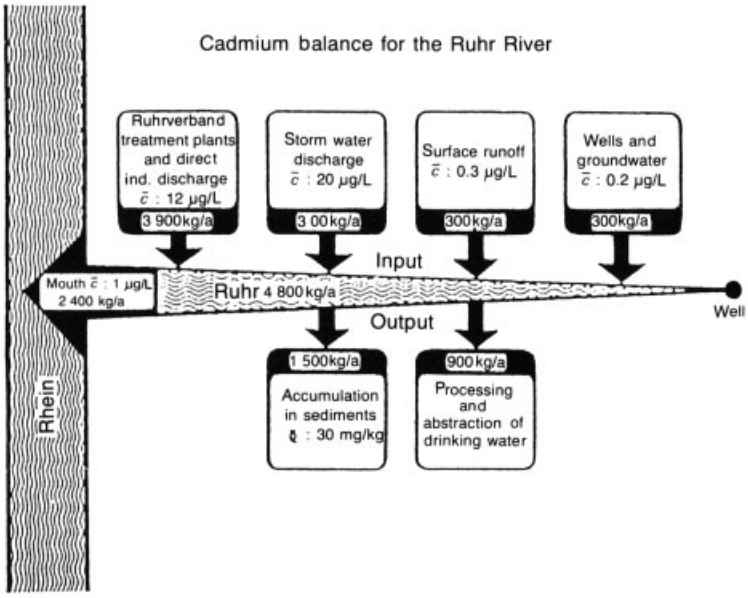

Figure 45. Origin, fate, concentrations, and annual input – output figures for cadmium in the river Ruhr (Germany) (after [393])

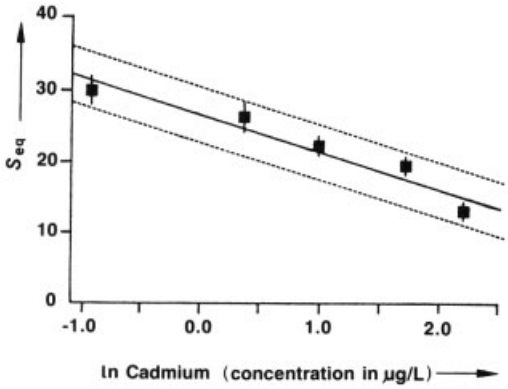

Figure 46. Microcosm analysis of the long-term effect of cadmium on the equilibrium number of species (S_{eq}) in aquatic biocenoses (after [380])

3.2.3.4. Ecological Risk Assessment of Xenobiotics

Ecological risk is a condition in which the normal functions of a population, ecosystem, or entire landscape are threatened by external forces or stress factors that, at present or in the future, may diminish the health, productivity, genetic structure, economic value, or aesthetic quality of the system [395]. In addition to xenobiotics, ecological risk may be caused by excessive inputs of nutrients; disturbance of normal energy flow by thermal pollution or global warming due to greenhouse gases; drastic changes in ecosystem structure from drying peatlands; and physical destruction of ecosystems by excessive grazing of rangelands, burning of forests, or covering with heavy loads of industrial dust.

Ecosystem risk assessment is defined as a set of procedures for measuring risk to the environment associated with the use of xenobiotics through an objective and probabilistic exercise based on empirical data and scientific judgment. The results of such a risk assessment can then be used to provide a consistent means of estimating the limiting concentrations of substances that will produce no unacceptable negative effects on ecosystems that may be exposed to the substance (see Section 3.2.5). The phases of making and reviewing an environmental risk assessment are presented in Figure 48. More about the method can be found in [68], [397], [398].

Concluding Remarks. The consequences of massive application of xenobiotics in Vietnam are drastic. Pesticide damage to nature involved the death of millions of trees and often their ultimate replacement by grasses, maintained in turn by subsequent periodic fires; deep and lasting inroads into the mangrove habitat; widespread site debilitation via soil erosion and loss of nutrients in solution; decimation of terrestrial wildlife, primarily through destruction of their habitat; losses in freshwater fish, largely because of reduced availability of food species; and a possible contribution to the decline of offshore fishery. Humans are suffering both direct and indirect effects. In addition to a variety of health problems of exposed humans, the abovementioned widespread damage to the inland and coastal forest ecosystems has had major indirect effects due to large-scale devastation of crops [67].

In the future, additional complications must be expected in the framework of xenobiotic application. These will be connected not only with the continuing adaptation of target organisms but also with changes in the total environment that are expected to be global. These include global warming accompanied by shifts in species composition, in food crops and production, and in conditions for human life. A mistake has been made in conceding to chemicals the right to be judged innocent until proven guilty. Therefore, a new strategy toward the application of xenobiotics is necessary, based on the following rules:

1) Minimize the number of newly invented chemicals in production and commercial use
2) Minimize the number of chemicals used in social and economic activities
3) Give priority to naturally occurring rather than synthetic chemicals
4) Give priority to the production of solid substances over liquids and liquids over gases
5) Minimize the spatial scope and time of their presence in the environment

3.2.4. Testing Approaches to Atmospheric Effects

Biotic Effects. For atmospheric hazard assessment, biotic and abiotic approaches are

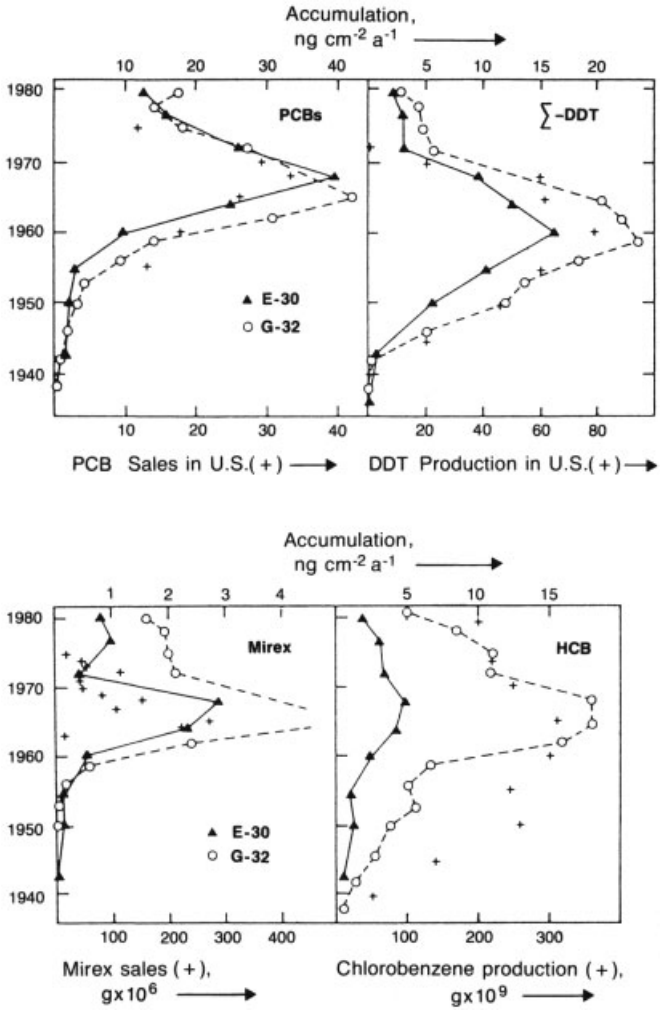

Figure 47. Annual accumulation rates of PCBs, DDT, mirex, and chlorobenzene in sediments of selected U.S. lakes (after [386])
E-30 and G-32 represent two different sampling stations (cores)

possible, although internationally accepted test guidelines to more precisely define the effects of airborne contaminants on organisms have not yet been developed. In most cases, only data on acute mammalian toxicity are available, which can be used for rough estimation of the risk a chemical poses for animals. However, atmospheric concentrations are normally not high enough to cause acute toxic effects. Therefore, data on long-term or chronic toxicity or on mutagenic and teratogenic effects would be more appropriate (→ Air, Chap. 3.).

Concerning the toxicity to terrestrial plants [2], no standardized test procedures are available for chemicals applied directly in gaseous or deposited form since a guideline for tests with herbaceous species has not yet been accepted. For some chemicals, enquiries into invertebrate toxicity – mostly insects (e.g., *Apis mellifera, Syrphus corollae*) – are available, which were conducted according to guidelines for the testing of plant protection agents.

Abiotic Effects. As regards abiotic effects, consideration of the following atmospheric

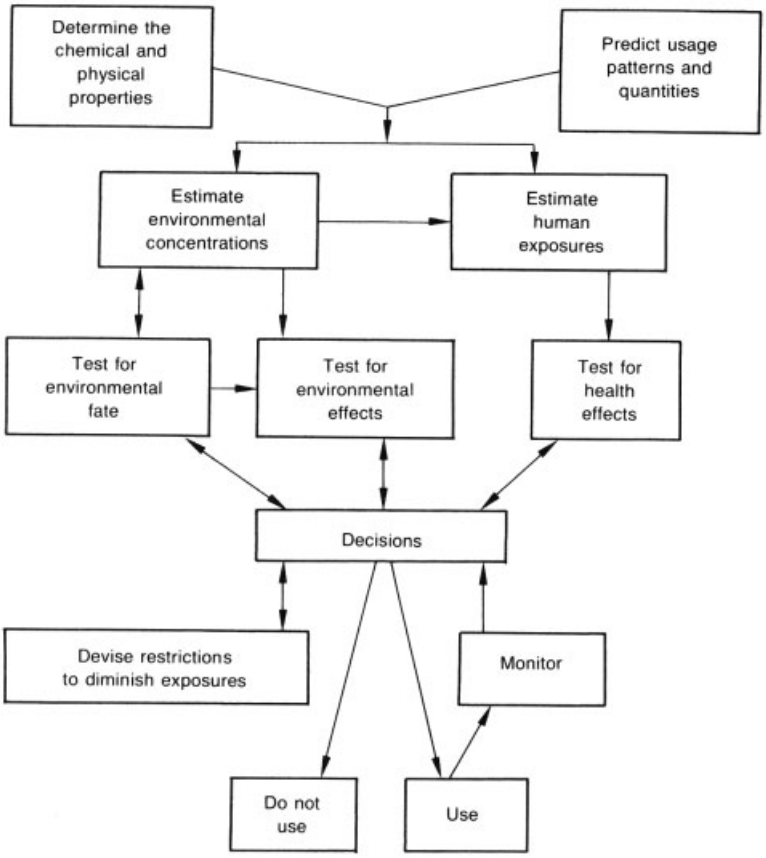

Figure 48. Environmental risk assessment (after [396])

processes is indicated [399]: global warming, stratospheric ozone depletion, tropospheric ozone formation, and acidification. A potential greenhouse gas has absorption bands in the 800–1200-cm^{-1} atmospheric window. Its *global warming potential* (GWP) is defined as the ratioof calculated warming for each mass unit of a gas emitted into the atmosphere relative to the calculated warming for a mass unit of the reference gas CFC 11 (CCl_3F). An approximate GWP value of a chemical S is given by

$$GWP = \frac{\tau_s M_{CFC11} S_s}{\tau_{CFC11} M_s S_{CFC11}} \tag{77}$$

where τ_s is the atmospheric lifetime ($\tau_{CFC11} = 60$ a); M_s the molar mass ($M_{CFC11} = 137$ g/mol); and S_s the IR absorption strength in the interval $800-1200\,cm^{-1}$ ($S_{CFC11} = 2389\,cm^{-2}$ bar).

In the above formula, the atmospheric lifetime is defined as

$$\tau = \frac{1}{k_d + k_w + k_c} \tag{78}$$

where k_d is the removal rate for dry deposition; k_w the removal rate for wet deposition; and k_c the pseudo-first-order chemical transformation rate.

A substance may have an impact on *stratospheric ozone* when its atmospheric lifetime is sufficiently long to allow for transport to the stratosphere and it contains chlorine or bromine substituents [290]. The ozone-depletion potential (ODP) is then defined as the ratio of calculated ozone change for a unit mass of gas emitted into the atmosphere relative to the depletion determined for an equal mass of the reference gas CFC 11 (ODP = 1). Thus, the ODP value can be approximated by

$$\mathrm{ODP} = \frac{\tau_s M_{\mathrm{CFC11}} n_{\mathrm{Cl}} + \alpha n_{\mathrm{Br}}}{\tau_{\mathrm{CFC11}} M_s 3} \quad (79)$$

where τ_s is the atmospheric lifetime ($\tau_{\mathrm{CFC11}} = 60$ a); M_s the molar mass of gas considered ($M_{\mathrm{CFC11}} = 137$ g/mol); $n_{\mathrm{Cl}}, n_{\mathrm{Br}}$ the numbers of Cl and Br atoms, respectively, per molecule; and α a measure of the effectiveness of Br in ozone depletion with respect to Cl ($\alpha = 30$).

The generation of *tropospheric ozone* depends on

1) The reactivity of the substance and the degradation pathway
2) The meteorological situation, with highest O_3 concentrations expected at high temperatures, intense solar radiation, and low wind speeds
3) The concentration of air pollutants, where the concentration of nitrogen oxides has to exceed several $\mu L/m^3$

Less reactive species (eg., CH_4 or CO) are important for long-term O_3 concentrations, while highly reactive species such as alkylenes or aldehydes contribute significantly to ozone peaks.

In analogy to the above depletion potential a photochemical ozone formation potential (POCP index) has been defined, based on ethylene, which has a POCP value of 100. Comparative studies showed, however, considerable variability in the POCP values assigned to each organic compound, which is difficult to explain by recourse to the basic chemical characteristics alone [2]. Therefore, ozone formation is best derived from an approximate reactivity scale based on the rate constant for the (hydrocarbon + OH·) reaction and molar mass [290]:

$$\mathrm{OH}^{\cdot}\, scale = \frac{k_s M_{\mathrm{ethylene}}}{M_s k_{\mathrm{ethylene}}} \cdot 100 \quad (80)$$

where k_s is the rate constant at $T = 298$ K for the reaction with OH radicals ($k_{\mathrm{ethylene}} = 8.5 \times 10^{-12}\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$) and M_s is the molar mass ($M_{\mathrm{ethylene}} = 28$ g/mol). For a presentation of pollutant impact on metals, organic and inorganic nonmetallic materials, see [2].

3.2.5. Appraisal

Hazard assessment of chemicals is based on stepwise determinations of predicted environmental concentrations and ecotoxicological no-effect concentrations followed by comparison of the two values. Figures 49 and 50 [239] indicate that whenever the PEC/PNEC ratio is <1.0, the chemical may be concluded not to present an environmental hazard under the conditions analyzed.

In addition to and as a substitute for specifically designed networks yielding spatially representative data, fate and exposure models are useful tools for estimating concentrations and are increasingly applied in many situations. In view of the considerable number of models available, careful selection is indicated to comply with specific problems. The fundamental problem is to define the set of questions a model is expected to answer and accuracy required at a given scale. Even if a suitable model has been chosen and valid input data have been obtained, the results can still be misinterpreted if the underlying principles are not understood properly. Thus, both the assumptions and the limitations of different modeling approaches must be recognized (see Section 3.3).

The first field of model application is the assessment of the fate of chemicals at a specific location (i.e., predicting concentrations related to well-defined sources of emission). The second is the screening of chemicals whose purpose is the evaluation of potential exposure, typically at larger or even global scales. Therefore, consideration of an average environment appears justified in the latter case.

As in the case of monitoring networks [2], the networks providing input data for the calibration and validation of models also require a geostatistical control to allow for interpolation and spatial extrapolation. Only in the case of site-specific modeling is strict validation possible, while for screening models proper validation cannot be ensured, because no specific situation is available as a reference [238]. Validation of probabilistic models is possible, but this requires a substantial amount of valid environmental concentration measurements in different locations to allow geostatistical analysis of the primary data.

As summarized in Figure 49, the principle of effects assessment is to use toxicity data from

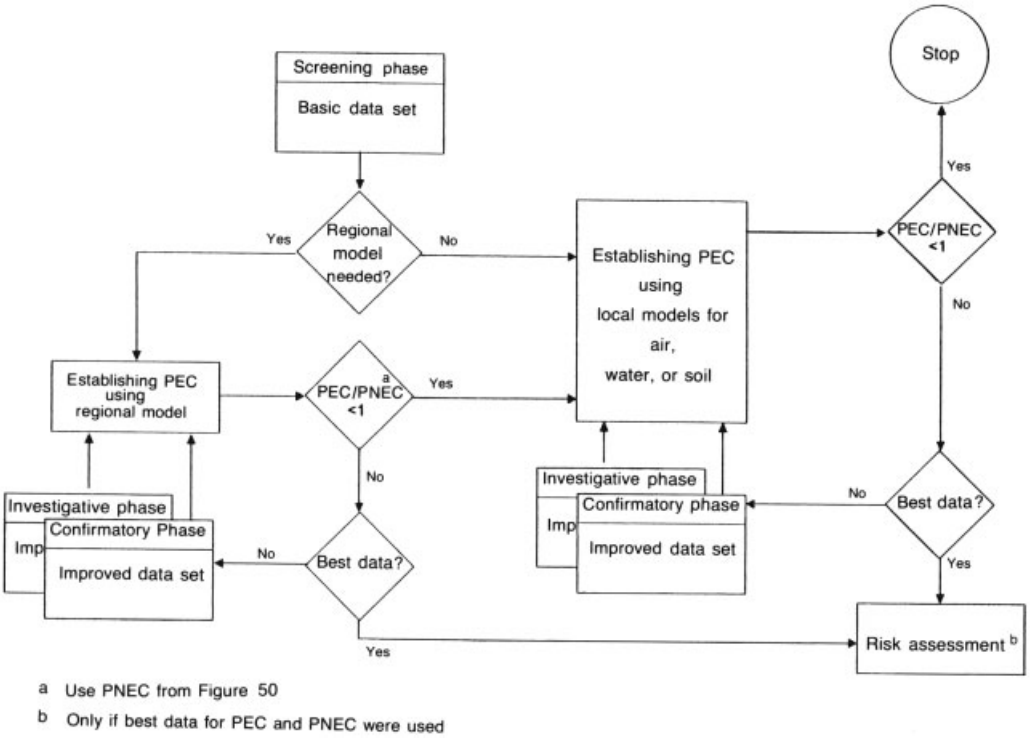

Figure 49. Exposure assessment scheme (after [239])

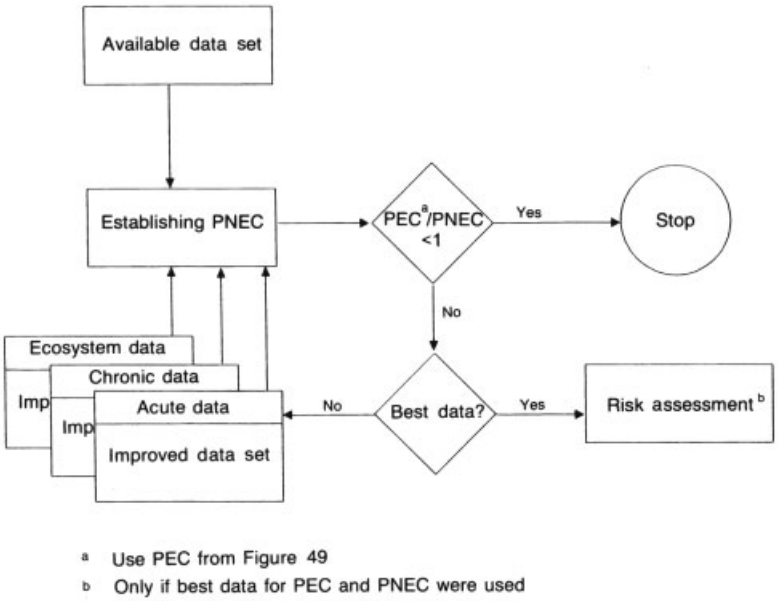

Figure 50. Effects assessment scheme (after [239])

short- and long-term experiments and appropriate assessment factors to calculate a PNEC value. In principle, this is accomplished separately for each environmental compartment, but practical difficulties may arise from the fact that most of the ecotoxicity data on existing chemicals were obtained for the aquatic compartment only.

Aquatic Hazard Assessment. According to the base set definition [289], [290], acute effect data for fish, *Daphnia*, and algae should be available. In addition, chronic toxicity investigations with standard or nonstandard organisms (see Section 3.2.1.3) may also be available. Multispecies tests, micro- and mesocosm studies, and field tests are comparatively rare for most chemicals [296]. The standard organisms used represent three trophic levels – primary producers and primary and secondary consumers – but in view of the great variability of aquatic phyto- and zoocenoses, their functional representativeness should be evaluated more precisely. In addition, a given population may represent more than one trophic level depending on the amount and spectrum of nutrition for the species; earlier life stages not infrequently live on completely different food than adults of the same species.

Therefore, ecotoxicological data collected in laboratory experiments can be difficult to translate into accurate predictions of effects in ecosystems. This applies in particular if the duration of a test is different from that of standard tests or if the test parameters (e.g., end points) are not comparable to those used in standardized test procedures (e.g., investigations of photosynthesis or behavior, studies conducted on the cellular or subcellular level).

In general, the use of laboratory test data will lead to an overestimate of chemicals' effects on complex aquatic ecosystems, which implies very conservative assessment factors, for the following reasons:

1) Bioavailability of chemicals is maximized in filtered-water laboratory systems, while in natural waters the higher concentrations of dissolved organic matter and suspended solids may reduce it substantially as a result of adsorption or complexation [291].
2) In laboratory tests, exposure concentrations are kept constant as far as possible, whereas in natural waters, exposure is likely to vary over time and degradation is greater. An overestimate is obvious when relatively long-term laboratory toxicity studies are the basis for extrapolation.
3) Laboratory experiments are normally conducted with a limited number of individuals or monospecific communites (e.g., *Daphnia magna*) and therefore exclude the various homoeotic mechanisms operating in complex natural multispecies communities.

The following features of laboratory experiments are likely to lead to an underestimate of toxic effects in ecosystems:

1) Few test species are used in laboratory test systems (see Section 3.2.1.3); therefore, they cannot be expected to encompass the most susceptible species in the field. In addition, the definition of sensitivity is dependent largely on the choice of the respective end point.
2) Information on the toxicity of substances normally relates to individual substances, while in nature complex mixtures are often present. Their effects may be additive, but the extent to which toxicant concentrations lower than the NOEC can contribute to toxicity is not yet clear [400–402].
3) In the laboratory, efforts are normally made to remove stressors other than the substance tested. In the field, however, a variety of stressors (e.g., temperature, drought, or disease) may contribute to increase the susceptibility to a chemical well above that found in the laboratory.
4) The use of fixed assessment factors between acute EC_{50} values and chronic NOECs excludes the possibility that differences in toxic mechanisms exist between acute and chronic exposure.
5) The same applies to fixed assessment factors between chronic NOECs and ecosystem NOECs since the history of the ecosystem in a monitoring study is not always known precisely enough. Therefore, an observed "anomaly" in ecosystem behavior cannot be conclusively attributed to the presence of a substance at the time of monitoring.

Hazard Assessment for Sediments. In comparison to aquatic hazard assessment, anal-

ogous approaches for sediments have not yet reached the same level of standardization. Various techniques (e.g., equilibrium partitioning, determination of interstitial water quality, or spiked sediment toxicity) are under discussion [292]. As long as ecotoxicological effect data on sediment-dwelling organisms are lacking, the $\mathrm{PNEC_{sed}}$ may be approximated by the PNEC for aquatic taxa and the sediment – water partitioning coefficient K according to the following formula:

$$PNEC_{sed} = K \cdot \mathrm{PNEC_{water}} \tag{81}$$

Terrestrial Hazard Assessment. Hitherto, hazard assessment schemes have been based predominantly on data obtained on aquatic organisms, while the amount of valid data on sediment and soil organisms, particularly of data obtained under realistic exposure conditions, is distinctly lower. Therefore, data obtained for aquatic test organisms should be used as an inital approach only to indicate whether terrestrial tests must be performed.

Aquatic organisms might display a similar susceptibility to chemicals compared to soil organisms exposed mainly to the pore water of the soil, but the transferability of aquatic effect values to pore water organisms requires further examination on the basis of comparative tests with chemicals of various groups. These have to make proper allowance for the fact that most chemicals relevant for soil either are resistant to degradation, show a strong adsorption to soil, or have a certain bioaccumulation potential. However, even readily biodegradable chemicals such as detergents may pose a problem for the terrestrial environment if applied in considerable amounts. Therefore, in many cases an adsorption potential may be more a supposition of biodegradibility than an exclusion criterion.

If no acute tests are available, one test with higher plants and one test with earthworms or microorganisms merit particular attention since they are involved in the transformation of organic matter and contribute considerably to the transformation potential of a terrestrial ecosystem.

Extrapolation of test results obtained for vertebrates (i.e., mice, rats, or birds) on the effects of a chemical in the field is difficult because of behavior- or stress-related individual or specific responses of the test organisms [255]. Furthermore, the assessment of critical environmental concentrations must be based on concentration estimates in the soil, which lead to effects via ingestion [403].

Fuzzy Set and Fuzzy Logic Applications in Ecological Research. A fuzzy set and a fuzzy logic approach [404] can be used to solve some problems of uncertainty in ecotoxicological research and practice, namely, the uncertainty of data and uncertainty of expert knowledge. Compared with the conventional methods of information processing, the application of the fuzzy set theory offers a better applicability of imprecise data and vague knowledge. This involves

1) The representation and processing of imprecise data in the form of fuzzy sets
2) The representation and processing of vague knowledge in the form of linguistic rules with imprecise terms defined as fuzzy sets

Two main applications of the fuzzy set theory in ecological and ecotoxicological research are data analysis and knowledge-based modeling. Conventional clustering methods definitely placed an object within only one cluster. With *fuzzy clustering*, an object need not be definitely placed within one cluster since the membership value of this object can be split up between different clusters. In comparison to conventional clustering methods, distribution of the membership value thus provides additional information from which the membership value can be interpreted as the degree of similarity between properties of a particular object and properties represented by particular clusters.

Kriging belongs to the most popular methods of spatial interpolation of data, but its application is sometimes restricted owing to an insufficient amount of data. The amount of available measurement data is often too low for conventional kriging methods, or exact data may be unavailable at some locations in the study area. In such a case, the data set can be completed by using additional imprecise data estimated subjectively by an expert. *Fuzzy kriging* is a modification of the conventional kriging procedure that utilizes exact (crisp) measurement data as

well as imprecise estimates obtained from an expert. These imprecise data are defined as so-called fuzzy numbers (a fuzzy number is a fuzzy set whose membership function is convex and takes its maximal value 1 for only one parameter value). Fuzzy kriging can be used to estimate interpolated values in the form of fuzzy numbers reflecting the imprecision of the input data.

Fuzzy knowledge-based modeling can be particularly useful when no analytical model of the relations to be examined exists or data are insufficient for statistical analysis [405]. To facilitate the construction of fuzzy knowledge-based models of complex ecological or toxicological systems, a modeling support system based on fuzzy logic (FLECO) has been developed [406–408]. The FLECO procedure employs fuzzy logic to handle inexact reasoning and fuzzy sets to handle the uncertainty of data. The main tasks of this system are

1) To facilitate the creation of the fuzzy knowledge base
2) To simplify changes within the fuzzy knowledge base
3) To create and facilitate the connections between submodels
4) To offer a set of inference methods
5) To facilitate the simulation process

3.3. State of the Art of Modeling in Ecotoxicology

3.3.1. Characteristics of Ecotoxicological Models

An increasing interest in management of pollution by toxic substances has emerged during the 1980s, and this has caused an equally large interest in toxic substance modeling. Models for toxic substances attempt to model the fate and effect of these substances in ecosystems. These models are management models, deterministic models, more frequently compartment models than matrix models, holistic models, and lumped models (because the difficulty of determining parameters is greater than for ecological modeling in general). Toxic substance models are most often biogeochemical models because they attempt to describe the mass flows of the toxic substances considered, although models of population dynamics also exist, which include the influence of toxic substances on birthrate or mortality and therefore should be considered toxic substance models.

Toxic substance models differ from other ecological models by the following:

1) The need for parameters to cover all possible toxic substance models is great, and general estimation methods are therefore used widely. Section 3.3.5 is devoted to this question.
2) The safety margin should be high when, for instance, expressed as the ratio between the actual concentration and the concentration that produces undesired effects.
3) The possible inclusion of an effect component, which relates the output concentration of a toxic substance to its effect, is easy. However, finding a well-examined relationship to base it on is often a problem.
4) The need is for simple models because of items (1) and (2), and because of limited knowledge of process details, parameters, sublethal effects, antagonistic and synergistic effects.

Several questions must be clarified before a toxic substance model can be developed in accordance with the general procedure used for development of ecological models. The following information must be available:

1) The best possible knowledge about the processes in which the toxic substance considered is involved in the ecosystem (as far as possible, knowledge about the quantitative importance of the processes should be obtained)
2) Obtaining parameters from the literature and/or from experiments (in situ or in the laboratory)
3) Estimation of all parameters by the methods presented in Section 3.3.5
4) Comparison of the results from items (2) and (3) and an attempt to explain the discrepancies
5) Estimation of which processes and state variables should be included in the model (if the slightest doubt exists at this early stage, too many processes and state variables should be included rather than too few)

6) Use of sensitivity analysis to evaluate the significance of individual processes and state variables, which in many cases will lead to further simplification

Section 3.3.2 reviews the characteristic features of ecotoxicological models and tries to classify them. Section 3.3.3 gives an overview of some of the ecotoxicological models published during the last 10 – 15 years. The description of the chemical, physical, and biological processes in the models will, in general, be in accordance with equations generally applied for physicochemical and biological processes. Some submodels of particular ecotoxicological interest are presented in Section 3.3.4. Section 3.3.5 deals with parameter estimation methods, which are of particular importance in ecotoxicological models. Section 3.3.6 is devoted to fugacity models in general, including their basic equations, and a case study in which the fugacity modeling approach is used is presented.

3.3.2. Classification of Ecotoxicological Models

Ecotoxicological models differ from ecological models in general by

1) Most often being more simple
2) More parameters are unknown, cannot be found in the literature
3) Wider use of parameter estimation methods
4) Possible inclusion of an effect component

Ecotoxicological models can be divided into six classes. The classification presented here is based on differences in the modeling structure. The decision on which model class to apply will be based on the ecotoxicological problem to be solved. The definitions of model classes are given below, along with the most appropriate use of each model type.

Class 1. Food Chain or Food Web Dynamic Models. Class 1 models consider the flow of toxic substances through the food chain or food web. Such models are relatively complex and contain many state variables. The models furthermore contain many parameters, which often have to be estimated by one of the methods presented in Section 3.3.5. These types of model are used when a great number of organisms are affected by the toxic substance, or the entire structure of the ecosystem is threatened by the presence of a toxic substance. Because of their complexity, these models have not been used widely. They are similar in many aspects to the more complex eutrophication models that consider the flow of nutrients through the food chain or even through the food web. In some instances, they are even constructed as submodels of a eutrophication model (see, for instance, [409]). Figure 51 shows a conceptual diagram of an ecotoxicological food chain model for lead. Lead from atmospheric fallout and wastewater enters an aquatic ecosystem, where it is concentrated through the food chain. Simplification is rarely possible for this model because the aim of the model is to describe and quantify the bioaccumulation and ecological magnification through the food chain.

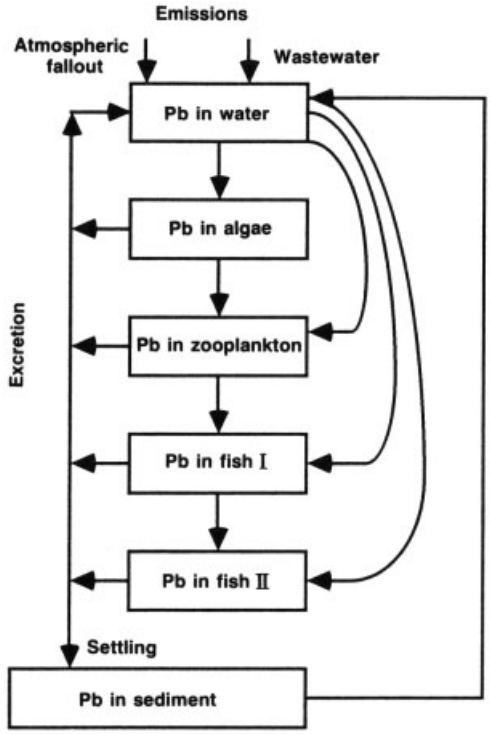

Figure 51. Conceptual diagram of lead accumulation in the food chain of an aquatic ecosystem

Class 2. Static Models of the Mass Flow of Toxic Substances. If seasonal changes are minor, or of minor importance, a static model of the mass flows is often sufficient to describe the

situation and even indicate the expected changes if the input of toxic substances is reduced or enlarged. This type of model is based on a mass balance as shown in the example in Figure 52. It often, but not necessarily, contains more trophic levels, and the modeler often is concerned with the flow of the toxic substance through the food chain. If some seasonal changes occur, this type, which in most cases is simpler than class 1, can still be used advantageously, for instance, if the modeler is concerned with the worst case of seasonal changes.

Class 3. Dynamic Model of a Toxic Substance in a Trophic Level. Often only the toxic substance concentration in one trophic level is of concern. This includes the zero trophic level, which is understood as the medium soil, water, or air. Figure 53 shows a model of copper contamination in an aquatic ecosystem as an example. The main concern is the copper concentration in the water, which may reach a toxic level for phytoplankton. Zooplankton and fish are much less sensitive to copper contamination, so the "alarm clock" rings first at the concentration level that is harmful to phytoplankton. However, only dissolved copper ions are toxic, and therefore the partition of copper in ionic, complex-bound, and adsorbed forms must be modeled. The exchange between copper in the water phase and in sediment is also included because sediment can accumulate relatively large amounts of heavy metals. The amount released from sediment may be significant under certain circumstances, for instance, at low pH.

Figure 54 gives another example. Here, the main concern is the DDT concentration in fish, which may be so high that, in accordance with WHO standards, the fish are unfit for human consumption. The model can therefore be simplified by including not the entire food chain but only fish. Some physicochemical reactions in the water phase are, however, important and are included in the diagram.

The so-called HESP model [260] estimates exposure, or the total concentration of a toxic substance for individuals living at a contaminated site. A number of possible exposure routes are identified and subsequently quantified in the model. The model can be used to initiate a hazard assessment process, which can provide a sound basis for decisions regarding possible measures.

As seen from these examples, simplifications are often feasible when the problem is well defined. Simplifications take into account which component is most sensitive to toxic matter and which processes are most important for concentration changes.

Class 4. Ecotoxicological Models in Population Dynamics. Population models are biodemographic models and therefore have numbers of individuals or species as state variables. The simple population models consider only one population. Population growth is the difference between natality and mortality:

$$dN/\mathrm{d}t = B \cdot N - M \cdot N = r \cdot N \tag{82}$$

where N is the number of individuals; B is the natality (i.e., the number of new individuals per unit of time and unit of population); M is the mortality (i.e., the number of organisms that died per unit of time and unit of population); and r is the increase in number of organisms per unit of time and unit of population (i.e., $B - M$). The concentration of a toxic substance in the environment or in the organisms influences natality and mortality, and if the relations between the concentration of a toxic substance and these population dynamic parameters are included in the model, it becomes an ecotoxicological model of population dynamics.

Population dynamic models may include two or more trophic levels, and ecotoxicological models will include the influence of the concentration of the toxic substance on natality, mortality, and interactions between these populations. In other words, an ecotoxicological model of population dynamics is a general model of population dynamics with the inclusion of the relation between toxic substance concentrations and some of the model parameters.

Class 5. Ecotoxicological Models with Effect Components. Although class 4 models include relations between concentrations of toxic substances and their effects, these are limited to population dynamic parameters. In comparison, class 5 models include more comprehensive relations between concentrations of toxic substances and effects. These models may include not only lethal or sublethal effects but also effects on biochemical reactions or on the enzyme

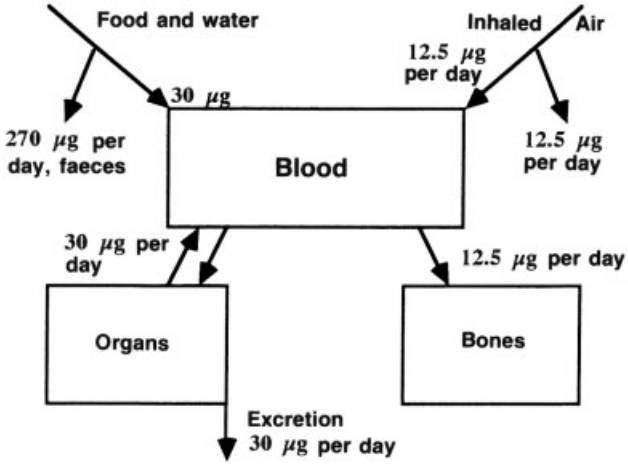

Figure 52. Static model of the daily lead uptake of an average Dane

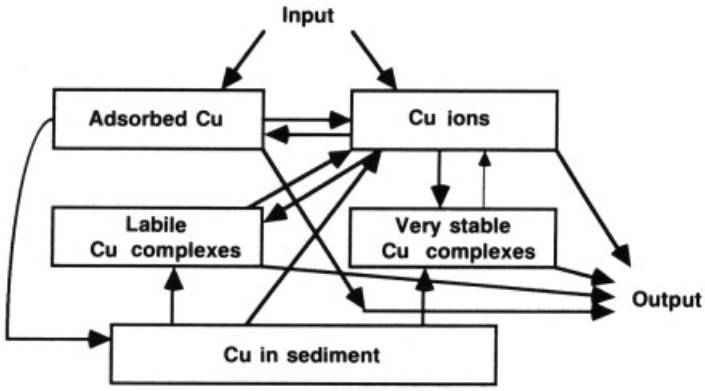

Figure 53. Conceptual diagram of a simple copper partition model

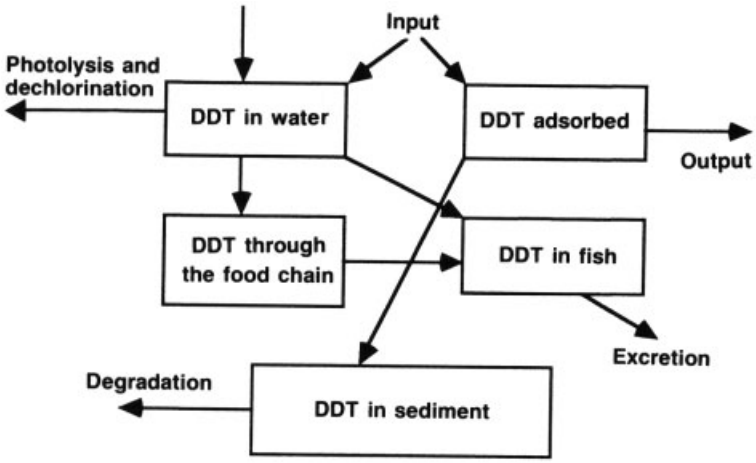

Figure 54. Conceptual diagram of a simple DDT partition model

system. As Figure 55 shows, the effects on various levels of the biological hierarchy may be considered from cells to ecosystems.

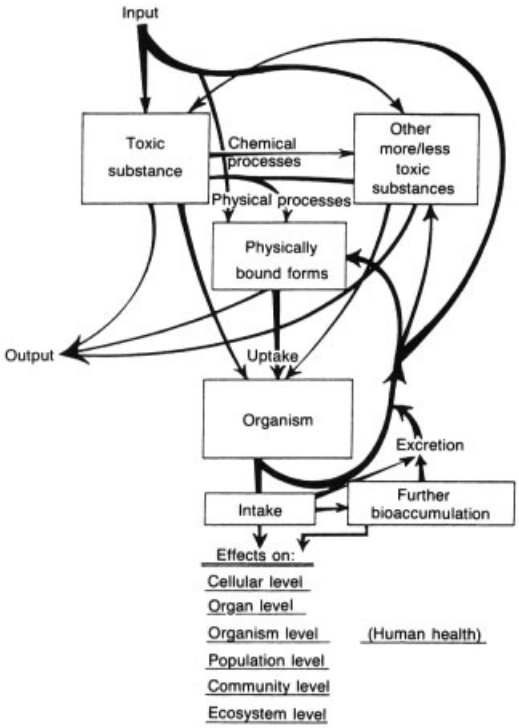

Figure 55. Conceptual diagram of ecotoxicological models with effect components on various levels of biological complexity

In many problems, more details about the effects may be required to answer the following relevant questions:

1) Does the toxic substance accumulate in the organism?
2) What will the long-term concentration in the organism be when uptake rate, excretion rate, and biochemical decomposition rate are considered?
3) What is the chronic effect of this concentration?
4) Does the toxic substance accumulate in one or more organs?
5) What is the transfer between various parts of the organism?
6) Will decomposition products eventually cause additional effects?

A detailed answer to all of these questions may require a model of the processes that take place in the organism and a translation of concentrations in various parts of the organism into effects. This implies, of course, that the intake = (uptake by the organism) × (efficiency of uptake) is known. Intake may be from either water or air, which can also be expressed by the use of concentration factors (the ratios between the concentration in the organism and that in air or water).

However, if all of the aforementioned processes are considered for just a few organisms, the model will become too complex, contain too many parameters to calibrate, and require more detailed knowledge than can be provided. Therefore, most models in this class do not consider too many details of the partition of toxic substances in organisms and their corresponding effects, but rather are limited to simple accumulation in the organisms and their effects. In most cases, accumulation is rather easy to model, and the following simple equation is often sufficiently accurate:

$$dC/\mathrm{d}t = (e_\mathrm{f} \cdot C_\mathrm{f} \cdot F + e_\mathrm{m} \cdot C_\mathrm{m} \cdot V)/W - Ex \cdot C = (INT)/W - Ex \cdot C \quad (83)$$

where C is the concentration of the toxic substance in the organism; e_f and e_m are the efficiencies for the uptake from the food and medium, respectively (water or air); C_f and C_m are the concentration of the toxic substance in the food and medium, respectively; F is the amount of food uptake per day; V is the volume of water or air taken up per day; W is the body weight as either dry or wet matter; and Ex is the excretion coefficient (d^{-1}). As seen from the equation, INT covers the total intake of toxic substance per day.

This equation has a numerical solution:

$$C/C\,(\max) = \{INT \cdot [1 - \exp(Ex \cdot t)]\}/(W \cdot Ex) \quad (84)$$

where C (max) is the steady-state value of C:

$$C\,(\max) = INT/(W \cdot Ex) \quad (85)$$

i.e., the concentration of the toxic substance approximates a steady-state value after a certain time interval.

Class 6. Fate Models with or without a Risk Assessment Component. The last type of ecotoxicological models focuses on the fate of toxic substances: Where in the ecosystem will the toxic substance be found? In what concentration? These models are sometimes called multimedia models because they focus on the distribution of toxic substances in the four spheres. The complete solution of an ecotoxicological problem requires in principle four (sub)models, of which the fate model may be considered the first model in the chain; see Figure 56. The four components are [410] (1) a fate or exposure model; (2) an effect model, translating the concentration into an effect (see class 5); (3) a model for human perception processes; and (4) a model for human evaluation processes. The first two submodels are in principle "objective," predictive models, whereas the latter two are value oriented. The development of submodels (1) and (2) is based on physical, chemical, and biological processes. They are very similar to other environmental models and are based on mass transfer; mass balances; and physical, chemical, and biological processes.

The second submodel requires good knowledge of the effects of the toxic components. Submodels (3) and (4) differ from the generally applied environmental management models and are presented in some details below in the discussion of risk assessment.

Fugacity models (see Section 3.3.6) are a special type of fate model. They attempt to answer the following questions: In which of these six compartments (air, water, soil, sediment, suspended sediment, and biota) can the greatest problem be expected for chemicals emitted to the environment? What concentration will be expected in each compartment? What are the implications of these concentrations? The fugacity models are to a large extent based on physicochemical parameters and the use of these parameters to estimate other required parameters. The applicability of the fugacity model is therefore very dependent on the use of the estimation methods presented in Section 3.3.5.

The risk assessment component, associated with the fate model comprises human perception and evaluation processes (see Fig. 56). These submodels are explicitly value laden but must of course build on objective information about concentrations and effects. Factors that may be important to consider in this context are

1) Magnitude and time constant of exposure
2) Spatial and temporal distributions of concentrations
3) Environmental conditions determining the process rates and effects
4) Translation of concentrations into magnitude and duration of effects
5) Spatial and temporal distribution of effects
6) Reversibility of effects

The uncertainties relating to the information on which the model is based and the uncertainties related to the development of the model, are crucial in risk assessment. The uncertainty in risk assessment may be classified in one or more of the following five categories:

1) Good direct knowledge and statistical evidence on the important components (state variables, processes, and interrelations of the variables) of the model are available.
2) Good knowledge and statistical evidence on the important submodels are available, but the aggregations of the submodels are less certain.
3) No good knowledge of the model components for the considered system is available, but good data are available for the same processes from a similar system, and these data are considered applicable for use directly or with minor modifications, to the model development.
4) Some, but insufficient, knowledge ist available from other systems. Attempts are made to use these data without the necessary transferability. Attempts are made to eliminate gaps in knowledge by the use of additional experimental data as far as possible within the limited resources available for the project.
5) The model is to a large extent based on the subjective judgment of experts.

Acknowledgment of the uncertainty is of great importance and may be taken into consideration, either qualitatively or quantitatively. Another problem is of course: Where and how should the uncertainty be taken into account? Should the economy or the environment benefit from the uncertainty? If the results of the management strategy are to be ensured, the un-

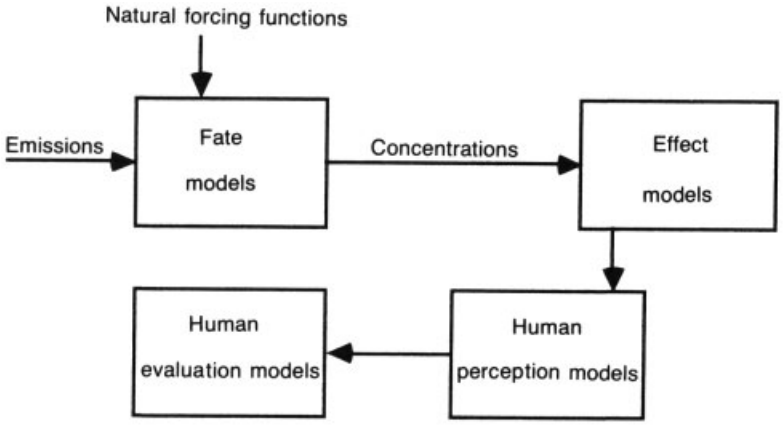

Figure 56. Components and general structure of comprehensive ecotoxicological modeling systems

certainty must be used to reduce the emission standards.

Until 10 – 15 years ago, researchers had very little understanding of the processes by which people actually perceive the exposures and effects of toxic chemicals, but these processes are just as important for risk assessment as the exposure and effect processes. The following list of risk characteristics summarizes the concerns:

1) Is the risk voluntary or involuntary?
2) Are the levels known to the exposed people or only to science?
3) Is the risk novel, or old and familiar?
4) Is it common or dreaded (e.g., does it involve cancer) ?
5) Does it involve death?
6) Are mishaps controllable?
7) Are future generations threatened?
8) Is the risk global, regional, or local?
9) Is the risk a function of time? How does it depend on time (e.g., increase or decrease) ?
10) Can it be reduced easily?
11) Is the expected effect immediate or delayed?
12) Are many or a few people involved?
13) Are the effects observable immediately?

A factor analysis has been performed by Slovic et al. [411] that shows, among other results, a not surprising emphasis on people's perception of dreadful and unknown risks. Broadly speaking, two methods of selecting the risks exist. The first may be described as the *rational actor model*, involving people that look systematically at all the risks they face and make choices about which they will live with and at what levels. For decision making, this approach would use some single, consistent, objective functions and a set of decision rules. The second method may be called the *political – cultural model*. It involves interactions among culture, social institutions, and political processes for the identification of risks and the determination of those that people will live with and at what level.

Both methods are unrealistic, since they are both completely impractical in their pure form. Therefore, a strategy for risk abatement must be selected that is founded on a workable alternative based on the philosophy behind both methods.

Several risk management systems are available. Here, some recommendations should be given for the development of risk management systems:

1) Consider as many of the characteristics listed above as possible and include human perception of these characteristics in the model.
2) Do not focus too narrowly on certain types of risks; this may lead to suboptimal solutions. Attempt to approach the problem as broadmindedly as possible.
3) Choose strategies that are pluralistic and adaptive.
4) Benefit – cost analysis is an important element of the risk management model, but it is far from being the only important element, and the uncertainty in evaluation of benefit

and cost should not be forgotten. The variant of this analysis applicable to environmental risk management may be formulated as follows :
Net social benefit = social benefits of the project minus "environmental" costs of the project

5) Use multiattribute utility functions, but remember that people in general have trouble thinking about more than two to three (at most, four) attributes in each outcome.

Presentation of the six classes of models above clearly shows the advantages and limitations of ecotoxicological models. The simplifications used in classes 2, 3, and 6 (at least without risk assessment components) often offer great advantages. They are sufficiently accurate to give a very applicable picture (overview) of the concentrations of toxic substances in the environment, because of the use of large safety factors. Application of the estimation methods presented in Section 3.3.5 makes feasible the construction of such models, even though knowledge of the parameters is limited. The estimation methods obviously have a high uncertainty, but the great safety factor helps in accepting this uncertainty. On the other hand, knowledge about the effects of toxic substances is very limited, particularly at the organism and organ level. Therefore, models with effect components must not be expected to give more than a first rough picture of what is known today in this area.

Because of the character of ecotoxicological models, a few questions should be clarified before beginning the modeling procedure:

1) Obtain the best possible knowledge about the processes of the toxic substance under consideration
2) Attempt to get parameters of the processes in which the toxic substance is involved in the environment from the literature
3) Estimate all parameters using methods presented in Sections 3.3.4 and 3.3.5
4) Compare the results from items (2) and (3) and attempt to explain discrepancies, if present
5) Use sensitivity analysis widely to estimate which processes and state variables would be feasible and relevant to include in the model

3.3.3. An Overview: Application of Models in Ecotoxicology

Some toxic substance models are reviewed in Table 20 to give an idea of the types of models available today. Most models reflect the proposition that good knowledge of the problem and the ecosystem can be used to make reasonable simplifications. Model characteristics indicated in Table 20 are state variables or processes considered in the model. Note the number of modeled toxic substances and the processes taken into account. The model class, according to the classification 1 to 6 discussed in Section 3.3.2, is given in brackets after the toxic substance. Only a few class 4 models are included in Table 20. Ecological modeling has been approached from two sides: population dynamics and biogeochemical flow analysis. Since the second approach has been most involved in environmental issues, the toxic substance problem has also been approached from this angle. The few class 4 models, including the example given in Table 20, are population dynamic models, with some additional equations to account for the influence of toxic substances on natality and mortality. Provided these relations are available, constructing this type of model should be relatively easy.

The most difficult part of modeling the effect and distribution of toxic substances is obtaining relevant knowledge about the behavior of toxic substances in the environment and using this knowledge to make feasible simplifications. The modeler of ecotoxicological problems is particularly challenged to select the proper, balanced complexity, and many examples can be found of rather simple ecotoxicological models that are able to solve the basic problem.

It can be seen from the overview in Table 20 that most ecotoxicological models have been developed during the last decade. Before ca. 1975, toxic substances were rarely associated with environmental modeling, since the problems seemed straightforward. The many pollution problems associated with toxic substances were believed to be easily solved simply by eliminating the source of the toxic substance. During the 1970s, the environmental problems originating from toxic substances were acknowledged to be very complex because of the interaction of

Table 20. Examples of toxic substance models

Toxic substance, model class	Model characteristics	Reference
Cadmium (1)	food chain similar to a eutrophication model	[412]
Mercury (6)	6 state variables: water, sediment, suspended matter, invertebrates, plant, fish	[413]
Vinyl chloride (3)	chemical processes in water	[414]
Methyl parathion (1)	chemical processes in water and microbial degradation and adsorption of benzothiophene, 2 – 4 trophic levels	[415]
Methyl mercury (4)	a single trophic level: food intake, excretion, metabolism, growth	[416]
Heavy metals (3)	concentration factor, excretion, bioaccumulation	[417]
Pesticides in fish, DDT and methoxychlor (5)	ingestion, concentration factor, adsorption on body, defecation, excretion, chemical decomposition, natural mortality	[418]
Zinc in algae (3)	concentration factor, secretion, hydrodynamic distribution	[419]
Copper in the sea (5)	complex formation, adsorption, sublethal effect of ionic copper	[420]
Lead (5)	hydrodynamics, precipitation, toxic effects of free ionic lead on algae, invertebrates, and fish	[421]
Radionuclides (3)	hydrodynamics, decay, uptake, and release by various aquatic surfaces	[69]
Radionuclides (2)	in grass, grains, vegetables, milk, eggs, beef, and poultry–state variables	[422]
SO_2, NO_x, and heavy metals on spruce forests (5)	threshold model for accumulation effect of pollutants (air and soil)	[369]
Toxic environmental chemicals in general tests (6)	hazard ranking and assessment from physicochemical and a limited number of laboratory data	[423]
Heavy metals (3)	adsorption, chemical reactions, ion exchange	[424]
Polycyclic aromatic hydrocarbons (3)	transport, degradation, bioaccumulation	[424]
Persistent toxic organic substances (3)	groundwater movement, transport and accumulation of pollutants in groundwater	[425]
Cadmium, PCB (2)	hydraulic overflow rate (settling), sediment interactions, steady-state food chain submodel	[409]
Hydrophobic organics (3)	gas exchange, sorption/desorption, hydrolysis compounds, photolysis, hydrodynamics	[426]
Mirex (3)	water – sediment exchange processes, adsorption, volatilization, bioaccumulation	[427]
Toxins (aromatic hydrocarbons, Cd) (3)	hydrodynamics, deposition, resuspension, volatilization, photooxidation, decomposition, adsorption, complex formation, (humic acid)	[428]
Heavy metals (2)	hydraulic submodel, adsorption	[429]
Oil slicks (3)	transport and spreading, influence of surface tension, gravity, and weathering processes	[430]
Acid rain (soil) (3)	aerodynamics, deposition	[70]
Acid rain	C, N, and S cycles and their influence on acidity	[431]
Persistent organic chemicals (6)	fate, exposure, and human uptake	[432]
Chemicals, general (6)	fate, exposure, ecotoxicity for surface water and soil	[433]
Toxicants, general (4)	effect on populations of toxicants	[434]
Chemical hazard (6)	basin-wide ecological fate	[435]
Pesticides (4)	effects on insect populations	[436]
Insecticides (2)	resistance	[437]
Mirex and lindane (6)	fate in Lake Ontario	[427]
Pesticides (3)	degradation in soil	[438]
Pesticides (3)	degradation in soil	[439]
Pesticides (3)	leaching to groundwater	[440]
Acid rain (5)	effects on forest soils	[70]
Acid rain (5)	cation depletion of soil	[441]
pH, calcium, and aluminum (4)	survival of fish populations	[442]
Photochemical smog and NO_x (6)	fate and risk	[443]
Nitrate (3)	leaching to groundwater	[444]
Oil spill (6)	fate	[445]
Toxicants (4)	effects on populations	[446]
Chromium (2)	distribution and accumulation in mussels	[71]
Toluene, benzene, Zn, Ni, Pb (3)	assess exposure levels of soil contaminants for humans	[447]

many sources and of many processes and components simultaneously. Several accidental releases of toxic substances into the environment have reinforced the need for models. Thus, various ecotoxicological models have been developed since the late 1970s. The list in Table 20 represents a comprehensive survey of available ecotoxicological models, but it should not be considered complete since it is not the result of an exhaustive literature review.

Table 21. Regression equations for estimation of concentration, bioconcentration, and ecological magnification factors

Indicator	Relationship	Correlation coefficient	Range (indicator)
K_{ow}	$\log CF = -0.973 + 0.767 \log K_{ow}$	.76	$2.0\times10^{-2} - 2.0\times10^{6}$
K_{ow}	$\log CF = 0.7504 + 1.1587 \log K_{ow}$	.98	$7.0 - 1.6\times10^{4}$
K_{ow}	$\log CF = 0.7285 + 0.6335 \log K_{ow}$	.79	$1.6 - 1.4\times10^{4}$
K_{ow}	$\log CF = 0.124 + 0.542 \log K_{ow}$	.95	$4.4\times100 - 4.2\times10^{7}$
K_{ow}	$\log CF = -1.495 + 0.935 \log K_{ow}$	.87	$1.6\times100 - 3.7\times10^{6}$
K_{ow}	$\log CF = -0.70 + 0.85 \log K_{ow}$	.95	$1.0 - 1.0\times10^{7}$
K_{ow}	$\log CF = 0.124 + 0.542 \log K_{ow}$	.90	$1.0\times100 - 5.0\times10^{7}$
S, µg/L	$\log BCF = 3.9950 - 0.3891 \log S$	.92	$1.2 - 3.7\times10^{7}$
S, µg/L	$\log BCF = 4.4806 - 0.4732 \log S$	.97	$1.3 - 4.0\times10^{7}$
S, µmol/L	$\log BCF = 3.41 - 0.508 \log S$	.96	$2.0\times10^{-2} - 5.0\times10^{3}$

3.3.4. Ecotoxicological Processes

Ecotoxicological processes differ from other ecological processes, and a review of the relevant processes and their quantitative description is therefore necessary as basis for model development. A survey of the most crucial ecotoxicological processes and the most applied mathematical descriptions of these processes is given below.

Concentration Factor. The concentration factor relates the concentration of the xenobiotic in the medium to the concentration in organisms, sediment, or suspended matter. The concentration in organisms is a result of uptake by respiration, and the concentration in sediment or suspended matter is caused by adsorption and sedimentation processes. The concentration factor *CF* can often be described with sufficient accuracy as a constant independent of the concentration in the medium (water):

$$CF = C_S / C_W \tag{86}$$

where C_S is the concentration in the organisms, suspended matter, or sediment and C_W is the concentration in water (or air). If the uptake includes the indirect route through food, the term bioconcentration factor *BCF* is often used ($BCF \geq CF$), which can be described by the same equation.

Table 21 gives equations relating water solubility and the *n*-octanol – water partition coefficient to *CF* and *BCF*. The *n*-octanol – water partition coefficient K_{ow} is defined in Section 2.4.1. An overview on relevant literature references dealing with the determination of the *n*-octanol –water partition coefficient of various chemicals in fish is given in Table 22. If the *n*-octanol –water partition coefficient is not known, it can be estimated from the water solubility of the corresponding substance (see Section 3.3.5 and Fig. 58).

Table 22. Determination of n-octanol – water partition coefficients in fish

Animal	Number of chemicals	References
Fish species	36	[448]
Mosquito fish	9	[449]
Mosquito fish	11	[450]
Trout	8	[451]
Fish species	26	[448]
Fathead minnow	59	[452]
Fathead minnow, bluegill	59	[453]
Fish species	13	[448]
Fish species	50	[448]
Mosquito fish	9	[449]
Fish species	36	[448]
Mosquito fish, whole	15	[450]

Biomagnification describes the process whereby pollutants are passed from one trophic level to the next and increasingly concentrated in organisms with higher trophic status (see Section 2.4.2.7). The magnification (i.e., the increasing concentration through the food chain) is caused by greater retention of toxic substances than of general food components by the organisms. The process may be described by the biological or ecological magnification factor, which indicates the magnification from one level in the food chain to the next.

Further important processes in ecological modeling are adsorption (see Section 2.4.2.4) and desorption.

Excretion of toxic substances by an organism can be shown to follow first-order kinetics to a

good approximation, and the rate coefficient depends on the chemical compound and the size of the organism considered (see [37]). This process is in most cases insignificant for the entire mass balance of the components considered but inclusion of the excretion process may be important for a model of a toxic substance on the organism level.

Biodegradation. It is rather difficult to estimate the reaction rate of biodegradation in ecotoxicology because a large number of factors influence the reaction and thereby also the concentration of microorganisms. The most important of these factors are: temperature, sorption, redox potential, ionic strength, adaptation of microbial populations, macronutrients and micronutrients available [454], [455].

Distribution of a Chemical Compound. Distribution of a chemical compound in the four spheres is of particular interest in fate models.

Distribution air – water

$$\frac{C^*}{To^*} \equiv F_{\mathrm{L}} = \frac{H18}{1000RT} \text{ (approximation)} \tag{87}$$

or at (103.33 kPa) and 20 °C

$$F_{\mathrm{L}} = 7.49 \cdot 10^{-4} H \tag{88}$$

where H is Henry's constant (bar); $C*$ and $To*$ are equilibrium concentrations (mass/volume) in air and water, respectively.

Rate of evaporation (distribution liquid – vapor phase)

$$Ev = -D \cdot A \frac{\mathrm{d}To}{\mathrm{d}x} = bK_{\mathrm{L}} \cdot A \cdot (To - To^*) \tag{89}$$

where D is the molecular diffusion coefficient (area/time), A is the area, and b is a coefficient that adjusts the evaporation coefficient K_{L} (length/time) to substances other than oxygen ($b = 1.0$ for oxygen); b can be found in [72] for other compounds. Several empirical equations for calculation of K_{L} can be found in [73], [74].

$$b = \frac{K_{\mathrm{L}}\,(\text{compound})}{K_{\mathrm{L}}\,(\text{oxygen})} = \frac{D\,(\text{compound})}{D\,(\text{oxygen})} \tag{90}$$

For lakes and streams,

$$To\,(t) = To\,(0) \cdot \mathrm{e}^{-b \cdot K_{\mathrm{L}} \cdot \mathrm{a}} \cdot t \tag{91}$$

where

$$a = \frac{A}{V} = \frac{\text{area}}{\text{volume}} \tag{92}$$

Distribution water – solid (sediment)

A distribution coefficient may be used at low concentrations (see also Langmuir's and Freundlich's adsorption isotherms):

$$F_{\mathrm{S}} = \frac{C^*}{To^*} \tag{93}$$

where $C*$ and $To*$ are the concentrations in the solid and water phases at equilibrium. Since the dimension of $C*$ is mass/mass (e.g., mg per gram dry matter) and $To*$ has the dimension of mass/volume (mg/L), F_{S} has the dimension volume/mass (L/g or L/kg).

F_{S} may be found from

$$F_{\mathrm{S}} = F_{\mathrm{oc}} \cdot f_{\mathrm{oc}} \tag{94}$$

where F_{oc} is the distribution coefficient for organic matter and f_{oc} is the fraction of organic carbon in the solid being considered.

The following relationship between F_{oc} and K_{ow} exists [456]:

$$\log F_{\mathrm{oc}} = -0.006 + 0.937 \log K_{\mathrm{ow}} \tag{95}$$

MATTER-MÜLLER et al. [457] found the following relationship for activated sludge:

$$\log F_{\mathrm{S}} = 0.39 + 0.67 \log K_{\mathrm{ow}} \tag{96}$$

where F_{S} has the units L/kg.

F_{oc} and F_{S} values for selected chemical compounds can be found in the literature.

Complex Formation. As for other chemical processes, the law of mass action is valid for complex formation reactions. For the process

$$Me^{n+} + \mathrm{L}^{m-} = \mathrm{MeL}^{(n-m)+} \tag{97}$$

(where Me denotes a metal and L a ligand), the law of mass action has the following form:

$$\frac{[MeL^{(n-m)+}]}{[\mathrm{Me}^{n+}]\,[\mathrm{L}^{m-}]} = K \tag{98}$$

K is termed the stability constant if the process – as in this case – is the formation of a complex.

The formation of coordination complexes plays a major role in modeling the distribution as well as the effect of metal ions in aquatic ecosystems because of the following influences:

1) Increase in metal solubility

$$MeY\,(S) + L = MeL + Y \tag{99}$$

where Me represents a metal that is bound in the sediment in the compound MeY; L denotes a ligand
2) Alteration of the distribution between oxidized and reduced forms of metals. The equilibrium constant for the process

$$Me^{n+} + m\mathrm{e}^- = \mathrm{Me}^{(n-m)+} \tag{100}$$

differs from the one for

$$MeL^{n+} + m\mathrm{e}^- = \mathrm{MeL}^{(n-m)+} \tag{101}$$

if the complexes, $\mathrm{MeL}^{\mathrm{n}+}$ and $\mathrm{MeL}^{(\mathrm{n}-\mathrm{m})+}$, have different stability constants, which often is the case
3) Alleviation of toxicity due to alteration of the availability to aquatic life
4) Change in adsorption and ion exchange on sediment and suspended matter
5) Change in stability of the metal-containing colloids

The presence of ligands in water generally increases the transfer of metals from sediment, soil, and suspended matter and thereby increases the solubility. However, since complexes show lower toxicity than metal ions, the harm to aquatic life in most cases does not increase in proportion to the increased solubility.

The application of coordination chemistry to environmental problems is very complex because many ligands are present in aquatic ecosystems and all ligands compete simultaneously for formation of complexes with metal ions.

Photolysis. The *Stark–Einstein law* states that if a species absorbs radiation, one particle is excited for each quantum of radiation absorbed. The energy of the particle is limited by the energy of the photon absorbed. This energy is given by

$$E = h\nu \tag{102}$$

where E is the energy of the photon, h is Planck's constant, and ν is the frequency of the photon. Thus, depending on the nature of the incident light and the property of the chromophore, different excited states of the chromophore may be formed. Although not exclusively true, in general only light absorption that brings about electronic excitation of the chromophore will result in photochemical reactions. Excitation of the vibrations and rotations of the absorbing species in its ground electronic state is less likely to induce chemical changes.

The effectiveness of photochemical reactions in the consumption or production of a given chemical species is given by the *overall quantum yield* of that process (see Section 2.4.2.2). The overall quantum yield is the number of molecules of a given reactant consumed for each photon absorbed. It may be further subdivided into the primary quantum yield (molecules consumed by primary reactions) and the secondary quantum yield (molecules consumed by secondary reactions).

If the quantum yield is not strongly wavelength dependent, for an incident beam of light with a certain spectral characteristic, the complete rate law for direct photolysis is

$$-\left(d\,[CA]\,/\mathrm{d}t\right) = k_{\mathrm{a}}\,[CA] \tag{103}$$

where CA is the concentration of the reactant and k_{a} is the sum of the k_{a} values of all the wavelengths represented by the photons in the incident light. Models for estimating the rates of direct photolysis in natural waters under restrictive conditions have been attempted by several investigators [458–466].

Acid – base reactions are of a great environmental interest because almost all processes in the environment depend on pH. A few illustrative examples are included in the following list:

1) Ammonia is toxic to fish and the ammonium : ammonia ratio depends on pH.
2) Carbon dioxide is toxic to fish and the ratio of hydrogencarbonate to carbon dioxide depends on pH.
3) The fertility of fish and zooplankton eggs is highly pH dependent.
4) All biological processes have a pH optimum, which is usually in the range 6 – 8. This implies that algal growth, microbiological decomposition, nitrification, and denitrification are all influenced by pH.

5) The release of heavy-metal ions from soil and sediment increases very rapidly with decreasing pH. Heavy-metal hydroxides have a very low solubility product, which implies that most heavy-metal ions are precipitated at pH 7.5 or higher.

Acid rain illustrates the problems related to a low pH value, but too high a pH is also of environmental concern because of items (1) and (4) above.

Other Chemical Processes. In addition, hydrolysis and redox processes may also be important for the speciation and general presence of toxic substances. Many metal hydroxides may be dissolved by hydrolysis, and the release of metal ions from lake sediments is highly dependent on the redox potential at the interface between water and sediment.

Organic compounds can also undergo hydrolysis reactions. Many compounds react with water, with introduction of a hydroxyl group into the molecule. Most toxic organic compounds are oxidized by oxygen, and carboxyl groups may be formed. For all of these reactions, a chemical pathway should be found and the application of simple first-order reaction kinetics is often sufficient to describe the process in an ecotoxicological model.

3.3.5. Parameter Estimation

About 50 000 chemicals are produced on a large scale (1000 kg or more per year), but only about 4000 chemicals are produced at a quantity of $\geq 1\times10^6$ kg/a. Approximately one-third of these chemicals are well-defined organics. Their environment-relevant properties are listed below:

Physical properties
vapor pressure, water solubility, Henry's constant, adsorption coefficient, partition coefficient octanol – water, mass-transfer coefficient, density, viscosity, diffusion coefficient in air and water, boiling and melting point

Chemical properties
rate of hydrolysis, rate of oxidation or reduction, rate of photolysis, acid dissociation coefficient(s)

Biological properties
bioconcentration factor, rate of biodegradation in soil and water under aerobic or anaerobic conditions

Most of the biological parameters are needed for all species or at least a representative fraction of the species on earth. This implies that for construction of all possible ecotoxicological models, about 20 000 parameters are needed for 50 000 chemicals (i.e., a total of ca. 1×10^9 parameters). A review of the literature reveals that hardly 1 % of these parameters, or 10×10^6 values are known. Providing the 9.9×10^8 parameters lacking will be an enormous task – a task that will take several decades to accomplish. How can ecotoxicological models be constructed under these circumstances? The answer is to estimate the parameters, which is acceptable, because ecotoxicological models do not require a high accuracy, or rather, one is forced to interpret the model results by the use of large safety factors, which makes it less necessary to demand great accuracy of the modeling results.

Many estimation methods are available for ecotoxicological parameters, and too many pages would be required to present even an overview of these methods. JØRGENSEN [37] presents a rather simple method that can be applied easily and still give sufficient accuracy for development of most ecotoxicological models.

The *physical properties* of interest in ecotoxicology are vapor pressure P, water solubility S, partition coefficient K_{ow}, and Henry's constant H; for determination of the transfer between air and water, adsorption isotherms K_{ac} and diffusion coefficients D are also necessary. For organic compounds, all these properties can be estimated just on the basis of their chemical structure. Some estimation methods are based on estimation of the boiling point and critical properties, which are then used later to estimate the aforementioned physical properties. Experience shows that generally better estimates can be given if the boiling point is known and not found by estimation methods. Simultaneous use of two or more estimation methods will always give a better overall estimation (an average value is used) than if only one method is applied. All relevant equations for the described estimation system are given in [37]. The system itself

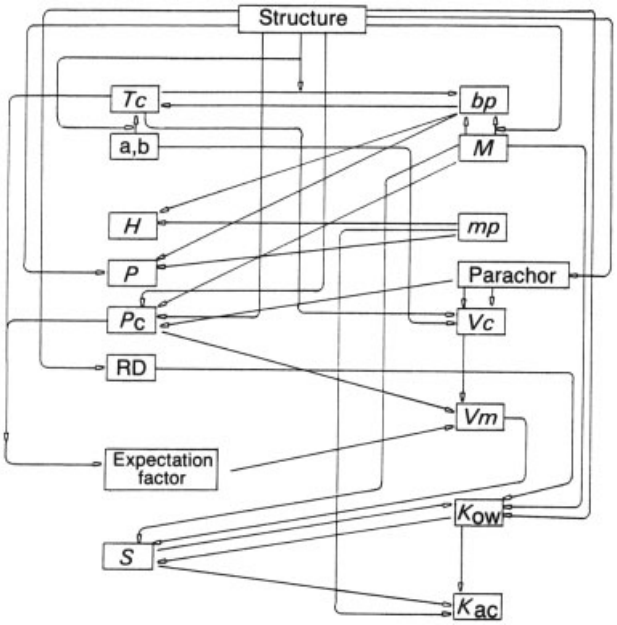

Figure 57. Conceptual diagram of an estimation system for ecotoxicological parameters (after [37])

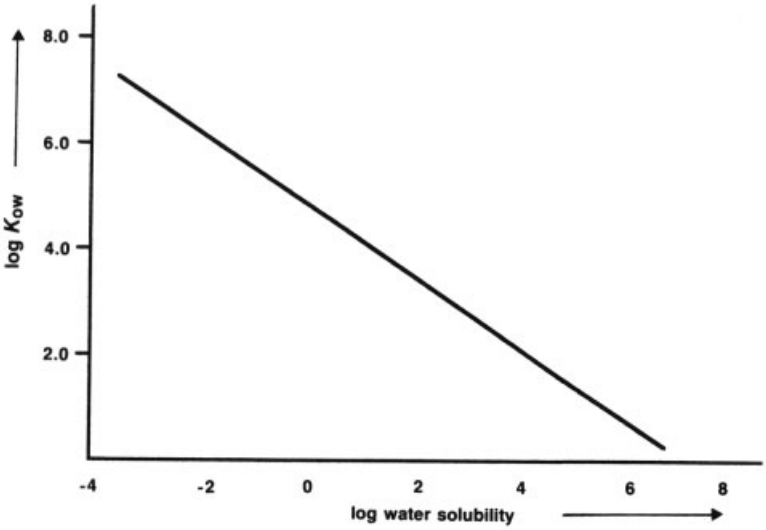

Figure 58. Linear regression between log K_{ow} and log S (water solubility in μmol/L), based on a range of organic compounds including aliphatic and aromatic hydrocarbons, aromatic acids, organochlorines, organophosphorus insecticides, and PCBs (after [467])

is shown conceptually in Figure 57 while the correlation between the partition coefficient *n*-octanol – water and the water solubility is given in Figure 58.

Chemical properties (i.e., rate of hydrolysis, rate of oxidation or reduction, and acid dissociation coefficients) can also be estimated.

Obviously, *biological properties* are very important in ecotoxicological modeling, and since these properties are generally even less well

known than physical and chemical properties, the estimation methods are even more important. Figure 59 shows how to estimate *CF*, *BCF*, and the ecological magnification *EM* by use of solubility data *S* and partition coefficients K_{ow}, see also the relation in Figure 60. From the structure, furthermore, biodegradation rate coefficients and toxicity can be estimated. Some of the methods work via molar mass *M* and estimations of *mp* and *bp* (see Fig. 59).

The estimation methods used in practical ecotoxicological modeling have been tested widely on compounds with known properties and have been found to have a standard deviation of 10 – 60 % for most parameters, although the standard deviations are higher for toxicological parameters. This is an acceptable accuracy since large safety factors are recommended and applied in ecotoxicological modeling. Although the number of relevant parameters that can be found in the literature is limited, review of the literature is advantageous to find the parameters that actually have been determined. Furthermore, a comparison of the parameters that have been estimated by various methods is helpful to get an idea of the accuracy of these methods for the compounds considered. The following references may be used as primary sources: [37], [47], [71], [72], [75].

3.3.6. Fugacity Models

The application of fugacity in environmental modeling has been discussed by Mackay and Paterson [215], [216]. Fugacity has the dimension of pressure and may be considered a measure of "escaping tendency" from the phase. When phases are at equilibrium, their fugacities are equal. Fugacity *f* is related to concentration by the following equation:

$$f = C/Z \tag{104}$$

where *C* is the concentration (mol/m^3) and *Z* is the fugacity capacity (mol m^{-3} Pa^{-1}). Equilibrium involves

$$C_a/C_w = Z_a/Z_w = K_p \tag{105}$$

where the indices a and w indicate air and water, respectively, as examples of the phases; K_p is a partition coefficient. Transfer rates between two phases by diffusion are expressed by the following equation (models per unit of area and time):

$$N = D \cdot \Delta f \tag{106}$$

where *N* is the rate of transfer, *D* is the diffusion coefficient, and Δf is the difference in fugacity. *D* is the total resistance for the transfer consisting of the resistances of the two phases in series. Note that *D* may be determined as $K \cdot Z$, where *K* is the transfer coefficient and *Z* is the fugacity capacity defined in Equation (104).

The so-called unit world model consists of six compartments: air, water, soil, sediment, suspended sediment, and biota. This simplified model aims at identification of the partition of toxic substances emitted to the environment among these six compartments. The conceptual diagram of the unit world model is shown in Figure 61.

The fugacity model may be used at four levels. The *first level* calculates the equilibrium distribution of a chemical between phases. It assumes that each compartment is well mixed, and no reaction or advection into or out of the system occurs.

If *M* is the total amount of a chemical in moles, m_i and V_i are the amounts of the chemical and the volumes in each compartment, then:

$$M = \sum m_i = \sum C_i \cdot V_i = \sum f_i \cdot Z_i \cdot V_i = f \cdot \sum Z_i \cdot V_i \tag{107}$$

Since *M* is known, *f* may be calculated and the amount of substance in each compartment is given by

$$m_i = f \cdot V_i \cdot Z_i \tag{108}$$

The percentage distribution is found as m_i/M and the concentrations are found by

$$C_i = Z_i \cdot f \tag{109}$$

Level two considers equilibrium but includes reaction and advection as well. Reactions comprise photolysis, hydrolysis, biodegradation, oxidation, and so on. All the processes are assumed to be first-order reactions. If the rate constants are not known or cannot be estimated, they are set to zero. Since process rates are expressed in first-order form, they are additive and the total

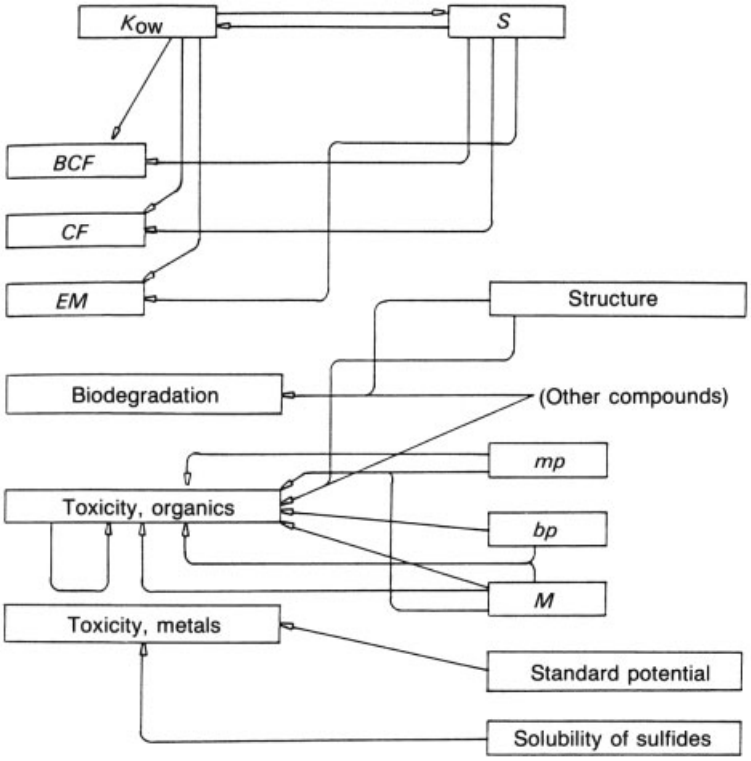

Figure 59. General structure – activity relationships of potentially toxic substances

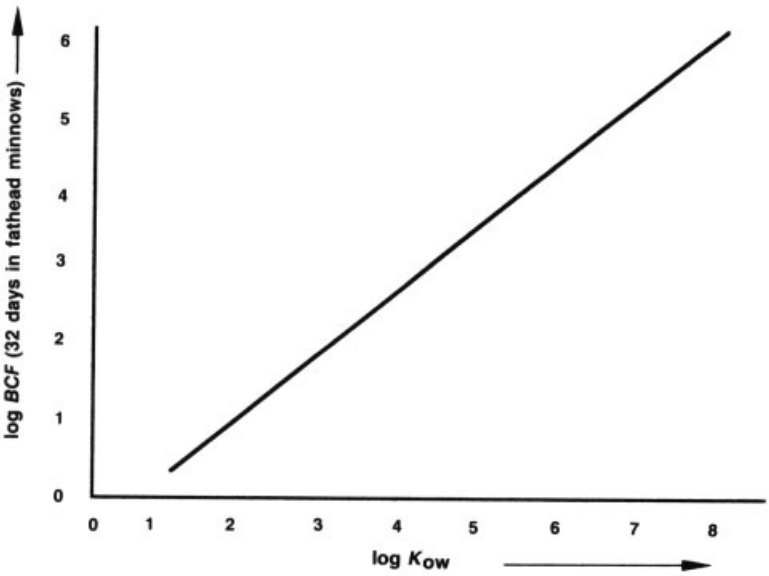

Figure 60. Relationship between log K_{OW} and log BCF, valid for fish ca. 25 cm in length

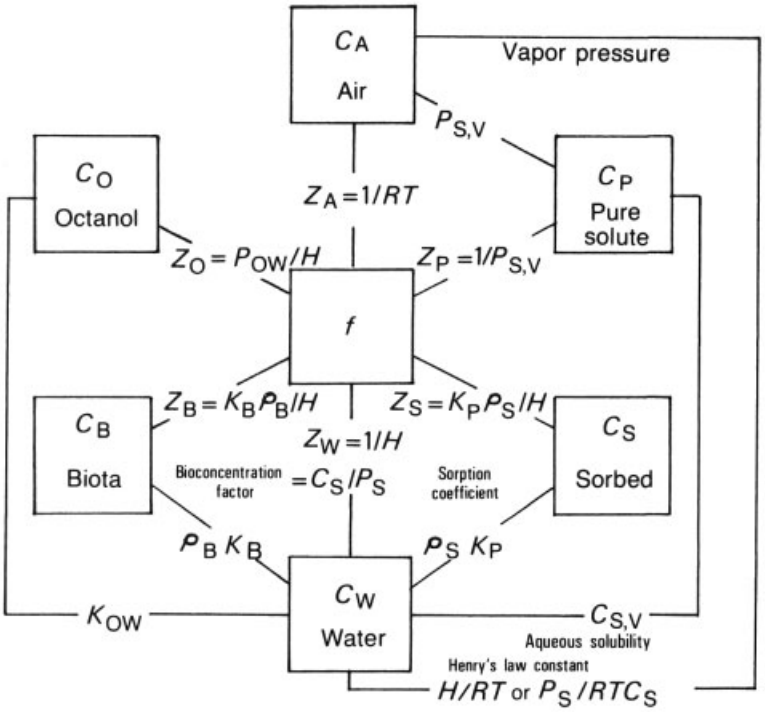

Figure 61. Conceptual diagram of fugacity models

removal rate of a chemical from a compartment is

$$\sum V_i \cdot C_i \cdot \sum k_j = \sum V_i \cdot C_i \cdot k_i \tag{110}$$

where k_j and k_i denote the rate constants of conversion of substances j and i.

As in level one, a common fugacity f prevails, and if $f \cdot Z$ is substituted for C the total rate E is

$$E = f \cdot \sum V_i \cdot Z_i \cdot k_i \tag{111}$$

The average residence time due to reactions t_r may be found by use of the following equation:

$$t_r = M/E \tag{112}$$

and the overall rate constant K is E/M or $1/t_r$.

The *third level* is devoted to a steady-state, nonequilibrium situation, which implies that the fugacities are different in each phase. Equation (105) is used to account for the transfer. The D values may be calculated from quantities such as interface areas, mass-transfer coefficients (as indicated above, D is the product of the transfer coefficient and the fugacity capacity: $D = K \cdot Z$), release rate of chemicals into phases such as biota or sediment, and Z values, or by use of the estimation methods presented in Section 3.3.5.

Level four involves a dynamic version of level three where emissions, and thus concentrations, vary with time. This implies that differential equations must be applied for each compartment to calculate the change in concentrations with time, for instance:

$$V_i \cdot \mathrm{d}C_i/\mathrm{d}t = E_i - V_i \cdot C_i \cdot k_i - \sum D_{ij} \cdot \Delta f_{ij} \tag{113}$$

This model level is similar in concept to the EXAMS model [251].

Levels one and two are sufficient in most cases, but if the environmental management problem requires the prediction of (1) the time required for a substance to accumulate to a certain concentration in a phase after emission has started, or (2) the length of time for the system to recover after the emission has ceased, the fourth level must be applied.

This approach has been used widely, and a typical example is given by MACKAY et al. [251]. It concerns the distribution of PCB between air

and water in the Great Lakes. In this case, H was 49.1 bar and the air – water distribution coefficient ($=H/RT$) 0.02. The fugacity capacity for water ($=1/H$) was found to be 0.0204 bar and the fugacity capacity for air ($=1/RT$) 0.000404 bar. The distribution coefficient of PCB between water and suspended matter in the water was estimated to be 100 000. Since the concentration of suspended matter in the Great Lakes has been found to be 2×10^{-6} on a volume basis (approximately 4 mg/L) the fraction dissolved was $1/(1+0.2)=0.833$.

The PCB concentration in water of the Great Lakes was 2 ng/L, and in the air 2 ng/m^3. The fugacity can be calculated in water and air as C/Z, and the fugacity in water was found to be 17 times higher than in air, which implies that volatilization will occur. If the transfer coefficient in water is assumed to be 10^{-5} m/s and in air 10^{-3} m/s, the volatilization rate can be calculated from the traditional two-resistance model, using the relation $D=K\cdot Z$ to find the overall diffusion coefficient, D:

$$1/D = 1/10^{-5}\cdot 0.0204\,\mathrm{bar} + 1/10^{-3}\cdot 0.000404\,\mathrm{bar} \tag{114}$$

D is 1.36×10^{-7}; N is calculated by use of Equation (106):

$$N = D\,(f_{\mathrm{w}} - f_{\mathrm{a}}) = D\left(2.8\times10^{-7}\,\mathrm{Pa} - 1.53\times10^{-8}\,\mathrm{Pa}\right) \tag{115}$$

N is found to be 35.9×10^{-15} mol m^{-2} s^{-1}.

The transfer with precipitation is negligible compared to the volatilization rate, while the washout of particles and dry deposition are important processes. If these processes are considered, the net flux to the atmosphere becomes about 75 % of the flux found above.

4. Appendix [291]

Risk Assessment Directive 93/67/EEC	Document XI/730/89 rev. 3 (Ispra)	Notes
Risk Assessment		
A generic term describing an administrative and technical process that entails some or all of the elements below	no equivalent term	much wider meaning of the term than at Ispra
Effects Assessment		
1) Hazard Identification	Hazard Identification	these two terms are synonymous
Identification of the adverse effects a substance has an inherent capacity to cause; and, where possible or appropriate, assessment of a particular effect		they describe inherent properties
2) Dose – response assessment		
Estimation of the relationship between dose (or level of exposure) and the incidence and severity of an effect		
Exposure assessment	Hazard assessment (part only)	
Determination of the emissions, pathways, and rates of movement of a substance and its transformation or deradation in order to estimate the concentrations/doses to which human populations or environmental compartments are or may be exposed	= ″environmental exposure″	this part of hazard assessment (Ispra) is synonymous with exposure assessment (draft directive)
		leads to PEC
	Hazard assessment (part only) = ″effect data with reference to the environmental compartment of concern″	this is aided by a knowledge of hazard identification (Ispra)
		eads to PEC
Effects assessment plus exposure assessment	Hazard assessment	synonymous terms
		comparison of PEC and PNEC
Risk characterization	Risk assessment (see below) plus hazard assessment	synonymous terms, i.e., PEC : PNEC ratios against a probability scale
Estimation of the incidence and severity of the adverse effects likely to occur in a human population or environmental compartment due to actual or predicted exposure to a substance and may include		
Risk estimation	Risk assessment	synonymous terms
Quantification of the likelihood of the incidence and severity of the adverse effects	estimation of the probability that a substance causes adverse effects as a result of its presence in the environment at a given concentration	
Risk reduction	Risk management	synonymous terms , e.g., special use and disposal instructions, emission control measures, restrictions on types of use, total ban of use
Measures that would enable the risks for man or the environment in connection with the marketing of the substances to be lessened	the taking of measures appropriate at least to diminish significantly the presence of a substance in the environmental compartments of concern	

5. References

General References

1. A. Leroi-Gourhan: *Le Fil du Temps,* Fayard, Paris 1983.
2. O. Fränzle: *Contaminants in Terrestrial Environments,* Springer Verlag, Berlin 1993.
3. L. Trepl: *Geschichte der Ökologie,* Athenäum, Frankfurt/Main 1987.
4. E. Haeckel: *Generelle Morphologie der Organismen,* Reimer, Berlin 1866.
5. C. J. Krebs: *Ecology,* Harper & Row, New York 1972.
6. M. Begon, J. L. Harper, C. R. Townsend: *Ecology: Individuals, Populations and Communities,* Blackwell, Boston – Melbourne 1990.
7. G. C. Butler (ed.): *Principles of Ecotoxicology,* J. Wiley, Chichester 1978.
8. H. Parlar, D. Angerhöfer: *Chemische Ökotoxikologie,* Springer Verlag, Berlin 1991.

9. H. Ellenberg: *Ökosystemforschung,* Springer Verlag, Heidelberg 1973.
10. S. L. Pimm: *Food Webs,* Chapman & Hall, London 1982.
11. R. V. O'Neill, A. R. Johnson, A. W. King "A Hierarchical Framework for the Analysis of Scale," *Landscape Ecology* **3** (1989) 193 – 205.
12. K. Dierßen: *Einführung in die Pflanzensoziologie,* Wissenschaftliche Buchgesellschaft, Darmstadt 1990.
13. W. Schröder, L. Vetter, O. Fränzle (eds.): *Neue statistische Verfahren und Modellbildung in der Geoökologie,* Vieweg, Wiesbaden 1994.
14. H. G. Gauch: *Multivariate Analysis in Community Ecology,* Cambridge University Press, Cambridge 1982.
15. W. Ziechmann: *Huminstoffe,* Verlag Chemie, Weinheim 1980.
16. P. Schachtschabel et al.: *Lehrbuch der Bodenkunde,* Enke Verlag, Stuttgart 1989.
17. K. Y. Kondratyev: *Radiation Processes in the Atmosphere,* WHO, Genever 1972.
18. M. I. Budyko: "Solar Radiation and the Use of it by Plants". Agroclimatological Methods, *Proc. Reading Symp.,* UNESCO, Paris 1968, 39 – 53.
19. H. Ellenberg, R. Meyer, J. Schauermann (eds.): *Ökosystemforschung. Ergebnisse des Solling-Projekts,* Enke Verlag, Stuttgart 1986.
20. B. J. Nebel: *Environmental Science. The Way the World Works,* Prentice Hall, Englewood Cliffs, N.J., 1990.
21. M. G. Barbour, J. H. Burk, W. D. Pitts: *Terrestrial Plant Ecology,* Benjamin/Cummings, Menlo Park, CA, 1980.
22. BUA (Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker): *Umweltrelevante Alte Stoffe III – Prioritätensetzung und eingestufte Stoffe der dritten Stoffliste,* VCH Verlagsgesellschaft, Weinheim 1992.
23. O. Hutzinger: *The Handbook of Environmental Chemistry,* 5 vols., Springer Verlag, Berlin 1989 – 1991.
24. T. Rosswall, R. R. Woodmanse, P. G. Risser (eds.): *Scales and Global Change,* J. Wiley, Chichester 1988.
25. F. W. Schaller, G. W. Bailey (eds.): *Agricultural Management of Water Quality,* Iowa State University Press, Ames, IA, 1983.
26. *FAO Yearbook: Fertilizer,* **vol. 39**. Statistical Analysis Service, Statistics Division, Food and Agriculture Organization, Rome 1989.
27. D. K. Northington, Y. R. Goodin: *The Botanical World,* Times Mirror/Mosby College Publ., St. Louis, MO, 1984.
28. M. Lieth, R. H. Whittaker: *Primary Productivity of the Biosphere,* Springer Verlag, Berlin 1975.
29. L. R. Pomeroy, J. J. Alberts: *Concepts of Ecosystem Ecology. A Comparative View,* Springer Verlag, New York 1988.
30. J. Cole, G. Lovett, S. Findlay (eds.): *Comparative Analyses of Ecosystems,* Springer Verlag, New York 1991.
31. H. Walter, S. W. Breckle: *Ökologie der Erde,* **vol. 1** , "Ökologische Grundlagen in globaler Sicht," G. Fischer Verlag, Stuttgart 1991.
32. D. Pimentel (ed.): *Handbook of Energy Utilization in Agriculture,* CRC Press, Boca Raton, FL, 1980.
33. K. E. F. Watt: *Understanding the Environment,* Allyn and Bacon, Boston 1982.
34. R. G. Wetzel: *Limnology,* Saunders College Publ., New York 1983.
35. H.-P. Blume (ed.): *Handbuch des Bodenschutzes,* Ecomed, Landsberg/Lech 1992.
36. W. J. Mitsch, J. G. Gosselink: *Wetlands,* Van Nostrand Reinhold, New York 1986.
37. S. E. Jørgensen: *Modelling in Ecotoxicology,* Elsevier, Amsterdam 1990.
38. T. H. F. Allen, T. B. Starr: *Hierarchy,* Univ. Chicago Press, Chicago 1982.
39. C. Elton: *The Ecology of Invasion by Animals and Plants,* Methuen, London 1958.
40. R. E. Ulanowicz. *Growth and Development: Ecosystems Phenomenology,* Springer Verlag, New York 1986.
41. P. B. Moyle, R. A. Leidy: "Loss of Biodiversity in Aquatic Ecosystems: Evidence from Fish Faunas," in P. S. Fielder, K. J. Subodh (eds.): *From Conservation Biology: the Theory and Practice of Nature, Conservation, Preservation and Management,* Chapman and Hall, New York 1992, pp. 127 – 169.
42. J. C. White (ed.): *Global Climate Change Linkages,* Elsevier, Amsterdam 1967.
43. A. K. Ward: "Geological Mediation of Stream Flow and Sediment and Solute Loading to Stream Ecosystems due to Climate Change," in P. Firth, S. G. Fischer (eds.): *Global Change and Freshwater Ecosystems,* Springer Verlag, Berlin 1991.
44. F. H. Borman, G. E. Likens: *Pattern and Process in a Forested Ecosystem,* Springer Verlag, New York 1979.

45. C. E. W. Steinberg, R. F. Wright (eds.): *Acidification of Freshwater Ecosystems. Implications for the Future,* J. Wiley, Chichester 1994.
46. R. J. Chorley, S. A. Schumm, D. E. Sudgen: *Geomorphology,* Methuen, London 1984.
47. K. Verschueren: *Handbook of Environmental Data on Organic Chemicals,* Van Nostrand Reinhold, New York 1983.
48. G. Mattheß: *Die Beschaffenheit des Grundwassers,* Borntraeger, Berlin – Stuttgart 1994.
49. H. Freundlich: *Kapillarchemie,* Akad. Verlagsgesellschaft, Leipzig 1930.
50. J. N. Israelachvili: *Intermolecular and Surface Forces. With Applications to Colloidal and Biological Systems,* Academic Press, London 1985.
51. S. L. Faust, J. V. Hunter (eds.): *Kinetics of Biologically Mediated Oxidation of Organic Compounds in Aquatic Environments,* Marcel Dekker, New York 1978.
52. J. W. Moore, S. Ramamoorthy: *Organic Chemicals in Natural Waters,* Springer Verlag, Berlin 1984.
53. F. Korte (ed.): *Lehrbuch der ökologischen Chemie,* Thieme Verlag, Stuttgart 1987.
54. R. Geiger: *Das Klima der bodennahen Luftschicht,* Vieweg & Sohn, Braunschweig 1961.
55. B. Ulrich (ed.): *Internationaler Kongreß Waldschadensforschung: Wissensstand und Perspektiven,* Kernforschungszentrum Karlsruhe, Karlsruhe 1990.
56. R. Dallinger, P. S. Rainbow: *Ecotoxicology of Metals in Invertebrates,* Lewis Publ., Boca Raton, FL, 1992.
57. J. A. Foran, L. E. Fink: *Regulating Toxic Substances in Surface Waters,* Lewis Publ., Boca Raton, FL 1993.
58. P. H. Howard: *Handbook of Environmental Fate and Exposure Data for Organic Chemicals,* **vols. II – III**, CRC Press, Boca Raton, FL, 1991.
59. K. C. Jones: *Organic Contaminants in the Environment,* Elsevier Applied Science, New York 1991.
60. R. A. Lewis: *Environmental Toxicology,* CRC Press, Boca Raton, FL, 1991.
61. M. Mansour (ed.): *Fate and Prediction of Environmental Chemicals in Soils, Plants and Aquatic Systems,* Lewis Publ., Boca Raton, FL, 1993.
62. S. Matsui, B. F. D. Barrett, J. Banerjee, (eds.): *Guidelines of Lake Management,* vol. 4, "Toxic Substances Management in Lakes and Reservoirs," International Lake Environment Committee Foundation, Otsu, Japan 1991.
63. J. F. McCarthy, L. R. Shugart: *Biomarkers of Environmental Contamination,* CRC Press, Boca Raton, FL, 1990.
64. R. G. Tardiff (ed.): "Methods to Assess the Impact of Pesticides on Non-target Organisms," SCOPE **49** (1992) IPCS Joint Symposia 16, SGOMSEC 7.
65. A. M. V. M. Soares, P. Calow (eds.): *Standardization of Aquatic Toxicity Tests,* Lewis Publ., Boca Raton, FL, 1993.
66. T. Colborn, C. Clement: *Chemically-induced Alterations in Sexual and Functional Development: The Wildlife/Human Connection. Advances in Modern Environmental Toxicology,* **vol. XXI**, Princeton Scientific Publ. Co., Princeton, N.J., 1993.
67. P. Bourdeau, J. A. Haines, W. Klein, C. R. Krishna Murti (eds.): "Ecotoxicology and Climate with Special Reference to Hot and Cold Climates," SCOPE **38** (1989).
68. S. M. Bartell: *Toxicological Risk Assessment in Aquatic Ecosystems,* CRC Press, Boca Raton, FL, 1991.
69. M. J. Gromiec, E. F. Gloyna: "Radioactivity Transport in Water," Final Report no. 22 to U.S. Atomic Energy Commission, Contract AT (11-1)-490, 1973.
70. P. Kauppi et al.: "A Model for Predicting the Acidification of Forest Soils: Application to Acid Deposition in Europe," IIASA Research Report 1984.
71. S. E. Jørgensen, S. Nors-Nielsen, L. A. Jørgensen: *Handbook of Ecological Parameters and Ecotoxicology,* Elsevier, Amsterdam 1991.
72. S. E. Jørgensen (ed.): *Modelling in Environmental Chemistry,* Elsevier, Amsterdam 1991.
73. S. E. Jørgensen, M. Gromiec: *Mathematical Submodels of Water Quality Systems,* Elsevier, Amsterdam 1989.
74. S. E. Jørgensen, I. Johnsen: "Principles of Environmental Science and Technology," *Stud. Environ. Sci.* **33** (1989).
75. Y. Samiullah: *Prediction of the Environmental Fate of Chemicals,* Elsevier Applied Science, London 1990. U. Schlottmann (ed.): *Prüfmethoden für Chemikalien,* S. Hirzel Verlag, Stuttgart 1994.

76. BUA (Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker): *Umweltrelevante Alte Stoffe – Auswahlkriterien und Stoffliste,* VCH Verlagsgesellschaft, Weinheim 1986.
77. BUA (Beratergremium für umweltrelevante Altstoffe der Gesellschaft Deutscher Chemiker): *Umweltrelevante Alte Stoffe II – Auswahlkriterien und Zweite Stoffliste,* VCH Verlagsgesellschaft, Weinheim 1988.
78. H. Bick: *Ökologie,* G. Fischer Verlag, Stuttgart 1989.
79. M. Straškraba, A. Gnauck: *Freshwater Ecosystems. Modelling and Simulation,* Elsevier, Amsterdam 1985.
80. J. Gilbert, M.-J. Dole-Olivier, P. Marmonier, P. Vervier in R. J. Naiman, H. Décamps (eds.): *The Ecology and Management of Aquatic-Terrestrial Ecotones,* UNESCO, Paris; Parthenon, Park Ridge, 1990, pp. 199 – 225.

Specific References

81. L. G. Freeman, *Am. Anthr.* **38** (1973) 3 – 44.
82. J. Seymour, H. Girardet: *Fern vom Garten Eden,* Fischer Verlag, Frankfurt/Main 1985.
83. E. H. Blair, *Proc. of the Workshop on the Control of Existing Chemicals under the Patronage of the Organisation for Economic Co-operation and Development,* Umweltbundesamt, Berlin 1981, pp. 252 – 260.
84. ECETOC: "Concentrations of Industrial Organic Chemicals measured in the Environment," Technical Report no. **29,** Brussels 1988.
85. Leitungsgremium des Projektzentrums "Ökosystemforschung im Bereich der Bornhöveder Seenkette" (ed.): "Arbeitsbericht 1988 – 91," *EcoSys,* **vol. 1,** Kiel 1992.
86. I. H. Koeman, *Ecotoxicol. Environ. Saf.* **6** (1982) 358 – 362.
87. W. Takken, F. Balk, R. C. Jansen, J. H. Koeman, PANS **24** (1978) 455 – 466.
88. Umweltbundesamt: Collection of Minimum Premarketing Sets of Data Including Environmental Residue Data on Existing Chemicals, Umweltbundesamt, Berlin 1982.
89. A. G. Tansley, *Ecology* **16** (1935) 284 – 307.
90. L. Finke: *Landschaftsökologie,* Westermann, Höller und Zwick, Braunschweig 1986.
91. C. Elton: *Animal Ecology,* Macmillan, New York 1927.
92. R. L. Lindeman, *Ecology* **23** (1942) 399 – 413.
93. O. Fränzle in B. Markert (ed.): *Environmental Sampling for Trace Analysis,* VCH Verlagsgesellschaft, Weinheim 1994, pp. 305 – 320.
94. G. G. Simpson: *The Major Features of Evolution,* Columbia University Press, New York 1953.
95. R. P. McIntosh, *Ecology* **48** (1967) 392 – 404.
96. C. E. Shannon, *Bell Syst. Tech. J.* **27** (1948) 379 – 423, 623 – 653.
97. O. Fränzle, W. F. Killisch, *Kieler Geographische Schriften* **50** (1979) 211 – 245.
98. L. Vetter: "Evaluierung und Entwicklung statistischer Verfahren zur Auswahl von repräsentativen Untersuchungsobjekten für ökotoxikologische Problemstellungen," Dissertation, Universität Kiel 1989.
99. R. M. G. Jongman, C. F. R. Ter Braak, O. F. R. Van Tongeren: "Data Analysis in Community and Landscape Ecology," PUDOC, Wageningen 1987.
100. K. R. Gabriel, *Biometrika* **58** (1971) 453 – 467.
101. O. Fränzle, W. F. Killisch in W. Schröder, L. Vetter, O. Fränzle (eds.): *Neuere statistische Verfahren und Modellbildung in der Geoökologie,* Vieweg & Sohn, Braunschweig 1994, pp. 129 – 143, 255 – 281.
102. O. Fränzle et al.: "Synoptische Darstellung möglicher Ursachen der Waldschäden: Untersuchungen zu Langzeitwirkungen von Kompensationskalkungen auf Buchen-, Fichten- und Kiefernforst-ökosysteme," Forschungsbericht 108 03 046/13 im Umweltforschungsplan des BMU, Kiel 1989.
103. W. Schröder: "Ökosystemare und statistische Untersuchungen zu Waldschäden in Nordrhein-Westfalen: Methodenkritische Ansätze zur Operationalisierung einer wissenschaftstheoretisch begründeten Konzeption," Dissertation, Universität Kiel 1989.
104. J. H. R. Gee, P. S. Giller (eds.): *Organization of Communities: Past and Present,* Blackwell, Oxford 1987.
105. J. H. Connell, R. O. Slatyer, *Am. Nat.* **111** (1977) 1119 – 1144.
106. G. Hemprich: "Landschaftsökologische Untersuchungen im Bereich des Belauer Sees und des Schmalensees," Diplomarbeit Geographie, Universität Kiel 1991.
107. V. Volterra: Variations and Fluctuations of the Numbers of Individuals in Animal Species Living Together (1926) (reprinted in R. N. Chapman: *Animal Ecology,* McGraw Hill, New York 1931).

108. J. H. Connell, *Am. Nat.* **122** (1983) 661 – 696.
109. P. Grubb, *Biol. Rev. Cambridge Philos. Soc.* **52** (1977) 107 – 145.
110. L. R. Lawlor, *Am. Nat.* **116** (1980) 394 – 408.
111. G. Müller, E. Foerster, *Acker-Pflanzenbau* **140** (1974) 161 – 174.
112. J.-F. Dhôte, F. Houllier in J.-D. Lebreton, B. Asselain (eds.): *Biométrie et Environnement,* Masson, Paris 1993, pp. 175 – 204.
113. P. Yodzis in J. Diamond, T. J. Case (eds.): *Community Ecology,* Harper & Row, New York 1986, pp. 480 –4 91.
114. C. D. Canham, *Ecology* **69** (1988) 786 – 795.
115. D. J. Levey, *Ecology* **69** (1988) 1076 – 1089.
116. J. L. Monteith, *Symp. Soc. Exp. Biol.* **29** (1965) 205 – 234.
117. C. T. de Wit: *Simulation of Assimilation, Respiration and Transpiration of Crops,* PUDOC, Wageningen 1978.
118. C. W. Thornthwaite, *Geogr. Rev.* **38** (1948) 85 – 94.
119. L. Turc, *Ann. Agrar.* **5** (1954) 491 – 596.
120. B. L. Dzerdzeevskii in: *Climat et Microclimatologie. Actes du Colloque de Canberra,* (Arid Zone Research/Recherches sur la Zone Aride XI), UNESCO, Paris 1958, pp. 315 – 325.
121. M. Fuchs, G. Stanhill, *Isr. J. Agric. Res.* **13** (1963) 63 – 78.
122. R. O. Slatyer in: *The Use of Soil Water Balance Relationships in Agroclimatology* (Natural Resources Res. VII) UNESCO, Paris 1968, pp. 73 – 87.
123. P. Benecke, R. R. van der Ploeg, *Forstarchiv* **49** (1978) 1 – 7, 26 – 32.
124. J. v. Hoyningen-Huene, S. Nasdalack, *Landschafts-ökol. Messen Auswerten* **1** (1985) 107 – 116.
125. B. Sevruk: "Methodische Untersuchungen des systematischen Meßfehlers des Hellmann-Regenmessers im Sommerhalbjahr in der Schweiz," Dissertation, ETH Zürich 6789, 1981.
126. H. Braden, *Mitt. Dtsch. Bodenkd. Ges.* **42** (1985) 294 – 298.
127. F.-J. Löpmeier: *Verhandlungen 45. Dtsch. Geographentages,* Steiner Verlag, Stuttgart 1987, pp. 434 – 438.
128. J. R. Philip in E. S. Hills (ed.): *Water Resources, Use and Management,* Melbourne University Press, Melbourne 1964, pp. 257 – 275.
129. O. Schwarz in G. Einsele et al. (eds.): *Das landschaftsökologische Forschungsprojekt Naturpark Schönbuch,* VCH Verlagsgesellschaft, Weinheim 1986, pp. 161 – 179.
130. R. E. Smith, J. Y. Parlange, *Water Resour. Res.* **14** (1978) 533 – 538.
131. A. C. Imeson, *CATENA Suppl.* **4** (1983) 79 – 89.
132. G. Einsele, G. Agster, M. Elgner in G. Einsele (ed.): *Das landschaftsökologische Forschungsprojekt Naturpark Schönbuch,* VCH Verlagsgesellschaft, Weinheim 1986, pp. 209 – 234.
133. J. Drescher, R. Horn, M. de Boodt (eds.): "Impact of Water and External Forces on Soil Structure", *CATENA Suppl.* **11** (1988).
134. E. Petzold, *Verh. Ges. Ökol.* **12** (1984) 487 – 498.
135. M. Renger, O. Strebel, *Wasser Boden* **32** (1980) 362 – 366.
136. Arbeitskreis Grundwasserneubildung, *Geol. Jahrb. Reihe C* **C 19** (1977) 3 – 98.
137. O. Fränzle et al.: "Auswahl der Hauptforschungsräume für das Ökosystemforschungsprogramm der Bundesrepublik Deutschland," Forschungsbericht 101 04 043/02 im Umweltforschungsplan des BMU, Universität Kiel 1987.
138. H. J. Heckmann, K. F. Schreiber, R. Thöle, *Mitt. Dtsch. Bodenkd. Ges.* **41** (1985) 353 – 356.
139. H. Sager, Informationsbericht, part 1, Bayer. Landesamt für Wasserwirtschaft, München 1983, pp. 59 – 72.
140. P. Greminger: "Physikalisch-ökologische Standortuntersuchung über den Wasserhaushalt im offenen Sickersystem Boden unter Vegetation am Hang," *Mitt. Eidgen. Anst. forstl. Versuchswesen* **60** (1985) no. 2.
141. I. D. White, N. D. Mottershead, S. J. Harrison: *Environmental Systems,* Allen & Unwin, London 1984.
142. M. Schroeder, *Dtsch. Gewässerkd. Mitt.* **20** (1976) 8 – 13.
143. C. W. Thornthwaite, B. Holzman, *Tech. Bull. U.S. Dep. Agric.* **817** (1942).
144. A. I. Dyer, *Q. J. R. Meteorol. Soc.* **87** (1961) 401 – 412.
145. J. L. Monteith, G. Szeicz, *Q. J. R. Meteorol. Soc.* **86** (1960) 205 – 214.
146. S. v. Stamm: "Untersuchungen zur Primärproduktion von Corylus avellana an einem Knickstandort in Schleswig-Holstein und Erstellung eines Produktionsmodells," *EcoSys,* **suppl. vol. 4** , Kiel 1992.

147. S. Gäb, H. Parlar, F. Korte, *Chemosphere* **3** (1974) no. 5, 187 – 192.
148. M. Straškraba, *Mitt. Int. Ver. Limnol.* **14** (1968) 212 – 230.
149. J. Blake in K.-H. Zwirnmann (ed.): "Nonpoint Nitrate Pollution of Municipal Water Supply Resources: Issues of Analysis and Control," *IIASA Collab. Proc. Ser.* **CP-82-S4** (1982) 231 – 245.
150. L. Procházková, P. Blažka, personal communication, 1993.
151. L. R. Brown: *State of the World,* W. W. Norton & Co., New York 1985.
152. H. Lieth, E. Box in R. Y. Mather (ed.): *Publications in Climatology,* **vol. 25,** C. W. Thornthwaite Associates, New Jersey 1972, pp. 37 – 46.
153. S. McNaughton, M. Oesterheld, D. A. Frank, K. J. Williams in [30, pp. 120 – 139].
154. I. Aselmann, H. Lieth, *Prog. Biometeorol.* **3** (1983) 101 – 119.
155. R. J. Naiman, H. Décamps, J. Pastor, C. A. Johnston, *J. North Am. Benthological Soc.* **7** (1988) 289 – 306.
156. F. E. Clements: *Research Methods in Ecology,* University Publ. Comp., Lincoln 1905.
157. W. Kluge, O. Fränzle, *Ges. Ökol.* **21** (1992) 401 – 407.
158. D. L. Urban, R. V. O'Neill, H. H. Shugart, *BioScience* **37** (1987) 119 – 127.
159. O. Fränzle in E. Bayer, H. Behret (eds.): "Bewertung des ökologischen Gefährdungspotentials von Chemikalien," *GDCh-Monogr.* **1** (1994) 45 – 90.
160. O. Fränzle in B. Markert (ed.): *Sampling of Environmental Materials for Trace Analysis,* VCH Verlagsgesellschaft, Weinheim 1994, pp. 217 – 227.
161. W. M. Stigliani: "Changes in Valued 'Capacities' of Soil and Sediments as Indicators of Nonlinear and Time-delayed Environmental Effects," IIASA Ecoscript **35** (1988), International Institute for Applied Systems Analysis, Laxenburg.
162. R. C. Wissmar, F. J. Swanson in R. J. Naiman, H. Décamps (eds.): *The Ecology and Management of Aquatic-terrestrial Ecotones,* UNESCO, Paris; Parthenon, Park Ridge 1990, pp. 171 – 198.
163. E. Pieczinska in R. J. Naiman, H. Décamps (eds.): *The Ecology and Management of Aquatic-terrestrial Ecotones,* UNESCO, Paris; Parthenon, Park Ridge 1990, pp. 103 – 140.
164. U. Schleuß: "Böden und Bodenschaften einer norddeutschen Moränenlandschaft," *EcoSys,* **suppl. vol. 2,** Kiel 1992.
165. M. M. Holland, D. F. Whigham, B. Gopal in R. J. Naiman, H. Décamps (eds.): *The Ecology and Management of Aquatic-terrestrial Ecotones,* UNESCO, Paris; Parthenon, Park Ridge 1990, pp. 171 – 198.
166. D. Scholle: "Vegetationskundliche Untersuchungen im Raum Bornhöved und deren Auswertung mit Hilfe eines Geographischen Informationssystems (GIS)," Diplomarbeit, Universität Saarbrücken, 1991.
167. G. Schernewski: "Raumzeitliche Prozesse und Strukturen im Wasserkörper des Belauer Sees," Diplomarbeit Geographie, in *EcoSys,* **suppl. vol. 1**, Kiel 1992.
168. C. S. Holling in E. Jantsch, C. H. Waddington (eds.): *Evolution and Consciousness,* Addison-Wesley, Reading, Mass., 1976, pp. 73 – 92.
169. I. Prigogine in E. Jantsch, C. H. Waddington (eds.): *Evolution and Consciousness,* Addison-Wesley, Reading, Mass., 1976, pp. 93 – 133.
170. W. E. Westman, *BioScience* **28** (1989) 705 – 710.
171. A. Gigon, Ber. Geobot. Inst. Eidg. Techn. Hochsch., Stift. Rübel Zürich **50** (1983) 149 – 177.
172. E. van der Maarel, *Ber. Dtsch. Bot. Ges.* **89** (1976) 415 – 443.
173. O. T. Solbrig, G. Nicolis: *Perspectives in Biological Complexity,* IUBS, Paris 1991.
174. R. H. MacArthur, *Ecology* **36** (1955) 533 – 536.
175. R. M. May, *Nature (London)* **238** (1972) 413 – 414.
176. R. E. McMurtrie, *J. Theor. Biol.* **50** (1975) 1 – 11.
177. J. E. Cohen, C. M. Newman, *Ann. Probab.* **12** (1984) 283 – 310.
178. C. P. Bachas, B. A. Huberman, *J. Phys. A Math. Gen.* **20** (1987) 4995 – 5014.
179. H. A. Ceccato, B. A. Huberman, *Physica Scr.* **37** (1988) 145 – 150.
180. P. Doreian, *J. Soc. Biol. Struct.* **9** (1986) 115 – 139.
181. F. Briand, J. E. Cohen, *Science (Washington D.C. 1883)* **238** (1987) 956 – 960.
182. T. Hogg, B. A. Huberman, J. M. McGlade, *Proc. R. Soc. London B* **237** (1989) 43 – 51.
183. M. Rejmanek, P. Stary, *Nature (London)* **280** (1979) 311 – 313.
184. F. Briand, *Ecology* **64** (1983) 253 – 263.
185. S. J. McNaughton, *Am. Nat.* **111** (1977) 515 – 525.

186. A. Hay, *Nature (London)* **302** (1983) 208 – 209.
187. R. V. O'Neill, *Ecology* **57** (1976) 1244 – 1253.
188. D. L. de Angelis, *Ecology* **61** (1980) 764 – 771.
189. J. Cairns Jr., B. R. Niederlehner, *Environ. Professional* **15** (1993) 116 – 124.
190. P. van Voris, R. V. O'Neill, W. R. Emanuel, H. H. Shugart, *Ecology* **6** (1980) 1352 – 1360.
191. D. W. Schindler, *Can. J. Fish. Aquat. Sci.* **44** (1987) Suppl 1, 6 – 25.
192. M. J. Crossey, T. W. LaPoint, *Hydrobiologia* **162** (1988) 109 – 121.
193. W. D. Crumby, M.-A. Webb, F. J. Bulow, H. J. Cathey, *Trans. Am. Fish. Soc.* **119** (1990) 885 –893.
194. J. H. Rodgers Jr., K. L. Dickson, J. Cairns, Jr., in R. G. Wetzel (ed.): "Methods and Measurements of Periphyton Communities," *ASTM,* Philadelphia, PA, 1980, pp. 142 – 167.
195. D. Uhlmann, H. Mihan, A. Gnauck, *Acta Hydrochim. Hydrobiol.* **6** (1978) 421 – 444.
196. A. Gnauck: "Strukturelle und funktionelle Änderungen in aquatischen Ökosystemen," Kongreß- und Tagungsberichte, Martin-Luther-Universität Halle-Wittenberg, 1982, pp. 335 – 344.
197. J. D. Yount, G. J. Niemi: "Recovery of Lotic Communities and Ecosystems from Disturbance: Theory and Applications," *Environ. Manage. (N.Y.)* **14** (1990) no. 5, 132 – 137.
198. G. M. Woodwell, *Science (Washington, D.C. 1883)* **156** (1967) no. 3774, 461 – 470.
199. T. Rosswall, R. G. Woodmanse, P. G. Risser: *Scale and Global Change: Spatial and Temporal Variability in Biospheric and Geospheric Processes,* J. Wiley, Chichester 1988.
200. J. Peñuelas in J. D. Ross, N. Prat (eds.): "Hommage to Ramón Margalef: or Why There is Such Pleasure in Studying Nature," Universitat de Barcelona Publications, Barcelona 1991, 367 – 385.
201. S. H. Schneider, N. J. Rosenberg in N. J. Rosenberg, W. E. Easterling, P. J. Crosson, J. Darmstadter (eds.): "Greenhouse Warming: Abatement and Adaptation," Proceedings of a Workshop held in Washington D.C., June 14 – 15, 1988, Resources for the Future, Washington 1989, pp. 7 – 34.
202. J. S. Clark in R. L. Wyman (ed.): *Global Climate Change and Life on Earth,* Routledge, Chapman and Hall, New York 1991, pp. 65 – 98.
203. J. D. Meisner et al., *J. Great Lakes Res.* **13** (1987) no. 3, 340 – 352.
204. P. Glansdorff, I. Prigogine: *Thermodynamic Theory of Structure, Stability, and Fluctuations,* Wiley-Interscience, New York 1971.
205. E. J. Fittkau: "Tropische Regenwälder – Ökologische Zusammenhänge," *Bensberger Protokolle* **66** (1991) 27 – 63 (Thomas-Morus-Akademie, Bensberg).
206. R. Herrera, C. F. Jordan, E. Medina, H. Kluge, *Ambio* **10** (1981) 109 – 114.
207. H. Sioli: *Amazonien. Grundlagen der Ökologie des größten tropischen Waldlandes,* Enke Verlag, Stuttgart 1980.
208. C. F. Jordan (ed.): *An Amazonian Rain Forest: the Structure and Function of a Nutrient Stressed Ecosystem and the Impact of Slash and Burn Agriculture,* UNESCO, Paris, Parthenon, Carnforth 1989.
209. D. L. de Angelis, *Ecology* **56** (1975) 238 – 243.
210. D. L. de Angelis, J. C. Waterhouse, *Ecol. Monogr.* **57** (1987) 1 – 21.
211. H. Remmert, *Verh. Ges. Ökol.* **16** (1987) 27 – 34.
212. H. M. Wilbur, *Ecology* **68** (1987) 1437 – 1452.
213. C. R. Townsend, *J. North Am. Benthological Soc.* **8** (1989) 36 – 50.
214. N. I. Sax: *Dangerous Properties of Industrial Materials,* Van Nostrand Reinhold, New York 1975.
215. D. Mackay, S. Paterson, *Environ. Sci. Technol.* **15** (1981) 1006.
216. D. Mackay, S. Paterson, *Environ. Sci. Technol.* **16** (1982) 654 A.
217. W. F. Spencer, W. J. Farmer in R. Haque (ed.): *Dynamics, Exposure and Hazard Assessment of Toxic Chemicals,* Ann Arbor Sci., Ann Arbor 1980, pp. 143 – 161.
218. O. Fränzle, *Z. Geomorphol. N.F.* **15** (1971) 212 – 235.
219. M. C. Lee et al., *Water Res.* **13** (1979) 1249 – 1258.
220. ECETOC: "Structure-activity Relationships in Toxicology and Ecotoxicology: an Assessment," *ECETOC Monogr.* **8** (1986).
221. A. Leo, C. Hansch, D. Elkins, *Chem. Rev.* **71** (1971) 525.
222. T. Mill in Umweltbundesamt (eds.): "Proc. Workshop Control of Existing Chemicals under the Patronage of the Organisation for Economic Cooperation and Development," June 10 – 12, 1981, Umweltbundesamt, Berlin 1981, pp. 207 – 227.

223. G. Mattheß et al.: "Der Stofftransport im Grundwasser und die Wasserschutzgebietsrichtlinie W 101 – Statusbericht und Problemanalyse," *Ber. Umweltbundesamt Ger.* **7** (1985).
224. W. Mabey, T. Mill, *J. Phys. Chem. Ref. Data* **7** (1978) 383 – 415.
225. ECETOC: "Experimental Assessment of the Phototransformation of Chemicals in the Atmosphere," Technical Report no. 7, Brussels 1983.
226. P. Schachtschabel, H.-P. Blume, G. Brümmer et al.: *Lehrbuch der Bodenkunde,* Enke Verlag, Stuttgart 1989.
227. R. Brdicka: *Grundlagen der physikalischen Chemie,* Deutscher Verlag der Wissenschaften, Berlin 1971.
228. G. Sandstede, *Z. Phys. Chem. (Munich)* **29** (1961) 120 – 133.
229. M. Alexander in R. Haque (ed.): *Dynamics, Exposure and Hazard Assessment of Toxic Chemicals,* Ann Arbor Sci., Ann Arbor 1980, pp. 179 – 190.
230. J. C. G. Ottow, *Landwirtsch. Forsch.* **35** (1982) 238 – 256.
231. OECD Guidelines for Testing of Chemicals, no. 301 A – E, OECD, Paris 1981.
232. OECD Guidelines for Testing of Chemicals, no. 302 A, OECD, Paris 1981.
233. OECD Guidelines for Testing of Chemicals, no. 303 A, OECD, Paris 1981.
234. OECD Guidelines for Testing of Chemicals, no. 304 A, OECD, Paris 1981.
235. H. Bernhardt (ed.): *NTA: Studie über die aquatische Umweltverträglichtkeit von Nitrilotriacetat (NTA),* Richarz, Sankt Augustin 1984.
236. A. G. Journel, C. J. Huijbregts: *Mining Geostatistics,* Academic Press, London 1978.
237. G. Verly, M. David, A. G. Journel, A. Maréchal (eds.): "Geostatistics for Natural Resources Characterization," *NATO Adv. Sci. Inst. Ser.,* Reidel, Dordrecht 1984.
238. ECETOC: "Estimating the Environmental Concentrations of Chemicals Using Fate and Exposure Models," Technical Report no. 50, Brussels 1992.
239. ECETOC: "Environmental Hazard Assessment of Substances," Technical Report no. 51, Brussels 1993.
240. HEDSET (1992): Harmonized Electronic Data Input Set for the Council Regulation on the Evaluation and Control of Existing Substances and the OECD, Existing Chemicals Programme (05/07/92).
241. T. R. Oke: *Boundary Layer Climates,* Methuen, London 1978.
242. T. R. Oke, *J. R. Meteorol. Soc.* 1982, 1 – 24.
243. D. O. Lee, *Prog. Phys. Geogr.* **8** (1984) 1 – 31.
244. E. Jauregui, *Erdkunde* **41** (1987) 48 – 51.
245. G. M. Hidy, C. S. Burton in G. M. Hidy et al. (ed.): *The Character and Origin of Smog Aerosols,* Wiley, New York 1980, pp. 385 – 433.
246. T. Schneider, L. Grant (eds.): *Air Pollution by Nitrogen Oxides,* Elsevier Sci., Amsterdam 1982.
247. B. Henderson-Sellers: *Modeling of Plume Rise and Dispersion – The University of Salford Model: U.S.P.R.,* Springer Verlag, Heidelberg 1987.
248. J. Pankrath, *Staub Reinhalt. Luft* **47** (1987) 239 – 244.
249. W. Klug in [255, pp. 243 – 248].
250. A. Eliassen, Ø. Hov, I. S. A. Isaksen, J. Saltbones in [255, pp. 347 – 356].
251. D. Mackay, S. Paterson, W. Y. Shiu, *Chemosphere* **24** (1992) 695 – 717.
252. E. W. Reiche: "Entwicklung, Validierung und Anwendung eines Modellsystems zur Beschreibung und flächenhaften Bilanzierung der Wasser- und Stickstoffdynamik in Böden", *Kieler Geogr. Schr.* **79** (1991) 1 – 150.
253. TA Luft, Aug. 28, 1974, GMBl I, pp. 426, 525, GMBl 1983, p. 98.
254. BUA (Beratergremium für umweltrelevante Altstoffe): "OH-Radikale in der Troposphäre: Konzentration und Auswirkung," BUA-Stoffbericht 100, Hirzel Verlag, Stuttgart 1993.
255. OECD: "Compendium of Environmental Exposure Assessment Methods for Chemicals," *Environ. Monogr.* **27** (1989).
256. OECD, Workshop on the Application of Simple Models for Environmental Exposure Assessment, Berlin, Sep. 11 – 13, 1991.
257. S. Wiegmann: "Die geogenen Schwermetallgehalte der Böden Schleswig-Holsteins," Diplomarbeit Geographie, Universität Kiel 1994.
258. M. Bonazountas in H. Barth, P. L'Hermite (eds.): *Scientific Basis for Soil Protection in The European Community,* Elsevier, London 1987, pp. 487 – 566.
259. DVWK (Deutscher Verband für Wasserwirtschaft und Kulturbau e.V.): *Abschätzen des Verhaltens organischer Chemikalien in Böden,* DVWK, Bonn 1989.

260. ECETOC: "Hazard Assessment of Floating Chemicals after Accidental Spill at Sea," Technical Report no. 35, Brussels 1990.
261. D. G. Farmer, M. J. Rycroft (eds.): *Computer Modelling in the Environmental Sciences,* Clarendon Press, Oxford 1991.
262. A. Dahmke: "Lösungskinetik von feldspatreichen Gesteinen und deren Bezug zu Verwitterung und Porenwasser-Chemie natürlicher Sander-Sedimente." *Berichte – Reports Geol. Paläont.-Inst. Universität Kiel* **20** (1988).
263. A. Petersen: "Laboruntersuchungen zum Einfluß organischer Komplexbildner auf die Kinetik der Feldspatverwitterung," *Berichte – Reports Geol. Paläont.-Inst. Universität Kiel* **21** (1988).
264. D. Boening: "Untersuchungen zur Modellierung des Stofftransportes in einem durch Deponiesickerwasser verunreinigten Grundwassergerinne," *Berichte – Reports Geol.-Paläont. Inst. Universität Kiel* **30** (1989).
265. R. D. Lindberg, D. D. Runnels, *Science* (Washington, D.C. 1883) 425 (1984) 925.
266. D. Langmuir in E. A. Jenne (ed.): "Chemical Modeling in Aqueous Systems," *ACS Symp. Ser.* **93** (1979).
267. D. J. Kirkner, T. L. Theis, A. A. Jennings, *Adv. Water Resour.* **7** (1984) 120.
268. P. G. Daniele, A. de Robertis, C. de Stefano, S. Sammarfano, *J. Chem. Soc. Dalton Trans.* 1985, 2353 – 2361.
269. R. Katlein: "Bestimmung von Komplexkapazitäten mit der Mangandioxid-Methode," Dissertation, Universität Marburg 1986.
270. W. Stumm, R. Kummert, L. Sigg, *Croat. Chem. Acta* **48** (1980) 291 – 312.
271. R. V. James, J. Rubin in E. A. Jenne (ed.): "Chemical Modeling in Aqueous Systems," *ACS Symp. Ser.* **93** (1979).
272. H. J. Hueck, D. M. M. Adema, W. C. de Kock, J. Kuiper, *Ber. Umweltbundesamt Ger.* **10** (1978) 159 – 167.
273. H. Biek, D. Neumann (eds.): "Bioindikatoren," *Decheniana Beih.* **26** (1982).
274. L. Steubing, H. J. Jäger (eds.): *Monitoring of Air Pollutants by Plants. Methods and Problems,* Junk, The Hague 1982.
275. OECD Guideline for Testing of Chemicals, no. 305 E, OECD, Paris 1981.
276. W. Ernst, *Decheniana Beih.* **26** (1982) 55 – 66.
277. E. A. Nusch, *Decheniana Beih.* **26** (1982) 87 – 98.
278. B. Hamburger, *Decheniana Beih.* **26** (1982) 78 – 81.
279. W. K. Besch, *Decheniana Beih.* **26** (1982) 67 – 77.
280. C. C. Brown in G. C. Butler (ed.): *Principles of Ecotoxicology,* J. Wiley, Chichester 1978, pp. 115 – 148.
281. P. Gerike, in [283, pp. 139 – 143].
282. R. N. Sturm, *J. Am Oil. Chem. Soc.* **50** (1973) 159 – 167.
283. R. Zahn, M. Wellens, *Chem. Ztg.* **98** (1974) 228 –232.
284. Umweltbundesamt: "Tests for the Ecological Effects of Chemicals," *Ber. Umweltbundesamt Ger.* **10** (1978).
285. J. S. Whiteside, M. Alexander, *Weeds* **8** (1960) 204 – 213.
286. A. Krotzky et al., *Z. Pflanzenernähr. Bodenkd.* **146** (1983) 634 – 642.
287. OECD: Data Interpretation Guides for Initial Hazard Assessment of Chemicals, Provisional OECD, Paris 1984.
288. P. Friesel, P. D. Hansen, R. Kühn, J. Trénel: "Überprüfung der Durchführbarkeit von Prüfungsvorschriften und der Aussagekraft der Stufe 1 und 2 des Chemikaliengesetzes," part VI, Umweltforschungsplan des Bundesministers des Innern, F+E-Vorhaben 106 04 001/08 (1984).
289. European Communities (EC): Technical Guidance Documents in Support of the Risk Assessment Directive (933/67/EEC) for New Substances Notified in Accordance with the Requirements of Council Directive 67/548/EEC, Brussels 1993.
290. Umweltbundesamt: "Technical Guidance on Environmental Risk Assessment of Existing Chemicals in Accordance with the Requirements of Council Regulation (EEC)" no. 793/93, draft, Berlin 1994.
291. ECETOC: "Aquatic Toxicity Evaluation," Technical Report no. 56, Brussels 1993.
292. OECD: "Report of the OECD Workshop on the Extrapolation of Laboratory Aquatic Toxicity Data on the Real Environment," *Environ. Monogr.* **59** (1992).
293. N. M. van Straalen, C. A. J. Denneman, *Ecotoxicol. Environ. Saf.* **18** (1989) 241 – 251.
294. H. Lokke in M. H. Donkers, H. Eijsackers, F. Heimbach (eds.): "Ecotoxicology of Soil Organisms," *SETAC Special Publication Ser.,* Lewis, Michigan, 1994, pp. 425 – 441.
295. T. L. Forbes, V. E. Forbes, *Funct. Ecology* **7** (1993) 249 – 254.

296. H. J. B. Emans et al., *Environ. Toxicol. Chem.* **12** (1993) 2139 – 2154.
297. T. Aldenberg, W. Slob, *Ecotoxicol. Environ. Saf.* **25** (1993) 48 – 63.
298. C. Wagner, H. Lokke, *Water Res.* **25** (1991) 1237 – 1242.
299. O. Fränzle et al.: "Grundlagen zur Bewertung der Belastung und Belastbarkeit von Böden als Teilen von Ökosystemen," Texte 59/93 des Umweltbundesamtes, Berlin 1993.
300. C. A. M. van Gestel in H. Becker et al. (eds.): *Ecotoxicology of Earthworms,* Intercept, Andover 1992, pp. 45 – 54.
301. OECD Guideline for Testing of Chemicals, no. 207, OECD, Paris 1984.
302. OECD Guideline for Testing of Chemicals, no. 208, OECD, Paris 1984.
303. NEN (Nederlandse Norm): Soil-determination of the Influence of Chemicals on Soil Nitrification, no. 5795, Nederlands Normalisatie Instituut, Delft 1988.
304. BBA (Biologische Bundesanstalt für Land- und Forstwirtschaft): "Bestimmung der Reproduktionsleistung von *Folsomia candida* (WILLEM) in künstlichem Boden," (Draft), Braunschweig 1990.
305. ISO (International Organisation for Standardisation): Soil Quality Effects of Pollutants on Earthworms (Eisenia foetida) – part 2: Method for the Determination of Effects on Reproduction. Draft Int. Standard 1993.
306. L. Moreth, E. Naton: Richtlinie zur Prüfung der Nebenwirkung von Pflanzenschutzmitteln auf *Aleochara bilineata* Gyll. (Col., Staphylinidae). Unveröff. Richtlinienvorschlag. 1992.
307. J. Römbke: "Entwicklung eines Reproduktionstests an Bodenorganismen – Enchytraen," UBA F+E-Vorhaben no. 106 03 051/01. Batelle-Institut, Frankfurt/Main 1991.
308. N. M. van Straalen, C. A. M. van Gestel: "Ecotoxicological Terrestrial Test Methods Using Terrestrial Arthropods," Detailed Review Paper for the OECD Test Guideline Programme, Amsterdam 1992.
309. A. Stavola: Detailed Review Paper on Terrestrial Ecotoxicology Test Guidelines, OECD Updating Programme, Periodical Review. OECD, Paris 1990.
310. F. Pedersen, L. Samsøe-Petersen: Discussion Paper Regarding Guidance for Terrestrial Effects Assessment (Draft), Water Quality Institute, Horsholm, Denmark, 1993.
311. H. Blanck, *Ecol. Bull.* **36** (1984) 107 – 119.
312. C. A. J. Denneman, C. A. M. van Gestel: "Bodemverontreiniging en bodemecosystemen: voorstel voor C-(toetsings)waarden op basis van ecotoxicologische risico's," Rijksinstituut voor Volksgezondheid en Milieuhygiene, Bilthoven 1990.
313. EPA (Environmental Protection Agency): "Pesticide Assessment Guidelines," Office of Pesticides and Toxic Substances, Washington D.C., 1982.
314. S. A. L. M. Kooijman, *Water Res.* **21** (1987) no. 3, 269 – 276.
315. W. Sloof, J. A. M. v. Oers, D. Zwart, *Environ. Toxicol. Chem.* **5** (1986) 841 – 852.
316. C. E. Stephan et al.: "Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their Uses," U.S. Environmental Protection Agency, Washington 1985.
317. K. Terytze: Qualitative and quantitative Aspekte der Bewertung von Böden. Manuskript 1994.
318. R. L. Metcalf, *Environ. Qual. Saf.* **5** (1976) 141 – 151.
319. C. Frehe, *Verh. Ges. Ökol.* **12** (1984) 507 – 510.
320. H. Kayser, *Helgol. Wiss. Meeresunters.* **30** (1977) 682 – 696.
321. A. M. Albert in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen," Endberichte der geförderten Vorhaben, Teil 2, Jül.-Spez. 503, A 1 – A 116, Jülich 1989.
322. G. Weidemann, K. Mathes, H. Koehler in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen," Jül-Spez. 439, 7 – 223, 1988.
323. L. Beck, K. Dumpert, U. Franke, W. Schönbor in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen," Jül-Spez. 439, 548 – 702, Jülich 1988.
324. T. Basedow, H. Rzehak, W. Liedtke in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen," Jül-Spez. 439, 223 – 369, Jülich 1988.
325. G. Kneitz, W. J. Kloft in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der

Belastbarkeit von Ökosystemen," Jül.-Spez. 439, 702 – 907, Jülich 1988.
326. G. Weigmann et al.: "Wirkung und Verbleib von PCP-Na in einem Brachland-Ökosystem," Abschluß-bericht 037 288, 037 289, 037 294 eines im Programm: "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen" geförderten Verbundvorhabens, 1986.
327. M. Runge et al. in B. Scheele, M. Verfondern (eds.): "Auffindung von Indikatoren zur prospektiven Bewertung der Belastbarkeit von Ökosystemen," Endberichte der geförderten Vorhaben, Teil 2, Jül.-Spez. 503, B 1 – B 218, Jülich 1989.
328. R. Guderian (ed.): *Air Pollution by Photochemical Oxidants,* Springer Verlag, Berlin – Heideberg – New York – Tokyo 1985.
329. L. Steubing, A. Fangmeier, *Environ. Pollut. Ser. A* **44** (1987) 297 – 306.
330. L. W. Kress, J. M. Skelly, *Plant Dis.* **66** (1982) 1149 – 1152.
331. H.-J. Jäger, E. Schulze: "Critical Levels for Effects of SO2," ECE Critical Levels Workshop, Bad Harzburg, March 14 – 18, 1988, Draft Report part II, pp. 15 – 50, UN Econ. Comm. Europe, Geneva 1988.
332. R. Guderian: "Critical Levels for Effects of NOx," ECE Critical Levels Workshop, Bad Harzburg, March 14 – 18, 1988. Draft Report part II, pp. 79 – 104, UN Econ. Comm. Europe, Geneva 1988.
333. R. Guderian: "Critical Levels for Effects of Ozone (O3)," ECE Critical Levels Workshop, Bad Harzburg, March 14 – 18, Draft Report part II, pp. 51 – 78, UN Econ. Comm. Europe, Geneva 1988.
334. D. C. Robinson, A. R. Wellburn, *Environ. Pollut. Ser. A* **32** (1983) 109 – 120.
335. E. A. Wright, P. W. Lucas, D. A. Cottam, T. A. Mansfield, *CEC Air Pollut. Res. Rep.* **4** (1986) 187 –200.
336. J. Mooi, *Forst. Holzwirt* **39** (1984) 438 – 444.
337. R. Guderian, K. Küppers, R. Six, *VDI-Ber.* **560** (1985) 657 – 701.
338. A. R. Wellburn, C. Higginson, D. Robinson, C. Walmsley, *New Phytol.* **88** (1981) 223 – 237.
339. J. A. Lee, M. C. Press, S. J. Woodin in CEC: "Study on the Need for a NO2 Long-term Limit Value for the Protection of Terrestrial and Aquatic Ecosystems," Final Report, EUR 10 546 EN, 1985.
340. A. C. Posthumus: "Critical Levels for Effects of NH_3 and NH_{+4}," ECE Critical Levels Workshop, Bad Harzburg, March 1988, pp. 117 – 127, UN Econ. Comm. Europe, Geneva 1988.
341. J. Bauch, P. Klein, A. Frühwald, H. Brill, *Eur. J. For. Pathol.* **6** (1979) 321 – 331.
342. S. Athari, *Mitt. Forstl. Bundesversuchsanst. Wien* **39** (1981) 7 – 27.
343. P. Nogler, *Allg. Forst Z.* **36** (1981) 709 – 711.
344. S. Athari, H. Kramer, *Forst. Holzwirt.* **38** (1983) 204 – 206.
345. G. K. Kenk, *Forst. Holzwirt* **39** (1984) 435 – 438.
346. Bundesministerium für Ernährung, Landwirtschaft und Forsten: Waldschadenserhebung 1985, Bonn 1985.
347. Bundesminister für Forschung und Technologie: "Umweltforschung zu Waldschäden," 2nd report, Bonn 1985.
348. O. Fränzle, W. Schröder, L. Vetter: "Saure Niederschläge als Belastungsfaktoren: Synoptische Darstellung möglicher Ursachen des Waldsterbens," Umweltforschungsplan des Bundesministers des Innern, Forschungsbericht 106 07 046/13, Berlin 1985.
349. Umweltbundesamt (eds.): "Wissenschaftliches Symposium zum Thema Waldschäden 'Neue Ursachenhypothesen'," *Schriftenr. "Texte" Umweltbundesamtes* **19** (1986).
350. Umweltbundesamt (eds.): "IMA-Querschnittsseminar zur Waldschadensforschung 'Belastung und Schäden auf Ökosystemebene und ihre Folgen'," *Schriftenr. "Texte" Umweltbundesamtes* **18** (1986).
351. H. H. Cramer in [348, pp. 308 – 312, 423 – 426].
352. H. Faulstich in [348, pp. 22 – 27, 381 – 382].
353. G. Rippen et al., *Environ. Technol. Lett.* **8** (1987) 475 – 482.
354. B. Ulrich, *Umschau* **11** (1984) 348 – 355.
355. B. Prinz, *Umschau* **18** (1984) 544 – 549.
356. E. F. Elstner, W. Osswald, R. J. Youngman, *Experientia* **41** (1985) 591 – 597.
357. A. Hüttermann, *Allg. Forst Jagdztg.* **4** (1978) 67 – 70.
358. U. Arndt, M. Kaufmann, *Allg. Forst Jagdztg.* **156** (1985) 16 – 20.
359. P. Benner, A. Wild, *J. Plant Physiol.* **129** (1987) 59 –72.
360. D. Murach, *Göttinger Bodenkd. Ber.* **77** (1984).
361. K. E. Rehfuess, C. Bosch, E. Pfankuch: "Nutrient Imbalances in Coniferous Stands in Southern Germany," Int. Workshop on Growth Disturbances of Forest Trees, Oct. 1982, IUFRO/FFRJ-Jyvaskyla, Finland.

362. B. Ulrich, *Forstwiss. Centralbl.* **100** (1981) 228 – 236.
363. B. Ulrich, *Allg. Forst Z.* **38** (1983) nos. 26/27, 670 –677.
364. W. U. Kriebitzsch: "Stickstoffnachlieferung in sauren Waldböden Nordwest-Deutschlands," *Scr. Geobot.* (1978).
365. R. Linthurst, *VDI Ber.* **500** (1983) 175 – 185.
366. L. S. Evans et al., *Water Air Soil Pollut.* **16** (1981) 469 – 509.
367. W. Schröder, O. Fränzle, L. Vetter, *Allg. Forst Z.* **41** (1986) no. 22, 543 – 544.
368. W. Saager: "Multivariate Bestimmung waldschadensdisponierender Faktoren in Schleswig-Holstein auf der Grundlage der terrestrischen Waldschadensinventur," Diplomarbeit Geographie, Universität Kiel 1990.
369. G. H. Kohlmaier et al., *Ecol. Modell.* **22** (1984) 45 – 65.
370. H. Bossel, W. Metzler, H. Schäfer (eds.): *Dynamik des Waldsterbens,* Springer Verlag, Berlin – Heidelberg – New York 1985.
371. G. H. Kohlmaier, M. Plöchl in [349, pp. 41 – 54].
372. H. Bossel, *Ecol. Modell.* **34** (1986) 259 – 288.
373. D. L. Godbold, A. Hüttermann, *Water Air Soil Pollut.* **31** (1986) 509 – 515.
374. R. Schultz: "Vergleichende Betrachtung des Schwermetallhaushalts verschiedener Waldökosysteme Norddeutschlands," Berichte des Forschungszentrums Waldökosysteme/Waldsterben, Reihe A **32** (1987).
375. W.-D. Grossmann, J. Schaller, M. Sittard, *Allg. Forst Z.* **39** (1984) no. 38, 837 – 843.
376. B. Prinz, G. H. M. Krause, H. Stratmann, *LIS Ber.* **28** (1982).
377. S. M. Adams et al., *Mar. Environ. Res.* **28** (1989) 459 – 464.
378. S. Gerlach: *Marine Pollution: Diagnosis and Therapy,* Springer Verlag, Berlin 1981.
379. A. Moubadan, J.-C. Pihan, *Hydroécol. Appl.* **5** (1993) 97 – 109.
380. J. Cairns, Jr., B. R. Niederlehner in J. Cairns, Jr., B. R. Niederlehner, D. R. Orvos (eds.): *Advances in Modern Environmental Toxicology,* **vol. XX**: "Predicting Ecosystem Risk", Princeton Scientific Publ. Co., Princeton, New Jersey, 1992, pp. 327 – 343.
381. P. H. Pritchard in R. A. Conway (ed.): *Environmental Risk Analysis for Chemicals,* Van Nostrand Reinhold, New York 1991, pp. 257 – 353.
382. P. H. Pritchard, H. W. Bourquin in H. H. White (ed.): *Concepts in Marine Pollution Measurements,* University of Maryland, 1984, pp. 117 – 138.
383. R. W. Howarth in J. Cole, G. Lowett, S. Findlay (eds.): *Comparative Analyses of Ecosystems,* Springer Verlag, New York 1991, pp. 169 – 195.
384. J. Cairns, Jr., J. R. Pratt: "Multispecies Toxicity Testing Using Indigenous Organisms – A New, Cost effective Approach to Ecosystem Protection," *Environmental Conference, TAPPI Proceedings,* TAPPI Press, Atlanta, Georgia, 1985, pp. 149 – 153.
385. W. B. Neely: *Chemicals in the Environment,* Marcel Dekker, New York 1980.
386. D. I. Swackhamer, S. J. Eisenreich in K. C. Jones (ed.): *Organic Contaminants in the Environment,* Elsevier Applied Sci., London 1991, pp. 33 – 86.
387. S. Griesbach, R. H. Peters, S. Youakim, *Can. J. Fish. Aquat. Sci.* **39** (1982) 727 – 735.
388. K. J. Macek, S. R. Petrocelli, B. H. Sleight in L. L. Marking, R. A. Kimerle (eds.): "Aquatic Toxicology," *Am. Soc. Test. Mater.* (1979) 251 – 268.
389. B. C. Patten in M. Higashi, T. P. Burns (eds.): *Theoretical Studies in Ecosystems: The Network Perspective,* Cambridge University Press, Cambridge 1991, pp. 288 – 351.
390. J. Roelof van der Meer, W. M. DeVos, S. Harayama, A. J. B. Zehnde, *Microbiol. Rev.* **56** (1992) 677 – 694.
391. R. Reuther, *Sci. Total Environ.* **115** (1992) no. 3, 219 – 237.
392. B. van Hattum, K. R. Timmermans, H. A. Govers, *Environ. Toxicol. Chem.* **10** (1991) 275 – 292.
393. R. R. Imhoff, P. Koppe, E. A. Nusch in S. Matsui, B. F. D. Barrett, J. Banerjee (eds.): "Toxic Substances Management in Lakes and Reservoirs," *Guidelines of Lake Management,* **vol. 4,** International Lake Environment Committee, UNEP, Shiga, Japan, 1991, pp. 127 – 159.
394. J. Plotner, R. Gunther, *Int. Rev. Gesamten Hydrobiol.* **72** (1987) no. 6, 759 – 771.
395. N. Nirmalakhandan et al., *Water Res.* **28** (1994) 543 – 552.
396. A. W. Maki, M. W. Slimak in W. Grodzinski, E. B. Cowling, A. I. Breymeyer (eds.): *Ecological Risks – Perspectives from Poland and the United States,* National Academy Press, Washington, D.C., 1990, pp. 77 – 87.

397. J. Cairns, J. R. Pratt, *Water Sci. Technol.* **19** (1987) no. 11, 1 – 12.
398. C. N. Haas, *Water Qual. Int.* **4** (1993) 30 – 32.
399. F.-A. A. M. de Leeuw, *Chemosphere* **27** (1993) no. 8, 1313 – 1328.
400. J. S. Alabaster, *Chem. Ind.* **1** (1981) 529 – 534.
401. J. Hermens, E. Broekhuyzen, H. Canton, R. Wegman, *Aquat. Toxicol.* **6** (1985) 209 – 217.
402. J. W. Deneer: "The Toxicity of Aquatic Pollutants: QSARs and Mixture Toxicity Studies, Chap. VI: The Joint Acute Toxicity to *Daphnia magna* of Industrial Organic Chemicals at Low Concentrations," Proefschrift 81 – 89, Rijsuniversiteit Utrecht, 1988.
403. ECETOC: "Assessment of Non-occupational Exposure to Chemicals," Technical Report no. 58, Brussels 1994.
404. L. A. Zadeh: "Fuzzy Sets," *Inf. Control* **8** (1965) 338 – 353.
405. A. Salski, *Ecol. Modell.* **63** (1992) 103 – 112.
406. A. Salski, P. Kandzia, *Springer Informatik Fachber.* **296** (1991) 303 – 310.
407. M. Schepers: "Ein Unterstützungssystem zur wissensbasierten Modellierung und Simulation von ökologischen Prozessen," Diplomarbeit, Inst. f. Informatik und Prakt. Mathematik, Universität Kiel 1991.
408. M. T. Bui: "Weiterentwicklung des Unterstützungssystems zur wissensbasierten Modellierung unter dem Einsatz der Fuzzy-Set-Theorie," Diplomarbeit, Inst. f. Informatik und Prakt. Mathematik, Universität Kiel 1991.
409. R. V. Thomann, *Ecol. Modell.* **22** (1984) 145 – 170.
410. M. G. Morgan in J. V. Rodricks, R. G. Tardiff (eds.): "Assessment and Management of Chemical Risks," *ACS Symp. Ser.* **239** (1984) chap. 8.
411. R. C. Schwing, W. A. Albers (eds.): *Societal Risk Assessment: How Safe is Safe Enough?* Plenum Press, New York 1982.
412. R. V. Thomann et al., *Water Res.* **8** (1974) 841 – 851.
413. D. R. Miller in G. C. Butler (ed.): "Principles of Ecotoxicology," *SCOPE* **12** (1979) 71 – 90.
414. J. W. Gillet et al.: *A Conceptual Model for the Movement of Pesticides Through the Environment,* National Environmental Research Center, U.S. Environmental Protection Agency, Corvallis, OR 1974, Report EPA 660/3-74-024, pp. 79.
415. R. R. Lassiter: *Principles and Constraints for Predicting Exposure to Environmental Pollutants,* U.S. Environmental Protection Agency, Corvallis, OR, 1978, Report EPA 118-127 519.
416. T. Fagerstrøm, B. Aasell, *Ambio* **2** (1973) 164 – 171.
417. I. Aoyama, Yos. Inoue, Yor. Inoue, *Water Res.* **12** (1978) 837 – 842.
418. D. K. Leung: *Modelling the Bioaccumulation of Pesticides in Fish,* Center for Ecological Modelling, Polytechnic Institute, Troy, NY, 1978, Report 5.
419. K. L. Seip, *Ecol. Modell.* **6** (1978) 183 – 198.
420. G. T. Orlob, D. Hrovat, F. Harrison, *Adv. Chem. Ser.* **189** (1980) 195 – 212.
421. D. C. L. Lam, T. J. Simons in J. O. Nriago (ed.): *Metals Transfer and Ecological Mass Balances. Environmental Biochemistry,* **vol. 2,** Ann Arbor Sci., Ann Arbor 1976, pp. 537 – 549.
422. T. B. Kirchner, F. W. Whicker, *Ecol. Modell.* **22** (1984) 21 – 44.
423. F. Bro-Rasmussen, K. Christiansen, *Ecol. Modell.* **22** (1984) 67 – 85.
424. S. M. Bartell, R. H. Gardner, R. V. O'Neill, *Ecol. Modell.* **22** (1984) 109 – 123.
425. C. G. Uchrin, *Ecol. Modell.* **22** (1984) 135 – 144.
426. R. P. Schwarzenbach, D. M. Imboden, *Ecol. Modell.* **22** (1984) 171 – 213.
427. E. Halfon, *Ecol. Modell.* **22** (1984) 213 – 253.
428. J. R. W. Harris et al., *Ecol. Modell.* **22** (1984) 253 – 285.
429. N. Nyholm, T. K. Nielsen, K. Pedersen, *Ecol. Modell.* **22** (1984) 285 – 324.
430. J. C. J. Nihoul, *Ecol. Modell.* **22** (1984) 325 – 341.
431. P. A. Arp, *Ecol. Modell.* **19** (1983) 105 – 117.
432. S. Paterson, D. Mackay, *Ecol. Modell.* **47** (1989) 85.
433. M. Matthies et al., *Ecol. Modell.* **47** (1989) 115.
434. J. T. de Luna, T. G. Hallam, *Ecol. Modell.* **35** (1987) 249.
435. T. Morioka, S. Chikami, *Ecol. Modell.* **31** (1986) 267.
436. G. B. Schaalje, R. L. Stinner, D. L. Johnson, *Ecol. Modell.* **47** (1989) 223.
437. B. C. Longstaff, *Ecol. Modell.* **43** (1988) 303.
438. D.-S. Liu, S.-M. Zhang, Z.-G. Li, *Ecol. Modell.* **41** (1988) 75.
439. D.-S. Liu, S.-M. Zhang, *Ecol. Modell.* **37** (1987) 131.
440. R. F. Carsel, L. A. Mulkey, M. N. Lorber, L. B. Baskin, *Ecol. Modell.* **30** (1985) 101.

441. G. R. Gorban, E. Bosatta, *Ecol. Modell.* **40** (1988) 25.
442. J. E. Breck, D. L. de Angelis, W. van Winkle, S. W. Christensen, *Ecol. Modell.* **41** (1988) 1.
443. D. S. Wratt et al., *Ecol. Modell.* **64** (1992) 185 – 204.
444. G. Wuttke, B. Thober, H. Lieth, *Ecol. Modell.* **57** (1991) 263 – 276.
445. J. K. Cronk, W. J. Mitsch, R. M. Sykes, *Ecol. Modell.* **51** (1990) 161 – 192.
446. T. C. Gard, *Ecol. Modell.* **51** (1990) 273 – 280.
447. ECETOC: Hazard Assessment of Chemical Contaminants in Soil, Technical Report no. 40, Brussels 1990.
448. E. E. Kenaga, C. A. I. Goring: "Relationship Between Water Solubility, Soil Sorption, Octanol-water Partitioning and Bioconcentration of Chemicals in Biota," Special Technical Publication (STP) 707, ASTM, Philadelphia, PA, 1980.
449. R. L. Metcalf, G. K. Sangha, I. P. Kopoor, *Environ. Sci. Technol.* **5** (1975) 709 – 713.
450. P.-Y. Lu, R. L. Metcalf, *Environ. Health Perspectives* **10** (1975) 269 – 284.
451. W. B. Neely, D. R. Branson, G. E. Blau, *Environ. Sci. Technol.* **8** (1974) 1113 – 1115.
452. G. D. Veith, D. L. Defoe, B. V. Bergstedt, *J. Fish. Res. Board Can.* **36** (1979) 1040 – 1048.
453. R. R. Lassiter: "Modelling Dynamics of Biological and Chemical Compartments of Aquatic Ecosystems," U.S. Environmental Protection Agency, Washington, D.C., 1975, EPA-660/3-75-012.
454. R. J. Larson, R. L. Perry, *Water Res.* **15** (1981) 697.
455. P. H. Pritchard, A. W. Bourquin, H. L. Fredrickson, T. Maziarz in A. W. Bourquin, R. H. Pritchard (eds.): "Proceedings of the Workshop: Microbial Degradation of Pollutants in Marine Environments," U.S. Environmental Protection Agency, Gulf Breeze, FL, 1979, p. 251.
456. D. S. Brown, E. W. Flagg, *J. Environ. Qual.* **10** (1981) 382 – 386.
457. C. Mater-Müller, W. Gujer, W. Giger, W. Stumm, *Water Technol. Sci.* **12** (1980) 299 – 314.
458. O. C. Zafiriou, M. B. True, *Mar. Chem.* **8** (1979) 9 – 32.
459. O. C. Zafiriou, M. McFarland, *J. Geophys. Res.* **86** (1981) 3173 – 3182.
460. R. G. Zika, L. T. Gidel, D. D. Davis, *J. Geophys. Res.* **11** (1984) 353 – 356.
461. C. van Baalen, J. E. Marler, *Nature (London)* **211** (1966) 951.
462. R. M. Baxter, J. H. Carey, *Freshwater Biol.* **12** (1982) 285 – 292.
463. W. R. Haag, J. Hoigne, E. Grassmann, A. M. Braun, *Chemosphere* **13** (1984) 631 – 640.
464. P. G. Zepp, *Environ. Sci. Technol.* **12** (1978) 327 –329.
465. P. G. Zepp in J. Calkins (ed.): *The Role of Solar Ultraviolet Radiation in Marine Ecosystems,* Plenum Press, New York 1982, pp. 293 – 307.
466. P. G. Zepp in O. Hutzinger (ed.): *The Handbook of Environmental Chemistry,* **vol. 2,** part B, Springer Verlag, Berlin 1982, pp. 19 – 40.
467. C. T. Chiou, V. H. Freed, D. W. Schmedding, R. L. Kohnert, *Environ. Sci. Technol.* **11** (1977) 475 – 478.

Eggs → Foods, 2. Food Technology
Eggs → Proteins
Einsteinium → Transuranium Elements
Elastomeric Fibers → Fibers, 4. Synthetic Organic
Elastomeric Fibers → Rubber, 2. Natural
Elastomeric Fibers → Rubber, 3. Synthetic
Elastomers, Thermoplastic → Thermoplastic Elastomers

My Profile Lo

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Toxicology

Standard Article

Wolfgang Dekant[1] and Spiridon Vamvakas[1]

[1]Institute of Toxicology, University of Wuerzburg, Germany

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Copyright © 2004 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b07_155.pub2
Article Online Posting Date: August 15, 2004

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction
1.1. Definition and Scope
1.2. Fields
1.3. History
1.4. Information Resources
1.5. Terminology of Toxic Effects
1.6. Types of Toxic Effects
1.7. Dose-Response: a Fundamental Issue in Toxicology
1.7.1. Graphics and Calculations
1.8. Dose-Response Relationships for Cumulative Effects
1.9. Factors Influencing Dose-Response
1.9.1. Routes of Exposure
1.9.2. Frequency of Exposure
1.9.3. Species-Specific Differences in Toxicokinetics
1.9.4. Miscellaneous Factors Influencing the Magnitude of Toxic Responses
1.10. Exposure to Mixtures
2. Absorption, Distribution, Biotransformation and Elimination of Xenobiotics
2.1. Disposition of Xenobiotics
2.2. Absorption
2.2.1. Membranes
2.2.2. Penetration of Membranes by Chemicals
2.2.3. Mechanisms of Transport of Xenobiotics through Membranes
2.2.4. Absorption
2.2.4.1. Dermal Absorption
2.2.4.2. Gastrointestinal Absorption
2.2.4.3. Absorption of Xenobiotics by the Respiratory System
2.3. Distribution of Xenobiotics by Body Fluids
2.4. Storage of Xenobiotics in Organs and Tissues
2.5. Biotransformation
2.5.1. Phase-I and Phase-II Reactions
2.5.2. Localization of the Biotransformation Enzymes
2.5.3. Role of Biotransformation in Detoxication and Bioactivation
2.5.4. Phase-I Enzymes and their Reactions
2.5.4.1. Microsomal Monooxygenases: Cytochrome P450
2.5.4.2. Microsomal Monooxygenases: Flavin-Dependent Monooxygenases
2.5.4.3. Peroxidative Biotransformation: Prostaglandin-synthase
2.5.4.4. Nonmicrosomal Oxidations
2.5.4.5. Hydrolytic Enzymes in Phase-I Biotransformation Reactions
2.5.5. Phase-II Biotransformation Enzymes and their Reactions
2.5.5.1. UDP-Glucuronyl Transferases
2.5.5.2. Sulfate Conjugation
2.5.5.3. Methyl Transferases
2.5.5.4. *N*-Acetyl Transferases
2.5.5.5. Amino Acid Conjugation
2.5.5.6. Glutathione Conjugation of Xenobiotics and Mercapturic Acid Excretion
2.5.6. Bioactivation of Xenobiotics
2.5.6.1. Formation of Stable but Toxic Metabolites
2.5.6.2. Biotransformation to Reactive Electrophiles
2.5.6.3. Biotransformation of Xenobiotics to Radicals
2.5.6.4. Formation of Reactive Oxygen Metabolites by Xenobiotics
2.5.6.5. Detoxication and Interactions of Reactive Metabolites with Cellular Macromolecules
2.5.6.6. Interaction of Reactive Intermediates with Cellular Macromolecules
2.5.7. Factors Modifying Biotransformation and Bioactivation
2.5.7.1. Host Factors Affecting Biotransformation
2.5.7.2. Chemical-Related Factors that Influence Biotransformation
2.5.8. Elimination of Xenobiotics and their Metabolites
2.5.8.1. Renal Excretion
2.5.8.2. Hepatic Excretion

2.5.8.3. Xenobiotic Elimination by the Lungs
2.6. Toxicokinetics
2.6.1. Pharmacokinetic Models
2.6.1.1. One-Compartment Model
2.6.1.2. Two-Compartment Model
2.6.2. Physiologically Based Pharmacokinetic Models
3. Mechanisms of Acute and Chronic Toxicity and Mechanisms of Chemical Carcinogenesis
3.1. Biochemical Basis of Toxicology
3.2. Receptor-Ligand Interactions
3.2.1. Basic Interactions
3.2.2. Interference with Excitable Membrane Functions
3.2.3. Interference of Xenobiotics with Oxygen Transport, Cellular Oxygen Utilization, and Energy Production
3.3. Binding of Xenobiotics to Biomolecules
3.3.1. Binding of Xenobiotics or their Metabolites to Cellular Proteins
3.3.2. Interaction of Xenobiotics or their Metabolites with Lipid Constituents
3.3.3. Interactions of Xenobiotics or their Metabolites with nucleic Acids
3.4. Perturbation of Calcium Homeostasis by Xenobiotics or their Metabolites
3.5. Nonlethal Genetic Alterations in Somatic Cells and Carcinogenesis
3.6. DNA Structure and Function
3.6.1. DNA Structure
3.6.2. Transcription
3.6.3. Translation
3.6.4. Regulation of Gene Expression
3.6.5. DNA Repair
3.7. Molecular Mechanisms of Malignant Transformation and Tumor Formation
3.7.1. Mutations
3.7.2. Causal Link between Mutation and Cancer
3.7.3. Proto-Oncogenes and Tumor-Suppressor Genes as Genetic Targets
3.7.4. Genotoxic versus Nongenotoxic Mechanisms of Carcinogenesis
3.8. Mechanisms of Chemically Induced Reproductive and Developmental Toxicity
3.8.1. Embryotoxicity, Teratogenesis, and Transplacental Carcinogenesis
3.8.2. Patterns of Dose-Response in Teratogenesis, Embryotoxicity, and Embryolethality
4. Methods in Toxicology
4.1. Toxicological Studies: General Aspects
4.2. Acute Toxicity
4.2.1. Testing for Acute Toxicity by the Oral Route: LD_{50} Test and Fixed-Dose Method
4.2.2. Testing for Acute Skin Toxicity
4.2.3. Testing for Acute Toxicity by Inhalation
4.3. Repeated-Dose Toxicity Studies: Subacute, Subchronic and Chronic Studies
4.4. Ophtalmic Toxicity
4.5. Sensitization Testing
4.6. Phototoxicity and Photosensitization Testing
4.7. Reproductive and Developmental Toxicity Tests
4.7.1. Fertility and General Reproductive Performance
4.7.2. Embryotoxicity and Teratogenicity
4.7.3. Peri- and Postnatal Toxicity
4.7.4. Multigeneration Studies
4.7.5. The Role of Maternal Toxicity in Teratogenesis
4.7.6. *In Vitro* Tests for Developmental Toxicity
4.8. Bioassays to determine the carcinogenicity of chemicals in rodents
4.9. *In vitro* and *in vivo* short-term tests for genotoxicity
4.9.1. Microbial Tests for Mutagenicity
4.9.1.1. The Ames Test for Bacterial Mutagenicity
4.9.1.2. Mutagenicity Tests in *Escherichia coli*
4.9.1.3. Fungal Mutagenicity Tests
4.9.2. Eukaryotic Tests for Mutagenicity
4.9.2.1. Mutation Tests in *Drosophila melanogaster*
4.9.2.2. *In Vitro* Mutagenicity Tests in Mammalian Cells
4.9.3. *In Vivo* Mammalian Mutation Tests
4.9.3.1. Mouse Somatic Spot Test
4.9.3.2. Mouse Specific Locus Test
4.9.3.3. Dominant Lethal Test
4.9.4. Test Systems Providing Indirect Evidence for DNA Damage
4.9.4.1. Unscheduled DNA Synthesis (UDS) Assays
4.9.4.2. Sister-Chromatid Exchange Test
4.9.5. Tests for Chromosome Aberrations (Cytogenetic Assays)
4.9.5.1. Cytogenetic Damage and its Consequences
4.9.5.2. *In Vitro* Cytogenetic Assays
4.9.5.3. *In Vivo* Cytogenetic Assays
4.9.6. Malignant Transformation of Mammalian Cells in Culture
4.9.7. *In Vivo* Carcinogenicity Studies of Limited Duration
4.9.7.1. Induction of Altered Foci in the Rodent Liver
4.9.7.2. Induction of Lung Tumors in Specific Sensitive Strains of Mice
4.9.7.3. Induction of Skin Tumors in Specific Sensitive Strains of Mice
4.9.8. Methods to Assess Primary DNA Damage
4.9.8.1. Alkaline Elution Techniques
4.9.8.2. Methods to Detect and Quantify DNA Modifications
4.9.9. Interpretation of Results Obtained in Short-Term Tests

4.10. Evaluation of Toxic Effects on the Immune System
4.11. Toxicological Evaluation of the Nervous System
4.11.1. Functional Observational Battery
4.11.2. Locomotor Activity
4.12. Effects on the Endocrine System
5. Evaluation of Toxic Effects
5.1. Acceptable risk, Comparison of Risks, and Establishing Acceptable Levels of Risk
5.2. The Risk Assessment Process
5.2.1. Hazard Identification Techniques
5.2.2. Determination of Exposure
5.2.3. Dose-Response Relationships
5.2.4. Risk Characterization
5.2.4.1. The Safety-Factor Methodology
5.2.4.2. Risk Estimation Techniques for Nonthreshold Effects
5.2.4.3. Mathematical Models Used in High- to Low-Dose Risk Extrapolation
5.2.4.4. Interpretation of Data from Chronic Animal Bioassays
5.2.4.5. Problems and Uncertainties in Risk Assessment
5.3. Future Contributions of Scientifically Based Procedures to Risk Assessment and Qualitative Risk Assessment for Carcinogens
5.4. Risk Assessment for Teratogens

[Top of Page]

1. Introduction

1.1. Definition and Scope

Chemicals that are used or of potential use in commerce, the home, the environment, and medical practice may present various types of harmful effects. The nature of these effects is determined by the physicochemical characteristics of the agent, its ability to interact with biological systems (hazard), and its potential to come into contact with biological systems (exposure).

Toxicology studies the interaction between chemicals and biological systems to determine the potential of chemicals to produce adverse effects in living organisms. Toxicology also investigates the nature, incidence, mechanisms of production, factors influencing their development, and reversibility of such adverse effects. Adverse effects are defined as detrimental to the survival or the normal functioning of the individual. Inherent in this definition are the following key issues in toxicology:

1. Chemicals must come into close structural and/or functional contact with tissues or organs to cause injury.
2. All adverse effects depend on the amount of chemical in contact with the biological system (the dose) and the inherent toxicity of the chemical (hazard). When possible, the observed toxic effect should be related to the degree of exposure. The influence of different exposure doses on the magnitude and incidence of the toxic effect should be quantitated. Such dose-response relationships are of prime importance in confirming a causal relationship between chemical exposure and toxic effect (for details, see Section Dose–Response: a Fundamental Issue in Toxicology).

Research in toxicology is mainly concerned with determining the potential for adverse effects caused by chemicals, both natural and synthetic, to assess their hazard and risk of human exposure and thus provide a basis for appropriate precautionary, protective and restrictive measures. Toxicological investigations should permit evaluation of the following characteristics of toxicity:

1. The basic structural, functional, or biochemical injury produced
2. Dose-response relationships
3. The mechanisms of toxicity (fundamental biochemical alterations responsible for the induction and maintenance of the toxic response) and reversibility of the toxic effect
4. Factors that modify response, e.g., route of exposure, species, and gender

For chemicals to which humans may potentially be exposed, a critical analysis, based on the pattern of potential exposure or toxicity, may be necessary in order to determine the risk-benefit ratio for their use in specific circumstances and to devise protective and precautionary measures. Indeed, with drugs, pesticides, food additives, and cosmetic preparations, toxicology testing must be performed in accordance with government regulations before use.

1.2. Fields

Toxicology is a recognized scientific discipline encompassing both basic and applied issues. Although only generally accepted as a specific scientific field during this century, its principles have been appreciated for centuries. The harmful or lethal effects of certain chemicals, mainly present in minerals and plants or transmitted venomous animals, have been known since prehistoric times. In many countries, toxicology as a discipline has developed from pharmacology. Pharmacology and toxicology both study the effect of chemicals on living organisms and have often used identical methods. However, fundamental differences have developed. Years ago, only the dependence on dose of the studied effects separated pharmacology and toxicology. Pharmacology focused on chemicals with beneficial effects (drugs) at lower doses whereas toxicology studied the adverse health effects occurring with the same chemicals at high doses. Today, the main interest of research in toxicology has shifted to studies on the long-term effects of chemicals after low-dose exposure, such as cancer or other irreversible diseases; moreover, most chemicals of interest to toxicologists are not used as drugs.

The variety of potential adverse effects and the diversity of chemicals present in our environment combine to make toxicology a very broad science. Toxicology uses basic knowledge from clinical and theoretical medicine and natural sciences such as biology and chemistry (Fig. 1). Because of this diversity, toxicologists usually specialize in certain areas.

Figure 1. Scientific fields influencing the science of toxicology

Any attempt to define the scope of toxicology must take into account that the various subdisciplines are not mutually exclusive and frequently are heavily interdependent. Due to the overlapping mechanisms of toxicity, chemical classes, and observed toxic effects, clear divisions into subjects of equal importance are often not possible.

The professional activities of toxicologists can be divided into three main categories: descriptive, mechanistic, and regulatory. The *descriptive toxicologist* is concerned directly with toxicity testing. Descriptive toxicology still often relies on the tools of pathology and clinical chemistry, but since the 1970s more mechanism-based test systems have been included in toxicity testing [1]. The appropriate toxicity tests in experimental animals yield information that is extrapolated to evaluate the risk posed by exposure to specific chemicals. The concern may be limited to effects on humans (drugs, industrial chemicals in the workplace, or food additives) or may encompass animals, plants, and other factors that might disturb the balance of the ecosystem (industrial chemicals, pesticides, environmental pollutants).

The *mechanistic toxicologist* is concerned with elucidating the mechanisms by which chemicals exert their toxic effects on living organisms. Such studies may result in the development of sensitive predictive toxicity tests useful in obtaining information for risk assessment (see Chap. Methods in Toxicology). Mechanistic studies may help in the development of chemicals that are safer to use or of more rational therapies for intoxications. In addition, an understanding of the mechanisms of toxic action also contributes to the knowledge of basic mechanisms in physiology, pharmacology, cell biology, and biochemistry. Indeed, toxic chemicals have been

used with great success as mechanistic tools to elucidate mechanisms of physiological regulation. Mechanistic toxicologists are often active in universities; however, industry and government institutions are now undertaking more and more research in mechanistic toxicology.

Regulatory toxicologists have the responsibility of deciding on the basis of data provided by the descriptive toxicologist and the mechanistic toxicologist if a drug or chemical poses a sufficiently low risk to be used for a stated purpose. Regulatory toxicologists are often active in government institutions and are involved in the establishment of standards for the amount of chemicals permitted in ambient air in the environment, in the workplace, or in drinking water. Other divisions of toxicology may be based on the classes of chemicals dealt with or application of knowledge from toxicology for a specific field (Table 1)

Table 1. Areas of toxicology

Field	Tasks and objectives
Forensic toxicology	diagnoses poisoning by analytical procedures
Pesticide toxicology	studies the safety of pesticides, develops new pesticides
Occupational toxicology	assesses potential adverse effects of chemicals used in the workplace, recommends protective procedures
Drug toxicology	studies potential effects of drugs after high doses, elucidates mechanisms of sideeffects
Regulatory toxicology	develops and interprets toxicity testing programs and is involved in controlling the use of chemicals
Environmental toxicology	studies the effects of chemicals on ecosystems and on humans after low-dose exposure from the environment

.

Forensic toxicology comprises both analytical chemistry and fundamental toxicologic principles. It is concerned with the legal aspects of the harmful effects of chemicals on humans. The expertise of the forensic toxicologist is invoked primarily to aid in establishing the cause of death and elucidating its circumstances in a postmortem investigation. The field of *clinical toxicology* recognizes and treats poisoning, both chronic and acute. Efforts are directed at treating patients poisoned by chemicals and at the development of new techniques to treat these intoxications. *Environmental toxicology* is a relatively new area that studies the effects of chemicals released by man on wildlife and the ecosystem and thus indirectly on human health.

Drug toxicology plays a major role in the preclinical safety assessment of chemicals intended for use as drugs. Drug toxicology also elucidates the mechanisms of side effects observed during clinical application. *Occupational toxicology* studies the acute and chronic toxicity of chemicals encountered in the occupational environment. Both acute and chronic occupational poisonings have exerted a major influence on the development of toxicology in general. Occupational toxicology also helps in the development of safety procedures to prevent intoxications in the workplace and assists in the definition of exposure limits. *Pesticide toxicology* is involved in the development of new pesticides and the safety of pesticide formulations. Pesticide toxicology also characterizes potential health risks to the general population caused by pesticide residues in food and drinking water.

1.3. History

Toxicology must rank as one of the oldest practical sciences because humans, from the very beginning, needed to avoid the numerous toxic plants and animals in their environment. The presence of toxic agents in animals and plants was known to the Egyptian and Greek civilisations. The papyrus Ebers, an Egyptian papyrus dating from about 1 500 B.C., and the surviving medical works of HIPPOCRATES, ARISTOTLE, and THEOPHRASTUS, published during the period 400–250 B.C., all included some mention of poisons.

The Greek and Roman civilizations knowingly used certain toxic chemicals and extracts for hunting, warfare, suicide, and murder. Up to the Middle Ages, toxicology was restricted to the use of toxic agents for murder. Poisoning was developed to an art in medieval Italy and has remained a problem ever since, and much of the earlier impetus for the development of toxicology was primarily forensic. There appear to have been few advances in either medicine or toxicology between the time of GALEN (131–200 A.D.) and PARACELSUS (1493–1541). The latter laid the groundwork for the later development of modern toxicology. He clearly was aware of the dose–response relationship. His statement that “All substances are poisons; there is none that is not a poison. The right dose differentiates a poison and a remedy,” is properly regarded as a landmark in the development of the science of toxicology. His belief in the value of experimentation also represents a break with much earlier tradition. Important developments in the 1700s include the publication of RAMAZZINI's *Diseases of Workers*, which led to his recognition as the father of occupational medicine. The correlation between the occupation of chimney sweepers and scrotal cancer by POTT in 1775 is also noteworthy.

ORFILA, a Spaniard working at the University of Paris, clearly identified toxicology as a separate science and wrote the first book devoted exclusively to it (1815). Workers of the later 1800s who produced treatises on toxicology include CHRISTISON, KOBERT, and LEWIN. They increased our knowledge of the chemistry of poisons, the treatment of poisoning, the analysis of both xenobiotics and toxicity, as well as modes of action and detoxication. A major impetus for toxicology in the 1900s was the use of chemicals for warfare. In World War I, a variety of poisonous chemicals were used in the battlefields of France. This provided stimulus for work on mechanisms of toxicity as well as medical countermeasures to poisoning. Since the 1960s, toxicology has entered a phase of rapid development and has changed from a science that was almost entirely descriptive to one in which the study of mechanisms has become the prime task. The many reasons for this include the development of new analytical methods since 1945, the emphasis on drug testing following the thalidomide tragedy, the emphasis on pesticide testing following the publication of Rachel Carson's *Silent Spring* and public concern over environmental pollution and disposal of hazardous waste.

1.4. Information Resources

Because of the complexity of toxicology as a science and the impact of toxicological investigations on legislation and commerce, a wide range of information on the toxic effects of chemicals is available. No single, exhaustive source of toxicological data exists; several sources are required to obtain comprehensive information on a particular chemical. Printed sources are often quicker and easier to use than computer data bases, but interactive online searching can rapidly gather important information from the huge number of sources present.

The information explosion in toxicology has resulted in a comprehensive volume dedicated to toxicological information sources:

P. Wexler, P. J. Hakkinen, G. Kennedy, Jr. F. W.Stoss, *Information Resources in Toxicology*, 3rd ed., Academic Press, 1999.

Textbooks. The easiest way to obtain information on general topics in toxicology and secondary references are a range of textbooks available on the market. Only a few selected books are listed below:

C. D. Klaasen, *Casarett and Doull's Toxicology; The Basic Science of Poisons*, 6th ed., McGraw-Hill, New York, 2001.
G. D. Clayton, F. E. Clayton (eds): *Patty's Industrial Hygiene and Toxicology*, Wiley, New York, 1993.
J. G. Hardman, L. E. Limbird, *Goodman and Gilman's, The Pharmacological Basis of Therapeutics*, 10th ed., McGraw-Hill, New York, 2001.
W. A. Hayes, *Principles and Methods of Toxicology*, 3rd ed., Raven Press, New York, 2001.
E. Hodgson (Ed.): *Textbook of Modern Toxicology*, 3rd ed., Wiley Interscience, 2004.
T. A. Loomis, A. W. Hayes, *Loomis's Essentials of Toxicology*, 4th ed., Academic Press, San Diego, 1996.

The huge volume by N. I. Sax and R. J. Lewis, *Dangerous Properties of Industrial Materials*, 7th ed.,Wiley, New York, 1999, contains basic toxicological data on a large selection of chemicals (almost 20 000) and may serve as a useful guide to the literature for compounds not covered in other publications.

Monographs. The best summary information on toxicology is published in the form of series by governments and international organizations. Most of these series are summarizing the results of toxicity studies on specific chemicals. The selection of these chemicals is mainly based on the extent of their use in industry (e.g. trichloroethene), their occurrence as environmental contaminants (mercury) or their extraordinary toxicity (e.g. 2,3,7,8-tetrachlorodibenzodioxin):

American Conference of Governmental Industrial Hygienists, Threshold Limit Values and Biological Exposure Indices (Cincinnati, OH). Published annually. MAK-Begründungen, VCH Publishers, Weinheim, Federal Republic of Germany. This German series includes detailed information on the toxicity of chemicals on the German MAK list (ca. 150 reports are available; the series is continuously expanded).

The Commission of the European Communities publishes the Reports of the Scientific Committee on Cosmetology and the Reports of the Scientific Committee for Food.

The Environmental Protection Agency (EPA) publishes a huge number of reports and toxicological profiles. They are indexed in “EPA Publications. A Quarterly Guide.”

The European Chemical Industry Ecology and Toxicology Centre (ECETOC) issues “Monographs” (more than 20 have been published) and “Joint Assessments of Commodity Chemicals.”

The monographs of the International Agency for Research on Cancer are definitive evaluations of carcinogenic hazards. The “Environmental Health Criteria” documents of the International Programme on Chemical Safety (IPCS) assess environmental and human health effects of exposure to chemicals, and biological or physical agents. A related “Health and Safety Guide” series give guidance on setting exposure limits for national chemical safety programs.

The National Institute for Occupational Safety and Health (NIOSH), has published 50 “Current Intelligence Bulletins” on health hazards of materials and processes at work.

The technical report series of the National Toxicology Program (NTP) reports results of their carcinogenicity bioassays, which include summaries of the toxicology of the chemicals studied. A status report indexes both studies that are under way and those that have been published. The program also issues an “Annual Review of Current DHHS [U.S. Department of Health and Human Services], DOE [U.S. Department of Energy] and EPA Research” related to toxicology.

A large number of internet-based resources are also available to collect information on toxic effects of chemicals and methods for risk assessment. Some information sites containing large amounts of downloadable information are listed below:

US Environmental Protection Agency (EPA), Integrated Risk Information System (IRIS), http://www.epa.gov/iris/index.html
US Environmental Protection Agency (EPA), ECOTOX Database, http://www.epa.gov/ecotox/
Organisation for Economic Co-operation and Development (OECD), test guidelines, http://www.oecd.org
Agency for Toxic Substances and Disease Registry (ATSDR), toxicological profile information sheet http://www.atsdr.cdc.gov/toxprofiles/
European Chemicals Bureau, http://ecb.jrc.it/
National Toxicology Programm, http://ntp-server.niehs.nih.gov/htdocs/liason/Factsheets/FactsheetList.html
United Nations Environment Programm, Chemicals http://www.chem.unep.ch/

Journals Results of toxicological research are published in more than 100 journals. Those listed below mainly publish research closely related to toxicology, but articles of relevance may also be found in other biomedical journals:

Archives of Environmental Contamination and Toxicology
Archives of Toxicology
Biochemical Pharmacology
Chemical Research in Toxicology
CRC Critical Reviews in Toxicology
Clinical Toxicology
Drug and Chemical Toxicology
Environmental Toxicology and Chemistry
Food and Chemical Toxicology
Fundamental and Applied Toxicology
Journal of the American College of Toxicology
Journal of Analytical Toxicology
Journal of Applied Toxicology
Journal of Biochemical Toxicology
Journal of Toxicology and Environmental Health
Neurotoxicology and Teratology
Pharmacology and Toxicology
Practical In Vitro Toxicology
Regulatory Toxicology and Pharmacology
Reproductive Toxicology
Toxicology
Toxicology and Applied Pharmacology
Toxicology and Industrial Health
Toxicology In Vitro
Toxicology Letters

Databases and Databanks. Electronic sources, such as computer data bases or CD-ROM are a fast and convenient way to obtain references on the toxicity of chemicals. Since on-line searching of commercial data bases such as STN-International may be expensive, CD-ROM-based systems are increasingly being used. The major advantages are speed, the ability to refine searches and format the results, and non-text search options, such as chemical structure searching on Beilstein and Chemical Abstracts.

Useful information about actual research on the toxicology of chemicals may be obtained by searching Chemical Abstracts or Medline with the appropriate keywords. Specific data banks covering toxicology are the Registry of Toxic Effects of Chemical Substances, which gives summary data, statistics, and structures; Toxline (available in DIMDI) gives access to the literature.

1.5. Terminology of Toxic Effects

Toxic effects may be divided according to timescale (acute and delayed), general locus of action (local, systemic, organ specific), or basic mechanisms of toxicity (reversible versus irreversible). *Acute toxic effects* are those that occur after brief exposure to a chemical. Acute toxic effects usually develop rapidly after single or multiple administrations of a chemical; however, acute exposure may also produce *delayed toxicity*. For example, inhalation of a lethal dose of HCN causes death in less than a minute, whereas lethal doses of 2,3,7,8-tetrachlorodibenzodioxin will result in the death of experimental animals after more than two weeks. *Chronic effects* are those that appear after repetitive exposure to a substance; many compounds require several months of continuous exposure to produce adverse effects. Often, the chronic effects of chemicals are different from those seen after acute exposure (Table 2). For example, inhalation of chloroform for a short period of time may cause anesthesia; long-term inhalation of much lower chloroform concentrations causes liver damage. Carcinogenic effects of chemicals usually have a long latency period; tumors may be observed years (in rodents) or even decades (in humans) after exposure.

Table 2. Toxic effects of different chemicals categorized by time scale and general locus of action

Exposure	Site	Effect	Chemical
Acute	local	lung edema	chlorine gas
	systemic	liver damage	carbon tetrachloride
		narcosis	halothane
Subchronic	local	sensitization	toluene diisocyanate
	systemic	neurotoxicity	hexane
Chronic	local	bronchitis	sulfur dioxide
		nasal carcinoma	formaldehyde
	systemic	bladder carcinoma	4-amino-biphenyl
		kidney damage	cadmium

Toxic effects of chemicals may also be classified based on the type of interaction between the chemical and the organism. Toxic effects may be caused by reversible and irreversible interactions (Table 3). When reversible interactions are responsible for toxic effects, the concentration of the chemical present at the site of action is the only determinant of toxic outcome. When the concentration of the xenobiotic is decreased by excretion or biotransformation, a parallel decrease of toxic effects is

observed.

Table 3. Reversible and irreversible interactions of chemicals with cellular macromolecules as a basis for toxic response

Mechanism	Toxic response	Example
Irreversible inhibition of		
Esterase	neurotoxicity	tri-*o*-cresylphosphate
Covalent binding to DNA	cancer	dimethylnitrosamine
Reversible binding to		
Hemoglobin	oxygen deprivation in tissues	carbon monoxide
Cholinesterase	neurotoxicity	carbamate pesticides

After complete excretion of the toxic agent, toxic effects are reduced to zero (see below). A classical example for reversible toxic effects is carbon monoxide. Carbon monoxide binds to hemoglobin and, due to the formation of the stable hemoglobin–carbon monoxide complex, binding of oxygen is blocked. As a result of the impaired oxygen transport in blood from the lung, tissue oxygen concentrations are reduced and cells sensitive to oxygen deprivation will die. The toxic effects of carbon monoxide are directly correlated with the extent of carboxyhemoglobin in blood, the concentration of which is dependent on the inhaled concentration of carbon monoxide. After exhalation of carbon monoxide and survival of the acute intoxication, no toxic effect remains (Fig. 2).

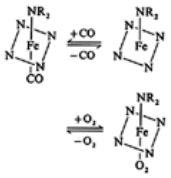

Figure 2. Reversible binding of carbon monoxide to hemoglobin and inhibition of oxygen transport

Irreversible toxic effects are often caused by the covalent binding of toxic chemicals to biological macromolecules. Under extreme conditions, the modified macromolecule is not repaired; after excretion of the toxic agent, the effect persists. Further exposure to the toxic agent will produce additive effects; many chemicals carcinogens are believed to act through irreversible changes (see Section Bioactivation of Xenobiotics).

Another distinction between types of effects may be made according to the general locus of action. *Local toxicity* occurs at the site of first contact between the biological system and the toxic agent. Local effects to the skin, the respiratory tract, or the alimentary tract may be produced by skin contact with a corrosive agent, by inhalation of irritant gases, or by ingestion of tissue-damaging materials. This type of toxic responses is usually restricted to the tissues with direct contact to the agent. However, life-threatening intoxications may occur if vital organs like the lung are damaged. For example, inhaled phosgene damages the alveoli of the lung and causes lung edema. The massive damage to the lung results in the substantial mortality observed after phosgene intoxication.

The opposite to local effects are *systemic effects*. They are characterized by the absorption of the chemical and distribution from the port of entry to a distant site where toxic effects are produced. Except for highly reactive xenobiotics, which mainly act locally, most chemicals act systemically. Many chemicals that produce systemic toxicity only cause damage to certain organs, tissues, or cell types within organs. Selective damage to certain organs or tissues by systemically distributed chemicals is termed organ- or tissue-specific toxicity [2]; the organs damaged are referred to as target organs (Table 4).

Table 4. Organ-specific toxic effects induced by chemicals that are distributed systemically in the organism

Chemical	Species	Target organ
Benzene	humans	bone marrow
Hexachlorobutadiene	rodents	damage to proximal tubules of the kidney
Paraquat	rodents, humans	lung
Tri-*o*-cresylphosphate	humans	nervous system
Cadmium	humans	kidney
1,2-Dibromo-3-chloropropane	humans, rodents	testes
Hexane	rodents, humans	nervous system
Anthracyclines	humans	heart

Major target organs for toxic effects are the central nervous system and the circulatory system followed by the blood and hematopoietic system and visceral organs such as the liver or the kidney. For some chemicals, both local and systemic effects can be demonstrated; moreover, chemicals producing marked local toxicity may also cause systemic effects as secondary responses to major disturbances in homeostasis of the organism.

1.6. Types of Toxic Effects

The spectrum of toxic effects of chemicals is broad, and their magnitude and nature depend on many factors such as the physiocochemical properties of the chemical and its toxicokinetics, the conditions of exposure, and the presence of adaptive and protective mechanisms. The latter factors include physiological mechanisms such as adaptive enzyme induction, DNA repair, and others. Toxic effects may be transient, reversible, or irrversible; some are deleterious and others are not. Toxic effects may take the form of tissue pathology, aberrant growth processes, or altered biochemical pathways. Some of the more frequently encountered types of injury constituting a toxic response are described in the following.

Immune-mediated hypersensitivity reactions by antigenic materials are toxic effects often involved in skin and lung injury by repeated contact to chemicals resulting in contact dermatitis and asthma. Inflammation is a frequently observed local response to the application of irritant chemicals or may be a component of systemic injury. This response may be acute with irritant or tissue damaging materials or chronic with repetitive exposure to irritants. Necrosis, that is, death of cells or tissues, may be the result of various pathological processes resulting from biochemical interactions of xenobiotics, as described in Chapter 3. The extent and patterns of necrosis may be different for different chemicals, even in the same organ. Chemical tumorigenesis or carcinogenesis (induction of malignant tumors) is an effect often observed after chronic application of chemicals. Due the long latency period and the poor prognosis for individuals diagnosed with cancer, studies to predict the potential tumorigenicity of chemicals have developed into a major area of toxicological research. Developmental and reproductive toxicology are concerned with adverse effects on the ability to conceive, and with adverse effects on the structural and functional integrity of the fetus. Chemicals may interfere with reproduction through direct effects on reproductive organs or indirectly by affecting their neural and endocrine control mechansims. Developmental toxicity deals with adverse effects on the conceptus through all stages of pregnancy. Damage to the fetus may result in embryo reabsorption, fetal death, or abortion. Nonlethal fetotoxicity may be expressed as delayed maturation, decreased birth weight, or structural malformation. The most sensitive period for the induction of malformation is during organogenesis; neurobehavioral malformations may be induced during later stages of pregnancy.

1.7. Dose–Response: a Fundamental Issue in Toxicology

In principle, a poison is a chemical that has an adverse effect on a living organism. However, this is not a useful definition since toxic effects are related to dose. The definition of a poison thus also involves quantitative biological aspects. At sufficiently high doses, any chemical may be toxic. The importance of dose is clearly seen with molecular oxygen or dietary metals. Oxygen at a concentration of 21% in the atmosphere is essential for life, but 100% oxygen at atmospheric pressure causes massive lung injury in rodents and often results in death. Some metals such as iron, copper, and zinc are essential nutrients. When they are present in insufficient amounts in the human diet, specific disease patterns develop, but in high doses they can cause fatal intoxications. Toxic compounds are not restricted to man-made

chemicals, but also include many naturally occurring chemicals. Indeed, the agent with the highest toxicity is a natural poison found in the bacterium *Clostridium botulinum* (LD_{50} 0.01 µ/kg).

Therefore, all toxic effects are products of the amount of chemical to which the organism is exposed and the inherent toxicity of the chemical; they also depend on the sensitivity of the biological system.

The term “dose” is most frequently used to characterize the total amount of material to which an organism is exposed; dose defines the amount of chemical given in relation to body weight. Dose is a more meaningful and comparative indicator of exposure than the term exposure itself. Dose usually implies the exposure dose, the total amount of chemical administered to an organism or incorporated into a test system. However, dose may not be directly proportional to the toxic effects since toxicity depends on the amount of chemical absorbed. Usually, dose correctly describes only the actual amount of chemical absorbed when the chemical is administered orally or by injection. Under these circumstances, the administered dose is identical to the absorbed dose; other routes of application such as dermal application or inhalation do not define the amount of agent absorbed.

Different chemicals have a wide spectrum of doses needed to induce toxic effects or death. To characterize the acute toxicity of different chemicals, LD_{50} values are frequently used as a basis for comparisons. Some LD_{50} values (rat) for a range of chemicals follow:

Ethanol	12 500
Sodium bicarbonate	4 220
Phenobarbital sodium	350
Paraquat	120
Aldrin	46
Sodium cyanide	6.4
Strychnine	5
1,2-Dibromoethane	0.4
Sodium fluoroacetate	0.2
2,3,7,8-Tetrachlorodibenzodioxin	0.01

Certain chemicals are very toxic and produce death after administration of microgram doses, while others are tolerated without serious toxicity in gram doses. The above data clearly demonstrate that the toxicity of a specific chemical is related to dose. The dependence of the toxic effects of a specific chemical on dose is termed dose–response relationship. Before dose–response relationships can be appropriately used, several basic assumptions must be considered. The first is that the response is due to the chemical administered. It is usually assumed that the responses observed were a result of the various doses of chemical administered. Under experimental conditions, the toxic response usually is correlated to the chemical administered, since both exposure and effect are well defined and can be quantified. However, it is not always apparent that the response is the result of specific chemical exposure. For example, an epidemiologic study might result in discovery of an “association” between a response (e.g., disease) and one or more variables including the estimated dose of a chemical. The true doses to which individuals have been exposed are often estimates, and the specificity of the response for that chemical is doubtful.

Further major necessary assumptions in establishing dose–response relationships are:

- A molecular site (often termed receptor) with which the chemical interacts to produce the response. Receptors are macromolecular components of tissues with which a chemical interacts and produces its characteristic effect.
- The production of a response and the degree of the response are related to the concentration of the agent at the receptor.
- The concentration of the chemical at the receptor is related to the dose administered. Since in most cases the concentration of an administered chemical at the receptor cannot be determined, the administered dose or the blood level of the chemical is used as an indicator for its concentration at the molecular site.

A further prerequisite for using the dose–response relationship is that the toxic response can be exactly measured. A great variety of criteria or end points of toxicity may be used. The ideal end point should be closely associated with the molecular events resulting from exposure to the toxin and should be readily determined. However, although many end points are quantitative and precise, they are often only indirect measures of toxicity. For example, changes in enzyme levels in the blood can be indicative of tissue damage. Patterns of alterations may provide insight into which organ or system is the site of toxic effects. These measures usually are not directly related to the mechanism of toxic action. The dose–response relationship combines the characteristic of exposure and the inherent toxicity of the chemical. Since toxic responses to a chemical are usually functions of both time and dose, in typical dose–response relationships, the maximum effect observed during the time of observation is plotted against the dose to give time-independent curves. The time-independent dose–response relationship may be used to study dose–response for both reversible and irreversible toxic effects. However, in risk assessments that consider the induction of irreversible effects such as cancer, the time factor plays a major role and has important influences on the magnitude or likelihood of toxic responses. Thus, for this type of mechanism of toxic action, dose–time–response relationships are better descriptors of toxic effects.

The dose–response relationship is the most fundamental concept in toxicology. Indeed, an understanding of this relationship is essential for the study of toxic chemicals.

From a practical point of view, there are two different types of dose–response relationships. Dose–response relationships may be quantal (all or nothing responses such as death) or graded. The graded or variable response involves a continual change in effect with increasing dose, for example, enzyme inhibition or changes in physiological function such as heart rate. Graded responses may be determined in an individual or in simple biochemical systems. For example, addition of increasing concentrations of 2,3,7,8-tetrachlorodibenzodioxin to cultured mammalian cells results in an increase in the concentration of a specific cytochrome P450 enzyme in the cells (for details of mechanisms, see Section Microsomal Monooxygenases: Cytochrome P450). The increase is clearly dose related and spans a wide range (Fig. 3). An example for a graded toxic effect in an individual may be inflammation caused by skin contact with an irritant material. Low doses cause slight irritation; as the amount increases, irritation turns to inflammation and the severity of inflammation increases.

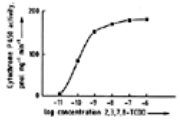

Figure 3. Dose-dependent induction of cytochrome P450 1A 1 protein in cultured liver cells treated with 2,3,7,8-tetrachlorodibenzodioxin [3]

In dose–response studies in a population, a specific endpoint is also identified and the dose required to produce this end point is determined for each individual in the population. Both dose-dependent graded effects and quantal responses (death, induction of a tumor) may be investigated. With increasing amount of a chemical given to a group of animals, the magnitude of the effect and/or the number of animals affected increase. For example, if an irritant chemical is applied to the skin, as the amount of the material increases, the numbers of animals affected and the severity of inflammation increases. Quantal responses such as death induced by a potentially lethal chemical will also be dose-dependent. The dose dependency of a quantal effect in a population is based on individual differences in the response to the toxic chemical. A specific amount of the potentially lethal xenobiotic given to a group of animals may not kill all of them, but as the amount given increases, the proportion of animals killed increases.

Althought the distinctions between graded and quantal dose–response relationships are useful, the two types of responses are conceptually identical. The ordinate in both cases is simply labeled response, which may be the degree of response in an individual, or the fraction of a population responding, and the abscissa is the range of administered doses.

1.7.1. *Graphics and Calculations*

Even with a genetically homogenous population of animals of the same species and strain, the proportion of animals showing the effect will increase with dose (Fig. 4A). When the number of animals responding is plotted versus the logarithm of the dose, a typical sigmoid curve with a log-normal distribution that is symmetrical about the midpoint, is obtained (Fig. 4B).

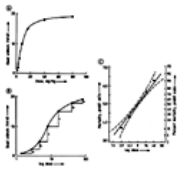

Figure 4. Typical dose – response curves for a toxic effect

Plots are linear – linear (A); log – linear (B); and log – probit (C) for an identical set of data

When plotted on a log-linear scale, the obtained normally distributed sigmoid curve approaches a response of 0% as the dose is decreased, and 100% as the dose is increased, but theoretically never passes through 0 or 100%. Small proportions of the population at the right- and left-hand sides of the curve represent hyposusceptible and hypersusceptible members. The slope of the dose–reponse curve around the 50% value, the midpoint, gives an indication of the ranges of doses producing an effect. A steep dose–response curve indicates that the majority of the population will respond over a narrow dose range; a shallow dose–response curve indicates that a wide range of doses is required to affect the majority of the population. The curve depicted in Fig. 4 B shows that the majority of the individuals respond about the midpoint of the curve. This point is a convenient description of the average response, and is referred to as the median effective dose (ED_{50}). If mortality is the endpoint, then this dose is referred as median lethal dose (LD_{50}).

Death, a quantal response, is simple to quantify and is thus an end point incorporated in many acute toxicity studies. Lethal toxicity is usually calculated initially from specific mortality levels obtained after giving different doses of a chemical; the 50% mortality level is used most frequently since it represents the midpoint of the dose range at which the majority of deaths occur. This is the dose level that causes death of half of the population dosed. The LD_{50} values are usually given in milligrams of chemical per kilogram of body weight (from the viewpoint of chemistry and for comparison of relative potencies of different chemicals, giving the LD_{50} in moles of chemical per kilogram body weight would be desirable). After inhalation, the reference is to LC_{50} (LC = lethal concentration), which, in contrast to LD_{50} values, depends on the time of exposure; thus, it is usually expressed as *X*-hour LC_{50} value. The LD_{50} or LC_{50} values usually represent the initial information on the toxicity of a chemical and must be regarded as a first, but not a quantitative, hazard indicator that may be useful for comparison of the acute toxicity of different chemicals [3].

Similar dose–effect curves can, however, be constructed for cancer, liver injury, and other types of toxic responses. For the determination of LD_{50} values and for obtaining comparative information on dose–response curves, plotting log dose versus percent response is not practical since large numbers of animals are needed for obtaining interpretable data. Moreover, other important information on the toxicity of a chemical (e.g., LD_{05} and LD_{95}) cannot be accurately determined due to the slope of sigmoid curve. Therefore, the dose–response curve is transformed to a log-probit (probit = probability units) plot. The data in the Fig. 4B form a straight line when transformed into probit units (Fig. 4C). The EC_{50} or, if death is the end point, the LD_{50} is obtained by drawing a horizontal line from the probit unit 5, which is the 50% response point, to the dose–effect line. At the point of intersection a vertical line is drawn, and this line intersects the abscissa at the LD_{50} point. Information on the lethal dose for 90% or for 10% of the population can also be derived by a similar procedure. The confidence limits are narrowest at the midpoint of the line (LD_{50}) and are widest at the two extremes (LD_{05} and LD_{95}) of the dose–response curve. In addition to permitting determination of a numerical value for the LD_{50} of a chemical with few groups of dosed animals, the slope of the dose–response curve for comparison between toxic effects of different chemicals is obtained by the probit transformation [4].

The LD_{50} by itself, however, is an insufficient index of lethal toxicity, particularly if comparisons between different chemicals are to be made. For this purpose, all available dose–response information including the slope of the dose–response line should be used. Figure 5 demonstrates the dose–response curves for mortality for two chemicals.

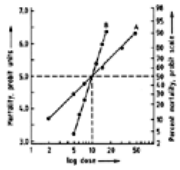

Figure 5. Comparison of dose – response relationships for two chemicals (log – probit plot)

Both chemicals have identical LD_{50} values, but different slopes of the dose – response curve

The LD_{50} of both chemicals is the same (10 mg/kg). However, the slopes of the dose–response curves are quite different. Chemical A exhibits a “flat” dose–response curve: a large change in dose is required before a significant change in response will be observed. In contrast, chemical B exhibits a “steep” dose–response curve, that is, a relatively small change in dose will cause a large change in response. The chemical with the steep slope may affect a much larger proportion of the population by incremental increases in dose than chemicals having a shallow slope; thus, acute overdosing may be a problem affecting the majority of a population for chemicals with steeper slopes. Chemicals with shallower slopes may represent a problem for the hyperreactive groups at the left-hand side of the dose–response curve. Effects may occur at significantly lower dose levels then for hyperreactive groups exposed to chemicals with a steep dose–response.

While the LD_{50} values characterize the potential hazard of a chemical, the risk of an exposure is determined by the hazard multiplied by the exposure dose. Thus, even very toxic chemicals like the poison of *Clostridium botulinum* pose only a low risk; intoxications with this compound are rare since exposure is low. Moreover, acute intoxications with other highly toxic agents such as mercury salts are rarely seen, despite detectable blood levels of mercury salts in the general population, since the dose is also low. On the other hand, compounds with low toxicity may pose a definite health risk when doses are high, for example, constituents of diet or chemicals formed during food preparation by heat treatment.

Therefore, for characterizing the toxic risk of a chemical, besides information on the toxicity, information on the conditions of exposure are necessary. When using LD_{50} values for toxicity characterisation, the limitations of LD_{50} values should be explicitly noted. These limitations include methodological pitfalls influenced by

1. Strain of animal used
2. Species of animal used
3. Route of administration
4. Animal housing

and intrinsic factors limiting the use of LD_{50} values

1. Statistical method
2. No dose–response curve
3. Time to toxic effect not determined
4. No information on chronic toxicity

The most serious limitation on the use of LD_{50} values for hazard characterization are the lack of information on chronic effects of a chemical and the lack of dose–response information. Chemicals with low acute toxicity may have carcinogenic or teratogenic effects at doses that do not induce acute toxic responses. Other limitations include insufficient information on toxic effects other than lethality, the cause of death, and the time to toxic effect. Moreover, LD_{50} values are not constant, but are influenced by many factors and may differ by almost one order of magnitude when determined in different laboratories.

1.8. Dose-Response Relationships for Cumulative Effects

After chronic exposure to a chemical, toxic response may be caused by doses not showing effects after single dosing. Chronic toxic responses are often based on accumulation of either the toxic effect or of the administered chemical. Accumulation of the administered chemical is observed when the rate of elimination of the chemical is lower than the rate of administration. Since the rate of elimination is dependent on plasma concentrations, after long-term application an equilibrium concentration of the chemical in the blood is reached. Chemicals may also be stored in fat (polychlorinated pesticides such as DDT) or bone (e.g., lead). Stored chemicals usually do not cause toxic effects because of their low concentrations at the site of toxic action (receptor). After continuous application, the capacities of

the storage tissues may become saturated, and xenobiotics may then be present in higher concentration in plasma and thus at the site of action; toxic responses result. Besides cumulation of the toxic agent, the toxic effect may also cumulate (Fig. 6).

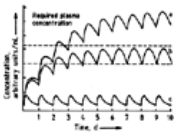

Figure 6. Accumulation of toxic chemicals based on their rate of excretion

a) The rate of excretion is equal to the rate of absorption, no accumulation occurs; b) Chemical accumulates due to a higher rate of uptake and inefficient excretion; the plasma concentrations are, however, not sufficient to exert toxic effects; c) The plasma concentrations reached after accumulation are sufficient to exert toxicity

For chemicals which irreversibly bind to macromolecules, the magnitude of toxic responses may be correlated with the total dose administered. In contrast to chemicals which act reversibly, the effect is not dependent on the frequency of dosing. Effect accumulation is often observed with carcinogens and ionizing radiation. In Figure 7 accumulation of effects is exemplified by the time- and dose-dependent induction of tumors by 4-(dimethylamino)azobenzene, a potent chemical carcinogen [5]. The TD_{50} values (50% of the treated animals carry tumors) are used to characterize the potency. Identical tumor incidences were observed after high doses and a short exposure time or after low doses and long exposure; the tumor incidence was only dependent on the total dose administered.

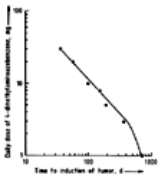

Figure 7. Time-dependent induction of tumors after different daily doses of 4-dimethylaminoazobenzene in rats [5]

Reversibility of toxic responses also depends on the capacity of an organ or tissue to repair injury. For example, kidney damage by xenobiotics is often, after survival of the acute phase of the intoxication, without further consequence due to the high capacity of the kidney for cell proliferation and thus the capacity to repair organ damage [6]. In contrast, injury to the central nervous system is largely irreversible since the differentiated cells of the nervous system cannot divide and dead cells cannot be replaced.

1.9. Factors Influencing Dose–Response

In animals and humans, the nature, severity, and incidence of toxic responses depend on a large number of exogenous and endogenous factors [7]. Important factors are the characteristics of exposure, the species and strain of animals used for the study, and interindividual variability in humans [8]. Toxic responses are caused by a series of complex interactions of a potentially toxic chemical with an organism. The type and magnitude of the toxic response is influenced by the concentration of the chemical at the receptor and by the type of interaction with the receptor. The concentration of a chemical at the site of action is influenced by the kinetics of uptake and elimination; since these are time-dependent phenomena, toxic responses are also time-dependent. Thus, the toxic response can be separated into two phases: toxicokinetics and toxicodynamics (Fig. 8).

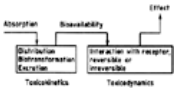

Figure 8. Toxicokinetics and toxicodynamics as factors influencing the toxic response

Toxicokinetics describe the time dependency of uptake, distribution, biotransformation, and excretion of a toxic agent (a detailed description of toxicokinetics is given in Section Biotransformation). Toxicodynamics describes the interaction of the toxic agent with the receptor and thus specific interactions of the agent (see below). Toxicokinetics may be heavily influenced by species, strain, and sex and the exposure characteristics [9-13]. Differences in toxic response between species, route of exposure, and others factors are often dependent on influences on toxicokinetics. Since toxicodynamics (mechanism of action) are assumed to be identical between species, this provides the basis for a rational interspecies extrapolation of toxic effects when differences in toxicokinetics are defined.

1.9.1. *Routes of Exposure*

The primary tissue or system by which a xenobiotic comes into contact with the body, and from where it may be absorbed in order to exert systemic toxicity, is the route of exposure. The frequent circumstances of environmental exposure are ingestion (peroral), inhalation, and skin contact. Also, for investigational and therapeutic purposes, intramuscular, intravenous, and subcutaneous injections may also be routes of exposure.

The major routes by which a potentially toxic chemical can enter the body are—in descending order of effectiveness for systemic delivery—injection, inhalation, absorption from the intestinal tract, and cutaneous absorption. The relationship between route and exposure, biotransformation, and potential for toxicity, may be complex and is also influenced by the magnitude and duration of dosing (Table 5).

Table 5. Toxicity of chemicals applied by different routes of exposure (data taken from [13])

Chemical	Species	Route of application	LD_{50}, mg/kg
DDT	rat	intravenous	68
	rat	oral	113
	rat	skin contact	1931
Atropine sulfate	rat	intravenous	41
	rat	oral	620
1-Chloro-2,4-dinitrobenzene	rat	oral	1070
	rat	intraperitoneal	280
	rabbit	skin contact	130
Dieldrin	rat	oral	46
	rat	intravenous	9
	rat	skin contact	10

The route of exposure has a major influence on toxicity because of the effect of route of exposure on the bioavailability of the toxic agent. The maximum tissue levels achieved, the time to maximum tissue levels, and thus the duration of the effect are determined by the rate of absorption and the extent of distribution within the system.

Direct injection into veins is usually restricted to therapeutic applications, but it is important for the toxicology of intravenously injected drugs in addicts. Chemicals applied by intravenous injection are rapidly distributed to well-perfused organs in the blood and thus may result in the rapid induction of toxic effects. The rapid dilution of a chemical after intravenous injection by venous blood permits even the injection of locally acting or corrosive chemicals which are well tolerated. The likelihood of toxicity from inhaled chemicals depends on a number of factors, of which the physical state and properties of the agent, concentration, and time and frequency of exposure are important. Major influences on the absorption and disposition of xenobiotics are exerted by species peculiarities since the anatomy of the respiratory tract and the physiology of respiration show major differences between rodents and humans. The water solubility of a gaseous xenobiotic has a major influence on penetration into the respiratory tract. As water solubility decreases and lipid solubility increases, penetration into deeper regions of the lung, the bronchioli, and the alveoli becomes more effective. Water-soluble molecules such as formaldehyde,are effectively scavenged by the upper respiratory tract and may have toxic effects on the eye and throat. In contrast, gases with low water solubility such as phosgene may penetrate through the bronchii and bronchioli to the

alveoli. Damage to the alveolar surface may initiate a series of events that finally results in lung edema. The degree to which inhaled gases, vapors, and particulates are absorbed, and hence their potential to produce systemic toxicity, depend on their diffusion rate through the alveolar mebrane, their solubility in blood and tissue fluids, the rate of respiration, and blood flow through the capillaries.

Uptake through the alimentary tract represents an important route of exposure for xenobiotics accumulated in the food chain, for natural constituents of human diet, and, drugs. Absorption from the gastrointestinal tract is dependent on the lipophilicity of a chemical, the molecular mass of the xenobiotic, and the presence of certain dietary constituents may influence the extent and rate of absorption. Chemicals absorbed from the gastrointestinal tract are transported to the liver via the portal vein; hepatic metabolism ("hepatic first-pass effect") may efficiently reduce the concentration of the xenobiotic available in the systemic circulation after oral uptake. Compounds undergoing bioactivation in the liver usually exhibit greater toxicity when given orally than when absorbed across the respiratory tract, due to the high proportion of material passing through the liver. In contrast, chemicals causing toxicity to extrahepatic, well-perfused organs such as the kidney often show a lower degree of toxicity to extrahepatic target organs when given orally.

Skin contact is an important route of exposure in the occupational and domestic environments. Local effects may include acute inflammation and corrosion, chronic inflammatory responses, immune-mediated reactions, and neoplasia. The percutaneous absorption of materials may also be a significant route for the absorption of systemically toxic materials. Factors influencing the percutaneous absorption of substances include skin site, integrity of skin, temperature, formulation, and physicochemical characteristics, including charge, molecular mass, and hydro- and lipophilicity.

1.9.2. *Frequency of Exposure*

The exposure of experimental animals may be categorized as acute, subacute, subchronic, and chronic. Acute exposures usually last less than 24 h, and all above-mentioned routes of exposure may be applied. With chemicals of low toxicity, repeated exposures may be used. Acute inhalation exposure is usually less than 24 h; frequently 4–8 h is chosen as timescale. Repeated exposure refers to application of the chemical for less than one month (subacute), one to three months (subchronic), and more than three months (chronic). Chronic exposures to detect specific toxic effects (carcinogenicity of a chemical) may span most of the lifetime of a rodent (up to two years). Repeated exposure may be by any route; the least labor intensive route is oral, by mixing the chemical with the diet; only for specific chemicals or to simulate likely routes of exposure for humans are application in drinking water, by gastric intubation, and by inhalation applied. These are more labor-intensive and require skilled personnel and/or sophisticated techniques and thus are more expensive.

The toxic effects observed after single exposure often are different form those seen after repeated exposure. For example, inhalation of high concentration of halothane causes anesthesia in animals and humans. In contrast, long-term application of halothane in lower doses causes liver damage in sensitive species The frequency of exposure in chronic studies is important for the temporal characterisation of exposure. Chemicals with slow rates of excretion may accumulate if applied at short dosing intervals, and toxic effects may result (see Section Types of Toxic Effects). Also, a chemical producing severe effect when given in a single high dose may have no detectable effects when given in several smaller doses. Interspecies and strain differences in susceptibility to chemical-induced toxicity may be due to heterogeneity of populations, species specific physiology (for example of the respiratory system), basal metabolic rate, size- and species-specific toxicokinetics and routes of metabolism or excretion (Table 6). In some cases, animal tests may give an underestimate, in others an overestimate, of potential toxicity to humans [14].

Table 6. Comparative LD_{50} values for four different chemicals in different animal species and estimated LD_{50} for humans

Chemical	Species	LD_{50}, mg/kg
Paraquat	rat	134
	mouse	77
	guinea pig	41
	human	32 – 48
Ethanol	rat	12 500
	mouse	8000
	guinea pig	5500
	human	3500 – 5000
Acetaminophen	rat	3763
	mouse	777
	guinea pig	2968
	human	42 800
Aspirin	rat	1683
	mouse	1769
	guinea pig	1102
	human	3492

1.9.3. *Species-Specific Differences in Toxicokinetics*

Species-specific differences in toxic response are largely due to difference in toxicokinetics and biotransformation. Distribution and elimination characteristics are quite variable between species. Both qualitative and quantitative differences in biotransformation may effect the sensitivity of a given species to a toxic response (Table 7).

Table 7. Species and sex differences in the acute toxicity of 1,1-dichloroethylene after oral administration and inhalation in rats and mice (data from World Health Organization, Geneva, 1990)

Species	Dosing criteria	Estimated LD_{50}/LC_{50}
Rat, male	inhalation/4 h	7000 – 32 000 mg/L
Rat, female	inhalation/4 h	10 300 mg/L
Mouse, male	inhalation/4 h	115 mg/L
Mouse, female	inhalation/4 h	205 mg/L
Rat, male	gavage	1550 mg/kg
Rat, female	gavage	1500 mg/kg
Mouse, male	gavage	201 – 235 mg/kg
Mouse, female	gavage	171 – 221 mg/kg

For example, the elimination half-life of 2,3,7,8-tetrachlorodibenzodioxin in rats is 20 d, and in humans it is estimated to be up to seven years [15]. An example for quantitative difference in the extent of biotransformation as a factor influencing toxic response is the species differences in the biotransformation of the inhalation anesthetic halothane. Both rats and guinea pigs metabolize halothane to trifluoroacetic acid, a reaction catalyzed by a specific cytochrome P450 enzyme [16-18]. As a metabolic intermediate, trifluoroacetyl chloride is formed, which may react with lysine residues in proteins and with phosphatidyl ethanolamine in phospholipids (Fig. 9).

Figure 9. Halothane metabolism by cytochrome P450 in rats, guinea pigs, and humans

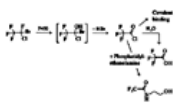

This interaction initiates a cascade of events finally resulting in toxicity. The metabolism of halothane in guinea pigs occurs at much higher rates than in rats, so guinea pigs are sensitive to halothane-induced hepatotoxic effects and rats are resistant. Qualitative differences in biotransformation are responsible for apparent differences in the sensitivity of rats and guinea pigs to the bladder carcinogenicity of 2-acetylamidofluorene. In rats, 2-acetylamidofluorene is metabolized by *N*-oxidation by certain cytochrome P450 enzymes. The *N*-oxide is further converted to an electrophilic nitrenium ion which interacts with DNA in the bladder; this biotransformation pathway explains the formation of bladder tumors in rats after long-term exposure to 2-acetylamidofluoren. In guinea pigs, 2-acetylamidofluorene is metabolized by oxidation at the aromatic ring; since nitrenium ions cannot be formed by this pathway, guinea pigs are resistant to the bladder carcinogenicity of 2-acetylamidofluorene (Fig. 10).

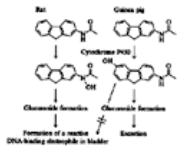

Figure 10. Biotransformation pathways of 2-acetylamidofluorene in rats and guinea pigs

With some chemicals, age may significantly affect toxicity, likely due to age related differences in toxicokinetics. The nutritional status may modify toxic response, likely by altering the concentration of cofactors needed for biotransformation and detoxication of toxic chemicals. Diet also markedly influences carcinogen-induced tumor incidence in animals [19] and may be a significant factor contributing to human cancer incidence.

The toxic response is influenced by the magnitude, number, and frequency of dosing. Thus, local or systemic toxicity produced by acute exposure may also occur by a cumulative process with repeated exposures to lower doses; also, additional toxicity may be seen in repeated-exposure situations. The relationships for cumulative toxicity by repetitive exposure compared with acute exposure toxicity may be complex, and the potential for cumulative toxicity from acute doses may not be quantitatively predictable. For repeated-exposure toxicity, the precise profiling of doses may significantly influence toxicity.

1.9.4. *Miscellaneous Factors Influencing the Magnitude of Toxic Responses*

A variety of other factors may affect the nature and exhibition of toxicity, depending on the conditions of the study, for example, housing conditions, handling, volume of dosing, vehicle, etc. Variability in test conditions and procedures may result in significant interlaboratory variability in results of otherwise standard procedures. For chemicals given orally or applied to the skin, toxicity may be modified by the presence of materials in formulations which facilitate or retard the absorption of the chemicals. With respiratory exposure to aerosols, particle size significantly determines the depth of penetration and deposition in the respiratory tract and thus the site and extent of the toxic effects.

1.10. Exposure to Mixtures

In experimental animals most data on the toxic effects of chemicals are collected after exposure to a single chemical; in contrast, human exposure normally occurs to mixtures of chemicals at low doses. Moreover, prior, coincidential, and sucessive exposure of humans to chemicals is likely. Interactions between the toxic effects of different chemicals are difficult to predict, effects of exposure to different chemicals may be independent, additive, potentiating (ethanol and carbon tetrachloride), antagonistic (interference with action of other chemical, e.g., as seen with antidotes administered in case of intoxications), and synergistic. Ethanol exerts a potentiating effect on the hepatotoxicity of carbon tetrachloride. In rats pretreated with ethanol, the hepatotoxic effects of carbon tetrachloride are much more pronounced than in control animals. This potentiation is due to an increased capacity for bioactivation (see Section Storage of Xenobiotics in Organs and Tissues) of carbon tetrachloride in pretreated rats due to increased concentrations of a cytochrome P450 enzyme in the liver [20]. Thus, an important considerations for the assessement of potential toxic effects of mixtures of chemicals are toxicokinetics and toxicodynamic interactions. Toxicokinetic interactions of chemicals may influence absorption, distribution, and biotransformation, both to active and inactive metabolites. Mixtures of solvents often show a competitive inhibition of biotransformation. Usually, one of the components has high affinity for a specific enzyme involved in its biotransformation, whereas another component has only a low affinity for that particular enzyme. Thus, preferential biotransformation of the component with the high affinity occurs. Different outcomes of enzyme inhibition are possible: if the toxic effects of the component whose metabolism is inhibited is dependent on bioactivation, lower rates of bioactivation will result in decreased toxicity; if the toxic effects are independent of biotransformation, the extent of toxicity will increase due to slower rate of excretion. Toxic effects of mixtures may also not be due to a major component, but to trace impurities with high toxicity. For example, many long-term effects seen in animal studies on the toxicity of chlorophenols are believed to be due to 2,3,7,8-Tetrachlorodibenzodioxin, which was present as a minor impurity in the samples of chlorophenols used for these studies.

[Top of Page]

2. Absorption, Distribution, Biotransformation and Elimination of Xenobiotics

2.1. Disposition of Xenobiotics

The induction of systemic toxicity usually results from a complex interaction between absorbed parent chemical and biotransformation products formed in tissues; the distribution of both parent chemical and biotransformation products in body fluids and tissues; their binding and storage characteristics; and their excretion.

The biological effects initiated by a xenobiotic are not related simply to its inherent toxic properties; the initiation, intensity, and duration of response are a function of numerous factors intrinsic to the biological system and the administered dose. Each factor influences the ultimate interaction of the xenobiotic and the active site (Section Factors Influencing Dose–Response). Only when the toxic chemical has reached the specific site and interacted with it can the inherent toxicity be realized. The route a xenobiotic follows from the point of administration or absorption to the site of action usually involves many steps and is termed toxicokinetics. Toxicokinetics influence the concentration of the xenobiotic or its active metabolite at the receptor. In the dose–response concept outlined in Section Dose–Response: a Fundamental Issue in Toxicology, it is generally assumed that the toxic response is proportional to the concentration of the xenobiotic at the receptor. However, the same dose of a chemical administered by different routes may cause different toxic effects. Moreover, the same dose of two different chemicals may result in vastly different concentrations of the chemical or its biotransformation products in a particular target organ. This differential pattern is due to differences in the disposition of a xenobiotic (Fig. 11).

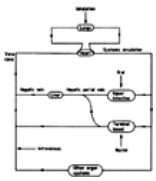

Figure 11. Possible fate of a xenobiotic in the organisms

The disposition of a xenobiotic consists of absorption, distribution, biotransformation, and excretion, which are all interrelated. The complicated interactions between the different processes of distribution are very important determinants of the concentration of a chemical at the receptor and thus of the magnitude of toxic response. They may also be major determinants for organ-specific toxicity.

For example, in the case of absorption of a xenobiotic through the gastrointestinal tract, the chemical proceeds from the intestinal lumen into the epithelial cells. Following intracellular transport, it passes through the basal membrane and lamina propria and enters the blood or lymph capillaries for transport to the site of action or storage. At that site, the xenobiotic is released from the capillaries, into an interstitial area, and finally through various membranes to its site of action, which may be a specific receptor, an enzyme, a membrane, or many other possible sites.

2.2. Absorption

The skin, the lungs, and the cells lining the alimentary tract are major barriers for chemicals present in the environment. Except for caustic chemicals, which act at the site of first contact with the organism, xenobiotics must cross these barriers to exert toxic effects on one or several target organs. The process whereby a xenobiotic moves through these barriers and enters the circulation is termed absorption.

2.2.1. *Membranes*

Because xenobiotics must often pass through membranes on their way to the receptor, it is important to understand membrane characteristics and the factors that permit transfer of foreign compounds. Membranes are initially encountered whether a xenobiotic is absorbed by the dermal, oral, or vapor route. These membranes may be associated with several layers of cells or a single cell. The absorption of a substance from the site of exposure may result from passive diffusion, facilitated diffusion, active transport, or the formation of transport vesicles (pinocytosis and phagocytosis). The process of absorption may be facilitated or retarded by a variety of factors; for example, elevated temperature increases percutaneous absorption by cutaneous vasodilation, and surface-active materials facilitate penetration. Each area of entry for xenobiotics into the organism may have specific peculiarities, but a unifying concept of biology is the basic similarity of all membranes in tissues, cells, and organelles.

All membranes are lipid bilayers with polar head groups (phosphatidylethanolamine, phosphatidylcholine). The polar groups predominate at the outer and inner surfaces of the membrane; the inner space of the membrane consists of perpendicularly arranged fatty acids [21]. The fatty acids do not have a rigid structure and are fluid under physiological conditions; the fluid character of the membrane is largely dominated by the fatty acid composition. The width of a biological membrane is approximately 7–9 nm. Figure 12 illustrates the concept of a biological membrane (fluid-mosaic model).

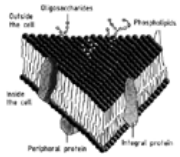

Figure 12. Simplified model of the structure of a biological membrane

Proteins are intimately associated with the membrane and may be located on the surface or inside the membrane structure, or extend completely through the membrane. These proteins may also form aqueous pores. Hydrophobic forces are responsible for maintaining the structural integrity of both proteins and lipids within the membrane structure. The ratio of lipid to protein in different membranes may vary from 5:1 (e.g,. myelin) to 1:5 (e.g., the inner membrane of mitochondria). Usually, pore diameters in membranes are small and permit only the passage of low molecular mass chemicals. However, some specialized membranes such as those found in the glomeruli of the kidney, which can have pore sizes of up to 4 nm, also permit the passage of compounds with molecular mass greater than 10 000.

The amphipathic nature of the membrane creates a barrier for ionized, highly polar compounds; however, changes in lipid composition, alterations in the shape and size of proteins, and physical features of bonding may cause changes in the permeability of membranes [22].

2.2.2. *Penetration of Membranes by Chemicals*

A chemical can pass through a membrane by two general processes: passive diffusion and active transport. Passive diffusion is described by Fick's law and requires no energy. Active transport processes involve the consumption of cellular energy to translocate the chemical across the membrane. Active transport may also act against a concentration gradient and result in the accumulation of a xenobiotic in a specific organ, cell type or organelle.

Diffusion of Chemicals through Membranes. Many toxic chemicals pass membranes by simple diffusion. Their rates of diffusion depend on their lipid solubility and are often correlated with the partition coefficient (solubility in organic solvents/solubility in water). Lipophilic chemicals may diffuse directly through the lipid domain of the membrane. However, a certain degree of water solubility seems to be required for passage since many poorly lipid soluble chemicals have been shown to penetrate easily. Once initial penetration has occurred, the molecule must necessarily traverse a more polar region to dissociate from the membrane. Compounds with extremely high partition coefficients thus tend to remain in membranes and to accumulate there rather than pass through them. Polar compounds that are insoluble in the nonpolar, fatty-acid-containing inner space of the membrane often cannot penetrate membranes, although some low molecular mass polar chemicals may slowly penetrate through the aqueous pores of the membranes.

The rates of movement of nonpolar xenobiotics through membranes can be predicted based on the assumptions from Fick's law of diffusion. Polar compounds and electrolytes of low molecular mass are believed to behave similarily. A first-order equation appears to be applicable to the majority of xenobiotics. The rate of diffusion of a xenobiotic is related to its concentration gradient across the membrane ($C_1 - C_2$), the surface area available for transfer A, the diameter of the membrane d, and the diffusion constant k. The latter is related to the size and structure of the molecule, the spatial configuration of the molecule, and the degree of ionization and lipid solubility of the xenobiotic.

$$\text{Rate of diffusion} = k\frac{A(C_1 - C_2)}{d}$$

As the xenobiotic is rapidly removed after absorption, C_2 can usually be ignored. and a log/linear plot of the amount of unpenetrated chemicals present over time should be linear. When relatively comparable methods have been used, calculation of the half-time of penetration $t_{1/2}$, is useful. The rate constant of penetration k is derived from

$$k = \frac{0.693}{t_{1/2}}$$

When the half-time of penetration after oral and dermal administration of several environmental contaminants were compared, rates were found to vary considerably. Clearly, rates of penetration by different routes in mammals show little or no correlation.

Ionization becomes particularly important when xenobiotics are introduced into the gastrointestinal tract, where a variety of pH conditions are manifest (see Section Gastrointestinal Absorption). Although many drugs are acids and bases and thus potentially ionizable form, most xenobiotics are neither acids nor bases and thus are unaffected by pH. The amount of a xenobiotic in the ionized or unionized form depends upon the pK_a of the xenobiotic and the pH of the medium. When the pH of a solution is equal to the pK_a of the dissolved compound, 50% of the acid or base exists in the ionized and 50% in the unionized form. The degree of ionization at a specific pH is given by the Henderson–Hasselbalch equation:

$$pK_a - \text{pH} = \log\frac{[\text{nonionized}]}{[\text{ionized}]}$$

$$pK_a - \text{pH} = \log\frac{[\text{ionized}]}{[\text{nonionized}]}$$

Since the unionized, lipid-soluble form of a weak acid or base may penetrate membranes, weak organic acids diffuse most readily in an acidic environment, and organic bases in a basic environment. There is some degree of penetration even when xenobiotics are not in the most lipid-soluble form, and a small amount of absorption can produce serious effects if a compound is very toxic.

2.2.3. *Mechanisms of Transport of Xenobiotics through Membranes*

Filtration. Passage of a solution across a porous membrane results in the retention of solutes larger than the pores. This process is termed filtration. For example, filtration of solutes occurs in the kidney glomeruli, which have large pores and retain molecules with molecular masses greater than 10 000. Elsewhere in the body, filtration by pores may only result in the passage of relatively small molecules (molecular mass ca. 100), and most larger molecules are excluded. Thus, uptake of xenobiotics through these pores is only a minor mechanism of penetration.

Special Transport Mechanisms. Special transport processes include active transport, facilitated transport, and endocytosis (Table 8). Often, the movement of chemicals across membranes is not due to simple diffusion or filtration. Even some very large or very polar molecules may readily pass through membranes.

Table 8. Special transport processes involved in the passage of xenobiotics through biological membranes

Type of transport	Carrier molecule required	Examples of substrates	Energy required	Against concentration gradient
Active transport	yes	organic acids in the kidney	yes	yes
Facilitated transport	yes	glucose	yes	no
Endocytosis	no	proteins	yes	?

Active transport systems have frequently been implicated in these phenomena. *Active transport* may be effected by systems that help transport endogenous compounds across membranes. Such processes require energy and transport xenobiotics against electrochemical or concentration gradients. Active transport systems are saturable processes and exhibit a maximum rate of transport; they are usually specific for certain structural features of chemicals. A carrier molecule (likely a protein) associates with the chemical outside the cell, translocates it across the membrane for ultimate release inside the cell. This is particularly important for compounds that lack sufficient lipid solubility to move rapidly through the membrane by simple diffusion. Active transport plays a major role in the excretion of xenobiotics from the body, and major excretory organs such as the liver or the kidney have several transport systems which may accept organic acids, organic bases, or even metal ions as substrates.

In contrast to other special transport processes, some carrier-mediated processes do not require energy and are unable to move chemicals against a concentration gradient. These processes are termed *facilitated transport*. Facilitated transport is particulary beneficial for compounds which lack sufficient lipid solubility for rapid diffusion through the membrane. Facilitated transport is more rapid than simple diffusion up to the point at which concentrations on both sides of the membranes are equal. For example, the transport of glucose through a variety of membranes occurs by facilitated transport. The mechanisms by which facilitated transport occurs are not well understood.

Pinocytosis (liquids) and *phagocytosis* (solids) are specialized processes in which the cell membrane invaginates or flows around a xenobiotic, usually present in particulate form, and thus enables transfer across a membrane. Although of importance once the xenobiotic has gained entry into the organism, this mechanism does not appear to be of importance in the initial absorption of a xenobiotic.

2.2.4. *Absorption*

Absorption is the process whereby xenobiotics cross body membranes and are translocated to the blood stream. The primary sites of absorption of environmental contaminants are the gastrointestinal tract (gastrointestinal absorption), the skin (dermal absorption), and the lung (respiratory absorption). Absorption of chemicals may also occur from other sites such as muscle, the subcutis, or the peritoneum after administration by special routes. In clinical medicine, many drugs are injected directly into the bloodstream to circumvent the problems of absorption posed by the peculiarities of the different routes.

2.2.4.1. *Dermal Absorption*

Human skin can come into contact with many potentially toxic chemicals. Skin is relatively impermeable to aqueous solutions and most xenobiotics present as ions. Therefore, it is a relatively good barrier separating the human body from the environment. However, skin is permeable in varying degrees to a large number of xenobiotics, and some chemicals may be absorbed through the skin in sufficient amounts to cause a toxic response [23]. A striking example of the significance of absorption through the skin is the large number of agricultural workers who have experienced acute poisoning from exposure to parathion (dermal $LD_{50} \approx 20$ mg/kg) during application or from exposure to vegetation previously treated with this pesticide.

The human skin is a complex, multilayered tissue with approximately 18 000 cm^2 of surface in an average human male. Chemicals to be absorbed must pass through several cell layers before entering the small blood and lymph capillaries in the dermis. Transport in blood and lymph then distributes absorbed chemicals in the body. The human skin consists of three distinct layers (Fig. 13) and a number of associated appendages (sweat and sebaceous glands, hair follicles).

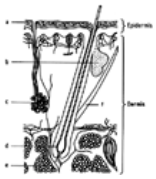

Figure 13. Cross section of human skin

a) Stratum corneum; b) Sebaceous gland; c) Sweat gland; d) Hair follicle; e) Fat; f) Muscle

The *epidermis* is a multilayered tissue varying in thickness from 0.15 (eyelids) to 0.8 mm (palms). This tissue appears to be the greatest deterrent to the absorption of xenobiotics. The epithelial tissues of the skin develop and grow divergently from other tissues. Proliferative layers of the basal cells (stratum germinativum) differentiate and gradually replace cells above them as surface cells deteriorate and are sloughed from the epidermis. Cells in this layer produce fibrous, insoluble keratin that fills the cells, and a sulfur-rich amorphous protein that comprises the cell matrix and thickened cell membrane. This cell layer, the stratum corneum, provides the primary barrier to the penetration of foreign compounds. It consists of several layers of flattened, stratified, highly keratinized cells. These cells are approximately 25–40 µm wide and have lost their nuclei. Although highly water retarding, the dead, keratinized cells of the stratum corneum are highly water absorbent (hydrophilic), a property that keeps the skin supple and soft. A natural oil covering the skin, the sebum, appears to maintain the water-holding capacity of the epidermis but has no appreciable role in retarding the penetration of xenobiotics. The rate-determining barrier in the chemical absorption of xenobiotics is the stratum corneum.

The *dermis* and *subcutaneous tissue* offer little resistance to penetration, and once a substance has penetrated the epidermis these tissues are rapidly traversed. The dermis is a highly vascular area that provides ready access to blood and lymph for distribution once the epithelial barrier has been passed. The blood supply in the dermis is subjected to complex, interacting neural and humoral influences whose temperature-regulating function can have an effect on distribution by altering blood supply to this area. Therefore, the extent of absorption of a chemical through the skin may be influenced by temperature, and relative humidity [24].

The *skin appendages* are found in the dermis and extend through the epidermis. The primary appendages are the sweat glands (epicrine and apocrine), hair, and sebaceous glands. These structures extend to the outer surface and therefore may play a role in the penetration of xenobiotics; however, since they represent only 0.1 to 1% of the total surface of the skin, their contribution to overall dermal absorption is usually minor.

Percutaneous absorption can occur by several routes, but the majority of unionized, lipid-soluble xenobiotics appear to move by passive diffusion directly through the cells of the stratum corneum. Important arguments for the importance of transepidermal absorption are that epidermal damage or removal of the stratum corneum increases permeability, the epidermal penetration rate equals whole-skin penetration, epidermal penetration is markedly slower than dermal, and the epidermal surface area is 100–1000 times the surface area of the skin appendages. Very small and/or polar molecules appear to have more favorable penetration through appendages or other diffusion shunts, but only a small fraction of toxic xenobiotics are chemicals of this type. Polar substances, in addition to movement through shunts, may diffuse through the outer surface of the protein filaments of the hydrated stratum corneum, while nonpolar molecules dissolve in and diffuse through the nonaqueous lipid matrix between the protein filaments.

Human stratum corneum displays significant differences in structure from one region of the body to the other, which affect the rate of absorption. Penetration at certain body regions thus varies according to the polarity and size of the molecule, but it is generally accepted that for most unionized xenobiotics the rate of penetration is in the following order: scrotal > forehead > axilla = scalp > back = abdomen > palm and plantar. The palm and plantar regions are highly diffuse, but their much greater thickness (100–400 times that of other regions) introduces an overall lag time in diffusion.

The condition of the skin greatly influences the absorption of xenobiotics. Damage to or removal of the stratum corneum cause a dramatic increase in the permeability of the epidermis for xenobiotics. Caustic and corrosive chemicals such as acids or alkali or burns will greatly enhance dermal absorption and thus influence the toxicity of a xenobiotic applied to the skin. Soaps and detergents are among the damaging substances routinely applied to skin. Whereas organic solvents must be applied in high concentrations to damage skin, 1% aqueous solutions of detergents increase the rate of penetration of solutes through human epidermis dramatically. Alteration of the stratum corneum by organic solvents may also be the cause of increased penetration.

Organic solvents can be divided into damaging and nondamaging categories. Damaging solvents include methanol, acetone, diethyl ether, hexane, and some solvent mixtures. These solvents and mixtures can extract lipids and proteolipids from tissues and are thus expected to alter permeability. Although the mechanical strength of the stratum corneum is unaltered, delipidization produces a more porous, nonselective surface. Solvents such as higher alcohols, esters, and olive oil do not appear to damage skin appreciably. On the contrary, the penetration rate of solutes dissolved in them is often reduced. Surprisingly, lipid-soluble xenobiotics may be markedly resistant to washing, even a short time after application. For example, 15 min after application, a substantial portion of parathion cannot be removed from

contaminated skin by soap and water.

When comparisons across species are made, human skin appears to be more impermeable, or at least as impermeable, as the skin of the cat, dog, rat, mouse, or guinea pig. The skin of pigs and guinea pigs in particular serves as a useful approximation to human skin, but only after a comparison has been made for each specific chemical.

Temperature, surface area of applied dose, simultaneous application of another xenobiotic, relative humidity, occlusion, age, and hyperthermia are among a number of chemical, physical, and physiological factors that may alter skin penetration.

2.2.4.2. *Gastrointestinal Absorption*

The oral route of entry into the body is specially important for accidental or purposeful (suicide) ingestion of poisonous materials. Food additives, food toxins, environmental xenobiotics accumulated in the food chain, and airborne particles excluded from passage to to alveoli are also introduced into the digestive system. The penetration of orally administered xenobiotics is primarily confined to the stomach and intestine [25].

The gastrointestinal tract may be viewed as a tube traversing the body. It consists of the mouth, esophagus, stomach, small and large intestine, colon, and rectum (Fig. 14). The digestive tract is lined by a single layer of columnar cells, usually protected by mucus, which do not present a barrier to penetration. The circulatory system is closely associated with the intestinal tract (30–50 µm from membrane to vasculature), and once xenobiotics have crossed the epithelium of the intestinal tract, entry into capillaries is rapid. Venous blood flow from the stomach and intestine rapidly removes absorbed xenobiotics and introduces them into the hepatic portal vein, which transports them to the liver.

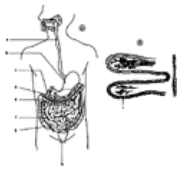

Figure 14. A) Intestinal tract in humans; B) Anatomy of the intestinal wall, the major site of absorption of xenobiotics

The lining of the small intestine is highly folded and has a special surface structure (brush-border membrane) to give a large surface available for the efficient uptake of nutrients.

a) Esophagus (4 – 7.2); b) Stomach (1.0 – 3.0); c) Duodenum (4.8 – 8.2); d) Pancreas; e) Colon (7.9 – 8.0); f) Jejunum (7.6); g) Ileum (7.6); h) Rectum (7.8); i) Brush-border membrane

Numbers in brackets represent pH in different parts of the intestinal tract.

Absorption of chemicals may take place along the entire gastrointestinal tract, but most xenobiotics are absorbed in the stomach and the small intestine. A major factor favoring absorption in the intestine is the presence of microvilli that increase the surface area to an estimated 100 m^2 in the small intestine (see Fig. 14) Because the intestinal area thus offers maximal opportunity for absorption, it is generally accepted that absorption of xenobiotics is greatest in this area of the gastrointestinal tract. Although the gastrointestinal tract has some special transport processes for the absorption of nutrients and electrolytes, most xenobiotics seem to enter the body from the gastrointestinal tract by simple diffusion. Exeptions are some heavy metals such as thallium and lead, which mimic the essential metals iron and calcium, respectively. They are thus absorbed by active transport systems developed for the uptake of these nutrients.

The gastrointestinal tract has areas of highly variable pH, which can markedly change the permeability characteristics of ionic compounds. For example, passive diffusion is greatly limited except for unionized, lipid-soluble chemicals. Although variable according to secretory activity, the pH of the stomach is ca. 1–3 and that of the intestine ca. 7. The measured pH of the intestinal contents may not be the same as the pH of the epithelium at the site of absorption, and this explains the entrance of compounds whose p*K*a would suggest less favored absorption. The variations in pH in the different sections of the intestinal tract may influence the absorption of acids and bases. Since most xenobiotics are absorbed by diffusion, only the unionized, membrane-permeable form may be absorbed. Weak organic acids are mainly present in the unionized, lipid-soluble form in the stomach, and predominantly in the ionized form in the intestine. Therefore, organic acids are expected to be more readily absorbed from the stomach than from the intestine. In contrast, weak organic bases are ionized in the stomach but present in the lipid-soluble form (unionized) in the intestine. Absorption of such compounds should therefore predominantly occur in the intestine rather than in the stomach.

However, other factors determining the rate of membrane penetration such as surface area available for diffusion, blood flow (influencing concentration gradients), and the law of mass action also influence the site of absorption of acids or bases from the gastrointestinal tract. For example, although only 1% of benzoic acid is present in the lipid-soluble, unionized form in the small intestine, the large surface area and the rapid removal of absorbed benzoic acid with the blood result in its efficient absorption from the small intestine.

Other factors contribute to gastrointestinal absorption Clearly a xenobiotic must be dissolved before absorption can take place. Particle size, organic solvents, emulsifiers, and rate of dissolution thus also effect absorption. In addition, the presence of microorganisms and hydrolysis-promoting pH offer opportunities for the biotransformation of many xenobiotics. Other factors affecting gastrointestinal absorption include binding to gut contents, intestinal motility, rate of emptying, temperature of food, effects of dietary constituents, health status of the individual, and gastrointestinal secretion.

2.2.4.3. *Absorption of Xenobiotics by the Respiratory System*

The respiratory system is an organ in direct contact with environmental air as an unavoidable part of living. A number of xenobiotics exist in gaseous (carbon monoxide, nitric oxides), vapor (benzene, carbon tetrachloride), and aerosol (lead from automobile exhaust, silica, asbestos) forms and are potential candidates for entry via the respiratory system. Indeed, the most important cause of death from acute intoxication (carbon monoxide) and the most frequent occupational disease (silicosis) are caused by the absorption or deposition of airborne xenobiotics in the lung.

The respiratory tract consists of three major regions: the nasopharyngeal, the tracheobronchial, and the pulmonary (Fig. 15). The nasopharynx begins in the mouth and extents down to the level of the larynx. The trachea, bronchii, and bronchioli serve as conducting airways between the nasopharynx and the alveoli, the site of gas exchange between the inhaled air and the blood. The human respiratory system is a complex organ containing over 40 different cell types. These cell types contribute to the pulmonary architecture and function over various zones of the lung, although to some extent, individual cell types can be found in several zones. The tracheobronchial system comprises airways lined with bronchial epithelium with associated submucosal glands and several different tissues with specific function and the lung vasculature.

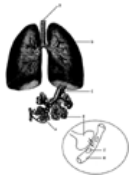

Figure 15. Anatomy of the human respiratory system

a) Trachea; b) Bronchii; c) Bronchioli; d) Alveoli; e) Capillary; f) Erythrocyte

The absorption of xenobiotics by the respiratory route is favored by the short path of diffusion, large surface area (50–100 m^2), and large concentration gradients. At the alveoli (site of gas exchange), the membranes are very thin (1–2 µm) and are intimately associated with the vascular system. This enables rapid exchange of gases (ca. 5 ms for CO_2 and ca. 200 ms for O_2). A thin film of fluid lining the alveolar walls aids in the initial absorption of xenobiotics from the alveolar air. Simple diffusion accounts for the somewhat complex series of events in the lung regarding gas absorption. The sequences of respiration, which involve several interrelated air volumes, define both the capacity of the lung and factors important to particle deposition and retention. Among the elements important in total lung capacity is the residual volume, that is, the amount of air retained by the lung despite maximal expiratory effort. Largely due to slow release from this volume, gaseous xenobiotics in the expired air are not cleared immediately, and many expirations may be necessary to rid the air in the lung of residual xenobiotic. The rate of entry of vapor-phase xenobiotics is controlled by the alveolar ventilation rate, and a xenobiotic present in alveolar air may come into contact with the alveoli in an interrupted fashion about 20 times per minute. The diffusion coefficient of the gas in the fluids of pulmonary membranes is another important consideration, but doses are more appropriately discussed in terms of the partial pressure of the xenobiotic in the inspired air. On inhalation of a constant tension of a gaseous xenobiotic, arterial plasma tension of the gas approaches the tension of gas in the expired air. The rate of entry is then determined by blood solubility of the xenobiotic and blood flow. For a high blood/gas partition coefficient, a larger amount must be dissolved in the blood to raise the partial pressure. Chemicals with a high blood/gas partition coefficient require a longer period to approach the same tension in the blood as in inspired air than less soluble gases.

Aerosols and Particulates. The entry of aerosols and particulates is affected by a number of factors. A coal miner inhales ca. 6000 g of coal dust particles during his occupational lifetime, and only ca. 100 g are found postmortem; therefore, effective protective mechanisms are operative. The parameters of air velocity and directional changes in air flow favor impaction of particles in the upper respiratory systems. Particle characteristics such as size, chemistry of the inhaled material,

sedimentation and electrical charge are important to retention, absorption, or expulsion of airborne particles. In addition to the other aforementioned lung characteristics, a mucous blanket propelled by ciliary action clears the respiratory tract of particles by directing them to the gastrointestinal system (via the glottis) or to the mouth for expectoration. This system is responsible for 80% of lung particulate clearance. The deposition of various particle sizes in different respiratory regions is summarized in Figure 16, which shows that particle size is important for disposition and particles larger than 2 µm do not reach the alveoli [26].

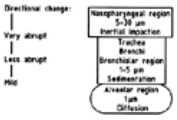

Figure 16. Effect of size on the disposition and sedimentation of particulates in the respiratory tract

The site of particle sedimentation is determined largely by particle size; only very fine particles are deposited in the alveoli; larger particles do not reach the lung but are deposited in the nasopharynx.

The direct penetration of airborne xenobiotics at alveolar surfaces or in the upper respiratory tract is not the only action of toxicological importance. Both vapors and particulates can accumulate in upper respiratory passages to produce irritant effects. Irritant gases may be deposited in the respiratory tract depending on their water solubility and may cause localized damage characterized by edema, swelling, mucus production, and increased d vascular permeability. If major airways are obstructed by these processes or important anatomical structures of the respiratory tract like the alveoli are damaged, life-threatening or deadly intoxications may be caused by the inhalation of irritant gases.

Despite the effectiveness of ciliary movement and phagocytosis, the cumulative effects of silica, asbestos, or coal dust ultimately cause chronic fibrosis even though direct absorption is of minor importance. Thus, phagocytosis prevents acute damage but may contribute to chronic toxicity. There is little evidence for active transport in the respiratory system, although pinocytosis may be of importance for penetration. The lung is an area of extensive metabolic activity; enzymes present in the lung may catalyze both activation and detoxication of xenobiotics (see Section Storage of Xenobiotics in Organs and Tissues).

2.3. Distribution of Xenobiotics by Body Fluids

After entering the blood by absorption or by intravenous administration, xenobiotics are available for distribution throughout the body. The initial rate of distribution to organs and tissues is determined by the blood flow to that organ and the rate of diffusion of the chemical into the specific organ or tissue. Uptake of xenobiotics into organs or tissues may occur by either passive diffusion or by special transport processes. Within tissues binding, storage, and/or biotransformation may occur. After reaching equilibrium, the distribution of a chemical among organs and tissues is largely determined by affinity; blood flow determines distribution only during the initial phase shortly after uptake.

Body fluids are distributed between three distinct compartments: vascular water, interstitial water and intracellular water. Plasma water and interstitial water are extracellular water. Plasma water plays an important role in the distribution of xenobiotics. Human plasma accounts for about 4% of the total body weight and 53% of the total volume of blood. By comparison, the interstitial tissue fluids account for 13% of body weight, and intracellular fluids account for 41%. The concentration of a xenobiotic in blood following exposure will depend largely on its apparent volume of distribution. If the xenobiotic is distributed only in the plasma, a high concentration will be achieved within the vascular tissue. In contrast, the concentration will be markedly lower if the same quantity of xenobiotic were distributed in a larger pool including the interstitial water and/or intracellular water.

Among the factors that affect distribution, apart from binding to blood macromolecules, are the route of administration, rate of biotransformation, polarity of the parent xenobiotic or biotransformation products, and rate of excretion by the liver or kidneys. Gastrointestinal absorption and intraperitoneal administration provide for immediate passage of a compound to the liver, whereas dermal or respiratory routes involve at least one passage through the systemic circulation prior to reaching the liver. The metabolism of most xenobiotics results in products that are more polar and thus more readily excreted than the parent molecules. Therefore, the rate of metabolism is a critical determinant in the distribution of a compound, since compounds that are readily metabolized are usually readily excreted, and thus are proportionally less prone to accumulate in certain tissues. The same principle applies to polarity, since very polar xenobiotics will be readily excreted. Chemicals may circulate either free or bound to plasma protein or blood cells; the degree of binding and factors influencing the equilibrium with the free form may influence availability for biotransformation, storage, and/or excretion [27].

Patterns of xenobiotic distribution reflect certain physiological properties of the organism and the physicochemical properties of the xenobiotics. An initial phase of distribution may be distinguished that reflects cardiac output and blood flow to organs. Heart, liver, kidney, brain, and other well-perfused organ- receive most of a lipophilic xenobiotic within the first few minutes after absorption. Delivery to the smooth muscles, most viscera, and skin is slower, and the time to reach a steady-state concentration of a xenobiotic in these organs may be several hours. A second phase of xenobiotic distribution may therefore be distinguished; it is limited by blood flow to an organ or tissue and involves a far larger fraction of body mass than the first phase of distribution (Fig. 17).

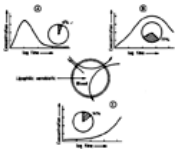

Figure 17. Uptake and redistribution with blood of lipophilic xenobiotics

Lipophilic xenobiotics in the blood are first distributed to well-perfused organs (A); after some time, they are redistributed to organs with lower blood flow representing a larger fraction of the body weight (B, C)

Only a limited number of xenobiotics have sufficient solubility in blood to account for simple dissolution as a route of distribution; the distribution of many xenobiotics occurs in association with plasma proteins. The binding of drugs to plasma proteins is of key importance in transport. Many organic and inorganic compounds of low molecular mass appear to bind to lipoproteins, albumins, and other proteins in plasma and are transported as protein conjugates. This binding is reversible. Cellular components may also be responsible for transport of xenobiotics, but such transport is rarely a major route. The transport of xenobiotics by lymph is usually quantitatively of little importance since the intestinal blood flow is 500–700 times greater than the intestinal lymph flow.

A large number of studies on binding of drugs by plasma protein have demonstrated that binding to serum albumin is particularly important for these chemicals. Only few studies on the reversible binding of toxic xenobiotics have been performed, but available evidence suggests a significant role of lipoproteins in plasma. These plasma proteins may bind xenobiotics as well as some physiological constituents of the body. Examples for plasma proteins which may bind xenobiotics are albumin, α- and β-lipoproteins, and metal-binding proteins such as transferrin. Lipoproteins are important for the transport of lipid-soluble endogenous chemicals such as vitamins, steroid hormones, and cholesterol, but they may also bind lipophilic xenobiotics. If a xenobiotic is bound to a protein, it is immobilized remote from the site of action. The extent of binding to plasma proteins varies considerably among xenobiotics. While some are not at all bound, for others more than 90% of administered dose may be bound to plasma proteins. These ligand–protein interactions are reversible and provide a remarkably efficient means for transport of xenobiotics to various tissues.The xenobiotic–protein interaction may be simply described according to the law of mass action as:

$$[T]_F + [\text{free sites}] \underset{k_2}{\overset{k_1}{\rightleftharpoons}} [T]_B$$

where $[T]_F$ and $[T]_B$ are the concentrations of free and bound xenobiotic molecules, respectively, and k_1 and k_2 are the rate constants for association and dissociation; k_2, which governs the rate of binding to the protein, dictates the rate of xenobiotic release at a site of action or storage. The ratio k_1/k_2 is identical with the dissociation constant K_{diss}. Among a group of binding sites on proteins, those with the smallest K_{diss} for a given xenobiotic will bind it most tightly.

In contrast to the covalent binding to proteins seen with many xenobiotics or their electrophilic metabolites (see Section Interaction of Reactive Intermediates with Cellular Macromolecules), the interaction of xenobiotics with plasma proteins is most often noncovalent and reversible. Noncovalent binding is of primary importance with respect to distribution because of the opportunities to dissociate after transport. Binding of xenobiotics to plasma proteins may be due to several types of interactions which are summarized in the following.

Ionic Binding. Electrostatic attraction occurs between two oppositely charged ions. The degree of binding varies with the chemical nature of each compound and the net charge. Dissociation of ionic bonds usually occurs readily, but some transition metals exhibit high association constants (low K_{diss} values), and exchange is slow. Ionic interactions may also contribute to binding of alkaloids with ionizable nitrogen groups and other ionizable xenobiotics.

Hydrogen Bonding. Generally, only the most electronegative atoms form stable hydrogen bonds. Protein side chains containing hydroxyl, amino, carboxyl, imidazole, and carbamyl groups may form hydrogen bonds, as can the nitrogen and oxygen atoms of peptide bonds. Hydrogen bonding plays an important role in the

structural configuration of proteins and nucleic acids.

Van der Waals forces are very weak interactions between the nucleus of one atom and the electrons of another atom, i.e., between dipoles and induced dipoles. The attractive forces are based on slight distortions induced in the electron clouds surrounding each nucleus as two atoms come close together. The binding force is critically dependent upon the proximity of interacting atoms and diminishes rapidly with distance. However, when these forces are summed over a large number of interacting atoms that "fit" together spatially, they can play a significant role in determining specificity of xenobiotic–protein interactions.

Hydrophobic Interactions. When two nonpolar groups come together they exclude the water between them, and this mutual repulsion of water results in a hydrophobic interaction. In the aggregate they represent the least possible disruption of interactions among polar water molecules and thus can lead to stable complexes. Some consider this a special case of van der Waals forces. The minimization of thermodynamically unfavorable contact of a polar group with water molecules provides the major stabilizing effect in hydrophobic interactions.

Consequences of the binding to plasma proteins are reduced availability of the free xenobiotic in the cells and a delayed excretion. The xenobiotic bound to plasma protein cannot cross capillary walls due to its high molecular mass. The fraction of dose bound is thus not available for delivery to the extravascular space or for filtration by the kidney. It is generally accepted that the fraction of xenobiotic that is bound may not exert toxic effects; however, many xenobiotics and endogenous compounds appear to compete for the same binding site, and thus one compound may alter the unbound fraction of another by displacement, thereby potentially increasing toxic effects. Plasma proteins that can bind endogenous chemicals and xenobiotics are listed below, together with examples of bonded xenobiotics:

α-Lipoproteins	vitamins A, K, D
	steroid hormones
	dieldrin
Albumin	salicylate
	tetracyclines
	phenols
	vitamin C

Binding of xenobiotics to plasma proteins is mostly reversible; the bound xenobiotic is in equilibrium with free xenobiotic, and thus binding usually slows excretion or delivery to cellular sites of action. Toxicological consequences of the reversible binding of a xenobiotic to plasma proteins may arise after saturation of the binding capacities of plasma proteins and by displacement of the bound xenobiotic by another chemical with higher affinity, which increases the free fraction of the formerly bound xenobiotic. This will result in an increased equilibrium concentration of the xenobiotic in plasma and in the target organ, with potentially harmful consequences.

2.4. Storage of Xenobiotics in Organs and Tissues

Absorbed xenobiotics may be concentrated in specific organs or tissues. The concentration of a xenobiotic in a specific tissue may cause toxic effects to that particular tissue; some xenobiotics actually attain their highest concentration at the site of toxic action. However, other xenobiotics may be concentrated in tissues without harmful consequences. Some tissues have a high capacity to accommodate certain xenobiotics and may release them only slowly. The compartment or tissue in which a chemical is concentrated can also be considered as a storage depot for this xenobiotic. If a chemical is stored in a depot and thus removed from the site of action (e.g., polychlorinated biphenyls in fat or lead in bone), no immediate manifestation of toxicity may be observed, even though a potential for adverse effects exists. For example, lead stored in bone does not cause a toxic response but has the potential for mobilization and thus for migration into soft tissues; toxic effects may appear after mobilization. As the xenobiotic in storage depots is in equilibrium with the free xenobiotic in plasma, mobilization is constant, and exposure of the target organ to low concentrations of the xenobiotic is constant. Some storage depots for specific chemicals follow:

Lead	bone
Fluoride	bone, teeth
Cadmium	kidney
Iron	transferrin (a blood protein)
Polychlorinated pesticides such as DDT	fat
Arsenic	skin

Liver, kidney, fat, bone, and plasma proteins may serve as storage depots for absorbed xenobiotics. Both liver and kidney have a high capacity to store xenobiotics and are major storage sites for a multitude of chemicals. Accumulation of circulating xenobiotics form the blood by active transport systems and binding to certain tissue constituents are major mechanisms involved in the renal and hepatic storage.

Several thiol-rich proteins present in liver and kidney have a high affinity for xenobiotics [28]. The binding protein ligandin, which as a glutathione *S*-transferase also has enzymatic activity (Section Glutathione Conjugation of Xenobiotics and Mercapturic Acid Excretion) and thus participates in xenobiotic biotransformation, binds organic acids, some azo dyes, and corticosteroids. Metallothionein, a cysteine-rich protein present in liver and kidney, serves as a binding and storage protein for metals such as cadmium and zinc. Its biosynthesis increases after exposure to metals,and this may result in storage of a considerable percentage of cumulative metal dose as metalothionein complex in liver and kidney.

Highly lipophilic chemicals rapidly penetrate membranes and are thus efficiently taken up by tissues. Lipophilic substances that are inefficiently biotransformed accumulate in the most lipophilic environment in the organism, fat. Most xenobiotics seem to accumulate by physical dissolution in neutral fats, which may constitute between 20 and 50% of the body weight in human males. Large amounts of lipophilic xenobiotics may therefore be present in fat; for xenobiotics that do not undergo biotransformation (e.g., 2,3,7,8-tetrachlorodibenzodioxin), determination of the concentration in body fat is a good measure for exposure.

Usually, xenobiotics stored in fat also do not induce toxic responses, since the xenobiotic is not readily available at the target site. However, during rapid mobilization of fat during disease or starvation, a sudden increase in the plasma concentration, and thus toxic effects in target organs, may occur. For example, signs of organochlorine pesticide intoxication have been observed after starvation in animals pretreated with persistent organochlorine pesticides.

Some xenobiotics have a high affinity for bone and may accumulate in the bone matrix. For example, 90% of the lead and a major part of the strontium present in the body after chronic exposure are stored in the skeleton [29]. Lead and strontium accumulate in bone due to their similarity with calcium; inorganic fluoride, which is also a "bone-seeker", accumulates in bone due to similarities in size and charge to the hydroxyl ion. The storage of xenobiotics in bone may or may not be responsible for toxic effects. Lead stored in the skeleton is not toxic to bone, but both stored fluoride and stored strontium cause toxic effects to bone (fluorosis, osteosarcoma). Xenobiotics stored in bone are also in equilibrium with the unbound xenobiotic circulating in plasma and may thus be released.

Certain plasma proteins have a high affinity for xenobiotics, binding of a chemical to plasma proteins may constitute a both transport form and a storage form. Globulins such as transferrin (involved in iron transport) and ceruloplasmin (copper) and α- and β-lipoproteins (lipophilic xenobiotics and endogenous chemicals) may be involved in binding.

Storage in tissue may greatly alter the rate of excretion of a xenobiotic. Only xenobiotics present in plasma are available for distribution, biotransformation, and excretion. However, excretion or biotransformation alters the plasma concentration of the xenobiotic, and some of the stored chemical is released into plasma from the site of storage. Owing to this mechanism, the rate of excretion of a xenobiotic stored in tissues can be very low.

2.5. Biotransformation

Most xenobiotics entering the body are lipophilic. This property enables them to penetrate lipid membranes, to be transported by lipoproteins in the blood, and to be rapidly absorbed by the target organ. However, the efficient excretory mechanisms of the organism require solubility of the xenobiotic in aqueous media and thus a certain degree of hydrophilicity is required for efficient excretion. Lipophilic substances can only be excreted efficiently by exhalation, but this is restricted to volatile xenobiotics. In the absence of efficient means for excretion of nonvolatile chemicals, constant exposure or even intermittent single exposures to a lipophilic chemical could result in accumulation of the xenobiotic in the organism.

Therefore, animal organisms have developed a number of biochemical processes that convert lipophilic chemicals to hydrophilic chemicals and thus assist in their excretion. These enzymatic processes are termed biotransformation, and the enzymes catalyzing biotransformation reactions are referred to as biotransformationenzymes. The biotransformation enzymes differ from most other enzymes by having a broad substrate specifity and by catalyzing reactions at

comparatively low rates. The low rates of biotransformation reactions are often compensated by high concentrations of biotransformation enzymes. For example, ca. 5 wt% of the protein in rat liver consist of cytochromes P450, which are major biotransformation enzymes.

The broad specifity of the biotransformation enzymes likely has evolutionary reasons. Biotransformation enzymes have evolved to facilitate the excretion of lipophilic chemicals present in the diet of animals. The broad substrate specificity helped to adjust to new dietary constituents and thus led to evolutionary advantages.

Biotransformation is generally the sum of several processes by which the structure of a chemical is changed during passage through the organism. The metabolites formed from the parent chemical are usually more water soluble; the increased water solubility reduces the ability of the metabolites to partition into membranes, restricts renal and intestinal reabsorption, and thus facilitates excretion with urine or bile.

2.5.1. *Phase-I and Phase-II Reactions*

Xenobiotic metabolism is catalyzed by a number of different enzymes. For solely operational purposes, the biotransformation enzymes are separated into two phases. In phase-I reactions, which involve oxidation, reduction, and hydrolysis, a polar group is added to the xenobiotic or is exposed by the biotransformation enzymes. Phase-II reactions are biosynthetic and link the metabolite formed by phase-I reactions to a polar endogenous molecule to produce a conjugate. Various endogenous molecules with high polarity and are utilized for conjugation; the resulting conjugates are often ionized at physiological pH and thus highly water soluble. Moreover, the moieties used for conjugation are often recognized by specific active transport processes, which assist in their translocation across plasma membranes and thus further enhance the rate of excretion.

The fate of a particular chemical and the participation of the various phase-I and phase-II biotransformation enzymes is determined by its chemical structure; biotransformation is usually complex and often integrated. Many chemicals bearing functional groups undergo conjugation without prior phase-I biotransformation, whereas others are oxidized or reduced prior to conjugation. However, chemicals lacking functional groups may also undergo phase-II biotransformations without being subjected to a prior phase-I reaction (examples are 1,2-dibromoethane and perchloroethene; see Section Phase-I Enzymes and their Reactions).

2.5.2. *Localization of the Biotransformation Enzymes*

The biotransformation enzymes are localized mainly in the liver. A significant fraction of the blood from the splanchnic area, which also contains xenobiotics absorbed from the intestine, enters the liver. Therefore, the liver has developed the capacity to enzymatically modify most of these chemicals before storage, release, or excretion. However, most other tissues also have the capacity to catalyse biotransformation reactions; indeed, most tissues tested have shown the presence of enzymes which can catalyze biotransformation reactions. The contribution of extrahepatic organs to the biotransformation of a chemical depends on many factors, including chemical structure, dose, and route of administration. However, biotransformation of a chemical within an extrahepatic tissue may have toxic effects on this specific tissue and may thus have important toxicological consequences.

Inside cells, phase-I enzymes are mainly present in the endoplasmatic reticulum, a myriad of lipoprotein membranes extending from the mitochondria and the nucleus to the plasma membranes of the cell. When an organ is homogenized, the endoplasmatic reticulum is broken and membrane fragments are sealed of to form microvesicles. These microvesicles can be sedimented by differential centrifugation, and the materials thus obtained is known as a microsomes. They are highly enriched in vesicles from the endoplasmatic reticulum and retain active biotransformation enzymes. Microsomes are often used to study the enzymatic biotransformation of xenobiotics in vitro.

The presence of phase-I enzymes within membranes has important implications, since lipophilic chemicals will preferentially distribute into lipid membranes; thus, high concentrations of lipophilic xenobiotics are present at this site of biotransformation. In contrast to phase-I enzymes, phase-II enzymes are often soluble, non-membrane-associated, and present in the cytoplasm of the cell. They are found in the supernatant (cytosol) obtained by ultracentrifugation of homogenized tissues. The subcellular localizations of enzymes responsible for biotransformation afollow:

Phase-I enzymes	
–Cytochrome P450	microsomal
–Flavin-dependent mono-oxygenase	microsomal
–Prostaglandin synthase	microsomal
–Epoxide hydrolase	microsomal/cytosolic
Phase-II enzymes	
–UDP-glucuronyl-transferases	microsomal
–Sulfotransferases	cytosolic
–*N*-acetyltransferases	cytosolic
–Glutathione *S*-transferase	cytosolic/microsomal

2.5.3. *Role of Biotransformation in Detoxication and Bioactivation*

The general purpose of biotransformation reactions is detoxication, since xenobiotics should be transformed to metabolites which are more readily excreted. However, depending on the structure of the chemical and the enzyme catalyzing the biotransformation reaction, metabolites with a higher potential for toxicity than the parent compound are often formed. This process is termed bioactivation and is the basis for the toxicity and carcinogenicity of many xenobiotics with a low chemical reactivity (see Section Bioactivation of Xenobiotics). The interaction of the toxic metabolite initiates events that ultimately may result in cell death, cancer, teratogenicity, organ failure, and other manifestations of toxicity. Formation of reactive and more toxic metabolites is more frequently associated with phase-I reactions; however, phase-II reactions and combinations of phase-I and phase-II reactions may also be involved in toxication.

2.5.4. *Phase-I Enzymes and their Reactions*

Phase-I reactions are catalyzed by microsomal monooxygenases and peroxidases, cytosolic and mitochondrial oxidases, reductases, and hydrolytic enzymes. All these reactions add or expose functional groups which can be conjugated later.

2.5.4.1. *Microsomal Monooxygenases: Cytochrome P450*

Microsomal monooxygenases are the cytochrome P450 enzymes and the mixed-function amine oxidase or flavin-dependent monooxygenase. Both enzyme systems add a hydroxyl moiety to the xenobiotic. Cytochrome P450, a carbon monoxide binding hemoprotein in microsomes, is the most important enzyme system involved in phase-I reactions. The name cytochrome P450 is a generic term applied to a group of hemoproteins defined by the unique spectral property observed when reduced cytochrome P450 (Fe^{2+}) is treated with carbon monoxide. The complex formed has a maximum absorption at 450 nm imparted by the presence of an axial thiolate ligand on the heme iron atom. This spectral characteristic is only present when the protein is intact and catalytically functional. Denatured cytochrome P450 shows, like other heme proteins, an absorbance maximum at 420 nm.

Cytochrome P450 enzymes are a coupled enzyme system composed of the heme-containing cytochrome P450 and the NADPH-containing cytochrome P450 reductase [30]. This flavoprotein has a preference for NADPH as its cofactor and transfers one or two electrons from NADPH to cytochrome P450. Cytochrome P450 and the reductase are embedded into the phospholipid matrix of the endoplasmatic reticulum. The phospholipid matrix is crucial for enzymatic activity since it facilitates the interaction between the two enzymes. The importance of the phospholipid matrix is indicated by the following: Both cytochrome P450 and cytochrome P450 reductase can be purified to apparent homogeneity; the enzymatic activity of the purified and recombined enzymes is dependent on the addition of phospholipids.

In vertebrates, the highest concentrations of cytochrome P450 are found in the liver, but cytochrome P450 enzymes are also present in lung [31-35], kidney, testes, skin, and gastrointestinal tract [36]. The presence of several forms of cytochrome P450 with different substrate specificity and different rates of biotransformation for certain xenobiotics was indicated by studies in the 1970s. In the early 1980s, several different cytochrome P450 enzymes from rodents were purified to apparent homogenicity. Moreover, a large number of cytochrome P450 enzymes have been purified from human organs. All these cytochrome P450 enzymes share the heme, but they differ in both the composition and thus the structure of the polypeptide chain and in the reactions they catalyze [37-40].

The individual enzymes are regulated in their expression by a variety of factors such as treatment with xenobiotics, species, organ, sex, and diet. Because of the multitude of enzymes present, the term “superfamily” of cytochromes P450 is frequently used. In mammals, two general classes of cytochrome P450 exist: six families involved in steroid metabolism and bile acid biosynthesis, and four families containing numerous individual cytochromes P450, mainly responsible for xenobiotic biotransformation. A complex nomenclature, based on amino acid sequence similarity, has been developed to designate individual cytochromes P450. The

genes for the individual enzymes are named by the root CYP followed by a number designating the family, a letter for the subfamily, and another number denoting the individual enzyme (see Table 9).

Table 9. Mammalian cytochromes P450 involved in xenobiotic biotransformation

Family	Number of subfamilies	Number of forms
CYP1	1	2
CYP2	8	59
CYP3	2	11
CYP4	2	10

Reactions catalyzed by cytochrome P450. Cytochromes P450 are monooxygenases. They utilize one of the oxygen atoms of molecular oxygen and incorporate it into the xenobiotic RH:

$$RH + O_2 + NADPH + H^+ \rightarrow ROH + H_2O + NADP^+$$

The other oxygen atom is reduced to water with consumption of NADPH as reducing cofactor. The likely mechanisms of electron transfer and xenobiotic oxidation is shown in Figure 18.

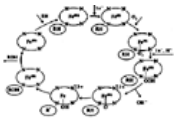

Figure 18. Mechanisms of electron transfer and xenobiotic oxidation by cytochrome P450

In the first step of the catalytic cycle, the xenobiotic combines with the oxidized form of cytochrome P450 (Fe^{3+}) followed by a one-electron reduction by NADPH-cytochrome P450 reductase to form a reduced (Fe^{2+}) cytochrome P450-substrate complex. This complex then combines with molecular oxygen, and another electron from NADPH is accepted. In a series of further steps, which are not completely understood, an oxygen atom from the intermediate is transferred to the substrate, while the other oxygen atom is reduced to water. In the last step of the catalytic cycle, the oxidized substrate dissociates and regenerates the oxidized form of cytochrome P450. Examples for oxidation reactions catalyzed by cytochromes P450 are shown in Table 10.

Table 10. Oxidations catalyzed by cytochrome P450

Type of reaction	Examples
Aliphatic hydroxylation	$R-CH_2-CH_2-CH_3 \xrightarrow[P450]{O_2} R-CH_2-CHOH-CH_3$
N-Dealkylation	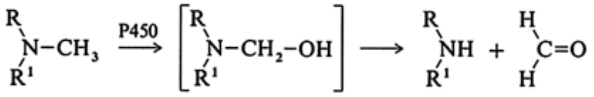
O-Dealkylation	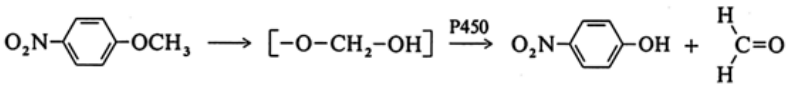
Epoxidation	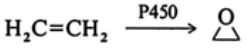

Cytochromes P450 may catalyze the hydroxylation of carbon–hydrogen bonds to transform hydrocarbons to the corresponding alcohols. In larger aliphatic chains, the (ω – 1) position is often a favored point of attack. Oxidative *N*-, *O*-, or *S*-dealkylation and oxidative dehalogenation are similar in mechanism to aliphatic hydroxylation, but, due to further reactions of the intermediate products, give different end products. Olefins are also oxidized by cytochrome P450, and with some substrates, epoxides are formed as products. The reaction, however, does not proceed in a concerted manner, but involves discrete ionic intermediates, which may also rearrange to products other than epoxides, as shown for chloroolefins in Figure 19.

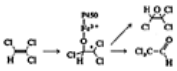

Figure 19. Mechanism of oxidation and rearrangement of trichloroethylene to chloral and trichlorooxirane, respectively

Oxidation at sulfur or nitrogen occurs by the addition of oxygen at the unshared electron pair on the sulfur or nitrogen atom. The products formed (sulfoxides or hydroxylamines) may be stable (many sulfoxides), may be further oxidized by other enzymes in the organism (e.g., hydroxylamines), or may decompose to sulfur and the corresponding oxo compound.

The above-mentioned reactions may be catalyzed by most cytochromes P450 involved in xenobiotic biotransformation; the type of reaction catalyzed seems to be more influenced by steric factors regarding the substrate-binding site of individual cytochromes P450 than by electronic factors. As shown in Table 11, which gives an overview of human cytochrome P450s involved in the biotransformation of xenobiotics and drugs, steric factors are likely major determinants of the substrate specificity of cytochrome P450 enzymes.

Table 11. Human cytochromes P450 identified as major catalysts in the biotransformation of specific xenobiotics that seem to play major roles in the oxidation of substrates listed

Cytochrome P450			
1A1	**1A2**	**2E1**	**3A4**
Benzo[*a*]pyrene	phenacetin	vinyl chloride	aflatoxin B_1
Other polycyclic hydrocarbons	1-aminofluorene	trichloroethylene	17- β-estradiol
	2-amino-3-methylimidazo-[4,5-*f*]quinoline	halothane benzene	6-aminochrysene sterigmatocystine
	2-naphthylamine	dimethylnitrosamine	nifedipine
		acetaminophen	ethinylestradiol

In addition to promoting oxidative metabolism, cytochrome P450 may also catalyze reductive biotransformation reactions [41, 42]. These reaction are favored under reduced oxygen pressure or occur with xenobiotics lacking oxidizable C—H bonds or olefinic moieties. In these cases,instead of oxygen, the xenobiotic accepts one or two electrons from NADPH-cytochrome P450 reductase or from cytochrome P450. Reductive biotransformation catalyzed by cytochrome P450 has been demonstrated with some azo dyes and several aromatic nitro compounds. The double bond in azo compounds may be progressively reduced to give amine metabolites; aromatic nitro groups may also be reduced via the nitrone and the hydroxylamine to the corresponding amine.

The reductive biotransformation of polyhalogenated alkanes is exemplified by the one-electron reduction of carbon tetrachloride to the trichloromethyl radical and chloride; reductive biotransformation of carbon tetrachloride by a two-electron reduction results in formation of chloroform:

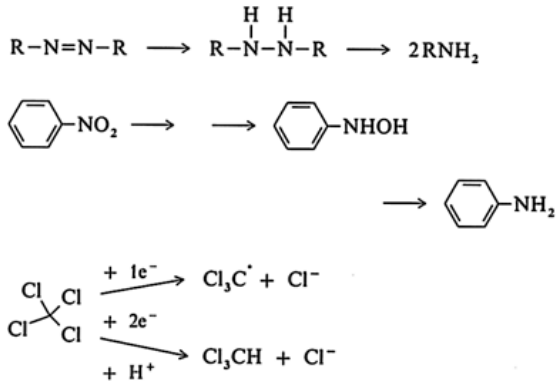

2.5.4.2. *Microsomal Monooxygenases: Flavin-Dependent Monooxygenases*

Tertiary amines and sulfur-containing drugs have been known to be metabolized to *N*-oxides or sulfoxides by a microsomal monooxygenase which is not dependent on cytochrome P450 [43, 44]. This enzyme, which is historically referred to as mixed-function amine oxidase, is a flavoprotein that is present in the endoplasmatic reticulum. It is capable of oxidizing nucleophilic nitrogen and sulfur atoms in xenobiotics. However, this enzyme shows a catalytic mechanism different from those of other heme- or flavin-containing enzymes. Like other monooxygenases, flavin-containing monooxygenases require molecular oxygen and NADPH as cofactors for oxygenation. In contrast to the other monooxygenases, flavin-dependent monooxygenases do not contain heme or iron, and the binding of the substrate is not required for the generation of the enzyme bound oxygenating intermediate (Fig. 20).

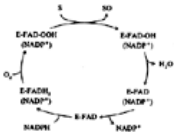

Figure 20. Mechanism of xenobiotic oxidation by the flavin-dependent monooxygenase

FAD = flavineadenine dinucleotide

The active, oxygenating form of the enzyme is present in the cell, and any soft, oxidizable nucleophile that can gain access to the enzyme-bound oxygenating intermediate will be oxidized. Precise fit of substrate to the enzyme is not necessary. This property seems to be largely responsible for the broad substrate specificity of flavin-dependent monooxygenases. Flavin-dependent monooxygenase catalyzes the oxidation of a wide variety of xenobiotics with few, if any, common structural features at maximum rate (Table 12).

Table 12. Oxidations catalyzed by flavin-dependent monooxygenases

Type	Examples
Amine oxidation	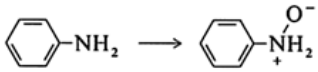
Hydroxylamine oxidation	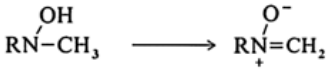
Thioamide oxidation	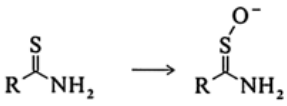
Thiol oxidation	$2R-SH \longrightarrow R-S-S-R$
Disulfide oxidation	$R-S-S-R \longrightarrow 2R-SO_2$

Many essential xenobiotics also bear functionalities that are oxidized by flavin-dependent monooxygenases. However, these enzymes apparently discriminate between physiologically essential and xenobiotic soft nucleophiles and seem to exclude the former.

As with cytochrome P450, species- and tissue-specific forms of flavin-dependent monooxygenase have been described. Species differences in hepatic flavin-dependent monooxygenase seem to be quantitative rather than qualitative, whereas tissue specific forms in the same species are clearly distinct enzymes. For example, hepatic and pulmonary flavin-dependent monooxygenases in rabbits exhibit distinct, but overlapping substrate specificities and are different gene products. Recent studies have shown that several isoforms of flavin-dependent monooxygenase exist; all isoforms show differences in their distribution in species and in organs within species [45].

The regulation of expression of flavin-dependent monooxygenase seems to be complex. Xenobiotics which were shown to increase the concentration of cytochrome P450 in mammals did not influence flavin-dependent monooxygenase concentrations. Recent evidence suggest that soft nitrogen and sulfur nucleophiles in the diet may act as inducers of flavin-dependent monooxygenases. Since the dietary inducers are continously taken up, flavin-dependent monooxygenases are present in maximal concentrations in rats on commercial rat chow.

2.5.4.3. *Peroxidative Biotransformation: Prostaglandin-synthase*

During the biosynthesis of prostaglandins, a polyunsaturated fatty acid, arachidonic acid, is oxidized to yield prostaglandin G_2, a hydroperoxy endoperoxide. This is further transformed to prostaglandin H_2. Both the formation of prostaglandin G_2 and its further transformation into prostaglandin H_2 are catalyzed by the same enzyme, prostaglandin synthase [46-48]. This enzyme is a glycoprotein with a molecular mass of approximately 70 000 Dalton and contains one heme per subunit. The enzyme is found in high concentrations in germinal vesicles and renal medulla, but also in several other tissues such as skin and adrenals [49]. The cyclooxygenase and peroxidase activity of prostaglandin synthase generate enzyme- and substrate-derived free-radical intermediates (Fig. 21).

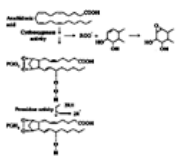

Figure 21. Cooxidation of xenobiotics during the biosynthesis of prostaglandins

Biotransformation of xenobiotics may be associated both with the cyclooxygenase and the peroxidase activity of prostaglandin synthase (Fig. 21). During the cyclooxygenase-catalyzed conversion of arachidonic acid to prostaglandin G_2 peroxy radicals are formed as intermediates. These lipid peroxy radicals represent a source of reactive oxygen metabolites (see below) and can in turn biotransform xenobiotics. Oxidation by the cyclooxygenase activity of prostaglandin synthase is important in the oxidation of diols derived from carcinogenic polycyclic aromatic hydrocarbons and transforms these diols to diol epoxides (Fig. 21, top).

During the reduction of the peroxide prostaglandin G_2 to prostaglandin H_2, the peroxidase undergoes a two-electron oxidation. To return to the ground state the enzyme requires two consecutive one-electron reductions, which are achieved by abstracting electrons from available donors. In addition to endogenous substrates, xenobiotics may act as electron donors and may thus be oxidized to radicals (Fig. 21, bottom). This process is termed cooxidation of xenobiotics. Classes of xenobiotics that undergo cooxidation during prostaglandin syntheses are aromatic amines, phenols, hydroquinones, and aminophenols. The role of prostaglandin synthase in the biotransformation of xenobiotics is somewhat unclear, since many of the end products of prostaglandin-synthase-mediated cooxidation of xenobiotic are identical to those formed by cytochrome P450. Therefore, it is assumed that prostaglandin synthase may contribute to the oxidative biotransformation of xenobiotics in tissues low in monooxygenase activity [50].

In addition to prostaglandin synthase, other peroxidases may also participate in the oxidation of xenobiotics [51-53]. For example, mammary gland epithelium contains lactoperoxidase and leucocytes contains myeloperoxidase. The general reaction catalyzed by this enzymes involves the reduction of hydroperoxide coupled to the oxidation of the substrate:

$$\mathrm{ROOH + XH \rightarrow ROH + XOH}$$

The availability of peroxides in tissues likely controls the extent of peroxidative biotransformation; however, the availability of hydrogen peroxide is usually low due to efficient scavenging by catalase and glutathione peroxidase. Therefore, peroxidative metabolism occurs mainly in tissues which can maintain an oxidizing environment.

2.5.4.4. *Nonmicrosomal Oxidations*

Several enzymes located in mitochondria or the cytoplasm of the cell may also catalyze the oxidation of xenobiotics. In contrast to cytochromes P450 with their broad substrate specificity, most of the nonmicrosomal oxidases have a more narrow substrate specificity and accept only xenobiotics bearing specific functional groups as substrates. Alcohol dehydrogenases catalyze the oxidation of alcohols to aldehydes or ketones:

$$\mathrm{RCH_2OH + NAD^+ \rightarrow RCHO + NADH + H^+}$$

The enzyme is mainly found in the soluble fraction of liver, but also in other organs such as the kidney and lung and is responsible for the oxidation of ethanol. The expression of alcohol dehydrogenase is under genetic control, which gives rise to a number of variants with differing activities. Usually, the oxidation of alcohols to aldehydes is reversible, since the reduction of aldehydes is also efficiently catalyzed by aldehyde reductases. However, in vivo, the reaction proceeds in the direction of alcohol consumption since aldehydes are further oxidized by aldehyde dehydrogenases. These enzymes catalyze the formation of acids from aliphatic and aromatic aldehydes:

$$\mathrm{RCHO + NAD^+ \rightarrow RCOOH + NADH + H^+}$$

The reaction may be catalyzed by aldehyde dehydrogenase, which has broad substrate specificity, and several isozymes of aldehyde dehydrogenase are found in liver cytosol, mitochondria, and microsomes with characteristic substrate specificities. Other enzymes in the soluble fraction of liver that can oxidize aldehydes are the flavoproteins aldehyde oxidase and xanthine oxidase.

Monoamine oxidases are a family of flavoproteins present in many tissues including liver, kidney, brain, and intestine. These mitochondrial enzymes have a broad and overlapping substrate specificity and oxidize a variety of amines:

$$\mathrm{RCH_2NH_2 + O_2 + H_2O \rightarrow RCHO + NH_3 + H_2O}$$

The monoamine oxidase found in the central nervous system is concerned primarily with neurotransmitter turnover.

2.5.4.5. *Hydrolytic Enzymes in Phase-I Biotransformation Reactions*

Many tissues contain enzymes with carboxylesterase and amidase activity. These enzymes are located both in microsomal and soluble fraction and hydrolyse ester and amide linkages in xenobiotics.

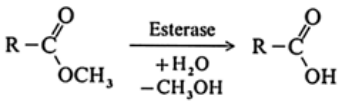

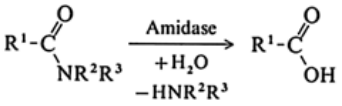

Although esterases and amidases were thought to be different enzymes, all purified esterases have been demonstrated to have amidase activity; similarly, all amidases have esterase activity. In general, esters are cleaved more rapidly than amides. The expression of many esterases is under genetic control; thus, extremes of high/low esterase activity and resistance/sensitivity to toxic effects mediated by esterases are known.

Epoxide hydrolase is an important enzyme cleaving aliphatic and aromatic epoxides. The enzyme hydrates arene oxides and aliphatic epoxides to the corresponding *trans*-1,2-diols [54]. Water is required as cofactor, and the catalytic mechanism of epoxide hydrolases involves ester formation of the oxirane with a carboxylic acid function at the active site of the enzyme and hydrolysis of this ester by water; no metals or other cofactors are required.

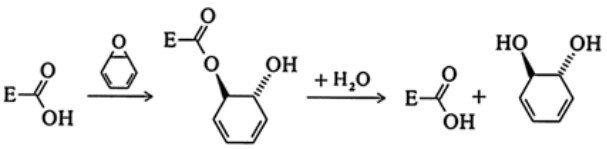

Microsomal epoxide hydrolases are thought to be present in close proximity to the microsomal cytochromes P450; in most cases, the conversion of the epoxide to the less reactive diol is considered to represent an important detoxication reaction for metabolically formed oxiranes. Epoxide hydrolases are found in many tissues such as liver, kidney, testes, and intestine. Their distribution is heterogenous between different cell types in a specific organ; in addition, several forms of microsomal epoxide hydrolases with broad substrate specificity have been found in different animal species. Moreover, in animals, in addition to membrane-bound epoxide hydrolase, a soluble epoxide hydrolase is present in cytoplasm of several tissues [55].

2.5.5. ***Phase-II Biotransformation Enzymes and their Reactions***

Products of the phase-I biotransformation reactions carrying functional groups such as hydroxyl, amino, or carboxyl often undergo a conjugation reaction with an endogenous substrate. The endogenous substrates may include sugar derivatives, sulfate, amino acids, and small peptides (glutathione). The conjugation products are usually more polar and thus more readily excreted than their parent compounds.

In contrast to phase-I biotransformation reactions, phase-II reactions are biosynthetic and require energy to drive the reaction. Energy is usually consumed to generate a high-energy cofactor or an activated intermediate and then utilized as cosubstrate. Thus, depletion of the cofactor or general interference with cellular energy status may interfere with the ability of cells to conduct phase-II biotransformation reactions.

2.5.5.1. *UDP-Glucuronyl Transferases*

Glucuronidation represents one or the main phase-II biotransformation reactions in the conversion of both endogenous and exogenous compounds to water soluble products [56-59]. The formed glucuronides are excreted with bile or urine. The formation of an activated glucuronide (uridine diphosphate glucuronic acid, UDPGA), is required for glucuronide formation. UDPGA is formed in a sequential reaction from uridine and glucose-1-phosphate.

$$\text{Uridine triphosphate} + \text{Glucose 1-phosphate} \xrightarrow[\text{Pyrophosphorylase}]{\text{UDPG}} \text{Uridine diphosphate glucose} + \text{Pyrophosphate}$$

$$\text{Uridine diphosphate glucose} + 2\,NAD^+ + H_2O \xrightarrow{\text{Dehydrogenase}} UDPGA + 2\,NADH_2$$

The enzymes that carry out the coupling of the xenobiotic with UDPGA are termed UDP-glucuronyl transferases. They couple D-glucuronic acid with a wide variety of xenobiotics carrying functional groups to give β-D-glucuronides [57]. These glucuronides are highly polar and ionized at physiological pH, and hence are rapidly excreted. The membrane-bound UDP-glucuronyl transferases are found in highest concentration in the liver, but also in most other tissues studied. The reaction catalyzed involves a nucleophilic displacement (S_N2) of the functional group of the substrate with Walden inversion (Fig. 22).

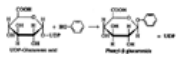

Figure 22. Conjugation of phenol to phenyl glucuronide catalyzed by UDP-glucuronyltransferase (UDP = uridine diphosphate)

UDP-glucuronyl transferases, like cytochrome P450, represent a family of enzymes, and at least ten individual forms are known. The various forms respond differently to inducers and have preferences for certain classes of chemicals. UDP-glucuronyl transferases catalyze the conjugation of numerous functional groups of xenobiotics with glucuronic acids. Some typical examples are shown in Table 13.

Table 13. Typical conjugation reactions catalyzed by UDP-glucuronyltransferases

Type of reaction	Examples
O-Glucuronide formation	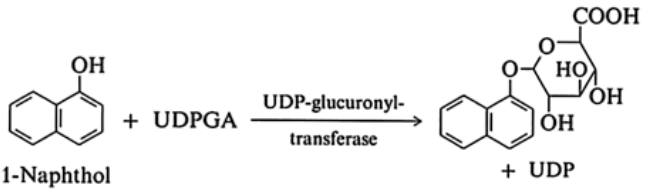
N-Glucuronide formation	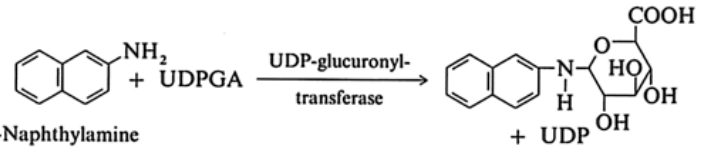
S-Glucuronide formation	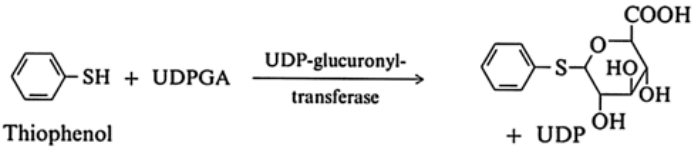

Glucuronides formed in the liver are excreted with urine or bile. Aglycones with molecular masses higher than 300 are transformed to glucuronides that surpass the molecular mass threshold for biliary excretion (see Section Hepatic Excretion) and will thus be excreted with bile into the intestine. Glucuronides may by cleaved there by β-glucuronidase present in intestinal microflora to the respective aglycone, which may be reabsorbed from the intestine and translocated back to the liver with the blood. The resulting cycle is called enterohepatic circulation; compounds that undergo enterohepatic circulation are only slowly excreted and usually have a longer half-life in the body.

2.5.5.2. *Sulfate Conjugation*

The formation of water-soluble sulfate esters is observed with many xenobiotics carrying functional groups such as alcohols, phenols, and arylamines. These reactions are catalyzed by sulfotransferases, a large group of soluble enzymes found in many tissues [60-62]. Sulfotransferases catalyze the transfer of a sulfate group from the “active sulfate” 3′-phosphoadenosine-5′-phosphosulfate to hydroxyl groups and amines. The resulting products are referred as sulfate esters of sulfamates.

The products of this reaction are ionized at physiological pH and may therefore be rapidly excreted in urine; thus, sulfate conjugation is an effective mechanism to enhance the rate of excretion of many xenobiotics. Sulfotransferases are also a family of enzymes, and at least four different classes of sulfotransferases are involved in biotransformation reactions: sulfotransferases, hydroxysteroid sulfotransferases, estrone sulfotransferases, and bile salt sulfotransferases; each class again has been divided into several distinct forms differing in substrate specificity, optimum pH and immunological properties.

Sulfate conjugation also requires a sulfate donor, 3′-phospho-adenosine-5′-phosphosulfate. 3′-Phospho-adenosine-5′-phosphosulfate is likely synthesized in the cytosol of most mammalian cells by a two-step reaction consuming ATP and utilizing inorganic sulfate originating from the catabolism of cysteine or from diet. In the first step of this sequence the sulfation of ATP to adenosine 5′-phosphosulfate is catalyzed by an ATP sulfurylase. Adenosine 5′-phosphosulfate is further transformed to 3′-phospho-adenosine 5′-phosphosulfate by an adenosine 5′-phosphosulfate kinase. However, the equilibrium concentration of 3′-phosphoadenosine 5′-phosphosulfate in mammalian cells may be low, and due to the tight coupling of two enzymes in 3′-phosphoadenosine 5′-phosphosulfate, biosynthesis may proceed rapidly.

$$SO_4^{2-} + ATP \xrightarrow{\text{Sulfurylase}} APS + \text{Pyrophosphate}$$

$$APS + ATP \xrightarrow[\text{phosphokinase}]{\text{APS-}} PAPS + ADP$$

Despite the rapid synthesis, the sulfation of xenobiotics may by limited by reduced availability of 3′-phosphoadenosine 5′-phosphosulfate. The availability of the cofactor of the synthesis reaction, sulfate, may be limited by consumption due to sulfation or the limited availability of free cysteine for transformation to sulfate. Therefore, the sulfation capacity for certain xenobiotics is dependent on dose. Following administration of low doses the compound may be excreted as sulfate; after high doses, the capacity of sulfate conjugation may be saturated, and other biotransformation reactions such as glucuronide formation may become more important for biotransformation.

2.5.5.3. *Methyl Transferases*

A large number of alcohols, phenols, amines, and thiols present both as endogenous and exogenous compounds may be methylated by several *N*-, *O*- and *S*-methyl transferases. The most common donor for the methyl group is *S*-adenosyl methionine, which is formed from methionine and ATP. Often, these reactions do not increase the water solubility of a xenobiotic, but they are regarded as phase-II reactions since they mask potentially toxic functional groups and thus may serve as detoxication reactions.

A large variety of enzymes catalyze methylations of xenobiotics. The more important enzymes involved in methylation reactions are usually found in many tissues and are present in the soluble fraction of tissues. Some of the enzymes have a high specificity for certain endogenous compounds such as histamine or noradrenaline; others, such as catechol *O*-methyl transferase, metabolize both endogenous catechols and certain xenobiotics carrying aromatic rings with catechol functionalities [63].

2.5.5.4. *N-Acetyl Transferases*

Aromatic amines, hydrazines, sulfonamides, and certain aliphatic amines are biotransformed into amides in a reaction catalyzed by *N*-acetyl transferases. Enzymes that catalyze the acetylation of amines are designated as acetyl CoA: amine *N*-acetyl transferases. These enzymes utilize acetyl coenzyme A as cofactor [64]. The acetylation reaction of arylamines occurs in discrete steps. In the first step, the acetyl group from acetyl coenzyme A is transferred to the *N*-acetyl transferase, which then acetylates the arylamine, thus regenerating the enzyme and forming the amide (Fig. 23).

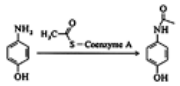

Figure 23. Mechanism of the *N*-acetylation of amines by *N*-acetyl transferase

N-Acetyl transferases are found in a number of different forms in cytosol of many tissues. In many species, the expression of *N*-acetyl transferases is under genetic control, and polymorphism for the expression of *N*-acetyl transferase has been found in several animal species and in humans. The transfer of an acetyl group to amines is reversible, and deacetylation of amides occurs in many species; as noted above, there are large differences between strains, species, and individuals in the extent of expression of amidases [65].

2.5.5.5. *Amino Acid Conjugation*

Exogenous carboxylic acids are conjugated with a variety of amino acids to form amides. Substrates for conjugation are mainly carboxylic acids containing aromatic rings. Glutamate and glycine appear to be the most common amino acids involved in these conjugation reactions in mammals; in other species such as reptiles and birds, ornithine is involved. The reaction proceeds in two steps and is catalyzed to two different enzymes. In the first step, the carboxylic acid is activated to form a coenzyme A derivative in a reaction involving coenzyme A and ATP. The enzymes that catalyze this reaction are called ATP-dependent acid coenzyme A ligases and are present in mitochondria. They appear to be identical to the intermediate-chain-length fatty acid:coenzyme A synthetase. The thus-formed coenzyme A thioester then transfers its acyl group to the amino group of the acceptor amino acid. This reaction is catalyzed by an *N*-acetyl transferase (Fig. 24). The enzymes catalyzing both steps in amino acid conjugation exist in several forms with different substrate specifities.

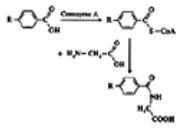

Figure 24. Amino acid conjugation of a xenobiotic carboxylic acid by ATP-dependent acid- coenzyme A ligases followed by *N*-acetyltransferase

2.5.5.6. *Glutathione Conjugation of Xenobiotics and Mercapturic Acid Excretion*

The conjugation of xenobiotics or their metabolites with the tripeptide glutathione is an important conjugation reaction. Glutathione is composed of the amino acids cysteine, glutamic acid, and glycine (ϒ-glutamylcysteinylglycine) and is present in many cells in high concentrations (up to 10 mM in liver cells) [66]. Since glutathione conjugation captures reactive electrophiles and transforms them into stable, often nontoxic thioethers, the formation of glutathione conjugates protects cells from the harmful effects of these electrophiles and thus serves as a major detoxication reaction (Section Detoxication and Interactions of Reactive Metabolites with Cellular Macromolecules). Glutathione conjugation is catalyzed by a family of enzymes termed glutathione *S*-transferases, which are present in the highest concentration in the liver, but are also found in high activity in the kidney, testes, and lung [67]. Glutathione *S*-transferases exist in both membrane-bound and soluble forms; with most substrates, the activity of soluble glutathione *S*-transferase is higher than that of microsomal glutathione *S*-transferase [68].

Cytosolic glutathione *S*-transferases exist in numerous different isoforms, each species being a dimer differing in subunit composition [67, 69]. The glutathione *S*-transferase gene family exists of at least six different families. In contrast to the multiple forms of soluble glutathione *S*-transferases thus possible, only one form of the membrane-bound enzyme is known. The glutathione *S*-transferases catalyze the reaction of the sulfhydryl group of glutathione with chemicals containing electrophilic carbon atoms (Table 14).

Table 14. Substrates for mammalian glutathione S-transferases

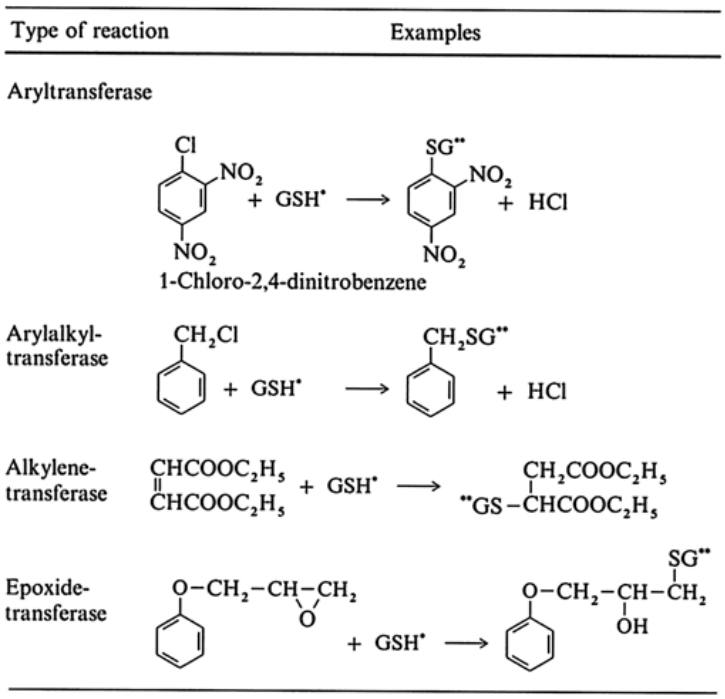

GSH = Glutathione. ** GS = Glutathionyl residue.

Thioethers are formed by reaction of the thiolate anion of glutathione with the electrophile; a spontaneous reaction, albeit at low rates, of the electrophile with glutathione without assistance by glutathione *S*-transferases is required for enzymatic catalysis. Glutathione thioethers formed in the organism are not excreted, but further processed to excretable mercapturic acids. Mercapturic acids are thioethers derived from *N*-acetyl-L-cysteine. Mercapturic acid formation is initiated by conjugation of the xenobiotic or an electrophilic metabolite with glutathione (Fig. 25).

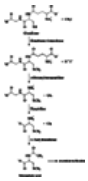

Figure 25. Formation of mercapturic acids by processing of glutathione *S*-conjugates as exemplified by the metabolism of methyl iodide

This is followed by transfer of the glutamate by ϒ-glutamyltranspeptidase, an enzyme specifically recognizing ϒ-glutamyl peptides and found in high concentrations in the kidney and other excretory organs. Dipeptidases catalyze the loss of glycine from the intermediary cysteinylglycine *S*-conjugate to give the cysteine *S*-conjugate which, in the final step of mercapturic acid formation, is *N*-acetylated by a cysteine-conjugate-specific *N*-acetyltransferase using acetyl coenzyme A as cofactor. The mercapturic acids formed are readily excreted into urine by active transport mechanisms in the kidney [70].

Glutathione conjugation is one of the most important detoxication reactions for reactive intermediates formed in organisms. Usually, metabolically formed intermediates are efficiently detoxified, but under specific circumstances, glutathione conjugation may be overwhelmed by high concentrations of electrophiles, which result in covalent binding of intermediates to cellular macromolecules, disruption of important cellular functions, and cell death and necrosis (see Section Hydrolytic Enzymes in Phase-I Biotransformation Reactions).

2.5.6. *Bioactivation of Xenobiotics*

Many xenobiotics with low chemical reactivity (e.g., the solvent carbon tetrachloride, the environmental contaminant hexachlorobutadiene, and the heat-exchanger fluid tri-*o*-cresyl phosphate) cause toxic effects. These toxic effects are initiated by covalent binding to macromolecules of metabolites formed in the organism by biotransformation enzymes. This process is termed bioactivation. With many chemicals, reactive metabolites formed during bioactivation may be efficiently detoxified; thus, toxic effects only occur when the balance between the production of reactive metabolites and their detoxication is disrupted. For example, toxic effects may be observed with a certain chemical only when the formation of reactive intermediates is enhanced or when the capacity for detoxication is diminished.

The mechanisms of bioactivation of xenobiotics may be classified into four categories describing the basic types of reactive intermediates formed and their potential reactivity (Table 15) [71-73].

Table 15. Basic mechanisms involved in the bioactivation of xenobiotics based on chemical reactivity of intermediates formed

Mechanism	Structure and reactivity of the intermediate	Examples
Biotransformation to stable but toxic metabolites	different structures, selective interaction of formed metabolite with specific acceptors, or disruption of specific biochemical pathways	dichloromethane, acetonitrile, parathion
Biotransformation to electrophiles	reactive electrophiles	dimethylnitrosamine, acetaminophen, bromobenzene
Biotransformation to free radicals	radicals	carbon tetrachloride
Formation of reactive oxygen metabolites	radicals	paraquat, aromatic nitro compounds

2.5.6.1. *Formation of Stable but Toxic Metabolites*

This mechanisms is limited to a few selected chemicals because few xenobiotic metabolites are both stable and toxic. The bioactivation of the solvents *n*-hexane and dichloromethane are examples of this mechanism. *n*-Hexane produces a characteristic neuropathy and peripheral nerve injury after chronic exposure. The same typical manifestations of toxicity are also observed when the *n*-hexane metabolites 2-hexanone and 2,5-hexanedione are administered to animals. The mechanism of *n*-hexane neuropathy thus involves oxidation of *n*-hexane by cytochromes P450 at both ends of the carbon chain (ω–1 hydroxylation) and further oxidation of the thus-introduced alcoholic function. The 2,5-hexanedione formed reacts with critical lysine residues in axonal proteins by Schiff base formation followed by cyclization to give pyrroles [74]. Oxidation of the pyrrole residues then causes crosslinking between two *n*-hexane-modified proteins; the resulting changes in the three-dimensional structures of proteins perturb axonal transport and function and cause damage to nerve cells (Fig. 26).

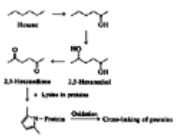

Figure 26. Bioactivation of hexane by cytochrome P450 to 2,5-hexanedione

Hexanedione reacts with lysine groups in proteins to form pyrroles; oxidation of two neighboring pyrrole residues causes the cross-linking of proteins.

Carboxyhemoglobin formation is observed after human exposure to dichloromethane. Dihalomethanes are oxidized by cytochrome P450, likely by P450 2E1, to carbon monoxide, which, due to its high affinity for iron(II)-containing porphyrins, binds to hemoglobin and interferes with oxygen transport in the blood. Other examples for the formation of stable, but toxic metabolites include the oxidation of acetonitrile to cyanide

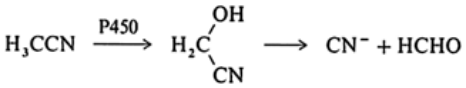

and the oxidative desulfuration of parathion.

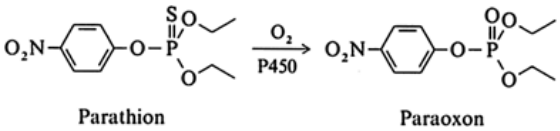

Parathion is a potent insecticide acting as an inhibitor of cholinesterase, but itself is only a weak cholinesterase inhibitor. Biotransformation of parathion by oxidative desulfuration to give the potent cholinesterase inhibitor paraoxone is responsible for the high insecticidal potency.

2.5.6.2. *Biotransformation to Reactive Electrophiles*

Biotransformation to reactive electrophiles is the most common pathway of bioactivation. The cytotoxicity and carcinogenicity of many chemicals is associated with the formation of electrophiles and the ensuing alkylation or acylation of tissue constituents such as protein, lipid, or DNA. Reactive intermediates include such chemically diverse functionalities as epoxides, quinones, acyl halides, carbocations, and nitrenium ions. The metabolic formation of electrophiles may be catalyzed by many different enzymes, although the majority of cases elucidated to date involve cytochrome P450-mediated oxidations.

Cytochrome P450 catalyzes the transformation of olefins to reactive and electrophilic oxiranes. For example, the carcinogenicity of the industrial intermediate vinyl chloride (Fig. 27) and the fungal toxin vaflatoxin B1 (Fig. 28) are dependent on their transformation to electrophilic oxiranes [75, 76].

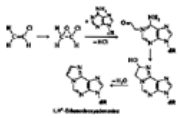

Figure 27. Bioactivation of vinyl chloride to chlorooxirane and reaction of the epoxide with critical macromolecules in the cell

dR = Deoxyribose

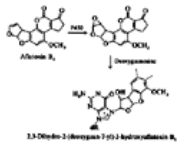

Figure 28. Bioactivation of aflatoxin B_1 to an electrophilic oxirane which results in the formation of DNA adducts and is believed to initiate tumor induction by aflatoxin B_1 in the liver

Carbocations are formed during the cytochrome P450-mediated oxidation of dialkyl nitrosamines. For example, the mutagen and potent carcinogen dimethylnitrosamine is hydroxylated by cytochrome P450 followed by loss of formaldehyde. Monomethylnitrosamine thus formed is unstable and rearranges to release an electrophilic carbocation (Fig. 29).

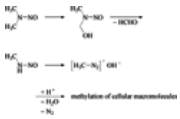

Figure 29. Bioactivation of dimethylnitrosamine to a methylating agent by cytochrome P450

Acyl halides are formed by the oxidation of carbon atoms bearing at least two halogen atoms. The initially formed products are unstable α-halohydrins, which lose hydrogen chloride and thus give reactive acyl halides. An example for the formation of acyl halides as reactive intermediates is the cytochrome P450-mediated oxidation of chloroform to phosgene [77] (Fig. 30).

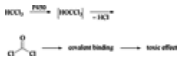

Figure 30. Bioactivation of chloroform by cytochrome P 450 mediated hydroxylation of a C — H bond

However, besides cytochromes P450, other monooxygenases such as flavin-dependent monooxygenase and of phase-II biotransformation enzymes such as UDP-glucuronyl transferases, sulfotransferases, or even the glutathione *S*-transferases may catalyze the bioactivation of xenobiotics [78]. For example, *N*-acetylamidofluorene is oxidized to *N*-hydroxyacetylamidofluorene by cytochrome P450. However, this metabolite is not electrophilic and requires further biotransformation via sulfate conjugation to the highly reactive *O*-sulfate ester. This sulfate ester fragments to a reactive intermediate (a nitrenium ion) which covalently binds to tissue constituents such as DNA (Fig. 31).

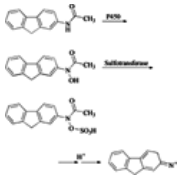

Figure 31. Bioactivation of acetylamidofluorene by cytochrome P450 and by UDP glucuronyl transferases. The glucuronide formed is acid-labile and decomposes to a nitrenium ion.

Some glutathione *S*-conjugates which are biosynthesized to detoxify electrophiles are toxic and mutagenic [79, 80]. 1,2-Dibromoethane is metabolized by glutathione conjugation to *S*-(2-bromoethyl)glutathione. Intramolecular displacement of the bromine on the adjacent carbon atom gives a highly strained, electrophilic episulfonium ion (Fig. 32).

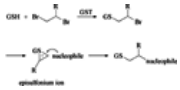

Figure 32. Bioactivation of 1,2-dibromoethane by glutathione conjugation to a reactive and electrophilic episulfonium ion

Other toxic glutathione *S*-conjugates require processing by the enzymes of mercapturic acid formation to give electrophiles. A minor pathway in perchloroethene biotransformation results in *S*-(1,2,2-trichlorovinyl)glutathione [81]. This glutathione *S*-conjugate is cleaved by γ-glutamyl transpeptidase and dipeptidases to *S*-(1,2,2-trichlorovinyl)-L-cysteine, which is a substrate for renal cysteine conjugate β-lyase and transformed to pyruvate, ammonia, and a reactive thioketene, binding of which to cellular macromolecules is likely responsible for the renal toxicity of perchloroethene (Fig. 33).

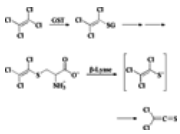

Figure 33. Bioactivation of perchloroethene by glutathione conjugation. The conjugate *S*-(1,2,2-trichlorovinyl)glutathione is biosynthesized in the liver, translocated to the kidney to be processed by γ-glutamyl transpeptidases and dipeptidases, and finally cleaved by cysteine-conjugate β-lyase to give dichlorothioketene.

Due to high concentrations of *S*-(1,2,2-trichlorovinyl)-L-cysteine obtained by active transport to the kidney, covalent binding of the dichlorothioketene formed via this pathway occurs only in the kidney; despite the presence of cysteine conjugate β-lyase in many other organs.

2.5.6.3. *Biotransformation of Xenobiotics to Radicals*

Free radicals are chemical species that may be formed by a one-electron oxidation to give a radical cation, by a one-electron reduction to give a radical anion, or by homolytic fission of a σ-bond to give a neutral radical.

Free radicals are highly reactive and, when formed in biological systems, are expected to react with a variety of tissue molecules. Radicals may abstract hydrogen atoms, undergo oxidation-reduction reactions, dimerizations and disproportionation reactions. Radicals may also participate in a chain mechanism, which is initiated by a reaction causing a free radical and propagated by a subsequence of reactions causing further radicals as products.

The toxic and tumorigenic solvent carbon tetrachloride is the outstanding example of a bioactivation reaction to a free radical. Carbon tetrachloride is biotransformed by a one-electron reduction to yield the trichloromethyl radical and chloride:

$$Cl_4C + e^- \rightarrow Cl_3C^\bullet + Cl^-$$

The trichloromethyl radical may abstract hydrogen atoms from tissue macromolecules to give chloroform, a proven metabolite of carbon tetrachloride, or may dimerize to give hexachloroethane, which is also a metabolite of carbon tetrachloride. Toxic effects of radicals formed during biotransformation reactions are lipid peroxidation and oxidative modification of proteins (see Section Formation of Reactive Oxygen Metabolites by Xenobiotics). Formation of radicals has been implicated in the bioactivation of many xenobiotics. Radicals may be formed by NADPH-dependent cytochrome P450 reductase, nitroreductases, or one-electron oxidations catalyzed by peroxidases such as prostaglanding synthetases. Formation of free radicals from tissue constituents also plays an important role in the toxic effects of ionizing radiation [82-85].

2.5.6.4. *Formation of Reactive Oxygen Metabolites by Xenobiotics*

Xenobiotic-induced formation of reduced oxygen metabolites such as the superoxide radical anion, hydrogen peroxide, and the hydroxyl radical has been implicated as a mechanism of producing cell damage, so-called oxidative stress [86-89]. The biotransformation of certain xenobiotics that are involved in redox cycles or undergo enzyme-catalyzed oxidation/reduction reactions may be associated with the production of reduced oxygen metabolites. 2-Methylnaphthoquinone (Menadione) has been intensively used to study the formation and cellular reactions of reduced oxygen metabolites. Menadione and other quinones undergo enzymatic redox cycling; these one-electron oxidation reactions are associated with the formation of the superoxide radical anion ($O^{\cdot 2-}$) by one electron reduction of triplet oxygen. In aqueous solution, superoxide is not particularly reactive, but dismutation or further reduction of superoxide may give rise to hydrogen peroxide (Fig. 34).

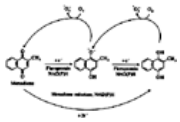

Figure 34. Biotransformation of menadione and induction of oxidative stress by reduction of triplet oxygen to the superoxide radical anion

Hydrogen peroxide is also a poor oxidant in biological systems, but sufficiently stable to cross biological membranes. The toxicity of hydrogen peroxide is attributed to the formation of the hydroxyl radical by the Fenton reaction, catalyzed by metal ions such as Fe^{2+} (M = transition metal):

$$M^n + H_2O_2 \rightarrow M^{(n+1)} + HO^\bullet + HO^-$$

The highly reactive hydroxyl radical may then initiate cellular damage by radical-based mechanisms. Besides menadione, oxidative stress may also be initiated by other xenobiotics such as the bis-pyridinium herbicide paraquat and nitroheterocycles. Moreover, the formation of reduced oxygen metabolites plays an important role in host defense against infectious agents and in the initiation and propagation of certain diseases such as arteriosclerosis and polyarthritis.

Since oxygen radicals are also formed in low concentrations during cellular respiration, efficient mechanisms for their detoxication exist (see Section Detoxication and Interactions of Reactive Metabolites with Cellular Macromolecules). Oxidative stress is thus only observed when the eqilibrium between oxidants and reductants is disturbed and detoxication mechanisms are overwhelmed.

2.5.6.5. *Detoxication and Interactions of Reactive Metabolites with Cellular Macromolecules*

Reactive intermediates formed inside cells may react with low and high molecular mass cellular constituents. These interactions may result in formation of less reactive chemicals and thus in detoxication, or may perturb important cellular functions and thus result in acute and/or chronic toxic effects such as necrosis or cancer. Usually, the interaction with low molecular mass constituents in the cell results in detoxication, whereas the irreversible interaction with cellular macromolecules results in adverse effects [70, 90-92].

Detoxication of reactive intermediates may be due to hydrolysis, glutathione conjugation, or interactions with cellular antioxidants. The reaction of electrophilic xenobiotics with the nucleophile water, present in high concentrations in all cells, is the simplest form of detoxication. Many of the products thus formed are of low reactivity and may be rapidly excreted. For example, acyl halides formed by the oxidation of olefins such as perchloroethylene are hydrolyzed rapidly to halogenated carboxylic acids; only minor amounts of the intermediate acyl halide reacts with protein and lipids (Fig. 35).

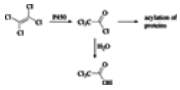

Figure 35. Biotransformation of tetrachloroethylene to trichloroacetyl chloride followed by hydrolysis to trichloroacetic acid, the major urinary metabolite formed from tetrachloroethylene. Only a small amount of the acyl halide formed reacts with proteins.

Glutathione-dependent detoxication is an important mechanism for metabolically formed electrophiles, free radicals, and reduced oxygen metabolites [93-96]. Electrophiles react with the nucleophilic sulfur atom of glutathione in a spontaneous or enzyme-catalyzed reaction. Spontaneous reactions are only observed at appreciable rates with soft electrophiles (glutathione is a soft nucleophile); the conjugation of hard electrophiles with glutathione requires enzymatic catalysis; usually, the rates of conjugation catalyzed by glutathione *S*-transferase differ between hard and soft electrophiles; soft electrophiles are conjugated more efficiently. For example, the hard electrophile aflatoxin B_1 8,9-oxide does not spontaneously react with glutathione; only in the presence of a certain glutathione *S*-transferase

enzyme is a glutathione *S*-conjugate of aflatoxin B_1 8,9-oxide formed. Species differences in the tumorigenesis of aflatoxin B_1 may serve to illustrate the important role of glutathione *S*-transferases in the expression of toxicity and carcinogenicity. Aflatoxin B_1 is a potent liver carcinogen in rats; in mice, aflatoxin B_1 is only weakly carcinogenic. The liver of mice contains a glutathione *S*-transferase which efficiently detoxifies aflatoxin B_1 8,9-oxide. This glutathione *S*-transferase enzyme is not present in rat liver; thus, the binding of aflatoxin B_1 8,9-oxide to rat liver DNA and liver carcinogenicity of aflatoxin B_1 are much higher in rats than in mice.

Glutathione also plays a major role in the detoxication of reactive oxygen metabolites and radicals. Selenium-dependent glutathione peroxidases are important enzymes catalyzing the detoxication of hydrogen peroxide. In the glutathione peroxidase catalyzed reaction, two moles of glutathione are oxidized to glutathione disulfide:

$$H_2O_2 + 2\,GSH \underset{\text{Glutathione reductase}}{\overset{\text{Glutathione peroxidase}}{\rightleftharpoons}} GSSG + H_2O$$

Glutathione can be recycled by the reduction of glutathione disulfide by glutathione reductase.

The copper- and zinc-dependent cytosolic and - manganese-dependent mitochondrial superoxide dismutases detoxify superoxide radical anions. Hydrogen peroxide formed by dismutation of superoxide is converted to water and oxygen and thus detoxified by catalase:

$$2\,H_2O_2 \xrightarrow{\text{Catalase}} 2\,H_2O + \tfrac{1}{2}\,O_2$$

Several cellular antioxidants also play a role in the detoxication of radicals. α-Tocopherol is an important lipophilic antioxidant, whose presence in lipid membranes prevents damage to lipid constituents (e.g. unsaturated fatt acids) by radicals. The hydroxyl radical, the superoxide radical anion, and peroxy radicals react with α-tocopherol to yield water, hydrogen peroxide, and hydroperoxides, which may be detoxified further by catalase and glutathione peroxidase. α-Tocopherol is transformed during these reactions to give a stable radical of comparatively low reactivity. Ascorbic acid is an important antioxidant present in the cytoplasm of the cell and may also participate in the detoxication of radicals.

2.5.6.6. *Interaction of Reactive Intermediates with Cellular Macromolecules*

Although a substantial body of information is available on the biotransformation of xenobiotics to reactive metabolites and the chemical nature of those metabolites, considerably less is known about how reactive intermediates interact with cellular constituents and how those interactions cause cell injury and cell death. The reaction of toxic metabolites may result in the formation of covalent bonds between the molecule and a cellular target molecule, or they may alter the target molecule without formation of a covalent bond, usually by oxidation or reduction [97].

Electrophilic metabolites may react with different nucleophilic sites in cells. Nucleophilic sites in cellular macromolecules are thiol and amino groups in proteins, amino groups in lipids, and oxygen and nitrogen atoms in the purine and pyrimidine bases of DNA. The formation of a covalent bond may permanently alter the structure and/or activity of the modified macromolecule and thus result in a toxic response. The complexity of the reaction of electrophilic metabolites with the various nucleophilic sites in cells may be interpreted on the basis of the concept of hard and soft electrophiles and nucleophiles (hard and soft acids and bases). The donor atom of a soft nucleophile is of high polarizability and low electronegativity, and is easily oxidized; the donor atom of a hard nucleophile is of low polarizability and high electronegativity. Hard electrophiles carry a high positive charge and have a small size; soft electrophiles are of low positive charge and large size. Soft electrophiles react predominantly with soft nucleophiles, and hard electrophiles with hard nucleophiles [98]. Thus, hard electrophiles formed during a biotransformation reaction (e.g., carbocations formed from dialkylnitrosamines) predominantly react with hard nucleophiles such as the oyxgen and nitrogen atoms of DNA, In contrast, soft electrophiles such as α, β-unsaturated carbonyl compounds (e.g., acrolein, benzoquinone) react predominatly with soft tissue nucleophiles such as the sulfhydryl groups of cysteine in proteins (Table 16).

Table 16. Metabolically formed electrophiles and their prime targets for covalent binding in cells

Soft		→		**Hard**
		Nucleophile		
SH of cysteine or glutathione	sulfur of methionine	primary or secondary nitrogen atoms in peptides (lysine, arginine, or histidine)	amino groups of purines and pyrimidines in RNA and DNA	oxygen in purines and pyrimidines in DNA and RNA
		Electrophile		
α, β-Unsaturated carbonyls, quinones	epoxides, alkyl sulfates, alkyl halides	nitrenium ions	benzylic carbocations	aliphatic and aromatic carbocations

Covalent interactions of xenobiotics with proteins occur with several nucleophilic nitrogen atoms; both alkylation and acylation reactions of amino acids have been reported as consequences of formation of reactive intermediates in cells. Besides the sulfur atom of cysteine, nitrogen atoms in the amino acids lysine, histidine, and valine are frequent targets for electrophilic metabolites. Consequences of the modifications may be inactivation of enzymes important for cellular function, changes in the tertiary structure of proteins, or changes in gene expression. Alkylation of the sulfhydryl-dependent enzymes of mitochondrial respiration is thought to play an important role in the initiation of mitochondrial dysfunction and thus cell damage.

Some modified proteins may also serve as immunogens, and hypersensitivity reactions, formation of immune complexes, and delayed hypersensitivities may be the consequences of protein adduct formation. Indeed, many drug- and chemical-related hypersensitivity reactions observed in clinical medicine are based on the formation of covalent protein adducts and their recognition as “foreign” by the immune system [99].

Oxidative stress produces mixed disulfides of proteins with low molecular mass thiols such as glutathione and thus alters protein structure and function. In addition, oxidants and radicals promote the oxidation of amino acids in proteins, which may increase the susceptibility of these proteins to proteolysis [100, 101]. Increased protein oxidation has been implicated in cellular aging and in the mechanisms of toxicity of several redox-active transition metals.

Radicals formed during the biotransformation of xenobiotics may abstract hydrogen atoms from cellular components [82]. The abstraction of hydrogen atoms from polyunsaturated fatty acids of lipids results in a process termed lipid peroxidation. The fatty acid radicals thus formed may react with molecular oxygen to give peroxy radicals and further to hydroperoxides. The initiated radical chain reactions cause the cleavage of carbon–carbon bonds in the fatty acids to short fragments such as α, β-unsaturated carbonyl compounds [102, 103] (see Section Binding of Xenobiotics or their Metabolites to Cellular Proteins).

The disruption of membranes and the formation of toxic hydroperoxides and α, β-unsaturated carbonyl compounds may cause disruptions in cellular calcium homeostasis and thus cause biochemical changes that ultimately lead to cell death [104, 105]. The reaction of electrophilic metabolites with DNA constituents results in the formation of altered purine and pyrimidine bases or other DNA damage such as DNA strand breaks or loss of single bases from the double helix. Many of these modifications are “premutagenic lesions”. After gene expression, these lesions may be translated into mutations [106]. Mutations in certain genes are considered to be the basis for the evolution of neoplastic cells and cancer and thus play a major role in chemical carcinogenesis. Other types of DNA damage may result in the activation of genes important for cellular differentiation or other regulatory functions. Electrophilic intermediates alkylate the nitrogen and oxygen atoms of the purine and pyrimidine bases in DNA; deoxyguanosine is often preferentially alkylated. The site of alkylation of a certain base in DNA is again dependent on the electrophilicity of the alkylation agent; hard electrophiles preferentially react with the oxygen atoms of guanosine, while soft electrophiles alkylate the exocyclic amino groups (Fig. 36).

Figure 36. Regioselectivity of DNA alkylation by different electrophiles

The pattern of base alkylation is different when DNA is modified in biological systems or when isolated DNA, nucleosides, and nucleotides are treated with the xenobiotic or its metabolite(s). Regioselectivity is further modified by solvents, buffer salts, and concentration of reactants. Certain modifications of deoxyguanosine result in the labilization of the glycosidic bond; loss of the deoxyguanosine derivate results in an “apurinic site” in DNA other chemical modifications may result in labilization of the five-membered ring and ring opening after reaction with water (Fig. 37).

Figure 37. Labilization of the bond of guanosine to the DNA backbone by alkylation of the N-7 position resulting in the loss of the modified guanosine from DNA and the formation of an apurinic site (AP)

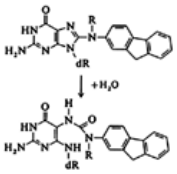

Figure 38. Ring opening of guanosine in DNA by alkylation of the N-7 position

Radicals formed as reactive intermediates may also cause DNA damage. Besides DNA strand breaks, which have been frequently observed, the reaction of oxygen-derived radicals may also result in the oxidation of purine and pyrimidine nucleotides. Due to the development of sensitive techniques for the characterisation of oxidative modifications in DNA, a number of modified bases have been identified:

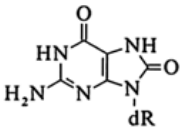

8-Hydroxydeoxyguanosine
(7,8-Dihydro-8-oxodeoxyguanosine)

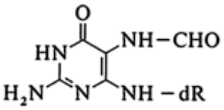

Formamidopyrimidine-deoxyguanosine

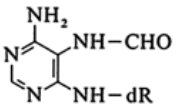

Formamidopyrimidine-deoxyadenosine

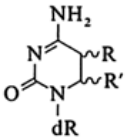

R,R′ = H, OH, OOH
dR = Deoxyribose

5,6-Dihydrodeoxycytidine

8-Hydroxydeoxyguanosine, a premutagenic modification, is considered as one of the more important lesions induced by oxidative DNA damage; because sensitive methods are available for its quantification, it can serve as a marker for the extent of oxidative DNA modification caused by a xenobiotic or by other processes. DNA oxidation has also been implicated in aging; an increase in oxidative DNA modifications may occur with age due the decreased availability of antioxidants in cells of aging mammals. Several theories suggest a correlation between increased oxidative DNA damage and the increased incidence of tumors in the aged population [107, 108].

2.5.7. *Factors Modifying Biotransformation and Bioactivation*

The biotransformation of xenobiotics may be modified by a variety of factors both intrinsic and extrinsic to the normal functioning of the organism. The changes in the extent of biotransformation may have profound effects on the toxicity of a specific chemical. When biotransformation results in detoxication and rapid excretion, increased rates of biotransformation will decrease toxicity. On the other hand, the toxicity of a chemical bioactivated to reactive intermediates will increase on enhancing biotransformation. A great variety of factors have been shown to influence the extent of biotransformation; many of the effects listed below have been primarily decribed in experimental animals. However, observations in humans (e.g,. after drug treatment) indicate that similar effects, albeit not of the same magnitude or duration, must occur in humans.

2.5.7.1. *Host Factors Affecting Biotransformation*

Enzyme Induction. The activity of biotransformation enzymes can be enhanced by pretreatment with a range of structurally different chemicals. These chemicals can be drugs, pesticides, natural products, environmental contaminants, and even ethanol. The enhanced enzyme activities and the increased enzyme concentrations may results from increased de novo synthesis of the protein, reduced degradation, or from other, often unknown effects. An increase in the concentration of an biotransformation enzyme in the organism, a certain organ or cell type is termed “enzyme induction” [109, 110]. Several hundred different chemicals have been demonstrated to increase the biotransformation of other xenobiotics and to act as enzyme inducers. The majority of these studies focused on the induction of microsomal monooxygenases, mainly cytochrome P450 enzymes; however, other membrane bound enzymes such as UDP-glucuronyl transferases may also be induced. Glutathione *S*-transferases are the only cytosolic biotransformation enzymes whose activities may be increased by the administration of inducers to experimental animals (Table 17).

Table 17. Inducers of the enzymes of biotransformation and enzymes whose cellular concentrations are increased by pretreatment

Inducing agent	Induced enzymes
2,3,7,8-Tetrachlorodibenzodioxin	cytochrome P450, UDPglucuronyltransferase
Ethanol	cytochrome P450
Phenobarbital	cytochrome P450, epoxide hydrolase, UDP-glucuronyltransferases
trans-Stilbenoxide	epoxide hydrolase
3-Methylcholanthrene	cytochrome P450, UDPglucuronyltransferases

The onset, magnitude, and duration of increases in the concentration of biotransformation enzymes after the administration of an inducer and the associated biochemical and morphological effects depend on the chemical nature of the inducing agent, dose, and time of administration. For example, the time required for maximum induction of specific cytochrome P450 enzymes by the classical inducers phenobarbital and 3-methylcholanthrene are different. Moreover, besides increases in the activity of hepatic monooxygenases, phenobarbital administration results in marked hepatic hypertrophy and proliferation of the smooth endoplasmatic reticulum; these effects are absent in animals treated with 3-methylcholanthren. Induction by parenteral application of 3-methylcholanthren results in maximum enzyme concentrations in the liver within 48 h, whereas maximal induction by parenteral application of hypnotic doses of phenobarbital requires up to 5 d.

Enzyme induction is reversible after withdrawal of the inducing agent, and the enzyme activities return to basal levels over a characteristic time span. Again, this time span is dependent on the chemical nature of the inducing agent. For example, cessation of phenobarbital treatment will result in a decline of enzyme activities to basal levels within one to two weeks. The mechanisms of enzyme induction are complicated and only partially understood. Apparently, different chemicals influence the activities of the biotransformation enzymes by different mechanims; even the effect of a specific chemical on different enzymes may be due to separate mechanisms (Table 18) [111].

Table 18. Mechanisms of cytochrome P450 enzyme induction by different xenobiotics

Cytochrome P450	Inducing agent	Mechanism of induction
1A1	2,3,7,8-tetrachlorodibenzodioxin	increased gene transcription
1A2	3-methylcholanthrene	stabilization of messenger RNA
2B1, 2B2	phenobarbital	increased gene transcription
2E1	ethanol, acetone	protein stabilization
3A1	dexamethasone	increased gene transcription, independent of glucocorticoid receptor
3A1	triacetyloleandomycin	protein stabilization
4A1	clofibrate	increased gene transcription, receptor mediated

Only a few mechanisms are well understood. Modulation of gene expression seems to be the basis for many inductive effects. For example, induction of cytochrome P450 1A1 is prevented by inhibitors of protein synthesis. Moreover, studies using the potent inducing agent 2,3,7,8-tetrachlorodibenzodioxin identified a high-affinity binding protein with the properties of a receptor for 2,3,7,8-tetrachlorodibenzodioxin in rat liver cytosol [112, 113]. Binding of 2,3,7,8-tetrachlorodibenzodioxin to this protein results, after further interaction with other proteins, in the translocation of the formed complex from the cytosol to the nucleus. This translocation is followed by interaction with specific recognition sites on the genome, transcription, and translation of the specific gene for cytochrome P450 1A1 and for other biotransformation enzymes such as UDP-glucuronyl transferases (Fig. 39).

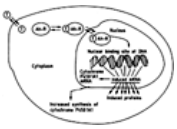

Figure 39. The Ah receptor (Ah-R) and mechanism of enzyme induction by 2,3,7,8-tetrachlorodibenzodioxin (T)

In contrast, the mechanisms of enzyme induction by phenobarbital and, for example, ethanol are not defined. A specific receptor for phenobarbital could not be demonstrated, but some experiments suggest involvement of the glucocorticoid receptor in phenobarbital-mediated enzyme induction. Ethanol and other inducers seem to stabilize the cytochrome P450 2E1 protein against degradation by an unknown mechanism.

Enzyme Inhibition. The decrease in the activity of specific biotransformation enzymes is termed inhibition of biotransformation. As noted above, inhibition of biotransformation may increase or decrease the toxicity of a xenobiotic. For example, the inhibition of cytochrome P450 by 2-(diethylamino)ethyl-2,2-diphenylpentanoate (SKF-525A) causes an increase in hexobarbital sleeping time, but a decrease in the hepatoxicity of carbon tetrachloride.

Several mechanisms responsible for inhibition of xenobiotic metabolizing enzymes are operative [114, 115]. Besides inhibition of protein synthesis and thus synthesis of the enzyme, xenobiotics may irreversibly bind to the active site of the enzyme. This process is termed suicide inhibition [116]. Following activation of the xenobiotic by the enzyme, the reactive intermediate formed reacts with constituents of the enzyme at or near the active site, thus blocking further catalytic activity. This effect has been demonstrated with several xenobiotics which are inhibitors of cytochrome P450, such as alkenes and compounds containing allylic and acetylenic derivatives. For example, ethylene oxide, the reactive metabolite formed by cytochrome P450 catalyzed oxidation of ethylene, may alkylate the pyrrole nitrogen atoms in the heme moiety and thus result in heme destruction.

During exposures to mixtures, chemicals with high affinity to certain biotransformation enzymes will be preferentially metabolized, and thus the biotransformation of other constituents of the mixture will be reduced or even totally inhibited. This reduced biotransformation may also alter the toxicity of a chemical present in a mixture compared with that of the pure chemical. For example, coadministration of ethanol markedly reduces the toxic effects of methanol (metabolic acidosis, reversible or even permanent blindness). These toxic effects are caused by the oxidation of methanol to formic acid as toxic metabolite; formic acid accumulates in the organism and damages the optical nerve. Oxidation of methanol to formaldehyde is competitively blocked by administration of ethanol; under these circumstances, the toxic metabolite formic acid can not be formed; methanol is excreted from the organism unchanged in the urine and by exhalation.

Inhibition of some biotransformation enzymes may also be caused by effects on the tissue levels of necessary cofactors. The availability of glutathione for conjugation is reduced by blocking glutathione biosynthesis; diethyl maleate and some other chemicals deplete intracellular glutathione concentrations by reacting with glutathione to give a glutathione *S*-conjugate. Pretreatment of experimental animals with these chemicals followed by the application of a xenobiotic which requires glutathione for detoxication will result in an increased toxic response. Moreover, large doses of nontoxic chemicals metabolized by sulfotransferases may deplete the cofactor for sulfate conjugation and may thus alter the disposition and, probably the toxicity, of other xenobiotics that undergo sulfate conjugation.

Genetic Differences in the Expression of Xenobiotic Metabolizing Enzymes. The ability of different animal species to metabolize xenobiotics is related to evolutionary development and therefore to different genetic constitution; thus, major species differences in the extent and pathways of biotransformation exist. These variations may be divided into qualitative and quantitative differences. Qualitative differences involve metabolic pathways and are related to species defects or peculiar reactions of a species. For example, guinea pigs do not have the enzymatic capacity to catalyze the last step in mercapturic acid formation, the *N*-acetylation of cysteine *S*-conjugates, and therefore excrete cysteine *S*-conjugates as end products of this pathway. Certain species such as cats do not have the capability to form glucuronides from xenobiotics.

Quantitative variations are often due to species differences in gene and enzyme regulation. For example, interindividual differences have been decribed in humans in the biotransformation of many drugs. The *N*-acetylation of the tuberculostaticum Isoniazid has a genetic basis. Some individuals are homozygous for a recessive gene, and this may result in the absence of isoniazid *N*-acetyltransferase, that is, they are “slow acetylators”. In normal homozygotes or heterozygotes, “rapid acetylators”, Isoniazid is rapidly transformed to the *N*-acetyl derivative. The *N*-acetyltransferase polymorphisms are correlated with different responses to Isoniazid-induced toxicities. This genetic polymorphism is also seen in human, 80% of the Japanese and Eskimos are “rapid acetylators”; whereas in some European populations only 40 to 60% are “rapid acetylators” [117]. Polymorphisms in the expression of cytochrome P450 1A1 and a specific glutathione *S*-transferase in human

lung have been implicated in an increased rate of lung cancer in individuals expressing high cytochrome P450 1A1 and deficient in this glutathione *S*-tansferase. Cytochrome P450 1A1 bioactivates aromatic hydrocarbons present in cigarette smoke to yield electrophiles; glutathione *S*-transferase detoxifies these metabolites.

Influence of sex on biotransformation reactions. Sex differences in the extent and pathways of biotransformation may be based on sex-dependent expression of certain biotransformation enzymes. For example, adult male rats metabolize many xenobiotics at higher rates than females; both phase-I and phase-II biotransformation reactions seem to be influenced by sex-dependent factors. With cytochrome P450, at least three different hepatic enzymes have been demonstrated to be under the control of sex-hormones (Table 19).

Table 19. Sex hormone-dependent hepatic cytochrome P450 in the rat

Enzyme	Sex specificity	Remarks
P450 2D	female	expressed at a hormone-independent basal rate, stimulated by estrogen, suppressed by androgen
P450 2C	male	
P450 3A	male	neonatally imprinted by androgen

In addition to the liver, sex differences in biotransformation are also found in extrahepatic tissues such as the kidney and may be responsible for sex-specific toxic effects of xenobiotics in these organs. For example, the kidneys of male mice contain a cytochrome P450 enzyme which bioactivates chloroform, acetaminophen, and 1,1-dichloroethene to reactive intermediates. The enzyme is present only in much lower activity in female mice, which are thus not susceptible to the renal toxicity of these chemicals.

In experimental animals, sex differences in the expression of biotransformation enzymes usually become apparent at puberty and are maintained throughout adult life. Despite the relatively large sex-dependent variations seen in animal studies, sex seems not to have a profound influence on the biotransformation of chemicals in humans.

Dietary constituents and the biotransformation enzymes. Nutritional factors influencing biotransformation may be mineral deficiencies, vitamin deficiencies, protein content, starvation, and natural substances in the diet. Mineral deficiencies (calcium, copper, zinc) have been shown to reduce the activities of cytochromes P450. On the other hand, an excess of dietary iron has been observed to increase monooxygenase activity. Dietary cobalt, calcium, and manganese may increase the hepatic levels of glutathione and may thus influence glutathione *S*-conjugate formation. Several vitamins are directly or indirectly involved in the regulation of cytochrome P450. For example, diets deficient in vitamins C and E reduce the activity of monooxygenases, whereas deficiencies in other vitamins increase monooxygenase activity. Moreover, several vitamins serve as important cellular antioxidants and influence the energy and redox state of the cell and thus also affect biotransformation reactions.

Low-protein diets generally reduce the activity of cytochrome P450 and certain phase-II biotransformation reactions. Thus, the nutrient status may also modify the toxicity of xenobiotics. For example, dimethylnitrosamine is a potent hepatocarcinogen in rats kept on a high-protein diet; but almost without effect in rats kept on a low-protein diet. Food deprivation reduces the hepatic concentration of glutathione by as much as 50% due to reduction of glutathione biosynthesis. Thus, xenobiotics detoxified by glutathione conjugation are more toxic in starved than in fed animals. Fasting has also been shown to increase the levels of cytochrome P450 2E1, but decrease the levels of cytochrome P450 2C11. Several natural ingredients in the diet of laboratory animals such as indoles, diallyl disulfid,e and psoralens may increase the activities of cytochrome P450. However, some of these compounds may selectively increase some cytochrome P450 enzymes, but inhibit others. For example, diallyl disulfide, a constituent of garlic, has been shown to induce cytochrome P450 2B in rats, but inhibits cytochrome P450 2E1. Vegetable ingredients present in broccoli are potent inducers of phase-II biotransformation enzymes and thus increase the capacity of the organism to detoxify reactive intermediates. These natural ingredients are thought to play a major role in the anticarcinogenicity of diets rich in those vegetables.

2.5.7.2. *Chemical-Related Factors that Influence Biotransformation*

Xenobiotic-related factors influencing biotransformation are the physiocochemical properties (i.e., chemical structure including the presence of functional groups) and dose. The major determinant of the rate of biotransformation is the concentration of the substrate at the active site of the enzyme. This concentration is determined by structure and lipophilicity and by dose. Lipophilic xenobiotics readily cross cell membranes and are rapidly absorbed and distributed in the organism. Moreover, lipophilic xenobiotics show a higher partitioning into lipid membranes. These factors contribute to higher concentrations of lipophilic xenobiotics at the active center, especially of membrane-bound enzymes. The presence of functional groups also influences rates and routes of biotransformation. Certain functional groups may compete for the same substrate for conjugation; also, specific functional groups may undergo different reactions, as indicated in Figure 40 for *p*-aminobenzoic acid.

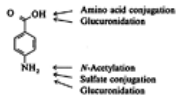

Figure 40. Possible biotransformation pathways for *p*-aminobenzoic acid

The presence of other specific functional groups may have a major effect on biotransformation and its regioselectivity. For example, the presence of trifluoromethyl groups in alkanes renders the adjacent methylene carbon atom almost inert to hydroxylation and strongly influences the regioselectivity of enzymatic hydroxylations on aromatic rings.

Dose is one of the most important factors determining rate and route of biotransformation for more complex molecules. Certain biotransformation enzymes have a high affinity but low capacity for a specific chemical, while others have a high capacity but low affinity. As dose increases, high-affinity, low-capacity enzymes will become staturated, and low-affinity, high capacity pathway(s) will biotransform a larger percentage of dose [118].

2.5.8. ***Elimination of Xenobiotics and their Metabolites***

The evolution of complex forms of life necessitated the development of specialized mechanisms to eliminate waste products formed from endogenous compounds and to prevent the accumulation of toxic xenobiotics present taken up in the diet. Excretion of wastes by the earlier forms of aquatic life was largely passive and involved the loss of large volumes of water and nutrients. For land-living animals, conservation of water, minerals, and nutrients was necessary for survival. Therefore, complex mechanisms for the elimination of both endogenous chemicals and xenobiotics evolved. A wide variety of xenobiotics can be handled by evolved excretory mechanisms and can thus be efficiently eliminated from the body.

Xenobiotics may be excreted as the parent compound, as metabolites, and/or as conjugates formed in phase-II biotransformation reactions. A major route of excretion of xenobiotics is via the kidney, and in some cases the urinary elimination of parent compound or metabolite can be used to determine absorbed dose. The kidney is the only organ which functions almost exclusively as an organ of elimination. The cells of the liver have more varied functions than those found in the kidney; however, the liver also plays an important role in the excretion of chemicals not effectively eliminated by the kidneys. Chemicals may be eliminated from the liver into bile and thus be finally excreted with feces. Active transport mechanisms, that is, transport against a concentration gradient, play a major role in renal and hepatic excretion of xenobiotics. In contrast, in most other organs which may serve as excretory systems for xenobiotics, passive excretion mechanisms are operative. For example, volatile chemicals and metabolites may be eliminated in expired air; this route is quantitatively significant for some solvents and inhalation anesthetics. Specific xenobiotics may also be excreted in sweat, saliva, and milk.

2.5.8.1. *Renal Excretion*

The kidneys are the only organs that are primarily designed for excretion. The function of these organs accounts for the elimination of most of the byproducts of normal metabolism and most of the polar xenobiotics and metabolites of lipophilic xenobiotics to which humans and experimental animals are exposed [119]. The kidney is a complex structure which consist of a number of different cell types [119-121]. Essentially, the kidney filters the bood and all components present in blood with a molecular mass of less than 50 000 (depending on structure and charge) enter the tubular system; there, important nutrients and most of the filtered water are recovered. Only a small fraction of the primary filtrate is excreted as urine (one to two liters per day). The human kidney consists of approximately two million nephrons, which are the functional units that filter the blood and the recover essential nutrients. The structure and components of a single nephron are shown in Figure 41.

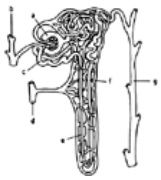

Figure 41. Structure of a nephron

a) Glomerulus; b) Renal artery; c) Proximal tubule; d) Renal vein; e) Loop of Henle; f) Distal tubule; g) Collecting duct

Glomerular Filtration. Renal excretion is the product of three complex and interactive processes: glomerular filtration, tubular reabsorption, and tubular secretion. Glomerular filtration is the passive filtering of the plasma as a result of its passage through glomerular pores (7–10 nm in diameter) under hydrostatic pressure generated by the heart. The average rate of glomerular filtration in adults is 125 ml/min or almost 200 liters/d. Glomerular filtration shows little specificity other than molecular size, and free solutes in the plasma that pass through the glomerular pores will all appear in the filtrate. Only protein-bound low molecular mass xenobiotics will not appear in the filtrate and remain in blood. Glomerular filtration is influenced by factors that affect the hydrostatic pressure or integrity of the glomerulus; thus, these factors may result in elevated plasma concentrations of excretory products formed from endogenous chemicals and from xenobiotics.

Tubular Reabsorption. The daily volume of glomerular filtrate exceeds that of the total body water by a factor of four and contains many necessary nutrients such as glucose, amino acids, and salt; therefore, most of the glomerular filtrate must be recovered. Thus, the second major process occuring in the kidney is tubular reabsorption. A number of discrete mechanisms, both active and passive and of varying degrees of specificity, are involved in tubular reabsorption. Many of these reabsorptive mechanisms are located in the cells of the proximal segments of the tubules. These cells account for the reabsorption of 65–90% of the glomerular filtrate. Glucose, certain cations, low molecular mass proteins, amino acids, and organic acids are actively reabsorbed. Water and chloride are passively reabsorbed as a result of the osmotic and electrochemical gradients generated by the active transport of sodium and potassium. The osmolarity of the fluid in the collecting duct is regulated in the loops of Henle; most of the remaining water and ions are reabsorbed in the distal tubules and collecting ducts. The rate of reabsorption in these segments of the proximal tubule is regulated to maintain the osmolar concentration of the blood. Most xenobiotics are also reabsorbed after glomerular filtration by passive diffusion during passage through the nephron. Passive tubular reabsorption of lipophilic xenobiotics is therefore greater than the reabsorption of polar xenobiotics or endogenous wastes.

Tubular Secretion. Xenobiotics present in blood may also be excreted by the kidney by tubular secretion. This secretion transports xenobiotics from the peritubular fluid (blood) to the lumen (urine) in the tubule. Tubular secretion is often selective; active transport mechanisms account for the secretion of many organic acids, including glucuronides and sulfates, and strong organic bases.

The secretion of weak bases and some weak acids may also occur by a passive mechanism that utilizes pH differences between peritubular fluid and urine. At the pH of the tubular lumen, these compounds become ionized and do not diffuse back across the cell wall.

Factors Affecting Renal Excretion of Xenobiotics. Xenobiotics are excreted by the same mechanisms which eliminate endogenous wastes; highly polar xenobiotics in plasma water are removed primarily by glomerular filtration and excreted in the urine with minimal involvement of tubular reabsorption or secretion. The rate of renal elimination of most of these xenobiotics is largely dependent on the rate of glomerular filtration. Since lipophilic compounds cross cell membranes more readily, they distribute into a much larger tissue volume than polar compounds, which are more likely to be restricted to the vascular volume. However, lipophilic xenobiotics metabolized to more polar compounds are usually rapidly returned to the circulation and are readily excreted. Therefore, the rate of metabolism of a xenobiotic may also play an important role in its rate of excretion.

2.5.8.2. *Hepatic Excretion*

Nutrients and xenobiotics are delivered from the gastrointestinal tract to the liver by the portal vein to be biotransformed there. Thus the liver is located between the intestinal tract and the general circulation and ideally located for the biotransformation of nutrients and xenobiotics taken up by this route. Besides participating in the biotransformation of xenobiotics, the liver is a major excretory organ and contributes to the excretion of many xenobiotics by eliminating them with bile into the gut and thus into feces [122-125].

The bulk of the liver consists of cells arranged in plates two cells thick. These plates are arranged radially around the terminal branches of the hepatic veins and are exposed to blood from the portal vein and hepatic artery flowing through interconnecting spaces referred to as hepatic sinusoids. The sinusoidal walls are freely permeable even to relatively large particles; special transport may only play a role in the uptake of certain anions from the blood. The epithelial hepatic cell is the smallest unit of the liver and accounts for most of the varied functions of this organ, including storage, secretion, biotransformation, and excretion (Fig. 42).

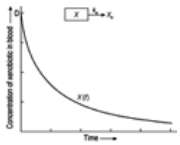

Figure 42. One-compartment model with first-order elimination and instantaneous absorption

Bile formation is thought to be the result of active transport of certain ionized compounds and passive transport of other solutes and water, which follow a concentration or electrochemical gradient. Active secretion of anions and cations appears to be controlled by different mechanisms, but the compounds actively excreted are usually amphipathic molecules and have both polar and nonpolar portions in their structures. Bile acids are the classical example of endogenous amphipathic molecules. Conjugates of lipophilic xenobiotics are also examples of amphipathic molecules, and many of these conjugates are ionized, a fact that facilitates excretion from the hepatocyte by active transport mechanisms.

As a result of both active and passive secretion into the bile, xenobiotics excreted into bile may be classified into different groups. Solutes found in bile may be divided according to their concentration in bile versus blood. For example, the excretion of Na^+, K^+, Cl^-, and glucose, for which the bile/blood ratio is close to unity, is thought to be passive. Conjugates of xenobiotics and bile acids have a bile/blood ratio of greater than 10 and are thought to be actively excreted from liver into bile. Other compounds such as proteins, inulin, sucrose, and phosphates do not cross the canalicular membrane of hepatocytes and have a bile/blood ratio of much less than one. Bile secreted by the liver cells into the bile canaliculi flows into the narrowest branches of the bile duct, the cholangioles, then into the hepatic duct, which carries the bile to the gallbladder and finally into the intestine. Lipophilic xenobiotics may appear in bile at low concentrations prior to metabolism. However, a compound absorbed from the stomach or intestine after ingestion is likely reabsorbed from the intestine if it is secreted in bile prior to structural modifications by biotransformation. Moreover, conjugates excreted into bile, such as glucuronides, may be hydrolyzed by enzymes present in the bacteria of the intestinal microflora, and the aglycone may be reabsorbed. Most xenobiotics reabsorbed from the intestine are returned to the liver. The process of excretion into bile, reabsorption from the intestine, and return to the liver is termed enterohepatic circulation. Enterohepatic circulation serves as an efficient physiological recovery mechanism for bile acids and certain hormones. When xenobiotics are trapped in this cycle, their rate of excretion from the body may be significantly reduced and their toxicity may be significantly increased.

Effect of Molecular Mass on Hepatic Excretion of Xenobiotics. From the liver cell, xenobiotics may be excreted into bile or returned to blood. The molecular mass of nonvolatile organic xenobiotics or their metabolites determines the primary route by which they are excreted from the hepatocyte. In the rat, xenobiotics, their metabolites and conjugates formed in phase-II reactions with molecular masses greater than 500 are most often excreted into bile, while xenobiotics and their metabolites with molecular masses of less than 350 return from the hepatocyte to blood and are thus delivered to the general circulation and to the kidney for excretion. In humans, the critical molecular mass threshold for biliary excretion is approximately 500.

2.5.8.3. *Xenobiotic Elimination by the Lungs*

Any xenobiotic present in blood with sufficient volatility will pass from the blood across the alveolar membrane into the air space of the lung and may be exhaled. The rate of elimination from blood of a volatile xenobiotic is dependent on the solubility of the xenobiotic in blood, the rate of respiration, and the blood flow to the lung. Xenobiotics like diethyl ether, which are highly soluble in blood, are rapidly eliminated by exhalation; their elimination may also be efficiently increased by a forced increase in the rate of respiration (hyperventilation). In contrast, volatile xenobiotics with low solubility in blood are only slowly cleared from the lung and thus from the body by exhalation, and their rate of exhalation may not be markedly influenced by hyperventilation. The proportionality among xenobiotic volatility, blood solubility, and the concentration of a volatile xenobiotic in the blood is utilized to quantitate blood alcohol content and thus estimate sobriety by breath analysis.

2.6. Toxicokinetics

The toxic response to chemical exposure depends particularly on the magnitude, duration, frequency, and route of exposure. These determine the amount of material to which an organism is exposed (the exposure dose) and hence the amount of material which can be absorbed (the absorbed dose). The latter determines the amount of material available for distribution and toxic metabolite formation, and hence the likelihood of inducing a toxic effect. Absorption and metabolite accumulation are opposed by elimination. All these factors define the disposition of the xenobiotic. The modeling and mathematical description of the course of disposition of a potentially toxic xenobiotic in the organism with time is termed toxicokinetics. Most of the methodologies and principles applied in toxicokinetics were first used to model the kinetics of chemicals applied as drugs (pharmacokinetics).

The goal of both toxicokinetics and pharmacokinetics is to quantitate the dynamic course of xenobiotic absorption, distribution, biotransformation, and elimination processes in living organisms with time. Both the whole process of disposition or individual steps such as elimination may be characterized.

2.6.1. *Pharmacokinetic Models*

A pharmacokinetic or toxicokinetic model is a functional representation that has the ability to describe the movement of a xenobiotic over time in a real biological system [126-128]. A common way to describe the kinetics of drugs is to represent the body as a number of interconnected compartments which may or may not have an anatomical or physiological reality. These compartments represent all tissues, organs, and fluids in the body that are kinetically indistinguishable from each other. A compartment might be represented by a cluster of cells within an organ, an organ, the blood, or the whole body taken together. More recently, so called physiologically based pharmacokinetic and toxicokinetic models have been developed. These models permit the modeling of the time course of the concentration of a chemical in a tissue on the basis of physiological considerations and may be particulary useful in interspecies extrapolations such as are necessary in risk assessment.

2.6.1.1. *One-Compartment Model*

This simplest toxicokinetic model depicts the body as a single homogeneous unit within which a xenobiotic is uniformly distributed at all times. The toxicokinetics of a xenobiotic may be analyzed by a one-compartment model if the determined plasma concentration after a single dose decreases exponentially (Fig. 43; the plot of the logarithm of the plasma concentration versus time yields a straight line). In this model, elimination of a chemical from the body occurs by first-order processes. A mathematical description of the first-order process is

$$\frac{dX}{dt} = -k_e X$$

where X is the amount in the body at time t and k_e is the rate constant for first-order elimination. In compartmental analysis, k_e is often referred to as an apparent first-order rate constant, to emphasize that the underlying processes may in reality only approximate first-order kinetics. For example, the xenobiotic might actually be eliminated by active biliary secretion, for which zero-order saturation kinetics would be observed under suitable conditions.

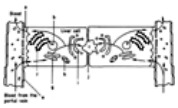

Figure 43. Structure of a liver cell (hepatocyte) and possible pathways for the uptake and elimination of xenobiotics and their metabolites

a) Space of Disse; b) Nucleus; c) Bile; d) Liver capillary; e) Sinusoidal wall; f) Rough endoplasmic reticulum; g) Smooth endoplasmic reticulum; h) Mitochondria; i) Golgy complex; j) Bile duct

If the xenobiotic is not present in the organism and a known amount X_0 is rapidly administered, the total amount of the xenobiotic initially in the body will be approximately X_0 = total dose. The amount of the xenobiotic present in the organism at a certain time after administration is then

$$X(t) = X_0 \times e^{-kt}$$

The first-order rate constant of elimination can be determined from the slope of the plot of the logarithmic plasma concentration versus time and can be used to estimate the half-life of elimination for the xenobiotic:

$$t_{1/2} = \frac{0.693}{k_e}$$

Within seven half-lives, a xenobiotic is almost completely eliminated (99.2%), although theoretically complete elimination will never be achieved. Important characteristics of the first-order elimination of a xenobiotic according to the one-compartment model are:

1. The half-life of the xenobiotic is independent of dose; the semilogarithmic plot of the plasma concentration of the xenobiotic versus time yields a straight line.
2. The concentration of the xenobiotic in plasma decreases by a constant fraction per time unit.

Xenobiotics responsible for toxic effects are usually not injected into blood and intake is thus not instantaneous compared to distribution and elimination and a lag period before maximal concentration in plasma is observed. If intake also approximates a first-order absorption process, then the rates of absorption and elimination will determine the time course of the plasma concentration (Fig. 44).

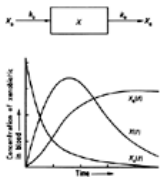

Figure 44. One- compartment model with first-order absorption and first-order elimination

The rate of change in the concentration of a xenobiotic in a one-compartment model with first-order absorption and elimination may be described by:

$$\frac{dX}{dt} = -k_e X_e + k_a X_a$$

where k_a is the rate constant of absorption, X_a is the amount of the xenobiotic present at the site of absorption, and k_e is the rate constant of elimination.

The rate constant of elimination can be experimentally determined by treating the elimination phase as if it occurred after the intravenous injection described above by extrapolation of the terminal straight line of the semilogarithmic plot to the ordinate at a putative X_0. By determing the plasma concentration of the xenobiotic at time points shortly after administration, the rate constant for absorption can be determined by plotting the difference between the early plasma concentrations and the extrapolated portion of the elimination curve.

2.6.1.2. *Two-Compartment Model*

If a xenobiotic does not distribute and equilibrate throughout the body rapidly, a two-compartment model provides a better description of the kinetics of disposition. In this case, the semilogarithmic plot of plasma concentration versus time does not yield a straight line. In two-compartment models, the main or central compartment is assumed to represent the blood and highly perfused organs and tissues such as the liver, the heart, and the kidneys, which are in rapid distribution equilibrium with the blood, while the second or peripheral compartment corresponds to poorly perfused tissues such as muscles and fat. This model is depicted in Figure 45 and is called a two-compartment open pharmacokinetic model.

Figure 45. Two-compartment model for the distribution of a xenobiotic

X_c = concentration in the central compartment; X_p = concentration in the peripheral compartment; X_0 = initial dose; k_{21}, k_{12}, k_{10} = rate constants

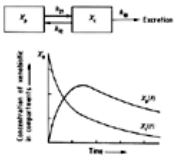

The time course of the plasma concentration of the xenobiotic can be decribed by two overlying monoexponential terms of the type:

$$X = Ae^{-kt} + Be^{-k't}$$

where *A* and *B* are proportionality constants and *k* and *k′* are rate constants. Thus, the pharmacokinetics of a two-compartment model describe a biexponential decay in the amount of xenobiotic in the body. Analogous to a one-compartment model, a distribution equilibrium is assumed to exist within each compartment that can be expressed in terms of blood concentration. According to this equation, a plot of the logarithm of blood concentration versus time will yield a biphasic decay from which the constants *A*, *B*, *k*, and *k′* can be estimated either graphically or by nonlinear regression analysis. Once these experimental constants have been determined, the pharmacokinetic parameters can be calculated. The numerical values of these parameters can aid in assessing the relative importance of tissue distribution and elimination to the disposition of the xenobiotic.

2.6.2. *Physiologically Based Pharmacokinetic Models*

Although compartmental analysis models are convenient to use and provide useful descriptions of the overall time course of xenobiotic disposition, the practical limitations of curve fitting to experimental data generally restrict such models to one or two compartments. Alternatively, it is often possible to develop a compartmental model from physiological principles and thereby circumvent the inherent limitations of curve-fitting analysis. A physiological pharmacokinetic or toxicokinetic model is a mathematical description of the disposition of a xenobiotic in the organisms or in a part thereof (e.g., a specific organ) Such a model is constructed by using physiological and biochemical parameters such as blood flow rates, tissue and organ sizes, binding, and biotransformation rates [129-131]. These are generally more complex and require the specification of many parameters; thus, physiological compartmental models are still simplified representations of biological systems. In addition, the accurate determination of physical and biochemical parameters is often both difficult and inaccurate. However, the physiological framework provides several advantages: the physical definition of compartments and transfer rates facilitates the incorporation of existing knowledge about the quantitative behavior of biological systems into the model; physiological changes with time during chronic exposure to a xenobiotic, such as those due to physical growth, or induction of biotransformation and changes in excretion rates, can be introduced (Fig. 46).

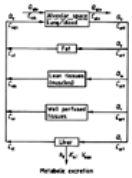

Figure 46. Schematic representation of a physiologically based toxicokinetic model

Most important, however, a reasonable basis exists for extrapolating the kinetics of a xenobiotic to predict the disposition of a xenobiotic following various types and patterns of exposure, and to extrapolate from experimental data obtained in one species of experimental animals to other animal species and humans.

Physiology-based models use the concepts of mass balance and flow-limited transport under the following definitions: Blood serves to distribute a xenobiotic from the site of absorption to the other parts of the body. In the normal sequence of events, a chemical entering the bloodstream is often rapidly distributed within the blood, and its blood concentration can be considered to be essentially uniform.

The chemical enters and leaves the compartment with the blood flow and diffuses or is transported from blood to tissue and back. The chemical may also undergo a variety of physical interactions, such as binding to macromolecules, and the result is a partitioning between tissue and blood that depends on the affinity of the chemical for each medium. Diffusion or transport directly between adjacent compartments, enzymatic biotransformations, and excretion may also occur. The net result of all of these changes is expressed by a mass-balance differential equation, which is simply a mathematical statement of the conservation of mass. Frequently, these mass balance equations can be greatly simplified, since many of the terms required may not apply to a particular compartment.

[Top of Page]

3. Mechanisms of Acute and Chronic Toxicity and Mechanisms of Chemical Carcinogenesis

3.1. Biochemical Basis of Toxicology

Since the 1960s, toxicology has moved from observing and classifying the harmful effects of chemicals in animals with the tools of pathology (descriptive toxicology), to a discipline able to explain the mechanisms of the basic changes in cell function responsible for toxic effects (mechanistic toxicology). This progress resulted from the widespread application of techniques and concepts from a range of basic sciences, most notably biochemistry and cell biology. Mechanistic explanations of toxic phenomena aid in the prevention of chemical or biological toxicity and provide a rational basis for the use of animal data to assess the anticipated consequences of human exposure to a particular chemical.

Many toxic compounds are chemically stable and produce their characteristic effects by interference with biochemical or physiological homeostatic mechanisms. Many adverse events are the consequence of disturbance of normal physiology and do not result in cell death. A xenobiotic may, for example, activate plasma membrane receptors and induce physiological signal transduction pathways. Induction of a normal cascade at the wrong time or to the wrong extent (too much) may cause undesirable or harmful effects. Toxic compounds interfering with homeostatic mechanisms do not necessarily cause cell death; however, the induced changes may have a harmful impact both on the altered cells and on the involved tissue or on the entire organ. Therefore, it is critical to have a proper knowledge on biochemical and molecular sites of action of the xenobiotics.

Cytotoxicity resulting in cell death is often the consequence of exposure to a harmful xenobiotic, but the number of cells which must be killed before the function of a tissue or organism is noticeably impaired is highly variable. Some cell types like the epithelia of the kidney and the liver have the ability to regenerate in response to damage, while others like neurons can not. Furthermore, some organs, such as the liver, lung, and kidney, have a substantial functional reserve capacity in excess of normal requirements, and normal function can be maintained even in the presence of extended necrosis.

In addition to cell death, disturbances in the regulation of cell division induced by toxic xenobiotics may have harmful long-term consequences for the organism affected. Nonlethal alterations in the genome of somatic cells can result in mutations and can lead to malignant transformation and tumor formation [132]. More recently, it has become clear that compounds not directly interacting with the genomic DNA can also produce cancer by so-called epigenetic mechanisms [133]. These may involve a proliferative response of epithelial cells to cytotoxicity, as is suggested to occur with high-dose carcinogens such as allyl isothiocyanate and chloroform, or a more direct action that enhances the rate of cell division in the absence of cytotoxicity, as is seen with 2,3,7,8-tetrachlorodibenzodioxin. Increased cell replication, whatever the cause, is accompanied by increased chance of unrepaired DNA lesions that may be fixed as mutations. Hyperplasia has long been suspected of preceding neoplasia, but the inevitability of such a progression has never been established on a pathological basis alone. Even in absence of foreign compounds, DNA is damaged to a considerable extent by reactive oxygen species formed during different biochemical processes in the cell, it may thus be speculated that nongenotoxic carcinogens act by enhancing the likelihood of this normal DNA damage being fixed as a mutation and leading to cancer [134].

Having established the fundamental differences between toxic chemicals which act through physiological inbalance, through cytotoxicity, or by causing alterations in cell proliferation, we now consider some sites or chemical reaction mechanisms of toxic action. The following major mechanisms are neither comprehensive nor mutually exclusive.

3.2. Receptor-Ligand Interactions

3.2.1. *Basic Interactions*

Some xenobiotics can interact with physiological receptors due to structural similarities to endogenous compounds. The toxic effects of these chemicals are related to

their ability to interfere with normal receptor–ligand interactions either as agonists or as antagonists of the physiological ligand. Receptors are macromolecular components, most often proteins, of tissues which interact with specific endogenous ligands or structurally related xenobiotics to induce a cascade of biochemical events and produce characteristic biological effects. The binding between a receptor (R) and a ligand (L) is usually reversible and can be described by the equilibrium reaction:

$$R + L \underset{k_2}{\overset{k_1}{\rightleftharpoons}} RL$$

The dissociation constant K_d that describes this relationship is thus given by:

$$K_d = \frac{k_1}{k_2} = \frac{[L][R]}{[LR]}$$

where [L], [R], and [LR] are the concentrations of ligand, unbound receptor, and ligand-bound receptor, respectively. The affinity of the ligand to a certain receptor is proportional to $1/K_d$. The ligand may be an endogenous substance that interacts with the receptor to produce a normal physiological response, or it may be a xenobiotic that may either elicit (agonist) or block (antagonist) the response.

Receptor–ligand interactions are generally highly stereospecific. Usually, small changes in the chemical structure of the ligand can Drastically influence its capability to bind to the receptor and thus reduce or completely abolish the effect elicited by the ligand–receptor interaction. In other cases, changes in the chemical structure may have an important impact on the response without altering the binding of the xenobiotic to the relevant receptor. Differences in the extent of activity are not only observed with structurally distinct chemicals but also with chemicals that are chiral and thus may be present as racemic mixtures of stereoisomers. Synthetic chemicals with chiral centers generally contain both enantiomers, often in a 1:1 ratio, yet in many instances only one of the enantiomers is biologically active. In pharmacological or toxicological studies, the inactive or weakly active enantiomer should be viewed as an “impurity” which may confound the interpretation of results obtained with the racemic mixture. Because the enzymes responsible for biotransformation of xenobiotics also contain active sites with specific steric requirements, stereoisomers may be differentially biotransformed. The selective biotransformation of only one enantiomer may markedly alter the potency and efficacy of one enantiomer compared with another. Stereoselective differences in action among enantiomers should not be surprising, if one considers that these chemicals are mirror images, much as the right hand is of the left, and that receptors have a physical orientation that can be compared with a glove. Although the left-hand and right-hand gloves look remarkably similar, they do not fit both hands equally well. However, these theoretical considerations do not imply that stereoisomers always elicit different responses. Since the chiral center of a stereoisomer is not necessarily involved with the active site of the receptor, chirality does not always results in differences in toxic response among enantiomers.

The acute toxic effects of many xenobiotics are related directly to their ability to interfere with normal receptor–ligand interactions. This is most clearly the case with neurotoxins, acting within and outside the central nervous system (CNS). The proper function of the nervous system is highly dependent on a diverse array of receptor–ligand interactions. For example, the belladonna alkaloids atropine and scopolamine bind to and block the cholinergic receptors, the binding site of the physiological agonist acetylcholine (Fig. 47). The interactions of acetylcholine and atropine are an excellent example demonstrating that interference of a xenobiotic with physiological regulatory mechanisms may result both in desired (pharmacological) and an undesired (toxic) effects.

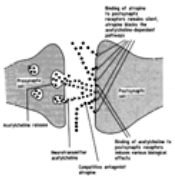

Figure 47. Induction of physiological response(s) by binding of the agonist acetylcholine to postsynaptic receptors; in contrast, binding of the antagonist atropine to the same receptors does not induce biological effects

Since atropine itself can not elicit the physiological responses mediated by acetylcholine binding to this receptor, toxic effects may result. In clinical therapy, atropine is often used as an antispasmodic to reduce the hypermotile state of the gastrointestinal tract or the urinary bladder. Its anticholinergic effects are also utilized in cardiovascular pharmacology to decrease pathologically elevated heart rates. At the same time, however, blockage of acetylcholine-mediated physiological cascades may cause dry mouth, blurred vision, and constipation. In addition, in the CNS toxic effects of atropine include restlessness, confusion, hallucinations, and delirium. Blockage of the receptors for the neutral amino acid glycine by strychnine represents a further example of the important and, in this particular case, life-threatening agonist–antagonist interactions. Binding of glycine to its receptor induces an increased permeability of the plasma membrane to chloride ions resulting in hyperpolarization and reduced activity of nerve cells, that is, glycine exerts an inhibitory effect on the nervous system. Blockage of the glycine receptors by the antagonist strychnine, that does not elicit a response, results in hyperactivity and severe convulsions.

In addition to acute toxicity, chronic influences exerted by xenobiotics on steroid hormone homeostasis may result in adaptive and proliferative responses in certain tissues. The long-term consequences may involve impairment or loss of the physiological function. Moreover, binding to a specific receptor may cause changes in gene expression and contribute to tumor formation. For example, binding to the Ah receptor (the aryl-hydrocarbon receptor, named on the basis of its interaction with planar, polycyclic aromatic compounds) mediates many of the effects observed with the highly toxic and carcinogenic 2,3,7,8-tetrachlorodibenzodioxin (TCDD). A cytosolic binding protein (cytosolic receptor) has also been identified for peroxisome proliferators, such as clofibrate and diethylhexyl phthalate [135, 136]. This protein has DNA-binding domains and appears to activate gene transcription in the nucleus in a manner analogous to the steroid hormones. The increased cell proliferation caused in target organs such as the liver may be responsible for the carcinogenicity of many peroxisome proliferators. Characteristic examples of receptor–xenobiotic interactions with harmful acute or chronic consequences are summarized in Table 20.

Table 20. Chemicals causing toxic response by interaction with receptors (Ah = aryl hydrocarbon)

Type of receptor	Chemical
Muscarinic receptor	atropine
Glycine receptor	strychnine
Ah receptor	TCDD
Peroxisome proliferator receptor	peroxisome proliferators (clofibrate, diethylhexylphthalate)

3.2.2. *Interference with Excitable Membrane Functions*

The maintenance and stability of excitable membranes is essential to normal physiology. Excitable membranes are critical to the function of nerves and muscles due to their ability to generate and propagate action potentials. Action potentials are elicited by the exchange of ions between the intra- and extracellular compartments, and hence they depend on the normal activity of ion channels and membrane ion pumps.

Xenobiotics may perturb excitable membrane functions in many different ways. Xenobiotics can block sodium channels resulting in toxic effects such as paralysis. For example, tetrodotoxin from the the puffer fish irreversibly blocks the sodium channel along the nerve axon and thus prevents the inward sodium current of the action potential while leaving the outward potassium current unaffected. The marine toxin saxitoxin, which is structurally quite different from tetrodotoxin, also produces its paralyzing effects by blocking sodium channels in excitable membranes in essentially the same manner. The insecticide dichlorodiphenyltrichloroethane (DDT) produces its neurotoxic action by interfering with the closing of sodium channels, which impairs the physiological repolarization of excitable membranes. This results in hyperactivity of the nervous system and repetitive discharges of neurons. The most striking symptoms in poisoned insects or mammals are persistent tremor and convulsive seizures. Other neurotoxins such as cyclodiene insecticides and picrotoxin interact with the γ-aminobutyrate (GABA) receptor in neurons. Blockage of the GABA receptor by these antagonists impairs the transport of chloride ions into the cell, which results in partial depolarisation of the membranes in the absence of adequate signals. The clinical symptoms of the pathophysiological changes can be summarized as a state of uncontrolled excitation. In addition, with cyclodiene insecticides this hyperexcitability is augmented by the inhibition of Ca^{2+}–Mg^{2+}–ATPase, which results in accumulation of intracellular free calcium and increased release of neurotransmitters from storage vesicles (Fig. 48).

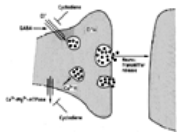

Figure 48. Sites of action of chlorinated cyclodiene insecticides on the plasma membrane of nerve cells

Organic solvents have a depressant effect on the CNS that results in narcosis by nonspecific alterations in membrane fluidity due to their lipid solubility rather than by interference with specific ion channels. This depressant effect is not based on certain structural features or specific interactions, that is, it is also observed after inhalation of the inert, but highly lipophilic gas xenon.

3.2.3. *Interference of Xenobiotics with Oxygen Transport, Cellular Oxygen Utilization, and Energy Production*

Continuous, sufficient production of energy in form of adenosine triphosphate (ATP) in specialized cellular organelles (mitochondria) is essential for cell survival and function. The process of cellular energy production requires oxygen, cofactors, and a wide array of specialized enzymes acting in concert. Many chemicals produce their toxic effects by interfering with the transport and cellular utilization of oxygen or with the oxidation of carbohydrates, which is coupled to the synthesis of ATP by oxidative phosphorylation.

The interference with cellular energy production may occur at different sites in the uptake, transport, and cellular utilization of oxygen. Oxygen transport from the lung to tissues may be reduced or blocked by xenobiotics which compete with oxygen for the binding site in the transport protein hemoglobin or which chemically modify this binding site. Carbon monoxide competitively blocks the binding of oxygen to the transport protein hemoglobin. Chemical oxidation of the iron in hemoglobin also impairs oxygen transport. Some xenobiotics oxidize the Fe^{2+} at the oxygen binding site in hemoglobin to Fe^{3+}. Hemoglobin with Fe^{3+}, known as, methemoglobin, can not reversibly bind oxygen. Thus, xenobiotics causing methemoglobinemia (e.g., nitrites, aromatic amines) effectively block oxygen transport.

The cellular utilization of oxygen in the tissues is blocked by cyanide, hydrogen sulfide, and azide because of their affinity for cytochrome oxidase. Cyanide exerts its toxic effects by interrupting electron transport in the mitochondrial cytochrome electron-transport chain. In addition, interferences with the enzyme systems producing ATP is a well-defined mechanism of toxicity. The ultimate formation of ATP in the cell by the oxidation of carbohydrates may also be blocked at other sites. For example, rotenone and antimycin A interfere with specific enzymes in the electron-transport chain, nitrophenols uncouple oxidative phosphorylation, and sodium fluoroacetate inhibits the citric acid (Krebs) cycle. Nitrophenols interfere with the production of high-energy phosphates in mitochondria. They prevent the phosphorylation of adenosine diphosphate (ADP) to ATP, but electron flow and oxygen consumption continue. In addition to inducing loss of cell function due to depletion of ATP, this type of uncoupler also causes a marked elevation of body temperature due to excess heat production. Toxicity resulting from blockage of the tricarboxylic acid cycle occurs in organs heavily relying on the availability of a continuous energy supply and results in cardiac and nervous system toxicity. A classic example for a mechanism inhibiting the Krebs cycle is termed “lethal synthesis” and is exemplified by fluoroacetate, which is incorporated into the Krebs cycle as fluoroacetyl coenzyme A, inhibits the aconitase-catalyzed conversion of citrate to isocitrate, and interrupts the Krebs-cycle (Fig. 49). The Krebs cycle is the major degradative pathway for the generation of ATP and also provides intermediates for biosynthesis. For example, the majority of the carbon atoms in porphyrins comes from succinyl CoA and many of the amino acids are derived from α-ketoglutarate and oxalacetate. Hence, interruption of the Krebs cycle by the incorporation of fluoroacetate may be lethal for the cell.

Figure 49. Lethal synthesis

Incorporation of fluoroacetate in the Krebs cycle; formation of fluorocitrate and inhibition of aconitase results in blocking of the tricarboxylic acid cycle and interruption of cellular energy production

The consequences of ATP depletion include impairment of membrane integrity, ion pumps, and protein synthesis. Depending on its extent, energy depletion will inevitably lead to loss of cell function and cell death.

3.3. Binding of Xenobiotics to Biomolecules

Many toxic chemicals exert their effects by covalent linkage of reactive metabolites to essential macromolecules of the cell. Xenobiotics may become covalently bound to the active site of enzymes or to other macromolecules whose function is critical to the cell. These include proteins (Section Binding of Xenobiotics or their Metabolites to Cellular Proteins) and lipids (e.g., as membrane structural elements; see Section Interaction of Xenobiotics or their Metabolites with Lipid Constituents) and nucleic acids (see Section Interactions of Xenobiotics or their Metabolites with nucleic Acids). The linkage generally involves binding of electrophilic metabolites to nucleophilic sites such as thiol, amino, and hydroxyl groups in the side chains of proteins and is essentially irreversible; it depends only on the turnover of the macromolecule in question for the repair of the lesion. In addition, other toxins may bind to proteins and impair their normal function without first being converted to reactive intermediates.

3.3.1. *Binding of Xenobiotics or their Metabolites to Cellular Proteins*

Many toxic substances exert their effects via binding to the active sites of enzymes or other proteins that are critical to cellular function. For example, hydrogen cyanide binds with high affinity to the Fe^{3+} ion in cytochrome oxidase and thus blocks the last step in the mitochondrial electron-transport chain, which is important for cellular energy production. This single and highly specific site of action is responsible for the rapid induction of the often fatal toxic effects of cyanide. Carbon monoxide binds to the reduced form of iron in hemoglobin, whose physiological function is to bind, transfer, and deliver oxygen to tissues. Since carbon monoxide has a 210-fold higher affinity to this binding site than oxygen, very low concentrations of carbon monoxide in the atmosphere are sufficient to displace the physiological ligand and produce severe toxic effects. Thus, 0.1% carbon monoxide in the atmosphere can occupy roughly 50% of the available hemoglobin binding sites, because on the basis of the 210-fold higher affinity to hemoglobin compared with oxygen, 0.1% carbon monoxide is equivalent to 20% oxygen in the atmosphere.

Another example for the importance of protein binding in toxicity is the binding of metal ions to protein thiol groups. Many toxic metals such as lead, mercury, cadmium, and arsenic bind to proteins with free sulfhydryl groups, which contributes by as-yet largely unknown mechanisms to the toxicity of these metals. Induction of porphyria by lead, mercury, and other metals is based in part on the inhibition of specific enzymes of heme biosynthesis and results in the accumulation of specific intermediates in heme synthesis. The binding of cadmium to the sulfhydryl-rich protein metalothionein results in active concentration of the metal-metalothionein complex in the proximal tubules of the kidney. Catabolism of the metal complex in the lysosomes of proximal tubule cells with concomitant liberation of the toxic metal results in nephrotoxicity, a common effect of cadmium exposure in humans and experimental animals [137].

Binding to active sites in enzymes may result in the inhibition of biochemical pathways vital to the cell. The induction of toxic effects may be due to accumulation of a specific enzyme substrate or insufficient amount of substrate available for normal physiological function. For example, many organophosphate pesticides inhibit cholinesterases by covalently binding to the active site, the amino acid serine of cholinesterase. Since the phosphorylated cholinesterase is stable, the covalent binding results in inhibition of enzymatic activity (Fig. 50). Due to the impaired cleavage, acetylcholine accumulates at cholinergic synapses and neuromuscular junctions. Organophosphates thus produce the typical signs of acetylcholine poisoning, such as increased salivation and lacrimation, abdominal cramps and diarrhea, cough, bronchoconstriction and breathing impairment, mental confusion, headaches, tremor, and coma.

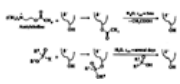

Figure 50. Mechanisms of cholinesterase inhibition by organophosphate pesticides

Compare the $t_{1/2}$ of the enzyme complex with the physiological substrate acetylcholine (a few milliseconds) to the $t_{1/2}$ of the enzyme complex with organophosphate pesticides (several days).

Protein modifications may also be the basis for immunosuppressive effects and chemical-induced allergy. For many chemicals causing necrosis, binding to cellular macromolecules is essential for the expression of toxic effects. However, in most cases the temporal sequence of events that is triggered by covalent binding and the cause–effect relationships between these events are not fully understood.

3.3.2. *Interaction of Xenobiotics or their Metabolites with Lipid Constituents*

Lipid peroxidation in biological membranes by free radicals initiates a series of events finally causing cellular dysfunction and cell death. The formation of radicals during peroxidation is a self-propagating process; the reaction may be started by organic radicals formed during biotransformation or by oxygen radicals formed by disruption of cellular energy metabolism. The initiation of lipid peroxidation by interaction of free radicals with polyunsaturated fatty acids to form lipid peroxy radicals, which then produce lipid hydroperoxides and other lipid peroxy radicals, has been proposed as a critical step leading to cell injury and death. Peroxidative damage to plasma membrane lipids may cause impairment of membrane integrity and, finally, rupture of the plasma membrane. In addition, breakdown of the membranes of subcellular organelles such as those of the mitochondria, the endoplasmic reticulum, and lysosomes may also contribute to induction of cell death. The end products of the breakdown of membranes, mainly unsaturated aldehydes, may also produce toxicity in distal tissues. Induction of lipid peroxidation is involved in the toxicity of many chemicals which are converted to free radicals. A well-studied example is carbon tetrachloride, which is converted by cytochrome P450 to the trichloromethyl radical (CCl_3) and the trichloromethylperoxyradical (CCl_3O_2). Radicals and reactive oxygen species may interact with the major soluble cellular thiols, glutathione (GSH), and thiol-containing proteins. Depletion of cellular glutathione and modification of thiol-containing proteins by oxidation or mixed disulfide formation with glutathione and other low molecular mass thiols may results in oxidative stress in the cell.

An important function of reduced glutathione is to protect the sulfhydryl groups of proteins by keeping them in the reduced state. Depletion of intracellular glutathione stores appears to be a prerequisite for the onset of significant oxidative stress. Many critical enzymes in the cell depend on reduced thiol groups (SH groups) to maintain their activity; hence, concentrations of reactive oxygen species in excess of that necessary to deplete intracellular glutathione can induce oxidation of protein thiols to form disulfide linkages, thereby impairing enzymatic activity. Although the direct covalent interaction of electrophilic chemicals with protein thiols may contribute to enzyme inhibition; it appears that reversible oxidation of the thiol group as a result of oxidative stress plays a more important role. One group of thiol-containing enzymes whose impairment as a result of oxidative inactivation may play a critical role in cell injury and death are the Ca^{2+}-transporting membrane systems.

3.3.3. *Interactions of Xenobiotics or their Metabolites with nucleic Acids*

Electrophilic compounds, usually formed by oxidative biotransformation of xenobiotics, may also interact with various nucleophilic sites in DNA, principally O-6, N-7, N-2, and C-2 of guanine. Moreover, other types of DNA damage may occur as a consequence of covalent binding of electrophiles or by interaction of DNA with reactive oxygen metabolites (Table 21).

Table 21. Types of DNA damage induced by xenobiotics

Type of DNA-damage	Examples
Modifications of purines and pyrimidines	alkylation by electrophiles, oxidation to 8-hydroxyguanosine and other oxidized purines, UV-induced formation of thymidine dimers
DNA strand breaks	oxygen radicals, activation of calcium-dependent endonucleases
AP lesions (apurinic, apyrimidinic sites)	loss of DNA bases through labilization after covalent binding of xenobiotics to specific sites in purine and pyrimidines
DNA cross-links	bifunctional electrophiles react with nucleophilic sites in both DNA strands (interstrand cross-link) or with two bases on one strand (intrastrand cross-link)

These interactions may alter gene expression (see Sections Nonlethal Genetic Alterations in Somatic Cells and Carcinogenesis and Molecular Mechanisms of Malignant Transformation and Tumor Formation). The changes in gene expression may be quantitative, that is, formation of the wrong amount of a protein in an inappropriate time period of cellular life, or qualitative, that is, formation of a protein with altered or impaired properties. In both cases, the changes may cause the death of the cell. However, interaction of xenobiotics with nucleic acids is more important in generating somatic mutations which can be the initiating event for a process ultimately leading to malignant transformation of cells and tumor growth. Alkylation of the O-6 position of guanine appears important in the mutagenicity and carcinogenicity of nitrosamines and other chemicals that readily form methyl carbonium ions. In addition, other sites such as the N-7, N-2, and C-2 positions of guanine may also play an important role in DNA adduct formation with other electrophilic chemicals. Ribonucleic acid (RNA) also contains nucleophilic sites, and thus critical intracellular functions of RNA such as protein synthesis may be perturbed by covalent interaction of electrophilic chemicals with RNA.

3.4. Perturbation of Calcium Homeostasis by Xenobiotics or their Metabolites

Intracellular calcium concentrations are rigorously maintained at around 10^{-7} M against an extracellular concentration of more than 10^{-3} M. Three principal buffer systems are important in maintaining this steep gradient in virtually all cells: the plasma membrane, the mitochondria, and the endoplasmic reticulum. All systems require ATP to transport calcium, directly or indirectly. Exposure of freshly isolated or cultured cells to numerous toxins such as *tert*-butyl hydroperoxide, quinones, paracetamol, and carbon tetrachloride induces a rapid and sustained rise in cytosolic calcium concentrations, which correlates well with the subsequent loss of cell viability [138, 139]. Prevention of this rise by omitting calcium in the extracellular buffer or by addition of calcium chelators may prevent cell death. A central event in calcium-induced toxicity is the activation of calcium-dependent degradative enzymes such as proteases, phospholipases, and endonucleases (Fig. 51).

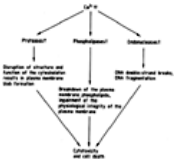

Figure 51. Increased concentrations of intracellular calcium may activate calcium-dependent degradative enzymes, resulting in the destruction of proteins and lipids and in DNA fragmentation, thus causing cell death

Cytoskeletal integrity is important for a number of cellular functions including motility, shape, secretion, and division. Activation of calcium-dependent proteases results in disruption of cytoskeletal components and formation of plasma membrane blebs (protrusions). In addition, activation of calcium-dependent phospholipases may induce membrane phospholipid breakdown, and calcium-dependent endonucleases may lead to DNA fragmentation by the formation of DNA double-strand breaks. These toxic biochemical cascades may act in concert to cause cell death [104].

3.5. Nonlethal Genetic Alterations in Somatic Cells and Carcinogenesis

The induction of cancer by chemicals is a complex multistep process involving interactions between environmental and endogeneous factors. Tumors are formed as a result of aberrant tissue growth due to loss of control mechanisms of cell division. A model traditionally used for an operational description of carcinogenesis is the initiation–promotion model. Initiation, the first stage, requires a genotoxic event such as binding of an electrophilic xenobiotic to DNA causing a premutagenic lesion in a single cell [140, 141]. After DNA replication, the premutagenic lesion may be transformed into a heritable mutation (i.e. by base-pair substitution; see Section Molecular Mechanisms of Malignant Transformation and Tumor Formation). In the promotion stage, several, primarily nongenotoxic, mechanisms facilitate (promote) the preferential proliferation of the initiated cell and finally resulting in the formation of a tumor. Initiators are usually genotoxic agents, whereas promoters generally act by interfering with extranuclear sites and processes; most promoters increase cell growth and cell proliferation.

In contrast to initiators, which are believed to result in irreversible changes in the cellular genome without thresholds, promoters show some reversibility, and thresholds for promoters can be discussed. This stepwise nature of carcinogenesis was first shown in mouse skin, and this model has now been well characterized (Fig. 52). In a typical experiment, an initiating chemical such as dimethylbenzanthracene is applied to mouse skin at a low dose so that very few tumors, if any, are produced in the animals lifetime. After an interval of one week to one year, the treated (initiated) skin is exposed to multiple applications of a promoter such as the phorbol esters found in croton oil. Tumors begin to appear within 5 to 6 weeks after application of the promotor, and all mice carry tumors by 10 to 12 weeks after the start of application (Fig. 52).

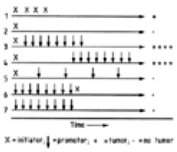

Figure 52. Schematic presentation of the initiation – promotion model of chemical carcinogenesis in the mouse skin

Promoters reduce the latency period and increase the number of tumors only when applied after the initiator (3, 4). The application frequency required for promoting effects to occur depends on the initiating carcinogen and the target tissue. If the application frequency of the promoter is too low, the promoting effects are not observed (i.e., no tumors are formed or the tumor yield is not increased, 5). This is also the case, when the promoter is applied prior to the initiator (6) or alone without initiating agent (7). Tumors can also be produced by the initiating agent alone if it is applied at sufficient concentration or above a certain application frequency (1). In this case, the initiator is termed a complete carcinogen.

The initiation–promotion experiments on mouse skin and similar experiments in the rodent liver have led to the following general rules of carcinogenesis:

1. The initiator must be given first; no tumors or very few tumors result if the promoter is given first.
2. The initiator, if administered once at a subcarcinogenic dose, does not produce tumors during the time of observation; however, repeated doses of the initiator may cause tumors even in the absence of the promoter (the initiator is a complete carcinogen in this case).
3. The action of the initiator is irreversible; tumors result in nearly the same yield if the interval between initiation and promotion is extended from one week to one year.
4. The initiator is an electrophile, or is bioactivated to an electrophile, which binds covalently to DNA causing a mutation after DNA replication.
5. In contrast, there is no evidence of covalent binding of promoters or their metabolites to DNA.
6. The action or the promoter is reversible at an early stage and usually requires repeated exposure; thus, there is probably a threshold level of exposure to promoters. However, threshold levels can not be reliably defined, as long as the biochemical mechanism involved in tumor promotion is not precisely known. In contrast to the established target of the initiation process (DNA), the molecular mechanisms of tumor promotion are still largely unknown.

In addition to promotion of skin carcinogenesis by the phorbol esters, there are known or suspected promoters for tumors in other organs [142, 143]. Bile acids are known to be promoters of colon carcinogenesis in experimental animals. In humans, there is a strong association between high intake of dietary fat and cancer of the colon; since ingestion of fat increases the amount of bile acids in the colon, the increased incidence of colon cancer may be due to the promoting effect of the bile acids on intestinal epithelia. In rat bladder, saccharin and cyclamate are promoters for tumors initiated by a single dose of dimethylnitrosourea; tryptophane is a promoter for urinary bladder tumors in dogs treated with an initiating dose of 4-aminobiphenyl or 2-naphthylamine. Hormones are also known modifiers of chemical carcinogenesis. Oral or intravenous administration of dimethylbenzanthracene produces mammary tumors in susceptible female mice. Prolactin will increase and accelerate tumor development, whereas ovariectomy results in reduced tumor yield.

In addition to these two stages a third stage, termed progression, has since been established as an integral part of the carcinogenic process. In this last stage additional mutational events increase the malignancy of the tumor, that is, the tumor grows in an invasive manner, destroying the surrounding tissues and forming metastases in other organs (Fig. 53).

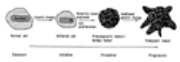

Figure 53. Multistep model of tumor formation

Genotoxic events resulting in heritable mutations cause the formation of an initiated cell (initiation). In the following stage of promotion, nongenetic (epigenetic) events contribute to preferential proliferation of the initiated cell. In the third stage of progression, additional genetic events increase the malignancy of the tumor tissue (i.e., its growth becomes increasingly destructive and metastases are formed in other organs).

The understanding of the molecular mechanisms on route from DNA interactions to a clinically observable tumor requires some knowledge on the basic roles of DNA biochemistry and biology in cellular function and replication. Therefore, a short chapter on DNA biochemistry and biology follows. Further and in-depth information is presented in textbooks on biochemistry and molecular biology.

3.6. DNA Structure and Function

3.6.1. *DNA Structure (* ⟶ Nucleic Acids – Structure of DNA*)*

With the exception of certain viruses, the genetic information of all cells is contained in deoxyribonucleic acid (DNA), whose structure and constituents permit the accurate storage of a vast amount of information. In the cell nucleus, the DNA is packaged with proteins to form chromatin.

The genetic code of the DNA is denoted by four letters: two pyrimidine nitrogenous bases, thymine (T) and cytosine (C), and two purine bases, guanine (G) and adenine (A). These are are functionally arranged in codons (or triplets). Each codon consists of a combination of three letters and codes for a specific amino acid.

The bases on one strand are connected together by a sugar (deoxyribose) phosphate backbone. DNA can exist in a single-stranded or double-stranded form. In the latter state, the two strands are held together by hydrogen bonds between the bases. The adenine on one strand binds to thymine on the sister strand, and guanine pairs with cytosine. Two hydrogen bonds are involved in the binding between adenine and thymine, while three hydrogen bonds are involved in the binding of guanine to cytosine.

Double-stranded DNA has the unique property that it can make identical copies of itself when supplied with precursors and the required enzymes. In simplified terms, two strands begin to unwind and separate as the hydrogen bonds are broken. This produces single-stranded regions. Complementary deoxyribonucleotide triphosphates then pair with the exposed bases under the control of the enzyme DNA polymerase.

The information in DNA is assembled in structural genes. A structural gene is a linear sequence of codons, which contain the information for a functional protein, consisting of a sequence of amino acids. Individual proteins may function as structural components of the cell, as enzymes, or may regulate important cellular functions. The DNA of eukaryotic cells contains repeated sequences of some genes. Also, eukaryotic genes, unlike prokaryotic (i.e., bacterial) genes, have noncoding DNA regions called introns between coding regions known as exons. This property means that eukaryotic cells have an additional processing mechanism at transcription.

3.6.2. *Transcription (* ⟶ Nucleic Acids – Transcription*)*

The linkage between the DNA in the nucleus and proteins in the cytoplasm is not direct (Fig. 54). The information contained in the DNA molecule is transferred to the protein-synthesizing machinery of the cell via another informational nucleic acid, called messenger RNA (mRNA), which is synthesized complementary to the relevant DNA sequence by RNA polymerase. Although similar to DNA, mRNAs are single-stranded, and contain the base uracil instead of thymine, and ribose instead of deoxyribose. The mRNA molecules act as transport vehicles for the information contained in the genes being expressed.

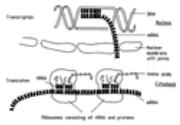

Figure 54. Scheme of the important steps in gene transcription and protein synthesis

In eukaryotic cells, the initial mRNA copy contains homologues of both the intron and exon regions. The intron regions are then removed and the exon regions are spliced together to form the active mRNA molecules, which are then transported through the pores of the nuclear membrane to the cytoplasm.

3.6.3. *Translation (* ⟶ Nucleic Acids – Translation*)*

The next process involves the translation of mRNA molecules into polypeptides. This procedure requires many enzymes and two further types of RNA: transfer RNA (tRNA) and ribosomal RNA (rRNA). There is a specific tRNA for each amino acids. The tRNA molecules are involved in the transport and coupling of amino acids into the resulting polypeptide (Fig. 54). Each tRNA molecule has two binding sites, one for the specific amino acid, the other containing a triplet of bases (the anticodon) which is complementary to the appropriate codon on the mRNA.

The rRNA is complexed with protein to form subcellular globular organelles called ribosomes. Ribosomes can be regarded as the reading heads, which allows the linear array of mRNA codons each to base-pair with an anticodon of an appropriate incoming tRNA amino acid complex.

3.6.4. *Regulation of Gene Expression*

All cells possess the same genetic information, but different types of cells exhibit distinct gene transcription patterns. These differences in gene expression are critical to the morphological and biochemical properties of the many thousands of cell types of the human and animal body. Hence, mechanisms are required that regulate gene expression, that is, determine which genes are expressed and to what extent and which genes are not expressed in a certain cell type at a particular time. The mechanisms involved in regulation of gene transcription are not entirely understood. The transcription of structural genes is regulated by a special set of codons, in particular, promotor sequences, the initial binding sites for RNA polymerase before transcription begins. Different promoter sequences have different affinities for RNA polymerases. Additional regulatory genes called operators regulate the activity of several genes or gene groups (operons). The activity of the operator itself is further controlled by a repressor protein, which stops the transcription of the whole operon by binding to the operator sequence. Due these regulatory mechanisms cells are able to express only the genes required at a given moment for their specialized function. This not only helps to conserve cellular energy, but is also critical for correct cellular differentiation, tissue pattern formation and function, and maintenance of the physiological integrity of the entire organism.

3.6.5. *DNA Repair*

All living cells possess several efficient DNA repair processes. DNA repair is crucial in protecting cells from spontaneous and exogenous lethal and mutating effects such as heat-induced DNA hydrolysis, UV radiation, ionizing radiation, DNA-reactive endogenous chemicals, free radicals, and reactive oxygen species. Among the various DNA repair mechanisms, the most comprehensively studied mechanism in eukaryotes is the excision repair pathway. This mechanism involves a group of enzymes acting cooperatively to recognize DNA lesions, remove them, and correctly replace the damaged sections of DNA.

The excision repair pathway is regarded as error-free and does not lead to the generation of mutations. However, this pathway may become saturated after excessive DNA damage. In this case, the cell may be forced to activate other repair mechanisms which do not operate error-free. Several of these mechanisms, such as error-prone repair, have been well characterized in bacteria, but their counterparts, if any, in mammalian cells have not been identified yet.

3.7. Molecular Mechanisms of Malignant Transformation and Tumor Formation

3.7.1. *Mutations*

Mutations are hereditary changes in genetic information, resulting from spontaneous or xenobiotic-induced DNA damage. The term mutation can be applied to point mutations, which are qualitative changes involving one or a few bases within one gene, and to larger changes involving parts of the chromosome detectable by light microscopy or even whole chromosomes and thus many thousands of genes (Table 22).

Table 22. Types of mutations

Gene mutations	base-pair substitutions, deletions, insertions, gene rearrangements, gene amplifications
Chromosomal mutations	
Structural	breaks, translocations
Numerical	loss or gain of an entire chromosome

Point mutations can occur when one base is substituted for another (base substitution) or when base pairs are deleted or inserted (deletions/insertions). Substitution of another purine for a purine base or of another pyrimidine for pyrimidine is called a transition, while substitutions of purine for pyrimidine or pyrimidine for purine are called transversions. Very small alterations in the chemical structure or the DNA bases may be sufficient for a base-pair substitution to occur. Guanine, for example, normally pairs with cytosine, while O^6-methylguanine (a frequent DNA modification seen with methylating agents such as dimethylnitrosamine) pairs with thymine (Fig. 55).

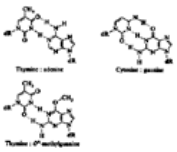

Figure 55. Formation of a base substitution

Guanine normally pairs with cytosine and adenine with thymine (upper part); in contrast, O^6-methylguanine (a frequent DNA modification induced by methylating agents such as dimethylnitrosamine) pairs with thymine (lower part), resulting in a hereditary change of the genetic information

These changes in certain codons may cause insertion of the wrong amino acid into a relevant polypeptide. In this case, the changes are named missense mutations. Such proteins may have dramatically altered properties if the new amino acid is close to the active center of an enzyme or affects the three-dimensional structure of an enzyme or a structural protein. Hence, the alterations may result in marked changes in the differentiation and proliferative characteristics of the affected cells. A base substitution can also result in the formation of a new inappropriate stop (or nonsense) codon. The result of nonsense mutations is the formation of a shorter and, most likely, inactive protein. Owing to the redundancy of the genetic code, about a quarter of all possible base substitutions will not result in amino acid replacement and will be silent mutations.

Bases can be also deleted or added to a gene. As each gene is of a precisely defined length, these changes, if they involve a number of bases that is not a multiple of three, result in a change in the reading frame of the DNA sequence and are known as frameshift mutations. Such mutations often have a dramatic effect on the polypeptide of the affected gene, as most amino acids will differ from the point of insertion or deletion of bases in the DNA strand onwards.

Some forms of unrepaired alkylated bases are lethal due to interference with DNA replication. Others, such as O^6-methylguanine lead to mutations if unrepaired. These differences indicate that not all DNA adducts are of equivalent importance. In fact, some adducts appear not to interfere with normal DNA functions or are rapidly repaired, others are mutagenic, and yet others are lethal. The most vulnerable base is guanine, which can form adducts at several of its atoms (e.g., N-7, C-8, O-6 and exocyclic N-2, see Section Interaction of Reactive Intermediates with Cellular Macromolecules).

Intrastrand and Interstrand Cross-Links. Xenobiotics with bifunctional alkylating properties can also form links between adjacent bases on the same strand (intrastrand cross-links) or between bases on different strands (intrastrand cross-links). The induction of frameshift mutation does not necessarily require formation of covalent adducts. Some compounds that have a planar structure, particularly polycyclic aromatic hydrocarbons, can intercalate between the strands of the DNA double strand. The intercalated molecules may interfere with DNA repair or replication and cause insertions and/or deletions of base pairs. The precise molecular event is still unclear, although several mechanisms have been proposed. Hot spots for frameshift mutations often involve sections of the DNA strand taht contain a run of the same base (e.g., the addition of a guanine to a run of six guanine residues).

DNA strand breaks result from the hydrolysis of the sugar–phosphate bond or a nucleotide. In a double-strand, both a single- and double-strand breaks may occur. DNA strand breaks are often induced by hydroxyl radicals, which are formed at high rates both spontaneously during normal cell life and in the presence of exogenous chemicals. AP lesions (apurinic/pyrimidinic sites) in the DNA strand result from spontaneous hydrolysis of the glucosidic bond and loss of the DNA base (see Fig. 37). Similar to the situation with DNA strand breaks, AP lesions are a common spontaneous event; however, the hydrolysis can be dramatically increased by various types of DNA adducts, such as N^7-substituted purines, a common target of alkylating chemicals (see Section Interaction of Reactive Intermediates with Cellular Macromolecules). Another common spontaneous event is the desamination of cytosine to uracil (Fig. 56); approximately 100 desaminations take place in each cell every day (⟶ Nucleic Acids – Transcription).

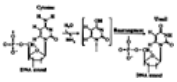

Figure 56. Deamination of cytosine to uracil

dR = Deoxyribose

3.7.2. *Causal Link between Mutation and Cancer (see also, ⟶ Carcinogenic Agents ; ⟶ Mutagenic Agents)*

The change from cells undergoing normal, controlled cell division and differentiation to cells that are transformed, divide without control, and are undifferentiated or abnormally differentiated does not occur in a single step. Malignant transformation is a multistage process. Evidence for the involvement of multiple stages comes from in vitro studies, animal models, and epidemiological observations. In humans, the latent period between exposure to a chemical carcinogen and the appearance of a tumor in the target tissue is approximately 10–25 years. Modern molecular-biology techniques enable thorough investigations of the genome of malignant cells in comparison with the genome of their normal counterparts. These studies clearly show that a single mutation is not sufficient to induce malignant transformation. The number of genetic changes varies between two and seven in different tumor types. Also, several types of mutations are usually formed in a malignant transformed cell (i.e., base-pair substitutions, gene rearrangements, chromosomal breaks, and deletions.

3.7.3. *Proto-Oncogenes and Tumor-Suppressor Genes as Genetic Targets*

Why should mutations be causally linked to cancer? The answer to this question has increasingly become clear since the 1980s with the study of proto-oncogenes and tumor-suppressor genes [144, 145]. It is now appreciated that normal control of cell division and differentiation is based on the interplay of two sets of genes, the proto-oncogenes and the tumor-suppressor genes. Abnormal activation of proto-oncogenes and/or inactivation of tumor-suppressor genes eventually leads to malignant transformation.

Oncogenes were originally discovered in the genome of transforming retroviruses and were therefore named v-oncogenes [144, 146]. Subsequent studies showed that these viral genes were originally derived from the mammalian genome. In the normal cell, these proto-oncogenes have important functions in signal transduction

pathways.

3.7.4. *Genotoxic versus Nongenotoxic Mechanisms of Carcinogenesis*

Oncogene activation and tumor-supressor gene inactivation induced by mutations provide strong evidence for the involvement of genotoxic mechanisms in tumor formation. However, it has been recognized for many years that cancers can arise without direct or indirect interaction between a chemical and cellular DNA, that is, in the absence of direct mutations. The distinction between nongenotoxic and genotoxic carcinogens was more sharply defined following the identification of a comparatively large number of nongenotoxic carcinogens by the U.S. National Toxicology Program [147]. These include a wide range of chemicals acting by a variety of mechanisms, such as disruption of normal hormonal homeostasis in hormone-responsive tissues, and peroxisome proliferation and proliferation of urothelial cells of the urether and urinary bladder following damage by kidney stones.

Genotoxic carcinogens tend to induce tumors in several tissues of both males and females in both rats and mice. In contrast, nongenotoxic carcinogens usually induce tumors only at high doses, in one tissue, in one sex, or only in one species. The experimental evidence available so far does not support the existence of real thresholds for DNA-reactive carcinogens, although very low concentrations may exist for which practically no clinically manifest tumors may be observed in the animal (or human) lifespan. However, these concentrations can not be considered as thresholds. In the case of carcinogens that operate via other biological effects, the carcinogenic activity would parallel dose–response relationships of the relevant biologic effects, a very important aspect for human risk assessment. Treatment regimens or exposure scenarios that do not elicit biological effects would not promote tumor formation. There are, however, two major problems: very few dose–response studies have been performed with nongenotoxic carcinogens, and in most cases the biochemical mechanisms responsible for the tumor promoting action are not understood.

Usually, non-enotoxic carcinogens are divided into two major categories. The first includes compounds that induce cytotoxicity and regenerative cell proliferation, e.g., 2,2,4-trimethylpentane and other branched-chain hydrocarbons in the proximal tubules. The second group of nongenotoxic carcinogens induce cell proliferation in the absence of cytotoxicity (they are directly mitogenic), relevant examples in this group are carcinogenic hormones or peroxisome proliferators such as di(2-ethylhexyl) phthalate and clofibrate. Induction of cell proliferation is involved in both categories of nongenotoxic carcinogens and may contribute to malignant transformation by increasing the number of spontaneous genetic errors, since DNA replication does not occur with 100% fidelity. Furthermore, in rapidly proliferating cells, DNA damage has a higher chance of being converted to heritable mutations. However, in all these cases cell proliferation is the final result of an as-yet unidentified molecular mechanism.

Genotoxic and nongenotoxic mechanisms are not mutually exclusive events. Rather, they cooperate in tumor formation, as can be seen with many genotoxic carcinogens. In most cases genotoxic carcinogens induce tumors only after applications of high doses, concomitantly causing cytotoxicity, cell death, and regenerative proliferation. Hence, a tumor is the final outcome of a complex, multistep interplay between genotoxic and extranuclear events.

3.8. Mechanisms of Chemically Induced Reproductive and Developmental Toxicity

The term reproductive toxicity covers any detrimental effect on the male and female reproductive system due to exposure to toxic chemicals. Developmental toxicity refers to detrimental effects produced by exposure to developing organisms during embryonic, fetal, and neonatal stages of development.The main phases of reproduction are listed below:

Germ cell production

Spermatogenesis (man)
Oogenesis (woman)

Preimplantation phase

Fertilization
Formation of the blastocyst
Implantation

Embryonic phase

Organogenesis (in humans the first 12 weeks, in rat the first 2 weeks of gestation)

Fetal phase

Functional maturation and growth of the organs, in humans week 12 to 30

Peri- and postnatal phase

Last week of gestation, birth and first period after birth

Such effects can be irreversible or reversible. Embryolethal effects are incompatible with survival and result in resorption or spontaneous abortion. Irreversible effects that are compatible with survival may cause structural or functional abnormalities in the offspring, and these are called teratogenic. Embryotoxic chemicals may also cause overall growth retardation or delayed growth of certain organs.

3.8.1. *Embryotoxicity, Teratogenesis, and Transplacental Carcinogenesis*

For an agent to be classified as a developmental toxicant, it must be harmful to the developing organism at exposure levels that do not induce severe toxicity in the mother, such as substantial reduction in weight gain, persistent emesis, or convulsions. Adverse effects on the developing organism under severe maternal toxicity may be secondary to perturbations in the maternal system. For practical purposes, however, the test compounds can be initially administered at maternally toxic doses to determine the threshold level for adverse effects on the offspring. At these exposure levels conclusions can be qualified to indicate that adverse effects of the conceptus were obtained at maternally toxic exposure levels, and may not be indicative for selective developmental toxicity.

The susceptibility of the developing organism to xenobiotic insults varies dramatically within the narrow time span of the major developmental stages (the preimplantation, embryonic, fetal, and perinatal periods), because developing organisms undergo rapid and complex changes within this short period. The major morphogenic events occuring during preimplantation development are formation of a compact mass of cells (the morula) and of the blastocyst. The latter already exhibits a certain degree of cellular differentiation. Considerable similarity exists in the timing of preimplantation development across several mammalian species, regardless of the total length of gestation [148]. At the time of blastocyst formation, cell division and metabolic capacity increase dramatically. During the preimplantation period, biochemical changes under progesterone and estrogen control render the endometrium sensitive to the blastocyst implantation. One important sign of blastocyst implantation is a prostaglandin-dependent increase in endometrial vascular permeability. Alterations in the hormonal milieu or direct excretion of specific xenobiotics into the uterine epithelia during this period can impair implantation and cause embryolethality. Limited data suggest that the preimplantation embryo appears to be susceptible to lethality but rarely to induction of structural aberrations (teratogenicity) with chemical insults. Sublethal exposures of preimplantation embryos have not yet been adequately explored.

Following implantation, organogenesis takes place, which is characterized by the division, migration, and association of cells into primitive organs. The most characteristic susceptibility of the embryo to xenobiotics during the organogenesis period is the induction of structural birth defects (terata). Within the organogenesis period (embryonic period), individual organ systems possess highly specific periods of vulnerability to teratogenic insult (Fig. 57). Administration of a teratogen on day 9 of rat gestation would result in a high level of brain and eye defects, while on day 15 structural abnormalities of the kidney and urinary bladder would predominate. As shown in Figure 57, critical periods for the susceptibility of different organs overlap, so that exposure to teratogens usually results in a spectrum of more or less severe malformations in a number of organ systems [149].

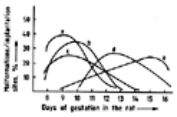

Figure 57. Degree of susceptibility of various organ rudiments of the developing organism to teratogenic xenobiotics (data from [144])

a) Eye; b) Brain; c) Heart and axial skeleton; d) Palate; e) Urogenital system

The critical phase for inducing anomalies in individual organ systems may be as short as one day or may extend throughout organogenesis. Urogenital defects, for example, can result from drug treatment from day 9 to 16 of gestation in the rat. However, the structural defects of the urogenital system depend upon the day(s) of exposure to the teratogenic antibiotic. The development of the urogenital system is multiphasic, and individual stages may have different sensitivities to chemical insult. The mechanisms governing embryonic differentiation are not well understood, but they are certainly involved in intrinsic susceptibility of individual organs to teratogenic insult.

Functional maturation and growth are the major processes occurring after organogenesis, during the fetal and perinatal periods. Insult at these late developmental stages leads to growth retardation or to more specific functional (but not structural) disorders and transplacental carcinogenesis. The fetal and perinatal period of life is highly susceptible to carcinogenesis, due to the high cellular replication rates, presence of xenobiotic biotransforming enzymes in the fetus, and immaturity of the immune system in the developing organism. Several childhood tumors occur so early after birth that prenatal origin is considered likely. These include acute lymphocytic leukemia, Wilms tumor (nephroblastoma), and neuroblastoma. Studies with direct-acting transplacental carcinogens such as ethylnitrosourea indicate that susceptibility to carcinogens begins after completion of the organogenesis period in rodents. Tumors in offspring occurred primarily when ethylnitrosourea was given during the fetal period, whereas birth defects and embryolethality predominated with exposures in the embryonic phase [150, 151]. However, this does not imply that teratogenesis and carcinogenesis are mutually exclusive processes. Teratogenesis and carcinogenesis can be regarded as graded responses of the embryo to injury, with teratogenesis representing the more gross response involving major tissue necrosis in early, relatively undifferentiated embryos, combined carcinogenicity–teratogenicity damage in older embryos, and finally, carcinogenicity alone in the fetus.

3.8.2. *Patterns of Dose–Response in Teratogenesis, Embryotoxicity, and Embryolethality*

The major toxic effects of prenatal exposure observed at the time of birth are embryolethality, malformations, and growth retardation. The relationship between embryolethality, malformations, and growth retardation is quite complex and depends on the type of agent, the time of exposure, and the dose. Some developmental toxins may cause malformations of the entire litter at exposure levels that do not cause embryolethality (Figure 58). If the dose is increased, embryolethality can occur, often in combination with severe maternal toxicity. Malformed fetuses are often more or less retarded in growth, and the curve for growth retardation is often parallel to and slightly displaced to the right from the curve for teratogenicity. Such a pattern of response is indicative of agents with high teratogenic potency.

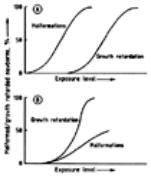

Figure 58. Possible dose – response relationship of teratogens.

A) Teratogens interfering with specific events in differentiation; B) Teratogens acting via general cytotoxicity and induction of cell necrosis

A more common dose–response pattern involves embryolethality, malformations, and growth retardation of surviving fetuses. Exposure to these chemicals results in a combination of resorbed, malformed, growth-retarded, and “normal” fetuses. Depending on the teratogenic potency of the agent, lower doses may cause predominantly malformations. As the dosage increases, however, embryolethality predominates until the entire litter is resorbed. Growth retardation can precede both these outcomes or parallel the teratogenicity curve.

A third dose–response pattern consists of growth retardation and embryolethality without structural abnormalities. Growth retardation of surviving fetuses usually precedes significant embryolethality. Agents producing this pattern of response would be considered embryotoxic but not teratogenic, and are also toxic to the maternal organism. In contrast, potent and specific teratogens often show only weak toxicity to the maternal organism.

The best known example is the hypnotic sedative thalidomide, which was rarely associated with severe undesired effects in adult humans but induced malformation in thousands of children whose mothers had taken the drug as a sleeping aid at the recommended therapeutic doses during gestation. The existence of these three general patterns of response indicate that for some agents embryolethality and teratogenicity are different degrees of manifestations of the same primary insult. For other agents, there is a qualitative difference in response, and the primary insult leads to embryotoxicity and embryolethality alone.

For practical purposes (for exact experimental procedures, see Section Embryotoxicity and Teratogenicity) a relatively small number of pregnant rodents (approximately eight per group) are exposed on days 6 through 15 of gestation to the test agent at doses up to those causing limiting maternal toxicity and/or severe embryotoxicity (death, severe growth retardation). The purpose of this dose-range finding study is to obtain a qualitative yes/no signal about the potential developmental toxicity of the agent, and information on doses causing severe impairment of the maternal organism. For the main evaluation of developmental toxicity the highest dose should cause measurable but slight maternal toxicity (i.e., significant depression of weight gain) or embryotoxicity (i.e., significant depression of fetal body weight, increased embryolethality, and/or structural malformations), and the low dose should cause no observable effects.

[Top of Page]

4. Methods in Toxicology

4.1. Toxicological Studies: General Aspects

The aim of toxicology is the assessment and management of potential hazards from exposure of humans and the general environment including animals and plants to chemicals. To achieve this objective, detailed knowledge on the inherent hazard of a xenobiotic (for definition, see Section Hazard Identification Techniques), that is, its acute and chronic toxicity, its no observed effect level (NOEL), and its teratogenic, mutagenic and carcinogenic effects, is required. This information can not be obtained from a single experiment. A battery of in vivo and in vitro toxicity tests must be utilized. As required by law in most industrialized countries, all toxicity testing must be performed under the rules of good laboratory practice with exact documentation of all relevant conditions and results. Adequate planning of toxicity tests for obtaining optimal information from the experiments may greatly improve the basis for the risk assessment of a chemical and may reduce the number of animals needed and the financial expense associated with toxicity studies. Therefore, all test batteries should be part of an integrated approach to toxicity studies and include not only methods to determine the toxic effects of a chemical and their dose dependence, but also toxicokinetics, biotransformation, and mechanisms of action. This chapter provides an overview of the currently used methods for the assessment of a chemical's toxic profile. For details, specific guidelines on practical aspects of toxicity studies and types of data required can be obtained from web sites of national and international organisations (see Section Information Resources). OECD guidelines for the toxicity testing of chemicals in vivo are listed in Table 23.

Table 23. OECD guidelines on short- and long term toxicity testing in vivo

No. Title	Original adoption	Updated
401 Acute Oral Toxicity	12 May 1981	20 Dec. 2002*
402 Acute Dermal Toxicity	12 May 1981	24 Feb. 1987
403 Acute Inhalation Toxicity	12 May 1981	
404 Acute Dermal Irritation/Corrosion	12 May 1981	24 April 2002
405 Acute Eye Irritation/Corrosion	12 May 1981	24 April 2002
406 Skin Sensitization	12 May 1981	17 July 1992
407 Repeated Dose 28-Day Oral Toxicity Study in Rodents	12 May 1981	27 July 1995
408 Repeated Dose 90-Day Oral Toxicity Study in Rodents	12 May 1981	21 Sept. 1998
409 Repeated Dose 90-Day Oral Toxicity Study in Non-Rodents	12 May 1981	21 Sept. 1998
410 Repeated Dose Dermal Toxicity:28-Day	12 May 1981	
411 Subchronic Dermal Toxicity: 90-Day	12 May 1981	
412 Repeated Dose Inhalation Toxicity: 28/14-Day	12 May 1981	
413 Subchronic Inhalation Toxicity: 90-Day	12 May 1981	
414 Prenatal Developmental Toxicity Study	12 May 1981	22 Jan. 2001
415 One-Generation Reproduction	Toxicity	26 May 1983
416 Two-generation Reproduction Toxicity Study	26 May 1983	22 Jan. 2001

417 Toxicokinetics	4 April 1984	
418 Delayed Neurotoxicity of Organophosphorus Substances Following Acute Exposure	4 April 1984	27 July 1995
419 Delayed Neurotoxicity of Organophosphorus Substances: 28-Day Repeated Dose Study	4 April 1984	27 July 1995
420 Acute Oral Toxicity — Fixed Dose Procedure	17 July 1992	17 Dec. 2001
421 Reproduction/Developmental Toxicity Screening Test	27 July 1995	
422 Combined Repeated Dose Toxicity Study with the Reproduction/Developmental Toxicity Screening Test	22 March 1996	
423 Acute Oral Toxicity–Acute Toxic Class Method	22 March 1996	17 Dec. 2001
424 Neurotoxicity Study in Rodents	21 July 1997	
425 Acute Oral Toxicity: Up-and-Down Procedure	21 Sept. 1998	17 Dec. 2001
426 Developmental Neurotoxicity Study	Draft New Guideline, October 1999	
427 Skin Absorption: In vivo method	Expected, Approved by WNT (May 2002)	
428 Skin absorption: In vitro method	Expected, Approved by WNT (May 2002)	
429 Skin Sensitization: Local Lymph Node Assay	24 April 2002	
430 In Vitro Skin Corrosion: Transcutaneous Electrical Resistance Test (TER)	Expected, Approved by WNT (May 2002)	
431 In Vitro Skin Corrosion: Human Skin Model Test	Expected, Approved by WNT (May 2002)	
432 In Vitro 3T3 NRU phototoxicity test	Expected, Approved by WNT (May 2002)	
433 Acute Inhalation Toxicity: Fixed Dose Procedure	Draft New Guideline, October 1999	
451 Carcinogenicity Studies	12 May 1981	
452 Chronic Toxicity Studies	12 May 1981	
453 Combined Chronic Toxicity/Carcinogenicity Studies	12 May 1981	

*Date of deletion.

For the evaluation of a new chemicals toxic effects in laboratory animals two types of studies are carried out: acute-toxicity and repeated-dosing studies.

Acute Toxicity. Following administration of a single dose of the test substance or of multiple doses given over a period of up to 24 h, potentially adverse effects are usually monitored during the following 14 d. Acute toxicity studies in animals aim to assess the human risk from single exposure to high doses, for example, in industrial accidents, after drug overdoses, or after suicide attempts.

Repeated-Dosing Studies: Subacute, Subchronic, and Chronic Toxicity. The purpose of repeated daily doses of a chemical for part of the animal's life span is to study subchronic and chronic effects. Studies on *subacute toxicity* are carried out for two to four weeks, while studies on *subchronic toxicity* usually last for a period of three months. These studies are helpful in assessing the human risk resulting from frequent exposure to household or workplace chemicals and from intake of chemicals used for therapeutic purposes. Studies to determine *chronic toxic effects* are carried out for at least six months; studies aiming to investigate the carcinogenic effects of a test compound are carried out over the animal's entire lifetime. Lifetime exposure of humans may occur to widespread environmental pollutants, food additives, or residues of agricultural chemicals in food.

In addition to the acute and repeated-dose toxicity studies, the reproductive and developmental toxicity as well as the genotoxicity of a new chemical must be investigated in separate experiments.

Many thousands of new and potentially toxic compounds are synthesized every year. It would be a waste of money, resources and manpower if the entire battery of toxicity tests were automatically performed for every new chemical. Therefore, toxicity testing is rather undertaken on the basis of a decision-point approach in several stages, as shown in Figure 59. At the end of every stage, the decision must be met, if the development will be continued or if, on the basis of the toxicity data available so far, the potential human risk of the exposure to this chemical is unacceptable. If the latter is true, the development and consequently the toxicity testing is stopped.

Figure 59. Evaluation of the toxicity profile of a new chemical compound on the basis of a decision point approach

At every step of toxicity testing, further development may be interrupted if, on the basis of data collected up to that point, the human risk is considered unacceptable.

Animal Husbandry. The use of standardized conditions for the housing of animals plays a major role in the planning, evaluation, and interpretation of toxicity tests. Animals must be kept in a controlled environment, i.e., constant temperature of 22 ± 3 °C, sufficient ventilation, relative humidity between 30 and 70% and a 12 h light/dark cycle. Diet composition and quality of drinking water must also be standardized and controlled throughout the experiment. Only healthy young adult animals should be enrolled in the studies, and the animals should be allowed to acclimatize to the experimental conditions for at least one week prior to first dosing. After the acclimatization period, animals with poor health or body weights varying by more than 20% of the the group's mean body weight are either excluded from the studies or randomized to ensure a homogenous population in the different control and treatment groups. The basic guidelines — choice of species, number of animals, dosing regimens, duration and frequency of observation, assessment of specific body functions — for acute, subchronic and chronic toxicity tests are summarized in Table 24.

Table 24. Basic guidelines for acute, subchronic, and chronic oral toxicity tests

	Acute oral	Subchronic oral	Chronic oral
Animals	rats preferred	rodent and nonrodent species	rodent and nonrodent species
Sex		males and females equally distributed per dose level	
Age	young adult, weight variation within 20 % of mean	rodents, 6 weeks; dogs 4 – 6 months old	
Number of animals	at least 10 (5 per sex)	at least 20 for rodents (10 per sex)	50 per sex group for rodents
Number of treatment groups	3; mortality rates between 10 and 90 % should be produced	3; mortality should not exceed 10 % in high-dose group	3; low dose should reflect expected human exposure and high dose must produce not more than 10 % mortality
Untreated control	not necessary	yes	yes
Vehicle control	yes, if vehicle of unknown toxicity is used	yes	yes
Dosing	gavage; single dose, same dose of vehicle; if necessary	diet, gavage, drinking water	diet, gavage, drinking water

	use divided doses over 24 h		
Duration of study	at least 14-d observation period	90 d	6 – 24 months in rats
Body weight determination	before dosing, weekly thereafter, and at death	weekly and at termination	weekly for first 13 weeks; every 2 weeks thereafter and at termination
Necropsy	all animals	all test animals; organ weights of liver, kidney, heart, lungs, brain, gonads, adrenals, and spleen	
Histopathology	examination of organs showing evidence of gross pathological change	all tissues high-dose and control groups; liver, kidney, heart, lungs, target organs, and any gross lesion in mid- and low-dose groups	all tissues of animals
Frequency of cage-side observations	frequently during day of dosing; once each morning and late afternoon thereafter	daily	daily
Observations, assessments	nature, onset, severity, and duration of any effect observed		
		ophthalmoscopy: pretest and at termination in control and high-dose groups	
		hematology/clinical chemistry:	
		pretest, dosing midpoint, termination	pretest and at 3, 6,12, 18, 24 months
		urine analysis: dosing midpoint, termination	pretest and at 3, 6,12, 18, 24 months

A variety of in vitro methods are in development to reduce the numbers of animals used in toxicity testing. Some of the developed methods have gained regulatory acceptance, and some may be used for a priority-determining process for further testing. All well-evaluated methods focus on local effects such as skin and eye irritation, where most progress in the development of nonanimal methods has been made. Regarding replacement of toxicity studies on systemic effects after repeated exposure by nonanimal methods, due to the complexities of interactions resulting in toxic responses, acceptable nonanimal tests equaling the predictive power of animal testing are unlikely to be available in the near future.

4.2. Acute Toxicity

4.2.1. *Testing for Acute Toxicity by the Oral Route: LD_{50} Test and Fixed-Dose Method*

The objectives of acute toxicity tests are

1. To assess the intrinsic toxicity of the test compound
2. To identify target organs of toxicity affected by the xenobiotic
3. To provide information concerning the dose selection and treatment regimens for repeated-dose studies
4. To provide information for human risk assessment after a single high-dose exposure to the chemical
5. To provide essential data for the classification, labeling, and transportation of the chemical (regulatory view report)

LD_{50} Test. The determination of the mean lethal dose (LD_{50}) is still often considered as the first step in the evaluation of the acute oral or inhalation toxicity of a new chemical; with the present knowledge and recent experiences, the formal determination of the LD_{50} is no longer considered as necessary, and alternative methods that are also to be used for classification and labeling have been developed. However, since the test is still widely used, it will be briefly described before treating the newer methods used in testing of acute toxicity and the reasons that led to these changes.

In the LD_{50} test, groups of animals (usually female rats) are treated with graduated doses of the test compound, and by using mathematical models the dose which causes death in 50% or more of the population is determined. Internationally accepted guidelines recommend the use of at least three dose groups with five males and five females for each dose or the use of three dose groups with five animals of one sex and one dose with five animals of the other sex [152]. The LD_{50} is then determined from data obtained as described in Section Dose–Response: a Fundamental Issue in Toxicology. Chemicals with LD_{50} values ≤ 25 mg/kg are considered very toxic, between 25 and 200 mg/kg as toxic, and between 200 and 2000 mg/kg as harmful. Some potentials and limitations of the LD_{50} test follow:

Potential

1. Useful as a first approximation of hazards in the workplace
2. Basis for the design of subchronic studies
3. Properly conducted test may give useful information on other relevant toxicity parameters
4. Rapid completion

Limitations and Problems

1. Lethality only criterion applied, other toxic effects not considered
2. Animal welfare is major point of concern because large number of animals are required to obtain statistically acceptable values
3. Large variations in LD_{50} in different laboratories with identical chemicals, many influencing factors
4. Species and strain differences cause difficulties in extrapolation
5. No information on chronic toxicity obtained (chronic effects are more important for regulating exposure)

The scientific significance of the LD_{50} test has been repeatedly questioned, not only because the lethal dose is not relevant for human risk assessment but also on the basis of the variability of the test results and last not but least for reasons of animal welfare. Comparative assessment of LD_{50} values in 60 laboratories under controlled conditions resulted for example in considerably different LD_{50} values (by a factor of up to 14). Therefore, numerous alternatives to the LD_{50} test have been proposed for the evaluation of acute toxicity that rely on signs of toxicity rather than on mortality. One of this procedures, the fixed-dose method - has recently gained acceptance by the OECD and the EU

Fixed-Dose Method. The fixed-dose method relies on the observations of clear signs of toxicity developed at one of a series of fixed-dose levels (i.e. 5, 50, 300, and 2000 mg/kg of the chemical oper kilogram body weight). The dose levels at which signs of toxicity but no deaths are detected are used to classify the test compounds according to their toxic potential (Table 25).

Table 25. Classification of toxicity of a xenobiotic with the fixed-dose method

Dose (oral), mg/kg	Results	Classification
5	less than 90 % survival	very toxic
	90 % or more survival but evident toxicity	toxic
	90 % or more survival, no evident toxicity	retest at 50 mg/kg

50	less than 90 % survival	toxic, retest at 5 mg/kg
	90 % or more survival, but evident toxicity	harmful
	90 % or more survival, no evident toxicity	retest at 500 mg/kg
500	less than 90 % survival or evident toxicity and no death	harmful, retest at 50 mg/kg
	no evident toxicity	retest at 2000 mg/kg
2000	less than 90 % survival	harmful
	90 % or more survival, with or without evident toxicity	unclassified, does not represent a significant acute toxic risk if swallowed, no further testing necessary

As can be seen in Table 25, the fixed-dose method allows a classification identical to that previously obtained in the LD_{50} test. Moreover, comparative investigations utilizing both the LD_{50} test and the fixed-dose procedure revealed that in the majority of the test compunds (80–90%), the toxicity class assigned by determining the LD_{50} was identical to that determined by the fixed-dose method [152, 153].

In the fixed-dose method, at least ten animals (five per sex) are used for each dose investigated. The initial dose chosen (5, 50, 300, or 2000 mg/kg body weight) is one that is judged likely to produce evident toxic effects, but no mortality. When such a judgement can not be made due to lack of information on the potential toxic effects of the xenobiotic, an initial “sighting” study should be carried out. If clear signs of toxicity do not occur at the starting dose of 300 mg/kg during the two weeks observation period, the dose is increased to the next level. A careful clinical examination of the animals is performed at least twice on the day of administration and once daily thereafter for the next two weeks. Animals obviously in pain or showing severe signs of distress or toxicity are humanely killed. Cage-side examinations include skin and fur, eyes and mucous membranes, respiratory system, blood pressure, somatomotor activity, and behavior (for procedures, see Section Toxicological Evaluation of the Nervous System). Particular attention is directed to observation of tremors, convulsions, hypersalivation, diarrhoea, and coma as indices of neurotoxicity. Food consumption and weight development are also monitored constantly. At the end of the observation period, all animals in the study are killed and subjected to gross autopsy. Organs showing macroscopic evidence of gross pathology are further subjected to histopathological examination.

The fixed-dose method offers several important advantages as compared with the traditional LD_{50} test:

1. The available evidence suggest that the fixed-dose method produces more consistent results without substantial interlaboratory variation.
2. It provides information on the type, time of onset, duration, and consequences of toxic effects. This information is more relevant for assessing the risks of human exposure to the chemical than the mean lethal doses of the LD_{50} test.
3. It requires fewer animals than the LD_{50} test (roughly 50%) and subjects the animals to less pain and distress.
4. It enables the classification of chemicals according to regulatory requirements.

For the “standard” acute oral and dermal tests the LD_{50} should be determined, except when the substance causes no mortality at the limit dose (usually 2000 mg/kg). Similarly, for an acute inhalation toxicity study the LC_{50} should be determined, unless no mortality is seen at the limit concentration (5 mg per L per 4 h for aerosols and particulates, 20 mg per L per 4 h for gases and vapors). In the fixed-dose procedure, the discriminating dose (the highest of the preset dose levels which can be administered without causing mortality) should be determined. For the acute toxic class and the up-and-down methods the final dose used in the study should be determined following the testing protocol, except when the substance causes no mortality at the limit dose.

Whichever approach is used in determining acute toxicity critical information must be derived from the data to be used in risk assessment. It is important to identify the dose levels at which signs of toxicity are observed, the relationship of the severity thereof with dose, and the level at which toxicity is not observed (i.e. the acute NOAEL). However, note that a NOAEL is not usually determined in acute studies, partly because of the limitations in study design.

4.2.2. *Testing for Acute Skin Toxicity*

Irrespective of whether a substance can become systemically available, it may cause changes at the site of first contact (skin, eye, mucous membrane/gastrointestinal tract, or mucous membrane/respiratory tract). These changes are considered local effects. A distinction can be made between local effects observed after single and after repeated exposure. For local effects after repeated exposure, see Section 3.9. Only local effects after single ocular, dermal, or inhalation exposure are dealt with in this section. Substances causing local effects after single exposure can be further classified as irritant or corrosive substances, depending on the (ir)reversibility of the effects observed.

Irritants are noncorrosive substances which through immediate contact with the tissue can cause inflammation. Corrosive substances are those which can destroy living tissues with which they come into contact.

Knowledge on the dermal toxicity of a new chemical is one of the prerequisites for assessment of the risks associated with human exposure to the chemical, because skin contact may represent a very important route of exposure in the occupational setting and in the home. Testing for dermal toxicity is usually performed in rabbits. Three types of application of the test chemical are employed: nonocclusive, semiocclusive, and occlusive. The test compound is applied uniformly to the back or a band around the trunk (clipped free of hair); approximately 10% of the body surface of the animal should be covered. Solid substances are pulverized and moistened to a paste with physiological saline or another appropiate solvent whose effects have been fully evaluated prior to the skin test. For occlusive or semiocclusive testing, the application site is covered with a plastic sheet (or other impervious material) or with a porous gauze dressing, respectively. For unocclusive exposure, the application site should be as close to the head as possible to prevent ingestion of the chemical by the animal licking the site of application. The duration of exposure varies between 4 and 24 h. If no test-chemical-related toxic effects on the skin or systemic toxicity are observed up after doses of up to 2 g/kg body weight, testing at higher doses is unnecessary. At the end of the exposure period, the compound is removed with cotton wool soaked in an appropiate solvent and the skin irritation is scored according to the Draize scoring system as shown in Table 26. In addition, any adverse systemic effects caused by percutaneous absorption of the test compound are monitored.

Table 26. Evaluation of skin reactions according to the Draize scoring system

Erythema	Score
No erythema	0
Very slight erythema (barely perceptible)	1
Well-defined erythema	2
Moderate to severe erythema	3
Severe erythema (beet redness)	4
No edema	0
Very slight edema (barely perceptible)	1
Slight edema (edges of area well defined by definite raising)	2
Moderate edema (area raised ca. 1 mm)	3
Severe edema (raised more than 1 mm and extended beyond area of exposure)	4

However, substantial differences exist in skin anatomy between humans and experimental animals. In general, the penetration of chemicals through the human skin is similar to that of pig, miniature swine, and squirrel monkey and clearly slower than that of the rat and rabbit. For example, administration of the insecticides lindane and parathion to rabbit skin results in an absorption of 51.2 and 99.5% of the dose, respectively; the corresponding absorption rates for human skin are 9.3 and 9.7%.

Since the 1980s, in vitro studies using human skin samples have been increasingly conducted to estimate percutaneous absorption of chemicals. The following

experimental design is commonly used: A piece of excised human skin is attached to a diffusion apparatus that has a top chamber for the test compound, an O-ring to hold the skin in place, and a bottom chamber to collect samples for analysis. The flow of a chemical across the skin can be calculated with models based on chemical thermodynamics, taking into consideration the octanol/water partition coefficients, the saturated concentration in aqueous solution, and the molecular mass of the test compound [154]. However, for routine applications, the method has not been sufficiently evaluated.

Furthermore, an increasing number of toxicokinetic models for estimating the extent of percutaneous absorption of chemicals has appeared in the literature. Among them, a physiologically based toxicokinetic model was recently developed to describe the percutaneous absorption of volatile and lipid-oluble organic contaminants in dilute aqueous solution [155]. This toxicokinetic model considers both physiological parameters such as volumes of body compartments and blood flow rate, as well as the properties of the test compound. Presently, these models do not play a role in regular toxicity testing and are therefore not discussed in depth here.

More recently extensive progress has been also made in developing in vitro systems for evaluating the dermal irritation potential of chemicals. An overview of the systems that have been evaluated so far for a range of compounds by comparison of their predictive accuracy with animal test results is presented in Table 27 (for a review see [156]).

When evaluating these studies, attention should be given to the occurrence of persisting irritating effects, even those which do not lead to classification. Effects such as erythema, oedema, fissuring, scaling, desquamation, hyperplasia, and opacity which are not reversible within the test period may indicate that a substance will cause persistent damage to the human skin and eye.

Table 27. In vitro test systems for detection of dermal irritation potential

System	End point
Mouse skin organ culture	leakage of LDH* and GOT**
Human epidermal keratinocytes	release of labeled arachidonic acid, cytotoxicity
Cultured BHK21/C13 cells	growth inhibition, cell detachment
SKINTEX – protein mixture	protein coagulation

* LDH = Lactate dehydrogenase.
** GOT = Glutamic acid oxalacetic transaminase.

4.2.3. *Testing for Acute Toxicity by Inhalation*

Studying the toxicity of a chemical by inhalation exposure requires a considerable technological input. Therefore, inhalation exposure is usually not tested if this absorption pathway is not expected to occur because the test chemical is not volatile or the physicochemical properties of solids do not allow the generation of respirable particles. Particles with diameters greater than 100 µm are unlikely to be inhaled, because they settle too rapidly. Particles with diameters of 10–50 µm are likely retained in the nose and the upper parts of the respiratory tract, while particles with diameters of less than 7 µm can reach the alveoli of the human lung. When performing toxicity studies with inhalation exposure, the differences in respiratory physiology between humans and the small laboratory rodents must be considered. In contrast to humans, the rat is an obligate nose breather with a complex nasal turbinate structure which filters many small particles. Therefore, the upper size limit for particles reaching the alveolar region in rats is in the range of 3–4 µm in diameter.

The duration of exposure in acute inhalation studies is usually 4–6 h, and they may be performed either as whole-body or head(nose)-only procedures in specific exposure chambers. A number of important considerations should be taken into account in planning and evaluating inhalation studies. The frequently made assumption that on the basis of their physical form, gases and vapors will be absorbed uniformly throughout the respiratory tract is incorrect. Many gases or vapors, such as ammonia, formaldehyde, and sulfur dioxide have high solubility in water and are rapidly absorbed by the humid epithelial surface of the upper respiratory tract. Therefore, toxic effects observed after inhalation of this type of chemical will generally be confined to these regions, especially to the nasal passages. In contrast, chemicals with low solubility in water such as nitrogen dioxide, phosgene, and ozone will penetrate readily to the low pulmonary regions, even at relatively low concentrations in the respiratory air. In mixed atmospheres containing both a gas or vapor and particulates, the vapor or gas may be absorbed on the particulate fraction, so that the deposition pattern of the vapor or gas is governed by the size of the particulate fraction and not by the water solubility of the vapor or gas.

The inhalation exposure of experimental animals may be performed in a dynamic or static mode. In dynamic systems, the test atmosphere is continuously renewed, ensuring atmospheric stability and constant concentrations of the chemical in the gas phase. This mode of inhalation exposure is more complicated and requires larger amounts of the test chemical and suitable systems to ensure complete mixing of the continuously applied test compounds with the stream of air flowing through the exposure system (Fig. 60).

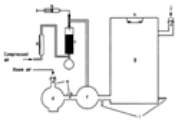

Figure 60. Example of an open inhalation chamber for exposure to volatile liquids

a) Meter; b) Syringe coupled to infuser; c) Heated beads of silica or glass; d) Compressor; e) Metering valves; f) Mixing chamber; g) Exposure chamber; h) Ventilation; i) Sampling valves for determination of atmospheric concentration; j) Exhaust

Static systems are sealed, and the atmosphere is circulated. The concentration of the chemical in the gas phase decreases during exposure due to its uptake and biotransformation by the experimental animals. Static systems are used predominantly in acute toxicity studies and in research laboratories because they are relatively inexpensive and consume only a small amount of the chemical in comparison to the amounts need to generate a dynamic exposure. In analogy to the LD_{50} values obtained in the acute oral studies the mean lethal air concentrations LC_{50} are assessed in acute inhalation studies, usually for an exposure time of 4 h (Fig. 61).

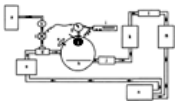

Figure 61. Closed exposure system

a) Oxygen cylinder; b) Metering valve; c) Solenoid valve; d) Mixing chamber; e) Thermometer; f) Oxygen sensor; g) Pressure gauge; h) Exposure chamber; i) Oxygen monitor; j) Injection port; k) Condenser; l) Flow meter; m) Carbon dioxide absorber; n) Gas chromatograph; o) Pump

4.3. Repeated-Dose Toxicity Studies: Subacute, Subchronic and Chronic Studies

Repeated-dose toxicity studies assess the effects resulting from the accumulation of a compound or its toxic effects in the organism and, unlike acute studies, they can also reveal toxic effects that appear after a latency period. A classic example is the delayed neuropathy caused by some organophosphorous insecticides and by cresyl phosphates, which is manifested several weeks after the first administration of the test compound. In contrast to the marked clinical symptoms observed in the course of delayed neuropathy, these compounds hardly cause any acute symptoms immediately after the first administration. Hence, false negative results may be obtained if an assessment were based only acute toxicity testing. The major aims of repeated-dose toxicity studies are the identification of starting points for the extrapolations required in human risk assessment such as NOAEL and benchmark doses, and the identification of critical end points to be carried over into the risk assessment process.

Testing for subacute toxicity is usually performed over a time period of 2–4 weeks at three dose levels as an aid in selecting the dose levels for the subchronic studies. Studies to determine subchronic effects are usually performed in rats and dogs over 10% of the animals life-span (3 months in rats, 12 months in dogs). At the start of the study, rodents should be 6–8 weeks old and dogs 4–6 months old. The animal numbers enrolled are between 10 and20 rats and 6 and 8 dogs per sex per dose group. Ideally, the lowest dose of the chemical should not induce toxic effects, the intermediate dose should induce slight toxicity, and the high dose should induce clear signs of toxicity without causing death in more than 10% of the animals.

In both subchronic and chronic toxicity studies (see below), the test compound is often incorporated into the diet or added to the drinking water. Food consumption varies from weanling to maturity, with younger animals consuming more food on a bodyweight basis. Therefore, it is necessary to predict the changes in body weight and food consumption on a weekly basis and adjust the concentration of the test compound in the diet in order to ensure constant dosing throughout the study. Compounds not stable in diet or water or not accepted by the animals may also be applied by gavage. Application by gavage directly into the stomach ensures constant dosing, but gavage studies require skilled personnel, and gavage-related trauma may reduce survival in all groups. Oral application with the feed is usually

performed on a 7 days per week basis, while a 5 days per week scheme is frequently used for gavage administration, skin application, and inhalation studies.

Studies on the chronic toxicity of chemicals are usually performed for at least six months in rodents and 12 months in dogs, the chronic toxicity studies can be combined with a carcinogenicity study. Dose levels are usually selected on the basis of the results of studies on acute and subacute toxicity. The highest dose applied should be toxic, i.e., suppress body weight up to 10% (maximum tolerated dose, MTD). The two other dose-levels are usually 1/4 and 1/8 of the MTD. Xenobiotics showing no adverse effects in the short-term studies are usually tested at doses which are 100–200 times higher than the expected human exposure.

In both subchronic and chronic studies, cage-side examination and clinical chemistry are performed routinely during the test period. After termination of the study, much emphasis is placed on the histopathological evaluation of treatment-induced adverse effects

Cage-side observations during the study

1. Body weight, food and water consumption
2. Skin and fur, eyes, mucous membranes
3. Respiration and blood circulation
4. Motor activity and behavioral pattern

Clinical chemistry during the study

1. *Blood:* erythrocyte, leukocyte, and differential leukocyte counts; hemoglobin concentration; hematocrit, platelet, and reticulocyte counts; electrolytes; inorganic phosphorus and alkaline phosphatase; glucose, protein, albumin, creatinine, urea, lipids, enzymes
2. *Urine:* volume and coloration/turbidity, osmolality and pH, glucose and protein, urine enzymes and cytology

Toxicologic pathology after termination of the study and in animals dying during the study

1. Organ weights and macroscopic evaluation
2. Histopathological examination of brain, liver, kidney, spleen, testes, and every organ with macroscopic changes

4.4. Ophtalmic Toxicity

The majority of injuries to the eyes by direct contact with a chemical occur with substances which are handled in an uncontrolled manner, e.g,. by children in the home; and this type of injury is easily prevented at the occupational setting by using simple protective procedures. However, many irritant gases and vapors may also produce ophtalmic toxicity, and these toxic effects are of practical importance in occupational medicine.

The conventional in vivo eye irritation test in the rabbit was formalized by Draize some 50 years ago and still remains the only fully validated method to assess ophtalmic toxicity. Increasing criticism primarily based on the discomfort and the deliberate injury caused to the animals led to the development of several in vivo and in vitro alternative methods, which will be described at the end of this chapter.

In the conventional in vivo test the chemical (0.1 ml of liquid or 100 mg of solid chemical) is instilled into one eye of each of the test rabbits, the contralateral eye serving as control. Eyes are then examined periodically (usually after 1, 24, 48, 72 h and 7 dd) and the ocular lesions are scored essentially according to Draize et al.:

Cornea	
A) Opacity	
No opacity	0
Scattered or diffuse area, iris details clearly visible	1
Easily discernible translucent areas, iris details slightly obscured	2
Opalescent areas, iris details not visible, pupil size barely discernible	3
Opaque, iris invisible	4
B) Area of cornea involved	
One-quarter or less, but not zero	1
Greater than one-quarter, but less than half	2
Greater than half, but less than three-quarters	3
Greater than three-quarters, up to whole area	4
Corneal score = (A)×(B)×5 (maximum total score = 80)	
Iris	
A) Normal	0
Folds above normal, congestion, swelling, circumcorneal injection (any or all), iris still reacting to light	1
No reaction to light, hemorrhage, gross destruction	2
Iris score = (A)×5 (maximum total score = 10)	
Conjunctivae	
A) Vessels normal	0
Vessels definitely injected, above normal	1
Diffuse, deep crimson red, individual vessels not readily discernible	2
Diffuse beefy red	3
B) No chemosis (swelling)	0
Any swelling above normal (includes nictitating membrane)	2
Obvious swelling with partial eversion of lids	2
Swelling with lids about half closed	3
Swelling with lids about half to completely closed	4
C) No discharge	0
Any amount of discharge different from normal	1
Discharge with moistening of lids and hairs adjacent to lids	2
Discharge with considerable moistening around eyes	3
Conjunctival score = [(A) + (B) + (C)]×2 (maximum total score = 20)	
Total maximum score (cornea + iris + conjunctiva) = 110	

There is increasing evidence that a volume of 0.01 ml of liquid xenobiotics is as sensitive as the conventionally used 0.1 ml and is probably more appropiate for comparison with human exposure situations. Ocular toxicity testing for exposure to gases, vapors, and aerosols is carried out in appropiate exposure chambers.

A number of in vivo and in vitro alternatives to the conventional eye irritation test have been suggested. The in vivo alternatives aim at reducing discomfort of the animals by employing lower doses of the test material and increasing the sensitivity of the test by using noninvasive objective measurements. Among them, the assessment of corneal thickness and of the intravascular pressure seem to be sensitive parameters to identify mild to moderately irritant chemicals. In spite of the large number of suggested in vitro tests, currently no single in vitro test has proved effective in predicting the eye irritation. Therefore, in vitro tests can not replace the rabbit eye test yet. However, in vitro assays are useful as screens for product development to reduce the number of tests performed in animals later. A number of these tests are presented in [157]

4.5. Sensitization Testing

Chemicals that have the potential to elicit allergic reactions are continously introduced into the human environment. Therefore, allergic reactions of the skin are becoming an increasingly important problem, especially in the workplace. Allergic contact dermatitis is one of the most common occupational diseases and may become debilitating unless the causative agent is identified and exposure stopped. While irritant dermatitis is generally produced by direct interaction of the chemical with skin constituents, allergic dermatitis is the result of a systemic immune reaction which in turn induces effects in the skin. An important characteristic of allergic reactions that must be taken into account when testing for allergenic potential, is that allergic responses usually have a biphasic course. The induction period between initial contact with the causative agent and the development of skin sensitivity may be as short as two days for strong sensitizers such as poison ivy extract, or may require several years for a weak sensitizer such as chromate; for most of the chemical compounds with allergenic potential the induction period usually takes from 10 to 21 d. After this initial development of sensitivity to a certain allergenic chemical, the time between reexposure to this agent and the occurrence of clinical allergic symptoms is generally between 12 and 48 h; in animal testing, this period is called the challenge phase.

The general objectives are to determine whether there are indications from human experience of skin allergy or respiratory hypersensitivity following exposure to the agent and whether the agent has skin sensitization potential based on tests in animals. There are two methods currently described in EU Annex V and OECD guidelines for skin sensitization in animals: the guinea pig maximisation test (GPMT) and the Buehler test. The GPMT is an adjuvant-type test in which the allergic state (sensitization) is potentiated by the use of Freund's Complete Adjuvant (FCA). The Buehler test is a non-adjuvant method involving topical application for the induction phase rather than the intradermal injections used in the GPMT (Table 28; for reviews see references [158, 159]. Although they differ by route and frequency of treatment, they all utilize the guinea pig as test species. In general, for the induction phase the chemical is administered to the shaved skin intradermally, epicutaneously, or by both routes several times over a period of two to four weeks. Freund's complete adjuvant (FCA, a mixture of heat-killed *Mycobacterium tuberculosis*, paraffin oil, and mannide monooleate) is often included to increase the immunological response. During the challenge phase, a nonirritating concentration of the chemical is applied. The concentration of the test chemical and the application route (epi- or intradermal) are often different between the two phases. Sensitization is assessed by examining the skin reactions (edema, erythema) following the challenge phase and comparing them with any skin reactions observed immediately after the induction phase; the latter reactions are considered to result from direct irritating (toxic) properties of the test chemical. Hence, the difference between the symptoms observed after the induction and after the challenge phase is attributed to the allergenic effects of the chemical.

Table 28. Guinea pig sensitization tests

Test	Induction: route/number of applications	Challenge: route/number of applications
Draize	intradermal/10	intradermal/1
Open epicutaneous	epidermal open/20	epidermal open/1
Buehler	epidermal occlusive/3	epidermal occlusive/1
FCA*	intradermal in FCA*/3	epidermal open/1
Split adjuvants	epidermal occlusive/4 + FCA* intradermal/1	epidermal occlusive/1
Optimization	intradermal + FCA*/10	intradermal/1 epidermal occlusive/1
Maximization	intradermal + FCA*/1 epidermal occlusive/1	epidermal occlusive/1

* FCA = Freund's complete adjuvant.

The guinea pig maximization test is the most widely used and is considered to be very sensitive. The first part of the induction phase includes simultaneous injection of FCA alone, the test compound in saline, and the test compound in FCA into three different areas in close proximity to each other. The second part of the induction phase 7 d later employs epicutaneous application of the chemical on a filter paper, which is occluded and left in place for 48 h. The challenge phase is conducted epicutaneously for 24 h, two weeks after the induction phase. The maximization test is very sensitive and may produce false positive results. The original procedure (injection of the test compound) does not allow testing of final product formulations. Therefore, a modified procedure has been developed. In the first week, the FCA is injected four times and the test product formulation is administered epidermally, as in the second induction week.

Both the GPMT and the Buehler test have demonstrated the ability to detect chemicals with moderate to strong sensitization potential, as well as those with relatively weak sensitization potential. These guinea pig methods provide information on skin responses, which are evaluated for each animal after several applications of the substance, and on the percentage of animals sensitized.

The murine local lymph node assay (LLNA) is another accepted method for measuring skin sensitization potential. It has been validated internationally and has been shown to have clear animal welfare and scientific advantages compared with guinea pig tests. In June 2001, the OECD recommended that the LLNA should be adopted as a stand-alone test as an addition to the existing guinea pig test methods.

Respiratory hypersensitivity is a term that is used to describe asthma and other related respiratory conditions, irrespective of the mechanism by which they are caused. When directly considering human data in this document, the clinical diagnostic terms asthma, rhinitis, and alveolitis have been retained.

There are currently no internationally recognised test methods to predict the ability of chemicals to cause respiratory hypersensitivity. Potentially useful test methods based on allergic mechanisms are the subject of research and development. However, there are currently no test methods under development which are designed specifically to identify chemicals that cause respiratory hypersensitivity by nonimmunological mechanisms.

4.6. Phototoxicity and Photosensitization Testing

The biologically active spectrum of light can be divided into UV (220–400 nm) and visible light (400–760 nm). The UV spectrum is further divided into UVA (315–400 nm), UVB (280–315 nm), and UVC (220–280 nm); the last-named is absorbed in the stratosphere and does not reach the surface of the earth. The primary source of toxic effects on the skin is UVB, although UVA may also play a critical role in some reactions.

Xenobiotics localized within the skin may be activated by UVB and induce phototoxicity and/or photosensitization (photoallergy). Photoallergy is similar, both mechanistically and clinically, to allergic contact dermatitis, the only difference being that the chemical must react with light to become allergenic. Photoallergic reactions are not necessarily dose-dependent and show great variability between individuals. In analogy, phototoxicity may be compared with irritant dermatitis. Many phototoxic reactions may be caused by the formation of free radicals followed by lipid peroxidation and localized inflammation. In addition to phototoxicity and photoallergy, light-induced activation of chemicals may cause depigmentation, induction of an endogenous photosensitizer or of a disease characterized by photosensitization such as lupus erythematodes or pellagra. However, these reactions are comparatively rare. The tests to identify photoallergenic and phototoxic chemicals are performed in analogy to the tests for skin sensitization and irritation, respectively.

Photoallergenic potential is evaluated by repeated application of the test compound on the skin of guinea pigs and exposure of the treated area with UV light after every application; the UV treatment should cause a very slight erythema. Several days after this induction phase, the challenge phase is conducted by treatment with a low dose of the test compound together with UV light. Phototoxic reactions can usually by observed after the first exposure to the test compound together with UV light; in addition to guinea pigs, mice and rabbits are also utilized in tests for phototoxicity.

4.7. Reproductive and Developmental Toxicity Tests

The general objectives of reproductive and developmental toxicity testing are to establish whether exposure to the chemical may be associated with adverse effects on reproductive function or capacity, and whether administration of the substance to males and/or females prior to conception and during pregnancy and lactation causes adverse effects on reproductive function or capacity. Another focus of these studies are induction of nonheritable adverse effects in the progeny and whether the pregnant female is potentially more susceptible to general toxicity.

Reproductive and developmental toxicity is a very broad term including any adverse effect on any of the following aspects:

1. Male or female sexual structure and function (fertility)
2. Development of the new organism through the period of major organ formation, organogenesis (embryotoxicity and teratogenicity)
3. Development of the new organism during the peri- and postnatal periods

The field of reproductive toxicology has become increasingly important with the recognition that viral and bacterial infections and xenobiotic chemicals can produce

severe and irreversible defects in the offspring. The first reports on malformations due to rubella virus infections, ionizing radiation, hormones, dietary deficiencies, and chemicals appeared in the 1930s and 1940s, but the potential impact of reproductive toxicity on public health was only recognized more than two decades later. In 1960, a large increase in newborns with specific limb malformations, which are rarely seen otherwise, was recorded in Germany and in other parts of the world. One year later, the sedative/hypnotic drug thalidomide was recognized as the causative agent. The thalidomide epidemic resulted in 10 000 malformed children and subsided after the drug was withdrawn from the market at the end of 1961. One important consequence of the thalidomide disaster was the introduction of requirements for the testing of potential new drugs for reproductive toxicity. All test batteries required include detailed tests for reproductive toxicity to prevent a repetition of the thalidomide disaster.

The first specific reproductive toxicity test to be conducted is usually the two-generation study. which should be initiated after the rat 90-d subchronic repeated-exposure study, since the results obtained may provide information necessary for selecting dose levels for the two-generation study. Additionally, repeated-exposure studies that can provide information relevant to reproductive toxicity should be used in the design of the two-generation study. For example, the observation of neurological effects may indicate the need to evaluate developmental neurotoxicity.

The first developmental toxicity study is normally performed after completion of the two-generation study. The design of the developmental toxicity study should use all information derived from the repeated-exposure and two-generation studies, in particular dose–response relationships and information on maternal toxicity. The preferred species for the two-generation study is the rat; the necessity of a developmental toxicity study in the rabbit is dependent on the outcome of the first study.

4.7.1. *Fertility and General Reproductive Performance*

Segment I experiments are usually conducted in rats (20 animals of each sex per dose) with three doses of the test chemical, most often administered with diet. The treatment must not cause general systemic toxicity in the parental organism; therefore, in dose selection the dose levels are chosen according to observations in studies of subacute and subchronic toxicity, which are usually performed before testing for reproductive toxicity. Young adult male rats are treated for 60–80 d prior to mating to cover a whole period of spermatogenesis. Female rats are pretreated for 14 d to cover three estrous cycles. Treatment of both sexes is continued during the mating period and that of females throughout pregnancy. Half of the females are sacrificed just before term, and the numbers of resorbed and dead fetuses as well as structural abnormalities in the developed fetuses are assessed. In the United States, pregnancy is interrupted midterm in half of the females. Treatment of the remaining females is continued through parturition and lactation until weaning of the newborns, usually 21 d after birth. The young animals (F_1 generation) are reared without receiving the test compound until sexual maturity, when their fertility is assessed. During the rearing period, the development of the young animals is monitored with cage-side carefully clinical observations. If an adverse effect on fertility, pregnancy, or development of the offspring is observed, it is necessary to evaluate whether the effect is due to toxicity to the male or female reproductive system or both. This information can be obtained by separate mating of treated males with untreated females and vice versa.

Furthermore, evaluation of toxic effects on the male and female reproductive systems with specific test systems may be required. The effect of the test compound on male reproductive performance may be evaluated by monitoring mating behavior (e.g., frequency of copulation). Structural and functional impairment of the male reproductive organs is assessed by conducting gross pathology and histology of the testes and sperm analysis (viability, motility, and morphology). Histological examination of the ovaries plays an important role in assessment of toxic effects on the female reproductive system.

4.7.2. *Embryotoxicity and Teratogenicity*

Segment II studies assess adverse effects during the period of organogenesis. Xenobiotics interfering with the developing organism during this extremely sensitive period may cause severe and irreversible structural malformations. These studies are carried out in two species, usually rats (20 per dose) and rabbits (10 per dose); in most cases, two dose levels and an untreated control group are included. Pregnant animals are tested during the period of organogenesis: days 6 to 15 for rats and 6 to 18 for rabbits. The fetuses are delivered by cesarean section one day prior to the estimated time of delivery: day 21 for rats and day 31 for rabbits. The main reason for avoiding natural delivery is to prevent loss of deformed or dead fetuses by cannibalism, which happens in rodents and rabbits. The uterus of the maternal animal is excised, weighed, and examined for implantation sites and resorbed fetuses. The pups are weighed, and one-half of each litter is usually examined for skeletal defects and the remaining one-half for soft-tissue defects.

4.7.3. *Peri- and Postnatal Toxicity*

For segment III studies, treatment of pregnant rats with three dose levels (10–12 animals per dose) begins on day 16 of gestation and is carried on through delivery and lactation, until weaning of the offsprings, normally on day 21 postpartum. Treatment during parturition and lactation is also performed in segment I studies; however, segment III studies may offer additional information since higher doses can be used than in segment I. The peri- and postnatal segment evaluates effects on birth weight and survival as well as development of the offspring in the postnatal period. However, extrapolation of results of segment III studies from the rat to the human situation should be performed with care and consideration of the specific circumstances in each species. In contrast to the human situation, where in addition to the mother the social environment takes care of the newborns, young rats depend completely on the functional integrity of the maternal organism. Furthermore, the organism of newborn rats is significantly less mature by the time of delivery than that of newborn humans. Hence, treatments that affect the maternal organism simply by causing sedation or fatigue may significantly impair the development of newborn rats in the first 21 days of life. Such effects should not be automatically interpreted as relevant for postnatal toxicity in humans.

4.7.4. *Multigeneration Studies*

Multigeneration studies assess the cumulative effects of continuous application of the test chemical on reproduction and development during two or three generations. This application mode is relevant for long-term exposure to chemicals in the environment, such as pesticide residues in food or contamination of drinking water with agricultural chemicals or nonbiodegradable solvents.

The two-generation study is a general test which allows evaluation of the effects of the test substance on the complete reproductive cycle including libido, fertility, development of the conceptus, parturition, postnatal effects in both dams (lactation) and offspring, and the reproductive capacity of the offspring. The two-generation study is preferable to the one-generation study because the latter has some limitation regarding assessment of post-weaning development, maturation, and reproductive capacity of the offspring. Thus, some adverse effects such as oestrogenic- or antiandrogenic-mediated alterations in testicular development may not be detected. The two-generation study provides a more extensive evaluation of the effects on reproduction because the exposure regime covers the entire reproductive cycle, permitting an evaluation of the reproductive capabilities of offspring that have been exposed from conception to sexual maturity. The prenatal developmental toxicity study only provides a focused evaluation of the potential effects on prenatal development.

Three dose levels are usually given to groups of 25 female and 25 male rats shortly after weaning at days 30 to 40 of age. In the multigeneration study, these rats are referred to as the F_0 generation. The F_0 generation is treated throughout breeding, which occurs at about 140 d of age, and the female animals also during pregnancy and lactation. Hence, the offspring (F_1 generation) has been exposed to the test compound in utero, via the maternal milk, and thereafter in the diet. In many protocols, the F_1 generation is standardized to include certain numbers of animals, e.g., eight animals per litter. In analogy to the F_0 generation, the F_1 generation is bred at about 140 d of age to produce the F_2 generation. In some of the parents (F_0 and F_1 generations), gross necropsy and histopathology is conducted with greatest emphasis on the reproductive organs. In addition, necropsy and histopathology are carried out in all animals dying during the study.

The percentage of F_0 and F_1 females that become pregnant, the number of pregnancies carried to full term, the litter size, and number of resorptions, stillborns, and live births are recorded. Viability counts and pup weights are recorded at birth, and at days 4, 7, 14, 21 and 28 of age. With these data, the following parameters are calculated for assessment of the long-term reproductive toxicity of the test compound:

$$\text{Fertility index}\,(\%) = \frac{\text{Number of pregnancies}}{\text{Number of matings}} \times 100$$

$$\text{Gestation index}\,(\%) = \frac{\text{Number of litters}}{\text{Number of bred females}} \times 100$$

$$\text{Birth index}\,(\%) = \frac{\text{Number of pregnancies resulting in live off spring}}{\text{Number of pregnancies}} \times 100$$

$$\text{Viability index}\,(\%) = \frac{\text{Number of animals alive at day 4 after birth}}{\text{Number of new borns}} \times 100$$

$$\text{Lactation index}\,(\%) = \frac{\text{Number of animals alive at day 28 after birth}}{\text{Number of animals alive at day 4 after birth}} \times 100$$

4.7.5. *The Role of Maternal Toxicity in Teratogenesis*

If an agent with selective developmental toxicity is administered throughout the organogenesis period (days 6 to 15 in the rat), identifying the most sensitive target organs becomes difficult. In addition, teratogenic effects may be masked by embryolethality with repeated dosing during the organogenesis period. Developmental toxicity in the form of increased resorption and decreased fetal body weight is generally accepted to occur at maternally toxic dose levels. The role of maternal toxicity in causing congenital malformations, however, is not clear. Doses causing maternal toxicity, as indicated by reduced maternal body weight, clinical signs of toxicity, or death, commonly cause reduction in fetal body weight, increased resorption, and rarely, fetal deaths. Three patterns of association between maternal toxicity and malformations can be observed: (1) for some compounds, maternal toxicity is not associated with malformations; (2) for others, maternal toxicity is associated with a diverse pattern of malformations, often including cleft palate; and (3) the maternal toxicity of still others is associated with a characteristic pattern of malformations.

Compounds in the second category are the most difficult to classify in terms of teratogenic potential. Cleft palate is the principal malformation resulting from food and water deprivation during pregnancy in mice; however, cleft palate is also a malformation specifically induced in mice by a number of teratogens, most notably the glucocorticoids, without apparent maternal toxicity. Complete determinations of food and water consumption, maternal body weights, and impairment of the maternal organism are necessary to distinguish between cleft palate caused by the teratogenic effect of a chemical on the embryo and that resulting from systemic maternal toxicity, which secondarily affects embryonic development. The association of maternal toxicity with major malformations, such as exencephaly and open eyes, is not generally accepted, although most investigations agree that maternal toxicity can cause minor structural abnormalities such as variants in the ribs.

4.7.6. *In Vitro Tests for Developmental Toxicity*

Models to elucidate the mechanism of embryogenesis have been under development for several decades, and therefore developmental toxicology is a fields in which alternative methods to animal experimentation are available. However, because of the complicated, multistep nature of the development of a new life, none of the in vitro systems presently available can replace the animal tests. In vitro tests rather serve for screening purposes, i.e., to preclude the extensive traditional whole-animal test protocol for compounds with marked toxicity on reproduction and development. The existing alternative test systems fall into six groups: lower organisms, cell-culture systems, organ-culture systems, whole-embryo cultures, embryos, and others (Table 29). Since none of the in vitro methods is sufficiently validated for a set of compounds for which the effects on humans or animals are known and the field is much too extensive to be comprehensively reviewed here, the reader is referred to two comprehensive reviews of this field [160].

Table 29. In vitro test systems for developmental toxicity

Group	Test system/organisms	End points monitored
Lower organisms and small animals	sea urchins	growth
	drosophila	
	trout, medaka (fish species)	
	plania	
	brine shrimp	
	animal virus	
Cell culture	pregnant mouse and chick lens epithelial cell	protein synthesis
	avian neural crest	cell differentiation
	neuroblastoma	cell differentiation
Organ culture	frog limb	regeneration
	mouse embryo limb bud	morphological and biochemical differentiation, toxicity
	metanephric kidney organ culture from day 11 mouse embryos	morphological and biochemical differentiation
Whole embryo cultures	chick embryo	embryotoxicity, malformations
	frog embryo teratogenesis assay	lethality, no observed-effect-level, development stage, attained growth, motility, pigmentation,
		gross anatomical malformations
	rat embryo culture (postimplantation embryo)	viability, growth and macromolecular content, gross structural and histological abnormalities

4.8. Bioassays to Determine the Carcinogenicity of Chemicals in Rodents

Despite the many available short-term in vivo and in vitro tests to determine the genotoxic and carcinogenic potential of chemicals and the vast amount of literature on the subject, the lifelong carcinogenicity bioassay remains the main instrument for reliable evaluation of the carcinogenic properties of a xenobiotic. The principal guidelines for performing bioassays were established some 25 a ago by the U.S. National Cancer Institute and have essentially been adopted with slight alterations by all regulatory authorities (Table 30).

Table 30. Basic procedures of rodent carcinogenicity bioassays

Species	rat (Fischer 344, Sprague – Dawley, Wistar)
	mouse (B6C3F1, CD)
Age of animals at the beginning	4 to 6 weeks (shortly after weaning)
Number of animals	50 per sex per dose for carcinogenicity
	10 – 20 for additional studies during the course of experiment
Dose	at least three doses and vehicle control
	maximum tolerated dose
	intermediate dose
	nontoxic dose
Duration	24 months
Application	gavage, in feed, drinking water, inhalation (only if absolutely necessary)
Toxicologic pathology	all animals: gross necropsy
	weight of all important organs
	histopathology of all tissues (ca. 40) and all tumors and preneoplastic lesions by two independent pathologists

A number of factors may interfere with the analysis and interpretation of data from animal carcinogenicity studies. A variety of statistical techniques has been developed to adjust for confounding factors and to estimate confidence intervals and significance of results. Significance tests are used to assess neoplastic response in treated groups as compared to control groups or historical controls (cancer incidence in the identical strain and species observed in control groups for other cancer bioassays under identical housing conditions in the same facility). To estimate the absolute cancer risk posed by a specific chemical, background or spontaneous cancer incidences (induction of neoplasms not related to the administration of the test chemical) must be well defined. In general, high background incidences of cancer such as liver cancer observed in specific strains of mice requires larger number of animals in the treatment groups to detect increases in cancer incidence induced by the administration of the test chemical and to obtain statistically significant results. The demonstration of a dose–response curve for the cancer incidence in groups of animals treated with different doses of the carcinogen will increase the confidence in positive results of an animal cancer bioassay. The same holds for identical results observed with cancer as an endpoint in an independent study.

However, despite the importance of animal cancer bioassays for characterizing chemical carcinogens, this approach has been criticized recently (see also Section Interpretation of Data from Chronic Animal Bioassays). Due to the influence of rodent carcinogenicity assay on the development of new chemicals and pharmaceutical drugs, the pros and cons of this type of assay are discussed in depth in the following.

Practically all chemicals identified as human carcinogens produce tumors in the rodent bioassay. Hence, the test has a very good predictive value, and every chemical exerting carcinogenicity in rodents should be handled as a potential human carcinogen. Considering the many new chemicals developed each year, two major disadvantages of the life-long assay are its high cost and long duration. A two-year gavage study in only one species amounts to approximately € 10^6 and takes 3–4 years or longer for complete evaluation. For optimized evaluation of carcinogenic properties, it is important to use animal species that are closest to humans with regard to biotransformation and toxicokinetics of the test compounds. However, for practical and financial reasons, only rodents can be used. Long-term carcinogenic bioassays in dogs or primates, for example, require seven to ten years for completion and are much more costly.

The application of the maximum tolerated dose (MTD) in rodent bioassays has been the subject of much controversy [161-167], but the experimental design limitations of in vivo studies make the application of high doses necessary. For example, if a specific dose of a chemical causes a 0.5% increase in human cancer incidence, this would result in several hundred thousands of additional cancer cases in a country such as Germany each year and would thus definitely pose an unacceptable risk. However, the identification of this 0.5% increase in cancer incidence with statistical confidence in the rodent bioassay would require a minimum of 1000 animals, provided the incidence of spontaneous tumors is zero. Therefore, there seems to be general agreement that the use of the MTD, although not an optimum solution, is necessary for risk assessment. According to the U.S. National Cancer Institute, the MTD is defined as “the highest dose that can be predicted not to alter the animals' normal longevity from effects other than carcinogenicity”. In practical terms “MTD is the dose which, in the subchronic three-months toxicity study causes not more than a 10% weight decrement as compared to the control groups and does not produce mortality, clinical signs of toxicity, or pathological lesions other than those which may be related to a neoplastic response that would be predicted to shorten an animal's life span”. As stated above, the MTD is determined in the preliminary three-month studies on subchronic toxicity, where it fulfills the above requirements. Ideally, this is exactly what should also happen in the 24-month carcinogenicity study. However, due to cumulation of toxic effects and/or alterations in toxicokinetics of the xenobiotics during the study, for example, induction of toxification or detoxification pathways by application of the xenobiotic in high doses, the MTD dose group often shows reduced survival rates in the life-long bioassay. This may invalidate the study, that is, make it inadequate for evaluation of the carcinogenic potential. Indeed, this is not a rare event in carcinogenicity studies. The opposite effect may also occur: due to toxicokinetic differences between the three- and 24-month studies, the MTD chosen may turn out too low in the long-term bioassay. In spite of all these problems and because of the absence of a satisfactory alternative solution, the use of the MTD is currently the only method to compensate for the fact that in relation to the human population exposed to potential carcinogens, the numbers of rodents used in the carcinogenicity bioassay are extremely low. The legitimate argument against the MTD is that any chemical given at a sufficiently high dose level will induce adverse effects. This understanding, which is beyond dispute in toxicology, has been tentatively generalized by several scientists in recent years by the notion that carcinogenic effects obtained at the MTD may exclusively result from target-organ toxicity, and the increased cell proliferation may contribute to tumor formation by increasing the rate of spontaneous mutations, since DNA replication does not take place with 100% fidelity. Furthermore, during increased cell turnover, the time available to repair DNA damage is reduced, so that an increased number of damaged DNA sites may be converted to heritable mutations. Although this may be the mechanism underlying the carcinogenic effects of some nongenotoxic chemicals, it can not be generalized to every tumor observed at the MTD. Toxicity and cell proliferation do not necessarily result in tumor formation. Table 31 summarizes the important differences between rodent bioassays and human exposure to carcinogens.

Table 31. Some important differences between carcinogenicity tests in rodents and human exposure to potential carcinogens

Rodent carcinogenicity test	Human exposure
High doses	(usually) low doses
Continuous exposure	(often) infrequent or not regular exposure
Single compound, no interactions	simultaneous exposure to several carcinogenic chemicals, interactions probable
Homogeneous population	heterogeneous population

In addition to these problems, which are inherent in the bioassay procedure, the evaluation of the toxicological pathology has repeatedly become an issue of debate, since differences in evaluations between pathologists are frequent. This does not necessarily indicate incompetence of one of the pathologists. The different evaluations may be the result of difference in terminology. Also, sometimes evaluations are conducted years apart, and in the interim period understanding of the pathogenesis of lesions may have changed. Thus, even the same pathologist may not come to the same conclusion when reevaluating tissue slices several years after the first examination.

Due to the uncertainties of rodent bioassays and the extremely high costs and personnel requirements, a multitude of short-term tests has been developed in recent years. These tests aim to predict the carcinogenic potential of chemicals. Most of these in vitro tests are based on damage to the genetic material (genotoxicity) by the chemical or its metabolites. Genotoxicity is without doubt the field in toxicology with the best established and validated in vivo and in vitro short-term tests.

4.9. *In Vitro* and *In Vivo* Short-term Tests for Genotoxicity

Genetic toxicology a comparatively new field of research that has rapidly grown since the 1960s, deals with mutagenicity and genotoxicity.

Mutagenicity is the induction of permanent transmissible changes in the genetic material of cells or organisms. Changes may involve a single gene or gene segment, a block of genes, or whole chromosomes. Effects on whole chromosomes may be structural and/or numerical.

Genotoxicity is a broader term and refers to potentially harmful effects on genetic material which may not be associated with mutagenicity. Thus, tests for genotoxicity include systems which give an indication of damage to DNA (no direct evidence of mutation). End points determined here are unscheduled DNA synthesis (UDS), sister-chromatid exchange (SCE), DNA strand breaks, formation of DNA adducts and mitotic recombination.

Evidence has increasingly accumulated that many carcinogens are also mutagenic, and a large number of short-term in vitro and in vivo tests were developed as predictive tools. Most of these tests are well validated and aim to assess genotoxic properties of chemicals. Today, the majority of potential carcinogens are first identified as mutagens or chromosome-damaging agents in short-term tests and subsequently as carcinogens in the rodent carcinogenicity bioassay. The numerous in vivo and in vitro assays can be categorized into two major groups:

1. Short-term tests detecting gene mutations
2. Short-term tests detecting structural and/or numerical chromosomal aberrations

A number of these in vitro test procedures have gained regulatory acceptance for toxicology testing (Table 32).

Table 32. OECD guidelines on genetic toxicology testing and guidance on the selection and application of assays

No.	Title	Original adoption	Updated
471	Bacterial Reverse Mutation Test	26 May 1983	21 July 1997
472	Genetic Toxicology: *Escherichia coli*, Reverse Assay	26 May 1983	21 July 1997*
473	In Vitro Mammalian Chromosome Aberration Test	26 May 1983	21 July 1997
474	Mammalian Erythrocyte Micronucleus Test	26 May 1983	21 July 1997
475	Mammalian Bone Marrow Chromosome Aberration Test	4 April 1984	21 July 1997
476	In Vitro Mammalian Cell Gene Mutation Test	4 April 1984	21 July 1997
477	Genetic Toxicology: Sex-Linked Recessive Lethal Test in *Drosophilia melanogaster*	4 April 1984	
478	Genetic Toxicology: Rodent dominant Lethal Test	4 April 1984	
479	Genetic Toxicology: In Vitro Sister Chromatid Exchange assay in Mammalian Cells	23 Oct. 1986	
480	Genetic Toxicology: *Saccharomyces cerevisiae*, Gene Mutation Assay	23 Oct. 1986	

481	Genetic Toxicology: *Saccharomyces cerevisiae*, Mitotic Recombination Assay	23 Oct. 1986	
482	Genetic Toxicology: DNA Damage and Repair, Unscheduled DNA Synthesis in Mammalian Cells In Vitro	23 Oct. 1986	
483	Mammalian Spermatogonial Chromosome Aberration Test	23 Oct. 1986	21 July 1997
484	Genetic Toxicology: Mouse Spot Test	23 Oct. 1986	

*Date of deletion (method merged with TG 471).

In vitro tests to detect gene mutations can be categorized into two groups: microbial and mammalian cell assays. An important step in the history of modern genetic toxicology was the development of genetically precisely defined strains of bacteria carrying mutations in particular genes coding for enzymes involved in the biosynthesis of amino acids. Among the numerous tests, the most widely used and best validated is the assay in *Salmonella typhimurium* developed by AMES et al.

4.9.1. *Microbial Tests for Mutagenicity*

4.9.1.1. *The Ames Test for Bacterial Mutagenicity*

The most common method to detect mutations in microorganisms is selecting for reversions in strains that have a specific nutritional (i.e., amino acid) requirement differing from wild-type members of the species; the tester strains are “auxotroph” for this particular nutrient. The *Salmonella typhimurium* mutant strains developed by AMES can not synthesize histidine, because each strain carries one of a number of mutations in in the operon (group of genes) coding for histidine biosynthesis [168, 169]. The result of this mutation is that the tester strains can not grow and form colonies in histidine-free medium. The mutation may revert to the wild-type sequence or a functionally equal sequence either spontaneously (a rare event) or by exposure of the tester strains to genotoxic compounds. The revertant colonies are, like the wild-type bacteria, capable of synthesizing histidine and form colonies in histidine-free medium. For the common tester strains, the DNA sequence at the site of the original mutation in the relevant histidine gene has been determined. According to the type of mutation leading to the inability to synthesize histidine, the strains can be categorized into two groups: base-substitution and the frameshift strains. The difference between these two categories can be illustrated with the following sentence, in which each letter represents a DNA base, each word a triplet coding for an amino acid, and the whole sentence a gene coding for a protein (i.e. enzyme of histidine biosynthesis). The correct sentence represents the gene in the wild-type strain.

THE NUN SAW OUR CAT EAT THE RAT	original sentence (wild-type)
Base-pair substitution:	
THE **S**UN SAW OUR CAT EAT THE RAT	missense mutation coding for a wrong amino acid
THE N**S**N SAW OUR CAT EAT THE RAT	nonsense mutation resulting in interruption of gene transcription
Frameshift mutation:	
THE NUN SA**S** WOU RCA TEA TTH ERA T + 1-frameshift mutation	

This example illustrates that a base substitution can be reverted by another base substitution and in analogy a frameshift mutation can be reverted by another frameshift mutation. Hence, the Ames test not only provides information on the genotoxic potential of the test compound but also on the nature of the DNA damage.

Genetic Makeup of the Salmonella typhimurium Strains Used in the Ames Test. *Histidine mutations of the tester strains*. The base-pair substitution strains can be categorized into two families.

1. The *Salmonella typhimurium* strains TA100 and TA1535 carry the sequence CCC (for leucine) instead of the wild-type sequence CTC (for proline). These missense mutations may be efficiently reverted by mutagens with alkylating properties.
2. The second group of base-pair substitution strains carry a nonsense mutation (TAA instead of CAA). These strains (TA2638, TA100) may detect mutations induced by radicals or oxidizing agents such as hydrogen peroxide and reactive oxygen matabolites.

The commonly used frameshift strains TA98 and TA1538 carry a +1 frameshift mutation near a GCGCGCGC sequence. This strain may be used to detect frameshift mutagens such as polycyclic aromatic hydroxycarbons, certain aromatic amines, and certain aromatic nitro compounds.

In addition to the mutation in one of the genes of histidine biosynthesis, the Ames strains carry additional genetic alterations that increase their sensitivity to detect mutagens (an overview of the genetic make-up of the *Salmonella typhimurium* tester strains is provided in Table 33).

Table 33. Genetic makeup of commonly used Salmonella typhimurium tester strains

Strain	Histidine mutation		Additional genetic alteration
	Location	Route	
TA100	his G46	CCC instead of CTC	*uvrB*
		base substitution results in proline instead of leucine	*rfa*
			*pkM*101
TA1535	his G46	CCC instead of CTC	*uvrB*
		base substitution results in proline instead of leucine	*rfa*
TA2638	his G428	ATT nonsense mutation	*rfa*
	his G8476	results in interruption of transcription	*pkM*101
TA102	his G428	multiple copies of plasmid pAQ1 carry the revertible nonsense	*rfa*
		mutation ATT	*pkM*101
TA98	his D3052	-1-frameshift mutation with a GCGCGCGC sequence	*uvrB*
			rfa
			*pkM*101
TA1538	his D3052	-1-frameshift mutation with a GCGCGCGC sequence	*uvrB*
			rfa

rfa Mutations. The *rfa* mutation results in a defective lipopolysaccharide membrane and thus increases the permeability of the cell wall to bulky hydrophobic chemicals. In addition, strains with defective cell walls are not pathogenic to experimental animals and humans.

uvrB Deletion. Wild-type bacteria possess several effective DNA repair systems, that operate practically error free and can repair DNA damage without allowing mutations to occur. To overcome the problem of DNA repair and thus the decreased sensitivity of the test system, AMES constructed a series of strains with a deletion of the *uvrB* gene, which codes for a subunit of an important enzyme (the cor-endonuclease-I) of the error-free excision repair system. This change increases the sensitivity of the tester strains to mutagens by several orders of magnitude.

The plasmid pkm101. Wild-type *Salmonella typhimurium* strains do not process an “error-prone” DNA repair which is found for example in *Escherichia coli* and some other members of the Enterobacteriaceae. In contrast to the above-mentioned error-free excision repair pathway, the error-prone system operates with low fidelity and introduces new mutations into the genome while repairing a damaged DNA site. To overcome this deficiency and to increase sensitivity to mutagens, the gene for the error-prone repair system has been introduced into some of the *Salmonella typhimurium* tester strains with the plasmid pkm101. This plasmid also carries the genetic information for ampicillin resistance, an important property for monitoring the physiological integrity of the tester strains in the course of the experiments.

The plasmid pAQ1. The *Salmonella typhimurium* strain TA 102 carries the revertible histidine mutation on multiple copies (approximately 30) of the plasmid pAQ 1 and not on the chromosome. Reversion of one of these copies returns the bacteria's capability to synthesize histidine. This increase in sensitivity is partly offset

because strain TA 102 has an intact excision repair system (it does not have the *uvrB* deletion). This strain has been constructed because DNA interstrand cross-linking agents such as the quinone mitomycin C or the combination of psoralens und UV light require an intact excision repair system to generate mutations. DNA interstrand cross-links must first be removed from one strand along with a small number of adjacent bases by excision repair. The gap left behind is repaired, and the remaining broken cross-link attached to the other strand is a premutagenic lesion that may give rise to a mutation by error-prone repair.

The Problem of Bioactivation in the Ames Test. The majority of the mutagenic xenobiotics require bioactivation (toxification) to reactive electrophiles to induce DNA damage. The main disadvantage of bacterial mutagenicity assays is that the tester strains do not express many of the enzymes that bioactivate xenobiotics in mammals. For example, cytochrome P450 activity is not detectable with most substrates in *Salmonella typhimurium*. Therefore, there is a need to simulate the biotransformation occurring in the intact animal by supplementing the test system with the enzymes of bioactivation and the necessary cofactors. For this purpose, numerous exogenous metabolic systems have been used in the last decades. They can be grouped into cell-free and cell-based systems, and among the cell-free systems the 9000 *g* supernatant (called S-9 fraction, the S-9-fraction with the cofactors necessary for enzyme activity is called S-9 mix) from rat liver is the most widely used and best validated.

Preparation of S-9 Fraction and S-9 Mix. S-9 fraction is usually prepared from the liver of male adult rats. Preparations from uninduced animals may contain only low activities of important enzymes of bioactivation such as cytochrome P450 1A1. Since these deficiencies could limit the use of the S-9 fraction as exogenous activating systems in bacterial assays, the rats are usually pretreated with enzyme inducers. Among them Aroclor 1254, a mixture of polychlorinated biphenyls, is the most widely used. Pretreatment of the rats with Aroclor 1254 results in induction of a broad range of cytochrome P450 enzymes. For specific purposes, more selective inducers of cytochrome P450 enzymes such as phenobarbital (cytochrome P450 2A1, P450 2B1) and 3-methylcholanthrene (cytochrome P450 1A1) may be also used.

The livers are removed from the animals after sacrifice at specific time intervals after application of the inducer, minced, homogenized, and centrifuged at 9000 *g* for 15 min. The monooxygenases contained in this S-9 supernatant require NADPH as cofactor, which is normally generated by glucose-6-phosphate dehydrogenase from glucose-6-phosphate by reducing $NADP^+$. Therefore, the S-9-fraction is also supplemented with these two cofactors and with magnesium and potassium salts to yield the final activating system, the S-9 mix. This standard S-9 mix is capable of performing phase-I biotransformation reactions (e.g. oxidations, reductions) but is deficient in most phase-II systems (conjugation reactions). The latter are often involved in detoxification reactions, while phase-I enzymes usually result in toxification of xenobiotics to more electrophilic metabolites. Hence, the S-9-mix may efficiently simulate bioactivation of xenobiotics in the liver, but not detoxification. This may be regarded as an advantage, since it usually increases the sensitivity of the system. On the other hand, the discrepancy between phase-I and phase-II enzymatic activities may be a source of false positive results because the genotoxicity observed in vitro may not reflect the in vivo situation, and detoxification may predominate in the intact animal. The deficiency in conjugation reactions may be partly overcome by adding appropiate cofactors for phase-II conjugation reactions, such asglutathione. The S-9 fraction from other organs may also be prepared and used for the bioactivation of organ specific carcinogens, for example, S-9-fraction of the renal cortex for compounds inducing renal cell tumors. Alternatively, cell-based systems may be used to bioactivate xenobiotics in bacterial mutagenicity assays. Freshly isolated hepatocytes or hepatocytes in primary culture retain the activity of the phase-I and phase-II enzymes of the intact liver and do not require the addition of cofactors for enzyme activity. However, besides greater technical difficulty in obtaining hepatocytes of good quality compared with S-9 fraction, additional problems arise when using intact cells because the electrophiles formed may have a very short half-life and may be trapped and react with macromolecules before diffusing out of the hepatocytes, resulting in false negative results.

An recent elegant approach to overcoming the deficiencies of the bacteria in biotransformation reactions is to clone genes of mammalian biotransformation enzymes into plasmids and to introduce these into tester strains. Although this approach has not been widely used so far, it seems to be a promising method for the future. For example, introducing the gene for *N,O*-acetyl transferase, an enzyme important in the bioactivation of aromatic amines, into *Salmonella typhimurium* TA 98 and TA 100 resulted in an approximately 100-fold increase in the corresponding enzymatic activity in the newly engineered strain. In contrast to the original strains, the *N,O*-acetyl transferase-proficient strains permitted efficient detection of the mutagenic activity of several nitroarenes and aromatic amines. A similar increase in the mutagenicity of bromo- and chloroalkanes was abtained with tester strains expressing a specific human glutathione *S*-transferase isoenzyme [170, 171]. Specific potentials and limitations of these tests follow:

•Potentials

- High reliability when testing known carcinogens and noncarcinogens
- Extensive data base available
- High sensitivity
- Information about the mechanism of mutations may be obtained by comparing results in different strains
- Rapid and inexpensive
- May be used as a bioassay to detect mutagenic components in complex mixtures

•Limitations

- Some important carcinogens are not active due to specific mechanisms of action (metals, particles, asbestos)
- Data are only qualitative
- Sample must be sterile
- Problems in testing bactericidal chemicals

Experimental Procedure for the Ames Test. Presently three different protocols are widely used for the Ames test: the plate-incorporation assay, the preincubation assay and the fluctuation test.

Plate Incorporation Assay. The basic procedure of the plate incorporation assay is illustrated in Figure 62. Briefly, 2.0 mL aliquots of soft agar overlay medium (Top Agar, 0.6% agar and 0.5% sodium chloride in distilled water) containing a trace of histidine and excess biotin and maintained in the liquid state at 45 °C, 100 µL of the tester strain, 20–100 µL of the test compound and, when necessary, 500 mL S9-mix (or another bioactivation system) are added. After mixing, the solution is poured onto dried Vogel–Bonner minimal medium plates. The plates carrying untreated, solvent, and positive (UV light or established mutagen) controls are incubated for 2 d at 37°C. At least ten concentrations of the test compound are usually tested with two plates per concentration.

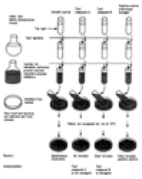

Figure 62. Basic procedure for the Ames test/plate incorporation assay

50 to 200 plates are usually prepared and evaluated per test. At least ten concentrations of the test compound are examined with two plates per concentration level

Preincubation Assay. Some mutagens, particularly those metabolized to short-lived reactive electrophiles, may not be detected in the standard plate incorporation assay due to reaction of the electrophile with constituents of the medium. This type of xenobiotic may be detected more efficiently by using the preincubation assay protocol, in which the bacteria are preincubated with the test compound and S-9 mix in suspension for 30–120 min in the absence of top agar. After the end of this preincubation period, 2.0 mL of soft agar is added to each tube and the reaction mixture is poured into Vogel–Bonner plates, which are incubated as described above.

Fluctuation Test. In the fluctuation test, the number of mutants in a series of small independent replicate cultures is detected. Overnight bacterial cultures are incubated with the test compound and the appropriate metabolizing system (when necessary) in the presence of a trace of histidine. As in the other protocols, this trace of histidine allows a few replication cycles, which are necessary for expression of the mutations following the initial DNA damage. The above mixture is then divided into a large number (usually 50–100) of test tubes or microtiter plates. When the histidine trace is consumed, only revertant cells can grow. Test tubes or wells of the microtiter plates containing revertants become turbid, and the media turn acidic as a result of acid release during growth. This pH drop to 5.2–6.8 can be demonstrated by the color change of a pH indicator such as bromoethyl blue.

Scoring for Colonies in the Plate Incorporation and Preincubation Assay The limited growth of nonrevertant colonies due to the trace of histidine added to enable mutational expression of DNA damage results in a slight background lawn of growth. Therefore, before scoring the plates for revertant colonies, the background lawn should be examined either macroscopically or under the low magnification of a light microscope. At toxic concentrations of the test chemical, the plates appear clear. Also, at toxic concentrations of the test chemical, tiny colonies (minicolonies) may be formed. This happens when, at bactericidal concentrations of the test chemical, very few bacteria survive. The histidine trace usually added to the system may then be sufficient for formation of these minicolonies (pseudorevertants). Revertant colonies resulting from gene mutations are clearly larger and can be counted either by hand or with an automatic colony counter.

Evaluation of Results. A typical mutagenic dose–response curve is illustrated in Figure 63. From the linear part of the dose–response curve, the mutagenic potency of the test compound can be calculated and is usually expressed in number of revertants per nanomole of test compound. At higher, toxic concentrations of the test chemical, the curve may flatten. At concentrations which are very toxic to the bacteria, the number of revertant colonies may start to decline after reaching a maximum. The significance of the results is analyzed with usual statistical methods.

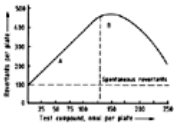

Figure 63. Typical dose – response curve in the Ames test

A) At nontoxic mutagen concentrations the number of revertants increases in a dose-dependent manner; B) Higher, toxic concentrations inhibit growth, which is reflected in flattening of the curve followed by decreasing numbers of obtained revertants

Role of the Ames Test in Evaluating Carcinogenic Properties.. The Ames test is a system to detect mutagenicity which has gained great practical importance as a short-term predictive test for potential carcinogenicity of a xenobiotic. Hence, the Ames test may predict the carcinogenic properties of genotoxic carcinogens, but not of nongenotoxic carcinogens. The very high correlation between mutagenicity and carcinogenicity found in previous reports (values of 90–95% were repeatedly estimated) are certainly not correct. One reason for this overestimation may be that in previous decades mainly structurally alerting chemicals (xenobiotics carrying functional groups present in other carcinogens and known or suspected to be metabolized to electrophiles) were tested for long-term rodent carcinogenicity, and most of these structurally alerting carcinogens are mutagenic. In contrast, more recently, more important environmental pollutants and occupational hazardous chemicals have increasingly been tested for carcinogenicity, independent on their chemical structure. As a result, an increasing number of carcinogenic compounds that do not react directly with DNA to cause mutations have been found. These xenobiotics are thought to act via additional and (unfortunately) only partly defined so,called epigenetic mechanisms (i.e., not involving genetic changes). Nongenotoxic carcinogens cannot be detected in the Ames test. In contrast to the genotoxic carcinogens, which are usually active in two species and/or at multiple sites, nongenotoxic carcinogens are often active in one species and at a single site, for example, branched chain hydrocarbons in the male rat kidney or peroxisome proliferators in the mouse liver. This is confirmed by comparing the carcinogenicity results with the bacterial mutagenicity. While more than 70% of the two-species/multiple-site carcinogens are positive in the Ames test, only 40% of the one-species/single-site carcinogens exert bacterial mutagenicity. Therefore, a reasonable evaluation of the results obtained in bacterial mutagenicity systems can only be conducted by taking into consideration additional information on the chemistry, toxicokinetics, biotransformation, and biological effects of the test compounds.

Presently, the Ames test is an important part of a screening battery for possible carcinogenic properties in the course of the development of new pharmaceutical drugs and chemicals. Depending upon the intended use of the chemical, the development may be interrupted if a compound is clearly positive in the Ames test and in an additional short-term test, often the in vivo micronucleus test (see Section *In Vivo* Cytogenetic Assays). With this “decision-point approach”, the number of unnecessary life-long bioassays in rodents is significantly reduced and large amounts of money and time may be saved.

4.9.1.2. *Mutagenicity Tests in Escherichia coli*

The *E. coli* tester strain WP2 and additional tester strains subsequently developed originate from the wild-type *E. coli* B strain; all tester strains are tryptophan(trp)-auxotrophic. For example, the *E. coli* WP2 trp E group has a terminating mutation at an AT base pair of the trp E gene which codes for anthranilate synthetase, an enzyme of the tryptophan biosynthetic pathway. Hence, the *E. coli* mutagenicity assays, like the Ames test, detect reversion of the deficiency to synthetize tryptophan. The *E. coli* tester strains have umuDC$^+$ genes coding for the error-prone repair system, and their membrane structure permit the passage of many large molecules. The experimental procedures and the inherent advantages and limitations are practically identical to those of the *Salmonella typhimurium* test. However, the *E. coli* test systems are less well validated than the Ames test, and this renders the selection of appropiate strains for the investigation of the mutagenicity of a specific xenobiotic and the interpretation of the obtained results more difficult.

4.9.1.3. *Fungal Mutagenicity Tests*

The most widely used species is *Saccharomyces cerevisiae*; the test detects reversion of the isoleucine auxotrophy induced by exposure to mutagens, in analogy to the Ames test.

4.9.2. *Eukaryotic Tests for Mutagenicity*

Although prokaryotic systems are fast, inexpensive and versatile, there is a requirement for mammalian test systems due to the important differences between prokaryotic and eukaryotic organisms in structure and function of the genome and in biotransformation and transport of xenobiotics. The eukaryotic systems for detecting mutations include tests in the fruit fly *Drosophila melanogaster*, in many different types of cultured mammalian cells, and in the intact animal in vivo.

4.9.2.1. *Mutation Tests in Drosophila melanogaster*

The fruit fly *Drosophila melanogaster* has a short generation time of 10 d and a similar cellular and chromosomal structure and function to mammalian cells. In addition, the genome of the fly is well characterized, and *Drosophila* species can perform many of the phase-I and phase-II reactions occuring in mammals. The *Drosophila* sex-linked recessive lethal test was a popular screening system in the 1970s and was used to screen many hundreds of compounds. However, since then the popularity of the test has waned drastically because of its poor performance in international collaborative trials that investigated the utility of mutagenicity assays as predictive tools. Another *Drosophila* test, devised in the 1980s, the somatic mutation and recombination test (SMART) seems promising but has not been sufficiently validated. Since both tests are not widely used for screening, the experimental procedures are not be described here (the interested reader may find them in reference [156]). Some potentials and limitations of the mutation tests in *Drosophila melanogaster* follow:

•Potentials

- Large number of test organisms can be raised
- Metabolic activation endogenous
- Genotoxic effects evaluated in germ cells
- Large data base on mutagens

•Limitations

- Some known carcinogens such as polycyclic hydrocarbons respond poorly
- Problems with toxic compounds (insecticides)

4.9.2.2. *In Vitro Mutagenicity Tests in Mammalian Cells*

Many short-time mutagenicity tests in mammalian cell lines have been developed. These systems have the advantage of determining genotoxic effects in relevant cell types and may permit the study of mechanisms of carcinogenicity. The potentials and limitations of these in vitro tests follow.

•Potentials

- Full range of mutational responses may be observed
- Low costs when compared to in vivo assays
- Respond to particulates and other types of compound not detected in the Ames test
- More relevant end point for mammals

•Limitations

- Most cell lines used are transformed cells and do not represent “normal” cells
- Only specific loci from the entire genome are monitored
- Most cell lines used are deficient in biotransformation enzymes

In contrast to popular bacterial assays that detect reversion mutations, the commonly used mammalian mutagenicity assays are based on the detection of forward mutations. A large number of cells is treated with the test compound. After a certain period of time, the cells are exposed to a toxic agent that is lethal to all cells not carrying mutations (i.e., only mutated cells can survive). Cultured mammalian cells are normally diploid and have two copies of each gene. Hence, recessive mutations may be missed if a normal copy is present on the homologous chromosome because the probability that a mutagenic compound alters genes is very low. Therefore, mutations are assessed in genes on the X chromosome in male cells, where only one copy of the gene is present. Alternatively, heterozygous cells are used, in which one copy of the gene is already inactivated.

Mammalian cells have genes that allow the cell to salvage nucleotides from the surrounding medium. These genes, although not essential for cell survival, save cellular energy, since the cell need not synthesize these molecules from simple precursors by energy-consuming pathways. If the medium is supplied with altered

nucleotides, the normal (not mutated) cell will incorporate them into the DNA, and this will result in cell death. However, if this salvage property is lost due to mutation, the mutated cells are not able to incorporate the toxic nucleotides from the surrounding medium and will survive and form colonies which can be detected as a parameter of genotoxicity. Although a gene may be inactivated by a mutation, the mRNA and the corresponding enzyme produced prior to the mutational event may be present for some time after exposure to the mutagenic test compound. Therefore, the cells have to be left for some time before challenging with the toxic nucleotide, this period is called expression time.

Basically, two genes of the salvage pathway are utilized in mutagenicity assays: the hypoxanthine–guanine phosphoribosyltransferase (HGPRT) gene and the thymidine kinase (TK) gene. The HGPRT gene is located on the X chromosome in humans and in the Chinese hamster, from which many useful cell lines have been developed. The genetic changes detected in the HGPRT gene as the target site are mainly point mutations because chromosomal deletions usually extend to flanking regions of the target gene, which may contain essential genes. Since only one copy of these essential genes exists in male cells, this results in cell death so that the cells in which the HGPRT gene is inactivated by deletion are usually lost for the test. This is not the case with the TK gene, because it is autosomal and heterozygous cells (TK^+/TK^-) are used in the assay. Here, both point mutations and changes at the chromosomal level can be detected. The HGPRT gene is the target gene in the Chinese hamster CHO and V 79 cell lines, while mouse lymphoma L 5178 Y $TK^+/^-$ cells detect mutations at the TK locus. The mechanistic background of these tests is shown in Figure 64.

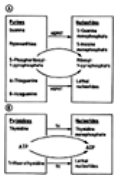

Figure 64. Assessment of mutagenicity in mammalian cells

A) The V 79/HGPRT (hypoxanthine – guanine phosphoribosyltransferase) assay; B) The L 5178 Y $TK^+/^-$ (thymidine kinase) mouse lymphoma assay

Mutant cells do not form the lethal nucleotides and survive in the presence of 6-thioguanine or trifluorothymidine.

The use of Chinese hamster cell lines for mutagenicity screening is limited due to low sensitivity. Chinese hamster cell lines grow on monolayers and, owing to metabolic cooperation, only a relatively small cell population can be used. In contrast, L5178Y $TK^-/^-$ cells grow in suspension, and the system is not impaired by the problems of metabolic cooperation, because intercellular bridges do not occur. As described above, a wide variety of genetic events including gene mutations, recombinations and mitotic nondisjunction can result in the formation of the $TK^+/^-$ genotype from the heterozygote $TK^+/^-$. Two protocols have been devised for mutation assays with mouse lymphoma L5178Y cells:the plate test in soft agar and the fluctuation test in suspension. For dose-finding, a preliminary cytotoxicity assay is usually conducted when testing new compounds. In this assay, the cloning efficiency (for the plate test) or the relative suspension growth (for the fluctuation test) is determined. The highest test concentration for the mutation assay is usually the concentration that reduces cloning efficiency or suspension growth to approximately 10–20% of the control values. In addition, a moderate concentration causing survival reduction to 20–70% and a low concentration (> 70% survival) are also used. After treatment with the test compound (3–7 h; in the presence or absence of an exogenous metabolic system, e.g., S-9 mix) the cells are incubated for 2 d for mutation expression before exposure to trifluorothymidine for 10–12 d. In the presence of trifluorothymidine, only mutant cells can survive. Thus, the relative growth of the cells at the end of the experiment can be used as a parameter for mutagenicity.

4.9.3. *In Vivo Mammalian Mutation Tests*

In contrast to the well-established and widely used micronucleus tests in rodents (see Section *In Vivo* Cytogenetic Assays), the existing in vivo tests to detect gene mutations are not used for routine genetic toxicology testing; therefore, only brief descriptions are given here (for exact experimental procedure, see [156]).

4.9.3.1. *Mouse Somatic Spot Test*

The mouse somatic spot test detects the inactivation of genes at a set of heterozygous loci controlling hair pigmentation. Inactivation can be induced by a wide spectrum of genetic alterations ranging from gene mutations to major chromosomal changes.

4.9.3.2. *Mouse Specific Locus Test*

The mouse specific locus test consists of treating parental mice homozygote for a marker gene controlling coat pigmentation, intensity, and pattern or size of the external ear. Treated mice are mated with a tester stock that is homozygous recessive at the marker loci. The resulting F_1 generation is normally heterozygous at the marker genes and therefore expresses the wild-type phenotype. In the case of mutations at any of these genes, the F_1 offspring express the recessive phenotype. The mouse specific locus test has a comparatively low sensitivity, thus requiring that many thousands of newborns be scored. However, it remains an important method for investigating heritable mutations.

4.9.3.3. *Dominant Lethal Test*

The dominant lethal test assesses embryonic death resulting from genetic changes in parental germ cells. The procedure is based on observing the viability of uterine implantations. Implantations that die at an early stage form a deciduoma or mole. The genetic event resulting in early death is predominantly chromosomal damage in the parental germ cells resulting in a dominant lethal mutation. Although the assay was popular in the 1960s and 1970s as a screen for germ cell damage, it is not widely used at present because it is considered relatively insensitive.

4.9.4. *Test Systems Providing Indirect Evidence for DNA Damage*

4.9.4.1. *Unscheduled DNA Synthesis (UDS) Assays*

Two major DNA repair mechanisms have been identified in mammalian cells. The first mechanism involves direct reversal of DNA damage, such as cleavage of pyrimidine dimers that have been induced by UV light or removal of potentially mutagenic methyl groups from the O^6 position of guanine by O^6-methylguanine-DNA methyltransferases. The second mechanism recognizes damaged DNA bases, removes them after nicking the DNA backbone, fills the gap formed with the correct bases using the opposite strand as a template, and finally seals the second strand break.

The most widely used method to study DNA repair is the “unscheduled” DNA synthesis assay. The principle of the UDS assay consists in detection of the incorporation of radioactive DNA bases or chemical analogues of nucleotides such as bromodeoxyuridine into DNA in the course of excision repair after damaged bases have been removed. This unscheduled DNA synthesis induced by DNA damage must be distinguished from the semiconservative DNA synthesis during DNA replication. DNA replication can occur in the same cells that undergo repair or, alternatively, in other cells of the test population and usually contributes to an incorporation of radiolabeled base or nucleotide analogue in concentrations several orders of magnitude higher than repair. Three approaches are currently available to overcome this problem and to detect selectively DNA repair-dependent incorporation.

1. Use of quiescent cell populations, that do not replicate
2. Suppression of DNA replication by incubation in serum-reduced and arginine-depleted medium and pretreatment with the ribonucleotide reductase inhibitor hydroxyurea
3. Evaluation of radioactivity by autoradiography of single cells, which allows the distinction between heavily labeled replicating cells and lightly labeled DNA-repairing cells

The UDS assay can be conducted in vitro and in vivo. The following description briefly outlines the in vitro assay in cultured mammalian cells. After inhibition of DNA replication by serum reduction, arginine depletion, and treatment with hydroxyurea (if replicating cell lines are used), the cells are exposed to the test compound in presence of radiolabeled thymidine. Among the various cell types used, primary cultures of rat hepatocytes are often preferred to permanent cell lines, because their biotransformation capabilities are closer to those of liver cells in vivo. In addition, primary cultures of hepatocytes do not divide, so that no suppression of cell proliferation is necessary.

Radioactivity incorporation can be measured by scintillation counting of the isolated DNA or by autoradiography of whole cells. For determination of UDS in vivo, the animals are treated with the test compound and radiolabeled thymidine (most often applied by implanted minipumps). Radioactivity in DNA is determined after isolation of the DNA. Alternatively, a combined in vivo/in vitro approach may be used. In this case, in vivo treatment of the animals is followed by isolation of cells from the appropiate organs and labeling of the cells with the radioactive nucleotide in vitro. Assessment of radioactivity incorporation after in vivo treatment is usually conducted by autoradiography.

Several factors influence the interpretation of UDS data. Detection of excision repair demonstrates that the test compound has caused repairable damage to DNA, but does not answer the question whether this has any significance for the mutagenic potential of the chemical. This is very important, because measurements of UDS are averages over all cells of the study population, whereas mutation is a rare event in an individual cell. In addition, extranuclear events such as mitochondrial toxicity resulting in oxidative stress and increased production of reactive oxygen species or increased calcium concentrations in the nucleus and activation of

endonucleases may also influence indirect DNA damage and induction of UDS.

4.9.4.2. *Sister-Chromatid Exchange Test*

Sister- chromatid exchanges are reciprocal exchanges between sister chromatids thought to occur at homologous loci. Although detection of sister-chromatid exchanges can also be conducted after in vivo treatment, the test is widely applied in established cell lines in vitro (Chinese hamster cell lines, human fibroblast cell lines) and in freshly isolated human lymphocytes. Any cell type that is replicating or can be stimulated to divide can be used. The test is carried out essentially according to the following basic procedure. Cells in exponential growth are usually exposed simultaneously to 5-bromo-2′-deoxyuridine and the test compound for a period equivalent to approximately two cell cycles (incubation time depends on the cell type used). In the final 1–2 h of incubation, a spindle inhibitor (colchicine or demecolcine) is added to arrest the cells in metaphase. For detection of sister-chromatid exchanges, the two chromatids are differentially stained with a fluorescent dye such as Hoechst 33258 plus Giemsa. With this method, sister-chromatid exchanges appear as color change on one chromatid (Figure 65).

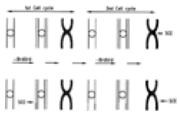

Figure 65. Sister-chromatid exchange (SCE)

The exchange is visualized by treating cells with 5-bromo-2′-deoxyuridine (BrdUrd) during two DNA replication cycles. The distribution of BrdUrd between chromatids is then determined by staining with a fluorescent dye and UV irradiation.

The sister-chromatid exchange test is one of the quickset, easiest, and most sensitive methods for detecting genetic damage. A large number of compounds have been evaluated with this method, which is usually part of the test battery for genotoxicity screening.

The problem is that the precise genetic event detected by this test is basically not known. sister-chromatid exchanges are obviously not related to chromosomal aberrations since they often fail to detect potent clastogens. They relate more to gene mutations, although there are examples of compounds that clearly induce sister-chromatid exchanges in the absence of mutations as well as the converse. Despite the uncertainty resulting from the unknown mechanism, the sister-chromatid exchange test is very useful because it detects important groups of potentially carcinogenic compounds such as alkylating agents and nucleotide analogues, as well as compounds inducing DNA single-strand breaks or acting through DNA binding.

4.9.5. *Tests for Chromosome Aberrations (Cytogenetic Assays)*

4.9.5.1. *Cytogenetic Damage and its Consequences*

In addition to gene mutations, most tumor cells investigated thus far exhibit structural and/or numerical chromosomal aberrations. Chromosomal aberrations such as deletions and translocations may result in activation of proto-oncogenes or inactivation of tumor-suppressor genes. In certain cancers of the lymphatic system (human Burkitt's lymphoma and mouse plasmocytoma), the normal proto-oncogene *c-myc*, which is involved in the regulation of cell proliferation, is translocated to the immunoglobulin locus. Immunoglobulin genes show very high transcription rates during development of B lymphocytes. The translocated *c-myc* is subjected to the same control mechanisms, resulting in inappropriately elevated expression of the gene and increased cell proliferation. In addition to these quantitative changes, chromosome aberrations may result in alterations of structure and function of cellular proteins. In many cases of human chronic myeloid leukemia, a piece of chromosome 9 carrying the proto-oncogene *c-abl* undergoes reciprocal translocation with a piece of chromosome 22. This results in joining of the proto-oncogene *c-abl* with a gene on chromosome 22 and production of a fusion protein encoded by both DNA sequences.

During cell division, chromosome segregation depends upon the functional integrity of the proteins of the spindle apparatus, that correctly divides the replicated genome into the two new nuclei. Impairment of this process, i.e. by chemical damage to the spindle apparatus, may result in nondisjunction, which means that both copies of a particular chromosome move into one daughter cell while the other receives none. Approximately 10% of the tumor types investigated so far are monosomic or trisomic for a specific chromosome. Aneuploidy is also an important cause of severe birth defects. Trisomy or monosomy of large chromosomes usually leads to early death of the embryo; trisomy of smaller chromosomes may allow the embryo to survive, but the newborn shows severe anatomical abnormalities and major physiological impairment. The most common syndrome of this group is Down's syndrome, which results from trisomy of chromosome 21.

Hence, structural and numerical chromosome aberrations may have detrimental effects on the health of the affected persons, and the following sections describe assays to detect chromosome aberrations in vitro and in vivo.

4.9.5.2. *In Vitro Cytogenetic Assays*

In the in vitro cytogenetic assays, proliferating cultured cells are treated with the test compound, and chromosomal aberrations are investigated after one or more cell cycles. Although cytogenetic assays can, in principle, be carried out in any cell type, Chinese hamster cell lines and human peripheral blood lymphocytes from healthy donors are generally used. In contrast to the spontaneously dividing Chinese hamster cell lines, peripheral blood lymphocyte cultures must be stimulated to divide by a mitogenic agent, such as phytohemagglutinin. Treatment is performed about 44 h after phytohemagglutinin stimulation, when the cells are proliferating. At least three doses of the test compound should be investigated, with the highest being in the low cytotoxic range. The usual recommendation is to treat the cells for ca. 1.5 normal cycles times, which is 15 h for Chinese hamster cells and 12–14 h for human peripheral lymphocytes. Many authors recommend longer treatment periods because some chemicals induce a mitotic delay at clastogenic doses (i.e., they increase the duration of the cell cycle). At the end of treatment, cells are induced to accumulate in the metaphase stage of mitosis with a spindle inhibitor such as colchicine or demecolcine, and at least 200 cells per treatment group are scored under the microscope for chromosomal aberrations. Chromosome breaks, fragments, and exchanges are considered specific structural aberrations, whereas gaps are usually excluded from the quantitation because the mechanism underlying their formation is not understood. Chromosome pulverization (complete fragmentation in small pieces) may also be regarded not as a parameter of clastogenicity, but rather as an indicator of severe cytotoxicity and cell death induced by the test chemical. Chemicals that are clastogenic in vitro at noncytotoxic concentrations are likely to be clastogenic in vivo. When bias resulting from lack of or insufficient biotransformation in the in vitro situation can be excluded, negative results in the in vitro assay provide a strong indication for absence of in vivo clastogenesis.

4.9.5.3. *In Vivo Cytogenetic Assays*

Induction of chromosome aberrations can be detected in intact animals either by examination of metaphases as described above for cultured cells or by the formation of micronuclei. Rats and Chinese hamsters are usually employed in metaphase analysis, while mice are commonly preferred in the widely used micronucleus test in bone marrow cells.

Rodent Micronucleus Test. The rodent micronucleus assay is an important part of the test battery for genotoxicity and is usually included in the toxicity evaluation of both pharmaceuticals and chemicals. The assessment of micronuclei is commonly conducted in polychromatic erythrocytes of bone marrow after treatment of young mice for six to eight weeks. Induction of numerical chromosome aberrations or chromosome breaks in the immature erythroblast of the bone marrow results in the formation of a micronucleus after cell division because the chromosome fragments are not participating in the chromosome segregation accomplished by the spindle apparatus. In the course of the physiological maturation of the red blood cells, the main nucleus is extruded from the erythroblast, while the micronucleus remains in the cell and can be detected in the young polychromatic erythrocyte after staining (Figure 66). The lack of nucleus in these bone marrow cells facilitates scoring for micronuclei under the microscope; however, peripheral blood cells and liver cells have also been used.

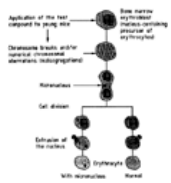

Figure 66. Mechanism of micronucleus formation in bone marrow erythroblasts

Usual treatment protocols include application of a single dose of the test compound and harvesting of bone marrow cells after 24, 48, and 72 h; or, alternatively, multiple applications (two or three doses 24 h apart) followed by a single harvest another 24 h after the last dose. The bone marrow is aspirated with a needle from the proximal end of the femur shafts and flushed with serum to obtain a homogeneous cell suspension, and after staining, the number of micronucleated immature polychromatic erythrocytes (not the number of micronuclei) is scored under a microscope. The immature erythrocytes stain bluish and can be distinguished clearly from the orange-red mature erythrocytes. Only tests with positive and negative control values within the historical control value of the laboratory are considered adequate. The minimum number of cells scored depends on the spontaneous frequency of micronuclei; 2000 cells should be scored if the control incidence is < 0.2% [172]. The ratio of polychromatic immature erythrocytes : monochromatic mature erythrocytes indicates the integrity of proliferation and differentiation processes in the bone marrow and is used as a measure of general toxicity. If the test compound has been administered at the maximum tolerated dose (as determined in the subacute toxicity test) and no increase in micronucleated erythrocytes can be detected, it is classified as negative for clastogenic activity in vivo. Clear positive results

in the in vivo assay indicate clastogenic potential also in humans.

Rodent Bone Marrow Metaphase Analysis. Metaphase analysis is usually carried out in rats at 6, 24, and 48 h after administration of a single dose or alternatively at 6 and 24 h after administration of multiple doses at 2 to 4 h prior to sampling. The animals are treated with spindle inhibitor such as colchicine and demecolcine to induce accumulation of cells in the metaphase. Metaphase analysis is much more time- consuming than scoring for micronuclei; therefore, only 50 cells per animal (500 per experiment with five male and five female animals) are usually analyzed [172]. In contrast to the rodent micronucleus test, the bone marrow metaphase analysis is not automatically included in the general screening scheme for genotoxicity; it is performed only when specific questions arise or for research purposes.

4.9.6. *Malignant Transformation of Mammalian Cells in Culture*

The term transformation in the present context means that a cultured cell line has taken on one or more morphological and/or malignant changes to give tumor cells. Among the various changes observed by exposure of cell lines to carcinogenic compounds, the following are two clearly established parameters indicating morphological transformation:

1. Inducation of anchorage independence (i.e., cells become capable of forming colonies in soft agar)
2. Loss of contact inhibition resulting in a tendency to grow in a piled-up criss-crossed pattern and to form foci

The term malignant transformation indicates that a cell line is capable of producing an invasive tumor in a suitable host. Hence, not every morphologically transformed cell line has necessarily undergone malignant transformation. On the other hand, human cancer cells are by definition malignantly transformed because they may metastasize under appropriate conditions.

The vast majority of cell transformation studies have been carried out with Syrian hamster cells [i.e., Syrian hamster embryo (SHE) cells and baby hamster kidney (BHK) cells] and with human fibroblasts.

SHE cells, isolated from 13-day-old embryos, can be induced to form piled-up criss- crossed colonies (foci) by treatment with carcinogens. Cells derived from these foci can produce tumors in athymic mice after extended subculturing (35–70 population doublings). Hence, the induction of malignantly transformed tumorigenic cells is a laborious, time- consuming procedure with several inherent problems (e.g., contamination of the plates in the course of the multiple passaging steps over several months). Therefore, the tendency developed to establish morphological transformation that can be induced in a comparatively short time as a short-term test for carcinogens. However, anchorage-independent cells showing criss- crossed growth pattern are usually not carcinogenic in animals. In addition, interlaboratory comparative elaboration studies indicated that scoring may be arduous and subjective, and in many cases the lack of dose-response and difficulties in obtaining consistently reproducible results in repeated assays may occur, even in the same laboratory [173].

Malignant transformation of cultured cells is currently an important tool to investigate the molecular mechanisms, i.e. mutations in oncogenes and tumor suppressor genes likely involved in the carcinogenic process. However, the use of cell transformation assays for screening purposes and for human risk assessment is very limited.

Two in vitro tests are required to provide a base level of information on the mutagenic potential of a chemical. These are a gene mutation test in bacteria and an in vitro mammalian cell test capable of detecting chromosome aberrations. For chemicals with significant toxicity to bacteria, an *in vitro* mammalian cell gene mutation test can be used as an alternative first test.

There are various options for selection of further test procedures in mammalian cells. An in vitro chromosome aberration test, i.e., a cytogenetic assay for structural chromosome aberrations using metaphase analysis will povide information on potential aneugenicity by recording the incidence of hyperdiploidy, polyploidy, and/or modification of mitotic index (e.g., mitotic arrest). A mouse lymphoma assay (L5178Y cells, TK locus) may detect gene mutations and structural chromosome aberrations but is not sufficiently sensitive for the detection of aneugens. These systems may be combined with an in vitro micronucleus tests, which is capable of detecting structural chromosome aberrations as well as aneuploidy.

4.9.7. *In Vivo Carcinogenicity Studies of Limited Duration*

In vivo tests of limited duration provide evidence for carcinogenicity in the whole animal in a short period (i.e., 52 weeks or less) compared to the lifelong bioassay. The experimental procedure usually involves administration of several doses of a known initiating carcinogen for a specific target organ to investigate promoting effects of the subsequently administered test compound. Alternatively, a known potent promoter for a specific tumor can be used to facilitate detection of initiating properties of the test compound. Also, many limited bioassays currently established or being developed aim to predict carcinogenicity by detecting cell lesions that consistently precede the appearance of the relevant tumor (i.e., induction of altered foci in the liver). In vivo carcinogenicity studies of limited duration are not automatically included in the usual test battery for carcinogenicity and are not accepted by regulatory authorities as a replacement for the conventional lifelong carcinogenicity study. The decision to carry out limited bioassays is made individually according to the available and required information on the test compound. Therefore, the present chapter includes only a brief description of the limited bioassays currently available.

4.9.7.1. *Induction of Altered Foci in the Rodent Liver*

In the course of rat liver carcinogenesis, cell foci exhibiting specific biochemical alterations precede the formation of liver tumors. These foci have, for example, abnormal concentrations of ϒ-glutamyl transpeptidase, glucose-6 phosphatase, adenosine triphosphatase, and the placental form of glutathione *S*-transferase and, unlike normal liver parenchyma, do not accumulate iron following iron loading. These parameters allow reliable and objective histochemical identification of preneoplastic foci as soon as three weeks after the initiation of treatment with known carcinogens. The yield is usually highest by the 12th to 16th week of exposure; therefore, the recommended approach is exposure for 12 weeks to the test compound with subcutaneous injection of iron during last two weeks to produce the iron load [174].

4.9.7.2. *Induction of Lung Tumors in Specific Sensitive Strains of Mice*

Certain strains of mice (e.g., the A/Heston) exhibit a high spontaneous incidence of lung tumors and are extremely sensitive to pulmonary carcinogens such as polycyclic aromatic hydrocarbons, certain nitrosamines, aflatoxin B_1 ethyl carbamate, hydrazines, and certain alkylating agents. The sensitivity of the test is highest 30 to 35 weeks after exposure initiation. Extension of treatment for a longer period is not recommended, because after 35 weeks the spontaneous lung tumor incidence increases rapidly in control animals as well, resulting in decreased sensitivity of the test.

4.9.7.3. *Induction of Skin Tumors in Specific Sensitive Strains of Mice*

The carcinogenic activities of several chemicals and crude mixtures can be readily revealed by their continuous application to the skin of a highly sensitive mouse strain, the Sencar mouse. Tars from coal, petroleum, or tobacco, as well as the pure polycyclic aromatic hydrocarbons and their congeners contained in such mixtures, give clearly positive results in this limited bioassay; in contrast, some arylamines and other established carcinogens do not elicit a positive response in this mouse skin test. These discrepancies are due mainly to differences in toxification or detoxification reactions between the systemic and the local applications. The mouse skin contains specific cytochrome P450 enzymes and peroxidases that can bioactivate polycyclic aromatic hydrocarbons, hence the differences in response. Therefore, this test cannot be considered a reliable predictor of the potential carcinogenicity of a xenobiotic or of human risk.

4.9.8. *Methods to Assess Primary DNA Damage*

4.9.8.1. *Alkaline Elution Techniques*

Many biochemical and analytical procedures are available to detect and quantitate the damage to DNA after contact with xenobiotics. Some of these procedures are very sensitive and thus permit detection and quantitation of DNA damage or DNA modifications after application of relevant doses of carcinogens to animals and after occupational or environmental exposure of humans.

The alkaline DNA filter elution was developed based on the observation that the rate at which large DNA single-strands pass through a membrane filter under alkaline denaturing conditions depends on the length of the strands. A broad spectrum of DNA damage types — both direct and indirect effects caused by effects of the toxicant on cellular function — can be detected with this techniques. The indirect effects include DNA single-strand breaks, DNA–protein crosslinks, and interstrand crosslinks (DNA–DNA).

The basic procedure (Fig. 67) operates essentially as follows: After treatment of cells in vitro or isolation of cells from an organ or a tissue of an animal treated in vivo, cell samples are placed onto membrane filters and lysed with a detergent- containing solution. This lysis solution is allowed to flow through the filter, thus removing most of the cellular protein and RNA, the intact DNA of cellular chromatin being retained on the filter. An elution solvent with a pH generally >12.0 is then pumped slowly through the filter to disrupt the hydrogen bonds between DNA strands. Treatment-induced DNA single-strand breaks produce DNA fragments with reduced molecular masses, thus increasing the rate at which DNA passes through the filters, whereas DNA–DNA or DNA–protein crosslinks decrease the rate compared to untreated controls. The amount of DNA eluted can be quantified either by using radiolabeled cell populations (by pretreating the cells with [^{3}H]thymidine) or by fluorogenic DNA-reactive compounds. The major limitations of the test are essentially the same as described in Section Test Systems Providing Indirect Evidence for

DNA Damage. DNA strand breaks do not necessarily indicate mutagenicity or interaction of a xenobiotic with DNA; they may also result from extranuclear damage such as increased production of reactive oxygen species due to impairment of mitochondrial functions. Hence, although a vast amount of data exists on the ability of mutagens and carcinogens to induce single-strand breaks, the test has only limited value as a screening method for mutagenicity.

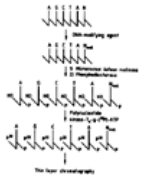

Figure 67. The ^{32}P-postlabeling procedure for detection of DNA damage ($N_{mod.}$ = modified deoxynucleotide)

4.9.8.2. *Methods to Detect and Quantify DNA Modifications*

The increased sensitivity of analytical methods has resulted in the development of several sensitive and selective methods to detect and quantify modifications of DNA bases induced by xenobiotics. These methods often rely on sophisticated chromatographic techniques to separate unchanged DNA constituents from modified DNA bases.

High-performance liquid chromatography (HPLC) with fluorescence detection may be used to detect and quantify DNA adducts formed from highly fluorescent xenobiotics, such as aflatoxin$_{B1}$ and polycyclic aromatic hydrocarbons. In addition, some DNA adducts of xenobiotics are highly fluorescent, such as alkylations of the N^7 atom of guanosine [175] and cyclic derivatives formed from vinyl chloride [176]. HPLC with electrochemical detection may detect oxidative DNA modifications, such as 8-hydroxydeoxyguanosine with high sensitivity and has been used to study DNA damage by cellular aging [177].

Gas chromatography/mass spectrometry (GC-MS) techniques are widely used in the detection and quantitation of DNA modifications. After derivatization to form volatile derivatives, DNA adducts formed by methylating and ethylating agents can be detected by GC-MS with selected-ion monitoring [178]. Due to the low detection limit (attomole range), GC-MS coupled with chemical ionization and negative-ion detection (after electrophore labeling with. e.g. pentafluorobenzyl bromide) has been applied in DNA adduct monitoring [179, 180]. However, the procedures are very time consuming and prone to artefact formation. At present, due to much simpler sample workup and high sensitivity LC-MS/MS methods promise to be useful for detecting DNA modifications.

A widely used method to detect DNA adducts of bulky organic substituents, which can not be transformed to volatile derivatives, is the ^{32}P-postlabeling technique [181]. The method can be widely used because it does not require the use of radiolabeled xenobiotic and sophisticated and expensive mass spectrometry. Due to the availability of ^{32}P-labeled adenosine triphosphate with high specific activity and the possibility of concentrating DNA adducts, this procedure may detect adduct frequences as low as one adduct in 10^{10} nucleotides. The method involves isolation of DNA from an animal or cells treated with the xenobiotic, enzymatic hydrolysis of the DNA to the 3′-nucleoside monophosphates and enzymatic phosphorylation to the 3′,5′-nucleoside diphosphates with ^{32}P-labeled adenosine triphosphate as phosphate donor. The obtained mixture of ^{32}P-labeled nucleosides and modified nucleosides is then separated by multidimensional TLC. Modified nucleosides are detected by placing the TLC plates on radiation-sensitive film; quantification can be performed by liquid scintillation spectrometry after cutting the adduct spots from TLC plates. Formation of artefacts and the poor resolution of the employed TLC method are problematic.

Several procedures to detect DNA adducts based on immunological methods have also been developed. These procedures have the advantage of being rapid and simple, but they are usually selective only for a specific type of adduct and require a time-consuming procedure to generate the antibody. In addition, sensitivity is often insufficient, quantification of adducts is complicated, and cross-reactivity generates artefacts.

Modern LC-MS/MS methods have sufficiently low detection limits to detect DNA modifications in a concentration range relevant for animal toxicity studies and do not require time-consuming sample preparation. Therefore, these methods have major potential for widespread use in the detection and quantitation of DNA modifications.

4.9.9. ***Interpretation of Results Obtained in Short-Term Tests***

Although no single rule applies in every case and each test compound should be evaluated individually, some general recommendations may be useful for a reasonable interpretation of results from short-term tests. Positive results in one in vitro test for mutagenicity (i.e., the Ames test) and one in vivo test (i.e., the rodent bone marrow micronucleus assay) indicate potential carcinogenicity. The test compound is highly suspect for potential carcinogenicity if clear- cut, dose-dependent evidence for genotoxicity in more than one in vivo and one in vitro test has been obtained, especially if the compound or its metabolites are structurally alerting for DNA reactivity.

This is supported by a recent evaluation on the activity of known human carcinogens in the *Salmonella* mutagenicity test and the rodent bone marrow micronucleus test [182]. As shown in Table 34, most of the human carcinogens identified so far are positive in these popular tests. Apart from hormones, no nongenotoxic organic chemical has been shown to cause cancer in humans thus far. In contrast, since the late 1980s an increasing number of organic chemicals has emerged that are clearly carcinogenic in rodents but are not mutagenic in *Salmonella* or clastogenic in rodent bone marrow. In addition, most of these nongenotoxic rodent carcinogens (and their metabolites) do not exhibit structural alerts for DNA reactivity. The reason for the differences between the epidemiological data in humans and the chronic rodent bioassays are not known. However, these considerations also suggest that although positive results in the short-term assays for genotoxicity indicate carcinogenic potential and may give reason to interrupt further development of a compound, negative results cannot be taken as a guarantee of the absence of carcinogenic activity in vivo, but rather as a green light to proceed to the lifelong rodent bioassay.

Table 34. Mutagenicity in Salmonella typhimurium and clastogenicity in rodent bone marrow of known human carcinogens [191]

Human carcinogens*	*Salmonella* mutagenicity	Rodent bone marrow chromosomal aberrations or micronuclei
Organic compounds		
Aflatoxins	+	+
4-Aminobiphenyl	+	+
Analgesics containing		
phenacetin	+	+
Azathioprine	+	+
Benzene		+
Benzidine	+	+
Betel quid and tobacco	+	+
Bis(chloromethyl) ether	+	(+)
Chlorambucil	+	+
Chlornaphazine	+	+
Cyclophosphamide	+	+
Melphalan	+	+
Mustard gas	+	ND(+)
Myleran	+	+

2-Naphthylamine	+	+
Tobacco, smokeless	+	+
Tobacco, smoke	+	ND(+)
Treosulphan	+	+
Vinyl chloride	+	+
Soot, tars, and oils		
Coal tar pitch	+	ND
Coal tar	+	ND
Mineral oil (untreated and mildly treated)	+	ND
Soot	+	ND(+)
Metals		
Arsenic compounds		+
Chromium compounds (hexavalent)	+	+
Nickel and nickel compounds		ND

+ = positive response; (+) = predicted positive response; ND = not tested so far.
* The carcinogenic hormones (essentially natural and synthetic estrogens) and fibers (asbestos, erionite, and talc containing asbestiform fiber) are not included; the members of these groups tested so far gave negative or inconclusive results in the *Salmonella* mutagenicity and rodent bone marrow clastogenicity assays.

4.10. Evaluation of Toxic Effects on the Immune System

In the last twenty years experimental and epidemiological evidence has accumulated that the mammalian immune system may be altered (i.e. suppressed or induced) by a wide range of environmental chemicals and drugs. Well-known examples are the polychlorinated and polybrominated biphenyls; dibenzo-*p*-dioxins; benzene; isocyanates; metals such as chromium, lead, and nickel; and organometallics such as di-*n*-octyltin chloride and tri-*n*-butyltin oxide.

Immunotoxicity is the ability of a substance to adversely affect the immune system: the immune response of affected individuals is altered. Immunotoxic responses may occur when the immune system is the target of the chemical insult; this in turn can result in either immunosuppression and a subsequent decreased resistance to infection and certain forms of neoplasia, or immune disregulation, which exacerbates allergy or autoimmunity. Alternatively, toxicity may arise when the immune system responds to an antigenic specificity of the chemical as part of a specific immune response (i.e., allergy or autoimmunity). Changes in immunological parameters may also be a secondary response to stress resulting from effects on other organ systems. Therefore, it must be recognized that in principle all chemical substances may be able to influence parameters of the immune system if administered at sufficiently high dosages. However, an immunotoxic effect should only be discounted when a thorough investigation has been performed. Although the immune system is considered as a target organ with regard to systemic toxicology, it consists of several different organ systems. A very large number of different cell types, present in practically all tissues and compartments of the human or animal body, participate in the immune response. Due to the complexity, it is not possible to describe the immune system in the context of the present chapter. Only a very brief description will be given to help the understanding of the current practical approaches of the evaluation of adverse effects on immune function induced by chemicals (see also ⟶ Immunotherapy and Vaccines – Introduction).

The overall immune system can be categorized in two major subsystems: the humoral and the cell-mediated subsystems. The effects of the humoral system are mediated by B lymphocytes producing antibodies that react with antigenic (usually foreign) material (i.e., antibodies attack bacteria and viruses before they can enter the host cell). The cell-mediated system involves primarily the mobilization of phagocytic leucocytes (macrophages) to ingest foreign organisms such as bacteria and the activation of T lymphocytes.

The two systems do not function independently in different situations; rather, they interact by complex feedback mechanisms that are only partly understood. One of the main properties of the immune system is the rapid production of a large number of cells capable to react with a specific antigen when this antigen is presented (again) to the organism. This important property is based on the presence of a wide variety of memory cells that were specifically adapted to the antigen at the time of initial contact.

Immunocompetent cells are required for host resistance, and thus exposure to immunotoxicants can result in increased susceptibility to any type of toxicity and disease including cancer.

The evaluation of adverse effects on the immune system can be carried out in two tiers [183, 184]. The first tier evaluates immune-related parameters (haematology, blood chemistry, organ weights and histopathology) that are included in revised standard testing protocols of repeated-dose (28 and 90 d) toxicity studies (Table 35).

Table 35. Immunological parameters that can be included in repeated-dose toxicity studies in rats for compounds for which no prior immunotoxic potential has been identified (tier I)

Parameters assessed	Tests
Hematology and blood chemistry	differential white blood cell counts
	bone marrow cellularity
	albumin : globulin ratio
	serum immunoglobulin classes
Organ weights	thymus, spleen, lymph nodes
Histopathology	thymus, spleen, lymph nodes, bone marrow, Peyer's patches (if oral administration)
	bronchus-associated lymphoid tissue (in case of pulmonary administration)

Compounds showing some immunotoxic properties in this first tier and also chemicals suspected of having immunotoxic effects based on information from prior studies or structure–activity relationships are further evaluated by functional assays that assess competence of immune cells (Table 36) For a detailed description of the experimental procedures see [196].

Table 36. Selected functional assays to assess immunotoxicity of compounds for which some immunotoxic properties have been implied in tier I evaluation (see Table 35) or in prior studies (tier II)

Parameter assessed	Tests
Cell-mediated immunity	mixed leukocyte response to antigenic determinants; induction of cell proliferation by the T-lymphocyte mitogen concanavalin A; T-lymphocyte cytotoxicity; delayed hypersensitivity response (DHR) to keyhole limpet hemocyanin (in contrast to the first three assays, that are carried out in cultured cells in vitro, the DHR is induced in vivo)
Antibody-mediated	antibody plaque-forming cell response; serum antibody titer after exposure to specific antigens determined by enzyme-linked immunosorbent assay (ELISA)

immunity	
Natural and induced host resistance	natural killer cell cytotoxicity; macrophage-mediated phagocytosis and intracellular killing; assessment of macrophage and T-lymphocyte competence by inoculation of *Listeria monocytogenes*; host resistance to melanoma cells

4.11. Toxicological Evaluation of the Nervous System

Neurotoxicity is the induction by a chemical of adverse effects in the central or peripheral nervous system, or in sensory organs. It is useful for the purpose of hazard and risk assessment to differentiate effects specific to sensory organs from other effects which lie within the nervous system. A substance is considered neurotoxic if it induces a reproducible lesion in the nervous system or a reproducible pattern of neural dysfunction.

The identification and characterization of neurotoxic properties of chemicals is one of the essential goals of every toxicity screening programm. Neurotoxicity has traditionally been associated with structural pathological modification of nervous system constituents (i.e. neurons, glial cells, and endothelial cells). Hence, the most important tool to assess neurotoxic effects in the past has been the histopathological examination of the nervous system of animals acutely or chronically exposed to xenobiotics. Over the last ten or twenty years, however, several reasons have called for the development and use of functional tests in neurotoxicity screening. First, detailed histopathological analysis of the nervous system is very time-consuming and requires experienced neuropathologists. Second, many chemical compounds disturb the nervous system without causing identifiable structural lesions. This clear need for functional tests to assess neurotoxicity has led to the development of a functional observational battery and an automated test to detect locomotor activity in rats based on methods that have been used by neuropharmacologists for many decades for the evaluation of psychoactive neurologic and autonomic pharmaceutical compounds.

4.11.1. *Functional Observational Battery*

The functional observational battery can be carried out both in acute and repeated-dose toxicity studies. The sequence of tests is usually arranged to progress from the least to the most interactive with the animal. The assessment begins with home cage observations followed by measurement made while handling the animals and assessment of activities in the open field. Assessment of reflexes, as well as physiologic and neuromuscular parameters is carried out at the end of the examination. Specific parameters assessed at these different stages are summarized in Table 37, and more detailed descriptions of some important tests are given below (reviews: [197-199]).

Table 37. General parts of the functional observational battery to assess neurotoxicity and specific examples of parameters recorded

A.	home cage observations	posture
		palpebral closure (eyelids wide open to completely shut)
		convulsions (clonic, tonic)
		biting
B.	observations while handling the animal	ease of removal from cage and of handling in hand, lacrimation, salivation, piloerection, fur appearance
C.	open-field activity (observations over 3-min period)	time to first step
		number of rears (supported and unsupported)
		number of urine pools and defacations
		mobility and gait
		tremors, ataxic gait, convulsions
D.	reflexes	approach response (e.g., to a pencil)
		touch response
		click response
		tail response
		pupil and eye blink response
		forelimb and hindlimb extension
		righting reflex: hold rat in supine position, drop from approximately 30 cm and note ease of landing
E	physiologic observations	catalepsy
		body temperature
		body weight
F	neuromuscular observations	rotarod performance grip strength
		hindlimb extension strength
		hindlimb foot splay

Catalepsy. Catelpsy can be measured by placing the rat on four corks (35 mm high, 40 mm in diameter, 100 mm between fore- and hindfeet, 60 mm between right and left feet). Duration of immobility at this position is measured for a period up to 60 s. For the assessment of catalepsy, a number of experimental modifications exist (i.e. placing the rats on a horizontal bar 12 cm above the ground). However, the four-corks procedure allows the best distinction between cataleptic and heavily sedated animals.

Grip Strength. The rat is allowed to grip a triangular ring with its forepaws and is pulled back along a platform until its grip is broken. As the pulling back continues, the animal's hindpaws reach a T-shaped rear limb grip bar, which it is allowed to grasp and then forced to release by continued pulling. Special devices are used to measure the maximum strain required to break forelimb and hindlimb grip.

Rotarod Performance (Rotating-Rod Test). This test requires preliminary training of the animals to walk on a rotating rod (7 cm in diameter, 5–12 rpm rotation rate). During the training period, the containers underneath the rod are filled with water to prevent the rats from jumping off. For testing, the time each rat remains on the rotarod is measured up to 2 or 3 min. Due to its objectivity and reproducibility this test is one of the most popular for the evaluation of adverse effects to the nervous system.

Hindlimb Foot Splay. The hind feet are painted, animals are dropped from a horizontal position 30 cm above a table onto paper and the distance between the middle of the ink spots is measured.

4.11.2. *Locomotor Activity*

Locomotor activity is not automatically part of the screening battery for neurotoxicity, the decision to carry out the test being rather made individually based on preliminary observations. Locomotor activity is assessed in special test chambers using a photocell detection procedure. Animal movement inside the chamber interrupts the photobeams and is translated into activity counts. Data on animal activity are usually recorded over three 5 min intervals.

In addition, a number of further methods to investigate neurotoxicity are available and standardized approach for an evaluation may not be given, but need to be decided on a case-by-case basis (Table 38).

Table 38. Methods for investigation of neurotoxicity

Effect	Methods available
Morphological changes	Neuropathology

	Gross anatomical techniques
	Immunochemistry
	Special strains
Physiological changes	Electrophysiology (e.g. nerve conduction velocity, NCV)
	Electroencephalogram (EEG, evoked potentials)
Behavioral changes	Functional observations
	Sensory function tests
	Motor function tests (e.g., locomotor activity)
	Cognitive function tests
Biochemical changes	Neurotransmitter analyses
	Enzyme/protein activity
	Measures of cell integrity

A major problem in neurotoxicity screening is that the general behaviour of rats is not stable and uniform, even when they are kept under strictly controlled conditions and not exposed to neurotoxic compounds. Therefore, many authors suggest to carry out the tests without knowledge of the treatment protocols. Furthermore, since many observations are subjective, it is important to use experienced and well trained scientists and technicians in the evaluation.

4.12. Effects on the Endocrine System

Although endocrine disruption is often regarded as a specific end point in toxicity testing, it is just a mechanism by which a chemical may induce adverse effects. Many endocrine-dependent toxicities will be detected in the course of the already available methods for toxicity testing. However, a number of specific tests for endocrine effects of chemicals are under discussion for introduction into toxicity testing.

The endocrine system consists of a number of glands such as the thyroid, gonads, and the adrenals, and the hormones they produce such as thyroxine, oestrogen, testosterone and adrenaline. These hormones may influence development, growth, reproduction and behaviour of animals and humans.

Endocrine disrupters are defined as:
- Chemicals that have properties that might be expected to lead to endocrine disruption in an intact organism, its progeny, or (sub)populations
- Chemicals that alter functions(s) of the endocrine system and consequently may cause adverse health effects in an intact organism or its progeny

Endocrine disrupters may to interfere with the endocrine system by several mechanisms:
- By mimicking the action of a naturally produced hormone such as oestrogen or testosterone and thereby inducing similar chemical reactions in the body
- By blocking the hormone receptors in cells and thus preventing the action of normal hormones;
- By affecting the synthesis, transport, metabolism, and excretion of hormones and thus altering their concentrations.

A variety of test systems to characterize the potential of chemicals for endocrine disruption, ranging from receptor-binding assays to complex in vivo measurements, have been proposed, but a specific testing approach has not yet been agreed.

Some testing guidelines (e.g., the 28-d study guideline) detect effects on endocrine function, but there are no test strategies/methods available which can detect all possible effects that may be linked to the endocrine disruption mechanism.

[Top of Page]

5. Evaluation of Toxic Effects

One of the major environmental and occupational issues of concern to both scientists and administrators is the control of potential health hazards to humans due to the production, use, and disposal of chemicals. The concern arises from the increasing numbers of chemicals in production and use and the increasing numbers of chemicals demonstrated to exert toxic effects in one or several of the sensitive toxicity testing systems available. This situation has afforded growing legislative control of the production and application of chemicals to ensure adequate protection of human health. Control measures based on the recognition of potential adverse health effects may limit the presence of hazardous chemicals in the environment or regulate the use of hazardous chemicals, thus reducing the potential health risks to humans (Table 39).

Table 39. Possible measures to reduce human exposure to hazardous chemicals

Application or exposure to chemical in question	Measures to reduce exposure
Industrial chemicals	reduction or cessation of application; protective measures in the workplace; alternative chemicals with lower hazard
Pharmaceuticals	cost – benefit analysis
Alcohol, smoking, drugs of abuse	education
Environmental chemicals	quantitation of exposure, strategies for avoidance or reduction of environmental pollution

The assessment of potential human health risks resulting from the exposure to chemicals provides the basis for appropriate regulatory and control measures. The health risk assessment determines whether a xenobiotic may cause adverse health effects, at what level and frequency of exposure, and the probability that adverse health effects will occur. The term “risk assessment” is increasingly used in the context of potentially toxic chemicals. Scientific risk assessment considers the available data on the toxicology of a specific chemical when judging which agents potentially pose a significant risk to the human population. Toxicology focuses on the identification and quantitation of potential hazards by using animal studies as surrogates for humans. Permissible exposure levels for humans are derived from the results of the animal studies by using margins of safety or defining “acceptable” incidences of adverse health effects in exposed humans [185].

Health risk assessment and its use in regulatory decisions have recently generated intense controversy. The debate over risk assessment is politically and emotionally charged, and creates an adverse atmosphere heightened by the extraordinary sums of money at stake. Industry complains that the costs of complying with possible overregulation based on inappropriate risk assessments may be excessive; moreover, lawsuits on potential environmentally caused diseases, especially in the United States, involve huge sums of money. On the other hand, environmentalists claim that risk assessment practices and policies do not adequately protect human health; moreover, health care costs for the treatment of environmentally caused diseases may also be very high. These considerations have led to an intensive rethinking of the health risk assessment process and have increased the awareness that in many cases, the scientific foundation for risk assessment is weak. This rethinking led to the conclusion that resolution of the controversies by the development of effective prevention strategies and rational priority setting may be achieved only by strengthening the scientific background and available data by research and by developing better methods to estimate risks due to chemical exposures [185-187].

Before considering the practice of health risk assessment, several terms frequently used and misused in risk assessment and its perception should be clarified. In discussions on health effects of potentially toxic chemicals, the terms “hazard” and “risk” are often used with an identical meaning, although they are clearly different. Hazard defines the intrinsic toxicity of a chemical and is not identical to risk. Risk is the estimated or measured probability of injury or death resulting from exposure to a specific chemical. Risk may be described either in semiquantitative terms such as high or low risk or in quantitative terms such as one person experiencing an adverse effect per 10 000 persons exposed. Risk may also be described in absolute terms (probability of adverse effects due to a specific chemical exposure) or in comparative terms by comparing the probability of adverse effects between a population exposed to an agent and an unexposed population.

The health risks due to the contact with potentially toxic chemicals are dependent on the conditions of exposure, since not only the intrinsic toxicity of a chemical determines the magnitude ot the adverse effect but also the dose. As noted in Chapter 1, the magnitude of the toxic effects is the product of the intrinsic toxicity of a chemical multiplied by the dose taken up by exposed animals or humans; thus, all toxic effects are dose-dependent and even very toxic chemicals may not cause toxic effects when the dose is low. If the dose is zero, despite a very high intrinsic toxicity of a specific chemical, the toxic effect and the risk of adverse health effects will be zero. On the other hand, chemicals with low intrinsic toxicity may induce toxic effects when the dose is high and may thus pose a significant risk. In toxicological terms, risk is therefore the product of the intrinsic toxicity of a chemical and the exposure characteristics.

5.1. Acceptable risk, Comparison of Risks, and Establishing Acceptable Levels of Risk

In earlier phases of risk assessment, the basic belief was that few chemicals are toxic and all of these toxic chemicals are derived from synthetic processes. To achieve a zero risk, chemical exposure must be reduced below a threshold level, under which it causes absolutely no risk. However, where such a threshold cannot be demonstrated, one must assume that a finite risk may occur at any exposure level, consequently, absolute control of risk is possible only if the source of exposure is eliminated altogether. These considerations resulted in the zero-risk concept. The Delaney Clause of the Federal Food and Drug Act in the United States is an example of a zero-risk approach in the regulation of food additives. This law states that no xenobiotic whose carcinogenic potency in animals has been demonstrated may be used as a food additive.

However, the more widespread testing of chemicals for toxicity, the increased sensitivity of analytical instruments to determine chemicals in the environment and at the workplace, and the developments in the science of toxicology put the basic assumption of the zero risk concept — that is, only synthetic chemicals are toxic — into question.

However, the more widespread testing of chemicals for toxicity, the increased sensitivity of analytical instruments to detect chemicals in the environment and the workplace, and developments in the science of toxicology put the basic assumption of the zero-risk concept — that only synthetic chemicals are toxic — in question. These developments led to the recognition that zero risk was unachievable and, perhaps, unnecessary for the regulation of chemicals. The observation was based mainly on a few facts: (1) all chemicals, both of synthetic and natural origin, are toxic under specific exposure conditions; (2) most of the hazardous chemicals routinely encountered by humans are of natural rather than synthetic origin; (3) most of the exposure to hazardous synthetic chemicals cannot be avoided entirely or be eliminated from the environment without changing profoundly the way of life in many countries; and (4) in the case of cancer risk assessment, DNA damage and mutations, assumed to be of major significance in the process of carcinogenesis, occur spontaneously, albeit at a low rate. Examples of endogenously occurring DNA damage are hydrolytic deamination, depurination, oxidative modification, and endogenously formed DNA adducts. Well-known examples of hazardous synthetic chemicals are benzene, which is present in the environment as a result of its emission from motor vehicles, cigarette smoking, and other sources, or 2,3,7,8-tetrachlorodibenzodioxin formed in forest fires. The effect of naturally occurring chemicals and chemical exposure due to life-style factors is best exemplified by the estimated contribution of different factors to the incidence of avoidable cancers in humans (Fig. 68).

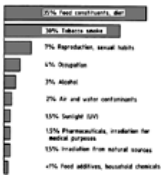

Figure 68. Contribution of chemical exposure and life-style factors to the incidence of avoidable cancer in human [200]

According to the large epidemiological study of DOLL and PETO (1981), natural chemicals in diet and chemicals inhaled by cigarette smoking are the major causative agents in human cancer. Occupational and environmental exposure to synthetic chemicals constitutes only a minor causative factor.

Given these facts, the acceptable risk concept was developed as an alternative. The acceptable risk concept realizes that it is not possible to eliminate all potential health risks associated with chemical exposure due to the life style. According to the concept of acceptable risk, safety — the reciprocal of risk — is no longer an absolute term but is redefined as a condition of certain, but very low and thus acceptable, risk. This conceptual change improves the ability to deal with potentially very low risks identified by the increased sensitivity of analytical instrumentation and with increasingly sensitive scientific methods to detect potential adverse effects of xenobiotics. The concept of acceptable risk also permits the definition of limits for the exposure to toxic chemicals that can be considered to have a negligible impact on the incidence of adverse effects in an exposed population. Risk assessment is therefore unavoidable and must implicitly or explicitly involve a balance of risk and benefit [185-189]. Some of the main factors considered in establishing acceptable risk levels for exposure to a chemical follow:

- **Beneficial aspects**
  - Economic growth
  - Employment
  - Increased standard of living
  - Increased quality of life
  - Taxes generated

- **Detrimental aspects**
  - Decreased quality of life
  - Health effects
  - Lawsuits
  - Loss of environmental resources
  - Loss of work
  - Medical expenses

5.2. The Risk Assessment Process

Several individual elements make up the risk assessment process. In the first step, the potential adverse health effects of a chemical, a mixture of chemicals or a specific technical process are evaluated by the application of toxicity tests (for details, see Chap. 4) and, if the chemical is already in widespread use and humans are exposed, by considering the data from epidemiological studies. The second step in risk assessment determines the dose–response for the observed adverse effects. In parallel, the exposure of humans to this xenobiotic is quantitated by analytical procedures or, if the chemical is not yet in widespread use, by the estimation of likely exposure scenarios. In the third step, the results obtained in the toxicity studies are extrapolated. This involves an extrapolation both from adverse effects seen in experimental animals to humans, and often an extrapolation from the effects seen after high doses in animals to the much lower doses humans usually encounter. The last step of risk assessment, risk characterization, involves the combination of steps one to three to judge the existence and magnitude of the public health problem and characterizes the uncertainties inherent in the risk assessment process

5.2.1. *Hazard Identification Techniques*

Hazard identification is the step in which the adverse effects of the xenobiotic are determined. Evaluation of both acute and chronic toxicity is performed by using animals as experimental models for humans. The use of animals as surrogate for humans is based on the following assumptions: (1) xenobiotics with a likely adverse effect in humans will manifest some degree of toxicity in other living systems when the dose is sufficiently high; (2) if a sufficient number of animal species are dosed with the xenobiotic, at least one should exhibit a similar pattern of biotransformation and toxicokinetics to that seen in humans; and (3) if a sufficient number of different animal species are dosed with the xenobiotic, at least one is likely to exhibit the toxic responses and clinical symptoms occurring in humans.

Typical end points in studies aimed at hazard identification in intact animals include mortality, reproductive and developmental effects, target organ toxicity, and cancer. Hazard identification studies at present also include the determination of a range of biochemical end points related to specific toxic effects such as toxicokinetics, routes and extent of biotransformation, structure of reactive intermediates, and binding of reactive intermediates to cellular macromolecules. In addition, many short-term tests for specific toxic effects such as DNA damage, mutagenicity, or clastogenicity are increasingly included in hazard identification procedures.

The acute toxicity of xenobiotics is evaluated by a number of procedures from which the LD_{50} may be calculated. With the more recently recommended fixed-dose method information on target organs affected and types of toxic effects may also be obtained. Repeated-dose toxicity studies last between two weeks (subacute toxicity studies) and 6–24 months (chronic toxicity studies), the lifespan of the animals, including post-mortem examination, histopathology, clinical chemistry, and

hematology at termination and at specified time points during the study. From the chronic toxicity studies, the lowest observed effect level (LOEL), also referred to as the lowest observed adverse-effect level (LOAEL), and the no observed effect level (NOEL) are obtained for noncancer endpoints. The NOEL is the highest does administered that does not induce observable toxic effects. The NOEL may not be identical to the no-effect level if insensitive methods are applied or the wrong end point is chosen. Moreover, the value obtained for the NOEL will depend on the number of animals used in the study and the spacing of the applied doses.

Short-Term Tests for Specific Toxic Effects In Vivo and In Vitro. A variety of short-term tests has been developed for the detection and quantification of toxic effects. Most of the more established and well-evaluated test systems are designed to evaluate the genotoxic activities of xenobiotics and employ well-defined genetic changes (DNA damage, gene mutations, chromosome defects, cell transformation) in prokaryotes, lower eukaryotes, and mammalian cells. Knowledge that a genotoxic chemical is active in vivo in the target organ of carcinogenesis enhances confidence that the genotoxicity of the chemical is important in the process of cancer induction by that chemical. Before any short-term test can be used with confidence to assess potential toxic effects, its validity should be thoroughly evaluated and its major drawbacks should be explicitly noted. Moreover, it should be kept in mind that most available short-term tests are well designed to give qualitative information and their major use for risk assessment purposes is therefore the confirmation or exclusion of a specific toxic response. The magnitude of toxic response in intact animals or humans depends on both toxicokinetics and toxicodynamics. The toxicokinetic phase of the toxic response is not considered in most of the in vitro short-term tests. Therefore, short-term tests in vitro may be used for hazard identification only in combination with studies on adverse effects in animals and studies on the toxicokinetics of a xenobiotic.

Biotransformation and Toxicokinetics. Studies on the extent of biotransformation including structural identification of the metabolites formed from the xenobiotic both in intact animals and in appropriate in vitro systems such as organ homogenates or fractions with enzymatic activity also contribute to hazard identification [190]. The structures of metabolites formed and the presence and extent of covalent binding of metabolites to macromolecules such as protein and DNA indicate the formation of electrophilic metabolites and thus a potential hazard.

For example, the structure of excreted mercapturic acids may give information on the structure and reactivity of the electrophilic metabolite and sites of cellular interactions [191]. Information on the rate of absorption and elimination may indicate a possible accumulation of the xenobiotic in humans, with the consequence of potential adverse effects. These studies should be performed on at least two animal species in vivo. For in vitro studies, human tissue samples should be included to confirm that biotransformation reactions are identical to those, observed in animals occur in humans. Different mechanisms of toxicity may operate at different dose levels; in these instances, toxicokinetic data may help in understanding dose-dependent mechanistic differences. The toxicokinetics of a xenobiotic in humans may be extrapolated by physiologically based pharmacokinetic models from the results obtained in experimental animals. Information on potential pathways of biotransformation may also be made by computerized structure analysis of the xenobiotics with specific computer programs [147, 192]. These are designed for predicting routes and rates of biotransformation based on the presence of functional groups in the molecule; however, at present, the available programs are far from perfect, and they should only be used in conjunction with expertise and as a basis for experimental planning.

Structure–Activity Relationships and Chemical Structure Analysis. Predictive data on the potential of xenobiotics, mainly organic chemicals, to induce adverse effects may be derived from relationships between chemical structure (physicochemical properties, presence of functional groups, atomic configuration) and biological activity, termed structure–activity relationships.

Two approaches to hazard prediction utilize structure–activity relationships. The first is essentially a qualitative approach and involves the comparison of the structure of the xenobiotic with that of other compounds already known to cause specific toxic effects. The comparison with known structures and the knowledge of biotransformation reactions and mechanisms of toxicity permits the identification of toxophores in the structure of xenobiotics. Toxophores are functional groups present in the molecule which are likely converted to toxic metabolites or posess chemical reactivity related to mechanisms of toxicity. Examples of toxophores are olefinic moieties, which may be oxidized to epoxides, and terminal carbon atoms in aliphatics bearing two halogen atoms, which may be oxidized to acyl halides. The scope and usefulness of structure–activity relationships depend on the availability of a sufficiently large database on the toxic effects of chemicals with common subgroups and structures [147], [192]. The main limitation of this approach is the qualitative nature of the resulting estimate of potential toxicity. The second approach involves quantitative structure–activity relationships and relies on the computer-aided analysis of databases on toxic effects of chemicals. A basic feature of the applied techniques is the use of pattern-recognition schemes or substituent weighting factors coupled with regression analysis. Some useful predictions have been made with these techniques, but due to the complex nature of toxic effects and the multitude of factors governing the toxic response, there are still severe limitations to the applicability of quantitative structure–activity relationships for predicting toxicity profiles.

Clinical and Epidemiological Studies. In principle, the best evidence for the toxic effects of a chemical in humans is derived from clinical and epidemiological studies. These studies assess effects in the species of interest for risk assessment and at relevant concentrations. Potentially confounding extrapolations from high to low dose and from animals to humans are not required. However, several weaknesses limit the applicability of these studies to hazard identification. In most cases, reliable exposure data are lacking, consequently, dose–response relationships cannot be established. In addition, the sensitivity of epidemiological studies to detect health problems is comparatively low. Unless the toxic effect of a particular xenobiotic is very unusual in control (unexposed) groups, it may pass unnoticed in a normal survey. Examples are the identification of asbestos exposure as a cause of mesothelioma or vinyl chloride as a cause of hemangiosarcoma, very rare forms of cancer in humans not exposed to asbestos or vinyl chloride, respectively. Moreover, the results of epidemiological studies, especially on cancer risk, may reflect the risk associated with exposure to chemicals decades ago because of the long latency period. Other limitations are confounding variables such as smoking and concomitant exposure to other xenobiotics, which often impair the interpretation of carcinogenicity data in humans. Thus, evidence based on epidemiological observations has identified only a limited number of chemicals as human carcinogens; many of the identified compounds are used in cancer chemotherapy and have the intrinsic property of genotoxicity (Table 40).

Table 40. Examples of established human carcinogens based on epidemiological observations

Chemical or agent	Site of tumor formation
Aflatoxin	liver
Alcoholic drinks	mouth, esophagus
4-Aminobiphenyl	bladder
Benzidine	bladder
2-Naphthylamine	bladder
Arsenic	skin, lung
Asbestos	lung, pleura, peritoneum
Azathioprine	reticuloendothelial system
Benzene	bone marrow
Bis(chloromethyl) ether	lung
Cadmium	prostate
Chlorambucil	bone marrow
Chlornaphazine	bladder
Chromium	lung
Cyclophosphamide	bladder
Bis(2-chloroethyl) sulfide	larynx, lung
Nickel compounds	nasal cavity, lung
Estrogens	endometrium, vagina
Phenacetin	kidney and lower urinary tract
Polycyclic aromatic hydrocarbons	skin, scrotum, lung
Steroid hormones	liver
Tobacco	mouth, pharynx, larynx, esophagus, lung, bladder
Vinyl chloride	liver

5.2.2. *Determination of Exposure*

The quantification of exposure, both in individuals and in populations, is a prerequisite for the quantification of risk. Reliable data on exposure are needed to assess the adverse effects of the xenobiotic and to recognize specific risk factors such as occupation, life style, and social status. The dimensions of exposure include

intensity, frequency, route, and duration; in addition, the nature, size, and makeup of the exposed population should be characterized. The assessment of exposure is a difficult and complex task, and is often neglected. Typically, estimations and field measurements are required. The estimation of human exposure to a particular xenobiotic involves an initial estimation of the possible sources of the chemical and the possibilities for exposure. A good inventory of sources may provide important information on critical pathways of exposure, populations at particular risk, and the levels of exposure.

In many cases, the duration and level of exposure, especially after chronic contact, may only be estimated from ambient levels of the xenobiotic in the environment, and estimations may thus be crude; owing to the large numbers of potentially exposed persons, only in special situations (e.g., occupational exposure, after disasters), will exposure data, including determination of the internal dose, be available.

Specific procedures to detect exposure to a certain xenobiotic and procedures to estimate exposure by determining biological effects may be used. These include:

- Direct measurement of the chemical in environmental samples such as water, air, and soil.
- Measurement of the chemical, its metabolites or products of the interaction of the chemical or its biotransformation products with cellular macromolecules (protein and/or DNA) in body fluids and tissues (biomonitoring).

Biological end points in exposure assessment may be:

- Assessment of biochemical indicators for specific adverse effects known to be caused by the xenobiotic, e.g., inhibition of specific enzymes such as cholinesterase activity in persons exposed to organophosphate pesticides.
- Observation of pathological evidence of exposure such as cytogenetic changes in lymphocytes from workers exposed to chromosome-damaging chemicals at the workplace. However, the use of biological end points for exposure assessment lacks the resolving power to discriminate between endogenous changes and the effects of xenobiotics, and therefore chemical specific indices of exposure should be favored for quantitation of exposure [178, 193].

Due to the time-consuming and cost- and labor-intensive procedures required, data on exposure to xenobiotics are usually limited. Difficulties in identifying concomitant exposures, interactions with other xenobiotics or activities, special risk groups such as the very old or very young and pregnant women, and patterns of exposure result in a high degree of uncertainty in exposure assessment in human populations.

A stepwise approach to exposure assessment using the following hierarchy can be applied. The most reliable exposure assessments are based on measured data, including the quantification of key exposure determinants. When these are not available, appropriate surrogate data may be used. Modeling may also be used in the absence of useful data, but the limitations of the modeling approach should be clearly stated.

5.2.3. *Dose-Response Relationships*

The establishment of the dose–response relationship for adverse effects in animals is the decisive step in risk assessment. This step quantifies the relationship between received or administered dose and biological response and may be performed on an individual or population basis. Dose–response assessment includes exposure intensity and duration and factors modifying toxic response such as sex, age, health status, and route of administration. Since the dose–response assessment can only rarely be performed in humans, extrapolation from data obtained in animals to humans is usually required. Moreover, many animal experiments, particularly carcinogenicity bioassays, are performed with high doses to increase the sensitivity of the assay. Therefore, besides species extrapolation, an extrapolation from effects seen after high doses in animals to the low doses usually encountered by humans is necessary. This extrapolation step from high, sometimes toxic doses in animals to low doses in humans is controversial.

The extrapolations form high to low dose are performed differently depending on the type of toxic response elicited by the xenobiotic. As explained in Chapter 1, toxic effects of a chemical may be caused by both reversible and irreversible interactions of the xenobiotic or its metabolite(s) with macromolecules in the organism. Many acute toxic responses such as carbon monoxide poisoning are based on reversible interactions and are associated with thresholds. Threshold doses are doses below which the probability of a response is zero. The biological basis for thresholds is well founded and may be demonstrated on a mechanistic basis. On the other hand, many chronic toxic responses, particulary chemical carcinogenesis, are often considered nonthreshold effects. Since a negative can never be proven, the absence of thresholds cannot be demonstrated by experiments and is based on consideration of the mechanisms of chemical carcinogenesis. Different approaches have thus been developed in establishing acceptable levels of exposure to threshold and nonthreshold responses (see below).

5.2.4. *Risk Characterization*

The hazard identification, the exposure assessment, and the dose–response assessment merge into risk characterisation. Risk characterisation estimates the incidence of expected adverse health effects in exposed populations. As noted in Section Hazard Identification Techniques, risk characterization and the establishment of acceptable exposure levels are handled differently for carcinogenic and noncarcinogenic xenobiotics. For chemicals that cause adverse effects by mechanisms with thresholds, the safety factor approach was developed. For nonthreshold responses such as cancer, both quantitative and qualitative risk assessment procedures are used.

5.2.4.1. *The Safety-Factor Methodology*

When the safety assessment of a chemical is based on animal toxicity testing, a different evaluation of the data is required compared to safety evaluations relying primarily on human observations. Since animal experiments to determine health hazards of chemicals were widely used in the 1940s and human health risks had to be extrapolated from these data, the idea of a safety factor, which was to be applied to the results from animal studies, was developed. The safety factor approach was first proposed in the 1950s by ARNOLD LEHMANN of the FDA as a protective measure for human health and is intended to compensate for uncertainties in the extrapolations of animal data to humans. This concept was applied to the determination of acceptable daily intakes (ADIs) by the World Health Organization from the 1960s onward. The ADI was defined as “the daily intake of a chemical which, during an entire lifetime, appears to be without appreciable risk on the basis of all known facts at that time.”

The safety factor, or as it is sometimes called, the uncertainty factor, has been introduced to consider both interspecies and interindividual differences in response to potential toxic effects of the chemical under consideration. The major purpose of the safety factor is the protection of human health by establishing safe exposure levels; the exposure levels defined do not mean that exposure above these levels will result in adverse effects. Despite several limitations and criticisms, the safety factor approach has been used for many years in Western Europe and the United States, and has proven useful and reliable. It was thus adopted internationally as the standard procedure for assessing the ADI.

This risk assessement approach is based on the establishment of a point of departure for the required extrapolation, which may be an NOEL or NOAEL or a benchmark dose (e.g. ED_{05}) in animal studies on chronic toxicity, and then defining permissible human exposures by the application of a safety factor. In this context, the NOEL is defined as the lowest dose of the xenobiotic in an animal experiment which produced no detectable effect in the most sensitive animal species treated. Once a NOEL has been determined, a safety factor for human exposure is introduced and, often, a permissible level of exposure of one hundredth of the NOEL in animals is defined for humans. The 100-fold safety factor is justified on the basis of a 10-fold difference to reflect an interspecies difference in susceptibility and a 10-fold difference to reflect possible interindividual variations in susceptibility in humans [188]. The acceptable daily intake is then obtained by dividing the NOEL from the study by the safety factor.

The safety-factor approach assumes that toxic effects exerted by the chemical exhibit a dose–response curve with a threshold, that the results of the toxicity studies in animals are relevant to humans, and that extrapolation of the dose is reliable.

A similar approach is the margin-of-safety (MOS) approach which determines by how much the derived NOAEL exceeds the determined or estimated exposure. The size of the MOS (< 100 or > 100) determines the extent of concern of a specific exposure.

As noted above, safety factors of 100 are frequently applied. However, when reliable data for adverse effects of the chemicals in humans are available, a safety factor of only 10 may be applied. On the other hand, a safety factor of more than 100 is appropriate when no or only limited data on the toxicity of the chemical in animals are available. Note that the safety factor is not based on scientific evidence. Additional data on the mechanisms of toxic effects in animals and on the toxicokinetics and biotransformation of the xenobiotic in animals and humans are required and add scientific credibility to the proposed safety factors. Therefore, the choice of a safety factor should consider, in addition to the NOEL, a series of further qualitative parameters:

•Evaluation of animal toxicity studies

- Number of studies and effects observed
- Type of toxic effects
- Time course for toxic effects
- Tumorigenicity

•Evaluation of biochemical end points
- Biotransformation and toxicokinetics
- Mechanism of action, covalent binding to macromolecules
- Short-term test for genotoxicity and other nonthreshold effects

•Evaluation of species differences
- Interspecies variations in biotransformation and toxicokinetics
- Influence on anatomical and physiological differences of toxic effects between species

Application of the most appropiate safety factor requires careful analysis of the data available on the toxic effects for every chemical, on a case-by-case basis. Consideration should be given to the quality and completeness of the data and the number and spread of the dose levels used. Besides the derived NOEL, all further information obtained in the long-term animal toxicity studies is valuable for characterizing toxic effects and determining safety factors. The type of toxic effect and the shape of the dose–response curve should also be taken into account in setting the size of the safety factor.

Also important are considerations of the ability to extrapolate from toxic effects seen in animals to human exposure scenarios and the results of human epidemiology studies, when available. Because of the complexity of information required, no universally accepted guidelines can be developed for determination of the precise magnitude of safety factors; therefore, expert judgment on an individual basis is a major factor contributing to the size of safety factors.

The safety factor approach has several problems. In experiments to determine the NOEL in animals, the experimental group size and the spacing of dose levels are major determinants of the numerical value of the NOEL obtained; the smaller the size of the individual dose groups and the larger the spacing of doses, the less likely is it that an effect will be observed. This phenomenon has the effect of rewarding poor experimental design because small group sizes in experiments will tend to produce higher NOELs. Moreover, nonthreshold effects may not be detected due to the low number of animals enrolled or the dose levels applied. In addition, the slope of the dose-response curve is often not considered or may not be determined with sufficient accuracy, and chemicals with steep and shallow dose–response curves are treated alike. Advantages and disadvantages of safety factors in the risk assessment process follow:

•Advantages
- Simple application
- Ease of understanding
- Flexibility of use
- Use of expert judgment

•Disadvantages
- Uncertainties of threshold values and size of safety factor
- No risk comparison possible
- Slope of dose–response curve not adequately considered
- Experimental NOELs are dependent on group size in the animal toxicity testing and end point selected

One of the most promising alternatives to the use of NOAELs or NOELs is the benchmark concept [194]. In the benchmark approach, a dose–response curve is fitted to the complete experimental data for each effect parameter. On the basis of the fitted curve, the lower confidence limit of the dose at which a predefined critical effect size is observed (i.e., the dose at which adverse effects start to arise or where 5% of the animals are predicted to be affected, effective dose, or ED_{05}) is defined as the benchmark dose. Advantages of this approach over the NOAEL are:
- The benchmark dose is derived by using all experimental data and gives a better reflection of the dose–response curve.
- The benchmark dose is independent of predefined dose levels and spacing of dose levels.
- The benchmark approach makes more reasonable use of sample size, and better study designs result in higher benchmark doses.

A disadvantage of this method is the uncertainty with respect to the reliability of the approach when results are obtained from toxicity studies performed according to the requirements defined in current guidelines. For the derivation of reliable dose–response relationships, the classical study design of three dose groups and a vehicle control group is limited, since adverse effects may only be observed at the highest dose level. An improved benchmark model fit could be achieved by increasing the number of dose groups without changing the total number of animals in the test.

At present, the determination of a NOAEL is mandatory for risk assessment in the EU. Nevertheless, the benchmark dose method can be used in parallel when a NOAEL cannot be established for the selected toxicological end point because only a LOAEL is available. In this case, benchmark modeling preferred over LOAEL–NOAEL extrapolation, which uses more or less arbitrary assessment factors. Benchmark dose software (BMDS) is available from the US EPA internet site (www.epa.gov).

5.2.4.2. *Risk Estimation Techniques for Nonthreshold Effects*

The procedures outlined below are most often applied to the low-dose risk estimation of human or animal carcinogens. Since dose–response data are not available for effects in animals at doses relevant to human exposure, extrapolations are required for determining the potential human cancer risk. The methodology employs mathematical modeling to characterize the relationship between exposure and response or to place an upper bound on the dose–response relationship. Dose–response data, available from specific study situations (mainly animal bioassays using high doses of the chemical and, sometimes, heavily exposed population groups), are extrapolated to the often much lower exposures of the general population in order to calculate the possible risk. Therefore, cancer risk assessment generally involves extrapolating risk from the relatively high exposure levels employed in animal studies, or occupational studies where cancer responses can be measured, to risks at the relatively low exposure levels that are of environmental concern. However, since the majority of carcinogenicity experiments use only two or three doses, it is impossible to assess the shape of the dose-response with a reasonable degree of precision. Risk assessment must therefore rely on some arbitrary assumptions about the shape of the dose-response relationship at low doses. Risk estimates thus obtained are not true or actual risk but values obtained by extrapolation well below the range of experimental observations. A summary of the advantages and disadvantages of quantitative risk assessment follows:

•Advantages
- Gives numerical values on risk that may be used for setting exposure limits.
- Permits the comparison of risks due to different chemicals.
- Provides a reasonable basis for setting exposure limits by identifying compounds with high risk.

•Disadvantages
- Extrapolation of data obtained at high doses to low doses relevant for human exposure by means of mathematical models which are not based on cancer biology and pathophysiology.
- Mechanistic and kinetic data are not used for the risk estimation process.
- Expensive and time- consuming lifelong bioassays are required.

The risk extrapolation techniques used are based on several conservative default assumptions, some unsupported by any direct empirical evidence. Conservatism is introduced to ensure maximum protection for those exposed to presumed chemical carcinogens. These assumptions were adopted to achieve some consistency in the application of risk-estimating techniques. The major assumptions are (1) carcinogenic risks are estimated from the data obtained in the most sensitive animal model only using positive responses (data from bioassays that do not show a treatment-related increase in cancer incidence are ignored); (2) linear, nonthreshold dose-risk models are applied; (3) statistical upper confidence limits are used rather than best estimates; and (4) a linear dose–response curve is assumed at low doses.

5.2.4.3. *Mathematical Models Used in High- to Low-Dose Risk Extrapolation*

Mathematical models for quantifying human cancer risk from exposure to carcinogens were first developed in the 1950s. These models were based on the one-hit or the multistage model of chemical carcinogenesis. A probit model was proposed in 1961 for assessing low-dose risk by extrapolation. An improved probit model was introduced in the 1970s by the FDA for computing the level of carcinogenic chemicals permissible in food. This procedure included the suggestion that the dose causing a very low risk (one additional cancer in 1 000 000 exposed people) to be considered as a “virtually safe dose.” Introduction of this mathematical model into

the risk assessment process stimulated the development of a variety of other mathematical models for carcinogen risk assessment. The basis of all the above methods is to apply a mathematical model to the tumor incidence observed in a long-term animal bioassay.

A major problem in dose– response and risk extrapolation is the determination of an appropriate mathematical model to predict effects at hypothetical low levels of exposure. Several models have been developed, and most of them contain analytical functions that appear to fit the experimental dose range quite well and also contain a dose–response functionality. Most models differ in the functions used to estimate response in the very low dose range (i.e., as dose approaches zero). The mathematical models in common use in carcinogen risk assessment are

- The linear model.
- Statistical or distribution models: log–probit, Mantel–Bryan, logit, Weibull.
- Mechanistic models: one-hit (linear), multihit, multistage (Armitage–Doll), linearized multistage, Moolgavkar model.
- Other models: statisticopharmacokinetic, time-to-tumor.

Linear Model. Linear extrapolation involves the intersection of a straight line between the origin (zero dose) and the upper confidence limit of the response at the single, lowest experimental dose. This model is based on the assumption that the increase in tumor incidence by the applied xenobiotic augments an already proceeding process.

Distribution models are based on mathematical functions of presumed population characteristics i.e. on the assumption that every member of a population has a critical dosage (threshold) below which the individual will not respond to the exposure in question. The probit model assumes that log dose–responses have a normal distribution. This model serves as the basis for the Mantel-Bryan risk extrapolation procedure. Other distribution models on which carcinogenicity dose–response models have been based include the logit and Weibull models.

Mechanistic models are based on the currently presumed mechanisms of chemical carcinogenesis. Each model reflects the assumption that a tumor originates from a single cell. The concept underlying the one-hit model is that a tumor can be induced by exposure of DNA to a single molecule of a carcinogen. This model is essentially equivalent to assuming that the dose–response is linear in the low-dose region and as a consequence tends to produce very low calculated “virtually safe doses” compared with the other currently applied models. The multihit model is a generalized version of the one-hit model which assumes that more than one hit is required at the cellular level to initiate carcinogenesis.

The biological justification for the multistage (Armitage–Doll) model is that cancer is assumed to be a multistage process that can be approximated by a series of multiplicative linear functions. It assumes that the effect of a chemical carcinogen occurs in multiple steps and that the effect of each step is additive. The dose–response predicted by this model is approximately linear at low doses and it results in estimates of potential risk that are similar to those of the one-hit model.

The Moolgavkar–Venson–Knudson model attempts a more comprehensive consideration of the biologic processes of cancer formation than the other mechanistic models and may provide a more accurate estimate of human risk by reducing some of the interspecies uncertainty. It is based on a two-stage growth model and considers the birth and death of cells, the effect of cell proliferation on the number of available cells for malignant transformation, and assumes two specific, irreversible, and rate-limiting mutational events to be necessary for cancer development. This model may quantitatively consider major phenomena influencing cancer formation such as cell proliferation, initiation/promotion, genetic predisposition, and synergism/antagonism. Moreover, model parameters may be obtained experimentally. At present, the major limitation for the application of this model is the lack of availability of many of the important biological parameters.

Other Models. The carcinogenicity of many chemicals is based on their biotransformation to reactive metabolites. The statisticopharmacokinetic model arises from consideration of competing metabolic activation and deactivation processes (e.g., detoxification and DNA repair) and estimates the “effective dose”, i.e., the level of reactive metabolites formed and interacting with critical macromeolecules, rather than the administered dose.

A modification of the Probit model relates it to the time at which a tumor is detected. For risk assessement, this time-to-tumor model uses the time to observance (latency) in addition to the proportion of animals bearing tumors at each dose.

5.2.4.4. *Interpretation of Data from Chronic Animal Bioassays*

The results of experimental animal studies may provide a variety of data in addition to the simple indication of the presence or absence of cancer. This information may be useful for the evaluation of potential human cancer risk and includes the number of neoplasms per animal, the number of different types of neoplasms observed, and the number of species affected. The organ or target tissue in which the carcinogenic response occurs is also important, as some rodents have extremely high and variable spontaneous incidence of certain tumor types. Where a chemical increases the number or accelerates the formation of neoplasms which spontaneously occur in high and variable incidence, the response carries less weight than the appearance of tumors with very low spontaneous rates. The time to development of tumor will also give an indication of potency. Further considerations on the interpretation of the obtained data are given in Section Bioassays to Determine the Carcinogenicity of Chemicals in Rodents.

5.2.4.5. *Problems and Uncertainties in Risk Assessment*

The practice of using tumor incidences obtained in long-term animal experiments with high doses (MTD, maximum tolerated dose, see Section Bioassays to Determine the Carcinogenicity of Chemicals in Rodents) for human cancer risk assessment and several of the conservative default assumptions used in the extrapolation processes have become the subject of major criticism. Some of the assumptions made when using experimental animal data for human risk assessment are not directly testable experimentally. Nevertheless, these assumptions are widely used because risk assessment would be difficult or impossible without them [188].

However, recent results on the mechanisms of carcinogenesis have revealed a complex biological process with many variables; a comprehensive consideration of all these parameters by a single, generally applicable mathematical model is not possible. Moreover, observations such as nonlinear toxicokinetics and delineated mechanisms of action for nongenotoxic carcinogens have indicated that a more scientifically based approach to carcinogen risk assessment, likely in a case-by- case examination by expert panels, may be more appropriate for defining actual risk than mathematical modeling. The major points of controversy in risk extrapolation from animal experiments and possible solutions to reduce uncertainties are described in Table 41.

Table 41. Uncertainties in quantitative risk assessment and the application of scientifically based methods for reduction of these uncertainties

Uncertainty	Reduction by
Extrapolation between species	physiologically based toxicokinetic modeling, comparative studies on toxicokinetics
Extrapolation from high to low dose	toxicokinetic modeling, quantitative determination of dose – response for biochemical effects of xenobiotics (e.g., DNA adducts, genetic changes), mechanisms of carcinogenesis
Extrapolation from controlled experimental conditions to variable human situations	none
Conservatism	elucidation of mechanism of actions
Toxicity and increased cell proliferation	dose – response for cell proliferation and cytotoxic effects
Dosimetry	toxicokinetic modeling and physiologically based toxicokinetic modeling
Mixtures	research on mechanisms of interactions
Bioassay-inherent factors in risk assessment	expert judgment

Extrapolation Between Species. In the absence of information on the possible mechanism of carcinogenesis by a particular xenobiotic, the tumor incidences observed in an animal experiment are often assumed to be useful in assessing human risk. However, the differences in the physiology and anatomy of laboratory animals and humans are well recognized. For example, the life span of the laboratory rodent is approximately two years whereas that of humans is approximately seventy years. Cancer appears to develop in rodents over a time scale that is proportional to the life span, and it is generally assumed that this will apply to all chemicals being assessed. The assumption that humans will respond in a similar fashion to laboratory animals is frequently shown to be inappropriate. Mechanistic studies demonstrate that qualitative and quantitative aspects of toxicokinetics and biotransformation, DNA repair and tissue susceptibility, and immune and other defense mechanisms may explain observed differences in the response of laboratory species and humans to exposure to carcinogens [195]. Expert judgement may be required to assess the nature of the end point or the mechanisms of carcinogenic action of the chemical in the experimental animal and to decide whether they are

relevant to man.

Extrapolation from High to Low Dose. The use of very high doses in animal cancer bioassays and the required extrapolation form high to low dose are a major point of critisim. Most cancer bioassays are performed in relatively small groups of rodents (between 50 and 100) and with administration of high doses. The highest dose selected is usually the “maximum tolerated dose” (MTD) in order to have maximum sensitivity and to ensure that the results obtained do not overlook a carcinogenic response simply because the dose was too low.

The following information that is obtained in bioassays with high doses in animals:

- Definite identification of compounds with carcinogenic activity in the species.
- Information about relative potency of different chemicals.
- Information about the carcinogenic activity of the test chemical, when administered alone.
- Characterization of tumor types, target organs, and presence or absence of dose–response relationship, which permit comparison of different chemicals and help in establishing structure–activity relationships to improve predictive capabilities.
- Information on the lack of carcinogenicity of many chemicals to assist priority setting in public health.

However, they also have the following limitations:

- No direct information about effects at doses lower than dose studied.
- No information on the mechanism of carcinogenicity.
- No information on the effects of the test compound, when administered together with other chemicals (synergistic/antagonistic effects).
- Use of high doses that may cause unspecific toxic effects contributing to carcinogenicity.
- Acute toxic effects that may prohibit long-term administration of specific chemicals in sufficiently high doses to cause tumors.

Human exposure to carcinogens usually occurs at doses several orders of magnitude lower than those used in the experiment. Clearly, the possible shape of the dose–response relationship is vitally important in establishing the likely effects at doses substantially below those in the observable range. Mathematical models used for extrapolation only give the upper limits of risk, the real risk may be somewhere between zero and the number calculated. Since the dose–response relationship can not be determined experimentally, application of different mathematical models results in markedly different slopes of the curves in the low-dose range (Fig. 69), and this has major implications for risk assessment and the establishment of ADI values.

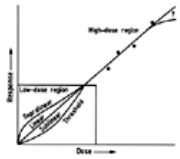

Figure 69. Possible slopes of dose – response curves in the very low dose range below the ability of experimental determination in cancer bioassays

• = Experimental data point

Only one study to determine the actual shape of the dose–response curve at low doses with group sizes large enough to constitute statistical significance was performed, the so-called ED_{01} study using the potent carcinogen 2-acetylaminofluoren as a model compound administered in the diet to mice. Urinary bladder and liver neoplasms were found to be related to 2-acetylaminofluorene administration. The resulting tumor incidences showed that even within one species and with one specific carcinogen, there are major differences in the shape of the dose–response curve at low doses taht depend on the target organ affected. The incidence of bladder tumors at all time points suggested the presence of a threshold for this end point, whereas the dose–response data for the liver tumors was nearly linear at all time points and did not indicate a threshold (Fig. 70).

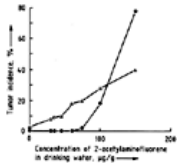

Figure 70. Dose – response data for the induction of liver (Δ) and urinary bladder (•) tumors induced by 2-acetylaminofluorene in mice after 24-month administration in diet

Extrapolation from Controlled Experimental Conditions to Variable Human Situations. In most animals studies, all critical factors are carefully controlled to guarantee the reproducibility of experimental observations. This is in contrast to the sometimes very large differences and time-dependent changes in the environmental exposure and other circumstances of the human population, which cannot be simulated in animal experiments. The complexity of these circumstances (i.e., changing life-style factors or workplaces) would necessitate an assessment of each human individually. Thus, risk assessment normally extrapolates directly from the animal data unless good evidence exists to suggest that an important confounding factor has been introduced. Besides the variable exposure situations, the human populations exposed to chemicals often differ in age, sex, and ethnic background and will certainly have a more heterogeneous genetic makeup than experimental animals. For the majority of carcinogenic risk assessments, individual variability can be taken into account only by the application of safety factors to compensate for these uncertainties.

Dosimetry. The problems of estimating environmental and/or occupational exposure to potential carcinogens have been described above. There are equally difficult problems in extrapolating the doses applied in animal experiments to those which might be encountered by humans due to environmental or occupational exposure and to distinguish between administered dose and effective dose.

Doses can be extrapolated from animal experiments to humans by five methods:

1. Expression of dose as a function of body weight (mg/kg or mmol/kg).
2. Expression of dose as a concentration in food or water (usually parts per million, i.e., µg/g or µL/L).
3. Expression of dose as a concentration in inhaled air (usually parts per million, i.e., µL/L).
4. Correction of dose for surface area. This is achieved in a complicated process by first raising the body weight of the laboratory animal to the power of 2/3 or 3/4. The correction converts doses that are expressed as concentrations in air or food to milligrams per kilogram of body weight in the laboratory species. The dose is then corrected for body surface area and converted back to concentrations in air or food by using appropriate conversion factors for humans. Corrections using surface area are based on the observation that metabolic rate is proportional to body surface area. For acute toxicity, this method may have some merit, but the rate of biotransformation may differently affect the potency of a carcinogen, depending on the role of biotransformation in detoxification or activation. Using correction factors for surface area provides data suggesting that humans may be more susceptible to potential carcinogens than laboratory animals compared to the data obtained by extrapolation on the basis of body weight.
5. Extrapolation assuming that the tissue dose is the primary determinant of carcinogenic response. This approach requires a study of the pathways of biotransformation and the kinetics of the pathway that generates the ultimate electrophile. In vitro studies may be necessary to obtain appropriate human data.

Further problems occur if the experiment is carried out by a protocol that is completely dissimilar from the human experience, for example, if animal exposure is carried out for a lifetime and human exposure is for a shorter period or if the animal experiment uses an exposure route that is not relevant for humans. As a general rule for chemical carcinogens, where the target tissue dose is presumably the most important determinant of the carcinogenic potency, the total body burden must be computed from the various routes of exposure to assess overall dose.

Mixtures. Cancer bioassays in animals are most often performed with a single chemical of definite high purity; exposure of experimental animals to other chemicals or other confounding factors such as vector-based disease is carefully avoided. In contrast, humans are continuously exposed to mixtures of chemicals and other agents. Thus, even when animals or humans are exposed to two or more chemicals, further complications of data interpretation are introduced. In general, the response of the animal and the human will depend on whether the chemicals' activities are interactive or not and whether the effects of the chemicals are similar or not. If the chemicals are not interactive and the response is similar, an additive effect may be assumed to occur. If they are not interactive, the overall response may be less than additive. In addition, a variety of metabolic or infectious diseases in the course of a human's life may have major impact on the development of a certain tumor due to exposure to the chemical under question. It is not feasible to simulate these real-life factors in the experimental situation [196].

Bioassay-Inherent Factors in Risk Assessment. A variety of biological factors exist that influence the assessment of carcinogenic hazard. Among these are the quality of the experiment, including the quality of the pathology, environmental control of animal facilities, estimation and standardization of chemical administration, purity of the chemical administered, and many other factors that together make up compliance with good laboratory practice. A further problem may be encountered if differences occur in outcome from different experiments. In the majority of cases, some aspect of an exceptional experiment may explain why the response was different (e.g., use of a different strain of animal). Since only positive data are used and negative data are ignored for carcinogenic risk assessment purpose to provide a conservative estimate of risk, this factor may also contribute to an overestimation of risk.

Conservatism and the Mechanisms of Chemical Carcinogenesis. To provide maximum protection, a conservative approach to risk assessment is preferred. This approach assumes that a single molecule of a carcinogen can interact with DNA and produce cancer; therefore, there can be no threshold for chemical carcinogenesis. Moreover, selection of the most sensitive response, irrespective of mechanism of action, and the assumption that the dose–response is linear at low doses may overestimate risk. This overestimation may be due to the unknown degree of conservatism at each step and to the amplification of previous bias in the assumption by the next step. The magnitude of overstatement of risk by these conservative assumptions is unknown, but is claimed to amount to several orders of magnitude.

Role of Toxicity and Increased Cell Proliferation. Critics point out that the high doses used in rodent carcinogenicity studies may have nonspecific effects such as an increase in the rate of cell proliferation due to cytotoxic effects [197]. These effects can be unique to high doses, and mitogenesis may itself be mutagenic in numerous ways, either by errors in replication or by conversion of endogenous DNA damage (e.g., by oxidative processes) and exogenous DNA damage to mutations before repair can occur. Moreover, sustained increases in the rate of cell proliferation may also yield secondary mutational events and could be important in the promotional phase of carcinogenicity by increasing the clonal expansion of initiated cells and thus increasing the chance that multiple critical mutational events will occur [198]. Thus, cell division will increase the chance of tumor formation. In this case, the tumor incidence for the same chemical applied at much lower doses is likely to be much lower than a linear model would predict and may even be zero [165, 199, 200]. Therefore, it would be important to add methods for determining cell division to animal cancer bioassays and apply the obtained results to estimate low-dose risks more adequately. Cell proliferation has been implicated as a major contributor to the cancerogenicity of several chemicals such as phenobarbital, 1,4-dichlorobenzene, *d*-limonene, and peroxisome proliferators.

Selection of Mathematical Model. The various models used in risk extrapolation fit the experimentally observed data well; however, they predict widely differing potential risks at low doses. Depending on the data used in the calculation, the predicted risks may differ by several orders of magnitude. These differences are inherent in the application of mathematical models, since only two or three doses are used in the experimental dose groups.

5.3. Future Contributions of Scientifically Based Procedures to Risk Assessment and Qualitative Risk Assessment for Carcinogens

Since quantitative cancer risk assessment can neither claim to be a scientific basis for development of historical models nor attempt to incorporate the large amount of scientific data on the mechanism of carcinogenesis into the assessment of risk, qualitative approaches considering all relevant data may be the best available solution for risk assessment. This approach, referred to a “weight-of-evidence determination of risk” is increasingly emphasized by regulatory authorities. The weight-of-evidence approach includes critical evaluation of the animal bioassay and all other available information on adverse effects of the chemical together with biotransformation, toxicokinetics, and expert judgment [201].

The criticims outlined above have demonstrated the need for further refinement of the extrapolation procedures by toxicokinetics and studies on the mechanisms of tumor formation. The role of animal experiments in predicting the potential risk of human carcinogen exposure will likely remain an important step of the risk assessement process despite the obvious limitations of this approach [197, 202]. However, the data obtained with animals must be interpreted with caution and in light of other data, both quantitative and qualitative, on the adverse effects of the compound (Table 42).

Table 42. Descriptive dimensions proposed as a framework to facilitate the use of mechanistic data in evaluation of carcinogenic risk to humans

Data set	Example of information required
Evidence of genotoxicity	DNA adduct formation, mutagenicity, bioactivation
Evidence of effects on the expression of genes relevant to the process of carcinogenesis	alterations of the structure or quantity of product of a proto-oncogene or suppressor gene
Evidence for effects on cell behavior	mitogenesis, cell proliferation, hyperplasia
Evidence of time and dose – response relationships and interactions	initiation, promotion, progression

Application of Toxicokinetic Models in Risk Assessment. Information on the way in which a chemical is absorbed, biotransformed, and excreted may be critical in extrapolating the relevance of the results obtained in experimental animals to humans [130, 203]. For the purposes of risk extrapolation, it is generally assumed that the administered dose is proportional to the effective dose. However, many chemicals are known to be carcinogenic only after they have been activated to reactive electrophiles by enzymatic reactions. The amount of reactive metabolite formed might not be directly related to dose because saturable enzymatic processes are involved in both bioactivation and detoxication. The extent of bioactivation and detoxication may therefore be highly dose dependent, and the relationship between administered dose and target dose may not be linear at all administered doses. For example, when high-affinity, low-capacity enzymes catalyze the detoxication of a xenobiotic and low-affinity, high-capacity enzyme catalyze its bioactivation, the amount of reactive metabolite formed is likely highly dependent on the administered dose (i.e., a tenfold higher administered dose may result in a 100-fold higher effective target dose). Moreover, there may be depletion of cosubstrates required by the enzymes catalyzing bioactivation or detoxication. After the reactive metabolite is formed, it is often deactivated by a second enzyme, such as epoxide hydrolase or glutathione *S*-transferase. These enzymes can also be saturated. The reactive metabolites that are not destroyed by these detoxication pathways may bind to DNA. The metabolites bound to DNA can be removed by saturable DNA repair systems. These effects are referred to as nonlinear toxicokinetics. Nonlinear kinetics are also seen with chemicals inhibiting or inducing drug-metabolizing enzymes. Application of a carcinogen at high doses may result in the induction of enzymes catalyzing its bioactivation and may leave detoxifying enzymes unaffected. Thus, the expected steady-state concentrations of reactive intermediates present in the cell and capable of binding to DNA are expected to be disproportionately higher after application of high doses which induce biotransformation enzymes than after low doses which leave the levels of biotransformation enzymes unaffected. Moreover, the metabolism of a xenobiotic may be changed as a consequence of the effects of long-term administration of the chemical [204].

To estimate the dose-dependent relationship between administered dose and the effective dose, toxicokinetic models incorporating saturable processes have been applied. The examples in Figure 71 show the theoretically derived relationships between administered dose and effective dose for the same chemical in animals by varying the kinetic parameters for bioactivation, detoxication, and DNA repair [205]: saturation of enzymatic bioactivation (A), saturation of detoxication and activation (B), and saturation of activation, detoxication, and DNA repair (C).

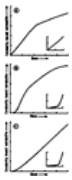

Figure 71. Possible relationships between administered dose and effective dose for the same data set including correction for nonlinear toxicokinetics

A) Saturation of enzymatic bioactivation; B) Saturation of detoxification and activation; C) Saturation of activation, detoxification, and DNA repair

The shapes of the curves in the low-dose range are shown in detail in the small sections.

The hockey-stick shape of the dose-response curve may remain unnoticed in animal experiments with high dose. The nonlinear correlation between effective dose and administered dose in the low, relevant dose region for human exposure may result in a decrease in the potential risk of exposure when compared to risk estimation based on experiments with high doses. Effective doses may be determined by measuring the amount of DNA adducts formed after administration of a carcinogen based on the assumption that for genotoxic agents the carcinogenic response is related to the extent of DNA adduct formation in target tissues. Unfortunately, nonlinear dose–response curves for effective doses have not yet been observed experimentally; all attempts to determine the dose-dependent formation of DNA adducts of potent carcinogens have shown linear relationships between administered dose and effect [206]. However, the increased sensitivity of analytical instruments currently available offers the opportunity to study the dose–effect curve for a wide range of carcinogens transformed to intermediates with different electrophilic reactivity in the low-dose range by quantifying the dose-dependent concentrations of DNA lesions. Quantitation of xenobiotic–hemoglobin adducts and chromosomal abnormalities in lymphocytes (e.g., siste-chromatid exchange frequency) offers other examples of biological markers that may prove useful in the definition of dose–response curves. The application of these procedures offers the advantage of obtaining quantitative information on a nonstochastic effect thought to be relevant in the mechanisms of carcinogenesis and thus in risk assessment.

As noted in Section Problems and Uncertainties in Risk Assessment, dosimetry may be a major cause of uncertainties in risk extrapolation form animals to humans. Besides saturable enzymatic reactions, the effective dose of a carcinogenic chemical may also be influenced by species differences in absorption, distribution and elimination, which may also be influenced by the dose administered.

To account for such factors, physiologically based pharmacokinetic models are increasingly used in the process of risk assessment. The principal purpose of the application of physiologically based pharmacokinetic models is to predict the concentration of carcinogen at the target site and describe the relationship between administered dose and target dose over a range of concentrations. By application of these models, a more accurate dose extrapolation is possible over broad ranges and may also incorporate nonlinearities in bioactivation and detoxification. Since the principles employed in the development of these models apply across species, the definition of the relationship between administered dose and effective dose and the important rate processes that cause a deviation from linearity permits a more accurate species extrapolation. The application of known physiological parameters may also enable the prediction of target organ concentrations in humans when direct measurements are not possible and may permit the comparison of different routes of application for effective dose. Physiologically based pharmacokinetic models are increasingly used to support quantitative risk assessments. Due to the risk estimates based on effective dose rather than external dose levels and the consideration of potential nonlinear relationships, an overestimation of risk obtained by linear extrapolations may be overcome and result in more scientifically founded risk assessments.

Mechanisms of Carcinogenicity and the Risk Assessment Process. Research on the molecular effects of particular agents has increased our understanding of the mechanisms of carcinogenicity. A range of biological processes has been implicated in carcinogenesis. Some of these mechanisms may be common to most carcinogens, others may be restricted to particular classes of chemicals or specific circumstances. In spite of the incomplete knowledge on carcinogenesis and a limited understanding of many processes, mechanism.based decisions are increasingly being introduced into risk assessment [207].

The emerging field of molecular toxicology may hold promise for the future of human risk assessment, and the application of methods developed for molecular biology may also offer more accurate ways of assessing human risk. The most promising application of mechanistic results for risk assessment are non-genotoxic carcinogens [208]. A number of xenobiotics are carcinogenic in animals but do not cause any detectable mutagenicity in in vitro studies. In contrast to other chemical carcinogens, which affect multiple organs or both sexes of both rat and mouse, many nongenotoxic carcinogens affect only a single organ in a single sex of a single species [133]. There are numerous examples of xenobiotics or treatment regimens in which production of tumors in animal experiments unlikely involves interaction of the xenobiotic or its metabolite with DNA. Examples include the induction of subcutaneous sarcomas by the repeated injection of glucose or saline solutions, the induction of skin cancer after chronic skin damage, and the induction of bladder cancer by the implantation of solid materials [209]. Recent experimental evidence also suggests that a number of xenobiotics that do not damage DNA may induce cancer in specific organs after systemic administration. Examples include chemicals binding to the circulating protein α_{2u}-globulin [210], peroxisome proliferators, and several structurally unrelated compounds interacting with receptors [211] (Table 43).

Table 43. Examples of nongenotoxic carcinogens and Their presumed mechanisms of action

Class of nongenotoxic carcinogen	Typical examples	Presumed mechanism of action
Peroxisome proliferators	diethylhexylphthalate	receptor-mediated increase in gene transcription, oxidative stress
α_{2u}-Globulin-binding agents	unleaded gasoline, limonene, decalin	regenerative cell proliferation due to cytotoxicity
Phenobarbital		mitogenic activity
Chlorinated dioxins	2,3,7,8-tetrachlorodibenzodioxin	receptor-mediated increase in gene transcription
Hormones	17- α-ethynylestradiol	receptor-mediated increase in gene transcription

The terms “genotoxic” and “nongenotoxic” were defined by Butterworth [133]: “A genotoxic agent is one for which a primary biological activity of the chemical or a metabolite is alteration of the information encoded in the DNA. These can be point mutations, insertions, deletions or changes in chromosome structure or number. Chemicals exhibiting such activity can usually be identified by assays that measure reactivity with the DNA, induction of mutations, induction of DNA repair or cytogenetic effects. Nongenotoxic chemicals are those that lack genotoxicity as a primary biological activity. While these agents may yield genotoxic events as a secondary result of other induced toxicity, such as forced cellular growth, their primary action does not involve reactivity with the DNA.”

Xenobiotics such as unleaded gasoline, limonene, and 1,4-dichlorobenzene induce renal cancer in male rats, but not in female rats or in either sex of mice [212]. Metabolites of these xenobiotics bind to the circulating protein α_{2u}-globulin, whose synthesis in the rat is under the control of androgens. The modified α_{2u}-globulin is concentrated in the kidney proximal tubules and is not degradable by the processes responsible for degradation of unmodified α_{2u}-globulin. Thus, accumulation of the modified protein causes cytotoxicity and regenerative cell proliferation, which has been implicated as the cause of tumor formation in the kidney.

Tumors in organs whose function is regulated by the endocrine system are often observed after hormonal therapies in humans and hormone treatment in animals and, due to the absence of genotoxicity of most hormones, nongenotoxic processes involving receptor-mediated transcription and activation or repression of specific genes have been implicated in hormonal carcinogenesis [213].

For both of the above cited examples, threshold mechanisms may be postulated. Kidney cancer induced by α_{2u}-globulin-binding agents requires cytotoxicity and cell death. Nontoxic low concentrations of these xenobiotics applied may not be tumorigenic. For receptor-mediated processes, a disproportional relationship between receptor occupancy and hormonally mediated cancer is also considered likely [214-216]. Hence, any assessment of carcinogenic risk of these chemicals must take into account the mechanisms by which they produce their effects.

The use of molecular toxicology in risk assessment is also advocated by the International Agency for Research on Cancer [207]. A monograph on the identification of cancer risks by mechanistic investigations suggested that cancer risks by mechanistic investigations suggests that “when available data on mechanisms are thought to be relevant to evaluation of the carcinogenic risk of an agent to humans, they should be used in making the overall evaluation, together with the combined evidence for animal and/or human carcinogenicity.” The consensus report further states that no definite guidelines for the inclusion of mechanistic data in the evaluation of carcinogens can be elaborated. However, a range of options are available: “First, information concerning mechanisms of action may confirm a particular level of carcinogen classification as indicated on the basis of epidemiological and/or animal carcinogenicity data. Second, for a particular agent, strong evidence for a mechanism of action that is relevant to carcinogenicity in humans could justify ‘upgrading’ its overall evaluation. Third, an overall evaluation of human cancer hazard on the basis of animal carcinogenicity data could be downgraded by strong evidence that the mechanism responsible for tumor growth in experimental animals is not relevant to humans. In keeping with the goal of public health, priority must be given to the demonstration that the mechanism is irrelevant to humans.”

5.4. Risk Assessment for Teratogens

The timing of exposure and patterns of dose–response from animal studies have important implications for extrapolating animal data to humans. A wide spectrum of end points can be produced, even under the controlled conditions of timing and exposure that can be achieved in animal studies. In some cases, the spectrum includes a continuum of responses, with depressed birth weight or functional impairment occurring at low doses, birth defects at intermediate doses, and lethality at high doses. Less commonly, birth defects alone or lethality alone are produced. Therefore, in estimating human risk, all exposure-specific adverse outcomes must be taken into consideration, and not just birth defects.

A similar response pattern is observed in humans exposed to developmental toxicants. The spectrum of responses is determined by the time and duration of exposure, magnitude of exposure, and interindividual differences in sensitivity. Hence, manifestations of developmental toxicity can not be expected to be identical across species; that is, an animal model can not be expected to forecast exactly the human response to a given exposure. For instance, an agent that induces cleft palate in the mouse may elevate the frequency of spontaneous abortion or retard growth in humans. However, any manifestation of exposure-related developmental toxicity in animal studies can be regarded as indicative of a spectrum of response in humans.

Epidemiological data suggest that the majority of human embryos with chromosomal and/or morphologic abnormalities are lost through early miscarriage and that relatively few survive to term. Consequently, determination of malformations or growth retardation at the time of birth alone (malformations, stillbirths, low birth weight) is likely to result in a substantial underestimate of the true risk, since the onset of embryolethality would be missed.

The sensitivity (ability to detect a true positive response in humans) and specificity (ability to detect a true negative response in humans) of laboratory animal studies have been evaluated [217]. Of 38 compounds with demonstrated or suspected teratogenic activity in humans, all but one (tobramycin, which causes otological

deficits in humans) tested positive in at least one animal species. Approximately 80% of the compounds were positive in multiple species. A positive response was elicited to 85% in the mouse, 80% in the rat, 60% in the rabbit, 45% in the hamster, and 30% in the monkey. These findings indicate that the usual laboratory animal species are highly sensitive for detecting human teratogens.

In contrast to the high sensitivity, laboratory animals show low specificity for predicting human teratogenesis; of 165 test compounds with no evidence of teratogenic activity in humans 65 (41%) were positive in more than one animal species. The high percentage of false positive results (i.e., compounds with no evidence for human teratogenicity inducing malformations in animals may be in part accounted for by the high doses usually applied in the animal tests. Also, human studies determine effects from the time of birth onward, which may result in underestimation of the true risk.

[Top of Page]

References

1. D. Henschler, *Trends Pharmacol. Sci. Fest. Suppl* (1985) 26 – 28.
2. W. Dekant, H.-G. Neumann: *Tissue Specific Toxicity: Biochemical Mechanisms*, Academic Press, London 1992.
3. D. O. Chanter, R. Heywood, *Toxicol. Lett.* **10** (1982) 303 – 307. Links
4. Y. Alarie, *EHP Environ. Health Perspect.* **42** (1981) 9 – 13.
5. H. Druckrey, D. Schmähl, W. Dischler, A. Schildbach, *Naturwissenschaften* **49** (1962) 217 – 228.
6. P. H. Bach, C. P. Ketley, I. Ahmed, M. Dixit, *Food Chem. Toxicol.* **24** (1986) 775 – 779. Links
7. P. D. Anderson, L. J. Weber, *Toxicol. Appl. Pharmacol.* **33** (1975) 471 – 483. Links
8. W. W. Weber, D. W. Hein, *Pharmacol. Rev.* **37** (1985) 25 – 79. Links
9. T. Green, M. S. Prout, *Toxicol. Appl. Pharmacol.* **79** (1985) 401 – 411. Links
10. D. A. Smith, *Drug Metab. Rev.* **23** (1991) 355 – 373. Links
11. J. H. Lin, *Drug Metab. Dispos.* **23** (1995) 1008 – 1021. Links
12. D. F. Lewis, C. Ioannides, D. V. Parke, *Environ. Health Perspect.* **106** (1998) 633 – 41. Links
13. M. R. Juchau, H. Boutelet-Bochan, Y. Huang, *Drug Metab. Rev.* **30** (1998) 541 – 68. Links
14. E. J. Calabrese, *J. Pharm. Sci.* **75** (1986) 1041 – 1067. Links
15. M. VandenBerg, J. DeJongh, H. Poiger, J. R. Olson, *CRC Crit. Rev. Toxicol.* **24** (1994) 1 – 74.
16. R. C. Lind, A. J. Gandolfi, P. d. I. M. Hall, *Anesthesiology* **70** (1989) 649 – 653. Links
17. R. A. van Dyke, J. A. Gandolfi, *Drug Metab. Dispos.* **2** (1974) 469 – 476. Links
18. I. G. Sipes, A. J. Gandolfi, L. R. Pohl, G. Krishna, B. R. Brown, *J. Pharmacol. Exp. Therap.* **214** (1980) 716 – 720.
19. A. Turturro, R. Hart, *Exp. Gerontol.* **27** (1992) 583 – 592. Links
20. K. O. Lindros, Y. A. Cai, K. E. Penttila, *Hepatology (N.Y.)* **12** (1990) 1092 – 1097.
21. S. J. Singer, G. L. Nicolson, *Science* **175** (1972) 720 – 731. Links
22. M. Eisenberg, S. M. McLaughlin, *Residue Rev.* **31** (1976) 1 – 8.
23. R. C. Scott, J. D. Ramsey, R. I. Ward, M. A. Thompson, C. Rhodes, *Br. J. Dermatol.* **115** (1986) 47 – 48.
24. P. H. Dugard, *Food Chem. Toxicol.* **24** (1986) 749 – 753. Links
25. T. R. Bates, M. Gibaldi in J. Swarbick (ed.): *Current Concepts in the Pharmaceutical Sciences: Biopharmaceutics*, **vol. 1**, Lea and Febiger, Philadelphia 1970.
26. A. R. Brody, *EHP Environ. Health Perspect.* **100** (1993) 21 – 30.
27. J.-P. Tillement, E. Lindenlaub: *Protein Binding and Drug Transport*, F. K. Schattauer Verlag, Stuttgart 1986.
28. M. P. Waalkes, P. L. Goering, *Chem. Res. Toxicol.* **3** (1990) 281 – 288. Links
29. R. A. Goyer, *EHP Environ. Health Perspect.* **100** (1993) 177 – 187.
30. F. P. Guengerich, T. L. Macdonald, *Acc. Chem. Res.* **17** (1984) 9 – 16. Links
31. G. S. Yost, A. R. Buckpitt, R. A. Roth, T. L. McLemore, *Toxicol. Appl. Pharmacol.* **101** (1989) 179 – 195. Links
32. F. M. Guengerich, *Annu. Rev. Pharmacol. Toxicol.* **29** (1989) 241 – 264. Links
33. F. P. Guengerich, *J. Biol. Chem.* **266** (1991) 10019 – 10022. Links
34. F. P. Guengerich, *Life Sci.* **50** (1992) 1471 – 1478. Links
35. F. P. Guengerich, *FASEB J.* **6** (1992) 745 – 748. Links
36. L. S. Kaminsky, M. J. Fasco, *CRC Crit. Rev. Toxicol.* **21** (1992) 407 – 422.
37. A. J. Paine, *Hum. Exper. Toxicol.* **14** (1995) 1 – 7.
38. F. P. Guengerich, *Chem. Res. Toxicol.* **14** (2001) 611 – 50. Links
39. F. P. Guengerich, *Drug Metab. Rev.* **34** (2002) 7 – 15. Links
40. X. Ding, L. S. Kaminsky, *Annu. Rev. Pharmacol. Toxicol. 43* (2003) 149 – 73. Links
41. D. R. Koop, *FASEB J.* **6** (1992) 724 – 730. Links
42. P. B. Danielson, *Curr. Drug Metab.* **3** (2002) 561 – 597.
43. D. M. Ziegler, *Drug Metab. Rev. 34* (2002) 503 – 511. Links
44. J. R. Cashman, *Curr Opin. Drug Discov. Dev.* **6** (2003) 486 – 93.
45. R. E. Tynes, R. M. Philpot, *Mol. Pharmacol.* **31** (1991) 569 – 574.
46. T. E. Eling, D. C. Thompson, G. L. Foureman, M. F. Hughes, *Annu. Rev. Pharmacol. Toxicol.* **30** (1990) 1 – 45. Links
47. B. J. Smith, J. F. Curtis, T. E. Eling, *Chem. Biol. Interact.* **79** (1991) 245 – 264. Links
48. C. Vogel, *Curr. Drug Metab.* **1** (2000) 391 – 404. Links
49. W. L. Smith, L. J. Marnett, *Biochim. Biophys. Acta.* **1083** (1991) 1 – 17. Links
50. B. H. Bach, J. W. Bridges, *CRC Crit. Rev. Toxicol.* **15** (1985) 217 – 329.
51. R. J. Duescher, A. A. Elfarra, *J. Biol. Chem.* **267** (1992) 19859 – 19865. Links
52. P. D. Josephy, T. E. Eling, R. P. Mason, *Mol. Pharmacol.* **23** (1983) 461 – 466. Links
53. A. J. Kettle, C. C. Winterbourn, *J. Biol. Chem.* **267** (1992) 8319 – 8324. Links
54. H. Thomas, F. Oesch, *ISI Atlas Sci.: Biochemistry* **1** (1988) 287 – 291.
55. S. A. Sheweita, A. K. Tilmisany, *Curr. Drug Metab.* **4** (2003) 45 – 58.
56. B. Burchell, W. H. Coughtrie, *Pharmacol. Ther.* **43** (1989) 261 – 289. Links
57. T. R. Tephly, B. Burchell, *Trends Pharmacol. Sci.* **11** (1990) 276 – 279. Links
58. R. H. Tukey, C. P. Strassburg, *Annu. Rev. Pharmacol. Toxicol.* **40** (2000) 581 – 616. Links
59. J. K. Ritter, *Chem. Biol. Interact.* **129** (2000) 171 – 193. Links
60. K. W. Bock, B. S. Bock-Hennig, G. Fischer, L. W. D. Ullrich in H. Greim et al. (eds.): *Biochemical Basis of Chemical Carcinogenesis*, **vol. 1**, Raven Press, New York 1984, pp. 13 – 22.
61. G. J. Mulder: *Sulfation of Drugs and Related Compounds*, CRC Press, Boca Raton, Fla., 1981.
62. M. W. Duffel, A. D. Marshal, P. McPhie, V. Sharma, W. B. Jakoby, *Drug Metab. Rev.* **33** (2001) 369 – 395. Links
63. E. Y. Krynetski, W. E. Evans, *Pharm Res.* **16** (1999) 342 – 349.
64. D. A. Evans, *Pharmacol. Ther.* **42** (1989) 157 – 234. Links
65. P. Meisel, *Pharmacogenomics* **3** (2002) 349 – 366. Links
66. I. M. Arias, W. B. Jakoby: *Glutathione: Metabolism and Function*, Raven Press, New York 1976.
67. B. Ketterer, D. J. Meyer, A. G. Clark in H. Sies, B. Ketterer (eds.): *Glutathione Conjugation: Mechanisms and Biological Significance*, Academic Press, London 1988, pp. 73 – 135.

68. R. Morgenstern, G. Lundqvist, G. Andersson, L. Balk, J. W. DePierre, *Biochem. Pharmacol.* **33** (1984) 3609 – 3614. Links
69. J. L. Cmarik, P. B. Inskeep, M. J. Meredith, D. J. Meyer, B. Ketterer, F. P. Guengerich, *Cancer Res.* **50** (1990) 2747 – 2752. Links
70. J. Caldwell, W. B. Jakoby: *Biological Basis of Detoxification*, Academic Press, New York 1983
71. M. W. Anders: *Bioactivation of Foreign Compounds*, Academic Press, Orlando 1985.
72. M. W. Anders in I. M. Arias, W. B. Jakoby, H. Popper, D. Schachter, D. A. Shafritz (eds.): *The Liver: Biology and Pathology*, 2nd ed., Raven Press, New York, 1988, pp. 389 – 400.
73. J. G. Bessems, N. P. Vermeulen, *CRC Crit. Rev. Toxicol.* **31** (2001) 55 – 138.
74. G. M. B. StClair, V. Amarnath, M. A. Moody, D. C. Anthony, C. W. Anderson, D. G. Graham, *Chem. Res. Toxicol.* **1** (1988) 179 – 185. Links
75. F. P. Guengerich, D. C. Liebler, *CRC Crit. Rev. Toxicol.* **14** (1985) 259 – 307.
76. D. C. Liebler, F. P. Guengerich, *Biochemistry* **22** (1983) 5482 – 5489. Links
77. L. R. Pohl, R. V. Branchflower, R. J. Highet, J. L. Martin, D. S. Nunn, T. J. Monks, J. W. George, J. A. Hinson, *Drug Metab. Dispos.* **9** (1981) 334 – 339. Links
78. M. W. Anders, W. Dekant: *Conjugation-dependent Carcinogenicity and Toxicity of Foreign Compounds*, Academic Press, San Diego, 1994.
79. T. J. Monks, M. W. Anders, W. Dekant, J. L. Stevens, S. S. Lau, P. J. van Bladeren, *Toxicol. Appl. Pharmacol.* **106** (1990) 1 – 19. Links
80. M. W. Anders, W. Dekant, S. Vamvakas, *Xenobiotica* **22** (1992) 1135 – 1145. Links
81. W. Dekant, S. Vamvakas, M. W. Anders, in W. Dekant, H.-G. Neumann (eds.): *Tissue Specific Toxicity: Biochemical Mechanisms*, Academic Press, London, 1992, pp. 163 – 194.
82. S. D. Aust, C. F. Chignell, T. M. Bray, B. Kalyanaraman, R. P. Mason, *Toxicol. Appl. Pharmacol.* **120** (1993) 168 – 178. Links
83. B. D. Goldstein, B. Czerniecki, G. Witz, *EHP Environ. Health Perspect.* **81** (1989) 55 – 57.
84. B. D. Goldstein, G. Witz, *Free Radicals Res. Commun.* **11** (1990) 3 – 10.
85. J. P. Kehrer, B. T. Mossman, A. Sevanian, M. A. Trush, M. T. Smith, *Toxicol. Appl. Pharmacol.* **95** (1988) 349 – 362. Links
86. P. A. Cerutti, B. F. Trump, *Cancer Cells* **3** (1991) 1 – 7.
87. E. Chacon, D. Acosta, *Toxicol. Appl. Pharmacol.* **107** (1991) 117 – 128. Links
88. R. H. Burdon, V. Gill, C. Rice-Evans, *Free Radicals Res. Commun.* **11** (1990) 65 – 76.
89. R. W. Pero, G. C. Roush, M. M. Markowitz, D. G. Miller, *Cancer Detect. Prev.* **14** (1990) 551 – 561.
90. W. B. Jakoby, D. M. Ziegler, *J. Biol. Chem.* **265** (1990) 20715 – 20719. Links
91. H. Sies, *Naturwissenschaften* **76** (1989) 57 – 64. Links
92. D. M. Ziegler in W. B. Jacoby (ed.): *Enzymic Basis of Detoxification*, **vol. I**, Academic Press, New York 1980, pp. 201 – 227.
93. E. Boyland, L. F. Chasseaud, *Adv. Enzymol. Relat. Areas Mol. Biol.* **32** (1969) 173. Links
94. T. A. Fjellstedt, R. H. Allen, B. K. Duncan, W. B. Jakoby, *J. Biol. Chem.* **248** (1973) 3702 – 3707. Links
95. R. C. James, R. D. Harbison, *Biochem. Pharmacol.* **31** (1982) 1829 – 1835. Links
96. L. I. McLellan, C. R. Wolf, J. D. Hayes, *Biochem. J.* **258** (1989) 87 – 93. Links
97. J. A. Hinson, D. W. Roberts, *Annu. Rev. Pharmacol. Toxicol.* **32** (1992) 471 – 510. Links
98. S. D. Nelson, P. G. Pearson, *Annu. Rev. Pharmacol. Toxicol.* **30** (1990) 169 – 195. Links
99. L. R. Pohl, *Chem. Res. Toxicol.* **6** (1993) 786 – 793. Links
100. E. R. Stadtman, C. N. Oliver, *J. Biol. Chem.* **266** (1991) 2005 – 2008. Links
101. E. R. Stadtman, *Science* **257** (1992) 1220 – 1224. Links
102. H. Esterbauer, H. Zollner, R. J. Schaur, *ISI Atlas Sci.: Biochemistry* **1** (1988) 311 – 317.
103. H. Kappus, *Biochem. Pharmacol.* **35** (1986) 1 – 6. Links
104. J. L. Farber, *Chem. Res. Toxicol.* **3** (1990) 503 – 508. Links
105. J. P. Kehrer, D. P. Jones, J. J. Lemasters, J. L. Farber, H. Jaeschke, *Toxicol. Appl. Pharmacol.* **106** (1990) 165 – 178. Links
106. K. Hemminki, *Carcinogenesis* **14** (1993) 2007 – 2012. Links
107. B. N. Ames, *Mutat. Res.* **214** (1989) 41 – 46. Links
108. R. Adelman, R. L. Saul, B. N. Ames, *Proc. Natl. Acad. Sci. USA* **85** (1988) 2706 – 2708. Links
109. K. W. Bock, H.-P. Lipp, B. S. Bock-Hennig, *Xenobiotica* **20** (1990) 1101 – 1111. Links
110. D. R. Koop, D. J. Tierney, *Bioassays*. **12** (1990) 429 – 435.
111. A. H. Conney, *Annu. Rev. Pharmacol. Toxicol.* **43** (2003) 1 – 30. Links
112. L. Wu, J. P. J. Whitlock, *Proc. Natl. Acad. Sci. USA* **89** (1992) 4811 – 4815. Links
113. J. P. J. Whitlock, *Chem. Res. Toxicol.* **6** (1993) 754 – 763. Links
114. B. Testa, P. Jenner, *Drug Metab. Rev.* **12** (1981) 1 – 117. Links
115. B. Testa, *Xenobiotica* **20** (1990) 1129 – 1137. Links
116. J. R. Halpert, F. P. Guengerich, J. R. Bend, M. A. Correia, *Toxicol. Appl. Pharmacol.* **125** (1994) 163 – 175. Links
117. M. Blum, D. M. Grant, W. McBride, M. Heim, U. A. Meyer, *DNA Cell Biol*. **9** (1990) 193 – 203. Links
118. M. E. Andersen, *CRC Crit. Rev. Toxicol.* **11** (1981) 105 – 150.
119. M. W. Anders, W. Dekant in M. W. Anders, W. Dekant, D. Henschler, H. Oberleithner, S. Silbernagl (eds.): *Renal Disposition and Nephrotoxicity of Xenobiotics*, Academic Press, San Diego 1993, p. 155 – 183.
120. R. M. Brenner, F. C. J. Rector: *The Kidney*, W. B. Saunders Co., Philadelphia 1981.
121. E. C. Foulkes, *Proc. Soc. Exp. Biol. Med.* **195** (1990) 1.
122. B. A. Saville, M. R. Gray, Y. K. Tam, *Drug Metab. Rev.* **24** (1992) 49 – 88. Links
123. C. D. Klaassen, J. B. Watkins, *Pharmacol. Rev.* **36** (1984) 1 – 67. Links
124. O. Fardel, L. Payen, A. Courtois, L. Vernhet, V. Lecureur, *Toxicology* **167** (2001) 37 – 46. Links
125. Y. Kato, H. Suzuki, Y. Sugiyama, *Toxicology* **181–182** (2002) 287 – 290.
126. J. G. Filser, *Arch. Toxicol.* **66** (1992) 1 – 10. Links
127. M. E. Andersen, *Toxicol. Lett.* **138** (2003) 9 – 27. Links
128. R. Dixit, J. Riviere, K. Krishnan, M. E. Andersen, *J. Toxicol. Environ. Health. B Crit. Rev.* **6** (2003) 1 – 40. Links
129. M. E. Andersen, M. L. Gargas, H. J. Clewell III, K. M. Seceryn, *Toxicol. Appl. Pharmacol.* **89** (1987) 149 – 157. Links
130. M. E. Andersen, D. Krewski, J. R. Withey, *Cancer Lett.* **69** (1993) 1 – 14. Links
131. R. B. Conolly, M. E. Andersen, *Annu. Rev. Pharmacol. Toxicol.* **31** (1991) 503 – 523. Links
132. S. M. Cohen, L. B. Ellwein, *Science* **249** (1990) 1007 – 1011. Links
133. B. Butterworth, *Mutation Res.* **239** (1990) 117 – 132.
134. Q. Chen, J. Marsh, B. Ames, B. Mossman, *Carcinogenesis* **17** (1996) 2525 – 2527. Links
135. R. A. Roberts, S. Chevalier, S. C. Hasmall, N. H. James, S. C. Cosulich, N. Macdonald, *Toxicology* **181–182** (2002) 167 – 170.
136. M. S. Denison, S. R. Nagy, *Annu. Rev. Pharmacol. Toxicol.* **43** (2003) 309 – 34. Links
137. J. Abel, D. Hohr, H. J. Schurek, *Arch. Toxicol.* **60** (1987) 370 – 375. Links
138. S. Orrenius, P. Nicotera, *Arch. Toxicol. Suppl.* **11** (1987) 11 – 19.
139. S. Orrenius, D. J. McConkey, G. Bellomo, P. Nicotera, *Trends Pharmacol. Sci.* **10** (1989) 281 – 285. Links
140. E. C. Miller, J. A. Miller, *Pharmacol. Rev.* **18** (1966) 805 – 838. Links
141. E. C. Miller, J. A. Miller, *Cancer* **47** (1981) 2327 – 2345.
142. P. A. Cerutti, *Science* **227** (1985) 375 – 381. Links
143. *International Agency for Research on Cancer: Models, Mechanisms and Etiology of Tumour Promotion*, IARC Sci. Publ., Lyon, 1984.

144. J. M. Bishop, *Science* **235** (1987) 305 – 311. Links
145. J. M. Bishop, *Cell* **64** (1991) 235 – 48. Links
146. J. M. Bishop, *Cell* **42** (1985) 23 – 38. Links
147. J. Ashby, R. W. Tennant, *Mutation Res.* **257** (1991) 229 – 306.
148. R. L. Brinster in T. H. Shepard, J. R. Miller, M. Marois (eds.): *Methods for Detection of Environmental Agents that Produce Congenital Defects*, Elsevier New York, 1975, pp. 113 – 124.
149. J. M. Manson, L. D. Wise, in M. O. Amdur, J. Doull, C. D. Klaassen (eds.): *Casarett and Doull's Toxicology. The Basic Science of Poisons*, Pergamon Press New York 1991, pp. 226 – 254.
150. R. P. Bolande in J. G. Wilson, F. C. Fraser (eds.): *Handbook of Teratology*, **vol. 2**, Plenum Press New York 1977, pp. 293 – 328.
151. S. Ivankovic: "Perinatal Carcinogenesis, DHEW Publication no. (NIH) 7a–1063", *Nat. Cancer Inst. Monogr.* **51** (1979) 103 – 116.
152. E. Schlede, U. Mischke, R. Roll, D. Kayser, *Arch. Toxicol.* **66** (1992) 455 – 470. Links
153. M. J. Van den Heuvel, *Human Exptl. Toxicol.* **9** (1990) 369.
154. B. Berner, E. R. Cooper in A. F. Kydonieu, B. Berner (eds.): *Transdermal Delivery of Drugs*, CRC Press, Boca Raton 1987, pp. 107 – 130.
155. J. A. Shatkin, H. S. Brown, *Environ. Res.* **56** (1991) 90 – 108. Links
156. B. Ballantyne, T. Marrs, P. Turner: *General and Applied Toxicology*, Macmillan Press, New York 1993.
157. A. P. Worth, M. Balls: *Alternative (non-animal) Methods for Chemical Testing: Current Status and Future Prospects*, FRAME, Nottingham, UK, 2002.
158. P. A. Botham, D. A. Basketter, T. Maurer, D. Mueller, M. Potokar, W. J. Bontinck, *Food Chem Toxicol.* **29** (1991) 275 – 86. Links
159. G. M. Henningen in D. W. Hobson (ed.): *Dermal and Ocular Toxicology. Fundamentals and Methods*, CRC Press, Boca Raton 1991, pp. 153 – 192.
160. W. E. Ecvam in European Center for the Validation of alternative methods *Alternative (non-animal) Methods for Chemical Testing: Current Status and Future Prospects*, **vol. 30**, Suppl. 1, FRAME, Nottingham, UK 2002, pp. 95 – 102.
161. D. B. Clayson, F. Iverson, R. Mueller, *Teratog. Carcinog. Mutagen* **11** (1991) 279 – 296.
162. J. K. Haseman, A. Lockhart, *Fund. Appl. Toxicol.* **22** (1994) 382 – 391.
163. D. G. Hoel, J. K. Haseman, M. D. Hogan, J. Huff, E. E. McConnell, *Carcinogenesis* **9** (1988) 2045 – 2052. Links
164. B. N. Ames, R. Magaw, L. S. Gold, *Science* **236** (1987) 271 – 80. Links
165. B. N. Ames, L. S. Gold in T. Slaga, B. Butterworth (eds.): *Chemically Induced Cell Proliferation: Implications for Risk Assessment*, Wiley-Liss, New York 1991, pp. 1 – 20.
166. L. S. Gold, T. H. Slone, B. R. Stern, N. B. Manley, B. N. Ames, *Science* **258** (1992) 261 – 265. Links
167. D. J. Waxmann, L. Azaroff, *Biochem. J.* **281** (1992) 577 – 592.
168. B. N. Ames, J. McCann, E. Yamasaki, *Mutation Res.* **31** (1975) 347 – 64.
169. D. M. Maron, B. N. Ames, *Mutat. Res.* **113** (1983) 173 – 215. Links
170. M. Watanabe, M. J. Ishidate, T. Nohmi, *Mutat. Res.* **234** (1990) 337 – 348. Links
171. T. P. Simula, M. J. Glancey, C. R. Wolf, *Carcinogenesis* **14** (1993) 1371 – 1376. Links
172. M. Richold, A. Chaudley, J. Ashby, D. G. Gatehouse, J. Bootman, L. Henderson in D. J. Kirkland (eds.): *UKEMS Sub-Committee on Guidelines for Mutagenicity Testing, report part Revised Basic Mutagenicity Tests, UKEMS Recommended Procedures*, Cambridge University Press, Cambridge 1990, pp. 115 – 141.
173. J. J. McCormick, V. M. Mahler, *Environ. Molec. Mutag.* **14**(S 15) (1989) 105 – 113.
174. G. M. Williams, J. H. Weisburger, *Mutat. Res.* **205** (1988) 79 – 90. Links
175. N. Ozawa, F. P. Guengerich, *Proc. Natl. Acad. Sci. USA* **80** (1983) 5266 – 5270. Links
176. H. Bartsch, A. Barbin, M.-J. Marion, J. Nair, Y. Guichard, *Drug Metab. Rev.* **26** (1994) 349 – 372. Links
177. C. Richter, J.-W. Park, B. N. Ames, *Proc. Natl. Acad. Sci. USA* **85** (1988) 6465 – 6467. Links
178. P. B. Farmer, H.-G. Neumann, D. Henschler, *Arch. Toxicol.* **60** (1987) 251 – 260. Links
179. J. A. Swenberg, N. Fedtke, F. Ciroussel, A. Barbin, H. Bartsch, *Carcinogenesis* **13** (1992) 727 – 729. Links
180. I. A. Blair, *Chem. Res. Toxicol.* **6** (1993) 741 – 747. Links
181. E. Randerath, T. F. Danna, K. Randerath, *Mutat. Res.* **268** (1992) 139 – 153. Links
182. M. D. Shelby, E. Zeiger, *Mutat. Res.* **234** (1990) 257 – 261. Links
183. J. G. Vos, *Arch. Toxicol Suppl.* **4** (1980) 95 – 108.
184. A. D. Dayau, R. F. Hertel, E. Hesseltine, G. Kazautzis, E. Smith, M. T. Van der Venne: *Immunotoxicity of Metals and Immunotoxicology*, Plenum Press, New York 1990.
185. D. M. Maloney, *Environ. Health Perspect.* **101** (1993) 396 – 401.
186. C. C. Harris, *Science* **262** (1993) 1980 – 1981. Links
187. J. V. Rodricks, *Environ. Health Perspect.* **102** (1994) 258 – 264. Links
188. F. R. Johannsen, *CRC Crit. Rev. Toxicol.* **20** (1990) 341 – 367.
189. A. M. Jarabek, W. H. Farland, *Toxicol. Ind. Health* **6** (1990) 199-. Links
190. D. W. Nebert, *Mutat. Res.* **247** (1991) 267 – 281. Links
191. R. T. H. van Welie, R. G. J. M. van Dijck, N. P. E. Vermeulen, N. J. van Sittert, *CRC Crit. Rev. Toxicol.* **22** (1992) 271 – 306.
192. R. W. Tennant, J. Ashby, *Mutat. Res.* **257** (1991) 209 – 227. Links
193. A. C. Beach, R. C. Gupta, *Carcinogenesis* **13** (1992) 1053 – 1074. Links
194. K. Crump, *CRC Crit. Rev. Toxicol.* **32** (2002) 133 – 53.
195. C. C. Travis, *Toxicology* **47** (1987) 3 – 13. Links
196. H. Vainio, M. Sorsa, A. J. McMichael: *Complex Mixtures and Cancer Risk*, International Agency for Research on Cancer (WHO), Lyon 1990.
197. J. A. Swenberg, R. R. Maronpot in (ed.): *Chemically Induced Cell Proliferation: Implications for Risk Assessment*, Wiley-Liss, New York 1991, pp. 245 – 251.
198. S. M. Cohen, L. B. Ellwein, *Cancer Res.* **51** (1991) 6493 – 6505. Links
199. B. N. Ames, L. S. Gold, *Science* **249** (1990) 970 – 1. Links
200. B. N. Ames, L. S. Gold, *Mutat. Res.* **250** (1991) 3 – 16. Links
201. G. M. Gray, J. T. Cohen, J. D. Graham, *Environ. Health Perspect.* **101** (1993) 203 – 208.
202. J. C. Barrett, *Environ. Health Perspect.* **100** (1993) 9 – 20. Links
203. M. E. Andersen, H. J. Clewell, 3rd, M. L. Gargas, F. A. Smith, R. H. Reitz, *Toxicol. Appl. Pharmacol.* **87** (1987) 185 – 205. Links
204. T. Green, *Xenobiotica* **20** (1990) 1233 – 1240. Links
205. D. G. Hoel, N. L. Kaplan, M. W. Anderson, *Science* **219** (1983) 1032 – 1037. Links
206. W. K. Lutz, *Carcinogenesis* **11** (1990) 1243 – 1247. Links
207. H. Vainio, P. N. Magee, D. B. McGregor, A. J. McMichael: Mechanisms of Carcinogenesis in Risk Identification, IARC Scientific Publications No. 116, Lyon 1992.
208. C. Shaw, H. B. Jones, *Trends Pharmacol. Sci.* **15** (1994) 89 – 93. Links
209. P. Grasso, M. Sharratt, A. J. Cohen, *Annu. Rev. Pharmacol. Toxicol.* **31** (1991) 253 – 87. Links
210. J. A. Swenberg, *Environ. Health Perspect.* **101** (1993) 39 – 44. Links
211. L. Poellinger, M. Göttlicher, J.-A. Gustafsson, *Trends Pharmacol. Sci.* **13** (1992) 241 – 245. Links
212. S. J. Borghoff, B. G. Short, J. A. Swenberg, *Annu. Rev. Pharmacol. Toxicol.* **30** (1990) 349 – 67. Links
213. R. M. McClain, *Toxicol. Lett.* **64/65** (1992) 397 – 408.
214. A. M. Tritscher, J. A. Goldstein, C. J. Portier, Z. McCoy, G. C. Clark, G. W. Lucier, *Cancer Res.* **52** (1992) 3436 – 3442. Links
215. C. Portier, A. Tritscher, M. Kohn, C. Sewall, G. Clark, L. Edler, D. Hoel, G. Lucier, *Fund. Appl. Toxicol.* **20** (1993) 48 – 56.
216. J. P. V. Heuvel, G. C. Clark, M. C. Kohn, A. M. Tritscher, W. F. Greenle, G. W. Lucier, D. A. Bell, *Cancer Res.* **54** (1994) 62 – 68.

217. V. H. Frankos, *Fund. Appl. Toxicol.* **5** (1985) 615 – 622.

[Top of Page]

Abbreviations:

Ah-R: arylhydrocarbon receptor

AP: apurinic/apyrimidinic site

APS: adenosine 5′-phosphosulfate

BHK: baby hamster kidney

BIBRA: British Industrial Biological Research Association

CoA: Coenzym A

DDT: 1,1′-(2,2,2-trichloroethylidene)bis-(4-chlorobenzene)

DHHS: U.S. Department of Health and Human Services

DHP: delayed hypersensitive response

ECETOC: European Chemical Industry Ecology and Toxicology Centre

ED: effective dose

ELISA: enzyme-linked immunosorbent assay

FCA: Freund's complete adjuvant

FAD: flavine adenine dinucleotide

GABA: ϒ-aminobutyrate

GC/MS: gas chromatography/mass spectroscopy

GOT: glutamic acid oxalacetic transaminase

GSH: glutathione

GSSG: glutathione disulfide

GST: glutathione *S*-transferase

GTP: guanosine 5′-triophosphate

HGPRT: hypoxanthine – guanine phosphoribosyltransferase

IPCS: International Programme on Chemical Safety

LDH: lactate dehydrogenase

LOAEL: lowest-observed-adverse-effect level

LOEL: lowest-observed-effect level

MIF: migration inhibition factor

mRNA: messenger RNA

MTD: maximum tolerated dose

NADPH: nicotinamide dinucleotide phosphate (H)

NOEL: no-observed-effect-level

NTP: National Toxicology Program

PAPS: 3′-phosphoadenosine-5′-phosphosulfate

PG: prostaglandin

rRNA: ribosomal RNA

SHE: Syrian hamster embryo

SMART: somatic mutation and recombination test

T, or TCDD: 2,3,7,8-tetrachlorodibenzodioxin

TD: tumor dose

TK: thymidine kinase

tRNA: transfer RNA

UDP: uridine diphosphate

UDPG: uridine diphosphate glucose

UDPGA: uridine diphosphate glucuronic acid

UDS: unscheduled DNA synthesis

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Legal Aspects

Jean-Marie Devos, CEFIC, Brussels, Belgium (Chaps. 1, 3)

Ludwig Krämer, Head of Legal Services, DG XI, European Commission, Brussels , Belgium(Chap. 2, Chap. 5 in part)

Gabrielle H. Williamson, LeBoeuf, Lamb, Greene & MacRae, Brussels, Belgium (Chap. 4)

Pascale Kromarek, Elf Aquitaine, Direction Environment, Paris-La-Defense, France (Chap. 5)

Turner T. Smith, Jr., Hunton & Williams, Washington, D.C., United States (Chap. 6)

Muneto Itoda, JCIA, Tokyo, Japan (Chap. 7)

1. Industrial Chemistry and Law . 3
1.1. Law, Society, Chemistry, and Ethics 3
1.2. Policy and Legal Principles Affecting Industrial Chemistry . . 5
1.2.1. Precautionary Principle 5
1.2.2. The Polluter-Pays Principle 7
1.3. Economic Integration and Environmental Protection 7
1.4. Other General Trends of Health, Safety and Environmental Laws Affecting Industrial Chemistry . 8
1.5. Scope of this Keyword 9
2. European Community Legislation Affecting Chemicals 9
2.1. The European Community Framework 9
2.2. Product-Oriented Legislation . . 10
2.3. Legislation on Chemical Installations 13
3. Towards the Adoption of New Civil Liability Rules: A New Challenge for the Chemical Industry 14
3.1. Introduction 14
3.2. General Trends and Debates Affecting Modern Civil Liability Systems 14
3.2.1. New Trends 14
3.2.2. Recognition of Ecological Damage in International and Community Law of Civil Liability 15
3.2.3. International Harmonization in the Context of Liability for Dangerous Activities 16
3.2.4. Actions and Requests by Organizations 17
3.2.5. Financial Security Schemes 17
3.2.6. Essential Requirements 17
3.3. Liability for Defective Products and European Community Law 17
3.3.1. Background and Legal Basis . . . 17
3.3.2. Nature of the Liability Regime . . 18
3.3.3. The Three Options of the Directive 19
3.3.4. Other important Aspects of the Directive 20
3.4. Community-Proposed Directive on Liability for Waste 21
3.4.1. The “Channelling” of Liability . . 21
3.4.2. Definition of the Producer 22
3.4.3. The Concept of Waste 22
3.4.4. Insurance and Compensation Fund 22
3.5. Civil Liability for Damage Caused during the Carriage of Dangerous Goods by Land: The Convention of October 10, 1989 23
3.6. Liability in Connection with the Carriage of Dangerous Substances by Sea (HNS) 24
3.7. The Council of Europe Convention on Compensation for Damage Caused by Dangerous Activities 24
3.7.1. Background 24
3.7.2. Key Features of the Council of Europe Convention 25
3.7.2.1. Liability Regime 25
3.7.2.2. Concept of Operator 25
3.7.2.3. Incident 25
3.7.2.4. Concept of Dangerous Activity . . 25
3.7.2.5. Concept of Damage 26
3.7.2.6. Environment 26
3.7.2.7. Right of Action of Private Groups 27

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b07_299

3.7.2.8. Causal Link and Administration of Proof 27
3.7.2.9. Financial Security Scheme 27
3.7.2.10. Access to Information 27
3.8. The 1993 European Commission Green Paper on Remedying Environmental Damage 27
3.9. Economic Aspects and Questions Bearing on the Insurance of the Industrial Risk 28
3.10. Conclusions 29
4. Waste Shipments in the European Union: The Interface between Environmental Regulation and the Internal Market 29
4.1. Introduction 29
4.2. European Community/Union Measures 30
4.2.1. The Internal Market, "1992" 30
4.2.2. The Single European Act 31
4.2.3. The Maastricht Treaty – Environmental Provisions 32
4.2.4. Other Applicable Provisions of the Treaty of Rome 33
4.3. Waste Shipments: Conflict between European Union Free Movement of Goods Principles and Environmental Regulation . 34
4.3.1. The New Regulation 259/93: Introduction 34
4.3.2. Specific Provisions of Regulation 259/93 36
4.3.2.1. Shipments of Waste between EU Member States 36
4.3.2.2. Exports of Waste from the EU . . 37
4.3.2.3. Imports of Waste into the EU . . . 37
4.3.2.4. Transit of Waste from Outside and Through the EU for Disposal or Recovery Outside the European Union 38
4.3.2.5. Common Provisions 38
4.3.3. The Trade – Free Movement and Environmental Protection Conflicts in Complying with EC Regulation 259/93 38
4.3.3.1. Definitions 38
4.3.3.2. Liabilities 41
4.3.3.3. Administrative Issues 41
4.3.3.4. Trade and Competition Issues . . . 42
4.4. Conclusion 43
5. The Seveso Directive: An Example of the Interface between Community Law and National Implementation Rules 43
5.1. Historical Background and Content 43
5.2. Implementation Requirements . 44
5.3. Conclusions 46
6. U.S. and European Environmental Legislation 47
6.1. Introduction 47
6.2. The Environmental Regulatory System 47
6.3. Methods of Lawmaking in the United States and the European Community 47
6.4. U.S. and European Environmental Laws 48
6.4.1. Environmental Impact Assessment Laws 49
6.4.2. Accident Planning and Public Information Laws 49
6.4.3. Air and Water Pollution Legislation 49
6.4.4. Solid and Hazardous Waste Regulation 50
6.4.5. Regulations on Chemical Products 51
6.5. Some Contrasts in U.S. and EC Environmental Legislation 52
6.5.1. Regulatory Gaps 52
6.5.2. Implementation of Legislation . . 52
6.5.3. Enforcement of Legislation 53
6.6. Outlook 54
7. Environment, Safety, and Chemical Safety in Japan 55
7.1. Current Situation Concerning Environmental Protection 55
7.1.1. Air Pollution Control 55
7.1.2. Control of Water Pollution 55
7.1.3. Revision of Compensation Systems for Patients with Pollution-Related Diseases 56
7.1.4. Measures for Waste Disposal . . . 56
7.1.5. Response to Global Environmental Problems 57
7.1.6. Basic Policy of the Japanese Chemical Industry Association about Environment and Safety 57
7.2. Current Situation Concerning Safety and Fire Prevention Measures 57
7.2.1. Revision of Fire Service Law . . . 57
7.2.2. Safety Measures for the Transport of Dangerous Goods 58
7.2.2.1. Marine Transport Standards for Hazardous Substances 58
7.2.2.2. Standards for Air Transport of Hazardous Cargoes 58

7.2.2.3. Standards for Transportation Containers for Poisonous and Toxic Substances 58
7.2.2.4. Safety Measures in the Use of Gases Under High Pressure 59
7.3. Present Situation Regarding Safety Measures for Chemical Substances 59
7.3.1. The Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances 59
7.3.2. Situation Regarding the Enforcement of Law Concerning Examinations and Regulations of Manufacture, etc. of Chemical Substances ("Chemical Substances Law") . . 60
7.3.3. International Cooperation on Chemical Substance Regulations . 60
8. References 61

The entire topic was coordinated by Jean-Marie Devos

1. Industrial Chemistry and Law

1.1. Law, Society, Chemistry, and Ethics

Law is the normative instrument that society uses to determine in a peaceful way the rules governing the relations between people (private law); people and authorities (public law); and states (international public law). It reflects to a large extent the values and ethics of a given society at a given time (→ Sustainability and the Chemical Industry – Ethical Aspects and Approaches). Normative activities have – at an early stage – addressed the question of humans and their interface with industrial and trade activities either to promote such activities, to control or restrict them, or in some cases, even to prohibit them.

Although the requirements of an organized society, as reflected in law, are supposed to be integrated, complementary, and consistent, they might actually conflict or at least appear contradictory. For example, the legal system might aim at economic objectives – for example, the free movement of products in an economic area – as reflected in one of the four basic freedoms of the European Communities Treaties. On the other hand, society and lawmakers might impose severe restrictions on the movement of products because of their impact on organized collectivities and the environment (see Section 4.3).

Law might actually impose restrictions and administrative burdens on research, production, transport, and distribution activities, either because of the "dangerous" nature of such activities or because they involve products that are, to various degrees, considered dangerous or simply perceived as negative or hostile to people or the environment. The end of the twentieth century has seen the emergence of the "environmental society". With the growing satisfaction of basic needs in the developed world and the collapse of traditional ideologies, growing awareness of environmental problems has become a major factor influencing sociology, politics, law, and economics. In parallel to the new "environmental" awareness, and in apparent contrast, industry and trade activities are more and more characterized by increased competition and "globalization." After the success of the General Agreement on Tariffs and Trade (GATT) round in Uruguay and the creation of a World Trade Organization, the dominant economic concepts and trade realities are clearly hostile to trade barriers and restrictions. The resolution of potential and actual conflicts between these two approaches should lie in a coherent and humanistic vision, based on human values, on responsibility, and on confidence in sustainable progress and science.

The main objective of this chapter is to focus on a few key trends which affect, or will influence, the socio-legal framework in which industry has to operate. As shown in the various chapters of the keyword Legal Aspects, innovation, production, distribution, and marketing strategies must integrate environmental and other ethical values which determine their legal framework.

At the same time, industry's response should be based on sound and credible science, on economic sense and on appropriate communication.

In other words, modern *legal systems* tend to address production, distribution activities and products from "cradle to grave" and respond to a growing social concern about perceived risks

generated by "dangerous" activities and products. This is particularly relevant in the area of chemicals which are subject to comprehensive sets of rules in the various regions of the world (see Chaps. 3, 6, 7). Perceived risks may be real and have to be seriously addressed. They may also result from fears generated by some media or pressure groups. In contrast, *science* has never been more advanced but, in a way, has become a victim of its own success as technology allows society to better measure even minor risk factors (or simply traces) which are then deemed to be unacceptable. [See for example the progress of detection limits of modern analytics using ultrasensitive equipment and the debate on the revision of the Drinking Water Directive [1]. The key issues are the allowance of greater flexibility for nonhealth parameters and reliance on toxicological data in setting maximum admissible concentrations (MAC) for health-related parameters. Progress in toxicological, analytical, and epidemiological knowledge and methods may imply that MAC should evolve, in line with WHO guidelines.]

The importance of innovation as a driving force of modern economy and society was defined by SCHUMPETER at the beginning of the 20th century as the entire process from the birth of an idea through to its widespread application in society. Innovation implies changed technologies, economic and social evolutions, and adaptability. The driving forces of innovation and changes are rooted in human curiosity and needs and are influenced by their evolving hierarchy.

Social psychology based on MASLOW's approach, distinguishes five hierarchical levels of needs

1) Physical survival and physiological needs
2) Safety
3) Social belonging
4) Respect and recognition
5) Self-fulfilment

According to classical marketing sciences, innovation was responding to "technology pushes" on the one hand and almost unlimited "market demands" on the other hand. Until the lower needs are satisfaction of human needs as described in MASLOW's ladder in industrialized countries and with the achievement of high standards of living, perceived priorities have evolved. Public awareness has become more and more attentive, and critical of "side effects" of innovation.

In other words, the better and the more social needs are fulfilled, the more acute is the perception and concern of the society about the instruments developed to respond to such needs and about their effects. This phenomenon has had a direct and massive impact on the lawmaking process and has led to a vast number of regulations, particularly in the health, safety, and environment area.

The Rio United Nations Conference on Environment and Development (UNCED) Declaration is probably the widest international effort to address environmental concerns on a global basis, although reflecting different Regional sensitivities and priorities. The growing environmental awareness has generated many "law – regulatory"-driven technologies and market developments (e.g., the compulsory introduction of catalytic converters for automobiles) and has led to the phased withdrawal of products considered as harmful to the environment (as for example, the phasing out of CFCs, initiated by the 1987 Montreal protocol on substances that deplete the ozone layer). Market and competition effects are far-reaching.

In the face of these developments and in parallel to the international, European and national lawmaking process, industry has gradually developed comprehensive "pro-active" programs. Such programs are meant to meet the public's and the authorities' concerns and to offer a sound managerial and operational response from industry. Such a response may take the form of general policy declarations or more operational programs. [See for example the "Business Charter for Sustainable Development" (November 1990) of the International Chamber of Commerce. The Charter lists sixteen principles for Environmental Management.] Typically, one leading feature of such policies is the improvement of information flows between industry, the public and the authorities.

On a more sectorial basis, the "Responsible Care" program developed by the chemical industry on a worldwide and a Regional basis reflects this sector's commitment towards achieving continuous and measurable progress in the health, safety, and environment areas. Association members of the International Coun-

cil of Chemical Associations, such as the European Chemical Industry Council (CEFIC), the Japanese Chemical Industry Association (JCIA), and the U.S. Chemical Manufacturers Association (MA) have identified Responsible Care as a priority and a contribution to sustainable development.

1.2. Policy and Legal Principles Affecting Industrial Chemistry

Years have seen the emergence of two major policy principles that influence and will continue to impact with special significance the lawmaking process affecting the exercise of activities related to chemicals.

1.2.1. Precautionary Principle

The "precautionary principle" has been used more and more frequently and quoted in public statements adopted in several international forums and according to some opinions has become a principle of (customary) international law [2–5]. Examples of international forums include:

1) Bergen Conference on Action for a Common Future, Norway, May 8 – 16, 1990. The Bergen Conference was organized by the government of Norway in cooperation with the UN Economic Commission for Europe as part of the preparation of the UN Rio Conference on Environment and Development. The Bergen Conference established that "in order to achieve sustainable developments, policies must be based on the Precautionary Principle. Environmental measures attack the causes of environmental degradation."
2) Principle 15 of Rio Declaration on Environment and Development, Rio de Janeiro, Brazil, June 13, 1992. "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. When there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environment degradation."

The precautionary principle is part of the wider concept of sustainable development and has been integrated in the European Communities' law since its introduction in the Maastricht Treaty [6]. Article 130 r, § 2, of the Treaty explicitly refers to the precautionary principle: "Community policy on the environment shall aim at a high level of protection . . . . It shall be based on the 'Precautionary Principle' and on the principles that preventive action should be taken, that environmental damage should as a priority be rectified at source and that the polluter should pay . . . ."

The new European Constitution is a major example of the fast recognition in a legally binding instrument of the two major principles for future environmental policies.

The precautionary principle institutionalizes caution when (sufficient) evidence exists that an activity or substance is likely to cause unacceptable harm to the environment. Its actual meaning, interpretation, and definition, however, are much debated because the implications of the principle could be far reaching, if not devastating. Definitions vary widely from encouraging a cautious and responsible attitude, based on scientific evidence, not to generate or discharge in an uncontrolled way elements hostile to humans and the environment, to a much more radical "anti-industrialization" attitude, which is leading to the prohibition of substances or activities even in the absence of conclusive scientific evidence. According to the precautionary principle, as defined by part of the doctrine and in several marine pollution conventions, a substance of activity posing a threat to humans or to the environment should therefore be prohibited (i.e., banned or restricted) even if there is no conclusive scientific proof of a causal link between that particular substance, or activity, and humans or environmental damage.

This rather radical approach was first used in the Ministerial Declaration of the Second North Sea Conference [7]. "The Parties agree to . . . accept the principle of safeguarding the marine ecosystem of the North Sea by reducing polluting emissions of substances that are persistent, toxic and liable to bioaccumulate at source by the use of the best available technology and other appropriate measures."

This applies especially when there is reason to assume that certain damage or harmful effects

on the living resources of the sea are likely to be caused by such substances, even where there is no scientific evidence to prove a causal link between emissions and effects ("the principle of precautionary action").

The overall goal was to achieve a substantial reduction in the total quantity of persistent, toxic and bioaccumulative substances reaching the North Sea. Actual policies of Member States are aiming at a target of 50 %. The national interpretations of the London Declaration have been far from homogeneous, most of them requiring some certainty of (scientific) evidence regarding environmentally damaging effects (i.e., France) or calling for actual evidence before taking action.

Those variations clearly reflect the embarrassment of public authorities who are confronted by complex situations in which multiple causation factors make scientific assessment and conclusions difficult or uncertain. [This embarrassment was reflected in the Third International Conference on the Protection of the North Sea, The Hague, March 7 – 8, 1990 (Third North Sea Conference). The final declaration confirmed the continuous application of the precautionary principle but limited it to those substances that are persistent, toxic, and liable to bioaccumulate.] A radical interpretation of the precautionary principle as reflected in the London Declaration is seriously objectionable – from both legal and scientific standpoints – because it conceptually opposes and rejects scientific evidence and causality as instruments for sound policy actions.

Such an interpretation necessarily leads, at the extreme, to an impasse and might be translated into a "paralysis principle." Indeed, a requirement based on the existence of "reasons to believe" that damage or harmful effects are "likely" to be caused becomes meaningless if opposed to or separated from scientific evidence. For example, a mere trace of a substance in the environment would equal pollution of that environment leading to possible prohibitons. The progress of analytical methods today allows the detection of minute quantities of a given substance (e.g., a piece of sugar in a lake). The progress of scientific and analytical methods without parallel progress of impact and risk assessment techniques could actually multiply purely "emotionally"-based policies. The exclusion of scientific evidence as a determining factor for shaping environmental policy might actually lead to the opposite of a precautionary and reasonable attitude. At the extreme, it might lead to decision-making processes entirely dominated by "beliefs" and "opinions" dominated by ideology without due regard for an objective and scientifically sound analysis of facts.

The radical interpretation of the precautionary principle seems to be disregarded by provisions of Article 130 r of the Maastricht Treaty. While endorsing the concept of the precautionary principle, the Treaty does not define it and obliges the Community to take account of available scientific and technical data [Article 130 r (§ 3)].

Realistically, complex situations and interactive systems might lead to a degree of uncertainty where scientific evidence is incomplete or even appears contradictory. The lack of full scientific evidence might encourage policy makers to take measures "where there are threats of serious or irreversible damage." Such measures must, however, be based on sufficient scientific evidence.

It remains to be seen if and how a moderate and reasoned interpretation will finally emerge in the international and national spheres, where scientific evidence and causation will actually serve and precondition the applicability of the precautionary principle.

[A dramatic and illustrative evolution has been the apparent change of policy of the U.S. administration, which, under the Reagan and Bush presidencies, carefully avoided endorsing the principle, while the EPA, under President Clinton, developed a strategy that would prohibit, reduce, or substitute the use of chlorine and chlorinated compounds on the basis of such a principle.

Clean Water Initiative Submitted to Congress, 1 February 1994. While aiming at a strategy for prohibiting, reducing, and substituting the use of chlorine and chlorinated compounds, the administration proposed a time schedule allowing federal agencies and "other experts" to comprehensively assess the use the environmental and health impacts of chlorine and chlorinated compounds, and availability and relative efficiency and safety of substitutes. Within 30 months, a plan "for our appropriate actions" should be developed. The National Academy of Sciences is expected to complete a

study for the Congress on the current knowledge of chemicals "that exhibit endocrine immune and nervous system health effects in human and wild life." The highly political and emotion-driven nature of this issue is demonstrated by the more moderate approach of the administration, which resulted from fierce reactions from many circles, including industry.]

The sole answer for policy makers is to develop more, not less, science and technology assessment and to assist policy and decision makers to take the most appropriate decisions.

1.2.2. The Polluter-Pays Principle

Historically, an economic and preventive principle, the "polluter-pays principle" is gradually influencing public and administrative law, or civil law.

The polluter-pays principle was initially adopted by the OECD in its Recommendations [8]. The 1991 OECD Recommendation [9] extended the polluter-pays principle to the "cost of damage" generated by pollution.

The Parliamentary Assembly of the Council of Europe summarized the extension of the polluter-pays principle to compensation of damage as follows:
"States, industry and all persons shall be liable to pay for the environmentally harmful consequences of their actions and development programmes ... The principle of the polluter's liability, or the "Polluter Pays Principle" as it is commonly known, shall be strictly applicable." [Recommendation 1130 (1990)]

The polluter-pays principle has been introduced into the European Communities Constitution via Article 130 r of the Single European Act [10] and later in the Treaty on the European Union. The extension of the polluter-pays principle as an environmental law principle has received priority consideration by the OECD. In its preventive form, the principle implies that those whose emissions could create a risk to the environment should bear the full cost of reducing or eliminating such emissions so as to prevent this risk from materializing.

The principle is thus also a means of internalizing environmental costs [11]. However, it is not yet clear what impact it will have in specific areas such as civil liability for damage, although it may be argued that it is based on a similar concept, i.e., the author of a damage, to persons or to property, has to compensate for such damage.

One of the important, and complex, questions raised by the polluter-pays principle within civil liability mechanisms is linked to the very question of "what is a damage?" since there is a tendency to extend the notion of damage to "damage to the environment", or "ecological damage". These are extremely difficult to define and to quantify. (For example, how is the degradation of the ozone layer to be evaluated in terms of damage and to which cause can it be rationally attributed? Who is the polluter? The "global climatic change" issue raises a number of complex questions of this nature.)

Other difficulties are linked to the establishment of causation, when several sources of pollution exist, and to the establishment of evidence. Should the plaintiff bring the full evidence that damage was caused by a given attributable incident?

Or is "prima facie" evidence sufficient? Should the burden of the proof be reversed with the risk of jeopardizing the very concept of civil liability? These, and other questions, have been debated in the context of the "Council of Europe" Convention and are still much argued within the European Union.

Efforts have also been made to make civil liability rules not only a legal instrument for compensation of injury to people and damage to property but also to use them as a (limited) tool for environment remediation. It is sometimes argued that strict civil liability could play a role in the prevention of environmental degradation and therefore be used as an instrument of environmental policy.

1.3. Economic Integration and Environmental Protection

The interface between economic integration and environmental protection is a major emerging issue at the end of the 20th century. Both on the world trade scene and within the European Union, law and policymakers have to reconcile potential conflicts between the logic of economic integration implying and enforcing similar rules and policies to facilitate trade and competition and, on the other hand, the logic of na-

tional/regional prohibitions and restrictions on products and activities. Global issues call for global solutions. Achieving cohesion is far from obvious, as illustrated by the present debates concerning the tensions between the Internal Market of the Community and environmental policies, but also in the debate within the World Trade Organization.

International trade of products which might have an impact on the environment is increasingly subject to severe restriction and/or to authorization and information procedures. For example, the case of waste movement restrictions and free trade is illustrative of such tensions. The EC Regulation on Control of Waste of February 1, 1993 [12] is intended to implement the Basel Convention on the "Control of Transboundary Movements of Hazardous Waste and their Disposal" (adopted and opened for signature on March 22, 1989). This Regulation completes and tightens the monitoring system already applicable in the Community. It covers five categories of transport: transfer between Member States, transfer within Member States, export to non-EC countries, imports from outside the EC and waste transfers. The role of the OECD, in seeking a workable solution especially for waste for recovery should be stressed.

The Regulation provides for a prior authorization system, recognizes a broad right for Member States to object to shipment and implements the self-sufficiency and the proximity principles which both aim at encouraging local waste disposal. Paradoxically, a piece of Community legislation, supposed to have direct effects, is opposing one of the fundamental freedoms of the EC Treaty, i.e., free circulation of goods. This has raised much criticism within the Community and the legal debate continues on the (impossible?) reconciliation of free movement of goods principles and restrictions of transboundary shipments of waste and recyclable materials (secondary raw materials) (see Chap. 4). Rather than concentrating its policy on the proximity principle, which may actually create inefficiencies and additional unnecessary costs, the Community should develop sound and uniform criteria for operating waste facilities and movements of waste.

Another example of restrictions to trade motivated by environmental concerns is the prior informed consent (PIC) procedure. The PIC procedure is based on the principle that international shipment of a chemical that is banned or severely restricted in order to protect human health and the environment should not proceed without the agreement, or contrary to the decision of, the designated national authority in the importing country [13–17]. These developments raise the question of their interface and compatibility with world trade liberalization agreements under the GATT. There is no doubt that the whole "trade and environment" debate will constitute a major global policy issue of the end of this century. For a comprehensive review, see [18].

1.4. Other General Trends of Health, Safety and Environmental Laws Affecting Industrial Chemistry

While legislations dealing with environmental aspects of industrial activities have to be situated within general public and civil law, the evolution of health, safety and, especially, environmental (HSE) laws has been characterized by the following elements:

1) A more autonomous, systematic system of law is gradually developing both in the national, European, and international spheres. The fast development of a European Community comprehensive environmental law is based on policy programs which aim at fully integrating environmental policies in other Community policies and in setting long-term objectives and targets [19], [20]. The primacy of Community law over national legislation, where sufficient, clear, and precise provisions exist, and the direct effect into national legal systems have far-reaching implications.
2) The gradual emergence of a "right to know" both in the European Union and in the United States is a good example of the impact of environmental law concepts on public and administrative law. This right to know gives citizens broad access to information obtained by public authorities in the environmental area, subject to some limitations [21].
3) The impact of environmental issues on competition policy. In a growing number of cases, "environmental benefits" are claimed in industrial and marketing strategies both

vis-à-vis consumers and vis-à-vis public authorities. This development raises many questions about ethics and (fair) competition. The eco-label concept is an illustration of this evolution [22].

Other aspects are linked to state aids given for environmental purposes which might distort competition in a given industrial sector.

The European Commission has issued guidelines describing its policy on state aid for environmental protection. Such guidelines would authorize up to 30 % state subsidy for (voluntary) investment related to the environment, while investment aimed at compliance with environmental legislation could be supported up to 15 % of the cost [23].

1.5. Scope of this Keyword

The keyword Legal Aspects does not aim at providing a comprehensive or exhaustive review of European, national, or international rules affecting chemicals. For a publication relating to international aspects of chemical control see [24]. The mere coverage of, for example, European Community legislation would require several volumes of Ullmann's. Many specific aspects have already been addressed in other keywords dealt with in volumes B 7/B 8 or in relevant sections of them. The main objective of the keyword Legal Aspects is to make better understood the interface between the legal system in general and chemical-related activities and to present in this respect the major trends characterizing the evolution of law on the eve of the year 2000.

2. European Community Legislation Affecting Chemicals

Definition of Terms Used in this Chapter. The European Union (EU) was set up by the Treaty of Maastricht, which came into force in November 1993. Part of the Treaty on the EU are the Treaties on the European Communities (EC), the EURATOM Treaty, and the Coal and Steel Treaty.

The EU does not have legal personality, but the EC has. Therefore, any legal instrument is adopted not by the EU but by the EC. Whenever references to legal texts are made, the term EC is correct, EU may only be used in a political context.

Prior to the coming into force of the Maastricht Treaty the correct name for the EC was European Economic Community (EEC). Thus, all legal texts adopted before November 1993 have in their title the reference EEC.

2.1. The European Community Framework

The European Community has, under different Treaties, the objective of creating an integrated economic community in which industrial frontiers are no longer economic frontiers, and which includes the free circulation of goods and services, capital, and labor as well as a high level of health, safety and environmental protection. The mandate to establish and monitor the Internal Market for goods and services, capital, and labor is in the Treaties given to Community institutions: Council and Commission, the European Parliament, the Economic and Social Committee, and the European Court of Justice (ECJ). Member States have the objective of facilitating achievement of the Community's tasks, and abstaining from any measures that could jeopardize the attainment of Community objectives [25]. In contrast to that there is the joint responsibility between the European Community and its Member States acting in isolation to ensure an appropriate protection of the environment [26]. This principle of subsidiarity, enshrined since autumn 1993 in the general provisions of the EC Treaty, has existed since the 1987 amendment of the environmental chapter of the EEC Treaty. Proposals for legislation made by the EC Commission have been adopted or rejected without express recurrence to this principle.

Chemical legislation in the European Community must find its way through this framework of responsibilities. In very broad terms, past legislation on product standards was generated mainly at Community level and thus had a considerable degree of uniformity. Legislation banning products or restricting their use was more often adopted at a national level; with the completion of the internal Community market, a marked trend exists towards integrated rules

for bans and restrictions, but progress is slow. With regard to production-orientated legislation, Community legislation in the past has been unsystematic and limited to framework rules. The tendency has developed to arrive at a more systematic approach, although the responsibility for rule making remains shared between Member States and the Community, the Community being limited to umbrella legislation.

2.2. Product-Oriented Legislation

Community legislation on chemical substances and preparations as well as on specific chemical products – food additives, pharmaceuticals, cosmetics, fertilizers – started early in the 1960s, shortly after the creation of the EEC. This legislation was based almost entirely on Article 100 of the EEC Treaty, which reads:
"The Council shall, acting unanimously on a proposal from the Commission, issue Directives for the approximation of such provisions laid down by law, Regulations or administrative action in Member States which directly affect the establishment or functioning of the common market. . .".

Total or Optional Harmonization? The approximation of national legislation aimed at diminishing the differences between national rules that created, or were capable of creating, barriers to the free circulation of goods. Since in all the above-mentioned sectors of chemical products, questions of health and safety considerably influenced rule setting, the EEC did not attempt the legislative technique called "optional harmonization" but rather used a "total harmonization" approach.

Under *optional harmonization*, Member States fix standards for national production at their discretion and at a level they consider appropriate. However, they are obliged to admit on their market those products from other Member States that comply with Community standards. This kind of approach was, for instance, largely used by the EEC in the area of technical legislation.

Under the *total harmonization* approach, which was used in the area of chemical products, national legislation is replaced by Community legislation so that, finally, one rule exists within the Community. This approach is obviously more ambitious, more delicate, and more time consuming. Since questions of health and safety are central to all Member States' national policy, progress in replacing the different national rules by one single Community rule has been slow and even slower in the areas of food and pharmaceuticals than in other areas.

In the area of chemical substances and preparations, national legislation in the 1960s was not very sophisticated. Therefore, attempts to create uniform rules for chemicals were relatively successful. Community legislation on chemical substances [27], solvents [28], pesticides [29], and paints and varnishes [30] was adopted soon after the transition period for setting up the EEC. This legislation set uniform rules for the classification, labeling, and packaging of substances and preparations, and ensured that chemicals conforming to these requirements were allowed to be circulated freely within the Community. Since this legislation in the 1960s and 1970s had to be adopted unanimously, Member States were able to block the adoption of Community legislation that did not conform to their national rules on health and safety, until their concern was met.

A Dynamic Cooperative System under the Control of the Commission. The possibility of blocking legislation served to protect the interests of Member States in terms of health and safety, before such legislation was adopted. However, this need could also arise after the adoption of Community legislation, for instancce, when an accident in a Member State demonstrated that the Community rule was not strict enough or where new scientific evidence was found. Since an amendment of Community legislation was slow and required an initiative from the EEC Commission, this mechanism was not considered able to adequately cover all cases. Therefore, the Community legislation on chemicals contained a safety clause: where a Member State found that a chemical, although complying fully with Community rules, presented a risk to health and safety, the Member State was allowed to take provisional measures to counteract that risk. These measures could even go so far as to remove the chemical from the market. The Member State that took such a measure had to inform the Commission and other Member

States of its action and give a reason for its decision. Then, a Community procedure was started to decide whether the national measure should be extended to the entire Community. Until such a decision, the national safeguard measure was maintained. By way of illustration, the text of such a safeguard clause is given here [31]:

"Where a Member State has detailed evidence that a substance, although satisfying the requirements of this Directive, constitutes a hazard for man or the environment by reason of its classification, packaging and labelling, it may provisionally prohibit the sale of that substance or subject it to special conditions in its territory. It shall immediately inform the Commission and the other Member States of such action and give reasons for its decision. The Commission shall consult the Member States concerned within six weeks, then give its view without delay and take the appropriate measures. If the Commission considers that technical adaptations to this Directive are necessary, such adaptations shall be adopted, either by the Commission or by the Council in accordance with the procedure laid down in Article 21; in such a case, the Member State which has adopted safeguard measures may maintain them until the adaptations enter into force."

In the mid-1970s, two elements were added to this situation – attempts to introduce Community-wide bans and restrictions on chemicals, and rising environmental concerns. Clearly, the feeling was that legislation on chemicals that tried to open up national frontiers and set uniform rules was insufficient since national prohibitions and restrictions for use continued to exist, and with growing concerns for health, safety, and the environment, new restrictions were being introduced at the national level. To ensure uniform standards, the EC felt the need to adopt Community-wide restrictions or bans. This legislation [32] was adopted on an ad hoc basis: whenever nationally – or indeed in international bodies – the need was felt to ban or restrict the use of a chemical product, the European Economic Community tried where appropriate, to reach Community-wide standards. The difficulty of reaching a consensus on such restrictions under the unanimity rule had, as a consequence, the adoption of a relatively limited number of such Community rules. Several attempts to accelerate the decision-making procedure at the Community level were unsuccessful, probably mainly because Member States were reluctant to see responsibility for health and safety questions shift more and more to the Community. Thus, between 1976 and 1992 some 30 substances were banned or severely restricted in use, covering only a part of the national bans or restrictions.

The second element was upcoming environmental concern in legislation. This concern gradually led some Member States to legislate in the area of chemicals, in order to protect, preserve, or improve the quality of the environment. The EEC reacted by including environmental considerations into its existing legislation on chemicals. The evolution of Directive 67/548/EEC, adopted in 1967 at Community level, is characteristic in this regard: in the mid-1970s, France informed the Commission of its intention to introduce new chemical legislation that took into a account environmental considerations. The Commission was of the opinion that such legislation would be capable of creating new obstacles to trade within the Community. Therefore, it suggested an amendment to Directive 67/548/EEC that finally led to the adoption of Directive 79/831/EEC, which, for the first time in such precise terms, talked of the need "to protect man and the environment" [33]. Since environmental policy is based, among other factors, on the principle that preventive action should be taken [34], that Directive introduced the principle that before new chemical substances were placed on the market, information had to be provided to competent authorities, together with test results on the substances, in order to enable the authorities to evaluate the foreseeable risk for humans and the environment. This evaluation is made Community-wide in concert between national and Community administrations. Conditions for marketing of new chemical substances based on health, safety, or environmental considerations are fixed for the entire Community.

A similar approach regarding notification prior to placing a substance on the market was used when the Community adopted legislation for genetically modified organisms on the market (deliberate release) [35]. Again, evaluation of the risk of such deliberate release is made Community-wide prior to the release.

However, this approach is by no means generalized; it is not applied to all chemicals, i.e., chemical preparations can be placed freely on the market without prior provision of test results and other information to competent authorities.

In 1987, the EC Treaty was amended. Specific rules were introduced to ensure achievement of the Internal Market by the end of 1992. For that purpose, a new Article 100 a provided that measures with the objective of achieving the Internal Market could be adopted by majority decision of Member States. At the same time, a new Article 100 a (4) was included in the Treaty, which reads:

"If, after the adoption of a harmonization measure by the Council acting by a qualified majority, a Member State deems it necessary to apply national provisions on grounds of major needs referred to in Article 36, or relating to protection of the environment or the working environment, it shall notify the Commission of these provisions. The Commission shall confirm the provisions involved after having verified that they are not a means of arbitrary discrimination or a disguised restriction on trade between Member States . . .".

Furthermore, Community environmental policy was considerably strengthened, by the introduction of an environmental chapter, Articles 130 r to 130 t, into the Treaty. Environmental measures were, in principle, to be adopted unanimously (Art. 130 s). Article 130 t stipulated, "The protective measures adopted in common pursuant to Article 130 s shall not prevent any Member State from maintaining or introducing more stringent measures compatible with this Treaty."

The borderline between the provisions of Article 100 a and Article 130 s seems, with regard to chemical products, to be rather clear: since chemical preparations are products and should –under the concept of an Internal Market, where national frontiers are no longer economical frontiers – circulate freely within the Community, the legal foundation for rules on their classification, labeling, or packaging – including their ban or restrictions on use – should be based on Article 100 a. Article 130 s would be reserved to production-related legislation that did not affect the establishment or the functioning of the Internal Market.

The majority of Community chemical legislation did indeed follow this approach. However, not all legislation respected this pattern. When the Council adopted legislation on chlorofluorocarbons (CFCs) and other ozone-depleting substances, this legislation was based on Article 130 s [36]; the same approach was followed on waste legislation, although, from the point of view of the environment, there exists little difference between poisoning by chemicals or by waste [37]. Thus, because of national initiatives remaining possible under Article 130 t, uniformity of rules within the Community was not achieved.

Even where new legislation on chemicals was based on Article 100 a, the establishment of uniform rules was questioned. Indeed, certain interpretations of the above-mentioned clause of Article 100 a (4) are equivalent to the wording of Article 130 t. In other words, this interpretation would allow Member States, after the adoption of Community standards, to introduce new, diverging rules at the national level, not merely to continue to apply national legislation that existed prior to the adoption of Community rules. Whether Article 100 a really has the same meaning as Article 130 t or should be interpreted more narrowly, remains to be decided by the European Court of Justice, which, until now, has not had occasion to pronounce itself on this question.

Pending this question and more generally, Member States remain responsible, together with Community institutions, for the protection of health, safety, and the environment. They have thus not lost the responsibility for banning or restricting the use of chemical products, as long as no comprehensive legislation setting up Community-wide bans and restrictions exists. Article 100 a, seen in more general context, enables the Community to draft uniform rules for chemicals that are based on a high level of protection of health, safety, and the environment. To the extent that the Community manages to elaborate, adopt, and permanently adapt such rules to technical and scientific progress, recurrence to national bans and restrictions will not be necessary; thus Article 100 a (4) would lose practical importance.

The Maastricht Treaty on European Union has not changed the principles on rule making at the Community level. Thus, the trends described

above are likely to remain applicable in the coming years.

2.3. Legislation on Chemical Installations

As stated above, Community legislation on chemical installations has not been very systematic in the past. The establishment and functioning of a common market were not in the past perceived to require the elaboration of Community-wide standards for industrial installations. Thus, rules affecting chemical installations were set up only with the introduction of an environmental policy at the Community level and under environmental protection auspices. Industrial installations that discharge dangerous substances into waters or into groundwater or emit pollutants into the air need specific authorization for such emission [38]. Such permitting rules were, however, far from covering all installations or all pollutants.

An industrial accident in Seveso, Italy, provoked Community legislation on the prevention of major accident hazards of certain industrial activities [39]. This so- called Seveso Directive, adopted in 1982 and amended several times, requested installations that process or stock chemicals to take measures to prevent industrial accidents, to inform the adjacent population of potential risks, and to elaborate on-site and off-site emergency plans. This legislation is described in detail in Chapter 5.

In 1985, the EEC adopted a Directive that required an environmental impact assessment for a number of public or private projects, before planning consent was given [40]. In 1993, the EEC adopted a Regulation on the voluntary participation in a system of eco-auditing [41]. This Regulation aims at the improvement of eco-management by industrial installations. The Regulation will take effect in spring 1995. Also, in 1993, the Commission proposed a legislative measure at the Community level on integrated pollution control, which aims at a generalized, integrated permitting system for industrial installations, in order to improve, rationalize, and harmonize authorizations for industry and adequately protect the environment [42].

These Community rules on chemical installations were either umbrella rules or left Member States relatively large discretion in complementary rule making at the national level. This reflects the fact that while chemical products should circulate freely within the EC, the chemical installations that produced these products are not movable. Therefore, where legislation on installations was adopted prior to the 1987 amendment of the EC Treaty, it was based on Articles 100 and 235 of the EC Treaty at the same time. After the single Act 1987 took effect, Community legislation on installations was based on Article 130 s; the one exception until now – legislation harmonizing national programs on disposal of waste from the titanium dioxide industry, which was finally based on Article 100 a of the EEC Treaty [43] – seems isolated. In general, legislation on industrial installations will for the foreseeable future probably be based on Article 130 s of the EEC Treaty. This is likely also to apply to accessory environmental legislation, such as emission of pollutants into the air or soil.

The consequence of the use of Article 130 s is that Member States, under Article 130 t, will remain entitled to maintain or introduce more protective legislation with regard to chemical installations. Different national legislation, however, might lead to different production costs and gradually influence the competitive situation of industrial installations. This, in turn, might have an influence on the free circulation of goods produced in such installations. Where Community rules on installations have a significant impact on the cost of production, good reasons exist to consider whether the establishment of an Internal Market does not require equivalent standards over the entire Community territory. This would suggest that for such rules, Article 100 a be chosen as a legal basis, since standards that are laid down on the basis of Article 130 s will – because of the existence of Article 130 t – never be capable of achieving uniformity. For example, Directive 82/501/EEC on the prevention of major accidents, which affects mainly industrial installations that produce or process chemical products, might well request such considerable investments for accident prevention that rules within different Member States would distort competition for the chemical products that are placed on the market. According to the content of such a revised Directive, good reason might thus exist to base such a Directive on Ar-

ticle 100 a and not on Article 130 s of the Treaty. Indeed, the precise content of the Directive will determine whether one can argue that the objective of such an accident prevention Directive was principally the protection of humans and the environment and whether the establishment of local competitive conditions among installations was only an accessory objective.

3. Towards the Adoption of New Civil Liability Rules: A New Challenge for the Chemical Industry

"Toutes les pertes et tous les dommages qui peuvent arriver par le fait de quelque personne, soit imprudence, légèreté, ignorance de ce que l'on doit savoir, ou autres fautes semblables, si légères qu'elles puissent être, doivent être réparées par celui dont l'imprudence ou autre faute y a donné lieu. Car c'est un tort qu'il a fait, quand même il n'aurait pas eu l'intention de nuire." Extract from "Lois Civiles, II" by Jean Domat (1625 – 1696). The original French text can be translated as follows: "All losses and all damage caused by the deeds of any person, through negligence, thoughtlessness, ignorance of what one should know or similar shortcomings, however trivial they may be, must be compensated by the person whose fault or negligence has created the damage. Because he generated a tort, even though this was unintentional."

3.1. Introduction

Significance of Civil Liability and its Evolution. A significant number of research, production, and processing activities are carried out by a host of individuals and companies. Once manufactured, the most diverse products – from the simplest to the most technologically complex – are put on the market and are subject, among other things, to transport, packaging, labeling, and storage operations. All of these activities may cause damage either to persons, to goods, or to the environment. Thus, one can easily imagine the importance of "civil liability law" or "third-party liability" and its economic impact on industry. "Manufacturer's liability" must be considered in the framework of general legal systems of third-party liability, or "responsabilité civile ou quasi-délictuelle" in French civil code systems.

Basics of Civil Liability. Traditionally speaking, civil liability for damage means an obligation, under certain conditions, to compensate for damage caused to persons, property, or proprietary rights. The concept of environmental damage is relatively new and refers only to the nature of the damage. It basically addresses damage to the unowned environment, for example, disappearance of a fish species in a river. The plaintiff has to prove the existence of the following factors: (1) that he/she has suffered damage to his/her person or property; (2) that the damage was caused by an offense of negligence by the defending party (i.e., was based on fault); and (3) that a causal link exists between the damage and the incident.

These fundamental principles of civil law are similar in most European countries but are subject to evolution as described below. Liability instruments of direct interest to industry have been drafted in relation to

1) Defective products
2) Transport of (hazardous) goods
3) Industrial activities

Chapter 3 addresses liability, and compensation questions, in the context of industrial and marketing or trade activities. In the following material, the major trends affecting the evolution of civil liability are outlined, and various major international legal instruments recently adopted or still under discussion in this area are reviewed.

3.2. General Trends and Debates Affecting Modern Civil Liability Systems

3.2.1. New Trends

When can a Person or Company Be Held Liable? In most European countries, existing civil liability laws are based on fault ("continental law systems") or on the "negligence concept" ("common law systems"). However, in some countries new laws with a "strict" liability regime as a basis for liability have been enacted or are under consideration. The strict liability regime usually applies to damage caused

by activities that are deemed especially dangerous, such as handling hazardous waste, dangerous substances, or nuclear energy. In other cases, the case law (theoretically based on the fault concept), has led to a similar result. Under common law, the "duty of care" has been considerably extended and case law developed in the direction of strict liability. Similarly, fault-based liability is interpreted by courts and jurisdictions in such a stringent manner that the difference between these regimes has shrunk. Even without such specific law, the case law of most countries has developed a strict liability regime for damage caused by dangerous activities, such as handling explosives or operating dangerous installations.

At European level, the Directive on defective products (see Section 3.3) is based on strict liability of producers for defective products put on the market. A similar trend toward strict liability has characterized international conventions in the context of land and maritime transport of dangerous goods [the so-called International Institute for the Unification of Private Law (UNIDROIT) convention, and the present "Hazardous Noxious Substances" (HNS) draft convention of the International Maritime Organization].

Caution is needed and "labels" can be misleading because a fault-based liability regime can actually be more severe for potentially liable parties than a well-defined strict liability system. Enormous conceptual and practical differences also exist between the so-called strict liability systems.

For example, little common ground exists between the "operator's liability" approach, which is the basis of the Council of Europe Convention (see Section 3.7), and the absolute "generator's liability," which was once proposed in the context of a draft protocol on liability for waste to the Basel Convention [44]. Both texts are supposed to be based on strict liability. The operator's liability approach focuses on actual responsibility of the person in control of certain activities and substances; the generator's liability approach merely indicates who should bear the damage consequences of products (here waste) whether this person is, or is not, in control. The generator can be held liable even for acts committed by third parties, professional or not.

Arguments in Defense. Certain arguments may be used to defend oneself. One cannot be held liable if one proves the existence of "force majeure" (i.e., the damage was caused by an act of war or hostilities or by a natural phenomenon of an exceptional, inevitable, and irresistible character). A similar defense is, for instance, a claim that the damage was caused by an intentional act by a third party or that the person who suffered the damage had by his own fault contributed to the damage.

Additionally, different laws contain defenses of a special nature typical of the area concerned. The state-of-the-art defense is one example of regulatory exclusion of liability. According to the Directive on defective products, producers are not held liable if they prove that the state of scientific and technical knowledge at the time they put the product into circulation was not such as to enable the existence of a defect to be discovered.

In the context of liability related to industrial activities, compliance with permits is taken into consideration although it does not lead to an absolute defense. Emissions within "tolerable levels under local circumstances" (Council of Europe Convention) also offer possible defenses depending on legislation or case laws.

The Emergence of the Polluter-Pays Principle in Civil Liability Law (see Section 1.2.2). The polluter-pays principle was introduced in the European Communities Constitution via Article 130 r of the Single European Act and later in the Treaty on the Economic Union. Scientific literature on the polluter-pays principle is considerable but in what concrete and practical manner it will be applied is still unclear because interpretations of questions such as "Who is the polluter?" still differ.

3.2.2. Recognition of Ecological Damage in International and Community Law of Civil Liability

One of the most significant aspects of the present debates and current work in the international sphere is the extension of the concept of damage (death, personal injury, damage to property) to ecological damage. This same ecological damage is defined in various ways, depending on the instruments analyzed. Some common features

can, however, be identified: (1) environmental damage concerns the community, not the individual; (2) environmental damage is difficult to quantify. For example, what is the value of a variety of flowers that has disappeared because of pollution? U.S. jurisdictions and legal professions have made considerable efforts to quantify the ecological damage. Another interesting source is the case law developed in connection with major maritime pollution cases (e.g., the Amoco Cadiz decision).

The European Commission Green Paper on Remedying Environmental Damage of 1993 (see Section 3.2) envisages (1) the harmonization of civil liability rules for environmental damage and (2) the possibility of remedying environmental damage not met by the application of civil liability principles. The Green Paper, being a discussion document only, does not propose any definition of environmental damage but refers to definitions contained in the Council of Europe Convention and in the amended proposal for a Council Directive for damage caused by waste. The fundamental importance of and difficulties related to the definition are emphasized.

The 1991 amended proposal for a Council Directive on civil liability for damage caused by waste (see Section 3.4) defines "impairment of the environment" as any significant physical, chemical or biological deterioration of the environment in so far as this is not considered to be damage within the meaning of "damage to property." "Damage" as such, in turn, is traditionally defined to mean death or physical injury and damage to property. The compensation obligation under the Directive would apply for both damage and impairment of the environment. The Council of Europe Convention refers to "loss or damage by impairment of the environment . . . [which] shall be limited to the costs of measures of reinstatement actually undertaken or to be undertaken." This definition includes the definition of the "CRTD Geneva Convention" of October 10, 1989 (see Section 3.5).

The word environment is itself defined very broadly in the Council of Europe Convention; it covers natural abiotic and biotic resources, such as air, water, soil, fauna, and flora, and the interaction among the same factors, assets making up the cultural heritage, and the characteristic aspects of the landscape. These last two elements of the definition illustrate the erosion of the traditional concept of damage, because it is now a matter of covering damage to collective property by civil liability mechanisms. Such damage is particularly difficult to quantify. The Commission, unlike the Council of Europe, has avoided giving a definition of the word environment as such in the mentioned proposal for the Directive on liability for waste.

3.2.3. International Harmonization in the Context of Liability for Dangerous Activities

Opinion is growing that civil liability rules should be addressed at the international level not only in relation to products and goods, but also in respect of industrial activities. International rules are said to be necessary because pollution is no respecter of national boundaries and Member States have been establishing different Regulations in this field that could distort competition and hamper the smooth operation of a single market. Protection of the environment is also one of the Community aims, equivalent to creation of the single market, as enshrined in the Treaty of Rome Articles 130 r – t.

This approach was much debated in the early 1990s. Although harmonization of civil laws is conceptually attractive from the point of view of the Internal Market, it is effectively desirable for industry only if it would simplify existing systems and if new international instruments do not simply add another layer of possible legal actions. Harmonization efforts are also seriously challenged by the existence of different procedural rules linked to national judicial systems.

The two most important efforts to introduce harmonization of liability for dangerous activities have been developed within the Council of Europe and the EEC. The Commission Green Paper on Remedying Environmental Damage envisages harmonization of (strict) civil liability rules for environmental damage, as well as the possibility of remedying environmental damage not met by the application of civil liability principles. The latter means creating compensation funds to be used when no liable party can be identified or in case of the insolvency of the liable party. Civil liability is also considered the domain of Community legislation because it is a legal principle closely linked to general Euro-

pean enviromental policy principles, such as the prevention of damage and the polluter pays.

Another important international document is the already-mentioned Council of Europe Convention on civil liability for damage resulting from activities dangerous to the environment. The convention was influenced by major pollution of the Rhine river by Sandoz that affected several countries. It is based on the belief that issues relating to civil liability and compensation for damage resulting from dangerous activities need to be addressed at the international level.

3.2.4. Actions and Requests by Organizations

The right to sue is normally given only to the party with a legal (i.e., direct and certain) interest in recovering compensation for injuries or for damage suffered to property. Where damage occurs to property that is not owned (i.e., environmental damage), no injured party with the right to bring legal action can be identified. At the European level the tendency is to grant such a right – although limited and under the control and at the discretion of the courts – to environmental associations. This right to protect the "public interest" normally rests with the public authorities that operate under public law and democratic institutions.

For example, according to Article 4(3) of the amended proposal for a Council Directive on civil liability for damage caused by waste, common interest groups or associations, which have as their objective the protection of nature and the environment, are given the right either to seek remedy or to join in legal proceedings that have already been brought. Also the Council of Europe Convention gives certain restricted rights to associations seeking to protect the environment; namely, the group may request of courts or jurisdictions (1) that an unlawful, dangerous activity, which poses a grave threat of damage to the environment be prohibited; (2) that the operator be ordered to take measures to prevent an incident or, after an incident, to prevent damage; or (3) that the operator be ordered to take measures of reinstatement. However, any party to the convention may reserve the right not to apply this provision.

3.2.5. Financial Security Schemes

Several proposals suggest compulsory insurance or other financial security schemes to cover the liability under various pieces of legislation. Article 11 of the amended proposal for a Council Directive on civil liability for damage caused by waste provides that the liability of the producer, who in the course of a commercial or industrial activity produces waste, and of the eliminator of waste, be covered by insurance or other financial security. The amended proposal for a Council Directive on the landfill of waste, in turn, contains an obligation on the operator of the landfill to have a financial guarantee or any other equivalent, whose purpose is to cover the estimated costs of the closure procedures and after-care operations of the landfill (Art. 17).

3.2.6. Essential Requirements

The trends described above represent a real evolution in fundamental civil liability rules. However, some of them are still quite flexible, and others imply more theoretical than practical changes compared to the existing situation. An example of this is the discussion of whether liability should be fault based or strict.

Nevertheless, some essential requirements exist for any legislation on civil liability, one of the most significant of which is maintaining the establishment of causation. It is of utmost importance to the entire system that the causal link between the incident and the damage suffered be shown by the plaintiff. Likewise, in the context of damage caused by industrial activities, causation must be proved to exist between the operator in charge of the activity and the incident that led to the damage.

3.3. Liability for Defective Products and European Community Law

3.3.1. Background and Legal Basis

The Directive on defective products was approved by the Council on July 25, 1985 [45]. European Community Member States had to implement the Directive by July 1988. In early

1994, most Member States had effectively introduced the Directive into their internal law. Only France and Spain had not yet enacted the Directive, which led the Commission to open legal proceedings in the European Court of Justice for the noncompliance of these two countries with their obligations under the EC Treaty.

The declared objective of approximation of the laws has been partially attained since the text allows Member States to introduce some variations to three specific provisions. Article 100 of the Treaty of Rome provides the legal basis for the Directive and lays down the following conditions:

1) Unanimous council decision, based on a proposal from the Commission
2) Compulsory consultation of the assembly (European Parliament) and the economic and social committee if there is a need to amend national laws. Both institutions were consulted and expressed their opinion.
3) Grounds for the Directive: According to indent one, the approximation of the laws of the Member States concerning the liability of producers for damage caused by the defectiveness of products is necessary because existing differences may distort competition, affect the free movement of goods within the Common Market, and lead to differing degrees of consumer protection.
4) Social – political justification: The preamble assumes that "liability without fault on the part of the producer is the sole means of adequately solving the problem, peculiar to our age of increasing technicality, of a fair apportionment of the risks inherent in modern technological productions."

The industry has challenged strict liability and expressed preference for liability regimes based on fault or negligence in accordance with traditional principles of the law of torts. This battle was lost for the chemical industry, however, when even Italy, once a strong opponent of strict liability, accepted this regime in the political compromise that led to the adoption of the Directive in 1985. Was this a real source of concern to industry and should strict liability a priori be perceived as hostile by industry? The answer very much depends on three interrelated factors:

1) The content of actual provisions of the Directive
2) The content of national laws implementing the Directive
3) Case law, including that of the EC Court of Justice, that could play an important role in clarifying ambiguities

As stressed in Chapter 5, more important than the "labels" or names used to qualify liability regimes are their specific features. Other elements are also important: the practice of the courts, the judicial organization and its cost, and the general attitude of the public toward the liability and compensation system. In this connection, the Directive, although far from perfect from the point of view of legal certainty, was probably a reasonable compromise acceptable to governments, industry, and consumers (although some would not share this view, both in consumer circles and in industry).

As shown below, the question of recognition of the state-of-the-art defense was an important aspect of the "package" finally adopted and illustrates quite well the conceptual tensions and arguments that prevailed during the European Commission and the Council of Ministers discussions.

3.3.2. Nature of the Liability Regime

Principle: No-fault Liability. The Directive provides for the following system: the injured party has to prove the damage, the defect of a product, and the causal relationship between the defect and the damage (Art. 4). Thus, negligence and fault are no longer relevant as sources of producer's liability. Suppression of the "fault" standard should ease the task of the victim. However, it should be stressed that the case law of most countries had already introduced a softening of the victim's position, for example, in reversing the burden of proof or in extending considerably the concept of fault or "negligence" of the producer. According to common law concepts, the duty of care is now very strict indeed, with the important exception of the state-of-the-art defense.

Lines of Defense and Exemptions from Liability (Art. 7 of the Directive). The producer

must prove that his/her is one of the cases of exemption provided for, namely,

1) That he/she did not put the product into circulation
2) That the defect did not exist when the product was put into circulation or that it came into being afterwards
3) That the product was neither manufactured for sale or any other form of distribution, nor manufactured or distributed by him/her in the course of his/her business
4) That the defect is due to compliance of the product with mandatory Regulations issued by the public authorities
5) State-of-the-art defense ("development risk", see Section 3.3.3)
6) In the case of a manufacturer of a component, that the defect is attributable to the design of the product in which the component has been fitted or to instructions given by the manufacturer of the product

This last point is very important and was stressed by the chemical industry. All sectors of the chemical industry should be fully aware of this defense. It will be more and more important to have clear agreements between purveyors and buyers about the purity and characteristics of component products.

3.3.3. The Three Options of the Directive

A fundamental question that had to be addressed by the authors of the Directive was the so- called harmonization question. Should the EEC Directive provide for strict harmonization ("maximum Directive") or should it create a "minimum system" allowing the Member States to establish different and additional standards, depending on their concepts of consumer protection? Real harmonization being one of the raisons d'être of this text was preferable. The EEC Member States delegations recognized that a maximum Directive would be the best solution from the Treaty of Rome point of view, but at the level of details, compromises became necessary and the maximum approach turned into a limited approximation of laws.

State-of-the-Art Defense. A highly important question is, should the manufacturer of a defective product be held liable if the state of scientific and technological development did not enable him to detect the existence of a defect at the time the product was put in circulation. The inclusion of development risks in the scope of liability was predicted to interfere with technological innovation. Until now, in the vast majority of Western legal systems, no party has had to accept the risk of the unknown and unforeseeable. As CEFIC wrote in a position paper addressed to the EEC Council:
"While it is perfectly understandable that a system of liability should oblige producers to take all possible precautions against any possible injury to the consumer, it seems totally unreasonable to require that these same producers be liable for the unknown, and for risks over which they have no control whatsoever. Injury caused by a defect in a product cannot be a source of liability if the defect was undetectable and unforeseen at the time of its being put into circulation because of the state of the art."

After long and difficult debates, the EC Council finally introduced in the Directive what is known as the state-of-the-art defense. Accordingly, the producer has the possibility of proving that the state of scientific and technical knowledge when the product was put into circulation was not such as to enable the existence of the defect to be discovered. This defense may therefore be introduced in the legislation of Member States and was effectively introduced in most national laws of implementation.

However, the Community regime allows individual Member States to maintain the liability of the producer for "development risk" if it already existed in their internal laws when the Directive was approved (i.e., July 1985). Only Germany was affected by this provision, because its Pharmaceutical Act did not allow producers to use the state-of-the-art defense.

The Community regime also allows Member States to introduce derogations to that principle in new legislation, although this possibility is tempered by a clause of Community "standstill" and the obligation to follow a specific procedure under the supervision of the Commission [Art. 15 (1) of the Directive]. The importance of recognizing a state-of-the art defense for sectors such as the chemical and related industries is clear if its impact on research and development is considered, which in certain sensitive

areas might otherwise be discouraged. Evidence already exists, especially in the United States, showing that the liability regime based on absolute liability and disregarding important factors such as the state of technology and the plaintiff's behavior has had adverse effects on new products.

In the United States, the impact of certain court cases based on "absolute producer liability" on insurance costs and the (non)availability of coverage has been tremendous. In the Federal Republic of Germany, the Pharmaceutical Law has not – thus far – generated similar experiences. However, nobody can predict what the consequences would be for the German pharmaceutical industry of a serious "developmental risk" case similar to the thalidomide affair. The compromise achieved in the Directive on the state-of-the-art defense seems to preserve a fair balance and efficiency.

Derogation in Respect of Financial Limitation of Liability. The Directive makes no provision for a Community system of financial ceilings to liability. It does, however, allow Member States to introduce on an individual basis a liability "ceiling" for personal injury, which may then not be less than 70×10^6 ECU (European Currency Units). The German and Belgian laws have, for example, introduced such ceilings.

The Commission has been instructed to submit a report to the Council ten years after the date of notification of the Directive (1995). Acting in the light of this report, the Council is to decide whether or not to cancel the option of introducing ceilings. As with the state-of-the-art issue, such a provision will mirror the Council's political embarrassment at the time of its adoption. The Council will, however, be free to decide to maintain the state-of-the-art and the ceilings option if a political majority so wishes.

Derogation in Respect of Agricultural Products. The Directive normally excludes agricultural products from its scope. However, Member States are authorized to include them within their laws and to apply the regime of the Directive.

3.3.4. Other important Aspects of the Directive

The Directive covers not only personal injury but damage to property (with a "lower threshold" of 500 ECU in the second stage). This provision has led to some divergent interpretations.

Definition of Producer. "Producer" is extensively defined in the Directive as

1) The manufacturer of a finished product
2) The producer of any raw material
3) The manufacturer of a component part
4) Any person who, by putting his name, trademark, or other distinguishing feature on the product, presents himself as its producer

Clearly, the concern of the Directive's drafters has been to avoid the possibility of an injured person finding himself in a situation for which no one is responsible. Hence, care is taken in defining the producer. So true is this that the Directive provides for treating the importer into the Community as a producer within the meaning of the Directive and, should it prove impossible to identify the actual producer, treating the supplier as the producer.

Product Defectiveness. For the purposes of the Directive, a product is defective not according to objective or material criteria but with reference to the "safety which a person is entitled to expect." Thus, § 6 of the recital states that defectiveness does not mean the product's lack of fitness for its intended use but the lack of safety that the public at large is entitled to expect. Such a reference creates difficulties of interpretation. In reality, it will be for the judge to assess. However, the Directive provides that all circumstances be taken into account in assessing the safety the product should provide.

Some of the criteria to be considered in assessing the safety of a product are: (1) the presentation of the product; (2) the use to which it could reasonably be expected to be put; and (3) the time when the product was put in circulation. Safety is to be assessed by taking into account the expected use of the product, its marketing, the information given and user instructions, and similar factors related to the product.

Conclusion. The new Community regime on liability for defective products contains both pros and cons and seems generally balanced. Despite imperfections and weaknesses, it genuinely contributes to the progressive establishment of similar legislations in Europe: one major weakness is that it leaves open the question of diverging procedural rules in the judicial systems of the Member States.

Another important question is the possible impact of other more recent Community legislation, i.e., the Product Safety Directive and the proposed Directive on Liability of Suppliers of Services on the Product Liability System.

The Product Safety Directive was adopted on June 29, 1992 and had to be implemented by the Member States in June 1994. The Directive was challenged on grounds of the Maastricht Treaty "subsidiarity" principle by the Federal Republic of Germany. A very wide and strict "general safety requirement" is introduced, although the Directive does not apply to sectors that are already subject to specific legislation. The Directive introduces a general obligation to market "safe products" only, which is not the same concept as a (non)defective product of the Product Liability Directive. Other provisions address the following issues:

1) Monitoring of products
2) Provision of information to consumers concerning possible risks involved in use of the products
3) Adoption of measures to ensure information of consumers
4) Ensuring uniform management of tests, checks, warnings, prohibitions, and market withdrawals

In practice, this legislation should be seen in relation to the already existing comprehensive laws on safety and should contribute to the highest safety standard for products marketed in the European Union.

The proposed Directive on the Liability of Suppliers of Services would make the supplier liable for damage to the health and physical integrity of persons or to property caused by a fault committed by the supplier. However, the proposal "revises the burden of proof" since the supplier would have to establish the absence of fault. The prospects of the Directive being adopted were not very good in 1994 because most Member States were reluctant to accept Community legislation in this area.

The practical impact of the new Community regime, particularly on insurance costs, has so far not been dramatic, but it is still uncertain and will depend on the content of national laws and their application by national courts under the control of the European Court of Justice.

3.4. Community-Proposed Directive on Liability for Waste

Implementation of the polluter-pays principle provides for consideration of the scope of civil liability to better define responsibility in the environmental field. So as to answer this requirement, also expressed in the previous Directive on the control of transfrontier movements of waste (84/631/EEC) and several other Community documents, the Commission in 1989 submitted a proposal for a Directive on civil liability for damage caused by waste. This proposal was amended two years later, in 1991 [46].

The proposal raised a number of arguments, one of them being that the Community Directive on civil liability should be compatible with the UNIDROIT Convention and, on a broader level, with the approach developed in the Council of Europe Convention. With the content of the amended proposal, this compatibility did not seem ensured. For example, contrary to the international conventions mentioned, the proposal channelled liability onto the waste producer regardless of whether or not the producer was in control of waste. To create a system that is workable and practicable, all other equivalent instruments must be studied and considered. Although the proposal was put aside pending the results of comprehensive discussions on the basis of the Green Paper on Remedying Environmental Damage, the main aspects of the draft Directive are described.

3.4.1. The "Channelling" of Liability

Article 3 of the proposal from 1991 poses the principle of the liability of the waste producer for damage and impairment of the environment caused by waste, irrespective of fault on the part

of the producer. Simultaneously, the proposal is intended to implement the polluter-pays principle. However, this principle – the legal ambit of which remains ill-defined – does not necessarily imply a strict liability regime. A liability regime based on the polluter's fault would be compatible with this principle if the regime ensures that the polluter actually suffers the consequences of the pollution. The Commission's choice of a strict liability regime channelled onto the producer is explained by the concern to use civil liability as an instrument to eliminate or reduce waste. Such an objective, however, has no direct relation to the function of civil liability, which is to ensure compensation of damage arising from wrong use or handling of products whether they are waste or not.

A positive feature of the proposal is Article 2.2.C., providing a transfer of liability whenever the producer delivers waste to a licensed or approved treatment or disposal facility. Applying the same logic, it is difficult to see how a similar transfer of liability would not occur upon delivery to a licensed carrier.

3.4.2. Definition of the Producer

The Commission proposal refers to any person who, in the course of a commercial or industrial activity, produces waste and/or anyone who carries out preprocessing, mixing, or other operations resulting in a change in the nature or composition of this waste. Article 2(2), however, states that the following are to be deemed the producer of the waste:

"The person who imports the waste into the Community;

The person who had actual control of waste when the incident giving rise to the damage to or impairment of the environment occurred if the producer is not identified;

The person responsible for the installation, establishment or undertaking where the waste was lawfully transferred to such installation, establishment or undertaking licensed pursuant to Directive 75/442/EEC on waste (amended thereafter), 75/439/EEC on waste oils or 78/319/EEC on hazardous wastes (amended thereafter)."

The method chosen in the draft leads to a complex definition based on a system of references to previous Directives, all of them providing for treatment, storage, and disposal operations subject to licensing. This leads to a contradiction in the underlying philosophy when it "channels" liability onto the producer while acknowledging the responsibility of the operator or the person having "actual control" of the waste, but only in cases where the producer has not been identified.

3.4.3. The Concept of Waste

The Commission proposal refers to the definition of the original waste Directive 75/442/EEC, which was amended in 1991. Waste is defined under Article 1 (a) of the amended Directive (91/156/EEC) as any substance or object in the categories set out in Annex 1 that the holder discards or intends or is required to discard. The concept of waste is rather broad by definition but excludes, for instance, gaseous effluents emitted to the atmosphere, radioactive waste, waste arising from mining operations, and decommissioned explosives. Given the fact that the definition is also fundamentally based on a subjective approach – the intent of the holder is decisive – judging whether or not a certain material is waste is far from simple. The categories set out in the mentioned annex do not make this easier, giving only rough examples of what should be considered waste.

Problems related to the definition of waste would complicate implementation of the Directive proposal. Different interpretations of waste in various Member States might also distort competition. Problems related to the definition have been under discussion for years, but a satisfying solution has not yet been found. Finally, one more question is still to be raised: Why is a harmonization of civil liability laws required especially for damage caused by waste? Would that not create an uneven situation when other possible sources of damage are left untouched?

3.4.4. Insurance and Compensation Fund

Article 11 of the draft Directive states the obligation of the producer, who produces waste in the course of a commercial or industrial activity, and of the eliminator of waste to be covered by insurance or other financial security. The article

also provides for the Commission to study the feasibility of establishing a European fund for compensating damage caused by waste when the person liable cannot be identified or is incapable of providing full compensation.

3.5. Civil Liability for Damage Caused during the Carriage of Dangerous Goods by Land: The Convention of October 10, 1989

The International Institute for the Unification of Private Law and the United Nations Economic Commission for Europe have actively cooperated in framing an international instrument on liability and compensation for damage caused during the carriage of dangerous goods by road, rail, and inland waterways. Prepared within the Institute of Rome, the draft was subsequently forwarded to the UN Economic Commission for Europe. CEFIC, along with other international organizations representing transport and insurance interests, participated as an observer in the work of UNIDROIT and the United Nations. The draft convention was finalized at the extraordinary session of the Inland Transport Committee of the UN Economic Commission for Europe, held in Geneva from October 2 – 10, 1989.

The main thrust of the Convention rightly establishes the linkage between civil liability and operational responsibility. This last, dated October 10, 1989, has been open for signature by the states since February 1990. Five ratifications are required for it to come into force. Among the countries that expressed at that time their active support are West European countries (especially The Netherlands, Denmark, the Federal Republic of Germany, and France with some reservations, Finland, and Austria) and also Central and Eastern European countries, and the United States.

The salient features of the Convention are as follows:

1) A "strict" liability regime is in effect.
2) For the purposes of the Convention, carriage comprises the period from the beginning of loading of goods to the completion of unloading operations.
3) Liability is channelled onto the carrier (save in the case of loading or unloading when such operations are carried out under the control of another person).
4) Reference is made to the ADR list for the definition of substances to which the Convention will apply (including hazardous wastes).
5) The principle of compulsory insurance (or any system of equivalent financial security) exists.
6) Total financial liability of the carrier (road, rail) is limited to a ceiling of 30×10^6 € per vehicle (18×10^6 € for personal injury; 12×10^6 € for damage to property and the environment). In the event of the "personal injury" ceiling being used up, the amounts of unmet claims for personal injury could compete with claims for damage to property within the "damage to property" ceiling. The liability ceiling specified for inland waterway transport is 15×10^6 € (8×10^6 € for personal injury; 7×10^6 € for other damage).
7) The carrier (road, waterway) is defined as being the person having use of the vehicle. The carrier is presumed to be the person on behalf of whom the vehicle was registered. This presumption may be reversed.
8) The rail carrier is defined as being the person(s) operating the railway on which the incident giving rise to damage occurred.
9) Grounds for exemption from liability are based on:
   Lack of information by the shipper or any other person as to the nature of the goods
   The fact that a third party acted or failed to act with intent to cause damage
   Deliberate fault of the injured party
   Occurrence of certain exceptional events: act of war, civil war, or natural phenomena of an exceptional and irresistible nature
10) The definition of "damage" is extended to cover ecological damage in addition to the more traditional categories of "death or personal injury" and "damage to property." Ecological damage is defined as "any loss or damage by contamination of the environment caused by dangerous substances, provided that the compensation for impairment of the environment . . . is limited to costs of reasonable measures of reinstate-

ment actually undertaken and to be undertaken."

The states will, however, be able to introduce some reservations if they so wish (particularly concerning defenses and the limiting of liability). From the economic and practical standpoints, it does not look as though the Convention will result in upheaval in land or rail transport. The Conseil des Bureaux (international insurers' body designed to facilitate the settlement of damage caused by automobile traffic – "green card system") has stated that the "green card" system will adapt fairly easily to the introduction of the Convention into the laws of Member States. As far as rail is concerned, the vast majority of railway companies have their own systems of insurance coverage. The Convention is, by contrast, a novelty for waterway transport, for which no compulsory insurance regime has existed up to now.

The authors of the draft have forwarded a resolution to the International Maritime Organization for a similar convention to be adopted for maritime transport, thereby complementing the work carried out in regard to land transport.

3.6. Liability in Connection with the Carriage of Dangerous Substances by Sea (HNS)

The adoption of an international draft convention on civil liability has for some years been included in the agenda of the IMO Legal Committee. An initial draft based on a "two-tier" system of liability – both complicated and costly – failed to achieve adoption in 1984. Despite this setback, the International Maritime Organization still sees the issue as a priority. The IMO Legal Committee is discussing several alternatives:

Alternative 1:
Sole liability of the shipowner coupled with a general raising of amounts of the total limitation stipulated in the 1976 Convention on the Limitation of Liability for Maritime Claims (CLLMC).

Alternative 2:
Sole liability of the shipowner coupled with an additional specific amount provided for in the CLLMC Convention for damage caused by HNS substances.

Alternative 3:
Shipowner liability limited in accordance with the provisions of the CLLMC and complemented by compulsory insurance, the cost of which would be borne by the shipper.

Alternative 4:
Shipowner liability limited in accordance with the provisions of the CLLMC and complemented by a fund financed by the shipper through either a pre- or a post-event financing mechanism.

Other options are possible, including the adoption, envisaged at the beginning of 1994, of a maritime carrier liability regime complemented by setting up an international compensation scheme, to come into operation beyond the limits provided by the shipowners' liability coverage or when the person responsible is not solvent. The system would be based on financial contributions calculated as a function of the type of cargoes and their hazard. Various options have been discussed, including the possibility of requesting importers, not shippers, to contribute. In any event, users of transport services would be asked to contribute.

The issue of adoption by the IMO of an international convention on liability and compensation is expected to be addressed in a 1996 diplomatic conference and will remain an important priority in years to come.

3.7. The Council of Europe Convention on Compensation for Damage Caused by Dangerous Activities

3.7.1. Background

In the 1986 Oslo Conference of European Ministers of Justice, the decision was taken to consider options to extend the reach of civil law to promoting environmental protection. Accordingly, the European Council of Ministers Committee set up a committee of experts in 1987 to propose measures for compensation for damage caused to the environment.

After its sixth meeting, the committee of experts instructed the Secretariat General and the Legal Affairs Directorate of the Council of Europe to prepare an interim activity report. This report was to be submitted to the European Committee on Legal Cooperation (CDCJ), a senior Council of Europe body in the legal field, and to Member States of the Council. The report enclosed preliminary draft rules on compensation for damage resulting from dangerous activities. The content of the Convention was adopted by the European Council of Ministers on March 24, 1993. The Convention was opened for signature on June 21, 1993 in Lugano and was signed by seven Council of Europe Member States (Cyprus, Greece, Finland, Italy, Liechtenstein, Luxembourg, and The Netherlands). Other countries were expected to follow suit, and debate within the European Union concerning the Commission's Green Paper has directly influenced this process.

The main aim of the Convention is to ensure prompt and effective compensation of damage to people, property, and the environment during the exercise of dangerous activities carried out in a professional capacity in installations or on sites. As stated in Article 1 of the Convention, it aims at providing adequate compensation for damage resulting from activities that are dangerous to the environment and also provides means of prevention and reinstatement.

3.7.2. Key Features of the Council of Europe Convention

3.7.2.1. Liability Regime

The Convention provides for a regime involving strict (no-fault) liability of "the operator of a dangerous activity." The operator's responsibility is not based on the concept of fault, but on the causal link existing between the damage caused and the incident occurring at the time or during the period when the operator was in charge of a dangerous activity. This general approach is in line with the one persistently advocated by the author and CEFIC, which have always wanted to see a strict link maintained between civil liability for damage and operational responsibility. The explanatory report to the Convention rightly points out major reasons for making the person in charge of the activity from which the damage resulted bear the cost of damage; the activity of this person is the source of the damage, and this person is best placed to prevent the damage and limit its extent.

3.7.2.2. Concept of Operator

The operator, as already discussed above, is defined as being "the person who exercises the control of a dangerous activity." Essential in determining who is in charge of a certain activity is to consider who has the power to decide on the way in which the activity is carried out.

3.7.2.3. Incident

Incident is defined as any sudden occurrence or any series of occurrences having the same origin that cause damage or create a grave and imminent threat of damage.

3.7.2.4. Concept of Dangerous Activity

Dangerous activity is defined as being "one or more of the following activities provided that it is performed professionally, including activities conducted by public authorities:

1) The production, handling, storage, use, or discharge of one or more dangerous substances or any operation of a similar nature dealing with such substances;
2) The production, culturing, handling, storage, use, destruction, disposal, release, or any other operation dealing with one or more:
   - Genetically modified organisms which as a result of the properties of the organism, the genetic modification and the conditions under which the operation is exercised, pose a significant risk for man, the environment or property;
   - Microorganisms which as a result of their properties and the conditions under which the operation is exercised pose a significant risk for man, the environment or property, such as those microorganisms which are pathogenic or which produce toxins;

3) The operation of an installation or site for the incineration, treatment, handling, or recycling of waste, such as those installations or sites specified in Annex II, provided that the quantities involved pose a significant risk for man, the environment or property;
4) The operation of a site for the permanent deposit of waste.

Dangerous substances , in turn, are defined as follows:

1) Substances or preparations which have properties which constitute a significant risk for man, the environment, or property. A substance or preparation which is explosive, oxidizing, extremely flammable, highly flammable, very toxic, toxic, harmful, corrosive, irritant, sensitizing, carcinogenic, mutagenic, toxic for reproduction, or dangerous for the environment within the meaning of Annex I, Part A to the Convention shall in any event be deemed to constitute such a risk;
2) Substances specified in Annex I, Part B to the Convention. Without prejudice to the application of sub-paragraph a above, Annex I, Part B may restrict the specification of dangerous substances to certain quantities or concentrations, certain risks or certain situations."

The operators of waste installations and sites for permanent waste deposit are explicitly subject to the new system. However, specific transitional provisions and liability rules are to be applied in the case of permanent waste deposit.

3.7.2.5. Concept of Damage

The concept of damage is defined as follows:

1) Loss of life or personal injury
2) Loss of or damage to property other than to the installation itself or property held under the control of the operator, at the site of the dangerous activity
3) Loss or damage by impairment of the environment in so far as this is not considered damage within the meaning of the preceding subparagraphs, and provided that compensation for impairment of the environment, other than for loss of profit from such impairment, is limited to the costs of measures of reinstatement actually undertaken or to be undertaken
4) The costs of preventive measures and any loss or damage caused by preventive measures, to the extent that the loss or damage referred to in subparagraphs (1) to (3) arises out of or results from the hazardous properties of the dangerous substances, genetically modified organisms, or waste.

The Convention covers legal categories traditionally included under the concept of "damage" in civil law (i.e., loss of life or personal injury). Additionally, the text introduces the concept of loss or damage by impairment of the environment, provided compensation is limited to the costs of measures of reinstatement actually undertaken or to be undertaken. Since "the environment" is defined in a remarkably broad way, the Convention potentially goes beyond the precedents established in the UNIDROIT Convention or in the EC draft Directive on liability for waste.

3.7.2.6. Environment

As already mentioned, the word "environment," in the confines of the Convention, includes the following factors:

1) Natural resources both abiotic and biotic, such as air, water, soil, fauna, and flora and the interaction among these factors
2) Property that forms part of the cultural heritage
3) Characteristic aspects of the landscape

The definition contained in the Convention is not exhaustive. On the contrary, the explanatory report to the Convention explicitly states that damage to any other aspect of the environment may also give rise to liability under the Convention. Even without this note, the definition is considered broad. Point (1) covers the factors traditionally understood to constitute "the environment." Points (2) and (3) refer to more immaterial values.

A legal definition of the environment (or damage to the environment) will influence the process of determining the type and scope of the necessary remedial action. The Commission Green Paper on Remedying Environmental

Damage refers to this definition of the environment and also notes the problems related to the matter.

3.7.2.7. Right of Action of Private Groups

Article 18 of the Convention to some extent authorizes actions by private associations that according to their statutes, aim at the protection of the environment, while it restricts the scope of such actions before the courts:

1) Requests for prohibition of any unlawful dangerous activity posing a grave threat of damage to the environment
2) Requests for an order to the operator to prevent an incident or damage
3) Requests for an order after an incident to prevent damage or reinstate the environment

3.7.2.8. Causal Link and Administration of Proof

According to Article 10, while maintaining the requirement of a causal relationship between the incident giving rise to damage and the activity of the operator, the court "shall take due account of the increased danger of causing such damage inherent in the dangerous activitiy."

3.7.2.9. Financial Security Scheme

According to Article 12 of the Convention, each Member State shall ensure that where appropriate, taking due account of the risks of the activity, operators conducting a dangerous activity on its territory be required to participate in a financial security scheme or to have and maintain a financial guarantee up to a certain limit, of such type and terms as specified by internal law to cover the liability under this Convention.

The article leaves a lot of freedom for national legislators, because it provides for a financial security scheme for operators covered by the Convention where appropriate, of the type and to such a limit as wanted, respectively. This solution is to be welcomed as leaving Member States free in a matter that does not require strictly prescribed international rules.

3.7.2.10. Access to Information

A system of access to information held by authorities or operators is also contained in the Convention. The same system explicitly recognizes the right to protect the confidentiality of data. The system is based broadly on the provisions of Community law governing the subject.

3.8. The 1993 European Commission Green Paper on Remedying Environmental Damage

In addition to international organizations, new civil liability regimes have been enacted by various EC Member States. In its Green Paper on Remedying Environmental Damage of 1993, the European Commission assumed that different approaches adopted might lead to distortion of competition and hamper the smooth operation of the single market. For this reason, the Commission suggested that a Community system of civil liability in relation to activities dangerous to the environment was necessary. Civil liability is also considered to belong to the domain of the Community legislation, since it presents many interfaces with EC environmental policy.

The purpose of the Green Paper, adopted in March 1993, is to inspire discussion on whether and how requirements to remedy environmental damage might be introduced appropriately and effectively within the Community to recover the costs of such restoration [47]. As its name envisages, the communication is a discussion document that does not formally propose a certain regime on the matter. Instead, it presents alternatives for the legislator and summarizes different arguments for and against different solutions. Nonetheless, the document outlines the main principles as follows:

1) The Commission aims at harmonization of civil liability systems for environmental damage in cases when a liable party can be identified, and it expresses a clear preference for strict (objective) liability, e.g., as defined in the Council of Europe Convention
2) Civil liability is not seen as a means for compensating damage when no liable party can be found; thus, the Green Paper looks for another kind of compensation mechanism, such as a collective compensation fund

Therefore, the Commission aims at a strict liability regime combined with a nonliability-related system (i.e., the cost of repair could be shared among several economic sectors).

The Green Paper suggests that such a joint compensation mechanism is needed in the event of the following:

1) Damage is unbounded or latent
2) Cumulative acts or incidents have occurred
3) Liable parties cannot be identified
4) No basis exists for liability
5) No causal link is determinable
6) No party has a legal interest to bring action

Questions to be answered are on whom the burden to contribute to such funds should rest and whether such funds are necessary in addition to existing mechanisms. The Green Paper does not provide the answer but asks whether the polluter-pays principle should be respected to the greatest possible degree so that contributions are requested from "the economic sectors most closely linked to the type of damage needing restoration." However, a number of difficulties exist in interpretation of the principle and in identification of those sectors. For example, should the "sector" that is today most closely linked to the damage pay for damage caused by operators in the past? In addition, the identification of sectors that are most closely linked to the damage raises serious principle and technical objections. It can be argued that postpollution, unrelated to faulty or illegal pollution, is the result of a given stage of development of society. Therefore, collective and public law answers and compensation are more suitable for such "historic" pollution.

Apart from the U.S. Superfund [Comprehensive Environmental Response Compensation Liability Act 1980 (CERCLA)] and the funds established by the oil industry [Brussels Convention on the Establishment of an International FUND for Compensation for Oil Pollution Damage (1971) and parallel financial structures, such as TOVALOP (Tankers Owners Voluntary Agreement Concerning Liability for Oil Pollution), CHRISTAL (Contract Regarding Interim Supplement to Tanker Liability for Oil Pollution), and OPOL (The Offshore Pollution Liability Agreement)], such compensation schemes are still unusual and of minor significance. In the Netherlands, a fund on damage from air pollution was created in 1972. In France, a fund for noise, which compensates persons living around Paris airports, was created in 1973. The U.S. Superfund, as is widely agreed, has proved inefficient and unfair and therefore should not serve as a model. Solutions for the oil industry, in turn, serve only for purposes characteristic of that industry (i.e., those resulting from major transport accidents).

The Green Paper considers a number of issues related to civil liability and aims at stimulating discussions among all interested parties. Therefore, it has a major role in opening discussions on the fundamental principles of Community environmental policy and also on the Community's role and need for harmonization in this area. The Council of Europe Convention recently adopted will, without doubt, have a great influence on the discussions since although it is binding on only a few states so far, it is the first existing international legal instrument covering civil liability in the context of dangerous activities for environmental damage as well as damage to persons and property.

The communication does not suggest any definition of environmental damage but merely refers to definitions contained in the Council of Europe Convention and in the amended proposal for a Council Directive for damage caused by waste. However, the fundamental importance of and difficulties related to the definition are emphasized.

3.9. Economic Aspects and Questions Bearing on the Insurance of the Industrial Risk

In the last few years, industry has encountered growing difficulties in renewing its "civil liability operation" insurance policies, in terms of both the cost of premiums and the extent of risks covered. Insurance companies, as well as the reinsurance market, have been very hesitant, chiefly for the following reasons:

1) The impact of the U.S. liability crisis, characterized by the number and magnitude of claims for damages
2) Fear of a similar situation developing in Europe, particularly in view of the introduction

of new liability laws and increasingly stringent technical Regulations
3) The potential liability of businesses involved in actions understood to be based, among other things, on damage to the environment whose cause is uncertain
4) Anticipated development at the national and international levels of new laws that would deal specifically with liability for pollution

This situation has prompted industry to study new additional methods of risk financing. In the United States, chemical and insurance companies alike have created insurance pools to increase available capacity.

In Europe, industry has found it increasingly difficult to find appropriate coverage, especially for "gradual pollution." Studies are therefore being undertaken to assess the "feasibility" of and arrangements for setting up a European Industrial Insurance Pool designed to provide added capacity and better conditions. Coverage of major industrial accident risks and of the risk of environmental pollution is reportedly being envisaged in particular. However, the actual adoption and operation of a pool depend largely on the development of the traditional insurance market. In many cases, conditions surrounding the renewal of policies can be said to have improved in recent years. One reason for this favorable trend could be that the envisaged creation of an insurance pool under direct industry control has served to make the insurance market more buoyant. The concept of a pool has rightly been perceived as mirroring the grave concern of industry and as offering an alternative to insurance mechanisms that no longer appear to be meeting industry's demand adequately.

The insurance crisis has not disappeared, however. Its acuteness has merely been mitigated a little. Major incidents or legislative initiatives unreasonably extending the scope of industrial civil liability could adversely affect the world insurance and reinsurance market, especially in the area of pollution.

3.10. Conclusions

Seeking harmonization of civil liability laws for all European Community member countries was the basis of the "product liability" Directive. Harmonization, or at least approximation of laws regarding liability for dangerous activities, was offered as a working hypothesis in the Green Paper published by the European Commission in 1993.

The issue could, however, be addressed in an even wider context, given that other economic or geographic areas might be affected by the transfrontier effects of pollution, as illustrated by major industrial accidents outside EC territory. With regard to the European Community, a harmonization of national laws is to be supported, if it eliminates trade distortions and simplifies existing provisions. A harmonization of this kind could, moreover, be achieved in a broader framework than that of the Community. However, real harmonization – a simplification of laws and a reduction of potential conflicts – is a most difficult and challenging task for lawmakers. No genuine harmonization can be achieved in this area without parallel progress in harmonization of the procedural rules governing civil liability. Therefore, this is the beginning, not the end, of a complex process.

The U.S. liability and insurance crisis has demonstrated that the dilution of basic legal requirements related, for example, to causation and evidence can lead to complete economic and legal failure. The question of need, justification, and possible creation of nonliability-related compensation mechanisms will have to be addressed carefully in the coming years, especially in relation to compensation for "historic" gradual nonattributable pollution damage and old dumping sites. Public authorities, the chemical industry, the insurance world, and the public at large have a joint interest in finding consensus-based solutions in a most important societal issue as the 20th century draws to a close.

4. Waste Shipments in the European Union: The Interface between Environmental Regulation and the Internal Market

4.1. Introduction

The last few years have seen an increase in the interaction between international trade policies and environmental protection. The first international convention on this subject was the

1933 Convention Relative to the Preservation of Fauna and Flora in Their Natural State. However, until the late 1980s, environmental regulators were developing their policies in large part separately from economic or trade considerations. This situation is changing rapidly, and today almost every international, intergovernmental organization has a program on trade and the environment. Moreover, environmental considerations are addressed as integral components of trade policy decisions, and trade or economic implications of environmental regulations are, conversely, being seen as key components of environmental regulatory decisions as well.

What is responsible for this dramatic change? First has been an increased recognition of the need to focus on transnational pollution issues such as the question of carbon dioxide emissions, chlorofluorocarbons (CFCs), and industrial accidents, for example, the Sandoz Rhine river pollution case. These transnational issues are increasingly addressed by multilateral environmental agreements, which also contain trade provisions. One example is the Montreal Protocol on CFCs of the United Nations Environment Programme (UNEP, 1987). Another example is the Basel Convention on Control of Transboundary Movements of Hazardous Wastes (UNEP, 1989).

In addition, focus has increased on the impact on competitiveness caused by some countries' having strict environmental standards or by attempts by such countries to erect barriers against imports produced in countries having less stringent environmental standards. This has led international, intergovernmental organizations, such as the General Agreement on Tariffs and Trade (GATT), to address the danger of countries taking unilateral actions under the guise of protection of environment, health, or safety, with the effect (and perhaps even the intention) of discriminating against imports. Included here would be such unilateral actions as discriminatory product standards, testing/certification requirements, and environmental subsidies.

A series of GATT disputes have been related directly to environmental regulation brought under the GATT nondiscrimination and national treatment provisions. In addition, the GATT, now the World Trade Organization (WTO), has instituted a Trade and Environment Program to facilitate the enforcement of "free" trade principles while simultaneously encouraging sustainable development.

Related subjects involve the effect of environmental policies of the developed or "Northern" countries on developing or "Southern" countries. Complicated issues of market access, technology transfer, commodity pricing of natural resources, debt levels, paternalism, and protectionism must be addressed here.

Regrettably, writings on the subject of free trade and environmental regulation tend to represent one polarized view or the other, either environmental protection or free trade. On the *free trade side*, the position is that environmental objectives do not justify violating the multilateral trading system rules of national and nondiscriminatory treatment. Moreover, the free traders believe that the sovereignty of countries in setting their own environmental priorities and requirements should be respected; that environmental policy should not impede the free flow of goods, services, persons, and capital; and that international environmental issues, while perhaps a component of trade policy, should be subservient to trade policy in the event of a dispute.

In contrast, *environmentalists* generally believe that economic expansion leads to environmental degradation; that harmonized standards reduce environmental protection; that unilateral trade-restrictive actions are valid or even desirable to promote domestic or global environmental objectives; and that environmental subsidies and other incentives or disincentives should be encouraged. The environmentalist perspective is not concerned with what it perceives as "outdated" issues of nondiscriminatory trade. Protection of the environment justifies almost any measures necessary.

4.2. European Community/Union Measures

In the European Community/Union, the interrelationship between trade and environmental regulation is best seen in the development of the Internal Market.

4.2.1. The Internal Market, "1992"

The Treaty of Rome was enacted in 1957 to establish the European Economic Community

and to guarantee, among members of the Community, four basic freedoms: the free movement of goods, persons, services, and capital between Member States.

The concept of 1992, "the Internal Market," was one of the best public relations coups in recent political history. Under the Treaty of Rome and other early documents of the European Economic Community, a single market and the elimination of physical, technical, and fiscal barriers (explanation see below) between Member States of the Community had already been anticipated. However, discouraged by the slow progress toward such a single market, the Community developed the 1992 program to accelerate its progress.

The first major document in the 1992 process was the White Paper on Completing the Internal Market [49]. It was published in 1985 and included a variety of proposals for legislation and timetables in which the 300-plus measures necessary to eliminate barriers to the free movement of persons, goods, services, and capital in the Community had to be enacted by the end of 1992. The barriers the Community was trying to eliminate among its Member States were physical, technical, and fiscal ones.

Physical barriers are such things as customs services, passports, and transport formalities. *Technical barriers* include health and safety requirements, standards, and product packaging. *Fiscal barriers* are among the most difficult to eliminate. Included here are harmonization of general tax rates and value-added tax (VAT) rates. Of the 300-plus measures necessary to implement the Internal Market, between one-third and one-half were environmental, health, and safety initiatives.

4.2.2. The Single European Act (see Section 2.1)

The Single European Act took effect on July 1, 1987 (Single European Act, signed at Luxembourg on February 17, 1986, and at The Hague on February 28, 1986). It contains numerous provisions amending the Treaty of Rome, many of which were necessary for the 1992 process to be successful. The most significant provisions affecting trade and waste shipments are the following:

1) Inclusion of a new Article 100 A in the Treaty of Rome providing that harmonization measures can generally be adopted by the Council by a qualified majority vote, rather than by unanimity as before
2) Inclusion for the first time in the Treaty of Rome of provisions specifically relating to environmental protection
3) Revision of the consultation procedure with the Parliament, giving Parliament a stronger role in the policy-making process

New Article 100 A of the Treaty of Rome (Article added by Article 18 of the Single European Act) provides that measures can be adopted by the Council by qualified majority, if their objective is the approximation of the provisions of national laws in order to further the establishment and functioning of the Internal Market, i.e., harmonization measures (see Section 2.1).

In the past, Council decisions had to be reached on a unanimous basis. As result, many Council actions, especially controversial ones, were deadlocked because one or another Member State could not agree with the rest. To facilitate action on measures necessary to implement the Internal Market, a new Article 100 A was included to permit the Council to act on Internal Market measures by qualified majority vote. (The votes of the Council are weighted. The total number of votes among the 12 Member States is 76; a qualified majority consists of 54 votes.) Exemptions from qualified majority voting exist here for measures related to fiscal provisions; worker rights issues; as well as health, safety, environmental, or consumer protection issues. Here, unanimity still had to prevail under the Single European Act.

Enactment of Community legislation by unanimous Council votes is usually more complicated and time consuming than reaching qualified majority decisions. One of the early steps in the EC legislative process under the Single European Act, therefore, always involves "jockeying for position" between the European Commission and Member States to determine the legislative basis of the measure in question. A dispute on the legal basis of a piece of Community legislation concerning titanium dioxide went as far as the European Court of Justice [50]. In this case, the Court of Justice ruled on the legal basis of Community legislation that was aimed both

at the achievement of the Internal Market and at environmental protection. The Court upheld the legal basis of the legislation as an Internal Market, rather than environmental, measure and annulled the Directive. The Court ruled that the choice of legal basis should be governed by the aim and content of the legislation in question.

As stated above, the Single European Act also specifically provides that protection of the environment is a valid area of Community legislation. In practice, the Community began legislating in the environmental area in the late 1960s. However, the original Treaty of Rome is primarily a trade agreement and does not specifically provide authority for such legislation. Under the Single European Act, the Community must now consider environmental implications in adopting its other legislation, as well [51]:

"Action by the Community relating to the environment shall be based on the principles that preventive action should be taken, that environmental damage should as a priority be rectified at source, and that the polluter should pay. Environmental protection requirements shall be a component of the Community's other policies."

The Member States were rather suspicious of this amended Treaty language, since environmental, health, and safety regulation had traditionally been within their own purviews. Therefore, the Treaty language now also includes two other provisions: (1) the issue of subsidiarity (see Section 2.1), and (2) unanimous voting on environmental matters. "Subsidiarity," a term that has received much attention in the context of the Maastricht Treaty, in this environmental context, means merely that environmental legislation or enforcement should be undertaken at the lowest feasible regulatory level.

New Article 130 T of the Treaty of Rome [51] specifies that Member States will not be prevented from adopting more stringent protective environmental measures than those enacted at the Community level, provided these measures are compatible with the Treaty of Rome (see Section 2.2). This is very useful for Member States, such as Denmark, the Netherlands, and Germany, that periodically adopt more stringent legislation. At the Community level, there has also been a trend to "harmonize up" (i.e., to harmonize toward more stringent environmental regulation). Invariably, the Community has usually taken existing legislation from one of the more stringently regulated Member States as a model for its own Community-wide legislation.

A number of procedural changes also occurred as a result of the Single European Act, the majority of which increased the power of the European Parliament. The general procedure of parliamentary advisory review for Community actions is one of the Commission proposing legislation and the Council acting on it. Prior to such Council action, Parliament and ECOSOC review the proposal and issue advisory opinions, which the Commission and/or Council can accept or reject.

The Single European Act introduced another procedural step in certain situations, the so-called Cooperation Procedure. Under this procedure, two readings in the Parliament and two Council reviews (under new Article 149, as replaced by Article 7 of the Single European Act) take place.

4.2.3. The Maastricht Treaty – Environmental Provisions [52]

Mention should also be made of the environmental provisions of the Maastricht Treaty. The Maastricht Treaty was agreed to by the Heads of State and Government of the European Community Member States in December 1991, to revise the Treaty of Rome to provide for such changes as gradual political integration, a common currency, and a Single European Bank.

The Maastricht Treaty is notable for its further integration of environmental considerations into all other policies of the European Community; the text goes even further than that of the Single European Act. (Since the entry into force of the Maastricht Treaty changed the name "European Community" to "European Union," henceforth, reference is to either the European Union or the EU, even though the measures described may have been initiated in EC times.) It also contains language on the interrelationship between transportation, energy, and agricultural policies and environmental protection. Moreover, the Treaty deals with the need to consider trade or competition and economic development issues within the framework of environmental considerations. In fact, protection of the environment is explicitly contained in the introductory section of the Maastricht Treaty (i.e.,

impact on the environment should be considered in the development of all other EU policies and the EU has competence to regulate in this area). The concept of "sustainable development" is also found in this Treaty.

The Treaty also eliminates the need for unanimous votes for environmental measures and adds a new "Cooperation Procedure" which expands the role of the European Parliament. In addition, the question of the meaning of "subsidiarity" is supposedly clarified. However, confusion about the exact definition of this term, and when it is to be applied, is still present under this new Treaty.

Finally, the Maastricht Treaty provides for the creation of a "Cohesion Fund" to assist the least financially sound of the European Member States, Ireland, Spain, Portugal, and Greece (including to help them improve environmental protection and enforcement measures).

4.2.4. Other Applicable Provisions of the Treaty of Rome

Notification and Standstill Procedure on Technical Standards and Regulations. Significant European Court of Justice case law exists on "mutual recognition"; i.e., goods lawfully produced and marketed in one EU Member State must be accepted in all other Member States (except where refusal is justified by Articles 30 and 36). Directive 83/189/EEC [53] specifies notification procedures for technical standards and regulations in the EU. The procedure was developed to discourage national, unilateral actions in this area. Member States must inform the Commission on each draft technical standard or regulation. The Commission then informs the other Member States. The Commission and the other Member States have the option of consenting to the draft technical standard or regulation text. The acting Member States must take their positions into account. Waiting periods for adoption of the measure(s) are also specified.

Article 30. Article 30 of the Treaty of Rome provides that "quantitative restrictions on imports and all measures having equivalent effect shall be prohibited." Quantitative restrictions have never constituted a serious obstacle to intra-EU trade. Therefore, the emphasis in practice has been on measures having an "equivalent effect." All legislative rules and administrative provisions, as well as administrative practices that form a barrier to importation or exportation, are governed. Included here are provisions and practices that render importation or exportation more expensive or difficult compared to sales of home products on the domestic market. The *Cassis de Dijon* decision [54] established the basic principles to be applied in judging the legitimacy of measures adopted by EU Member States that hinder the free movement of goods among them. In the absence of harmonized EU law, Member States are free to establish rules affecting the free movement of goods.

However, goods lawfully produced or marketed in one Member State must be accepted in other Member States, except where refusal is based on the "rule of reason" or on exceptions provided in Article 36. The rule-of-reason exceptions are applicable only to noneconomic matters, i.e., (1) environmental protection, (2) consumer protection, (3) prevention of unfair commercial practices or (4) effectiveness of fiscal supervision – provided, however, that the national measure in question is not unreasonable, disproportionate, or protectionist. Exceptions provided in Article 36 include national restraints on trade, if necessary to protect (1) public morality, public policy, or public security; (2) health and life of persons, animals, or plants; (3) national treasures of artistic, historic, or archaeological value; and (4) industrial or commercial property.

However, not all trade restrictions based on the "rule of reason" or on Article 36 exemptions are acceptable. A restriction may not, for example, constitute a means of arbitrary discrimination or disguised restriction on trade between Member States. These concepts are referred to as the principles of "nondiscrimination" and "proportionality;" the restriction must apply equally to nationally produced goods and goods imported from other Member States and must be necessary to achieve legitimate, proportional objectives.

For environmental cases, there should also be a balance between environmental and free trade purposes. A regulatory measure must be necessary to achieve a stated goal, and no less restrictive measures can exist to achieve this goal.

A case in point is the Danish Bottles case [55], which demonstrates the complexities of justifying the compatibility of environmental measures with rules on the free movement of goods. In question was a Danish law providing that mineral water, soft drinks, and beer could be sold only in refillable containers. A deposit and return system to ensure refilling of a large portion of containers was also included. The result of this law was that beverages lawfully sold in other Member States could not be sold in the same containers in Denmark, a clear violation of Article 30 of the Treaty of Rome.

The European Commission took Denmark to the European Court of Justice to contest this legislation. The Court found that the Danish system of mandatory deposit on containers for mineral water, soft drinks, and beer, although placing a greater burden on non-Danish producers, did not infringe Article 30. The Court held that environmental protection is an objective that may justify restraints on trade. The Danish deposit system appeared necessary to attain the desired objective, and no other less restrictive means existed.

A more recent example of one State's unilateral trade action, ostensibly to protect the environment, is the German Packaging Waste Regulation [56]. Under it, a complicated take-back and recycling system for packaging and packaging waste (set up and funded by industry because of the constraints of the Regulation) imposes, in practice, significantly disproportionate problems on importers into Germany than on the domestic industry.

A number of complaints have been filed with the European Commission alleging Germany's violation of Article 30 of the Treaty of Rome. One benefit of the German Regulation is that it has resulted in the EC's accelerating the development of its own Directive on Packaging and Packaging Waste [57] to harmonize all EU Member State rules on this subject. Germany, however, has steadfastly opposed the text of this EC draft Directive and refused to amend its own Regulation. As a result, the European Commission decided, on July 6, 1994, to initiate legal proceedings against Germany because its unilateral packaging Regulation creates an unjustified barrier to trade between EU Member States. In all likelihood, this is merely a political ploy to get Germany to drop its opposition to the EC's draft Packaging and Packaging Waste Directive and to bring the German regulatory requirements for waste packaging into line with the EC Directive's requirements (i.e., reduce the German recycling and reuse targets and build up domestic German recycling capacity more rapidly).

4.3. Waste Shipments: Conflict between European Union Free Movement of Goods Principles and Environmental Regulation

4.3.1. The New Regulation 259/93: Introduction

Recent regulatory initiatives on waste shipments provide perhaps the most dramatic example of the conflict between the European Union's principles of free movement of goods and environmental regulation. On May 6, 1994, the European Union Regulation "On the Supervision and Control of Shipments of Waste within, into and out of the European Community" [58] became effective (hereinafter Regulation 259/93). Regulation 259/93 imposes a complex system of shipment notification and consent on all shipments of waste, not just hazardous waste, that cross borders in the European Union, as well as all waste imports into and waste exports from the EU. Although the notification and consent procedures vary depending on the type of shipment, its destination, and its purpose; waste shipments for recovery or recycling operations are covered, as is waste for disposal.

Regulation 259/93 must be viewed in the context of other key EC waste legislation, most notably, Directive 75/442 EEC on Waste [59], as amended by Directive 91/156/EEC [60]; and Directive 91/689/EEC on Hazardous Waste [61], replacing Directive 78/319 on Dangerous Waste [62]. For these pieces of legislation, see → Waste, Chap. 2. International commitments of the EU on this subject (e.g., Basel Convention) are also interrelated.

On the specific topic of transfrontier movements of waste, the EEC had Regulations as early as 1984, with Council Directive 84/631/EEC, as amended [63], "On the Supervision and Control within the European Community of the Transfrontier Shipments of Hazardous Waste." Directive 84/631 required Mem-

ber States to take measures to ensure the supervision and control of transfrontier shipments of hazardous waste within, into, and out of the Community to protect human health and the environment.

Under Directive 84/631, all cross-border movements of hazardous waste were subject to a notification requirement to be effected by a uniform Consignment Note and typically made to the Competent Authority of the Member State of destination. Procedures for acknowledgment of and objections to the shipment, if any, were provided. However, given the narrow definition of "hazardous waste," Directive 84/631 applied to only a limited number of transfrontier shipments of waste. According to 1992 figures of the Organization for Economic Cooperation and Development, ca. 200×10^6 t of waste cross OECD borders each year (140×10^6 t cross EU borders). Of these 200×10^6 t, < 1 % was estimated by the OECD to involve so- called hazardous waste.

In addition, since this 1984 legislation was a Directive, EC Member States tended to implement it inconsistently. To fill this regulatory gap, address implementation inconsistencies, and adapt the rules regulating transfrontier movements of hazardous wastes to the requirements of the March 22, 1989 Basel Convention on Transboundary Movements of Hazardous Waste and Their Disposal and the Lomé Convention of December 15, 1989 (the Fourth ACP – EEC Convention), the European Council adopted the Regulation "On the Supervision and Control of Shipments of Waste Within, Into and Out of the Community" (Regulation 259/93). Since this legislation is in the form of a Regulation, as opposed to a Directive, it is directly applicable and binding on EU Member States.

Regulation 259/93 was also drafted to ensure consistency with the OECD, March 1992 Council Decision Concerning the Control of Transfrontier Movements of Waste Destined for Recovery Operations, specifically its "Green," "Amber," and "Red" Lists of Waste. These lists are categorized according to hazard ("Green" being least hazardous, with "Red" most hazardous) and appear at Annexes II through IV of Regulation 259/93. Regulation 259/93 was also deemed necessary to ensure continued and improved environmental protection in EU waste management, in view of completion of the Internal Market and its resultant removal of border controls.

Finally, the issues of waste management in general in the EU, and waste shipments in particular, are highly emotional and political ones. Regulation 259/93 is at least partially the result of successful political pressure by environmental groups, some Member State regulatory authorities, and "the public" for stricter and harmonized European Union legislation on this subject.

The overriding purpose of the Regulation is, of course, to ensure prior notification and consent for waste shipments to Competent Authorities, Member States, enabling them to be informed of the particular type, movement, and disposal or recovery of waste, so that they can take all necessary measures for the protection of health and the environment, including the possibility of raising reasoned objections to the shipments (e.g., de facto "PIC" or prior informed consent).

Regulation 259/93 provides for a Consignment Note notification system for the covered waste shipments to the Competent Authorities of destination, dispatch, and transit and to the consignee (see Sections 4.3.2.1 and 4.3.2.2). The content of the notification and the subsequent regulatory avenues available to the Competent Authorities to "control" these shipments depend on the nature and composition of the waste (i.e., its hazard); the shipment's purpose (disposal or recovery); and its destination (to or through other countries in the EU, to non-EU OECD Decision signatories, to other parties to the Basel Convention, to countries having equivalent bilateral or multilateral agreements with EU Member States – the possibilities are numerous).

The bases on which Member States may object to a waste shipment for disposal include the principles of proximity (waste should be disposed of near the point of generation – a policy encouraging local waste disposal); self-sufficiency (to ensure that each Member State can dispose of its waste in its own territory); and priority for recovery. Differing positions have been taken by commentators as to whether the first two of these principles can or should be extended to waste shipments for recovery under the Regulation. Political pressures by local authorities, NIMBY("not-in-my-backyard")-ism by the neighboring public if such an interpretation is

upheld, and resulting negative competitive impacts on industry in the EU can be anticipated if this interpretation is sustained.

4.3.2. Specific Provisions of Regulation 259/93

Exclusions from the scope of Regulation 259/93 are:

1) Offloading to shore of waste generated by normal operation of ships and offshore platforms, including wastewater and residues, provided that such waste is subject to a specific binding international instrument
2) Shipments of civil aviation waste
3) Shipments of radioactive waste
4) Shipments of waste already covered by other relevant legislation
5) Shipments of waste into the EU subject to the environmental protocol of the Antarctic Treaty
6) Shipments of waste contained on Annex II (the OECD Green List of Waste) destined only for recovery, subject to certain conditions

The main text of the Regulation is divided into sections, depending on whether the shipments are between EU Member States or involve imports or exports and, in each section, whether disposal or recovery is involved.

4.3.2.1. Shipments of Waste between EU Member States

Shipments of Waste for Disposal. If a notifier intends to ship waste for disposal from one EU Member State to another or pass it in transit through one or several other Member States, the Competent Authority of destination must be notified, and a copy of the notification must be sent to the Competent Authority of dispatch and of transit and to the consignee. Notification must also cover any intermediate stage of the shipment from the place of dispatch to its final destination. The notification is to be effected by means of a Consignment Note, which is to be issued by the Competent Authority of dispatch. In making the notification, the notifier must complete the Consignment Note and must supply any additional information and/or documentation requested by the Competent Authorities.

Moreover, the notifier must enter into a contractual agreement with the consignee for disposal of the waste. Specific provisions are included as acknowledgments of the Consignment Note and authorization of the shipments. The shipment may be effected only after the notifier has received authorization from the Competent Authority of destination.

To implement the principles of (1) proximity, (2) priority for recovery, and (3) self-sufficiency at the EU and national levels, Member States may take appropriate measures generally or partially to prohibit shipments of waste for disposal.

Shipments of Waste for Recovery. The notification requirements of Regulation 259/93 with respect to shipments of waste between Member States for recovery are generally consistent with the requirements described above for shipments between Member States for purposes of disposal, with certain key deviations. First, the content of the Consignment Note differs. Moreover, tacit consent for a shipment is deemed to exist if 30 days have passed following notification, and no objection has been lodged.

In addition, Competent Authorities with jurisdiction over specific recovery facilities may decide, notwithstanding the notification and consent requirements of the Regulation, that they will not raise objections concerning shipments of certain types of waste to specific recovery facilities. Such decisions may be limited for a specific time period and may be revoked at any time. The Commission must be notified of any such special arrangements; it will inform the Competent Authorities of the other EU Member States and the OECD. However, Member States of transit or dispatch may object.

Shipments of Waste within one Member State. The notification and consent provisions described above do not apply to shipments of waste within one Member State. Here, the individual Member State in question has the role of establishing an appropriate system for the supervision and control of shipments of waste, "coherent with" the system established by Regulation 259/93. The Member States must inform the Commission of these individual systems. However, shipments between Regions are covered if

the Region has jurisdiction over waste management (see also Section 4.3.3.1).

4.3.2.2. Exports of Waste from the EU

Exports for Disposal. Exports of waste from the EU for disposal are prohibited under Regulation 259/93, except to European Free Trade Area (EFTA) countries, which are also parties to the Basel Convention. Exports of waste for disposal to an EFTA country may also be banned if the EFTA country of destination prohibits imports of such waste, if it has not given its written consent to the specific import in question, or if the Competent Authority of dispatch in the EU has reason to believe that the waste will not be managed in accordance with environmentally sound methods in the EFTA country of destination.

The notification and consent requirements with respect to exports of waste from the EU for disposal generally track those described above for intra-Union shipments for disposal, with certain deviations. For example, the notification and consent time periods are significantly expanded here. The Regulation also contains language as to obligations of EU Customs offices of departure.

Exports of Waste for Recovery. Exports from the EU of waste destined for recovery are prohibited under Regulation 259/93, except for those exports to countries to which the OECD Decision applies and other countries that are parties to the Basel Convention (or with which the EU or the EU and its Member States have concluded bilateral, multilateral, or regional agreements/arrangements compatible with the Basel Convention and EU legislation). The Regulation also provides specific requirements that such agreements or arrangements must meet.

In addition to the procedures above, exports of waste from the EU for recovery to countries to which the OECD Decision, Basel Convention, and/or bilateral or multilateral agreements or arrangements apply may also be banned if the country in question prohibits all imports of such waste, if it has not given its consent to the specific import in question, or if the European Union Competent Authority of dispatch has reason to believe that the waste will not be managed in accordance with environmentally sound methods in the country of destination.

Generally, waste specified on Annex II of Regulation 259/93 (OECD Green List of Waste) for recovery is not subject to the notification and consent requirements of the Regulation, again with certain exceptions. First, the Commission is required to inform every country to which the OECD Decision does not apply of the waste categories contained in Annex II and request written confirmation that such waste is not subject to control in the country of destination and that the latter will accept such waste to be shipped without recourse to the control procedures. Annex II waste must be exported for recovery operations within a facility that is locally authorized and operating in the importing country. Moreover, such exports will be subject to surveillance.

As to exports of Annex III and IV waste for recovery (OECD Amber and Red Lists, respectively), the notification and consent procedures discussed earlier with respect to intra-Union shipments generally apply.

Finally, the Regulation provides that all exports of waste to ACP (African Caribbean, Pacific) States will be prohibited. This prohibition does not prevent an EU Member State to which an ACP State has decided to export waste for processing from returning that processed waste to the ACP State of origin.

4.3.2.3. Imports of Waste into the EU

Imports for Disposal. Imports of waste into the EU for disposal will be prohibited, except those from EFTA countries that are Parties to the Basel Convention, other countries Party to the Basel Convention, or countries that have concluded appropriate bilateral or multilateral agreements or arrangements with the EU as a whole or with individual Member States. Such agreements or arrangements must be notified to the European Commission, which will review them.

As in the case in the intra-EU and export shipment situations discussed above, notification of imports of waste for disposal will be subject to the Consignment Note procedure. Notification must be made to the Competent Authority of destination, and the Consignment Note will be issued by that Competent Authority. Again, spe-

cific deadlines for notification, objections, and decisions are provided.

Imports for Recovery. Imports of waste into the EU for recovery are prohibited under Regulation 259/93, with the exception of imports from countries to which the OECD Decision applies and, again, countries Party to the Basel Convention or with which the EU/Member States have concluded appropriate bilateral, multilateral, or regional agreements or arrangements.

The notification and control procedures of Regulation 259/93 also apply to waste imports into the EU for recovery from countries and through countries to which the OECD Decision applies for waste listed on Annexes III and IV of the Regulation or for waste that has not yet been assigned to an Annex.

4.3.2.4. Transit of Waste from Outside and Through the EU for Disposal or Recovery Outside the European Union

Here again, the Consignment Note and control procedures apply. The notification is to be effected by means of the Consignment Note sent to the last Competent Authority of transit within the EU, with copies to the consignee, the other Competent Authorities concerned, and the Customs offices of entry into and departure from the EU.

The shipment may be admitted into the European Union only if the notifier has received the written consent of the last Competent Authority of transit, which authority will also issue the Consignment Note. Specific time deadlines for notifications and decisions are also provided. Transits of waste listed on Annexes III and IV for recovery through one or more EU Member States from a country, and transit of waste for recovery to a country to which the OECD Decision applies require notification of all Competent Authorities of transit in the Member States concerned. Such shipments may be effected only in the absence of objections.

4.3.2.5. Common Provisions

Regulation 259/93 contains certain common provisions applicable to all shipments, whether they are for disposal or recovery, or involve waste imports, intra-EU shipments, or exports from the European Union. These cover general notification possibilities, illegal traffic, financial guarantees and insurance, inability to facilitate recovering disposal and return, regulatory authority designation, documents retention, etc.

4.3.3. The Trade – Free Movement and Environmental Protection Conflicts in Complying with EC Regulation 259/93

4.3.3.1. Definitions

One of the most difficult issues to anticipate in compliance with this new EC Regulation is a definitional one. Even though much progress has been made in the definitions of "waste" and "hazardous waste" in the OECD and in recent EC Directives on the subject, clear EC definitions have still not been developed, and differing EU Member State definitions continue to exist. Therefore, regulatory inconsistencies on this subject can be expected to remain.

One case in point is the German distinction between something having commercial value (Wirtschaftsgut) and waste, and the German treatment of secondary raw materials. However, the new German Federal Comprehensive Waste Reform Law (Kreislaufwirtschaftsgesetz) intends to change German waste definitions to comply with EU definitions. This law was finally passed by the Bundesrat before its summer 1994 recess and is expected to come into effect, at the latest, by 1996.

The EC Waste and Hazardous Waste Directives attempt to define waste in their texts. The definition of waste in the 1991 Waste Directive (91/156/EEC) is as follows:

"Waste" shall mean any substance or object in the categories set out in Annex I which the holder discards or intends or is required to discard" (Article 1).

The "hazardous" waste definition under the 1991 Hazardous Waste Directive (91/689/EEC) is much more complicated and includes various references to Annexes, origin, composition, properties, and a final umbrella category of "any other waste" considered by a Member State to be hazardous, in effect negating the usefulness of the harmonized European-level definition.

In addition, both the 1991 Waste and the Hazardous Waste Directives mandate the compilation of lists or catalogues of waste under their domain. The EU has published its Waste Catalogue [64] under the 1975 and 1991 Waste Directives [65], and the so-called Waste Catalogue II or List of Hazardous Wastes under the 1991 Hazardous Waste Directive was published on December 31, 1994 [66].

The process of developing both European Commission lists has been highly controversial. Officials at the Commission have taken varying positions as to whether these should be listings of products and substances or whether they should, instead, be criteria-based. Moreover, a number of key Member States, most notably Germany and France, are still lobbying to have the new European List of Hazardous Waste be similar to those in place in their countries (i.e., a listing of substances or products that have been found to be hazardous; therefore, waste containing any of these products or substances in whatever quantities would, likewise, be categorized as hazardous).

As was the case with Waste Catalogue I, the scope and content of the EU Hazardous Waste List (Waste Catalogue II) are of major importance to a number of industries and companies operating in Europe. The content of this new List will affect such commercial questions as whether and where a material can be recycled or remediated, whether it can be shipped across borders, and what other regulatory constraints and liabilities will apply vis-à-vis manufacturers, generators, and transporters. Sales price and contractual and marketing questions of customers are intricately involved here as well.

The new List of Hazardous Waste Pursuant to Article 1 (4) of Directive 91/689/EEC on Hazardous Waste (January, 1994), while still product related, deals only with broad product categories and is keyed to the thresholds and concentrations developed for hazardous properties under EU chemicals legislation [e.g., the 0.1 % concentration for category 1 and 2 carcinogens in the 1988 Preparations Directive (88/379/EEC)]. The inclusion of these thresholds and concentrations has been the subject of some controversy.

Wallonian Waste Decree. As illustrated in the German "Wirtschaftsgut" example above, another definitional issue of concern in complying with Regulation 259/93 is whether a material is a "product" or "good," hence coming under the Treaty of Rome's free movement of goods principles, or whether it is something else entirely. Under the new EC Regulation 259/93 and such recent legal interpretations as the July 1992 ECJ Judgment on Wallonian Waste Imports [67], waste should be granted special treatment for environmental protection purposes, tantamount to at least a partial exemption from the European Union's free movement of goods and Internal Market principles. This case was particularly interesting in that the Wallonian Decree at issue not only impeded the flow of waste across EU Member State borders, but also impeded intra-Belgian trade on a regional basis. While the ECJ held that the challenged Wallonian Decree did not conform to the 1984 Hazardous Waste Movements Directive (the precursor to Regulation 259/93), it found that the Decree's banning of the disposal of wastes in Wallonia that were generated outside Wallonia was not "discriminatory" within the meaning of the Treaty of Rome. This decision, was based on the "particularité" of wastes, in that Article 130 S of the Treaty of Rome requires that, as a priority, "environmental damage" should be rectified at the source. The Court also referred to the Basel Convention's principle of waste self-sufficiency.

The ECJ appears to have erroneously interpreted the Treaty of Rome provisions here. The Wallonian Decree clearly violates not only the 1984 Hazardous Waste Shipments Directive, but also the free movement of goods provisions of the Treaty of Rome. Relying on Article 130 provisions is also questionable since a Directive that deals with shipments of materials and border transfers within the EU should more appropriately be considered an Internal Market measure and, therefore, judged under Article 100 A of the Treaty of Rome, not Article 130 S.

In addition, because the Wallonian Judgment did not address all of the issues on this subject raised by Regulation 259/93, a definitive legal interpretation is needed on the free trade versus environmental protection argument for waste shipments in the EU.

Article 130 S vs. Article 100 A. The debate continues – the ECJ has just rejected the European Parliament's request to annul Regulation 259/93 because it was improperly based

on Article 130 S (environment) of the Treaty of Rome. The Parliament said that this Regulation was essentially an Internal Market (trade) measure, hence based on Article 100 A of the Treaty of Rome and subject not only to the Treaty's free movement of goods provisions, but also to Article 113 of the Treaty dealing with external trade (It is not just happenstance that an Article 100 A determination for the new Regulation would give the Parliament greater powers as to the form and content of the Regulation.) However, the ECJ rejected the Parliament's arguments and upheld the Article 130 S basis of the Regulation. Waste, it seems, is indeed special and, at least for now, will not be called a "good" in the EU.

The European Court of First Instance has also just rejected as inadmissible a request by two French waste management companies and one Luxembourg one that it annul Article 4 of Regulation 259/93, which permits individual Member States to ban imports of waste, even nonhazardous waste, from other EU countries. (At issue was a French Decree banning imports into France of household waste for disposal in landfill.) The Court of First Instance did not address the substance of the plaintiffs' complaint. Instead, the case was dismissed on the procedural ground of lack of "standing" of the plaintiffs, since they were not "directly concerned." In this ruling, the President of the Court said that Article 4 of the 1993 Council Regulation was not aimed at companies, "except in their objective quality as economic operators in the waste management and transport sector, the same as for any other operators in an identical situation." The Court's President also rejected the request for damages and interest. Case T-475/93, SA Buralux, SA Satrod, SA Ourry vs. Council of the European Union, judgment published in O.J.E.C. C188, July 9, 1994.

Disposal and Recovery. Another definitional concern with Regulation 259/93 is seen in the question of how to define "disposal" and "recovery." Member States of the European Union have, to date, taken inconsistent positions as to whether operations to be carried out are either of the two. This is obviously a critical point for purposes of defining the type of notification and the content of information to include in the Consignment Note under Regulation 259/93. Defining an operation as "recovery" or "recycling/further use," as opposed to "disposal," will also have a critical effect on how the proximity principle would apply. This issue is particularly difficult for operations involving energy generation and incineration.

Perhaps the most difficult practical problem resulting because of these vague definitions is in the area of waste for recovery. The impact on byproducts, coproducts, secondary raw materials, and product or container take-back programs must be clarified.

The practical, commercial difficulties posed by all of the above definitional issues for EU companies in attempts to adhere to Regulation 259/93 are discussed in Sections [**?**]. These problems are magnified even more in the international arena.

In addition to the dilemmas caused by inconsistencies between the EU-level definitions and those in some EU Member States, OECD and Basel Convention strictures, obviously, also have to be adhered to. It is already evident that definitional divergencies between these two organizations and the EU may cause problems in the future.

In addition, many affected companies are multinational enterprises, with manufacturing, waste-generating plants in and outside the EU. Company attempts to impose one comprehensive waste management scheme throughout all of such companies' international activities will be stymied. A case in point is seen in the differences in the U.S. and EU definitions of "waste," "hazardous waste," "disposal," and "recovery," and the shipment restrictions imposed. For example, unlike the ECJ, the U.S. Supreme Court has determined that waste is an article of commerce (i.e., a "good") meriting constitutional protection against discriminatory state and local Regulations that impede its free movement and that nothing special or distinctive exists in waste that should permit discriminatory restriction of its flow across state lines to apply in the United States.

A second example is seen in the fact that the United States attempts to define waste in a way that encourages beneficial recycling (exemption of certain kinds of materials destined for certain recycling). The EU definition follows a less desirable, inconsistent approach: It covers all secondary materials as "waste" as well, regardless

of whether disposed or recycled, but then differs in the regulatory treatment of conventional waste and recyclable waste.

Some parties, frustrated by these definitional differences and the length and complexity of the debate believe that "waste prevention" should be the focus in the future, instead of waste shipment restrictions, especially when the words "waste" and "hazardous waste" cannot even be defined properly.

4.3.3.2. Liabilities

An attempt must also be made to address liability issues under Regulation 259/93 at the EU level to ensure some degree of consistency in determining liable parties and in enforcement. From past experience, this issue cannot be left solely to the Member States. Adoption of EU legislation on Civil Liability for Damage Caused by Waste is a first step (see Section 3.4), but continued Member State inconsistencies on this subject will further confuse companies in compliance.

4.3.3.3. Administrative Issues

European industry, while favoring harmonized regulation of waste shipments throughout the EU, has raised some of the issues noted in the previous sections. However, it appears to be focusing currently on the failure of the EU and its Member States to provide the administrative structures and procedures to allow industry to comply with Regulation 259/93 by its effective date of May 6, 1994, or shortly thereafter.

Among the most active EU-level industry associations on this subject have been UNICE, Eurométaux (which has particular concerns about disruption of shipments of scrap metals as secondary raw materials), the European Chemical Industry Council (CEFIC), and the European Federation of Waste Management (FEAD). According to a UNICE discussion paper on the implementation of Regulation 259/93,

"... the adminstrative backup necessary for the proper implementation of the Regulation is widely lacking in most of the Member States. It is both paradoxical and disastrous for a Community Regulation, directly applicable according to Community law to be paralyzed in this way ...."

The concern about the lack of administrative systems in the Member States to implement the Regulation and the enormous burden to be placed on Competent Authorities in the Member States in homogeneous or integrated implementation is well justified. As noted in Sections [**?**], the notification and approval system under Regulation 259/93 is very complex and inconsistent, depending on the types of shipments, their purposes and destinations, and quite bureaucratic and paper-intensive. It is questionablewhether the Competent Authorities will be able to deal with the large number of notifications to be expected in the time anticipated by the Regulation. Given the large degree of discretion still left to the Member States in implementation, it is also questionable whether a truly harmonized system will evolve. Therefore, it cannot be expected that a significant degree of increased environmental protection will ensue as a result of Regulation 259/93. (Of course, on the other hand, trade, will be severely hampered.)

Moreover, given the fact that Regions having responsibility for waste management will also be in a position of appointing Competent Authorities, an additional administrative question is raised. It is rare that the proper Regional infrastructure exists in each of the Regions throughout the European Union which may appoint such Competent Authorities.

If this administrative situation is not solved quickly, it could lead at worst to failure of the entire system or, at a minimum, to the need to store a large number of wastes prior to disposal or recovery, pending clearance. Such storage may result in more environmental, health, and safety problems than the new requirements in the Regulation will solve.

Other administrative concerns include the fact that on the eve of implementation of the Regulation, the standard Consignment Note to be drafted by the Commission was not yet available; the list of competent authorities to be designated by Member States was also not available (indeed, most of the Member States had not even submitted their Competent Authority designations to the Commission); the list of customs offices called for had not yet been published; and the list of wastes on the OECD Green List that will be subject to specific controls in certain Member States was not available, nor was

information on which third countries have bilateral agreements with which Member States or the EU authorizing such trade.

Thus, according to one industry observer, "chaos" will result as companies attempt to comply with the terms of the Regulation, with clearly inadequate and inconsistent administrative structures and procedures in place.

The Commission Staff, while admitting certain transitional administrative problems, does not appear to believe that all of these concerns are fully justified, and international environmental organizations believe that the administrative and compliance concerns of industry are excessive. Reports that illegal traffic in toxic waste, particularly in a West-to-East direction in Europe and to Asian countries, is increasing have led to growing environmental group pressure to strengthen EC legislation even further on this subject. Greenpeace, in particular, is reportedly concerned with increased shipments to Asian countries, as markets in Africa and in Central and Latin America become increasingly difficult to enter, and with reports of mislabeling waste shipments as bound for recycling or recovery facilities, when, in fact, they are being sent to developing countries for disposal.

4.3.3.4. Trade and Competition Issues

Since broad disparities exist among EU Member States on the administrative structures and procedures to implement Regulation 259/93 and definitional and liability disputes continue, competitive differences will arise state-to-state in the EU. In addition, possible uses and abuses of the proximity and self-sufficiency principles by Competent Authorities could give rise not only to political problems, but also to major anticompetitive ones as well.

Although disposal of waste at sites as close as possible to the point of generation is generally preferable, in some circumstances moving waste over greater distances to ensure its proper disposal may be desirable. Moreover, many companies in Europe have one or a limited number of waste disposal or recycling sites to ensure the proper disposal or recovery of their company's waste throughout the EU. Sending waste to these sites to be disposed of will involve movement, possibly over considerable distances, from one site to another and perhaps across national borders in the EU. This growing tendency for large, usually multinational, companies to be as self-sufficient as possible in the integrated management of their own waste should be encouraged rather than penalized by possibly inappropriate application of the proximity principle.

How the growing number of contractual arrangements involving manufacturers of products taking back contaminated products or packaging from customers (under a life-cycle philosophy) will be merged with the requirements of Regulation 259/93 is also unclear. Therefore, under this new Regulation, industry could well face disruption of its waste disposal and recovery programs, delays in operations, and the need to store waste at significant cost and potential environmental risk, pending administrative resolution of these questions.

In addition, some would argue that it is unwise for the EU, given the current economic situation in particular, to adopt more stringent legislation on this subject than its major trading partners have and than is required by its international commitments.

The Parties to the Basel Convention agreed on March 25, 1994, to adopt the EU Council Common Position to amend the Basel Convention to impose additional restrictions on exports of hazardous waste to developing countries (potentially non-OECD states), for disposal immediately and for recycling or recovery as of December 31, 1997. As a result, at the Second Conference of the Parties to the Basel Convention (Basel COP 2) recently held in Geneva, the Basel signatory countries agreed, in Decision II/12, to ban trade of hazardous waste between OECD and non-OECD countries. The U.S. Government and U.S. industry oppose this action and estimate that such a ban on trading waste for recycling between OECD countries and non-OECD (developing) countries could cost the United States about $\$2.2\times10^9$ a year in commodity trade.

A major indusry sector affected by this ban would be the scrap metal industry. In addition, many non-OECD countries are desperate for raw materials to fuel their economies and may, in response, redefine what is considered a Basel waste in order to circumvent the Treaty.

In the European Union, Greenpeace has pressured the European Commission immediately to

bring it into conformance with these Basel Convention commitments. The Commission, on the other hand, does not believe it is necessary to modify the Regulation and certainly not at this time.

Such a ban on hazardous waste shipments between OECD and non-OECD countries, especially for recycling or recovery, would arguably not be compatible with the GATT. In addition, if the hazardous waste definition is overly broad, valuable goods not usually considered waste – for example, chemical byproducts, secondary raw materials, and used electronic goods that constitute legitimate trade between the EU and non-OECD countries – would be blocked. Again, all of these issues and the differing ways in which the EU and its trading partners handle them will likely result in competitive disadvantages for European industry vis-à-vis its non-EU competitors.

4.4. Conclusion

With its Regulation 259/93, the European Community has undertaken an ambitious regulatory program in a much needed area of harmonized EC legislation, with laudable environmental protection goals. However, the scope and complexity of the Regulation, without adequate administrative implementation procedures in place at the EC level and in the Member States, will give rise to significant legal, trade and competition, and administrative compliance problems for European industry. Moreover, the undesirable inconsistencies on this subject among EU Member States are likely to continue to exist in the application and enforcement of this Regulation. In addition, despite ECJ determinations to the contrary, Regulation 259/93 is primarily a free movement of goods/Internal Market measure, not an environmental one. Wastes, especially waste for recovery, should be treated as "goods" under the free movement provisions of the Treaty of Rome. Failure to do so for intra-Community movements of waste violates the letter and spirit of the Treaty of Rome, as amended – without providing additional harmonized environmental protection.

Finally, inconsistent ratifications and implementation of Basel Convention and OECD commitments by the EU and its trading partners will further complicate the competitive problems facing European industry. Recent Basel Convention amendments on this subject, especially with respect to shipments of waste for recovery, arguably violate GATT rules. It will be interesting to see whether, and how, this issue will be resolved in the GATT's – WTO's new Trade and Environment Programme or perhaps in a new GATT "Green Round."

5. The Seveso Directive: An Example of the Interface between Community Law and National Implementation Rules

5.1. Historical Background and Content

Council Directive 82/501/EEC of June 24, 1982 on the major accident hazards of certain industrial activities owes its existence to a number of industrial accidents. Subsequent to accidents in Flixborough in 1974 (United Kingdom), Beek (The Netherlands) in 1975, and in particular at an ICMESA plant in Seveso (Italy) in July 1976, the Commission undertook to draft legislation for the prevention of major industrial accidents, having been encouraged by the intentions of the United Kingdom, The Netherlands, France, and Italy to adopt national legislation in order to improve accident prevention.

The Commission's proposal was submitted to the Council in July 1979, about three years after the Seveso accident. The proposal was largely inspired by measures that existed for nuclear energy installations to avoid nuclear accidents. While the general lines of the proposed Directive were agreed to relatively quickly by the Council, adoption of the Directive was held up because France, in particular, was afraid that information on risks and safety measures to be given to other Member States might create difficulties for the siting and activity of its nuclear installations. Under pressure from the European Parliament, which had from its beginnings been in favor of the adoption of the Directive, and the Benelux countries, a compromise was finally found that enabled the Directive to be adopted in June 1982.

At the end of 1984, a leak of methyl isocyanate from a Union Carbide plant in Bhopal,

India caused the death of more than 2000 persons, and some 20 000 persons were injured. Following that accident, the Commission proposed an amendment to Directive 82/501/EEC, which was adopted in early 1987 (Directive 87/216/EEC). The Parliament, in its opinion on this proposal, had suggested a wide-ranging review of the Directive. However, the accident of Sandoz in Basel in October 1986 and the public concern raised by it necessitated immediate action. The Commission submitted in spring 1988 a proposal for a further amendment that considerably enlarged the existing provisions for accident prevention related to the storage of dangerous substances. That proposal was adopted within six months, a record time, as Directive 88/610/EEC.

The announced general revision of Directive 82/501/EEC was inserted into the Commission's legislative program 1993 but has, as of April 1995, not yet been adopted. Directive 82/501/EEC aims at preventing major accidents and limiting the consequences thereof. Approximating the legislation of the Member States is a means of attaining this objective.

Industrial activity is defined in Article 1 of this Directive. Industrial activity means any operation in an industrial installation listed in Annex I to the Directive that involves one or more dangerous substances and is capable of presenting major accident hazards, as well as storage of dangerous substances under conditions laid down by Annex II of the Directive.

Major accidents are also defined in Article 1: "an occurrence such as a major emission, fire or explosion resulting from uncontrolled developments in the course of an industrial activity, leading to a serious danger to man, immediate or delayed, inside or outside the establishment, and/or to the environment, and involving one or more dangerous substances."

The Commission thought it necessary to publish guidelines for the interpretation of this concept. Directive 82/501/EEC provides for a number of obligations:

1) Manufacturers who process or store certain chemicals have a general duty to prevent major accidents and limit their consequences for humans and the environment (Art. 3 and 4). Should a major accident occur, manufacturers have a duty to report it (Art. 10). The Competent Authorities of Member States have a general duty to organize inspections or other methods of control [Art. 7(2)].
2) Manufacturers, furthermore, have specific duties that apply either when any of 180 dangerous substances are used in quantities above specified threshold levels or when any of 28 dangerous substances are stored. In such cases, the manufacturer must produce and update periodically a safety report (Art. 5) and an on-site emergency plan.
3) The Competent Authorities must ensure that an off-site emergency plan is produced (Art. 5 and 7). They must further ensure that "persons liable to be affected by a major accident" are informed of safety measures and of the correct behavior to adopt in case of accident.

5.2. Implementation Requirements

Member States had to adapt their national legislation to the requirements of the Directive by January 8, 1984. This national legislation and the above mentioned amendments of the Directive phased in gradually. The following shows how three Member States – France, Germany, and Greece – have complied with the requirements under Directive 82/501/EEC.

Situation in France. On August 16, 1982 (i.e., only seven weeks after the adoption of the Directive), France issued an internal circular, by which it drew the attention of the Commissaires de la Republique (the Préfets) to the existence of the Directive and requested that specific care be given to the information requirements contained in its Article 8. It listed the installations that came under the Directive and were, in France, the subject of the legislation on "classified installations" (loi sur les installations classées) of July 19, 1976 and of several decrees issued under that legislation, in particular a decree of September 21, 1977.

The legislation of 1976 provides that all types of installations, listed in a specific register, that might present a risk must obtain an administrative license for operation. The corresponding list includes more than 300 activities and thus goes far beyond the number of installations that are covered by Directive 82/501/EEC.

To fully adapt its existing legislation, in 1987, France adopted legislation on the organization of civil protection, forest protection against fire, and the prevention of major accidents. This legislation requested the establishment of national, regional, or departmental offices for organizing help in cases of accidents and natural or man-made catastrophes. Emergency plans among others were to be prepared by the departmental prefects around certain industrial installations.

These are the essential pieces of French legislation to which workers' safety provisions must be added in order to cover the area of the Directive. What is striking is the fact that the legislation on industrial installations is, to a large extent, umbrella legislation. The major part of the obligations that the Directive places on the individual manufacturer is to be determined by individual administrative decision. Some 500 000 French plants are registered and subject either to a system of declaration or, for some 55 000, to a system of authorization procedure under the legislation of 1976. Most of the elements of Directive 82/501/EEC are found in the 1976 legislation, although this legislation does not seem to cover all elements of Articles 3 and 4.

The requirement for safety reports (risk analysis) was enlarged after adoption of the Directive to cover existing installations. These safety reports under Article 5 of Directive 82/501/EEC refer to some 320 types of installations. On-site emergency plans were, prior to adoption of the Directive, mandatory only for oil industry installations; since the first circulars of 1982 and 1983, they have also been required for installations that come under Article 5 of the Directive. Off-site plans have been requested gradually since 1985, subsequent to legislation adopted that year.

The French system is a clear case of existing means being amended in order to comply better with a Directive that postdates the basic national legislation. Furthermore, the French system gives considerable authority to the departmental, regional, and national administrations, which operate via individual administrative decisions.

Situation in Germany. Germany effected Directive 82/501/EEC and its subsequent amendments by a great number of provisions, under essentially four different categories:

1) Legislation on workers' safety
2) Legislation on chemical substances and preparations
3) Legislation on catastrophes
4) Environmental legislation

Overall, more than 40 pieces of legislation, adopted at the federal or regional level, apply to the subject matter covered by the Directive.

The most relevant provisions to Directive 82/501/EEC are contained in the so-called "Störfallverordnung," the 12th Regulation on the implementation of the Federal Act on Quality Standards (Bundesimmissionsschutzgesetz). For political reasons, Germany thought it necessary to have its national legislation in place when Community legislation on this subject was adopted. This led the German government to submit a proposal for the 12th Regulation on February 20, 1980, while the Commission proposal for Directive 82/501/EEC dates from July 16, 1979. In 1980, the 12th Regulation was adopted in Germany, while the Community Directive was politically agreed to, though not yet formally adopted. The same coincidence occurred after the Sandoz accident in 1986, when both the German government and the Commission submitted proposals. This time the Community acted more quickly; Directive 88/610/EEC was adopted in 1988, while the German amendment was finally adopted in 1991.

The multiplicity of legal texts makes it difficult to identify discrepancies between the 12th Regulation and Directive 82/501/EEC, without being confronted with the objection that other legal provisions explicitly or implicitly cover the aspect in question. The text of the German 12th Regulation is more precise than that of the French legislation; it gives greater legal certainty to industrial installations and less discretion to the administration.

The manufacturers' general obligations are listed specifically, as well as the obligation to establish an on-site emergency plan and a safety analysis. The obligation to inform persons outside the plant has been placed on manufacturers. Off-site measures are to be agreed to with the competent regional or federal authorities. Generally, the threshold levels in German legislation are more stringent than those in the Directive. Also, the list of substances that make the pro-

visions applicable is longer than that of Directive 82/501/EEC.

Situation in Greece. Greece is a Member State that, at the time of its accession to the Community in 1981, had relatively little legislation on the environment. Accident prevention measures for installations were a rather new item, and the Greek administration was not really accustomed to monitor manufacturers' behavior with regard to safety. These rather sketchy lines might help explain why national measures to transpose the amendment of Directive 82/501/EEC took several years. A similar attitude is to be observed regarding the amendment made by Directive 88/610/EEC: only in 1993 after the Commission had filed an application to the European Court of Justice for noncommunication of national implementation measures, Greece transformed the requirements of Directive 88/610/EEC into national law.

The national legislation reproduces the requirements of Directive 82/501/EEC and its amendments correctly. However, the responsibility of monitoring the legislation is given to no less than four bodies [environmental and industrial ministries, prefects at local – regional level, and a Coordinating Interministerial Committee (CIC)]. These rather dispersed responsibilities might have contributed to certain doubts being raised about effective application of the legislation. Indeed, several accidents suggest that effective application, safety analysis, and on-site or off-site emergency plans have not been made for all installations that come under the Directive, nor is regular inspection always made. One reason for this situation is that the different central and local – regional administrations did not receive any supplementary financial or human resources, when the new monitoring tasks were assigned to them in 1988. Thus, the administration does not act as counterpart interlocutor to manufacturers. Since the requirement of informing the public of measures to be taken in case of accident has been placed on prefects, a manufacturer does not incur too serious risks regarding compliance with the Directive's requirements.

5.3. Conclusions

If one looks horizontally at Directive 82/501/EEC and its implementation, a number of observations can be made:

1) The Directive owes its existence and its tightening-up to industrial accidents and public concern; these accidents made possible the adoption of preventive provisions that had not been acceptable at the national or Community level prior to the accidents.
2) Of the three Member States considered, France undoubtedly gives the greatest discretion to the administration and Germany the greatest legal certainty to manufacturers.
3) Effective application of the requirements of the Directive requires flexible cooperation between manufacturers and administration regarding inspection, monitoring, safety analysis, and risk assessment. The administration itself needs to be trained and experienced to be able to contribute to accident prevention without unduly hampering manufacturers' activity. The absence of a trained, effective, and active administration that effectively monitors the practical application seems to be the greatest handicap in Greece.
4) The Directive is not entirely clear on the question of who has to inform the public on safety measures and on behavior in case of accident (Art. 8) – the administration or the manufacturer. This sometimes leads to practical difficulties. Furthermore, if this kind of information is too detailed, it might provoke negative reactions among the population concerned.
5) The cost of prevention might be significant, while the benefits – the absence of major industrial accidents – are very difficult to estimate. Both France and Germany had, prior to adoption of the Directive, taken a number of steps to prevent major accidents, although these measures focused more on human health than on environmental protection.
6) It is not clear under the Directive whose task it is to set up the off-site emergency plan, to ensure its consistency with the on-site emergency plan, and to manage off-site emergency operations. Obviously, the manufacturer is not capable of providing for all details of the off-site plan such as blocking roads or

evacuating the population. Therefore, active cooperation by the administration is necessary, not only in case of accident, but also to learn about cooperative mechanisms and gain experience for any accidental event.

7) French administration has a tendency to ask for measures to prevent accidents in the worst possible scenario, while German administration normally is satisfied with the degree of accident prevention that is reached by using the best available technology.
8) German administration seems to attach limited importance to the question of siting of industries, whereas this aspect plays an increasing role in the French administration's approach to accident prevention.
9) Safety reports play an essential role among all measures required by the Directive. They are considered a means to obtain more precise information on installations, to identify their accident potential, and to prepare preventive measures more adequately. However, which role the administration has to play in this context is not made clear in the Directive. The differences among Member States seem to be important. The administration in Greece seems to be easily satisfied with the safety reports submitted. In France, the authorities examine the safety reports and sometimes require other preventive measures, but never "accept" the report, thus leading to a kind of uncertainty by manufacturers. It might be useful to systematically check, as German administration does, the reports' completeness, correctness, and their effective monitoring and updating.

The hope is that further amendments of the Directive will not be the result of new industrial accidents.

6. U.S. and European Environmental Legislation

6.1. Introduction

This Chapter discusses the nature and role of environmental Regulation, particularly as it applies to the chemical industry. It makes no attempt to summarize the vast and rapidly evolving body of current U.S. and EC environmental legislation. Other sources can be found for this (see, e.g., [68], [69]). The environmental legal provisions applicable to all phases of the chemical industry are now so extensive, detailed, and complex that they cannot be treated exhaustively under a single keyword of Ullmann's.

Rather, the attempt here is to step back from the system and compare briefly the style and context found in the world's two major centers of regulatory activity, the United States, and Europe, to note some of the key features of the legislation in both systems; and to comment briefly on what may lie ahead.

6.2. The Environmental Regulatory System [70]

The United States and Western Europe have had significant environmental laws on the books since the early 1970s. The United States has erected an elaborate edifice of Regulations. At the EC level and in many European countries (Germany being a noteworthy exception), however, a great deal has not happened until recently. Over the past several years, the European Community and its Member States have been generating a flow of environmental initiatives of ever-increasing importance. Now developing in Europe is an extensive framework of environmental regulatory and liability law (see Chaps. 3, 4) that presents business operating in Europe with many of the same kinds of risks found in the United States. The ways these laws are developed in the European Community and in the United States differ fundamentally.

6.3. Methods of Lawmaking in the United States and the European Community

United States. In the United States, the law and lawyers have been central to the development of environmental policy, and highly structured efforts at the federal level have led the way. U.S. Regulation relies heavily on generic, federal administrative rule making to achieve specificity and uniformity, and to gather the technical, scientific, and economic data that form the

premise for public decision making. Administrative rule making proceedings provide public, formal, and structured opportunities for citizens and industry to participate in the development of Regulation through comments, public hearings, and sometimes judicial challenges to rules. The "political" aspects of the U.S. environmental regulatory process take the form of lobbying directed at the U.S. Congress, as well as less formal efforts to lobby the administrative agencies that establish, implement, and enforce Regulations.

In short, in the United States, contending interests use legal processes in the judicial arena and before administrative agencies to influence the formulation and execution of policy. This is particularly true of environmental groups who, with some frequency, use citizen suit provisions in federal environmental legislation as a means of influencing environmental policy.

The result of this dynamic, contentious law and rule making process is a highly articulated, but often inflexible, regulatory system, i.e., a regulatory system that is at times too rigid, complex, and confusing.

EC Regulation (see also Chaps. [**?**]). Regulation in Europe, on the other hand, has traditionally been regarded as "technical" or "political," rather than "legal" in nature, although this attitude appears to be changing. European industry, as a whole, has not in the past looked to legal counsel to assist it in handling environmental matters, but has relied primarily on its technical staff, who often operate through national and European-wide trade associations. The EC regulatory process, furthermore, is largely nonadversarial and is the result of a slow, nonpublic, lawmaking structure designed to achieve political consensus among the Member States. A close industry – government relationship exists. Industry operates through contacts with national authorities; through "expert" advisory committees at the Commission of the European Community; and through lobbying of the Commission, Parliament, and COREPER (the permanent representation of Member States, which assists the Council on day-to-day matters). Proposed Directives are generally made publicly available and are commented on by Parliamentary committees, but until recently, this was not required at the early stages. This situation has been changed, at least in theory, in 1992 – 1993 with the adoption of Council and Commission Resolutions calling for early consultation of interested bodies, including industry associations. No structured open administrative process for promulgating Directives or Regulations exists, however, such as the U.S. notice and comment rule making process.

Implementation of EC Directives is left to the authorities of individual Member States. The degree to which Directives are effectively implemented varies widely from country to country. While some Member States do not give full effect to Directives until long after adoption, other states adopt even stricter measures than those existing at the Community level. For example, Denmark, The Netherlands, and Germany have well-developed national environmental policies. Ireland, Greece, Portugal, Spain, and to some extent, Italy, are in earlier stages of development. Belgium, France, and the United Kingdom are somewhere in the middle, although they have had environmental policies for many years. British law is now developing especially rapidly. As a result of the wide disparity in Member State national environmental laws, chemical companies operating in Europe must be familiar with both Community and national environmental legislation in order to minimize the risk of unexpected environmental Regulation or liability.

6.4. U.S. and European Environmental Laws

Environmental law, to date, has progressed chiefly in two broad streams – one regulating the manufacturing process and the other regulating the products themselves (→ Chemical Products: Safety Regulations). A third stream of natural resource law regulates extraction of the oil, gas, coal, and various minerals that provide the raw material for the chemical manufacturing process. (A number of minor subjects in environmental law deal, for example, with the special protection of certain types of resources, such as wetlands, endangered species, wild and scenic rivers, archaeological sites, and historic preservation. They are not dealt with here since they are less important to and less specific to the chemical industry.) This third broad stream

of natural resource Regulation will be treated in → Raw Materials and Energy. The focus in this section is on chemical process and chemical product Regulation.

Chemical process Regulation encompasses chiefly pollution control Regulation – the classic areas of air emissions, water effluents, and solid and hazardous wastes. It centers on the chemical manufacturing plant, although it typically also applies to all related commercial or industrial facilities (e.g., storage depots, means of transport, and disposal facilities). Generally, the basic tools are construction or operating permits that impose relevant operating constraints, monitoring and reporting requirements, etc. Beyond these standard provisions, however, environmental impact assessment laws, accident planning laws, and public information laws have also been developed, sometimes implemented in permits or the permit process and sometimes separately.

6.4.1. Environmental Impact Assessment Laws

Both the United States and the EC have environmental impact assessment (EIA) laws [71]. These apply chiefly during the licensing of new chemical or other manufacturing facilities or the major expansion of existing plants. They require a systematic, detailed, and written analysis of the likely environmental impacts of the proposed project. This analysis is designed to force both chemical industry management and government regulators to consider environmental factors in their project decision making. In both the United States and the EC, however, the EIA laws themselves mandate no particular result. Rather, they are procedural laws premised on the assumption that a better decision-making process leads to better decisions. U.S. law goes beyond EC law in many ways. It requires consideration of all reasonable alternatives to the project, and has more detailed guidance and "teeth" than its EC counterpart.

6.4.2. Accident Planning and Public Information Laws

The United States and the EC likewise have accident planning and public information laws of great significance for the chemical industry. The EC developed broad-gauged legislation on this subject earlier than the United States, and most European countries have a more systematic approach to risk assessment and analysis built into the new (and modified) plant licensing process than does the United States to date. The EC law (the Seveso Directive, see Chap. 5) is designed to prevent and limit the consequences of major accidents resulting from industrial operations involving dangerous substances [72].

The corresponding U.S. law, the Emergency Planning and Community Right-to-Know Act, requires (1) emergency planning, (2) emergency release notification, (3) community right-to-know reporting, and (4) toxic chemical release inventory reporting [73]. While the EC does not yet have an analogue to the U.S. toxic chemical release inventory requirements, one is on the drawing board. For many years, the United States has also had separate oil spill legislation requiring the well-known Spill Prevention, Control and Countermeasure (SPCC) plans [74] and has much better developed legislation on aboveground and belowground storage tank Regulation than either the EC or most European countries (Germany being an exception in this regard).

6.4.3. Air and Water Pollution Legislation

The air and water pollution legislation in both the United States and the EC is extensive [75]. In both cases, the legislation uses "best-technology"-type substantive emission and effluent standards (emission limits in German terms) as well as case-specific emission and effluent limits to be set at levels necessary to attain air and water quality requirements (the latter called immission standards in German terms).

U.S. water pollution legislation has for years built its control program around a comprehensive permit system, and U.S. air pollution legislation is now adopting the same technique. The EC legislation has had no comprehensive permit system requirements to date, although the Commission is now developing a multimedia permit system modeled after the British Integrated Pollution Control system. The U.S. air and water legislation has regulated new plants

more stringently than existing sources through technology-based new source performance standards for industry categories. The legislation has evolved industry-specific technology-based limits for existing sources on the water side and is now moving to do likewise on the air side, and has developed complex antidegradation requirements (in effect, special limits on degradation of an air or water resource that meets existing ambient requirements). The EC legislation has tended, in practice, chiefly to apply technology-based limits for new sources in a few large industrial categories, and the pervasive air and water quality-based Regulation of existing sources called for by EC legislation has never been put in place. Indeed, not even the air and water quality monitoring systems necessary to implement such ambient-based Regulation have been developed. The Commission is still struggling to come up with a politically and administratively feasible vehicle for setting air and water technology-based limits for the industry, such as are pervasive in the United States and in Germany. Antidegradation requirements, to the extent theoretically present under EC law, have been largely ignored.

In both the United States and the EC, efforts are now being made to move beyond Regulation of the few parameters dealt with in both the air and the water context in the early years and to tackle the difficult and complex job of effectively regulating a multitude of toxic pollutants.

6.4.4. Solid and Hazardous Waste Regulation

U.S. Regulations. In the field of solid and hazardous waste Regulation, the United States has for years had a much more effective regulatory system in place for hazardous waste than the EC (although the German waste regulatory system is comprehensive, except for a loophole in the case of "recyclable" wastes, now to be closed by the new German Eco-Cycle Waste Law of 1994) [76]. The U.S. cradle-to-grave regulatory system uses registration requirements, permits, and transport manifests to regulate virtually all aspects of the generation, storage, transport, treatment, and disposal of a broad range of hazardous wastes (including many destined for recycling), under an extremely complex set of rules and both written and unwritten regulatory practices. All types of treatment, storage, and disposal facilities must meet specific design and operating performance standards, including statutory minimum technology and groundwater monitoring requirements and a prohibition on the land disposal of untreated hazardous wastes. The U.S. system has complex rules for mixtures of wastes; for wastes derived from other wastes; and for distinguishing wastes from products, coproducts, byproducts, chemical intermediates, and secondary raw materials. It regulates some wastes destined for recycle or reuse and does not regulate others. Despite all this hazardous waste Regulation, however, the U.S. system has virtually no comprehensive or systematic Regulation of solid waste at the national level, this area being left almost entirely to state and local governments.

EC Regulations (see Section 4.3). The EC waste management framework is more coherent, but less effective, than that in the United States. The European system is built on a basic framework Directive that deals with all forms of waste and is supplemented by a more specific Directive dealing with hazardous waste [77]. The entire system is now in flux (see Section 4.3).

Further, the European waste regulatory system has depended, from the outset, on Regulation of the *shipment of waste* (see Section 4.3.1). Both the United States and the EC countries have Regulations governing how hazardous materials must be transported, but EC waste shipment Regulation governs whether hazardous waste can be shipped across national borders at all, and under what conditions. As early as 1984, the Directive on Transfrontier Shipments of Hazardous Wastes regulated hazardous waste shipments across Member State borders and into and out of the EC by creating prior notification, manifesting, and packaging requirements [78]. More recently, the community has adopted a new Regulation on Shipment of Waste that implements the Basel Convention in Europe (the Basel Convention governs shipment of hazardous waste on an international scale) and that imposes more stringent requirements (including prior authorization in some cases) on the shipment of all forms of waste, both hazardous and normal, across Member State borders and into or out of the EC (see Section 4.3.1). Shipments of waste

for recovery operations would, depending on how they were listed in three categories (Red, Amber, and Green; see Section 4.3.1) on lists prepared by the "Committee Procedure," be (1) largely excluded from Regulation (green), (2) subject to a prior notification regime (amber), or (3) subject to a prior authorization regime (red; waste not listed would be subject to the authorization system). This Regulation could balkanize the EC on waste matters, with local authorities lodging objections motivated by protectionism of local disposal or recovery facilities and NIMBY-ism. Chemical facilities may be faced with disruption of their waste disposal and recovery arrangements, and those dependent on off-site waste disposal or recovery (and having only limited on-site waste storage capacity) may find their operations at risk of interruption.

To implement the EC's Waste and Hazardous Waste Directives as recently amended, the Commission has drawn up a comprehensive list of solid and hazardous wastes (the European Waste Catalogue). A separate list has been adopted by the Council, which will be critical to EC waste legislation because it will determine the threshold question, for particular wastes, of whether they must be regulated as "solid" or as "hazardous" (see Section 4.3.3). The Commission is also having great difficulty establishing and implementing the definition of "waste" found in relevant Directives.

U.S. Superfund and EC Packaging Legislation. Two idiosyncratic areas of waste Regulation remain to be discussed – one in the United States and one in the EC. In the United States *Superfund legislation* uses government-funded cleanup backed up by a civil liability scheme to deal with soil and groundwater contamination [79]. Originally designed for old dump sites, this broad-reaching legislation extends in fact to all instances of such contamination, both old and new and both on-site at chemical manufacturing facilities and off-site. The government-funded cleanup process has been slow and has been required by Congress to use extremely stringent cleanup levels. The liability regime imposes retroactive, joint and several, strict liability regarding site cleanup liability and responsibility for natural resources. This liability is imposed on virtually everyone connected with the contaminated site since its contamination, including lenders in some cases. The primary Superfund program has been widely criticized and drives broad secondary efforts of private remediation at sites where the potentially responsible parties hope to avoid becoming entangled in the main cleanup and liability program. The European Commission has been considering legislation to deal with soil and groundwater contamination in Europe, and the Council of Europe has adopted a convention on this subject (see Chap. 3).

The second piece of idiosyncratic legislation is the effort by the EC and a number of Member States to regulate one form of solid waste, *packaging*, through product take-back (or advance collection, sorting, recycle, or disposal fees) and thus to develop a new paradigm for regulating all forms of solid waste . The Commission, reacting to the so-called Töpfer Ordinance on packaging waste in Germany and to similar efforts in France and differently structured efforts in other Member States, has proposed a Directive on Packaging and Packaging Waste, which is aimed at eliminating and preventing trade barriers arising from varying national packaging waste initiatives. The essence of this legislation is the ability to impose product charges to force those in the packaging chain, from manufacturer to consumer, to begin to internalize some elements of the costs of collection, sorting, recycle, or disposal of solid waste. The German system achieves the same end by imposing on the commercial parts of the packaging chain a legal "product take-back" obligation. By forcing internalization of some of the costs of the end-of-life disposition of products, the scheme aims at correcting market conditions so as to favor recycling, while letting the market rather than direct Regulation sort out how much recycling is warranted and in which cases. Particularly at the EC level, however, extensive direct regulatory measures, which may well not be environmentally wise or economically efficient, are also being considered.

6.4.5. Regulations on Chemical Products

EC Regulations. Environmental legislation regarding chemical products was one of the first areas pursued by the European Community, although initially its motivations had more to do

with the Common Market than with environmental protection. The first major piece of chemical legislation was adopted in 1967, and established a common Community system for the classification, packaging and labeling of dangerous chemical substances [80]. This basic Directive was modified in 1979 by the so-called Sixth Amendment, which established a premarketing clearance requirement [81]. Under this amendment, companies marketing new "dangerous substances" are required to submit the new chemical substance to a prior notification procedure that involves testing the substance for toxicity and eco-toxicity and preparing technical dossiers to provide information on the substance's characteristics (see Section 2.2). In 1988, the Community adopted similar requirements for dangerous preparations (i.e., mixtures of dangerous substances) [82].

The Commission has also recently adopted material safety data sheet (MSDS) rules and has extended its Regulation to include the marketing and use of certain dangerous substances and preparations, as well as enacting Directives in the field of biotechnology designed to control the handling and release of genetically modified organisms. The Commission is also moving to impose on manufacturers and importers of existing chemical substances some significant data reporting obligations for assessment of their risk characteristics.

U.S. Regulations. In the United States, chemical product Regulation on a comprehensive basis began with the Toxic Substances Control Act in 1976 [83]. That Act places on manufacturers the responsibility of providing data on the health and environmental effect of chemical substances and mixtures and gives the Environmental Protection Agency comprehensive authority to regulate the manufacture, use, distribution in commerce, and disposal of chemical substances. The central regulatory mechanism is the requirement for premanufacture review of new chemical substances prior to their commercial production and introduction to the marketplace. The Environmental Protection Agency also has authority to require testing of chemicals, although no requirement for a "base level" of tests exists as a precondition to premanufacture review as in the Sixth Amendment to the EC Regulation.

Regulation of environmental harms from chemicals in the workplace takes place in the United States under the Occupational Safety Health Act. Under that act, the Hazard Communication Standard requires that employees be informed about the potential chemical exposures involved in their work and be trained to deal with them [84]. Chemical manufacturers and importers are required to evaluate the hazards of the chemicals they produce or import, and to use MSDS to transmit this information to their own employees and to employers to whom they ship the chemicals. Further, chemicals shipped must contain labels on the containers with similar information.

Beyond these core areas, both the U.S. and the European systems regulate use of polychlorinated biphenyls, asbestos, and agricultural chemicals. Further, hazardous chemicals are regulated to varying degrees and in varying ways in consumer products, foods, drugs, and other end uses such as cosmetics and medical devices as well.

6.5. Some Contrasts in U.S. and EC Environmental Legislation [70]

For obvious constitutional and political reasons, the scope, specificity, and rigor of EC legislation do not equal those of U.S. federal legislation.

6.5.1. Regulatory Gaps

Important regulatory gaps exist in the EC (e.g., on wetlands, underground storage tanks, and until recently, municipal sewage), along with weaknesses (e.g., Member State ability to determine, on a relatively subjective basis, the applicability of certain EC Directives on fish protection and bathing water). Nonetheless, in some areas, EC provisions or proposals go beyond present U.S. practices or requirements (e.g., eco-labeling, eco-auditing, packaging waste, and economic and fiscal measures such as a carbon or fuel tax).

6.5.2. Implementation of Legislation

Implementation of the EC Directives is spotty and frequently ineffective. Directives,

particularly the early ones, tend to be general policy statements, with little of the detailed standards and permit requirements of their U.S. counterparts. For example, the European Community as yet has no specific standards for surface impoundments (e.g., pits, ponds, lagoons), underground tanks or pipes, waste piles, and various forms of waste treatment. Where Member State implementation takes place, standards and permit conditions are often negotiated on a case-by-case basis, and systematic monitoring and reporting obligations are not always imposed.

United States. In the United States, the implementation of legislation at the state level is reasonably well assured by U.S. federal legislation through a series of devices lacking in EC legislation. First, a well-staffed, trained, and funded agency (the EPA) can be found at the federal level in the United States. Also available is a great deal of public, comparable monitoring data and other information, as well as effective freedom-of-information legislation at the federal and state levels. Second, the EPA may as a formal matter withdraw a state's right to implement many of the federal regulatory programs if it judges the proposed implementation to be inadequate. The EPA also normally has direct prior approval authority over state-implemented Regulations, standards, and individual permits before those measures can go into effect under federal law (although they can and do sometimes go into effect under state law without regard to EPA approval at the federal level). In the absence of an EPA-approved state implementation program or Regulations, standards, or permits, the EPA is frequently required to promulgate directly applicable federal provisions. Should it fail to supervise state implementation where it has a mandatory duty to do so, it may be sued in federal court by citizens and environmental groups and forced to act. Lastly, the EPA has frequently threatened to withhold federal funding on which the state programs depend if state implementation is deemed inadequate.

6.5.3. Enforcement of Legislation

The enforcement of EC legislation is perhaps the weakest element of the EC regulatory framework. All the European Community can really do is insist that Member States enact legislation (i.e., transpose EC legislation into national law). The European Court of Justice has new power to impose penalties on Member States for failing to implement EC environmental law, but it remains to be seen how effective this tool will prove. An enforceable obligation on the part of Member States to undertake effective, concrete implementation through more specific regulatory standards and requirements, comprehensive permit or approval systems, and an adequately staffed and trained bureaucracy with sufficient funding and legal authority is generally lacking in many countries. Furthermore, the European Environment Agency, which is only just getting started has only a data collection and dissemination function. It does not, for the time being, have enforcement powers.

Enforcement in the United States is vigorous and takes a number of forms. Nearly all of the major U.S. environmental regulatory statutes require some form of publicly available self-monitoring and reporting. The system thereby automatically highlights lapses in compliance and the need for enforcement. When the EPA deems state enforcement inadequate, it normally has the authority at any time to override the state's failure to enforce and to issue an administrative order or, in some cases, to assess civil penalties on an administrative basis. In either case, the EPA may enforce its action in federal court or take the violator directly into federal court for injunctive relief and civil or criminal penalties. Should the EPA fail to do so, citizens and environmental groups can themselves normally take the violators into federal court.

Further Development in the EC. The EC institutions currently have none of these tools at their disposal. At the level of enforcement of specific standards, requirements, or permit conditions against an individual regulated facility, the Community plays virtually no role. Such enforcement is left up to the Member States. Enforcement is difficult, even when attempted, since self-monitoring and reporting obligations have not always been imposed either by Regulation or by permit conditions. Government failure to prosecute generally cannot effectively be

challenged except through publicity and the political process.

The Commission is now engaged in a series of moves to surmount the difficulties it has encountered in trying to develop an effective "command-and-control" EC regulatory system. It is considering the use of fiscal and economic measures instead of administratively driven control requirements. It is trying to make greater use of directly applicable Regulations in the place of Directives that must be implemented by Member States; and in any event, EC legislation, overall, is becoming more specific, thereby providing the detail necessary for ensuring the consistent implementation that has been absent in earlier legislation. The Commission is drafting an Integrated Pollution Control Directive, modeled on British legislation, that would impose comprehensive permitting requirements covering all environmental media, i.e., air, water, soil, groundwater (but, initially, applying only to a restricted list of larger chemical, metallurgical, and paper plants). This Directive would impose substantive and management requirements as permit conditions and would implement on a case-by-case basis the existing media-specific air and water quality legislative requirements. The Commission has also adopted an eco-audit Regulation that provides a (presently "voluntary") comprehensive, universal system that is a partial, periodic substitute for the self-monitoring and reporting provisions that are lacking in existing Directives. These eco-audits would have to be verified by independent outside auditors, with public disclosure of certain information. The likely result of these eco-audits will be public and political pressure for more effective standard setting, implementation, and enforcement.

At a fundamental level, for the first time, the Treaty on Political Union (the Maastricht Treaty), would provide the Court of Justice with limited powers to sanction with penalties those Member States failing to implement EC law. Also, the Court of Justice has been developing the following legal principles to increase the significance of EC law: (1) obligations on national bureaucrats at all levels to apply EC Regulations, even where the national legislature has failed to implement that EC Regulation adequately (i.e., "direct applicabibility"); (2) "direct effect" of provisions in EC Directives that are sufficiently precise and unconditional, even where the national legislatures have failed to implement adequately; and (3) state responsibility for harm to individuals arising from the government's failure to implement EC legislation adequately. The cumulative effect of these developments will be to increase the importance of EC environmental law and to render more likely its effective, if not consistent, implementation and application throughout the European Community.

6.6. Outlook [70]

While the basic framework of U.S. environmental law has been fairly settled (the U.S. Clean Air Act and new initiatives by the first Republican-controlled Congress in 40 years being exceptions), environmental law in Europe is changing rapidly. The past and present roles of European environmental law, of course, indicate that the EC environmental proposals may not necessarily have more than limited importance for business. As is well known, a gap has existed between the rhetoric and the reality of European environmental law. The European Community and many Member States have a good deal of environmental "law" on the books, but much of the EC legislation requires little real action. Moreover, an inability at the EC level and, in some cases, a reluctance at the Member State level to implement and enforce existing environmental law, have been evident. Further, a remarkable lack of comparable data to document many European environmental problems, as well as a lack of data on government actions and enforcement in regard to them, are noticeable.

The politics of the environment appear to be changing, however, As a result of such crises as Seveso, Chernobyl, and Sandoz, and of the palpable, long-neglected environmental problems facing Western and Eastern Europe, public attitudes toward environmental risks in the European Community are evolving quickly. Environmental groups, although perhaps not yet as sophisticated and well funded as their U.S. counterparts, are becoming increasingly well organized and active. Adoption of the EC Directive on Freedom of Information on the Environment and creation of the European Environment Agency, by improving the quantity and quality of publicly available environmental information, will in all likelihood result in heightened public

awareness, increased effectiveness of environmental groups, and consequently, increased political pressure for consistent and effective environmental legislation and enforcement.

7. Environment, Safety, and Chemical Safety in Japan

7.1. Current Situation Concerning Environmental Protection

Since the early 1970s, Japan has made a strong effort to tackle the problem of environmental preservation. Various measures, such as reinforcement of pollution control Regulations and amplification of pollution control facilities have been implemented.

As a result, some serious environmental conditions have been alleviated and the situation has generally improved since the 1980s. Moreover, because the social demand for environmental preservation, particularly in metropolitan areas, tends to be more and more diversified and sophisticated, the Environment Agency is promoting various measures in response to this situation.

In addition, finding solutions to the recent environmental problems that are developing into international and global issues and global climate problems are considered important political subjects by the Japanese government as in other countries. In November 1993 the basic law for environmental pollution control was put in force in response to the result of the UNCED held in June 1992.

7.1.1. Air Pollution Control

Regulatory standards to protect human health (environmental standards) have been set for SO_2, NO_2, CO, photochemical oxidants, and airborne particles based on the Basic Law for Environmental Pollution Control. For Regulation of exhaust gas emission accompanying industrial activities, emission control Regulations are in force for SO_x, NO_x, cadmium and its compounds, Cl_2 and HCl, F_2, HF and SiF_4, lead and its compounds, and exhaust dust. The Regulations are based on the Air Pollution Prevention Law. With regard to SO_x and NO_x, Regulations on total amounts of emissions are also valid within the designated area in the country. In addition to these nationwide uniform standards, special emission standards must be met by newly constructed or expanded facilities in the specified area. In certain areas, additional emission standards have been set by local governments (in municipalities designated by government ordinance).

To meet these standards, industries have taken measures such as desulfurization of exhaust gas [176×10^6 m^3/h (STP) at 1810 facility units], denitration of exhaust gas [142×10^6 m^3/h (STP) at 379 units], development of combustion processes that emit lower amounts of NO_x, absorption of HCl and HF, electrostatic dust separation, etc.

The regulatory system for total NO_x emission control for each industrial installation was introduced into three areas with a high population density, and the system was implemented in Yokohama and in Kawasaki in April 1982, and in Osaka and Tokyo in November 1982.

However, NO_2 concentration in the environment shows an increasing trend, and reinforcing the Regulations on NO_2 mobile emission sources such as automobiles is considered necessary.

In June 1989, the Air Pollution Prevention Law was amended. Fibrous materials such as asbestos that are hazardous to human health have been treated as specific dangerous forms of dusts in this amendment.

7.1.2. Control of Water Pollution

Wastewater Regulation standards were implemented in 1970 based on the Water Pollution Prevention Law. Together with additional standards implemented by the respective local governments (in municipalities designated by government ordinance), water pollution has been controlled with regard to the following items:

1) **Substances Posing a Health Hazard (Toxic Substances).** Concentration limits are set for cadmium and its compounds, cyanides, organic phosphorus compounds, lead and its compounds, chromium(VI) compounds, arsenic and its compounds, mercury, alkylmercury and other mercury

compounds, polychlorinated biphenyls, trichloroethylene, and tetrachloroethylene.

2) **Items Affecting Living Environment.** Limits have been set on pH range, BOD, COD, suspended solids, concentration of oily substances, phenols, copper, zinc, soluble iron, soluble manganese, chromium, fluorine, and a number of *Escherichia coli* colonies.

Industries have promoted various measures such as wastewater treatment and process rationalization. As a result, the environmental standards on toxic substances have been satisfied by 99.9 %, and a general trend of improvement is observed for household effluents. However, in closed water areas such as rivers, inner bays, lakes, and swamps, pollution was still proceeding and eutrophication was advancing. Since the cause of such situations was the insufficent installation of sewage systems for treatment of household wastewater, the Water Pollution Prevention Law was revised.

The Law on Special Measures for Preservation of Water Quality of Lakes and Swamps was enacted in July 1984, and an environmental standard and wastewater discharge control standard for nitrogen and phosphorus concentration were introduced.

The Regulation on the total quantities of contaminants in water was implemented in 1978 for three closed sea areas, namely, Tokyo Bay, Ilse Bay, and Seto Inland Sea, and all-around measures for reduction of the total COD were implemented. In January 1991 the third Regulation for the reduction of total quantities of contaminants was prepared. The countermeasures for water quality control of the closed sea will become more important in the future because tremendous growth of algae resulting from eutrophication will seriously damage fishery by causing red tide.

To counteract groundwater pollution, trichloroethylene and tetrachloroethylene were classified as hazardous substances under the Ordinance for Enforcement of the Water Pollution Prevention Law in April 1989; in June, upon amendment of the Water Pollution Prevention Law, percolation of specific water containing hazardous materials into the groundwater was prohibited. In this connection, the ordinances for reinforcement of the Law related to the Disposal and Cleaning of Wastes were also revised.

7.1.3. Revision of Compensation Systems for Patients with Pollution-Related Diseases

The compensation system for patients suffering from pollution-related diseases was implemented in September 1974, and compensation has been paid, at the cost of polluting parties, to the approved patients in Class no. 1 (directed to the four designated diseases caused by air pollution) and Class no. 2 (directed to diseases caused by water pollution or soil contamination). For Class no. 1, the Central Pollution Study Committee submitted a report in October, 1986 proposing a system that reflects a substantial improvement in environmental pollution.

Based on the above report, a revision bill of the Law on Compensation of Pollution-Related Patients was approved by Parliament. The contents of the revision are (1) all-out cancellation of class I area designation due to the improvement of the environmental situation; (2) continued compensation to approved patients; and (3) promotion of comprehensive environmental hygiene measures (establishment of a fund). As a result, a long-awaited rational revision of this system was implemented in March 1988. By the end of 1993 the number of approved patients was 84 471 and expenses totaled 93.4×10^9 Yen.

7.1.4. Measures for Waste Disposal

Since 1970, the Law on Waste Treatment and Cleaning has been in force, with the two groupings of "Domestic Wastes" and "Industrial Wastes." Section 3 of the law stipulates that an entrepreneur shall be responsible for, e.g., self-accountability, appropriate treatment, discharge reduction, not making products difficult to treat, etc. With respect to industrial waste, the law provides for standards on collection, transportation, and disposal and Regulations concerning the industrial waste treatment industry, industrial wastes treatment facilities, and hazardous industrial wastes.

As far as domestic wastes are concerned, awareness is growing of the need for waste reduction, including recycling of waste. This is due to the fact that the quantities of wastes discharged have been increasing in recent years,

existing incineration and final treatment facilities have reached their maximum capacity, and need exists to utilize limited resources more effectively. The Law Concerning the Utilization of Recycled Resources was published in April 1991. The bill for revision of the Law on Waste Treatment and Cleaning was deliberated by Parliament in September 1991.

7.1.5. Response to Global Environmental Problems

The Law Concerning Protection of Ozone Layers by Regulation of Specified Substances was enforced in Japan in May 1988. The law specifies measures that are to be taken to regulate the manufacture of specific substances, to restrict their discharge, and to rationalize the use of these substances.

In June 1990, the Montreal Protocol was revised to strengthen Regulations by placing a complete ban on the use of certain fluorocarbons by the year 2000 and adding to the list substances such as carbon tetrachloride and 1,1,1-trichloroethane. The related industries in Japan are making efforts to meet these requirements by developing alternative substances for use in place of existing ones.

In August 1990, IPCC of United Nations Environment Program (UNEP)/WHO (the intergovernmental panel regarding climate fluctuation) issued a report about the forecast influences and countermeasures of global warming, which had been studied for three years since 1988. In Japan, corresponding studies have been conducted under the leadership of relevant government authorities. The issue of global warming is also of utmost importance to industries.

It is estimated that ca. 244×10^6 t/a of CO_2 (in terms of C) is discharged in Japan, accounting for 4.3 % of the amount discharged worldwide. Of this amount, about 4.9 % originates from the chemical industry. Since the chemical industry has curtailed energy consumption by about 40 % since the 1980s, little room is available for further energy saving. Thus, reduction of CO_2 discharge appears to be a difficult problem.

7.1.6. Basic Policy of the Japanese Chemical Industry Association about Environment and Safety

The Executive Committee of the International Council of Chemical Associations (ICCA) has set up the ICCA Principles of Responsible Care with the aim of promoting a program for the improvement of voluntary environmental and safety measures.

The Japan Chemical Industry Association concluded that the chemical industry clearly realizes its social responsibility and continues its endeavors to undertake voluntary measures in order to achieve sound growth of the chemical industry at large, at a time when humans are faced with new international problems such as global-scale environmental issues, waste disposal, and the safety of chemical substances. Japan's chemical industry has already made great progress in environmental pollution countermeasures and is at a very advanced level compared with to its U.S. and European counterparts.

7.2. Current Situation Concerning Safety and Fire Prevention Measures

As a result of a series of explosions and fires that occurred in 1973 – 1974 in chemical and oil refinery complexes in Japan, the chemical and related industries have implemented safety and fire prevention measures. Due to such efforts, large-scale explosion accidents in the plants and works have decreased drastically since 1978.

However, the number of accidents is gradually leveling off, and further safety and fire prevention measures should be taken and promoted.

7.2.1. Revision of Fire Service Law

Ignitable and flammable materials are classified as Hazardous Materials based on the Fire Service Law.

On May 24, 1988, the law for partial revision of the Fire Service Law (Law no. 55 of 1988) was promulgated, and in this relation, the revision of the government and ministerial ordinances. The essential part of this revision is the "review of the scope of hazardous materials" to define

the hazardous material more clearly and introduce methods to test whether a substance is hazardous or not. The tests should be performed and judged by the enterprises themselves, following and considering these laws and Regulations.

Another important point is a review of the scope of the term "dangerous articles."

1) The term dangerous article is clearly defined, and judgment of a system by testing is introduced. The test methods and criteria are established by taking into account the definition and international test methods.
2) A designated quantity is determined by government ordinance by taking into account the danger of the articles concerned.
3) Quasi-hazardous materials and special combustible materials are regulated under the category of combustible materials.

The provisions relating to the revision of the scope of dangerous materials in this law were enforced as of May 23, 1990.

7.2.2. Safety Measures for the Transport of Dangerous Goods

For the air and marine shipments involved frequently in international transport, international Regulations have been introduced so that Japanese domestic legislation is consistent with international transport standards.

7.2.2.1. Marine Transport Standards for Hazardous Substances

Breakbulk Transport. The Ministry of Transport provided Rules for the Marine Transport and Storage of Dangerous Goods with respect to the breakbulk transport. The rules were determined with reference to the International Marine Dangerous Goods Code (IMDG Code) set by the International Maritime Organization. The notice of revision incorporating the 25th revision of the IMDG code was enforced in January 1991. With regard to container testing, the test system and UN label indication system have been internationally enforced since January 1991, and the use of qualified containers having UN labels is obligatory for the transport of dangerous goods.

Transportation in Bulk. With regard to the marine transport in bulk of hazardous liquids by chemical tankers and liquid gas tankers, the Rules for the Massive Transport and Storage of Dangerous Substances were amended as of July 1986, and the International Bulk Chemical Code (IBC) and International Gas Carrier Code (IGC) of the IMO were officially adopted.

The amendment of IBC Codes was adopted in March 1989. In response to this amendment, the Ocean Pollution Prevention Law, the Regulations on the Marine Transport and Storage of Dangerous Materials, etc. were revised and enforced in October 1990. In particular, as for the Ocean Pollution Prevention Law, it is not necessary to make prior assessment for pollution classification of goods mixed with pollutants already assessed or that contain pollutants already classified by the International Marine Organization.

7.2.2.2. Standards for Air Transport of Hazardous Cargoes

With respect to air transport, standards for the transport of hazardous cargoes (notification) in Japan are determined on the basis of the Aviaiton Act. These standards are in accordance with the Regulations on hazardous cargoes (International Civil Aviation Convention Annex no. 18) and the technical guidelines of the International Civil Air Organization (ICAO). These standards (notice) were revised on December 20, 1990, and enforced on January 1, 1991. Testing of containers and the use of containers with UN labels have been obligatory since 1991 as in the case of marine transport.

7.2.2.3. Standards for Transportation Containers for Poisonous and Toxic Substances

The Control Act on Poisonous and Toxic Substances stipulates the handling of poisonous and toxic substances.

The Pharmaceutical and Supply Bureau of the Ministry of Health and Welfare issued the standard on transportation containers for poisonous and toxic substances (no. 3; a standard

for small-size transportation containers having an inner volume of 450 L or less). The standard went into effect on October 1, 1991.

This standard is in line with the United Nations Hazardous Substances Experts Committee Recommendation-no.6 Edition from the standpoint of international harmonization. Also on October 1, notice was given of the establishment and revision of standards concerning each of the poisonous and toxic substances with respect to emergency measures in transportation accidents as well as disposal.

7.2.2.4. Safety Measures in the Use of Gases Under High Pressure

Safety Regulations with respect to high-pressure gas are in accordance with the High-Pressure Gas Control Act and governmental ordinances associated therewith.

The Industrial Location and Environmental Protection Bureau of the Ministry of International Trade and Industry has been holding meetings of the High-Pressure Gas Safety Policy Committee since September 1989 to study the best possible way of organizing high-pressure gas safety administration in the future. In June 1991, a report was published, proposing the following:

1) To promote self-security (i.e., self-protection measures and attitude for security)
2) To strengthen safety measures in response to diversification of applications
3) To review import Regulations from the standpoint of internationalization
4) To promote international cooperation in the field of safety
5) To strengthen execution systems by making full use of private-sector inspection organizations
6) To enhance systems supporting voluntary security activities

Since July 1991, the Ministry of International Trade and Industry has been holding meetings of the High-Pressure Gas and Explosives Safety Review Committee to study the future administration of high-pressure gas safety, including revision of the High-Pressure Gas Control Act.

7.3. Present Situation Regarding Safety Measures for Chemical Substances

In Japan, two laws exist for maintaining the safety of chemical substances under the notification system for new chemical substances prior to the manufacture thereof: (1) The Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances, which is designed to regulate substances that are hazardous to human health because of contamination of the environment and (2) the system of Investigation of Toxicity of Chemical Substances under the Industrial Safety and Health Law, which is to regulate the concentration of carcinogenic substances in the workplace.

7.3.1. The Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances

The Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances had been, until its revision in April 1987, implemented effectively for the 14 years since its enforcement in 1973 to regulate PCBs and similar substances.

However, as Regulations became stricter in Europe and the United States, the need arose to coordinate Regulations on chemical substances in various countries and to expand Regulations on harmful chemical substances that are slowly degraded in the environment. The Japanese government then enacted a Law for the Partial Revision of the Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances in May 1986 and implemented it in April 1987. The gist of the revised law is as follows:

1) The system of prior examination of new chemical substances was reinforced by demanding that toxicity tests, namely, mutagenicity tests (Ames test, chromosonal aberrations test) and repeated dose tests are to be conducted for 28 days, in addition to the biodegradability and accumulation tests now in force. The tests must be performed, at the time the government is notified of the chemical substances, prior to their manufacture and import.

2) Substances that are difficultly degradable and are suspected to have chronic toxicity are to be termed "Designated Chemical Substances" and must be put under surveillance with respect to their production volume and application after manufacture and import.
3) When such "Designated Chemical Substances" are suspected to be contaminating the environment and be hazardous to human health, the manufacturer or importer will be advised to perform a study on hazards of such substances.
4) If such studies reveal that the substances are hazardous, they are termed "Class II Specified Chemical Substances," and their production volume and handling methods must be regulated.
5) "Specific Chemical Substances" with respect to the old Law Concerning Examination and Regulation of Manufacture, etc. of Chemical Substances are now termed "Class I Specified Chemical Substances," and the same control measures, such as ban on their manufacture in principle and on their use in open systems, are used as in the past.
6) As a result of tests conducted by the government on existing chemical substances (which correspond to a prior examination for new substances), some of them may be termed "Designated Chemical Substances."

7.3.2. Situation Regarding the Enforcement of Law Concerning Examinations and Regulations of Manufacture, etc. of Chemical Substances ("Chemical Substances Law")

Under the old Chemical Substances Law (until March 1987) the reported number of new chemical substances admitted for manufacture and import was 3864; under the new Chemical Substances Law (April 1987 – December 1993), which applies more stringent Regulations, it is 1492.

The designated chemical substances as of the end of September 1993 amounted to 97, including chlorohydrocarbons and organotin compounds. Among them are 135 substances (excluding the 23 substances that were regrouped to the "Class II Specified Chemical Substances") whose manufacture and import quantities must be reported. The 23 specified chemical substances inlcude chlorohydrocarbons (trichloroethylene, tetrachloroethylene, carbon tetrachloride) and organotin compounds (e.g., triphenyl- and tributyltin compounds). Furthermore, safety measures for each substance such as the preparation of proper use manuals and labeling of containers are being taken by individual industry organizations.

As far as existing chemical substances are concerned, the government continues to conduct systematic safety monitoring. The Ministry of International Trade and Industry checked the safety of 970 substances from 1974 through 1993. The result was that 358 substances decomposed in solvents to nonhazardous compounds, 594 substances showed low and 18 substances high accumulation of hazardous compounds upon decomposition. Of these, the nine substances or classes of substances listed below are designated Class I Specified Chemical Substances. In addition, the Ministry of Health and Welfare conducts toxicity testing of each substance.

Polychlorinated biphenyls
Polychlorinated napththalenes
Hexachlorobenzene
Aldrin
Dieldrin
Endrin
DDT
Chlordane
Bis(tributyltin)oxide

More careful handling and discharge, not only of those substances that are regulated specifically but of all chemical substances in open systems, is desired.

7.3.3. International Cooperation on Chemical Substance Regulations

In Japan, the Good Laboratory Practice (GLP) system was introduced in 1986, and the coordination of testing methods with those of OECD and the mutual acceptance of test data were promoted.

Moreover, the OECD started to promote the work of international cooperation for investigating the safety of existing chemical substances.

Specifically, the preparation of data bases (EXICHEM) on the testing status in member countries and joint investigation of chemical products of high production volume are being put in force. Concerning the joint investigation of chemical products with a high production volume, 53 substances were allocated in April 1990 to each country as Phase 1 in accordance with the investigation priority substances (P 1) list. Japan took charge of nine substances and collected data on them. Based on the result of this investigation, tests in 1991 were conducted with respect to characteristics on which no data were available or the data were indefinite. In addition, investigation of about 100 substances (Phases 2 and 3) was initiated toward the end of 1991. In this program Japan is to evaluate data on 24 substances. Furthermore, until the year 2000 about 5000 chemical substances of high production volume including the above figures will be surveyed. Japan will evaluate one-fourth of those substances.

8. References

1. Council Directive 80/778/EEC Relating to the Quality of Water Intended for Human Consumption, OJL 229, August 30, 1980, p. 11.J.
2. J. Cameron, J. Aubouchay: "Precautionary Principle: a Fundamental Principle of Laws and Policy for the Protection of the Global Environment," Boston College International and Comparative Law Review, 1991.
3. E. Rehbinder, "The Precautionary Principle in an International Perspective," in Miljørettens grundspøgsmål – bidrag til en nordisk forskeruddannelse – G-E-C Gad, København 1994.
4. M. Rémond-Gouilloud, "Le risque de l'incertain: la responsabilité face aux avancées de la science" – La Vie des Sciences, *Comptes rendus, série générale* vol. 10 (1993) no. 4, 341 – 357 – Archives de l'Académie des Sciences, Paris.
5. Oslo and Paris Conventions, Paris, 21 – 22 Sep. 1992, vol. 14 (1991) no. 1, p. 16.
6. Treaty on European Union, Maastricht, 7 Feb. 1992.
7. "London Declaration," *2nd Int. Conf. on the Protection of the North Sea,* 24 – 25 Nov. 1987.
8. [C(72)128], OECD 1972; [C(74)223, OECD 1974.
9. [C(90)177 (final)], OECD 1991.
10. O.J.E.C., L 169 1 – 29, June 26, 1987.
11. H. Smets: "Le principe pollueur payeur, un principe de l'environment," *Rev. Gen. Droit International Public* **2** (1993) April/June.
12. Regulation 93/259/EEC on the Supervision and Control of Shipments within, into and out of the Community.
13. International Information Exchange and the PIC Procedure (UNEP/FAO).
14. UNEP (1989): London Guidelines for the Exchange of Information on Chemicals in International Trade (1987, amended 1989), UNEP, Nairobi.
15. FAO (1989): International Code of Conduct on the Distribution and Use of Pesticides (1985, amended 1989, FAO, Rome).
16. FAO/UNEP (1991): Guidance for Governments: Operation of the Prior Informed Consent Procedure for Banned or Severely Restricted Chemicals in International Trade, FAO/UNEP, Rome/Geneva.
17. Council Regulation (EEC) No 2455/92 or July 23, 1992 (amended on January 11, 1994) concerning the Export and Import of Certain Dangerous Chemicals.
18. Rege, "GATT Law and Environment-Related Issues Affecting the Trade of Developing Countries," in Journal of World Trade Law (June 1994).
19. "Towards Sustainability," European Commission Fifth Environmental Programme [*COM* (92) 93 final]
20. Resolution of the Council and the representatives of the Governments of the Member States on the fifth Community Policy and Action Programme for the Environment and Sustainable Development, adopted on March 18, 1992, OJC 138, May 17, 1993.
21. Council Directive 90/313/EEC On the Freedom of Access to Information on the Environment of June 7, 1990 and Communication to the Council, Parliament and the Economic and Social Committee 93/C 156/05 on Public Access to the Institutions' Documents.
22. Council Regulation 880/92/EEC on a Community Award Scheme for an Eco-Label (OJL 99, April 11, 1992)
23. OJC 72/3, March 10, 1994.
24. R. Lönngren: *International Approach to Chemical Control: A Historic Overview,* Kemi, The National Chemical Inspectorate, Stockholm 1992.

25. EEC Treaty, Article 5.
26. EEC Treaty (1987 – 1993), Article 130 r (4). Maastricht Treaty on European Union, Article 3b.
27. Directive 67/548/EEC Relating to the Classification, Packaging and Labelling of Dangerous Substances, O.J. 1967, no. 196, p. 1 with later amendments.
28. Directive 73/173/EEC relating to the Classification, Packaging and Labelling of Dangerous Preparations (Solvents), *O.J.* 1973, no. L 189, p. 7, with later amendments.
29. Directive 78/631/EEC Relating to the Classification, Packaging and Labelling of Dangerous Preparations (Pesticides), *O.J.* 1978, no. L 206, p. 13, with later amendments.
30. Directive 77/728/EEC Relating to the Classification, Packaging and Labelling of Paints, Varnishes, Printing Inks, Adhesives and Similar Products, *O.J.* 1977, no. L 303, p. 23 with later amendments.
31. Example taken from Directive 67/548/EEC (note 3) in its version of the 6th amendment, Directive 79/831/EEC, *O.J.* 1979, no. L 259, p. 10.
32. Directive 76/769/EEC on the Approximation of the Laws, Regulations, and Administrative Provisions of the Member States Relating to Restrictions on the Marketing and Use of Certain Dangerous Substances and Preparations, *O.J.* 1976, p. 201 with later amendments.
33. Directive 79/831 (note 7), recital 1.
34. EEC Treaty, Article 130 r (2).
35. Directive 90/220/EEC on the Deliberate Release in to the Environment of Genetically Modified Organisms, *O.J.* 1990, no. L 117, p. 15.
36. Regulation 594/91/EEC on Substances that Deplete the Ozone Layer, *O.J.* 1991, no. L 67, p. 1; now replaced by Regulation 3093/94, O.J. 1994 no. L 333, p. 1.
37. Directive 91/156/EEC on Waste, amending Directive 75/442/EEC, O.J. 1991, no. L 78, p. 32; Directive 91/689/EEC on Hazardous Waste, *O.J.* 1991, no. L 377, p. 20.
38. Directive 76/464/EEC on Pollution caused by Certain Dangerous Substances Discharged into the Aquatic Environment of the Community, O.J. 1976, no. L 129, p. 23; 80/68/EEC on the Protection of Groundwater Against Pollution Caused by Certain Dangerous Substances, *O.J.* 1980, no. L 20, p. 43; 84/360/EEC on the Combating of Air Pollution from Industrial Plants; *O.J.* 1984, no. L 188, p. 20.
39. Directive 82/501/EEC on the Major Accident Hazards of Certain Industrial Activities, *O.J.* 1982, no. L 230, p. 1 with later amendments.
40. Directive 85/337/EEC on the Assessment of the Effects of Certain Public and Private Projects on the Environment, *O.J.* 1985, no. L 175, p. 40.
41. Regulation 1836/93 allowing Voluntary Participation by Companies in the Industrial Sector in a Community Eco-management and Audit Scheme, *O.J.* 1993, no. L 168, p. 1.
42. Commission Proposal for a Council Directive on Integrated Pollution Prevention and Control, O.J. 1993, no. C 311, p. 6.
43. Directive 92/112/EEC, *O.J.* 1992, no. L 409, p. 11.
44. UNEP, "Liability and Compensation for Damage resulting from the transboundary Movement of Hazardous Wastes and their Disposal." Basel Convention, Montevideo, Dec. 1992, draft articles.
45. Directive on Civil Liability caused by Defective Products, *O.J.E.C.* no. L 210, August 7, 1985.
46. Amended proposal for a Directive on Civil Liability for Damage caused by Waste, *O.J.E.C.,* no. L 192/6 (July 23, 1991).
47. Communication to Council and Parliament: Green Paper on Remedying Environmental Damage, *COM* (93), 47, March 1993.
48. Sections of this chapter have been taken, in amended form, from two previous publications by G. H. Williamson: "Environmental Protection versus Free Trade: A European Community Perspective," in: *Private Investments Abroad,* Southwestern Legal Foundation, Dallas, Texas, 1993, Chap. 3. G. H. Williamson: "The European Union's New Regulation on Waste Shipments: Its Content and Practical Implementation Concerns," *Environment Watch, Western Europe,* May 6, 1994.
49. "Completing the Internal Market: White Paper from the Commission to the European Council," Commission of the European Communities, June 1985.
50. Commission vs. Council, Judgment of June 11, 1991, on Dir. 89/428 of the Council.
51. Paragraph 2, new Article 130 R, added to Part Three of the Treaty by Article 25 of the Single European Act.
52. Treaty on European Union, signed at Maastricht on Feb. 7, 1992 (*O.J.E.C.* C191, July 25, 1992).

53. *O.J.E.C.* 1983, L 109/8, amended by Directive 88/182/EEC, O.J.E.C. 1988, L 81/75.
54. Rewe-Zentral AG vs. Bundesmonopolverwaltung für Branntwein (Case 120/78), Judgment of Feb. 20, 1979. ECR (1979) p. 649.
55. Commission vs. Denmark (Case 302/86), Judgment of Sep. 20, 1988, ECR (1988) p. 4607;1 CMLR 408.
56. Verordnung über die Vermeidung von Verpackungsabfällen, Bundesgesetzblatt 1991, part I, p. 1234.
57. Common Position No. 13/94 with a View to Adopting a European Parliament and Council Directive on Packaging and Packaging Waste, 94/C/137/08, *O.J.E.C.* C 137/65, May 19, 1994.
58. Regulation 259/92, *O.J.E.C.* L 30/1, Feb. 2, 1993.
59. Directive 75/442/EEC, *O.J.E.C.* L 194/39, July 25, 1975.
60. Directive 91/156/EEC, *O.J.E.C.* L 78/32, March 26, 1991.
61. Directive 91/689/EEC, *O.J.E.C.* L 377/21, Dec. 31, 1991, as amended.
62. Directive 78/319/EEC, *O.J.E.C.* L 84/43, March 31, 1978.
63. Directive 84/631/EEC, *O.J.E.C.* L 326/31, Dec. 13, 1984.
64. Commission Decision 94/3/EC, *O.J.E.C.* L 5/15, Jan. 7, 1994.
65. 75/442/EEC and 91/156/EEC.
66. 91/689/EEC; List published at O.J.E.C. L 356/14 seq., Dec. 31, 1994.
67. Commission vs. Belgium, E.C.J. Case no. C-2/90, decided July 9, 1992.
68. S. M. Novick et al. (eds.): *Law of Environmental Protection,* Environmental Law Institute, Washington, D.C. 1987.
69. T. T. Smith, Jr., R. D. Hunter (eds.): *European Community Deskbook,* Environmental Law Institute, Washington, D.C. 1992.
70. T. T. Smith, Jr., et al.: "Understanding European Environmental Regulation," *Conference Board Monograph,* report no. 1026, New York 1993.
71. 42 U.S.C. §§ 4321-70C; Directive 85/337, O.J.E.C. L 175/40, July 5, 1985.
72. Directive 82/501, *O.J.E.C.* L 230/1, Aug. 5, 1982.
73. 42 U.S.C. $ 11001 (a) – (c).
74. 33 U.S.C. § 1321.
75. The Clean Air Act, 42 U.S.C. §§ 7401 ff, as amended; the Clean Water Act, 33 U.S.C. § 1251 – 1376. The EC Air and Water Legislation is Summarized and Cited in Ref. [69].
76. The Resource Conservation and Recovery Act, 42 U.S.C. §§ 6901 ff. The EC Waste Legislation is Summarized in Ref. [69] pp. 24 – 28.
77. Directive 75/442, *O.J.E.C.* L 194/26 (July 25, 1975), as amended; Directive 78/319, *O.J.E.C.* L 78/32 (Mar. 31, 1978), as amended.
78. Directive 84/631, *O.J.E.C.* L 326/31 (Dec. 13, 1984), as amended.
79. 42 U.S.C. §§ 9601 et seq.
80. Directive 67/548, *O.J.E.C.* L 196/1 (Aug. 16, 1967), as amended.
81. Directive 79/831, *O.J.E.C.* L 259 (Oct. 15, 1979).
82. T. T. Smith, Jr., R. D. Hunter (eds.):*European Community* in [69] pp. 28 – 31.
83. 15 U.S.C. §§ 2601 – 29.
84. 29 C.R.R. § 1910.1200, issued under 5 U.S.C. 553.
85. T. T. Smith, Jr., et al.: in [70] p. 7 – 10.
86. T. T. Smith, Jr., et al.: in [70] p. 10.

Leucine → Amino Acids

Environmental Management in the Chemical Industry

Louis Jourdan, European Chemical Industry Council (CEFIC), Brussels, Belgium

Filip Jonckheere, European Chemical Industry Council (CEFIC), Brussels, Belgium

Ulrich Geffarth, European Chemical Industry Council (CEFIC), Brussels, Belgium

Jaques Busson, European Chemical Industry Council (CEFIC), Brussels, Belgium

1. Introduction 2
1.1. Historical Development 2
1.2. Factors Influencing Companies to Improve Their Environmental Performance 3
1.3. Scope of This Contribution 5
2. Environmental Management in Chemical Companies 5
2.1. Preamble 5
2.2. Environmental Management Systems 5
2.3. Environmental Policy 5
2.4. Organization and Structure 6
2.4.1. Typical Organization 6
2.4.2. Staffing of HSE Functions 7
2.4.3. Employee Motivation and Training . . 7
2.4.4. Research and Development 8
2.5. Standards, Regulations, Monitoring, and Environmental Programs . 8
2.5.1. Register of Standards 8
2.5.2. Measuring Environmental Performance 8
2.5.3. Environmental Programs 9
2.6. Environmental Audits 9
2.6.1. Definition 9
2.6.2. The Audit Plan 10
2.6.3. Environmental Auditors 10
2.6.4. Audit Methodology and Process . . . 11
2.6.5. Audit Report and Action Program . . 11
2.7. Environmental Management Reviews 11
3. Environmental Communication . . 11
3.1. Environmental Communication: A New Aspect of Management 11
3.2. Audiences and Vehicles 12
3.3. The Environmental Report 13
3.3.1. Audiences 13
3.3.2. Corporate or Site Reports 13
3.3.3. Contents of the Report 13
3.3.4. Links to Environmental Management Systems 14
3.3.5. Trends 14
3.4. Stakeholder's Reactions 14
3.5. Who Communicates? 15
4. The Impact of Legislation on Environmental Management 15
4.1. Operating Permits 15
4.2. Process Safety Legislation 16
4.3. The EC Eco-Management and Audit Regulation 16
4.4. Mandatory Environmental Communication 17
5. Voluntary Action by Industry 18
5.1. Internal Statements 18
5.2. Voluntary Agreements 18
5.3. Public Commitments 19
5.4. Responsible Care 20
5.4.1. Responsible Care; What Is It? 20
5.4.2. The Guiding Principles 20
5.4.3. National Programs 20
5.4.4. The Scope of Responsible Care; Product Stewardship 21
5.4.5. The Challenges of Responsible Care . 21
6. Standardization and Certification of Environmental Management Systems 22
6.1. Standards for Environmental Management Systems 22
6.1.1. Generic or Specific Standards 22
6.1.2. EMS Standards 23
6.1.3. The ISO 9000 and 14 000 Standards . 23
6.2. Certification of Environmental Management Standards 23
6.2.1. Certification of an EMS 23
6.2.2. Certification Against the EC Regulation 24
7. International Aspects of Environmental Management 24
7.1. Environmental Requirements in Different Countries 24
7.2. Technological Cooperation 25
8. References 25

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b07_359

1. Introduction

1.1. Historical Development

A few decades ago, the term "environmental management," not to mention "environmental management system," would have raised eyebrows. Despite the fact that some advanced companies were already behaving quite properly as far as environmental protection was concerned – at least relative to the prevailing standards – the term had not yet been coined, and companies did not have the type of environmental organization considered necessary today.

Nevertheless, environmental considerations had been taken into account, consciously or unconsciously, in the management of chemical industry for a long time: one might quote, for example, a French imperial decree from 1810 controlling "dangerous, noxious and infectious establishments" which could be considered as a precursor to the EC "Seveso" Directive. Closer to the present, the Solvay process for the manufacture of soda ash, which in the second half of the nineteenth century replaced the Leblanc process, brought substantial environmental improvements.

In fact, safety considerations were integrated into industrial management long before the need to protect the environment was recognized. Following dramatic accidents and severe health damage that were common in the early stages of the industrial revolution, governments, supported by a few enlightened industrialists, established the first of many standards for process, product, and human safety. The efforts of certain politicians to protect workers against the effects of white lead, and of Chancellor Bismarck to lay the foundations for accident prevention and social security, are well known examples. Some major companies, such as E.I. du Pont de Nemours, have built their reputations on the basis of very strict safety standards.

But it was not before the 1960s or 1970s that the impact of chemical products and chemical industry on the environment were widely publicized – first in 1962 with Rachel Carson's "Silent Spring," then with the UN Conference in Stockholm in 1972. Since that time, local and national governments as well as intergovernmental organizations, in particular the Organization for Economic Cooperation and Development (OECD) and various United Nations Agencies, have established a number of legislative actions, conventions, or recommendations that have necessarily been integrated into the management of chemical companies.

The concept of "Sustainable Development" was introduced in the 1987 report "Our Common Future" from the UN Commission for Environment and Development, led by G. H. Brundtland [1]. The report recognized the threats imposed on the environment by both increasing population and ongoing industrial development, but it also identified challenges that must be met if such development is to be sustained without compromising the needs of future generations. Industry has an indispensable role to play if these objectives are to be achieved, in particular with respect to energy consumption, food production, and chemicals.

The Brundtland report paved the way to organization of the UN Conference on Environment and Development (UNCED), held in Rio de Janeiro in June 1992 [2]. Whatever can be said about the meager immediate results of the Conference, its products, in particular the "Agenda 21", will have to be taken into consideration by industrial managers. Two chapters of Agenda 21 are of direct relevance to chemical industry, and a United Nations Commission for Sustainable Development (UNCSD) has been set up to implement the Conference's decisions.

The integration of environmental considerations into industrial management is influenced most obviously by the level of development of the country in question: even within the European Community, after more than twenty years of legislative actions to harmonize ("approximate") national legislation, some differences still exist between the behavior of certain companies in different Member States. When one looks at developing countries, the differences are often even greater. Political regimes also have an influence on environmental awareness and requirements: the state of the environment left by the centrally planned economy regimes of the former communist countries clearly shows that free-market economy, democracy, and a high state of development all contribute to superior environmental protection in industrialized or industrializing countries.

1.2. Factors Influencing Companies to Improve Their Environmental Performance

While companies undoubtedly must satisfy first their "bottom line" (profitability), as well as a second crucial one – socio-economic responsibility – they now must satisfy a third one as well, namely environmental responsibility. Several factors are pushing industrial managers to meet this third challenge and to improve the environmental performance of their companies. Some of these are listed below, in an order that might vary depending on a particular manager's personal commitment, the company's culture, the local legal and economic conditions, etc.

Legal Requirements and Regulations. In all industrialized countries, environmental regulations are now being established at an ever increasing pace (Fig. 1), and they are setting ever tighter standards. Although enforcement requirements may differ between countries, no reputable company can afford not to comply with at least national and local regulations.

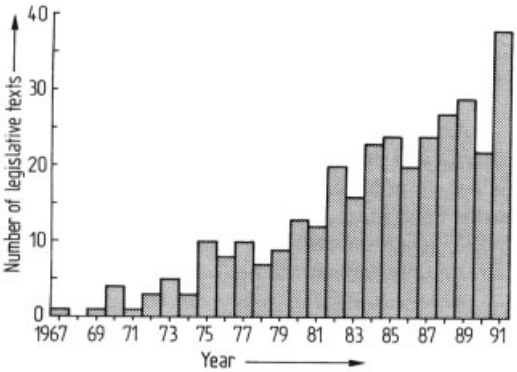

Figure 1. Evolution of environmental legislation in the European Community (EC)

Environmental Liability. Fears of prosecution triggered by civil or even criminal liability for environmental damage are a strong incentive toward excellence in environmental performance. Some well-known "environmental scandals" have been settled, either between the parties involved or by the courts, at levels that could endanger the financial survival of a company. A matter of especially great concern is "past pollution"; i.e., the ecological damage created by an industrial operation that was complying fully with the legal requirements of some past period. Even if not yet common, criminal liability for environmental damage may soon become a reality, as is already the case for safety failures. Several pieces of legislation (Council of Europe, European Community) based on the concept of "strict" liability will make the matter even more important (see Chap. 4 for more detailed information).

Image. As the public, under the influence of the media and many environmental nongovernmental organizations (NGOs), becomes increasingly aware of the need to protect the environment, companies will need to build public images as "good environmental citizens" if they hope to maintain their licenses to operate. There have already been cases of companies boycotted by consumer organizations in response to some environmental affair. "Eco-labels" or "Green Awards" have become strong elements in marketing strategies, and environmental performance criteria are becoming subject to more and more consideration in relations between suppliers and consumers.

Employees. Employees and unions have traditionally been particularly concerned by labor relations, and especially wages, working conditions, and safety, but they are becoming increasingly interested as well in the environmental performance and reputation of the companies with which they are associated. Some companies have established "Environment Committees" at the shopfloor level paralleling the long-established Safety Committees.

There are also clear indications that a good environmental image helps a company recruit high-quality personnel, especially at the management and executive level.

Finance. It is now recognized by the financial community – shareholders, bankers, insurers – that good environmental performance is an asset, even if in the short term it entails some capital expenditures and increased operating costs. There are more and more cases of corporate annual general meetings of shareholders where questions on environmental performance are debated. Moreover, insurance companies now adjust their premiums according to the environ-

mental reputations of their clients. Some organizations have proposed to their customers – mostly in the financial community – an "eco-rating" of major industrial corporations, similar to financial ratings.

Professional Associations. Industrial associations at the national or international level have no real power over their member companies, but they are nevertheless an effective platform for establishing common principles and guidelines that set a high-level yardstick of environmental performance for all their members.

Ethical Considerations. Some industry leaders are already well-known for their personal commitments to environmental protection, and they manage their companies accordingly. Many other company executives are beginning to transfer to their management activities their personal ethics with respect to the environment.

1.3. Scope of This Contribution

When speaking about environmental management it has become current practice to discuss health-protection and safety-management aspects together with purely ecological matters. This contribution devoted to "environmental management" will therefore include various aspects of health, safety, and environmental (HSE) management. It must be recognized, however, that in some companies these elements are not integrated, and that health protection (H), safety (S), or health protection and safety (H + S), are managed separately from environmental protection (E). Conversely, some companies include quality under the same management structure. Because of their specificity, safety and environmental aspects of transport and distribution are integrated into an HSE management structure by only a few companies.

2. Environmental Management in Chemical Companies

2.1. Preamble

As noted in the previous chapter, the concept of environmental management is relatively new; it has developed without established standards, in a "learning-by-doing" process. It is therefore no wonder that the environmental management structures adopted by individual companies differ significantly according to company size, geographical location, and coverage; the legislative framework of the country(ies) in which they operate; their culture; etc.

What is described below is the typical structure, organization, and content of an environmental management system adopted by a typical large or multinational company with headquarters in one of the industrialized countries. Companies of smaller size or operating under different conditions should adapt it accordingly.

2.2. Environmental Management Systems

Basically, the functioning of a company relies on two systems:

1) An *operational system*, which achieves the basic purposes of the company (e.g., the manufacture of a certain range of products under certain financial and social conditions)
2) A *management system*, which acts on the operational system to help it function and improve

Management systems applied to the various aspects of management are an important key to the long-term success of any company. Fundamentally, such a system necessitates defining a policy, establishing standards and objectives, planning for achievement, and allocating whatever resources are required, as well as documenting, measuring, auditing, and reviewing for improvement. An environmental management system (EMS) is built along this "loop," as represented by Figure 2 (see next page). Its main elements are described below.

2.3. Environmental Policy

Although it must be recognized that establishing and publicizing a policy for HSE matters is not, or at least not yet, a part of the culture of some companies, a clearly stated environmental policy is undoubtedly the foundation of any environmental management system.

Figure 2. The environmental management system (EMS) "loop"

An *environmental policy* can be defined as a public statement of the intentions, principles of action, and objectives that govern the management of a company with respect to the environment. Although the details and extent of such a policy may vary from company to company, the policy itself is generally expressed in a concise statement. Typical elements include:

1) A commitment to conduct the company's activities in a manner that safeguards the environment, the health, and the safety of its personnel, local communities, consumers, and the general public
2) Compliance with environmental legislation, or even an intention to go beyond legislative compliance when appropriate
3) Communication of environmental information to concerned stakeholders within society
4) Information for customers on the nature and safe use of the company's products
5) Environmental education and training of employees and contractors
6) Support for environmental research
7) Access to environmentally sound technologies

A corporate environmental policy is a document approved and endorsed by the highest level of management; the President, Chief Executive Officer (CEO), or Executive Committee. The policy is also made available to the public and communicated to all interested parties, in particular the company's employees, shareholders, and major customers.

2.4. Organization and Structure

To implement a company's environmental policy the necessary human resources must be allocated and organized in such a way that they work with the greatest possible effectiveness, in particular in their relationship with the operating system. Organization implies a structure that identifies areas of responsibility and authority for key personnel.

2.4.1. Typical Organization

Again, the personnel structure of an environmental management system may differ quite considerably from company to company, but a typical organization is that described by Figure 3 (note, however, that in some companies the functions H + S and E are separated). The basic principles are as follows:

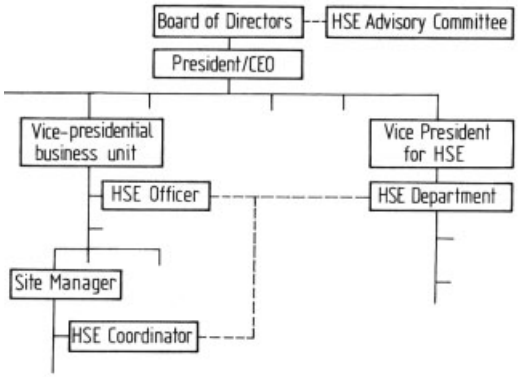

Figure 3. Structure of a typical health, safety, and environmental (HSE) organization

The *Board of Directors* is responsible for defining the company's environmental policy. Some companies have established an "HSE Advisory Committee" to assist the Board in this function; members of the Advisory Committee are members of the Board or persons external to the company selected for their interest, competence, or representativity in HSE matters.

The *Managing Director* (President, CEO) is responsible for the adoption and implementation

of the company's environmental policy. His personal involvement should be made clearly visible across the company (in the famous words of DuPont's E. WOOLARD: "as Chief Officer of this company, I am also its Chief Environmentalist" [3]).

A *senior manager* (Vice-President, Director) reporting directly to the President, heads the company's central HSE department. This centralized function is not directly responsible for policy implementation, but ensures the proper working of the various elements of the EMS. It includes specialized activities such as toxicology/ecotoxicology, occupational health, emissions, waste management, environmental audits, product safety, regulatory affairs, etc. Most of the members of the central HSE departments act in an advisory capacity to support the "line" management.

Business units (product divisions, manufacturing companies, geographical areas) are responsible for policy implementation. Here again, HSE directors or coordinators at the level of the division or site function in an advisory capacity.

2.4.2. Staffing of HSE Functions

The number of employees holding a full-time HSE position varies widely, depending on such factors as process and product hazards, local regulations, etc. An approximate range is from 2 to 25 HSE staff members per 1000 employees, with an average around ten.

Allocation of professionals between a central corporate HSE department and the HSE departments in various divisions and sites is again a matter to be decided by the company: Some have large centralized departments, with up to 150 experts competent in all aspects of the corporation; others prefer to decentralize to the greatest extent possible at the level of divisions or sites, maintaining only a light core group with mainly coordinating functions.

Whether the HSE functions should be staffed by generalists or by professionals educated in HSE matters is much debated, especially since the educational system now provides explicit technical or university degrees in HSE disciplines. Equally important are concerns about the possibility of career development for specialists in HSE. Many companies regard an HSE position as a necessary step toward general management; others ensure that HSE specialists have career development opportunities similar to those offered to generalists.

2.4.3. Employee Motivation and Training

As noted previously, employees in business units and operational sites are responsible for the implementation of a company environmental policy. This requires that staff at all levels be clearly aware of their accountability, as well as properly trained and motivated.

Most companies hold employees accountable through evaluation of their individual environmental performance, or at least safety performance, which is more easily measurable. Plant and site managers are often required to set HSE objectives for their specific areas of responsibility, with possible monetary bonuses linked to achieving the goals. Several companies have prize programs with monetary awards or other forms of recognition for environmental improvements.

Training in HSE matters is a key element in the effectiveness of an EMS. Training is especially necessary when a generalist is appointed to an HSE position. Several companies include HSE elements in their training programs for managers at various levels. Training of all employees in HSE matters is more and more considered to be a major priority, even if for the moment only a few companies have implemented such comprehensive programs.

The case of line middle-managers deserves special consideration, because these persons sometimes perceive environmental demands as coming from people outside the plant with little knowledge about what is happening in practice. It is therefore important that senior management demonstrate its commitment to environmental quality, for example, by facilitating a two-way communication channel on HSE matters, allocating the necessary human and financial resources, and properly rewarding positive achievements. Also, programs that include a strong motivation element, such as Responsible Care (see Section 5.4), are helpful in convincing line managers, as well as staff people like sales managers, of their environmental duties.

Another difficult case is that of contractors working on behalf of a company for construction or maintenance work, waste disposal, transport, etc. It is normally part of a company's policy to insist that contractors work under the same HSE standards as the company's own personnel, but accountability, communication lines, training possibilities, etc. are obviously different. The problem requires special attention on the part of all those involved.

2.4.4. Research and Development

Most major chemical companies have R & D programs devoted exclusively to HSE matters and representing some 10 – 25 % of all their R & D expenditures. Environmental R & D topics are linked most obviously to a company's products and processes, and to its HSE priorities and programs.

2.5. Standards, Regulations, Monitoring, and Environmental Programs

A fundamental element of an EMS is a comprehensive register of applicable legislation and company standards. It is against these governmental or company standards that environmental performance will be measured, permitting the necessary programs for correction or improvement to be developed.

2.5.1. Register of Standards

The number and complexity of environmental statutes at the local, national, and international level is constantly increasing. It is the role of a central business unit or individual site HSE functions to determine those that are applicable in the various sectors of the company. It may also happen that, because of gaps in the applicable legislation or difficulties in adapting it to the specialties of a certain activity, a given company is forced to establish its own standards, which would of course go beyond the legislated criteria. In general such standards are established at the highest level of the company, and are applicable to all its sectors.

To ensure proper implementation by line personnel, legislation and company standards may require more detailed guidelines or guidance notes. Such guidelines, generally drafted by professionals in the central HSE department on the basis of their own experience and outside information, provide site managers with key principles that are to be incorporated into local procedures.

Maintaining the currency of all documentation of applicable laws and standards, guidelines, and local procedures, and communicating modifications and additions to the affected personnel, is a difficult task; a written procedure with clear allocation of responsibilities may be necessary.

2.5.2. Measuring Environmental Performance

"You can't manage what you can't measure." Therefore, to measure the effectiveness of an environmental management system, real data are required on the precise effects of the company's activities on the environment and individual people. It is thus necessary to conduct a planned, regular sampling and monitoring program at the various company sites as well as outside the company (e.g., regarding the effects of certain products).

It is generally accepted that the parameters to be monitored should include:

- Point sources and fugitive emission to the atmosphere
- Point sources and fugitive discharges to water or sewers
- Solid and other wastes, particularly hazardous wastes
- Contamination of land
- Use of water, fuels, and energy
- Discharge of thermal energy, noise, odor, and dust
- Effects on specific parts of the environment and ecosystems
- On-site accidents and near-misses
- Personnel injuries
- Transport accidents
- Complaints from community residents
- Complaints from customers related to HSE matters

All these data should be obtained by prescribed or accepted (standardized) measurement techniques and collected in registers established at various levels in the company. Some companies collect only absolute figures, while others combine certain data with appropriate weighting factors into an index or several indices. Such indices may be useful management tools for fixing objectives and demonstrating improvements.

Results of all measurements are communicated to the various interested parties, from plant operators to senior management, at the appropriate frequency and in the relevant form.

2.5.3. Environmental Programs

On the basis of global, long-term targets, set at the highest achievable level (e.g., 50 % reduction in waste generation within the next 10 years), together with data from the register of environmental effects, individual environmental programs are established for the various sites and units of a company. Programs usually cover a one-year period within a multi-year global program.

Proposals for individual programs are checked for feasibility, relevance, and priority, usually by the central HSE department. A proposal for a corporate environmental program is then submitted to senior management, which decides on the proposals to be accepted and allocates the necessary human and financial resources. The corresponding investments are included in the global investment program of the company.

It is not easy to determine the amount of capital investment currently devoted to HSE by chemical companies, in part because there is no established methodology for computing these expenses. Furthermore, while it may be reasonably easy to determine the cost of end-of-pipe equipment, how is one to isolate the environmental component in the cost of a "cleaner," integrated process? Estimates of capital expenditure related to HSE average 10 – 20 % of total capital investment for most companies. Operating expenses for HSE are in a similar range, and these are expected to increase, at least as long as processes intrinsically safe and "clean" are not generalized.

2.6. Environmental Audits

All companies have long been accustomed to having their financial management systems audited by accredited experts. A few decades ago the auditing of safety practices became a current practice for many companies. Only a few are so far familiar with the concept of "environmental auditing," but a trend in this direction is spreading rapidly across industry.

2.6.1. Definition

The International Chamber of Commerce (ICC) has proposed the following definition for an environmental audit [4]: "A management tool comprising a systematic, documented, periodic, and objective evaluation of how well environmental organization, management and equipment are performing with the aim of helping to safeguard the environment by:

1) Facilitating management control of environmental practices
2) Assessing compliance with company policies, which would include meeting regulatory requirements"

Such a definition might appear somewhat narrow when one considers that many corporate policies tend to cover unregulated risks. Indeed, the United States Environmental Protection Agency (EPA) has included such unregulated risks in its definition of environmental auditing.

Environmental audits are therefore a necessary component of a company's environmental management system, permitting a determination of whether or not the system is implemented effectively and is suitable for fulfilling the company's environmental policy.

Given the many activities of a company that have a bearing on the environment, environmental audits may serve various purposes:

Compliance audits verify that an industrial activity meets its legal obligations (permits, emission limits, etc.). They do not differ in principle from inspections by government inspectors.

Health and safety audits are still carried out separately in many companies, even though the tendency now is to integrate them into global environmental audits.

Site audits are the most common type of environmental audits. They are designed to check individual site activities against those elements of a company's environmental policy and management system applicable to the site in question. For large sites, audits may be limited to a single plant or a series of units.

Specific issues may also be the subject of an environmental audit at the corporate or site level, such as waste management, transport and distribution activities, etc.

2.6.2. The AuditPlan

To ensure proper verification of a global environmental management system, auditing of the various facets of a company's activities should be carefully planned.

The central HSE department is generally responsible for establishing an appropriate *audit plan*, which should deal with the following points:

1) Setting priorities
2) Establishing specific areas and activities to be audited and the frequency for auditing each particular site as a function of the nature and risk of its activities as well as conclusions reached in the previous audit; experience shows that a frequency of one to five years covers most situations encountered in the chemical industry
3) Objectives of the various audits
4) Personnel requirements
5) A protocol for conducting the audits
6) Procedures for reporting audit findings

2.6.3. Environmental Auditors

It is generally accepted, even in legislation (see Section 4.4), that environmental audits should be performed either by personnel from within the company (but wherever possible independent of the unit being audited) or people external to the company. A growing number of organizations and consulting companies are offering their expertise for environmental audits, which may be useful when internal expertise or capacity is lacking. Because of the rapid development of this market, great care should be exercised in the selection of an external consultant. Simultaneously with the publication of the ISO 14 001 Standard for Environmental Management Systems, ISO issued in 1996 a series of guidelines for environmental auditing:

– ISO 14 010: General principles on environmental auditing
– ISO 14 011: Audit procedures – Auditing of environmental management systems
– ISO 14 012: Qualification criteria for environmental auditors

A large number of environmental auditors have already been accredited to carry out third party ISO 14 001 audits according to the abovementioned auditing requirements.

In its efforts to make the ISO 9000 and ISO 14 000 series of standards more compatible, ISO decided in 1998 to develop one common auditing standard for quality and environmental management systems, replacing ISO 10 011 and ISO 14 010/14 011/14 012.

Companies may want to carry out environmental audits using their own teams of auditors, supplied either by the central HSE department or drawn from various line activities within the company. It is highly recommended that the auditing requirements expressed in ISO 14 010/14 011/14 012 should be taken into consideration. A typical team includes three to four persons under the recognized authority of a leader. At least two members of the team should be experts in environmental matters, and one or more should be familiar with the particular type of activity to be audited. All members should have the necessary skills to obtain, from observation and interviews, clear, unbiased information on the operation in question without being regarded as "policemen" by the local employees. To facilitate contacts and technical arrangements an employee of the unit subject to audit may be nominated to act as a liaison with the auditing team.

The time required for an environmental audit varies considerably according to the size of the site, the nature of the operation, the cooperation of local personnel, etc. It may range from a few days to several weeks.

2.6.4. AuditMethodology and Process

The detailed actions required to carry out an environmental audit are highly dependent on the type and nature of the activities to be audited, but the following features are common to most audits:

1) Identification of elements of the management system, including instruction manuals, written procedures, communication, and reporting schemes
2) Evaluation of the control system
3) Assessment of strengths and weaknesses of the management system, especially deviations from written procedures and failures in the reporting lines
4) Examination of records of data on controlled or accidental emissions, energy use, complaints, etc.
5) Evaluation of training and motivation of personnel

Auditors collect the necessary information through direct observation and interviews of personnel at the plant, drawn from all levels. Cross-checking of information is a common practice, but care should be exercised to avoid friction with and between employees. Contacts may be facilitated by starting the audit activities with a brief meeting with plant personnel, where the audit objectives, plan, and methods are explained. The full cooperation of local management is in any case required.

Immediately at the conclusion of an audit the leader of the auditing team should present the major findings to the local (plant or site) management. Already at this stage a number of immediate corrections can often be achieved.

2.6.5. AuditReport and Action Program

The final report is drafted by the team leader, and usually checked by the site manager before being sent to senior management (President and/or Board).

A typical report includes three parts:

1) An executive summary of major findings
2) A section on detailed observations, the part most useful to local management for improving the operating procedures
3) A proposal for corrective actions, including capital investment, major equipment modification, or changes to managerial procedures when appropriate

The corrective action program is discussed at senior management level. When approved, in conjunction with an appropriate time scale, it is integrated into the global environmental program of the company.

2.7. Environmental Management Reviews

The suitability and effectiveness of an environmental management system must be assessed at regular intervals. This is accomplished at regular management meetings involving senior managers of the company together with the head of the central HSE department. Such meetings are intended to review the latest data on environmental performance, trends, deviations from predetermined objectives and targets, how effectively the environmental program has been implemented, etc.

When necessary, a review meeting may lead to decisions to modify or adapt the environmental objectives, the program, or even the policy if this seems justified.

With this last element the environmental management system "loop" is closed.

3. Environmental Communication

3.1. Environmental Communication: A New Aspect of Management

Traditionally, industry and particularly the chemical industry have not been models of good communication with society; indeed, secretiveness and confidentiality were deeply entrenched in the cultures of many reputable companies.

Starting with the disclosure of financial information, however, enterprises have increasingly adopted a more open attitude.

A similar evolution has occurred in the environmental domain: a few decades ago, no company would have considered going beyond minimal "instructions for use" on the labels of their products. Then came recognition of the "duty-to-inform," mostly in the form of information provided to the local community in the vicinity of an industrial site regarding actions to be taken in case of an accidental emergency at the plant. Finally, the public "right-to-know" was recognized by industry, first in North America, then in Europe, and ultimately even in Japan, where public-opinion pressures are less severe.

This is not to say that there is now full transparency on HSE matters between industry and society. It is worth noting in this context that environmental organizations talk in terms of "corporate disclosure," whereas industry prefers to talk in terms of "corporate communication and reporting." Some companies still believe that the public is interested only in being told about risks to which they are exposed, not absolute figures regarding emissions of dangerous substances. Industrial property rights are often used as an argument to limit disclosure.

Nevertheless, environmental communication is now featuring in the environmental policy of a growing number of companies, and it is developing rapidly across the world chemical industry.

A number of factors explain this trend:

1) The disclosure of environmental information has become compulsory under some existing national legislation (see Section 4.4). In the future more legal requirements are expected in many countries.
2) Within the broad concept of "sustainable development" (→ Sustainability and the Chemical Industry – Ethical Aspects and Approaches) environmental reporting has been identified as a major challenge for industry. As mentioned in the Agenda 21 document [5], "companies are encouraged to report annually on their environmental records, as well as on their use of energy and natural resources."
3) The credibility of industry, in particular of the chemical industry, is poor in the eyes of the public. Providing data on environmental performance (showing not only positive achievements but also difficulties and failures) can help restore industry's image and build credibility.
4) Customers have begun to integrate environmental considerations, including the environmental performance of suppliers, into their purchasing decisions.
5) Shareholders, financial institutions, and the insurance branch require information to help them assess the real present and future financial risks to which companies they are dealing with are exposed.
6) Employees and local communities need information about their working and living environments.
7) Environmental NGOs are likely to campaign more aggressively against corporations that give the impression they have something to hide about their environmental performance.

In fact, more and more companies consider environmental reporting to be a definite competitive advantage and a major factor in their "license to operate".

3.2. Audiences and Vehicles

Many different stakeholders in society have an interest in environmental information supplied by industry. The key ones include:

Local and national governmental agencies
Local communities
The general public
Media (general and specialized)
Employees
Shareholders
Banks, financial institutions, and insurers
Domestic and trade customers
Schools, colleges, universities, and business schools
Environmental NGOs

Satisfying the needs of these various audiences may require specific vehicles for environmental communication. Some examples of appropriate communication forms are:

Labels. The chemical industry has for several decades been using regulatory or standardized labels, in particular for hazardous products. Specification sheets, such as the standardized "Material Safety Data Sheets" (MSDS), provide

supplementary information on the hazards and risks of chemicals for various user categories.

Product advertising increasingly includes some environmental information. To limit the use of ill-founded or misleading statements, certain industrial organizations have drafted special guidelines on the matter [6]. The eco-labeling schemes operative in the EC and other countries represent a form of regulated environmental advertisement.

Media, including the general press and TV, are becoming increasingly interested in the environmental performance of enterprises. However, they are attracted more by bad news than good news, and the publicity afforded by the media to unfortunate accidents or incidents may seriously endanger a company's reputation.

Company newsletters and *magazines* offer a privileged form of support to environmental information directed at employees or shareholders.

"*Open house"* days and *plant visits* provide the best opportunity to convey environmental information about a site to local communities (including local media and environmental organizations), to employees, and to schools and other educational bodies.

"Hot lines", toll-free telephone numbers, and *information centers* are being made available by some companies to their customers and the public.

3.3. The Environmental Report

The concept of an environmental report, especially an annual environmental report, is becoming more and more popular among corporations in North America and Europe, and also Japan. Many companies consider the annual environmental report as a key element in their environmental communication strategy.

Because this concept is still in its infancy (the first full free-standing environmental reports were published in the late 1980s or early 1990s), there is as yet no standard or commonly accepted format for such reports. What follows is therefore simply a review of current trends, which may evolve significantly in the coming years.

3.3.1. Audiences

The annual environmental report is certainly the communication vehicle best suited to the needs of a number of audiences, in particular the employees, local communities, shareholders and financial institutions, environmental organizations, etc.

This underscores a major difficulty in the concept of the annual environmental report: finding the right style and the right balance between technical information, financial data, and easy-to-understand words, graphs, and figures. From the experience gained so far it is apparent that some reports would benefit from an improvement in style and presentation.

3.3.2. Corporate or Site Reports

Perhaps for fear of local adverse reactions, some corporations publish only a report covering the company as a whole, with consolidated data. However, if the report is to meet the needs of local communities and local audiences it is obvious that data pertaining to the site in question must also be published.

This fact is accepted by a growing number of companies. Some publish a single report with a section for the entire corporation together with various sections dealing with individual sites; others publish a corporate report supplemented by free-standing individual reports for their (main) sites.

3.3.3. Contents of the Report

Some reports are still mainly textual, providing only a few numeric figures for principal indicators, generally consolidated at the corporate level. Certainly reports should include a significant textual part, presenting, for example, the company's environmental policy, its environmental program and organization, the major achievements, objectives, and targets. However, a growing need is perceived to publish numerical data for both corporate and site reports as a response to the now recognized "right-to-know" of the public. Indeed, publishing such numerical data within a report offers a definite advantage

over provision of the "dry" figures required under certain legislation: the figures are put in the perspective of a real business (size of the site, products manufactured, etc.), thus avoiding – at least partially – easy misinterpretation by certain groups.

It is mentioned in Section 2.5.2 that some companies establish "environmental indices" from measured data on emission. Certainly such indices are useful for demonstrating the trends in environmental performance, but using them as a substitute for the publication of absolute figures can generate harsh criticism from some concerned stakeholders.

Another issue of concern is whether reports of all companies (at least in the same industrial sector) should include data covering the same parameters for a company or a site. This is certainly reasonable in terms of environmental relevance. Furthermore, one cannot deny the right on the part of many stakeholders (including regulatory authorities) to undertake certain comparisons between companies and sites, and this consideration would also militate in favor of a certain comparability of data. At least a "core list" of common parameters should be included in all reports.

In order to improve transparency and consistency of the reports the European Chemical Industry Council (CEFIC) issued in 1998 the Responsible Care – Health, Safety and Environmental guidelines. It lists 16 core parameters and their definitions according to which all chemical companies will report. The collation of comparable data will allow monitoring continual improvements.

3.3.4. Links to Environmental Management Systems

Environmental communication is not *stricto sensu* an element of the EMS. However, it is becoming more and more accepted that the results of environmental audits – at least those not containing confidential information, the disclosure of which could endanger competitiveness – must be made public. The EC Regulation on EcoAudits (see Section 4.3) requires the publication of a "public statement" after every audit. Site environmental reports would be a perfectly appropriate support vehicle for these statements.

3.3.5. Trends

There are many indications that corporate environmental reports, published at regular intervals, will eventually become obligatory for all major companies, in particular in the chemical industry.

The quality and quantity of information disclosed in the reports is likely to evolve. A research study [7] has identified five stages in the way environmental reports develop to meet the stakeholders' information needs:

1) Green "glossies," newsletters, videos; short statements in annual reports
2) One-off environmental reports, often linked to a first formal policy statement
3) Annual reporting, linked to an environmental management system but providing more text than figures
4) Provision of full environmental performance data on an annual basis, available on diskette or on-line, with an environmental report cited in the annual report
5) Sustainable-development reporting linking environmental, economic, and social aspects of corporate performance and supported by indicators of sustainability

Presently, most examples of environmental reports meet the definition of stage 3; only a few have reached stage 4. To attain stage 5 companies will need to develop and adopt methodologies for full environmental cost pricing and life-cycle assessment of their activities and products.

3.4. Stakeholder's Reactions

The argument most commonly used by those reluctant to embark upon environmental communication is a fear of adverse reactions from various groups within society. Experience shows, however, that with the necessary skills and precautions, the release of environmental information and in particular of annual environmental reports generates only occasional negative reactions, and that the benefits definitely outweigh the difficulties.

For example, local communities and employees always welcome environmental information, especially that related to specific sites. Shareholders and financial analysts have begun to

take an interest in environmental reports, but are likely to demand deeper assessments. The media have reacted positively to the release of environmental reports, but may be critical of reports that are too "textual." Environmental organizations are for the moment granting the benefit of doubt to corporations strongly involved in environmental reporting.

Certainly the credibility of information released by companies, in particular by those with a poor public image, may become an issue, and it will take some time for companies to demonstrate that the information they publish is candid and accurate. This has prompted some companies to request validation of their information, in particular of their environmental reports, by independent third parties, such as accredited inspection organizations. For the moment, few reports have received such external validation, but several companies envisage having future reports validated by reputable, undisputed third-party organizations.

3.5. Who Communicates?

Many companies still have a policy of relegating the responsibility for environmental communication to their professional communicators, backed as far as technical information is concerned by staff people in the HSE department. Professional communicators will always be needed to coordinate a company's information policy and define its rules and practices, but there is a recent trend toward inviting line and staff managers to participate in environmental communication. For example, open house days obviously involve many employees, who must be prepared to answer questions about a company's environmental performance. Site managers or superintendents may also be interviewed by journalists or media reporters, and employees are the company's best ambassadors in the local community.

Involving various levels of HSE managers in communication becomes all the more important when a communication crisis bursts following an accident in a plant, for example, or at any event attracting the attention of the media. All companies and sites should establish crisis management plans, duly tested and identifying all the major players, especially with respect to communications. There have been too many examples of accidents or isolated incidents that have ruined several years of effort by reputable companies to build trust with the public, simply because communications have been poorly managed during the event.

Many companies now include communication as an element in their training programs, in particular in the training of their HSE specialists.

4. The Impact of Legislation on Environmental Management

Complying with regulatory standards is clearly an important element of any environmental management system. The purpose of this section is not to describe all legislative texts applicable to industrial activities, but rather to identify major trends in environmental legislation influencing the management of chemical companies.

4.1. Operating Permits

Any industrial operation is subject to the granting of a permit by the competent authority. The dossier submitted with an application should include all information required to evaluate the environmental impact of the planned operation. It is then up to the authority to establish conditions that may be attached to the permit. These conditions might in some cases mandate a revision of certain processes or procedures as they were initially envisaged by the applicant. The permit process therefore implies a dialogue between the authority and the applicant, and it also implies the release of information to members of the public who might be affected by the planned operation. Even if the power of decision remains with the authority, the result of such consultation might be quite important for some "sensitive" operations, such as waste disposal sites. The whole permit process can extend over several years, which underscores the extent to which environmental questions can impact on industrial management decisions.

Certain conditions attached to such a permit are of a managerial nature (e.g., emergency plans, reporting lines, etc.). Most of them, however, establish emission limit values (ELVs) for

a list of relevant substances or indicators (waste, energy, noise, etc.) that the planned operation is expected to meet at all times.

In theory, these ELVs should result from a risk assessment process, starting with a definition of environmental quality objectives (EQOs) for the area concerned. Defining an EQO is a political decision by the authority: the fact that the area in question is of a residential nature, for example, or that rivers should continue to sustain aquatic life, and so forth. EQOs lead to the determination of environmental quality standards (EQSs) for every environmental compartment, such as the maximum concentration of SO_2 permitted in the atmosphere or heavy-metal limits in water, the maximum level of noise, etc., which in turn make it possible to achieve the EQOs. EQOs are fixed at the local level, or more often internationally, on the basis of scientific assessments. Finally, a model is used to calculate various ELVs of the planned operation to ensure that the EQOs are met, taking into account the existing levels of various parameters in the area concerned. This process, although complicated and possibly leading to discrepancies between the conditions of similar operations in different areas, is the only one that can assure that the environmental risk of an industrial activity is limited to an acceptable level while at the same time being optimally cost-effective.

Nevertheless, another process is becoming increasingly popular for establishing ELVs: the use of *"best available technique*" (BAT). A BAT may be defined as that set of technical and managerial processes and procedures available at a given time which provides for the most cost-effective manufacture of a certain product with minimal impact on the environment. The use of BAT is attractive to the authorities because it simplifies the process of establishing ELVs. It may also attract industrialists, because it ensures a "level playing field" for all companies in a particular business. However, it does raise certain questions, such as how long a BAT remains a BAT before becoming outdated by a new development, the possibility that some developers might enjoy a monopolistic situation, or the best way of ensuring that EQOs are met in heavily industrialized areas.

Whatever the process chosen for determining the ELVs, it is necessary that an integrated approach be adopted so that the impact on all compartments of the environment is minimized: taking care that atmospheric pollution, for example, is not transferred to the aquatic environment. A series of modern legislative texts follows this trend, including the EC Directive on Integrated Pollution Prevention and Control (IPPC).

4.2. Process Safety Legislation

Safety, because of its probabilistic aspects, cannot be regulated simply by standards. Assessing the accidental risk of an activity with respect to both equipment and management practices is therefore the key to safe operation (→ Plant and Process Safety).

In principle, regulation is impossible because, at least in the chemical industry, every case is a special one. However, legislation can require an operator to carry out a risk assessment and present the results to the authorities. Here again, a dialogue between the competent authority and the applicant will determine the level of acceptable risk and various conditions the activity should satisfy to meet this acceptable risk.

EC Directive 96/82/EC, the "Seveso" Directive, now being revised, is typical of legislation that fixes not a standard, but a management practice (i.e., the risk assessment procedure) in order to achieve its objective.

4.3. The EC Eco-Management and AuditRegulation

In 1993, the EC Council adopted a Regulation (EEC/1836/93) introducing a "Voluntary Community Scheme of Eco-Management and Audit (EMAS)." The initial intent was to facilitate development of environmental auditing practices within industry, but it was soon recognized that environmental audits could not be isolated from environmental management systems, and the title and content of the regulation were modified accordingly.

An enterprise deciding to take part in the scheme is committed to take the following actions:

1) Carry out an initial review of its environmental performance

2) Establish an EMS
3) Conduct periodic environmental audits of its various sites
4) Produce a "public statement" after each audit
5) Have the public statement "verified" by an independent and accredited third party

When the competent national authorities have satisfied themselves that a particular enterprise has complied with the requirements of the regulation they then grant what is known as a "statement," which the company is free to use at its discretion.

The regulation includes annexes specifying the elements of an EMS, the content and methodology of environmental audits, the role of verifiers, and the requirements for accreditation.

At the end of 1998, some 2 000 industrial sites won EMAS registration. Around two thirds of these registrations are in Germany, whereas the take-up elsewhere has been patchy. Many companies went to the ISO 14 001 certificate instead and were not willing to duplicate part of the efforts in order to get an EMAS registration, which is limited to Europe and does not offer tangible additional benefits.

The Regulation is currently revised and is likely to show the following features:

– EMAS will no longer be limited to "sites" but will be open to "organizations", which is in line with ISO 14 001
– The environmental management requirements of ISO 14 001 will be taken over as this will facilitate the step towards EMAS
– EMAS will retain a number of specific aspects which, in addition to the (extended) environmental statement, will differentiate it form ISO 14 001

Industry welcomes the alignment with ISO 14 001 but remains concerned that in the absence of any regulatory relief for EMAS registered companies many companies will stay with ISO 14 001

4.4. Mandatory Environmental Communication

Operating permits require an operator to monitor emissions and all other parameters subject to ELVs, and to report the measured results to the competent authority. In many countries, including the EC Member States, legislation of the "Freedom of Information" type make any information (especially environmental information) collected by governmental authorities available to the public, either upon request or directly at the initiative of the authorities themselves. In addition, many statutes related to accidental risk, such as the EC Directive on Major Hazards (the "Seveso" Directive) require active informing of any population likely to be affected by an accident from a chemical site [8]. Through this mechanism a form of mandatory environmental communication already exists in many countries.

Some countries have gone further in requiring industrial activities to measure the emissions of a specific list of substances and then report them to a "register" freely available to the public, sometimes on-line or on diskettes. In the United States, for example, the 1986 Emergency Planning and Community Right-to-Know Act (EPCRA) introduced under Section 313 of Title III the concept of the Toxics Release Inventory (TRI) [9]. Under this legislation, companies employing more than 10 full-time employees must provide the EPA with annual emission data for some 300 toxic chemicals manufactured, processed, or used in excess of certain threshold amounts. Data on emissions released and waste generated by individual sites are made available to the public through printed reports and electronic databases.

Certainly the introduction of the TRI has met the expectations of many stakeholders; it has also been a powerful incentive to industry in the United States to monitor and reduce its emissions and waste generation. It should be recognized, however, that making raw data available out of context opens the door to considerable misinterpretation, for example, identifying the largest companies or sites as "major polluters."

Certain other countries, in Europe and elsewhere, also maintain registers of the TRI type, albeit ones that are less comprehensive and less readily available. The EC has been considering introduction of a "Polluting Emission Register," possibly modeled along the lines of the TRI.

5. Voluntary Action by Industry

Environmental legislation, however comprehensive and detailed it may be in many industrialized countries, cannot cover all the cases encountered in day-to-day industrial practice. Moreover, legislation is rigid, and it cannot be adapted to follow the pace of technological and managerial development. It is also not the most cost-effective approach, because it requires that resources be committed to inspection and enforcement. Finally, the pressure of legislation is not perceived as a real motivation on the part of the workforce.

This has induced industry, led by a few large, advanced companies, to develop voluntary environmental activities and programs. These steps have been further encouraged by industrial associations and supported by certain enlightened governments. Some of the resulting acts can be seen as full substitutes for legislation. More often they complement existing legislation in facilitating its implementation and bridging gaps that legislation cannot cover.

Voluntary corporate action can take several forms. These are described briefly below, before focusing on the worldwide chemical industry initiative known as Responsible Care.

5.1. Internal Statements

Several elements of the environmental management systems discussed in Chapter 3 constitute voluntary actions, provided they are not made mandatory, for example, through an extension of the EC Eco-Management and Audit Regulation (see Section 4.3). These include in particular:

1) The HSE policy (see Section 2.3); this policy, endorsed at the highest level of corporate management, states the principles and objectives that govern management as far as HSE matters are concerned
2) The environmental program (see Section 2.5.2), which defines actions to be taken in a certain area to reach an objective within the framework of a set timetable

Both overall environmental policy and specific environmental programs are internal management tools, even if many companies publicize them (or at least a few of their environmental programs, avoiding the disclosure of competitively sensitive information) in their annual environmental reports. Experience shows that the drafting of a set of environmental programs contributes greatly to the development of environmental awareness within the workforce.

5.2. Voluntary Agreements

At the other extreme of the spectrum are "voluntary agreements" signed between a company and an authority for the purpose of achieving a target, with the choice of means left to the company.

The term "voluntary agreement" may not appear strictly appropriate: in one sense, entering into an agreement is always a voluntary action; on the other hand, the alternative to not signing such an agreement is, in most cases, a law. Other terms are sometimes used as well, in particular "covenant."

The components of a voluntary agreement are:

1) The *aim*; e.g., an emissions reduction target, a labeling scheme, the phaseout of a certain product, etc.
2) The *parties to the agreement*:
   a) On the governmental side, a local authority (e.g., for reduction of discharges to water from a site), a national government (for a global plan involving an industrial sector), or the EC Commission (although monitoring the implementation may be left to national authorities).
   b) For industry, the most straightforward case is when the agreement is signed by a company, but when the agreement covers an entire industrial sector at the national or regional level, the responsible party is usually the trade association concerned. This immediately raises the problem of the extent of the authority an association may be able to wield over its members.
3) The *penalties* in case of failure to achieve the aims of the agreement. Some agreements include financial penalties; this may lead to serious difficulties if the agreement has been signed by a trade association. More often the penalty is simply that a statute is subsequently enacted, which is more burdensome and less cost-effective for both parties.

4) A *monitoring system* and an *administration* to facilitate implementation of the agreement. Experience has shown that this can be done through light, cost-effective organizations.

The following examples are representative of voluntary agreements within the chemical sector:

1) The "33/50 Program" signed between the U. S. Environmental Protection Agency and a number of chemical companies for the voluntary reduction of their emissions of 17 of the most polluting substances by 33 % in 1992 and by 50 % in 1995 relative to 1988 levels
2) EC Commission Recommendation 90/437/EEC of June 1990 formalizing a voluntary program of reduction in the use of chlorofluorocarbons by the foam plastics industry, represented by its European associations
3) EC Commission Recommendation 89/542/EEC introducing a voluntary system of detergent labeling under the control of the European industry federation of detergent manufacturers
4) The various "covenants" signed between the Dutch government and Dutch industry associations for implementing the National Environment Policy Plan (NEPP) of 1989, which established a long-term policy and a program of reduction of emissions, negotiated within the various sectors

5.3. Public Commitments

Half-way between internal statements and binding agreements are the various forms of public commitments by which a company, through its President or CEO, commits itself to adopting an environmental behavior expressed in a set of principles.

There are indeed certain similarities between this form of voluntary action and the voluntary agreements discussed above:

1) The principles define the aims of the commitment. In general they are drafted in the form of a set of 10 – 15 short guidelines. For effective, practical implementation these must be supplemented by Codes of Practice or Guidance Notes.
2) The principles are drawn up by an organization such as the national or international trade association of the industrial sector concerned, and then endorsed publicly by the senior management of individual member companies.
3) The other party is the public, in general represented by environmental non-governmental organizations (NGOs) that check company environmental performances against the content of the corresponding guiding principles. Companies may suffer a severe penalty when failure to implement the principles is publicized by a credible NGO.

The best known example of such a commitment is the "Business Charter for Sustainable Development" launched by the International Chamber of Commerce (ICC) in April 1991, and now endorsed by over 1100 major corporations around the world [10]. The Charter is a set of 16 principles covering various aspects of the integration of environmental considerations into management practice, such as: environmental policy, continuous improvement of environmental performance, employee education, research, technology transfer, impact assessment, product and services stewardship, sound design and operation, contractors and suppliers, emergency preparedness, dialogue with stakeholders, and reporting.

The ICC Business Charter for Sustainable Development is recognized as a major driving force for continuous improvement in the environmental performance of enterprises through voluntary public commitment. Similar sets of guiding principles have been established by a number of national and international organizations. Some are less comprehensive than the ICC Business Charter, especially with respect to the requirements on research, product stewardship, and reporting; others stress specific aspects, such as emergency preparedness or overseas operations. Two of these Codes are perhaps worthy of special mention.

One is the set of CERES principles (formerly the "Valdez" principles), drawn up in 1989 by the Coalition for Environmentally Responsible Economies (CERES) after the Exxon Valdez oil spill [11]. Signatory companies, based mostly in

the United States, are particularly committed to environmental reporting (see Chap. 3) and compensation for environmental damage.

A second is the Global Environment Charter, launched in 1991 by Keidanren, the Japanese Federation of Economic Organizations [12]. The Keidanren Charter emphasizes the role of corporations in technology transfer, including transfer to developing countries, but does not require public reporting.

5.4. Responsible Care

Several chemical companies are signatories to the ICC Business Charter or other sets of principles like those mentioned above. But most also take part in "Responsible Care," the chemical industry's own initiative that originated in 1984 in Canada and is now spreading throughout the world [13].

5.4.1. Responsible Care; What Is It?

Responsible Care relies on two fundamental features:

1) A commitment by company leaders to seek continuous improvement in the environmental performance of a company's activities, including employee involvement and relations with customers and the society's stakeholders. This commitment is expressed by the signature of the company President or CEO on a set of Guiding Principles.
2) A program established and operated by national chemical industry associations to assist their member companies in implementation of the Guiding Principles and verification of the true achievements resulting from their commitments.

5.4.2. The Guiding Principles

There is not a single set of Responsible Care Guiding Principles: these are instead established on a national basis by the corresponding chemical industry associations with consideration directed to local circumstances. However, the Principles everywhere are remarkably similar in content and format; indeed, they are all modeled along the lines of other well-known principles, such as the ICC Business Charter for Sustainable Development or the CEFIC Guidelines for Protection of the Environment.

5.4.3. National Programs

Similarly, the details involved in facilitating and monitoring specific programs are left to national chemical industry associations, who must adapt them to local considerations such as the national legislative framework, the state of industrial development, the companies' cultures, etc. However, to ensure consistency in the initiative in the various countries, all national programs are expected to include seven fundamental features:

1) A formal *commitment* on behalf of each company to a set of guiding principles (see above. In the majority of cases the commitment is signed by the company CEO, but in some countries it is fulfilled by endorsement by the Board of the national association).
2) A series of *codes, guidance notes*, and *checklists* to assist companies in implementing the commitment. In some countries (e.g., the United States) a set of detailed codes has been drafted to cover the following areas: control of emissions, process safety, occupational health and safety, transport and distribution, community awareness and emergency response, and product stewardship. Elsewhere, especially in Europe, relevant associations had already developed a number of guidance documents on environmental management long before the introduction of Responsible Care. Producing and updating such codes and guidance documents is a continuing process.
3) The progressive development of *indicators* against which improvements in performance can be measured. In general, these indicators include sets of emission data for a number of selected substances and data on waste generation and recycling, as well as data on energy consumption and accidents.
4) An ongoing process of *communication* on HSE matters with interested parties outside the industry. The issue of environmental communication, discussed in Chapter 3,

is therefore an integral part of Responsible Care. Obviously the recommendations provided by the association to its member companies regarding the communications process influence the selection of indicators of performance.

5) Provision of *fora* in which companies can share views and exchange experiences on implementation of their commitments. Some associations provide extensive service to their member companies in the organization of seminars, workshops, local cells, etc. This is particularly valuable for increasing the motivation of people who in general do not respond easily to environmental concerns, such as sales managers or line middle-managers.
6) Adoption of a *title* and *logo* that clearly identify a national program as being consistent with and part of the concept of Responsible Care. The standard logo (Fig. 4) has been registered in many countries, and is used throughout the world by the chemical industry. Many associations, even in non-English speaking countries, have retained the registered term "Responsible Care;" others have adopted a different expression that communicates in translation the spirit of the term Responsible Care.

Figure 4. The Responsible Care logo

7) Consideration of the best ways to encourage all member companies to *commit to* and *participate in* Responsible Care. Many associations have made or will make company commitments to Responsible Care a condition of membership. This implies that if a company does not fulfill its commitment, the association has the right to expel it. Before taking this extreme action, however, strong peer pressure will be exerted by other members on the noncomplying company.

5.4.4. The Scope of Responsible Care; Product Stewardship

Responsible Care is sometimes understood as covering nothing more than the manufacturing activities of a company. Fundamentally, however, the Guiding Principles of Responsible Care apply to all aspects of industrial operation, from basic research to product development, manufacture, distribution, and use. They are thus more nearly consistent with a "cradle-to-grave" approach.

In this respect, product stewardship is a vital component of Responsible Care. Product stewardship may be defined as "the responsible and ethical management of the health, safety, and environmental aspects of a product throughout its life cycle, from its design to manufacture and distribution, including resource consumption, transport, and final disposal options."

Responsible Care therefore aims at developing an "environmental culture" among all employees of a company, whatever their position and function. Researchers are motivated to assess the HSE effects of a new product at an early stage of development, and to find substitutes if the risk appears unacceptable. Engineers develop processes with a concern for minimal consumption of resources and the lowest possible impact on the environment. Production people operate their plants safely, and with full respect for environmental standards. People in charge of transport and distribution seek the safest means of transport. Sales managers advise their customers on the safe use of products, including problems raised after normal use to avoid creating waste problems.

5.4.5. The Challenges of Responsible Care

The successful development of the Responsible Care initiative by chemical industry in many

countries throughout the world entails two major challenges:

1) Finding ways to maintain the integrity of the initiative as national associations adapt their programs to local conditions and cultures – which may vary significantly among countries like the United States, Japan, and the nations of South America, Europe, or Africa. In addition, Responsible Care is a chemical industry initiative, but other industrial sectors more or less closely related to the chemical industry would benefit from similar programs adapted to their specific needs.
   The International Council of Chemical Associations (ICCA) has established a procedure for checking the consistency of programs developed by national associations against the seven fundamental features, both at the time of introduction of the program and periodically thereafter. The ICCA has also established rules for "partnership" with other industrial sectors that deal with chemicals (transport and distribution companies, waste disposal contractors, etc.) and are also interested in Responsible Care.
2) Finding ways to make Responsible Care acceptable and credible to the public, regulatory authorities, and environmental organizations. Responsible Care is not simply a publicrelations exercise: it is based strictly on improved performance – not on words, but on actions. Credibility therefore will be gained by demonstrable achievements, which will need to be communicated effectively. Many national programs include "public advisory panels" consisting of representatives of the various stakeholders; others have developed alternative channels of communication, but all companies are encouraged to adopt an open attitude toward information.

Honest dialogue between industry and the public and between plants and local communities is a crucial element in bringing about the industry-wide cultural change required by Responsible Care, and in building public confidence and trust, a key to enabling the chemical industry to retain its "license to operate."

6. Standardization and Certification of Environmental Management Systems

As discussed in Chapter 2, companies first establish policies for health, safety, and environmental matters, and then set up environmental management systems to ensure that objectives defined in the policies are effectively achieved.

An EMS cannot really be defined by law, even if the EC Eco-Management and Audit Regulation (see Section 4.3) does describe the basic elements of an EMS and encourages companies to establish such EMSs. Conversely, leaving the companies without adequate guidance on how to establish an EMS may result in serious discrepancies in the structure and organization of EMSs as adopted by individual firms, with the accompanying risk of some stakeholders coming to question the credibility of the EMSs of certain companies.

This dilemma has led to the concept of standardization of environmental management systems. In addition to helping companies achieve their objectives, appropriate standards can, through certification, offer assurance to others that the stated objectives will be achieved.

6.1. Standards for Environmental Management Systems

6.1.1. Generic or Specific Standards

Although there is a long history associated with the drafting of standards for product specifications, testing methods, etc., the concept of standards for management systems is quite new: it indeed first appeared in the mid 1980s with the issuance of the series of standards for quality management systems ISO 9000 (and its national or regional equivalents, EN 29 000, BS 5750).

Whereas the ISO 9000 standards were restricted in their initial scope to that which was necessary to achieve product quality, it was soon recognized that the same requirements could also have a generic interpretation, and that, for example, the ISO 9001 and ISO 9004 standards could be adapted to cover environmental management systems.

Nevertheless, a few standardization bodies undertook the drafting of standards specific to

EMS, such as BS 7750 (see below), issued in 1992. A few similar national standards have since been issued elsewhere, and the International Organization for Standardization (ISO) is taking the same route.

A certain number of standardization bodies share with some industrialists the view that there should be further development of the concept of a generic management-system standard. The standard would include appendices covering specific requirements for product quality, environmental management, etc. The vision is that this structure would eventually replace the emerging sets of separate management-systems standards, but it probably will not come to pass for several years.

6.1.2. EMS Standards

In March 1992 the British Standards Institute (BSI) issued the first standard specifically designed for environmental management systems, under the reference BS 7750. In April 1993 the Association Française de Normalisation (AFNOR) published an experimental standard (X30-200) which, although somewhat different in its wording, follows the same structure as the British Standard.

Under the impulse of S. Schmidheiny from the Business Council for Sustainable Development, the International Organization for Standardization (ISO) established in 1991 a Strategic Advisory Group on the Environment (SAGE). The ISO-SAGE in turn created six subgroups, one of them in charge of making recommendations on a standard for EMS. In June 1993 the work of ISO-SAGE was taken over by a new Technical Committee (TC): ISO/TC 207, "Environmental Management."

Six subcommittees (SC) were set up

- SC 1 – Environmental management systems (ISO 14 00x series of standards)
- SC 2 – Environmental auditing (ISO 14 01x series of standards)
- SC 3 – Environmental labeling (ISO 14 02x series of standards)
- SC 4 – Environmental performance evaluation (ISO 14 03x series of standards)
- SC 5 – Life cycle assessment (ISO 14 04x series of standards)
- SC 6 – Terms and definitions ISO 14 05x series of standards)

The first standards were issued in 1996. Amongst them ISO 14 001 (Environmental management systems – Specification with guidance for use) which has since substituted all national EMS standards.

6.1.3. The ISO 9000 and 14 000 Standards

There has been a growing perception that although separate standards may be required for quality and environmental aspects for management systems, these standards should be made as compatible as possible. Compatibility is understood to mean that the common elements (e.g., training, auditing, document control, etc.) of the standards can be implemented without unnecessary duplication or the imposition of conflicting requirements.

In 1998, ISO therefore decided that ISO 14 001/14 004 should be revised in advance of its envisaged revision schedule, in an effort to make it as compatible as possible with ISO9001/9004 which are currently subject to fundamental revision. Both series of standards would then be simultaneously issued in 2000.

6.2. Certification of Environmental Management Standards

Standards for management systems, in addition to their "educational" aspects, also allow for "certification" (or "registration") of the activity covered by the standard by an independent, accredited third party, thus giving confidence to the company and outside stakeholders that the objectives of the management system will be achieved.

6.2.1. Certification of an EMS

Certification is a process in which a certification body (often a commercial firm) assesses a particular system against the relevant standard and then issues a certificate of compliance. This is followed up with a series of regular surveillance visits (usually once or twice a year). Certification bodies may themselves be accredited for this

function by a governmental or similar agency, lending additional credibility to their work and resulting in an "accredited certification."

ISO has developed standards for the auditing of an EMS and has furthermore decided to develop a common auditing standard for quality and environmental management systems (see Section 2.6.3). The latter will lead to significant savings for the industry as one single accredited certification body will be able to certify both management systems.

It can be expected that the existence of a certificate awarded by an international or national accredited body will become increasingly important to regulators and stakeholders as a form of assurance regarding the professionalism of the company and its EMS. This will assist many companies in meeting the key objective of being able to demonstrate and promote their management of environmental issues in a tangible and credible way. Lack of certification will in time increasingly expose a company to multiple audits by customers seeking to satisfy themselves with regard to the company's environmental probity.

6.2.2. Certification Against the EC Regulation

One article in the EC Eco-Management and Audit Regulation stipulates that "Companies implementing national, European or international standards for environmental management systems and certified, according to appropriate certification procedures, as complying with those standards, shall be considered as meeting the corresponding requirements of this Regulation . . ." (certification, however, will not exempt the "environmental statement" from being validated by a certified verifier).

The introduction of this article was of crucial importance as a way of linking the Regulation and voluntary participation on the part of companies to certification against the standards for EMS; because the public statement is an integral part of the scheme, it will quite normally be "verified" by the certification body as a standard part of the certification process.

The revised EMAS Regulation will now formally take on board the environmental management system requirements of ISO 14 001 (see Section 4.3).

7. International Aspects of Environmental Management

Most chemical companies, at least above a certain size, operate in several countries. In some cases all the countries may be at a similar state of development, but both industrialized and developing countries may also be involved. Especially in the latter case, special consideration must be given to environmental management to avoid the problems that too often have tarnished the images of "multinational" or "transnational" companies.

7.1. Environmental Requirements in Different Countries

Many reasons, mostly of a commercial nature, may prompt a company to establish an operation in another country, one of which might be a less stringent environmental legislative atmosphere in the host country. There have been a few cases in which "relocation" of a company's activity from its home country to another was perfectly justified because of the sudden introduction of exceptionally stringent environmental legislation not considered sensible or appropriate by most countries, even the most developed ones. Unfortunately, however, there have been more cases in which companies have established operations in countries with little environmental legislation simply to benefit from standards less demanding and costly than those reasonable standards in effect in their home country. Such "double-standard" practices, including in the environmental domain, have contributed significantly to the bad reputation associated with "multinationals" by some environmental organizations and the governments of some developing countries.

These practices are no longer considered to be ethical by reputable corporations that, in their foreign operations, comply not only with local legislation but also with their own company standards, which are generally equivalent to or exceed the requirements of their home country.

This is a basic principle expressed by many companies in their environmental policies.

It must be recognized, however, that companies of foreign parentage following this course of action put themselves in a position of competitive disadvantage vis-à-vis local companies satisfied with complying, at lower cost, only with national environmental requirements. Foreign companies can sometimes overcome this difficulty by "educating" both national companies (e.g., through the intermediacy of national chemical industry associations) and local governments in modern aspects of environmental protection. The role of transnational corporations in raising the environmental awareness and behavior of developing countries is now clearly recognized [1, Chap. 3.III].

7.2. Technological Cooperation

Technological cooperation between countries at different levels of development has been identified as a key component of sustainable development.

Technological cooperation between companies in industrialized and developing countries can take several forms:

Technical Assistance. A company or an organization sends experts to a company requesting expertise and assistance, perhaps for an audit of environmental performance, a troubleshooting mission, an education or training session, etc. The mission might be accomplished on a commercial basis, but more often it is carried out under the aegis of a voluntary program operated by an industrial association, a government, or an international organization.

Transfer of Technology. A company makes available, within the terms of a commercial contract, a process or a technique to which it holds title, generally as the result of a research and development program on its own premises. Such a transfer of technology should obviously include technological cooperation before, during, and after actual start-up of the operation.

Foreign Investment. A company may elect to set up its own operation in a foreign country, either in the form of a fully owned subsidiary or as a partnership with some other company at an appropriate level of equity. The industrial operation might be built from a green site, or the foreign investor might acquire an existing facility. In general, any such foreign investment implies a transfer of technology from the home company to the foreign one.

Several industrial associations have produced guidelines for an environmentally sound transfer of technology [15].

8. References

1. The World Commission on Environment and Development: *Our Common Future,* Oxford University Press, Oxford 1987.
2. S. P. Johnson: *The Earth Summit: The United Nations Conference on Environment and Development (UNCED),* Graham and Trotman Ltd., London 1993.
3. S. Schmidheiny (Business Council for Sustainable Development): *Changing Course,* MIT Press, Cambridge 1992.
4. International Chamber of Commerce: *Environmental Auditing,* Publication 486, ICC Publishing, Paris 1989.
5. S. P. Johnson: *The Earth Summit,* Graham and Trotman, London 1993, Agenda 21.
6. Confederation of Finnish Industries: *Guidelines for Environmental Arguments in Marketing,* Helsinki 1990.
7. Deloitte Touche Tohmatsu International, International Institute for Sustainable Development, Sustainability: *Coming Clean, Corporate Environment Reportin* g, Deloitte Touche Tohmatsu, London 1993.
8. M. S. Baram, D. G. Partan: *Corporate Disclosure of Environmental Risks, US and European Law,* Butterworth Legal Publ., Salem 1992.
9. U.S. Environmental Protection Agency: *Emergency Planning and Community Right-to-Know Fact Sheet,* EPA, Washington D.C., 1988.
10. International Chamber of Commerce: *Business Charter for Sustainable Development,* Publication 210/356A, ICC Publishing, Paris 1991. J. O. Willums, U. Goluke: *From Ideas to Action,* Publication 504, ICC Publishing, Paris 1992.
11. The Coalition for Responsible Economies: *The CERES Principles,* CERES, Boston 1989.

12. Keidanren: *The Global Environmental Charter,* Keidanren, Tokyo 1991.
13. European Chemical Industry Council: *Responsible Care,* CEFIC, Brussels 1993.
14. Chemical Industries Association: *Responsible Care Management Systems, Guidelines for Certification to ISO 9001,* CIA, London 1992.
15. *CEFIC Guidelines on Transfer of Technology (Safety, Health and Environment Aspects),* CEFIC, Brussels 1991.

Noise and Vibration

Wolfgang Daiminger, Müller-BBM Schalltechnisches Beratungsbüro, Planegg bei München, Federal Republic of Germany

Klaus R. Fritz, Müller-BBM Schalltechnisches Beratungsbüro, Planegg bei München, Federal Republic of Germany

Edwin Schorer, Müller-BBM Schalltechnisches Beratungsbüro, Planegg bei München, Federal Republic of Germany

Burkhard Stüber, Müller-BBM Schalltechnisches Beratungsbüro, Planegg bei München, Federal Republic of Germany

1. Introduction . . . 1
2. Definitions . . . 2
2.1. Sound and Vibration Levels . . . 2
2.2. Level Calculations . . . 3
2.3. Spectra . . . 3
3. The Impact of Noise and Vibration 4
3.1. Sound Perception and Noise Effects on Humans . . . 4
3.2. Effects of Vibration on Humans and Equipment . . . 5
4. Legal Aspects . . . 6
4.1. Occupational Noise . . . 6
4.2. Occupational Vibration . . . 6
4.3. Environmental Noise and Vibration . . . 6
5. Vibration . . . 8
5.1. Measurement . . . 8
5.2. Principal Sources . . . 8
6. Sound Emission . . . 9
6.1. Determination of Sound-Power Levels . . . 9
6.2. Flow-Induced Noise . . . 10
6.3. Major Sources of Noise . . . 11
6.3.1. Burners . . . 11
6.3.2. Fans . . . 12
6.3.3. Compressors . . . 12
6.3.4. Pumps . . . 13
6.3.5. Electric Motors . . . 14
6.3.6. Steam Turbines . . . 14
6.3.7. Gears . . . 14
6.3.8. Valves . . . 15
6.3.9. Exhaust Noise . . . 15
6.3.10. Granule-Conveyor Piping . . . 16
6.3.11. Water-Splash Noise from Cooling Towers . . . 16
6.3.12. Coal Mills . . . 16
6.3.13. Belt Conveyors . . . 17
6.4. Sound-Power Levels Associated with an Entire Plant . . . 17
7. Sound Immission . . . 18
7.1. Measurement of Sound Levels in the Vicinity of an Industrial Complex . . . 18
7.2. Outdoor Sound Propagation . . . 19
7.3. Typical Sound Levels at the Workplace and Near an Industrial Complex . . . 20
8. References . . . 20

The entire topic was coordinated by Burkhard Stüber

1. Introduction

Noise abatement is an essential consideration with respect to chemical and petrochemical plants for the protection of people living in the vicinity of such a plant as well as employees working at the facility. Reduction of workplace noise within a plant also increases the intelligibility of public announcements or warning signals, at the same time that it facilitates communication among personnel on the plant floor and in the control rooms. Moreover, extraordinary noise events are more conspicuous in a quiet environment, so measures that limit background noise contribute to improved safety conditions. Effort to counter excess vibration of plant equipment in general has the further advantage that it can prevent unnecessary damage and consequent disruption of productive activities.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b07_383

2. Definitions

2.1. Sound and Vibration Levels

Audible sound refers to mechanical vibration with components falling in the approximate frequency range 16 Hz – 16 kHz. Sound in gases and liquids is always related to vibrational motion of the particles constituting the fluid. Important *field quantities* with respect to sound are the time variation in the associated alternating pressure $p(t)$ (which is superimposed over some static pressure) and the time variation of the velocity $v(t)$ with which the affected particles move back and forth.

The root mean square (rms) value p of the alternating pressure $p(t)$ is commonly expressed in terms of a logarithmic function, the *sound pressure level* L, which is measured in decibels (dB). L is defined by the relationship

$$L = 20\log(p/p_0)\ \mathrm{dB} \tag{1}$$

The reference value p_0 (for airborne sound) has been internationally fixed at $p_0 = 2\times10^{-5}\ \mathrm{N/m^2}$, a value that corresponds to the human auditory threshold at medium frequencies. A compelling reason for selecting a logarithmic scale is the fact that the alternating pressures of interest from a technical standpoint cover a range of eight orders of magnitude.

Other field quantities used to describe sound energy or sound-energy transfer can be derived through multiplication. Thus, the *intensity* I

$$I(t) = \overline{p(t)\cdot v(t)} \tag{2}$$

specifies the power (i.e., energy per unit time) that will penetrate a standard surface perpendicular to the velocity vector. The overbar in Equation 2 indicates a time-average value. The *sound power* $W(t)$ can be derived by forming the surface integral:

$$W(t) = \oint_S \overline{p(t)\cdot v(t)}\cdot\cos\varphi\,\mathrm{d}S \tag{3}$$

where φ is the angle between the velocity vector $v(t)$ and a line normal to the surface element $\mathrm{d}S$.

The rms value W of the sound power $W(t)$ is also expressed in terms of a logarithmic measure (again with the unit decibels) known as the *sound-power level* L_W:

$$L_W = 10\log(W/W_0)\ \mathrm{dB} \tag{4}$$

For airborne sound sources, the value $W_0 = 10^{-12}$ W is normally used as a reference point.

Mechanical vibrations in solids can be characterized by three motion quantities and their variations as a function of time. These are the time variations of the displacement $s(t)$, velocity $v(t)$, and acceleration $a(t)$. The three quantities can be interconverted by integration or differentiation:

$$a(t) = \frac{\mathrm{d}v(t)}{\mathrm{d}t} \quad \text{and} \quad v(t) = \frac{\mathrm{d}s(t)}{\mathrm{d}t} \tag{5}$$

All three measurable variables cover large ranges, so – just as in the case of sound pressure – rms values are often reported rather than the quantities themselves, in logarithmic form (once again in decibels) as the corresponding levels L. These levels are defined by the relationships

$$\begin{aligned} L_s &= 20\log(s/s_0)\ \mathrm{dB} \\ L_v &= 20\log(v/v_0)\ \mathrm{dB} \\ L_a &= 20\log(a/a_0)\ \mathrm{dB} \end{aligned} \tag{6}$$

The following have been established as reference values for computing velocity and acceleration levels (according to international standards; see ISO 1683 [1]):

$$\begin{aligned} v_0 &= 10^{-9}\ \mathrm{m/s} \\ a_0 &= 10^{-6}\ \mathrm{m/s^2} \end{aligned} \tag{7}$$

However, in the field of acoustics it is usually recommended that a somewhat different set of reference values be utilized instead:

$$\begin{aligned} s_0 &= 8\times10^{-12}\ \mathrm{m} \\ v_0 &= 5\times10^{-8}\ \mathrm{m/s} \\ a_0 &= \pi\times10^{-4}\ \mathrm{m/s^2} \end{aligned} \tag{8}$$

Using these reference values, the three structure-borne sound levels L_s, L_v, and L_a all share a common numerical value at a frequency of 1000 Hz, and the average sound-pressure level directly in front of a vibrating structure corresponds to the average structure-borne sound level.

2.2. Level Calculations

Studies in the field of sound and vibration usually involve individual sound signals that are not coherent relative to one another (i.e., they do not depend on each other). The power quantities in such a case are found to be additive according to the following addition rule:

$$L_{\text{total}} = 10\log\left(\sum_{i=1}^{n} 10^{0.1\cdot L_i}\right)\ \text{dB} \tag{9}$$

where

L_{total} = total level
L_i = ith individual level
n = number of individual levels to be added

This relationship can of course also be reversed. If, for example, level L_1 is to be subtracted from level L_2, the following calculation leads to the level L_{Diff} representing the differential value:

$$L_{\text{Diff}} = 10\log\left(10^{0.1\cdot L_2} - 10^{0.1\cdot L_1}\right)\ \text{dB} \tag{10}$$

Just as in addition, any averaging of incoherent signals must also be carried out using power quantities. The following expression can be used for obtaining an energetic mean value:

$$L_{\text{m}} = 10\log\left(\frac{1}{n}\sum_{i=1}^{n} 10^{0.1\cdot L_i}\right)\ \text{dB} \tag{11}$$

where

L_{m} = energetic mean level (level of the average value),
L_i = ith individual level,
n = number of individual levels to be averaged

The above arithmetic rules apply to all the various level quantities, including the sound-power level, acceleration level, etc.

The average level of a time-dependent (variable) sound-pressure level $L(t)$, expressed over an average time T, is called the (energy) *equivalent continuous level* L_{eq}, defined according to the following expression:

$$L_{\text{eq}} = 10\log\left(\frac{1}{T}\int 10^{0.1\cdot L(t)}\,\text{d}t\right)\ \text{dB} \tag{12}$$

The significance of L_{eq} is that it permits a time-dependent noise function to be replaced by an equivalent constant noise with the same energy.

If a time-dependent variable noise consists of n time periods with individual durations T_i and corresponding sound-pressure levels L_i, where the latter are constant over the course of the stated durations, then the following expression is valid:

$$L_{\text{eq}} = 10\log\left(\frac{1}{T}\sum_{i=1}^{n} T_i\, 10^{0.1\cdot L_i}\right)\ \text{dB} \tag{13}$$

where

$$T = \sum_{i=1}^{n} T_i$$

The equivalent continuous level L_{eq} is the function most often used to assess the effect of time-dependent sound on human beings.

2.3. Spectra

In addition to considering its variation as a function of time (i.e., sound pressure vs. time), a particular vibration event can also be described in terms of its spectrum, in which case the vibrational quantity is expressed as a function of frequency f. Measuring instruments with filters – so-called fast Fourier transform (FFT) analyzers – are generally used for this purpose, since they make it possible to establish levels within discrete frequency ranges. Most important for practical applications is filtering in octave bands, where the transmission band selected has a range of one octave. The relationship between the lower and upper frequency limits for such a transmission band is 1 : 2. The octave-band central frequencies are uniformly determined; sound-pressure levels L_{oct} in octave bands are generally provided for the following central frequencies f_n:

$$f_n = 31.25 \times 2^{n-1} \text{ with } n = 1 - 9 \tag{14}$$

In addition, it is also common to use third-octave bands or narrow-band spectra with constant bandwidths.

3. The Impact of Noise and Vibration

3.1. Sound Perception and Noise Effects on Humans

Audible sound covers a frequency range of 3 decades and a sound-pressure range of ca. 7 decades; in a double-logarithmic presentation of sound pressure versus frequency (Fig. 1), the *audible range* is represented by the area between the threshold in a quiet situation ("threshold in quiet," the lowest audible sound level for a pure tone) and the threshold of pain.

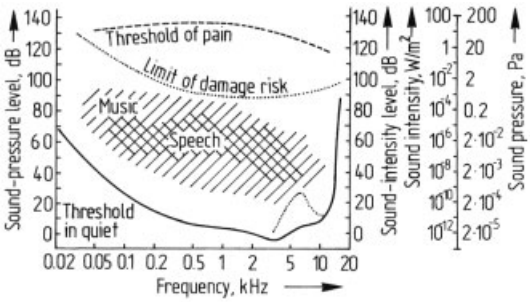

Figure 1. Hearing area; i.e., the area between the threshold in quiet and the threshold of pain; the area encompassed by music and speech and the limit of damage risk are also indicated.
The ordinate scale is expressed not only in terms of sound-pressure level but also with respect to sound intensity and sound pressure. The dotted part of the threshold in quiet relates to subjects who frequently listen to very loud music (based on ZWICKER and FASTL [2]).

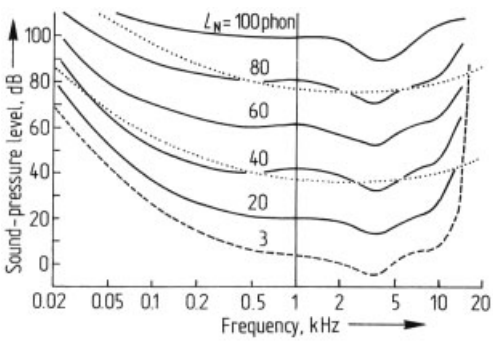

Figure 2. Equal-loudness contours for pure tones in a free sound field.
The parameter is expressed in terms of loudness level, L_N. The 3-phon curve also represents the threshold in quiet. A-Weighting curves have been provided for comparison (dotted) at two levels (based on ZWICKER and FASTL [2]).

Vibrations at very low frequencies ($f < 16$ Hz, infrasound) and very high frequencies ($f > 20$ kHz, ultrasound) cannot be perceived by the human ear. In Figure 1 the region encompassed by speech and music is hatched, and a limit of hearing-damage risk has been indicated for a noise exposure of eight hours per day.

At a constant sound level, sound with a low or very high frequency will not elicit the same loudness sensation as that in the medium-frequency range. This is illustrated by curves of equal loudness (Fig. 2). Figure 2 shows that a 100-Hz tone at L=50 dB, for example, is perceived as being of the same loudness as a 1000-Hz tone at $L = 40$ dB.

In order to simulate this hearing characteristic in the measurement of sound levels, a *weighting filter network* (*A-filter*) with defined attenuation at low and high frequencies is used to approximate a loudness-equivalent measuring result. This A-weighting curve stems from the early days of sound measurement; examples of weighting curves are provided in Figure 2.

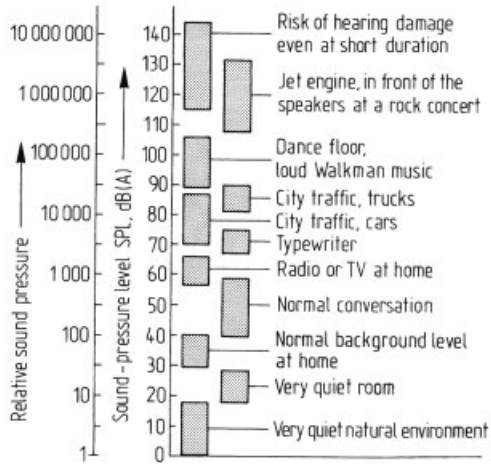

Figure 3. Typical sound-pressure levels for everyday noise sources

The A-weighting curve is constant at all levels, while the loudness sensation (i.e., the shape of the equal-loudness contours) varies with level. From Figure 2 it is apparent that an A-weighted sound-pressure level approximates the loudness sensation better in the medium-level range than at high sound levels. In spite of this, A-weighted sound-pressure levels, indicated by dB (A), are generally used for the measurement and assessment of sound events. The

A-weighting curve is internationally standardized (IEC 651 [3]). Additional weighting curves are defined for special applications (e.g., B-, C-, D-weighting). Currently accessible, more accurate measurement methods for ascertaining hearing-adjusted loudness have not yet gained general acceptance (e.g., ISO 532 [4]). Figure 3 illustrates a number of typical everyday sound-pressure levels.

Noise can impair or destroy human hearing, and it can also cause physical or mental stress (increased pulse rate and/or blood pressure, nervousness, sleep disorders). Furthermore, noise can increase the risk of an accident if warning signals go undetected because of excessive background noise. These effects are particularly important to consider in the assessment of workplace noise.

There is no evidence to indicate that noise with a sound-pressure level below 85 dB (A) can directly impair the ear. In terms of continuously present noise and its average A-weighted sound-pressure level $L_{\mathrm{A,m}}$, biological effects are now commonly accepted as follows:

disturbance of a night's rest	$L_{\mathrm{A,m}} > 40\,\mathrm{dB\,(A)}$
disturbance of concentration	$L_{\mathrm{A,m}} > 45\,\mathrm{dB\,(A)}$
disturbance of communication	$L_{\mathrm{A,m}} > 50\,\mathrm{dB\,(A)}$
increased risk of cardiovascular disease	$L_{\mathrm{A,m}} > 65\,\mathrm{dB\,(A)}$

These noise effects are relevant factors in any assessment of noise in residential neighborhoods located near industrial zones or roads.

The hearing-impairment effect of noise depends heavily on the sound-pressure level of the noise itself as well as on the corresponding exposure time. Irreparable damage can be caused even by single transient sound events (shots, explosions) with peak values exceeding ca. 140 dB (A). Long-duration noise exceeding 85 dB (A) (e.g., at the workplace) leads to short-term impairment of hearing capacity (*Temporary Threshold Shift,* TTS), from which the ear will recover during periods of silence. Based on empirical data, ISO 1999 [5] provides a mathematical model for calculating damage-probability distributions for populations exposed to noise. The calculation model uses the following input quantities:

1) The A-weighted equivalent continuous noise level over an 8-h period (noise-rating level) expressed in dB (A)
2) Exposure time in years
3) Age in years
4) Sex

A rule of thumb for calculating a maximum permissible continuous noise level at the workplace, one that will not lead to permanent hearing impairment even after years of exposure, is expressed by the following formula:

$$L_{\mathrm{A,m}} \leq 85 + 10\log\left(24/T_{\mathrm{n}}\right)\,\mathrm{dB\,(A)} \tag{15}$$

where

T_n = daily noise-eposure time in hours

According to this model, no irreparable hearing damage is to be expected at levels up to 90 dB (A) with a working time of eight hours per day.

If the ear is exposed to noise levels higher than those recommended, sensory cells (hair cells) located in the inner ear will be destroyed, leading to permanent hearing impairment (*Permanent Threshold Shift,* PTS).

Wearing ear protectors at continuous noise levels > 85 dB (A) can prevent or reduce temporary shifts of the hearing threshold, and the danger of permanent damage is also considerably reduced.

3.2. Effects of Vibration on Humans and Equipment

Vibration becomes distinctly noticeable above a certain magnitude. The threshold for vibration perception depends on the frequency. In the range 5 – 100 Hz, the perception of vibration is nearly proportional to velocity. Vibrations in this frequency range will be felt if their velocity exceeds 0.1 mm/s. Vibrations with velocities of ca. 0.4 – 1.6 mm/s are clearly recognizable, and those up to ca. 6.3 mm/s are strongly felt. Vibrations with frequencies < 5 Hz must be more intense before they will be perceived.

Increasing vibration impact results in an observable impairment of well-being and fitness; higher values endanger the health and may even lead – depending upon the exposure time – to

permanent injury. The effect produced by vibration depends on the point at which the vibrations penetrate the human body. Sitting or standing individuals may suffer alteration of the skeleton (spine) and internal organs. Vibrations passing through the hands and arms can cause peripheral circulatory problems, nervous disease, and alterations of the bones, joints, and muscles of the hand – arm system (white-finger syndrome).

The vibration of plant equipment may also lead to physical damage or work disruption if it exceeds a certain magnitude (e.g., vibrational cracks at ducts, mechanical breakdown, structural damage, etc.).

4. Legal Aspects

In all highly industrialized countries, noise is recognized as a dominating incident factor in the human environment. Impairments caused by the effects of other vibration are of minor importance, but they do occur. Measures designed to reduce noise and vibration in industry and trade are likely to prevail only if appropriate laws are enacted, because achieving the desired end entails extra costs, and it may sometimes complicate plant operation and maintenance. Regulations aimed at protecting individuals from noise and vibration are directed toward immission levels in the vicinity of industrial plants as well as immission at the workplace. It is only natural that such rules have developed differently in various countries, excluding the possibility of our providing a detailed examination in the context of this article of the many specific regulations.

4.1. Occupational Noise

Regulations directed toward noise within the plant (e.g., noise at the workplace) are essential to protect workers from hearing impairment. As indicated in Section 3.1, most countries have established 85 dB (A) as the limit for permissible noise at the workplace, and this value should not be exceeded if it can be avoided by technical means. Ear protection should be worn at any workplace with a noise-rating level ≥ 85 dB (A). Moreover, workers exposed to such high noise levels should have their hearing regularly checked by a specialist in industrial medicine.

4.2. Occupational Vibration

Rules related to the human effects of vibration at the workplace are necessary in order to protect the staff from injury and restrictions with respect to their performance and fitness, particularly at workplaces associated with plant safety (e.g., control rooms, etc.). ISO 2631 [6] defines limits for vibration effects based on the criteria "reduced comfort boundary," "fatigue-decreased proficiency boundary," and "exposure limits" for whole-body exposure as a function of frequency and daily exposure time (see Fig. 4). ISO 5349 [7] deals with criteria for assessing vibrational effects on the hand – arm system.

4.3. Environmental Noise and Vibration

Regulations with respect to noise in the general vicinity of industrial plants are concerned not with potential hearing damage but rather with protection from disturbance and noise pollution. Although the annoyance caused by industrial noise and vibration from industrial plants is relatively minor compared to road and rail-traffic noise, industrial noise sources tend to be overrepresented from the standpoint of complaints.

ISO 1996 [8], consisting of three parts, provides a set of international regulations for noise protection in residential neighborhoods located near industrial areas. The basic quantity used for describing sound immission is the A-weighted energy-equivalent continuous-level L_{Aeq}. The noise-rating level L_{Ar} that has been proposed assumes the probable need for additional charges K_1 related to tonality and K_2 for impulsive noise components as follows:

$$L_{\mathrm{Ar}} = L_{\mathrm{Aeq}} + K_1 + K_2 \,\mathrm{dB\ (A)} \tag{16}$$

ISO 1996 lists additional measurement quantities that permit a detailed description of sound immission even for fluctuating or single-sound events. National or local authorities must establish appropriate noise-limiting values that should not be exceeded by the noise-rating levels in affected neighborhoods. The magnitude of the limiting value, the extent of additional charges

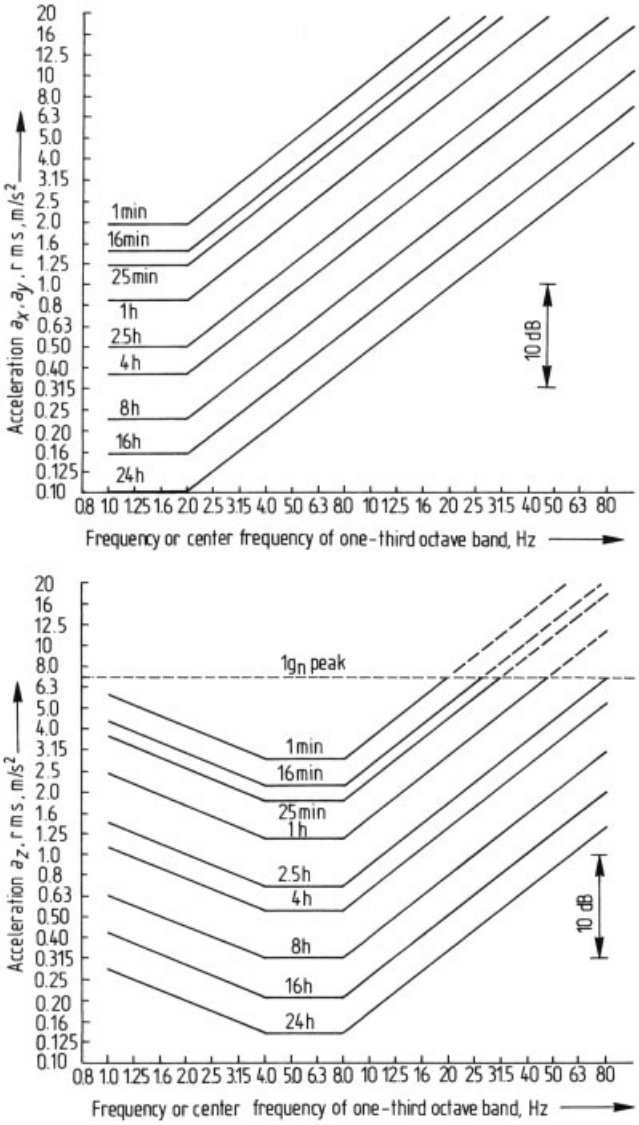

Figure 4. Transverse (A) and longitudinal (B) acceleration limits as a function of frequency and exposure time. "Fatigue-decreased proficiency boundaries" (based on ISO 2631-1 [6]); to obtain "exposure limits," multiply the acceleration values by 2 (6 dB higher); for "reduced-comfort boundaries" divide acceleration values by 3.15 (10 dB lower)

for tonality and impulsive noise, and the evaluation procedure itself vary not only from country to country but sometimes even within different states or regions of the same country. In general, nighttime noise limits are 10 – 15 dB lower than those for the daytime; some countries also define intermediate values for the early morning and evening hours.

Noise-limiting values also depend on the type of structures present in the immission area. In some countries, the lowest limit is 35 dB (A) at night for residential areas. The highest allowed value is typically 70 dB (A) for both day and night for industrial zones in which appartments are permitted only in exceptional cases. It can generally be expected that residential use without major disturbance will be possible if a value

of 45 dB (A) is not exceeded at night. Compliance with respect to the limits is usually checked by taking measurements 0.5 m in front of an open bedroom or living room window.

Like the limiting values for sound immission, limits for vibrational immission also differ from country to country. The vibrational immissions of chemical plants operating up to current technical standards are considerably below the threshold for vibration perception, so problems with vibrational immission are quite unlikely in the vicinity of such plants.

5. Vibration

5.1. Measurement

Generally speaking, all three of the motion quantities described in Section 2.1 can be used for the measurement of vibration.

The displacement of a vibrating structure has an easily measurable magnitude only at very low frequencies. For this reason measurements of displacement itself are rarely performed. It is much easier to measure the related motion quantities (velocity and acceleration) and then convert these into the quantity of interest.

Geophones are used for detecting vibrational velocity. These are moving-coil sensors with characteristic frequencies tuned to the lower end of the transmission range. Such characteristic frequencies often fall within the relevant measuring range, so appropriate correction networks are required. Because of the way geophones are constructed, their industrial range of application is rather narrow.

Accelerometers are piezoelectric transducers with characteristic frequencies tuned to the upper end of their transmission range. These are universally applicable for measuring vibrations, and many designs and models are available for use in a wide variety of situations. Accelerometers are almost always utilized for modern industrial vibration measurements.

Measurement techniques are prescribed in such individual regulations as ISO 2631 [6] (effects on man) and ISO 2372 [9] (machine vibrations). Modern digital instruments are used for registering and evaluating the signals, a process that may include frequency analysis, use of weighting filters, etc. Continuously registering instrumentation is recommended for monitoring vibrations from critical equipment in which standstill might lead to expensive downtime or safety-relevant incidents.

5.2. Principal Sources

Vibration is produced by all types of machinery involving rotating or reciprocating masses, including pumps, compressors, turbines, motors, centrifuges, and mills.

ISO 2372 [9] provides severity criteria for assessing machine vibration as a function of the nature and size of the device. Equipment with strong vibrational emissions must be isolated from areas subject to protection by specially installed foundations or resilient mountings. This is especially true for machines associated with extensive vibration, such as rocking conveyors, oscillating grates, etc. Figure 5 illustrates severity criteria for various types of machine vibration.

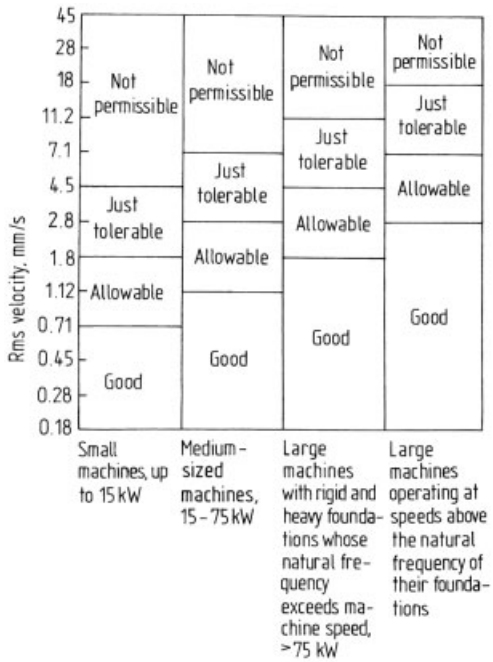

Figure 5. Vibration-severity criteria for the range 10 – 1000 Hz (based on ISO 2372 [9])

Pipelines may be subject to high vibration levels due to pulsating flow (e.g., piping associated with piston or screw compressors). Such

vibration may even be amplified by system resonance caused by an unfavorable arrangement of pipe supports. Tear-offs and, in the case of continuous stress, fatigue fracture of the pipes themselves can be the result. Furthermore, vibrating pipes may cause vibrations in other parts of a plant. Figure 6 illustrates criteria applicable to the evaluation of piping vibration.

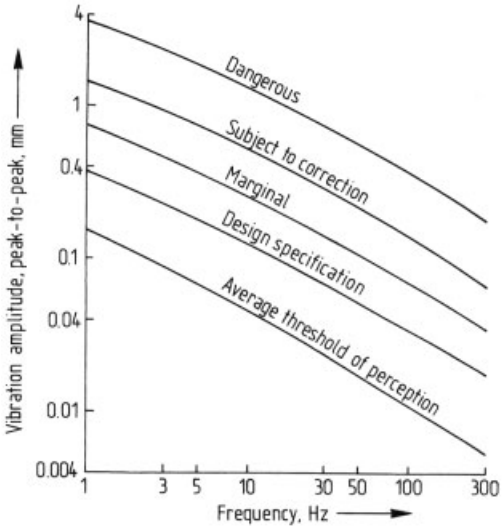

Figure 6. Criteria for piping vibration levels

6. Sound Emission

The sound emission from a sound source can be described fundamentally in terms of a corresponding sound-pressure level L. Since this measure depends upon distance from the sound source, some indication must also be provided of the measuring distance. However, it is always preferable to work with a single numerical value for the description of a sound emission, namely the sound-power level L_W (see Section 2.1), which is generally supplemented with a spectrum (e.g., in octave bands).

6.1. Determination of Sound-Power Levels

There are various methods available for laboratory or field measurements intended to establish sound-power levels. The most important procedures applicable to industrial plants are described below.

Measurement in Hemianechoic Space. The hemianechoic space method can be applied to sound sources or sound-source groups characterized by free sound propagation. The corresponding temporally and locally averaged sound-pressure level L_{m} is measured on a surface with a total area S that encompasses the sound source and terminates at a set of reflecting boundary surfaces (e.g., the floor). The measuring surface is in most cases a cuboid with surfaces located (preferably) 1 m from the contour of the sound source. Based on the assumption that sound waves pass perpendicularly through the measuring surface, the sound-power level L_W is then:

$$L_W = L_{\mathrm{m}} + 10 \log\,(S/S_0)\ \mathrm{dB} \tag{17}$$

where $S_0 = 1\ \mathrm{m}^2$

For more details (exact position of the measuring points and correction for surroundings, noise level, etc.), ISO 3744 [10] should be consulted.

Method Applicable to Multi-Source Industrial Plants. This method, described in ISO/DIS 8297 [11], can be applied to large industrial plants with numerous sound sources provided the major extension of the plant (up to ca. 300 m) is horizontal. Sound-pressure levels are determined on a single measuring path that encloses the entire plant. The required height of the microphone depends on the size of the plant and the distance between the plant contour and the measuring path; a typical value is 5 – 15 m. A unidirectional microphone can be used to reduce the influence of background noise (for more details see ISO/DIS 8297 [11]). This measuring approach is easy to apply (measuring time ca. 4 h for a plant occupying 40 000 m^2), and it establishes both a sound-power level and an octave-band spectrum for all noise radiated by the plant.

Vibration Measurement. The airborne noise power radiated by a vibrating surface can be determined by measuring the corresponding vibration. This technique is most often applied when a direct measurement of airborne sound is impractical because of excessive background

noise. A vibration-measuring instrument is used to establish a temporally and locally determined velocity level L_{vm} at the surface of the radiating structure (surface S). The radiation behavior of a structure can be described either by the radiation efficiency σ or the logarithmic measure $10 \log \sigma$. The radiation efficiency depends on the ratio of the radiated airborne wavelength to the size of the radiator, the geometric form of the radiator, and the phase distribution of structure-borne sound over the radiator surface. The sound-power level L_W is obtained from the quantities L_{vm}, S, and σ as follows:

$$L_W = L_{\mathrm{vm}} + 10 \log\,(S/S_0) + 10 \log \sigma \,\mathrm{dB} \tag{18}$$

where $S_0 = 1\ \mathrm{m}^2$

For more details see ISO/TR 7849 [12]. This procedure is generally not as accurate as the direct measurement of airborne sound because of an often insufficient knowledge of the radiation efficiency. A value of $\sigma = 1$ provides an upper limit for the sound-power level.

Measurement of Sound Intensity. In contrast to indirect methods for the determination of sound intensity and sound power by sound-pressure measurements, sound intensity [see Eq. (2)] can also be measured directly with the aid of a two-microphone method and digital two-channel FFT analyzers for data processing. Details regarding the measurement of sound power by the intensity method are available in ISO 9614 [13]. Part 1 of this standard describes the method for obtaining data at discrete points on the measuring surface. Part 2 (in preparation) describes a scanning method that is often easier to apply. With the scanning method, a measuring probe is led continuously along a set of adjacent measuring paths that cover the entire measuring surface. Intensity measurements are seldom applicable to industrial plants.

6.2. Flow-Induced Noise

In refineries and chemical or petrochemical plants much of the noise generated is due to the flow of gases or liquids. For this reason it is appropriate at this point to provide a brief introduction to sound generation by flow mechanisms.

Part of the flow energy in any flow field is converted into sound energy. At subsonic flow speeds, flow noise is generated by sources of three types: monopole, dipole, and quadrupole.

A *monopole* source exists if an unsteady volume flow is responsible for the noise generation. Typical examples result from periodic afflux of fluid into a pipe or collapsing cavitation bubbles. Periodic afflux into a pipe is responsible for the low-frequency noise emitted by piston and screw compressors. If the local pressure in a flowing liquid falls below a certain critical value (approximately equal to the vapor pressure of the liquid), then the presence of appropriate nuclei can cause gas- and vapor-filled bubbles or cavities to develop and grow. This phenomenon, which is typical of all liquids, is called (hydrodynamic) *cavitation*. Such bubbles can collapse violently if the external pressure increases above the critical pressure, giving rise to very high local pressure peaks (sometimes exceeding 10^5 bar with strong cavitation, sufficient to cause material damage) and also to high noise levels. Cavitation usually generates a rattling, broadband type of noise. Noise generation by cavitation is important with respect to valves and pumps.

In the case of a *dipole* source the volume flow is locally constant, but alternating forces are present as well, such as at the surface of a solid object encountering a stream generated by vortex shedding or interacting with a turbulent incoming flow or a turbulent efflux. This type of noise generation is particularly important in the case of fans and valves.

If the alternating forces also tend to compensate each other, then the corresponding noise sources are of the *quadrupole* type, as observed with free turbulence. Such quadrupole sources are sometimes important in conjunction with valves or vents to the open air.

Theoretical considerations have shown that with respect to the sound power W of the three different types of elementary sources, velocity dependencies can be anticipated as listed below provided the frequencies typical for the noise behavior and the characteristic velocity U (e.g., the flow velocity) are proportional to each other and the sources are also "compact."

Source	*Sound power*
Volume flow (monopole)	$W \sim \varrho c^3 M^{2\ldots4}$
Alternating force (dipole)	$W \sim \varrho c^3 M^{4\ldots6}$
Free turbulence (quadrupole)	$W \sim \varrho c^3 M^{6\ldots8}$

where

ϱ = density of the flow medium
c = sound velocity in the flow medium
U = characteristic velocity
M = U/c (the Mach number)

This table shows that in a sound field composed of different types of sources that are active simultaneously, the sources of higher order become more conspicuous with increasing Mach number compared with those of lower order. This is the reason why quadrupole sources are of no significance in liquids, where the Mach number is small.

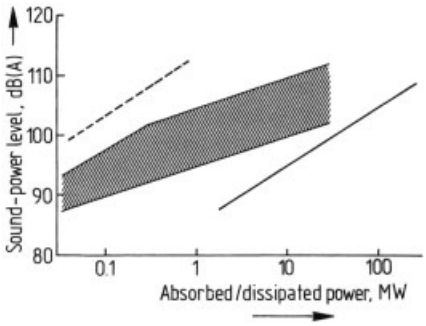

Figure 7. A-weighted sound-power levels of important noise sources in chemical and petrochemical plants based on state-of-the-art noise abatement; solid line = process furnaces, steam boilers; broken line = induced-draft cooling towers; hatched area = fans, compressors, pumps, steam turbines, gears, valves, coal mills

6.3. Major Sources of Noise

The main individual sources of noise in chemical and petrochemical plants are:

- Burners (process furnaces, steam boilers, and flares)
- Fans
- Compressors
- Pumps
- Electric motors
- Steam turbines
- Gears
- Valves
- Exhausts to the open air
- Granule conveyors and silos
- Airborne splash noise from cooling towers
- Coal mills
- Belt conveyors

Reference values for the A-weighted sound-power levels of noise generated and radiated to the outside by these sound sources as a function of installed power are illustrated in Figure 7 based on state-of-the-art noise-control measures.

6.3.1. Burners (→Combustion)

The noise-radiation level of a process furnace, steam boiler, or flare is determined by the nature of the burner. Burner noise is generated by turbulent combustion of gas and oil, turbulent efflux of gaseous fuel through the burner nozzles, and the efflux of steam added to improve combustion. Important quantities associated with combustion noise include heat release and the extent of turbulence in the flame.

In the case of *furnaces* and *boilers*, the burner nozzles are located inside a room. The resulting noise is of a broadband nature with peaks attributable to characteristic frequencies of the combustion room and the gas or air supply of the burner. Noise radiation takes place via air intakes from the surroundings (natural-draft burners), air flow into the air ducts and then into the open air (forced-draft burners), the furnace walls, and flue-gas ducts into the stack and then into the open air. Stack walls themselves are generally of minor importance.

The following noise-control measures are used to reduce burner-noise radiation:

1) Installation of silencers in the air inlets of natural-draft burners
2) Acoustically optimized furnace-wall construction
3) Acoustical treatment of the air ducts in forced-draft burners

Newly designed furnaces and boilers outfitted with currently available noise-control facilities are capable of meeting A-weighted burner-noise sound-power levels of

$$L_{WA} = 83 + 10 \log N \, \mathrm{dB\ (A)} \tag{19}$$

where N = heat release in MW (0.1 . . . 10 MW)

Flares are used for burning off small masses of gas resulting from leakage and changes in operating conditions, burn-off during controlled start-up and shut-down procedures, and burn-off occasioned by emergency shut-down of a process plant. Only elevated flares can be used for the emergency elimination of large masses of gas. Elevated flares are designed for gas-mass flows up to ca. 500 t/h. Ground flares are often used for smaller mass flows because combustion is more satisfactory with a ground flare than with an elevated flare. This applies not only to environmental stress occasioned by unwanted light radiation but also with respect to the emission of residual hydrocarbons in the flue gas. Furthermore, a ground flare is less noisy with the same gas-mass flow relative to the corresponding elevated flare [by about 10 – 20 dB (A)]. Ground flares are designed for gas masses up to ca. 70 t/h.

With a ground flare, noise is radiated via the air-intake openings and the flare mouth. It is customary to install silencers at the air-inlet side, which can reduce the A-weighted sound-power level to

$$L_{WA} = 102 + 15 \log q \, \mathrm{dB\ (A)} \tag{20}$$

where q = gas-mass flow in t/h (3 . . . 70 t/h)

6.3.2. Fans

Process plants most often utilize centrifugal and axial fans. The noise from a fan consists of a broadband component accompanied by specific tones. Noise sources of the dipole type are responsible for the broadband noise; these are caused by vortex shedding from the moving blades and turbulent flow acting upon solid construction parts. The specific tonal components are produced mainly by forces resulting from interactions between the moving blades and guide vanes in the case of axial fans and between blades and cut-off with centrifugal fans, both of which lead to periodically fluctuating streams flowing toward solid components.

The noise is radiated primarily via free openings on the suction or pressure sides of the fan, by ducts connected to the fan, and by the fan casing.

The following secondary noise-control measures are useful for reducing this radiated noise:

1) Installation of silencers at the suction and pressure sides of the fan, particularly for fans with free openings on either the suction or the pressure side
2) Acoustical enclosures, or wrapping of the fan casing
3) Acoustical wrapping of connected ducts

Such noise-control measures permit the achievement of an A-weighted sound-power level of

$$L_{WA} = 75 + 10 \log N_a \, \mathrm{dB\ (A)} \tag{21}$$

where N_a = rated electric power in kW (1 . . . 3000 kW)

Air coolers and induced-draft cooling towers are equipped with single-stage axial fans. In the case of low-noise fans (broad blades, large diameter, tip speed < 35 m/s), the corresponding fan noise can be brought down to an A-weighted sound-power level of

$$L_{WA} = 78 + 10 \log N_a \, \mathrm{dB\ (A)} \tag{22}$$

where N_a = rated electric power in kW (0.1 . . . 500 kW)

6.3.3. Compressors (→ High-Pressure Technology, Chap. 3.4.; → Refrigeration Technology, Chap. 1.1.2.)

Noise radiation from a compressor depends on the type of construction, the power, and the rate of revolution. The most common compressor types in chemical and petrochemical plants are piston compressors, turbocompressors (axial and centrifugal compressors), and Roots and screw compressors.

Piston Compressors. A piston compressor conveys a fixed mass of gas with each piston stroke. This periodic conveyance is in turn related to periodic pressure variations on the suction and pressure sides of the compressor. The noise spectrum that results is responsible for the noise level in the low-frequency range. High-frequency noise is produced by relief within the pressure valves, where flow rates often reach sonic speeds.

Sound radiation of the noise takes place via the compressor casing and the connected suction and pressure pipes, including pulsation dampers.

Turbocompressors. In axial-flow compressors, compression of the gas is achieved by rotating blade rings and rows of stationary or adjustable guide vanes. The corresponding noise spectrum consists of specific tones superimposed over broadband noise. The broadband noise is mainly caused by (a) alternating forces on the rotating blades due to vortex shedding at the back edge and by turbulent flow down stream from struts and intake nozzles, as well as (b) alternating forces on guide blades in the turbulent flow down stream from the rotor blades. Individual tones in the spectrum result mainly from periodic alternating forces due to interaction of the rotor blades with the guide vanes.

The noise generated by such a compressor is radiated via the connected suction and pressure piping (major portion), the compressor casing, and interstage coolers and separators.

Roots and Screw Compressors. Compressor noise in this case is due mainly to the periodic conveyance of gas and the flow and exhaust noise accompanying pressure-compensation phenomena at the pressure outlet.

Noise generated by such a compressor is radiated via the connected piping (major portion), the compressor casing, and the inlet (for air compressors). A sound-power spectrum for noise introduced into the connected piping can be determined quite accurately provided a time-dependent volume flow can be derived from the compressor's construction data.

Important secondary noise-control measures applicable to compressors include installation of pulsation absorbers and silencers in the suction and pressure pipes as well as acoustical wrapping of the connected pipes. These noise-control measures can result in the following A-weighted sound-power levels in the case of piston and turbocompressor noise:

Piston compressor

$$L_{WA} = 86 + 5 \log N \, \mathrm{dB\ (A)} \tag{23}$$

where N = required power in kw (50 . . . 1000 kW)

Axial turbocompressor

$$L_{WA} = 105 + 5 \log N \, \mathrm{dB\ (A)} \tag{24}$$

where N = required power in MW (0.3 . . . 20 MW)

If stringent acoustical requirements are in force, installation of a compressor within a building requires provisions for a suitable acoustical enclosure.

With Roots and screw compressors, auxiliary sound-absorbing enclosures are common, leading to an A-weighted sound-power level of

$$L_{WA} = 72 + 10 \log N \, \mathrm{dB\ (A)} \tag{25}$$

where N = required power in kW (10 . . . 300 kW)

6.3.4. Pumps (→ High-Pressure Technology, Chap. 3.3.)

Pumps can be divided into two main categories: hydrostatic and hydrodynamic. *Hydrostatic* pumps are used for the generation of very high pressures. One characteristic of these pumps is a mechanical separation between the suction and pressure sides. Furthermore, depending on the extent of this separation, the discharge pressure is almost independent of the counterpressure or lift of the pump. The most important types of hydrostatic pumps are the vane cell pump, the gear pump, the axial piston pump, and the spindle pump.

In contrast to hydrostatic pumps, *hydrodynamic* pumps (rotary pumps) provide no mechanical separation between the inlet and discharge ports. The pressure difference between the two ports must be maintained by hydrodynamic forces. Rotor blades introduce kinetic energy into the axially flowing liquid, energy that is transformed mainly into pressure energy. Fluid conveyance is continuous.

The major cause of noise radiation from pumps is alternating pressure within the pumped liquid. With respect to hydrostatic pumps, three different mechanisms can be responsible for noise generation:

1) Periodic alternation of the discharge
2) Impulse-like pressure-compensation processes at the point of intersection of liquid volumes that differ in pressure
3) Cavitation and gas exhaust during the suction process

Hydrodynamic pumps are associated with noise-generation mechanisms similar to those of fans. Furthermore, cavitation may arise under certain operating conditions in conjunction with the suction process at the rotating blades.

The major noise radiation from a pump occurs via connected pipes and the pump casing. An important secondary noise-control measure is acoustical wrapping of the pipes, leading to an A-weighted sound-power level of

$$L_{WA} = 70 + 10 \log N \, \mathrm{dB\ (A)} \tag{26}$$

where N = required power in kW (1 . . . 2000 kW)

If the acoustical requirements are rather strict, pumps must be installed within sound-absorbing enclosures or in separate buildings.

6.3.5. Electric Motors

Three-phase line-voltage electric motors are commonly employed for rated powers up to ca. 400 kW. Three-phase high-voltage electric motors (6000 V) are used for rated powers > 160 kW. The noise generated by electric motors is due mainly to associated blowers (in the case of air-cooled motors), magnetostriction, and the motor bearings. The extent of noise radiation from a modern electric motor depends on the motor's power and rate of revolution as well as the extent of built-in noise-control measures. In the case of 3000-rpm motors, level differences > 10 dB (A) may distinguish "normal" from "low-noise" types. Enclosures are necessary for large motors when the acoustical requirements are stringent.

For air-cooled line-voltage motors the following A-weighted sound-power level can be achieved assuming the use of low-noise blowers:

$$L_{WA} = 75 + 7 \log N \, \mathrm{dB\ (A)} \tag{27}$$

where N = nominal power in kW (3 . . . 400 kW)

For high-voltage motors with integrated noise control or a suitable enclosure, an appropriate target for the A-weighted sound-power level is

$$L_{WA} = 70 + 7 \log N \, \mathrm{dB\ (A)} \tag{28}$$

where N = nominal power in kW (160 . . . 3000 kW).

6.3.6. Steam Turbines

The noise generated by a steam turbine is due mainly to flow noise from within the turbine itself. This noise is probably caused by the same mechanisms operative in turbocompressors. Condensing turbines generate additional noise in the condensers.

Noise generated by a steam turbine is radiated via the casing and the steam-feed pipe as well as the steam-exhaust pipe and any condensers that may be present. The following secondary noise-control measures are effective for reducing noise radiation:

1) A sound-absorbing enclosure for the turbine casing
2) Combined heat and noise wrapping of the steam-feed pipe
3) Wrapping of the steam-exhaust pipe (and condensers) or installation of a silencer in the steam-exhaust pipe adjacent to the steam turbine

These measures can lead to an A-weighted turbine sound-power level of

$$L_{WA} = 95 + 7 \log N \, \mathrm{dB\ (A)} \tag{29}$$

where N = rated power in MW (0.3 . . . 20 MW)

6.3.7. Gears

Many investigations of gear noise and its origin have been described in the literature. The evidence suggests that there is a close relationship between dynamic load on the gear teeth (and therefore the service life) and resulting noise. In some cases it is possible to derive information regarding gear quality and gear defects from the nature of the radiated gear noise.

Geometric deviations and elastic deformation of the teeth are major sources of gear-noise generation. Uneven spots on the surface of the teeth arise during manufacture, and these cannot be avoided. As the teeth engage, the uneven spots are "played" like a phonograph record. Although the shape deviations are very small (ca. 10 µm), they can cause significant noise, which is generally of a broadband nature.

Flank deformations of the teeth are often due to defects in the driving gear of the tooth-cutting machine, and therefore display spatial periodicity related to tooth frequency and the rate of revolution of the gear-cutting machine. Flank defects generally cause frequency maxima that are related not to the meshing frequency of the gear in question but rather to the gear-cutting machine used for its manufacture.

A special case of tooth defect is *pitch defect*, caused by unequal spacing between the teeth. The noise produced by pitch defect displays peak levels at the harmonics of the rate of revolution. Under loaded condition, the teeth of a gear wheel are deformed slightly as a result of their elastic properties. This is an important factor with respect to noise generation, because such deformation changes as a function of time.

The most important factors associated with the generation of gear noise are geometrical deviations, alterations in tooth stiffness, specific tooth power, circumferential speed, alignment of the shafts, shaft bearings, and stiffness of the casing. These parameters interact in a complicated way that depends to some extent on design data (tooth geometry, wheel diameter, tooth width, etc.).

In order to reduce sound radiation, gears are often provided with acoustical wrappings and mounted resiliently. These measures permit the following A-weighted sound-power level to be met:

$$L_{WA} = 74 + 7 \log N \, \mathrm{dB\ (A)} \tag{30}$$

where N = rated power in kW (1 . . . 20 000 kW)

6.3.8. Valves (→ High-Pressure Technology, Chap. 4.)

Gas flow through control valves generates noise mainly as a result of the flow field down-stream from the narrowest cross section, especially via vortex shedding and turbulent-flow impact on solid construction elements (with alternating intensity, thereby constituting a dipole source), free radiation in the mixing zone (a quadrupole source), and compression shock when sonic speeds are reached or exceeded within the valve.

Noise control is generally unnecessary with valves for liquids so long as there is no cavitation or evaporation (flashing).

Noise generated by control valves is radiated mainly via the connected piping, not the valve casing itself. Noise-reduction features have been successfully incorporated into gas control valves in recent years, but process plants still typically require extensive additional secondary valve-noise control measures. These include sound-absorptive wrapping of the pipes (and sometimes the valve casing) and/or the installation of silencers between a valve and the connected pipes. These noise- control measures can lead to an A-weighted sound-power level for control-valve noise of

$$L_{WA} = 70 + 10 \log N_a \, \mathrm{dB\ (A)} \tag{31}$$

where N_a = power dissipated in the valve in kW (0.1 . . . 3000 kW)

6.3.9. Exhaust Noise

Chemical and petrochemical plants contain numerous valves for venting gaseous materials to the atmosphere. The corresponding noise is often characterized by very high sound power. Many of these are emergency valves engaged only in the case of an unforeseen incident, but during start-up or shut-down of a plant it often becomes necessary to blow out superfluous steam or air via so-called start-up control valves. Pressure relief and noise generation with respect to such installations is confined mainly to the valve and the vent opening.

The major sound-transmission paths are blow-out openings, pipes, and valve casings. Important noise-control measures include installation of valves with low-noise design characteristics, acoustical wrapping of valve casings and connected piping, installation of silencers (which generally consist of separate pressure-relief and absorption units) at the blow-out openings, as well as addition of sound-absorbing cladding to the silencer casing. These measures make it possible to achieve an A-weighted sound-power level of

$$L_{WA} = 95 + 10 \log q \, \mathrm{dB\ (A)} \tag{32}$$

where q = mass flow in t/h

6.3.10. Granule-Conveyor Piping (→ Solids Handling, Chap. 6.)

Pneumatic conveyance is widely used for transporting plastic granules between polymerization plants and silo facilities. The extent of sound radiation from a pneumatic conveyor line depends on the piping parameters (diameter, wall thickness, material), transport parameters (product throughput, transport speed), and parameters associated with the material conveyed (density, granule size, hardness).

Noise is produced by impact between the conveyed material and the pipe walls. Maximum excitation of the pipe occurs at bends because of the large impulse interchange that accompanies a turn.

The frequency spectrum of noise produced by granule conveyance is broadband in nature, consisting mainly of high-frequency components. Common noise-control measures include use of special steels instead of aluminum for the pipes (especially at bends), together with acoustical wrapping of both pipes and bends. If these measures are taken, the sound-power level of a granule-conveyor pipe can be reduced to less than

$$L_{WA} = 75 + 10 \log l \, \mathrm{dB\ (A)} \tag{33}$$

where l = pipe length in m

6.3.11. Water-Splash Noise from Cooling Towers

Apart from air coolers, most cooling towers in process plants are of a modular, rectangular design for the purpose of water recooling. The towers generally operate according to the principle of countercurrent or crosscurrent air flow, and are equipped with axial fans (see Section 6.3.2). Cooling of descending water takes place by evaporation and convective heat transfer (→ Cooling Tower Technology).

Fan drive is accomplished with an electric motor connected via a gear box. The fan hub itself is flanged directly to the output shaft of the gear box. In addition to noise produced by the fans, gears, and drive motors, much of the emitted noise is caused by water splash in the basin. Radiation of these water noises occurs via air intakes and the diffuser opening. A reasonable estimate of the A-weighted sound-power level of water noise is provided by the expression

$$L_{WA} = 69 + 10 \log q \, \mathrm{dB\ (A)} \tag{34}$$

where q = water throughput in m^3/h

In general, secondary noise-control measures are not necessary with the cooling towers found in chemical or petrochemical plants.

6.3.12. Coal Mills (→ Size Reduction)

Coal mills are an important source of sound in coal processing plants (power stations and coal gasification or liquification facilities). The following types of mill are used for crushing and pulverizing coal, depending on the quality of the crude coal and the technical requirements: ball triturators; bowl mills, cylinder crushers, and roller mills; and impact rotor mills or air-jet impact pulverizers.

The milling itself occurs according to the following principles:

In a ball triturator, steel balls are lifted by friction in a rotary drum at the interior steel cladding, causing coal to be ground by the impact of backfalling balls [A-weighted sound-power level: 114 . . . 129 dB (A)].

In bowl, cylinder, and roller mills the coal is crushed between rolling bodies and an appropriate grinding bowl [A-weighted sound-power level: 105 . . . 118 dB (A)].

In principle, an impact rotor mill resembles a centrifugal fan with blades designed to act as baffle plates. Crushing of the coal occurs as a result of impact with the rotating baffle plates and the armored casing [A-weighted sound-power level: 105 . . . 111 dB (A)].

Noise radiated by coal mills is derived from the following sound sources:

1) Impact noise produced as grinding balls strike the interior armoring of the milling drums (ball triturators)
2) Impact noise from coal striking a rotating impact rotor and the armored casing (impact rotor mills and air-jet impact pulverizers)
3) Crushing and rolling noise (bowl mills, cylinder mills, and roller mills)
4) Impact noise produced at the crude coal inlet
5) Driving motors and gears

6) Fan noise (impact rotor mills)

In order to reduce noise transmission, coal mills are installed in their own separate buildings. Provided careful attention has been paid to acoustical planning in the building concept, it is possible to achieve a coal-mill A-weighted sound-power level of

$$L_{WA} = 85 + 5\log N \,\mathrm{dB\,(A)} \tag{35}$$

where N = driving power in kW (10 . . . 2000 kW)

6.3.13. Belt Conveyors (→ Solids Handling, Chap. 4.)

Belt conveyors are usually used for transporting coal and slag in coal-processing plants. Such a conveyor consists of an endless belt that passes over two drums, the driving pulley and the tail pulley. Between the two pulleys the belt is supported by support rollers. One or more driving stations may be located at the head of the conveyor.

In defining the sound radiation due to a belt conveyor it is necessary to differentiate between noise from the driving station and noise due to motion of the band. The noise produced by the driving station is a function of the corresponding gears and driving motors. Sound radiation from motion of the band is due mainly to rolling noise generated by belt contact with the support rollers, together with bearing noise from the support rollers themselves. The belt speed is an important parameter with respect to overall noise level.

With a belt conveyor system an A-weighted sound-power level of

$$L_{WA} = 84 + 5\log q \,\mathrm{dB\,(A)} \tag{36}$$

where q = output in t/h
can be achieved if the following measures are taken:

1) Limiting the band speed to a maximum of 3 m/s
2) Use of low-noise bearings for the support rollers
3) Regular maintenance of the support rollers (i.e., replacement of defective bearings and support rollers)
4) Provision of an acoustical enclosure for the driving station

6.4. Sound-Power Levels Associated with an Entire Plant

Large-scale industrial plants typically exhibit the following noise characteristics:

1) Many individual sound sources spread over a large surface area
2) Relatively constant noise emission as a result of 24-h operation
3) Constant broadband background noise that provides little specific information since it results from the interaction of many discrete sound sources

For large-scale industrial plants, especially refineries and petrochemical complexes, the area-specific A-weighted sound-power level (also called the specific sound-power level) is an important parameter for estimating community noise levels, one that is being increasingly accepted, particularly with respect to land-utilization and building codes.

In order to avoid misunderstanding it is essential that uniform definitions be adopted. We therefore suggest the following definitions of specific sound-power levels $L_{W''A}$:

For a single plant within a larger complex:

$$L_{W''A} = L_{WA} - 10\log\left(\frac{S_a}{S_0}\right) \tag{37}$$

where

L_{WA} = A-weighted sound-power level of the plant in question
S_a = "Plant area"; i.e., the area occupied by an individual process plant or an auxiliary plant, including access areas within the plant but not the surrounding areas (access areas include open areas, roads, and main pipelines within the plant)
S_0 = 1 m^2

For an overall industrial complex:

$$L_{W''A} = 10\log\left(\sum_{\nu=1}^{n} 10^{0.1\cdot L_{WA\nu}}\right) - 10\log\left(\frac{S_C}{S_0}\right) \tag{38}$$

where

n = number of individual plants within the factory

$L_{\mathrm{A}v}$ = fraction of A-weighted sound-power level attributable to the vth individual plant within the complex

S_{C} = "Complex area"; i.e., the total area occupied by the complex, including storage tanks, administrative buildings, traffic areas, and car parks. Large open areas (fallow land) within the limits of the complex are not included.

On average, the following area ratios are typical:

$$\sum S_{\mathrm{a}}/S_{\mathrm{C}} \approx 0.05 \text{ for refineries} \approx 0.15 \text{ for petrochemical complexes} \quad (39)$$

For individual plants, production facilities, and auxiliary plants subject to currently available noise-abatement measures, specific sound-power levels fall within the following limits:

$$68 \leq L_{W''\mathrm{A}} \leq 82\,\mathrm{dB\ (A)} \quad (40)$$

The dispersion reflected in the specific A-weighted sound-power levels $L_{W''\mathrm{A}}$ for individual plants is relatively large. Therefore, acoustic predictions applicable to individual plants in relation to their area-specific sound-power levels are possible only for plants with similar operational conditions (including throughput, installed power, etc.).

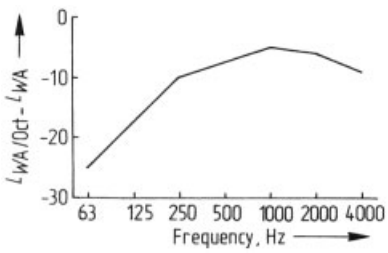

Figure 8. Typical A-weighted sound-power spectrum in octave bands for refineries as well as chemical and petrochemical complexes
$L_{W\mathrm{A}}$ = A-weighted overall sound-power level; $L_{W\mathrm{A/oct}}$ = A-weighted sound-power level in octave bands

In the course of developing zoning and land-utilization plans, and for the determination of minimum distances applicable to refineries and chemical or petrochemical complexes, an A-weighted sound-power level ($L_{W''\mathrm{A}}$) of 62 dB (A) represents a reasonable estimate for assignment to the overall premises.

Figure 8 provides a standardized, typical octave sound-power spectrum of 0 dB (A) for refineries and chemical or petrochemical complexes.

7. Sound Immission

The sound immission associated with an industrial plant in its surroundings depends not only on sound emission within the plant and the distance from the point of immission but also on various topographical factors and, especially at distances of more than 200 m, factors related to the weather. Direct measurement of the noise-immission level due to an individual plant is often impossible as a result of inevitable background noise; in this case, immission calculations must be based on emission data collected within the plant itself.

7.1. Measurement of Sound Levels in the Vicinity of an Industrial Complex

The sound-pressure levels generated by a distant sound source vary considerably; e.g., by more than 20 dB (A) at a distance of 1000 m. This is due largely to changeable weather conditions, which are in turn associated with highly variable sound propagation. The highest sound-power levels are generally recorded when the wind is blowing directly from the plant to the receiver (i.e., under downwind conditions). The dispersion of measured values under downwind conditions is relatively small, so an *average* sound-pressure level under downwind conditions is the appropriate quantity for describing noise immission due to a plant. A downwind situation is defined as one in which the wind direction deviates no more than 60° from the line joining the sound source and the point of measurement.

Noise immission in the vicinity of a continuously operating industrial plant is generally measured at night because of the correspondingly lower traffic background noise levels. In general, all measurements should be conducted under downwind conditions, and an energetic mean value should be used based on at least three measurements. Addition of appropriate charges for

tonality and impulsive noise components leads to a corresponding noise-rating level (see Section 4.3). Such an assessment of noise immission should normally be based on the weather situation that actually prevails in the vicinity of the plant. If a downwind situation is to be expected at least 25 % of the year, then this weather situation, which supports sound propagation, is the one that must be taken as prevailing. If a downwind situation is relatively rare, a reduction of up to 5 dB (A) is permissible in establishing the long-term noise-rating level as a function of wind direction and speed, the frequency of temperature inversions, and the distance between the plant and the immission point. It is important to verify, however, whether relevant governmental regulations include provisions for such long-term assessments.

7.2. Outdoor Sound Propagation

Sound immission to be anticipated at a greater distance from the sound source can be determined through a sound-propagation calculation based on sound emission at the source. This calculation method is useful for an acoustic prognosis with respect to a plant still in the planning stage, as well as for an existing facility if background noise at the immission point prevents direct measurements.

ISO/DIS 9613-2 [14] suggests an international calculation approach that could be described as a semiempirical method. The complex physical phenomenon of sound propagation is divided into a set of individual steps described in empirical ways, based partly on physical considerations. Real sound sources are represented by point sound sources characterized by sound-power levels L_W. The sound-pressure level L at the immission point of interest under downwind conditions is then calculated according to the following equation:

$$L = L_W - A \tag{41}$$

where A is the sum of a series of attenuation terms, the most important of which are described below.

1) $A_{\mathbf{div}}$ *: Attenuation due to geometrical divergence.*

Sound from a point sound source propagates within a hemispherical space. This leads to a geometrically based level reduction:

$$A_{\mathrm{div}} = 10\log\left[4\pi\times\left(\frac{d}{d_0}\right)^2\right] = 20\log\left(\frac{d}{d_0}\right) + 11\,\mathrm{dB} \tag{42}$$

where d = the distance (sound source – immission point) in m, and $d_0 = 1$ m.
This means that a level reduction of 6 dB can be anticipated with a doubling of the distance.

2) $A_{\mathbf{atm}}$*: Attenuation due to atmospheric absorption.*
During airborne sound propagation, part of the sound energy is transformed into heat. This airborne absorption depends on the ambient temperature and the relative humidity of the air as well as on the wavelength (or frequency) of the sound. It is represented by the absorption coefficient α. The following expression characterizes the attenuation due to atmospheric absorption:

$$A_{\mathrm{atm}} = \alpha\times d\,\mathrm{dB} \tag{43}$$

The absorption coefficient increases with increasing frequency (e.g., at 500 Hz, $\alpha \approx 0.21$ dB/100 m).

3) $A_{\mathbf{ground}}$*: Attenuation due to ground effects.*
At larger distances, sound-level reductions are observed as a result of interference between the direct ray and sound rays reflected by the ground. Sound scattering and absorption in the surface layer of air is another important factor. The two effects are difficult to differentiate, and are expressed by the damping attenuation A_{ground}, which in turn depends on the height of the sound source, the height of the receiver, and the distance by which the two are separated. A_{ground} increases from 0 dB to 5 dB with increasing distance.

4) $A_{\mathbf{screen}}$*: Attenuation due to screening.* A "sound shadow" is always observed behind a physical barrier (e.g., a building, a row of houses, a wall, or a dam). Thus, the corresponding sound-pressure level is reduced relative to the value predicted for free sound propagation. A_{screen} can be calculated on the

basis of geometric quantities (height of the source, screen height, receiver height, distance between source and screen and between screen and receiver).

Apart from these damping attenuations, other influences also lead to level reductions. A level *increase* can be caused by the presence of reflecting barriers, an effect that can be calculated by the use of mirror sound sources.

The calculations are carried out in accordance with ISO/DIS 9613-2, generally in octave bands. It is common to use PC programs in performing calculations of this type. The following formula is useful for an approximate immission calculation:

$$L_A = L_{WA} - 27 \log d + 4\,\text{dB} \tag{44}$$

where

L_A = A-weighted sound-pressure level under downwind conditions
L_{WA} = A-weighted sound-power level of the sound source
d = distance (sound source – immission point) in m

This relationship is applicable to free sound propagation over an even surface in the absence of screening, and for the average octave spectrum shown in Figure 8.

7.3. Typical Sound Levels at the Workplace and Near an Industrial Complex

It is necessary to take modern acoustics into account during the technical planning phase of plant development. This is the only way of meeting acoustical requirements established by the responsible authorities (see Section 4) for sound immission limits both at the workplace and in the immediate neighborhood.

In new chemical plants equipped with state-of-the-art noise-control measures, sound-pressure levels exceeding 85 dB (A) under normal operating conditions are found only inside machinery structures and accessible enclosures. Outside the plant, sound-pressure levels should not exceed 80 dB (A).

With modern noise-control measures, the minimum distance between a plant boundary and an adjacent residential neighborhood required to meet an immission-level target of 45 dB (A) is ca. 500 m for a complex occupying 1 km^2, and ca. 350 m for a plant size of 0.3 km^2. These distances are sufficiently small to permit the resolution of acoustic problems related to protection of the neighborhood for most refineries and chemical or petrochemical plants (→ Oil Refining, Chap. 4.2.3.; → Iron, Chap. 4.3.).

8. References

1. ISO 1683 (E), Acoustics – Preferred Reference Quantities for Acoustic Levels, 1983.
2. E. Zwicker, H. Fastl: *Psychoacoustics – Facts and Models,* Springer Verlag, Berlin – Heidelberg 1990.
3. International Electrotechnical Commission, IEC Standard, Publication 651, Sound Level Meters, 1979.
4. ISO 532 (E), Acoustics – Method for Calculating Loudness Level, 1975.
5. ISO 1999 (E), Acoustics – Determination of Occupational Noise Exposure and Estimation of Noise-Induced Hearing Impairment, 1990.
6. ISO 2631-1 (E), Evaluation of Human Exposure to Whole-Body Vibration, part 1: General Requirements, 1985; ISO 2631-2 (E), Evaluation of Human Exposure to Whole-Body Vibration, part 2: Continuous and Shock-Induced Vibration in Buildings (1 to 80 Hz), 1989; ISO 2631-3 (E), Evaluation of Human Exposure to Whole-Body Vibration, part 3: Evaluation of Exposure to Whole-Body z-Axis Vertical Vibration in the Frequency Range 0.1 to 0.63 Hz, 1985.
7. ISO 5349 (E), Mechanical Vibration – Guidelines for the Measurement and the Assessment of Human Exposure to Hand-Transmitted Vibration, 1986.
8. ISO 1996-1 (E), Acoustics – Description and Measurement of Environmental Noise, part 1: Basic Quantities and Procedures, 1982; ISO 1996-2 (E), Acoustics – Description and Measurement of Environmental Noise, part 2: Acquisition of Data Pertinent to Land Use, 1987; ISO 1996-2 (E), Acoustics – Description and Measurement of Environmental Noise, part 3: Application to Noise Limits, 1987.
9. ISO 2372 (E), Mechanical Vibration of Machines with Operating Speeds from 10 to 200 rev/s – Basis for Specifying Evaluation Standards, 1974.

10. ISO 3744 (E), Acoustics – Determination of Sound Power Levels of Noise Sources – Engineering Methods for Free-Field Conditions over a Reflecting Plane, 1981.
11. ISO/DIS 8297, Acoustics – Determination of Sound Power Levels of Multi-Source Industrial Plants for the Evaluation of Sound Pressure Levels in the Environment – Engineering Method, 1988.
12. ISO/TR 7849 (E), Acoustics – Estimation of Airborne Noise Emitted by Machinery Using Vibration Measurement, 1987.
13. ISO 9614-1 (E), Determination of Sound Power Levels of Noise Sources Using Sound Intensity, part 1: Measurement at Discrete Points, 1993: part 2: Measurement at Planned Points (incomplete).
14. ISO 9613-2 (E), Acoustics – Attenuation of Sound during Propagation Outdoors, part 2: A General Method of Calculation, 1992.

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Air

Standard Article

John Wolf[1], Gerhard Manier[2], Thomas Eikmann[3], Gerhard Scholl[4], Gerhard Feldhaus[5], Klaus Grefen[6], Peter Bruckmann[7], Hans-Ulrich Pfeffer[8], Klaus Lützke[9], Eberhard Schmidt[10], Torsten Schmidt[11], Theo Pilhofer[12], Helmut Krill[13], Harald Menig[14], Stephan Schirz[15], Helmut Pollack[16], Walter Simmler[17]

[1]Kommission Reinhaltung der Luft im VDI und DIN, Düsseldorf, Federal Republic of Germany
[2]Institut für Meteorologie der Technischen Hochschule, Darmstadt, Federal Republic of Germany
[3]Universität Gießen, Hygiene-Institut, Gießen, Federal Republic of Germany
[4]Bad Oeynhausen, Federal Republic of Germany
[5]Bonn, Federal Republic of Germany
[6]Kommission Reinhaltung der Luft im VDI und DIN, Düsseldorf, Federal Republic of Germany
[7]Landesumweltamt Nordrhein–Westfalen, Essen, Federal Republic of Germany
[8]Landesumweltamt Nordrhein-Westfalen, Essen, Federal Republic of Germany
[9]RWTÜV Anlagentechnik GmbH, Essen, Federal Republic of Germany
[10]Universität Karlsruhe, Institut für Mechanische Verfahrenstechnik und Mechanik, Karlsruhe, Federal Republic of Germany
[11]Engelhard Process Chemicals GmbH, Hannover, Federal Republic of Germany
[12]QVF Glastechnik GmbH, Wiesbaden, Federal Republic of Germany
[13]Lurgi Umwelt GmbH, Frankfurt/Main, Federal Republic of Germany
[14]Fachhochschule Frankfurt am Main, Frankfurt/Main, Federal Republic of Germany
[15]Fachhochschule Münster, Fachbereich Versorgungstechnik, Steinfurt, Federal Republic of Germany
[16]Gelsenkirchen, Federal Republic of Germany
[17]Bayer AG, Leverkusen, Federal Republic of Germany

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b07_403.pub2
Article Online Posting Date: December 15, 2006

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Introduction: Emissions due to the Chemical Industry
1.1. Greenhouse Gases
1.2. Nitrogen and Sulfur Oxides
1.3. Volatile Organic Compounds
1.4. Emission Inventories
2. Physics and Chemistry of the Atmosphere
2.1. Chemistry of the Atmosphere
2.1.1. Formation of the Primeval Atmosphere
2.1.2. Formation of the Present Atmosphere
2.1.3. Composition of the Present Atmosphere
2.1.4. Trace Gas Cycles
2.1.5. The Aerosol
2.2. Energetics of the Atmosphere

2.2.1. Solar Radiation
2.2.2. Infrared Radiation
2.2.3. Radiation Equilibrium
2.2.4. Energy Balance
2.3. Dynamics of the Atmosphere
2.3.1. Forces
2.3.2. Motions
2.3.3. Stability
2.4. Air-Pollution Control
2.4.1. Methods of Calculating the Dispersion of Pollution
2.4.1.1. Empirical Formulas
2.4.1.2. Gauss Formulas
2.4.1.3. Diffusion Equation
2.4.1.4. Stationary Solution of the Diffusion Equation
2.4.1.5. Particle Models
2.4.2. Use of the Diffusion Calculation
2.4.3. Environmental Compatibility and Simulation of the Environment
3. Effects of Air Pollutants
3.1. Effects on Humans
3.2. General Effects
3.2.1. Specific Pollutants
3.2.2. Persons Needing Special Protection
3.2.3. Carcinogenicity, Mutagenicity, and Teratogenicity
3.3. Environmentally Relevant Air Pollutants
3.3.1. Sulfur Dioxide
3.3.2. Nitrogen Oxides
3.3.3. Ozone
3.3.4. Carbon Monoxide
3.3.5. Dusts
3.3.6. Lead
3.3.7. Cadmium
3.3.8. Arsenic
3.3.9. Benzene
3.3.10. Polycyclic Aromatic Hydrocarbons
3.3.11. Diesel Motor Emissions
3.3.12. Dioxins and Furans
3.4. Effects on Plants
3.4.1. Relationship between Immission and Effect
3.4.2. Effect of Individual Pollutants
3.4.2.1. Sulfur Dioxide
3.4.2.2. Hydrogen Fluoride
3.4.2.3. Hydrogen Chloride and Chlorine
3.4.2.4. Nitrogen Dioxide and Ammonia
3.4.2.5. Hydrogen Sulfide
3.4.2.6. Ethylene
3.4.2.7. Ozone and Other Photooxidants
3.4.2.8. Dust
3.4.2.9. Acid Rain
4. Laws and Technical Regulations
4.1. Introduction
4.2. Federal Antipollution Law
4.3. Federal Statutory Antipollution Regulations Relating to Air Pollution and Plant Safety
4.3.1. First BImSchV
4.3.2. Second BImSchV
4.3.3. Third BImSchV
4.3.4. Fourth BImSchV
4.3.5. Seventh BImSchV
4.3.6. Eleventh BImSchV
4.3.7. Twelfth BImSchV

4.3.8. Thirteenth BImSchV
4.3.9. Seventeenth BImSchV
4.3.10. Nineteenth BImSchV
4.3.11. Twentieth BImSchV
4.3.12. Twenty-First BImSchV
4.3.13. Twenty-Second BImSchV
4.3.14. Twenty-Third BImSchV
4.3.15. Twenty-Fifth BImSchV
4.3.16. Twenty-Seventh BImSchV
4.3.17. Other Federal Regulations
4.3.18. Smog Regulations of the Länder
4.4. General Administrative Regulations for the Prevention of Air Pollution
4.4.1. TA Luft
4.4.2. Fourth BImSchVwV: Determination of Air Pollution in Test Areas
4.4.3. Fifth BImSchVwV: Register of Emissions in Test Areas
4.5. Technical Regulations for the Prevention of Air Pollution
4.5.1. National Regulations
4.5.2. European Regulations
4.5.3. International Regulations
5. Air Pollution Monitoring
5.1. Measurement Planning
5.1.1. Types of Air Pollution Monitoring
5.1.2. Planning of Pollution Measurements
5.1.2.1. Measurement Networks (General Air Pollution Monitoring)
5.1.2.2. Evaluation of Data; Period and Frequency of Measurements
5.1.2.3. Planning Measurements on Industrial Plants
5.2. Quality Assurance
5.2.1. Procedure-Based Methods
5.2.2. User-Based Methods
5.3. Measuring Methods
5.3.1. Measuring Methods for Inorganic Gases
5.3.1.1. Automatic Measuring Methods
5.3.1.2. Calibration of Automatic Measuring Methods
5.3.1.3. Manual Methods
5.3.2. Measurement and Analysis of Particulate Atmospheric Impurities
5.3.2.1. Measurement of Particle Mass and Particle Fractions
5.3.2.2. Measurement of Particle Numbers and Size Distributions
5.3.2.3. Measurement of Inorganic Components in Suspended Particle Matter
5.3.2.3.1. Measurement of Major Ionic Compounds in Suspended Particulate Matter
5.3.2.3.2. Measurement of Heavy Metals and Metalloid Compounds in Suspended Particulate Matter
5.3.2.3.3. Measurement of Soot (Elemental Carbon)
5.3.2.3.4. Measurement of Fibers
5.3.3. Measurement of Deposition
5.3.3.1. Measurement of Total (Bulk) Deposition
5.3.3.2. Measurement of Wet Deposition
5.3.3.3. Special Sampling Devices for Deposition Measurements
5.3.3.4. Analysis
5.3.4. Optical Methods of Remote Measurement
5.3.4.1. Fourier Transform Infrared Spectroscopy
5.3.4.2. Differential Optical Absorption Spectroscopy
5.3.4.3. Long-Path-Length Absorption Spectroscopy with Tunable Diode Lasers
5.3.4.4. Differential Adsorption LIDAR
5.3.4.5. Correlation Spectroscopy and Gas Filter Correlation Spectroscopy
5.3.4.6. Significance of Remote Measurement Methods
5.3.4.7. Measurement of Unstable Trace Substances
5.3.5. Measurement of Organic Atmospheric Pollutants
5.3.5.1. Overview
5.3.5.2. Summation Methods (Total-C Measurements)

5.3.5.3. Chromatographic Methods
5.3.5.3.1. Chromatographic Methods for Volatile Organic Compounds (VOCs)
5.3.5.3.2. Chromatographic Methods for Low-Volatility Organic Compounds
5.3.5.4. Determination of Organic Compounds with Functional Groups
5.3.6. Olfactometric Methods
5.3.7. Screening Measurements
6. Emission Measurements
6.1. Introduction
6.2. Objectives and Planning of Measurements
6.3. Requirements for Measurement Techniques and Measurement Stations
6.3.1. Quality of Measurement Methods
6.3.2. Choice and Equipment of Measurement Stations
6.4. Requirements for Continuous Measurement and Associated Computer Systems
6.4.1. Minimum Legal Requirements
6.4.2. Testing the Suitability of Continuous Measurement Methods and Evaluation Computers
6.4.3. Testing and Calibration of Measurement Equipment
6.5. Physical State Parameters, Reference Value Measurements, and Evaluation Computers
6.5.1. Temperature
6.5.2. Moisture Content
6.5.3. Carbon Dioxide Content
6.5.4. Oxygen Content
6.5.5. Evaluation Computers
6.6. Measurement of Gas Velocity and Flow Rate
6.7. Measurement of Particles
6.7.1. Sampling
6.7.2. Discontinuous Methods
6.7.3. Continuous Methods
6.7.3.1. Optical Methods
6.7.3.2. Radiometric Methods
6.7.4. Particle Size of Dusts
6.7.5. Heavy Metals as Dust Constituents
6.8. Measurement of Gaseous Inorganic Emissions
6.8.1. Sampling
6.8.2. Sulfur Oxides
6.8.2.1. Discontinuous Methods
6.8.2.2. Continuous Methods
6.8.3. Nitrogen Oxides
6.8.4. Carbon Monoxide
6.8.5. Fluorine Compounds
6.8.6. Chlorine Compounds
6.8.7. Hydrogen Sulfide
6.8.8. Ammonia
6.9. Measurement of Gaseous Organic Compounds
6.9.1. Sampling
6.9.2. Measurement of Total-C
6.9.3. Chlorinated Hydrocarbons
6.9.3.1. Polychlorinated Dibenzodioxins and Polychlorinated Dibenzofurans (PCDD/F)
6.9.3.2. Polychlorinated Biphenyls
6.9.3.3. Chlorobenzenes and Chlorophenols
6.9.3.4. Volatile Chlorinated Hydrocarbons
6.9.3.5. Chlorofluorocarbons
6.9.4. Hydrocarbons
6.9.5. Other Compounds
6.10. Measurement of Odor Emissions
7. Separation and Purification
7.1. Separation of Solid and Liquid Particles from Gases
7.1.1. Introduction

7.1.2. Characterization of Particle Separation
7.1.3. Gas Cyclones
7.1.3.1. Mode of Operation and Basic Designs
7.1.3.2. Collection Efficiency and Pressure Drop
7.1.3.3. Operational Characteristics
7.1.4. Wet Scrubbers
7.1.4.1. Mode of Operation and Basic Designs
7.1.4.2. Collection Efficiency and Pressure Drop
7.1.5. Filters
7.1.5.1. General
7.1.5.2. Deep-Bed Filters
7.1.5.3. Surface Filters
7.1.5.4. Granular-Bed Filters
7.1.6. Electrical Precipitators
7.1.6.1. Mode of Operation and Basic Designs
7.1.6.2. Design Calculations
7.2. Waste-Gas Purification by Thermalor Catalytic Reactions
7.2.1. Fundamentals
7.2.1.1. Types of Reaction
7.2.1.2. Reaction Engineering
7.2.1.2.1. Homogeneous Gas-Phase Reactions
7.2.1.2.2. Heterogeneous Gas-Phase Reactions, Catalytic Waste-Gas Purification
7.2.1.3. Process Selection and Plant Concepts
7.2.1.4. Operational Check
7.2.2. Thermal Processes for Waste-Gas Purification
7.2.2.1. Oxidation Processes
7.2.2.1.1. Waste-Gas Disposal in Furnaces
7.2.2.1.2. Thermal Combustion
7.2.2.1.3. Combustion in Flares
7.2.2.1.4. Safety Engineering in Thermal Waste-Gas Purification
7.2.2.2. Reduction Processes
7.2.2.2.1. Staged Combustion
7.2.2.2.2. Thermal Reduction of Nitrogen Oxides
7.2.3. Catalytic Waste-Gas Purification Processes
7.2.3.1. Catalytic Reduction of Nitrogen Oxides in Flue Gases and Process Off-Gases
7.2.3.2. Engine Exhaust Emission Control
7.2.3.3. Gas-Turbine Exhaust Emission Control
7.2.3.4. Catalytic Purification of Claus Plant Off-Gases
7.2.3.5. Catalytic Oxidation in Industrial Off-Gas Streams
7.3. Absorption
7.3.1. Introduction
7.3.2. Basic Principles of Absorption Processes
7.3.3. Operating Principles of Absorption Apparatus
7.3.4. Absorption Apparatus
7.3.5. Selection of Equipment
7.4. Adsorption
7.4.1. Fundamentals
7.4.2. Industrial Adsorbents
7.4.3. Design Criteria for Adsorbers
7.4.4. Adsorber Designs
7.4.5. Plant Concepts
7.4.6. Adsorption Processes
7.4.7. Recycling and Disposal of Spent Adsorbents
7.5. Biological Processes
7.5.1. Overview and Classification
7.5.2. Process Engineering Principles
7.5.2.1. Biological Conditions
7.5.2.2. Physicochemical Conditions
7.5.3. Biofilters

7.5.4. Bioscrubbers
7.6. Flue-Gas Purification in the Incineration of Hazardous Waste
8. Photochemical Degradation
8.1. Introduction
8.2. Natural and Anthropogenic Emissions
8.3. Reaction Cycles in the Atmosphere
8.3.1. The Troposphere
8.3.2. The Stratosphere
8.4. Interaction with Surface Layers of Aquatic and Terrestrial Environments
8.5. Measurement and Computation of Degradability
8.5.1. Lifetime in the Homogeneous Gas Phase
8.5.2. Lifetimes of Adsorbates Aerosols
8.5.3. Assessment and Evaluation
9. Acknowledgement

[Top of Page]

1. Introduction: Emissions due to the Chemical Industry

John Wolf

1.1. Greenhouse Gases

The protection of the earth's atmosphere is one of the greatest environmental-political challenges worldwide. International agreements are needed to reduce man-made emissions and to introduce effective countermeasures. Precautionary measures on a national scale are certainly necessary, but alone are not sufficient to prevent the global changes to the earth's atmosphere that could result from the greenhouse effect or the ozone hole. Using the chemical industry as an example, the following chapters describe measures taken in Germany, and discuss the proportion of the total emissions for which the chemical industry is responsible.

The air — the natural mixture of gases that forms the earth's atmosphere — contains 78 % nitrogen and 21 % oxygen at sea level, excluding water vapor. The trace gases present in air are extremely important. Without them, the temperature at the earth's surface would be ca. – 18 °C.

Concentrations of these trace gases have been measured continuously only in recent times. The longest series of records is for carbon dioxide (CO_2). The continuous increase in CO_2 concentration since the beginning of industrialization over 200 years ago is shown in Figure 1. The growth rates for nonmethane hydrocarbons (NMHC), nitrogen oxides (NO_x), carbon monoxide (CO), methane (CH_4), dinitrogen oxide (N_2O), and ozone (O_3) are given in Table 1. As a result of the tendency of these concentrations to increase, some scientists have predicted a temperature increase of 3 – 5 °C, based on mathematical models. It might be thought that such a small temperature rise would not make any action necessary, as temperature changes have often been observed in the earth's history. The most recent change was a reduction of 6 °C following a temperate period, which led to an ice age. Today we are in a warm period, and a further temperature increase of 3 – 5 °C could change the precipitation distribution, shift climatic and vegetation zones, and hence contribute to an increase in climate-induced damage, degradative changes to the soil, and deterioration in the world food supply. The main contributors to this greenhouse effect are carbon dioxide (ca. 50 %), followed by methane (19 %) and the chlorofluorohydrocarbons (17 %). Other contributors include the ozone of the troposphere (ca. 8 %), nitrogen oxides (ca. 4 %), and stratospheric water vapor (ca. 2 %). The increase in the methane concentration is often directly related to the food requirements of the increasing world population (e.g., a byproduct of animal husbandry). The increase in CO_2 concentration correlates with the growth of energy production and traffic, associated with industrialization. The aim of German legislation is to reduce the energy-related emission of carbon dioxide by 25 % by the year 2005. The volume emitted during the year 1987 is the reference figure: emission from the pre-unification Federal Republic in 1987 was 726×10^6 t; total emission from the reunified Federal Republic in 1990 was 1031×10^6 t [1], [2].

Table 1. Mean global concentrations and rates of increase of the most important trace gases

Compound	Concentration, ppb	Rate of increase, %/a
CO_2	350×10^3	0.5
C_2H_6	1	0.8*
NO_x	very variable	1
CO	100	0.8*
CH_4	1640	1
N_2O	310	0.3

O_3	30	1*

* Increase observed in the northern hemisphere only.

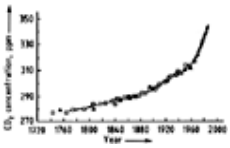

Figure 1. Increase in the CO_2 content of the atmosphere

The main international agreement to reduce these man-made emissions is the United Nations Framework Convention on Climate Change (UNFCCC), adopted in New York on May 9, 1992. The ultimate objective of this convention is to achieve stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system and to enable economic development in a sustainable matter. The conference of the Parties, as the supreme body of this convention, selected Bonn, Germany, to host the permanent secretariat of the Convention [13]. As of January 28, 1998 the Convention had received 174 instruments of ratification. In 1995 in Berlin, the UNFCCC member nations agreed to negotiate setting new commitments for limiting greenhouse gas emissions after the year 2000 for developed nations. The EU proposed a 7.5 % emission reduction by 2005 and a 15 % reduction from the 1990 level by 2010. Germany declared that it would reduce its 1990 CO_2 emissions by 25 % by 2005, which represents a further reduction relative to the initial decision made in 1990 [1]. In the third session of the conference of the parties, held in December, 1997, the Kyoto Protocol was adopted. In Kyoto the parties assured that their aggregate anthropogenic carbon dioxide equivalent emissions of greenhouse gases (CO_2, CH_4, N_2O, fluorohydrocarbons, perfluorocarbons, and SF_6) do not exceed their assigned amounts and will be reduced to at least 5 % below 1990 levels in the commitment period 2008 to 2012. As of 7 January 1999, 71 countries have signed the Kyoto Protocol.

The chemical industry will contribute to the reduction in CO_2 emissions by utilizing industrial waste heat, by increasing the efficiency of its fuel-fired equipment (and other energy-saving measures), and by rationalizing energy utilization by matching supply to demand. The German chemical industry committed itself to reducing its CO_2 emissions by the year 2005 to more than 30 % below 1990 levels, in line with the voluntary agreement reached within the industry in February 1996. The objective of this agreement is to fulfill the climate change Convention and outperform the target of the German Government. By 1997 the German chemical industry reduced the specific energy consumption to 22 % and the CO_2 emission to 26 % below 1990 levels. The improvement made from 1990 to 1994 is listed in the following (CO_2 emissions due to power generation in 10^6 t) [14]:

1990	67.8
1991	56.3
1992	51.4
1993	48.0
1994	45.2

In the year 2000, the CO_2 emission of the German chemical industry due to power generation will be reduced by 30 % below 1990 levels.

1.2. Nitrogen and Sulfur Oxides

Whereas the environmental consequences of increased CO_2 emission cannot at present be estimated exactly, information is available on the effect of atmospheric impurities which have accumulated in the past as a result of the limited technology then available for reducing emissions (see Chap. Effects of Air Pollutants). The environmental effects first became evident as damage to forests caused by acid rain [3].

A considerable proportion of the acid in acid rain is derived from nitrogen oxides (NO_x), most of which is produced by combustion processes in factories and internal combustion engines by oxidation of nitrogen in the fuel and combustion air. A relatively small proportion is produced in the production of nitric acid.

The emission of nitrogen oxides in the Federal Republic of Germany in 1990 was ca. 2.6×10^6 t in the 11 old Länder and ca. 550×10^3 t in the 5 new Länder [3], [4]. Oxidation of the sulfur in fuel during combustion processes gives rise to sulfur dioxide emissions, e.g., from sulfuric acid production, from natural gas desulfurization plants (Claus plants), and from oil refineries. The emission of sulfur dioxide in 1990 was ca. 10^6 t (old Federal Republic). The new Länder emitted 4.8×10^6 t SO_2. The corresponding emissions of carbon monoxide in 1990 were 9×10^6 t (3.8×10^6 t), and of airborne particulate 480×10^3 t (1.8×10^6 t) [5].

Nitrogen oxide emissions in Germany in 1994 were 2.2×10^6 t and sulfur dioxide emissions 2.9×10^6 t [15]. In 1995, the German chemical industry emitted 63.4×10^3 t of sulfur dioxide and 39.6×10^3 t of nitrogen oxides, which includes the emissions of the old and new Länder. In the former Democratic Republic, the emissions of sulfur dioxide by the chemical

industry were reduced by 74 %, and those of nitrogen oxides by 57 %.

1.3. Volatile Organic Compounds

Emissions of volatile organic compounds (VOCs) in 1995 were 2.1×10^6 t. About 40 % of this resulted from incomplete combustion in internal combustion engines, mainly in road and rail vehicles. Approximately 20 % of the VOCs are emitted during storage and handling, during the filling of vehicles, and by evaporation from vehicles.

The annual total emission of 31×10^3 t of benzene in 1994 is mainly (> 90 %) due to motor vehicles. The benzene content of gasoline for Otto engines is currently 1.7 – 2.5 vol %. The main use of benzene is in industrial chemistry for the synthesis of aromatic intermediates and cycloaliphatic compounds. Approximately 80 % of the benzene is used in the production of ethylbenzene and isopropylbenzene (cumene), and as a precursor for styrene, phenol, and cyclohexane, which is further processed to give intermediates for polyamide production.

Approximately 50 % of VOC emission results from the use of solvents, and nearly 40 % of this is emitted in the course of painting. Next in order of quantity come degreasing in the metal industry (8 %), printing (8 %), and adhesives (7 %). A high proportion of these emissions are of halogenated hydrocarbons (HHCs), especially the chlorinated hydrocarbons (CHCs) such as perchloroethylene, trichloroethylene, dichloromethane, and 1,1,1-trichloroethane. In 1991, a total of 259×10^3 t were produced, of which 202×10^3 t were exported. The main uses of CHCs are metal degreasing, solvents for adhesives, plastics processing, solvents in chemical syntheses, paint removal, and textile cleaning.

The use of CHC solvents decreased from 180×10^3 t in 1986 to 53×10^3 t in 1992 following the introduction of recycling systems. Increased use of recycled materials reduced consumption to 30×10^3 t by 1995. In Western Europe, total consumption of CHC solvents in 1990 was 600×10^3 t (510×10^3 t new material, 90×10^3 recycled material). In the EU, 75 % of all CHC solvents used are lost to the atmosphere, while the corresponding figure for Germany is 53 % owing to legal restrictions on emissions [6]. The possibilities of finding substitutes for these applications are limited, so that continued efforts are being directed to the improvement of waste-gas purification techniques.

Production of the following compounds was discontinued as early as the 1980s, mainly in order to reduce emissions of dioxin: trichlorophenol (TCP), 2,4,5-trichlorophenoxyacetic acid (2,4,5-T), pentachlorophenol (PCP), and polychlorinated biphenyls (PCB).

Other man-made greenhouse gases and compounds that contribute to the destruction of ozone in the stratosphere include the chlorofluorohydrocarbons (CFCs) and halons, both of which are products of the chemical industry. Fully halogenated CFCs are stable in the troposphere. However, they can also reach the stratosphere where they are photolytically cleaved by UV radiation. This produces free chlorine atoms which can break down the ozone in a catalytic chain reaction.

A detailed treatment of the photochemical reactions of these and other trace gases in the atmosphere can be found in Chapter Photochemical Degradation. As a result of the Montreal Protocol of 1987 and the follow-up London Conference of 1990, the production of carbon tetrachloride, fully halogenated CFCs, and halons is to cease worldwide by the year 2000 [16]. Production was discontinued in Germany in 1994 and in the EU in 1997. Production of hydrogen-containing CFCs, CH_3CCl_3, and CHF_2Cl will cease in the year 2005 [9]. In the meantime, emissions of hydrogen-containing CFCs in Germany will be much reduced as a result of voluntary agreements with the producers of expanded polystyrene foams for thermal insulation, who have committed themselves to producing these foams without HCFCs by January 2000. In most areas, suitable replacements for CFCs are available [6], [10]. CFCs have not been used since 1988 in hair sprays, deodorants, and other domestic aerosol spray dispensers. CFC-free medicaments are being developed to replace the 900 t of CFCs which are currently used in medical aerosol sprays (e.g., asthma sprays). Since mid-1994 household appliances in Germany have been produced without CFCs; propane – butane mixtures or fluorohydrocarbons (FHCs) are now used as refrigerants. The latter compounds (e.g., R 134 a, CH_2FCF_3) are more short-lived in the atmosphere because of their hydrogen content, and have lower ozone depletion potential (ODP). Pentane is a suitable substitute for producing foamed plastics.

Halons were mainly used in fire extinguishers and industrial fire-fighting equipment. They have now been replaced by CO_2, water, foam, and powder. In 1990, 80 683 t of CFCs were produced in the Federal Republic of Germany. The production of halons was discontinued in 1991.

Guidelines for the limitation of VOC emissions are being prepared in the EU. Also, solvent management plans will be set up. The above-mentioned EU Guidelines will considerably accelerate the conversion to low-solvent water-based paints or even solvent-free powder paints in the paint industry (annual production of paints in 1997: 1.99×10^6 t). The production of high-solids coatings was increased from 6.8×10^3 t (1990) to 31.4×10^3 t (1996). The corresponding production figures for low-solvent paints (the sum of powder, dispersion, electrophoretic, and high-solids paints) are 130×10^3 t (1990) and 186×10^3 t (1997). Meanwhile, the major German automobile manufacturers now have plants available for painting motor vehicles with water-based paints [11]. As halogenated hydrocarbons are not used in these plants, VOC emissions have been considerably reduced. Initial tests show that medium-sized and small automobile painting and repair workshops, numbering around 12 000, can also use water-based paints. In any case, it should be possible to reduce the emission of organic solvents by up to 70 %. The printing and paper industries, which emit ca. 60×10^3 t VOCs (mainly hydrocarbons) could also make a considerable contribution to the planned reduction in emissions by changing to cleaning agents of vegetable origin, (e.g., in offset printing) and by finding substitutes for the solvent-containing paints and adhesives used in finishing processes.

In these and many other cases, further reductions in the emission of hydrocarbon solvents will be achieved by the use of substitute materials or by minimizing the solvent content of products rather than by improving waste-gas cleaning

technology, which can only be considered as secondary or end-of-the-pipe technology.

Approximately 26 × 10^3 t VOCs were emitted in 1995 by chemical plants [12], mainly during the production of plastics, rubber, pharmaceuticals, dyes, paints, solvents, and adhesives.

1.4. Emission Inventories

The European Environment Agency (EEA) was launched by the EU in 1993 and is based in Copenhagen, Denmark. It has a mandate defined by Council Regulation (EEC) No. 1210/90 to ensure the supply of objective, reliable, and comprehensive information at the European level. One task is to work on air pollution related problems [17], [18]. The Cooperative Programme for Monitoring and Evaluation of the Long Range Transmission of Air pollutants in Europe (EMEP) formed by a Protocol under the Long Range Transboundary Air Pollution Convention has established task forces to develop guidelines for the estimation and reporting of emission data. The UNECE Task Force on Emission Inventories (TFEI) will work out an Atmospheric Emission Inventory Guidebook, which will give quantitative information on emissions to the atmosphere by industrial processes, power plants, refineries, chemical plants, and other sources. The draft of this Guidebook is available on the Internet [19]. The final official version will be published on the website of the EEA [17].

The emissions due to the German chemical industry for 1979 – 1990 and 1995 are listed in Table 2 [12]. These data refer to the 11 Länder of the pre-unification Federal Republic. Comparable data from the Democratic Republic for this period are not available. The following data on emissions have been published by the new Länder for the year of reunification (1990): sulfur dioxide 460 × 10^3 t, nitrogen oxides 35 × 10^3 t, carbon monoxide 25 × 10^3 t, airborne particulates 150 × 10^3 t, and VOCs 90 × 10^3 t.

Table 2. Emissions from the German chemical industry in 10^3 t (1979 – 1995)

	1979	1981	1983	1984	1985	1986	1987	1988	1989	1990	1995
SO_2	178.0	158.5	131.0	125.0	111.0	104.0	101.0	91.0	71.0	56.0	65.4
NO_2	114.0	105.0	91.5	89.5	88.0	85.0	80.0	69.0	63.5	52.0	39.6
Dust	27.5	18.0	15.5	14.5	13.0	11.0	9.0	7.5	6.5	6.5	
CO	92.0	60.5	51.0	43.0	44.5	47.0	47.5	56.0	54.5	57.5	
VOCs	76.0	57.5	41.5	34.0	31.5	30.5	20.5	21.0	27.0	25.5	25.9

Unfortunately, it has not been possible to obtain comparable data for the emissions due to the chemical industries of other European countries.

Table 2 shows that there has been a drastic reduction in the pollutants emitted into the atmosphere by the chemical industry (pre-unification Federal Republic), which on average accounts for less than 3 % of total air pollution. This success is due in large measure to the changes brought about by the TA Luft 1986 (see Chap. Effects of Air Pollutants) [7], which has led to an extensive reorganization of the regulations for the control of emissions, and to the updating of the waste gas cleaning technology to correspond to the current state of the art. The TA Luft states that all old plants must be provided with state-of-the-art waste gas cleaning equipment within eight years. Similar regulations (Großfeuerungsanlagenverordnung, 13. BImSchV), were implemented in 1986 for power stations (flue-gas purification, especially by desulfurization and removal of nitrogen compounds). These regulations applied to a total of ca. 50 000 installations.

The costs of environmental protection were correspondingly large, accounting for an average of 15 % of the total investment in chemical plants. Of these costs, ca. 40 % are for air purification [8]. From 1987 to 1996 the German chemical industry invested € 7 × 10^9 in environmental protection equipment.

Since ca. 1970, the German chemical industry has increased its production by ca. 200 %, while at the same time emissions of air pollutants have been reduced by over two-thirds. This shows that industrial growth and the conservation of the environment need not be in conflict with each other. The following chapters consider which techniques can be used (Chap. Flue-Gas Purification in the Incineration of Hazardous Waste) and the legal measures necessary to achieve the desired standard of environmental protection.

[Top of Page]

2. Physics and Chemistry of the Atmosphere

Gerhard Manier

2.1. Chemistry of the Atmosphere

2.1.1. *Formation of the Primeval Atmosphere*

The primeval atmosphere was formed by the evolution of gas from the molten planet earth. As oxygen was chemically combined, the gases of this first atmosphere must have been largely in reduced form. The atmosphere consisted mainly of methane, with some ammonia, hydrogen, and water. In the molten planet, a gigantic blast furnace process started, in which nickel and iron oxides were reduced to nickel and iron. These components eventually formed the earth's core. Oxygen was liberated and increased the degree of oxidation of the earth's mantle and of the outgassed substances. The atmosphere then contained water vapor, carbon dioxide, and nitrogen. It is possible that this atmosphere had the same

composition as the gases produced today by volcanic action, which have the following composition:

H_2O	80 %
CO_2	10 %
H_2S	7 %
H_2	0.5 %
CO	0.5 %
CH_4 and NH_3 traces	

2.1.2. *Formation of the Present Atmosphere*

The present atmosphere was formed along with the hydrosphere, lithosphere, and biosphere. This can be envisaged as follows:

The earth cooled by radiating energy into space. The upper layers of the atmosphere became cool enough to be supersaturated with water vapor, and clouds formed. Droplets that were formed in the clouds led to raindrops. The rain did not yet reach the earth's surface, but vaporized in the lower atmosphere, which was still very hot. The water vapor so formed was carried back to high altitude by convection currents, where it again condensed to form clouds. Rain formed, and again fell back into the lower levels of the atmosphere and evaporated.

This water cycle was associated with a very large transfer of energy, as the heat needed to evaporate the water in the lower levels of the atmosphere was later liberated in the upper levels by condensation to form clouds. This is known as latent heat flow, as the heat is latent in the water vapor transported upwards. This latent heat flow led to further radiative cooling. After ca. 100 000 years, the rain reached the earth's surface, and a period of rain began which lasted ca. 40 000 years. Initially, the rain evaporated immediately at the earth's surface. In the course of time, the earth's surface cooled more and more, so that liquid water collected in the deeper regions, eventually forming the hydrosphere. The rain washed carbon dioxide out of the atmosphere, forming the carbonate sediments of the lithosphere in the seas.

A precondition for the formation of the biosphere was the presence of oxygen in the atmosphere.

The oxygen was formed by plants according to the following equation:

$$6\,CO_2 + 6\,H_2O + \text{light} + \text{chlorophyll} _ C_6H_{12}O_6 + 6\,O_2$$

This process is known as assimilation. At the same time, plant respiration and decay of the plant mass convert oxygen into carbon dioxide. A continuous increase in the oxygen content of the atmosphere can only happen if part of the plant mass does not decay, but is preserved with exclusion of air (i.e., forms sediments). It has been calculated that 1×10^{22} g organic carbon—the amount of carbon preserved with exclusion of air— corresponds to 2.7×10^{22} g oxygen. This is 20 times the amount present in the atmosphere today. The difference was consumed in oxidizing the earth's crust and the primeval atmosphere.

The chronology of these events was probably as follows:

The first appearance of cells in an oxygen-free environment (energy gain by fermentation processes) occurred 4×10^9 years ago.

The first oxygen was produced in water (2 – 3) $\times 10^9$ years ago at a depth great enough to avoid damage to the organisms by UV radiation from the sun, but with adequate availability of light. Part of this oxygen was released from the ocean into the atmosphere and formed the ozone layer. Part of the UV radiation was absorbed by this ozone layer, and life formation moved closer to the surface of the water, the oxygen content of the atmosphere increased, and the ozone layer grew. The ozone layer protected the earth from the UV radiation of the sun to an ever-increasing extent, until eventually life became possible on land. This process was complete ca. 600×10^6 years ago.

2.1.3. *Composition of the Present Atmosphere*

It is convenient to consider the composition of the atmosphere in three parts:

1. Dry air
2. Water in all three states of aggregation
3. Atmospheric aerosol

Composition of the Dry Atmosphere (Table 3). Nitrogen and oxygen together account for >99 % of the atmosphere. If the noble gas argon is included, this figure becomes 99.96 %.

Table 3. Composition of the dry atmosphere [20]

Component	Concentration	Probable residence time in the atmosphere

Principal components		
N_2	78.09 %	2×10^7 a
O_2	20.95 %	
Ar	0.93 %	
CO_2	0.033 %	5 – 10 a
Other components (*constant concentration*)		
Ne	18 ppm	
He	5 ppm	3×10^6 a
Kr	1 ppm	
Xe	0.09 ppm	
CH_4	1.5 ppm	3 a
CO	0.1 ppm	0.35 a
H_2	0.5 ppm	
N_2O	0.25 ppm	<200 a
Other components (*variable concentration*)		
O_3	≤10 ppm (stratosphere)	
	5 – 50 ppb (pure air)	
	≤500 ppb (polluted air near ground)	
H_2S	0.2 ppb (over land)	10 d
SO_2	0.2 ppb (over land)	5 d
NH_3	6 ppb (over land)	1 – 4 d
NO_2	1 ppb (over land)	2 – 8 d
	100 ppb (polluted air near ground)	
CH_2O	0 – 10 ppb	

The carbon dioxide is present at a much lower concentration which varies somewhat, especially close to the surface. If the carbon dioxide is added to the above three gases, the total then accounts for 99.997 % of the dry atmosphere. The small remainder of 0.003 % (30 ppm) is nevertheless extremely important, and includes both highly aggressive gases and gases upon whose presence the very existence of life earth depends, such as ozone. The concentrations of some of these gases can fluctuate considerably.

Variation in Composition with Altitude. If there is a concentration gradient within a given volume of a gas, Brownian motion and diffusion of the gas eventually lead to a uniform spatial distribution. In addition, the gases in the atmosphere are subject to gravitational acceleration towards the earth's center. In the equilibrium state, this force is compensated by a pressure gradient force acting upwards. This force results from the decrease in air pressure with altitude. The partial pressure of a light gas decreases more slowly with altitude than that of a dense gas, a phenomenon known as segregation.

For example, the light gas helium (relative molecular mass 4 g/mol) is present at the earth's surface at a partial pressure of 5.24×10^{-3} hPa. The other air components (mean relative molecular mass 29 g/mol) have a total partial pressure at the earth's surface of 1000 hPa. At an altitude of 113 km, the partial pressures of the helium and of the other air components might be expected to be identical, but this is not the case. Measurements have shown that the composition remains constant in the lower 80 – 100 km, but above 100 km the concentration of the helium decreases less rapidly than that of the heavier constituents of the atmosphere. Segregation in the lower atmospheric region is prevented by other mechanisms, mainly turbulent mixing of the atmosphere by rising and falling air streams. This region, in which the composition of the atmosphere is constant, is named the homosphere, and extends to an altitude of ca. 100 km. In the overlying atmospheric layers, the concentration of the light gases decreases more slowly with altitude than those of the heavy gases, and segregation takes place. The topmost layer of the atmosphere consists almost entirely of hydrogen and helium. The atmosphere above 100 km is known as the heterosphere.

The Ozone Layer. At an altitude of 20 – 30 km, there is a layer in which the concentration of ozone is very high. There are several reactions that lead to the formation and destruction of atomic oxygen and ozone [21].

$$O_2 + h\nu \rightarrow 2O \quad (1)$$

$$O_3 + h\nu \rightarrow O_2 + O \quad (2)$$

These are the primary photochemical reactions. Equation (1) represents cleavage of the oxygen molecule into two oxygen atoms by UV radiation. In Equation (2), the action of radiation up to the IR range of the solar spectrum on an ozone molecule leads to formation of an oxygen molecule and an oxygen atom. Two further, secondary reactions are involved in the formation and destruction of ozone.

$$O + O_2 + M \rightarrow O_3 + M \quad (3)$$

$$O + O_3 \rightarrow 2\,O_2 \quad (4)$$

In Equation (3), a neutral collision partner M must be present to remove the excess energy and momentum liberated by ozone formation.

The probability of a particular reaction taking place depends on the available radiation and the density of the atmosphere. Both of these parameters vary with altitude, the density decreasing, and the radiation increasing with increasing altitude. As these two parameters have opposite effects, there is an altitude at which the ozone concentration is a maximum. This altitude can be calculated, and is ca. 20 km at a latitude of 10° N. At higher latitudes, the maximum is displaced upwards and the concentration decreases. Measurements have shown that at lower latitudes the calculated concentrations are considerably higher than the measured concentrations. At higher latitudes, the measured concentrations are greater and occur at lower altitudes than calculations predict. These differences between measured and calculated values can only be explained by transport of ozone from lower to higher latitudes. Human activities (e.g., CFC production, air traffic) increase the concentration of nitrogen and chlorine radicals (NO_x and ClO_x) in the stratosphere. Both substances lead to ozone destruction, but because the individual reaction cycles are interlinked and extremely complex, the prediction of future changes in the ozone layer is difficult [21].

Uppermost Layers of the Atmosphere. Above 500 km, the atomic forms of the atmospheric gases dominate, and above 1000 km, the light atoms helium and hydrogen are virtually the only species present. Moreover, above 500 km, the mean free path between collisions becomes so large that the particles follow ballistic trajectories. This is known as the spray region, and the atmospheric region is known as the exosphere. When the individual atoms follow ballistic trajectories, they can overcome the attraction of the earth and escape into space. Only hydrogen can reach the escape velocity of 12 km/s to an appreciable extent, so that the amount of hydrogen in the atmosphere is very small. Hydrogen is produced by the dissociation of water vapor. The probability of escape of an oxygen atom is very low, so that it can be assumed that all the oxygen formed on the earth is still present.

2.1.4. *Trace Gas Cycles*

Sulfur Cycle. The following sulfur compounds are present in the atmosphere: H_2S, SO_2, SO_3, and SO_4^{2-}.

All these compounds are chemically reactive. There is a link to the water (hydrological) cycle. Their residence times are short, so that their concentrations can show very large spatial and time-dependent variations.

The most important natural source is the bacterial degradation of organic substances in coastal regions and marshy areas, forming hydrogen sulfide (ca. 32×10^{12} g/a, as sulfur). High wind speeds and the breaking of ocean waves lead to the formation of small water droplets which then evaporate, leaving behind salt (NaCl) as the main component. Sulfates (44×10^{12} g/a) are also formed. Volcanic activity causes the transport of 3×10^{12} g/a S into the atmosphere, mainly as H_2S and SO_2. The most important man-made source of sulfur compounds is the burning of coal and mineral oil, which forms sulfur dioxide (65×10^{12} g/a).

The most important sink for the sulfur dioxide and sulfates is wet deposition (106×10^{12} g/a) by the atmospheric processes of washout and rainout. Gases and sulfate can dissolve in raindrops and be transported back to the earth's surface, a process known as washout. The condensation nuclei from which cloud droplets are formed, and which can then become raindrops, contain gases and sulfates. This process is known as rainout.

In addition to wet deposition, dry deposition of sulfates and sulfur dioxide on plants and on the surface of the earth or water (3.8×10^{13} g/a) can occur. The total turnover of sulfur, is 144×10^{12} g/a.

Nitrogen Cycle. The nitrogen molecule is the main constituent of the atmosphere. Nitrogen is chemically unreactive, and its contribution to the nitrogen compounds cycle is of minor importance, due to the enormous size of the reservoir.

Dinitrogen oxide, N_2O, has a long residence time in the atmosphere. It is formed by the reduction of nitrates and nitrites by bacteria in the soil. The compound probably disappears by photochemical dissociation:

$$N_2O + h\nu \rightarrow N_2 + O$$

There is no link to the other nitrogen cycles. N_2O takes part in catalytic ozone degradation reactions.

Ammonia and its compounds are chemically reactive. There is a link to the water cycle. Because the residence time of these gases in the atmosphere is short, the concentration can show large spatial and time-dependent variations.

The main source is the degradation of organic substances in the soil. Figures for the amount (as nitrogen) that enters the atmosphere fluctuate considerably [$(113 – 244) \times 10^{12}$ g/a]. Dry and wet deposition carry NH_3 back to the earth's surface. Reactions with water vapor, sulfur dioxide, and carbon dioxide can form ammonium sulfate or ammonium bicarbonate as aerosol particles. A small sink for ammonia is provided by the reaction with the radical ·OH, which forms NO_x and water. This represents the only link between the ammonia cycle and the NO_x cycle.

Nitrogen oxides, NO_x are chemically reactive. There is a link to the water cycle. The residence time in the atmosphere is short, and the concentrations can therefore show very large spatial and time-dependent variations. The most important

substances are nitrogen monoxide (NO) and nitrogen dioxide (NO_2). Other compounds include dinitrogen trioxide (N_2O_3), nitrous acid (HNO_2), dinitrogen pentoxide (N_2O_5), nitric acid (HNO_3), nitrites (NO_2^-) and nitrates (NO_3^-). The nitrates have the highest oxidation state, and are therefore the stable final state after the reaction of NO via NO_2 to form NO_3^-. Each oxidation step occurs by reaction with ozone (O_3).

The main sources of nitrogen oxides are bacterial nitrate degradation in the soil (2.1 – 8.9 × 10^{13} g/a) and the combustion of oil in industry and by motor vehicles (19 × 10^{12} g/a). Further sources are the link to the ammonia cycle and the formation of nitrogen oxides by lightning. Organic nitrogen compounds are very important in the soil and in the ocean, but they can be harmful constituents of the atmosphere.

Carbon Cycle. Four atmospheric carbon cycles exist, involving the substances CO_2, CO, CH_4, and other carbon compounds.

Carbon dioxide is one of the most important constituents of the atmosphere. It has a long residence time. The main sources are the combustion of coal, the respiration of living organism, and the degradation of organic substances. The main sink is photosynthesis, a further sink being the dissolution of carbon dioxide in the oceans and formation of calcium carbonate sediments. The carbon dioxide content of the atmosphere increases by ca. 0.75 ppm annually.

The greenhouse effect of the atmosphere (solar radiation passes through, thermal radiation is retained) depends on the concentrations of water vapor and carbon dioxide. Therefore, an increase in the carbon dioxide content can lead to a change in the greenhouse effect, and hence to climatic change. However, because these processes are extremely complex, predictions are difficult [22].

Carbon monoxide can be oxidized to carbon dioxide, but this reaction does not proceed in the atmosphere, so that CO has a relatively long residence time. Sources of carbon monoxide include incomplete combustion and generation by microorganisms on the surface of the ocean. Bacteria in soil consume carbon monoxide. It also reacts with ·OH radicals. Carbon monoxide concentrations can show very large spatial and time-dependent variations, especially in towns.

Methane is a relatively common gas and has low chemical reactivity. It is produced and consumed as part of the biological cycle. Its concentration is presently increasing by ca. 1.5 %/a. Reaction with oxygen in the stratosphere leads to the formation of the methyl radical and ·OH, which also contributes to catalytic ozone breakdown. Methane also contributes to the greenhouse effect.

Water Cycle. The total mass of water on the earth is 1.36 × 10^{21} kg, corresponding to 2.66 × 10^6 kg water per m^2, or a layer of water 2660 m thick. Of this water, 97 % is contained in the oceans, and 2.3 % in ice. Subterranean water (groundwater) accounts for 0.7 %, and lakes and rivers 0.01 %. The atmosphere contains 0.001 % or 26.6 kg/m^2. This amount of water corresponds to a depth of 2.66 cm. As the average annual precipitation is 857 mm, and only 26.6 mm is present in the atmosphere, the equivalent of the total atmospheric water content must be deposited on the earth's surface in 10 d to give the above mean annual precipitation figure, i.e., the cycle time of the water in the atmosphere is 10 d. The water cycle causes cleansing of the atmosphere, as all substances are eventually converted to water-soluble products and return to the earth's surface with the rain. Evaporation produces only water vapor.

2.1.5. *The Aerosol*

The atmosphere contains particles as well as gaseous components. These particles are much larger than single molecules, but are still so small that their settling rate is generally negligible. Their size varies between 5 × 10^{-3} and 20 µm, and their number density between 10^2 and 10^6 cm^{-3}. Particles with a radius of 0.1 µm are especially abundant. Particles larger than 10 µm are extremely rare, as their settling rate is so great that they become deposited on the earth's surface close to their source. Very small particles also do not occur, because they agglomerate rapidly to form larger particles.

Sources and Chemical Composition of the Aerosol. There are four mechanisms by which aerosols are produced:

1. Combustion (industrial combustion, forest fires, traffic), forming salts, carbonates, and lead.
2. Reactions in the gas phase (sulfur and nitrogen cycle) to form sulfates and nitrates.
3. Whirling up of solid material, whereby silicates and salts of calcium, potassium, and sodium enter the atmosphere.
4. Whirling up of liquid. At high wind speeds, air bubbles form on the ocean. When these bubbles burst, many small water droplets are formed. These then evaporate, leaving salt as an aerosol.

Sinks for the Aerosol. Large particles leave the atmosphere owing to their high rate of settling, becoming deposited preferentially on vegetation. Larger particles are also washed out of the atmosphere by rain.

The residence time of large particles (>10 µm) amounts to minutes or hours. Particles of 2 –10 µm have a residence time of hours or days, and those of 0.1 µm a residence time of several days. Particles smaller than 0.05 µm agglomerate to form larger particles and have a short residence time (< 1 h).

Effect of Altitude on Aerosol Concentration. In the lowest 5 km, the concentration of the aerosol decreases much more rapidly than that of the other atmospheric constituents. Whereas the altitude of the homogeneous atmosphere is 8 km, it is found that a hazy atmosphere is homogeneous to an altitude of 3 km. (The altitude of a homogeneous atmosphere is that for which the density of the air or of the aerosol, independent of the altitude, is equal to the density at the earth's surface.) Above an altitude of 5 km, the aerosol behaves like air, i.e., the densities decrease at the same rate. Between 20 and 30 km, the aerosol concentration remains constant. Aerosol particles are produced in this region.

2.2. Energetics of the Atmosphere

2.2.1. *Solar Radiation*

The sun is the sole source of energy for all processes on the earth. At the upper limit of the atmosphere, the radiation flux density is 1376 W/m^2. This radiation is attenuated in the atmosphere by scattering and absorption caused by air molecules, certain gases, the aerosol, and clouds.

Attenuation by the gases is proportional to λ^{-4}. With increasing path length through the atmosphere, the energy maximum shifts from the blue to the red region. One consequence of this is the red appearance of the sun at sunrise and sunset.

The aerosol also attenuates the shorter wavelengths more strongly than the longer (~ λ^{-1}). Different particle size distributions of the aerosol can lead to very different directional distributions of the scattered radiation. Forward scattering dominates, in contrast to scattering by air molecules.

The absorption of solar radiation by gases leads to excitation. Owing to the high density of the air, the collision number is so large that the energy of excitation can be transferred as kinetic energy to other air molecules. This leads to an increase in the temperature of the gas. Thus, unlike scattering, absorption leads to the transformation of radiant energy into thermal energy.

The nitrogen molecules and the oxygen atoms or molecules absorb UV radiation (<200 nm) even at high altitudes. Ozone absorbs at wavelengths of <300 nm, so that radiation in this range does not reach the earth's surface. Ozone absorbs only weakly in the visible region. Water vapor absorbs in the infrared.

Even a thin cloud layer prevents the sun from being visible, i.e., direct solar radiation is zero.

Part of the direct solar radiation is absorbed at the earth's surface. The energy yield increases with the altitude of the sun and also with increasing darkness of the earth's surface. Forests and water surfaces are very dark, while snow has the brightest surface.

Diffuse solar radiation is produced by scattering of direct solar radiation by air molecules and the aerosol. The sum of the direct and diffuse solar radiation is the global radiation. In the following, some values of the global radiation F_G for a cloudless sky at various zenith angles ζ (measurements were carried out in Frankfurt/Main, Germany) are given:

ζ,	30°	40°	50°	60°	70°	80°
F_G, W/m^2	963	830	670	481	293	119

If the sky is completely covered with clouds, the following percentages of the global radiation of a cloudless sky are obtained:

Fog	15 %
Rain clouds	19 %
Low clouds	25 – 33 %
Medium altitude clouds	40 – 50 %
High ice clouds	80 %

2.2.2. *Infrared Radiation*

All bodies whose temperature is above absolute zero emit radiation. In the atmosphere, these includes the following emitters: the earth's surface, clouds, water vapor, and carbon dioxide.

The earth's surface and clouds radiate like black bodies, i.e., the radiation flux density F depends on the surface temperature T and the longwave emissivity ϵ of the material and is given by:

$$F = \varepsilon \times 5.67 \times 10^{-8} \times T^4$$

For most surfaces, ϵ lies between 0.92 and 0.98.

Unlike solid or liquid bodies, gases do not radiate at all in certain wavelength ranges, and radiate like a black body in other ranges.

2.2.3. *Radiation Equilibrium*

It can be assumed that the planet earth is in thermal equilibrium. Thus, the absorbed short-wave radiation from the sun is equal to the long wave radiation emitted from the earth. In the absence of an atmosphere, this would occur if the earth's surface temperature were – 21 °C. But the average temperature of the earth's surface is + 14.3 °C, and a temperature of – 21 °C is found at an altitude of ca. 6 km. This altitude corresponds to the average altitude of the upper surface of the clouds. The atmosphere and the clouds provide effective protection from radiation. This is also known as the greenhouse effect, whereby shortwave radiation is allowed to reach the earth, while water vapor, carbon dioxide, aerosol, and clouds only allow the partial transmission of long-wave radiation.

2.2.4. *Energy Balance*

The radiation balance of the earth's surface is positive, and that of the atmosphere negative. The energy balances of the earth and the atmosphere are evened out by the turbulent flows of sensible and latent heat.

A turbulent transport process, whether of temperature, water vapor, or pollutants can be pictured as follows: The air flow is split up into several densely packed flow filaments. In each flow filament, the mean flow rate, temperature, water vapor content, and concentration of pollutants is constant. There are differences between the individual flow filaments. In a turbulent flow, there is an intensive exchange of air parcels between the individual flow filaments. These air parcels transfer the property of one flow filament into another. If there is a temperature difference between the two flow filaments, this exchange of air parcels implies a transfer of heat. The transfer is always in such a direction that temperature differences are evened out. The same considerations also apply to water vapor (latent heat flow) or any other substance.

If we consider the effect of geographical latitude on the energy balance, there is a surplus at the lower latitudes, while, at high latitudes where the sun does not shine for a large part of the year or the angle of incidence of the solar radiation is low even in summer, the energy balance is negative. As the temperature does not increase continuously in the lower latitudes or decrease continuously in the polar regions, there must be a transfer of energy from the lower latitudes to the higher. This transfer exhibits a maximum in the geographical latitude at which the energy balance is zero. This is the case at a latitude of ca. 40° N or 40° S. At this latitude, the energy transferred must be 4.2×10^{15} W. This transfer takes place as a result of normal weather processes, in high latitudes by the interchange of low- and high-pressure areas, and in low latitudes by the circulation of trade winds.

2.3. Dynamics of the Atmosphere

2.3.1. *Forces*

Five forces act in the atmosphere and determine wind flows.

The gravitational force causes all objects, including the air, to accelerate towards the center of the earth.

The pressure gradient force operates wherever pressure differences exist. It accelerates the air from high- to low-pressure regions. In the vertical direction, it opposes the gravitational force.

The Coriolis force is an apparent force that must be introduced because the air movements are observed on a rotating earth. The Coriolis force diverts flows to the right in the northern hemisphere, and is proportional to the flow rate.

The centrifugal force acts in an outwards direction on curved flow paths (e.g., low-pressure areas or tornadoes). It is proportional to the square of the flow rate, and is inversely proportional to the radius of the flow path.

The frictional force is always present if there is a spatial change in flow rate, as is always the case in the vicinity of the earth's surface. The frictional force is in the opposite direction to the direction of the flow and its magnitude increases with increasing flow rate and with decreasing stability of the atmosphere (see also Section Stability.)

2.3.2. *Motions*

There is very often an equilibrium between two or three forces in the atmosphere, with the other forces being of minor importance. (This is true even though continuously observable changes in the flow pattern can only be explained by nonequilibrium conditions.) In the free atmosphere, pressure gradient forces and Coriolis forces dominate. In the equilibrium condition, flow takes place parallel to the isobars (lines of equal atmospheric pressure). In the northern hemisphere the winds in low-pressure regions rotate anticlockwise, and in high-pressure areas the rotation is clockwise.

Close to the earth's surface, the frictional force is also important.

Two different layers close to the earth's surface can be distinguished. In the lower (Prandtl) layer, which extends to a height of ca. 100 m, the wind velocity increases with altitude, but the wind direction remains constant. In the overlying layer, planetary boundary layer, which can reach an height of 1500 m, the wind velocity increases further and the wind direction changes.

The variation of wind speed with height in the Prandtl layer over a land surface with constant roughness can be expressed by the following equation:

$$\bar{u}(z) = u^*/\varkappa[\ln(z/z_0)] - \psi(z/L)$$

$\bar{u}$ = mean wind speed, z = altitude coordinate, u^* = friction velocity, $\varkappa$ = 0.4, z_0 = roughness height, Ψ = stability function, L = Monin – Obukhov length.

Disregarding the stability function, the flow rate varies with the logarithm of the altitude. At the roughness height ($z = z_0$), the velocity is zero. z_0 depends on the height of the roughness elements. It varies from a few millimeters over ice and water surfaces to several meters over large towns and forests. The friction velocity u^* represents the effect of friction. u^* is ca. $0.1 \times \bar{u}$ (10 m). The stability function represents the change with height as a function of the stability of the layer (see Section Stability). For a stable layer, the function is <0, and for an unstable layer >0.

The flow direction in the Prandtl layer deviates from the direction of the ground flow at the top of the planetary boundary layer. There is a shift to the left, which increases with increasing roughness and stability (the extent of shift being 10 – 40°). At the top of the planetary boundary layer, the wind vector changes into the equilibrium flow resulting from the pressure gradient force and the Coriolis force. That is, with increasing height in the planetary boundary layer, the wind

rotates towards the right and the velocity increases.

Deviations from the ideal conditions (flat surface and constant roughness) can lead to considerable changes in the wind field.

For very large ground roughness values (e.g., in large towns or forests) the wind speed does not become zero at height z_0, but at greater heights. A zero point shift, ($z = z + d_0$) must be introduced. The zero point shift d_0 is somewhat less than the mean height of the obstructions.

The flow is very strongly deviated by topographical obstructions. The deviation depends on the orientation of the topographical structures relative to the ground flow (flow at the top of the planetary boundary layer). In valleys, the flow is channeled in the direction of the axis of the valley, independent of the direction of the ground flow. The channeling effect becomes greater as the depth of the valley increases, but even very wide valleys such as the Rhine valley at Ludwigshafen, Germany (width ca. 30 km, height of sides ca. 300 m) cause channeling of the flow. The wind speed at the base of the valley is lower for larger differences between the ground flow direction and the valley direction. The air preferably flows around individual high and steep mountains, but goes over the top of wide high ridges. In basins, the flow rate is significantly lower, and in open locations on mountains it is higher than over flat terrain.

Single buildings modify the flow considerably. On the side facing the wind, a ground eddy forms with a horizontal axis and a countercurrent flow near ground level. A disturbed region also forms on the downwind side, i.e., an eddy with a horizontal axis and a countercurrent flow at ground level. Further downwind, the flow changes back to the undisturbed state. The distance to which this disturbance extends in the flow direction can be many times the height of the building. Undisturbed flow is reestablished at a distance equal to 100 times the height of the building [23].

In streets, an eddy is produced whose intensity increases as the ground flow component perpendicular to the street increases.

The flow field in cities and industrial areas is extremely complex owing to the complexity of the configurations of the buildings.

Flow fields of this kind cannot be adequately understood by measurements in the field. Wind tunnel tests [23] or mathematical models [24] must be used. An example of the mathematical modeling of the wind field in a city area is shown in Figure 2.

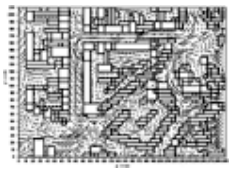

Figure 2. Calculated wind field 1.5 m height in a city area for undisturbed flow at 60°, showing 10-s trajectories

2.3.3. *Stability*

Thermodynamics plays a part in many atmospheric processes. This is very important in the suppression or production of vertical movement. The most spectacular manifestation of this is a thunderstorm, in which intense vertical movements of the air are a characteristic feature. On clear nights, the thin clusters of very long, meandering smoke trails that can be observed owe their existence to a much diminished vertical movement.

In each case, the vertical temperature gradient is decisive. Normally, the decrease in temperature with altitude is 0.7 °C/100 m. On cloudless nights, reversal of the temperature gradient occurs such that the temperature can increase by several degrees per 100 m (inversion). On sunny days and during the summer months the temperature decrease can be > 1 °C/100 m.

These temperature changes in the atmosphere can be compared to the temperature changes experienced by a parcel of air that is ideally thermally isolated from its surroundings, and either rises or falls like a body floating in a liquid. The temperature change of the air parcel can be deduced from the First Law of Thermodynamics, which gives a temperature decrease of 1 °C for a 100 m ascent, and an equal temperature increase if the air parcel sinks by 100 m. In the normal case, the rising air parcel is colder and hence heavier than its surroundings. As it is heavier, it accelerates downwards, back to the starting level. The air parcel that has sunk is warmer than its surroundings and hence lighter, so that it accelerates upwards, again towards the starting level. Thus, vertical movements involving organized flows (e.g., upward moving or “anabatic” winds) or turbulence are damped, i.e., the layer is stable. The layer is especially stable if temperature increases with altitude (inversion).

If the temperature decrease in the atmosphere is > 1 °C/100 m the temperature of the air parcel above the starting level is higher than that of its surroundings. It is therefore lighter and becomes accelerated upwards, away from the starting level. This also holds when an organized or accidental displacement in a downwards direction occurs. Thus, vertical movements are amplified, and the layer is unstable.

The effect of stability or instability on the turbulence of the flow is especially relevant to the dilution of harmful substances in the atmosphere (see Section Air-Pollution Control). With turbulent fluctuations, momentum is also transferred, and this transfer determines the friction of the flow. Hence, the friction is greater for unstable layers, and the wind profile has a different form (see Section Motions).

Measures of stability are, e.g., the Richardson number, the Monin – Obukhov stability length, and stability classes.

The Richardson number *Ri* is defined as [25]:

$$Ri = \frac{g}{\bar{\Theta}} \frac{\partial\bar{\Theta}/\partial z}{(\partial\bar{u}/\partial z)^2}$$

where *g* = gravitational acceleration, $\bar{\Theta}$ = potential temperature.

It is a dimensionless number that represents the ratio of buoyancy to frictional forces.

The Monin – Obukhov stability length *L* is defined as [26]:

$$L = \frac{T}{g} \frac{u_*^3}{\overline{w'T'}x}$$

where T is the temperature of the air in Kelvin, *w′* the turbulent fluctuation of the vertical velocity, *T′* the turbulent fluctuation in temperature, and $\overline{w'T'}$ he turbulent heat transfer.

This number also represents the ratio of the frictional forces to the buoyancy. Its dimension is length.

The Richardson number and the Monin – Obukhov stability length are negative for unstable layers and positive for stable layers.

The formerly much used stability classes (e.g., those due to Klug – Manier [27]) should no longer be employed, as these do not correspond to the present state of the art.

2.4. Air-Pollution Control

Pollutants are always present in the atmosphere, and may be of natural origin (see Section Composition of the Present Atmosphere). The only effective method of reducing additional pollution of human origin is to reduce emissions.

The subject of this section is not the reduction of emissions, but what happens to pollutants after they have been released into the atmosphere, and what can be done to prevent threshold values from being exceeded. The observance of threshold values can be monitored by measuring pollution concentrations. In general, point measurements are taken, so that a result is only applicable to that location. Threshold values are often statistical quantities which must be determined from a number of single measurements. To include all variations of the influential factors (source parameters and meteorological conditions), measurements must be carried out over long periods (one year or more). Only then can it be decided whether threshold values are being exceeded or observed. The measurements provide only a present time and past for each measuring location. To obtain spatial information, and to enable the effects of preventive measures to be forecast (new sources, reduction of emissions, relocation of sources), the extent of the pollution must be calculated.

To obtain statistical information, the concentrations for time series, with the corresponding variations of the input quantities, must be calculated. The calculation is highly time consuming. Frequency distributions of the emissions and the meteorological input quantities are therefore used. The use of several one-dimensional frequency distributions is only possible if the individual parameters are independent of each other. For the most important meteorological input quantities, i.e., wind direction, wind velocity, and stability, this is certainly not the case, so that multidimensional frequency distributions are required.

2.4.1. *Methods of Calculating the Dispersion of Pollution*

There are four methods of calculating pollution concentration:

1. Empirical formulas
2. Gauss formulas
3. Solution of the diffusion equation
4. Particle models

2.4.1.1. *Empirical Formulas*

Empirical formulas give the relationship between measured emissions, immissions, meteorological influences, and geometrical parameters. The simplest formula expresses the fact that the concentration is proportional to the source intensity, is inversely proportional to the wind velocity, and decreases in proportion to a negative index of the distance between the source and the point under consideration. Using this method, cause and effect can be investigated and, within limits, concentration distributions at ground level can be obtained. The method is not transferable, and the formulas must be recalculated if there are changes in the emissions. It is not suitable for forecasting the effects of preventive measures.

2.4.1.2. *Gauss Formulas*

Diffusion experiments using continuous sources at various heights led to the Gauss formula:

$$\bar{C}(x, y, z) = Q/[2\pi\sigma_y(x)\sigma_z(x)\bar{u}] \cdot \exp\left(-\frac{y^2}{2\sigma_y^2(x)}\right) \exp\left(-\frac{z^2}{2\sigma_z^2(x)}\right)$$

where $\bar{C}$ is the concentration (time average), *x*, *y*, *z* are space coordinates (*x* in the wind direction), *Q* is source intensity,

σ_y horizontal spread of concentration perpendicular to the wind direction, σ_z the spread of concentration in the vertical direction, and $\bar{u}$ the mean wind velocity.

Experiments show that the concentration distribution perpendicular to the mean direction of propagation can be represented by Gauss functions [28]. The experiments measure the spreads and show that these depend on the distance from the source, the height, and the stability of the atmosphere.

There are numerous Gauss formulas, all derived from an original formula, which applies to a spreading spontaneous emission that drifts with the wind:

$$C(x,y,z,t) = Q/\left[(2\pi)^{3/2}\sigma_x(t)\sigma_y(t)\sigma_z(t)\right] \cdot \exp\left(-\frac{x^2}{2\sigma_x^2(t)} - \frac{y^2}{2\sigma_y^2(t)} - \frac{z^2}{2\sigma_z^2(t)}\right)$$

t = traveling time of the cloud, $\sigma_{x,y,z}$ = diffusion in the *x*, *y*, *z* directions.

Other formulas can include the effects of influence such as:

1. Height *h* of the source
2. Reflection at the earth's surface, at an inversion, or at a wall
3. Chemical conversions
4. Settling rate of the pollutant
5. Time-dependent emissions
6. Point, line, two-dimensional, and volume sources

Gauss formulas are still extensively used. They are easily understood and simple to use, and the computation is not complicated. However, the conditions under which the Gauss formula can be used are so restrictive (constant meteorological conditions over space and time, flat surface of uniform roughness) that the results calculated can only be used as a guide, and the formula is only suitable for rough calculations. Moreover, better methods are available for calculating concentrations.

2.4.1.3. *Diffusion Equation*

The balance equation for a given substance (diffusion equation) is:

$$\frac{\partial \bar{C}}{\partial t} = -\bar{u}\frac{\partial \bar{C}}{\partial x} - \bar{v}\frac{\partial \bar{C}}{\partial y} - \bar{w}\frac{\partial \bar{C}}{\partial z} + \frac{\partial}{\partial x}\left(K_x\frac{\partial \bar{C}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y\frac{\partial \bar{C}}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z\frac{\partial \bar{C}}{\partial z}\right) + P$$

where time averages are indicated by a bar, K_x, K_y, K_z are exchange coefficients and *P* is the source term.

The change in concentration with time at a point (*x, y, z*) (left-hand side of equation) is determined by:

1. Mass transport to or from the test point by an average wind ($\bar{u}$, $\bar{v}$, $\bar{w}$) at a nonzero concentration gradient (1st to 3rd summand on the right-hand side of the equation).
2. Turbulent transport (4th to 6th summand on the right-hand side of the equation). Mass is transported from the region with an excess to the region with a deficit.
3. Sources and sinks for masses of material.

Analytical solutions to the diffusion equation exist (the Gauss formula is a solution of the diffusion equation given certain preconditions for the relationship between diffusion and distance from the source). However, they cannot be used in practice if meteorological fields vary over space and time, and topographical structures and buildings are to be included. The diffusion equation can then only be solved numerically.

Computation Grid. To obtain a numerical solution of the differential equation, a three-dimensional grid is superimposed on the region in which the concentration calculation is to be carried out. Depending on the mathematical model, the region being studied, and its size, the grid can be rectangular or can follow the topography. The spatial solution depends on the size of the region, the size of the structures to be analyzed, and the speed and storage capacity of the computer. The grid element size lies between 1m and some kilometers. Regions studied can vary from a few meters (e.g., a back yard) to many kilometers (e.g., transport over a large area).

Finite Difference Equation. To obtain a numerical solution, the differential equation must be converted into a difference equation. This gives the following:

$$\frac{\Delta C_{i,j,k,t}}{\Delta t} = \frac{1}{\Delta x}\left\{ u_{i-\frac{1}{2},j,k} C_{i-\frac{1}{2},j,k,t} - u_{i+\frac{1}{2},j,k} C_{i+\frac{1}{2},j,k,t} + K X_{i+\frac{1}{2},j,k} \frac{C_{i+1,j,k,t} - C_{i,j,k,t}}{x_{i+1,j,k} - x_{i,j,k}} - K X_{i-\frac{1}{2},j,k} \frac{C_{i,j,k,t} - C_{i-1,j,k,t}}{x_{i,j,k} - x_{i-1,j,k}} \right\} + \frac{1}{\Delta y}\left\{ v C_{i,j-\frac{1}{2},k,t} - \mathrm{v}_{i,j+\frac{1}{2},k} C_{i,j+\frac{1}{2},k,t} + K Y_{i,j+\frac{1}{2},k} \frac{C_{i,j+1,k,t} - C_{i,j,k,t}}{y_{i,j+1,k} - y_{i,j,k}} - K Y_{i,j-\frac{1}{2},k} \frac{C_{i,j,k,t} - C_{i,j-1,k,t}}{y_{i,j,k} - y_{i,j-1,k}} \right\} + \frac{1}{\Delta z_k}\left\{ w_{i,j,k-\frac{1}{2}} C_{i,j,k-\frac{1}{2},t} - w_{i,j,k+\frac{1}{2}} C_{i,j,k+\frac{1}{2},t} + K Z_{i,j,k+\frac{1}{2}} \frac{C_{i,j,k+1,t} - C_{i,j,k,t}}{z_{i,j,k+1} - z_{i,j,k}} - K Z_{i,j,k-\frac{1}{2}} \frac{C_{i,j,k,t} - C_{i,j,k-1,t}}{z_{i,j,k} - z_{i,j,k-1}} \right\} + \frac{P_{i,j,k,t}}{\Delta x \Delta y \Delta z_k}$$

where i, j, k are the indices of the grid element for the coordinate directions x, y, z and $\Delta C_{i,j,k,t}$ is the concentration change in the grid element with indices i, j, k, in time Δt.

To determine all the summands on the right hand side of the equation, the following are required:

1. The three-dimensional wind field u, v, w
2. The three-dimensional field of the exchange coefficients K_x, K_y, and K_z
3. The three-dimensional concentration field at time $t = 0$
4. The three-dimensional field of the production and dissipation terms $P_{i,j,k,t}$

Also, the boundary conditions and the time interval Δt must be fixed.

The wind field largely determines the concentration field. Over a flat surface of uniform roughness, the wind field is horizontally homogeneous, and only changes with height (see Section Dynamics of the Atmosphere). If the land is topographically organized, the land use diverse, and buildings present, the wind field becomes highly structured and cannot be determined from measurements. A mathematical model must be used to determine the three-dimensional wind field.

Only one equation is necessary to determine the concentration. Wind velocity is a vector, and three components u, v, and w must be calculated to define it (i.e., three differential equations must be solved). Other factors are the pressure-gradient force (which depends on the air pressure), the buoyancy forces determined from the temperature field, and the air density. Hence, the number of variables increases to six, and a coupled system of six equations (4 prognostic and 2 diagnostic) must be solved. The method is used in a number of models. The computation is very sophisticated, and only a limited number of structures can be solved, even using a large computer.

If the extensive information available from wind tunnel experiments and field measurements on wind flows over obstacles is utilized, a wind distribution roughly corresponding to reality can be arrived at. To use this in the propagation calculation, this wind field must simply be made to be free of divergence, which can be realized by the iterative solution of a single differential equation. The calculation time is considerably shorter, and very large regions with many obstacles can be used [29].

Whichever of the two methods is used, a three-dimensional wind field is obtained for the diffusion calculation.

Turbulence Field. The variance in the Gauss formulas corresponds to the exchange coefficients in the diffusion equation, i.e., they represent the dilution of pollutants by the turbulence of the atmosphere. Measurements of the exchange coefficients for the transport of pollutants in the atmosphere are not available. The exchange coefficients cannot be calculated from the variances, as the Gauss formulas used are not solutions of the diffusion equation, and the environmental conditions during the measurements were horizontally homogeneous.

There are theoretical approaches to finding the exchange coefficients for the vertical transport momentum. This method is valid for flat land of uniform roughness, and is established by measurements of the exchange coefficients. It is assumed that the theoretical relationships for any oriented surfaces can be used, and that the exchange coefficients for pollutants can be determined from the exchange coefficients for momentum. The three-dimensional field of exchange coefficients is

then obtained from the wind field and the distance to the solid boundary.

Concentration Field. In general, it is assumed that the concentration field at time $t = 0$ is zero. However, this assumption is not obligatory.

Pollution sources have various spatial structures and behavior patterns with time, depending on their nature. An accident can often be regarded as a point source, and is sometimes an instantaneous emission of material. Emissions from vehicles are linear and vary with time. A sewage treatment plant is a two-dimensional source with an almost constant emission. The spatial structure of the source is fitted to the grid. Thus each discrete source becomes a volume source with at least the size of a grid element. The temporal behavior of the source corresponds to the emissions in the time interval Δt.

Pollution Sinks. Pollutants can disappear by chemical reactions, deposition on the ground, and washout by precipitation; either another substance is formed or the original substance is transferred to the soil or groundwater. The quantity disappearing per unit of volume and time interval is always a certain percentage of the concentration.

Boundary Conditions. Solid boundaries (the earth's surface, walls, and roofs) are impermeable, but material can be deposited there. The side boundaries are permeable, and the upper boundary can be permeable or impermeable (in the case of an inversion). These conditions can be represented by appropriate limits to the flow components and exchange coefficients at the boundaries.

Time Interval. A mathematically stable solution is only obtained if the time interval is sufficiently short. The time interval becomes shorter as the grid element becomes smaller and the wind velocity and exchange coefficient increase. It can vary between 0.5 s and 1 min.

Results. The result of the calculation is the variation with time of the structure of the pollutant cloud. An example is shown in Figure 3, which shows a section at a height of 2.5 m through a calculated pollutant gas cloud 400 s after the onset of emission.

Figure 3. Pollutant cloud from a continuous source, 400 s after start of emission

Horizontal cross section at 2.5 m height, undisturbed flow at 135°

Accuracy. There are a number of sources of error that affect the accuracy of the calculations:

1. Inaccurate emission data
2. Errors in the numerical determination of the wind field
3. Errors in the determination of the exchange coefficients
4. Numerical errors due to the method of computation

As it is not possible to estimate the individual errors, the reliability of the computation must be found by comparing measured with computed figures. Investigations using wind tunnel experiments and field measurements [29] show that the accuracy can be given by the factor 2 or 1/2.

2.4.1.4. *Stationary Solution of the Diffusion Equation*

Continuous sources of constant emission must often be considered. In these cases, only the stationary concentration distribution is of interest. An equation for the determination of this concentration distribution is obtained if the change with time $\partial C/\partial t$ is set at zero in the differential equation. In the stationary case, advective and turbulent flows counterbalance production and dissipation. In this case, too, an analytical solution for normal practical use is not possible, and the differential equation must be converted into a difference equation. The stationary solution is obtained by iteration of the initial field with zero concentration and constant sources and sinks.

Figure 4 shows results for the propagation of car exhaust emissions.

Figure 4. Diffusion of car exhaust emissions

Cross section at height 1.5 m, undisturbed flow at 320°, wind speed at height 10 m 1.7 m/s; traffic only in through streets, 500 vehicles per hour (1000 h^{-1} in upper part of longitudinal street)

2.4.1.5. *Particle Models*

An alternative method for determining concentration fields is by following the paths of a large number of individual particles in a given wind and turbulence field (Lagrange particle model). The use of particle models avoids numerical errors originating from the computing method.

2.4.2. *Use of the Diffusion Calculation*

TA-Luft. For plants requiring a license, a diffusion calculation must be carried out in certain circumstances according to the guidelines of the TA-Luft [30]. A Gauss formula that gives the concentration distribution for a continuous source is used. The concentration at ground level is determined by the intensity of the source, its height (including the effects of the buoyancy of a hot gas if applicable), wind direction, wind speed, and diffusion class. The computation method was developed before 1985, and no longer corresponds to the state of the art. This is especially true for topographically organized land or if land use changes. Some important influences cannot be taken into account, e.g., inversions, chemical changes, and variations in meteorological parameters with location and time. An EC Regulation may possibly supersede TA-Luft.

Safety analysis is compulsory for many plants, especially in the chemical industry [31]. A propagation calculation may be required in the course of such a safety analysis.

Two different questions can arise:

In the case of a flammable gas, a worst-case distance and an average distance from the source at which the gas is still capable of ignition (lower ignition limit) must be stated. With toxic gases, the region in which there is the possibility of a hazard must be established. Here, too, the worst case and the average case are required.

The substances released can be light, density-neutral, or heavy gases, or can be particulate. A gas is assigned to one of these three groups according to its density at the place of release relative to the density of the air. The heavy gases therefore also include cryogenic gases that under normal conditions are not heavier than air and would be grouped with the light gases, and high molecular mass gases at high temperatures are regarded as light.

The spread of *heavy gases* cannot be calculated by the method described above, as in this case propagation is controlled by the force of gravity, and the sudden change in density at the interface with the surrounding air prevents mixing. A heavy gas spreads in a flat layer close to the ground, like a liquid. Obstacles such as fences, walls, houses, and dams can prevent propagation. In streets and on sloping country, channeling takes place, extending propagation.

Analytical or numerical models to determine ignition limits cannot be used, as the effect of obstructions cannot be included. Therefore, VDI guidelines recommend the use of the results of systematic wind tunnel investigations [32]. Dimensional analyses enable the results of small-scale experiments to be transferred to the full scale.

For a continuous source of heavy gas, three characteristic quantities are obtained:

$L_{cc} = (V_0^2/g_e)^{1/5}$ (length scale)
$T_{cc} = (V_0/g_e^3)^{1/5}$ (time scale)
$U_{cc} = (V_0/g_e^2)^{1/5}$ (velocity scale)

where V_0 is the source volume flow in m^3/s in the wind tunnel or in reality, and g_e is the acceleration due to gravity, normalized for the relative density difference.

Wind tunnel experiments show that, in the worst case, the dilution depends only on the normalized distance x/L_{cc}. As the lower limit of flammability is defined by a particular dilution, a constant value is obtained for the normalized distance. If this is multiplied by the characteristic length, which can be determined from the source data, the distance of the lower limit of flammability on the full scale can be found, and it can then be seen whether a safety hazard exists. The VDI guidelines list the results of a large number of experiments with various shapes of obstacles.

In the case of *gases of neutral density*, a modified Gauss model is used in the VDI guidelines [33]. Each emission is divided up into a number of individual clouds emitted sequentially, so that emissions that change with time can also be considered. The worst case is given by:

1. The test point lying on the downwind side of the source
2. A stable temperature stratification
3. A barrier layer at a height of 20 m, or at the height of the upper edge of the highest building

As an average case, the most common wind velocity and diffusion class is used, with no barrier layers.

The variation with time of the concentration at a given test point is calculated. From this, depending on the substance, the relevant parameters, e.g., average value or dose, can be calculated, or the highest value can be found. Point sources, line sources, flat sources, and volume sources can be treated.

Heavy gases are treated like density-neutral gases if the concentration of the heavy gases is < 1 % and the gas can have toxic effects. In this case, the propagation model for heavy gases gives the input values for the diffusion calculation for density-neutral gases.

Light gases have either a relative molecular mass lower than that of the gas mixture air, or the temperature of the gas is significantly higher than that of the surrounding air. Light gases experience a buoyancy. They are accelerated upwards, and the gas only finds its way back to the earth's surface at considerable distances from the source by turbulent diffusion. The maximum concentration is shifted to greater distances. In the guidelines [33], this is accommodated by a displacement of the source in an upwards direction. The height of the source is obtained from the heat emission, and is calculated as

described in the TA-Luft Guidelines [32]. The worst case propagation situation is given by the combination of wind velocity and diffusion class that gives the highest concentration at the test point.

Odor. Factory farming, sewage works, composting plants, waste dumps, and chemical plants can be sources of unpleasant odors. Immissions are directly perceived by people, and can constitute a nuisance. The most important factors determining the degree of the nuisance are frequency, duration, intensity, and nature of the odor, and the opinions of the people affected. Gauss models or solutions of the diffusion equation are not in themselves suitable for calculating odor nuisance, as they give average values for stationary diffusion conditions. The averaging times in the diffusion experiments used to produce the diffusion figures were 30 – 60 min, and the average figures calculated by the Gauss model are therefore also averaged over this time period. In the human nose, the test volume is a few cubic centimeters, and the averaging time is one to a few seconds. Variations in concentration due to atmospheric turbulence lead to large fluctuations in the intensity of the odor.

A VDI guidline [34] enables the frequency of occurrence of the odor of a particular material above a given concentration of the odoriferous substance in the vicinity of single or multiple sources to be determined. This is based on a model that gives average concentrations. It is assumed that these average values are derived from a frequency distribution of concentrations of short averaging time. The frequency distribution can be represented by an exponential function. As well as the probability of detecting an odor, the probability that an odor sensation in excess of the nuisance level will enter the respiratory tract is also obtained.

Accidental Release (⟶ Plant and Process Safety). The main problem with accidental releases of flammable or toxic gases is detecting the accident. Only if an accident is immediately recognized can the appropriate measures be taken to prevent hazards to personnel. The region in which threshold values have been exceeded and the expected duration can be predicted by means of a diffusion calculation. Preconditions for this include:

1. Immediate establishment of the location of the accident and of the temporal behavior of the pollutant source
2. Computational methods significantly more rapid than real time (by a factor of > 10) for calculating the future concentration of the pollutant
3. Criteria for the noxiousness of immission concentrations resulting from the accident

The methods currently used to detect accidental discharges are inadequate, especially where the transport of hazardous goods is involved. In the future, it will be possible to use remote monitoring systems which, when fully developed, will

1. Immediately detect an accident
2. Identify the pollutant
3. Give the spatial distribution of the pollutant concentration
4. Give the changes in the pollutant source with time

Diffusion calculations can already be carried out by personal computers appreciably more rapidly than real time, especially as the accuracy of the diffusion calculation need not be better than the accuracy with which the source term can be determined.

Threshold values for toxic substances are only partially suitable for assessing short-term accidental immissions.

After accidental release a reconstruction of events is often required, or the amount of pollutant material deposited on the ground may have to be determined. An accidental-release calculation can enable the propagation process to be reconstructed, and the emissions can be assessed from the measured immission concentrations. The amount of material deposited and the extent of the region affected can also be determined.

2.4.3. *Environmental Compatibility and Simulation of the Environment*

If the earth is to remain habitable, all human activities must be investigated in terms of whether their consequences are environmentally tolerable. Also, all man-made objects are exposed to the effects of the environment, so that they become contaminated and are eventually destroyed. Consequently, repair, replacement, and even disposal may be necessary, so causing further pollution of the environment.

Attempts can be made to solve the first problem by means of environmental compatibility studies, experiments, and tests [35]. For this, all the effects on the soil, water, and the atmosphere must be taken into account. The effects on the atmosphere can be divided into effects on air quality and effects on the climate.

Environmental simulations aim to provide information on the durability of manufactured objects. Here, too, the effects of air pollution and climatic influences are investigated [36].

Environmental Compatibility. The question of the effects of air pollutants is closely linked with the diffusion calculation. For given emissions, immission concentrations and amounts of material deposited or absorbed can be found. From this, the effects on humans, animals, plants, monuments, and natural features can be derived. Synergistic effects can never be excluded, but assessment of these is difficult. Climatic effects (air temperature, air humidity, rain, sunlight, and wind) on the human population are difficult to quantify. Normally, humans, animals, and plants have a wide tolerance range within which they can live without harm. The lack of specific receptors for the individual components of the climate and the variable susceptibility of people (e.g. due to age and state of health) means that different effects can produce the same stress.

Combinations of meteorological parameters can be regarded as separate entities, and subjective descriptive terms can be applied to them, such as harmful, stress-inducing, causing nuisance, neutral, beneficial, or useful.

The effects of corrective measures on, e.g., air temperature and humidity, ventilation by the wind, and sunlight can be predicted by mathematical models. The combined effects of climate and pollution are sometimes very important, but it is not at present possible to quantify these.

Environmental Simulation. Simulation of the environment aims to predict the functionality and lifetime of products. In environmental simulation experiments, the essential characteristics of the environment must be modeled, and a time-compression technique is necessary. Synergistic effects must be included.

In laboratory experiments, only one parameter is usually investigated, and the time compression can be achieved by increasing the dose. Limits are imposed by the necessity for a linear dose – effect relationship. In field experiments, the magnitude of the causal influences are often not known, and only the effects can be analyzed. Also, forecasts of future events must be made from minimal changes and short exposure times. Numerical environmental simulation is only in the early stages of development. The degree of time compression depends on solely computer capacity. However, there is in particular a lack of essential information about processes occurring at the atmosphere – object interface.

[Top of Page]

3. Effects of Air Pollutants

Thomas Eikmann and Gerhard Scholl

3.1. Effects on Humans

Air pollution has always been an undesirable byproduct of human activities, and has presumably had harmful effects on human health ever since the cave dwellers lit their first fire. The problem increased with the growth of human settlements, the practice of crafts and trades, and the extensive development of large industrial concentrations starting in the mid-1800s. Smog episodes, e.g., in 1930 in the Maas Valley (Belgium), in 1948 in Donora, Pennsylvania, and in 1952 in London attracted the attention of a wide section of the public to the problems of health damage by air pollutants and the consequent increase in the death rate of especially vulnerable sections of the community.

Continuing structural change in post-industrial society, including a decrease in classical industrial operations (e.g., the coal and steel industries) and a simultaneous increase in the importance of the service industries, has led to considerable changes in the release of emissions, with consequent reductions in the extent to which the population is exposed to air pollutants.

Land can no longer be divided up into highly polluted industrial regions and so-called clean air regions according to the amount of pollution, mainly because of the increasing impact of traffic and the reduction in the release of harmful substances from stationary emitters (e.g., power stations, domestic heating, and coking plants). Today, areas of a high traffic density in rural regions can sometimes have higher immission concentrations of pollutants than areas of low traffic density in large cities.

The increasing importance of hitherto unknown emissions has led to increasing effects of these other pollutants on the health of the population. As well as the substances typical of motor vehicles (mainly benzene, nitrogen dioxide, and ozone), persistent chlorinated organic compounds play an ever-increasing role in the exposure of the human population to pollution (including secondary accumulation in foodstuffs).

Furthermore, there has been a significant increase in indoor air pollution. Changes in living conditions (longer periods spent in enclosed spaces), the use of new types of construction materials, and changes in behavior patterns (e.g., inadequate ventilation) have increasingly led to health damage (sick building syndrome).

As a result of this, and also of the general population's continuously increasing awareness of the environment, a tendency towards individual risk assessment has become noticeable. The statistical observation of health hazards is carried out by (epidemiological) methods different from those used to investigate individual risks, so that it has been necessary to develop new techniques of observation and investigation. Examples of this are the provision of out-patient departments specializing in environmental medicine for the clinical examination of patients and the rapid development of environmental medicine as a branch of general medicine.

3.2. General Effects

The effects of air pollutants on humans can range from simple nuisance (e.g., an odor) to serious health damage (e.g., periods of smog in London in 1952 caused over 4000 deaths). A large number of substances cause air pollution, and their concentration in the air varies greatly with time and location, depending on weather conditions (transmission effect) and the various types of emitters involved. Today, their concentrations are in general very low (e.g., compared with those at the workplace), so that an assessment of health damage can only be carried out by considering combination effects in the low ose range. Competing effects from the private actions of individuals (e.g., smoking, medication) and exposure by other routes (e.g., water and food) must also be considered. Especially when long-term (chronic) effects are being investigated, epidemiological studies on representative population groups are carried out, (e.g., investigations of health effects in the “clean air” zone of North Rhine-Wesphalia, Germany).

The respiratory tract, with its large inner surface area, is the main site of damage by air pollutants. However, there can also be effects on the heart and circulatory system, blood formation, the kidneys, the immune and nervous systems, and the skin. Air pollutants relevant to environmental medicine include sulfur dioxide, nitrogen dioxide, ozone, carbon monoxide, dusts, heavy metals, and a wide range of hydrocarbons and chlorinated hydrocarbons. Of special significance are air pollutants that are regarded as potential carcinogens. Important substances of this type include arsenic, benzene, cadmium, diesel motor emissions (DME), polycyclic aromatic hydrocarbons (PAH), and 2,3,7,8-TCDD (dioxin).

3.2.1. *Specific Pollutants*

It is considered that an effect on humans by an environmental pollutant has been caused if reversible or irreversible changes to normal physiological processes are caused in the human organism. The scheme of continuous biochemical activities is changed by the substance in question, and this can be detected and observed. In the case of substances with a reversible effect, it is possible to specify a dose or exposure concentration below which no effect is produced or can be observed. The presence of the substance at concentrations below this threshold can therefore be regarded as harmless to health.

In the case of substances that have an irreversible, carcinogenic, mutagenic, or allergenic effect, it is assumed that an effect threshold does not exist. Since a risk is present even at low concentrations, special efforts must be made to reduce or limit such pollutants.

The specific characteristic effects of a substance are mainly a function of a combination of dose, exposure time, the nature of the effects, and the fundamental mechanism of these effects. In the context of environmental pollutants, “effect” means any change which is brought about by a substance after acute or chronic exposure. Damage to health means reversible or irreversible undesired changes caused by a substance or factor. Toxicity is the ability of a substance or factor to cause such damage, depending on the applied dose and exposure time. Risk is defined as the probability that a given damage occurs in that part of a population exposed to a harmful agent.

The effects of environmental pollutants are assessed by using the results of animal experiments, in vitro experiments on biological material, human case studies, and epidemiological investigations. The derivation of standard values for the protection of human health is mainly based on case studies of humans.

The toxicological basis for such values is the lowest concentration which leads to detectable harmful effects on humans and/or animals (LOAEL = lowest observed adverse effect level) and/or the highest dose at which no known adverse effects have so far been detected in humans or animals (NOAEL = no observed adverse effect level). In accordance with the WHO definition, the LOAEL is the lowest concentration of a substance at which definite detrimental changes to morphology, functionality, growth, development, and life expectancy of the organism concerned is to be expected.

When standard values are derived from the LOAEL, safety and protection factors must be used, including the following:

1. A safety margin in the estimation of a no effect level
2. Interindividual differences (toxicokinetic and toxicodynamic differences)
3. Interspecies differences (toxicokinetic and toxicodynamic differences)
4. The simultaneous effect of several substances, or their combined effects
5. Variation in the sensitivity of persons (special risk groups)
6. Severity of the effects
7. Exposure by other routes (e.g., food contamination)

3.2.2. *Persons Needing Special Protection*

Because of their general sensitivity to the pollutants, children, pregnant women, old and ill persons, and persons with a genetic predisposition must receive special consideration when assessing the effects of (air) pollutants. For example, persons with chronic bronchitis or bronchial hypersensitivity are especially susceptible to substances which attack the respiratory tract or lungs. This is exacerbated in children, because they have higher volume per minute respiration rates per unit body weight than adults, and are more sensitive to infections.

Persons sensitive to (air) pollutants also include allergic persons for whom the frequency and severity of allergic reactions can be intensified by certain nonallergenic foreign substances in the air. Thus, the percentage of children with hay fever is significantly higher in areas of high traffic density than in traffic-free areas, even though the pollutants themselves do not have an allergenic effect [37].

Setting guideline values for the protection of especially sensitive persons (e.g., those with congenital immune deficiency or other special idiosyncrasies) is difficult. Thus it is not possible to specify a dose of an allergen that will have no effect on an allergic person. For these persons, the standard values for pollutants derived for a normal population cannot guarantee freedom from complaints. Here, further measures should be taken in addition to those recommended in the standard regulations.

3.2.3. *Carcinogenicity, Mutagenicity, and Teratogenicity*

Carcinogenicity is the ability of a substance or physical factor to cause cancer in humans or animals (⟶ Carcinogenic Agents). Cancer-risk factors can lead to the uncontrolled and unorganized formation of neoplasms, which include both benign and malignant tumors. Carcinogenic factors can include ionizing and UV radiation, numerous chemical substances, and oncogenic viruses. Also, immune deficiencies, hormonal influences, and inherited predispositions affect the formation of cancer.

The development of malignant tumors takes place in three phases, according to the multi-stage model:

1. Initiation: in general, the sudden appearance of a persistent cell modification that is transmitted to the descendants of the cell. It gives the affected cell the ability to form a tumor when acted upon by a promoter.
2. Promotion: a lengthy process (weeks to years) in which the action of promoters (cancer-risk factors) cause cells modified by initiation to multiply. Promoters by themselves are not sufficient for the production of malignant cells, but increase the risk of progression to a malignant tumor.
3. Progression: increase in autonomy and malignity (manifestation of a tumor).

Processes that intensify the initiating action of a substance are known as cocarcinogenic processes. Syncarcinogenicity

means the cooperation of several carcinogenic substances, leading to the additive or superadditive intensification of the action of the individual substances. Anticarcinogenic processes are those in which cancer formation is inhibited or prevented by the effects of chemical factors.

Humans are exposed to a large number of carcinogenic substances via various environmental media (e.g., water, soil, air, food), and physical and biological factors can also play an important role. The risk of cancer can be obscured by general environmental factors, individual behavior patterns (e.g., smoking, diet, and hobbies), and/or by stress at the workplace. Risk estimates in the United States indicate that the effect of environmental factors in the narrower sense (exposure to pollutants in water, soil, and air) on the overall death rate due to cancer is quite small (ca. 5 %).

Causative factors in cancer death rates in the United States [38], [39], given as percentages of all cancer deaths, are as follows:

General anthropogenic pollution (e.g., indoor air, air, drinking water, soil, abandoned polluted areas, pesticides)	2 %
Geophysical factors (e.g., exposure to sunlight, pollution of indoor air by radon, general radiation)	3 %
Workplace	4 %
Industrial products	<1 %
Tobacco consumption	30 %
Alcohol	3 %
Food	35 %
Food additives	<1 %
Reproductive and sexual behavior	7 %
Medication and medical procedures	1 %
Infections	10 ?
Unknown	?

Although there is clearly a risk of cancer even when the above-mentioned effects are excluded any additional burden from the environment should be minimized. An additional cancer risk of 5 % would mean an additional 22 000 cancer deaths per year in the United States.

The carcinogenicity of a substance is usually expressed as the unit risk (risk coefficient), which indicates the risk of cancer from lifelong exposure, e.g., to 1 µg of a carcinogenic air pollutant per cubic meter of respired air. Generally, the “upper-bound risk value” is given; i.e., there is a probability of 95 % that the actual risk is less than the unit risk. The virtually safe dose (VSD) is the daily dose based on a lifetime's exposure (70 years) that leads to an acceptable theoretical additional risk over and above the background risk.

Mutagenicity is the ability of a substance or physical agent to cause permanent structural changes in the genetic information of cells in vivo and in vitro (⟶ Mutagenic Agents). The changes can involve one or more genes, one or more chromosomes, or sections of chromosomes. Mutations in gametes are transmitted to their descendants. Mutations in body cells remain in the organism in which they were induced.

Teratogenicity is the ability of certain chemical substances, microorganisms, or ionizing radiation to cause nontransmissible malformations, mainly in mammalian embryos (⟶ Teratogenic Agents). According to German chemical legislation on hazard data for substances and preparations, a substance with teratogenic effects has the ability to harm prenatal human life, causing irreversible disturbance to fetal development in the mother's body or irreversible impairment of the post-natal development of the child. In theory, all substances can have a harmful effect in all phases of reproduction. Reproductive toxicology deals with the effects of compounds on the sperm and ova, on conception and implantation, on organ formation, and on fetal maturity. It is not yet known precisely how the formation of defects in embryos occurs. It is thought that teratogenic effects arise through gene damage due to chromosome aberrations, disturbances to mitosis, or inhibited or faulty synthesis of nucleic acids.

3.3. Environmentally Relevant Air Pollutants

3.3.1. *Sulfur Dioxide*

Sulfur dioxide (SO_2) is an irritating gas which can be retained to 90 % in the nasal passages on inhalation owing to its high solubility in water. It can reach the lower respiratory passages if its concentration is very high, if it is adsorbed on dust particles, or if breathing is mainly through the mouth. Dust particles can then catalyze the production of sulfur trioxide (SO_3). SO_2 is absorbed from the respiratory tract into the blood. After biological conversion to sulfate, it is mainly excreted in the urine. Only a few minutes after exposure, the distribution of sulfates by the blood into the tissues and organs and their excretion in the urine can be detected. Radioactively labelled sulfur compounds can be found in the lungs up to one week after exposure.

No systemic toxicity or damage to organs due to SO_2, including mutagenic or carcinogenic effects, have so far been observed. The effects of sulfur dioxide and of the acids formed from it vary in severity, depending on the state of health of the exposed person. Persons already suffering from lung and bronchial diseases must be regarded as high-risk groups. Furthermore, short- and long-term exposure, and the effects of physical stress under certain conditions of exposure require special assessment.

Numerous studies have shown that changes in lung function only occur in healthy subjects at very high SO_2 concentrations, untypical of the environment. However, in the case of respiratory systems already affected by chronically obstructive diseases and asthma, sensitivity to SO_2 can be many times higher. SO_2 concentrations of > 10 000 µg/m^3 can lead to restriction of ciliary motion and slowing or cessation of mucociliary transport, with consequent impairment of lung clearance. Acute effects are bronchoconstriction, bronchitis, and tracheitis. Physical exertion, with increased respiration rate and volume, can significantly lower the threshold concentration of 10 000 µg SO_2/m^3 [40], [41].

Young asthma sufferers exposed to SO_2 by inhalation through a mouthpiece, with simultaneous physical exertion, show noticeable effects on lung function at the low concentration of 100 µg/m^3. However, with asthmatic adults, a concentration of at least 350 µg SO_2/m^3 is necessary. Simultaneous exposure to 300 µg SO_2/m^3, 100 µg NO_2/m^3, and 50 µg ozone/m^3, leads to an increase in bronchial sensitivity in the provocation test with histamine, carbachol, acetylcholine, etc. [42].

In chronic exposure to SO_2, the main effect is an increased incidence of bronchitis and other symptoms of infection. Furthermore, investigations of the effects of exposure to SO_2 at typical environmental concentrations using spirometric tests showed that lung function was affected. An investigation of 200 schoolchildren in an industrial area showed that exposure to increasing SO_2 levels (up to 281 µg SO_2/m^3) and particle concentrations led to impairment of the enforced vital capacity and the 1 s capacity. Also, ca. 25 % of the children studied showed a ca. four-fold sensitivity to pollution [43].

Reduction of the enforced lung function was observed in adult patients with chronic obstructive respiratory ailments on a 5-d exposure to 190 µg SO_2/m^3 (peak figure 720 µg SO_2/m^3) and 44 µg dust particles/m^3. Also there was a significant increase in the observable effects at low air temperatures (< 0 °C). Impairment of lung function (oscillatory resistance) in normal population samples was observed on increasing the 24-h mean value of 60 µg SO_2/m^3 to 256 µg SO_2/m^3 with simultaneous exposure to airborne particles in concentrations of 69 µg/m^3 and 179 µg/m^3 respectively (24-h mean values) [44].

In Japan, using normal population samples, an increase in the incidence of bronchitis in men from 5.5 % to 6.7 % was observed, and in women from 1.1 % to 3.9 % when the sulfur dioxide level increased from 60 µg SO_2/m^3 to 140 µg SO_2 /m^3. The corresponding concentrations of airborne particles were 100 µg/m^3 and 200 µg/m^3, respectively. The smoking behavior of the sample groups investigated were standardized. Investigations on the normal population in Germany showed that when the sulfur dioxide level increased by ca. 50 µg/m^3, a 10 % increase in the incidence of bronchitis occurred [45].

In children, an increase in the long-term chronic exposure to 181 – 275 µg SO_2/m^3 and 230 – 310 µg airborne dust/m^3 led to increased incidence of diseases of the lower respiratory tract compared with children from a control region with lower air pollution. Also, in other studies of children from variously polluted regions, a correlation between the frequency of diseases of the lower respiratory tract and the degree of pollution was found [46].

A series of epidemiological investigations into the effect of SO_2 on health were carried out on chronically exposed population groups in polluted areas during smog episodes in which the concentrations of pollutants (mainly SO_2 and dust) were extremely high for brief periods. Here, it was shown that exposure over several days led to an increase in the death rate of older people with chronic obstructive lung diseases or heart diseases when the mean 24-h concentration of SO_2 reached ca. 700 µg/m^3 and the airborne dust concentrations simultaneously rose to 500 µg/m^3 [46].

During the smog periods in North Rhine-Westphalia in January 1985, a significant increase (8 %) in the death rate occurred in the polluted area. Hospital admissions for respiratory and cardiovascular diseases increased by 17 %, and a marked correlation with age was noticeable, the increase being greater in older people. The highest 24-h mean values in the polluted area were 830 µg SO_2/m^3, and, in a control area, 320 µg SO_2/m^3. The corresponding airborne dust concentrations were 600 µg/m^3 and 190 µg/m^3 [47].

When setting threshold values for SO_2 in the air, it should be noted that in many cases SO_2 immission values were used as indicators of air quality when classifying effects from epidemiological investigations, whereas other pollutants that were also present as immissions in high concentrations were disregarded. However, especially in humans, a combination effect with particulate air pollutants present at the same time must be assumed.

In Germany, the TA-Luft Regulations for sulfur dioxide specify an arithmetic annual mean value of 140 µg/m^3 (IW 1) and a 98 % value of 400 µg/m^3 “for protection against health hazards”. The maximum immission concentrations (MIK values) specified by the VDI Air Pollution Commission are 1000 µg/m^3 for the 30-min value and 300 µg/m^3 for the 24-h mean value. The World Health Organization (WHO) has recommended the following guidelines: 500 µg/m^3 as the 10-min maximum and 250 µg/m^3 as the 1-h maximum. When SO_2 is present with airborne dust (black smoke), the specified 24-h mean value is 125 µg/m^3 for both SO_2 and airborne dust, and the 1-a mean value is 50 µg/m^3 [48-50].

3.3.2. *Nitrogen Oxides*

The nitrogen oxides NO and NO_2 are important air pollutants. Health hazards are almost exclusively associated with NO_2

and its reaction products, while NO acts as a reservoir from which nitrogen dioxide is produced.

Nitrogen dioxide is a strongly irritant gas. On inhalation, the mucous membranes of the respiratory tract are attacked, the gas reacting immediately with the moisture present there and on alveolar surfaces. It probably does not reach the lung capillaries, although lung function is affected. On inhalation, 80 – 90 % of NO_2 is absorbed in the respiratory tract. Absorption possibly occurs by formation of nitrous and nitric acids or their salts, as indicated by the detection of nitrite and nitrate in the urine and blood after NO_2 inhalation.

The odor threshold lies in the range 200 – 440 µg/m^3. Under controlled conditions, effects on the lung function of healthy experimental subjects are detectable above a concentration of 3800 µg NO_2/m^3. In the case of asthma sufferers under physical exertion, effects on lung function occur above 560 µg/m^3. The combined effects of NO_2, SO_2, and ozone at low concentrations increase the sensitivity of the respiratory tract to acetylcholine [40].

In the case of bronchitic patients, a fairly rapid increase in the respiratory tract resistance takes place in the concentration range 3000 –3760 µg NO_2/m^3. Moreover, at higher concentrations, a decrease in the alveolar – arterial oxygen pressure difference, increased sensitivity to acetylcholine, and an increase in the respiratory tract resistance caused by carbachol is observed. The results of epidemiological studies on the relationship between NO_2 exposure and health effects should not be used for determining threshold limit values for the protection of personnel because other pollutants are also usually present [51].

The effects observed in animal experiments are of only limited applicability to humans, as combination effects with other pollutants are generally not considered. Short-term and single exposures of rats in the low concentration range (380 µg NO_2/m^3) led to inhibition of prostaglandin metabolism. At concentrations of ten times this figure, metabolic impairment was of the same order of magnitude, although the recovery time was significantly extended (<70 h). The observed decrease in prostaglandin breakdown can lead to changes in the respiratory tract and disturbance of the blood flow through the lungs [52].

Long-term exposure of experimental animals at 190 – 1900 µg NO_2/m^3 led to thickening of the alveolar capillary membrane, loss of the ciliary epithelium, and collagen formation in untypical areas of the lungs. In rats, the inhalation of 900 µg NO_2/m^3 for 4 h led to the degranulation of fat cells, followed by excretion of histamine [53].

In a series of studies, increased sensitivity towards viral and bacterial infections after chronic exposure to NO_2 has been demonstrated. Prolonged inhalation of 940 µg NO_2/m^3 by mice leads to increased sensitivity towards *Klebsiella pneumoniae*, the magnitude of this effect appearing to depend more on the concentration than on the total dose. On the basis of the results of animal experiments, it can be assumed that humans will also experience weakening of the immune system on exposure to NO_2 or combinations thereof with other air pollutants.

In Germany, the TA-Luft Regulations specify the following levels of nitrogen dioxide “for protection against health hazards”: 80 µg/m^3 as the annual arithmetic mean (IW 1) and 200 µg/m^3 (IW 2) as the 98 % value. The maximum immission concentrations (MIK values) laid down by the VDI commission on air pollution are 200 µg/m^3 as the 30-min mean value and 100 µg/m^3 as the 24-h mean value. The following guidelines are recommended by the World Health Organization (WHO): 400 µg/m^3 as the 1-h mean value and 150 µg/m^3 as the 24-h mean value [48], [50], [54] (see also Chap. Laws and Technical Regulations).

3.3.3. *Ozone*

Ozone (O_3) is a highly reactive gas that can change or decompose enzymes, coenzymes, and proteins by oxidizing thiol groups, and can oxidize fatty acids to toxic fatty acid peroxides. Protein and lipid membranes are the main sites of attack by ozone.

Ozone is mainly absorbed via the respiratory system, where rapid reaction with the tissues occurs. In many animals, acute exposure to high concentrations of the gas (440 – 1000 µg/m^3) causes rapid, shallow breathing, increased resistance of the respiratory system, reduced vital capacity, and bronchial hyperreactivity. Younger animals show higher sensitivity to these concentrations of ozone. An 8 °C increase in the air temperature doubles the toxic effects. Investigations have shown that concentration is more significant than exposure time.

Numerous animal experiments have demonstrated that chronic exposure to concentrations between 200 and 1000 µg O_3/m^3 has a number of harmful effects, including inflammation of the bronchiole region leading to lung emphysema with fibrosis; reduction of the inhalation and exhalation capacity per second; and decreased lung compliance (elasticity). Inhalation capacity per second is the most sensitive indicator of lung function, and a significant and detectable reduction is apparent at 240 µg O_3/m^3. When rats were exposed for 93 d to as little as 120 µg O_3/m^3, an increase in the ketosteroids in the urine and a decrease in the cholinesterase activity occurred. At 200 µg O_3/m^3, biochemical changes in the lung tissue take place, and above 1000 µg/m^3, a decrease in the surfactant concentration can be detected [40], [55], [56].

Both short- and long-term exposure to ozone cause increased mortality rates in animal experiments (60 – 160 µg/m^3). This has been attributed to decreased antibacterial activity of the lung tissue [55], [56].

The odor threshold of ozone for humans is 20 – 40 µg/m^3. Lower concentrations lead to odor acclimatization. Subjective effects such as lachrymation (caused by other substances present with the ozone), irritation of the respiratory system, coughing, headache, and respiratory discomfort are observed above a concentration of ca. 200 µg/m^3. In school children and adults, changes in the lung function parameters (e.g., reduction of enforced exhaled air volume and increase in the resistance of the respiratory system) appear at concentrations between 160 and 300 µg/m^3. In clinical exposure experiments with volunteer subjects, detectable changes to lung function parameters were detected after 6.6-h exposure to 160 µg O_3/m^3 or 1 – 3 h exposure to 240 µg/m^3 combined with physical exertion [55], [56].

Deterioration in physical fitness was observed at 240 – 740 µg O_3/m^3. With maximum physical exertion (competitive sport), a decrease in performance was detected at considerably lower ozone concentrations. Inflammatory reactions of the tissues of the respiratory tract appear above 160 µg/m^3 after 6.6-h exposure with intermittent physical activity. Also, an increase in the incidence of asthma attacks was noted at 240 – 300 µg/m^3 [55], [56].

At ozone concentrations of >1000 µg/m^3, marked changes to lung function occur, with a significant reduction in the diffusion capacity and one-second capacity. The effects of ozone on lung function show considerable individual differences. Approximately 20 – 30 % of the population show pronounced reactions. Also, there is an increase in the number of cells of the immune system, the fragility of the erythrocytes increases, and visual impairment occurs (e.g., reduction in visual acuity and changes in the field of vision).

In general, children show similar changes in lung function to those exhibited by adults after exposure to ozone. Older persons generally react less often than young adults, but their symptoms are often more severe. In epidemiological investigations, a correlation between typical environmental concentrations of ozone and temporary changes to lung function are reported. When respiratory diseases are already present, a deterioration in health can occur. Investigations on children show a connection between levels of the daily maxima of the ozone immission concentrations and impairment of lung function. Such changes were observed above 120 µg/m^3. This exposure concentration is significantly below those used in experimental investigations. The difference is probably due to the simultaneous presence of other pollutants. Symptoms such as irritation of the eyes, nose, and throat; chest pain, coughing; tiredness; and nausea were observed in children subjected to ozone concentrations of >200 µg/m^3. In asthmatic patients, as the severity of the disease increased, a correlation between the frequency of attacks and the ozone concentration was observed. People with obstructive lung diseases and smokers generally do not react to ozone more sensitively than healthy subjects. However, patients with limited lung function show more severe reactions compared to persons not so affected [56].

Carcinogenic effects of ozone have not so far been proved. However, ozone can have effects resembling those caused by radioactivity, and even low ozone concentrations (200 µg/m^3) in cell cultures can cause chromosome rupture resembling the effects of X-rays. Therefore, although there is no indication of increased cancer rates in occupational groups exposed to ozone, a carcinogenic effect of ozone cannot definitely be excluded.

High-risk population groups include all persons frequently exposed to prolonged strenuous physical activity in the open air on hot summer days. These are mainly healthy children and adults who take part in games or sport, and persons involved in strenuous work in the open air.

In Germany, the VDI Air Pollution Commission has set a 30-min mean value of maximum immission concentrations (MIK value) of 120 µg/m^3. Exposures of up to 400 µg/m^3 for a period of up to 30 min should not occur more frequently than once per week. The environmental ministers' conference has specified a maximum figure of 180 µg/m^3. If this is exceeded, a recommendation should be made to the public along these lines: “Persons who have found from experience that they are especially sensitive to ozone (or other pollutants) should not take prolonged strenuous exercise in the open during the afternoon.” The WHO recommends a 1-h figure for ozone of 150 – 200 µg/m^3. To prevent chronic damage to health, an 8-h figure of 100 – 200 µg/m^3 is recommended.

3.3.4. *Carbon Monoxide*

Carbon monoxide (CO) is a colorless, odorless, and tasteless gas, with low solubility in water. It is formed by the incomplete combustion of carbon or carbon-containing compounds.

Carbon monoxide affects humans by combining with hemoglobin (Hb) to form carboxyhemoglobin (COHb). Carbon monoxide blocks the divalent iron center of Hb by displacing oxygen without change of oxidation state. Carbon monoxide forms a bond with hemoglobin that is ca. 240 times as strong as that formed by oxygen. Therefore, even at comparatively very low concentrations it blocks the oxygen-binding site, and prevents oxygen transport by the blood. The bond between CO and hemoglobin is reversible, and when an intact organism breathes CO-free air, all the CO is eventually released and replaced by O_2, and Hb functionality is fully restored. If pure oxygen is breathed, the regeneration process is much faster. As CO is formed endogenously in normal metabolism, unexposed persons have 0.8 % COHb in their blood. On average, smokers have ca. 7 % COHb, and this rises to 15 % COHb for heavy smokers [40].

All the symptoms of carbon monoxide intoxication are due to oxygen deficiency in the tissues and the accumulation of carbon dioxide during normal metabolism. Unlike reduced hemoglobin, COHb cannot carry away the carbon dioxide. The most important effects are on the central nervous system and the cardiovascular system. The symptoms of carbon monoxide intoxication depend primarily on the COHb concentration in the blood, which is in turn affected by a number of factors including the CO concentration in the respired air, the respiration-time-volume, the duration of the exposure, the oxygen demand of the tissues, and the Hb concentration.

The symptoms of carbon monoxide intoxication with increasing COHb content of the blood [58] are as follows:

5 – 10 %	Slight, measurable visual impairment (threshold of fusion frequency)
10 – 20 %	Slight headache, lassitude, malaise, shortness of breath with exertion, palpitations
20 – 30 %	Vertigo, partial loss of consciousness, limpness and paralysis of limbs
30 – 40 %	Pink skin color, loss of consciousness, shallow breathing, circulatory collapse
40 – 60 %	Deep unconsciousness, paralysis, Cheyne-Stokes respiration, decrease in body temperature
60 – 70 %	Death within 10 min to 1 h
>70 %	Death within a few minutes

Within limits, the healthy organism can compensate for oxygen deficiency caused by CO, in the long term by an increase in hemoglobin, and in the short term by increasing the flow of blood through the organs at risk. Neurophysiological investigations and experimental behavior studies show a significant decrease in the ability to concentrate in healthy persons with 2.5 % COHb. This effect can be important in regions of high exposure concentrations (e.g., traffic jams in road tunnels).

In persons with cardiovascular diseases (heart muscle diseases) and sclerotic changes in the blood vessels of the heart and brain, harmful effects appear at relatively low concentrations (2.8 – 3 %). For instance, in 1000 metallurgical workers exposed to CO for at least 4.2 years, a linear correlation between CO dose and angina pectoris symptoms was established. These industrial medical findings were supported by results of experimental investigations. For example, angina pectoris patients, after a 90-min journey in an open vehicle through Los Angeles, showed a significant reduction in the time required for the appearance of typical chest pain compared with the time required after a journey in which synthetic air was inhaled. The COHb level in the patients after the journey in the open vehicle was 5.08 %, and 0.65 % after breathing synthetic air [58], [59].

According to calculations, a 24-h exposure to air containing 10 mg/m^3 CO produces a maximum of 1.5 % COHb, a 30-min exposure to 30 mg/m^3 CO produces 2.3 % COHb, and a 30-min exposure to 50 mg/m^3 CO gives a maximum of 2.8 % COHb [58], [60].

In Germany, the TA-Luft Regulations for carbon monoxide “for protection against health hazards” specify an arithmetical annual mean of 10 mg/m^3 (IW 1), and 98 % figure of 30 mg/m^3 (IW 2). The maximum immission concentrations (MIK values) specified by the VDI commission for air pollution are 50 mg/m^3 for the 30-min figure and 10 mg/m^3 for both the 24-h and annual mean figures. The WHO recommends the following figures: 15 min: 100 mg/m^3, 30 min: 60 mg/m^3, 1 h: 30 mg/m^3, and 8 h: 10 mg/m^3 [48], [50], [61].

3.3.5. *Dusts*

Dusts are defined as dispersions of solid materials in gases. The particles can be of any shape, structure, and density, and can be up to ca. 100 µm in size. They are formed either by entrainment or mechanical processes. Like smoke and fog, dusts are types of aerosol.

The health hazard due to dusts depends on chemical composition, concentration, exposure time, and especially particle size. This last parameter distinguishes dusts from gases and vapors. Dusts are principally absorbed by respiration. The transport and deposition of a dust in the respiratory tract are mainly determined by the behavior of particles in flowing gases. The characteristic parameter is the aerodynamic diameter d_{ae} of a particle. The aerodynamic diameter of a particle of a given shape and density is given by the diameter of a sphere of unit density (1 g/cm^3) which has the same sedimentation rate as that of the particle in still air or air moving with laminar flow. Fine particles of size <10 µm are regarded as airborne dust [62].

Dust fractions are classified according to their particle size as follows: coarse dust: diameter >50 µm, medium-size dust: 10 – 50 µm, fine dust: 0.5 – 10 µm, and very fine dust: <0.5 µm. So-called respirable dust (airborne dust), which consists of particles of size <10 µm, is of relevance to health.

Particulate pollution (dust) is a mixture of many chemical substances that can have a wide range of effects both separately and in combination. They can be of natural or anthropogenic origin. The main constituents of atmospheric dust are sulfates, carbonates, nitrates, halides, silicates, salts and oxides of calcium, iron, potassium, sodium, aluminum, magnesium, and organic substances. The components with toxic potential include heavy metals, arsenic, polycyclic aromatic hydrocarbons (PAH), and chlorinated hydrocarbons.

Depending on their chemical and physical properties and the duration of their effect, dusts can cause morphological damage and changes to the function of the lungs or to their susceptibility to infection. The particular part of the respiratory tract affected by dusts depends mainly on their particle size. Particles of >10 µm diameter only reach the nose and throat area, and do not enter the middle to lower regions. They can also be carried into the digestive tract by the lung clearance mechanism. The tracheobronchial dust is the dust fraction deposited in the area of the mucociliary clearance apparatus of the tracheobronchial system, and this too can be carried into the digestive tract. Alveolar dust is the dust fraction deposited in the alveolae and in the region of the bronchioles without mucociliary clearance. Via the tracheobronchial system, it can appear both in the digestive tract and in the interstitial tissue of the lungs.

The effectiveness of dust transport mainly depends on the presence of an intact bronchial mucous membrane. The

transport rate can be adversely affected by medication or toxic substances. Effects of gaseous substances such as SO_2 and NO_2 on ciliary activity and mucociliary clearance have also been observed.

Dusts can contribute to the acute effects of smog and can also cause chronic effects. Acute dust pollution always gives rise to health effects that can be unambiguously detected in experimental and epidemiological investigations. The most important first acute effects of dust are impairment of the clearance mechanism of the respiratory system and irritation of the bronchial mucous membrane. The lowest dust concentrations, derived from periods of smog in London and New York, at which an effect on death rate due to disturbances of respiratory function can be detected are in the region of 0.15 mg/m^3. The first changes to the lung function of children were detected at an airborne dust concentration of 0.18 mg/m^3 in the middle of the day [40], [63].

Longer exposure leads to cytotoxic effects and tissue damage. The first effects are observed at an annual average concentration of airborne dust of > 0.09 – 0.15 mg/m^3. More recent industrial medical investigations have shown that dust at the workplace can promote chronic bronchitis, although cigarette smoke has a significantly greater influence. At today's low dust concentrations in the outside atmosphere and also indoors, it may prove very difficult to detect long-term effects on the incidence of chronic bronchitis by epidemiological methods against the background of active and passive smoking.

In Germany, the TA-Luft Regulations for airborne dust (irrespective of its constituents) specify an arithmetic annual average of 0.15 mg/m^3 (IW 1) and a 98 % figure of 0.30 mg/m^3 (IW 2) “for protection against health hazards”. For “protection against significant detriment and nuisance”, the following figures for dust deposits (nonhazardous dusts) are prescribed: 0.35 g m^{-2} d^{-1} (IW 1) and 0.65 g m^{-2} d^{-1} (IW 2). The VDI commission for air pollution recommends the following maximum immission concentrations (MIK values): 1-a average: 0.100 mg/m^3, 24-h average: 0.250 mg/m^3, 3-h average: 0.500 mg/m^3. The following standard figures for simultaneous exposure to sulfur dioxide and airborne dust are laid down by the WHO: total suspended particulates (24-h figure): 0.120 mg/m^3, thoracic particles (deposition characteristics 50 % < 10 µm, 24-h figure): 0.070 mg/m^3 [48], [50], [62].

3.3.6. *Lead*

Lead is one of the most extensively investigated pollutants both in the environment and the workplace. There have been many investigations into the occurrence of lead in the various environmental media (air, water, soil, and food) and into the toxicity of lead to humans and animals. These are documented in the literature in a large number of comprehensive reports.

In the general population of Germany, the most important source of lead pollution is food. The daily intake has been estimated by the registry and advisory center for environmental chemicals of the German Health Office (ZEBS). The figure for men is 147.2 µg/d (1.03 mg/week), and for women 104.22 µg/d (0.73 mg/week). Of this, ca. 5 – 10 % is absorbed in the digestive tract by adults, while small children have an higher absorption rate amounting to ca. 50 % [64].

Of the total lead absorbed by adults (in air, food, and drinking water), the fraction absorbed by inhalation is ca. 17 – 25 %. Because of the higher lead absorption rate in the digestive systems, this figure is much lower for children (2 – 3 %).

Lead absorption by the human organism depends on the solubility of the lead compounds and the composition of the food. Deficiencies of calcium and vitamin D increase the absorption. The amount of inhaled lead-containing dust taken up by the organism is determined by the deposition, elimination, and absorption of the inhaled dust particles. It is assumed that 90 % of inhaled lead is absorbed.

The absorbed lead first enters the blood, and is distributed among the various organs and tissues. Approximately 90 % of the lead in the blood is bound to components of the erythrocytes. Lead inhibits various enzymes and thus affects several stages of hemoglobin synthesis. Several precursors of hemoglobin appear in the blood in increasing amounts, and are excreted in the urine (e.g., δ-aminolevulinic acid and coproporphinogen-III). Effects of lead on the erythrocyte membrane and reduction of the lifetimes of erythrocytes have also been described.

Lead is mainly accumulated in the bones. It can pass through the placenta and into the fetal blood circulatory system. Lead can also pass through the blood – brain barrier to a small extent, especially in children, in whom this barrier is not fully formed. The excretion of lead takes place via the feces and kidneys, a smaller fraction being eliminated via the hair and nails.

Acute intoxication is rare, as most lead compounds are sparingly soluble. Symptoms such as vomiting, colic, constipation, and acute kidney failure occur. Sometimes, lead encephalopathy is diagnosed, which leads to vomiting, apathy, vertigo, coordination disturbances, hyperactivity, and other effects.

The best known effects of chronic lead intoxication include weakness, loss of appetite, sleeplessness, nervousness, constipation, headache, muscular and joint pains, shivering, encephalopathy, and colic. Objective signs include pallor, weight loss, anemia, increased excretion of δ-aminolevulinic acid in the urine, increased blood lead level, decreased level of δ-aminolevulinic acid dehydrase in the blood, and changes in peripheral nerve conduction velocity.

Many investigations into damage to the central and peripheral nervous system caused by exposure to lead in the course of employment or from the environment have been carried out. There is a clear correlation between increased exposure to lead in the prenatal period and/or in early childhood and neuropsychological effects such as hyperactivity, fine motor disturbances, and reduction in intelligence and learning ability.

In the typical exposure situation of the general population, the following observed effects can be correlated in quantitative terms with the lead concentration in whole blood (internal lead burden). At lead concentrations of > 10 µg/100 mL in the

blood of children and > 15 µg/100 mL in adults, the first effects that appear are changes in the biosynthesis of hemoglobin by inhibition of certain enzymes. At blood concentrations of > 30 µg/100 mL, there is a statistically significant increase in arterial blood pressure. In children with blood lead concentrations of > 20 µg/100 mL, statistically significant impairment of individual functions of the CNS can be detected in neuropsychological tests. Impairment of fetal and infant development becomes statistically significant at concentrations of >10 to 15 µg/100 mL [65], [66].

Lead concentrations in the blood of the general population in Germany lie between 6 and 7 µg/100 mL. Lead concentrations of > 15 µg/100 mL are observed in only <1 % of the population. The WHO requires that 98 % of the normal population of adults and children in Europe should have a lead concentration in the blood of < 20 µg/100 mL [67].

There is no evidence of carcinogenic or cocarcinogenic effects of lead on humans, although the appearance of cancer in animals is observed at high doses. Mutagenic effects (chromosome aberrations) can occur with lead intoxication. Apart from its toxic effect on the fetus, there is no evidence that lead has teratogenic properties. In Germany, the TA-Luft Regulations for lead and inorganic lead compounds as components of airborne dusts “for protection against health hazards” specify an arithmetic annual average of 2.0 µg/m^3 (IW 1). For “protection against significant detriment and nuisance”, the figure 0.25 mg m^{-2} d^{-1} is specified as the IW 1 value for lead and inorganic lead compounds. The VDI commission has prescribed the following maximum immission concentrations (MIK values): 1-a average: 1.5 µg/m^3, 24-h average: 3.0 µg/m^3. The WHO recommends a standard range of 0.5 – 1.0 µg/m^3 [48], [50], [68].

3.3.7. *Cadmium*

The heavy metal cadmium is a nonessential trace element. It is absorbed orally and by inhalation, skin absorption being practically insignificant. Absorption of cadmium by the German population is mainly through food. According to information provided by the registry and evaluation center for environmental chemicals in the German Health Office (ZEBS) for the year 1984, the average cadmium intake via food by men was 34 600 ng/d (0.242 mg/week) and by women 26 900 ng/d (0.189 mg/week). The amount taken in via respired air is small, and is estimated at ca. 40 – 50 ng/day. Smoking 20 cigarettes per day results in a cadmium absorption of ca. 360 – 2200 µg/a. In food cadmium is mainly chemically bound to proteins [64].

The effects of cadmium depend on the quantity of cadmium absorbed into the blood. The absorption rate by the stomach and intestines is usually ca. 5 %, but this can be higher in cases of iron-deficiency syndromes. In contrast, cadmium entering the alveolar region of the lungs by inhalation is almost completely absorbed.

Cadmium has a marked tendency to accumulate in the human organism, mainly in the kidneys and liver. Thus, cadmium concentrations in the liver and kidneys increase with increasing age, and also in the blood and the urine. The most important factor is cigarette smoking. Cadmium concentrations in the blood of smokers are 3 – 5 times as high as those for nonsmokers, and cadmium concentrations in the renal cortex of smokers are ca. twice those for nonsmokers.

The health effects of increased chronic cadmium exposure in the workplace and the environment consist mainly of disturbances to kidney function, increase in arterial blood pressure, and the production of bronchial carcinomas.

The toxic effect of cadmium on the kidneys is mainly a result of the increase in the cadmium concentration in the kidney tissue. An early symptom is the increased excretion of low molecular mass proteins in the urine due to a disturbance to reabsorption in the proximal tubular section (tubular proteinuria). The diagnostic criterion is the increased excretion of β_2-microglobulin and retinol-binding protein. As well as tubular proteinuria, cadmium-induced glomerular proteinurias also occur, characterized by the excretion of high-molecular proteins. This symptom is very probably due to the deposition of immune complexes in the blood capillaries of the glomeruli [40].

If cadmium levels in the renal cortex exceed 200 µg/g, the first indications of a tubular proteinuria can appear in susceptible persons. At concentrations of > 250 µg Cd/g tissue, the appearance of a tubular proteinuria can be expected in 50 % of exposed persons.

Animal experiments have shown that the cardiovascular system is also sensitive to cadmium. Increase in blood pressure and myopathic heart damage can occur without any sign of damage to the kidneys. There are no corresponding results for humans at present. Lung and liver damage, anemia, and changes to the mineral balance of the bones (itai-itai disease) are also known.

There is no evidence of teratogenic effects of cadmium on humans, probably due to the low permeability of the placenta to cadmium. Information on the mutagenic effects of industrial or environmental exposure is contradictory, although genotoxic changes to human and animal cells can be caused by cadmium salts.

It is considered certain that a number of cadmium compounds show carcinogenic potential after inhalation. The German MAK Commission classifies cadmium, cadmium chloride, cadmium oxide, cadmium sulfate, cadmium sulfide, and other compounds that could enter the human organism (in the form of respirable dusts/aerosols) as materials with a definite carcinogenic effect (Group II A 2). Orally ingested cadmium, however, has no carcinogenic effect [69].

The German TA-Luft Regulations for cadmium and inorganic cadmium compounds as constituents of airborne dust specify an arithmetic annual average of 0.04 µg Cd/m^3 (IW 1) “for protection against health hazards”. For “protection against serious detriment and nuisance”, the figure for cadmium and inorganic cadmium compounds is 5 µg m^{-2} d^{-1} (IW 1). The WHO recommends that in urban and industrial areas, cadmium concentrations of 0.01 to 0.02 µg/m^3 can be tolerated. However, any increase in existing concentrations should be avoided so as to minimize the entry of cadmium into the food chain via deposition of dust. The Länder committee for immission protection (LAI) gives a “unit risk” for cadmium of 1.2×10^{-2} and an orientation figure of 1.7 ng/m^3. This figure represents the upper limit of background concentration in

rural areas in Germany, and is regarded as a standard for future immission reductions in urban areas [48], [50], [70].

3.3.8. *Arsenic*

Arsenic is a metalloid which is ubiquitous in the earth's crust in various forms. Toxicologically, inorganic arsenic compounds, organic arsenic compounds, and arsenic hydride (arsine) must be considered separately.

Arsenic is mainly ingested orally by humans in food and drinking water, but is also inhaled. The extent of absorption depends on the type of arsenic compound. Arsenic can also be absorbed through the skin. Organic arsenic compounds are ingested mainly during the consumption of fish, mussels, and crustaceans, the main compounds concerned being arsenobetaine, arsenocholine, and trimethylarsonium lactate and its derivatives.

Arsenic is rapidly transported within the body by the blood. It becomes distributed in human tissue in varying concentrations, being more concentrated in the skin, hair and nails, the lungs, the bones, and the brain. Inorganic arsenic compounds in the human body are largely methylated to dimethylarsinic acid, and to a lesser extent to methylarsonic acid. These methylated compounds are excreted more rapidly in the urine, and are less toxic. The excretion of arsenic is mainly performed by the kidneys, ca. 75 % of the arsenic in the urine being present as dimethylarsinic acid and a lower percentage as monomethylarsonic acid and inorganic arsenic.

The toxic effect of arsenic is mainly due to its ability to form covalent bonds with thiol groups. Various enzymes, such as monooxidase, urease, glucose oxidase, choline oxidase, can be inhibited by arsenic. Acute effects, chiefly on the gastrointestinal tract, are only likely with high doses. A dose of 70 – 180 mg arsenic trioxide is fatal for humans. The main symptoms of acute intoxication are difficulties in swallowing, abdominal pain, nausea, vomiting, chest pains, and watery diarrhea. Inorganic arsenic compounds, especially the trivalent oxides and chlorides, cause irritation of lungs and skin.

Chronic arsenic intoxication leads to disturbances to the peripheral and central nervous system, and lesions to skin and mucous membranes. Skin lesions include irregular pigmentation of the skin (melanism and leucoderma), hyperkeratosis, the appearance of diagonally striped white fingernails (Mees line), and warts. Damage to the mucous membranes mainly takes the form of conjunctivitis, nasopharyngeal catarrh, perforations of the nasal septum, bronchitis, diarrhea and constipation, and stomach cramps.

Effects on the PNS take the form of paraesthesia, motor disfunction, and a decrease in nerve conduction velocity. Lassitude and apathy, and more rarely encephalopathy indicate damage to the CNS. The optic nerves and the organ of balance can also be affected. The liver can become inflamed, and morphological changes such as cirrhosis and carcinoma can occur. Other effects related to exposure to arsenic have been observed, including damage to the heart (ECG changes), disturbance of red corpuscle production, (anemia and agranulocytosis), allergic contact dermatitis, and peripheral vascular disorders, leading to gangrene (black foot disease).

Indications of long-term occupational or environmental exposure include hematological, dermatological, and neurological symptoms. The critical threshold dose for a daily intake of arsenic is stated by various authors to be 400 µg/d. Inhalation of arsenic can lead to increased incidence of cancer of lung and skin, liver carcinoma, and tumors of the brain and nervous system. Oral ingestion mainly leads to skin carcinoma [40].

Arsenic has a teratogenic effect, as it can penetrate the placenta of both animals and humans.

The carcinogenic effect of inorganic arsenic on the human skin, lungs, and occasionally the liver has been demonstrated by epidemiological studies. It is explained by the inhibition of repair processes. Possibly, the phosphate groups of nucleotides are replaced by arsenate, as the two ions are of similar structure. Organic arsenic compounds do not appear to have a carcinogenic effect. The available data to some extent show dose – effect correlations for lung cancer (bronchial carcinoma) due to inhalation and for skin cancer due to oral ingestion.

Immission limit values are not specified for arsenic either in the TA-Luft Regulations or in the Air Quality Guidelines of the WHO. The Länder Committee for Immission Protection (LAI) gives a “unit risk” for arsenic of 4×10^{-3} and an orientation value of 5 ng/m^3. The latter figure represents the upper limit for the background concentration in rural areas of Germany, and can be used as a measure for future immission reductions in densely populated areas [70].

Inorganic arsenic compounds are essentially negative in point mutation tests, but have been demonstrated to be effective inducers of chromosome aberrations and sister chromatid exchanges. Chromosome changes were also observed in the lymphocytes of therapeutically or occupationally exposed persons.

3.3.9. *Benzene*

Benzene is the simplest aromatic hydrocarbon and is used as a solvent for a wide range of substances. It occurs in nature only in very low concentrations, e.g., as a constituent of crude oil (max. 0.4 g/L). It can also be formed by the incomplete combustion of organic compounds. In Germany, the main source of the benzene to which the population is exposed in the outdoor air is traffic.

The amount of benzene absorbed by the population of densely populated areas can be 200 µg/day, and can reach 600 µg/day in regions close to benzene emitters. Benzene absorption can also occur through smoking (20 cigarettes per day corresponds to a maximum of 600 µg), and can sometimes be due to food sold in gas stations (up to 5 µg/d) [71].

Prolonged inhalation of benzene leads to ca. 50 % absorption, oral ingestion probably to 100 %, and skin contact to <0.2 %. It is distributed throughout the organism, and becomes concentrated in fat-containing organs. Retained benzene is eliminated unchanged to a very variable extent by respiration (depending on the exposure concentration and individual circumstances) or is oxidatively metabolized and excreted with the urine.

The carcinogenic effect results from the metabolization of benzene to benzene epoxide by the monooxigenases in the endoplasmatic reticulum. Benzene epoxide can be converted nonenzymatically to phenol, but can also be metabolized by enzymatic action to catechol. Other metabolites include hydroquinone and hydroxyhydroquinone, benzyl glycol, and *trans-trans*-muconic acid. Also, a reaction of benzene epoxide with glutathion mediated by glutathion-*S*-transferase can occur. This thioether, after further metabolization, is then excreted as phenylmercaptic acid in the urine. Benzene epoxide can also react nonenzymatically with nucleic acids (to nucleic acid adducts, e.g., *N*-7-phenylguanine). In both humans and animals, the principal metabolite of benzene is phenol, which is mainly excreted in the urine in conjugated form. The ultimate toxic and carcinogenic metabolites of benzene are the quinones and hemiquinones [72].

Like most solvents, benzene has a depressant effect on the CNS. Local irritant effects on the skin and mucous membranes can occur. The most important effect of benzene is damage to the blood-forming system, especially on long-term exposure. The classical symptoms of acute and chronic benzene intoxication are generally only observed at very high or medium concentrations and usually after prolonged exposure. The immission concentrations normally present in the environment are insufficient to cause these symptoms. It is unanimously accepted that chronic exposure to benzene can lead to a very rapidly accelerating depression of all bone marrow functions, which then gives rise to changes in the peripheral blood picture. The various bone marrow functions can be damaged either simultaneously, in succession, or in isolation.

There is some disagreement over the question whether damage to bone marrow functions caused by chronic exposure to benzene is a precondition for the appearance of leukemia. The more recent view is that, in addition to the frequently described acute myeloid leukemia in epidemiological studies, there are also indications of an increased occurrence of chronic myeloid and chronic lymphatic leukemias, and a dose-dependent increase in lymphatic and hematopoietic neoplasms.

In vitro and in vivo tests have given no indication of point mutations due to benzene. However, numerical and structural chromosome aberrations were observed in exposed humans and in animal experiments. Carcinogenic effects in rodents (leukemias, lymphomas, and an increase in the various types of tumor and tumor locations) could be found in long-term inhalation experiments. However, high concentrations were required.

Immission limit values are not prescribed either by the TA-Luft Regulations or the Air Quality Guidelines of the WHO. Epidemiological studies gave unit risk estimates of 6×10^{-6} to 20×10^{-6}, with the median figure of 9.2×10^{-6}. The LAI gives 9×10^{-6} as the unit risk, and an orientation figure of 2.5 µg/m^3. This figure represents the upper limit of the background concentration in rural areas of Germany, and is the standard for future immission reductions in densely populated areas [70].

3.3.10. *Polycyclic Aromatic Hydrocarbons*

Polycyclic aromatic hydrocarbons (PAH) are formed pyrolytically in all incomplete combustion processes and in the low-temperature carbonization of organic materials. The relative concentrations of the many types of PAH (PAH profiles) mainly depend on the extent of heating. When different organic materials are pyrolyzed, the profiles obtained at a given temperature are very similar. Thus, as a general rule for the assessment of the carcinogenic hazard due to PAH in the open air, it is assumed that, as in the ideal case, the PAH profiles are identical, and that the concentration of an individual PAH can be used as a measure of the carcinogenic potential of the whole group. Therefore, the assessment of PAH concentrations in the air is usually based on the parameter benz[*a*]pyrene (BaP), as the most experience is with this substance. However, the BaP concentration in cigarette smoke or diesel engine emissions is not suitable for estimating the carcinogenic potential of the PAH in these media.

Compounds containing at least four rings are particularly relevant to the environment, as a number of these have carcinogenic effects. PAH have been detected in many vegetable and animal foodstuffs. Particularly high BaP concentrations are found in roasted or grilled products. Various vegetables such as kale, lettuce, carrots, spinach, and some types of grain sometimes contain high concentrations of PAH.

The absorption of PAH takes place mainly by the oral route (in food), while absorption by inhalation is almost exclusively associated with particles. The daily amount of BaP ingested in food in Germany is thought to be 1000 – 10 000 ng per capita. The daily intake by inhalation is estimated to be 10 – 200 ng per capita. Ingestion via drinking water is negligibly small, and cutaneous absorption in general plays no role. The amount of BaP ingested orally is certainly greater than that absorbed by inhalation in terms of quantity, but the intake of particles by inhalation is of considerably greater significance because of the formation of bronchial carcinomas [70].

The PAH compounds adhering to the particles are partly dissolved in the lungs and metabolized there. The metabolites can act on the cells, but can also be taken into the blood and detoxified in the liver, or can be excreted unchanged, especially in the bile, via the gut. Only a very small proportion of the PAH compounds is excreted with the urine in either a metabolized or an unchanged state.

PAH compounds have an inducing effect on microsomal enzymes. Little is known of acute or chronic toxic effects of the most important PAH compounds. Clinical signs of intoxication first appear after comparatively high doses. For example, newborn mice after a subcutaneous injection of 30 000 ng dibenz[*a, h*]anthracene (= 15 mg/kg body weight) showed no acute effects. This shows that the carcinogenic effect predominates over other effects [73].

The mutagenic effects of some PAH compounds have been detected in many test systems. There is a good correlation between the results of in vitro mutagenicity tests and the carcinogenic effect. BaP has a toxic effect on rat and mouse embryos, and a reduction in the fertility of mice has been reported.

The carcinogenic effects (mainly skin and lung carcinomas in persons exposed to chimney soot and coal tar) have long been known, and the mechanism of the carcinogenic action of BaP has been well researched. In general, the most widely

accepted view is that it is not the substance itself that is active, but the reactive epoxides formed by microsomal monooxigenases. The epoxides are converted by epoxidases to *trans*-dihydrodiols, which are the final carcinogenic metabolites. In laboratory animals, the carcinogenic effect is independent of the mode of application, and systemic tumors were also observed in other tissues.

Immission limit values are not specified for PAH compounds or BaP either in the TA-Luft Regulations or the WHO Air Quality Guidelines. From the risk estimates based on epidemiological studies, a unit risk of 7×10^{-2} can be derived for BaP. The LAI specifies an orientation value of 1.3 ng/m^3 for BaP. This figure represents the upper limit of the background concentration in rural areas of Germany, and is a standard for future immission reductions in densely populated areas [70].

3.3.11. *Diesel Motor Emissions*

Diesel motor emissions (DME, diesel engine exhaust) consist of a large number (>1000) of gaseous and particulate substances. DME particles, due to their small diameter (median diameter: 0.2 – 0.3 µm), mostly reach the alveoli, but are then to the extent of ca. 50 % expelled in the exhaled air. The mass of diesel soot particles absorbed by inhalation by the general population can only be estimated at present. Exposure in the open air in rural areas is up to 1.5 µg/m^3, in densely populated areas 5 – 10 µg/m^3, and in regions close to the emitters 15 – 25 µg/m^3 [70].

Apart from the particles, which carry adsorbed organic substances, the main toxic components of DME are nitrogen oxides and aldehydes. The acute toxicity of DME is insignificant. In animal experiments, the inhalation of DME with high particle concentrations led to a reduction in lung clearance. In inhalation experiments, exposure of rats to DME caused benign and malignant lung tumors after exposures that were ca. 75 – 100 times as high as those in densely populated areas in the middle of the week.

The carcinogenic effect is clearly linked to the particle phase. The very low level of polycyclic aromatic hydrocarbons (PAH) in DME can account for at most 1 % of the carcinogenic effect.

Particles and particle extracts from DME exhibit a mutagenic effect in several in vitro and in vivo test systems. Although this demonstrated a genotoxic mechanism, other mechanisms are also under discussion.

Epidemiological investigations on persons heavily exposed to DME in the course of their employment have so far provided “limited evidence” for carcinogenic effects of DME on humans, because insufficient attention was paid to extraneous effects.

DME is classified as a substance definitely known to cause cancer (MAK III A 2), and is classified with the Group II cancer-producing substances (“serious hazard”) in the hazardous substance regulations. Based on inhalation experiments on rats, a unit risk of 7×10^{-5} has been derived [70].

3.3.12. *Dioxins and Furans*

Polychlorinated dibenzodioxins (PCDD) and polychlorinated dibenzofurans (PCDF) are tricyclic, almost planar, aromatic compounds with similar physical, chemical, and biological properties. Depending on the degree of chlorination a number of isomeric compounds (congeners) can exist. The theoretical possible number of congeners is 75 for PCDD, and 135 for PCDF. None of the congeners are produced industrially, except as analytical standards. PCDD and PCDF are formed only as undesirable byproducts of chemical and thermal processes, and enter the environment from many sources. Their persistence, buildup in the food chain, ubiquitous occurrence, and the marked toxicity of some congeners of both groups can constitute a health hazard for humans. They are present as impurities at very low concentrations in many chemicals (e.g., pentachlorophenol) [40].

Adequate data are unavailable for most of the PCDD and PCDF congeners of toxicological significance. The most biologically active compound of the PCDD group is 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD or “Seveso” dioxin). This does not occur in isolation, but can be detected in PCDD/PCDF mixtures in a wide variety of compositions. The toxic potential of PCDD/PCDF mixtures is estimated as TCDD equivalent figures (TEq), derived partially from subchronic toxicity studies and from certain biochemical properties. TEq figures represent the relative toxicity of a mixture in terms of the toxic effect of 2,3,7,8-TCDD. The scientific basis for the TCDD equivalence factors (TEF) is that there is a common effect mechanism, and there is therefore a strict correlation between the binding of the various PCDD/PCDF compounds to the cytoplasmic arylhydrocarbon (Ah) receptors and the biological effects resulting from this. The following properties of PCDD and PCDF based on animal experiments were taken into account [74]:

1. The binding affinity to the Ah receptors
2. The ability to induce various enzyme systems
3. The toxic effects on certain cell cultures (e.g., keratinization)
4. The acute and chronic toxicity

The strongest counterargument is that the isomers so compared with each other differ strongly in their toxicological behavior, and this cannot be adequately represented by the TEF, and TEq figures are not suitable for estimating the potential cancer risk of exposure to PCDD and PCDF. Any statements made about the possible risk to the population from other effects must be made with considerable reservations, especially concerning long-term exposure to very low concentrations.

In the environment PCDD and PCDF are almost exclusively associated with fine dust, and hence can be absorbed by inhalation. However, the amount absorbed by inhalation is less than 5 % of the total. An inhaled volume of 20 m^3/d corresponds to an inhaled dose of 0.4 – 6 pg TEq per person per day. With 50 % deposition in the lungs and 100 % absorption, this represents 0.2 – 3 pg TEq absorbed per person per day [74].

The highest proportion of the daily intake of PCDD and PCDF by the general population comes from foodstuffs of animal origin. The average daily intake by an adult is estimated by BGA/UBA (1993) to be ca. 2 pg TEq/kg body weight. Bearing in mind variations in eating habits, an average intake by 4 – 9 year old children of 2 – 3 pg TEq/kg body weight can be derived. For one year old children, an average daily intake of ca. 4 pg TEq/kg body weight is calculated. Gastrointestinal absorption from foodstuffs is almost 100 %. Intake with drinking water is small [75].

2,3,7,8-TCDD is rapidly distributed in the organism, and causes systemic damage to various organs; it can cross the placental barrier. 2,3,7,8-TCDD affects lipid metabolism, vitamin A balance, and thyroid gland function. Some of these disturbances are due to binding to the Ah receptors and the resulting induction of various enzyme systems. Various glucuronide and sulfate conjugates have been shown to be metabolites. The metabolism takes place through P450-dependent enzymes. There is a high percentage of nonmetabolized 2,3,7,8-TCDD. The PCDD compounds are mainly eliminated in the feces. Dioxins accumulate in the fatty tissue and are mobilized in the milk of nursing mothers. Animal experiments and human experience have shown that excretion decreases with increasing degree of chlorination. The mechanism of excretion in humans is unknown. The half-life of PCDD in humans is 5 – 10 years, but only a few weeks in rats [76].

Acute and subchronic intoxication by TCDD-containing PCDD mixtures can be recognized from a large number of organ manifestations. The skin and liver are particularly affected by disturbance of the porphyrin and fat metabolism, and by the induction of mixed function oxidases. The PNS, CNS and immune system can also be damaged. Direct contact can cause lesions of the skin. TCDD has strong acute toxicity compared with PCB. Acute exposure is first followed by irritation of the eyes and the upper respiratory tract, sometimes accompanied by vertigo, nausea, and vomiting. In later stages, reddening and swelling of the affected parts of the skin can occur. Typical skin lesions known as “chloracne” develop. The most affected organ is the liver. Swelling and liver parenchyma damage occurs, leading to an increase in the transaminases in the blood. Neurological damage following TCDD exposure seems to occur more often than has generally been assumed up to now [74], [76].

Animal experiments indicate fetotoxic and teratogenic effects, and impairment of reproduction by TCDD is also known. Following maternotoxic doses, teratogenic effects are observed in mice, but fetotoxic effects in rats and monkeys. At present, there is no indication of teratogenic effects in primates. Animal experiments also indicate immunotoxic effects such as thymic atrophy and suppression of cell-mediated humoral immune responses. TCDD does not appear to be mutagenic in the Ames test. Absence of mutagenicity is also shown in many other in vivo and in vitro tests, the binding of TCDD to DNA being very weak. In exposed industrial workers, no indication of chromosomal damage has been found either.

The carcinogenic effect of TCDD following oral and parenteral application in animal experiments has been thoroughly investigated and neoplasms at multiple sites have been recorded. Carcinogenic effects of other PCDD or PCDF compounds are presently unknown.

Persons exposed in Vietnam to TCDD-containing herbicides or to TCDD in Seveso were not found to have any increase in chromosome damage. Comprehensive epidemiological studies in the United States, Sweden, Canada, and New Zealand did not provide any solid evidence for carcinogenic effects from the use of herbicides containing TCDD or PCDD as impurities. However, an increased incidence of stomach carcinomas was reported in exposed workers. More recent investigations of this group revealed a slight increase in the standardized mortality rate, but no statistically significant increase in the incidence of tumors. Case control studies of workers who had worked with TCDD-containing herbicides for some years revealed an increased incidence of rare soft tissue sarcomas (hemangiosarcomas). The apparent absence of a mutagenic effect, the lack of DNA adduct formation, and the tumor-promoting activity in two-stage models in liver and skin indicate that TCDD must be regarded as a nongenotoxic carcinogen. This allows the conclusion that a threshold of carcinogenicity exists, below which promotion is not maintained [74-77].

In view of the lack of any indication of a direct genotoxic effect, and assuming a threshold concentration for the promoting effect, a temporarily tolerable daily uptake has been derived by national and international authorities. A figure of 1 pg TEq/kg body weight has been proposed for Germany [74].

If, however, contrary to the present state of knowledge, it is assumed as a precautionary measure that there is no effect threshold, linear extrapolation from animal experiments with oral application of 2,3,7,8-TCDD yields a putative unit risk of 1.4 for inhalative exposure. Combination effects with other PCDD compounds were not included.

Immission limit values for PCDD/PCDF have neither been specified in the TA-Luft Regulations nor in the WHO Air Quality Guidelines. The LAI gives an orientation figure of 16 fg/m^3 for 2,3,7,8-TCDD. This figure represents the upper limit of the background concentration in rural areas of the Federal Republic of Germany, and is the standard for future immission reductions in densely populated areas [70].

3.4. Effects on Plants [78-81]

The effects of air pollutants on plants are of interest both from the point of view of their economic exploitation and because of their great ecological significance.

Before they act on plants, air pollutants are subject to the effects of a large number of fortuitous and nonfortuitous factors (weather, etc.). The plants themselves have defense mechanisms that can influence the progress and severity of the action of a pollutant. They react towards air pollutants with a specific resistance behavior of genetic origin, which is also affected by the plants' state of development at the time and by the external growth conditions. Since plants are open systems, they are highly sensitive towards foreign substances present in the air.

3.4.1. *Relationship between Immission and Effect*

A complex structure of relationships exists between immission and effect. The reaction of the plants depends primarily on

the presence of one or more immission components. The type and extent of the effect are essentially determined by the genetic structure, the growth stage of the plant at the time of exposure, and the modifying influence of various external factors. Depending on the severity and circumstances of the pollution, the stage of growth, and environmental conditions such as soil, weather or climate, and biotic effects, individual plants can suffer acute, chronic, or insidious damage. Plant communities react with changes in their composition and structure, up to complete destruction. The plants' economical and ecological viability can be affected, and eventually gene erosion by extinction of the species can occur.

Acute damage is mainly caused by the short-term effects of immissions at higher concentrations. This is manifested first on plant leaves by loss of turgescence and subsequent drying of the affected parts of the tissue. In general, a few days later, ivory colored, brown, red-brown, or blackish leaf necroses occur. Repeated effects of this kind weaken the vitality of plants. Severe, acute damage can lead to their complete destruction.

Chronic damage occurs after prolonged effects of low concentrations of immissions. Typical external signs include yellowing, or reddish-brown, turgescent marks on the leaves. In the course of continuing exposure, the symptoms of damage become more severe, causing decreased vitality, restriction of growth, and finally the death of the plant.

In addition to acute and chronic damage, erosive damage is also observed. This is produced by the action of acid fog or water-soluble deposits of reactive dusts. In these cases, the high ion concentrations that can be formed on moist leaf surfaces are very important.

Indirect damage can occur when air pollutants accumulate in soil penetrated by roots and change its properties as growth substrate (e.g., pH value). Damage can also be caused by absorption of accumulated substances through the roots.

Finally, immissions can cause a general reduction in the resistance of the plants towards plant diseases (e.g., from fungi and insects) and frost.

The solubility, reactivity, and state (gaseous or particulate) of the substance are the main properties that determine whether a pollutant can affect the plant surface on contact. The spectrum of activity is broad. Not all species and varieties of plants react in the same way to a harmful immission component. Also, some harmful substances produce specific reactions in certain plants. This peculiarity in the spectrum of effects on vegetation has been recognized from early times, and used to detect various air pollutants [82].

Effect Criteria. The large number of possible effects of air pollutants can be characterized by the corresponding effect criteria. These range from accumulation of harmful substances in the plants, disturbance of the physiological and biochemical processes in the cell, and submicroscopic changes to the cell organelles, to morphological changes to organs and impairment of the entire organism. On the level of a plant community, effects on structure and function, possibly including complete destruction, must be mentioned. Other effect criteria include reduced vitality, reduced yields, loss of quality, increased susceptibility to abiotic and biotic influences, and death of the entire plant. An overview of the criteria for the detection of the effects of immissions on plants is given in Table 4 [83].

Table 4. Criteria for determining the effects of immissions on plants [83]

Criterion	Level of investigation
Species spectrum	ecosystem
Abundance of species	biocenosis
	population
Growth form	population
Necroses	organism
Chloroses	organ
Leaf fall	
Biomass	
Accumulation of materials	
Photosynthesis	organism
Gas metabolism	organ
Water balance	
Nutrient status	
Cell damage	tissue
Chloroplast damage	cells
Tannin losses	
Tissue infiltration	
Enzyme activity	organism
Metabolism	tissue
(protein, amino acids,	cells
carbohydrates, pigment	organelles
status, energy balance,	macromolecules
buffering capacity,	

mineral balance)

Proof of Effects. In addition to the measurement of emissions and immissions, the use of plants as information systems for the protection of the environment from air pollution is indispensable. To obtain a complete survey, it is essential that every effect should be determined on both higher and lower plants. However, in practice, this requirement can only be met to a certain extent. As a rule, relevant effects on the most representative plants possible are chosen [84]. Representative plants are those for which the effects are broadly comparable with those on other plants, or for which an effect can be predicted from another effect on one and the same plant. Since the many factors cause a wide variety of harmful effects, the causal relationship between air pollution and effects often cannot be established with certainty in field investigations. For the qualitative and quantitative determination of immission effects, standardized methods whose results enable reliable conclusions to be drawn have now been in use for some years [85-88].

Dose – Effect Relationship. Greenhouse and open land experiments under defined boundary conditions and observations in open land enable the relationship between immission concentration and effect to be described. As absorption of the pollutants usually precedes their effect, an attempt is also made to determine the relationship between these two parameters rather than a relationship between immission concentration and effect. The amount of pollutant taken up from the immission is defined as the immission dose (milligrams pollutant taken up per kilograms acceptor substance or milligrams pollutant taken up per squared meter acceptor surface) and as the immission rate (milligrams pollutant taken up per kilograms acceptor substance per unit time, or milligrams pollutant taken up per squared meter acceptor surface per unit time) (see VDI 2309, Sheet 1). The use of the terms immission dose and immission rate is subject to the limitation that the analytically determined concentration in an acceptor is always a function of the immission flow (see VDI Guideline 2450, Sheet 1), as well as being the result of absorption by other routes, and processes of growth, destruction, excretion, and redistribution in the organism.

Immission dose, immission rate, and immission concentration are known as immission values. The relationships determined between one of these values and its effects on plants is the basis for the establishment of maximum immission values by the VDI commission for air pollution. The aim and significance of these values, which give the decision maker the essential information for fixing limit values, are described in the VDI Guideline 2310, Sheet 1 [89].
Concentration/dose – effect relationships determined in gas exposure experiments on test plants can only be translated to open land relationships to a limited extent, as in climatic chambers, greenhouses, or other closed experimental arrangements, the changing immission concentrations of an open location and the varying meteorological growth factors such as light, temperature, air humidity, and wind can only be approximated. Under such “artificial” conditions, plants can react towards air pollutants more sensitively than in open land conditions [90]. Observations under natural open land conditions or in open land experiments also enable only approximate conclusions to be drawn about the relationship between exposure to a harmful substance and its effect. Furthermore, a generally valid derivation of the relationship between immission concentrations and effect cannot be obtained with certainty from observations in open land or from the results of open land experiments. This is because in the open air very variable air pollution occurs, several immission components normally being present. This leads to effects that are not comparable to specific damage by a single pollutant.

Open top chambers occupy a position somewhere between open land and climatic cabinets [91]. These large chambers, which are open at the top, are placed in the open air. The dose – effect relationship is determined experimentally by introducing air from outside to the test plants. In practice, the immission values are deduced by careful investigation of the results of one or all three methods. Responsibility for determining immission values lies with various working groups of the German Association of Engineers (VDI). The maximum immission values obtained are not limit values, but standard figures at which the risk of damage to vegetation can be limited.

3.4.2. *Effect of Individual Pollutants*

3.4.2.1. *Sulfur Dioxide*

Many publications deal with the effect of sulfur dioxide on plants. Important knowledge was obtained in the 1930s, especially in the United States and Canada. A bibliography by the VDI for 1974 – 1978 lists almost 250 references [92].

Uptake and Plant Reaction. Sulfur is an essential nutrient for plants. It is taken up via the roots from the soil in the form of SO_4^{2-} ions. However, plants can also use the SO_2 from air pollutants to make up for sulfur deficiency.

To affect the leaf [93], SO_2 must penetrate via the stomata into the leaf tissue and cells. This leads to the formation of sulfite and hydrogensulfite ions as dissolved products. A significant part of the absorbed SO_2 is oxidized to sulfate. In general, it is known that damage can consist of reduced activity of various enzymes or increased peroxidase activity and breakdown of chlorophyll. However, these reactions are not specific to the effects of SO_2, and can also result from the action of ozone or phytopathogenic fungi. Reactions specific to SO_2 include the breakdown of tryptophan or indolylacetic acid and the rupture of S – S bonds in proteins.

Photosynthesis reacts to the effects of SO_2 with high sensitivity. However, any retardation of this process can be compensated to a certain extent by an increase in the concentration of CO_2.

Detoxification reactions include neutralization of the acid sulfite and oxidation to sulfate and the formation of organic sulfur compounds. This explains why the concentration or, more accurately, the SO_2 uptake per unit time determines the extent of damage. The physiological activity of a plant is therefore a decisive parameter for the risk of acute damage. This is at its greatest in the spring and summer months, and especially in the morning hours. Chronic damage, mainly nonspecific in character, is caused by the gradual accumulation of sulfur until the damage threshold is exceeded.

Plant species vary in their sensitivity towards SO_2. In general, plants are classified as very sensitive, sensitive, and less sensitive. However, even the most resistant plant species are not totally resistant to SO_2. The resistance behavior of various economically important cultivated plants is given in Table 5 [94].

Table 5. Resistance of plants to SO_2

Very sensitive plants	Sensitive plants	Less sensitive plants
Deciduous trees and bushes		
Walnut (*Juglans regia*)	Lime (*Tilia spec.*)	Plane (*Platanus spec.*)
Red currant (*Ribes rubrum*)	Red beech (*Fagus silvatica*)	Oak (*Quercus spec.*)
Gooseberry (*Ribes uva-crispa*)	Hornbeam (*Carpinus betulus*)	Poplar (*Populus spec.*)
	Apple (*Malus domestica*)	Maple (*Acer spec.*)
	Hazelnut (*Carylus avellana*)	Alder (*Alnus spec.*)
		Lilac (*Syringa vulgaris*)
		Woad (*Salix spec.*)
		Robinia (*Robinia pseudoacacia*)
		Birch (*Betula spec.*)
		Cherry, plum (*Prunus spec.*)
		Vine (*Vitis vinifera*)
		Rhododendron (*Rhododendron spec.*)
Conifers		
Fir (*Abies spec.*)	Pine (*Pinus silvestris*)	Black pine (*Pinus nigra austriaca*)
Spruce (*Picea spec.*)	Weymouth pine (*Pinus strobus*)	Arbor vitae (*Thuja spec.*)
Douglas fir (*Pseudotsuga menziesii*)	Larch (*Larix sepc.*)	Yew (*Taxus baccata*)
		Juniper (*Juniperus spec.*)
		Japanese Cypress (*Chamaecyparis spec.*)
Agricultural and horticultural plants		
Clover (*Trifolium spec.*)	Barley (*Hordeum vulgare*)	Potato (*Solanum tuberosum*)
Lucerne (*Medicago sativa*)	Oats (*Avena sativa*)	Tomato (*Lycopersicon esculentum*)
Lupin (*Lupinus spec.*)	Rye (*Secale cereale*)	Maize (*Zea mays*)
Vetch (*Vicia sativa*)	Wheat (*Triticum sativum*)	Turnips (*Beta vulgaris ssp.*)
Horse bean (*Vicia faba*)	Lettuce (*Lactuca sativa*)	Cabbage (*Brassica oleracea ssp.*)
Spinach (*Spinacia oleracea*)	Bean (*Phaseolus spec.*)	Carrot (*Daucus carota*)
Pea (*Pisum sativum*)	Rape (*Brassica napus*)	Strawberry (*Fragaria chiloensis*)
		Garlic (*Allium spec.*)
Ornamental plants		
Begonia (*Begonia spec.*)	Dahlia (*Dahlia variabilis*)	Rose (*Rosa spec.*)
Vetch (*Lathyrus odorata*)	Geranium (*Pelargonium zonale*)	Gladiolus (*Gladiolus spec.*)
	Petunia (*Petunia hybrida*)	Tulip (*Tulipa gesneriana*)
	Carnation (*Dianthus caryophyllus*)	Iris (*Iris spec.*)
	Fuchsia (*Fuchsia hybrida*)	Narcissus (*Narcissus pseudonarcissus*)
		Chrysanthemum (*Chrysanthemum leucanthemum*)
		Heather (*Calluna vulgaris*)

This classification is carried out by using as effect criteria visible damage and impairment of physiological vitality, adversely affecting the economic exploitation of the plant. Results of open land observations and experiments in combination with continuous and noncontinuous SO_2 measurements in the air yield the essential information for the establishment of maximum immission parameters for the protection of vegetation [94]. In risk assessment from epidemiological data on the effects of SO_2, the presence of other pollutants is not taken into account. The composition of the pollutants, their mixing ratio in the air, and also hitherto unrecognized factors can give rise to synergistic or additive

effects on plants. Various findings concerning the combined effects of SO_2 with NO_2, HCl, HF, ozone, and heavy metals are reported in the literature.

Relationship Between Concentration/Dose and Effect. Investigations into exposure to SO_2 under experimental conditions have shown that most plants show considerably greater compatibility than in open land conditions [94]. In the last decade, short-term high SO_2 concentrations have in general been limited to accidental releases, so that there is now more interest in the harmful effects of SO_2 on long-term exposure to low concentrations. Investigations into tolerance thresholds carried out in various countries give figures of 0.04 mg/m^3 for very sensitive plants, >0.08 mg/m^3 for sensitive plants, and up to 0.12 mg/m^3 for less sensitive plants.

There are many publications on the subject of combined effects, especially of SO_2 and ozone. It is often difficult to link visible symptoms of damage with the pollutant causing them. The nature of the effect is also dependent on the degree of exposure to the pollutant. A combination of SO_2 and ozone in a concentration at which each individual gas would cause serious leaf damage often has an antagonistic effect. At lower concentrations, synergistic effects predominate [95]. The effect of combining SO_2 with other gases, according to available knowledge, can be antagonistic, additive, or synergistic. The effect of a combination of substances depends on the ratio of the concentrations of the two components and their absolute concentrations [92].

Threshold values for SO_2 for the protection of vegetation have not yet been established. The VDI Guideline 2310, Sheet 2, of Aug. 1978, in which the maximum immission values for SO_2 for the protection of vegetation are specified has not so far been finalized.

3.4.2.2. *Hydrogen Fluoride*

As early as the 1800s, hydrogen fluoride was known to be a pollutant harmful to plants. It has the strongest phytotoxic action of all gases. The gaseous compounds SiF_4 and H_2SiF_6 are as harmful as HF, while particulate fluorine compounds are less so. Among the many publications on the effect of fluoride immissions on plants. There is a comprehensive review of the subject in VDI Guideline 2310, Sheet 3 [96].

Plant Uptake and Plant Reaction. Plants can take up fluoride both from the air and from the soil. The fluoride absorbed from the soil is carried to the parts of the plant above ground by transpiration. Transport via the xylem and plasmatic transport causes fluoride to be deposited preferentially in the tips or edges of the leaves.

Absorption of gaseous fluorine compounds from the air takes place mainly through the stomata of the leaves. Particulate immissions containing water-soluble fluorides can cause local erosive damage, mostly of the epidermis, on plant surfaces covered with dew. The absorption of gaseous fluorine compounds through the stomata is influenced by the width of the openings, which depends on, e.g., light, temperature, atmospheric humidity, and the state of development of the plant. Damage to the plants takes place by direct action, which is concentration-dependent, and indirectly by accumulation in the leaf, especially in the tips and edges. In these parts of the tissue, the amount of fluoride deposited is often many times higher than in the other parts of the plant. Because of the marked concentration effects, plants are also damaged by low concentrations if they are exposed to the pollutant for prolonged periods. Increased fluorine contents are usually a sign of the effects of fluorine compounds, and are therefore an important diagnostic criterion.

Variable sensitivity towards fluoride immissions can occur with different plant species, varieties, and origins. However, absolute resistance does not exist. The species-specific resistance differences are utilized to build up resistant series and resistant groups. For practical protection against immissions, it is convenient to divide plant species into the groups “very sensitive”, “sensitive”, and “less sensitive” [96]. The very sensitive species include spruce, vines, apricots, onions, gladioli, and other bulbs, and corms. The group of sensitive plants includes pines, firs, larch, beech, walnut, cherry, lucerne, grain, fodder barley, strawberries, and caryophyllaceous plants. The less sensitive plants include yew, black pines, maple, oak, birch, pear, potatoes, tomatoes, turnips and beet, geraniums, and ornamental compositae. However, the various varieties and subspecies within a group (e.g., the gladioli) can differ greatly in sensitivity among themselves.

Relationship between Concentration/Dose and Effect. Extensive documentation of the relationship between HF concentrations in the air and the effects on higher plants, as a basis for hazard predictions, is provided in the VDI Guideline 2310, Sheet 3. This contains results of experiments on exposure to gases, and data from open land observations. The relationship between immission and effect does not generally obey the quantitative law of erosion, which states that equal concentrations/doses cause equal effects. A demonstration of the difference in reaction between individual species was provided by gladioli and ponderosa pines that were exposed for 2 weeks in the vicinity of an aluminum works [97]. At an average HF concentration of 0.49 ppb, the gladioli showed a damage index (ratio of necrotized leaf tips to total length × 100) of 25.4, compared with 32.7 for the pines exposed to a concentration of 2.36 ppb (i.e., 5 times as high). The damage to the pines increased progressively with the concentration. The difference in reaction may essentially be caused by differences in the secondary deposition of the absorbed fluorine at the leaf edges and tips, and also by an effect on the diffusion resistance of the leaves.

The combination of fluorine compounds with other gaseous air pollutants can lead to increased, decreased, or unchanged fluorine accumulation in plants. At present, there is no consistent model that can account for the combined effects of HF with other gaseous components. However, it is possible to make qualitative predictions for certain boundary conditions [96].

Maximum Immission Values. The necessary preconditions for the derivation of maximum immission values for hydrogen fluoride in combination with other pollutants have not been fulfilled. Hence, the VDI Guideline 2310, Sheet 3, of December 1989, only gives results of gas exposure experiments and open land investigations in which the amount of harmful

substance was defined by the concentration and duration of its effect. Based on protection of the species and not of the individual, the maximum immission values for hydrogen fluoride are as follows (in µg/m^3):

Very sensitive plants	1.0 [a]	0.3 [b]	0.2 [c]
Sensitive plants	2.0	0.6	0.4
Less sensitive plants	7.5	2.5	1.2

[a] Exposure time: 1 d

[b] Exposure time: 1 months

[c] Exposure time: 7 months (vegetation period)

3.4.2.3. *Hydrogen Chloride and Chlorine*

The harmful effects of hydrogen chloride on plants are comparable to those of SO_2. In gas exposure experiments, it was found that the lower limit for damage to sensitive plants after prolonged exposure lies between 0.1 and 0.3 mg/m^3.

The results of HCl gas exposure experiments cannot be applied directly to open land conditions, as the atmosphere often contains large quantities of neutral chlorides (e.g., salt) in the form of aerosols as well as HCl. These cannot be distinguished from HCl by measurement techniques. Salt aerosols, e. g., in coastal regions, can cause plant damage, but their effect is small compared with that of HCl. Laboratory gas exposure experiments have shown that salt aerosols have less than 10 % of the harmful properties of HCl aerosols or HCl gas [98].

The VDI Guideline 2310, Sheet 4, gives the following limit values (mg/m^3) based on gas exposure experiments lasting 24 h and 1 month respectively:

Very sensitive plants	0.80	0.10
Sensitive plants	1.20	0.15

Very sensitive plants include vines, spruce, horse beans, and berberis, while sensitive plants include red beech, pine, oats, and primulas, and the less sensitive plants include oak, blue spruce, kale, and fuchsia.

Chlorine is as harmful as hydrogen chloride, but is only of relatively small significance for vegetation because of its very rare occurrence.

Several species of pine showed needle necroses after exposure to 3 mg Cl_2/m^3 for 3 h [98]. Very sensitive herbaceous plants, including lucerne, tobacco, and radish were damaged after exposure to as little as 0.3 mg/m^3 for 2 h.

3.4.2.4. *Nitrogen Dioxide and Ammonia*

Nitrogen dioxide is the most harmful of the oxides of nitrogen, but it is less phytotoxic than other acidic gases. Nevertheless, prolonged exposure of sensitive plants to >0.4 mg NO_2/m^3 leads to growth retardation and signs of chlorosis. In gas exposures of several hours, concentrations of 10 – 20 times this figure are necessary to cause any damage.

The VDI Guideline 2310, Sheet 5, for sensitive plants such as maple, spruce, wheat, and fuchsia, gives a limit value of 0.35 mg/m^3 as the mean figure for the vegetation period, and 6 mg/m^3 for a single exposure of 30 min. For very sensitive plants such as apple, birch, larch, oats, and rose, slight damage is possible at these concentrations. Less sensitive plants include beech, yew, brassicas, and gladioli.

Ammonia is absorbed through the stomata of the leaves, and can be used for protein synthesis in the cells provided no cell damage is caused at the concentration present. Damage caused by ammonia differs from that caused by acid gases in that the leaves and needles first turn brownish-red, later dark brown, and finally black. Relatively little is known about the dose – effect relationship. Gas exposure experiments by the LAI showed that the first signs of damage to cress, a sensitive plant, were visible after exposure to 1 mg NH_3/m^3 for 2 – 3 days [99].

3.4.2.5. *Hydrogen Sulfide*

Hydrogen sulfide was formerly thought to be a gas that causes relatively slight plant damage, plants in general being regarded as much more resistant to H_2S than to SO_2 [78 p. 99]. Following more recent investigations, this conclusion must be amended. Exposure of spinach and peas to 1.2 mg H_2S/m^3 led to an increase in transpiration which caused drooping of the leaves, often followed by the formation of necroses [101]. Also, the efficiency of photosynthesis can be reduced, leading to a reduced formation of biomass.

H_2S has been reported to be even more harmful than SO_2 [102]. In a gas exposure test lasting 14 d, damage occurred even at concentrations of 0.45 mg H_2S/m^3 in various garden plants (radishes, dwarf beans, and cress) comparable to that

produced by SO_2 at 0.70 mg/m^3. Only red clover was more sensitive to SO_2 than to H_2S. However, it can be assumed that H_2S damage to vegetation only occurs on accidental release from industrial installations.

3.4.2.6. *Ethylene*

Ethylene is produced in the plant itself. It acts as a ripening hormone, and at the same time automatically stimulates increased ethylene production. Anthropogenic ethylene concentrations are also effective in extremely low concentrations. This differentiates it fundamentally from the other harmful gases. Necroses hardly ever occur. Instead, inhibition of the lengthwise and surface growth of leaves occurs, abnormal ripening such as the curling of younger leaves and the downwards curling of the leaf stalks (epinastia), dying off of end buds, premature aging, yellowing, and leaf fall.

The following results are selected from the many gas exposure investigations: 0.1 mg C_2H_4/m^3 caused a 30 % reduction in the dry substance of dwarf beans and radishes after 14-d exposure. The blooms of ornamental plants lost their color at the same concentration after the second day. The first reactions can occur at a concentration of little more than 0.01 mg/m^3. As well as the above plant species, peach, clover, tomatoes, peas, potatoes, and roses are sensitive to ethylene. However, relatively insensitive plants include the monocotyledonous varieties, with the exception of orchids and the conifers, which are very sensitive.

The flat dose – effect curves are striking. An effect appears immediately at very low concentrations, but increases only very slowly with increasing concentration. This makes it very difficult to establish immission threshold values [99]. A comprehensive study of the effect of ethylene and other hydrocarbons on higher land plants was carried out by the Deutsche Gesellschaft für Mineralölwissenschaft und Kohlechemie (German Association for Mineral Oil Science and Carbon Chemistry) [100].

3.4.2.7. *Ozone and Other Photooxidants*

Ozone has a strong phytotoxic effect on plants. It is a component of photochemical smog, being an intermediate product of photooxidation which eventually leads to the production of the strongly phytotoxic peroxyacetyl nitrate as well as other smog components. The photochemical oxidants are taken up exclusively by direct absorption from the air. In the absorption of ozone, it has been found that the stomatic resistance is a decisive factor, and determines the severity of the effect of the gas in each individual case [103]. The action of ozone leads to a change in the permeability of the cell wall, which leads to an increased loss of the dissolved substances in the cell. The efficiency of photosynthesis is reduced by ozone to a measurable extent, often before the appearance of visible damage. Plants react to ozone with varying sensitivity.

It is estimated that in the United States ca. 90 % of plant damage is due to photooxidants. The direct economic damage is mainly due to the effects on growth, yield, and quality. Plant products used as foods and feeds and ornamental plants show reduced yields, changes in the composition of the useful plant substances, and lower perceptible quality. Because photooxidants are always in gaseous form, damage to ornamental plants is also widespread in the United States [106].

Considerable damage is also reported in The Netherlands. The occurrence of damaging ozone concentrations in Germany was detected in many urban areas with the aid of highly sensitive indicator plants (e.g., the tobacco species Bel-W-3 or the small nettle). Also, the contribution of ozone to the so-called new forest damage is generally acknowledged. The phenomenon of extensive forest damage observed in the mid-1970s is linked to air pollution, e.g., by ozone, and is supraregional. A large range of effects appear over a narrow concentration range. Higher concentrations are found especially in the areas of Central Europe remote from sources of pollution and at higher altitudes. Moreover, ozone pollution has considerably increased [107].

The concentration – effect relationship is the subject of many investigations. In VDI Guideline 2310, Sheet 6 [95], air quality criteria for the protection of vegetation were derived from results from open land, climatic chambers, and greenhouses. According to this, plants react with varying sensitivity to the action of O_3. The more sensitive plants include apricot, walnut, Weymouth pine, lucerne, oats, lettuce, and tomato. The less sensitive plants include red beech, red oak, Norway maple, mustard, turnips, strawberries, celery, and cabbage lettuce.

Combination effects with other harmful substances are known in which the effects of ozone are intensified by the presence of, e.g., SO_2 and NO_2. However, in the present state of knowledge, it is not yet possible to quantify the effect of SO_2 and NO_2 on the overall effect.

The maximum immission concentrations for the protection of vegetation (see VDI Guideline 2310, Sheet 6) apply to O_3 as a single component acting for a brief period, i.e., 0.5 – 8 h. In general, very sensitive, sensitive, and less sensitive plants are protected up to concentrations of 320 µg/m^3, 480 µg/m^3, and 800 µg/m^3 respectively for an exposure time of 0.5 h. The concentration decreases with increasing exposure time, and, for a period of 8 h, is 70 µg/m^3, 160 µg/m^3, and 320 µg/m^3, respectively. There is no figure for prolonged exposure, as the necessary data are not available. The WHO specifies a figure of 60 µg/m^3 for the protection of vegetation for long periods, as average value for the vegetation period [108].

3.4.2.8. *Dust*

The term dust refers to particulate air pollutants of very heterogeneous composition. Their phytotoxic effect is considerably less than that of harmful gaseous pollutants. The direct effect of dusts depends very much on their solubility in water. Indirect effects on roots depend on the condition of the soil. Dust deposits on plant surfaces do not usually have a serious effect. However, damage to vegetation can be caused by salt, gypsum, cement, and chalk dusts (erosion and incrustation). Heavy deposits, e.g., of fly ash, on leaf surfaces can impair the photosynthesis efficiency of the assimilation

organs. As well as physical methods of determining the constituents of sedimentary (i.e., deposited) dust and airborne dust, chemical methods are also very important. Dust deposits containing the heavy metals lead, cadmium, and zinc are of particular significance. Zinc is an essential trace element for plants. Plants with an inadequate supply of zinc show deficiency symptoms, but the metal is toxic in excess. Lead and cadmium are not essential for plant life, and have a much greater toxic effect on plant tissue.

These harmful substances enter the plant directly from polluted air, and also indirectly via the roots from soil in which they have accumulated. The main subject of interest is not the slight phytotoxic effects, which are only observed at higher heavy metal concentrations. More important is the fact that plants used for both human and animal food can become contaminated. This contamination can pass down the food chain and endanger the health of humans and animals [109], [110].

3.4.2.9. *Acid Rain*

It had already been established by the mid-1800s that smoke from industrial installations and domestic heating can affect the chemical composition of rain, and that sulfuric acid was one of the pollutants concerned. After 1850, rainwater analyses were carried out in Germany and other countries to determine the effect of air pollution on the penetration of substances into the soil. These investigations were carried out because it was thought that chronic pollution of the soil could also affect plants by this route.

From 1950 onwards, following worldwide discussions on acid rain, regional measurement networks were installed in large parts of Europe and North America [110]. The main measurements made were of the sulfate, nitrate, and chloride levels in the rain, though other substances, e.g., ammonium compounds, were also determined. The highest acid content was found in the industrialized areas of Central Europe, where a pH of 4 was recorded. Rain from areas without anthropogenic pollution had an average pH of 5.6 – 4.6.

Absorption and Plant Reaction. The action of acid rain on plants has mainly been investigated under laboratory conditions. Plants show a considerably greater reaction under laboratory conditions than has been found in the smaller number of open land observations. The absorption of acid rain is influenced by morphological and structural properties of leaves and needles. Under comparable exposure conditions, broad-leaved plants are usually more sensitive than narrow-leaved plants. Damage to the leaf organs takes the form of fossules (pitting), dent-like deformations, and punctiform necroses. The effect of acid rain apparently is to produce more rapid weathering of the cuticle. Chlorophyll degradation is also possible, and the rate of washout of the plant's own nutrients is increased by proton – cation exchange. However, the negative effects of the added hydrogen ions can sometimes be more than compensated for by plant nutrients in the acid rain. The many possible effects of acid rain have been reported in comprehensive studies carried out by VDI [110]. Investigations in the United States have shown that treatment with simulated acid rain led to reductions in the yield of maize and soya beans, these being the most economically important crops in the United States.

Acid rain can cause indirect effects on vegetation via the soil. Changes to the chemical and physical soil condition are particularly important, as borne out by the intensity of the discussion of this complex question. In particular, toxic metal ions that can damage the root system may be liberated. The most important aspect of the action of acid rain on the soil is the increase in the plant availability of toxic aluminum and heavy metal ions due to pH reduction.

A considerable increase in forest damage took place in the 1970s in Germany, and acid rain was immediately thought to be the cause. The first damage was observed to fir trees in so-called pure air regions of Southern Germany. Since 1980, widespread damage to spruce trees has also been recorded. There has also been increasing damage to deciduous trees, in which the symptoms, chronological development, and distribution of the damage were different from those for fir and spruce trees. At first, many different conclusions were drawn by experts on the possible causes of the forest damage. These ranged from the one-sided assumption that the effects were due to exclusively natural factors to the view that they were due only to anthropogenic factors. In the course of the continuing discussion and consideration of the results of new investigations, it appeared more and more probable that forest damage was linked to air pollution whose effects could be changed by natural and/or silvicultural influences. Three possible effect mechanisms are favored [110]:

1. Nutrients are washed out of contaminated soil by acid rain, and toxic metal ions (e.g., aluminum) are released. These damage the plant roots, whose ability to absorb nutrients and water is reduced.
2. The plants are directly damaged in the crown region by acid rain.
3. There is a direct effect of gaseous air pollutants, leading to a certain susceptibility of the whole organism to the subsequent effects of acid rain.

The causes of forest damage are still under discussion. There are other possible factors, e.g., combination effects of various air pollutants, toxic substances as yet unknown, and various stress factors.

[Top of Page]

4. Laws and Technical Regulations

Gerhard Feldhaus and Klaus Grefen

4.1. Introduction

In Germany, the maintenance of air purity and plant safety have their legal basis in the Federal Antipollution Law or *Bundes-Immissionsschutzgesetz* (BImSchG). This is the most important legislation for the protection of the environment from the point of view of both environmental and industrial politics. It is a demanding law of great importance in environmental politics, aiming to prevent air pollution, to provide protection from noise and vibration, to prevent environmental damage and other hazards, and to promote safety in production plant. Its significance for industrial politics is that almost all industrial and business enterprises require a license before starting operation, and must remain under

continuous environmental supervision by the state during their operation. The BImSchG is therefore effectively a plant licensing law into which extensive environmental protection is integrated.

The law lays down the basic requirements which the plant must satisfy. In doing so, it makes considerable use of so-called undefined legal terms such as harmful effects on the environment, pollution, emissions, hazards, plant, state of the art, precautions, and compatibility. These legal terms are characterized by the fact that their meaning is only partially defined by the law. They therefore admit scope for discretion on the part of the authorities and courts, and thus affect legal security. However, they are unavoidable in law concerning technology in order to be able to keep abreast of scientific and technological developments. “The setting of a given legal safety standard by making rigid rules is more likely to limit both technological development and the preservation of basic rights (if it allows these at all) than to enhance them. It would constitute a backward step at the cost of safety” [decision of the German Federal Constitutional Court (Bundesverfassungsgericht) of Aug. 8th, 1978: the so- called Kalkar Ruling]. The use of undefined legal terms is thus asserted to be fundamentally and constitutionally permittable.

However, undefined legal terms constantly require further definition to give consistent legal applicability. For this, there are several possibilities:

1. Regulations enacted on the basis of BImSchG. The most important of these are those regulations issued by the Federal Authorities for the enforcement of the Federal Antipollution Law (BImSchG). Others are issued by the Länder (e.g., smog regulations). Regulations are legally binding on persons to whom they apply.
2. General administrative regulations. These define the undefined legal terms, with legally binding effect on the executive authorities, or they restrict the discretion of these authorities. The most important of these regulations are the technical instructions for the maintenance of air purity (TA Luft) and protection against noise (TA Lärm).
3. Standards set by groups or committees of nongovernmental technical legal authorities. These can be National Standards (DIN, VDI), EU Standards (CEN, CENELEC), or International Standards (ISO). These standards are only binding if government regulations refer to them. Otherwise, they generally have the status of generalized expert opinion.

Further development in the areas of air pollution prevention and plant safety will be increasingly determined by the following:

1. National legislation will be increasingly controlled by laws made by the EU.
2. In the national, supranational, and international areas, regulations relating to individual environmental pollutants are becoming increasingly incorporated into general antipollution regulations. Examples of the latter include environmental criminal law, the law dealing with tests for environmental compatibility, and the law of liability for environmental matters.

4.2. Federal Antipollution Law

The law relating to the protection of the environment from the harmful effects of air pollutants, noise, vibrations, and similar effects (Bundes-Immissionsschutzgesetz, BImSchG) has been in force since April 1st, 1974. It has been much modified since then, mainly by the 2nd amending law of Oct. 4th, 1985, the 3rd amending law of May 11th, 1990, the law of April 22nd, 1993 relating to capital investment facilitation and residential building land, which has been in force since May 1st, 1993, as well as the law for the acceleration and simplification of antipollution licensing procedures of October 9th, 1996.

The BImSchG brought about a comprehensive legal reorganization of the law on pollution. In particular, it superseded the regulations of the §§ 16 ff. Trade and Industry Code (Gewerbeordnung) and a large number of regulations made by the Länder; included nonlicensable installations; and created new regulations (e.g., for planning, traffic, and products).

The BImSchG makes regulations for the building and operation of installations, but not for effects on the environment that depend simply on human behavior, which are covered by law at the Länder level. As far as air purity and noise reduction are concerned, BImSchG applies to installations in all areas (e.g., industry, trade, crafts, agriculture, and forestry) and includes nationalized and private enterprises (§§ 2). Antipollution Law in the wider sense applies only to profit-making enterprises. It basically does not apply to airports and nuclear power stations.

The law has three aims (§§ 1):

1. The most important aim is protection against harmful effects on the environment.
2. For licensable installations, protection is extended to other hazards such as fire and explosion and risk of disturbances which may be associated with the production of chemical products in particular.
3. The long-term aim is prevention. Even if there are no harmful environmental effects, the aim is to improve the protection of human living space, especially by judicious utilization of modern technology (state of the art).

Pollution prevention law is divided into five main areas:

1. Installation-related pollution prevention
2. Product-related pollution prevention
3. Region-related pollution prevention
4. Traffic-related pollution prevention
5. Organization-related pollution prevention

Installation-related pollution prevention mainly relates to stationary installations, i.e., factory premises. A distinction is made between licensable installations (§§ 4 ff.) and nonlicensable installations (§§ 22 ff.).

Licensable installations are those whose nature or size render them especially likely to bring about harmful effects on the environment or to cause other hazards. They include stationary waste disposal plants in which waste materials are stored or treated. Licensable installations are specified in detail by regulations (4. BImSchV).

Licensable installations may only be erected, operated, or significantly modified if a license has been granted (§§ 4, 15).

Licensable installations incur four fundamental duties (§§ 5):

Duty to Protect. The construction and operation of the installation must not lead to any possibility of harmful effects on the environment or to any other hazards, or to serious detriment or nuisance. This duty extends also to the prevention of pollution. In general the limits of harmful pollutants are more exactly defined mainly in general administrative regulations, especially by the immission values of TA Luft and TA Lärm. These specify at what level of regular exposure there is a harmful effect on the environment. The immission figures are always determined by well-defined measurement and assessment methods. More precise definition is usually provided by nongovernmental regulations, such as DIN standards and VDI guidelines. Duty to protect is more precisely specified for installations where there is a hazard from potential disturbances. Here, methods of preventing such disturbances or of limiting their effects are prescribed in the accident regulations (12. BImSchV).

Duty to Take Precautions. When licensable installations are constructed and operated, precautions must be taken against harmful effects on the environment, in particular by state-of-the-art measures to limit emissions. These are usually assessed by comparison with emission limits as prescribed principally in the TA Luft Regulations, Power Station Emission Regulations (13. BImSchV), and Waste Incineration Plant Regulations (17. BImSchV).

Duty to Avoid, Reuse, or Dispose of Residues. Licensable installations are among the most significant producers of waste. The goal of waste prevention is best served by preventing the generation of residues, or, failing that, by the lawful and nonhazardous utilization of any residues generated. Residues are substances that arise during production and processing that are not desired by the proprietor of the business, who is only absolved from the duty of preventing or reusing these residues if this is technically impossible or unreasonable. The proprietor is then obliged to dispose of the residues as waste materials without impairment of the well-being of the general public.

Duty to Utilize Waste Heat. In 1985, the duty to utilize waste heat produced in an installation was introduced as the fourth fundamental duty. The prerequisite is that such utilization is technically possible and feasible for the type and location of the installation. In 1990, the duty to utilize waste heat was widened to include supplying it to a willing third party. This fundamental duty only applies to installations specified by order of the Federal Government. No such a law has so far been enacted.

The license is always granted by following a formal procedure, i.e., with participation of the public. The procedure is laid down in §§ 10 and 9th BImSchV, and is initiated by a written application to which the documents required for assessment should be attached. These include descriptions of installations and business enterprises, data on emissions, and proposed measures for preventing harmful effects to the environment and other hazards, including a safety analysis in the case of certain installations. The licensing authority receives the statement from the authority whose field of duties is affected by the proposal. If the application documents are completely in order, the proposal is published, and the application and documents must be exhibited for public inspection for one month. Any person then has the right to raise objections to the proposal up to two weeks after the end of this period, i.e., a total of about six weeks. If objections are raised, the licensing authority must discuss these with the objectors and the applicant. In those types of installation specified in column 2 of the Appendix to 4th BImSchV, a simplified license application procedure without the participation of the public is used (§ 19). Discussion is concluded without notification or display of the documents.

In order to speed up the licensing procedure, preliminary decisions can be made for individual licensing prerequisites and for the location of the installation (§ 9). Large installations can be licensed in stages by means of partial licenses (§ 8). With the introduction of the amending law of 1996, modifications to existing plant only require authorization only if they are expected to cause negative effects (§ 16, Section 1). The authorities must be notified of other modifications at least one month prior to implementation (§ 15, Section 1). If an installation is to be significantly altered, the licensing authority, before the required license is granted, can permit commencement of construction of the installation, if at that point in time it can be assumed that a license will be granted, if there is public or private interest in the earlier building of the installation, and the proposer of the project accepts the risk of subsequent refusal of the license (§ 8 a).

If the basic duties of § 5 are fulfilled and the installation satisfies other public legal requirements and concerns of industrial safety the licensing authority must grant the license (§ 6), and the applicant has a legal right in this respect. In accordance with other regulations, the license for the installation includes most of the other licenses (e.g., planning permission) and permits (concentration effect of § 13).

The license authorizes the construction and operation of the proposed installation. Subsequent restrictions to the licensed business are only allowed to a limited extent, especially by subsequent regulations under § 17, but they may not be excessive.

Installations which do not require a license under § 4 are in the category of nonlicensable installations (§§ 22 ff.), e.g., small and medium-sized commercial enterprises, domestic heating, and construction machinery. Less stringent basic duties apply to these (§ 22). For some types of installation, the duties are detailed in official regulations, e.g., 1st BImSchV (small furnaces), 2nd BImSchV (installations that contain halogenated hydrocarbons), 7th BImSchV (wood processing plants).

For both licensable and nonlicensable installations, supervisory procedures may be prescribed, usually according to § 52, or, with special reference to the determination of emissions and pollutants, according to §§ 26 ff. Since Sep. 1st, 1990,

anyone who operates a licensable installation must provide a statement of emissions, and this must be updated every four years (§ 27, 11. BImSchV). In the statement of emissions, figures must be provided for the harmful substances emitted from the installation during the previous calendar year.

Following the 3rd amending law of 1990, BImSchG has been upgraded to the status of a plant safety law. The authorities are empowered to order certain technical safety checks and safety reviews (§ 29 a). The law also provides for the appointment of a technical committee for plant safety (§ 31 a), an accident committee (§ 51 a), and accident representatives (§ 58 a ff.).

Regulations can demand the prevention of harmful effects on the environment from air pollutants, plant noise or vibration, combustible materials, fuels, and other materials and products (§§ 32 ff.). Relevant regulations include: 3rd BImSchV (sulfur content of light heating oil and diesel fuel), and 19th BImSchV (chlorine and bromine compounds used as fuel additives).

A section of the BImSchG is dedicated to the prevention of pollution from traffic. The design and operation of motor vehicles can be regulated in accordance with § 38 with respect to air purity and noise prevention. In still weather, traffic restrictions can be ordered by the Land authorities; the smog regulations issued by the Länder provide the details. Most of the smog regulations so far enacted have since been repealed, since the pollutant load of the air in Germany has been lowered to such an extent by the numerous protective measures that even under smog weather conditions the critical values are no longer attained. However, so-called summer smog — a short-term increase in atmospheric ozone concentrations near the ground under the influence of sunlight — is regarded as a new problem. The amending law of July 19th, 1995 introduced §§ 40 a to 40 e into law. They provide for traffic bans in the event of high ozone concentrations. Since September 1st, 1990, the street traffic authorities can order traffic restrictions, whether in still weather or not, to prevent harmful air pollution, especially in the inner cities (§ 40 Section Physics and Chemistry of the Atmosphere). The twenty-third statutory regulation for the implementation of the Federal Antipollution Law (23. BImSchV) of December 16th, 1996 specifies concentration values for traffic-related air pollution. If these values are exceeded, the responsible authorities of the Länder must determine whether traffic restrictions are necessary, and if so, which measures are to be taken. A complete system of noise prevention for the design of streets, railways, and roads is provided by §§ 41 – 43 in conjunction with § 50. Some details are covered by the Traffic Noise Prevention Order (16. BImSchV).

The prevention of pollution in individual regions is controlled by §§ 44 ff. and § 49. These regulations control regional or local problems of pollution prevention. For areas in which especially hazardous air pollutants occur or are likely to occur, a continuous comprehensive control of air polluting substances is prescribed (§ 44). All significant sources of air pollution are included in an emission register, i.e., industrial and other business enterprises, domestic heating, and motor vehicles (§ 46). The emission register is based on the statements of emissions made in accordance with § 27. If investigations show significantly high concentrations or a continuous increase in air pollution over long periods of time, the setting up of air antipollution plans must be considered for the whole area of investigation or for parts thereof (§ 47). Under certain conditions, the Land authorities must prepare a clean-up plan. In other cases, according to their judgment, they may prepare a precautionary plan. Noise reduction plans may be prepared under certain circumstances in accordance with § 47 a. For areas in special need of protection (highly polluted areas or areas receiving special protection), specially stringent controls may be prescribed (§ 49).

As well as stipulating requirements for environmental technology, the law also lists a number of duties in connection with management structure. Proprietors of businesses that have environmental implications resulting either from emissions or from the products manufactured must appoint an antipollution representative (§ 53). He is the company expert in questions of pollution. His most important task is, on his own initiative, to promote the introduction of environmentally friendly processes and products, and to support the managers of the business therein. Control of the working environment within the company is also entrusted to him (§ 54). He has a special position within the company, having the right to make direct representations to management in important questions. He may not be dismissed as a result of his actions (§§ 56, 57, 58). For especially hazardous installations, the appointment of an accident representative is prescribed (§ 58 a). His rights and duties essentially correspond to those of the antipollution representative (§ 58 b – 58 d). The German Recycling and Waste Act (Kreislaufwirtschafts- und Abfallgesetz) prescribes the appointment of a waste materials representative, also modeled on the antipollution representative, and similarly the water resources policy act prescribes a water pollution representative. These various functions are often assigned to a single member of staff, who then has the function of an environmental representative.

Since Sep. 1st, 1990, proprietors of licensable installations have been obliged to make reports of their management structure to the competent authorities (§ 52 a). In corporations whose authorized representation includes several members, and in unincorporated firms with several legally appointed staff representatives, the authority must be informed who within the management is responsible for pollution prevention. All proprietors of licensable installations must also inform the authority of their method of ensuring that regulations and orders intended to prevent harmful effects on the environment and other hazards will be observed in their organization. In this connection, the management structure must be made known to the authority. The law does not demand a more detailed management structure, but more details emerge, for example, from the stipulations concerning management representatives, accident regulations, and general rules of management structure. Further elements are contained in the EU Regulation of June 29th, 1993 concerning the voluntary participation of business undertakings in a cooperative system for environmental management; investigation of the environmental aspects of a company (ABl. Nr. L 168/1) are regulated by the Environmental Audit Law of December 7th, 1995 and DIN-EN-ISO 14001, 14004, 14010, 14011, and 14012.

4.3. Federal Statutory Antipollution Regulations Relating to Air Pollution and Plant Safety

4.3.1. *First BImSchV*

The first statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning small furnaces – 1. BImSchV) of July 15th, 1988, mainly based on § 23 BImSchG, controls the construction, structure, and operation of furnaces that do not require a license under BImSchG (§ 1), i.e., installations with an output of up to 1 MW

using solid fuel, 5 MW using light heating oil, or 10 MW using gaseous fuel. Small furnaces are important from the point of view of pollution prevention because of their abundance and their relatively high contribution to pollution (low-altitude gas discharge).

Main contents: Limit values for particulates, CO, and oil derivatives (§§ 4 ff.); limitation of NO_x emissions from oil- and gas-fired furnaces by means of state-of-the-art combustion technology (§ 9); limitation of waste-gas emissions for oil- and gas-fired furnaces (§ 11); restriction to certain fuels (§ 3); inspections (§ 14); annual measurements during supervision by the regional stack inspector (§§ 15 ff.); discharge conditions for waste gases (§ 18). Existing furnaces must be modified after a transition period to comply with requirements for new installations (§ 23).

4.3.2. *Second BImSchV*

The second statutory regulation for the implementation of the Federal Antipollution Law (regulation to limit emissions of volatile halogenated hydrocarbons – 2. BImSchV) of Dec. 10th, 1990, based on § 23 BImSchG, applies to nonlicensable installations in which volatile halogenated hydrocarbons are used as solvents, especially surface treatment plants, dry cleaning plants, textile finishing plants, and solvent extraction plants. The regulation incorporates new knowledge about the potential hazards of volatile halogenated hydrocarbons and the ability of volatile chlorofluorocarbons (CFCs) to destroy ozone. This regulation implements The Montreal Protocol of Sep. 17th, 1987 concerning CFCs.

Main contents: The only permitted volatile halogenated hydrocarbons are tetrachloroethyl-ene, trichloroethylene, and dichloromethane, and these may be used only in certain installations (§ 2). Authorization for CFCs expired on Dec. 31st, 1994. Technical requirements, handling methods, and emission limit values are prescribed to limit the emission of permitted solvents. Existing installations must be renovated by Dec. 31st, 1994 at the latest (§ 6). This regulation also contains instructions for measurement and supervision (§§ 10 ff.).

Disposal of solvents from these or similar installations is regulated in the regulations for used solvent disposal (Oct. 23rd, 1989). Other restrictions and prohibitions relating to CFCs are contained in the EU Regulation regarding substances with an ozone-depletion potential of December 15th, 1994. and the CFC – Halon Prohibition Order (May 6th, 1991), which is based on the Chemicals Act.

4.3.3. *Third BImSchV*

The third statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning the sulfur content of light heating oil and diesel fuel – 3. BImSchV) of Jan. 15th, 1975, based on § 34 BImSchG.

Emissions from domestic and small commercial furnaces and diesel vehicles have a major influence on SO_2 emissions in densely populated areas because of their low emission altitude. To reduce these emissions, this regulation limits the sulfur content to 0.20 wt % in light heating oil (following the amending regulation of Dec. 14th, 1987) and in diesel fuel to 0.05 wt % from October 1st, 1996 in accordance with the amending law of September 26th, 1994. The EU Guideline 75/716/EEC of March 30th, 1987 prescribes the lowest possible limit value. Earlier limit values were: May 1st, 1975: 0.55 wt %; May 1st, 1976: 0.50 wt %; Jan. 1st, 1979: 0.30 wt %. Since Oct. 1st, 1996, a new limit value of 0.05 wt % applies for diesel fuel.

4.3.4. *Fourth BImSchV*

The fourth statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning licensable installations – 4. BImSchV) in the version of March 14th, 1997, based on § 4, Section 1, Clause 3, BImSchG specifies the range of licensable installations. Only those types of installations named in the Appendix require a license, usually on the grounds of maintaining air purity. They include almost all industrial installations, e.g., those involving chemicals (see Nos. 4.1 – 4.10). The obligation to be licensed often depends on the capacity or size of an installation and extends to adjacent equipment which is associated with it geographically and operationally and which can have an important bearing on the emission or immission behavior or the safety of the installation, such as the storage of raw materials, finished products, and residues, and equipment for transport, waste materials, packaging, and shipping.

For installations named in Section 1, the license is granted in a formal procedure (§ 10 BImSchG) with participation of the public, and for installations named in Section 2 by a simplified procedure (§ 19 BImSchG), i.e., without participation of the public.

4.3.5. *Seventh BImSchV*

The seventh statutory regulation for the implementation of the Federal Antipollution Law (regulation to limit the emission of wood dust – 7. BImSchV) of Dec. 18th, 1975, based on § 23 BImSchG applies to nonlicensable installations for the working or processing of wood or wood- containing materials if wood dust or shavings are emitted from these installations.

Main contents: Wood dust and shavings must be stored in bunkers, silos, or other closed spaces. These must be emptied in such a manner that the emission of wood dust or shavings is prevented as far as possible (§ 3). Emission limit values are specified for wood dust and shavings. If the exhaust air contains particulates, a limit of 50 mg/m^3 in the exhaust air is imposed (§ 4).

4.3.6. *Eleventh BImSchV*

The eleventh statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning the declaration of emissions – 11. BImSchV) of Dec. 12th, 1991, based on § 27, Section 4 BImSchV, concretizes § 27 BImSchG, which obliges proprietors of all licensable installations to provide a statement of emissions to the competent authority, i.e., information on the type, amount, and distribution over space and time of atmospheric pollutants emitted annually from the installation. The emissions statement is the basis for the emissions register, and for planning and clean-up procedures. Installations with low emissions may be excepted (§ 1). The time period for the statement is the even-

numbered calendar year (§ 3). In general a complete statement of emissions must be supplied, with quantities of these emissions. In certain cases, a simplified statement may be provided, for which operational data and a list of materials handled are sufficient (§ 4). According to § 27, Section 1 of BImSchG in the version of the amendment law of 1996, the emission statement must be updated every four years (formerly every other year). The authority may publish the figures from the emission statement, unless trade and operational secrets can be deduced therefrom (§ 27, Section 3 BImSchG).

4.3.7. *Twelfth BImSchV*

The twelfth statutory regulation for the implementation of the Federal Antipollution Law (accident regulation – 12. BImSchV) of Sep. 20th, 1991, mainly based on § 7 BImSchG, basically applies to all licensable installations in which materials referred to in Appendices II and III or other materials that are highly toxic, toxic, able to support combustion, explosive, combustible, highly flammable, or flammable (Appendix IV) are present or can be formed in an accident. It does not apply if negligible quantities are involved (§ 1). The regulation is intended to prevent accidents and to limit their effects as far as possible. Accidents are disturbances to normal operation in which materials as specified in Appendices II, III or IV lead to a serious hazard from large-scale emissions, fires, or explosions, i.e., danger to life, serious effects on health, or environmental hazards affecting the well-being of the population (§ 2).

Main contents: Basic safety duties (§ 3), detailed safety duties (§§ 4 – 6), safety analysis (§ 7), duty to report accidents and near accidents (§ 11), information to the public (§ 11 a), duty of notification (§ 12). Safety duties requiring qualified personnel (e.g., safety analysis) need only be carried out for installations named in Appendix I (§ 1, Section 2); these are mainly chemical installations.

4.3.8. *Thirteenth BImSchV*

The thirteenth statutory regulation for the implementation of the Federal Antipollution Law (power station emission regulation – 13. BImSchV) of June 22nd, 1983, mainly based on § 7 BImSchG, applies to the construction and design of large furnaces, i.e., installations with a combustion heat output of ≥ 50 MW, or ≥ 100 MW for gaseous fuels (§ 1).

Main contents: Emission limit values for particulates, CO, NO_x, SO_2, and halogen compounds (§§ 3 ff.). A continuous monitoring method is desirable for measurement and assessment (§§ 21 ff.). For NO_x, emission limit values were not finalized, as the technical development was still in a state of flux. Here, so- called dynamic clauses were introduced, requiring the use of the existing state of art. These have since been specified in detail to enable them to be carried out by management. The part with most relevance to air purity consists of instructions for cleaning up old installations (§§ 17 ff.). The regulation gave the proprietors until July 1st, 1984 to decide whether they wished to reequip their installations or shut them down by April 1st, 1993. If they opted for reequipment, this had to be completed by July 1st, 1988. Most furnaces have now been reequipped. Compared with 1980 figures, emissions of SO_2 in 1993 from large furnaces in the “old” Länder of the Federal Republic have been reduced by ca. 80 %, of NO_x by ca. 70 %, and of particulates by ca. 75 %. In the area of power generation, > 11×10^9 € were invested for these measures, of which 7.3×10^9 € were for desulfurization of flue gas and 3.5×10^9 € for reduction of NO_x emissions.

4.3.9. *Seventeenth BImSchV*

The seventeenth statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning incineration plant for combustible waste and similar materials – 17. BImSchV) of Nov. 23rd, 1990, based on § 5 Section 2, and § 7 BImSchG, applies to all licensable installations in which solid or liquid waste or similar solid or liquid combustible materials are burnt, including installations whose main purpose is different (e.g., cement furnaces).

Waste incineration plants are of great significance for the environment because of their emissions of gaseous inorganic chlorine and fluorine compounds, and also because of some constituents of emitted particulates, especially heavy metals and polyhalogenated aromatic compounds. The emission limit values laid down earlier in TA Luft 1986 were made much more stringent, and the requirements of the EU Guidelines of June 1989 with respect to existing and new incineration plants for domestic refuse were incorporated in national law.

Main contents: Emission limit value for dioxins and furans 0.1 ng TEq/m^3 (i.e., reduction in dioxin emissions from waste incineration plants by > 99 %); emission limit values for other air pollutants (§ 5); constructional and operational requirements, e.g., dwell time and minimum temperature (§ 4); and the annual provision of information to the public (§ 18). For existing installations there were a series of intermediate deadlines until March 1st, 1994 or Dec. 1st, 1996 (§ 17).

4.3.10. *Nineteenth BImSchV*

The nineteenth statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning chlorine and bromine compounds – 19. BImSchV) of Jan. 17th, 1992, based on § 34 BImSchG, seeks to reduce pollution of the environment by dioxins and furans by prohibiting the addition of chlorine and bromine compounds (scavengers) to motor vehicle fuel. The marketing of chlorine and bromine compounds as fuel additives and the marketing of fuels that contain such additives is forbidden (§ 2).

4.3.11. *Twentieth BImSchV*

The twentieth statutory regulation for the implementation of the Federal Antipollution Law (regulation to limit hydrocarbon emissions during the transfer and storage of gasoline – 20. BImSchV) of Oct. 7th, 1992, based on § 23 BImSchG and § 24 Trade Regulation.

The object of the 20th BImSchV is to reduce the emission of hydrocarbons during the supply of gasoline. These substances may be directly hazardous to health (e.g., the carcinogen benzene). However, hydrocarbons are also important precursors for the formation of ozone and other photooxidants. The regulation applies to nonlicensable installations to which deliveries of gasoline are made, or from which gasoline is taken, e.g., filling stations, small supply

tanks, and road and rail tankers (§§ 1, 2).

Main contents: In principle, the gas displacement method must be used in filling operations from storage tanks and during deliveries to filling stations. If this is technically infeasible, the gasoline vapors must be passed to a gas- cleaning device (§ 3). The filling of vehicles at filling stations is regulated by 21. BImSchV.

4.3.12. *Twenty-First BImSchV*

The twenty-first statutory regulation for the implementation of the Federal Antipollution Law (regulation to limit hydrocarbon emissions during the filling of motor vehicles – 21. BImSchV) of Oct. 7th, 1992, based on § 23 BImSchG and § 24 Trade Regulation.

The object of 21st BImSchV is to reduce the release of gasoline vapors during the filling of motor vehicles at filling stations. Customers, filling station personnel, and persons who live close to filling stations are exposed to gasoline vapors, which contain the carcinogen benzene.

The regulation applies to nonlicensable filling stations where motor vehicles are filled with gasoline (§ 1). In the filling operation, a gas recycling system must be used which collects the gasoline vapors from the tank of the vehicle and feeds them back to the storage tank of the filling station (§ 3). The regulation applies with immediate effect to new filling stations. Existing filling stations must be equipped with a gas recycling system within not more than five years (§ 9).

4.3.13. *Twenty-Second BImSchV*

The twenty-second statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning levels of pollutants – 22. BImSchV) of Oct. 26th, 1993, based on § 48 BImSchG, incorporates three EU Guidelines into Federal Law: of July 15th, 1980 (80/779/EEC amended by June 21st, 1989 89/427/EEC) on limit values for SO_2 and suspended particulates, of Dec. 3rd, 1982 (82/884/EEC) on the limit value for the lead content of air, and of March 7th, 1985 (85/203/EEC) on air quality standards for NO_x. The regulation became necessary after the European Court of Justice in May 30th, 1991 judged that the Federal Republic of Germany had not duly incorporated these guidelines. The limit values named in the guidelines are established as pollution limit values by the regulation (§ 1). The following were also established: the main reference time periods (§ 2), criteria for the location of measuring points (§ 3), and the methods of measurement and analysis (§ 4). The actions necessary if the limit values are exceeded are specified in § 6 (air purification plans, official orders). TA Luft is unaffected according to § 5.

4.3.14. *Twenty-Third BImSchV*

The twenty-third statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning the stipulation of concentration values – 23. BImSchV) of December 16th, 1996 specifies, for certain streets or areas in which especially high traffic-related pollution levels are to be expected, concentration values for nitrogen dioxide, soot, and benzene. If these values are exceeded, the need for traffic-restricting measures must be determined.

4.3.15. *Twenty-Fifth BImSchV*

The twenty-fifth statutory regulation for the implementation of the Federal Antipollution Law (regulation to limit emissions from the titanium dioxide industry – 25. BImSchV) of November 8th, 1996, based on § 48 a, Section 1, Clause 1, BImSchV, applies to the production of titanium dioxide, the concentration of weak acid, and the cleavage of sulfate-containing salts. It specifies values for dust, sulfur dioxide, sulfur trioxide, and chlorine (§ 3, 4).

4.3.16. *Twenty-Seventh BImSchV*

The twenty-seventh statutory regulation for the implementation of the Federal Antipollution Law (regulation concerning crematoriums – 27. BImSchV) of March 19th, 1997 is based on §§ 4 and 19, BImSchG. The regulation applies to crematoriums and contains requirements regarding the use of fuels, technical requirements, emission limits (§§ 3, 4), and measurement specifications (§§ 7 – 10).

4.3.17. *Other Federal Regulations*

The Law on Testing for Environmental Compatibility of Feb. 12th, 1990 is based on the EU Guideline of June 27th, 1985 on testing for environmental compatibility for certain public and private projects (85/337/EEC).

Testing for environmental compatibility must be carried out for the most important licensable installations, nuclear installations, projects involving water, mining installations, trade installations, and other installations (Appendix to § 3 of the law).

The proposer of the project determines and describes the foreseeable effects of the project on the population and the environment. This involves a process of communication and consultation with the relevant authorities and the general public. The competent authority then presents and appraises all the information on the project in a comprehensive report. The findings concerning licensing of the project are considered using the prescribed official procedure, and must be made available to the public. For licensable installations, the test for environmental compatibility is carried out as part of the licensing procedure. Details of this are given in 9th BImSchV.

The Law of Liability for the Environment of Dec. 10th, 1990 establishes liability for damage to the environment caused by installations specified in Appendix I (§ 1). These include the most important (according to the BImSchG) licensable installations. Liability is not limited to accidents or disturbances, but also includes normal operation (§ 5, § 6 Section 2). Liability is considerably increased by the presumption of cause in § 6, i.e., if an installation is liable to cause the damage produced, it is presumed that the damage was in fact caused by the installation. The authorities can set aside the presumption of cause on receipt of proof that the installation was operated in accordance with regulations.

Environmental Sanctions. According to the 18th Criminal Law amendment (law for combating environmental crime) of March 28th, 1980, modified as of June 27th, 1994 a special section on criminal offences against the environment was introduced into the penal code. The penal code of §§ 324 – 330 d includes the following punishable offences: § 325 (air pollutants hazardous due to the improper operation of installations), § 327 (illicit operation of installations), § 329 (imperilment by air pollutants of areas needing protection), § 330 (serious hazard to the environment), § 330 a (serious hazard due to the release of toxic substances).

4.3.18. *Smog Regulations of the Länder*

By the powers given to them in §§ 49 Section 2 and 40 Section 1 of the BImSchG, the individual Länder (with the exception of Mecklenburg-Vorpommern) have issued smog regulations which control the designation of smog zones, and specify measures to be taken to restrict traffic and reduce emissions in the case of a smog alert. The hazard levels (warning, first-stage alert, second-stage alert) correspond with the following measures:

1. Temporary restriction of the operation of plant (including complete shutdown)
2. Restriction of the use of certain fuels (e.g., high-sulfur coals and heating oils) in industrial and business enterprises and communal and private households
3. Prohibition of traffic for certain periods, certain vehicles, and certain zones

The smog regulations of the Länder follow the so- called model regulation developed by the Länder antipollution commission and therefore have essentially the same structure, although each Land has its own special features. All smog regulations, as well as defining and prescribing the above warning stages, contain certain smog event criteria. These criteria are based on both the actual meteorological conditions and the air pollution situation. Whereas the definition of still weather conditions can differ from Land to Land (e.g., for North Rhine Westphalia the definition is an average wind speed of < 1.5 m/s at ground level over a period of 12 h, but for Bremen 4 m/s), in all smog regulations of the Länder uniform criteria apply to the decision as to when a given concentration of harmful substances corresponds to one of the alert stages. However, the number and distribution of the measuring stations in the network of pollution monitoring stations in the smog zones of the Länder are different, so that ultimately the methods of deciding on the various alert stages in the individual Länder also differ. For example, different rules apply to the issuing of a smog alert if the concentration figures found exceed the limit values (for SO_2 in combination with suspended particulates, for SO_2 alone, for NO_2, or for CO) at more than 50 % of the measuring stations, at all measuring stations, or at, e.g., two neighboring measuring stations.

A smog alert is issued if the concentration limit values for only one of the above-mentioned harmful substances or for SO_2 in combination with suspended particulates are exceeded (e.g., the 24-h average and the last 3-h average of the sum of the concentration of the SO_2 and twice the concentration of suspended particles, or if the 3-h average for SO_2, NO_2, or CO is exceeded).

4.4. General Administrative Regulations for the Prevention of Air Pollution

The Federal Government is empowered according to § 48 BImSchG, after hearing the parties concerned and with the approval of the Bundesrat, to issue general administrative regulations. These regulations, which in principle are binding on the administrative authority only, serve to implement BImSchG and the statutory regulations made under this law.

4.4.1. *TA Luft*

The first general administrative regulation (Allgemeine Verwaltungsvorschrift – 1. AVwV) of the Federal Antipollution Law (Technical Directions for the Prevention of Air Pollution, or Technische Anleitung zur Reinhaltung der Luft, TA Luft) is an important instrument for the implementation of the Federal Antipollution Law. As described in Part 1 of the four-part TA Luft, this is to be applied in particular

1. When considering applications for a license for the construction and operation of an industrial installation, for any significant change in its location, design, or operation, or for a cleanup of such an installation
2. When considering applications for a preliminary decision
3. For supplementary regulations
4. When ordering the determination of the nature and extent of the emissions from an installation and of the extent of pollution in the zone of influence of the installation

The first TA Luft regulation, issued in 1964, has since been amended several times. The chronological development was as follows:

Sep. 8th, 1964	General administrative regulation for licensable installations according to §§ 16 industrial code (TA Luft 64).
Aug. 28th, 1974	First general administrative regulation (TA Luft 74) of the BImSchG. This was brought into operation in March 15th, 1974.
March 4th, 1983	Amendment of TA Luft. This was largely restricted to a revision of the regulations for licensing procedures and for the assessment of pollution for “defense against hazards”.
March 1st, 1986	Amendment of TA Luft following amendment of the BImSchG (Oct. 4th, 1985) and the earlier (July 24th, 1985) amendment of 4th BImSchV for licensable installations, with adjustment of the installation catalog to conform to technical developments. The “principle of prevention” was increasingly taken into account, and the cleanup concept for old installations was first established.

The four parts of TA Luft are as follows:

1. Scope
2. General instructions for the prevention of air pollution
3. Requirements for the limitation and determination of emissions
4. Requirements for old installations (cleanup)

Part 2 of TA Luft (see above for further details of Part 1) specifies in more detail the mainly undefined legal terms included in the expression “harmful effects on the environment due to atmospheric pollution”. Explanations are given of, e.g., the assessment of health hazards, significant detriment, significant pollution, assessment in those cases where limit values have not been established, and abnormal occurrences. In No. 2.5, in explanation of § 5 of 1 BImSchG, the term “harmful effects on the environment” is defined for important pollutants by means of limit values. A distinction is made between pollution levels at which there are health hazards and levels at which significant detriment and pollution occur. In the latter case, an abnormal occurrence investigation (No. 2.2.1.3 TA Luft) is required if a pollution limit value is exceeded, and this must also be carried out for substances for which no pollution limit values are specified in TA Luft but where there is “reasonable suspicion of harmful effects on the environment”. Other fundamental explanations in Part 2 of TA Luft concern, e.g., the term “precautions”, and give information about methods of limiting carcinogenic substances, the consideration of applications for the granting of an interim action, a partial license, modifications to a license, and the discharge of waste gases. In No. 2.6, “determination of pollution parameters”, important aspects of pollution measurement (measurement planning, sampling, and analysis) are described, and guidelines can be obtained from the VDI handbook of air purity prepared by the KRdL (Kommission Reinhaltung der Luft, or Commission for the Prevention of Air Pollution) formed by the VDI (Verein Deutscher Ingenieure, or Association of German Engineers) and DIN.

In Part 3 “limitation and determination of emissions”, emission limit values corresponding to state-of-the-art technology for licensable installations are specified, and emission minimization requirements are laid down. Large furnaces are not included, as these are covered by 13th BImSchV. Methods for the measurement and control of emissions are specified in Part 3. Emission measurement techniques are described in more detail in the VDI Guidelines listed in Appendix G. Finally, No. 3.1.1 of TA Luft refers to continuing detailing of the state of the technology as defined by the BImSchG, and to the guidelines for process and gas cleaning in the VDI handbook on the prevention of air pollution. These are listed in Appendix F.

A comprehensive plan for cleaning up old installations is set up in TA Luft 86, Part 4. Old installations must be renovated within a given time limit (between “immediately” and 10 years), depending on the type, quantity, and hazardous nature of the emissions and on any special features of the installation. A distinction is made between “supplementary regulations” with immediate effect for the prevention of harmful effects on the environment under certain conditions (if pollution limit values are exceeded or carcinogenic substances are emitted) and precautionary supplementary regulations. As for new installations, the severity of the precautionary requirements (time limits, technical measures) depends on the potential hazard of the emitted substance. Measures to minimize emissions are in all cases fundamentally preferable to improving the condition of discharges.

Unlike laws and regulations, TA Luft is an instruction to the authorities, and does not directly affect individual citizens. However, as its provisions affect antipollution licensing procedures, they have an indirect effect on individual citizens or proprietors of installations with respect to the principle of equality. TA Luft sets measuring standards that are uniform nationally and therefore ensure uniform treatment of proprietors and uniform legal security. Any deviation from its provisions is possible only in exceptional circumstances, e.g., if these provisions no longer correspond to a technological development.

4.4.2. *Fourth BImSchVwV: Determination of Air Pollution in Test Areas*

The general administrative regulation (AVwV) of Oct. 26th, 1993 is based on § 44 Section 1 BImSchG and certain guidelines by the European Parliament. It contains definitions of terms and instructions on the object of measurement, measurement areas (test areas as defined by § 44 Section 3 BImSchG), number and position of measuring points, measuring methods, measuring equipment, and the assessment of results. In order to inform the public about the pollution situation, the data must be provided in a prescribed form suitable for publication. In Appendices A – E of this AVwV, directions are given on EU Guidelines to be considered, the choice of measuring points and instructions to be followed, and the DIN Standards and VDI Guidelines that provide measuring techniques and the determination of process parameters (see also Section TA Luft). The data must be recorded in a standard form, i.e., following the pattern given in Appendix E, and must be transmitted in the form of a report to the EU Commission. Furthermore, the data must be transferred to the German Parliament in accordance with § 61 BImSchG and, in accordance with the highest Land Authorities, to the Federal Environment Office.

The EU (or EEC) Guidelines cited in Appendix C1 and the corresponding Federal Guidelines mentioned in C2 for the monitoring of pollution are of special importance:

1. Guidelines for the construction and testing of the suitability of measuring equipment for the continuous monitoring of pollution (Gemeinsames Ministerialblatt (GMBl), 1981, p. 355)
2. Guidelines for selecting the location and method of construction of automatic measuring stations in telemetric measurement networks (GMBl, 1983, p. 78)
3. Guidelines for establishing reference methods, choosing methods of measuring equivalence, and using calibration procedures (GMBl, 1988, p. 191)

The AVwV will make it possible to establish a uniform federal practice and increasingly a European practice in the determination and monitoring of pollution for the assessment and improvement of the state of air pollution prevention.

4.4.3. *Fifth BImSchVwV: Register of Emissions in Test Areas*

The 5th BImSchVwV of April 24th, 1992 (amended on March 24th, 1993) contains legal requirements which the competent authorities must take into account regarding the establishment and regular updating of emission registers. These registers must be established for the test areas created in accordance with § 44 Section 3 of the BImSchG, in particular for the preparation of air antipollution plans (§ 47 BImSchG). Groups of emitters particularly affected include licensable furnaces (as specified in 1. BImSchV), and other furnaces. This AVwV also applies to “other non-licensable installations” with emissions of volatile halogenated hydrocarbons (as specified in 2. BImSchV) and wood dust (as specified in 7. BImSchV).

They also apply to nonlicensable installations where organic solvents are used if 2nd BImSchV does not apply, e.g., plants for recycling residues, painting, printing, smoke- curing or fumigation, also roasting plants, filling stations, stationary internal combustion engines, surface-treatment plants, and chemical cleaning plants. Other groups to be considered as emitters include livestock, installations of the Federal Railroads, Federal Postal Administration, and Federal Armed Forces, road vehicles, railroad locomotives, shipping, aircraft, and agricultural and military traffic “if inclusion of the emissions is necessary for preparation of the emission register”.

The following substances or groups of substances produced by the emitters listed above are to be considered: dust, lead and inorganic lead compounds, sulfur dioxide, nitrogen oxides (expressed as NO_2), carbon monoxide, and 11 other substances or groups of substances including asbestos, diesel exhaust particles, ammonia, and especially those substances specified in No. 2.3 of TA Luft (carcinogenic substances).

The measurements are carried out either continuously or as individual measurements.

The setting up and presentation of the emissions register (style and scope) is carried out in accordance with established principles to make the data consistent and readily assessible, and to enable and facilitate the updating of the emission register. Thus, the style, scope, and system of presentation are regulated for the individual groups of emitters. The conclusions of the emissions statement order are also important here. The use of 5th BImSchVwV should provide the authorities and other participants with effective instruments for developing measures for reducing emissions and for calculating the spread of pollution.

4.5. Technical Regulations for the Prevention of Air Pollution

The term “technical rules” includes not only technical rules (TR) referred to explicitly under that name, but also standards, guidelines, directions, instructions, worksheets, leaflets, and recommendations produced by various private institutions. However, a technical rule in the wider sense is not only to be regarded as the responsibility of the private maker of rules, but can also require Governmental regulation (e.g., TA Luft). In this section, the issuing of technical rules for the prevention of air pollution by the private institutions DIN, VDI, CEN (Comité Européen de Normalisation), and ISO is described in more detail in the national, European, and international context. These private institutions make an important contribution to the prevention of pollution within the State by converting so- called anticipated expert reports into the format of technical rules. All areas of technology except electrical technology are included. In the DIN catalog of technical rules in the DITR (Deutsches Informationszentrum für Technische Regeln, or German Information Center for Technical Rules), more than 130 private organizations are listed as rule makers. Over 200 regulatory handbooks are recorded. In 1998, ca. 230 000 documents were entered as “technical rules” in the DITR databank in Berlin. Every year, ca. 6000 technical rules are added, and ca. 3000 are withdrawn and hence invalid. The DITR also operates and makes available the multinational and multilingual standards databank PERINORM. This contains the bibliographical information for all valid national standards and proposed standards for Germany, the United Kingdom, France, Austria, Switzerland, and The Netherlands, all European and International Standards, all other valid technical rules in Germany and France, and information on technical rules originating from Government and other authorities, including EU Guidelines. In total, PERINORM 1998 contains over 236 000 entries. PERINORM data can be obtained directly from the DITR, or by using the computer disks available from DITR. The DITR is thus an important instrument for providing reviews and information on technical rules, including those relevant to environment protection.

4.5.1. *National Regulations*

In mid-1990 in Germany, the then existing organizations (the VDI Commission for the Prevention of Air Pollution and the DIN Standards Committee for the Prevention of Air Pollution) merged to form the *Kommission Reinhaltung der Luft (KRdL) im VDI und DIN Normenausschuß* (Commission for the Prevention of Air Pollution in the VDI and DIN Standards Committee). The task of this committee is the preparation of VDI Guidelines, DIN Standards, DIN Provisional Standards, DIN – EN Standards, and DIN – ISO Standards. The office of the KRdL took over the secretariat of the ISO/TC (Technical Committee) 146 “Air Quality” (see also Section International Regulations.) of the DIN Standards Committee for the Prevention of Air Pollution. Furthermore, on the initiative of the KRdL, the CEN/TC 264 “Air Quality” was established in March 1991 (see Section European Regulations.). The secretariat of the KRdL is also responsible for this technical committee. In the establishment of technical rules for the prevention of air pollution, the principles, rules of procedure, and priorities of the DIN (in accordance with DIN 820) and also those of CEN and ISO apply to the procedural guidelines created for the KRdL. In the preparation of VDI Guidelines, the procedural guidelines of the VDI (in accordance with VDI 1000) must be followed.

The KRdL comprises over 200 working groups made up of ca. 1700 specialists in economics, science, and administration working together in an honorary capacity. The KRdL is subdivided into several specialist groups:

Group I. environmental protection techniques
Group II. environmental meteorology
Group III. environmental quality
Group IV. environmental measurement techniques

The work of the KRdL is supported financially by the Federal Government, whose interest in the work of the Commission is

apparent from the statement of the case for the budget, according to which “the Federal Ministry for Environment, Nature Conservation and Nuclear Safety is supported by the KRdL in the performance of tasks in the area of prevention of air pollution as defined by § 1 of the Federal Antipollution Law such that the KRdL, in voluntary self-responsibility and together with the participating authorities, science, and industry, determines the current state of the art in science and technology and records this in guidelines and standards whose contents can become part of the function of the executive and the legislature”.

The KRdL is in charge of ca. 500 VDI Guidelines and DIN Standards, which form the six-volume VDI/DIN Handbook “Prevention of Air Pollution”. The actualization and upkeep of and assistance with ca. 40 DIN/ISO Standards also fall within its sphere of responsibility. The first European Standards (EN) for the prevention of air pollution are in preparation.

4.5.2. *European Regulations*

The protection of the environment is becoming more and more a political topic of the European Union (EU). The 5th Antipollution Action Program of the EU came into force on Jan. 1st, 1993, and will continue until the year 2000. The EU is politically active in the area of technical regulations. In its resolution (85/C 136–01) of May 7th, 1985, which refers to a “New Concept” in the field of technical harmonization and standardization, the European Parliament resolved to limit the harmonization of official regulations of the EU to basic necessities, and commissioned the existing private legislative bodies CEN (Comité Européen de Normalisation) and CENELEC (Comité Européen de Normalisation Electrotechnique) with the preparation of European Standards (EN) for the regulation of technical/scientific matters taking into account the current state of the art of science and technology. The European Standards Committee, CEN, includes the national standards organizations of the 18 countries of the EU and EFTA and the Czech Republic.

DIN officially represents the Federal Republic of Germany as a member of CEN (office in Brussels). A European Standard (EN) is developed with general consent, and is adopted by qualified majority vote. It must then, irrespective of the way in which the individual member country has voted, be included unchanged in all national standards, and becomes a DIN – EN Standard in the Federal Republic of Germany. National standards that conflict with it must be withdrawn in all CEN member countries. During the development phase of an EN under the umbrella of CEN, the so- called Standstill Agreement applies to the national standards may be written or published on the subject of this EN.

To avoid trade restrictions that could arise through national technical rules, CEN, in cooperation with EU and EFTA, provides an information service for standardization work carried out in the European countries (Guideline 83/EEC – Information Service on Standards and Technical Regulations, updated in 1998). This openness about national standardization plans enables possible conflicts between member countries to be largely avoided, and allows coordinated standardization programs to be set up. Moreover, following an agreement on cooperation made in 1984 between the EU, EFTA, and CEN/CENELEC, the European Standardization Committees of the EU and EFTA are mandated to carry out special standardization work. In applying the above mentioned “New Concept” to the field of technical harmonization and standardization, the results of this standardization work form the necessary supplementation for the completion of the EU Guidelines and their incorporation in national law. The European standardization work on air quality is an important example of this.

In March 1990, the Commission for the Prevention of Air Pollution (KRdL) in VDI and DIN proposed to CEN a technical committee “Air Quality”. In October 1990, this application was approved by the CEN member organizations. The office of the KRdL was empowered to take over the tasks of the secretariat. The inaugural meeting of CEN/TC 264 “Air Quality” took place in March 1991 in Bonn.

The scope of CEN/TC 264 “Air Quality” was defined at the inaugural meeting in Bonn as follows: “Standardization of methods for the characterization of air quality with respect to emissions, pollution, air within buildings, gases in the soil and emitted by it, and waste dumping, especially:

- the establishment of uniform measuring procedures for air pollutants (e.g., particles, gases, odors, and microorganisms) and methods for determining the effectiveness of systems for cleaning waste gas, but excluding:
- the establishment of limit values for pollutants
- work areas and clean rooms
- radioactive substances.”

The subject of “air within buildings” is expressly mentioned in the definition of the scope of CEN/TC 264. This topic will become even more important in connection with the EC building materials guideline. The EU Guideline on the incineration of hazardous waste is of special relevance to the European standardization work for the prevention of air pollution. The object of this EU Guideline is “to preserve the environment and improve its quality, and to contribute to the protection of human health”. In contrast to the former EC Guidelines on domestic waste, this EU Guideline sets stringent requirements which must be adhered to by all EU Member States. These stringent requirements prescribe specific techniques of both pollution limitation and measurement to achieve the low limit values. The guideline thus sets new standards in the field of EU Regulations for the prevention of air pollution.

Technical details of measurement are not discussed in the guideline itself. Instead, for information on measurement techniques, reference is made to European Standards (EN), or, if these are not available, to ISO Standards.

The measurement techniques used in existing national standards were often designed for considerably higher emission concentrations than those to be expected when European pollution reaches the low levels predicted for the future in the above EU guideline. There is therefore a need to review procedural parameters or even to develop completely new reference methods. Thus the CEN/TC 264 Technical Committee is required in many cases to provide improvements to measurement techniques or to carry out comparative testing of existing methods. For this, the standardization work for verification of procedural parameters will be accompanied by extensive European measurement programs (interlaboratory

testing). For example, in association with the standardization work, elaborate comparative tests of various methods of measuring dioxin levels at domestic waste incinerators in Germany and Austria have been carried out. Six institutes from EU and EFTA countries contributed to these investigations. The measuring methods for threshold/guideline value monitoring of 0.1 ng/I-TEq m^3 for the dioxins or 10 mg/m^3 for HCl (24-h average 5 mg/m^3) must be validated. These two projects were mandated and financially supported by the EU (100 % support for the measurements, 50 % support for preparing the standards). This financial support ensured that the work was carried out in the time set by the EU (1996).

4.5.3. *International Regulations*

The ISO (International Organization for Standardization), founded in 1947 and based in Geneva is a worldwide association of national standards organizations. Only one national standards organization of any country can be a member of ISO at any one time. Over 70 % of the national member organizations of the ISO are linked with the government, or become so linked by incorporation into the law of the member country. The national member of ISO for Germany is DIN. The results of the work of ISO are published as International Standards (ISO Standards). Standardization for electrical technology is covered by the IEC (International Electrotechnical Commission).

The work of ISO worldwide is carried out by ca. 20 000 experts in over 200 technical committees (TCs). Each TC has a secretariat which is based at one of the national member organizations. Important environmental protection secretariats of ISO are assigned to DIN.

Up to 1998, ca. 11 000 ISO Standards have been published. An ISO Standard is the result of the cooperative work of experts in the individual national member organizations, and is carried out according to fixed rules. It can be valid both as an ISO Standard in its own right and after conversion into a national standard (e.g., a DIN – ISO Standard).

Every year, ca. 10 000 working documents worldwide are sent for further processing in the national working committees for agreement and to obtain the national votes. Meetings in the international area are called as necessary. ISO Standards are reviewed every five years for their validity or in case of scientific or technological developments. The ISO/TC 146 “Air Quality” was established in 1971. The secretariat was taken over by the office of the Standards Committee for the Prevention of Air Pollution (N Luft) of DIN. As the “KRdL of the VDI and DIN” was formed in 1990 (see Section National Regulations), the office of the KRdL in the VDI and DIN in Düsseldorf took over the secretariat of the ISO/TC 146.

The scope of the ISO/TC 146 “Air Quality” is as follows: “Standardization in the area of air quality including definition of terms, air sampling, measurement techniques, and reporting of characteristic values for air. Limit values for air quality are excluded.” The ISO/TC 146 comprises ca. 40 working groups from which four subcommittees (SCs) are formed. The meeting of the ISO/TC 146 in the autumn of 1993 in Philadelphia set up two new subcommittees for “Indoor Air” and “Meteorology.”

To prevent trade restrictions in third party countries, close cooperation and the performance of standardization work by harmonization of the standardization programs of CEN and ISO are extremely important. The cooperative work between ISO and CEN is therefore regulated by the Vienna Agreement of July 1991. Also, the cooperative work of the national standardization organizations with CEN and of CEN with EU/EFTA is carried out according to fixed rules and on the basis of formal agreements. New European standardization projects can be proposed by any CEN/CENELEC member, by any European organization, and by the Commission of the European Union.

[Top of Page]

5. Air Pollution Monitoring

Peter Bruckmann and Hans-Ulrich Pfeffer

Air pollution monitoring is the systematic measurement of impurities in the atmosphere at locations where harmful and/or adverse effects on “acceptors” such as humans, animals, plants, or materials may occur (immission measurements). Atmospheric impurities are trace substances that are not natural components of the atmosphere [128] or substances whose concentration in the atmosphere deviates significantly from the natural level as a result of human activities [129].

Assessment is made by comparing measured levels of substances (e.g., concentrations, m/V; amounts deposited $m\,A^{-1}\,t^{-1}$, or doses, m (harmful substance/m (body weight) × t or $m\,V^{-1}\,t^{-1}$) with limit values based on their effects. These limit values are specified to protect acceptors from harmful environmental effects and to maintain natural living conditions [130], [131], [132], [133-137].

However, attempts to provide the fullest possible monitoring of air pollution in a given area encounter the following main difficulties:

1. In densely populated and industrial areas, the number of hazardous substances that can be emitted into the atmosphere and are listed in the emissions registers can run into hundreds. These substances have very different chemical, physical, and toxicological properties, and their concentrations can extend over 12 orders of magnitude (from mg/m^3 to fg/m^3). This precludes any simple, quick measurement and assessment.
2. The dynamics of the atmosphere and the variation in the emissions over space and time lead to large variations in the concentration of atmospheric impurities over space and time. The measurement of all possible air quality conditions in an area (statistically, the “population”) is not practicable, so that measurements of pollution always have the character of spot checks.

The above difficulties have led to the development of a wide range of specific measuring techniques and make detailed measurement planning necessary.

In this chapter, both measuring methods and measurement planning are described for the most important fields of air pollution monitoring. Related topics such as the monitoring of air in buildings and the measurement of ground air are not discussed. Reviews can be found in [138], [139].

5.1. Measurement Planning

5.1.1. *Types of Air Pollution Monitoring*

The starting point of every measurement plan is the precise definition of the questions to be addressed by the results. Examples of important pollution measurement tasks are:

a. General air pollution monitoring [132], [133], [137] for the assessment of effects on public health and/or the environment and for compliance checking
b. Measurements for the authorization of industrial facilities [140]
c. Real-time measurements of acute episodes of high atmospheric pollution (winter smog [133], high ozone concentrations [135])
d. Orientational (screening) measurements [141], [142]
e. Measurements following complaints around industrial plants
f. Source identification and source apportionment [143], [144]
g. Measurements in cases of fire and accidents [145]
h. Measurements for research purposes

Each of these tasks requires its own dedicated measurement plan.

The most important area of activity is the assessment of air quality in given areas for compliance checking. This applies in particular to countries that have imposed limit values on pollution levels [130], [133-135]. In their antipollution policies, these countries usually employ the dual strategy of imposing limits on both emissions and levels of environmental pollution. Limit values for air quality are intended to reduce both harmful effects on the environment and hazards to health (smog), and can also have preventive effects [132]. If these limit values are exceeded, emission reduction must follow. They are therefore an element of pollution control.

5.1.2. *Planning of Pollution Measurements*

The measurement of environmental pollution should be planned in such a way that the results will answer at minimum costs all the questions they are designed to meet with given requirements for the uncertainty (accuracy) [146], [153], [133], [154] and representativeness of the data. The measurement plan therefore includes an assessment of technical, financial, and personal aspects. The important parameters that must be fixed in any measurement plan are listed in Table 6 [147].

Table 6. Parameters of pollution measurements to be established for a measurement plan [147]

Parameter	Topics which have to be addressed
Problem analysis	questions to be answered by the measurements, background information, available data, required uncertainty, pollutants to be assessed
Organization	project leader, qualification of personnel, time schedule, available budget
Measurement methods	requirement of standardized methods, performance characteristics (e.g., detection limit, temporal resolution), documentation and data processing, costs
Measurement strategy	measurement area; spatial and temporal resolution, sampling periods and measuring times; sites of the measurements; sampling height
Data evaluation and reports	algorithms to calculate characteristic values, treatment of values below the detection limit and of data gaps, treatment of outliers, structure of report(s)
Quality assurance	quality assurance measures, also for subcontractors; determination of measurement uncertainty

5.1.2.1. *Measurement Networks (General Air Pollution Monitoring)*

The requirements of the European directives [132], [133-137] as one outstanding example for general air pollution monitoring can be summarized as follows:

- Air quality must be assessed over the whole territory of a Member State.
- Air quality for the protection of human health shall be monitored:
  - in areas which are representative for the exposure of the general population (urban background)
  - in areas where the highest concentrations occur to which the population may be exposed for a significant period (hot-spot locations)
- Air quality for the protection of ecosystems shall be monitored with a spatial resolution of at least 1000 km^2.

Bearing in mind the large variation of atmospheric pollutants over space and time and the great number of possible hot-spot locations (busy streets, surroundings of industrial facilities), it is apparent that these ambitious requirements cannot be achieved by measurements alone. A combination of continuous measurements with other assessment techniques such as diffusion modeling, indicative spot-check measurements, analysis of emission inventories, and intelligent data

interpolation is called for, to get at least an approximation of the real air quality. Guidance has been published for the design of monitoring networks [156] and for the assessment of air quality [157].

The recommended approach uses a stratification of measurement locations. The population of possible measurement locations within the area to be investigated is divided into strata of elements with common characteristics [155]. Typical strata of measurement locations, or microenvironments, are urban background areas, sites exposed to traffic, rural areas, and industrial sites. One or more measurement stations are then placed at those locations which are representative for the pollution burden of these typical microenvironments. The monitoring results of these stations can, at least in principle, be transferred to similar locations of the same microenvironment.

This exposure-related approach is also often used in health-related monitoring studies [158]. Information on the position and extent of the microenvironments can be obtained from the analysis of emission inventories, geographical information (location of busy streets combined with traffic density and adjacent built-up areas), and diffusion modeling.

One of the problems this approach must cope with is the widely differing spatial structure of the air pollution, which makes different spatial resolutions of the assessment necessary. Whereas air quality in rural areas is quite homogeneous (a spatial resolution of several hundred or thousand square kilometers is sufficient), pollution levels within a busy street canyon and in a backyard only 50 m apart can differ by up to a factor of two. Consequently, the necessary spatial resolution of the assessment comes down to several hundred square meters. Urban background locations in typical residential areas are in between, with necessary spatial resolutions of several square kilometers. Also industrial sources with fugitive emissions such as cokeries or steel mills [144] can have a high spatial gradient of their air pollution, and a spatial resolution of 250 × 250 m^2 is recommended for assessment [136], [157].

The measurement uncertainty caused by the dynamics of the atmosphere and the emissions can be reduced to the uncertainty of the spatial distribution by using continuous monitors (see Automatic Measuring Methods). In addition, continuous monitors for SO_2, NO_2, or ozone can be used to survey smog-alert thresholds [133], [135]. However, for pollutants such as most organic species, heavy-metal compounds, or other contents of suspended particulate matter, discontinuous sampling with subsequent analysis in the laboratory must be applied. Consequently, in addition to the uncertainty caused by the spatial variability of air pollution, a second part of the uncertainty budget caused by insufficient time coverage of the spot-check sampling comes into play. Statistical methods have been developed to estimate this uncertainty and to determine the necessary number of measurements to obtain a given uncertainty [159], [160].

The results of diffusion modeling, particularly in the case of busy streets, and of indicative measurements with comparably cheap equipment such as passive sampling [161], [142], [162], can be used to find locations for monitoring stations which are representative for the microenvironment in question.

To safeguard a comparable monitoring approach in Europe, the directives [133-136] require a minimum number of fixed sampling sites, which depend on the pollution burden (monitoring efforts must be higher near or above the limit values than at low pollution levels). In addition, the number of costly monitoring stations can be reduced if additional assessment tools such as diffusion modeling or an analysis of emission inventories have been applied.

The ultimate aims of this approach are pollution maps. However, the uncertainty of this spatial representation of the pollution burden will depend inter alia on two factors:

- How complete have hot-spot locations been detected and taken into account?
- How uncertain is the transfer of typical monitoring results within a microenvironment to other locations of the same type of microenvironment?

Bearing in mind the limited amount of technical and human resources available for general air pollution monitoring, it is evident that a good approximation of the air pollution over a given area is all that can be hoped for. Statistical methods [148], were developed to assess the standard error expected when figures determined at the measurement points are used to estimate corresponding figures at points in the surrounding area.

An alternative approach to designing monitoring networks starts with the aim of describing the pollution burden in a given area as representatively as possible, irrespective of the location of specific microenvironments or goods to be protected. The principal idea of this approach is to arrange the measurement sites in a regular grid, e.g. 1 × 1 or 4 × 4 km. If this regular structure of the measurement network is independent of the structure of the investigation area (e.g., location of streets, residential or industrial areas), the grid arrangement ensures a representative sample of all theoretically possible measurement sites [148]. This strategy makes allowance for the fact that air pollution monitoring of areas necessarily has the character of spot check measurements, at least with respect to the spatial distribution of air pollutants. The density of measurement points on the network determines the resolution of the system; hence structures smaller than the network unit area (e.g., 1 km^2) cannot be detected (e.g., points of high pollution in streets). The frequency of sampling at the measurement points determines the temporal resolution. If continuous measurement is performed (see Section Automatic Measuring Methods), the temporal resolution can be practically perfect, except for periods when the equipment is out of commission.

When using random measurements, not only must the size of the statistical sample be adequate, but also measurement times must be randomly selected to ensure accurate representation of the time dependence. As atmospheric pollution often shows marked variation during the day and correlation with the seasons, measurements only at certain times of the day or certain times of the year would not give representative results.

Random sample measurements are particularly well suited for the determination of annual averages or percentiles not higher than 98 %. For these characteristic values, the additional uncertainty introduced by cost effective random sampling

can be kept reasonably small [163].

Outstanding examples of this basic concept of grid measurements combined with random sampling were the pollution measurements prescribed for the German licensing procedure before 2002.

For general air pollution monitoring larger spacings of the network were chosen, with continuously operating measurement stations 4 – 16 km apart. Several German air pollution networks were designed in this way, e.g., for the Rhine-Ruhr area [148], but were restructured following the “stratification approach” of the European air quality directives [132-137] after 1998.

5.1.2.2. *Evaluation of Data; Period and Frequency of Measurements*

The choice of measurement period and frequency and the evaluation of the results depend mainly on the definition of the assessment criteria (limit values) used. The assessment and measurement systems must complement each other. If the criteria are related to a one-year period, the measurement period must be the same. If, however, alerts or short-term abatement measures are based on periods of 1 – 3 h (e.g., as in the Smog Regulations of the EC Directives [133], [135]) the temporal resolution of the measurements must be smaller than the assessment interval, which makes continuous measurement necessary.

For regulatory purposes (e.g., comparison with limit values), it is generally sufficient to derive two characteristic figures from the statistical sample of measured values: the average pollution (arithmetic mean or median) and the peak pollution (e.g., a 90 or higher percentile). Very high percentiles such as the short-term limit values for SO2 and NO2 in the EC legislation [133] afford continuous measurements, whereas annual means, for example, of benzene [134] or heavy metals [133], [136], can be monitored more economically by spot-check measurements. The additional uncertainty introduced by the incomplete time coverage of the monitoring data can be estimated [159].

The question whether and how the uncertainty of measurement of pollution (e.g., expressed as a confidence interval with a given confidence level) should be taken into account when comparing measurements with limit values is the subject of controversy. An ISO Guideline contains recommendations on this subject [149]. The results of immission measurements are generally compared directly with the limit values.

However, the data-quality objectives for the measurement uncertainty, for minimum time coverage and minimum data capture prescribed in the EC legislation [133-137] limit the overall uncertainty of the measurements.

5.1.2.3. *Planning Measurements on Industrial Plants*

Careful measurement planning is needed for monitoring of air pollution near industrial facilities [164], such as measurements specifically relating to cause analysis in connection with complaints or with plans and programmes to abate air pollution [144]. Such planning can greatly reduce measurement costs compared with schematic use of measurement grids. The assessment of the validity of complaints or of the results of abatement measures for single sources often requires one or two measurement points only, these being located at the place specified by the complainant or the point of the expected maximum pollution and exposure when determining peak and average figures due to a particular source. Such an exposure-related approach is also used to measure the air pollution in the vicinity of a planned or existing industrial facility as part of the German authorization procedure [140]. Generally, two measurement points are selected which are representative for the areas of maximum exposure (combination of background levels and surplus burden of the industrial source calculated by diffusion modeling). One monitoring point should reflect the highest average exposure, and the other the highest peak levels. For cause analysis and the identification of sources of pollution, simultaneous measurements are made upwind and downwind and also evaluated as a function of wind direction. Methods of multivariate statistics, fingerprint or cluster analysis may be used in the case of multicomponent measurements for source identification and quantification [143, 150, 151]. Especially when sources emit several substances simultaneously, with a characteristic emission pattern (e.g., ratios of different hydrocarbons or heavy metals to one other), analysis of the pollution pattern, e.g., as a function of the wind direction, can often enable the main cause or causes to be determined, and in favorable cases even the approximate relative immission contributions [143].

5.2. Quality Assurance

Ambient air measurements can be very costly with respect to capital investment, management, and personnel. Based on the results, very far-reaching decisions can often be made, with considerable economic or ecological consequences. Therefore, quality assurance forms an inseparable part of the measurement system. Quality assurance can be divided into two different categories, which are described in more detail in Sections Procedure-Based Methods and User-Based Methods.

5.2.1. ***Procedure-Based Methods***

Every comprehensive ambient air monitoring method consists of several steps: sampling, sample preparation and cleanup, isolation of the pollutant, detection, data collection, and data evaluation. In a few cases, especially in the continuous measurement of inorganic gases, certain steps of the sequence can be omitted (e.g., separation of the pollutants).

In most types of pollution measurement, extensively documented, comprehensive procedures, covering all steps from sampling to data evaluation and standardized in DIN – ISO Standards, CEN Standards, or VDI Guidelines [152], are recommended (see Chap. Measuring Methods).

For many years, an important part of the documentation of measuring procedures has been the specification of performance characteristics such as

1. Detection limit: the smallest concentration or amount of a substance that can be distinguished from zero with a given probability (e.g., 95 %)

2. Trueness: the degree of agreement between a measured value and the “true” value or an accepted reference value
3. Precision: the degree of agreement between values measured under given conditions by the repeated application of a measuring procedure
4. Measurement uncertainty: a comprehensive indicator for possible deviations of a measurement result from the “true” value (see below)

There are various definitions of these and other statistical concepts. Details can be obtained from [153, 155, 165].

In the past, the comprehensive VDI Guidelines have been of the greatest importance in the standardization of pollution measurement in Germany. However, as the EU became more and more firmly established, the importance of European and international standardization grew rapidly.

The main reason for this development was the new European legislation [133-137], which defines so-called data-quality objectives for minimum time coverage, minimum data capture, and, especially, for maximum measurement uncertainty. As a consequence, all newly developed standards describing reference methods laid down in the EU directives mentioned above must deal explicitly with the evaluation of measurement uncertainty in order to demonstrate that these data quality objectives can be met. The fundamental concept for these evaluations is given by the “Guide to the Expression of Uncertainty in Measurement” (GUM) [146]. The GUM describes a very abstract concept, so many papers were published to give practical guidance for using this concept [154], [166], [167]. But according to European legislation it is not mandatory for the Member States to use the reference methods standardized by CEN, because all directives contain a sentence like this: “A Member State may also use any other methods which it can demonstrate gives results equivalent to the above (reference) method”. A guidance document developed by a European working group describes how equivalence of monitoring methods with the reference method within the required maximum measurement uncertainty may be demonstrated [168]. This concept was recommended for use by the European Commission.

All automated instruments for the measurement of ambient air pollutants should be type-approved. The procedures for this type approval are described in various standards, e.g, in Germany [169-171] and in the UK [172]. Some of the new CEN standards for inorganic gases [173-176] and for benzene [177] also contain corresponding regulations.

5.2.2. *User-Based Methods*

The use of standardized and tested equipment and procedures for the measurement of pollution cannot of itself ensure an acceptable quality of measurement if adequate systems of quality management are not operated by the user.

The ISO 9000 family of standards has been developed to assist organizations in implementing and operating effective quality management systems [178], [179]. Together they form a coherent set of quality management system standards facilitating mutual understanding in national and international trade. But conformity of the quality management system within which a laboratory operates to the requirements of ISO 9001 does not of itself demonstrate the competence of the laboratory to produce technically valid data and results.

For that reason, in the field of ambient air measurement the standard ISO/IEC 17025 (General requirements for the competence of testing and calibration laboratories) is of much more relevance [180]. Over several years this standard and those of the ISO 9000 series were harmonized. That means that testing and calibration laboratories that comply with the 17025 International Standard will therefore also operate in accordance with ISO 9001.

The 17025 standard specifies the general requirements for the competence to carry out tests and/or calibrations, including sampling. It covers testing and calibration performed using standard methods, nonstandard methods, and laboratory-developed methods.

The standard consists of two main parts. The part “management requirements” deals, e.g., with organization, document control, contracting and subcontracting, customer service, preventive and correcting actions, control of records, and auditing. “Technical requirements” include personnel, environmental conditions, test and calibration methods and method validation, estimation of measurement uncertainty, data control, measurement traceability, use of reference standards and materials, and reporting.

Many technical tasks in implementation of the European air quality directives may be performed only by laboratories working in accordance with this international standard. A formal accreditation based on ISO/IEC 17025 is internationally recognized as proof of competence.

Another important element of quality assurance is the national reference laboratories (NRL), which had to be notified by all European Member States within the implementation of the air quality directives. The NRL must coordinate and harmonize quality assurance measures on national level and participate in European programmes organized by the European Commission. The NRL are organized in AQUILA, the Network of Air Quality Reference Laboratories [181]. AQUILA includes the EFTA countries and the EU candidate countries.

A further important element of quality assurance is interlaboratory comparison, [182], [183]. In air pollution measurement technology, this is mainly carried out with the aid of air sample distribution equipment [184], [183] which enables standard gas mixtures and aerosols of widely different compositions and qualities to be produced and supplied. A number of participating laboratories take samples repeatedly from the same gas/aerosol supply and analyze them chemically. These data are then statistically analyzed, enabling important conclusions to be drawn regarding measuring procedures and participants (measurement uncertainty, accuracy, precision, repeatability, comparability). Interlaboratory comparisons are performed on national levels [182], [183] as well as on international level, e.g. in the framework of AQUILA [185].

Also, interlaboratory testing is an important element of notification by the Länder of institutes to carry out measurements in accordance with § 26 of the Air Pollution Control Act (BImSchG). This is a special variation of an accreditation procedure

based on the International Standard ISO/IEC 17025.

Last but not least, measures of quality assurance and quality control are an extremely important element in routine monitoring. As well as the systematic repair and maintenance of all measuring equipment and rectification of faults, regular calibration and monitoring of functions are essential. Many of those measures are mandatory within the implementation of the new CEN standards described in Measuring Methods [173-176].

5.3. Measuring Methods

5.3.1. *Measuring Methods for Inorganic Gases*

A very wide range of automatic and manual measuring methods is available for inorganic gases [152, 173-176], [186], [187], [188].

5.3.1.1. *Automatic Measuring Methods*

A large number of continuously measuring monitors are available for the substances sulfur dioxide, nitrogen oxides, carbon monoxide, and ozone. These are very widely used, especially in monitoring networks. As a consequence of the European Directives on ambient air quality [133-135], these methods were standardized by CEN [173-176]. The main parts of these standards describe a procedure of systematic type approval (see also Procedure-Based Methods). The German Ministry for the Environment, Nature Conservation, and Nuclear Safety declares on a regular basis that certain types of equipment, produced by various manufacturers, are suitable for air pollution measurement [188], [189]. The main features of the individual methods are listed in Table 7.

Table 7. Automatic measurement methods for inorganic gases

Substance	Measurement principle	Comments	Refs.
SO_2	UV fluorescence; excitation of SO_2 molecules by UV radiation (200 – 220 nm); measurement of the fluorescence radiation (240 – 420 nm)	spurious reactions (e.g., with hydrocarbons) must be eliminated by filters	[174]
NO/NO_2	chemiluminescence; reaction of NO with O_3 to form NO_2; detection of chemiluminescence radiation by photomultiplier	NO_2 only indirectly measurable; use of converters to reduce NO_2 to NO; measurement of NO and NO_x (= NO + NO_2) in separate measurement channels; converter efficiency must be determined	[173]
CO	nondispersive infrared absorption; measurement of IR absorption by substance-specific detector with reference cell	reference: a) cell with N_2; b) sample air after selective removal of CO by catalytic/chemical reactions; c) optical removal of the CO absorption (gas filter correlation)	[176]
O_3	UV absorption; stabilized low-pressure Hg lamp as radiation source (253.7 nm); recording of the UV absorption according to the Lambert – Beer law; detector: vacuum photodiode, photomultiplier	single- and double-cell instruments; reference: air sample after selective removal of the ozone by chemical/catalytic reactions (e.g., Cu/MnO_2)	[175]

According to the regulations foreseen in the European Directives, measurement results generally must be reported at 20 °C and 101.3 kPa. Therefore, most automated instruments are equipped with sensors measuring temperature and pressure in the measuring cell for appropriate correction.

Another important gas for which an automated measuring method is available for is ammonia. In a multistep process NH_3 penetrates Teflon membranes and is finally dissolved in purified water. The increase in conductivity of this solution is a measure for the ammonia content in the air [190]. A commercial instrument is available.

All automatic measuring methods based on physicochemical principles must be calibrated using calibration gases traceable to national standards. This even applies to the measurement of ozone by UV absorption, although this in principle is itself a reference method under controlled conditions. Experience shows that knowledge of extinction coefficients at 253.7 nm coupled with the validity of the Lambert – Beer law does not ensure that measured values will be automatically correct. For example, soiling of the gas passages, malfunction of the internal ozone scrubber for the quantitative decomposition of the ozone in the reference measurement cycle, or electronic defects can lead to significant false readings.

5.3.1.2. *Calibration of Automatic Measuring Methods*

All automated instruments for the measurement of inorganic gases must be calibrated using calibration gases traceable to national standards. The new CEN standards defined as reference methods for the implementation of the European Directives on air quality [173-176] require such calibration at least every three months. Various methods for the generation

of calibration gases are described in ISO standards or German VDI guidelines [191-202], [228]. The appropriate application of these methods requires specific equipment and experienced staff. In calibration laboratories for ambient air measurement especially the following methods are used in practice: static dilution [190], thermal mass flow controllers [198], diffusion [199], and permeation [202]. In addition, certified gas standards are commercially available in pressure cylinders.

Automatic measuring equipment is mainly used in measurement stations, and these often form a part of a pollution measurement network (see Section Measurement Networks (General Air Pollution Monitoring)).

As the equipment is operated under these conditions for long periods of time (sometimes several weeks) without direct personal supervision, it is essential to monitor their correct function by using an automated procedure on a regular basis. An essential requirement for this supervision is the availability of test gas generators, which supply blank samples and standard gases to the measuring equipment.

A blank sample gas is one in which the concentration of the substance to be measured and for which the equipment is being calibrated is present at only trace concentrations, below the detection limit of the measuring method. Blank sample gases must not contain substances that react with the component to be measured or interact in any other way. A test gas is a gas mixture which usually consists of a carrier gas (nitrogen, air) with one or more added components [186-188], preferably produced by dynamic test gas generators [228].

Nitrogen or synthetic air in pressurized gas cylinders can also be used as blank gases. Such blank gases are relatively easy to handle, but often have the disadvantage in pollution technology that their properties differ from those of the matrix air being analyzed (e.g., in moisture content). There is also the necessity for preparing and changing the gas cylinders. Therefore, blank gases are preferably prepared immediately before use from the air to be analyzed, e.g., by filtration through sorbents.

For many substances, standard gases can be obtained commercially in pressure cylinders. If the gas cylinders contain the standard gas in the concentration required, dilution is unnecessary. However, with such standard gases there are problems of high gas consumption and sometimes poor storage stability, which usually decreases with decreasing concentration. Alternatively, gas cylinders can be used in which concentrations are 50 – 100 times as high as that required in the standard gas. In this case, controlled dilution must be carried out with the base gas. A standard gas generator operating on this principle is more expensive, and the dilution stage is a additional source of measurement uncertainty. Advantages include the generally better storage stability and lower rate of consumption of the bottled gas.

The standard gas mixture can also be prepared by permeation, or, in the case of nitrogen oxides, by gas-phase titration [229], [230]. The most important methods used for preparing standard gas mixtures containing components measured in pollution measurement networks on a routine basis are indicated in Table 8.

Table 8. Production of test gases for calibration

Component		Atmosphere	Pressure cylinder (direct)	Pressure cylinder (with dilution)	Permeation	Gas-phase titration
SO_2	blank gas	x	x			
	test gas		x	x	x	
NO	blank gas	x	x			
	test gas		x	x		x
NO_2	blank gas	x	x			
	test gas		x	x	x	x
CO	blank gas	x	x			
	test gas		x	x		
O_3	blank gas	x	x			
	test gas	x				

According to the CEN standards [173-176] all these transfer standards used as test gases in field measurements must be checked at least every six months using calibration gases traceable to national standards. Zero and span checks have to be performed at least every two weeks; 23 or 25 h are recommended.

5.3.1.3. *Manual Methods*

In manual measuring methods for inorganic gases, the substance to be measured is generally absorbed in a solution or adsorbed on a solid material during sampling. In most cases, this is followed by a color reaction and photometric determination. Because the personnel requirement is comparatively high, these are seldom used for field measurements if alternative automatic methods are available. The most important methods are briefly outlined in Table 9.

Table 9. Manual measurement methods for inorganic gases

Substance	Method	Details	Comments*	Refs.
SO_2	TCM method	absorption in tetrachloromercurate solution (wash bottle); reaction with formaldehyde and *p*-rosaniline to give red-violet sulfonic acid; photometric determination	DL = 0.2 µg SO_2; s = 3 µg/m^3 at 140 µg/m^3	[203], [204], [205]
NO_2	Saltzman method	absorption in reagent solution to form a red azo dye (wash bottle); photometric determination	calibration with sodium nitrite; DL = 3 µg/m^3;	[206], [207], [208]
F^-	silver bead method; variant 1	sampling with dust removal; concentration of F^- on silver beads coated with sodium carbonate; elution and measurement with ion-sensitive lanthanum fluoride electrode	simultaneous detection of an unknown proportion of particulate fluoride pollutants	[209]
F^-	silver bead method; variant 2	sampling with heated membrane filter; concentration of F^- on silver beads coated with sodium carbonate; determination by electrochemical method (variant 1) or photometrically (alizarin-complexone method)	danger of low results due to partial sorption of gaseous fluoride pollutants on membrane filter; DL = 0.5 µg/m^3	[210]
Cl_2	methyl orange method	decolorizing reaction with methyl orange solution (wash bottle); photometric determination	DL = 0.015 mg/m^3	[212], [213]
NH_3	indophenol method	absorption in dilute H_2SO_4 (impinger/wash bottle); reaction with phenol and hypochlorite to form indophenol dye; photometric determination	DL = 3 µg/m^3 (impinger); partial simultaneous detection of NH_4 compounds and amines	[214]
NH_3	Nessler method	absorption in dilute H_2SO_4 (impinger/wash bottle); distillation and reaction with Nessler's reagent; photometric determination	DL = 2.5 µg/m^3 (impinger); partial simultaneous detection of NH_4 compounds and amines	[219]
NH_3	Denuder with H_3PO_4 coating	flow injection analysis	DL = 0.1 µg/m^3	[216]
NH_3	Denuder with citric acid coating	analysis with indophenol method	DL = 0.1 µg/m^3	[217]
H_2S	molybdenum blue method	sorption as silver sulfide on glass beads prepared with silver sulfate and potassium hydrogen sulfate (sorption tube); liberation as hydrogen sulfide and reaction to form molybdenum blue; photometric determination	DL = 0.4 µg/m^3	[218]
H_2S	methylene blue method	absorption in cadmium hydroxide suspension to form CdS; liberation of H_2S; reaction to form methylene blue; photometric determination	DL = 0.3 µg/m^3	[219]

* DL = detection limit; *s* = standard deviation.

For several decades, sulfur dioxide was regarded as the most important atmospheric pollutant. Consequently, this substance was included in most pollution measurement programs, and some methods for measuring pollution by sulfur dioxide exist, e.g., the tetrachloromercurate (TCM) method [204], [205]. However, due to significant decrease of SO_2

concentration in air in most parts of Europe, these methods are of minor importance today.

The Saltzman method [206-208] was an important manual method for nitrogen dioxide in the past.

Samples for measuring pollution by hydrogen fluoride and fluorides can be taken by sorption tubes [209], [210], special filters, or denuders (diffusion separation tubes). Analytical determinations are mainly based on photometric, electrochemical, or ion chromatographic methods [186], [211].

Pollution by chlorine and chlorine compounds is seldom measured in practice, as chlorine itself occurs very rarely and in very few locations because of its high reactivity.

Ammonia in the atmosphere is in chemical equilibrium with ammonium compounds, so that NH_3 and NH_4^+ compounds are often determined together. For a long time, the most important manual methods were the indophenol method [214] and the Nessler method [215].

A special sampling method, mainly used in conjunction with manual measurement, is based on the diffusion separation tube (denuder), which separates the gaseous phase from the particulate phase by means of different diffusion rates. It is therefore mainly used for difficult separation problems involving systems of this type [220]. In the classical denuder technique, the air sample is drawn through a glass tube with a coating which depends on the substance or substances to be determined [221-224]. For example, in the determination of basic components such as ammonia and ammonium compounds, tubes coated with an acidic substance (e.g., H_3PO_4, oxalic acid or citric acid) are used [216], [217]. The gaseous substances are deposited and fixed on the tube wall by chemical reaction, while the particles mainly pass through the tube and can be collected on a filter.

Many variations of the denuder technique have been developed, some of which have been automated (annular gap denuders, wet denuders containing solutions of reagents, thermodesorption denuders) [225], [226]. Denuders are the preferred method for sampling systems such as ammonia/ammonium compounds, nitrogen oxides/nitric acid/nitrates, sulfur oxides/sulfuric acid/sulfates, or hydrogen halides/halides. The denuder technique has much extended the possibilities of taking differentiated samples. However, depending on the particular version of the method, it needs a comprehensive laboratory, and correct use of the technique requires great experience.

5.3.2. *Measurement and Analysis of Particulate Atmospheric Impurities*

For the assessment of particulate atmospheric impurities, not only their concentration — expressed as mass per volume or number of particles per volume — but also their size distribution is of paramount importance. In addition, parameters such as the surface area and the composition of the aerosol are currently strongly discussed in relation to health effects [227].

The size distribution of particles in ambient air varies from aerodynamic diameters [231] of a few nanometers to more than 100 µm [232]. The very coarse particles at the upper end of this distribution are deposited quickly on the ground by virtue of their size and density and are therefore not homogeneously dispersed in ambient air; this fraction is characterized as deposited dust (Section Measurement of Total (Bulk) Deposition). The atmospheric particles, which are quasihomogeneously dispersed, are classified as suspended particulate matter (SPM).

The borderline between the two regimes cannot be exactly drawn and depends in a complex fashion on many exterior conditions. However, as a rule of thumb it can be stated that suspended particulate matter mainly consists of particles smaller than 30 µm [233] in aerodynamic diameter.

5.3.2.1. *Measurement of Particle Mass and Particle Fractions*

How deeply particulate matter (SPM) may penetrate into the human respiratory system depends mainly on the size of the particles. Experiments have shown that all fractions of SPM are inhalable (inhalable fraction) [231], but that coarser particles with aerodynamic diameters > 10 µm are mainly deposited within the nose and throat [234]. Particles smaller than 10 µm can penetrate beyond the larynx (thoracic fraction) [231], whereas particles finer than about 4 – 5 µm in the case of healthy adults and with median diameters below 2.5 µm in the case of children or infirm persons with narrower respiratory passages are fully respirable (respirable fraction [231]).

In health-related sampling of particles, there is no clear cut off between the different fractions, and conventions defining the size distributions of the fractions are needed. Conventions which are valid likewise for ambient air and the workplace atmosphere have been laid down in ISO 7708 [231]. The definition of inhalable more or less corresponds to SPM, the thoracic convention (PM 10) comprises particles with a median aerodynamic diameter of up to 10 µm, and particles designated as respirable for persons at high risk (PM 2.5) have a median diameter of 2.5 µm [231].

In Figure 5, these conventions are compared with measured size distributions of particles in ambient air, which are typically bimodal. The thoracic convention (PM 10) cuts the peak of the coarse particles, whereas the PM 2.5 convention fully includes the fine-particle mode.

Figure 5. Measured size distribution of particles in ambient air (Bilthoven, Netherlands) in comparison with the conventions laid down in ISO 7708

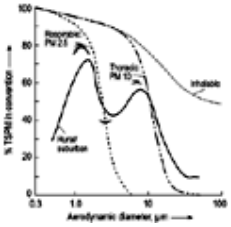

The bimodal distribution points to different processes of particle generation in the atmosphere [232]. Whereas the coarse mode is predominantly caused by mechanical processes such as the resuspension of dust from surfaces, the fine mode consists of particles which stem from nucleation and condensation.

Bearing in mind the deeper penetration of finer particles into the airways, it is not surprising that the PM 10 and particularly the PM 2.5 burden have been shown to correlate better with health effects than the concentrations of coarser particles [231], [235], although recent research has shown that the effects of coarser particles cannot be neglected [236]. Consequently, air quality standards in European countries [133] and in the USA [237] are based on particle fractions and hence afford the fractionating measurement of SPM for compliance checking.

The majority of dust measurements, both fractionating and nonfractionating, are based on a mass determination of particles on glass-fiber or membrane filters [230]. This is followed by a gravimetric determination (manual methods) or, in the case of continuously operating instruments, by radiometric measurements or oscillating microbalances (Table 10). Different validated sampling heads are used to select the particle fractions to be measured (TSP, PM 10, PM 2.5, and PM 1.0). These sampling heads use the phenomenon of impaction (separation based on inertial forces) to preseparate the coarse particles. The air stream within the sampling head is diverted in such a way that only particles smaller than the selected diameter can follow the air stream and are collected on a backup filter. The coarser particles are trapped in fractionators before the filter. The fractionators necessary for the different particle fractions can be calculated and tested with model aerosols, but the collection efficiencies of the different sampling heads do not exactly match the conventions laid down in ISO 7708 [231].

As the object of measurement — SPM or a fraction thereof — is not sharply defined, and variations in concentration, particle-size distribution, and composition can occur with time and location, there is no absolute measuring method. To a certain extent, each applied method defines its measurement object, and the interconversion of measured values obtained with different methods is only possible under carefully controlled conditions. Consequently, reference methods have been laid down which give by convention the “correct” results, and the equivalence of other methods has to be validated by test measurements in comparison with the reference procedures [168]. The European union [133] has laid down the gravimetric WRAC instrument as reference method for PM 10. A CEN standard [238] describes how validation measurements with other methods against this standard have to be performed. One low-volume sampler (Table 10) and one high-volume sampler are also described in the CEN standard as reference methods [238] for field tests after comprehensive validation measurements in comparison with the WRAC sampler. A CEN reference method for PM 2.5 measurements [239] describes the design of a high-volume and a low-volume gravimetric sampler.

Table 10. Measurement methods for SPM

Method (examples)	Measurement principle (examples)	Comments	Refs.
Low-volume sampler	fractionating and nonfractionating sampling; air flow rate 2.3 – 2.8 m^3/h (small filter equipment) or 1 m^3/h (Partisol); filter diameter 47 – 50 mm; gravimetry; volume measurement by revolving vane anemometer	low noise; easy operation due to small size and low weight; control by time-control switch; also available with PM 10, PM 2.5 and PM 1.0 sampling heads (impactors), European reference method (PM2.5)	[254], [255], [239]
High volume sampler	detection of particles < ca. 30 µm diameter; air flow rate ca. 30 – 100 m^3/h; filter diameter 130 – 257 mm; gravimetry; volume measurement by flow meter	recovery of large amounts of dust, useful for analysis of dust components; high noise level; also available with PM 10 and PM 2.5 sampling heads (impactors), European reference method (PM 2.5)	[256], [257], [239], [258]
Radiometric beta gauge instrument	continuous radiometric dust measurement equipment; fractionating and nonfractionating sampling; air flow rate 1 or 3 m^3/h; dust collection on filter band; dust measurement by attenuation of β-radiation on passing through exposed filter (ionization chambers) or Geiger counter	gravimetric calibration by known amounts of dust on single filters and intercomparison with gravimetric reference method; operation with PM-10 and PM2.5 sampling heads possible	[259], [260] [261], [262], [263]

Oscillating microbalance	continuous dust measurement equipment; fractionating and nonfractionating sampling; air flow rate 1 m^3/h; dust collected on filter (which is part of a vibrating system with resonance frequency) in side stream (3 L/min); recording of frequency decrease with increase in dust load on the filter	relationship between frequency decrease and mass of dust must be established by calibration and intercomparison with gravimetric reference method; operation with PM 10 and PM 2.5 sampling heads possible	[264], [265]
Optical analyzer	light-scattering device using a laser light source orthogonal to the sampled air stream. Particles (> 0.3 μm) are counted and classified. Variant: parallel radiometric dust measurements for in situ calibration	PM 10, PM 2.5 and PM 1.0 can be monitored simultaneously. Conversion of counts to particle mass depends on nature of particles. Suited for screening measurements if there is no on situ calibration	[266, 267]
WRAC (wide-range aerosol classifier)	fractionating sampling; four impactors (10 µm, 21.8 µm, 40.3 µm, 65.4 µm) running parallel; air flow rate 77.9 m^3/h (PM 10 stage)	European PM 10 reference method; difficult handling in the field, not for routine operation	[238]

The continuous monitoring of PM 10 and PM 2.5 fractions (either radiometric, by microbalances, or by optical methods; see Table 10) is still subject to some difficulties, because validation measurements with the reference methods have revealed deficiencies on the order of 30 wt % or more [232], [268], [269]. Hence, correction factors or equations must be established by intercomparison measurements with the reference methods, and the results have to be corrected accordingly [133], [268].

It has been demonstrated that the mass losses of PM 10 or PM 2.5 of automatic instruments compared with the reference method are mainly due to the evaporation of semivolatile aerosols such as ammonium nitrate in the heated sampling tube or on the filter [269], [270]. Some heating is necessary (typically 50 °C for the older instruments) to avoid condensation of water vapor. Also the gravimetric reference methods are not free from artefacts and may loose some semivolatiles when sampling head and tube heat up on hot summer days [232].

Various recent improvements have reduced or allowed the quantification of the loss of semivolatiles in continuously operating instruments. The instruments were modified to operate at lower temperatures. In the case of the TEOM, a diffusion dryer was placed before the filter unit to remove moisture and to enable operation of the filter at 30 °C (TEOM SES [271]). New versions of the beta gauge instruments (e.g., FH62IR or MP101M) were equipped with controlled heating devices (TRS) which keep sampling line and filter unit only several degrees Celsius above ambient temperature [267]. The differential TEOM uses two sensor units, which are alternatively switched on and off out of phase, so that alternatively one sensor is driven with particle free air, and the other with particle-burdened air [272]. The sensor in the particle-free air stream measures the “artefacts” (e.g., evaporation), and the output signal can be corrected accordingly. A further improvement uses a cooled filter unit upstream of only one sensor unit. Alternatively, the air stream bypasses the cooled filter or is cleaned from particles. The same sensor thus alternatively measures the particle mass and possible artefacts (TEOM FDMS) [273].

Discontinuous gravimetric sampling of SPM or fractions thereof is facilitated by automatic filter changers that store a large number of filters and supply them in sequence to the sampling device. The exposed filters are stored on site in a cooled magazine [274].

The black smoke method [138], [275], [276], formerly widely applied in the United Kingdom and other European countries, is a combination of filter precipitation of dust and an optical method. Air is drawn through a filter paper at a low rate (ca. 2 m^3/d), and the darkening of the filter is then measured with a reflection photometer. This black smoke value is converted into gravimetric units (µg/m^3) by means of a calibration curve [277], [278], [279]. However, this calibration curve depends on the composition of the particles (e.g., their soot content) and is both site- and season-specific. Notwithstanding these shortcomings, the results of black smoke measurements are still of value, because many epidemiological studies were based on them [280] and show a clear correlation with negative health effects.

5.3.2.2. *Measurement of Particle Numbers and Size Distributions*

Particle mass is not the only metric which correlates with health effects. Some studies suggest that high numbers of particles, especially of ultrafine particles (< 0.1 μm), may be of equal or more importance for some detrimental effects such as cardiovascular disease [240], [241]. Consequently, measurements of the size distribution and the number of particles are gaining in importance. As particle numbers are dominated by the ultrafine fraction, which on the other hand adds almost nothing to particle weight [232], ultrafine particles are generally measured as numbers.

Condensation particle counters (CPC) [242], [243], [244] measure particle numbers directly in the range of 3 nm to 2 mm. Ambient air passes through a chamber filled with supersaturated vapor (butanol or water) and then through a second cooled tube, where the fluid condenses on the particles. The particles grow to approximately the same size and can be counted by an optical detector. However, information on particle size is lost by the condensation step. Information on the size distribution can be obtained in combination with a separation unit upstream of the CPC.

Without a condensation step, optical particle counters can only be used down to a particle size of about 100 nm [245],

[246]. They measure the light which is scattered by single particles and count the events. Lasers or high-pressure Xe lamps are used as light source. The intensity of the scattered light is proportional to the particle size, so that information on size distribution and number can be obtained simultaneously (laser aerosol spectrometer, LAS [247]).

Size distributions in the range from several nanometers up to several micrometers can be measured by a scanning mobility particle analyzer (SMPS) [248], [249], [244]. The polydisperse aerosol is charged by a beta radiation source (85Kr) and then passes into a cylindrical electrostatic classifier (DMA). The wall of the cylinder serves as outer electrode, whereas the inner electrode is situated in the center of the cylinder. The inner cylindrical part of the DMA is flushed with particle-free air, and the sample air stream is introduced near the outer electrode. The potential difference between the electrodes allows only particles with a given mobility and diameter to pass through a narrow outlet at the end of the cylinder. Small particles with a large mobility are collected by the inner electrode, whereas large particles with low mobility are swept through by the flushing air stream. The exiting particles within a narrow size range are counted by a condensation particle counter (CPC). The particle size range is scanned by changing the electric field strength.

Another particle size analyzer applicable in the size range > 500 nm up to 10 μm uses a critical orifice to accelerate the air stream. The acceleration of the particles and their velocity depends on their aerodynamic diameter. After the orifice, the velocity of the particles is measured by two partly overlapping laser light beams (aerodynamic particle sizer, APS [250]), which directly gives information on the aerodynamic particle diameter.

Other principles which are used to measure particle size distributions are electrical low-pressure cascade impactors (down to particle diameters of 30 nm) and diffusion batteries combined with condensation particle counters [245], [251], [252]. In order to cover the full particle size range, generally a combination of instruments (e.g., APS or LAS and SMPS) must be applied.

5.3.2.3. *Measurement of Inorganic Components in Suspended Particle Matter*

Suspended particulate matter in ambient air is a complex mixture of miscellaneous components such as elemental carbon, high-boiling compounds (organic carbon), partly water soluble ionic compounds, silicates and other crustal material, sea salt, and heavy-metal compounds. Typical compositions of particles can be found in [250], [253]. Whereas crustal material is considered as relatively inert with respect to human health [235], other toxic and carcinogenic components such as soot, heavy-metal compounds, and fibers can significantly enhance the health effects of particles [235]. Consequently, measurement methods have been developed to analyze these toxic and carcinogenic ingredients of the aerosol separately. In addition, the analysis of the composition of the atmospheric aerosol gives valuable information on the origin of the aerosols and is an important tool for source apportionment [250]. In the following section, measurement methods for inorganic components are described, whereas the analysis of important organic constituents of the aerosol is dealt with in Measurement of Organic Atmospheric Pollutants.

5.3.2.3.1. Measurement of Major Ionic Compounds in Suspended Particulate Matter

Although the major ionic compounds in suspended particles are nontoxic (sea salt) or less toxic (ammonium sulfate), these compounds are of special interest because they are constituents of secondary aerosols (ammonium, nitrate, sulfate) and of sea salt aerosols (sodium, chloride) so that analysis of these compounds can give valuable information on the origin of the collected aerosols [250].

For the analysis of the water-soluble compounds ammonia, nitrate, chloride, and partially sulfate the filter with the absorbed particles is extracted with water, e.g., in an ultrasonic bath. The obtained solution can be analyzed by ion chromatography according to the methods of water analysis [281]. Sodium, potassium, magnesium, and calcium can be analyzed together with the other metals (see below).

5.3.2.3.2. Measurement of Heavy Metals and Metalloid Compounds in Suspended Particulate Matter

Metal compounds in SPM samples are often determined on a routine basis; minimum requirements for sampling and analysis are described in the EU council directive on ambient air quality assessment and its daughter directives [133], [136].

The SPM and its PM 10 fraction is first collected on filters (see Measurement of Particle Mass and Particle Fractions) and is then usually chemically digested to bring the metal and metalloid compounds into aqueous solution before analysis by standard methods. The most important of these are inductively coupled plasma mass spectrometry (ICP-MS) [282], inductively coupled plasma optical emission spectroscopy (ICP-OES) [283], and atomic absorption spectrometry (AAS) [284].

For ICP-MS the inductively coupled plasma is produced in an ionized gas (argon) in a high-frequency field and acts as the ion source for mass spectrometry of the elements of the solution, sprayed into the argon stream. Although this method is expensive it is now the standard method for the analysis of heavy metals in environmental samples because of its high sensitivity and its ability to analyze many elements in a very short time. In addition, it is the European reference method for the analysis of arsenic, cadmium, lead, and nickel in PM 10.

The inductively coupled plasma can also be used as the atomizing and excitation source. In the ICP-OES technique the emission of the excited elements is analyzed. Compared to ICP-MS the method is less expensive but less sensitive.

Atomic absorption spectrometry is a reliable and relatively simple routine method, and is still an important standard method in the field of pollution measurement [285]. The flame technique is only rarely used because of its low sensitivity, but the graphite furnace technique has also been selected as a European reference method for the analysis of arsenic, cadmium, lead, and nickel in PM 10

Other methods for determining metal compounds in SPM include X-ray fluorescence analysis (XRF), polarography, and

neutron activation analysis, but they are only rarely used [285]

Analysis of metal compounds in SPM by conventional methods does not distinguish between the various oxidation states of the elements or between their different compounds. In many cases, a determination of the total metal present is sufficient. However, with certain metals, especially those with carcinogenic properties (As, Cr, and Ni), this further information would be desirable.

There are often large differences in carcinogenicity not only between elements but also between different compounds of the same element (e.g., only chromium compounds of oxidation state VI are carcinogenic). In such cases, specific determination of individual compounds (specification) would be desirable. In spite of the importance of this problem, such specification in pollution measurement is only at the experimental stage. Some progress has been made in speciation of arsenic, chromium, and nickel [286-289].

5.3.2.3.3. Measurement of Soot (Elemental Carbon)

Soot emitted from diesel engines has become relevant since it has been classified as carcinogenic [290], [291]. In addition, soot (elemental carbon) is an important component of the aerosol.

Owing to the variable and complex composition of diesel soot and because some of its constituents are also emitted from other sources, there is no specific method for measuring it. However, existing methods used for workplace monitoring have been adapted to enable reliable figures for the concentrations to be expected in outdoor air to be obtained [292-294]. The fraction of PM 10 in a certain volume of air is separated by a preliminary separator and collected on a glass- or quartz-fiber filter. After solvent extraction and thermodesorption of organic constituents, the elemental carbon content as a measure for the soot concentration the carbon is determined by combustion in a stream of oxygen and coulometric titration of the carbon dioxide formed. In Germany this coulometric method is now the reference method for measurement of soot. at sites influenced by road traffic [296]. In the USA a more sophisticated method is used. Charring of organic material during the thermodesorption step is measured by optical reflectance and the amount of thus-formed elemental carbon is subtracted from the result [297].

Optical determination of the soot content of the dust can be carried out using an aethalometer, based on the light transmission of an exposed dust filter compared with a reference measurement on the unexposed filter. The concentration can be calculated from the light attenuation, using an empirically determined absorption coefficient for particulate elemental carbon and the volume of air passed through the filter. The aethalometer values show good correlation with the coulometric method, but the absorption coefficient is dependent on the specific site [298].

5.3.2.3.4. Measurement of Fibers

Some fibers are classified as carcinogens [299]. Carcinogenic fibers have a diameter of $d \leq 3$ μm and a length $l \geq 5$ μm, where $l : d \geq 3$. These fibers may consist, for example, of asbestos, minerals, or ceramic material (MMVF, man-made vitreous fibers). In the past, asbestos fibers were most important. As a consequence of the prohibition of manufacturing and use of asbestos-containing products, measurements of asbestos fibers in ambient air are now of only minor significance. Fiber concentrations are mostly below the detection limit of the measurement methods (300 fibers/m^3). During recent years, interest has shifted more to other types of fibers, e.g., the MMVFs mentioned above. The concentration of other (i.e., non-asbestos) inorganic fibers may be on the order of some hundreds to some thousands of fibers per cubic meter of ambient air.

Measuring methods for fibrous materials depend on collecting the fibers on filters and counting them under a microscope [300], [301].

Methods of measuring fibers in the atmosphere use the electron microscope. Samples are taken on gold-coated capillary membrane filters and ashed on the filter by oxygen plasma treatment to remove organic substances. The fibers on a randomly selected part of the filter surface are then counted according to defined fiber counting rules under the scanning electron microscope (SEM) and classified according to their fiber type. Classification is based on energy-dispersive X-ray analysis (EDXA), which can distinguish between chrysotile, amphibole asbestos, calcium sulfate, and other inorganic fibers. The fiber numerical concentration is then calculated using the fiber count result, the examined filter area, and the volume of sample air drawn through this filter area.

Under the usual conditions of measurement and counting recommended in the guideline (1 m^3 sampled air per cm^2 filter surface, 1 mm^2 filter surface inspected), the detection limit is 300 fibers per cubic meter. The whole procedure is very complex and requires great care to obtain reliable results.

5.3.3. ***Measurement of Deposition***

5.3.3.1. *Measurement of Total (Bulk) Deposition*

A common method of measuring deposited dust is the Bergerhoff method. The entire material deposited from the atmosphere (dry and wet deposition) is collected in vessels for 28 – 32 d at ca. 1.5 – 2.0 m above ground level (bulk deposition). The collection vessels are then taken to the laboratory and processed (filtration, evaporation of the water, drying, weighing). The result is quoted per unit area of the collecting surface of the vessel and per unit of exposure time, i.e., in grams per square meter per day. According to, the detection limit is 0.035 g m^{-2} d^{-1}.

Total deposition can also be collected by using funnel – bottle combinations, as are frequently used in the EMEP measuring network [302].

All measurements of deposited dust are relative values that depend on the apparatus used, as the recovery of dust is influenced by air flow behavior and other apparatus-dependent parameters. Measurements by different methods can differ

by as much as 50 % [186].

5.3.3.2. *Measurement of Wet Deposition*

Wet deposition (in rain, snow, fog, and dew) and dry deposition are the two routes which harmful substances can pass from the air to the ground, groundwater, or the surface of vegetation. The measurement of wet deposition has therefore been carried out for some decades as part of the study of environmental problems such as the acidification of soil and lakes and the destruction of forests. For example, the European EMEP measurement network [302] has measured the rate of deposition of sulfate and nitrate anions and some cations, and also the pH of the precipitation at ca. 90 measurement stations. There are also many measurement grids in the United States [303]. Unlike the concentration in the air and the directly correlated dry deposition, the fluctuations in the concentration of atmospheric pollution in rain are relatively small, so that a representative measurement of wet deposition can be performed with a wide-mesh grid. Wet deposition in Germany is traditionally measured by the wide-mesh grid of the Federal Environmental Office [304], rather than by the air measurement stations of the Länder.

To completely separate the measurement of wet deposition by rain and snow (fog and dew represent special problems) from the measurement of total deposition (bulk deposition) and dry deposition, rain collectors are used whose sampling orifices are covered during rain-free periods (wet only samplers [305]). Rain sensors, mostly depending on the principle of conductivity change, trigger the removal of the cover at the beginning of a period of rain and its replacement at the end. A review can be found in [306].

The collecting vessel consists of a funnel (collection area ca. 500 cm^2 or more [305]) and the collection vessel itself, which should be thermostatically controlled and kept in dark conditions to avoid artifacts. (Rain in collecting vessels exposed to light stimulates the growth of algae, and organic micropollutants may undergo photochemical degradation by, for example, H_2O_2 or other peroxides.) Special methods must be used to avoid artifacts and the decomposition of sensitive substances. If the collected precipitation is to be analyzed for inorganic substances, the funnel and collection vessel should be composed of materials such as polyethylene or polypropylene. If organic constituents are also to be measured, borosilicate glass is preferable.

5.3.3.3. *Special Sampling Devices for Deposition Measurements*

Totalizators are used to determine both the wet deposition and the approximate dry deposition (on a standardized model surface). The totalizator consists of a funnel and a sample receiver for measuring the wet deposition and a vessel (often a standardized glass jar for measuring dust deposits, see Section Measurement of Total (Bulk) Deposition) for collecting the dry deposition. A rain sensor is used to switch from wet to dry collection and vice versa.

Snowfall cannot always be detected with certainty by the sensors of rain collection equipment. The best method for measuring snow is therefore still to collect it on an exposed plastic sheet.

The sampling of fog, clouds, and dew requires special equipment. Fog and cloud water can be collected actively by impactors [307] or passively by mesh screens [307]. The latter consist of plastic fibers from which the collected water from the fog or cloud runs into the collection vessel. Dew samples are collected on cooled glass surfaces [308], although it is doubtful whether this method corresponds exactly to the behavior of natural surfaces.

Dry deposition can be collected on adhesive films [309] or in more complex systems [310].

5.3.3.4. *Analysis*

If only dry deposition is collected or if the collected deposition is dried before analysis, analytical methods used are fundamentally the same as those used in the analysis of SPM.

Collected rain water can usually be analyzed for inorganic constituents without preliminary sample preparation, and samples need preliminary centrifugation or filtration only if there is visible cloudiness. Heavy metals can be analyzed by the methods described in Measurement of Heavy Metals and Metalloid Compounds in Suspended Particulate Matter. Routine measurements are also made of conductivity, pH, important anions (NaO_3^-, Cl^-) and cations (Ca^{2+}, K^+, Mg^{2+}, Na^+, NH_4^+, etc.) [304], and also, for research purposes, unstable trace substances and intermediates such as H_2O_2 and HSO_3^-.

The analytical methods used are those generally available for aqueous solutions such as conductivity measurement [311], pH measurement by electrodes [311], [281].

Organic components are extracted from rainwater (e.g., with dichloromethane or toluene). Analysis is then carried out by HPLC or gas chromatography (see Measurement of Organic Atmospheric Pollutants).

5.3.4. ***Optical Methods of Remote Measurement***

Whereas the methods so far described enable atmospheric pollution to be measured at a single location, remote methods of optical measurement usually enable integrating measurements to be carried out with light beams up to several kilometers in length. They exploit the absorption properties of atmospheric gases in the UV, visible, or IR spectral regions, and are based on the Lambert – Beer law, which states that the measured extinction is proportional to the product of the path length and the concentration. If the wavelength of the radiation source and the detector in the measuring equipment can be varied, several components can be measured in parallel or in sequence with a single device.

Remote optical sensing methods thus enable measurements to be performed in areas to which access is difficult or impossible, or to trace diffusive emissions from large-area sources such as landfills or petrochemical plants with extended

piping systems and storage tanks.

The five most commonly used measuring systems — FTIR, DOAS, TDLAS, DAS-LIDAR (or DIAL), and COSPEC — are discussed below. Reviews can be found in [312], [313]; minimum requirements for type approval of instruments have been standardized [314].

5.3.4.1. *Fourier Transform Infrared Spectroscopy*

The Fourier transform infrared spectroscopy (FTIR) method is a true multicomponent system which simultaneously measures the spectra of all compounds present in the light beam that absorb in the IR range typically between 500 and 3500 cm^{-1}. The heart of the equipment is an interferometer, which produces interferograms from which the IR spectra are obtained by Fourier analysis [312]. In most applications close to ground level, artificial light sources are used, which are either at a distance of as much as several hundred meters from the spectrometer (bistatic arrangement), or which direct the light beam into the spectrometer by means of reflecting mirrors (monostatic arrangement). If white cells are used, measurements at almost point locations are also possible. A large number of organic components at concentrations of a few ppb can be measured in a light beam ca. 500 m in length. The FTIR method can thus be used as an “optical fence” for the effective monitoring of large-area sources of organic compounds such as petrochemical plants or waste dumps [315], [316]. If the light source and the spectrometer face towards each other and are mobile, gridlike surveys of large-area sources can be carried out [313]. Light paths of several hundred meters yield detection limits which are low enough for measurements in urban background locations [317]. Another important application is the monitoring of gaseous compounds contributing to climate forcing such as CH_4, N_2O, and CO_2 [318]. The FTIR method, its fields of application, and quality assurance have been described in a national standard [319] and in an upcoming CEN standard [320]

The sun can be used instead of an artificial light source, especially for measurements from aircraft and balloons. Integrated column concentrations of trace gases in the atmosphere (e.g., ozone) can be determined in this way from ground level, as well as pollutants present in high smoke plumes [312], [313]. If hot exhaust gases (e.g., from aircraft engines) are observed, the FTIR system can be used to detect the emission spectra of the thermally excited molecules without using a light source.

5.3.4.2. *Differential Optical Absorption Spectroscopy*

Differential optical absorption spectroscopy (DOAS) is probably the most widely used method of pollution monitoring on account of its comparative ease of operation. There are several manufacturers of commercial equipment [312]. It operates in the UV and visible spectral regions, usually with a high-pressure xenon lamp as light source. The light source and spectrometer can either be in the bimodal arrangement, as used in the FTIR method, or the light beam can be directed into the spectrometer by a reflecting mirror. By using beam lengths of up to 10 km, some harmful substances (e.g., HONO), can be measured at concentrations of < 100 ppt [313], [321]. However, this requires very good atmospheric visibility, as fog and rain, for example, attenuate the light beam so much that the beam length must be limited to <100 m for practical purposes [322]. The light is collected by a telescope and focused onto the slit of a diffraction spectrometer. A photomultiplier acts as detector and measures the differential absorption coefficients, which are a measure of the difference between the absorption I measured at the wavelength and the broad band background I_0. Because the light source has a broad bandwidth, a wide spectral range is available, allowing several characteristic lines of the trace gas to be measured.

The sequential measurement of several components is possible, as is the use of several spectrometers simultaneously, although the number of components that can be measured is appreciably smaller than is possible with FTIR. However, the DOAS system has very low detection limits, enabling unstable trace substances such as NO_3 (see Section Measurement of Unstable Trace Substances) to be measured. The DOAS system is used for measuring line sources (e.g., streets) and area sources (e.g., airports), but also for general air-pollution monitoring as a supplement or alternative to conventional measurement stations [313], [317]. The DOAS method has been standardized [323].

5.3.4.3. *Long-Path-Length Absorption Spectroscopy with Tunable Diode Lasers*

Tunable cooled diode lasers can be tuned specifically to the absorption bands of atmospheric pollutants of interest and can provide monochromatic, coherent, narrow light beams in the visible to IR range (3 – 30 μm) for use in absorption spectrometry (TDLAS); sensitivities are in the ppt range [324]. By using long path lengths in low-pressure white cells (a few millibar), it is possible to measure unstable trace substances that are important in air pollution chemistry (e.g., ·OH radicals) [325]. In open absorption paths, the detection limits increase to about 1 ppb, because of pressure broadening of the absorption bands. When compared to FTIR, TDLAS is superior in respect to sensitivity and selectivity. However, because of the narrow light beams of the laser, only one compound per laser can be measured. Multilaser arrays enable the measurement of several compounds [326].

TDLAS has found wide application as a research tool [324]. However, the relatively high capital costs, the need to cool the laser to 20 – 110 K, the difficult operation of the equipment, and the uneven quality of the laser have prevented routine application in pollution monitoring up to now, but improvements such as better white cells with paths lengths of up to 1 km and high-frequency modulation of the diode current are underway [326]. A simplified mobile absorption spectrometer in which the laser is replaced by a different light source in the infrared range is commercially available [327] and is used for monitoring large-area sources. However, it cannot achieve the low detection limits of laser systems.

5.3.4.4. *Differential Adsorption LIDAR*

If a laser beam is directed into the atmosphere without a reflector, part of the radiation is scattered by particles and air molecules, whereby a very small fraction is reflected in the direction of the laser source and can be detected by an adjacent telescope. If pulsed laser light is used and the time taken for the reflected light to return is measured, the distance of the back-scattering particles can be determined. This technique is known as LIDAR (light detection and ranging). If, in addition, the laser beam is tuned to the wavelength of an absorption band in the visible, UV, or IR range and compared

with a reference wavelength adjacent to the absorption band, it is possible to determine the spatial distribution of the pollution. This measuring technique is known as differential absorption LIDAR (DAS-LIDAR or DIAL) [312], [313]. The measuring system can be stationary or mobile (detection vehicle, aircraft), and has been used for large-area pollution measurements in urban regions [328], measurements of the vertical profile of pollutants such as ozone [328], large-area sources such as waste dumps and refineries [327], [328], loading facilities for petrochemical products [327], for monitoring combustion gas plumes [327], and for the remote measurement of elemental mercury emissions from chlor-alkali plants [329]. If the laser beam is periodically swung over a particular area or rotated in a circle, a two-dimensional picture of the pollution can be obtained in which point sources, for example, show up clearly as concentration peaks [313]. The remote monitoring of emissions from exhaust gas plumes is hindered by the fact that only the concentration of a pollutant is measured, and not other important parameters such as volume flow rate, temperature, or the oxygen content of the combustion gas.

Monitoring techniques using DAS-LIDAR have been standardized [330], and several instruments are used in the field [330], [329], [331]. However, DAS-LIDAR has not become widely used on a routine basis, probably because of high equipment costs and the difficulty in operating the laser, which affords highly trained personnel.

5.3.4.5. *Correlation Spectroscopy and Gas Filter Correlation Spectroscopy*

In correlation spectrometers, a spectral region of the undisturbed incident light is correlated with a prepared etched slit mask spectrum of the gas to be measured (COSPEC) [313] or with a reference cell which contains the gas to be measured (GASPEC) [312]. Both the slit mask spectrum and the reference cell rotate in front of the detector, so that the measured signal and the reference signal are detected alternately.

COSPEC passively measures the incident light, and determines the vertical concentration column of the components to be measured (SO_2 and NO_2). Both stationary and mobile operation (aircraft, detector vehicle) are possible, and commercial equipment is available.

The main applications are in studying atmospheric chemical reactions (e.g., conversion of NO to NO_2 in combustion gas plumes), and in the measurement of cross-border pollution from aircraft or detector vehicles [332].

5.3.4.6. *Significance of Remote Measurement Methods*

Remote measurement methods offer a number of advantages over conventional point-measurement technology:

1. Measurements at places and areas to which it is difficult or impossible to gain access (e.g., factory land, waste dumps, chimney stacks).
2. Continuous or quasi-continuous integrating measurement over an area (e.g., factory fence, factory land). This enables, for example, localized gas emissions from diffuse sources to be detected immediately, whereas single-point measuring equipment can often be at the “wrong” location. Moreover, integrating measurements provide a more representative picture than point measurements, for example in urban background locations, which should be representative for an area of several square kilometers.
3. Measurement of vertical profiles of air pollutants (e.g., O_3).
4. Measurement without physical contact (e.g., no spurious results caused by the sampling process itself), so that the measurement of highly reactive trace substances (e.g., $NO^{\cdot}_3$ radicals) is also possible.
5. Measurement of many pollutants simultaneously (FTIR), and measurement of substances that are very difficult to measure by conventional methods (FTIR, DOAS, TDLAS).

These advantages must be balanced against the disadvantages of high costs; operation difficulties, especially of the laser systems; high personnel requirement; the unacceptably high detection limits in many applications (especially FTIR and DIAL); and the difficulty in incorporating the remote measuring method into regulations for air pollution monitoring, which were not conceived for integrating measurements but for conventional point measurements. Some examples of this are listed below:

1. The incompleteness of the results obtained by some remote measurement methods in darkness or fog (e.g., COSPEC, DOAS) does not comply with the requirement for measurements to be representative of all times of the day (see Section Planning of Pollution Measurements).
2. For routine monitoring in some countries (e.g., Germany, the USA, EC Guidelines), the equivalence of measuring methods to reference methods, which are always based on point measurements, must be proved, and the uncertainty of the results must be demonstrated to fulfill given criteria (e.g., expected uncertainty of 25 %) [133]. In open areas, this is only approximately possible [312], but the application of absorption cells make traceability to a gas-based standard and the regular use of test gases for quality assurance possible [317].

As a result of these difficulties, remote monitoring methods are so far mostly confined to special and research applications, such as monitoring diffusive sources or ozone profiles. However, the standardization of several methods and their quality assurance as well as their type approval in association with continuing improvements to the methods may lead to wider practical use in the future.

5.3.4.7. *Measurement of Unstable Trace Substances*

The use of long light paths for remote measurement gives high sensitivity and lack of contact with the substance during “sampling”, thereby enabling important unstable trace substances in the atmosphere to be measured (see Section Significance of Remote Measurement Methods). Examples include $NO^{\cdot}_3$ and HNO_2 (DOAS) [333]; $OH^{\cdot}$ radical and H_2O_2 (TDLAS) [325], [334]; and HNO_2, N_2O_5, CH_2O, and organic intermediates (FTIR) [335]. Although remote measuring techniques do not involve contact with the substance, they can nevertheless lead to spurious results. Thus, in

addition to OH· radicals being present at low concentrations (< 1 ppt), unless special countermeasures are taken, laser light at 308 nm causes photolysis of ozone followed by reaction with water vapor, producing additional OH· radicals in the measurement path [325].

Conventional point measurement techniques are also used for this group of substances, e.g., gas chromatography for peroxyacetyl nitrate [336] and fluorescence measurements on H_2O_2 [337].

5.3.5. *Measurement of Organic Atmospheric Pollutants*

5.3.5.1. *Overview*

The main problem with the measurement of organic atmospheric pollutants is the large number of substances of this type. Several hundred individual substances are listed in emission registers and air quality management plans for densely populated areas. These have a very wide range of toxicological, chemical, and physical properties. TA Luft alone lists ca. 120 organic pollutants and 15 organic carcinogens with varying degrees of relevance to air pollution [140].

The large number of substances suggests that the high cost of measuring the individual substances could be met by summation measurements, e.g., of parameters such as the total content of organically bonded carbon (total C) in the atmosphere. This was the first route taken in monitoring atmospheric pollution.

However, it subsequently became clear that because the effects and concentrations of the various organic compounds were so variable, measurements of total C were not very useful. For example, a benzene concentration of 100 µg/m^3 in the atmosphere represents a considerable cancer risk [299], but the same concentration of *n*-hexane is considered harmless. This applies even more strongly to highly toxic compounds.

Total C measurements were therefore increasingly supplemented or replaced during the 1980s by single substance determinations of the relevant organic components.

This means that the analysis for organic compounds in the atmosphere is an exercise in “applied chromatography” (see ⟶ Basic Principles of Chromatography), as it is necessary to separate the complex mixture of organic compounds present. Highly sensitive detectors and mass spectrometers are available for a wide range of concentrations extending over ten orders of magnitude, i.e., from the mg/m^3 range (e.g., tetrachloroethene in close proximity to emitters) down to the fg/m^3 range (dioxins, furans). Chromatographic methods are therefore the methods of choice if the organic compounds are fairly stable both thermally and chemically. However, there are also special methods for a number of organic substances with reactive functional groups. Here, the physical properties or chemical reactions of the functional groups are utilized to detect the substances.

5.3.5.2. *Summation Methods (Total-C Measurements)*

The standard method [338] used to determine the total C content of the atmosphere consists of introducing the air sample directly into a flame ionization detector (FID). The hot hydrogen flame of the FID produces a ionization current in the presence of organic compounds, which is measured. The sensitivity (response factor) of FID is not the same for all organic compounds. Whereas the response factors of hydrocarbons lie close to unity relative to the calibration standard propane, this figure can greatly decrease for compounds with several heteroatoms (e.g., oxo compounds, halogenated hydrocarbons). Compounds of low volatility, largely associated with the aerosol, are detected incompletely or not at all. Therefore, the FID method underestimates the content of organic substance not consisting of high- to medium-volatility hydrocarbons.

Apart from total C determination, FID is an important detector with a wide range of applications in gas chromatography. As FID can be operated by continuously feeding the air sample, it is suitable for use in automated stations.

Even if results are difficult to interpret from the point of view of air pollution control, they are always of some significance when linked to the problems of increased ozone concentrations and to the detection of unusual pollutants in areas occupied by the petrochemical industry.

5.3.5.3. *Chromatographic Methods*

In the analysis of organic compounds in air gas chromatography (GC, see ⟶ Gas Chromatography) and high pressure liquid chromatography (HPLC, see ⟶ Liquid Chromatography) are widely used for separating the complex mixtures of substances. In gas chromatography high-resolution capillary columns with small cross sections (ca. 0.2 – 0.3 mm) and lengths of 30 – 60 m are now used almost exclusively (HRGC).

After separation on the column, the compounds are detected individually by various detectors, including FID, electron capture detectors (ECD, specifically for electrophilic substituents such as halogens), photoionization detectors (PID, which are very sensitive to aromatic hydrocarbons and other π-electron systems), and thermoionic detectors (NPD, specifically for nitrogen and phosphorus compounds). The HPLC system incorporates special flow-through detectors, similar to the flow-through cells of a UV spectrometer.

An especially powerful method is to use a mass spectrometer as the detector. Knowledge of the mass spectrum of an organic compound often enables it to be identified with certainty, especially in unknown mixtures of substances. Qualitative knowledge of the retention time from chromatograms obtained with conventional detectors, is generally only useful for mixtures whose composition is largely known.

When high-resolution gas chromatography is combined with mass spectrometry (HRGC – MS), the specific detection of individual components using mass fragmentograms is combined with high detection sensitivity, so that difficult analyses such as the determination of polyhalogenated dibenzodioxins and dibenzofurans in the atmosphere then become possible.

Analysis is preceded by sampling, for which various techniques are used, as organic compounds can differ so widely with respect to volatility, polarity, and concentration range.

The choice of sampling method is mainly determined by volatility. It is therefore useful to divide the analytical techniques into those for volatile and those for low-volatility compounds, methods for the latter being subdivided into those based on concentration measurement and those based on deposition measurement.

5.3.5.3.1. Chromatographic Methods for Volatile Organic Compounds (VOCs)

In spite of the high sensitivity of gas chromatographic detectors, direct measurement of VOC in the air is not possible without prior concentration on sorbents during sampling or in traps prior to analysis (with the exception of methane), as detection limits in the range ng/m^3 to µg/m^3 must be achieved. Sampling and preconcentration can be performed either by actively pumping an air stream through tubes or traps filled with sorbents or by passive sampling. The collected VOCs are then desorbed thermally or with solvents, and the resulting gas streams or solutions are injected into a gas chromatograph. The preconcentration step with subsequent thermal desorption can be part of the chromatographic system, so that air samples (e.g., from cylinders) can be introduced directly.

Both manual and automated methods are widely used, depending on the compounds to be analyzed and the time resolution and detection limits to be achieved. An overview of the usual methods for active pumped sampling and subsequent chromatographic analyses is given in Table 11. Two of the examples presented in Table 11 have been standardized by CEN as reference methods for benzene measurements: active pumping with adsorption on organic polymers followed by thermal desorption [339] and active pumping followed by adsorption on charcoal with subsequent desorption by solvents (CS_2) [340].

Table 11. Widely used chromatographic methods of determining organic pollutants (pumped sampling)

Substance group	Concentration range	Sampling, treatment	Final analysis	Refs.
Low-boiling hydrocarbons, volatile halogenated hydrocarbons	ng/m^3 – µg/m^3	evacuated passivated stainless steel cylinders (also for measurement of background air); samples fed via gas loops; cold trapping; thermal desorption	PID GC/FID/ECD	[346], [349]
Organic compounds, C_6 – C_{30}, *bp* range 60 – 350 °C	µg/m^3	adsorption on activated carbon; desorption with CS_2 or other solvents	capillary GC/FID	[350], [351], [353], [340]
Organic compounds, *bp* range 20 – 300 °C	ng/m^3 – µg/m^3	adsorption on organic polymers (e.g., Tenax) or carbon molecular sieve (Carbopack), thermal desorption with cold trapping before capillary column or solvent extraction	capillary GC/FID/ECD/MS	[341], [347], [348], [339]
Modification for low-boiling substances (< – 120 °C)	ng/m^3 – µg/m^3	adsorption on cooled polymers (e.g., Thermogradient tube); cooling to – 120 °C; use of Carbopack	capillary GC/FID/ECD/MS	[343]
High-boiling organic substances partly bonded to particles (e.g., PAH, PCB, PCDD/PCDF); high sample volume	fg/m^3 – ng/m^3	sampling on filters (e.g., small filter equipment) or high volume sampler with attached polyurethane cartridge for the gaseous fraction; solvent desorption of filter and sometimes of polyurethane; various purification and other treatment stages; sublimation of PAH	capillary GC – GCMS (PCDD/PCDF); capillary GC – FID or MS (PAH); HPLC fluorescence detector (PAH)	[385], [382], [365], [366], [367, 369]
High-boiling organic compounds, e.g., CHC, PCDD, PCDF, PBDD, PBDF; low sample volume	fg/m^3 – ng/m^3	adsorption on organic polymers (e.g., polyurethane foam cylinders) with filters (e.g., glass fibers) or inorganic adsorption (e.g., silica gel); extraction with solvents; various purification and other treatment steps (e.g., multicolumn	HRGC/ECD (CHC) capillary GC-GCMS (PCDD, PCDF)	[341], [373], [368], [369]

		chromatography); formation of derivatives of chlorophenols		
High-boiling organic compounds bonded to particles, e.g., components of organic aerosols; deposition samples	ng/m^3 (aerosol: ng/g – µg/g) (pg – ng)/m^2 d^{-1}	Collection of the aerosol on glass fiber filters (e.g., high- or low-volume samplers) or collection of dust on standard surfaces; extraction with solvents; various purification and other treatment steps	HRGC/MS (for PCDD/F) HPLC (for PAH)	[382], [374], [375], [383]

The most common adsorbents used for sampling with subsequent thermal desorption are organic polymers such as Tenax [poly(diphenyl oxide)] and Chromosorb (styrene/divinylbenzene copolymer) or carbonaceous sorbents such as Carbotrap (graphitized carbon), carbonized molecular sieves, and charcoals. A description of sorbents and guidance on sorbent selection and use is given in [348]. If compounds with a wide range of volatility have to be analyzed, the sorbents can be used in series. Care must be taken that no break through of VOCs occurs during sampling. This can be achieved by limiting the sampling volumes to safe levels [348].

This sampling procedure is appropriate for a wide range of VOCs, including hydrocarbons, halogenated hydrocarbons, esters, ketones, and alcohols. Very low boiling compounds may only be partially retained. However, the use of cooled carbon molecular sieves, which have a higher adsorption power than Tenax, enables the sampling of compounds with *bp* < 40 °C [347], [343] such as Freons.

The sampling of VOCs on solid adsorbents followed by thermal desorption and gas chromatography has been automated. Suitable process chromatographs are available from several suppliers (e.g. Chrompack, Airmotec, Siemens, Horiba, and Air Instruments), which can be run in automated measuring networks and which thus enable the quasicontinuous monitoring of VOC. The instruments have FID or PID detectors and are mainly used for the determination of volatile hydrocarbons such as benzene, toluene, and xylene (BTX aromatics). Several automated instruments have been tested and approved by the German authorities [357], [358]. Examples of automated instruments for the monitoring of benzene and the procedure for type approval have been standardized by CEN as reference method [177]. Remote measuring methods for organic compounds, which can also be applied for continuous monitoring, are described in Section Optical Methods of Remote Measurement.

The solid adsorbent is contained in tubes that can be sealed at both ends, and through which a definite amount of air is drawn by a pump (ca. 0.1 – 1 L). Depending on the flow rate and quantity of air, sampling times can be varied from a few minutes to a few hours.

For analysis, the tubes are connected to the sampling system of a gas chromatograph, and the adsorbed compounds are thermally desorbed in a stream of inert carrier gas. Direct feeding to the capillary column is not possible, as the gas volume of several milliliters would lead to loss of the separation effect of the capillary. Intermediate condensation in the gas chromatograph (cold trapping) is therefore carried out, followed by controlled thermal desorption to feed the capillary separation column (cryofocusing). Detection is performed by FID (hydrocarbons), ECD (halogenated hydrocarbons), or mass spectrometer.

This method is also suitable for the analysis of ground air [354], e.g., over abandoned polluted areas and waste dumps.

As an alternative to thermal desorption, solvents such as toluene, carbon disulfide, and pentane [350], [351] can be used, followed by injection of the solution into the gas chromatograph. This method enables thermally labile halogenated hydrocarbons such as dibromoethane to be analyzed [355]. The measurement of organic soil pollutants by the solvent desorption technique has also been described [352].

In the United States, active sampling using specially passivated stainless steel canisters is more widely used than collection on solid adsorbents. This method has been standardized by the EPA [346]. Containers with a capacity of ca. 1 – 15 L are evacuated, taken to the test site, and filled with the air to be analyzed via a critical nozzle or an active pump in the sampling system. A disadvantage of using a pump is that it must be attached to the pressure side of the canister, so that there is a danger of contaminating the sample. However, some problems associated with the use of adsorbents are avoided, such as impurities leading to high blank figures and the decomposition of substances being measured. The volume flow rate during filling of the canister can be adjusted to such low values by means of flow regulators that sampling times of up to 24 h are possible. If the containers are at subatmospheric pressure before analysis, they are brought to atmospheric pressure with high-purity nitrogen, and a definite volume of air, e.g., 100 – 500 cm^3 is then drawn from the container into a gas sample loop by a pump. Moisture in the air can be removed by passing the sample through a dryer (risk of absorption of polar compounds). The substances being measured are then collected in a cold trap and released in a controlled manner as in the thermodesorption method described above, and gas chromatography is then performed.

If stainless steel containers with a blank value of zero are used, this method is also suitable for measuring background concentrations of Freons in the ppt range [346]. Terpenes and isoprene can also be measured in this way [356].

Sampling can also be carried out using bags made of Tedlar or Teflon. However, there is a considerable risk of contamination associated with this system, which is therefore mainly used for analysis at higher concentrations, e.g., in gases from waste dumps [345]. Polar compounds are partially adsorbed on the walls of the bag.

Special gas mice can also be used very successfully for sampling volatile $C_2 – C_9$ hydrocarbons. These are sealed with septums [344]. Some of the collected air is removed with a gas-tight syringe, fed into a gas-sample loop, and cryofocused [344].

All the sampling techniques described above are particularly suitable for measuring hydrocarbons and halogenated hydrocarbons. For the analysis of strongly polar compounds such as alcohols or ketones, the suitability of the methods must be tested in each case.

Solid adsorbents such as Tenax, charcoal, or Chromosorb in specially designed tubes or badges can also be applied for passive (diffusive) sampling of VOCs (see Section Screening Measurements). The great advantage of passive samplers is their easy and cheap handling in the field without the need for access to electricity. They are simply exposed for one or more weeks. After transfer to the laboratory, the samplers can be thermally desorbed [359] or, in the case of charcoal as adsorbent, desorbed with solvents like CS_2 [360]. Subsequently, the samples are analyzed by gas chromatography as described above for active sampling. Passive samplers are commercially available (e.g., Perkin Elmer or Chrompack with thermal desorption, Dräger with solvent desorption) and widely applied in ambient air for the determination of aromatic hydrocarbons such as benzene, toluene, and xylene (BTX) [361], [362]. Comparative measurements with active and passive sampling in the field have shown overall uncertainties of 15 – 25 % for diffusive samplers [363]. In combination with their easy handling, this makes them ideally suited for indicative measurements.

5.3.5.3.2. Chromatographic Methods for Low-Volatility Organic Compounds

Concentration Measurements. The difficulties in analyzing and sampling compounds of this class are due not only to their low concentrations (from a few ng/m^3 to the fg/m^3 range) but also to the fact that low-volatility organic compounds can exist in the atmosphere both bonded to particles and in gaseous form. In sampling, several cubic meters of air must be concentrated, so that if only the particles are collected by the sampling procedure, “blow-off” effects from the filters can be produced by the air flow being sampled [364].

This applies especially to the sampling of semivolatile compounds with vapor pressures above 10^{-8} kPa [365] such as polycyclic aromatic hydrocarbons (PAH) with three or four rings, whereas compounds with vapor pressures below 10^{-9} kPa, such as benzo[*a*]pyrene, can be often sampled satisfactorily by particle collection on filters [370], [366]. However, at ambient air temperatures above 30 °C, 20 % of benzo[*a*]pyrene and perylene passed through the filter [371].

The equipment used for sampling compounds with $bp \leq 500$ °C (e.g., polychlorinated biphenyls; hexachlorohexane; polyhalogenated dibenzodioxins and furans, PAH) therefore includes filters (usually of glass fiber) for collecting particles, followed by equipment to adsorb gases and other components that pass through the filter.

For the collection of particles in the first stage of the sampling train, conventional high- [376], medium- [368], or low-volume samplers [369] are used, which are standardized for SPM measurements with the correct sampling rate. This is necessary to avoid particle size discrimination. The second stage, which can be integrated into the sampling head [368], [369], consists of an adsorption unit, usually a polyurethane foam plug [368], [369] or an organic polymer such as XAD/2 (Amberlite) [372], in which all compounds passing through or blown off the filter are trapped. For the analysis of PAH, a procedure using high-quality samplers has been standardized by ISO [365]; standard procedures for the determination of PCDD/F and PCB by VDI/DIN are available [367-369].

Before sampling, standards (e.g., ^{13}C-labeled congeners) can be added to determine whether losses occur during the sampling operation.

After sampling, both the filter and the adsorbent are extracted, and further ^{13}C-labeled standards are added for calibration purposes. Workup and chromatographic cleaning of the extracts are then carried out prior to analysis by high-resolution gas chromatography coupled with mass spectrometry (GC/MS; PXDD/F), or using detectors (ECD, FID) with spot checking by GC/MS. Labeled standards are added before or immediately after sampling for quantification purposes.

High pressure liquid chromatography (HPLC) can be used as an alternative to HRGC, especially for the analysis of PAH, with detection by fluorescence detectors [366] or UV detectors. PAH derivatives with nitro and oxo groups are determined by HRGC [381].

Examples of these procedures are listed in Table 11.

These analytical techniques can be used to determine both brominated and mixed halogenated PXDD/F. The analytical problem is not only the large number of possible congeners (44 different homologue groups with 5020 single congeners) but also the lack of standard compounds for calibration. Therefore, only a few approximate analyses for PXDD/F in air have been undertaken so far [377], [378].

Very little information on the uncertainty and comparability in the measurement of pollution by PXDD/F have so far been published. In cooperative tests between four laboratories in 1989 [373], the experimental results (expressed as toxicity equivalents) showed standard deviations of 16 % in the first series of measurements and 23 % in the second.

Deposition Measurements. As ingestion in food accounts for most of the daily dose of PXDD/F or PCB absorbed by the general population, it is more important when studying the effects of low-volatility chlorinated compounds to measure their transfer into the food chain (e.g., by deposition on surfaces) than to measure their concentration in the atmosphere.

As with deposition measurements in general, wet deposition (rain, fog, snow), dry deposition, and total (bulk) deposition must be measured. Bulk deposition is measured with standardized dust deposition measurement apparatus (e.g.,

Bergerhoff method, see Section Measurement of Total (Bulk) Deposition) that has been simply modified (e.g., by shading from strong light). Some serious measurement problems (e.g., the resuspension of already precipitated particles, evaporation, and possible photolytic decomposition) have not been systematically researched, so that results are of an approximate character. Standardized methods for the measurement of bulk deposition of PCDD/F and PAH have been developed [374], [375], [379].

Bulk deposition of PCB has been determined in an analogous manner [384].

Wet deposition of PAH and chlorinated compounds such as PCB [386] or PCDD/F [387] has been investigated by using special rain and fog collectors, in connection with the transport of these compounds into relatively unpolluted areas.

5.3.5.4. *Determination of Organic Compounds with Functional Groups*

Organic compounds with functional groups (e.g., aldehydes and phenols) are difficult to analyze by chromatography because of their polarity, unless they are derivatized. However, the physical properties of the functional groups can often be utilized for detection by optical methods, such as infrared absorption or differential optical absorption spectrometry (Section Optical Methods of Remote Measurement). The chemical reactivity of functional groups can be used in reactions to form dyes, which are then determined photometrically, or derivatives are produced for chromatographic analysis.

Thus, aldehydes react readily and specifically with amines to form hydrazones. At concentrations of > 4 µg/m^3, formaldehyde can be determined by the color produced in the sulfite – *para*-rosaniline reaction followed by photometric measurement [388].

Reaction with 2,4-dinitrophenylhydrazine to form the hydrazone, followed by HPLC separation, is more sensitive [389], [390]. This method can also be used to determine higher aldehydes and ketones [390], and can be automated [390].

Similarly, phenol and the cresols react with *p*-nitroaniline in dilute NaOH to give azo dyes which can be determined photometrically [391]. In a method used for measuring pollution in Berlin in 1989, solid-phase extraction from the NaOH solution was followed by HPLC.

The alkylnitrosamine compounds present more serious problems of pollution measurement. There is in principle a gas chromatographic method based on a nitrosamine-specific detector (TEA) and concentration of the alkylnitrosamines on Tenax tubes followed by thermal desorption [392]. However, at the relevant low atmospheric concentrations (a few ng/m^3), there is considerable risk of the unwanted formation of nitrosamines from amines and NO_x on the concentration or separation columns.

Future developments in the analysis of organic atmospheric pollutants will mainly be directed towards inexpensive screening methods, so that the costly analysis of single substances can be concentrated on the most important cases and air pollution monitoring can be kept within economically acceptable limits.

5.3.6. ***Olfactory Methods***

Olfactory methods are used in pollution monitoring to enable the nuisance caused by odors to be objectively assessed in terms of their frequency, intensity, duration, and/or hedonic effects [395], [396]. Fields of application are licensing procedures for installations emitting odorous compounds, cause analysis in the case of complaints, estimation of odor emissions, and calibration of dispersion models [397].

In practice, there are two methods to assess the odor impact in terms of odor frequency:

1. Measurement of the emission concentration (number of odor units) with an olfactometer (see Chap. Emission Measurements), followed by diffusion modeling [398], [399]
2. Odor determination by field inspections with panelists, either by grid measurements or by plume measurements [397]

Method (1) combines olfactometric emission measurements with diffusion modeling and cannot strictly be regarded as pollution measurement. In the method (2), measurements on site are carried out by panelists. In both measurements, the human nose is used as the detector. However, the direct relevance of this detector to the effect to be assessed must be balanced against its considerably lower precision compared with physicochemical methods of pollution measurement. On the other hand, direct measurements of main odor carriers by physicochemical methods are infeasible in most cases due to the complex mixture of odorants, although gas chromatographs equipped with olfactory detection ports parallel to normal detectors can be used to identify the main odor carriers [400].

Field inspections by grid measurements are carried out in accordance with precisely laid down measurement plans by a group of panelists, who record at the locations and times specified whether odors are recognizable and can be identified as characteristic for the source to be studied (an odor event). From the number of odor events during the measurement time, the number of odor hours per year (odor impact) can be determined, and these give a measure of the significance of the odor problem [401]. The locations to be tested are usually arranged in a regular grid of 250 × 250 m in the surroundings of one or several odor source(s) or in an area to be assessed; the sampling times are chosen to be representative for the measurement period (12 or 6 months). The measurement plan thus corresponds closely to the grid arrangement of monitoring stations outlined in Section Measurement Networks (General Air Pollution Monitoring). Field inspections of odors are particularly suited for license procedures and have been standardized [402]. In addition to the frequency of odor events, the intensity of odors and their hedonic properties such as “pleasant”, “unpleasant”, or in very rare cases “nauseating” can be taken into account [401], and standardized methods to do so are in preparation [403]. Questionnaires can be used to link the measured frequency of odor events with the nuisance percieved by residents in an area [404].

German Antipollution Laws do not contain scales of judgment for assessing the significance of odor problems, but this gap is being filled by administrative regulations prepared by the Länder [405], [401].

Field inspections by plume measurements [406], [407] examine the odor plume of a source or group of sources, depending on the current emissions and dispersion conditions. Typically, three cross sections are established perpendicular to the axis of the plume, and each cross section consists of at least five measurement points at which the frequency of odor events and, where appropriate, also the odor intensity and hedonic properties can be assessed by the test panel. Plume measurements are performed to identify emitters and, in combination with dispersion modeling [399], they can be used to estimate the source strength of emitters. If the source strength is known, for example, by emission measurements, plume measurements can be applied the other way round to validate and calibrate dispersion models.

5.3.7. *Screening Measurements*

Simplified measuring methods, possibly with a higher measurement uncertainty compared to more sophisticated methods described above, are used for instance for preliminary investigations (screening). Examples include diffusive samplers, detector tubes, biological methods, and sensors.

In the past, such methods were mainly used for the monitoring of work environments (personal sampling) and air quality in buildings, and sometimes for epidemiological investigations [408]. Development and implementation of the new European Air Quality directives [132-135] led to a considerable need for relatively inexpensive and comparably simple methods of ambient air measurements. Three CEN standards were developed for diffusive sampling where general and specific requirements were defined [409], [410]. The third part of this standard provides guidance for selection, use, and maintenance of such methods [411].

The CEN standards define a diffusive sampler as “a device which is capable of taking samples of gases or vapors from the atmosphere at a rate controlled by a physical process such as gaseous diffusion through a static air layer or a porous material and/or permeation through a membrane, but which does not involve the active movement of air through the device”.

In diffusive samplers, the substance to be determined is collected for example, on adsorption media like charcoal, synthetic organic material, or impregnated filters and meshes. This contrasts with active sampling processes in which air is drawn through a sampling medium by a pump. After a certain exposure time, the collected substance is determined analytically, and its original concentration is calculated based on the relevant physical laws (e.g., of diffusion), taking into account the substance-specific diffusion coefficient, collection time and the geometrical parameters of the sampler.

At present, diffusive samplers — especially for nitrogen dioxide and benzene — are widely used under the European Air Quality Directives, e.g., for preliminary assessment, indicative and screening measurements, network design and optimization, assessment in the vicinity of point sources (traffic, industry), and for assessments in ecosystems [411], [412], [413], [414], [161]. An overview on ambient air measurement by diffusive sampling is given in [415].

Passive collectors are very cheap compared with active measurement devices, a further advantage being that no energy source is required. Disadvantages may be founded on nonideal behavior of the samplers and the dependence on meteorological parameters such as temperature, humidity, and wind speed. In such cases, validated uptake rates must be determined by on-site calibrations, e.g., by parallel measurements with a reference method [161].

Detector tubes are used for sampling and rapid analysis of gases [416]. A hand pump or automatic sampling device is used to draw a given volume of air through a glass tube containing a specific reagent supported on a carrier. The reagent changes color in a characteristic manner depending on the concentration of the substance tested in the air. Detector tubes are mainly used for workplace monitoring or as rapid measurement methods in accidents. They are not used for routine pollution measurements because the detection limits are normally too high and selectivity is poor. Detector tubes are available for many substances over a wide range of concentrations.

Two biological methods are routinely used in pollution monitoring. In the standard lichen exposure method [418], the death rate of the lichen is determined over an exposure time of 300 d. In another method, a species of rye grass is exposed for 28 ± 1 d [419], and the growth of the grass is determined. Both methods enable the general effect of pollution to be determined.

Further development can be expected in the field of sensors which determine chemical states or parameters [420]. They often have the disadvantage of unacceptably high detection limits, poor selectivity and stability, drifting phenomena, and sensitivity to environmental conditions such as temperature. Sensors are already used in higher concentration ranges where great accuracy is not required (workplace monitoring, warning equipment, leak tracing).

[Top of Page]

6. Emission Measurements

Klaus Lützke

6.1. Introduction

To protect the environment, air-polluting emissions must be limited using state-of-the-art technology. Here, the word emissions is used in the legal sense, in accordance with § 3 Section 3 of the Federal Antipollution Law (BImSchG) [421]. Emissions can pass into the atmosphere in exhaust gases or waste air from fixed sources, such as chimney stacks. However, they can also be produced diffusely, e.g., from openings in buildings or waste-storage areas. Emissions are usually expressed as mass concentrations, but also as mass flow rates or mass ratios (mass of emitted material per unit mass of substance produced or processed). Atmospheric impurities are changes to the natural composition of the air due to smoke, soot, dust, gases, aerosols, vapors, or odoriferous substances (odoriphores). Methods for determining

emissions comprise both manual methods of measuring spot samples and continuous measurement techniques. Depending on the components measured and the measurement requirements, emission concentrations can be from the g/m^3 range down to the trace range (pg/m^3). As environmental protection standards have become stricter, new measurement methods have been developed with lower limits of detection. The demand for continuously operating measurement equipment has greatly increased. The methods described here therefore represent only an overview of the present state of measurement technology.

6.2. Objectives and Planning of Measurements

The definition of objectives must specify the quantities to be determined, the boundary conditions, and the required accuracy of the results. There can be many different reasons for determining emissions. Typical objectives include: monitoring compliance with emission limit values; determination of agreed guaranteed figures for production plants and exhaust gas purification plants (acceptance measurements); calibration of continuously recording emission measurement equipment; testing the functioning of continuously recording emission measurement equipment; emission control monitoring fitness for use; analysis of the causes of noncompliance with guaranteed values or emission limit values; determination of changes in the emissions from a plant after process changes, operational disturbances, or design modifications.

Pollution control regulations require that compliance of plants with emission limit values be monitored. In Germany, the main regulations are the Federal Antipollution Law (BImSchG) [421], TA Luft [422], 13th BImSchV [423], 17th BImSchV [424], and 2nd BImSchV [425] (see Chap. Laws and Technical Regulations). In the licensing procedure, the requirements included in the individual regulations form part of the conditions and supplementary provisions under which a plant must be operated (requirements of the authorities). The rules concerning the monitoring of performance characteristics (proof of guarantee) of items of plant and gas purification installations are civil law agreements. The monitoring of compliance with emission limit values is already regulated by the Federal Antipollution Law. According to this, the authorities can order that, in the case of licensable plants [426], measurements should be carried out after bringing into service, after a significant change, or after certain periods of time. As well as random sample measurements (spot measurements), continuous measurement and recording of certain emissions may be required.

In measurement procedures, the measurement plan is of great importance. Measurements for determining plant emissions should be carried out such that the results correctly represent these emissions and are comparable to similar plants and operating conditions [422]. Before the measurement plan is produced, the objectives of the measurements must be exactly defined [427]. These objectives are usually based on licensing information, official regulations, or contractual documents (acceptance measurements). Alternatively, guidance can be obtained from relevant administrative regulations or VDI Guidelines. According to these, the necessary prior knowledge of the plant under investigation can be used, e.g., technical data about the plant, information about operational behavior, materials used, positions of the measurement points, etc. This is supplemented by knowledge about the measurement methods to be used, the personnel plan, and the time period. In the measurement plan, the following must be specified:

1. Where will the measurements be carried out?
2. What must be measured?
3. How are the measurements made?
4. What accuracy is required?
5. How will the results be determined?
6. Who will carry out the investigations?
7. When will the measurements by carried out?

VDI 2448 describes the planning of spot sampling measurements of stationary source emissions [427]. The measurement plan should be produced at an early stage. For example, the planning of acceptance measurements must be started at the planning stage of the plant. The type of measurements to be carried out, the arrangement of measuring ports, and the equipment for the measurement points must be considered and fixed. If this work is done at a later stage, the installation costs are often higher. A detailed plan is especially important if the equipping of measurement points will be impractical at a later stage, as in the case of coated exhaust gas ducting, or if the measurement points are no longer accessible or are unsuitable. In planning, the accessibility and the minimum requirements that apply to exhaust gas ducting must be considered. VDI 2066, Sheet 1 gives advice on the minimum length required for the exhaust gas ducting upstream and downstream of the measurement point [428]. For measurements at existing plants, the measurement plan must conform with plant conditions. The arrangement of measurement points then cannot always be freely chosen. In every case, it must be judged whether a measurement can be regarded as representative. If it is not, modifications to the measurement technique to improve its accuracy must be considered.

6.3. Requirements for Measurement Techniques and Measurement Stations

Emission measurements must be carried out using state-of-the-art techniques and equipment. TA Luft prescribes that emission measurements must be carried out in conformity with certain guidelines and standards of the Commission for Atmospheric Pollution in the VDI and DIN. Alternative or supplementary measurement methods are permitted, especially if they are accepted by the Federal Environmental Ministry with agreement with the highest Land authorities. The guidelines by the Commission for Atmospheric Pollution in the VDI and DIN describe the established state of the art in the measurement field.

6.3.1. *Quality of Measurement Methods*

The quality of the measurement method must meet the requirements of the measurement task. To specify measurement performance, quantities characteristic of the measurement method are used, such as detection limits and determination limits as defined in VDI 2449, Sheets 1 and 2 [429], [430]:

Detection limits and *determination limits* must lie significantly below the lower emission limit. In general, the detection limit

should be less than one-tenth of the lower emission limit. Where there are no emission limits, but compliance with other constraints must be monitored, requirements are analogous.

The *measurement uncertainty* can be determined from the reproducibility found in duplicated determinations. The reproducibility of a measurement method is determined from a series of simultaneous determinations using two complete measurement procedures. The reproducibility can be based on the mean value of the sample of measured values or on classes of measured values. Thus, the standard deviation of the measurement process can be estimated from the number of pairs of measured values and the concentration of the substance.

The *selectivity* of the measurement method chosen must be sufficiently high with respect to other substances present in the exhaust gas, i.e., the cross-sensitivity must be low.

If continuously recording measurement methods are used, other parameters are important, e.g., linearity of the measurement signal, dead time, rise time, drifting of zero point and sensitivity, and setting-up times. The measurement method must cover the expected range of values. Similar considerations apply to the choice of methods for the measurement of reference quantities (O_2 content) and the state quantities temperature, humidity, and pressure.

6.3.2. *Choice and Equipment of Measurement Stations*

“Measurement stations should be of adequate size, easily accessible, and so designed and selected that technically problem-free emission measurements that correctly represent the emissions from a plant can be carried out” [422].

Measurement and sampling stations must be included at the planning stage of new plants. With existing plants, the requirements for measurement stations should be considered to enable the necessary modifications to be carried out and/or equipment to be installed during planned shut-downs. For all measurement stations, the following are fundamental requirements:

The association of the measurement station with the emission source must be indicated by unambiguous labeling. For this, the source numbers used in the description of the emissions or the company source numbers can be used. It should also be possible to a make an assignment to the relevant part of the plant or operation.

Flow into the measurement line should be as smooth and undisturbed as possible. Recommendations for the measurement of dust in flowing gases can be found in VDI 2066, Sheet 1 (dust measurements). The flow properties in the exhaust gas duct can be affected by merging gas flows, changes in direction, changes in cross section, inserted items of equipment, and valves. To ensure a flow which is as undisturbed as possible, the measurement/sampling point should be in a straight section of the exhaust gas duct with a constant size and shape. VDI 2066 specifies uninterrupted inlet and outlet pipes whose length is each at least three times the hydraulic cross section. To prevent sedimentation (e.g., when taking measurements in the unpurified exhaust gas prior to the dust filter), the measuring point for the determination of particulate emissions should if possible be located in a vertical part of the ducting. The sampling points should be so arranged that grid-based measurements can be made for determining both dust emissions and emissions in gaseous form.

The size of the measurement openings is determined by the type of measurement equipment and sampling probes. The number of openings depends on the size and shape of the cross section under investigation. The arrangement, number, and dimensions of openings usually used for dust measurement has also been found to be suitable for measuring gaseous components where measurement grids are required. A good solution consists of large circular openings with closable gas-tight lids which can be modified to suit the measurement task. This possibility of modifying the lid at a later stage is especially important if it is not possible to modify the ducting under investigation (e.g., because of the presence of a rubber lining for protection against acid).

The operating platforms for sampling and measurement must be sufficiently large to accommodate the sampling probes, sampling equipment, and associated measurement apparatus. Access to measurement stations must be provided in accordance with the relevant regulations. There must be no extraneous emission sources (e.g., bursting disks, safety valves) in the operating area of the measurement station. The measurement station must at least be provided with protection from rain and weather if the operating platforms are in the open. Adequate, correctly sized, and fused electrical supply points must be installed at the measurement station. Depending on the type of measurement, compressed air and water supply points may also be required.

6.4. Requirements for Continuous Measurement and Associated Computer Systems

For a number of licensable plants, systems for continuously monitoring emissions are legally required. In addition, owners of industrial plants in their own interest may carry out continuous monitoring of emissions to show that emission limit values are being complied with, e.g., if local residents complain about plant emissions. The regulation on power-station emissions (13. BImSchV) states that power stations must be provided with equipment for continuously monitoring emissions of dust, carbon monoxide, nitrogen oxides, and sulfur oxides, and that the results must be continuously and automatically evaluated [423]. The regulation covering incineration plants for waste and similar combustible materials (17. BImSchV) requires the continuous determination, recording, and evaluation of emissions of carbon monoxide, total dust, organic substances (as total carbon), gaseous inorganic chlorine compounds (as hydrogen chloride), gaseous inorganic fluorine compounds (as hydrogen fluoride), sulfur dioxide and sulfur trioxide (expressed as sulfur dioxide), and nitrogen monoxide and nitrogen dioxide (as nitrogen dioxide). Additionally, other exhaust gas parameters such as oxygen content, exhaust gas volume flow rate, and temperature must be determined and recorded [424].

For other licensable plants, TA Luft prescribes the conditions under which quantitatively significant emissions of dust and gaseous atmospheric pollutants must be continuously automatically measured and evaluated. The measuring equipment is specified in publications by the Federal Ministry responsible for antipollution. Testing for suitability, installation, calibration, and maintenance of measurement equipment are described in binding guidelines by the Environment Minister

in conjunction with the Länder so as to provide quality assurance and to ensure uniform federal practice in the monitoring of emissions.

6.4.1. *Minimum Legal Requirements*

A series of parameters, the so-called minimum requirements, ensure that continuously recording measurement equipment does not fall below a certain standard of performance [431]. These minimum requirements are from time to time reassessed in the light of the latest state of the art in measurement technology. They are subdivided into various general requirements, e.g., position of reference and zero points (live zero), recording range, and availability. Other requirements refer to measurement principles (e.g., photometry and radiometry), substance-specific properties such as dust content and smoke number, required degree of reproducibility, drift with time, and detection limits. Minimum requirements are also set for detection limits, availability, and the continuous measurement of reference values, e.g., oxygen content and waste gas volume flow rate. A separate guideline describes the requirements that apply to computers for the calculation of emissions [432].

There is a further guideline on the evaluation of continuous emission measurements on incineration plants (17. BImSchV) [433].

6.4.2. *Testing the Suitability of Continuous Measurement Methods and Evaluation Computers*

When testing for suitability, compliance with minimum requirements must be determined in laboratory tests and also in a long-term test lasting at least three months on an emitting plant. Standards of performance, but not measurement principles or any specific measurement equipment, are prescribed in the minimum requirements. Any kind of measuring equipment can be used, whatever its measurement principle, provided that it complies with the minimum requirements. Testing for suitability usually begins with the determination of characteristic performance figures in a laboratory test. After this, a long-term test under practical conditions is carried out on full-scale plants that require continuous monitoring of pollution emissions in accordance with legal and administrative regulations of the Federal Antipollution Law (13. and 17. BImSchV, TA Luft). The suitability test should be carried out with two similarly constructed measurement devices operating in parallel, with separate sampling, sample preparation, and data processing. The three-month test should be carried out under the severest known boundary conditions. If important minimum requirements are not complied with, the test is discontinued. The applicant can then carry out improvements, repairs, and modifications. Consideration must then be given to the question to what extent the changes to the measurement system make it necessary to retest any of the minimum requirements. On conclusion of the test for suitability, the testing institute produces a test report on the results. This is sent to the Federal Environmental Office and to the Länder Committee for Antipollution. If the Federal Ministry for the Environment, the relevant Land authorities, and the testing institutes agree that the overall judgment is positive, the certification of suitability is issued at the instigation of the Federal Ministry for the Environment in the general ministerial paper issued by the Federal Ministry [434], [435].

6.4.3. *Testing and Calibration of Measurement Equipment*

Quality assurance tests for continuously recording measurement methods can be divided into tests of internal equipment function, maintenance, annual functional verification, and calibration [436].

Internal equipment functions tested can include automatic changeover of mesh filters, automatic electrical balancing at given time intervals, or the regular functional checking of certain components (e.g., pumps, flowmeters).

Normally, the equipment must be maintained and tested at definite time intervals by the operator. This requires functional testing with calibration aids such as mesh filters and test gases.

For in-situ measurement equipment, the optical boundary surfaces must be cleaned, air filters must be cleaned or changed, and the zero point, the sensitivity, and data recording must be checked. In methods involving extractive sampling, the heater for the sampling system, the leak-tightness of the tubing, the sampling flow rate, and the condensate flow must be tested. The zero point and sensitivity must be tested and if necessary adjusted by using test gases.

For continuously recording measurement equipment, an annual functional test must be carried out on the complete measurement system, including sampling. In a general test of the state of maintenance, and, in the case of extractive sampling systems, a test of the leak-tightness of the measurement system, various testing activities are required. These can include: testing for soiling of the optical components or of the sampling system; determination of the changes with time of zero and reference points; determination of cross-sensitivity towards CO_2, CO, NO, NO_2, SO_2, water vapor, etc.; testing of the functioning of the equipment with test gases, test solutions, filters, or other equipment-specific standards with three or four values over the measured range. Also, specific equipment functions relevant to operational dependability are tested, e.g., constancy of sample stream offtake, duration of measurement cycle, metering and composition of reagents. The transfer of the measured figures to the data collection system must also be tested [437].

Continuously recording measuring equipment must be calibrated after a new plant has been put into service, after significant alterations, and at intervals of 3 – 5 a. Calibration is the determination of the relationship between the readings of the complete measuring equipment and the actual concentration of the substance present in the exhaust gas from the plant as determined by comparison measurements with a conventional analytical method [438], [439]. Checks on the maintenance and functioning of the system must be carried out beforehand. The figures determined simultaneously by the two methods are compared and statistically evaluated. To calibrate dust measurement devices, at least 12 – 15 comparative measurements are usually necessary [428], [440], these being carried out at a number of points arranged in a grid pattern covering the cross section of the ducting. This gives the grid-based analytical function.

With gas emission measurement equipment, there are two main procedures for determining the analytical function. In the first of these, which is analogous to that used in dust measurement, comparative measurements are carried out as grid analyses using a standard method suitable for a bulk phase, e.g., the H_2O_2 – thorine method for SO_2 determination (VDI

2462, Sheet 8). Sampling at the individual measurement points on the grid must be carried out at a rate proportional to the gas velocity and if possible, for various concentrations of the measured substance. This gives the grid-based analytical function. In the second method, a single point analysis is carried out. With continuously recording measurement methods using extractive sampling, ca 20 comparative measurements are carried out in the first stage, with the conventional method and the continuously recording measurement method being operated at adjacent sampling points. The statistical evaluation gives the single-point analytical function for sampling by both methods at the same place [438], [439]. The difference between the equipment-specific characteristic line and the single point and grid-based analytical function is shown schematically in Figure 6.

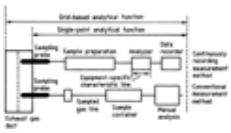

Figure 6. Calibration scheme for continuously recording emission measurement equipment

For many applications where the concentration distribution over the cross section is approximately uniform, this single-point analysis is sufficient. It includes the entire exhaust gas matrix and the variation of concentration with time. It might be thought that this procedure would also be valid for linear sampling by in-situ measuring equipment. However, even when the single-point analytical function is used, the evaluation requires additional grid measurements to verify that the concentration distribution is approximately uniform. In all cases where the concentration distribution is not uniform, this single-point analytical function must be supplemented by additional measurements to give a grid-based analytical function. Therefore, in a second stage, a correction factor must be obtained by means of approximate grid-based measurements, enabling the representativity of the measurement point to be improved. This stepwise procedure is recommended for calibration with single-point sampling. It has the advantage that the cost of the grid-based measurement can be reduced, because instead of using a conventional analytical method to determine the correction factors, a second continuous emission measurement apparatus is used whose results have previously been compared with those from the equipment being calibrated, which also operates by extractive sampling [438]. For the measuring equipment to be calibrated, the single-point analytical function has already been determined by a conventional analytical method, taking into account the complete exhaust gas matrix. Thus, these results can be transferred to the second emission measurement equipment. Grid-based measurements are then carried out. The results of the grid-based measurements, in combination with the single-point analytical function of the equipment to be calibrated, lead to a grid-based analytical function. The arrangement of sampling probes for determining the equipment-specific factor for a second recording method and the procedure for extending the single point analytical function to a grid-based analytical function are shown schematically in Figure 7.

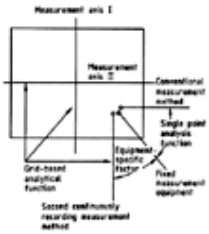

Figure 7. Arrangement of sampling probes to determine the equipment-specific factor for a second continuously recording method, and a representation of the relationship between single point and grid-based analysis

6.5. Physical State Parameters, Reference Value Measurements, and Evaluation Computers

The physical quantities temperature and humidity, and the reference values oxygen content and carbon dioxide content must be determined at the same measurement point as the measured components within the length of ducting under investigation. The temperature of the exhaust gas and sampled gas and the pressure of the exhaust gas must be known to enable the mass concentration of the measured components under standard conditions (273 K, 0.1013 MPa) to be calculated. Figures for the mass concentration based on dry exhaust gas are generally also required, so that the moisture content of the exhaust gases must also be determined. The standard density of a gas mixture is found from the standard densities of the individual components and the quantitative composition. The standard density of each component is found by dividing the molecular weight by the molar volume in the standard state (DIN 1306 and 1871) [441], [442]. The oxygen, carbon dioxide, water vapor, and nitrogen contents enable the density of the exhaust gas (used in dynamic pressure measurements) to be calculated precisely. The oxygen content is an important reference value from which the measured concentrations can be calculated according to the equation

$$E_B = \frac{21 - O_B}{21 - O_M} \cdot E_M$$

where E_M = measured emission, E_B = emission based on the oxygen content, O_M = measured oxygen content, O_B = reference oxygen content.

6.5.1. *Temperature*

The exhaust gas temperatures at which emission measurements are carried out are generally in the range 50 – 400 °C, which can conveniently be measured by resistance thermometers or preferably by thermocouples. The relevant regulations on temperature measurement should be referred to [443-446]. The furnace temperature is now of great importance in refuse incineration plants, and must be monitored as prescribed in 17th BImSchV. This requires that the minimum temperature after the final addition of combustion air must be maintained at least between 850 or 1200 °C. These temperatures must be determined continuously, recorded, and evaluated. In this temperature range, only thermocouples have become established for long term use. As there is significant radiation in this temperature range, the thermocouples must be calibrated. Gas extraction pyrometers can measure gas temperature without radiative contribution.

The gas is sampled at a high rate by a pump and passed over a thermocouple. The thermocouple is covered with a protective tube to prevent errors due to thermal radiation. However, continuous long-term measurement is not possible at present [443].

Recording measuring equipment based on acoustic pyrometry has been marketed. This relies on the fact that the speed of sound in gases depends mainly on the temperature of the gas and only slightly on its composition.

6.5.2. *Moisture Content*

Several methods are in principle suitable for determining the moisture content in off-gases, including psychrometric difference measurement, the lithium halide dew point method, and condensation and adsorption methods. In off-gases, the H_2O content usually lies between 2 and 30 vol %, corresponding to a water dew point temperature of 20 – 70 °C. The determination of water content has hitherto been carried out by manual analysis, as continuous measurement was not technically feasible.

The regulation that deals with waste incineration plants (17. BImSchV) prescribes continuous measurement, recording, and evaluation. Initial investigations into the use of continuously operating LiCl moisture measurement equipment were carried out in an R & D project intended to formulate the minimum requirements that would apply to moisture measuring equipment. The LiCl sensor was replaced by an LiBr sensor to give good service availability [447]. Legally binding minimum requirements for the measurement of moisture content have existed since 1992 [448]. A multicomponent (HCl, SO_2, NO, CO) photometer which uses the single-beam, dual-frequency method, produced by Perkin – Elmer, has been approved for measurements on waste-incineration plants. It can also measure the H_2O content in the range 0 – 40 vol % [449]. Another recommended measuring device manufactured by Ultrakust operates as a back-scattering psychrometer [450]. This is suitable for the exhaust gases from furnaces and from incinerators for process and domestic refuse. The lowest measurement range during the test for suitability is 300 hPa water vapor partial pressure.

Other measuring equipment operating on the principle of a difference measurement on “dry” and “wet” zirconium oxide probes is currently being investigated for suitability.

6.5.3. *Carbon Dioxide Content*

During the 1970s oxygen content replaced the formerly used reference value CO_2 for standardizing measured emission concentrations. There are therefore no minimum requirements and suitability tests for CO_2 measurement equipment. Investigations have been carried out to determine to what extent CO_2 measurement could be used as a substitute reference value [447]. In the combustion of a single substance, the two reference values CO_2 and O_2 are interchangeable without appreciable error. However, in the combustion of mixtures, the discrepancies are significant because of differences in the maximum CO_2 content associated with the fuel. In random sample emission measurements the CO_2 content is usually also measured to enable the density of the exhaust gas to be better characterized. In principle, the Orsat method is suitable, but the measuring methods usually used are based on instruments such as the IR absorption photometer. Electrochemical methods are also used.

6.5.4. *Oxygen Content*

All discontinuous and continuous measure methods include a mean-value calculation followed by conversion to STP, and this is usually followed by a further conversion to a definite oxygen content. Because errors in O_2 determinations have an immediate impact on the results of emission measurements, minimum requirements apply. A number of types of measurement equipment have been approved [431, 434, 435, 449-451]. Sampling can be performed by extraction or in situ. In extractive sampling, the exhaust gas together with the other pollutants is passed through a cooling system, so that measurement results based on dry exhaust gas and 273 K are immediately obtained. Suitable measurement principles include the magnetodynamic method using a torsion balance (manufacturers: Bühler Mess- und Regeltechnik, Mannesmann/H.u.B., MC Products Analysentechnik, Maihak, Rosemount), the equal pressure or variable-pressure measuring method with an auxiliary gas (manufacturers: Maihak, Siemens), and the electrochemical method with a zirconium dioxide solid-state electrolyte (manufacturers: ASEA BBC, Enotec, Westinghouse Controlmatic, Yokogawa). Using the last-named measuring principle, the O_2 content is determined either as “wet” O_2 content directly in the exhaust gas duct, or as “dry” O_2 content by extractive sampling followed by cooling and condensate removal.

6.5.5. *Evaluation Computers*

To comply with the requirement to evaluate measured results continuously and automatically and to document them, the measured values from the recording equipment (dust, SO_2, CO, NO, etc.) are converted into the appropriate physical quantities based on the regression curves determined in the calibration, and are averaged over the reference time (usually 30 min) [432]. After converting to STP and reference oxygen content (e.g., 11 % O_2 for refuse incineration equipment), the concentration figures are divided into 20 classes of equal range. A scheme for the production and evaluation of measured values is shown in Figure 8. These values are not classified in the true sense, but are counted in the computer memory. The emission limit value corresponds to the upper limit of Class 10. Values that exceed twice the limit value are counted in Classes 21 and 22.

Figure 8. Collection and processing of measured values

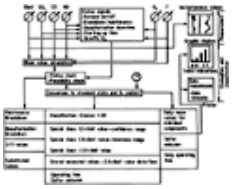

The legally prescribed evaluation calculations should be carefully noted. For example, according to TA Luft and 13th BImSchV, emission values are considered to have been complied with if

1. Overall daily mean values do not exceed the emission limit value
2. 97 % of all half-hour mean values do not exceed six-fifths of the emission limit value
3. Overall half-hour mean values do not exceed twice the emission limit value

For refuse incineration plants a different kind of evaluation is required [424], [433].

The limits of these two classes correspond to the tolerance range established in the calibration. In Class 23 is stored the number of half-hour mean values that exceed 1.2 times the limit value, up to the limit of the confidence range. The special classes are used to store the half-hour mean values which did not represent true measurement values for more than one-third of the measurement time because of breakdowns or maintenance work on the analyzers (Class 24). The classification of half-hour mean values is shown schematically in Figure 9.

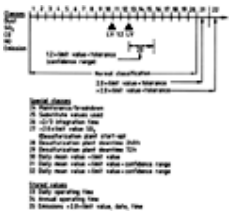

Figure 9. Classification scheme for half-hour mean values

The total number of these values, including those values that were integrated over less than two-thirds of the normal measurement interval owing to plant startup or shut-down are also counted in Class 26. In the absence of the reference value of oxygen due to a breakdown of the measurement equipment, the evaluation computer uses a substituted figure for O_2 which is established in the calibration. All half-hour mean values that were obtained using a substituted figure are classified in the normal way and also counted in Class 25. As well as counting the daily mean values in three classes (30, 31, and 32), the operating times of the plant during the day and the year are stored and reported (Classes 33 and 34). Class 35 is used to store all measured values where twice the limit value has been exceeded.

The time and date of each event is also recorded. Transmission of the data, i.e., the numbers of measurements counted in the 20 normal classes and special classes and the stored data takes place daily automatically at a fixed time. Furthermore, all the data is also transmitted at the end of each calendar year. The daily print-outs give continuous documentation of the emissions from the plant. The annual print-out provides information for the supervisory authority.

Evaluation computers tested for suitability include those manufactured by Durag, Gesytec, Industrie Electronic Schmitz, Maihak, NIS Ingenieurgesellschaft, Nukem, Sick, and RWE.

6.6. Measurement of Gas Velocity and Flow Rate

Normally, only the mass concentrations of the pollutants are measured for the control of emissions. However, in some cases, the emission mass flow rate of a pollutant must be measured discontinuously or continuously. It is determined as the product of the mass concentration and the exhaust gas volume flow rate. The mass flow rate can usually be calculated with sufficient accuracy from known plant parameters such as fuel consumption or steam production rate. If the operating parameters of the plant fluctuate widely or if the quality of the fuel is not known accurately, a direct determination of the volume flow rate of the exhaust gas must be carried out. Thus, for example, the regulation covering waste incineration plants usually requires continuously recording volume flow rate measurement [424]. The minimum requirements should be referred to (see Section Minimum Legal Requirements).

Discontinuous Methods. To calculate the volume flow rate $\dot{V}$, the mean velocity $\bar{v}$ in the measured cross section must be determined. The requirements that apply to the properties of the measured ducting (undisturbed inlet and exit) are described in the VDI Guideline 2066 [428]. The mean flow rate is determined by point velocity v_i measurements carried out in a grid arrangement covering the entire measured cross section. For emission measurements, the arithmetic mean is obtained in accordance with the instructions given in VDI 2066, which also specifies the number of measurement points (see Section Measurement of Particles).

The value of $\bar{v}$ is given by

$$\bar{v} = \frac{1}{n}\sum_i vi$$

The mean volume flow rate $\dot{V}$ is given by $\dot{V} = \bar{v} \cdot A$, where A is the measured cross section.

Pitot tubes are used for manual measurement. The dynamic pressure is measured as the difference between the total pressures p_{tot} and the static pressure, and the local velocity is then given by

$$v = \sqrt{\frac{2}{\varrho}(p_{tot} - p_{stat})}$$

The Prandtl modification of the Pitot tube is widely used (VDI 2066). In the case of different Pitot tube designs, correction figures determined in fluid-dynamic experiments in a wind tunnel are sometimes used. The measurement range of the Prandtl tube is ca. 2 – 50 m/s. Anemometers are used for lower velocities; when used in stack gases special calibration conditions and the influence of gas temperature and gas properties on the reading must be considered.

Continuous Methods. Methods for the continuous measurement of gas velocity and volume flow rate include dynamic pressure difference methods with transducers, ultrasound measurements by the time-difference method, and determination of turbulence frequency by ultrasound measurement.

Dynamic pressure measurements usually require several openings on the upstream side which enable the mean total pressure over the cross-sectional area of the exhaust gas duct to be measured. Using the arrangement of measurement points specified in VDI 2066, with equal grid unit areas, the probe must be fitted at each of these points. For velocity measurement based on the determination of the turbulence frequency by its modulating effect on an ultrasound beam, precalibration must be carried out by the manufacturer for each user. In time difference measurement by ultrasound, the signals pass through the gas stream at an angle α in two directions. Each of the two transducers acts as both a transmitter and a receiver.

Equipment that operates according to these principles has now been officially approved. Note that all methods and equipment in the grid measurement system must be calibrated by carrying out simultaneous velocity measurements with a reference method, usually dynamic pressure measurement by the Prandtl method. Calibration is discussed in Section Testing and Calibration of Measurement Equipment.

6.7. Measurement of Particles

Both manual and continuously recording methods can be used for measuring particulate materials. The manual methods and some of the recording methods are described in VDI 2066. When using recording methods, the minimum legal requirements (see Section Requirements for Continuous Measurement and Associated Computer Systems) and the certification of suitability must be complied with. The manual methods are proven, but problems can occur with very low dust contents (<1 mg/m^3) and moist, supersaturated exhaust gas (e.g., after it has passed through water scrubbers). Although methods of continuous measurement are available, not all are equally suitable. For example, it is not possible to work with optical transmission equipment on supersaturated exhaust gases. Also, it is unsuitable for gases with low dust content inducts providing small optical path lengths. With exhaust gas saturated with water vapor, extractive sampling and reheating techniques are used.

The measurement principles used include the measurement of extinction, light scattering, and radiometry. For low dust contents, the more sensitive method based on scattered light measurement is used.

6.7.1. *Sampling*

The sampling of particles in a flowing exhaust gas must take place at the same velocity as that of the gas to prevent segregation or disturbance of the particle-size distribution due to the inertia of the particles, leading to falsification of the measured solids content. If the sampling rate is too high, the measured dust content will be too low, and vice versa. This mechanism depends on the particle size distribution. For particles of aerodynamic diameter <0.5 µm, the effect of this inertia is practically negligible. A schematic diagram of equal velocity sampling is given in Figure 10.

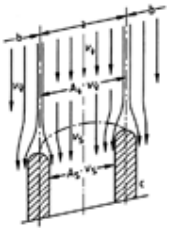

Figure 10. Scheme for equal velocity sampling

a) Sampled gas stream $\dot{V}$; b) Main gas stream $\dot{V}$; c) Sample probe

A = cross-sectional area; $A_{\dot{V}}$ = cross-sectional area of sampled gas stream = effective cross-sectional area of probe; v = velocity

Equal velocity condition: $v_{\dot{V}} = v\dot{V}$; Equal flow rate condition: $A_{\dot{V}} \cdot v\dot{V} = A_S \cdot v_S$

Another important aspect of sampling is that only part of the flowing gas in the measurement grid is removed. As the gas velocity and the dust flow can only be measured at certain points in the measurement cross section, the measurement points chosen must be representative. The gas and dust flow rates must be determined as accurately as possible using a suitable number and position of measurement points within acceptable time scales and at acceptable labor cost.

For this measurement process, the total area is subdivided into smaller regions of equal area whose mid-points are used as the measurement points (grid measurement). It is assumed that at each point the gas velocity and the mass flow rate represent the mean for the small area. For rectangular cross sections, the areas should be similar. The arrangement of measurement points in a rectangular cross section is shown in Figure 11. With a circular cross section, the measurement plane is divided into annular regions of equal area. The typical measurement points for circular cross sections are illustrated in Figure 12. Measurements are generally carried out on two perpendicular axes (diameters). For very uniform flow, a single measurement axis may suffice, although in exceptional cases more than two axes may be required.

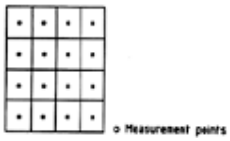

Figure 11. Arrangement of measurement points in a grid-based measurement in a rectangular cross section

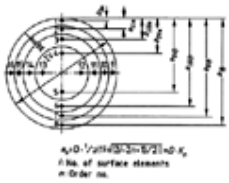

Figure 12. Arrangement of measurement points in a grid-based measurement in a circular section with two measurement axes

$$a_n = D^1/_2\left(1 \pm \frac{2i-2h+1}{2i}\right) = DK_n$$

where i = number of surface elements, n = ordinal number

The necessary number of equal rectangles or annular areas depends on the flow distribution and the area of the measured cross section. Increased nonuniformity of gas and dust distribution means that more measurement points must be provided.

For a cross-sectional of 0.03 – 1 m^2, four measurement points are required, and for >1 m^2, four per square meter.

However, as a rule, 20 measurement points are sufficient for larger cross-sectional areas.

6.7.2. *Discontinuous Methods*

Particles can be measured discontinuously by various gravimetric methods, by optical methods in which the DIN smoke number is estimated visually, and by visual estimation of the gray value (Ringelmann method).

Gravimetric Determination. A probe removes a sample stream of dust-laden gas from the total volume flow, and the particles present in the sampled gas are collected on filters made of glass, quartz fiber, or quartz wadding. The dust content is found from its mass and the volume of sampled gas [428], [452-454]. In this method, particles are recovered under the conditions prevailing in the exhaust gas duct, and the filter is located immediately after the sampling probe.

These methods are used to determine particulate emissions, to investigate the efficiency of particle separators, and as reference methods for calibrating recording dust content measurement equipment.

A sampling arrangement for moist, saturated off-gases is described in VDI 2066, Sheet 3, in which the filter is located outside the duct. The sampling probe and the filter are heated so that the particles are collected at temperatures above the exhaust gas temperature (i.e., “quasi-dry”).

A method for determining the dust content of the unpurified gas (up to 200 g/m^3), i.e., before the dust filters, is described in VDI 2066, Sheet 3. A sampling flow rate of ca. 4 m^3/h is used with a double filter cartridge with higher dust storage capacity.

The measurement method described in Sheet 2 is used in practice for dust contents of 1 – 1000 mg/m^3. For concentrations of <20 mg/m^3, a packed cartridge/flat filter combination is used. The arrangement described in Sheets 2 and 7 is illustrated in Figure 13.

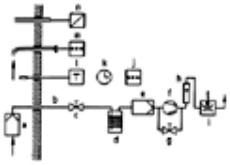

Figure 13. Sampling equipment (VDI 2066, Sheets 2 and 7)

a) Filter head with probe and diffuser; b) Gas sampling pipe; c) Shut-off valve; d) Drying tower (if necessary); e) Protective filter for suction equipment; f) Suction equipment (leak-proof); g) Bypass control valve; h) Flowmeter (rotameter); i) Gas meter with thermometer; j) Barometer; k) Timing instrument; l) Temperature sensor with display; m) Prandtl tube with micromanometer (alternatively anemometer); n) Gas measurement equipment (if necessary)

This arrangement requires leak-proof suction equipment with a by-pass and a flow-control valve. If the suction equipment follows the gas meter, the volume flow rate can be controlled by bleeding in air.

The method of VDI 2066, Sheet 7 is suitable for measuring low dust contents (0.1 – 20 mg/m^3). For the range 0.1 – 5 mg/m^3, weighing must be very accurate. The system is designed for sampling rates up to 4 m^3/h. A typical flat filter head is shown in Figure 14. The same method is used for measuring asbestos dust [455].

Figure 14. A typical flat filter head

a) Exchangeable sampling probe; b) Inlet cone; c) Union nut; d) Seal; e) Tension ring; f) Filter holder; g) Flat filter; h) End housing and connection piece

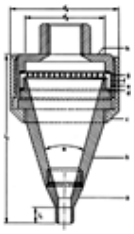

d_1 = diameter of housing (65 mm); d_2 = diameter of filter (50 mm); l_1 = length of inlet tube (10 mm), l_2 = length of flat filter head (100 mm); α = angle of inlet cone (30°)

The relative detection limits for the various dust measurement methods with the usual operating time of 30 min are 1 – 2 mg/m^3 with the 12 or 4 m^3/h filter head, 0.3 – 0.5 mg/m^3 with the 40 m^3/h filter head, and 0.1 – 0.2 mg/m^3 with the flat filter head.

Lower detection limits can be achieved by increasing the gas sampling rate or using longer running times. With a dust content of ca. 10 mg/m^3, duplicate determinations indicate approximate measurement uncertainties *U* (95 %) of between ± (5 – 10) % with the 4 – 40 m^3/h equipment, and between ± (4 – 8) % with the flat filter head equipment.

Optical Methods. In the past, visual methods were prescribed for monitoring dust and smoke emissions. Official regulations at one time specified the Ringelmann method for checking the acceptability of exhaust gas plumes. However, now that dust contents are much lower, this requirement is now regarded as obsolete.

The Ringelmann scale is a series of gray values between white and black in five steps from 0 % (white) to 100 % (black). Estimation is carried out with a chart having four areas covered with black hatched lines that occlude 20, 40, 60, and 80 % of the area, giving the impression to the eye of different gray values.

For combustion plants running on light heating oil, smoke emission limits must be complied with. Smoke emission is represented by the smoke number (DIN 51 402, Part 1), which is determined with a manual smoke pump. The value is determined from the change in the reflectivity of the measuring filter. An increase in the smoke number of 1 unit is defined as a decrease of 10 % in the reflectivity of the measurement filter.

This method is mainly designed for monitoring small furnaces. For plants fired with light heating oil with a thermal capacity ≥ 5 MW, TA Luft requires continuous monitoring of the smoke number. These recording devices must be calibrated against a conventional measurement method. The DIN method using manual smoke pumps is unsuitable. A modified method is described in VDI 2066, Sheet 8 [456], based on DIN 51 402, Part 1. It is a photometric method developed from a technique for measuring low dust contents [454].

6.7.3. *Continuous Methods*

Where there is a legal requirement for equipment for continuously measuring emissions, the conditions described in Section Requirements for Continuous Measurement and Associated Computer Systems must be observed. Measurements carried out at the instigation of internal plant management are not affected.

6.7.3.1. *Optical Methods*

Optical monitoring of dust emissions is widely used, based on transmission/opacity, extinction, or scattered light.

In in-situ photometric dust measurement, the intensity of a beam of light passed through the off-gas duct is decreased as a result of absorption and scattering by the particles. The ratio of attenuated to emitted light is the optical transmission *T*. The quantity 1 – *T* is known as the opacity. The value log (1/*T*) is the extinction *E*.

The relationship between the length of the light beam *L* and the transmission *T* for constant dust properties in the exhaust gas is expressed by the Lambert – Beer law:

$$T = \exp(-E) = \exp(-\varepsilon L)$$

The extinction coefficient ϵ depends on the properties of the light used, the properties of the dust to be measured (e.g., shape, color, and particle size distribution), and on the dust content *c*. It has been found that, within certain limits, there is a linear relationship between dust content *c* and extinction coefficient ϵ if other parameters are constant. If ϵ' is the proportionality constant:

$$T = \exp(-\varepsilon' cL) \text{ or } E = \varepsilon' cL$$

Depending on the application, the equipment can be designed for smoke density measurement or dust content measurement. In smoke density measurement equipment, only the transmission is measured. The object is to monitor the functioning of filter equipment and to determine the opacity of the exhaust gas, e.g., by determining its visibility threshold or by using the Ringelmann scale. There are several approved types of measuring equipment [431]. Manufacturers include Durag and Sick.

In dust content measurement equipment, the extinction is measured. An individual calibration by means of a separate determination of the dust content enables the measured extinction values to be calibrated, taking into account the effect of particle properties. Particle-size distributions, densities, colors, and refractive indices differ from plant to plant, but are comparatively constant for the same plant.

The usual arrangement for photometric dust content measurement apparatus in situ is illustrated in Figure 15. The measurement head is installed on one side of the exhaust gas duct and the reflector on the other. The light source and photoelectric detector are located close together in the measurement head. The reflector consists of an arrangement of triple mirrors (retroreflectors) which reverse the direction of the light beam. Thus, the light beam passes through the measurement path to the reflector and is returned. The reference light beam passes through a dust-free reference path inside the measurement head. A half-silvered mirror inserted alternately into the measurement and reference light paths causes the two light beams arrive at the photoelectric detector out of phase. The electrical signal produced by the detector is processed so that the final signal is proportional to the extinction.

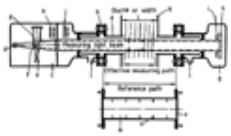

Figure 15. Schematic diagram of in situ equipment for photometric dust measurement

a) Mounting flange; b) Adjustment element; c) Measurement head; d) Light source; e) Light receiver; f) Half-silvered mirror; g) Reflector head; h) Reflector; i) Reflector diaphragm; j) Zero point reflector; k) Control filter; l) Purge air connection; m) Dust-free reference path; n) Scattered light diaphragm; o) Blank flange; p) Reference light beam; q) Exhaust gas duct

Several types of measurement equipment have been approved [431]. Manufacturers include Durag, Mannesmann/H.u.B., and Sick.

Scattered light measurement exploits the fact that when a collimated light beam is passed through a dust-laden measured volume, part of the light is scattered in various directions. The intensity of the light scattered at a given angle to the original direction depends on the intensity, wavelength, and polarization of the incident light, on the angle itself, and also on the shape, particle-size distribution, and refractive index of the particles. The principle of scattered light measurement is illustrated in Figure 16. As with extinction measurement, a linear relationship exists between the dust content and scattered light intensity within certain limits. The sensitivity of light scattering measurement equipment is considerably higher than that of extinction measurement equipment. Depending on the method used, the detection limit can be as low as 0.1 mg/m^3. In practical applications, dust concentrations as high as 100 mg/m^3 have been determined. However, with both extractive sampling and in-situ equipment (small measurement volumes), care must be taken to ensure that sampling is representative.

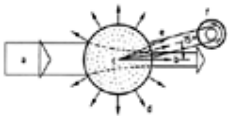

Figure 16. Principle of scattered light measurement

a) Incident light beam on sample; b) Light beam transmitted through sample; c) Sample; d) Scattered light; e) Measurement of selective scattered light beam; f) Photocell

The function scheme of the KTN/KTNR photometer manufactured by Sigrist is illustrated in Figure 17. This operates according to the double beam system, with forward scattering at 15° and automatic compensation.

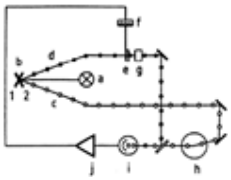

Figure 17. Functional scheme of a scattered light photometer

a) Light source; b) Flicker mirror; c) Measurement beam; d) Reference beam; e) Light attenuator; f) Display instrument; g) Reference standard; h) Measurement chamber; i) Light detector; j) Measurement amplifier

1) Measurement position of flicker mirror; 2) Reference position of flicker mirror

The beam from a light source (a) passes through an optical path to the “flicker” mirror (b). In position (1), this deflects the incident light as the measuring beam (c) via an optical path into the measuring chamber (h). Part of the scattered light produced by the sample is received at an angle of 15° to the forward direction by a light detector (i) and measured. In position (2), the incident light is passed as a reference beam (d) through a light attenuator (e) and a reference standard (g) to the light detector. The signal currents produced by the light detector are compared in a measurement amplifier (j) and converted into a control signal which changes the reference beam by means of a light attenuator until its intensity corresponds to the intensity of the light which is scattered by the sample and falls on the light detector.

The measurement method and calibration are described in VDI 2066, Sheet 6 [457]. The equipment operates with an extractive sampling system. The sampling system can be heated to 180 °C, so that even water vapor saturated exhaust gases can be continuously measured [431], [434]. In another approved variation, the equipment is used to measure the smoke number [458].

In-situ scattered light photometers are produced by Durag and Sick for measuring both low dust contents and smoke numbers.

A schematic diagram of equipment for in-situ scattered light measurement is shown in Figure 18 [456]. This is suitable for measuring both dust contents and smoke numbers. Light is passed directly through the exhaust gas duct and received and absorbed on the opposite side by a light trap. The scattered light from a definite measured volume in the exhaust gas duct is received and measured at a certain angle (e.g., 90° scattering or back-scattering). Because of the geometry of the equipment, the measured volume lies near to the duct wall. Problems can therefore arise on installing this equipment in thick-walled ducts (e.g., brick-lined) or double-walled pipe. In-situ light scattering equipment is sensitive and suitable for measuring low dust contents (<10 mg/m^3) and low smoke numbers (< 3 or < 1). Unlike extinction equipment, it can be operated in narrow ducts. Various types of equipment have been approved [434], [435], [449], [458].

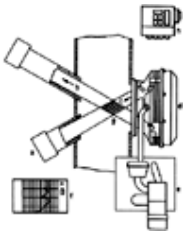

Figure 18. Schematic diagram of in-situ scattered light measurement equipment

a) Light trap; b) Direct light; c) Scattered light; d) Scattered light measurement; e) Purge air unit; f) Pen chart; g) Measurement volume; h) Display

6.7.3.2. *Radiometric Methods*

Radiometric measurement equipment is manufactured by the companies Verewa (dust content, smoke number) and FAG Kugelfischer (dust content), and has been officially approved [451]. (Smoke number measurement equipment is based on reflection photometry rather than β-radiation.) In dust measurement based on β-radiation absorption, a sample gas stream is taken by a probe and passed through a strip of filter paper which collects the dust particles. The paper strip is moved along in a stepwise fashion. The quantity of dust deposited on the filter is measured by the attenuation of β-radiation (source: ^{14}C or ^{85}Kr) passed through the filter. The attenuated radiation is detected by a Geiger counter, and the attenuation is a measure of the mass of dust collected (Fig. 19).

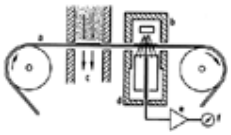

Figure 19. Principle of radiometric measurement of dust by β-radiation absorption

a) Filter band; b) β-Emitter; c) Partial stream; d) Detector; e) Amplifier; f) Display

As with the other methods, calibration is based on a conventional measurement method, carried out as a grid measurement over the duct cross section.

In the Beta-Staubmeter F 904 equipment (Verewa), a sample stream is taken isokinetically from the gas flowing in the duct, and preheated dilution air is added to reduce the dew point of saturated exhaust gases. The gas is then filtered and the dust collected is measured using β-radiation. The result is the mean value of the dust content based on the cycle time used. The dust measuring equipment FH 62 E – NA (FAG) measures the actual dust content radiometrically. After the zero has been automatically set, a sample stream of gas is removed from the stack isokinetically. It is then heated, diluted with clean air, and aspirated through a measuring filter. The continuously increasing amount of dust on the filter causes the integral signal obtained to increase with time. An instantaneous dust concentration is obtained from the signals for rate of mass increase and flow rate.

In another continuous measurement method, the gas is extracted by a sampling probe which simultaneously measures the gas velocity. The dust-laden gas is trapped on a filter, which is changed at a time interval depending on the amount of dust collected. The dust is determined by measuring the differential pressure across the filter (LSP apparatus, Becker Verfahrenstechnik).

6.7.4. ***Particle Size of Dusts***

Laboratory methods for determining the particle size of dusts include microscopy, sieving, air classification, sedimentation, and scattered light [459]. Dust quantities of 2 – 20 g are necessary. With low dust contents, relatively large volumes of the exhaust gas must be filtered. However, the particle size distribution can be changed by the sampling and collection processes (e.g., by agglomeration). Redispersion in the laboratory may result in the measured particle size differing from that in the gas [460]. It is therefore advantageous to measure the particle-size distribution in the gas. Various types of measurement equipment have been developed, e.g., light scattering, diffuser batteries, electrostatic aerosol analyzers, and aerosol centrifuges. Sampling must be carried out with a suction probe, as the equipment cannot be directly installed in the exhaust gas duct. A high proportion of particles of diameter > 3 µm are lost during collection by deposition on the walls of the probe and on the inlet to the measuring equipment. The method is therefore limited to particles of diameter < 3 µm, although particles in the range <1 to 100 µm can be expected in emissions from industrial and commercial plants. Aerosol measurement equipment generally cannot be used in hot and chemically aggressive exhaust gases. The sampled gas stream should not be strongly cooled because condensation of moisture and other substances can change the particle-size distribution. Cascade impactors are suitable for sampling a partial gas stream and separating the particles into mass fractions. They can be used directly in hot, chemically aggressive gases [461]. Gravimetric evaluation of cascade impactor measurements gives mass distributions as a function of aerodynamic diameter of the particles. However, this method does not replace the measurement of total dust content (VDI 2066, Sheets 2, 3, and 7). The cascade impactor method is described in detail in VDI 2066, Sheet 5 [462]. Cascade impactors can separate particles into several fractions between ca. 0.3 µm and 15 µm, depending on the design, the type of gas, and the particle properties. Larger particles (> 10 µm) interfere with the fractionation process and must be removed by a preliminary separation process [463]. Low-pressure impactors are used for particles with a lower size range.

The impactor exploits differences in the inertia of particles to separate them into fractions. An impactor stage consists of a jet and an impact plate. The particles are accelerated in the jet and those having sufficient inertia strike the impact plate and are collected (see Fig. 20).

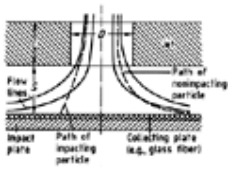

Figure 20. Principle of impaction of particles

The particle fractionation depends on jet width *D*, jet length *L*, and the distance *S* between the impact plate and the jet. Cascade impactors consist of several impactors in series, whereby particles of decreasing inertia are separated in the successive stages, giving fractions of decreasing particle size. Particles not collected on the impact plates are trapped on a filter.

Evaluation of impactor measurements is based on experimentally determined median values of the Stokes number St_{50} and the limiting aerodynamic particle diameter d_{ae50} of an impact stage. For an impactor with several circular jets per stage, a uniform value of St_{50} = 0.26 is used for all impactor stages for reasons of comparability [461].

The measurement setup is specified by VDI 2066, Sheet 7. The gas passes through the sampling probe, the cascade impactor, and then the angle bend. Without a preliminary separation stage, the cascade impactor described in VDI 2066, Sheet 5 can be used for dust contents of 1 – 2 mg/m^3. The impactor must not be overloaded. The maximum permissible load per stage is ca. 100 mg. If there is a preliminary separation stage, the dust content can be between 5 mg/m^3 and 25 g/m^3, depending on the fraction of coarse particles.

6.7.5. *Heavy Metals as Dust Constituents*

It is a legal requirement that all emissions of metals, metalloids, and their compounds should be measured [422-424]. Their emission limit values lie in the range 0.05 – 5 mg/m^3. These limits are defined as the sum of the concentrations of the substances in the form of dust, gas, and vapor. The duration of an individual measurement should normally be 30 min and should not exceed 2 h. The most important elements are antimony, arsenic, beryllium, lead, cadmium, chromium, cobalt, copper, manganese, nickel, palladium, platinum, mercury, rhodium, selenium, tellurium, thallium, vanadium, zinc, and tin. A proven measurement method is described in VDI 3868, Sheets 1 and 2 [464], [465]. It consists of a particle collection system (VDI 2066) and an absorption stage made up of a battery of wash bottles, as used for measuring emissions of gases such as SO_2 and HCl (Fig. 21). The exhaust gas is sampled isokinetically by a probe. The particles are retained by a filter system as specified in VDI 2066, Sheet 2 or Sheet 7. The materials passing through the filter are drawn through a heated sampling tube. A bypass stream is passed through one or more absorption systems in parallel, consisting of at least three gas-absorption bottles in series. The method enables the mass of the dust and of its constituents to be determined both per unit of time and per unit of volume. Two types of absorption solution are used:

1. Solution A: 3 volumes hydrochloric acid mixed with 1 volume nitric acid, 1 volume of this mixture being diluted with 9 volumes water
2. Solution B: 1 volume nitric acid mixed with 1 volume hydrogen peroxide, 2 volumes of this mixture being diluted with 13 volumes water

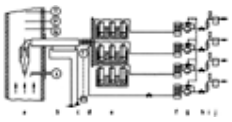

Figure 21. Equipment for the simultaneous collection of particulate materials and materials passing through the filter

a) System for retaining particles; b) Sample pipe with controlled heating; c) Temperature measurement and control for sampled gas stream; d) Adapter with insulation and heating if necessary, e) Absorption system; f) Condensate separator; g) Leak-proof pumping system with controllers; h) Isolation valve; i) Gas volume flow rate measurement; j) Gas meter with thermometer

T = temperature measurement; P = pressure measurement; G = gas analysis; v = gas velocity measurement

Mercury is determined with a solution of 20 g $KMnO_4$ in 1 L dilute sulfuric acid (VDI 3868, Sheet 2).

Under the usual sampling conditions, the detection limit for metals is usually <0.01 mg/m^3. The method described in Guideline VDI 3868, Sheet 2 for mercury is suitable for concentrations of >5 µg/m^3 [465]. Both methods are especially suitable for measurements on exhaust gases from thermal processes in which the constituent elements can accumulate on the smallest particles as well as being present in gaseous and vapor form.

The most important methods of analysis include X-ray fluorescence (XRF), atomic emission spectrometry with inductively coupled plasma source (ICP – AES), atomic absorption spectrometry (F-AAS and G-AAS), and instrumented neutron activation analysis (INAA). Analytical determination of the elements is described in VDI 2268 [466-469]. Because of the required detection sensitivity, atomic absorption is used preferentially for many components. Methods used include flame AAS, graphite-tube AAS, the hydride and cold vapor technique, and ICP – AES. In cases where even atomic absorption spectrometry does not enable all trace elements to be determined with sufficient sensitivity, INAA can be used if available.

6.8. Measurement of Gaseous Inorganic Emissions

6.8.1. *Sampling*

A measurement grid is necessary when sampling gaseous emissions if the concentration is not uniformly distributed over the measured cross section [439]. There are two methods of determining the mean concentration. Either the local

emission concentration c_i and the local velocity u_i (together with the oxygen content) are determined at each measuring point, or a partial mass m_i is determined by using a collection phase, with a sampling rate proportional to the velocity. However, this does not necessarily mean isokinetic sampling (as with dust measurement), but a sampling rate proportional to the velocity or the mass.

Generally, gaseous emissions are sampled at a velocity considerably lower than that of the exhaust gas. Also, the probe is not parallel to the flow direction as with dust measurement, but at an angle <90° to the flow direction. The sampling volume rate is typically 30 – 120 L/h. The probe is made of, e.g., glass, quartz, titanium, stainless steel, or PTFE. The probes and associated pipework are usually heated to prevent condensation.

With recording analytical equipment that operates extractively, correct functioning and the results obtained depend on the processing of the sampled gas, which must be dust-free, dry, and free of corrosive materials before it enters the analyzer. Furthermore, the volume flow rate, pressure, and temperature must generally lie within limits imposed by the measurement method. If the dew point of the sampled gas is higher than the temperature of those parts of the analysis system with which it comes into contact, the gas must be cooled at the exit of the sampling equipment before it enters the analyzer to condense the moisture. The possibility as the substances being measured dissolving in the condensate must be considered.

6.8.2. *Sulfur Oxides*

When sulfur-containing fuels and wastes are burnt, sulfur dioxide is formed. Important emission sources are industrial furnaces, chemical plant, the cement industry, metal smelting, ceramics, the cotton industry, domestic heating, and traffic. Part of the SO_2 may be oxidized to SO_3. It is normally a legal requirement that the total sulfur oxides (SO_2 + SO_3) should be determined and expressed as SO_2.

6.8.2.1. *Discontinuous Methods*

At present, four manual analytical methods are available for determining SO_2, and these are also used as reference methods for the calibration of continuously measuring equipment. They are described in the Guideline VDI 2462, Sheets 1, 2, 3, and 8 [470-473] (Table 12).

Table 12. Determination of SO_2 (VDI 2462)

Sheet	Brief title	Measurement principle	Remarks
1	Iodine – thiosulfate method	bubbling into I_2, oxidation to H_2SO_4, back-titration with sodium thiosulfate, detection limit 60 mg SO_2/m^3	cross-sensitivity towards hydrogen sulfide, nitrogen oxides, halogens; suitable for higher SO_2 concentrations
2	Hydrogen peroxide oxidation + titration	oxidation to H_2SO_4 by H_2O_2, titration with NaOH, detection limit 30 mg SO_2/m^3	cross-sensitivity towards components that can form acids or bases in the absorption solution; suitable for higher SO_2 concentrations
3	Hydrogen peroxide oxidation + gravimetric determination	oxidation to H_2SO_4 and precipitation of $BaSO_4$, detection limit 40 mg SO_2/m^3	cross-sensitivity towards other sulfur-containing components except SO_3; suitable for higher SO_2 concentrations
8	H_2O_2 – thorine method	oxidation to H_2SO_4, titration with $Ba(ClO_4)_2$ with thorine as indicator, detection limit 0.72 mg SO_2/m^3	cross-sensitivity towards certain anions and metallic cations and other sulfur-containing components except SO_3; suitable for low SO_2 concentrations; no cross-sensitivity towards nitrogen oxides

Because of the low limits of detection and lack of cross-sensitivity towards nitrogen oxides, the method described in Sheet 8 is often used for emission measurements and as a reference method for the calibration of continuous measuring equipment. The gas is sampled through a heated probe, and is then passed through a heated silica wool filter to remove dust and through two wash bottles in series containing a 3 % solution of hydrogen peroxide. The temperature of the heated probe and the silica wool filter is chosen so that condensate formation (acid dew point) is prevented. A temperature of 200 – 220 °C is adequate. The SO_2 in the exhaust gas is quantitatively oxidized to H_2SO_4. This method also determines SO_3 and is thus a summation method for sulfur oxides (SO_2 + SO_3) expressed as SO_2. A schematic diagram of the sampling apparatus is given in Figure 22. Analysis is by titration against a solution of barium perchlorate. A color change takes place in the presence of the added metal indicator thorine when the solubility product of barium sulfate is exceeded.

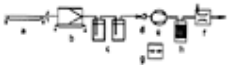

Figure 22. Schematic of sampling equipment for the measurement of sulfur oxides

a) Sampling probe, heated; b) Filter with silica wool packing, heated; c) Gas wash bottle, 250 mL, with grade G 2 frit; d) Flow control valve; e) Suction pump; f) Gas meter with thermometer; g) Barometer; h) Drying tower (only used with dry gas meter)

The separate determination of SO_3, described in Sheet 7 [474], is also based on the thorine reaction. The SO_3 is absorbed in isopropanol, which does not absorb SO_2. The SO_3 in isopropanol solution is titrated with barium perchlorate with thorine as indicator, as described for SO_2. The relative detection limit is ca. 2 mg/m^3. Volatile sulfates that pass through the filter are also determined.

6.8.2.2. *Continuous Methods*

Several photometric measurement principles can be used for continuously measuring SO_2 emissions, with extractive and in-situ sampling. Conductivity methods are also used, with extractive sampling.

Photometry with Extractive Sampling. All heteroatomic molecules such as CO, SO_2, and NO have a characteristic absorption spectrum in the infrared; SO_2 and NO also absorb in the ultraviolet.

The simplest arrangement for an extractive absorption photometer is illustrated in Figure 23. The light from a radiation source passes through an optical filter to limit its spectral range to that required for the measured component, passes through an optical cell through which the gas being analyzed flows, and falls on a photodetector linked to an electronic signal processing device. Part of the light is absorbed by the pollutant molecules. The light attentuation is therefore a measure of the concentration of the pollutant. In this simple arrangement, very small changes in the radiation emission and the receiver sensitivity lead to unacceptably high zero point errors. Measures to avoid this problem are based either on a periodic zero point correction or a reference standard in the form of a second reference filter or a reference gas. This reference standard can either be inserted from time to time (and thus out of phase) into the radiation path, or it can be located in a parallel reference beam. The first arrangement is known as a single-beam photometer, and the second a double-beam photometer. The photometric analysis equipment can be caused to be sensitive to a selected component using either dispersive or nondispersive systems.

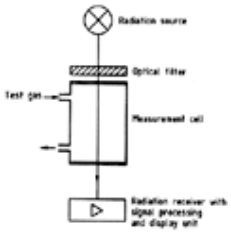

Figure 23. Absorption photometer (schematic)

Dispersive Systems. In dispersive systems, the light from a broad band radiation source is broken down into its spectral components by a spectrometer before the measurement is carried out. Prisms, gratings, or interference filters can be used to perform the dispersion.

Nondispersive Methods. In nondispersive methods, wavelength selection is carried out by the substances being measured, which are stored in the equipment. There are three methods of storage:

1. In the nondispersive infrared (NDIR) method, a gas-filled optical cell is used for storage (Fig. 24). The radiation is deflected to pass through the gas-filled chambers, and modulation by a rotating wheel produces periodic pressure fluctuations in the detector chambers. These are detected either by a membrane capacitor or a microflow detector which detects a flow between the two chambers caused by the pressure fluctuations, and converts this into an electrical signal.
2. In the gas filter correlation (GFC) method (Fig. 25), a gas-filled optical cell attached to a rotating wheel acts as the storage device. This is periodically brought into the radiation path, alternating with a nitrogen-filled gas filter.
3. In the nondispersive ultraviolet (NDUV) method, the component to be measured is stored in the radiation source. Gas-filled discharge lamps are used which emit spectral lines characteristic of the component being measured. UV photometers operate with 1 or 2 beams and with 1 or 2 photoelectric detectors. Some photometers use a combination of two of the optical selection methods described above.

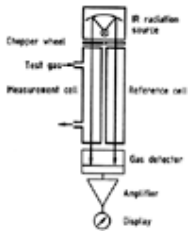

Figure 24. NDIR photometer (schematic)

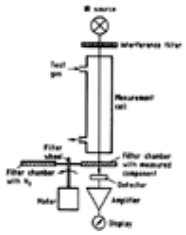

Figure 25. Gas filter correlation method (schematic)

In-Situ Photometry. The absorption measurement path in in-situ photometers is located directly in the exhaust gas duct, so that the gas sample need not be transported to the optical cell by a sampling system. The photometer, consisting of a photometric detector, wavelength selection equipment, and electronics to calculate the result is situated outside the exhaust gas duct. In the UV range, wavelength selection is by diffraction grating. In the IR range, interference filters and gas-filled filter chambers (as with the GFC method) are also used.

Two arrangements are illustrated in Figure 26. In both cases, the windows on each side of the absorption path are protected from dust soiling by an air curtain, as in the photometric measurement of dust.

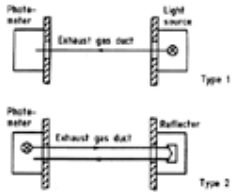

Figure 26. In-situ photometer (schematic)

Conductometry. In conductometric gas analysis, the gas sample is passed into a liquid reagent, and the change in conductivity after reaction of the liquid with the gas is measured.

In continuous conductometry, the sampled gas and the reagent liquid are continuously fed into one end of a reaction tube. As the change in conductivity also depends on the ratio of the volume flow rates, both flow rates must be kept constant by a control device. The conductivity before addition of the gas phase and after its removal is measured by two pairs of electrodes (Fig. 27). The conductivity difference, which is proportional to the concentration of the measured component, is calculated by an electronic system. As conductivity is strongly temperature dependent, the effect of temperature must be compensated.

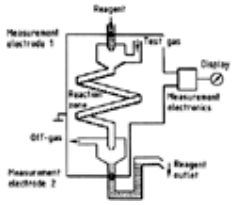

Figure 27. Conductometric measurement (schematic)

Details of officially approved equipment for measuring SO_2 have been published [475] (Table 13).

Table 13. Equipment for continuous measurement of SO_2

Manufacturer	Type	Principle
Extractive sampling		
NDIR method		
Afriso	Infralyt 1210	double-beam equipment with two radiation sources
Mannesmann/H.u.B	URAS 3 G, 3 E	double-beam difference measurement method with one radiation source and double-layer detector
Maihak	UNOR 6 N So_2	double-beam difference measurement method with one radiation source and double-layer detector
	UNOR 6 N - F	
	UNOR 600/610	single-beam analyzer with double-layer detector
Perkin – Elmer	MCS 100 HW	single-beam photometer, bifrequency method, two interference filters per measured component, sequential measurement

Siemens	Ultramat 32	double-beam difference measurements method with one radiation source and double-layer detector
	Ultramat 5	
	Ultramat 21, 22	single-beam analyzer with double-layer detector
NDUV method		
Maihak	Defor 3	double-beam difference measurement method with one radiation source
Rosemount	SO_2 – UV Binos	double-beam difference measurement method with one radiation source
Conductivity		
Wösthoff	Mikrogas – SO_2	absorption in H_2O_2, change in electrical conductivity
In situ sampling		
Opsis	Opsis AR 600	single-beam photometer, absorption spectrum in certain UV wavelength range (diffraction grading)
Sick	GM 21, GM 30	single-beam photometer with reflector, absorption spectrum in certain UV wavelength range, spectral dispersion of the resulting light beam

6.8.3. *Nitrogen Oxides*

Nitrogen oxides are formed in all thermal combustion processes in air and by oxidation of the combined nitrogen in fuel. Emission sources include furnaces burning solid, liquid, and gaseous fuels; internal combustion engines; power plants; cement works, metallurgical plant; waste incinerators; and chemical plant. In general, the concentration of NO_2 is < 10 %, but in a few cases (e.g., fertilizer production), it can be considerably higher.

Discontinuous Methods. Four manual analytical methods are described in VDI 2456. The phenoldisulfonic acid method and the titration method described in Sheets 1 and 2, respectively, are suitable for high concentrations in the g/m^3 range. The sodium salicylate method and the dimethylphenol method described in Sheets 8 and 9 are often used as reference methods for calibrating continuously measuring equipment on account of their low detection limits [476-479].

In the sodium salicylate method, total NO + NO_2 is determined. Sampling can be performed by a gas collection tube or an evacuated absorption vessel. When the gas collection tube is used, a volume equal to about ten times the volume of the tube is drawn through the sampling apparatus.

If an evacuated vessel is used, it is connected via a three-way tap between the solid filter and the pump. Its volume must be corrected for temperature and pressure. Oxidation to nitric acid is then carried out for 24 h. The nitric acid is reacted with sodium salicylate in a solution of H_2O_2 acidified with sulfuric acid to form nitrosalicylic acid. The anion is then measured photometrically. The relative detection limit, using 1 L gas collection tubes, is ca. 2 mg/m^3 (as NO_2). The total time between sampling and obtaining the result is 2 d.

In the dimethylphenol method, ozone is used as oxidizing agent. NO and NO_2 are oxidized by ozone in the gas phase to form dinitrogen pentoxide. The reaction takes place in 5 – 10 min. The N_2O_5 is absorbed in water, forming nitric acid. Samples are taken through a critical jet into an evacuated gas sample container. A schematic of the sampling and oxidation equipment is given in Figures 28 and 29.

Figure 28. Sampling equipment for measurement of nitrogen oxide emissions (schematic)

a) Sampling probe, heated if necessary; b) Solid filter, heated if necessary; c) Test gas ducting, heated if necessary; d) T-piece of glass, heated if necessary; e) Critical jet, heated if necessary; f) Gas sample container with isolation valves; g) Suction pump; h) Thermometer; i) Pressure measurement (vacuum gauge); j) Tubing as short as possible; k) Vacuum pump

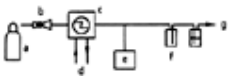

Figure 29. Example of oxidation equipment

a) Oxygen cylinder; b) Pressure reducer; c) Ozone generator; d) Water cooling (if necessary); e) Gas sample container with isolation valve; f) Wash bottles, the last of which containing; water g) Waste gas

The evacuated gas sample container is filled in the bypass to ca. 500 mbar with the sampled material, and, after temperature adjustment (ca. 5 min), is then filled with ozone until the pressures are equalized. Oxidation is complete after ca. 5 min. The nitric acid formed is reacted with 2,6-dimethylphenol (DMP) in a solution containing sulfuric and phosphoric acids to form 4-nitro-2,6-dimethylphenol. The 4-nitro-2,6-dimethylphenolate anion formed in alkaline solution is measured photometrically at ca. 430 nm. The relative detection limit is 15 mg/m^3, expressed as NO_2.

Ammonia at levels up to 150 ppm does not cause problems. If high concentrations of SO_2 are present in the exhaust gas, there is a risk of formation of other nitrogen-containing products (e.g., N_2O). It is then preferable, after pressure measurement, first to fill the evacuated gas sample container with ozone to ca. 250 – 300 mbar, then to add the gas being measured to ca. 700 mbar, and then to fill with ozone until the pressure is equalized.

As the time required to obtain an analytical result is short, this method is widely used for calibrating continuously measuring equipment in situ.

Continuous Methods. There are several methods of continuous measurement which use both extractive and in-situ sampling techniques. The measurement principles used are infrared and ultraviolet absorption, chemiluminescence reaction, and amperometric measurements with electrochemical cells. As a rule, only the component NO is selectively measured. However, equipment for selectively measuring NO_2 is also commercially available. By combining two analyzers in a single item of equipment, NO and NO_2 can be measured simultaneously. Various types of equipment measure NO and NO_2 sequentially. In the chemiluminescence method, a converter can be incorporated for reducing NO_2, so that total NO + NO_2 is measured [480]. Legal requirements are concerned with the determination of total nitrogen oxides (NO + NO_2) only. The manual method described above can be used to determine the concentration of NO_2 for calibration purposes if NO-selective equipment is used for continuous monitoring, provided that the proportion of NO_2 does not exceed 5 – 10 %. Measurement by IR and UV absorption is described in Section Sulfur Oxides.

The chemiluminescence method is based on the reaction of nitrogen monoxide with ozone:

$$NO + O_3 \rightarrow NO_2 + O_2$$

Some of the nitrogen dioxide molecules formed (ca. 10 %) are in an excited state (NO^*_2) The excitation energy is partially released in the form of chemiluminescence

$$NO_2^* \rightarrow NO_2 + h\nu$$

At constant pressure, constant volume flow rate, and with a sufficient excess of ozone, measurement of the intensity of the radiation in the wavelength range 600 – 660 nm enables the concentration of nitrogen monoxide to be selectively measured. Commercial equipment operates at a pressure equal to or less than atmospheric pressure in the measurement chamber [481], [482]. Equipment legally approved for NO/NO_2 measurement is listed in Table 14.

Table 14. Equipment for continuous measurement of NO/NO_2

Manufacturer	Type	Principles
Extractive sampling		
NDIR method		
Maihak	UNOR 6 N-F	double-beam difference measurement method with one radiation source and double-layer detector
	UNOR 6 N - NO	single beam analyzer with double-layer detector
	UNOR 600/610	
Mannesmann/H.u.B.	URAS 3G NO	double-beam difference measurement method with one radiation source and double-layer detector
	URAS 3K NO	
	URAS 10 P	
Siemens	Ultramat 32	double-beam difference measurement method with one radiation source and double-layer detector
	Ultramat 5	single-beam analyzer with double-layer detector
	Ultramat 21, 22	
Perkin Elmer	Spectran 647 IR	single-beam photometer, gas filter correlation method with two miniature cells per measured component, sequential measurement
	MCS 100 HW	
	MCS 100 CD	
Rosemount	NO – IR Binos	double-beam difference measurement method with one radiation source and measurement and compensation chamber
NDUV method		
Mannesmann, H.u.B.	Radas 1 G	gas filter correlation with moving filter cell
	Radas 1 G NO	
Chemiluminescence		
Rosemount	951	chemiluminescence

Eco Physics	CLS 700 El ht	chemiluminescence, two-channel equipment for NO and NO_2 (with converter), separate reaction chambers
Electrochemical cells		
AEG	NO_x-Monitor 4000	three gas diffusion cells with SO_2 absorption filter (cross-sensitivity)
MSI/Elektronik	MS/5600	three electrode micro fuel cells
In-situ sampling		
Opsis	Opsis AR 602 Z for SO_2, NO and NO_2	single-beam photometer measurement of the absorption spectrum in certain wavelength regions of the UV
Sick	GM 30	single-beam photometer with reflector, measurement of the absorption spectrum in certain wavelength regions of the UV

Selective measurement of NO_2 can be carried out by the Perkin – Elmer multicomponent system MCS 100 CD, the Opsis multicomponent system AR 602 Z, the Rosemount NO_2-UV Binos, and the combined equipment Binos 1004 (selective for both NO and NO_2). In the chemiluminescence equipment, it is possible to determine NO_2 from the difference sum (NO + NO_2).

6.8.4. *Carbon Monoxide*

Carbon monoxide can be produced by the incomplete combustion of solid, liquid, and gaseous fuels and waste and is therefore likely to occur in the exhaust gases of all combustion plants. Other sources include internal combustion engines and gas turbines. Depending on the raw materials and the technological processes used, carbon monoxide emissions are produced by installations in the cement industry, the metal smelting industry, iron and nonferrous metal treatment plants, and ceramics.

Discontinuous Methods. The iodine pentoxide method described in the VDI 2459, Sheet 7 for the manual determination of carbon monoxide concentration is a reference method [483]. The principle is the oxidation of CO in the sampled gas to CO_2 with liberation of an equivalent quantity of iodine. The sample is passed over iodine pentoxide (I_2O_5) in a furnace at high temperature.

The iodine is absorbed in a solution of sodium hydroxide and oxidized with bromine to iodate. The iodate is reacted with potassium iodide to form iodine, which is titrated against a solution of sodium thiosulfate. Interfering substances are removed by various adsorption methods during sampling.

Either the exhaust gas is passed via the sample probe directly into the CO combustion furnace, or, more commonly, sampling is carried out with a gas collecting container (gas mouse), and the analysis is then carried out in the laboratory. The schematic design is shown in Figure 30. The working range of the method is between 10 mg/m^3 and 1.25 kg/m^3 CO. The relative detection limit is 2.5 mg/m^3 CO with a sample volume of 4 dm^3.

Figure 30. Arrangement of sampling equipment for CO with collecting vessel

a) Sampling probe, heated if necessary; b) Coarse dust filter, heated if necessary; c) Wash bottle with KOH solution; d) Absorption tube packed with soda – lime; e) Pump; f) Flow control valve for bypass; g) Absorption tube packed with IBr and activated carbon; h) Connection to gas sample container

Continuous Methods. The continuous measurement of CO by NDIR absorption is described in Sheets 1 – 5 of VDI 2459, and the measurement principle in Sheet 6. These sheets are now partially obsolete, having been superseded by BMU (Bundesministerium für Umwelt) approval of certain measurement equipment, including that for the determination of CO concentration by the IR absorption method. Two manufacturers use the gas filter correlation system; one type of equipment includes an electrochemical cell (Table 15).

Table 15. Equipment for continuous measurement of CO

Manufacturer	Type	Principle
Extractive sampling		
NDIR method		
Afriso	Infralyt 1210	double-beam apparatus with two radiation sources
Maihak	UNOR 6 N – CO	double-beam difference measurement method with one radiation source and double-layer detector
	UNOR 6 N – F	

	UNOR 600/610	single-beam analyzer with double-layer detector
Mannesmann/H.u.B.	URAS 3 G/E	double-beam difference measurement method with one radiation source and double-layer detector
	URAS 3 K	
	URAS 10 E	
Perkin – Elmer	Spektran 647 IR	single-beam photometer
	MCS 100 HW	gas filter correlation method
	MCS 100 CD	two miniature cells per measured component, sequential measurement
Rosemount	CO-IR-Binos	double-beam difference measurement method with one radiation source and measurement and compensation chamber
Siemens	Ultramat 32	double-beam difference measurement method with one radiation source and double-layer detector
	Ultramet 5	single-beam analyzer with double-layer detector
	Ultramat 21/22	
Electrochemical cell		
MSI	MSI 5600	three-electrode micro fuel-cells
In-situ sampling		
Sick	GM 900/Modell 9200	single-beam photometer, gas filter correlation, cell wheel with four cells in the radiation source

6.8.5. *Fluorine Compounds*

Gaseous fluorine compounds can occur in the exhaust gases of combustion plants burning solid, liquid, and gaseous fuels and waste, chemical plants, metal treatment plants, and plants for processing rocks and earth.

Discontinuous Methods. Fluorine and gaseous inorganic fluorine compounds can be determined by absorption in sodium hydroxide solution followed by measurement with fluoride-sensitive electrodes or by photometry (method of VDI 2470, Sheet 1) [484]. The exhaust gas is sampled by a heated quartz probe, and is then filtered through a heated quartz wool filter and, if necessary, a heated fine filter or membrane filter to remove dust. It is then passed through two or three bottles filled with sodium hydroxide solution. After steam distillation, the fluoride content of the solution is determined photometrically by the Alizarin Complexone method or potentiometrically with a fluoride-ion-sensitive electrode. With the photometric method, cross-sensitivity effects rarely occur, as the very specific detection reaction is carried out after removal of interfering substances. Measurement with electrode chains gives high selectivity, provided fluoride complex formers such as Fe^{3+} or Al^{3+} are not present in the solution in large amounts. The detection limit is 0.05 mg F^{-}/m^3.

Continuous Methods. Two types of equipment have been officially approved for continuous potentiometric measurement. The gas is passed through a buffered electrolyte solution, and the ion concentration, changes by the substances being measured, is determined by an ion-sensitive electrode system.

In the Bran & Luebbe HF analyzer Sensimeter G, the gas is continuously extracted, passed the probe, the sampling line, the fine filter, and the absorption vessel where it is bubbled through a volume of buffer solution which is renewed every six minutes. The buffer solution is then transported to the ion-sensitive measurement cell. Measurement is thus only semicontinuous. To obtain an accurate result, the reagent solution must be carefully conditioned and its volume accurately measured, and the flow rate of the sampled gas must be constant during the concentration phase.

The Compur Ionotox HF equipment of Compur/Bayer for measuring HF emissions continuously measures the concentration of hydrogen fluoride and other inorganic, gaseous, hydrolyzable fluorine compounds. It is equipped with an electrochemical measurement cell containing an ion-sensitive electrode (fluoride electrode) and a reference electrode (silver/silver chloride). The gas being analyzed is passed continuously through an atomizer where it is mixed with the absorption solution. The latter is fed continuously into the gap between the electrodes whereby the electrode potential is measured. The interfering substances HCl, SO_2, H_2O, CO_2, CO, and NO lead to cross-sensitivity, so that the concentration of these substances must be limited.

6.8.6. *Chlorine Compounds*

Gaseous inorganic chlorine compounds are present in the exhaust gases from combustion plants burning solid fuels and waste materials, and also in the exhaust gases from rock and earth processing plants, the metallurgical industries, and the chemical industry.

Discontinuous Methods. Two methods for measuring chlorine are described in the VDI 3488, Sheets 1 and 2 [485], [486]. The methyl orange method is based on the bleaching effect of elemental chlorine on an acidified solution of methyl orange.

In the bromide – iodide method, bromine is liberated by chlorine from an acidified potassium bromide solution and determined iodometrically. These methods are generally unsuitable for exhaust gases from incineration plants because of

their cross-sensitivity to nitrogen dioxide and sulfur dioxide.

A conventional method for measuring gaseous inorganic chlorine compounds is described in VDI 3480, Sheet 1 [487], with three alternative chloride determination methods: (1) Mohr titration, (2) potentiometric titration, and (3) photometric determination with mercury thiocyanate. The exhaust gas is sampled through a heated silica or borosilicate probe and a heated quartz wool filter, and is then passed into an absorption vessel containing distilled water. Methods (2) and (3) are suitable for HCl contents up to 100 mg/m^3. Methods (1) and (2) are suitable for higher concentrations. The relative detection limits are as follows: method 1: 20 mg/m^3, method 2: 2 mg/m^3, method 3: 2.5 mg/m^3.

Continuous Methods. Several proven types of equipment for the continuous measurement of gaseous inorganic chlorine compounds are commercially available. These make use of potentiometry, conductivity, and NDIR gas filter correlation. The Ecometer and Spectran 677 IR equipment are described in detail in VDI 3480, Sheets 2 and 3, in which the principle, design, and function of the equipment, the construction of the measurement station, and the calibration, evaluation, and parameters of the method are discussed [488], [489]. The Mikrogas conductivity equipment for measuring SO_2 is described in VDI 2462, Sheet 5. A summary of the commercially available equipment is given in Table 16.

Table 16. Equipment for continuous measurement of inorganic chlorine components

Manufacturer	Type	Principle
Bran and Luebbe	Sensimeter G	ion-sensitive electrode, cyclic measurement
	Ecometer	ion-sensitive electrode, absorption path, continuous measurement
Perkin – Elmer	Spectran 677 IR MCS 100 HW	NDIR single-beam photometer, gas filter correlation
Wösthoff	Mikrogas HCl	conductivity, absorption in H_2O_2,
	Mikrogas HCl/SO_2	difference measurement (SO_2 + HCl)–SO_2 with two conductivity cells

The Sensimeter G for HCl operates like the HF equipment of the same name, with a timed collection phase in the measurement cycle. The other equipment described here operates continuously and is based on extractive sampling. In the conductometric method there is cross-sensitivity to HF, which must be calibrated for.

6.8.7. *Hydrogen Sulfide*

Hydrogen sulfide can occur in exhaust gases from coking plants, refineries, synthetic fiber plants, and cupola furnaces.

Discontinuous Methods. Two methods for random sample measurement are described in VDI 3486, Sheets 1 and 2. These are manual analytical methods with potentiometric and iodometric titration [490], [491].

In the potentiometric titration, the test gas is first passed through a solution of hydrogen peroxide and sulfuric acid to remove any sulfur dioxide present, and the hydrogen sulfide is then absorbed in sodium hydroxide solution. The sulfide content of the solution is determined by potentiometric titration with silver nitrate. The relative detection limit is 0.5 mg H_2S/m^3. The method is suitable for determining hydrogen sulfide in exhaust gases. Because hydrogen cyanide and carbon disulfide interfere, the method cannot be used for coke oven gas. When determining hydrogen sulfide in combustion gases, the carbon dioxide content should be taken into account. (The CO_2 is absorbed in the alkaline wash bottle and must be subtracted from the sample volume.)

In the iodometric titration, hydrogen sulfide is reacted with cadmium acetate solution to form cadmium sulfide, which is determined iodometrically after removing the liquid phase and dissolution in hydrochloric acid. The relative detection limit is 1 mg H_2S/m^3. The method is generally applicable and is especially suitable for gases with high concentrations of hydrogen sulfide. Collecting the cadmium sulfide precipitate by filtration and analyzing the filter cake eliminates cross-sensitivity effects.

Continuous Methods. Hydrogen sulfide can be continuously measured and recorded by the Maihak Monocolor equipment. It is based on the colorimetric principle and operates semicontinuously, using a collecting phase and a dry reaction. Selectivity is achieved by means of chemically specific color indicators for hydrogen sulfide (lead acetate, silver nitrate, and silver sulfate), evaluated photometrically. The Monocolor equipment described in VDI 3486, Sheet 3 operates with a reagent paper strip impregnated with color indicators chemically selective for the components to be determined. The paper strip passes stepwise through an optical scanning device [492].

6.8.8. *Ammonia*

Emission sources of ammonia include coking plants and chemical fertilizer plants. In furnaces equipped with reduction catalysts and NH_3 addition (SCR) or with NH_3 addition alone (SNCR) to reduce sulfur emissions, unreacted NH_3 can be measured for monitoring purposes.

Discontinuous Measurement. Two methods are described for ambient air measurements in the µg/m^3 range, namely the

indophenol method and the Nessler method [493], [494]. In both, the sampled gas is passed through dilute sulfuric acid, trapping ammonia as ammonium sulfate. In the indophenol method, the sulfate is reacted with phenol and sodium hypochlorite to form a blue indophenol dye whose concentration is determined photometrically at 630 nm.

In the Nessler method, interfering substances (e.g., SO_2, NO, NO_2, and HCl) are removed by distillation of the ammonia from alkaline solution. If H_2S and formaldehyde are present, they are decomposed by potassium manganate(VII) before distillation. This is followed by reaction with the Nessler's reagent and photometric measurement at 450 nm.

Emission measurements are carried out in accordance with VDI 2461. Sampling is performed with heated quartz tubes leading to quartz or glass absorption vessels. The temperature of the sampling system is adjusted to the exhaust gas temperature, but should not exceed 330 °C. The dust is removed by a heated filter after the probe. The absorption solution is 0.1 N sulfuric acid. A flow rate of ca. 2 m^3/h can be achieved by using two impingers in series. The rate is lower if wash bottles are used. Distillation, reaction with Nessler's reagent, and photometric determination at 450 nm are described in [494]. Determination with a sensitive electrode is also possible. For determinations in accordance with Sheet 2 [494], reproducibility is improved by buffering the reaction solution [495], enabling a detection limit of <0.1 mg/m^3 to be attained.

Continuous Methods. Two types of measuring equipment that fulfil the minimum legal requirements are commercially available. Neither TA Luft, 13th BImSchV, nor 17th BImSchV prescribes the continuous measurement of ammonia in the exhaust gas. Official approval of the equipment is therefore unnecessary [435], [496]. The equipment Opsis AR 602 Z for NH_3 operates in situ as an NDUV single-beam photometer with radiation source and receiver. The absorption spectrum is determined in certain wavelength ranges of the UV (differential optical absorption spectroscopy). The smallest measurement range checked was 0 – 20 mg/m^3. The detection limit is ca. 0.5 mg/m^3.

The Mipan equipment of Siemens is based on absorption of electromagnetic radiation in the microwave range (microwave process gas analyzer). The method is suitable for gases with a permanent dipole moment. The equipment uses extractive sampling. The smallest measurement range checked was 0 – 15.4 mg/m^3, and the detection limit is ca. 0.2 mg/m^3.

6.9. Measurement of Gaseous Organic Compounds

TA Luft and 17th BImSchV regulations impose emission limits on ca. 150 organic substances. The regulations refer not only to individual compounds, but also groups of substances such as polychlorinated dibenzodioxins (PCDD), polychlorinated dibenzofurans (PCDF), and polychlorinated biphenyls (PCB). Almost all installations, including domestic heating, traffic, commerce and industry, can be sources of organic substances.

6.9.1. *Sampling*

The recommendations in Section Sampling regarding velocity-proportional sampling, heating of the probe, construction materials, etc., also apply here. In addition, chemical changes to organic compounds during sampling due to heat, oxidation, nitration, chlorination, etc. must be prevented. As a rule, the substances must be collected at low temperature. Therefore, the sampled gas should be cooled during sampling by a heat exchanger or by adding purified air or nitrogen to reduce its temperature by dilution. An aqueous condensate is usually formed on cooling. Dust is removed from the sampled gas by a filter. For high-boiling substances such as PCDD/F, PAH, and PCB, the dust contains components that must be analyzed. Here, both gaseous and solid phases must be collected and analyzed to give the total figure. The relative proportion of gaseous and solid materials in the total sample depends on the exhaust gas temperature and sampling conditions. The gas can be collected in sample tubes, or in absorption apparatus with wash bottles or impingers. Solid adsorbents such as active carbon, silica gel, Florisil, resins, or various gas chromatographic media are also used. Depending on the detection limit or emission limit value, the collection apparatus must handle between a few dm^3 and 20 m^3 of sampled exhaust gas, and the sampling time can be up to several hours.

6.9.2. *Measurement of Total-C*

The TA Luft and 17th BImSchV Regulations state that volatile organic substances should be measured and recorded as total carbon.

Discontinuous Measurement. The silica gel method of VDI 3481, Sheet 2 [497] is a manual analytical method. The sampled gas is passed via a sampling probe (sometimes heated) through two adsorption tubes in series packed with silica gel.

The organic substances are adsorbed, though this may not be quantitative in some cases (i.e., CH_4, C_2H_6, C_2H_4, C_2H_2, C_3H_8, C_3H_6, C_3H_4, C_4H_{10}, C_4H_8, C_4H_6, C_5H_{12}, CH_2Cl_2, C_2H_3Cl, C_2H_5OH). Desorption is carried out in a stream of oxygen at elevated temperature followed by combustion to form CO_2, which is determined titrimetrically (acidimetric back-titration of barium hydroxide solution after precipitation of barium carbonate) or coulometrically (neutralization of a barium hydroxide solution). As this method does not adequately measure substances with very low boiling points, discrepancies with other methods can occur, depending on the composition of the exhaust gas. Additional gas chromatographic measurements are advisable. With water vapor concentrations of >20 g/m^3 (dew point ca. 20 °C) the adsorption may no longer proceed quantitatively, and the temperature of the sampled gas should not exceed 25 °C prior to the first sorption stage.

The relative detection limit for the titrimetric determination is 17 mg C/m^3, and for coulometric determination 2 mg C/m^3. When using the silica gel method, even at high temperatures and moisture contents, a condensate collecting vessel or a cooler with a condensate collecting vessel is fitted in front of the adsorption tubes [498]. The condensate must then also be analyzed, as it may contain some of the organic compounds. The concentration of CO_2 in the exhaust gas should be low, as it can dissolve in the condensate and lead to false results.

Continuous Methods. Continuous measurement of total-C concentration is usually carried out with flame ionization detectors (FID). These are described in VDI 3481, Sheets 1 and 2 [499], [500]. The limits of application of the FID process and a comparison with other methods for determining the carbon contents of exhaust gases are discussed in VDI 3481, Sheet 6 [501].

In the flame ionization detector, organically bonded carbon is ionized in a hydrogen flame. The ionic current produced in an electric field is amplified and displayed. It is proportional, over a wide range, to the number of carbon atoms introduced into the flame per unit time. The structure of the molecule (e.g., single and double bonds, number and nature of substituted atoms, chain length, or ring structure) considerably influences the oxidation behavior of the carbon and hence the magnitude of the detector signal. Thus, organic compounds with heteroatoms (e.g., N, O, S, or Cl) are generally detected with considerably lower sensitivity than hydrocarbons with the same number of carbon atoms per molecule.

The varying sensitivities to the various organic compounds are expressed as response factors. These are equipment-specific, depending, for example, on jet shape, electrodes, combustion space, and combustion air mixing, and must not be assumed to apply to other FID equipment.

A schematic diagram of an FID is given in Figure 31. The flame requires fuel gas (hydrogen or hydrogen-containing mixtures) and combustion air free of organic substances. The fuel gas and sample gas are mixed before they enter the flame. The combustion air is fed to the flame separately. The flow of combustion air is adjusted so that the flame burns steadily, while at the same time the dew point is kept as low as possible by dilution of the fuel gas, preventing undesired condensation in the exhaust gas duct of the detector. As the pressure and the volume flow rate at the FID affect the measured value, these must not be altered after calibration. Depending on the sensitivity (measurement range) and equipment type, quality requirements for the fuel gas, combustion air, test gas, and composition of the sampled gas (hydrogen/oxygen cross-sensitivity) vary.

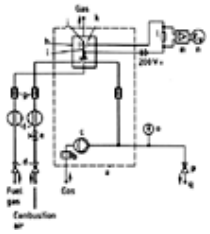

Figure 31. Flame ionization detector (schematic)

a) Heated housing; b) Fine dust filter; c) Pump for sampled gas; d) Pressure regulator; e) Flow control valve; f) Flowmeter; g) Capillary; h) Combustion chamber; i) Jet; j) Flame; k) Electrode; l) Resistor; m) Amplifier; n) Display of measured value; o) Manometer; p) Nonreturn valve; q) Bypass

An overview of equipment is given in Table 17. Monitoring limit values for total-C specified in 17th BImSchV (10/20 mg C/m^3) requires a measurement range of 0 – 30 mg/m^3.

Table 17. FID equipment

Manufacturer	Equipment name	Smallest measurement range tested, mg/m^3	Year of test report
Bayer	Compur FID	0 – 50	1989
Diagnostic	Compur Multi FID 100 E 17	0 – 25	1992
	Compur Multi FID 100 FE 17	0 – 25	1992
Bernath Atomic	BA 3001/BA 3004	0 – 160	1987/1988
	BA 3002 RC	0 – 160	1990
	BA 3002 RC	0 – 25	1991
J.U.M. Engineering	FID VE 7	0 – 25	1990
Mannesmann/H.u.B.	FIDAS 2 T	0 – 645	1977/1982
	FIDAS 2 T	0 – 50	1987/1988
	FIDAS 3 E	0 – 50	1989
	FIDAS 3 E	0 – 25	1992
Siemens	FIDAMAT I/K	0 – 220	1977/1979
	FIDAMAT Ausf. M 52 044-A 10	0 – 50	1990
Testa	TESTA FID 123	0 – 50	1992
Ados*	KM 2-CnHmEM-ADOS	0 – 226	1977

* Heat evolution.

One type of equipment operates on the principle of heat evolution. Catalytic oxidation of combustible components of the gas produces heat which is detected by temperature measurement. There is cross-sensitivity towards CO. The catalysts are poisoned by halogen, lead, and silicon compounds.

6.9.3. *Chlorinated Hydrocarbons*

The various chlorinated organic compounds that can occur in emissions can exist as gases, vapors, solids, materials adsorbed on solids, or as mixtures thereof. Therefore, different sampling and analysis methods are required, depending on the boiling points of the individual compounds and the exhaust gas temperature. Boiling ranges of halogenated compounds of Classes I – III (taken from Subparagraph 3.1.7 of TA Luft) are illustrated graphically in Figure 32 [502]. The boiling ranges of PCDD, PCDF, and PCB compounds are shown for comparison.

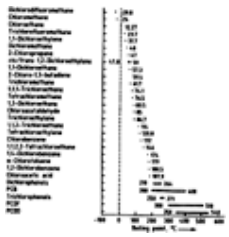

Figure 32. Boiling point ranges of organic compounds (taken from TA Luft)

The high-boiling PCDDs and PCDFs have been thoroughly investigated owing to their environmental implications. Emission measurement methods are described in VDI 3499. A CEN Standard is in preparation, and investigations into PCBs are also in progress. The chlorobenzenes and chlorophenols in the middle boiling range have only been investigated individually from the point of view of emissions. This group of substances can be determined by the sampling and analysis methods described for high-boiling substances. However, there have been no systematic investigations of any significance.

6.9.3.1. *Polychlorinated Dibenzodioxins and Polychlorinated Dibenzofurans (PCDD/F)*

PCDDs and PCDFs are emitted at exhaust gas temperatures between 50 and 400 °C, depending on the design and exhaust gas purification technology of the plant. In this temperature range, the compounds exist in gaseous, solid, and adsorbed form, depending on boiling point. Because volume flow rates of exhaust gases from industrial plants are very high, analysis usually has to be based on extractive sampling, and, because of the presence of the particle phase, the sampling must be isokinetic, using a measurement grid for reasons of representativity. For limit values of ≤0.1 ng/m^3 as a toxicity equivalent, sampling volume rates of 10 – 20 m^3/h are necessary. To ensure mild collection of organic compounds, sampling must take place at a low temperature (if possible <50 °C). At higher temperatures, the sample gas stream must be cooled by dilution with air or by cooling.

The dilution method is described in VDI 3499, Sheet 1 [503]. Cooling is carried out by adding dried air to the gas stream. This prevents the temperature from falling below the dew point of water and causing condensation. The design of the sampling equipment is illustrated in Figure 33.

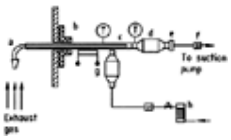

Figure 33. Schematic diagram of a sampling system using the dilution method

a) Probe; b) Heated suction tube (thermostated); c) Mixing tube; d) Sample filter; e) Solid adsorbent; f) Gas meter; g) Cooling air filter; h) Drying tower

The gas is sampled isokinetically by a probe heated to the temperature of the exhaust gas. The sampled gas stream is cooled to below 50 °C by adding dried, filtered air. The diluted exhaust gas is then passed through a filter which traps the solids and the condensed fraction of organic compounds, and the fraction adsorbed on the particles. The PCDDs and PCDFs are collected on the filter at ca. 40 °C almost quantitatively. A solid sorbent such as Florisil, XAD adsorber resin, Porapak, or PU foam, can be included to ensure complete filtration. Alternatively, the above system can be replaced by a flat filter with PU foam as the collecting phase, as is commonly used for the measurement of PCDD and PCDF as atmospheric pollutants.

The sampling system described in Sheets 2 – 4 involves condensation methods with direct cooling [504-506].

Sheet 2. Condensation – adsorption method with filter
Sheet 3. Condensation method, cooled sampling probe
Sheet 4. Condensation method with polyurethane foam for adsorption

In the condensation method of Sheet 3 the hot exhaust gas is cooled as it is withdrawn through the water-cooled probe, and the condensate formed is collected in a cooled separator. Filtration through a solid medium is thus avoided. The sampled gas is then passed through two parallel impinger lines filled with organic solvent, or through a solid adsorbent. The sampling apparatus is shown schematically in Figure 34.

In the method of Sheet 4, PU foam is used as an additional adsorbent. For inclusion of this sampling method in the CEN Guideline, a solid filter is planned prior to the last absorption or adsorption stage.

The first analytic step is extraction of the organic compounds from the collecting phase. For example, in the dilution method, a sampling filter, and, if necessary, a solid absorbent (XAD-2, Florisil, Porapak PS, and polyurethane foam) are used as collecting devices. For the solvent extraction stage, the whole glass fiber filter including loose dust particles is transferred to an extraction thimble. For isomer-specific measurement of the PCDD/F, ca. 5 ng of the following ^{13}C-

labelled standards (one per isomer group) added before extraction

^{13}C-2,3,7,8-tetraCDD/F, ^{13}C-1,2,3,7,8-pentaCDD/F, ^{13}C-1,2,3,7,8,9-hexaCDD, ^{13}C-1,2,3,4,7,8-hexaCDF, ^{13}C-1,2,3,4,6,7,8-heptaCDD/F, ^{13}C-1,2,3,4,6,7,8,9-octaCDD/F.

Extraction with toluene is then carried out in a Soxhlet apparatus for at least 20 h. The toluene extract is reduced to ca. 10 mL in a rotary evaporator at ca. 70 mbar and a bath temperature of ca. 45 °C. The adsorbents XAD-2 and Porapak PS are also extracted with toluene in a Soxhlet apparatus, for at least 20 h. Florisil is eluted with dichloromethane in a glass column. Before further treatment, the extract or eluate is reduced to ca. 5 mL under a controlled vacuum (70 mbar, 45 °C). The extracts can be combined for determination of total PCDD/F, or treated separately. Sample treatment is illustrated in Figure 35.

The ^{13}C-labelled standards are added to the fraction which experience shows is likely to have the higher PCDD/F content. This is either the filter (VDI 3499, Sheet 1) or the condensate (Sheet 2). The extracts are transferred to a column packed with 50 g aluminum oxide and 10 g sodium sulfate, which is eluted with benzene and then with hexane/dichloromethane (98 : 2). This eluate is collected and checked to ensure that at least 60 % of the ^{13}C-2,3,7,8-TCDD is present. The PCDD/F is then eluted with 200 mL hexane/dichloromethane (1 : 1). The eluate is reduced to ca. 2 mL under vacuum, transferred to a mixed column, and eluted with 250 mL hexane. This eluate is reduced to ca. 5 mL, and then chromatographed on a column of Bio-Beads (PCDD/F fraction 120 – 180 mL). The solvent is removed from this fraction under vacuum. The residue is taken up in a solution of 5 ng (or less) $^{13}C_6$-1,2,3,4-TCDD in 100 µL benzene, reduced if necessary in a stream of nitrogen to 30 – 50 µL, and analyzed by GC/MS. The quantitative evaluation is based on the added ^{13}C-labelled standard, identical response factors for all PCDD/F isomers being assumed for all degrees of chlorination. The analysis is carried out “isomer-specifically” for the 2,3,7,8-substituted PCDD/F. The capillary columns are charged with Silar C or an equivalent phase such as Sil 88, SP 2330, or SP 2331. Coupling with the mass spectrometer can be carried out either direct or open. Analysis is by mass fragmentography, with at least two masses being recorded both for the tetra- to octa-CDD/CDF and for the ^{13}C-labelled standard. The absolute detection limit depends on the mass spectrometer used and the amount of material injected, and is normally ca. 0.001 – 0.1 ng per individual isomer and sample. For a gas sample volume of ca. 20 m^3, the detection limit is 0.00005 – 0.005 ng/m^3 for a single isomer.

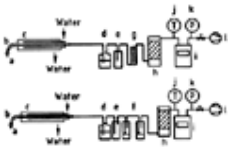

Figure 34. Schematic of a sampling system by the condensate method with cooled suction pipe

a) Probe; b) Bend; c) Cooled suction pipe; d) Condensation vessel; e) Droplet separator; f) Impinger with methoxyethanol; g) Solid adsorbent; h) Drying tower; i) Gas meter; j) Temperature at gas meter; k) Pressure at gas meter; l) Pump

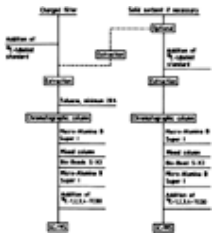

Figure 35. Scheme of sample preparation for PCDD/F

6.9.3.2. *Polychlorinated Biphenyls*

Sampling of PCB can be performed by the same technique as used for PCDD/F. However, there is at present no mandatory guideline with proven parameters. The stages of the analysis are the same as those used in dioxin analysis, the choice of separation and extraction materials being based on the specific properties of PCB (polarity, solution properties). Before extraction of the solid filter (flat filter/PU foam, see VDI 3499, Sheet 1) or of the condensate (VDI 3499, Sheet 3) with toluene, internal standards (^{13}C-labelled for subsequent GC – MS analysis or PCB 28 and PCB 202 for subsequent GC – ECD analysis) are added.

After addition of 6 mL decane, the extract is reduced with the aid of a rotary evaporator and charged to the mixed column (see VDI 3499 and VDI 3498). It is eluted with 250 mL hexane, and the eluate is then reduced to 6 mL and charged to a alumina column. This is then eluted with 60 mL *n*-hexane, 90 mL toluene, and 250 mL *n*-hexane/dichloromethane (1 : 1) [507], [508]. The toluene contains the PCB compounds and the hexane – dichloromethane mixture the PCDD/F compounds. The fractions can be purified if necessary by a further cleanup with, e.g., Florisil or HPLC. A labelled standard is then added, e.g., for PCB, $^{13}C_{12}$-2,2′,4,5,5′-pentachlorobiphenyl, and, for PCDD/F, $^{13}C_6$-1,2,3,4-tetrachlorodibenzodioxin. It is then reduced to a final volume of 10 – 20 µL, and analyzed by GC – MS. The relative detection limit for the total content of PCB is 10 – 20 ng/m^3. For the individual components it is ca. 1 – 2 ng/m^3.

6.9.3.3. *Chlorobenzenes and Chlorophenols*

Sampling of chlorinated benzenes and phenols can in principle be carried out by the same methods as those for PCDD/F and PCB. There is no guideline with approved method parameters. Also, individual tests should be carried out to determine to what extent the more volatile mono- and dichlorocompounds are quantitatively recovered. Suitable collecting phases include solid filters, PU foam, methoxyethanol, and adsorber resins. Another possibility is absorption in cold solvents (e.g., methyldiglycol), as described in the draft version of VDI 2457, Sheet 2 [509].

Chlorobenzene. Toluene is used as an extraction medium for separating chlorobenzenes and chlorophenols from the collecting phase. The toluene extract is reduced to a few milliliter, and chromatographed on alumina. Elution is then carried out with hexane/dichloromethane (98 : 2), and this fraction is again reduced in volume. Analysis for chlorobenzenes is carried out, for example, by GC – ECD or GC – MS. The quantification of the chlorobenzenes is carried out externally using a mixture of di- to hexachlorobenzenes. Analysis can also be carried out with ^{13}C-labelled standards, using one standard for each degree of chlorination. The relative detection limit is ca. 1 ng/m^3 for each component.

Chlorophenols. The chlorophenols are determined by extracting the toluene extract with K_2CO_3 solution. Acetic anhydride is added, the samples are agitated, and the chlorophenol acetates are then extracted with hexane. The hexane extract is dried over Na_2SO_4, and analyzed by GC/ECD or GC/MS. Quantification of the chlorophenols is carried out externally using a mixture of di- to pentachlorophenols derivatized under the same conditions. Alternatively, ^{13}C-labelled standards can be used.

6.9.3.4. *Volatile Chlorinated Hydrocarbons*

Volatile chlorinated hydrocarbons (VCH) such as trichloroethane, trichloroethylene, tetrachloroethane, and tetrachloroethylene are used in industrial surface treatment, chemical cleaning, and solvent extraction. Limit values are imposed by 2nd BImSchV on emissions of the above VCHs and dichloromethane [425]. The use of trichloroethane is now prohibited in some cases.

Measurement of emissions is described in the VDI 2457, Sheets 2 – 4 of 1974 – 1976 for 1,1,1-trichloroethane, trichloroethene, and tetrachloroethene. They are absorbed in a liquid sorbent and determined by gas chromatography. However, in some cases the detection limit is relatively high.

Sampling and analysis are therefore now carried out by a method based on that used for ambient air pollution measurement [510], [511]. Small test tubes packed with activated carbon are used to collect the sample, and desorption is then carried out with solvents or by heating. The sampling itself is carried out with quartz probes provided with dust filters, if necessary heated to the temperature of the exhaust gas. With higher water vapor contents, a condensate trap follows the dust filter. The condensate is also analyzed for VCHs by gas chromatography. Another solid sorbent for pollutants is 2,6-divinylbenzene-*p*-phenylene oxide polymer (Tenax TA) [512]. The VCHs can be desorbed for gas chromatographic analysis and recondensed in a cooled container. Detection is by ECD or mass spectrometry. Another method is absorption in a cold solvent [509].

6.9.3.5. *Chlorofluorocarbons*

The regulation 2nd BImSchV referred to in Section Volatile Chlorinated Hydrocarbons above also imposes limits on the emission of chlorofluorocarbons (CFCs) the only permitted CFCs being 1,1,2,2-tetrachloro-1,2-difluoroethane (R-112), 1,1,2-trichloro-1,2,2-trifluoroethane (R-113), and trichlorofluoromethane (R-11). Sampling is performed by the method described in Section Volatile Chlorinated Hydrocarbons, the substance being adsorbed on, e.g., activated carbon. With low-point components, gas collecting tubes can be used. Analysis is carried out by GC/ECD.

6.9.4. ***Hydrocarbons***

TA Luft imposes emission limits on a number of hydrocarbons, depending on the substance group and mass flow rate. Total emission concentrations of 20 – 150 mg/m^3 are permitted, depending on the substance classification. Substances such as benzene, acrylonitrile, or polycyclic aromatic hydrocarbons, are controlled separately in Subparagraph 2.3 of TA Luft.

Polycyclic Aromatic Hydrocarbons. Polycyclic aromatic hydrocarbons (PAH) are present in many natural materials, for example, mineral oil. PAH structures exist in coal, but the compounds themselves are only formed after chemical transformation. In industrial processes, unreacted PAH residues and newly formed PAH are emitted during the incomplete combustion or pyrolysis of organic materials, especially fossil fuels. Some polycyclic aromatic hydrocarbons, especially tetracyclic to hexacyclic compounds, have a mutagenic and/or carcinogenic effect on human and animal organisms. A limit of 0.1 mg/m^3 is imposed on benz[*a*]pyrene and dibenz[*a, h*]anthracene by TA Luft. VDI 3873, Sheet 1 describes a method of measurement for use in industrial and commercial plants [513]. The sampling method used is the dilution process described in VDI 3499 (see Section Polychlorinated Dibenzodioxins and Polychlorinated Dibenzofurans (PCDD/F)). The four- to seven-ring PAH compounds have approximately the same boiling range as PCDD/F. They therefore occur in exhaust gases as gases, solids, and materials adsorbed on particles. The method of VDI 3873, Sheet 1 can deal with all states of aggregation. The four- to seven-ring compounds are quantitatively collected on a filter. The more volatile PAH compounds with two or three rings can be recovered by a downstream solid sorbent (e.g., Porapak PS). It is possible to collect PAH, PCDD/F, and PCB at the same time. The sampling times (collecting times) should not exceed ca. 2 h to ensure that the PAH profile remains the same, because some PAH compounds are very reactive. A detection limit of ca. 10 ng per component can be achieved. With a sample volume of 5 m^3 a relative detection limit of 2 ng/m^3 per component can be attained. The exposed filter is extracted with toluene for 30 min after adding the internal standard. The sample preparation scheme is illustrated in Figure 36. A detailed description can be found in VDI 3873, Sheet 1. Analysis is then carried out by capillary GC.

Figure 36. Scheme of sample preparation for PAH

Aromatic and Volatile Hydrocarbons. The sampling technique for aromatic hydrocarbons (benzene, toluene, xylene, and ethylbenzene: BTXE) is similar to that used for pollution measurement described in Section Volatile Chlorinated Hydrocarbons. The most important collecting phases used are activated carbon, Tenax, and cold solvents [510], [511]. Materials are desorbed from activated carbon mainly with solvents [514] or by thermal methods. Identification is by GC – FID or sometimes GC – MS. If condensate is formed during sampling, this must also be analyzed. With a sampling volume of 20 L, the detection limit is 5 µg/m^3 per component. The Guideline VDI 2457, Sheet 1, which describes the fundamental principles of sampling and determination of organic compounds by gas chromatography, should be consulted [515]. The method of absorption in nitrobenzene described in VDI 2457, Sheet 5 is less suitable because of the high detection limits and various cross-sensitivity effects [516].

Volatile organic compounds (VOC) such as aliphatic, cyclophatic, and aromatic hydrocarbons and solvents can be determined by the methods given for BTXE or VCH compounds. Activated carbon or other solid sorbents are mainly used.

1,3-Butadiene. Emissions of 1,3-butadiene, like those of acrylonitrile, must not exceed 5 mg/m^3 according to Subparagraph 2.3 of TA Luft. The Guideline VDI 3953 describes a measurement method by adsorption on activated carbon and desorption with CS_2 [524]. Gas chromatographic analysis is based on the GC –FID head space technique (vapor space analysis). The determination limit is 0.2 mg/m^3 with a sampled gas volume of 30 L. The limited storage time of the samples must be borne in mind.

6.9.5. *Other Compounds*

Phenol and Cresols. Phenol and the cresols can be collected on silica gel using the sampling arrangement described in Section Volatile Chlorinated Hydrocarbons for active carbon. Identification can be made by HPLC with a fluorescence detector. With a sampled gas volume of ca. 30 L, a detection limit of ca. 1 µg/m^3 is achieved.

Amines. Aliphatic amines such as methylamine, dimethylamine, ethylamine, propylamine, butylamine, and hexylamine can be quantitatively determined, especially in exhaust air, by the method described in VDI 2467 for ambient air pollution measurement [517]. The amines are dissolved as alkylammonium chlorides in dilute hydrochloric acid, and are then reacted to form dinitrophenylamines. Determination is by HPLC, with UV absorption analysis. Aromatic amines can be sampled with silica gel tubes and determined by HPLC – UV. Detection limits, e.g., in the case of 2-naphthylamine and toluidine, are ca. 5 µg/m^3.

Aldehydes. Short-chain aldehydes are formed by the incomplete combustion of hydrocarbons or alcohols, e.g., in internal combustion engines including diesel engines.

A method of measuring these emissions is described in the VDI 3862, Sheet 1 [518]. Aliphatic C_1 – C_3 aldehydes are absorbed in solutions by the MBTH (3-methyl-2-benzothiazolinone hydrazone) method, and determined photometrically. The determination limit is 0.05 mg/m^3 if the volume of exhaust gas sampled is 25 L. If SO_2 is present at concentrations of >30 mg/m^3, the method is unsuitable because of interference effects.

In Sheet 2 (in preparation) of VDI 3862 the determination of long-chain and aromatic aldehydes and ketones is described [519]. These substances are reacted with 2,4-dinitrophenylhydrazine (DNPH method) to give the hydrazone and determined as individual components. Sampling is carried out with wash bottles containing acetonitrile to dissolve the hydrazine. The absorption solution is directly chromatographed without further treatment. Determination is by HPLC with UV absorption analysis.

Sheet 3 of the VDI 3862 (in preparation) describes sampling with the aid of DNPH dissolved in hydrochloric acid solution [520]. The hydrazones are extracted with carbon tetrachloride, the organic phase is chromatographed using HPLC, and analysis is performed by UV absorption. Depending on the substance, detection limits between 2 and 100 µg/m^3 have been quoted for a sampled volume of 50 L. A disadvantage is that the samples can be stored for a short time only.

Acrylonitrile. Three methods for the determination of acrylonitrile are described in VDI 3863, Sheets 1 – 3 [521-523].

Sheet 1: Sampling by a gas sample container. The detection limit with a flame ionization detector (FID) is 0.5 mg/m^3. With a phosphorus – nitrogen detector (PND), the limit is 0.05 mg/m^3. The samples can be stored only briefly. With high water vapor contents, the methods of Sheets 2 and 3 are used.

Sheet 2: Sampling in cold solvents, detection limit with 15 L sampled volume using FID 0.2 mg/m^3, or using PND 0.05 mg/m^3. Samples can be stored only briefly.

Sheet 3: Sampling on activated carbon, desorption with dimethylformamide (DMF), detection limit 12 µg/m^3 with 50 L sampled volume. GC with column switching and two FIDs.

All sampling equipment is fitted with a sampling probe and solid filter (heated if necessary).

Vinyl Chloride. The determination of vinyl chloride by gas chromatography of samples collected in gas sample containers is described in the Guideline VDI 3493, Sheet 1, of Nov. 1982 [525]. Analysis is by GC – FID, the determination limit being 0.3 mg/m^3. Subparagraph 2.3 of TA Luft imposes an emission limit value of 5 mg/m^3. By adapting the method of VDI 3494, Sheets 1 and 2, for determining vinyl chloride ambient air pollution by adsorption on activated carbon, it should be possible to achieve improved sensitivity [526], [527].

6.10. Measurement of Odor Emissions

Odor emissions can be produced in many industries, e.g., cocoa, coffee, and tobacco processing, fish production and processing, rubber production and processing, and the chemical industry. The aims of the measurement technique are to determine the odor intensity (based on determination of the odor threshold), to estimate nuisance potential, to estimate the effectiveness of techniques of odor emission reduction (crude/purified gas), and to establish data to enable spreading rates to be calculated and odor pollution to be forecast. Odors are determined using the human nose as a sensor (olfactometry). Olfactometry consists of the controlled presentation of odoriphores (odoriferous substances) and the assessment of the sensations which these produce in human subjects. An odor sensation can be caused even when the concentration of an odoriphore is less than the limit of detection by chemical or physical methods. The qualitative and quantitative assessment of mixtures with a large number of components is difficult. Even if analysis is possible, there may be no connection between this and the odor sensation.

The concentration of the odoriphore in the gas sample (single substance or mixture of substances) under investigation is determined by diluting with neutral air until the odor threshold is reached. One odor unit (OU) is the quantity of odoriphore which when uniformly distributed in 1 m^3 of neutral air produces an odor sensation in accordance with the definition of the odor threshold. The unit of odoriphore concentration is one odor unit per unit volume (OU/m^3). The odor threshold is the concentration of a substance or mixture of substances that gives an odor sensation in 50 % of a statistical sample of test subjects. The fundamental principles of odor threshold determination are described in the VDI 3881, Sheet 1 [528]. Olfactometers are used for measurement. A gas sample containing an odoriphore is diluted with various amounts of neutral air in a controlled manner and submitted to test subjects for odor testing.

Sampling can be dynamic or static. In dynamic sampling, a stream of exhaust gas or exhaust air is continuously removed. The olfactometer is attached by a T-piece to the sample line, so that a test volume can be fed into the olfactomer as required. The scheme for dynamic sampling is illustrated in Figure 37.

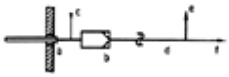

Figure 37. Dynamic sampling

a) Sampling probe; b) Dust filter (if necessary); c) Air addition for dilution (if necessary); d) Pipeline for gas samples; e) Excess to exhaust air; f) To the olfactometer

In static sampling, the sample of odoriferous substance is charged into a container, transported to the olfactomer, and connected to it for the experiment. Possible methods of static sampling are shown schematically in Figure 38 [529]. If there is a possibility that water or other substances in the sample could condense inside the bag, the sample should be diluted. In this case, the sample bag is filled with a definite amount of odor-free air in an apparatus as shown in Figure 38. A definite amount of the test gas is then added. The dynamic sampling method is advantageous for constant emission condition because of the short time delay between sampling and analysis. The method is not suitable if concentrations change rapidly, as the sampled material changes during the course of an individual measurement, and there can then be a wide distribution of measured values.

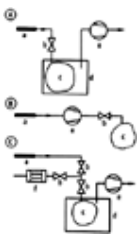

Figure 38. Various methods of static sampling

A) Bag filling by suction pump; B) Bag filling by pressure pump; C) Bag filling by suction pump with additional dilution with filtered air

a) Sampling probe; b) Valve; c) Bag; d) Bag container; e) Pump; f) Activated carbon filter

With static sampling of varying concentrations, a representative mean value over the sampling time is obtained. Evaluation of the results is described in [528], [530]. Various types of olfactometer are described in [531]. The parameters of the method and an example of an odor measurement record are given in [532].

[Top of Page]

7. Separation and Purification

Eberhard Schmidt, Torsten Schmidt, Theo Pilhofer, Helmut Krill, Harald Menig, Stephan Schirz, Helmut Pollack

7.1. Separation of Solid and Liquid Particles from Gases [533-541] (see also ⟶ Dust Separation)

7.1.1. *Introduction*

Many industrial processes inevitably generate particles entrained by a carrier gas. These particles may be solid (dust) or liquid (droplets or mist). As a rule, the particles must be separated from the transporting gas, either because they are the desired product or because they are an undesirable emission. Product recovery by dust separation has been practiced long and intensively in process engineering. The importance of emission control has grown tremendously in the last two decades, because of rapid increases in industrial production, and better knowledge of potential hazards and harmful effects.

Dust collection methods used in practice can be classified as inertial and centrifugal separation (gas cyclones), wet collection (scrubbers), filtration (deep-bed filters, surface filters, and granular-bed separators), and electrostatic collection (precipitators). All of these collectors separate particles from gas by a common principle: forces acting on the particles cause them to enter regions from which the gas cannot transport them away. These regions may be the inner wall of a cyclone, the droplet surface in a scrubber, the fiber or grain surface in a filter, or the collecting electrode in a precipitator.

One difficulty in selecting and designing separation processes is that the size distribution of the particles to be collected generally extends over a wide range, from <0.1 µm to >100 µm. The knowledge about the dependence of the operational characteristics of each separator on particle size distribution is very important and is discussed widely.

7.1.2. *Characterization of Particle Separation*

To evaluate dust collection equipment, both the energy consumption (or pressure drop Δp) and the collection efficiency must be determined. The latter can be found by integration techniques (for the totality of particles) or differential methods (for each particle size fraction).

The overall collection efficiency *E*, also known as the total or cumulative collection efficiency, is given by

$$E = \frac{\dot{M}_c}{\dot{M}_e} = \frac{\dot{V}_e c_e - \dot{V}_f c_f}{\dot{V}_e c_e} = 1 - P \tag{1}$$

where *E* is collection efficiency, $\dot{M}$ is mass flow rate of particles, $\dot{V}$ is volume flow rate of gas, *c* is particle concentration, and *P* is penetration. The subscript c stands for collected (coarse), e for entry, and f for penetrating (fine).

Note that if a collection device is rated in terms of overall collection efficiency, this quantity depends not only on the equipment design and operating conditions, but also on the type of dust and, especially, the particle-size distribution. The stated overall collection efficiency or overall penetration thus always refers to a particular case. A more useful and informative measure is the fractional collection efficiency *T* (*x*), which measures collection as a function of particle size *x*:

$$T(x) = \frac{\mathrm{d}\,\dot{M}_c(x)}{\mathrm{d}\,\dot{M}_e(x)} = 1 - \frac{\dot{V}_f \mathrm{d}\,c_f(x)}{\dot{V}_e \mathrm{d}\,c_e(x)} \tag{2}$$

With $\dot{V}_f = \dot{V}_e$ and particle size distribution *q*:

$$T(x) = 1 - \frac{P_{q_f}(x)}{q_e(x)} = \frac{E_{q_c}(x)}{q_e(x)} \tag{3}$$

The cut size x_t is the particle size at which a given fractional collection efficiency *T* is achieved. A value of 50 % is often used ($x_{50,t}$).

7.1.3. *Gas Cyclones*

7.1.3.1. *Mode of Operation and Basic Designs*

The simplest way to separate particles from gases is to allow them to settle out in so-called gravity settlers into zones of low gas velocity. Because this method is technically feasible only for coarse particles that are significantly larger than 100 µm, gravity settlers now have negligible importance and are not discussed here.

Collection by means of centrifugal forces is much more efficient than simple gravity separation, and this led to the development of centrifugal collectors. Because of their simple design, reliable operation, small space requirement, and low cost, gas cyclones are used widely in many branches of industry. They are frequently employed in material recovery from gas recycle systems, pneumatic conveying, and other areas. They are often used as first-stage collectors. Cyclones can be operated at pressures of 1 kPa to 10 MPa and temperatures exceeding 1200 °C. The advantages of cyclones over other collectors are offset by the fact that they are less efficient collectors for particles below ca. 1 – 5 µm in size.

Figure 39 shows the most common type of cyclone with flow reversal. The gas path is designed to impart a twist to the particle-laden gas entering the rotationally symmetrical apparatus. A turbulent, three-dimensional, rotational flow is produced. In the separating chamber spiral, downward flow takes place at larger radii. Flow is then reversed, and the gas spirals upward at smaller radii, and reaches the exit duct (vortex finder). The particles entrained in the revolving gas stream experience centrifugal forces hundreds to thousands of times greater than the force of gravity. Thus, the larger particles migrate outward to the cyclone wall, where they collect as strands; they are carried down the tapered body and out of the separating chamber by boundary-layer flow. The conical shield, often built into the bottom of the tapered body, prevents already collected material from being reentrained. The particles that are not collected are carried inward by the gas flow and removed through the exit duct.

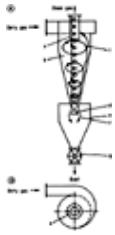

Figure 39. Schematic of a reverse-flow cyclone

A) Vertical section; B) Top view

a) Exit duct; b) Separating chamber; c) Gas flow pattern; d) Conical shield; e) Dust hopper; f) Dust; g) Lock

Cyclones vary in the form of the swirl-producing inlet (tangential, axial with swirl vanes), the shape of the separating chamber (cylindrical, conical), the design of the exit duct, and their dimensions. Diameters range from 0.02 m to 5 m and flow rates from 1 to 100 000 m^3/h.

7.1.3.2. *Collection Efficiency and Pressure Drop*

Important parameters in cyclone design and operation are the fractional collection efficiency $T(x)$ and the energy consumption, measured in terms of the pressure drop Δp.

Several models have been devised for the approximate calculation of collection efficiency at low dust loads. They describe the flow field in more or less simplified terms. The separation-zone model is now employed widely because of its simplicity and ease of use. However, this approach does not give the complete fractional efficiency curve; only a critical particle diameter x_{cr} can be calculated. According to this concept, all particles whose diameter x is less than x_{cr} pass through the device; larger particles are collected.

In theory, the fractional efficiency curve jumps from 0 to 1 at x_{cr}. It is more reasonable to assume that x_{cr} corresponds to the median diameter $x_{50,t}$ of the real separation function $T(x)$, i.e., the 50 % value on the effective S-shaped fractional efficiency curve. Based on experimental results the suggestion has been made that the actual cut diameter can be estimated as $x_{50,t} \approx 1.3 \cdot x_{cr}$.

In a more detailed approach, particle motion is treated as a process in which diffusional particle transport is superimposed on a predetermined motion. Particle flux calculations are based on a modified flow field. The advantage of this model is that the complete fractional efficiency curve can be obtained.

At higher particle concentration (> 10 g/m^3), additional collection mechanisms contribute to particle separation. By analogy with pneumatic conveying a limiting dust load c_m is proposed. The excess particle mass cannot be transported by the gas and is accelerated to the wall immediately after entering the cyclone; the particles move down into the dust trap in the form of a strand. Along with particle collection due to excess loading, agglomeration of small particles with large ones can be assumed to occur at high particle concentration; this also improves overall collection efficiency.

The total pressure drop of a cyclone consists of pressure losses in the inlet, separating chamber, and exit duct; it is usually calculated as

$$\Delta p = (\xi_{\text{inlet}} + \xi_{\text{separator}} + \xi_{\text{exit}}) \cdot \frac{\varrho_f v_i^2}{2}$$

where ξ is the pressure drop coefficient, ϱ_f is the gas density, and v_i is the axial velocity in the exit duct. Experience has shown that the pressure drop in the exit duct accounts for as much as 90 % of the total cyclone pressure drop. Typical values for Δp are in the range of 500 – 2000 Pa.

7.1.3.3. *Operational Characteristics*

The collection performance of a cyclone depends chiefly on its geometry, gas throughput, inlet gas concentration, and properties of the material to be separated. For example, the increase in gas viscosity when the temperature is raised decreases performance. If the exit-duct diameter or inlet cross section is diminished, the tangential velocity component at the exit-duct radius increases, so that a greater centrifugal force acts on the particles. As a result, the cut size $x_{50,t}$ decreases and particle collection improves. Collection efficiency is also improved if the volume flow rate through the cyclone increases. Theoretically the cut size varies according to $\dot{V}^{-0.5}$.

When the tangential velocity in the cyclone increases, however, the pressure drop Δp also increases. To achieve good collection performance and acceptable energy consumption in cleaning gases with high flow rates, several small cyclones connected in parallel are preferable to a large cyclone because the reduction in cyclone diameter shifts the cut size toward smaller particle sizes. Figure 40 presents typical fractional efficiency curves for three cyclones with different diameters but the same pressure drop. If geometrically similar cyclones are assumed and Δp is constant, the result for cleaning a gas with a given flow rate $\dot{V}$ is that the cut size for n cyclones in parallel is proportional to $n^{-0.25}$.

Figure 40. Fractional efficiency curves for cyclones of varying diameter d at a constant pressure drop

$\dot{V}$ = volume gas flow rate

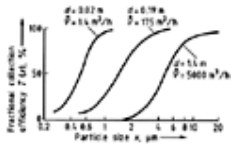

7.1.4. *Wet Scrubbers*

7.1.4.1. *Mode of Operation and Basic Designs*

In wet collectors, both particulate and gaseous components can be separated from the carrier gas.

Even very fine particles (<0.1 µm) can be collected in scrubbers. The particles are brought into contact with a scrubbing liquor and are removed from the gas stream in bound form as a slurry, which must be subjected to wastewater treatment or liquor recycling. In most scrubbers the scrubbing liquor is in droplet form. The first operation can, therefore, be regarded as the agglomeration of small solid particles and larger liquid particles. These droplets are readily collected from the gas stream.

The transport step is decisive in separation. Diffusion and condensation can take place, but in general inertial forces appear to dominate. The wettability of the particles does not play a crucial role in binding; wettable particles penetrate the interior of the droplet, while nonwettable particles adhere to the surface.

Efficient dust collection in a scrubber requires separation of the dust-laden droplets in a separator that is either integrated into the scrubber or connected downstream. Because droplet sizes generally range from 1 µm to several hundred micrometers, commercial separators utilize inertial forces to separate droplets from the gas stream. Layers of wire mesh or fabric, stacks of plates, or centrifugal devices are employed.

Scrubbers are relatively easy to adapt to various operating conditions. They are employed mainly when products are to be recovered from a wet phase or when the danger of dust explosion or other hazards forbids dry collection. They have proved suitable chiefly for moderately large-volume gas flow rates (often <30 000 m^3/h). Many scrubber designs are available. However, most designs fall into five basic types, shown in Figure 41. Typical operating data are listed in Table 18.

Table 18. Operating data for scrubbers (see also Figure 41

	Packed-bed scrubber	Ejector venturi scrubber	Self-induced spray scrubber	Rotational scrubber	Venturi scrubber
Cut size, µm*	0.7 – 1.5	0.8 – 0.9	0.6 – 0.9	0.1 – 0.5	0.05 – 0.2
Relative velocity, m/s	1	10 – 25	8 – 20	25 – 70	40 – 150
Pressure drop, kPa	0.2 – 2.5		1.5 – 2.8	0.4 – 1.0	3 – 20
Water consumption, L/m^3	0.05 – 5	5 – 20*		1 – 3*	0.5 – 5
Energy consumption kW · h/1000 m^3	0.2 – 1.5	1.2 – 3	1 – 2	2 – 6	1.5 – 6

* Particle density ϱ_p = 2.42 g/cm^3.

* For each stage.

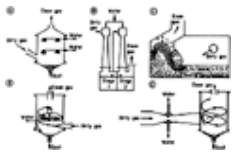

Figure 41. Basic types of scrubbers

A) Spray tower; B) Ejector venturi scrubber; C) Self-induced spray scrubber; D) Rotational scrubber; E) Venturi scrubber

The spray tower (Fig. 41 A) is the oldest type of scrubber. The dirty gas flows upward through a large tube at a relatively low speed of ca. 1 m/s. The washing fluid is dispersed countercurrent to the gas by spray nozzles mounted at different levels. The dust-laden liquid leaves the bottom of the scrubber. Packings (e.g. glass beads) are sometimes used to improve collection efficiency. The droplet collector can be integrated in the upper region.

The ejector venturi scrubber (Fig. 41 B) is basically a water-jet pump. Water is sprayed into the scrubber at a speed of ca. 25 – 35 m/s and provides a draft for moving the gas, which often attains a flow rate of 10 – 20 m/s. Two successive stages are often used because the collection efficiency of ejector venturi scrubbers is not much higher than that of spray towers.

In self-induced spray scrubbers (Fig. 41 C) the dust-laden gas is fed through the wash fluid. The wash fluid is thus atomized, entrained, and mixed with the gas. This effect is reinforced by a redirecting vane. Entrained, dust-laden fluid

collects inside the scrubber in the collection tank.

In rotational scrubbers (Fig. 41 D), the washing fluid is dispersed radially in the gas by rotating spray nozzles. The gas spirals upward and the dirty water is removed from the bottom of the scrubber. The droplet collector is located downstream.

Venturi scrubbers (Fig. 41 E) have the highest collection efficiency. The gas is accelerated up to 150 m/s in the venturi throat. The washing fluid enters the middle of the throat just in front of its narrowest point or is fed radially into the throat. The liquid is dispersed by the inflowing gas. The dust-laden water droplets are usually separated from the gas in a cyclone.

7.1.4.2. *Collection Efficiency and Pressure Drop*

Development and sizing of scrubbers are often purely empirical because the processes that occur in these complicated pieces of equipment are complex and difficult to simulate with theoretical models. However, the proposed models are quite helpful because they allow clear identification of important parameters and trends.

In most models, the scrubbing liquor is assumed to be in droplet form, and inertial forces are considered crucial for the transport of dust particles to the droplet. The calculation consists of three steps: calculation of single-droplet collection, calculation of the gas volume cleaned by the single droplet, and calculation of the change in dust concentration due to the action of all droplets. Important parameters are the Reynolds number and the inertial parameter (Stokes number). The effort required for the numerical computations is shortened substantially by fitting the equations to experimental values, so that the equation of motion for the droplets need not be solved. For the venturi scrubber, such an equation for the fractional collection efficiency $T(x)$ is given by

$$T(x) = 1 - \exp\left[-\frac{2\,\dot{V}_l d_l \varrho_l v}{55\,\dot{V}_g \mu} \cdot F(\psi, f)\right] \tag{4}$$

where $\dot{V}_l$ is the volume flow rate of scrubbing liquor, $\dot{V}_g$ is the volume flow rate of gas, d_l is droplet diameter, ϱ_l is droplet density, v is maximum gas velocity, μ is gas viscosity, and F a function of the inertial parameter Ψ and an empirical fitting parameter f ($0.25 < f < 0.5$). The following approximation is given for the function F in the range $1 \leq \Psi \leq 4$.

$$F(\psi, f) \approx 0.312 \cdot \psi f^2 \tag{5}$$

Equation (4) can also be used, along with a pressure-drop equation, for optimization. In general, model calculations indicate that higher relative velocities and increased addition of scrubbing liquor lead to better collection efficiency; this is also observed experimentally.

Calculation of the pressure drop is of interest chiefly for venturi scrubbers. Empirical formulas and model calculations are available for this purpose. The empirical Equation (7) is based on the procedure for evaluating flow in pipes.

$$\Delta p = \left(C_1 + C_2 \frac{\dot{V}_l}{\dot{V}_g} 10^3\right) \frac{\varrho_f v 2}{2} \tag{7}$$

C_1 and C_2 are empirical constants, with typical values $0 \leq C_1 \leq 0.5$ and $0 \leq C_2 \leq 1.7$; ϱ_f is gas density.

7.1.5. ***Filters***

7.1.5.1. *General*

Filters constitute an important aid in modern particle collection technology, especially when highly efficient collection in the finest particle-size range is required. Many versions are available. Of all collectors, filters have the widest spectrum of applications and thus a large share of the market.

With regard to application, design, and mode of operation, filters for dust collection can be classified in different but overlapping groups. Three main classes are the so-called deep-bed filters, surface filters, and granular-bed filters.

7.1.5.2. *Deep-Bed Filters*

Deep-bed filters (fibrous filters) are used where dust loads are low, on the order of several milligrams per cubic meter of gas. Typical applications include air conditioning and ventilation, laboratory exhaust systems, clean rooms, and respiratory protection. The filters consist generally of a relatively loose fiber mat with a pore fraction over 90 % (often > 99 %). As the gas flows through the fibrous layer, particle collection takes place in the interior of the layer, where the dust accumulates (deep-bed filtration). After becoming saturated with dust, these filters are usually discarded; some can be cleaned by washing or blowing. Typical gas velocities are 0.1 – 3 m/s.

7.1.5.3. *Surface Filters*

Mode of Operation and Basic Designs. Surface filters (fabric filters) are used for high dust loads, on the order of several grams to hundreds of grams per cubic meter. This is the typical range for industrial dust removal and air-pollution control. The principal surface-filter media are nonwovens and (needled) felts composed of randomly arranged glass, plastic, or ceramic fibers. Woven materials, chiefly constructed from glass-fiber yarns or metal wires, are also employed. Only at the beginning of filtration do these media trap particles in the interior of the fibrous layer. As more dust is deposited, it bridges the fibers and forms a dust layer (filter cake) on the surface.

The fractional efficiency curves in Figure 42 show the improvement in collection performance as more dust is collected and pressure drop increases. The new, clean filter gives a typical depth-filter curve. Since the cake-formation phase lasts only

a short time for the dust loads usually encountered in practice, the interception effect inside the cake is dominant. This process is called surface filtration. Emission levels of <1 mg/m^3 can be achieved.

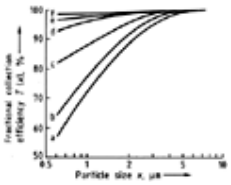

Figure 42. Variation in fractional collection efficiency due to dust deposition on a polyester needled felt

Amount of dust deposited, g/m^2: a) 4, Δp = 33 Pa; b) 14; c) 29; d) 47; e) 75; f) 95, Δp = 110 Pa

As the amount of dust deposited in and on the filter grows, the pressure drop also increases. After a certain time, or when a predetermined pressure-drop value is reached (ca. 1000 – 2500 Pa), the dust must be removed (cleaned off) to regenerate the filter.

In modern commercial filter media, surface coating or treatment is used with the aim of collection on the surface right from the start and to prevent particles from penetrating into the medium. This approach yields high initial collection efficiencies, avoids the danger of plugging, and facilitates removal of the filter cake.

The three most important surface-filter designs are round bag, envelope, and cartridge. A filter unit contains the requisite number of filter elements connected in parallel. Designs differ primarily in the method of cleaning and sometimes in the way the dirty gas is introduced.

Round bag filters are used most widely. Typical bags are 100 – 300 mm in diameter and 1.5 –10 m long. The classical design is a multichamber filter with shaker cleaning. Dirty gas is fed into the bags from below; it then flows through the filter medium and is led to a clean-gas duct at the top of the device. Particle collection takes place on the inside of the bags. For cleaning, the dirty-gas supply is cut off from one chamber at a time, and the dust is loosened by shaking or rapping the bag support. The dust removed from the bags falls into a dust hopper; purge air drawn into the bags from outside aids this process.

Pulse-jet filters (Fig. 43) became more widely used as progress was made in needle felts. The gas always flows from outside to inside, so the bags must be pulled over supporting frames. For cleaning, a jet of compressed air is admitted to the bag through its open top end. The compressed-air jet suddenly raises the pressure in the bag, which is rapidly inflated and gas flow is reversed. The dust is thus dislodged from the outside of the bag. Because single bags or single rows of bags must be removed from the filtration process for only a short time (ca. 0.1 – 0.3 s), the filter need not be divided into chambers, and cleaning can be done on-line.

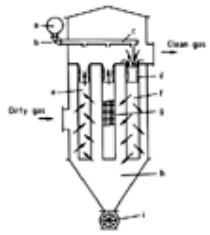

Figure 43. Round bag filter with reverse-pulse cleaning

a) Compressed-air tank; b) Right-angle diaphragm valve; c) Compressed-air duct with nozzles; d) Venturi nozzle; e) Filter bag in operation; f) Filter bag being cleaned; g) Retainer; h) Dust hopper; i) Air lock; j) Deflector shield

Design Calculations. The main purpose of surface-filter design is to determine the required filtration area A, which is given by the ratio of the volume flow rate of dirty gas $\dot{V}$ and the filtration velocity or filter load v. The filtration velocity is selected on the basis of empirical values because of a lack of generally valid models.

The pressure drop Δp is assumed to be the sum of the contributions of the medium and the filter cake. This leads to Equation (8), where K_1 is the residual resistance of the filter medium after cleaning, K_2 is the specific resistance of the filter cake, W is the mass of dust collected per unit filter area, and μ is the gas viscosity. Since W is proportional to filtration time, the second term describes the time dependence of Δp.

$$\Delta p = K_1 \mu v + K_2 \mu W v \tag{8}$$

Important parameters in this equation are the resistances K_1 and K_2; as a rule, they are not constants but functions of several variables. Prediction of these parameters is still problematic because the functions are not sufficiently well known. Even when K_1 and K_2 can be determined experimentally, great care must be taken to ensure that conditions in the experimental device and the full-scale plant are comparable.

Operational Characteristics. Particle collection in a surface filter occurs mainly in the dust layer formed on the surface. Therefore, this highly efficient cake must be retained on the surface as long as possible.

Figure 44 shows that the pressure drop Δp increases as more dust is deposited, and periodic cleaning is necessary. After cleaning, the pressure drop falls to a value that depends on the flow resistance of the filter medium (including residual dust); this value is higher than the initial pressure drop of the new, clean filter. For stable filter operation the residual pressure drop must reach an approximately constant value after an induction period. The filter medium, filtration velocity, and the method and intensity of cleaning determine whether stable filtration is attained for given gas and dust properties.

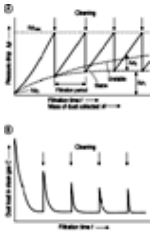

Figure 44. Pressure drop (A) and dust content (B) in clean gas versus filtration time at constant filtration velocity

The values Δp_1 and Δp_2 denote the pressure-drop contributions of the filtration medium and filter cake, respectively

During and immediately after cleaning, the particle content of the clean gas increases. This occurs partly because the filter cake, which acts as a layer of filter aid, has been removed. Especially with reverse-pulse cleaning, some of the dust passes through the filter medium because after the pressure pulse inside the bag has decayed, the filter medium falls back against the cage and knocks remaining dust through the filter.

7.1.5.4. *Granular-Bed Filters*

In granular-bed filters, the dust-laden gas stream flows through a bed of gravel, sand, ceramics, activated carbon, etc. and is thereby cleaned. The granular-bed functions as a filter medium and can be regarded as a depth or surface filter. Besides pure particle collection, simultaneous removal of gaseous components can be performed in such filters. Granular-bed filters are used chiefly for the removal of hot, chemically aggressive, abrasive, or adhesive dust or when danger of fire exists.

Granular-beds can be either fixed or moving. Filters with fixed beds are operated batchwise; when dust deposition has raised the pressure drop to a predetermined limit, the dirty gas stream is cut off and the bed is regenerated. The dust is removed internally by reverse flushing and agitation or stirring, or externally by screening or other means. In filters with moving beds the granules creep through the filter chamber by gravity, either steadily or at intervals. Fresh granules are added continuously from above. Regeneration takes place in an external loop. This filtration method operates without interruption of the dirty-gas stream.

The design of granular-bed filters is mainly empirical (as in collection generally). Typical initial fractional collection efficiency curves (no dust load in the bed) are shown in Figure 45.

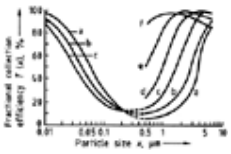

Figure 45. Fractional efficiency curves for a bed of glass beads at various filtration velocities

Bed height H = 20 mm; bead diameter d_G = 0.58 mm

Filtration velocity u_0, m/s: a) 0.05; b) 0.10; c) 0.25; d) 0.5; e) 1.0; f) 1.5

7.1.6. ***Electrical Precipitators***

7.1.6.1. *Mode of Operation and Basic Designs*

An efficient way of separating particles from gases is based on the forces exerted on charged particles in an electric field. This principle is important when inertial forces are no longer effective, i.e., for fine particles with diameters < 1 µm. Because the required capital investment is rather large, the main application of these precipitators is in dust removal from large gas streams (up to several million cubic meters per hour). Emission levels significantly lower than 20 mg/m^3 can be achieved if the dust properties are favorable. Pressure drops are generally slight (< 500 Pa).

The collection principle used in electrical precipitators is depicted in Figure 46 and involves three successive steps: charging of particles, transport of particles to the collecting electrode, and removal of particles gathered on the collecting electrode.

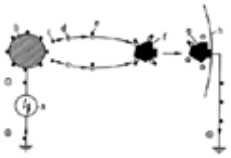

Figure 46. Schematic diagram of charging and collection process

a) High-voltage power supply; b) Emission electrode; c) Electron; d) Neutral molecule; e) Ionized molecule; f) Charged dust particle; g) Collected dust particle; h) Collecting electrode

The ions needed for particle charging are produced by corona discharge at the emission electrodes, which are made of thin wires or tapes with barbs. To obtain optimal collection efficiencies, high field strengths and high corona currents are required. The d.c. voltage is, therefore, adjusted to establish a condition as close as possible to electrical breakdown (ca. 60 – 80 kV), which depends on electrode geometry and gas properties.

Depending on particle size, the attachment of gas ions to particles occurs by one of two mechanisms: field charging or diffusion charging. Field charging dominates in the particle-size range above ca. 1 µm. The ions move toward the collecting electrodes along the electric field lines. As they move, the ions collide with particles and become attached to them. Diffusion charging results from random thermal motion of the ions (Brownian motion) and is important chiefly for particles smaller than 1 µm.

Charged particles are transported to the collecting electrode and must adhere to it. The electrical properties of the particles, particularly resistance, are important for the behavior of both the particle on impact and the resulting dust layer.

The favorable resistivity range for collection is 10^4 – 10^{11} Ωcm. Particles with a lower resistance give up their charge quickly when they collide with the collecting electrode; they change their polarity and are repelled from the electrode back into the gas stream. If the resistance is too high, the particles form a layer that continues to accumulate more charge, thus leading to a weakening of the electric field and discharge effects in the porous dust layer (back corona).

There are two main designs of precipitators. Wire-in tube precipitators are employed for small quantities of gas and especially for the collection of liquids. The wire emission electrodes are stretched along the central axis of parallel tubes. The inside walls of the tubes serve as collecting electrodes. If the tubes are set up vertically, the collected liquid runs down the walls as a film. If solid particles are collected, they are frequently rinsed out by flushing the tubes with water. The tubes are usually 0.2 – 0.3 m in diameter and 2 – 5 m long.

Parallel-plate precipitators contain many plane or profiled plates suspended at uniform intervals (200 – 600 mm) as collecting electrodes (Fig. 47). Gas flows horizontally in the channels between plates. Emission electrodes hang in the midplane between neighboring plates. Particle removal may be either dry or wet. In dry electrostatic precipitators, the dust layer collected on the plates is removed periodically by rapping the plates with hammers, whereby the dust falls into the hopper beneath the plates. The total plate length is divided into several zones for two purposes: dust reentrainment is reduced because the zones can be cleaned at different times; and separate zone-by-zone power controls allow the voltage to be adjusted in accordance with dust concentration.

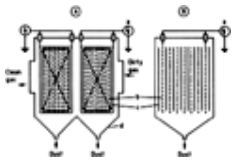

Figure 47. Schematic of a two-zone parallel-plate electrical precipitator (vertical sections A and B are shown at right-angles to each other)

a) High-voltage power supply; b) Emission electrode; c) Collecting electrode; d) Hopper

In wet electrostatic precipitators, the collecting electrodes are cleaned by flushing with water. The dust flows down off the plates with the water; dust reentrainment is thus effectively prevented.

7.1.6.2. *Design Calculations*

Calculations for precipitator design are concerned with collection efficiency; the very small pressure drop is not important. The well-known Deutsch precipitator equation is still used as the basis for dimensioning precipitators and for evaluating and interpreting measurements. If turbulent flow in the precipitator is assumed and if reentrainment is neglected, the Deutsch equation for fractional efficiency is

$$T(x) = 1 - \exp\left[-\frac{2L}{av}w(x)\right] \tag{9}$$

where L is the collection length, a is the plate spacing (parallel plate precipitators) or the radius of the tube (tube precipitators), v is the gas velocity, and $w(x)$ is the so-called migration velocity. Equation (9) can also be written as

$$T(x) = 1 - \exp\left[-\frac{A}{\dot{V}}w(x)\right] \tag{10}$$

where A is the surface area of the collecting electrodes and $\dot{V}$ is the volume gas flow rate.

The migration velocity $w(x)$ depends on the particle size and charge, the electric field, and the gas viscosity. It can be derived theoretically by solving the equation of particle motion. In practice migration velocity is calculated from the measured fractional efficiencies, by means of Equation (10); it often lies in the range of 10 – 30 cm/s. The parameter $w(x)$ is basically a transport coefficient that characterizes particle transport to the collecting electrode; it is thus vital for dimensioning electrical precipitators.

7.2. Waste-Gas Purification by Thermalor Catalytic Reactions

In thermal and catalytic waste-gas purification processes, gaseous air pollutants are converted by chemical reactions (oxidation, reduction, or decomposition) into harmless substances. These reactions can occur in flames (combustion [542]) or combustion chambers (e.g., thermal oxidation), or over catalysts (e.g., catalytic oxidation). In the following sections these processes are discussed with regard to their possible applications and the process and plant concepts.

7.2.1. ***Fundamentals***

7.2.1.1. *Types of Reaction*

Air pollutants can be disposed of in the gas phase at the site of emission by thermal and catalytic reactions. By utilizing the reaction energy of the exothermic reactions, thermal recycling is possible but not the recycling of material. The latter can be achieved by absorption, adsorption, condensation, or membrane processes.

The disposal effect in the case of thermal and catalytic processes consists in the conversion of the air pollutants into harmless reaction products by chemical reactions. These can be divided into oxidation reactions, reduction reactions, and decomposition reactions.

Substances regarded as harmless here include species natural to the air such as CO_2, H_2O, and N_2. Some examples of

reactions with their characteristic quantities and applications follow:

Oxidation reactions	
Example	catalytic hydrocarbon combustion
	$C_mH_nO_p + (m + n/4 - p/2)\ O_2 \longrightarrow m\ CO_2 + n/2\ H_2O$
Temperature	200 – 600 °C
Residence time	0.42 – 0.02 s
Space velocity	5000 – 50 000 h^{-1}
Application	catalytic afterburning, engine exhaust gas cleaning
Reduction reactions	
Example 1	selective catalytic nitrogen oxide reduction (SCR)
	$4\ NH_3 + 4\ NO + O_2 \longrightarrow 6\ H_2O + 4\ N_2$
Temperature	200 – 400 °C
Residence time	0.2 – 0.3 s
Space velocity	1000 – 5000 h^{-1}
Application	$DeNO_x$
Example 2	nonselective catalytic nitrogen oxide reduction (NSCR)
	$CO + NO \longrightarrow 1/2\ N_2 + CO_2$
	$C_mH_n + 2\ (m + n/4)\ NO \longrightarrow (m + n/4)\ N_2 + n/2\ H_2O + m\ CO_2$
Temperature	200 – 600 °C
Residence time	0.42 – 0.02 s
Space velocity	5000 – 50 000 h^{-1}
Application	engine exhaust gas cleaning
Decomposition reactions	
Example	catalytic ozone decomposition
	$2\ O_3 \longrightarrow 3\ O_2$
Temperature	20 – 100 °C
Residence time	0.7 – 0.3s
Space velocity	5000 – 10 000 h^{-1}
Application	water purification

Thus additional reactants (oxidizing agents, e.g., O_2, and reducing agents, e.g., NH_3) are required for conversion of air pollutants.

On the basis of the oxygen demand in oxidation processes, it is technologically reasonable to classify the gaseous emission streams as waste air streams (high O_2 content, differing little from the natural composition of the air and sufficient for safe burner operation) or waste gases (inert, low O_2 content, burners require additional fresh air for operation). In the following, however, in conformity with the definition of [543], carrier gases with solid, liquid, or gaseous emissions are referred to as waste gases.

7.2.1.2. *Reaction Engineering*

7.2.1.2.1. Homogeneous Gas-Phase Reactions

The course of homogeneous gas-phase reactions can be broken down into three process steps:

1. Preparation (e.g., atomization, swirling)
2. Mixing (diffusion, convection)
3. Reaction of the reactants

High- and low-temperature mechanisms have been formulated for the oxidation of hydrocarbons [544-551]:

1. High-temperature mechanisms: radical-chain mechanism at $T > 1000$ K
   hydrocarbon $\longrightarrow$ intermediate products
   intermediate product $\longrightarrow$ carbon monoxide
   carbon monoxide $\longrightarrow$ carbon dioxide
   (the last step is rate-determining)
2. Low-temperature mechanisms: $T < 1000$ K, formation of partially oxidized intermediate products, such as aldehydes, ketones, alcohols, peroxides, hydroperoxides, and carboxylic acids

The rate of the overall process is determined by the design parameters temperature, time, and turbulence.

The exothermic oxidation and reduction reactions (e.g., combustion) used for waste-gas purification are in practice carried out at temperatures so high that the reaction step is not rate-determining. The main task in carrying out the reaction and designing the equipment is mixing the reactants [552].

Flame Combustion. The course of reaction in flames is characterized [553] by high temperatures (>1200 °C), high radical concentrations, and luminescence phenomena.

Flame combustion is used for waste gases containing pollutants in concentrations near the explosion limit. Combustion takes place in flares after addition of air or oxygen, without energy supply (autothermally) or with a small energy supply [554]. By using staged oxygen supply the combustion can be carried out such that reduction reactions occur [555].

Thermal Reactions. Thermal waste-gas purification (e.g., thermal postcombustion or thermal oxidation) is generally used for waste gases whose pollutant concentration is well below the flammability limit. If the explosion limit can be exceeded as a result of output variations, safety devices must be installed [556]. An auxiliary heater provides the required minimum operating temperature in the combustion chamber [544, 545, 552, 553]. The processes occurring there are influenced both by the reactivity of the pollutants [546-551] and by conditions depending on plant design, such as the combustion chamber geometry [548], nature of the auxiliary heating, and flow and temperature fields in the combustion chamber. The operating temperatures of thermal oxidizer plants for reasonable residence times (0.1 – 2 s) are between 700 and 1200 °C, so heat-resistant materials must be used. The thermal energy of the combustion flue gases can be utilized by heat exchange [557], [558].

7.2.1.2.2. Heterogeneous Gas-Phase Reactions, Catalytic Waste-Gas Purification

Kinetics. In catalytic waste-gas purification pollutant gases and vapors are converted to harmless substances by chemical reaction over a solid inorganic catalyst.

Most industrial catalysts are highly porous materials with a large surface for chemical reaction. Because the reactants are present in a different phase from the catalyst, physical transport processes occurring on the catalyst are also important.

The important individual steps of such heterogeneously catalyzed reactions are shown in Figure 48.

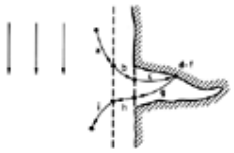

Figure 48. Individual steps of heterogeneous catalysis

a) Transport of the reactants through the gas stream to the boundary layer; b) Diffusion of the reactants through the boundary layer to the catalyst's external surface; c) Diffusion of the reactants through the pores to the catalyst's internal surface; d) Adsorption on the catalyst surface; e) Chemical reaction; f) Desorption of the products from the catalyst surface; g) Diffusion of the products through the pores to the catalyst's external surface; h) Diffusion of the products through the boundary layer; i) Transport of the products in the gas stream

The use of solid catalysts in gas reactions (heterogeneous catalysis) raises the rate of the reaction by condensation of the reactants on the catalyst surface and, as a result, increased collision frequency, and by decreasing the activation energy (see Fig. 49).

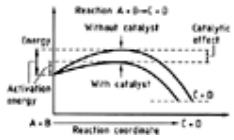

Figure 49. Simplified potential energy curves for homogeneous and heterogeneous exothermic reactions

Due to the increased reaction rate the conversion of air pollutants by chemical reaction is achieved at lower temperatures and with smaller reactors (shorter residence time) compared with thermal processes.

Requirements of Waste-Gas Purification Catalysts. The *activity* of a catalyst is its ability to increase the rate of a reaction in question under defined conditions. Activity is therefore not a fundamental property, but a property specific to a reaction and process parameters. The following are regarded as alternative measures of activity:

1. The percentage conversion at a given temperature and contact time and inlet concentration
2. The temperatures and contact times necessary to maintain certain percentage conversions or certain concentrations in the purified gas.

Only those reactions should occur on the catalyst that lead to the required products (*selectivity*). For example, during the catalytic total oxidation of hydrocarbons, partially oxidized intermediate products are converted with higher activity than the hydrocarbon. Intermediate products are therefore not emitted [559]. A catalyst should exhibit *stability towards deactivation,* i.e., its properties must remain unchanged for the longest possible lifetime. When selecting a catalyst, special attention must be paid to potential deactivation. It must be verified that

1. The catalyst is resistant to the substances present (often in traces) in the waste gas
2. The catalyst undergoes no structural changes in the operating temperature range
3. The catalyst is sufficiently resistant to abrasion and fracture by vibrations and stresses occurring in operation

Although the catalyst is not consumed in the chemical reaction, a change of properties during operation under industrial conditions must be expected. This leads to reversible or irreversible decrease in the catalytic activity (deactivation). Chemical and thermal factors can limit the effective life of catalysts.

Catalyst poisoning occurs when the waste gas contains substances that react with the catalyst support or the active component.

These undesirable reactions can lead to changes in chemical composition, reduction of surface area, loss of active components, or masking of active centers (e.g., coking).

The rate and frequency of structural changes (crystal growth, spinel formation, sintering) increase greatly with increasing temperature. These structural changes can lead to reduced internal surface area and decreased degree of dispersion of finely-divided active components.

For industrial catalysts, therefore, upper limits are specified for the temperature of use:

Metal oxide catalysts	
bulk catalysts	500 °C
supported catalysts	700 °C
Noble metal catalysts	
aluminum oxide supported	700 °C
stabilized aluminosilicate supported	1000 °C

Catalyst overheating during periods of intense heat of combustion release due to high pollutant concentrations can be avoided by introducing cold air or reducing the preheating.

Catalyst damage caused by mechanical influences include abrasion and fracture. Abrasion can be caused by vibration of the plant, pressure surges, loosening of the catalyst bed, and solid particles (abrasive dusts); and fracture by thermal stress owing to rapid temperature changes, and mechanical stress due to thermal expansions of the reactor.

Mechanical stress in reactors filled with honeycomb catalysts can be compensated by elastic internal seals.

The risk of catalyst beds being fluidized can be avoided by a vertical inlet flow from above. At the same time, however, the formation of a funnel-like bed surface by too high an inlet flow rate or by a swirling incident flow must be avoided.

The *pressure drop* in catalytic waste gas purification reactors can be influenced by selecting an appropriate geometry and type (e.g., honeycomb) for the catalyst as well as by the height of the catalyst bed.

Typical values of the specific pressure drops, per unit bed height, are 10 kPa/m for bulk catalysts and 1 kPa/m for honeycombs.

Reactor Design. Two methods can be used for reactor design.

Computing based on reaction engineering models [560], which are applicable if

1. The following kinetic data, necessary for determining the reaction rate, are known: activation energy, collision factor, reaction order, and kinetics of mass transfer.
2. There is no influence of interfering components.
3. Flow and temperature conditions are known. Simple reactor models are based on assumptions such as isothermal reactor, adiabatic reactor, or pseudohomogeneous reaction (i.e., no influence of mass-transfer processes).

Experimental simulation (*pilot-plant trials*) is more important in practice since:

1. Mixtures of pollutants are present, often with variable concentrations
2. Conditions of flow and temperature vary greatly
3. No kinetic data are available
4. The influence of further crude gas constituents is unknown
5. A pilot-plant trial is often necessary for detecting the presence of catalyst poisons

The results of these experiments are plotted in diagrams showing, for specific pollutants on specific catalysts, the dependence of the percentage conversion on inlet temperature, space velocity (under standard conditions), and inlet concentrations (Fig. 50).

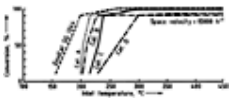

Figure 50. Curves of percentage conversion [559]

Activity comparison of various metal oxide catalysts (pollutant: 1000 ppm ethyl acetate) in air; catalyst aged for 168 h at 500 °C; space velocity = 15 000 h^{-1})

These diagrams can be used as the basis for design according to the following scheme:

1. Determination of the percentage conversion required
2. Determination of the catalyst inlet temperature at the expected inlet concentration (taking into account the temperature rise due to the heat of reaction)
3. Determination of the permissible space velocity under normalized design conditions (SVN)
4. Calculation of the catalyst volume required

$$V_c = \frac{\Phi v, n}{\mathrm{SVN}} = \frac{\text{volumetric flow rate(STP)}}{\text{space velocity(STP)}}$$

If mixtures of contaminants are present, the temperature and space velocity should be determined from the reaction behavior of the least reactive component.

During the catalytic afterburning of hydrocarbon mixtures the individual components are selectively oxidized, so the proportion of less reactive components increases with increasing temperature. The overall fractional conversions therefore exceed the value for the less reactive component.

When designing the catalytic waste-gas purification reactors, it must be ensured that the thoroughly mixed reactants are distributed homogeneously over the entire inlet flow cross section of the catalyst bed. This can be achieved by mixing blades, baffle plates, and straightening blades.

7.2.1.3. *Process Selection and Plant Concepts*

The process is selected primarily from knowledge of the appropriate type of reaction for pollutant conversion under the prevailing waste-gas conditions of temperature, pressure, composition (components and concentration) and required conversion.

The choice of process is also determined by economic considerations [563] such as investment costs; operating costs (energy supply, waste-heat utilization [557], [564], pressure drop); and the life of components such as catalysts, heat exchangers, and valves; formation of byproducts (e.g., HCl during the oxidation of chlorinated hydrocarbons), waste disposal (e.g., of catalysts).

The following process steps can be regarded as the basic operations of waste-gas purification processes:

1. Transfer of the waste gas (e.g., by suction)
2. Mechanical transport of the exhaust air (e.g., by ventilators, compressors)
3. Separation of interfering components (e.g., by filters, adsorbers, absorbers)
4. Heating of the waste gas by heat exchange
5. Supply of external energy (e.g., by burners, electrical resistance heating)
6. Mixing with oxidizing or reducing agents (e.g., oxygen, air, ammonia)
7. Pollutant conversion (e.g., by flame, thermal, or catalytic reaction)
8. Waste heat utilization (e.g., for heating steam or heat-transfer oil)
9. Separation of reaction products (e.g., HCl)
10. Discharge of purified gases (e.g., from the stack)

Economically, integration of waste-gas purification into the production process should be aimed for. In oxidation processes this can be achieved by waste-heat utilization [557], [560], [564], [576].

7.2.1.4. *Operational Check [566], [567]*

Waste-gas purification plants lead to reduced emission mass flow rates.

If waste gas mass flow rates are known, concentration measurements give information on the efficiency of the plant and are prescribed as emission values in permits.

In measuring emission concentrations, it must be taken into account for processes with pollutant conversion that the type of substance can change. Therefore, with thermal combustion processes, for example, the measurement of CO content in the purified gas is specified. Catalytic processes usually operate selectively.

The frequency of emission monitoring is prescribed for the specific plant and adapted to plant size and fluctuations in operating conditions.

For chemical conversions the reaction temperature largely determines the rate of the process. Temperatures (combustion chamber outlet, catalyst inlet/outlet) are therefore frequently used as control variables.

It is not useful to specify standard oxygen contents for oxidative waste-gas purification processes, since the oxygen concentration of the waste gas is predetermined by the emitting process and reduced as little as possible by an energy optimized conduct of the process (i.e. one with reduced fuel requirement) [552].

7.2.2. ***Thermal Processes for Waste-Gas Purification***

7.2.2.1. *Oxidation Processes*

Thermal oxidation processes associated with flames are known as combustion. Complete combustion converts the carbon bound in the atmospheric pollutant or in auxiliary fuels to carbon dioxide. In the treatment of exhaust air streams containing

sulfur or halogens, the formation of SO_2 and HCl or Cl_2 necessitates downstream separation. If chlorine-containing compounds are present, attention must be paid to the formation of polychlorinated compounds. In the presence of precursor compounds (e.g., chlorobenzene), polychlorinated dibenzodioxins and dibenzofurans can be formed under conditions such as local oxygen deficiency, low temperatures (300 – 600 °C), or inadequate residence time in the presence of oxygen.

Therefore in the design of combustion systems (intensive mixing) and waste-heat recovery (rapid cooling), these conditions must be avoided in order to avoid “de novo” synthesis from hydrocarbon radicals and organic or inorganic chlorine [552], [568], [569]. If organic nitrogen compounds or NH_3 are present, nitrogen oxides are formed [546], [549]. These can be avoided by control of the combustion or converted by waste gas aftertreatment.

The sequence of the following process descriptions follows the specific CO_2 formation to be expected, taking into account pollutant conversion, external energy supply, and waste-heat utilization. The amount of CO_2 formed is least if the combustible substances in the waste gas are used as replacement fuel in existing furnaces or stationary motors. With these processes, however, attention must be paid to nitrogen oxide formation. In separate waste-gas purification units, specific CO_2 formation decreases with increasing waste-gas preheating or with increased cooling of the purified gas by waste-heat utilization before discharge to the atmosphere.

7.2.2.1.1. Waste-Gas Disposal in Furnaces [557], [564]

Waste gases containing oxygen can be added to the combustion air of furnaces provided it is ensured, if necessary, by fitting (see Fig. 51), that

1. Deposits from the waste gas (e.g., cracking products) in the air supply system of the burner are avoided
2. The resulting waste-gas stream is always supplied to the heater
3. Fresh air is available for the generation of steam

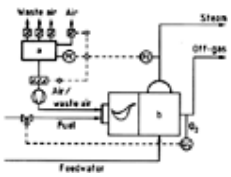

Figure 51. Waste gas thermal purification in a boiler furnace [557], [564]

a) Collection chamber; b) Steam boiler

Conditions (2) and (3) can be realized by using a vacuum-controlled waste-gas collecting chamber with a fresh-air valve [557].

With regard to condition (2) it must be checked whether there is simultaneity of waste gas production and furnace operation in the correct order of magnitude (approximately 800 – 1000 m^3/h waste gas for 1 t/h steam) [564].

7.2.2.1.2. Thermal Combustion

Waste-Gas Preheating by Regenerative Heat Exchangers [558, 570, 571]. Regenerative heat exchangers are characterized by:

1. The presence of a heat-storage material
2. Position- and time-dependent temperature profiles in the storage material for fixed beds (see Fig. 52)
3. An alternating flow direction (fixed bed) or moving storage materials (moving bed)
   Advantages of these systems are:
4. High degrees of heat exchange

$$n = \Phi_E/\Phi_E\max = \frac{T_{\text{hot,inlet}} - T_{\text{hot,outlet}}}{T_{\text{hot,inlet}} - T_{\text{cold,inlet}}} > 90\ \%$$

5. Homogenization of the flow

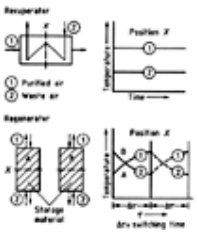

Figure 52. Difference between recuperators and regenerators [558]

The following features of this system must be taken into account:

1. Periodically changing mixture of waste gas and purified gas, since the waste gas path is the same as the purified gas path. Therefore, at high preheating temperatures the use of uneven-numbered tower layouts for carrying out a

purging phase with recycle of the unpurified gas is advisable. By opening and closing the respective valves, the tower in purge mode is connected to a line leading to the suction side of the blower. After purging the crude gas from the void fraction of the bed, the hot purified gas from the burner chamber is connected to the purged tower. The bed is heated by heat transfer from the hot gases. Thus the cycle of modes is completed without giving untreated gas access to the purified gas duct.
2. Large weight.
3. Mechanical stress on the moving parts (valves, rotors, seals) with cycle times of 30 – 200 s [558].
4. Fouling of the storage material by waste gas constituents [571].

The same applies to catalytic oxidation with heat exchanger designs of this type [589].

Figure 53 shows a regenerative thermal oxidation plant [572]. The process consists of the following steps:

1. The process exhaust stream enters the oxidizer through the common inlet manifold and passes into a regenerative chamber through a butterfly valve.
2. The exhaust stream then passes through a bed of ceramic heat-exchange matrix, which raises it to near oxidation temperature.
3. The contaminated air then enters the combustion (oxidation) chamber, which is maintained at ca. 800 °C by burners. The heat released by oxidation of the VOCs lowers the fuel consumption of the burners required to maintain the oxidation temperature.
4. The air stream then exits the oxidation chamber through another bed of ceramic heat-exchange matrix. The clean exhaust then gives up 95 % of its thermal energy to the matrix to be reused for preheating on the next inlet cycle.
5. The clean air is then drawn through an outlet valve and exhaust manifold for discharge to the atmosphere.

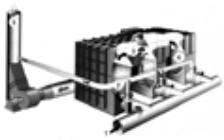

Figure 53. Regenerative thermal oxidation plant [572]

1) Exhaust inlet; 2, 4) Ceramic heat-exchange matrix; 3) Oxidation chamber; 5) Exhaust manifold

Figure 54 shows various designs of regenerative combustion systems.

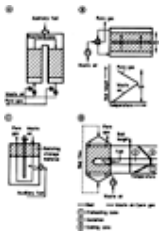

Figure 54. Schemes of various regenerative combustion systems [558]

A) Fixed-bed regenerator; B) Combu-Changer; C) Ljungstrom regenerator; D) Moving-bed regenerator

The Combu-Changer and moving bed systems carry out the thermal oxidation in the bed of heat-storage material.

In the other systems, the combustion chamber and heat exchanger are separate [572]. The dimensions are based on a given residence time in the combustion chamber (e.g., 0.5 s) [571]. Since the reactions also occur in the heat bed, the residence time in the total reaction volume is ca. 1 s [567]. Processes that mainly involve thermal oxidation without the action of a flame are known as regenerative thermal afterburning or regenerative thermal oxidation (RTO, thermo-reactor, etc.). At high concentration peaks, thermal energy must be withdrawn from the reaction zone by a bypass to avoid buildup of the temperature [558], [571]. Due to the relatively high combustion space temperature, the large excess of air, and small influence of flame, only small amounts of CO and NO_x are formed.

Typical process parameters and range of application [555]:

Temperature range	800 – 1000 °C
Residence time	ca. 1 s
Conversion	>99.5 %
Heat utilization, waste-gas preheating	90 – 97 %
Waste-gas volume flow rates (STP)	5000 – 10 000 m^3/h
Pollutant concentrations	≤10 % of lower explosion limit
Suitable for continuous waste-gas emission	
Unsuitable for discontinuous waste-gas emission	

Waste-Gas Preheating by Recuperative Heat Exchangers. In these processes the combustion system consists of a burner and an afterburner chamber.

Burner designs are:

1. Flat burner (see Fig. 55)
2. Swirl combustion chamber (see Fig. 56)
3. Lance burner (see Fig. 57)

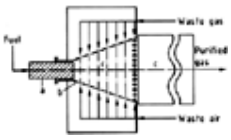

Figure 55. Flat burner [552]

a) Feeding and monitoring elements; b) Mixing cone; c) Mixing and reaction space

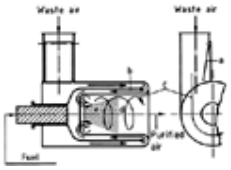

Figure 56. Swirl combustion chamber [552]

a) Swirl flap; b) Annular gap; c) Swirling flow; d) Combustion space; e) Swirling flame

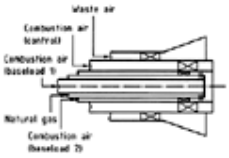

Figure 57. Lance burner [555]

In the burner the auxiliary fuel (preferably natural gas) is burnt. The necessary oxygen can be provided by oxygen-containing waste gases (diffusion burner, types 1 – 3) or, if the oxygen content of the waste gas is insufficient, from combustion air fed to the burner together with the fuel (premixing burner, type 3). The added combustion air increases the waste gas mass flow rate, resulting in increased extraneous energy requirement [557]. Initial pollutant conversion occurs during the mixing of the waste gases with the hot flue gases of the burner, partly by contact with the flame.

The afterburner chamber provides additional residence time at high temperatures for the oxidation reactions to proceed to completion, ensuring total combustion. In the swirl combustion chamber system, burner and afterburner chamber are combined in a single unit. The adjustability of the combustion system must ensure matching to different waste-gas mass flow rates.

For waste-gas preheating the following recuperative gas – gas heat exchangers are used (see Figs. 52, 58):

1. Circular-disk recuperator
   advantage: = space-saving construction, low weight; suitable for compact units
   disadvantage: = sensitivity to fouling, and thermal stress as a result of hindered differential expansions

2. Duct recuperator
   advantage: = robust construction, easily maintained, good compensation of thermal expansions
   disadvantage: = high weight

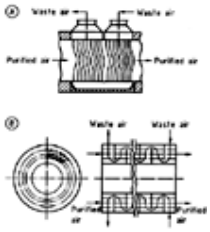

Figure 58. Recuperator designs [557]

A) Duct recuperator; B) Circular disk recuperator

Units can be classified on the basis of the structural arrangement of these components as compact units (see Fig. 59), which are often integrated into production plants [563], and units with separate arrangement of the components (see Fig. 59).

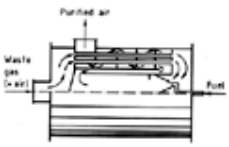

Figure 59. Compact thermal afterburning units [557]

For units that are started up and shut down frequently, an all-steel construction with ceramic-fiber lining is preferable.

Waste-heat utilization can be carried out by downstream, steam generation, heat-transfer medium heating, or hot-air generation.

Typical process parameters and range of application [555]:

Temperature range	800 – 1200 °C
Residence time	0.3 – 2 s
Conversion	>99.5 %
Heat utilization,	
waste-gas preheating	50 – 55 %
waste-heat utilization	50 – 80 %
Pollutant concentration, % of LEL	
with safety engineering	0 – 100 %
without safety engineering	0 – 50 %
Volume flow rate (STP)	1000 – 50 000 m^3/h
Suitable for continuous waste-gas emission	
Less suitable for discontinuous waste-gas emission	

7.2.2.1.3. Combustion in Flares

Combustion in flares is used mainly in the chemical industry and in plants for oil and natural gas recovery and treatment (e.g. refineries), for waste gases originating from the following operating conditions of the associated production plants [554]:

1. Start-up and shut down
2. Disturbances with controllable operating conditions (e.g., safety valves)
3. Disturbances with uncontrollable operating conditions (e.g., power cut or cooling-water failure)

The construction of a high flare is shown in Fig. 60. The combustion processes occur in the flare stack tips which can be of three types (see Fig. 61) free flame flares, screen flares, and muffle flares.

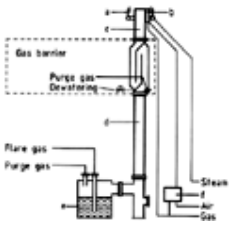

Figure 60. High flare system [554]

a) Steam jets; b) Pilot burner; c) Flare head; d) Flare stack; e) Water seal; f) Igniter

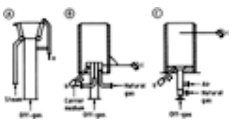

Figure 61. Types of flare [555]

A) Free-flame flare; B) Screen flare; C) Muffle flare

a) Pilot burner; b) Ignition burner; c) Thermoelement

There is no utilization of heat from flares.

Typical process parameters and range of application [555]:

Temperature range	> 800 °C
Residence time	1 – 2 s
Conversion	> 99.5 %
Heat utilization	not provided
Pollutant concentration, % of LEL	
with safety engineering	0 – 100 %
without safety engineering	0 – 50 %
Volume flow rate	≤ 1000 m^3/h

7.2.2.1.4. Safety Engineering in Thermal Waste-Gas Purification

The combustion chamber of waste-gas thermal purification plants constitutes a permanent ignition source. If the waste-gas

composition can be explosive permanently (Zone 0), occasionally (Zone 1), or rarely (Zone 2), then three, two, or one independent protective measures, respectively, must be provided for introduction of the waste gas [556], [573]: (1) low-flashback inlet to burner, (2) explosion safeguard, (3) immersion safeguard.

7.2.2.2. *Reduction Processes*

7.2.2.2.1. Staged Combustion

Waste gases that contain organic nitrogen compounds or NO_x can be purified by combustion with staged oxygen supply. The nitrogen bound in the pollutants is emitted in molecular form (N_2). In the first step an atmosphere with reducing compounds (e.g., CO, H_2, NH_3) is generated by substoichiometric combustion of a fuel (oxygen deficiency). NO_x formation therefore stops, or that which is already present is reduced. This effect can be intensified by cooling (steam addition) in the first step [555]. The unreacted combustibles emerging from this step are afterburnt in the second step with an excess of air ($\lambda > 2$).

For establishing substoichiometric conditions in oxygen-containing waste gases, these processes sometimes require considerable amounts of fuel, with which an energy release is associated. This is a stoichiometric requirement of the process and cannot be decreased by preheating the waste gas [555].

7.2.2.2.2. Thermal Reduction of Nitrogen Oxides

The reactions:

$$4\,NH_3 + 4\,NO + O_2 \rightarrow 6\,H_2O + 4\,N_2$$

$$CO(NH_2)_2 + 2NO + \frac{1}{2}O_2 \rightarrow 2\,N_2 + CO_2 + 2\,H_2O$$

occur as homogeneous gas-phase reductions in a temperature range 900 – 1000 °C. This temperature range can be established in furnace and waste-gas purification plants. Ammonia [555] or urea solution [574] is injected into this temperature zone by lances. By injecting additional chemicals, the temperature range of the urea process has been extended to 500 – 1200 °C, allowing NO_x conversions of > 80 % to be attained [574].

7.2.3. ***Catalytic Waste-Gas Purification Processes***

7.2.3.1. *Catalytic Reduction of Nitrogen Oxides in Flue Gases and Process Off-Gases*

Catalytic reduction is used for the removal of nitrogen oxides from flue gases and process off-gases (e.g., furnaces, nitric acid plant, glass melting, refuse incineration plants). The process is known as $DeNO_x$ or SCR (selective catalytic reduction).

In the SCR process, ammonia is added as a selective reducing agent (i.e., one that is effective even in presence of oxygen), and the nitrogen oxides are converted to molecular nitrogen:

$$4\,NH_3 + 4NO + O_2 \rightarrow 6\,H_2O + 4\,N_2$$

Process description: SCR for power station flue gases

Two process variants can be distinguished. In the *crude-gas circuit* (see Fig. 62).

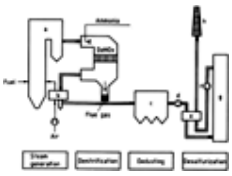

Figure 62. $DeNO_x$: crude-gas circuit [560]

a) Steam generator; b) Air preheater; c) Filter; d) Induced-draught blower; e) Gas heater; f) Blower; g) Washer; h) Stack

The flue gases are taken off directly from the steam generator (after the boiler feedwater preheater, before the combustion air preheater) and, after addition of the reducing agent, passed over the catalyst.

The flue gas, now largely free of nitrogen oxides is introduced into the combustion air preheater to preheat the combustion air (heat recovery). Other off-gas purification stages (dedusting, desulfurization) follow.

Crude-gas preheating or addition of fuel prior to the catalyst is not necessary since the flue gas is removed from a point in the steam generator where it is sufficiently hot (350 – 450 °C).

The catalysts are either bulk honeycombs in the fixed-bed process (Ti, V, Fe, W, Ni, Co, Cu, Cr, U, Mo, Sn as oxides, Ag, Be, Mg, Zn, Al as metals); these can also be used for removing dioxins from flue gases in municipal waste incineration plants [575], or activated coke in the moving bed process [576].

Catalyst requirements are selectivity ($SO_2 \longrightarrow SO_3$ conversion must be avoided), activity (low NH_3 consumption), and high abrasion resistance owing to the dust content of the flue gases.

In the *clean-gas circuit* (see Fig. 63), dust and sulfur oxides are removed by electric precipitators or flue-gas scrubbers before the crude gas/reducing agent mixture reaches the catalyst. The crude gas must be preheated to the required

reaction temperature of 300 – 400 °C.

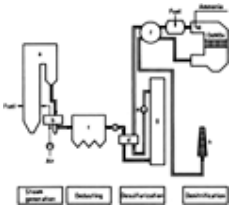

Figure 63. $DeNO_x$: pure-gas circuit [560]

a) Steam generator; b) Air preheater; c) Filter; d) Gas heater; e) Blower; f) Gas heater; g) Washer; h) Stack

Honeycomb catalysts are used, similar to those of the crude-gas process, but with higher activity. Prior separation of interfering components means that demands on the selectivity are lower and a longer effective life can be expected.

Typical design parameters

Space velocity	1000 – 10 000 h^{-1}
Temperature	250 – 450 °C
NH_3 addition ratio	
of the stoichiometric ratio	≤ 100 %

7.2.3.2. *Engine Exhaust Emission Control*

Exhaust emissions from internal combustion engines contain pollutants of the following three classes: hydrocarbons (C_mH_n), carbon monoxide (CO), and nitrogen oxides (NO, NO_2).

By using special catalysts and a suitable air-to-fuel ratio in spark-ignited internal combustion engines the three classes of pollutants can be removed together in a single catalytic converter.

The main reactions, occurring in parallel, are [577]:

1. Oxidation reactions

$$C_mH_n + (m + n/4)O_2 \rightarrow mCO_2 + n/2\,H_2O$$

$$CO + 1/2\,O_2 \rightarrow CO_2$$

$$H_2 + 1/2\,O_2 \rightarrow H_2O$$

2. NO-reducing reactions (nonselective catalytic reduction)

$$CO + NO \rightarrow 1/2\,N_2 + CO_2$$

$$C_mH_n + 2(m + n/4)NO \rightarrow (m + n/4)N_2 + n/2\,H_2O + mCO_2$$

$$H_2 + NO \rightarrow 1/2\,N_2 + H_2O$$

The catalyst is integrated into the exhaust system of the engine.

Owing to the predetermined exhaust gas temperature, preheating or external energy supply is unnecessary. The required concentrations of oxidizing agents (O_2, NO) and reducing agents (C_nH_m, CO, H_2) are established by feedback to the air-to-fuel ratio in the engine [578] using an electronic lambda probe. In the case of block power stations driven by gas engines, the thermal energy of the purified gases is recovered. Noble-metal-coated honeycomb catalysts are used. The noble metals are mainly platinum, palladium, and rhodium. Coating and support are optimized for high thermal stability.

Typical design parameters

Start-up temperature	ca. 200 °C
Space velocity	up to ca. 10^5 h^{-1}
Pressure drop	<1 kPa

Exhaust emissions from diesel engines can lead to problems in industrial safety and environmental protection owing to their content of hydrocarbons and derivatives e.g., aldehydes) (HC), carbon monoxide (CO), nitrogen oxides (NO, NO_2), and particles (soot, polycyclic aromatics).

Oxidizable constituents (hydrocarbons, CO) can be removed from the exhaust gas by oxidation catalysts.

High catalyst selectivity is required since the oxidation of SO_2 to SO_3 and of NO to NO_2 must be avoided [579]. For retaining the particles various mechanical and electrostatic systems are known. The honeycomb filter has a catalytic coating and can be regenerated by burning off at ca. 400 °C.

Owing to the high oxygen content of diesel exhaust, the three-way catalysts cannot be used to decrease the NO_x content. The NO_x can be decreased by primary measures (e.g., exhaust gas recycling and filtration) or selective secondary measures (SCR: see Section Catalytic Reduction of Nitrogen Oxides in Flue Gases and Process Off-Gases).

7.2.3.3. *Gas-Turbine Exhaust Emission Control*

SCR catalysts are available for gas turbine cogeneration and combined cycle installation [580]. A newly developed zeolite catalyst is available for the simple cycle gas turbine applications which require higher catalyst operating temperatures. A typical gas turbine SCR system is outlined schematically in Figure 64.

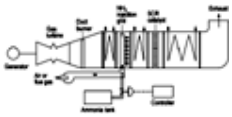

Figure 64. Typical cogeneration heat recovery steam generator (HRSG) installation [580]

7.2.3.4. *Catalytic Purification of Claus Plant Off-Gases*

Sulfreen Process [581], [582]. Typical Claus plant off-gases contain 1 % H_2S and 0.5 % SO_2.

In the Sulfreen process (Fig. 65) the reactions of the Claus process occur at fairly low temperatures (120 – 140 °C). Beds of modified alumina catalyst are used. The elemental sulfur formed by reaction of H_2S with SO_2 is adsorbed in the pores of the aluminum oxide, reversibly poisoning the catalyst. Cyclic purging with hot inert gas is used to regenerate the catalyst and recover sulfur.

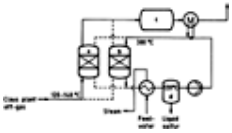

Figure 65. Sulfreen process [560]

a) Adsorption, reaction; b) Desorption; c) Afterburning; d) Condensation

Catalytic Afterburning [583]. Sulfur compounds (H_2S, COS, CS_2) remaining in the waste-gas stream must be oxidized to SO_2 before the gas is discharged to the atmosphere. At high waste-gas volume flow rates the energy consumption is distinctly reduced, compared with thermal combustion methods, by using alumina-containing catalysts.

7.2.3.5. *Catalytic Oxidation in Industrial Off-Gas Streams*

Total Oxidation. Industrial and commercial exhaust air streams that contain oxidizable impurities can be purified by catalytic total oxidation (catalytic afterburning) [579], [584-587]. Table 19 shows some examples of the reactions. The use of catalytic afterburning is restricted by the presence of catalyst poisons and by the need to prevent catalyst overheating with highly variable pollutant loadings.

Table 19. Combustion reactions and heating values of some air pollutants [575]

Compound	Combustion reaction	Heating value, kJ/kg	ΔT_{ad}, K, in air per 1 g/m^3
Toluene	$C_7H_8 + 9\ O_2 \longrightarrow 7\ CO_2 + 4\ H_2O$	40 935	30.6
Methanol	$CH_3OH + 1.5\ O_2 \longrightarrow CO_2 + 2\ H_2O$	21 119	15.8
Acetone	$C_3H_6O + 4\ O_2 \longrightarrow 3\ CO_2 + 3\ H_2O$	29 065	21.9
Ethyl acetate	$C_4H_8O_2 + 5\ O_2 \longrightarrow 4\ CO_2 + 4\ H_2O$	23 530	18
Carbon monoxide	$CO + 0.5\ O_2 \longrightarrow CO_2$	10 423	7.5
Ammonia	$NH_3 + 0.75\ O_2 \longrightarrow N + 1.5\ H_2O$	18 424	13.9
Hydrogen sulfide	$H_2S + 1.5\ O_2 \longrightarrow SO_2 + H_2O$	15 355	11

Noble metal and metal oxide catalysts can be used for total oxidation [586].

Example 1: catalytic afterburning unit with recuperative waste-gas preheating [588].

Styrene-containing exhaust air is emitted from a plant for producing glass fiber reinforced plastic moldings. After reducing the exhaust-air volume flow rate by enclosing the production machines, the catalytic afterburning unit was designed for a volumetric flow rate (STP) of 15 000 m^3/h with a styrene loading (STP) of ca. 5 g/m^3. The heat exchanger allows the unit to operate autothermally above a loading (STP) of ca. 2.5 g/m^3. At higher loadings, heat recovery for heat-transfer fluid heating is possible.

Operating data:

Temperature	
Crude gas	30 °C
Catalyst inlet	320 °C
Catalyst outlet	470 °C
Stack	100 °C
Heat recovery (heat-transfer oil heater)	465 kW
Average natural gas consumption (for heating at low pollutant concentration)	6 m^3/h
Electrical load connected	50 kW
Catalyst volume	800 L
Catalyst type	honeycomb EnviCat KCO WK 4100 with 1 g Pt/L

The concentrations in the purified gas are less than 10 mg/m^3, and odor nuisance in the surroundings of the plant is hence avoided [560].

Example 2: regenerative catalytic oxidation (RCO [561]) with waste-gas preheating (Fig. 66).

1. The VOC-laden process exhaust stream enters the oxidizer through an inlet manifold and is allowed to enter a regenerative chamber by means of control valves.
2. The airstream passes upwards through a bed of hot ceramic heat-sink material [562] and is heated to a temperature which allows oxidation to occur in the presence of catalyst.
3. The VOCs passing through the upper catalyst layer are oxidized.
4. The combustion chamber is used to regulate the temperature of the airstream so that, in conjunction with thermal transfer to the heat sink and catalyst oxidation, the airstream is regulated to the correct temperature for complete oxidation to occur.
5. The airstream passes downwards through an exit layer of catalyst [561] to complete the oxidation process.
6. As the hot oxidized airstream passes downwards through the next regenerative chamber, the heat is absorbed by the cooler bed of ceramic heat-sink material. After the heat sink in the inlet chamber has been heated and cooled, and the heat sink in the outlet chamber heated, the airflow control valves are reversed, so that the chambers operate alternately as inlet or outlet chambers.
7. The cleaned airstream is drawn through the outlet manifold and discharged to atmosphere by an induced-draft blower.
8. The transition duct is used to provide smooth pressure conditions while the flow direction is being reversed.

Figure 66. Catalytic oxidation with waste-gas preheating [561]

1) Exhaust inlet; 2) Bed of ceramic heat-sink material; 3) Catalyst layer; 4) Combustion chamber; 5) Exit layer of catalyst; 6) Regenerative chamber; 7) Outlet manifold; 8) Transition duct

Combination of Thermal and Catalytic Afterburning [560]. To the operator of thermal afterburning units it often seems desirable to reduce the combustion chamber temperatures, to reduce energy consumption (natural gas) and the fatigue or wear of components such as burners or exhaust air heat exchangers.

However, this also leads to formation of intermediate products of hydrocarbon oxidation. Relatively high combustion chamber temperatures (>750 °C) are necessary to ensure that the aldehydes and carbon monoxide formed as

intermediate products in the combustion chamber are converted to carbon dioxide within the predetermined residence time.

If conditions of flow, temperature, or residence time are unfavorable (e.g., during capacity changes), the emission of secondary pollutants (e.g., CO) can be detected in the flue gas of thermal afterburning units.

Increasing the combustion chamber temperature is a suitable countermeasure, but this often runs up against equipment limitations. Moreover, it must be investigated how far the temperature can be raised without significantly increasing the NO_x emissions.

The use of waste-gas purification catalysts enables secondary pollutant emission to be avoided. A lower combustion chamber temperature that ensures oxidation of the primary pollutants (solvents) with reliable observance of the emission limits can then be used.

The object of combining homogeneous and heterogeneous pollutant combustion (thermal and catalytic afterburning) is the conversion of hydrocarbons in the gas phase, and of carbon monoxide and unburnt hydrocarbons on the catalyst.

By dividing the reaction stream into two reaction zones as well as by matching the operating temperature to the thermal ignition temperature of the hydrocarbons, the catalyst can be designed for a higher space velocity than in the case of a single-stage catalytic combustion unit. As a result of the lower concentrations at the entrance to the catalyst stage, smaller temperature differences between inflow and outflow are to be expected.

Owing to the relatively high inlet temperatures for waste-gas purification catalysts, a high thermal stability is demanded of the catalyst. In addition, particularly when refitting old plants, only a relatively small pressure drop is available for flow through the catalyst stage. Therefore, honeycomb catalysts with a large surface area for incident flow are preferred.

For example, in an adhesive tape processing plant a thermal afterburning unit has been retrofitted with a catalyst stage, which allows the exhaust air flow of 30 000 m^3/h (STP) and contaminated with ca. 10 g/m^3 solvent, to be purified at a combustion chamber temperature of ca. 500 °C [560]. The installed catalyst volume is 1 m^3. For hydrocarbons and CO the concentrations measured in the purified gas at the stack are distinctly below the emission limits.

Droplets of high-boiling hydrocarbon compounds (aerosols), which are present in the waste gas and on contact with the catalyst surface could lead to coking or the masking of active centers, are vaporized in the combustion chamber. These compounds therefore burn in the vapor state without residue.

Rotary Concentrator Adsorption (RCA) System (Fig. 67) [562]. An RCA system increases the fuel value of VOC-laden gas streams and the fuel economy of the oxidizer by raising the VOC concentration in the inlet gases. The RCA uses a rotating adsorption bed of zeolite or activated carbon to trap VOCs and vents the cleaned air to the atmosphere. As the bed revolves, the trapped VOCs are subjected to a warm stream of desorption air and are released. This much-reduced volume passes into the oxidizer for final treatment. Savings vary, depending on the nature of the emission flow and operating conditions. A concentrator will affect the selection of the oxidizer type, and the combination of concentrator and oxidizer must be evaluated as an integrated system.

Figure 67. Rotary concentrator adsorption system [562]

7.3. Absorption

7.3.1. *Introduction*

Absorption is a process for removing some components of a gas stream by contact with a liquid phase (solvent). The desired selectivity with respect to the component(s) to be removed (solute) depends either on purely physical interactions (physical absorption) or on strong chemical bonding (chemical scrubbing or stripping). So-called physicochemical scrubbing occupies an intermediate position. Here, dissolution takes place in the absorbing liquid due to physical interaction accompanied by a reversible chemical reaction.

There are three major industrial applications of absorption processes. In one large group applications, undesired components such as H_2S or organic sulfur compounds, and, in some cases, water vapor, are removed from natural gas, refinery gases, synthesis gases, or process gases. Another major application is the removal of gaseous pollutants such as hydrogen halides or volatile organic solvents from waste air. The third group includes the removal of SO_2 and NO_x from the flue gases of power stations or domestic waste incinerators.

This section treats the industrial use of equipment for absorption processes. The discussion is limited to processes in which only absorption takes place. Apparatus principally intended for wet scrubbing of solids and only incidentally performing an additional absorption function, such as venturi or rotary scrubbers, is described in Section Separation of Solid and Liquid Particles from Gases .

7.3.2. *Basic Principles of Absorption Processes*

The most important element of an absorption plant is the absorber (Fig. 68), in which the gas stream being treated and the

solvent are brought into contact. The purified gas and the loaded solvent then leave the absorber. In some cases (e.g., absorption of HCl and NO_x with water; absorption of NH_3 with sulfuric acid) the loaded solvent is itself a usable product, so that a solvent regeneration stage is not required.

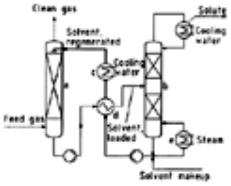

Figure 68. Flowsheet of an absorption process

a) Absorber; b) Desorber; c) Cooler; d) Recuperator; e) Boiler

However, in most cases it is necessary to regenerate the solvent for closed-loop recycling to the absorption stage. This takes place in the desorption stage, which can be a distillation step, as shown in Figure 68. Alternatively, steam or inert-gas stripping is sometimes used. For thermodynamic reasons, desorption should be operated at high temperature and low pressure. The efficiency of the recuperator (d) is thus important for the economics of the process. In physicochemical scrubbing, a desorption process can be considered if the chemical bonding can be broken by pressure reduction or temperature increase, but if the chemical bond is very strong the solvent cannot be regenerated. In most cases where this happens, the charged absorbing liquid is itself a usable product.

The absorption and desorption stages can be designed according to established methods (see ⟶ Distillation and Rectification) based on knowledge of the equilibrium between the partial pressure of the gaseous component in the gas phase and its concentration in the liquid phase. Details of these solution equilibria can be found in the comprehensive reference books [589], [591]. For newer processes, such as the absorption of organic substances in high-boiling solvents, special literature should be consulted [592].

The operating costs of a combined absorption/desorption plant mainly depend on the recirculating flow of solvent. As heat recovery in the recuperator is incomplete, the recycled solvent must be further cooled. To maintain thermal balance, it is therefore necessary to supply energy to the desorption stage. This occurs at the bottom of the desorption column if distillative treatment is used, or by a superheater before the inlet to the desorption column in the case of stripping. Note that desorption (with the exception of inert gas stripping) requires a certain minimum supply of steam, depending on the substance being desorbed, and this requires additional energy input to the process.

The heat of absorption is liberated in the absorption process; it is usually slightly higher than the enthalpy of condensation of the absorbed gas. However, in the case of strongly nonideal equilibria, the heat of absorption may be very high (e.g., 2100 kJ/kg for HCl). This fact, together with the possibility of heat and mass transfer between the gas and liquid streams due to different inlet conditions, must be taken account of in the design, especially of the absorption stage.

7.3.3. *Operating Principles of Absorption Apparatus*

The thermodynamic design of an absorption apparatus usually leads to a requirement for a certain number of theoretical plates needed to produce the required quality of the purified gas. This can only be achieved by use of the countercurrent principle, whereby the two process streams pass through the apparatus in opposite directions. However, in this context it is immaterial whether the flow in the individual stages is countercurrent or cocurrent (see Fig. 69). The concentration change represented in the distribution diagram shows that the same degree of separation can be produced in both cases with the same number of theoretical stages. The essential difference is that in cocurrent flow, one stream (5) in a unit flows in the opposite direction to the principal flow direction. Therefore individual stages always represent separate units of apparatus. With countercurrent flow, the required number of theoretical stages can be achieved in a single column. In special cases, a semicontinuous mode of operation is used in which an absorption process takes place in each stage, and after a certain period of time the solvent is pumped in countercurrent flow to the next stage, with fresh solvent being added to the first stage.

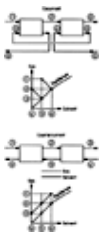

Figure 69. Internal flow in multistage co- and countercurrent processes

Representation of concentrations in McCabe – Thiele diagrams

The removal of large amounts of heat of absorption requires modifications of the apparatus (Fig. 70). In design (A) the heat removal takes place in the apparatus. Good results have been obtained using film absorbers or tray columns operating on this principle (see Section Absorption Apparatus). In the alternative version (B) the liquid is pumped through a recirculating system containing a cooler which removes the heat of absorption. If the heat of absorption is very large and the solution equilibrium is favorable, as in HCl absorption, water can be evaporated in the apparatus. The water vapor leaves the top of the apparatus and is condensed in a downstream condensation stage.

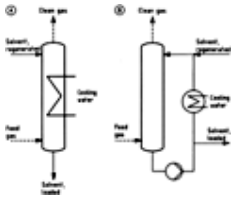

Figure 70. Absorption equipment with removal of heat of absorption

A) Removal takes place in the apparatus; B) Removal through recycling system

For certain purposes, it can be advantageous or necessary to link together different designs of apparatus or completely different processes (see Fig. 71). For high heats of absorption, as in the removal of ammonia from waste air, the stage into which the gas is fed (Fig. 71 A) should be operated in cocurrent flow with heat removal, and final purification should be carried out in countercurrent flow apparatus [593]. The flow sheet of Figure 71 B is used in, e.g., the absorption of HCl – SO_2 mixtures, where SO_2 has fairly low but nevertheless unacceptably high solubility in aqueous HCl. A stripping stage is then included in which SO_2 is removed by a stripping agent to give the required purity [594].

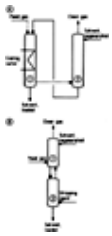

Figure 71. Combination of different process stages

A) Without stripping stage B) Including stripping stage

7.3.4. *Absorption Apparatus*

The apparatus described below are characterized by the largest possible interface between the gas and the liquid. In some types of apparatus, one phase is dispersed in the other. In another design, the liquid is spread over a surface to form a very thin film. This method has the advantage that only a small amount of energy is required to produce the interface, and the pressure drop is relatively low. Furthermore, the risk of entrainment of fine droplets is low, in contrast to processes in which one phase is dispersed in the other. In this case, droplet separation or demisting equipment must be included to prevent entrainment of the liquid phase in the gas stream.

Spray Absorbers. In spray absorbers (Fig. 72) the solvent is broken down into droplets by nozzles arranged in one or more planes. The momentum of the liquid is not high enough for hydraulic interactions to take place between the phases. Also, the absence of internals means that the pressure drop across the apparatus is almost zero and soiling is negligible. Spray absorbers are preferred for applications which require approximately one theoretical stage. These are mainly chemical absorption processes.

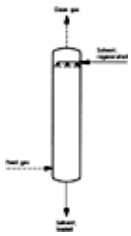

Figure 72. Spray absorber

Jet scrubbers [595] represent a further development of spray absorbers (see Fig. 73). The gas and liquid pass through the jet tube in cocurrent flow. The absorbing liquid enters the interaction space at high velocity. The gas stream is accelerated by the momentum of the liquid, and the pressure increase produced can in some cases overcome the flow resistance of the apparatus including the associated pipework. The momentum transfer also causes the liquid phase to break up into droplets, so that intensive mass and heat transfer occur. The downstream separation space is also used as receiving vessel for the pump. Generally, a demister is included at the gas outlet.

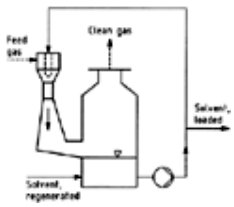

Figure 73. Jet scrubber

Jet scrubbers can be used in a multistage arrangement if several theoretical stages are necessary. This also gives the possibility of semicontinuous operation using a timed system for pumping the liquid from stage to stage. As the energy for dispersing the liquid is contained in the liquid itself as it enters the apparatus, the separating power is largely independent of fluctuations in the gas flow rate. If heat must be removed, a cooler can be included in the liquid pumping circuit.

Packed Columns. In packed columns the interface is produced by forming a film on the surface of internals (see Fig. 74 A). This enables a continuous countercurrent flow of gas and liquid to be produced. Random packings like Berl saddles, which provide a large surface for the liquid to cover, have been widely used. More recent packings are provided with holes and interior structures which not only give film formation but to some extent break the liquid up into droplets.

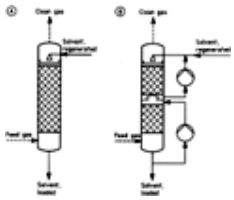

Figure 74. Packed column absorbers

A) For continuous countercurrent flow; B) With packings in series

For some time, packings first used in distillation technology have been widely used for absorption [597]. These consist of parallel toothed strips of various types of material. The teeth are inclined with respect to the axis of the column, and neighboring strips incline in opposite directions. Regular packings produce a considerably lower pressure drop per theoretical stage compared with random packings. Also, the minimum liquid loads are lower than with random packings. The separation effect only slightly depends of the column diameter. Both types of design require distribution devices that uniformly distribute the liquid over the cross section.

Because of the low liquid content, packed columns are not very suitable if fluctuations occur in the inlet gas velocities or in the gas concentrations. If it is not desired to use a different apparatus such as jet scrubbers or tray columns, another alternative is to use several packings in series, with recirculation (see Fig. 74 B). Fresh solvent is fed to the top of the column, and the product is withdrawn from the base. This design is relatively costly and should only be considered in special cases.

Tray columns used in absorption processes (see Fig. 75) are in principle identical to those used in distillation. A fundamental disadvantage is that typically the pressure drop in tray columns is high compared to packed columns. However, the following requirements or operating conditions can make the use of a tray column worthy of consideration:

1. In absorption processes where the loaded solvent is a useful product, high concentrations are often required. If the concentration of the component to be absorbed from the feed gas stream is low, this may mean that for mass balance reasons the necessary minimum loading of packed columns can no longer be maintained. In contrast the liquid holdup in tray columns can be reduced almost to zero.
2. With fluctuating gas flow rates or gas concentrations, a certain volume of liquid must be stored in the apparatus. This can be achieved by the use of bubble caps with chimneys. However, there is competition from other equipment, e.g., jet scrubbers.
3. A feature of the design is that contact between the gas and the liquid is concentrated on the trays. It is therefore possible to remove the heat of absorption liberated here (e.g., during the absorption of NO_x) by using tray columns fitted with cooling coils [598]. In this way, an absorption temperature that ensures optimum absorption can be maintained.

Figure 75. Tray column absorbers

Tray columns are now commercially available in all the usual construction materials, including highly corrosion-resistant versions made of graphite [599] and glass [600].

Film absorbers (Fig. 76) are used where the heat of absorption must be removed in the apparatus. The design is illustrated in Figure 76. In principle, they are heat exchangers through which the gas and liquid pass in cocurrent flow. Distributors must be included to ensure that each tube is fed with the same amount of liquid. The liquid forms a film on the inner surface of the tube, while the gas flows through the remaining interior space. The heat of absorption is removed via the outer surface of the tubes.

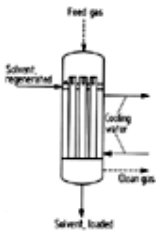

Figure 76. Film absorber

Film absorbers represent one theoretical stage at the most. They are therefore only used where absorption must take place at low temperature to achieve high product concentrations. In all cases, a conventional final purification stage for the gas stream must be included (e.g., a packed column).

7.3.5. *Selection of Equipment*

The various types of absorption equipment described above have preferred areas of use, depending on their characteristics. These characteristics, subdivided into process and operational parameters, are reviewed below. However, an unequivocal choice of equipment is not always possible.

Process Parameters. One important process parameter is the efficiency of separation required, expressed in terms of the number of theoretical stages. For demanding applications, packed columns or tray columns are used. Jet scrubbers are usually designed as only two or three stages, and spray absorbers in practice represent only one theoretical stage.

A further important consideration affecting the choice of apparatus is the question of heat removal. If it is possible for process reasons to remove heat by evaporation of a liquid, tray or packed columns can be used. If the temperature increase in the column leads to no phase change of the liquid, it is possible to use jet scrubbers or packed columns with exterior recirculation of the liquid through coolers. Film absorbers are also used. Tray columns having cooling devices in the trays are preferentially used in NO_x absorption.

If for reasons of mass balance the liquid flow rate is very low compared to an appreciably higher gas flow rate, the use of tray columns should be considered first. If packed columns, preferably with ordered packing, are considered, uniform distribution of the liquid over the cross section of the column is essential.

Operational Parameters. Intermittent operation with a variable total gas flow rate or variable concentrations of the substances to be absorbed should be considered first. The apparatus should be able to store a certain amount of liquid, and, the efficiency should vary as little as possible with the gas flow rate. Tray columns are highly suitable in this case. Jet scrubbers and packed columns with an external liquid circulation system can also be used, although practical problems with large number of stages must be borne in mind.

A further important operational parameter is the pressure drop across the apparatus. Jet scrubbers are a special case because they can create a pressure increase in the gas stream. The pressure drop across spray absorbers and packed columns is usually low, whereas tray columns always give rise to a high pressure drop.

7.4. Adsorption

Research into the use of porous solids for removing vapors from gas mixtures began with the construction of the first industrial plants for the production of activated prowdered carbon in 1900 to meet the demands of the sugar refining industry. The physicist J. DEWAR, who worked with activated carbon, silica gel, and aluminum oxide gel, was awarded a patent on the subject in 1905. The first process for the adsorptive removal of organic solvents from waste air and recovery by steam desorption was patented in 1916 for the Friedrich Bayer dyeing company. The use of adsorption for purifying town gas increased the importance of this gas separation technique in the 1920s. Since the 1960s, the development of adsorption catalysis and the introduction of new adsorbents have opened up further applications [601-603].

7.4.1. *Fundamentals*

In the application area considered here, adsorption means the adherence of molecules of gas or vapor (the adsorptive) to the inner surface of a highly porous solid (the adsorbent). The latter contains atoms or groups of atoms with free valencies (active centers) to which foreign molecules can become attached, and which are known as the adsorpt when in the adsorbed state. The complete boundary surface phase is known as the adsorbate (see Fig. 77). The reverse of adsorption, namely desorption, yields the desorpt and the desorption medium, which together are known as the desorbate.

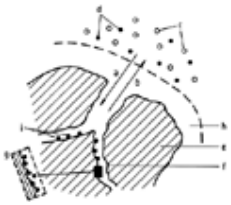

Figure 77. Adsorption and desorption processes at the boundary layer

a) Adsorption, exothermic; b) Desorption, endothermic; c) Carrier gas; d) Adsorptive; e) Adsorbent; f) Inner surface (sum of pore walls); g) Adsorbate; h) Boundary surface phase; i) Adsorpt

In addition to reversible physical sorption with purely physical bonding between adsorbent and adsorptive (van der Waals forces), there is also the phenomenon of chemisorption in which stronger interaction energies cause the adsorptive to undergo chemical change, which can lead to either reversible or irreversible attachment.

The adsorption capacity of an adsorbent for a given adsorptive at equilibrium at constant temperature depends on the size and structure of the inner surface of the adsorbent and is usually represented as a function of the concentration in the carrier gas, i.e., an adsorption isotherm $x = f\,(p/p_S)_T$. Figure 78 shows the equilibrium loading x as a function of the concentration in the carrier gas, expressed in terms of the ratio of partial pressure p to the saturation pressure p_S at constant temperature T. In large-scale industrial plants attainable breakthrough or additional loadings are considerably lower for the following reasons:

- Coadsorption of other compounds, e.g., water vapor
- The shape of mass-transfer zone (adsorption zone)

– Irreversible damage of the adsorbent
– Residual load due to incomplete desorption

Theoretical adsorption isotherms can be derived from various models [604], [605]:

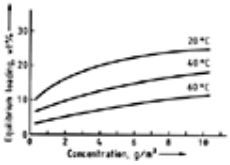

Figure 78. Adsorption isotherms for CS_2 at different temperatures on activated carbon

The Freundlich model

$$x = k \cdot p_i^n$$

applies to low partial pressures of the adsorptive.

The Langmuir model

$$x = \frac{x_{max} \cdot b \cdot p_i}{1 + b \cdot p_i}$$

applies to homogeneous adsorbent surface and monomolecular covering.

The Brunauer – Emmet – Teller (BET) model

$$\frac{p_i}{V(P_s - P_i)} = \frac{1}{V_{MC} \cdot C} + \frac{C-1}{V_{MC} \cdot C} \cdot \frac{p_i}{p_s}$$

applies to homogeneous surface, multilayer adsorption and capillary condensation.

The Dubinin – Raduskevic model

$$\log V = \log V_s - 2.3 \cdot \frac{RT}{\varepsilon_0 \beta} \cdot \left(\log \frac{p_s}{p_i} \right)^n$$

applies to potential theory for n = 1, 2, 3, e.g., n = 2 for adsorption of organic solvents to microporous activated carbon.

The following symbols are used in the above equations:

k, n = specific constants of the adsorptive
x = equilibrium loading of the adsorbed substance, g/100 g
x_{max} = saturation value of the isotherm for monomolecular covering, g/100 g
b = adsorptive coefficient
C = constant
p_i = partial pressure of the adsorptive, Pa
p_s = saturation vapor pressure of the adsorptive, Pa
V = adsorbed volume, mL
V_{MC} = adsorbed volume for monomolecular covering, mL
V_s = adsorbed volume at saturation pressure, mL
T = temperature, K
R = gas constant, J K^{-1} mol^{-1}
ϵ_0 = characteristic adsorption energy, J/kg
β = affinity coefficient

Figure 79 illustrates schematically the five basic types of adsorption isotherms for gases and vapors according to the IUPAC classification.

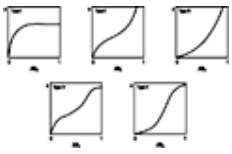

Figure 79. Basic types of adsorption isotherms

At adsorption temperatures below the critical temperature of the adsorptive, the pores of the adsorbent can become filled with liquid adsorpt. This capillary condensation increases the adsorption capacity of the adsorbent. Assuming that the

pores are cylindrical, it can be represented quantitatively by the Kelvin equation, which indicates that the degree of saturation is inversely proportional to the pore radius:

$$\ln \frac{p}{p_s} = -\frac{2\, V_M \sigma \cos\theta}{rRT}$$

The following additional symbols are used in the Kelvin equation:

p/p_S = relative equilibrium pressure

V_M = molar volume, m^3/mol

σ = surface tension adsorpt/adsorbent, N/m

θ = contact angle adsorpt/adsorbent

r = minimum pore radius, nm

As the transition from the fluid to the adsorbed phase is associated with release of energy, adsorption is an exothermic process. The enthalpy of adsorption liberated is a measure of the binding energy of the adsorpt, and this decreases as the quantity adsorbed increases. The energy required for solvent recovery is ca. 1.5 times the evaporation energy, resulting in effective temperature increases of 10 – 20 °C. The energy liberated during the complete adsorption process is known as the integral enthalpy of adsorption. The differential enthalpy of adsorption represents thermal processes at different stages of the adsorption process.

Because of the poor mechanical properties of most adsorbents, adsorption is usually carried out in a fixed bed, and the adsorption zone (mass-transfer zone) migrates through the adsorber in the same direction as the gas flow (Fig. 80). Below this zone the adsorbent is fully charged, and above it the adsorbent contains only the residual loading from the previous adsorption cycle. Within the adsorption zone, mass transfer takes place until equilibrium between the adsorbed phase (adsorpt) and the adsorptive concentration in the gas phase is reached. In adsorbents with wide pores, these complex processes are controlled less by mass transfer than is the case with adsorbents with narrow pores.

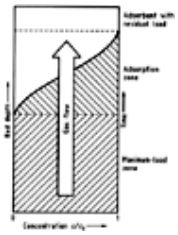

Figure 80. Loading profile in a fixed-bed adsorber

In practical applications, the waste gases to be purified often consist of a carrier gas containing mixtures of substances from which one or more components must be selectively removed. Separation can be achieved by utilizing the following effects:

1. Variation in the affinity of the adsorbent for the various gaseous components, this being apparent from the adsorption isotherms (equilibrium effect)
2. Differences in the critical molecular diameters of the adsorptives, enabling separation to be carried out if the effective pore diameter of the adsorbent (molecular sieve) is greater than that of the molecules of the component to be selectively adsorbed and smaller than those of the carrier gas (steric effect)
3. Different rates of diffusion of the adsorptives through the boundary layer and the pore structure to the active centers (kinetic effect)

Industrial adsorbents are regenerated when the breakthrough of the adsorptive occurs. Regeneration involves driving off the adsorpt as a gaseous or liquid phase (desorption), and drying and/or cooling treatment to restore the adsorbent to the optimum condition for reuse. Table 20 lists industrial regeneration methods. In waste-air purification the temperature-swing process is usually applied. The regeneration method and equipment have a major influence on the plant concept and investment and operating costs.

Table 20. Industrial regeneration methods

Method	Features	Typical applications
Temperature-swing desorption	desorption with slightly superheated steam, inert gas, or electrical energy (resistance heating of graphitic activated carbon or induction heating) up to 300 °C followed by cooling to adsorption temperature, if necessary after drying	purification of process gas, solvent recovery
Pressure-swing desorption	desorption by total pressure reduction between adsoption and the desorption steps, sometimes combined with purge gas feeding	separation of permanent gases, e.g., O_2 separation from air; CH_4 separation from

		water, hydrogen, or biogas
Extraction	a. with liquid solvents which are subsequently desorbed by the temperature-swing process (with steam) b. with NaOH followed by water wash and drying c. with supercritical CO_2	a. extraction of sulfur with carbon dioxide b. phenol-contaminated activated carbon from water purification c. activated carbon loaded with organic compounds
Wet oxidation	heating of the loaded and water-saturated adsorbent under pressure (p > water vapor pressure) with air feeding	spent activated carbon from drinking-water treatment

As it is not possible to drive off the adsorpt completely at acceptable cost, a residual loading is left behind, reducing the capacity of the adsorbent. If side reactions occurring during adsorption or desorption change the chemical structure of the adsorpt such that the residual loading increases, it is still often possible to reactivate the adsorbent in a separate plant. The process conditions for this treatment correspond to those of activation, i.e., the adsorpt is removed by oxidation. For activated carbons, this involves a selective gas treatment of the adsorbents with water vapor or CO_2 at 700 – 900 °C without appreciably changing the carbon structure. Oxidic adsorbents are regenerated with hot air.

7.4.2. *Industrial Adsorbents*

The adsorbents used in gas technology are not usually finely divided, but are highly porous granules or fibers with a large inner surface on which the foreign molecules become concentrated. Depending on the raw material and method of manufacture, they have different adsorption properties, which can be classified according to the following criteria:

The extent of the inner surface formed by all the pore walls is the main factor determining the adsorption capacity. It may be determined by the method of DIN 66 131 (BET method) [606], which is based on the adsorption isotherm for nitrogen at its boiling point and the space required by nitrogen molecules for monomolecular occupancy. The method is valid only for isotherms of types II and IV (Fig. 79).

The pore structure is important for the diffusion of the adsorptive to the active centers of the inner surface. Unlike molecular sieves, in which the inner cavities are linked by pores of uniform diameter, other adsorbents have a wide spectrum of pore size (Fig. 81). The pore radii (assuming a cylindrical shape) determined by different methods depending on the pore diameter (macropores: mercury intrusion porosimetry; mesopores and micropores: Kelvin equation) can extend over several orders of magnitude, and are classified according to the IUPAC Standard into

macropores	$d > 50$ nm
mesopores	$50 > d > 2$ nm
micropores	$2 > d > 0.4$ nm
submicropores	$d < 0.4$ nm.

The pore radii distribution of meso- and micropores is normally determined by evaluating the desorption profile of an isotherm, and that of macropores by mercury intrusion porosimetry.

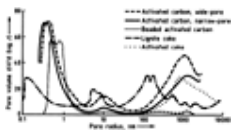

Figure 81. Differential pore distribution of carbonaceous adsorbents

The porosity is the pore volume as a fraction of the adsorbent volume, or per gram of adsorbent.

The adsorption characteristic of an adsorbent is revealed by the isotherms obtained with different adsorptives. It yields important information on the concentration dependence of the adsorption properties and the establishment of equilibrium.

The hydrophobicity or hydrophilicity of an adsorbent often determines whether it can be used in waste-gas purification, because water vapor is usually present at a considerably higher concentration than the substances to be removed. Figure 82 shows the different adsorption behaviors of some important adsorbents for water and toluene as typical examples for polar and nonpolar adsorptives.

Figure 82. Adsorption isotherms for water (A) and toluene (B)

a) Activated carbon; b) Activated alumina; c) Molecular sieve, hydrophilic; d) Silica gel; e) Zeolite, hydrophobic; f) Polymeric resin

The selectivity of an adsorbent can be a result of chemical effects (hydrophobicity, hydrophilicity) or steric effects (similarly shaped pore openings through which only molecules smaller than a particular diameter are able to penetrate).

Catalytic properties resulting from the chemical constitution of the inner surface may enable certain tasks to be performed, but can also lead to undesired side reactions and irreversible formation of chemical bonds (oxidation, polymerization, hydrolysis).

The industrial adsorbents can be divided into the categories shown in Table 21 in accordance with the raw materials and treatment methods used in their production.

Table 21. Industrial adsorbents

Adsorbent	Specific surface, $m^2\ g^{-1}$	Specific heat, $J\ kg^{-1}\ K^{-1}$	Compacted apparent density, $kg\ m^{-3}$	Typical applications
Activated carbons				
fine-pored	1000 – 1200	850	400 – 500	waste-air treatment, deodorization, separation of low-boiling substances
medium-pored	1200 – 1400	850	350 – 450	solvent recovery, adsorption of medium-boiling substances
wide-pored	1000 – 1500	850	300 – 400	adsorption of high-boiling substances
Activated coke	<400	850	500 – 600	separation of dioxines, furanes, SO_x and NO_x
Carbon molecular sieves	<100	1000	620	nitrogen and oxygen separation, methane adsorption from biogas
Aluminum oxide	100 – 400	920	700 – 800	desulfurization of gases
Silica gel				
fine-pored	250 – 350	920	700 – 800	deodorization, gas dehydration
wide-pored	600 – 850	920	400 – 800	adsorption of hydrocarbons
Molecular sieve				
zeolites				
hydrophilic	500 – 1000	900	600 – 800	gas dehydration, fractionation of hydrocarbons
hydrophobic	500 – 700	900	400 – 500	waste-air treatment, solvent recovery, solvent upgrading
Polymer adsorbents (styrene – divinylbenzene resins)	1000 – 1500	350	300 – 400	solvent recovery, gasoline recovery

The activated carbons (⟶ Carbon – Activated Carbon) are hydrophobic adsorbents with a wide range of specific surface areas, pore structures, and pore volumes, over 90 % of which consist of carbon with a graphitic crystal structure. They are produced in the form of shaped, powdered, and fibrous activated carbon from raw materials of vegetable origin by low temperature carbonization (primary carbon), grinding, shaping, and activation, and play a major role in the adsorptive purification of waste gas.

Activated coke is an adsorbent produced from finely hard coal preoxidized with air, shaped using pitch as a binding agent, and activated by treatment with gas (CO_2, H_2O). The inner specific surface area, pore volume, and porosity so produced do not reach the figures typical of activated carbons.

Coke produced by the low temperature carbonization of lignite from the Rhine region in a hearth type furnace contains up to 87 % carbon. Its adsorptive and catalytic properties enable the gaseous impurities Hg, SO_2, HCl, HF, Cl_2, NO_x, dioxins and furans to be efficiently removed. This cheap material is not regenerated and is used as a crushed or powdered adsorbent to remove small quantities of impurities.

Carbon molecular sieves have micropores with diameters of 0.3 – 1.0 nm. These very narrow pores mean that these adsorbents cannot be used for emission reduction.

Activated aluminum oxide (> 99 % Al_2O_3), an oxidic hydrophilic adsorbent, is mainly used for removing sulfur compounds.

Molecular sieve zeolites are crystalline aluminum silicates with an SiO_2 – Al_2O_3 lattice. They contain cavities linked by pores of uniform diameter, so that their adsorption capacity combined with their steric separating effect can be used in gas purification. If a high proportion of Al_2O_3 is replaced by SiO_2, this gives a hydrophobic zeolite (trade name: Wessalith DAY) [607] suitable for selectively adsorbing organic components and mercury from waste gases with a high moisture content, and is therefore known as “white activated carbon”. As it is a mineral, it can be regenerated at high temperatures with air.

Spheroidal adsorber resins made of nonpolar macroporous polymers are produced by polymerization of styrene in the presence of a cross-linking agent, and have the advantage of ease of regeneration at low temperatures when used for the purification of solvent-containing waste air. They also have a low affinity for water vapor.

In silica gels, the SiO_2 particles form a continuous crystalline structure containing pores. Silica gels with small pores are used as drying agents, while the wide-pore types are suitable for removing hydrocarbons from waste gases.

The properties of some adsorbents can be modified by impregnation with organic or inorganic substances (Table 22) to give additional separation effects (e.g., the adsorption of H_2S by activated carbon impregnated with potassium iodide).

Table 22. Reactions of impregnated activated carbon with contaminants of a gas stream

Contaminant	Impregnating agent	Active substance	Conversion
Ammonia	copper	$CuSO_4 \cdot 5\,H_2O$	$CuSO_4 \cdot 5\,H_2O + 4\,NH_3 \longrightarrow [Cu(NH_3)_4]SO_4 + 5\,H_2O$
Arsenic hydride	silver, zinc	Ag, Ag_2O, ZnO, Na_2ZnO_2	$2\,AsH_3 + 3\,O_2 \longrightarrow As_2O_3 + 3\,H_2O$
Hydrogen cyanide	copper	CuO, Cu_2O	$2\,HCN + Cu_2O \longrightarrow 2\,CuCN + H_2O$
	zinc	ZnO, Na_2ZnO_2	$2\,HCN + ZnO \longrightarrow Zn(CN)_2 + H_2O$
Phosgene	copper	CuO, Cu_2O	$COCl_2 + CuO \longrightarrow CuCl_2 + CO_2$
	zinc	ZnO, Na_2ZnO_2	$COCl_2 + ZnO \longrightarrow ZnCl_2 + CO_2$
Mercury	sulfur	S_8	$Hg + S \longrightarrow HgS$

The following are the principal criteria used when selecting an adsorbent for waste-gas purification:

1. High selectivity with fast kinetics for the component to be removed
2. High adsorption capacity even at low concentrations
3. Ease of desorption
4. Stability towards temperature change and chemical attack
5. Hydrophobicity and attrition resistance

The design of an adsorption plant for waste-gas treatment depends not only on the type of adsorbent but also on some properties of the substances to be removed. The design engineer should pay attention to the aspects listed in Table 23.

Table 23. Properties of substances to be removed and associated design measures

Pollutant properties	Design aspects
Molecular mass	adsorption capacity generally increases with increasing molecular mass
Boiling point or boiling range	adsorbability increases with increasing boiling point
Wetting ability	the surface tension of the component to be adsorbed must allow good adsorbent wetting
Molecular structure	its significance depends on the pore structure of the adsorbent
Critical temperature	increasing the adsorption capacity by capillary condensation is only possible if the operating temperature is below the critical temperature of the adsorptive (condensation cannot take place at supercritical temperatures)
Desorbability	frequently the equipment and process steps needed for desorption determine the system configuration and economics

Water solubility	crucial factor in steam desorption, may entail process steps for recovering a water-free desorpt or a hot-gas desorption process
Corrosion behavior	governs material selection
Explosion limits	determine safety measures

7.4.3. *Design Criteria for Adsorbers*

To enable the gas that is to be purified to pass through the adsorbent bed as effectively as possible, the latter must be arranged as a fixed bed, moving bed, or other arrangement in the adsorber. The dimensions of the equipment are determined by the maximum permissible flow rate of the waste gas and bed depth of the adsorbent. For adsorbents in powder form, fluidized beds or entrained flow systems give intensive contact with the adsorptive. Most adsorbers are of the fixed-bed type, because of the poor mechanical stability of the adsorbents. The gas velocity is about 0.5 m/s with respect to the free cross section and should be well below the velocity at which the granules in the upper region of the bed will be in motion. With most activated carbons, this velocity is 0.8 m/s for the carrier gas air at ca. 20 °C. In some cases, the adsorption and/or reaction rates make the residence time decisive, which can lead to larger adsorber cross sections or/and bed depth.

Flow resistance figures for various granular adsorbents under normal gas velocities are shown in Figure 83. They depend on the properties of the fluid phase (density, viscosity, velocity) and of the adsorbent bed (grain size, grain shape, surface area, voids fraction). The calculation method often quoted in the literature [608] also contains an experimentally determined resistance number. Uniform flow through the bed not only gives optimum utilization of its adsorption capacity, but tends to give efficient heat removal, preventing hot spots which can lead to spontaneous ignition of carbonaceous adsorbents. Equipment for monitoring CO in the impure and purified gas serves as a warning device. Bed temperature measurements provide inadequate warning owing to the poor heat-transfer properties.

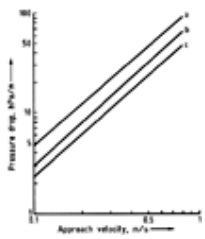

Figure 83. Pressure drop for various adsorbents (p = 1 bar abs.; T = 20 °C)

a) Activated carbon cylinders, 3 mm; b) Activated carbon cylinders, 4 mm; c) Zeolite cylinders, 4 mm

For a given diameter, the bed depth of the adsorber depends on the amount of adsorption medium charged. This is usually determined from the equilibrium loading, after subtraction of factors that reduce the capacity such as residual loading and irreversible damage. This lowers the operating capacity of the adsorbent to a fraction of the total loading at equilibrium. In solvent recovery by activated carbons, this fraction is ca. 30 %. Moreover, a minimum bed depth must be provided to accommodate movement of the adsorption zone. This also ensures minimum pressure drop required for uniform gas distribution.

The effective bed depth can be derived from the profile of the adsorption zone [604], [605]. As the adsorption zone usually migrates through the bed at constant velocity in the same direction as the gas flow, breakthrough curves determined at the same flow rate in experimental adsorbers can be transferred to large-scale industrial plants. In breakthrough curves, the ratio of the concentration of the adsorptive at the exit of the adsorber to that at the inlet (c/c_0) is plotted as a function of adsorption time t. As shown in Figure 84, for bed depths h_1 and h_2, the half-value times t_{h_1} and t_{h_2} (where c/c_0 = 0.5) are determined. This gives the migration rate u of the adsorption zone:

$$u = \frac{h_2 - h_1}{t_{h_2} - t_{h_1}}$$

Multiplication by the time interval $t_2 - t_1$ gives the length h_z of the adsorption zone:

$$h_z = u(t_2 - t_1)$$

Assuming a symmetrical profile of the adsorption zone, the effective height h_{eff} of the bed is given by $h_{eff} = h - 0.5\, h_z$. The amount of adsorbate x (t) taken up by the adsorbent from the carrier gas $\dot{V}$ up to time t is given by

$$x(t) = \frac{\dot{V}}{m}\int_0^t (c_0 - c_t)\mathrm{d}t$$

Figure 84. Determination of effective bed depth from breakthrough curves

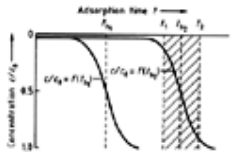

Other methods are described in the literature [609], [610] for dimensioning an adsorber under given conditions from mass and energy balances.

The design of the adsorber calculated from adsorption criteria must also be suitable for the regeneration conditions, and this must be investigated separately.

7.4.4. *Adsorber Designs*

Specific types of adsorber have been developed for the various tasks of waste-gas purification [611]. These can be divided into fixed-bed and moving-bed equipment, and special designs (Fig. 85).

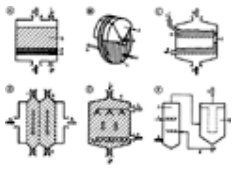

Figure 85. Types of adsorber

A) Fixed-bed adsorber

a) Raw gas; b) Clean gas; c) Desorption medium; drying and cooling medium; d) Desorption medium and desorpt; e) Carrier grid; f) Intermediate layer; g) Adsorbent packing

B) Rotating adsorber

a) Raw gas; b) Clean gas; c) Adsorbent; d) Desorption segment; e) Cooling segment

C) Fluidized-bed adsorber

a) Raw gas; b) Clean gas; c) Adsorbent, regenerated; d) Adsorbent, loaded; e) Fluidized bed; f) Down pipe

D) Moving-bed adsorber

a) Raw gas; b) Clean gas; c) Adsorbent, unloaded; d) Adsorbent, loaded; e) Moving bed

E) Countercurrent adsorber

a) Raw gas; b) Clean gas; c) Adsorbent, unloaded; d) Adsorbent, loaded; e) Moving ded

F) Entrained-flow adsorber

a) Raw gas; b) Clean gas; c) Adsorbent, unloaded; d) Adsorbent, partially loaded; e) Adsorbent, transferred outward; f) Tubular reactor; g) Woven filter

Because of the poor mechanical properties of most adsorbents, the fixed-bed design (Fig. 85 A) is usually used in industrial plants. This consists of a random packing on a carrier grid covered with perforated sheet, woven steel wire, or a perforated ceramic plate. Vertical or horizontal adsorbers may be preferred depending on the type of waste gas to be treated and the space available on site. In the vertical designs, the adsorbent is sometimes contained in a ring-shaped shaft to save floor space. Even for longer service life of the adsorbent, the apparatus should include equipment for rapid charging and emptying. If desorption with water vapor is used, the adsorber must be designed as a pressure vessel.

Moving-bed adsorbers (Fig. 85 D) are almost always designed vertically to save floor space and to facilitate the demands made on the mechanical charging and discharging equipment. Virgin or regenerated adsorbent is fed from above while the same amount of exhausted material is removed from below. Locks are provided if necessary at both feed and discharge points. Slotted walls or screens are provided at the gas inlet and outlet sides to guide the adsorbent bed. The bed can be subdivided into several layers arranged in series, containing different adsorbents with different functions which can be moved at different rates. The mechanism for removing the adsorbents must ensure that they move downwards as a single mass, including the edge regions. In another design, a circular moving bed is used through which the gas flows from the inside to the outside or vice versa. In the design illustrated in Figure 85 E, the waste gas is routed through the adsorber in countercurrent flow to the moving bed.

The fluidized-bed technique (Fig. 85 C) requires hard adsorbent granules with high abrasion resistance, which has considerably restricted the use of this method. The apparatus contains a perforated sheet or nozzle tray on which the adsorbent is maintained in a fluidized state by the waste gas entering from below. The adsorbent passes from the inlet to the outlet perpendicular to the gas flow. The purification effect can be increased by using several fluidized beds in series. Like scrubbing columns, they are equipped with inlet and outlet weirs and are connected by down tubes. The comparatively small amount of adsorbent makes the removal of some gaseous components difficult, and the additional loading capacity of the adsorbent is lower than that obtainable with fixed beds. An advantage is the continuous operation of the adsorption and regeneration stages.

Continuous operation is also achieved by rotating adsorbers (Fig. 85 B). The rotor has a large specific surface area formed by granular activated carbon or ceramic trays which carries a layer of finely divided activated carbon or fibrous carbon. The rotor rotates round a horizontal or vertical axis inside a housing, passing through the adsorption, desorption, and cooling zones successively. The gas can pass through the rotor axially or radially. Because desorption is carried out with hot air at a flow rate that is only a fraction of that of the waste gas, the adsorber is not suitable for the final purification stage, but

only for a preliminary increase in concentration.

In the entrained-flow system (Fig. 85 F), the adsorbent powder is carried by the flow of waste gas. The main function of the apparatus is to provide a mixing space, while the adsorption of the impurities mainly takes place on a woven filter downstream.

Some special designs of adsorbers include adsorbent mats made of activated carbon fibers, filter elements in which activated carbon granules are bonded together, or powdered activated carbon bonded to highly porous carriers such as polyurethane foam.

7.4.5. *Plant Concepts*

Apart from those plants intended purely for waste-gas separation, in which no regeneration is provided for or possible because of the long service life of the adsorbent (several months), the concept of an adsorption plant for waste-gas purification is very much influenced by the regeneration method. Regeneration can be achieved in various ways (see Table 20).

Steam desorption, which exploits the temperature dependence of the adsorption equilibrium and the high heat of condensation of the steam, is the most widely used process for solvent recovery (Fig. 86). The adsorber is heated with slightly superheated steam at ca. 150 °C, which also acts as a purging medium. The condition and consumption of steam (3 – 5 kg/kg solvent) depend mainly on the type of adsorptive and the desired residual loading of the adsorbent, according to the required purity of the waste gas. The desorbate produced, consisting of solvent vapors, steam, and condensate, is condensed, cooled, and separated. If the solvent is immiscible with water, the separation equipment used is based on density difference. As a rule of thumb the higher the mutual solubility of solvent and water, the more complex separation process, and a rectification column may even be required. An overview of methods used for processing the desorbate is given in Figure 87. After desorption, the adsorber is dried and cooled with purified waste gas or clean air to the temperature of the adsorption step. As the cooling gas contains small quantities of solvent it is passed through the operating adsorber to prevent emission peaks. This alternating adsorption/desorption system means that there must be at least two adsorbers to ensure continuity of the waste-gas purification process.

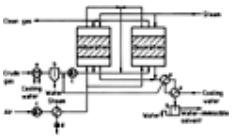

Figure 86. Flow sheet for activated carbon plant with steam desorption

a) Cooler; b) Separator; c) Fan; d) Steam trap; e) Condenser

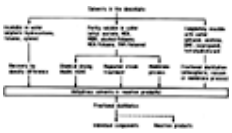

Figure 87. Methods of processing desorbates

In special cases, an inert gas or air is used to heat the adsorber directly. This can be carried out in three ways (Fig. 88). The closed-loop system (Fig. 88 A) is advantageous if enough of the desorbed substances can be condensed to ensure that their residual partial pressures do not cause an unacceptably high residual loading. Otherwise, if oxygen-free waste gases are being purified or a noncombustible adsorbent is used, regeneration can be carried out in an open loop (Fig. 88 B). Here a certain amount of the purified waste gas is fed through the adsorber being cooled and that being heated, and, after partial condensation of the desorbed substances, passed into the impure gas and thus into the adsorber during the adsorption cycle. If the desorbed substances cannot be adequately condensed, or if they consist of a mixture of substances whose recovery is not feasible, a recirculation system cannot be used, and the desorpt is then treated by thermal or catalytic combustion. In this case, the adsorption plant functions as a concentration plant, using preheated air for desorption (Fig. 88 C).

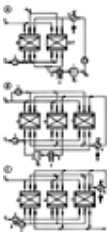

Figure 88. Plant design concepts with hot-gas desorption

A) Closed-loop system

B) Open-loop system

C) Plant using preheated air

a) Raw gas; b) Clean gas; c) Blower; d) Adsorption (adsorber); e) Desorption (adsorber); f) Cooling (adsorber); g) Condenser/cooler; h) Cooling, water; i) Separator; j) Desorpt; k) Heater; l) Heating steam; m) Purge gas, inert; n) Inert gas, makeup; o) Air

▷ ◁ = Open valve; ▸ ◂ = Closed valve

Desorption by pressure reduction (pressure-swing process) requires a significant pressure difference between the adsorption and the desorption steps so that desorption has to be carried out under vacuum combined with a small purge gas stream which carries the desorpt out of the adsorber.

In a few special cases, e.g., sulfur-loaded carbon treated with CS_2, desorption is carried out by solvent extraction, after which the solvent is desorbed by steam.

Plant sizes range from compact plants with throughputs of a few hundred m^3/h to large plants in which waste gas can be purified at a rate of several hundred thousand m^3/h.

7.4.6. *Adsorption Processes*

The availability of several adsorbents with various physical, chemical, and mechanical properties in association with the wide range of waste-gas purification tasks has led to the development of a variety of industrial techniques.

In the oldest application — the recovery of organic solvents from waste air — the most widely used processes are based on fixed beds of activated carbon (see Fig. 85). The waste air is extracted from hooded production areas conditioned by cooling, partial condensation of the water vapor and heating to reduce the relative humidity in order to minimize coadsorption of water and then fed at ca. 40 °C to the adsorber on loading duty. At an approach velocity of ca. 0.5 m/s it passes upwards through the activated carbon bed, which has a height of 0.8 – 1.5 m. The granular adsorbent takes up the solvent vapors, and the purified waste air passes out of the adsorber. After the solvent breakthrough limit has been reached, the adsorption process is interrupted, after the freshly regenerated adsorber has taken over the purification of the waste air. The solvent is then desorbed by passing superheated steam downwards through the activated carbon. The steam heats the bed to > 100 °C and carries away the desorpt. The desorbate is condensed, cooled, and passed to the gravity separator. This is a settling vessel in which the mutually insoluble organic and aqueous phases separate as a result of the density difference. Approximately 4 kg steam/kg solvent is required for this process step, which accounts for most of the operating costs. The subsequent drying and cooling steps are carried out with purified waste air or clean air, which are then cleaned by passing them through the adsorber on operating duty. If the recovered solvent is miscible with water, fractional distillation may be required.

If the time required for regeneration is appreciably shorter than that for adsorption, higher loading of the activated carbon can be achieved. In this case, the regenerated adsorber can be connected after the operating adsorber (two-stage adsorption). The waste air passes through the two vessels connected in series, the first adsorber being charged beyond breakthrough.

In activated carbon plants using hot gas desorption, the operation must be performed by purging the loaded adsorbent with an inert gas in a closed circuit, consisting of a cooler-condenser and separator, a heater, and a drying stage for the regeneration gas (see Fig. 88 A).

With the hydrophobic zeolites used for removing solvents from waste air, dry desorption with hot air can be used, preferably in a closed recirculating system, provided that partial condensation of the solvent is sufficient for reuse of the waste air in, for example, a drying process.

Polymer adsorber resins loaded with solvents can be regenerated at temperatures as low as 80 °C, either by steam or hot gas [612]. They are typically employed in moving-bed applications (Fig. 89, and permit continuous operation of fluidized bed adsorbers in combination with moving-bed regenerators (e.g., Polyad process) [612].

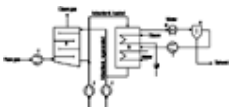

Figure 89. Continuous adsorption system for recovery of solvent from exhaust air

a) Fluidized-bed adsorber; b) Moving-bed regenerator; c) Raw-gas blower; d) Condenser; e) Separator; f) Blowers

The Kontisorbon process for continuous adsorption and desorption (Fig. 90) [613] uses attrition-resistant activated carbon beads of diameter 0.7 mm in a multistage fluidized bed, maintained in the fluidized state by waste air at a velocity of 0.8 – 1.2 m/s. The carbon is regenerated by feeding it into a heat exchanger with vertical tubes underneath the adsorber. Indirect steam heating in combination with nitrogen purging drives off the solvent, which is partially recovered as condensate. The remaining solvent is adsorbed from the nitrogen purge cycle by the activated carbon above the desorber which has passed through the fluidized bed. Cooling of the activated carbon after desorption is also indirect. It is then pneumatically returned by air lift to the upper perforated tray of the fluidized bed.

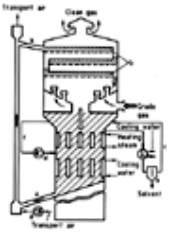

Figure 90. Kontisorbon process

a) Adsorbent, regenerated; b) Fluidized bed; c) Inert gas; d) Fan; e) Separator

Rotary adsorbers also enable the process to be operated continuously. The rotor is packed with granular activated carbon or consists of a ceramic honeycomb structure coated with fine-grained activated carbon or fibrous carbon. It rotates round its horizontal axis in a housing at 1 – 2 rpm passing through an adsorption zone, a desorption zone, and a cooling zone in succession (Fig. 85 B). The waste air passes through the adsorption segment at 1 – 2 m/s, the pressure drop being only 2 – 3 mbar [614]. Desorption can be carried out with steam, but preferably by hot air at a flow rate of 2 – 20 % of the waste gas flow rate. This passes out of the plant as higher concentrated secondary waste air and requires purification.

The plant concept to clean waste air from e.g. gasoline tanks with high concentration of organic vapors up to saturation involves a scrubber or a membrane unit combined with a downstream pressure-swing adsorption plant.

For the removal of low-boiling and medium-boiling hydrocarbon vapor mixtures, pressure-swing adsorption with activated carbon is followed by an absorption stage to treat the compressed gases from vacuum desorption with gasoline as scrubbing liquid, which can be returned to the tank farm.

In another option the hydrocarbon-saturated air is freed from the bulk of its vapor load by a membrane permeation stage operated at a pressure of 3 – 4 bar. The recovery efficiency depends on the specific permeation rates of the organic compounds. The resulting concentrated mixture of organics can then be condensed.

For final abatement of the hydrocarbons the exhaust air leaving the membrane system enters a pressure-swing adsorption unit, which consists of two alternately operating parallel lines, each equipped with two fixed-bed adsorbers in series. The first is packed with activated carbon to retain higher boiling hydrocarbons while the downstream apparatus contains molecular sieves for adsorbing low-boiling organics. The desorbate air stream is returned to the front end of the treatment system and mixed with the raw gas.

Sulfur compounds are removed from waste gases using either activated carbons or activated aluminum oxide as the adsorbent or adsorption catalyst. In the Sulfosorbon process [615] for purifying waste air containing H_2S and CS_2 from the viscose process, each adsorber is usually packed with two layers of activated carbon having two different effects. The lower part of the bed consists of wide-pore carbon impregnated with potassium iodine, which catalyzes oxidation of the H_2S and adsorbs the elemental sulfur formed. Oxidation to sulfuric acid is largely prevented. The CS_2 is removed mainly in the upper bed of narrow-pore activated carbon. As soon as the CS_2 concentration in the purified waste air exceeds the limit value, a regenerated adsorber takes over the purification process, and after purging with an inert gas, desorption of the CS_2 by steam at 110 – 130 °C commences. The organic and aqueous phases are separated in a gravity separator. Regeneration of the lower, sulfur-loaded bed is required only in intervals of several months and comprises the following process steps: water washing to remove sulfuric acid and/or ammonium sulfate formed as side products, extraction of the elemental sulfur with CS_2, separation of sulfur and CS_2 by distillation, desorption of CS_2 with steam, and drying and cooling the adsorber with air.

The Thiocarb process [616] developed for the same purpose uses a single bed of wide-pore activated carbon with a very low metal content (to suppress the catalytic formation of H_2SO_4). Any acid which is formed despite this is neutralized by adding NH_3. The regeneration process is as described above.

The Sulfacid process [617] can be used with SO_2-containing, dust-free process waste gases using activated carbon as an adsorption catalyst to purify the gas in a fixed bed, provided the O_2 content is sufficient to convert SO_2 and water to sulfuric acid. Washing with water yields dilute H_2SO_4 (10 – 15 %), which can be concentrated to ca. 30 % in combination with the waste-gas cooling in the venturi scrubber. For SO_2 concentrations of 0.1 – 1.0 vol % the removal efficiency is 90 – 95 %.

Desulfurization and removal of nitrogen oxides from combustion waste gases is carried out in the BF – Uhde process [618] by adsorption catalysis at 120 – 150 °C on 6 mm activated coke in a two-stage moving-bed reactor with transverse gas flow (Fig. 91). As well as desulfurization, the first passage of the gas also removes HCl and HF. Ammonia is then added to convert the nitrogen oxides to N_2 and H_2O by selective catalysis in the second passage of the gas, through the upper part of the reactor. Ammonium sulfate is formed in side reactions. Regeneration is carried out by indirectly heating the dedusted activated coke to ca. 450 °C, and a gas with a high SO_2 content is formed by reaction of the sulfuric acid with the carbon to give a SO_2 rich gas which can be converted to sulfuric acid, elemental sulfur or liquid SO_2.

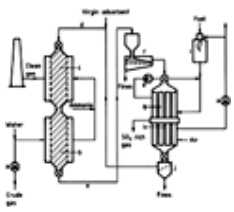

Figure 91. BF – Uhde process

a) Fan; b) Desulfurization; c) NO_x removal; d) Adsorbent, regenerated; e) Adsorbent, loaded; f) Air classifier; g) Desorber; h) Cooler; i) Combustion chamber; j) Sieve

Both activated carbon and activated aluminum oxide catalyze the Claus reaction

$$2\,H_2S + SO_2 \rightleftharpoons 3S + 2\,H_2O$$

at ca. 130 °C and readily adsorb the sulfur formed. This is the basis of the Sulfreen process [619] for desulfurizing Claus waste gases. In the single-stage process (Fig. 92), a fixed bed of activated aluminum oxide gives removal efficiencies of 80 – 93 %, depending on the inlet concentration of H_2S and SO_2. During desorption by heated clean gas at ca. 300 °C, most of the sulfur driven off can be condensed from the recirculating gas stream. Clean gas is also used for cooling to the adsorption temperature. For higher efficiencies, the two-stage variant (the Carbosulfreen process) includes an activated carbon bed in which H_2S reacts with O_2, and the elemental sulfur formed is adsorbed. Preliminary catalysis by titanium dioxide allows COS and CS_2 to be hydrolyzed to H_2S (Hydrosulfreen process).

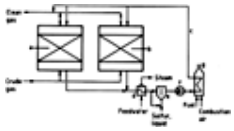

Figure 92. Sulfreen process

a) Adsorption (adsorber); b) Regeneration (adsorber); c) Regeneration gas; d) Condenser/cooler; e) Separator; f) Fan; g) Heater

Another important application of adsorption is the removal of heavy metals (in particular mercury vapor) as well as dioxins and furans from the flue gases from incineration plants for municipal waste, hazardous waste, and sewage sludge. Adsorbents used include activated carbon, activated coke, hearth-furnace coke, and hydrophobic zeolites. The efficiency of removal of mercury by activated carbon or zeolites can be considerably improved by impregnation with sulfur or sulfuric acid, and the same effect is produced by SO_2-containing flue gases. The long residence times required are a disadvantage, requiring the use of high-volume vertical adsorbers [620]. The exhausted adsorbent is removed discontinuously or continuously, and is often used as a fuel in combustion plants. The pollutants are recovered by means of dedusting filters or washing processes (Hg, $HgCl_2$, SO_2, HCl, and HF) or are incinerated (dioxins and furans). A further alternative for oxidic adsorbents is thermal regeneration, which can be more cost-effective than combustion in the case of expensive adsorbents [621]. If the space for large fixed-bed filters is not available, powdered hearth-furnace coke can be added to the wash suspension for desulfurization. High removal efficiencies have been obtained using powdered hearth furnace coke and lime in a pilot plant consisting of a fluidized-bed reactor and fabric filter in an hazardous-waste incineration plant (Fig. 93) [622]. The results correspond to those obtained on a full-scale plant for treating flue gases from municipal waste incineration operating on almost the same principle [623].

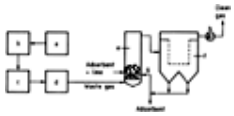

Figure 93. Circulating fluidized bed for removal of mercury, dioxins, and furans as tail end cleaning stage.

a) Waste incineration plant; b) Electrostatic precipitator; c) Scrubbers for $HgCl_2$, HCl, HF, SO_2; d) DENOX; e) Fluidized bed; f) Fabric filter; g) Recirculation of adsorbent; h) Fan

Powdered carbonaceous adsorbents injected in entrained-flow systems and separated downstream by fabric filters are gaining increased significance in removing heavy metals, dioxins and furans from waste gases and hazardous waste incineration plants, steel production plants, recycling plants for Pb- and Zn-containing filter dust, sintering plants and other processes. They collect those substances from the off-gas streams at high selectivity in spite of their low concentrations [624-626]

Natural zeolites boast the advantage of greater safety in operation. This means that the safety equipment normally required, addition of limestone, extinguishing facilities, measuring and control devices etc. can be omitted. There is no self-ignition hazard [627].

Circulating fluidized bed pilot plant tests with natural zeolites have demonstrated that nitrogen oxides can be successfully converted to nitrogen and water by selective catalytic reduction.

7.4.7. *Recycling and Disposal of Spent Adsorbents*

Depending on the type of adsorbent, the following alternatives are available for recycling or disposal of spent adsorbents:

1. Reactivation of carbon-based and oxidic adsorbents with recovery of the adsorbed substances, if possible
2. Hydrogenation of carbon-based adsorbents and polymer adsorber resins, together with other organic wastes such as spent oils, paint sludge, plastics, and solvents to produce base stocks for the oil industry
3. Co-burning of carbon-based adsorbents in the combustion plant producing the flue gas
4. Co-burning of carbon-based adsorbents and polymer adsorber resins in hazardous-waste incineration plants;
5. Disposal of oxidic adsorbents

Figure 94 shows a reactivation plant for treating granular activated carbons from waste-gas cleaning and drinking-water treatment loaded with organics. Instead of a rotary kiln a multistage reactor or fluidized-bed reactor may be installed in which a reducing atmosphere is maintained. With carbon retention time in the kiln of ca. 15 min and 800 °C operating temperature, the gasification losses amount to ca. 5 % including losses due to mechanical and thermal stresses (undersize). During the heating step the reversibly adsorbed substances are desorbed. The irreversibly bound compounds are removed by selective gasification with water vapor and carbon dioxide, which to a certain extent changes the pore structure. The waste gases of the reactivation process are cleaned by thermal oxidation of the pollutants in the incinerator and passage through a downstream caustic scrubber. The adsorption power of oxidic adsorbents can be restored by controlled oxidation with air or oxygen fed to the adsorber at high temperature.

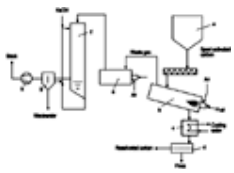

Figure 94. Flowsheet of a plant for the reactivation of activated carbon

a) Hopper; b) Rotary kiln c) Cooler; d) Screen; e) Incinerator; f) Scrubber; g) Separator; h) Blower

7.5. Biological Processes

7.5.1. *Overview and Classification*

In biological methods for purifying waste gas, the hazardous and/or odorous constituents are reacted or degraded biologically. This does not take place in the waste gas itself, but after the materials to be treated have been sorbed by solids or liquids in the biological purification equipment. This equipment can be classified according to the process used for conversion or reaction as follows:

1. Biofilters containing organic and inert materials
2. Bioscrubbers containing fixed or suspended microorganisms
3. Combined processes in which biofilter and bioscrubber units are arranged in series.

The selection of the process to be used depends on several criteria which must be considered in each individual case [628].

7.5.2. *Process Engineering Principles*

To correctly operate and optimize a process, biological, physical, and chemical effects have to be considered. The main process engineering task is to establish and maintain optimum living conditions for the microorganisms.

7.5.2.1. *Biological Conditions*

Microorganisms are so-called destruents, which break down organic materials such as carbohydrates, fats, and proteins into their basic components (CO_2, CH_4, NO_3^-, PO_4^{3-}, etc.). These components are then used by the microorganisms to form their own biomass. The required energy is derived from the biochemical oxidation (i.e., the degradation of the organic substances). Microbial degradation is therefore associated with an increase in the biomass. If the required carbon is derived from an organic substance, the term “heterotrophic habit” is used. Most gaseous substances are available to microorganisms, although their biodegradability properties can range from easy to very difficult, depending on their chemical structure.

Inorganic gaseous compounds such as ammonia and hydrogen sulfide can also be used by microorganisms as sources of energy or as nutrients. The carbon required for growth is derived from atmospheric carbon dioxide (carbon autotrophism) [629].

There are ca. 3500 naturally occurring microorganisms which, together with specialized lower life forms, make it possible to break down even difficultly available substances. The microorganisms are either sessile (i.e., attached to organic filter materials or inert packings) or suspended freely in a scrubbing liquid. In both cases, the gaseous substances must be absorbed in a liquid to make them available for degradation.

The following conditions must be fulfilled to enable the equipment to function:

1. The components of the waste gas must be biologically degradable
2. The waste gas must not contain substances toxic to microorganisms
3. The components of the waste gas must be soluble in water
4. The products formed by the microorganisms must not have a detrimental effect on the biological degradation process.

The conditions in the medium must be optimized for the microorganisms. In biological processes, the following factors affect growth:

Oxygen Demand. In most waste gases not derived from purely chemical processes, sufficient oxygen is present to enable processes based on fixed microorganisms to be used. In bioscrubbers which use suspended microorganisms, adequate amounts of oxygen must be added at the regeneration stage.

Water Demand. Living microorganisms contain 75 – 85 % water, which is needed to transport nutrients into and metabolic products out of the cells. Thus they have a water demand which can be met by the water in which they are suspended or by water-containing solid substrates. The actual quantity of water is not as important as its availability in the substrate or medium. A measure of this is the water activity a_w, which is the ratio of the vapor pressure of an aqueous solution to the vapor pressure of the gas dissolved in water. The water activity of pure water is $a_w = 1.0$. The water requirement of bacteria is approximately $a_w = 0.91$, of molds $a_w = 0.80$, and of osmophilic yeasts $a_w = 0.60$ [629].

Nutrient Demand. The approximate composition of a bacterial cell on a dry basis is 50 % carbon, 14 % nitrogen, 2 – 6 % phosphorus, and 1 % sulfur and other elements.

Carbon, nitrogen, and phosphorus are essential for cell growth and producing energy. As a rule, a well-balanced nutrient contains these elements in the approximate ratios (C : N : P) 100 : 5 : 1 (bioscrubbers) or 200 : 10 : 1 (biofilters).

Temperature. Three temperature ranges, in which different types of microorganisms are active, are of interest for biological waste-gas purification processes:

Psychrophilic	10 – 20 °C
Mesophilic	20 – 37 °C
Thermophilic	50 – 65 °C

The activity of the microorganisms increases with temperature, but the solubility of gaseous substances in water decreases. It is important for the reliable operation of a biological waste-gas purification plant to keep its temperature within the appropriate range as consistently as possible to prevent depressions due to population changes. In practice, most installations are operated in the mesophilic range.

The temperature range 37 – 50 °C should be avoided if possible, as a continuous fluctuation between mesophilic and thermophilic microorganisms can occur, reducing the efficiency of the installation.

pH Value. The hydrogen ion concentration in the medium affects the growth and metabolism of the microorganisms. Whereas most heterotrophic bacteria prefer the neutral range, fungi prefer the acid range. The growth and metabolism of microorganisms also cause changes in the pH of the medium. Such changes can be prevented by controlled additions of H_2SO_4 and NaOH (buffering), or by operating multistage equipment at various pH ranges.

7.5.2.2. *Physicochemical Conditions*

Since microorganisms can exist only in or on moist media, they can only metabolize the components of the gas to be treated if these are absorbed in the liquid phase.

Biological waste-gas purification is a biochemical absorption, the result of which is not exclusively determined by the physical solution equilibrium because of the reactive conversion of the dissolved gas [630]. Henry's solubility coefficient plays an important role, and is affected by the temperature and pH of the washing water [631].

As the component dissolved in the wash liquid are biologically degraded by the microorganisms present in a suspended state or fixed in the biofilm, thermodynamic equilibrium cannot be established. The resulting fall in concentration influences diffusion and accelerates absorption. This process depends mainly on the biological reaction rate.

However, with increasing biological activity, the absorption of pollutants depends less and less on the reaction rate and increasingly on the diffusion rate [632].

There are thus two areas in which the absorption process and hence the effectiveness of biological waste-gas purification processes can be influenced:

1. Diffusion
   For a constant concentration in the waste gas, the concentration difference can be increased by increasing the water supply or by additional biological purification of the wash water.
   The solubility can be increased by solubilizing agents (e.g., high-boiling oils) [633].
   The specific exchange surface (phase boundary) in the absorber must be as large as possible.
2. Kinetics
   The reaction conditions (temperature and pH) must be optimized.
   The residence time can be increased by additional aeration stages.
   The biological activity (cell concentration) can be influenced by adding nutrients at appropriate times.

As the microorganisms in a biofilter settle on the moist filter material, two types of transfer occur: adsorption on a solid substance (van der Waals forces) and biochemical absorption. The ratio of the two types of sorption in the filter bed depends on the moisture content of the materials. In purely dry zones only adsorption takes place up to the point of saturation, but no biological degradation can occur.

Biofilters are operated in a moist condition without water circulation, so that the absorption capacity and hence the efficiency of the equipment depends on a homogeneous and uniform flow through the filter bed. The flow behavior of the waste gas through the filter bed is thus determined by the void volume, the gas volume, and the associated relative pressures. The void volume of the material should be distributed uniformly over the filter cross section to ensure uniform flow rates and residence times at all points [634], [635].

Other important factors are the water retention properties of the filter material which determine the water content during operation. With increasing water content, the free void volume in the filter bed decreases and the gas flow rate and the pressure drop increase. The residence time and efficiency of the biofilter therefore both decrease.

7.5.3. ***Biofilters [636]***

Designs. The classical biofilter is the flat filter (see Fig. 95). The conditioned crude gas is compressed in a chamber under a tray provided with holes or slits. The distance between the bottom of the pressure chamber and the grid above should be at least 0.7 m to ensure uniform distribution of the gas flow. The air then flows through the filter bed whose height depends on the material chosen, and should be at least 1 m in a settled state. Flat filters are used where enough floor space is available, and where there are large flow rates with relatively small concentrations of pollutant gas or odoriferous material (e.g., slaughterhouses, composting plants, sewage works). If floor space is restricted, multistage filters of equal area can be stacked one above the other. These can be operated either in parallel or, with multistage operation, in series. In highly loaded waste gases with a low flow rate, e.g., from pharmaceutical and chemical plants especially paint manufacture and processing and in printing works, open or closed container plants (compact plants) can be used.

Figure 95. Schematic of a flat biofilter

a) Filter material; b) Slotted floor; c) Bottom seal

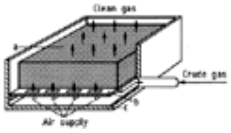

Design Parameters. Because biofilter plants are used to treat complex (odoriferous) gas mixtures, it is seldom possible to perform theoretical calculations. Plant design usually involves comparison with other plants and/or the results of pilot plant experiments. The most important design criterion is the volumetric filter loading, i.e., the amount of waste gas in m^3/h that can be passed through each m^3 of filter material while maintaining a given efficiency.

The bed volume determines the residence time of the waste gases in the filter bed, and depends on the structure and type of filter material. The following materials are used at present:

1. Composted garbage (20 – 40 mm) is sieved out of composted plants and organic domestic waste. This material is biologically highly active, but gives problems with flow obstruction by settled material and uncontrolled gas breakthrough. It is therefore mainly used in combination with inert material (clay, Styropor, and activated carbon) in compact equipment. The bed height is 1.0 m.
2. Coarse material (40 mm, 80 mm, 120 mm) sieved from composted plant material with a bed height of 1.0 – 1.5 m (depending on the humus fraction). The material is biologically very active, but has to be sieved annually.
3. Mixtures of brushwood/fibrous peat, brushwood/coconut fiber, or brushwood/heather are very suitable for flat filters, but can be difficult to procure. The bed height is 1.0 m.
4. Bark mulch from spruce and fir can be used if the material is homogeneously shredded to 50 – 150 mm. The bed height is ca. 1.2 m. Pine bark is unsuitable because it is friable, swells with moisture, and hence obstructs the flow.
5. Cracked root wood pieces, 60 – 180 mm long, can be used if the bed height is at least 1.5 – 1.8 m and is provided with a covering layer of ca. 0.2 m bark mulch for pressure equalization and protection from rain.

The function of a biofilter can be optimized by ensuring that the material has the following properties [637]:

1. Uniform structure (uniform flow and low pressure drop)
2. Adequate void volume (low pressure drop, low specific filter resistance, low energy consumption, good drainage and oxygen supply)
3. High portion of organic material (good additional supply of nutrients, long service life of the filter)
4. Large surface of the carrier material (large specific surface area for occupation by microorganisms)
5. No extreme pH fluctuations in the filter material (good buffer for variations in the waste gas)
6. Good water retention properties (constant moisture)
7. Low rate of decay (long service life, long maintenance intervals)
8. Low odor (low dispersion range of the odor, no unpleasant odor quality)

The balance of these properties is an important condition for optimum functioning of the filter.

The density of the filter material in the wet state is ≤1 t/m^3, depending on the type of material and its condition.

Design Details. Biological methods have only limited ability to smooth out peaks in the concentration of impurities in the waste gas. It is therefore preferable to operate the equipment continuously or to add buffers such as activated coke, to the system.

Biofilter materials become clogged by dust and fats from the waste gas. Also, certain operating ranges of temperature and moisture must be maintained. The crude gas must therefore be conditioned upstream of the biofilter (see Fig. 96). To prevent breakthrough of crude gas at the edge of the filter, a grid floor covering should be provided, extending to 1.0 – 1.5 m from the edge. At a bed height of ≥1.5 m, another edge screen 0.5 m wide should be added.

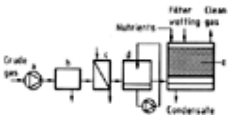

Figure 96. Preconditioning for a biofilter

a) Fan; b) Dust separator; c) Temperature controller; d) Scrubber; e) Filter bed

On concrete surfaces where biologically active material is present under pressure, e.g., on prefabricated slatted floors used in agricultural equipment, biological corrosion can take place. These surfaces should be sealed with liquid resin. Corrosion has so far not been found on the concrete walls or other equipment under the grid.

It is not necessary to provide a cover on flat filters, but a duct under the grid is required for seepage water. This water can be supplied to the scrubber for humidifying the crude gas, or can be sprinkled on the filter. Thus, the biofilter operates without producing waste water.

Maintenance and Repair. The filter bed must be kept free of plant growth, and must be regularly checked for settled material and gas breakthroughs. Fine filter materials has to be loosened and sieved at least once a year. In the case of cracked root packings, it is sufficient to loosen the top layer. As degradation of organic substances takes place in the filter material, an addition of ca. 5 – 10 % of the material volume must be made annually. Complete replacement of the material

is only necessary if the degradation process has proceeded to such an extent that the required air flow rate can no longer be achieved. To enable replacement or sieving of the filter material, the biofilter equipment has to be designed and constructed such that the grid is accessible for vehicles (wheel loader) or the filter can be maintained with a mobile gripper. The biofilter should be divided into several segments so that the entire production need not be shut down for maintenance work.

Process Monitoring. Biological waste-gas purification processes require a start-up period of 4 – 8 weeks, depending on the composition and concentration of the waste gas. It is only necessary to inoculate the filter if it is filled with an inert material. If waste-gas compositions match those of the associated sewage plants, sewage sludge can be sprinkled on the biofilter to increase its efficiency and to reactivate the biological activity rapidly after a long shutdown. The water content and pH of the material in the filter must be continuously monitored and controlled.

Additions of nutrients are only necessary if there is a deficiency in carbon, nitrogen, and phosphorus in both the waste gas and the filter material. Such a deficiency can occur after a long shutdown period or production changeovers.

Another important operating parameter is the continuous monitoring of the pressure drop in the filter bed. This indicates whether the filter bed has settled or whether surface densification (e.g., due to heavy rain) has occurred. Permeability properties are tested with smoke cartridges or fog generators.

7.5.4. *Bioscrubbers [630]*

Designs. Biological gas scrubbing can be carried out in bioscrubbers containing suspended microorganisms (see Fig. 97). They include an absorber in which the constituents of the waste gas are dissolved in the scrubbing liquid. This is collected in the absorber, and fed to an aerated activating tank. The regenerated scrubbing liquid is then recirculated to the absorber. The degradation efficiency depends on the active sludge mass formed from the microorganisms suspended in the scrubbing liquid and the dissolved pollutants. The limiting factor for this process is the distribution coefficient: for sparingly water-soluble substances in the waste gas extremely large amounts of water – sludge mixture have to be available and kept in circulation. For economic operation, the distribution coefficient should be higher than the limit value of 0.0005 g/m^3 gas to 1 g/m^3 liquid. The addition of a high-boiling oil (e.g., a methylsilicone fluid) as a 10 – 30 % dispersion to the water – sludge mixture can increase the absorption rate by a factor of 100 – 1000 [633, 640, 641].

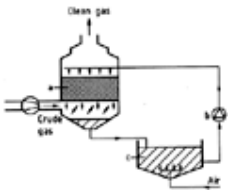

Figure 97. Bioscrubber with suspended microorganisms

a) Absorber; b) Activated sludge circulation; c) Activating tank

In a trickling-bed reactor, absorption and biological degradation take place in the same part of the equipment (see Fig. 98). The scrubbing liquid is circulated over an inert packing material and collected in a pump sump. A biological film forms on this material, which can consist of loose filling material or a structured packing. The oxygen needed for biological degradation and the constituents of the waste gas to be degraded are absorbed by the water phase and transported to the biofilm where the reaction takes place. The degradation capacity depends on the specific surface of the packing, i.e., of the biofilm, and the residence time in the trickle bed.

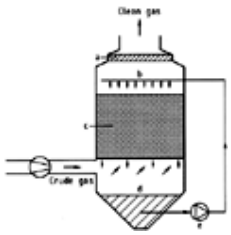

Figure 98. Trickling-bed reactor with fixed microorganisms

a) Demister; b) Nozzles; c) Packing material with biological coating; d) Pump sump; e) Water circulation

Because mass transfer from the gas to the liquid phase takes place at the same time as the reaction on the biofilm of the absorber, the trickling-bed reactor can be used if the distribution coefficient (g/m^3 gas to g/m^3 liquid) of the components to be degraded is higher than 0.0005 [637].

Design Parameters. There are more design criteria to be considered with bioscrubbers and trickling-bed reactors than with biofilters. A correctly dimensioned and operated biofilter has a large buffering capacity which a bioscrubber cannot have, as the latter has to give the same purification efficiency with only a fraction of the biologically active volume. Whereas in biofilters only small parts can become overloaded or sour and respond to toxic substances, disturbances in bioscrubbers have an immediate effect on the entire biosystem.

Therefore, it has to be determined at a preplanning stage whether certain components of the waste gas that are harmful to the microorganisms or can clog the packing need to be removed by dedusters or prescrubbers before they reach the bioscrubbers. It must also be decided whether the composition of the waste gas makes a two-stage operation (acid/neutral) necessary.

The reaction products and cell mass become concentrated in the circulating scrubbing liquid. Part of the scrubbing liquid therefore has to be removed at regular intervals and replaced by fresh water or solution. Any solubilizing agents used have

to be separated before the spent liquid is disposed of.

Thus, bioscrubbers are waste-gas purification devices producing wastewater (according to the German Water Management Act they are not wastewater purification devices). As a rule, the mixture of water and suspended material which is continuously removed from the installation does not contain heavy metals and other hazardous substances, and can therefore be disposed of in domestic sewage plants without problems. Ranges of characteristic parameters for bioscrubbers can be found in [630, 633, 638, 639].

If the necessary design parameters are unknown because the mixture of substances to be treated cannot be precisely determined, as in the case of emissions of odorous substances, pilot plant investigations have to be carried out. If the waste-gas load cannot be reproduced, reference substances may be used to test the purifying efficiency of the pilot plant.

Design Details. Materials that come into contact with intensive biological activity have to be protected from (biological) corrosion. Metals inside a scrubber, including stainless steels and seawater-resistant aluminum, have to be coated with plastic.

Packings in absorbers and trickling filters have to be designed and constructed such that a biofilm can form while a low pressure drop is guaranteed and the packing does not become clogged with activated sludge. A good impingement force of water will prevent clogging and flow obstructions, and there large inspection openings are necessary to enable an easy addition and removal of the packing material.

The scrubbing liquid distribution system above the packing material needs to be designed to operate without blockages. The water circulation pumps have to be suitable for transporting activated sludge, and need to run slowly so that the sludge flocks formed in the activating tank are not broken down. Displacement or slow-running plunger pumps are therefore recommended.

Process Monitoring. All the parameters relevant to operation of the process, such as temperature, pH of the scrubbing liquid, pressures in the scrubbers, oxygen and nutrient concentrations, and characteristics of the spent scrubbing liquid have to be determined continuously or at regular intervals for process control. Maintenance at regular intervals is therefore unnecessary, as the packed bed need only be cleaned when the pressure drop indicates flow obstruction. When biological scrubbing equipment is in the initial start-up phase, e.g., after inoculation with activated sludge from an associated sewage treatment plant, natural selection of the microorganisms (adaptation) takes place for 4 – 6 weeks. Significant changes or fluctuations in the crude gas composition may require re-adaptation, which can also take several weeks.

During plant shutdowns when no pollutants are added via the waste gas, nutrients have to be added to the scrubbing liquid by metering devices. The biological oxygen demand must be continuously supplied by aeration of the activating tank or by passing indoor air through the absorber.

Process Optimization. The purification efficiencies of biological plants can be improved by connecting a trickling-bed reactor with a biofilter to form a two-stage process. The first stage acts as a neutralizer, reacting with acid or basic components of the waste gas. The biofilter is then used without special control techniques to reduce the residual pollution.

Further developments have led to the biosorber principle [642], a biofilter and a bioscrubber combined in a single unit. This consists of a packing of inert material on the surface of which nutrient media specially suitable for the application are grown.

In the biomembrane principle [643], strongly hydrophobic waste gas components diffuse under pressure through a silicone tube located in a mixture of activated sludge, water, and, if necessary, solubilizing agent. The tube is coated on the outside by a biofilm, so that the biological reaction takes place in the biofilm and the waste gas components need not be dissolved in a liquid. Waste gas components that are not picked up by the biofilm can then be treated in the activating tanks.

Measurement and Assessment. Determination of the gaseous components of a waste gas before and after biological purification is carried out by a wide range of methods, including gas chromatography, and spectrometry. To determine odoriferous substances, olfactometry is used for measuring emissions [644], and field inspection with panelists is used for ambient air pollution [645]. The following criteria should be taken into account:

1. Determination of the odor concentration in the waste gas is by itself not a criterion for odor nuisance. The odor intensity [646] and the position on the hedonic scale [647] also need to be determined.
2. The odor of the crude gas is changed by the biological treatment. This “biological smell”, which is undetectable a few meters away from the equipment, also has an odor concentration, so that the calculated efficiency gives a false impression of the purification performance of the equipment.
3. Biofilter material, e.g., composted domestic waste, can have its own specific odor.

Assessment of the ability of a biological process to remove odor emissions therefore has to be carried out in accordance with the VDI Guidelines [630], [636], which state that the aim of the purification process is achieved when the characteristic odor of the crude gas is undetectable after purification in the biofilter or bioscrubber.

7.6. Flue-Gas Purification in the Incineration of Hazardous Waste

In the incineration of hazardous waste, the methods and equipment described in Sections Separation of Solid and Liquid Particles from Gases , Waste-Gas Purification by Thermalor Catalytic Reactions, Absorption, Adsorption, Biological Processes are used together with additional technical and operational procedures to enable high standards of environmental protection to be met. As a result of their composition, condition, or quantity, such materials can endanger health, the atmosphere, or water supplies, or they can constitute an explosive hazard. Typical substances include:

1. Solvent mixtures
2. Industrial sludges
3. Paint residues
4. Plastic waste
5. Inorganic waste (salts)
6. Halogenated hydrocarbons
7. Petrochemical residues
8. Halogenated acids/alkalis
9. Pharmaceutical waste
10. Acid sludge
11. Waste oil
12. Pesticides
13. Herbicides

If hazardous wastes cannot be treated or reused and direct dumping is impossible, they must be incinerated to destroy their organic constituents and produce inert combustion products that can be reused or safely dumped. By far the largest quantities of hazardous waste are incinerated in rotary furnaces. This enables the incineration of solids, liquids, and pastes and the thermal treatment of substances that produce molten slags during the incineration process. Solids and pastes are added by crane loaders and liquids through burners and lances, or barrels are charged to the furnace intact.

A simplified flow diagram of a complete plant is shown in Figure 99 and a section through an actual plant in Figure 100.

Figure 99. A) Simplified flow diagram of a hazardous-waste incineration plant

B) Flow diagram of a rotary furnace

Figure 100. Hazardous-waste incineration plant

With kind permission of the L. & C. Steinmüller GmbH, Gummersbach, Federal Republic of Germany

In the incineration process, the elemental constituents of the waste are retained, but their chemical and physical forms are changed. To prevent unacceptable contamination of the environment by hazardous substances, either by their entry into the atmosphere or as solids, the operation of hazardous-waste incineration plants must include comprehensive measures to reduce emissions.

Typical concentrations of pollutants in crude gas produced by “publicly” operated incineration plants used to treat waste from trade, business, and industry are:

Dust	1000 – 10 000 mg/m^3
Inorganic chlorine compounds (as HCl)	3000 – 10 000 mg/m^3
Inorganic fluorine compounds (as HF)	50 – 550 mg/m^3
Sulfur compounds (SO_2 + SO_3)	1500 – 5000 mg/m^3
Nitrogen oxides (as NO_2)	100 – 300 mg/m^3
Water vapor (H_2O)	6 – 20 vol %

In plants operated by the chemical industry for the incineration of production residues with high contents of reactive substances, a much wider range of concentrations can occur in the waste gas. Table 24 lists the legally required emission limit values for existing and newly constructed plants in Germany. The emission limit values in an EC Guideline proposed in 1993 and given legal force in the 1994 plan are also given. These are to be incorporated in National Law within two years.

Table 24. Waste incineration emission limit values simplified synopsis of EG Guideline/17th BImSchV, Germany (limit values in mg/m^3 based on STP and 11 % O_2)

Component	EG Guideline 17. BImSchV

Carbon monoxide		
Daily mean value	50	50
Hourly mean value		100
90 % of all measurements (peak concentration) in 24 h		150
95 % of all 10 min mean values in 24 h	150	
Daily mean values (DMV)		
Total dust	10	10
Organic substances (total-C)	10	10
Hydrogen chloride	10	10
Hydrogen fluoride	1	1
Sulfur dioxide	50	50
Nitrogen dioxide	(none)	200
Half-hour mean values (HMV)		
Total dust	30	30
Organic substances (total-C)	20	20
Hydrogen chloride	60	60
Hydrogen fluoride	4	4
Sulfur dioxide	200	200
Nitrogen dioxide	(none)	400

The emission limits for conventional air pollutants have been reduced considerably, and a new limit value for the emission of polychlorinated dibenzodioxins and dibenzofurans (PCDD/PCDF) has been introduced. Whereas emissions of conventional atmospheric pollutants are essentially a function of the type, amount, and composition of the waste materials incinerated, dioxins and furans and other organic hydrocarbons are formed by complex decomposition and side reactions in the combustion and cooling processes.

According to current knowledge, most of the dioxins and furans found in the flue gas are formed from unchlorinated organic material in the presence of a chlorine source. In the temperature range 450 – 250 °C, catalytic formation of chlorine from metal chlorides in the presence of oxygen is the essential initiation reaction. The mechanism of formation assumes the presence of chlorine and oxygen, and this is unavoidable in the flue gases produced by incineration of hazardous waste. Dioxins are formed in a solid state reaction, and are mainly bonded to particles.

The primary emission control measures are meticulous planning of the incineration process and control of the firing process. The incineration plan determines how the waste materials are metered and added. If possible, they are size reduced and mixed to ensure that the furnace feed has consistent properties, including calorific value, content of components such as halogens, sulfur, nitrogen, heavy metals and organic compounds, and other properties relevant to the incineration process such as ease of ignition and burning rate. Drums to be added intact to the incinerator for reasons of operational safety must be of a size and energy content suitable for the firing equipment. Owing to the heterogeneous nature of the fuel, the usual systems for controlling the combustion air are unsuitable. The amount of combustion air is often fixed and the waste material is added in a controlled manner to give the required temperature, controlling the CO content at the same time. Emission control measures in the incineration process affect almost exclusively the carbon monoxide level and the total content of organic compounds.

The special characteristics of the rotary furnace, such as the frequent change of fuel for the burners and sludge lances and feeding of drums lead to large, brief fluctuations in the flue gas composition. High air pollutant concentrations that change according to the feed and the combustion conditions must be dealt with by secondary emission control measures. For a flue-gas cleaning system for a modern hazardous-waste incineration plant, not only must the capacity to remove pollutants be high, but also the following requirements must be met:

1. The process must be designed so that no wastewater is produced
2. Residues from the various processes used in waste-gas purification must be reused as far as economically possible
3. Residues that cannot be reused must be safely dumped

The costs of dumping depend on the nature of the waste material. For example, the disposal of residues containing heavy metals, dioxins, and soluble salts in an underground dump is approximately five times as expensive as disposal in an open hazardous-waste dump supervised in accordance with modern technical regulations. Therefore, on cost grounds, every operator endeavors to concentrate potentially environmentally hazardous substances as far as possible, so that only small amounts of residues will need expensive treatment. Finally, flue-gas purification must function with a high degree of safety and at low operating cost.

Flue-gas purification equipment for hazardous-waste incineration plants usually incorporates a dry dedusting stage and several wet scrubbing stages. A design of such elements including spray tower scrubbers is shown in Figure 101. The combustion exhaust gases are cooled in a steam-generating heat exchanger to ca. 250 °C. The design concept requires

1. Preliminary dedusting of the waste gases by a double-field electrostatic filter
2. An initial scrubbing stage to cool the combustion gas to the cooling limit temperature (quench), with removal of HF, HCl, and heavy metals
3. A second scrubbing stage to remove SO_2, producing the required waste-gas purity
4. Wastewater treatment including gypsum production and external evaporation of the $CaCl_2$ solution produced

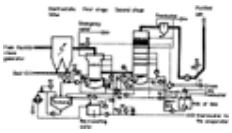

Figure 101. Dedusting and wet scrubbing stages as elements of flue-gas purification system

The first scrubbing stage, also termed the prescrubber, is operated at pH <2 to remove HCl, HF, and heavy metals. At this pH, the chloride content of the wastewater can be high, which is indispensible or advantageous for the external evaporation step. In the first scrubbing stage, only a limited amount of SO_2 is removed. The sulfite formed is oxidized to sulfate by aeration of the sump. The second scrubbing stage, also termed the main scrubber, predominantly removes SO_2. In the spray tower, the suspension used for the scrubbing is introduced via nozzles at different three levels. The droplet content on the pure gas side should be as low as possible. The droplets are removed by an internal and an external demister. Sulfur dioxide scrubbers are operated at pH 6.5 – 7; the absorption medium is milk of lime. The salts (mainly gypsum) formed by reaction with SO_2 are removed in the wastewater treatment stage. To improve the ease of water removal from the insoluble salts and to ensure high purity of the separated sulfate (>99 % $CaSO_4$ in the solid), the sulfite formed during the absorption is oxidized to sulfate by passing air into the sump of the scrubber. Fine crystals must be avoided as far as possible in the sulfate crystallization. Therefore, a solids content of ca. 8 % is maintained in the wash liquor suspension to ensure that enough seed crystals are produced during the crystallization and spontaneous nucleation is avoided. The $CaSO_4$ formed is removed from the scrubber overflow by a thickener and band filter. The clear fraction is returned to the scrubbers, and a small percentage is taken directly to wastewater treatment.

Special technical measures must be taken to deal with fluctuations in the concentration of the waste-gas constituents. By using rapidly responding control systems and by directly feeding the absorption medium into the scrubber inlet pipes, an adequate quantity of absorption medium is provided at all operating points. The rate of addition of the suspension is controlled in accordance with the concentrations measured in the crude gas after the electrostatic filter. Both scrubbers include normal methods for measuring and controlling pH, density of the suspension, and flow rates by cascade controllers. Temporary supersaturation in the first scrubbing stage can lead to incrustation and scaling. This is prevented by spraying the top of the scrubber with fresh water. This mechanical and chemical rinsing also provides the water required for the process. Hydrogen chloride peaks not eliminated in the prescrubber are reduced to permissible levels in the main scrubber. Any significant chloride loading in the main washer is undesirable because it has an unfavorable effect on the equilibrium conditions for SO_2 removal. Therefore, an external demister is included between the prescrubber and the main scrubber. As well as the cascade control system in the main scrubber, additional equipment is necessary in the liquid circulation system to maintain the liquid balance and achieve the necessary oxidation and crystallization of the insoluble solids simultaneously with the removal of SO_2.

The spray tower scrubber can be used as an example of the problems to be solved when transferring a process technology to full industrial scale. In other full scale plants, spray scrubbers, rotary scrubbers, venturi scrubbers, or packed tower scrubbers are used. Sodium hydroxide solution can be used as an alternative to a milk of lime as absorption medium. Sodium salts are soluble in water, which avoids the danger of incrustations in pipework and equipment. Instead of the electrostatic filter, a cloth filter can also be used. The choice depends on the material to be incinerated. Cloth filters, which are usually operated at ≤ 200 °C, are unsuitable, if there is a danger of clogging by fine dust produced in the incineration process, or if the dust is hygroscopic and forms coatings on the filter that cannot be removed. Electrostatic filters are unsuitable if the dust cannot be electrostatically precipitated, if the electrical resistance of the dust layer is too high, or if the dust has such a low density (e.g., in the incineration of chlorosilane) that it is not removed on trapping the electrostatic elements. Therefore, in some areas of the chemical industry dry removal of dust is infeasible. For wet-dust removal, single-stage or multistage venturi or rotary scrubbers can be used (see Section Wet Scrubbers). These operate on the principle of contacting the dust-laden waste gas with a relatively finely dispersed liquid. The dust particles come into contact with the liquid droplets, are wetted, and are collected along with the liquid. The dust must therefore be wettable.

The liquid from the scrubbers goes to a water-treatment stage which includes the precipitation of heavy metals, flocculation, and solids separation. The heavy metal sludge also contains dioxins bound to the dust collected by the scrubbers. The clarified liquid is evaporated externally (if the dry product can be sold) or internally by a spray dryer that forms part of the main waste-gas stream. Internal evaporation produces a relatively large amount of a dust – salt mixture containing dioxin and heavy metals. This residue is usually disposed of by underground dumping.

In the process described above, mercury is removed in the first, acid scrubbing stage. The residual concentration in the purified gas depends on the vapor pressure of mercury. If mercury concentration peaks occur or if mercury is present in the scrubber in metallic form (e.g., caused by dissolved sulfur dioxide), this can lead to the breakthrough of mercury, which can then exceed emission limit values. This system also does not include any process to limit the emission of nitrogen oxides. If large amounts of aerosols are present, there is an increase in the dust content and hence also in the dioxin content of the purified gas.

Thus, the process technology for waste-gas purification must be improved for new plants. There are many proposals based on laboratory or pilot plants. Emissions of nitrogen oxides can be reduced by the selective catalytic reduction (SCR) process as used in power station technology. Only the low-dust variant is relevant. The waste gas must be reheated to ca. 250 – 300 °C before it enters the catalyst, by heat transfer from the hot to the cold waste gas or by heating with burners. Experiments have shown that when catalysts are used to reduce emissions of nitrogen oxides, organic trace substances such as dioxin can also be catalytically oxidized. Reaction efficiencies (removal efficiencies) of ca. 90 % for PCDD and PCDF have been found in semitechnical experiments. The oxidation catalyst would then perform two of the additional emission-control tasks required. A particular advantage is that no additional residue is produced in this process.

In many hazardous-waste incineration plants, nitrogen oxide concentrations in the crude gas are < 300 mg/m^3. To achieve the legally required value of 200 mg/m^3, a reduction of 30 % is necessary. This can be achieved either by the selective noncatalytic reduction (SNCR) process in which ammonia or aqueous ammonia solution is added to the waste gas at 850 – 1000 °C. As the reaction does not proceed to completion, ammonium chloride can be formed as an aerosol. Furthermore it can be assumed that some of the ammonia will go into solution in the first scrubbing stage and must then be removed in a stripping operation.

To remove aerosols, a wet electrostatic filter can be used as final purification stage. The purified gas leaving the last scrubbing stage is cooled until droplets form and agglomerate, and the droplets are removed by an electric field. Decreasing the dust content of the purified gas automatically reduces dioxin emissions. If cooling before the wet electrostatic filter gives a temperature of up to 20 °C below the cooling limit temperature, the mercury content is also reduced because of the lower vapor pressure.

Multifunctional separations are achieved by activated carbon filters (see Section Adsorption), which have undergone rapid development. They have been tested in full-scale experiments, and are already used in one hazardous-waste incineration plant in Germany. Activated carbon filters, operated as fixed or moving beds (Fig. 102), adsorb heavy metals and organic trace substances simultaneously. They also adsorb any hydrogen chloride and sulfur dioxide that have passed through the earlier stages. If they are used on the purified-gas side, the consumption of activated carbon is very low. Entrainment reactors can also be used, in which a mixture of activated calcium carbonate with a proportion of activated carbon is blown into the waste gas and then removed on a cloth filter. Using this as the final stage waste-gas purification gives results as good as those obtained with moving or fixed bed activated carbon filters. Examples of analysis figures for the purified gas in full-scale plants are given in Table 25. Technical-grade hydrochloric acid can be produced from the chloride solution generated in the first scrubbing stage. For this, the first scrubbing operation is carried out in three stages, in which dust and heavy metals are removed in the first stage, the required concentration of HCl is obtained in the second stage, and the third stage performs the final purification. The hydrochloric acid is removed and distilled. The distillation process would replace an external evaporation of chlorides, which are generally difficult to market.

Table 25. Comparison of emission limit values according to EC Regulations and 17th BImSchV (Germany) with initial results from a hazardous waste incineration plant with a carbon filter with an activated coal – lime cloth filter. Values are in mg/m^3 at STP based on 11 % O_2

	Emission limit (mean values)		Daily mean values Dec. 91 – Feb. 92
	17. BImSchV		
	Daily	Half-hourly	
Total dust	10	30	<1 [a]
$\sum C_{org}$	10	20	<2
CO	50	100 [c]	25
HCl	10	60	<2 [a]
HF	1	4	<0.5 [a]
So_2 (SO_2, SO_3)	50	200	8
NO_x (NO, NO_2)	200	400	120

Heavy metals		
	∑0.05 Cd, Tl	<0.007 [b]
	0.05 Hg	<0.002 [a]
	∑0.5 Sb, As, Pb, Cr, Co, Cu, Mn, Ni V, Sn	<0.2 [b]
Dioxins and furans as ∑TE[d]	0.0000001	0.000000028 (=0.028 ng)

[a] Below detection limit.
[b] Sometimes below detection limit.
[c] Hourly.
[d] TE = Toxic equivalent values.

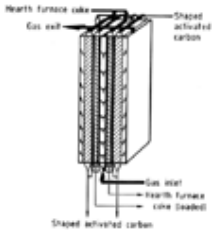

Figure 102. Hearth furnace coke adsorber – shaped activated carbon catalyst (countercurrent flow)

All these processes for additional waste-gas purification and reuse of residues can lead to complex design and process technology (Fig. 103). Plants of this type have not yet been realized. In Germany the first plants of this type will be put into operation in 1996 and 1997.

Figure 103. Possible plant design for hazardous-waste incineration plant

a) Material supplied in containers; b) Charging of drums; c) Rotary furnace; d) Post-combustion equipment; e) Waste-heat boiler; f) Evaporative cooler; g) Electrostatic scrubber; h) HCl scrubber; i) SO_2 scrubber; j) Suction pump 1; k) Heat exchanger; l) Burner; m) Aqueous ammonia addition; n) Combination catalyst; o) Addition of activated coke – lime mixture; p) Filter/adsorber; q) Suction pump; r) Stack; s) Fly ash vitrification; t) Hydrochloric acid recovery; u) Container for hydrochloric acid; v) Gypsum dewatering; w) Gypsum silo

It was formerly intended that hazardous-waste incineration plants should be designed for the treatment of all types of hazardous waste using all techniques available. These requirements now appear to be changing. It is apparent that the cost of incineration is increasing rapidly, and cheaper methods for disposing of hazardous waste will be sought.

The positive results obtained using the entrainment method and the fact that the SNCR method for removing nitrogen oxides has been very effective have led to a reassessment of the situation. It is likely that in future hazardous waste will be more carefully sorted to match the material to the incineration method used. Wastes that do not require high-quality gas purification will then be treated in the appropriate plants using simpler techniques like dry or semidry absorption methods.

[Top of Page]

8. Photochemical Degradation

Walter Simmler

8.1. Introduction

In terms of mass and energy turnover, photochemical degradation is one of the largest chemical reactions in nature. It is ultimately an oxidation process involving numerous physicochemical intermediates. It takes place in the troposphere and makes an essential contribution to global recycling because of its mineralizing effects, leading to the formation of carbon dioxide and water. In simple terms, photochemical degradation starts with the formation of reactive species, usually radicals, when the molecules of certain atmospheric trace substances are cleaved by sunlight. These species then react with other, mineralizable trace substances.

These in turn can be of natural and anthropogenic origin: organic and inorganic components of sea spray, exudates from forests, products of biotic transformation processes, agriculture, industrial emissions, etc. [648-650].

The most important single process which brings about natural purification of the atmosphere is the formation of hydroxyl radicals, which oxidize hydrocarbons and their derivatives, e.g.,

$$O_2 + NO_2 \xrightarrow{h\nu} NO + O_3$$

$$O_3 + H_2O \xrightarrow{h\nu} O_2 + 2\,{}^{\bullet}OH$$

$$2\,{}^{\bullet}OH + CO \longrightarrow H_2O + CO_2$$

Here, as in most other cases, it is indirect photochemical degradation that is taking place. Direct photochemical degradation — the absorption of a light quantum followed immediately, or after rearrangement, by oxidation — is comparatively rare and plays only a minor role in natural purification of the atmosphere and in special cases of industrial air purification. The term photochemical degradation is often used in the literature for reasons of simplicity when indirect photochemical degradation is really meant.

The hydroxyl radical is the most important oxidizing agent formed photochemically, although it is present at a concentration of only 0.5×10^6 molecules/cm^{-3}. However, it reacts rapidly and fairly unselectively, is continuously formed, and is involved in the first stage of most oxidative transformations in the troposphere. In contrast the activation energy for direct oxidations by molecular oxygen in the troposphere is so high that measurable effects are not accomplished in reasonable periods of time.

For determining the lifetimes of chemical substances in the environment or of their residence times in individual environmental compartments, atmospheric conditions and processes are of great importance, especially in the case of emissions into the atmosphere of homogeneous gases or substances bound to particles. This also holds for deposition and other immissions from the air. Interactions with energy fluxes in the atmosphere and with atmospheric cycles of trace substances are particularly important [651-656].

The role of the atmosphere has many aspects: air acts as a container, vehicle, reactor, reactant, and filter, these functions sometimes being mutually interdependent. Here, the decisive feature is the characteristic temperature profile with increasing elevation (Fig. 104) [649], [657], [658].

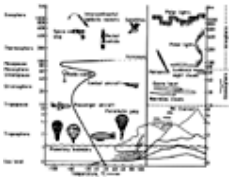

Figure 104. Atmospheric strata with temperature profile

(Origin: Funds of Chemical Industry “Ecological Compartment Air”, 1987)

From ground level, the temperature at first decreases linearly by ca. 6.5 °C/km. This zone, the troposphere, extends to an altitude of ca. 10 km (9 km at the poles, 15 km at the equator). It is formed by rising masses of air heated by the earth's surface and then subsequently cooled. Above the boundary zone (the tropopause), the stratosphere begins, and temperature now increases with altitude as a result of solar radiation; short-wave sunlight is directly absorbed. The absorbing substance is ozone in the lower stratosphere (altitude ca. 25 km) where the O_3 concentration is ca. 10^{12} molecules/cm^3. The ozone concentration decreases rapidly both above and below this zone. The presence of ozone in the lower parts of the troposphere is a result of other formation mechanisms (see Section Reaction Cycles in the Atmosphere).

The inversion of the temperature gradient above the tropopause is essential for maintaining the stability of the stratosphere. It causes the formation of a zone which acts as a barrier layer that prevents rapid ascent of trace gases from the earth's surface into the stratosphere. The vertical diffusion coefficient is at a minimum at this point, and penetration through this layer takes several years.

The composition, i.e., the nature and amount of the principal reactants in the zones of the troposphere close to the earth's surface, determines the processes that lead to degradation (Fig. 105).

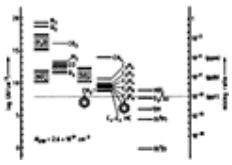

Figure 105. Characteristic concentrations (in molecules/cm^3) and volume mixing ratios (= mole fraction) of tropospheric main and selected trace gases

The hatched areas indicate common variabilities

HC = hydrocarbons

Apart from concentration, the various reaction rate constants of the trace substances are decisive (see Section Measurement and Computation of Degradability). The only source of energy is the sun. The energy flux for vertical radiation has a mean value of 1378 W/m^2, corresponding to a total annual energy intake of 1.56×10^{18} kW · h, with the radiation maximum at 480 nm (yellow-green). Of this, ca. 70 % is absorbed at the earth's surface on a cloudless day in Central Europe (Fig. 106).

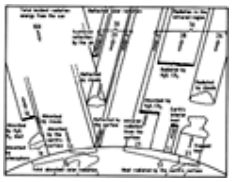

Figure 106. Percentage energy balance of the earth's surface

(Origin: Funds of Chemical Industry “Ecological Compartment Air”, 1987)

In the lower troposphere the fraction of the radiation absorbed is 17 %. This corresponds to a daily amount comparable to annual global energy consumption by humans today. After subtraction of the dissipated light from the incident light, a proportion remains for tropospheric photochemistry [659]. Figure 107 shows the spectral distribution of sunlight and the limiting wavelengths for important photolysis reactions.

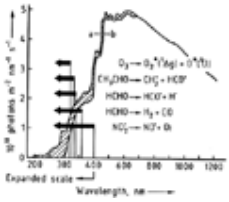

Figure 107. Spectral distribution of sunlight and wavelength limits for some important photolytic reactions

a) Incident sunlight reaching the outer atmosphere; b) Incident sunlight reaching the earth's surface

Hatched area: portion of light absorbed in higher atmospheric strata

The formation of photooxidants and their reaction with pollutants entering the troposphere primarily depends on the photodissociation of O_2 and O_3. The fundamental reactions were formulated as early as 1930, albeit for the stratosphere [660].

8.2. Natural and Anthropogenic Emissions

Photoinduced degradation of emissions that enter the atmosphere from the earth's surface is achieved by physical processes, particularly those occurring in the zones of the troposphere near ground level, and by trace gases that are already present there, especially carbon monoxide, methane, and nitrogen oxides. Via formation of photooxidants, they determine the rate of degradation of other trace substances emitted into the atmosphere. Several billion tonnes of such substances are emitted from the earth's surface annually. They are chemically transformed in the troposphere, mainly by mineralization to form oxidation products, and return as such to the earth's surface [661], [662].

The concentration of a trace substance is determined by its rates of emission and degradation. The most important of such substances (Table 26), originating from various sources, are measured and calculated [649, 651, 663-669].

Table 26. Global annual emissions*

	Emission rate, 10^9 kg/a	Main sources, 10^9 kg/a	
Carbon dioxide	830 000	respiration, biological degradation	370 000 – 520 000
	(700 000 – 1 000 000)	oceans	300 000 – 450 000
		combustion of fossil fuels and forest clearance	20 000 – 30 000
Carbon monoxide	3400	combustion of biomass	400 – 1600
	(1700 – 5800)	combustion of fossil fuels	500 – 800
		oxidation of hydrocarbons in the atmosphere	800 – 2200
Hydrocarbons except CH_4 (calculated as C)	1000	trees (terpenes, isoprenes)	600 – 1300
	(640 – 1400)	motor vehicles	30 – 50
		solvents	10 – 20
Methane	500	swamps and geothermal activity	100 – 300
	(350 – 1600)	rice paddies	50 – 300
		animal husbandry (ruminants)	100 – 200
		termites	100 – 150

Sulfur compounds (calculated as SO_2)	400	combustion of coal and mineral oil	160 – 240
	(290 – 500)	oxidation of sulfur compounds from oceans, swamps, etc.	80 – 200
		volcanoes	10 – 20
Nitrogen oxides excluding N_2O (calculated as NO_2)	160	combustion of fossil fuels	40 – 80
	(70 – 320)	combustion of biomass	10 – 80
		lightning	7 – 70
		atmospheric oxidation of ammonia	3 – 30
Ammonia	100	biological degradation, soil	up to 1200
	(30 – 1200)	animals	20 – 30
		wastewater treatment	10
Dinitrogen oxide	50	soil	up to 600
	(10 – 600)	oceans	6 – 15
Hydrogen	50	hydrocarbon oxidation	20 – 50
	(40 – 130)	industry and traffic	10 – 60

* This table lists the probable total annual quantities of emissions of various substances. The spread of values quoted in the literature is given in parentheses. Because the estimates of the main emission sources show such wide divergence, only the ranges of these estimates is shown.

In many countries of the world (Table 27), certain emissions are continuously measured [663]: CO_2, CH_4, and N_2O because of the greenhouse effect, and CFCs because of their effect on the ozone layer.

Table 27. Main emissions by countries (10^3 t/a)

	CO_2			CH_4	CFCs	N_2O
	From energy use	**From industrial processes**	**Total**			
Canada	435 400	25 700	461 100	3 700	14	92
United States	5 035 000	31 950	5 066 950	29 000	90	428
Japan	1 078 900	46 600	1 125 500	1 338	111	47
Germany	956 900	25 000	981 900	6 060	15	220
Italy	418 500	92 078	510 578	3 646	25	42
Spain	236 200	25 764	261 964	1 400	11	
Sweden	56 300	3 250	59 550	490	2	11
United Kingdom	608 400	5 981	614 381	4 844	31	109
OECD	10 438 660	372 880	10 811 540	63 840	350	
World	21 613 300	733 850	22 347 150	274 930	480	

By 1978, over 1300 trace substances had been identified in the atmosphere [667]. Substances of low vapor pressure enter the atmosphere bound to dusts [653], [670-672], and take part in photochemical degradation like other aerosols (see Sections Natural and Anthropogenic Emissions and Lifetimes of Adsorbates Aerosols). Horizontal transport processes occur in this way, and often removal of transformation products from the troposphere by precipitation. Smog, which can have a variety of compositions, modes of formation, and reaction paths is also included here, especially in connection with reaction cycles, atmospheric lifetime, and quality assessment.

So-called London smog differs from Los Angeles smog, also referred to as photosmog. London smog was formed when the smoke from domestic fires typical of the technological state of the art in the early 1900s combined with fog. Owing to the use of sulfur-containing fuels, the smog was acidic, and catalytic oxidation by heavy metal containing dusts converted sulfur dioxide to the trioxide. This type of smog occurred particularly under conditions of atmospheric inversion typical of the winter. Today, improved combustion technology and low-sulfur fuels have led to its virtual disappearance worldwide.

Los Angeles smog is also formed in inversion conditions but is caused by intense solar radiation acting on emissions from motor vehicles. Volatile organic compounds (VOCs) behave similarly, with the exception of methane. Systems of this general type are referred to as photosmog [649], [673]. It consists of photooxidants — products of the oxidation of VOCs by nitrogen oxides. Generally, these are organic nitrates and peroxo compounds such as hydrogen peroxide, methyl

hydroperoxide, peroxyacetic acid, peroxynitric acid, and peroxyacetyl nitrate. The most important component of photosmog is ozone, which far exceeds all other compounds in concentration [674]. The formation and reactivity of photosmog depend very much on the time of day, and follow complex mechanisms (see Section The Stratosphere).

Information about the earlier condition and behavior of the atmosphere is provided by investigations of sample cores from boreholes drilled in the ice of polar regions, especially Greenland and the Antarctic, and also from cores obtained from the seabed. Contained in the occluded gas bubbles are trace substances from earlier times, including materials deposited as solid particles from the atmosphere, e.g., from volcanic eruptions or meteorite impacts. The mightiest continental ice shield is found in the South Polar regions, with a thickness profile corresponding to a time span of over 100 000 a. The composition of the troposphere can thus be traced back over this time period. The carbon-containing compounds, especially methane, carbon monoxide, and carbon dioxide, are the most meaningful as they are partly of anthropogenic origin and participate in the most important tropospheric reaction cycles. They have been thoroughly, and in some cases continually investigated, as shown by the example of emission data for methane:

Sources (10^{12} g/a)

1. *Fossil deposits*
   As 18 – 24 % of the methane does not contain any radioactive ^{14}C, the amount from fossil deposits is (100 ± 20) × 10^{12} g/a:

from coal mines	25 ± 5
from methane hydrates	~5
from losses of natural gas	70 ± 15

   Thus, 6 – 9 % of natural gas production is lost as CH_4 emissions
2. *Biogenic sources*

Ruminants	~80
Dumps	50 ± 20
Ocean and termites	~30
Wet regions and rice paddies	215 ± 50
Combustion of biomass	30 ± 15
Total sources	505 ± 105

Sinks

OH oxidation	420 ± 80
Uptake by the soil	30 ± 15
Stratospheric degradation by Cl and O (^1D)	10 ± 5
Subtotal	460 ± 100
Annual increase	45 ± 5

The main sources of methane are in the undeveloped countries [675], and the main sources of nitrogen oxides in the industrialized countries [676].

Emissions of NO_x (10^6 t/a, calculated as NO_2) can be summarized as follows:

Ground sources	124
Combustion of fossil fuels	69
Emissions from the soil	18
Combustion of biomass	37
Atmospheric sources	30
NH_3 oxidation	10.2
Lightning	16.4
Aircraft exhaust gases	1.0
NO_x from the stratosphere	2.0

Total	154 ± 76

Emissions in the form of aerosols make a large contribution to the input to the troposphere. In total, 1 – 2 × 10^6 t are added annually, of which 10 % is of anthropogenic origin [653] (Fig. 108). The inorganic material in aerosols alone is considered to have a total surface area of 2 – 10 times that of the earth [677] and a residence time of 1 – 14 d [678].

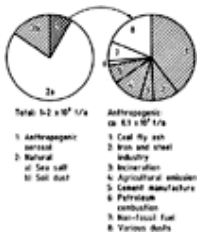

Figure 108. Source strength of the inorganic fraction of the atmospheric aerosol

The anthropogenic portion dominating in industrialized countries is split into various contributions in the right-hand diagram. The hatched areas denote light-absorbing aerosol materials with semiconductor properties.

Trends are of particular interest, especially those that influence anthropogenic activities or are initiated or directly caused by them. Of primary concern today is the increase in CO_2 emissions. The carbon dioxide content of the atmosphere has increased continuously since 1700, and currently by ca. 1 ppm/a. The volume concentration today is 355 ppm, the total amount added annually being 22 × 10^9 t [663]. As with CO_2, the global concentrations of CH_4, H_2, CO, and O_3 are also increasing, but only in the northern hemisphere (Table 28) [662].

Table 28. Concentration trends of chemically reactive trace substances [X]*

X	[X], ppbv	Increase above preindustrialized period, %	Annual increase, %
CO	NH: 100 – 150		NH: 1 [679]
	SH: 50 – 70		SH: ≈0 [679]
CH_4	1700	130	1 [679]
H_2	512		0.6 [679]
O_3	NH: 40 – 45**	160	NH: 1 [679]
	NH: 20 – 30***		SH: ≈0
Nitrate		100 [680]	
Sulfate		100 [680]	
H_2O_2	0.2	50 [681]	
·OH	ca. (2 – 3)×10^{-5} ?		?

* NH = Northern hemisphere; SH = Southern hemisphere.
** Free troposphere.
*** Near ground level.

8.3. Reaction Cycles in the Atmosphere (see also Chap. Physics and Chemistry of the Atmosphere)

The atmosphere contains a total of ca. 5 × 10^{18} kg air, of which 3 × 10^{18} kg make up the troposphere. The reactions above and below the tropopause differ, and there is almost no mass transfer between troposphere and stratosphere. Emissions take part in chemical transformations in the troposphere, and eventually return to the earth's surface, partly by transport reactions, but mainly as mineralized substances.

The atmosphere obtains its energy from solar radiation, which amounts to 3.7 × 10^{23} kW, or 3.2 × 10^{27} kW · h/a. Of this a light flux of 1.6 × 10^{18} kW · h/a is incident on the stratosphere. After passing through the ozone layer, where light is absorbed in the wavelength range below 300 nm, 8 × 10^{17} kW · h/a passes into the troposphere and to the earth's surface.

8.3.1. *The Troposphere*

Material cycles in the troposphere are partly mass-transfer cycles with a high energy requirement and partly chemical transformation processes involving reactive intermediates formed photochemically. These cycles are partly interlinked, especially the water cycle and the carbon cycle.

The amount of water entering the troposphere annually is 4.6 × 10^{17} kg, of which 4.0 × 10^{17} evaporates from the oceans and 0.6 × 10^{17} from the land surfaces. This corresponds to an energy of 3 × 10^{17} kW · h, which is released as heat of condensation on cloud formation and precipitation. The largest part of the carbon cycle is accounted for by CO_2, with a concentration in air of 0.034 vol % at sea level and a turnover of 8.3 × 10^{14} kg/a. Of this, 3.7 × 10^{14} kg CO_2 takes part in a

direct partial cycle with the ocean.

The degradation of organic compounds, as well as other oxidation processes there, in the troposphere is initiated by only a small number of photochemical primary reactions. Very few reactions act as radical sources, as only light of wavelength ≥ 300 nm reaches the lower atmosphere [662].

The fundamental reactions are:

1. Photolysis of nitrogen dioxide at 400 nm

   $$NO_2^{\bullet} \xrightarrow{h\nu} NO^{\bullet} + O:$$

   followed by the formation of ozone

   $$O: + O_2 \xrightarrow{M} O_3$$

   where M is a collision partner, an inert molecule that absorbs energy without itself being chemically changes.
2. Photolysis of ozone at 300 nm with formation of excited oxygen atoms

   $$O_3 \xrightarrow{h\nu} O_2^*(^1\Delta g) + O^*(^1D)$$

   followed by formation of the hydroxyl radical

   $$O_*(^1D) + H_2O \rightarrow 2 \cdot OH$$

3. Oxidation of carbon compounds, ultimately of CO

   $$HO\cdot + CO \rightarrow H\cdot + CO_2$$

   followed by formation of the peroxyhydroxyl radical and regeneration of nitrogen dioxide and hydroxyl

   $$H\cdot + O_2 \rightarrow HO\cdot_2$$

   $$HO\cdot_2 + NO\cdot \rightarrow \cdot OH + NO\cdot_2$$

This regeneration ensures that the oxidation rate in the troposphere is maintained.

Together with VOCs complex reaction mechanisms result.

Radical reactions in the system CH_4 + CO + air + NO_x + $h\nu$ are listed in the following Radical source

$$O_3 + H_2O + h\nu \rightarrow 2 \cdot OH + O_2$$

Radical transformation: HO· ⟶ $HO\cdot_2/CH_3O_2$

$$\cdot OH + CO + O_2 \rightarrow CO_2 + HO\cdot_2$$

$$\cdot OH + CH_4 + O_2 \rightarrow CH_3O\cdot_2 + H_2O$$

$$2\,CH_3O\cdot_2 + O_2 \rightarrow 2\,HO_2 + 2HCHO$$

$$\cdot OH + O_3 \rightarrow HO\cdot_2 + O_2$$

Radical regeneration: $HO\cdot_2/CH_3O_2$ ⟶ ·OH

$$HO\cdot_2 + O_3 \rightarrow \cdot OH + 2\,O_2$$

$$HO\cdot_2 + NO \rightarrow \cdot OH + NO_2$$

$$CH_3O_2 + 2NO + O_2 \rightarrow \cdot OH + 2\,NO_2 + HCHO$$

$$(\text{via } CH_3O\cdot + O_2)$$

Radical loss

$$\cdot OH + NO_2 \rightarrow HNO_3$$

$$HO\cdot_2 + HO\cdot_2 \rightarrow H_2O_2 + O_2$$

$$HO\cdot_2 + CH_3O\cdot_2 \rightarrow CH_3OOH + O_2$$

$$2\,CH_3O\cdot_2 \rightarrow CH_3OH + HCHO$$

Ozone formation

$$NO_2 + O_2 + h\nu \rightarrow NO + O_3$$

Some intermediates react with one another, with already formed photooxidants, or with direct photolysis products of certain VOCs, e.g.,

$$H_2CO \xrightarrow{h\nu} H^\bullet + HCO^\bullet$$

$$HCO^\bullet + O_2 \longrightarrow HO_2^\bullet + CO$$

These radical-loss reactions lead to oxidation products which are washed out by precipitation.

Oxidation of VOCs during the day proceeds via radical chains (Fig. 109 A), and at night via the $NO\cdot_3$ radical (Fig. 109 B).

Figure 109. VOC oxidation by ·OH/·O_2H radical chain

A) Day time chemistry ; B) Night time chemistry

The atmospheric oxygen that takes part in the reaction is present in excess and is not shown in the reaction scheme

The continuous regeneration of oxidizing species consumed in oxidation, especially the ·OH radical, depends basically on the concentration of the nitrogen oxides. With increasing emissions of organic compounds, emissions of NO_x ought also to increase to ensure their degradation [661]:

$$HO\cdot_2 + NO\cdot \rightarrow \cdot OH + NO\cdot_2$$

$$RO\cdot_2 + NO\cdot \rightarrow RO\cdot + NO\cdot_2$$

$$HO\cdot_2 + NO\cdot_3 \rightarrow \cdot OH + NO\cdot_2 + O_2$$

$$RO\cdot_2 + NO\cdot_3 \rightarrow RO\cdot + NO\cdot_2 + O_2$$

In this context, tropospheric ozone plays an ambivalent role. Depending on its concentration and that of other trace substances, especially nitrogen oxides, it is an essential provider of excited oxygen atoms, which form hydroxyl radicals and hence cause degradation of VOCs. However, ozone formed in the same way in photosmog causes formation of the nitrogen-containing components of the photooxidants. At lower NO_x concentrations (i.e., <10 ppt), ozone is consumed, and at higher NO_x concentrations (i.e., up to 10 ppb), O_3 and ·OH are continuously regenerated. Above 10 ppb, ·OH radicals are captured by NO_2 radicals and the chain is broken.

Ozone-forming and ozone-decomposing reactions are summarized in the following:

Ozone formation

$$NO_2 + O_2 + h\nu \rightarrow O_3 + NO$$

$$O_3 + NO \rightarrow O_2 + NO_2$$

$[O_3] = c\,J\,[NO_2]/[NO]$, photochemical equilibrium, J = photolysis frequency

$$RO\cdot_2 + NO \rightarrow RO\cdot + NO_2$$

$$HO\cdot_2 + NO \rightarrow \cdot OH + NO_2$$

These two reactions cause the $[NO_x]/[NO]$ ratio to increase without consuming O_3 Critical competitive reactions which retard the oxidation of NO:

$$HO\cdot_2 + HO\cdot_2$$

$$HO\cdot_2 + RO\cdot_2$$

$$RO\cdot_2 + R'O\cdot_2$$

$HO\cdot_2 + O_3$

$RO\cdot_2 + O_3$

$RO\cdot_2 + NO \rightarrow RONO_2$

Destruction of radicals on surfaces and in raindrops $RO_2 + NO_2 \longrightarrow RO_2NO_2$ (only transitory, as thermal decomposition occurs; stable alkyl nitrates can be formed on surfaces)

		Ozone loss
$O_3 + h\nu$ (λ < 310 nm) $\longrightarrow O(^1D) + O_2$		30 %
$O(^1D) + H_2O \longrightarrow 2 \cdot OH$		
$\cdot OH + O_3 \longrightarrow HO\cdot_2 + O_2$		
$HO\cdot_2 + O_3 \longrightarrow \cdot OH + 2\,O_2$	(HO_x cycle)	30 %
Overall: $2\,O_3 \longrightarrow 3\,O_2$		
$RO\cdot_2 + O_3 \longrightarrow RO\cdot + 2\,O_2$	(with low NO_x)	
Alkenes + $O_3 \longrightarrow$ products		<2 %
$NO_2 + O_3 \longrightarrow NO\cdot_3 + O_2$	(night time)	<1 %
Dry deposition		ca. 25 %

In the case of photochemical degradation in the troposphere above densely populated areas, the lifetimes and residence times of photooxidants are important evaluation criteria together with the emission rates of nitrogen oxides and VOCs [661]. With large-area emission sources such as dense motor traffic and combustion in general, the chemistry and physics of the photosmog form the decisive basis for technical and administrative action.

Some lifetimes τ (298 K, 1 bar) for the reaction $RO_2NO_2 \longrightarrow RO_2 + NO_2$ are:

Ethyl peroxynitrate ($C_2H_5O_2NO_2$)	0.3 s
Acetyl peroxynitrate ($CH_3C(O)O_2NO_2$)	2174 s (0.6 h)
Acetonyl peroxynitrate ($CH_3C(O)CH_2O_2NO_2$)	0.4 s

The mean ozone volume concentration in the free troposphere of the northern hemisphere is 40 ppb (80 µg/m^3), which is three times that of 100 a ago, and is mainly due to anthropogenic emissions of nitrogen compounds. However, these are needed to degrade the increasing emissions of organic compounds, including methane and carbon monoxide, by continuously forming oxidizing species in the troposphere.

The ozone in the troposphere takes part in a complex series of reactions, and is to a degree indispensable to maintain the purity of the lower atmospheric strata. However, these chemical processes and their significance are entirely different from those of the ozone of the stratosphere. If an interaction between the two strata does exist, the possible role of ozone must again be considered. Although this substance is formed differently in the two strata, has different effects, and is not involved in mass-transfer processes, there is an indirect relationship between the ozone in the troposphere and that in the stratosphere (see Section The Stratosphere).

A further contribution to atmospheric photochemical degradation is made by halogen chemistry. Halogenated alkanes and volatile halogenated organic compounds in general are mostly degraded fairly rapidly, similar to chlorine-free organic compounds, though some are degraded relatively slowly [652], [661], [682]. These processes are directly relevant to questions raised by some trace substances which are not degraded particularly rapidly in the troposphere, have high vapor pressures (e.g., fluorinated hydrocarbons; see Section Measurement and Computation of Degradability), and as a result can therefore pass into the stratosphere. Degradable halogen is converted to hydrogen halide, and is washed out of the troposphere by precipitation.

A direct contribution to photochemical degradation in the lower strata of the atmosphere above the oceans and coastal regions (extending as far as 100 km inland, depending on the wind) is made by entrained sea spray. An aerosol is formed and a photochemical reaction, mainly with NaCl, then forms Cl atoms, which degrade organic trace substances in the troposphere more rapidly than · OH radicals [653].

Of the large mass cycles, that of carbon and its compounds forms the foundation for living matter. It is dominated by carbon dioxide [649], which at a concentration of 0.034 vol % in the atmosphere, amounts to 2.6×10^{12} t, with an annual turnover of 830×10^9 t and a mean residence time of 3.3 a.

The turnovers of photochemical partial cycles are comparatively small.

Entry of carbon compounds into the atmosphere (in 10^9 t) is as follows:

CO_2 from respiration of biomass	220	
CO_2 from putrefaction of biomass	220	
(formation on biomass by photosynthesis)		440 (removal)
CO_2 from the sea	365	
Returned to the sea (equilibrium)		365 (removal)
CO_2 from forest clearance	7	
CO_2 from combustion processes	18	
CO_2, total emission	830	
CO, total emission	3.4	
CH_4, total emission	0.5	
C (in VOCs), total emission	1.0	

Worldwide efforts to reduce CO_2 emission are intended to

1. Reduce the greenhouse effect
2. Save energy
3. Conserve resources of carbon-containing fossil fuels, partly to retain them as chemical raw material

Since, the majority of anthropogenic CO_2 emissions are produced by combustion processes, which simultaneously produce NO_x, a decrease in CO_2 would also be associated with a decrease in NO_x and hence in tropospheric ozone. Decreases in the concentration of hydrocarbons and nitrogen oxides would immediately lead to a noticeable reduction in the concentration of ozone, as its lifetime in the lower atmosphere is short.

8.3.2. *The Stratosphere*

The most important trace gas in the stratosphere is ozone. It is contained in a layer 25 –30 km thick in the lower stratosphere at a latitude-dependent concentration of 10^{12} cm^{-3}. This accounts for 90 % of the total atmospheric ozone [657, 658, 679, 683, 684]. Ozone is mainly produced above the equatorial regions and is transported towards the polar regions. Its existence depends on a delicate equilibrium between formation and degradation, and it is essential that its concentration in the stratosphere should be maintained for the benefit of climatic and living conditions on the earth's surface. Stratospheric photochemical degradation of volatile substances emitted into the troposphere but are degraded so slowly there that they migrate to the stratosphere would lower the stratospheric ozone concentration and lead to impaired conditions at the earth's surface.

Ozone is formed in the stratosphere exclusively by the photolysis of oxygen.

$$O_2 \xrightarrow[\lambda \leq 250\ nm]{h\nu} 2\,O$$

$$O + O_2 \xrightarrow{M} O_3$$

The inert collision partner M is provided from the surrounding supply of O_2 or N_2 molecules. Photolytic degradation then follows, but caused by light of a longer wavelength

$$O_3 \xrightarrow[\lambda \geq 300\ nm]{h\nu} O_2 + O$$

$$O + O_3 \longrightarrow 2\,O_2$$

A chemical degradation process also exists

$$Y + O_3 \rightarrow O_2 + YO$$

$$YO + O \rightarrow O_2 + Y$$

The net reaction is thus

$$O + O_3 \rightarrow 2\,O_2$$

In the natural stratosphere, Y = ·OH and YO = $HO\cdot_2$. The natural ozone concentration is formed in this way.

The exhaust gases from supersonic aircraft, rockets, and the undecomposed components of emissions from the earth's surface ascending through the troposphere can provide further reactive species that can take part in ozone degradation and affect the above-mentioned equilibrium concentration:

$$Y = NO, Cl\cdot, Br\cdot$$

$$YO = NO_2, ClO\cdot, BrO\cdot$$

In this system, the transformation of ·OH into $HO\cdot_2$ and back again takes place very rapidly (Fig. 110).

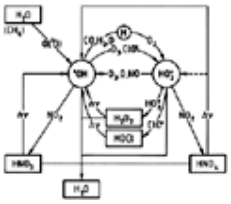

Figure 110. HO_x chemistry of the stratosphere

H_2O_2 and HOCl are formed and act as temporary sinks, while nitric acid and water act as permanent sinks. If these degradation reactions consume ozone more rapidly than it is regenerated or replaced by transport mechanisms, which is regularly the case in the region of the antarctic polar cap, this leads to a transitory but periodically recurring decrease in the concentration of stratospheric ozone (Fig. 111).

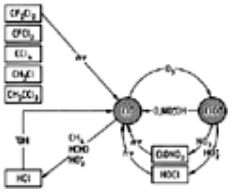

Figure 111. Scheme of the chlorine-induced cycle causing degradation of ozone in the global stratosphere

This can lead to a 50 % depletion of the normal ozone layer over an area the size of Antarctica — the so-called ozone hole (Fig. 112).

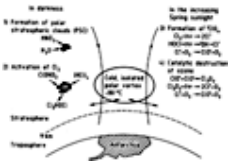

Figure 112. Meteorology and chemistry of antarctic ozone depletion

The chlorine compounds HOCl, $ClONO_2$, HCl, and elemental chlorine are formed from ascending CFCs and chlorine injections from the boosters of rocket carriers [685].

The ozone hole starts to form at the end of the antarctic winter by mid September and disappears again by mid November. It is preceded by a rapid cooling of the stratosphere to – 80 °C in June and July. This causes formation of polar stratospheric clouds (PSCs), one type of which consists of nitric acid, hydrochloric acid, and water, and the other exclusively of water. The latter type is larger and leads to precipitation and hence dehydration of the polar stratosphere. In a reaction that takes place at the surfaces of the PSCs, $ClONO_2$ forms Cl_2, HOCl, and HNO_3. This amounts to condensation of the nitrogen species into sinks but to a redistribution of the chlorine species from less active sinks into more reactive species (see Figs. 110 and 111).

The degradation of emissions that have entered the stratosphere also leads to the degradation of the stationary reaction partner ozone. The direct result is an increase in the UV-B radiation at the earth's surface, and an indirect result is increased permeability of the stratosphere towards other emissions ascending from the troposphere. This would be enabled by cooling of the stratosphere due to the loss of the ozone.

8.4. Interaction with Surface Layers of Aquatic and Terrestrial Environments

Under the influence of sunlight, active oxidizing agents are formed in water [686] with the following concentrations (mol/L):

Hydroperoxyl ($HO\cdot_2$)	$10^{-9} – 10^{-8}$
Peroxyalkyl (ROO·)	$10^{-11} – 10^{-10}$
Singlet molecular oxygen	$10^{-14} – 10^{-13}$
Hydrated electron	$10^{-17} – 10^{-15}$
Hydroxyl (·OH)	

$10^{-18} – 10^{-19}$

Of these, the hydroxyl radical, with 10^6 radicals per cubic centimeter down to a depth of ca. 1 m [687], is the most significant for photochemical degradation in water as well as in air. The concentration is comparable to that in air, but can be lower in pure water or pure natural water bodies [688]. In natural surface water, the most important source of ·OH is the photolysis of nitrate.

Dissolved or dispersed organic compounds or inorganic compounds in lower oxidation states are also oxidized in the aqueous phase by indirect photochemical degradation. These processes are similar to those that take place in the mixture of hydrogen peroxide and iron(II) salts known as Fenton's reagent, in which intermediate stages also involve radicals (see ⟶ Wastewater). Reactions of this type can already take place with constituents of natural surface water. For example, in seawater the bromide ion is oxidized to hypobromite, and in the layers near the surface of fresh water the humic matter present at a concentration of 1 – 10 mg/L is oxidized to form oxidizing radicals such as RO·, $RO·_2$, and $HO·_2$.

Most rate constants for the reaction of organic compounds with ·OH radicals in water have values of the order of $k_{OH} \approx 10^9$ L mol^{-1} s^{-1} [689]. The rate constants of other reactions show a similar scatter around the mean figure [687], e.g., K_{OH} values in air (see Section Lifetime in the Homogeneous Gas Phase).

k_{OH} (L mol^{-1} s^{-1}) values for some compounds in water follow:

Urea	2.1×10^6
Acetonitrile	3.5×10^6
Chloroform	8.5×10^6
Ethylene oxide	4.0×10^7
Methanol	8.9×10^8
Thiophene	3.3×10^9
Phenol	1.6×10^{10}
Aniline	1.7×10^{10}

The dependence on solar radiation is more pronounced than in the case of air (Table 29) [688].

Table 29. Lifetime of ·OH in surface (fresh) water

Substance	τ (sun) [a], h	τ (average) [b], h
Aniline, phenol	14, 5	6.37×10^3
Methanol	320	10.43×10^4
Chloroform	33 000	12.46×10^6

[a] Optimum conditions (intense sunlight, at water surface) [·OH] ≈ 10^{-15} mol/L.
[b] With annual average values ⟨[·OH]⟩ ≈ 3×10^{-18} mol/L. The time for (a) is hours of intense sunlight, for (b) it is "real" time.

In water droplets, fog, dew, etc., or on the earth's surface, the situation is similar, and the substances in these environmental compartments do not have a high vapor pressure. In the vapor-pressure range from 10^{-7} to 1 Pa they are referred to as semivolatile organic compounds (SOCs) [690]. For the photochemical degradation of these substances, not only the formation but also the distribution of the photooxidants is of direct relevance [691]. In the interaction between biotic and abiotic degradation steps of the aquatic systems of rivers, lakes, and wet ground, and mass transfer into the troposphere at ground level, the photochemical degradation of organic substances in water makes a contribution to the total mineralization [654, 656, 687, 690, 691].

8.5. Measurement and Computation of Degradability

To derive ecological relationships from the properties of chemical substances, their biotic and abiotic interactions must be considered. The behavior of a substance in the environment is determined and assessed by drawing on the results of ecotoxicological investigations (see ⟶ Ecology and Ecotoxicology). However, only since the late 1970s has it been possible to answer a fundamental aspect of this question, i.e., how long a substance exists in the atmospheric environment. This came with the development of reliable measurement techniques and equipment.

With regard to the biotic degradation pathways in the carbon cycle (see Chap. Physics and Chemistry of the Atmosphere), the reaction chains, mechanisms, and rates in the water and soil environmental compartments are well known in many cases, and provide a working knowledge for science, economy, administration, and legislation. An abiotic process in the

atmosphere was only confirmed with the development of laboratory methods that supplemented earlier field tests of mainly meteorological interest. Besides benefiting from progress in other fields, especially the various types of laser spectroscopy, a number of techniques have been developed specifically for investigating photodegradation.

In the environmental compartment troposphere, degradability (or its converse, chemical persistence) is characterized by abiotic processes which are photochemically induced and then proceed in oxidative steps. Degradability is determined by

1. The concentration of the oxidizing agent in the troposphere, sometimes in a particular locality and at a particular time
2. The rate constants of the reaction with a reaction partner which is volatile and homogeneous in the gas phase or is bonded to particles as an aerosol

There is now much data on both factors. A number of experimental methods have been developed for determining rate constants, and figures for K_{OH} are now available for 600 compounds in air, as are rate constants for other oxidizing agents in other media and in heterogeneous phases. Rate constants for photochemical degradations are available in spreadsheets [693-695] and data banks [696]. Values for K_{OH} can also be derived from bond energy increments of individual structural groups and classes of substances [697-699].

The chemical lifetime is found from the rate constants of bimolecular reactions. The degradation rate of substance X

$$-\frac{\mathrm{d}\,c_{(X)}}{\mathrm{d}t} = k\,c_X c_Y$$

and the rate constant k are both expressed as cm^3 per molecule per second (cm^3/s). It depends on temperature and pressure. It is simplified by incorporating c_Y, the time-averaged concentration of the oxidizing agent, into the constant, and a new monomolecular rate constant K is derived:

$$K = k\,c_Y$$

The degradation rate is then given by

$$-\frac{\mathrm{d}\,c_X}{\mathrm{d}t} = K\,c_X$$

Sometimes, the less obvious term “chemical half-life” ($\tau_{1/2} = 0.69\ \tau_X$) is used, i.e., the time for the concentration of the starting material to reach half its initial value. However, this is dependent on the physicochemical environment of a substance, and is thus not a property inherently characteristic of a substance like the half-life for radioactive decay, which is unaffected by outside influences. The residence time in the atmosphere is usually smaller than the chemical lifetime because in some cases part of the undecomposed material can be removed by transport processes.

The main primary oxidizing agents in the troposphere are ·OH, $HO\cdot_2$, O_3, and $NO\cdot_3$. The ·OH radical is present at rather low concentration, but reacts so much more rapidly than the other oxidizing agents that it has been found possible to deduce the chemical lifetime of a substance in the troposphere solely from its rate constants [690].

The limits of measurability are:

K_{OH} values	Corresponding chemical lifetime
10^{-10} cm^3/s for rapid reactions	1 h
10^{-16} cm^3/s for slow reactions	130 a

To measure rate constants of the photochemical degradation of a substance in the atmosphere, a vapor pressure of >1 Pa is required in the laboratory to give a sufficiently high concentration of the substance in the photoreactor [689], [699]. Nonvolatile compounds are first adsorbed on an inorganic carrier, and the adsorbate is then investigated as an aerosol in a smog chamber [653], [672]. This method has the advantage that it simulates a real dust-laden atmosphere, and any catalytic acceleration of the photochemical degradation by traces of heavy metals in the carriers can be measured. This also applies to the use of TiO_2 as a carrier for the test aerosol [672].

If only the rate constant K_{OH} is to be determined, this can be carried out in aqueous solution with H_2O_2 as the source of ·OH radicals [689] at a stationary concentration of $c_{OH} = 5 \times 10^{-14}$ mol/L or 4×10^7 cm^{-3}. For SOCs, which usually are only sparingly soluble in water, 1,1,2-trichlorotrifluoroethane can be used as solvent [692], with H_2O_2 as ·OH source.

A guideline for the standardization and international harmonization of test methods and standards for photochemical degradation in air [699] and water [700] has been produced by the OECD countries. Methods for the derivation of rate constants from structural increments have also been produced.

8.5.1. *Lifetime in the Homogeneous Gas Phase*

The various important naturally-occurring oxidizing agents can be present in widely different concentrations. The effects of

the intensity and duration of the solar radiation can also differ widely (Table 30) [690].

Table 30. Concentration of oxidizing agents

Molecules/cm^3	Day	Day/night	Night
·OH	1×10^6	0.5×10^6	
O_3	1×10^{12}		
NO·$_3$	instable	1.5×10^8	2.5×10^8

The NO·$_3$ radical is rapidly photolyzed and is therefore mainly active at night. It forms nitric acid by hydrogen abstraction and undergoes addition to double bonds.

In the medium lifetime range (ca. 10 d) the fluctuation in ·OH concentration caused by irregular radiation is sufficient to cause significant differences in rate of degradation (Table 31) [661], [687].

Table 31. Chemical lifetimes of some volatile organic compounds in the troposphere*

Compound	Chemical formula	*k* values for ·OH, 10^{-12} cm^3/s	Lifetime, d	
			Tropics	Global
Ethane	C_2H_6	0.29	20	61
Methanol	CH_3OH	0.80	6.5	15
Acetylene	C_2H_2	0.90	5.5	13
Benzene	C_6H_6	1.31	4.5	12
Propane	C_3H_8	1.59	3.6	11
n-Butane	C_4H_{10}	2.4	2.1	7.5
Isobutane	C_4H_{10}	2.3	2.2	8
Dimethyl ether	CH_3OCH_3	2.6	1.9	7
Ethyl alcohol	C_2H_5OH	3.0	1.8	5.9
n-Pentane	*n*-C_5H_{12}	3.7	1.5	4.8
Toluene	$C_6H_5CH_3$	6.19	0.9	2.8
Ethylbenzene	$C_6H_5C_2H_5$	7.73	0.75	2.5
Ethene	C_2H_4	7.85	0.7	1.9
Formaldehyde	HCHO	10	0.5	0.9
o-Xylene	$C_6H_4(CH_3)_2$	14	0.35	0.7
Acetaldehyde	CH_3CHO	16	0.3	0.6
Propene	C_3H_6	29	0.2	0.4

* ·OH concentration: 2×10^6 cm^{-3} at 24 h mean in tropical regions, 6.5×10^5 cm^{-3} as global mean; O_3 concentration: 30 ppbv.

A relationship between K_{OH} values and chemical structure is only observed with homologues [653], [693]. The first member reacts much more slowly than the rest (Table 32). Photochemical degradation by ozone also shows no dependence on chemical structure [653], [701]. Alkenes and other substances with olefinic double bonds are significantly degraded by ozone, but other compounds containing multiple bonds and aromatics are degraded rather slowly (Table 33).

Table 32. Rate constants k (in 10^{-12} cm^3 s^{-1}) for the reactions of ·OH with various organics at room temperature

Compound	Ref. [653]	Ref. [693]	Compound	Ref. [653]	Ref. [693]
Propane	1.3	1.18	2,2-Dimethylhexane	4.9	
n-Butane	2.5	2.53	2,2,4-Trimethylpentane	3.4	3.66
n-Pentane	4.1	4.04	2,3,5-Trimethylhexane	7.9	

n-Hexane	5.7	5.58	Cycloheptane	11.7	13.1
n-Heptane	7.3	7.2	Cyclooctane	14.1	
n-Octane	8.8	8.72	1,1,3-Trimethylcyclohexane	8.7	
n-Nonane	10.3	10.0	Adamantane	21.9	22.7
n-Decane	12.4	11.2	Propene	28	26.3
n-Undecane	13.3		Acrylonitrile	4.3	4.8
n-Dodecane	13.9		Propionitrile	0.8	0.19
n-Tridecane	15.5		Butyronitrile	1.1	
Isobutane	2.5	2.37	Benzene	1.3	1.28
Isopentane	4.0	3.9	Toluene	5.6	6.19
3-Methylpentane	5.7	5.6	*o*-Xylene	10.3	14.7
4-Methyloctane	9.7		*p*-Xylene	11.6	15.2
2-Methyloctane	10.1		Chlorobenzene	1.0	0.7
2,3-Dimethylbutane	5.8	6.2	Vinyl chloride	5.6	6.6
2,3-Dimethylpentane	6.0	5.2	1,1-Dichloroethene	7.0	
2,4-Dimethylhexane	7.8		1,2-Dichloroethene	2.2	
2,2-Dimethylbutane	2.4	2.59	Trichloroethene	2.2	2.36
2,2-Dimethylpentane	3.4		Carbon tetrachloride	<0.06	
3,3-Dimethylpentane	3.7				

Table 33. Rate constants, k (in 10^{-17} cm^3 s^{-1}), for the reaction of ozone with various organics at room temperature

Compound	Ref. [653]	Ref. [701]
Propene	1.1	1.13
Isobutene	1.3	1.21
1-Pentene	0.95	1.1
1-Hexene	1.1	1.17
1-Heptene	1.15	0.8 – 1.7
1-Octene	1.25	0.8
1-Nonene	1.45	
Vinyl chloride	0.014	0.017 – 0.2
Acrylonitrile	0.002	
Propionitrile	<0.001	
Butyronitrile	<0.001	
o-Xylene	<0.007	7×10^{-5}
p-Xylene	<0.01	4×10^{-5}

The rates of degradation of chlorinated hydrocarbons also do not show any correlation with structure [702]. They include very slowly degraded substances with lifetimes of several years and typical ones with lifetimes of around 14 d (Table 34).

Table 34. Tropospheric concentrations and lifetimes of some chlorinated hydrocarbons

Compound	Troposheric concentrations, pptv	Tropospheric lifetimes, a
C_2HCl_3	16	0.04
C_2Cl_4	40	0.4
$CHCl_3$	15	1
CH_3Cl	600	2 – 3
CH_3CCl_3	110	5 – 10
CCl_4	120	60 – 100

The large number of K_{OH} values known show that degradability rate constants of organic compounds can have a range of six orders of magnitude. Most organic compounds have a photochemical lifetime within a range of two orders of magnitude, i.e., a lifetime of 1 – 100 d. Some are degraded within hours, and a few require more than a year (Fig. 113). If

the lifetime exceeds 10 a, ascent from the troposphere into the stratosphere has to be taken into consideration [661].

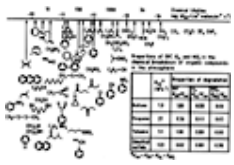

Figure 113. Chemical lifetimes of organic compounds in the atmosphere

Standardized graphical and tabular presentations are usually based on a concentration of 0.5 × 10^6 ·OH radicals/cm^3 as a basis for K_{OH} values and lifetimes.

8.5.2. *Lifetimes of Adsorbates Aerosols*

A portion of natural and anthropogenic emissions reach the troposphere bound to aerosol particles. Experimental investigations in an aerosol/smog chamber with a volume of 2.4 m^3 [653], [672] have shown that the degradation rate of nonvolatile substances transported into the atmosphere in adsorbed form is comparable to that of volatile substances in a homogeneous phase.

The aerosols have a multilayer structure. Surrounding a water-insoluble core of, for example, desert dust, fly ash, volcanic ash, is a salt-containing aqueous envelope onto which the organic trace substances are adsorbed, with the hydrophilic ends of the molecules directed inwards. The hydrophobic groups directed outwards hinder the penetration of hydrophilic gases such as acids and tend to keep the aerosol in suspension [704].

Most of the global total of (1 – 2) × 10^9 t of aerosol is formed by sodium chloride from sea spray. Some of the chlorine is transformed photolytically to chlorine-containing reactive species. Hence a photo-induced degradation of emissions in the troposphere also takes place in this way. The reaction can be measured in a smog chamber [653], [672]. The k_{Cl} rate constants found are in agreement for various Cl sources [703]. The degradation process is about two orders of magnitude more rapid than that via OH· radicals (Table 35).

Table 35. Rate constants k (in 10^{-10} cm^3 s^{-1}) for the reaction of chlorine atoms with various organics

Compound	Source*			Ref. [703]
	I [653]	II [653]	III [653]	
Propene	1.09		1.02	1.34
n-Butane**	1.97	1.97	1.97	1.97
n-Pentane	2.65	2.50	2.35	2.52
n-Hexane	3.15	3.00	3.20	3.03
n-Octane	4.25	3.65		
n-Nonane	4.40			
Isobutane	1.35		1.30	1.37
Isopentane	2.10		1.75	2.03
3-Methylpentane	2.70		2.80	
2,2-Dimethylbutane	1.60	1.75		
2,3-Dimethylpentane	3.00			
2,2-Dimethylhexane	3.30			
2,2,4-Trimethylpentane	2.25	2.25		
2,2,3,3-Tetramethylbutane		1.55		
Propene	3			2.44
Vinyl chloride	2.05	1.55	1.90	1.97
1,1-Dichloroethene	2.10			
Benzene	0.05			0.15
Toluene	0.74	0.66		0.589

* Source of Cl: NaCl aerosol in photosmog (I); photolysis of Cl_2 (II), photodegradation of chlorobenzene, vinyl chloride, trichloroethene and 1,1-dichloroethene (III).
** Reference compound.

A further decrease in the lifetime is observed if the airborne dusts of the aerosols contain heavy metals, have semiconducting properties, or can react in redox systems. With pure titanium dioxide as the carrier particle, the ·OH concentration is increased by a factor of ten, and the degradation rate is correspondingly increased (Fig. 114) [653], [672].

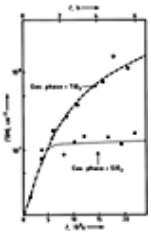

Figure 114. Typical formation of ·OH in an aerosol smog chamber in the presence of SiO_2 or TiO_2 aerosol

In the absence of aerosol the formation of ·OH is similar to the curve observed in the presence of SiO_2

This reaction principle is also used in aqueous suspensions [705] (see ⟶ Wastewater). Multiphase systems are important in addition to airborne aerosols in the earth – air boundary region (see Section Reaction Cycles in the Atmosphere).

8.5.3. *Assessment and Evaluation*

Photochemical degradation is of major and direct importance because it eliminates substances that enter the atmosphere as emissions from natural and anthropogenic sources or from products used in the open. Some of these substances are considered to affect the climate, others to be detrimental to the filtering effects of the ozone layer on the UV-B fraction of sunlight, and many others to be pollutants of soils and bodies of water on the earth's surface, from where they can return to the atmosphere.

The question of the whereabouts of a substance in a given environment [687], [706] is largely tantamount to enquiring about its compatibility with conditions prevailing in that particular environmental compartment; holistic considerations contribute greatly [707-710].

Moreover, it is processes much more than states which furnish the determinants. Already the environmental concentration of a substance is merely the difference between two rates, that of input and that of output. In that sense, cross sections of an environment are better expressed by mass and energy fluxes than by properties of any particular substance or a group of substances.

Data on individual substances are much less suitable for describing the status of the environmental compartment “air”. Instead they express the oxidative sensitivity of a substance in sunlight as a first step towards mineralization [711].

[Top of Page]

9. Acknowledgement

The entire topic was coordinated by John Wolf

[Top of Page]

References

1. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (eds.): *Beschluß der Bundesregierung vom 7. November 1990 zur Reduzierung der CO_2-Emissionen in der Bundesrepublik Deutschland bis zum Jahr 2005*, Druck-Service E. Böhm, Hansen 1991.
2. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (eds.): *Klimaschutz in Deutschland, Nationalbericht der Bundesregierung für die Bundesrepublik Deutschland im Vorgriff auf Artikel 12 des Rahmenübereinkommens der Vereinten Nationen über Klimaänderungen*, 1993.
3. Dritter Bericht der Enquete-Kommission “Schutz der Erdatmosphäre” – “Klimaschutz durch umweltgerechte Landwirtschaft und Erhalt der Wälder”, Drucksache 12/8350, 1994.
4. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (eds.): *Bericht der Bundesregierung an den Deutschen Bundestag*, Fünfter Immissionsschutzbericht der Bundesregierung, 1993.
5. Umweltbundesamt (UBA) (eds.): *Daten zur Umwelt 1990/91*, Erich Schmidt Verlag, Berlin 1992.
6. Bericht der Enquete-Kommission “Schutz des Menschen und der Umwelt – Bewertungskriterien und Perspektiven für umweltverträgliche Stoffkreisläufe in der Industriegesellschaft”, Drucksache 12/8260, 1994.
7. P. Davids, M. Lange. *Die TA Luft 1986 – Technischer Kommentar,* VDI-Verlag, Düsseldorf 1986; *Die Großfeuerungsanlagen – Technischer Kommentar,* VDI-Verlag, Düsseldorf 1984.
8. Verband der Chemischen Industrie e.V.: *Chemiewirtschaft in Zahlen*, V. edition, 1993, M. Erhardt KG, Frankfurt.
9. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (eds.): *Zweiter Bericht der Bundesregierung an den Deutschen Bundestag über Maßnahmen zum Schutz der Ozonschicht*, 1992.
10. Der Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (eds.): *Tagungsband Alternativen zu FCKW und Halonen*, Berlin, Feb. 24 –26, 1992.
11. VDI 3455: Anlagen zur Serienlackierung von Automobilkarossen, Beuth-Verlag, Berlin 1996.
12. O. Lenz, Verband der Chemischen Industrie e.V. (VCI), personal communication, 1994.
13. Climate Change Secretariat, P.O.Box 260124, D-53153 Bonn, Germany; website: http://www.unfccc.de.
14. Verantwortliches Handeln. Daten zu Sicherheit, Gesundheit, Umweltschutz, Bericht 1996, Verband der Chemischen Industrie, Frankfurt.
15. Daten zur Umwelt, Der Zustand der Umwelt in Deutschland, Ausgabe 1997, Umweltbundesamt, Erich Schmidt

Verlag, Berlin, 1997.

16. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit, Umweltpolitik — 10 Jahre Montrealer Protokoll, Köllen Verlag, Bonn 1997.
17. Air Pollution in Europe, European Environment Agency, Kongens Nytorrg, DK-1050 Copenhagen; homepage: http://www.eea.eu.int.
18. Europe's Environment: The Second Assesment — An Overview, EEA, Copenhagen, 1998.
19. Atmospheric Emission Inventory Guidebook, 1st ed.; homepage: http://tiger.eea.eu.int/aegb/.
20. J. v. Iribarne, H. R. Cho: *Atmospheric Physics*, D. Reidel Publ., Dortrecht 1980.
21. P. Fabian: *Atmosphäre und Umwelt*, Springer Verlag, Berlin 1984.
22. C. D. Schönwiese, B. Dieckmann: *Der Treibhauseffekt*, Rowohlt-Verlag, Hamburg 1989.
23. R. P. Hosker, Jr. in D. Randerson (ed.): “Flow and Diffusion near Obstacles, Atmospheric Science and Power Production,” *Tech*, Inf. Cent. U.S. Dep. Energy "DOE/TIC-27 601 (DE 84 005 177)".
24. A. Knaps, G. Manier, R. Röckle: *Untersuchung der Strömungsverhältnisse im Bereich von Industrieanlagen zur Beurteilung von Störfallemissionen*, Bundesministerium für Forschung und Technologie, no. RGB 8606, Darmstadt 1989.
25. L. F. Richardson: “The Supply of Energy from and to Atmospheric Eddies,” *Proc. R. Soc. London A* **97** (1920) 354 – 373.
26. A. M. Obukhow: “Turbulence in an Atmosphere with Non-uniform Temperature,” *Boundary Layer Meteor.* **2** (1946) 7 – 29.
27. G. Manier: “Vergleich zwischen Ausbreitungsklassen und Temperaturgradienten,” *Meteorol. Rundsch.* **28** (1975) 6 – 11.
28. H. E. Cramer, F. A. Record, H. C. Vaughan: “The Study of the Diffusion of Gases in the Lower Atmoshere,” M.I.T. Dept. of Meteorology, *Final Report under Contract,* no. AF 19 (604) 1058. J. S. Hay, F. Pasquill: “Diffusion from a Continuous Source in Relation to Spectrum and the Scale of Turbulence. Atmospheric Diffusion and Air Pollution,” in F. N. Frenkiel, P. A. Sheppard (eds.), *Adv. Geophys.* **6** (1959) 345. K. Vogt, H. Geiss, G. Polster: “New Sets of Diffusion Parameters Resulting from Tracer Experiments with 50 and 100 m Release Hight,” *Proc*, 9th Int. Meet. Air Poll. Modelling Appl. **No. 103**, Nato Comm. (1978) 221 – 239.
29. S. Blinda, A. Hopf, G. Manier, K. Marburg: Entwicklung und Verifizierung eines Rechenmodells zur Simulation der Ausbreitung von Kfz-Emissionen im Bereich komplexer Gebäudekonfigurationen. Im Auftrage des Ministers für Umwelt, Raumordnung und Landwirtschaft des Landes Nordrhein-Westfalen, Institut für Meteorologie, Darmstadt 1993.
30. Technische Anleitung zur Reinhaltung der Luft (TA-Luft), Erste Allgemeine Verwaltungsvorschrift zum Bundesimmissionsschutzgesetz, C. Heymanns Verlag, Köln 1980.
31. Zweite Allgemeine Verwaltungsvorschrift zur Störfall-Verordnung (2. Störfall VwV), April 27, 1982, Gemeinsames Ministerialblatt A 33, no. 14, pp. 205 – 216.
32. VDI 3783, Blatt 2: Umweltmeteorologie – Ausbreitung von störfallbedingten Freisetzungen – schwerer Gase – Sicherheitsanalyse, Beuth Verlag, Berlin 1990.
33. VDI 3783, Blatt 1: Umweltmeteorologie – Ausbreitung von störfallbedingten Freisetzungen – Sicherheitsanalyse, Beuth Verlag, Berlin 1987.
34. VDI 3782, Blatt 4: Umweltmeteorologie – Ausbreitung von Geruchsstoffen in der Atmosphäre, Beuth Verlag, Berlin 1994.
35. H. Schirmer, W. Kuttler, J. Löbel, K. Weber: *Lufthygiene und Klima. Ein Handbuch zur Stadt- und Regionalplanung*, VDI-Verlag, Düsseldorf 1993.
36. VDI 3958, Blatt 1: Umweltsimulation, Grundlagen und Methoden, Beuth Verlag, Berlin 1994.
37. H. Behrendt in: N. H. Seemayer, W. Hadnagy (eds.) *Environmental Hygiene* II, 1991 , 253 – 256.
38. R. Doll, R. Peto, *JNCI J. Natl. Cancer Inst.* **66** (1981) 1191 – 1308.
39. M. Gough, J. *Am. Chem. Soc.* **23** (1989) 925 – 930.
40. T. Eikmann in W. Bachmann (ed.): *Das Grüne Gehirn*, Schulz-Verlag, Starnberg 1991, E 2, pp. 1 – 57.
41. D. Sheppard, *Am. Rev. Respir. Dis.* **123** (1981) 486–491. Links
42. G. Ericsson, P. Canner, *Scand. J. Work Environ. Health* **9 Suppl. 3** (1983) 1 – 52.
43. D. S. Dockery, *J. Air Pollut. Control Assoc.* **32** (1982) 937 – 942. Links
44. A. Cohen, A. Nelson, C. J. Bromberg, *Am. J. Public Health* **64** (1974) 442 – 449. Links
45. S. Tani, *Nippon Koshu Eisei Zasshi* **22** (1975) 431 – 438.
46. H. E. Wichmann, D. Sugiri, G. Herold, E. Knülle, *Zentralbl. Bakteriol Mikrobiol. Hyg. Abt. 1 Orig.* **B 185** (1988) 509 – 519.
47. H. E. Wichmann, C. Spix, G. Mücke: *Bericht für den Minister für Arbeit, Gesundheit und Soziales*, Medizinisches Institut für Umwelthygiene, Düsseldorf 1987.
48. Technische Anleitung zur Reinhaltung der Luft – TA Luft vom 27. 02. 1986, Erste Allgemeine Verwaltungsvorschrift zum Bundes-Immissionsschutzgesetz der Bundesrepublik Deutschland.
49. VDI-Richtlinie Schwefeldioxid, no. 2310, Blatt 11, Verein Deutscher Ingenieure, Düsseldorf 1984.
50. WHO, Air Quality Guidelines for Europe, Europ. Series no. 23, Copenhagen 1987.
51. H.-W. Schlipköter, N. Manolovic, R. Dolgner, *Luftverunreinigung*, Medizinisches Institut für Umwelthygiene, Jahresbericht 1985, Düsseldorf 1985, 1 – 6.
52. D. B. Menzel, in T. Schneider, L. Grant (eds.): *Air Pollution by Nitrogen Oxides*, Elsevier Scientific Publ., Amsterdam 1982, pp. 417 – 426.

53. WHO, Environ. Health Criteria no. 4, Geneva 1977.
54. VDI-Richtlinie Stickstoffoxid, no. 2310, Blatt 12, Verein Deutscher Ingenieure, Düsseldorf 1985.
55. EPA – U.S. Environmental Protection Agency, Air Quality Criteria for Ozone and Other Photochemical Oxidants, 1986.
56. VDI-Richtlinie Ozon (und photochemische Oxidantien), no. 2310, Blatt 15, Verein Deutscher Ingenieure, Düsseldorf 1987.
57. WHO, Environmental Health Criteria 13, Geneva 1979.
58. W. S. Aronow, M. W. Isbell, *Ann. Intern. Med.* **79** (1973) 392 – 395. Links
59. W. S. Aronow, J. Cassidy, *Ann. Intern. Med.* **83** (1975) 496 – 499. Links
60. W. S. Aronow, E. A. Stemmer, M. W. Isbell, *Circulation* **49** (1974) 415 – 417. Links
61. VDI-Richtlinie Kohlenmonoxid, no. 2310, Blatt 13, Verein Deutscher Ingenieure, Düsseldorf (Entwurf) 1992.
62. VDI-Richtlinie Schwebstaub, no. 2310, Blatt 19, Verein Deutscher Ingenieure, Düsseldorf 1992.
63. J. W. B. Douglas, R. W. Waller, *Br. J. Prev. Soc. Med.* **32** (1966) 1 – 8.
64. ZEBS – Zentrale Erfassungs- und Bewertungsstelle für Umweltchemikalien des Bundesgesundheitsamtes, *ZEBS-Hefte* **1** (1984).
65. WHO, Environmental Health Criteria no. 3, Geneva 1977.
66. G. Winneke, A. Brockhaus, U. Ewers, U. Kauer, M. Neut, *Neurotoxicol. Teratol.* **12** (1990) 553 – 559. Links
67. U. Ewers, H.-W. Schlipköter in E. Merian (ed.): *Metalle in der Umwelt*, Verlag Chemie, Weinheim, Germany 1984, pp. 351 – 374.
68. Verein Deutscher Ingenieure, *VDI Ber.* **203** (1974).
69. DFG – Deutsche Forschungsgemeinschaft, VCH-Verlagsgesellschaft, Weinheim 1992.
70. LAI – Länderausschuß für Immissionsschutz: *Krebsrisiko durch Luftverunreinigungen,* Düsseldorf 1992.
71. T. Eikmann, M. Kramer, H. Goebel, *Zbl. Hyg.* **193** (1992) 41 – 52.
72. T. Eikmann in H. Wichmann, H.-W. Schlipköter, G. Füllgraf (eds.): *Handbuch der Umweltmedizin*, ecomed-Verlag, Landsberg 1992.
73. F. Pott, *Staub Reinhalt. Luft* **45** (1985) 369 – 379.
74. BGA – UBA – Bundesgesundheitsamt – Umweltbundesamt, Bundesgesundheitsbl., Sonderheft 1993.
75. H. Beck, *Bundesgesundheitsbl.* **33** (1990) 99 – 104.
76. J. Abel, *VDI Ber.* **634** (1987) 487 – 501.
77. Y. P. Drogan *et al.*, *Carcinogenesis* **13** (1992) 1389–1395. Links
78. J. A. Stöckhardt: “Untersuchungen über die schädlichen Einwirkungen des Hütten- und Steinkohlenrauches auf das Wachstum der Pflanzen, insbesondere der Fichte und Tanne,” *Forstl. Jahrb.*, 1871, p. 218 ff.
79. K. Garber: *Luftverunreinigungen und ihre Wirkungen*, Borntraeger, Berlin-Nikolassee 1967.
80. R. Guderian: *Air Pollution Phytotoxicity of Acidic Gases and its Significance in Air Pollution Control, Ecological Studies,* vol. 22, Springer Verlag, Berlin 1977.
81. Umweltbundesamt Berlin (eds.): *Medizinische, bio-logische und ökologische Grundlagen zur Bewertung schädlicher Luftverunreinigungen*, Sachverständigenanhörung Berlin, Feb. 20 – 24, 1978, Aug. 1978, pp. 33 – 99.
82. R. Guderian, K. Reidl: “Höhere Pflanzen als Indikatoren für Immissionsbelastungen im terrestrischen Bereich,” *Dechema-Beih*, (Bonn) **26** (1982) 6 – 22.
83. Biologische Meßverfahren zur Ermittlung und Beurteilung der Wirkung von Luftverunreinigungen. Grundlagen, VDI-Richtlinie 3957, Blatt 1, Vorentwurf, Stand July 1994.
84. G. Scholl, D. Schwela: “Kriterien zur Auswahl von Akzeptoren und Meßgrößen des Wirkungskatasters,” *Schriftenr. Landesanst. Immissionsschutz Landes Nordrhein-Westfalen, Essen* **40** (1977) 49–52.
85. H. Schönbeck: “Eine Methode zur Erfassung der biologischen Wirksamkeit von Luftverunreinigungen durch transplantierte Flechten, *Staub Reinh. Luft* **29** (1969) 14 – 18.
86. G. Scholl: “Die Immissionsrate von Fluor in Pflanzen als Maßstab für eine Immissionsbegrenzung,” *VDI-Ber.* **164** (1971) 39 – 45.
87. Bioindikation. Wirkungsbezogene Erhebungsverfahren für den Immissionsschutz, *VDI-Ber.* **609** (1987) 377 – 393.
88. Bioindikation: Ein wirksames Instrument der Umweltkontrolle, *VDI-Ber.* **901** (1992) vols. 1 and 2.
89. Zielsetzung und Bedeutung der Richtlinie Maximale Immissions-Werte, VDI 2310, Blatt 1 (10/88).
90. A. C. Posthumus, *VDI-Ber.* **314** (1978) 225 – 230.
91. A. S. Heagle, D. E. Bady, W. W. Heck: “An Open-Top-Field Chamber to Assess the Impact of Air Pollution on Plants,” *J. Environ. Qual.* **2** (1973) 265–268.
92. VDI-Fachdokumentation Reinhaltung der Luft: *Sauerstoffhaltige Schwefelverbindungen. Bibliographie 1974 – 1978*, VDI-Verlag, Düsseldorf 1978, pp. 286 ff.
93. J. Ziegler, H. Ziegler, *VDI-Ber.* **314** (1978) 219 – 223.
94. Maximale Immissions-Werte zum Schutz der Vegetation. Maximale Immissions-Werte für Schwefeldioxid, VDI-Richtlinie 2310, Blatt 2 (Entwurf) Stand Aug. 1978.
95. Maximale Immissions-Werte zum Schutz der Vegetation. Maximale Immissions-Konzentrationen für Ozon. VDI-Richtlinie 2310, Blatt 6, April 1989.
96. Maximale Immissions-Werte zum Schutz der Vegetation. Maximale Immissions-Konzentrationen für Fluorwasserstoff. VDI-Richtlinie 2310, Blatt 3, Dezember 1989.
97. D. F. Adams *et al.*: “Relationship of Atmospheric Fluoride Levels and Injury Indexes on Gladiolus and Ponderosa

Pine," *J. Agric. Food Chem.* **4** (1956) 64 – 66. Links

98. H.-J. Crecelius, W. Forwerg, R. Zahn, *Staub Reinh. Luft* **36** (1976) 201 – 205.
99. E. Brennan, J. A. Leone, R. H. Daines, *For. Sci.* **12** (1966) 286 – 390.
100. H. van Haut, B. Prinz, *Staub Reinh. Luft* **39** (1979) 408 – 414.
101. L. Steubing, *Staub Reinh. Luft* **39** (1979) 161 – 164.
102. G. H. Krause, *Staub Reinh. Luft* **39** (1979) 165 – 167.
103. Deutsche Gesellschaft für Mineralölwissenschaft und Kohlechemie, Forschungsbericht 174–1, Hamburg 1978.
104. D. T. Tingey, G. E. Taylor, jr.: "Variation in Plant Response to Ozone: A Conceptional Model of Physiological Events," in D. P. Orurod (ed.): *Effects of Gaseous Air Pollutants in Agriculture and Horticulture,* Butterworth Scientific 1982, p. 111 ff.
105. W. A. Feder: "Plant Response to Chronic Exposures to Low Levels of Oxidant Type Air Pollution," *Environ. Pollut.* **1** (1970) 73 – 79.
106. Umweltbundesamt Berlin (eds.): Berichte 5/83, Luftqualitätskriterien für photochemische Oxidantien, Erich Schmidt Verlag, Berlin 1983.
107. WHO Regional Office for Europe: "The Effects of Ozone and Other Photochemical Oxidants on Vegetation," in: *Air Quality Guidelines for Europe,* WHO Regional Publications, Eur. Ser. No. 23, Copenhagen 1987.
108. Minister für Arbeit, Gesundheit und Soziales des Landes Nordrhein-Westfalen (eds.): *Umweltprobleme durch Schwermetalle im Raum Stolberg*, Düsseldorf 1975.
109. Minister für Arbeit, Gesundheitheit und Soziales des Landes Nordrhein-Westfalen (eds.): *Umweltbelastung im Raum Datteln, Analyse und Problemlösung*, Düsseldorf 1977.
110. Verein Deutscher Ingenieure, Kommission Reinhaltung der Luft (eds.): *Säurehaltige Niederschläge – Entstehung und Wirkungen auf terrestrische Ökosysteme*, Düsseldorf 1983.
111. P. Davids, M. Lange, Die Großfeuerungsanlagenverordnung, Komm., 1984. H. Engelhardt, Bundesimmissionsschutzgesetz, Komm., 4th ed., 1997.
112. G. Feldhaus, Bundesimmissionsschutzrecht, Losebl.-Komm., 1998. K. Hansmann, TA Luft, Komm. 1987. H. D. Jarass, BImSchG, Komm., 1995. H.-J. Koch, D. Scheuing, BImSchG, Gemeinschafts-Komm., Losebl., 1998.
113. Landmann, Rohmer, vol. III, *Umweltrecht*, Losebl.-Komm. 1995. C. Offermann-Clas, *Luftreinhaltung in der Bundesrepublik Deutschland*, 1984. H. Schmatz, M. Nöthlichs, *Sicherheitstechnik*, **vol. VIII**, Losebl.-Komm. 1998. D. Sellner, *Immissionsschutz und Industrieanlagen*, 1988.
114. R. Stich, K. W. Porger, *Immissionsschutzrecht des Bundes und der Länder*, Losebl.-Komm. 1998. C. H. Ule, W. H. Laubinger, *Bundesimmissionsschutzgesetz*, Losebl.-Komm., 1998.
115. H. Reihlen: *Technische Normen – Freiheit und Bindung*, DIN-Mitteilung 70, 1991, no. 10, D. 527–532.
116. DIN-Normenheft 10: Grundlagen der Normungsarbeit des DIN, Beuth-Verlag, Berlin 1991.
117. Richtlinie des Rates vom 21. Dez. 1988 zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedsstaaten über Bauprodukte (89/106/EWG) A Bl. no. 40, Feb. 12, 1989, p. 12.
118. Amtsblatt der Europäischen Gemeinschaften: "Richtlinie des Rates vom 28. März 1983 über ein Informationsverfahren auf dem Gebiet der Normen und technischen Vorschriften," 83/109/EWG.
119. C. Mohr: *ISO and CEN*, DIN-Mitteilung 70, 1991, no. 6, pp. 319 – 320.
120. Europäische Normen für 1992 – Ein Leitfaden des DIN, DIN Deutsches Institut für Normung e.V., Berlin.
121. H. Pohle: *Chemische Industrie – Umweltschutz, Arbeitsschutz, Anlagensicherheit. Rechtliche und Technische Normen, Umsetzung in die Praxis*, VCH Verlagsgesellschaft, Weinheim, Germany 1991.
122. R. Kleinert: *Marktzugang durch Normenkonformität*, DIN-Mitteilung 71, 1992, no. 1, pp. 17 – 23.
123. K. Grefen: "Harmonisierung technischer Regeln im EG-Binnenmarkt – Beitrag der Kommission Reinhaltung der Luft im VDI und DIN," *Staub – Reinhalt. Luft* **51** (1991) 199 – 205.
124. C. A. Aronds: "International Efforts for Standardization of Measuring Methods in Air Pollution Control – Requirements and Aims," *Tagung Kommission Reinhaltung der Luft im VDI und DIN über Aktuelle Aufgaben der Meßtechnik in der Luftreinhaltung*, Heidelberg 1990, pp. 3 – 10.
125. Richtlinie (EWG) des Rates über die Verbrennung gefährlicher Abfälle, Bundesrats-Drucksache 290/92, Ratsdok. 5761/92, Heger-Verlag, Bonn.
126. Bauproduktenrichtlinie 89/106/EWG: Richtlinie des Rates vom 21. 12. 1988 zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedsstaaten über Bauprodukte, geändert durch die Richtlinie 93/68/EWG des Rates vom 22. Juli 1993.
127. G. Feldhaus: *Grundlagen des Umweltschutzes – Bericht über die Arbeit des NAGUS*, DIN-Mitteilung 73. 1994, no. 7, pp. 456 – 460.
128. R. C. West, M. J. Astle: *Handbook of Chemistry and Physics*, CRC Press, Palm Beach 1979.
129. T. E. Graedel: *Chemical Compounds in the Atmosphere*, Academic Press, New York 1978.
130. L. S. Lochran, R. A. Pielke, E. Kovács, *J. Air Waste Manage. Assoc.* **42** (1992) 1567 – 1572.
131. P. Watkiss, D. Forster, A. Hunt, A. Smith, T. Taylor: "Comparison of EU Air Pollution Policies and Legislation with other Countries", *AEA Technology Environment*, Abingdon, UK, 2004.
132. Council Directive 96/62/EC of 27. September 1996 on ambient air quality assessment and management. *Official Journal of the European Communities*, No. L296, 21.11.1996, pp. 55 – 63.
133. Council Directive 1999/30/EC of 22.4.1999 relating to limit values for sulphur dioxide, nitrogen dioxide and oxides of nitrogen, particulate matter and lead in ambient air, Official Journal of the European Communities L 163/41, 29.6.1999.
134. Directive 2000/69/EC of the European Parliament and of the council of 16.11.2000 relating to limit values for

benzene and carbon monoxide in ambient air, Official Journal of the European Communities L 313/12, 13.12.2000.

135. Directive 2002/3/EC of the European Parliament and of the council of 12.2.2002 relating to ozone in ambient air, Official Journal of the European Communities L 67/14, 9.3.2002.
136. Directive 2004/107/EC of the European Parliament and of the council of 15.12.2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air, Official Journal of the European Communities L 23/3, 26.1.2005.
137. 22. Verordnung zur Durchführung des Bundes-Immissionsschutz-Gesetzes (Verordnung über Immissionswerte für Schadstoffe in der Luft — (22. BImSchV) vom 11.9.2002, BGBl. I, S. 3626.
138. E. Lahmann: *Feststellung und Bewertung von Immissionen — Leitfaden zur Immissionsüberwachung in Deutschland*, 2. Auflage, UBA-Texte 34/9, Umweltbundesamt, Berlin 1997.
139. D. L. Fox, H. E. Jeffries, *Anal. Chem.* **51** (1979) 22 R – 34 R.
140. Technische Anleitung zur Reinhaltung der Luft (TA Luft) vom 30.07.2002 Gemeinsames Ministerialblatt 25 — 29 (2002) p. 511 — 605.
141. D. H. F. Atkins, I. Quirino: *A Survey of Nitrogen Dioxide in Paris*, Report EUR 13 369 EN, Commission of the European Communities (eds.), Brussels 1992, pp. 1 – 59.
142. E. De Saeger, M. Gerbolés, P. Perez Ballesta, L. Amantini, M. Payrissat: Air Quality Measurements in Brussels (1993 — 1994) - NO_2 and BTX Monitoring Campaigns by Diffusive Samplers", EUR Report 16310 EN, 1995.
143. P. Bruckmann, R. Beier, S. Krautscheid, *Staub Reinhalt. Luft* **43** (1983) 404 – 410.
144. Workshop PMX-Quellenidentifizierung: Ergebnisse als Grundlage für Maßnahmenpläne. T. Kuhlbusch, A. John, S. Top (eds.), Duisburg 22 — 23.1. 2004.
145. M. Buck, S. Krautscheid, *Umwelt* 1988, no. 5, 188–190.
146. DIN V ENV 13005: Guide to the Expression of Uncertainty in Measurement. German version EN 13005: 1999.
147. VDI-Richtlinie 4280, Bl. 1 (11/96), in [152].
148. R. Beier, *Staub Reinhalt. Luft* **50** (1990) 265 – 270.
149. ISO/CD 11 222.2, Air Quality – Comparison of Mean Values Determined from Air Quality Measurements with Prescribed Values.
150. B. Schneider, *Atmos. Environ.* **21** (1987) 1275 – 1283.
151. W. A. McClemey, K. D. Oliver, J. D. Plell, *Environ. Sci. Technol.* **23** (1989) 1273 – 1379.
152. Kommission Reinhaltung der Luft im VDI und DIN (eds.): *VDI-Handbuch Reinhaltung der Luft*, VDI-Verlag, Düsseldorf (loose-leaf book).
153. DIN ISO 5725, Genauigkeit (Richtigkeit und Präzision) von Meßverfahren und Meßergebnissen.
154. DIN EN ISO 20988: 2004: Guide to Estimating Measurement Uncertainty; German version pr EN ISO 20988: 2004.
155. DIN 32 645 (05/94).
156. R. van Aalst, L. Edwards, T. Pulles, E. De Saeger, M. Tombrou, D. Tonnesen: Guidance Report on Preliminary Assessment under EC Air Quality Directives, *European Commission, Jan.* 1998. Technical report No. 11, http://binary.eea.eu.int/t/tech11.pdf.
157. Guidance on Assessment under the EU Air Quality Directives, www.europa.eu.int/comm/environment/air/background.htm.
158. D. Schwela, L. Morawska, D. Kotzias: *Guidelines for Concentration and Exposure: Response Measurement of Fine and Ultra Fine Particulate Matter for use in Epidemiological Studies*, World Health Organization, Geneva 2002.
159. ISO 11222 (2002) Air Quality-Determination of the Uncertainty of the Time Average of Air Quality Measurements.
160. R. Shukla, I. Luo, G. K. Le Masters, S. Grinshpun, D. Martuzevicius, *J. Environ. Monit.* 2005, 603 – 607.
161. U. Pfeffer, T. Zang, R. Beier, *Gefahrstoffe-Reinhalt. Luft* **66** (2006) 38 – 44.
162. M. Gerbolez, D. Buzica, L. Amantini, *Atmos. Environ.* **39** (2005) 2579 – 2592. Links
163. Immissions- und Wirkungsuntersuchungen in Aalen und Heidenheim 1994/95. UMEG Gesellschaft für Umweltmessungen und Umwelterhebungen, Karlsruhe, 1995.
164. VDI-Richtlinie 4280, Bl. 3 (06/2003) in [166]
165. DIN EN ISO 9169 (09/2006).
166. Air quality — Approach to Uncertainty Estimation for Ambient Air Reference Measurement Methods, CEN report CR 14377:2002E.
167. EURACHEM/CITAC Guide Quantifying Uncertainty in Analytical Mesurement, 2nd ed.
168. Demonstration of Equivalence of Ambient Air Monitoring Methods. Report by an EC Working Group on Guidance for the Demonstration of Equivalence (2005). www.europa.eu.int/comm/environment/air/cafe/index.htm
169. VDI 4202, Part 1 (2002-06): Minimum Requirements for Suitability Tests of Automated Ambient Air Quality Measuring Systems — Point-Related Measurement Methods of Gaseous and Particulate Pollutants, in [166].
170. VDI 4203, Part 1 (2001-10): Testing of Automated Measuring Systems — General Concepts, in [166].
171. VDI 4203, Part 3 (2004-08): Testing of Automated Measuring Systems — Test Procedures for Point-Related Ambient Air Quality Measuring Systems of Gaseous and Particulate Pollutants, in [166].
172. http://www.npl.co.uk/environment/mcerts.html.
173. Ambient Air Quality — Standard Method for the Measurement of the Concentration of Nitrogen Dioxide and Nitrogen Monoxide by Chemiluminescence; German version EN 14211:2005, Beuth-Verlag.
174. Ambient Air Quality — Standard Method for the Measurement of the Concentration of Sulphur Dioxide by Ultraviolet Fluorescence; German version EN 14212:2005, Beuth-Verlag.
175. Ambient Air Quality — Standard Method for the Measurement of the Concentration of Ozone by Ultraviolet

Photometry; German version EN 14625:2005, Beuth-Verlag
176. Ambient Air Quality — Standard Method for the Measurement of the Concentration of Carbon Monoxide by Nondispersive Infrared Spectroscopy; German version EN 14626:2005, Beuth-Verlag
177. Ambient Air Quality – Standard Method for Measurement of Benzene Concentrations — Part 3: Automated Pumped Sampling with In Situ Gas Chromatography; German version EN 14662-3:2005
178. Quality Management Systems — Fundamentals and Vocabulary (ISO 9000:2005); Trilingual version EN ISO 9000:2005
179. Quality Management Systems — Requirements (ISO 9001:2000); trilingual version EN ISO 9001:2000
180. General Requirements for the Competence of Testing and Calibration Laboratories (ISO/IEC 17025:2005); German and English version EN ISO/IEC 17025:2005
181. http://ies.jrc.cec.eu.int/Units/eh/Projects/Aquila/
182. H.-U. Pfeffer: "Qualitätssicherung in automatischen Immissionsnetzen,", *LIS-Ber.* **52** (1984) 1 – 73.
183. U. Pfeffer, G. Nitz: *Interlaboratory Comparisons and Control of Private Laboratories, QA/QC in the Field of Emission and Air Quality Measurements: Harmonization*, Standardization and Accreditation, Prague, May 21 – 23. 2003. European Commission, Joint Research Centre, Report EUR 20973 EN, ISBN 92-894-6523-9, 2004, pp. 150 – 158.
184. H. Manns, H. Gies, G. Nitz: "Anlage zur Qualitätssicherung von Immissionsmessungen," *LIS-Ber.* **27** (1982) 1 – 139.
185. EC Harmonization Programme for Air Quality Measurements. Intercomparison Exercises 1999/2000 for SO2, CO, NO2, and O3 (ERLAP Report EUR 19629 EN, 2000).
186. E. Lahmann: *Luftverunreinigung – Luftreinhaltung*, Verlag Paul Paray, Berlin 1990.
187. G. Baumbach, Luftreinhaltung, 3. Aufl., Springer, Berlin 1994.
188. D. Eickelpasch, G. Eickelpasch, Determination and Evaluation of Ambient Air Quality, Manual of Ambient Air Monitoring in Germany, 3rd, revised ed., UBA-Texte 07/2004, ISSN 0722-186X.
189. http://www.umweltbundesamt.de/messeinrichtungen/.
190. J. W. Erisman *et al*: "Instrument Development and Application in Studies and Monitoring of Ambient Ammonia," *Atmos. Environ.* **35** (2001) 1913 – 1922. Links
191. ISO 6142, Publication date: 2001-04, Gas Analysis — Preparation of Calibration Gas Mixtures — Gravimetric Method.
192. ISO 6143, Publication date: 2001-05, Gas analysis — Comparison Methods for Determining and Checking the Composition of Calibration Gas Mixtures.
193. DIN ISO 6144, Publication date: 1985-05, Gas Analysis — Preparation of Calibration Gas Mixtures; Static Volumetric Methods.
194. ISO 6145-1, Publication date: 2003-11, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 1: Methods of Calibration.
195. ISO 6145-4, Publication date: 2004-06, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 4: Continuous Syringe Injection Method.
196. ISO 6145-5, Publication date: 2001-04, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 5: Capillary Calibration Devices.
197. ISO 6145-6, Publication date: 2003-05, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 6: Critical Orifices.
198. ISO 6145-7, Publication date: 2001-04, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 7: Thermal Mass-Flow Controllers.
199. ISO 6145-8, Publication date: 2005-02, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 8: Diffusion Method.
200. ISO 6145-9, Publication date: 2001-04, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 9: Saturation Method.
201. ISO 6145-9 Technical Corrigendum 1, Publication date: 2002-05, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 9: Saturation Method; Technical Corrigendum 1.
202. ISO 6145-10, Publication date:2002-02, Gas Analysis — Preparation of Calibration Gas Mixtures using Dynamic Volumetric Methods — Part 10: Permeation Method.
203. P. W. West, G. C. Gaeke, *Anal. Chem.* **28** (1956) 1816 – 1819. Links
204. VDI-Richtlinie 2451, Bl. 3 (09/96) [152].
205. ISO 6767, 09/1990.
206. B. E. Saltzman, *Anal. Chem.* **26** (1954) 1949 – 1955. Links
207. VDI-Richtlinie 2453, Bl. 1 (10/90) in [152].
208. ISO 6768, 1998-08.
209. VDI-Richtlinie 2452, Bl. 2 (02/75) in [152].
210. VDI-Richtlinie 2452, Bl. 3 (07/87) in [152].
211. K. Svoboda, H. Ixfeld, *Staub Reinhalt. Luft* **31** (1971) 1 – 8.
212. H. Kettner, W. Forwerg, *Atmos. Environ.* **3** (1969) 215 – 220.
213. VDI-Richtlinie 2458, Bl. 1 (12/73) in [152].
214. VDI-Richtlinie 2461, Bl. 1 (03/74) in [152].
215. VDI-Richtlinie 2461, Bl. 2 (05/76) in [152].
216. VDI 3869, Part 3 (1996-05) in [166]
217. VDI 3869, Part 4 (1996-05) in [166]

218. VDI-Richtlinie 2454, Bl. 1 (03/82) in [152].
219. VDI-Richtlinie 2454, Bl. 2 (03/82) in [152].
220. D. Klockow, *Fresenius Z. Anal. Chem.* **326** (1987) 5 – 24.
221. M. Ferm, *Atmos. Environ.* **13** (1979) 1385 – 1393.
222. M. Ferm, *Atmos. Environ.* **20** (1986) 1193 – 1201.
223. J. Slanina *et al.*, *Int. J. Environ. Anal. Chem.* **9** (1981) 59 – 70. Links
224. N. A. Dimmock, G. B. Marshall, *Anal. Chim. Acta* **185** (1986) 159 – 169. Links
225. M. P. Keuken, A. Wayers-IJpelaan, J. J. Mols, R. P. Otjes, J. Slanina, *Atmos. Environ.* **23** (1989) 2177 – 2185.
226. J. Slanina, G. P. Wyers, *Fresenius J. Anal. Chem.* **350** (1994) 467 – 473. Links
227. A. Kappos et al., Int. *J Hyg. Environ. Health* **207** (2004) 399 – 407.
228. VDI-Richtlinie 3490, Bl. 1 – 17 in [152].
229. VDI 2453, Part 2 (2002-10) in [166].
230. J. P. Lodge, Jr. (ed.): *Methods of Air Sampling and Analysis*, 3rd ed., Lewis Publishers, Chelsea, Michigan, 1989.
231. J. Schwartz, D. W. Dockery, L.M. Neas: "Is Daily Mortality Associated Specifically with Fine Particles?", *J. Air Waste Manag. Assoc.* **46** (1996) 927–939. Links
232. Second Position Paper on Particulate Matter, CAFE Working Group on Particulate Matter, 6.4. 2004. www.europa.eu.int/comm/environment/air/cafe/pdf/working_groups/ wg_particulate_matter.htm.
233. VDI-Richtlinie 2463, Bl. 1 (11/99) in [152].
234. Task Group of the IRCP: IRCP Publication 66. Human Respiratory Tract Model for Radiological Protection. Chapter 4: Deposition, Ann IRCP **24** (1994) 36 – 65.
235. Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide. Report on a WHO Working Group, Bonn, 13 – 15.1. 2003. WHO Regional Office for Europe, Copenhagen 2003.
236. B. Brunekreef, B. Forsberg, *Eur Respir. J.* **26** (2005) 309 – 318. Links
237. United States EPA. Air Quality Criteria for Particulate Matter. Environment Protection Agency, Research Triangle Park, NC, USA (10/2004).
238. EN 12 341 (12/97): Luftbeschaffenheit - Ermittlung der PM10-Fraktion von Schwebstaub - Referenzmethode und Feldprüfverfahren zum Nachweis der Gleichwertigkeit von Meßverfahren und Referenzmeßmethode.
239. EN 14907 (8/2005): Ambient Air Quality — Standard Gravimetric Measurement Method for the Determination of the PM2.5 Mass Fraction of Suspended Particulate Matter.
240. G. Oberdorster, J. Ferin, B. Lehnert, *Environ. Health Perspect.* **102** (1994) 173 – 179. Links
241. A. Peters, S. von Klot, M. Heier, *New Engl. J. Med.* **351** (2004) 1721 – 1730. Links
242. VDI-Richtlinie 3867 Bl. 2, in preparation, to be published in [166].
243. A. Wiedensohler, *Aerosol Sci. Technol.* **27** (1997) 224 – 242.
244. H. G. Horn: "Neue Entwicklungen bei der Messung und Beurteilung der Luftqualität", *VDI-Ber. 1885* (2005) 47 – 61.
245. VDI-Richtlinie 3867 Bl. 1 (E) (2/2006) in [166].
246. C. Helsper: *Partikelmesstechnik* in: L. Gail, H.-P. Hortig: Reinraumtechnik, Springer, Heidelberg 2002, pp. 57 – 82.
247. J. Cyrys, M. Stötzel, J. Heinrich, W. G. Kreyling, N. Menzel, K. Wittmaack, T. Tuch, H.-E. Wichmann, *Sci. Total Environ.* **305** (2003) 143 – 156. Links
248. X. Querol, A. Alastney, C. R. Ruiz, B. Artinano, H.C. Hansson, R. M. Harrison, E. Buringh, H. M. Ten Brink, M. Lutz, P. Bruckmann, P. Straehl, J. Schneider, *Atmos. Environ.* **38** (2004) 6547 – 6555. Links
249. VDI-Richtlinie 3867 Bl. 3, in preparation, to be published in [166].
250. U. Quass, T. Kuhlbusch, M. Koch: Identifizierung von Quellgruppen für die Feinstaubfraktion, IUTA Report Nr. LP 15, 2004.
251. J. Keskimmen, K. Pietarinen, M. Lehtimäki, *J. Aerosol Sci.* **23** (1992) 353 – 360.
252. D. Sinclair, R. J. Countess, B. Y. H. Liu, D. Y. H. Pui: "Automatic Analysis of Submicron Aerosols" in D. A. Lundgren (ed): *Aerosol Measurement*, University Presses of Florida, Gainesville 1979.
253. H. Fissan, D. Hummes, F. Stratmann, P. Büscher, S. Neumann, *Aerosol Sci. Technol.* **24** (1996) 1 – 13.
254. D. Bake, L. Laskus, *Staub Reinhalt. Luft* **38** (1978) 21 – 24.
255. VDI-Richtlinie 2463, Bl. 7 (08/82) in [152].
256. G. A. Jutze, K. E. Foster, *J. Air Pollut. Control. Assoc.* **17** (1967) 17 – 25.
257. US Environmental Protection Agency: Quality Assurance Handbook for Air Pollution Measurement Systems, vol. II: Ambient Air Specific Methods, Research Triangle Park, N.C., 1977.
258. VDI-Richtlinie 2463, Bl. 11 (10/96) in [166]
259. VDI-Richtlinie 2463, Bl. 5 (12/87) in [152].
260. VDI-Richtlinie 2463, Bl. 6 (11/87) in [152].
261. H. G. Wirweitzky, *VDI-Ber.* **1059** (1993) 533 – 569.
262. H.-U. Pfeffer, M. Buck, *J. Aerosol Sci.* **17** (1986) 290 – 292. Links
263. K. Williams, C. Fairchild, J. Jaklevic: "Dynamic Mass Measurement Techniques" in: K. Willeke, P. A. Baron (eds.): Aerosol Measurement — Principles, Techniques, and Applications, Van Nostrand Reinhold, New York 1983.
264. H. Patashnick, E. G. Rupprecht, *J. Air Waste Manage. Assoc.* **41** (1991) 1079 – 1083. Links
265. G. P. Ayers, *Atmos. Environ.* **38** (2004) 3453 – 3458. Links
266. D. J. Eatough et al.: Measurement of Fine Particulate Mass, Including the Semi-Volatile Constituents, with a GRIMM

Monitor, 97th Annual AWMA Conference and Exhibition, Pittsburgh 2004.

267. Thermo ESM Andersen: “Technische Spezifikation ZT-030-0498, Beheiztes Probenahmerohr (komplett) zu den Staubmessgeraten FH 62 I-R und FH 62 I-N”, Erlangen, 29.04.2002.
268. M. Williams, P. Bruckmann: “Guidance to Member States on PM10 Monitoring and Intercomparisons with the Reference Method”. EC Working Group on Particulate Matter, 2001. http://europa.eu.int/comm/environment/air/pdf.
269. A. Charron, R.M. Harrison, S. Moorcroft, J. Booker, *Atmos. Environ.* **38** (2004) 415 – 423. Links
270. J. C. Chow, J. G. Watson, E. M. Fujita, Z. Lu, D. R. Lawson, L. L. Ashbaugh, *Atmos. Environ.* **28** (1994) 2061 – 2080. Links
271. *Rupprecht & Patashnick, Operating Manual — Sample Equilibration System with Relative Humidity Module*, Rupprecht & Patashnick Co. Inc., Albany, NY, USA, 2000.
272. H. Patashnick, G. Rupprecht, J. C. Ambs, M. B. Meyer, *Aerosol Sci. Technol.* **34** (2001) 42 – 45.
273. C. Asbach, T. A.J. Kuhlbusch, H. Fissan: “Online Monitor for Measurement of Real Airborne Particle Mass Concentrations”, *Proceedings of QA/QC in the Field of Emission and Air Pollution Control Conference, Prague* 2003.
274. VDI-Richtlinie 2463, Bl. 10 (9/96) in [166].
275. Organisation for Economic Cooperation and Development (OECD): *Methods of Measuring Air Pollution*, Paris 1964.
276. *Instruction Manual for the National Survey of Smoke and Sulfur Dioxide*, Warren Spring Laboratory, Stevenage, Hertfordshire, 1966.
277. D. L. R. Balley, P. Clayton; Report LR 325 (AP), Warren Spring Laboratory, Stevenage, Hertfordshire, 1980.
278. L. Laskus, D. Bake, L. Armbruster, *Staub Reinhalt. Luft* **40** (1980) 18 – 26, 101 – 105.
279. G. E. Pashel, D. R. Egner, *Atmos. Environ.* **15** (1981) 919 – 927.
280. Particulate Matter in the United Kingdom, Air Quality Expert Group, Defra, London 2005.
281. DIN EN ISO 1034-1 (D19), Bestimmung der gelösten Anionen Fluorid, Chlorid, Nitrit, Orthophosphat, Bromid, Nitrat und Sulfat mittels Ionenchromatographie, Beuth-Verlag GmbH, Berlin, April 1995.
282. VDI 2267 part 15, Determination of Suspended Matter in Ambient Air — Measurement of the Mass Concentration of Al, As, Ca, Cd, Co, Cr, Cu, K, Mn,. Ni, Pb, Sb, V, Zn as Part of Dust Precipitation by Mass Spectrometry (ICP-MS).
283. VDI 2267 part 5, Determination of Suspended Matter in Ambient Air — Determination of the Mass Concentration of Be, Cd, Co, Cr, Cu, Fe, Mn, Ni, Pb, Sb, V, Zn by Optical Emission Spectroscopy (ICP-OES) after Sampling on Filters and Digestion in an Oxidising Agent.
284. VDI 2267 part 1, Determination of Suspended Matter in Ambient Air — Measurement of the Mass Concentration of As, Be, Cd, Co, Cr, Cu, Mn,. Ni, Pb, Sb, Tl, Zn by Atomic Absorption Spectroscopy (AAS) after Sampling on Filters and Digestion in an Oxidising Agent.
285. VDI-Richtlinie 2267, Bl. 3, 4, 6, 7, 11, 12, in [152].
286. S. M. Talebi, “Speciation of Arsenic in Airborne Particulate Matter”, *Fresenius Environ. Bull.* **8** (1999) 59 – 64.
287. D. Metze, H. Herzog, B. Gosciniak, D. Gladtke, N. Jacubowski: “Determination of Cr(VI) in Ambient Airborne Particulate Matter by a Species-Preserving Scrubber-Sampling Technique”, *Anal. Bioanal. Chem.* **378** (2004) 123 – 128.
288. L. Füchtjohann, N. Jakubowski, D. Gladtke, D. Klockow, J. A. C. Broekart: “Speciation of Nickel Compounds in Airborne Particulate Matter by Means of Sequential Extraction in a Microflow System and Determination by Graphite Furnace Atomic Absorption Spectrometry and Inductively Coupled Plasma Mass Spectrometry”, *J. Environ. Monit.* **3** (2001) 681 – 687.
289. K. C. Galbreath, D. L. Toman, C. J. Zygarlicke, F. E. Huggins, G. P. Huffman, J. L. Wong: “Nickel Speciation of Residual Oil Fly Ash and Ambient Particulate Matter using X-ray Absorption Spectroscopy”, *J. Air Waste Manag. Assoc.* **50** (2000) 1876 – 1886. Links
290. F. Pott, *VDI-Ber.* **888** (1991) 211 – 244.
291. Technische Regeln für Gefahrstoffe, BArbBl. 2/92, 57.
292. G. Elbers, S. Muratyan, *VDI-Ber.* **888** (1991) 143 – 170.
293. M. Buck, G. Elbers, *Erdöl Kohle Erdgas Petrochem.* **45** (1992) 219 – 223.
294. G. Elbers, S. Lutz, *VDI-Ber.* **1059** (1993) 337 – 356.
295. A. Petzold, R. Niessner, *Gefahrstoffe Reinhaltung Luft* **56** (1966) 173 – 177.
296. VDI 2465, Blatt 1: Messen von Ruß (Immissionen); Chemisch-analytische Bestimmung des elementaren Kohlenstoffes nach Extraktion und Thermodesorption des organischen Kohlenstoffs
297. M. E. Birch, R. A. Cary: “Elemental Carbon-Based Method for Monitoring Occupational Exposures to Particulate Diesel Exhaust”, *Aerosol Sci. Technol.* **25** (1996) 221 – 241.
298. A. Petzold, R. Niessner, *Mikrochim. Acta* **117** (1995) 215 – 237. Links
299. Guidelines for Air Quality, WHO, Geneva, 2000
300. VDI 3492 (2004-10) in [166]
301. BS ISO 14966, Publication date: 2002-12-11: Ambient air — Determination of numerical concentration of inorganic fibrous particles — Scanning electron microscopy method.
302. Norwegian Institut for Air Research (NILU), “Annual Summary Report from the Chemical Co-ordinating Centre”, *EMEP-CCC-Report 6/88,* 1988.
303. V. Mohnen, J. Wilson, J. Kadlecek in H. W. Georgii, W. Jaeschke (eds.): *Chemistry of the Unpolluted and Polluted Troposphere*. D. Reidel Publ., Dordrecht 1982, pp. 395 – 423.
304. W. Fricke, S. Beilke, *Environ. Pollut.* **75** (1992) 121 – 127. Links
305. VDI 3870 Bl. 1 (12*1997): Messen von Regeninhaltsstoffen. Probenahme von Regenwasser, in [152].

306. P. Winkler, *Staub Reinhalt. Luft* **4** (1985) 256 – 260.
307. H. W. Georgii, G. Schmitt, *Staub Reinhalt. Luft* **45** (1985) 260 – 264.
308. F. G. Römer, B. H. TE Winkel, L. H. J. M. Janssen in G. Restelli, G. Angeletti (eds.): *Physico-chemical Behaviour of Atmospheric Pollutants*, Kluwer Acad. Publ., Dordrecht 1990, pp. 444 – 451.
309. E. Schultz, *Staub Reinhalt. Luft* **3** (1983) 482 – 489. Link
310. M. Odabasi, A. Sofuoglu, N. Vardar, Y. Tasdemir, Th. M. Holsen, *Envir. Sci. Technol.* **33** (1999) 426 – 434.
311. U. Bartels, J. Block, *J. Plant Nutr. Soil Sci.* **148** (1985) 689 – 698.
312. W. B. Grant, R. H. Kagann, W. A. McClenny, *J. Air Waste Manage. Assoc.* **42** (1992) 18 – 30. Links
313. K. Weber, V. Klein, W. Diehl: "Optische Fernmeßverfahren zur Bestimmung gasförmiger Luftschadstoffe in der Troposphäre," *VDI-Ber.* **838** (1990) 201 – 246.
314. VDI-Richtlinie 4202 Bl. 2 (7/2004): Mindestanforderungen an automatische Immissionsmesseinrichtungen bei der Eignungsprüfung — Optische Fernmesseinrichtungen zur Messung von gasförmigen Immissionen.
315. W. M. Vaughan, J. O. Zwicker, R. H. Dunaway, P. Martinoin Proceedings of Optical Sensing for Environmental Monitoring. Air & Waste Management Association, Pittsburgh 1994, pp. 517 – 528.
316. K. Weber, T. Lamp, J. Schulte, G. Weidemann, G. van Haren, A. Gärtner, G. Bröker in CD-Proceedings of the 89th Annual Meeting & Exhibition of the Air & Waste Management Association, Pittsburgh 1996, 96-TA 26A.02.
317. K. Weber, A. Ropertz, T. Lamp, *G. van Haren, VDI-Ber.* **1443** (1999) 269 – 306.
318. R. Hans, K. Schäfer, W. Bautzer, J. Heland, H. Mosebach, H. Bittner, T. Eisenmann, *Appl. Opt.* **33** (1994) 5682 – 5689.
319. VDI-Richtlinie 4211 Bl. 1 (12/2000): Fernmessverfahren — Messungen in der bodennahen Atmosphäre nach dem FTIR-Prinzip, Messen gasförmiger Emissionen und Immissionen, Grundlagen, in [166].
320. DIN EN 15483 (draft, in preparation): Ambient Air Quality — Atmospheric Measurements Near Ground with FTIR Spectroscopy.
321. U. Platt et al., Nature **285** (1980) 312 – 314.
322. Eignungsprüfung Mehrkomponentenmeßsysteme AR 500 Komponente SO2 im Auftrag der Fa. Opsis AB. Prüfbericht nr. 33–01/93. UMEG GmbH (ed.), Karlsruhe 1993.
323. VDI-Richtlinie 4212 Bl. 1 (E, 11/2005): Fernmessverfahren — Messungen in der bodennahen Atmosphäre nach dem DOAS-Prinzip. Messen gasförmiger Emissionen und Immissionen, Grundlagen.
324. K. Josek, W. Junkermann, P. Werle, F. Slémr: "Erfahrungen an Absorptionsspektrometern mit abstimmbaren Diodenlasern für die Mehrkomponentenmessung von Luftschadstoffen," *VDI Ber.* **838** (1990) 181 – 200.
325. A. Hofzumahans, H.-P. Dorn, U. Platt: "Tropospheric OH Radical Measurement Techniques: Recent Developments," in G. Restelli, G. Angeletti (eds.): *Physico-chemical Behaviour of Atmospheric Pollutants,* Kluwer Acad. Publ., Dordrecht 1990, pp. 103 – 108.
326. J. Kames, R.W. H. Schmidt, *Labor* 2000) 12 – 23.
327. P. T. Woods, R. H. Partridge, M. J. Milton, A. J. Davenport, *Bull. Int. Org. Legal Metrology* **131** (1993) 11 – 25.
328. K. Fritzsche, G. Schubert, *Laser Optoelektronik* **29** (1997) 56 – 61.
329. R. Grönlund, M. Sjökolen, P. Weibring, H. Edner, S. Svanberg, *Atmos. Environ.* **39** (2005) 7474 – 7480. Links
330. VDI-Richtlinie 4210 Bl. 1 (06/1996): Fernmessverfahren — Messungen in der Atmosphäre nach dem LIDAR-Prinzip, Messen gasförmiger Luftverunreinigungen mit dem DAS-LIDAR.
331. C. Weitkamp et al.: "Das mobile, ortsauflösende Schadgasfernmeß-System ARGOS" in C. Werner, V. Klein, K. Weber (eds.): *Laser in der Umweltmeßtechnik in Remote Sensing,* Springer Verlag, Berlin 1992, pp. 197 – 200.
332. S. Beilke et al.: "Air Sampling Flights at Low Altitudes along the Border between the FRG and its Neighbours," in G. Angeletti, G. Restelli (eds.): *Physico-Chemical Behaviour of Atmospheric Pollutants,* D. Reidel Publ., Dordrecht 1987, pp. 685 –703.
333. U. Platt, D. Perner, A. M. Winer, G. W. Harris, J. N. Pitts, *Geophys. Res. Lett.* **7** (1980) 89 – 92. Links
334. F. Slemr, G. W. Harris, D. H. Hastic, G. I. Hackay, H. I. Schiff, *J. Geophys. Res.* **91** (1986) 5371. Links
335. E. C. Tuazon *et al.*, *Environ. Sci. Technol.* **18** (1984) 981 – 984. Links
336. P. Bruckmann, W. Mülder, *Schriftenr. Landesanstalt Immissionsschutz d. Landes Nordrhein-Westfalen, Essen* **47** (1979) 30 – 41.
337. T. E. Kleindienst *et al.*, *Environ. Sci. Technol.* **22** (1988) 53 – 61. Links
338. VDI-Richtlinie 3483 Bl. 1 (12/79): Messen der Summe organischer Stoffe mit einem Flammen-Ionisations-Detektor (FID). Grundlagen, in [152].
339. DIN EN 14662-2 (08/2005) Ambient Air Quality — Standard Method for Measurement of Benzene Concentrations — Part 1: Pumped Sampling Followed by Thermal Desorption and Gas Chromatography; German version EN 14662-1:2005.
340. DIN EN 14662-2 (08/2005) Ambient air quality — Standard Method for Measurement of Benzene Concentrations — Part 2: Pumped Sampling Followed by Solvent Desorption and Gas Chromatography; German version EN 14662-2:2005.
341. P. Bruckmann *et al.*, *Chemosphere* **17** (1988) 2363 – 2380. Links
342. DIN EN ISO 16017-1 (10/2001): Indoor, Ambient and Workplace Air — Sampling and Analysis of Volatile Organic Compounds by Sorbent Tube/Thermal Desorption/Capillary Gas Chromatography — Part 1: Pumped Sampling (ISO 16017-1:2000); German version EN ISO 16017-1:2000.
343. VDI 2100 Blatt 5 (08/2005): Messen gasförmiger Verbindungen in der Außenluft — Messen von Innenraumluftverunreinigungen — Gaschromatographische Bestimmung organischer Verbindungen — Messen von leicht flüchtigen organischen Verbindungen, insbesondere Ozon-Vorläufersubstanzen.

344. W. Schneider, J. C. Frohne, H. Bruderreck, *J. Chromatogr.* **155** (1978) 311 – 327. Links
345. H. Kern, N. Kirshen: "Halogenierte organische Verbindungen in der Umwelt," *VDI-Ber.* **745** (1989) 335 – 344.
346. W. McClenny, J. Pleil, G. Evans, K. Oliver, M. Holdren, W. Winbery, *J. Air Waste Manage. Assoc.* **41** (1991) 1308 – 1318. Links
347. VDI 2100 Blatt 2 (11/2002): Messen gasförmiger Verbindungen in der Außenluft - Messen von Innenraumluftverunreinigungen - Gaschromatographische Bestimmung organischer Verbindungen; Aktive Probenahme durch Anreicherung auf Sorbenzien; Thermodesorption.
348. DIN EN ISO 16017-1 (10/2001): Indoor, ambiant and workplace air - Sampling and analysis of volatile organic compounds by sorbent tube/thermal desorption/capillary gas chromatography - Part 1: Pumped sampling (ISO 16017-1:2000); German version EN ISO 16017-1:2000.
349. J. Rudolph, A. Khedim, *Int. J. Environ. Anal. Chem.* **20** (1985) 265 – 282. Links
350. K. Grob, G. Grob, *J. Chromatogr.* **62** (1971) 1–13. Links
351. VDI 2100 Blatt 2 (06/2001) Determination of Gaseous Compounds in Ambient Air; Determination of Indoor Air Pollutants — Gas Chromatographic Determination of Organic Compounds — Active Sampling by Accumulation on Activated Charcoal; Solvent Extraction.
352. VDI-Richtlinie 3865 B1.3 (6/1998): Messen organischer Bodenverunreinigungen — Gaschromatographische Bestimmung von niedrigsiedenden organischen Verbindungen in Bodenluft nach Anreicherung an Aktivkohle oder XAD-4 und Desorption mit organischem Lösungsmittel, in [166].
353. VDI 2100 Blatt 2 (06/2001) Messen gasförmiger Verbindungen in der Außenluft; Messen von Innenraumluftverunreinigungen - Gaschromatographische Bestimmung organischer Verbindungen - Aktive Probenahme durch Anreicherung auf Aktivkohle; Lösemittelextraktion.
354. H. P. König: "Halogenierte organische Verbindungen in der Umwelt," *VDI-Ber.* **745** (1989) 321 – 334.
355. M. Zingg, C. Schlatter: "Halogenierte organische Verbindungen in der Umwelt," *VDI-Ber.* **745** (1989) 245 – 255.
356. M. A. K. Khalil, R. A. Rasmussen, *J. Air Waste Manage Assoc.* **42** (1992) 810 – 813. Links
357. Bericht über die Eignungsprüfung einer Immissionsmeßeinrichtung für Benzol im Auftrag der Fa. Chrompack GmbH. TÜV Umwelttechnik, München, 1996.
358. Bericht über die Eignungsprüfung des Immissionsmeßgerätes für Benzol, Modell Airmo BTX 1000, der Fa. Airmotec AG, RW TÜV, Essen, 1996.
359. R. H. Brown: "The use of Diffusive Samplers for Monitoring of Ambient Air", *Pure Appl. Chem.* **65** (1993) 1859 – 1874. Links
360. K. H. Pannwitz: "Influence of Air Currents on the Sampling of Organic Solvent Vapours with Diffusive Samplers" in: *Diffusive Sampling—An Alternative Approach to Workplace Air Monitoring*, Royal Society of Chemistry, EC Publ. No. 10555 EN, Brussels/Luxembourg (1987), pp. 157 – 160.
361. DIN EN 14662-5 (08/2005) Ambient Air Quality — Standard Method for Measurement of Benzene Concentrations — Part 5: Diffusive Sampling Followed by Solvent Desorption and Gas Chromatography; German version EN 14662-5:2005.
362. DIN EN 14662-4 (08/2005) Ambient Air Quality — Standard Method for Measurement of Benzene Concentrations — Part 4: Diffusive Sampling Followed by Thermal Desorption and Gas Chromatography; German version EN 14662-4:2005.
363. H. U. Pfeffer, L. Breuer, K. Ellermann: Validierung von Passivsammlern für Immissionsmessungen von Kohlenwasserstoffen, Materialien Nr. 46, Landesumweltamt NRW (Ed.), Essen (1998).
364. J. König *et al.*, *Atmos. Environ.* **14** (1980) 609 – 613.
365. DIN ISO 12884 (12/2000): Ambient Air — Determination of Total (Gas and Particle Phase) Polycyclic Aromatic Hydrocarbons — Collection on Sorbent-Backed Filters with Gas Chromatographic/Mass Spectrometric Analysis (ISO 12884:2000).
366. DIN ISO 16362 (01/2006) Determination of the Particle-Phase Polycyclic Aromatic Hydrocarbons by High Performance Liquid Chromatography (ISO 16362:2005).
367. VDI 2464 Blatt 1 (02/2005): Messen von Immissionen — Messen von Innenraumluft — Messen von polychlorierten Biphenylen (PCB); GC/MS-Verfahren für PCB 28, 52, 101,138, 153, 180.
368. VDI 3498 Blatt 1 (07/2002): Messen von Immissionen — Messen von Innenraumluft — Messen von polychlorierten Dibenzo-p-dioxinen und Dibenzofuranen; Verfahren mit großem Filter.
369. VDI 3498 Blatt 2 (07/2002): Messen von Immissionen — Messen von Innenraumluft — Messen von polychlorierten Dibenzo-p-dioxinen und Dibenzofuranen; Verfahren mit kleinem Filter.
370. R. Westerholen, U. Stenberg, T. Alsberg, *Atmos. Environ.* **22** (1988) 1005 – 1010.
371. H. Yamasaki, K. Kuwata, H. Miyamoto: "Effects of Ambient Temperature on Aspects of Airborne Polycyclic Aromatic Hydrocarbons", *Envir. Sci. Technol.* **16** (1982) 189 – 194.
372. C. Rappe, L. O. Kjeller, P. Bruckmann, K. H. Hackhe, *Chemosphere* **17** (1988) 3 – 20. Links
373. P. Bruckmann *et al.*, *Chemosphere* **27** (1993) 707–720. Links
374. VDI 2090 Blatt 1 (01/2001): Messen von Immissionen — Bestimmung der Deposition von schwerflüchtigen organischen Substanzen — Bestimmung der PCDD/F-Deposition; Bergerhoff-Probenahme und GC/HRMS-Analyse.
375. VDI 2090 Blatt 2 (12/2002):Messen von Immissionen — Bestimmung der Deposition von schwerflüchtigen organischen Substanzen — Bestimmung der PCDD/F-Deposition; Trichter-Adsorber-Probenahme und GC/HRMS-Analyse.
376. G. T. Hunt, B. E. Maisel, *J. Air Waste Manage Assoc.* **42** (1992) 672 – 680. Links
377. Umweltforschungsplan des Bundesministers für Umwelt Naturschutz und Reaktorsicherheit, Forschungsbericht 104 03 364/01: "Weiterführende Untersuchung zur Bildung von polybromierten Dioxinen und Furanen bei der

thermischen Bela-stung flammengeschützter Kunststoffe und Textilien," Feb. 1991.

378. E. Hiester: "Messung von polyhalogenierten Dioxinen und Furanen in Luft", *Aus der Tätigkeit der LIS* 1991, Essen 1992.
379. DIN 19739-2 (03/2003: Luftbeschaffenheit und Bodenbeschaffenheit — Messen der atmosphärischen Deposition organischer Spurenstoffe; Trichter-Adsorber-Verfahren — Teil 2: Bestimmung von polycyclischen aromatischen Kohlenwasserstoffen.
380. DIN 19739-2 (03/2003: Luftbeschaffenheit und Bodenbeschaffenheit — Messen der atmosphärischen Deposition organischer Spurenstoffe; Trichter-Adsorber-Verfahren — Teil 2: Bestimmung von polycyclischen aromatischen Kohlenwasserstoffen.
381. B. Zielinska, J. Arey, R. Atkinson, A. M. Winer, *Atmos. Environ.* **23** (1989) 223 – 229.
382. M. Oehme, S. Man, A. Mikalsen, P. Kirschmer, *Chemosphere* **15** (1986) 607 – 617. Links
383. H. Knoeppel *et al.*, in: *Proceedings of the First European Symposium on Physico-Chemical Behaviour of Atmospheric Pollutants*, Ispra (1980) 25 – 40.
384. A. R. Fernandez *et al.*, *Chemosphere* **25** (1992) 1311 – 1316.
385. W. N. Billings, T. F. Bidleman, *Environ. Sci. Technol.* **14** (1980) 679. Links
386. D. A. Kurtz (ed.): *Long Range Transport of Pesticides*, Lewis Publishers, Chelsea, Michigan, 1990.
387. C. Tashiro *et al.*, *Chemosphere* **18** (1989) 777–782.
388. VDI-Richtlinie 3484 Bl. 1 (11/2001): Messen von gasförmigen Immissionen — Messen von Innenraumluftverunreinigungen — Messen von Prüfgasen; Bestimmung der Formaldehydkonzentration nach dem Sulfit-Pararosanilin-Verfahren.
389. H. I. Schiff *et al.*, *Atmos. Environ.* **22** (1988) 1931 – 1939.
390. P. Kirschmer, P. Eynck: "Meßverfahren mit automatisierter Probenahme zur Bestimmung von Aldehyden in der Luft," *LIS-Ber.* **92** (1989) 1–34.
391. VDI-Richtlinie 3485 Bl. 1 (12/88): Messen von Phenolen, p-Nitroanilin-Verfahren, in [152].
392. B. Spiegelhalder, R. Preussmann, *IARC Sci. Publ.* **84** (1984) 411 – 414. Links
393. DIN EN 14662-4 (08/2005): Luftbeschaffenheit — Standardverfahren zur Bestimmung von Benzolkonzentrationen — Teil 4: Diffusionsprobenahme mit anschließender Thermodesorption und Gaschromatographie; Deutsche Fassung EN 14662-4:2005.
394. DIN EN 14662-5 (08/2005): Luftbeschaffenheit — Standardverfahren zur Bestimmung von Benzolkonzentrationen — Teil 5: Diffusionsprobenahme mit anschließender Lösemitteldesorption und Gaschromatographie; Deutsche Fassung EN 14662-5:2005.
395. W. Medrow in: *Gerüche,* Schriftenreihe der VDI-Kommission Reinhalt. Luft, vol. 12, Düsseldorf 1990.
396. M. T. Hellmann, F. H. Small, *JAPCA* **24** (1974) 979 – 982.
397. VDI-Richtlinie 3940 (10/1993): Bestimmung der Geruchsstoffimmissionen durch Begehungen, in [166].
398. EN 13725: 2003: Air Quality — Determinaton of Odor Concentration by Dynamic Olfactometry, CEN, Brüssel, 2003.
399. L. Janicke, U. Janicke, D. Ahrens, U. Hartmann, W. J. Müller: "Development of the Odour Dispersion Model AUSTAL 2000 G in Germany" in *Environmental Odour Management, VDI-Ber.* **1850** (2004) 411 – 417.
400. M. Reiser, M. Terrel, K. Fischer, M. Krauert, *VDI-Ber.* **1885** (2005) 133 – 140.
401. R. Both, K. Sucker, G. Winneke, E. Koch, *Water Sci. Technol.* **50** (2004) 83 – 92. Links
402. VDI-Richtlinie 3940 Bl. 1 (11/2003): Bestimmung von Geruchsstoffimmissionen durch Begehungen — Bestimmung der Immissionshäufigkeit von erkennbaren Gerüchen, Rastermessung in [166].
403. VDI-Richtlinie 3940 Bl. 3 (in preparation): Bestimmung von Geruchsstoffimmissionen durch Begehungen; Ermittlung von Geruchsintensität und hedonischer Geruchswirkung im Feld.
404. VDI-Richtlinie 3883 Bl. 1 (7/1997): Wirkung und Bewertung von Gerüchen — Psychometrische Erfassung der Geruchsbelästigung — Fragebogentechnik, in [166].
405. GIRL, 2004: Feststellung und Beurteilung von Geruchsimmissionen (Geruchsimmissions-Richtlinie — GIRL) in der Fassung vom 21.9.2004 mit Begründung und Auslegungshinweisen. Ministerium für Umwelt und Naturschutz, Landwirtschaft und Verbraucherschutz Nordrhein-Westfalen, Düsseldorf, Erlass: V-4-8851.4.4 vom 11.10.2004, 35 S.
406. M. Kamp, Gefahrstoffe **61** (2001) 133 – 137.
407. VDI-Richtlinie 3940 Bl. 2 (11/2003): Bestimmung von Geruchsstoffimmissionen durch Begehungen — Bestimmung der Immissionshäufigkeit von erkennbaren Gerüchen, Fahnenmessung.
408. R. H. Brown, *Pure Appl. Chem.* **65**, no. 8 (1993) 1859 – 1874. Links
409. Ambient Air Quality — Diffusive Samplers for the Determination of Concentrations of Gases and Vapours; Requirements and Test Methods — Part 1: General Requirements; German version EN 13528-1:2002
410. Ambient Air Quality — Diffusive Samplers for the Determination of Concentrations of Gases and Vapours; Requirements and Test Methods — Part 2: Specific Requirements and Test Methods; German version EN 13528-2:2002
411. Ambient Air Quality — Diffusive Samplers for the Determination of Concentrations of Gases and Vapours — Part 3: Guide to Selection, Use and Maintenance; German version EN 13528-3:2003
412. H.-U. Pfeffer, L. Breuer, K. Ellermann: Validierung von Passivsammlern für Immissionsmessungen von Kohlenwasserstoffen, Materialien Nr. 46, Hrsg.: Landesumweltamt Nordrhein-Westfalen, Essen 1998, S. 95 (ISSN 0947-5206).Validation of Passive Samplers for Measurements of Hydrocarbons in Ambient Air (English Translation), Materialien No. 46, Ed.: Landesumweltamt Nordrhein-Westfalen (North Rhine-Westphalia State Environment Agency), Essen 1998, pp. 92 (ISSN 0947-5206).

413. H.-U. Pfeffer, L. Breuer: “BTX Measurements with Diffusive Samplers in the Vicinity of a Cokery: Comparison Between ORSA-Type Samplers and Pumped Sampling”, *J. Environ. Monit.* **2** (2000) 483 – 486.
414. U. Pfeffer, L. Breuer: “Calibration and Use of Diffusive Samplers for BTX Measurements in Urban and Industrialised Areas of Germany” *International Conference Measuring Air Pollutants by Diffusive Sampling*, Montpellier, France, 26 – 28. September 2001, Abstract Book, p. 110.Conference Proceedings, European Communities 2002, pp. 241 – 245.
415. *International Conference Measuring Air Pollutants by Diffusive Sampling*, Montpellier, France, 26 – 28. September 2001, Conference Proceedings, European Communities 2002.
416. K. Leichnitz: *Prüfröhrchen-Meßtechnik*, Ecomed-Verlagsgesellschaft, Landsberg/Lech 1981.
417. *Dräger-Röhrchen Handbuch*, 11th ed., Drägerwerk Aktiengesellschaft, Lübeck 1994.
418. VDI-Richtlinie 3799, Bl. 2 (10/91), in [152].
419. VDI-Richtlinie 3957, Bl. 2 (01/2003), in [166].
420. M. Birkle, *VDI-Ber.* **1059** (1993) 49 – 65.
421. Gesetz zum Schutz vor schädlichen Umwelteinwirkungen durch Luftverunreinigungen, Geräusche, Erschütterungen und ähnliche Vorgänge (Bundes-Immissionsschutzgesetz-BImSchG) vom 15. März 1974, Bundesgesetzblatt, Teil 1, no. 27, Jahrgang 1974, pp. 721 ff., last version 11. September 2002, BGBl. I, pp. 3622 ff.
422. Erste Allgemeine Verwaltungsvorschrift zum Bundesimmissionsschutzgesetz (Technische Anleitung zur Reinhaltung der Luft – TA Luft) vom 27. 2. 1986, Gemeinsames Ministerialblatt, Ausgabe A, 37. Jahrgang, Feb. 28, 1986, pp. 95 ff.
423. Dreizehnte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes über Großfeuerungsanlagen – 13. BImSchV vom 22. Juni 1983, Bundesgesetzblatt I, pp. 719 ff.
424. Siebzehnte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes über Verbrennungsanlagen für Abfälle und ähnliche brennbare Stoffe – 17. BImSchV vom 23. 11. 1990, Bundesgesetzblatt 1990, no. 64, pp. 2545 ff.
425. Zweite Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Verordnung zur Emissionsbegrenzung von leichtflüchtigen Halogenkohlenwasserstoffen – 2. BImSchV) vom 10. Dezember 1990 BGBl. I, pp. 2694 ff.
426. Vierte Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes über genehmigungsbedürftige Anlagen – 4. BImSchV in der Fassung vom 24. 7. 1985 (BGBl. I, p. 1586) last version 28. Aug. (BGBl. I, pp. 1838, 2044).
427. Richtlinie VDI 2448, Blatt 1, April 1992, Planung von stichprobenartigen Emissionsmessungen an geführten Quellen.
428. Richtlinie VDI 2066, Blatt 1, Oct. 1975, Messen von Partikeln, Staubmessungen in strömenden Gasen, Gravimetrische Bestimmung der Staubbeladung – Übersicht.
429. Richtlinie VDI 2449, Blatt 1, Entwurf, Dec. 1991, Prüfkriterien von Meßverfahren, Ermittlung von Verfahrenskenngrößen für die Messung gasförmiger Schadstoffe (Immission).
430. Richtlinie VDI 2449, Blatt 2, Jan. 1987, Grundlagen zur Kennzeichnung vollständiger Meßverfahren, Begriffsbestimmungen.
431. RdSchr. vom 1. 3. 90, Bundeseinheitliche Praxis bei der Überwachung der Emissionen – Richtlinien über die Eignungsprüfung, den Einbau, die Kalibrierung und die Wartung von Meßeinrichtungen für kontinuierliche Emissionsmessungen, Gemeinsames Ministerialblatt, 41. Jahrgang 1990, pp. 226 ff.
432. RdSchr. d. BMU vom 26. 7. 1988 – IGI 2–556 134/4, Bundeseinheitliche Praxis bei der Überwachung der Emissionen, Richtlinie über die Auswertung kontinuierlicher Emissionsmessungen, GMBl. 1988, pp. 426 ff.
433. RdSchr. d. BMU vom 16. 10. 1992 – IGI 3–51 134/3, Bundeseinheitliche Praxis bei der Überwachung der Emissionen, Richtlinien über die Auswertung kontinuierlicher Emissionsmessungen nach der Verordnung über Verbrennungsanlagen für Abfälle und ähnliche brennbare Stoffe (17. BImSchV), GMBl. 1992, pp. 1137 ff.
434. RdSchr. d. BMU vom 28. 10. 1992 – IGI 3–51 134/2, Bundeseinheitliche Praxis bei der Überwachung der Emissionen und der Immissionen, Eignung von Meßeinrichtungen zur kontinuierlichen Überwachung von Emissionen/Bezugsgrößen, GMBl. 1992, pp. 1140 ff.
435. RdSchr. d. BMU vom 1. 7. 1992 – IGI 3–51 134/2, Bundesheinheitliche Praxis bei der Überwachung der Emissionen und der Immissionen. Eignung von Meßeinrichtungen zur kontinuierlichen Überwachung von Emissionen, Bezugsgrößen (Abgasvolumenstrom, Sauerstoff), Eignung von elektronischen Systemen zur Auswertung kontinuierlicher Emissionsmessungen, GMBl. 1992, pp. 294 ff.
436. K. Lützke: “Leitlinien zur Messung und Bewertung von Emissionen,” *VDI-Ber.* **608,** Aktuelle Aufgaben der Meßtechnik in der Luftreinhaltung, Kolloquium Heidelberg, Sept., 17 – 19, 1986, VDI-Kommission Reinhaltung der Luft, VDI-Verlag, Düsseldorf 1987.
437. Luftreinhaltung, Leitfaden zur kontinuierlichen Emissionsüberwachung, Vorschriften und Verfahren der Emissionsmeßtechnik unter Berücksichtigung der TA Luft 86 und Datenblätter eignungsgeprüfter Meßgeräte, 4th ed., *Ber. Umweltbundesamt (Ger.)* **11/90** (1990).
438. Richtlinie VDI 3950, Blatt 1, Entwurf Januar 1991, Kalibrierung automatischer Emissionsmeßeinrichtungen.
439. K. Lützke, H.-D. Burk: “Kalibrierung automatischer Emissions-Meßeinrichtungen – Das Konzept der Richtlinie VDI 3950”, VDI-Ber. 1059. *Aktuelle Aufgaben der Meßtechnik in der Luftreinhaltung*, Tagung Heidelberg, June 2 – 4, 1993, Kommission Reinhaltung der Luft im VDI und DIN, VDI-Verlag, Düsseldorf 1993.
440. Richtlinie VDI 2066, Blatt 4, Jan. 1989, Staubmessung in strömenden Gasen, Bestimmungen der Staubbeladung durch kontinuierliches Messen der optischen Transmission.
441. DIN 1306, Dichte (Begriffe), Dec. 1971.
442. DIN 1871, Gasförmige Brennstoffe und sonstige Gase, Teil 1 (1981), Teil 2 (1990).
443. Richtlinie VDE/VDI 3511, Feb. 1967, Technische Temperaturmessung.

444. Richtlinie VDE/VDI 3512, Blatt 2, Sept. 1972, Meßanordnungen für Temperaturmessungen.
445. Richtlinie VDE/VDI 3511, Blatt 5, Entwurf, June 1993, Technische Temperaturmessungen, Einbau von Thermometern.
446. Richtlinie VDE/VDI 3511, Blatt 1, Entwurf, Nov. 1991, Technische Temperaturmessung, Grundlagen und Übersicht über besondere Temperaturmeßverfahren.
447. K. Lützke, R. Wilkes: “Erprobung von Meßverfahren zur Durchführung der Großfeuerungsanlagen-Verordnung, Umweltforschungsplan des Bundesministers für Umwelt, Naturschutz und Reaktorsicherheit”, *Luftreinhaltung Forschungsbericht 88–104 02 164.*
448. RdSchr. d. BMU vom 1. 4. 1992, Bundeseinheitliche Praxis bei der Überwachung der Emissionen, Richtlinien über die Eignungsprüfung, den Einbau, die Kalibrierung und die Wartung von Meßeinrichtungen für kontinuierliche Emissionsmessungen, Mindestanforderungen bei der Eignungsprüfung, Messen von Bezugsgrößen, Feuchtegehalt, GMBl. 1992, pp. 377 ff.
449. RdSchr. d. BMU vom 30. 10. 1991 – IGI 3–51 134/4, Bundeseinheitliche Praxis bei der Überwachung der Emissionen und Immissionen, Eignung von Meß-einrichtungen zur kontinuierlichen Überwachung von Emissionen, GMBl. 1991, pp. 1045 ff.
450. RdSchr. d. BMU vom 7. 6. 1991 – IGI 2–51 134/4, Bundeseinheitliche Praxis bei der Überwachung der Emissionen und der Immissionen, Eignung von Meßeinrichtungen zur kontinuierlichen Messung von Bezugsgrößen, GMBl. 1991, pp. 526 ff.
451. RdSchr. d. BMU vom 1. 6. 1990 – IGI 2, Bundeseinheitliche Praxis bei der Überwachung der Emissionen und der Immissionen, Eignung von Meß-einrichtungen zur kontinuierlichen Überwachung von Emissionen, Eignung von Meßeinrichtungen zur kontinuierlichen Messung von Bezugsgrößen, GMBl. 1990, pp. 400 ff.
452. Richtlinie VDI 2066, Blatt 2, Aug. 1993, Messen von Partikeln, Manuelle Staubmessung in strömenden Gasen, gravimetrische Bestimmung der Staubbeladung, Filterkopfgeräte (4 m3/h, 12 m3/h).
453. Richtlinie VDI 2066, Blatt 3, 1993, Messen von Partikeln, Manuelle Staubmessung in strömenden Gasen, Gravimetrische Bestimmung der Staubbeladung, Filterkopfgerät (40 m3/h).
454. Richtlinie VDI 2066, Blatt 7, Aug. 1993, Messen von Partikeln, Manuelle Staubmessung in strömenden Gasen, Gravimetrische Bestimmung der Staubbeladung, Planfilterkopfgeräte.
455. Richtlinie VDI 3861, Blatt 1, Dec. 1989, Messen faserförmiger Partikeln, Manuelle Asbest-Staubmessung im strömenden Reingas, IR-spektro-graphische Bestimmung der Asbeststaub-Massenkonzentration.
456. Richtlinie VDI 2066, Blatt 8, Entwurf, 1993, Staubmessungen in strömenden Gasen, Messung der Ruß-zahl an Feuerungsanlagen für Heizöl EL.
457. Richtlinie VDI 2066, Blatt 6, Jan. 1989, Staubmessungen in strömenden Gasen, Bestimmung der Staubbelastung durch kontinuierliches Messen des Streulichtes mit dem Photometer KTN.
458. RdSchr. d. BMU vom 7. 11. 1990 – IGI 2–40 038–2, Bundeseinheitliche Praxis bei der Überwachung der Emissionen und Immissionen, Eignung von Meßeinrichtungen zur kontinuierlichen Überwachung von Emissionen.
459. Richtlinie VDI 2031, Oct. 1962, Feinheitsbestimmungen an technischen Stäuben.
460. W. Muhr: “Feinstaub – Bedeutung und Messung”, *Schriftenreihe RW-TÜV,*13. Heft.
461. K. Lützke, W. Muhr: “Erprobung von Emissionsmeßverfahren zur Feststellung von Korngrö-ßenfraktionen”, *RWTÜV Forschungsbericht* 81–104 02 121.
462. Richtlinie VDI 2066, Blatt 5, 1993, Staubmessung in strömenden Gasen, Fraktionierende Staubmessung nach dem Impaktionsverfahren – Kaskadenimpaktor.
463. K. Lützke: “Mit Kaskadenimpaktoren – Feinstaubmessungen an Industrieanlagen”, *Ind. Anz.* **104** (1982) no. 4, 22 – 25; no. 19, 24 – 26.
464. Richtlinie VDI 3868, Blatt 1, Entwurf, Oct. 1992, Messen der Gesamtemission von Metallen, Halbmetallen und ihren Verbindungen. Manuelle Messung in strömenden, emittierten Gasen, Probenahmesystem für partikelgebundene und filtergängige Stoffe.
465. Richtlinie VDI 3868, Blatt 2, Vorentwurf, Feb. 1992, Bestimmung der Gesamtemission von Metallen, Halbmetallen und ihren Verbindungen, Messen von Quecksilber, Atomabsorptionsspektrometrie mit Kaltdampftechnik.
466. Richtlinie VDI 2268, Blatt 1, Stoffbestimmung an Partikeln: Bestimmung der Elemente Ba, Ca, Cd, Co, Cr, Cu, Ni, Pb, Sr, Zn in emittierten Stäuben mittels atomspektrometischer Methoden.
467. Richtlinie VDI 2268, Blatt 2, Stoffbestimmung an Partikeln; Bestimmung der Elemente Arsen, Antimon und Selen in emittierten Stäuben mittels Atomabsorptionsspektrometrie nach Abtrennung über ihre flüchtigen Hydride.
468. Richtlinie VDI 2268, Blatt 3, Stoffbestimmung an Partikeln; Bestimmung des Thalliums in emittierten Stäuben mittels Atomabsorptionsspektrometrie.
469. Richtlinie VDI 2268, Blatt 4, Stoffbestimmung an Partikeln; Bestimmung der Elemente Arsen, Antimon und Selen in emittierten Stäuben mittels Graphitrohr-Atomabsorptionsspektrometrie.
470. Richtlinie VDI 2462, Blatt 1, Feb. 1974, Messen gasförmiger Emissionen, Messen der Schwefeldioxid-Konzentration. Jod-Thiosulfat-Verfahren.
471. Richtlinie VDI 2462, Blatt 2, Feb. 1974, Messung gasförmiger Emissionen, Messen der Schwefeldioxidkonzentration, Wasserstoffperoxid-Verfahren, Titrimetrische Bestimmung.
472. Richtlinie VDI 2462, Blatt 3, Feb. 1974, Messung gasförmiger Emissionen, Messen der Schwefeldioxidkonzentration, Wasserstoffperoxid-Verfahren, gravimetrische Bestimmung.
473. Richtlinie VDI 2462, Blatt 8, March 1985, Messen gasförmiger Emissionen, Messen der Schwefeldioxid-Konzentration, H2O2Thorin-Methode.
474. Richtlinie VDI 2462, Blatt 7, March 1985, Messen gasförmiger Emissionen, Messen der Schwefeltrioxid-Konzentration, 2-Propanol-Verfahren.
475. RdSchr. d. BMU vom 26. 3. 1991, Bundeseinheitliche Praxis bei der Überwachung der Emissionen, Eignung von

Meßeinrichtungen zur kontinuierlichen Überwachung von Emissionen, GMBl. 1991, pp. 470 ff.

476. Richtlinie VDI 2456, Blatt 1, Dec. 1973, Messung gasförmiger Emissionen, Messen der Summe von Stickstoffmonoxid und Stickstoffdioxid, Phenoldisulfonsäure-Verfahren.
477. Richtlinie VDI 2456, Blatt 2, Dec. 1973, Messung gasförmiger Emissionen, Messen der Summe von Stickstoffmonoxid und Stickstoffdioxid, TitrationsVerfahren.
478. Richtlinie VDI 2456, Blatt 8, Jan. 1986, Messung gasförmiger Emissionen, Analytische Bestimmung der Summe von Stickstoffmonoxid und Stickstoffdioxid, Natriumsalicylat-Verfahren.
479. Richtlinie VDI 2456, Blatt 10, Nov. 1990, Messen gasförmiger Emissionen, Analytische Bestimmung der Summe von Stickstoffmonoxid und Stickstoffdioxid, Dimethylphenol-Verfahren.
480. Richtlinie VDI 2456, Blatt 6, May 1978, Messen gasförmiger Emissionen, Messen der Summe von Stickstoffmonoxid und Stickstoffdioxid als Stickstoffmonoxid unter Einsatz eines Konverters.
481. Richtlinie VDI 2456, Blatt 7, April 1981, Messen gasförmiger Emissionen, Messen von Stickstoffmonoxid-Gehalten, Chemiluminiszenz-Analysatoren (Atmosphärendruckgeräte).
482. Richtlinie VDI 2456, Blatt 5, May 1978, Messen gasförmiger Emissionen, Messen von Stickstoffmonoxid-Gehalten, Chemiluminiszenz-Analysator, Thermo Electron Modell 10.
483. Richtlinie VDI 2459, Blatt 7, Entwurf, Jan. 1990, Messen gasförmiger Emissionen, Messen der Kohlenmonoxid-Konzentration, Jodpentoxid-Verfahren.
484. Richtlinie VDI 2470, Blatt 1, Oct. 1975, Messen gasförmiger Emissionen, Messen gasförmiger Fluor-Verbindungen, Absorptions-Verfahren.
485. Richtlinie VDI 3488, Blatt 1, Dec. 1979, Messen gasförmiger Emissionen, Messen der Chlorkonzentration, Methylorange-Verfahren.
486. Richtlinie VDI 3488, Blatt 2, Nov. 1980, Messen gasförmiger Emissionen, Messen der Chlorkonzentration, Bromid-Jodid-Verfahren.
487. Richtlinie VDI 3480, Blatt 1, July 1984, Messen gasförmiger Emissionen, Messen von Chlorwasserstoff, Messen der Chlorwasserstoff-Konzentration von Abgas mit geringem Gehalt an chloridhaltigen Partikeln.
488. Richtlinie VDI 3480, Blatt 2, Jan. 1992, Messen gasförmiger Emissionen, Messen von Chlorwasserstoff, kontinuierliches selektives Messen von Chlorwasserstoff mit dem SPECTRAN 677 IR.
489. Richtlinie VDI 3480, Blatt 3, Jan. 1992, Messen gasförmiger Emissionen, kontinuierliches Messen von gasförmigen anorganischen Chlorverbindungen mit dem ECOMETER.
490. Richtlinie VDI 3486, Blatt 1, April 1979, Messen gasförmiger Emissionen, Messen der Schwefelwasserstoff-Konzentration, Potentiometrisches Titrations-Verfahren.
491. Richtlinie VDI 3486, Blatt 2, April 1979, Messen gasförmiger Emissionen, Messen der Schwefelwasserstoff-Konzentration, Jodometrisches Titrations-Verfahren.
492. Richtlinie VDI 3486, Blatt 3, Nov. 1980, Messen gasförmiger Emissionen, Messen der Schwefelwasserstoff-Konzentration, Colorimetrisches Verfahren (Monocolor-Analysator).
493. Richtlinie VDI 2461, Blatt 1, March 1974, Messung gasförmiger Immissionen, Messen der Ammoniak-Konzentration, Indophenol-Verfahren.
494. Richtlinie VDI 2461, Blatt 2, May 1976, Messung gasförmiger Immissionen, Messen der Ammoniak-Konzentration, NESSLER-Verfahren.
495. Bericht Nr.: 936/800 010, Modellhafte Eignungsprüfung der Mehrkomponenten-Emissionsmeßeinrichtung OPSIS AR für NH_3, Institut für Umweltschutz und Energietechnik, Köln.
496. RdSchr. d. BMU v. 2. 6. 1993 – IGI 3–51 134/2 – Bundeseinheitliche Praxis bei der Überwachung der Emissionen und der Immissionen, Eignung von Meßeinrichtungen zur kontinuierlichen Überwachung von Emissionen, GMBl. 1993, pp. 467 ff.
497. Richtlinie VDI 3481, Blatt 2, April 1980, Messen gasförmiger Emissionen; Bestimmung des durch Adsorption am Kieselgel erfaßbaren organisch gebundenen Kohlenstoffs.
498. Richtlinie VDI 3481, Blatt 4, Vorentwurf (Arbeitspapier), Bestimmung des durch Adsorption an Kieselgel erfaßbaren organisch gebundenen Kohlenstoffs in Abgasen mit höherem Wassergehalt.
499. Richtlinie VDI 3481, Blatt 1, Aug. 1975, Messung gasförmiger Emissionen, Messen der Kohlenwasserstoff-Konzentration, Flammen-Ionisations-Detektor (FID).
500. Richtlinie VDI 3481, Blatt 3, Entwurf, Sept. 1992, Messen gasförmiger organischer Verbindungen, insbesondere von Lösemitteln, mit dem Flammen-Ionisations-Detektor (FID).
501. Richtlinie VDI 3481, Blatt 6, Entwurf, Sept. 1992, Messung gasförmiger Emissionen, Auswahl und Anwendung von C-Summenverfahren.
502. K. Lützke: “Emissionsmessung leicht- und schwerflüchtiger halogenierter Verbindungen; Halogenierte organische Verbindungen in der Umwelt: Herkunft, Messung, Wirkung, Abhilfe-Maßnahmen,” *VDI-Ber.* **745,** Tagungsbericht der VDIKommission Reinhaltung der Luft; Kolloquium Mannheim, April 25 – 27, 1989, VDI-Verlag, Düsseldorf 1989.
503. Richtlinie VDI 3499, Blatt 1, 1993, Messen von Emissionen – Messen von Reststoffen, Messen von polychlorierten Dibenzodioxinen und -furanen im Rein- und Rohgas von Feuerungsanlagen mit der Verdünnungsmethode, Bestimmung in Filterstaub, Kesselasche und in Schlacken.
504. Richtlinie VDI 3499, Blatt 2, Entwurf, 1993, Messen von Emissionen, Messen von polychlorierten Dibenzo-*p*-dioxinen und Dibenzofuranen, Filter/Kühler-Methode.
505. Richtlinie VDI 3499, Blatt 3, Entwurf, 1993, Messen von Emissionen, Messen von polychlorierten Dibenzo-*p*-dioxinen und Dibenzofuranen an indu-striellen und gewerblichen Anlagen, Kondensationsmethode – Gekühltes Absaugerohr.
506. Richtlinie VDI 3499, Blatt 4, Entwurf, 1993, Messen von Emissionen, Messen von polychlorierten Dibenzodioxinen

und Dibenzofuranen in Emissionen von Verbrennungsanlagen und bei anderen Verbrennungsprozessen, Polyurethan-Adsorptions-Methode.

507. P. Kirschmer, W. Mülder, P. Eynk: “Comparison of Sample Preparation and Extraction Procedure for the Analysis of Polychlorinated Dibenzo-p-dioxins in Deposited Particulate Matter and Balkdeposition,” *Chemosphere* **24** (1992) no. 5, 575–580. Links
508. Richtlinie VDI 2464, Blatt 1, Vorentwurf, 1992, Messen von Immissionen, Messen von Innenraumluft, Messen von polychlorierten Biphenylen.
509. Richtlinie BDI 2457, Blatt 2, Vorentwurf, Messen gasförmiger Emissionen, Gaschromatographische Bestimmung organischer Verbindungen, Probenahme durch Absorption in tiefkaltem Lösemittel, (2-(2-Methoxylethoxy)ethanol, Methyldiglykol).
510. Richtlinie VDI 3482, Blatt 4, Nov. 1984, Messen gasförmiger Immissionen, Gaschromatographische Bestimmung organischer Verbindungen mit Kapillarsäulen, Probenahme durch Anreicherung an Aktivkohle – Desorption mit Lösemittel.
511. Richtlinie VDI 3482, Blatt 6, July 1988, Messen gasförmiger Immissionen, Gaschromatographische Bestimmung organischer Verbindungen – Probenahme durch Anreicherung – Thermische Desorption.
512. Richtlinie VDI 3864, Blatt 2, Vorentwurf, April 1993, Messen gasförmiger Immissionen, Gaschromatographische Bestimmung von leichtflüchtigen halogenierten Kohlenwasserstoffen, Probenahme durch Adsorption, Thermische Desorption.
513. Richtlinie VDI 3873, Blatt 1, Nov. 1992, Messen von Emissionen, Messen von polycyclischen aromatischen Kohlenwasserstoffen (PAH) an stationären industriellen Anlagen – Verdünnungsmethode (RWTÜV-Verfahren) – Gaschromatographische Bestimmung.
514. Richtlinie VDI 3482, Blatt 5, Nov. 1984, Messen gasförmiger Immissionen, Gaschromatographische Bestimmung von aromatischen Kohlenwasserstoffen, Probenahme durch Anreicherung an Aktivkohle – Desorption mit Lösemittel.
515. Richtlinie VDI 2457, Blatt 1, Entwurf, March 1991, Messen gasförmiger Emissionen, Gaschromatographische Bestimmung organischer Verbindungen – Grundlagen.
516. Richtlinie VDI 2457, Blatt 5, June 1981, Messen gasförmiger Emissionen, Gaschromatographische Bestimmung von Benzol, Toluol und Xylol.
517. Richtlinie VDI 2467, Blatt 2, Aug. 1991, Messen gasförmiger Immissionen, Messen der Konzentration primärer und sekundärer aliphatischer Amine mit der Hochleistungs-Flüssigkeits-Chromatographie (HPLC).
518. Richtlinie VDI 3862, Blatt 1, Dec. 1990, Messen gasförmiger Emissionen, Messen aliphatischer Aldehyde (C1 bis C3) nach dem MBTH-Verfahren.
519. Richtlinie VDI 3862, Blatt 2, Entwurf, Messen gasförmiger Emissionen, Messen aliphatischer und aromatischer Aldehyde und Ketone nach dem DNPH-Verfahren, Acetonitril-Verfahren.
520. Richtlinie VDI 3862, Blatt 3, Entwurf, Messen gasförmiger Emissionen, Messen aliphatischer und aromatischer Aldehyde und Ketone nach dem DNPH-Verfahren, Tetrachlorkohlenstoff-Methode.
521. Richtlinie VDI 3863, Blatt 1, April 1987, Messen gasförmiger Emissionen, Messen von Acrylnitril, Gaschromatographisches Verfahren, Probenahme mit Gassammelgefäßen.
522. Richtlinie VDI 3863, Blatt 2, Feb. 1991, Messen gasförmiger Emissionen, Messen von Acrylnitril, Gaschromatographisches Verfahren, Probenahme durch Absorption in tiefkalten Lösemitteln.
523. Richtlinie VDI 3863, Blatt 3, Entwurf, Oct. 1988, Messen gasförmiger Emissionen, Messen von Acrylnitril, Adsorption an Aktivkohle, Desorption durch Dimethylformamid (DMF).
524. Richtlinie VDI 3953, Blatt 1, Entwurf, April 1991, Messen gasförmiger Emissionen, Messen von 1,3-Butadien, Gaschromatographisches Verfahren, Probenahme durch Adsorption an Aktivkohle, Dampfraumanalyse.
525. Richtlinie VDI 3493, Blatt 1, Nov. 1982, Messen gasförmiger Emissionen, Messen von Vinylchlorid, Gaschromatographisches Verfahren, Probenahme mit Gassammelgefäßen.
526. Richtlinie VDI 3494, Blatt 1, Entwurf, May 1988, Messen gasförmiger Immissionen, Messen von Vinylchlorid-Konzentrationen, Gaschromatographische Bestimmung, Manuelle und automatische Dampfraumanalyse.
527. Richtlinie VDI 3494, Blatt 2, April 1986, Messen gasförmiger Immissionen, Messen von Vinylchlorid-Konzentrationen, Gaschromatographische Bestimmung mit der Trennsäulenschalteinrichtung für Live-Chromatographie.
528. Richtlinie VDI 3881, Blatt 1, May 1986, Olfaktometrie, Geruchsschwellenbestimmung – Grundlagen.
529. Richtlinie VDI 3881, Blatt 2, Jan. 1987, Olfaktometrie, Geruchsschwellenbestimmung – Probenahme.
530. Richtlinie VDI 3882, Blatt 1, Oct. 1992, Olfaktometrie, Bestimmung der Geruchsintensität.
531. Richtlinie VDI 3881, Blatt 3, Nov. 1986, Olfaktometrie, Geruchsschwellenbestimmung, Olfaktometer mit Verdünnung nach dem Gasstrahlprinzip.
532. Richtlinie VDI 3881, Blatt 4, Entwurf, Dec. 1989, Olfaktometrie, Geruchsschwellenbestimmung, Anwendungsvorschriften und Verfahrenskenngrößen.
533. F. Baum: *Luftreinhaltung in der Praxis*, Oldenbourg Verlag, München 1988.
534. H. Brauer (ed.): *Handbuch des Umweltschutzes und der Umweltschutztechnik*, **vol. 3**, Springer Verlag, Heidelberg 1996.
535. R. C. Brown: *Air Filtration*, Pergamon Press, Oxford 1993.
536. F. A. L. Dullien: *Introduction to Industrial Gas Cleaning*, Academic Press, San Diego 1989.
537. W. Licht: *Air Pollution Control Engineering*, Marcel Dekker, New York 1988.
538. F. Löffler: *Staubabscheiden*, Thieme-Verlag, Stuttgart 1988.
539. F. Löffler, H. Dietrich, W. Flatt: *Dust Collection with Bag Filters and Envelope Filters*, Vieweg, Braunschweig 1988.

540. K. R. Parker: *Applied Electrostatic Precipitation*, Blackie, London 1997.
541. J. Seville: *Gas Cleaning in Demanding Applications*, Blackie, London 1997.
542. VDI-Richtlinie 2442, Abgasreinigung durch thermische Verbrennung, VDI-Handbuch Reinhaltung der Luft, VDI-Verlag, Düsseldorf 1985.
543. Erste allgemeine Verwaltungsvorschrift zum BImSchG (Technische Anleitung zur Reinhaltung der Luft, TA-Luft), Feb. 27, 1986.
544. H. Meier zu Köcker: “Abgasreinigung durch thermische Verbrennung,” part I, “Thermische Stabilität im Wärmetauscherrohr,” *Staub Reinhalt. Luft* **38** (1978) 101 – 105.
545. H. Meier zu Köcker: “Abgasreinigung durch thermische Verbrennung,” part II, “Thermische Abluftverbrennung mit indirekter Beheizung,” *Staub Reinhalt. Luft* **38** (1978) 167 – 206.
546. H. Meier zu Köcker, F. Weger: “Abgasreinigung durch thermische Verbrennung,” part V, “Reaktionsverhalten stickstoffhaltiger Schadstoffe,” *Staub Reinhalt. Luft* **40** (1980) 245 – 249.
547. H. Meier zu Köcker, G. Müller: “Abgasreinigung durch thermische Verbrennung,” part VI, “Reaktionsverhalten halogenhaltiger Schadstoffe,” *Staub Reinhalt. Luft* **40** (1980) 280 – 283.
548. H. Meier zu Köcker, F. Weger, R. Vögtlin: “Abgasreinigung durch thermische Verbrennung – Reaktionen im Röhrenwärmetauscher einer TNV-Anlage,” *Staub Reinhalt. Luft* **42** (1982) 287 – 292.
549. H. Meier zu Köcker: “Verbrennung N-haltiger Verbindungen,” *VDI-Ber.* **525** (1985) 333 – 346.
550. J. K. Lesmana, E. Schütt, H. Meier zu Köcker: “Abgasreinigung durch thermische Verbrennung – Reaktionsmodell der Carbonsäureverbrennung,” *Staub Reinhalt. Luft* **42** (1982) no. 5, 208 – 214.
551. H. Meier zu Köcker, J. Lesmana: “*Stoffspezifische Einflüsse auf den Reaktionsverlauf der thermischen Nachverbrennung,*” VDI Ber. 423, (1981) 61 – 65.
552. O. Carlowitz: “Grundlagen der thermischen Abgasreinigung,” *Tech. Mitt.* **82** (1989) no. 5, 325–332.
553. H. Meier zu Köcker, J. Dühr: “Abgasreinigung durch thermische Verbrennung,” part III, “Thermische Abluftverbrennung mit Flammenbeheizung,” *Staub Reinhalt. Luft* **38** (1978) 370 – 374.
554. D. Siegel: “Anwendung von Fackeln und Ergebnisse von Emissionsmessungen,” *VDI-Ber.* **525** (1985) 365 – 380.
555. H. Seifert, R. Becker, G. Hemmer: “Verfahrenstechnische Lösungen bei der thermischen Abgasreinigung,” *VDI-Ber.* **1034** (1993) 39 – 68.
556. W. Hünig: “Sicherheitstechnische Maßnahmen bei der thermischen Abluftreinigung,” *Tech. Mitt.* **82** (1989) no. 5, 333 – 343.
557. O. Carlowitz: “Wirtschaftlichkeitsbetrachtungen bei der Gestaltung von thermischen Abgasreinigungsanlagen,” *Tech. Mitt.* **82** (1989) no. 5, 344 – 351.
558. St. Jauns, O. Carlowitz, M. Jäger: “Realisierung hoher Abluftvorwärmtemperaturen in thermischen Nachverbrennungsanalgen mit Hilfe von Festbettregeneratoren,” *Gas Wärme Int.* **40** (1991) no. 1, 59 – 65.
559. T. Schmidt, M. Hoffmeister, H. Falke: “Entfernung organischer Lösemittel und Ammoniak aus Abluft durch Adsorption und katalytische Oxidation,” *VDI-Ber.* **1034** (1993) 139 – 175.
560. T. Schmidt: “Anwendungsbeispiele für die katalytische Abgasreinigung,” *Tech. Mitt.* **82** (1989) no. 5, 360 – 369.
561. J. Chen, “Lower operating temperatures oxidizing VOCs,” *Pollution Engineering* (1996).
562. Engelhard Corporation: “Focused Clean Air Solutions for Industry,” company brochure, Jselin, NJ 1996.
563. W. Golibrzuch: “Thermische Verfahren der Abgasreinigung – Einsatzbereiche, Anwendungsbeispiele, Wirtschaftlichkeitsbetrachtungen,” *VDI-Ber.* **1034** (1993) 89 – 108.
564. O. Carlowitz: “Anlagen zur thermischen Abgasreinigung – Ansatzpunkte zur Umsetzung der Wärmenutzungsverordnung,” *Staub Reinhalt. Luft* **52** (1992) 325 – 330.
565. F. H. Ackermann: “Wirtschaftlicher Einsatz von thermischen Abluftreinigungsanlagen,” *VDI-Ber.* **525** (1985) 275 – 298.
566. W. Bitter: “Einführung in die TA-Luft im Hinblick auf den Einsatz von Verbrennungsverfahren zur Emissionsminderung,” *Tech. Mitt.* **82** (1989) no. 5, 320 – 324.
567. J. Guggenberger, A. Brandl: “Meßtechnische Überwachung von Verfahren zur thermischen und katalytischen Reinigung von Abgasen,” *VDI-Ber.* **525** (1985) 91 – 119.
568. J. M. Freidel: “Giftresistante Katalysatoren in der Abgasreinigung,” *VDI-Ber.* **1034** (1993) 177–186.
569. H. Müller, K. Deller: “Katalytische Reinigung CKW-haltiger Abgase aus dem VCM-Herstellungsverfahren,” *VDI-Ber.* **1034** (1993) Anhang.
570. K. H. Reitze: “Der ”Thermo-Reaktor“: Ein neuartiges System der thermischen Abgasreinigung,” *VDI-Ber.* **525** (1985) 323 – 332.
571. O. Ebbecke, K. Noll: “Abgasreinigung mit regenerativer thermischer Oxidation – Leistungsbereiche, Betriebsergebnisse,” *VDI-Ber.* **1034** (1993) 69 – 88.
572. Engelhard Corporation: “Regenerative Thermal Oxidation,” company brochure, Jselin, NJ 1997.
573. Braunschweiger Flammenfilter Armaturen- und Apparatebau GmbH, company brochure, Braunschweig 1990.
574. T. von Harpe: “Nichtkatalytische Reduktion von NOx mittels Harnstoff – Erste Ergebnisse,” in: *Aktuelle Probleme der Abluftreinigung,* VDI-Kommission Reinhaltung der Luft, April 12/13, 1989.
575. G. Mayer-Schwining, R. Knoche, G. Schaub: “Abscheidung von Dioxinen, Furanen und Schwermetallen bei der thermischen Abfallverwertung,” *VDI-Ber.* **1034** (1993) 107 – 122.
576. E. Richter: “Die katalytische NO_x-Reduktion – eine Übersicht,” *VDI-Ber.* **730** (1989) 121 – 156.
577. E. Koberstein: “Katalytische Motorabgasreinigung,” *VDI-Ber.* **525** (1985) 217 – 246.
578. Engelhard Kali-Chemie AutoCat GmbH, company brochure, Hannover 1985.

579. B. Engler, E. Koberstein: “Katalysatoren in der Abgasreinigung,” *VDI-Ber.* **730** (1989) 97 – 120.
580. Engelhard Corporation: “Engelhard NO_x Reduction Catalyst”, company brochure, Jeslin, NJ 1997.
581. E. Ruhl: “Katalytische Reinigung von Claus-Anlagen-Abgasen,” *VDI-Ber.* **730** (1989) 255 – 274.
582. R. Lell: “Katalytische Reduktion von SO_2-Emissionen mit Wertstoffgewinnung,” *VDI-Ber.* **1034** (1993) 203 – 226.
583. R. Kettner, T. Lübcke: “Betriebserfahrungen mit der KNV an Claus-Anlagen,” *VDI-Ber.* **730** (1989) 255 – 274.
584. K. Kirchner, A. Kaiczik: “Grundlagen der katalytischen Nachverbrennung,” *VDI-Ber.* **525** (1985) 119 – 145.
585. T. Schmidt: “Katalytische Verbrennung chlorierter Kohlenwasserstoffe,” *VDI-Ber.* **730** (1989) 201 – 237.
586. J. J. Spivey: “Complete Catalytic Oxidation of Volatile Organics,” *Ind. Eng. Chem. prod. Res. Dev.* **26** (1987) 2165 – 2180.
587. C. Herion, R. Meißner: “Katalytische Abgasreinigung in der chemischen Industrie,” *VDI-Ber.* **1034** (1993) 123 – 138.
588. EnviCat, Solvay Catalysts GmbH, company brochure, Hannover 1992.
589. U. Tanger: “Energiesparendes KNV-Verfahren durch neuartige Prozeßführung mit regenerativen Wärmeaustauschern,” *VDI-Ber.* **730** (1989) 189 – 200.
590. *Landolt-Börnstein*, **IV, 4 c.**
591. D. Tiegs *et al.*: “Activity Coefficients at Infinite Dilution,” *DECHEMA Chem. Data Ser.* **IX** (1986).
592. M. Schiller, J. Gmehling, *J. Chem. Eng. Data* **37** (1992) 37.
593. G.-G. Börger, *Chem. Ing. Tech.* **64** (1992) 905.
594. R. Schmidt, D. Wolf: “Absorption von Chlor und Chlorwasserstoff aus Abgasen,” *VDI Ber.* **253** (1976) 77 – 84.
595. K. H. Schaber, *Chem. Ing. Tech.* **59** (1987) 376.
596. R. Billet, J. Mackowiak, *Chem. Eng. Technol.* **11** (1988) 213.
597. G. K. Chen, *Chem. Eng.* **91** 1984, March 5, 40.
598. J. Roiron: “How to Solve the Problem of NOx Pollution from Nitric Acid Plants,” The British Sulphur Corp., London 1986.
599. Sigri GmbH, company brochure, Meitingen.
600. QVF Process Systems Ltd., Catalogue, Sect. 6, “Column Components,” Stone, United Kingdom, 1992.
601. H. Krill, H. Menig: “Aktivkohlen im Umweltschutz,” *Umwelt Aktuell* **1** (1990) 46 – 55.
602. H. Krill: “Adsorptive Abgasreinigung – eine Bestandsaufnahme,” *VDI-Ber.* **1034** (1993) 339 – 371.
603. M. Bank: *Basiswissen Umwelttechnik*, 1st ed., Vogel Verlag, Würzburg 1993.
604. W. Kast: *Adsorption aus der Gasphase – Grundlagen und Verfahren*, VCH, Weinheim, Germany 1988.
605. H. Jüntgen: “Physikalisch-chemische und verfahrenstechnische Grundlagen von Adsorptionsverfahren,” Haus der Technik, no. 404 (1977) 5 – 22.
606. DIN 66 131, Bestimmung der spezifischen Oberfläche von Feststoffen durch Gasadsorption nach Brunauer, Emmet und Teller, Beuth Verlag, Berlin.
607. W. Otten, E. Gail, T. Frey: “Einsatzmöglichkeiten hydrophober Zeolithe in der Adsorptionstechnik,” *Chem. Ing. Tech.* **64** (1992) no. 10, 915 – 925.
608. VDI-Richtlinie 3674, Abgasreinigung durch Adsorption – Prozeß- und Abgasreinigung.
609. A. Mersmann, G.-G. Börger, S. Scholl: “Abtrennung und Rückgewinnung von gasförmigen Stoffen durch Adsorption,” *Chem. Ing. Tech.* **63** (1991) no. 9, 892 – 903.
610. W. Otten, W. Kast: “Entwicklung von Auslegungsdiagrammen für Festbettadsorber,” *Chem. Ing. Tech.* **60** (1988) no. 11, 929 – 933.
611. H. Menig: *Emissionsminderung und Recycling – Grundlagen und Technologien*, 3rd ed., ecomed, Landsberg/Lech 1995.
612. R. Backes, H. Plinke: “Mikroporöse Polymere zur Reinigung lösemittelhaltiger Abluft im Fließbettverfahren,” *VDI Ber.* **1034** (1993) 473 – 486.
613. H. W. Bräuer: “Adsorptive Abluft- und Abgasreinigung mit der Festbett- und Wirbelschichttechnik,” *VDI Ber.* **1034** (1993) 445 – 460.
614. G. Konrad, G. Eigenberger: “Rotoradsorber zur Abluftreinigung und Lösemittelrückgewinnung,” *Chem. Ing. Tech.* **66** (1994) no. 3, 321 – 331.
615. K. Storp: “Abluft- und Abgasentschwefelung durch Adsorption und Katalyse an Aktivkohlen,” *Dechema Monogr.* **64** (1970) 91 – 94.
616. H. W. Bräuer, F. Fischer: “Die Rückgewinnung von Schwefelkohlenstoff bei gleichzeitiger Entfernung von Schwefelwasserstoff,” *Dechema Monogr.* **59** (1968) 173 – 189.
617. Lurgi AG, Sulfacid-Process, Desulfurization of SO2-laden Waste Gases with Direct Production of Dilute Sulfuric acid, company brochure, no. 1657 e/2.94/10, Frankfurt/Main 1994.
618. K. Schneiders, J. Zimmermann: “Erste Betriebserfahrungen mit dem BF/Uhde-Verfahren im Kraftwerk Arzberg,” *VGB Kraftwerkstech.* **69** (1989) no. 1, 66 – 69.
619. R. Lell: “Katalytische Reduktion von SO_2-Emissionen mit Wertstoffgewinnung,” *VDI Ber.* **1034** (1993) 187 – 202.
620. K.-A. Theis, W. Dolkemeyer, H. Kreusing: “Braunkohlenkoks – Anwendungsgebiete in der Umwelttechnik,” *ET – Z. Energiewirtsch. Recht Tech.* **4** (1989) 2 – 6.
621. E. Richter: “Die Abscheidung von polychlorierten Dioxinen und Furanen aus Abgasen mit Aktivkoksverfahren,” *Chem. Ing. Tech.* **64** (1992) no. 2, 125 – 136.
622. M. Hefni-Omar, E. Jury: “Abscheidung von Schadstoffen aus den Rauchgasen von Müllverbrennungsanlagen in

einem zirkulierenden Wirbelschichtreaktor," *Chem. Ing. Tech.* **64** (1992) 442–446.

623. P. Janositz: "Kat knackt Dioxin," *VDI Nachr*. (1992) no. 10, 26.
624. M. Eicken, W. Esser-Schmittmann, J. Lambertz, G. Ritter: "Braunkohlenkoks zur Rauchgasreinigung und Reststoffminimierung von Abfallverbrennungsanlagen," *BWK-TÜ-UT-special* Nr. 10/90, R 37 – 49.
625. J. Gottschalk: "Flugstromverfahren – ein Verfahren der Kohlenstoffadsorptionstechnik zur Minimierung der Restemissionen, " *Abfallwirtschaftsjournal* **4** (1992) Nr. 12, 997 – 1001.
626. E. Richter: "Aktivkokse als Adsorbentien/Katalysatoren in der Rauchgasreinigung," *Chemie-Technik* **19** (1990) Nr. 4, 122 – 126.
627. G. Mayer-Schwinning, H. Herden, H. W. Bräuer: "Zeolithe zur Dioxin/Furan- und Schwermetallabscheidung," *Staub Reinhaltung der Luft* **55** (1995) 183 – 188.
628. G. Feldhaus: *Feststellen und Beurteilen des Standes der Technik bei der Richtlinienarbeit*, no. 10, VDI-Kommission Reinhaltung der Luft, Düsseldorf 1980.
629. D. Bardtke: "Mikrobiologische Voraussetzungen für die biologische Abluftreinigung," *Kontakt Stud.* **212** (1990) 1 – 12.
630. VDI 3478 E, Biologische Abgasreinigung – Biowäscher und Rieselbettreaktoren, VDI-Verlag, Düsseldorf 1994.
631. W. J. Moore, D. O. Hummel: *Physikalische Chemie*, De Gruyter, Berlin – New York 1973.
632. K. Kirchner, P. Krämer: "Zur Absorption und Oxidation von Schadstoffen mittels Bakterien-Suspensionen," *VT Verfahrenstech.* **17** (1983) no. 12, 675 – 681.
633. E. Schippert: "Biowäschertechnologie," *VDI Ber.* **1104** (1994) 39 – 56.
634. F. Sabo *et al.*: "Entwicklung von Hochleistungsbiofiltern – Optimierung der Materialdurchströmung," *Entsorgungspraxis* **9** (1992) 576 – 578.
635. F. Sabo: "Verfahrenstechnische Grundlagen der biologischen Abluftreinigung," *Kontakt Stud.* **212** (1990) 13 – 34.
636. VDI 3477, Biologische Abgasreinigung – Biofilter, VDI-Verlag, Düsseldorf 1991.
637. R. M. M. Diks, S. P. P. O. Ottengraf: "Verification Studies of an Simplified Model for the Removal of Dichloromethane from Waste Gases Using a Biological Trickling Filter," *Bioprocess Eng.* **6** (1991) part I: pp. 93 – 99, part II: pp. 131 – 140.
638. C. A. Gossen: "Abgasreinigung mit fixierten Bakterien im Rieselbett," Ph. D. thesis, TU München 1990.
639. R. M. M. Diks: "The Removal of Dichloromethane from Waste Gases in a Biological Trickling Filter," Dissertation, Eindhoven University of Technologie, Eindhoven 1992.
640. E. Schippert: "Abscheidung von Lösemitteln im Biowäscher," *VDI Ber.* **1034** (1993) 577 – 610.
641. S. Fröhlich: "Entscheidungshilfen und Kriterien für den Einsatz von Biowäschern zur Abluftreinigung und Geruchsminderung," *VDI Ber.* **1104** (1994) 81 – 93.
642. P. Kunz *et al.*: "Biosorber – ein neues Prinzip für die biologische Abgasreinigung," *VDI Ber.* **1104** (1994) 239 – 250.
643. M. Reiser *et al.*: "Kombination aus Biowäscher und Biomembranverfahren zur Reinigung von Abluft mit hydrophilen und hydrophoben Inhaltsstoffen," *VDI Ber.* **1104** (1994) 103 – 112.
644. VDI 3881, Blatt 1 – 4, Olfaktometrie, VDI-Verlag, Düsseldorf 1986.
645. VDI 3940, Bestimmung der Geruchsstoffimmission durch Begehung, VDI-Verlag, Düsseldorf 1993.
646. VDI 3882, Blatt 1, Bestimmung der Geruchsintensität, VDI-Verlag, Düsseldorf 1992.
647. VDI 3882, Blatt 2, Bestimmung der hedonischen Geruchswirkung, VDI-Verlag, Düsseldorf 1994.
648. B. J. Finlayson-Pitts, J. N. Pitts, Jr. *Atmospheric Chemistry: Fundamentals and Techniques*, John Wiley & Sons, New York 1986.
649. Umweltbereich Luft, Fonds der Chemischen Industrie, Frankfurt 1987.
650. P. Warneck: *Chemistry of the Natural Atmosphere*, Academic Press, San Diego 1988.
651. R. Zellner, K. J. Niemitz in *Dechema Monogr.* **104** (1987) 5.
652. P. Ott, D. Helf, K. Kirchner in [651] p.91.
653. C. Zetsch in [651], p.187.
654. W. Simmler: "Air – an Environmental Compartment," *Proceedings of the GDCh-Hauptversammlung*, Berlin 1987.
655. W. Simmler in [651], p.1.
656. W. Simmler: "Abiotic Degradation," *Chemie-Report 10/91*, Verband der Chemischen Industrie, Frankfurt 1991, p. 3.
657. R. Zellner, *UWSF – Z. Umweltchem. Ökotox.* **3** (1991) no. 1, 52.
658. R. Zellner, *Chem. Unserer Zeit* **27** (1993) no. 5, 230.
659. I. M. Campbell: *Energy and the Atmosphere*, John Wiley & Sons, New York 1977.
660. S. Chapman, *Mem. R. Meteor. Soc.* **3** (1930) 103.
661. K. H. Becker, *UWSF – Z. Umweltchem. Ökotox.* **3** (1991) no. 1, 48.
662. K. H. Becker, *UWSF – Z. Umweltchem. Ökotox.* **4** (1992) no. 1, 25.
663. Environmental Indicators, OECD Core Set, Organisation for Economic Cooperation and Development, Publications Service, Paris 1994.
664. DFG Deutsche Forschungsgemeinschaft: *Atmosphärische Spurenstoffe*, VCH Verlagsgesellschaft, Weinheim, Germany 1987.
665. J. D. Nriaga (ed.): "Gaseous Pollutants: Characterization and Cycling," *Advances in Environmental Science and Technology*, **vol. 24**, John Wiley & Sons, New York 1992.
666. P. Fabian: *Atmosphäre und Umwelt*, 4th ed., Springer Verlag, Berlin 1992.
667. T. E. Graedel: *Chemical Compounds in the Atmosphere*, Academic Press, New York 1978.

668. K. H. Becker in: *Atmosphärische Spurenstoffe und ihr physikalisch-chemisches Verhalten,* Springer Verlag, Berlin 1985.
669. O. Hutzinger, M. Blumich, M. Berg, K. v. d. Olie, *Chemosphere* **14** (1985) 581. Links
670. G. W. Israel in [651], p.147.
671. K. R. Spurny in [651], p.161.
672. C. Zetsch, *UWSF – Z. Umweltchem. Ökotox.* **3** (1991) no. 1, 59.
673. "Photochemical Smog," *OECD Report*, Organisation for Economic Cooperation and Development, Publications Service, Paris 1982.
674. K. H. Becker, W. Fricke, J. Löbel, U. Schurath in: *Air Pollution by Photochemical Oxidants,* Springer Verlag, Berlin 1985.
675. X. J. Crutzen, *Nature* **350** (1991) 380.
676. D. H. Ehalt, J. W. Drummond in: *Tropospheric Ozone,* Reidel Publ., Dordrecht 1988.
677. H. Güsten in: *Chemistry of Multiphase Atmospheric Systems,* Springer Verlag, Heidelberg 1986, p. 567.
678. R. Jaenicke, *Ber. Bunsen-Ges. Phys. Chem.* **82** (1978) 1198.
679. "Scientific Assessment of Stratospheric Ozone," *World Meteorological Organization Global Ozone Research and Monitoring Project*, Report 20, vol. 1, WMO, Geneva 1990.
680. P. A. Mayewski *et al.*, *Nature* **346** (1990) 554. Links
681. A. Sigg, A. Neftel, *Nature* **351** (1991) 557. Links
682. K. Bächmann, G. Fuchs in [651], p. 79.
683. "WMO and the Ozone Issue," *World Meteorological Organization*, Geneva 1992.
684. "Scientific Assessment of Stratospheric Ozone," *World Meteorological Organization* Report 25, WMO, Geneva 1992.
685. M. J. Prather, R. Watson, *Nature* **344** (1990) 729. Links
686. T. Mill: Reprint Extended Abstract, Division of Environmental Chemistry, Am. Chem. Soc., Boston, April 1990.
687. W. Klöpffer in G. Rippen (ed.): *Handbuch der Umweltchemikalien*, part 2, suppl. 26, no. 10/1994, ecomed-Verlag, Landsberg 1991.
688. R. G. Zepp, J. Hoigné, H. Bader, *Environ. Sci. Technol.* **21** (1987) 443. Links
689. W. Klöpffer, E.-G. Kohl, *Ecotoxicol. Environ. Saf.* **22** (1991) 67. Links
690. W. Klöpffer, *Sci. Total Environ.* **123/124** (1992) 145.
691. J. Hoigné: "Bildung und Bedeutung von Photooxidantien in der wässrigen Phase," *Proc. Int. Symp. Verteilung Wirkung Photooxidantien Alpenraum*, 1988.
692. W. Klöpffer, E. G. Kohl, *Ecotoxicol. Environ. Saf.* **26** (1993) 346. Links
693. R. Atkinson, *Chem. Rev.* **86** (1986) 69. Links
694. R. Atkinson, *J. Phys. Chem. Ref. Data Monograph 1*, Am. Inst. Physics, New York 1989.
695. W. Klöpffer: Contract R-67.233, Batelle Europe to VCI Frankfurt and UBA Berlin, Batelle Institute, Frankfurt 1989/90.
696. W. Klöpffer, B. Daniel: "Reaktionskonstanten zum abiotischen Abbau," *texts 57*, Umweltbundesamt, Berlin 1991.
697. R. Atkinson, *Environ. Sci. Technol.* **21** (1987) 305. Links
698. W. L. Dilling, S. J. Gonsior, G. V. Boggs, C. G. Mendoza, *Environ. Sci. Technol.* **22** (1988) 1447. Links
699. OECD-Test "Guideline on Photochemical Oxidative Degradation in the Atmosphere," Umweltbundesamt Berlin, draft, 2nd ed. 1987, final draft 1991.
700. OECD-Test "Guideline on Photochemical Transformation of Chemicals in Water," Umweltbundesamt Berlin, draft for OECD workshop, Berlin 1987.
701. R. Atkinson, W. P. L. Carter, *Chem. Rev.* **84**, (1984) 437. Links
702. K. Bächmann, G. Fuchs in [651], p. 79.
703. R. Atkinson, S. M. Aschmann, *Int. J. Chem. Kinet.* **17** (1985) 33. Links
704. C. J. Weschler, T. E. Graedel: "Theoretical Limitations on Heterogeneous Catalysis by Transition Metals in Aqueous Atmospheric Aerosols," in "Heterogeneous Atmospheric Chemistry," *Geophys. Monogr. Am. Geophys. Union* **26** (1982) 196.
705. V. Brezova, *et al.*, *J. Photochem. Photobiol. A* **59** (1991) 385.
706. W. Klöpffer in G. Rippen (ed.): *Handbuch der Umweltchemikalien,* part 1, ecomed-Verlag, Landsberg 1991.
707. W. Klöpffer: "Environmental Hazard Assessment of Chemicals and Products," part I, *ESP Environ. Sci. Pollut. Res.* **1** (1993) 47.
708. W. Klöpffer: "Environmental Hazard Assessment of Chemicals and Products," part II, *ESP Environ. Sci. Pollut. Res.* **1** (1994) 108.
709. W. Klöpffer: "Environmental Hazard Assessment of Chemicals and Products," part III, *ESP Environ. Sci. Pollut. Res.* **1** (1994) 179.
710. W. Klöpffer: "Environmental Hazard Assessment of Chemicals and Products," part IV, *ESP Environ. Sci. Pollut. Res.* **1** (1994) 272.
711. F. Korte, et al. *Lehrbuch der ökologischen Chemie,* Thieme-Verlag, Stuttgart 1987.

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

My Profile
WILEY InterScience
HOME
ABOUT US
CONTACT US
HELP
Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Soil

Standard Article

Stephen Nortcliff[1], Herwig Hulpke[2], Claus G. Bannick[3], Konstantin Terytze[3], Gerhard Knoop[2], Michael Bredemeier[4], Hubert Schulte-Bisping[5], Karl Auerswald[6], Norbert Litz[3], Robert Mayer[7], Alexander Stoy[8], Kassem Alef[9], Helmut Kerndorff[3], Gerd Crössmann[10], Thomas Eikmann[11], Volker Franzius[3], Detlef Grimsky[3], Norbert Möhlenbruch[12], Peter-Christoph Storm[13], Ulf Dworshak[14]

[1]University of Reading, Department of Soil Science, Reading, United Kingdom
[2]Bayer AG, WV Umweltschutz, Leverkusen, Federal Republic of Germany
[3]Umweltbundesamt, Berlin, Federal Republic of Germany
[4]Universität Göttingen, Forschungszentrum Waldökosysteme, Göttingen, Federal Republic of Germany
[5]Universität Göttingen, Institut für Bodenkunde, Göttingen, Federal Republic of Germany
[6]Technische Universität München, Lehrstuhl für Grünlandlehre, Freising, Federal Republic of Germany
[7]Universität Gh Kassel, Fachbereich Stadtplanung/Landschaftsplanung, Kassel, Federal Republic of Germany
[8]Fachhochschule Kiel, Fachbereich Landbau, Rendsburg, Federal Republic of Germany
[9]Universität Bayreuth, Lehrstuhl für Ökologische Chemie, Bayreuth, Federal Republic of Germany
[10]Landwirtschaftliche Untersuchungs- und Forschungsanstalt, Münster, Federal Republic of Germany
[11]Universität Gießen, Hygiene-Institut, Gießen, Federal Republic of Germany
[12]Rheinbraun AG, Abt. Forstwirtschaft, Erftstadt, Federal Republic of Germany
[13]Wangen im Allgäu, Federal Republic of Germany
[14]Rheinbraun AG, Abt. Forstwirtschaft, Erfstadt, Federal Republic of Germany

Copyright © 2006 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b07_613.pub2
Article Online Posting Date: December 15, 2006

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Definition
2. Ecological and Political Aspects of Soil Protection
3. Function and Utilization of Soil
3.1. Soil Ecosystems
3.1.1. Introduction
3.1.2. Ecological Soil Functions
3.1.2.1. Regulating Functions
3.1.2.2. Habitat Function for Organisms in and upon the Soil
3.1.3. Productive Function
3.1.3.1. Agriculture
3.1.3.2. Forestry
3.1.4. Other Soil Functions
3.1.4.1. Utilization of Wastes
3.1.4.2. Raw Material Deposits
3.2. Land Utilization and Soil Destruction (Statistics)
3.2.1. Introduction
3.2.2. Land Utilization
3.2.3. Soil Destruction
3.2.4. Forecast
4. Physical Damage to Soils
4.1. Compaction
4.2. Erosion
4.3. Building Over and Sealing
5. Effects of Anthropogenic Substances on Soil
5.1. Persistent Pollutants
5.1.1. Definition and Criteria
5.1.2. Sources and Entry Paths
5.1.3. Processes in Terrestrial Soils
5.1.4. Persistent Organic Compounds
5.1.4.1. Chlorinated Aliphatic Hydrocarbons
5.1.4.2. Aromatics and Chlorinated Aromatics
5.1.4.3. Nitroaromatics and Aromatic Amines
5.1.4.4. Polychlorinated Dibenzodioxins (PCDDs) and Dibenzofurans (PCDFs)
5.1.4.5. Polycyclic Aromatic Hydrocarbons (PAHs)
5.1.4.6. Polychlorinated Biphenyls (PCBs)
5.2. Acid Precipitation
5.2.1. Definitions
5.2.2. Sources of Acidity in Atmospheric Deposition
5.2.3. Source and Sink Processes for Acidity in the Atmosphere and in the Soil Environment
5.2.4. Direct Effects of Acid Deposition on Soils
5.2.5. Effects Mediated by Organisms (Indirect Effects)
5.2.5.1. Soil Acidification by Canopy Leaching
5.2.5.2. Acidification by Uptake of Nitrogen Compounds
5.2.5.3. Acidification by Accumulation of Organic Matter
5.2.5.4. Acidification by Humus Disintegration
5.2.6. Consequences of Soil Acidification
5.2.7. Field Observations of Soil Acidification
5.3. Fertilizers and Agricultural Chemicals
5.3.1. Fertilizers
5.3.1.1. Development of Fertilization
5.3.1.2. Legal Framework of Fertilization
5.3.1.3. Forms of Fertilizers
5.3.1.4. Effects of Fertilizers in the Soil
5.3.1.4.1. Nutrient Contents and Nutrient Losses

5.3.1.4.2. Soil Reaction
5.3.1.4.3. Humus Content
5.3.1.4.4. Heavy Metals and Organic Pollutants
5.3.2. Pesticides
5.3.2.1. Introduction
5.3.2.2. Side-Effects
5.4. Effects of Abandoned Waste Disposal Sites
5.4.1. Introduction
5.4.2. Material Contents in Groundwater Affected by Abandoned Waste Disposal Sites
5.4.3. Contamination Criteria and their Quantification
5.4.3.1. Contamination Factor
5.4.3.2. Emission Detection Frequency
5.4.3.3. Mean Emission Concentration
6. Measurement and Evaluation of Soil Pollution
6.1. Sampling and Investigation of Soil Pollutants
6.1.1. Introduction
6.1.2. Sampling
6.1.3. Sample Pretreatment
6.1.4. Sample Preparation
6.1.5. Measurement
6.1.6. Evaluation
6.1.7. Quality Assurance
6.1.8. Analytical Procedures for Selected Inorganic and Organic Soil Pollutants
6.1.8.1. Analysis of Cadmium and Thallium
6.1.8.1.1. Sample Preparation
6.1.8.1.2. Measurement
6.1.8.2. Analysis of Organic Soil Pollutants
6.1.8.2.1. Analysis of Polycyclic Aromatic Hydrocarbons
6.1.8.2.2. Petroleum Hydrocarbons
6.2. Evaluation Concept
6.2.1. Uptake Paths to Humans
6.2.1.1. Inhalation Uptake
6.2.1.2. Oral Uptake
6.2.1.3. Uptake via the Skin
6.2.2. Relevant Soil Uses
6.2.3. Evaluation Criteria
7. Remediation of Soil and Land
7.1. Remediation Methods
7.1.1. Introduction
7.1.2. Methods and Processes for Remediation
7.1.2.1. Overview
7.1.2.2. Soil Treatment Processes
7.1.2.2.1. Thermal Soil Treatment
7.1.2.2.2. Physicochemical Soil Treatment/Soil Washing
7.1.2.2.3. Microbiological Soil Treatment
7.1.2.2.4. Soil Vapor Extraction Processes
7.1.2.3. Safeguarding Technologies
7.1.3. Costs for Soil Remediation
7.1.4. Outlook
7.2. Recultivation
7.2.1. Definition
7.2.2. Demands on Recultivation
7.2.2.1. Recultivation Goals
7.2.2.2. Initial Substrata
7.2.3. Soil-Physical Properties
7.2.3.1. Silvicultural Recultivation
7.2.3.2. Agricultural Recultivation
7.2.4. Soil-Chemical Properties
7.2.4.1. Silvicultural Recultivation
7.2.4.2. Agricultural Recultivation
7.2.5. Soil Biology
8. Legal Aspects of Soil Protection
8.1. Foundations of Soil Protection Law
8.1.1. Land Law
8.1.2. Soil Protection Law
8.2. Foundations of Soil Protection Policy
8.2.1. Soil Protection in Germany
8.2.2. International Soil Protection
8.3. Constitutional Soil Protection
8.4. Soil Protection under Administrative Law
8.5. Soil Protection Goals
8.5.1. Soil
8.5.2. Protection of Soil Functions
8.5.3. Impairments of the Soil
8.5.4. Soil Protection by Provision and Aftercare
8.5.5. Direct and Indirect Soil Protection
8.5.6. Relative and Absolute Soil Protection
8.6. Soil Protection Measures
8.6.1. Types of Measure
8.6.2. Classification of Soil Protection Measures
8.6.3. Measures against Material Soil Impairments
8.6.4. Measures against Nonmaterial Soil Impairments
8.7. Development Trends
8.7.1. Further Development of German Soil Protection Law since 1985
8.7.2. The German Soil Protection Act 1998
8.7.3. Soil Protection in the Draft of an Environmental Code (UGB)

[Top of Page]

1. Definition

Stephen Nortcliff

Soil is normally considered as the fine earth which covers land surfaces as a result of the in situ weathering of rock materials or the accumulation of mineral matter transported by water, wind, or ice. The distinctive feature of soil is that to this weathered mineral material is added organic material. This organic material may be both living and dead. The dead organic matter will include little altered and freshly added dead plant roots and leaf and other plant litter, dead fauna, and organic material in various stages of decomposition from little modified relatively fresh materials to the complex decomposed material called humus. It is this mixture of mineral and organic material which gives the soils their distinctive characteristics. Across the surface of the earth there are many different types of soil which reflect, at least in part, varying combinations of mineral and organic matter and their differing responses — both separately and often in complex association — to different environmental conditions. Indeed soil (and the soil constituents), together with the plant life it supports, the rock on which it lies, and the climate it experiences, forms a finely balanced system.

Depending upon the context, the word “soil” may have very different meanings. A simple definition of soil is the material that plants grow in and which provides them with physical support, water, and nutrients. There are other more particular uses of the term soil. To the engineer, soil is the finely divided and relatively loose “rock” material at the earth's surface, and this overburden is often considered an inconvenience because it must be removed. The geologist calls this layer the regolith and geological investigations frequently begin below it. The hydrologist looks on the soil as if it were a large reservoir storing water to supply streams and rivers. The most widely held view of the soil, however, is as a medium for plant growth and the provision of food and fiber directly or through the intermediate stage of animals.

Soil is important as a medium for plant growth and for the support of much animal and human activity. The soil acts as a reservoir for nutrients and water providing the plants' needs for these requirements throughout their growth. The soil may also provide an environment for the breakdown and immobilization of materials added to the surface (in addition to the aforementioned plant and animal remains) such as fertilizers and pesticides and waste products such as sewage sludge, animal wastes and slurries, and composted refuse materials. The soil is a complex dynamic system in which the interactions of the biological, chemical, and physical environments results in the transformation of materials, possibly rendering initially harmful materials less dangerous and immobilizing others as a result of the interactions between these added materials and the organic and inorganic soil constituents. This immobilization may enable breakdown of the potentially dangerous materials to less dangerous forms. These interactions and transformations may be long term, over decades, medium term over months or years, short term between individual events such as rainstorms, or almost instantaneous.

The soil usually consist of a vertical sequence of layers or horizons differentiated by physical, chemical, or biological characteristics. This sequence of layers is normally described as the soil profile. Surface horizons are usually characterized by accumulations of organic materials as a result of the addition of plant and animal residues (it is, however, becoming increasingly apparent that organic additions in deeper layers from roots may have a significant role to play in the overall additions of organic material and in the operation of the soil processes). Surface horizons with a mixture of mineral and organic materials are usually identified as *A horizons* and may be further described by the amount and nature of the organic matter present within the horizon. Where there is a substantial proportion of organic material in the surface layer it may be called *O horizon*. The A and O horizons are generally considered the biologically most active part of the soil and are also subject to considerable changes as a result of the weather, being frequently exposed to wetting and drying cycles and to a wide range of temperatures, which results in high rates of activity in physical and chemical processes. Both organic and mineral materials are transformed in these surface layers and frequently some of the byproducts of these transformations are released and removed to lower layers within the soil (eluviation). Below the A horizons usually *B horizons* are encountered where the balance shifts from predominantly biological processes as in the A horizons to predominantly chemical and physical processes. Pedologists normally consider this zone to be that of peak activity of the pedological or soil-forming processes, and it is for this reason that many of the currently used soil classifications pay particular attention to the characteristics of the B horizon in the allocation of soils to classes. The B horizons are frequently subject to the maximum intensity of weathering processes and transformations and may also exhibit the accumulation of materials translocated from surface layers (illuviation) such as clays, organic matter, and iron and aluminum hydrous oxides. The description of B horizons frequently emphasises the degree of weathering and the nature of the accumulated materials. The *C horizon*, below the B, is the relatively unaltered material from which the A and B horizons may have developed.

The importance of the incorporation and breakdown of organic matter and the production of complex organic materials, as well as the weathering of mineral materials to produce secondary products (such as clay minerals) and to release minerals which may be taken up by plants or leached out of the system, provides the distinctive nature of the soil. In particular the large surface area associated with soil materials and their ability both to retain and exchange ions (both those added in fertilizers and those released during weathering and biological decomposition) and to absorb water are important properties which distinguish soil from rock materials.

The soil performs many functions. These include functions related to natural ecosystems, agricultural productivity, environmental quality, soil as source of raw materials and as base for buildings. Of these the agricultural productivity function is probably the most widely recognized and understood. This function of soil is to support plant and animal productivity whilst maintaining or enhancing water and air quality and also supporting human health and habitation. To perform this production function, the soil must be able to provide the following: a physical, chemical and biological context suitable for the survival and development of living organisms; the conditions for the regulation and partitioning of water flow, storage and recycling of nutrients and other elements; conditions to support biological activity and diversity for plant growth and animal productivity; an environment to filter, buffer, degrade, immobilize, and detoxify organic and inorganic substances; and provide mechanical support for living organisms and their structures. This brief list of requirements illustrates the complex demands placed upon soil if it is to maintain the productivity which is essential if the food demands of an ever growing global population are to be met. It further illustrates the need to understand the complex interactions between the various components within the soil and the interactions with other components of the terrestrial ecosystems. The productivity functions of soil is probably the most widely studied and understood, but the soil is increasingly expected to perform functions related to the maintenance of environmental quality, such as the filtering of toxic materials applied to the soil surface. If the soil is to perform these varied functions effectively and with no reduction in its overall quality the role of the soil, its components and processes must be investigated and understood. It must be ensured that in performing one function the soil's ability to perform other functions is not reduced or removed.

Soil is essential for many human activities. It is also a basic part of the natural environment. The development of humans and society since prehistoric times has been closely linked with an increasing ability to manage the soil to human benefit. This progress has been achieved by adjusting the balance between the soil and its natural environment. These adjustments have not always resulted in positive responses and benefits, and there are records throughout history describing soil destruction as a result of mismanagement and misuse of the soil, often as a result of the failure to understand the nature and complexity of the soil – environment relationships. Future use of the soil must take account of these relationships even in a society where we are increasingly able to manipulate our environment. As yet we know too little either of the nature of the soil – environment relationships or of the changes in these relationships and the response of the soil as the environmental conditions are changed. It is essential that this potential for change is acknowledged in any attempts to manipulate the soil – environment system.

[Top of Page]

2. Ecological and Political Aspects of Soil Protection

Herwig Hulpke

Soils are complex ecosystems that are strongly cross-linked to the other environmental compartments (water and air). Because of these linkages far reaching interactions with the other environmental compartments exist, through which soils have an important ecological mediator function. With respect to material transport processes, soils can be both sources and sinks.

In significant contrast to the other environmental compartments (air, water systems) humans took possession of soil from early times. The concept of “owned land”, which implies “possession of soil” underlines this fact. In some languages (e.g., German) land (Land) and soil (Boden) are often used as synonyms. Soil therefore represents the most important resource for human life in both its material and cultural aspects. Because soil is possessed as property, it has long been assumed that special care of the soil by the state could be dispensed with, since the soil's owners would already take care of it. This assumption may well be true in many cases for specific forms of use, such as agricultural use. But less care is felt when the utilization of the soil does not serve a single aim but several parallel or even conflicting purposes. Such multiple uses are typically the result of an increasing density of human population.

Soil is utilized by exploiting its functions. The great importance of soils for the existence of humankind is clearly suggested by the term “culture”, which is derived from the Latin word “colere” (to till).

The exploitation of certain soil functions does not only create benefits but can also cause damage. There are many examples for such conflicts. Agricultural use as opposed to planned use for housing or industrialization is a typical conflict of this type.

In the course of this, the use of the soil for nature conservation purposes or as a groundwater reservoir for the production of drinking water can be restricted or even endangered. Often, however, with cross-linked and complex ecosystems the conflicting interactions are more complicated and often neither transparent nor easy to understand. Nevertheless, when exploiting the soil it must always be borne in mind that the benefit from the use of one function can at the same time be detrimental to the use of another function. Soil protection is therefore chiefly a task whereby the utilization of soil functions has to be prudently selected. The properly targeted management of these uses must then be carried out in such a way that no unreasonable and unaccepted conflicts of the various intended uses will occur. At the same time the kind of use and its extent must also be taken into account. The understanding of soil protection as the management of competing claims for use by human beings is undoubtedly an anthropocentric view. But criticism raised against this view frequently fails to understand that this approach to action implies a large number of optional actions that can extend from the utilization of one or a few soil functions to the use of all conceivable or available functions. In the limiting case the utilization of any soil function can be entirely dispensed with. The use can therefore admit the range from “no functions” to “all.” As a result of this, and despite the anthropocentric approach, even extensive ecological requirements are taken into account. A system designed for the protection of soil functions consequently offers starting points for practicable as well as ecologically demanding measures.

With regard to its utilization and exploitation, the soil is the memory and archive of these actions. In this respect, pollutions arising from earlier uses (contaminated sites) play an important part in soil protection. As a result of the exploitation of soils in the past, restrictions often arise today on the utilization of some or all soil functions. Thus a soil that functions as a reservoir for groundwater regeneration and was contaminated in the past cannot be used as a water protection zone if through this contamination, owing to displacement processes, concentrations of contaminants appear in the groundwater making it unacceptable for drinking water catchment. The problem of exploitations in the past (contaminated sites) for soil protection is that the actual results of this exploitation often cannot be undone. Care must then be taken that this exploitation in the past has no harmful effects for present-day or future uses. The selection of appropriate uses therefore plays an important part. The precautionary principle, generally important in environmental protection, cannot apply here since, understandably, actions in the past cannot be cancelled by means of precautionary measures.

In the case of contaminated sites, the protection of soil functions can therefore not be managed by applying the precautionary principle. Instead, it must be orientated to the reduction or elimination of risks to soil functions intended for future utilization.

[Top of Page]

3. Function and Utilization of Soil

Claus G. Bannick, Konstantin Terytze, Gerhard Knoop, Michael Bredemeier, Hubert Schulte-Bisping

3.1. Soil Ecosystems

3.1.1. *Introduction*

Pedogenesis — the formation of soil — involves interactions between the lithosphere, hydrosphere, atmosphere, and biosphere (Fig. 1) [1]. At the earth's surface, the lithosphere becomes the pedosphere under the influence of weathering forces and organisms.

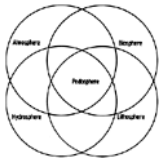

Figure 1. The soil in terms of the interaction of various spheres [1]

Thus soil is a complex, multiphase mixture of minerals, water, air, and organisms, together with products of their transformation and degradation. Depending on the climatic conditions and the nature of the underlying rocks, widely differing types of soil can be formed, whereby the rate of formation depends on how fast the rocks are weathered in the lower layers and on how rapidly biomass is degraded in the upper layers.

Lithosphere. The lithosphere is the rock mantle of the continents and the earth's surface beneath the oceans. The soil — the upper, inhabited part of the earth's crust — is formed from rocks by weathering, mineralization, decomposition, humification, texture formation, and displacement, under certain conditions (e.g., climate, leaf fall from vegetation). Weathering is the chemical, physical, and biological transformation of inorganic material such as basalt, granite, gneiss, carbonate, loess, and sand at the earth's surface by size reduction, degradation, transformation, and dissolution. With progressing soil development, the petrogenic starting materials are increasingly replaced by pedogenic materials such as clay minerals, which are more stable at the earth's surface.

Nine groups of minerals can be distinguished on the basis of chemical composition [2]:

Class I	Elements: metals (Ag, Au, . . .), metalloids (As, Sb, . . .), nonmetals (C, Se, . . .)
Class II	Sulfides: galena (PbS), sphalerite (ZnS)
Class III	Halides: halite (NaCl), sylvinite (KCl)
Class IV	Oxides, hydroxides: corundum (Al_2O_3), quartz (SiO_2), hematite (Fe_2O_3)
Class V	Carbonates, nitrates, borates: calcite ($CaCO_3$)
Class VI	Sulfates, chromates, molybdenates, tungstates: baryte ($BaSO_4$)
Class VII	Phosphates, arsenates, vanadates: apatite [$Ca_5[(F, Cl; OH)/(PO_4)_3]$]
Class VIII	Silicates: quartz (SiO_2)
Class IX	Organic compounds

As a result of the various soil-forming factors, characteristic horizontal layers, known as horizons, are formed, which differ in their biological, chemical, and physical properties. At the soil surface, the horizons contain larger proportions of organic material (O/A horizons), whereas the mineral component increases with increasing depth.

The horizons together make up the solum. The vertical distribution of the horizons in a two-dimensional cross section through the soil (soil profile) is the main criterion for classification of soils.

The smallest, undividable unit of a soil body is known as a pedon. Neighboring peda of the same soil type form a polypedon. The properties of the peda of a polypedon vary only slightly (e.g., horizon depth). On a larger scale, polypeda may be collected together as pedocomplexes.

Atmosphere and Hydrosphere. The spacial arrangement of the voids, the pore system, and the solid soil components determines the soil structure. The structures are classified in two main categories: primary structures and secondary structures (aggregates). In primary structures, aggregation is not apparent, due to uniform binding forces between the primary particles. Only in coherent structures are the primary particles permanently attached, forming a nonorganized mass, as in humus and clay soils. In secondary structures, primary particles are bound to one another far more strongly than to neighboring groups, thus forming secondary particles. Aggregates such as crumbs (1 – 10 mm) are formed by aggregation of individual grains. Segregates form by shrinkage from a coherent bulk. Interparticle binding can be by pure adhesion, or by cementation by mineral precipitates, inorganic colloids, roots, fungal hyphae, or bacterial slime. The soil solid phase can be separated into various particle sizes by classification processes such as sieving and sedimentation. The mass composition of the solid phase as a function of particle size is known as granularity or texture. Particles >2 mm make up the coarse fragment, and those <2 mm, the fine earth, which is subdivided into sand (63 – 2000 µm), silt (2 – 63 µm), and clay (<2 µm) fractions.

Soil pores are also classified according to size: large pores (>50 µm), medium pores (0.2 –50 µm), and fine pores (<0.2 µm). Large pores bind soil water only weakly and are easily drained. Medium pores can be penetrated by root hairs and microorganisms and store capillary water available to plants. Fine pores bind water so strongly that it is no longer available to plants. The soil contains a pore system formed by mineral particles generated by weathering. The two main types of particle packing (see Fig. 2), cubic loose and hexagonal close packed, influence various soil processes.

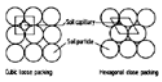

Figure 2. Packing of spherical particles [3]

Biosphere. The totality of the ecosystems of the earth's crust, in close interaction with the hydro-, litho-, and atmospheres, makes up the biosphere. The total organic fraction of the solid phase of the soil is divided into biomass, which includes all living organisms (soil flora and fauna), and the necromass, which contains all dead organisms and their transformation products. The latter category is derived from both underground (e.g., roots) and overground sources (e.g., leaf fall). The degradation products of the necromass are the humic substances which, depending on their solubility, are divided into the alkali-insoluble humins and the alkali-soluble humic acid precursors and humic acids (see Section Regulating Functions).

Within their biotope, the soil organisms form a community known as an edaphon. The edaphon is divided according to the size and way of life of the organisms. In the division according to morphological size [4] the edaphon is divided into microflora (bacteria, fungi, algae), microfauna (protozoa), mesofauna (nematoda, rotifera, tardigradae, mites, and collembola), macrofauna (annelids, snails, insects and their larvae), and megafauna (vertebrates); see Figure 3.

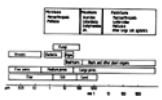

Figure 3. Classification of soil organisms by size [4]

The particular living conditions in the soil are coupled with typical modes of behavior of the organisms. Thus the group of soil burrowers (fodient edaphon) is formed mainly of organisms from the mega- and macrofauna (annelids, insects and their larvae) which generate their own living environment by burrowing and agitation of the soil. Serpentine edaphon move through the pores of the soil (rhizopods, annelids, collembola). They are restricted to open, air-containing pores and are found mainly in the upper soil layers. The soil swimmers (natant edaphon) are adapted to life in the water of the soil. They include some microflora and the majority of the mesofauna (rotifera, nematodes, tardigradae). Bacteria and fungi, which rapidly cover the walls of pores or enter soil voids belong to the sessile edaphon [5].

The soil flora and fauna introduce organic material into the soil system. Almost 100 % of the elemental composition is accounted for by C, O, H, N, and S. The central importance of the soil is due to the fact that fundamental ecological processes interact there. Thus the soil is an essential basis for all life processes. Soil functions are summarized in Figure 4.

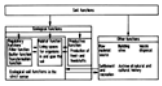

Figure 4. Soil functions

From the environmental viewpoint, the most important functions are the regulating and productive functions, known together as ecological soil functions. Other soil functions include raw material resources, land for building, recreational areas, and waste disposal.

The filter function is the mechanical retention of particules and colloids by the soil-pore system.

The buffer function includes the binding of dissolved substances by absorption and desorption, the inclusion and release of substances (including water) by pore spaces, and the in- and excorporation of substances by soil organisms. The habitat function is the ability of the soil to support colonies of organisms within and upon the soil. This includes the function of the soil as substrate for higher plants.

The productive function differs from the habitat function in that here the economic exploitation of the biomass is of primary importance.

The other soil functions are mainly or exclusively of anthropogenic nature. Here, the ecological functions are of minor importance compared to the anthropogenic exploitation or are indirectly exploited, for example, in the degradation or sorption of pollutants. In contrast to the other soil functions (raw material deposits, building, recreation, waste disposal), the regulatory and habitat functions can be defined solely on the basis of the natural pedological and ecological properties of the soil, regardless of the use to which it is put. The productive function can be distinguished on the basis of both ecological aspects (habitat for cultivated plants) and socio-economic criteria, since it is directly interlinked with agriculture and forestry.

3.1.2. *Ecological Soil Functions*

3.1.2.1. *Regulating Functions*

The regulating function of the soil depends on the regulation of natural material and energy cycles. It includes all material and nonmaterial introductions to the soil and the process therein which they initiate.

Filter Function. Filtration is defined as mechanical retention of material >100 nm in size [6]. Filtration occurs during the flow of aqueous suspension, through the pore systems.

The behavior of the soil water (hydraulic potential ϕ) is determined by the gravitation potential z (downwards) and the matrix potential Ψ (upwards). At equilibrium, these two potentials are equal, i.e., $\phi = (-\Psi) + (+z) = 0$. Addition of water (e.g., by precipitation) increases the matrix potential and displaces the equilibrium. This infiltration (⟶ Ecology and Ecotoxicology) is one of the most important effects counteracting erosion.

Decreased filter performance can arise through pore blockage due, for example, to compaction or clogging with mud. Pore blockage can also have deleterious effects on the transformation and habitat functions (stagnant water, air deficiency).

Buffer Function. The buffer function encompasses the following major areas:

1. Buffering of water balance
2. Buffering of acid inputs
3. Buffering of nutrients and pollutants

The buffering of water balance is largely a result of the ability of soils to store water in pore spaces by means of capillary forces. Since the pressure of a liquid (tension) is lower than atmospheric pressure, the liquid rises in a capillary until hydrostatic equilibrium is attained. The capillary pressure and hence the height to which the liquid rises increase with decreasing capillary diameter.

The water-storage capacity of a soil is characterized by the amount of water that can be retained for a given value of capillary pressure (pF characteristic). The pF value is defined as the logarithm of the negative capillary force ($-\Psi$), or the pressure in centimeters water column. Of particular importance for the habitat and productive functions is water available to plants, with pF = 1.8 – 4.2 (exploitable field capacity). For each soil type, at a given pF value, the quantity of water that can be retained against the force of gravity can be determined. The exploitable field capacity (in mm/dm) for some substrates is given in the following [7]:

Fine sand	13
Silt	25
Sandy loam	19
Clay	15

Buffering of acid inputs is achieved by various soil buffer systems, depending on the pH range:

Neutral (6.2 – 8.0)	carbon dioxide/ carbonate buffer
Acidic (4.2 – 6.2)	exchange buffer
Highly acidic (2.8 – 4.2)	aluminum buffer

Buffer systems consist of a weak acid or base and the corresponding salts, and maintain constant pH within a certain range. All soils are exposed to acids, both of natural (e.g., absorption of nutrient ions by plants with release of protons) and anthropogenic origin (e.g., fertilization, atmogenic inputs).

Carbonate-containing soils can buffer protons by formation of hydrogencarbonate

$$CaCO_3 + H^+ \rightleftharpoons Ca^{2+} + HCO_3^-$$

After exhaustion of the carbonate buffer, buffering is performed by the exchange buffer system, whereby all cations of the mineral and organic ion exchangers are replaced by protons. Below pH 4.2 in the aluminum buffer system, destruction of the crystal lattice structure and release of aluminum ions occurs

$$AlO(OH) \cdot H_2O + 3H^+ \rightarrow Al^{3+} + 3H_2O$$

The buffering of nutrients and pollutants affects most ecological soil functions. It is made possible by the presence of organic and inorganic exchangers, whereby the following adsorption phenomena can occur:

1. Van der Waals forces: relevant for organic pollutants, humus
2. Hydrogen bonds: relevant for organic pollutants, mineral surfaces
3. Electron donor – acceptor complexes: relevant for organic pollutants, humus
4. Hydrophobic bonding: relevant for organic pollutants, humus
5. Ion bonding: relevant for organic pollutants, inorganic pollutants and nutrients, humus, clay minerals
6. Covalent bonding: relevant for organic and inorganic pollutants

For nutrients and inorganic pollutants, ion formation is the most important process. Ion exchangers have positive or negative excess charge, thus binding anions or cations. Sorption of this type is reversible and an ion-concentration-dependent equilibrium is established in the soil solution. The sorption strength depends on the distance between the bonding partners, which is largely determined by their hydration shells, decreasing with increasing distance according to the Coulomb law $P \sim Q_1Q_2/r^2$.

In addition, the strength of sorption increases with increasing ion charge, which is taken into account in the Hofmeister ion series of decreasing hydration shell and increasing exchange potential:

$$Li^+ < Na^+ < k^+ < Rb^+ < Cs^+$$

$$Mg^{2+} < Ca^{2+} < Sr^{2+} < Ba^{2+}$$

The input of nutrients and inorganic pollutants into the soil, their dissolution in soil water, buffering by adsorption and desorption, uptake by plants, and loss of eluates can be depicted in nutrient and pollutant cycles (⟶ Ecology and Ecotoxicology – Nutrient Cycles and Productivity of Terrestrial and Aquatic Ecosystems). The ecological cycle for heavy metals is shown in Figure 5.

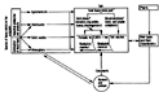

Figure 5. Heavy-metal cycle in the ecosystem [8]

[a] Can be used as indicator for accumulation. [b] Potentially biorelevant. [c] Momentary biorelevant increase of metal concentration in the dissolved phase as indicator for biotoxicity and bioacumulation.

Heavy metals can be bound in the soil in the following ways (R represents a solid/dissolved inorganic/organic compound, e.g., iron oxide, clay mineral, humic substance):

Solid state

Unspecifically adsorbed, e.g., $R\langle^{O}_{OCd^{2+}}$ (aq.)

Specifically adsorbed, e.g., $R-OCu^{+}$

Metal salts, e.g., $PbCO_3$, $Pb_3(PO_4)_2$, PbS

Oxides, e.g., As_2O_3

Bound in silicates e.g., tin in Hornblende

Dissolved

Aquo complexes, e.g., $Cd^{2+}_{(aq.)}$, $Pb^{2+}_{(aq.)}$

Inorganic complexes, e.g., AsO_4^{3+}, $Zn(OH)_3^{-}$, $Pb(CO_3)_2^{2-}$, $CdCl^{+}$

Organic complexes and chelates, e.g., $CH_3Hg^{+}_{(aq.)}$, R–Cu–R

The pollutant content of the soil is characterized by background values, which are representative values for the general background content of a substance or group of substances in the soil. When statistically determined they can be used to give the local, regional, or nationwide pollution of the soil. Table 1 lists values based on the evaluation of over 16 000 soil samples collected nationwide in Germany [9-11]. The samples were collected by a uniform method, and the 50th and 90th percentiles of the heavy-metal content determined for three types of land usage. The shown values are representative for the rural area.

Table 1. Heavy-metal contents in rural areas for arable land (90th percentile)

Substrate	As	Cd	Cr	Cu	Hg	Ni	Pb	Sb	Zn
Sand	3.0	<0.3		3.0	0.03	<3.0	13	<0.3	14
Sandstone	7.0	0.9		15	0.11	30	75	0.6	43
Loess	9.6	<0.3	149	25	0.16	44	51	1.1	89
Boulder clay	5.4	<0.3		14	0.08	19	32	0.9	76
Marshes									
Claystone	10	1.1		27	0.14	74	61	0.6	121
Limestone									
Basalt	5.0	0.8		71	0.1	339	49		168

Soil has to fulfill a buffer function to protect humans, plants, and groundwater from soil pollutants. Soil pollutants and the corresponding pathways are listed in Table 2.

Table 2. Priorities of relevant soil pollutants

Relevant pollutants in the pathway		
Soil–human	Soil–plant	Soil–groundwater
Arsenic	arsenic	heavy metals
Lead	lead	crop-protection agents
Cadmium	cadmium	volatile chlorinated hydrocarbons
Chromium	chromium	
Nickel	copper	
Mercury	mercury	
Vanadium	thallium	
Zinc	zinc	
Benzo[*a*]pyrene	aldrin/dieldrin	
Benzene	benzo[*a*]pyrene	
DDT (total)	endrin	
Dieldrin	hexachlorobenzene	
PCB (total)	pentachlorophenole	
PCDD/PCDF*	DDT (total)	
Phthalates	nitrate	
1,1-Dichloroethane	PCDD/PCDF*	
Dichloromethane		
Tetrachloroethylene		
Toluene		
Trichloromethane		
Xylene		
Vinylchloride		
Cyanides		

* PCDD = polychlorinated dibenzodioxins; PCDF = polychlorinated dibenzofurans.

Transformation Function. The behavior of chemical compounds in the pedosphere depends on the properties of the substance, the location, and in some cases is affected by anthropogenic activities. The transformation function provides for self-cleaning of the soil. Degradation and transformation of substance in soil can occur abiotically (e.g., reactions on mineral surfaces, photochemical reactions) or biotically. One of the most important forces of the self-cleaning capacity of soils are the microorganisms, which provide the enzymes necessary for biotic transformation processes. Microorganisms use substances that enter the soil as nutrients or energy sources, whose exploitability increases with decreasing persistence of the compounds. After a substance has entered the soil, the microflora require a latent period (lag phase) for the development of the specific species spectrum required for its metabolization. This time period depends on the xenobiotic properties of the substance. If the substance is introduced more often, the lag phase becomes shorter and the persistence decreases. In addition, all processes that stimulate microbial activity in the soil are suitable for increasing its metabolic potency. These include application of organic and inorganic nutrients and cultivation of the soil.

The degradation and transformation of postmortal plant and animal material to humic substance is also attributed to the transformation function. Humic substances are naturally occurring biogenic substances with characteristic yellow to black color, high particle density, and persistence. Humic substances have been found in materials that have undergone coalification, such as peat, liquites, and lignites. The humic substances can be classified according to their solubility. Humic substance fractions and their solubility, as determined by various authors, are as follows:

1. Classification according to SCHEFFER, ULRICH [12] :

Crenic acids	acids, alkali, water, alcohol
Apocrenic acids	acids, alkali, water, alcohol
Hymatomelanic acids	alkali, alcohol, acid (partially)

Gray humic acids	alkali
Brown humic acids	alkali

2. Classification according to FLAIG [13] :

Fulvic acids	acids, alkali
Hymatomelanic acids	alkali, alcohol
Humic acids	alkali
Humines	insoluble

3. Classification according to SWIFT [14] :

Fulvic Acids	acids and alkali
Humic acids	alkali
Humines	insoluble

In contrast, ZIECHMANN [15] classifies the humic acid fractions according to their reactivity and their ability to undergo further humification with uptake of oxygen:

Starting materials (nonhumic)	organic substances undergoing humification lightly colored to colorless
Humic acid	reactive fraction, gradually transforming into humic acid light brown to brown
Humic acids	fraction containing carboxylic acid groups, phenolic OH group base-soluble, precipitate in acid intense brown color
Humines	inactive fraction insoluble intense brown color

The principle route for the formation of humic substances is shown in Figure 6. In reality, the formation mechanism presented here may be far more complex. This is due firstly to the wide variety of substance groups that can be involved in the humification process, and secondly to the activities of the micro- and macrofauna, which are responsible not only for producing these substances but also for mineralization, which competes with humification. Substances produced by microorganisms during the degradation of organic matter, as well as partially degraded, but considerably modified, substances, are involved in the humification process. In contrast to this incorporation of degraded biopolymers in a preexisting humic substance system, it is now being discussed whether such degraded biopolymers should themselves be designated as humic acids.

Figure 6. Formation of humic substances [15]

HAP = humic acid precursors; HA = humic acid; Hm = humines

Degradation and transformation of nutrients (⟶ Ecology and Ecotoxicology – Nutrient Cycles and Productivity of Terrestrial and Aquatic Ecosystems) is another important transformation process in soils and also involves microorganisms. Of major importance for the cleansing of the soil is the degradation and transformation of pollutants. The microbial degradation of various groups of chemical substances with ultimate mineralization is a first-order reaction. This has been shown, among others, for some chlorinated hydrocarbons, triazine derivatives, organic phosphorus compounds, carbamates, thiocarbamates, thiurams, urea derivatives. Under comparable conditions, the rate of microbial degradation is given approximately by

$$c = e^{-kt}$$

where *c* is the concentration of the compounds, *t* is time, and *k* is the mineralization rate.

Plotting log *c* against *t* gives a straight line whose gradient is the mineralization rate *k*, which allows predictions of the persistence of the compound under given soil and other environmental conditions to be made. Strongly sorbent, dry, low-nutrient soils with relatively low microbial activity exhibit small gradients, whereas soils in which the nutrient supply, air and water balance and hence the microbial activity are optimal give a steep gradient.

Continuation of the microbial transformation process in the soil is essential for plant growth and hence for the existence of animals and humans. In this context, the question of whether the increasing production of chemicals in the last few decades, which has led to intentional or accidental introduction of intermediate, final, and waste products into the soil, can seriously affect the biological functions of the soil, is becoming of increasing importance.

To investigate this, it is necessary to study the behavior of numerous chemical substances in the soil and to elucidate their effects on soil organisms. Depending on their properties, chemical pollutants can persist in the soil or be degraded. Thus a distinction can be made between persistent and nonpersistent pollutants. Both groups can affect soil microorganisms and their activity in various ways (Fig. 7). Persistent pollutants accumulate in the soil, acting on the soil microflora either alone or as complexes with clay minerals, humic acids, and oxides. Nonpersistent pollutants undergo a decrease in concentration due to chemical or physicochemical, but mainly biological/biochemical processes. If such processes result in mineralization of the pollutant, a detrimental effect on soil microorganisms is unlikely. However, since complete mineralization is generally the exception, it can be expected in many cases that intermediate degradation products or the final products of transformation reactions could adversely affect the soil microorganisms and their activity. Such effects can change the condition of the total biomass as well as the species composition of the microorganism population and its mineralization and synthesis activity.

Figure 7. Behavior of pollutants in the soil and their interaction with microorganisms [16]

3.1.2.2. *Habitat Function for Organisms in and upon the Soil*

The organic material of the soil consists of 85 % postmortal substance, 10 % plant roots, 4 % bacteria and fungi, and 1 % soil fauna.

Figure 8 shows the soil microorganisms and their important ecological functions.

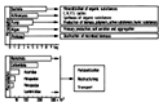

Figure 8. Soil microorganisms and their ecological functions

Bacteria and fungi account for 90 % of degradation, while soil animals are responsible for only ca. 10 %. However, soil animals are highly effective in accelerating microbial degradation and can increase the availability of nutrients by up to 50 %. Field studies have shown that decreasing their numbers decreases the degradation performance of the soil by up to 40 %. In general, single-cell organisms, worms, and anthropods each make up one-third of the soil biomass. The number of organisms per square meter in the upper 30 cm of the soil is as follows [17]:

Bacteria

	60×10^{12}
Fungi	10^9
Algae	10^6
Single-cell organisms	500×10^6
Threadworms	10×10^6
Mites	150 000
Collembola	100 000
White ringworms	25 000
Earthworms	200
Snails	50
Spiders	50
Isopods	50
Millipedes	150
Centipedes	50
Beetles	100
Fly larvae	200
Vertebrates	0.001

Respiration of an organism is a measure of its metabolic activity and, therefore, its material turnover. Protozoa account for 70 % of respiration — considerably higher than their biomass fraction of about one-third. It is not general knowledge that these soil-inhabiting microfauna produce approximately as much biomass as earthworms. In the food chain, protozoa represent an ecologically important link between bacteria and small metazoans (e.g., rotifera and nematoda) and thus exhibit pronounced catalytic activity towards bacterial processes and remineralization processes in the soil.

Taxonomically, soil protozoans can be divided into three main groups: small flagellates of size 3 – 10 µm; 2 µm flat naked amoebas with diameter 10 – 300 µm and testate amoebas of size 20 – 200 µm; and ciliates (20 – 300 µm).

Their occurrence is mainly limited to the upper 10 cm of the soil. Due to their small size or flat, flexible form, flagellates and naked amoebas can enter even the smallest soil pores (≥ 2 µm). In terms of ecological function, biomass (Table 3), and distribution, the three groups of protozoans (flagellates, amoebas, and ciliates) can be regarded as being of equal importance.

Table 3. Distribution of "naked" protozoa in different soils and in the rhizosphere (numbers given: individuals/g dry matter [18] and, below, relative biomass fraction (%) calculated according to [19])

	Arable soil		Forest soil, subtropical	
	Uncultivated soil	**Rhizosphere**	**Soil (total)**	**Rhizosphere**
Flagellates	570 – 14 000	1600 – 65 000	22 000 – 3 127 000	4000 – 8 231 000
	4 – 16	8 – 15	19 – 58	12 – 16
Naked amoebas	10 500 – 50 000	18 000 – 181 000	14 800 – 823 000	1100 – 30 750 000
	79 – 56	43 – 92	5 – 39	3 – 61
Ciliates	30 – 280	0 – 1750	10 – 132 600	280 – 132 600
	17 – 29	0 – 41	3 – 76	23 – 84

The broad-spectrum activity of pollutants in soils, in particular their effects on entire colonies of organisms, has remained largely unregarded, partly due to the lack of suitable test systems. Classical laboratory tests (e.g., LC_{50} values) are unsuitable for predicting the true toxicity in open country. To make more accurate predictions for both existing and new substances, current laboratory tests, which are restricted to a few selected animal species, generally on standardized substrates, must be extended to further soil animal species, and additional tests on organisms, organs, single cells, and cell compartments are also required. Novel test methods based on enzymes as biomarkers avoid the need for animal tests and can also be used at sublethal concentrations. Although such tests provide important ecological data on the lethal and sublethal action of pollutants, they are not suitable for predicting effects on species colonies in ecosystems. This information can only be obtained by means of field experiments designed to take into account temporal, spatial, and genetic population variations in the ecosystem and the varying bio-availability of pollutants from species to species.

3.1.3. *Productive Function*

The productive function is the ability of the soil to act as substrate for cultivated plants (food and feed plants, renewable resources). Effects on the productive function are generally assessed in economic terms, rather than ecologically, since cultivation in agriculture, forestry, and horticulture is intended to produce a profit. Nevertheless, a number of ecologic soil functions are affected by cultivation.

3.1.3.1. *Agriculture*

Plants obtain their nutrient elements, but also pollutants, mainly from the soil. The supply of nutrients is generally typical for the type of soil and region. However, in modern agriculture external measures are applied to increase the harvest: inorganic and organic crop-protection agents, organic and mineral fertilizers, and utilization of organic and mineral wastes, all of which can lead to pollution of the soil and can damage soil functions.

Crop-Protection Agents. The entry of CPA into the soil as a result of agriculture is a potential influence on soil function. The CPAs are designed to alter the balance of nature in favor of the cultivated plants.

The behavior of CPA in soil depends on their properties, the way in which they are used, and on local conditions. The soil determines important properties of the location, especially the hydrological properties.

The amount of CPA applied, the frequency of application, and the time at which it is applied are relevant to movement of the agents in the soil and their potential entry into water systems. The manner in which a CPA is used affects its degradation, which is faster at the higher temperatures of spring and summer and leads to a more rapid decrease in the amount of CPA present. In autumn and winter degradation is generally slower. However, general statements valid for all CPAs cannot be made.

Fertilization and Waste Utilization. Fertilization and the utilization of waste materials can lead to the introduction of both organic and inorganic pollutants into the soil. In contrast to the application of crop-protection agents, this is an unintentional contamination. The aim of fertilization, which can include utilization of waste, is to improve conditions for the growth of cultivated plants, either by direct action on the plants themselves or by effecting the soil. The macronutrients nitrogen, phosphorus, and potassium are the most important for the nutrition of cultivated plants. Thus in the sustainable development of agriculture, both nutrient and pollutant aspects must be taken into account.

Nutrient Aspect. The high productivity in countries of Central Europe, compared to average world harvests, is only attainable by intensive application of fertilizers. The concept of location-specific soil fertility formulated by KLAPP [20], who defined this as the soil's own level of natural, persistent performance which distinguishes it from other soils, is not applied in conventional agriculture. According to KLAPP the most important aspect is the sustainability: the long-term ability of the soil to provide all requirements for plants to flourish, that is, to provide constant or only slowly decreasing harvests without fertilization. In contrast, in modern agriculture, fertilization is regarded as guaranteeing the nutrient requirement of the plants, taking into account the nutrients already available in the soil. The natural productivity of the soil is thus completely disregarded. In modern agriculture, the target of fertilization to an extent appropriate to the requirements of the plants is not met. If a balance is determined for nitrogen, the result is a considerable fertilizer excess. The starting point for sustainable use of the soil should be a nutrient supply that is appropriate to the needs of the plants and which avoids an excess in the soil and release into other environmental compartments. Furthermore, in intensive agriculture, the proportion of nutrients introduced into organic materials should be restricted to 60 % of the total nitrogen requirement. The reference points for fertilization planning are the exploitable components in the following plant population, which increase with increasing frequency of utilization and the amount of organic material utilized (see Transformation Function.). Residual material remaining on the field after harvesting must also be included in the balance, and manure should be used in preference to external waste materials.

Pollutant Aspect. Mineral fertilizers, organic fertilizers, and wastes all introduce inorganic and organic pollutants into the soil and can cause adverse effects there and in other environmental compartments. Mineral fertilizers can contain various amounts of pollutants, depending on the origin of their raw materials and the production process used.

The main pollutants entering soils via fertilization are Cd and Cr, and to a lesser extent Pb, Ni, and As. The Mn, Cu, and Zn contents of fertilizers are, with the exception of the high Cu and Zn content of liquid pig manure, low, lying in the nutrient element/trace nutrient element range. Most of the heavy metal load comes from phosphate rock and P fertilizers derived therefrom (super and triple phosphate, partially digested and soft phosphate rock), Thomas phosphates, and phosphate-containing multinutrient fertilizers. P fertilizers derived from phosphate rock can exhibit high Cd contents, depending on the origin of the phosphate rock. Thomas phosphates have high Cr contents, resulting from the production process. The specific quantity of heavy metals introduced is strongly influenced by the various land-use and fertilization systems. According to BOYSEN [21], the Cd entry is up to 4.1 g/ha in

conventional farming, up to 4.8 g/ha for meadowland, but only 0.9 g/ha for ecological agriculture. Data on organic pollutants in fertilizers are not yet available in sufficient quantity. In terms of quantity, the most important organic fertilizers are liquid manure, sewage sludge, and compost.

In Europe only sewage sludge is subject to EC regulations which set values for heavy metals in sewage sludge and soil. In addition, the methods of application are specified, allowing theoretical permissible pollutant loads to be calculated. For compost, numerous quality marks exist, whereby the pollutant concentration is specified. They do not, however, set limits on quantity. If an average application of 10 t ha^{-1} a^{-1} is assumed, a cadmium load of 10 – 40 g/ha can be calculated [22].

However, depending on the region and the spectrum of discharge or compost starting materials, these theoretical pollutant loads are not always fully exploited.

The inputs can be compared to theoretical outputs, as exemplified below for the Cd and Hg balance. For the average German wheat harvest of 5 t ha^{-1} a^{-1} and a tolerable Cd content of 0.2 mg Cd/kg dry matter wheat (marketability criterion), a potential output of ca. 1 g ha^{-1} a^{-1} results. The Hg output is 0.3 g ha^{-1} a^{-1} (0.06 mg Hg/kg dry matter).

If the values specified e.g., by the Drinking Water Regulations (Trinkwasserverordnung), are taken as the tolerable pollution of the leachate, then for the average leachate quantity in Germany of 200 mm/a (2×10^6 L ha^{-1} a^{-1}), a potential Cd output of 10 g ha^{-1} a^{-1} and an Hg output of 2 g ha^{-1} a^{-1} result. Thus the potential outputs are already met by the potential inputs from the utilization of waste materials alone.

3.1.3.2. *Forestry*

After agriculture, forestry is the second largest use of land. In forestry, atmospheric nutrients and pollutants are of major importance.

The problem of pollution in forestry is exemplified by a comparison between Cd inputs and outputs (in g ha^{-1} a^{-1}) for a forest soil (spruce trees, ca. 120 years old) in a mountainous region of northwest Germany [23]:

Total input	6
Precipitation	2.7
Interception	3.3
Total output	15.3
Leachate	13.5
Crop (trunk wood without bark)	1.8

Thus the mean annual output is 2.5 times higher than the atmospheric input by precipitation and dry deposition. It is questionable whether the limit values of the German drinking water regulations are suitable quality standards for assessing outputs in the leachate. For a limit value of 0.005 mg Cd/L water, a leachate flow of 424 mm/a, and given the requirement that the leachate water quality must comply with the drinking water regulations, a tolerable Cd output of 21.2 g ha^{-1} a^{-1} can be calculated. This is considerably higher than the actual leachate output of 13.5 g ha^{-1} a^{-1}. In analogy to the specified quality target that the output should not exceed the input, according to this model calculation, an additional Cd input of 9.3 g ha^{-1} a^{-1} would be possible without exceeding limit values specified for drinking water.

The fact that for the material balance for cadmium in this example, the output is more than twice as high as the input shows that this soil has already lost its function as a filter medium for the formation of new groundwater. The mobilization of cadmium accumulated in the forest soil as a result of acidification demonstrates the problems that arise when the filter capacity of the soil is exhausted. The filtration capability of the soil can only be regarded as undisturbed if the output (at the background level) is lower than the current input. If changes in soil properties lead to release of accumulated materials, then pollution of the groundwater or of the habitat for organisms cannot be excluded.

3.1.4. ***Other Soil Functions***

3.1.4.1. *Utilization of Wastes*

The exploitation of waste is possible in cases where it is both safe and useful. Benefits to the soil occur if at least one soil function is improved without permanent detriment to other natural soil functions or other environmental compartments. In the use of organic wastes in agriculture, the benefits can be assessed in terms of the fertilizer effect. Utilization of waste should be safe with regard to nutrients and pollutants. An important part of the recycling and waste business in Europe is the material utilization of waste with an increasing pressure to use wastes in soil-related applications.

In Germany for example, the distinction between mineral and organic wastes is defined in the regulation TA Siedlungsabfall, which excludes wastes containing >5 % TOC (organic wastes) from dumps to minimize the pollution of the leachate associated with the degradation of organic materials.

Soils should be included under mineral wastes because the organic matter (up to 30 %) they contain differs strongly from organic wastes in type and constitution. They therefore represent a special case. Organic wastes are used in agriculture, gardening, and landscaping, whereby a distinction can be made between repeated application, mainly in agriculture, and one-time use, mainly in recultivation and landscaping. Mineral wastes are mainly used in landscaping and in recultivation and building projects; smaller amounts are utilized in agriculture. An exception is soil material, which is used in considerable amounts for improving soil functions.

An important criterion for assessing the harmlessness of such measures are the background concentrations of pollutants (see Section Regulating Functions), which should not be exceeded by the input pollutant load of organic wastes. In the utilization of organic wastes, the pollutant load is the most important factor, since they remain almost entirely on the surface after mineralization of the organic component, contributing to an increase in the pollutant potential.

In the case of mineral wastes, there is generally no loss of mass after use in the soil, so that a different pollution assessment concept must be used.

The pollutant concentrations in the waste should not exceed the background concentrations in the location of use. If the values are compatible, application is possible, since a deterioration in soil quality with respect to pollutant concentration will not occur. For the application of mineral wastes that exceed the background values, eluate investigations must be undertaken, which indicate whether risks to groundwater quality are present.

The use of soil material in the topsoil region is subject to strict requirements for the soil properties. This is due to the requirements of plants for water retention capacity, aeration, rootability, warmability, and nutrient storage capacity. For this reason, the soil should have similar composition and properties to that at the site of use. Poorer soil materials should not be applied to soils of higher quality.

3.1.4.2. *Raw Material Deposits*

Soils that act as raw material deposits include sands for building and clays for ceramic production, but the largest areas are taken up by lignite mining. Other areas that generally require recultivation are the slagheaps from salt, ore, and coal mining dumps and military training areas. All of these lack a plant-supporting upper layer and have lost the natural soil functions.

Table 4 compares the three main lignite mining areas in Germany.

Table 4. Breakdown of areas under mining law [24]

	Lignite mining districts, ha		
	Rhenish	Central German	Lausatia
Total area used	24 500	46 900	48 999
Area subjected to recultivation	15 600	22 000	24 655
Operated area	8 900	24 900	24 344

Especially in the former German Democratic Republic, there are a large number of areas still to be recultivated. In the Land Brandenburg alone, as of 1994, the total area affected by open-cast mining of lignite was 50 000 ha. About half of this has so far been rendered reusable. The recultivation of devastated and mined areas is intended to restore or improve the natural soil function. This includes building up a plant-bearing soil layer in sufficient quality and depth. Other targets are:

1. Increasing the content of organic substances in the topsoil
2. Introduction of plant nutrients
3. Regulation of the pH value
4. Stimulation of soil life

At the same time, deterioration of soil function due to pollution input, addition of inappropriate nutrients, and compaction must be avoided.

3.2. Land Utilization and Soil Destruction (Statistics)

3.2.1. *Introduction*

The consumption and destruction of soils, i.e., the destruction of the surface soil or of the topsoil [25], is one of the biggest ecological problems of the world. The soil is the basis for food production and will remain so in future [26]. While in the highly developed and industrialized countries problems of soil acidification and of contaminated sites (see Section Effects of Abandoned Waste Disposal Sites) are discussed [27], [28], soil erosion (Section Erosion), soil salination, and desertification in other parts of the earth progress unabatedly.

Many thousands of hectares of cultivated land are lost each year through erosion, salination, and desertification. Thus, in 1975 the soil of half of the Indian subcontinent, 150×10^6 ha, was seriously affected. In the United States in 1975, erosion had made nearly 40×10^6 ha unusable for agricultural cultivation [26]. In Argentina in 1980, 22×10^6 ha, nearly 16 % of the cultivated land, had been destroyed by wind erosion. In the 1970s – 1990s, according to estimates, the farmers lost 480×10^9 t of topsoil, which equals the entire topsoil of India [29]. These few examples should prove the seriousness of the problem of soil destruction.

3.2.2. *Land Utilization*

Table 5 shows a breakdown by continents with regard to the land area and its use [30].

Table 5. Land area and use (in 10^6 ha; source [30])

	Land area	Arable land incl. permanent crops	Permanent pasture	Forest
Africa	2963	211	900	650
North and Central America	2144	272	349	549
South America	1753	127	516	886
Asia	3098	573	1110	548
Europe	2260	304	183	1039
Oceania	849	54	413	198
World	**13 067**	**1541**	**3472**	**3870**

In the following, the land utilization and land use are considered in more detail for the Federal Republic of Germany, as an example of an industrialized country. Figure 9 shows the land use distribution in the Federal Republic of Germany in 1994. Settlement and traffic areas, which are composed of residential and office buildings, industrial facilities, administrative and educational establishments, as well as roads, sports fields and airfields, cemeteries and parks, in 1994 occupied about 12 % [31]. For comparison, in 1950, the proportion of residential, industrial, and traffic land was 7.2 %. Figure 10 [32] shows the daily change in the years 1981 to 1985.

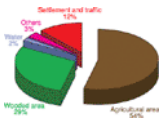

Figure 9. Land use in the Federal Republic of Germany (source: [31])

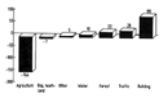

Figure 10. Daily change in land use from 1981 to 1985 in hectares (source: [32])

The settlement area in Germany increases by ca. 105 ha daily [33].

3.2.3. *Soil Destruction*

Proceeding on another scale than the sealing of land mentioned above are the destruction of soil by erosion, salination, desertification, and destruction of forests.

In the frequently very labile ecosystems in the non-industrialized countries, where 80 % of the world population growth occurs, industrial methods of agriculture lead much more quickly to the destruction of the valuable topsoil. In the subtropics and tropics, this leads rapidly, because of erosion, to the expansion of dry areas and desertification after a short phase of meager yields [34].

Erosion. Erosion by wind or water plays the most important part in soil destruction. Worldwide, $25\,000\times10^6$ t of soil are washed away by water erosion every year [38]. The extent of erosion is shown in Tables 7 and 6.

Table 6. Estimates of global rates of soil erosion and likely future trends (source [39])

Region	Soil erosion by water				Soil erosion by wind	
	Area, 10^6 ha	Denudation rate[2], mm/a	Dissolved load, 10^6 t/a	Future trends[*]	Area, 10^6 ha	Future trends[*]
Africa	227	0.023	201	+	186	+
Asia	441	0.153	1592	+	222	+
South America	123	0.067	603	+	42	–
Central America	46	0.055[**]	758[**]	+	5	–
North America	60	0.055[**]	758[**]	+/–	35	–
Europe	114	0.032	425	+/–	42	+/–
Oceanica	83	0.39	293	+	16	+
World	**1094**	**0.079**	**3872**	**+**	**548**	**+**

[*] + = Increased risks; – = decreased risks;
[**] total value for Central America and North America

Chinese soil scientists estimate that in their country since the 1960s almost 10 % of the arable land has been lost through erosion and desertification [40]. In the United States in the last 200 years about one-third of the surface soil has been destroyed by erosion [41].

At the 7th Conference of the International Soil Conservation Organization in 1992 in Sydney, the following figures were reported by the project for the worldwide estimation of soil degradation: Between the 57th degree of latitude South and the 72nd degree of latitude North, 164×10^6 ha, i.e., 15 % of the earth's surface, show more or less strong signs of degradation. In addition, 295×10^6 ha, corresponding approximately to the size of India, have to be classed as highly degraded. About 124×10^6 ha of these are in African and 106×10^6 ha in Asian countries [35].

Chemical Degradation. The agricultural lands rendered usable by artificial irrigation are threatened, after some time of use, by salination. Table 7 shows the significant proportion of soil areas degraded by salination [35-37].

Two further aspects of chemical degradation, i.e., contamination and acidification of soils, are closely related to anthropogenic industrial pollution. While their total extent is relatively small compared to the areas influenced by the other chemical degradation processes, there are numerous “hot spots” with severe problems in the heavily industrialized regions.

With respect to area, loss of nutrients is the most important chemical degradation process. Nutrient losses result from current or historical overexploitation of soils by non-sustainable cultivation methods.

Table 7. Extent of human-induced soil degradation for the main types of soil degradation (10^6 ha, source [37])

Types of degradation	Light	Moderate	Strong/extreme	Total

Water erosion	343	527	224	1094
Wind erosion	269	254	26	549
Chemical degradation	93	103	43	239
Loss of nutrients	52	63	20	135
Salinization	35	20	21	76
Contamination	4	17	1	22
Acidification	2	3	1	6
Physical degradation	44	27	12	83
Total	**749**	**911**	**305**	**1965**

Desertification. About 25 % of the earth's surface is subject to the spread of deserts and every sixth human being is affected [42]. Every year, about 6×10^6 ha of pastureland become desert [43]. The deserts have in this way expanded by 120×10^6 ha from 1970 to 1990. This corresponds to a greater land area than is cultivated in China today [47].

In the late 1970s the United Nations estimated the risk that non-desert land will become desert as follows:

50 % desertification of non-desert land in Africa
32 % desertification in Asia
19 % desertification in South America [48].

The Sahara alone expanded in the 50 years before 1977 by 650 000 km^2 [42].

Forests. The felling of trees in forests leads indirectly to the destruction of soil. Through the felling of trees the soil is often exposed to erosion by wind or water.

More than 5×10^6 ha of the tropical rain forest are wiped out each year to bring land into cultivation. Since 1970, farmers and ranchers in South America have converted more than 20×10^6 ha of wet tropical forests into cattle pasture [43].

The global current annual change of forest cover is depicted in Figure 11 [44]. According to this estimation, forest land is dramatically shrinking in South America (particularly Brazil), Central Africa, Southeast Asia and the Pacific Area, and in Russia. Rates of losses in the affected regions range from 100 000 to several million hectares per year. There is no indication that deforestation trends have been halted or reversed in those affected regions (see time series on Amazon deforestation in [45]).

Figure 11. Average annual change of forest per country (source: [44])

3.2.4. *Forecast*

The amount of soil available on the globe cannot be increased, whereas the human population is still increasing at a rate of 2 % or more annually in developing countries. Pressure on the resource soil is highly related to population growth. A map of the number of inhabitants per hectare agricultural land is shown in Figure 12 [30], [46]. It illustrates that pressure on the soil will be highest, where human population is already high and growing fast, while the area of cultivated land cannot be increased much further. A dramatic situation of almost continental extent exists in entire South East Asia. Some countries in Africa exhibit the same constellation of factors. In South America the situation is still less dramatic, because along with an intense population growth, the cultivation of land is increasing at the expense of a loss of forests.

Figure 12. Capita per hectare agricultural land (source: [30], [46])

Currently (2001) there are no indications of a major turn in the trend of increasing soil consumption and destruction. However, this trend cannot be continued arbitrarily long into the future, because the resource consumed is finite, and methods of enhancing yield from a given area of soil are limited. Critical ranges of soil consumption are attained in many parts of the world (Fig. 11, 11). The only way out of this dilemma seems to be the stabilization of human populations plus improved, sustainable methods of soil utilization.

[Top of Page]

4. Physical Damage to Soils

Karl Auerswald

Soils play a central role in ecosystems, e.g., for plant growth or the storage of rainwater, since soils exchange matter and energy with the atmosphere, the biosphere, and the hydrosphere. This exchange depends on their physical structure, e.g., on their depth and the structure of their pore system. Physical damage therefore impairs all functions of soils in the ecosystem.

4.1. Compaction [49-51]

Process. Natural compaction is caused by the soil's own weight, which leads to higher bulk densities at greater depths. Menisci of the water in the soil pores excert forces on the pore wall. These forces increase strongly with the drying-out of the soil and therefore stabilize aggregates. The force excerted by the water menisci is 1.5 kPa when the soil has been dried to the point where plants begin to wilt irreversibly. Further natural compaction is caused by the burden of the vegetation cover, especially under woodland; and by the intercalation of illuvial material (clay, iron oxides, carbonates) in the course of pedogenesis (luvizol, podzol, gleyzol). With increasing compaction of the soil, its shear strength, its bearing capacity and the capillary rise height increase. Soils loosened by cultivation are therefore reconsolidated, e.g., by rolling, after sowing.

The compactibility of soils decreases with increasing content of sand, stones, and organic matter; grain roughness; increasing aggregation (from blocky via prismatic and angular to subangular aggregates); rooting; and preconsolidation.

Damage. Only above a certain limit does compaction harm plant growth. In what follows, this compaction is called *excessive compaction*. When a compaction becomes excessive depends, apart from the soil properties, greatly on the weather and thus changes in the course of the year and from year to year [52]. In dry seasons a fairly high soil density is advantageous, because of the faster capillary water supply. In wet seasons, on the other hand, a low density is advantageous, because the soil then contains more coarse pores draining the excess water. Since the weather changes, it is also impossible to quote an optimum soil density. If the bulk density deviates from those shown in Figure 13, however, on average reduced yields have to be reckoned with under Central European climatic conditions.

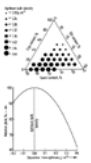

Figure 13. (A) Optimum bulk density for soils of different texture; (B) change in yield of grain, potatoes, and sugar beet on deviation from the optimum bulk density (according to [71], modified)

Excessive compaction harms plants in all crop stages, but especially in the early stages. Compacted soils warm up more slowly (increased thermal conductivity, increased evaporation) to the temperatures necessary for germination and N mineralization, and the young plants develop only a weak root system. In the later stages, root development in compacted soils is more shallow and less dense. This lowers the nutrient uptake and impedes root respiration. The susceptibility to root diseases increases and in leguminous plants fewer root nodules are formed. Even with sufficient water and nutrient supply, plants react to compaction with a modified hormone balance, which reduces their growth [53].

The damage due to compaction is difficult to quantify, since large differences occur over the years and even within one field. Estimates with grain crops assume mean yield reductions of 5 –15 %. Other crops, especially root crops, react more strongly than grains. In addition to the losses of yield, the harvest is more difficult (clods during the potato harvest). The products have a lower quality (more clinging dirt in sugar beet, deformations in carrots). The tractive force requirement and machine wear during soil cultivation rise. (If compaction

causes only 5 % more slip, the tire wear already increases by 20 %.) The soil compaction can persist many years, especially if the subsoil has been compacted. Up to 50 years persistence has been reported [54].

Excessive compaction does not occur only in mechanized agriculture but in all soil uses, also during forestry [54] and as the result of cattle tracks on pastures [55]. In modern agriculture the loads and the number of machine passes have risen, increasing the compaction risk. The loads rose because the yields increased through the intensification of agriculture and an increased stocking of pastures with cattle became possible. In addition, to reduce the working costs the operating widths were increased and larger tractors, which can move the greater loads and are more convenient, were used. Particularly high loads on the soil occur in lime and slurry spreading and in the harvesting of crops of high water content (sugar beet, potatoes, green maize). The number of passes increased, because a well-devised system of agriculture necessitates, among other things, intensive tillage, fertilization in many small doses and the regular application of pesticides and growth regulators. In particular the fungicides are frequently sprayed after rainstorms because of the high risk of infection at that time. But at that time also the bearing capacity of the soil is low. In grain cultivation, 10 – 15 passes are customary today and in the case of more intensive crops (especially, e.g., hops) considerably more. Five passes, which are necessary for the sowing of barley, already cover 91 % of the field with tire tracks [50]. Compaction is a problem particularly in perennial crops, such as hops and wine, and in fruit plantations, since the same tire tracks are always used and very many passes are necessary.

Loosening. Freezing and thawing, swelling and shrinking, soil animals (the edaphon), and roots loosen the soil naturally. Roots displace the soil mainly to the side (radial pressure: 0.5 –1 MPa) and because of that compact it in addition, but create the vertical pores necessary for gas exchange and percolation. Earthworms also mainly displace and compact the soil. Because of their small axial force (0.1 MPa) they require zones of weakness (cracks), into which they penetrate in order to widen them. If the soil strength is higher than the radial forces that the earthworms can exert (0.2 MPa) [56], this strategy fails and the earthworms then loosen by ingesting compacted soil and transporting it to the soil's surface. Since this is exhausting, they loosen only slowly and only in cases when displacement is no longer possible but at the same time their other living conditions (e.g., availability of food) are advantageous.

Frequency and intensity of frost (up to 200 MPa at –20 °C) and swelling and shrinking (up to 2 MPa in Central Europe) decrease with increasing soil depth. In Central Europe the soil at a depth > 0.4 m freezes on average less than 10 times per year. Since freezing and thawing or shrinkage must occur many times before a compaction is loosened, the compaction in the subsoil persists for a long time. It is therefore especially problematic there.

In the surface soil, which is loosened considerably faster by natural processes, humans also intervene with tillage. In doing so they create above all coarse pores between the clods, but cannot loosen the clods themselves. Natural processes are necessary for that.

Protection Measures. Excessive compaction can be avoided by larger supporting contact areas of agricultural machines (dual wheels, cage wheels, wide-section and ultra-low pressure tires, more axles, tracked vehicles). The contact area can be raised further by reducing the tire inflation pressure, the use of radial instead of cross-ply tires and a small tire profile. Since as a result of larger contact surfaces the compacted fraction of the surface increases and pressure decays slower with depth , an improvement is achieved only if the contact pressure is decreased by more than 50 % [50]. The critical contact pressure is considered to be 60 – 80 kPa.

Above all, heavy loads should be avoided, since with heavy loads the compaction depends more on the load than on the contact pressure and penetrates deep into the soil. The critical wheel load is considered to be 30 kN [50]. Faster passing compacts less deeply (e.g., at 1 km/h, 28 cm deep; at 12 km/h, 16 cm deep). However, the compaction and structure deterioration at the soil surface rises as a result, since the shearing of the soil increases.

Because the optimum density to obtain high yields depends on the weather, a well settled soil with rapid capillary water supply is advantageous, if large vertical pores, e.g., earthworm channels, regulate excess water and gas exchange and a mulch cover simultaneously reduces evaporation losses. These conditions are ideally present in the case of the no-till cultivation system. In the no-till system tillage is replaced by chemical weeding with herbicides. The crop residues accumulate at the settled soil surface where they supply food for earthworms over a long period. Under this system the soils are also more stable to compaction because loads are well supported in the settled soil and the compaction of vertical coarse pores is more difficult than that of horizontal ones. Despite these advantages, the no-till system has hitherto been successful only in isolated cases because, among other reasons, the lack of working-in of crop residues causes problems with plant diseases which survive in the crop residues.

4.2. Erosion [57], [58]

Soil erosion is the loss of soil through the action of wind (wind erosion) or water (water erosion). The water can originate from rain, dew, or snowmelts. Transportable soil particles are initially detached from the assemblage and then transported away by water or wind. Additionally, special types of erosion are caused by tillage (tillage erosion) and creeping snow.

Wind Erosion. Small aggregates and in particular loose sand particles are caught and dragged along by the wind and roll or saltate along the soil surface. On impact these particles loosen further particles that initially are still bound and accelerate them. As a result these particles also can be caught and transported by the wind. The amount detached and transported increases in the wind direction. The detached fines (clay, organic material) are transported very far as aerosols, while the sand, in rolling and jumping motion, is usually soon redeposited. Consequently a loss of fines, which are important for the soil's productivity, occurs while sand accumulates.

Particularly susceptible to erosion are fine sandy soils, because they are only weakly aggregated, and peaty soils which, because of their low specific gravity, can easily be blown away when they dry out at the surface. The high wind speeds necessary occur particularly in spring and autumn storms. In these seasons the fields are often inadequately covered with vegetation. Large fields without fairly high structures (hedges, ridges) at the edges favor high wind speeds at the soil surface.

Water Erosion. Transportable particles are detached by striking raindrops, air entrapment by rapid wetting, dispersion, and shear forces of the runoff. (When water is imbibed from both sides of a capillary pore, it compresses the entrapped air. Air pressure may be strong enough to disrupt the aggregate.) The detached particles are transported by lateral splash and above all by the water running off. Sheet erosion occurs if the detachment by the runoff is small. With increasing detachment the runoff cuts its way in. If this takes place at many points simultaneously, *rill erosion* occurs (depth of the rills is usually <10 cm). If the runoff concentrates, the rills become fewer and deeper. It then comes to the point of *ephemeral gully erosion*. In this case the channels can still be regularly refilled by the tillage. Still deeper cutting-in (>30 cm) leads to *gully erosion*. *Tunnel erosion* occurs when a readily dispersible subsoil is washed out below a relatively stable topsoil, e.g., one interlaced with roots.

Water erosion is especially high for soils rich in silt and poor in organic matter, which have a low stability in water and easily form transportable particles. The erosivity of the rain increases approximately as the square of the rain intensity. For this reason summer thunderstorms with their high intensities (>10 mm/h) are particularly erosive. In early summer they impact — above all in the case of row crops such as maize, beet, potatoes, and soya — on a surface which still has only a sparse vegetation cover, and can cause large soil losses. With increasing slope and size of the surface drainage area (dependent on slope length and slope shape) and with decreasing roughness of the soil surface, the transport capacity of the runoff rises. In the absence of cover, snowmelts likewise can lead to erosion, since the frost destabilizes the soil.

Extent. Natural ("geological") erosion is at a maximum in semi-arid to arid regions because of the scattered vegetation cover there. Rare but in some cases violent rains or strong winds can therefore trigger off considerable erosion. High erosion losses are natural in the mountains also, owing to the steep slopes and frequently unfavorable vegetation conditions. But erosion is principally a problem of farming, because this at times leaves soils bare, and wind and water can then attack. Wind erosion occurs above all near coasts, where high wind speeds and sandy soils are common. Leaving aside a few areas, it is less important than water erosion. The amounts of soil lost through wind in the United States are assumed to be less than 1/10 of those lost through water [59]. During farm use in moderate latitudes, soil erosion by water is frequently of the order of 10 t ha^{-1} a^{-1}, but soil losses of up to 200 t ha^{-1} a^{-1} have also been observed over fairly long periods [60]. Probably the highest soil losses, of the order 1000 t ha^{-1} a^{-1} [61], occur in South-East Asia (see Section Soil Destruction) [62]. High soil losses are reported for other tropical regions also (Brazil, West and East Africa, and India) [58].

Damage. In wind erosion, saltating grains of sand, above all, damage the plants by exposing, covering, or by injuring the cuticle. In water erosion, rainfall flows unproductively away and cultivation is made difficult by uneven crop development, and by gullies and channels. In both forms of erosion, nutrients and pesticides are lacking at the eroded sites and overdosed at the deposition sites.

Aside from these short-term forms of damage, cumulative, irreversible, long-term damage occurs because the soil depth and the water- and nutrient-holding capacity decrease. This damage is particularly marked where unfavorable parent material, such as gravel, hard rock, or acid subsoil, is covered with a fertile but shallow material such as loess.

To the on-site damage to the eroded areas must be added the off-site damage at the points where the sediment and runoff accumulate. Examples are the soiling of roads and houses, the siltation of waters, and the input of nutrients, pesticides, and heavy metals into waters and other neighboring ecosystems.

In some cases erosion can also be advantageous. Thus the soil fertility rises if an impoverished or acidified topsoil is eroded and exposes more favorable material. In addition the input of soil into waters can reduce the concentration of dissolved nutrients by sorption if high nutrient concentrations are caused by point sources.

Erosion Protection [63]. A permanently covered soil surface offers the best protection against erosion. 30 – 50 % covering is usually sufficient, since a vegetation cover reduces erosion more than proportionate. The covering protects from drop impact, reduces wind speed and runoff velocity, raises the soil moisture at the surface, and promotes vertical pore formation by earthworms. The cover need not be present permanently, but only during the seasons of erosive rain or winds. In conventional farming the soil is least covered directly after sowing. In minimum or conservation tillage a sufficient cover can also be guaranteed after sowing.

Besides that, runoff and wind can be slowed down by a rough surface, which is obtained by cultivating transversely to the slope or to the main wind direction. A small-scale, varied land-use system in which areas with different degrees of erosion protection adjoin each other prevents that too much runoff or eroding saltation material is collected and that high wind velocities arise. This mutual protection of different crops is deliberately made use of in strip cropping, in which the individual crops are grown in strips of any length and a width of usually 10 – 50 m, running transversely to the slope or to the main wind direction. The runoff can be slowed down further by terraces, and the wind by wind breaks.

Models. Because of the great variability of erosion in time and space, many prediction models have been developed to estimate its extent and to plan protection measures. Far more and better supported models exist for water erosion than for wind erosion (Table 8). The oldest, best known, and most used model of water erosion is the Universal Soil Loss Equation (USLE) [64], on which many more recent models are based. Its latest version is the Revised Universal Soil Loss Equation (RUSLE) [72]. These models differ in the processes considered (wind, water erosion; sheet, gully erosion), in the time scale (single-event, long-term models), in the spatial scale (field, drainage area), and in the modeling technology (stochastic vs. deterministic; process oriented vs. lumped). Table 8 assembles the most common models with their main applications. Many have been extended beyond this.

Table 8. Main areas of application for five common erosion models (special adaptations were not considered) [a]

Areas of application	Model					
	USLE/RUSLE [b] [64], [72]	CREAMS [c] [73]	ANSWERS [d] [74]	EPIC [e] [75]	WEPP [f] [76]	EUROSEM [g] [77]
Time scale						
Long-term	•			•	•	
Annual	o	•		o	•	
Single event	o	•	•	o	•	•
Spatial scale						
Field	•	•		•	•	•
Drainage area	o		•		•	•
Applicability in the field	•				•	
Soil use						
Agriculture	•	•	•	•	•	
Rangeland	•				•	
Forest	•				•	
Construction sites	•				•	
Erosion control structures	•		•		•	
Types of erosion						
Gully erosion		•	•			
Ephemeral gullies				•	•	
Sheet and rill	•	•	•	•	•	•
Snowmelt	o	•		•	•	•
Irrigation					•	•
Wind erosion				•		
Sedimentation		•	•	•	•	•
Nutrient loss		•		•		
Pesticide loss		•				
Water movement (runoff, infiltration, percolation)		•	•	•	•	
Weather generator				•		
Soil temperature				•		
Plant growth				•	o	
Economic evaluation				•		

[a] • = applicable; o = restricted applicability.
[b] USLE = Universal Soil Loss Equation; RUSLE = Revised Universal Soil Loss Equation.
[c] CREAMS = Chemicals, Runoff, and Erosion from Agricultural Management Systems.
[d] ANSWERS = Areal Non-point Source Watershed Environment Response Simulation.
[e] EPIC = Erosion Productivity Impact Calculator.
[f] WEPP = Water Erosion Prediction Project.
[g] EUROSEM = European Soil Erosion Model

Tillage Erosion [65]. Tillage lifts up the soil normally to the soil surface. The soil then falls back vertically due to gravitation. On sloping land, the upward movement thus follows a different path than the downward movement (Fig. 14). This causes a net movement downslope. This movement occurs in any case where the soil is lifted by tillage, i.e., also in cases where the tillage direction is perpendicular to the slope or where tillage operations are carried out with changing directions. Additional movement results where tillage direction is downslope only. This is the case when traction forces do not allow upslope tillage on very steep slopes or in poorly developed agricultures where tillage is carried out with the help of animals or with hand hoes.

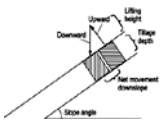

Figure 14. Movement of a representative soil block during tillage

The amount of soil translocation increases with tilling depth, slope angle, lifting height and numbers of tillage operations. Tillage translocation can thus be reduced by reducing tillage frequency and depth. On a straight slope, the downslope loss at a certain point is fully compensated by the gain from upslope. A net loss hence only results at the upper field edge where no input from upslope occurs. Correspondingly, a net gain occurs at the lower field edge which receives input from upslope without downslope losses. This characteristic distribution of loss and gain areas leads to the development of terraces after long-lasting land use in landscapes with small field layout like in Middle Europe. Loss and gain are also not balanced on slopes where slope steepness changes. When slope steepness increases, downslope soil transport is larger than the input from upslope resulting in a net loss. This causes the characteristic strong soil truncation of slope shoulders and crests which cannot be explained by water erosion because erosion should be larger downslope where increasing slope gradient and amount of runoff both cause an increase in transport capacity. Where slope gradient decreases downslope like at toeslope positions, soil input from the steeper upslope area is larger than the soil movement downslope. This surplus soil material then accumulates. In contrast to water and wind erosion, in total no soil is lost from a field and no selective loss of the fines occurs. Damages are only caused by the uneven soil distribution with tillage erosion. In slope positions prone to tillage erosion (upper field edge, crests, shoulders) soil losses can be much larger than the losses by water and wind erosion.

4.3. Building Over and Sealing [66-68]

Building over and artificial sealing extensively damage all soil functions; infiltration and gas exchange, for example, are impeded, temperature buffering is impaired, and plant growth is scarcely possible any more. Even a minor amount of sealing may have consequences. Thus a land reconsolidation with asphalting of only 6 % of the surface by field roads increased peak runoff twelvefold [69]. Despite the higher runoff, the net percolation is in some cases greater under sealed areas (e.g., asphaltic surfacings) because of the greatly reduced evaporation. This impairs the microclimate in sealed areas (higher temperatures and lower atmospheric humidities by day, lower temperatures by night). Moreover, additional damage to the neighboring soils often emanates from sealed surfaces such as roads, owing to the drift of combustion residues or thawing salt, for example. The soils in the vicinity of sealed areas are often greatly changed by deep mixing, input of artificial products such as bricks, asphalt or concrete, or by covering with natural or anthropogenic substrates. Therefore, a significantly larger proportion of soils is damaged by urban activities than the sealed area itself.

In contrast to compaction and erosion, sealing affects limited areas only, but there frequently leads to a complete loss of soil. In some industrialized countries more soil is being lost annually by urban development than by erosion on arable areas, although the area affected by development is only 0.2 % of that affected by erosion.

Whereas compaction and erosion have natural causes also and are merely increased by humans, they alone are responsible for building over. Nevertheless, just this kind of damage is increasing more and more, especially in industrialized countries. There, the settlement area rises by ca. 1 % per year, since the area needs of the individual increase and there is a trend to styles of residential construction that use up much soil area. In order to limit the damage or partly to cancel it, the following measures are necessary in case of existing or planned building over:

1. More compact styles of construction with smaller consumption of soil area per unit of useful area
2. Land recycling, i.e., redevelopment of formerly developed areas instead of allocation of new construction sites
3. Replacement of coverings that greatly impede mass transfer by more permeable coverings (Fig. 15)
4. Removal of sealing where it is unnecessary (e.g., central strip between the track grooves of roadways)

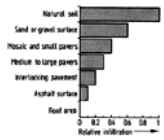

Figure 15. Relative infiltration of precipitation with different kinds of sealing (according to [66], modified)

As a rule these measures are also perceived as an esthetic gain (e.g., [70]).

[Top of Page]

5. Effects of Anthropogenic Substances on Soil

Norbert Litz, Robert Mayer, Alexander Stoy, Kassem Alef, Helmut Kerndorff

5.1. Persistent Pollutants [78-81]

The growing production and use of organic chemicals, their worldwide dissemination, and ignorance about the behavior of substances in the environment are some of the reasons why persistent organic substances can be detected in the environment, and not only in industrial conurbations. About 10×10^6 chemical substances are registered with the CAS. Day after day about 1500 new substances are registered, that is, 500 000 new substances are developed or registered every year. But of these 10×10^6 substances only about 100 000 are produced and traded industrially at present, and of those, according to information from the German Association of the Chemical Industry (Verband der Chemischen Industrie, VCI) about 5000 substances are produced in amounts of more than 10 t and 1080 in amounts of more than 1000 t/a [82]. By comparison with the number of substances produced and traded, the number of persistent substances known up to now is small.

5.1.1. *Definition and Criteria*

Persistent compounds are durable substances that are stable in the biosphere under natural climatic and soil conditions and also towards soil flora and fauna (the edaphon). Substances can be called persistent whose total degradation in the biosphere does not occur in one vegetation period. To be classified as nonpersistent half of the amount of a substance must be degraded within 1.5 months. Such criteria are also drawn up for plant-protection agents [83]. OTTOW defines persistence as the time required for 95 % degradation [84]. Persistent compounds are frequently xenobiotic and are among the environmental chemicals.

5.1.2. *Sources and Entry Paths*

Chloroorganic compounds in particular represent the bulk of persistent compounds in the environment. At present, about 3×10^6 t of chloroorganic compounds are produced annually in Germany. Mainly low-boiling chlorinated hydrocarbons are emitted into the environment in noteworthy amounts, i.e., between 60 and 90 %. Pollutants are released not only from production sites but also during the consumption and application of products [85].

Pollutants can enter the soil via the *air path* (dry deposition); the *water path* (sewage water, precipitation, surface water, irrigation water, i.e., wet deposition); and via *organic solids*, such as sewage sludges, composts, organic residual substances, or agricultural inputs (soil-improvement agents, fertilizers, manure, plant-protection agents).

Air Path. The importance of the air path for the entry of organic compounds into soils is recognized more and more. For 1986 the total emission of volatile organic substances into the atmosphere originating from the Federal Republic of Germany was estimated at 2.45×10^6 t. According to a description of SCHLEYER et al. for the Federal Republic of Germany before 1990 (area 248 798 km^2), 7000 t/a benzene, 250 000 t/a toluene, and 250 000 t/a nonvolatile chlorinated hydrocarbons were released and gave calculated mean concentrations in the rainwater of 336 µg/L, 1200 µg/L, and 1200 µg/L, respectively [86].

Organic substances can enter the soil in gas form, dissolved, or particle-bound. The relative importance of individual kinds of deposition is determined by the substance properties, the particle supply, the kind of emission, and meteorological parameters. Substances with low vapor pressure, e.g., polychlorinated aromatic biphenyls (PCBs) are present in the air mainly in the gas phase and hardly at all in the dissolved aqueous phase [87]. Similar results have been obtained for polycyclic aromatic hydrocarbons (PAHs) and other chloroorganic compounds [88]. Dry deposition means the deposition from the atmosphere onto the earth's surface. Substances can be deposited from the gaseous or particle-bound phase.

Wet Deposition. Studies on wet deposition (i.e., entries as a result of rain, snow, mist and dew) have been performed in the case of low-boiling chlorinated hydrocarbons and aromatics [89-92]. The behavior of PAHs and PCBs in rainwater has also been studied [91], [93-95]. A survey can be found in [96]. CAPEL et al. [97] determined depositions of *n*-alkanes in Switzerland at concentrations of 150 – 170 µg m^{-2} a^{-1} by mist and at 400 – 4000 µg m^{-2} a^{-1} by rain and depositions of PAHs at 90 – 900 µg m^{-2} a^{-1} by mist and at 100 – 600 µg m^{-2} a^{-1} by rain. In the United States, PAH depositions of 350 µg m^{-2} a^{-1} for urban locations and 100 µg m^{-2} a^{-1} for rural areas have been determined.

Surface Waters. The pollution of soils by organic pollutants through infiltration of contaminated surface water (caused mainly by sewage waters) tends to decrease. However, the pollutants concentrations in surface waters are distinctly higher than the concentrations in rainwater, sometimes by a factor 10 or more [98]. When flooded and irrigated by surface water, soils come into contact with its constituents. Areas of flooding therefore often constitute endangered areas [99].

Sewage Water and Sewage Sludge. Some substances, e.g., surfactants, enter the soil environment exclusively via sewage water and sewage sludge. It is estimated, for example, that large amounts of linear alkyl benzene sulfonates remain in the sewage sludge [101], [102]. Leakages in sewage water systems can also contribute significantly to soil pollution.

Within the framework of measures for groundwater enrichment, old shutdown irrigation fields are increasingly being reactivated and rainwater percolating areas (infiltration ponds) used. Studies on substance retaining powers of the upper active soil layers and on the contamination of the underground and groundwater showed that an accumulation of PAHs, PCBs, and dibenzodioxins/dibenzofurans in particular occurs in the mud zone of infiltration ponds [103]. In the sediment of such a pond, 15 µg/kg toluene, <10 µg/kg benzene, and <1 µg/kg trichloroethylene and tetrachloroethylene were determined. The PCB contents were up to 260 µg/kg and the PAH contents between 1 and 2 mg/kg d.m. (dry matter). Hexachloro- and pentachlorobenzene also were detected in traces.

The revision of the German TA Siedlungsabfall (Technical Instructions for Estate Waste) will lead to an increasing consideration of *organic solids*, such as sewage sludge, composts, organic residues, etc. In Germany, ca. 30 % of all municipal sewage sludges are spread in agriculture. The concentration of some pollutants in sewage sludges is therefore subject to limit values: thus, e.g., the dibenzodioxins and dibenzofurans to 100 ng TEQ (Toxicity Equivalents of the BGA)/kg d.m.; the PCB congeners 28, 52, 101, 138, 153, and 180 each to 0.2 mg/kg d.m.; and AOX (adsorbable organic chloro compounds) to 500 mg/kg d.m. The pollutant load of sewage sludges — on which limit values are imposed in the German regulation on sewage sludge (Klärschlammverordnung) [104] — has continuously decreased since the 1970s. The extreme and mean values of different pollutants in more than eighty-nine different German municipal sewage sludges in 1989 are listed in Table 9 [105].

Table 9. Extreme and mean values of various organic pollutants in sewage sludges (data in mg/kg d.m.) [105]

Compound	Minimum	Maximum	Mean value
Fluoranthene	0.26	9.38	1.495
Benzo[*b*]fluoranthene	0.13	4.07	0.665
Benzo[*k*]fluoranthene	0.05	1.97	0.36
Benzo[*a*]pyrene	0.07	5.16	0.57
Indano[1,2,3-*cd*]-pyrene	0.03	3.07	0.36
Benzo[*g,h*]perylene	0.03	2.50	0.36
Di(2-ethylhexyl)phthalate	15	314	110.9
AOX	42	4400	360.9
PAHs (total)	1	21.81	5.508

In the case of PAHs the proportions by mass were between 1 and 21.81 mg/kg d.m. for individual PAHs, with the exception of fluoranthene (0.26 – 9.38 mg/kg d.m.). Benzo[*a*]pyrene reached values between 0.07 and 5.16 mg/kg d.m. [105]. Contamination of the sewage sludges by surfactants (linear alkyl benzene sulfonates and nonylphenol ethoxylate) in somecases reaches the g/kg range and can strongly influence the behavior of other pollutants. The mineral oil hydrocarbons also reach similar orders of magnitude (g/kg d.m.).

5.1.3. *Processes in Terrestrial Soils*

Distribution. The processes in terrestrial ecosystems determine the distribution of chemicals in the soil. The distribution via the *air path* is influenced by climatic/meteorological factors (main wind direction, air temperature, frequency and intensity of precipitation, terrain topography, distance from emitter). The spread via the *water path* by seepage and infiltration is influenced by the intensity and frequency of the water input, the permeability of the soil, the properties of chemicals and the chemical properties of the soil. If an ecosystem is regarded as a compartment, the input and output of substances must be considered. Both processes are time-dependent. Between the two processes fixation, reaction, or mineralization occurs.

Sorption. Organic pollutants are subject in the soil to the same processes as, e.g., pesticides. After the pollutant enters the soil, it may be sorbed on solids (soil matrix) or remain in the gas phase or solution phase. Substances that have a higher density than air move downwards in the soil, while substances that are lighter than air can also be transported upwards.

The *bioavailability* is closely connected to the effect thresholds. The toxic threshold value is below the no observed effect level (NOEL, ⟶ Toxicology – Toxicological Studies: General Aspects). For estimating inputs over area and time and their effect on the ecosystem, the concept of critical load and critical level [106] is used. This means the pollutant loads that, for certain organisms in the soil and plants, still cause no effects.

Organic pollutants are bound in the soil as a function of their water solubility, molecular structure, and functional groups. The soil is basically a multiphase system, with solid, liquid, and gas phase. The liquid phase can be broken down spatially into a hydrophobic and a hydrophilic phase and therefore considerably changed in its properties by dissolved organic substances, organic complexes, and suspended particles [107]. Contaminated soil water can contain miscible or immiscible organic phases of cosolvents that can considerably influence the sorption behavior of contaminants [108-111]. The soil sorption coefficients of chemicals decrease semilogarithmically with increasing concentration of the cosolvent in water. In addition to the type of bonding (van der Waals, hydrophobic, hydrogen-, charge-transfer, complex, and ionic bonding, induced ion – dipole or dipole – dipole interactions, see also [112]), the soil properties (contents of humus, clay, oxides, and hydroxides, the pH value, redox potential, and charge [113]), and the properties of the chemicals (water solubility, molecular size and structure, lipophilicity, acidity, polarity, vapor pressure, [114]) are also important. The sorbents contained in the soil have very different properties. The organic matter is an important sorbent for all kinds of organic compounds, e.g., cationic or anionic, polar or nonpolar, because it can interact with the functional groups contained in these substances. The degree of decomposition of the organic matter has to be taken into account. Thus, the sorption capacity increases in the sequence mor, moder, and mull. Oxides and hydroxides can be further important sorbents, especially for anionic organic compounds. Clay minerals, because of their negative surface charge, are capable of binding cationic compounds as well as negatively charged compounds, such as basic nitrophenols and anilines after protonation [115]. The bonding relationships are mainly determined in batch experiments. However, more realistic predictions on sorption are obtained in leaching tests. Sorption constants are usually smaller in leaching tests since the penetration into the soil particle is less. The sorption is also time dependent, organic compounds are initially bound at the bonding sites at the surface of clay minerals and humus. Transposition and incorporation into the sorbents can occur gradually. Only in this way can it be explained that bound residues can scarcely be desorbed from the soil matrix with aqueous solution [116], [117].

Elimination here means the sum of the chemical, physical, and biological degradation processes in the soil, including volatilization and loss because of uptake of the substances by plants and seepage into the groundwater (nonextractable substance residues are often considered being part of the eliminated portion). The first degradation step is often a *hydrolysis*, which can proceed only in the presence of water. The rate of hydrolysis is influenced by the pH value and by the presence of metal cations. Hydrolysis is especially important in inputs via the air (precipitation) and surface waters. In the soil, *reductive and oxidative processes* can lead to chemical alteration of substances. Oxidative reactions occur with “reactive oxygen species” or molecular oxygen (autoxidations) in soil and bodies of water. Abiotic reductions possibly occur in the degradation of pentachloronitrobenzene under anaerobic conditions to pentachloroaniline [118].

The *physical degradation* is based on photochemical and photophysical processes, which are mainly restricted to the soil surface. Dechlorination reactions of chlorohydrocarbons are typical photoreactions. The photomineralization of adsorbed organic pollutants is based on the absorption of UV light of fairly long wavelengths (>290 nm), as is assumed for pentachlorophenol (PCP) [119]. Dibenzodioxins and -furans were 90 % dechlorinated within a week in soils [120].

Organic compounds are mainly degraded biologically (metabolic or cometabolic). Chemical and physical degradation plays a part with substances that are degradable with difficulty, for example, highly chlorinated dibenzodioxin congeners (octa- and heptachlorodibenzodioxins).

The persistence of a substance depends on its physicochemical properties and the soil properties, including the soil's microbiological potential. According to OTTOW [84], the degradation of near-natural compounds is mainly metabolic and that of xenobiotics mainly cometabolic. Persistent substances usually have unphysiological, chemically stable groups. These include, e.g., C–Cl, C–SO_3H, C–NO_2 groups, trisubstituted N atoms, quaternary carbon atoms, ether bridges, polychlorinated compounds, chlorine substituents in benzene rings in the *ortho* or *meta* positions, pyridines, and long-chain hydrocarbons. Furthermore, these substances are persistent because of their low water solubility and bioavailability and large molecular size. They are mainly cometabolic and frequently only to very stable metabolites (“dead-end” metabolites).

In soils that contain various persistent organic pollutants, the degradation may be reduced. In saturated, anaerobic soils, other microorganisms are usually necessary for the degradation of a substance than under aerobic conditions. Many chlororganic substances, for instance, are degraded more easily by reductive dechlorination under anaerobic than under aerobic conditions [121]. Alternately moist and dry conditions therefore increase the degradation efficiency of soils. The oxygen supply of the soil influences the activity of aerobic microorganisms that not only use oxygen as an energy donor but bring about hydroxylations of aliphatics and aromatic compounds by means of microbial oxygenases [122-124]. The degradation rate usually decreases with increasing soil depth, but a heterogeneous distribution of the organism density can lead to quite different degradation rates [125], as has been shown in the cases of methanol, phenol, and *tert*-butanol. The degradation is in general distinctly slower in winter because of lower temperatures [100].

Volatility. The transfer of substances from the soil into the atmosphere is known as volatilization or volatility. For aqueous solutions Henry's law constant describes the ratio of the concentration of a substance in the gas phase to its concentration in the solution phase. The volatility of highly volatile adsorbed substances is liquid-phase controlled and that of low-volatility substances is gas-phase controlled. In addition, the volatility is strongly influenced by the water solubility of the substance, the vapor pressure, the temperature, the soil moisture, and the humus content of the soil. Substances can as a rule be more easily released from moist soils, because they are more easily desorbed from the aqueous phase. The soil water can moreover make an important contribution to the evaporation of substances at the soil surface that are delivered subsequently from deeper soil layers. Diffusion plays a certain part in the subsequent release at the soil surface and therefore in the volatility of chemicals. Even low-volatility compounds, such as dibenzodioxins/dibenzofurans, with extremely persistent properties can be volatilized into the atmosphere [126].

Uptake by Plants. The uptake of organic pollutants in plants is determined by their water solubility, molecular mass, and lipophilicity, as well as their concentration in the soil and the soil properties. The uptake can occur via the root with the transpiration stream; as a result of uptake via the leaves (dry and wet deposition); and by dust particles, degassing and spraying of the soil by the impact of rain.

5.1.4. *Persistent Organic Compounds*

The persistent organic substances also include organochlorine compounds, such as hexachlorocyclohexane (HCH) hexachlorobenzene (HCB), DDT, and 1,1-dichloroethenylidenebis-(4-chlorobenzene) (DDE), that have been used as pesticides (⟶ Insect Control – Organochlorine Insecticides). These substances and those dealt with below have in some cases been detected worldwide in various environmental media (water, sediment, soil, organisms, and the atmosphere). Their contents in environmental media decrease only slightly, even over fairly long periods.

5.1.4.1. *Chlorinated Aliphatic Hydrocarbons*

Some haloalkanes (used as organic solvents) behave as persistent compounds. The low-boiling chlorinated hydrocarbons occur ubiquitously, they are partly of biological origin. Algae in particular are capable of forming halogen compounds. It is assumed that ca. 10 000 t/a of organohalogen compounds are emitted into the atmosphere worldwide from macroalgae [127]. In 1987 about 250 000 t/a of low-boiling chlorinated hydrocarbons of industrial origin were consumed in the Federal Republic of Germany and at the same time partly emitted into the environment [128]. Relative to the average rainfall of 837 mm/a, this would yield a hypothetical mean concentration of 1.2 g per milliliter rain water, i.e., at uniform deposition ca. 1 g per square meter soil surface [86].

The most important low-boiling chlorinated hydrocarbons are dichloromethane, 1,1,1-trichloroethane, trichloroethylene, and tetrachloroethylene. Some of the low-boiling chlorinated hydrocarbons, e.g., tetrachloroethylene, have a carcinogenic potential. They reach the environment via the air and via the sewage water path, from contaminated sites, and by accidents. Some of the low-boiling chlorinated hydrocarbons are quite stable in the atmosphere (e.g., 1,1,1-trichloroethane has a tropospheric half-life of 63 a [129]) and are photooxidized (e.g., tetrachloroethylene to trichloroacetyl chloride). After hydrolysis, the water-soluble, low-volatility trichloroacetic acid [130] is formed in amounts of ca. 6 g d^{-1} km^{-2}, which yields an estimated deposition of 4 g ha^{-1} a^{-1}. SCHLEYER et al. [86] estimate depositions of 23 g ha^{-1} a^{-1} (in a spruce stand) and 6 g ha^{-1} a^{-1} (in open land) (Table 10) [86]. Trichloroacetic acid is also formed in the photochemical degradation of trichloroethylene. Carbon tetrachloride is especially stable in the atmosphere (half-life 60 a). Many low-boiling chlorinated hydrocarbons accumulate in soils relative to the atmosphere (see Table 11, e.g., $CHCl_3$, CH_3CCl_3, C_2Cl_4), but their outgassing as well as their metabolites also have to be reckoned with [131].

Table 10. Mean concentrations of selected organic parameters in rainwater calculated from four measuring stations and deposition rates determined from them; basis of calculation is a mean rainfall quantity in Germany of 837 mm/a [86]

Substance	Spruce stand		Open land	
	Mean concentration, ng/L	Deposition rate, g ha^{-1} a^{-1}	Mean concentration, ng/L	Deposition rate, g ha^{-1} a^{-1}
AOX	29 000	243	4700	39
Chloroform	56	0.5	14	0.1
Trichloro- acetic acid	2800	23	730	6.1
Toluene	47	0.4	48	0.4
ϒ-HCH (lindane)	36	0.3	39	0.3
Dibutyl phthalate	310	2.6	265	2.2
Di(2-ethyl- hexyl) phthalate	790	6.6	580	4.9

Table 11. Contents of various low-boiling chlorinated hydrocarbons (LBCHs) in soil atmosphere and in the atmosphere of a forest in the Black Forest in 1989 (means of six measurements) [92]

LBCH	Soil atmosphere, µg/m^3	Atmosphere, µg/m^3
$CHCl_3$	41.8	0.167
CH_3CCl_3	2.27	1.3
CCl_4	0.5	0.66
$CHClCCl_2$	0.23	0.58
C_2Cl_4	2.98	1.85

The persistent low-boiling chlorinated hydrocarbons, the half-lifes of which in soils are more than 5 months, include in particular trichloroethylene, dichloromethane, trichloromethane, carbon tetrachloride, and vinyl chloride. Under *water-saturated conditions* the dominant degradation processes are reductive dehalogenation (e.g., with tetrachloroethylene), dehydrohalogenation (e.g., hexachloroethane to 1,1,1-tetrachloroethane), and hydrolytic dehydrohalogenation (thiolysis of the C–Cl bond in dichloromethane). In *anaerobic media* mainly higher halogenated alkanes are dehalogenated. Reductive dehalogenation depends on the redox potential of the reactants and requires an anaerobic medium [132]. *cis*-1,2-Dichloroethylene and vinyl chloride frequently accumulate as degradation products in groundwater, because of the slow transformation of vinyl chloride to ethylene. 1,1,1-Trichloroethane proves to be extremely persistent in water-saturated soils (half-life in the groundwater 2.8 – 19 a [133]). The assumption of slow degradation in water-saturated soil is moreover supported for tri- and dichloromethane [134], [135].

Under *aerobic conditions* haloalkanes are oxidatively dehalogenated by monooxygenases. In the case of C_1 haloalkanes, the corresponding aldehydes and carboxylic acids are formed successively. Mono- and dihaloalkanes in particular are degraded more rapidly under aerobic than under anaerobic conditions. The degradation of carbon tetrachloride is quite slow, even under aerobic conditions. During degradation by the various oxygenases the metabolite trichloroethylene epoxide, as a cell poison, inhibits the further elimination of trichloroethylene and trichloromethane [136].

The degradation should not lead to toxic “dead-end” metabolites such as *cis*-1,2-dichloroethylene or vinyl chloride. In the degradation of dichloromethane, which is regarded as easily degradable, the presence of 1,2-dichloroethane causes a competitive inhibition of the dichloromethane-degrading enzymes [137]. Dichloromethane is degraded more strongly (faster) in presence of acetate than in its absence [138]. The presence of formate also had a stimulating effect [139]. Formate can thus be advantageously used in in situ remediations. Low-boiling chlorinated hydrocarbons are weakly bound in the soil, because of their high water solubility (up to 20 g/L) and high vapor pressure. The main sorbent is the organic matter of the soil. The Freundlich sorption constants K (K_{OC} relative to the humus content) vary from <1 to 300. The low binding by the soil also explains the large number of nearly 3000 cases of groundwater pollution by chlorinated hydrocarbons reported to the German water authorities in 1990 [140]. Remediation measures that enhance the bioavailability of chlorinated hydrocarbons by increasing the desorption from the soil matrix, e.g., by surface active substances, should only be used if no additional risk arises to soil and groundwater.

5.1.4.2. *Aromatics and Chlorinated Aromatics*

Aromatic hydrocarbons are among the most widespread organic compounds [141]. Thus, e.g., in the Federal Republic of Germany in 1985 alone, ca. 76.6×10^6 t of crude oil were distilled and subsequently processed to energy carriers and fuels. The motor vehicle fuels produced contained ca. 680 000 t of benzene, 2.4×10^6 t of toluene, and ca. 3.1×10^6 t of xylenes. Simultaneously, 1.6×10^6 t of pure benzene, 360 000 t of toluene, and ca. 470 000 t of xylenes were produced for further chemical processing [142]. According to an estimate of the Umweltbundesamt (UBA, Federal Environmental Agency) for 1982, the total emisson of benzene from motor vehicles, heating plants, and industrial production was ca. 60×10^6 t/a, but distinctly smaller amounts can be assumed today. At the end of the 1980s, the 70 000 t/a of benzene and 250 000 t/a of toluene were emitted into the environment by motor vehicle traffic. Despite the enormous amounts introduced into the environment, the benzene, toluene, xylene (BTX) aromatics can only be detected in small concentrations (0.1 – 1 µg/L) in water bodies. The concentrations in municipal sewage water are greater by a factor ten and industrial wastewaters can be loaded with 1 – 10 g/L. More than 90 % of the BTX is eliminated in sewage treatment plants. Annually, 3.5 t benzene are said to be introduced to the environment via sewage sludges and ca. 80 t into bodies of water [143]. The benzene contents of the air in urban zones are between 5000 –50 000 ng/m^3, in rural zones 500 – 5000 ng/m^3, and in remote areas <2000 ng/m^3. In the case of toluene, the pollution is 5000 – 10 000 ng/m^3 for urban zones, 1000 – 7000 ng/m^3 for rural areas, and ca. 750 ng/m^3 for remote areas. The concentration of the various xylene isomers varies from 800 – 70 000 ng/m^3 and from 200 – 1200 ng/m^3, no data being available for areas remote from sources [86], [144]. The ethylbenzene contents are also relatively high. In precipitation waters the pollution due to aromatics is higher by a factor of ca. 1000, depending on the distance from emitters. Thus, e.g., <220 ng/L toluene, <260 ng/L xylene (*m*-, *p*-), and up to 72 ng/L ethylbenzene are found in rainwaters [145]. Rain loadings of a similar order were determined in the Rhine – Main area and in the hills of Hesse, the concentrations being distinctly higher under spruce stands than in open land. This is a result of the effect of trees in “mopping up” air pollution. In both cases the loads were higher in winter than in summer.

No representative data are available for background concentrations in soils for BTX, but data exist for higher-chlorinated benzenes and phenols [99]: an average of 0.1 µg/kg d.m. was measured for pentachlorobenzene and 1 – 2 µg/kg d.m. for hexachlorobenzene.

In Switzerland the background values vary between 0.15 and 0.5 µg/kg d.m. for hexachlorobenzene [146]. According to studies of [147] and [148], the pentachloro- and hexachlorobenzene contents in the humus cover of forest sites (6.5 – 11.5 µg/kg d.m.) are distinctly higher than those of the mineral horizons (1 – 2 µg/kg d.m.) or of sites on open land (<5 µg/kg [149]). In forest soils near Linz [150], the hexachloro- and pentachlorobenzene contents were 0.38 – 0.4 µg/kg d.m. and in the humus cover 0.147 µg/kg d.m.

In the soils themselves, benzene, toluene, and xylene are only moderately bound. K_{OC} values are <100 for benzene, <500 for toluene, and <1400 for xylene (Table 12). The aromatics are bound mainly to the organic matter. The aerobic degradation of BTX in the soil is very rapid (half-lifes 2 – 16 d [151]). In water-saturated soils, on the other hand, these substances prove to be more stable; apparently volatilization often makes the main contribution to elimination [152]. The biradical oxygen molecule oxidizes BTX aromatics to phenols. Subsequently the ring opens. This process is inhibited under anaerobic conditions and only occurs at a sufficient rate if the ring is substituted by carboxyl groups. Thus, a strong benzene contamination in anaerobic underground was not degraded even after 4 years [153]. Strong persistence and small degradation under anaerobic aquifer conditions was also established for aromatics [154], [131]. This behavior is strictly related to the lack of molecular oxygen. After dissolved oxygen was supplied, a rapid degradation occurred [155]. On the other hand, a degradation could occur when ferulic acid (4-hydroxy-3-methoxycinnamic acid) [156] was present or iron reduction was possible.

Table 12. Sorption of aromatics and haloaromatics to organic matter [110], [130]

Substance	K_{OC}	Substance	K_{OC}
Benzene	60 – 100	Phenol	<100
Chlorobenzene	210 – 390	Chlorophenol	100 – 300
Dichlorobenzene	230 – 970	Dichlorophenol	300 – 700
Trichlorobenzene	4300	Trichlorophenol	100 – 570
Tetrachlorobenzene	85 000	Tetrachlorophenol	150 – 750
Pentachlorobenzene	11 400	Pentachlorophenol	300 – 950
Hexachlorobenzene	3900 – 11 800	Toluene	300 – 510
Ethylbenzene	300 – 600	*p*-Xylene	400 – 1400

Before aromatics are illuviated into the soil, they can evaporate into the air. Owing to the high vapor pressure of BTX, high volatilization rates are to be expected. In the atmosphere, benzene, for example, undergoes a photochemical degradation by OH radicals, NO_x, and ozone (⟶ Air – Photochemical Degradation).

The volatility of aromatics is used especially in the remediation of contaminated areas. Volatility depends greatly on the porosity and the moisture of the soil. The relative humidity (rH) at a certain water content of the soil determines the composition of the vapor phase of BTX, etc. very strongly. A stepwise increase of the relative humidity from 0 – 20 % produced a reduction of the binding of *p*-xylene to kaolinite [114]. At 67 rH only 50 % of the xylene remained bound. Compounds such as BTX with Henry's law constants of >10^{-3} bar mol^{-1} m^{-3} can be removed from the soil and groundwater by stripping [157], they can also be reduced relatively rapidly in bioreactors with 100 % turnover. Under field conditions the remediation measure was distinctly slower.

Chlorinated Aromatics. Chloro-substituted aromatics behave distinctly differently in the environment, especially in terrestrial ecosystems, from the pure aromatics. The sorption in the soil increases with increasing degree of chlorination. The solubility of the BTX and chlorinated aromatics in the soil can be raised by addition of vegetable oils or by biosurfactants formed by microorganisms, so that a more efficient remediation is possible via the water phase [158]. But the degradation can also be reduced if the toxic threshold concentration for microorganisms is exceeded. It has been reported that biopolymer-bound trichlorophenol was stable to degradation, and a negative effect of the surfactant on the microorganisms also being possible [159].

5.1.4.3. *Nitroaromatics and Aromatic Amines*

Nitroaromatics play a notable role in regard to the amount of inputs from the atmosphere. They are formed from BTX after photochemical reaction with NO_2 and OH radicals in the troposphere. Through rainfall nitroaromatics enter soil, leachates, and spring waters in low concentrations over a wide area [86]. As intermediate and end products in the production of plant-protection agents, pharmaceuticals, plastics, dyes, and explosives, they reach the environment in the course of their processing and use. Most important are the various nitrophenols, nitrotoluenes, nitroanilines, and pyridines, as well as chlorine-derivatives of these compounds. It is estimated for Germany that ca. 20 000 t of nitrophenols are emitted annually into the atmosphere or formed there [160]. Aromatic amines such as chloroanilines are determined in surface water in concentrations of 1 – 5 µg/L but also distinctly higher

values are found [161].

Because of their considerable toxicity to microorganisms and humans and their mutagenic potential, which is in some cases considerable, nitroaromatics and aromatic amines are described in more detail here. In the the soil aromatic amines and nitrocompounds get into contact with the clay minerals and the organic matter via their nitro and amino groups. Anilines react as bases and are often present in protonated form, depending on the pH value. With decreasing pH value, increasing sorption to the mineral matter occurs because protonated amino groups are more strongly bound than the free bases. Anilines can form bonds with exchangeable protons of silicate surfaces or positively charged hydroxide surfaces. They react with hydrated exchangeable cations of silicates and can be indirectly bound via hydrogen bridges. They can also be coordinately bound by means of electron transfer [162]. In soils, aromatic amines are oxidized by free radicals, e.g., at the amino group. The reaction is catalyzed by clay minerals. Aromatic amines can form complex compounds with metals. At fairly high pH, pyridines and other compounds occur as neutral species and can then also bind to the organic matter. Ionic bonding to the soil matrix occurs especially when the pH of the soil is near the pK_a of the amine [163]. Pyridines can also be hydrophobically bound [164]. With few exceptions, the binding of aromatic amines and nitro compounds to the organic matter is only moderately strong (see Table 13). Pyridines [164], [166] and also nitro compounds [167] are therefore repeatedly found in mg/L quantities in the groundwater. Particularly in acid soils nitrobenzenes are reduced to anilines, by sulfides, especially by Na_2S in anaerobic soils [113].

Table 13. Degradation sorption and volatilization in % of pyridine (-derivatives) in soil suspensions after 30 d [165]

Pyridine (-derivatives)	Position	Bound	Degraded	Volatilized	Loss*
Pyridine		15.6	6.3	16	100
-OH	2,3-	2.4	<1		4
-COOH	2,3-	3.0	81		100
-CCl	2,3-	15.1	<1	50	79
$-NH_2$	2,3-	14.3	<1		4
$-CH_3$	2,6-	7.7	<1	23	29

* Indicates the recovery rate.

The nitrotoluene concentrations at a blasting site of a former explosives factory were in the mg/kg range and led to the closing of a drinking water well [168]. As follows from subsequent analyses, the concentrations clearly decreased (Table 14). Under anaerobic conditions a distinctly slower degradation occurs. A trinitrotoluene contamination of groundwater near Grand Island, Nebraska was also reported [167].

Table 14. Concentrations of nitrotoluenes in the surface soil of a blasting site near Torgau in µg/kg d.m. in 1992 as a function of time (data by the Regional Environmental Agency, Torgau)

	January	February	March	April	May	August	October
2,4-DNT [a]	390	18	n.d.*	96	n.d.*	<5	n.d.*
2,6-DNT [a]	360	8	<5	26	<5	7	n.d.*
TNT [b]	2860	156	31	260	10	8	n.d.*
2,4-Dinitro-6-aminotoluene	1220	38	33	560	<5	7	n.d.*
2,4-DN-4-AT [c]	2240	82	46	920	<5	8	n.d.*
Cyclonite	2700	78	56	210	8	24	n.d.*

[a] DNT = dinitrotoluene.
[b] TNT = trinitrotoluene.
[c] DN-4-AT = dinitro-4-aminotoluene.
* n.d. = not detected.

The degradation of 2-nitroaniline and 4-nitrophenol was studied [169]. The bacterium *Nocardia sp*. can degrade these substances to 4-hydroxy-2-nitroaniline and nitrocatechol. The removal of a nitro or amino group associated with a hydroxylation has been observed with 2-nitrophenol, 4-chloroaniline, and 4-chloro-2-nitrophenol. An extremely slow degradation of chloroanilines by reductive dehalogenation by microorganisms under anaerobic aquifer conditions was observed [170]. Under methanogenic conditions the degradation rates were slightly greater (K < 0.01 – 1.36 after 8 months) than under sulfate-reducing conditions (K <0.01). The only exception was provided by 2,3,4-trichloroaniline, which was completely transformed after 10 months. Under some circumstances, pyridines are rapidly degraded; thus, e.g., pyridine itself is eliminated within 7 d [163]. However, also very persistent pyridine derivatives exist (Table 13), that frequently are also rather volatile. The differences between aerobic and anaerobic degradation are small. The persistence of the pyridine derivatives decreases in the following order: pyridinecarboxylic acid < pyridine = monohydroxypyridine < methylpyridines < aminopyridines = chloropyridines [165].

5.1.4.4. *Polychlorinated Dibenzodioxins (PCDDs) and Dibenzofurans (PCDFs)*

Dibenzodioxins and dibenzofurans are classes of aromatic ethers, with totals of 75 and 135 congeners, respectively. There are congeners substituted with 1 to 8 chlorine atoms (mono- to octachlorinated), which can also differ in the position of the chlorine atoms in the organic rings. Chlorine can be replaced by bromine, some substances can contain chlorine and bromine, so that finally 5020 different compounds of this type are possible. The toxicity of the dioxins and furans is rated by quoting a toxicity equivalent (TEQ) according to the BGA [171].

Chlorinated and/or brominated dibenzodioxins and dibenzofurans reach the environment mainly through thermal reactions (motor traffic, refuse incineration plants, industrial incinerators and blast furnaces, domestic heating systems, as byproducts of syntheses), etc. [172]. Also flame retardants that consist of bromine-containing phenols, benzenes, and biphenylene contain dibenzodioxins and -furans. Precursor compounds for chlorinated dibenzodioxins and dibenzofurans are chlorinated phenols, benzenes, polychlorinated biphenyls, and other aromatics in combination with oxygen and chlorine. Sulfur-containing octachlorodibenzodioxin (OCDD) analogues reach the environment via tire abrasion. Since dioxins and furans reach the environment mainly particle-bound or gaseous, degradation in the atmosphere is important. The hepta- and octachlorinated congeners especially are strongly bound and carried in the soil via dust particles or dust precipitation. The lower-chlorinated congeners seemingly undergo a more rapid photochemical degradation as a result of their higher vapor pressure. The annual PCDD/PCDF inputs to the environment of Germany, Sweden, Switzerland, and Norway have been estimated [173], [174]; inputs of ca. 0.928 kg (Germany), 0.14 – 0.82 kg (Sweden), 0.285 – 0.37 kg (Switzerland) and 0.430 kg (Norway) were determined. This calculation disregarded the large number of industrial blast furnaces and incineration plants with considerable PCDD/PCDF emissions. Only after these plants have been re-equipped can large reductions of the input be expected. At the end of the 1980s, 48 g of PCDD/PCDF were said to be introduced into soils annually through sewage sludges used in agriculture and amounts of the same order through liquid manure and composts (2.6 – 47 and 32 –80 g TEQ, respectively) [175]. Today the inputs are presumably distinctly less because the loads of sewage sludges and composts are no longer as high as in the 1970s or 1980s. Dioxin contents in surface waters of Hesse (the Main river) were 0.1 – 0.2 pg/L TEQ, in the corresponding suspended material 8 – 20 ng/kg d.m. TEQ, and in the sediments 9 – 19 ng/kg d.m. TEQ [176].

The immission air loads of rural regions are 0.025 to 0.07 pg TEQ/m^3, with a deposition of 5 – 20 pg TEQ m^{-2} d^{-1}. In urban regions the load is 0.07 – 0.35 pg TEQ/m^3 and 10 – 85 pg TEQ m^{-2} d^{-1}. In the neighborhood of emitters, immission values of 0.35 – 1.6 pg TEQ/m^3 and deposition values of ≤1000 pg TEQ m^{-2} d^{-1} are reached [177]. Loads in the rainwater (Dübendorf, Switzerland) were 0.03 – 0.1 and in mist 0.3 – 3 ng/L [97].

Viewed over time, the inputs of dioxins and furans to soils distinctly rose from 1920 and reached their maximum in the 1970s [178]. Studies in England on lake sediments show increases of dioxin since 1940 [179] (Fig. 16). Core drillings of lake sediment in Berlin indicate inputs falling overall by now.

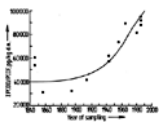

Figure 16. Development of PCDD and PCDF contents in the surface soil of Rothamsted, United Kingdom [178]

The background load for soils was quoted as 0.8 ng TEQ/kg d.m. for Baden-Württemberg [180]. The contents in Bavaria are of a similar order [181]. Loads in urban areas and areas with greater industrial use are 3.4 and 8.4 ng TEQ/kg d.m., respectively [129]. On the other hand, the dioxin loads in North Rhine-Westphalia in rural areas were 0 – 10 ng/TEQ/kg d.m. (50 % < 5 ng TEQ/kg d.m.), and in the conurbations 10 – 15 ng TEQ/kg d.m. [99]. The dioxin load in forest soils reaches values of 50 – 80 ng TEQ/kg d.m. in the litter cover. Table 15 gives a compilation. The soils (0 to 5 cm) in the surroundings of metal works have distinctly higher loads (e.g., in Linz, Maulach, and Rastatt with up to 332 ng TEQ/kg d.m., 7926 ng TEQ/kg d.m., and 29 039 ng TEQ/kg d.m., respectively [150]. In the grassland young growth near such plants, 20 – 40 ng TEQ/kg d.m. were determined [182]. Correspondingly, increased dioxin loads were found in milk.

Table 15. Range of the dioxin loads of soils in Germany, compiled from data of the different states (situation at May 15, 1991; TEQ-calculation according to BGA [177])

Loaded area		Mean (ng/kg TEQ)	Range of the means	Minimum value	Maximum value	Number of samples	Sampling depth, cm
Soils in rural areas (background)	forest soil	5	1.4–14	0.04	46	33	[a]
	forest litter	35	7–56	0.3	140	41	
	grassland	3	0.6–7	0.04	10.7	95	1 state 0–5 4 states 0–10
	arable land	2	0.2–5	0.01	7.7	178	0–30
	other	5 (1 state)	2–7	0.03	35	34	0–2
	soils treated with sewage sludge	5	3–7	0.025	21	46	grassland 0–10 arable land 0–30
Variously polluted soils in conurbations	forest	2 (1 state)		0.8 (1 state)	7.4	48	1 state 0–5 1 state 0–15
	forest litter	10			75.1	29	
	arable land	2.8			32	98	0–30
	grassland	3.8		0.5	100	223	1 state 0–5 2 states 0–10
	residential areas	9 (1 state)			30	10	
	parks and children's playgrounds			5	1 594	126	1 state 0–10 1 state 0–5 & 0–40
	kindergartens	12		2.5	27.3	86	1 state 0–30 1 state 0–5 & 0–40
	other	32 (1 state)		8.6 (1 state)	73.0 (1 state)		
Soils in the area around defined emitters	refuse incineration plants	43		0.01	202	118	[b]
	landfills			0.8	188 (1 state)	5	
	old sites			3	200 000		1 state 0–10 1 state 0–5 & 0–40
	chemical industry	8	5–13	0.05	1 658	57	1 state 0–10 1 state 0–5 & 0–40
	metal processing			0.07	4 958	262	[c]
	road traffic	10	10–20		200	30	1 state 0–5 1 state 0–30
	other therm. processes	65	1–160	0.02	800	47	[d]
	other chem. processes	7		0.05	89	27	1 state 0–10 1 state 0–10

[a] 1 state 0–4 cm; 2 states 0–5 cm; 1 state 0–12 cm; 1 state 0–15 cm (horizon-dependent).
[b] 1 state 0–10 cm; 1 state 0–5 and 0–30 cm; 1 state 0–40 cm.
[c] 1 state 0–5 cm; 2 states 0–10 cm; 1 state in addition 10–30 cm; 2 states 0–30 cm.
[d] 1 state 0–5 and 0–40 cm; 3 states 0–10 cm; 1 state 0–25 cm; 1 state 0–30 cm.

Because of the low water solubility and high lipophilicity of dioxins and furans, they are sorbed to the organic matter (Table 16). The K_{OC} values increase with increasing degree of chlorination, in this case from 10 000 to 57×10^6. Because of this strong sorption the higher-chlorinated congeners in particular have a low bioavailability and are nearly immobile in the soil. This is evident also in the congener distribution in soil, where the hexachloro- (HCDD) and octachlorodibenzodioxins (OCDD) dominate. The origin of the dioxins and furans can in some cases be deduced from the congener patterns. High proportions of octa- and hexachlorinated dioxins point to inputs from pentachlorophenol (PCP) sources.

Table 16. Sorption constants, water solubility (WS) and Henry's law constant of various dioxins

p-Dioxin	WS, µg/L	log k_d	log K_{OC}	log K_{OW}	Henry's law constant
Dibenzo-	842		3.91*	4.3**	12.29
2-Chloro-	278	3.3	4.60	5.00	14.82
2,3-Dichloro-	14.9		5.19	5.60	6.61
1,2,4-Trichloro-	8.41		5.93	6.35	3.84
1,2,3,4-Tetrachloro-	0.63	4.65	6.18	6.60	3.77
2,3,7,8-Tetrachloro-	0.2	4.48	6.38	6.80	1.63
1,2,3,4,7-Pentachloro-	0.118		6.97	7.40	0.264
1,2,3,4,7–6-Hexachloro	0.004		7.37	7.80	4.52
1,2,3,4,6,7,8-Heptachloro	0.002		7.57	8.00	0.133
Octachloro-	0.0004		7.76	8.20	0.683

* Calculated from log K_{OC} = 0.989 log K_{OW} – 0.346 [183].
** log K_{OW} of [184].

In a sandy-loamy silt, increased mobility of low-chlorinated dioxins (down to 40 cm depth) was observed over 12 months [185]. Deeper displacement of 2,3,7,8-tetrachloro-*p*-dioxin (2,3,7,8-TCDD) down to a depth of 5.40 m was determined at a contaminated site [186]. After the addition of lauryl sulfate and at pH values of 8 instead of 7, the K_{OC} fell by a factor 10. Such behavior of a substance can be influenced by cosolvents (e.g., oil) [187], [188].

Near the soil's surface (0.06 – 0.13 mm) the higher-chlorinated dioxins and furans in particular are subject to a photolysis at 250 – 360 nm [189], [190] and volatilization, that can be enhanced by solvents. Photodegradation and volatilization on plant surfaces can also contribute to the elimination of dioxins [191]. The degradation of dioxins and furans proceeds extremely slowly. The half-lifes vary from 10 – 14 years [192] to several decades. In other investigations scarcely any losses, but only small lateral displacements were found after 17 years [193-195].

The dichlorobenzofurans are less persistent than the dichlorobenzodioxins. While the degradation routes of the PCDDs have not yet been elucidated, microorganisms exist in polluted waters that can degrade dibenzofurans [196]. The organisms capable of degradation include *Pseudomonas spp*. [197]. The degradation of dibenzofuran leads after dehalogenation in an anaerobic environment finally to branched and unbranched paraffins [198]. Diphenyl ether, 2-methoxydiphenyl, and 2,3-dihydroxydibenzofuran are formed as intermediates. FORTNAGEL et al. [199] assume a primary degradation by *Pseudomonas sp*. *HH69*, in which 1-hydroxybenzo-*p*-dioxin is formed via an unstable dihydrodiol and after ring cleawage is degraded to 2,2,3-trihydroxybiphenyl which is transformed to catechol after a second cleavage. The inability of the microorganisms to degrade TCDD is a result of TCDD's high toxicity to microorganisms [200], its low ability to permeate membranes, its strong sorption, and its low water solubility or bioavailability.

Studies on the degradation of dioxins showed that succinate- and biphenyl-adapted microorganisms degrade monochlorinated dioxins because succinate and biphenyl bring about the enzyme induction required for this [201]. The relative conversion rate compared to biphenylene is >72 % for monochlorodibenzodioxin and 8.9 % for 1,2,4-trichlorodibenzodioxins. The degradation product is 1,2-dihydroxydibenzo-*p*-dioxin. In the case of more highly chlorinated dioxins a steric hindrance of the enzymes is to be expected. It is assumed that in the presence of other carbon sources (e.g., naphthalene, benzene, phenol) the microorganisms utilize these first. It can furthermore be assumed that the concentration range of the dioxins between enzyme induction and toxicity is very small. Furans are as a rule taken up in distinctly greater amounts than dioxins by soil-activating organisms or organisms living in the sediment [202]. The ovaries, liver, and intestine of *Fundulus heteroditus* (a species of fish) which take up higher amounts of dioxins provide an exception.

The estimation of the loading of the food chain by dioxins and furans plays a part in the determination of limit values. Over 90 % of the PCDD/PCDF uptake of humans occurs via food [203]. The dioxin uptake by dairy cattle, sheep, etc. as a result of polluted soils can amount to 10 % or more of the amount available in the soil. On average, 20 – 56 % of the dioxins absorbed by dairy cattle enter the milk [204]. Hexachlorodibenzodioxin is more rapidly metabolized than OCDD. Compared with other dioxins/furans, the 2,3,7,8-substituted tetra- and hexachlorinated dibenzodioxins and -furans are absorbed particularly in the milk and not mainly in adipose tissue of dairy cattle. The consumption of eggs is another source of PCDD/PCDF uptake by humans.

In rats 2,3,7,8-TCDD was 100 % bioavailable after transpulmonary uptake. The transfer factors via the air – plant path were estimated for leaf vegetable species to be 0.2 ng I-TEQ/kg d.m. (International Toxicity Equivalent Factor according to NATO-CCMS, 1988) [177], [205]. Deviations: green cabbage 0.4 ng I-TEQ/kg d.m.; pasture land 0.1 – 0.2 ng I-TEQ/kg d.m. A value of 0.01 ng I-TEQ/kg d.m. was assumed for the soil – plant transfer path. Such a definition often overlooks the fact that the transfer factors frequently vary just as strongly between the plant as between the plant species. The transfer of PCDD/PCDF via the water – sediment – fish path, is relatively high (30 % transfer). Thus the congener profile found in eels, for example, is reflected in that of the sediments.

Because of the enormous toxicity, the BGA has defined use-dependent limit values for soils [171], that do not, however, take account of the soil properties.

5.1.4.5. *Polycyclic Aromatic Hydrocarbons (PAHs)*

(⟶ Hydrocarbons, ⟶ Hydrocarbons – Toxicology and Occupational Health)

Polycyclic Aromatic hydrocarbons are formed during combustion processes (domestic heating, road traffic, fires, refineries, coking plants, waste incineration plants, etc.). They include a whole series of carcinogenic compounds. In all, there are assumed to be several hundred such PAHs. They enter the environment via the air, sewage water, and sewage sludges. Usually 18 PAHs in all are considered. On lands used for agriculture the contents of benzo[*a*]pyrene (BaP) were mainly <0.2 mg/kg d.m. and of fluoranthene <0.3 mg/kg d.m. soil [149]. In urban backyards and allotments, 10 – 20 % of the soils have contents of >1 mg BaP and >2 mg fluoranthene. In Bavaria, the concentration of six PAHs (fluoranthene; benzo[*b*]- and benzo[*k*]fluoranthene; benzo[*a*]pyrene, indeno[1,2,3-*c,d*]pyrene; and benzo[*g,h*]perylene) was determined in various soils [206]. The mean of the benzo[*a*]pyrene was a factor of ten higher than the reference value of 0.05 mg/kg d.m., and two soils out of twelve were above the B value (limit value for soil investigation [207]) of 1.0 mg/kg d.m. Soils in urban center areas are more heavily polluted. These areas consist in some cases of rubble and burnt building material and contaminated sites, such as gasworks. The total PAH load of soils frequently amounts to 10 times the BaP value.

With the exception of locally concentrated inputs, it is assumed today that PAH inputs into the soil occur mainly via the air path. Loads of PAHs (21 compounds) in Dübendorf (Switzerland) were 0.1 – 0.6 µg/L in the rain and up to 60 µg/L in the mist, with deposition rates of 10 – 60 and 90 ng cm^{-2} a^{-1} [97]. The annual increases in the soil of Rothamsted (United Kingdom), at 4.56 mg/m^2, were higher by a factor 2 – 3 [208] (see Fig. 17). Since the 1890s the concentration of PAH components in soil increased by the following factors: for benzo[*b*]fluoranthene, benzo[*k*]fluoranthene, and benzo[*a*]pyrene >15; for pyrene, benzo[*a*]fluoranthene, benzo[*e*]pyrene, and indenopyrene 10 – 15; for chrysene, benzo[*g,h,i*]pyrene, acenaphthylene, acenaphthene, 5 – 10; and for phenanthrene, anthracene, fluoranthene, and naphthalene, <5. Long-term sewage sludge application from 1940 to 1960 caused a remarkable increase of PAHs (see Fig. 18) compared to manured areas or the control area [208]. Since 1960 the PAHs contents have decreased gradually due to increasing quality of the sewage sludge.

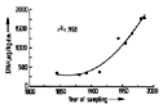

Figure 17. Development of PAH contents in the surface soil of Rothamsted, United Kingdom [208]

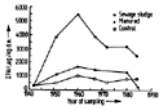

Figure 18. PAH pollution of experimental soils fertilized by sewage sludge and of a soil treated with farmyard manure [208]

The strength of binding of the PAHs in the soil differs, depending on their water solubility. The K_{OC} values increase with the number of condensed benzene rings, and specifically from ca. log 2.8 to log 7 [209]. Because of the chemical structure of this class of compound, mainly hydrophobic bonding can be assumed.

The degradation proceeds very slowly in some cases [210]. The half-lifes in an arable soil vary for less condensed PAHs from 14 – 48 d and for more highly condensed PAHs from 603 to 2030 d, depending on the stability of the compound and its molecular mass (Table 17). The results of this study show that laboratory experiments permit conclusions on the behavior under natural conditions only to a very limited degree. In smoked soils the sorption is weaker and the degradation more rapid because the sorption is not yet at equilibrium. Bioavailable PAHs can be more rapidly degraded. In two soils treated in 1968 with sewage sludge rich in heavy metals, the half-lifes of PAHs were 2 – 9 a [211], i.e., degradation was impeded. The biological degradation is important in the case of naphthalene, which is degraded via the dihydrodiol to dihydroxynaphthalene and a hydroxybenzopyruvate derivative and finally via salicylic acid to catechol [124]. The degradation of phenanthrene is basically similar, although *o*-phthalic acid is cleaved in addition. Aside from *Pseudomonas spp.*, *Alcaligenes denitrificans* also is capable of degrading PAHs [212]. The more highly condensed PAHs (>4 rings) are considered to be extremely resistant to enzymatic degradation.

Table 17. Half-lifes (d) of various PAHs in soils from microcosm studies (losses in % in brackets) [200]

Compound	RA 1 [a]		RA 2 [b]		CF [c]		RS [d]		Mean SD	
	Soil	R^2 [e]	Soil	R^2 [e]	Soil	R^2 [e]	Soil	R^2 [e]		
Naphthalene	33 (100)	0.72	16 (100)	0.86	14 (95)	0.67*	48 (91)	0.54*	28	16.0
Acenaphthene/fluorene	74 (86)	0.68	70 (84)	0.88	73 (100)	0.90*	44 (96)	0.94	65	14.3
Phenanthrene	108 (85)	0.88	193 (50)	0.8*	112 (56)	0.63*	83 (94)	0.55*	124	47.8
Anthracene	48 (61)	0.88*	210 (22)	0.34	105 (90)	0.54*	200 (59)	0.98	141	77.9
Fluoranthene	110 (78)	0.88	184 (35)	0.69	143 (60)	0.93	110 (73)	0.6	137	35.1
Pyrene	285 (36)	0.69	167 (49)	0.78	320 (20)	0.57*	127 (57)	0.90*	225	92.4
Benzo[*b*]chrysene	199 (66)	0.68	106 (76)	0.79	313 (34)	0.42*	240 (67)	0.94*	215	86.3
Benzo[*b*]fluoranthene	135 (90)	0.65	113 (44)	0.62	279 (8)	0.22*	282 (34)	0.93*	202	90.8
Benzo[*k*]fluoranthene	350 (53)	0.93	143 (57)	0.76*	351 (8)	0.23	359 (30)	0.97*	301	105.2
Benzo[*a*]pyrene	120 (85)	0.62	270 (18)	0.9	197 (44)	0.43*	258 (48)	0.92*	211	68.7
Benzo[*g,h,i*]pyrene	460 (15)	0.75*	365 (20)	0.7	460 (16)	0.83	535 (25)	0.73*	455	69.6
Coronene	2030 (16)	0.65	603 (14)	0.79	753 (17)	0.5	669 (7)	0.41	1014	680.3
PAHs (total)	146	0.90	165	0.85*	192	0.79*	177	0.82	170	19.4

[a] RA 1 = rural arable soil, no preceding sewage sludge application.
[b] RA 2 = rural arable soil.
[c] CF = rural coniferous forest.
[d] RS = roadside soil.
[e] R^2 = linear regression.
* =significance.

According to studies the half-lifes of >4-ring PAHs exceed 300 d [213]. PAHs are not degraded under anaerobic conditions [170]. In addition to photolytic degradation and chemical oxidation, volatilization also plays a certain part during the elimination in and on soils. Simple PAHs, such as naphthalene and methylnaphthalene, volatilize in proportions of 20 – 30 %, and are also degraded abiotically to 2 – 20 %. Neither is true of more highly condensed compounds.

Analyses of grass samples since 1900 prove that the highest PAH values occurred between 1931 and 1935, whereas distinct reductions occurred after 1955.

The “light” PAHs stay largely in the gas phase and can be found in the leaves or parts of fruit, whereas the heavier ones remain in the soil [214] or root parts. They can therefore be found mainly on root surfaces, such as those of carrots [211]. Eleven years after the application of sewage sludge 30 µg ∑PAH/kg d.m. were found in potatoes. The light PAHs were dominant in sugar beets, beets, and carrots [215]. There is little indication that the contents of PAHs were raised by the sewage sludge in comparison with the control.

The mobility of PAHs in soils can be distinctly raised by increased sorption on the dissolved organic matter. It has been that naphthalene and other PAHs can be displaced in the order in which they are bound to dissolved organic matter (DOM) [107]. From 100 mg/L DOM, hydrophobic substances with K_{OW} values $>10^{4.5}$ can be transported at the same rate as the dissolved organic matter. Under aquifer conditions, the mobility of the mobile phase of a PAH-containing coal tar in sandy material amounted to barely 7 cm in 54 d ca. 14 % of the naphthalene, 0.2 % of the fluoranthene and 0.06 % of the benzo[*a*]anthracene having been transformed in 91 d [216]. The transport is attributed essentially to molecular diffusion.

5.1.4.6. *Polychlorinated Biphenyls (PCBs)*

Polychlorinated biphenyls can occur in 209 congeners. About 60 individual substances with chlorine contents between 21 and 68 % have been detected. PCBs have a very low water solubility and are chemically very stable. They are therefore detected ubiquitously. About 2×10^6 t of PCBs have been produced worldwide since 1929, but they can also occur as byproducts in organic syntheses and in the pyrolysis of chlorobenzene. The coplanar PCBs 77, 120, and 169 have proved to be more toxic than the nonplanar *ortho*-substituted PCBs. PCBs are present in used oil, condensers, adhesive and sealing compounds, chipboards, ceiling panels, flame retardants, hydraulic fluids, films (as plasticizers), etc., but an extensive ban has existed on their use for open systems, in the Federal Republic of Germany since 1972 and EC-wide since 1985.

Wide-area pollution occurs mainly via the air path. In conurbations, 180 g km^{-2} a^{-1} of wet and dry deposition of PCBs is assumed. Low-chlorinated congeners distributed more by the gas phase, and high-chlorinated ones are mainly distributed particle-bound. According to CAPEL et al., PCBs (31 compounds) can be discharged at 7 – 22 µg/L in the mist and 0.002 – 0.003 µg/L in the rain, which results in a deposition of ca. 110 –330 g km^{-2} a^{-1} [97].

Municipal sewage sludges contain 0.1 – 1.0 mg/kg d.m. [105] and river sediments 0.01 – 1.0 mg/kg d.m. of PCBs. The 90 % percentile values are about 4.8 µg/kg d.m. for rural areas (grassland) and ca. 15.5 µg/kg d.m. of PCBs (13 congeners) for conurbations. The three coplanar PCBs usually amount to 10 % of the whole. The PCB contents were raised by a factor 10 – 15 especially in the flooded area of the Rhine [99]. In forest soils, the mean PCB content in the humus cover was ca. 85 µg/kg d.m. The contents in the subsoil were as high as in the surface soil [176].

PCBs are bound in soils as a function of their water solubility and lipophilicity, i.e., sorption increases from the singly chlorinated to the polychlorinated biphenylene. The K_{OC} values rise from 10^3 to 8×10^5 (see Table 18). At low humus contents, the content of clay and allophanes is also important in PCB sorption [288].

Table 18. Physicochemical data of PCBs

Biphenyl	IUPAC, No.	WS, mg/L	K_{OC}*	k_d**
Biphenyl		7.5	1400	
2-Chloro-	6	0.062	4460 – 7940	
4,4′-Dichloro-	15	0.062	20 000 – 100 000	
2,4-Dichloro-	20		151 000	7 – 20
2,4,4-Trichloro-	28	0.113	166 000	
2,2,5,5′-Tetrachloro-	52	0.113	257 000	21
2,2,4,4′-Tetrachloro-		0.085	17 000	
2,2′,3′,4,5-Pentachloro-	87	0.099	490 000	740
2,2′,4,5,5′-Pentachloro-	101	0.029	425 000	810
2,2′,3,4,4′,5-Hexachloro-	138		70 800	2420
2,2′,4,4′,5,5′-Hexachloro-	153	0.001	724 000	2500
2,2′,3,4,4′,5,5-Heptachloro-	180		794 000	4640
2,2,3,4,4,4′,5,6-Heptachloro-	183	0.00047	620 000	3620
2,2′,3,3′,4,5,5′,6-Octachloro-	201			8700

* [217], low clay contents.
** [218], low humus contents.

The degradation of PCBs in soils under aerobic conditions proceeds slowly, and under anaerobic conditions very slowly [219]. Anaerobic degradation occurs as a reductive dehalogenation. After input over many years, the hepta- and octachlorinated compounds accumulate, because of extremely small degradation and more powerful sorption than the low-chlorinated compounds [150]. After 300 d no reduction of PCBs was found in soils treated with sewage sludge [220]. In other field experiments only after application of 50 t sewage sludge per hectare was a five- to seventeenfold enrichment observed, a value of 0.05 mg/kg d.m. not being exceeded [221]. Mono- and dichlorinated PCBs degrade microbially, e.g., by *Achromobacter* or *Pseudomonas spp.* to benzoic acid via the dihydrodiol and *meta* cleavage of the rings.

With increasing chlorination, the degradation proceeds more slowly, especially if a chlorine atom occupies a position *ortho* to the second ring [222]. The volatilization of low-chlorinated PCBs is regarded as a possibility of elimination, especially because of their great persistence [223]. Aroclor 1254 was released from the soil by 40 – 50 % within 2 – 4 months. Hexa- and heptachlorobiphenyls were released only half as well as trichlorobiphenyl.

In case of high PCB pollution and long-term action, PCBs can reach even deep soil layers and the groundwater (distance from ground elevation: 1 – 1.2 m). The migration depends on the solubility of the PCB [224]. If used oils or an oil phase are present in the underground, PCBs can also move laterally. The greater solubility of PCBs in the oil phase than in the dissolved organic matter causes a tenfold enrichment in the oil phase [225].

For many applications (e.g., risk assessment) it is necessary to estimate the behavior of PCBs and other pollutants in soil. For landscape planning, for groundwater protection, for sewage sludge application on farmland, etc. a site-specific prediction system to evaluate the behavior of such pollutants is important for the administration to protect the environment. Results from field trials with more than ten different soil types are in good agreement with estimates obtained for a qualitative evaluation [226].

PCBs are absorbed by plants only in small amounts. Studies with soils treated with sewage sludge containing PCBs showed that carrots in particular, because of their lipoid-containing roots, accumulated PCBs in their skin [227]. The bioconcentration factors (⟶ Ecology and Ecotoxicology – Bioconcentration, Bioaccumulation, and Ecological Magnification) were: in the roots, 0.02 – 0.04; in the skin, 0.02 – 0.1; and into the foliage, <0.002 – 0.0008. The low-chlorinated congeners in particular are absorbed. Small uptakes were also determined in various grasses, grains, and vegetable species [221]. The bioconcentration factors were between 0.2 and 0.02. The action of PCBs in other useful plants was also studied [228]. At PCB concentrations above 10 000 mg/kg d.m., growth inhibition was established in soya and beets. No metabolization and only minor xylene transport was found [229]. Only 20 – 25 % of PCBs absorbed by rats via the soil was not excreted. Different soil properties scarcely affected the excretion.

Within an animal food chain (oak leaves, earthworm, great tit's egg) an accumulation (biomagnification) of 1 ng/g to 170 ng/g of PCB 153 occurred [230].

5.2. Acid Precipitation

5.2.1. *Definitions*

Used in a strict sense, the term acid precipitation is defined as increased hydrogen ion content of precipitation water due to anthropogenic sources [231]. From a chemical point of view this term must be extended, and other rain components included that are subject to acid – base reactions after deposition. As a consequence of these reactions, hydrogen ions may be generated when rain water comes into contact with vegetation and soil surfaces. The term “potential acidity” therefore takes into account all acidic components carried in atmospheric deposition fluxes.

Meteorology distinguishes different forms of atmospheric deposition such as rain, snow, dry particulates, aerosols, gases. Assessment of deposition fluxes and separation of single fluxes are very difficult to accomplish for methodological reasons [232]. Nevertheless, in order to estimate effects of acid precipitation on soils and ecosystems, assessment of all forms of atmospheric deposition is indispensable.

With regard to the impact of acid deposition on soils, it is also necessary to define the terms acidity and acidification for soils, soil water, groundwater, and surface water. *Soil acidity* is usually expressed as the hydrogen ion activity (pH) measured in an aqueous suspension of a soil sample. The pH is a key parameter for most soil-chemical reactions and for the conditions in which soil organisms live and plant roots are growing. However, pH is an intensity parameter which gives no information on the extent to which acid – base reactions can be maintained in a given soil over time. Therefore, when soil changes in time are considered soil acidity must be described using capacity parameters (such as mass of acids and bases available in the surface soil layers).

According to [233], *soil acidification* is related to (1) a decrease in acid neutralizing capacity (ANC, common units mol H^+ per kilogram soil or per square meter of a defined soil horizon) of the soil by loss of base cations such as Ca^{2+}, Mg^{2+}, and K^+, and/or (2) an increase in base neutralizing capacity by retention and accumulation of H^+ and acid cations such as Al [Al^{3+}, $Al(OH)^{2+}$, $Al(OH)_2^+$] and heavy-metal ions. These changes are accompanied with a decrease in soil pH when the acid strength of the acidic groups of the solid phase in the soil increases. Acidification may take place in one or more soil horizons, depending on the balance between acid input (or production) and acid output (or consumption).

In soil water, groundwater, and surface water chemistry, acid – base conditions are usually described in terms of alkalinity. Alkalinity can be defined as the sum of

$$2\left[CO_3^{2-}\right] + \left[HCO_3^-\right] + \left[OH^-\right] - \left[H^+\right] - 3\left[Al^{3+}\right] - 2\left[Al(OH)^{2+}\right] + \left[Al(OH)_2^+\right]$$

where brackets denote molar concentrations of each ion species [234]. Alkalinity defined in this way may take positive or negative values. Water acidification is then defined as a loss in alkalinity.

Thus, the choice of appropriate parameters to characterize acidity of precipitation in relation to its effects on soils and ecosystems depends on the time scale in which the impact of acid deposition is considered. Processes occurring in the soil can be classified into different categories, *direct* and *long-term processes.* The first category of processes is directly and instantaneously related to acid intensity. Examples for such processes are: dissolution of solid mineral phase, ion exchange on mineral particles and humic substances, formation and destruction of ionic complexes and, as a consequence, toxic effects on soil organisms and plant roots. These processes instantaneously follow changes in thermodynamic conditions at any given point of the soil or ecosystem. Here acidity influences soil-chemical processes which may be connected to life processes on the organism or ecosystem level, but must not necessarily have any visible effects on soil components (soil minerals, aggregates, humus). For these processes acidity is appropriately described by the pH of solution. The presence of cation acids and the redox potential may be additional conditions to be considered.

The second category of acidity-related soil processes becomes important in long-term considerations. As a result of this type of processes, changes in composition and structure of soil minerals and soil organic matter, in nutrient availability, and soil biology are observed, followed by changes on the ecosystem level (changes in tree population, plant, and animal

communities). The underlying processes — acid – base and redox reactions — are the same as those described in the first category and are occurring naturally in soils. Acid impact often results in an acceleration and intensification of natural processes such as weathering and alteration of minerals, leaching of nutrient cations, aluminum, iron, and manganese, depletion of the exchange complex, formation of secondary minerals, redistribution of elements in the soil profile, build-up of organic matter at the soil surface, retreat of roots from deeper soil layers, changes in the edaphon (soil animals) and in the herb layer. All these changes tend to be irreversible at least within a human life span. In contrast to the first category, these processes result in visible changes in the structure and composition of soil itself. In order to characterize acid deposition with respect to the described effects, capacity parameters such as deposition, capacity parameters such as deposition fluxes of acidic substances (expressed in ion equivalent per unit surface and time, e.g., meq m^{-2} a^{-1}) are appropriate.

5.2.2. *Sources of Acidity in Atmospheric Deposition*

Most acidic components of atmospheric deposition may originate both from natural and from anthropogenic sources. In order to justify control measures for soil protection and to set priorities, it is necessary to quantify the contribution of single sources of acidity, natural as well as anthropogenic.

Table 19 shows substances which form protons in cloud droplets and precipitation water, or which carry acidity into soil. The substances are listed together with their most important sources. In industrialized countries, carbonic acid (1) and strong mineral acids (2) are generally the most important contributors to precipitation acidity, the acid intensity (pH of precipitations; see below) being determined mainly by (2). In class (2), man-made sources by far exceed natural sources, at least in the industrialized countries of the northern hemisphere [231]. Ammonium (3) plays an important role in areas with intensive agriculture mainly with a high density of livestock. Hydrosulfuric acid (4) and its oxidation products may be important contributors to acidity in bogs, swamps, and on mangrove coasts. Organic and cation acids [substances (5) and (6)] are usually of minor importance in precipitation chemistry.

Table 19. Acidic components of atmospheric deposition and their main sources

Substances	Anthropogenic sources	Natural sources
(1) CO_2 and H_2CO_3	combustion of fossil fuels	atmospheric CO_2, respiration, mineralization of organic matter
(2) Strong mineral acids and their precursors		
SO_2 and H_2SO_4	combustion of fossil fuels, incineration, smelting of ores	volcanic exhalation
NO_x and HNO_3	combustion of gasoline, agriculture (livestock)	electric discharge (lightning), microbial denitrification in soils, forest and grassland fires
HCl_{gas}	incineration of refuse, chemical industry	
(3) NH_4^+, when transformed into ammonia or when taken up by organisms to form organic N-compounds	agriculture, sewage treatment, chemical industry	decomposition of organic N-compounds (proteins)
(4) H_2S	chemical industry	biogenic emissions from marine algae and sulfate-reducing bacteria
(5) Organic acids	chemical industry	
(6) Cation acids (as Al^{3+}, heavy metals)	soil and industrial dust particles when dissolved in acid precipitation	soil particles when dissolved in acid precipitation

The rate of deposition of acidic components varies in a wide range, depending upon local pollution situation, exposition, and surface characteristics. Mean annual desposition rates in Germany were around 270 meq m^{-2} a^{-1} on average, with values between 80 in the southwest, and 1000 in the eastern part of the country in 1993 [237]. About 60 % of the total is made up of sulfuric oxides. As a consequence of technical measures to reduce acid deposition from the atmosphere, deposition rates of sulfur compounds have been gradually reduced in the 1980s. Nitrogen deposition does not show this trend [237], [238].

As pointed out above, when impact of acid deposition on soils is studied, alkaline components of precipitation must be taken into account. The most important basic components are listed in Table 20, together with their sources.

Table 20. Components contributing basicity to atmospheric deposition and their main sources

Substances	Anthropogenic sources	Natural sources
(1) HCO_3^-	see (1), Table 19	see (1), Table 19
(2) NO_3^-, when taken up by organisms to form organic N-compounds	see NO_x in (2), Table 19	see NO_x in (2), Table 19
(3) NH_3, when transformed to ammonium NH_4^+ (see Table 19)	agriculture (livestock)	decomposition of organic matter (proteins)
(4) dust particles from limestone and other carbonates	cement industry, agriculture	soil dust

5.2.3. *Source and Sink Processes for Acidity in the Atmosphere and in the Soil Environment*

For the effect of acid deposition on soils it is unimportant whether the hydrogen ions are already present in the cloud droplet or in precipitation, or are generated by chemical reaction of a pollutant at vegetation surfaces or in the soil. Chemical reactions not only depend on presence and amount of reaction agents, but also on thermodynamic conditions at the site of reaction. Therefore, in order to understand the effects these conditions must be specified. The most important variable in acid – base reactions is the pH at the site of reaction, but other factors such as redox potential or temperature must also be taken into account.

Source processes are listed in Table 21 together with the range of physicochemical conditions — in a generalized form — under which they take place. Under these conditions atmospheric substances undergo chemical transformation (indicated by an arrow), generating hydrogen ions in the atmosphere (cloud droplets, precipitation water), at the contact of rain with surfaces (vegetation, soil), or in soil water.

Table 21. Hydrogen ions generation processes in atmosphere and in soil

Process	Conditions
(1) Formation of carbonic acid from carbon dioxide originating	
in atmosphere: from atmospheric CO_2	
in soil: from soil respiration	
$CO_2 + H_2O \longrightarrow H_2CO_3 \longrightarrow H^+ + HCO_3^-$	pH >5.0
(2) Formation of strong mineral acids	
$SO_2 + 1/2\ O_2 \longrightarrow SO_3 \longrightarrow H_2SO_4 \longrightarrow 2\ H^+ + SO_4^{2-}$	independent of pH; oxidizing conditions with an oxygen source (free atmosphere)
$NO_2 + H_2O \longrightarrow HNO_3 \longrightarrow H^+ + NO_3^-$	independent of pH
$HCl_{gas} \longrightarrow H^+ + Cl^-$	independent of pH
$H_2S + 3/2\ O_2 \longrightarrow H_2O + SO_2$ (see Table 19)	oxidizing conditions
nitrification $N_{org} \longrightarrow HNO_3$ (see Table 19)	oxidizing conditions
$NH_4^+ \longrightarrow HNO_3$ (see Table 19)	
(4) Formation of organic acids	presence of organic matter and soil microorganisms
(5) Hydrolysis/protonation of weak metal hydroxides, e.g., $Al^{3+} + H_2O \longrightarrow Al(OH)^{2+} + H^+$	pH >4

(6) Formation of biomass by uptake of nutrient cations (most important: ammonium)	

Finally, Table 22 gives a list of *sink processes* which consume hydrogen ions, thus neutralizing acidity.

Table 22. Hydrogen ion consumption processes in atmosphere and in soil

Process	Conditions
(1) Transformation of bicarbonate into carbonic acid $HCO_3^- + H^+ \longrightarrow H_2CO_3$	pH <5
(2) Dissolution of carbonates $CaCO_3 + H^+ \longrightarrow Ca^{2+} + HCO_3^-$	
(3) Weathering of silicates	
(4) H^+ exchange on cell walls of leaf surface and on the exchange complex of the soil against (base) cations	
(5) Weathering of clay minerals	
(6) Formation of biomass by uptake of nutrient anions (most important: nitrate)	
(7) Transformation of ammonia into ammonium $NH_3 + H^+ \longrightarrow NH_4^+$	pH >6.5

5.2.4. *Direct Effects of Acid Deposition on Soils*

Traditionally, direct effects are studied separately for different environment compartments, such as vegetation cover, soils, subsoils, groundwater bodies, lakes, buildings, etc. In landscapes subject to acidic deposition, chemical reactions occurring at any point are interrelated, mostly by fluxes of water and dissolved matter, but also by transfer in the solid and gaseous phase, or by organisms. The ecosystem approach helps to find the linkage between individual processes, and to close the gap of understanding between chemical characterization of acid rain and damage to the environment.

When precipitation reaches the soil, acids or acidifying components usually undergo chemical reactions with the soil matrix which is not in a (chemical) equilibrium with rain water. The kind of reaction and the effect of acid input depend very much on type of soil and underlying bedrock. It has been suggested to group soil types according to criteria which reflect the type of buffering system in effect, and its buffering capacity [235].

In soils on calcareous bedrocks that are not totally decalcified and contain rather unweathered lime particles in the surface horizons, acidity is readily neutralized by weathering of carbonates, resulting in the release and leaching of Ca^{2+} with seepage water to the subsoil, together with bicarbonate [reaction (2) in Table 22]. This reaction is usually fast enough to buffer even high atmospheric input of acids. Only if limestone is distributed very disproportionate over the soil volume, even this soil type may acidify in the fine fraction of upper soil layers.

In soils free of limestone the weathering of primary silicates leads to the release of base cations such as Na^+, K^+, Mg^{2+} and Ca^{2+}, which then buffer the protons [reaction (3) in Table 22]. The consequence is again cation loss by leaching, soil is depleted from essential plant nutrients. The rate of silicate weathering depends on contents, type, and grain size of silicates in the mineral fraction of soil. The rate is estimated to range between 20 (soils on sandy sediments, sandstone, quartzitic bedrocks) and 200 (soils on basaltic rock) meq m^{-2} a^{-1} for a soil depth of 1 m [239]. These figures represent the maximum proton consumption rate. In comparison, the acid load from atmospheric deposition is 50 to 600 meq $m^{-2}a^{-1}$ [240].

In cases where acid deposition exceeds silicate weathering rate, or when silicate content in mineral soil is low or zero (sandy soils, soils on acid loess loam, quartzitic bedrocks), protons are neutralized by displacement of exchangeably bound Ca^{2+} and Mg^{2+} from the surfaces of clay minerals, iron and manganese oxides, and humic matter [reaction (4) in Table 22]. As a consequence, soil suffers further leaching losses of plant nutrients. The total ANC of the exchange complex in the rooting zone of vegetation ranges between 0.7 (acid sandy soils) and 70 eq/m^2 (well-buffered clay soils). Comparing these figures with total acid input during a 10-year period of 0.5 – 6 eq/m^2 (see above) it is obvious that acid input may lead to a complete depletion of the exchange complex on less-buffered soils within a short period of time. This depletion is reversible after a decrease in acid load from atmospheric input, but at low weathering rates (see above) complete recovery may require decades or centuries.

When the ANC decreases because of leaching of buffering substances to deeper soil layers, protons begin to attack the crystal lattice of clay minerals at pH <4.2 [reaction (5) in Table 22]. The consumption of protons is then balanced by release of cation acids (mainly monomeric or polymeric Al ions) from the clay minerals which, in turn, tend to displace base cations from exchange sites. Saturation of the exchange complex with Al ions is followed by an increase of Al and heavy metal ions in soil solution. This process is ecologically harmful because Al and most heavy metal ions are toxic in their ionic form for plant roots and soil microorganisms [241], and acidification of subsoil and groundwater is promoted with leaching of these cation acids to deeper soil layers.

When the exchange sites are saturated with Al ions, the concentration of Al ions in soil solution is mainly determined by the concentration of sulfate, which is usually the dominant anion in acid soils in regions subject to acid precipitation. Under these conditions aluminum hydroxosulfates are formed and accumulate in acidified soil horizons [242]. These sulfates have to be considered as a quite labile pool of cation acids, which is readily mobilized under changing chemical conditions in the soil (lowering of pH), and then may be carried to deeper soil layers or groundwater with seepage water.

5.2.5. *Effects Mediated by Organisms (Indirect Effects)*

Indirect effects are mediated by organisms, mainly through plant roots and microorganisms. In such cases, atmospheric input of acidic components initiates processes which produce acidity and transmit it to the soil. This gives rise to further soil acidification with reactions described in Section Direct Effects of Acid Deposition on Soils.

5.2.5.1. *Soil Acidification by Canopy Leaching*

When acid precipitation hits vegetation, hydrogen ions may be buffered on leaf surfaces by exchange with Ca^{2+} or Mg^{2+} bound to acidic groups in the cell walls (reaction (4) in Table 22) [243], [244]. The basicity used up for neutralization of atmospheric acidity must be replenished from the exchangeable pool of Ca^{2+} and Mg^{2+} in the soil. Consequently, soil is acidified by this process and its ANC decreases. Such effects are predominantly observed on slightly acid soils in the main rooting zone of the trees (B horizon). Surface soil (A horizon) and organic surface layer (O horizon), which receive base cations from canopy leaching and litterfall, remain in a state of high base saturation. This shows how acid deposition reaches soil via root surfaces. Buffering in the canopy does not prevent the soil being acidified in the rhizosphere within very short time.

5.2.5.2. *Acidification by Uptake of Nitrogen Compounds*

Atmospheric deposition of nitrogen compounds occurs mainly in the form of nitrate and ammonia. The plants usually take up these compounds wholly or partially, nitrogen being one of the most important plant nutrients. The uptake of nitrate by organisms and transformation into organic molecules results in the equivalent consumption of protons (reaction (2) in Table 20).

In contrast, if ammonium (NH_4^+) is taken up to form organic nitrogen compounds, an equivalent amount of protons is generated [Table 19, reaction (3)]. If nitrate and ammonium are taken up in equivalent amounts, consumption and production of protons are balanced and no net acidification takes place. But when ammonium is deposited in excess, as in the vicinity of intensive agriculture (livestock), and is taken up by plants or transformed to nitrate and leached, this results in a considerable acidification of the soil horizon in which ammonium uptake is located.

5.2.5.3. *Acidification by Accumulation of Organic Matter*

The emission of nitrogen oxides (from industrial processes and combustion of oil and gasoline) and of ammonia (mainly from intensive agriculture/livestock) leads to considerable deposition of nitrogen compounds on large areas (20 – >50 kg N ha^{-1} a^{-1}). In many forests on poor soils, nitrogen would represent a minimum factor in the nutritional supply of the trees without this input. The consequence of nitrogen deposition is an enhancement of tree growth and accumulation of biomass and litter at the soil surface. During the buildup phase of this pool of biomass, these processes result in a depletion of the mineral soil from nutrients (Ca, Mg, K), and a decrease of buffering capacity in the root zone, connected with a temporary or permanent acidification of the root zone.

Another effect of soil acidification is the inhibition of mineralization of litter and other forms of dead organic matter. In forests, an organic layer is built up at the soil surface, in which the activities of soil animals are reduced, and the microorganism population shifts from a dominance of bacteria to one of fungi. The mineral horizons underlying the organic surface layer are subject to intensified podzolization, with leaching of base cations, low-molecular organic matter and metal ions (Al, Fe, Mn) in complex form. In shallow soils, this can result in

groundwater and surface water acidification.

5.2.5.4. *Acidification by Humus Disintegration*

Forests on well buffered, noncalcareous soils, e.g., on basaltic rocks under deciduous forests in favorable climatic conditions, usually show a soil profile with a deeply developed A horizon very rich in humus. The amounts of nitrogen stored in the organic fraction of such soils may be considerable. Under the conditions of high acid input, these stores of organic matter tend to disintegrate, i.e., the buildup of new humic substances is delayed (possibly through inhibition by Al ions) compared with the mineralization through accelerated degradation of high polymeric humic substances [245]. As a consequence of humus disintegration, the organically bound stores of nitrogen are transformed into nitrate (nitrification by soil bacteria). If the nitrates thus formed are not taken up by roots or soil organisms, but are leached with seepage water, a considerable acidification of the soil according to reaction (2), Table 21, results. This effect may be estimated by keeping in mind that a loss of 1000 kg N by nitrification and leaching would be equivalent to 7 eq/m^2. The same process, however, may also lead to the release of gaseous nitrous oxides, mainly NO, into the atmosphere [246], hereby not only avoiding the release of protons to soil solution, but returning to the atmosphere a gaseous component which is acidic (Table 19) and is relevant with respect to the "greenhouse effect". This example elucidates the complexity of acid-base-processes in the soil environment which can only be understood as part of functional relationships between ecosystems. It makes the distinction between direct and indirect effects to be rather artificial one.

5.2.6. *Consequences of Soil Acidification*

The most important consequences of soil acidification for structure and function of soil itself, for the ecosystem of which soil is part, and for adjacent systems such as groundwater, lakes, and rivers are summarized below:

1. Loss of plant nutrients by enhanced leaching of basic cations which function as buffering against acidity — as secondary effects disturbance in the nutrition of trees and other plants, destabilization of forest ecosystems
2. Mobilization of cation acids, mainly Al and heavy metal ions from native soil minerals and/or atmospheric pollutants accumulated in the upper soil layers — as secondary effects toxicity to fine roots of trees and other forest vegetation, to soil animals and microorganisms, destabilization of forest ecosystems
3. Acidification of subsoil, underlying bedrock, groundwater and surface water, and concentration increase of undesired or toxic ions (nitrate, Al and heavy-metal ions) in the water
4. Formation of dissolved organic matter [247] which may interfere in many ways with inorganic components, thus making the judgement of acdifying effects upon hydrosphere very difficult.

5.2.7. *Field Observations of Soil Acidification*

Since the beginning of industrialization of Western Europe, and more intensively from the 1950s onward, soil acidification has increased progressively. This is indicated by a number of studies, refering to different interrelated acidification parameters (see Sections Definitions, Sources of Acidity in Atmospheric Deposition, Source and Sink Processes for Acidity in the Atmosphere and in the Soil Environment, Direct Effects of Acid Deposition on Soils, Effects Mediated by Organisms (Indirect Effects), Consequences of Soil Acidification).

Decrease of Soil pH. The earliest widespread pH measurements in forest soils were made in the late 1920s. These data were reexamined in 1978 [248]. Comparison with new measurements suggests that strong soil acidification has occurred in areas close to industrial zones and on mountain ridges with high acid atmospheric deposition. Soil acidification in soils under new planted spruce stands is attributed to ecosystem-internal acidification. pH measurements were repeated in a forest soil inventory in Bavaria after 20 – 30 a [249]. Depending on the initial acidity status, a decrease in soil pH of 0.3 to 1.8 units was observed on almost 50 % of 1000 cases. A review of soil pH changes found in a number of European countries is given in [250]. BERDEN et al. 1987 critically reviewed pH and base saturation data from Scandinavian and Central European forest soil samples since the 1940s, resampled after 20 to 50 years [251]. The drop in pH after this time span decreases, as expected, with decreasing pH at the time of first sampling. Similar results are reported for North Rhine-Westphalia, which suggest acid deposition as cause for soil acidification in the 1960s –1980s [252]. A good correlation was found between acidification parameters in soils and exposition of the site resulting in very different atmospheric deposition rates. Further evidence for soil pH decrease under forest vegetation is given for the Berlin region [253], mountain areas of North Rhine-Westphalia [254] Hesse [255], and the Hamburg region [256].

Decrease in Exchangeable Bases (Ca and Mg) and Base Saturation. In forest soils in North Rhine-Westphalia the exchangeable Ca has decreased by about 50 % since the 1950s [257]. In 1987 an average of 50 % decrease in base cation stores was found in Swedish forest soils since the 1970s to 1950s [258]. Similar results were found for the Harz [259], the Oberpfalz [260], [261], and South Sweden [262] and in a large scale survey of forest ecosystems in Western Germany [263] . A large number of field studies showed increased soil acidification around the trunks of beech trees, where large amounts of stemflow with high acid load infiltrates the soil [264-267]. Beech stands in remote areas of Southern Europe that are little influenced by acid deposition do not show this pattern [268].

Increase of Protons and Cation Acids at the Exchange Complex [269], [270]. For south Swedish forest soils an increase in the pool of exchangeable Al of 100 % was shown [262].

The appearance of cationic acids, mainly ionic species of Al, in groundwaters and surface waters as a consequence of soil acidification was reported in many studies, for example, in forest-covered Central German mountain areas such as Kaufunger Wald [271], Harz [259], [272], Hesse [273], Fichtelgebirge [274]. For a review, see [275], [276]. Distinction between naturally occurring acidification of ground- and surface water and acidification caused by atmospheric deposition is usually made by using sulfate concentration as indicator (mobile anion concept, see [234]). When sulfate becomes the dominant anion in solution, the cause for acidification is attributed to atmospheric input, provided the underlying bedrock does not contain sulfate or sulfide minerals. In the mountains of the Harz (Lower Saxony), rivers and creeks of different acidification types exist in close vicinity, depending on acid load from atmospheric deposition and types of soil and bedrock [277], [278].

Reversibility of Soil Changes. The undesired effects of soil acidification are traditionally avoided in agriculture by liming. Due to the progressive acidification of forest soils in the last decades, liming has become common practice in forestry on soils with high atmospheric acid input and low base saturation [279]. This is done in order to improve forest nutrition and nutrient turnover as well as to protect hydrosphere from the input of potentially toxic ions with seepage water. On the other hand, great efforts have been made to reduce acidity in the atmosphere by legislative measures, mainly directed towards control of sulfur and nitrogen emissions from combustion of fossil fuels and refuse. Related to ecosystem research, investigation of the ability of ecosystems to recover from the negative effects has become an important topic. This found its expression in "clean rain roof experiments" in different countries, in which forest ecosystems have been excluded from acid deposition impact, observing the change in acidity parameters and recovery of plant and animal communities [280], [281].

5.3. Fertilizers and Agricultural Chemicals

5.3.1. *Fertilizers*

5.3.1.1. *Development of Fertilization*

The development of fertilization in the 20th century is marked by an increased production of manures and a large rise in the consumption of inorganic fertilizers. From the beginning of the century until the mid-1980s, the use of nutrients in Germany in agriculture developed as follows: Inorganic fertilizers increased from a few kg/ha to an average of 130 kg N, 65 kg P_2O_5 and 90 kg K_2O per hectare of agricultural areas. In the 1980s the annual consumption of manures of animal origin was 100 kg N, 60 kg P_2O_5, and 110 kg K_2O per hectare [282]. In Germany, for example, from 1985 to 1997 the annual consumption of inorganic fertilizers fell to 101 kg/ha N, 24 kg/ha P_2O_5, and 37 kg/ha K_2O [283]. Similar results are reported for The Netherlands and the United Kingdom, whereas a much lower reduction is found in southern Europe. The major causes of this decline in Germany were the reorganization of the agricultural market ordinance (reduction of selling prices to world market level, payment of land grants), structural change, and the increasingly selective use of manures. These and other causes (limitation of the amounts of fertilizers by legislation, smaller nutrient losses and greater utilization of the fertilizers, higher fertilizer prices in combination with decreasing selling prices of agricultural products, increase of ecological farming with abandonment of the use of fast-acting inorganic fertilizers, declaration of water protection and conservation areas) will have the effect that the trend, of a declining use of inorganic fertilizers with simultaneously more selective use of the manures will also continue, in weakened form, in future.

Because of the use of composts and of secondary raw material fertilizers (e.g., sewage sludge), an additional input of nutrient into soils will occur in future. While the proponents consider this a possibility of closing the nutrient cycle of nature to the greatest possible extent and of dispensing with the feeding of inorganic fertilizers, critics fear a contamination of the soil by heavy metals and organic pollutants.

5.3.1.2. *Legal Framework of Fertilization*

In many countries, especially in the EU, fertilization is governed by various laws and regulations. First of all users have to prevent damage to the environment as a result of improper use of fertilizer. But during the 1990s the regulation of fertilization was discussed in another context: To what extent can the possibility of lasting cultivation of the soil and land be guaranteed as it is demanded by the AGENDA 21? In terms of long-term effects fertilization can have favorable as well as negative effects, therefore it is the responsibility of government and science to come to balanced conclusions. In some countries support and subsidies are provided for farmers to comply with special restrictions of fertilization.

In Germany all laws and regulations affecting fertilization were amended or reenacted during recent years. This fact reflects the increased attention of society to the application of substances to soil. As an example the essential contents and objectives of the German laws and regulations are described below.

The fertilizer law (Düngemittelgesetz, DümiG of November 15, 1977, last amended by law on December 17, 1999) defines fertilizers and thereby lays down what substances may be applied to agricultural lands. Beyond that the law governs the application of fertilizers and defines the expression "fertilization according to good technical practice" as follows: "The fertilization must be directed, as regards type, quantity and time, to the needs of the plants and of the soil, taking into consideration the nutrients and organic matter available in the soil as well as the site and cultivation conditions; the nutrient needs of the plants in consideration of their productivity under the prevailing site and cultivation conditions and to the standards of quality of the produce".

Further details are regulated by a fertilizer decree (Düngeverordnung of January 26, 1996). Aside from restrictions and prohibitive rules on the application of substances, it is intended to achieve and guarantee a selective fertilization that avoids any overprovision of nutrients by obliging farmers to make soil analyses and written records and checks of fertilizer use. From the ecological point of view the fertilization decree is help up as an example for regulations in other fields of agricultural production.

The recycling and waste law (⟶ Waste) can stipulate prohibitive rules and restrictions concerning contaminants for the application of manures and secondary raw material fertilizers.

The German soil protection law (Bundesbodenschutzgesetz, BBodSchG), which is in draft form, is intended to protect the soil from damages. The landowner is forced to practice precautionary soil protection, which naturally applies also to the application of materials such as fertilizers.

5.3.1.3. *Forms of Fertilizers*

(⟶ Fertilizers)

The fertilizers applied can be divided into inorganic and organic fertilizers.

Inorganic fertilizers contain inorganic nutrients or release them after a reaction. Since the nutrients are frequently water-soluble salts or salt mixtures, they are usually available to plants immediately after application or their period of availability can at least be calculated. However, adverse effects can occur when these fertilizers are not properly used; in particular, too high a supply of readily soluble nutrients can lead to increased nutrient loads of the soils and nutrient losses, in addition to reductions in crop yield and quality.

Organic Fertilizers. Two types of organic fertilizers exist, (1) *commercial organic fertilizers*, which are produced by the processing of peat (⟶ Peat – Peat Processing Technologies) or of vegetable and animal materials (⟶ Fertilizers – Organic Fertilizers (Secondary Raw Material Fertilizers)) and are important in the backyard sector at most, and the *organic manures* whose origin is animal (liquid manure, slurry, solid manure) (⟶ Fertilizers – Manure). The nutrients of organic fertilizers are organically combined to different degrees, depending on the type of fertilizer and must the mineralized in the soil in order to be taken up by the plants. Since mineralization processes are highly dependent on temperature and soil moisture, the nutrients are not always available specifically and the amounts taken up by the plants cannot be calculated. The feeding of crops via organic fertilizers therefore makes great demands on the farmer, especially because the nutrient contents of organic fertilizers can vary within wide limits and the precision of application is not optimum.

With the amendment to the German Fertilizer Law of September 27, 1994, *sewage sludge and other wastes* also are for the first time assigned to the fertilizer sector as “secondary raw material fertilizers.”

5.3.1.4. *Effects of Fertilizers in the Soil*

Fertilization pursues several goals. On the one hand, the real task of fertilization — feeding the crop plants with the aim of improving the harvest in quantity and quality — must be carried out; and on the other the soil's fertility must not be threatened. Fertilization can affect various parameters of the soil and therefore the soil's fertility. Strongly influenced are chemical characteristics of the soils (nutrient concentration, soil reaction, pollutants and heavy-metals content), followed by biological parameters (humus content, soil edaphon). Soil-physical parameters are relatively little affected by fertilization.

5.3.1.4.1. Nutrient Contents and Nutrient Losses

(⟶ Fertilizers – Soil Science)

Fertilization should supply the soil with plant nutrients without enriching it with nutrients beyond a tolerable extent. The tolerability in this connection corresponds less to the concentration above which toxicity symptoms appear than to the contents above which increased nutrient discharges occur from the soil to groundwater, surface waters, and the air. If soil pollution by nutrient enrichment can be spoken of at all, it is by possible environmental loads as a result of losses from the soils. In addition, disproportionately high concentrations of certain nutrients in particular impair the composition of the plants via the nutrient uptake and reduce yield and quality.

Nitrogen. More than 90 % of the nitrogen present in soils is combined in the organic matter. Depending on the depth of topsoil and humus content, between 2000 and 10 000 kg nitrogen per hectare are stored under Central European conditions. This nitrogen is mineralized by microbial conversion processes and released as ammonium (NH_4^+); at an average annual mineralization rate of 2 % about 40 –200 kg of nitrogen per hectare and year are supplied. The NH_4^+ formed can be bound by the exchange complex in the soil, but in aerated soils usually undergoes nitrification by bacteria to nitrate (NO_3^-). Nitrate is not adsorbed in the soil, is thus freely mobile, and is the most important nutrient form of N for plants. High nitrate concentrations in the soil solution are advantageous if they are provided specifically in the period of uptake by the plants and can be largely utilized by these plants.

In the absence of uptake by plants or if the nitrate supply exceeds the uptake capacity of the plants, NO_3^- is displaced from the root space as a result of the movement of water in and out of the soil. A surface runoff or discharge via drainage water contributes to the eutrophication of surface waters and the seas; increased contents of nutrients (N in particular) bring about increased algae growth which results in decreasing oxygen contents in the water. A far greater part in public discussion is played by the washing-out of nitrate from the soil and its translocation into the groundwater, which can greatly impair the drinking water quality. Both during the nitrification and especially during the denitrification in the soil nitrous oxide as an intermediate product of these processes can escape and contribute to the greenhouse effect [284].

Before the influence of fertilization on the nitrogen content and nitrogen dynamics in the soil as well as on potential nitrogen losses is described, it has to be pointed out that the agricultural use itself already decisively interfers with the natural nitrogen cycle. Because of the complexity of the effects of individual variables, the influence of fertilization on nitrogen loads in soil will not be exactly quantified in the following.

In long-term nitrogen fertilization the amount of combined nitrogen in the soil is influenced both by the level of fertilization and by the form of fertilizer chosen. Compared with similar, but unfertilized soils, soils exploited agriculturally have higher humus contents with greater depths of topsoil and higher nitrogen contents. The extent of the accumulation of organic matter and therefore also of organically combined nitrogen as a result of long-term fertilization is very strongly influenced by temperature and rainfall [285]. Not only the productivity but particularly the reliability of yield of soils, which is independent of current fertilization, can be improved by raising the N contents and the additional N supply. With the exception of crops that react to higher N supplies from the soil with a decrease especially in processing quality (brewing barley, sugar beet, etc.), the positive aspect of the feeding of crop plants predominates. Moreover, N fertilization positively affects the soil biota and the microbial biomass in the soil [286]. The nitrogen-fixing microorganisms whose activity falls with increasing nitrogen supply are an exception [287].

In addition to the level of N compounds supplied, the type of fertilizer especially influences the amount of nitrogen stored in the soil. Since organic nitrogen fertilizers already contain the nitrogen in the storage form found in the soil, they can raise the N contents in the soil more rapidly and — via higher reproduction rates of the humus in the soil — also more effectively than inorganic nitrogen fertilizers [288]. Long-term fertilization with liquid manure improves the humus and nitrogen contents by about the same order of magnitude as an equivalent dosage of farmyard manure [289]. The chain of action following inorganic fertilization — more plant growth bringing about a greater additional supply of organic matter — also can have a positive influence on the nitrogen concentration in the soil [290], [291]. A combination of farmyard manure and inorganic fertilization can increase the N content more than farmyard manure alone [285]. Hydrolyzable organic N compounds give the most important contribution to the higher total N contents [292].

With increasing additional nitrogen supply from the organic matter or with increased direct N fertilization, there is an increased risk of nitrogen losses from the soil which exceed an unavoidable basic emission [293], [294]. The factor usually responsible for increased nitrate discharges is a discrepancy between the amount and the period of time of the NO_3^- supply in the soil and the amount extracted by the plants during the same period of time. In this connection, the form of the applied nitrogen plays a minor part: thus, even with ecologically cultivated lands, high nitrate translocations are observed under certain conditions [295]. The limit of the N fertilization level above which environmental loads occur depends on many factors (soil, weather, rotation of crops, etc.) and cannot be defined in general. At present the nitrate content in the soil is determined after the harvest of a crop at the start of the translocation period in the autumn and used as a criterion for a balanced NO_3^- supply during the period of uptake by plants [296]. Nitrogen balances also can give clues to a possible risk potential from a N fertilization measure. However, they do not permit direct conclusions on the nitrate discharge [297], [298]. Beyond that, computer simulation models of the nitrogen cycle are being developed that shall reproduce the processes of N dynamics in the soil [299] and thus can provide concrete informations for an N fertilization that accords with requirements and is environmentally acceptable. Even though, the models are not yet suitable for practical use at the present time [300] because of the large number of input parameters, they will become important in future.

The influence of N fertilization on losses in the form of nitrous oxide cannot be estimated at present. All fertilization measures that trigger nitrification and denitrification processes in the soil contribute to nitrous oxide formation in the soil. Even if N_2O is released in any conversion of organic matter (also in the absence of fertilization) [301], it can be assumed that these losses will be greater after N fertilization [302]. Organic and ammonia-containing fertilizers increase N_2O production at a higher extent than nitrate fertilizers [303]; in all, the N_2O emission level probably depends less on the form of fertilizer than on its amount.

Phosphates. Fertilizer phosphates are absorbed in the soil, especially by iron oxides, and thereby fixed. Because of their low mobility, phosphates accumulate mainly in the topsoil. With rising P fertilization the translocation into the subsoil increases, however, losses from the root zone can be observed only under special conditions (peat soils, [304]). Phosphates from organic fertilizers are more mobile in the soil and, compared with phosphates from inorganic fertilizers, can be translocated into deeper soil layers. However, the phosphate contents in groundwaters near the surface is not increased [282], [305]. The removal of phosphates from the surface soil by erosion and drainage water can more likely be of ecological relevance.

Phosphate fertilization raises the P contents in the topsoil, as is shown by comparison with nonfertilized forest sites. If the P concentrations necessary for plant nutrition are not substantially exceeded no harmful effects in the soil from P fertilization are to be expected.

Potassium. Fertilization with potassium can raise the potassium contents, depending on the buffering capacity of the soil. A disproportionately high K supply can hinder the magnesium uptake by plants and so induce a Mg deficiency [306]. In isolated cases, too high potassium concentrations which, however, occur in practice only as the result of a long-term excessive potash fertilization, can impair the quality of crop products. Avoidable losses from the soil as a result of fertilization measures can occur on light soils [294], [307], but no ecological relevance is known.

5.3.1.4.2. Soil Reaction

Agricultural production is possible only on soils with a soil reaction that is optimum for crop plants. This optimum reaction depends on the kind of plant, the nutrient contents, the kind of soil, and the soil struture. From the plant nutrition standpoint, the optimum pH values are considered to be 6 – 7. In this range most microorganisms [286] and the soil edaphon also show their highest activity. As a result of the natural acidification of the soil by translocation of alkali metal and alkaline-earth metal ions from the soil, which is further reinforced by agricultural production, it is usually necessary to raise the pH value.

Fertilization measures can decisively influence the extent of soil acidification. Ammonium-containing fertilizers, which are nitrified in the soil and in the course of this release H^+, have a physiologically acidifying effect. Since inorganic N fertilization is carried out worldwide mainly using NH_4^+ containing fertilizer, N fertilization contributes considerably to the lowering of the pH value and involves an increased lime requirement. Phosphate-containing fertilizers that are obtained by solubilizing apatites or are combined with Ca in the course of production can promote the soil reaction. Organic manures show no short-term effect on the soil reaction but in the long term can lead to an increase of the pH value by replacement of cations [308].

Through liming, the natural soil acidification and the lowering of the pH value as a result of fertilization measures and acidic inputs from the air can be compensated. This is true not only for agricultural but also, e.g., for forestry lands.

5.3.1.4.3. Humus Content

Humus influences the chemical, physical, and biological properties of soils and is thus an important factor in soil fertility. The humus content is essentially determined by quantities typical of the site, such as nature of soil and climate as well as the conditions for conversion of the organic matter and the possibilities of formation of durable humus materials. Within the limits set hereby, the fertilization can influence the humus content.

Obviously, the humus content of soils can best be influenced by organic fertilizers. Long-term application of solid or liquid manure improves the humus content for a long time afterward. Similar effects are produced by compost and sewage sludge. The humification rates in this connection decrease with increasing supply of organic matter and increasing humus content of the soil [318]. Inorganic fertilization can improve the humus content to a limited extent by way of an increase in the vegetation cover and the resulting crop residues (straw, foliage, green manure) or the humus content can at least be maintained with regular working-in [310]. Only in crop rotations with high proportions of root crops, crop and root residues even in increased amounts cannot prevent a decrease of the humus content and fertilization in organic form is necessary [311].

Very close relations exist between the organic matter content of the soil and its nitrogen pool and nitrogen supply capacity. It is today very questionable from the ecological standpoint to improve the humus content typical of the site by intensive fertilization measures, because with this the risk of N losses detrimental to the environment increases.

5.3.1.4.4. Heavy Metals and Organic Pollutants

Even though heavy metals occur naturally in all soils, a possible accumulation in the soil as a result of fertilization measures must be avoided. The ecological relevance of an accumulation in that connection consists not so much in increased emissions but rather in an increased uptake of heavy metals by plants with the risk of an impact on the health of humans and animals. Similar effects are to be exptected from an application of toxic organic pollutants.

Inorganic fertilizers and farm-produced manures usually contain no organic pollutants and only occasionally small amounts of heavy metals that are not hazardous to the soil. Phosphate fertilizers, which are made from crude phosphates of organogenic origin, can contain considerable amounts of cadmium. By avoiding the use of highly contaminated crude phosphates, the Cd content of P fertilizers has been reduced and currently is usually not more than 40 g/t P_2O_5; hence only an extremely small Cd enrichment occurs in the soil [312]. In the EU countries limit values have been drawn up for further reducing the undesirable Cd intake. The level of Cd intake can in addition be influenced by the form of fertilizer chosen [313].

Farm-produced manures also can contain heavy metals; high Cu and Zn contents in pig manure originating from the pig feed considerably exceed uptake by plants and these metals can accumulate in the soil over the long term. The risk of a possible load has been reduced as a result of the restriction of fertilization by liquid manure (see Section Legal Framework of Fertilization) and to the restriction of copper feed supplements by the German law on feed materials (Futtermittelgesetz).

With the parity of legal treatment between sewage sludge and farm-produced manures, a possible load of soils by “fertilization” with sewage sludge could occur in the future. The present stage of the discussion is given below:

1. If disposal areas are limited, there is a high sociopolitical pressure to use sewage sludge alternatively on agricultural lands
2. From the scientific standpoint, the pollutant tolerance levels of the soil and the permissible pollutant contents are controversal subjects of discussion and in some cases called in question
3. In agricultural practice the reservation against the use of sewage sludge prevail, despite considerable financial incentives

Sewage sludges have a relatively small fertilizing and soil-improving effect and are not necessary as fertilizers from the plant nutrition standpoint. Possible soil loads from the application of sewage sludge have to be justified not by agriculture but by society and the political world in their desire for waste disposal using agricultural areas.

5.3.2. ***Pesticides (*** → Crop Protection; → Acaricides; → Fungicides, Agricultural; → Insect Control; → Molluscicides; → Nematicides; → Rodenticides; → Weed Control***)***

5.3.2.1. *Introduction*

Pesticides are in general formulations of active substances and formulation aids. *Active substances* are chemically defined substances or mixtures of substances and are the actual carriers of the plant-protecting action. The *formulation aids* (e.g., solvents and extenders, deposit builders, emulsifiers, wetting agents, stabilizers, and warning dyes) provide the mixtures with the physical properties which make them suitable for application as sprays, dusting agent, etc. in solid, liquid, or gaseous form. Pesticides arrive, either directly or via the crop plants, onto and into the soil or in groundwater and surface water.

Pesticides are used in agriculture, forestry, industry, and in urban areas. They are frequently used for medical and military purposes. Agriculture forms the largest sales outlet for pesticides. Thousands of tonnes are applied worldwide each year, including 110 000 t in the Federal Republic of Germany alone in 1981 [314-316], the use of herbicides and insecticides being predominant internationally. Various pesticides and their chemical structures are given below. They have been selected mainly from chemical considerations and less from their economic importance or biological action (I = insecticide, H = herbicide, F = fungicide, A = acaricide).

Alicyclic insecticides

Endosulfan

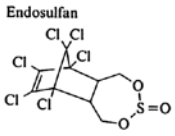

1

Aldrin

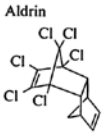

2

Dieldrin

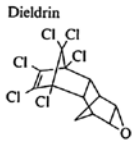

3

Heptachlor

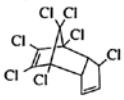

4

cis-Chlordane

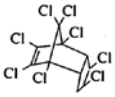

5

Lindane

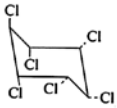

6

Anilides

Propanil (H)

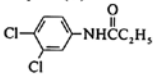

7

Metolachlor (H)

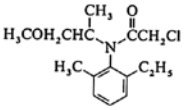

8

Homocyclic herbicides

2,3,6-TBA

9

Dinoseb

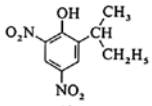

10

DNOC

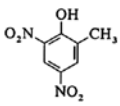

11

Pentachlorophenol

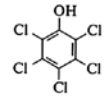

12

Nitralin

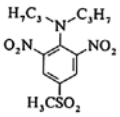

13

Trifluralin

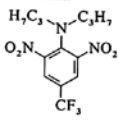

14

Heterocyclic herbicides

Paraquat

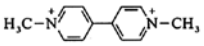

15

Prometryn

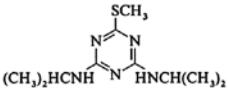

16

Atrazine

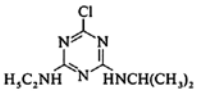

17

Simazine

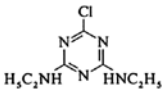

18

Picloram

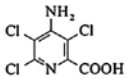

19

Methyl and phenyl carbamates

Triadimenol

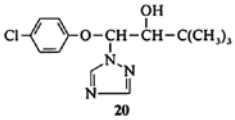

20

Carbaryl (I)

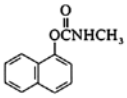

21

Carbofuran (I)

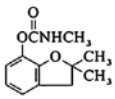

22

Aldicarb (I)

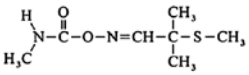

23

Thiophanatemethyl (F)

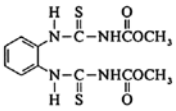

24

Carbendazim (F)

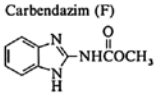

25

Captan (F)

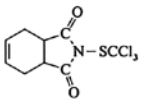

26

Organophosphorus compounds

Malathion (I)

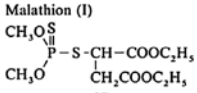

27

Diazinon (I)

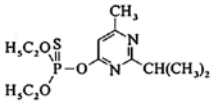

28

Derivatives of phenoxyacetic acid

2,4-D

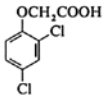

29

MCPA

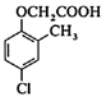

30

2,4,5-T

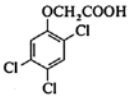

31

Substituted ureas

Monuron

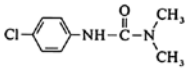

32

Diuron

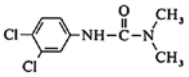

33

Monolinuron

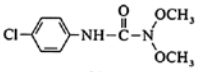

34

Methabenzthiazuron

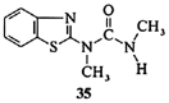

35

Linuron

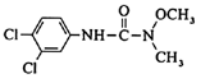

36

The utilization and the toxicity of pesticides is decisively influenced by their behavior in the environment. They can be adsorbed/desorbed, vaporized, leached out, biodegraded and changed photochemically [314], [316], [317]. These processes are clearly influenced by abiotic and biotic factors. Important factors are physicochemical properties of the pesticides such as solubility, volatility, mobility, and degradability [318-320], but also the soil characteristics and the boundary conditions.

Mobility means the translocation of the pesticides in the soil. The mobility is strongly dependent on the water solubility and the sorbability of the pesticides. The water solubility of the individual pesticides varies within wide limits. While triazoles, phenoxy fatty acids, and the benzoic acids have a high water solubility of more than 500 mg/L, chlorinated hydrocarbons such as DDT, aldrin (**2**), and heptachlor (**4**) are only slightly soluble (<1 mg/L, Table 23). The mobility of pesticides is determined under laboratory conditions using percolation tests with soil columns. The organic matter and the clay content of a soil and the pH value have to be taken into account as important factors in this connection. The herbicides 2,4-(dichlorophenoxy)acetic acid [2,4-D, (**29**)] and 2,4,5-(trichlorophenoxy)-acetic acid [2,4,5-T, (**31**)], for example, percolate as rapidly in a low-humus subsoil (pH = 6.6, 0.26 % C_{org}, 19 % clay) as tritium-labeled water, because at pH values above 4 they are not bound to clay ([314]). In the cases of monolinuron (**34**) and chloridazon, a translocation in humus-containing soil (pH = 6.4, 3.2 % C_{org}, 8 % clay) could not be detected. Generally, the adsorption on soils with a very low content of organic matter is small. With high content of organic matter, an increase of adsorption is usually found. It can be assumed that both very low and very high contents of soil organic matter retard the degradation of active substances and reduce their effect. In the case of numerous active substances, 60 – 95 % of the substances used can be immobilized by adsorption.

Table 23. Properties and mode of action of crop-protection agents in soil at doses used in practice and favorable conditions of degradation (11 – 16 °C, loose, loamy arable soils, pH 5.5 – 6.5, 2 – 4 % humus, [314]) [a]

Common name (action) [b]	Chemical name	Prohibited, permitted (FRG) [c]	Solubility	Volatility	Bonding by			Degradation		Mobility	Formula
					Humus	Clay, Fe-Ox	pH influence	Aerobic	Anaerobic		
Cation-active											

Bipyridyl derivatives											
Deiquat (H)	1,1′ethylene-2,2′-bipyridylium	+	4	1	5	4 – 5	+	3		1 – 2	
Paraquat (H)	1,1′-dimethyl-4,4′-bipyridylium	+	4	1	5	4 – 5	+	1		1	(**15**)
Weakly basic diazines											
Chloridazone (H)	5-amino-4-chloro-2-phenyl-(2*H*)-pyridazinone	+	3	1	3	2	(–)	4 – 5		2	
Bentazone (H)	3-isopropyl-2,1,3-benzothiadiazin-4-one-2,2-dioxide	(+)	4	1				4 – 5			
s-Triazines											
Prometryn (H)	4,6-diisopropylamino-2-methylthio-*s*-triazine		4	1	3	2	(–)	4		2	(**16**)
Atrazine (H)	4-ethylamino-2-chloro-6-isopropylamino-*s*-triazine	(+)	2 – 3	1	2	1	–	3 – 4	3 – 4	3	(**17**)
Simazine (H)	chlorobis(ethylamino) triazine	(+)	2	1	2	2	(–)	3 – 4		3	(**18**)
Propazine (H)	2-chloro-4,6-diisopropylamino-*s*-triazine	(+)	2	1	2		(–)	3 – 4		3	
Anilazine (F)	2,4-dichloro-*N*-(6-chlorophenyl)-1,3,5-triazine-2-amine	+	2	1			(–)	5			
Terbuthylazine (H)	4-ethylamino-6-*tert*-butylamino-2-chloro-*s*-triazine	+	2	1	2 – 3	2	(–)	3 – 4			
s-Triazoles											
Amitrole (H)	3-amino-1,2,4-triazole	(+)	4	1	2	2	(–)	4		4	
Propiconazol (F)	1-[2-(2,4-dichlorophenyl)-4-propyl-1,3-dioxolan-2-yl (methyl)]-1*H*-1,2,4-triazole	+	3	1			(–)				
Triadimenol (F)	1-(4-chlorophenoxy)-3,3-dimethyl-(1*H*-1,2,3-triazole-2-yl)-butane-2-ol	+	3	1	(1)		(–)	(1)			(**20**)
Triazinones											
Metribuzin (H)	4-amino-6-*tert*-butyl-3-methylthio-1,2,4-triazine	(+)	4	1	2	(2)	(–)	4			
Metamitron (H)	4-amino-3-methyl-6-phenyl-1,2,4-triazine	+	4	1			(–)	5			
Acid											
Phenoxy fatty acids											
2,4-D (H)	2,4-(dichlorophenoxy) acetic acid	+	4	1	1 – 2	1	–	4		4	(**29**)
MCPA (H)	4-(4-chloro-*o*-tolyloxy) acetic acid	+	3	1	2	1	–	4		4	(**30**)
Mecoprop (H) CMPP	2-(4-chloro-2-methylphenoxy)propionic acid	+	4	1	2	1	–	4		3	
2,4,5-T (H)	2,4,5-(trichlorophenoxy) acetic acid	+	3	1	2 – 3	1	–	4	3	3	(**31**)
Fluazifop-butyl (H)	butyl-2-[4(*S*-trifluoromethyl-2-pyridyloxy)-phenoxyl] propionate	(+)	1 – 2	1	2		(–)	5			
Dichlorprop (H)	2-(2,4-dichlorophenoxy) propionic acid	(+)	3 – 4	1			(–)	4 – 5			
Benzoic acid											
Dicamba (H)	3,5-dichloro-2-methoxybenzoic acid	(+)	4	1	2	1	–	5		4	
Chlorambene (H)	3-amino-2,5-dichlorobenzoic acid	(+)	4	2			(–)	2		4	
2,3,6-TBA (H)	2,3,6-trichlorobenzoic acid		4	2	2	1	–	3		4	(**9**)
Picolinic acid derivatives											
Picloram (H)	3-amino-3,5,6-trichloropicolinic acid	(+)	3 – 4	1	2	1	–	2 – 3		3	(**19**)
Phenols, benzenes											
Dinoseb (H)	4,6-dinitro-2-(1-methylpropyl)phenol	(+)	4	2	2		(–)	5		4	(**10**)
DNOC (H)	4,6-dinitro-2-methylphenol	(+)	3	1	2 – 3		–	4		3	(**11**)
PCP (I)	pentachlorophenol	86	2	1	4	2	–	3 – 4	4	1	(**12**)
Quintozene (F)	pentachloronitrobenzene (PCNB)	+	1	2			0	2 – 3	4	1	
Hexachlorobenzene (F)	hexachlorobenezene	77	1	1	4	2	0	2	1	1	
Apolar											
Halogenated hydro-carbons											
Endosulfan (I)	6,7,8,9,10,10-hexachloro-1,5,5*a*,6,9,9*a*-hexahydro-6,9-methano-2,4,3-benzodioxathiepin 3-oxide	+	1	2	4		0	3	(2)	1	(**1**)
Aldrin	1,2,3,4,10,10-hexachloro-1,4,4*a*,5,8,8*a*-hexahydro-1,4-*endo*,*exo*-5,8-dimethanonaphthalene	81	1	1	(5)	(3)	0	3	(2)	1	(**2**)
Dieldrin (I)	3,4,5,6,9,9-hexachloro-1*a*, 2,2*a*,3,6,6*a*,7,7*a*-	71	1	1	5	3	0	1 – 2	1 – 2	1	(**3**)

	octahydro2,7 : 3,6-dimethanonaphth[2,3*b*] oxirene										
Heptachlor (I)	1,4,5,6,7,8,8-heptachloro-3*a*,4,7,7*a*-tetrahydro-4,7-methanoindene	86	1	2	5	3	0	2 – 3	2 – 3	1	(**4**)
Chlordane (I)	1,2,4,5,6,7,8,8-octachloro-2,3,3*a*,4,7,7*a*-hexahydro-4,7-methano-1*H*-indene	71	1	2	5		0	3		1	(**5**)
Lindane (ϒ-HCH) (I)	ϒ-1,2,3,4,5,6-hexachlorocyclohexane	(+)	2	1	3 – 4	3	0	2 – 3	5	1	(**6**)
DDT (I)	1,1′-(2,2,2-trichloroethylidene)bis(4-chlorobenzene)	77	1	1	5	4	0	1 – 2	3	1	
Dichloropropene (N)	1,3-dichloropropene	(+)	2	2	2	2	0	5	4	4	
Methyl bromide (I, N)	methyl bromide	(+)	4	3			0				
Phosphate esters											
Malathion (I)	*O,O*-dimethyl-*S*-(1,2-dicarbethoxy)ethyl										
	phosphorodithioate	+	3	2	2 – 3		0	3	3 – 4	2	(**27**)
Parathion (I, A)	*O,O*-diethyl-*O*-(4-nitrophenyl)-thionophosphate	+	2	2	4 – 5	2 – 3	0	4 – 5	4	2	
Phorate (I, A)	*O,O*-diethyl-*S*-[(ethylthio)-methyl] phosphodithioate		2 – 3	2	4	2	0	4		2	
Diazinon (I)	*O,O*-diethyl-*O*-[(2-isopropyl-4-methyl) pyrimidin-6-yl)]-phosphorothioate		2	2	3	2	0	4	4	2 – 3	(**28**)
Carbamates											
a) Methylcarbamates											
Carbaryl (I)	1-naphthyl methylcarbamate	86	2 – 3	2	2 – 3		0	3 – 4	4	2 – 3	(**21**)
Carbofuran (I, A, N)	2,3-dihydro-2,2-dimethyl-7-benzofuranyl										
	methylcarbamate	(+)	3	1	3		0	4	3 – 4	2 – 3	(**22**)
Propoxur (I)	2-(1-methylethoxy) phenyl methylcarbamate	(+)	4	2	2		0	(4)			
Phenmedipham (H)	3-[(methyloxycarbonyl)-amino]-phenyl-*N*-(3-methylphenyl)carbamate	+	2	1	3 – 4		0	4			
b) Phenylcarbamates											
Propham (H)	isopropyl *N*-phenylcarbamate	+	2	1	2	1 – 2	0	4 – 5		3	
Chlorpropham (H)	isopropyl *N*-(3-chlorophenyl)carbamate	+	3	1	3	1 – 2	0	4		2	
c) Thiocarbamates											
Diallate (H)	*S*-2,3-dichloroallyl *N,N*-diisopropylthiocarbamate		2	1	4		0	4 – 5		2	
Triallate (H)	2,3,3-trichloroallyl diisopropylthiocarbamate	+	2	1	4	2	0	3 – 4		2	
Butylate (H)	*S*-ethyl *N,N*-diisobutylthiocarbamate	+	2	2	3		0	4		2	
EPTC (H)	*S*-ethyl *N,N*-dipropylthiocarbamate	+	3	2	2		0	4		2	
d) Dithiocarbamates											
Maneb (F)	manganese(II)-1,1-ethylene-bis (dithiocarbamate)	+	1	1			0	5	5		
Metam-Na (F, N, H)	*N*-methyldithiocarbamate	(+)	4	1	1 – 2		0	4		3	
Zineb (F)	zinc-ethylene-bis (dithiocarbamate)	+	2	1	3 – 4		0	4 – 5		2	
Thiram (F)	tetramethylthiuram disulfide	+	3	1	3		0	4 – 5			
Propanil (H)	*N*-(3,4-dichlorophenyl)-propionamide		3	1 – 2	2	1 – 2	0	4			(**7**)
Mancozeb (F)	ethylene-bis (dithiocarbamate)	+	1	1			0	4 – 5			
e) Oxicarbamates											
Aldicarb (I, N)	2-methyl-2-(methylthio) propanal *O*-[(methylamino)carbonyl] oxime	(+)	4	1	2	1	0	3 – 4	4	4	(**23**)
Acetanilides											
Propachlor (H)	2-chloro-*N*-isopropyl-*N*-phenylacetamide	(+)	4	1	2	1 – 2	0	4		3	
Carboxin (F)	5,6-dihydro-2-methyl-1,4-oxathiin-3-carboxanilide	+	3	1	5	3	0	4 – 5		2	
Metolachlor (H)	chloroacetanilide	+	4	1	2 – 3		0	4			(**8**)
Metazachlor (H)	*N*-(2,6-dimethylphenyl)-*N*-(1-pyrazolylmethyl) chloroacetanilide	(+)	3	1	2 – 3		0	4 – 5			
Urea derivatives											
Fenuron (H)	1,1-dimethyl-3-phenylurea		4	1	2 – 3	2	0	3		3	
Monuron (H)	3-(4-chlorophenyl)-1,1-dimethylurea	+	3	1	3	2	0	4		3	(**32**)
Diuron (H)	3-(3,4-dichlorophenyl)–1–1-dimethylurea	+	4	1	3	2	0	3		2	(**33**)
Monolinuron (H)	3-(4-chlorophenyl)-1-methoxy-1-methylurea	+	3 – 4	1	3	2	0	4		2	(**34**)

Methabenzthiazuron (H)	1-(benzthiazol-2-yl)-1,3-dimethylurea	+	3	1	3	2	0	3	2	(**35**)
Metoxuron (H)	3-(3-chloro-4-methoxyphenyl)-1,1-dimethylurea	+	4	1	2	(2)	0	4 – 5	2	
Linuron (H)	3-(3,4-dichlorophenyl)-1-methoxy-1-methylurea		3	1	3 – 4	2	0	4	2	(**36**)
Chlortoluron (H)	3-(3-chloro-4-methylphenyl)-1,1-dimethylurea	+	3 – 4	1	2 – 3		0	4 – 5		
Isoproturon (H)	*N*-(4-isopropylphenyl)-*N′*,*N′*-dimethylurea	+	3	1	2		0	5		
Dinitroanilines										
Nitralin (H)	4-(methylsulfonyl)-2,6-dinitro-*N*,*N*-dipropylaniline	+	1	1	3 – 4	(3)	0	3 – 4	1	(**13**)
Trifluralin (H)	2,6-dinitro-*N*,*N*-dipropyl-4-trifluoromethylaniline	+	1	1	4 – 5	(3)	0	3 – 4	1	(**14**)
Pendimethalin (H)	*N*-(3-phenyl)-3,4-dimethyl-2,6-dinitroaniline	+	1	1			0	4		
Benzimidazoles										
Benomyl (F)	methyl-1-(butylcarbamoyl)-2-benzimidazolcarbamate	+	2	1	5	3	0	4	1	
Thiophanatemethyl (F)	1,2-bis[(3-methoxycarbonyl)-2-thioureido]-benzene	+	2	1			0	4 – 5	1	(**24**)
Carbendazim (F)	methyl(1-*H*-benzimidazol-2-yl) carbamate	+	2	1	4		–	3 – 4	1	(**25**)
Prochloraz (F)	*N*-propyl-*N*-2,4,6-trichlorphenoxy-ethylimidazole-carboxamide	+	2	1						
Phthalimide derivatives										
Captan (F)	*N*-trichloromethylthio-cyclohex-4-ene-1,2-dicarboximide	+	1	1	1 – 2		0	4 – 5	2	(**26**)
Captafol (F)	*N*-(1,1,2,2,2-tetrachloroethylthio)-3,6,7,8-tetrahydrophthalimide	+	1 – 2	1	1 – 2		0	4 – 5		
Aromatic cyanides										
Dichlobenil (H)	2,6-dichlorobenzocyanide	+	2	2	2		0	3	2	
Ioxynil (H)	3,5-diiodo-4-hydroxybenzonitrile	+	2	1	2		0	5		
Bromoxynil (H)	3,5-dibromo-4-hydroxybenzonitrile	+	3	1	2		0	4 – 5		
Amino acid derivatives										
Glyphosate (H)	*N*-(phosphonomethyl) glycine	+	4	1	3			3 – 4		
Metalaxyl (F)	methyl-D,L-*N*-2,6-dimethylphenyl-*N*-(2′-methoxyacetyl) alanate	(+)	4	1	3			4		
Aldehyde oligomers										
Metaldehyde (M)	2,4,6,8-tetramethyl-1,2,3,5,7-tetraoxacyclooctane	+	3	2	2					
Morpholine derivatives										
Fenpropimorph (F)	*cis*-4-[5-(4-*tert*-butylphenyl)-2-methylpropyl]-2,6-dimethylmorpholine	(+)	2	1						
Quaternary ammonium compounds										
Chlormequat (H)	chlorocholine chloride	+	4	1						
Uracil derivatives										
Bromacil (H)	5-bromo-3-*sec*-butyl-6-methyluracil	+	4	1				3		
Benzofuran derivatives										
Ethofumesat (H)	(2-ethoxy-2,3-dihydro-3,3-dimethylbenzofuran-5-yl) methanesulfonate	+	3	1	2			3 – 4		

[a] Ranking: **0** practically not; **1** very slight, very weak; **2** slight, weak; **3** medium to increased; **4** high, strong; **5** very high, very strong; ( ) doubtful data; solubility (mg/L H_2O by 20 °C): **1** <1; **2** 1 – 50; **3** 50 – 500; **4** >500; strength of binding: Freundlich sorption coefficient of humus and clay: K_{OC}: **1** 1 – 100; **2** 100 – 300; **3** 300 – 1000; **4** 1000 – 10 000; **5** >10 000; K_{Clay}: **1** 0.5 – 50; **2** 50 – 150; **3** 150 – 500; **4** 500 – 5000; **5** >5000; influence of bonding increases with (+) or decreasing (–) pH; 0 no effect; degradation (>75 %): **1** >3 years; **2** 1 – 3 years; **3** 18 weeks to 1 year; **4** <18 weeks; **5** <6 weeks; volatility (vapor pressure in hPa at 20 – 25 °C): **1** $<10^{-3}$; **2** 10^{-3} – 100; **3** 100 – 5×10^3; **4** $<5\times10^3$ (data for pure substance). Mobility (annual depth of translation by 1500 mm rain, at 25 °C in loamy soil with ca. 2 % humus and pH 6 in the Ap horizon: **1** <10; **2** 5 – 20; **3** 15 – 40; **4** >35 cm.
[b] (A) acaricide; (F) fungicide; (H) herbicide; (I) insecticide; (N) nematocide; (R) rodenticide; (M) molluscicide.
[c] Since 1988 prohibited, + permitted, ( ) prohibited in water protection zones since September 1, 1988 (Pflanzenschutzanwendungsverordnung).

Transformation. Pesticides can be biologically transformed (cometabolized) in the soil under aerobic or anaerobic conditions and/or degraded to CO_2 and H_2O (mineralized). In the course of the degradation process, metabolites, which can be harmless or toxic, can be formed.

In the presence of microorganisms pesticides undergo oxidations, hydrolyses, and aerobic and anaerobic ring cleavage [318], [320]. The stability of the pesticides in the soil is described by their half-life ($t_{1/2}$). The half-lifes are regarded as ecologically relevant band-widths of stability. If deviations from an average band-width occur, this is in most cases due to soil properties or soil/environment conditions that lead to particularly low or particularly high stability of active substances [316]. Degradation rates usually mean the disappearance of the pesticides within a certain period of time. The majority of the test systems previously proposed and used for pesticide degradation in the soil assume that the active substances are retained in the Ap horizon and are also degraded there. However, because of the detection of certain pesticides in the groundwater [321], [322] the reactions and effects in the subsoil were investigated to an increased extent [323].

The *biodegradability* of the active substances is influenced by biological factors such as the quantity of the microbial biomass, the physiological state (activities), and adaptability of the microorganisms.

Factors such as temperature, salinity, water content, oxygen availability and the supply of organic substances determine the composition of the microbial populations at the site and

have therefore to be considered during the application of the pesticides.

On the soil surface, pesticides are *photochemically degraded* and *chemically converted*. Pentachlorophenol and DDT can be degraded photochemically, e.g., in presence of UV radiation to CO_2 and hydrochloric acid. Chemical conversions of the pesticides also occur by hydrolysis, oxidation, and isomerization.

Table 23 shows that the pesticides differ greatly in their biological and chemical degradability. While the phenoxy fatty acids can be degraded in arable soils, mainly in their topsoil, in about 18 weeks, pentachlorobenzene and the majority of the chlorinated hydrocarbons remain detectable for many years.

Behavior in the Environment. The applied amount, the formulation, and the type of application are very important for the action and the environmental behavior of pesticides. The interplay of the various factors finally determines the actual behavior of the pesticides in the environment.

Composition and properties of pesticides and their behavior in soils at doses used in practice and favorable degradation conditions are listed in Table 23 [314].

Legal Aspects, Test Systems. The use of pesticides causes a contamination of the soils; the magnitude of the enrichment depends on the persistence of an active substance, the amount and frequency of application, and also on the sorption, degradation, and translocation conditions of the soil [317], [319]. The formation of pesticide residues in agricultural soils as well as the contamination of groundwater and surface water cannot be excluded. According to the German law on crop protection (Pflanzenschutzgesetz), pesticides, if applied properly, must have no impact on humans, animals, or groundwater and not affect the natural economy in a way that does not comply with the present state of scientific knowledge. In order to guarantee these requirements, pesticides are subjected before use to extensive studies, so that sufficient information is available on their behavior (physicochemical properties, adsorption/desorption, degradability, etc.) and possible side-effects on plants, selected animals, and microorganisms under standardized laboratory conditions. Test systems for determining the behavior of pesticides, such as the degradability or the percolation behavior in soil columns have been introduced in the industrialized countries and by international organization (OECD, EU, WHO, FAO). Test systems such as those for acute and chronic toxicity, mutagenicity, carcinogenicity, and bioaccumulation are used for estimating the risk potential [324-330]. The information gained is only to a limited extent applicable to the ecosystem. According to experience with aquatic organisms, the informative value of tests can certainly be increased by compilation of test batteries, but the uncertainties remain considerable.

In the soil area the action of the pesticides on the whole microbial population can be estimated. With the exception of the oxidation of ammonia to nitrite (nitrification) the microbial indicators used hitherto, such as respiration, enzyme activities, or bacterial counts, have not been very sensitive. This was actually to be expected, because the ecological relevance of the result of a measurement generally increases dramatically with multi-species test system, usually at the expense of sensitivity [324]. Investigations of the action of pesticides under field conditions are certainly desirable. At present this seems difficult for logistical and financial reasons and because of the low reproducibility of the results.

Almost all known tests investigate the action of environmental chemicals under aerobic conditions. Anaerobic test systems are hardly available [325], [331]. It is to be expected, though, that aerobic degradation will differ greatly from anaerobic degradation, with regard both to the composition of the microbial populations and to degradation pathways. An action of the pesticides differing according to O_2 availability is therefore to be expected. ALEF et al. [332] have developed the arginine mineralization as a simple and reproducible method for estimating the action of environmental chemicals on anaerobic microbial populations. Both the NOEL and the EC_{50} values can be determined from the dose – response relationship (Fig. 19; the greatest inhibition of the arginine mineralization by the environmental chemicals was usually found on the third day of incubation). The new method can be used for determining the acute and chronic effects of environmental chemicals [332].

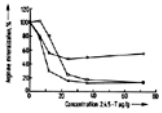

Figure 19. Action of 2,4,5-T on the anaerobic arginine mineralization; data in % of the control values [332]

Spheres = 2 days; triangles = 3 days; squares = 4 days

To obtain registration for a pesticide, generally studies carried out under defined “worst-case” conditions must be submitted investigating the effect on soil microorganisms and on selected organisms. The results obtained give information on the behavior of pesticides in the environment; they should be reasonably comprehensive so that the authorities can make a decision taking into account the intended use. The objective of the investigations can never be a complete picture of the behavior of the pesticides in all environmental compartments.

5.3.2.2. *Side-Effects*

Pesticides can interfere with various processes in the cell. Their mechanisms of action differ [333]. While photosynthesis is the main point of attack for herbicides, fungicides attack nuclear division, nucleic acid and protein synthesis, respiration, biogenetic pathways, and the membrane function. Insecticides influence respiration and the membrane structure.

To be able to use pesticides profitably, active substances are required whose selective action on crop plants is as different as possible from that on target organisms. A very high specificity of action, though, is not always achievable. Herbicides, for example, can influence respiration and biogenetic pathways aside from photosynthesis, and also inhibit germination. Undesirable effects on soil organisms, plants, animals, and humans can therefore be expected. Pesticides can reach humans via the food chain or drinking water.

Side-Effects on Soil Organisms. The action of pesticides on soil organisms is to be regarded as negative if the soil biota is so extensively disturbed that essential functions of the ecosystem such as decomposition and humification, aggregation, nutrient immobilization and mobilization, loosening, and bioturbation are lastingly impaired. In this connection, negative effects can be called “critical” that last for more than one month [316], [334]. From the standpoint of species preservation, action of pesticides should preferably also not lead to any complete change of species or at least guarantee a rapid recolonization after the specific use of soil is finished. The potential regenerability of the soil biota has in any case to be preserved.

The effects of herbicides, fungicides, and insecticides are not completely specific. Their action depends on the concentration used, on the mode of application, and on the properties of the site. Fungicides such as benomyl and thiabendazole, for example, can have a negative effect on earthworms. Other pesticides can influence bacteria and stimulate ammonification.

Herbicides such as paraquat, linuron, and 2,4-D can stimulate bacterial activities (e.g., respiration), but on the other hand injure algae [335-337]. Insecticides can in some cases also act as nematocides and then be used as such. Carbamates and phosphate esters are in some cases toxic to earthworms and their activity even at a normal dose [335].

Different site conditions not only modify the expected impact of pesticides but can also lead to different side-effects, a topic which has been little studied up to now. It is impossible to consider all environmental factors in the laboratory test; notwithstanding that, basic information on the possible ecotoxicological action of pesticides is indispensable in practice.

Various groups have proposed uniform procedures or tests of estimating the action of pesticides on soil microorganisms (respiration, substrate-induced respiration, nitrification, N mineralization, dehydrogenase activity, [334]). A series of data is available in the literature, on the effect of pesticides on the composition of microbial populations (bacterial counts, fungal counts, or plate counts). This information should be regarded with caution, especially because no uniform procedure is given and comparability of the data is therefore scarcely possible. Furthermore, reliable information on the action of pesticides on the interactions between the various organisms in the soil is scarcely obtainable. There is a great need for research in this field.

Uptake, Accumulation, and Side-Effect on Plants. In principle, the uptake and accumulation of pesticides can occur in all parts of plants. This depends on the plant species, on the physicochemical properties of the substances and on the site (pH value, content of organic substances in the soil, soil structure).

In *soil application*, pesticides are absorbed with the soil water by the plant roots and transported in the xylem. With high vapor pressure, the active substances can also reach the roots in the soil airspace.

In *leaf application*, the pesticides can penetrate into the leaf by diffusion. The uptake is considerably easier for lipophilic than for hydrophilic substances. Both in soil and in leaf application the active substances can accumulate in the plant leaf. However, pesticides can also remain at the site of application, and not be transported in plants. Aside from the positive effects of pesticides, negative effects can also occur. The side-effects can extend as far as to serious damage [338]. Under extreme weather conditions, damage, e.g., to crop plants cannot be ruled out. After the use of urea derivatives, for example, an increased attack by fungi has occasionally been observed in grain cultivation. The selectivity of the active substances is of great importance and is therefore tested during the registration process.

A consequence of the use of pesticides in agriculture is that occasional residues in foods cannot be prevented. In order to be able to estimate the risk potential, both the acute and the chronic action on animals (e.g., rats, guinea-pigs) are tested during the registration process. In these studies, the NOEL value (⟶ Toxicology – Toxicological Studies: General Aspects) for the experimental animal is determined. In application to humans this maximum inactive dose is still for safety's sake reduced by a factor of at least 100. From this, there is obtained the highest still acceptable daily dose in milligrams per kilogram food (ADI; ⟶ Foods, 3. Food Additives – Safety Aspects; ⟶ Insect Control – Toxicology of Insecticide Residues). To prevent that the maximum quantity regulation is exceeded when using pesticides, the concept of “waiting time” is introduced. This means how many days before the harvest of the crop, at the latest, a pesticide may be used to make sure that the maximum amount is not exceeded. As a result of the maximum quantity regulation, the marketing of foods with excessive pesticide residues is forbidden [339]. The strict conditions cannot, however, completely prevent human foods from containing pesticide residues [340].

Contamination of Groundwater and Surface Waters. Because of their mobility, pesticides can be translocated after their application into the groundwater and surface water. Despite high dilution in the surface water, however, remobilization, can occur. The pollution of surface water by polluted groundwater depends on the length and bonding strength of the filter section, the groundwater flow rate, and the degradation rate of the pesticides in the groundwater. With long-term and intensive use of mobile active substances on the soil, pollution of the groundwater must be reckoned with.

In order to protect water bodies from contamination by pesticides, national and international regulations have come into force in which the maximum permitted concentrations have been stipulated. The EC drinking water regulations, for example, provide that drinking water may contain no more than 0.5 µg/L of pesticides in total and 0.1 µg/L of a single active substance.

5.4. Effects of Abandoned Waste Disposal Sites

5.4.1. ***Introduction***

The genesis of contaminated sites is different from site to site. A distinction can be made, however, between those sites in which the chemical nature, quantity, number and distribution of the contaminants are largely production-specific and those in which they are waste- and deposition-specific.

The first group includes contaminated plant sites (*contaminated land*). Aside from different localities and extents, these sites also differ in the arrangements of manufacturing plants, handling facilities for water-endangering substances, intermediate storage facilities, waste deposits, etc., depending on the type of plant (gasworks, oil refinery, a factory for manufacturing explosives, weapons, plant-protection agents, dyes, etc.). Old sites are contaminated from specific manufacturing operations or production residues. At present no statistically supported results on the extent of this contamination exist. In addition, usually no data are available on the persistence and mobility of specific contaminants in soil and groundwater and on their possible ecotoxicity or human toxicity.

The situation is usually different in the case of abandoned waste disposals (*abandoned sites*). By contrast with the contaminated land, from which dangers often arise only from changes in ground use (e.g., excavation of foundation pits) or by direct contact with "easily accessable" hazardous substances (for example by children at play), dangers originating from old deposits can be attributed almost entirely to their emissions. Especially important are the leachates, loaded with considerable quantities of pollutants that reach the soil and subsequently the groundwater and can then endanger its use as drinking water. In Germany alone account must be taken of more than 100 000 old deposits. This kind of contamination of soil and groundwater therefore represents a problem also with regard to the area concerned, which is steadily increasing because the pollutant plumes continue to expand in the groundwater. This section describes the qualitative and quantitative effects of leachate emissions from abandoned waste disposals on groundwater.

5.4.2. *Material Contents in Groundwater Affected by Abandoned Waste Disposal Sites*

Inorganic Substances. The natural substance contents of groundwater are changed considerably by the leachates of abandoned sites [341]. Table 24 assembles the mean, median, and other percentile values calculated for contaminated groundwaters from unconsolidated rocks, together with the maximum values of the various inorganic parameters.

Table 24. Inorganic constituents of groundwaters contaminated by abandoned waste disposal sites and their statistical parameters

Parameters	DL [1]	Number of measured values [2], *n*	*n* >DL	DF[3], %	Mean A[4]	Mean B[5]	Median 50 % percentile	75 % percentile	Maximum
Main cations									
Calcium, mg/L	5.0	318	317	99.7	177.	175.	156.	229.	785.
Magnesium, mg/L	2.0	321	320	99.7	39.3	38.1	22.0	42.3	436.
Sodium, mg/L	5.0	338	337	99.7	140.	128.	45.6	130.	3600.
Potassium, mg/L	1.0	217	217	100.	34.4	32.9	12.0	36.0	350.
Ammonium, mg/L	0.02	276	231	83.7	14.4	11.0	0.41	6.2	945.
Iron, mg/L	0.01	335	330	98.5	10.3	9.6	2.0	11.0	240.
Manganese, mg/L	0.01	329	311	94.5	1.6	1.5	0.43	1.2	33.1
Main anions									
Hydrogencarbonate, mg/L	18.0	233	233	100.	527.	519.	408.	661.	2458.
Chloride, mg/L	1.0	360	360	100.	218.	202.	74.2	168.	6020.
Sulfate, mg/L	1.0	352	350	99.4	218.	194.	122.	223.	8560.
Nitrate, mg/L	0.5	347	244	70.3	58.7	25.6	2.5	24.3	11 500.
Trace elements									
Aluminum, µg/L	40.0	168	135	80.4	947.	596.	80.0	573.	59 600.
Lead, µg/L	0.05	334	129	38.6	7.6	6.3	<0.05	2.1	450.
Cadmium µg/L	0.05	335	126	37.6	40.3	1.5	<0.05	0.3	13 000.
Chromium (total), µg/L	10.0	279	179	64.2	39.5	21.2	<10.	<10.	5 120.
Copper, µg/L	10.0	275	189	68.7	26.5	24.5	<10.	20.0	577.
Nickel, µg/L	10.0	249	187	75.1	116.	22.9	13.0	24.0	23 200.
Mercury, µg/L	0.2	232	39	16.8	<0.2	<0.2	<0.2	<0.2	<0.2
Strontium, µg/L	40.0	205	201	98.0	604.	580.	399.	764.	5 580.
Zinc, µg/L	10.0	328	276	84.1	921.	410.	64.5	220.	168 000.
Trace anions									
Arsenic [arsenates(III)/(V)], µg/L	0.2	253	172	68.0	61.0	45.4	1.3	6.0	4 000.
Nitrite, mg/L	0.1	285	107	37.5	0.77	0.49	<0.1	0.3	79.0
Selenium [selenates(I)/(VI)], µg/L	0.2	112	20	17.9	0.22	0.10	<0.2	<0.2	14.0
Boron (borates, total),µg/L	20.0	259	253	97.7	949.	832.	181.	530.	31 200.
Fluoride, mg/L	0.5	97	57	58.8	<0.5	<0.5	<0.5	0.50	1.7
Phosphate (total), mg/L	0.1	207	113	54.6	2.3	1.8	0.1	0.60	115.

[1] DL = Detection limit.
[2] In some cases the mean values of one well are used, therefore *n* is the number of measured wells.
[3] DF = Analytical detection frequency (percentage from *n* > DL to n_{total}).
[4] Arithmetic mean of all measured values including maximum value.
[5] Arithmetic mean rejecting maximum value.

Organic Substances. As regards the number of substances the possible contamination of groundwater in the outflow from abandoned sites by organic compounds is far greater than that by inorganic compounds. About 1200 organic groundwater contaminants have up to now been detected by extensive GC – MS analytical studies in Germany and the United States [342] in the outflow from abandoned disposal sites. Only a relatively small number of these contaminants, however, are detected fairly frequently and at the same time have concentrations distinctly above 1 µg/L (United States) or above 0.1 µg/L (West Germany), respectively. The majority of the identified organic contaminants (>1000) have a detection frequency (DF, see Section Emission Detection Frequency) of less than 0.1 %; i.e., in 1000 contaminated groundwater samples polluted by abandoned waste deposits they are detected on average less than once.

A bias for substances or classes of substances detected in the contaminated groundwater due to the analytical method used can be almost ruled out because the sample preparation (substances extractable under acid or basic/neutral conditions, highly volatile substances, and pesticides) and the analytical method (essentially GC – MS) were aimed at the detection of the largest number of groundwater contaminants possible [343], [344].

Table 25 lists, in the order of decreasing *detection frequency*, the important organic groundwater contaminants determined in Germany (former Western Germany) with various statistical parameters of their concentrations. Of the 19 contaminants with detection frequencies ≥10 %, 12 (63 %) are volatile halogenated hydrocarbons. Of these, 5 are alkane derivatives, 4 are alkene derivatives, and 3 are chlorosubstituted benzenes. Among the 7 halogen-free contaminants, benzene and its alkyl derivatives (4 compounds) are most important; in addition 2 phenols and naphthalene have detection frequencies ≥10 %.

Table 25. Organic constituents of groundwater contaminated by abandoned waste disposal sites and their statistical parameters

Parameters	DL[1], µg/L	Number of measured values[2], *n*	*n* > DL	DF[3] %	Nonlogarithmic values				Mean of logarithmic values	
					Mean µg/L	Median 50 % percentile, µg/L	75 % percentile, µg/L	Maximum µg/L	log	µg/L
Tetrachloroethylene	0.1	277	195	70.4	56.1	1.4	3.7	6 500.	0.62	4.2
Trichloroethylene	0.1	277	154	55.6	1 010.	2.3	11.0	128 000.	1.06	11.5

cis-1,2-Dichloroethylene	30.1	22 100.	166.	1 150.	411 000.	2.37	234.			
Benzene	1.0	127	38	29.1	141.	13.5	38.0	1 795.	1.24	17.4
1,1,1-Trichloroethane	0.1	206	47	22.8	16.5	1.0	7.4	270.	0.25	1.8
m/*p*-Xylene	0.1	92	21	22.8	39.9	2.8	4.9	447.	0.53	3.4
Trichloromethane	0.1	236	52	22.0	76.2	1.4	10.1	2 800.	0.32	2.1
1,2-Dichloroethane	5.0	16	3	18.8	107.	<5.	<5.	210.	1.86	72.4
Vinyl chloride	1.0	136	24	17.7	1 690.	99.5	1 950.	12 000.	1.93	85.1
Toluene	0.1	127	21	16.5	73.2	3.5	11.0	911.	0.61	4.1
Dichloromethane	10.0	114	17	14.9	38 100.	438.	16 800.	499 000.	2.94	871.
Tetrachloromethane	0.1	201	29	14.4	1.2	0.2	0.3	23.0	0.48	3.0
4-Methylphenol (*p*-Cresol)	1.0	124	17	13.7	42.0	8.6	20.7	283.	1.08	12.0
Chlorobenzene	0.1	93	12	12.9	52.9	2.3	6.5	388.	0.67	4.7
2-Methylphenol (*o*-Cresol)	0.5	124	16	12.9	1.0	5.9	9.2	63.3	0.66	4.6
1,2-Dichlorobenzene	0.1	90	11	12.2	1.4	0.9	2.1	6.6	0.08	0.83
1,4-Dichlorobenzene	0.1	90	11	12.2	31.9	2.2	37.5	265.	0.53	3.4
Naphthalene	0.1	124	15	12.1	2.2	1.1	2.1	12.6	0.16	1.5
Ethylbenzene	0.1	124	14	11.3	32.2	4.4	62.5	160.	0.82	6.6
o-Xylene	0.1	127	12	9.5	13.8	4.7	19.6	69.0	0.77	5.9
2,4,6-Trichlorophenol	0.1	124	11	8.9	3.2	0.6	1.5	24.1	–0.15	0.71
3,5-Dimethylphenol	0.1	124	10	8.1	16.2	6.9	35.6	61.0	0.60	4.0
Phenol	0.1	124	10	8.1	2.2	1.5	3.6	5.6	0.24	1.7
1,1,2-Trichloroethane	0.5	140	11	7.9	36.0	3.0	70.0	190.0	0.88	7.6
1,3-Dichlorobenzene	0.1	90	7	7.8	11.5	1.1	38.2	74.0	0.23	1.7
trans-1,2-Dichloro-ethylene	5.0	134	10	7.5	57.1	50.0	95.0	135.	1.61	40.7
Cumene	0.1	90	5	5.6	2.4	3.1	4.5	4.7	0.20	1.6
1,1-Dichloroethane	10.0	130	7	5.4	52.7	53.0	90.0	110.	1.64	43.7
Acenaphthene	0.1	124	6	4.8	6.3	1.3	17.1	32.0	0.19	1.6
2,4-Dichlorophenol	0.1	124	6	4.8	3.5	0.9	9.6	17.2	–0.07	0.85
3-Chlorophenol	0.1	124	6	4.8	12.7	12.3	20.8	22.8	1.02	10.5
1,2,4-Trimethylbenzene	0.1	90	4	4.4	1.9	1.5	2.6	3.5	0.22	1.7
2-Ethyltoluene	0.1	90	4	4.4	0.6	0.6	0.8	1.0	–0.21	0.62
2,4,5-Trichlorophenol	0.1	127	5	3.9	7.1	1.3	16.6	31.9	0.23	1.7
Mesitylene	0.1	90	3	3.3	1.7	<0.1	<0.1	4.0	–0.02	0.95
Phenanthrene	0.1	124	4	3.2	1.5	0.6	2.6	4.4	–0.48	0.33
Tribromomethane	1.0	130	4	3.1	3.0	2.5	4.5	6.0	0.39	2.5

[1] Detection limit.
[2] In some cases mean values of one well are used, therefore *n* is the number of measured wells.
[3] Analytical detection frequency (percentage from *n* > DL to n_{total}).

Highly volatile halogenated hydrocarbons also show the highest *mean concentrations* in groundwaters, both when evaluated from the concentrations themselves and when evaluated from the logarithms of the concentrations. Dichloromethane (38 mg/L and 871 µg/L, respectively), *cis*-1,2-dichloroethylene (22 mg/L and 234 µg/L, respectively), vinyl chloride (ca. 1.7 mg/L and 85 µg/L, respectively) and trichloroethylene (ca. 1 mg/L and 11.5 µg/L, respectively) should be mentioned in the first place, so confirming quantitatively the dominance of these compounds as groundwater contaminants from abandoned waste disposal sites [345].

5.4.3. *Contamination Criteria and their Quantification*

The influence of leachates from abandoned sites on the groundwater can be characterized and quantified by comparing for relevant substances the statistical parameters of data records of contaminated and noncontaminated groundwaters calculated from the concentrations determined. In the case of inorganic substances, the geogenic fraction must be taken into account and in the case of a few organic substances the ubiquitous contamination by these compounds.

5.4.3.1. *Contamination Factor*

The contamination factor (*CF*) of a substance represents the ratio of the observed concentration in the contaminated area to the concentration in the uncontaminated area of a site [346]. The contamination factor indicates how many times the concentration of a substance in the groundwater at the contaminated area is increased compared with that in the unaffected area. It is therefore a measure of the intensity of the effect of a substance on groundwater. It is furthermore also an indirect measure of the mobility and persistence of a contaminant on the emission path from the waste site into the groundwater, because only substances that are highly water soluble, have a low sorbability on C_{org} and clay minerals (i.e., a low geoaccumulability), and that are not or not rapidly degradable can ocur in significantly enhanced concentrations in groundwaters.

As a measure of the strength of influence of the individual *inorganic parameters*, the ratio is formed of the mean, rejecting the maximum value, from the data record of the contaminated groundwater (see Table 24) to the mean from the data record of groundwaters that are unpolluted or affected as little as possible (see [341]). In principle a contamination factor can be calculated for other statistical parameters also (median, percentile values, maximum value). A subscript is used (as in CF_{mean} in this case) for differentiating the contamination factors. For all inorganic parameters studied, with the exceptions of selenium and nitrate, the contamination factors CF_{mean} are >2 (Table 26). Values >20 are reached by ammonium (CF_{mean} 64.7), arsenic (CF_{mean} 37.8), chromium (CF_{mean} 36.6), and boron (CF_{mean} 23.9). Values >10 are found for phosphate (CF_{mean} 12.9), aluminum (CF_{mean} 11.9), cadmium (CF_{mean} 11.5), and potassium (CF_{mean} 11.3). This verifies the strongly increased concentration of these substances in groundwaters as compared with their natural concentrations.

Table 26. Inorganic groundwater contaminants from abandoned waste disposal sites listed in the order of decreasing contamination factors

Substance	Mean*, contaminated**	Mean, uncontaminated [341]	Contamination factor (CF_{mean})
Ammonium, mg/L	11.0	0.17	64.7
Nitrite, mg/L	0.49	<0.01	>49.0
Arsenic, µg/L	45.4	1.2	37.8
Chromium, µg/L	21.2	0.58	36.6
Boron, µg/L	832.	34.8	23.9
Phosphate, mg/L	1.8	0.14	12.9
Aluminum, µg/L	596.	50.	11.9
Cadmium, µg/L	1.5	0.13	11.5
Potassium, mg/L	32.9	2.9	11.3
Manganese, mg/L	1.5	0.18	8.3
Nickel, µg/L	22.9	2.8	8.2
Sodium, mg/L	128.	17.5	7.3
Chloride, mg/L	202.	33.4	6.0
Iron, mg/L	9.6	2.0	4.8
Lead, µg/L	6.3	1.4	4.5
Zinc, µg/L	410.	117.	3.5
Copper, µg/L	24.5	7.0	3.5
Sulfate, mg/L	194.	59.0	3.3

Magnesium, mg/L	38.1	14.3	2.7
Strontium, µg/L	580.	240.	2.4
Hydrogencarbonate, mg/L	519.	238.	2.2
Calcium, mg/L	175.	83.2	2.1
Nitrate, mg/L	25.6	13.3	1.9
Selenium, µg/L	0.10	0.55	0.2
Fluoride, mg/L	<0.5	<0.1	
Mercury, µg/L	<0.2	<0.1	

* Arithmetic mean, rejecting maximum value.
** see Table 24

5.4.3.2. *Emission Detection Frequency*

The emission detection frequency of a substance is by definition the percentage of detections of a substance that has entered the groundwater with the leachate from the abandoned site [for example, if the boron contents of 259 contaminated groundwater samples (different wells) is determined and the concentration in 222 samples is higher than the background, the emission detection frequency is 85.7 %]. It is important in this connection whether an emitted substance (contaminant) can occur naturally in the groundwater or not. Naturally occurring are almost all inorganic contaminants, but not the purely anthropogenic organic substances. To recognize an increase of contents of the natural inorganic constituents of water, their geogenic content must be known.

In the case of the anthropogenic organic contaminants, the natural geogenic “background” concentration is zero; i.e., analytical detection frequency and emission detection frequency are usually identical. In the case of the detection frequency, it is not important — as in the case of the contamination factor — how large the increase is but only that the value is increased.

Inorganic Substances. In the case of inorganic substances, the arithmetic mean value from the data record of the “least affected” groundwaters from unconsolidated rock aquifers is by definition the concentration for the average geogenic background [341]. The number of measured values above this geogenic background value is determined from the data record of the groundwaters affected by abandoned waste disposal sites and the percentage detection frequency calculated. If the background value is below the detection limit, the corresponding detection limit is used instead (fluoride, nitrite, mercury). The results are listed in Table 27. Hydrogencarbonate shows the highest emission detection frequency (85.8 %), followed by boron (85.7 %), sodium (83.4 %), chloride (79.2 %), and magnesium (76.0 %). A lower detection frequency is found for lead (38.3 %), nitrite (35.4 %), aluminum (33.0 %), selenium (22.3 %), cadmium (14.9 %), and mercury (14.7 %).

Table 27. Inorganic groundwater contaminants from abandoned waste disposal sites listed in the order of decreasing detection frequencies

Substance	“Background” (mean unaffected [341])		Number of samples affected*	Number of samples with value > “background”	Detection frequency, %**
Hydrogen-carbonate	238.	mg/L	233	200	85.8
Boron	34.8	µg/L	259	222	85.7
Sodium	17.5	mg/L	338	282	83.4
Chloride	33.4	mg/L	360	285	79.2
Magnesium	14.3	mg/L	321	244	76.0
Calcium	83.2	mg/L	318	237	74.5
Potassium	2.9	mg/L	217	161	74.2
Strontium	240.	µg/L	205	152	74.1
Sulfate	59.	mg/L	352	227	64.5
Nickel	2.8	µg/L	249	160	64.3
Manganese	0.18	mg/L	329	206	62.6
Arsenic	1.2	µg/L	253	155	61.3
Chromium	0.58	µg/L	279	157	56.3
Ammonium	0.17	mg/L	276	148	53.6
Fluoride	<0.1	mg/L	97	51	52.6
Copper	7.0	µg/L	275	144	52.4
Nitrate	13.3	mg/L	347	176	50.7
Phosphate	0.14	mg/L	207	86	41.5
Zinc	117.	µg/L	328	135	41.2
Iron	2.0	mg/L	335	136	40.6
Lead	1.4	µg/L	334	128	38.3
Nitrite	<0.01	mg/L	285	101	35.4
Aluminum	50.	µg/L	168	57	33.9
Selenium	0.55	µg/L	112	25	22.3
Cadmium	0.13	µg/L	335	50	14.9
Mercury	<0.1	µg/L	232	34	14.7

* see Table 24.
** Emission detection frequency.

Comparison of the detection frequencies (Table 27) with the contamination factors (Table 26) shows that a high contamination factor does not necessarily also imply a high detection frequency. Thus aluminum and cadmium have relatively high contamination factors of 11.9 and 11.5, respectively, but comparatively small detection frequencies (33.9 % and 14.9 %, respectively). The reverse situation also occurs (e.g., HCO_3^-, Mg). For inorganic substances high contamination factors in combination with high detection frequencies can be regarded as typical of emissions of abandoned sites into the groundwater. Contamination factors >10 and detection frequencies >50 % are found for arsenic (37.8, 61.3 %), ammonium (64.7, 53.6 %), boron (23.9, 85.7 %), potassium (11.3, 74.2 %), and chromium (36.6, 56.3 %).

Organic Substances. For some organic substances the analytical detection limit is so low that even ubiquitous loadings not attributable to old deposit emissions are recorded. Thus the analytical detection limit for trichloroethylene, tetrachloroethylene and trichloromethane is 0.1 µg/L, whereas the ubiquitous loadings are in the order of up to 0.4 µg/L [345]. Thus, an analytical detection of these substances in the concentration range of the ubiquitous loading cannot be interpreted as the detection of emission from an abandoned site.

For determining the emission detection frequency, the so-called “relevance limit” was created. This is the concentration above which a detection of an organic substance can be attributed with considerable certainty to the emission from an old deposit. In nearly all cases the relevance limit corresponds to the detection limit: only for a few volatile chlorinated hydrocarbons was the relevance limit set at 1 µg/L.

Since in the case of tetrachloroethylene, for example, of 195 measured values above the analytical detection limit (0.1 µg/L), 67 measured values (24.2 %) fall below the relevance limit of 1 µg/L, the emission detection frequency is only 46.2 % (Table 28), compared with the analytical detection frequency of 70.4 % (Table 25).

Table 28. Organic groundwater contaminants from abandoned waste disposal sites listed in the order of decreasing emission detection frequencies

Substance	Number of analyses	Relevance limit, µg/L	Number of measured values > relevance limit	Detection frequency, %
Tetrachloroethylene	277	1.0	128	46.21
Trichloroethylene	277	1.0	113	40.79
cis-1,2-Dichloroethylene	153	4.0	46	30.07
Benzene	127	1.0	37	29.13
p/m-Xylene	92	0.1	21	22.83

1,1,1-Trichloroethane	206	0.1	47	22.82
Trichloromethane	236	0.1	52	22.03
1,2-Dichloroethane	16	5.0	3	18.75
Vinyl chloride	136	1.0	24	17.65
Toluene	127	0.1	21	16.54
Dichloromethane	114	10.0	17	14.91
Tetrachloromethane	201	0.1	29	14.43
4-Methylphenol (*p*-Cresol)	124	1.0	17	13.71
2-Methylphenol (*o*-Cresol)	124	0.5	16	12.90
Chlorobenzene	93	0.1	12	12.90
1,2-Dichlorobenzene	90	0.1	11	12.22
1,4-Dichlorobenzene	90	0.1	11	12.22
Naphthalene	124	0.1	15	12.10
Ethylbenzene	124	0.1	14	11.29
o-Xylene	127	0.1	12	9.45
2,4,6-Trichlorophenol	124	0.1	11	8.87
3,5-Dimethylphenol	124	0.1	10	8.06
Phenol	124	0.1	10	8.06
1,1,2-Trichloroethane	140	0.5	11	7.86
1,3-Dichlorobenzene	90	0.1	7	7.78
trans-1,2-Dichloroethylene	134	5.0	10	7.46
Cumene	90	0.1		5.56
1,1-Dichloroethane	130	10.0	7	5.38
2,4-Dichlorophenol	124	0.1	6	4.84
3-Chlorophenol	124	0.1	6	4.84
Acenaphthene	124	0.1	6	4.84
2-Ethyltoluene	90	0.1	4	4.44
1,2,4-Trimethyl-benzene	90	0.1	4	4.44
2,4,5-Trichlorophenol	127	0.1	5	3.94
Mesitylene	90	0.1	3	3.33
Phenanthrene	124	0.1	4	3.23
Tribromomethane	130	1.0	4	3.08
1,1,2,2-Tetrachloroethane	124	1.0	3	2.42
2,6-Dichlorophenol	124	0.1	3	2.42
1,2,4-Trichlorobenzene	90	0.1	2	2.22
2-Methyl-naphthalene	90	0.1	2	2.22
3-/4-Ethyltoluene	90	0.1	2	2.22
Fluoranthene	123	0.1	2	1.63
2-Chlorophenol	124	0.1	2	1.61
1,2,3-Trichlorobenzene	90	0.1	1	1.11
1,3,5-Trichlorobenzene	90	0.1	1	1.11
1-Methylnaphthalene	90	0.1	1	1.11
2-Chlorotoluene	90	0.1	1	1.11
Pentachlorobenzene	93	0.1	1	1.08
Fluorene	124	0.1	1	0.81
Pentachlorophenol	124	0.1	1	0.81
Pyrene	124	0.1	1	0.81
Dibromochloromethane	125	1.0	1	0.80
Acenaphthylene	124	0.1	0	0.00
Anthracene	124	0.1	0	0.00
Benz[*a*]anthracene	124	0.1	0	0.00
Benzo[*a*]pyrene	124	0.1	0	0.00
Benzo[*b*]fluoranthene	124	0.1	0	0.00
Chrysene	124	0.1	0	0.00
Hexachlorobenzene	93	0.1	0	0.00

Table 28 shows, in the order of decreasing emission detection frequencies, the organic substances that have been detected in groundwaters affected by old deposits. Those of prime importance are tetrachloroethylene (DF 46.21 %) and trichloroethylene (DF 40.79 %). Altogether only 7 substances have a detection frequency >20 % and 19 substances a detection frequency >10 %. “Only” 49 substances have a detection frequency of >1 %.

5.4.3.3. *Mean Emission Concentration*

A suitable measure of the quantitative influence of a substance is its mean emission concentration. This is the difference between the mean concentration from affected groundwaters and the mean concentration from unaffected groundwaters of a given origin. In the case of organic substances, the small ubiquitous concentration has a negligible influence on the difference, so that here the mean emission concentration is equal to the mean concentration in the affected groundwaters.

Inorganic Substances. The mean emission concentrations of inorganic substances from abandoned sites are listed in Table 29. For better comparison, all concentrations are listed in the unit µg/L and sequenced in the order of decreasing mean emission concentration. Hydrogencarbonate has the highest mean emission concentration (281 mg/L), followed by chloride (169 mg/L), sulfate (135 mg/L), sodium (111 mg/L), and calcium (92 mg/L). In the old deposit (which functions as a “bioreactor”) CO_2 is formed in significant amounts and converted to HCO_3^-. Thus, hydrogencarbonate must be a principal emitted species, just like Na^+ and Cl^-, which are emitted together (both originating from NaCl). Since construction waste rubble makes up a considerable part of the volume of waste, SO_4^{2-} and Ca^{2+} are also among the substances that are emitted most frequently and in high quantities from abandoned waste disposal sites.

Table 29. Inorganic groundwater contaminants from abandoned waste disposal sites listed in the order of decreasing mean emission concentrations

Substance	Mean[+] affected, µg/L	Mean unaffected [341], µg/L	Mean emission concentration, µg/L
Hydrogen-carbonate	519 000	238 000	281 000
Chloride	202 000	33 400	168 600
Sulfate	194 000	59 000	135 000
Sodium	128 000	17 500	110 500
Calcium	175 000	83 200	91 800
Potassium	32 900	2 900	30 000
Magnesium	38 100	14 300	23 800
Nitrate	25 600	13 300	12 300
Ammonium	11 000	170	10 830
Iron	9 600	2 000	7 600
Phosphate	1 800	140	1 660

Manganese	1500	180	1 320
Boron	832	34.8	797
Aluminum	596	50.0	546
Nitrite	490	<10.	480
Strontium	580	240	340
Zinc	410	117	293
Arsenic	45.4	1.2	44.2
Copper	24.5	7.0	17.5
Chromium	21.2	0.58	20.6
Nickel	22.9	2.8	20.1
Lead	6.3	1.4	4.9
Cadmium	1.5	0.13	1.4
Selenium	0.10	0.55	0.
Fluoride	<500.	<100	
Mercury	<0.2	<0.1	

* Mean rejecting maximum value (cf. Table 24).

Organic Substances. To calculate the mean emission concentration of organic substances, the logarithms of the individual concentrations are first taken, because a few measured values are so high that the arithmetic mean values of concentrations before taking their logarithms are meaningless and make comparison impossible. Since logarithms cannot be taken of the concentrations below the detection limit (exact concentrations cannot be determined, but the substance is present in small amounts), only the measured values above the detection limit are included in the mean emission concentration of the organic substances. This procedure is reasonable also because the detection frequencies of most organic substances are very small (see Table 28), which means that a high proportion of measured values are below the detection limit. If these measured values were to be included in the calculation, most mean concentrations would fall below the detection limit.

Table 30 assembles the organic substances in the order of decreasing mean emission concentrations. The list is headed by dichloromethane and a number of other chlorinated and highly volatile compounds, a fact which again confirms the predominance of this group of organic substances in emissions from abandoned sites.

Table 30. Organic groundwater contaminants from abandoned waste disposal sites listed in the order of decreasing mean emission concentrations

Substance	Number of analyses	Number of measured values > relevance limit	Mean of log. concentrations*	Mean emission concentrations, µg/L
Dichloromethane	114	17	2.94	870.96
cis-1,2-Dichloroethylene	153	46	2.37	123.42
Vinylchloride	136	24	1.93	85.11
1,2-Dichloroethane	16	3	1.86	72.44
1,1-Dichloroethane	134	10	1.61	40.74
trans-1,2-Dichloroethylene	130	7	1.64	43.65
Benzene	127	37	1.24	17.38
4-Methylphenol (*p*-Cresol)	124	17	1.08	12.02
Trichloroethylene	277	113	1.06	11.48
3-Chlorophenol	124	6	1.02	10.47
1,1,2-Trichloroethane	140	11	0.88	7.59
Ethylbenzene	124	14	0.82	6.61
o-Xylene	127	12	0.77	5.89
Chlorobenzene	93	12	0.67	4.68
2-Methylphenol (*o*-Cresol)	124	16	0.66	4.57
1,1,2,2-Tetrachloroethane	124	3	0.63	4.27
Tetrachloroethylene	277	128	0.62	4.17
Toluene	127	21	0.61	4.07
3,5-Dimethylphenol	124	10	0.60	3.98
p-/*m*-Xylene	92	21	0.53	3.39
1,4-Dichlorobenzene	90	11	0.53	3.39
Tetrachloromethane	201	29	0.48	3.02
Tribromomethane	130	4	0.39	2.45
Trichloromethane	236	52	0.32	2.10
1,1,1-Trichlorethane	206	47	0.25	1.78
Phenol	124	10	0.24	1.74
1,3-Dichlorobenzene	90	7	0.23	1.70
2,4,5-Trichlorophenol	127	5	0.23	1.70
Pseudocumene	90	4	0.22	1.66
Cumene	90	5	0.20	1.58
Acenaphthene	124	6	0.19	1.55
Naphthalene	124	15	0.16	1.45
Mesitylene	90	3	–0.02	0.95
2,4-Dichlorophenol	124	6	–0.07	0.85
1,2-Dichlorobenzene	90	11	–0.08	0.83
2,6-Dichlorophenol	124	3	–0.10	0.79
2,4,6-Trichlorophenol	124	11	–0.15	0.71
2-Ethyltoluene	90	4	–0.21	0.62
Phenanthrene	124	4	–0.48	0.33
1,2,4-Trichlorobenzene	90	2	<–1	<0.1
2-Methylnaphthalene	90	2	<–1	<0.1
3-/4-Ethyltoluene	90	2	<–1	<0.1
Fluoranthene	123	2	<–1	<0.1
2-Chlorophenol	124	2	<–1	<0.1
1,2,3-Trichlorobenzenel	90	1	<–1	<0.1
1,3,5-Trichlorobenzene	90	1	<–1	<0.1
1-Methylnaphthalene	90	1	<–1	<0.1
2-Chlorotoluene	90	1	<–1	<0.1
Pentachlorobenzene	93	1	<–1	<0.1
Fluorene	124	1	<–1	<0.1
Pentachlorophenol	124	1	<–1	<0.1

Pyrene	124	1	<–1	<0.1
Dibromochloromethane	125	1	<–1	<0.1
Acenaphthylene	124	0	<–1	<0.1
Anthracene	124	0	<–1	<0.1
Benz[*a*]anthracene	124	0	<–1	<0.1
Benzo[*a*]pyrene	124	0	<–1	<0.1
Benzo[*b*]fluoranthene	124	0	<–1	<0.1
Chrysene	124	0	<–1	<0.1
Hexachlorobenzene	93	0	<–1	<1

* Only measured values > relevance limit.

[Top of Page]

6. Measurement and Evaluation of Soil Pollution

Gerhard Knoop, Gerd Crössmann, Thomas Eikmann

6.1. Sampling and Investigation of Soil Pollutants

6.1.1. *Introduction*

Sampling and investigation of samples are integral parts of procedures for the hazard and risk assessment of soil pollution. Soil pollution here means the presence and usually uneven spatial distribution (both horizontally and vertically) of certain substances in soil. These substances either do not occur naturally in these soils or their natural inventory has been enriched to an undesirably high degree by inputs via various exposure paths. The first substance group includes high- and low-volatility halogenated hydrocarbons, aromatic and polycyclic aromatic hydrocarbons, petroleum hydrocarbons, but also organically bound residues of pesticides. Nonessential heavy metals such as lead, cadmium, mercury, and thallium are to be assigned to the second group. Procedures for the detection and recording of soil pollution are combinations of various methods and techniques. The most important individual steps are given in Figure 20.

Figure 20. Individual steps of object-related procedures for detection and identification of soil contaminants

Sampling has a central importance in this process (see also, ⟶ Sample Preparation for Trace Analysis). All other individual steps of investigation processes depend on planning and performing of sampling. Compared with the endeavour and successes in the field of chemical and instrumental analysis, the sampling, but also the sample pretreatment for the preparation of a representative laboratory sample, have up to now been underestimated.

For international harmonization of test standards the ISO has undertaken the task of collecting the national standards for soil investigations, and — after thorough tests, improvements, and adaptations to the current state of analysis — of introducing them into international ISO standard. The Technical Committee “Soil Quality” (ISO TC 190 – Soil Quality) has already submitted standards for numerous test methods, including methods for sampling and sample pretreatment. A basic valuable reference compendium is the German *Handbuch der Bodenuntersuchung* [347] (“Handbook of Soil Investigation”), in which national and international standards, methods and rules in the field of soil investigations have been collated in abridged version in order to update this collection continually.

6.1.2. *Sampling*

Sampling is the first and most important step of soil investigation procedures. Not only the quality of the test data but also decisions that can have far-reaching consequences at many levels (economic, health, and environmental policy) depend on the well-designed planning and careful performing of sampling. The soil sample once taken is seldom reproducible with respect to its chemical and physical composition, i.e., a second sample from the same sampling point may not be identical.

The basis for the development of an object-related sampling strategy is the definition of the objective of the investigation, i.e., the future use of an area, taking into account all site-specific features that should be protected. If knowledge about the nature, position, and extent of soil pollution is lacking or inadequate, a preliminary investigation is required. In the preliminary investigation sampling points are rather randomly selected and smaller in number: both selection and number of sampling points must suffice to delimitate spatially the area under suspicion. Information about previous uses of the land or soil and about material inputs and their paths is very helpful in this phase. If the suspicion of soil pollution is confirmed, the sampling for the detailed investigation has to be planned.

For this, criteria according to DIN ISO 10381–1/4 [348], which are listed in brief below, have to be observed:

1. *Definition of the sampling site*
   The sampling site is delimitated and defined, taking into account the nature of the relevant substances and their lateral spread and vertical distribution. This information has been obtained in the preliminary investigation.
2. *Selection of sampling patterns and sampling points*
   Sampling patterns can be based on
   - Known “hot spots” and substance distribution
   - Distribution along a line
   - Strip-wise distribution
   - No specific estimation of substance distribution

   The design of patterns includes the setting up of single sampling points to complex patterns developed on the basis of statistical and computer-aided methods.
   The following patterns can be taken [348], [349]:
   - Systematic sampling (regular grids), e.g., sampling on a circular grid or triangular grid; sampling along a linear source
   - Nonsystematic (irregular) sampling, e.g., random sampling; stratified random sampling, unaligned random sampling

   The location and the number of sampling points in the site depend on the aim of investigation and the results of preliminary studies.
3. *Selection of the test parameters*
   Depending on the objective, the test parameters are selected in accordance with the results from the preliminary investigation.
4. *Single or composite samples*
   Single samples are necessary during the preliminary investigation to characterize the heterogeneity of the soil, to delimit the contaminated area, and to determine the vertical substance distribution. Single samples are taken from one sampling point. Composite samples are individual samples from different sampling points but from the same sampling depth. It has to be determined for each individual case which single samples shall be combined to give composite samples.
5. *Number of samples*
   The number of samples of one site follows from the results of preliminary investigation and is linked with the objective of the investigation. Because of the general complexity of the sampling, it is advisable to take more single or composite samples than are required for the laboratory investigations. The reserve samples can be investigated subsequently on request.
6. *Sampling from disturbed or undisturbed soil stratification*
   For most laboratory tests for soil pollutants, sampling from disturbed stratification is sufficient. For certain soil-physical and soil-biological parameters, a sampling from undisturbed stratification is necessary. For this, the arrangement of soil prospecting (trial pits) can also be useful.
7. *Sampling equipment*
   Depending on the problem, various equipments or techniques be used, such as hand or power augers, percussion drilling systems, tube borers, but also mobile hydraulic drilling systems. For sampling from undisturbed stratification, metal cylinders and cutting frames (e.g., for sampling of forest soils) are used. DIN ISO 10381–2 [350] contains an assortment of the techniques concerned. Cross contaminations of the samples by abraded metal and by transfering of soil material from sample to sample, must be avoided.
8. *Sampling depth*
   The sampling depth depends on the type of soil use, but can also relate to the particular route of effect for special kinds of acceptors (e.g., children's playgrounds). In cultivated soils, samples are usually taken from the top soil, i.e., to a depth of 20 – 30 cm for arable soils and to a depth of max. 10 cm for pasture soils. In uncultivated soils, samples are usually taken to a soil depth of max. 10 cm. For refilled sites and in analogous cases, sampling by soil layers is appropriate.
9. *Sample quantity*
   The sample quantity per sampling point and layer is determined by the kind and size of the laboratory investigation, the condition of the soil material (e.g., mineral grains, sand, pebble, or foreign materials), and required reserve samples. For chemical analysis 500 g of fine soil should be obtained.
10. *Use of containers*
    The containers must ensure that the sample material is not contaminated by foreign substances and that no losses or changes of material occur by adsorption, volatilization, or degradation processes. For soils containing organic pollutants, glass vessels with tight closures should be used. For preserving soil samples after sampling on the spot, cooled transport containers are useful. They should be used for samples whose constituents to be tested are volatile or degradable. The soil samples can also be temporarily stored cool in these containers for a limited time. The sample should be clearly and unmistakably labeled on both the container and its lid.
11. *Safety precautions*
    The personal protection of the sampling staff also has to be included in the sampling strategy [351].
12. *Sampling report*
    The sampling report should contain the following informations [348]:

Description of site and surrounding area, time of sampling
Cartographic sketch of the area and the sampling points
Execution of the sampling in all details
Deviations from plan and their cause
Delivery to the laboratory

The quality of the sampling and of the reporting is closely related to the sampling staff. It has to be ensured that the sampling, as a decisive quality parameter of the laboratory data, does not put that data in question. Qualification and advanced training of the sampling staff are therefore of utmost importance.

6.1.3. *Sample Pretreatment*

After their entry into the laboratory or after appropriate temporary storage, soil samples are usually pretreated for analysis in various ways, depending on the analytical problem.

Like sampling, the sample pretreatment also is a step of the analytical method, to which little attention often is paid, yet which nevertheless can contribute considerably to the total error of a method. Especially, contamination or losses of materials to be tested must be avoided or minimized. Contamination can arise, e.g., during mechanical size-reduction by grinding devices as well as by the carry-over of soil material. Material losses or changes often occur through inappropriate drying procedures, the effects of light, prolonged periods of storage, or change of storage conditions.

It is advisable to pretreat the whole field sample. The sample is dried sometimes after preceding manual or mechanical preliminary size-reduction with sieves. A mild thermal drying is carried out at up to 40 °C. Freeze-drying is to be preferred for tests on organic compounds (see Section Analysis of Organic Soil Pollutants) as well as on volatile inorganic substances (e.g., mercury). For this, though, the often large initial sample has first to be reduced, because the capacity of the freeze-drying units is often limited.

After drying, the whole sample is mechanically size-reduced. For sample material from the freeze-drying size-reduction is sometimes not necessary.

Afterwards, sample division is advisable if reserve samples have to be formed. This is performed in mechanical sample-dividers, because manual sample division is not without errors.

To prepare the laboratory sample ready for analysis, the size-reduced soil material is sieved to obtain different grain-size fractions. In general, fractions with a grain size < 1 mm are used for the laboratory sample. Other grain fractions are, however, also customary. Sieving to definite grain sizes is performed to obtain samples of the same chemical and physical composition, and also for improving the extraction yield in the analysis.

During the storage of dried soil samples, the soil reabsorbs water, depending on the room temperature and humidity. The water content of so-called air-dried samples is usually about 3 – 5 %. In individual cases a further drying can be required. An DIN ISO standard exists for the handling of samples [352]. The laboratory is in charge of the correct care during sample pretreatment. Documentation of the laboratory operations is mandatory for reasons of quality assurance.

6.1.4. *Sample Preparation*

The sample preparation is usually the first individual step of an analytical method and an important part of combined analytical methods. Its objective is

1. To separate the elements/compounds to be tested from interfering matrix constituents and to convert them into a form accessible to the measurement (e.g., aqueous solution)
2. To render low substance concentrations (ng/g or lower) detectable or determinable by preconcentration steps

In the case of instrumental direct methods (e.g., X-ray fluorescence), this preparation is omitted; at the most, melt digestion is used. The relatively low power of detection of direct methods is disadvantageous [353]. Wet-chemical combined methods, on the other hand, enable the use of various measuring systems as well as simple calibration by aqueous standard solutions (⟶ Trace Analysis – Methods of Modern Trace Analysis).

Systematic sources of error can originate from contamination of equipment, vessels, and reagents (blank values), as well as from the working environment (laboratory air). Material losses occur through volatilization and adsorption.

The digestion techniques are becoming increasingly powerful, because of their degree of automation, ensuring optimal decomposition conditions. Consequently they are more economical than previous techniques, e.g., because of higher sample throughput. In principle, the following modes of processing can be used:

1. Digestion
2. Extraction
3. Elution

In *inorganic analysis* the terms digestion and extraction are frequently used synonymously. According to the recent ISO methods, *digestion* means the total digestion of soils, but also of geological samples with the aid of strong acids (mixtures). The digestion yield should amount to 100 %.

The term *extraction* means that relatively weak acids, acid mixtures, complexing agents, neutral salt solutions, etc. are applied. Extraction is used when relatively low material yields can be tolerated or a certain element fraction or form of element binding (e.g., plant-available compounds, exchangeable cations) has to be determined selectively. Environmentally relevant material mobilities in soils can be characterized approximately by various extractants. Figure 21 shows the corresponding relationships. Extraction is usually used for the detection of organic substances in soils.

Figure 21. Relations between the state of mobility of heavy metals in soils and state-related digestion or extraction solutions [354]

DTPA = diethylenetriaminepentaacetic acid; EDTA = ethylenediaminetetraacetic acid

The *elution* of soil contaminants is frequently carried out from the standpoint of groundwater management. However, elution with water, which is still practiced today, is very controversial, since the significance of the results can fluctuate greatly owing to considerable influences from the soil matrix.

For sample preparation for the detection of *inorganic substances* in soils, high-performance digestion techniques are available today that enable an efficient, low-cost, environmentally compatible, and largely low-error operation:

1. Wet digestion (extraction) by means of open systems with convective heat transfer (heating block, reflux) [354]; closed high-pressure digestion techniques [355]; or microwave technique [356], [357]
2. Dry decomposition [358] by means of combustion systems such as pressureless decomposition in open systems [flask combustion (e.g., Schöniger digestion), Wickbold combustion with oxyhydrogen gas flame], cold-plasma ashing and dry ashing or melt digestion.

Of the wet-chemical digestion techniques, the microwave technique is becoming increasingly important. It is time-saving, environmentally compatible and, with regard to the reproducibility of the material yield, often more favorable than, e.g., pressure digestion techniques.

For the determination of *organic compounds* in soils, there are standard procedures for the preparation of measuring solutions. The separation processes are based on partitioning processes between phases, which obey physicochemical laws. For organic soil pollutants the following technical processes are available in routine analysis [359], [360]:

1. Head-space processes (⟶ Gas Chromatography – Static and Dynamic Head Space Analysis) for highly volatile halogenated hydrocarbons and aromatic hydrocarbons (benzene, toluene, xylenes) that are also suitable for difficultly vaporizable and polar compounds after converting them to suitable derivatives
2. Soxhlet extraction for low-volatility halogenated hydrocarbons
3. Extraction with supercritical fluids (SFE) as a relatively new process route for the detection of PAHs, PCBs, and pesticides in soils

In individual cases, preconcentration steps are sensible and necessary in order to determine the true substance concentration (e.g., preconcentration by ion exchange or in cartridges).

6.1.5. *Measurement*

Inorganic Substances. Inorganic soil pollutants, which are found matrix-free and without interfering accompanying substances in the measuring solution, can be determined by different instrumental methods. Of these, atomic absorption spectrometry (AAS, ⟶ Atomic Spectroscopy – Atomic Absorption Spectrometry) and its various special techniques are used most widely:

1. Flame technique for alkali metals and heavy metals (F – AAS)
2. Graphite furnace technique (GF – AAS) for determining elements in the lower ng/g range
3. Hydride technique for hydride-forming elements such as As, Se, or Sb
4. Cold vapor technique for mercury

All of the techniques are used widely in soil and environmental analysis, depending on the kind of element and its concentration and on the desired sensitivities of detection. Correction of the nonspecific background by deuterium or Zeeman compensation leads to improvement of the data in the trace and ultratrace range [361].

Atomic emission spectroscopy with inductively coupled plasma impulse (AES – ICP, ⟶ Atomic Spectroscopy – Inductively Coupled Plasmas) is already widely used in the routine

analysis of inorganic substances in soils because of its great technical advantages. The important advantages include high measurement frequencies and the simultaneous determination of several elements. Disadvantages are at present relatively low sensitivities of detection. However, measurement of most relevant elements (with the exception of cadmium, mercury, and thallium) is possible because of their relatively high geogenic background contents. With AES – ICP, survey analyses for a wide range of elements in unknown samples also can be carried out.

The flow injection analysis (FIA) is applied with the mentioned AAS systems to improve analytical and economical aspects by automation of, e.g., sample introduction, dilution or preconcentration, or addition of internal standards [361].

For element measurement other techniques exist, not or scarcely inferior in their specificity and sensitivity of measurement to the AAS methods, but less widely used in routine analysis for the detection of soil pollutants. They include electrochemical methods (e.g., polarography, voltammetry) and molecular spectroscopic methods (e.g., UV/VIS spectroscopy). Multielement determinations also are possible with X-ray fluorescence spectroscopy after melt digestion.

Organic Compounds. The determination of organic compounds in polluted soil samples is performed according to the following process steps [359]:

1. Chromatographic separation
2. Detection
3. Evaluation

The *vapor-phase separation* of organic substances is today carried out mainly in capillary columns, usually with high resolution and sensitivity. Substance-specific separating columns are used, e.g., for the analysis of PAHs and PCBs. Aside from the “miniaturization” of separating columns, further advances in the separation of substances are emerging in connection with the development and adaptation of microchips and chemosensors [360].

Many substance- and structure-specific detectors are suitable for the detection of organic compounds. Examples are flame-ionization detectors (FID), thermionic detectors (TID), or electron-capture detectors (ECD) (⟶ Gas Chromatography – Detectors). The coupling of gas chromatography with mass spectrometry (GC – MS) offers an improved sensitivity of detection as well as a reliable identification of individual compounds or substances.

High-performance liquid chromatography (HPLC) has developed into one of the important and efficient separation techniques for environmental analysis [362]. This technique enables the determination of difficultly vaporizable, thermally labile, or ionogenic substances in extracts or eluates. Columns both for general uses (e.g., in reversed-phase chromatography) and for special applications (e.g., for the determination of PAHs) have a high degree of standardization with good reproducibility. Improvements in detection are possible with the photodiode array detector, which permits simultaneous measurement at different wavelengths. Coupling techniques with the mass spectrometer (HPLC – MS) having the same effects as those of GC – MS are awaited.

6.1.6. *Evaluation*

Instrumental measuring systems are generally equipped with microprocessors and computers. The equipment is used for controlling apparatus and operating parameters as well as the entire procedure from the input of the sample up to the documentation of the measurement signals. Editing, processing, and documentation of the data — as far as to adoption in dispatchable test certificates and in data banks — become more and more computerized. Interfaces allow the individual measuring systems to be included in networks such as the laboratory information and management system (LIMS). This makes it possible to record all information on sampling and sample pretreatment.

6.1.7. *Quality Assurance*

Analytical data are basically subject to error. Systematic and random errors result from the various process steps, from sampling to measurement. While errors during the individual analytical steps can more easily be identified, minimized, or eliminated, little is yet known about sources of error in sampling and sample pretreatment and their proportion in the total error, or the analyst may often not be conscious of them.

The inhomogeneity of contaminated soil samples is unquestionably the most frequent cause of a sometimes high systematic error. Therefore, as already explained in Section Sampling, the sampling on the spot has to be carried out very carefully. In all later steps — from the immediate storage in containers on the spot, via transport, to the conclusion of the sample pretreatment — the minimization or elimination of errors is more easy if the sampling standard operation procedures (SOP) are exact and obeyed. The documentation of the sampling in all its details is very important here.

Features of the procedure such as selectivity, accuracy, and precision as well as limits of detection and determination are indispensable for the statistical evaluation of analytical procedures [363]. Often they have still to be worked out.

Quality assurance systems should comprise the following main issues:

Sample and laboratory management
Quality assurance manual according to DIN EN ISO 17 025
Participation in round robin tests
Staff qualification
Maintenance of equipment
Sample handling
Methodology
Statistic control
Recording and reporting of data

The realization of these requirements is the basis for the competence and reliability of the laboratory carrying out the test. It is therefore recommendable to entrust with the investigation of soils only such test laboratories as meet all conditions mentioned.

6.1.8. *Analytical Procedures for Selected Inorganic and Organic Soil Pollutants*

From the multiplicity of analytical procedures for determining soil pollutants in this section detailed procedures are described for the elements cadmium and thallium as well as for PAHs and petroleum hydrocarbons. These are analytical procedures that are already far advanced in their degree of standardization — thus the extraction of trace elements soluble in aqua regia (DIN ISO 11 466), the measurement of cadmium and other elements by F – AAS and GF – AAS (DIN ISO 11 047).

First drafts for national or international standardization are available for the extraction of mobile trace elements from soils with ammonium nitrate solution (DIN 19 730) and for petroleum hydrocarbons (ISO TR 11 046).

6.1.8.1. *Analysis of Cadmium and Thallium*

Both heavy metals occur ubiquitously in soils, although cadmium is more widespread than thallium, which has a more local importance. They are nonessential trace substances which, however, have toxicological relevance because of their mobility behavior in soil – plant transfer and their enrichment capacity in food and forage plants. Previously only the so-called total content after extraction with aqua regia or equivalent acid mixtures was used as the basis for evaluation, this is today supplemented by the determination of that part of the total content which is mobile, i.e., short- or mediumterm available by plants or translocatable by percolate water into deeper soil layers (leaching).

6.1.8.1.1. Sample Preparation

According to DIN ISO 11 466 *cadmium* is extracted from air-dried fine soil (grain size <0.1 mm) with aqua regia under reflux for a mean duration of ca. 2 h. The supernatant suspension (filtrate) in the measuring flask after sedimentation of the undissolved fraction is ready for measurement. A detailed description is given in [364].

Owing to the formation of volatile thallium compounds, HNO_3 and H_2O_2 are used as an extraction mixture for the determination of *thallium* [365]. Pressure digestion systems should be used, with closed vessels of PTFE (polytetrafluoroethylene), PFA (perfluoroalkoxy resin) or other suitable materials and with thermal heating devices or with microwave heating systems. Undissolved substances must immediately be removed from the extraction solution, since otherwise deficiency of the true concentration is possible.

The extraction with *ammonium nitrate* detects nearly all environmentally relevant trace elements, including cadmium and thallium, in one solution. The mobile element fractions, i.e., those available or leachable in a short time, are determined. Extraction is carried out by shaking overhead at room temperature. After the filtration, the extract is stabilized with HNO_3.

6.1.8.1.2. Measurement

In instrumental analysis, the relevant heavy metals are determined with, among others, F – AAS, electrothermal AAS (ET – AAS), and AES – ICP, which are part of the basic equipment of laboratories working in environmental analysis. Other measuring systems will consequently not be discussed.

Flame AAS and electrothermal AAS are described in detail in DIN ISO 11 047. They differ above all in the sensitivity of detection for the elements named therein. Thus with expected cadmium contents >2 mg/kg soil, the use of F – AAS is recommendable and with contents <2 mg/kg soil that of ET – AAS.

Because the thallium contents in soils usually are very low, ET – AAS (graphite tube furnace) and appropriate background compensation (e.g., Zeeman method) are recommended for the measurement [365].

The same instrumental systems can be used for the measurement of the heavy metal contents from the ammonium nitrate solution. Since the elements often occur in very low concentrations in these extracts, measurement with GF – AAS is advisable. At higher assumed contents, F – AAS also can be used. Efforts to determine cadmium concentrations also with AES – ICP have at present not yet reached the routine stage, because here the sensitivity is still unsatisfactory.

Taking into account instrumental equipment, extraction agents, and laboratory staff, the following limits of determination specific to different procedures can be assumed (Table 31)

[347].

Table 31. Limits of detection of different AAS methods with respect to cadmium and thallium [347]

Element	F – AAS, mg/L	GF – AAS, µg/L	AES – ICP, µg/L
Cadmium	0.02	0.1	10
Thallium	0.2	2	100

Information is at present still being compiled on procedural characteristics, such as precision and accuracy.

6.1.8.2. *Analysis of Organic Soil Pollutants*

In this chapter the analytical methods for the determination of two typical organic soil pollutants, PAHs and petroleum hydrocarbons are presented.

6.1.8.2.1. Analysis of Polycyclic Aromatic Hydrocarbons

Polycyclic aromatic hydrocarbons occur ubiquitously in the environment (see Section Polycyclic Aromatic Hydrocarbons (PAHs)). While previously their main anthropogenic source originated from coal combustion and coal processing, today they are mainly emitted by scattered sources, such as motor vehicle traffic, or domestic fuel.

The EPA has stipulated 16 PAHs — some of which are suspected of being carcinogenic — in its “priority pollutant list” for monitoring and restriction (see below):

Naphthalene	($C_{10}H_8$)
Acenaphthylene	($C_{12}H_{10}$)
Acenaphthene	($C_{12}H_{10}$)
Fluorene	($C_{13}H_{10}$)
Phenanthrene	($C_{14}H_{10}$)
Anthracene	($C_{14}H_{10}$)
Fluoranthene	($C_{16}H_{10}$)
Pyrene	($C_{16}H_{10}$)
Benz[*a*]anthracene	($C_{18}H_{12}$)
Chrysene	($C_{18}H_{12}$)
Benzo[*k*]fluoranthene	($C_{20}H_{12}$)
Benzo[*b*]fluoranthene	($C_{20}H_{12}$)
Benzo[*a*]pyrene	($C_{20}H_{12}$)
Indeno[1,2,3-*cd*]pyrene	($C_{22}H_{12}$)
Dibenzo[*a,h*]anthracene	($C_{22}H_{14}$)
Benzo[*g,h,i*]perylene	($C_{22}H_{12}$)

After extraction with an organic solvent PAHs are determined by capillary gas chromatography with mass-selective detector (quadrupole or ion trap). In addition HPLC with fluorescence detection (FLD) is becoming increasingly established [347], [368-371].

The ISO has published in the Technical Committee 190 “Soil Quality” a draft (ISO/CD 13 877) on the analysis of the PAHs [372].

Samples from lightly loaded and heavily loaded soils are prepared differently. *Lightly loaded soils* are extracted with acetone. The extracts are purified by mixing with cyclohexane, extraction of acetone with water, and subsequently purification via an aluminum oxide column.

The *heavily loaded soils* are extracted with toluene in a Soxhlet apparatus, and the PAHs measured, without further purification, with an HPLC – FLD or diode array detector (DAD) (⟶ Liquid Chromatography).

In the case of lightly loaded soils the purification is labor-intensive, while in the second method, only one Soxhlet extraction taking 4 – 8 h is required.

A HPLC method [373] has been developed to improve the efficiency of the extraction process, its reproducibility, practicability, profitability, industrial safety, and environmental compatibility. In this method an ultrasonic extraction is performed with tetrahydrofuran or acetonitrile, and a purification carried out on hydrophilized silica gel [373].

Samples for gas chromatography (GC) are subjected to a Soxhlet extraction with toluene. Purification is carried out on silica gel or aluminum oxide.

Because of the sample preparation by air drying which is generally used, the measurements for low-molecular PAHs are unreliable. This applies particularly to the four easily volatile components of the EPA list, naphthalene, acenaphthylene, acenaphthene, and fluorene. To avoid these possible losses, the mild freeze-drying method should be used preferably [374]. In this technique the samples are dried at ca. –40 °C under vacuum.

Other laboratories use a cold acetone – cyclohexane mixture for well homogenized, fieldfresh samples. The measurements are made with HPLC – FLD or – DAD, or GC – MS. This method was developed and validated in collaboration with state and private laboratories for the analysis of soil, waste, and contaminated site samples [375].

Immunoassay test methods have also been offered as screening processes [376].

The use of SFE is also customary for soil samples (see Section Sample Preparation). In this, supercritical carbon dioxide is passed through the sample at high pressure. The carbon dioxide is subsequently depressurized and collected in a solvent receiver. The advantage of this method is the low solvent requirement and short extraction times. An obstacle to routine application is the different extraction yields under the same process conditions but different soil samples. An optimized method must be developed here for each type of soil [377-379].

6.1.8.2.2. Petroleum Hydrocarbons

The risk of pollution exists anywhere where petroleum hydrocarbons are processed, stored, reloaded, transported, or used [380].

In some countries standards and draft standards exist for the determination of petroleum hydrocarbons in soils [381], [382].

In the extraction of petroleum hydrocarbons since 1981, carbon tetrachloride (which had usually been used up to that time) has been dispensed with for reasons of industrial safety in favor of 1,1,2-trichlorofluoroethane. However, the use of chlorofluorocarbons (CFCs) is controversial because of their ozone-depleting effects in the atmosphere. Whether CFCs will be available in the future for analytical purposes is therefore uncertain.

For quantitative determination of petroleum hydrocarbons an *infrared spectrum* is recorded from wave number 3125 cm^{-1} to 2800 cm^{-1}. The CH_2 absorption band, the CH_3 band at 2925 cm^{-1}, the CH_3 band at 2958 cm^{-1}, and the aromatic CH absorption at 3030 cm^{-1} are used for quantification.

An alternative to this is offered by a *GC – FID method*, whereby the total area of the peaks from decane ($C_{10}H_{22}$) to *n*-tetracontane ($C_{40}H_{82}$) is evaluated as petroleum hydrocarbons content.

The ISO TC 190 “Soil Quality” has submitted a draft technical report [ISO/TR 11 046: 1994 (E)] [382]. The procedure recommends a shaking method or a Soxhlet extraction with 1,1,2-trichlorofluoroethane. The “clean-up” step uses a column filled with magnesium silicate or aluminum oxide.

The validity of the aforementioned GC method has been checked by an interlaboratory comparison. According to this the difficulty of the procedure is in the external calibration with a petroleum standard whose petroleum hydrocarbon distribution does not in every case agree with that of the soil sample. Sample injection is the largest source of error in the procedure. Split or splitless injection (⟶ Gas Chromatography – Split Injection) techniques should not be used in any case. The results determined with the GC method should be supported by means of IR [383].

With the draft method 3560/8440 [384] the EPA has proposed an extraction with *supercritical carbon dioxide*. The sample is dried with magnesium sulfate before the extraction. The SFE is carried out at 80 °C and 34 MPa. Tetrachloroethylene is used as collection solvent. The “clean-up” is carried out with a silica gel column or a solid-phase cartridge. The SFE

method is equivalent to a Soxhlet extraction in respect of reproducibility and recovery.

Accelerated solvent extraction (ASE) is a newer development. Conventional solvents are used at up to 200 °C. In order to maintain the solvent in the liquid state above its atmospheric boiling point, the extraction is carried out at 0.3 – 20 MPa in pressure-tight tubes [385], [386].

Outlook. In soil analysis, the development of digestion procedures, extraction procedures, and methods of measurement is aimed at efficiency, as much automation as possible, and at little solvent consumption.

Soil sampling remains the most relevant area for accuracy because it is subject to relatively large errors. Inaccuracy of the analytical values caused by the sampling can only to a limited extent be compensated by improving the analytical procedure [387].

6.2. Evaluation Concept

6.2.1. *Uptake Paths to Humans*

The hazard potential of pollutant-containing soils to humans is normally assessed from a number of different basic conditions. In the foreground — aside from the investigation of the nature and concentration of the pollutants in the soil — are principally the uptake paths (possibilities of transfer) by which pollutants can enter the human body. In addition the present use of the soil areas is important, because conclusions can be drawn from this on predominant uptake paths or special exposure situations. The relevance of the pollutant contents of the soil can only be assessed by considering all these parameters. On the other hand, the evaluation of pollutant concentrations on the basis of clinical-toxicological data alone without reference to the soil environment must be rejected as inadmissible [388].

6.2.1.1. *Inhalation Uptake*

Nonvolatile Pollutants. The uptake of particle-bound, nonvolatile pollutants is always assumed when fairly large parts or even the whole of a contaminated site have no vegetation or other kind of cover (e.g., asphalt). The lack of vegetation is observed relatively frequently on foundry heaps in which high concentrations of phytotoxic metals are present [388]. Here dust can be blown up and subsequently inhaled by human beings.

The uptake by inhalation depends on the concentration, exposure time, substance-dependent toxic effect, and especially on the particle size. The fine dust (particles with an aerodynamic diameter of 1 – 7 µm) which can pass into the alveoli must be regarded as especially biologically relevant. These dust particles reach as far as the pulmonary vesicles. From there easily absorbable toxic agents can pass directly into the bloodstream.

It can be presumed from literature data that hazardous effects on humans as a result of the blowing-up of dusts are encountered only very rarely [389]. Effects on humans have to be assumed chiefly when there is an active release of dusts from contaminated material, e.g., during activities on sports grounds with surface made of blast-furnace ash or on leisure grounds in the vicinity of contaminated sites. In dry weather, these activities can result in severe dust formation. This leads to a potential hazard not only for the participants themselves, but also for spectators or neighbors [390].

Volatile Pollutants. The uptake of volatile pollutants from gas emanating from the soil is considerably more important than the inhalation of particles. In the case of volatile pollutants, a genuine exposure of residents and users of contaminated sites and people living nearby is frequently demonstrable [391].

The exposure to gases or vapors from contaminated sites is possible when, for example, a building is constructed directly on the contaminated soil. Taking into account that residents stay up to 24 h a day in the polluted area, a considerable uptake of pollutants is certainly possible. But even people living near contaminated sites can be at considerable risk as a result of the migration of gases and vapors beyond the immediate area of the contaminated site. However, the concentrations of the volatile pollutants are usually less in the neighborhood than directly on the contaminated site itself. The possible accumulation of gas in indoor air is particularly relevant. The occurrence of odor nuisances and/or the detection of methane and carbon dioxide in cellars of buildings on contaminated sites is always a indication of this risk [392].

6.2.1.2. *Oral Uptake*

The risk of oral uptake of pollutants from the soil appears to be greatest for young children. Many authors report an oral uptake, above all of heavy metals by hand-to-mouth activity, but give no information on the amount of dust or soil thereby ingested. Estimates in the literature are in a range between 200 mg and 10 000 mg of soil per day. Generally it is assumed that children at play take up 200 mg to 1000 mg of soil per day. Taking this amount as the basis for children aged three to five years, the uptake of pollutants via the soil plays an important part for this group when compared with the other contamination paths and must definitely be taken into account in the estimates of effect.

The oral exposure for adults originates in most cases from the uptake of contaminated edible plants from backyards or garden plots. The uptake of heavy metals by plants has been proved in a large number of studies. As a result, the agricultural use of lands in the neighborhood of heavy-metal emitters has been restricted. Aside from the uptake of pollutants present in the soil via the roots, additional contamination of the plants can occur by deposition of pollutant-containing dust.

The direct oral uptake of pollutants by adults from the soil of backyards and garden plots can be rated as negligible. The danger to small children is certainly considerably less than that of playgrounds, because the use of backyards and garden plots, especially for the cultivation of fruit and vegetables on patches, usually allows no play activities that could lead to noteworthy contamination.

Another exposure path for oral uptake of pollutants can be contamination of groundwater, which frequently occurs at contaminated sites. However, the problem is not the public supply of drinking water of adequate quality, but rather the possible use of contaminated groundwater from domestic wells [390].

6.2.1.3. *Uptake via the Skin*

The contamination of the skin has to be rated as rather unimportant. The theoretically possible uptake of fat-soluble substances or solutions by direct contact with the skin requires very high pollutant concentrations, such as are not normally found on the surface of contaminated sites or other contaminated soils. The absorption of gas or vapor via the skin must be rated as almost impossible even under extreme exposure conditions and can therefore be ruled out in this context [389].

6.2.2. *Relevant Soil Uses*

To enable an evaluation of soil contamination with regard to the health hazard for persons concerned and for the establishment of tolerable soil values, it is reasonable to reduce the huge number of possible uses of the soil to a small number of use scenarios. Starting points could be on the one hand particular user groups (playing children or young people, sportsmen and sportswomen, garden plot holders, etc.) and on the other the complex uptake paths of nonvolatile or semivolatile substances (predominantly oral uptake, mixed inhalation and oral uptake, etc.). It does not appear sensible to consider all possible transfer modes from the soil to humans in detail but rather to use simple models and scenarios as a basis for a general assessment. Table 32 shows use scenarios according to the assessment system of EIKMANN and KLOKE with parameters characterizing use and protected subjects [393].

Table 32. Use criteria for various types of soil use (objective: protection of humans)

Type of use	User group/ protected subject	Uptake path	Soil areas	Soil depth
Children's playgrounds	small children and escorts	oral	sand pits vegetation-free environment	35 cm
Backyards and garden plots	children adults	oral by inhalation	patches low-vegetation areas	35 cm
Sports and recreation grounds	sportsmen and sportswomen young people	 by inhalation	gravel surfaces vegetation-free areas	10 cm
Parks and leisure grounds	adults children	oral by inhalation	unpaved, low-vegetation areas	10 cm
Industrial and commercial areas	adults of working age	by inhalation water path	unpaved, vegetation-free areas	10 cm

6.2.3. *Evaluation Criteria*

In the development of regulative values or values for the soil, there is always a need to establish conventions, for example, for the comparability of measurement methods and of measurement results. Although these conventions are certainly based on scientifically supported knowledge, they are nonetheless to be regarded within a certain tolerance range as “arbitrary”.

The calculation of tolerable soil values depends to a considerable degree on the particular conventions chosen. The influence of these conventions on the quantification of the uptake is clearly greater than that of toxicologically derived criteria (body loads). Moreover the tolerable soil values calculated as a function of the chosen assumptions on exposure are frequently far above or far below the usual background concentrations of these substances in soils.

In the assessment methodology, owing to the considerable uncertainties and difficulties which have been described, there is a trend of describing the exposure — especially of persons on contaminated soils — by characteristic simplified individual scenarios for types of use and protected property and of establishing a quantitative assessment on this basis or to orient oneself according to the main uptake path. Thus the base values (BW I) stated by EIKMANN and KLOKE for (harmful) substances in soils are derived essentially from the corresponding background concentrations in the soils in Germany. The tolerance (test) values (BW II) and intervention values (BW III) are defined as a function of the toxic properties of the substances involved and the specific exposure situation (inhalation) by application of appropriate factors [393].

Other evaluation models (e.g., the EPA model or the Holland List) attempt to avoid these methodical difficulties. Here, either the risk is estimated by using evaluation models referring to individual cases without defining soil values; or for defining acceptable soil values, only difficultly understandable correction factors or so-called reliability scores are applied. In addition, all estimation parameters used in these models (e.g., according to information from the authors of the Holland List) have uncertainties that frequently can amount to a factor of 10 or more [394].

The repeated demand for “strictly toxicologically derived” soil values, taking into account the various kinds of soil use, cannot be satisfied owing to the paramount influence of the conventions chosen for quantifying the (inhalation) uptake of pollutants. When different evaluation models are applied in the context of an individual-case assessment according to the classical derivation model, this can result in the definition of different soil concentrations for a certain pollutant. However, different intervention or remediation target values, depending on the particular contaminated site, are certainly unacceptable. In particular the frequently existing necessity for immediate action — in relation to a further possible use of contaminated soil areas — requires at least the definition of (standardized) intervention values. In addition, different remediation target values for the same kind of uses will certainly not be accepted by the public because of the incomprehensibly higher remediation costs in single cases.

In the preparation of balances for the transfer of pollutants from the environment to humans the pollutant contaminations of the various environmental media (atmospheric and indoor air, drinking water, and food) are considered. No attention is usually paid to the direct exposure from the soil, because it must be rated unimportant in relation to the total uptake. The indirect uptake from the soil via food plants and animals is normally taken into account by recording the contamination of the food [395].

The direct intake of pollutants from the soil — which is frequently considered the main problem of contaminated sites — is a very specific pollution problem for only part of the population, and therefore also cannot be considered in a general balance. If tolerable body doses are reached or exceeded as a result of the general environmental pollution with particular pollutants, such as heavy metals, in theory also a reduction of the soil contamination can be considered to lower the metals content in the environment. However, because of the practically nonexistent discharge from soil and the nondegradability of metals, a reduction cannot normally be achieved here, by contrast, e.g., with air (by reduction of emissions). An inclusion of the soil in such balances as an equivalent medium appears therefore to be unjustified and leads to an assessment dead-end [396].

Extensive contamination of the soil by heavy metals (caused by immissions) frequently occurs, especially in the conurbations, and, in the use of garden plots or backyards, results in (officially recommended) restrictions of the cultivation of various useful plants. This is one of several possibilities of reacting to such wide-ranging soil contamination. On the other hand, an extensive remediation of the soils of whole regions to achieve tolerable heavy metal concentrations must certainly be regarded as difficult, if not quite impossible. For the purposes of general soil protection (above all also for the persistent organic compounds), any further input must rather be kept as small as possible.

[Top of Page]

7. Remediation of Soil and Land

Volker Franzius, Detlef Grimsky, Norbert Möhlenbruch, Ulf Dworshak

7.1. Remediation Methods

7.1.1. *Introduction*

The effects of point sources polluting soil and groundwater are particularly obvious when regarding the problem of contaminated sites. Based on the importance of this problem during the 1980s, administration, industry and commerce have equipped themselves with strategies and techniques to solve it. The methods for soil purification described in this section mainly relate to the remediation of contaminated sites. Application of remediation processes to the treatment of extended soil contaminations caused by airborne emissions or agricultural measures will not be treated.

As of October 31, 1998 about 238 000 suspected contaminated sites were registered in Germany (Table 33), divided into about 100 000 abandoned waste disposal sites and about 138 000 abandoned industrial sites. In addition, registration has been carried out for about 34 000 suspected contaminated sites on 1026 military properties, that had been used by the former West Group of the Soviet Troops (WGT) in East Germany and 3240 suspected former armament sites (status 1995). Altogether about 10 000 properties used by the military are being examined for suspected contaminated sites. The number of actual contaminated sites is uncertain because the risk assessment which follows the registration has not yet been completed for the whole area of Germany. On a cautious forecast, about 10 % of suspected contaminated sites are expected to be contaminated and to require some remediation. As a rule, suitable remediation measures for individual cases can be put into concrete terms only after risk assessment.

Table 33. Suspected contaminated sites in Germany*

Berlin	673	5541	6214
Brandenburg	5585	8580	14 165
Bremen	105	4000	4105
Hamburg	460	1701	2161
Hesse	198	160	358
Mecklenburg-Westpommerania	4332	7462	11 794
Lower Saxony	8957	n.a.	8957
North Rheine-Westphalia	17 155	14 874	32 029
Rheinland-Palatinate	10 578	n.a.	10 578
Saarland	1801	2442	4243
Saxony	9382	22 197	31 579
Saxony-Anhalt	6936	13 295	20 231
Schleswig-Holstein	3076	14 497	17 573
Thuringia	6192	12 368	18 560
Germany total	**100 229**	**137 798**	**26 038**

*The figures reflect the status of registered suspected contaminated sites in the Federal States. The individual figures are not comparable to each other since the Federal States do not use compatible definitions for the terms of suspected contaminated sites and abandoned industrial sites (for example, in Hesse only sites with proved contaminations are included).

7.1.2. *Methods and Processes for Remediation*

7.1.2.1. *Overview*

The final objective of any remediation measure is the elimination of risks and dangers that may arise from soil pollution when using the site for specific purposes. Principally, this can be achieved by the treatment of soil or groundwater to remove the pollutants or to reduce their concentration (decontamination measure); or by containment or fixation of the pollutants to interrupt or reduce the spreading of pollutants towards the relevant targets (safeguarding measure). Both types of remediation methods are applied and they are specifically defined as remediation measures in the German Soil Protection Act.

Safeguarding technologies were originally designed, developed, and extensively used for the remediation of abandoned waste disposal sites. However, in the 1990s they became more and more important also for the remediation of abandoned industrial sites.

Safeguarding measures include:

Surface sealing,
Vertical sealing (slurry wall),
Immobilization measures,
Passive hydraulic and pneumatic measures

Decontamination measures for soil include:

Thermal soil treatment,
Physicochemical soil treatment,
Microbiological soil treatment,
Soil vapor extraction.

Furthermore several *groundwater remediation* measures are applied. They are generally based on well known technologies used in the treatment of municipal or industrial sewage. For groundwater these measures are mainly performed aboveground after pumping (pump and treat). A presently intensive discussed, developed, and increasingly applied technology for the purification of groundwater is the placement of absorptive barriers into the groundwater stream, in order to remove the contaminants by absorption. These absorptive barrier wall systems are applied with and without additional hydraulic measures and work in two ways. On the one hand they act as a less permeable vertical barrier that prevents the contaminants from spreading into the broader groundwater regime, on the other hand they serve as a filter that slowly removes the contaminants from the groundwater by absorption. The advantage of the process is its cost effectiveness. The disadvantage is the very long period of time that is necessary for the purification of groundwater below the required concentration.

Soil decontamination can be principally applied in three ways:

1. Treatment at the natural location of the contamination without excavation of soil (in situ)

2. Treatment after soil excavation or groundwater pumping at the site (on-site)
3. Treatment after excavation and transportation to another site (off-site)

In situ soil treatment has the advantage that the soil compartment remains in its natural location and an excavation and subsequent backfilling is not necessary. Under this aspect, in situ measures are regarded as rather inexpensive compared to on-site/off-site technologies. In the practical application, however, the natural conditions at a specific site must be taken into account. The lower cost factor is certainly true for cases with homogeneous soil conditions and a limited pollutant spectrum. In the majority of cases with inhomogeneous subground conditions and a wide spectrum of pollutants, the effectiveness of in situ processes is restricted. In practice therefore, in situ applications are limited primarily to soil vapor extraction and with restrictions also to microbiological in situ treatments.

Further promising options for in situ soil treatment are in the research and development. These approaches for in situ thermal treatment aim at the vaporization of volatile organic compounds by in situ heating by high-frequency energy or hot air injection. The soil vapor is then collected at the surface and led to a surface treatment facility. Such techniques have been applied in pilot scale with quite promising results. However, they are rather in an outsider position compared to available and best demonstrated techniques and do not play a significant role on the remediation market. The same is true for an in situ soil washing concept that has been successfully applied on a large scale in Germany. The technology involved the high pressure injection of water into the soil, effecting the separation of pollutants from the soil particles in situ. On-site and off-site treatment is the standard procedure of soil decontamination now in Germany, whereas off-site treatment in stationary facilities is the main method.

Since the beginning of the 1990s the number of stationary soil treatment plants has tremendously increased (Fig. 22). In 1990, 22 soil treatment plants were either in operation or in planning. In September 1993 the number of operating treatment plants increased to 47 facilities with a total treatment capacity of 1.25×10^6 t/a (1 thermal treatment plant, 12 soil washing plants, and 34 biological treatment plants). In 1995 the number of operated soil treatment plants was 95 (3 thermal treatment plants, 23 soil washing plants, and 69 biological treatment plants). The total available treatment capacity was about 2.5×10^6 t/a. This figure already exceeded the request by the customers. According to information of the operating companies, the average degree of usage was only about 50 %.

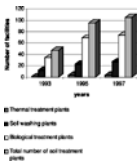

Figure 22. Development of soil treatment plants in Germany

The latest figure on soil treatment facilities and capacities dated October 1997 [398]. At that time, 104 soil treatment facilities were offering their services all over Germany:

4 thermal treatment plants providing a total capacity of 168 000 t/a,
24 soil washing plants providing a total capacity of 1.5×10^6 t/a, and
81 biological treatment plants providing a total capacity of 2×10^6 t/a.

These plants provide an available soil treatment capacity of ca. 3.8×10^6 t/a in 1997. The average degree of usage was about 64% (Table 34).

Table 34. Soil treatment technologies and capacities in Germany in 1996

Technology	Capacity, t	Soil treated, t (average degree of usage)
Biological treatment	1 897 850	1 243 678 (65.5 %)
Soil washing	1 382 900	854 333 (61.8 %)
Thermal treatment	168 000	109 200 (65.0 %)
Total	3 448 750	2 207 211 (64.0 %)

In the next sections, the main processes for soil decontamination are briefly described followed by selected examples of decontamination plants that are working or have been working in Germany. Since safeguarding measures are becoming increasingly relevant for the remediation of contaminated soil, their main effects will be briefly described.

7.1.2.2. *Soil Treatment Processes*

7.1.2.2.1. Thermal Soil Treatment [400]

Thermal soil treatment processes are based on the principle of scrubbing pollutants from the soil into gaseous phase by supply of thermal energy; the elimination of pollutants occurs either by a pyrolytic process, by incineration, or by evaporation of pollutants, while avoiding a melting of the mineral soil matrix.

In general, two types of processes can be distinguished:

Thermal treatment by direct combustion of the material (incineration),
Thermal treatment by indirect combustion of the material (pyrolysis).

Both processes have in common the treatment of soil in a furnace (e.g., rotary kiln, fluidized bed, flystream reactor or sintering belt) to which, depending on the procedure, a receiving dryer can be connected as well as flue gas treatment, consisting of high-temperature postincineration and flue gas purification.

In the case of high-temperature postincineration, the waste gas is heated to 900 – 1300 °C so that the existing organic pollutants are completely oxidized. The flue gases from high-temperature postincineration are generally subjected to a multi-step purification in order to comply with the emission limit values required.

Thermal treatment is an available and best demonstrated technology. In practice, thermal soil treatment is mainly performed using medium temperatures of 600 – 700 °C. High-temperature treatment is exclusively applied to highly contaminated soil.

Thermal soil treatment processes are suitable for almost all soil types and pollutant groups (see Table 35 and 36). As a rule, they are only limitedly applicable for the treatment of heavy-metal contamination. At high treatment temperatures, however, heavy-metal compounds become largely immobilized in the soil matrix. High degrees of cleaning, usually > 90 % are achievable for aliphatic hydrocarbons/petroleum products, polycyclic aromatic hydrocarbons (PAHs), volatile chlorinated hydrocarbons, aromatic hydrocarbons, polychlorinated biphenyls (PCBs), halogenated hydrocarbons, cyanides, organic nitro compounds, pesticides, as well as dibenzodioxins and dibenzofurans, depending on treatment temperatures (ca. 350 – 1200 °C) and residence times in the decontamination process (15 – 60 min).

Table 35. Suitability of thermal processes depending on the soil/grain size

Soil/grain size, mineral content	Suitable	To some extent	Not suitable
Gravel	X		
Medium sand	X		
Fine sand	X		
Silt	X		
Clay		X	
Loam	X		
Building rubble	X		
Sludge		X	
Fine grain residues from soil washing			

Table 36. Suitability of thermal procedures depending on the pollutant

Type of pollutant	Suitable	To some extent	Not suitable

Mineral oil hydrocarbons	X		
Polychlorinated aromatic hydrocarbons	X		
BTEX[a]	X		
Phenol compounds	X		
Cyanides	X		
Chlorinated compounds[a]	X		
Organic compounds containing N, S, O, P[b]		X	
Hg, Cd, As, Se, Zn		X	
Other heavy metals[c]			X

[a] BTEX: Benzene, toluene, ethylbenzene, xylene; these compounds include all organic chlorinated compounds such as PCB, PCP, CF, chlorinated pesticides and PCDD/PCDF. Other halogenated compounds can, in general, be treated. At present, they are not significant in practical terms.
[b] NSOP compounds include all organic compounds that are not halogenated and contain at least one of the four elements N, S, O, P. They include, e.g., pesticides, TNT, etc.
[c] Non-volatile heavy metals.

Direct Heating Process of Umweltschutz Deutzen GmbH. The first thermal soil treatment plant in the new Federal States was erected and operated by Broerius Bodensanierung und Umwelttechnik (Borna) on a former coal briquette factory site in Deutzen in December 1994. The plant is now owned by the Umweltschutz Nord Group and operated by the company Umweltschutz Deutzen. The treatment plant has been licensed under the Federal Mining Act and the Federal Technical Instructions on Air Quality Control. The direct heated rotary kiln is operated at average treatment temperatures between 520 and 600 °C. The treatment capacity is about 80 000 t/a (150 000 t/ a are planned). The plant includes a roofed storage for contaminated soil with a capacity of 300 000 t. The same direct heating process has been operated since 1989 by the Broerius company in The Netherlands.

Terravec Process of Harbauer Umwelttechnik. The Terravec process of Harbauer Umwelttechnik, Berlin is an example for an innovative technology that was used very successfully in on-site operation. The plant described below has now disappeared from the market since it gained not enough interest for further applications because of the high costs of the process and sufficient treatment alternatives in existing stationary plants.

The Terravac process was used for the treatment of highly contaminated fine grain fractions, especially from wet mechanical soil cleaning (soil washing). This vacuum distillation plant worked in the low temperature range and was brought into operation in June 1993 in Marktredwitz (Germany). This plant was especially designed for the treatment of soil contaminated with mercury and mercury products. Figure 23 shows the plant under construction. The plant combines a wet mechanical soil separation, concentrating pollutants in the fine grain fraction, with a subsequently thermal distillation treatment. The fine grain suspension coming from the soil washing is dewatered and dried and then fed via vacuum locks to the vacuum distillation plant. In the vacuum distillation reactor the soil is heated to temperatures between 350 and 450 ° C at 5 to 15 kPa. Mercury evaporates under these conditions and can be recovered by recondensation. After passing the vacuum distillation unit the soil is cooled and mixed with washed coarse grained material from the soil washing process. The plant was in operation from 1993 to 1996, treating about 50 000 t of mercury-contaminated soil.

Figure 23. Vacuum distillation plant (Terravac Process) of Harbauer Umwelttechnik Berlin under construction at Marktredwitz

7.1.2.2.2. Physicochemical Soil Treatment/Soil Washing [401]

Physicochemical processes are mainly washing processes. The applied soil washing processes mainly use water as extractant, adding solubilizers, e.g., surfactants, if necessary. Moreover, there are processes that use inorganic and organic solvents.

Common to all soil washing processes is that the pollutants are stripped from the grain by means of mechanical energy and transferred to the liquid and gaseous stage or eliminated with the finest soil fraction. In doing so, water is used as dispersion, solvent, and transport agent; the efficiency can be increased by additives, depending on the type of pollutant. The pollutant quantity remains constant in all the processes outlined. After the process the pollutants are present in a concentrated form.

In general, the following steps are required in the soil washing facilities:

Preparation of the feedstock,
Wet decomposition,
Separation of light/high-density solids (sorting),
Finest grain separation,
Separation of process water and rinsing of washed soil,
Process water circulation,
Wastewater purification,
Waste air purification

The basis for the separation of pollutants (soil washing) is wet cooking. The aim is to achieve a maximum dispersion of the solids and to promote the separation of pollutants by supply of kinetic energy.

In addition, chemical additives, which facilitate overcoming the adhesive forces between soil particles and pollutants, are frequently used.

The application of soil washing processes mainly depends on the content of extremely fine grains in the soil. Fine-grained soil types (e.g., loam, clay, loess) can be washed sufficiently with these processes only to a limited extent since the fractions with a small particle diameter contain the largest pollutant load as compared to their mass. Thus, the energy required for washing increases with decreasing grain diameter, i.e., depending on the process, the energy necessary for the separation of pollutants cannot be provided economically.

Likewise, the limits of washing processes depend on the type and distribution of pollutants in the soil. Tables 37 and 38 provide orientation for the assessment of possible applications for soil washing processes. Moreover, there are limits to this process if the pollutants are viscous (e.g., tar) since they could cause malfunctions by smearing plant components. It is also not reasonable to use wash processes for pollutants that are incorporated into porous components and/or in cases where the pollutant is glued together with the mineral grain due to the washing medium and cannot be separated.

Table 37. Suitability of soil washing processes depending on the type of soil

Type of soil	Not suitable	To a limited extent	Suitable
Gravel[*]			X
Medium-grained sand[*]			X
Fine-grained sand[*]			X
Silty sand[*]			X
Loam		X	
Clay		X	
Rubble		X	
Sludge	X		
Ash	X		

[*] in line with DIN 18 196.

Table 38. Suitability of soil washing processes depending on the pollutant

Pollutant	Not suitable	To some extent	Suitable
Mineral oil hydrocarbons			X

Polycyclic aromatic hydrocarbons	X	
Highly volatile chlorinated hydrocarbons		X
Aromatic hydrocarbons		X
Dioxins and furans	X	
Cyanides	X	
Heavy metals	X	

High-Pressure Washing Process of NORDAC. Since l986 the high-pressure washing process of Klöckner – Oecotec, Duisburg has been used by the soil washing company Afu, Berlin, at various Berlin sites. This process has been developed in The Netherlands. The centerpiece of the washing plant consists of a high-pressure steel jet tube, in which water under high pressure up to ca. 35 N/mm^2 impinges on the pollutant-laden soil and bring about a separation of pollutant and soil particles. This process is usable for a broad pollutant spectrum. A development of Klöckner – Oecotec based on this principle started up in April 1989 on the former site of the Mannesmann Röhrenwerke in Düsseldorf-Lierenfeld. The Oecotec high-pressure soil washing plant 2000, which consists of 50 standard containers and is equipped among other things with a three-stage high-pressure steel tube uses steam injections to partially dissolve the pollutants before the treatment in the high-pressure steel tube. This plant was subsequently transferred to contaminated sites in Hamburg, Stuttgart-Bad Cannstatt, Munich, and Thale. In 1992 the plant was sold to the GFA company in Thale. Figure 24 shows this plant at the site of use in Düsseldorf-Lierenfeld.

Since 1991 a further development of this process has been in continuous operation on the site of the North-German Contaminated Sites Remediation Center (Norddeutsches Altlasten-Sanierungszentrum, NORDAC) in Hamburg. The technical capacity of the plant is about 100 000 t/a. Since the beginning of operation about 500 000 t of contaminated material has been purified with the NORDAC soil washing plant.

Figure 24. Oecotec high-pressure soil washing plant at the first site of application in Düsseldorf-Lierenfeld

7.1.2.2.3. Microbiological Soil Treatment [402] (⟶ Bioremediation)

Microbiological decontamination processes make use of microorganisms' ability to use organic compounds, in particular of biogenic origin, as nutrients (substrate). Under favorable conditions, the microorganisms convert these organic compounds to carbon dioxide, water, and biomass (mineralization), using the substrates as energy source and to build up their cells.

The processes mainly used in practice include biopiles and in situ procedures.

Biopile procedures are divided into *static* and *dynamic* processes. As opposed to static processes, in *dynamic processes* the soil, aggregated in layers of 0.5 to 3 m, is taken up periodically with special turning arrangements, homogenized and aggregated again, if necessary by adding water and nutrients. Figure 25 shows the tiling of soil at the microbiological treatment plant Ramstein (Bicoterra process) of the Sotec company.

Figure 25. Tiling of soil at the microbiological treatment plant Ramstein of Sotec (Bicoterra process)

In the *static process*, the necessary environmental conditions are created using irrigation and venting systems.

The degradation process in biopiles generally takes several months. All of the 81 stationary large-scale biological soil treatment plants in Germany use the biopile procedure. Figure 26 shows an example of the biological treatment of soils contaminated with petroleum hydrocarbons and polycyclic aromatic hydrocarbons according to the Terraferm process in roofed piles by Umweltschutz Nord. This process is used at 11 stationary biological treatment plants in Germany.

Figure 26. Terraferm heap process of Umweltschutz Nord for the biological treatment of soils contaminated with petroleum hydrocarbons and polycyclic aromatic hydrocarbons in Frankfurt/Main

In the microbiological in situ procedures, the soil is not excavated. Depending on the depth of contamination, different in situ methods are possible. In the case of *superficial contaminations* in the unsaturated zone soil can be loosened up in order to make it possible to add nutrients or water to create optimum environmental conditions (landfarming).

In the case of *deep-seated contaminations*, infiltration of microorganisms and nutrients through a rinsing cycle in combination with hydraulic measures are used.

However, the applications of biological in situ processes are relatively limited compared to on-site/off-site processes. The site is not usable during the relatively long period of remediation and in many cases, the in situ boundary conditions (inhomogeneous pollutant distribution, deficient hydraulic conductivities, inadequate moisture and oxygen conditions, unfavorable conditions of temperature and pH) are such that on-site biological treatment, like biopiles or other measures are preferred. As a rule, suitability of in situ biological treatment depends on the bioavailability of pollutants in the soil, their microbial degradability, and certain soil qualities (homogeneity, content of fine grains). Table 39 shows the microbial degradability of substances classes, provided that these substances are bioavailable.

Table 39. Microbial degradability of substances relevant for decontamination

Class of substances	In principle, easily degradable	In principle, difficult to degrade
Aliphatic hydrocarbons, petroleum hydrocarbons and their derivatives	+	
Monocyclic aromatic (e.g., BTX aro-matics) and heterocyclic hydrocarbons (e.g., pyridine, quinoline)	+	
Polycyclic aromatic hydrocarbons (PAH)	+[a]	
Highly volatile halogenated — in particular chlorinated — hydrocarbons	+	+[b]
Alicyclic chlorinated hydrocarbons and their derivatives	+	
Polychlorinated biphenyls (PCB)		+[c]
Polychlorinated dibenzodioxins and dibenzofurans (PCDD and PCDF)		+[c]
Pesticides and their derivatives		+[c]
Heavy metals		not degradable

[a] Up to ring 4 PAH.
[b] Ring 5 and 6 PAH, trichloroethylene, polychloroethylene (PCE).
[c] Some low-chlorinated congeners are in principle degradable/dehalogenatable. For highly chlorinated congeners degradation cannot be confirmed at present.

7.1.2.2.4. Soil Vapor Extraction Processes

Soil vapor extraction procedures are applied for the remediation of soils that are contaminated with highly volatile compounds. These extraction procedures are primarily applied in the unsaturated zone of the soil as in situ processes. The contaminated soil vapor is extracted by a low pressure, generated by a vacuum well. At the same time, evaporation is promoted by increasing the flow of air along the pollutant molecules. The exhausted soil vapor is purified by activated carbon.

Soil vapor extraction can be combined with air sparging of the saturated zone. The pressurized air introduced into the groundwater causes a stripping process and thus converts the volatile pollutants to gas. The gaseous pollutants are collected and extracted by the soil vapor extraction device and decontaminated with activated carbon.

Soil vapor extraction has been used very successfully for a long time on a large scale for the remediation of soil contaminated with volatile compounds (volatile halogenated hydrocarbons, aromatic hydrocarbons, gasoline hydrocarbons). The advantage of this process is that high purification capacities can be achieved with relatively simple design. Figure 27 shows a soil vapor extraction arrangement for the remediation of underground contamination at a plant site. The compact system, consisting of an exhaustion unit with water separator, electrical control unit, and an activated carbon filter, has only a small space requirement.

Figure 27. Compact plant for soil vapor extraction (HUT company) on a plant site

7.1.2.3. *Safeguarding Technologies*

Surface sealing systems seal off the contaminated soil from atmospheric influence. Thus, they prevent infiltration of rain water in the contaminated areas and leaching/migration of pollutants into groundwater. Surface sealing systems consist of the actual sealing component, surface water, and gas drainages.

Vertical sealing systems are mostly carried out as slurry walls. Slurry walls are designed to seal off the contaminated soil from the surrounding and thus prevent lateral leaching of pollutants. They are inserted up to a depth where a natural compact layer is present; in combination with surface sealing they encapsulate the contaminated soil.

The requirements for the type and the mass of the slurry wall are determined for each individual case, depending on the chemical and hydraulic strain. The most common types of slurry walls are sheet walls, thin walls, and diaphragm walls.

Immobilization Measures. With immobilization measures, the pollutants in the soil are contained in such a way that leaching into the environment is prevented or reduced. Although various immobilization processes are available (compaction, sealing, canopy, etc.), in most cases processes for pollutant stabilization by solidification are applied for the treatment of contaminated soil. In the solidification process, the contaminated material is mixed with a bonding agent and water and then filled into molds; or installed and compacted on the spot; or pressed into molds and pelletized. This process is aimed at manufacturing a water-proof and solid monolith as final product of the solidification process.

There is no best available technology for solidification procedures with qualitative requirements for the solidified material. The following criteria are used to determine the suitability of solidification procedures:

mixed contamination of soil,
low to medium organic contamination,
low to high organic contamination,
fine-grain material.

The suitability has to be tested in advance.

Passive Hydraulic and Pneumatic Measures. *Hydraulic measures* have been the most frequently used technology in hydraulic engineering and foundation engineering for a long time. These measures are employed to prevent contact between the contaminated material and the groundwater. This is achieved by changing the hydrodynamic situation of the groundwater, for example by lowering the groundwater level with extraction wells or reversing the groundwater flow with infiltration wells. Hydraulic measures are used for existing depositions and hazardous waste sites.

In contrast to that, *pneumatic measures* are only relevant for existing depositions that form a hazardous gaseous phase. Pneumatic measures imply installing a gas drainage to control the disposal of gas and thus to prevent the gas from migrating into the surrounding.

7.1.3. ***Costs for Soil Remediation***

As a matter of fact, the costs incurred for soil decontamination in Germany are strongly influenced by the market conditions, i.e., by the treatment capacities offered and the amount of soil that is made available for decontamination. As previously mentioned, in Germany the available treatment capacity already exceeds the request for soil cleaning. Consequently, competition strongly influences the costs for soil treatment. In 1990 the companies charged for the clean up of 1 t of contaminated soil about

$ 150 – 300 for thermal treatment
$ 120 – 200 for soil washing
$ 100 – 200 for biological treatment

Since then, the prices for soil decontamination have decreased significantly, as the figures from Germany from June 1998 indicate. Table 40 reflects the results of a survey on the costs of the main decontamination techniques for three defined types of soil [399] (3000 t soil, moisture content 15 % contamination profile given as PAHs and TPH (Total Petroleum Hydrocarbons).

The costs for *soil washing* are constantly falling, but rather slowly. The upper limit of cost in terms of an economic operation of this technique seems to be reached. The same is true for *thermal treatment*. However the cost for the thermal treatment of lightly contaminated soil are now on the same reasonable level as the biological treatment.

Table 40. Costs for soil treatment in Germany

	Average costs in 1995, €/t			Average costs in 1996, €/t			Average costs in 1997, €/t			Average costs in 1998, €/t		
Type of soil	**Thermal treatment**	**Soil washing**	**Biological treatment**	**Thermal treatment**	**Soil washing**	**Biological treatment**	**Thermal treatment**	**Soil washing**	**Biological treatment**	**Thermal treatment**	**Soil washing**	**Biological treatment**
Contaminated soil type 1[a]	121	124	75	111	112	70	97	100	62	90	93	56
Contaminated soil type 2[b]	101	96	75	94	86	67	84	81	54	84	68	51
Contaminated soil type 3[c]	81	83	64	74	75	56	83	64	46	83	53	38

[a] Fine grain content < 30 %, PAH 2000 ppm, TPC 10 000 ppm.
[b] Fine grain content 15 %, PAH 200 ppm, TPC 5000 ppm.
[c] Fine grain content: 15 %, PAH none, TPC. 5000 ppm.

7.1.4. ***Outlook***

In Germany, the standard for the treatment of contaminated soil is relatively high. Expenses spent by the public sector for research and development are estimated being added up to 153×10^6 € for about 200 projects within the last 20 years. Industry has invested approximately 460×10^6 € only for soil decontamination. Meanwhile, there are more than 100 soil treatment plants in operation providing a total treatment capacity of almost 4×10^6 t/a. Further development of soil decontamination technologies will be restricted to very specific problems like the biological treatment of TNT or TNT derived compounds. While during the 1980 and in the first half of the 1990s the main remediation strategy had been focused on soil decontamination, nowadays the approaches move to a broader concept. Land remediation must be seen in a close relation to the reuse of land. In this context, land remediation is not only the elimination of dangers by the clean-up of soil and groundwater. Besides the purification it encompasses the whole process of land development for future uses. This means, the remediation or reclamation process will become more complex and interdisciplinary.

7.2. Recultivation

7.2.1. ***Definition***

The term *recultivation* has become established as a collective name for all human activities for the restoration of a landscape. Following definition is given in [404]: “Recultivation denotes all the initiatives and measures that serve for the restoration of the scenic and economic structure of those parts of the cultivated land whose natural productivity have been considerably reduced or destroyed by human interventions, primarily by the mining of raw materials.” In this definition the economic, ecological, and structural factors are regarded as of equal rank.

The term *renaturation* has been introduced for the treatment and development of abandoned mining areas, preference here being given to free succession without significant additional shaping of the landscape. As a rule, renaturation areas are small. In recultivation, by contrast, various, extensive landscape elements are combined with one another. At the same time the main characteristics of the cultivated land — that is its settlement and use by human beings — are deliberately incorporated.

Successional processes are integrated also in the modern notion of recultivation. If a certain nature conservation goal has to be followed, priority is given to the taking over of geological and ecological conditions exposed by mining or to the conscious shaping with initial substrata that are not immediately cultivable. A landscape made largely uniform by agriculture and forestry, especially in Central Europe, can require this.

7.2.2. *Demands on Recultivation*

The mining of raw materials is usually accompanied by a radical change of the earth's surface, and ecological equilibria that have become established over thousands of years and have been more or less modified by human use are destroyed. For this reason the ultimate goal of recultivation is considered the restoration of a balanced ecological state and fitting of the recultivated area into the surrounding landscape. Negative examples from many countries of the world show that this cannot be achieved without a target for soil development or establishing an ecologically compatible chemical and physical state. Those examples include the tertiary sand waste heaps of the former East German lignite open mines, the erosion-threatened and nutrient-free dumps of lateritic soils in Asia, and the heavymetals contents of waste materials from mines worldwide.

7.2.2.1. *Recultivation Goals*

The recultivation goals for mined areas have to be as various as the settlement, industrial, and countryside areas concerned themselves. However, some basic principles can be generalized:

1. Recultivation, in contrast to renaturation, requires such a large area of land that its organization becomes the subject of district planning and regional planning procedures. Recultivation thus becomes a task of regional policy and is therefore exposed to political processes even in the goal setting.
2. Interventions in the economic area necessitate interference with property. Depending on the initial situation, decisive courses are often set here. Thus, in former West Germany, for example, clear preferences exist for the reestablishment of agricultural or subsequent commercial uses.
3. The nature of the intervention to be compensated for, especially its technical or commercial layout, creates far-reaching conditions that exclude particular recultivation goals and favor others. If, for example, the landscape is changed by the creation of inner and outer dumps, only woods can be planted because of the resulting slopes, especially also with regard to the erosion dangers. Microclimatic changes due to the intervention also occur, but their influence is weaker. Furthermore, the leaving of final voids will inevitably lead to formation of water bodies. Altogether, changes of the groundwater conditions are far-reaching for the shaping of the landscape in certain regions of the world.
4. Important factors for the formulation of the recultivation goal are the geological conditions and thereby the possibility of selecting adequate soils. The lignite commission mining plans in former West Germany, which were drawn up for the Rhenish lignite district, can serve as examples for the drawing up of recultivation goals.

In the Rhineland, in accordance with land use before coal mining, attention is centered on *agricultural* reutilization. With this, the careful choice of initial substrate and the development of soil fertility become very demanding tasks. The application of the tillable cultivated layer, mostly consisting of loess or loam, is followed by seven years of interim farming, which leads, via the first sowing and maintenance of alfalfa, after three years to the main crops barley, wheat, rye and sugar beet or even vegetables.

In the winning of other raw materials, agricultural recultivation after termination of extraction is not an attainable goal anyway, since no soil substrata at all are available originally: examples are the hard coal heaps which even after applying native soil, merely allow planting with trees or grasses, but on the other hand are well suitable for the restoration of recreation areas.

In other climates, especially in the tropics, agricultural recultivation is started best with agroforestry, especially in order to remove the initial food bottlenecks and solve the erosion problems.

For *reafforestation*, a much wider range of different substrata can be used because of the greater ranges of site requirements of the wood plants.

For the *moderate climatic regions of former West Germany*, determined by loam – gravel soils, the possibilities of forestry can be described with the following types of management objective.

1. Common oak – small-leaved lime – hornbeam mixed forests with transitions to sessile oak forest on dry locations, 40 % of recultivated area
2. Beech forests with strong inclusion of valuable broad-leaved trees (e.g., ash, red oak, walnut, lime, maple, and common elder) and a minor amount of European larches, 30 %
3. Deciduous forests with a high proportion of valuable broad-leaved trees distributed in groups or small stands, 10 %
4. Coniferous woods, principally Douglas fir, but also pine and fir, 10 %
5. Meadows, succession areas, hedges, bodies of water, 10 %

Things are quite different in the drier and sandy *East German recultivation areas*. In former East Germany the following types definitely predominate:

1. Pine – sessile oak forest
2. Pine – common oak forest
3. Birch – heath succession stages

Water bodies can be laid out as a further important building block of recultivation landscapes. These bodies of water can be economically important as retention areas, but in arid regions they take over the important function of water reservoirs. Water bodies are regarded in Central Europe as undesirable for water management owing to the exposure of the groundwater strata and the risk of entry of substances. As recreation areas and above all as nature conservation areas, however, they are outstandingly important. If recultivation areas are declared as national preserves in former West and East Germany this is based mostly on the wealth of species on the bodies of water.

Owing to the enormous decline of species in the cultivated land it must be assumed that at least 10 % of the recultivated area are used for the protection of species in highly developed settlement and economic areas. In other areas, e.g. in former East Germany, that have a considerably smaller demand for agricultural use, the proportion of areas used for the protection of endangered species amounts to more than 50 % of the total area. By this means recultivation comes ever closer to its goal of being a compensation for an intervention in the ecosystem.

7.2.2.2. *Initial Substrata*

A key role in recultivation planning attaches to the analysis and selection of the geological strata exposed during mining. Materials from great depths often have on the one hand advantages in nutrient content and grain size structure, but on the other hand outright poverty or absence of biogenic contents. As a rule, a topsoil application is impossible for technical reasons. The main problem in this connection is that because bulldozers and lorries must be used for the application of topsoil, the possible positive effects are nullified again by soil compaction.

The displacement process inevitably entails the destruction of the live substratum “soil” and its specific bedding. The recultivation cannot start to develop a pedosystem and finally an ecosystem from spent soils but merely from introduced substrata.

For Europe, three initial materials in particular play an important part: loess and loess loam; river gravel; Pleistocene sands and marl.

Loess is a yellow-brown to gray-brown, limey, eolian loose sediment from the dry-cold climatic phases of the Pleistocene. Over a period of about 20 000 years, the upper layer of this silt-rich material has developed through pedogenetic processes (clay formation, clay displacement, deliming) as a rule to para-brown earths, similigleys (including their transitional phases) and, if groundwater influence is present, to gleys.

This loess loam grades downward into lime-rich, loose loess. Until into the present century, this loose loess was extracted as an agricultural fertilizer in small underground “marl pits”. Below this “marl” there follows a layer, of Riss date, of carbonate-poor, partly clayey or sandy-gravelly loam.

This stratigraphic sequence found in the Rhineland is of variable thickness, because during the deposit of wind-borne loess erosion phenomena also naturally occurred, and may occasionally have removed a whole bed again.

The terrace sediments lying under the loess beds are melt water deposits (sands, gravels, and boulder flint) from the Pleistocene. Also found are oligotrophic, weakly sorbing clays from the Tertiary, consisting mainly of kaolinite.

7.2.3. *Soil-Physical Properties*

7.2.3.1. *Silvicultural Recultivation*

The investigations carried out by the State Geological Office of North Rhine-Westphalia led in 1967 to the laying-down of directions on the application of tillable soil material for silvicultural recultivation in the lignite districts of the Rhineland. After a certain trial period, these directions were revised in 1973, 1984 and finally once more in 1996 in order also to support a larger variety of different soil substrates. Thus the variation of ecological stations is increased.

The essential elements of the directions are: the rough heaps must be so shaped that no water damage or damming occurs. Their top 1 – 2 m should consist of sufficiently permeable sandy – gravelly material. Tillable soil material should be applied about 3 – 5 m thick on the rough heap. Possible tillable material is a mixture of sandy – gravelly beds of the high and middle terraces and loess or other cohesive types of soil called forest gravel. Also usable are highly loamified loesses unsuitable for agricultural recultivation as well as loess beds affected by groundwater and dam water. On level or gently inclined surfaces the tillable material should include 25 – 66 % of loess or loess loam. On slopes (i.e., areas of an inclination exceeding 1 : 3) the loess fraction should be limited to 20 – 25 %. Instead of a 4 m thick mixture of loam and gravel, 2 m of loess or loess loam also can be dumped. Processing the surface thus applied with earthworking machines is not required.

After the last revision of these directions it is also allowed to use pure sand or clay and different other substrates on small areas in order to increase pedological variability.

In 1984, STRATMANN carried out profitability studies on recultivation areas in the Rhenish lignite district for the Niedersächsische Forstliche Versuchsanstalt (Lower Saxony Forestry Research Institute) [405]. One of the purposes of the investigation was to assess in detail the relationships between tree growth and soil. The soils included were mixtures of gravels and loams: i.e., substrata that corresponded to the instructions of the State Supreme Office of Mining for the Rhineland, even if these came from times when the directives for the

deliberate blending of such mixtures were not yet laid down. However, among others locations with pure loams, sands, and silts having more or less small additions of fly-ash and coal were also investigated.

According to the results, the prospects for future oak cultivation on forest gravel must be judged favorable. In accordance with the specific site requirements of the two oak species, the sessile oak should be placed on the drier sites and the common oak on the wetter. The hornbeam or the small-leaved lime can be mixed in instead of the beech for the clearing of the stems.

In the final assessment it is stated that sandy forest gravels can lead to a considerable reduction of the yield of forest stands. Soil compaction significantly restricts the growth. The effects of soil compaction were investigated in detail in the following years. On spreading the forest gravel, more or less strongly marked ridges arise which, according to forestry opinion of the 1970s, had to be removed by leveling. The standards of the final operation schedules (especially the required slope inclinations) made in many areas an enormous amount of leveling necessary; and the resulting compactions, which in some cases were high, could not be removed satisfactorily, even by subsequent ripping.

The compactness of the forest gravel was subjected to a detailed investigation. Finally the decision was made to preferably dump the material such that small ridge heights are obtained and leveling is therefore omitted [406]. The tipped substrata then have the highest possible pore volumes which gives the best prerequisites for good rooting, a sufficient water storage and subsequent water supply, as well as an assured gas exchange for plant growth.

The basis of these considerations was that each excavation process destroys the soil structure, and as a result of the deposition of both loess and boulder gravel, a coherent structure that tends to instability or *compaction* is finally formed. This relates especially to compaction of the coarse pores, whereas the distribution of the fine pores is little changed.

The total pore volume fluctuates between 40 and 25 %, forest gravels of high loam content show relatively high pore volumes. The influence of the leveling operation was investigated and established in detail [407]. It is interesting that merely a single aftertreatment shows a positive effect, while several operations are again accompanied by negative compaction effects (Table 41).

Table 41. Effects of various forms of tillage before afforestation on soil structure, void ratios per measuring depth

	Surface			
	Leveled, ripped and harrowed	Leveled, ripped and smoothed	Leveled, ripped	Small spreader, not treated
Measuring points	1 – 8	1 – 6	1 – 6	1 – 6
Depth	Void ratio %			
0.3 m	30.1 ± 2.9	28.0 ± 1.9	29.2 ± 1.0	35.2 ± 1.0
0.6 m	27.7 ± 2.6	27.8 ± 1.2	29.0 ± 2.3	33.5 ± 3.5
0.9 m	28.1 ± 2.5	28.3 ± 1.5	30.5 ± 2.4	35.3 ± 1.1
1.2 m	28.5 ± 3.0	30.1 ± 1.5	32.5 ± 1.8	35.0 ± 3.5

It remains to be seen whether the better starting conditions on unleveled forest soils persist: future research will have to keep an eye especially on the biological regeneration and structuring of the soils. Previous investigations show that there is apparently a connection between amount of initial compactness and earthworm settlement.

In forest gravel investigations the compactness ranges from 1.5 – 2 g/cm^3, and therefore approaches that of pure main terrace gravels. Forest gravels richer in loess, with their densities of 1.7 – 1.85 g/cm^3, correspond to the old forest sites on similigleys.

Recultivation soils must be generally classified as “dense” to “very dense”, an increase of loess and loam fractions enhances pore space and loosening. For typical physical soil characteristics see Table 42.

Table 42. Most frequent values for consolidation, pore volume and void ratio of the soils (105°C dry)

Soil type	Consolidation, g/cm^3	Pore volume, %	Void ratio (or ϵ)*
Sandy	1.67 – 1.19	37 – 55	0.58 – 1.22
Loamy	1.96 – 1.19	26 – 55	0.25 – 1.22
Silty	1.53 – 1.19	42 – 55	0.72 – 1.22
Clayey	1.32 – 0.92	50 – 65	1.00 – 1.85
Organic	0.48 – 0.12	60 – 90	1.50 – 9.00

* ϵ = porosity.

A peculiarity shown in the investigations on forest gravel was a strong tendency to a natural compaction of the materials when the admixture of loams and loesses in the forest gravel was such that the proportion of cohesive components taken up was between 15 and 18 % [407]. In these cases a complete filling of the pore space can be assumed due to the composition and grain size of the material, so that compaction occurs even without fairly strong external action.

The investigations in the Rhineland show that with loess, but in the end also with forest gravel with the appropriate admixture of cohesive components, the moisture of the material to be excavated is of great importance. It is generally accepted that wet areas should not be excavated. Wet areas are also often encountered where as a result of preliminary work soil materials are stockpiled and are then wetted for a fairly long period. The materials scarcely release this moisture even after digging.

The erosion risk on the numerous slopes leads to the conclusion that the forest gravels should contain a maximum of 25 % of cohesive components. However, investigations in open pits that have limey loess available showed, that on increasing the lime content of the admixed loess the stability of the forest gravel increases considerably even on the slopes, so that 30 – 40 % admixtures of loess in the forest gravel are acceptable without considerable erosion problems occurring.

7.2.3.2. *Agricultural Recultivation*

The greatest demands on the choice of soil naturally arise for agricultural recultivation. If loess materials are available, it is clearly preferable here to feed them into the agricultural recultivation. The following procedure has been successful [408]. The quality of the loess to be spread is analyzed by sampling the mining slope and hand drilling in the zone immediately in front of the working boundary. During the building up of the rough heaps and the spreading of the loess, technical aids are used (e.g., rotary lasers) in order to produce the most accurate possible surface even with the high-capacity equipment and to minimize the leveling expenditure. Bulldozers with extremely low specific soil pressure are used. In the loess leveling of the arable land, surface roughnesses can be allowed to remain, provided they do not prevent subsequent treatment with agricultural machines. The loess must not be leveled immediately after spreading, but only after complete drying out. To obtain a small ridge height during the spreading of the loess, high-capacity equipment must be operated at reduced power, if required. Wherever possible the dumping height of the topmost spreader must be so reduced that a horizontal position of the discharge boom is maintained. By this means a more precise filling is made possible.

The questions of the aeration and rooting of the soil are most important, especially in the technical design of the measures. The comparison of yield between old and new land allows a positive assessment of the loess recultivations (Table 43). The variation of the basic soil-physical data in loess recultivations has been pointed out [409] (see below):

Soil type and grain size distribution	lU (medium-loamy silt) with 14 – 17 % clay, 70 – 80 % silt, 2 – 5 % sand
Humus content	0.5 to >1.0 %; average increase ca. 0.03 %/a
Bulk density	1.55 – 1.75 g/cm^3
Total pore volume	35 – 52 vol %
Proportion of airconveying voids (>30 µm)	4 to >12 vol %
Utilizable water capacity relative to 1 m root space	high, 160 to >200 L/m^2
Consistency limits	
yield point	26.0 – 29.0 wt %
plastic limit	18.0 – 22.0 wt %

plasticity index	5.5 – 8.0 wt %
shrinkage limit	5.0 – 10.0 wt %

Table 43. Comparison of crop yields on recultivated areas and neighboring old land in Neuss county (t/ha)

Crop type	1990		1991	
	Recultivated area	Old land	Recultivated area	Old land
Wheat	6.1	7.1	8.5	8.5
Rye	5.4	6.1	6.7	6.0
Barley	6.3	5.1	6.9	6.8
Rape	4.1	3.1	4.2	4.0
Sugar beet	71.4	58.6	66.0	58.4

The suitability of different starting materials for different recultivation goals is shown in Table 44.

Table 44. Use of various substrates

	Loess			Marl	Pleistocene river gravel	Pleistocene sands	Tertiary sands	Clay	Marl – sand mixture	Loess – gravel mixture
	Parabrown earth	Similigley – gley	Carbonate-containing underlying loess							
Agriculture	x	▪	x	▪	o	▪	o	o	▪	o
Forestry	x	x	x	o	o	x	o	o	x	x
Protection of species	x	x	x	x	x	x	▪	x	x	x
Recreation areas	x	▪	x	▪	o	x	o	o	x	x

x = suitable; ▪ = of limited suitability; ○ = unsuitable.

7.2.4. *Soil-Chemical Properties*

7.2.4.1. *Silvicultural Recultivation*

Investigations on nutrients content at various problem sites and on nutrient uptake by their vegetation have yielded fundamental information.

Thus it has been found for oaks and beeches on comparable difficult stands that as regards soil nutrient supply, only a few nutrients such as phosphorus and magnesium showed a deficiency. In addition, the pH values were too low in many cases, because at the problem sites the forest gravels were usually very poor in loam and loess. The leaf nutrient values found there were sufficient, except for nitrogen in the case of the oak and potassium and magnesium in the case of the beech. This finding runs through all other investigations, which have shown that the leaf nutrient contents from the dumping of forest gravels are sufficient.

If, on the other hand, the nutrient contents of the soils analyzed at the problem sites are considered, it becomes clear that in some cases a deficiency in the nutrients phosphorus and magnesium and in many areas also nitrogen occurred. With the other nutrients, the supply could not be called particularly good, but it could not in the end be held responsible for causing the harm. In the case of some soils it was also shown that the supply of calcium is inadequate and a strong acidification was found as far as into the aluminum buffer range.

FRIEDRICH has demonstrated a relationship between the nutrient conditions and the different proportions of loess in the forest gravel [408]. Thus, for the forest gravel group with ≥50 % loess and the soils with 25 – 33 % loess neutral to weakly acidic reactions are found. At unlimited exchange capacity calcium and magnesium almost exclusively make up the exchangeable cations. The reserve of phosphorus in the loess-rich forest gravels can also be regarded as high, but naturally in young forest gravels the proportions of carbon and nitrogen in the dry material are low. The forest gravels with 20 or even 10 % loess appear distinctly worse. The pH values of these gravels are in the acid range (Table 45).

Table 45. Chemical state of soils of forest gravel with varying proportions [406]

Surface profile no.	Proportion of loess, %	Depth, 10^{-1} m	pH ($CaCl_2$)	AK_E	%C/d.m.	%N/d.m.	C : N ratio	P, mg/100 g
62	50 – 66	0 – 4	7.35	145.692	0.37	0.03	12.23	38.0
61	25 – 33	0 – 4	7.8	271.761	0.51	0.017	29.82	26.0
		4 – 8	7.7	137.350	0.64	0.021	30.24	18.0
21	20	0 – 4	5.6	31.752	0.098	0.01	9.80	10.0
72	20	0 – 4	7.7	118.008	0.14	0.01	14.30	14.5
		4 – 8	7.4	58.129	0.076	0.008	9.50	11.0
45	20	0 – 4	5.8	61.104	0.44	0.036	12.28	22.0
11	20	0 – 4	5.3	24.611	0.22	0.02	13.06	11.0
22	10	0 – 4	5.4	26.008	0.077	0.009	8.56	12.0
65	10	0 – 4	3.75	9.500	0.14	0.011	12.45	11.0
41	10	0 – 4	4.95	106.509	0.05	0.006	7.67	8.0
31	10	0 – 4	3.55	23.843	0.14	0.012	11.58	13.0

AK_E = effective exchange capacity; d.m. = dry matter.

Especially in forest gravels with less than 10 % loess, aluminum and in some cases iron and hydrogen can have a high to predominant proportion of the cation exchange. The exchange capacities fall in each case by 30 % relative to the preceding soil groups (10 – 20 % loess, 25 – 33 % loess, etc.) that for forest gravels with small proportions of loess nutrient problems must occur.

In the bad forest gravels, possible heavymetals concentrations have been suggested as the cause of poor growth. Here, though, the analyses have shown that the heavy metals copper, silver, chromium, and zinc are present only in very small concentrations and thus cannot be the cause for the impairment of plant growth.

7.2.4.2. *Agricultural Recultivation*

DUMBECK has assembled the most important chemical parameters for initial soils in the Rhineland (see below).

Carbonate content	4.5 – 11.5 %
Nitrogen	total N = 0.02 – 0.05 %
Potassium	total K_2O = 300 – 700 mg/100 g soil
	calcium lactate soluble K_2O = 6 – 12 mg/100 g soil
Phosphorus	calcium lactate soluble P_2O_5 = 3 – 6 mg/100 g soil
Magnesium	$CaCl_2$ soluble Mg = 12 mg/100 g soil

Cation exchange capacity 8 – 10 mval/100 g soil

It is significant that, because of the admixture of carbonate-containing loesses, the initial pH values (7.5 – 8.0) are in the weakly alkaline range. The initial soils often show excellent nutrient supply. However, it can be observed in the course of recultivation when intermixing the soils that the nutrient contents in the recultivation substrata are decidedly low for agricultural use. Here, appropriately increased amounts of fertilizer are supplied, which exceed the uptake by plants and therefore lead to an enrichment in the soil. Since the humus contents are low, the nitrogen mineralization is also lower than on old sites. A 15 – 30 % nitrogen supply compensates the nitrogen release resulting from the lower mineralization.

Great problems have occurred in the recultivation of poor tertiary material, especially sands, in the mining districts of Central and former East Germany. The content of marcasite and pyrite led to an enrichment of the soil with acids, which can only be contained by extensive liming using power station ashes or lime fertilizers. Above all, the tertiary sands having a total sulfur content >2 % also must be viewed critically for improvement measures, since often with them only a superficial remediation is possible; and, with deep rooting of plant species, this can lead to considerable growth reductions.

In the discussion of improving agricultural or silvicultural use on poor substrata, the recycle of recycling substances also has been considered. It is being discussed above all in the case of recultivation, to what extent sewage sludge, wood chips, composts, and other recycling products can be used. The previous studies show that for difficult recultivation situations a distinct improvement can be seen, especially in the medium term.

7.2.5. *Soil Biology*

Soil is a basic compartment of terrestrial ecosystems (see also Chap. Function and Utilization of Soil). All nutrients are recycled in the soil through a cascade of smaller and smaller animals, fungi and bacteria. Many of the aspects of stability in ecosystems depend on the performance of these organisms. In the first stage the reclaimed substrates are colonised by airborn species: microorganisms, aeronautic spiders, flying insects, algae and plants with winddriven seeds. The further establishment of more demandig species without farreaching propagules or flight mechanisms depends much on the quality of soils. (For further discussion see [410].)

It has been demonstrated, that adverse chemical properties of the soil — especially acidic substrates — prevent soil dwelling earthworms from colonization [411]. An other effect is depending on the physical properties: compaction has proved to be the main factor impeding the development of a stable soil fauna. In compacted soils areation and percolation of water are reduced. The burrowing activity of the common earthworm (*Lumbricus terrestris*) is stopped at bulk densities above 1.6 g/cm^3 [412] and rooting of plants is impeded. Therefore, the careful selection of soil substrates and gentle dumping are most important for the establishment of a diverse and efficient soil fauna and flora.

On forest gravel in the Rhineland soil texture exerts a major effect on establishment of earthworm populations: higher contents of clay and silt favore soil dwelling species. If loess content of the forest gravel is at minimum 40 %, earthworm communities — compareable to those on undisturbed sites — are able to establish within ten years [413].

Colonization is much faster than suggested by studies on the dispersal of earthworms [414]. Therefore it is assumed, that at least cocoons or even living animals are translocated when topsoil is mixed in the soil substrates [415].

Land-use after substrate placement is crucial for the biological development. On afforested reclaimed stands with loess soil, total carbon reaches levels of more than 2 % at depth 0 – 10 cm and 0.8 % (depth: 10 – 30 cm) within 20 years, compared to about 0.4 % (depth: 0 – 30 cm) under agricultural use [416].

Recently, scientific studies focused on the ecological effect of the highly divers microrelief on unleveled reclaimed land. Preliminary observations suggest, that in the depressions of the rough surface organic material and fine soil is accumulated. Thus favoring hot spots of biological activity and facilitating the establishment of species that need a moist environment, rich in biomass.

[Top of Page]

8. Legal Aspects of Soil Protection

Peter-Christoph Storm

8.1. Foundations of Soil Protection Law

8.1.1. *Land Law*

Following Roman legal tradition, the open air and flowing water are not things in possession in the legal sense and not accessible to action under (private) law. In contrast, the partition of the land, is considered an original law of the earth. The taking of land is regulated by land law: this allocates the land to persons in law as bearers of rights and duties and governs legal relationships with regard to the land. It is expected that the user of the land will also protect, maintain, and preserve it. Until recently the land has been regarded not only as immovable and nonincreasable but also as indestructible. Since the 1970s, however, concern about human's natural fundamental requirements of life has developed on the one hand from the recognition that natural resources are limited, and on the other from a growing awareness both of the increasing anthropogenic pollution (or other harmful change of the natural condition) of soil, water, and air and of their dwindling renewability. The protection and maintenance of human's natural fundamental requirements of life, i.e., care of the environment, has developed into a public task for public benefit going far beyond the interests of the individual for private benefit protected by neighbor law. Within the framework of care of the environment, care of the soil has been added as a new terrestrial primary law. The soil is no longer indestructible. The supposed identity of interests between soil use and soil care appears no longer adequate to maintain the soil as a part of the natural fundamental requirements of life for present and future generations.

8.1.2. *Soil Protection Law*

Land law is therefore supplemented by a soil care law or soil protection law [417-420], that

1. as a collective concept gathers together the legal provisions that serve the protection, maintenance, and development of the soil,
2. can be described as a legal system of measures for intervention, incentives, and planning for the realization of certain soil-protection goals, including the necessary processes, organization and sanctions — and therefore as a part of environmental law, and
3. as a result of its primary purpose of protection, differs from regulations for land use in the narrow sense, without it being possible to avoid the overlapping of the two spheres in the face of the subject of regulation named.

8.2. Foundations of Soil Protection Policy

8.2.1. *Soil Protection in Germany*

The environmental program of the government of the Federal Republic of Germany of 1971 [421] already designated the soil in its “triad of goals” as an environmental medium, alongside water and air, that had to be protected from harmful human actions. The soil protection concept of the German Federal Government of 1985 [422] described soil protection as a cross-sectional task of environmental protection and emphasized two central approaches for enhanced soil protection:

1. Minimization of the qualitatively or quantitatively problematic entry of substances and
2. A “trend reversal” in land use [423]

8.2.2. *International Soil Protection*

Already in 1972 in the regional framework of the Council of Europe a “European Soil Charter”, which formulates a Europe-wide soil protection policy in twelve articles, was adopted as a recommendation for the Member States [424].

In 1981, the protection of the soil was recognized worldwide as a problem in environmental law by the Montevideo Environmental Law Program of the United Nations [425]. The program describes as goals of the legal measures in the soil protection subsection at national, regional and global levels

1. To prevent the impairment as a result of human actions of the sustained productivity of the soil, resulting in erosion, desertification, salination, deforestation, overuse, contamination, or excessive land use, and
2. If the need arises, to remediate polluted soils

Aside from the application of appropriate worldwide basic principles, guidelines, and recommendations at all levels of control (“World Soil Charter”, “World Charter for Nature”), a stronger emphasis on soil protection in the laws at national level is encouraged.

The Conference of the United Nations for Environment and Development of 1992 in Rio de Janeiro worked out in the action program “Agenda 21” among other things instructions for an integrated approach to the planning and management of land resources (Chap. 10 of the Agenda) and the combating of desertification (Chap. 12 of the Agenda), which need national enforcement [426]. In the following sections the soil protection law in Germany is described as an example.

8.3. Constitutional Soil Protection

Soil protection is part of the national task of care of the environment, even if this national task was not expressly enshrined in the German “Grundgesetz” until 1994 (Art. 20a GG) . With regard to intervention measures, soil protection experiences constitutional barriers. The measures must be suitable, necessary, and proportionate.

The soil protection legislator has correctly to value the realities underlying his decisions, to which end scientifically based data are drawn upon, especially for the definition of soil-protecting environmental standards (soil values) to quantify soil damage. This does not exclude precautionary measures if there is not (yet) sufficient scientific knowledge. In addition the remaining constitutional standards of competence and of guaranty have to be observed, above all the warranty of property under the Grundgesetz (Art. 14 Grundgesetz). The social obligation of land ownership is of increased importance because of the nonincreasability of the land. Where constitutions of the German “Bundesländer” (federal states) contain state objectives in favor of care of the environment, these also include soil protection within the scope of the federal state's spheres of jurisdiction.

8.4. Soil Protection under Administrative Law

The protection of the soil is found in a large number of legal regulations from all stages of legislation, supplemented since March 1999 by the Federal Soil Protection Act (Bundesbodenschutzgesetz – BBodSchG) and since July 1999 by the Federal Ordinance on Soil Protection and Contaminated Sites (Bundes-Bodenschutz- und Altlastenverordnung – BBodSchV) [429] . Where a federal state has its own soil protection law, this is applied, provided regulations under federal or state law do not contain provisions to the same effect or to the contrary [427].

The soil protection standards are mainly administrative law. Environmental private law (e.g., environmental liability, Umwelthaftungsgesetz) and environmental criminal law, on the other hand play only a supporting role in soil protection owing to their essentially sanctioning character. Soil protection law is accordingly principally administrative law and as such a part of environmental administrative law, which in turn is part of special administrative law. If no special regulations exist under the law of a particular federal state, soil protection follows the provisions of the remaining environmental administrative law.

8.5. Soil Protection Goals

8.5.1. *Soil*

The meaning of the term soil has in each case to be obtained from the context of legal rules in which it is used; the regulatory purpose of the rules can circumscribe the connotation. However, as a starting point for a legal description of the soil as a value to be protected it can be defined as the topmost layer of the earth's surface, inhabited by animals, plants, and microorganisms, on the firm ground and below the water surface of waters above ground. Beyond that, the protected zone of standards significant for the soil can also cover the uninhabited underground, as far as it is accessible to anthropogenic influences. The new federal act on soil protection [429] defines the soil as the topmost layer of the earth's surface, as far it is bearing soil functions (see Section Protection of Soil Functions), including the fluid or gaseous parts of the soil, but excluding the groundwater and the ground below the surface of water.

8.5.2. *Protection of Soil Functions*

The soil is protected legally as a part of hu-man's fundamental requirements of life for the sake of its functions, especially as a habitat for biomass, as a filter, as a stock of raw materials, and as building land. The legal protection has to guarantee that the soil can provide its varied functions in the required quality and quantity for a long time. The reference to the soil as part of the natural fundamental requirements of life shows the ecological approach to soil protection, which not only views the soil sectorally and as a medium comparable with the environmental media water and air, but also sees it together with these media and the sunlight as the basis of all life [428]. Futhermore the reference makes clear that soil protection is not an end in itself, but rather, like care of the environment as a whole, in the end always occurs for human's sake.

8.5.3. *Impairments of the Soil*

The protection of the soil is directed against impairments of the soil quality and quantity that are caused in the soil by material and nonmaterial anthropogenic intrusions in the soil.

Material intrusions in the soil can be brought about by the directly or indirectly caused entry of substances (soil change by substance entry). Soil impairment by substance entry means the effects of chemical elements or chemical compounds that are capable of detrimentally changing the natural condition of the soil.

Nonmaterial intrusions in the soil include intrusions through an alteration of the soil's use, whereby the efficiency of the soil or of the ecosystem or of the landscape can be impaired considerably or for a long time (soil consumption). In addition they include other non-material intrusions into the soil, which, without being soil consumption in the sense just named, can bring about corresponding impairments.

8.5.4. *Soil Protection by Provision and Aftercare*

Protection of the soil signifies not only the warding-off of impairments of the soil. Protection of the soil also comprises maintenance and development in the sense of shaping, improvement and restoration (remediation of contaminated sites) and provision that impairments do no occur at all and that the soil can perform its diverse functions for a long time. Soil protection accordingly means both precautionary care and aftercare.

8.5.5. *Direct and Indirect Soil Protection*

Soil protection can occur directly as the main or subsidiary goal of regulation, if the soil is identified explicitly as the value to be protected. Direct soil protection can also be spoken of when the legal system uses terms such as environment, natural fundamental requirements of life, natural resources, ecosystem, or nature and landscape, because these notions include the soil as protected property. Soil protection can also occur indirectly, however, by other goods and values being protected, the soil being co-protected without itself expressly being protected property.

8.5.6. *Relative and Absolute Soil Protection*

In view of the variety of the soil's functions and of the multitude of demands that are made upon it, the legal system has to provide for an appropriate solution of possible conflicts of goal. When weighing different interests, consideration is accorded to soil protection (relative soil protection). Soil protection can also be given priority, either generally or in the individual case (absolute soil protection), for instance, by regulations on sewage sludge application.

8.6. Soil Protection Measures

8.6.1. *Types of Measure*

The set of instruments under administrative law, consisting of soil protection measures for intervention, incentives, and planning (triad of measures) is used to realize soil protection goals. The measures interfering with the legal rights of the individual are orders and prohibitions or restrictions directed to the carrying out, toleration, or neglect of particular actions. They are supplemented by instruments that, by means of specific incentives such as the grant of tax allowances or of user advantages, indirectly guide human behavior to the realization of goals serving soil protection. With the aid of the incentive measures the environmental administration influences corporate behavior. This involves the erection and maintenance of supply and waste-disposal equipment, public consultation and information, the public financial support of activities appropriate to soil protection, and public compensation. Within the triad of measures the measures of environmental planning have a special function, because they above all are capable of realizing the precautionary principle.

The measures can be restricted spatially to certain areas (protected areas). They can distinguish chronologically between old soil contaminated sites and new soil impairments.

8.6.2. *Classification of Soil Protection Measures*

The standards significant for soil protection can be classified according to the type of soil intrusion into protective standards against impairments that are brought about by material intrusions and those that are brought about by nonmaterial intrusions.

8.6.3. *Measures against Material Soil Impairments*

The soil receives direct or indirect protection as a result of environmental law from impairment(s) of the soil brought about by directly or indirectly caused entry of substances in the environmental care areas of the protection of soil (Bundes-Bodenschutzgesetz), of waters (Wasserhaushaltsgesetz, Water Resources Policy Act), avoidance and disposal of wastes (Abfallgesetz, Federal Waste Act), emmission protection (air pollution control; Bundesimmissionsschutzgesetz, Benzinbleigesetz, etc.), radiation protection (Atomgesetz, Atomic Act), protection from harmful materials (ChemG) and genetically engineered organisms (Gentechnikgesetz, Genetic Engineering Act), and nature conservation (nature protection and landscape preservation, soil protection; Bundesnaturschutzgesetz). These are supplemented by the Building Regulations Law (Landesbauordnungen) and the General Police and Regulatory Law (Landespolizeigesetze) but also by the Regional Planning Law and Federal Planning Laws (Raumordnungsgesetz, Landesplanungsgesetze) so far as this can control entries of substances by appropriate allocation of soil uses. The main emphasis of the regulations that are intended to protect the soil from the direct entry of substances is in protection from harmful substances and organisms modified by gene technology, avoidance and disposal of wastes, and protection of waters. The immission protection regulations, in particular, give protection from the entry of substances via the air path by measures that operate at the source of emissions. In the area of general environmental care, the environmental impact assessment prescribes for particular projects the determination, description, and evaluation of their effects, also on the soil, and requires consideration of the result of the evaluation during the licensing of the project.

8.6.4. *Measures against Nonmaterial Soil Impairments*

Standards governing nonmaterial soil impairments are stipulated in laws of nature protection and landscape preservation, and of soil protection (Bundesnaturschutzgesetz, Bundeswaldgesetz, Bundes- und Landesbodenschutzgesetze). This core area is supplemented by the standards controlling soil protection of environmental planning and regional planning (Raumordnungsgesetz, Landesplanungsgesetz), of zoning (Baugesetzbuch), and of individual technical plans. For projects subject to an environmental impact assessment the observance of soil protection is assured. This is supplemented by direct and indirect soil-protection standards of agricultural law (e.g., land consolidation, Flurbereinigungsgesetz) and of the Federal Mining Law (Bundesberggesetz).

8.7. Development Trends

8.7.1. *Further Development of German Soil Protection Law since 1985*

At first sight, the soil protection law in force in Germany has an appearance of strong fragmentation. But this alone must not lead to the conclusion that the soil is unprotected or only very incompletely protected. The legal translation of the ecopolitical demands for a reduction of the direct and indirect entry of substances into the soil and for a reduction of soil consumption certainly signals weaknesses of standardization of the law in force. But this is up to now combated less by a special soil protection law than by a strict application of the soil protection law in force and its development.

8.7.2. *The German Soil Protection Act 1998*

The German Soil Protection Act 1998 came into force on 1 March 1999 [429]. It contains in its Article 1 an “Act for the Protection from Harmful Changes to the Soil and for the Remediation of Contaminated Sites (Bundes-Bodenschutzgesetz, BBodSchG)”; Articles 2 to 3 contain amendments to pertinent environmental laws to ensure the dovetailing of the soil protection law with the remaining environmental law.

The Federal Soil Protection Act aims to protect the soil from harmful changes and to make provisions against the build-up of harmful soil changes. Aside from its precautionary function, comprising the defense against hazards and the reduction of detrimental effects on the soil, the law also stipulates soil protection by aftercare for the removal of harmful soil changes and contaminated sites.

The purpose of the Act is highlighted by two principles, which present soil protection as soil function protection:

1. Preservation or restoration of the natural functions of the soil as a fundamental requirement of life and as living space for humans, animals, plants, and soil organisms, as part of the ecosystem, as a degradation, equilibration, and synthesis medium for substances based on its filtering, buffering, and matter conversion properties
2. Preservation or restoration of the soil in its functions as a raw material deposit, as a site for agricultural and silvicultural use, as an area for settlement and recreation, as a site for economic uses, traffic, supply, and waste disposal as well as an archive of natural history and the history of civilization

The Act defines *harmful soil changes* as changes of the physical, chemical, or biological condition of the soil that, depending on their nature, extent, or duration, are capable of impairing the soil functions and thereby of bringing about dangers, considerable disadvantages, or considerable nuisances for the individual or the general public. Here the ecological functions of the soil are included as protected property of the general public. The law also takes into account whether the possibility of a sustained impairment of the natural functions of the soil exist which cannot be reconciled with the protective and user interests of the individual or with the common good.

Obligations to provide against harmful soil changes, for accident prevention, and for the elimination of harmful soil changes support the objective of the law. Special regulations are made for the registration and remediation of contaminated sites. The obligation of remediation of contaminated sites is put in concrete form by rules and orders on soil-dependent pollution limits and test values as well as requirements for removal of pollutants. The federal states can also prepare soil protection plans for the remediation of areas in which harmful soil changes occur, or are to be expected.

8.7.3. *Soil Protection in the Draft of an Environmental Code (UGB)*

The German federal government wants to create an Environmental Code (Umweltgesetzbuch – UGB) that brings together, unifies, harmonizes, develops and updates the environmental law, which is constituted by more than 20 special environmental acts; a draft has been prepared by an independent expert commission in 1997 [430]. In the specific part of this draft of an Environmental Code, chapter ten will regulate soil conservation (Art. 326–354) on the basis of the Federal Soil Protection Act.

[Top of Page]

References

1. C. Bosch in D. Rosenkranz, G. Bachmann, G. Einsele, H.-M. Harreß (eds.): *Bodenschutz*, Erich Schmidt Verlag, Berlin 1988, p. 1480.
2. H. Strunz: *Mineralogische Tabellen 3*, 7th ed., Akademische Verlagsgesellschaft Egeestu. Portig., Leipzig 1970.
3. C. Ahl *et al.*: *Aspekte und Grundlagen der Bodenkunde*, Eigenverlag, Göttingen 1994.
4. V. Wolters: "Bodenökologische Aspekte der Freisetzung gentechnisch veränderter Organismen," *UBA-Texte* **20/93** (1993) 182 – 196.
5. N. Knauer, D. Trautz: "Böden als Lebensraum von Organismen," in H. P. Blume (ed.): *Handbuch des Bodenschutzes: Bodenökologie und Bodenbelastung; vorbeugende und abwehrende Schutzmaßnahmen,* Ecomed Verlag, Landsberg/Lech 1990, pp. 59 – 70.
6. H.-P. Blume: *Handbuch des Bodenschutzes*, ecomed, 2nd ed., Landsberg 1990.
7. AG Bodenkunde: *Bodenkundliche Kartieranleitung*, 3rd ed., E. Schweitzerbart'sche Verlagsbuchhandlung, Stuttgart 1982.
8. S. K. Gupta, F. X. Stadelmann, F. X. Häni, A. Rudaz: "Interrelation of Cd-Ion Concentrations and the Growth and Activities of Microorganisms in two Growth Media." *Toxicol. Environ. Chem.* **8** (1984) 173 – 184.
9. R. Hindel, W. Fleige: "Schwermetalle in Böden der Bundesrepublik Deutschland – geogene und anthropogene Anteile," *UBA-Texte* **10/91** (1991).
10. M. Gruppe, H. Kuntze: "Bodenuntersuchungen im Bereich des UBA-Meßnetzes als Bestandsaufnahme zur Beweissicherung von Umweltveränderungen," *UBA-Texte* **60/93** (1993).
11. R. Hindel *et al.*: "Kartiertechnisches Konzept zur flächenhaften Erfassung von Schwermetallgehalten in Böden," *UBA-Texte*, in press.
12. F. Scheffer, B. Ulrich: *Humus und Humusdüngung*, Enke Verlag, Stuttgart 1960.
13. W. Fleig, H. Beutelspacher, E. Rietz in J. E. Gieseking (ed.): *Soil Components*, vol. **1**, Springer Verlag, New York 1975, pp. 1 – 211.
14. R. S. Swift in G. R. Aiken, D. U. McKnight, R. L. Wershaw, P. McCarthy (eds.): *Humic substances in Soil*, Sediment and Water, J. Wiley & Sons, New York 1985, pp. 387 – 408.
15. W. Ziechmann: *Huminstoffe*, Verlag Chemie, Weinheim, Germany 1980.
16. Z. Filip: "Bodenorganismen und ihre Aktivitäten als Anzeiger ökologisch bedeutender Auswirkungen anthropogener Bodenbelastungen," in: *Tagungsbericht-Symposium mit osteuropäischen Staaten "Untersuchungsmethoden, Bewertungsmaßstäbe und staatliche Regelungen für den Bodenschutz",* BMU, 1993, pp. 47 – 56.
17. Verband der Chemischen Industrie (ed.): *Chemie und Umwelt: Boden*, VCI, 1987, 28 pp.
18. J. D. Stout, S. S. Bamforth, J. D. Lousier: "Protozoa," in: *Methods of Soil Analysis 2 — Chemical and Microbiological Properties,* American Society of Agronomy, Madison, Wis., 1982.
19. E. Aescht, W. Foissner: "Effects of Mineral and Organic Fertilizers on the Microfauna in a HighAltitude Reafforestation Trial," *Biol*, Fertil. Soils (1992) 17 – 24.
20. E. Klapp: *Lehrbuch des Acker- und Pflanzenbaus*, Paul Parey, Hamburg 1958.
21. P. Boysen: "Schwermetalle und andere Schadstoffe in Düngemitteln," *UBA-Texte* **55/92** (1992).
22. C. Bannick, *Mitt. Dtsch. Bodenk. Ges.* **73** (1994) 23 –26.
23. R. Schulz: "Vergleichende Betrachtung des Schwermetallhaushaltes verschiedener Waldökosysteme Norddeutschlands," *Ber. Forschungszentr. Waldökosysteme/Waldsterben der Universität Göttingen* **32** (1987) 1 – 27.
24. M. Abo-Radi, H. Meyer-Steinbrenner, *Wasser Boden* **46** (1994) 58 – 65.
25. DIN 4047, part 3, Landwirtschaftlicher Wasserbau, Begriffe, Beuth Verlag, Berlin 1985.
26. FAO, Guidelines for Control of Soil Degradation, Rome 1983.
27. D. Rosenkranz, G. Einsele, H.-M. Harreß (eds.): *Bodenschutz*, supplement, Erich Schmidt Verlag, Berlin 1988.
28. Der Rat der Sachverständigen für Umweltfragen: Altlasten, Sondergutachten, Stuttgart, Dec. 1989.
29. L. Brown: *State of the World*, W. W. Norton, New York - London 1991.
30. FAO — Food and Agriculture Organization of the United Nations: FAOSTAT, Land Use. Computerized Information Series. FAO, Rome, 2004; http://faostat.fao.org/.
31. German Advisory Council on GlobalChange (ed.): *World in transition. The threat to soil*, Economica-Verlag, Bonn 1997.
32. Bundesanstalt für Landeskunde und Raumordnung, quoted in VCI (ed.): *Chemie und Umwelt: Boden*, 1986, p. 7.
33. Umweltbundesamt: Reduzierung der Flächeninanspruchnahme durch Siedlung und Verkehr – Materialienband. Berlin 2003. http://www.umweltbundesamt.org/fpdf-l/2587.pdf.
34. Bundesminister für Forschung und Technologie: *Unsere Erde im Wandel*, Bonn 1993.
35. L. Brown: *State of the World 1990*, W. W. Norton, New York - London 1990.
36. E. Fleischhauer: "Menschen schützen ihr Land," *Umwelt* **1** (1993) 25 – 27.
37. L. R. Oldeman: Global extent of soil degradation. ISRIC-annual report 1990 – 1992, Wageningen 1992, pp. 19 – 39.
38. FAO, Protect and Produce – Putting the Pieces Together, Rome 1992, p. 5.
39. Intergovernmental Panel on Climate Change (IPCC) (ed.): *Climate Change 1995 — Impacts*, Adaptations and Mitigation of Climate Change: Scientific technical Analyses. Contribution of Working Group II to the second Assessment Report of the Intergovermental Panel on Climate Change, University Press, Cambridge – New York 1996.
40. R. W. Arnold, Iszabocs, V. O. Targulian (eds.): *Global Soil Change*, Laxenburg 1990, p. 78.
41. *Entsorgungspraxis* **3** (1993) 169.
42. FAO, The Conservation and Rehabilitation of African Land, an International Scheme, ARC/90/4, p. 9.
43. L. Brown: *Zur Lage der Welt 1987/88*, World Watch Institute, Washington, D.C., S. Fischer Verlag, Frankfurt/Main 1987, p. 93.
44. FAO – Food and Agriculture Organization of the United Nations: State of the World's Forest. FAO, Rome 1997, p. 200.
45. http://www.mongabay.com/brazil_deforestation.htm; All figures derived from official National Institute of Space Research (NIPA) data.
46. UNFPA (ed.): *Weltbevölkerungsbericht 1996. Welt im Wandel: Bevölkerung, Entwicklung und die Zukunft der Stadt*, UNO-Verlag, Bonn 1996.
47. G. O. Barney (Study director); prepared by the Council of Environmental Quality and the Department of the State, *The Global 2000 Report to the President of the U.S. Entering the 21st Century,* **vol. 1**: The Summary Report, Pergamon Press, New York 1980.
48. B. Roberts: "The Health of the World's Land: A Perspective," *7th ISCO Conference Sydney, Proc.,* **vol. 1**, 1992.
49. R. Horn, *Soil Technol.* **6** (1993) 47 – 75.
50. B. D. Soane: "Compaction by Agricultural Vehicles: A Review," *Scott. Inst. Agric. Eng. Tech. Rep.* **5** (1983) 95 pp.
51. W. B. Voorhees, *Adv. Soil Sci.* **19** (1992) 73 – 95.
52. J. Eriksson, I. Håkansson, B. Danfors: "The Effect of Soil Compaction on Soil Structure and Crop Yields," *Swed. Inst. Agric. Eng. Bull.* **354** (1974) 101 pp.
53. J. B. Passioura, *Aust. J. Soil Res.* **29** (1991) 717 – 728.
54. E. L. Greacen, R. Sands, *Aust. J. Soil Res.* **18** (1980) 163 – 189.
55. R. Horn, *Z. Kulturtech. Flurbereinig.* **26** (1985) 42 – 51.
56. B. M. McKenzie, A. R. Dexter, *Biol. Fert. Soils* **5** (1988) 328 – 332.
57. R. F. Follett, B. A. Stewart: *Soil Erosion and Crop Productivity*, Soil Sci. Soc. Am., Madison, Wis., 1985.
58. R. Lal: *Soil Erosion in the Tropics – Principles and Management*, McGraw-Hill, New York 1990.
59. Committee on Conservation: *Assessing the National Resources Inventory*, vol**s. 1 and 2**, National Academy Press, Washington 1986, 114 pp., 314 pp.
60. U. Schwertmann, F. Schmidt in M. DeBoodt, D. Gabriels (eds.): *Assessment of Erosion*, J. Wiley, Chichester 1980, pp. 403 – 406.

61. K. Auerswald, *Z*, Geomorphol. N.F. Suppl. 83 (1991) 155 – 160.
62. H. E. Dregne, *J. Soil Water Conserv.* **47** (1992) 8 – 13.
63. U. Schwertmann, R. J. Rickson, K. Auerswald: "Soil Erosion Protection Measures in Europe," *Soil Technol. Ser.* **1** (1989) 216 pp.
64. W. H. Wischmeier, D. D. Smith: "Predicting Rainfall Erosion Losses - A Guide to Conservation Planning," *USDA Agric. Handbook*, Washington 1978, pp. 537.
65. G. Govers, K. Vandaele, P. Desmet, J. W. A. Poesen, K. Bunte, *Europ. J. Soil. Sci.* **45** (1994) 469 – 478.
66. H.-P. Blume: *Handbuch des Bodenschutzes*, ecomed, 2nd ed., Landsberg 1990.
67. D. Rosenkranz, G. Einsele, H.-M. Harreß: *Bodenschutz*, Erich Schmidt Verlag, Berlin 1990.
68. P. Bullock, P. J. Gregory (eds.): *Soils in the urban environment*, Blackwell, London 1991.
69. G. Luft, G. Morgenschweis, *Z. Kulturtech. Flurbereinig.* **25** (1984) 138 – 148.
70. R. Hoisl, W. Nohl, S. Zekorn, G. Zöllner: *Landschaftsästhetik in der Flurbereinigung*, Materialien zur Flurbereinigung 11, Bayer. Staatsministerium für Ernährung, Landwirtschaft und Forsten, München 1987, 161 pp.
71. H. Petelkau, *Agrartechnik* **41** (1991) 54 – 57.
72. K. G. Renard, G. R. Foster, G. A. Weesies, D. K. McCool, D. C. Yoder, *USDA Agricultural Handbook* **703** (1997) 404.
73. W. G. Knisel: "CREAMS: A Field-Scale Model for Chemicals, Runoff, and Erosion from Agricultural Management Systems," *Conserv*, Res. Rep. (U.S. Dep. Agric.) 26 (1980) 643 pp.
74. D. B. Beasley, L. B. Huggins: *ANSWERS User Manual*, Purdue University Chicago, Agric. Eng. Dep., 1981.
75. A. N. Sharpley, J. R. Williams: "EPIC-Erosion Productivity Impact Calculator: I. Model Documentation," *Tech. Bull. U.S. Dep. Agric.* 1768 (1990) 235 pp.
76. L. J. Lane, M. A. Nearing (eds.): *USDA - Water Erosion Prediction Project: Hillslope Profile Model Documentation,* National Soil Erosion Research Laboratory, report 2, West Lafayette, 1990.
77. R. P. C. Morgan *et al.*: *Earth Surface Processes and Landforms* **23** (1998) 527 – 544. Links
78. F. Scheffer, P. Schachtschabel: *Lehrbuch der Bodenkunde*, 13th ed., Enke Verlag, Stuttgart 1992.
79. H.-P. Blume: *Handbuch Bodenschutz*, ecomed, Landsberg 1990.
80. F. Korte: *Lehrbuch der Ökologischen Chemie*, 2nd ed., Thieme Verlag, Stuttgart 1987.
81. U. Gisi: *Bodenökologie*, Thieme Verlag, Stuttgart 1990.
82. B. O. Wagner, W. Mücke, H.-R. Schenk: *Umweltmonitoring – Umweltkonzentrationen organischer Chemikalien/Literatur-Recherche und -Auswertung. Angewandter Umweltschutz*, ecomed, Landsberg 1989.
83. W. Koch, K. Hurle: *Grundlagen der Unkrautbekämpfung*, Ulmer Verlag, Stuttgart 1978.
84. J. C. Ottow: "Umwandlung von Xenobiotika durch Bodenmikroorganismen und ihre potentielle Steuerung," in: "Biotechnologische In-situ-Sanierung kontaminierter Standorte," *Schriftenr. Ver. Wasser Boden Lufthyg.* **80** (1988) 61 – 92.
85. T. M. Sack, D. H. Steele: "A Survey of Household Products for Volatile Organic Compounds," *Atmos. Environ.* **26 A** (1992) no. 6, 1063 – 1070.
86. R. Schleyer, I. Renner, D. Mühlhausen: "Beeinflussung der Grundwasserqualität durch luftgetragene organische Schadstoffe," *WaBoLu-Ber.* 1991, no. 5, 1 – 96.
87. J. S. Eisenreich, B. Looney, J. D. Thoronton: "Airborne Organic Contaminants in the Great Lake Ecosystem," *Environ. Sci. Technol.* **15** (1981) 30 – 33. Links
88. T. F. Bidleman: "Atmospheric Processes," *Environ. Sci. Technol.* **22** (1988) 361. Links
89. V. Neumayr: "Verteilungs- und Transportmechanismen von chlorierten Kohlenwasserstoffen in der Umwelt," *WaBoLu-Ber.* 1981, no. 3, 24 – 40.
90. G. Milde, P. Friesel, V. Neumayr: *Verhalten flüchtiger halogenierter Stoffe in Böden auch im Hinblick auf Grundwassergefährdung*, Final Report to the Ministry of Science and Technology of the Federal Republic of Germany, no. 037210 (1983).
91. J. Förster: "Roof Runoff: A Source of Pollutants in Urban Storm Drainage Systems?," *Proc. Int. Conf. Urban Storm Drainage*, 5th, Osaka 1990, p. 469.
92. H. Frank, W. Frank, D. Thiel: "C_1- and C_2-Halocarbons in Soil-Air of Forests," *Atmos. Environ.* **23** (1989) 1333 – 1335.
93. H. Hellmann: "Leichtflüchtige Chlorkohlenwasserstoffe in den Gewässern der Bundesrepublik Deutschland – Auftreten und Bilanz," *Gesund.Ing.* **105** (1984) no. 5, 169 – 278.
94. H. D. Winkeler, U. Puttins, K. Levsen: "Organische Verbindungen im Regenwasser," *Vom Wasser* **70** (1988) 107 – 117.
95. J. Daub, T. Striebel: "Niederschlagsbedingte Schadstoffe im Straßenabfluß," in: *Schriftenreihe des Instituts für Siedlungswasserwirtschaft der Universität Karlsruhe,* no. 58: Schadstoffe im Regenabfluß aus städtischen Gebieten (1990).
96. R. Rippen *et al.*: "Do Airborne Nitrophenols Contribute to Forest Decline?" *Environ. Technol. Lett.* **8** (1987) 475 – 482.
97. P. D. Capel, T. Lichtensteiger, P. H. Brunner: "The Use of Sludge-only Landfills as Historical Records of Persistent Organic Chemicals and Heavy Metals in Sewage Sludge," *Water Research* **23** (1989) 525 – 527. Links
98. U. Pagga: "Biologischer Abbau von Stoffen bei geringen Konzentrationen," *Z. Wasser Abwasser Forsch.* **20** (1987) 101 – 107.
99. NRW-Meßprogramm: "Chloraromaten – Herkunft und Transfer, Abschlußbericht 1991," Ministerium für Raumordnung und Landwirtschaft des Landes Nordrhein-Westfalen, Düsseldorf 1991.
100. N. Litz, M. Thiele, H. W. Döring, H.- P. Blume: "The Behaviour of Linear Alkylbenzenolsulfonates in Different Soils: A Comparison between Field and Laboratory Studies," *Ecotoxicol. Environ. Saf.* **14** (1987) 103 – 106. Links
101. R. Leschber, N. Litz: "Klärschlammverwertung und -beseitigung im Licht der Klärschlammbelastung mit organischen Schadstoffen. Verfahrenstechnik der mechanischen, thermischen, chemischen und biologischen Abwasserbehandlung," *2nd GVC-Congress*, Würzburg, Oct. 19 – 21, 1992, VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen, pp. 477 – 482.
102. N. Litz: "Abschätzung des Verhaltens von Umweltchemikalien in Böden," 2. Internationale Tagung und Ausstellung über Umweltinformation und Umweltkommunikation, Bayreuth, Sep. 14 – 18, 1993, *ECOINFORMA* **2** (1992) 323 – 335.
103. D. Pernak, R. Leschber: "Tiefenverteilung persistenter Schadstoffe im Sediment einer Regenwasserversickerungsanlage," *Vom Wasser* **79** (1992) 321 –331.
104. Klärschlammverordnung (AbfKlärV) of April 28, 1992, Bundesgesetzblatt, part 1, pp. 912 – 928.
105. R. Leschber, R. Mergler-Völkl, U. Zimmermann, N. Litz: *Untersuchung von relevanten organischen Schadstoffen im Klärschlamm*, Forschungsbericht 10303521 an das Umweltbundesamt, 1992.
106. K. R. Bull: "The Critical Loads/Levels Approach to Gaseous Pollutant Emission Control," *Environ. Pollut. Ser. A* **69** (1991) 105 – 123.
107. A. T. Kan, M. B. Tomson: "Groundwater Transport of Hydrophobic Organic Compounds in the Presence of Dissolved Organic Matter," *Environ. Toxicol. Chem.* **9** (1990) 253 – 263. Links
108. R. L. Wershaw: "Application of a Membrane Model to the Sorptive Interactions of Humic Substances," *EHP Environ. Health Perspect.* **83** (1989) 191 – 203.
109. P. Nkedi-Kizza, P. Rao, A. Hornsby: "Influence of Organic Solvents on Sorption of Hydrophobic Organic Chemicals by Soils," *Environ. Sci. Technol.* **19** (1985) 975 – 979. Links
110. B. T. Walton *et al.*: "Soil Sorption of Volatile and Semivolatile Organic Compounds in a Mixture," *J. Environ. Qual.* **21** (1992) 552 – 558. Links
111. K. B. Woodburn, P. S. C. Rao, M. Fukui, P. Nkedi-Kizza: "Solvophobic Approach for Predicting Sorption of Hydrophobic Organic Chemicals on Synthetic Sorbents and Soils," *J. Contam. Hydrogeology* 1986, 227 – 241.
112. B. Oepen, W. Kördel, W. Klein: "Sorption of Nonpolar and Polar Compounds to Soils: Processes, Measurements and Experience with the Applicability of the Modified OECD-Guideline 106," *Chemosphere* **22** (1991) 285 – 304.
113. Y. S. Yu, G. W. Bailey: "Reduction of Nitrobenzene by Four Sulfide Minerals: Kinetics, Products and Solubility," *J. Environ. Qual.* **21** (1992) 86 – 94. Links
114. K. D. Pennell, R. D. Rhue, A. G. Hornsby: "Competitive Adsorption of Para-xylene and Water Vapors on Calcium, Sodium, and Lithium-saturated Kaolinite," *J. Environ. Qual.* **21** (1992) 419 – 426. Links
115. J. D. Ortego, M. Kowalska, D. L. Cocke: "Interactions of Montmorillonite with Organic Compounds – Adsorptive and Catalytic Properties," *Chemosphere* **22** (1991) no. 8, 769 – 798. Links
116. I. Scheunert, M. Mansour, F. Andreux: "Binding of Organic Pollutants to Soil Organic Matter," *Int. J. Environ. Anal. Chem.* **46** (1991) 189 – 199.
117. I. Scheunert, C. Meer-Beek, F. Korte: "Distribution and Biodegradation of ^{14}C-Residues Bound in Various Soilfractions after Treatment of the Soil with Modern ^{14}C-Chemicals," *Proceedings of the final Research Co-ordination Meeting on Isotopic Tracer-aided Studies of Unextractable or Bound Pesticide Residues in Soil, Plant, and Food*, International Atomic Energy Agency, Vienna 1986, pp. 31 – 40.
118. N. Kamal, I. Scheunert, F. Korte: "Mass Balance of Pentachloronitrobenzene and Metabolite in a Closed Aerated Soil-plant- or Soil-system," *Bull. Environ. Contam. Toxicol.* **31** (1983) 559 – 565. Links
119. D. G. Crosby, A. S. Wong: "Environmental Degradation of 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)," *Science* (Washington, D.C. 1883) 195 (1977) 1337 – 1338.
120. I. Gebefügi, R. Baumann, F. Korte: "Photochemical Degradation of 2,3,7,8-TCDD under Simulated Environmental Conditions," *Naturwissenschaften* **64** (1977) 486.
121. M. L. Rochkind-Dubinski, G. S. Sayler, J. W. Blackburn: "Microbial Decomposition of Chlorinated Aromatic Compounds," *Microbiol. Ser.* **18** (1987) 314 pp.
122. R. C. Bayly, M. G. Barbour: "The Degradation of Aromatic Compounds by the Meta and Gentisate Pathways: Biochemistry and Regulation," in D. T. Gibson (ed.): "Microbial Degradation of Organic Compounds," *Microbiol. Ser.* **13** (1984) 253 – 294.
123. L. Britton: "Microbial Degradation of Aliphatic Hydrocarbons," in D. T. Gibson (ed.): "Microbial Degradation of Organic Compounds," *Microbiol. Ser.* **13** (1984) 89 – 130.
124. D. T. Gibson, V. Subramanian: "Microbial Degradation of Aromatic Hydrocarbons," in D. T. Gibson (ed.): "Microbial Degradation of Organic Compounds," *Microbiol. Ser.* **13** (1984) 181 – 252.
125. G. Hickman, J. Novak, M. Morris, M. Rebhuhn: "Effects of Site Variations on Subsurface Biodegradation Potential," *J. Water Pollut. Control. Fed.* **61** (1989) 1564 – 1575.
126. A. L. Young: "Long-term Studies on the Persistence and Movement of TCDD in a Natural Ecosystem," *Environ. Sci. Res.* **26** (1983) 173 – 190.
127. P. Geschwend, S. Wu: "On the Constancy of Sediment-water Partition Coefficients of Hydrophobic Organic Compounds," *Environ. Sci. Technol.* **19** (1985) 90 – 96.
128. K. E. Quentin *et al.*: *Halogenorganische Verbindungen in Wässern*, Forschungsbericht 102 04 323 im Auftrag des Umweltbundesamtes, Berlin 1987.

129. K. Ballschmiter: “Transport und Verbleib organischer Verbindungen im globalen Rahmen,” *Angew. Chem.* **104** (1992) 501 – 528. Links
130. H. Seip, J. Alstad, G. Carlberg, K. Martinsen: “Measurement of Mobility of Organic Compounds in Soils,” *Sci. Total Environ.* **50** (1986) 87 – 101. Links
131. B. H. Wilson, G. B. Smith, J. F. Rees, *Environ. Sci. Technol.* **20** (1986) 997. Links
132. T. Vogel, C. Criddle, P. McCartey: “Transformations of Halogenated Aliphatic Compounds,” *Environ. Sci. Technol.* **21** (1987) 722 – 736. Links
133. T. Vogel, P. McCartey: “Biotransformation of Tetrachloroethylene to Trichloroethylene, Dichloroethylene, Vinyl Chloride, and Carbon Dioxide under Methanogenic Conditions,” *Appl. Environ. Microbiol.* **49** (1985) 1080 – 1083. Links
134. B. Zoeteman *et al.*: “Persistent Organic Pollutants in River Water and Groundwater of The Netherlands,” *Chemosphere* **9** (1981) 231 – 249.
135. M. Nerger, R. Mergler-Völkl: “Biologischer Abbau von leichtflüchtigen Chlorkohlenwasserstoffen in Grund- und Abwasser,” *Z. Wasser Abwasser Forschg.* **21** (1988) 16 – 19.
136. H. Scholz-Muramatsu, H.-C. Flemming: “Unter welchen Milieubedingungen erfolgt ein mikrobieller Abbau leichtflüchtiger chlorierter Kohlenwasserstoffe (LCKW) ?” in R. Wagner (ed.) *Wasserkalender 1991,* Erich Schmidt Verlag, Berlin 1991, pp. 135 – 158.
137. V. Schneider, H. Scholz-Muramatsu, S. Gaiser: “Hemmung der biologischen Elimination von Dichlormethan durch kontaminiertes Grundwasser,” *GWF Gas Wasserfach. Wasser/Abwasser* **129** (1988) 755 – 758.
138. A. B. Zehnder, G. Schraa: “Biologischer Abbau xenobiotischer Verbindungen bei niedrigen Konzentrationen und das Problem der Restkonzentration,” *GWF Gas Wasserfach. Wasser/Abwasser* **129** (1988) 369 – 373.
139. L. Semprini, D. Grbic-Galic, P. L. McCartey, P. V. Roberts: “Introduction,” in P. V. Roberts et al. (eds.) “In situ Aquifer Restoration of Chlorinated Aliphatics by Methanotrophic Bacteria,” *Tech. Rep. Dep. Civil Eng. Univ. Stanford* **310** (1989) 65 – 90.
140. B. Toussaint, H.-R. Langguth: “Das Reinigungsvermögen des Untergrundes – Vorstellungen, Wünsche und Wirklichkeit,” *Wiss. Umwelt* 1991, 1 – 2.
141. World Health Organization: “Chlorobenzenes other than Hexachlorobenzene,” *Environ. Health Crit.* **128** (1991)
142. BUA-Report, (Beratergremium für Umweltrelevante Altstoffe) *24: Benzene,* VCH Verlagsgesellschaft, Weinheim 1992.
143. ATV (Abwassertechnische Vereinigung e.V.): “Zusammenfassender Bericht über einige ökotoxikologisch gefährliche Stoffe,” *Dokumentation und Schriftenreihe aus Wissenschaft und Praxis 27*, St. Augustin 1991.
144. R. Schleyer, G. Milde: “Zur Bewertung luftbürtiger Grundwasserqualitätsbeeinflussungen,” *VDI Ber.* **837** (1990) 663 – 684.
145. M. P. Ligocki, C. Leuenberger, J. F. Pankow: “Organic Compounds in Rain – II. Gas Scavenging of Neutral Organic Compounds,” *Atmosp. Environ.* **19** (1985) 1609 – 1617.
146. Bundesamt für Umweltschutz: “Polychlorierte organische Verbindungen in Rheinfelden (Schweiz),” *Schriftenreihe Umweltschutz* **18** (1983) .
147. R. Seibold, W. Arnold, R. Schmid: “Persistente Chlorkohlenwasserstoffe in Böden, Klärschlämmen und Flußsedimenten,” *Hohenheimer Arb.* 1986, 65 – 76.
148. A. Reischl, M. Reissinger, O. Hutzinger: “Organische Luftschadstoffe und ihre Bedeutung für die terrestrische Vegetation,” *UWSF Z. Umweltchem. Ökotox.* **2** (1989) 32 – 41.
149. W. König, A. Hembrock-Heger, M. Wilkens: “Persistente organische Chemikalien im Boden – Eintragspfade und Vorkommen,” *UWSF Z. Umweltchem. Ökotox.* **3** (1991) no. 1, 33 – 36.
150. P. Weiss, A. Riss: “Bodenuntersuchungen von Grünland- und Waldböden im Raum Linz auf Schwermetalle, polyzyklische aromatische Kohlenwasserstoffe und chlorierte Kohlenwasserstoffe,” in: “Schadstoffe im Raum Linz,” *Monographie,* **vol. 20**, Bundesministerium für Umwelt, Jugend und Familie, Wien 1992.
151. T. A. Anderson, J. J. Beauchamp, B. T. Walton: “Organic Chemicals in the Environment,” *J. Environ. Qual.* **20** (1991) 420 – 424. Links
152. Y. Jin, G. O'Connor: “Behavior of Toluene Added to Sludge-Amended Soils,” *J. Environm. Qual.* **19** (1990) 573 – 579.
153. E. Merian, M. Zander: “Volatile Aromatics,” in O. Hutzinger (ed.) *Handbook Environ. Chem.* **3** (1982) 117 – 161.
154. M. Reinhard, N. L. Goodman, J. F. Barker, *Environ. Sci. Technol.* **18** (1984) 953. Links
155. C. Y. Chiang *et al.*: “Aerobic Biodegradation of Benzene, Toluene, and Xylene in a Sandy Aquifer – Data Analysis and Computer Modeling,” *Ground Water* **27** (1989) no. 6, 823 – 834. Links
156. D. Grbic-Galic, T. M. Vogel, *Appl. Environ. Microbiol.* **52** (1986) 200. Links
157. J. P. van der Hock, L. G. C. M. Urlings, C. M. Grobben: “Biological Removal of Polycyclic Aromatic Hydrocarbons, Benzene, Toluene, Ethylbenzene, Xylene, and Phenolic Compounds from Heavily Contaminated Ground Water and Soil,” *Environ. Technol. Lett.* **10** (1989) 185 – 194.
158. D. M. Falatko, J. T. Novak: “Effects of Biologically Produced Surfactants on the Mobility and Biodegradation of Petroleum Hydrocarbons,” *Water Environ. Res.* **64** (1992) no. 2, 163 –169. Links
159. K. G. Robinson, W. S. Farmer, J. T. Novak: “Availability of Sorbed Toluene in Soils for Biodegradation by Acclimated Bacteria,” *Water Res.* **24** (1990) 345 – 350. Links
160. R. Herterich, R. Herrmann: “Comparing the Distribution of Nitrated Phenols in Atmosphere of German Hill Sites,” *Environ. Technol.* **11** (1990) 961 – 972. Links
161. B. Scholz, N. Palauschek: “The Determination of Substituted Aromatic Amines in Water and Sediment Samples,” *Fresenius Z. Anal. Chem.* **331** (1988) 282 – 289.
162. J. Zachara, J. Lawrence, J. Sauer: “Sorption on Aniline on Selected Alfisols from the Eastern Coal Region,” *Soil Sci.* **138** (1984) 209 – 219. Links
163. G. Sims, L. E. O'Loughlin: “Degradation of Pyridines in the Environment,” *CRC Crit. Rev. Environ. Control.* **19** (1989) 309 – 340.
164. J. A. Leenheer, H. A. Stuber: “Migration Through Soil of Organic Solutes in Oil-shale Process Water,” *Environ. Sci. Technol.* **15** (1981) 578. Links
165. G. K. Sims, E. Sommers: “Biodegradation of Pyridine Derivates in Soil Suspension,” *Environ. Toxciol. Chem.* **5** (1985) 503 – 509.
166. D. H. Stuermer, D. J. Ng, C. J. Morris: “Organic Contaminants in Groundwater Near an Underground Coal Gasification in Northeastern Wyoming,” *Environ. Sci. Technol.* **16** (1982) 582. Links
167. R. F. Spalting, J. W. Fulton: “Groundwater Munitions Residues and Nitrate Near Grand Island, Nebraska, USA,” *J. Contam. Hydrol.* **2** (1988) 139 – 153.
168. A. Oberholz: *Tödliche Gefahr aus der Tiefe. Bittere Erkenntnisse zu Kriegs- und Rüstungsaltlasten*, Kommunal-Verlag, Düsseldorf 1991.
169. B. Ehring, J. Prassler, P. R. Wallnöfer, W. Ziegler: “Mikrobielle Transformation von o-Nitroanilin und p-Nitrophenol,” *UWSF Z. Umweltchem. Ökotox.* **4** (1992) 81 – 85.
170. E. P. Kuhn, J. M. Suflita: “Microbial Degradation of Nitrogen, Oxygen, and Sulfur Heterocyclic Compounds under Anaerobic Conditions: Studies with Aquifer Samples,” *Environ. Toxicol. Chem.* **8** (1989) 1149 – 1158. Links
171. W. Rothard: “Aktuelles zur Dioxinproblematik – Ableitung von Dioxinrichtwerten für die Bodensanierung,” *Bundesgesundheitsbl.* **4** (1991) 155 – 158.
172. Environmental Health Criteria 88: *Polychlorinated Dibenzopara-dioxins and Dibenzofurans*, World Health Organization, Geneva 1989.
173. O. Hutzinger, H. Fiedler: “From Source to Exposure: Some Open Questions,” *Chemosphere* **27** (1993) 121 – 129. Links
174. H. Kaune, H. Fiedler, O. Hutzinger: *Dioxine und Furane – Quellen, Einträge in die Umwelt und Aufnahme durch den Menschen*, Im Auftrag der Hamburger Umweltbehörde, Bayreuth 1990.
175. H. Fiedler, O. Hutzinger: “Sources and Sinks of Dioxins: Germany,” *Chemosphere* **25** (1992) no. 10, 1487 – 1491. Links
176. M. Büchen *et al.*: *Dioxine und Furane in der hessischen Umwelt – Meßergebnisse aus Hessen*, Hessische Landesanstalt für Umwelt, Umweltplanung, Arbeits- und Umweltschutz, no. 126, Wiesbaden 1991.
177. First VGA/UBA-Evaluation of the Internationale Dioxin Symposium and of the Public Expert Hearing of 09. – 13. Nov. 1992, Berlin, 28. Jan. 1993.
178. L. O. Kjeller, C. Rae, C. K. C. Jones, A. E. Johnston: “Evidence for Increases in the Environmental Burden of Polychlorinated Dibenzo-p-Dioxins and Furans (PCDD/PCDFs) over the last Century,” *Chemosphere* (1990) in K. Meyer: Bodenverschmutzung in der Schweiz. Themenbereich des Nationalen Forschungsprogrammes “Boden,” Liebenfeld – Bern, 1991.
179. R. M. Smith *et al.*: “Measurement of PCDFs and PCDDs in Air Samples and Lake Sediments at Several Locations in Upstate New York,” *Chemosphere* **25** (1992) nos. 1 – 2, 95 – 98. Links
180. Ministerium für Umwelt Baden-Württemberg (eds.): *Umweltdaten 89/90*, Stuttgart 1990.
181. Bayerisches Geologisches Landesamt: “Dioxine in Böden Bayerns,” *GLA Fachberichte* **7** (1991) 1 – 96.
182. A. Riss, H. Hagenmaier, W. Rotard: “Wirkungen von Dioxinimmissionen auf Boden, Grünlandaufwuchs und Kuhmilch – Fallstudie anhand einer Metallrückgewinnungsanlage in Österreich,” in: *VDI Ber.* **837** (1990) 831 – 848.
183. S. Karickoff: “Semiempirical Estimation of Sorption of Hydrophobic Pollutants on Natural Sediments and Soils,” *Chemosphere* **10** (1981) 833 – 346.
184. W. Y. Shiu *et al.*: “Physical-Chemical Properties of Chlorinated Dibenzo-p-dioxins,” *Environ. Sci. Technol.* **22** (1988) 651 – 658. Links
185. S. Kapila *et al.*: “Field and Laboratory Studies on the Movement of Tetrachlorodibenzo-p-dioxin in Soil,” *Chemosphere* **18** (1989) nos. 1 – 6, 1297 – 1304.
186. H.-J. Jürgens, R. Roth: “Case Study and Proposed Decontamination Steps of the Soil and Groundwater Beneath a Closed Herbicide Plant in Germany,” *Chemosphere* **18** (1989) nos. 1 – 6, 1163 – 1169.
187. R. K. Puri *et al.*: “Studies of Parameter Affecting Translocation of Tetrachlorodibenzo-p-dioxin in Soil,” *Chemosphere* **18** (1989) 1291.
188. D. Dougherty, L. Peters, M. Overcash, R. Carbonelle: “Transport of 2,3,7,8-Tetrachlorodibenodioxin in Soil Containing Organic Solvents,” *Organohalogen Compounds* **3** (1991) 315 – 318.
189. R. Orth, C. Ritchie, F. Hileman: “Measurement of the Photoinduced Loss of Vapor Phase TCDD,” *Chemosphere* **18** (1989) 1275 – 1282.
190. G. C. Miller *et al.*: “Photolysis of Octachlorodibenzo-p-dioxin on Soils: Production of 2,3,7,8-TCDD,” *Chemosphere* **18** (1989) nos. 1 – 6, 1265 – 1274.
191. J. K. McCrady, S. P. Maggard: “Uptake and Photodegradation of 2,3,7,8-Tetrachlorodibenzo-p-dioxin Sorbed to Grass Foliage,” *Environ. Sci. Technol.* **27** (1993) 343 – 350. Links
192. S. Cerlesi, A. Di Domenico, S. Ratti: “2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD) Persistence in the Seveso (Milan, Italy) Soil,” *Ecotoxicol. Environ. Saf.* **18** (1989) 149 – 164. Links
193. A. Yanders, S. Kapila, R. Schreiber in C. Rappe, G. Choudhary, L. Keith (eds.): *Chlorinated Dioxins and Dibenzofuran in Perspective*, Lewis Publ., Chelsea, Michigan, 1986, pp. 237 – 239.
194. R. A. Freeman, J. M. Schroy, *Chemosphere* **18** (1989) 1305.
195. A. L. Young, L. G. Cockerham, C. E. Thalken: “A Long-term Study of Ecosystem Contamination with 2,3,7,8-Tetrachlorodibenzodioxin,” *Chemosphere* **16** (1987) 1791 – 1815. Links
196. E. Arvin, B. Jensen, J. Aamand, C. Jorgensen: “The Potential of Free Living Ground Water Bacteria to Degrade Aromatic Hydrocarbons and Heterocyclic Compounds,” *Water Sci.*

Technol. **20** (1988) 109. Links
197. G. Springer, H. G. Rast: "Biologischer Abbau mehrfach halogenierter mono- und polyzyklischer Aromaten," *GWF Gas Wasserfach*: Wasser/Abwasser 129 (1988) 70.
198. D. Frahne *et al.*: "Zum biologischen Abbau von Dibenzofuran im Hinblick auf doxinähnliche Verbindungen," *Fat Sci. Technol.* **93** (1991) no. 9, 355 – 360.
199. P. Fortnagel, R.-M. Wittich, H. Harms, S. Schmidt: "New Bacterial Degradation of the Biaryl Structure," *Naturwissenschaften* **76** (1989) 523 – 524. Links
200. M. F. Arthur, J. I. Frea: "2,3,7,8-Tetrachlorodibenzo-p-dioxin: Aspects of its Important Properties and its Potential Biodegradation in Soils," *J. Environ. Qual.* **18** (1989) no. 1, 1 – 11. Links
201. G. Klecka, D. Gibson: "Metabolism of Dibenzo-p-dioxin and Chlorinated Dibenzo-p-dioxins by a Beijerinckia Species," *Appl. Environ. Microbiol.* **39** (1980) 288 – 296. Links
202. K. R. Cooper, A. Christini, P. Bergqvist, C. Rappe: "Bioavailability and Bioconcentration of Polychlorinated Dioxins (PCDD) and Furans (PCDF) to Organisms Inhabiting a Heavily Contaminated Estuarine Ecosystem," *Chemosphere* **25** (1992) nos. 1 – 2, 25 – 28. Links
203. G. F. Fries: "Assessment of Potential Residues in Food Derived from Animals Exposed to TCDD-Contaminated Soil," *Chemosphere* **16** (1987) 2123 – 2128. Links
204. M. S. McLachlan, H. Thoma, M. Reissinger, O. Hutzinger: "PCDD/F in Agricultural Food Chain Part 1: PCDD/F Mass Balance of a Lactating Cow," *Chemosphere* **20** (1990) nos. 7 – 9, 1013 – 1020. Links
205. C. S. Nessel *et al.*: "Transpulmonary Uptake and Bioavailibility of 2,3,7,8-TCDD from Respirable Soil Particles," *Chemosphere* **25** (1992) nos. 1 – 2, 29 – 23. Links
206. W. Feige, R. Schmalzried: *Stoffbestand von Gartenböden*, Bayerisches Staatsministerium für Landesentwicklung und Umweltfragen, München 1991.
207. VROM: Leidraad Bodembescherming, Staatsuitgeverij, Den Haag 1983/1990.
208. K. C. Jones *et al.*: "Increase in the Polynuclear Aromatic Hydrocarbon Content of an Agricultural Soil over the Last Century," *Environ. Sci. Technol.* **23** (1989) no. 1, 95 – 101. Links
209. A. Sablijic: "On the Prediction of Soil Sorption Coefficients of Organic Pollutants from Molecular Structure: Application of Molecular Topology Model," *Environ. Sci. Technol.* **21** (1987) 358 –365.
210. S. R. Wild, K. C. Jones: "Biological and Abiotic Losses of Polynuclear Aromatic Hydrocarbons (PAHs) from Soils Freshly Amended with Sewage Sludge". *Environ. Toxciol. Chem.* **12** (1993) 5 – 12.
211. S. R. Wild, M. L. Berrow, K. C. Jones: "The Persistence of Polynuclear Aromatic Hydrocarbons (PAHs) in Sewage Sludge Amended Agricultural Soils," *Environ. Pollut. Ser. A* **72** (1991) 141 – 157.
212. W. D. Weissenfels, M. Beyer, J. Klein: "Degradation of Phenantrene, Fluorene, and Fluoranthene by Pure Bacterial Cultures," *Appl. Microbiol. Biotechnol.* **32** (1990) 479 – 484. Links
213. K. S. Park et al.: "Fate of PAH Compounds in Two Soil Types: Influence of Volatilization, Abiotic Loss and Biological Activity", *Environ. Toxicol. Chem.* **9** (1990) 187 – 195.
214. S. R. Wild, K. C. Jones, A. E. Johnston: "The Polynuclear Aromatic Hydrocarbon (PAH) Content of Herbage from a Long-term Grassland Experiment," *Atmos. Environ.* **26 A** (1992) no. 7, 1299 – 1307.
215. S. R. Wild, M. L. Berrow, S. P. McGrath, K. C. Jones: "Polynuclear Aromatic Hydrocarbons in Crops from Long-termed Field Experiments Amended with Sewage Sludge," *Environ. Pollut. Ser. A* **76** (1992) 25 – 32.
216. W. Pyka, C. Schüth, P. Grathwohl: "Dissolution and Transport of Coaltar Constituents and their Impact on Groundwater Quality," *Int. Symp. Environ. Contam. Central Eastern Europe*, Budapest 1992, Oct. 12 – 16, 1992.
217. A. Cortes, J. Riego, A. B. Paya-Perez, B. Larsen: "Soil Sorption of Co-Planar and Non-Planar PCB's," *Toxicol. Environ. Chem.* **31 – 32** (1991) 79 – 86.
218. A. B. Paya-Perez, F. Pelusio: "Sorption of 3,4,5-Trichloroaniline in a Soil Profile," *Sci. Total Environ.* **123/124** (1992) 519 – 529.
219. J. F. Brown *et al.*: "Polychlorinated Biphenyl Dechlorination in Aquatic Sediments," *Science* (Washington, D.C., 1883) 236 (1987) 709 – 712.
220. A. Marcomini et al.: "Behaviour of Aromatic Surfactants and PCB's in Sludge-treated Soil and Landfills," *J. Environ. Qual.* **18** (1989) 523 –528.
221. W. Kampe: "Organische Stoffe in Böden und Pflanzen nach langjährigen, intensiven Klärschlammanwendungen," *Korresp. Abwasser* **43** (1987) no. 8, 820 – 827.
222. S. W. Hooper, C. A. Pettigrew, G. S. Sayler: "Ecological Fate Effects and Prospects for the Elimination of Environmental Polychlorinated Biphenyls," *Environ. Toxicol. Chem.* **9** (1990) 655 –667. Links
223. D. Pal, J. Weber, M. Overcash: "Fate of Polychlorinated Biphenyls (PCB) in Soil-plant System," *CSIR Publ. Rev.* **74** (1980) 45 – 98.
224. S. Ketterer, J. Michel, W. Holzwarth: "Verhalten von polychlorierten Biphenylen (PCB) in der wasserungesättigten Bodenzone eines kontaminierten Standortes", Bayreuth, Sep. 14 – 18, 1992 *ECOINFORMA* **2** (1992).
225. S. Boyd, S. Sun, *Environ. Qual.* **24** (1990) 1920 –1923.
226. G. A. O'Connor, R. L. Chaney, J. A. Ryan: "Bioavailability of Plants of Sludge-Borne Toxic Organics," *Rev. Environ. Toxicol.* **121** (1991) 129 – 155.
227. N. Litz: "Verhalten von Umweltchemikalien in Böden – Bewertungsverfahren für die Umweltplanung," *Z. Pflanzenernähr. Bodenkd.* **156** (1993) 203 – 208.
228. G. F. Fries, G. S. Marrow: "Influence of Soil Properties on the Uptake of Hexachlorobiphenyls by Rats," *Chemosphere* **24** (1992) 109 – 113. Links
229. Q. Ye *et al.*: "Studies on the Transport and Transformation of PCBs in Plants," *Chemosphere* **25** (1992) 1475 – 1479. Links
230. S. Winter, B. Streit: "Organochlorine Compounds in a Three Step Terrestrial Food Chain," *Chemosphere* **24** (1992) no. 12, 1765 – 1774. Links
231. VDI (ed.): *Acidic precipitation – Formation and Impact on Terrestrial Ecosystems*, VDI Verlag, Düsseldorf 1983, 1987, 281 p.
232. H.-W. Georgii (ed.): *Mechanisms and Effects of Pollutant-Transfer into Forests*, Kluwer Academic Publ., Dordrecht 1989, p. 361.
233. B. Ulrich, *Z. Pflanzenernähr. Bodenkd.* **149** (1986) 703 – 717.
234. J. O. Reuss, D. W. Johnson: *Acid Deposition and the Acidification of Soils and Waters*, vol. **8**, Springer Publ., New York 1986. , pp. 247 – 256.
235. B. Ulrich, *Z. Pflanzenernähr. Bodenkd.* **144** (1981) 289 – 305.
236. H. D. Nagel, H.-D. Gregor: *Ökologische Belastungsgrenzen — Critical Loads & Levels*. Springer Verlag, Berlin 1999, p. 259.
237. Umweltbundesamt (ed.): *Daten zur Umwelt 1992/93*, Erich Schmidt Verlag, Berlin 1994.
238. B. Manderscheid, E. Matzner, K. J. Meiwes, Y. Xiu: "Long-term Development of Element Budgets in a Norway Spruce Forest of the German Solling Area," *Water, Air and Soil Pollution* **79** (1995) 3 – 18.
239. B. Ulrich, *Forstarchiv* **58** (1987) 232 – 239.
240. R. Mayer, *VDI Ber.* **837** (1990) 295 – 303.
241. D. L. Godbold: "Aluminum Toxicity in Trees — 15 years on," *Ber. Forschungszentr. Waldökosysteme/Waldsterben Göttingen A* **154** (1998) 154 – 164.
242. J. Prenzel in B. Ulrich, J. Pankrath (eds.): *Effects of Accumulation of Air Pollutants in Forest Ecosystems*, D. Reidel Publ. Comp., Dordrecht 1983, pp. 157 – 170.
243. B. Ulrich in B. Ulrich, J. Pankrath (eds.): *Effects of Accumulation of Air Pollutants in Forest Ecosystems*, D. Reidel Publ. Comp., Dordrecht 1983, pp. 1 – 20.
244. W. Flückiger in G. Glatzel (ed.): *Möglichkeiten und Grenzen der Sanierung immissionsgeschädigter Waldökosysteme*, Institut für Forstökologie der Hochschule für Bodenkultur, Wien 1987, pp. 65 –81.
245. B. Ulrich, *Experentia* **41** (1985) 578 – 584.
246. H. Papke, H. Papen: "Influence of Acid Rain and Liming on Fluxes of NO and NO_2 from Forest Soil," *Plant Soil* **199** (1998) 131 – 140. Links
247. T. Gjessing, G. Riise, E. Lydersen: "Acid rain and natural organic matter (NOM)," *Acta hydrochimica et hydrobiologica* **26** (1998) 131 – 136. Links
248. B. Ulrich, H. Meyer, *Ber. Forschungszentrum Waldökosysteme Göttingen* **6** (1978) 133 pp.
249. O. Wittmann, K. D. Fetzer: "Aktuelle Bodenversauerung in Bayern," *Bayer. Staatsmin. Landesentw. Umwelt. Materialien* **20** (1982) 71 pp.
250. J. Nilsson (ed.): *Critical Loads for Sulfur and Nitrogen*, Nordisk Ministerrad, Miljo Rapport 1986.
251. M. Berdén, S. I. Nilsson, K. Rosen, G. Tyler: "Soil Acidification: Extent, Causes, Consequences," *Natl. Swed. Environ. Protection Board Rep.* **3292** (1987) 164 pp.
252. E. v. Zezschwitz, *Allg. Forst. Jagdztg.* **158** (1987) 136 – 147.
253. H.-P. Blume: "Alarmierende Versauerung Berliner Forsten," *Berliner Naturschutzblätter* **75** (1981) 713 – 715.
254. H. Butzke: "Versauern unsere Wälder? Erste Ergebnisse der Überprüfung 20 Jahre alter pH-Messungen in Waldböden Nordrhein-Westfalens," *Forst Holzwirt* **36** (1981) 542 – 548.
255. R. Riebeling, Ch. Schaefer: "Jahres- und Langzeitentwicklung der pH-Werte von Waldböden in hessischen Fichtenbeständen," *Forst Holzwirt* **39** (1984) 177 – 178.
256. N. Rastin, B. Ulrich: "Bodenkundliche Standortscharakterisierung zur Beurteilung des Stabilitätszustands von Waldökosystemen in Hamburg," *Ber. Forschungszentrum Waldökosysteme Göttingen* **10** (1985) 1 – 91.
257. E. v. Zezschwitz, *Geol. Jahrb. Reihe F.* **20** (1985) 3 – 41.
258. U. Falkengren-Grerup, *Environ. Pollut. Ser. A* **43** (1987) 79 – 90.
259. M. Hauhs, *Ber. Forschungszentr. Waldökosysteme Göttingen* **17** (1985) 1 – 206.
260. R. Grimm, K. E. Rehfuess, *Allg. Forst. Jagdztg.* **157** (1986) 205 – 213.
261. K. Rost-Siebert, *Texte Umweltbundesamt* 1, Deutsches Umweltbundesamt, Berlin 1986.
262. L. Hallbäcken, *Z. Pflanzenernähr. Bodenkd.* **155** (1992) 51 – 60.
263. H. Meyer-Wenke: "Bodenversauerung in naturnahen Waldökosystemen der Mittelgebirge Westdeutschlands zwischen 1950/70 und 1990/91," *Ber. Forschungszentr. Walsökosysteme/Waldsterben Göttingen A* **140** (1996) 129 S.
264. H. Koenies, Ph. D. Thesis, Universität Kassel 1982.
265. G. Glatzel, M. Kazda, L. Lindebuer, *Düsseldorfer Geobot. Kolloq.* **3** (1986) 15 – 32.
266. R. Wittig, H. Neite, *Vegetatio* **64** (1985) 113 – 119.
267. A. Schulte, M. Spiteller, *Forst Holzwirt* **42** (1987) 150 – 154.
268. H. Jochheim, H. Schäfer, *Z. Pflanzenernähr. Bodenkd.* **151** (1988) 81 – 85.

269. W. Hetsch, K. Meiwes, B. Ulrich, *Mitt. Forstl. Bundesversuchsanst. Wien* **140** (1981) 223 – 230.
270. E. E. Hildebrand, *Forstwiss. Centralbl.* **105** (1986) 60 – 76.
271. J. Puhe, B. Ulrich, *Arch. Hydrobiol.* **102** (1985) 331 – 342. Links
272. H. Heinrichs *et al.*, *Neues Jahrb. Mineral. Abh.* **156** (1986) 23 – 62.
273. M. Brechtel, *Forst Holzwirt* **43** (1988) 298 – 302.
274. R. Hantschel, *Bayreuther Bodenkd. Ber.* **3** (1987) pp. 219.
275. S. Linkersdörfer, P. Benecke: *Auswirkungen von sauren Depositionen auf die Grundwasserqualität in bewaldeten Gebieten. Eine Literaturstudie*, Materialien des Umweltbundesamtes no. 4, Erich Schmidt Verlag, Berlin 1987, pp. 170.
276. M. Hauhs, R. F. Wright (undated): "Acid Deposition: Reversibility of Soil and Water Acidification," *Air Pollution Research Report* 11, Commission of the European Communities, Luxemburg pp. 42.
277. M. Schmidt, *Allgem. Forst Z.* **30/31** (1990) 790 – 792.
278. V. Malessa: "Depositionsbedingte Tiefengradienten der Bodenversauerung in der Sösemulde (Westharz)," *Ber. Forschungszentr. Waldökosysteme Göttingen* **98** (1993) 223 pp.
279. F. Beese, K. I. Meiwes: "10 Jahre Waldkalkung — Stand und Perspektiven," *AFZ* **50** (1995) 946 – 949.
280. F. Wright, L. Rasmussen: "Introduction to the NITREX and EXMAN projects," *For. Ecol. Manag.* **101** (1998) 1 – 8.
281. H. Hultberg, R. A. Skeffington (eds.): *Experimental Reversal of Rain Effects: Gardsjoen Roof Project*, John Wiley & Sons, Chichester 1998, p. 466.
282. H. Vetter, G. Steffens: *Wirtschaftseigene Düngung*, DLG-Verlag, Frankfurt/Main 1986.
283. Bundesministerium für Ernährung, Landwirtschaft und Forsten (ed.): *Statistisches Jahrbuch über Ernährung*, Landwirtschaft und Forsten, Landwirtschaftsverlag Münster, Münster 1994.
284. Deutscher Bundestag, ENQUETE-KOMMISSION: "Vorsorge zum Schutz der Erdatmosphäre," 3. Bericht zum Thema Schutz der Erde, Drucksache 11/8030, 1990. J. Jung, K. Isermann, G. Henjes: "Der Einfluß von cadmiumhaltigen Düngerphosphaten auf die Cadmiumanreichungen von Kulturböden und Nutzpflanzen," *Landwirtsch. Forsch.* **32** (1979) 262 –282.
285. J. Baier, W. Werner: "Einfluß langjähriger Stallmistdüngung auf chemische Parameter der Bodenfruchtbarkeit unter verschiedenen Klimabedingungen," *VDLUFA-Schriftenr.* **33** (1991) 177 – 182.
286. M. Bosch, A. Amberger: "Einfluß langjähriger Düngung mit verschiedenen N-Formen auf pH-Wert, Humusfraktionen, biologische Aktivität und Stickstoffdynamik einer Ackerbraunerde," *Z. Pflanzenernaehr. Bodenkd.* **146** (1983) 714 –724.
287. W. Kolb: "Nitrogenaseaktivität und Populationsdichte von luftstickstoffbindenden Bakterien in Abhängigkeit von Umweltfaktoren," Ph. D. Thesis, Universität Hoheneim 1984. H. Kuhlmann: "Ursachen und Ausmaß der N-, P-, K- und Mg-Ernährung der Pflanzen aus dem Unterboden," Habilitationsschrift, Universität Hannover 1988.
288. A. Klansink, G. Steffens: "Der Einfluß unterschiedlicher organischer Düngung auf N-Bedarf, Ertrag und Qualität von Ackerkulturen sowie auf den N- und Humusgehalt im Boden," *VDLUFA-Schriftenr.* **37** (1993) 269 – 272.
289. W. Heymann: "Humusgehalt verbessert," *Neue Landwirtschaft* **7** (1993) 40 – 42.
290. M. Körschens, *Archiv Acker Pflanzenbau Bodenkd.* **22** (1978) no. 3, 175.
291. T. Beck, *VDLUFA-Schrifenr.* **16** (1986) 31 – 46.
292. H. W. Schwerer, W. Werner, U. Ortseifen: "N-Haushalt und N-Nachlieferungsvermögen einer Lößparabraunerde nach langjähriger Strohdüngung in Kombination mit unterschiedlichen N-Düngerformen," *VDLUFA-Schriftenr.* **32** (1990) 199 – 203.
293. P. Foerster: "Einfluß hoher Güllegaben und üblicher Mineraldünger auf die Stoffbelastung (NO_3, NH_4, P und SO_4) im Boden und Grundwasser in Sandböden Nordwestdeutschlands," *Z. Acker Pflanzenbau* **137** /1973) 270 – 286.
294. P. Boysen: "Nährstoffauswaschungen aus gedüngten und ungedüngten Böden in Abhängigkeit von Standorteigenschaften und Nutzung der Moränen- und Sandergebiete Schleswig-Holsteins," Ph. D. Thesis, Universität Kiel, 1977.
295. I. Hess, H. Franken: "Über die Reduzierung von Nitratverlusten nach Leguminosen nach Anbau von Cruziferen," *Mitt. Dtsch. Bodenkdl. Gesellschaft* **57** (1988) 55 – 60.
296. J. Bischoff, K.-W. Krüger: "Einfluß von N-Düngermenge und Applikationszeitpunkt auf die Restmengen an Nitrat-N im Boden nach dem Anbau von Kartoffeln," *VDLUFA-Schriftenr.* **38** (1994) 125 – 128.
297. T. Appel, M. Schmücker, U. Schultheiss: "Möglichkeiten und Grenzen schlagbezogener N-Bilanzen zur Reduzierung der Nitratbelastung in Wasserschutzgebieten," *VDLUFA-Schriftenr.* **37** (1993) 137 – 140.
298. O. Günther-Borstel, M. Finck, H. Hanus: "N-Dynamik im Boden und N-Aufnahme von Winterweizen, Wintergerste und Winterraps unter besonderer Berücksichtigung von mineralischer Stickstoff- und Gülledüngung – Versuch einer Bilanzierung," *VDLUFA-Schriftenr.* **37** (1993) 125 –128.
299. K.-C. Kersebaum: "Die Simulation der Stickstoffdynamik von Ackerböden," Ph. D. Thesis, Universität Hannover 1989.
300. H. Hersemann: "Stickstoffsimulationsmodelle zur Unterstützung einer umweltgerechten Landbewirtschaftung und Düngebedarfsermittlung," *VDLUFA-Schriftenr.* **38** (1994) 223 – 226.
301. Council for Agricultural Science and Technology: "Effect of Increased Nitrogen Fixation on Stratospheric Ozone," *CAST-Report* **53** (1976).
302. W. Seiler, R. Conrad: "Mikrobielle Bindung von N_2O (Distickstoffoxid) aus Mineraldüngern – ein Umweltproblem," *Forum Mikrobiologie* **6** (1981) 322 – 328.
303. M. J. Eichner: "Nitrous Oxide Emissions from Fertilized Soils, a Summary for Available Data," *J. Environ. Qual.* **19** (1990) no. 2, 272 – 280.
304. H. Munk: "Zur vertikalen Wanderung mineralischer Phosphate bei starker Phosphatdüngung," *Landwirtsch. Forsch. Sonderh.* **27** (1972) 192 –199.
305. H. Kuntze: "Düngung und Gewässergüte," *Landwirtsch. Forsch. Sonderh.* **35** (1979) 14 – 24.
306. B. Seggewiss: "Magnesiumdynamik im wurzelnahen Boden und Magnesiumverfügbarkeit unter dem Einfluß von Kalium," Ph. D. Thesis, Universität Göttingen 1986.
307. P. Foerster, H. Neumann: "Stoffbelastung kleiner Fließgewässer in landwirtschaftlich genutzten Gebieten Norddeutschlands," *Mitt. Niedersächs. Wasseruntersuchungsamt* 7, Hildesheim 1981.
308. E. Rübesam: "Die Wirkung der Stallmistdüngung dargestellt an Dauerversuchen in Thyrow und Müncheberg," *Tagungs-Ber. Dtsch. Akad. Landwirtschafts. Wiss.*, 1962, 23 – 37.
309. F. Asmus: "Einfluß organischer Dünger auf Ertrag, Humusgehalt des Bodens und Humusreproduktion," *Landwirtsch.* **4** (1993) 127 – 139.
310. H. Sturm, A. Buchner, W. Zerulla: *Gezielter Düngen*, DLG-Verlag, Frankfurt/Main 1994.
311. F. Asmus: "Zeit, Menge und Art bestimmen den Erfolg," *Neue Landwirtschaft* **4** (1994) 47 – 51.
312. J. Jungk, K. Isermann, G. Henjes: "Einfluß von cadmiumhaltigen Düngerphosphaten auf die Cadmiumanreicherung von Kulturböden und Nutzpflanzen," *Landwirtsch. Forsch. Sonderh.* **32** (1979) 262.
313. A. Finck: "Phosphat-Düngung – Aktuelle Fragen aus agronomischen und ökologischen Perspektiven," *VDLUFA-Schriftenr.* **37** (1993) 133 –136.
314. H.-P. Blume: "Pflanzenschutzmittel (Pestizide)," in H.-P. Blume (ed.): *Handbuch des Bodenschutzes,* ecomed, Landsberg 1990, pp. 311 – 340.
315. C. M. Himel, H. Loats, G. W. Bailey: "Pesticide Sources to the Soil and Principles of Spray Physics," in H. H. Cheng (ed.): *Pesticides in the Soil Environment: Processes, Impacts, and Modeling,* Soil Science Society of America, Madison, Wis., 1990, pp. 7 – 50.
316. K.-H. Domsch: *Pestizide im Boden, mikrobieller Abbau und Nebenwirkungen auf Mikroorganismen*, VCH Verlagsgesellschaft, Weinheim, Germany 1992.
317. H. H. Cheng: "Pesticides in the Soil Environment – An Overview," in H. H. Cheng (ed.): *Pesticides in the Soil Environment: Processes, Impacts, and Modeling,* Soil Science Society of America, Madison, Wis., 1990, pp. 1 – 5.
318. J.-M. Bollag, S.-Y. Liu: "Biological Transformation Processes of Pesticides," in H. H. Cheng (ed.): *Pesticides in the Soil Environment: Processes, Impacts, and Modeling,* Soil Science Society of America, Madison, Wis., 1990, pp. 169 – 212.
319. B. L. Sawhney, K. Brown (eds.): *Reactions and Movement of Organic Chemicals in Soils*, Soil Science Society of America, American Society of Agronomy, Madison, Wis. 1989.
320. R. W. Wallnöfer, G. Engelhardt: "Microbial Degradation of Pesticides," *Chem. Plant Prod.* **2** (1989) 1 – 115.
321. J. M. Dougherty, G. R. Lanza: "Anaerobic Subsurface Soil Microcosms: Methods to Monitor Effects of Organic Pollutants on Indigenous Microbial Activity," *Toxic. Assess.* **4** (1989) 85 – 104.
322. R. J. Hance: "Herbicide Behaviour in the Soil, with Particular Reference to the Potential for Ground Water Contamination," *Prog. Pestic. Biochem. Toxicol.* **6** (1987) 223 – 247.
323. G. T. Hickmann, J. T. Novak: "Relationship Between Subsurface Biodegradation Rates and Microbial Density," *Environ. Sci. Technol.* **23** (1989) 525 – 532.
324. K. Alef: "Ökotoxikologische Verfahren," in K. Alef (ed.): *Biologische Bodensanierung,* VCH Verlagsgemeinschaft, Weinheim 1994, pp. 171 – 186.
325. K. Alef: "Anaerobic Microbial Activities in Soil," in K. Alef, P. Nannipieri (eds.): *Methods in Applied Soil Microbiology and Biochemistry,* Academic Press, London 1995, pp. 271 – 310.
326. Chemikaliengesetz (1982) Gesetz zum Schutz vor gefährlichen Stoffen, BGBl. I, novelliert 14. 03. 1990.
327. Pflanzenschutzgesetz – PflSchG (Gesetz zum Schutz der Kulturpflanzen), Sep. 15, 1986 (BGBl. I, S. 1505), in C. H. Beck (ed.): *Umwelt-Recht,* Verlag C. H. Beck, München 1987, Chap. 9.2.
328. P. Rudolph, R. Boje: *Ökotoxikologie. Grundlagen für die ökotoxikologische Bewertung von Umweltchemikalien nach dem Chemikaliengesetz*, ecomed, Landsberg 1986.
329. OECD-Guidelines for Testing of Chemicals, Paris 1984.
330. C. A. M. Van Gestel, N. M. Van Straalen: "Ecotoxicological Test System for Terrestrial Invertebrates," in M. H. Donker, H. Eijsackers, F. Heimbach (eds): *Ecotoxicology of Soil Organisms,* Lewis Publishers, Boca Raton 1994.
331. K. Alef: *Methodenhandbuch Bodenmikrobiologie*, ecomed, Landsberg 1991, pp. 135 – 150.
332. K. Alef, J. C. Forster, A. Hauck, G. Bergmann: "Arginin-Mineralisation zur Beurteilung der Wirkung von Chemikalien in Umweltproben," in K. Alef, H. Fiedler, O. Hutzinger (eds.): *Umweltmonitoring und Bioindikation,* Ecoinforma Press, Universität Bayreuth, Bayreuth 1994, pp. 325 – 334.
333. J. R. Corbett: *The Biochemical Mode of Action of Pesticides*, Academic Press, London 1974.
334. L. Somerville, M. P. Greaves: *Pesticide Effects on Soil Microflora*, Taylor & Francis, London 1987.

335. K. E. Lee: *Earthworms*, Academic Press, Sydney 1985.
336. G. Simon-Sylvester, J. C. Fourner: "Effects of Pesticides on the Soil Microflora," *Adv. Agron.* **31** (1979) 1 – 92.
337. B. Weischer, J. Müller: "Nebenwirkungen von Pflanzenschutzmitteln auf Nematoden und ihre Antagonisten," in: *Pflanzenschutzmittel und Boden,* Paul Parey, Berlin 1985, pp. 159 – 175.
338. B. Hock E. F. Elstner: *Pflanzentoxikologie*, Wissenschaftverlag, Mannheim 1984.
339. E. Grigo: "Pflanzenschutz," in J. Oehmichen (ed.): *Pflanzenproduktion, 2. Produktionstechnik,* Paul Parey, Hamburg 1986, pp. 131 – 219.
340. H. W. Berg, G. F. Diehl, H. Frank: *Rückstände und Verunreinigungen in Lebensmitteln*, D. Steinkopff Verlag, Darmstadt 1978.
341. R. Schleyer, H. Kerndorff: *Die Grundwasserqualität westdeutscher Trinkwasserressourcen*, VCH Verlagsgesellschaft, Weinheim, Germany 1992.
342. R. H. Plumb, Jr.: "The Importance of Volatile Organic Compounds as a Disposal Site Monitoring Parameter," in S. Lesage, R. E. Jackson (eds.): *Groundwater Contamination and Analysis at Hazardous Waste Sites,* Marcel Dekker, New York 1992.
343. R. H. Plumb, Jr.: "Disposal Site Monitoring Data: Observation and Strategy Implications," in B. Hitchon, M. Trudell (eds.): "Hazardous Wastes in Groundwater: A Soluble Dilemma," *Proceedings of the Second Canadian/American Conference on Hydrology,* Banff, AB, Canada 1985, pp. 66 – 67.
344. R. H. Plumb, Jr., A. M. Pitchford: "Volatile Organic Scans: Implications for Groundwater Monitoring," in: *Proceedings of the National Water Well Association/American Petroleum Institute Conference on Petroleum Hydrocarbons and Organic Chemicals in Groundwater,* Housten, TX, Nov. 13 – 15, 1985, pp. 1 – 15.
345. H. Kerndorff, R. Schleyer, H. H. Dieter: "Bewertung der Grundwassergefährdung von Altablagerungen – Standardisierte Methoden und Maß-stäbe," *WaBoLu-Hefte 1/1993*.
346. H. Kerndorff *et al.*: "Erfassung grundwassergefährdender Altablagerungen – Ergebnisse hydrogeochemischer Untersuchungen," *WaBoLu-Hefte* **5/** 1985.
347. *DIN: Handbuch der Bodenuntersuchung: Terminologie*, Verfahrensvorschriften und Datenblätter; physikalische, chemische und biologische Untersuchungsverfahren; gesetzliche Regelwerke, Wiley-VCH, Weinheim und Beuth Verlag, Berlin 2000.
348. DIN ISO 10 381–1/4: Soil Quality - Sampling, Part 1: Guidance on the Design of Sampling Programmes 1994; Part 4: Guidance on the Investigation of Natural, Near Natural, and Cultivated Sites, 1994.
349. N. Nothbaum, R. W. Scholz, Th. W. May: *Probenplanung und Datenanalyse bei kontaminierten Böden*, Erich Schmidt Verlag, Berlin 1994.
350. DIN ISO 10 381–2: Soil Quality – Sampling, Part 2: Guidance on Sampling Techniques, 1994.
351. DIN ISO 10 381–3: Soil Quality – Sampling, Part 3: Guidance on Safety, 1994.
352. DIN ISO 11 464: Soil Quality – Pretreatment of Samples for Physico-chemical Analysis, 1994.
353. P. Tschöpel: "Fehlerquellen beim Aufschluß," in M. Stoeppler (ed.): *Probenahme und Aufschluß,* Springer Labormanual, Springer Verlag, Berlin 1994, pp. 111 – 120
354. O. Fränzle: *Methodik für Schadstoffuntersuchungen in Böden (draft)*, Geograph. Institut der Universität Kiel, 1994.
355. E. Jackwerth, M. Würfels: "Der Druckaufschluß – apparative Möglichkeiten, Probleme und Anwendungen," in M. Stoeppler (ed.): *Probenahme und Aufschluß,* Springer Labormanual, Springer Verlag, Berlin 1994, pp. 121 – 138
356. L. Dunemann: "Mikrowellenaufschluß," in M. Stoeppler (ed.): *Probenahme und Aufschluß,* Springer Labormanual, Springer Verlag, Berlin 1994, pp. 139 – 150
357. G. Knoop, B. Pehl, D. Schlösser: "Vergleichende Untersuchungen zum Feststoffaufschluß mit der Mikrowellentechnik," *Korresp. Abwasser* **41** (1994) 1836 – 1839.
358. L. Dunemann: "Aufschlußmethoden für die Schwermetallanalytik (Marktübersicht)," *Nachr. Chem. Tech. Lab.* **39** (1991) no. 10, M 2 – M 16.
359. H. Hein: "Aktuelle chromatographische Umweltanalytik," *CLB Chem. Labor Betr.* **43** (1992) 7 – 15, 65 – 75.
360. W. Engewald: "Gaschromatographie (Marktübersicht)," *Nachr. Chem. Tech. Lab.* **42** (1994) no. 4, 356 – 366; **43** (1995) no. 3, M 1 – M 6.
361. G. Schwedt: *Analytische Chemie (Grundlagen*, Methoden, Praxis), Thieme Verlag, Stuttgart 1995.
362. K. Unger: "HPLC-Quo vadis?" *Labor 2000*, 1994, 90 – 93.
363. W. Funk, V. Dammann, G. Donnevert: *Qualitätssicherung in der Analytischen Chemie*, VCH Verlagsgesellschaft, Weinheim, Germany 1992.
364. VDLUFA: "Aufschluß mit Königswasser," in: *Umweltanalytik,* Methodenbuch **vol. VII**, Kennziffer 2.1.2 (draft), VDLUFA-Verlag, Darmstadt 1995.
365. VDLUFA: "Bestimmung von Thallium," in: *Umweltanalytik,* Methodenbuch **vol. VII**, Kennziffer 2.2.3, VDLUFA-Verlag, Darmstadt 1996.
366. DIN 19730: Bodenbeschaffenheit – Extraktion von Spurenelementen mit Ammoniumnitratlösung, 1997.
367. P. Dominik, A. Paetz: *Methodenbuch Bodenschutz I*, Forschungsbericht UFOPLAN no. 10705001/07, Umweltbundesamt, Berlin 1994.
368. W. Jennings: *Analytical Gas Chromatography,* Academic Press, New York 1987.
369. K. J. Hyver, P. Sandra: *High Resolution Gas Chromatography*, Hewlett-Packard, Palo Alto 1989.
370. H. Engelhardt: *Practice of High Performance Liquid Chromatography*, Springer Verlag, Heidelberg 1986.
371. J. W. Dolan, L. R. Snyder: *Troubleshooting LC Systems*, Humana Press, London 1989.
372. ISO Standard Compendium: Environment – Soil Quality – General Aspects, Chemical and Physical Methods of Analysis: Biological Methods of Analysis, 1st ed., 1994.
373. Merkblätter no. 1, Bestimmung von polyzyklischen aromatischen Kohlenwasserstoffen in Bodenproben, Landesumweltamt Nordrhein-Westfalen, Essen 1994.
374. J. Bortlisz, L. Velten: "Herstellung und Bestimmung der Trockenmasse von Klärschlämmen, Sedimenten und Böden mit Hilfe der Gefriertrocknung," *Abwassertech. Abfalltech. Recycl.* 1991, no. 4, 35 – 37.
375. I. Blankenhorn, D. Meijer, R. J. van Delft: "Interlaboratory Comparison of Methods Used for Analysing Polycyclic Aromatic Hydrocarbons (PAHs) in Soil Samples," *Fresenius J. Anal. Chem.* **343** (1992) 497 – 504. Links
376. C. Herziger-Möhlmann, A. Manns: "Erkundung und Bewertung von Bodenkontaminationen mit immunochemischen Meßverfahren," *Draegerheft* **357** (1994) 2 – 6.
377. S. B. Hawthorne: "Analytical Scale Supercritical Fluid Extraction," *Anal. Chem.* **62** (1990) no. 11, 633.
378. F. Höfler: "Überkritische Fluidextraktion, Teil I," *LaborPrax* is 16 (1992) 350.
379. F. Höfler: "Überkritische Fluidextraktion," *LaborPraxis* **16** (1992) 506.
380. E.-G. Zeschmann: *Beurteilung und Sanierung ölverunreinigter Standorte*, Expert Verlag, Böblingen 1987.
381. EPA Method 9071 A/8440, Total Recoverable Petroleum Hydrocarbons by Infrared Spektrophotometry, Revision Draft, 1992.
382. ISO/TR 11 046, Determination of Mineral Oil Content-Method by Infrared Spectrometry and Gas Chromatography Method, 1994.
383. H. Fischer *et al.*: "Bestimmung von Mineralöl-Kohlenwasserstoffen im Boden," *UWSF-Z Umweltchem. Ökotox.* **6** (1994) no. 4, 189 – 195.
384. EPA Method SW 846/3560, Supercritical Fluid Extraction, Revision Draft 1, 1992.
385. F. Höfler, J. Ezzell, B. Richter: "Beschleunigte Lösemittelextraktion (ASE)," *LaborPraxis* **19** (1995) nos. 3, 4.
386. EPA M 3545, Proposal, Accelerated Solvent Extraction, 1994.
387. M. Stoeppler: *Probenahme und Aufschluß*, Springer Verlag, Heidelberg 1994.
388. S. Michels, T. Eikmann, *VDI Ber.* **837** (1990) 509.
389. H. J. Einbrodt *et al.*, *Wiss. Umwelt* **2** (1985) 149 –157.
390. T. Eikmann in V. Franzius, R. Stegmann, K. Wolf (eds.): *Handbuch Altlastensanierung*, Section 1.3.2.2, R. v. Decker's Verlag, Heidelberg 1988, p. 1.
391. T. Eikmann, S. Michels: "Wirkung von Luftverunreinigungen auf Böden," *VDI-Ber.* **837** (1990) 489.
392. H. Heil *et al.*, *Vom Wasser* **72** (1989) 321 – 348.
393. T. Eikmann, A. Kloke in D. Rosenkranz, G. Einsele, H.-M. Harreß (eds.): *Handbuch Bodenschutz 3590*, 2nd ed., Erich Schmidt Verlag, Berlin 1993.
394. T. Eikmann, *Altlasten-Spektrum* **1** (1993) 31 – 38.
395. B. Hudec *et al.*, *Altlasten-Spektrum* **3** (1993) 164 – 168.
396. S. Eikmann, U. Lieser, T. Eikmann, *Altlasten-Spektrum* **2** (1993) 75 – 84.
397. P. Doetsch, P. Dreschmann: "Verfahrensdokumente wichtiger Dekontaminationsverfahren als Leitfaden für die Sanierungsuntersuchung bzw. -planung bei Altlasten," in V. Franzius, R. Stegmann, K. Wolf, E. Brandt (eds.): *Handbuch der Altlasten-Sanierung,* R. V. Decker's Verlag, G. Schenck, Heidelberg, 1993.
398. Landesanstalt für Umweltschutz Baden-Württembergw (ed.): "Stationäre Bodenreinigungsanlagen in der Bundesrepublik Deutschland," *Texte und Berichte zur Altlastenbearbeitung* **31** (1997).
399. "ITVA-Preisspiegel," *altlasten spektrum* **5** (1998) 310 ff.
400. Ingenieurtechnischer Verband Altlasten, Arbeitshilfe H - 1 - 6: "Thermische Verfahren der Bodendekontamination", May 1996.
401. Ingenieurtechnischer Verband Altlasten, Arbeitshilfe H - 1 - 1: "Dekontamination durch Bodenwaschverfahren", Nov. 1994.
402. Ingenieurtechnischer Verband Altlasten, Arbeitshilfe H - 1 - 3: "Mikrobiologische Verfahren zur Bodendekontamination", Dec. 1994.
403. NATO, Committee on the Challenges of Modern Society: The Environmental Aspects of Reusing Former Military Lands, Technical Report, Phase I, **Vol. 2**: Annexes, Dec. 1996.
404. G. Schmitz: "Rekultivierung," in *Taschenwörterbuch der Umweltplanung,* vol. 1614, F. Malz, München 1974, p. 463.
405. I. Stratmann: "Ertragskundliche Untersuchungen auf Rekultivierungsflächen im rheinischen Braunkohlegebiet," *Braunkohle (Düsseldorf)* **37** (1985) 484 – 491.
406. E. Friedrich: "Forststandörtliche Eignung des ' Forstkieses' aus Löß und Rheinschotter im Hinblick auf Lagerungsdichte, Mischungsverhältnis, Nährstoffe," Untersuchung im Auftrag der Rheinbraun AG, Köln 1985 (unpublished).
407. K. Winter: "Bodenmechanische und technische Einflüsse auf die Qualität von Neulandflächen," *Braunkohle (Düsseldorf)* **42** (1990) no. 10, 15 – 23.
408. R. Starke: "Wiedernutzbarmachung im rheinischen Braunkohlerevier," *Braunkohle (Düsseldorf)* **42** (1990) no. 10, 4 – 10.
409. G. Dumbeck: "Bodenkundliche Aspekte bei der landwirtschaftlichen Rekultivierung im rheinischen Braunkohlerevier," *Braunkohle (Düsseldorf)* **44** (1992) no. 9, 8 – 11.
410. J. D. Majer (ed.): *Animals in Primary Succession*, Cambridge University Press, Cambridge 1989, p. 547.
411. W. Dunger, *Pedobiologia* **9** (1969) 146 – 151.
412. S. P. Rushton, *Pedobiologia* **29** (1986) 85 – 90.

413. U. D. Dworschak, *Eur. J. Soil Biol.* **33** (1997) 75 – 81.
414. M. B. Bouché: “The Establishment of Earthworm Communities”, in J. D. Satchell (ed.): *Earthworm Ecology*, Chapman & Hall, London 1983, pp. 431 – 448.
415. E. Westernacher-Dotzler, G. Dumbeck, *J. Agronomy Crop Sci.* **169** (1992) 298 – 309.
416. R. Schneider: *Gefügeentwicklung in Neulandböden aus Löß und Hafenschlick und deren Auswirkungen auf bodenphysikalische und -mechanische Parameter*, Verlag Shaker, Aachen 1994, p. 236.
417. M. Kloepfer: *Umweltrecht*, Verlag C. H. Beck, 2nd ed., München 1998, p. 761 f.
418. P.-C. Storm: “Bodenschutzrecht,” in O. Kimminich, H. v. Lersner, P.-C. Storm (eds.): *Handwörterbuch des Umweltrechts,* 2nd ed., vol. I, Erich Schmidt Verlag, Berlin 1994. W. Burhenne: *Umweltrecht*, **loose-leaf collection, 7 vols.**, Erich Schmidt Verlag, Berlin.
419. M. Kloepfer: *Umweltschutz*, Verlag C. H. Beck, München 1981.
420. P.-C. Storm: *Umweltrecht, Wichtige Gesetze und Verordnungen zum Schutz der Umwelt*, Textausgabe 13th ed., Deutscher Taschenbuch Verlag, München 2000.
421. Umweltprogramm der Bundesregierung (Environmental Program of the Federal Government), 1971 (Bundestags-Drucksache 6/2710).
422. Bundesinnenministerium (eds.): *Bodenschutzkonzeption der Bundesregierung (Soil Protection Concept of the Federal Government)* (Bundestags-Drucksache 10/2977), Kohlhammer Verlag, Stuttgart 1985. Bericht der Bundesregierung, Maßnahmen zum Bodenschutz, 1988 (Bundestags-Drucksache 11/1625).
423. H. Ebersbach: *Rechtliche Aspekte des Landverbrauchs am ökologisch falschen Platz*, Erich Schmidt Verlag, Berlin 1985 (UBA-Berichte 1/85).
424. D. Rosenkranz, G. Einsele, H.-M. Haress (eds.): *Bodenschutz*, Ergänzbares Handbuch, Erich Schmidt Verlag, Berlin 1988, no. 8902.
425. P.-C. Storm: “Umweltrechtsprogramm von Montevideo,” in O. Kimminich, H. v. Lersner, P.-C. Storm (eds.): *Handwörterbuch des Umweltrechts,* 2nd ed., **vol. II**, Erich Schmidt Verlag, Berlin 1994.
426. Bundesumweltministerium (eds.): *Bericht der Bundesregierung über die Konferenz der Vereinten Nationen für Umwelt und Entwicklung Rio de Janeiro*, 1992, Bonn 1992, pp. 30.
427. Bodenschutzgesetz (BodSchG, Soil Protection Act) § 3, Baden-Württemberg, 24th June 1991 (GBl. S. 434).
428. in [422] p. 21.
429. Gesetz zum Schutz vor schädlichen Bodenveränderungen und zur Sanierung von Altlasten (Bundes-Bodenschutzgesetz BBodSchG, Federal Soil Protection Act), 17th March 1998 (Bundesgesetzblatt Part I p. 502); Bundes-Bodenschutz- und Altlastenverordnung (BBodSchV, Federal Ordinance on Soil Protection and Contaminated Sites), 12th July 1999 (Bundesgesetzblatt Part I, p. 1554, in [420], Secs. 3.4, 3.4.1.
430. *Environmental Code (Umweltgesetzbuch – UGB)*, Draft, Verlag Duncker & Humblot, Berlin 1998, "Chap. Soil Conservation, p. 203".

[Top of Page]

About Wiley InterScience | About Wiley | Privacy | Terms & Conditions
Copyright © 1999-2007John Wiley & Sons, Inc. All Rights Reserved.

Wastewater

Walter Simmler, Bayer AG, Leverkusen, Federal Republic of Germany (Chaps. 1, 4)

Theo Mann, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 2,2.1, 2.1.2, 2.1.3, 2.1.4, 5.1)

Michael Berger, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.1.1)

Günther Kern, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 2.1.4.2 in part)

Joachim Lemke, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.1.4.2 in part)

Dieter Klockner, Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Section 2.1.4.2 in part)

Elmhardt Neuber, Hüls Aktiengesellschaft, Marl, Federal Republic of Germany (Section 2.1.4.2 in part)

Gerhard-Wilhelm Hebbel, Merck AG, Darmstadt, Federal Republic of Germany (Section 2.1.4.1, 2.1.4.2 in part)

Ulrich Werthmann, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.1.4.2 in part)

Günter Klinsmann, Knoll AG, Minden, Federal Republic of Germany (Section 2.1.4.2 in part)

John F. Lawson, Miles Inc. now Bayer Corporation, Kansas City, Missouri, United States (Section 2.1.4.2 and 2.1.4.6 in part)

Hans Georg Meyer, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.1.4.3)

Manfred Müller, DEA Mineralöl AG, Wesseling, Federal Republic of Germany (Section 2.1.4.3)

Klaus Balser, Wolff Walsrode AG, Walsrode, Federal Republic of Germany (Section 2.1.4.3)

Walter Maier, Hoffmann-La Roche AG, Grenzach-Wyhlen, Switzerland (Section 2.1.4.4)

Joachim Frieser, Hoechst Aktiengesellschaft, Werk Kalle-Albert, Wiesbaden, Federal Republic of Germany (Section 2.1.4.4)

Markus Thüer, Ciba-Geigy AG, Basel, Switzerland (Section 2.1.4.5)

Jürgen Malaszkiewicz, Henkel KGaA, Düsseldorf, Federal Republic of Germany (Section 2.1.4.6 in part)

Alexander Aivasidis, Forschungszentrum Jülich, Jülich, Federal Republic of Germany (Section 2.2)

Bernd Koglin, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 3)

Hartmut von Kienle, Chemviron Carbon GmbH, Neu-Isenburg, Federal Republic of Germany (Section 5.1)

Uwe Wegmann, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 5.2)

Walter Weisbrodt, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 5.2)

Wolfgang Moldenhauer, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 6)

Günter Mischer, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 6)

1. Introduction 4
2. Biological Wastewater Treatment 6
2.1. Aerobic Treatment 8
2.1.1. Biological Principles 10
2.1.1.1. Organisms 11
2.1.1.2. Biocenosis 17
2.1.1.3. Metabolism 19
2.1.2. Basic Biotechnological Considerations 32
2.1.3. Technology and Operation of a Wastewater-Treatment Plant 41
2.1.4. Examples of Industrial Wastewater-Treatment Plants Based on a Variety of Centralized Processes 45
2.1.4.1. Single-Stage and Multistage Activated-Sludge Plants 47
2.1.4.2. Oxygen-Based Facilities 57

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_001

2.1.4.3. Treatment Facilities with Selective Carbon and Nitrogen Elimination . 62
2.1.4.4. Trickling-Filter Facilities 68
2.1.4.5. Biofilter Systems 71
2.1.4.6. Indirect-Discharge Facilities 72
2.2. Anaerobic Treatment 77
2.2.1. Introduction 77
2.2.2. Basic Principles 77
2.2.3. Thermodynamics 80
2.2.4. Kinetics 81
2.2.5. Mass and Heat Transfer 86
2.2.6. Reactor Design 87
2.2.7. Biomass Retention/Recycling 88
2.2.7.1. The Anaerobic Contact Process . . 90
2.2.7.2. Upflow Anaerobic Sludge Blanket (UASB) Reactors 90
2.2.7.3. Fixed-Bed Reactors (Anaerobic Filters) 92
2.2.7.4. Fluidized-Bed Reactors 93
2.2.7.5. Fixed-Bed Loop Reactors 95
2.2.7.6. Engineering Considerations 96
2.2.8. Control 100
2.2.9. Future Prospects 101
3. Aerobic Processing Steps in Process Combinations 102
3.1. Flotation 103
3.1.1. Basic Principle 103
3.1.2. Contact and Adhesion of Particles to Gas Bubbles;Flotation Kinetics . . . 105
3.1.3. Implementation, Fields of Application, and Operational Experience . . 108
3.1.4. Costs and Benefits 110
3.2. Biofilters and Bioscrubbers 110
3.2.1. Basic Principle 110
3.2.2. Prerequisites for Biological Waste-Gas Purification 111
3.2.3. Implementation, Fields of Application, and Operational Experience . . 111
3.2.4. Costs and Benefits 113
4. Chemical Wastewater Treatment 114
4.1. Oxidative Degradation by Hydrogen Peroxide 115
4.2. Oxidative Degradation by Ozone 117
4.3. Photocatalysis in Oxidative Degradation by Air 118
4.4. Wet Oxidation by Oxygen Sources in the Aqueous Phase 119
4.4.1. Low-Pressure Wet Oxidation 120
4.4.2. Oxidation by Nitric Acid 127
4.4.3. High-Pressure Wet Oxidation 127
4.5. Hydrogenation 129
5. Physical Wastewater Treatment . 130
5.1. Adsorptive Purification 130
5.1.1. Adsorption 130
5.1.2. Adsorption Isotherms 130
5.1.3. Adsorption Kinetics 132
5.1.4. Adsorbents 133
5.1.5. Adsorption Processes 133
5.1.6. Adsorbent Recovery and Regeneration 134
5.1.6.1. Regeneration of Carbon Adsorbents 135
5.1.6.2. Regeneration of Aluminum Oxide . 136
5.1.6.3. Regeneration of Adsorber Resins . 136
5.1.7. Applications of the Adsorptive Purification of Wastewater 136
5.1.7.1. Purification with Activated Carbon 136
5.1.7.2. Purification with Lignite Coke . . . 137
5.1.7.3. Purification with Aluminum Oxide 137
5.1.7.4. Treatment with Adsorber Resins . . 138
5.1.7.5. Purification with Isopolyacids . . . 138
5.2. Mechanical Treatment 138
5.2.1. Screening Processes 138
5.2.2. Gravity Processes 142
5.2.2.1. Sedimentation 142
5.2.2.2. Floatation 144
5.2.2.3. Flotation 145
5.2.2.4. Foaming-Out Process 146
5.2.2.5. Centrifugal Processes 146
6. Wastewater Incineration 146
6.1. Introduction 146
6.2. Legal Aspects 147
6.3. Distinctions with Respect to Other Processes, and Evaluation Criteria 147
6.3.1. Calorific Value 148
6.3.2. Wastewater Constituents 149
6.4. Incineration Plants 149
6.4.1. Combustion Chambers 149
6.4.2. Immersion Heating Chambers . . . 151
6.4.3. Fluidized-Bed Furnaces 151
6.4.4. Rotary Kilns 152
6.5. Emissions: Off-Gases, Water, and Solid Waste 152
7. References 153

The entire topic was coordinated by Walter Simmler

1. Introduction

As mankind eventually adopted a more settled, non-nomadic way of life, people became increasingly involved in the technical aspects of

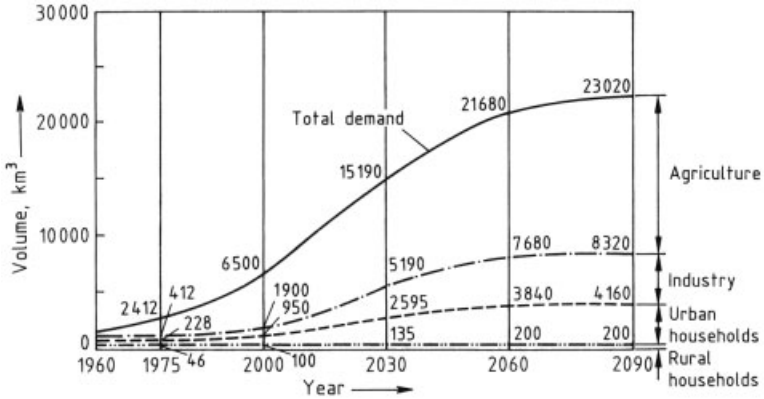

Figure 1. The increasing worldwide demand for water, projected into the future

water. To a large extent, the primary concerns in the beginning were utilization and improvement of existing water resources, together with protection against the hazards and potential harm associated with uncontrolled natural water. It was only toward the end of the nineteenth century that wastewater became an issue in science, technology, and legislation: specifically, its production and treatment, in terms of both municipal and industrial sources [1], [2].

As early as 4500 years ago the first prerequisites were met for urban and agricultural water management. This occurred in the Indus Valley, and it encompassed irrigation and drainage systems, canals, and sewage facilities. Even so, the treatment of wastewater in formal waste-treatment plants by means of the microbial degradation of wastewater components was reported for the first time in 1892 in London, followed in 1895 by similar efforts in Frankfurt/Main. In both cases it was municipal wastewater, including that from artisans, craftsmen, and small factories, typical of larger cities of that time. Already in the Middle Ages communal wastewater from Milan was being used directly to irrigate the meadows along the outskirts of town. Much earlier, in Greco – Roman times, extensive facilities were erected and maintained for supplying drinking water to cities. Wastewater in those days presented no major problems, and sewage systems were regarded less as a means of collecting water for reuse than as a way of draining off potential sources of hazard and preventing pollution of the streets, with the attendant risk of a spread of vermin and epidemics.

Development of the organized utilization of water as an essential resource for human beings, animals, and plants led to further technical strides, such as dams against flooding or for storage purposes, waterways designed for transport, and harbors on the seacoast and along inland waterways. Problems of wastewater arose gradually during the same period in conjunction with the increase in urban population as the natural self-purification capacity of surface waters proved no longer able to keep pace with development. Risks related to groundwater contamination are associated not only with emissions in the form of wastewater; the ground – air cycle plays a role as well via atmospheric deposition [3]. The state of a particular body of water can be described by a set of code numbers, but the core of the problem is in fact a sum of all the processes leading to the observed state. In this case, the essential element is the kinetics of two opposing processes: the rate of pollution, and the rate of cleansing. Each of these is in turn a combination of natural and anthropogenic phenomena applicable to the site in question. Most of the problems are derived not from absolute numbers, but rather from population densities, production densities, productivities, etc., in the various urban centers of the industrialized world, all of which actually have access to an adequate supply of natural water.

Utilization of water power generated by falls, tides, or waves, and then transformed into the power of mills, pumping plants, and turbines,

has been responsible for very little of the wastewater problem, which in the early days of industrialization was due largely to hammer and stamping mills in the metalworking industry and various ore-dressing facilities. Only when the use of water led to its exploitation as a production medium the greatest specific wastewater load was reached. This was especially the case with process wastewater derived from tanneries and dye works. To this day, street names associated with these activities can be found along inner-city watercourses in the old city centers of such medieval towns as Nuremberg and Strasbourg.

High densities of population and industry – as in the Ruhr area of Germany during the nineteenth century – necessitated the development of a complex water economy. Thus, the Ruhr – Emscher water system provided simultaneously for drinking water, industrial water, and inland waterways, and it even included a wastewater disposition system fully consistent with the knowledge and technological potential available at that time.

In the terminology of water economics, *consumption* of water refers to a loss of quantity, not a decrease in quality. In this sense, consumption represents that part of the water supply that is lost in the course of use, primarily through evaporation. This fraction of the water is permanently withdrawn – at least from the local water cycles – and is thus no longer available for further utilization, so it must be replenished with water from precipitation, springs, or wells [2], [3].

As industrialization proceeded, two unique characteristics of water acquired rapidly increasing importance: its high specific heat capacity, and its rather high solvent power with respect to many inorganic and some organic substances. The consequences of these factors in the context of the production and disposal of wastewater are quite different, however. Water that is intended to serve as a heat reservoir, cooling agent, steam source, etc., must be cleaned before use in order to prevent corrosion and erosion in turbines and heat exchangers, and it is subsequently returned to the environment in a purified state, albeit at a higher temperature. On the other hand, water in its function as a reaction medium, or even a reaction partner, has now developed into the most significant source of wastewater in industry. Entire branches of manufacturing are based on production processes carried out in the aqueous phase, where water is used as a solvent, dispersing agent, transport medium, and reagent. This is perhaps most evident in the case of breweries, in sugar, paper, and pulp factories, in dye manufacture by chemical concerns, and in dye works, tanneries, and the like where the real problem is one not only of the actual content of the wastewater, but also its quantity [1], [4].

For the year 2000, a global balance of quantities and fluxes shows an overall water supply of 25 000 km^3 and a water demand of 6000 km^3, representing 24 % of the directly usable supply [1]. When this is allocated among the major consumers and account is taken of the corresponding levels of specific water consumption (as defined above; i.e., evaporation), several trends with regard to quantities and types of waste water are discernible:

Consumption category	Percentage of total demand (6000 km^3)	Consumption (evaporation) as a percentage of demand
Domestic	8	20 – 30
Industry	29	15 – 20
Agriculture	59	75
Storage losses	4	100

Water circulation through the atmosphere occurs relatively rapidly. The atmosphere contains 13 000 km^3 (ca. 0.02 %) of the overall (liquid) global water reserve of 1.3×10^9 km^3. The annual quantity subject to evaporation (which is equal to the annual amount of precipitation) is estimated at 475 000 km^3. This corresponds to a 35-fold annual turnover of the atmospheric content, which means water exchange between the atmosphere and the surface of the earth is complete every 9.5 days. In some cases local circumstances lead to considerable deviations from these global values.

Wastewater treatment becomes especially important in times of water scarcity. This is particularly true when agriculture, domestic water needs, and industry find themselves in vigorous competition. The greatest increase by far (Fig. 1) is anticipated for agriculture [1], assuming for the year 2000 a world population of $6 - 6.5 \times 10^9$.

Water is regarded as the chief raw-material problem of the future, elevating wastewater treatment to the status of a recycling technology. This involves "substance protection" in the sense of sustainable development (→ Sustainability

and Industrial Chemistry – Ethical Aspects and Approaches), together with the maintenance of an adequate supply of drinking water, an emphasis that goes beyond the earlier concern directed almost exclusively toward environmental protection, especially the protection of natural waters.

Each of the processes described below has its place in the broad spectrum of technical possibilities. The question of what is the "best process" should thus be replaced by a search for the most suitable process in a particular circumstance, taking fully into account the nature of certain definable problem cases. Modern technological development of wastewater treatment has occurred largely in and with the aid of the chemical industry. However, the wastewater problem and its treatment is of interest not only to this particular aspect of industrialized society. Numerous other branches are affected as well, especially municipalities, often to an even greater extent.

There is no shortage today of diverse international experience in the construction and operation of waste-treatment plants. In fact, in some places this has developed into its own separate branch of process engineering. It is still worth noting with respect to terminology, however, that some of the reasoning applied to wastewater concepts and standards of evaluation has been borrowed from neighboring disciplines, especially limnology and biology.

2. Biological Wastewater Treatment

Large waste-treatment plants intended for the purification of industrial or municipal wastewater utilize a combination of processing techniques:

1) *Chemical treatment* – insofar as necessary – for the neutralization and precipitation of wastewater components
2) *Mechanical treatment* for the elimination of undissolved substances, including screening, clarification, and filtration
3) *Biological treatment* for the removal of dissolved substances, in the course of which colloidally dispersed materials are also destabilized and flocculated; substances can also be eliminated by adsorption onto the biomass

Biological wastewater treatment involves the biotechnological application of processes that occur in nature as well, albeit very slowly. These are the processes known broadly as the "selfpurification of surface water."

In the presence of dissolved oxygen (DO), biological metabolism on the part of certain microorganisms always found in a body of water proceeds in an *aerobic* fashion. In this process, organic wastewater components may be oxidized all the way to carbon dioxide (*mineralization*) or some other metabolic product, whereas ammonium ion is transformed into nitrate (*nitrification*). In the case of excessive contamination with oxidizable substances, all available dissolved oxygen ultimately becomes consumed. Replenishment of the oxygen supply by diffusion through the air – water interface is limited, for which reason aerobic metabolism slowly ceases, and a change occurs via the *anoxic* state to *anaerobic* metabolism. Under anoxic conditions, oxygen required for aerobic metabolism is acquired through the biological reduction of some oxygen donor (e.g., nitrate to nitrogen, known as *denitrification*, or sulfate to hydrogen sulfide). Purely anaerobic metabolism is limited primarily to sediments at the bottom of a natural body of water. Here organic substances are first converted to carboxylic acids (*acidification* step) and then disproportionated to carbonic acid, methane, and hydrogen (*methanogenic fermentation*).

Targeted biological wastewater treatment takes advantage of all three types of metabolic process: aerobic, anoxic, and anaerobic. The choice of a particular process or perhaps a combination of processes depends on the quality of the wastewater in question and the goal of the treatment. The advantage of biological wastewater treatment is the enormous adaptability of microorganisms to a wide variety of nutrient media. Such adaptation to different wastewater components and wastewater qualities may occur very rapidly or it may require longer periods of time. During adaptation, metabolism is altered through the synthesis of new enzymes or exoenzymes, or else new metabolic pathways may become available. Successful adaptations are often coded into *plasmids* (see Section 2.1.1.1), which are relatively small DNA units not permanently integrated into the bacte-

rial genome. This metabolic activity, which frequently provides a competitive advantage, can be transferred to other microorganisms by *plasmid transfer*. Since adapted microorganisms are able to metabolize the available substrates more effectively and more rapidly, they also produce more cell substance and undergo multiplication by cell division. Due to plasmid transfer and rapid growth of the microorganisms, biological systems exhibit a high degree of adaptability, even to greatly fluctuating wastewater compositions.

Since anaerobic wastewater treatment has an efficiency <85 % [5], [6], it is essentially used only for pretreating wastewater characterized by a high organic load (>5000 mg/L) and, insofar as possible, a constant quality (e.g., that from the production of citric acid, yeast, alcohol, sauerkraut, sugar, paper, or pulp). It offers the advantages of low energy consumption, production of an energy-rich sewage gas that may be amenable to further use after some preparation, and relatively little clarification sludge that must be subjected to subsequent treatment (see Section 2.1.2). On the other hand, this same low rate of sludge accumulation is responsible for the high sensitivity of anaerobic systems to toxic components of the inflowing wastewater, and for increased discharge of microorganisms (see the discussion of bacterial retention in Section 2.1.2).

Aerobic wastewater treatment generally represents the last biological step. It offers the advantage of a high rate of sludge growth that not only facilitates handling the various components of inflowing toxic wastewater but also provides an efficiency > 85 %. The disadvantages are high energy consumption in supplying oxygen to the microorganisms and the production of considerable clarification sludge (see Section 2.1.2). Table 1 provides a comparison of the two processes.

Processes that might be described as "close to nature" were important in the early stages of the development of biological wastewater treatment [7]. These include landbased treatment (spray distribution, irrigation, infiltration fields), use of aerated and nonaerated wastewater ponds [8], and fish ponds. Although these never played a significant role in the treatment of wastewater from the chemical industry, they have been used for municipal sewage and the wastewater from sugar factories and facilities processing animal cadavers. Today such processes are rarely applied, because they give rise to such insoluble problems as hygienic concerns, nuisance due to flies, and unpleasant odors – a result of the anaerobic conditions encountered at least part of the time. Nevertheless, wastewater ponds have experienced something of a renaissance as "polishing ponds" [9], especially in the area of municipal wastewater treatment. Processes of a more technical nature were developed in the late 19th and early 20th centuries, primarily in England. These include intermittently submerged reactors, the trickling-filter process, and the activated-sludge process (see Section 2.1.2).

The method of choice in the chemical industry today is the activated-sludge process, occasionally replaced by the trickling filter in individual cases (see Section 2.1.4).

In the interest of solving specific problems, the process-engineering aspects of the activated-sludge process have been optimized and incorporated into closed systems as a way of avoiding the potential emission of odors [10]. These special processes utilize deep water columns in order to minimize the off-gas volume, taking advantage of the increased solubility and uptake of oxygen under pressure. They include biotowers (Turmbiologie, Hochbioreaktor) and deep-shaft systems, as described in Section 2.1.3.

Prerequisites for the successful implementation of biological wastewater treatment are a system-dependent microbial degradability of the wastewater components and the absence of effective concentrations of substances that would inhibit degradation. These considerations apply to all aspects of the wastewater regime.

In the chemical industry, economic factors usually dictate the inclusion of a wastewater-separation system. Thus, wastewater that is not in need of treatment (*clean water*), especially cooling water, is separated from that which does require treatment. Clean water can be discharged directly into the receiving stream. If the wastewater requiring treatment fails to meet the quality specifications for biological waste treatment it must first be subjected to decentralized chemicophysical pretreatment, after which it can be fed into the central waste-treatment plant for purification [7], [11] (Fig. 2).

Specific wastewater loads can be reduced or even avoided through measures of the type

Table 1. Advantages, disadvantages, and specific process parameters associated with anaerobic, anoxic, and aerobic biology

Parameter	Anaerobic	Anoxic	Aerobic
DO (mg O_2/L)	0	0	>0
Energy consumption	low	low	high
Sludge accumulation	low	high	high
Sensitivity to toxic substances	high	low	low
Removal efficiency (COD)	<85 %	varies depending upon denitrification	>85 %
Removal efficiency for nitrogen	0	45 – 90 % (nitrification required)	0
Suitable as pretreatment	yes	yes	yes
Suitable as last stage	no	no	yes

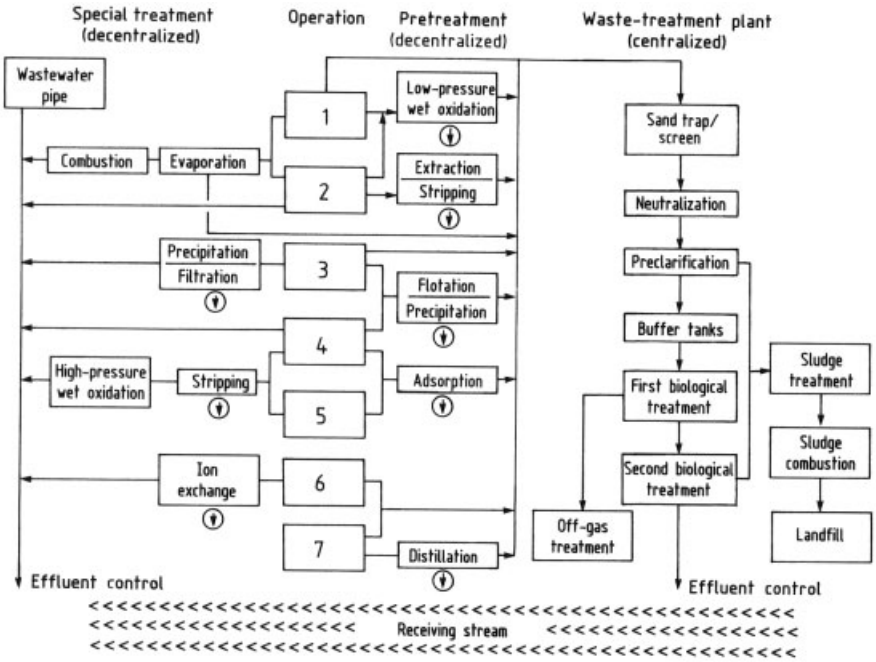

Figure 2. Decentralized and centralized aspects of wastewater treatment

recommended in conjunction with process-integrated environmental protection [7], [11], [12].

The entire wastewater regimen is also reflected in the approach taken to wastewater decision making. For each type of wastewater, all the following questions must be rigorously addressed to ensure consistent and safe disposal (see Fig. 3):

1) Can the amount and contamination level of the wastewater be reduced or even eliminated by process-integrated means?
2) Does all the wastewater in question in fact require treatment (i.e., would a separation system be useful)?
3) Is the wastewater suitable in its present form for biological treatment, or should it be subjected to decentralized pretreatment?

These considerations apply equally with respect to both direct discharge (for companies that have their own central wastewater-treatment plants) and indirect discharge (for companies that dispose of their wastewater via a public wastewater facility).

2.1. Aerobic Treatment

In the self-purification of any body of water, oxidative metabolic processes tend to be very slow:

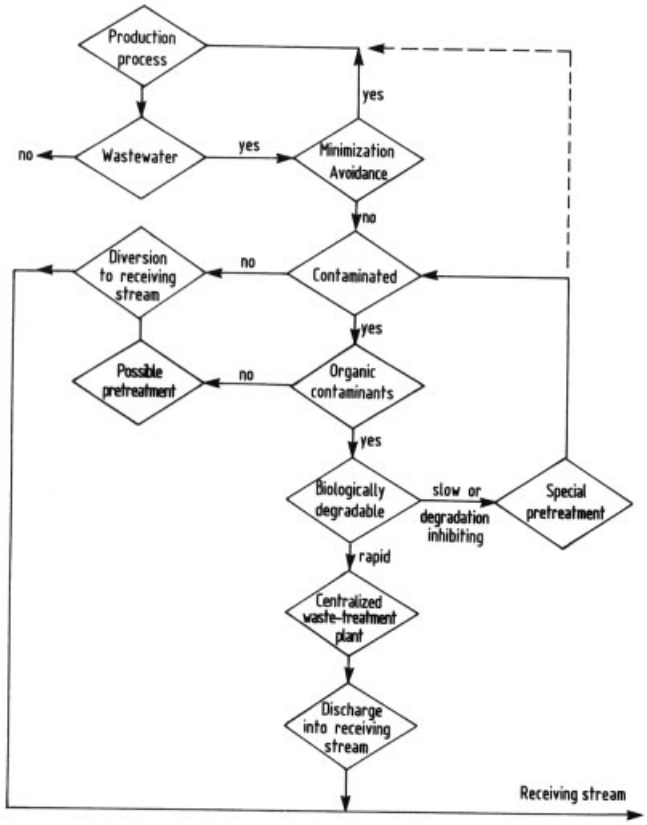

Figure 3. Decision diagram dealing with wastewater formation, avoidance, separation, and treatment

on one hand because the concentration of nutrients is low, but also because almost no optimization of the process engineering is possible apart from aeration of the water.

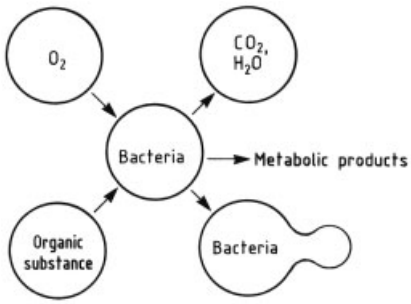

Figure 4. Aerobic metabolism

In processes, such as these, heterotrophic organisms oxidize organic substances to CO_2 or some other metabolic product, thereby utilizing organic materials as a carbon source for creating new cell substance, with rapid multiplication by cell division (Fig. 4). Meanwhile, autotrophic organisms oxidize ammonium ion (NH_4^+) to nitrate (NO_3^-), where cell substance must be derived from CO_2 as the carbon source, resulting in slow multiplication by cell division.

The biotechnological application of this self-purification scheme requires optimization of the process engineering in such a way as to attain the highest possible space – time yield with adequate efficiency. A few of the relevant factors are as follows:

1) The concentration of nutrients in the wastewater should not be too low. This particular requirement can normally be met.
2) As many suitable organisms as possible should be present within the bulk (bacterial retention) to ensure stable biocenosis (mixed culture; see Section 2.1.1.2).

3) The bacteria must be provided with an adequate supply of oxygen to support aerobic metabolism (aeration).
4) The three reaction partners – bacteria, substrate, and dissolved oxygen – must be brought together in a targeted way.
5) The growth rates for the different bacteria must be well matched (sludge age).
6) The optimal environmental conditions (pH, temperature) for biological degradation must be established.
7) The C: N: P ratio required for the growth of cell substance must be assured [13], and all required trace elements must be present.

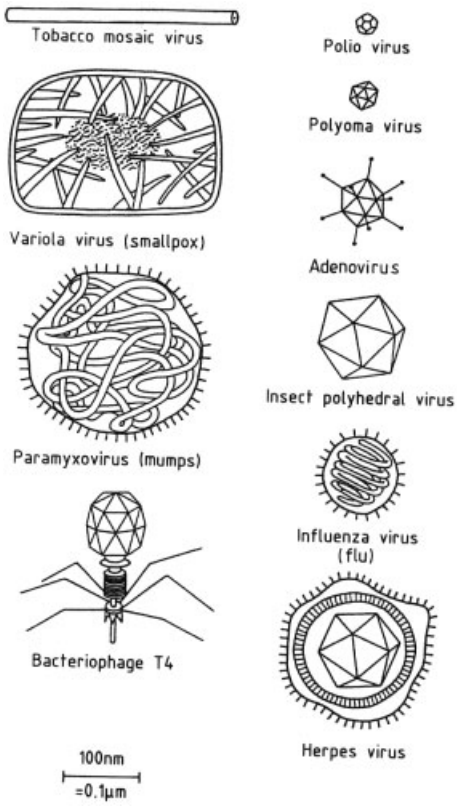

Figure 5. Relative size and shape of various viruses [17]

2.1.1. Biological Principles

Broadly speaking, only chemical reactions or physical separation processes lend themselves to the elimination of unwanted wastewater components. Living creatures, especially microorganisms, have developed a tremendous potential for chemical reactivity over the course of time (see Section 2.1.1.1). Their cells in fact constitute tiny "bio"chemical reactors. Aerobic, heterotrophic organisms oxidize organic compounds with the aid of oxygen. Moreover, the microorganisms themselves represent natural sorbents that bind both organic substances and heavy metals. The flocs, threads, and films formed by the aggregation and intergrowth of these organisms behave like small particle filters that clarify the wastewater mechanically. Associations of this type (*biocenoses*, see Section 2.1.1.2) can be established in many kinds of reactors – activated-sludge stages, trickling filters, intermittently submerged reactors, waste ponds – to assist in wastewater treatment. The conversions leading to actual degradation of unwanted wastewater components constitute integral parts of the phenomenon of biological metabolism (Section 2.1.1.3).

2.1.1.1. Organisms

Organisms can be regarded as unicellular or multicellular compartments that are separated from the environment, have their own metabolism, and are capable of growth and potentially of multiplication. Derived from the Greek word *organon* (tool, instrument, organ) via the Late Latin word *organicus* (well-ordered) the designation "organism" indicates that despite their apparently trivial accomplishments, such creatures have developed complicated, closely regulated systems for performing their tasks, and also that they act on one another and on their environment in a complex manner [14].

As a rule, organisms involved in biological wastewater treatment are invisible to the naked eye. Depending upon their structures and cellular components they can be subdivided into bacteria, fungi, plants, and animals, with viruses representing a special group of their own.

Viruses. The term "virus" is borrowed from Latin, and means "poison." Viruses lack an independent metabolism, and are unable to grow and multiply on their own. Viruses are not organisms; they are only modified, independent building blocks derived from organisms [15], [16]. Since viruses do not have their own metabolism

they are necessarily intracellular parasites, dependent upon true living organisms for their replication. A distinction is made between deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) viruses depending on the nature of the carrier of genetic information. The corresponding DNA or RNA is associated with proteins (the *core*) packed in a shell (*capsid*) of mostly uniformly arranged proteins (*capsomeres*), which may in turn be surrounded by another shell (the *envelope*). Most viruses are polyhedral, tubular, or bag-shaped. Some bacteria-destroying bacteriophages display a rocket-shaped projection in addition to a polyhedral capsid (Fig. 5) [17].

Viruses are found in all municipal wastewaters, and they represent a serious problem, especially with respect to hygiene. Tiny pathogens with lengths of up to 2 µm or diameters of up to 200 nm, pass through bacterial filters, and are separable by sedimentation or flotation only when aggregated or bound to other particles [18]. Viruses or their fragments are detectable in treated and untreated wastewaters [19–21], including infectious doses of the pathogenic viruses responsible for poliomyelitis, hepatitis, and gastroenteritis. However, viruses are partially eliminated [22] during wastewater treatment, and they are destroyed by enzymes such as proteinases and nucleases that are formed by bacteria, algae, and other organisms ubiquitously present in the environment. Moreover, their sensitivity to UV light [22] and to photochemically formed reactants is well known [18].

Bacteria. The term bacteria, derived from the Greek word *bakterion* (little rod), refers to a series of very diverse unicellular microorganisms possessing no true cell nucleus; for this reason they are also called *prokaryotes* (after the Greek word *karyon*: nut, kernel) [14–16], [21].

Structure and Morphology [23], [24]. Only in exceptional cases do bacteria exceed a length of 10 µm. Under the light microscope, one can recognize spherical cocci, straight short rods or bent "vibrions," corkscrew-shaped spirilla, rootlike and amoeboid forms, etc. Bacteria also occur in pairs (e.g., diplococci), in fours, in threads, in nets, or in clumps, all encompassed by a common slimy shell. Although actinomycetes form mycelia and myxobacteria form fruiting bodies, cell differentiation is at most quite limited. The so-called wastewater fungus, *Sphaerotilus natans*, a chlamydobacterium, covers all nutrient-rich substrates with mycelia. Actinomycetes mycelia are undesirable in wastewater-treatment facilities because of sludge bulking [25], [26].

The bacterial cell [21] is contained within a cytoplasmic membrane, usually enclosed in a cell wall, that may in turn be surrounded by a capsule or a slime layer. The cell walls of most bacteria consist of a single-layer or multilayer murein sack to which teichoic acids or lipoproteins and lipopolysaccharides are attached. Bacteria with only one or a few murein layers lose stain most readily in Gram's method of staining, and are thus described as "gram-negative." By contrast, bacteria with many murein layers retain their stain (gram-positive). The capsule or slime layer is usually made up of polysaccharides and proteins, and in the case of *Salmonella*, for example, it causes virulence, resistance to bacteriophages, and resistance to drying. The lactic acid bacterium *Leuconostoc mesenteroides* leads to slime formation, especially in wastewater from sugar factories. At the same time, however, slime is what holds together the activated sludge floc, which always involves several types of bacteria [25].

Many bacteria possess one or more *flagella*, which serve as their means of locomotion. Velocities per second of up to 50 times the bacterial body length have been observed. Some bacteria are adapted to sliding on suitable surfaces.

Substances dissolved within the bacterial cell generate an osmotic pressure corresponding to that of a 10 – 20 % sucrose solution. If a stable cell-wall skeleton fails to form (e.g., due to penicillins), or is destroyed (e.g., by lysozyme), or if the internal pressure is increased by heating, the cellular contents escape, resulting in death of the bacterium.

In comparison with the cells of fungi, plants, and animals (eukaryotes), the bacterial cell itself shows little compartmentalization. Organelles like mitochondria, chloroplasts, the Golgi apparatus, and the endoplasmic reticulum are absent entirely, and bacterial ribosomes are smaller than those in eukaryotes [16]. Much of the cell is often filled with invaginations of the cytoplasmic membrane (for photosynthesis), gas vacuoles, or storage substances.

The well-ordered genetic material lies in the nuclear region of the cytoplasm, the *nucleo-*

plasm. There is no nuclear membrane, so bacteria lack a true cell nucleus. The DNA that serves as the genetic material (genome) forms a ring-shaped closed thread, typically with a contour length of up to 3 mm, the native state of which is called the bacterial chromosome. Smaller circular DNA fragments called plasmids are also frequently found in the cell [21]. These may become integrated into the bacterial chromosome – sometimes to be eliminated once again, occasionally taking genome fragments with them in the process (excision) [16], [23], [27], [28]. Although the information stored in plasmids is usually not essential for basic metabolism, it is of great significance with respect to resistances, stress tolerance, and vitamin and nutrient requirements (auxotrophs), as well as for the biological degradation of wastewater components.

Multiplication [16]. Bacteria usually multiply by fission, less frequently by budding. Depending on the species or strain, ten minutes to several hours are required for bacterial division under favorable conditions. The shorter the time between one division and the next (*generation time*), the faster the bacteria fill their habitat, and the better the activated-sludge formation.

If the conditions for life become very unfavorable, many types of bacteria form spores. These can survive extreme environmental conditions, but they make no contribution to material degradation in the wastewater-treatment plant. To assure optimal degradation potentials and to avoid unwanted losses, wastewater with a fluctuating composition should always be tempered.

Composition. The water content of the bacterial cell ranges between 70 and 85 %, and it decreases with increasing concentrations of reserve substances, such as lipids, polysaccharides, poly(β-hydroxybutyric acid), polyphosphates, and sulfur. The solids consist of protein (50 %), cell wall (10 – 20 %), RNA (10 – 20 %), DNA (3 –4 %) and lipids (10 %). The solid matter contains the percentages indicated of the following elements (among others): carbon (50), oxygen (20), nitrogen (14), hydrogen (8), phosphorus (3), sulfur (1), potassium (1), sodium (1), calcium (0.5), magnesium (0.5), chlorine (0.5), and iron (0.2) [15]. The density of bacteria is influenced greatly by the type and concentration of reserve substances. A high protein content results in a high density. The density may fall below 1 kg/L due to the deposition of lipids – to be expected in wastewater facilities subject to high loading.

Metabolism. Most bacteria utilize organic substrates as their sources of energy, hydrogen, and carbon; i.e., they are *chemoorganoheterotrophic*. These bacteria constitute the most important group of organisms useful for the degradation of organic substrates. Photosynthesizing cyanobacteria – formerly known as blue-green algae – absorb light as a way of meeting their primary energy requirement, utilize water or other (inorganic) compounds to meet their hydrogen requirement, and withdraw carbon dioxide from the environment to meet their carbon requirement. They are described as *photolithoautotrophic*. Many nitrifying bacteria trap inorganic sources of energy, hydrogen, and carbon, thereby classing among the *chemolithoautotrophic* organisms. Thiobacilli oxidize sulfides and other inorganic sulfur compounds as a way of obtaining energy and electrons. The obligatory chemolithoautotrophic thiobacilli utilize only carbon dioxide as a source of carbon, whereas the facultative autotrophic bacilli are also able to metabolize organic compounds [16], [21].

Ecology. Oxygen content is of tremendous importance in the activated-sludge stage, because oxygen is not only a substrate for respiration but also a source of toxic oxygen radicals. In addition, the carbon dioxide content also plays a role, especially because CO_2 lowers the pH. Most bacteria prefer pH values near the neutral point (pH 7). A few specialized bacteria tolerate extremes; e.g., sulfuric acid-forming thiobacilli tolerate a pH of 1 (acidophilicity). Most bacteria grow best at temperatures between 20 and 45°C, and are referred to as *mesophilic* organisms. Bacterial growth is retarded considerably in the cold, a factor that can interfere with the efficiency of wastewater treatment. Many bacteria are able to survive for years in the frozen state. Some extremely thermophilic archaeobacteria continue to grow at temperatures above 100°C, but die at low temperatures.

An optimal salt content for many freshwater bacteria appears to be ca. 3 %. Halobacteria thrive in salt concentrations >30 % [25].

Classification. In contrast to the natural or phylogenetic system, the common "artificial" system of bacterial classification presented in the

list below [29] is based not on degree of relationship but on morphological, physiological, biochemical, ecological, and other characteristics. These include shape, flagella, spore-forming ability, cell inclusions, pigmentation and capsule substances, staining properties, oxygen requirement, energy recovery, temperature and pH optima, applicable nutrients, reserve substances, cell-wall structure, sensitivity to antibiotics, pathogenicity and other relationships with respect to foreign organisms, and serological properties (for examples, see [30]). The nomenclature is binomial; i.e., each type is specified on the basis of a generic name followed by a specific name, the *epitheton* (e.g., *Escherichia coli*). The species are assembled in *genera*, the genera in *families* (with names ending in "-aceae"), and the families in *orders* ("-ales").

Group no.	Description
1	Spirochetes
2	Aerobic/microaerophilic, motile, helical, vibrioid gram-negative bacteria
3	Nonmotile (or rarely motile), Gram-negative, curved bacteria
4	Gram-negative aerobic rods and cocci
5	Facultatively anaerobic Gram-negative rods
6	Anaerobic gram-negative straight, curved, and helical rods
7	Dissimilatory sulfate- or sulfur-reducing bacteria
8	Anaerobic gram-negative cocci
9	Rickettsias and chlamydias
10	Mycoplasmas
11	Endosymbionts
12	Gram-positive cocci
13	Endospore-forming gram-positive rods and cocci
14	Regular, nonsporing gram-positive rods
15	Irregular, nonsporing gram-positive rods
16	Mycobacteria
17	Nocardioforms
18	Anoxygenic phototrophic bacteria
19	Oxygenic photosynthetic bacteria
20	Aerobic chemolithotrophic bacteria and associated organisms
21	Budding and/or appendaged bacteria
22	Sheathed bacteria
23	Nonphotosynthetic, nonfruiting, gliding bacteria
24	Fruiting gliding bacteria: the myxobacteria
25	Archaeobacteria
26	Nocardioform actinomycetes
27	Actinomycetes with multilocular sporangia
28	Actinoplanetes
29	Streptomycetes and related genera
30	Maduromycetes
31	Thermonospora and related genera
32	Thermoactinomycetes
33	Other genera

Fungi. The terms *mycete* and *fungus* are derived from the respective Greek and Latin names for pileate fungi (mushrooms). These typical fruiting bodies are found on many treated clarification sludges, but not in the actual wastewater-treatment regimen.

Morphology [21], [25]. Fungi possess true cell nuclei, and like all the groups of organisms mentioned hereafter they are eukaryotes. Since fungi have cell walls and vacuoles, they are frequently classified as plants, even though they differ from plants with respect to a number of important characteristics. For example, fungi possess no photosynthetic apparatus, and they are always chemoorganoheterotrophic. Their poorly differentiated vegetative body (*thallus*) consists of branched threads (*hyphae*) ca. 5 µm in diameter. These in turn colonize the nutrient substrate and form the *mycelium*.

Reproduction. Fungal hyphae grow at their tips. Even small fragments of mycelium are capable of growing. Fungi multiply asexually by means of spores that are abstricted from the ends of the hyphae or formed in special receptacles. Yeasts multiply by budding. Almost all fungi also reproduce sexually. Higher fungi (see below), with a still unknown sexual form, are called "fungi imperfecti."

Classification. Distinctions are made between, among others, slime fungi (Myxomycota), Chytridiomycota, Oomycota, Zygomycota, and higher fungi (Eumycota), including ascomycetes, basidiomycetes, and imperfect fungi (deuteromycetes).

Ecology. Fungi colonize any space where other organisms exist or have left behind organic remains. As saprophytes they decompose the remains of plants and animals, spoil food, and infest wood, leather, paper, etc. As parasites they are dreaded pathogens, and their spores cause severe allergic reactions. In symbiotic relationship with algae or cyanobacteria they form lichen, which colonize cold environments. In symbiosis with trees (mycorrhiza) they act to improve the nutrient and water supply. In symbiosis with animals they metabolize otherwise indigestible foods. Their food is usually absorbed in dissolved form, and very seldom engulfed in particle form.

Only about 2 % of fungi live in water [21]. The Chytridiomycota are mainly aquatic fungi, and act as parasites of algae and water plants.

Among the aquatic Oomycota the genera *Saprolegnia* and *Leptomitus* form what is known as "water mold." Yeasts are ascomycetes, which among other things ferment nutrientrich liquids. Higher fungi play a role in the formation of the activated-sludge floc and the biological film, although excessive fungal growth leads to sludge bulking and clogging [25]. In the biological self-purification of surface waters, organic substances are degraded by such species as *Aspergillus niger, Fusarium aquaeductum, Leptomitus lactus*, and *Rhizopus nigricans* [26].

Plants. The term "plant" (Latin *planta*: seedling, shoot) refers not only to typical vascular plants (tracheophytes) such as club mosses, horsetails, ferns, and seed plants, but also to algae and mosses, which as prostrate plants (thallophytes and bryophytes) produce only a small number of differentiated tissue types.

Morphology. The plant cell is typically surrounded by a cell wall composed mainly of cellulose. Cell organelles that are especially conspicuous include the pigmented plastids (e.g., green chloroplasts for photosynthesis) and the vacuoles, which are the storage and waste compartments of the plant cell.

Metabolism. Algae absorb mineral nutrients and, to a small extent, organic substances across their entire surface [31]. Some flagellates engulf pieces of food. On the other hand, mosses and vascular plants take up water and salts by rhizoids or roots. As a result of their photosynthetic productivity, photolithoautotrophic plants constitute the basis of virtually all food chains. Roughly 60×10^9 t of carbon are channeled from the air and water into the biosphere by photosynthesis every year [32]. A few plants have lost their ability to perform photosynthesis, and now live as saprophytes or parasites. The flagellates are the closest organisms to the common ancestors of both animals and plants. Animal zooflagellates often differ from plant phytoflagellates only by the absence of chloroplasts.

Ecology. Most algae live in water, and mosses are usually limited to moist environments. Although horsetails, ferns, and especially seed plants have increasingly colonized drier locations, this category also includes various water plants (hydrophytes), marsh plants (helophytes), and plants that thrive in moist places (hygrophytes) [15].

Especially the algae are important with respect to water-pollution control. Typical algae populations grow in surface waters as a function of the pH, rate of flow, temperature, nutrient supply, lighting conditions, salt content, competitors, phytophagic species, and many other factors. A massive development of single-cell algae (e.g., *Ceratium fusus, Ceratium fruca, Chrysochromulina polylepis, Noctiluca miliaris, Phaeocystis pouchetii*) or cyanobacteria is referred to as "algal bloom" or, if red in color, red tide. They are responsible for the death of fish as a consequence of toxins they generate, oxygen deficiencies caused by respiration or their own biological degradation, poisonous mussels, and excessive foam formation. Algal bloom is sometimes associated with the eutrophication of a body of water (corresponding to an oversupply of fertilizer), but a regular succession of algal blooms in the North Sea [15] and other bodies of water has been found to be independent of the anthropogenic nutrient supply.

Euglenophyta are common in heavily polluted waters, oxidation ponds, and other nutrient-rich sites. They are tolerant of both large fluctuations in pH and anaerobic conditions, and are regarded as marker organisms for contaminated waters. The golden algae (Chrysophyta) include the genus *Synura*, which colonizes trickling filters and wastewater ponds. Many green algae of the order Volvocales live in eutrophic waters, including the family Chlamydomonadaceae, and they may even colonize specific types of wastewater (e.g., *Pandorina* and *Spondylomorum*, which colonize the wastewater from paper mills). Only a few types of green algae grow in the well-mixed, turbid activated-sludge stages of wastewater-treatment plants due to the lack of light. They are sometimes found inhabiting the inner walls near the water level. Upper layers of trickling filters that are exposed to light may be colonized by green algae of the genera *Chlorella, Chlorococcum, Stigeoclonium*, and *Ulothrix* [25].

In the presence of light, algae slightly improve the oxygen supply of the biological film. Oxygen produced by algae also substantially reduces the required residence time for wastewater in an oxidation pond. Small, spherical, green algae of the genus *Chlorella* are found in such ponds at all times of the year [25].

Mosses are typical inhabitants of environments that are poor in nutrients. Only in exceptional cases are they found at receiving streams contaminated with wastewater (*Marchantia* of the stone liverwort family), on slimy banks of ponds and rivers, and on trickling filters (*Leptodictyum* of the class Musci).

Higher plants do not normally grow in activated-sludge stages. Various rushes, bulrushes, grasses, and herbage thrive in polishing ponds, plant-wastewater treatment facilities or other wastewater ponds [14], [33], [34]. In particular, the sedges (Cyperaceae), with aerating tissue in their stems, are able to supply nutrient-rich sediments with atmospheric oxygen. In addition, bactericidal emissions from the roots have been reported. The unobtrusive sedge *Schoenoplectus lacustris* is already being introduced into numerous wastewater ponds, less frequently the reed *Phragmites australis*, one of the sweet grasses. Reeds are also used for the dewatering of clarification sludge because of their potential for oxygen transfer via the roots [35].

Animals. As a rule, animals are more complex organisms than plants, although they lack a cell wall, photosynthetic apparatus, and vacuoles [36]. Most animals are mobile, having become sessile only in water and as parasites. Various groups of small animals are of great importance in wastewater biology since they consume and metabolize bacteria, fungi, algae, and other types of biomass. The most important are as follows [21], [25], [36]:

1) Subkingdom Protozoa
   1. Flagellata
      - Protomonadina (often sessile)
      - Rhizomonadina (cilia as sensory organs, sessile)
   2. Rhizopoda
      - Amoebina (flowing-creeping, naked amoeba)
      - Testacea (shelled amoeba)
      - Heliozoa (skeletons with radiating axial rods)
   5. Ciliata
      - Holotricha (largely covered by cilia)
      - Spirotricha (cilia especially in the mouth region)
      - Peritricha (bearing a distinctive wreath of cilia)
      - Suctoria (tentacles)
2) Subkingdom Metazoa
   14. Nemathelminthes (round worms)
      - Rotifera (see text)
      - Nematoda (see text)
   20. Annelida (segmented worms)
   25. Arthropoda (segmented, joint-footed animals)
      - Crustacea (including crayfish, lobsters, crabs)
      - Insecta

Protozoa. Morphology [24], [25], [28], [36], [37]: Unicellular protozoans include flagellates, which are closely related to algae; rhizopodes (Sarcodina) which are distinguished by slimy projections (pseudopodia); and ciliates with two types of nuclei, which move and acquire food by means of cilia.

Ecology: Virtually all protozoans are capable of ingesting particulate food, especially bacteria, with the aid of specialized cell regions. Among the ciliates, for example, the slipper animalcule (*Paramecium*) whirls food into its mouth opening, Didinium numbs its prey with toxins from transformed cilia (trichocysts), and suctorians capture their prey with tentacles. Flagellates and ciliates are responsible above all for ingesting individual suspended bacteria and small particles that would otherwise enter into the effluent from a wastewater-treatment plant. Amoebas consume bacteria in activated-sludge floc and retard sludge formation.

Metazoa. The multicellular metazoans exhibit sensory organs as well as nervous, circulatory, digestive, and excretory systems, which frequently appear in reduced form in small organisms or highly specialized parasites [36].

Rotifera. Morphology [24], [25], [37]: The wheel animalcules (rotifers) are seldom larger than 0.5 mm, and their number of cells is fixed at a level usually attained after a few hours of growth. This is followed by an approximately one-day period of maturity. Rotifers acquired this name as a consequence of a corona of cilia surrounding the mouth area, which gives the impression of a rotating wheel. Within the group, the wheel organ (corona) can be transformed or simplified in diverse ways. The trunk is often sack-shaped, armored, and prickly. The head with its wheel organ and a telescope-shaped foot with adhesive glands can be withdrawn into the

trunk. Many species have a conspicuous masticatory (chewing) apparatus in the gizzard that operates constantly and is readily discernible.

Ecology: Many rotifers acquire nutrients by swirling; some also feed on activated-sludge floc, thus stimulating floc renewal. Rotifers frequently colonize small freshwater habitats that exist only for a short time. They are seldom found in the sea. When the water is drained off they survive by drying out and assuming a state of dry torpor. Free-swimming, fluke-like creeping, and sessile species are all known. In such wastewater-treatment facilities as aeration tanks, trickling filters, and oxidation ponds, the two orders Bdelloidea and Monogononta are particularly common.

Nematoda. Morphology: The free-living threadworms (nematodes) are spindle shaped, and only rarely longer than 3 mm [33], [37]. They are surrounded by a coarse shell (*cuticula*), which serves as an antagonist for the longitudinally oriented muscles. No transverse muscles are required. Nematodes wriggle in a strange manner; this mixes their body fluids, rendering a blood circulatory system unnecessary. A nematode can be identified above all by its mouth, mouthparts, and foregut.

Ecology: Nematodes live as parasites or saprophytes in substrates rich in nutrients and poor in oxygen, such as host tissue, fungi, sediments, or feces. They are poorly adapted to life in aqueous medium on activated-sludge floc. They reside in aerobic layers in the sediments of wastewater ponds or in the biofilms of trickling filters, supporting aeration and regeneration by their feeding activity [25].

Annelida. The segmented worms (annelids) are common in trickling filters. The earthworm is the best known representative of this class. Important species include *Enchytraeus coronatus* and *Lumbricillus lineatus*, which loosen the biological film and can even protect it from becoming clogged with sludge. The population density of annelids often depends on the population of trickling-filter flies (*Psychodae*). The fly larvae eat the worms and their eggs.

By contrast, the worm species *Tubifex* is encountered only rarely in trickling filters. It is regarded primarily as an indicator of the quality of surface waters [25].

Crustacea. Various small crustaceans are found in grade II and III surface waters, in the outflow from slightly loaded trickling filters, and in wastewater ponds. These creatures live as plankton and feed on detritus, algae, or other animals. They indicate water of relatively high quality [25].

Insecta. Two orders of insects are commonly found in trickling filters: springtails (Collembola) and flies (Diptera). The Collembola are omnivorous and colonize the biological film. The adult trickling-filter fly *Psychoda* is poorly equipped for flying, and actually belongs to the family of moth gnats, known for their fluttering flight. *Psychoda* larvae bore into the biological film and counteract clogging [33]. The mass reproduction of trickling filter flies is a nuisance, and can be controlled by a day-long flooding of the trickling filter [25].

High population densities of chironomids can also become problematic. The eggs, larvae, and pupae swim in jelly-like clumps on the surface of the water, interfering with the efficiency of the activated-sludge tanks [25].

2.1.1.2. Biocenosis

Not only free-swimming individual organisms, but also aggregates of organisms are found in wastewater and in treatment plants. To some extent these aggregates constitute highly organized communities (biocenoses), that are responsible for much of the biological degradation efficiency of a wastewater-treatment plant. All the various organisms, abiotic particles, and other suspended wastewater components in the thickened wastewater constitute the *activated sludge*, which typically contains about 1 % dry matter. The most important process-engineering parameter with respect to activated sludge is the sludge age, obtained by dividing the total mass of sludge dry matter by the amount discharged daily [33].

Activated sludge always contains bacteria. In industrial wastewater-treatment plants, for example, the following genera have been identified: *Acinetobacter, Aeromonas, Agrobacterium, Alcaligenes, Azotobacter, Bacillus, Brevibacillus, Chromobacterium, Enterobacter, Escherichia, Flavobacterium, Microbacterium, Micrococcus, Moraxella, Mycobacterium, Nocardia, Pseudomonas*, and *Zoogloea* [30]. Depending on the age of the sludge and its

load, populations of protozoans and metazoans may be established. In older sludge, the slow-growing nitrifying bacteria increase in importance (Fig. 6).

Some organisms swim freely in the wastewater, and others form typical aggregates that function as biocenotic teams. Depending upon appearance, a distinction is made between suspended activated-sludge floc, fixed biofilm, and (as undesirable forms) floating bulking sludge and fast-peeling biofilm.

Activated-Sludge Floc. Activated-sludge floc represents the smallest unit of biocenosis with a high potential for biodegradation of organic substances. The activated floc consists of bacteria, fungi, protozoans, and metazoans. The organic composition varies depending on the relevant ecofactors,such as operating conditions,wastewater composition, climate, season, etc. [24]. The diameter of the irregularly formed, spongy activated-sludge floc is typically 50 – 200 µm [39]. The floc is normally gray to brown in color. Floc is produced by bacteria that form slime shells or capsules, as well as by fungi. Especially valuable in this respect are, for example, *Bacterium aerogenes, Bacterium cercus, Bacterium proteus, Enterobacter aerogenes, Nocardia actinomorpha*, and *Zoogloea ramigera* [25]. Parts of the floc constantly become lost due to both turbulence and feeding. Flocs are regenerated by regrowth from fragments, and they are colonized by free-swimming organisms. Increasing floc size leads to a decrease in the oxygen and nutrient supply for organisms in the interior of the floc, causing their growth to stagnate. Ciliates are indicators of healthy flocs [39]. These can be determined on the basis of microscopic and physiological characteristics [24], [40]. Indicator organisms for monitoring floc status are identified in Figure 7 [25]. For municipal wastewater-treatment plants, a rule of thumb states that organisms associated with water-quality classes II and III are found in healthy activated-sludge flocs, and these for water classes III and IV are present in poorly functioning activated-sludge flocs. Organisms of the latter type include amoebas, flagellates, spirilla, rod-shaped bacteria, and slipper animalcules [25].

Bulking sludge is activated sludge in which the extent of activated-sludge floc is diminished relative to filament-like organisms [41]. Although the purification power of bulking sludge is good, its presence is undesirable in wastewatertreatment plants because of its poor separability [25], [33]. A rich food supply is available in a heavily loaded treatment plant. In the event that the carbon/nitrogen ratio is high [15], filament-forming microorganisms grow especially rapidly, storing large quantities of lipids, polysaccharides, and poly(β-hydroxybutyric acid). Moreover, the branched bacterial filaments and fungal hyphae readily retain gas bubbles. Malodorous gases are formed in the anaerobic zones associated with wastewater activation. The term "bulking sludge" refers to floating sludge that rises with visible gas bubbles, and is often related to odor production. Filament-forming bacteria, especially those belonging to the group actinomycetes, as well as fungi are responsible for the unwanted appearance of a viscous top scum [42]. Apart from *Spaerotilus natans, Microthrix parvicella,* representatives of the genus *Nocardia*, and other bacteria [43], fungi of the genera *Leptomitus* and *Geotrichum* have been identified [25]. Upon exhaustion of other nutrients, actinomycetes generate poly(β-hydroxybutyric acid) from excess carbon compounds that are present. Appearance of this phenomenon may also indicate the presence of high levels of sulfides [44], hydrocarbons, surfactants, organic acids, or alcohols in the wastewater, or perhaps the absence of adequate aeration. Filaments exhibit a relatively large surface area relative to flocs, so organisms living in the filaments are able to absorb dissolved substances more easily then those living in flocs. On the other hand, this hypothesis suggests that flocs have the advantage of more effective particle absorption, so an equilibrium should be established between flocs and filaments, the position of which would be a function of the state of the nutrient supply [25]. Sludge bulking is promoted by high temperatures and finely dispersed aeration [45].

Biofilm. A microbial coating forms on all solid surfaces that are in permanent contact with wastewater. This coating is called *biofilm* as a result of its manner of formation, or *microbial lawn* on the basis of its appearance. Biofilm has

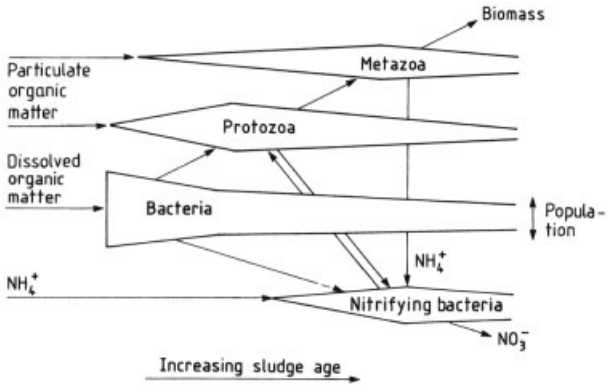

Figure 6. Simplified schematic representation of the qualitative compositions of microbial biocenoses as a function of sludge age [38]

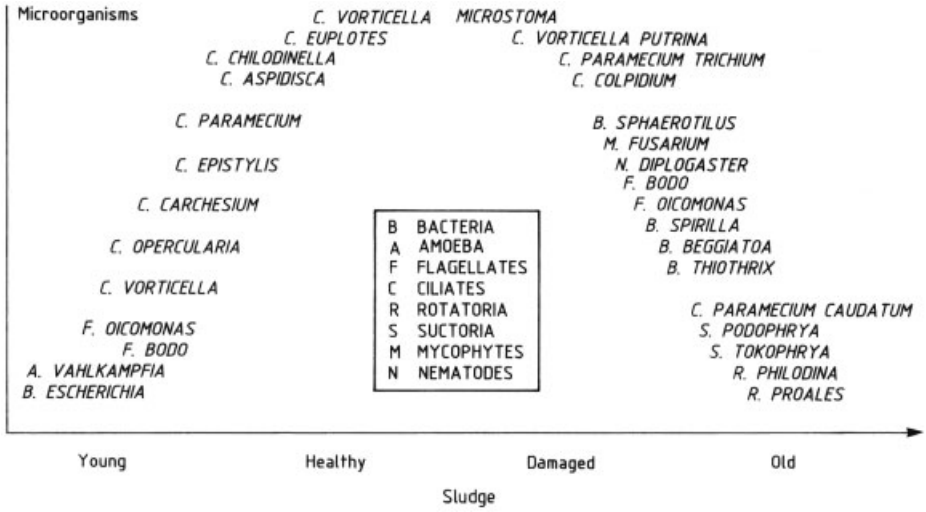

Figure 7. Indicator organisms present in activated-sludge floc [25]

a slimy feel. It is a gel consisting of extracellular organic polymers that enclose various microorganisms and other particles.

The formation of a biofilm, its strength, and its structure are determined by the medium (wastewater composition, temperature, pH), hydrodynamic conditions (shear forces, pressure), substrate (chemical composition, solubility, roughness, porosity, surface charge, hydrophobicity), and nature of the microorganisms present [44]. The only surfaces that remain largely uncolonized are those exposed to a very strong flow, surfaces that are extremely smooth, and surfaces from which biocides constantly diffuse into the water.

The symbiosis within a biofilm consists of bacteria, fungi, protozoans, and in most cases also metazoans and algae. Metazoans such as rotifers, threadworms, segmented worms, springtails, and fly larvae develop especially in lightly loaded treatment plants. Some metazoans are able to filter particles out of the wastewater, thereby improving the effluent quality. In particular, metazoans graze on the biofilm, thereby improving its aeration, sometimes increasing its biological efficiency, and reducing sludge formation. However, excessive amounts of metazoans cause the biofilm to become detached. Algae develop on exposure to light, and under favorable conditions they form distinguishable layers, similar to the colorful striped muds found,

for example, along the North Sea [46], which are formed and stained as a function of the supply of light, oxygen, and nutrients. Algae are also transported passively into the unlighted areas of trickling filters where some of them survive heterotrophically (e.g., the silica algae *Navicula cryptocephala*) [25].

In trickling filters, intermittently submerged reactors, oxidation ditches, wastewater ponds, or even natural waters, biofilms degrade a large share of the organic compounds present, and contribute significantly to self-purification. Consistent with the cleaning effect of trickling filters or intermittently submerged reactors, indicator organisms found on the influent side are characteristic of higher loads than those on the effluent side [25]. The biofilm itself is also not homogeneous. The concentrations of oxygen, nutrients, nitrate, and sulfate decrease from the surface to the base. The aerobic surface layer, with a thickness ≤ 3 mm, consumes the largest share of the nutrients and oxygen [33]. Below this layer is an oxygen-free (anoxic) layer containing nitrate, in which denitrification occurs. A sulfate-reduction layer may exist underneath this layer [14].

2.1.1.3. Metabolism

Metabolism (derived from the Greek word *metabole*: to turn about, change, convert [21]) encompasses the sum of all chemical conversions that occur within organisms, and even some reactions that occur outside the organisms as a consequence of excreted enzymes [47–49]. A distinction is made between *anabolism* (the synthesis of organic substance) and *catabolism* (degradative metabolism). *Primary* metabolism refers to metabolism directly required for life. *Secondary* metabolism refers to conversions that are not directly necessary for life, such as the synthesis of antibiotics. Metabolism in the context of ecochemistry is the degradation of substances that makes possible the growth of organisms, while *cometabolism* comprises degradation that occurs only in the presence of other easily usable substances [50].

The physical and chemical parameters in a wastewater-treatment plant must be adjusted in such a way as to permit a full spectrum of metabolic processes to occur. These biological conversions are by nature chemical reactions, and they are subject to the same natural laws as chemistry in the test tube. Conditions that must be maintained and the ways in which important metabolic processes proceed are described below.

Reactions and Catalysts. Almost all metabolic reactions are catalyzed by biogenic enzymes [48]. Enzymes, like industrial catalysts, only accelerate the establishment of equilibrium by reducing the corresponding activation energy; they do not change the position of equilibrium. Therefore, just as with industrial reactions, starting materials are always "left over" in metabolic reactions as residual materials. According to the law of mass action, there is never a complete conversion of material. As in industrial systems, however, the amount of residual starting material can become very small if the products are constantly removed or in some way utilized. Entire reaction sequences proceed according to this principle, with the result that only small concentrations of the many intermediates appear, even though the starting materials or end products may be present in large amounts. This is referred to as a dynamic equilibrium process. For example, every day a human being uses up about his or her own weight of adenosine triphosphate, a chemical equivalent of energy. However, at any given time the body contains only tiny amounts of this compound, which is continually regenerated [15]. This principle has significant advantages. Enzymes and cofactors required for the various conversions are used much more frequently, and are thus present in smaller amounts. Metabolism itself can be closely regulated via enzyme activity, appropriate coupled reactions, the introduction of intermediates (anaplerotic reactions), and regeneration of the cofactors.

Most enzymes (*holoenzymes*) consist of a protein component (*apoenzyme*) together with a permanently bound prosthetic group, or else they are reversibly bound to typical cosubstrates (*coenzymes*). Coenzymes serve as reaction partners for the substrate, and contain electrons or chemical groups present in an easily convertible form. Thus, instead of utilizing free hydrogen in reductions, enzymes take advantage, for example, of hydrogen available in the coenzyme nicotinamide adenine dinucleotide (NAD) in its

reduced form. Enzyme structures undergo reversible changes during the course of a reaction [48], [51]. The part of the enzyme that binds the substrate and catalyzes the reaction itself is called the *active site*.

The most important parameters influencing the activity of enzymes and organisms in wastewater-treatment plants are pH, temperature, and oxygen content.

pH. The ability of an active site to bind substrate and coenzyme depends on the states of charge or protonation of both the enzyme and its reaction partners. In addition, pH determines the water solubility of important gases, such as carbon dioxide, hydrogen sulfide, and ammonia. The pH also affects the bioavailability of heavy-metal ions and other materials. Extreme pH values can lead to the denaturation of proteins.

Temperature. Reaction partners reach the active site by diffusion, which increases with temperature. According to the Arrhenius equation, it is possible in the case of typical biochemical reactions to conclude that the fraction of the reacting components with energy equal to or greater than the activation energy roughly doubles with a 10 K increase in temperature. This is the case, for example, with many fermentations and reactions controlling bacterial growth.

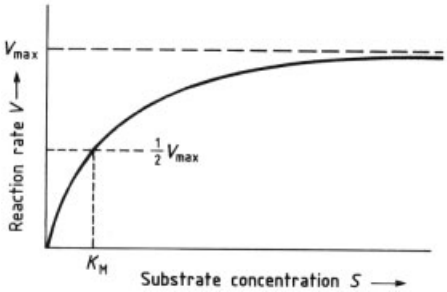

Figure 8. Relationship between conversion rate and substrate concentration [48]

Oxygen. Oxygen is the most important electron acceptor for most organisms, and it is the medium that makes respiration possible in what is known as the *aerobic* or *oxic* state. If oxygen is lacking, the state is called *anoxic*. In the anoxic state, many microorganisms are able to utilize substances such as nitrate rather than oxygen (see Denitrification p 21). If nitrate too is lacking, the medium is described as *anaerobic*, and sulfate, carbon dioxide, or protons become subject to reduction, or organic substrates are fermented [52]. A table characterizing the metabolic activities of activated sludge as a function of the available electron acceptor is presented in Section 2.1.2 (Table 4).

Regulation and Adaptation. The degradation efficiency of a wastewater-treatment plant may continue to increase even after years of operation. The reasons for increasing adaptation of the microorganisms to their substrates and environmental conditions include:

1) Regulation or activation of the enzymes present; enzyme activities are subject to regulation within seconds or minutes.
2) Induction or the new formation of enzymes, with the time required being on the order of hours.
3) Appearance of new bacteria by mutation and/or plasmid transfer; the individual processes require seconds to hours, although effects are first discernible much later.
4) Selection and growth of adapted microorganisms, which occurs in the course of days to years.

Enzyme Regulation. The conversion rates of materials can be controlled via the catalytic activity of enzymes. In this context it is appropriate to take note of a few kinetic terms:

The *maximum rate* $V_{\max}$ of an enzyme-catalyzed reaction for the conversion of a substrate A to the product B ($A \rightarrow B$) is obtained with an unlimited excess of substrate (Fig. 8).

The substrate concentration that corresponds to one-half the maximum reaction rate $V_{\max}$ is equivalent to the *Michaelis constant*, K_{m}. The lower the value of K_{m}, the better the conversion at low substrate concentrations.

The *turnover number* or *molecular activity* indicates how many molecules of substrate are converted per molecule of enzyme per second.

A great excess of substrate does not lead to a correspondingly high level of material conversion. In a living organism, conversion rates depend rather on the enzyme characteristics K_{m} and activity. For this reason, regulation of many metabolic pathways starts with the enzymes:

In the case of *competitive* inhibition, a molecule similar to the substrate competes for the active center. The resulting inhibitory effect can be overcome by an excess of substrate.

Competitive inhibition increases the magnitude of K_m.

In *noncompetitive* inhibition the active center becomes damaged, and an excess of substrate is incapable of overcoming the observed inhibition. The value of V_{max} decreases in proportion to the fraction of damaged enzyme molecules.

Allosteric inhibition or activation is based on changes in the structure of the enzyme, whereby *effectors* (i.e., regulatory or inhibitory substances) bind outside the active site at what are known as allosteric sites.

Cooperativity is observed with enzymes that consist of several identical subunits. Cooperativity means that the subunits are turned on and off simultaneously whenever a single subunit is affected.

Enzyme Induction. Moving beyond the control of enzymic activity, entire metabolic pathways in organisms can be regulated by adapting the amount of available enzyme to the demands of the cell. Some level of enzymes is always present in the cell, catalyzing maintenance metabolism among other reactions. However, many enzymes are actually formed only in response to some chemical or physical signal. This process is called *induction*. For instance, bacterial enzymes that digest various sugars are inducible, as are heat-shock proteins. The advantage of induction is that it economizes on building blocks, but there is the disadvantage that it takes a relatively long time before newly formed enzymes become available. New formation can proceed only if an efficient primary metabolism is established, along with sufficient nutrient and reserve substances. Thus, in the case of the batchwise accumulation of wastewater components in a wastewater-treatment plant, an induction adaptation time of as much as several days should be anticipated.

Formation of Adapted Microorganisms. Bacteria are able to transfer DNA from one cell to another, thereby conferring certain characteristics such as degradation potentials.

In *conjugation*, a plasmid and sometimes also chromosomal DNA is transferred from a donor cell to a recipient cell. Conjugation is dependent upon "fertility" genes.

In *transformation*, a "competent" recipient cell accepts a DNA fragment from the medium and incorporates it into its own genome. Transformation is especially common in rapidly growing bacterial populations.

In *transduction*, a bacterial DNA fragment is transferred by means of bacteriophages [23].

DNA transfer has been observed not only between bacteria of the same species, but also between different species [53]. Bacteria from a receiving stream close to the discharge point display more resistance characteristics than bacteria from a load-free receiving stream [27].

Mutations may change genes in such a way as to generate improved enzymes. For example, as a result of mutation a *Pseudomonas* strain was better able to degrade polychlorinated biphenyls than the original (wild) type [54].

Selection. Growth and DNA transfer form the basis for the bacterial utilization of wastewater components and the degradation of this material in a biological wastewater-treatment facility. Given optimum living conditions, bacteria that contain plasmids grow more slowly than bacteria that do not. Consequently, wastewater components metabolized via a plasmid-coded degradation pathway must always remain in the biological stage of a treatment plant for longer periods of time in order to permit the build-up of plasmid-carrying bacterial populations. This process corresponds to natural selection. If the selected wastewater components subsequently disappear (e.g., because production is interrupted), the plasmid-free bacteria outgrow the plasmid carriers. Plasmids can also be eliminated "spontaneously," or become lost during fission [23], [28], [53], [55].

Carbon Oxidation. In the aerobic degradation of such organic substrates as fats (lipids), carbohydrates (polysaccharides), and proteins, the materials are converted with consumption of oxygen either to other organic substances (*biotransformation*) or to inorganic substances such as water, carbon dioxide, and ammonia (*mineralization*).

The degradation of organic compounds is linked through intermediary metabolism to the synthesis of cellular building blocks (Fig. 9). The degradation and synthesis of even complicated substances proceeds via a series of common C_4, C_3, C_2, and C_1 compounds. Two of the most important intermediates are shown in Figure 9: pyruvate, the salt of pyruvic acid, and acetyl CoA, an acetic acid residue bound to

coenzyme A [33], [48]. Carbon entering into metabolism in the form of usable organic compounds is not entirely discharged through respiration; at least a portion of it is incorporated into cellular substance [49].

Added carbon is converted almost completely to carbon dioxide only when the organisms in question have largely exhausted their own internal reserves. In a normally loaded activated-sludge stage, far more than one-third of the usable organic carbon is retained in the sludge, about the same amount is released through respiration, and a small part remains in the treated wastewater in the form of nonseparable organic substance. The latter comprises unconverted starting materials, free-swimming organisms, and degradation products of substrates and organisms (e.g., soluble or suspended humic matter).

Ammonification. Ammonification is a mineralization process that leads to the release of ammonia or ammonium ions from organic nitrogen compounds. In particular, bacteria and fungi liberate ammonia in the presence of a rich supply of organic nitrogen. This can be suppressed by the feeding of readily assimilable carbon compounds (see also Nutrientspage 25 and [38]). Plants lose ammonia in photosynthesis via small ventilation openings, the *stomata* [56]. Ammonotelic animals such as fish and amphibians also excrete ammonia into the environment [36].

Domestic sewage contains nitrogen largely in the form of urea, this being the anthropogenic excretory product since humans are ureotelic [36]. Indeed, nitrogen compounds are metabolized and ammonia is formed even in the waste-discharge system itself [57]. An undesirable ammonification process is the conversion of nitrate to ammonia (see also Denitrification page 23 and [49]).

Nitrification. The biological oxidation of ammonia via nitrite to nitrate is called *nitrification*. The most familiar example is that of the chemolithoautotrophic nitrifying bacteria. These bacteria meet their energy requirements by oxidizing ammonia to nitrite or nitrite to nitrate. No known bacterium is capable of oxidizing ammonia chemolithotrophically directly to nitrate. Nitrite-forming bacteria such as *Nitrosomonas, Nitrosococcus, Nitrosospira, Nitrosolobus*, and *Nitrosovibrio* [38] are ammonia oxidizers that produce nitrous acid or nitrite:

$$2\,NH_3 + 3\,O_2 \rightarrow 2\,HNO_2 + 2\,H_2O + 636\,kJ$$

Nitrite-forming bacteria are always associated with nitrate-forming bacteria. These in turn oxidize the toxic nitrite to nitric acid or nitrate:

$$2\,HNO_2 + O_2 \rightarrow 2\,HNO_3 + 184\,kJ$$

Nitrate bacteria belong to the genera *Nitrobacter, Nitrococcus, Nitrospina*, and *Nitrospira* [38]. When *Nitrobacter* encounters an inadequate supply of oxygen during nitrification, nitrite is instead reduced, with the formation of nitrous oxide, N_2O [58], [59].

Nitrification results in the consumption of ca. 4.6 g of oxygen per gram of ammonia nitrogen, a small part of which is recovered in anabolic biosynthesis. Since the final product is a strong acid, nitrification is also linked to the acidification of an activated-sludge stage [57], [60].

Only a few lithoautotrophic nitrifying bacteria are also able to live heterotrophically. Nitrifying bacteria fix carbon dioxide via metabolic pathways like those found in green plants. Indeed, the low usable energy output associated with ammonia and nitrite oxidation necessitates metabolic pathways that are not very energyefficient, such as reverse electron transport. Thus, the growth of chemolithoautotrophic nitrifying bacteria is slow, and actively nitrifying bacteria can only be established in sludge that is highly aged. The operator of a wastewater-treatment plant is faced with another problem as well: nitrite oxidizers are more sensitive than ammonium oxidizers to many (toxic) substances and to any decrease in the oxygen content or pH. Unfavorable conditions lead to the accumulation of nitrite.

The heterotrophic nitrifying organisms include such bacterial species as *Pseudomonas, Streptomycetes*, and *Arthrobacter*, as well as fungi of the order Aspergillacea. In addition to ammonia and nitrite, these are also able to oxidize organic nitrogen compounds. However, heterotrophic nitrification appears to proceed only at a low rate in the form of a cooxidation involving other substrates, so it results in no appreciable contribution to energy metabolism [61].

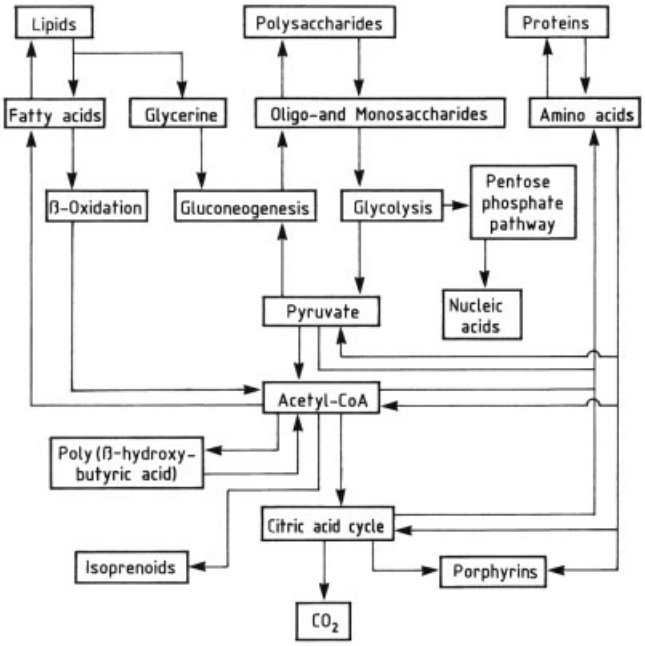

Figure 9. Examples of pathways and linkages in carbon metabolism

Denitrification. Microorganisms can process nitrate via a number of metabolic pathways, thereby reducing it to ammonia, nitrous oxide, or molecular nitrogen (dinitrogen).

The uptake of nitrate is called *nitrate assimilation* when the reduction product ammonia flows into an anabolic pathway to be used, for example, in amino acid synthesis. Nitrate is assimilated by many aquatic bacteria, fungi, and plants. Closely related to nitrate assimilation is the *dissimilative reduction* of nitrate, also known as *nitrate ammonification*. Nitrate assimilation by *Escherichia coli* leds to the liberation of excess ammonia [62]. In strict anaerobic nitrate ammonification, nitrate functions as the terminal electron acceptor [21]. Nitrate ammonification is capable of replacing sulfate reduction in sulfate-reducing bacteria [63].

Denitrification actually refers to the reduction of nitrate to atmospheric nitrogen:

$$10\,[H] + 2\,HNO_3 \rightarrow N_2 + 6\,H_2O$$

The intermediates are nitrite, nitrogen monoxide, and dinitrogen monoxide. The enzymes involved are reductases with names derived from the corresponding starting substances or appropriate intermediates [64].

Nitrate serves in heterotrophic energy metabolism as the terminal electron acceptor in an anoxic respiratory chain. In contrast to the nitrate reductases in nitrate assimilation, most denitrification enzymes are sensitive to oxygen (for an exception see [65]). Consequently, wastewater denitrification is practical only in the anoxic state.

Many types of bacteria are capable of supporting heterotrophic denitrification, though not all of them have the full complement of necessary enzymes. In the absence of nitrate reductase only nitrite can be denitrified. If dinitrogen oxide reductase is absent, nitrous oxide accumulates [38], [58], [66]. The liberation of dinitrogen monoxide, which affects stratospheric ozone, is also to be anticipated when:

1) At very high nitrate concentration the dinitrogen oxide reductase activity is no longer sufficient to reduce in situ all the dinitrogen monoxide.
2) As a result of sequential or, in this case, late enzyme induction, the amount of dinitrogen oxide reductase available is not yet sufficient.
3) The biosynthesis of dinitrogen oxide reductase is inhibited, for example, by oxygen.

4) Dinitrogen oxide reductase activity is inhibited by the toxic effects of such species as hydrogen sulfide, acetylene, or oxygen, or because the pH is too low [58], [67].

Easily degradable carbon sources are required for heterotrophic denitrification [68]. Thus, in the case of glucose:

$$5\,C_6H_{12}O_6 + 24\,HNO_3 \rightarrow 12\,N_2 + 30\,CO_2 + 42\,H_2O$$

In practice, nitrified wastewater is either returned to the raw wastewater, or it is suplemented with a substrate like methanol. Since heterotrophic denitrification consumes acid, denitrifying wastewater can be used to accomplish the neutralization of acidic wastewater. It also reduces the energy expenditure required for oxygen transfer, and compensates for a part of the energy expenditure associated with nitrification.

Apart from heterotrophic denitrifying bacteria, there also exist autotrophic bacteria that accomplish denitrification with the aid of inorganic electron donors. These include *Thiobacillus denitrificans* and *Paracoccus denitrificans*, which utilize sulfide, thiosulfate, or hydrogen:

$$8\,NO_3^- + 5\,S^{2-} + 8\,H^+ \rightarrow 4\,N_2 + 5\,SO_4^{2-} + 4\,H_2O$$

$$8\,NO_3^- + 5\,S_2O_3^{2-} + H_2O \rightarrow 4\,N_2 + 10\,SO_4^{2-} + 2\,H^+$$

$$2\,NO_3^- + 5\,H_2 \rightarrow N_2 + 4\,H_2O + 2\,OH^-$$

One species of *Beggiatoa* can oxidize sulfide, and *Gallionella ferruginea* is able to oxidize iron(II) [49]. Autotrophic denitrifying bacteria are probably irrelevant in the context of wastewater treatment.

Biological Elimination of Phosphorus. Phosphorus in the form of phosphate is an essential nutrient for all organisms. Bacterial biomass contains 2 – 3 % phosphorus in the dry matter; fungi contain 0.4 – 4.5 % [69]. Consequently, some phosphorus is always eliminated from wastewater together with the excess sludge. However, conventionally operated municipal sewage-treatment plants remove only about one-fourth of the phosphorus present in wastewater.

In addition to normal phosphate uptake, which is a function of growth, microorganisms assimilate phosphate beyond their actual requirements. Phosphates are stored by bacteria, fungi, and algae [31] in the form of polyphosphate [57], and by plants also as *phytin*, a calcium – magnesium salt of inosite hexaphosphoric acid [70]. Polyphosphates are microscopically visible in bacteria as inclusion bodies, also known as *volutin* or *metachromatic granules*.

A distinction is made between two types of polyphosphate storage: "overplus accumulation" and "luxury uptake." In overplus accumulation, restoration of the phosphate supply after a period of phosphate deficiency causes the microorganisms to take up more phosphate than they immediately require, but the excess is soon reduced in the course of further growth. In luxury phosphate uptake, especially if growth is limited by a nitrogen or sulfur deficiency, cells take up more phosphate than is required for growth even without the need for a prior phosphate deficiency. The resulting polyphosphates function as a source of phosphate and, under anaerobic conditions, also of energy. During growth, polyphosphate reserves normally accumulate. Polyphosphate storage of this type is found with such facultative-anaerobic genera of bacteria as *Acinetobacter* and *Moraxella*.

To increase the stock of polyphosphate-storing bacteria in a wastewater-treatment plant the activated sludge is exposed alternately to anaerobic and aerobic conditions. Polyphosphate-storing bacteria cannot grow under anaerobic conditions, but they are able to take up organic acids or alcohols, which they can metabolize and store [e.g., as poly(β-hydroxybutyric acid)]. The source of energy in this case is polyphosphate, which undergoes hydrolysis [49]. Phosphate is not retained in the cell, but is released into the wastewater. This process is called *phosphate redissolution*. The level of redissolution depends largely on the extent to which reserve substances are formed in the anaerobic phase. The anaerobic phase normally lasts 1 – 3 h.

In a subsequent aerobic phase the bacteria form polyphosphate, thereby withdrawing phosphate from the wastewater. In this aerobic phase the reserves laid down anaerobically serve as a source of both energy and carbon for accomplishing phosphate uptake. Thus, bacteria of this type grow more rapidly than aerobic bacteria, which are dependent on an uptake of organic compounds in the aerobic phase. The

more anaerobic redissolution of phosphate, the greater is the aerobic uptake of phosphate. Aerobic phosphate uptake always exceeds anaerobic phosphate redissolution, which means that a net elimination of phosphate is achieved [71], [72].

Nutrients. Optimal degradation is achieved in the treatment of wastewater only if in addition to usable carbon compounds the other necessary nutrients are available as well. Domestic sewage normally contains a balanced supply of nutrients. Industrial wastewater is clarified only after compensation for deficient substances or – better still – the addition of nutrient-rich domestic wastewater. An attempt should be made to reach an optimal BOD_5: N: P ratio of 100: 5: 1, and a C: N: P ratio of 100: 12: 3. Summaries suggesting somewhat different numerical values are presented in [44], [69].

Nitrogen. Nitrate, nitrite, ammonium, amino acids, and other metabolites can be utilized by most types of bacteria, even though group-specific preferences do exist. Some industrial wastewater-treatment plants employ readily degradable nitro compounds, aniline, or organic ammonium derivatives as sources of nitrogen. Denitrifying bacteria are able to accept nitrous oxide. Some bacteria reduce atmospheric nitrogen to ammonia. As in the case of nitrate assimilation, nitrogen compounds are taken up and metabolized to ammonia, which is incorporated first into amino acids, and then into a broad spectrum of nitrogen-containing organic metabolic products. In addition to the assimilative uptake of nitrogen, the dissimilative pathway is also significant (see also Denitrification).

Phosphorus. Phosphate is an essential part of the anabolism and catabolism of all organisms. Chemical species that can serve as sources include phosphate, oligophosphates, and simple phosphoric acid esters subsequent to enzymatic hydrolysis. In wastewater-treatment plants containing adapted organisms a number of fairly stable phosphorus compounds can be used as well.

Sulfur. All organisms require sulfur-containing amino acids like methionine and cysteine for protein synthesis, together with other sulfur compounds in small amounts [47]. Many bacteria, fungi, and plants take up sulfur in the form of sulfate, less frequently as sulfide, and synthesize amino acids on their own. Numerous animals are dependent on food sources supplying both amino acids and sulfur-containing vitamins. Bacteria dissimilatively utilize sulfate as an electron acceptor primarily when dissolved oxygen and nitrate supplies in the vicinity have been exhausted. Conversely, purple sulfur bacteria (among others) employ molecular sulfur and sulfides as electron donors [23].

Other Macronutrients. Ions such as magnesium, potassium, and calcium stabilize the three-dimensional structures of proteins and play a role in the regulation of osmotic properties. Magnesium is the central atom of the chlorophyll molecule, an important component for photosynthesis. Iron and other metals are found in the cytochromes of the respiratory chain, and in the oxidases of various degradation pathways. Animals also require substantial amounts of sodium, chloride, and fluoride ions. All of these substances are usually available in a wastewater-treatment plant at adequate concentrations. Iron and aluminum salts are frequently added to wastewater to induce the chemical precipitation of phosphate.

Trace Elements. Small amounts of several other elements are required for a number of purposes and by several enzymes. Examples include: boron (complexation of polyhydroxy compounds?), vanadium (nitrogen fixation), manganese (photosystem II), cobalt (ribonucleotide reductase, nitrogen fixation), nickel (methyl coenzyme M reductase), copper (cytochrome oxidase, phenol oxidase), zinc (dehydrogenases), and molybdenum (xanthine oxidase, nitrate reductase, nitrogenase) [17], [48]. These elements are sometimes required only by a subset of the organisms (e.g., boron, vanadium, and nickel), or by certain species in a bound organic state in the form of a vitamin (e.g., cobalt). An insufficient supply of trace elements is less common in wastewater-treatment plant than levels exceeding the toxic limiting concentrations.

Toxicity. "Poisonous character" and toxicity refer in a general way to effects experienced by organisms in the presence of specific chemical substances. Various toxic materials can be present in wastewater [73], resulting in undesirable consequences with respect to wastewater treatment. These include:

1) Inhibition in terms of individual enzyme activities

2) Inhibition of specific metabolic pathways
3) Inhibition of growth (a *microbiostatic* effect) – of nitrifying bacteria, for example
4) Death of certain organisms (a *microbiocidal* effect)
5) Selective pressures favoring undesirable organisms as manifested in sludge bulking

Representative reported data with respect to threshold concentrations for detectable damage to activated sludge are as follows [44]:

Substance	Inhibiting concentration (mg/L)
Cadmium (Cd^{2+})	2 – 5
Chromium (CrO_4^{2-})	3 – 10
Copper (Cu^{2+})	1 – 5
Nickel (Ni^{2+})	2 – 10
Zinc (Zn^{2+})	5 – 20
Chlorine (Cl_2)	0.2 – 1
Cyanide (CN^-)	0.3 – 2
Mineral oils	>25
Phenols	200 – 1000
Hydrogen sulfide (HS^-)	5 – 30

The potential effect, and thus the toxicity, of a substance within a wastewater-treatment plant is not a predetermined constant, but is a function instead of the conditions of exposure and the organisms present (Table 2).

There is no such thing as toxicity per se. The term toxicity really refers to an interaction between some substance and a specific organism, which means it is subject to environmental influences. The substance itself merely represents the potential for a reaction.

The amount of a substance taken up by a particular organism is defined as the dose. Under stipulated exposure conditions it is sometimes more convenient to report a concentration in the case of gaseous or liquid media rather than a dose. Not the presence of a substance but the dose taken up of its biologically active form determines the effects. Many substances have never been shown to have any favorable effects whatsoever on organisms. However, in the case of heavy metals, several elements are now known to be essential at trace levels, despite the fact that they are toxic in high doses. The simultaneous action of several substances may lead to an alteration of effects (*coergism*), in which the resulting effect is either larger (*synergism*) or smaller (*antagonism*) than the sum of the expected effects of the individual compounds [15], [75], [76].

The physical distribution and chemical bonding state of a substance in the environment determines its rate of uptake and metabolism by an organism; this is known as its *bioavailability*. Poorly soluble (precipitation) products, substances adsorbed tightly on minerals, and other inert materials usually show no significant toxic effects.

The way in which a wastewater load is fed into the system is often decisive with respect to the behavior of a wastewater-treatment plant. A distinction can be made between two important feeding patterns [25]:

1) Irregular introduction of slugs of highly contaminated wastewater, usually a consequence of some disturbance of industrial batch production or, in an agricultural region, of seasonal production.
2) Continuous feeding of material residues in low concentrations, for example, following some form of wastewater pretreatment.

In the case of irregular feeding, even readily degradable, soluble organic compounds may be responsible for increased effluent values or sludge bulking. With continuous feeding of low concentrations of "toxics," the inhibitory effect soon diminishes due to adaptation, leading in fact to the growth of microorganisms displaying increased resistance and a higher degradation potential.

Degradation of Xenobiotics. The term xenobiotic is derived from the Greek words *xenos* (foreign) and *bios* (life), implying that xenobiotics are not formed biologically, but instead arise through industrial processes [14]. More and more metabolic pathways and organisms have been found to synthesize typical xenobiotics, especially organohalogen compounds [15], [77], [78]. Based on the assumption of an exclusively abiotic origin it is frequently deduced that xenobiotics cannot be biodegraded; nonetheless, the biodegradation of several such "xenobiotics" will be described below. Among all the substances still referred to as xenobiotic, the term as defined above probably applies only to a small subset of compounds. Used in this loose way it serves primarily as evidence of human ignorance. The term is employed in this contribution only to indicate that the substances

Table 2. Factors influencing effects of a substance (adapted from [74])

Substance	Environment	Organisms		
		Cells	Process	Effect
Concentration	pH	Metabolism	Adsorption	Promotion
Distribution	Temperature	Degradation potential	Accumulation	Neutrality
Reaction potential	Oxygen	Sensitivity	Enzyme regulation	Inhibition
Chemical, physical behavior	Nutrients	Resistence	Enzyme induction	Death
	Sludge concentration	Multiplication	Adaptation	Tolerance
Exposure to other substances	Sludge age	Adaptation capacity		
	Chemical load			

Table 3. Aerobic biodegradability of organic compounds (adapted from [49])

Readily biodegradable [a]	Moderately biodegradable [b]	Slightly biodegradable under studied conditions [c]	Refractory under studied conditions [d]
Cyclohexane	1-Decanol	Decane	Dodecane
Octane	1-Dodecanol	1,3-Dichloropropane	Dichloromethane
Phenol	Acetone	Ethyl ether	Chloroform
Methanol	Ethylbenzene	Phenanthrene	1-Chloropropane
1-Propanol	2-Furaldehyde		1-Chlorobutane
1-Butanol	Benzonitrile		1-Chloropentane
1-Pentanol	4-Bromophenol		1-Chlorohexane
1-Hexanol	Hydroquinone		1-Chlorodecane
1-Octanol			1,2-Dichloroethylene
2-Chloropropionic acid			Isopropyl ether
Diethanol amine			Trichloroacetic acid
Acetonitrile			Chlorobenzene
Acrylonitrile			1,3-Dichlorobenzene
Benzene			1,4-Dichlorobenzene
Toluene			1,2,4-Trichlorobenzene
Xylene			1,2,4,5-Tetrachlorobenzene
Benzyl alcohol			Hexachlorobenzene
Nitrobenzene			Benzidine
Naphthalene			
Pyridine			
Quinoline			
m-Cresol			
p-Cresol			
4-Chlorophenol			
4-Nitrophenol			

[a] Biochemical oxygen demand (BOD)/theoretical biochemical oxygen demand (TOD), >50 %.
[b] BOD/TOD, 25 – 50 %.
[c] BOD/TOD, 10 – 25 %.
[d] BOD/TOD, <10 %.

in question are not derived from the same organisms that degrade them. In this sense, nutrients and vitamins are examples of xenobiotics for organisms that require them [15].

Like toxicity, biodegradability is also not a function exclusively of the substance itself, but is instead determined by a combination of organisms and abiotic environmental factors. In the presence of ordinary activated sludge, for instance, cellulose is not microbiologically degradable under normal conditions [33]. Some of the factors upon which degradation depends are:

1) The organisms, their arsenal of enzymes [79], and their adaptation potential [80].
2) The physiological state of the organisms and their supply of nutrients, trace elements, and vitamins.
3) The growth status of the organisms.

4) Environmental parameters such as pH, salt content, supply of electron acceptors, and temperature.
5) Substance concentration, since the rate of degradation exhibits an optimum value at a particular concentration of substrate; at very low concentration, for example, enzyme induction does not occur [81], whereas high concentrations result in toxicity [82].
6) The presence of other organic compounds.

If several substances are subject to degradation simultaneously, the observed degradation rates are usually not identical [33]. In fact, degradation of diverse but similar substrates seldom occurs simultaneously, but is most often successive [83] because of the ecological principle that energy and raw-material expenditure is minimized. This means that only one set of enzymes is synthesized rather than two. Consequently, processing of the more easily degradable substance initially inhibits the degradation of other substances (*diauxie*) [33]. On the other hand, in the case of cometabolism a poorly degradable material is degraded alongside a suitable, more easily degradable substance.

To determine a level of biodegradability, the substance to be examined (the test substance) is added to cultures of specific microorganisms, activated sludges, or other biocenotic systems. In a *direct* degradation test, disappearance of the test substance is measured after a period ranging from a few days to months. Tests of this type are also called "die-away tests." In *indirect* tests one measures the consumption of a substrate like oxygen, or the formation of some degradation product like carbon dioxide, hydrochloric acid, or ammonia rather than the test substance itself. The *biological oxygen demand* (BOD) is often determined as a wastewater quality criterion that is not substrate specific. The BOD is the mass of oxygen required for biological oxidation of the organic wastewater components that can be degraded by microorganisms in standardized wastewater [15], [84–87]. Indirect tests have the disadvantage that many substances are consumed also through anabolism in the course of forming endogenous substance, resulting in data reflecting an unrealistically low degree of degradation. If the test substance is the only usable carbon source present, its biodegradability can be established from the increase in the number of microorganisms. Biodegradability can also be predicted on the basis of structural characteristics and water solubility [88].

A few examples of substances degradable in wastewater-treatment plants are presented in Table 3 [49]; for further examples see [89], [90].

Enzymes that incorporate oxygen or hydroxyl groups into organic compounds play an important role in aerobic degradation. Depending upon the reactions catalyzed, the following distinction can be made [where X = xenobiotic, D = (other) hydrogen donor, A = (other) hydrogen acceptor] [15], [91–93]:

Oxygenases (dioxygenases)

$$X + DH_2 + O_2 \rightarrow X(OH)_2 + D$$

$$XH_2 + O_2 \rightarrow X(OH)_2$$

Monoxygenases (hydroxylases)

$$XH_2 + DH_2 + O_2 \rightarrow XHOH + D + H_2O$$

Dehydrogenases

$$XH_2 + A + H_2O \rightarrow XHOH + AH_2$$

Lyases

$$X + H_2O \rightarrow XHOH$$

Some xenobiotics can be channeled into primary metabolic pathways after a few preliminary reaction steps; for example, alkanes are first oxidized to alcohols and then processed further by β-oxidation (Fig. 9). Cleavage of appropriate substituents often permits carbon compounds bearing functional groups or substituents to be transformed into substances that can be dealt with by primary metabolism. For this reason, the discussion here will focus on the elimination of selected substituents rather than complete degradation pathways.

Aromatic Compounds. The key substance for the microbial degradation of aromatic compounds is catechol, which is in turn subject to catabolism and anabolism after an initial ring cleavage (Fig. 10). Several of the enzymes responsible for oxidizing aromatic compounds

show little specificity. These polyfunctional enzymes are able to convert various aromatics to catechol [94], its derivatives [95–97], or its precursors [98]. The enzyme-producing organism is thus able to metabolize a broad spectrum of aromatic compounds with little effort, which is significant, because aromatic compounds often occur in groups.

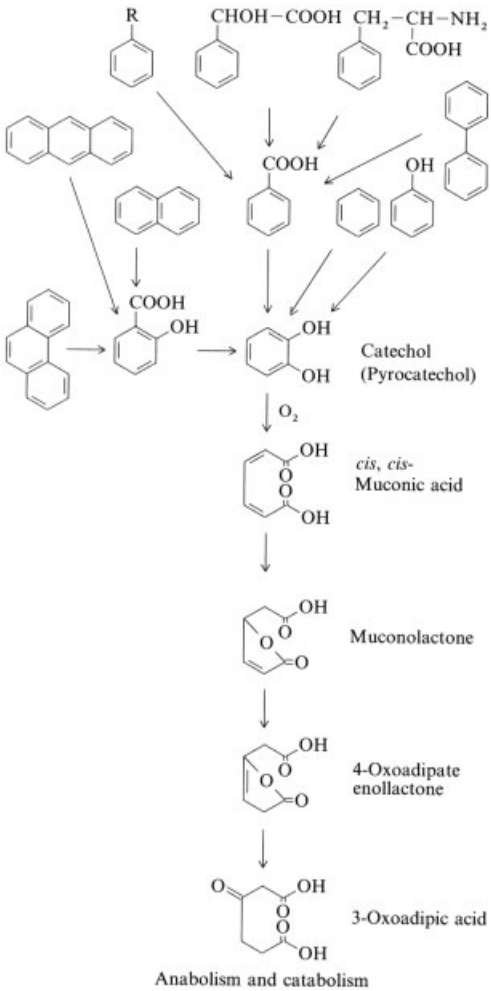

Figure 10. Microbial degradation of aromatic compounds [94]

Resonance, inductive effects, and steric effects all play a significant role in the degradation of aromatic compounds. The pH is important especially in the case of derivatives of aniline and benzoic acid. The rate of degradation of substituted aromatic compounds increases mostly in the presence of hydroxyl, carboxyl, carbonyl, and, often, alkyl substituents. It usually decreases in the presence of nitro, amino, chloro, and sulfonyl substituents. As far as disubstituted benzene derivatives are concerned, *para*- and *ortho*-substituted systems are usually degraded most readily.

Organic Halogen Compounds. Halogen removal (dehalogenation) is catalyzed by various enzymes, including oxygen-dependent dioxygenases and monoxygenases as well as oxygen-independent dehalogenases (reductases) and hydrolases [15], [55], [99]. Reductive dehalogenation, especially of the type occurring in anaerobic media, permits organohalogen compounds to replace oxygen as the electron acceptor. Various dioxygenases that degrade aromatic organohalides first attack the aromatic ring, usually at neighboring unsubstituted positions, which are then hydroxylated (Fig. 11) [99–101]. The biodegradation of polychlorinated dibenzo-*p*-dioxins and other organohalides can be catalyzed by lignin-degrading enzymes, where radicals formed from hydrogen peroxide serve as the oxidizing agent [102–104].

Nitro Compounds. Two degradation pathways are important here: reduction of the nitro group and direct cleavage of the C – N bond.

In the case of reduction, amino groups remain in the product, and their further degradation depends on the availability of hydrogen at the α-carbon. The presence of such hydrogen facilitates oxidative formation of an imino group, which is in turn hydrolyzed through the addition of water and elimination of ammonia. Reactions of this type also occur in amino acid metabolism. Many aromatic nitro compounds are reduced to aniline derivatives [105]. Figure 12 shows an example of the degradation of an aromatic nitro compound; the reduction of 4-nitrotoluene, which proceeds via nitroso and hydroxylamino intermediates [106].

Monoxygenases and dioxygenases catalyze direct removal of the nitro group in the form of nitrite [107–109]. Lignin-degrading enzyme systems similarly eliminate nitro groups as nitrite [110]. Examples of the degradation of industrially important nitroaromatic compounds are provided in [111].

Aromatic Sulfonic Acids. Dioxygenases cleave aromatic sulfonic acids to produce sulfite, simultaneously forming a catechol derivative (Fig. 13) [112–114]. Various bacteria and

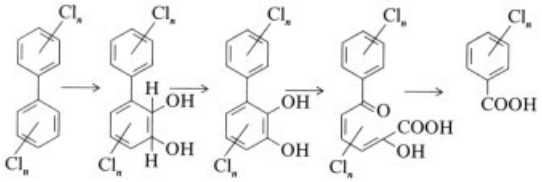

Figure 11. Proposed pathway for the degradation of polychlorinated biphenyls ($n = 1-4$) [100]

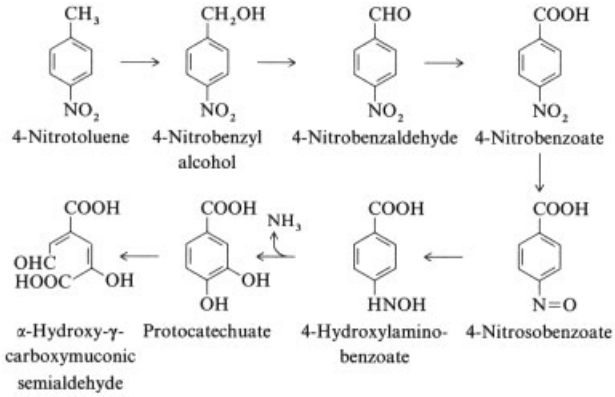

Figure 12. Proposed pathway for the degradation of 4-nitrotoluene (4-NT) by *Pseudomonas* sp. strain 4 NT [106]

green algae accept naphthalenesulfonic acids as sources of sulfur [107]. Toluenesulfonic acid is also degraded by aquatic plants [115].

Biological Elimination of Heavy Metals. Heavy metals may be present in wastewater in both the dissolved and undissolved state. In the dissolved state they form complexes with such species as water, carbonate, cyanide, and chloride, or with chelating agents like ethylenediamine-tetraacetic acid (EDTA), nitrilotriacetic acid (NTA), and citric acid. In undissolved form they are adsorbed onto particles, precipitated as salts, or included within other precipitation products. Suspended heavy metals are frequently bound to humic substances whose particle size and surface charge prevent their sedimentation. Several heavy metal complexes are always present simultaneously. These complexes are subject to interconversion, and are influenced by pH, temperature, redox potential, and the presence of other substances. In wastewater, precipitation, adsorption, and complex formation all occur in parallel with dissolution, desorption, and coordination change.

The biological elimination of heavy metals has been attributed to several different mechanisms [101]:

Sorption. Cell walls, slimes, and capsules associated with living organisms, together with their degradation products, macromolecules like proteins and polysaccharides, and humic substances all act as adsorbents, especially as a consequence of their polar groups. Carboxyl, sulfonic acid, amino, and phenolic groups act as ion exchangers that bind heavy-metal ions as a function of the pH and neutral salt concentrations, among other factors. Living cells adsorb various heavy metals and accumulate them. Large amounts of organic substances are bound to activated sludge by sorption [116–119].

Precipitation. Sulfide is formed in anaerobic zones of biofilms, leading to poorly soluble precipitates. Nevertheless, heavy metals are able to reach these sulfidic zones only slowly. More soluble precipitation products are formed in aerobic layers or activated sludges, but these precipitates are easily included in the growing biomass.

Complexation. Heavy metals existing as stable complexes are neither biologically available nor the cause of toxic reactions. Many microor-

ganisms, especially the heavy-metal-resistant organisms, synthesize and store complexing agents such as the iron-containing protein ferritin and the siderochromes (iron trihydroxamates) like ferrichrome [75], [120].

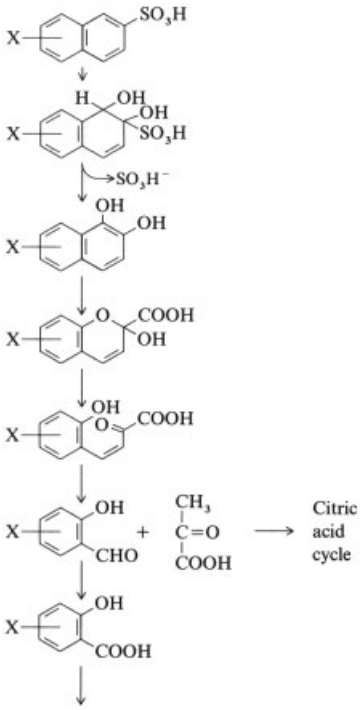

Figure 13. Degradation of 5-, 6-, 7-, or 8-substituted aminonaphthalene- or hydroxynaphthalenesulfonic acids by *Pseudomonas* sp. strain BN 6, and the assimilation of salicylates by other bacteria harboring complementary catabolic sequences [113]
X = $-NH_2$ or $-OH$

Methylation. Methylcorrinoid derivatives, better known as vitamin B_{12} derivatives, are capable of transferring methyl groups to such heavy-metal ions as mercury(II) [75]. These reactions would lead to the formation of toxic compounds in a wastewater-treatment plant, some of which would contaminate the off-gases. A number of reactions and metabolic pathways have been postulated for the biological methylation of heavy metals, but their relevance is still disputed [121].

Other Reactions. Microorganisms can metabolize and accumulate selenium in place of sulfur, arsenic in place of phosphorus [75], and cadmium in place of calcium. The reduction of vanadium, selenium, silver, and other heavy-metal ions is regarded as a mechanism by which these ions are eliminated from wastewater [122]. Microorganisms also take up heavy metals as trace elements necessary for life, incorporating them into coenzymes and other compounds [48].

The great importance of biological heavy-metal elimination is apparent from the brownish-black biofilms found in siderotrophic waters poor in nutrients, in iron drinking-water pipes, and in acidic wastewater from the mining industry. These films are due mainly to bacteria that oxidize iron(II) and, less frequently, manganese(II). Some "sham" iron bacteria have been found to live chemoorganotrophically, storing iron hydroxide in their capsules. True iron bacteria like *Thiobacillus ferrooxidans, Sulfolobus acidocaldarius,* and *Gallionella ferruginea* thrive chemolithotrophically [19], [123].

2.1.2. Basic Biotechnological Considerations

The early 1950s witnessed an economic upswing that led to increased production and the creation of more jobs, but on the other hand also resulted in increasing loads of industrial wastewater. Growing prosperity also meant greater public demand for hygiene. Sewage systems were expanded in many municipalities, but often without corresponding increases in wastewater-treatment capacity. All these factors resulted in more pollution of surface waters, and the resulting self-purification phenomena altered the natural oxygen balance. With the ensuing oxygen deficit came a deterioration in water quality, a decrease in the variety of aquatic organisms, and the death of many fish.

As a consequence, political pressure grew for the prevention of water pollution. Concerted efforts were initiated to somehow eliminate those biodegradable substances that led to a depletion of dissolved oxygen in surface waters.

Aerobic biological wastewater treatment is now the method of choice for elimination of these materials, whose levels are usually measured in terms of the biological oxygen demand (BOD) [87]. This is a long-established approach for the treatment of municipal sewage, but adaptation was required before it could be applied to industrial wastewater, especially wastewater from the chemical industry. This in turn dictated

that special demands be imposed regarding the wastewater itself.

Demands Applicable to Wastewater Quality. Intensive study of the problem revealed that a specific wastewater regime would be required from both the technical and ecological points of view. This led on one hand to separation of the wastewater into one waste stream requiring special treatment and another for which this was unnecessary, as well as separate transport systems. Qualitative considerations argued in favor of decentralized pretreatment based on physicochemical processes (discussed in detail below) as a way of eliminating those substances that disturb or inhibit biodegradation [7], [11], as well as centralized chemical and mechanical pretreatment within the wastewater-treatment plant in order to neutralize and eliminate undissolved substances.

Decentralized wastewater treatment facilitates the specific removal of substances that inhibit or disturb biodegradation based either on precipitation and subsequent separation or detoxification by oxidative treatment.

Over the course of time, increased knowledge has permitted requirements related to water-pollution control to become ever more strict. Oxygen balance is now not the only goal of water protection measures. Requirements have also become necessary to meet the demand for drinking water, mandating the virtual exclusion of organic substances – even substances subject to slow biodegradation. Consequently, the key parameter was no longer the BOD, but rather the chemical oxygen demand (COD) [124]. In addition, ecological concerns required the elimination of nitrogen and phosphorus compounds that tend to encourage eutrophication.

Thus, the overall demands currently imposed on wastewater quality are quite stringent. Substances that undergo slow biodegradation must now be removed in decentralized facilities engaging in specific pretreatment measures, or else converted into other, more readily biodegradable substances by means such as partial oxidation. Furthermore, it became necessary to introduce decentralized pretreatment processes directed toward the specific elimination of nitrogen and phosphorus, or else modify the biological wastewater-treatment stage in such a way as to ensure further removal of nitrogen (nitrification [125], denitrification [126]) and phosphorus (biological elimination [127] or simultaneous precipitation [52]), which entailed enormous developmental effort.

Demands Related to Microorganisms. The prerequisite for successful biological wastewater treatment is the presence of suitable microorganisms capable, perhaps after adaptation, of providing metabolic pathways for the degradation of various substances, and suitable for safe establishment at adequate levels in the biocenotic system of a biological wastewater-treatment plant (mixed sludge). Growth kinetics and mechanisms of flocculation and adsorption play a considerable role, and it is crucial that these organisms not be displaced by others in the necessarily nonsterile conditions characterizing a wastewater-treatment plant.

Hence, the biotechnological processes for aerobic wastewater treatment are based on the ability of bacteria on one hand to adapt to diverse wastewater components as their source of nutrients, but also to undergo aggregation (floc formation) in times of low nutrient supply [7].

Consider the way a biological system develops with time in the course of a batch experiment (discontinuous). In what is known as the "lag phase" the bacteria become accustomed to the nutrient substrate (*adaptation*). During a subsequent "log phase" the number of bacteria increases rapidly, and the nutrient substrate is quickly depleted. In this phase the bacteria are present as free-swimming entities, because there is still a sufficient amount of substrate available. Only when the substrate has largely been consumed will the bacteria tend to agglomerate.

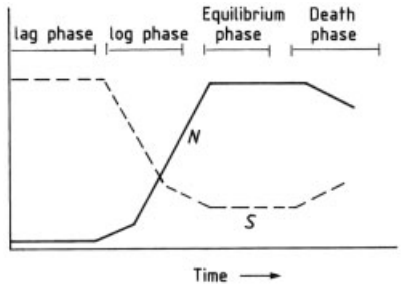

Figure 14. Developmental patterns for the number of bacteria N and the substrate concentration S over the course of a batch (discontinuous) experiment

Biological wastewater-treatment plants are operated near the equilibrium-phase/death-phase boundary. This facilitates retention of the microorganisms while maintaining the lowest possible outlet concentration of the substrate, consistent with substrate degradation being the primary goal of wastewater treatment (Fig. 14).

Demands Made in Terms of Process Engineering. With respect to process engineering, three requirements must be met in taking advantage of the aerobic metabolism of microorganisms for biological wastewater treatment in such a way as to produce the highest possible space – time yield (i.e., treatment in the smallest possible volume over the shortest possible time):

1) The number of suitable bacteria must be as high as possible inside the available volume; i.e., the bacteria must be specifically retained by either internal or external means, depending on the process.
2) The bacteria must be provided with a sufficient supply of dissolved oxygen (i.e., there must be adequate aeration), where oxygen diffusion is enhanced by maximizing the extent of the gas – water interface.
3) The three reaction partners – bacteria, organic substrate, and dissolved oxygen – must be brought together, which requires carefully engineered turbulence.

These boundary conditions are not completely independent, and they have provided the impetus for various biotechnological approaches to the aerobic treatment of wastewater, including the trickling filter and the activated sludge-process.

Bacterial Retention. For economic reasons, efforts must be made to achieve the highest possible space – time yield in any industrial operation. Therefore, one of several process-engineering requirements associated with a biological wastewater-treatment plant is the presence of the highest possible number of active microorganisms within the reaction volume. Normally, bacterial growth in the course of the biological degradation of wastewater components is insufficient to guarantee a high concentration of bacteria in the bioreactor. Most biological wastewater-treatment plants are operated in a continuous mode, and thus constitute dynamic systems. Bacteria are constantly being discharged along with the purified effluent water. Bacterial growth within the reactor will compensate for loss of bacteria in the outflow only in the special case of a wastewater stream very heavily loaded with rapidly biodegradable substances. However, a system of this type would be in the log phase with respect to bacterial growth; i.e., the bacteria would be free-swimming and not separable, and degradation of the substrate would be incomplete. Such a fermenter can therefore be used only for biological pretreatment [128].

For this reason, all wastewater-treatment plants are instead operated with deliberate bacterial retention and bacterial separation. This is necessary as well for assessment of the effluent and efficiency of the plant from the standpoint of discharge licensing requirements. Bacteria in the effluent are monitored analytically in terms of substance removable by filtration, COD, and nitrogen and phosphorus content.

Bacterial retention and separation are both positively influenced by the growth and agglomerating power of the microorganisms in the equilibrium/death phase. Bacterial retention also results in the bacteria remaining in contact with the wastewater and its components over long periods of time, permitting the organisms to adapt accordingly.

Internal Bacterial Retention. If the bacteria are presented with the largest possible appropriate surfaces for growth within the bioreactor, then growth on these surfaces will in fact occur, leading to biofilms. These surfaces may be permanently installed as fixed or movable arrays, or they may consist of free-floating objects that are confined within the bioreactor in some convenient way. Processes based on this principle include the trickling filter, the intermittently submerged reactor, the biofilter, the loop reactor, and the Linpor device. Especially suitable growth surfaces are porous and rough, as for example lava slag, certain plastics (PVC, PU), sand, coke, and expanded clay [129]. Plant-based wastewater-treatment works [122], [130–132] also fall within this category, although they are applicable only for small amounts of wastewater. The biodegradation of organic substances is carried out in this case by microorganisms that colonize the surfaces of the plants.

Among the prerequisites for implementation of such biotechnological process are exclusion of solids from the inflow to the bioreactor, taking advantage, for example, of a prior mechanical purification step and targeted control of the hydraulic and waste loads as a way to prevent clogging due to bacterial growth.

Excess bacterial mass is detached from the growth surfaces by means of a hydraulic shock load, and then isolated from the effluent in the form of excess sludge, often by sedimentation, for further preparation in the sludge-treatment stage (see Fig. 29).

The *sequencing batch process* represents a special case of internal bacterial retention. This can be regarded as a transition process in the direction of external bacterial retention, because it also takes advantage of the floc-building capacity of microorganisms, which agglomerate into flocs in the equilibrium phase. The sequencing batch reactor (SBR) is operated discontinuously, as suggested by its name. Mechanically pretreated waste-water is introduced into a prepared bacterial suspension. After the completion of biological elimination processes all mixing is discontinued, providing the suspended flocculated bacterial mass with sufficient time to sediment. This is followed by phase separation within the reactor and subsequent withdrawal and redirection of the clear phase. Processing is resumed with renewed introduction of wastewater into the remaining bacterial suspension.

External Bacterial Retention. With an aerobic biological process in the absence of specific bacterial retention in the reactor, a continuously operated system soon becomes depleted of bacterial mass, because more bacteria are ordinarily discharged in the effluent than are formed through biological degradation. To prevent this negative effect, the process is divided into two steps. Biological degradation – the first process step – is followed by a second step: solid – liquid separation of the effluent on the basis of sedimentation, flotation, or filtration. This approach is feasible because the bioreactor operates in its equilibrium phase, so the bacteria aggregate into flocs. The externally retained bacterial mass is returned as required to the bioreactor in the form of recycled sludge (Q_{RS}). Excess bacterial mass is withdrawn as surplus sludge, which must then be appropriately prepared in the sludge-treatment stage (Fig. 15).

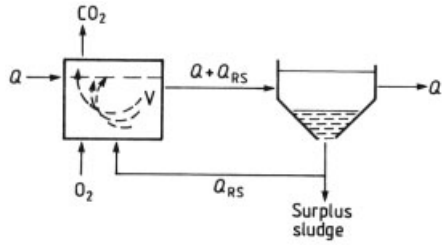

Figure 15. The activated-sludge process Q = wastewater; Q_{RS} = recycled sludge

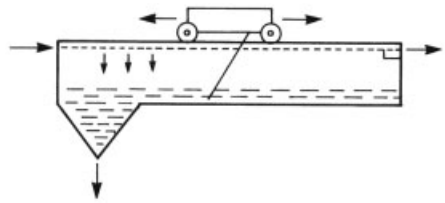

Figure 16. Horizontal-flow post-clarification conducted in a rectangular basin

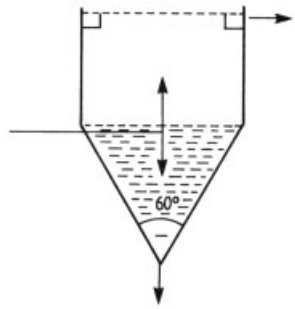

Figure 17. Vertical-flow post-clarification based on a Dortmund tank

Sedimentation. The sedimentation process takes advantage of a difference in density between the activated-sludge floc and purified wastewater (see also Section 5.2.2.1). The sedimentation rate is a function not only of difference in density, but also of the particle size and the viscosity of the wastewater (Stokes law).

Sedimentation may be allowed to proceed in flow aggregates arranged either horizontally (rectangular or circular sedimentation basins) or vertically (Dortmund tanks).

With a horizontal-flow clarifier the sedimented clarification sludge is compacted on the bottom by sludge collectors and either returned to the activation step or withdrawn as excess

sludge (Fig. 16). In a Dortmund tank, one example of a vertical-flow clarifying system, mechanical collection is unnecessary, because the geometric design of the cone (angle ca. 60°) permits the activated sludge suspension to flow by itself (Fig. 17).

The quality of the activated sludge is a decisive factor with respect to its possible concentration. This quality can be assessed on the basis of the *sludge volume index* (reported in cm^3/g after a settling period of 30 min) [133]. Flocculation and sedimentation characteristics can sometimes be improved by the addition of flocculation aids as necessary.

The potential bacterial concentration in an aeration tank is a function of the concentration process. If the activated sludge can be concentrated to a value of x (g/L), and the reflux ratio is 100 % [return flow (Q_{RS}) = inflow (Q)], an activated sludge concentration equal to $0.5\,x$ (g/L) plus the bacterial mass produced in the course of biological degradation is established in the bioreactor. Consequently, it is the sedimentation and concentration properties of the activated sludge that determine the space – time yield of the process. For this reason sedimentation is applicable only to activated sludges with low sludge-volume indices (~50 – 200 cm^3/g). With higher sludge indices (bulking sludge, threadlike growth [41], [42]) or to meet special requirements, flotation or filtration is used instead.

Flotation [134–136] (see also Sections 3.1 and 5.2.2.3) takes advantage of the tendency for substances suspended in water to adsorb gas bubbles at their surfaces. The suspended substances thereby become lighter than water, float, accumulate at the water surface, and can be skimmed off.

There are two prerequisites that must be met for successful flotation:

1) Adequate adhesion of gas bubbles to the surface of the suspended substances, encouraged as required by the addition of flotation aids.
2) Creation of the finest possible gas bubbles in the water by aeration flotation (Königsbacher [137]), dissolved-air flotation (Bitburger 2.134]), or pressurized dissolved-air flotation.

Advantages of activated-sludge separation by flotation relative to sedimentation are discussed in the literature, and include higher concentration of the activated sludge as well as reduced sludge residence time in the second process step, solid – liquid separation, which helps to prevent anoxic conditions.

Filtration. Ultrafiltration represents a special case of sludge retention downstream from a bioreactor [139]. Especially high concentrations of activated sludge can be achieved with this approach. When combined with a closed biological system this method is also applicable to the degradation of volatile materials (e.g., chlorinated substances). However, the effective range of application of biology coupled with ultrafiltration so far limits the approach to relatively small wastewater streams.

Provision of Oxygen. In order for aerobic biological degradation to be successful the bacteria in the bioreactor must have access to an adequate supply of oxygen. This normally entails oxygen dissolved in the water (DO), but in special cases, oxygen-containing materials such as nitrate (denitrification) or hydrogen peroxide may serve the same purpose [140].

Effective mass transport through the interfaces gas – water and water – bacterium is crucial for supplying the microorganisms with dissolved oxygen. Oxygen for conversion by bacterial metabolism must first be transported through the gas – water interface and dissolved in the water before it becomes available to the bacteria. This mass transport is in turn a diffusion-controlled process (Fig. 18).

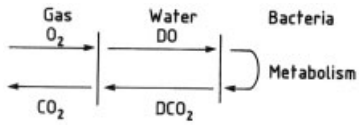

Figure 18. Material interchange at the gas – water and water – bacteria interfaces
DO = dissolved oxygen; DCO_2 = dissolved carbon dioxide

The solubility of oxygen in water depends on both the temperature and the oxygen partial pressure of the gas. According to Henry's law:

$$p = K\,c$$

where p = partial pressure, K = a temperature-dependent constant, and c = concentration in the liquid.

Dissolved-oxygen values for air (O_2 = 20.8 vol %) at 1 bar in salt-free water are as follows [141]:

°C	Dissolved oxygen (mg/L)
5	12.37
10	10.92
15	9.76
20	8.84
25	8.11
30	7.53

Oxygen transport through the gas – water interface depends on the extent of the available interface and on the concentration gradient required for diffusion. It is described by the expression [136]:

$$G = k_L A \Delta c$$

where G = mass flow through the phase boundary (e.g., kg of O_2/h), k_L = oxygen transfer coefficient on the liquid side (e.g., m/h), A = area of the phase interface (e.g., m^2), Δc = concentration gradient (e.g., g of O_2/m^3) = $C_S - C_O$ (in the simplest case), with C_S = saturation concentration and C_O = current system concentration.

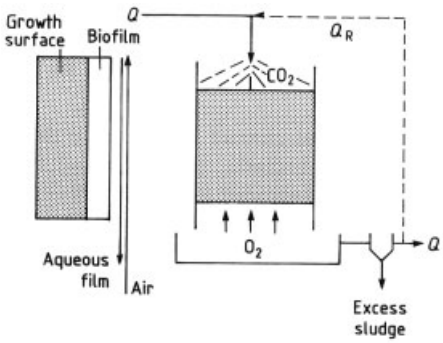

Figure 19. The trickling-filter process
Q = wastewater; Q_R = recycle water

Oxygen input normally occurs directly within the bioreactor. The requisite exchange surface between the gas – usually air; in exceptional cases technical-grade oxygen – and the wastewater is produced deliberately in the presence of the microorganisms. With a trickling filter (a type of packed column), wastewater is sprayed in from the top onto a set of internal fittings (growth surfaces). It then flows downward in a thin layer over the previously developed biofilm, permitting oxygen exchange with gas that is flowing countercurrently (Fig. 19).

In the intermittently submerged reactor process – utilizing, for example, round discs mounted on a rotatable axis – material exchange occurs as a result of alternating exposure of the growth surface to wastewater and air as the biofilm is subjected to continuous rotation (Fig. 20).

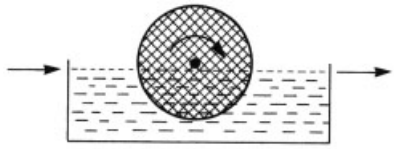

Figure 20. The intermittently submerged reactor process

In the activated-sludge process, development of the phase boundary required for adequate mass transport is facilitated either by surface aeration or volume aeration [142], [143]. In the case of surface aeration the wastewater is sprayed in the form of fine droplets by means of aggregates rotating horizontally or vertically at the water surface (Fig. 21). Volume aeration relies instead on fine gas bubbles produced in the aeration zone by an appropriate gas-introduction system. With volume aeration it is necessary to take into account as well the phenomenon of coalescence in the course of oxygen transfer [144]. Depending on the wastewater composition, finely dispersed gas bubbles may be stabilized, but they might also coalesce and immediately form large bubbles. Under coalescence-promoting conditions, aeration with fine bubbles should therefore be avoided in the interest of saving energy.

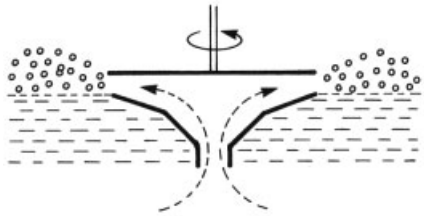

Figure 21. A surface-aeration facility

Only in a few exceptional cases is oxygen transfer accomplished outside the bioreactor in a separate aeration chamber. This approach is applicable to biofilters designed for upward flow

and operated either as fixed or fluidized beds (e.g., the Oxitron process).

Supplying necessary oxygen to the microorganisms via oxygen-containing compounds presents no particular problems. Nitrate need not be added specifically for this purpose: it is either already present in the wastewater, or is formed by ammonium oxidation (nitrification). Nitrate is deliberately converted to other forms through denitrification in the interest of nitrogen removal.

The use of hydrogen peroxide is of essentially no practical significance. However, if H_2O_2 were to be introduced it would be important to adequately condition the organisms so that they would form enough catalase for enzymatic cleavage of the otherwise bacteriotoxic hydrogen peroxide.

Mixing. Mutual contact between the three reaction partners bacteria, substrate, and dissolved oxygen is established automatically in the course of oxygen transfer, because the process inevitably produces considerable turbulence in the aqueous phase. Only in denitrification, which proceeds with the exclusion of oxygen, is there a need for special mixing of the water – sludge mixture – with the aid of a mixer or through injection of an inert gas (which can be recirculated if necessary).

Carbon, Nitrogen, and Phosphorus Elimination. Biological treatment of wastewater eliminates not only organic substances (measured as BOD, COD, or total organic carbon, TOC), but also nitrogen and phosphorus compounds. The oxidative degradation of organic compounds by heterotrophic organisms causes the organic matter to be mineralized to CO_2, NH_4^+, and PO_4^{3-}, with concurrent formation of bacterial mass.

Two different metabolic processes are involved in nitrogen removal:

1) Nitrification, which occurs with autotrophic organisms under oxic conditions

$$NH_4^+ + 2\,O_2 \rightarrow NO_3^- + H_2O + 2\,H^+$$

2) Denitrification, observed with heterotrophic organisms in the anoxic range in the presence of readily biodegradable organic compounds (e.g., acetic acid, methanol, etc.), assuming there is thorough mixing

$$6\,NO_3^- + 5\,CH_3OH \rightarrow 3\,N_2 + 6\,OH^- + 5\,CO_2 + 7\,H_2O$$

Heterotrophic organisms are responsible for phosphorus elimination, incorporating polyphosphate into their reserves under alternating anoxic – oxic conditions in a process known as "luxury uptake" [52].

Biotechnology makes it possible to combine these various elimination processes in a meaningful way provided appropriate attention is directed to the corresponding diversity in environmental prerequisites and growth characteristics, as illustrated in Table 4 for the case of the activated-sludge process.

Table 4. Essential conditions for bacterial metabolism and growth

Elimination process	Bacteria	Environ. conditions	Growth
Carbon	heterotrophic	oxic	rapid
Nitrogen			
Nitrification	autotrophic	oxic	slow
Denitrification	heterotrophic	anoxic	rapid in the presence of a carbon source
Phosphorus	heterotrophic	anoxic/oxic	dependent upon carbon elimination

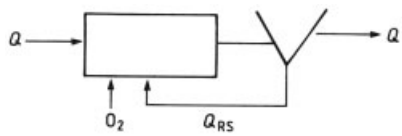

Figure 22. Single-step carbon elimination
Q = wastewater; Q_{RS} = recycled sludge

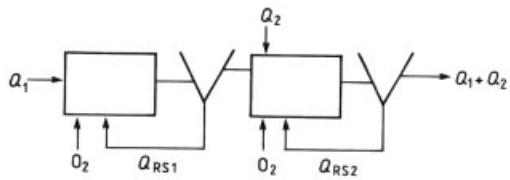

Figure 23. Two-step carbon elimination
Q_1, Q_2 = wastewater streams; Q_{RS1}, Q_{RS2} = recycled sludge streams

Carbon Elimination. The biological elimination of carbon may take place as a single step, in

the form of a cascade, or as two steps, each with its own discrete sludge cycle.

The one-step approach (Fig. 22) is normally preferred under low-load biological conditions. The two-step mode of operation is usually conducted as a combination of a high-load first step with a low-load second step. In order to stabilize the second step (i.e., ensure sufficient sludge growth), the wastewater stream is sometimes split, with a portion of it being fed directly into the second stage (step feeding) (Fig. 23) [145]. In the cascade mode of operation (Fig. 24), utilized only in special cases, a second stream of wastewater can be fed directly into the second cascade.

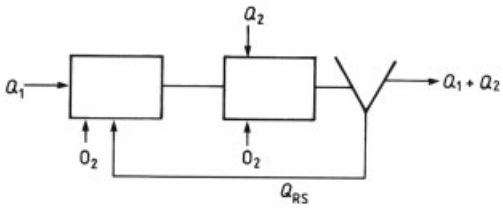

Figure 24. Cascade system for carbon elimination
Q_1, Q_2 = wastewater streams; Q_{RS} = recycled sludge

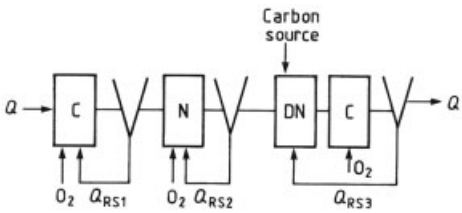

Figure 25. Three-step elimination of carbon and nitrogen
Q = wastewater stream; Q_{RS1}, Q_{RS2}, Q_{RS3} = recycled sludge streams; C=carbon-elimination; N=nitrification; DN = denitrification

Nitrogen Elimination. The elimination of nitrogen is based on two different processes: nitrification (N), the oxidative conversion of ammonium ion to nitrate by autotrophic organisms, and denitrification (DN), the reduction of nitrate to molecular nitrogen by heterotrophic organisms in the presence of easily biodegradable organic compounds, as described above.

Wastewater is usually contaminated primarily by organic compounds, which must also be eliminated biologically, so the nitrogen-elimination process can be extended to cover the elimination of both carbon and nitrogen. This can be accomplished, for example, in a three-stage plant with three separate sludge cycles (Fig. 25). The last step must be conducted as a cascade, because adequate anoxic denitrification requires that an appropriate source of carbon be introduced in excess. Unreacted residue must than be degraded oxically.

However, this particular biotechnological approach to the problem entails a major capital investment coupled with high operating costs, so less costly solutions are being sought. For example, efforts have been made to achieve carbon elimination and nitrification in a single stage. For this purpose, the stage in question should be subjected to the lowest load level possible, permitting adjustment of the growth of heterotrophic organisms by a shift into the death phase, thereby establishing a proper match with the slow growth of the autotrophic organisms. This is essential to ensure that there will be no change in the sludge composition accompanying the withdrawal of excess sludge. Excessive growth or loss of various organisms through washing must be prevented to the greatest extent possible. This in turn imposes a demand for a larger bioreactor to ensure concurrent carbon elimination and nitrification, but at the same time eliminates the need for a third sludge-separating system (Fig. 26).

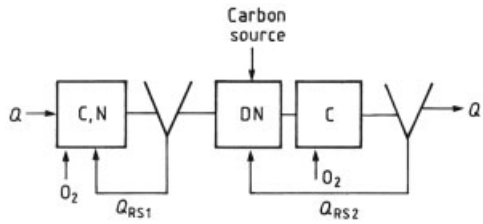

Figure 26. Two-step elimination of carbon and simultaneous nitrogen
Q = wastewater stream; Q_{RS1}, Q_{RS2} = recycled sludge streams; C = carbon-elimination step; N = nitrification; DN = denitrification

Here, too, an external source of carbon must be introduced for denitrification. In the absence of adequate amounts of low-nitrogen wastewater it becomes necessary to resort to a substance like methanol, which obviously increases the operating costs (see the first example in Section 2.1.4.3). The demand for methanol can be minimized or even eliminated by introducing a denitrification step prior to the first stage of joint carbon elimination and nitrification. Such a process combination can be operated either as two

steps with separate sludge cycles (Fig. 27) or as a cascade (Fig. 28), requiring only a final clarification. After nitrification, the nitrate that is to be eliminated is sent back for denitrification along with the recycled sludge (Q_{RS}), and/or introduced into the cascade first stage via recirculation (Q_{RC}) or backpass (Q_{BP}).

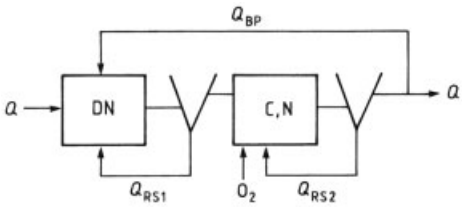

Figure 27. Two-step simultaneous elimination of carbon and nitrogen (nitrification) preceded by denitrification
Q = wastewater stream; Q_{RS1}, Q_{RS2} = recycled sludge streams; Q_{BP} = wastewater backpass; C = carbon elimination; N = nitrification; DN = denitrification

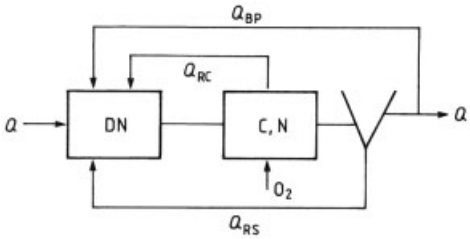

Figure 28. Cascade arrangement for carbon and nitrogen elimination preceded by denitrification
Q = wastewater stream; Q_{RS} = recycled sludge stream; Q_{BP} = wastewater backpass; Q_{RC} = recirculated wastewater; C = carbon elimination; N = nitrification; DN = denitrification

The potential extent of nitrogen elimination depends on the amount of nitrate returned, as well as on the carbon source available for denitrification [146]. A preceding denitrification step leads inevitably to unloading, and thus stabilization of the subsequent process of joint carbon elimination and nitrification.

Should the nitrogen elimination achieved in a preceding denitrification step prove insufficient, a second denitrification can still be imposed subsequent to carbon elimination and nitrification. Introduction of a suitable carbon source is again required in the form of a nitrogen-free wastewater stream or perhaps a stoichiometric excess of methanol. For this reason, final denitrification must take the form of a cascade so that any excess of the carbon source present at the end of the process chain will be eliminated biologically (Fig. 29) [147].

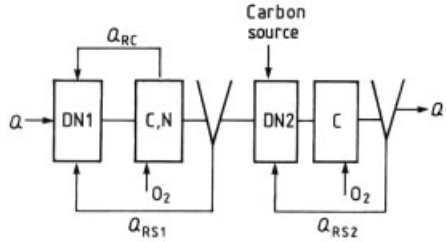

Figure 29. Two-stage cascade for simultaneous elimination of carbon and nitrogen, with denitrification both before and after
Q = wastewater stream; Q_{RS1}, Q_{RS2} = recycled sludge stream; Q_{RC} = recirculated wastewater; C = carbon-elimination step; N = nitrification; DN 1, DN 2 = denitrifications

Phosphorus Elimination. Biological phosphorus elimination accomplished in conjunction with chemical precipitation is found today primarily in the treatment of municipal sewage. The change from an anoxic to an oxic state, required for luxury uptake, is achieved by providing a preceding anoxic stage. Insofar as possible, reintroduction of dissolved oxygen or nitrate is avoided in this step, because every oxygen donor leads to a decrease in performance [52]. The desired end is met by a very high rate of recirculation (Q_{RC}) (Fig. 30).

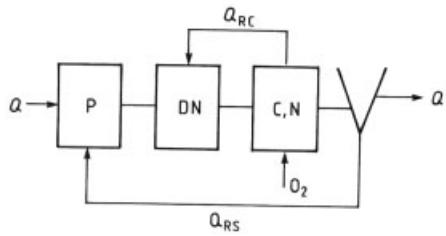

Figure 30. Cascade arrangement for the elimination of carbon, nitrogen, and phosphorus, where in volumetric terms $Q_{\mathrm{RC}} \gg Q_{\mathrm{RS}}$
Q = wastewater stream; Q_{RS} = recycled sludge stream; Q_{RC} = recirculated wastewater; C = carbon-elimination step; N = nitrification; DN = denitrification; P = phosphorus elimination

2.1.3. Technology and Operation of a Wastewater-Treatment Plant

The central wastewater-treatment plant of a direct-discharge system can be subdivided into the following process segments [148]:

1) Feed
2) Chemical purification (e.g., neutralization)
3) Mechanical purification
4) Quantity and quality adjustment by means of buffer and balancing volumes
5) Separate detection of unusual influents in the catch volumes
6) Single- or multistep biological wastewater treatment, together with any necessary off-gas treatment
7) Sludge treatment and targeted disposal [149]
8) Discharge into a receiving stream, with appropriate account taken of potential high-water situations

Assessment of the individual process stages and selection from among possible process variants requires a precise knowledge of the particular wastewater situation in question (district wastewater register, wastewater regime, possible production-related variations as a consequence of economic factors and production mode), as well as a consideration of local circumstances.

The hydraulic dimensioning of the individual process sections constituting the treatment plant must be based on short-term maximum influxes of wastewater (m^3/s). However, likely increments such as those that would accompany a heavy rain dare not be neglected in the course of planning. Possible compensating measures within the plant itself can be considered for the successive stages assuming adequate equalization and buffer volumes will always be available.

In terms of sizing the various load-based process steps (e.g., the biological steps) it is the maximal influx in t/d that is decisive, taking into account the possible need for compensation on a daily or weekly basis. The actual goal of the process is also important: whether the plant is designed only for carbon elimination or for the elimination of both carbon and nitrogen.

The task of any central wastewater-treatment plant is to fulfill at all times the legal requirements stipulated in the corresponding official discharge license. For this reason, heavy emphasis must be placed on process reliability [150] in the individual processing steps. A disruption in one part of the plant (e.g., in the event of repairs) should not present a risk to the overall disposal function. Therefore, depending on its size, a wastewater-treatment plant should be designed and constructed as a two – stream or even better as a multistream operation. In order to detect weak points and define the requisite redundancies it has been found useful to conduct an *availability analysis*. This must in turn be based on the requirement that safe operation of the wastewater-treatment facility will be possible at all times.

In addition to quantity-oriented factors (volumetric flow rate, load), the quality of the incoming wastewater is also important in the dimensioning of a wastewater-treatment plant. In many cases, use of data acquired from other wastewater-treatment facilities is possible only to a limited extent. Wastewater qualities from different companies and different manufacturing sites vary tremendously due to differences in factors related to production, process design, and process integration. For this reason the relevant parameters must always be taken from experiments conducted at the laboratory and pilot-plant levels. It is thus understandable that wastewater-treatment plants associated with chemical operations differ greatly from one to another. Each must be custom-designed to suit the wastewater conditions at a particular location consistent with current state-of-the-art technology. For this reason several different plants are described in detail in Section 2.1.4 as a way of illustrating the potential diversity.

Buffer, Equalizing, and Receiving Volumes. To ensure that required effluent levels are reliably maintained, a biological wastewater-treatment plant should be operated insofar as possible without major quantitative or qualitative fluctuations in the inflowing wastewater. Buffering the production against short-term (daily) and long-term (weekly) variations imposes the need for equalization volumes. This buffering can be achieved in a decentralized way at the various production sites or centrally in the wastewater-treatment plant. The appropriate capacity is a function of the anticipated fluctuations. Figure 31 illustrates a typical buffering

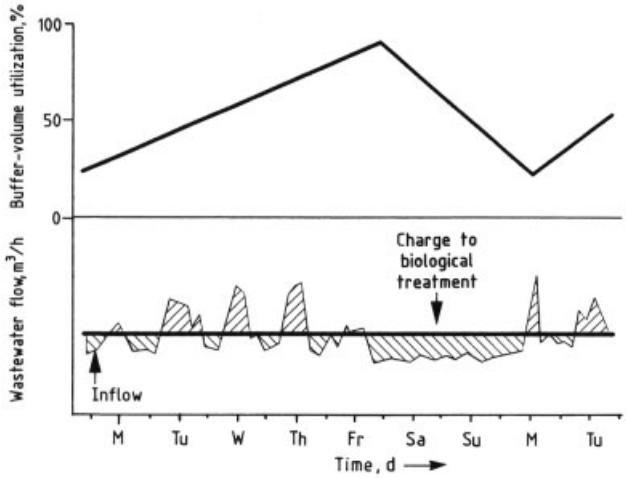

Figure 31. Buffering strategy for dealing with a variable wastewater flow

strategy that takes into account inflow, utilization of the available volume, and loading of the subsequent biological stages from the standpoint of quantitative equalization over the period of one week. In practice, however, this approach is eclipsed by qualitative considerations.

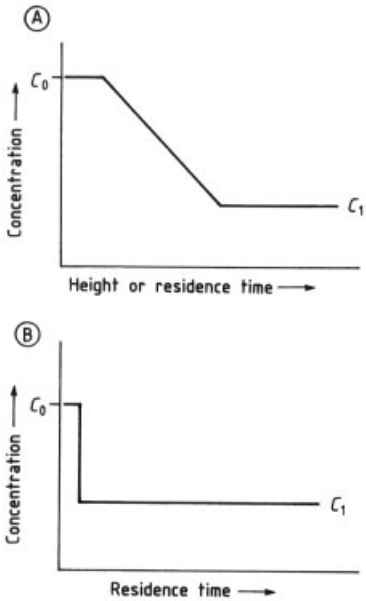

Figure 32. Concentration-change patterns for biological wastewater purification
(A) Trickling filter with plug flow; (B) Activated-sludge process, total mixing
c_0 = influent concentration; c_1 = effluent concentration

Controlling unusual influxes into the wastewater-treatment plant, such as those resulting from an operational failure or fire, creates the need for centralized or decentralized receiving volumes. Their capacity depends in turn on the extent of the hazard potential. Of great importance with respect to the operation of a barrier system of this type is timely detection of an unusual event [151] with the aid of both analytical and organizational measures.

Biological Wastewater Treatment. Biological treatment of wastewater in the chemical industry usually takes the form of an activated-sludge process, in exceptional cases a trickling-filter process. Deciding which of these to implement requires consideration of two primary factors, the advantages and disadvantages of which are immediately apparent:

1) The activated-sludge process is conducted either oxically for carbon elimination and nitrification, or anoxically for the purpose of denitrification and phosphorus removal. The operating conditions are therefore predetermined. On the other hand, in the biofilm of a trickling filter (see Fig. 19), the oxic conditions at the surface are supplemented by anoxic or even anaerobic conditions in the interior, depending on the film thickness. Thus, it is likely that methane will be generated in an aerated trickling filter, perhaps even in large quantity.

2) In a trickling filter, wastewater (diluted, if necessary) flows via the return flow Q_R (see Fig. 19) through a packed column according to the principle of plug flow. This results in formation of a concentration gradient over the height of the trickling filter. Microorganisms at the top of the trickling filter come into contact with high concentrations of contaminants, which may inhibit degradation depending on the quality of the wastewater. On the other hand, if an activated-sludge process is operated as a thoroughly mixed system (see Fig. 15), the influent concentration C_0 is diluted immediately to the effluent concentration C_1, reducing the sensitivity of the system to toxic influents (Fig. 32).

Dimensioning of the biological stage of a wastewater-treatment plant for the chemical industry should be based preferentially on results from laboratory- and pilot-scale experiments. Sizing data from "comparable" treatment plants or municipal sewage-treatment plants can be used as a rough guide, but special account must be taken of the effects of low degradability and inhibition of the degradation process by substances that might be toxic at ambient concentrations.

The proper sizing of a trickling filter [152] depends for biological degradation on the load, expressed in terms of the loading rate per unit area (kg BOD or COD/$m^2 \times d$), and from a hydraulic standpoint on the washing efficiency (with return flow Q_R, if necessary) to prevent plugging and ensure elutriation of excess biomass.

The sizing of an activated-sludge plant [153] is based on the load expressed as the loading rate ("food") per unit volume [F/V = kg BOD (COD)/$m^3 \times d$] or, better still, on the sludge loading [food/microbe ratio F/M = kg BOD (COD)/kg STS$\times$d], where sludge total solids (STS) takes into account the sludge properties as well (see Table 5). The hydraulic aspect depends on the throughput efficiency [superficial velocity, $m^3/m^3 \cdot h = m/h$] of the external biomass separation as a function of the sludge properties [154], [155]. The design can reflect either sedimentation or flotation [155], [137].

Only the organic portion of the sludge parameters should be factored into the dimensioning of an activated-sludge plant. This is significant, especially in the case of simultaneous phosphorus precipitation, because considerable amounts of iron or aluminum phosphates accumulate in the activated sludge. In exceptional cases, simultaneous phosphorus precipitation may even prevent nitrification, because the nitrifying bacteria are unable to multiply to a sufficient extent (see Nitrogen elimination in Section 2.1.2).

Special Process-Engineering Solutions. Odor emission accompanying start-up of a wastewater-treatment plant has frequently been described, and it is regarded as a significant nuisance by residents in the immediate vicinity. The main components of the problem are quality, plant size, technology, and the location of the facility. Odor emission may be attributed to two causes:

1) Formation of aerosols during the aeration process, which acquire their odor only after evaporation.
2) Stripping of malodorous substances in the course of oxygen transfer, a process that competes with sorption (Fig. 33).

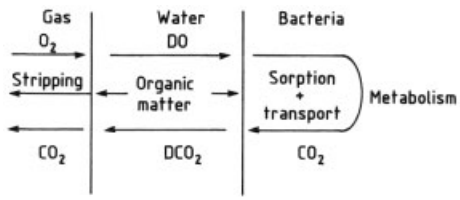

Figure 33. Mass transport at the gas – water and water – bacteria interfaces
DO = dissolved oxygen; DCO_2 = dissolved carbon dioxide

The solution to this problem is to build closed wastewater-treatment plants that retain the aerosols, simultaneously minimizing the quantity of off-gas through optimal oxygen utilization [10], [154]. This in turn opens the way to targeted off-gas treatment.

The Unox Process. The Union Carbide Corporation (UCC) has developed in the United States an activated-sludge process (Unox) based on the use of technical-grade oxygen. It consists of a closed cascade in which wastewater and gas flow in cocurrent fashion to ensure optimal uti-lization of the oxygen. Gas dispersal is accomplished with aeration mixers. The biological stage is followed by sedimentation (Fig. 34) [138], [157–159].

Table 5. Sizing of the activated sludge process for domestic sewage

Parameter	Loading rate per unit volume		Sludge loading		Activated sludge concentration (total solids), g/L	Sludge age, d
Treatment goal	kg BOD_5 / $m^3 \cdot d$	kg COD / $m^3 \cdot d$	kg BOD / kg STS · d	kg COD / kg STS · d		
Carbon elimination	<1	<2	ca. 0.3	ca. 0.6	3	ca. 5
Carbon elimination and nitrification	<0.5	<1	<0.15	<0.3	3	>10
Carbon elimination, nitrification, and aerobic sludge stabilization	<0.2	<0.4	<0.05	<0.1	4	>20

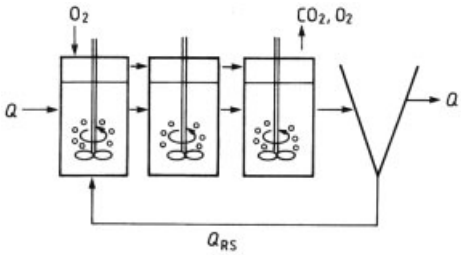

Figure 34. The Unox process
Q = wastewater stream; Q_{RS} = recycled sludge

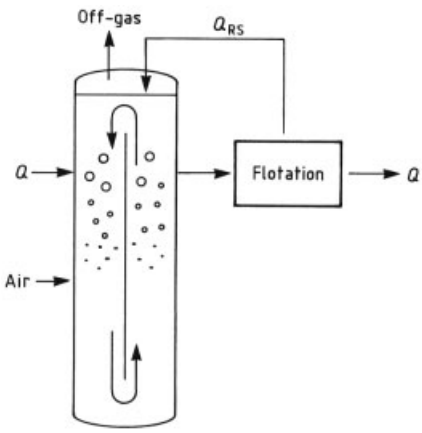

Figure 35. The ICI deep-shaft process
Q = wastewater stream; Q_{RS} = recycled sludge

The Deep-Shaft Process [160]. The British concern ICI is credited with developing the deep-shaft process. This is based on a loop reactor with a depth as great as 200 m (Fig. 35), which is aerated with fine air bubbles during downward flow. At the high pressures characterizing the lower region, both oxygen and nitrogen are completely dissolved. The pressure drop accompanying subsequent upward flow causes gas bubbles once again to form, and this in turn serves both as the driving force for the loop reactor and as the basis for sludge separation by dissolved-air flotation.

Bayer Tower Biology (Turmbiologie) Process [7], [10], [161–163]. The Turmbiologie process was developed by the German company Bayer. This utilizes a bubble column associated with a head of water corresponding to as much as ca. 30 m. Aeration is provided by fine bubbles created with injectors distributed across the bottom. Any off-gases that accumulate are monitored and specifically diverted. These gases may in turn be fed into an appropriate treatment plant as required. Sludge separation is accomplished through sedimentation, or else by gassing or degassing flotation. Prior sedimentation the water – sludge mixture is deliberately degassed and flocculated in cyclones. Sedimentation occurs in Dortmund tanks placed either next to the reactor or around the upper portion of the bioreactor in the form of a collar (Fig. 36).

The Tall Bioreactor [163], [164]. Another German company, Hoechst, developed what is known as the Biohochreaktor ("tall bioreactor"). This is a loop reactor with a water column of ca. 20 m. Oxygen is introduced via radial jets installed below the internal flue. Sludge separation is accomplished by sedimentation in clarifiers, which are integrated into the upper portion of the bioreactor. The possibility of flotation is also under discussion (Fig. 37).

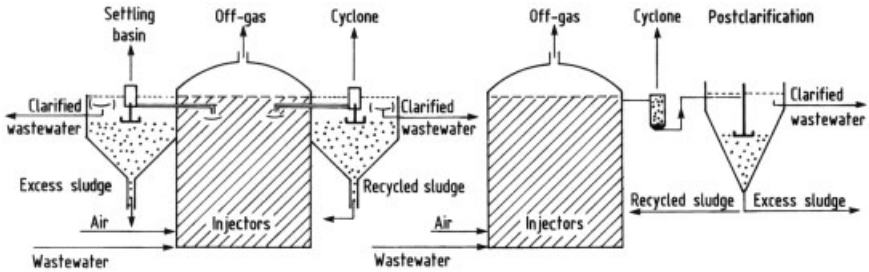

Figure 36. The Bayer biotower (Turmbiologie) process

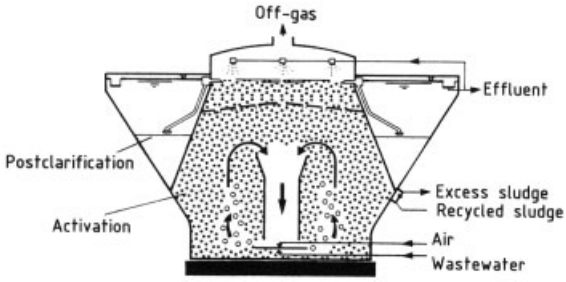

Figure 37. The Hoechst tall bioreactor (Biohochreaktor) process

PACT Process [165–167]. The Powdered Activated Carbon Technology (PACT) process was developed in the United States by DuPont. In this approach to biological wastewater treatment, powdered activated carbon is introduced directly into the activated-sludge process. It has been shown that odor emission caused by stripping can in this way be reduced, and that the sludge properties are influenced positively as well. The resulting sludge becomes more dense, leading to a lower sludge volume. The space – time yield is thus improved, because there is more bacterial mass present in the activation system, and the sludge loading is reduced. The same effect can be achieved with other growth surfaces, including lignite – coke dust [168] and plastics. For economic reasons, the use of activated carbon is potentially significant only in cases where stabilization of the biological system is achieved through the selective adsorption of degradation-inhibiting wastewater components (see also Section 5.1.7.1).

Reliable Operation of a Wastewater-Treatment Plant. Apart from careful planning of a wastewater-treatment plant and a concern for redundancy in the design, it is necessary that provision also be made for specific monitoring and balancing of the individual stages and their reliable control [169]. Proper operation of a wastewater-treatment plant demands constant surveillance and the targeted adjustment of various process parameters. Adherance to the terms of a discharge license [150], [151] depends upon both the proper mode of operation for the wastewater-treatment plant itself and the quality of the inflowing wastewater. Monitoring of the relevant parameters can be accomplished through on-line measurements that facilitate direct intervention and control, or analytical results derived from wastewater samples, which reflect longer-term trends and are also essential for monitoring and documentation purposes. On-line measurements are available for parameters such as the following: quantity of wastewater (Q_R), temperature, pH, electrical conductivity, bacterial toxicity [170], [171], DO, density and/or solids content, TOC, total nitrogen, NH_4^+, NO_3^-, and total phosphorus and PO_4^{3-}.

Both the needs with respect to on-line measurements and the availability of suitable equipment for the purpose must be analyzed separately for every contemplated application. A survey of possible measurements and analyses applicable to the various process stages associated with a wastewater-treatment plant is provided in Table 6. Specific requirements may vary in individual cases.

2.1.4. Examples of Industrial Wastewater-Treatment Plants Based on a Variety of Centralized Processes

Differences in product spectra, differently configured production systems, and unique local conditions determine the nature of the water

Table 6. Control and monitoring of a wastewater-treatment plant

Parameter	Influent	Neutralization	Preclarification	Buffer influent to biology	Biology	Effluent	Sludge treatment
Analyses from							
On-line measurements							
Wastewater flow	c					c	
pH	c	c	c	c	c	c	c
Temperature	c					c	
Bact. toxicity	c						
TOC				c	c	c	
Nitrogen				c	c	c	
Phosphorus				c	c	c	
Total solids, g/L			c		c		
Dissolved oxigen					c	c	
Random samples							
Total solids, g/L			i		i		i
Settling volume			i		i	i	
Average samples							
TOC	i			i		i	
COD	i			i		i	
BOD	i			i		i	
Adsorbable organic halides (AOX)	i			i		i	
Total nitrogen	i			i		i	
NH_4^+-nitrogen	i			i		i	
NO_3^--nitrogen	i			i		i	
Total phosphorus	i			i		i	
PO_4^{3-}-phosphorus	i			i		i	
Individual substances							

c = continuous outline; i = inhalatory samples.

regime associated with a particular chemical manufacturing plant.

Most such plants are "direct dischargers," i.e., they themselves operate centralized wastewater-treatment facilities and then discharge the purified effluent into the appropriate watercourse, taking into account the terms of an official discharge permit.

Some are indirect dischargers, releasing their wastewater – subsequent to applying decentralilzed treatment measures, complete with appropriate safeguards – into a public wastewater-disposal facility for further purification. The quality requirements pertaining to these wastewaters are established on the basis of widely accepted guidelines for indirect discharge together with local statutes.

The descriptions provided here of a number of different wastewater regimes, each with its own specific technical features, should illustrate the wide variety of possible solutions to the effluent problem. The plant technologies selected for presentation were determined on the basis of the quantity and contamination level of the waste-water, local conditions, and the technology available at the time of construction. Every waste-disposal plant is tailored to fit a specific application. The plants associated with direct dischargers have been further classified from a biotechnological perspective into

1) Single- and multistage activated-sludge plants
2) Oxygen plants
3) Plants with selective carbon and nitrogen elimination
4) Trickling-filter plants
5) Biofilter plants

2.1.4.1. Single-Stage and Multistage Activated-Sludge Plants

The Wastewater-Treatment Plant of BASF at Ludwigshafen, Germany [172–175]. The wastewater-treatment facility of BASF Aktiengesellschaft (Fig. 38) processes wastewater derived from the 350 production plants located in the city of Ludwigshafen and the neighboring communities of Frankenthal and Bobenheim/Roxheim. The industrial component makes up about 90 % of both the wastewater volume and the organic pollution load. In dry weather the daily influx may be as great as 700 000 m^3, with a maximum BOD_5 load of 375 000 kg. The facility removes 97 % of the BOD, 89 % of the COD, and 98 % of the nitrate.

Figure 38. Aerial view of the BASF AG wastewater-treatment plant at Ludwigshafen, Germany

The industrial wastewaters are first neutralized by introducing milk of lime (a calcium hydroxide suspension) into the main sewer (Fig. 39). After passing a coarse screening facility that retains bulky material, the wastewater is delivered by a pumping station (b) through two pressurized lines via a fine screen (c) into four circular primary clarification tanks (d). At this point, because of a short retention time (20 min), only rapidly sedimenting solids are separated.

A second section of the activated-sludge step consists of five activated-sludge "carrousel tanks" operated in parallel, with a total capacity of 300 000 m^3. These are outfitted with 110 surface aerators for supplying oxygen, which alone consume in the course of an hour as much as 15 000 kW of electrical power. The activated-sludge tanks are completely covered with glass-fiber reinforced plastic elements to minimize odor emission.

In addition to the degradation of organic matter, a degradation of nitrates introduced with the raw water also occurs in the activated-sludge unit. To ensure denitrification, one section of the circulation-tank system is kept oxygen-free, such that microorganisms there utilize oxygen from the nitrates for respiration. Consumption of roughly 50 t/d of nitrate in this way leads to a significant saving in the power demand for aeration.

The secondary clarification system (e) consists of 15 circular tanks with horizontal liquid

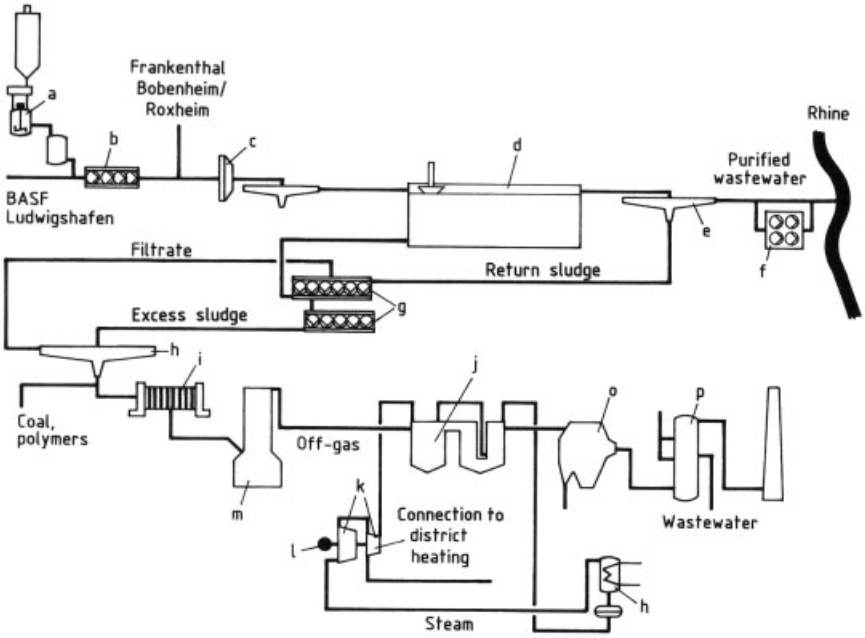

Figure 39. Schematic diagram of the BASF Ludwigshafen wastewater-treatment plant
a) Neutralization; b) Wastewater pumping station; c) Fine screen; d) Activated-sludge treatment; e) Clarification; f) High-water pumping station; g) Sludge pumping station; h) Static thickener; i) Filter press; j) Steam generation; k) Two-stage steam turbine; l) Generator; m) Fluidized-bed furnace; n) Condensation; o) Electrostatic filters; p) Wet off-gas treatment

flow and a total capacity $>100\,000\,m^3$. After the sludge has settled out in the secondary clarification system the purified wastewater flows into the Rhine. At high-water periods in the absence of a natural gradient for outflow, a wastewater pumping station (f) is capable of discharging $\leq 14\,m^3/s$ with a discharge head of 8 m.

A sludge pumping station (g) feeds most of the sedimented activated sludge back from the secondary settling tanks into the aeration tanks. A portion of the activated sludge – the so-called excess sludge – is sluiced out together with the primary sludge for sludge treatment.

The BASF wastewater-treatment plant has a central safety system for capturing toxic and poorly degradable batches of wastewater as well as water released for fire protection. The entire inflow can be led into two special aeration tanks as required, which serve as buffer volumes. A storage tank is also available for wastewater-buffering purposes. This tank is equipped with a special treatment system in which the wastewater can also be purified with activated carbon, for example. The overall storage capacity at the BASF wastewater-treatment plant amounts to $180\,000\,m^3$.

Excess sludge and primary sludge, which accumulates to the extent of $30\,000\,m^3$ per day containing up to 320 t of solids, emerges from the sludge-pumping station into thickening tanks (h). Here about 77 % of the sludge water separates by gravitational compression. After the addition of coal slurry, ash, and polymeric flocculants the concentrated sludge is subsequently dewatered in chamber filter presses (i) at pressures ≤ 15 bar to a moist-solid filter cake containing about 45 % solids.

The coal introduced into the clarifier sludge as a filter aid provides fuel for subsequent incineration. Five fluidized-bed furnaces (m) are capable of burning up to 1000 t of filter cake daily to carbon dioxide and ash. Organic constituents of the sludge decompose completely at a furnace temperature of ca. 1000°C.

The heat generated during incineration is used in waste-heat boilers (j) for steam production, air preheating, and boiler-feed preheating. The steam supplies a two-stage turbine (k) which generates 14 MW of electric power (l), provides

heat for the wastewater-treatment plant, and also supplies a portion of Ludwigshafen with domestic heat. In this way, clarification sludge serves to replace valuable sources of primary energy at many different points.

Figure 40. Aerial view of the Bayer AG wastewater-treatment plant at Leverkusen, Germany

Off-gases are dedusted in electrostatic filters (p) prior to further purification in wet scrubbers and subsequent stack discharge to the atmosphere. Resulting ash is used both in mining and as a reinforcing material at the BASF Flotzgrün landfill. A part of the ash is also recycled as a filter aid during dewatering of the clarification sludge.

The Wastewater-Treatment Plant of Bayer AG at Leverkusen, Germany [176], [177]. The Leverkusen wastewater-treatment plant is used for combined biological cleaning of the production wastewater from Bayer together with municipal wastewater from the lower Wuppertal catchment area (Fig. 40). The plant was erected in two stages: the first stage with open aeration tanks came into operation in 1971, and the second, which includes four biological towers (Bayer Turmbiologie), in 1980. Design parameters are as follows, with data applicable in 1995 in parenthesis:

Influent

a) Production wastewater 90 000 (67 000) m^3/d, BOD_5 load 95 (55) t/d
b) Municipal wastewater 70 000 m^3/d, BOD_5 load 14 t/d

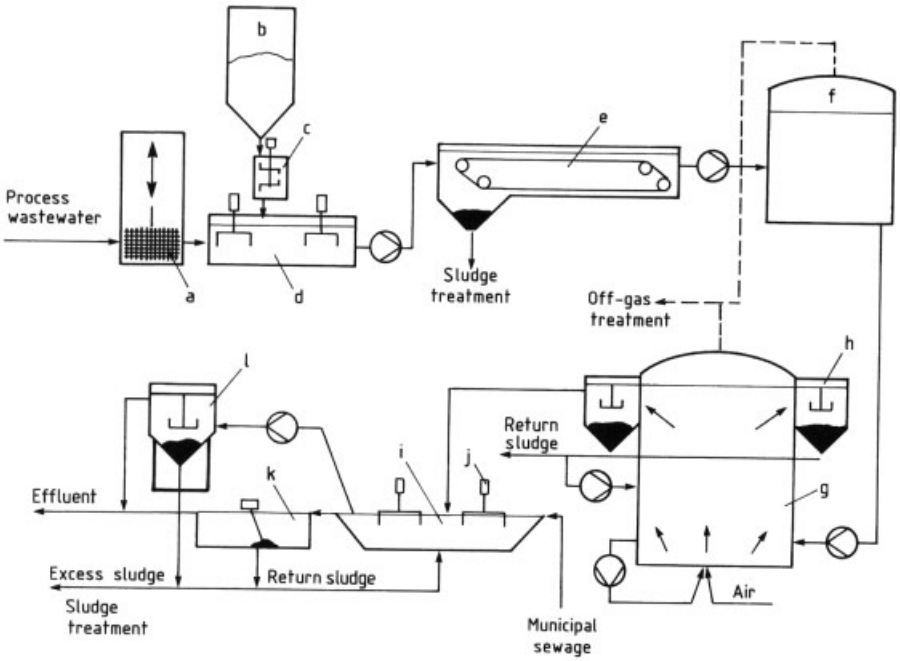

Figure 41. Schematic overview of the Bayer Leverkusen wastewater-treatment plant
a) Screen; b) Lime silo; c) Milk-of-lime preparation; d) Neutralization; e) Presedimentation tank; f) Holding (buffer) tank; g) Biotower (Turmbiologie); h) Elevated settling system; i) Biotank; j) Surface aerators; k) Final clarification tanks; l) Dortmund tanks

Primary clarification. Surface loading $1.0\,m^3/(m^{-2}\cdot h^{-1})$, retention time 2.5 h

6 Buffer tanks. Usable volume $3\times10\,000\,m^3$, $1\times15\,000\,m^3$, $2\times25\,000\,m^3$

4 Biotowers (tower biology). Usable volume $4\times13\,000\,m^3$, water height 26 m, retention time 14.5 h, oxygen feed 4×30 t/d, food/volume ratio 1.8 kg $BOD_5 m^{-3}\,d^{-1}$, food/microbe ratio 0.32 kg BOD_5/kg solids · d

Intermediate clarification. Surface loading (superficial velocity) $0.95\,m^3$, retention time 7 h

2 Biotanks. Usable volume 11 500 and $25\,000\,m^3$, water depth 3.5 m, retention times 3.4 and 7.5 h, oxygen feed 60 t/d, food/ volume ratio 0.8 – 1.7 $BOD_5\,m^{-3}\,d^{-1}$, food/microbe ratio 0.32 kg BOD_5/kg solids · d

Secondary treatment (rectangular clarifier). Surface loading (superficial velocity) $0.8\,m^3$, retention time 3 h

Final clarification (Dortmund tank). Surface loading $1.1\,m^3$, retention time 6.3 h

Wastewater entering the treatment plant by gravity flow (induced by a 10-m drop from the production facilities) is neutralized (d) after passage through a screen, lifted with centrifugal pumps, and then sent through a primary clarification system (e) equipped with chain-driven scrapers (Fig. 41). The mechanically cleaned, neutral wastewater is then homogenized in buffer tanks (f) with respect to the production-dependent parameters quantity and pollutant load. It is also possible here to introduce municipal wastewater for dilution purposes in the interest of optimal operation of the tower biology.

Wastewater is pumped from the buffer tanks directly into the biotower (g). Oxygen is fed through injectors arranged on the tank bottom. In order to optimize the oxygen-transfer efficiency, a constant volume of water is admitted to the injectors from a separate pumping circuit. Wastewater treated in the towers in the first process step flows via degassing cyclones into the elevated intermediate settling system (h), where activated sludge and water are separated. Recycle sludge is returned to the biotowers. Air required for the activation process is obtained from the production plant through a separate pressure line. Off-gases from the buffer tanks and biotowers are thermally deodorized in an off-gas treatment plant. All sections of the plant are covered to prevent odor emissions.

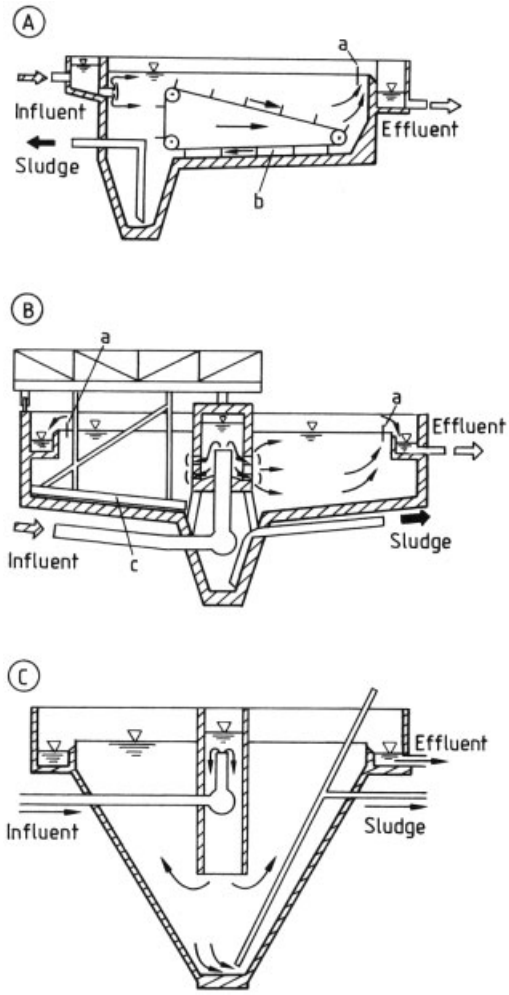

Figure 42. Various types of settling tanks
A) Rectangular flat tank, equipped with flight scraper; B) Circular flat tank, equipped with a bottom scraper; C) Hopper-bottom ("Dortmund") tank; a) Inverted buffle; b) Flight scraper; c) Bottom scraper

Effluent water from the intermediate biotower clarification process is treated further together with preclarified municipal wastewater from the Wupper Regional Authority in biotanks (i) from the first stage of construction in the form of a second processing step. Conventional surface aerators (j) supply the necessary oxygen to the aeration tanks. The available rectangular clarifiers (k) and a series of "Dortmund tanks" (l) (see Fig. 42 C) serve for the final clarification.

Three different sludges arise during the course of this wastewater purification, and all three are disposed of as follows:

1) Neutralization of primary sludge together with municipal primary sludge and excess sludge are introduced into a thickener where they are thickened with the aid of a polymeric additive to a level of 7 % solids.
2) The thickened mixed sludge is conditioned with lime and an iron salt solution and dewatered by means of chamber filter presses to ca. 65 % residual moisture.
3) The resulting filter cake is incinerated in a multiple-hearth furnace, and the ash formed is deposited in a nearby landfill.

Figure 43. Aerial view of the Hoechst AG wastewater-treatment plant at Frankfurt-Höchst, Germany

The Central Wastewater-Treatment Plant of Hoechst AG at Höchst, Germany [178], [179]. The central wastewater-treatment plant of the Hoechst at Höchst (Fig. 43) comprises three phases of construction, the first of which was brought into operation in 1967. The overall capacity of the plant is 160 t/d of BOD_5. Only industrial wastewater is treated in this plant, supplied via a special sewerage network 50 km in length. Separate systems are available for dealing with cooling water and rainwater on one hand, and sanitary sewage on the other.

Industrial wastewater flows into the treatment plant under gravity, reaching the neutralization tank after passing through a coarse screen. The predominantly acidic wastewater is neutralized with milk of lime, which is metered in through a mixing conduit in the sewer ahead of the neutralization tank. A flocculating agent ($FeSO_4$ solution) is then added to flocculate finely-divided insoluble constituents.

Wastewater pumps lift the neutralized wastewater to the surface from a depth of about 10 m. After flowing through a sand filter the water passes a fine screen and enters the presedimentation tanks of the individual purification stages, flocculation aids being added to improve sedimentation. Sludge separated in the primary settling tanks is discharged and pumped into thickener units for further concentration.

The chemically and mechanically pretreated wastewater, which is largely free of solids, is pumped into a safety and equalizing tank with a volume of 40 000 m^3, one-half of which is used to buffer the concentration and pollutant levels while the other half is kept in reserve for use in the case of a major disruption.

Subsequent to this buffering the wastewater enters an aeration tank where dissolved organic matter is eliminated by microorganisms. The biological part of the first phase of the plant is equipped with two such tanks, each with a depth of 3.6 m operating on the basis of surface aerators. This stage is at present not in operation. The activated-sludge section of the second phase, with a depth of 8.3 m, occupies considerably less space, has five rectangular tanks, and employs radial-flow nozzles for aeration. The third phase (Fig. 44) utilizes two tall bioreactors (Biohochreaktor; d and e). The concentrically arranged circular aeration space with external sedimentation tanks (f) has a water height of 20 m, and is equipped, like the second expansion stage, with radial-flow nozzles. Seven guide tubes installed above the nozzles cause the tank to function as a loop reactor. A perforated plate separates the upper region of the reactor, providing a stilling and degasification zone.

Compared with the conventional approach to construction, the tall bioreactor (Biohochreak-

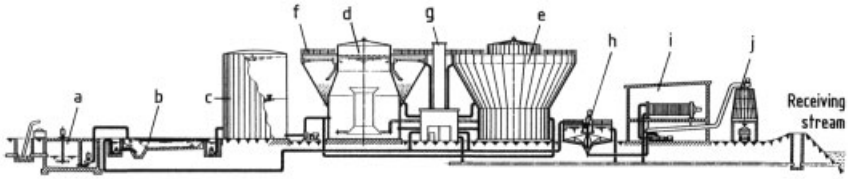

Figure 44. Schematic overview of the third development phase of the Hoechst central wastewater-treatment facility
a) Neutralization; b) Preclarification; c) Safety and equalizing tank; d) Tall bioreactor (Biohochreaktor; cross section); e) Tall bioreactor (overview); f) External sedimentation; g) Power house; h) Sludge thickener; i) Filter station; j) Filter-cake bunker

tor) offers the following important advantages (see Section 2.1.3):

1) Improved introduction of oxygen as a result of the radial-flow nozzles and the great depth of the water.
2) Less off-gas because of the low oxygen requirement.
3) Reduced space demands as a result of compactness.
4) Optimum monitoring of leaks.

Excess sludge together with primary sludge is conditioned with calcium hydroxide and dewatered with membrane filter presses. After storage in bunkers the filter cake – containing >40 % solids – was formerly deposited in a controlled landfill. After 1994 a sludge-incineration plant assumed responsibility for disposal of the filter cake.

The neutralization facility as well as the primary clarification tanks in the second and third stages are encased, as are all aeration tanks.

The Wastewater-Treatment Plant of Hoechst AG at Griesheim, Germany [180], [181]. The wastewater-treatment plant at the Griesheim works of Hoechst Aktiengesellschaft is designed for a daily BOD_5 load of 5 t, and treats industrial wastewater containing a particularly wide range of difficultly degradable and intensely odorous materials. Noteworthy features of the plant are the fact that activated carbon is added to the activated-sludge process and that all the off-gases with their accompanying volatile materials are drawn off from the various tanks to prevent odor emission. The activated carbon improves the biological efficiency of the plant as a result of its adsorption characteristics, and it stabilizes the biological system. Conceptually this plant, illustrated in Figure 45, represents a third stage in the expansion of the Hoechst wastewater facilities.

Figure 45. Aerial view of the Hoechst AG wastewater-treatment plant at Frankfurt-Griesheim, Germany

Process wastewater entering through a main sewer is first subjected to a neutralization and precipitation step, followed by a primary clarification for the separation of solids. The liquid is then collected in an equalizing tank 25 m high with a storage capacity adequate to accommodate all the wastewater collected in a single day. In the inlet section of the activated-sludge plant the raw water is diluted with cooling water, and fine-grain activated carbon is added (20 – 100 g per cubic meter of influent). The secondary wastewater-treatment stage consists of two tall bioreactors (Biohochreaktor) 24 m high, each with an aerated volume of 8000 m^3. The tanks for both the aeration stage and the integrated secondary clarification are constructed of steel, protected by several coats of isocyanate-based paint. The tall bioreactors operate on

the basis of a sludge consisting of activated carbon and biomass at a concentration of 10 – 20 kg of solids per cubic meter of aerated volume. Sludges obtained from primary clarification and the tall bioreactor are concentrated separately in two thickening units. The primary sludge is dewatered with centrifuges and the reactor sludge with chamber filter presses. This dewatered sludge is stored in bunkers prior to further treatment. To avoid problems with odor, all parts of the plant exposed to raw water, including the tall bioreactors and thickeners, are enclosed. Off-gases are removed by suction and incinerated in a combustion chamber, which operates as a separate unit within the sludge-incineration plant. Deodorization is assured by incineration with gas and fresh air at 900°C.

The incineration plant consists of an incineration furnace, a combustion chamber, and appropriate auxiliary facilities. Here the sludge is incinerated in a fluidized-bed furnace at a temperature >900°C. Natural gas is injected as a supplementary fuel, along with fresh air (preheated in a flue-gas recuperator). A cyclone is used to separate the ash, while the remaining residues are precipitated in a venturi scrubber. Hydrogen chloride and sulfur dioxide are absorbed here simultaneously by the addition of aqueous alkali. The sludge water is returned to the treatment plant, whereas ash goes to a landfill.

Wastewater Treatment of Hüls AG at the Marl Works, Germany [182–185]. The 120 different plants associated with Hüls in a closely integrated system at the Marl site produce about 2.8×10^6 t of saleable products annually. The product range includes, among others, basic and speciality chemicals, performance products, thermoplastics, and rubber, along with polymeric materials. The nearby Lippe is a rather small river, and that places heavy demands with respect to the water management of a large chemical processing operation. It has been necessary from the outset to collect all wastewater separately from unpolluted cooling and rainwater, and to process it in a wastewater-treatment plant incorporating state-of-the-art technology. For this purpose roughly 32 km of wastewater collecting sewers are embedded in roads passing through the plant.

Wastewater-Treatment Facilities. Mining operations play an important role in the runoff conditions at the 6.5 km^2 plant site, and it proved necessary to construct two mechano-biological wastewater-treatment plants at the outflow of the wastewater sewer system. These facilities purify all the polluted process wastewater arising from the production processes, amounting to ca. 77 000 m^3/d.

The wastewater treatment facilities have been expanded as production has increased, and consist now of a screening plant, a wastewater pumping station, a mechanical primary-clarification facility, two-step neutralization, an activated-sludge plant (the first biological step), mechanical secondary purification, and a settling pond (the second biological step), as shown in Figure 46.

From the standpoint of capacity these treatment facilities would be sufficient to meet the domestic sewage demands of a city of about 800 000 inhabitants. The activated-sludge plants are designed for a low food/microbe ratio ($\leq$0.1 kg BOD/kg of activated sludge solids per day), and to operate with nitrification. The nitrogen content of the plant effluent does not exceed 5 mg of ammonium nitrogen per liter. Each of the four large aeration tanks constituting the first biological stage has a volume of 6000 m^3, and is equipped with four surface aerators.

Experiments associated with the planning process indicated that short-term concentration changes would be without detrimental consequences if the wastewater is mixed rapidly with all the other water already present in the tank. Given the high hydraulic efficiency of the installed aerators, the tank contents circulate completely within 2 min. Moreover, as a result of the high activated-sludge concentration in the aeration tanks (6 – 8 kg of activated-sludge solids per cubic meter), concentration peaks are buffered by adsorption onto activated sludge flocs. Reaching these high activated-sludge concentrations required allocating to each aeration tank two humus tanks, operated with a surface-loading rate $\leq$0.5 m/h.

The clarified sludges (both primary sludge and excess activated sludge from the two wastewater-treatment plants) are collected in gravity thickeners and pre-dewatered. This thickened sludge, when mixed with organic polyelectrolytes, coagulated and passed over filterpress

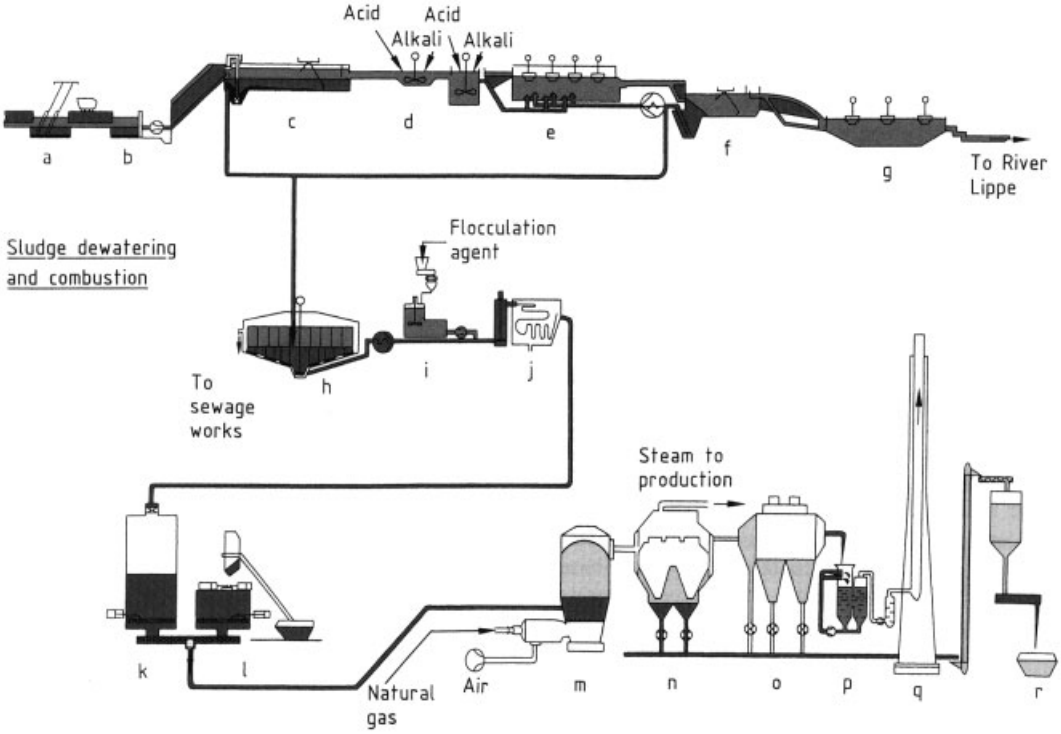

Figure 46. The Hüls wastewater-treatment plant and the sludge-treatment plant at Marl
a) Screening plant; b) Pumping station; c) Mechanical preclarification; 3 sedimentation basins; d) Neutratization; e) Aeration basins, three; f) Mechanical secondary clarification, six basins; g) Aerated settling basin; h) Thickeners, 2; i) Flocculation preparation; j) Filter press belt; k) Sludge silo; l) Carbon "station" black silo; m) Fluidized-bed furnace; n) Waste heat boiler; o) Electrostatic filter; p) Flue gas scrubber; q) Stack; r) Ash silo and discharge

belts, yields a filter cake with a solids content of 20 – 30 %. The dewatered, clarified sludges are incinerated in a fluidized-bed furnace (Fig. 46), in which the energy deficit is balanced out by a combination of solid fuels and natural gas.

Selective purification of split wastewater streams at the production plants has led to an annual decline in the load reaching the central wastewater-treatment facility and the River Lippe (Fig. 47). The range of split-stream treatments applied extends from sedimentation with and without coagulation, via stripping, to thermal wastewater treatment. The central wastewater-treatment facilities essentially perform the function of downstream purification of a mixed wastewater, as well as monitoring discharges into the Lippe. COD concentrations in the treatment-plant effluents are currently (1995) below 75 mg/L, and the corresponding BOD_5 concentrations below 10 mg/L. Beyond that, various steps have been taken in the production sphere to reduce the quantity of wastewater generated. In the period 1989 – 1995 alone the hydraulic load on the wastewater-treatment plants fell by more than 20 %.

The quality of the water discharged into the Lippe is checked continuously on the basis of extensive analyses conducted in specially equipped laboratories.

Safety Features. Apart from the wastewater, the two collecting mains for cooling and rainwater are monitored continuously as well. Here there is also continuous control of the corresponding water quality.

In the event of a remote control-panel indication that the cooling water has become contaminated, quick-acting conduit slide valves stop all discharges into the Lippe within one minute, and pumps deliver the water instead into large reservoirs. Two reservoirs, each with a capacity of 25 000 m^3, are available in the vicinity of each of the treatment plants, providing a total stor-

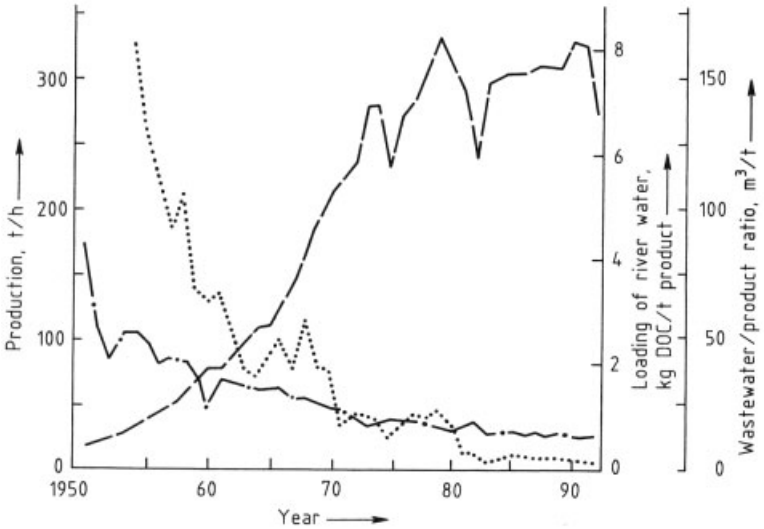

Figure 47. Product-based loading of the River Lippe from Hüls
– – Production; · · · · Loading of river water; –·–·– Wastewater/product ratio
DOC = Dissolved organic carbon.

age volume of 100 000 m^3. Another function of the storage systems is to protect the wastewater-treatment plants from extreme loads that could place the biological treatment steps in jeopardy or prevent the self-set purification goals from being met. Intensive area-wide monitoring of the sewer system in the plant complex and automated analytical monitoring of the wastewater-treatment plant influents ensures rapid detection of unusual loads (e.g., due to organic materials), permitting the influent to the wastewater-treatment facilities to be diverted as necessary with the aid of pumps into the reservoirs. At the conclusion of each such storage event the water collected is analyzed and then purified in accordance with the analytical results.

The Merck Wastewater-Treatment Plant at Darmstadt, Germany. The central wastewater-treatment facility at the Merck headquarters in Darmstadt (Fig. 48) was designed for a maximum BOD_5-removal capacity of 30 t/d, and it processes a daily average of 12 500 m^3 of wastewater with a BOD_5 of 12 t. Included in these figures is the domestic sewage (3 vol %) generated by the approximately 8000 employees, rainwater (7 vol %), and cooling water (30 vol %).

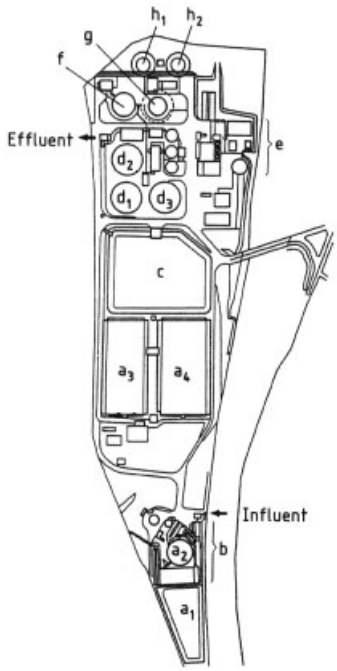

Figure 48. Plan of the Merck wastewater-treatment facility at Darmstadt
a) Retention basins 1 – 4; b) Presedimentation; c) Polishing pond; d) Secondary clarifiers 1 – 3; e) Sludge treatment; f) Equalizing tank; g) Tall bioreactor; h) Biological tanks 1 and 2

A unique feature of this wastewater-treatment facility, devoted exclusively to wastewaters from the plant complex, is the fact that it must process wastewaters derived from an unusually varied spectrum of production processes related to numerous organic and inorganic chemicals and pharmaceuticals, usually prepared on a batchwise basis. As a result of this typical polyproduction operation, there are large fluctuations over time in the quantity of wastewater collected along with the pollutant load, the pH level, and other significant characteristics. An additional challenge is presented by the relatively weak receiving stream and, related to that, the requirements and targets associated with the quality of effluent released from the wastewater-treatment plant.

Fluctuations due to polyproduction are damped in the case of COD and BOD_5 by a relatively uniform load from a production facility that operates continuously throughout the year, and by a long retention time of 5 d with intensive wastewater mixing in the equalizing tank (f), the biological stage (g), and the polishing pond (c).

In-plant wastewater-pretreatment facilities are used to keep most heavy metals, haloorganic compounds, and difficultly biodegradable materials away from the central wastewater-treatment plant, and in some cases to convert them to degradable materials. Surpluses of ammonium nitrogen and phosphate in the various wastewaters are temporarily stored, and at times of nitrogen or phosphorus shortage they are fed as nutrient solutions to the wastewater-treatment plant itself or distributed to other consumers.

As a result of in-plant pretreatment (hydrolysis) of difficultly biodegradable materials it has become possible to increase the COD-removal rate in the wastewater-treatment plant to 92 %. In the case of BOD_5, 99.5 % degradation is achieved.

Effluent concentrations expressed in terms of the most important parameters averaged (at the end of 1993): COD 120 mg/L, $BOD_5 < 10$ mg/L, adsorbable organic halogen compounds (AOX) 0.3 mg/L, inorganic nitrogen 5 mg/L, total phosphorus 2 mg/L, and heavy metals and fish toxicity equal to or below threshold values established by the German water-pollution control act. Daphnia, algae, and luminescent bacteria generally survive under the standardized test conditions in wastewater that has been diluted 1: 1. Attempts are being made to reduce the phosphorus content by simultaneous precipitation.

In order to provide for the retention of water used for fire fighting, wastewater contaminated by plant breakdowns, or extreme amounts of rainwater, four retention basins (with capacities of 1000 m^3, 3000 m^3, 8000 m^3, and 11 000 m^3) are available at the wastewater-treatment plant itself (a), and there is an additional basin (5600 m^3) located at the production plant as well. There is also the 10 000 m^3 equalizing tank (f) that is on average only half full, and this too can be used for the separate collection of contaminated wastewater.

Wastewater collected in a combined sewer system is treated based on the following steps (Fig. 49): coarse screening (a), grit chamber (b), neutralization with hydrochloric acid or lime (c), primary settling in a 1500-m^3 tank (d), equalization (e), optionally either one- or two-step biological treatment in a tall bioreactor (f; first stage; Hoechst system; 9000 m^3, of which 5000 m^3 is aerated volume and 4000 m^3 is for final sedimentation) and a biotank (g; second stage: Lurgi system; 2 tanks, each with aerated volumes of 7500 m^3 and the possibility for selective nitrification and denitrification), final sedimentation (h; three tanks, each 2100 m^3), and aerated pond polishing (i; 20 000 m^3). Based on a parallel mode of operation in the biological step, the average loading rate per unit volume is 0.6 kg BOD_5 m^{-3} d^{-1}, and the sludge loading fluctuates between 0.1 and 0.2 kg BOD_5 per kg solids per day as a function of the sludge properties (index 40 – 150 mL/g) and sludge concentrations (3 – 6 g solids/L) prevailing in the activation. In both steps, aerobic treatment consistent with the activated-sludge process is accomplished with compressed-air aeration.

Sludge formed in the course of wastewater purification (0.6 t/d of primary sludge solids, 3.5 t/d of excess activated sludge solids, corresponding to 0.3 kg of solids per kilogram of BOD_5 removal), together with excess sludge from a branch plant (ca. 1.4 t/d of solids) is thickened in two 800 m^3 thickeners from a level of ca. 5 g/L in the recycled sludge to a solid content of ca. 30 g/L. The thickened sludge is either conditioned with iron(III) chloride and lime [$Ca(OH)_2$] and dewatered in two chamber filter presses to a solid content of about 35 %, or

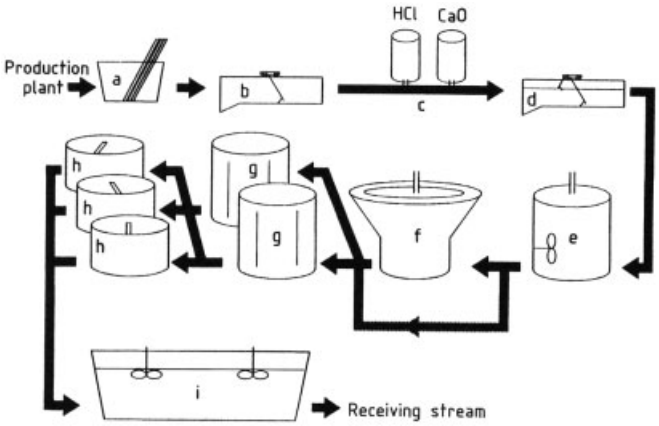

Figure 49. Schematic diagram of the Merck wastewater-treatment plant
a) Screen; b) Grit chamber; c) Neutralization; d) Presedimentation; e) Equalization; f) Activation stage 1; g) Activation stage 2; h) Final sedimentation; i) Polishing pond

dried after organic conditioning (with polyelectrolytes) and dewatering (decanter).

The dewatered (ca. 25 t/d) or dried (6 t/d) clarification sludge is utilized externally insofar as possible either as a material or as a fuel, or else it is deposited in a landfill. The contaminant load in this clarification sludge is comparable to that of a good municipal sludge, and falls well below all limiting values specified in the German ordinance on sewage sludge.

Additional very small amounts of residue are generated as screenings (0.02 t/d, external incineration), grit-chamber material (ca. 0.4 t/d, landfill), and floating sludge from the primary settling tanks (ca. 0.05 t/d, external incineration).

The wastewater-treatment plant is active and under the supervision of operating personnel around the clock, 365 days per year. Selfmonitoring is assured by numerous continuous online instruments and analysis in the company laboratory, which is staffed on a daily basis.

2.1.4.2. Oxygen-Based Facilities

The Bayer Wastewater-Treatment Plant at Wuppertal-Rutenbeck, Germany [186]. The Wuppertal-Rutenbeck wastewater-treatment plant was built to dispose of wastewater originating from the Elberfeld Works of Bayer AG (see Fig. 50). It was first put in operation in the spring of 1977.

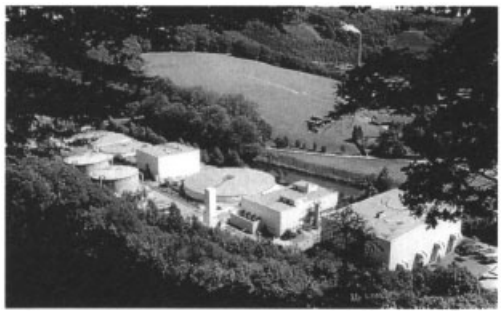

Figure 50. Aerial view of the Bayer AG wastewater-treatment plant at Elberfeld, Germany

Construction of this facility, which for reasons of space was necessarily located in Wuppertal-Rutenbeck, downstream from the production facility, demanded a compact design due to space limitations, as well as the laying of a sewer within a tunnel ca. 1 km in length.

The biological segment of the wastewater-treatment plant was intended to operate in two stages, both based on the oxygen process. Thus, instead of air, 98 % oxygen was to be injected from injectors as a way of supplying the activated sludge with sufficient oxygen for aerobic metabolism. Small quantities of off-gas discharged deliberately from the enclosed plant were deodorized in a thermal off-gas treatment plant.

A large reduction in the amounts of both pollutant and water has in the meantime been achieved by decentralized means at the produc-

tion plants themselves. As a result, aeration by oxygen is now necessary only in the first aeration stage, with normal air being used in the second. Off-gas incineration previously required has been dispensed with as well, since there are no longer any offensive odors.

Wastewater collected within the Elberfeld Works is fed to the Rutenbeck plant (Fig. 51) through a sewer leading to the influent pumping station (a). Compatibility of the wastewater with activated sludge present in the plant is checked continuously with a respirometer, and the measurements are transmitted to a control room where they are recorded.

Centrifugal pumps force wastewater from the inlet structure via two buffer tanks (c) and a subsequent neutralization facility into the first aeration stage (b). The quantity and extent of the organic load are continuously measured and recorded in the control room. This in turn permits centralized control of the entire plant by a process computer, ensuring that the activated sludge always receives an organic pollutant load that is as constant as possible, corresponding to optimum purification efficiency at all times. Load fluctuations can be levelled out with the aid of the buffer tanks. Appropriate nutrient solutions can be added to correct the carbon: nitrogen: phosphorus ratio in the wastewater as a way of optimizing the efficiency of biological degradation. Fluctuations of the influent pH are levelled out by neutralization.

The first aeration stage (b) consists of two separate trains, each with a volume of 6000 m^3 and divided into three cascades, one of 1500 m^3 and two of 750 m^3. The oxygen content of the gas phase in these two systems is >50 %. Return sludge is introduced into the first of the cascades (f). Injectors (e) installed in several rows on the tank bottom ensure very intensive aeration with oxygen, simultaneously guaranteeing thorough mixing of the reactor contents. Approximately 94 % of the organic constituents are removed from the wastewater in this very first stage.

The first aeration stage is followed by intermediate settling (i), in which activated sludge separates by sedimentation from the partially purified wastewater. Flocculation cyclones (d) to improve the settling properties of the activated sludge are installed between the aeration stages and the associated intermediate settlers or secondary clarifiers.

Sedimented sludge is pumped back into the aeration tank (b) in the form of recycle sludge, whereas excess sludge is conveyed off for sludge treatment.

Wastewater from the intermediate settler is transferred to the second aeration stage (g) which, as noted above, is oxygenated with air. This stage also incorporates two trains, each with a total volume of 1600 m^3 consisting of two cascades of 800 m^3. The second stage is designed to eliminate slowly degraded materials that have passed through the first stage. The activated sludge is separated in a subsequent secondary clarifier (k) and pumped back into the second stage as return sludge, or removed as excess sludge for sludge treatment.

Overall BOD_5 elimination is >99 %, and COD elimination amounts to 96 %. As a result of altered wastewater composition, increased retention time, and increased sludge age it has now become possible to pursue nitrification in the second aeration stage as well.

Until 1991, excess sludge was dewatered in decanters (j) to a solids content of ca. 15 %, and granulated in a mill dryer to a granular powder, which was then discarded in a landfill. With a reduction in the amount of sludge produced this process became uneconomical, so now the sludge is dewatered only to the extent of ca. 11 % dry matter, and the decanter discharge is processed further at another wastewater-treatment plant.

Both aeration stages incorporate simultaneous precipitation for phosphorus removal, whereby the phosphate is precipitated by metal salt solutions and removed together with the excess sludge.

The Wastewater-Treatment Plant of Knoll AG at the Minden Works, Germany [187]. The wastewater-treatment plant illustrated in Figure 52 began operating in 1980. It differs from conventional plants mainly in the fact that the activated sludge is aerated with technical-grade oxygen rather than air.

Mixed wastewater arriving from the production facility (process and cooling water from the manufacture of pharmaceuticals, sanitary water, and surface water, totalling about 8000 m^3/d) is pumped first into the mixing and equalizing tanks (b), with volumes of 8000 and 10 000 m^3, respectively. Their function is primarily that of

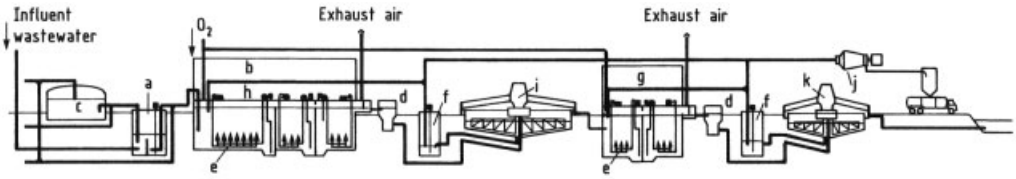

Figure 51. Schematic representation of the Bayer Elberfeld wastewater-treatment plant
a) Influent pumping station; b) Aeration stage 1; c) Buffer tank; d) Flocculation cyclones; e) Injectors; f) Return-sludge pumping station; g) Aeration stage 2; h) Turbine blower; i) Intermediate settling tank; j) Decanter; k) Secondary clarifier

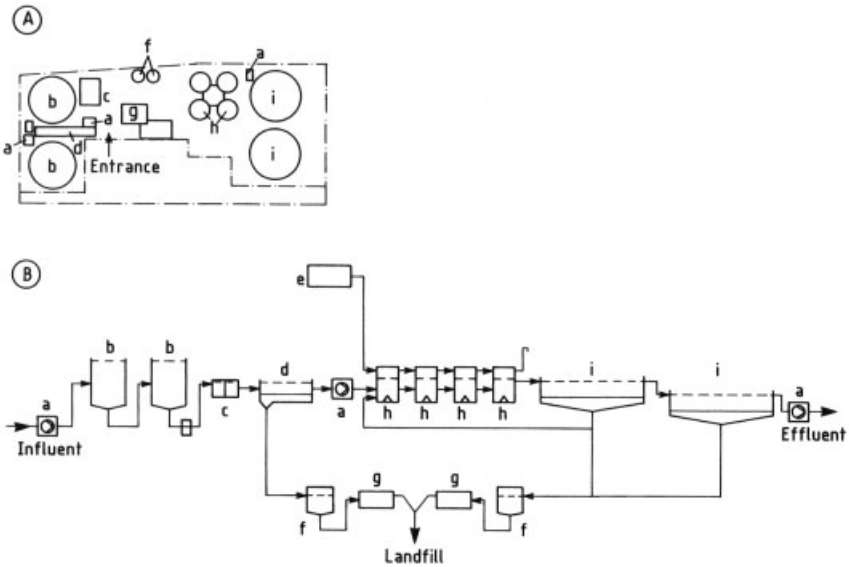

Figure 52. The Knoll AG wastewater-treatment plant at Minden, Germany
A) Sketch of the plant layout; B) Schematic diagram
a) Pumping stations; b) Mixing and equalizing tanks; c) Neutralization; d) Primary settling; e) Oxygen supply; f) Sludge thickeners; g) Sludge dewatering; h) Biological tanks; i) Secondary clarification basins

equilibrating the wastewater with respect to both constituents and pH, and of ensuring throttled inflow to the biological stage during production-free periods (weekends, public holidays).

The wastewater is then transported via a two-stage neutralization plant (c), in which milk of lime can be introduced to raise the pH, to the primary settling facility (d). Here, in a covered rectangular tank (volume 510 m^3), readily settleable materials are separated.

The mechanically cleaned wastewater is then delivered to the heart of the plant: the biological treatment stage. This consists of four enclosed circular tanks (h), all equal in size, through which the water passes successively. Because of the high oxygen demand in the first tank, it is here that the oxygen is introduced (e). Gas and the water – sludge mixture flow in cocurrent fashion through the subsequent stages. The oxygen content of off-gas from the final stage is 40 %.

The aerated volume can be varied between 4200 and 5100 m^3 by adjusting an overflow weir in the effluent shaft after the fourth stage. Gas is introduced through jet nozzles.

The water/sludge mixture then flows through two circular secondary clarifiers (i; each with a volume of 2500 m^3), in which sludge is once again separated by sedimentation. The biologically treated wastewater ultimately enters the Weser River via a penstock.

Part of the settled sludge is returned to the first aeration tank, while the remainder is transported as excess sludge to the thickener (f). Sludge obtained in the primary settling tank is thickened in an analogous way.

The sludge, to which flocculation aids have been added, is dewatered with membrane chamber filter presses (g) and fed to a landfill.

The extent of contaminant elimination based on the change in BOD_5 is ca. 98 % (90 % based on COD). Because of a shortage of phosphorus in the original wastewater, this nutrient is introduced directly in the form of phosphoric acid.

There are a number of advantages that result from aerating the activated sludge with technical-grade oxygen instead of air. Thus, the amount of off-gas requiring attention is definitely less than from air-driven plants, thereby virtually eliminating odor nuisances in the surroundings. The oxygen transfer per unit time is also greater, and the required aeration volume and therefore the land requirement turns out to be smaller as well. One important advantage associated with the four-stage cascade system is great adaptability of the biomass to fluctuations in loading. These advantages are partly offset by the need for an oxygen supply (with the attendant costs), and the added expense of installing the necessary safety equipment.

Wastewater-Treatment System of Bayer Corporation (Miles, Inc. before April 1, 1995), Agricultural Chemicals Division, at Kansas City, Missouri, USA. The Kansas City pesticide production facility of Bayer Corporation protects surface waters in the Missouri River with the aid of its terminal wastewater-treatment operations (Fig. 53). The system selected was the Unox process biological system, which utilizes high-purity oxygen rather than air. This system was chosen in preference to an air-fed system because the oxygen-enriched biological system is enclosed, thereby reducing odors and minimizing size constraints, and such a system need be only 20 % as large as a typical air-fed aerobic reactor.

All process wastewaters from the site, together with storm and rain waters, enter the primary clarifiers (d) for solids removal (Fig. 54). The wastewater is then pumped through overhead lines to the equlization system prior to secondary treatment. Two equalization tanks (f), 11 355 m^3 each, are piped in such a way that either tank can be operated in the lead or lag position; moreover, each tank can be isolated should a spill or a high level of pesticides mandate additional hydrolysis time.

The bioreactors (h) are of the Unox design, consisting of two parallel three-stage reaction trains. Each reactor has a capacity of 511 m^3. Pure oxygen is added to the first stage of each train, and the off-gas from the final stage is vented. Return activated sludge is recycled from the first final clarifier. Surface aerators in each bioreactor are used to completely mix the biomass. Typical operating parameters for the biosystem are listed in Table 7.

Oxygen for the bioreaction is produced in an air separation plant with an output capacity of 45 400 kg/d. Byproduct nitrogen vapor from this plant is piped to all operating units for use as an inert gas.

The two final clarifiers (i; 1836 m^3 each) are operated in series for solids removal. Return activated sludge from the first clarifier is split between the two reaction trains, and the second clarifier operates in a sludge-wasting mode. When solids inventories rise above the allowable operating level, sludge wasting is commenced from the lead clarifier.

Final clarifier effluent is pumped through two parallel carbon adsorbers (k) for tertiary treatment. Tertiary treatment by carbon absorption was selected because of carbon's great affinity for high molecular mass hydrocarbons. Such compounds tend to be nonbiodegradable, or at least not totally eliminated by biodegradation. Tertiary treatment has also been found to reduce toxic properties of the effluent with respect to aquatic life. Prior to installation of the tertiary plant, LC_{50} fish toxicities averaged between 1: 3 and 1: 6. Subsequent to start-up of the tertiary plant this was reduced to between 1: 0 and 1: 2. The adsorbers operate in an up-flow expanded-bed mode. Spent carbon is pulsed on a time basis ○Bfrom the bottom of the adsorbers, while regenerated carbon is added to the tops of the adsorber beds. Each adsorber has a capacity of 136 080 kg of granular activated carbon.

Carbon is transported hydraulically through the adsorption system to the spent-carbon furnace for regeneration. The plant has a fluidized-bed regeneration furnace capable of regenerating 18 145 kg/d of carbon. This furnace is designed with two beds, an upper bed for drying the carbon and volatilizing light organics and a lower one for pyrolysis and reactivation. The regeneration system has an afterburner for oxidizing as much as possible of the organics, a

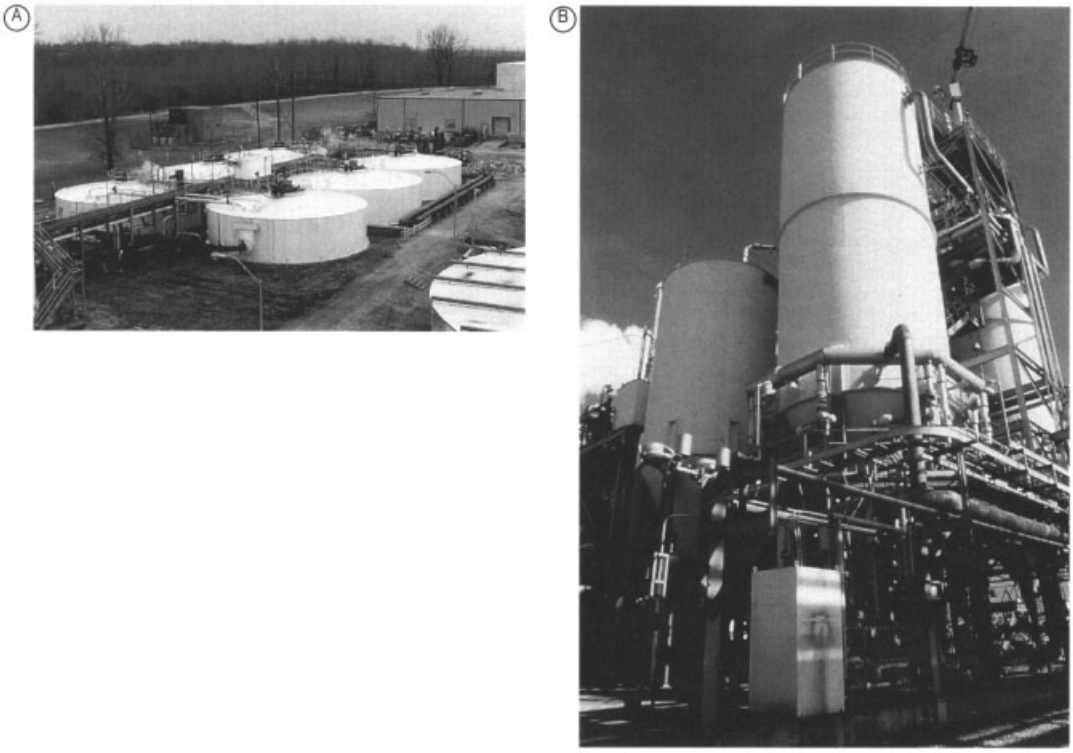

Figure 53. The Bayer Corporation wastewater-treatment plant at Kansas City, Missouri
A) Biooxidation tanks; B) Adsorption column (filter)

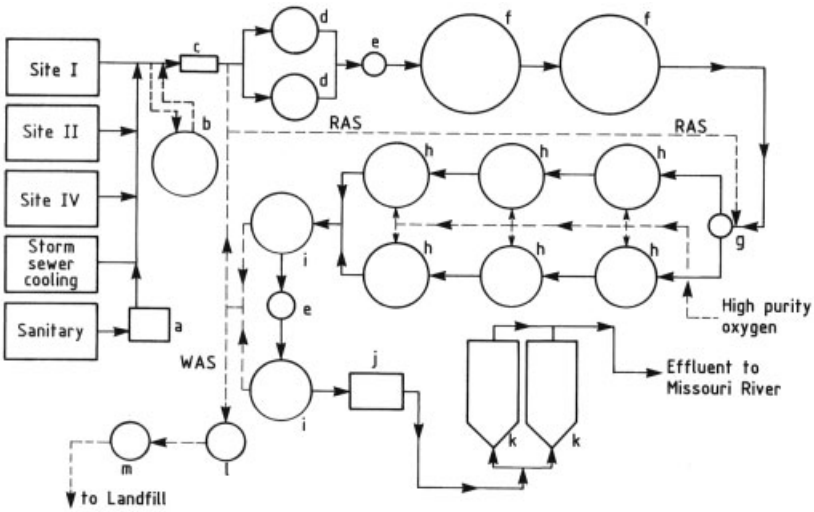

Figure 54. Schematic diagram of the Bayer Corporation wastewater-treatment plant
a) Wet well; b) Emergency tank; c) Neutralization; d) Primary clarifiers; e) Surge tanks; f) Equalization tanks; g) Splitter; h) Biooxidation tanks; i) Secondary clarifiers (adsorber feed); j) Adsorber feed sump; k) Carbon adsorbers; l) Sludge thickener; m) Centrifuge

Table 7. Biosystem operating data for the Bayer Corp. wastewater-treatment facility at Kansas City, Missouri

	Total dissolved solids, %	Total oxygen demand, mg/L	MLSS ∗, mg/L	MLVSS ∗∗, mg/L	Dissolved oxygen, ppm	Total oxygen demand removal, %
Equalization tank discharge	3.2	4100				
Bioreactor, stage I			9 500	6 650	22	
Bioreactor, stage II			9 500	6 650	22	
Return activated sludge			1 500	1 050	<1	
Carbon adsorber feed	2.6		94	66	7	40 – 50

∗ Mixed liquor suspended solids.
∗∗ Mixed liquor volatile suspended solids.

high-energy venturi for particulate removal, and a scrubber for the conversion of sulfur dioxide to sodium sulfite. Sulfite-to-sulfate conversion takes place in a liquid-phase oxidation tower that is shared with the plant's hazardous-waste incineration system.

Effluent from the treatment system is analyzed by means of a magnetic flowmeter and a suspended-solids monitor. A continuous-flow compositor samples the effluent prior to its discharge to surface waters.

All process vessels are covered and vented by means of induced draft to a central location prior to off-gas incineration as a way of minimizing odors.

Particularly noteworthy is the level of total dissolved solids (TDS) in the biofeed of this system: 3.2 %. The biomass has been acclimated to a high salt concentration, a precaution that exceeds standard practice. One other interesting facet of the operation is the high level of dissolved oxygen (DO) present in the system. It has been reported that high DO levels in an oxygen-enriched system result in a light floc with poor settling characteristics, but this is not consistent with experience gained from this particular system. The removal efficiencies for the system based on influent versus effluent levels are >99 % for pesticides and >90 % for BOD.

2.1.4.3. Treatment Facilities with Selective Carbon and Nitrogen Elimination

Wastewater Treatment of Bayer AG at the Dormagen Plant, Germany. The Dormagen production site covers about 6 km^2, and is divided by a railroad line. It is the home of ca. 50 chemical plants associated with Bayer and about 20 petrochemical plants of EC Erdölchemie GmbH, a 50 % subsidiary of Bayer. Wastewaters are discharged into two separate sewer systems. All wastewater from Bayer and EC Erdölchemie requiring special treatment is merged for processing purposes with industrial wastewaters from a neighboring sugar factory and a Dormagen brewery.

The first experimental wastewater-treatment plant at this location was erected by Erdölchemie in 1958. Mechanobiological wastewater treatment began in 1963, and by 1973 the facility had expanded to become the largest industrial wastewater-treatment plant at that time in all of Europe, with a purification capacity of 90 000 m^3/d. This was joined in 1978 by a second, completely covered wastewater-treatment plant in the Dormagen West Works, with a purification capacity of 25 000 m^3/d for the purification of intensely odorous wastewaters, mainly derived from the production of crop-protection agents.

Since then the amount of wastewater subject to treatment has continually been reduced, as has the pollutant load. For example, of the 63 t of COD originally generated daily by Bayer, 24 t/d was eliminated in 1991 directly at the production plants, leaving 34 t/d to be subsequently removed in the two centralized wastewater-treatment facilities, with the remaining 5 t/d being discharged. The total amount of wastewater fell at the same time to 58 000 m^3/d. Simultaneously, it became essential that new requirements be met for the reduction of nitrogen and phosphorus loads, necessitating the construction of production-integrated wastewater-treatment facilities together with additional treatment capacity. The fact that excess capacity was now available at the smaller of the two centralized

treatment plants because of diminishing wastewater volumes and loads presented an opportunity to develop an effective and low-cost solution, one which was realized in 1992.

The plan consisted of converting the wastewater-treatment plant at the West Works into a facility for the pretreatment of split streams for nitrogen- and phosphorus-containing waste, with conversion of the existing facility into a two-stage biological wastewater-treatment plant. For this purpose, two aeration tanks previously operated in parallel were reconfigured for connection in series, while two of the three existing Dortmund tanks originally intended for secondary treatment were allocated to the first aeration stage, the third being associated with the second aeration stage.

COD removal with simultaneous phosphate precipitation now takes place in the first stage, whereas the nitrification of ammonium ion (originally present or formed during first-stage treatment), via nitrite and nitrate, occurs in the second stage. These aerobic processes are carried out with the aid of the bacterial strains *Nitrosomas* and *Nitrobacter*, releasing both heat and H^+ ions. The biological treatment step is a sensitive one, and it demands effective control over both temperature and pH. The effluent from this primary treatment plant with its high nitrate load, together with the remaining wastewater, proceeds to the central treatment facility, where targeted denitrification is accomplished with the aid of a deliberately anoxic operation at the outset of the biological treatment. The two plants have also been separated organizationally. Because of the high sensitivity of the primary nitrification step, the primary treatment plant for nitrogen- and phosphorus-containing wastewaters must rely upon close cooperation with the relevant Bayer production facilities. The plant is therefore under the supervision of Bayer's environmental protection department. As before, the joint secondary treatment of all wastewater subject to obligatory treatment is accomplished under EC's guidance. The flow diagram in Figure 55 helps to explain the current mode of operation.

Both the primary treatment plant and the central facility are equipped with neutralization stages (a), holding tanks (b) for optional use, primary clarifiers (c), and two- or single-stage biological treatment based on the activated-sludge process and conducted in the aeration tanks (d) and the secondary clarifiers (e). Aeration is accomplished with the aid of injectors specially developed by Bayer, and secondary clarification is carried out in Dortmund tanks 16 m deep. Design data are as follows:

Primary Treatment Plant with Nitrification	
Wastewater flow	$8000\,m^3/d$
O_2-transfer capacity	30 t/d
COD influent	20 t/d
COD removal	17 t/d
Nitrification capacity	
(NO_3^--nitrogen)	2.8 t/d
Neutralization	
Volume	$760\,m^3$
Retention time	2.3 h
Holding tanks	
Volume	$9100\,m^3$
Primary clarification	
Volume	$3000\,m^3$
Retention time	9 h
Surface loading	$0.3\,m^3/(m^{-2}\cdot h^{-1})$
Aeration	
Volume	$5600\,m^3$
Retention time	17 h
Specific O_2 transfer	2.6 kg
Wastewater space loading	$1.4\,m^3\,m^{-2}\,d^{-1}$
Secondary clarification (aeration)	
Volume	$2000\,m^3$
Retention time	6 h
Surface loading	$1.1\,m^3/(m^{-2}\cdot h^{-1})$
Nitrification	
Volume	$5600\,m^3$
Retention time	17 h
Specific O_2 transfer	$2.6\,kg\,m^{-3}\,d^{-1}$
Wastewater space loading	$1.4\,m^3\,m^{-3}\,d^{-1}$
NH_4^+-nitrogen load	0.1 kg/kg solids · d
Secondary clarification, nitrification	
Volume	$4000\,m^3$
Retention time	12 h
Surface loading	$0.6\,m^3/(m^{-2}\cdot h^{-1})$
Off-gas (for incineration)	
Volume	$15\,000\,m^3/h$
Central Treatment Plant with Denitrification	
Wastewater flow	$80\,000\,m^3/d$
Max. O_2-transfer capacity	80 t/d
Nitrate-removal capacity	
(NO_3^--nitrogen)	4.3 t/d
COD influent	90 t/d
COD removal	75 t/d
Neutralization	
Volume	$960\,m^3$
Retention time	18 min
Holding tanks	
Volume	$4500\,m^3$
Primary clarification	
Volume	$8000\,m^3$
Retention time	2 h
Surface loading	$2.25\,m^3/(m^{-2}\cdot h^{-1})$
Aeration	
Volume	$32\,000\,m^3$

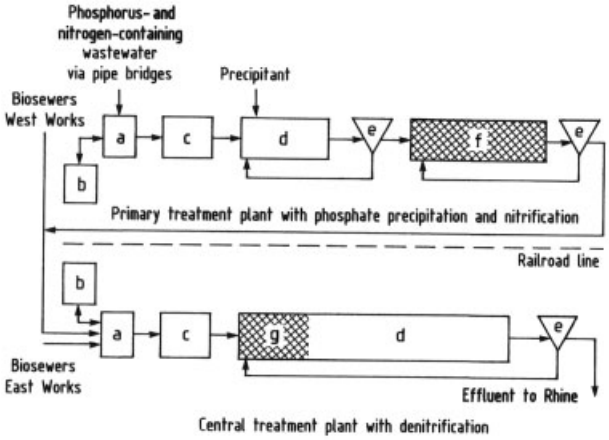

Figure 55. Schematic diagram of the Bayer AG wastewater-treatment facilities at Dormagen, Germany
a) Neutralization; b) Holding tanks; c) Primary clarification; d) Aeration tank; e) Secondary clarification; f) Nitrification; g) Denitrification

(Continued)

Retention time	12 h
Specific O_2 transfer	$2.5\,kg\,m^{-3}\,d^{-1}$
Wastewater space loading	$2.6\,m^3\,m^{-3}\,d^{-1}$
Sludge loading	0.6 kg BOD_5/kg solids · d
Sludge growth	0.3 kg solids/kg COD removal
Secondary clarification	
Volume	$18\,500\,m^3$
Retention time	6 h
Surface loading	$1\,m^3/(m^{-2}\cdot h^{-1})$

As a result of the various reconfiguration measures described, all requirements under German water legislation are met, including nitrogen and phosphate elimination.

Biological Wastewater-Treatment Plant of DEA Mineraloel AG at the UK Wesseling Facility, Germany [188–190]. The DEA Wesseling Works is a refinery and petrochemical complex with an annual crude-oil throughput of 6×10^6 t. The site is a direct discharger into the Rhine, and it must therefore meet the corresponding regulatory requirements for wastewater quality. The permissible inorganic nitrogen concentration (defined as the sum of all nitrogen from NH_4^+, NO_2^-, and NO_3^-) was previously 60 mg/L, which could be met with nitrogen removal at the 94 % level in the wastewater-treatment plant. As of 1 January 1995 the permissible value was reduced to 40 mg/L.

A modern five-stage wastewater-treatment plant based on the Bayer tower-biology principle became operational in 1981, replacing a single-stage aerobic plant of flat construction (Fig. 56). An average of $8000\,m^3$ of wastewater passes through daily, containing 6.5 t of BOD_5 (ca. 800 mg/L) and 8.1 t of NH_4^+ ion (ca. 1000 mg/L).

Figure 56. Aerial view of the DEA Mineraloel AG wastewater-treatment plant at Wesseling, Germany

In addition to contamination by hydrocarbons, wastewaters from the individual plants contain large amounts of nitrogen and sulfur in the form of NH_3 and H_2S, which enter the wastewater as a result of hydrocracking and refining of various crude-oil fractions. Separable oils are removed by phase separation in a coupled upstream physical purification stage. Stripping permits the H_2S level of the wastewater to be lowered to a maximum of 50 ppm. The stripped H_2S

is then converted to elemental sulfur by the Claus process (→ Sulfur, Chap. 6.).

The first of the five reaction stages in the biological treatment train (Fig. 57) is a denitrification (a), in which 80 % of the BOD_5 from the raw wastewater is degraded. To ensure anaerobic nitrate respiration, a nitrate-containing stream is recycled to this stage from the last nitrification stage. Gasification is accomplished with a mixture of the plant's own off-gas (N_2 and CO_2) and pure CO_2 withdrawn from the plant network. One important function of this treatment is protection of the subsequent nitrification from inhibitors present in the raw wastewater.

Nitrification is carried out in two stages, the first of which – activation with partial nitrification (c) – is a combination of aerobic degradation of the remaining BOD_5 and nitrification of 50 – 85 % of the ammonium nitrogen present. For removal of the remaining ammonium ion a nitrate-containing recycle stream flows from the outlet of the second nitrification stage (d) to the first denitrification (a_1). The two nitrification stages are aerated with air from a turbocompressor passing through a system of tubes on the reactor bottom that is outfitted with Bayer injector nozzles. The air requirement is 10 800 m^3/h at an oxygen utilization of 60 – 75 %. Nitrate is converted via nitrite to molecular nitrogen in the second denitrification stage (a_2), with methanol added as a carbon source.

The last activation stage (e) serves as protection against a BOD_5 breakthrough.

Wastewater flows through the entire plant under gravity, the water level falling from a height of 24 m at the first reactor to 17 m in the last secondary clarifier (f). All reaction stages are incorporated into closed tanks, with the intermediate and final clarifiers being Dortmund tanks.

After centrifugation, excess sludge is disposed of by incineration in a residue incinerator associated with the power station.

The Wolff Walsrode AG Wastewater-Treatment Plant at Bomlitz, Germany. The Bomlitz joint wastewater-treatment plant serving Wolff Walsrode and the community of Bomlitz was built in the years 1975 – 1977 at a cost of 7×10^6 €. It utilizes the principle of denitrification [191], and was one of the first industrial wastewater-treatment plants to incorporate this mode of operation.

As early as 1963 – 1968, extensive preliminary studies directed toward this end were carried out as a joint venture by Wolff Walsrode and the Institute for Community Water Management and Waste Engineering of the Technical University of Hannover. The objective was to test the so far undeveloped biological purification principle of denitrification as a way of utilizing as an oxygen source the nitrate content of wastewater derived from nitrocellulose production.

Preliminary experiments were initiated on a laboratory scale (total volume of the experimental plant, 53 L), pursued at pilot-plant scale (total volume 3.9 m^3), and completed with an industrial-scale facility (total volume 123 m^3). As a result, the potential for applying microbiological principles to a practical bioengineering process was substantiated, and dimensioning parameters were worked out for the design of a large plant [192], [193].

The most important findings in the course of this effort were:

1) Activated sludge from a municipal wastewater-treatment facility has already reached its full denitrification capability after only ten days. An operational change from anoxic denitrification to aeration (in the absence of nitrate) has no significant effect on contaminant-removal efficiency. The mode of operation can thus be changed as required.
2) The BOD_5 removal efficiency does not differ from that of a conventional activated-sludge plant.
3) The small amount of oxygen (<2 mg/L) introduced in the form of large bubbles during circulation of the activated sludge by aeration has no negative effect on denitrification.
4) The nitrate concentration falls and the alkalinity rises in proportion to the extent of BOD_5 removal.
5) A low phosphate content in industrial wastewater necessitates phosphate addition – to an average level of 3.7 g of phosphorus per kilogram of BOD_5.
6) Sludge flotation in the secondary clarifier due to the evolution of nitrogen gas can be avoided by brief, intensive aeration prior to secondary clarification.

Since at the same time an expansion of the municipal wastewater-treatment plant for the community of Bomlitz also was pending, it was

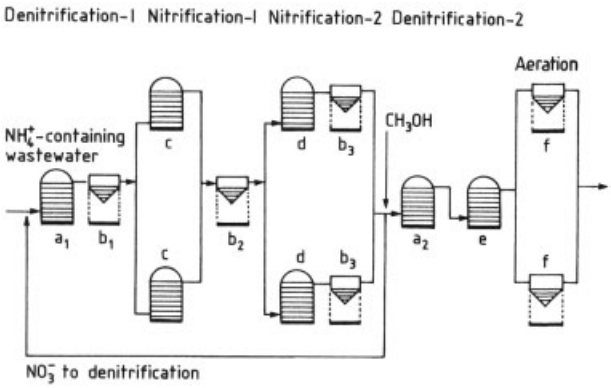

Figure 57. Schematic diagram of the DEA Mineraloel AG wastewater-treatment plant
a) Denitrification; b) Intermediate clarification; c) Activated-sludge treatment with partial nitrification; d) Nitrification; e) Activated-sludge treatment; f) Secondary clarification

decided to construct a "Bomlitz joint biological wastewater-treatment plant" outside the boundaries of the Wolff Walsrode production facility. The plan was based on the wastewater and pollutant loads indicated in Table 8.

Table 8. Projections utilized in planning the Bomlitz joint wastewater-treatment facility

	Wolff Walsrode	Municipal system	Overall projection
Wastewater flow, m^3/d	16 000	2 000 ∗	18 000
BOD_5 load, kg/d	8 000	465	8 720 ∗∗
Population equivalent	200 000	11 625	218 000

∗ Including rainwater from the municipal combined sewer system.
∗∗ Including 255 kg/d of BOD_5 from sludgefiltrate liquor.

The treatment plant become operational in 1978, and reached constant levels of purification by 1980. In the course of time the various removal percentages in the biological stage have stabilized themselves as follows:

BOD_5	97 %
COD	76 %
DOC ∗	73 %
Nitrate-nitrogen	95 %
Ammonium-nitrogen	95 %
Total nitrogen	73 %
AOX ∗∗	70 %

∗ DOC = dissolved organic carbon.
∗∗ AOX = adsorbable organic halogen compounds.

Not taken into account by these data are load reductions achieved as a result of in-plant avoidance and pretreatment measures.

Overall, the following advantages have become apparent in the denitrification process [194]:

1) Energy savings attained through the use of nitrate instead of atmospheric oxygen (7 W/m^3 of installed capacity during nitrate respiration instead of 47 W/m^3 during oxygen respiration).
2) An increase in acid-binding capacity as a result of alkalinity developed during denitrification (i.e., a saving in terms of neutralizing agents).
3) Little excess sludge production (ca. 0.5 g of solids per kilogram of BOD_5 eliminated).
4) Reduction in the nitrogen load of the receiving water.

The flow diagram in Figure 58 shows the individual steps in the wastewater-treatment process. Water from the local authority is mechanically preclarified at the treatment-plant site using a screen (a), a grit chamber (b), and a settling tank (c). This water is then mixed in the inlet structure with recycled sludge and wastewater from the production facility. The industrial wastewater is adjusted with milk of lime to pH 5 – 6 in a neutralization plant at the factory (which has been in operation since 1956), preclarified mechanically, and pumped to the inlet structure.

The activated-sludge plant consists of four circular tanks (d) connected as a cascade, each 20 m in diameter, 6.2 m deep, and with 1800 m^3 of usable capacity. Thus, including the interme-

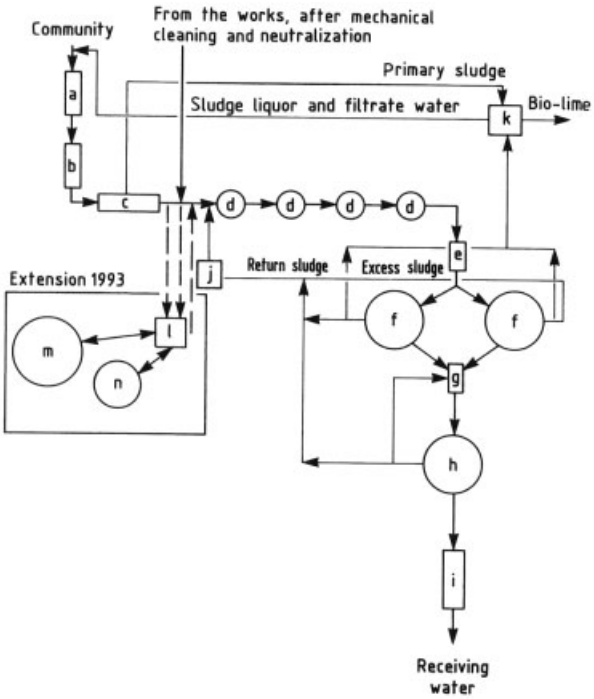

Figure 58. Schematic diagram of the Wolff Walsrode AG wastewater-treatment plant at Walsrode, Germany
a) Screen; b) Grit chamber; c) Settling tank; d) Aeration tanks; e) Intermediate aeration with hydrocyclone; f) Secondary clarification tanks; g) Intermediate aeration; h) Final clarification tank; i) Discharge cascade; j) Return-sludge pumping station; k) Sludge treatment; l) Pumping station; m) Equalization tank (16 000 m^3); n) Holding tank (5 000 m^3)

diate aeration section, ca. 8000 m^3 of total reaction volume is available. During the denitrification phase, circulation is achieved as a result of large bubbles introduced through a perforated tubular grid on the tank bottom. In aerobic operation the required input of air is ensured by injectors producing fine bubbles.

Intermediate aeration is essential for stripping nitrogen-saturated wastewater from the activation step down to the solubility limit of nitrogen relative to air. An aeration tank (e) with a capacity of 350 m^3 guarantees an aeration time of only 10 min even at peak load. A coupled hydrocyclone located downstream is used for degasification as well, but also for agglomeration of the sludge flocs. Defoaming agents or coagulation aids can also be introduced at this point.

The first secondary clarifier stage consists of two circular tanks (f) connected in parallel, each 38 m in diameter and with a surface area of 1000 m^2 and a capacity of 3600 m^3. In order to minimize the floating of activated sludge, suction dredges for the rapid conveyance of recycle sludge together with skimming troughs for clearing away any floating sludge have been installed on the scraper bridge.

A further secondary clarifier stage consists of a third secondary clarifier (h) connected to another intermediate aeration unit (g) of the same design as those used in the first stage. Clarified wastewater enters the receiving stream (the Bomlitz) via a step cascade (i) in which it is enriched with oxygen.

The return-sludge pumping station (j) includes four pumps, permitting operation with a reflux ratio >100 %. Excess sludge from the biological step is thickened together with primary sludge from the municipal sedimentation tank, conditioned with lime, and dewatered with chamber filter presses (k) to 40 – 45 % solids; the resulting "bio-lime" is subsequently used in agriculture.

Even today this operation still meets all the standards of the most recent wastewater legislation. Nevertheless, because of the largely anoxic

mode of operation the purification capacity with respect to ammonium is no longer sufficient to cover the increased demand, given the nature of the available receiving stream. Trials were therefore conducted in 1987 with a pilot-scale plant (total volume 6.4 m^3) aimed at additional nitrification. The results form the basis for a plan for increasing the efficiency of the joint facility. The proposal envisions introducing into the circuit (upstream) a mixing and equalizing volume of 16 000 m^3 capacity with a variable fill level as a way of smoothing out fluctuations in quantity and load over the course of the week, together with a second holding tank (n) with a constant fill of 5000 m^3. Feed and discharge to the aeration tank will be controlled via a new pumping station (l).

In addition to a smoothing-out in the quality and quantity of influent water and the resulting constant operating mode in the plant, the additional volumetric capacity should make it possible to conduct an at least partial nitrification operation in the existing activation cascade.

2.1.4.4. Trickling-Filter Facilities

Wastewater-Treatment Plant of Hoffmann-La Roche AG at Grenzach-Wyhlen, Germany. The plant utilizes large-scale equipment to prepare not only a wide range of pharmaceutical speciality products but also vitamins. This leads to as much as 200 000 m^3/d of cooling water and 8000 m^3 of wastewater contaminated with organic chemicals.

Separation of the Wastewater Streams. Individual wastewater streams are maintained in 3 separate sewage systems with a total length of 5 km as follows:

1) Wastewater heavily contaminated with organic material but containing only small amounts of inorganic salts and traces of organic solvents and acids (chemical wastewater 1).
2) Wastewater contaminated mainly with inorganic chemicals but containing some organic contaminants, as well as strongly acidic wastewaters (chemical wastewater 2).
3) Cooling water that has been strictly confined to closed circuits for the purpose of removing heat from chemical processes, and is chemically uncontaminated.

The first two streams described are fed to the biological wastewater-treatment plant, where they undergo appropriate treatment depending upon the specific nature of the contaminants (Fig. 59).

The plant is dimensioned in terms of the following capacities:

Chemical wastewater flow	8400 m^3/d
COD load	46 t/d
BOD_5 load	27 t/d

Neutralization. The weakly acidic "chemical wastewater 1" is first adjusted to pH 7 by neutralization (a) with aqueous sodium hydroxide.

Primary Sedimentation. Suspended matter is next removed in a primary sedimentation unit (b). The suspended matter consists essentially of all the activated carbon consumed in the production plant. The presence of this activated carbon makes it possible to take advantage of two positive effects: the carbon adsorbs pollutants up to the point of 100 % saturation, and it also acts as a welcome filter aid when the resulting sludge is later dewatered in a chamber filter press. In order to reduce fluctuations in concentration and quantity, "chemical wastewater 1" is led after primary sedimentation into a 15 000 m^3 buffer tank (c).

Subsequent biological treatment consists of two steps, one utilizing a trickling filter and the other an activated-sludge system.

Trickling Filter, Biological Stage I. The first partial purification by biological means is carried out in a trickling filter (d) with a usable volume of 2250 m^3. The water is passed over plastic fittings (Hydropak) with a specific surface area of 200 m^2/m^3. This stage is capable of removing 30–50 % of the BOD_5 load depending on the level of contamination. The requisite oxygen is derived from air circulating at the rate of 24 000 m^3/h, with a fresh-air component of 8000 m^3/h. Any solvents that do not undergo decomposition are simultaneously blown out and removed from the exhaust air by means of an activated-carbon adsorber system (e). The enclosed character of the construction in combination with this adsorption facility guarantees the absence of odors. Because of the short retention time of the water in the trickling filter, the most easily degradable materials are the ones

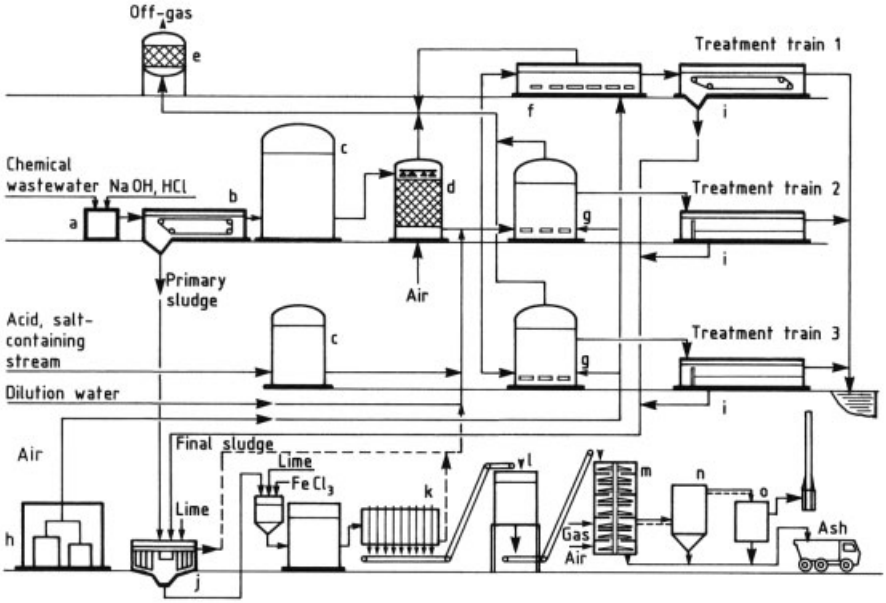

Figure 59. The Hoffmann – La Roche AG wastewater-treatment plant at Grenzach-Wyhlen, Germany
a) Neutralization; b) Primary sedimentation unit; c) Buffer tanks; d) Trickling filter; e) Carbon adsorber; f) Aeration basin; g) Aeration tanks; h) Compressors (4); i) Secondary clarifiers; j) Thickeners (2); k) Filter press; l) Silo for dewatered sludge; m) Sludge furnace; n) Afterburning chamber; o) Off-gas purification

mainly removed at this stage, which has a positive effect on sludge quality in the subsequent aeration stage.

The trickling-filter effluent is mixed with the acidic "chemical wastewater 2," which is characterized by a light organic load and is supplied from a second 5000 m^3 buffer tank, as a way of smoothing out fluctuations in concentration.

Activated-Sludge Plant, Biological Stage II. The second biological stage comprises three aeration units connected in parallel, over which the combined wastewater is distributed. One of these units is a flat aeration basin (f) constructed of concrete with a water depth of 4.4 m and a usable volume of ca. 3000 m^3. Oxygen is introduced by means of submerged aeration. The other two aeration units are identical closed steel tanks (g), each with a water depth of 20.3 m and a usable volume of 6300 m^3. Oxygen introduction is via injectors. Liquid from the tanks is recirculated to serve as a propellant for these nozzles. The propellant pumps are switched on and off in a cyclical fashion as a way of controlling the nitrification and denitrification processes.

Four turboblowers (h) generate the required compressed air. Because of differing water depths, single-stage blowers are used for the aeration tanks and two-stage blowers for the aeration basins. The blower output is adjusted as a function of the oxygen demand in the corresponding biological process.

Secondary Clarification, Sludge Dewatering, and Sludge Incineration. Each of the three aeration units is associated with its own secondary clarifier (i). Settled sludge is collected with chain-driven scrapers or suction dredges, and then pumped to two mutually independent thickeners, where it is thickened to a solids content of 2 – 6 %.

After conditioning with iron chloride and lime, the thickened sludge is dewatered in chamber filter presses (k) to a solids content of 40 – 45 %. In this form its heat of combustion is 3700 kJ/kg.

The dewatered clarified sludge is incinerated in a multiple-hearth furnace (m) with a supporting flame of natural gas. The resulting ash can be disposed of in the same way as domestic refuse.

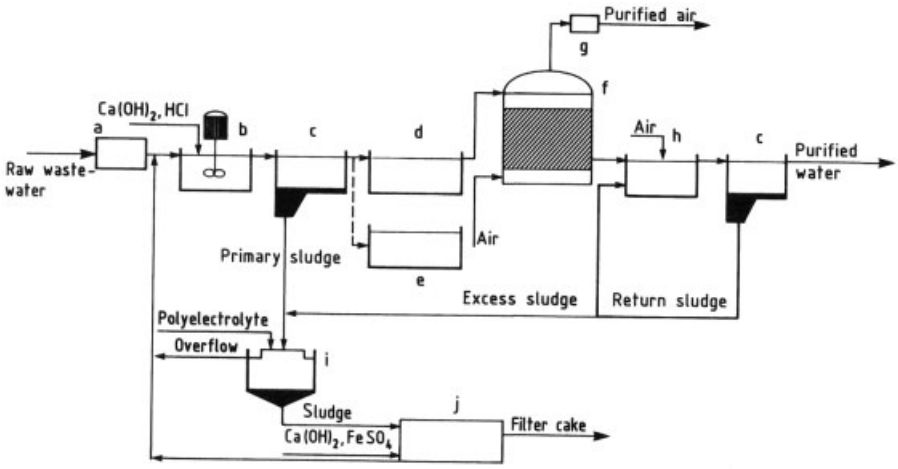

Figure 60. Schematic diagram of the Hoechst AG Kalle-Albert wastewater-treatment plant
a) Pumping station; b) Neutralization; c) Clarifiers; d) Equalizing tank, buffer stage; e) Safety basin; f) Trickling filter; g) Thermal off-gas post-combustion; h) Active-sludge treatment; i) Thickener; j) Sludge dewatering

Exhaust-Air Treatment. All parts of the plant are either enclosed or covered to prevent odor problems. They are at all times kept under a slight underpressure by the application of suction. Exhaust air from the ventilators is purified via activated-carbon adsorbers or cross-flow washers.

The wastewater-treatment plant of the Hoechst AG Kalle-Albert Works at Wiesbaden, Germany, is designed for a wastewater flow of up to 20 000 m^3/d, with an influent organic BOD_5 load of up to 33 t. The facility consists of a chemicomechanical stage, followed by two biological stages, with subsequent sludge treatment. The plant is laid out in two processing trains (Fig. 60).

The design of the plant is characterized by the removal of orthophosphate present in the influent through precipitation with slaked lime as well as nitrate denitrification in both the equalizing tank and the heavily loaded trickling filters incorporated as a subsequent stage.

Chemicomechanical Stage. The predominantly acidic wastewater reaches the inlet of the neutralization unit via a pumping station (a) that includes a screening system, and it is neutralized (b) with milk of lime. Alkaline peaks are balanced out with concentrated hydrochloric acid. The pH is adjusted to 8.5 as a way of precipitating out phosphate present in the influent. The inorganic sludge is separated along with primary sludge in the primary clarifiers (c).

Under normal operation this chemically and mechanically pretreated wastewater, which is largely free of solids, flows directly to the equalizing tank (d). Water that is abnormally heavily contaminated (e.g., water resulting from the extinguishing fire) is drained as necessary into a safety basin (e) to prevent possible harmful effects in the subsequent biological systems.

Biological Stages. Biological purification is accomplished in two stages. The first consists of two heavily loaded trickling filters (f). Trickling-filter effluent passes immediately into the two aeration tanks (h) that constitute the second stage. Nitrate present in the plant influent is denitrified to some extent in the equalizing tank of the chemicomechanical stage and in part in the two trickling filters of the first biological stage.

Sludge Treatment. Excess and primary sludges are combined and thickened (h), and after conditioning with iron salt and slaked lime the thickened mixed sludge is dewatered in filter presses (j).

Design specifications for the facility are as follows:

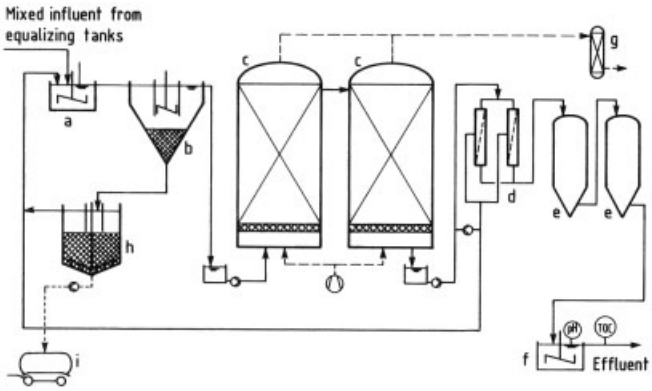

Figure 61. Schematic diagram of the Ciba Geigy AG wastewater-treatment plant at Münchwilen, Switzerland
a) Neutralization, flocculation, and precipitation; b) Sedimentation (preclarification); c) Two of the three fixed-bed bioreactors; d) Microfiltration; e) Activated carbon adsorbers; f) Postneutralization, final monitoring; g) Off-gas treatment; h) Thickener; i) Thickened-sludge disposal

Plant influent	
Wastewater flow	20 000 m^3/d
BOD_5 load	33 t/d of neutralization influent 30 t/d of influent to the biological stages
Trickling filters	
Space loading	2.8 kg BOD_5 m^{-3} d^{-1}
BOD_5 degradation	70 – 75 %
Aeration	Forced aeration via ventilators
Air flow	20 000 m^3/h, of which 10 000 m^3/h is fresh air and 10 000 m^3/h is recycled air
Exhaust air	Deodorized by thermal afterburning
Aeration basin	
Space loading	1.3 kg BOD_5 m^{-3} d^{-1}
Sludge loading	0.26 kg BOD_5/kg (dry solid matter) · d
BOD_5 degradation	90 – 95 %
Aeration	Diffused aeration via Hoechst radial-flow nozzles
Overall plant	
BOD_5 degradation	98 – 99 %
Specific excess sludge production	0.25 kg (dry solid matter)/kg BOD_5 degraded

2.1.4.5. Biofilter Systems

The Wastewater-Treatment Plant of the Ciba Geigy AG Agro-Development Center at Münchwilen, Switzerland [195–198]. This wastewater-treatment plant has been in operation since 1991. It is used exclusively for treating wastewater obtained during the development and small-scale production of organic intermediates and crop-protection agents. The flow of true process wastewater is about 15 m^3/d, to which must be added ca. 60 m^3/d of slightly contaminated wash water, for a total of ca. 75 m^3/d. Characteristic features of this wastewater are the presence of crop-protection agents at the level of a few mg/kg, organic compounds (in some cases as emulsions) in the range of a few grams per liter of total organic carbon, and occasionally a very high salt content (several tens of grams per liter). The design of the plant was based on the following parameters:

Wastewater flow	
Process water	15 m^3/d
Wash water	60 m^3/d
Total	75 m^3/d (ca. 20 000 m^3/a)
BOD	
Load (estimate for design purposes)	200 kg/d
Concentration	2700 mg/L
TOC	
Load	225 kg/d
Concentration (90 % degradable)	3000 mg/L
Refractory (for estimating the activated-carbon requirement)	300 mg/L
Activated-carbon requirement	
Design	3 kg/m^3
Observed	1 kg/m^3

Process water and wash water are collected in separate distribution systems and pumped to the treatment plant. The process wastewater is collected in a total of three 40-m^3 tanks. The tanks are equipped with mixing nozzles and interconnected in such a way that a relatively high

degree of homogeneity is achieved, equivalent to a quantitative and qualitative smoothing-out over the course of an entire week. The washwater is held temporarily in a 60-m^3 tank. An attempt is made by suitable mixing of the wastewaters to achieve the highest possible degree of constancy in the organic loading of the treatment plant, as well as relatively minor fluctuation in the salt concentration.

In the first treatment stage (see Fig. 61), solids and sludges already present in the wastewater, as well as those formed as a result of treatment, are removed by neutralization (a), flocculation (b), and sedimentation (see Fig. 61).

The second stage consists of three tower-form bioreactors (c) connected in series, which are filled with an orderly packing made of plastic. The liquid flows upward through the first reactor and downward through the others. Reactors 2 and 3 can also be operated in parallel. Microorganisms fixed to the plastic surfaces degrade roughly 90 % of the TOC and at least 98 % of the BOD. Fixing the bacteria in this way makes it easier to cope with a high salt content while at the same time facilitating long-term retention in the system of special slow-growing bacteria. The latter, by degrading more persistent substances, reduce the load remaining for the subsequent activated-carbon treatment.

Wastewater discharged from the bioreactors contains free-floating bacteria, and this is pumped via two parallel crosscurrent microfiltration units (d) in an attempt to remove as much of the solid matter as possible.

The wastewater is next passed over two activated-carbon adsorbers (e) connected in series, in which traces of nondegradable organics and crop-protection agents still present are removed by absorption.

The treated wastewater effluent is monitored (f) continuously by an on-line TOC analyzer. All other wastewater parameters, including levels of crop-protection agents, are determined in the laboratory on the basis of composite samples.

Once the adsorption capacity of the granular activated carbon has been exhausted the charge is thermally reactivated by the supplier.

The bioreactors are backwashed approximately once a week with purified wastewater. The resulting mixture of biological sludge and water is collected in a tank, and from there dosed continuously into the influent of the coagulation stage. The sludge obtained is thickened (h) and transported in this form to another chemical wastewater-treatment plant where it is dewatered with centrifuges and incinerated in a fluidized-bed furnace.

The treatment plant operates largely automatically seven days a week, although staff are present only on the five workdays during normal working hours. Labor input for operation of the plant (excluding analysis) amounts to roughly one-half of a full-time employee.

The following performance data refer to experience gained during the first year of plant operation:

Wastewater processed:	
Lightly laden (TOC 50 mg/L)	60 m^3/d
Heavily laden (TOC 2000 mg/L)	15 m^3/d
Total (TOC 1400 mg/L)	75 m^3/d
Critical substances present (concentration)	1 – 100 (30 – 50 mg/L)
Extent of removal:	
TOC	98 %
COD	99 %
Critical substances	>99 %

2.1.4.6. Indirect-Discharge Facilities

Optimization of Indirect Wastewater Discharge by Henkel KGaA at its Düsseldorf-Holthausen Works, Germany [199–201]. The firm, with its principal production facility at Düsseldorf, is a major indirect discharger of chemical wastewater. The plant site, including production facilities, laboratories, administration buildings, storage areas, and connecting routes, comprises a total paved area of ca. 10^6 m^2.

Structure of the Site Sewage System. Wastewaters and precipitation are drawn from the entire complex via a highly segmented drainage system ca. 70 km in length. This could be described as a combined sewer system designed according to standard specifications. True wastewater from production operations is in most cases pretreated at the place of origin prior to discharge into the plant sewage system. All processes employed are designed to meet code requirements and other official administrative regulations. Before discharge into the sewagetreatment plant of the city of Düsseldorf, all waste-

water is first accumulated in a central wastewater-treatment facility located outside the boundaries of the actual site, from which it is subsequently drained off continuously into the municipal sewage system. Total wastewater flow from the plant amounts to ca. 7×10^6 m^3/a, most of it production wastewater. This figure includes ca. 300 000 m^3 of sanitary wastewater and ca. 400 000 m^3 of rainwater. The daily COD discharge averages 32 t at a COD/BOD_5 ratio of ca. 2.

Structure and Function of the Central Wastewater-Treatment Facility. This plant (see Figs. 62 and 63), which first went into service in 1978 and was considerably expanded in the late 1980s, consists of a screw pumping station [Fig. 63 (e)] with a total capacity of 6100 m^3/h, a screen (f) for the removal of coarse materials, two identical 1200-m^3 mixing and equalizing tanks (i) with continuous flow that are equipped for both neutralization and grease removal, and an additional wastewater-storage basin (a) with a capacity of 4800 m^3. All units for conveying wastewater have odor-tight coverings of glass-fiber reinforced plastic panels. The plant is controlled largely automatically with the aid of a process computer.

Figure 62. Overview of the Henkel KGaA central wastewater-treatment facility at Düsseldorf, Germany

Under normal circumstances all the mixed wastewater is conveyed through one of the two U-shaped mixing and equalization tanks (i). A central trough distributor makes it possible to allocate the influent wastewater uniformly through six inflow openings into the available tank volume. Variability in retention time is used to ensure optimum mixing and equalization. Grease, oil, and other floating matter not completely removed by the stream-splitting operation separate on the water surface. These are extracted by a skimming device and conveyed into special floating-sludge collection tanks (h). After removal of residual water the sludge is transported as necessary with the aid of a gantry dredger system into transport containers for conditioning with sawdust and subsequent transfer to a waste-incineration plant.

Maintenance of the required pH range in the mixing and equalization tanks is assured by introduction of acid or alkali through an automatic neutralization system. At the bottom of these tanks is a propellant system driven with filtered wastewater, which generates a uniform degree of turbulence, thereby keeping precipitated material in a state of suspension. The fluid level in the tanks is automatically held constant by adjustable overflow weirs at the tank outlets.

In normal service the second mixing and equalizing tank and the wastewater-storage tank are not charged with wastewater, but are nonetheless maintained in a constant state of readiness. In the event of a disruption in wastewater flow, a fire, or some similar emergency threatening to produce highly contaminated and perhaps toxic wastewater, up to 6000 m^3 of influent can be accommodated in these two tanks (Fig. 62).

Wastewater from the production facility is monitored on a continuous basis, both in the sewer system and in the influent and effluent of the central wastewater-treatment plant. Control stations are equipped with on-line analytical instrumentation for continuous monitoring of the organic load (TOC), biochemical oxygen demand (BOD), potential toxicity (bacterial toximeter), conductivity, pH, and anionic surfactant content. Processing of the on-line measurements is attended to by a "wastewater-control expert system," which takes into account not only a scientific database, but also previous experience gained in the course of plant operation. This system is capable whenever necessary of independently activating empty tank units or sounding an alarm.

The "Sewage-Treatment Plant – South" Maintained by the City of Düsseldorf. Effluent from the Henkel wastewater-treatment plant is connected via the so-called Düsseldorf – South Main Sewer, which also drains the southern parts of the city, with the Düsseldorf – South

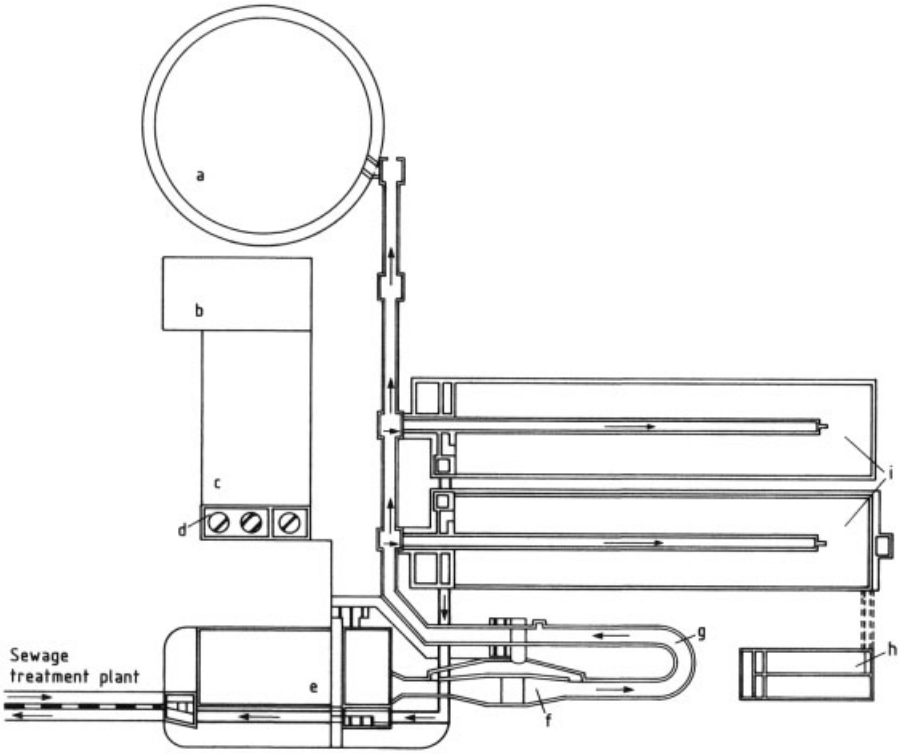

Figure 63. Schematic diagram of the Henkel wastewater-treatment facility
a) Storage basin (4800 m^3); b) Warehouse; c) Operations center; d) Tank farm; e) Wastewater lifting facility; f) Screening plant; g) Volume measurement; h) Floating-sludge tank; i) Flow-mixing tanks (1200 m^3 each)

sewage-treatment plant. The wastewater flow time from the Henkel delivery shaft to the municipal sewage-treatment plant is 3 – 4 h over a distance of ca. 8 km.

The sewage-treatment plant itself was designed as a typical single-stage mechanobiological plant, and it began operating in 1975. The dry-weather flow to this facility is roughly 100 000 m^3/d, with a daily COD load of 106 t. Therefore, from a volume standpoint the impact of wastewater from the Henkel works amounts to only about 18 %, but ca. 30 % based on pollutant load. One screw pumping station (consisting of 8 segments) supplies the plant with a maximum of 23 m^3/s of wastewater from which undissolved and mineral constituents are removed with the aid of coarse and fine screens and two circular degritters. A primary clarification stage (27 400 m^3) is followed by a biological-treatment stage (23 800 m^3) fed at a maximum rate of 2.85 m^3/s, into which air is introduced by surface aerators.

Since 1992, the final sedimentation unit (surface area 15 120 m^2, depth 2.70 m) has been accompanied downstream by a contact filtration plant (24 chambers with a capacity of 60 m^2 each, filter height 2 m), in which the concentration of total phosphorus in the treatment-plant effluent is constantly maintained below 1 mg/L by the addition of precipitants.

Because of increasingly stringent legal requirements applicable to the purification performance of the plant (e.g., Appendix 1 of the German Wastewater Administration Guidlines, size class 5) as well as factors related to sewage technology in the urban sector, considerable expansion of the wastewater-treatment plant has become necessary, affecting primarily the aeration basins and the secondary clarifiers. A future dry-weather inflow of ca. 144 000 m^3/d is anticipated, with an increase in the COD to

ca. 180 t/d, a total phosphorus increase of ca. 800 kg to 2200 kg/d, and a maximum total nitrogen load of 11 t/d. This demand is to be met by the installation of eight aeration basins for simultaneous nitrification and denitrification (240 000 m^3), equipped with diffusion aerators, together with nine secondary clarifiers (surface area 16 300 m^2, depth 4 m). In addition, the primary clarifiers currently in use will be partially rebuilt for conversion into a two-stage biological phosphorus removal plant. The existing sludge-incineration plant will also be replaced by a drying plant, and in the interest of odor control, individual segments of the plant will be covered and the exhaust air will be cleaned with compost filters.

Figure 64. Aerial view of the Miles Inc. anaerobic wastewater-treatment plant at Elkhart, Indiana

The Wastewater-Treatment System of the Miles Inc. (Bayer Corporation since April 1, 1995), Pharmaceuticals Division at Elkhart, Indiana, USA. The Miles Inc. pharmaceutical production operation pretreats all wastewater prior to discharging it into the City of Elkhart's municipal wastewater-treatment facility. The pretreatment system of choice in this case was a Bacardi Corporation "Anaerobic Filter Treatment System." This system removes biodegradable wastes while at the same time producing a methane-rich biogas that is captured and used as fuel for the Miles power plant.

Approximately 6625 m^3/d of chemical-processing wastewater is treated by the Elkhart facility's wastewater-treatment plant (Fig. 64). This plant was designed to remove two types of contaminants. The first is suspended solids, which can be eliminated by gravity settling. These solids include biomass from fermentation together with filter media, carbon, and solids from washings in the manufacturing area. The second type of contaminant removed is organic compounds, which would impart an oxygen demand on the receiving waters if left untreated. The organics are treated by the Bacardi process after solids have been removed by gravity settling in the primary clarifiers.

Primary (Organic-Rich) Waste System. Approximately 1540 m^3/d of organic-rich wastewater containing an average 27 200 kg of COD is processed through two anaerobic reactors (b) connected in series (see Fig. 65). The waste stream is equalized for flow and COD-concentration in two 3030 m^3 equalization tanks (a). Overflowing the first equalization tank into the second and operating the second tank at a 50 % level permits about three days worth of equalization to be achieved.

About 90 % of the water feed is directed to the first anaerobic reactor, while the remaining 10 % is passed into the second anaerobic reactor along with effluent from the lead reactor. The anaerobic reactors are 3030 m^3 tanks packed with a plastic, channeled medium to which the anaerobic microorganisms become attached. The combined recycle and feed is pumped to the top of the reactor and allowed to flow downward through the medium. The system removes about 90 % of the organic compounds present while producing approximately 14 200 m^3 of biogas per day, about 60 % of which is methane. The biogas is compressed to ca. 2.4 bar and pumped into a natural gas line used to fuel the boilers at the power plant. The biogas pipeline can be heated to prevent freezing of the condensables during cold weather. Should the gas recovery system happen to fail, a backup flare system burns the gas before releasing it to the atmosphere.

The design loading rate for the lead reactor was originally 8 kg of COD per cubic meter daily. This rate was later increased to 12 kg m^{-3} d^{-1} on the basis of operational experience. It was originally anticipated that the system would achieve a 60 % organic removal rate in each reactor, equivalent to an 80 % removal overall. Actual demonstrated efficiencies have been in the range 70 – 80 % in each reactor, for an overall removal efficiency of ca. 90 %.

Secondary Waste System. Approximately 6625 m^3/d (5110 m^3 of secondary waste com-

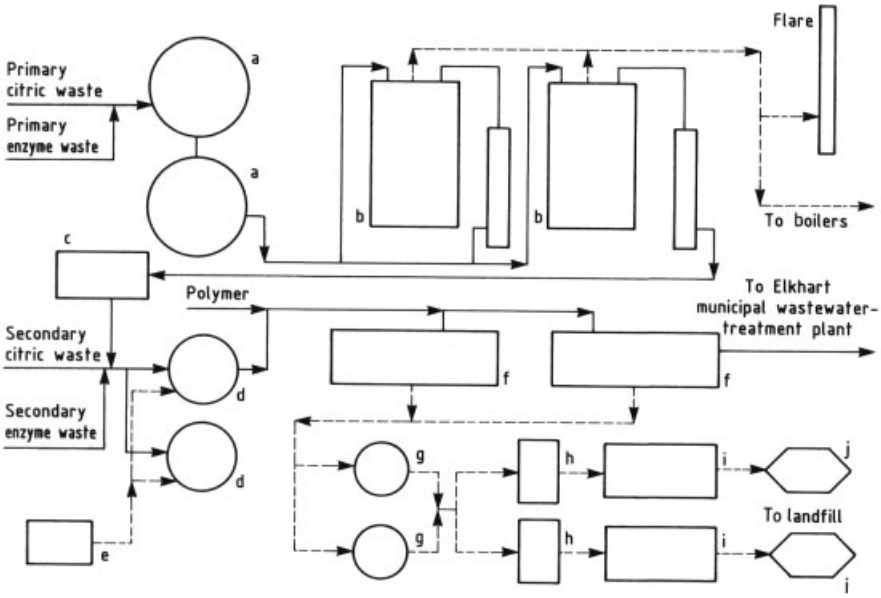

Figure 65. Schematic diagram of the Miles anaerobic wastewater-treatment plant
a) Equalization tanks; b) Anaerobic reactors; c) Chlorination tank; d) Neutralization tanks; e) Caustic tank; f) Clarifiers; g) Sludge-holding tanks; h) Mixing tanks; i) Belt filter presses; j) Sludge hoppers

bined with 1515 m^3 of anaerobic effluent), containing ca. 13 600 kg of suspended solids and 13 600 kg of COD, are treated in the secondary system. These waste streams are initially processed in two staged neutralization tanks (d), each with a capacity of 39 m^3. The pH is adjusted as necessary, although the incoming stream is usually neutral. Polymers are then added to enhance the solids-settling characteristics, after which the wastewater is clarified in two parallel 380-m^3 clarifiers (f). The clarified effluent passes into the interceptor system of the City of Elkhart for subsequent treatment at the municipal wastewater-treatment plant. Solids removed from the clarifier are pumped to two 227-m^3 sludge-holding tanks (g). The sludge is subsequently pumped to one of two belt filter presses (i) where, with the aid of a polymer flocculant, it is dewatered to produce a cake with a solids content of ca. 35 %. The latter is conveyed to hoppers (j) for transport to a local landfill.

In the course of the overall wastewater-pretreatment regimen, Miles' average discharge of COD and suspended solids is reduced from 41 000 and 14 500 kg/d, respectively, to 10 990 and 1370 kg/d. This pretreatment-system effluent meets all regulatory requirements of the City of Elkhart.

The entire process is monitored and controlled by a central control system. Four microprocessors receive the input and relay their output to various field instruments that in turn control flow rates, tank levels, etc. A series of active graphic display panels indicates current conditions at either of two operator stations in the control room. Any parameter falling outside the control limits is printed out in the control room or, if it is critical, announced over a loudspeaker by a computer-generated voice synthesizer.

Odor Control. The Miles chemical wastewaters contain high levels of sulfates, so extensive odor control is necessary to control emissions. Special equipment is used to add chlorine to the anaerobic effluent as a way of reducing microbial activity, which would otherwise result in the production of objectionable sulfur compounds. In addition, a 3400-m^3/min scrubber collects air from the covered clarifiers, sludge tanks, belt filter press area, and primary waste zone and

treats it chemically to remove odor-producing constituents.

2.2. Anaerobic Treatment

2.2.1. Introduction

In *aerobic fermentation* (e.g., yeast production, isolation of single-cell bacterial protein) the production of biomass is usually the most important single factor (→ Single Cell Proteins ; → Yeasts). Although a high degree of substrate conversion is the primary aim in aerobic wastewater treatment (i.e., the activated-sludge process), large amounts of biomass (clarification sludge) also form simultaneously because of the aerobic nature of the phenomenon. The fundamental advantage of an aerobic treatment is that oxidative degradation of the carbon substrates provides the energy required for propagation of the microorganisms that act as the biocatalyst. From the standpoint of achieving a high space – time yield the limiting factor is often the supply of oxygen, a gas that is only slightly water soluble.

In *anaerobic fermentation* it is the formation of product that usually occupies the spotlight (generation of methane, or of alcohols, ketones, and organic acids). The main advantage of biological fermentation as a method of synthesis is that inexpensive substrates are converted to valuable disproportionation products with the simultaneous formation of carbon dioxide. Here it is the energy that is usually limiting, resulting in relatively little growth of the catalytic microorganisms compared to the aerobic case. Since the microorganisms serve as both reaction product and catalyst, it is advantageous to maintain a high catalyst concentration under an anaerobic regime as well. The fact that energy is seldom a limiting factor in the case of the alternative aerobic fermentation must be paid for with a considerable energy expenditure in the course of providing the oxygen supply. The energy limitation associated with anaerobic fermentation is a function of the formation of energy-rich products.

In analyzing the microbial formation of methane, two important aspects stand out: the fact that the methane is derived from waste materials (clarification sludge and agricultural wastes), and its applicability in the treatment of industrial wastewater. Waste-material utilization in this sense amounts essentially to a form of biological stabilization (odor reduction, destruction of pathogenic microbes, dewaterability) in the interest of transforming the waste into a fertilizer. Biogas that arises as a byproduct is utilized in meeting the energy requirements of the associated bioreactors, where surplus gas is either flared off or burned as a source of heat for agricultural barns [202–205]. The anaerobic treatment of industrial wastewater has become increasingly important in recent years as a result of environmental-protection legislation, rising energy costs (which also have an impact on the cost of producing oxygen), and problems with the disposal of excess sludge formed in aerobic treatment processes. Efforts are now being made to remove organic contaminants insofar as possible without external sources of energy, taking advantage of the biogas produced, where the desired level of purification is ultimately achieved with the aid of a subsequent aerobic biological-clarification step.

2.2.2. Basic Principles

Developments in production technology have frequently resulted in the concentrated local accumulation of highly organic-laden wastewaters, a factor that must then be taken seriously into account with respect to disposal technology. With conventional aerobic wastewater treatment (the activated-sludge process), however, the greater the organic load, the more problematic the process becomes, because the required level of oxygen transfer rapidly approaches limits that are both technical and economic in nature. The key objective associated with the biogas process described here is thus the development of a mature alternative solution to the overall problem of wastewater, including water that is heavily contaminated with organic substances.

Efforts are currently being made to find a solution based not on increasingly elaborate aeration techniques, but rather on a concept already practiced in nature. *Biogas* forms wherever organic material decomposes microbially in the absence of air (i.e., anaerobically). This gas contains methane and carbon dioxide in a 1: 1 to 3: 1

ratio, together with hydrogen sulfide to an extent that depends on the sulfate content of the substrate, as well as traces of nitrogen and hydrogen. Whereas the activated-sludge process in principle represents a "cold (biological) combustion," the biogas process might be regarded as a "cold (biological) pyrolysis." Thus, anaerobic wastewater treatment represents an industrial application of a natural principle, one in which there is no trade-off between ecology (reduction of the pollutant load) and economics (energy-rich oxygen transfer is avoided, and a useful product is formed: biogas). Ecology and economics instead become mutually interdependent, and in an optimal way (effective contaminant reduction is accompanied by the production of large amounts of biogas).

The fundamental advantages of anaerobic wastewater treatment are:

1) Energy-intensive oxygen transfer is avoided.
2) The energy content of the resulting biogas can be put to further use.
3) Very little excess sludge is formed, while at the same time up to 95 % of the organic contami-nation is converted into a combustible gas (see Fig. 66), constituting a case of true disposal.
4) Heavy metals are not subjected to oxidative mobilization, but rather to reductive precipitation (as heavy-metal sulfides), facilitating their orderly disposal rather than wider dispersal.
5) Aerosol formation accompanying oxygen transfer is avoided, as is the stripping out of volatile components.
6) Space – time yields (bioreactor performances) reach a level fundamentally unattainable in the activated-sludge process because of the limitations imposed by oxygen transfer.

Given the same organic load, excess sludge formation is lower by a factor of 10 in an anaerobic process, because the exclusion of oxygen means that only ca. one-tenth as much metabolic energy is available to the microorganisms compared to an aerobic process. To be able to exist with a limitation of this kind, anaerobic microorganisms find themselves under tremendous selection pressure in favor of mutants with high metabolic activity (thereby compensating for a low energy yield by increased throughput). From the point of view of waste disposal, "high throughput" with little excess sludge formation is particularly interesting. Until now, anaerobic microorganisms have always been considered "sluggish" because of their low rate of multiplication, but this characterization ignores their high metabolic activity.

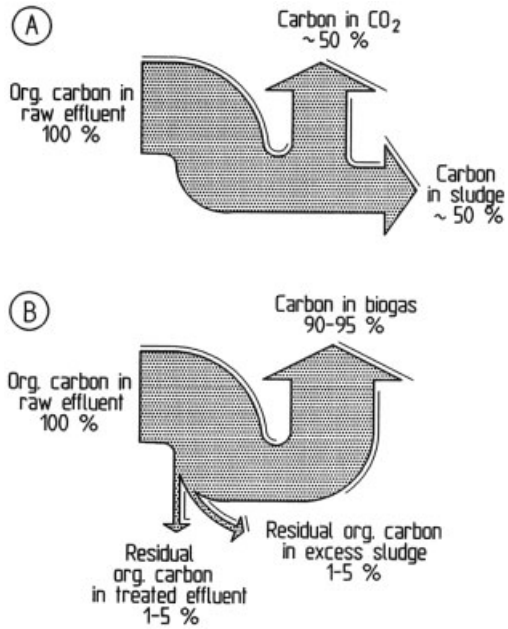

Figure 66. Carbon balances in the (A) anaerobic and (B) aerobic microbial degradation of organic compounds

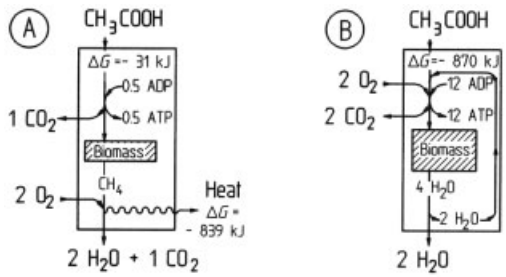

Figure 67. Energy balances in the (A) aerobic and (B) anaerobic degradation of acetic acid

Figure 67 uses the biological transformation of acetic acid – a central intermediate in anaerobic degradation – as a vehicle for comparing anaerobic and aerobic processes. Taking into ac-

count external combustion of the biogas formed in the anaerobic case, the limitations associated with the two approaches become comparable. In anaerobic degradation the products consist of a small amount of biomass and a combustible, gaseous substance. The aerobic case leads to more biomass, and the simultaneous generation of heat. In principle, the newly formed biomass is also combustible, but only after a considerable expenditure of energy for separation and drying.

Acquiring a more than empirical understanding of anaerobic wastewater treatment forces one to examine more closely the interactions between the various microorganisms in a food chain. The basic scheme is depicted in Figure 68. In simplified terms, at least three groups of microorganisms are involved in the degradation of complex organic molecules. To begin with, biopolymers are first hydrolytically degraded in the context of acidogenesis to give soluble monomers. This is followed by acid formation, with the simultaneous generation of hydrogen and carbon dioxide, a step referred to as acetogenesis. It is at this point that methane formation can commence. Indeed, tracer experiments have shown that ca. 70 % of the methane produced in the overall process comes from the degradation of acetic acid [206], whereas the remainder arises from the reaction of hydrogen with carbon dioxide. The complete degradation chain is in reality much more complicated, and requires a balanced cooperation among the various microorganisms present, ensuring that there will be neither a shortage nor an accumulation of intermediates (see Section 2.2.3).

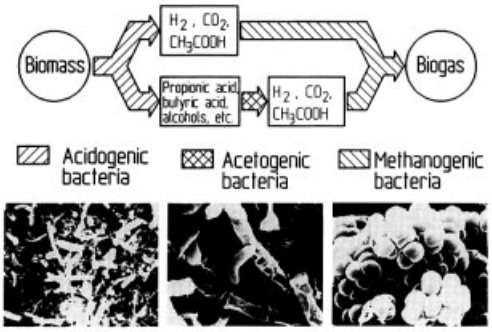

Figure 68. Schematic representation of the course of anaerobic methane generation from complex organic substances

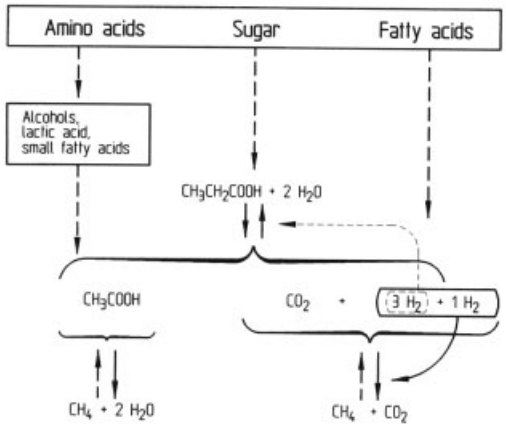

Figure 69. Methane formation from acetic acid and the combination of hydrogen and carbon dioxide: the influence of hydrogen on propionic acid degradation

Another erroneous preconception – that anaerobic microorganisms are only able to degrade "chemical plain fare" – has been refuted by results demonstrating that the spectrum of degradable compounds is very wide indeed, encompassing even chlorinated aromatics [207–210]. This is not surprising when one considers that the decomposition of waste in nature occurs largely in the absence of oxygen. Thus, the observed low rate of natural degradation is a function of the low rate at which the relevant organisms multiply, although it was long misinterpreted as being due to poor degradability of the waste itself.

The central problem from a reaction-engineering point of view is finding a way to achieve space – time yields that are high, as in the anaerobic case. The more complex the substrate mixture from a chemical structure and particle size standpoint, the more difficult it becomes to meet this challenge. The growth of anaerobic microorganisms is a function of numerous factors, including residence time, temperature, redox potential, pH, and nutrient composition. From a thermodynamic perspective, hydrogen is a particularly important "intermediate." If hydrogen utilization is not sufficiently effective, then the equilibrium position of the preceding steps is influenced in an unfavorable way. The kinetics of the most important individual steps (Fig. 69) are so far quite inadequately understood. This is in part because of the complexity of the sub-

strate mixtures usually encountered in practice, but also because collecting reliable kinetic data is a very time consuming process, and possible only in the case of continuous fermentation.

Since the generation time for the microorganisms involved is comparatively long, the residence time of the catalyst must be uncoupled from the residence time of the substrate if one hopes to increase the space – time yield. To this end, various approaches are available for retention and recycling of the catalytic biomass. Furthermore, biological systems react sluggishly to changes in the supply of substrate, so it is necessary that there be developed an "intelligent" control technique taking into account the various factors mentioned. This is the reason why it is important to discuss here certain fundamental principles of reaction engineering relevant to the anaerobic mode of operation, with special emphasis on biomass retention and recycling as a way of achieving high space – time yields. From the standpoint of wastewater treatment it is also advisable that easily separable solid particles be removed from the influent stream in order to avoid an unwanted accumulation of solids in the course of subsequent efforts to retain or recycle the biomass.

2.2.3. Thermodynamics

As shown in Figure 68, the anaerobic degradation of organic 'matter proceeds via a series of intermediate steps. In terms of both thermodynamics and kinetics, further degradation may be severely limited if intermediates are allowed to accumulate. The concentration of hydrogen takes on a regulative role by influencing the distribution of metabolic products formed by the fermentative bacteria [211–214]. This point is emphasized in Figure 68 and reinforced by Equations 2.1 – 2.5 (Scheme 1) involving the products of acidogenesis (acetic acid, propionic acid, and butyric acid) and the hydrogen produced from them.

While the decomposition of acetic acid is independent of the partial pressure of hydrogen, the degradation of propionic acid and butyric acid would be inhibited by a high partial pressure of hydrogen. At the same time, an adequate hydrogen partial pressure is necessary for the formation of methane from hydrogen and CO_2 (either directly or via acetic acid). This results in a narrow "thermodynamic window" within which the decisive reactions are able to proceed simultaneously (see Fig. 70). It follows that there must be a close correlation between microorganisms that form hydrogen and those that instead utilize it (*interspecies hydrogen transfer*). Hydrogen transfer presupposes a spatial proximity on the part of the relevant microorganisms. In fact, if the population density is inadequate, vigorous stirring can actually interfere with this synergistic effect.

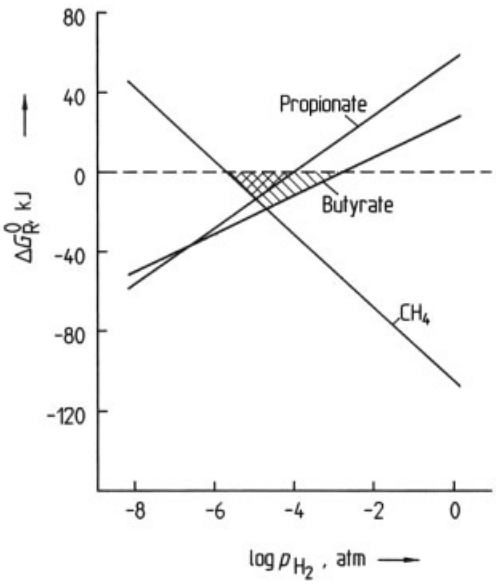

Figure 70. "Thermodynamic window" for simultaneous methane formation from propionic acid, butyric acid, and H_2/CO_2 as a function of hydrogen partial pressure ($T = 25°C$, pH = 7.0)

Figure 70 shows that propionic acid in particular is poorly degradable in the presence of even a low partial pressure of hydrogen. If hydrogen utilization in a mixed culture is not sufficiently high, accumulation is observed, especially of propionic acid (and the wastewater reactor "turns acidic"). Since the activity of methanogenic microorganisms decreases significantly at a pH below 5, a vicious circle is thus likely to result. Relevant data collected by LETTINGA et al. are presented in Figure 71 [115], which shows the acid concentrations at the out-

$$CH_3-COOH \rightleftharpoons CH_4 + CO_2 \quad (2.1)$$

$$CH_3-CH_2-COOH + 2H_2O \rightleftharpoons CH_3-COOH + CO_2 + 3H_2 \quad (2.2)$$

$$CH_3-CH_2-CH_2-COOH + 2H_2O \rightleftharpoons 2CH_3-COOH + CO_2 + 2H_2 \quad (2.3)$$

$$4H_2 + CO_2 \rightleftharpoons CH_4 + 2H_2O \quad (2.4)$$

$$4H_2 + 2CO_2 \rightleftharpoons CH_3-COOH + 2H_2O \quad (2.5)$$

Scheme 1. Products of acidogenesis

let of an anaerobic wastewater reactor in the first few days after startup. It will be noted that acetic acid and butyric acid become increasingly subject to degradation, whereas the concentration of propionic acid remains approximately constant. In such a case it is essential that the growth of hydrogen-utilizing microorganisms be encouraged by controlling the pH or prolonging the residence time to ensure adequate degradation of hydrogen and organic acids. The residence time can later be reduced again, using the partial pressure of hydrogen and the concentration of propionic acid as indicators of any threat of renewed overloading of the reactor (see Section 2.2.8).

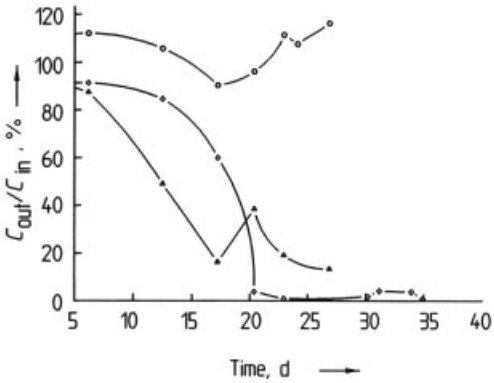

Figure 71. Relative distribution of lower fatty acids after startup of an anaerobic fermenter (based on [115]) triangles = acetic acid; circles = propionic acid; diamonds = butyric acid

In practice, however, the relationships are much more complex, because it is not only propionic acid already present that may serve as an input to acetogenesis, but also additional propionic acid derived from prior steps in the context of acidogenesis. The equations below show that fermentative conversion of carbohydrates in the presence of hydrogen is directed mainly toward propionic acid, because this step is linked to the greatest change in free energy (ΔG_R^0: -358.1 kJ/mol).

$$C_6H_{12}O_6 + 4H_2O \rightarrow 2CH_3COO^- + 2HCO_3^- + 4H^+ + 4H_2$$
$$\Delta G_R^0: -206.3\,\text{kJ}$$

$$C_6H_{12}O_6 + 2H_2 \rightarrow 2CH_3CH_2COO^- + 2H_2O + 2H^+$$
$$\Delta G_R^0: -358.1\,\text{kJ}$$

$$C_6H_{12}O_6 + 2H_2O \rightarrow CH_3CH_2COO^- + 2HCO_3^- + 3H^+ + 2H_2$$
$$\Delta G_R^0: -254.6\,\text{kJ}$$

$$C_6H_{12}O_6 + 2H_2O \rightarrow 2C_2H_5OH + 2HCO_3^- + 2H^+$$
$$\Delta G_R^0: -225.9\,\text{kJ}$$

$$C_6H_{12}O_6 \rightarrow 2CH_3CHOHCOO^- + 2H^+$$
$$\Delta G_R^0: -198.3\,\text{kJ}$$

2.2.4. Kinetics

Anaerobic wastewater treatment can be accomplished under either *mesophilic* or *thermophilic* conditions. The advantages and disadvantages of the two alternatives are indicated in Table 9. With regard to activity of the microorganisms, thermophilic treatment offers no general

advantages. In fact, stability is greater in the mesophilic mode of operation. Although thermophilic fermentation does offer certain advantages from the standpoint of wastewater sanitization this is offset by an increased thermal requirement for preheating the substrate. However, thermophilic reaction conditions can be considered for wastewater that has a temperature at the bioreactor inlet of at least 60°C. Thermophilic conditions apparently result in the selection of a narrower spectrum of microorganisms, leading to more clearly defined but less stable conditions. Even though mesophilic conditions are accepted most widely today, the question of temperature selection still cannot be clearly answered, because results presented in the literature are frequently contradictory [216–224]. These discrepancies are caused not only by the various types of wastewater investigated, but also by a still inadequate understanding of thermophilic microorganisms.

Table 9. A comparison between mesophilic and thermophilic anaerobic fermentation

Characteristic	Mesophilic *	Thermophilic *
Activity	o	o
Stability	+	−
Sanitization/ "sterility" **	o	+
Thermal demand	+	−

* + = advantageous; − = unfavorable; o = average.
** Sanitization = elimination of pathogens; "sterility" = reduction of living-cell numbers.

In practice, only continuously operated systems can be used for kinetic measurements, because batch-reactor experiments (shaken cultures) provide results that are poorly reproducible. This is due to the fact that in a batch experiment the reaction conditions change constantly over the course of the test period resulting in the appearance and disappearance of various limitations. In a chemostatic operating mode the reactor is maintained at a steady state with respect to outlet conditions. The influence of various parameters can thus be studied by varying either the residence time or the substrate composition. The basic procedure for collecting kinetic data can be illustrated by using as an example the anaerobic degradation of acetic acid with a pure culture of *Methanosarcina barkeri*. This example is also of practical importance, because most biogas comes from the degradation of acetic acid, and this reaction is regarded as the limiting step. The extent of the kinetic data available for anaerobic treatment of wastewater itself with pure cultures is still much too small to be of use.

In studying the growth of *M. barkeri* (DSM 804), acetic acid was used as the sole source of carbon, with a defined mineral growth medium of the composition specified in the caption to Figure 72. An elemental analysis of *Methanosarcina barkeri* growing on this medium revealed carbon (46.85 %), nitrogen (12.16 %), hydrogen (6.69 %), sulfur (1.05 %) and phosphorus (0.91 %). Equation 2.1 for the methanation of acetic acid shows that a pH shift should be expected during the conversion process. Since optimal growth of this particular microorganism is possible only in a very narrow pH range (Fig. 73), it is necessary that the pH be held constant. Making a definitive statement regarding the course of the pH requires an examination of the dissociation equilibria for acetic acid and carbonic acid:

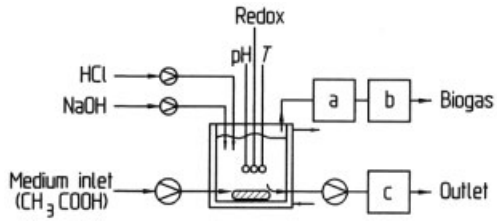

Figure 72. Schematic representation of the experimental setup for obtaining kinetic data regarding the anaerobic degradation of acetic acid by *M. barkeri*
a) Gas flow meter; b) Gas analyzer (CH_4, CO_2); c) Analysis of acetic acid and biomass
Growth medium (concentrations in parentheses in mmol/L): Acetic acid (1000), Na^+ (35); NH_4^+ (10), K^+ (5), Ca^{2+} (3), Mg^{2+} (2), S^{2-} (1.25), $HPO_4^{2-}/H_2PO_4^-$ (1), L-Cysteine (0.85), Fe^{2+} (0.55); Trace elements: Co, Ni, Mn, Se, Zn; Mo (<10 µmol/L); Vitamins: Wolfe's vitamins (≤1 µmol/L); pH_R (pH in the reactor) = 6.3; T = 37°C; V = 5 L [225]

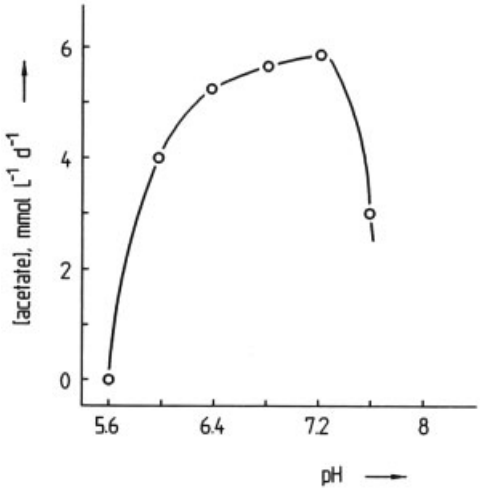

Figure 73. Relationship between acetate degradation and pH in a culture of the mesophilic species *Methanosarcina barkeri*

$$CH_3COOH \rightleftharpoons CH_3COO^- + H^+$$

$$CO_2^g \rightleftharpoons CO_2^{aq} \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons 2\,H^+ + CO_3^{2-}$$

It is clear that the pH should increase with increasing degradation of acetic acid and decrease with the formation of CO_2. In addition, one must also take into account the solubility equilibrium of carbon dioxide. Figure 74 A shows the total concentration of carbonic acid (i.e., the sum of dissolved CO_2, hydrogencarbonate, and carbonate) as a function of pH. With increasing pH the concentration increases rapidly as a result of a shift in the dissociation equilibrium in favor of hydrogencarbonate and carbonate. The extent of CO_2 dissolved physically was calculated according to Henry's law for a partial pressure of 0.45 bar at 37°C, assuming that the saturation concentration is independent of pH.

$$[CO_2] = [CO_2\ \text{aq.}] + [HCO_3^-] + [CO_3^{2-}]$$

$$[CO_2\ \text{aq.}] = 11.4\ \text{mmol/L};\ T = 37°\text{C},\ P_{co^2} = 0.45\ \text{bar}$$

With the help of the dissociation constants for the various species involved, and taking into account the solubility of carbon dioxide, the expected course of the pH can be calculated. Figure 74 B shows a set of corresponding pH – conversion curves determined for various starting pH values for the conversion of a 0.3-mol/L acetic acid solution. If one assumes a conversion of 90 % and a pH optimum of 6.4, then in the case of an 0.3-mol/L solution the reaction should be initiated with a pH of 4.0 at the inflow in order to reach the pH optimum of 6.4 in a continuously operated stirred-tank reactor or in a reactor sys-tem with complete backmixing.

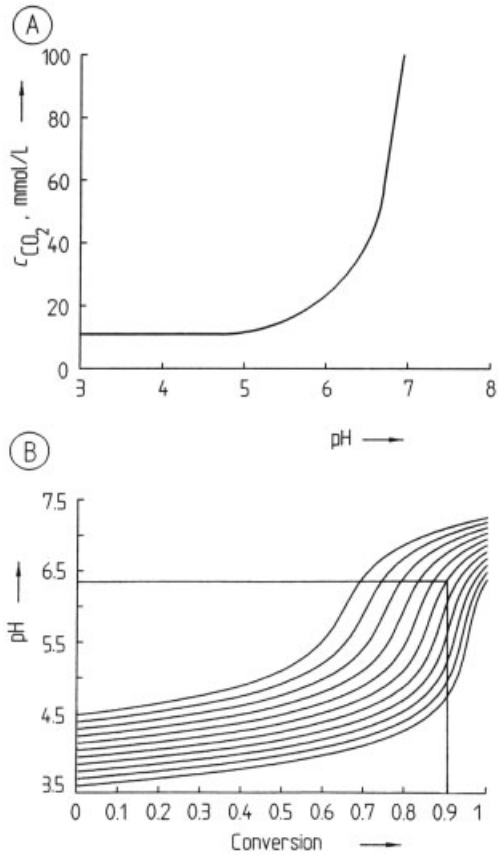

Figure 74. A) Overall concentration of carbonic acid as a function of pH ($T = 37°$C, $C_{CO_2\,(aq)} = 11.4$ mmol/L, $p_{CO^2} = 0.45$ bar) [226], [227]
B) A series of pH – conversion curves for acetate degradation ($S_0 = 0.3$ mol/L; $T = 37°$C; $p = 1$ bar; no buffer) [226], [227]

A schematic representation of the experimental arrangement used in this study is shown in Figure 72. The stirred-tank reactor was equipped with pH-measuring and control devices, a redox indicator, two coordinated pumps

for substrate dosing and product withdrawal, thermostatic control, a thermal mass-flow meter (a), and a gas analyzer (b) for methane and carbon dioxide. Substrate flow was determined with the aid of a balance.

To establish a set of model parameters it was necessary to solve simultaneously the substrate and the biomass balances for a stationary state in a continuous stirred-tank reaction without biomass retention:

Mass balance for the substrate (S = substrate concentration; X = cell dry-mass concentration):

$$0 = -\frac{1}{\tau}(S_0 - S) + \frac{\mu_{\max}}{Y_{\max}} \cdot \frac{S}{K_m + s} \cdot X + mX$$

Mass balance for the biomass:

$$0 = -\frac{1}{\tau} \cdot X + \mu_{\max} \cdot \frac{S}{K_m + s} \cdot X$$

Model parameters (acetate degradation by *M. barkeri*):

- $\mu_{\max}$ = maximum growth rate = 0.206 d^{-1}
- $Y_{\max}$ = maximum growth-yield coefficient = 1.533 g/mol
- K_m = Michaelis – Menten constant = 4.01 mmol/L
- m = maintenance factor = 2.88 mmol/L
- $t_{2/1}$ = generation time = 80.66 h
- τ_{crit} = critical residence time = 116.84 h
- $\tau_{\max}$ = residence time for maximum space-time yield = 124.24 h

Reaction conditions for continuous measurement:
S_0 = initial substrate concentration = 1000 mmol/L
pH = 6.3
T = 37°C

This leads to the following equations for the two variables X and S, obtained by solving for the biomass concentration X and the substrate concentration S:

$$X = f(\tau) = \frac{S_0 - S}{\tau \cdot \left(\frac{\mu_{\max}}{Y_{\max}} \cdot \frac{S}{K_m + S} + m\right)}$$

$$S = f(\tau) = \frac{K_m}{\mu_{\max} \cdot \tau^{-1}}$$

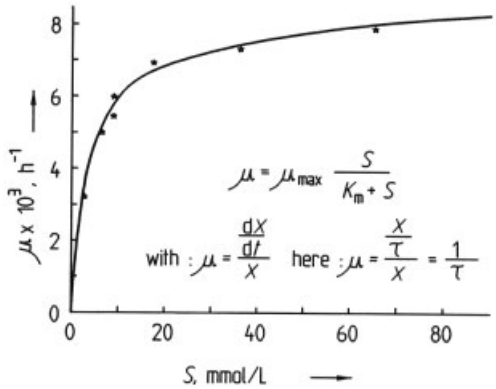

Figure 75. Maximum specific growth rate $\mu_{\max}$ established from a plot of the observed growth rate μ vs. the stationary substrate concentration S, where X = cell dry-mass concentration; $\mu_{\max} = 8.593 \times 10^{-3}$ h, $K_m = 4.01 \times 10^{-3}$ mol/L [228]
—— Calculated; ∗ Measured

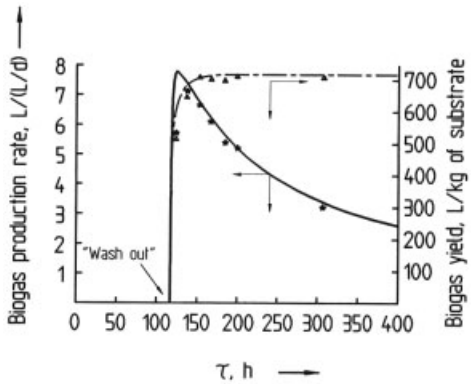

Figure 76. Comparison between measured and observed volume-specific gas formation (space – time yield) and biogas yield in the degradation of acetic acid (S_0 = 1 mol/L, τ_{cr} = 117 h) [228]

Simple Michaelis – Menten kinetics were assumed for the growth of the microorganisms. The model parameters $\mu_{\max}$ (maximum growth rate), $Y_{\max}$ (maximum growth – yield coefficient), K_m (Michaelis – Menten constant), and m (maintenance factor) above were obtained by nonlinear regression from seven different experiments, each involving a different residence time, waiting always to make certain that stationary conditions had been established. The concentrations of acetate and biomass and the

formation of biogas were measured at equilibrium ($dS/dt = 0$, $dX/dt = 0$). Based on the measured quantities it was possible to achieve a 100 % closed-carbon balance. The generation time ($t_{2/1}$), the minimum residence time (τ_{cr}) leading to biocatalyst being washed out, and the residence time corresponding to a maximum space – time yield (τ_{max}) are derived quantities arising from this model.

The maximum specific growth rate μ_{max} and the Michaelis – Menten constant K_m can be determined by nonlinear regression by plotting the specific growth rate μ (in the form of the reciprocal of the residence time in the stationary state) against the stationary substrate concentration S (Fig. 75). With the help of the model parameters, the volume-specific biogas production rate and the biogas yield are also subject to calculation as a function of residence time. Calculated results are compared with the corresponding experimental data in Figure 76. It will be noted that the residence time for a maximum space –time yield is exceptionally predictable with an adequate degree of accuracy. It is also clear that even a slight variation in the residence time can lead to washing out of the microorganisms. Residence times that are only slightly larger than the critical residence time τ_{cr} results in a high biogas yield, because the K_m value for acetate (see above) is only 4 mmol/L, so a zero-order reaction is observed over most of the conversion range. The carbon balance associated with conditions of maximum space – time yield is further elaborated in Figure 77. In principle this represents an example of substrate carbon selectivity. Of the converted substrate carbon, only 3.12 % is incorporated into new biomass, while about 97 % is recovered in the biogas. Most of the latter is formed in the course of providing sufficient ATP for carbon incorporation (growth). Additional ATP is required to assure energy metabolism (maintenance). The resulting numbers clearly reveal the energy-limiting nature of the anaerobic fermentation process, but also the fact that true "disposal" occurs in the course of anaerobic wastewater treatment, since roughly 97 % of the carbon source is converted to biogas.

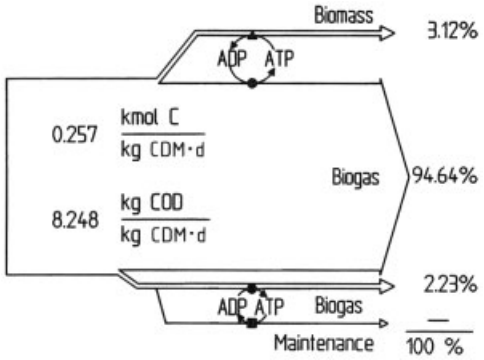

Figure 77. Selectivity in the utilization of substrate carbon, as illustrated by the example of aerobic methanization of acetic acid at maximum space – time yield [228]
$\tau = 124.24$ h; yield = 94.06 %; $x = 1.41$ g/L ● Biomass; △ Growth; ■ Maintenance

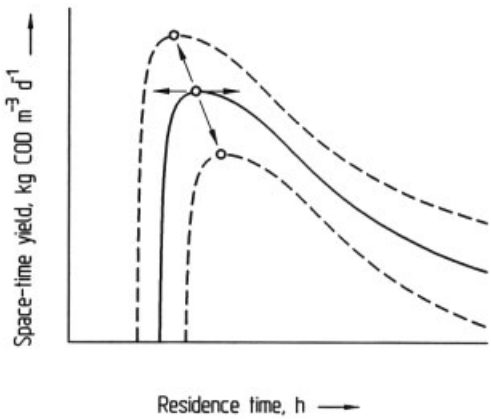

Figure 78. Optimization of selection stress by maximizing the space – time yield

A further increase in biomass activity could be achieved by successively reducing the residence time in the chemostats. This fundamental procedure also illustrates the basic idea behind the concept of selection-stress optimization by maximizing the space – time yield (Fig. 78), thus permitting an optimally adapted pure or mixed microbial population with maximum activity to be obtained in the shortest possible time [229]. The residence time should continue to be reduced so long as the space – time yield increases.

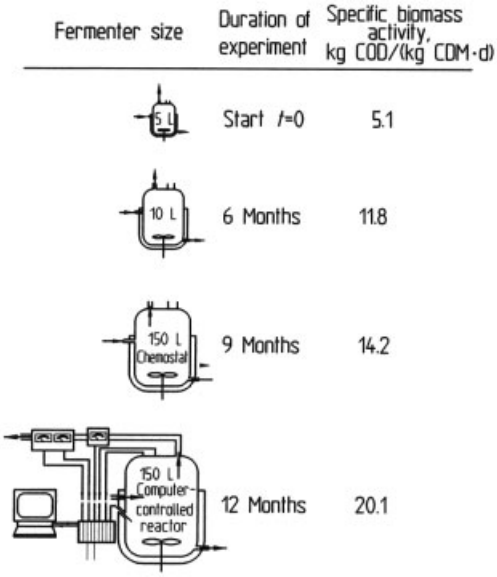

Figure 79. Increase in biomass activity [kilograms of COD per kilogram of cell dry mass (CDM) per day] of methane-producing bacteria through manual and computer-controlled reduction in the residence time

The residence time must be increased slightly if washing-out occurs – and not simply with respect to less active parts of the population, but for parts with higher activity as well. Continuous attempts should then in any case be made to increase the flow once again as soon as the population has recovered. Figure 78 illustrates the changes that may occur with increasing or decreasing residence time (horizontal arrows). A process computer outfitted with a control strategy for flow adjustment represents the best possibility for actually carrying out such an iterative approach to stress optimization (see Section 2.2.8). It has proven feasible in this way within six months to more than double the specific activity of a conventional enrichment culture relative to a 5-L or 10-L stirred-tank reactor. Over the course of another three months this value was improved further to 14.1 kg of COD per kilogram of cell dry mass (CDM, biomass) per day by manual adjustment (Fig. 79). Finally, after another three months the value had been increased to 20.1 kg COD/(kg CDM · d) with the aid of computerized set-point control. Compared with conventional experiments in a 5-L chemostats, the overall specific activity of the methanogenic population was ultimately quadrupled.

2.2.5. Mass and Heat Transfer

The anaerobic treatment of wastewater occurs essentially without any heat of reaction, which means that heat-exchanging surfaces are unnecessary. Nevertheless, since a bioreactor operates above ambient temperature, insulation of the outer walls is advisable. In fact, simply warming the substrate up to reaction temperature meets the primary heat requirement. To some extent it is possible to utilize the heat capacity of the effluent flow for heating purposes. Heat losses can be compensated without difficulty by combustion energy derived from the biogas produced. Thus, operation of an anaerobic wastewater-treatment plant is at least self-sustaining from an energy standpoint.

A biogas plant becomes all the more self-sufficient in energy the higher the organic-contaminant level of the wastewater and the higher its initial temperature. At a wastewater temperature of 10°C the concentration of organic components, expressed as COD (chemical oxygen demand), must be at least 7 kg/m^3. At a temperature of 30°C a COD level of 2 kg/m^3 is sufficient to cover the process-energy demands. Additional biogas produced at higher levels of organic contamination can, after desulfurization, be burned as necessary in a gas boiler for direct generation of steam.

Since no gaseous substrate (oxygen) is required, problems related to creating a large surface for material exchange are avoided. However, it is necessary that provision be made for the rapid removal of gaseous reaction products to ensure that as little of the "catalyst surface" as possible is blocked by gas bubbles. As a result of the residence times involved and the difference in density between the emerging biogas and the liquid phase (which provides a stirring effect), mechanical stirring aimed at mixing the liquid phase can be more or less dispensed with. When mechanical stirrers are used their primary purpose is to ensure homogeneous distribution of the catalyst or the suspension of solids. The energy requirement associated with fermenter circulation is 5 W/m^3.

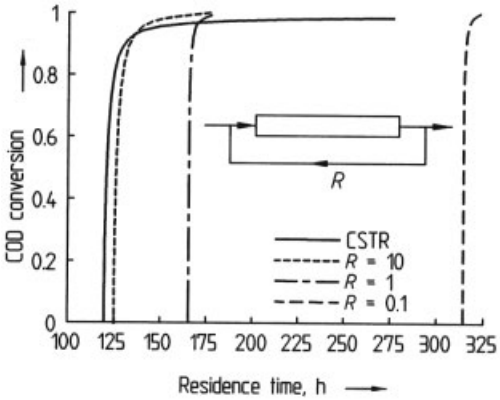

Figure 80. Yield as a function of residence time in a backmixing reactor (tubular reactor with a relative return flow *R*) [227]
CSTR = Continuous stirred-tank reactor.

2.2.6. Reactor Design

Fermentation displays typical autocatalytic behavior (i.e., a portion of the product functions simultaneously as a catalyst), and since K_m values for the most important substrates are very small, every effort must be made at least with soluble substrates to approximate stirred-tank conditions. The literature includes both theoretical treatments and experimental examples suggesting the applicability of plug-flow conditions, but most of this information relates to secondary effects (biomass retention, partial utilization of solid substrates). Plug-flow behavior has also been recommended for the reduction of product inhibition [230], but there have to date been no convincing experimental applications.

The superiority of the stirred-tank reactor or a similar reactor system with complete backmixing has been clearly documented by simulated data based on the kinetic constants for *M. barkeri*, plotted in the form of conversion vs. residence time (Fig. 80). It is immediately obvious that the stirred-tank reactor offers the lowest reaction volume up to high levels of conversion. In the case of the backmixing reactor with low return flow ($R = 0.1$), the same conversion requires a residence time that is longer by a factor of 3. As the extent of back mixing increases, the conversion behavior approaches that of a continuous stirred-tank reactor (CSTR). Achieving approximately the same conversion, however, requires a return-flow ratio of 10: 1. It is nevertheless remarkable that even higher conversions can be attained with a still-finite return-flow ratio and a reasonable residence time (e.g., $R = 10$ with a 150-h residence time, as shown in Fig. 80).

Another approach to reducing the reaction volume entails the use of reactors connected in series, with a tubular reactor serving as the second stage. The stirred-tank reactor should then be operated in such a way that conversion in the first stage corresponds to the maximum value along the reaction-rate curve, equivalent to a minimum in the curve presented in Figure 81. However, given the pronounced substrate affinity of methanogenic microorganisms, the reaction is zero order over most of the conversion range relevant to actual practice (>90 %), so the effort required for the second reaction rules out any commercial interest. Strictly speaking, however, these observations apply only to a continuous stirred-tank reactor in which there is no biomass retention, because – as described later (see Section 2.2.7.6, "Influence of Residence – Time Distribution on the Microbial Degradation Kinetics") – the maximum space – time yield in a reactor system with biomass retention occurs at a significantly lower level of COD conversion (70 – 85 %), and use of a cascade leads to a marked reduction in the required reaction volume.

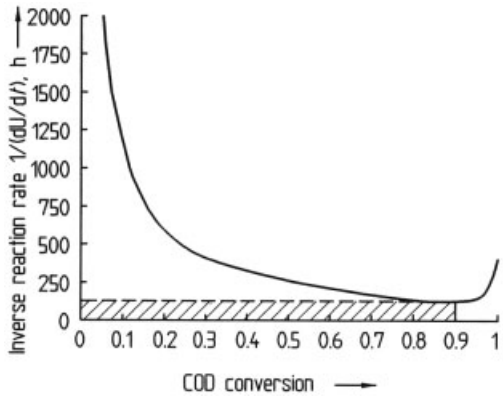

Figure 81. Inverse of the reaction rate plotted against COD conversion for a stirred-tank reactor [227]

As a result of stepwise degradation involving various bacterial populations, together with the complex interactions linking fermentative acid formation with methanogenesis, it has been suggested that these two individual steps be separated in space by the use of a multistage reactor combination [231–233]. The most widely applied variant is a two-stage mode of operation (Fig. 82), in which acidogenic microorganisms are most active in the first stage and mainly methanogenic microorganisms in the second [234–239]. The latter is usually implemented in the form of an upflow anaerobic sludge blanket (UASB; see Section 2.2.7.2) or fixed-bed reactor (Section 2.2.7.3). Single-stage operation typically entails a lower initial investment than a two-stage system, presumably the most important reason for the proliferation of single-stage commercial facilities in the 1970s and early 1980s. Today, however, the trend in anaerobic treatment of highly contaminated wastewater is clearly toward a two-stage process design. A comparison of these two variants based on the example of anaerobic treatment of wastewater from ethanol distillation (spent wash) is discussed in Section 2.2.7.6. The generation time of microorganisms in the first stage is clearly shorter than that in the second stage, so incorporation after the first stage of a sedimenter with partial biomass return is sufficient to maintain stable conditions in the first stage. Nevertheless, in a large-scale operation it is extremely important to ensure effective separation of preacidification biomass from the inlet of the second stage to prevent solids loading either in the methanogenic stage or at the outlet of the plant complex as a whole.

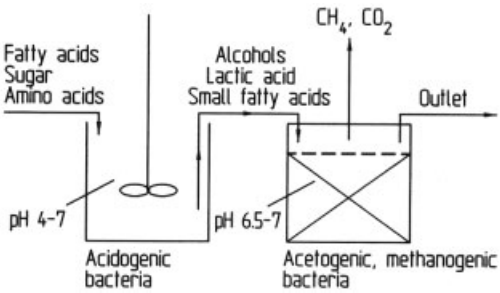

Figure 82. Schematic representation of a two-stage anaerobic fermentation with acidification followed by methanization

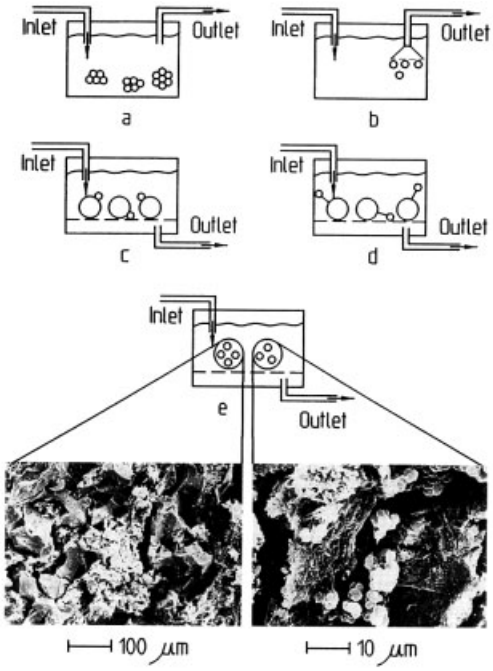

Figure 83. Schematic representation of various approaches to biomass retention
a) Flocculation – pelletization; b) Filtration; c) Immobilization through adsorption; d) Immobilization through covalent bonding; e) Immobilization through colonization on a macroporous carrier

2.2.7. Biomass Retention/Recycling

Achieving high performance in a reactor is not simply a function of the biomass activity. Because of the autocatalytic factor the biomass concentration is important as well. The biomass concentration cannot be adjusted independently of the wastewater residence time in the bioreactor. Given the long generation time of anaerobic microorganisms, important components of the mixed population are always lost along with the partially treated water even at a residence time of several days if no precautions are taken with respect to biomass retention or external separation and return. In general, this requires increasing the particle size and/or the density difference relative to the liquid phase.

The most important methods for decoupling the residence times for liquid substrate and biomass in anaerobic wastewater treatment via

either internal retention or external separation and recycling can be summarized as follows:

A. Biomass retention (internal)
1. Sedimentation by pellet formation
1.1 Upflow Anaerobic Sludge Blanket (UASB) [239–244]
2. Filtration
2.1 Membrane Anaerobic Reaction System (MARS) [245]
2.2 Rotor-fermenter [237], [246]
3. Immobilization by adsorption or colonization of open-pored carrier materials
3.1 Fixed-bed reactor [247–256]
3.2 Fixed-bed loop reactor [257–268]
3.3 Anaerobic film reactor [269–278]
3.4 Fluidized-bed reactor [245], [279–293]
3.5 Hybrid concepts (UASB/fixed bed) [294–297]
3.6 Stirred-tank reactor with suspended carriers [298], [299]
4. Immobilization by inclusion or covalent bonding [300], [301]

B. Biomass recycling (external separation)
5. Sedimentation by chemical or physical separation
5.1 Anaerobic contact process (flocculation/lamella separator) [302–307]
5.2 Centrifugation
6. Flotation [308], [309]

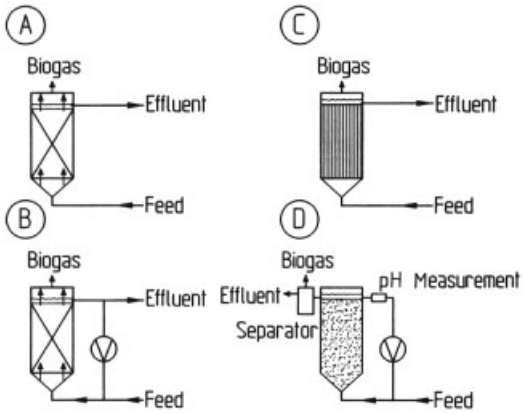

Figure 84. Reactors for the immobilization of microorganisms in anaerobic wastewater treatment
A) Fixed-bed reactor; B) Fixed-bed loop reactor; C) Fixed-film reactor; D) Fluidized-bed reactor

Various approaches to biomass retention are illustrated in Figure 83.

A selection in favor of either flocculating or pellet-forming microorganisms that sediment easily in a bioreactor is still accomplished in most cases by empirical optimization of the reaction conditions. This principle of biomass retention is the basis for the upflow anaerobic sludge blanket (UASB) reactor [240–244]. It is difficult to reproduce pelletization, however, and there is also the risk that the biogranulates will subsequently redissolve. Free biomass or pelletized biomass can be separated with the help of filtration techniques, depending on the pore size of the filter [227], [245], [246], although such filters easily become clogged. Moreover, activity losses must be anticipated due to the mechanical shear stress involved in recirculating a microbial suspension. The practical applicability of this approach to wastewater reactors is thus severely restricted. At present, microfiltration is not utilized anywhere in large-scale anaerobic wastewater treatment.

One way to achieve an increase in the particle size, with a corresponding improvement in the separability of the liquid phase, is to immobilize the biomass on a carrier material. Evidence for the potential of this approach to microbial retention is available from nature, because most microorganisms in the environment are present in a fixed state. Nevertheless, adsorption presents the risk that the biomass might become detached as a consequence of shear forces. Efforts to counter this obstacle include binding biomass covalently to a carrier and enclosing it in a gel. The method has so far resulted in essentially no practical wastewater-treatment applications, however, because of frequent simultaneous inactivation of the biomass and the unacceptably high costs involved in covalent fixation. The colonization of macroporous carriers with high surface areas lowers the risk of biomass detachment and circumvents the cost entailed in covalent fixation, while at the same time substantially increasing the amount of biomass that can be accumulated per unit volume. Proper adjustment of the pore size of the carrier material permits the flow partially to penetrate the interior of the particles without also detaching the biomass. This has the further advantage that it facilitates the rapid removal of biogas as it is formed. A summary of common reactor configurations based on immobilized microorganisms is provided in Figure 84 [245], [247–293].

In the case of external biomass recycling, sedimentation is achieved either by flocculation induced by chemical additives, or by increasing the separation area with the help of lamella separators. This technique is also known as the *anaerobic contact process* [302–307]. It is also possible to increase the density difference by flotation, or to increase the gravitational field by centrifugation. The principles of exter-

nal biomass recycling are illustrated schematically in Figure 85. Practical bioreactor designs for the anaerobic treatment of wastewater are discussed in the sections that follow.

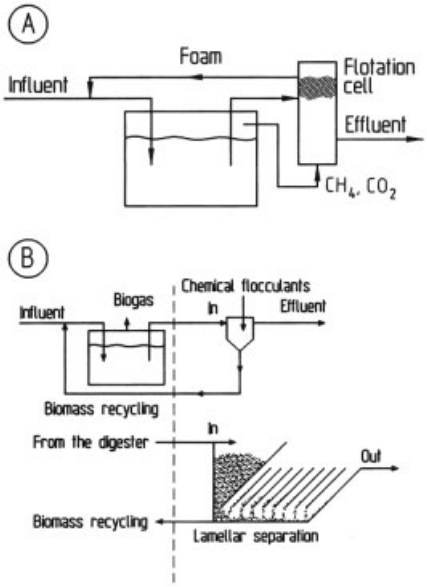

Figure 85. Schematic representation of methods for biomass recycling in the anaerobic contact process
A) Foam removal through flotation; B) Separation by external chemical or physical sedimentation

2.2.7.1. The Anaerobic Contact Process

This particular variant is based on external sludge separation and return to the methane reactor, and it attracted attention for practical wastewater treatment at an early date because of its limited technical requirements [302], [303]. The natural settleability of microorganisms is in this case increased either by the addition of organic flocculation aids or through the use of lamellar surfaces. Both approaches are employed in industrial applications. Continuous centrifugal separators might in principle be utilized as well, but this has not yet been implemented in practice. A microorganism concentration of 5 – 10 g/L (based on solids) is achievable through biomass recycling. Two major problems arise in the recycling of flocculent sludge, however:

1) Suspended sludge tends to form at sludge loads greater than 1 kg COD/(kg CDM · d).
2) Sedimentation in the settling tank is greatly impeded by secondary fermentation and gas formation. One remedy involves keeping the temperature low in the settling tank; alternatively, the sludge must be degassed under vacuum before entering the settling tank [310].

As a result of a low biomass concentration and low sludge load, the maximum achievable space – time yield is 3 – 5 kg COD m^{-3} d^{-1}. The superficial velocity in the sedimentation tank is ca. 0.5 m/h. Large plants based on this type of reactor have been built mainly for processing concentrated wastewater from the food industry (ANAMET process [306], [307]). About 80 large-scale contact-process plants are currently in operation worldwide. A major disadvantage is the considerable space requirement. Figure 86 is a photograph of a plant using the ANAMET process and built for use in the sugar industry. The lamella separators are clearly visible in the foreground.

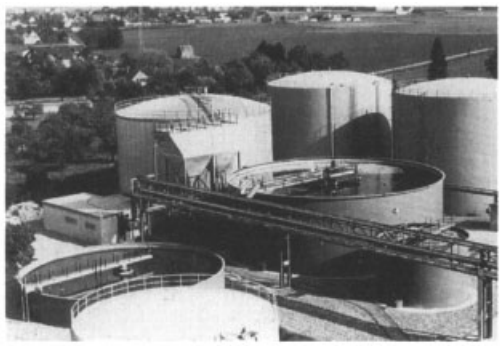

Figure 86. Photograph of an anaerobic treatment plant based on the contact process and utilized in the sugar industry (COD load: 15 000 kg/d)

2.2.7.2. Upflow Anaerobic Sludge Blanket (UASB) Reactors

A simplified flow diagram of the UASB process is presented in Figure 87. Wastewater flows vertically through the system starting at the bottom. Microbial pellets sediment on the reactor floor, forming a sludge bed. Enrichment of the biomass in the reactor is made possible by a three-phase separation system designed especially to minimize any disruptive influence on the part of biogas during sludge separation. Apart from granulate sludge, a considerable amount of flocculated

sludge is formed as well, which in turn constitutes a basis for the formation of new pellets.

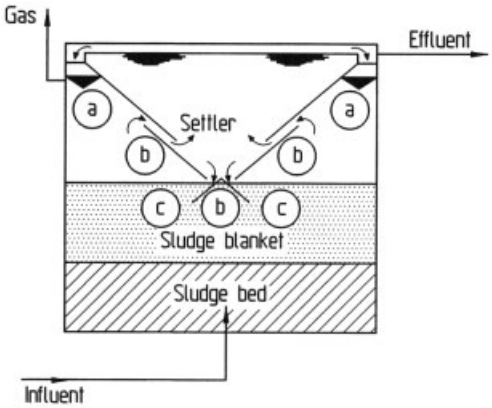

Figure 87. Schematic representation of the upflow anaerobic sludge blanket (UASB) process
a) Sludge – liquid inlet; b) Gas screens; c) Settled-sludge return opening

A distributor on the floor of the reactor is used for uniform liquid feeding, and it must necessarily meet special design requirements. This distributor consists of a system of polypropylene tubes laid in a defined way and fixed to a concrete plate. Each tube assumes a prescribed hydraulic load determined by the cross section of the tube and the number of outlets provided. Recirculation of partially purified wastewater to the UASB reactor is required only in the case of high COD concentrations in the wastewater in order to avoid pronounced concentration gradients as a function of height. Experience has shown that the height of the system should not exceed ca. 6 m, because uncontrollable hydraulic conditions would otherwise result owing to gas coalescence. This in turn generally also entails a considerable amount of surface area.

Wastewater fed into a UASB reactor should be acidified as completely as possible. The process chain is enhanced by a preceding preacidification step, which simultaneously ensures hydraulic equilibration. Granulate formation, of crucial importance for the proper functioning of a UASB plant, is influenced in part by the solids content of the reactor. A high proportion of inert solids may prevent pellet formation and thus the successful application of the concept.

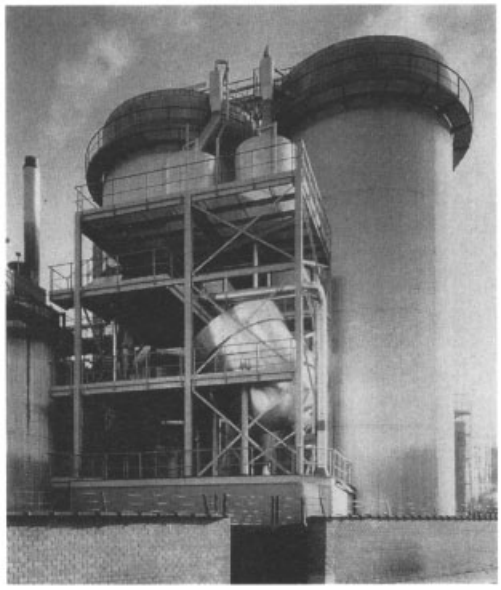

Figure 88. A two-stage UASB treatment facility used in the starch industry (COD load: 28 800 kg/d)

Most of the large-scale wastewater-treatment plants designed for anaerobic operation – 129 as of January 1989 [311] – were conceived as UASB plants. This speaks well for the idea, but there are also disadvantages to be considered. For example, the process of microbial pellet formation is still very empirical, and it is neither applicable to all types of wastewater nor uniformly reproducible. The formation of granulated sludge may occasionally require several months or even years, and sometimes it never takes place; moreover, existing pellets may redissolve and be washed out of the system due to a change in process conditions. For this reason, pellet sludge is extracted from existing plants for the purpose of "seeding" newly commissioned UASB reactors, and material may even be transported overseas to this end. An examination of the utilization efficiency of the existing reactor volume shows that only 25 % of the space is actually devoted to bioconversion, while the rest functions simply as a phase-separating device. The biosolids concentration approaches 75 – 85 g/L in the granular sludge region, but the specific sludge activity with respect to acetate conversion (methanogenic activity) amounts to only 0.2 kg COD/(kg CDM · d) as a result of hindered mass transport. This in turn limits the

COD-elimination capacity (at maximum space–time yield) to ca. 6–14 kg COD m^{-3} d^{-1} [312].

The minimum realizable residence time in a UASB reactor is 2–6 h. However, such short residence times are actually achieved only with lightly contaminated wastewaters (1–2 kg COD/m^3). A residence time of 8–20 h is required for a moderately concentrated effluent (COD 3–20 kg/m^3). In the years 1976–1982 UASB reactors were installed primarily in the Netherlands for the treatment of wastewater from the sugar industry. However, starting about 1985 such plants also began to appear in conjunction with the starch, beverage, and paper industries. Figure 88 is a photograph of a two-stage UASB plant constructed for the starch industry.

2.2.7.3. Fixed-Bed Reactors (Anaerobic Filters)

Reactors with this configuration, in which microorganisms are fixed onto stationary carrier materials, were developed primarily in the United States as a response to problems encountered with sludge retention in the contact process [247]. As a result of pressure drop in the bed there is a lower limit of practicality for the size of the carrier particles utilized. In addition, a fixed bed easily becomes clogged from biomass growth or the accumulation of solids from the wastewater. It is advisable to use coarse, lumpy carrier materials made of stone, ceramic, or plastic with a diameter of 6–10 cm. The specific surface available for immobilization is thus restricted to 150–300 m^2 per cubic meter of reactor volume.

Other disadvantages are a consequence of the plug-flow nature of the process, which is undesirable from a reaction-engineering point of view (see Section 2.2.6). The associated pH and substrate-concentration profiles over the bed depth result in corresponding inhomogeneities with respect to the type, density, and activity of the microorganism population. Near the bottom of the fixed bed, fermentative (acidogenic) microorganisms predominate at a pH of 3.8–4.5 before acid degradation and increased pH in the upper regions combine to produce acceptable environmental conditions for methanogenic microorganisms. A low flow rate may encourage the accumulation of a high biomass concentration on the biofilms due to minimal shear stress, but the specific activity of the biocatalyst nevertheless remains extremely low on account of serious restrictions on mass transport. Increasing the length of the reactor causes a steadily increasing fraction of the cross section to be blocked by emerging biogas. For this reason, the capacity with respect to COD degradation for industrial-scale reactors of this type is restricted to 4–10 kg COD m^{-3} d^{-1}.

Figure 89. Anaerobic fixed-bed reactor of the Bacardi Corporation at Puerto Rico, designed for the treatment of residues left behind after alcohol distillation [248]

The residence time in an anaerobic filter is ca. 2–8 d. The first commercial plant of this type was built in the United States as early as 1973 for the anaerobic treatment of wastewater from the wheat-starch industry [313]. Other plants followed in the early 1980s in Europe, Canada, and South America for the purification of industrial wastewater and animal residues. A plant installed in Puerto Rico by the Bacardi Corporation for treating residues from alcohol distillation is especially impressive because of the size of the reactor [248]. With a tank volume of 12 500 m^3 this anaerobic-filter system is the largest single-stage biomethanation plant in the world. Figure 89 is a photograph of the Bacardi fixed-bed reactor. It is designed for downward flow, and is filled with 9900 m^3 of PVC honeycomb structures. A wastewater inflow of ca. 1400 m^3/d results in an average residence time of 9 d and an inflow COD load of 95 kg/m^3 at a load rate per unit volume of 10 kg COD m^{-3} d^{-1}. The ca. 25 000 m^3 of biogas produced each day is used for steam generation. Start-up of the reactor required ten months.

Figure 90. Two-stage anaerobic fluidized-bed reactor for treating wastewater from baker's yeast manufacture by Gist Brocades, Delft, Holland [314]

2.2.7.4. Fluidized-Bed Reactors

With this bioreactor concept, carrier particles covered with biomass are maintained in a fluid state by an upward flow of the wastewater. The disadvantages of anaerobic filters outlined above do not apply here because a superficial liquid velocity of 30 – 50 m/h provides almost complete back mixing and virtually unimpeded liberation of biogas. Further advantages of the fluidized-bed technique include:

1) The possibility of treating wastewater containing solids without the need for physical pretreatment
2) Simultaneous continuous discharge of excess sludge, which makes periodic sludge removal unnecessary
3) Increased specific biomass activity because of a reduction in mass transport resistance in the relatively thin biofilms owing to an increase in the convection current
4) Low space requirements associated with a compact design (height-to-depth ratio $\gg 1$ with a height of 15 – 20 m)

In general, natural minerals like quartz sand, meerschaum (sepiolite), and other silicates with particle sizes of 0.15 – 0.30 mm can be used as carrier materials. Since these present smooth surfaces, and the microorganisms are subject to immobilization only by adhesion (electrostatic and van der Waals forces, hydrophobic interactions), excessive shear stress causes considerable abrasion and the loss of biomass. This factor represents the most serious disadvantage of the fluidized-bed reactor configuration.

Pioneering work in the field of anaerobic fluidized-bed technology was conducted by the Dutch company Gist Brocades in Delft. It was found that the microbial tendency towards adhesion could be enhanced by maintaining the residence time in the reactor below the generation time of the bacteria, in this way providing selection pressure in favor of sessile mutants [300]. As early as the late 1970s pilot plants (V_R = 3.6 m^3) of this type began anaerobic treatment of wastewater from the baker's yeast industry. The first analogous large-scale plants were constructed in the early 1980s. Mixed wastewater

from the baker's yeast industry consists largely of condensate from the evaporation of a heavily contaminated split stream mixed with contaminated spray water and cleaning water. A typical COD load is 4.5 kg/m^3. Figure 90 is a photograph of such a two-stage plant, comprising two identical fluidized-bed reactors. Each of the reactors has a volume of 390 m^3 and is 21 m high. Quartz sand (140 t/reactor) serves as the carrier. The fluidized-bed volume is 215 m^3. An important design factor in the reactor head is a phase-separating device for the removal of biogas adhering to the biomass. The onset of methanation required four months. Peak values as high as 52 kg m^{-3} d^{-1} were subsequently achieved during a residence time of 3 h for COD elimination. It must also be noted, however, that it was the fluidized-bed volume – interesting from a chemical engineering viewpoint – that serve as the basis for calculating the performance data, not the total reactor volume, which is so crucial with respect to plant construction. Moreover, total volume is the decisive factor in process-engineering design for anaerobic wastewater treatment, so it would have been preferable if this had been taken as the basis for reporting performance data. In the preceding example, the space – time yield would then be reduced by almost 50 %, and the residence time would correspondingly be doubled.

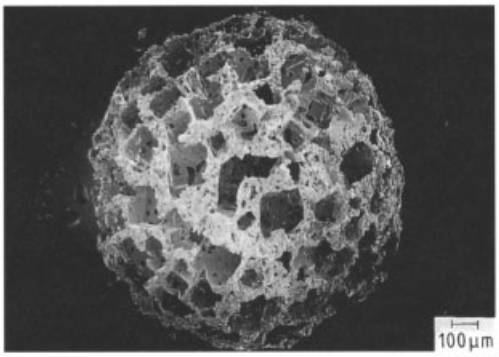

Figure 91. Scanning electron micrograph of a Siran sphere (particle size ca. 2 mm; porosity 55 %; pore diameter 60 – 300 µm)

Problems subsequently arose in reproducing pilot-plant performance under full-load operation. Inadequate colonization of the carrier was found to be the cause, which resulted in a subsequent expansion of the facility. According to available verbal information, the plant is now being operated as a UASB system.

Another example of the practical application of anaerobic fluidized-bed techniques was announced in 1982 [293]. This report refers to the treatment of brewery water in ANAFLUX reactors based on the carrier Biolite, an aluminum magnesium silicate.

The use of porous carrier materials as an alternative to sand represents an especially interesting development in the fluidized-bed concept, although it has so far only been put into practice on a pilot scale. In conventional fluidized-bed reactors, loss of biomass is common due to abrasion related to inevitable shear forces, because only the outer surface of the carrier particles is available for the adsorption of biomass. Especially in conjunction with this reactor concept a macroporous carrier material like Siran, which microorganisms can colonize in a three-dimensional way, is far superior to materials subject only to external surface coverage [315], [316]. Figure 91 is a scanning electron micrograph of such a Siran pellet (for further details about pellet preparation and pore structure, see Section 2.2.7.5).

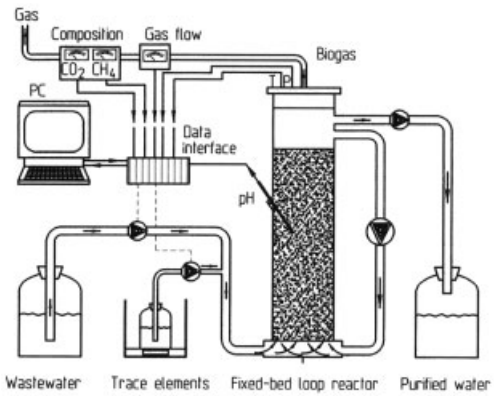

Figure 92. Schematic illustration of a fixed-bed loop reactor

Using as an example the anaerobic treatment of vapor condensates from the pulp industry,

an elimination rate of 250 kg COD $m^{-3}\,d^{-1}$ has been achieved with a loading rate per unit volume >300 kg COD $m^{-3}\,d^{-1}$. The corresponding biogas productivity was 160 $m^3\,m^{-3}\,d^{-1}$, from which the potential of this technique is abundantly clear. Nevertheless, practical applications in the wastewater sector are probably still rather remote because of the high cost of the carrier material.

2.2.7.5. Fixed-Bed Loop Reactors

The principle underlying this type of reactor combines the advantages of fixed-bed and fluidized-bed reactors, but avoids several of the disadvantages of both (Fig. 92). Most of the liquid responsible for the upward flow in a fixed-bed loop reactor [257–268] leaves the fixed bed and returns once again through an external loop, where it is mixed with fresh influent. Circulation is gradually increased until concentration gradients no longer play a role. Establishing the correct rate of circulation requires that biomass accumulation in the intergranular volume also be taken into account. This in turn means that the void volume must be neither rinsed entirely free of biomass nor allowed to become clogged. In this way it is possible to establish a very high biomass concentration. Thus, despite the low propagation rate of anaerobic microorganisms, provision must be made for periodically washing out excess sludge by briefly increasing the circulation rate. Simultaneously, small suspended solid particles not trapped in the course of physical pretreatment of the wastewater, and which may have accumulated in the fixed-bed loop over time, are also flushed out.

This particular anaerobic-reactor concept has become especially familiar in the literature in conjunction with the use of macroporous carrier materials for microbial immobilization with three-dimensional colonization. The technique itself leads to an increase in particle size and an increase in the density difference between the particles and the fluid, making it possible to decouple the residence time for the biocatalyst and the fluid in the reactor without the addition of chemicals (e.g., for carrier fixation, flocculation, or flotation) or supplementary energy (e.g., for filtration or centrifugation), and also without the need for constructing additional facilities (e.g., sedimenters or lamellar separators).

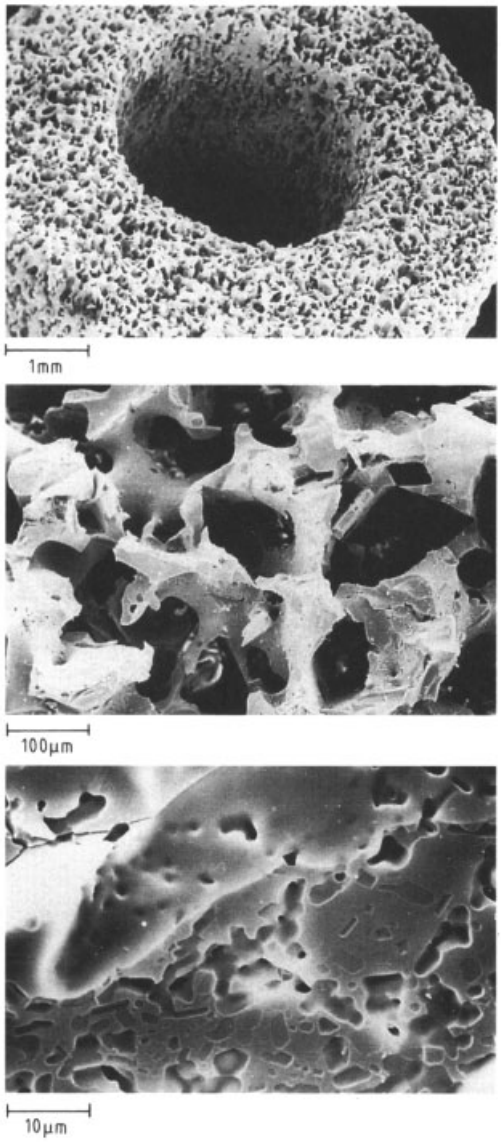

Figure 93. Scanning electron micrographs at various magnifications showing an open-pored sintered glass that has been formed into Raschig rings

On account of their low density, organic polymers frequently used for carrier purposes, such as polyurethane foams [267], [317], are not advantageous in this case, because such a carrier would float on the surface and require confinement with nets. The circulating stream would then need to be directed downward, impeding

the release of biogas. For this reason inorganic materials are much more effective, including lava slag, a naturally occurring macroporous carrier. An alternative is Siran, a commercial sintered-glass material with adjustable properties, manufactured by Schott Glaswerke, Mainz. A homogeneous mixture of the starting materials, glass, and salt powder of a defined particle size is first sintered, and the cooled material is subsequently extracted with water, resulting in structures with porosities as high as 70 %, together with a defined pore-size distribution. The production process lends itself to carriers of almost any shape (e.g., Raschig rings, pellets). Figure 93 is a scanning electron micrograph of a Raschig ring made of sintered glass depicted at various enlargements.

Basic process-development work in this area was initiated in the 1980s at the Forschungszentrum Jülich (Jülich Research Center, Germany), leading eventually to establishment of the "biogas high performance process" for treating heavily organic-laden wastewaters. This process was successfully commercialized over a period of several years by licensing it to various parties interested in building treatment plants.

The first large-scale plant of this type was for a potato-starch factory, and it took only ten months to build, with operation commencing in September, 1987. Figure 94 provides both a flow diagram and a photographic overview of the two-stage facility. It incorporates a preacidification tank with a working volume of 1000 m^3. Four fixed-bed loop reactors (d; 300 m^3 each) for methanization are connected in parallel. These are filled with lava slag that serves as the porous carrier. The plant is designed for a hydraulic load of 2000 m^3/d, resulting in a residence time of 14 h for the methanization reactor (second stage) at a loading rate per unit volume of 30 kg COD $m^{-3}\,d^{-1}$. Operation under full hydraulic load leads to the anaerobic elimination of more than 20 t of COD per day. At a COD/BOD ratio of 1.7, this corresponds to about 200 000 population equivalents. The process in turn generates ca. 12 000 m^3 of biogas per day. The exceptional efficiency of microbial colonization in conjunction with a macroporous carrier is apparent from the fact that this plant was able to operate at full capacity within six weeks. Seven large-scale plants of this type are currently active. Apart from the one described above, similar plants have been built for the pulp and paper industry, baker's yeast fermentation, and feed manufacture. In the latter application peak values $\leq$106 kg COD $m^{-3}\,d^{-1}$ have been achieved without difficulty in facilities with capacities of 20 and 170 m^3 through the use of Siran Raschig rings with projected space – time yield of 35 kg COD $m^{-3}\,d^{-1}$.

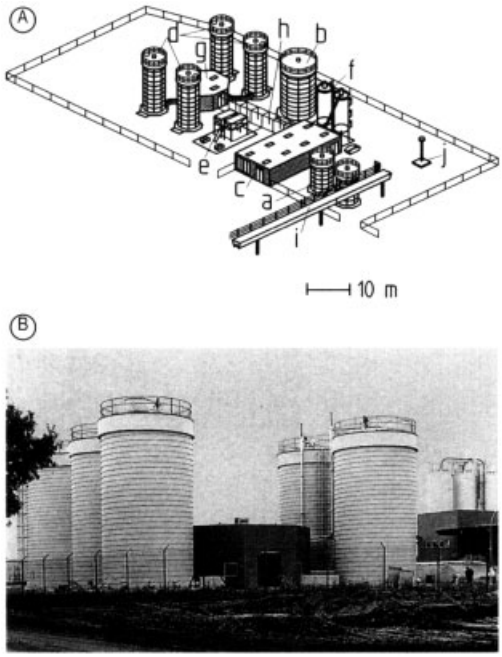

Figure 94. Flow diagram (A) and photograph (B) of a two-stage anaerobic wastewater-treatment facility of the fixed-bed loop-reactor type, used in the potato starch industry (COD load: 35 000 kg/d)
a) Entrance tanks; b) Acidification tank; c) Operations center; d) Methanization reactors; e) Excess-sludge separation; f) Gas purification; g) Pumping station; h) Pipe framework; i) Pipe bridge; j) Gas flare

2.2.7.6. Engineering Considerations

Scale Transfer and Dimensioning of Anaerobic Fixed-Bed Loop and Fluidized-Bed Reactors. Little information is available in the literature regarding scale transfer and dimensioning for anaerobic reactors with integrated provision for biomass retention (e.g., UASB reactors, fixed-bed and fluidized-bed reactors). Apart from systems based on the fluidized-bed technique, all the plants are relatively flat, with a *H/D* ratio $\leq$1. As a result of a variety of limi-

tations (e.g., low biomass concentration and activity in the contact sludge process, considerable loss of biomass with increasing gas velocity in UASB plants, and the formation of concentration and pH gradients in consequence of the plug-flow behavior of filter reactors) the space – time yields with respect to COD degradation do not exceed 15 kg COD m^{-3} d^{-1} from a design standpoint.

Industrial high-load plants [space – time yield ≤ 80 kg COD m^{-3} d^{-1}] operating under conditions comparable to those associated with the large-scale projects already realized (e.g., the biogas high-performance process) remain a rarity. According to [311], it is possible to achieve loading rates per unit volume >60 kg COD m^{-3} d^{-1} under optimized conditions in a UASB plant. The corresponding operating conditions would however require simultaneous development of new sludge-separation devices for installation at the reactor head to ensure the efficient retention of biomass pellets. Such a facility might be described as a "floating – settling granular-sludge UASB plant."

The trend toward further establishment of high-load treatment plants – more than 130 industrial facilities are currently operating throughout the world – is clear from efforts now underway to perfect this especially promising approach to anaerobic wastewater treatment. Nonetheless, the scale-up of anaerobic wastewater treatment has never been studied in a systematic way and for a variety of reasons (e.g., the absence of free exchange of experience, inadequate basic knowledge on account of a "black box" attitude). Of all the methods devised for model transfer in chemical process engineering, especially the similarity principle and modelling theory, the latter is most applicable to the conditions associated with anaerobic wastewater treatment, and it has also been accepted as the most useful starting point.

From a chemical reaction-engineering viewpoint, anaerobic fixed-bed loop reactors and fluidized-bed reactors behave almost like well-mixed, continuously operated stirred-tank reactors under discharge conditions (i.e., under circumstances where there is no appreciable pH or substrate-concentration gradient). The result is a source term for emerging product that remains constant over the entire height of the reactor. But since the product is a gas mixture (CH_4/CO_2) that is only slightly soluble in water, there is a height-dependent linear gas accumulation that increases as the height of the reactor increases. In the course of scaling up to the next larger version of a bioreactor (increased height or increased biogas production) it is necessary to anticipate extensive biogas flooding in the upper bed regions to an extent dependent on the *H/D* ratio. This means that the effective hydraulic residence time in the reactor will decrease in an indefinite way, leading to a corresponding reduction in conversion, and inevitably also to a decrease in the space – time yield in the larger unit relative to a laboratory-scale plant.

The following central question thus arises in the practical scaling up of the reactor types discussed here: In a large scale plant, what is the maximum practical bed height, or how small might the volume-based reactor cross-section be so that biogas release will not have a limiting effect on the overall bioprocess?

It is possible to approach a solution by introducing a process-dependent influencing parameter that takes adequate account of the issues raised above. The "surface gas-loading rate" (or "gas loading per unit area", GLA) with dimensions m^3 m^{-2} h^{-1} can be regarded as such a quantity. This characterizes the number of cubic meters of newly-formed biogas released from the anaerobic reactor per cubic meter of reactor cross-sectional area per hour. The surface gas-loading rate increases in a linear way with increasing column height, and reaches a maximum that limits the further release of biogas, thus inevitably establishing a maximum permissible bed height. This maximum height is a function of the nature of the process in the case of the fixed-bed or fluidized-bed technique, and of the particle size if a resting carrier is used. The requisite design-dependent volume can then be achieved simply by adopting the corresponding cross section.

Appropriate experiments with fixed-bed loop reactors (containing 7- and 25-mm Siran Raschig rings and 32 – 50-mm lumps of broken lava slag) as well as fluidized-bed reactors (Siran pellets with a diameter of 1 – 2 mm) have provided the desired maximum values for surface-gas loading rate from the observed gas hold-up. The reactor heights together with the carrier-bed heights were increased stepwise until sharp decreases in performance became apparent. The

following surface-gas loading rates were established in this way:

7 mm Raschig ring/fixed-bed loop reactor:	3.19 $m^3\,m^{-2}\,h^{-1}$
25 mm Raschig ring/fixed-bed loop reactor:	6.9 $m^3\,m^{-2}\,h^{-1}$
32 – 50 mm lava slag/fixed-bed loop reactor:	8.5 $m^3\,m^{-2}\,h^{-1}$
1 – 2 mm pellets/fluidized-bed reactor:	19 $m^3\,m^{-2}\,h^{-1}$

The relationship between the maximum permissible bed height and volume-specific biogas production (space – time yield) is illustrated in Figure 95. It is clearly apparent that when the volume-specific biogas production assumes a value equal to the maximum gas-loading rate per unit area (GLA), the permissible bed height shrinks to only 1 m. Disregarding the 7-mm carrier particles (which are not suitable for practicalapplication in a fixed bed because the realizable bed height is too low and the potential for clogging is too great), use of carrier particles >25 mm in diameter definitely opens the way to plant heights of 8 – 14 m operating in a practical range of volume-specific biogas production.

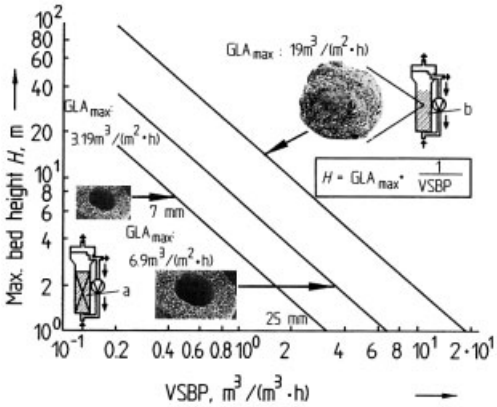

Figure 95. Scaling (A) a fixed-bed loop reactor or (B) a fluidized-bed reactor: the relationship between maximum permissible height and volume-specific biogas production (VSBP)

Comparison of Single-Stage and Two-Stage Process Operation. Taking the anaerobic treatment of distillery residue as an example of dealing with a complex, highly organic-laden effluent from the food industry, it proved possible to define realistic limits for single-stage processing [318], [319]. A fixed-bed loop reactor like that shown in Figure 93 was used in this particular study. Very large fluctuations were observed in the fatty-acid spectrum in the reactor effluent, as well as in the hydrogen content of the biogas, despite the unchanging nature of the wastewater and an approximately constant residence time (Fig. 96). The observed propionic acid peaks clearly correlate with the hydrogen concentration, which serves to verify the relationship suggested by Figures 70 and 71. It should also be noted that an acetic acid peak accompanies each propionic acid peak (suggesting the conversion of propionic acid to acetic acid!).

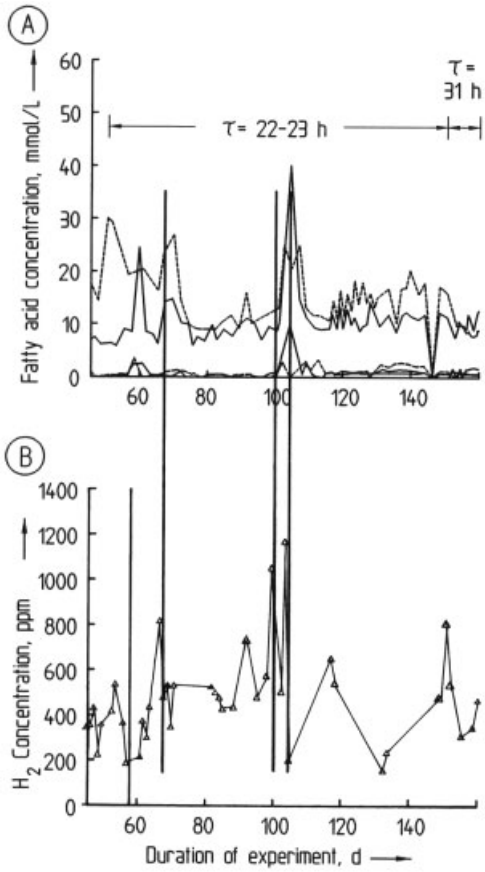

Figure 96. Fluctuations in (A) fatty acid and (B) hydrogen concentrations in a single-stage fixed-bed loop reactor with a feed consisting of distillery residues
—- Acetic acid; · · · · · · Propionic acid; -·-·- Isobutyric acid; - - - - - Butyric acid; ——— Valeric acid

In view of these unsatisfactory results, a reactor configuration corresponding to that shown in Figure 83 was installed, along with a fixed-bed loop reactor for methanation. This particular reactor concept facilitated independent adjustment of the pH in the course of preacidification. A close relationship was discovered between the product spectrum in the effluent and the preacidification step as a function of pH. The formation of propionic acid can pose problems in the subsequent methanation stage, and the propionic acid level proved lowest when a pH of 4 was established in the preacidification stage. At this pH the organisms of the methanation stage are no longer active, however, which explains the absolute necessity for a two-stage procedure. Results for both single-stage and two-stage processes in the anaerobic treatment of distillery residues are presented in Figure 97. For every spaceload tested the attainable COD conversion was found to be greater in the case of the two-stage process, and process stability was greatly improved as well.

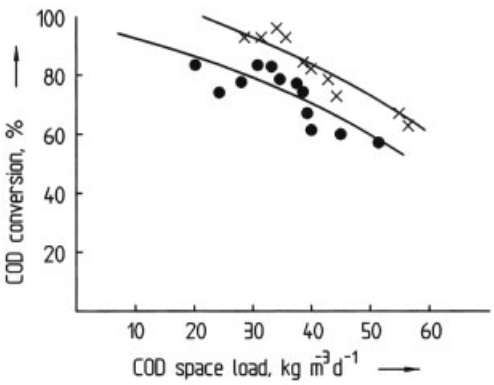

Figure 97. The effect of COD loading on COD conversion in single-stage and two-stage configurations for the degradation of distillery residues (COD loading = 34.5 kg m^{-3} d^{-1}; single stage configuration: 75 % conversion with $\tau = 25$ h, 89 % with $\tau = 35$ h; two stage configuration: 93 % with $\tau = 17$ h)
●–● Single-stage; ×—× Two-stage

Influence of Residence-Time Distribution on the Microbial Degradation Kinetics. Since fermentation displays typical autocatalytic behavior, with part of the product simultaneously serving as a catalyst, and because K_m values (Michaelis – Menten constants) for the most important substrates are very small, efforts have been made to approximate stirred-tank behavior by virtually complete backmixing. The need for this remedy applies only so long as the stationary substrate concentration in the reactor is significantly higher than the semisaturation concentration, however. Should a higher level of degradation be desired, or if the K_m values increase (by an apparent decrease in substrate affinity) as a consequence of integrated measures taken to decouple the mass-transport dependent relationship between substrate residence time and biomass, a transition would be observed from a zero-order reaction to a first-order reaction. Under these conditions, use of a reactor system with limited backmixing (e.g., a cascade) would be feasible.

Wastewaters from the production of sauerkraut and from ethanol distillation were used to compare the two design variants in both single-stage and cascade methanation. Fixed-bed loop reactors were employed in all cases. A preceding stirred tank was used for preacidification of the wastewater. The investigation covered residence times of 5 – 16 h. COD effluent values from the cascade arrangement were in all cases clearly lower than those from effluent data for single-stage methanation. In fact, the observed differences were more pronounced (by up to 4 kg of COD per cubic meter) in the case of the more contaminated distillery residues compared with the moderately contaminated wastewater from sauerkraut fermentation. The clear superiority of the process involving a two-stage methanation cascade relative to a fixed-bed loop reactor is especially apparent from curves of the residual COD content as a function of residence time (Fig. 98). For a COD reduction of 80 % ($COD_e/COD_o = 0.2$), a residence time of ca. 13 h is required in the case of single-stage methanation. The residual content in effluent from a two-stage cascade with the same residence time is only 0.1, corresponding to a conversion of 90 %. A residence time of as little as 8 h is sufficient to degrade 80 % of the COD in the case of the two-stage process. Under these conditions, the observed volume reduction in a methanation cascade compared with the single-stage mode of operation is ca. 40 %.

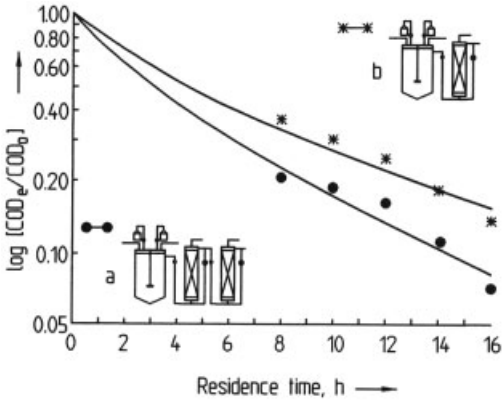

Figure 98. Change in residual COD as a function of residence time in the anaerobic treatment of distillery residues via a two-stage process (a) with and (b) without a cascade arrangement
COD_e = effluent COD; COD_o = initial COD.

Results from all the reaction-engineering investigations so far reported for process design optimization with fixed-bed loop reactors in the anaerobic treatment of highly organic-laden effluents from the food industry are summarized in Figure 99, which shows the course of the minimum residence time and the associated relative methanation volume for a COD reduction of ca. 85 % at various developmental stages in the corresponding "process evolution." Accordingly, the volume of a two-stage cascade is only 3.5 % of that required for a continuously operated stirred-tank reactor (CSTR) lacking provisions for residence-time decoupling.

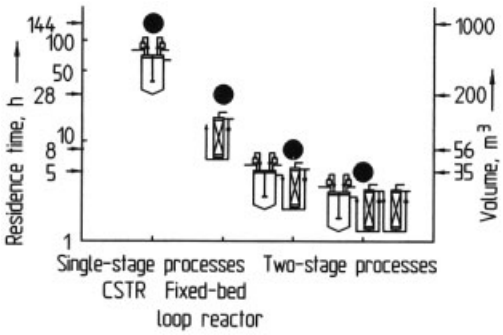

Figure 99. Decrease in the minimum residence time and relative methanization volume for an 85 % COD reduction in the treatment of complex residues from the food industry as anaerobic fixed-bed loop technology has evolved

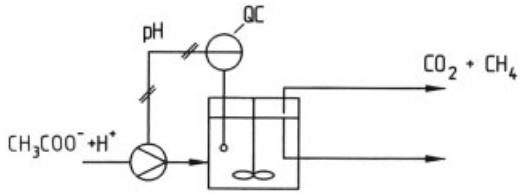

Figure 100. Schematic representation of pH-auxostatic regulation

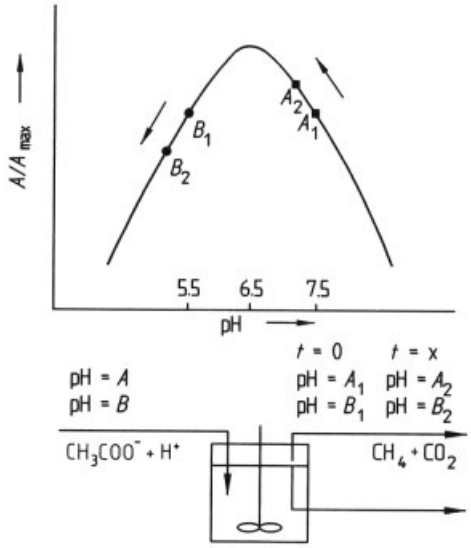

Figure 101. Schematic representation of chemostatic regulation

2.2.8. Control

The control concepts that are potentially applicable to anaerobic degradation can be illustrated with the case of acetic acid. To prevent acidification of the reactor due to insufficient substrate degradation, the substrate pump can be controlled via a pH electrode in such a way that substrate feed is interrupted as soon as the pH falls below a certain value. With this pH-auxostatic mode of operation (Fig. 100), wastewater feed is governed by a pH controller. The cleavage of acetic acid to CH_4 and CO_2 corresponds formally to a titration with an alkaline solution (see Fig. 74), so as the pH in the reactor rises this effect must be compensated by the addition of acidic substrate. Use of a proportional controller means that there will always be a difference between the actual pH and the desired value, and

the amount of substrate added is proportional to the magnitude of this difference. Since substrate introduction is in effect controlled by the degradation efficiency of the microorganisms, the microorganisms themselves determine how much wastewater will be fed to the reactor. 101

Alternatively, an alkaline pH might also be assured in the reactor by establishing a suitable pH for the influent. The target pH in the reactor must be sufficiently alkaline so that, in the case of a change in residence time or concentration, a slight decrease in the pH can be associated with an increase in the activity of the microorganisms. This chemostatic concept results in automatic stabilization so long as the operation is conducted on the alkaline side of the pH optimum with respect to the microorganisms. It is advantageous to safeguard the process by ensuring additional pH control on the acidic side of the pH optimum to prevent excessive drift into the acidic range.

The first approach has the advantage that the bioreactor always operates at maximum load, but it requires a buffer volume at the inlet to permit the controller to interrupt the substrate feed. On the other hand, the second approach excludes the possibility of continuous operation corresponding to maximum space – time yield. Nevertheless, it does avoid the need for a buffer volume, and there is no risk of uncontrolled preacidification.

Because of the low propagation rate of anaerobic microorganisms, considerable importance is attached to an optimal time-course for start-up of the biogas reactor, as well as to preventing damage to a cultivated biomass. Both objectives are most easily achieved with the help of modern computer-coupled control technology (Fig. 92). Thus, the set-points for the feedback control systems associated with the relevant parameters (residence time, temperature) are established in advance with the help of a personal computer based on the criterion of a maximum increase in the space – time yield. The amount of biogas generated, the pH, and the derivatives of these quantities as a function of time serve as indication values with this form of control. A definite level of degradation for the wastewater contaminants can be established as a secondary condition. Actual degradation can be determined online on the basis of the residence time and the extent of biogas production. A decision matrix is constructed from a combination of directly measurable and derived quantities, on the basis of which the computer adjusts the externally predetermined set-points consistent with observed trends.

The process configuration illustrated in Figure 102 made it possible to demonstrate in the anaerobic treatment of exhaust-steam condensates from the pulp industry, for example, that a greater than 3-fold increase in performance could be achieved within less than four weeks starting with a space – time yield of ca. 40 kg COD $m^{-3}\,d^{-1}$.

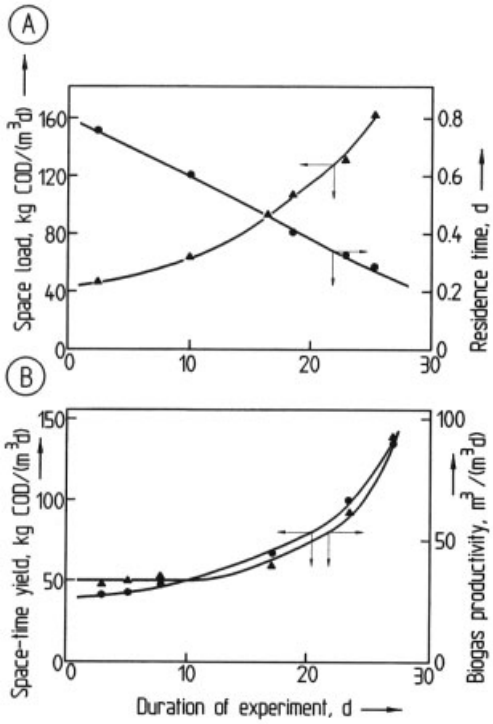

Figure 102. Computer-controlled startup of a fixed-bed loop reactor for the anaerobic degradation of exhaust-steam condensates in the pulp industry
A) Space load and residence time; B) Space – time yield and biogas productivity

2.2.9. Future Prospects

The anaerobic treatment of wastewater represents another example of the fact that innovations may result from the joint application of

biological and industrial principles. Despite sophisticated optimization of the oxygen transfer process, aerobic wastewater treatment becomes less competitive with the anaerobic technique the higher the contamination level of the wastewater. The principle of carrier fixation as a way of preventing the washing-out phenomenon is also derived from a model frequently observed with sessile microorganisms in nature. On the other hand, minimizing limitations associated with diffusion problems by increasing the convection current is a typical industrial approach to problem-solving. Since energy required for this purpose can, in principle, be provided directly by gas-driven motors, the approach in question is justifiable even over the long term. In the case of highly contaminated wastewater, anaerobic treatment is not only energy self-sufficient, but also capable of providing excess energy for use in other applications. In addition, the resulting marked reduction in excess sludge production should constitute a weighty argument in favor of the anaerobic process.

Concerns related to heavy metal deposition in a wastewater reactor operating under anaerobic conditions have not yet been accorded sufficient attention despite the fact that the distribution of clarification sludge (e.g., from the activated-sludge process) on agricultural land has been curtailed because of excessive heavy-metal contamination. Wastewater treatment is an industrial process associated with one of the largest levels of mass throughput known, and for this reason it has provided a major impetus to further developments in bioprocess technology in general.

3. Aerobic Processing Steps in Process Combinations

This chapter represents a collection of the various processing steps in which the liquid – gas interface plays an important part in the purification of water or air.

Flotation is a technique for separating undissolved constituents from water, whereby the particles attach themselves to gas bubbles. Particles rising with the gas bubbles are removed at the surface as foam. As a rule, flotation processes compete with sedimentation processes for accomplishing a particular separation. Flotation can offer significant advantages over sedimentation, especially for poorly sedimenting and poorly thickening water constituents.

One application of flotation is in drinking-water treatment [320], especially for the clarification of surface waters contaminated with such poorly sedimenting particles as algae. Flotation is often used in this context as a preliminary purification step prior to filtration [321]. Later in the course of drinking-water treatment, flotation can be used for treating waterworks sludges or sludge-containing waters that arise from rapid filtration, sand washing, microstraining, coagulation and precipitating plants, or decarbonation [322].

In the purification of commercial or industrial wastewaters, including the recovery of valuable materials, flotation is useful especially for the separation of fibrous, flocculent, greasy, or oily materials [323]. Flotation is normally used in this context only to carry out decentralized partial cleansing upstream from a central biological wastewater-treatment unit [324], [325].

Flotation also forms a part of the overall process of biological wastewater treatment for separating and concentrating sludge [326].

Biofilters and *bioscrubbers* are devices used for the biological cleaning of waste air. Such biological waste-air cleaning processes compete most notably with adsorption, absorption, and thermal and catalytic processes. Biological waste-air treatment entails transferring some of the constituents of the air in question to an aqueous solution that is then oxidized by microorganisms, analogous to biological wastewater treatment. Another portion is consumed in the formation of cell mass [327–330].

In biological waste-air treatment a distinction is made between processes involving fixed microorganisms (biofilters and *trickle-bed bioreactors*) and those based on suspended microorganisms (bioscrubbers). In the case of fixed microorganisms, the waste-air flows through some type of microorganism-support layer, which in the case of a biofilter may consist of twigs or compost, whereas with a trickle-bed bioreactor system it would be a porous packing. In a bioscrubber the air contaminants are absorbed into water or an aqueous suspension of microorganisms in an absorption column. The absorbed contaminants are then degraded inside the regenerator, a bioreactor like that used in biolog-

ical wastewater treatment. The bacteria used in biological waste-air purification may originate from a wastewater treatment plant or from the soil. Their composition must first be adapted – as a rule over the course of several weeks – to the waste air subject to treatment. However, a regenerator unit might also be integrated through a recycle system with a biological wastewater-treatment facility, such as one designed for treating industrial effluents. If the air contaminants to be degraded happen also to be present in the wastewater, then microorganisms from the wastewater-treatment facility will already be sufficiently adapted to perform this supplementary function [331].

Biological waste-air purification was originally called upon most often for odor-reduction assignments, in conjunction for example, with intensive management of livestock or other agricultural problem areas. Wastewater-treatment plants also represent one of the traditional fields of application [332]. Biofilters find use in the influent treatment zones of wastewater-treatment plants, in the biological stages of both trickling filters and activated-sludge tanks, in sludge-treatment facilities, and in conjunction with such other equipment as flotation tanks or buffer tanks.

Biological waste-air purification has been increasingly applied to problems of other organic emissions in various industrial and commercial fields.

3.1. Flotation (see also Section 5.2.2.3; → Flotation)

3.1.1. Basic Principle

The principle underlying flotation is that gas bubbles rising through water are capable of transporting to the surface particles that may adhere to them, where these particles can then be separated with the foam. Flotation has been used in mineral technology for about a hundred years as a technique for sorting, i.e., for the separation of particles of different materials. A prerequisite for such sorting – for separating a valuable ore fraction from its matrix, for example – is that only one of the two components has a hydrophobic surface, thus permitting it to adhere to the gas bubbles, while the surface of the other component is hydrophilic and well wetted by the water. Since the hydrophilic particles adhere only weakly or not at all to the bubbles, they will not be transported to the surface, but instead remain in suspension or sink to the bottom.

In mineral technology, specially selected combinations of flotation aids are introduced for each specific separation. For example, materials known as *collectors* assume the task of conferring a hydrophobic character on the hydrophilic surface of certain components. These collectors are surface-active substances whose polar regions adhere to the solid surface, leaving the nonpolar parts oriented toward the surrounding water. *Foaming agents* have the function of preventing the rising particle-laden bubbles from bursting at the surface, instead causing the particles to bind in a foam that is sufficiently stable for separation. Foaming agents are also surface-active substances, but ones whose polar group is surrounded by water, permitting a nonpolar group to project into the air bubble. Finally, *modifying reagents* are added to influence the flotation process by way of the pH, or to intensify hydrophobicity differences between the components to be separated by passivation or activation of the collecting effect.

Various methods have been used to generate the corresponding gas bubbles. In convective bubble generation (*aeration* or *induced air flotation*) the gas stream (usually air that has been sucked or pumped into the system) is caused to disperse. Devices for this purpose include rotor–stator systems and nozzles, occasionally also porous aeration surfaces. In *pressurized dissolved-air flotation* the water is first saturated with gas at an elevated pressure and then depressurized to a lower pressure prevailing in the flotation apparatus. According to Henry's law, the solubility of a gas in a liquid is proportional to the pressure, so bubbles are produced in the course of this depressurization. Another important bubble-generation variant is *electroinduced air flotation*, in which gas bubbles are generated electrolytically. Some types of flotation apparatus take advantage of a combination of various principles. An important advantage of dissolved-air flotation is that the gas bubbles form preferentially at the surfaces of hydrophobic particles. For these particles, a transport step conveying them to the gas bubble is avoided, and

this is often the rate-determining phenomenon in aeration flotation.

Aeration flotation is most widely applied in mineral technology, in which it has long been regarded as a standard method for various sorting functions. In the past, mechanical flotation devices based on rotor – stator systems played a greater role than hydraulic or pneumatic flotation systems, into which energy must be introduced with a liquid pump or a gas compressor. *Countercurrent flotation columns* have become increasingly important since the early 1980s. In this case air is fed into the column from below, while wash water is charged from above. The suspension to be separated is introduced into the column about two-thirds of the way up from the bottom. This countercurrent principle confers greater selectivity on the separation process.

Bubbles generated by the dispersion technique in mineral technology aeration flotation are about 1 mm in diameter. Materials that are to be sorted by flotation are as a rule ground to particle sizes ranging from a few tens to a few hundreds of micrometers. Particles that are too coarse are no longer capable of being transported by the gas bubbles. In the fine-particle range, selectivity decreases; in particular, the fraction $<5\,\mu m$ enters a foam unselectively, since very fine particles will adhere to gas bubbles even if they have a low degree of hydrophobicity. Smaller bubbles in the approximate size range $50-100\,\mu m$ are formed in dissolved-air flotation, and in electroflotation a typical bubble size is only ca. $20\,\mu m$.

Flotation has become an accepted method for water-purification purposes since about the 1960s. The problem here differs from that in mineral technology in two important respects: first, flotation is used not for sorting but for clarification. Thus all the particles present must be transported to the surface by gas bubbles for separation with the foam. Second, the particle size involved is below the range of typical applications in mineral technology.

In water treatment, flotation represent an alternative to sedimentation. For mineral particles larger than about $40\,\mu m$, which have a settling rate >5 m/h and can thus be separated rapidly in settling tanks, no alternative method is necessary or suitable. For the fines fraction below about $5\,\mu m$, on the other hand, with settling rates <0.1 m/h, separation is no longer practical in a settling tank without flocculation. Flocs, even when composed of mineral particles, often have such a high fraction of pore volume that their settling rate is unsatisfactory irrespective of dimensions. Moreover, many of the particles to be separated in water treatment differ in density only slightly from the surrounding liquid. Examples include such organic contaminants as oil droplets or microorganisms, and especially activated sludge, which must be concentrated in the course of biological wastewater treatment.

Flocculants used in the flotation process for water treatment have another important function aside from producing flocs with a size and structure favorable to flotation: they serve to render the particles hydrophobic, and therefore to guarantee adhesion of the flocs onto the gas bubbles. The individual steps involved in particle flocculation and adhesion of the particles to gas bubbles often proceed simultaneously. The result is the formation of voluminous flocs in which gas bubbles have also been incorporated.

The same basic flotation principles developed for mineral technology are also used for water treatment. Mechanical flotation devices have never found application here, however – only hydraulic or pneumatic flotation systems, since bubble sizes $\leq 100\,\mu m$ have proven advantageous. Electroflotation has been applied successfully to certain types of problems, but by far the most important flotation method for water-treatment purposes is dissolved-air flotation.

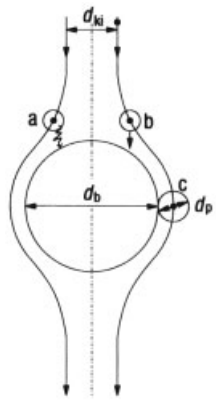

Figure 103. Mechanisms for establishing contact between particles and gas bubbles
a) Diffusion; b) Sedimentation; c) Interception

3.1.2. Contact and Adhesion of Particles to Gas Bubbles;Flotation Kinetics

In both aeration flotation and electroflotation, contaminant particles must be transported to the gas bubble before the two surfaces can make contact with each other. The flow pattern around a rising gas bubble is sketched in Figure 103. Three important mechanisms can lead to contact between particles and bubbles suspended in a liquid.

First, a particle may make contact with a gas bubble through interception without its center ever leaving the streamline, provided the particle's radius is at least as large as the shortest distance from the streamline to the bubble. Since transport of the particles out of their streamline is not required, this contact mechanism cannot be classified as a transport mechanism.

Alternatively, particles might be transported by either sedimentation or diffusion (Brownian motion) away from a streamline and toward a gas bubble. (For particles in the size range of interest for water treatment by flotation – well under 100 µm – transport by inertial effects plays essentially no part relative to interception or sedimentation.)

For interception Figure 103 shows an example of the limiting streamlines for contact of a particle of diameter d_p with the bubble. These streamlines constitute a rotationally symmetrical tube around the vertical axis. All particles whose centers move toward the bubble within this limiting-streamline tube will inevitably make contact with it. On the other hand, particles with centers of gravity outside this limiting-streamline tube flow past the bubble without touching it. The cross-sectional area of the limiting-streamline tube at a great distance from the bubble is described as the *collisional cross-sectional area* A_ki for the contact mechanism of interception (i). The ratio of this collisional cross-sectional area to the projected area P_b of the bubble is described as the *single-bubble collision efficiency* η_i:

$$\eta_\mathrm{i} = \frac{A_\mathrm{ki}}{P_\mathrm{b}} = \left(\frac{d_\mathrm{ki}}{d_\mathrm{b}}\right)^2 \tag{3.1}$$

where d_ki is the collisional cross-sectional diameter for interception and d_b the bubble diameter. Another collisional cross-sectional area, A_ks, can be defined correspondingly for the sedimentation contact mechanism, in which particles are actively transported toward the bubble from the streamline. A collisional cross-sectional area A_kd must also be formally defined for the diffusion contact mechanism, although the transport mechanism in this case is not determinate, but stochastic. Single-bubble collisional efficiencies for sedimentation (s) and diffusion (d), respectively, can be expressed as:

$$\eta_\mathrm{s} = \frac{A_\mathrm{ks}}{P_\mathrm{b}} = \left(\frac{d_\mathrm{ks}}{d_\mathrm{b}}\right)^2 \tag{3.2}$$

$$\eta_\mathrm{d} = \frac{A_\mathrm{kd}}{P_\mathrm{b}} = \left(\frac{d_\mathrm{kd}}{d_\mathrm{b}}\right)^2 \tag{3.3}$$

The overall single-bubble collisional efficiency η_k is interpreted as the sum of contributions from interception, sedimentation, and diffusion:

$$\eta_\mathrm{k} = \eta_\mathrm{i} + \eta_\mathrm{s} + \eta_\mathrm{d} \tag{3.4}$$

Relationships describing these individual contributions are known from the theory of deep-bed filtration:

$$\eta_\mathrm{i} \sim \left(\frac{d_\mathrm{p}}{d_\mathrm{b}}\right)^2 \tag{3.5}$$

$$\eta_\mathrm{s} \sim \frac{g\,(\varrho_\mathrm{p} - \varrho_\mathrm{l})\,d_\mathrm{p}^2}{\mu \cdot u} \tag{3.6}$$

$$\eta_\mathrm{d} \sim \left(\frac{kT}{d_\mathrm{p}\,d_\mathrm{b}\,u}\right)^{2/3} \tag{3.7}$$

where g is the acceleration due to gravity, ϱ_p is the particle density, ϱ_l is the liquid density, μ is the dynamic viscosity, k is the Boltzmann constant, and T is the absolute temperature. Here u represents the incident velocity of flow on a spherical collector of diameter d_b. If u is defined as the bubble-rising velocity for bubble sizes $d_\mathrm{b} \leq 100\,\mu\mathrm{m}$,

$$u \sim \frac{g \cdot \varrho_\mathrm{l} \cdot d_\mathrm{b}^2}{\mu} \tag{3.8}$$

then the following modified relationships are obtained for η_s and η_d [333]:

$$\eta_\mathrm{s} \sim \frac{\varrho_\mathrm{p} - \varrho_1}{\varrho_\mathrm{l}} \left(\frac{d_\mathrm{p}}{d_\mathrm{b}}\right)^2 \tag{3.9}$$

$$\eta_\mathrm{d} \sim \left(\frac{kT}{g\,\varrho_\mathrm{l}\,d_\mathrm{p}}\right)^{2/3} \cdot \frac{1}{d_\mathrm{b}^2} \tag{3.10}$$

For relatively large particles whose densities differ very little from that of the liquid, interception is decisive, but with large density differences sedimentation predominates. For relatively small particles, diffusion represents the dominant contact mechanism.

Figure 104 shows the single-bubble capture efficiency in the case of a negligible density difference as a function of particle size, with the bubble size as a parameter [333], [334]. Taking into account both diffusion, dominant in the fines range, and interception, which dominates in the coarse range, it develops that there is a minimum capture efficiency for particle sizes around 1 µm. In order to avoid unsatisfactory flotation with particles in this size range, fine primary particles should be flocculated insofar as possible. In practice, therefore, interception and sedimentation are most often the contact mechanisms of interest.

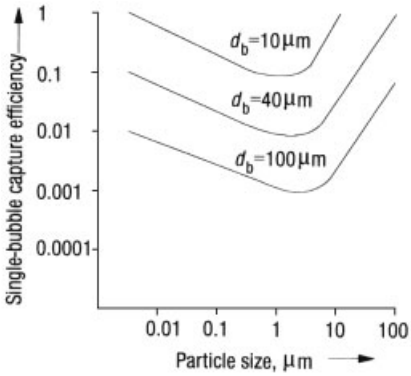

Figure 104. Single-bubble capture efficiency as a function of particle size [333]

For all three contact mechanisms, the single-bubble collision efficiency decreases in inverse proportion to the square of the bubble diameter. For all particle sizes, therefore

$$\eta_k \sim \frac{1}{d_b^2} \tag{3.11}$$

Apart from the temperature and various material parameters, the proportionality constant contains only the particle size d_p. In other words, the collisional cross-section A_k is independent of the bubble size.

A second prerequisite for successful separation is adhesion of the particles to the gas bubble subsequent to collision. Decisive with respect to such adhesion is on the one hand van der Waals forces acting between the particles and the liquid, and on the other electrostatic interactions between the particles and the gas bubbles. These quantities are difficult to access by direct measurement. A reasonable starting point for evaluating the effectiveness of adhesion is the contact angle Θ between the liquid – gas interface and the solid surface at the three-phase contact line. Figure 105 illustrates the contact angle Θ for a drop on a planar solid surface as well as for a spherical solid particle at a planar gas – liquid interface. The contact angle Θ is established such that the total interfacial free energy of the solid – liquid (sl), solid – gas (sg), and liquid – gas (lg) interfaces assumes a minimum value. Young's equation

$$\gamma_{lg} \cos\Theta = \gamma_{sg} - \gamma_{sl} \tag{3.12}$$

describes the relationship between the corresponding interfacial free energies per unit area γ, also known as the *interfacial tensions*, and the contact angle Θ.

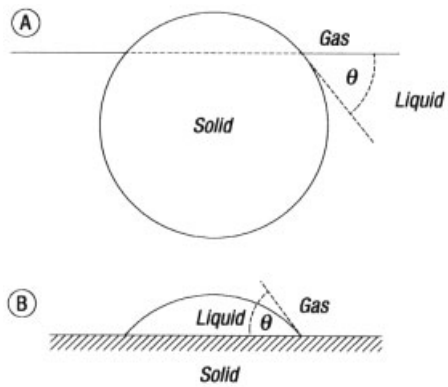

Figure 105. Contact angles Θ for (A) a spherical solid particle at a gas – liquid interface, and (B) a drop resting on a solid surface

As Figure 105 A illustrates, a particle penetrates more deeply into a gas bubble the larger the contact angle. Hydrophilic particle surfaces form a small contact angle with water, hydrophobic particles a large one. The strength of the bond between particle and gas bubble is greater the more hydrophobic the particle and the greater the contact angle.

The *adhesion probability* in the event of a collision between a particle and a bubble is designated here as η_h. This approaches zero as the contact angle Θ approaches zero. With increasing contact angle Θ the adhesion probability η_h approaches the value one. A quantitative description is difficult, since it requires comparing adhesive forces with disruptive hydrodynamic forces. The shear stress such a particle – gas bubble agglomerate experiences in the flow field increases more rapidly with the particle size than does the strength of the binding. For this reason small particles, which are the target of separation in water treatment, will adhere sufficiently well to gas bubbles even if they are only slightly hydrophobic and develop only a small contact angle with the liquid.

The single-bubble capture efficiency η is the product of the adhesion probability η_h and the single-bubble collision efficiency η_k:

$$\eta = \eta_h \cdot \eta_k \tag{3.13}$$

If the single-bubble capture efficiency η is known, the kinetics of particle capture by gas bubbles can be calculated. Over the time interval dt, gas bubbles with a rising velocity u pass through a layer whose thickness can be represented as

$$\mathrm{d}h = u \cdot \mathrm{d}t \tag{3.14}$$

within the flotation tank. The relative concentration reduction for freely suspended particles in the course of this ascent is

$$\frac{\mathrm{d}C}{C} = -\eta \cdot P_v \cdot C_{vg} \cdot u \cdot \mathrm{d}t \tag{3.15}$$

where

$$P_v = \frac{3}{2} \cdot \frac{1}{d_b} \tag{3.16}$$

P_v is the projected area per unit volume of the gas bubbles, and C_{vg} is the volume concentration of gas bubbles in the flotation tank. This is obtained as a ratio of the superficial gas velocity u_g and the rising velocity u of the gas bubbles:

$$C_{vg} = \frac{u_g}{u} = \frac{q_g/A}{u} \tag{3.17}$$

The superficial gas velocity can be described as ratio of the volumetric flow rate of the gas q_g and the cross-sectional area A. From integration of the equation

$$\frac{\mathrm{d}C}{C} = -\frac{3}{2}\frac{u_g \cdot \eta}{d_b} \cdot \mathrm{d}t \tag{3.18}$$

it follows that

$$\ln(C/C_0) = -\frac{3}{2}\frac{u_g \cdot \eta}{d_b} \cdot t \tag{3.19}$$

If a constant adhesion probability η_h is assumed, it follows from Equations 3.11 and 3.13 that

$$\eta \sim \frac{1}{d_b^2} \tag{3.20}$$

and

$$\ln(C/C_0) \sim -\frac{u_g}{d_b^3} \cdot t \sim -\frac{q_g}{d_b^3} \cdot t \tag{3.21}$$

The flotation-velocity constant

$$k = \frac{\ln(C/C_0)}{t} \tag{3.22}$$

is thus proportional to the gas throughput, and inversely proportional to the third power of the bubble diameter (i.e., inversely proportional to the bubble volume).

From this point of view, the bubble size should therefore be kept as small as possible. On the other hand, the minimum size of the bubble is limited by the requirement of adequate buoyancy and sufficiently high rising velocity even in the presence of attached particles. The bubble volume must as a rule be at least one order of magnitude greater than the volume of the largest particles to be separated. In practice, bubble sizes in the range of ca. 10 – 100 µm have proven advantageous for water-treatment flotation [335–338].

If, during aeration flotation, the bubbles are broken up by turbulent-flow processes, the ultimate bubble size is determined by the power per unit volume $\varepsilon = P/V$:

$$d_b \sim \varepsilon^{-2/5} \tag{3.23}$$

It follows that

$$\ln(C/C_0) \sim q_g \cdot \varepsilon^{6/5} \cdot \mathrm{t} \tag{3.24}$$

The flotation-velocity constant k thus increases in proportion to the gas throughput, and slightly more than proportionally relative to the power per unit volume.

3.1.3. Implementation, Fields of Application, and Operational Experience

The mechanical flotation devices widely used in mineral technology, in which power for dispersing the gas bubbles is introduced by rotor–stator systems, play virtually no role in water treatment.

Figure 106 shows an example of a hydraulic flotation device developed for wastewater treatment in the chemical industry [339] as well as for the separation and thickening of sludge in biological wastewater treatment [326]. Small bubbles about 150 µm in diameter are formed by a flared nozzle. Flotation tests with systems of varying material composition, in which the two process parameters gas throughput and liquid throughput were decoupled from each other by throttling the gas throughput, agreed in yielding the common relationship

$$k \sim q_g \cdot q_l^2 \tag{3.25}$$

Since the power per unit volume ε introduced by the jet of liquid is proportional to the square of the volumetric flow rate of the liquid q_l^2 it follows that

$$k \sim q_g \cdot \varepsilon \tag{3.26}$$

This experimental result approximately confirms the relationship found in the model calculation (Equation 3.24).

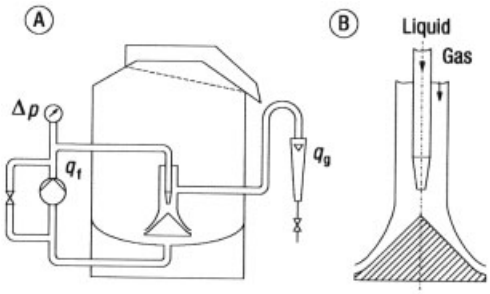

Figure 106. Aeration flotation [326]
A) A flotation cell with automatic flotate discharge; B) Detailed view of the flared nozzle (self-priming, with radial bubble generation)

Numerous examples of the application of a flotation cell with a flared nozzle like that illustrated (Fig. 106) can be found for wastewater treatment in the chemical industry, including the separation of an azo pigment or a crop-protection agent from wash water, separation of toluene from a toluene – water emulsion, and the removal of printer's ink from paper waste [339]. In the separation and thickening of sludge from biological wastewater treatment, biomass with an initial dry-matter concentration of 2 – 10 g/L was removed almost entirely from treated wastewater on a pilot-plant scale, resulting in a flotate with 20 – 30 g of dry matter per liter [326].

Aeration flotation is regarded as unsuitable for the treatment of drinking water [336]. In the wastewater field as well the advantages associated with dissolved-air flotation (see below) predominate more often than not [325].

A special form of dissolved-air flotation is dissolved-air flotation for the separation of sludge from wastewater treated in an aeration chamber exceeding 10 m in height (e.g., a Bayer tower-biology unit). Figure 107 shows a flotation cell developed for this particular purpose [326]. The aeration chamber is arranged concentrically within a cylindrical tank. Effluent from the tower biology flows through the flared nozzle in the aeration chamber, and the bubbles required for flotation form spontaneously as the external pressure is released from gas originally dissolved under the static pressure of the tower. Dissolved-air flotation therefore entails no additional energy costs. The flotate so obtained contains dry matter ranging from 20 g/L (brewery wastewater) to 80 g/L (wastewater from chemical plants) at permissible area-loading rates up to 8 m/h [$(m^3/h)/m^2$] [326], [340].

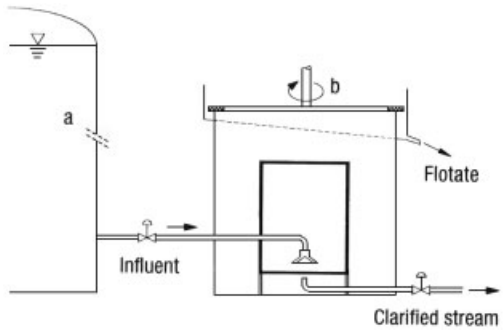

Figure 107. Flotation cell developed for dissolved-air flotation in conjunction with a tower-biology system [326]
a) Tower biology; b) Flotation cell

Under certain conditions, gas arising in the course of microorganism metabolism also can be utilized directly for flotation. For example,

gas generated during the anaerobic degradation of nitrate has been used for intermittent flotation of activated sludge with a flotation time between 4 and 48 h [341]. As a rule, however, dissolvedair flotation is applied in conjunction with a split stream of previously treated water that has been saturated with a gas in the pressure range 0.4 – 0.8 MPa, and subsequently led back via pressure-releasing elements into the flotation tank, as shown in Figure 108 [342].

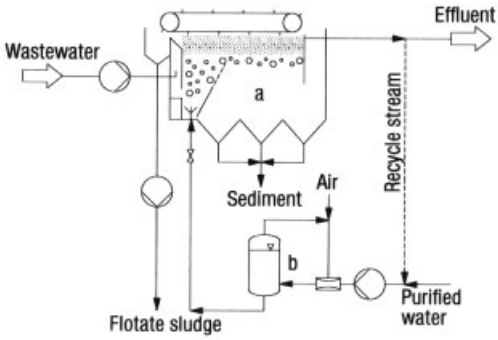

Figure 108. Schematic representation of a typical dissolved-air flotation system
a) Flotation tank; b) Water pressurization

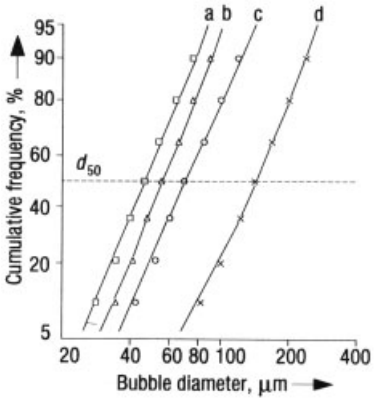

Figure 109. Air-bubble diameter distribution in dissolved-air flotation as a function of saturation pressure
a) 0.8 MPa; b) 0.6 MPa; c) 0.4 MPa; d) 0.2 MPa

Figure 109 illustrates the bubble-diameter distribution for air in water as a function of saturation pressure [342]. In the preclarification of chemical wastewater from the manufacture of organic intermediates, a relatively uniform solids-removal efficiency of 90 – 98 % was achieved by this technique at an area-loading rate ≤4 m/h by the addition of 20 % pressurized water at a saturation pressure of 0.4 MPa. Figure 110 shows the superiority of the process over sedimentation in a thickener [343]. This superior contaminant-removal efficiency in the preclarification stage of wastewater treatment leads to a noteworthy reduction in the burden placed on the subsequent biological treatment.

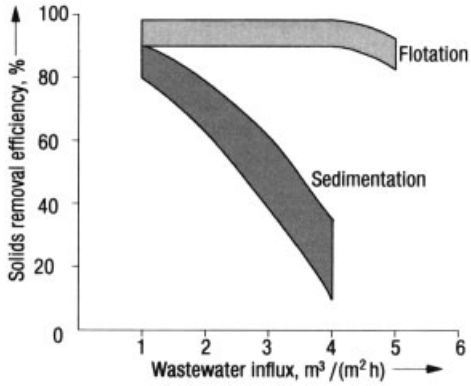

Figure 110. Comparison of the separation efficiency of dissolved-air flotation versus sedimentation [343]
Insolubles content = 90 – 450 mg/L; pressurized water addition for flotation = 20 %

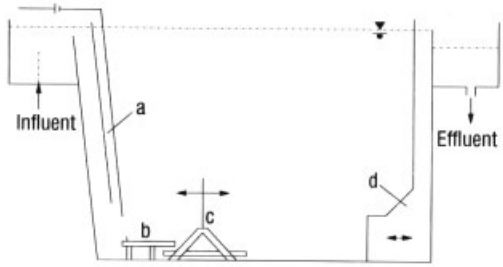

Figure 111. Schematic representation of an electroflotation system [348]
a) Influent structure; b) Electrodes; c) Partition; d) Effluent structure

Dissolved-air flotation has proven particularly worthwhile in the context of grease- or oil-containing wastewaters from a wide variety of industrial and commercial sources, including the foods sector [344–347]. Dissolved-air flotation has also had wide success in drinking-water treatment where prior flocculation of the raw water is necessary [336].

Electroinduced air flotation is reserved for smaller mass fluxes associated with special types of problems, such as wastewaters derived from electroplating (see Fig. 111) [348] or oil – water

emulsions [349]. Electroflotation is uneconomical for large-scale applications [348] because of the high cost of the energy required for generating the gas bubbles.

3.1.4. Costs and Benefits

As an example of the operating costs (without capital debt service) of a flotation plant in wastewater treatment, the various process steps of neutralization, precipitation of organics with polymeric precipitants, flocculation, dissolved-air flotation, dewatering of the flotate with a filter press, and incineration of the dewatered sludge have been examined for a medium-sized decentralized wastewater stream of 50 m^3/h [325]. In an unfavorable case the overall operating costs may be as high as € 5 per cubic meter of wastewater. The largest shares are associated with the costs of chemicals and incineration: ca. € 1.5 per cubic meter each. Annual operating costs for 50 m^3/h of wastewater may amount to € 2.3×10^6.

The cost of flotation, which is the most universal approach to solid – liquid separation in decentralized wastewater treatment, is lower than that of evaporation or incineration by a factor of about 10. But the benefits of evaporation and incineration are also greater, since they result in complete disposal of highly contaminated effluent, whereas the precipitation and flotation processes discussed in the above example represent only a partial treatment of a moderately contaminated split stream. Since the value of this partial treatment is also a subject of some dispute it must remain an open question whether it is wise to spend € 2.3×10^6 annually for the pretreatment of a stream that comprises only a small fraction of the total wastewater from a large chemical plant [325].

Compared with sedimentation, flotation offers considerable advantages in many cases, not only in water treatment and recovery of valuable materials but also in the separation and thickening of sludge. As a rule, it leads to a higher dry-matter content for the resulting concentrate. Because of higher area-loading rates and shorter residence times, smaller apparatus volumes are required as well. This in turn generally implies lower investment costs, albeit at the expense of higher operating costs. A process comparison for drinking-water treatment of water for the River Ruhr yielded for sedimentation a space requirement 50 times greater than that for flotation. On the other hand energy costs for flocculation – flotation were found to be about 50 times greater than for flocculation – sedimentation [321].

Better opportunities exist with flotation than with sedimentation for control and adaptation to fluctuating operating conditions, although taking advantage of these opportunities requires the availability of more highly trained personnel.

3.2. Biofilters and Bioscrubbers

3.2.1. Basic Principle

Biological waste-gas purification, just like biological wastewater treatment, takes advantage of the metabolic capacity of microorganisms as a way of degrading and oxidizing organic compounds: to carbon dioxide and water, for example. Microorganisms are fundamentally able only to utilize materials dissolved in water. In aerobic biological wastewater treatment one of the most important technical challenges is dissolving sufficient amounts of oxygen in the water to permit propagation and metabolism on the part of the bacteria. With biological waste-gas purification there is the further problem that the compounds to be degraded must somehow be transferred into aqueous solution.

In a *biofilter* or a *trickle-bed bioreactor*, absorption and biological degradation of waste-gas components occur in a single step. With a *bioscrubber* the waste-gas components are first dissolved in water in an absorption column and then largely degraded in a regenerator coupled downstream. The operating principles of the biofilter, trickle-bed bioreactor, and bioscrubber are illustrated schematically in Figure 112 [350]. The biofilter and trickle-bed bioreactor both operate with fixed microorganisms. The organisms are fixed in the case of the biofilter on layers of such materials as twigs, heather, peat, compost, or other natural organic matrices, whereas the trickle-bed bioreactor utilizes an inert porous packing. Column fittings constructed from plastic, for example, are used in bioscrubbers. An aerated stirred tank or a hydraulic or pneumatic bioreactor can serve as the regenerator.

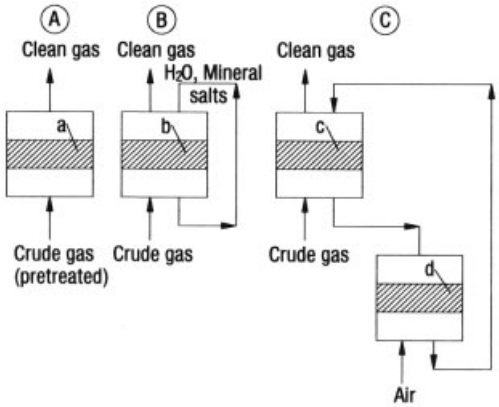

Figure 112. Basic operating principles associated with various biological waste-gas purification alternatives [350]
A) Biofilter; B) Trickle-bed bioreactor; C) Bioscrubber
a) Filter layer; b) Biocatalyst; c) Column; d) Regenerator

Biodegradation generally occurs with the aid of a mixture of various species of bacteria, actinomycetes, and fungi. In a biofilter the biomaterial serves not only as a support for the biofilm, but also as a source of moisture, and in some cases also of nutrient. Fungi and actinomycetes often predominate. In a bioscrubber, bacteria usually predominate. All nutrient is provided via the wash liquid containing the dissolved off-gas compounds that are to be degraded [351].

3.2.2. Prerequisites for Biological Waste-Gas Purification

Prerequisites for the growth and metabolic activity of microorganisms are a sufficient supply of nutrients and water as well as a suitable range of temperature and pH. Biological waste-gas purification with a bioscrubber requires supplemental addition of nutrients depending on the waste-gas composition, since biological regeneration of the scrubbing liquid would otherwise be impossible. In a biofilter, on the other hand, sufficient amounts of nutrients are available in the filter material. A precondition for adequate purification capacity with a biofilter is a sufficient and uniform supply of moisture. In practice, the moisture content is in the range 40 – 65 % depending on the filter material. The temperature range for most bioscrubbers and biofilters is 20 –40°C. The optimal and tolerable pH ranges vary considerably depending on the organisms involved. Most heterotrophic bacteria prefer neutrality, but mold fungi and yeasts are partial to slightly acidic conditions [351], [352].

3.2.3. Implementation, Fields of Application, and Operational Experience

Biofilters. The most widely used biofilters are single-level filters. Figure 113 illustrates the basic principle [329], [350]. If space is at a premium, several filter beds can be arranged above each other in tiers. The waste gas subject to treatment enters the filter bed from below through an air distribution system (e.g., perforated or slotted elements made of special concrete, plastic, or wood; gratings or ventilating bricks). The filter-bed height is typically 0.5 – 1.5 m. Filter area-loading rates fall in the approximate range of 20 – 300 $m^3/(m^{-2} \cdot h^{-1})$. Such single-level filters are constructed with surface areas up to ca. 2000 m^2 for the treatment of waste-gas streams as large as 100 000 m^3/h. Pressure loss in the biofilter itself is in the range 200 – 2000 Pa per meter of filter height [353]. Particularly in the case of relatively highly contaminated waste-gas, the biofilter is sized according to the specific filter loading, calculated from the extent of waste-gas constituents that will flow through the filter per unit of filter volume and time. The corresponding data (for organic carbon, for example) are reported in $g\,m^{-3}\,h^{-1}$. If the filter is to be used for odor removal, the loading can also be reported in odor units per cubic meter per hour.

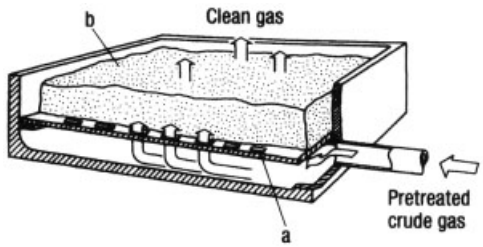

Figure 113. Overview of a single-level filter utilized for biological waste-gas purification
a) Concrete base; b) Filter layer

The first biofilters were used as early as the 1920s for reducing odor emissions from waste-treatment plants. The odor problem in waste-treatment business has in the meantime assumed

an important role in licensing laws and ordinances. In the past it was necessary with many waste-treatment plants to add special facilities for the collection and treatment of waste air. Today, construction and operating licenses are granted for new plants near residential areas only if proposals for the reduction of possible odor emissions are submitted in the early planning stages [332]. Apart from scrubbers, whose mode of operation is exclusively chemical, biofilter units are the most important devices for waste-air treatment in the wastewater-treatment sector. Biofilters are utilized in conjunction with almost every waste-gas source: from the influent zone through the biological stage, sludge treatment, and even such areas as flotation facilities. Recommended standards include a maximum specific filter-loading rate of 2×10^6 odor units per cubic meter per hour, a reaction time of 50 – 90 s, and a filter area-loading rate of ca. $100\,m^3/(m^{-2}\cdot h^{-1})$ [332].

The use of porous support materials makes it possible to achieve higher cell densities, and therefore in some cases higher reaction rates, with a trickle-bed bioreactor than with a biofilter. A schematic overview of a trickle-bed bioreactor is presented in Figure 114 [350]. Reaction products harmful to the microorganisms (such as hydrochloric acid, formed during the conversion of chlorinated hydrocarbons) can be removed from the reactor.

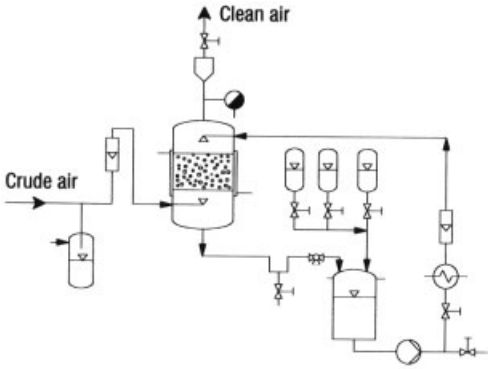

Figure 114. Schematic representation of a trickle-bed bioreactor [350]

The greater resistance of fixed microorganisms to the influence of poisons, acids, or temperature fluctuations is regarded as an advantage common to both biofilters and trickle-bed bioreactors [350].

Bioscrubbers. A bioscrubber consists of two separate pieces of equipment. Gas constituents that are to be degraded are first absorbed in a scrubbing column, after which these substances are biologically degraded in aqueous solution in a regenerator. The size of the scrubbing column and the quantity of scrubbing liquid must be such that the requisite level of clean-gas concentration is reached even at maximum concentration and quantity of crude gas and maximum stationary residual concentration of organic substances in the solvent. The spray density must be set sufficiently high to avoid buildup of a biological film and consequent clogging of the packing. Based on experience, a spray density $>20\,m^3/(m^{-2}\cdot h^{-1})$ is required. Moreover, the regenerator must be so designed that organic contaminants are almost totally degraded. At the same time, the requisite oxygen content of at least 1.55 mg/L for biological degradation must be maintained in the activated-sludge tank by suitable aeration units [354].

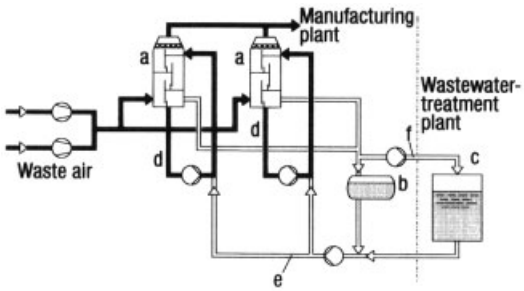

Figure 115. A bioscrubber facility integrated into a wastewater-treatment plant for the purpose of odor reduction [331]
a) Scrubbing columns (H = 9 m; D = 3,2 m; three sieve trays each); b) Regeneration tank ($V = 30\,m^3$); c) Biological stage of a wastewater-treatment plant; d) Biosuspension recycle system ($200\,m^3/h$); e) Exchange recycle system for the regeneration tank ($60\,m^3/h$); f) Exchange recycle system involving the wastewater-treatment plant (ca. $5\,m^3/h$)

Compared with biofilters, bioscrubbers are distinguished by lower space requirements, higher space – time yields, and greater flexibility and adaptability. On the other hand, maintenance and operating costs are usually higher with bioscrubbers, as is energy consumption. Waste-gas constituents that are only slowly biodegradable demand a long biosuspension residence time, and thus a correspondingly large regenerator. Providing such conditions makes it advanta-

geous to integrate into the system the biological stage of an existing wastewater-treatment plant, which can serve as a superior reservoir for fresh biosuspension. Even with high contaminant concentrations and slowly degradable constituents it is then possible to get along with a relatively small regenerator. Figure 115 is a schematic illustration of such an integrated plant designed for odor reduction of a waste-gas stream of 60 000 m^3/h [331].

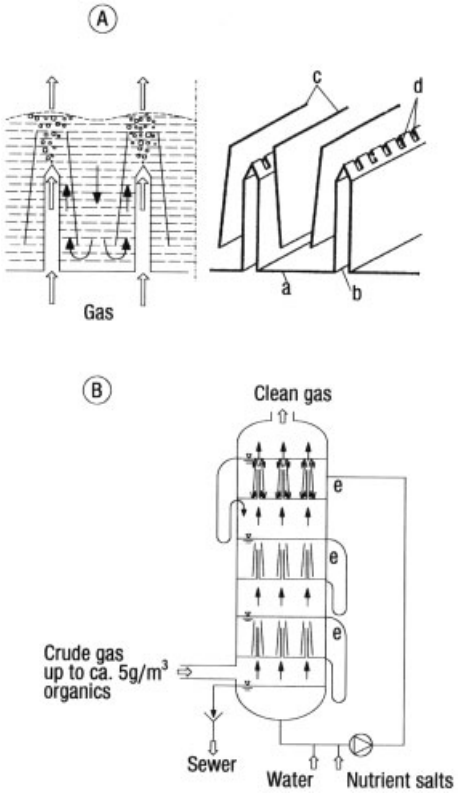

Figure 116. Bioscrubber in which the bioreactors are incorporated into the column plates
A) General operating principle [355]; B) Three-stage implementation [356]
a) Tray; b) Gas channel; c) Baffles; d) Slits; e) Bioscrubber trays with short absorption zones and long degradation zones

A new bioscrubber concept (Fig. 116) distinguishes itself by the fact that absorption and biological degradation occur not in separate devices, but in a single plate column [355], [356]. The required bioreactor volume is accommodated on the plates of the column. This concept, which has already proven its worth on a large scale, is characterized by compact construction, high elimination capacity, and a good space–time yield.

3.2.4. Costs and Benefits

Specific investment costs for a biofilter plant (without taking waste-gas collection into account) are estimated at ca. € 230/m^2 for a large facility (ca. 1000 m^2) and as much as ca. € 500/m^2 for a small facility (ca. 100 m^2) [332]. The operating costs associated with a biofilter plant are determined chiefly by the energy costs for fans and waste-air humidification (ca. 1.8–2.5 kWh/1000 m^3 of exhaust air). Costs related to personnel, maintenance, biomass replenishment and exchange, and water must also be taken into account. Figure 117 illustrates the effect on specific operating costs of the waste-air volumetric flow rate. In the waste-treatment sector, where it usually becomes necessary to treat quantities of waste-air exceeding 3000 m^3/h, operating costs are in the range of ca. 0.40–0.65 DM/1000 m^3 of waste air [357].

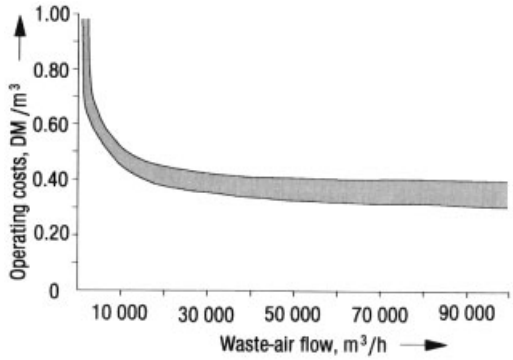

Figure 117. Operating costs associated with a biofilter plant, including costs for energy, personnel, water, compost, and maintenance
(1 DM = 1.95583 €)

Investment costs of ca. € 5–10 per m^3/h of waste-air capacity have been cited in conjunction with bioscrubbers. The corresponding operating costs are in the range € 0.08–0.18/1000 m^3 of waste air. In this context energy costs account for 80–90 % of the overall expenditures [358].

The fraction of organic emissions for which biological waste-gas purification systems are appropriate, taking all factors into account, has been estimated for Germany to represent about

10 % of total emissions [331]. The potential for such biological processes applies especially to waste air containing very low concentrations of odorous substances or solvents. Compared with catalytic and thermal processes, the biological processes offer the advantage that compounds present in the ppm range can be removed in an economically justifiable way provided they are water-soluble and sufficiently readily biodegradable.

4. Chemical Wastewater Treatment

Quite apart from much older methods for the oxidative and reductive elimination of wastewater components in the course of microbial metabolism, processes have been developed especially in the last twenty years that utilize chemical reactions directly for the degradation of water contaminants. Some have been developed to the point of technical maturity, and without exception these are oxidative processes (→ Oxidation, Chap. 4.4.2., → Oxidation, Chap. 5.2.1.).

The active oxidizing agents are oxygen- containing radicals, in the vast majority of cases the radical $H-O^{\bullet}$. The oxidizing agents actually introduced are usually air, oxygen, hydrogen peroxide, or ozone. Process-engineering characteristics distinguishing the individual methods vary quite considerably. What differentiates one process from another, and distinguishes all of these approaches from the microbial route, is a consequence largely of the nature of the wastewater itself and the extent to which its components should or must be degraded. It should also be noted that advantage is sometimes taken of several either direct chemical or biochemical methods, as well as combinations of the two together.

The generally more drastic reaction conditions or more aggressive reactants associated with a purely chemical method, and the often higher degradation rates achieved, lead easily to the impression that direct chemical methods are inevitably superior to biochemical methods. This is not necessarily the case, however. Material degradation is first and foremost a kinetic problem, and reaction mechanisms and catalysis play decisive roles [359–361]. To describe a substance as "not degradable" is basically an unwarranted oversimplification, because it implies that "degradability" is a fixed property of a substance, whereas it is in fact a matter of behavior; i.e., some degradation will indeed occur, but at a rate that is unacceptably low. In drawing any methodological comparison, various direct chemical methods must be taken into account, and especially from the following points of view [359–364]:

1) Energy costs – and not only with regard to requisite temperature levels, but also the required duration of action and any energy expenditures associated with separation and enrichment steps, especially in the case of dilute wastewater.
2) Irregular waste flows and how to cope with peak volumes or concentrations. Microbial processes function best when volumes and concentrations are relatively constant. Large fluctuations, especially toward higher values, can often be managed with the aid of subsequent or parallel chemical process steps.
3) Mixed wastewater and the separation of partial ("split") streams containing slowly degradable or interfering wastewater components [359], [360], [364]. Major grounds for the separation and individualized treatment of specific production wastewater streams include the following:
   a) A certain partial stream is characterized by high volume and the presence of contaminants amenable to degradation by means of the simple and inexpensive low-pressure wet-oxidation process (Section 4.4.1), resulting in a secondary wastewater suitable for introduction into a central biochemical wastewater-treatment plant.
   b) A partial stream contains wastewater components that degrade only slowly, suggesting that high-pressure wet oxidation (Section 4.4.2) might be feasible. This process is quite elaborate and involves relatively high operating costs, so a more economical alternative might be to pursue chemical degradation only up to the point of suitable fragmentation, assigning the responsibility for further degradation, as above, to a central biochemical wastewater-treatment plant.

c) A partial stream contains such a high level of salts that a cell culture would no longer be capable of functioning. Wet oxidation on the other hand might not be feasible either if in the case of certain salt-like companion substances the threat of corrosion rules out both low- and high-pressure oxidation. Such a case could be dealt with (after appropriate concentration) only by strictly thermal oxidation based on combustion (see Chap. 6).

Generally speaking, chemical treatment is restricted to special wastewaters characterized by components that are degraded too slowly in conventional waste-treatment plants or that interfere with the biochemical degradation of other substances. Chemical waste degradation processes then open to consideration include the following, listed in order of increasing operating temperature (and pressure):

1) Atmospheric-pressure wet oxidation by means of hydrogen peroxide, ozone, or air, with iron oxide or titanium dioxide as catalyst.
2) Low-pressure wet oxidation with air, based on an iron/quinone catalyst.
3) High-pressure wet oxidation with air, using copper as the catalyst.
4) Thermal oxidation; i.e., evaporation of water and combustion of the residue.

There is no such thing as a single "best" process, one clearly characterized by an ecological maximum and economic minimum. What is instead required is extensive experimental investigation to establish the limits and possibilities associated with each individual case, thereby providing a sound basis for meaningful comparisons. The question of how one should proceed with a particular wastewater source can then be answered reliably.

Thermal processes must generally be ruled out in cases involving low concentrations of degradable substances because of excessively high specific energy costs per unit volume. The use of hydrogen peroxide is often advantageous here, especially if the substances in question accumulate on an irregular basis. With higher contaminant concentrations, either high-pressure or low-pressure wet oxidation can be implemented. Evaporation followed by combustion becomes the method of choice if there is the added complication of high concentrations of inorganic salts. Wastewater of this type is derived largely from the chemical and materials-processing industries.

4.1. Oxidative Degradation by Hydrogen Peroxide

The direct oxidation of organic compounds by hydrogen peroxide under acidic conditions is a well-known but relatively seldom-employed process. The oxidation potential of H_2O_2 can be increased above that of ozone through catalysis, usually with Fe^{2+}. Oxidation with hydrogen peroxide was investigated as early as 1968 in conjunction with the treatment of municipal sewage [365] and wastewater from industrial production [366–368], especially effluents containing sulfur and phenols [359–372]. Successful hydrogen-peroxide treatment of wastewater from hardening plants [373] and tanneries [374] has also been described.

The special advantage of this process is that the technological effort required is small, an especially positive factor in the case of dilute wastewater (i.e., < 10 g COD/L). Nonbiochemical treatment is of special importance for wastewater fractions with low concentrations of contaminants, because the correspondingly large volumes determine the size of the required treatment facility, and thermal processes are excluded because of the enormity of the associated energy requirement.

Oxidation with H_2O_2 proceeds at atmospheric pressure and room temperature within 60–90 min. Hydrogen peroxide is not toxic, is easy to handle, and decomposes into the environmental products oxygen and water. Large-scale use of this otherwise virtually universal oxidizing agent is restricted by the high price of H_2O_2 itself, together with the fact that although self-decomposition proceeds with the formation of oxygen, the resulting oxygen contributes almost nothing to the oxidation process.

Hydrogen peroxide is not merely an oxygen-transfer agent that facilitates work in a homogeneous aqueous phase; indeed, one should make every effort to avoid conditions favoring the self-decomposition of hydrogen peroxide, i.e.:

$$H_2O_2 \longrightarrow H_2O + 1/2\,O_2$$

More to the point is the fact that catalysis by iron (molar ratio $H_2O_2 : Fe^{2+} = 15:1$, pH = 3.0), leads – as in the reaction of the Fenton reagent – to the formation of the $H-O^{\bullet}$ radicals, and it is these that constitute the actual oxidizing agent.

The redox potential E_0 for the reaction sequence

$$H_2O_2 \longrightarrow HO^+ + HO^-$$

$$Fe^{2+} + HO^+ \longrightarrow Fe^{3+} + HO^{\bullet}$$

$$HO^{\bullet} + H^+ + e^- \longrightarrow H_2O$$

has been determined to be 2.28 V [359], [375]. For the analogous reaction with ozone the redox potential is 0.21 V lower:

$$O_3 + 2\,H^+ + 2\,e^- \longrightarrow H_2O + O_2 \; (E_0 = 2.07\,V)$$

In the case of substances that oxidize only slowly even under these conditions, or via oxidation chains starting from them, it has been observed that self-decomposition of H_2O_2 – which is of no value for COD degradation – becomes appreciable. It is therefore necessary to establish the extent of utilization of introduced H_2O_2, and to compare this with the applied dosage in that particular case. The dosage should then be decreased until a utilization of nearly 100 % is achieved [375]. If a dosage is calculated based on the stoichiometry corresponding to complete oxidation all the way to CO_2, an almost equivalent COD degradation can usually be achieved with 60 % of the calculated dose, resulting in almost 100 % utilization of the added H_2O_2 [375].

The heat of reaction has also been determined under these conditions, leading to the following values established with a flow calorimeter (isothermal) [366]:

$$Q_1 = 4.59\,\text{cal/mg COD} \triangleq 19.22\,\text{J/mg COD} \; (\pm 3\,\%)$$

$$Q_2 = 2.48\,\text{cal/mg } H_2O_2 \triangleq 10.39\,\text{J/mg } H_2O_2 \; (\pm 4\,\%)$$

The extent to which different classes of substances differ in their behavior is illustrated by the data in Table 10. From an economic standpoint, the selected conditions in this study were appropriate only for the last two substances listed. The process can be regarded as nonspecific with respect to the various classes of organic substances typically encountered in wastewater. It also shows little dependence on the salt load.

The degradation of more than 100 individual compounds as well as authentic industrial wastewaters has been described [366], [375]. As a result of adsorption on flocculated material during the neutralization step the process leads to an effective reduction of COD roughly 10 – 15 % greater than the level of true degradation.

The process can be operated continuously in the form of a cascade (Fig. 118). On a pilot-plant scale ($\leq$ 500 L) the residence time is 60 – 90 min with a dosage of 20 – 80 % H_2O_2/COD. Partial oxidation with maximum utilization of H_2O_2 is pursued only to the point at which the wastewater can be subjected to subsequent degradation in a biological wastewater-treatment plant without impairing biochemical processes there. The type and level of salts present is of some importance, even though the actual oxidation with H_2O_2 is largely insensitive to salts [376]. Considering only the COD value of the wastewater, partial oxidation with H_2O_2 is still technically feasible with dilute effluents in the concentration range 0.5 – 10 g COD/L, where thermal processes must be excluded because of the associated energy demand.

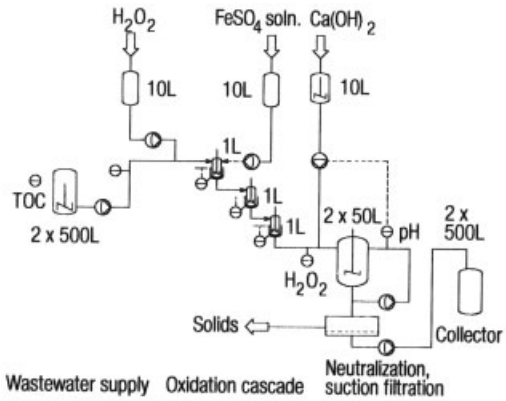

Figure 118. System for the continuous treatment of wastewater with H_2O_2

Radical formation based on hydrogen peroxide has also been achieved by UV irradiation for wastewater-treatment purposes [362].

Table 10. COD elimination of various substances as a function of H_2O_2 utilization

Substance	Dosage H_2O_2/COD, % of stoichiometric amount	COD elimination, % of initial value	H_2O_2 utilization, %
Morpholine	100	20	20
2-Aminoethanol	100	37	37
Diethyleneglycol	100	45	45
Polyethyleneglycol	100	35	35
Hexamethylenetetramine	100	32	32
2,4-Difluoro-5-chloro-6-methylpyrimidine	100	30	30
Phenyltrifluoroethylcarbamide	80	75	94
Ammonium trifluoroethyldithiocarbamate	80	79	99

4.2. Oxidative Degradation by Ozone

Use of ozone for the oxidative elimination of wastewater components has also been known for a long time [377]. Ozone was used even earlier for the sterilization of air and the treatment of drinking water. Its effectiveness is highly dependent on the pH (Fig. 119), and is based essentially on two mechanisms:

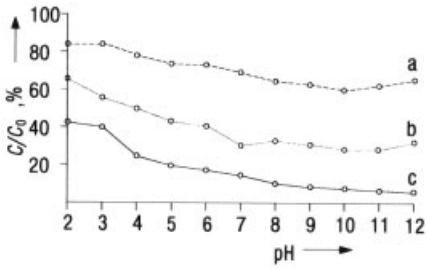

Figure 119. Ozonization of *p*-chlorophenol at an initial concentration C_0 of 250 mg/L (TOC = 140 mg/L, COD = 420 mg/L, AOX = 69.0 mg/L)
a) TOC; b) COD; c) AOX

1) So-called direct oxidation occurs under acidic conditions. This is a fairly slow process, but the conversion can be accelerated greatly if the energy necessary for radical formation is provided in the form of UV light from a Hg low-pressure radiation source [378–381].

$$O_3 \xrightarrow{h\nu} O^{\bullet} + O_2$$
$$O^{\bullet} + H_2O \longrightarrow H_2O_2$$
$$H_2O_2 \xrightarrow{h\nu} 2\,HO^{\bullet}$$

2) Alkaline oxidation also takes place via the intermediate formation of hydroxyl radicals [379]:

$$O_3 + HO^- \longrightarrow HO_2^- + O_2$$

$$HO_2^- + O_3 \longrightarrow {}^{\bullet}O_2^- + O_2 + HO^{\bullet}$$

$$H_2O + O_3 + {}^{\bullet}O_2^- \longrightarrow HO^- + 2\,O_2 + HO^{\bullet}$$

Ozonization of organic halogen compounds leads to both inorganic halide (from dehalogenation at C–Cl bonds) [377], [382] and the ensuing formation of new organic halides [381].

The formation of organic halides has also been detected in wastewater containing a combination of chlorine-free organic materials and inorganic halides originating from the accompanying salt load [381], [383], [384]. Compounds formed in this way generally reflect the presence of oxidizing agents, especially ozone or hydrogen peroxide.

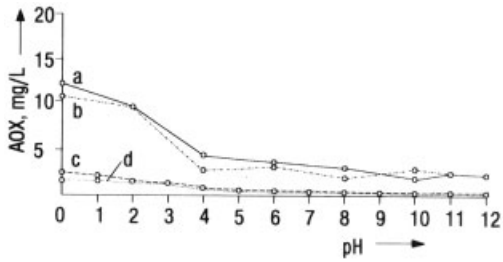

Figure 120. Ozonization of phenol (176 mg/L) in the presence of chloride ion
a) 2.0 g/L Cl^-; b) 2.0 g/L Cl^- + UV; c) 50.0 g/L Cl^-; d) 50.0 g/L Cl^- + UV

In this system, oxidative degradation of C–H and C–C bonds and the formation of C–Hal bonds in the presence of the appropriate salts is strongly pH dependent (see Fig. 120). Radical reactions produce the intermediate hypochlorite [383], which in alkaline medium also acts as an

oxidizing agent. Chlorination at carbon is clearly apparent under acidic conditions, as implied by an equilibrium shift in the system

$$Cl_2 + H_2O \rightleftharpoons HOCl + HCl$$

This effect is even more pronounced in the case of the higher halides (Fig. 121) [383], [385], [386], as shown by the following data on C–Hal formation in conjunction with TOC oxidation:

Halide	Threshold pH
Chloride	< 5
Bromide	< 10
Iodide	< 12

Oxidative cleavage predominates in all three cases at a pH > 4.

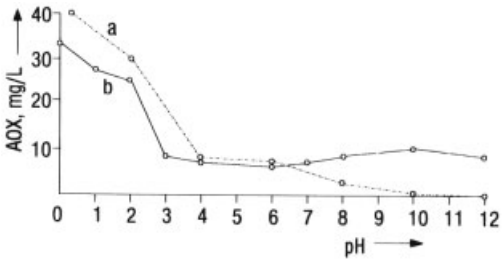

Figure 121. Ozonization of phenol (176 mg/L) in the presence of bromide and iodide ions
a) 500 mg/L I^-; b) 500 mg/L Br^-

The oxidative treatment of wastewater containing salts requires the prevention of simultaneous and persistent halogenation. A high chloride content must be anticipated in seepage water from landfills and in certain effluent streams derived from production lines in chemical processing plants.

Bromide in drinking water leads to the formation of organobromine compounds upon ozonization [385]. Special measures must be instituted in dealing with bromide- and iodide-containing wastewaters derived from the processing of color negative films and the production of printing plates [381]. In the case of iodide- containing wastewater it is useful to remove the iodide prior to oxidation of other constituents. With bromide- containing wastewaters it is possible to degrade the adsorbable organic halide (AOX) formed in situ during the course of oxidation by introducing an excess of oxidizing agent.

Radicals derived from natural ozone, and from natural, extracellular sources of hydrogen peroxide in the soil, especially the hydroxyl radical $H-O^{\bullet}$ [387–389], also produce halogenated intermediates in the presence of ubiquitous salts, including halogenated metabolites. Some of the latter are intermediate products with a limited lifetime in the mineralization process, whereas others are persistent excretion products in the form of humic acids with halogenated aromatic groups. Similar processes are observed in the ocean and in inland waters.

Corresponding process engineering in industry (i.e., oxidative degradation reactions with the simultaneous formation of intermediate organic halogen derivatives) has so far not been reported, although oxidative degradations in the form of natural decomposition and decay induced by natural oxygen sources together with biochemical catalysts (e.g., oxidases, peroxidases) have been simulated in the laboratory [389]. Specific proposals were developed in conjunction with this study for the industrially very important degradation of lignin in the wastewater released from paper and pulp mills.

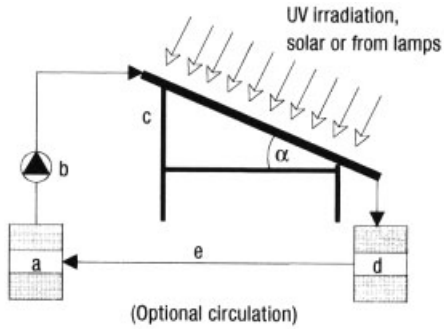

Figure 122. Schematic representation of a thin-film fixed-bed reactor for solar detoxification
a) Wastewater source; b) Pump; c) Catalytic surface; d) Effluent; e) Recirculation (optional)

4.3. Photocatalysis in Oxidative Degradation by Air

Semiconductor catalysts such as the oxides of titanium, iron, or zinc, together with atmospheric oxygen as the oxidizing agent and sunlight as a source of the requisite activation energy [362], [390] constitute systems for the treatment of dilute wastewater. These systems have already

been investigated in the context of photochemical degradation of organic substances in the atmosphere (→ Air, Chap. 8.) [391], [392].

Apart from studies involving individual substances [393–395], the behavior of actual wastewaters [396] has also been investigated in a thin-film fixed-bed reactor (Fig. 122), including groundwater, seepage water, influents to biochemical wastewater-treatment plants, and the corresponding effluents [362]. A decrease in TOC is observed even under neutral conditions, and at a pH of 3 or less the residual TOC may be as little as 30 % of the initial value (Fig. 123).

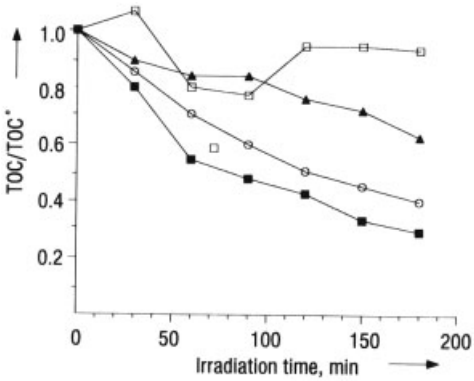

Figure 123. Relative TOC content as a function of irradiation time in the photocatalytic treatment of seepage water (volume = 1,4 L; $[TiO_2] = 1$ g/L; $[H_2O_2]^0 \approx 1500$ mg/L; $[TOC]^0 \approx 320$ mg/L; $[COD]^0 \approx 900$ mg/L; $[BOD_5]^0 \approx 10$ mg/L; O_2 flow rate ≈ 35 L/h; $T = 20-23$ °C; UV irradiation ≈ 50 W)
□ pH 11; ▲ pH 7; ○ pH 5; ■ pH 3

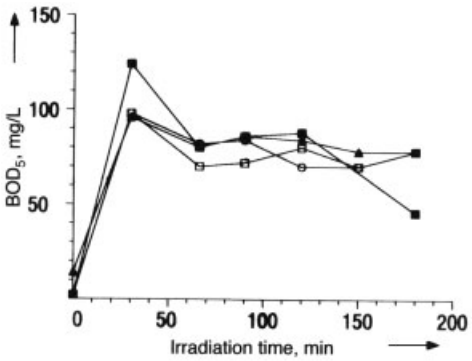

Figure 124. BOD_5 as a function of irradiation time in the photocatalytic treatment of seepage water (volume = 1.4 L; $[TiO_2] = 1$ g/L; $[H_2O_2]^0 \approx 1500$ mg/L; $[TOC]^0 \approx 320$ mg/L; $[COD]^0 \approx 900$ mg/L; $[BOD_5]^0 \approx 10$ mg/L; O_2 flow rate ≈ 35 L/h; $T = 20-23$ °C; UV irradiation ≈ 50 W)
□ pH 11; ▲ pH 7; ○ pH 5; ■ pH 3

The early stages of the process are characterized by an increase in biological oxygen demand, with the concurrent formation of biodegradable fragments (Fig. 124).

A consideration of the associated energy factors shows that titanium dioxide absorbs light only at wavelengths below 390 nm because of its relatively large band gap of 3.2 eV. Accessible sunlight with a radiant power of 20 – 30 W/m^2 in the wavelength range 300 – 400 nm makes available 0.2 – 0.3 photons per m^2 per h. Quantum yields of 0.1 – 10 % have been reported at TiO_2 surfaces. Assuming an average quantum yield of 1 %, this energy source would suffice for the degradation of 2 – 3 mmol of wastewater components per square meter of surface area per hour. For substances with molecular masses of ca. 100, present in concentrations of ca. 1 mg/kg, the potential wastewater throughput would be 200 –300 L/h for an absorption surface area of one square meter. Lamp panels equipped with standard fluorescent lamps of the type developed for tanning benches have also been used as sources of near-UV radiation. The corresponding degradation reactions proceed more rapidly with hydrogen peroxide as a sole or supplementary source of oxygen.

Variants such as parabolic channel reactors have been found to be much less effective than thin-film fixed-bed reactors [362]. Such chemical variants as hydrogen peroxide in conjunction with mixed oxides based on Fe(III)/Ti(IV) have led to better utilization of the available visible light, because α-Fe_2O_3 absorbs at wavelengths under 540 nm [362], [397]. Photoreactivity has been found to increase in the presence of mixed oxides of this type as a result not only of photooxidation but also photoaddition [398], which may prove to be of some importance in the context of degradation intermediates.

4.4. Wet Oxidation by Oxygen Sources in the Aqueous Phase

Oxidative degradation of wastewater components by means of air or oxygen under pressure and at elevated temperature is generally referred to as "wet oxidation." The essential features of this process have been known since 1912, and wet oxidation in the form of the Zimpro process has been in use since 1958 [399], mainly

in the treatment of wastewater from the pulp industry and municipalities [400–422]. The number of such plants currently in operation worldwide is rather small, but of all the non-biological wastewater treatment processes, wet oxidation is at present the most important. As a result of its ready adaptability to a wide variety of conditions relevant to the wastewater problems of both industry and municipalities, this process is attracting increasing interest.

Apart from its use in the oxidation of inorganic components, wet oxidation is also being implemented increasingly for the oxidation of organic constituents in wastewater, although the requisite temperatures and pressures are significantly higher than in the otherwise customary approach of clarification-sludge conditioning [359], [366]. As with other alternative processes, the advantage of wet oxidation lies in its potential for the treatment of partial wastewater streams containing components that degrade only slowly in a biological wastewater-treatment plant. Another advantage is the possibility of dealing with large quantities of wastewater containing contaminants of high concentrations, often with reasonable investment and operating costs. Characteristics of the process include some tolerance for salt-like companion substances, slightly acidic operating conditions, and catalysis by heavy metals, primarily Fe(II,III). Assuming complete degradation of the starting materials, the reaction products from oxidation of compounds containing the elements carbon (organic), hydrogen, nitrogen, sulfur, phosphorus, and chlorine are the corresponding mineralization products: CO, CO_2, H_2O, N_2, SO_4^{2-}, PO_4^{3-}, Cl^-, and NH_4^+. Operating experience with such systems has been accumulating in the chemical industry since 1978, when the first of many wet-oxidation plants installed by Bayer first went into operation at facilities in Germany and elsewhere.

With air or oxygen itself as the oxygen source, a pressure in excess of 0.4 MPa, and temperatures > 60 °C, pronounced degradation can be achieved for certain classes of compounds. However, with other substance categories even the originally reported higher temperature of 170 – 220 °C proved insufficient for achieving satisfactory degradation rates over short residence times [366], [399].

In order better to assess the potential range of applications for wet oxidation it has proven advantageous to develop two process variants with differing reaction conditions. The corresponding distinctions with respect to temperature and pressure are also associated with important differences in process engineering factors, material considerations, and plant safety requirements. Low-pressure wet oxidation has been the subject of especially extensive development, as a result of which the applicability and potential of the process have been greatly expanded. Based on experience collected to date, a line can be drawn at 200 °C and 0.5 – 2 MPa between low-pressure wet oxidation and high-pressure wet oxidation.

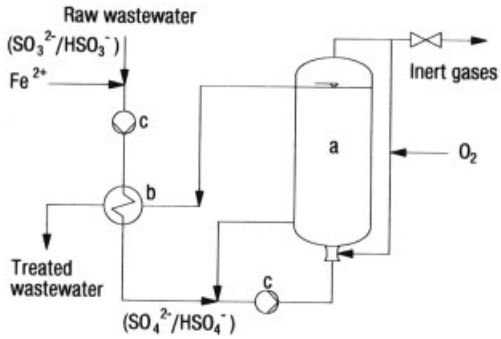

Figure 125. Flow diagram of a wet-oxidation plant for SO_3^{2-}/HSO_3^- - containing wastewater
a) Bubble column; b) Heat exchanger; c) Pumps

4.4.1. Low-Pressure Wet Oxidation

Inorganic substances also count among the oxygen-consuming wastewater components provided they are present in low valence states. Usually this involves reductive sulfur compounds in which sulfur has an oxidation number between − 2 and +4, most often wastewater containing the sulfite – hydrogensulfite system. The oxidation itself proceeds as follows:

$$SO_3^{2-}/HSO_3^- + O_2 \xrightarrow[\text{pH } 1-5]{Fe^{2+/3+}} SO_4^{2-}/HSO_4^-$$

Sulfite/hydrogensulfite-containing wastewater is continuously pumped into an oxidation bubble column by way of a countercurrent heat exchanger. Here it is oxidized by pure oxygen at temperatures between 60 and 100 °C and a pressure of 0.4 – 0.6 MPa to give a sulfate – hydro-

gensulfate mixture (Fig. 125). An aqueous solution of iron sulfate is used as the catalyst. The required oxygen is recirculated by means of a pump-driven ejector.

Provided no inert gases like nitrogen or carbon dioxide are present in the wastewater, no gas will accumulate in the gas space, and no off-gases are evolved. Only additional oxygen necessary for the oxidation must be supplied over the course of time via a pressure holding valve.

This mode of operation makes it possible to respond immediately to fluctuations in the sulfite concentration of the influent without the need for elaborate concentration measurements. The rate of oxidation is highly dependent on pH (Fig. 126). Oxidation is even possible without a catalyst; catalysis with Fe permits a doubling of the throughput.

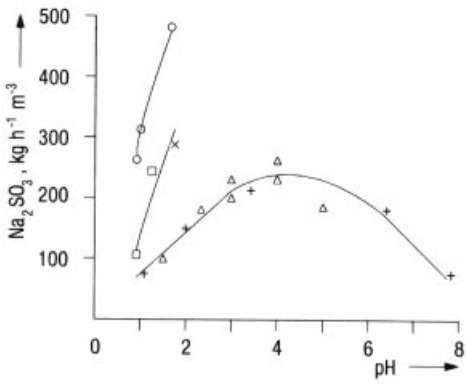

Figure 126. Influence of pH, temperature, and Fe^{2+} concentration on the rate of oxidation of Na_2SO_3 (p_{O_2} = 0.4 MPa)

Symbol	[Na_2SO_3], g/L	T, °C	[Fe], mg/L
+	35	70	
△	62	70	
□	62	90	
×	100	90	
○	100	90	50

Numerous processes important in chemical industries utilize sodium sulfide or hydrogen sulfide as a reducing agent. The wastewater that results contains not only sulfide, but also both sulfite and thiosulfate ions. Wastewater of this type is treated in a two-stage wet-oxidation plant like that shown in Figure 127, which operates at a temperature of 120 – 140 °C and a pressure of 0.5 – 0.6 MPa.

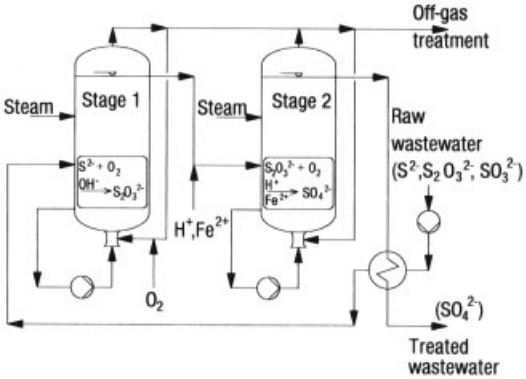

Figure 127. Flow diagram of a two-stage facility for the wet oxidation of wastewater containing reductive sulfur compounds

The first stage of such a facility accomplishes the rapid alkaline oxidation of sulfide to thiosulfate:

$$4\,S^{2-} + 3\,O_2 \xrightarrow{OH^-} 2\,S_2O_3^{2-}$$

The second stage operates under acidic conditions, thiosulfate being oxidized to sulfates.

$$S_2O_3^{2-} \xrightarrow{H^+} S^0\ \text{(colloidal)} + SO_3^{2-} \xrightarrow[Fe^{2+/3+}]{\frac{1}{2}O_2} SO_4^{2-}$$

In both cases the oxidizing agent is pure oxygen, recirculated by means of an ejector system.

The oxidation of thiocyanate demands more drastic conditions:

$$SCN^- + 3\,O_2 \xrightarrow[Fe^{2+/3+}]{H^+} SO_4^{2-} + NH_4^+ + CO_2$$

Catalytic amounts of iron will not suffice for this oxidation at 130 – 150 °C; stoichiometric amounts must be used instead. Catalytic amounts are adequate only in the temperature range 180 – 200 °C.

The oxidation of nitrogen-containing inorganic wastewater components is easily achievable with wet oxidation in the case of nitrite or hydrazine. The oxidation of nitrite takes the form

$$NO_2^- + \tfrac{1}{2}\,O_2 \xrightarrow{H^+} NO_3^-$$

so nitrite-containing wastewaters can be treated in the same way as sulfite-containing material

in a single-stage process that evolves no off-gas. The oxidation can be accomplished without a catalyst at 30 – 50 °C. With hydrazine- containing wastewater the oxidation can be represented as

$$N_2H_4 + O_2 \xrightarrow[Fe^{2+/3+}]{H^+} N_2 + 2\,H_2O$$

Nitrogen formed during the process must be removed along with the off-gas from the single-stage wet oxidation. Reaction conditions range from 50 – 120 °C at 0.4 – 0.5 MPa [423].

Organic wastewater components represent a much more fertile field of application. Many problematic wastewater streams arise during the course of syntheses carried out in the aqueous phase, some in separation steps associated with the isolation of intermediates, and others during the isolation and purification of final products. Some of this wastewater contains substances that degrade only very slowly in a biological wastewater-treatment plant, especially certain aromatic derivatives that can be degraded biologically only with difficulty, and then only with special bacterial cultures (see Section 2.1).

The treatment of wastewater containing such aromatic compounds can be accomplished at temperatures below 200 °C in the presence of an appropriate redox catalyst [423]; especially suitable for the purpose is the system quinone ⇌ hydroquinone Fe(II) – Fe(III):

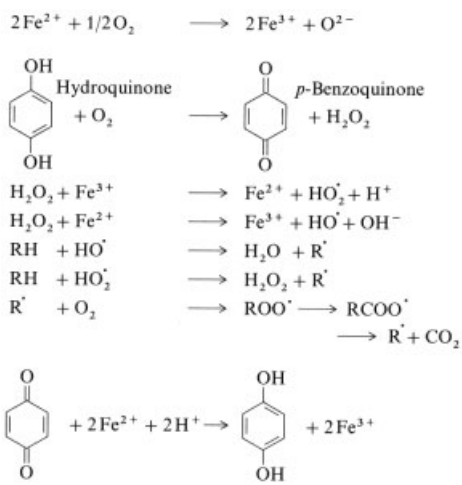

Quinone-forming agents and substances containing partial quinoid structures are also suitable. In the course of the catalytic wet oxidation of aromatic compounds, 60 – 90 % of the carbon is oxidized to CO_2 depending on the nature of the compound and the reaction temperature. Residual organic matter remaining in the wastewater consists chiefly of metabolites like acetic acid and acetone, which are subject to further biological degradation. Organic nitrogen is converted mainly to ammonia.

A time vs. conversion curve (Fig. 128) based on the wet oxidation of phenol (2 g/L, 140 °C, 0.3 MPa, Fe 0.3 g/L, quinone 0.3 g/L, pH 2) reveals this to be a first-order reaction. After a reaction time of 90 min the TOC value is reduced by 70 %; indeed, phenol is no longer detectable by gas chromatography after only 45 min. No phenol oxidation whatsoever is found to occur under otherwise identical conditions in the absence of the catalyst combination.

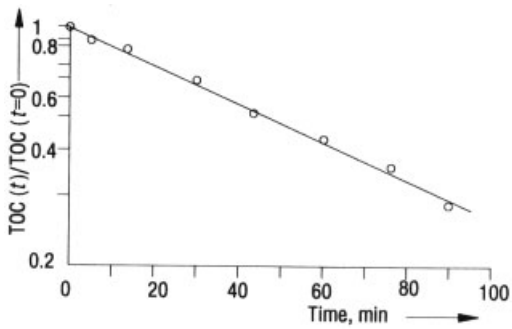

Figure 128. TOC elimination from a phenol solution by catalytic wet oxidation at 140 °C (see text for details)

Studies based on other aromatic compounds indicate that amino- and hydroxyl-substituted phenols, naphthols, and their analogues are readily degraded by low-pressure oxidation at 120 – 150 °C with Fe – quinone catalysis, whereas sulfo and nitro derivatives are less readily degraded [359], [366], [423]. Most other compounds require an operating temperature of 180 – 190 °C.

When wastewater derived from the production of anthraquinone dyes is oxidized in a multistage bubble column (Fig. 129) at ca. 190 °C and 1.6 – 1.8 MPa, a COD reduction of 85 % is typically achieved. As much as 90 % of what remains is subject to subsequent degradation in a biological waste-treatment plant. Taken together these results amount to a COD reduction of 98 % in the combined process, equivalent to almost

complete degradation. Wastewater of this type inevitably contains quinone-forming agents at a level high enough such that the further addition of quinone is unnecessary. In other cases it frequently suffices to introduce alkali-treated brown coal as a quinone-forming agent.

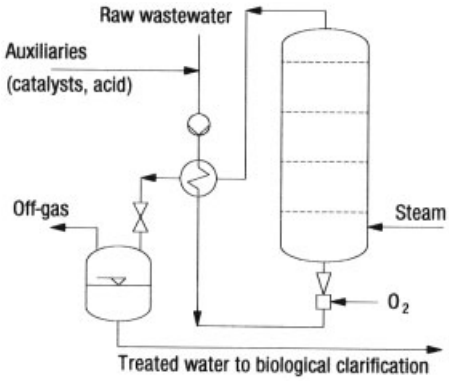

Figure 129. Flow diagram of a Loprox facility operated at 120 – 200 °C and 0.3 – 2.0 MPa with a retention time < 3 h

The degradation of aromatics is of great importance not only because a considerable fraction of chemical wastewater is derived from the production and further processing of aromatic compounds, but also because operators of waste-treatment plants are becoming increasingly interested in wet oxidation as an aid to the disposal of clarification sludge. Like many types of soil, however, clarification sludge always contains humic acids formed in the ubiquitous presence of sodium chloride, and characterized by aromatic nuclei bearing 1 – 3 atoms of chlorine [389].

With respect to aliphatic compounds, methanol and many carboxylic acids prove to be only slowly degradable by low-pressure wet oxidation, and in the case of amino compounds all that is observed is cleavage of the amino groups. Nevertheless, even this modest change results in a clear improvement in biodegradability.

The process of low-pressure wet oxidation is in fact rather substance-specific. Given that the extent of COD reduction for a wide variety of organic compounds ranges from 15 – 90 %, a targeted investigation is required to establish whether the process is suitable for use in a specific case. If tests are conducted on the basis of a batch operation it is necessary to take into account the fact that both pH and redox potential may shift under the test conditions. In a continuous operation various measures can be implemented to ensure that the reaction conditions remain constant at their optimal values.

One adaptation of wet oxidation with wide applicability, especially in conjunction with biological waste treatment, is the Loprox (Low-Pressure Oxidation, see Fig. 129) process developed by Bayer. This process accomplishes partial oxidation of a range of compounds that are only slowly biodegradable, and it is thus useful for preparation of a split wastewater stream prior to its introduction into a conventional waste-treatment plant [424]. Experience has now been collected from Loprox plants over a period of more than 15 years, including plants of various sizes (with wastewater throughputs of 4 – 60 m^3/h) [425], [426]. Residence time in the pressure reactor is 1 – 3 h. The preferred influent is wastewater with a COD range of 5 – 50 g/L. The oxidation reaction is exothermic, with the temperature rising during the course of degradation. A wastewater COD value of 6 – 8 g/L is sufficient to permit oxidation to proceed autothermically. Once the plant has been brought up to temperature no further input of heat energy is required.

Based on data from wastewater contaminated with difficult to degrade chloroaromatics, it is apparent that the COD and the AOX both decrease with increasing reaction temperature during the course of the Loprox process (Fig. 130), and that the decrease in AOX values is usually more rapid and of greater magnitude than the decrease in COD. The BOD_5 value increases simultaneoulsy.

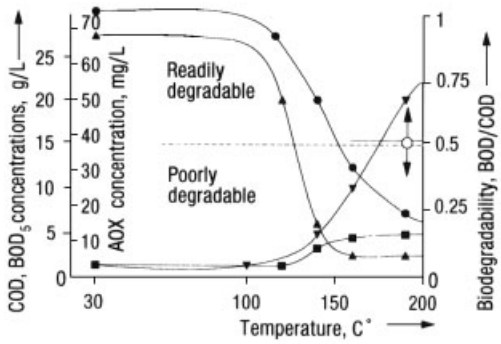

Figure 130. Typical course of reaction in the Bayer Loprox process
• COD; ▲ AOX; ■ BOD_5; ▼ BOD/COD

At a reaction temperature of 190 °C the AOX falls by more than 90 %, and the COD by more than 75 %. The ratio BOD_5/COD increases from less than 0.1 to above 0.5, a value characteristic of a readily biodegradable wastewater [424–426].

The conditions of the reaction impose certain restrictions on materials that can be used for the hot parts of the plant. Equipment that is enamelled or lined with PTFE is suitable for operating temperatures < 160 °C. Metal parts must be fabricated from titanium or its alloys with palladium for temperatures ≤ 200 °C. This also ensures corrosion resistance at high chloride-ion concentration. However, fluoride ion at a concentration > 10 ppm must be removed by precipitation with Ca(II) and subsequent filtration. Apart from the quantity of wastewater subject to treatment, operating costs also depend heavily on the reaction temperature. For a single-stage plant these costs range from ca. € 2.6/m^3 at 100 °C to € 10 – 20/m^3 at 200 °C [423], [427].

The applicability of the process in both industrial and municipal settings is largely dependent on the availability of a suitable process and disposal combination [426–429]. Several illustrative examples are examined in detail below.

Wastewater from Pulp Production. The Loprox process makes it possible under fairly mild reaction conditions to partially oxidize long- chain wastewater components to such an extent that they become biodegradable. Crude wastewater from pulp prodution usually has a COD of 20 – 30 g/L, of which at most 50 % represents biodegradable substances. At a temperature of ca. 150 °C and a residence time of one hour, the Loprox treatment results in a reduction in the COD to 9 g/L and a marked improvement in the biodegradability of the outflow based on a single-stage mode of operation. The residual COD can be reduced further to the extent of 92 % by subsequent biological treatment, resulting in an overall COD reduction > 97 %.

Clarification Sludge from a Biological Wastewater-Treatment Plant. Excess organic sludge is separated by sedimentation in the course of final clarification. It is then concentrated in a screen-belt drum to a solids content of ca. 5 % for subsequent degradation in the Loprox process. At a reaction temperature of 190 °C only 10 % of the starting solids remain. These solids consist mainly of mineral components that can be easily separated by sedimentation and then dewatered. Sludge destined for final disposal is thus reduced to 6 – 7 % of the amount that would be generated in a conventional operation.

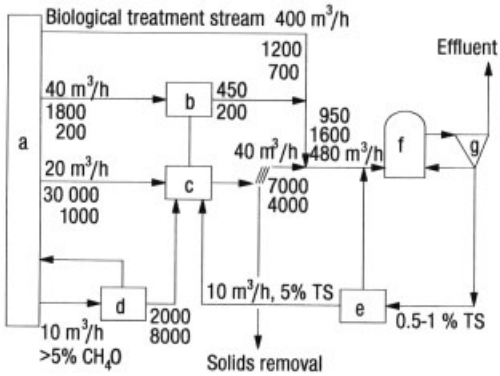

Figure 131. Wastewater treatment with integrated disposal of clarification sludge; numbers below the arrows correspond to values (mg/L) for COD and BOD, respectively (TS = total solids)
a) Production plant; b) Precipitation/flocculation; c) Loprox unit; d) Stripping; e) Sludge concentration; f) Biological treatment tower; g) Sludge separation

Combination Disposal Networks in the Context of Chemical Plants. Wastewater treatment at large chemical plants normally occurs in centralized treatment facilities following decentralized pretreatment. The central facility usually consists of a neutralization stage, a preclarification stage, and an aerobic biological wastewater-treatment system with final clarification (see Section 2.1.4). Inclusion of a Loprox plant [428–430] facilitates the handling of problematic split streams from specific production areas as well as from the disposal area itself (Fig. 131). In one particular case, a split stream with a COD concentration of 1.8 g/L, too dilute for direct inflow into a wet-oxidation system, is subjected to a preatreatment consisting of precipitation and flocculation. The resulting concentrate is fed into the Loprox wet-oxidation stage (c), and the overflow into the biological clarification stage (f). The wet-oxidation facility simultaneously receives split streams rated at 20 m^3/h and 30 g/L COD as well as 10 m^3/h and 8 g/L COD, along with concentrated excess sludge from the biological clarification stage. Inorganic matter from

the sludge residue after wet oxidation together with the iron catalyst is discharged, dewatered, and placed into the waste depository. Wastewater subjected to this pretreatment is pumped directly into the biological clarification stage. Coupling the process in this way offers many advantages with respect to both disposition and operation. Nevertheless, the COD cycle associated with clarification-sludge oxidation requires the introduction of 20 % more oxygen into the biological clarification stage than would be the case in a conventional operation.

Split Pretreatment in Conjunction with Dye Manufacture. A two-stage Loprox plant with a throughput of 20 m^3/h is used by Bayer to oxidize primarily inorganic reductive sulfur compounds at a temperature of 120 – 140 °C and a pressure of 0.6 MPa. The pH is subsequently adjusted to a value of 1 – 2 by the addition of sulfuric acid, and a second oxidation is performed in a separate stage at 140 °C in the presence of a catalyst. At the time this process was instituted (1982) it utilized for the second stage the largest one-piece enamel tower ever built, with a height of 10.5 m and a diameter of 1.8 m (Fig. 132).

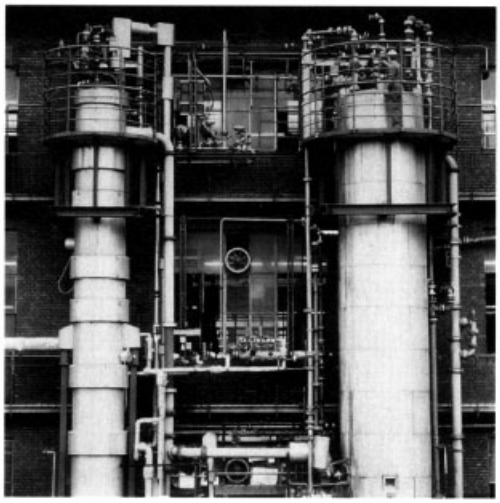

Figure 132. Two-stage Loprox unit for the pretreatment of 20 m^3/h of wastewater from dye production at the Bayer plant in Leverkusen, Germany

Split Pretreatment in the Production of Organic Intermediates. Using a facility with the same throughput as above but a single-stage mode of operation at 190 °C and 2 MPa it is possible to achieve a chemical wastewater COD reduction of 85 % by wet oxidation, followed by a further 90 % reduction in the biological clarification step, for a total reduction of 98 %. The single-stage column was lined with titanium because of an anticipated chloride content as high as 6 % in the inflow.

Split Pretreatment for Wastewater from Pharmaceutical Production. Wet oxidation proceeds in this case with a throughput of ca. 4 m^3/h at an operating temperature of ca. 160 °C in a two-stage plant with enamel towers, where each tower has a diameter of 1 m and a height of 10 m (Fig. 133). This particular plant was put in operation in 1993.

Figure 133. A Loprox unit for the pretreatment of 4 m^3/h of wastewater from pharmaceutical production at a Bayer subsidiary in La Felguera, Spain

Combination Disposal in a Municipal Wastewater-Treatment Plant. Even if excess sludge that accumulates constantly from aerobic degradation of wastewater components is subsequently degraded further anaerobically (see Sec-

tion 2.2), there still remains a certain amount of sludge to be disposed of. Such sludge typically has a solids content of almost 50 %, 20 to 30 % after dewatering. Until recently sludge of this type was largely consigned to landfill or used for agricultural purposes. In the western part of Germany, for example, only about 10 % of the sludge has been routinely incinerated. Nevertheless, both waste deposition and agricultural use are subject to strict laws and regulations. Combustion is inherently expensive, and it will become more expensive still as increasingly stringent flue-gas restrictions take effect.

Residual sludge disposal by wet oxidation can be conducted as a high-pressure process in the temperature range 250 – 300 °C and at a pressure > 15 MPa, as has been demonstrated by facilities in the United States, Japan, and The Netherlands [430]. Pilot plants have also been operated on the basis of low-pressure wet oxidation for the treatment of both pure biomass derived from fermenters and excess sludge from industrial wastewater-treatment plants [424]. Sludge removed from digester towers is first subjected to wet milling to give particles less than 100 – 200 µm in diameter, after which the mass is adjusted to a pH of 2 and subjected to the Loprox process in the form of a material with a dry-weight solids content of 12 – 15 % (Fig. 134).

The reactor itself is a bubble column without internal fittings, in which the sludge is brought into contact with pure oxygen at 190 °C and 2 MPa over the course of a residence time of 2 – 3 h. It is ordinarily not necessary to add mixed catalyst, because sludge of this type usually contains sufficient amounts of both iron and quinoid substances. After cooling and expansion, solids separation is accomplished by means of filter presses or decanters. The remaining sludge is almost entirely inorganic, and it can easily be dewatered to less than 50 % dry matter. The organic content will have been reduced by about 85 %; almost of the original organic carbon is converted to CO_2, while 1/3 is returned to the biological clarification stage in the form of readily degradable wastewater. Most of the nitrogen originally present resides in this return flow as ammonium ion. If the biological clarification stage includes nitrification, this NH_4^+ is degraded here; otherwise it is eliminated by stripping and subsequent reuse. The ammonia-free return flow from the Loprox process can be utilized as a carbon source and fed directly into the denitrification step of the biological clarification stage [426].

Communities hosting a large number of small and medium-sized factories often encounter problems with poorly degradable wastewater derived from industrial production. In many cases the quantities of waste generated are rather small, and economically viable pretreatment processes might not exist. It has been proposed that such problematic wastewater might be transported in tank trucks to a municipal Loprox plant (Fig. 134) [426]. Addition of a catalyst would not be required, because sufficient catalytic matter is present in the clarification sludge. Experience [426], [427] suggests that current specific operating costs for the treatment of excess sludge from a waste-treatment plant amount to about DM 250 – 350 per ton of dry matter. In an average situation it is possible in this way to achieve a 75 % reduction in sludge volume and an 85 % reduction in the weight of organic dry matter.

Decontamination of Dredger Sludge and Soil. Low-pressure wet oxidation has been used successfully in the treatment of dredger sludge from harbors [426]. The contamination in this case is due mainly to mineral oils and heavy metals. The solids are classified according to particle size directly on the dredging vessels with the aid of wet- cyclone classifiers. Most of the organic contamination is associated with the finest components (< 0.5 mm in diameter). These consist largely of petroleum products, polycyclic aromatic hydrocarbons (PAHs), and chlorinated hydrocarbons. The situation is similar with contaminated soil derived from ground washing. The organic components can be eliminated almost entirely subsequent to Loprox treatment. Thus, an analysis (Table 11) covering the 16 compounds on the United States Environmental Protection Agency (EPA) Register for PAHs shows that more than 99 % of the material in question is destroyed. After separation of the oxidatively treated solids, heavy metals that pass into solution during acidic oxidation are precipitated by the addition of alkali, filtered, and disposed of separately.

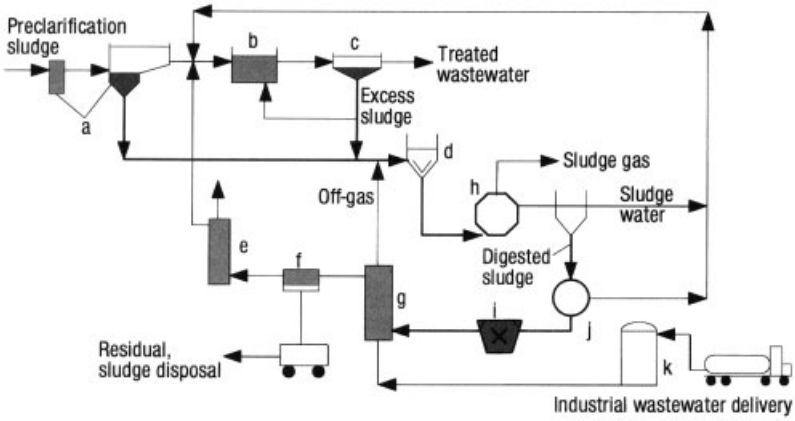

Figure 134. Flow diagram of a municipal waste-treatment plant with an integrated Loprox stage for treating clarification sludge
a) Preclarification; b) Biological treatment (carbon, nitrogen, and phosphorus removal); c) Post- clarification; d) Concentrator; e) Ammonia stripping (as required); f) Filter press; g) Loprox unit; h) Digester tower; i) Homogenization; j) Decanter; k) Storage tank

Table 11. Polyaromatic hydrocarbons (PAHs) in the solids (<0.5 mm) of contaminated dredger sludge before and after oxidative degradation in the Bayer Loprox process

Substance	Before treatment, mg/kg	After treatment, mg/kg
Naphthalene	1.1	< 0.2
Acenaphthalene	–	–
Acenaphthene	44	0.3
Fluorene	11	< 0.2
Phenanthrene	98	0.6
Anthracene	44	< 0.2
Fluoranthene	200	< 0.2
Pyrene	162	0.2
Benz[*a*]anthracene	108	< 0.2
Chrysene	122	< 0.2
Benzo[*b*]fluoranthene	108	< 0.2
Benzo[*k*]fluoranthene	44	< 0.2
Benz[*a*]pyrene	112	< 0.2
Dibenz[*ah*]anthracene	19.6	< 0.2
Benzo[*ghi*]perylene	66	< 0.2
Indeno[1.2.3.*cd*]pyrene	74	< 0.2
Total 16 EPA substances	1213.7	< 3.5

4.4.2. Oxidation by Nitric Acid

Nitric acid and its salts have been used as a source of oxygen at the borderline between high- and low-pressure wet oxidation [431]. For TOC degradation of wastewater the preferred reaction temperature is in the range 180 – 220 °C. Ammonia is converted to nitrogen. Here, too, the oxidation can be carried out partially and subsequently completed in a biological clarification stage.

4.4.3. High-Pressure Wet Oxidation

With some types of wastewater associated with dye manufacture or the further processing of organic intermediates, the level of degradation achievable in low-pressure wet oxidation is insufficient. Much more severe reaction conditions are required to attain satisfactory results. The problem is aggravated by the fact that accompanying inorganic salts (which need not in themselves represent an environmental or material problem) become extremely corrosive under the conditions of high-pressure wet oxidation, leading to serious problems with construction materials. The reaction temperature is inevitably 280 °C, and the most important parameters are the pH and the nature of the catalysts. Depending on the situation, either of two forms of high-pressure wet oxidation may be considered:

1) An alkaline procedure in the absence of heavy-metal catalysts like iron and copper (which would precipitate under these conditions).
2) An acidic process that does take advantage of heavy-metal catalysis, with special attention to the materials problem.

The line is drawn between these two approaches essentially on the basis of the concentration of chloride ions, which is in turn a function of the inorganic salt load of companion substances. In the case of wastewater with a chloride concentration ≤ 35 g/L, high-pressure wet oxidation of the acidic type can be conducted in titanium autoclaves, but only in the absence of fluoride ions. Even a low concentration of F^- consistent with the solubility product of CaF_2 is sufficient to initiate corrosion at a pH < 5. Construction materials that have proven suitable include Ti-3.7035 for heated areas and a chloride-resistant stainless steel such as 1.4439 (Antinit, Novonox, Remanit) for cold parts (Fig. 135).

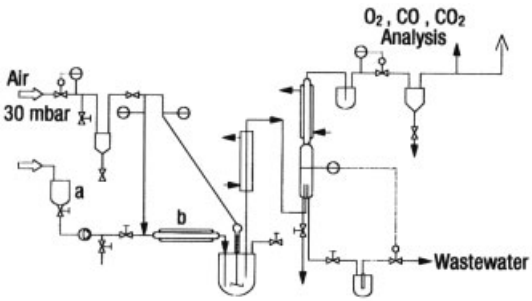

Figure 135. Pilot plant for continuous wastewater treatment with the aid of high-pressure wet oxidation
a) Wastewater; b) Heat exchanger

The following general rules apply to the expected extent of oxidative degradation as a function of classes of substances and reaction conditions [359], [366]:

1) For a given residence time and otherwise comparable conditions, TOC degradation proceeds further under acidic conditions than under alkaline conditions, with the difference becoming less pronounced as the temperature increases (Table 12).
2) Under acidic conditions with a fixed residence time and a temperature $\leq 280\,°C$, TOC degradation can be increased by an average of ca. 12 % by copper catalysis. A Cu(II) concentration of 500 ppm is sufficient. Catalysis loses its importance with increasing temperature. Above 300 °C, a TOC degradation > 90 % is achievable even without heavy-metal catalysis (Table 13).
3) Irrespective of the pH, nitrogen- containing compounds not subject to protonation are degraded to N_2, CO_2, CO, and H_2O. The same applies to protonatable substances under alkaline conditions. In an acidic process, nitrogen compounds that can be protonated are degraded only to ammonia, which undergoes almost no oxidation under these conditions.
4) The COD/BOD ratio after partial oxidation attains a value < 2.5, consistent with the requirements for biodegradable wastewater.

Table 12. Degradation by uncatalyzed high-pressure wet oxidation as a function of pH

No.	Contaminant		pH	TOC reduction at 280 °C, % of initial value	
1	Aromatic	10		76	
	amine		1.1		83
3	Nitrogen	10		85	
	heterocycle		1.2		92
6	Azo	10		97	
	compound		1.5		99
10	Nitrosulfonic	10		76	
	acid		1.5		83

An important factor in estimating the costs associated with high-pressure wet oxidation [426], [427] is the concentration of organic wastewater components, especially in view of increasing energy prices. Above 50 g COD/L of wastewater, heat recovered from the enthalpy of oxidation is comparable to the total energy requirement for the process. Several plants are already in operation in various parts of the world [359], [366], [430], [432], [433] (Fig. 136).

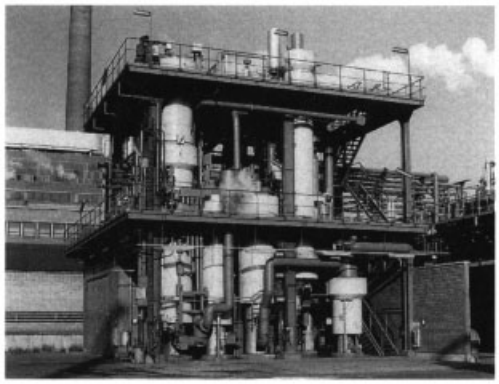

Figure 136. Prototype facility for the high-pressure (≤ 20 MPa) wet oxidation of industrial wastewater (cyclohexanone oxidation at the Bayer plant in Uerdingen, Germany)

Table 13. Degradation by uncatalyzed high-pressure wet oxidation as a function of temperature

No.	Contaminant	Concentration, g TOC/L	pH	TOC reduction, % of initial value	
				280 °C	300 °C
1	Aromatic amine	14.6	10	76	76
2	O heterocycle	52	11.5	80	87
3	N heterocycle	14.3	10	85	91
4	Cl- containing sulfonic acid	3.5	1.6	90	93
5	Aminosulfonic acid	23.8	1.4	80	86
6	Azo compound	13.1	10	97	98
7	Aminosulfonic acid	20.2	2.2	87	88
8	Aminopolysulfonic acid	47.6	0.5	64	65
9	Nitro compound	6.0	9.3	93	94
10	Nitrosulfonic acid	14.0	10	76	84
11	N heterocycle	1.1	3.1	77	88
12	Oxy compound	9.2	2.2	75	81
13	Mixed wastewater	11.4	1.9	77	81

The extreme case of a high-pressure wet oxidation occurs at the critical point (i.e., 374 °C and 21.8 MPa), and this has been the subject of investigation on a laboratory scale [434]. Organic wastewater components were found to be almost completely degraded in the course of extremely short residence times. This result may be due to unlimited solubility of the materials in "water" under these conditions. On the other hand, salts precipitate. The most serious problems under these reaction conditions relate to construction materials.

4.5. Hydrogenation

Degradation of wastewater components with the goal of mineralization leads to carbon dioxide and water in the case of the elements carbon, hydrogen, and oxygen. Other elements such as phosphorus and sulfur are oxidized in an analogous way to sulfate and phosphate. In the case of nitrogen, it is the element itself to which the corresponding compounds are to be converted in the mineralization process, a result that cannot be achieved directly by biochemical means (see the discussions of nitrification and denitrifiction in Section 2.1.1.3).

The degradation of nitrates directly to elemental nitrogen is accomplished by using hydrogen. This process is of special interest with respect to avoiding groundwater contamination, or in the pretreatment stage of drinking water preparation. Roughly $^2/_3$ of the nitrate present in crude water can be reduced in this way to nitrogen in the presence of a noble-metal- coated carrier.

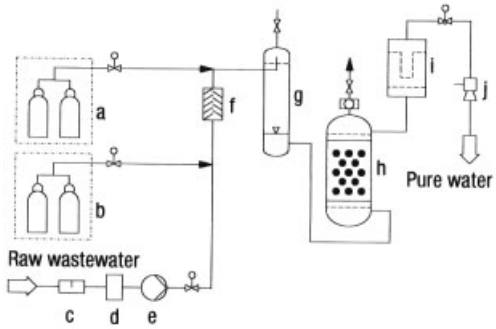

Figure 137. Process flow diagram for a wastewater hydrogenation system
a) Hydrogen; b) Carbon dioxide; c) Filter; d) UV treatment; e) Pump; f) CO_2 mixer; g) Hydrogen saturator; h) KNR reactor; i) Filter; j) Aerator

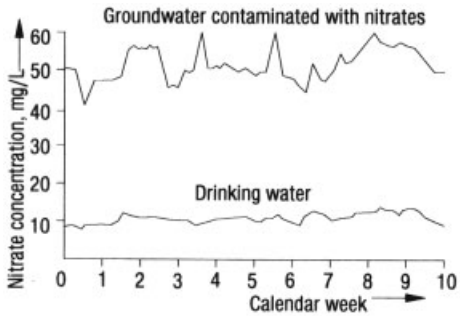

Figure 138. Variation in nitrate concentration as a function of time

The so- called KNR process has been implemented in a pilot plant that operates with a

throughput of 3 m^3/h and consistent long-term nitrate degradation at a rate of 45 mg NO_3^-/L (Fig. 137). This would correspond to a degradation of up to 70 kg of nitrate per cubic meter of catalyst per day.

The nitrite concentration in clean effluent water amounts to < 0.1 mg/h, and the nitrate concentration ≤ 17 mg/L (Fig. 138).

5. Physical Wastewater Treatment

5.1. Adsorptive Purification [435–437]

Adsorption processes have been successfully applied in the chemical industry for the purification of water generally, as well as for various solutions and individual wastewater streams.

In the chemical, physical, and biological purification of wastewater dissolved constituents are eliminated not only by use of special absorbents, but also in many cases by adsorption on the surface of undissolved substances already present. Elimination occurring by the latter mechanism cannot be examined separately, so it is simply ascribed to the treatment process as a whole.

The deliberate application of adsorption is always in competition with other chemical, physical, and biological approaches to the purification of wastewater. Ecological and economic considerations are decisive in the choice of any particular process or process combination. The question as to whether adsorption is technically and economically feasible, either alone or in combination with other processes, must be examined separately in each individual case. To this end it is necessary to consider the adsorption phenomenon itself in conjunction with the preparation and regeneration of the adsorbent, since from a process-engineering standpoint the two processes constitute a single entity. In addition, reuse or disposal of materials arising in the course of the process should be taken into account as appropriate.

Adsorption causes dissolved organic wastewater constituents to accumulate at the surface of one or more adsorbents. Adsorption must therefore be regarded as a physical concentration process taking place at a liquid – solid phase boundary, one that competes with other concentration processes occurring at liquid – liquid (extraction or membrane processes) and liquid – gas boundaries (evaporation, distillation, stripping). In the unified examination that follows, it was necessary that some attention also be devoted to such oxidative processes as aerobic biological treatment, chemical oxidation, wet oxidation, and combustion (treated elsewhere in this contribution in detail).

It may be assumed in general that adsorption processes have already established themselves as valuable for dealing with fairly small wastewater streams containing low concentrations of adsorbable substances. Adsorption is especially common in decentralized wastewater treatment, both alone and in combination with other processes.

5.1.1. Adsorption

Dissolved wastewater constituents are potentially subject to attachment at the surfaces of solids. An accumulation of this type is known as *adsorption*, and it is attributable mainly to van der Waals forces, particularly dipole – dipole interaction, though coulombic forces also often play an important role (→ Adsorption). The fixing of a dissolved substance to a sorbent can sometimes lead to an enrichment factor of 10^5 or greater. In most cases the adsorbent – adsorbate mixture must then be worked up in a second processing step. Regeneration of a contaminated sorbent frequently takes advantage of the reverse counterpart to adsorption: *desorption*.

5.1.2. Adsorption Isotherms

The equilibrium state that is established after a sufficient amount of time has elapsed defines the equilibrium concentration of a solute in a liquid and the loading of an adsorbent. This state can be described in terms of so-called *adsorption isotherms* (→ Adsorption, Chap. 4.). An isotherm is applicable only to a specific, defined temperature, although the influence of temperature is very small in liquid – solid systems in contrast to gas – solid systems. The curves themselves are usually determined empirically, but their shape can also be described mathematically.

Langmuir Isotherms. One of the first attempts at developing an isotherm equation was reported by LANGMUIR (1918), who was concerned originally with the adsorption of gases, basing his work on methods of statistical thermodynamics. The Langmuir equation assumes the presence of a homogeneous adsorbent surface, and it is therefore valid only to the point of a monomolecular covering of the relevant surface. In general, Langmuir isotherms are described by the equation:

$$X = X_{\mathrm{m}} \frac{C}{b+C}$$

where X = equilibrium loading, X_{m} = loading for monomolecular coverage, C = residual solution concentration remaining after establishment of the adsorption equilibrium, and b = a constant.

Freundlich Isotherm [438]. The isotherm equation proposed by FREUNDLICH as early as 1906 was established empirically. Because of its relatively simple form it is an excellent device for describing adsorption from an aqueous solution:

$$\frac{X}{M} = kc^n$$

where X = amount of adsorbed substance, M = weight of adsorbent used, C = residual concentration remaining after establishment of the adsorption equilibrium, and k and n are constants specific to a given adsorbent – adsorbate mixture.

Plotting this equation in double-logarithmic form [log (X/M) = log $k + n$ log C] produces a straight line, which greatly facilitates practical application of the method (Fig. 139).

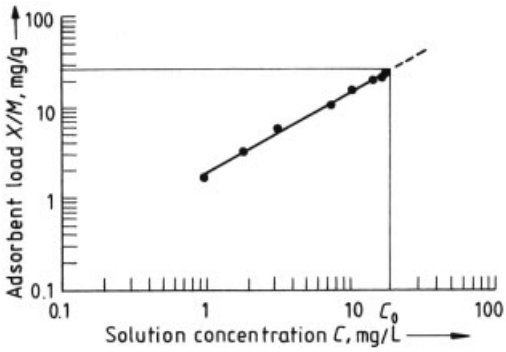

Figure 139. Example of a well-behaved Freundlich isotherm

The Langmuir equation is clearly limited to low degrees of loading, whereas in the case of the Freundlich equation there is no formal concentration-dependent limitation to the expression's validity. In practice, however, a flattening of the Freundlich isotherm does become evident with increasing concentration, which means that the Freundlich exponent (n) increases with decreasing concentration. Thus, rigorous application of the Freundlich equation is limited to a rather narrow concentration range.

Isotherms for Multicomponent Mixtures. Multicomponent mixtures are the rule in practical wastewater treatment, and these can be characterized only by such cumulative parameters as COD, AOX, or UV absorbance. Individual substances included in a collective analysis compete for occupation of the available adsorption sites. Since different substances usually have very different adsorption behaviors, the Freundlich isotherms actually obtained in such cases are curved, or even bent at sharp angles (Fig. 140).

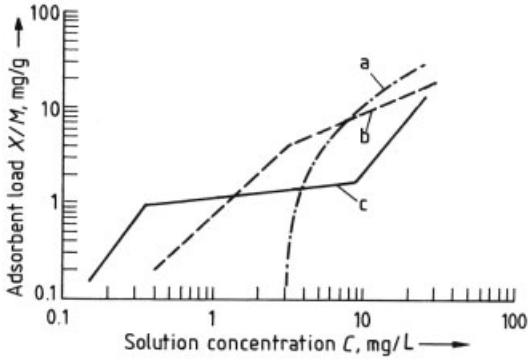

Figure 140. Typical non-linear Freundlich isotherms (a, b, c) obtained with different mixtures of solutes

Despite the complications created by isotherms whose course is not straight, the Freundlich isotherm method is frequently used as a way of rapidly acquiring information regarding adsorbability on adsorbents. Moreover, such isotherms make it possible to estimate adsorbent consumption rates as well as the achievability of particular concentration targets.

There have been very many methods proposed for approximating such mixture isotherms mathematically. All of them are based on the ideal adsorbed solution (IAS) theory. According to this theory of the "ideally adsorbable

solution," mixture isotherms can be computed from single-substance data. A standard method in drinking-water treatment for the arithmetic construction of mixture isotherms is based on approximating the comprehensive results in sections. An extensive description of various approaches to a solution is available in et al. [439].

5.1.3. Adsorption Kinetics

The adsorption process comprises a series of diffusion processes that together determine the overall rate of the phenomenon:

1) Transport of a dissolved molecule from solution to the laminar boundary layer of the adsorbent
2) Diffusion through the laminar boundary layer surrounding the adsorbent
3) Grain diffusion into the interior of the pore system all the way to the active centers

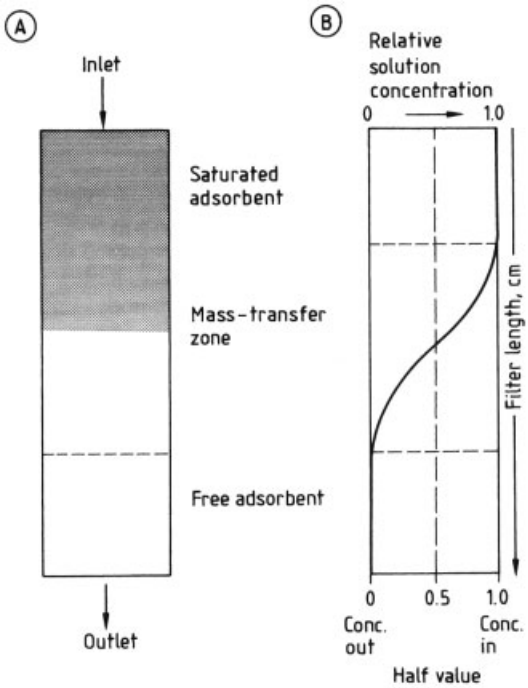

Figure 141. Mass-transfer zone in an adsorption filter bed
A) Schematic depiction; B) Typical breakthrough curve

While step (1) generally proceeds rapidly, step (3) is largely responsible for determining the kinetics of the overall process. Grain diffusion depends, among other things, on the relationship between the size of the adsorbate molecules and the pore diameter, but especially on the length of the diffusion path (i.e., the grain size of the adsorbent).

In a fixed-bed absorber with continuous flow, a combination of all the active diffusion processes leads to the development of an *adsorption* or *mass-transfer zone* (Fig. 141 A). As adsorption proceeds, this mass-transfer zone migrates through the adsorbent bed in the direction of flow. When the front reaches the outlet, *breakthrough* of the dissolved substances begins. Figure 141 B shows an example of a typical breakthrough curve.

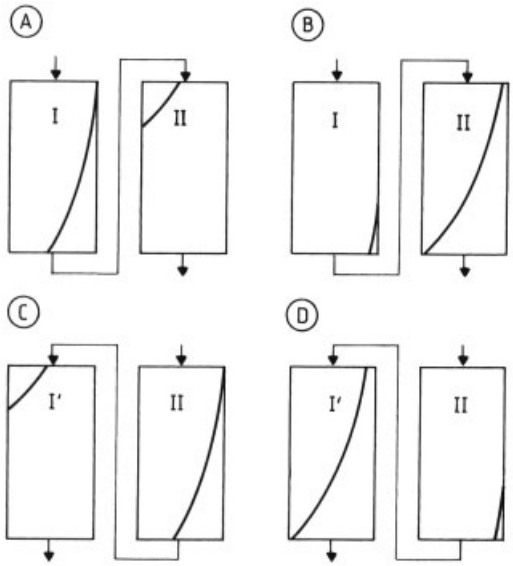

Figure 142. Operation of two filters (I and II) connected in series
A) Sequence I → II, early in the process; B) Sequence I → II, when filter I is due to be replaced; C) Sequence II → I′, shortly after the replacement of I; D) Sequence II → I′, when filter II is due for replacement

The length of the mass-transfer zone is a key criterion in establishing the optimum adsorber size for a prescribed wastewater stream. For any given starting solution and pore distribution with respect to the adsorbent, the length of this zone depends exclusively on the grain size and the contact time; i.e., on the ratio of the throughput to the adsorbent filling volume. The objective of preliminary experiments conducted in experimental adsorbers with continuous flow is always the adaptation of adsorber particle size to the size

of the mass-transfer zone and its migration rate in order to optimize adsorption efficiency and adsorbent utilization.

In the case of a relatively short mass-transfer zone, a single-stage adsorber can be used, since the ratio of unused adsorbent to the total volume can be small. If the wastewater of interest contains components that produce a long mass-transfer zone it is advisable instead to connect two or more adsorbers in series. The purpose of this arrangement is to ensure virtual saturation of adsorbent in the first reactor before breakthrough is observed at the end of the sequence (Fig. 142 A). After the first adsorber becomes exhausted (Fig. 142 B and D) it is refilled with fresh or regenerated material, or regenerated in situ, at which point it is moved to the last position in the sequence (Fig. 142 C) so that it may assume responsibility for final purification.

In some cases a more sensible solution is to operate two adsorbers in parallel but with staggered operating times, at least so long as the resulting mixture of purified wastewaters meets the necessary criteria. This method of operation is particularly applicable in situations involving extended mass-transfer zones. When the adsorber subject to the largest period of use must be exchanged or regenerated, either the feed stream is throttled temporarily or the remaining adsorber is used alone for a short time assuming it still has sufficient residual capacity to meet the abruptly increased load.

5.1.4. Adsorbents

Potential adsorbents for adsorptive wastewater purification include activated carbon, lignite coke, activated aluminum oxide, adsorber resins [440], and the isopolyacids of iron, aluminum, or silicon. Useful forms of these materials, their fields of application, and possible elimination mechanisms are summarized in Table 14.

Characteristic properties of the adsorbents, such as specific surface areas, pore volumes, and bulk densities, are collected in Table 15. Isopolyacids cannot, however, be characterized according to these parameters, since their properties are influenced by the conditions under which they form in the wastewater.

5.1.5. Adsorption Processes

Adsorption processes can be conducted either batchwise or continuously. In principle they can also be subdivided into *mixing* and *filter* (*percolation*) processes.

The Mixing Process. Mixing is usually used for batchwise wastewater treatment. It is essentially single-stage in nature, since adsorption and adsorbent separation together constitute one processing step. If technical or economic considerations dictate the use of a multistage mixing process, several discrete processing steps must be arranged in series. This usually entails countercurrent flow, but it can also assume a cocurrent form.

The Percolation Process. The percolation process, which normally involves a fixed-bed adsorber, is applicable to the continuous treatment of wastewater. When conducted in a countercurrent mode it is theoretically a multistage process, with effectiveness being a function of filtering speed and residence time.

The Pulse-Bed Process. In the pulse-bed process, which is a special case of continuous percolation, wastewater and adsorbent are both led in countercurrent fashion through the filter. The wastewater awaiting purification flows steadily from below into the adsorber packing. Adsorptively purified wastewater is withdrawn from the top. Adsorbent charging, in the form of either new material or regenerate, occurs at the top of the filter directly onto the filter bed. Contaminated adsorbent is withdrawn through a specially designed lock at the base of the filter, and then treated appropriately.

Prerequisites and Selection Criteria. A prerequisite to the use of adsorption for wastewater purification is complete wetting of the adsorbent surface by water. This is frequently accomplished in special ways prior to adsorption itself.

To avoid clogging of the filter and blockage of the adsorbent active surface, the wastewater subject to treatment must be freed as completely as possible from solids, if necessary through prior filtration.

Table 14. Adsorbents useful in wastewater purification

Adsorbent	Form of use	Fields of Application		Elimination mechanisms
		Organic	Inorganic	
Activated carbon	powdered, granular, or formed	in molecular solution; hydrophobic	in ionic solution, for heavy metals	adsorption (ion exchange)
Lignite coke	granular, powder	in molecular solution; hydrophobic	in ionic solution, for heavy metals	adsorption (ion exchange)
γ-Al_2O_3	granular, powder	in molecular or ionic solution; hydrophobic, hydrophilic	in ionic solution	adsorption, surface precipitation
Adsorber resins	granules	in molecular or ionic solution; hydrophobic, hydrophilic		adsorption, solid solution – extraction
Isopolyacids (formed within the water)	Fe^{3+}, Al^{3+}, Si^{4-} solutions	in molecular or ionic solution; hydrophobic, hydrophilic, dispersed	in ionic solution dispersed	adsorption, precipitation ion exchange, destabilization

Table 15. Properties of important adsorbents

Adsorbent	Form of use	Specific surface area, m^2/g	Pore volume, cm^3/g	Bulk density, g/L
Activated carbon	granular, powder	500 – 1000	0.3 – 0.8	300 – 550
		600 – 1500	0.3 – 1.0	
Lignite coke	granular, powder	200 – 250	< 0.1	ca. 500
γ-Al_2O_3	granular, powder	300 – 350	0.4 – 0.5	700 – 800
Adsorber resins	granules	400 – 1500	porosity 35 – 65 vol %	650 – 700

Table 16. Summary of the features associated with various adsorption processes

Adsorbent	Mixing process	Percolation process	Special techniques
Activated carbon	powder	granular, formed	PACT $*$, countercurrent, falling bed
γ-Al_2O_3	powder	granular	countercurrent, falling bed
Adsorber resins	granules (fine)	–	–
	–	granules	countercurrent
Isopolyacids	Fe^{3+}, Al^{3+} solutions		formation of the adsorbent within the wastewater by pH adjustment;
	Si^{4-} solutions		preliminary precipitation of activated silicic acid
Attendant problems	solid – liquid separation, processing, regeneration, reactivation, waste disposal	regeneration, reactivation, waste disposal	sedimentation, flotation, filtration, extraction, thermal treatment

$*$ PACT = Powdered activated- carbon treatment.

Table 16 offers a survey of the most important adsorbents and their uses in special process applications.

5.1.6. Adsorbent Recovery and Regeneration

Once adsorptive wastewater purification has occurred, the contaminated adsorbents must then be removed from the purified wastewater by one of the standard solid – liquid separation methods (→ Solid – Liquid Separation, Introduction, → Filtration, → Hydrocyclones → Centrifuges, Filtering, → Centrifuges, Sedimenting, → Sedimentation). This is followed by appropriate reprocessing, the purpose of which is either safe disposal of the contaminated adsorbents or (for economic reasons) their preparation for reuse. Some aspects of adsorbent recovery and regeneration have already been noted in Table 16 under the heading "Attendant problems."

In the case of the mixing technique, added powdered adsorbents or isopolyacids formed in the aqueous solution are subsequently separated

by sedimentation, flotation, and/or filtration. In the percolation process, adsorbents are separated from the purified wastewater by filtration and subsequent draining. This dehydration may be carried out either directly in the filter or in some remote location depending on what is to occur next. Countercurrent pulse-bed adsorbers are particularly convenient if external drying is contemplated.

Direct landfill disposal of contaminated adsorbents is subject to numerous restrictions. Important considerations are water content, elutability of the adsorbed substances, and content of organic material subject to decay. For this reason a thermal treatment based on calcining or incineration is frequently carried out. The residues from such treatment can normally be disposed of without difficulty.

In many cases, contaminated adsorbents are reprocessed for purely economic reasons, permitting their subsequent reuse for adsorptive wastewater purification. In the course of such a regeneration, wastewater constituents that have been eliminated by adsorption can also be recovered selectively. This is most often accomplished by extractive regeneration with solvents, or by a change in pH. With thermal regeneration, adsorbed organic materials are simply combusted.

The various adsorbents are subject to specific regeneration processes that take into account the properties of the materials involved and the nature of the corresponding reaction mechanisms.

5.1.6.1. Regeneration of Carbon Adsorbents

In this category only the activated carbons are candidates for regeneration. In the case of lignite coke the associated costs would exceed the price of new material, so spent coke is normally burned after use.

Thermal Reactivation (→ Carbon, Chap. 5.5.). The carbon regeneration process most frequently encountered in practice is thermal reactivation. This is conducted under the same conditions as fresh carbon production: i.e., at 900 – 1000 °C in the presence of water vapor. Thermal reactivation can be divided into four consecutive steps:

1) Evaporation of adhering water and water resident in the pores; in this step, which occurs during the heating phase, readily volatile substances also can be desorbed.
2) With rising temperatures, desorption of higher-boiling substances occurs.
3) Many organic constituents of wastewater can only be desorbed with simultaneous decomposition. This cracking begins to take place above ca. 400 – 500 °C. The process releases mainly CO and CO_2, but some low molecular mass organic cracking products are also released, while decomposition coke is left behind in the pore system.
4) In the last step of the process, above ca. 850 °C, decomposition coke (lustrous coke) is gasified by the water-gas reaction, which reopens the micropores that are very important for adsorption. This largely selective gasification is made possible by the amorphous structure of lustrous coke, since the more orderly activated- carbon structure withstands attack by water vapor by reason of a higher degree of graphitization.

The thermal units employed for reactivation (such as shelved kilns, rotary kilns, or fluidized-bed furnaces) are identical to the furnaces used for the production of fresh activated carbon. For technical reasons, this particular gas – solid reaction is not used for reactivating powdered activated carbons.

Only with very large wastewater streams is it practical to maintain a reactivation plant directly in conjunction with an adsorption facility. Reactivation plants can only be operated economically at throughputs exceeding several tons per day; moreover, high-temperature units of this type should not be subjected to frequent start-up and shutdown. In most cases, therefore, reactivation is carried out elsewhere by the manufacturer, where unavoidable losses can be also compensated for by fresh carbon. A further reason for encouraging large, centrally operated reactivation plants is the complicated problem of off-gas cleaning, which must ensure that all desorbed organic materials are completely combusted, and that inorganic breakdown products such as HCl or SO_2 are eliminated in appropriate ways.

Chemical Regeneration. In practice, chemical regeneration is carried out only in a few special cases where adsorbed substances can be desorbed by a pH change. This is true, for example, with phenols or organic acids, which can be adsorbed in neutral and acidic medium and subsequently desorbed again with alkali (sodium hydroxide solution). Conversely, compounds like aniline can be adsorbed under alkaline conditions and then desorbed from the adsorbent with acids.

Biological Regeneration. Biological regeneration, frequently described in the literature, is in practice also conducted as a separate process step. Often, however, activated carbon colonized with microorganisms is used in such a way that adsorption of pollutants and biodegradation of the adsorbate occur under equilibrium conditions. In this case the activated carbon frequently assumes a buffering role during concentration peaks.

5.1.6.2. Regeneration of Aluminum Oxide

Aluminum oxide contaminated with organic substances is regenerated thermally. In this case the organic substances are oxidized at a temperature of ca. 750 °C in the presence of excess oxygen. Oxidation occurs either directly on the adsorbent surface or after evaporation into the gas phase.

Aluminum oxide contaminated with inorganic substances is regenerated chemically. For example, sorbed phosphate can be eluted with caustic soda solution.

5.1.6.3. Regeneration of Adsorber Resins

Adsorber resins contaminated only with organic substances are always regenerated chemically. The sorbed substances are eluted by such solvents as methanol or acetone, in combination as required with caustic soda solution.

5.1.7. Applications of the Adsorptive Purification of Wastewater

Potential applications involving the various possible adsorbents for wastewater purification are determined on the one hand by the stated purification objectives and on the other by technical and economic considerations. This point will become clearer with the examples that follow.

5.1.7.1. Purification with Activated Carbon [435–437], [441–445]

Activated carbon is useful in the purification of wastewater as a way of removing most of the nonpolar organic compounds that may be present. Adsorption can be called upon to accomplish this task alone, or it may be used in combination with other processes. Technical and economic considerations determine whether a mixing process with carbon powder or percolation with granular or formed carbon is preferable. Activated carbon is used only in exceptional cases for the elimination of heavy-metal ions [441].

Activated carbon is used preferentially for the adsorptive purification of individual wastewater streams from production plants. Adsorption in this context may be used exclusively, or it may be combined with other processes, as in the two-step approach of stripping followed by adsorption. Decentralized water-protection measures (integrated environmental protection) of this type may be warranted for such purposes as water reuse, product recovery, or wastewater purification. However, reuse of wastewater after an adsorptive purification is frequently limited by the accumulation of non-adsorbable organic or inorganic contaminants.

Product recovery by adsorption on activated carbon may prove feasible if the wastewater stream is contaminated mainly with a single organic compound, and if this compound is readily adsorbable and desorbable. Examples worth noting include aniline, which is readily adsorbed from alkaline solution onto activated carbon and then desorbed by acid extraction in the form of an ammonium salt, as well as phenol, which can be correspondingly adsorbed from an acidic solution and desorbed by alkaline extraction as a phenolate salt.

Wastewater purification with activated carbon is especially appropriate whenever it is possible to discharge the treated wastewater directly into a receiving stream, or when activated-carbon treatment is used to remove substances

that would otherwise interfere with subsequent biological processing, perhaps in a centralized wastewater-treatment facility.

One special application is the use of activated carbon in biological treatment systems [442]. Thus, activated carbon is employed in biofilters (Section 3.2) [444] both as a growth surface for microorganisms and as an adsorbent for wastewater constituents. One prerequisite in this case is that the organic constituents be biodegradable. With batch filters the activated carbon is regenerated biologically after adsorption has occurred, whereas with a continuous filter the filter output mut be monitored regularly to assess the rate of biodegradation. Growth surfaces other than adsorbents can also be utilized in biofilters [443]. Powdered activated carbon is often introduced into activated-sludge operations, producing a combination known as the powdered activated-carbon treatment (or PACT) process. It is especially suitable for situations in which small amounts of readily adsorbable substances tend to inhibit biological degradation, substances that are not eliminated by other pretreatment measures in affect. The PACT process has also been considered as a way of adapting existing activation plants to meet overload conditions. In this context the powdered activated carbon serves as a growth surface for the activated sludge, causing the latter to proliferate so that the bioreactor can be operated with a higher biomass concentration. Other powdered substances can be called upon to play a similar role as well, including lignite coke dust (see the description of the PACT process under "Special Process-Engineering Solutions" in Section 2.1.3).

Treatment with activated carbon downstream from a central biological clarification plant is practiced only in exceptional cases, such as for reducing the toxicity of a wastewater effluent toward fish. This particular variant was once considered quite attractive, but it has been found to be subject to serious operational problems. Biological growth on the activated carbon can lead to anoxic conditions, with the resulting mercaptans and hydrogen sulfide constituting a considerable source of odor emission. Serious corrosion problems are also encountered in the filters themselves for related reasons. Downstream adsorption has therefore been abandoned in most cases in favor of appropriate pretreatment.

Landfill leachates are today often subjected to a process combination consisting of biological treatment, ultrafiltration, and activated- carbon adsorption [445]. In this case biological problems related to the activated carbon are largely avoided as a result of the ultrafiltration step.

Powdered carbon can be reactivated by wet oxidation. Granular and formed carbons are thermally reactivated in rotary or shelved kilns as well as in fluidized beds. This reactivation can be undertaken by either the user or the producer, with economic considerations normally dictating the choice.

5.1.7.2. Purification with Lignite Coke

In many cases, lignite coke is utilized instead of activated carbon for wastewater purification. Technical considerations play a role, but economic factors are most important. Lignite coke is certainly less expensive than activated carbon, but this is offset by lower adsorption efficiency. If relatively large amounts of lignite coke are to be employed it is necessary to take into account process-engineering considerations, logistics, and disposal technology.

5.1.7.3. Purification with Aluminum Oxide [435], [437], [446–450]

Treatment with activated aluminum oxide, especially γ-Al_2O_3, results in hydrophilic substances [446], [447] being sorbed at the solid surface, where they are retained by surface precipitation. This effect is exploited for the treatment of wastewaters from pulp production, above all from chlorine bleaching [446], [448].

Contaminated aluminum oxide was at one time regenerated thermally, but serious operational problems were encountered both in the adsorption filters and in the rotary kilns. Adsorptive wastewater purification of this type was largely abandoned after it became possible to improve the quality of wastewater by altering the corresponding production processes.

There has been some discussion of using α-Al_2O_3 for the elimination of fluoride, but especially of phosphate [449], [450]. Surface precipitation on the α-Al_2O_3 is then removed by extraction with sodium hydroxide solution, in the course of which residues can be recovered.

5.1.7.4. Treatment with Adsorber Resins [437], [451–462]

Adsorber resins are used in wastewater treatment for the targeted removal of both hydrophobic and hydrophilic organic compounds. This is possible since adsorption is in this case overshadowed by the formation of solid solutions in the adsorber matrix. In process-engineering terms the procedure can be described as adsorption and extraction with a solid solvent. It must always be borne in mind when using such adsorbers that they tend to swell in the course of taking up organic solutes.

Adsorber resins are normally regenerated by extraction with methanol, acetone, and other solvents. Appropriate processing of the extracts permits the solvents to be recovered and the eliminated organic materials to be isolated. Adsorber resins are therefore the adsorbents of choice for final purification of wastewater streams containing low concentrations of chlorohydrocarbons [455–457] and wastewater purification designed to facilitate recovery of the extractant and isolation of specific organic compounds [451–455].

5.1.7.5. Purification with Isopolyacids

The addition of iron(III) salt solutions, aluminum salt solutions, alkaline sodium aluminate solutions, or water glass to wastewater leads, depending on the pH, not to pure metal hydroxides, such as $Fe(OH)_3$ and $Al(OH)_3$, but rather to macromolecular hydroxides linked together via oxygen bridges. This can be illustrated as follows for formation of the first oxygen bridge:

$$2\,Fe^{3+} + 5\,OH^- \longrightarrow (OH)_2{-}Fe{-}O{-}Fe{-}(OH)_2 + H^+$$

If isopolyacids are used in a targeted way for wastewater purification, several elimination mechanisms can be in effect simultaneously:

1) Colloids may be destabilized and flocculated
2) Hydrophobic organic compounds may be bound adsorptively
3) Hydrophilic organic compounds may be precipitated or sorbed by ion exchange
4) Metal ions may be incorporated into the isopolyacid by coprecipitation

For this reason, metal salt solutions can be usefully incorporated prior to neutralization, concurrent with biological purification (for phosphate precipitation) and/or subsequent to biological purification.

The resulting contaminated isopolyacids must ultimately be separated from the wastewater and appropriately treated. Since they usually arise in conjunction with preclarification and excess sludge in central clarification plants, they are typically attended to as a part of clarification-sludge processing.

5.2. Mechanical Treatment

The goal of mechanical processing is to separate undissolved components from wastewater. The range of targets extends from debris and sand in the case of municipal sewage to coagulated sludge and oil emulsions in industrial wastewaters.

All the mechanical separation processes in current use can be related to one of two principles: *sieving* or *gravity*. Screens, sieves, and filters operate on the basis of sieving; sedimentation plants, floatation plants, flotation plants, foaming-out plants, and centrifugal plants are driven by gravity. Crucial to the choice of a particular process and to its effectiveness, aside from a difference in density between the components of interest and the wastewater matrix, is particle size.

5.2.1. Screening Processes [459–462]

A *bar screen* consists of parallel rods (usually made of steel) that are inserted in the wastewater channel in such a way as to intercept coarse and bulky materials (e.g., pieces of wood, fabric, boxes, etc.). The screens must then be cleaned, either manually or mechanically. Bar screens are used less for the removal of pollutants than as a way of protecting various downstream installations (e.g., pumping stations) from damage and blockage. The collected screenings are disposed of by consignment to a landfill or incineration. Figures 143 and 144 are schematic illustrations of such screens installed in wastewater channels, including provisions for both manual

and automatic cleaning. In the case of mechanical cleaning a distinction is made between *bar screens, curved bar screens*, and screens cleaned with *circulating rakes*. The gap between bars is generally 1 – 10 cm wide. Mechanical cleaning is initiated by a pressure differential and/or the elapse of a certain amount of time. Because of the risk of the screens icing in winter, and also because of the possibility that odors might develop, these facilities must be enclosed within a housing of some type. Screens can also be combined with *comminutors*, which crush the intercepted screenings in such a way that they can pass through the bars. This provision has the disadvantage that it increases the wastewater pollutant load, however, and comminutors do not always work satisfactorily with large amounts of textile residue or other long-fibered materials.

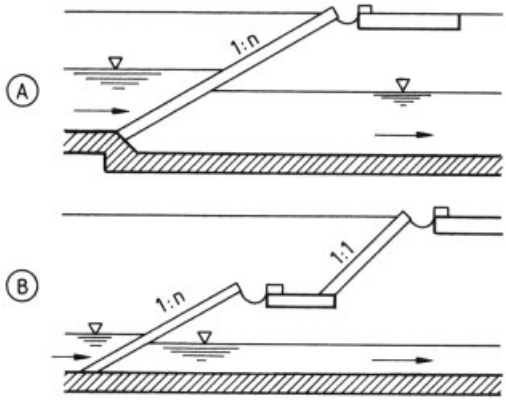

Figure 143. Channel installation of screens designed for manual cleaning [460]
A) Bar screen for a small channel (gradient 1 : 2 to 1 : 3; maximum 45 – 60°); B) Bar screens for a large channel, with provision for both dry-weather flow and stormwater run-off

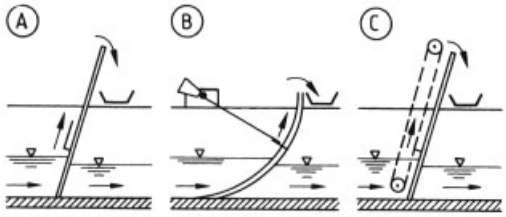

Figure 144. Channel installation of screens designed for mechanical cleaning [460]
A) Rake screen, inclination 60 – 80°; B) Curved-bar screen; C) Screen cleaned with circulating rakes, inclination 60 – 80°

Sieving plants (Fig. 145) are applicable above all in the industrial field for the separation and potential recovery of fine solid particles. Sieve apertures are normally 0.1 – 5.0 mm. The construction material may be high-grade steel or plastic.

Belt sieves, drum sieves, and *disk sieves* are circulating or rotating devices. The corresponding sieving surface consists of fabric or of slotted or perforated metal sheets. After emerging from the water the separated solids are removed by rinsing or shaking. Continuous cleaning is necessary in order to minimize losses of efficiency due to blockage.

Belt sieves (Fig. 145 A) are especially well-suited to very large and highly variable flow rates. With most drum sieves (Fig. 145 B) the water runs from the interior outward through the rotating sieve drum. Material collected on the sieve is sprayed loose from the outside toward the center, where it falls into a collecting channel.

A curved sieve (Fig. 145 D) consists essentially of a feeding device and a concave surface, and it is self- cleaning in operation. Wedge-shaped profiled rods are arranged perpendicularly to the flow direction of the water. A relatively steady overflow ensures that the screen will clean itself. The various screen segments are all interchangeable. Typical gap widths are 0.02 – 2 mm for sieve areas of 0.1 – 3.0 m^2 [maximum throughput 300 $m^3/(m^2 \cdot h)$]. Curved screens are frequently used in preserve factories and slaughterhouses as well as in the paper and pulp industries.

Microstraining (Fig. 146) is used mainly for more complete elimination of suspended material in wastewater-treatment effluents. A rotating drum fitted internally with fine gauze (e.g., a high-quality alloy-steel gauze) serves in this case as the sieve. The mesh width is usually 5 – 100 µm, with peripheral speeds of ca. 50 m/min. The process stream runs from the inside of the rotating drum outward. About one-third of the drum projects above the water. In this region the gauze screen is continuously cleaned from the outside by water sprayed through nozzles (at ca. 3 bar). The water required for this purpose is equivalent to 2 – 5 % of the volume of the treated effluent. The hydraulic surface-loading of a microstrainer is 10 – 20 $m^3/(m^2 \cdot h)$. Micros-

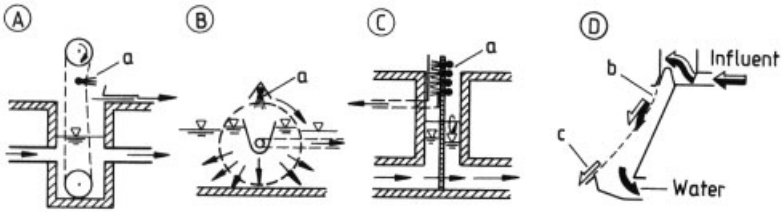

Figure 145. Sieves for wastewater treatment [460]
A) Belt sieve (water flows through the circulating sieve); B) Drum sieve (water flows through the rotating drum); C) Disk sieve (water flows through the rotating disk); D) Curved sieve (water flows over the inclined screen surface)a) Rinse water; b) Sieving surface; c) Solids

training represents a borderline case that could also be classed with the filtration processes.

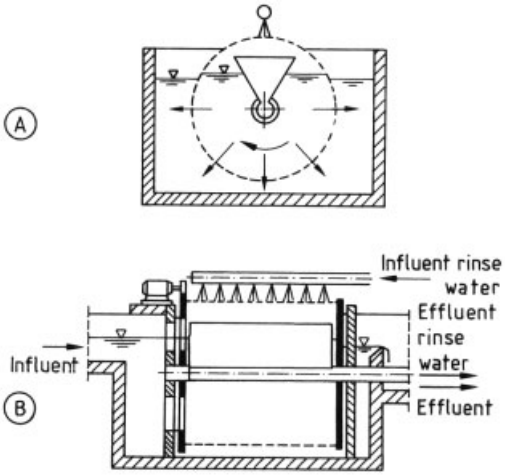

Figure 146. Schematic section through a microstrainer unit [462]
A) Cross-section; B) Longitudinal section

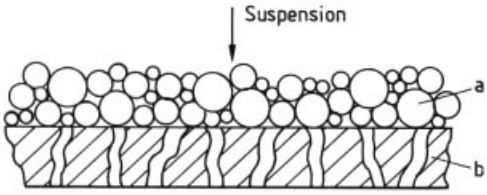

Figure 147. Mechanism of cake filtration [462]
a) Cake; b) Medium

In *filtration*, a fundamental distinction is made with respect to the filtration mechanism between *cake filtration* (Fig. 147) and *deep-bed filtration* (Fig. 148). With cake filtration, the uppermost layer of the filter cake operates as a sieve, so that even particles with diameters smaller than the pore size of the medium are also separable. In deep-bed filtration (e.g., a sand filter), separation takes place in the channels and pores of the filter bed itself.

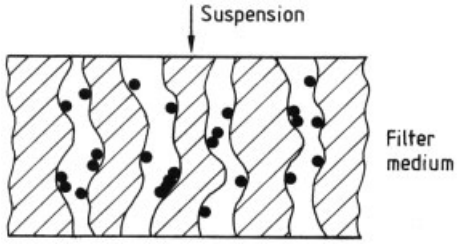

Figure 148. Mechanism of deep-bed filtration [463]

In designing a filtration plant it is necessary to take into account the fact that during operation the pressure drop increases to a degree related to the extent of the solids retained. Once a filter becomes loaded the collected material must be released by backwashing. The resulting backwash water then usually becomes subject to treatment as well.

Very high levels of purity can be achieved by filtration, for which reason such processes are commonly used as a preliminary treatment for the removal of interfering suspended matter prior to other purification measures (adsorption, ion-exchange, etc.), or as a final purification step analogous to microstraining (e.g., for achieving established limits for suspended matter in the discharge from a biological treatment plant).

Deep-bed filtration entails less frequent backwashing than cake filtration, and thus a lower requirement for backwash water. For this reason, cake filtration is applicable only in exceptional cases in wastewater treatment.

Sand with a grain size between 0.5 and 3 mm is usually used as the filter material in the *sand*

filter version of deep-bed filtration (Fig. 149). The process stream usually passes downward through the filter at a surface-loading rate of $5-10\,m^3/(m^2 \cdot h)$. High solids concentrations lead to short filtering times and a heavy requirement for backwash water, so the solids concentration in the feed to a sand filter should not be allowed to exceed ca. 40 mg/L.

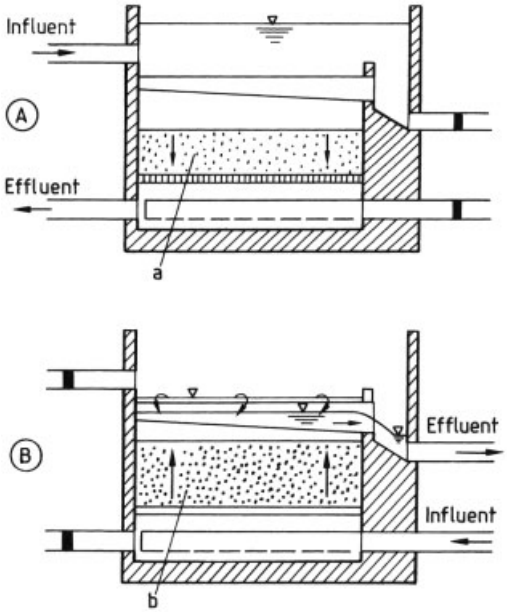

Figure 149. Design and operation of an open sand filter
A) Filtration; B) Backwash
a) Filter bed; b) Expanded filter bed

1 – 2 % of the filter effluent is used for the backwashing of a sand filter. Backwashing is conducted as soon as some predetermined pressure difference is exceeded, or at preselected time intervals. Backwashing can be expected to take 20 – 50 min, and is effectively accomplished by alternate introduction of air and water.

In the case of continuously working sand filters a distinction is made between *moving-bed filters* (Fig. 150) and *gravity filters* (Fig. 151). In a moving-bed filter the raw water runs from below, countercurrent to a sand bed that is sliding downward. In a gravity filter the backwashing process is initiated automatically by differential pressure via a water siphon. The amount of water required for backwashing a continuously working sand filter is about 5 % of the filtrate.

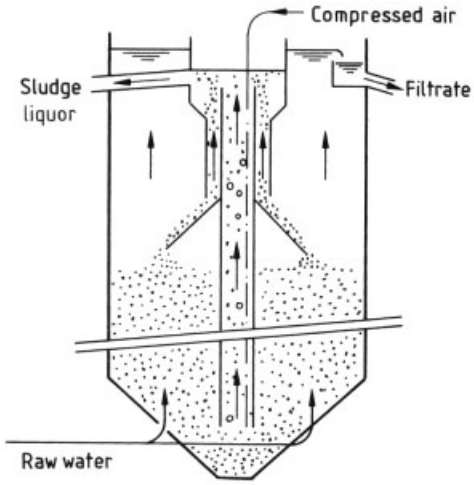

Figure 150. A continuous moving-bed filter [462]

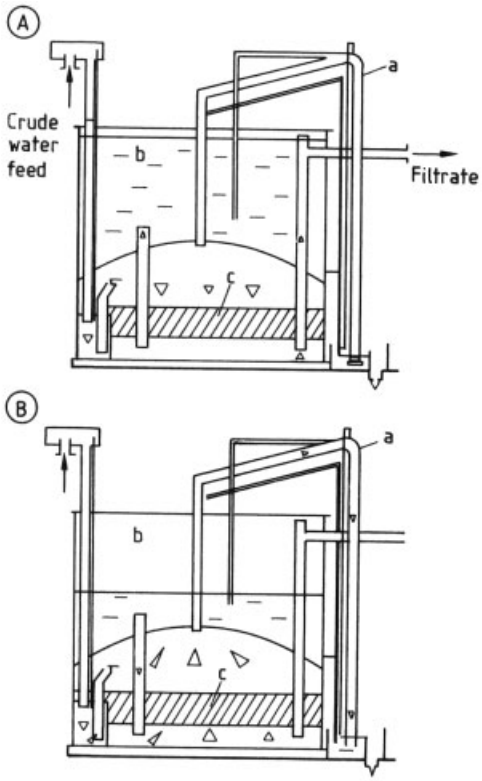

Figure 151. A continuous gravity filter [462]
A) Filtration; B) Backwash
a) Backwash system; b) Backwash-water storage vessel; c) Filter bed

Filter presses operate with an overpressure of up to 15 bar, generated by piston diaphragm pumps that compress the suspension onto a system of filter plates (Fig. 152). These plates are covered with a stretched filter cloth and arranged in parallel array at a distance corresponding to the cake thickness. After compression is complete, the plates are pushed apart one after another, either mechanically or manually, permitting the cakes to fall out.

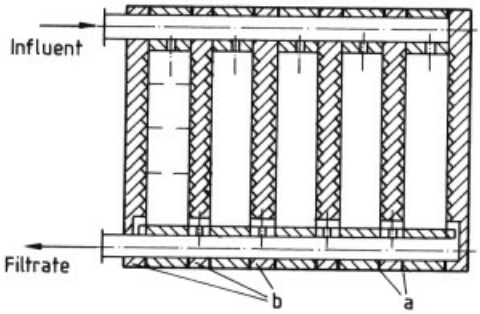

Figure 152. Schematic illustration of a filter press [463]
a) Filter cloth; b) Filter plates

Vacuum filters dehydrate a suspension with the aid of reduced pressure. Rotary-drum filters are frequently used for this purpose. The suspension is fed continuously into a trough in which the drum, covered with stretched permeable cloth, is about half submerged. Solids are sucked onto the cloth by reduced pressure, simultaneously removing water from the suspension. As the drum slowly rotates, the filter cake is doctored off the filter cloth by a removal device.

5.2.2. Gravity Processes [459], [460], [463–465]

The most important of the mechanical wastewater-treatment processes are gravity processes, especially sedimentation.

5.2.2.1. Sedimentation (→ Sedimentation)

Grit chambers (sand traps) are indispensible if a waste-treatment plant is also obliged to deal with rainwater, which normally entrains a considerable amount of sand. This might otherwise be deposited in many inconvenient places, and it could also lead to rapid abrasion in the pumps. Grit chambers in a wastewater-treatment plant are usefully situated immediately downstream from the screen that protects them from coarse and fibrous materials. Grit chambers are designed to handle a horizontal flow rate of ca. 30 cm/s. As a result, only sand is separated, whereas the lighter sludges are carried on further by the flowing liquid.

Three different types of grit chambers can be distinguished. The *channel-shaped horizontal-flow* grit chamber (Fig. 153) has a cross section of such a shape that a flow rate of 30 cm/s can be maintained in combination with a venturi flume located immediately downstream, even with variable wastewater streams. Alternatively, a similarly constant rate of flow can be achieved in a highly fluctuating wastewater stream by incorporating into the effluent region a baffle plate adapted to conform to the cross-section. With small fluctuations, a grit chamber with a parabolic or trapezoidal cross-section will suffice. Small grit chambers are cleaned manually, larger ones mechanically with the aid of scraper systems or air-jet lifts.

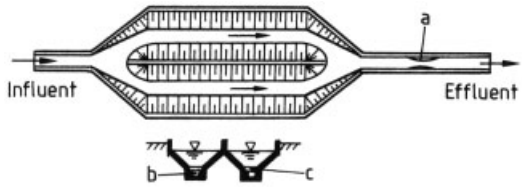

Figure 153. Channel-shaped horizontal-flow grit chamber with a venturi flume
a) Venturi flume; b) Drainage; c) Sand storage

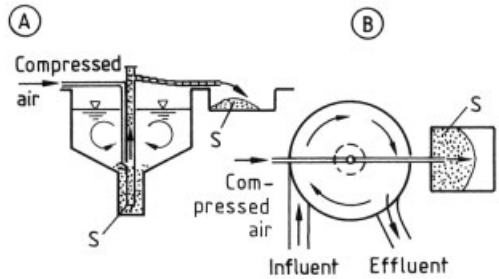

Figure 154. Circular grit chamber
A) Cross-section; B) Top view
S = Sand

In the case of a *circular* grit chamber (Fig. 154), the water is introduced tangentially, causing the contents to circulate. As a result,

grains of sand are washed to the center where they can be removed with an air-jet lift. However, this type of grit chamber is less suitable for highly fluctuating flow rates.

In the *aerated* grit chamber (Fig. 155), circulation of the contents is achieved by an injection of air such that a flow rate of 30 cm/s results at the bottom. The bottom of the tank should be rounded in order to ensure undisturbed rotation. Aerated grit chambers can be combined effectively with a means for the separation of floating matter, and they also operate flawlessly with fluctuating flow rates.

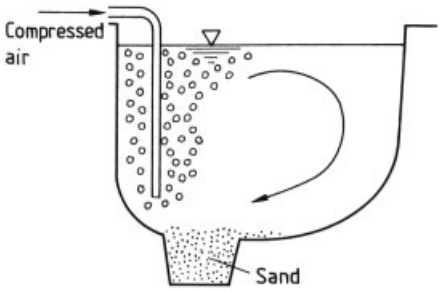

Figure 155. Aerated grit chamber

Settling tanks (Fig. 42) are not usually used alone, but are rather integrated into a sequence of treatment steps, for example, as primary and secondary clarifiers in mechanobiological treatment facilities or as components of precipitation-flocculation-sedimentation plants.

Again, three main types of construction can be distinguished:

1) Rectangular tanks with horizontal flow (Fig. 42 A)
2) Circular tanks with horizontal flow (Fig. 42 B)
3) Hopper-bottom ("Dortmund") sedimentation tanks with vertical flow (Fig. 42 C)

In the circular and rectangular flat tanks the settled sludge is removed with such continuously working devices as suction dredges or sludge scrapers. Hopper-bottom sedimentation tanks usually have no provision for mechanical sludge removal. Oil, grease, and other floating materials are retained by inverted baffles set ahead of the tank outlet. An important consideration in the sizing of a settling tank is the settling rate of the suspended solids. Figure 156 shows settling tracks for (A) discrete and (B) flocculated particles in an ideal settling tank. The discrete particles settle at a constant rate, whereas flocculated particles settle at an ever-increasing rate, because the floc continuously grows as a result of flocculation and agglomeration phenomena. At high flocculated-sludge concentrations, such as those found in secondary clarifier tanks associated with activated-sludge plants, settling is impeded by mutual interaction of the sludge flocs. Properly designing a settling tank requires that experiments be conducted to establish the anticipated settling rate. Once the settling rate is known, the maximum permissible hydraulic surface loading of the proposed tank can be ascertained. The hydraulic surface loading, expressed in m/h, is the quotient obtained by dividing the flow rate of effluent water by the active surface area (clarification surface) of the sedimenter [$(m^3/h)/m^2 = m/h$]. In the case of nonflocculating particles the effectiveness of the treatment depends only on the surface area of the tank, not on the tank depth. With flocculating ingredients the depth of the settling tank becomes important as well, since the further the suspended solids settle the larger they become, and thus the faster they settle. Aside from the surface area, and therefore the hydraulic surface loading, other crucial parameters are the tank volume and the residence time. Table 17 gives an indication of maximum permissible hydraulic surface loadings. For use as a secondary clarifier tank in conjunction with an activated-sludge facility another parameter taken into account for sizing the tank surface area is the solids-surface loading. This is obtained by dividing the rate of biomass settlement in the secondary clarifier tank (in kg/h) by the tank surface area in m^2, and is thus expressed in units of $kg\,m^{-2}\,h^{-1}$. Satisfactory values generally fall below 3 kg of total suspended solids per square meter per hour.

Table 17. Hydraulic surface loading, in m/h; based on [464]

	Primary clarifier	Secondary clarifier	Intermediate clarifier
Mechanical treatment only	1.5 – 0.8		
With chemical precipitation	4 – 2.5	1.5 – 1.0	>2.5
With trickling-filter plant	1.5 – 0.8	1.5 – 1.0	>2
With an activated-sludge plant	4 – 2.5	1.2 – 0.5	>2.5

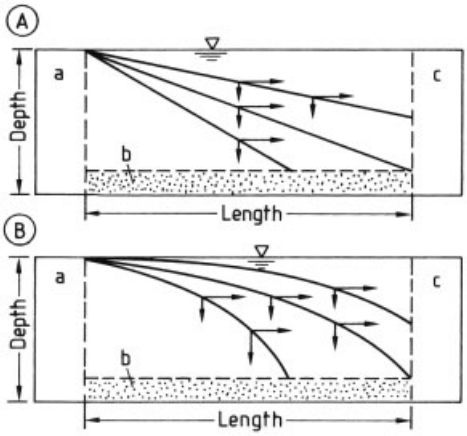

Figure 156. Settling tracks in an ideal settling tank for (A) a discrete particle and (B) a flocculated substance
a) Inlet zone; b) Sludge zone; c) Outlet zone

The principle underlying the *parallel-plate* (or *lamella*) *separator* (Fig. 157) is the fact that separation efficiency for a nonflocculating sludge is independent of the tank depth. The active clarification surface can be increased by installing inclined parallel plates in the separating chamber, thereby permitting the device as a whole to be smaller and more compact. The plates are set at an angle of inclination between 45° and 60°, with a plate spacing of a few centimeters. Owing to the small plate spacing, coarse contaminants can lead to blockage. Separation of biological sludge with the aid of plate separators is associated with certain risks, since the plates can also serve as growth surfaces for microorganisms. If such growth becomes too severe it may roughen the plates to such an extent that sliding of the sludge is impeded.

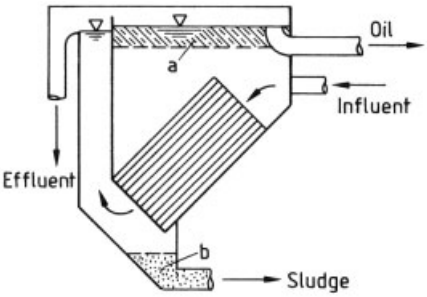

Figure 157. Parallel-plate (lamella) separator
a) Oil; b) Sludge

In practice, the separating efficiency of conventional settling tanks is often greater than that of parallel-plate separators, but the latter offer the advantage of reduced investment costs and above all a lower space requirement.

Parallel-plate separators are used both for the removal of settleable solids and for the removal of light, floating materials, such as oil in water, and especially for combining the two processes (see Fig. 157).

5.2.2.2. Floatation

Substances that are lighter than water, usually in suspended form, can be removed at the surface by floatation: in separating pits (Fig. 158 A), grease separators (consistent with DIN 4043; see Fig. 158 B), oil separators (DIN 4041), or gasoline separators (DIN 1999). The standards cited rely on the feed rate as a sizing criterion for the various separators. The separation effect is expected to be 95 % in each case. For a gasoline separator the sizing recommendations are based on a permissible surface loading of 18 $m^3/(m^2 \cdot h)$. Most separators described are available in ready-made form. Tipping troughs and overflow channels are usually used for removing collected residue.

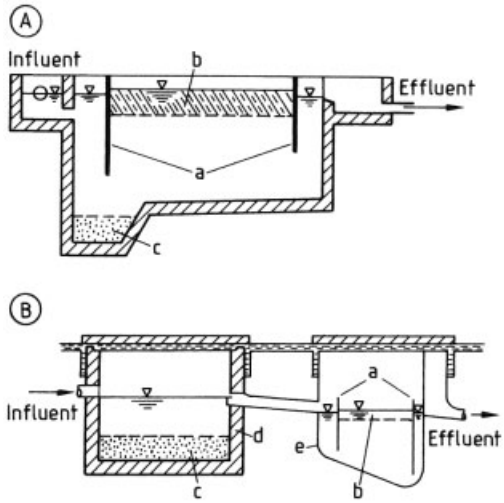

Figure 158. Examples of flotation separators
A) Separating pit; B) Grease separator equipped with a sludge trap
a) Inverted buffles; b) Light materials; c) Sediment (sludge); d) Sludge trap; e) Grease separator

The parallel-plate type of separator has already been described in the preceding section (Fig. 157). On the underside of the inclined plates, separated oil droplets rise to the water surface, whereas material collecting on the upper sides slides off to the tank bottom.

5.2.2.3. Flotation (see also Section 3.1; → Flotation)

In the process known as flotation (not to be confused with flo*a*tation; "flotation" is an artificial type of floatation) suspended particles are *caused* to float by becoming attached to very fine gas bubbles. In this way suspended solids that would otherwise be difficult to separate because of a small density difference relative to the wastewater matrix are raised to the surface. Flotation is applicable to both small and large amounts of wastewater, especially for the recovery of oils and greases, but also for the pretreatment of wastewaters from slaughterhouses and paper and cellulose production. Flotation facilities are also sometimes used for sludge thickening, and for easing the burden on (or even replacing) secondary clarifier tanks in biological wastewater-treatment plants.

Depending on the technique used for bubble generation, a distinction is made between induced air flotation, electro-air flotation, and pressurized dissolved-air flotation (see Section 3.1.3). Wastewater technology takes advantage of pressurized dissolved-air flotation almost exclusively. A distinction is made in this context between full-flow treatment (Fig. 159), in which all of the wastewater influent is saturated with air at a certain pressure, and the circulation or recycling process (see Fig. 108), in which a portion of the treated water is isolated, saturated with air, and returned to the flotation vessel. The flotation effect can be enhanced by a preliminary flocculation.

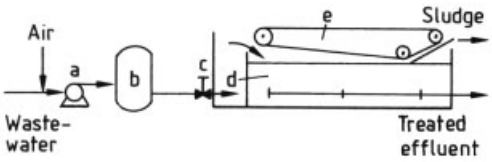

Figure 159. Flotation system for full-flow treatment
a) Pressurizing pump; b) Pressure chamber; c) Reducing valve; d) Flotation tank; e) Sludge remover (skimmer)

Flotation tanks can be either rectangular or circular. They must be designed to permit a flow pattern in the flotation vessel that is as uniform as possible, with a minimum amount of turbulence. Apart from the hydraulic surface loading, the solid-surface loading and the air/solid ratio represent important design criteria. Optimum design values depend heavily on the nature of the suspended particles subject to treatment, and must be established with the aid of special test apparatus. The air/solid ratio must be between 10 and 30 g of air per kilogram of total suspended solids.

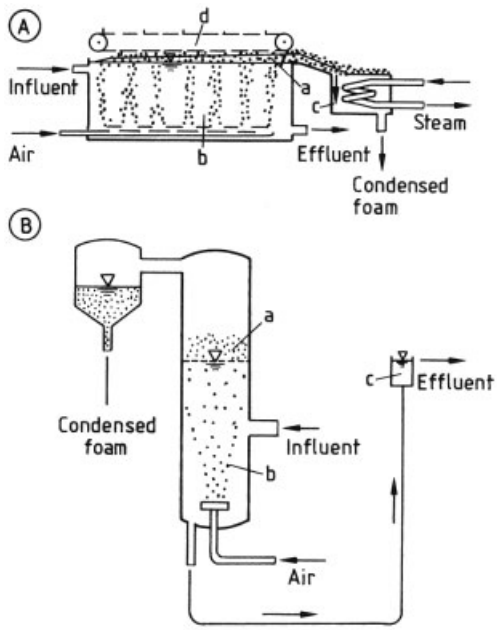

Figure 160. Foaming-out processes [462]
A) Aeration tank; B) Aeration tower
a) Foam; b) Water; c) Foam breaker; d) Foam remover (skimmer)

The flotate is usually removed from the surface mechanically, and some type of discharge provision is also usually required for solids that sediment in the flotation system.

Compared with sedimentation separators, flotation devices are characterized by low structural volumes but high operating costs attributable to energy expenditures.

5.2.2.4. Foaming-Out Process

Foaming-out (Fig. 160) is a way of removing various foam-forming substances (e.g., surfactants and saponins), which because of their surface-active character tend to accumulate at the water – air interface. Two modes of operation can be distinguished. With systems based on an *aeration tank* (Fig. 160 A), foam is swept out with a belt foam remover and destroyed in a foam breaker; if an *aeration tower* is employed instead (Fig. 160 B), foam reaches the foam breaker without assistance.

5.2.2.5. Centrifugal Processes

Centrifugation is a way of increasing dramatically the effect of gravity on suspended substances. As a result, even suspended solids that are incapable of settling under normal conditions can be induced to separate.

Decanting centrifuges (also known as decanters), *filtering centrifuges*, and *separators* all find application in wastewater technology. With a decanting centrifuge (Fig. 161) thickened solid is discharged continuously from the interior of the centrifuge by means of a screw. The filtering centrifuge relies on centrifugal force to enhance the filtering effect. Separators are able effectively to isolate both suspensions and other materials with densities differing from that of the matrix medium. Centrifuges are most often employed in sludge treatment, especially in the form of decanters for sludge dewatering.

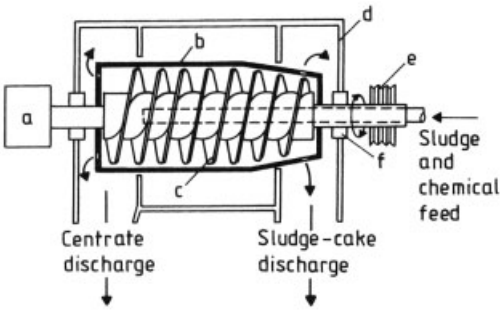

Figure 161. Decanting centrifuge (base omitted) [465]
a) Differential-speed gearbox; b) Rotating bowl; c) Rotating conveyor; d) Cover; e) Main drive shaft; f) Bearing

Hydrocyclones (→ Hydrocyclones) could also be classified as centrifugal separators, and these are called upon for such purposes as desanding, classification and thickening of industrial slurries and sludges, and the separation of granular suspensions. In the case of a hydrocyclone, the suspension subject to treatment is fed into the device tangentially, as a result of which the mass of liquid is induced to circulate and solids are flung onto the wall, where they migrate downward. Hydrocyclones are adapted to suit the size of the particles to be separated: small particles demand higher operating pressures and smaller cyclone diameters. Hydrocyclones for wastewater-treatment applications should have a diameter of at least 300 mm to minimize the risk of clogging. The separating efficiency of a hydrocyclone is strongly affected by fluctuations in the feed rate.

6. Wastewater Incineration

6.1. Introduction

Wastewaters for which biological, chemical, and physical purification processes either fail or prove too expensive can, if necessary, be purified by combustion of their organic constituents. In principle, all organic substances can be oxidized in a flame, and volatile inorganic substances formed can be separated in a flue-gas scrubber. Oxidation in a flame and the associated breakdown of organic materials depends on temperature, residence time, oxygen content, and turbulence in the combustion chambers.

The incineration of organic constituents is a specialized process for treating those industrial wastewaters in which the organic constituents, and under some circumstances the inorganic constituents as well, are chemically oxidizable at high temperature with the aid of atmospheric oxygen; the accompanying water matrix is simultaneously vaporized. The expression "gas-phase oxidation" [466] is also used in conjunction with this type of combustion to distinguish it from processes such as wet oxidation (see Section 4.4).

In the course of this treatment organic materials are converted almost completely into inorganic reaction products, mainly carbon dioxide and water, together, in some cases, with nitrogen oxides, sulfur oxides, or halogen compounds. During the incineration process such inorganic

constituents of the wastewater as salts lead (depending on the temperature and the formation of eutectic mixtures) to slags, ashes, or ash sludges, which can either be reused or (depending on the nature of the flue-gas treatment) be diverted for collection or disposed of as waste. Dust and gaseous substances foreign to air that may be present in the flue gases include heavy-metal compounds, sulfur dioxide, hydrogen halides, and phosphates, and these must be removed by appropriate means.

The process of gas-phase oxidation is particularly suitable [467], [468] if:

1) Organic constituents in the water cannot be reused, or their processing is unprofitable
2) The contaminants constitute a multicomponent mixture in which both the concentration of the mixture and the mixing ratio of the components vary continuously
3) Apart from organic materials the wastewater also contains considerable amounts of inorganic matter
4) The materials present are poorly biodegradable or biologically toxic
5) Contaminant levels, time demands, and economic considerations justify the erection of such a plant
6) The salt content is so high that the wastewater in question is treatable biologically only after considerable dilution
7) Incineration would make it possible to produce valuable products, or to recycle feed materials – salts, for example

Many other processes compete with gas-phase oxidation, including adsorption (Chap. 5), stripping, low- or high-pressure wet oxidation (Section 4.4), reverse osmosis, etc. The choice of a particular method usually depends on the environmental situation at the treatment site together with investment and operating costs [469].

Wastewater incineration is an exothermic process. Nevertheless, an incineration that is self-sustaining from an energy standpoint is achievable only if the organic load is sufficient to ensure adequate support for vaporization and superheating of the water present [469], [470]. If the organic burden is low, a wastewater incineration plant must therefore be operated with a supporting flame [471], [472]. The requirement for additional energy can be reduced by lowering the water content of the feed. This can be accomplished, for example, with a single- or multistage evaporation unit situated upstream. Installation of a heat-recovery section (boiler) may be a convenient way of generating steam from the resulting heat of combustion, steam that might in turn be used for evaporation purposes (Fig. 162) [466].

Wastewater incineration is highly demanding with respect to both technology and energy. In the development of a suitable furnace chamber, for example, account must be taken of the chemical effects of the wastewater subject to thermal treatment in terms of corrosion behavior, the technology envisioned, and possibly the refractory lining.

Thus, depending on the nature of the individual organic and inorganic constituents in the wastewater and relevant local factors, very different plant designs might result, and each plant must generally be tailored to a specific situation.

6.2. Legal Aspects [473]

Construction of a wastewater- combustion facility requires compliance with all applicable laws and statutes. In the Federal Republic of Germany, for example, such facilities must be in conformance with provincial wastewater law (Landesabwasserrecht). In the event that combustion of liquid or solid wastes other than wastewater is contemplated it is necessary to acquire permission under the German Federal Pollution Control Act (17th BImSchV, regulating combustion facilities for wastes and similar combustible materials).

The costs of wastewater combustion are subject to dramatic increases as a result of the need to meet specified emission limits and lowest acceptable temperatures, and in some cases because of administrative expenses. Similarly, account must be taken of the inevitable time lapse between the filing of an application and the granting of a permit.

6.3. Distinctions with Respect to Other Processes, and Evaluation Criteria

Organic constituents are destroyed during the incineration of wastewater, so it is always important to investigate the alternative possibility

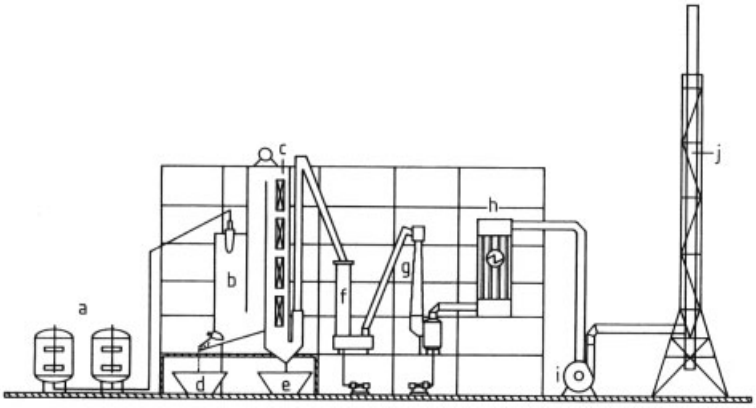

Figure 162. Schematic view of a typical wastewater-incineration (gas-phase oxidation) plant with flue-gas washing
a) Wastewater storage tanks; b) Combustion chamber; c) Process-gas cooler; d) Molten salt; e) Discharged salts; f) Quenching system; g) Jet washer; h) Electrical precipitator; i) Fan; j) Stack

of recovery by such processes as adsorption, ion exchange, distillation, evaporation, stripping, extraction, reverse osmosis, or crystallization, which in some cases may even offer significant advantages [474], [475].

Important criteria related to wastewater incineration and for selection of the type of facility to be constructed or the recovery process to be instituted for heat and/or valuable substances include:

1) The calorific value (heat equivalent) of the wastewater, which is an essential quantity for assessing the supplementary firing (fuel) requirement. A self-sustaining combustion can be anticipated with a calorific value of ca. 6000 – 12 000 kJ/kg.
2) The nature, amount, and variability of the wastewater and its organic and especially inorganic constituents. This strongly influences such factors as the preferred type of furnace, the construction materials to be used, the potential for heat utilization, and flue-gas treatment.
3) Legal aspects, including probable requirements under licensing law, the role of location, and possible consequences under tax law.

6.3.1. Calorific Value

The first consideration in terms of thermal disposal of wastewater should be the calorific value, because it is upon this that fuel consumption and therefore overall operating costs will depend. For example, if the calorific value of a wastewater is zero, then for 1 t of wastewater at a combustion- chamber temperature of 1000 °C and with a 40 % excess of air at 6 % residual O_2 and 5 % radiation loss, ca. 250 kg of light fuel oil would be consumed for evaporation and superheating. If the air is somehow preheated from 20 °C to 300 °C this estimate can be lowered to just under 200 kg of fuel oil. With heat recovery by steam generation (exit-gas temperature 300 °C), the fuel-oil requirement can be reduced substantially, to 91 – 100 kg per ton of water. For easily combusted constituents, complete burn-out can be achieved even at low temperatures and with little excess air or residual O_2. The advantages, apart from an energy saving, are less NO_x formation as well as a correspondingly smaller facility [469], [474], [476].

For wastewaters with low calorific values it is advisable to install multistream burners, introducing light fuel oil or waste solvent separately from the water via flame-vaporization burners.

Destroying the organic constituents quantitatively requires that optimum distribution be assured in the furnace chamber. This is achieved by the addition of an atomizing medium, such as steam or compressed air (2 – 15 bar). The viscosity of the wastewater must also be taken into account in the process, along with such parameters as particle size and particle-size distribution.

6.3.2. Wastewater Constituents

During the course of incineration, specific wastewater constituents can create various problems both in terms of the combustion process itself and in subsequent flue-gas treatment, thereby influencing the choice of the incineration facility and the associated treatment equipment [477]. Volatile organic materials in dissolved form and devoid of heteroatoms other than oxygen are unproblematic. Their combustion leads to virtually no ash, dust, or troublesome emissions. Vapor-phase feed is possible, and there is no problem with heat recovery. The same is true of dissolved, suspended, or emulsified organic matter (e.g., carbohydrates, fats or greases, lignosulfonic acids, condensates, and polymers). Emission problems arise only with halogen- or sulfur- containing materials. Vapor-phase introduction is in this case not possible, however.

In the incineration of wastewaters containing inorganic salts, and sometimes also with mineral acids or alkalis, one must be prepared to deal with aerosols and dusts that are difficult to remove. Special separators are required, and heat recovery in the form of steam is technically more difficult. In addition, damage can be expected with respect to ceramic materials (refractory lining) from alkalis and salts, especially phosphates and carbonates. Any melts, ashes, or dusts formed must be removed and discarded. Aside from emission problems and the presence of dust and ash, there is also a risk of incrustations and blockages in the combustion- chamber inlet and in the waste-heat boiler from salts or oxides of calcium, magnesium, and iron. Acids present the further challenge of metal corrosion.

6.4. Incineration Plants

The design of an incineration plant is influenced primarily by the nature of the wastewater subject to disposition. In this connection the most important factors are quantity, calorific value, salt and solid content, and the content of such heteroatoms as sulfur, nitrogen, the halogens, and phosphorus in the organic materials present, not to mention excessive amounts of acidic or alkaline constituents. The processes envisaged for treatment, removal or recovery, and possibly steam production also play a role, as do legal requirements that must be met.

In the case of strict wastewater incineration the choice will probably be between a simple or a cooled combustion chamber and a fluidized-bed furnace, whereas for combined incineration with solid wastes a facility such as a rotary kiln may be an option. Co- combustion in a boiler for steam generation is also a possibility in special cases [478], although rather severe restrictions may apply as a result of the salt content and its melting or softening point.

It is always important to ascertain whether a preliminary evaporation prior to incineration might be less expensive than direct incineration of the wastewater itself [472]. Even if the vapors still must be combusted it can be advantageous to concentrate the wastewater with the aid of waste heat obtained from the actual combustion process.

With wastewaters containing low concentrations of contaminants it may be possible to reduce the specific steam consumption from 1 kg of steam per kilogram of water to ca. 0.2 kg of steam per kilogram of water by a five-stage evaporation. In order that the flexibility of the incineration not be too seriously restricted by preliminary evaporation, an intermediate tank should be installed between the corresponding facilities.

6.4.1. Combustion Chambers

When incinerating a wastewater stream, perhaps in conjunction with waste solvents or waste air, a combustion chamber fixed in a horizontal, inclined, or vertical position is most appropriate depending on the salt and solids content [479], [480]. This is especially the case at relatively high incineration temperatures (>900 °C, "high-temperature combustion"). Compared with a facility designed for lower temperatures, such as a fluidized-bed furnace, a combustion chamber of this type offers greater flexibility with respect to fluctuations in throughput and wastewater composition.

A vertical mounting has been found to be particularly advantageous if salts must be removed in molten form at the bottom of the combustion chamber [466], [481], [482]. A horizontal arrangement with a downstream waste-heat system is preferable if salt removal is not required.

Inclined combustion chambers have also been used.

Wastewater and fuel are charged via burners or special nozzles, sometimes at several points within the combustion chamber. Air – both "primary" or combustion air added jointly with the fuel, and "secondary" or mixing air supplied, for example, via an annular conduit – is likewise introduced at several points, usually with a spinning impulse. The temperatures in the individual combustion zones can be regulated via the supply of fuel or air. Thus, if the combustion chamber is large enough it is possible to conduct a deliberate two-stage combustion (e.g., wastewater in a first stage, and waste gas in a second stage, each with different residence times and temperatures. The combustion chamber must be of such dimensions that the residence time of the flue gases is sufficient for complete combustion. Allowance must also be made for a sufficient excess of air).

A combustion-chamber temperature of at least 800 °C is generally necessary for the complete incineration of wastewater constituents. An increase in the combustion temperature not only raises the costs (energy consumption, maintenance) but also the proportion of nitrogen oxides in the flue gas, so the temperature should not be higher than absolutely necessary. However, it is important to ensure that all organic constituents are oxidized almost completely. In some cases it is possible to take advantage of special design features or air-supply techniques (e.g., "recycle burners", a stepwise mode of operation in the combustion chamber) as a way of suppressing the formation of nitrogen oxides [471], [483], [484].

Formation of such acidic flue-gas constituents as SO_2, HCl, and HF can be prevented by ensuring sufficient alkalinity in the wastewater, a precaution that can also greatly simplify and reduce the cost of subsequent flue-gas purification [466].

If substances other than "typical fuels" are also used as energy carriers in such plants, higher incineration temperatures (e.g., 1200 °C) and longer residence times may be stipulated (e.g., 17th BImSchV).

The casing of a combustion chamber is usually made of steel, and frequently equipped with a refractory lining. The lining is subject to attack if alkali-metal carbonates or other alkaline substances are formed during the combustion process. For this reason a hard skin of slag should, if possible, be maintained on unlined portions of the wall (by ensuring that there is a temperature gradient toward the outside), over which the melt can flow. A refractory lining is less sensitive to acidic constituents, so no problems are likely to arise if acidic wastewaters are injected and incinerated. Nevertheless, a lining may be subject to attack by such acidic flue-gas constituents as hydrogen fluoride and phosporus pentoxide, as well as by saline vapors and saline melts. Appropriate preventive measures should be taken in the form of a suitable choice of materials and operating procedures. Damage to internal linings tends to be less with designs that incorporate cooling of the combustion- chamber wall, heat-transfer oils, or steam generation [466], [481].

If steam generation is envisaged then a proper choice of pressure and temperature is important with respect to low- and high-temperature corrosion attributable to chloride. Depending on the chloride content, corrosion can be kept, for example, to 0.1 mm/a by operating at 15 –30 bar, whereas at 6 bar and < 150 °C a corrosion rate of 1 mm/a must be anticipated. Above 300 °C there is a risk of stress corrosion.

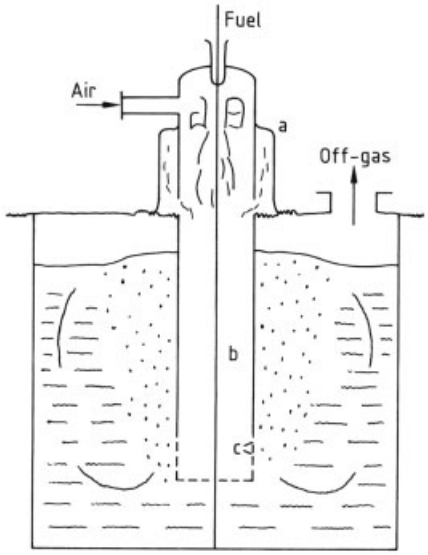

Figure 163. Schematic view of an immersion heating chamber
a) Burner and combustion chamber; b) Immersion tube; c) Exit slits

With relatively large amounts of wastewater, heavy-duty combustion chambers characterized by high turbulence are recommended, where the air and wastewater are likely to be charged tangentially into the combustion chamber [485], [486]. Especially high throughputs can be achieved with flame-vaporization burners developed specifically for wastewater incineration [487].

In case of a high salt content, some portion of the salts (depending on the precise salt composition and the mode of operation) can almost certainly be withdrawn in the form of a melt at the bottom (exit) of the first chamber, but a large fraction of the salt usually leaves the combustion chamber along with the flue gases, and must subsequently be separated in filters and/or quenchers or flue-gas scrubbers. The salt content of the flue gas can however be reduced, for example, with filtering devices or quenchers.

6.4.2. Immersion Heating Chambers

A different type of thermal wastewater disposal is possible with what is known as an *immersion burner* [485], [488]. An immersion heating chamber is a vertical combustion chamber the lower part of which is immersed in a water bath (Fig. 163). It is here that the flue gases are quenched and precleaned.

For exclusively volatile organic wastewater constituents, situations sometimes arise in which the prior stripping of these particular substances is advantageous (e.g., via an immersion burner), with subsequent normal or catalytic combustion of the vapors. In this case evaporation of the aqueous portion can largely be omitted [489].

6.4.3. Fluidized-Bed Furnaces

Wastewater, especially in concentrated form, just like sludge or other solids-rich material, can be incinerated directly, either alone or together with other substances (Fig. 164) [486], [490]. However, such incineration in a fluidized-bed furnace is usually conducted only as a low-temperature process at 700 – 900 °C. The wastewater is fed into the furnace from above, or from the side. Preheated air is blown in from below. A supplementary firing is usually useful or even necessary. The water fraction is rapidly vaporized in the fluidized bed by intensive exchange of heat and material, and the organic material is oxidized.

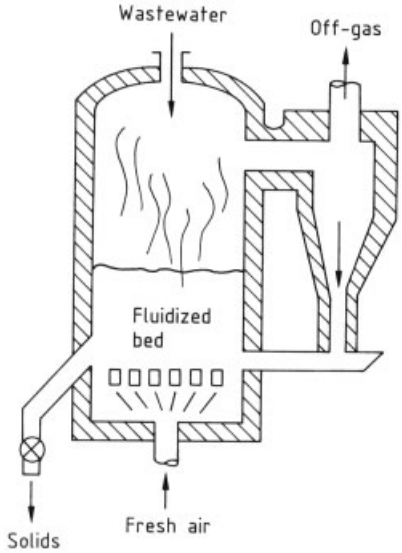

Figure 164. Example of a fluidized-bed furnace [486]

The fluidized bed may consist of an inert heat-transfer agent such as sand, or of incombustible inorganic constituents from the wastewater itself. The latter is a feasible alternative only if the wastewater has a uniform composition, since a consistent supplementation of the fluidized bed is essential.

The fluidized bed must be kept at a temperature such that the solid constituents do not melt. Saline components of the wastewater such as sodium chloride, sodium carbonate, sodium sulfate, etc. (all of which have low melting points in pure or eutectic form) may thus lead to problems.

In the ideal case, inorganic constituents remaining after vaporization of the water and oxidation of organic matter will agglomerate, facilitating their removal in pellet form. A temperature $< 800\,°C$ for the fluidized bed and the salt discharge is usually necessary to ensure that the process will function smoothly, but this means that not all organic substances will be completely combusted, placing restrictions on the applicability of the process. Should a combustion temperature of 900 – 1200 °C be formally required (e.g., by law) then the energy savings of ca.

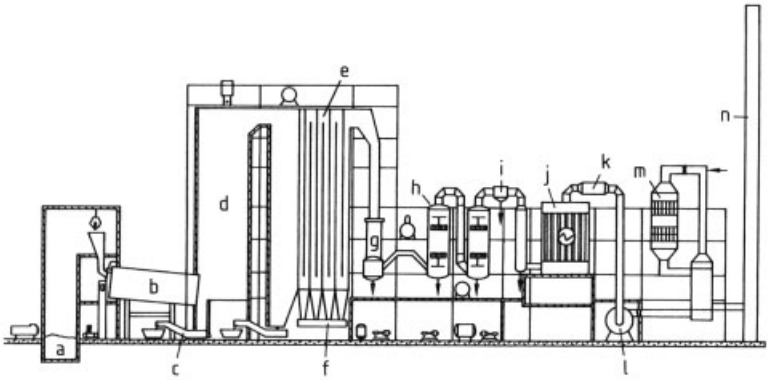

Figure 165. Example of a rotary-kiln system with flue-gas washing
a) Bunker; b) Rotary kiln; c) Deslagger; d) Afterburner chamber; e) Process-gas cooler; f) Deasher; g) Quenching system; h) Rotary disk washers; i) Condensor; j) Electrical precipitator; k) Gas heater; l) Fan; m) Selective catalytic-reduction unit for NO_x and dioxins; n) Stack

2500 kJ per kilogram of wastewater associated with low-temperature combustion must be sacrificed [474].

Fluidized-bed furnaces play a significant role especially in the combustion of pulp-digester lyes, including chemical recovery. Because of the uniform and high content of inorganic constituents in this case, all the prerequisites for fluidized-bed processing are met, and plants of this type with throughputs of 10 – 30 t/h have operated satisfactorily for years in Germany and elsewhere. Taken on balance, however, the fluidized-bed process is less flexible than combustion in a combustion chamber.

6.4.4. Rotary Kilns

Rotary kilns are cylindrical, rotating steel furnaces, usually with refractory linings. They are arranged approximately horizontally, with a slight downward inclination (e.g., ca. 3°) toward the exit end. The material to be incinerated is usually charged at the front end (see Fig. 165).

Rotary kiln facilities are more expensive than fixed combustion chambers, but also more versatile in their applications, so that in addition to solid or viscous wastes (including containers), waste solvents, other liquid residues, and solvent- containing or odoriferous air can be incinerated as well [491]. Moreover, in special cases even wastewaters can be injected directly into such a furnace or the post combustion chamber.

Given such a wide variety of potential feeds and the corresponding variations in pH, combustion temperature, and other parameters, increased attentiveness and operating experience are necessary to ensure an adequate life time for the furnace lining. An attempt is made to keep corrosion and erosion within reasonable limits by selecting the highest-quality ceramic-lining material available (stability to 1400 °C, in some cases 1700 °C) and always ensuring the presence of a protective skin of molten slag. Despite these precautions, the lining of the rotatory kiln often must be renewed once or twice annually, with a shutdown each time lasting ca. 3 – 4 weeks.

6.5. Emissions: Off-Gases, Water, and Solid Waste

Requirements for flue-gas cleaning are set in Germany by the 17th BImSchV (see Section 6.2). This prescribes a limiting value for polychlorinated dibenzodioxins and -furans of 0.1 mg/m^3 of total emissions (at 0 °C, 1.013 bar), a value that is usually met by such special gas-phase oxidation plants. It is particularly important that the requirements regarding the parameters "dust" (particles) and "NO_x" be strictly observed.

Wastewater. Depending on the nature of the outlet or discharge, wastewater from a combustion plant in Germany is subject to either the

22nd or the 47th VwV (Administrative Regulation). Plants currently in the design phase should also be considered in the light of EC regulations.

Solid Waste. Solid residues from a combustion plant are frequently soluble salts, which cannot simply be discarded in above-ground disposal sites. In the case of a single stream of wastewater, reuse should certainly be considered, and it may even prove economically justifiable. With mixed salts, underground disposal remains the only practical solution.

7. References

1. G. Garbrecht: *Wasser* : Vorrat, Bedarf und Nutzung in Geschichte und Gegenwart, Rowohlt, Reinbeck 1985.
2. J. Marcinek, E. Rosenkranz: *Das Wasser der Erde,* Verlag Harri Deutsch, Frankfurt 1989.
3. H. J. Voigt: *Hydro-Geo-Chemie,* Springer Verlag, Berlin 1990.
4. Umweltbereich Wasser, Fonds der Chemischen Industrie, Frankfurt 1992.
5. "Abwasserreinigung durch anaerobe biologische Behandlung," VCI-Verfahrensberichte zur Abwasserbehandlung, 11th report, Verband der Chemischen Industrie, Frankfurt 1986.
6. "Anaerobe Verfahren zur Behandlung von Industrieabwässern," Arbeitsbericht des ATV-Fachausschusses 7.5, Abwassertechnische Vereinigung, Hennef, *Korresp. Abwasser* **37** (1990) no. 10, 1247 – 1251.
7. T. Mann: "Die Entwicklung der Abwassertechnik und der Wasserreinhaltung," *Chem. Unserer Zeit* **25** (1991) no. 2, 87 – 95.
8. "Abwasserteiche für kommunale Abwässer," Arbeitsbericht der ATV Arbeitsgruppe 2.6, Abwassertechnische Vereinigung, Hennef, *Korresp. Abwasser* **26** (1979) no. 8, 403 ff.
9. K. R. Imhoff, G. W. Annen: "Neuere Erfahrungen mit Schönungsteichen," *GWF Gas – Wasserfach: Wasser/Abwasser* **134** (1993) no. 5, 282 – 285.
10. H. Fuhr, T. Mann: "Abgedeckte Kläranlagen," *Chem. Ind. (Düsseldorf)* **31** (1979) Jul., 477 – 482.
11. M. Bueb, M. Finzenhagen, T. Mann, K. Müller: "Verminderung der Abwasseremission in der Chemischen Industrie durch Verfahrensumstellungen und Teilstrombehandlungen," *Korresp. Abwasser* **37** (1990) no. 5, 542 – 558.
12. J. Wiesner: *Produktionsintegrierter Umweltschutz in der Chemischen Industrie,* DECHEMA, Frankfurt/Main 1990.
13. G. Baumann: "Zu viel und doch essentiell Phosphatbedarf von Belebtschlammanlagen," *Entsorgungstechnik* Mar./Apr. (1993) 52 – 57.
14. H. Hulpke, H. A. Koch, R. Wagner: *Römpp Lexikon Umwelt,* Thieme Verlag, Stuttgart 1993.
15. J. Falbe, M. Regitz (eds.): *Römpp Chemie Lexikon,* 9th ed., Thieme Verlag, Stuttgart 1989 – 1992.
16. H. Dellweg, R. D. Schmid, W. E. Trommer: *Römpp Lexikon Biotechnologie,* 1st ed., Thieme Verlag, Stuttgart 1992.
17. A. I. Laskin, H. A. Lechevalier: *CRC Handbook of Microbiology,* 2nd ed., **vol. 4**, CRC Press, Boca Raton 1982.
18. J. A. Borchardt, J. K. Cleland, W. J. Redmann, G. Olivier: *Viruses and Trace Contaminants in Waster and Waste Water,* Ann Arbor Science Publ., Ann Arbor 1977.
19. S. A. Ansari, S. R. Farrah, G. R. Chaudhry: "Presence of Human Immunodeficiency Virus Nucleic Acids in Wastewater and their Detection by Polymerase Chain Reaction," *Appl. Environ. Microbiol.* **58** (1992) 3984 – 3990.
20. D. O. Cliver: "Significance of Water and the Environment in the Transmission of Virus Disease," *Monogr. Virol.* **15** (1984) 30 – 42.
21. *Herder Lexikon der Biologie,* Herder, Freiburg 1983 – 1992.
22. C. J. Hurst: "Fate of Viruses during Wastewater Sludge Treatment Processes," *Crit. Rev. Environ. Control* **18** (1989) 317 – 343.
23. H. G. Schlegel: *Allgemeine Mikrobiologie,* 6th ed., Thieme Verlag, Stuttgart 1985.
24. Bayerisches Landesamt für Wasserwirtschaft: *Das mikroskopische Bild bei der Abwasserreinigung,* 2nd ed., Bayerisches Landesamt für Wasserwirtschaft, München 1992.
25. L. Habeck-Tropfke, H.-H. Habeck-Tropfke: *Abwasserbiologie,* 2nd ed., Werner-Verlag, Düsseldorf 1992.
26. J. Schwoerbel: *Einführung in die Limnologie,* 4th ed., G. Fischer, Stuttgart 1987.
27. A. M. Baya et al.: "Coincident Plasmids and Antimicrobial Resistance in Marine Bacteria Isolated from Polluted and Unpolluted Atlantic Ocean Samples," *Appl. Environ. Microbiol.* **51** (1986) 1285 –1292.

28. H. G. Rast, W. Springer: "Abbau von Mischsubstraten durch Spezialkulturen," *GWF Gas – Wasserfach: Wasser/Abwasser* **131** (1990) 516 – 521.
29. *Bergey's Manual of Systematik Bacteriology,* 4 vols., Williams and Wilkins, Baltimore 1984.
30. H. Blaim: "Floraanalysen an Abwasseranlagen der chemischen Industrie," Technische Universität München 1985.
31. C. van den Hoek: *Algen,* Thieme Verlag, Stuttgart 1978.
32. Umweltbereich Luft, Fonds der chemischen Industrie, Frankfurt/Main 1987.
33. K. Mudrack, S. Kunst: *Biologie der Abwasserreinigung,* 3rd ed., G. Fischer, Stuttgart 1991.
34. Abwassertechnische Vereinigung (ATV): "Abwasserbehandlung in Anlagen mit Sumpfpflanzen," *Korresp. Abwasser* **29** (1982) 161 – 163.
35. B. Zacher et al.: "Untersuchungen über den Einsatz von Pflanzen zur Klärschlamm-Entwässerung," *Korresp. Abwasser* **34** (1987) 922 – 931, 1197 – 1203.
36. A. Remane, V. Storch, U. Welsch: *Kurzes Lehrbuch der Zoologie,* 3rd ed., G. Fischer, Stuttgart 1978.
37. P. Brohmer: *Fauna von Deutschland,* Quelle und Meyer, Heidelberg 1977.
38. G. Reinheimer, W. Hegemann, J. Raff, I. Sekoulov: *Stickstoffkreislauf im Wasser,* Oldenbourg, München 1988.
39. R. Pudill, K. Kuczera: "Belebtschlamm mikroskopisch erfaßt," *Labor 2000* (1994) 64 – 73.
40. D. H. Eikelboom, H. J. J. Buijsen: *Handbuch für die mikroskopische Schlammuntersuchung,* 2nd ed., Hirthammer, München 1987.
41. "Verminderung und Bekämpfung von Blähschlamm und Schwimmschlamm," Arbeitsbericht der ATV-Arbeitsgruppe 2.6.1, *Korresp. Abwasser* **35** (1988) no. 2, 152 ff.
42. "Blähschlammbildung und -bekämpfung – Bestimmungstabelle für fadenförmige Mikroorganismen im Belebtschlamm," Arbeitsbericht der ATV-Arbeitsgruppe 2.6.1, *Korresp. Abwasser* **24** (1977) no. 1, 12 ff.
43. D. H. Eikelboom: "Filamentous Organisms Observed in Activated Sludge," *Water Res.* **9** (1975) 365 – 388.
44. Abwassertechnische Vereinigung (ATV): "Die Biologische Reinigung von Industrie-Abwasser nach dem Belebungsverfahren," Berichte zur Abwasser- und Abfalltechnik, GFA, Bonn 1970.
45. J. A. Soddell, R. J. Seviour: "Microbiology of Foaming in Activated Sludge Plants," *J. Appl. Bacteriol.* **69** (1990) 145 – 176.
46. H.-E. Reineck: *Das Watt – Ablagerungs- und Lebensraum,* 3rd ed., W. Kramer, Frankfurt/Main 1982.
47. J.-H. Fuhrhop: *Bio-organische Chemie,* Thieme Verlag, Stuttgart 1982.
48. P. Karlson: *Kurzes Lehrbuch der Biochemie für Mediziner und Naturwissenschaftler,* 13th ed., Thieme Verlag, Stuttgart 1988.
49. D. H. Zitomer, R. E. Speece: "Sequential Environments for Enhanced Biotransformation of Aqueous Contaminants," *Environ. Sci. Technol.* **27** (1993) 226 – 244.
50. F. Korte: *Lehrbuch der ökologischen Chemie,* 3rd ed., Thieme Verlag, Stuttgart 1992.
51. J. Kraut: "How do Enzymes Work?" *Science (Washington D.C.)* **242** (1988) 533 – 540.
52. "Biologische Phosphorentfernung," Arbeitsbericht der ATV-Arbeitsgruppe 2.6.6, *Korresp. Abwasser* **36** (1989) no. 3, 337 ff.
53. G. S. Sayler, S. W. Hooper, A. C. Layton, I. M. H. King: "Catabolic Plasmids of Environmental and Ecological Significance," *Microb. Ecol.* **19** (1990) 1 – 20.
54. B. D. Erickson, F. J. Mondello: "Enhanced Biodegradation of Polychlorinated Biphenyls after Site-Directed Mutagenesis of a Biphenyl Dioxygenase Gene," *Appl. Environ. Microbiol.* **59** (1993) 3858 –3862.
55. M. Boyle: "The Environmental Microbiology of Chlorinated Aromatic Decomposition," *J. Environm. Qual.* **18** (1989) 395 – 402.
56. M. G. Berger, R. E. Klaus, H. P. Fock: "Assimilation of Gaseous Ammonia by Sunflower Leaves During Photosynthesis," *Aust. J. Plant Physiol.* **13** (1986) 211 – 220.
57. J. Bever, A. Stein, J. Teichmann: *Weitergehende Abwasserreinigung,* 2nd ed., Oldenbourg, München 1993.
58. K. Krauth: "N_2O in Kläranlagen," *Korresp. Abwasser* **40** (1993) 1777 – 1791.
59. I. C. Anderson, J. S. Levine: "Relative Rates of Nitric Oxide and Nitrous Oxide Production by Nitrifiers, Denitrifiers, and Nitrate Respirers," *Appl. Environ. Microbiol.* **51** (1986) 938 – 945.
60. H. A. Painter: "Nitrification in the Treatment of Sewage and Waste-waters," *Spec. Publ. Soc. Gen. Microbiol.* **20** (1986) 185 – 211.
61. D. D. Focht, W. Verstraete: "Biochemical Ecology of Nitrification," *Adv. Microb. Ecol.* **1** (1977) 135 –214.

62. R. S. Bandurski: "Biological Reduction of Sulfate and Nitrate," in: J. Bonner, J. E. Varner (eds.): *Plant Biochemistry,* Academic Press, New York 1965, pp. 467 – 490.
63. T. Dalsgaard, F. Bak: "Nitrate Reduction in a Sulfate-reducing Bacterium Desulfovibrio Desulfuricans, Isolated from Rice Paddy Soil: Sulfide Inhibition, Kinetics, and Regulation," *Appl. Environ. Microbiol.* **60** (1994) 291 – 297.
64. L. I. Hochstein, G. A. Tomlinson: "The Enzymes Associated with Denitrification," *Annu. Rev. Microbiol.* **42** (1988) 231 – 261.
65. L. A. Robertson, E. W. J. van Niel, R. A. M. Torremans, J. G. Kuenen: "Simultaneous Nitrification and Denitrification in Aerobic Chemostat Cultures of Thiosphaera Pantotropha," *Appl. Environ. Microbiol.* **54** (1988) 2812 – 2818.
66. R. Körner, G. Benckiser, J. C. G. Ottow: "Quantifizierung der Lachgas-Freisetzung aus Kläranlagen unterschiedlicher Verfahrensführung," *Korresp. Abwasser* **40** (1993) 514 – 525.
67. S. Geywitz-Hetz, M. Bußmann, G. Schön: "Einfluß einiger Umweltfaktoren auf die N_2O-Bildung durch Belebtschlamm unter anoxischen Bedingungen," *Acta Hydrochim. Hydrobiol.* **21** (1993) 258 – 266.
68. T. Hillenbrand, E. Böhm: "Untersuchungen zur Verbesserung der Prozeßstabilität bei der Stickstoffelimination," *Korresp. Abwasser* **40** (1993) 536 – 554.
69. G. Baumann: "Zu viel und doch essentiell," *EntsorgungsTechnik* 1993, 52 – 57.
70. W. Nultsch: *Allgemeine Botanik,* 8th ed., Thieme Verlag, Stuttgart 1986.
71. J. S. Cech, P. Hartman, J. Wanner: "Competition between PolyP and non-PolyP Bacteria in an Enhanced Phosphate Removal System," *Water Environ. Res.* **65** (1993) 690 – 692.
72. N. Jardin: "Einfluß der biologischen P-Elimination auf die Schlammbehandlung," *Mitt. Oswald-Schulze-Stiftung* **16** (1993) 9/1 – 9/34.
73. D. Kaltenmeier: "Abwasserreinigung nach dem Stand der Technik in chemischen Großbetrieben," *Korresp. Abwasser* **37** (1990) 534 – 541.
74. P. Kunz, G. Frietsch: *Mikrobiozide Stoffe in Biologischen Kläranlagen,* Springer Verlag, Berlin 1986.
75. D. Schlee: *Ökologische Biochemie,* Springer Verlag, Berlin 1986.
76. N. Nirmalakhandan, V. Arulgnanendran, M. Mohsin, B. Sun, F. Cadena: "Toxicity of Mixtures of Organic Chemicals to Microorganisms," *Water Res.* **28** (1994) 543 – 551.
77. E. de Jong et al.: "Significant Biosynthesis of Chlorinated Aromatics by Fungi in Natural Environments," *Appl. Environ. Microbiol.* **60** (1994) 264 –270.
78. K. Naumann: "Chlorchemie in der Natur," *Chem. Unserer Zeit* **27** (1993) 33 – 41.
79. M. Alexander: "Biodegradation of Chemicals of Environmental Concern," *Science (Washington D.C.)* **211** (1981) 132 – 138.
80. N. Nyholm et al.: "Removal of Organic Micropollutants at ppb Levels in Laboratory Activated Sludge Reactors under Various Operating Conditions: Biodegradation," *Water Res.* **26** (1992) 339 –353.
81. M. Alexander: "Biodegradation of Organic Chemicals," *Environ. Sci. Technol.* **18** (1985) 106 – 111.
82. W. G. Haltrich: "Abbaubarkeit und tatsächlicher Eliminationsgrad von Einzelstoffen und Stoffgruppen bei der Chemieabwasserbehandlung," *Z. Wasser Abwasser Forsch.* **19** (1986) 4 – 9.
83. H.-M. Hwang, R. E. Hodson, D. L. Lewis: "Microbial Degradation Kinetics of Toxic Organic Chemicals over a Wide Range of Concentrations in Natural Aquatic Systems," *Environ. Toxicol. Chem.* **8** (1989) 63 – 74.
84. R. Wagner: *Methoden zur Prüfung der biochemischen Abbaubarkeit chemischer Substanzen,* VCH Verlagsges., Weinheim 1988.
85. *Amtsblatt der Europäischen Gemeinschaften* **35, L383A** (1992) 107 – 235.
86. Ministry of International Trade and Industry: *Handbook of Existing and New Chemical Substances,* Chemical Daily Co., Tokyo 1986.
87. "BSB", DIN 38 409, H 51, ISO 5815.
88. J. R. Parsons, H. A. J. Govers: "Quantitative Structure-activity Relationship for Biodegradation," *Ecotoxicol. Environ. Saf.* **19** (1990) 212 – 227.
89. P. Pitter, P. Lischke: "Biochemische Abbaubarkeit organischer Stoffe," *Acta Hydrochim. Hydrobiol.* **17** (1989) 217 – 226.
90. W. W. Eckenfelder, Jr.: "Aerobic Biological Treatment," in P. W. Lankford, W. W. Eckenfelder, Jr. (eds.): *Toxicity Reduction in Industrial Effluents,* Van Nostrand Reinhold, New York 1990.

91. F. S. Sariaslani: "Microbial Enzymes for the Oxidation of Organic Molecules," *Crit. Rev. Biotechnol.* **9** (1989) 171 – 257.
92. J. R. Mason, R. Cammack: "The Electron-transport Proteins of Hydroxylating Bacterial Dioxygenases," *Annu. Rev. Microbiol.* **46** (1992) 277 – 305.
93. S. Harayama, M. Kok, E. L. Neidle: "Functional and Evolutionary Relationships Among Diverse Oxygenases," *Annu. Rev. Microbiol.* **46** (1992) 565 –601.
94. D. T. Gibson: "Microbial Degradation of Aromatic Compounds," *Science (Washington D.C.)* **161** (1968) 1093 – 1097.
95. F. Rojo et al.: "Assemblage of Ortho Cleavage Route for Simultaneous Degradation of Chloro- and Methylaromatics," *Science (Washington D.C.)* **238** (1987) 1395 – 1398.
96. P. Fortnagel et al.: "Metabolism of Dibenzofuran by Pseudomonas sp. strain HH 69 and the Mixed Culture HH 27," *Appl. Environ. Microbiol.* **56** (1990) 1148 – 1156.
97. W. Springer, H. G. Rast: "Biologischer Abbau mehrfach halogenierter mono- und polyzyklischer Aromaten," *GWF Gas – Wasserfach: Wasser/Abwasser* **129** (1988) 70 – 75.
98. V. Strubel et al.: "Enrichment of Dibenzofuran Utilizing Bacteria with High Co-metabolic Potential Towards Dibenzodioxin and Other Anellated Aromatics," *FEMS Microbiol. Lett.* **58** (1989) 233 –238.
99. D. J. Hardman: "Biotransformation of Halogenated Compounds," *Crit. Rev. Biotechnol.* **11** (1991) 1 – 40.
100. W. Reineke, H.-J. Knackmuss: "Microbial Degradation of Haloaromatics," *Annu. Rev. Microbiol.* **42** (1988) 263 – 287.
101. D. L. Bedard et al.: "Extensive Degradation of Aroclors and Environmentally Transformed Polychlorinated Biphenyls by Alcaligenes eutrophus H 850," *Appl. Environ. Microbiol.* **53** (1987) 1094 –1102.
102. J. A. Bumpus, M. Tien, D. Wright, S. D. Aust: "Oxidation of Persistent Environmental Pollutants by a White Rot Fungus," *Science (Washington D.C.)* **228** (1985) 1434 – 1436.
103. D. K. Joshi, M. H. Gold: "Degradation of 2,4,5-Trichlorophenol by the Lignin-degrading Basidiomycete Phanerochaete chrysosporium," *Appl. Environ. Microbiol.* **59** (1993) 1779 – 1785.
104. Abwassertechnische Vereinigung (ATV): "Polychlorierte Dibenzo-p-dioxine and Polychlorierte Dibenzofurane," *Korresp. Abwasser* **41** (1994) 108 –116.
105. A. Grote, B. Hamburger, R. Kanne, M. Olivier: "Zum biologischen Abbau von 3,4-Dichlor-1-nitrobenzol unter den Bedingungen industrieller Kläranlagen," *Vom Wasser* **60** (1983) 191 –196.
106. B. E. Haigler, J. C. Spain: "Biodegradation of 4-Nitrotoluene by Pseudomonas sp. strain 4 NT," *Appl. Environ. Microbiol.* **59** (1993) 2239 – 2243.
107. C. Bruhn, H. Lenke, H.-J. Knackmuss: "Nitrosubstituted Aromatic Compounds as Nitrogen Source for Bacteria," *Appl. Environ. Microbiol.* **53** (1987) 208 – 210.
108. M. Luther: "Nitro- and Amino-substituted Aromatic Compounds as Nitrogen Sources for the Green Alga Scenedesmus obliquus," *Chemosphere* **21** (1990) 231 – 241.
109. J. C. Spain, D. T. Gibson: "Pathway for Biodegradation of p-Nitrophenol in a Moraxella sp., " *Appl. Environ. Microbiol.* **57** (1991) 812 – 819.
110. K. Valli, B. J. Brock, D. K. Joshi, M. H. Gold: "Degradation of 2,4-Dinitrotoluene by the Lignin-degrading Fungus Phanerochaete chrysosporium," *Appl. Environ. Microbiol.* **58** (1992) 221 – 228.
111. Beratergremium Umweltrelevante Altstoffe: *BUA-Stoffberichte* 11, 12, 41, 59, 102, VCH Verlagsges., Weinheim 1987 – 1991, Hirzel, Stuttgart 1993.
112. B. Nörtemann, H.-J. Knackmuss: "Abbau sulfonierter Aromaten,"*GWF Gas – Wasserfach: Wasser/Abwasser* **129** (1988) 75.
113. B. Nörtemann, J. Baumgarten, H. G. Rast, H.-J. Knackmuss: "Bacterial Communities Degrading Amino- and Hydroxynaphthalene-2-sulfonates," *Appl. Environ. Microbiol.* **52** (1986) 1195 – 1202.
114. B. Nörtemann et al.: "5-Hydroxyquinoline-2-carboxylic Acid, a Dead-end Metabolite from the Bacterial Oxidation of 5-Aminonaphthalene-2-sulfonic Acid," *Appl. Environ. Microbiol.* **59** (1993) 1898 –1903.
115. Beratergremium Umweltrelevante Altstoffe: *BUA-Stoffbericht* 63, VCH Verlagsgesellschaft, Weinheim 1991.
116. B. Gutekunst: "Sielhautuntersuchungen zur Einkreisung schwermetallhaltiger Einleitungen," *Korresp. Abwasser* **37** (1990) 1000 – 1004.
117. L. Wang, R. Govind, R. A. Dobbs: "Sorption of Toxic Organic Compounds on Wastewater

Solids: Mechanism and Modelling," *Environ. Sci. Technol.* **27** (1993) 152 – 158.
118. A. C. Petrasek et al.: "Fate of Toxic Organic Compounds in Wastewater Treatment Plants," *J. Water Pollut. Control Fed.* **55** (1983) 1286.
119. M. Tsezos, W. Seto: "The Adsorption of Chloroethanes by Microbial Biomass," *Water Res.* **20** (1986) 851 – 858.
120. H. Dieckmann: "Siderochromes [Iron(III)-trihydroxamates]," in A. I. Lastin, H. A. Lechevalier (eds.): *CRC Handbook of Microbiology,* 2nd ed., vol. 5, CRC Press, Boca Raton 1984, pp. 783 – 792.
121. J. H. Weber: "Review of Possible Paths for Abiotic Methylation of Mercury(II) in the Aquatic Environment," *Chemosphere* **26** (1993) 2063 – 2077.
122. D. R. Lovley: "Dissimilatory Metal Reduction," *Annu. Rev. Microbiol.* **47** (1993) 263 – 290.
123. "Neue Erkenntnisse zum Bau und Betrieb von Pflanzenkläranlagen," *Korresp. Abwasser* **39** (1992) 872 ff.
124. "CSB", DIN 38 409, H 41, ISO 6060.
125. "Elimination von Ammonium aus Abwässern," VCI-Verfahrensberichte zur Abwasserbehandlung, 12. Bericht, Frankfurt/Main 1989.
126. "Abwasserreinigung durch Denitrifikation," VCI-Verfahrensberichte zur Abwasserbehandlung, 13. Bericht, Frankfurt/Main 1990.
127. "Verfahren zur Elimination von Phosphor aus Abwasser," ATV-Regelwerk Abwasser-Abfall, Arbeitsblatt A 202, GFA, Hennef, Oct. 1990.
128. G. Jaekel: "Introduction of Fermentation in the Biological Purification of High Loaded Wastewater," *Münch. Beitr. Abwasser-Fisch. Flußbiol.* **22** (1972) 125 – 140.
129. I. Sekoulov: "New Developments in Industrial Wastewater Treatment 1991," *NATO ASI Ser., Ser. E* **191** (1991) 25 – 31.
130. "Abwasserbehandlung in Anlagen mit Sumpfpflanzen," Arbeitsbericht des ATV-Fachausschusses 2.10, *Korresp. Abwasser* **30** (1983) 161 ff.
131. "Behandlung von häuslichem Abwasser in Pflanzenbeeten," ATV-Regelwerk Abwasser-Abfall, Arbeitsblatt H 262, GFA, Hennef, August 1989.
132. R. G. Smith: "The Overland-Flow Process;" *Environ. Prog.* **1** (1982) no. 3, 195 – 205.
133. *Lehr- und Handbuch der Abwassertechnik,* 3rd ed., **vol. IV,** Ernst & Sohn, Berlin 1985 pp. 268 – 269.
134. M. Zlokarnik: "Neue Wege bei der flotativen Aufbereitung und Abwasserreinigung in der chemischen Industrie," *Chem. Ing. Tech.* **53** (1981) no. 8, 600 – 606.
135. D. C. Hempel: "Flotation in der kommunalen und industriellen Abwassertechnik," *Abwassertechnik* **6** (1993) 476 – 482.
136. M. Zlokarnik: "Belüftung und Abtrennung des Klärschlammes in Kläranlagen," *Umwelt* 1985, no. 3, 237 – 251.
137. E. Schmitt: "Vollbiologische Betriebskläranlage," *Brauwelt* **123** (1983) no. 42, 1830 – 1841.
138. G. Kühbeck: "Das neue Abwasserreinigungskonzept der Bitburger Brauerei – Verfahren und erste Erfahrungen," *Forum Brau.* **6** (1984) no. 6, 126 – 131.
139. K. Krauth: "Erhöhung der Biomasse durch Ultrafiltration bei aerob und anaerob betriebenen Bioreaktoren," *Korresp. Abwasser* **38** (1991) no. 9, 1150 – 1164.
140. K. H. Hunken, I. Sekoulov, D. Bardtke: "Intensive Sauerstoffversorgung von Belebungsanlagen mit Wasserstoffperoxid," *GWF Gas – Wasserfach: Wasser/Abwasser* **114** (1973) no. 4, 176 – 178.
141. D. A. Truesdale, A. L. Downing, G. F. Lowden: The Solubility of Oxygen in Pure Water and Seawater," *J. Appl. Chem.* **5** (1955) 53 – 62.
142. M. Zlokarnik: "Eignung und Leistungsfähigkeit von Belüftungsvorrichtungen für die biologische Abwasserreinigung," *Chem. Ing. Tech.* **52** (1980) no. 4, 330 – 331.
143. M. Zlokarnik, *Adv. Biochem. Eng.* **11** (1979) 157 –180.
144. M. Zlokarnik: "Koaleszensphänomene im System gasförmig/flüssig und deren Einfluß auf den O_2-Eintrag bei der biologischen Abwasserreinigung," *Korresp. Abwasser* **27** (1980) no. 11, 728 – 734.
145. T. Mann: "Die Abwassersituation der Bayer AG," *Korresp. Abwasser* **33** (1986) no. 10, 954 – 958.
146. "Umgestaltung 2-stufiger biologischer Kläranlagen zur Stickstoffelimination," Arbeitsbericht des ATV-Fachausschusses 2.6, *Korresp. Abwasser* **41** (1994) no. 1, 95 – 100.
147. I. Pascik: "Biochemische Stickstoffelimination aus ammoniumreichen Raffinerie-Abwässern in der Bayer-Turmbiologie," *Chem. Ing. Tech.* **54** (1982) no. 12, 1190 – 1192.
148. T. Mann, P. Reher, M. Willig: "Die Entwicklung der Abwasserentsorgung in

einem Großbetrieb der Chemischen Industrie," *Korresp. Abwasser* **35** (1988) no. 8, 799 – 810.
149. T. Mann: "Die Entwicklung der Klärschlammbehandlung im Gemeinschaftsklärwerk Leverkusen," *awt* (Abwassertechnik, Abfalltechnik und Recycling), 1989, no. 6.
150. "Verfahrenssicherheit industrieller Abwasserbehandlunganlagen," Arbeitsbericht der ATV-Arbeitsgruppe 7.02, *Korresp. Abwasser* **38** (1991) no. 9, 1239 – 1247.
151. J. Fritzmann: "Vorsorgekonzept der Chemischen Industrie; Maßnahmen zur Verhinderung gewässerrelevanter Unfälle, betriebsinterne Warnsysteme," Vortrag am 29. April 1993 bei Wasser, Berlin.
152. A 135 "Grundsätze für die Bemessung von Tropfkörpern und Tauchkörpern mit Anschluß-werten über 500 Einwohnergleichwerten," Gesellschaft zur Förderung der Abwassertechnik (GFA), Hennef 1989.
153. A 131 "Bemessung von einstufigen Belebungsanlagen ab 5000 Einwohnerwerten," Gesellschaft zur Förderung der Abwassertechnik, Hennef 1991.
154. M. Freund, R. Otterpohl, M. Dohmann: "Dynamische mathematische Modelle von Nachklärbecken – Übersicht und Vergleich," *Korresp. Abwasser* **40** (1993) no. 5, 738 – 746.
155. A. Deininger, V. Rehbein, P. A. Wilderer: "Untersuchungen zum Einfluß von Mischwasserzufluß auf die Leistungsfähigkeit von Nachklärbecken," *Korresp. Abwasser* **41** (1994) no. 2, 240 – 246.
156. T. Mann: "Abwasserreinigung in der Chemischen Industrie," *Entsorgungspraxis Spezial, Chemie und Umwelt* (1988) no. 3.
157. "Anwendung der Sauerstoffbegasung beim Belebungsverfahren," Teil 1 Arbeitsbericht des ATV-Fachausschusses 2.6, *Korresp. Abwasser* **9** (1975) 278 ff.
158. "Anwendung der Sauerstoffbegasung beim Belebungsverfahren," Teil 2 Arbeitsbericht der ATV-Arbeitsgruppe 2.6.2, *Korresp. Abwasser* **10** (1984) 855 ff.
159. J. R. Lemke, K. Mack: "Anwendung von reinem Sauerstoff in der Abwassertechnik," DECHEMA Monographie no. 1670 – 1692, **vol. 81**, Biotechnologie, Verlag Chemie, Weinheim 1977, pp. 233 –242.
160. R. P. Fields: "Deep or Semideepshaft Technology," *Symp. Pap.-Inst. Chem. Eng. North West Branch* 1992, no. 3, 14.1 – 14.13.
161. G. Diesterweg, H. Fuhr, P. Reher: "Die Bayer-Turmbiologie," *Industrieabwässer* 1976, no. 6, 7 – 13.
162. G. Diesterweg: "Turmbiologie – die Wirtschaftlichkeit eines modernen Verfahrens zur Abwasserreinigung," *Chem. Ind. (Düsseldorf)* **35** (1983) no. 2, 100 – 102.
163. M. Zlokarnik: "Tower-Shaped Reactors for Aerobic Biological Waste Water Treatment," in H. J. Rehm, G. Read (eds.): *Biotechnology,* VCH Verlagsges., Weinheim 1985, chap. 22, pp. 537 – 569.
164. G. Leistner, G. Müller, G. Sell, A. Bauer: "Bio-Hochreaktor – eine biologische Abwasserreinigungsanlage in Hochbauweise," *Chem. Ing. Tech.* **51** (1979) no. 4, 288 – 294.
165. Du Pont de Nemours, US 3 904 518, 1975.
166. Du Pont de Nemours, US 4 069 148, 1978 (D. G. Hutton, F. L. Robertacci).
167. T. Mann: "Biologisch-adsorptive Abwasserreinigung," *Korresp. Abwasser* **36** (1989) no. 9, 968 –974.
168. "Braunkohlenkoks in der biologischen Abwasserreinigung," *Wasser Abwasser Prax.* 1993, no. 5, 330 – 331.
169. M. Freytag, T. Mann: "Sichere Steuerung und Überwachung der Kläranlage Leverkusen," *Wasser Boden* **37** (1985) no. 10, 494 – 496.
170. U. Pagga: "Biologische Abwasserreinigung in einem Großbetrieb der Chemischen Industrie," *Vom Wasser* **67** (1986) 83 – 98.
171. R. Kanne: "Überwachung von Kläranlagenzuläufen mit Hilfe von gentechnisch konstruierten Leuchtbakterien," *Vom Wasser* **67** (1986) 195 – 199.
172. K. G. Malle, P. Tonne: "Die Abwasserreinigung der BASF AG in Ludwigshafen," *Korresp. Abwasser* **39** (1992) no. 4, 532 – 545.
173. R. Römer: "Klärschlammverbrennung BASF AG Ludwigshafen," Reprints GVC-Tagung "Entsorgung von Sonderabfällen durch Verbrennung," Baden-Baden, Dec. 4 – 6, 1989.
174. R. Römer: "Klärschlammverbrennung – Erfahrungen und Rahmenbedingungen," Reprints 2nd GVC-Kongreß, Würzburg, Oct. 19, 1992.
175. R. Römer, M. Neuwirth: *Zentrale Schwermetallfällanlage der BASF AG für Abwässer aus Rauchgaswäschen, Verfahrenstechnik der mechanischen, thermischen, chemischen und biologischen Abwasserreinigung,* **vol 1**: "Mechanische und chemische Verfahren," GVC Gesellschaft für Verfahrenstechnik und

Chemieingenieurswesen, Baden-Baden, Oct. 1988.
176. H. Fuhr, T. Mann: "Abgedeckte Kläranlagen," *Chem. Ind. (Düsseldorf)* **31** (1979) no. 7, 477 – 482.
177. T. Mann, P. Reher, M. Willig: "Die Entwicklung der Abwasserentsorgung in einem Großbetrieb der Chemischen Industrie," *Korresp. Abwasser* **35** (1988) no. 8, 799 – 810.
178. A. Bauer, G. Leistner: "Die zentrale Abwasserreinigunganlage des Werkes Hoechst der Hoechst AG," *Chem. Tech. (Heidelberg)* **7** (1978) 51 – 65.
179. G. Leistner, G. Müller, G. Sell, A. Bauer: "Der Biohochreaktor – eine biologische Abwasserreinigungsanlage in Hochbauweise," *Chem. Ing. Tech.* **51** (1979) no. 4, 288 – 294.
180. A. Bauer, G. Sell, L. Schäfer: "Einsatz von Aktivkohle im Biohochreaktor," *Chem. Tech. (Heidelberg)* **11** (1982) no. 6, 3 – 9.
181. G. Laubert: "Einsatz von Aktivkohle bei der Abwasserbehandlung," *Gewässerschutz Wasser Abwasser* **112** (1990) 443 – 449.
182. G. Baumann, H. Haffner, K. Hussel: "Sedimentieren, Neutralisieren, Reinigen, Entwässern, Verbrennen," *Lichtbogen* **199** (1981) 40 – 49.
183. G. Baumann: "Speichersysteme für Kühlwässer und Produktionsabwässer zum Schutz des Gewässers und der Kläranlage," *Forum Städte Hyg.,* **Special Issue, vol. 42** (1991) Jan./Feb., pp. 36 –40.
184. G. Baumann: "Behandlung von Industrieabwasser, Hüls AG – Wasserwirtschaft und Abwasserreinigung im Werk Marl," *Entsorgungs-Tech.* June/July (1992) 14 – 16.
185. Hüls AG: Wassernutzung und Abwasserbehandlung im Werk Marl; Informationen zum Umweltschutz, Marl 1993.
186. J. R. Lemke, K. Mack: "Biologische Abwasserreinigung mit technisch reinem Sauerstoff in der Kläranlage Bayer AG, Werk Wuppertal-Elberfeld," *Chem. Tech. (Heidelberg)* **7** (1978) 193 –196.
187. G. Klinzmann, D. Stammann, *GWF Gas – Wasserfach: Wasser/Abwasser* **122** (1981) 258 – 262.
188. H. Scholtyssek: "Biologische Stickstoffelimination aus ammoniumreichen Abwässern mittels Nitrifikation/Denitrifikation," *GVC-Tagung "Verfahrenstechnik der mechanischen, thermischen, chemischen und biologischen Abwasserreinigung,"* Baden-Baden, Oct. 17 – 19, 1988.
189. H. Scholtyssek: "Nitrifikation und Denitrifikation; kein Ammonium in den Vorfluter," *Chem. Ind. (Düsseldorf)* **43** (1990) 17 – 20.
190. I. Pascik: "Biochemische Stickstoff-Elimination aus ammoniumreichen Raffinerie-Abwässern in der Bayer Turmbiologie," *Chem. Ing. Tech.* **54** (1982) no. 12, 1190 – 1192.
191. K. Mudrack, S. Kunst: *Biologie der Abwasserreinigung,* G. Fischer Verlag, Stuttgart 1985, p. 105 ff.
192. K. Mudrack, E. Hellwig: "Reinigung eines Chemieabwassers unter Ausnutzung seines Nitratgehaltes," *Chem.-Ztg. Chem. Appar.* **93** (1969) no. 10, 362 – 369.
193. K. Mudrack: "Die Anwendung der mikrobiellen Denitrifikation zur biologischen Reinigung von Industrieabwasser," *GWF Gas – Wasserfach: Wasser/Abwasser* **112** (1971) no. 1, 33.
194. K. Mudrack, E. Hellwig: "Erfahrungen beim Einsatz des 'Denitrifikationsverfahrens' zur biologischen Reinigung von Industrieabwasser," *GWF Gas – Wasserfach: Wasser/Abwasser* **124** (1983) no. 10, 478 – 483.
195. M. Thüer, K. Coers: "Biologische Behandlung kombiniert mit Aktivkohle-Adsorption für schwierige Abwässer aus der Formulierung von Pflanzenschutzmitteln," *GWF Gas – Wasserfach: Wasser/Abwasser* **133** (1992) no. 3, 113 – 118.
196. K. R. Ronner, M. Thüer: "Neue Lösungen lassen alte Probleme verschwinden," *Ciba-Geigy-Magazin* **1** (1992).
197. M. Thüer: "Innovatives, maßgeschneidertes Verfahren zur Reinigung chemischer Problem-Abwässer," *Internationaler Kongreß für Umwelttechnologie und -forschung,* Basel, Oct. 5 – 7 1992.
198. M. Thüer: "Entwicklung eines Verfahrens zur Behandlung von Problemabwässern aus Formulierung und chemischer Produktion. Verfahrenstechnik der mechanischen, thermischen, chemischen und biologischen Abwasserbehandlung," 2nd *GVC-Kongreß,* Würzburg, Oct. 19 – 21, 1992.
199. J. Malaszkiewicz: "Können Indirekteinleiter den neuen Anforderungen genügen?", *ATV-Infoshop Indirekteinleiter,* Henkel KGaA, Essen, Nov. 26, 1992.

200. W. Schmitt, *Chem. Rundsch.* **45** (1993) no. 10 (Sonderausgabe Umweltschutz/Sicherheitstechnik).
201. The information regarding the Sewage-Treatment Plant-South was kindly placed at our disposal by Mr. Thomas Tscherner at the Sewer and Water Office of the city of Düsseldorf.
202. H. H. Würch, *Grundlagen Landtechnik* **31** (1981) no. 2, 55 – 60.
203. R. Kloss, *Grundlagen Landtechnik* **31** (1981) no. 2, 64 – 74.
204. H. W. Orth, *Grundlagen Landtechnik* **31** (1981) no. 2, 60 – 64.
205. F. Schuchardt, *Grundlagen Landtechnik* **31** (1981) no. 2, 42 – 47.
206. P. H. Smith, R. A. Mah, *Appl. Microbiol.* **14** (1966) 368 – 371.
207. G. Brune, S. M. Schoberth, H. Sahm, *Process Biochem.* **17** (1982) no. 3, 20 – 24, 35.
208. G. Brune, S. M. Schoberth, H. Sahm, *Appl. Environ. Microbiol.* **46** (1983) 1187 – 1192.
209. R. Hakulinen, S. Woods, I. Ferguson, M. Benjamin, *IAWPRC Symposium on Forest Industry Wastewaters,* Tampere, Finland, June 4 – 9, 1984.
210. M. Salkinoja-Salonen, R. Hakulinen, R. Valo, J. Apajalahti, *Water Sci. Technol.* **15** (1983) 309 –319.
211. M. P. Bryant, E. A. Wolin, M. J. Wolin, R. S. Wolfe, *Arch. Microbiol.* **59** (1967) 20 – 31.
212. R. E. Hungate et al., *J. Bacteriol.* **102** (1970) 389 –397.
213. R. K. Thauer, K. Jungermann, K. Decker, *Bacteriol. Rev.* **41** (1977) 100 – 180.
214. M. J. Wolin: *Microbial Formation and Utilization of Gases,* Verlag Goltze, Göttingen 1976, pp. 141 –150.
215. G. Lettinga, L. van Velsen, W. de Zeeuw, S. W. Hobma: *Anaerobic Digestion,* Applied Science Publ., London 1980, pp. 167 – 168.
216. V. H. Varel, H. R. Isaacson, M. P. Bryant, *Appl. Environ. Microbiol.* **33** (1977) 298 – 307.
217. V. H. Varel, A. G. Hashimoto, Y. R. Chen, *Appl. Environ. Microbiol.* **40** (1980) 217 – 222.
218. H. O. Buhr, J. F. Andrews, *Water Res.* **11** (1977) 129 – 143.
219. H. H. Therkelsen, D. A. Carlson, *J. Water Pollut. Control Fed.* **51** (1979) 1949 – 1964.
220. J. T. Pfeffer, *Biotechnol. Bioeng.* **16** (1974) 771 – 787.
221. F. G. Pohland, D. E. Bloodgood, *J. Water Pollut. Control Fed.* **35** (1963) 11 – 42.
222. W. F. Garber, G. T. Ohara, J. E. Colbaugh, S. K. Raksit, *J. Water Pollut. Control Fed.* **47** (1975) 950.
223. G. T. Ohara, *Proc. 2nd Nat. Conf. Munic. Sludge Manage. Disposal 1975.*
224. K. J. Kennedy, L. van den Berg, *Biotechnol. Lett.* **4** (1982) no. 3, 171 – 176.
225. G. Brune, Dissertation, Universität Düsseldorf 1982.
226. Kh. Bastin, C. Wandrey, *Ann. NY. Acad. Sci.* **369** (1980) 135 – 145.
227. Kh. Bastin, Dissertation, TU Clausthal 1983.
228. A. Aivasidis, C. Wandrey, *Ann. N.Y. Acad. Sci.* **413** (1983) 489 – 500.
229. C. Wandrey, A. Aivasidis, Kh. Bastin, *Ann. N.Y. Acad. Sci.* **469** (1986) 421 – 432.
230. W. J. Jewell et al. in: *Anaerobic Fermentation of Agricultural Residue: Potential for Improvement and Implementation,* U.S. Department of Commerce, Springfield, Virginia 1978.
231. S. Gosh, J. R. Conrad, D. L. Klass, *J. Water Pollut. Control Fed.* **47** (1975) 30 – 45.
232. M. L. Massey, F. G. Pohland, *J. Water Pollut. Control Fed.* **50** (1978) 2204 – 2222.
233. F. G. Pohland, S. Ghose, *Environ. Lett.* **1** (1971) 255.
234. P. Pipyn, W. Verstraete, *Biotechnol. Lett.* **1** (1979) 495 – 500.
235. P. Pipyn, W. Verstraete, *Biotechnol. Bioeng.* **23** (1981) 1145 – 1154.
236. A. Cohen, R. J. Zoetemeyer, A. van Deursen, J. G. van Andel, *Water Res.* **13** (1979) 571 – 580.
237. A. Cohen, A. M. Breure, J. G. van Andel, A. van Deursen, *Water Res.* **14** (1980) 1439 – 1448.
238. A. Cohen, A. M. Breure, J. G. van Andel, A. van Deursen, *Water Res.* **16** (1982) 449 – 455.
239. A. Bories, *Ann. Technol. Agric.* **29** (1980) no. 3, 509 – 528.
240. L. W. Hulshoff et al., *Water Sci. Technol.* **15** (1983) 291 – 304.
241. G. Lettinga et al., *Biotechnol. Bioeng.* **22** (1980) 699 – 734.
242. K. C. Pette, A. I. Versprille in D. E. Hughes et al. (eds.): *Anaerobic Digestion,* Elsevier Biomedical Press, Amsterdam 1981, pp. 121 – 133.
243. P. Pipyn, W. Verstraete, *Biotechnol. Lett.* **1** (1979) 495 – 500.
244. G. Lettinga et al., *Water Sci. Technol.* **15** (1983) 177 – 195.
245. P. M. Sutton, A. Li, R. R. Evans, S. Korchin: "Dorr-Oliver's Fixed Film and Suspended Growth Anaerobic Systems for Industrial Waste Water Treatment and Energy Recovery," *Proc. Ind. Waste Conf.* **37th** (1982) no. 10, 213.

246. A. Margaritis, C. R. Wilke: *Engineering Analysis of the Rotorfermentor. Developments in Industrial Microbiology,* **vol. 13**, American Institut of Biological Sciences, Washington D.C. 1972.
247. J. C. Young, P. L. McCarthy, *J. Water Pollut. Control Fed.* **41** (1969) 160 – 171.
248. L. M. Szendrey, *Environ. Prog.* **3** (1984) no. 4, 222 – 228.
249. Y. C. Wu, J. C. Kennedy, E. D. Smith, *Proc. Int. Conf. Fixed Film Biol. Processes 1st* 1982, 1495 –1520.
250. B. Frostell, *Vatten* **35** (1979) 169 – 175.
251. J. C. Jennet, N. D. Dennis, *J. Water Pollut. Control Fed.* **47** (1975) no. 1, 104 – 121.
252. P. Huysman, P. van Meenen, P. van Assche, W. Verstraete, *Biotechnol. Lett.* **5** (1983) no. 9, 643 –648.
253. Corning Glass, US 4 286 061, 1981 (R. A. Messing, R. A. Oppermann).
254. K. Wulfert, P. Weiland, *Chem. Ing. Tech.* **57** (1985) no. 5, 472 – 473.
255. Linde AG, DE 3 032 869, 1980 (U. Fuchs, H. Reimann).
256. P. Weiland, K. Wulfert, *BTF Biotech. Forum* **3** (1986) no. 3, 151 – 158.
257. Forschungszentrum Jülich, DE 3 345 691, 1986 (A. Aivasidis, C. Wandrey).
258. Forschungszentrum Jülich, US 4 532 042, 1985 (A. Aivasidis, C. Wandrey).
259. Forschungszentrum Jülich, Schott Glaswerke, US 5 096 814, 1992 (A. Aivasidis, C. Wandrey, W. Kiefer).
260. Forschungszentrum Jülich, US 4 670 140, 1987 (A. Aivasidis, C. Wandrey).
261. A. Aivasidis, C. Wandrey, *Korresp. Abwasser* **35** (1988) 216 – 230.
262. A. Aivasidis, *Bioprocess Eng.* **4** (1989) 19 – 25.
263. A. Aivasidis, C. Wandrey, *Chem. Ing. Tech.* **61** (1989) no. 6, 484 – 487.
264. A. Aivasidis, C. Wandrey, *Ann. N.Y. Acad. Sci.* **589** (1990) 599 – 615.
265. M. Radke, *Entsorgungs-Praxis* **10** (1989) 506 –510.
266. R. Schraewer, *Abwassertech. Abfalltech.* **39** (1988) 40 – 43.
267. M. Morper, R. Schönberger, *Abwassertech. Abfalltech.* **39** (1988) no. 2, 27 – 31.
268. F. B. de Walle, E. S. K. Chian, *Biotechnol. Bioeng.* **18** (1976) 1275 – 1295.
269. L. van den Berg, K. J. Kennedy, *Biotechnol. Lett.* **3** (1981) no. 4, 165 – 170.
270. L. van den Berg, C. P. Lentz, *Proc. Ind. Waste Conf.* **34th** (1979) 319 – 325.
271. K. J. Kennedy, L. van den Berg, *Agric. Wastes* **4** (1982) 151 – 158.
272. K. J. Kennedy, L. van den Berg, *Water Res.* **16** (1982) 1391 – 1398.
273. K. J. Kennedy, L. van den Berg, *Biotechnol. Lett.* **4** (1982) no. 3, 131 – 176.
274. K. J. Kennedy, R. L. Droste, *Biotechnol. Bioeng.* **27** (1985) 1152 – 1165.
275. K. J. Kennedy, R. L. Droste, *Water Res.* **20** (1986) no. 6, 685 – 695.
276. M. F. Hamoda, K. J. Kennedy, *Biotechnol. Bioeng.* **30** (1987) 272 – 281.
277. R. Samson, K. J. Kennedy, *J. Biotechnol.* **2** (1985) 95 – 106.
278. K. J. Kennedy, M. Muzar, G. H. Copp, *Biotechnol. Bioeng.* **27** (1985) 86 – 93.
279. W. J. Jewell, US 4 284 508, 1981.
280. Gist-Brocades, EP 0 028 846, 1983 (J. J. Heignen).
281. Gist-Brocades, EP 0 090 450, 1986 (J. J. Heignen, P. A. Lourens, A. J. Vixseboxse).
282. J. J. Heignen, *Chem. Ing. Tech.* **56** (1984) no. 7, 526 – 532.
283. I. Pascik, *Korresp. Abwasser* **34** (1987) no. 10, 1053 – 1057.
284. M. Salkinoja-Salonen et al., *Water Sci. Technol.* **17** (1985) 77 – 88.
285. P. M. Sutton, A. Li, *Water Sci. Technol.* **15** (1983) 333 – 344.
286. M. A. Bull, R. M. Sterritt, J. N. Lester, *Biotechnol. Bioeng.* **26** (1984) 1054 – 1065.
287. S. J. Chen, C. T. Li, W. K. Shieh, *J. Chem. Technol. Biotechnol. Biotechnol.* **35B** (1985) 101 – 109.
288. S. J. Chen, C. T. Li, W. K. Sieh, *J. Chem. Technol. Biotechnol. Biotechnol.* **35B** (1985) 183 – 190.
289. J. S. Rockey, C. F. Forster, *Chem. Eng. Res. Des.* **63** (1985) 300 – 304.
290. F. Toldra, A. Floss, J. L. Lequerica, S. Valles, *Appl. Microbiol. Biotechnol.* **23** (1986) 336 – 341.
291. W. Trösch, H. Chmiel, *Chem. Ind. (Düsseldorf)* **37** (1985) no. 8, 543 – 546.
292. R. A. Binot, H.-P. Nareau, E.-J. Nyus, *Water Sci. Technol.* **15** (1983) 103 – 115.
293. E. Oliva, J. C. Jacquart, C. Prevot, *Proc. IAWPRC Conf. "Technical Advances in Biofilmreactors,"* Nizza 1989, pp. 489 – 498.
294. R. Samson, S. R. Guiot, *J. Chem. Technol. Biotechnol. Biotechnol.* **35B** (1985) 63 – 74.
295. R. Fiebig, H.-W. Dellweg, *Biotechnol. Lett.* **7** (1985) no. 7, 487 – 492.
296. S. R. Guiot, L. van den Berg, *Biotechnol. Lett.* **6** (1984) 161 – 164.

297. K. J. Kennedy, S. R. Guiot, *Water Sci. Technol.* **18** (1986) 71 – 86.
298. L. Martensson, B. Frostell, *Water Sci. Technol.* **15** (1983) 233 – 246.
299. G. H. Fyun, T. N. Whitmore, *Biotechnol. Lett.* **4** (1982) no. 9, 577 – 582.
300. P. Scherer, M. Kluge, J. Klein, H. Sahm, *Biotechnol. Bioeng.* **23** (1981) 1057 – 1065.
301. I. Karube, S. Kuriyama, T. Matsunaga, S. Suzuki, *Biotechnol. Bioeng.* **23** (1981) 847 – 857.
302. G. J. Schröpfer, N. R. Ziemke, *Sewage Ind. Wastes* **31** (1959) 164 – 190, 697 – 711.
303. S. Schlegel, K.-H. Kalbskopf in D. E. Hughes et al. (eds.): *Anaerobic Digestion,* Elsevier Biomedical Press, Amsterdam 1981, p. 169.
304. S. Schlegel, Veröffentlichungen des Instituts für Siedlungswasserwirtschaft der TU Hannover, no. 39, 1974.
305. L. van den Berg, C. P. Lentz, *Proc. Ind. Waste Conf.* **33rd** (1978) 185 – 193.
306. L. Huss: "Applications of the Anamet Process for Waste Water Treatment," in D. E. Hughes et al. (eds.): *Anaerobic Digestion,* Elsevier Biomedical Press, Amsterdam 1981, pp. 137 – 150.
307. B. Frostell, *Chem. Ind. (London)* 1981, 465 – 469.
308. J. Kriese, Dissertation, Technische Universität Hannover 1989.
309. M. Zlokarnik, *Chem. Ing. Tech.* **53** (1981) no. 8, 600 – 606.
310. G. J. Schroepfer et al., *Sewage Ind. Wastes* **27** (1955) 460.
311. G. Lettinga, L. H. Pol, Preprints Verfahrenstechnik der Abwasserreinigung, **vol. 2**: "Biologische Verfahren," GVC Gesellschaft für Verfahrenstechnik und Chemieingenieurwesen-VDI (1988) 187 –202.
312. M. S. Schwitzenbaum, *Enzyme Microbiol. Technol.* **5** (1983) 242 – 250.
313. D. W. Taylor, R. J. Burm, *AIChE Symp. Ser.* **69** (1973) 30 – 37.
314. J. J. Heijnen et al., *GWF Gas – Wasserfach: Wasser/Abwasser* **126** (1985) no. 2, 81 – 87.
315. P. Keim, M. Luerweg, A. Aivasidis, C. Wandrey, *Korresp. Abwasser* **36** (1989) no. 3, 675 – 687.
316. P. Keim, Dissertation, Universität Bonn 1990.
317. P. Huysman, P. van Meenen, P. van Assche, W. Verstraete, *Biotechnol. Lett.* **5** (1983) no. 9, 643 –648.
318. A. Aivasidis, *GWF Gas – Wasserfach: Wasser/Abwasser* **130** (1989) no. 3, 163 – 174.
319. A. Aivasidis, C. Wandrey, E. Hilla, *Bioprocess Eng.* **4** (1989) 63 – 74.
320. *DVGW-Schriftenr. Wasser* **206** (1987).
321. H. Eckardt: "Sedimentation und Flotation," in [320], pp. 6.1 – 6.20.
322. F. Sarfert: "Entsorgung von Wasserwerksschlämmen," in [320], pp. 21.1 – 21.19.
323. K. Kermer: "Physikalisch-chemische Verfahren," part 1, *Schriftenreihe zur rationellen Wasserverwertung in der Industrie,* Verlag für Bauwesen, Berlin 1992.
324. GVC VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen (eds.): *Verfahrenstechnik der mechanischen, thermischen, chemischen und biologischen Abwasserreinigung,* **2 vols.,** Düsseldorf 1988.
325. T. Melin: "Flotation in der dezentralen Abwasserreinigung," in [324] vol. 1, pp. 67 – 87.
326. M. Zlokarnik: "Neue Flotationstechniken zur Abtrennung und Eindickung bei der biologischen Abwasserreinigung," *Chem. Ing. Tech.* **57** (1985) no. 5, 442 – 446.
327. VDI (eds.): "Biologische Abgasreinigung," *VDI-Ber.* **735** (1989).
328. VDI (eds.): "Fortschritte bei der thermischen, katalytischen, sorptiven und biologischen Abgasreinigung," *VDI-Ber.* **1034** (1993).
329. VDI-Richtlinie 3477, Biofilter, Dec. 1991.
330. VDI-Richtlinie 3478, Biowäscher, July 1985.
331. S. Stockhammer, W. Schäfer-Treffenfeldt, K. Zetzmann: "Einsatz von Biowäschern in der chemischen Produktion," *Chem. Ing. Tech.* **64** (1992) no. 2, 148 –155.
332. U. Kersting: "Biofilteranlagen im Bereich von Kläranlagen, Einsatzbereiche – Konstruktion – Kosten," in [327], pp. 313 – 329.
333. A. S. Ward: "Dissolved Air Flotation for Water and Wastewater Treatment," *Chem. Eng. Res. Des.* **70** (1992) part B, Nov, 214 – 218.
334. J. K. Edzwald, J. P. Malley, C. Yu: "A Conceptual Model for Dissolved Air Flotation in Water Treatment," *Water Supply* **9** (1991) 141 – 150.
335. K. J. Ives (ed.): *The Scientific Basis of Flotation,* Martinus Nijhoff Publ., The Hague 1984.
336. T. Zabel: "Flotation in Water Treatment," in [335], pp. 349 – 377.
337. H. H. Hahn: "Wastewater-Treatment," in [335], pp. 379 – 414.
338. H. H. Hahn: "Verfahrenstechnische Gesichtspunkte der Flotation in der Siedlungswasserwirtschaft," in [324], vol. 1, pp. 35 – 52.

339. M. Zlokarnik: "Neue Wege bei der flotativen Aufbereitung und Abwasserreinigung in der Chemischen Industrie," *Chem. Ing. Tech.* **53** (1981) no. 8, 600 – 606.
340. M. Zlokarnik: "Anwendung der Flotation bei der Reinigung der Prozeßabwässer der Chemischen Industrie sowie bei der Klärschlammabtrennung nach biologischer Abwasserreinigung," *Kem. Ind.* **34** (1985) no. 1, 1 – 6.
341. S. Cizinska, V. Matejo, C. Wase, Y. Klasson, J. Krejci, G. Dalhammer: "Thickening of Waste Activated Sludge by Biological Flotation," *Water Res.* **26** (1992) no. 2, 139 – 144.
342. H. Bennoit: "Anwendung der Entspannungsflotation bei der Abwasserreinigung in der Chemischen Industrie," in [324], vol. 1, pp. 53 – 65.
343. H. Bennoit: "Vorklärung von Chemieabwasser durch Sedimentation und Entspannungsflotation," *Chem. Ing. Tech.* **57** (1985) no. 2, 158 – 159.
344. S. M. Travers, D. A. Lovett: "Pressure Flotation of Abattoir Wastewaters Using Carbon Dioxide," *Water Res.* **19** (1985) no. 12, 1479 – 1482.
345. D. A. Lovett, S. M. Travers: "Dissolved Air Flotation for Abattoir Wastewater," *Water Res.* **20** (1986) no. 4, 421 – 426.
346. C. C. Ho, Y. K. Tan: "Comparison of Chemical Flocculation and Dissolved Air Flotation of Anaerobically Treated Palm Oil Mill Effluent," *Water Res.* **23** (1989) no. 4, 395 – 400.
347. D. C. Hempel: "Flotation in der kommunalen und industriellen Abwassertechnik," *Entsorgungspraxis* **11** (1993) no. 6, 476 – 482.
348. H. Kohler, W. Ernst: "Behandlung von chromhaltigen Spülwässern und Konzentraten aus einer Galvanik mit Hilfe der Elektroflotation," *GWF Gas – Wasserfach: Wasser/Abwasser* **130** (1989) no. 11, 578 – 582.
349. A. Y. Hosny: "Separation of Oil from Oil/Water Emulsions Using an Electroflotation Cell with Insoluble Electrodes," *Filtr. Sep.* **29** (1992) Sep./Oct., 419 – 423.
350. K. Kirchner et al.: "Grundlagen der biologischen Abgasreinigung," in [328], pp. 487 – 509.
351. K. Fischer: *Biologische Abluftreinigung,* Expert-Verlag, Ehningen 1990.
352. D. Bardtke: "Mikrobiologische Voraussetzungen für die biologische Abluftreinigung," in [351], pp. 1 –12.
353. K. Fischer: "Biofilter: Aufbau, Verfahrensvarianten, Dimensionierung," in [351], pp. 35 – 54.
354. E. Schippert: "Abscheidung von Lösemitteln im Biowäscher," in [328], pp. 577 – 609.
355. F. Wolff: "Biologische Abluftreinigung mit einem neuen Biowäscherkonzept," in [327], pp. 99 – 107.
356. G.-G. Börger: "Abscheidung organischer Dämpfe," *Chem. Ing. Tech.* **64** (1992) no. 10, 905 – 914.
357. D. Eitner: *Biofilter in der Abluftreinigung; Biomassen-Planung-Kosten-Einsatzmöglichkeiten, Umwelt, Special Sonderlösungen der Luftreinhaltung,* **vol. 19,** no. 3, VDI-Verlag, Düsseldorf 1989, pp. 224 – 229.
358. H. Kohler: "Biowäscher-Aufbau, Verfahrensvarianten, Dimensionierung," in [351], pp. 74 – 97.
359. W. Simmler, *Chem. Ing. Tech.* **56** (1984) 117.
360. W. Simmler, K. Holzer, *PdN-Ch. Praxis der Naturwissenschaften: Chemie* **8** (1986) 7.
361. W. Simmler, Proceedings Opening Conference, Nov. 14, 1990, Ost-West Wirtschafts Akademie, Berlin.
362. D. Bahnemann, *Nachr. Chem. Tech. Lab.* **42** (1994) 378.
363. W. Simmler, *Ecotoxicol. Environ. Saf.* **6** (1982) 399.
364. W. Simmler: *Chemistry of the Future,* Pergamon Press, Oxford 1984, pp. 63 – 80.
365. D. F. Bishop, G. Stern, M. Fleischmann, C. S. Marshall, *Ind. Eng. Chem. Proc. Des. Dev.* **7** (1968) 110.
366. K. Holzer et al., BMFT Forschungsbericht T-82-87, 1982.
367. W. G. Strunk: "Hydrogen Peroxide Treats Diverse Wastewaters," *Ind. Wastes* **25** (1979) no. 1, 32 – 35.
368. P. Schulte, M. Volkmer, F. Kuhn: "Aktiviertes Wasserstoffperoxid zur Beseitigung von Schadstoffen im Wasser (H_2O_2/UV)," *Wasser Luft Boden* **35** (1991) no. 9, 55 – 58.
369. W. H. Kibbel, Jr. et al.: "Hydrogen Peroxide for Industrial Pollution Control," *27th Annual Purdue Industrial Waste Conference* Part 2, West Lafayette, Ind., 1972.
370. W. H. Kibbel, Jr., *Ind. Water Eng.* **13** (1976) no. 4, 6.
371. I. Sekoulov, *Dechema-Monogr.* **81** (1977).
372. E. J. Keating, R. A. Brown, E. S. Greenberg: "Phenolic Problems Solved with Hydrogen Peroxide Oxidation," *Ind. Water Eng.* **15** (1978) no. 7, 22 –27.

373. H. Knorre: "Entgiftung von Abwässern aus Härtereien mit H_2O_2," *HTM Haerterei-Tech. Mitt.* **34** (1979) no. 1, 38 – 42.
374. E. T. O'Neill, H. M. Castrantas, E. J. Keating: "Tannery Waste Streams meet Effluent Regulations using Hydrogen Peroxide," *Proc. Ind. Waste Conf.* **33** (1978) (Publ. 1979) 471 – 478.
375. Bayer, DE-OS 2 927 911, 2 927 912, 1979 (K.-G. Wilms, H. Waldmann).
376. Forschungsvorhaben 02-WA-150 des BMFT, Bearbeiter Bayer AG; UMPLIS, Umweltforschungskatalog '82.
377. E. Gilbert: "Über den Abbau von organischen Schadstoffen im Wasser durch Ozon," *Vom Wasser* **43** (1975) 275 – 290.
378. D. M. Gurol, R. Vatistas: "Oxidation of Phenolic Compounds by Ozone and Ozone + UV Radiation," *Water Res.* **21** (1987) no. 8, 895 – 900.
379. C. von Sonntag: "The Chemistry Behind the Upgrading with UV Light," *Wasser Berlin* 1989, International Ozone Association V-1-1.
380. W. H. Glaze, J.-W. Kang, D. H. Chapin: "The Chemistry of Water Treatment Processes Involving Ozone, Hydrogen, Peroxide and Ultraviolet Radiation," *Ozone: Sci. Eng.* **9** (1987) 335 – 352.
381. J. Rudolph, *Korresp. Abwasser* **40** (1993) 1298.
382. E. Gilbert: "Ozonolysis of Chlorphenols and Maleic Acid in Aqueous Solution," 2nd International Ozone Symposium, Montreal, Canada, May 11 – 14, 1975.
383. G. Jayson, B. J. Parsons, A. J. Swallow: "Simple, Highly Reactive, Inorganic Chlorine Derivatives in Aqueous Solution," *J. Chem. Soc. Faraday Trans. 1* **69** (1973) no. 9, 1597 – 1607.
384. B. Höger: "Untersuchungen zum Abbau organischer Modellsubstanzen bei Anwendung des Systems O_3/H_2O_2 und Einfluß von Chlorid," Diploma Thesis, Universität Karlsruhe 1991.
385. W. R. Haag, J. Hoigné, H. Bader: "Ozonung bromidhaltiger Trinkwasser: Kinetik der Bildung sekundärer Bromverbindungen," *Vom Wasser* **59** (1982) 237 – 251.
386. M. Gratzel, M. Halmann: "Photosensitized Oxidation of Bromide in Dead Sea Water," *Mar. Chem.* **29** (1990) 2 – 3, 169 – 182.
387. E. de Jong et al., *Appl. Environ. Microbiol.* **60** (1994) Jan., 264.
388. E. de Jong et al., *Appl. Environ. Microbiol.* **60** (1994) Jan., 271.
389. K. Naumann, *Nachr. Chem. Tech. Lab.* **42** (1994) 389.
390. D. F. Ollis, H. Al-Ekabi (eds.): "Photocatalytic Purification and Treatment of Water and Air," *Trace Metals in the Environment,* vol. 3, Elsevier Science, London 1993.
391. C. Zetsch, *UWSF – Z. Umweltchem. Ökotox* **3** (1991) no. 1, 59.
392. W. Simmler: "Abiotic Degradation," Chemie Report no. 10, Verband der Chemischen Industrie, Frankfurt/Main 1991, p. 3.
393. W. H. Glaze, J. F. Kennecke, J. L. Ferry, *Environ. Sci. Technol.* **27** (1993) 177.
394. D. Bahnemann et al. in: *Aquatic and Surface Photochemistry,* Lewis Publishers, Boca Raton, FL 1994, p. 349.
395. D. Bahnemann, D. Boekelmann, R. Goslich, *Sol. Energy Mat.* **24** (1991) 564.
396. D. Bahnemann et al. in L. Schimmelpfeng (ed.): *Altlasten – Deponietechnik – Kompostierung: Praktizierter Umweltschutz als Vorsorge und Nachsorge,* Academia-Verlag, St. Augustin 1993, p. 294.
397. M. R. Hoffmann et al., *Angew. Chem.* **106** (1994) 1148.
398. M. Mitani et al., *J. Org. Chem.* **59** (1994) 1279.
399. F. J. Zimmermann, *Chem. Eng.* **65** (1958) 117 –120.
400. Sterling Drug, US 2 665 249, 1954 (F. J. Zimmermann).
401. Sterling Drug, US 2 824 058, 1958 (F. J. Zimmermann).
402. U. Lohmann, A. Tilly, *Chem. Ing. Tech.* **37** (1965) 913 – 960.
403. Bayer, DE-OS 2 260 114, 1972 (R. Thiel).
404. Sterling Drug, US 3 912 626, 1975 (R. B. Ely).
405. U. Fenchel, C. Ulrich, *Wochenbl. Papierfabr.* **19** (1976) 733 – 734.
406. Hoechst, DE-AS 2 657 011, 1976 (A. Topp).
407. B. L. Flynn, *Chem. Eng. Prog.* **75** (1979) April, 66 –69.
408. A. R. Wilhelmi, P. V. Knopp, *Chem. Eng. Prog.* (1979) Aug., 46 – 52.
409. F. J. Zimmermann, *Ind. Water Wastes* **6** (1961) 102.
410. E. Hurwitz et al., *Water Sewage Works* **112** (1965) 298.
411. J. J. A. Ploos van Amstel, Dissertation, T. H. Eindhoven 1971.
412. J. J. A. Ploos van Amstel, K. Rietema, *Chem. Ing. Tech.* **42** (1970) 981.
413. W. J. Fisher, *Water Res.* **5** (1971) 187 – 201.
414. J. J. A. Ploos van Amstel, *Ingenieur (The Hague)* **84** (1972) G 31 – G 36.

415. D. Wouda, *Ingenieur (The Hague)* **84** (1972) G 36 – G 45.
416. R. Randolf, *Wasserwirtsch. Wassertech. WWT* **18** (1968) 342.
417. Mitsui Toatsu Chemicals, DE-OS 2 425 587, 1974 (M. T. Okabe).
418. Produits Chimiques Ugine Kuhlmann, DE-OS 2 637 266, 1976 (M.-C. Daude-Lagrave).
419. Houston Research, DE-OS 2 618 338, 1976 (H. W. Prengle, Jr.).
420. Asahi Kasei Kogyo, DE-OS 2 534 458, 1975 (Y. Tagashira).
421. Osaka Gas, DE-OS 2 735 892, 1977 (N. Okada).
422. Bayer, DE-OS 2 640 603, 1976 (R. Thiel).
423. O. Horak, *Chem. Ing. Tech.* **62** (1990) 555.
424. K. Holzer, O. Horak, VDI-Gesellschaft Verfahrenstechnik u. Chemieingenieurwesen, Reprints 2nd GVC Kongreß, Würzburg 1992,**vol. 1**, C 1, p. 177.
425. K. Holzer, O. Horak, *Chem. Ind. (London)* 1994, nos. 7/8, 14.
426. K. Holzer, *27. Essener Tagung f. Wasser- und Abfallwirtschaft,* March 1994; Preprints, ed. by Inst. f. Siedlungswasserwirtschaft der RWTH Aachen 1995, Section 43, 1 – 16.
427. A. Schulz-Walz, R. Braden: "Abwasserreinigung durch Naßoxidation, Kosten des Verfahrens," *VDI-Ber.* **409** (1981), *Chem. Ing. Tech.* **53** (1981) 295.
428. T. Mann, I. Pascik, *Umwelt,* Suppl. 3 (1985) 19.
429. M. Bueb et al., *Korresp. Abwasser* **37** (1990) 542.
430. J. Tränkler et al., *Abwassertechnik* **44** (1993) 339.
431. Hoechst, DE-OS 2 748 638, 1977 (E. Paszthory).
432. H. Bretscher et al., *Chimia* **32** (1978) 180.
433. G. Friedhofen et al.: BMFT Forschungsbericht T80-132, 1980.
434. "Supercritical Water will be Used on Wastes," *Chem. Week* **130** (1982) no. 2, 50. "Using Supercritical Water to Destroy Tough Wastes, Chem. Week 130 (1982) no. 16, 26.
435. VCI-Verfahrensbericht: "Adsorptive Abwasserreinigung," Verband der chemischen Industrie, Frankfurt/Main 1988.
436. M. Bueb, M. Finzenhagen, T. Mann, K. Müller: "Verminderung der Abwasseremissionen in der chemischen Industrie durch Verfahrensumstellungen und Teilstrombehandlungen," *Korresp. Abwasser* **37** (1990) 542 – 558.
437. P. Kaiping: "Adsorptionsverfahren zur Abwasserbehandlung," *Verfahrenstechnik der mechanischen, chemischen und biologischen Abwasserbehandlung,* Preprints, 2nd GVC-Kongreß, Würzburg 1992, 157 – 176.
438. ASTM Standard D 2355-65 T.
439. H. Sontheimer et al.: *Adsorptionsverfahren zur Wasserreinigung,* ZfGW-Verlag, Frankfurt 1985, pp. 140 ff.
440. R. Brost, L. Hartinger, H. Hoffmann, W. Junge, T. Mann, R. Nagel: "Abwasserbehandlung mit Ionenaustauschern und Adsorberharzen," *Korresp. Abwasser* **41** (1994).
441. K. Periasamy, C. Namasivayan: "Process Development for Removal and Recovery of Cadmium from Wastewater by a Low-cost Adsorbent: Adsorption Rates and Equilibrium Studies," *Ind. Eng. Chem. Prod. Res. Dev.* **33** (1994) no. 2, 317 – 320.
442. T. Mann: "Biologisch-adsorptive Abwasserreinigung," *Korresp. Abwasser* **36** (1989) 968 – 974.
443. I. Sekoulov, S. Görg: "Festbettechnologie in der Industrieabwasserreinigung," Publication of the Institut für Siedlungswasserwirtschaft und Abfalltechnik der Universität Hannover (1994), Heft 87, 5. Hannoversche Industrieabwassertagung, pp. 95 –110.
444. W. Lin, A. S. Weber: "Aerobic Biological Activated Carbon (BAC) Treatment of a Phenolic Wastewater," *Environ. Prog.* **11** (1992) no. 2, 145 – 154.
445. ATV-Arbeitsbericht "Deponiesickerwasser-Behandlung," *Korresp. Abwasser* **40** (1993) no. 3, 365 –404.
446. I. Neff: "Reinigung organisch belasteter Abwässer mit Aluminiumoxid," *Chem. Tech. (Heidelberg)* **12** (1983) no. 1, 41 – 43.
447. F. H. Frimmel, B. Leisenheimer, M. Weis: "Adsorption Kinetics of Aquatic Fulvic Acids at Alumina. A Detection Comparison," *Vom Wasser* **77** (1991) 1 – 11.
448. Feldmühle AG, DE-OS 2 320 799, 1973 (Th. Ploetz, V. Franzen).
449. U. Donnert, S. H. Eberle: "Rückgewinnung von Phosphat aus Abwässern mittels Aluminiumoxid," *Chem. Ztg.* **105** (1981) no. 4, 113 – 114.
450. D. Gleisberg: "Phosphates and the Environment," *Kommunalwirtschaft* 1989, no. 4, 97 – 102.
451. F. Martinola: "Ion Exchangers and Adsorbers – Versatile Aids in the Chemical Industry," *Chem. Ing. Tech.* **51** (1979) no. 7, 728 – 736.

452. K. H. Radeke et al.: "Adsorption of Organic Materials from Aqueous Solutions on Synthetic Adsorption Resins," *Chem. Tech. (Leipzig)* **42** (1990) no. 8, 335 – 338.
453. A. Klassert: "Removal of Phenols and Polycyclic Aromatics from Wastewater with Adsorber Resin," *Gewässerschutz, Wasser, Abwasser* **125** (1991) 333 –355.
454. G. M. Guster et al.; "Sorption of Organics from Aqueous Solution onto Polymeric Resins," *Ind. Eng. Chem. Prod. Res. Dev.* **32** (1993) no. 11, 2727 – 2735.
455. R. Wagner, T. Augustin: "Neue Entwicklungen von Adsorberharzen und deren Einsatzgebiete für die Abreicherung von organischen Stoffen aus gewerblichen und industriellen Abwässern," lecture on the occasion of the Essener Tagung, March 9 – 11, 1994; published in *Gewässerschutz, Wasser, Abwasser* 1994.
456. G. Dummer, L. Schmidhammer: "Neues Verfahren zur Entfernung von aliphatischen Chlorkohlenwasserstoffen aus Abwässern mittels Adsorberharzen," *Chem. Ing. Tech.* **56** (1984) no. 3, 242 – 243.
457. G. Dummer, L. Schmidhammer: "Entfernung von aliphatischen Chlorkohlenwasserstoffen aus Abwässern mittels Adsorberharzen," *Chem. Ind. (Düsseldorf)* **39 (Special issue Chemie und Umwelt** (1984).
458. VCI-Verfahrensbericht: "Fällung und Flockung," Verband der Chemischen Industrie, Frankfurt/Main, March 1977.
459. K. Imhoff, K. R. Imhoff: *Taschenbuch der Stadtentwässerung,* 28th ed., R. Oldenbourg Verlag, München 1993.
460. F. Pöpel (ed.): *Lehrbuch für Abwassertechnik und Gewässerschutz,* Deutscher Fachschriften-Verlag, Wiesbaden 1975/1993.
461. Abwassertechnische Vereinigung e. V. (ATV): Arbeitsblatt A 203, Abwasserfiltration, Hennef 1993 (draft).
462. VCI-Verfahrensbericht: "Filtration und Flockungsfiltration," Verband der Chemischen Industrie, Frankfurt/Main 1992.
463. L. Svarovsky (ed.): *Solid-Liquid Separation,* 3rd ed., Butterworths, London 1990.
464. Abwassertechnische Vereinigung e. V. (ATV): *Lehr- und Handbuch der Abwassertechnik,* 3rd ed., **vol. 3**, Ernst & Sohn, Berlin 1983.
465. W. W. Eckenfelder: *Industrial Water Pollution Control,* 2nd ed., McGraw-Hill, New York 1989.
466. G. Mischer, K. P. Kiesewetter, R. Krebber, *Chem. Ing. Tech.* **57** (1985) no. 10, 859 – 863.
467. W. Leuckel, R. Roemer, *VDI-Ber.* **286** (1977) 141 –149.
468. P. Jaeger, W. Jarosch, *Chem. Ing. Tech.* **52** (1980) no. 6, 527 – 530.
469. Verband der Chemischen Industrie, Verfahrensbericht Abwasserbehandlung, Frankfurt/Main, March 1988.
470. D. F. Othmer, *Dechema Monogr.* **86** (1979) nos. 1775 – 1801, 597.
471. K. Hess, W. Leuckel, E. Minx, *Chem. Ing. Tech.* **44** (1972) no. 5, 325 – 332.
472. Bayer, EP 0 531 864 A1, 1992 (D. Tegtmeyer, F. W. Baumann, G. Stephan).
473. A. Juhnke, legal department Bayer AG (personal communication).
474. G. Hörner, *Wärme* **83** (1977) no. 5, 77 – 87.
475. K. G. Malle, *Korresp. Abwasser* **29** (1982) no. 7, 460 – 463.
476. K. Steinebrunner, H. Seifert, R. Becker, *Chem. Ing. Tech.* **MS 2073/92.**
477. G. Hebbel, *Chem. Ing. Tech.* **48** (1976) no. 1, 12 –15.
478. K. Lohhoff, T. Burkard, *Wärme Gas Int.* **32** (1983) no. 5, 206 – 211.
479. T. Burghard, *Industriefeuerung* **13** (1979) 43 – 49.
480. H. Salomon, *Vom Wasser (Jahrbuch)* **39** (1972) 223 – 230.
481. H. P. Jäger, *Tech. Mitt.* **80** (1987) no. 6, 384 –391.
482. G. Winkler, *Chem. Ind. Int. (Engl. Transl.)* 1991, no. 6, 18 – 22.
483. G. Kurz, H. Guthoff, *Chemieanlagen und Verfahren* **15** (1982) no. 9, 55 – 62.
484. BASF, DE 41 01 497 A1, 1991 (R. Becker et al.).
485. G. Kurz, R. Rud, *Chem. Ing. Tech.* **53** (1981) no. 5, 379 – 383.
486. M. V. Bernadiner et al., *Int. Chem. Eng.* **13** (1973) no. 2, 256 – 262.
487. J. Recknagel, R. Roemer, *Wärme Gas Int.* **28** (1979) no. 5, 252 – 259.
488. K. Schlömer, *Chem. Ing. Tech.* **48** (1976) no. 1, 41 – 44.
489. G. Kurz, *Chem. Rundsch.* **25** (1972) 1681.
490. Bertrams, EP 0 131 116 A1, 1984 (H. Kühnlein, H. J. Mildenberger).
491. H. Fuhr, *Hazard. Waste Hazard. Mater.* **2** (1985) 1 – 5.

Raw Materials and Energy

Rolf Kola, B.U.S. AG, Eschborn, Federal Republic of Germany (Chaps. 1, 2)

Otto von Elsner, Wintershall AG, Kassel, Federal Republic of Germany (Sections 3.1–3.3)

Wolfgang Riepe, Universität Paderborn, Fachbereich 13, Paderborn, Federal Republic of Germany, Universität Salzburg, Institut Chemie und Biochemie, Salzburg, Austria (Section 3.4)

Knud Reuter, Bayer AG, Krefeld-Uerdingen, Federal Republic of Germany (Chap. 4)

1. Introduction . . . 2
2. Mineral Raw Materials . . . 2
2.1. Formation of Deposits . . . 3
2.2. Mineralogy and Raw Material Groups . . . 4
2.3. Development of Deposits . . . 4
2.4. Exploitation of Raw Materials . . . 4
2.5. Availability of Mineral Raw Materials . . . 5
2.6. Environmental Factors in the Processing of Raw Materials . . . 6
2.7. Secondary Raw Materials . . . 6
3. Oil, Gas, and Coal . . . 8
3.1. Introduction . . . 8
3.2. Oil . . . 9
3.2.1. Exploration . . . 9
3.2.2. Production . . . 9
3.2.3. Transportation . . . 9
3.2.4. Loading and Storage . . . 11
3.2.5. Processing . . . 11
3.2.6. Uses . . . 11
3.2.7. Cost . . . 12
3.3. Gas . . . 12
3.3.1. Processing . . . 12
3.3.2. Transportation . . . 12
3.3.3. Storage . . . 12
3.3.4. Uses . . . 13
3.3.5. Cost . . . 13
3.4. Coal . . . 13
3.4.1. Mining Engineering . . . 13
3.4.2. Processing . . . 13
3.4.3. Utilization . . . 14
4. Renewable Resources . . . 19
4.1. Significance . . . 19
4.2. Important Renewable Resources 20
4.3. Industrial Use . . . 20
4.3.1. Natural Fats and Oils . . . 20
4.3.2. Cellulose and Its Derivatives . . . 21
4.3.2.1. Cellulose . . . 21
4.3.2.2. Cellulose Esters . . . 22
4.3.2.3. Cellulose Ethers . . . 23
4.3.3. Starch . . . 23
4.3.4. Low-Molecular Carbohydrates – Sugars . . . 24
4.3.5. Other Renewable Resources . . . 24
4.3.5.1. Polyisoprene . . . 24
4.3.5.2. Natural Resins . . . 25
4.3.5.3. Plant Polysaccharides (“Gums”) . . 25
4.3.5.4. Active Substances Based on Renewable Resources . . . 25
4.3.5.5. Flavors and Fragrances . . . 25
4.4. Energy Use . . . 26
4.4.1. Biofuels . . . 26
4.4.2. Biogas . . . 26
4.4.3. Heat from Biomass . . . 27
4.5. Appraisal . . . 27
4.5.1. Ecobalancing of Renewable Resources . . . 27
4.5.2. Opportunities and Limits of Use of Renewable Resources . . . 27
5. References . . . 27

1. Introduction

“Raw materials” is the name given to those materials whose constituents – after physical or chemical treatment – can be utilized by humans as a whole or in part, or whose stored binding energy can be released in the form of thermal energy.

Naturally occurring materials (e.g., ore, rock, or fossil material) are known as *primary raw materials*. Those materials that have already been in use or subjected to a physicochemical process, but whose constituents can be recovered and consequently returned to the material cycle, are termed *secondary raw materials*. Examples

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_153

of these are scrap materials, metallurgical dust and slag, waste glass, and wastepaper.

Renewable resources are raw materials of vegetable or animal origin whose availability can be secured by fresh planting or breeding.

The need for raw materials results from the volume of all commodities and consumer goods that must be produced and from the amount of energy that must be used in the extraction of raw materials, the production of intermediate and end products, and the use of goods during their service life. From this, the important factors affecting raw material consumption derive directly: (1) population density or growth, and (2) standard of living or degree of industrialization.

The extraction of natural raw materials and the production and use of consumer goods are basically associated with effects on the natural environment. The effects can be described in a simplified way as follows:

*Extraction of Raw Materials*Impact on the countryside and frequently on the water resources of a region, disturbance of natural biotopes, consumption of energy
*Processing of Raw Materials*Production of emissions in air and water, formation of unusable residues, consumption of chemical and physical energy
*Manufacture of Consumer Goods*Production of emissions, formation of further residues, consumption of energy
*Use of the Goods*Consumption of further materials and energy, creation of waste at the end of the service life, and consumption of disposal site volume

Against the background of a steadily increasing world population and the rise of industrialization and of living standards in the newly industrialized and developing countries, effects on the environment caused by the totality of human activities and questions about the nature and availability of raw materials are becoming increasingly important.

2. Mineral Raw Materials

Mineral is a collective name for all elements and their compounds, mainly inorganic and rarely organic in nature, that occur naturally as constituents of the earth's crust. Solidified minerals, as an aggregate, form the rocks of the earth's crust; 99.4 % of the earth's crust is composed of compounds of only 11 of the 92 natural elements:

SiO_2	59.12 %
Al_2O_3	15.34 %
CaO	5.08 %
Na_2O	3.84 %
FeO	3.81 %
Fe_2O_3	3.08 %
MgO	3.49 %
K_2O	3.13 %
H_2O	1.15 %
TiO_2	1.05 %
P_2O_5	0.30 %

The distribution of chemical elements in the earth's crust is not homogeneous. Dynamic geological changes led to element regroupings (i.e., enrichment and depletion). If a regional element enrichment reaches a mineable concentration, this is called a deposit. Deposits form the natural basis of the raw materials industry.

2.1. Formation of Deposits

The most important deposits are the magmatic deposits. When the magma cooled down slowly, dark, low-quartz or quartz-free rocks were formed in this *early crystallization phase*. These rocks are based mainly on magnesium – iron silicates. Simultaneously, metals such as Cr, Pt, V, Ni, and Cu separated from the liquid magmatic phase and accumulated in the magma chambers (e.g., Ni – Cu – Pt deposits in Sudbury, Canada; → Platinum Group Metals and Compounds, Chap. 3.3.). The subsequent *main crystallization phase* of the magma led to the formation of quartz-bearing eruptive rocks, such as granite. Further cooling and the *late crystallization* occurring at the same time resulted in superheated aqueous solutions (hydrothermal springs), which absorbed a number of the metals important today such as Pb, Zn, Cu, Mo, U, Bi, Hg, Ag, and Au, as well as Ta, Zr, and Nb. Because of temperature and pressure gradients, the solutions diffused into the cavities, veins, and clefts formed during the consolidation stage. Because of appropriate physical and chemical conditions, the metals separated in the form of mixed crystals (e.g., the aurifer-

rous veins of Idaho Springs, United States or the zinc – lead deposits in Trepca, Serbia; → Lead). If, as a result of volcanic processes, hydrothermal springs emerged on the sea bottom, the dissolved metals were precipitated and formed volcanogenic – sedimentary deposits (Zn – Pb – Cu deposits in Kidd Creek, Canada or Pb – Zn deposits in Mount Isa, Australia). If metal-bearing rocks were brought up to the surface by emergence of land and therefore into the weathering zone, they were eroded as a result of physical or chemical processes and redistributed. In this way the so-called sedimentary deposits, in some cases with considerable enrichment, were formed (e.g., the copper belt of Zambia – Zaire; → Copper, Chap. 4.2.; gold placers in South Africa).

Through the settling of rock formations to greater depths, partial melting and new enrichment occurred, and *metamorphic* deposits were formed (e.g., the Zn – Pb deposits in Sulitjelma, Norway or Broken Hill, Australia).

Table 1. Mineral raw materials from the viewpoint of the raw materials industry

Raw material group	Examples
1. Ores	
1.1 Iron ores	hematite, magnetite, pyrite, goethite, siderite
1.2 Nonferrous ores	lead glance, zinc blende, copper pyrite, bauxite, cassiterite
1.3 Ores of alloying metals	molybdenum glance, chromite, vanadinite, cobalt glance, wolframite, rutile
2. Noble metals	metallic gold, platinum metals
3. Natural rocks	
3.1 Magmatic rocks	granite, basalt, pumice
3.2 Sedimentary rocks	limestone, sandstone, tufa
3.3 Metamorphic rocks	marble, slate, serpentine
4. Salt rocks	rock salt, sylvine, Glauber's salt
5. Quartz minerals	quartz, chalcedony, kieselguhr, flint
6. Clay minerals	kaolin, montmorillonite, illite, chlorite
7. Precious and decorative stones	beryl, jade, opal, topaz, turquoise, garnet
8. Rare earths, reactor fuels	monazite, cerite, thorite, pitchblende
9. Other minerals	
Nitrogen minerals	Chile saltpeter, saltpeter
Phosphate minerals, boron minerals	apatite, phosphorite, borax, colemanite

2.2. Mineralogy and Raw Material Groups

In *systematic mineralogy* the minerals are classified by the criteria of chemical composition and crystal structure into nine main groups:

Class I	elements
Class II	sulfides, compounds of As, Sb, Br, Se, and Te
Class III	halide salts
Class IV	oxides, hydroxides
Class V	nitrates, borates, carbonates
Class VI	sulfates, chromates, molybdates, tungstates
Class VII	phosphates, arsenates, vanadates
Class VIII	silicates
Class IX	organic compounds

From the viewpoint of the raw materials industry a grouping such as that shown in Table 1 is appropriate.

2.3. Development of Deposits

Deposit research is concerned with ascertaining the formation of individual deposits and presenting the results in the form of geoscientific (metallogenetic) models. Such a metallogenetic model, for example, led to the discovery of porphyry copper ores on the West Coast of North and South America, in the Southeast Asian island arc, and in the Balkans. About 50 % of present-day copper production originates from such deposits.

The *prospecting and development* of raw material deposits is divided into three main stages.

The first stage, *prospecting*, seeks to determine geological structures with potential enrichments in a relatively large target area, and possibly broadly to delimit these structures. During the second stage, *exploration*, the area in which recoverable deposits are suspected is explored in more detail with regard to the concentration, extent, and character of the mineral body. The following phase is the *exploitation* of the deposit. In this, the type of ore or mineral is more closely identified and the content of the entire mineral body calculated. Noteworthy effects on the environment do not appear up to the development of the deposit, if the negligible effects of trial borings are disregarded. This can change, however, drastically when *mining* begins. In many countries, therefore, extensive environmental compatibility tests must be carried out before a min-

ing permit is granted. In a feasibility study, the deposit is evaluated with regard to both technical and economic aspects. This means that the mineral must be minable by the application of available technologies, the necessary energy supply must be guaranteed, and transportation problems must be solved. The value of the ore body must cover the cost of exploration, mining, treatment, and marketing, as well as amortization of the investment within an acceptable period of time. If the profitability study leads to positive conclusions, the deposit is opened up. The time between exploration and opening is between five and ten years. The investment can be as much as $\$2\times10^9$ under difficult conditions, such as extreme climatic conditions, a lack of infrastructure, and long transportation routes [e.g., in the case of the copper – gold deposit of Ok Tedi (Papua New Guinea)]. Such projects can be handled only by financially strong international consortia.

2.4. Exploitation of Raw Materials

Depending on the structure and position of the mineral body, mining is carried out either two-dimensionally above ground (e.g., in the case of the saltpeter deposits in Atacama, Chile) or in the form of the technically simple and also very economical open-pit mining (e.g., the gigantic iron ore deposits in Minas Gerais, Brazil or Hamersley Range, Australia). In many cases, the position and structure of the ore body require technically complex underground mining.

With the exception of natural rocks such as granite, basalt, limestone, and marble, and of salt minerals such as rock salt, saltpeter, or sylvine, the substance of value in the extracted ore can be assumed to have too low a concentration to make direct metallurgical processing economically possible. In the case of low-content ores – particularly mixed ores with complex interstratification – expensive physicochemical methods are used: crushing, classification, pulverizing, flotation, or the sink-float process. In this way, not only can the gangue be removed, but concentrates of specific metals can also be produced. With decreasing metal content, however, the energy requirements rise exponentially; in many cases, therefore, the availability and price of energy determine the feasibility of a deposit.

The disposal of worthless rock, the creation of cavities or gigantic pits aboveground, the pumping away of mine water and, associated with it, the lowering of groundwater level – aside from intrusion into the landscape and disturbance of natural biotopes – represent the main environmental impacts from the mining of mineral law materials.

The ecological effect profile can therefore be classified as follows:

1) Interference with the natural water distribution in an area and, resulting from this, microclimatic change for plants and ground animals
2) Formation of emissions by deflation of heaps and dumps and from the operation of mechanical processing plants
3) Creation of wastewater through drainage of mine water and operation of flotation plants
4) Formation of residues as a result of overburden and separated gangue
5) Alteration of the natural appearance of the landscape (especially by open-pit mining)

With the exception of emissions into air and water, the remaining effects generally remain regionally limited and can be minimized by appropriate measures.

The depression funnel formed by pumping off mine water is generally many times the actual area of mining. Formation of the depression funnel can be partially compensated by extensive seepage of the water, instead of pumping it off. In many cases, previous treatment of the mine water is necessary. Iron and manganese, for example, must be removed, and in the case of nonferrous deposits, heavy metals (lead, copper, zinc, etc.), and possibly anions such as sulfide and sulfate, must be removed as well.

In open-pit mining, ca. 16×10^9 t of rock is produced per year worldwide, of which about 10×10^9 t is disposed as waste. By refilling, usually after having been stored in bulk for three to eight years, the greater part of a mined area can be recultivated. A mass deficiency remains, however, which can either be covered by filling with other, suitable residues or remain in the form of artificial lakes.

The ground surface must be protected from extensive subsidence, especially in underground mining. Subsidence occurs either through sagging of the overburden rock into the cavities

formed by mining or by extensive change of the flows of groundwater and associated subrosive effects (i.e., dissolution of rocks or rock constituents). The effects can be largely limited by a controlled water-distribution system and refilling the galleries and shafts.

2.5. Availability of Mineral Raw Materials

The question of availability and resources arises primarily with reference to the most important commodity metals such as iron, aluminum, copper, zinc, chromium, or nickel.

Salt minerals such as halides, phosphates, nitrates, and sulfates, apart from their recovery by the mining of natural raw materials, can be produced without difficulty by appropriate processes or are obtained to a certain extent as byproducts (e.g., desulfurization gypsum or ammonium sulfate during flue-gas scrubbing).

In structural and civil engineering, *natural rocks* such as granite, basalt, sandstone, and marble have been replaced to a large extent by concrete and steel and are currently used mainly for decorative purposes. Natural rocks based on quartz, clay, and feldspar minerals represent by far the predominant constituent of the earth's crust, and their availability is practically unlimited.

Geochemistry shows that most *metals* occur in the earth's crust with a statistically normal distribution, if the concentration is plotted logarithmically against the quantity, as shown in simplified form in Figure 1. According to this, most metals occur in concentration ranges that are not recoverable under present conditions because of high costs. According to McKelvey, by taking into account the actual basic technical and economic conditions, deposit reserves are classified with a constant confidence level of 90 % as follows [7]:

Extent and content of the ore body are known with a limit of error of
≈10 %: proven reserves
≈20 %: probable reserves
≈30 %: possible reserves I
≈50 %: possible reserves II

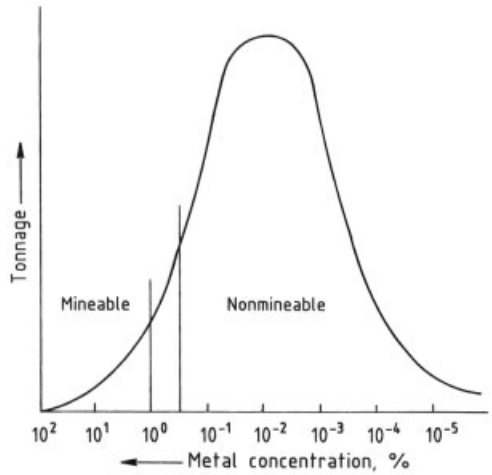

Figure 1. Distribution of metals in the earth's crust

The boundary lines drawn in Figure 2 (according to McKelvey) are not to be regarded as definite. More accurate exploration methods widen the field of known reserves; further development of mining methods and preparation processes, energy costs, and raw materials prices influence the boundary line between economic and subeconomic fields. The terms reserves and resources are thus dynamic quantities that are influenced by technoeconomic factors.

Given the reserves known today and the resources presumed with high probability, the possibility of replacing metals with plastics or ceramics, and advancing technical development, the supply of metallic raw materials is ensured for the foreseeable future.

2.6. Environmental Factors in the Processing of Raw Materials

The type of raw material, its recovery method, questions of location, the treatment process, the provision of environmental protection equipment, and other items naturally have a considerable effect on the environment. Table 2 shows the important environmental factors associated with aluminum, for example [8]. The effects of the mining and transportation of bauxite are not considered here. For comparison with the recovery of primary aluminum, the corresponding val-

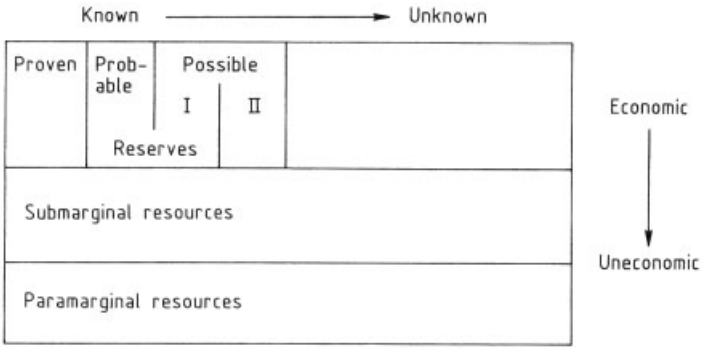

Figure 2. Classification of mineral resources (according to McKelvey)

ues for secondary aluminum production are also given.

Table 2. Factors with environmental impacts in the production of primary and secondary aluminum

Factor	Primary aluminum	Secondary aluminum
Energy Consumption, MJ/t Al		
Alumina production	27 683	
Scrap preparation		1 752
Melting down, preparation of dross, sweepings, etc.		10 060
Salt slag preparation		3 376
Fused-melt electrolysis	129 618	
Refining	4 473	
Casting	1 956	2 016
Total	163 730	17 204
Residues, kg/t Al		
Red mud	3 175	
Electrolysis residues	328	
Ash		47
Other	147	46
Total	3 650	93
Emissions, kg/t Al		
From primary energy generation	173	
Process specific	31	6
From melt energy consumption		6
Total	204	12

Supplying mineral raw materials to the treatment sites concerned and marketing the products are unavoidably combined with transportation processes. Table 3 lists the environmentally relevant factors of various means of transportation based on a study by the Schweizerisches Bundesamt für Umwelt, Wald und Landschaft

(Swiss Federal Office for Environment, Forests and Countryside) [9].

2.7. Secondary Raw Materials [10]

The demand for raw materials and energy has risen steadily with advancing industrialization and growing demand for products and goods, together with the increasing world population. As world population has increased since 1900 from ca. 1.5×10^9 to ca. 6.2×10^9 people in 1980 – that is, by a factor of about 4.1 – energy consumption has risen by a factor of 15; steel production by a factor of 20; the production of typical commodity metals zinc and copper by factors of 15 and 18, respectively; and the production of the rather novel metals nickel and aluminum by factors of 100 and 3000, respectively.

Aside from the production of consumer materials and goods from natural raw materials, the reprocessing of residues and wastes is becoming increasingly important. The return of materials into the material cycle is called recycling, and the substances which are recycled are known as secondary raw materials. The important aspects of recycling are

1) Conservation of natural resources
2) Reduction of total energy consumption
3) Reduction of the disposal site volume required

The reuse of valuable resources has a long tradition, activities being centered on metals. Examples are the remelting of noble metals such as gold and silver, the traditional processing of scrap iron and scrap nonferrous metals, and the processing of high-content metal dross, sweepings, ashes, dusts, etc. The recycling ratios of metals are therefore also distinctly higher than those of other materials, as shown in Table 4,

Table 3. Specific emissions of various means of transportation (g km^{-1} t^{-1}) and corresponding energy consumption [9]

	Rail	Truck	River boat	Sea-going vessel
Particles	0.006	0.080	0.038	0.011
CO	0.011	0.398	0.220	0.004
HC	0.069	0.199	0.055	0.035
NO_x	0.041	0.995	0.384	0.026
SO_2	0.081	0.080	0.165	0.234
Energy consumption, MJ km^{-1} t^{-1}	0.117	0.850	0.470	0.200

Table 4. Rate of recycling in the Federal Republic of Germany in 1989

	Consumption, 1000 t	From recycled materials *	
		1000 t	Percentage
Lead	374	184	49
Zinc	567	216	38
Copper	1 128	553	49
Aluminum	1 979	689	35
Steel **	39 700	17 356	44
Plastics	8 100	500	6

* Without internal recycling.
** Basis: production in 1988 including cast iron.

with the Federal Republic of Germany as an example.

Energy saving is a decisive factor, especially in the recycling of metals from scrap, since the expensive energy of reduction has already been provided and only the usually low energy melting must be expended again. To produce a marketable grade, refining or realloying may still be necessary. Figure 3 compares the energy requirements of metals from primary and secondary production.

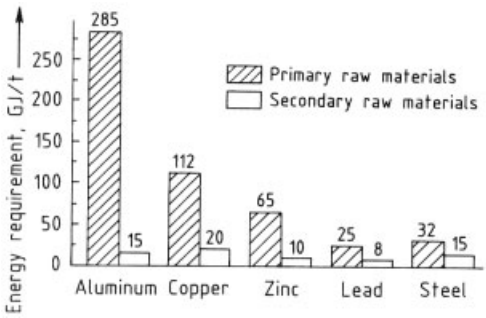

Figure 3. Energy requirements of metal production from primary and secondary raw materials

The essential goal of conventional recycling is the recovery of valuable substances from relatively high-grade residues. The overall "*conventional recycling*" system can be described as follows:

1) The value of the recovered materials or substances exceeds processing costs
2) Recycling is carried out mainly within the process or the industry concerned
3) Considerable quantities of residues of complex composition exist, since the selection of materials and the design of goods have been solely production or application orientated
4) An overall economic – ecological concept is lacking.

Especially in highly industrialized countries, extensive, complex environmental protection legislation requires – aside from the use of efficient waste gas and wastewater purification systems – the conservation of natural resources and, to the greatest possible extent, the avoidance of residues and wastes (→ Legal Aspects, Chap. 2., → Legal Aspects, Chap. 3., → Legal Aspects, Chap. 4.). This has led to the development of new technologies especially suitable for processing industrial residues. For a modern raw materials industry with integrated recycling, therefore, some additional key features apply to the so-called *new recycling*:

1) The environment is incorporated in national economic thinking as a resource to be conserved.
2) Residues are also processed whose processing costs exceed the value of the recoverable components.

3) Recycling is not limited to the industry that produces the residue but is also carried out in other or new established sectors of industry.

Examples of this kind of recycling are the processing of salt slag from the secondary aluminum industry; the processing of zinc-containing filter dust from electrosteel works; the regeneration of used foundry sand; the use of building rubble and slag in civil engineering; the processing of electronic scrap; and the dismantlement of automobiles, household appliances, and other consumer goods. An important precondition for raising the recycling ratios for raw materials from vehicles and appliances is a "recycling-friendly" choice of material and design, a task to which designers and producers are increasingly dedicating themselves.

3. Oil, Gas, and Coal

3.1. Introduction

For millions of years, oil, gas, and coal have been parts of our natural environment. Thus, special environmental protection problems are not expected in connection with the natural occurrence of these substances, nor have they occurred. Only when these materials are used by humans must environmental protection be taken into account: special environmental protection measures are necessary during exploration, production, transportation, storage, processing (→ Oil Refining, Chap. 4.) and use (→ Heating Oil, Chap. 1.6.; see also → Automotive Fuels, Chap. 8.) of these materials and the products obtained from them. Special dangers arise because these materials and their associated substances can form easily flammable mixtures with air. A series of spectacular accidents have shown that at all stages of processing – from the well to the consumer – warning and accident prevention plans must be drawn up in agreement with the authorities and emergency services. Some constituents of these substances and of the products obtained from them are considered environmentally hazardous in the widest sense, and in some cases they are classified as harmful to humans and animals (e.g., carcinogenic [11], mutagenic, or teratogenic) or as substances dangerous in water [12].

For some years, protection of the earth's atmosphere worldwide has been considered one of the greatest ecological challenges. In the scientific discussion, methane [13] present in natural gas and carbon dioxide [14] formed by combustion of oil, gas, and coal play a special part. Within the United Nations, a framework climate convention has been signed by more than 150 governments [15].

3.2. Oil

3.2.1. Exploration

Exploration is carried out by seismic reflection surveying of potential structures after promising areas have been selected by geological studies (→ Oil and Gas). In the surveying, which is carried out with heavy equipment, sometimes even in impassable terrain, the reflections of shock waves are analyzed. The shock waves are generated by vibrators, falling weights, or explosions in 10 to 20 m deep boreholes or by pressure pulses in water. In inhabited areas, the effects of noise and vibration on the population must be minimized. Cultivated areas must be recultivated and the farmers and foresters compensated.

If the result of these geophysical studies is positive, an exploratory well provides final confirmation of the presence of a reservoir; this is then explored further and developed by subsequent wells. If reservoirs must be opened up below towns, military exercise terrain, or areas that have to be protected for environmental reason, this can be done by "directional drilling" from less sensitive locations (→ Oil and Gas). Sometimes one location is sufficient. The same applies to the development of offshore fields from a platform (Fig. 4). Pollution control of seawater in the case of offshore wells in coastal waters is particularly sensitive ecologically. All wastewater, superfluous drilling mud, and drill cuttings must be collected in a barge, transported to land, and processed before the residues are then dumped.

3.2.2. Production

The legal framework governing oil production differs greatly in different countries. Whereas in

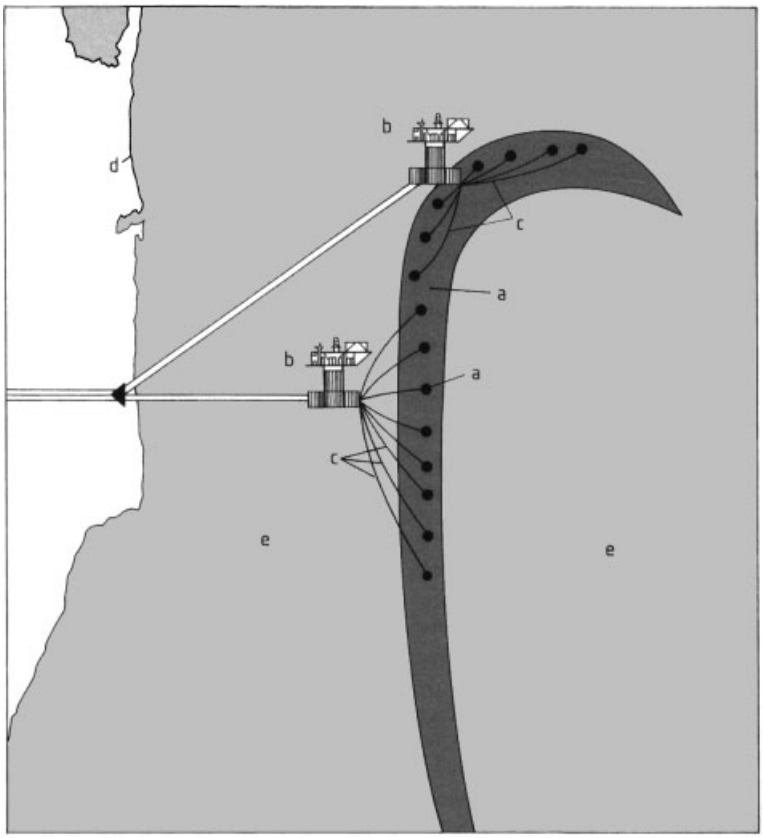

Figure 4. Directional drilling in the Schwedeneck See Oilfield in the German Baltic Sea
a) Oilfield; b) Platform; c) Directional drillings; d) Coast; e) Baltic Sea

highly developed Western countries a multiplicity of laws (on nature protection, soil protection, water pollution control, immission protection, and occupational safety, as well as mining regulations) accord high priority to ecological goals, in the countries of the Third World in the absence of legislation, responsibility for the environment falls to the development companies. The highest environmental standards apply to drilling and production platforms for offshore production, especially near the coast (Gulf of Mexico, North Sea, Baltic Sea) since particularly sensitive ecosystems can be threatened here (Wattenmeer). In particular cases, both feces and rainwater must be completely captured and disposed of on land.

On its way from the earth's surface to the oil field, the well is sunk through a series of strata and groundwater horizons. The oil and its mineral, aqueous, and associated gaseous substances must be guaranteed not to penetrate into these strata and groundwater horizons during extraction. Furthermore, the groundwater horizons themselves must not be brought by the well into contact with other horizons or strata. This is achieved by casing the hole and continuously sealing the outside by injecting cement (→ Oil and Gas, Chap. 3.3.). Production occurs through additional riser pipes in the hole.

On *abandoning production* and shutting down operation, the holes must be filled and the processing site recultivated. These measures are usually carried out on the basis of plugging plans approved by the mine inspectorate. If, at this time, oil contamination of soil or groundwater is observed in the area of the processing site, remedial action must be carried out according to a previously drawn-up plan. The improvement achieved thereby must be followed by analysis of soil and water samples, and completion of the action must be documented.

3.2.3. Transportation

Normally, a series of production locations can be found in an onshore oil field. From these production locations, oil flows in pipelines to collection points. In inhabited areas, these pipelines are laid in the ground and – for safety reasons – covered with at least 1 m of soil and protected against damage by means of concrete slabs or protective tubes if required. Steel lines are coated with plastic and provided with cathodic corrosion protection.

After being processed in central processing plants (see Section 3.2.5), the crude oil is transported further by land (usually by steel pipeline) to the port of loading and unloading or to a refinery near the consumer. Ecological questions have an increasing influence on the route, because both the pipeline and the refilled pipe ditch can affect the microclimate: e.g., the usual sand bed in the pipe ditch causes undesirable drainage of moist biotopes.

In some cases, special environmental protection measures are necessary: e.g., double-walled pipelines in sensitive areas or leading a heated and insulated crude oil line over several thousand kilometers on stilts in order to keep the crude oil liquid and pumpable but not thaw permafrost soils and thereby upset the sensitive balance of nature (Alaska).

Pipelines require regular *monitoring, testing, and maintenance*. During operation, pressures and flows are monitored at several points on the line. If the deviation of the actual value from the target value exceeds defined limits, the line is shut down and depressurized. The line is inspected regularly by patrolling on foot or by flying over with helicopters in order to ensure regular monitoring, for example, if excavation works are carried out in the vicinity of the line. The line is tested at regular fixed intervals by pressure tests and by running through with cleaning scrapers, and condition monitoring by intelligent scrapers developed especially for this purpose. Maintenance of the measurement and control equipment and of the safety and shutoff valves should be carried out according to a fixed maintenance plan.

In individual cases, for smaller fields, transportation can be made by road tankers or railroad tank cars to the refinery or, in the case of offshore fields near the coast, by lighters (double walled for environmental reasons). Sea passages over great distances are carried out in special large tanker vessels (very large crude carriers, VLCCs) between the loading ports of the producing country and the unloading piers of recipient countries. Tanker vessels have single-walled construction: after a series of serious oil accidents at sea, transportation in double-walled vessels has increasingly been demanded since the early 1990s.

3.2.4. Loading and Storage

During the *loading* of processed crude oil into road tank trucks, railroad tank cars, lighters, or oceangoing ships, as well as during the filling of storage tanks, environmental protection measures should be used to protect soil, bodies of water, and air from contaminants: filling operations must be monitored with alarm and cutoff devices; pumps have to be installed in an oil-collecting trough; and air pollution by volatile hydrocarbons must be minimized by gas displacement and gas recycling systems.

The oil can be *stored* in abandoned salt mines, salt caverns, granite caverns (Sweden), and caverns in permafrost soil or in tanks aboveground. For storage in steel tanks, a double tank bottom is increasingly required. The tanks must stand in an oiltight, oil-collecting trough or tank bund or must have a double wall. Floating roof tanks have to be provided with especially effective edge seals or be constructed as fixed roof tanks with internal floating ceilings.

3.2.5. Processing

The crude oil that emerges from the well is contaminated with minerals, water, and gas, which must be separated (→ Oil and Gas, Chap. 3.4.3.1.). The separated *aqueous constituents* with their dissolved salts are reinjected into the reservoir and in this way disposed of while preserving the environment and maintaining pressure.

Depending on local conditions, the separated *gas* is flared, reinjected into the reservoir, or – after treatment – used for energy generation. Flaring has to be deprecated for environmental reasons. Further processing of crude oil is carried out in oil refineries (→ Oil Refining), where

it is processed to motor fuels, combustibles, and chemical raw materials.

Upon combustion of the refinery products, *sulfur compounds* present in the crude oil will be oxidized to sulfur oxides (mainly SO_2). To protect the environment, these compounds are increasingly being removed from the individual intermediate products by hydrogenation and reduced to hydrogen sulfide. Hydrogen sulfide is processed in Claus plants to elemental sulfur (→ Sulfur, Chap. 6.3., → Sulfur, Chap. 6.4.), which is stored and transported to end users (usually as a liquid). For environmental protection reasons, downstream of the Claus plants, tail gas plants are connected which guarantee that 99 % of the hydrogen sulfide is converted to elemental sulfur (→ Natural Gas, Chap. 2.7.1., → Natural Gas, Chap. 2.7.2.).

Nitrogen-containing compounds present in crude oil are hydrogenated in the aforementioned process to ammonia and fed together with hydrogen sulfide to the Claus plant. Here, ammonia is converted largely to elemental nitrogen.

Organometallic compounds present in crude oil remain in the heavy fuel oil or the petroleum coke. When these fuels are burned the metals remain in the ash – as oxides, sulfates, etc. – or in the flue gas (flue-gas scrubbing is therefore necessary).

3.2.6. Uses

The main constituents of all petroleum products are hydrocarbons, which are used as motor fuels, combustibles, and chemical raw materials. These substances and the products arising from them burn to water and carbon dioxide:

$$C_nH_{2m} + xO_2 \rightarrow nCO_2 + mH_2O$$

The carbon dioxide produced during combustion is a "greenhouse gas" that is judged by some scientists and governments to contribute substantially to global warming. In 1991 a national program for CO_2 reduction was started by the government of Germany. A reduction of 25 – 30 % between 1987 and the year 2005 was fixed as a target, and a catalogue of measures was laid down. From this, a reduction of CO_2 emissions of about 300×10^6 t is expected. The effectiveness of this measure is estimated to be slight, because the building of coal-based power stations in China alone in one year counterbalances German reduction measures. An internationally coordinated policy is therefore being aimed at.

Used *engine and gear oils* were previously disposed of by burning. Today these materials are increasingly reprocessed because legislation in some countries prohibits burning and requires recycling. The reprocessed materials then come onto the market as secondary raffinates.

In the use of *motor fuels,* special emission reduction measures are required for reduction of the environmental pollution caused by road traffic).

3.2.7. Cost

In the case of environmental protection costs, a distinction generally is made between the pure investment for environmental protection and the environmental protection operating costs. A large part of the environmental protection costs results from legal requirements. These regulations require observance of "generally recognized engineering standards" or the "state of the art.". The environmental protection operating costs excluding capital costs of the refineries of the EC states for 1998 have been estimated at about 2 to 10 ECU (European Currency Units) per tonne of product, depending on the state [16]. The environmental protection operating costs of the German petroleum industry, including service of capital, have been determined for 1987 and 1989 as about € 5.6/t in each case [17].

3.3. Gas

Natural gas and petroleum belong to the same class of substances: both are hydrocarbons (→ Natural Gas) and in almost all cases occur together in nature. Exploration and production are therefore carried out in the same way as in the case of oil (see Sections 3.2.1 and 3.2.2). Special environmental protection and safety measures are necessary in case of sour gas occurrences, since the hydrogen sulfide they contain is toxic to humans, animals, and plants. For production from such wells, subsurface safety valves (SSSV), which are installed from 30 – 100 m below the earth's surface or below the seabed, are a

proven technology. In the course of production, soil subsidences may occur, which can make special measures necessary in flat country to protect bodies of water.

3.3.1. Processing

The natural gas produced is processed at the well locations. The undesirable solid constituents, reservoir water, and heavy hydrocarbons are separated and the off-gas is sweetened to remove any hydrogen sulfide and carbon dioxide and dehydrated to avoid corrosion in the process installations connected downstream (→ Natural Gas, Chap. 2.3., → Natural Gas, Chap. 2.4.).

Some gases contain small amounts of mercury, which require further processing since this substance is hazardous to health and therefore must be removed (→ Natural Gas, Chap. 2.3.2.).

3.3.2. Transportation

Natural gas is transported at high pressure in steel pipes laid below ground or in liquid form by special ships as liquefied natural gas (LNG) (→ Natural Gas, Chap. 3.5., → Natural Gas, Chap. 4.1.). In ecologically sensitive or especially valuable areas (Wattenmeer, other biotopes, rivers, and roads), the pipelines can be laid by trenchless technology or horizontal drilling. Monitoring, testing, and maintenance are carried out similarly to oil pipelines (see Section 3.2.3).

3.3.3. Storage

Large amounts of gas are stored underground. For this purpose, natural porous reservoirs and artificial salt caverns can be used (→ Natural Gas, Chap. 4.2.). No special environmental considerations apply in the operation of these reservoirs. During creation of the salt caverns by solution mining, environmentally acceptable disposal of the saline brine must be ensured.

3.3.4. Uses

Natural gas consists mainly of methane (→ Natural Gas, Chap. 1.1.). It is used as a source of energy and a chemical raw material. Both methane itself and the carbon dioxide produced during its combustion are climatically active gases, i.e., they absorb and reemit thermal radiation and so contribute to the greenhouse effect. Different calculations exist which [14] characterize the effectiveness of various greenhouse gases. The data available in the literature deviate greatly from each other, because much detailed knowledge is lacking and many effects have still been inadequately explained. Thus, for example, the lifetime in the earth's atmosphere of the gas molecules relevant to climate is subject to considerable uncertainty.

To minimize the greenhouse effect of the gases mentioned, in the case of *methane* the transport, storage, and distribution of natural gas must be as nearly loss-free as possible. *Carbon dioxide emissions* from the combustion of natural gas can be reduced by increasing the overall efficiency of heat production by technical improvements of the firing installations, combined heat and power schemes, and other energy saving measures. To hasten these savings, in various countries reduction programs have been drawn up or the introduction of taxes ("carbon tax") has been considered.

3.3.5. Cost

The costs of a carbon dioxide limitation (stabilization of emissions in the year 2010) are estimated for the United States as ca. $ 10^{12} [13].

3.4. Coal (→ Coal) [18][19][20]

3.4.1. Mining Engineering

Mining plants need large surface areas for their installation. In planning new mines, hauling shafts and coal preparation installations of existing mines are used as much as possible.

Underground mining of coal causes damage at the surface. Mining damages include subsidence, dislocated inclined planes, stretch, and wrench faults. Subsidence changes the slope of rivers and causes the formation of swamps, fissures in buildings, and thresholds in streets. Mining dam-age may be minimized by application

of underground replacement techniques (refilling the cavities with waste rocks or washery tailings).

The *open-cast mining* of coal often requires drainage of the seam and overlying strata. Since open-cast mining may reach depths of 600 m, lowering of the groundwater level in areas adjacent to the mine is also necessary to stabilize the scarp of the mine. For this purpose, wells are installed at the border of the mine and at the entrance area of the groundwater, these wells are arranged to galleries. Modern surface mining techniques demand not only full mechanization with high efficiency (excavators, etc.) but also extensive automation of the sequence of work. The current mining technique using an equipment system consisting of bucket–wheel–excavator, conveyor belt and stacker has a performance of some 100 000 m^3/d. The relation of removed soil to coal (stripping ratio) increases with depth of 700 – 1100 m from about 3 to 5 m^3/t. The removed waste material is used directly for replenishment of the coal-spent area and reclamation.

3.4.2. Processing

Coal normally cannot be mined as clean coal. Run-of-mine coal is generally associated with different amounts of mineral matter. For some purposes, especially power and heat generation, coal may be used directly as so-called ballast coal. However, since mineral matter associated with coal has a considerable negative environmental impact (see Section 3.4.3), preparative measures must be used to separate the mineral matter from the coal. Preparation of coal is thus a primary measure of environmental protection, which separates a major part of harmful components (sulfur, trace elements) from the coal at convenient cost. Preparation lowers the pyrite sulfur level of the run-of-mine coal; separation of organically bound sulfur is not possible with the usual mechanical preparation methods. Pyrite crystals are intergrown with coal to different degrees. To obtain separation, coal must be crushed to grain sizes that correspond to the grain sizes of pyrite. This becomes more difficult the finer the pyrite grains and their distribution within the coal are. An extensive separation of pyrite from coal is connected with a lowering of the amount of the trace elements lead, cadmium, copper, and zinc in coal. The concentration of the trace elements – arsenic, barium, lead, cadmium, chromium, copper, mercury, strontium, and zinc – in the preparation waste increases with increasing ash content of the preparation wastes. Traces of beryllium, cobalt, molybdenum, vanadium, and nickel are bound to the organic matter itself.

Dust emissions resulting from preparation measures are decreased by filtering air and suffocating vapors. The amount of *wastewater* that leaves the preparation installation can be kept low by intermediate cleaning and reuse of part of the water from the preparation. Wastewater from the preparation installation is usually led to clearing ponds. For the elimination of flotation wastes, the deposition of washery tailings on dumps is mainly used. By special construction of the dump, oxidations processes (occurring especially with pyrite) can be suppressed to a great extent, thereby preventing the mobilization of heavy metals.

3.4.3. Utilization

Coal is used as a raw material in coal conversion processes, which can be divided into three main groups (see Fig. 5):

1) Mechanical conversion, including coal preparation and briquetting
2) Processes to obtain secondary fuels, including coking, gasification, liquefaction, and combustion
3) Other processes for the generation of energy (e.g., absorption on active coke and active carbon) including the recovery of byproducts (e.g., benzene, phenols)

Coal is an organic rock, consisting of two essential parts: the organic and the inorganic (mineral) matter. Its components may affect the environment in a typical manner during coal utilization. According to the particular conversion process, different products are formed in different quantities.

The *mineral matter*, which is termed ash or slag, may be used in commercial products. The main portion of trace elements present in coal is bound to the mineral matter, especially the sulfidic components. Care must be taken that these

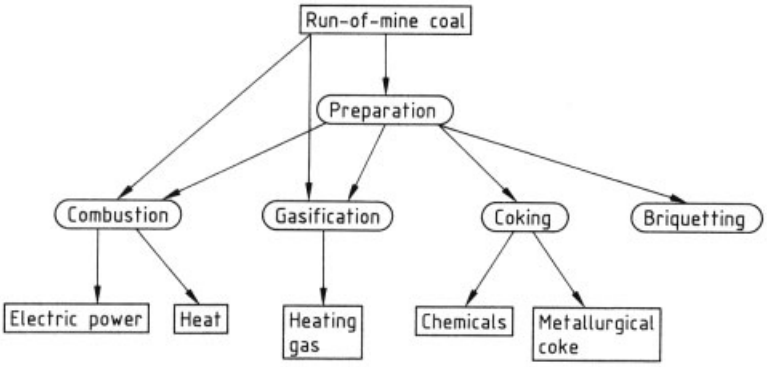

Figure 5. Utilization of coal

trace elements are not released uncontrolled into the environment.

The aforementioned environmental influences are introduced directly by the coal into the corresponding processes of utilization and valorization. Many other products that affect the environment are formed during conversion itself.

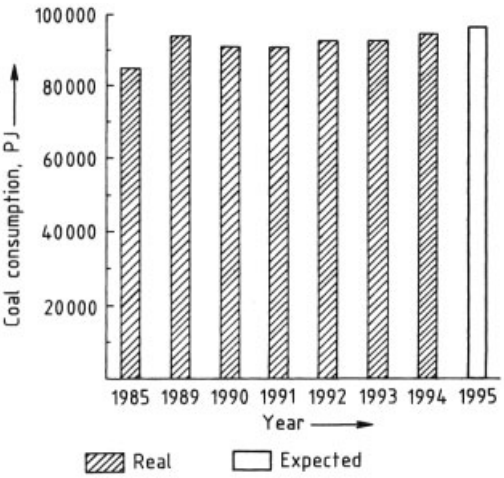

Figure 6. Development of world coal consumption

Heat and Energy Production. In the combustion processes used essentially for the production of electric power and heat, CO_2 is a climate-influencing emission, as are SO_2 and NO_x (including N_2O). In unfavorable conditions, toxic compounds may even be formed; a representative example is dioxin.

The development of world coal consumption in petajoules is shown in Figure 6. Consumption is expected to increase only slightly by about 2 % per year in the near future. That means that in the next few years, coal consumption will not exceed 100 000 PJ.

To avoid or reduce environmental problems, fundamental knowledge must be obtained for the development and optimization of technical processes. The main product formed by coal combustion for heat and electricity production is CO_2. As depicted in Figure 7, 3 t of CO_2 results from 1 t of hard coal, that is 1500 m^3 of CO_2. Worldwide coal consumption of 100 000 PJ will result in the generation of 10^{16} m^3 CO_2. Since the CO_2-free combustion of coal is not envisioned, emphasis must be placed on strategies of avoidance (i.e., economical use of energy ressources, as well as an increase in efficiency, which is demanded of engineering). By gas and steam processes, efficiencies of more than 50 % appear possible. A corresponding reduction of SO_2 and NO_x can be expected from increasing the efficiency of combustion.

The specific coal consumption for energy generation has been reduced from 1200 to 300 g/kW · h today by technical progress, and by more efficient dust precipitation, dust emission has been lowered from 800 to 50 mg per cubic meter of flue gas. At the same time the emission of detrimental matter bound to particles (trace elements, polyaromatic hydrocarbons, etc.) has decreased accordingly.

Emissions from coal combustion can be reduced by primary and secondary measures (Fig. 8). Primary measures include coal preparation, special arrangements at the dust entry (e.g., in a cyclone), improvement of the firing system, lowering of combustion temperature and

shortening of combustion time, flue-gas recycling, and eventually the use of additives, which may reduce the emission of dust, SO_2, NO_x, HCl, and HF as well. Yet all of these primary measures together are usually not sufficient, to meet the demands of environmental protection. Accordingly, secondary measures must also be involved in the processes. These are, in terms of the flue gas, mainly the precipitation of dust, but also include gas cleaning by desulfurization and DENOX processes (→ Fixed-Bed Reactors, Chap. 3.). In terms of heat, overall reduction of noxious emissions may be reached by increasing efficiency and better utilization of waste heat.

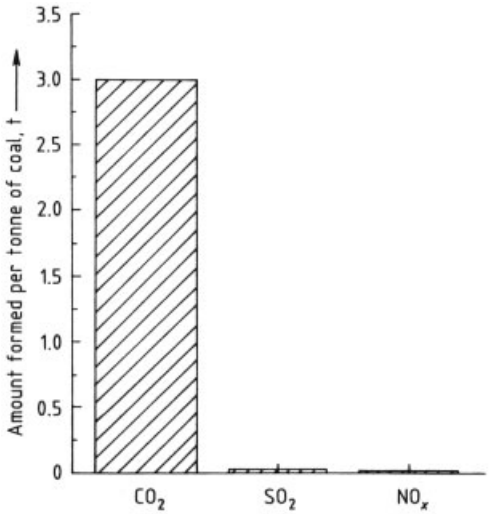

Figure 7. Amount of CO_2, SO_2, and NO_x formed during combustion of 1 t of coal

The schematic of a wet flue-gas desulfurization process, employed for some years in an electric power and heat generating plant in Germany, is shown in Figure 9. Wet washings for desulfurization are accomplished with lime or suspensions of limestone at a temperature near the dew point of the flue gas. The washing liquids are recycled. The best degrees of desulfurization are obtained with hydrate of lime; slaked lime is diluted before entering the absorber (c). The ratio of sprayed liquid and gas to be cleaned is usually between 3 and 10 L/m^3. The absorbent is added in about stoichiometric ratio, because of stoichiometric reaction with the noxious gas (molar ratio Ca: S = 1). Flue gas and washing liquid move countercurrently. The flue gas is thus mixed with finely distributed lime absorbent and sulfur dioxide is dissolved in the washing liquid before it reacts with the absorbent. The contacting time between gas and liquid is ca. 5 min. Part of the washing liquid flow is led to an oxidation stage (f), which is mostly integrated into the washery sump. The calcium sulfite formed is oxidized here to calcium sulfate. From this partial flow the utilizable end product gypsum is brought out, which is crystallized on added crystallization nuclei. Downstream of the absorber, the flue gas has to be reheated to 100°C before it is led to the chimney. Elimination degrees for sulfur dioxide of more than 85 % are obtained with this process, i.e., concentrations of <400 mg SO_2 per cubic meter in the cleaned gas.

On a technical scale, NO_x diminution in the flue gas is achieved by selective reaction of NO_x with added ammonia in the presence of a solid catalyst (DENOX process). Nitrogen oxides are reduced thereby to water and nitrogen in the presence of oxygen. The elimination of NO_x occurs at 300 – 500°C. Titanium dioxide or vanadium pentoxide can be used as catalyst (Fig. 10).

Another technique for desulfurization and denitrogenation is the so-called activated carbon process. Flue gas is led to an absorber stage where SO_2 is absorbed by activated carbon. NO_x is converted to N_2 with 80 – 90 % yield.

The success of SO_2 reduction due to the introduction of flue-gas desulfurization can be seen in Figure 11. In the Federal Republic of Germany, emission of SO_2 was reduced from about 900 000 t/a in 1983 to <100 000 t/a in 1989. The same is true for NO_x emissions, which were reduced by DENOX installations from about 500 000 t/a in 1983 to less than 100 000 t/a in 1992 (Fig. 12). If this reduction could be achieved worldwide when utilizing coal for the generation of electric power and heat, it would be an eminently important contribution to the protection of the environment. The processes are available, but introduction of these techniques often fails because of an unwillingness to invest in environment protection.

Coke Production. The second largest use of coal production after the generation of electric power and heat is in the production of coke. Cokefaction is also connected with environmental impacts, especially gaseous emissions, as

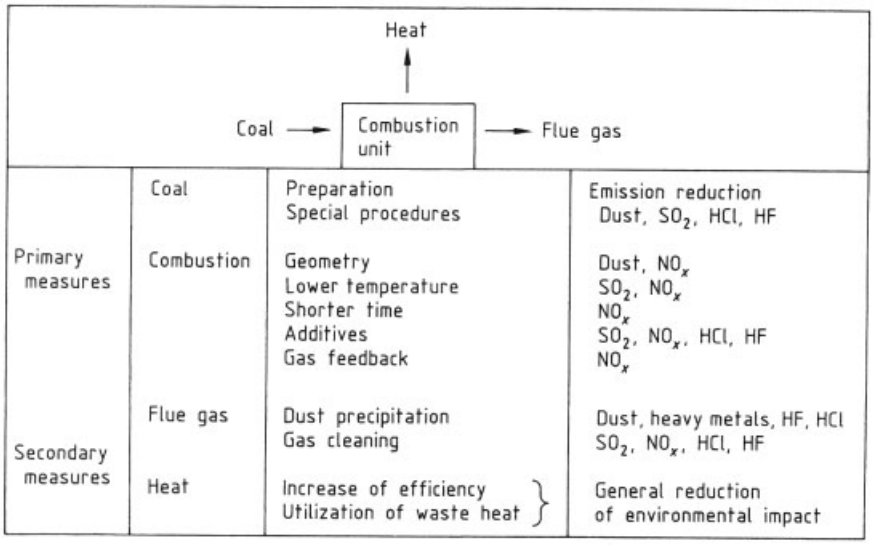

Figure 8. Measures for reduction of emissions formed during combustion of coal

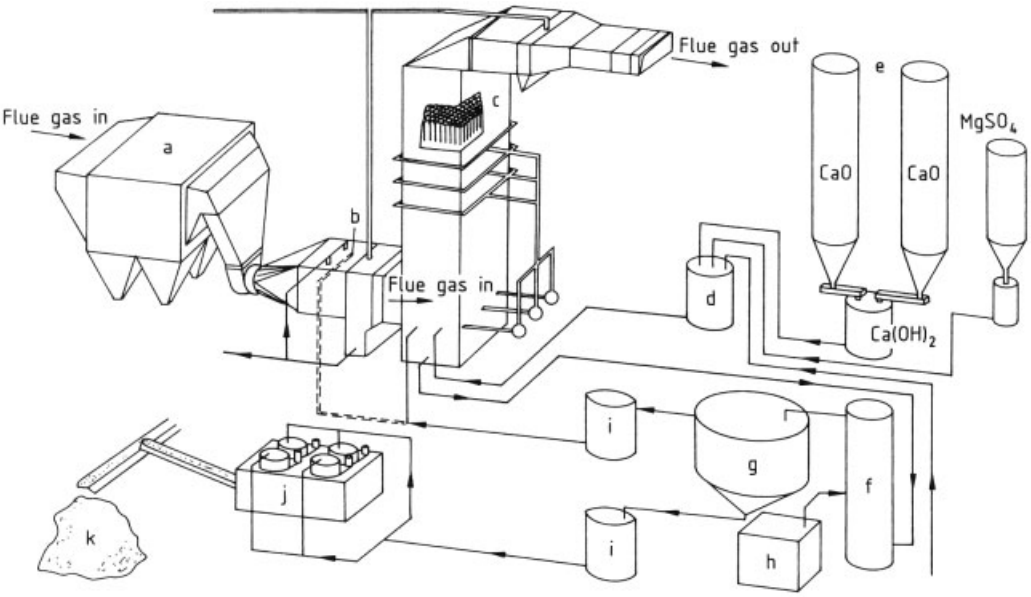

Figure 9. Schematic of a wet flue-gas desulfurization process
a) Electric precipitator; b) Prewasher; c) SO_2 absorber; d) Regenerator; e) Lime preparation station; f) Oxidator; g) Settling tank; h) Compressor; i) Gypsum buffer container; j) Gypsum dryer; k) Gypsum deposit

well as charging of wastewater. Avoiding emissions from a coking plant is especially difficult because of its large-faced construction. Since the installations stand in the open air and many sources of emissions are distributed over a large area, enclosing them is difficult. Moreover, these diffuse sources (leakages) are related to operations such as filling, pressing, quenching, classifying, and so on. The type and quantity of emissions depend on various influences that occur during cokefaction at various points. Some points and types of emission, as well as limit values and measures that minimize environmental impact, are given below.

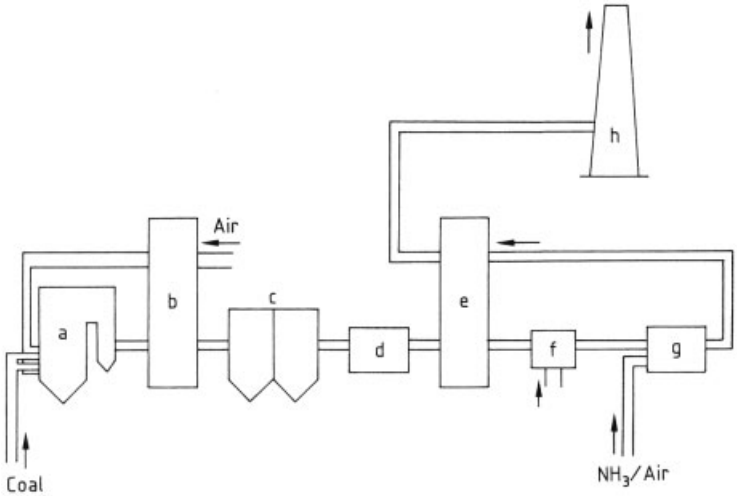

Figure 10. Schematic of a DENOX plant
a) Steam generator; b) Preheater; c) Dust precipitator; d) Flue-gas desulfurization plant; e) Heat exchanger; f) Heater; g) DENOX unit; h) Chimney

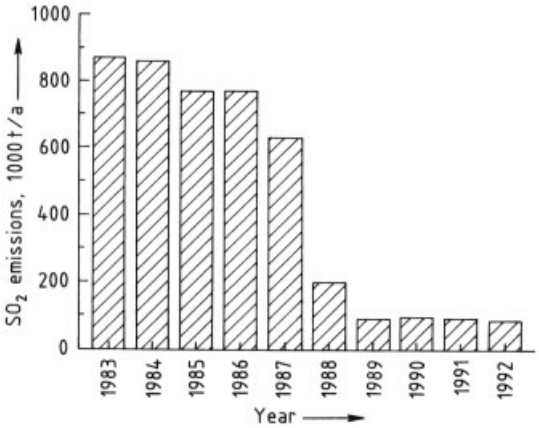

Figure 11. Reduction of SO_2 emissions in the Federal Republic of Germany

Point and type of emission	Limit values and preventive measures
Combustion gas	total sulfur $<0.8\,g/m^3$
Flue gas	NO_2 $<0.5\,g/m^3$ (5 % O_2)
Filling	dust $<25\,mg/m^3$
Cooling	dust $<20\,mg/m^3$

The wastewater of coking plants with wet quenching is alkaline (pH 8.5 – 9.5); it contains ammonia, phenols, sulfur, sulfanes, and cyanides. It must be rid of special contaminants before delivery to the purification plant. Intensively smelling substances such as mecaptans are separated from wastewater. In addition to the elimination of toxic components (phenols and cyanides), oils and tars must be separated.

For the elimination of organic compounds, biochemical processes, extractions, absorption by activated coke, and combustion of wastewater are well suited. Some modern coking plants use the dry coke cooling process in which problems with organic componds no longer occur.

Coal Valorization Processes (Gasification and Liquefaction). Particularly in the case of *gasification* (→ Gas Production, Chap. 4.), emissions may really affect the environment during sluicing of material streams, especially of residues. Environmental pollution may also originate from flue dust in the raw gas, the raw gas itself, and the process water. The same sub-

stances are found here as in combustion processes, and the techniques of diminution are essentially the same. Another problem is that these residues cannot be deposited in an environmentally protective manner without more ado, especially in the case of alkaline catalysts. A solution could be to introduce a stage of inertization before deposition, for example, thermal treatment. The developments required are similar to those for power station byproducts.

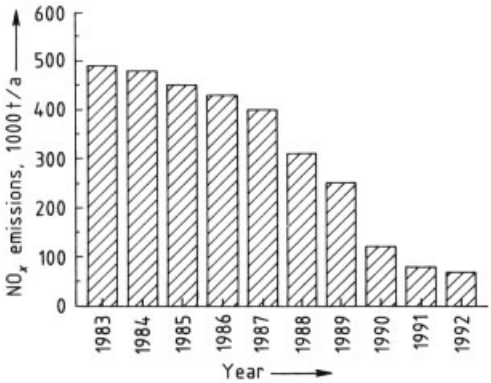

Figure 12. Reduction of NO_x emissions in the Federal Republic of Germany

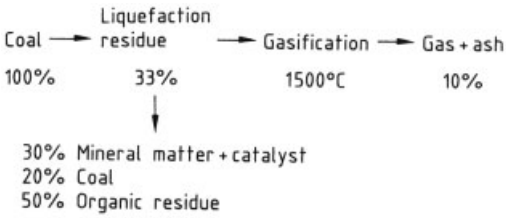

Figure 13. Reduction of residues formed during coal liquefaction by subsequent gasification of the residues

In the case of *coal liquefaction* (→ Coal Liquefaction), e.g., in liquid-phase hydrogenation, damage to the environment can be minimized and easily controlled because of the closed system. The formation of phenols – which formerly was a problem – and the environmental problems caused thereby may be avoided by appropriate operating conditions. The compounds H_2S and NH_3 contained in process water and gas can be eliminated without any problems. Only elimination of the mineral residue of hydrogenation in an environment-protecting manner remains. One may proceed here in the same way as suggested for gasification residues.

The combination of liquefaction and subsequent gasification of the liquefaction residues at 1500°C does not produce more than ca. 10 % ash, which is inert and ready for deposition (Fig. 13).

4. Renewable Resources

Renewable resources are animal and vegetable resources (biomass) that are used as energy carriers or raw materials in the widest sense. The term "renewable" means that the biomass of the agricultural ecosystem on average "survives," i.e., it is used only to the extent that it is regenerated. Agricultural ecosystems are therefore not of themselves lasting; rather, the conditions for their persistence are protection of the ecosystem and abandonment of predatory exploitation.

4.1. Significance

Renewable resources were for centuries virtually the only energy carriers and resource base for chemical products. Although in the 19th and 20th centuries they were driven back, first by coal and later by petroleum, they have never completely lost their importance for chemistry, because many of the structures and properties provided by nature are difficult, uneconomic, or practically impossible to reproduce in a chemical synthesis. Examples are the drying oils, with their importance for paint raw materials, or cellulose with its ethers and esters with inorganic and organic acids.

Renewable resources received greater attention again after the oil crises of the 1970s had made clear the dependence on petroleum and the limited nature of oil reserves. In the recent past, environmental reasons have been added as a motivation for renewed interest in renewable resources, since evidence exists that the global CO_2 balance is affected significantly by burning fossil energy carriers. In addition, this discussion also plays a part – although a considerably less important one – for chemical raw materials and their participation in the carbon cycle.

Not to be underestimated is the fact that many industrialized countries are running short

of space for waste disposal. For this reason, the possibilities of waste disposal, especially of plastics and packaging materials, via composting are being discussed increasingly. Biodegradable materials from renewable as well as fossil resources may make an important contribution here.

Finally, agricultural policy reasons exist for using renewable resources. The cultivation of plants that provide raw materials usable for energy production or internal combustion engines ("energy plants") or for the chemical industry ("industrial plants") has been publicized by agricultural interests, especially in countries with excess production, such as the EU countries. The possibility is seen here of cutting back excess production and yet keeping useful agricultural land alive.

The cultivation and use of renewable resources have been encouraged on national and European levels with the aid of subsidies and tax reliefs, and their political importance has thus been emphasized.

Within the framework of the Altener program [21,22], the proportion of regenerative energy (in which biomass is included) to total energy consumption within the European Union is to be increased from 4 to 8 % by the year 2005, for example, with EU grants.

In Germany renewable resources are encouraged on the national level via the " Fachagentur Nachwachsende Rohstoffe" ("Special Agency for Renewable Resources"). The purpose of this institution is to seek practical opportunities for increased application of renewable resources via improved coordination among the government, science, and industry.

4.2. Important Renewable Resources

Table 5 (modified according to [23]) surveys the quantitatively most important sources of renewable resources.

In the renewable resources used by the chemical industry, natural fats and oils play the most important part. Table 6 shows the estimated consumption of various groups of resources by the chemical industry [24].

4.3. Industrial Use

Industrial application relates above all to use in the chemicotechnical field, where renewable resources are used for the production of intermediate and end products. The most important product lines from renewable resources are based on fats, oils, starch, sugar, and cellulose.

4.3.1. Natural Fats and Oils

World production of natural fats and oils has increased steadily in the last decades and in 1990 reached ca. 80×10^6 t [25,26]. Vegetable oils, at 60.7×10^6 t, make up the main part, but animal fats (20×10^6 t) also occupy an important place in oleochemistry, especially in the utilization of inedible tallow and fat. In all, probably ca. 14 % of world fat production is used chemically [24, 26], ca. 80 % for human food, and the remaining 6 % as feed material. Castor oil is an exception: because of its unpalatability and special structure it is used exclusively for industrial purposes, mainly for paint raw materials and plastics (e.g., polyamide 11). The 1990 production figures for the most important oils and fats (in 10^6 t) [26] are listed below; for further information also regarding composition, see → Fats and Fatty Acids.

Soybean oil	16.5
Palm oil	10.8
Rapeseed oil (erucic acid free)	8.1
Sunflower oil	8.0
Cottonseed oil	3.8
Peanut oil	3.7
Coconut oil	2.8
Olive oil	1.8
Palmkernel oil	1.3
Linseed oil	0.7
Castor oil	0.4
Others (corn, sesame, etc.)	2.8
Vegetable oils, total	60.7
Tallow, lard	12.3
Butter	6.3
Fish oil	1.4
Animal fats, total	20.0
Total	80.7

Table 7 (after [24]) surveys the most important fields of application of oleochemical products.

Table 5. Important renewable resources

Cellulosecontaining resources	Starchcontaining resources	Sugarcontaining resources	Oilcontaining resources
Wood	potato	sugarcane	rape
Straw	corn (maize)	sugar beet	flax
Hulls/shells	grain	fodder beet, etc.	soya
Bagasse *	cassava	sweet sorghum	sunflower
Cotton		molasses	oil palm
			coco palm
			castor-oil plant
			peanut
			cotton

* Wastes from cane sugar production.

Table 6. Present use of natural raw materials (in 1000 t) by the chemical industry, estimated

Raw material	Germany (excluding former German Democratic Republic)	EU	World
Oil and fat	700	2700	9500
Cellulose	220	600	5000
Starch	115	390	1750
Sugar	15	65	800

Table 7. Most important fields of application of oleochemical products

Product	Application
Fatty acids and derivatives	metallic soaps, detergents, soaps, cosmetics, alkyd resins, paints; textile, leather, and paper industries; rubber, lubricants
Fatty acid methyl esters	cosmetics, detergents
Glycerol and derivatives	cosmetics, toothpaste, pharmaceuticals, foods, paints, plastics, synthetic resins, tobacco, explosives, cellulose processing
Fatty alcohols and derivatives	detergents, cosmetics; textile, leather, and paper industries; duplicator stencils, petroleum additives
Fatty amines and derivatives	fabric softeners, mining, road building, biocides, textile and fiber industries, petroleum additives
Drying oils	paints, varnish, linoleum
Castor oil, ricinoleic acid	polyamide 11, alkyd resins

Alkyd resins (→ Alkyd Resins) are produced on a large scale from natural triglycerides and the fatty acids made from them. Soybean oil is the most significant in this connection as the most important (semi-) drying oil. Linseed oil, which was previously dominant in the production of alkyd resins, is currently used mainly in the manufacture of linoleum.

The most important conversion reactions and products are shown in Figure 14 (according to [26]).

Surfactants. After soaps and petrochemically produced linear alkylbenzenesulfonates, fatty alcohol ethoxylates are the most important surfactants, as measured by consumption in Western industrialized countries (→ Surfactants, Chap. 14.) [27]. Fatty alcohol ethoxylates figured in public debate, being raw materials for fatty alcohol ether sulfates. Fatty alcohol ether sulfates play a large part in cosmetic formulations. Under the acidic conditions of their production by sulfation, traces of dioxane (a suspected carcinogen) can appear as a degradation product of the polyether chain (→ Surfactants). Today, process technology measures enable the dioxane content to be lowered to less than 10 ppm [28].

For nonionic surfactants, greater consideration of exclusively natural structural elements can be expected in the future. Here, the new alkyl polyglucosides, in particular, are interesting for cosmetic applications because of their mucous membrane compatibility, low toxicity, and easy biodegradability [26, 29] (→ Surfactants).

The potential for renewable resources in the surfactant field is increased further by the combination of oleochemical and carbohydrate structural units. A survey of existing carbohydrate surfactants, their significance, and their properties is given in [29]. For a short description of fatty acid esters of sorbitol and sucrose, see → Surfactants, → Surfactants.

Rapeseed oil has recently acquired practical importance as a *noncirculating lubricant*. Mineral-oil-based lubricants from motor saws lead to contamination of forest soil, which can be avoided by the introduction of rapeseed oil

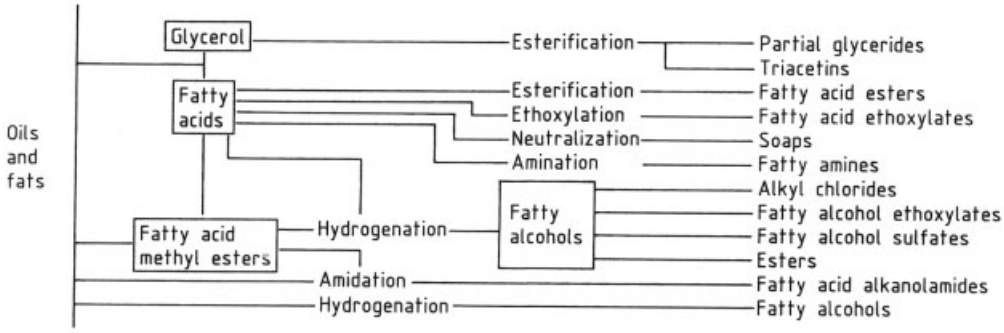

Figure 14. Most important conversion reactions of natural oils and fats and products formed

products because of their rapid and complete biodegradability [30].

4.3.2. Cellulose and Its Derivatives

4.3.2.1. Cellulose

The chemical industry worldwide uses ca. 5.1×10^6 t of cellulose (dissolving pulp; EEC: 0.6×10^6 t, Germany: $>0.2\times10^6$ t). This constitutes about 3.5 % of the total cellulose production of ca. 150×10^6 t (1990 – 1991), of which the vast majority – more than 98 % – goes into "nonfood use" [31].

Cellulose is the most important product of chemical exploitation of wood ("chemical pulp"); between 30 and 50 % of hardwoods and softwoods consist of cellulose. A particular problem of this wood exploitation is association of the cellulose with lignin and hemicelluloses, each of which makes up 20 – 25 % of the wood.

As a result of the process for producing cellulose from wood, particularly the customary use of sulfur-containing compounds in the sulfite and the kraft (sulfate) processes, lignin and the hemicelluloses must be burned and thus can be utilized only for generation of process heat. A detailed survey of various pulping processes is given in [32] (→ Paper and Pulp, Chap. 1.3.).

Most of the world chemical pulp production (nearly 90 %) is produced by kraft (sulfate) pulping, in which lignin and hemicelluloses are dissolved out of the wood with an aqueous mixture of sodium sulfide, sodium carbonate, and sodium hydroxide solution (→ Paper and Pulp, Chap. 1.3.1.) [33]. The complicated bleaching with chlorine – hypochlorite or chlorine dioxide, which is required because of the dark color of the sulfate (kraft) cellulose, can lead to considerable environmental pollution because a number of organic, nondegradable, and unusable chlorine compounds are formed (→ Paper and Pulp, Chap. 1.6.1., → Paper and Pulp, Chap. 1.6.2.). Even though lower wastewater loads are definitely achieved with chlorine dioxide, reservations remain that – aside from the uneconomical size of the operating units – have led to shutdown of the kraft pulping process (e.g., in Germany).

The *sulfite process* [34], which is stagnating worldwide, is less polluting and is used on a large scale in Germany (→ Paper and Pulp, Chap. 1.3.2.). A few technical drawbacks must be accepted, however, particularly poorer mechanical strength of the (paper) cellulose so prepared; besides, only low-resin woods can be used as raw materials.

In a few processes, an attempt is made to avoid the problems associated with bleaching; examples are the ASAM process as a modification of the sulfite process (→ Paper and Pulp), and the Organocell (→ Paper and Pulp) and Acetosolv processes, which both work with organic lignin solvents, the Organocell process requiring ClO_2 [35].

Regenerated cellulose (cellophane), as obtained via alkali cellulose and the xanthate (produced from alkali cellulose by reaction with carbon disulfide), is easily biodegradable [36, 37] and permitted under food laws. Nevertheless, this product is declining as a film material since it is increasingly being replaced by polypropylene which is more resistant to mechanical stress and is considerably cheaper.

Detailed information on fibers from regenerated cellulose (viscose fibers) is given in → Cellulose, Chap. 3.1.

Regenerated cellulose, which is made by the copper oxide – ammonia (cuoxam) process in declining production volumes, is an important membrane material for dialyzers ("artificial kidneys") (→ Cellulose, Chap. 3.3.). Copper from the cuoxam process is almost completely recovered with ion exchangers. From an ecological viewpoint, the viscose process is inferior to the cuoxam process. The recovery rate for carbon disulfide used in the viscose process is ≤70 % (→ Cellulose, Chap. 3.4., → Cellulose, Chap. 3.1.4.).

4.3.2.2. Cellulose Esters

Inorganic Cellulose Esters. Cellulose nitrate, the oldest inorganic cellulose ester (nitrocellulose), has great importance as an explosive and a paint raw material, whereas the mixture with camphor is produced still as thermoplastic (celluloid) to a minor extent (e.g., for combs, table tennis balls, spectacle frames). The world capacity for cellulose nitrate amounts to ca. 165 000 t, with declining utilization.

Lacquers made from cellulose nitrate are excellent for open-pore wood (furniture) coatings, but usually contain considerable amounts of organic solvents [38].

For a description of safety aspects in the production, storage, and transport of cellulose nitrate, see → Cellulose Esters, Chap. 1.2.8. The potential hazard of cellulose nitrate depends on the degree of nitration and on the wetting with, e.g., ethanol or *n*-butanol (at least ca. 20 wt %). Since cellulose nitrate with high nitrogen content is explosive in the dry state, it may only be transported moistened.

Organic Esters. Of the cellulose esters of organic acids, acetic, propionic, and butyric acid esters, as well as the corresponding mixed esters (so-called acetopropionates and acetobutyrates), are especially important technically. The mixed ester of acetic and phthalic acid (monophthalate, from partially hydrolyzed cellulose triacetate and phthalic anhydride) is used to a small extent, especially in galenicals as tablet coatings for regulating drug absorption (dissolving first in the alkaline intestinal medium). For a short description of industrial uses of mixed esters, see → Cellulose Esters, Chap. 2.2.5.

For the production of moldings, thermoplastic cellulose esters of organic acids play a considerably greater part than cellulose nitrate, whose importance has diminished because of its flammability. Cellulose acetate and acetopropionate are important for spectacles (lenses and frames), sunglasses, and sports goggles, see also → Cellulose Esters, Chap. 2.4.9. [39].

During the debate on the biodegradability of plastics, especially packaging materials, thermoplastic cellulose esters have been studied in more detail. Cellulose triacetate proved to be not or not sufficiently rapidly biodegradable [40, 41]. However, the compostability of cellulose esters improves with decreasing degree of substitution. Cellulose-2,5-acetate ("secondary acetate") is apparently biodegraded slowly [41]. Cellulose diacetate, on the other hand, is designated compostable and is available as a commercial product, optionally also for food packaging [42], mixed with, e.g., citrate ester plasticizers that support degradation [43, 44].

Cellulose acetate butyrate (CAB) continues to be important in automobile metallic paints. As a result of the rapid viscosity increase of CAB-containing paints upon evaporation of organic solvents, and despite the very dilute solutions necessary to achieve planar orientation of the aluminum bronze, flawless processing is possible. Because of the high solvent emissions of such paint systems, considerable efforts are being made to replace them by waterborne metallic paints [45].

4.3.2.3. Cellulose Ethers

Cellulose ethers play an important part industrially. In 1987, world consumption – without the state trading countries – was 230 000 t and the EC share 116 000 t [46, 47], with a tendency to rise (for more detailed data see → Cellulose Ethers, Chap. 10.). The most important ethers are carboxymethyl cellulose (or its sodium salt) and methyl cellulose, which together make up more than three-fourths of the world consumption. The hydrophilic character of the cellulose ethers (hydrophilicity depending on the degree of substitution) leads to water-soluble products that are used as thickeners, water absorbers, and adhesives in the building industry; in surface

coatings (including methyl cellulose as a wallpaper adhesive); as an auxiliary and additive in detergents (largest use [47]); in foods, cosmetics, and pharmaceuticals; as paper and textile auxiliaries, in petroleum production; etc. Hydroxyethyl cellulose also, as the third largest product of this group, has similar fields of application but is not permitted as a food additive (→ Cellulose Ethers, Chap. 5.1.2.).

4.3.3. Starch

The mixture of amylose (essentially straight chain) and amylopectin (branched), occurring naturally as "starch," has a fluctuating percentage composition depending on its raw material source. The amylopectin fraction usually predominates (ca. 3/4) (→ Starch, Chap. 1.).

Table 8. Uses of starch in different industries

Branch of industry	Products	Starch content ∗
Paper manufacture	wrapping paper	3.0 – 5.0 %
	newsprint	1.5 – 2.0 %
	graphic paper	1.5 – 2.5 %
Paper processing	corrugated cardboard	10 – 14 g/m^2
	laminated paper	12 – 25 g/m^2
Building materials, etc.	gypsum plasterboard	5 – 7.0 %
	mineral fiberboard	3 – 6.0 %
	granulated coal	2 – 5.0 %
Chemistry, cosmetics, pharmaceuticals	washing starch	15 – 25 %
	washing powder	3 – 5 %
	washing raw materials	30 – 45 %
	toothpaste	6 – 70 %
	tablets, pills	0.5 – 3.0 %

∗ Based on dry matter.

Exceptions with a predominant amylose fraction are, e.g., starches from peas and high-amylose corn, grown especially for its high amylose content [48]. Since time immemorial, starch has been used outside the food field (e.g., as a paper and textile auxiliary). In total, ca. 500 000 t of starch are consumed annually in the industrial and technical fields in Germany (excluding the former German Democratic Republic) and more than 2×10^6 t in the EU. Its use in the chemical industry as a chemical raw material in the narrower sense, is considerably less [49], about 600 000 t in the EU [50]. Important applications are as an additive for building materials (e.g., in gypsum wallboards), cosmetics, and detergents, as well as an auxiliary for dyes and pharmaceuticals (→ Starch, Chap. 1.7.).

Table 8 shows the content and applications of starch in selected products in the chemicotechnical field [51].

Starch has recently attracted great interest as a blend raw material for thermoplastics, especially for packing materials, refuse sacks, or mulch films that are claimed to be biodegradable. Such degradation cannot occur in disposal sites, so only composting in appropriate systems is possible [52].

In the composting of blends of polymers customary in the packaging sector with starch, a macroscopically rapidly advancing degradation is initially achieved, based on microbial degradation of the starch contained in the blend. The thermoplastic component of the blend (e.g., polyethylene) is not completely degraded chemically despite the complete disappearance of the film or molding (→ Plastics, Recycling, Chap. 6.2.). This is clear from early product examples with degradation curves in which the weight loss is followed and only an asymptotic approach to a maximum value (= starch content) is visible [53, 54]. Thus, a final evaluation of the biological compostability depends also on the second blend component [e.g., polyethylene, ethylene – acrylic acid copolymer [55, 56], partially hydrolyzed poly(vinyl acetate) [57–59]. Increasing degradation of synthetic polymer blend components is aimed at with prooxidants [60]. Composting conditions are decisive for the chronological sequence of degradation. To overcome drawbacks in biodegradability a more recent development uses poly(ε-caprolactone) – instead of, e.g., partially hydrolyzed poly(vinyl acetate) or poly(vinyl alcohol) copolymers – in combination with starch. These blends are totally compostable and have a reduced thermomechanical stability.

4.3.4. Low-Molecular Carbohydrates – Sugars

Products containing both carbohydrate structures and oleochemical building blocks are mentioned in Section 4.3.1.

Mono- and disaccharides, as well as simple derivatives, are produced in consider-

able amounts. The world production of low-molecular carbohydrates in 1991 – 1992 (in 1000 t/a) is listed below [61]:

D-Sucrose	113 614
D-Glucose	5 000
D-Sorbitol	650
D-Lactose	180
D-Fructose	50
D-Gluconic acid	40
L-Sorbose	25
D-Isomaltulose	20
D-Mannitol	8
D-Maltose	3
Xylitol	3
D-Xylose	1
Dianhydrosorbitol	1

These products are used predominantly in the food sector. Worldwide more than 800 000 t of them go into the non-food area (1991 figures), that is, less than 1 % (EC: 153 000 t; Germany in the chemical industry sector ca. 29 000 t).

In the chemical industry these carbohydrates are used mainly in the pharmaceutical field, as building blocks in polyurethane raw materials (sugar polyethers), and to a considerable extent as a carbon source in biotechnological processes (e.g., for the production of citric acid). To some extent, waste materials, such as molasses and bagasse, are also important here (→ Sugar, Chap. 22.2.). Molasses is used to a large extent, e.g., as raw material for the production of citric acid via fermentation by *Aspergillus niger* [62]. The fermentation residues (vinasse), with only a low remaining sugar content, can be used in agriculture as a cattle feed additive or fertilizer, see, e.g., [63] (→ Sugar, Chap. 22.3.).

4.3.5. Other Renewable Resources

4.3.5.1. Polyisoprene

Natural rubber is described in detail in → Rubber, 2. Natural. Aside from natural rubber, which consists essentially of *cis*-1,4-polyisoprene, its *trans*-isomer, gutta percha, is also used industrially – however, to a considerably lower extent [64].

4.3.5.2. Natural Resins

Natural resins are described in → Resins, Natural. For the most important toxicological and legal aspects, see → Resins, Synthetic, Chap. 4. Although the economic importance of natural resins has declined, almost 1×10^6 t/a of rosin is still produced, chiefly in the United States and the People's Republic of China [65,66] (→ Resins, Natural, Chap. 5.6.).

Important applications are paint and varnish resins and printing inks.

4.3.5.3. Plant Polysaccharides ("Gums")

The polysaccharides that can be used with water for the production of highly viscous gels are grouped under the name gums (see, e.g., [67]).

Some of the most used gums in approximate order of their quantitative importance are guar gum from the seed of the guar plant (→ Polysaccharides); gum arabic (→ Polysaccharides, Chap. 7.), an acacia tree resin; locust bean gum (carob gum) (→ Polysaccharides) from the seed of the carob tree; alginates and carrageenans (→ Polysaccharides, Chap. 4., → Polysaccharides, Chap. 5.) from seaweed, etc.

Aside from their importance in the food industry as emulsifiers, stabilizers, thickeners, etc., gums have been widely used industrially: guar gum (a branched mannose – galactose polysaccharide) in mining, in the textile industry, and above all in the paper industry; gum arabic (alkali – alkaline-earth salts of complex branched polysaccharides) for writing inks, adhesives, and textiles; locust bean gum (branched galactomannan) for textile sizes and dyes, and for the paper industry. Alginates (essentially salts of polysaccharides from guluronic and mannuronic acid units) are used chiefly for foods, pharmaceuticals, dental molding compounds, etc.; and carrageenans (galactans partially esterified with sulfuric acid) are used essentially for foods [67].

4.3.5.4. Active Substances Based on Renewable Resources

The active substances based on renewable resources include the antibiotics produced by fermentation (→ Antibiotics). The most important fermentation raw material is sugar, at ca. 50 000 t/a (EC 1990).

Some vitamins form a second important group of pharmaceutical preparations for which renewable resources and their biotransformation play a part. The most important are Vitamin C (ascorbic acid, worldwide ca. 36 000 t from 60 000 t glucose or sorbitol), Vitamin B_2 (riboflavin, worldwide ca. 2000 t/a), and in considerably smaller amounts Vitamin B_{12}. An important raw material source, among others, is beet molasses.

Finally, very large amounts of L-glutamine (ca. 400 000 t/a worldwide, EU ca. 31 000 t/a) and L-lysine (ca. 40 000 t/a worldwide, EU ca. 10 000 t/a) are produced from agricultural raw materials for use in foods, pharmaceuticals, and feeds.

So-called natural product insecticides play a growing part in public debate. However, they were recommended and used early. Examples are nicotine (from tobacco leaves), pyrethrum (from chrysanthemum flowers), and other plant extracts, in some cases used for centuries (→ Insect Control, Chap. 3.).

However, unmodified natural substances cannot usually meet the criteria for modern plant protection agents – uniform and controllable action, toxicological acceptability, practicability, and economy. For example, the toxicity of nicotine to warm-blooded animals and the poor stability of pyrethrum make them unable to meet present-day standards. As a rule, therefore, ideas for the use of renewable resources for plant protection begin with modification of the natural model – example: pyrethroids.

The insect repellent azadirachtin occurring in the oil of the neem tree seed is attracting increasing interest. The oil has been used in India from time immemorial for insect control [68].

4.3.5.5. Flavors and Fragrances

Flavors and fragrances, such as those employed in foods, spices, drinks, medicine, toilet articles, cosmetics, and perfume, are often based on natural "renewable resources." In the Federal Republic of Germany, of the 40 000 t of flavors and fragrances produced in 1989 ca. one-third were of natural origin. Ca. two-third were produced synthetically, most of them to natureidentical substances. Fully synthetic aroma substances that do not correspond to natural products are the exception. Aroma substances of both animal and vegetable origin are employed, yet the products of animal origin, e.g., in perfumery (ambergris, musk, civet, etc.), are ever more costly and less used because of their increasing scarcity. A comprehensive survey is given in → Flavors and Fragrances.

4.4. Energy Use

The energy use of renewable resources comprises the use of biogas from industrial and agricultural wastes (e.g., for feeding into the public gas grid or for electricity generation); the cultivation of biomass for gasification, fermentation, or direct combustion; and the use of biofuels.

4.4.1. Biofuels

Biodiesel. The use of natural oils such as rapeseed oil or of rapeseed oil fatty acid methyl ester as a fuel for motor vehicles is increasingly discussed. While conventional diesel engines must be converted to use rapeseed oil, rapeseed oil methyl ester ("biodiesel"; produced from rapeseed oil by transesterification with methanol) is suitable directly for diesel engines. Although the ecological and economic evaluation of biodiesel is not yet complete [69, 70], a few aspects can be stressed:

In Germany alone, the area cultivated for rape is about 1×10^6 ha. Of this, up to now, only 42 000 ha (1991) has been used for the non-food area; yet by 1992, ca. 70 000 ha had been contracted for this purpose [70]. The potential is considered to be a maximum of ca. 400 000 ha [69]. Replacement of about 5 %, or in the medium term not more than 10 %, of diesel fuel is therefore achievable [70]. Environmental advantages then arise from the sulfur-free product and – even on allowing for emission of other greenhouse gases (N_2O, etc.) from the cultivation of the rape – from the reduction of CO_2 emissions due to the virtually closed carbon cycle. If 1 kg of diesel fuel is replaced by the required amount of 1.06 – 1.08 kg [70] rapeseed oil methyl ester, from 0.6 – 2.2 kg [69] to 2.3 – 2.8 kg [70] less CO_2 (including all other climatically active gases) is emitted. In addition savings in crude petroleum use per kilogram diesel fuel

of 0.76 – 0.91 kg [70] or 0.55 – 0.94 kg [69] occur.

Ethanol. Molasses as a waste material from cane sugar production is important for the production of ethanol as a motor vehicle fuel ("gasohol"). During fermentation, enormous amounts of slops are formed and must be coped with. An overall estimation of emission reductions as a result of ethanol addition is disputed [71, 72]; while the CO emissions are reduced, NO_x values in the exhaust gas increase [73]. Economically, the high production costs make this alternative fuel, even at only ca. 20 % addition, appear uncompetitive ("proalcool" program in Brazil) (→ Ethanol, → Automotive Fuels, Chap. 8.). For efforts in this regard in the United States, see, e.g., [74]. Bagasse (sugarcane residue) is either burned [e.g., for heating purposes (→ Sugar, Chap. 20.5.3.)] or used for the production of high-quality paper or cardboard. For a description of the extremely important fermentative ethanol synthesis, which clearly predominates over the synthesis from ethylene, see → Ethanol, Chap. 5.

4.4.2. Biogas

One possible means of waste utilization is the production of methane (biogas) from organic materials in sewage waters. This anaerobic enzymatic conversion (biomethanation) leads by oxidative and reductive processes to a mixture of ca. 60 mol % methane, ca. 40 vol % carbon dioxide, and small proportions of other gases. Biomethanation can also be carried out with " energy plants" as raw materials. The analogous conversion of organic landfill constituents likewise provides methane – CO_2 (landfill gas). A detailed description, including economic and ecological aspects, can be found in → Methane; → Gas Production, Chap. 1.1.4.

4.4.3. Heat from Biomass

Biomass firing in heating plants (e.g., with waste wood; whole-plant combustion of Miscanthus reeds, for example) plays an increasing role in projects and small decentralized plants [75]. Its profitability is not yet confirmed. However, of the energy sources bioethanol, biodiesel, and whole-plant combustion, the last has the smallest subsidy requirement at the present time [76].

4.5. Appraisal

4.5.1. Ecobalancing of Renewable Resources

Ecobalances seek to follow the life cycle of a product and to register the environmental influences associated with it [77]. Ecobalances in the field of production of renewable resources have already been drawn up [78, 79]. In the course of this, specific balancing problems occur with renewable resources compared to commercial processes. These concern especially the feedstock concept, the description of the emission situation, the land requirement and alternative land use, the dependence of the "production process" on conditions specific to the location (soil, weather, seasons), the description of alternative paths for waste use, and emissions into the soil.

4.5.2. Opportunities and Limits of Use of Renewable Resources

Renewable resoures have opportunities especially in those areas where the product of natural synthesis can be used directly or after simple processing and a lasting, economically advantageous supply of these raw materials is guaranteed. Conversely, renewable resources have their limits in fields of use where the aforementioned conditions are not met.

Thus, the use of renewable resources is reasonable when technical, economic, and ecological advantages over other materials exist.

5. References

General References

1. Metallgesellschaft Frankfurt: *Metallstatistik* **80** (1993).
2. *Statistisches Jahrbuch der Stahlindustrie 1993,* Verlag Stahleisen mbH, Düsseldorf, 1993.
3. *Metal Bulletin Handbook,* The Metal Bulletin Ltd., London 1993.
4. G. J. S. Govett, M. M. Govett: *World Mineral Supplies-Assessment and Prospectives,* Elsevier, Amsterdam 1976.

5. V. M. Goldschmidt: "Die geochemischen Verteilungsgesetze der Elemente," *Skr. Nor. Vidensk. Akad. Kl. 1 Mat. Naturvidensk. Kl.* 1923.
6. H. Kirsch: *Technische Mineralogie,* Vogel Verlag, Würzburg 1965.

Specific References

7. V. E. McKelvey: *The Mineral Potential of the United States 1975 – 2000,* Univ. Wisconsin Press, Madison 1973.
8. K. Krone et al.; "Ökologische Aspekte der Primär- und Sekundäraluminiumerzeugung in der Bundesrepublik Deutschland," *Metall (Berlin)* **44** (1990) no. 6.
9. Bundesamt für Umwelt, Wald und Landschaft (BUWAL): Schriftenreihe Umwelt Nr. 132, Ökobilanzen von Packstoffen Stand 1990, Bern, Feb. 1991.
10. H. Maczek, W. Massion: "Umwelt, Energie und Recycling," *Erzmetall* **44** (1991) no. 12.
11. World Health Organization (WHO) International Agency for Research on Cancer (IARC); *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Occupational Exposures in Petroleum Refining; Crude Oil and Major Petroleum Fuels,* vol. 45, Lyon 1989.
12. Council Directive of December 17, 1979 on the protection of groundwater against pollution caused by certain dangerous substances (80/68/EEC) (List I, no. 7), O.J. no. L 20/43, 26.1.1980.
13. International Petroleum Industry Environmental Conservation Association (IPIECA): *Global Climate Change,* London 1991.
14. DGMK (German Society for Petroleum and Coal Sciences and Technology): Project 448-2 "Approaches and Potentials for Reducing Greenhouse Effects from Fossil Fuels," Hamburg 1992.
15. Conference of the United Nations on Environment and Development (UNCED), Rio de Janeiro 1992.
16. Commission of the European Communities: "Study on the Cost Born by the Refining Industry in the Member States in order to Comply with Environmental Legislation," prepared by Chem. Systems International Ltd., London 1986.
17. Estimation of Mineralölwirtschaftsverband e.V., Hamburg (Petroleum Industry Association in the Federal Republic of Germany, Hamburg).
18. J. F. Unsworth, D. J. Barratt, P. T. Roberts, *Coal Quality and Combustion Performance* in: Coal Science and Techology, vol. 19, Elsevier, Amsterdam 1991.
19. D. van Velzen (ed.) Sulfur Dioxide and Nitrogen Oxides in Industrial Waste Gases: *Emission, Legislation and Abatement* in: Chemical and Environmental Science, vol. 3, Kluwer Academic Publishers, Dordrecht 1991.
20. F. H. Friedich, K. J. Guntermann, M. J. Paersch "Kohle und Umwelt," in: Bergbau, Rohstoffe, Energie, vol. 26, Verlag Glückauf GmbH, Essen 1989.
21. EEC, O.J. no. L 235/41, Sep. 18, 1993.
22. *Ökologische Briefe* **34** (1992) 14 – 15.
23. D. Osteroth: *Von der Kohle zur Biomasse,* Springer Verlag, Berlin 1989.
24. H. Baumann et al., *Angew. Chem.* **100** (1988) 41.
25. H. J. Richtler, J. Knaut, *Fat. Sci. Technol.* **93** (1991) 1.
26. H. Eierdanz in M. Eggersdorfer, S. Warwel, G. Wulff (eds.): *Nachwachsende Rohstoffe – Perspektiven für die Chemie,* VCH Verlagsgesellschaft, Weinheim 1993, p. 23.
27. H. J. Richtler, J. Knaut, *Seifen Öle Fette Wachse* **117** (1991) 545.
28. H. Baumann, M. Biermann in M. Eggersdorfer, S. Warwel, G. Wulff (eds.): *Nachwachsende Rohstoffe – Perspektiven für die Chemie,* VCH Verlagsgesellschaft, Weinheim 1993, p. 44.
29. M. Biermann, P. Schulz, *GIT Fachz. Lab.* **35** (1991) 963.
30. *Tribol. Schmierungstech.* **37** (1990) 193.
31. H. U. Woelk, personal communication, see also [48]
32. D. Schliephake et al.: *Nachwachsende Rohstoffe,* Verlag J. Kordt, Bochum 1986.
33. D. Fengel, G. Wegener: *Wood, Chemistry, Ultrastructure, Reactions,* Walter de Gruyter, Berlin 1984.
34. L. Greg et al.: International Bioenergy Directory no. 353, The Bioenergy Council, Washington 1982.
35. R. Patt, *Nachr. Chem. Tech. Lab.* **36** (1988) 26.
36. P. Finch, J. C. Roberts: "Enzymatic Degradation of Cellulose," in: T. P. Nevell, S. H. Zeronian (eds.): *Cellulose Chemistry and its Applications,* Horwood, Chichester 1985, pp. 312 – 343.
37. M. Korn: *Nachwachsende und bioabbaubare Materialien im Verpackungsbereich,* Verl. Roman Kovar, München 1993, pp. 119, 463 – 464.

38. Lacke und Lösemittel, Verlag Chemie, Weinheim – New York 1979, S. 8 – 10.
39. M. Wandel, C. Leuschke, Kunststoffe **74** (1984) 589.
40. in [37], p. 122.
41. J.-D. Gu, D. T. Eberiel, S. P. McCarthy, R. A. Gross, *J. Environment. Polym. Degradation* **1** (1993) 143.
42. in [37], p. 122.
43. Tubize Plastics: *Biocell 163,* company publication, Tubize 1991.
44. *Chem. Ind. (Düsseldorf)* **114** (1991) no. 3, 64.
45. A. J. Backhouse, *J. Coat. Technol.* **54** (1982) 83.
46. L. Jeromin: *Nachwachsende Rohstoffe für die Chemische Industrie – eine neue verfahrenstechnische Herausforderung?* VDI-Verlag, Düsseldorf 1994.
47. Chemic. Economics Handbook 1989, cited in [46]
48. G. Tegge: *Stärke und Stärkederivate,* 2nd ed., B. Behr's Verlag, Hamburg 1988.
49. H. Baumann, *GIT Fachz. Lab.* **35** (1991) 957.
50. H. Koch, H. Röper, R. Höpke, *Spec. Publ. R. Chem. Soc.* **134** (1993) 157.
51. H. U. Woelk: "Die industrielle Verwendung von Stärke," in: *Stärke im Nichtnahrungsbereich,* Schriftenreihe des BML, Reihe A/Heft 380, Landwirtschaftsverlag, Münster 1990, pp. 5 – 19.
52. V. T. Breslin, *J. Environment. Polym. Degradation* **1** (1993) 127.
53. in [37], pp. 37 – 55.
54. R. P. Wool, S. M. Goheen, Intern. Symposium on Biodegradable Polymers, Tokyo 1990.
55. F. H. Otey, R. P. Westhoff, C. R. Russel, *Ind. Eng. Chem. Prod. Res. Dev.* **16** (1977) 305.
56. in [36], pp. 72 – 75.
57. Novamont: "The Biodegradable Nature of Mater-Bi," company publication, Kunststoff-Information ki-Nr. 29558, Bad Homburg 1990.
58. C. Bastioli, V. Bellotti, L. Del Giudice, G. Gilli, *J. Environment. Polym. Degradation* **1** (1993) 181.
59. in [37], pp. 76 – 85.
60. K. Steinbach, *Umwelt* **23** (1993) 546.
61. F. W. Lichtenthaler, A. Boettcher in : M. Eggersdorfer, S. Warwel and G. Wulff (eds.): *Nachwachsende Rohstoffe – Perspektiven für die Chemie,* VCH Verlagsgesellschaft, Weinheim 1993, pp. 151 ff.
62. The Association of American Feed Control Officials, Official Publication, 1977, Section 36, pp. 87 –88.
63. W. Dimmling, *CZ – Chemie Technik* **2** (1973) 425.
64. Winnacker-Küchler: *Chemische Technologie,* 4th ed., **vol. 6,** Carl Hanser-Verlag, München 1982, pp. 520 – 521.
65. *Farbe Lack* **95** (1989) 832.
66. *Naturwiss. Rundsch.* **46** (1993) 192.
67. R. L. Whistler: *Industrial Gums,* Academic Press, New York 1973.
68. W. Kraus, A. Klenk, M. Bokel, B. Vogler, *Liebigs Ann. Chem.* 1987, 337.
69. Umweltbundesamt (ed.): Ökologische Bilanz von Rapsöl bzw. Rapsölmethylester als Ersatz für Dieselkraftstoffe (Ökobilanz Rapsöl), Jan. 1993.
70. K. Scharmer, G. Golbs, I. Muschalek, Union zur Förderung von Oel- und Proteinpflanzen e. V. (UFOP)-Studie "Pflanzenölkraftstoffe und ihre Umweltauswirkungen – Argumente und Zahlen zur Umweltbilanz," April 1993.
71. *Naturwiss. Rundsch.* **47** (1994) 76.
72. *New Sci.* **137** (1993) 22.
73. D. Schliephake: "Alkohole für Kraftstoffe," in: D. Schliephake et al. (eds.): *Nachwachsende Rohstoffe,* Verlag J. Kordt, Bochum 1986.
74. *Chem. Ind. (Düsseldorf)* **32** (1980) 84, 748.
75. nawaros – Nachwachsende Rohstoffe/Produktion, Projekte und Politik, no. 2/1993.
76. P. Breloh: "Die Bedeutung nachwachsender Rohstoffe aus Sicht des BML," in: Symposium Nachwachsende Rohstoffe – Perspektiven für die Chemie, Tagungsband, Landwirtschaftsverlag, Münster 1993, p. 19.
77. H. Hulpke, M. Marsmann, *Nachr. Chem. Tech. Lab.* **42** (1994) 11.
78. Dr. Wintzer et al.: Technologiefolgenabschätzung zum Thema "Nachwachsende Rohstoffe," Schriftenreihe des Bundesministers für Ernährung, Landwirtschaft und Forsten, Landwirtschaftsverlag, Münster 1993.
79. G. A. Reinhardt: *Energie- und CO_2-Bilanzierung nachwachsender Rohstoffe – Theoretische Grundlagen und Fallstudie Raps,* Verl. Vieweg, Braunschweig 1993.

Rayon → Cellulose

Chemical Products: Safety Regulations

Goffredo Del Bino, European Commission, Brussels, Belgium (Chap. 1)

Bernard Broecker, formerly Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany (Chaps. 2 and 4)

Roberto Binetti, Istituto Superiore di Sanità, Rome, Italy (Chap. 3)

Norman King, Department of the Environment, London, United Kingdom (Chaps. 5 and 6)

1. Introduction 2
2. Definitions 2
3. Hazard Identification 3
3.1. Significance of Parameters 3
3.1.1. Physicochemical Properties 3
3.1.2. Toxicological Properties 6
3.1.3. Ecotoxicological Properties 10
3.2. Interaction of Parameters 12
3.2.1. Hydrosolubility 12
3.2.2. Liposolubility 13
3.2.3. Corrosion/Irritation and Oxidizing Power 13
3.3. Classification Schemes: The European Union System 13
3.3.1. Transportation 14
3.3.2. The EU Classification for the Marketing of Dangerous Substances 14
3.4. Hazard Identification of Preparations: Calculation Systems for Different Use Categories (EU Approach) 20
3.4.1. Classification Based on Physical and Chemical Effects 20
3.4.2. Classification Based on Biological Effects 21
3.4.3. Classification Based on Environmental Effects 21
4. Risk Assessment 21
4.1. General Methodology 21
4.2. Qualitative and Quantitative Aspects: Toxicological and Ecotoxicological Properties 23
4.2.1. Hazard Identification 23
4.2.2. Dose (Concentration) – Response (Effect) Assessment 24
4.2.3. Exposure Assessment 24
4.2.4. Risk Characterization 24
4.2.5. Physicochemical Properties 25
4.3. Risk – Benefit Considerations 25
4.4. Definition of Acceptable Risk 25
4.5. Summary of Risk Assessment 26
5. Risk Management 26
5.1. Supplying Information (Risk Communication) 27
5.2. Codes of Practice 28
5.3. Product Standards 28
5.4. Control of Supply 28
5.5. Control of Use 28
5.6. Voluntary Agreements 28
5.7. Economic Instruments: Harnessing Market Forces 29
5.8. Integration of Pollution Control and Chemicals Control 29
5.9. Monitoring Effectiveness 29
6. Ecolabeling 30
6.1. Philosophy 30
6.2. The EU Scheme 31
7. References 32

The entire topic was coordinated by Goffredo Del Bino

1. Introduction

About 30 years ago, governments started to adopt measures to control chemicals with the aim of protecting people and the environment. Since then, rules and regulations have been introduced in many industrialized countries to assess and manage the risks of chemicals. It was recognized from the outset that harmonization of controls is important if the need for protection was to be reconciled with the aim of avoiding technical barriers to trade. The European Union (EU) took the first step in 1967 with the adoption of Directive 67/548 on classification, packaging, and labeling of dangerous substances, since then amended seven times to introduce appropriate risk assessment measures.

In the United States, the Toxic Substances Control Act (TSCA) became law in 1976. It introduced pre-manufacturing/pre-marketing assessment of the effects of chemicals.

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_181

In the 1970s, the OECD, grouping for 24 most industrialized countries in the world, launched a Chemicals Programme which is still the most important international initiative in the field of chemicals control.

All chemicals are toxic. The difference between them is the dose required to cause toxic effects: sugar and salt, if consumed by the kilogram, may be as lethal as milligrams of cyanide or strychnine. In addition to their toxicity, chemicals may be explosive, flammable, or corrosive. Others may have deleterious effects on our environment, e.g., chlorofluorocarbons.

Given that all chemicals are potentially dangerous, it is reasonable that precautions be taken to assess each chemical and to ensure that appropriate steps are taken to reduce the potential risks associated with its use to acceptable levels.

Control of chemicals includes the different steps illustrated in Figure 1.

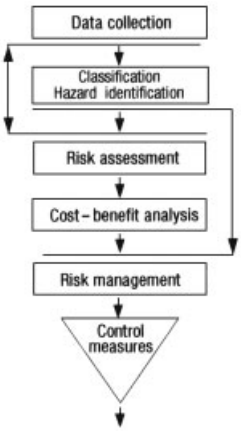

Figure 1. Schematic of the process of chemicals control

In recent years, governments and industry have come to realize the part that consumer choice can play in protecting the environment from chemicals and other hazards. The need to develop more environmentally friendly products is now well recognized, and this article also deals with the objectives and function of ecolabeling systems.

2. Definitions

Risk assessment is a stepwise, iterative process, comprising the following steps:

1) Hazard identification: the identification of the adverse effects which a substance has an inherent capacity to cause
2) Dose (concentration) – response (effect) assessment: estimation of the relationship between the dose, or level of exposure to a substance, and the incidence and severity of its effect
3) Exposure assessment: determination of the emissions, pathways, and rates of movement of a substance, and its transformation or degradation in order to estimate the concentrations/doses to which human populations or environmental compartments are or may be exposed
4) Risk characterization: estimation of the incidence and severity of the adverse effects likely to occur in a human population or environmental compartment owing to the actual or predicted exposure to a substance
5) Risk estimation: quantification of the estimated likelihood in a risk characterization

3. Hazard Identification

The risk assessment of a chemical substance is a complex procedure based on comparison of its potential adverse effects with the reasonably foreseeable exposure of people and the environment to that substance. The identification of the potential adverse effects of a substance is the first step of this procedure [1].

Hazard identification is mainly based on a critical evaluation of the intrinsic properties of a substance, i.e., the physicochemical, toxicological, and ecotoxicological properties. For each of these groups of properties it is possible to define a number of specific end points, and for each of them internationally recognized guidelines [2], [3], to determine specific parameters. The interpretation of such parameters leads to the hazard identification of the chemicals.

3.1. Significance of Parameters

Intrinsic properties characterize the behavior of a chemical substance. They are directly linked to the structure of the molecule, although sweeping conclusions about behavior and properties

cannot be drawn simply because a substance belongs to a chemical class or family. It is enough to substitute one radical by another to induce specific properties which are not shared by the whole family.

Consequently, although the relationship between structure and activity is now an area of great scientific interest and ever-growing endeavors are being made to replace compulsory, costly, and long-term experimental studies by mathematical models, it is still necessary to use experimental methods to determine the basic properties from which the potential effects of a given substance can be deduced.

The following paragraphs deal with physicochemical, toxicological, and ecotoxicological properties, attempting briefly to clarify the significance of each of these and the principles underlying the methods of determining them, with direct reference to existing internationally recognized methods.

3.1.1. Physicochemical Properties

Physicochemical properties play an important role in the potential of a substance to produce adverse effects. There is a close correlation between physicochemical properties and the occurrence of toxic and environmental effects.

Each physicochemical property plays a specific role. For instance, the higher the molecular mass, the lower the possibility of absorption by the organism; the higher the melting point, the greater the physical stability of the substance; the lower the boiling point, the more readily the substance can be inhaled; liposoluble substances can penetrate more readily through the skin, facilitating their bioaccumulation, while hydrosoluble substances have a preference for gastric and pulmonary channels; substances able to lower surface tension are more readily absorbed and promote the absorption of substances which are difficult to absorb on their own. Similar considerations apply to the various sectors of the environment since substances can enter, move among, persist, and bioaccumulate in natural species.

Molecular Mass. The size of the molecule gives an idea of whether the substance is bioavailable, i.e., can be absorbed, assimilated, and metabolized by the organism. A comparatively small molecule can enter the organism and produce its effects much more readily than a large molecule. A high polymer cannot generally be absorbed and assimilated and is therefore much more inert as regards interaction with the human organism or other species.

Melting Point. The melting point characterizes the physical state of the substance at ambient temperature, and normal pressure. If the substance is solid at room temperature, it is less mobile than a substance in the liquid or gaseous state.

Boiling Point. Like the melting point, the boiling point characterizes the physical state of the substance under normal pressure conditions, providing important information on its ability to affect, for instance, channels of inhalatory exposure. In case of diffusion into the environment, it qualitatively indicates how easily the substance can be transferred into the air, and to what extent it diffuses into the environment. The higher the boiling point, the smaller the risk of inhalatory effects.

Relative density represents the mass ratio between a solid or liquid substance and water or, in the case of gaseous substances, between the substance and air.

When referred to water, it is normally expressed as d_4^{20} (mass of the substance measured at 20 °C and mass of water measured at 4 °C). If the substance does not dissolve in water, its density indicates whether the substance can be expected to stratify on the surface (density <1) or to sink to the bottom (density >1), which is important in relation to mobility in the environment and the diffusion of pollution into different compartments, including indirect human exposure.

When referred to air, substances which are denser than air (density >1) tend to stratify around the source of emission, constituting a major risk of exposure at relevant concentrations, while at density <1, diffusion into the air is easier, and the risk is generally reduced since it is less persistent in the vicinity of the emission site.

Vapor pressure is also an important way of finding out whether the substance, at ambient

temperature, may readily become volatile from a solution, another liquid, or solid matrix. It is defined as the pressure of saturation above a solid or a liquid. A substance with a high vapor pressure chiefly affects inhalatory and dermal channels of exposure. As regards diffusion into the environment, higher values are linked to an increased transfer potential. As regards physical danger, higher vapor pressure may be associated with risks of explosion and the need for greater precautions as regards storage, packaging, and transportation. In brief, greater risks are attached to high vapor pressure values.

Surface Tension. The ability of a substance to lower the surface tension of water is closely linked to its increased potential for absorption, permeation of barriers (membranes), and diffusion into water, even when the substance is scarcely hydrosoluble. A substance which can lower the surface tension of water may also make it easier for liposoluble substances which are insoluble or only slightly soluble in water to dissolve in water. A typical example is provided by surfactants which facilitate absorption of substances such as polycyclic aromatic hydrocarbons which are not hydrosoluble.

The surface tension value which is of interest is not that of the substance per se, but of its solution in water. The greatest risks are associated with low surface tension values measured in this solution.

Water solubility is defined as the concentration of saturation in water at a defined temperature. It is generally expressed as mass per unit volume of solution.

Hydrosolubility has direct implications as regards both potential toxicological and ecotoxicological effects. Hydrosolubility shapes adsorption by gastric or pulmonary channels, the possibility of dissolution in surface waters, and therefore the mobility in water and soil, as well as the precipitation of substances onto the soil in rain.

Fat solubility (liposolubility) expresses a substance's potential to dissolve and eventually accumulate in human and natural species' fats. The more liposoluble a substance, the more easily it can penetrate through the skin and produce local or systemic effects. It is normally measured using standard triglycerides with defined fatty acids content. Alternatively, the solubility in organic (possibly nonpolar) solvents may be useful information.

Partition Coefficient *n*-Octanol/Water. The problem of bioaccumulation may reach very high proportions in natural species. Aquatic species, which filter enormous quantities of water, accumulate liposoluble substances on a scale which increases with the species' physical size and life expectancy, playing a part in "magnifying" pollution of the trophic chain.

An idea of the bioaccumulation capacity of a substance may be gained from the *n*-octanol/water partition coefficient. This is obtained by establishing the ratio at which a given substance is distributed when brought into contact with a system of two immiscible phases formed by normal octyl alcohol (*n*-octanol) and water. The *n*-octanol simulates the fatty phase: if the substance dissolves to a greater extent in the aqueous phase the partition coefficient assumes negative values (it is normally expressed as a decimal logarithm, $\log P_{ow}$), showing that the substance is more hydrosoluble than liposoluble; positive values in the case of substances which dissolve to a greater extent in the octanol phase, indicating that the substance can bioaccumulate (normally this requires $\log P_{ow} > 3$).

The partition coefficient is particularly important for substances which are scarcely soluble in water and fats. It might be thought that low liposolubility entails a low level of risk of bioaccumulation, whereas if the substance is more liposoluble than hydrosoluble (i.e., high partition coefficient), bioaccumulation may also occur, but to a lower extent because of the low absolute liposolubility. The partition coefficient is independent of the absolute values for liposolubility and hydrosolubility, expressing the ratio between them.

Flammability. This property, like the following physicochemical properties, provides direct information on the possibility of a so-called physical effect (it is better to talk of physicochemical effects, since oxidation is a typical chemical effect). While the previously described properties "characterize" the substance, giving indirect information on the potential bioavailability and mobility into the environment, the

properties regarding physical effects provide information on immediate and direct risks. Most damaging accidents are linked to physical effects, often with consequences which are disastrous in terms of their size, persistent effects, and number of casualties.

The risk of flammability may derive from a low vapor ignition temperature in the presence of an ignition source (primer) or from the possibility of the substance igniting spontaneously in air with no ignition point (autoflammability).

For liquid products, the flash point is the parameter which allows quantification of the flammability risk. It is defined as the lowest temperature, at a corrected pressure of 101.325 kPa, at which a liquid emits vapors in such quantity as to produce a flammable vapor/air mixture. Experimentally, the substance is slowly and progressively heated and, at regular intervals, its vapors are exposed to a spark: the lowest temperature at which the vapors/air mixture flashes is the flash point. This type of test is appropriate for liquids and low-melting-point substances. A different type of test is carried out for solid substances, yielding a simple positive/negative response, by evaluating the substance against a reference flammable product.

If the substance decomposes during heating, the flash point is still of interest, as the ignition of the vapors may be due to the mixture of decomposition product vapors rather than the substance per se. Moreover, account has to be taken of the case in which the substance per se is not flammable, but releases flammable gases in contact with other substances, e.g., water.

Some gaseous substances ignite in the presence of a primer at low or very low temperature; in this case it is useful to know the range of concentration in air at which flame propagation occurs.

Information on autoflammability can prevent serious accidents, especially during long-term storage of chemical products. It is vital to know whether a substance may autoignite at temperatures which are foreseeable under bad storage conditions.

Explosive Properties. A substance is considered to have explosive properties if, following heating, shock, or friction, it may decompose abruptly, releasing large quantities of energy and/or gas. Explosive properties are tested by methods which take account of the three different possibilities (heat, shock, friction). If the substance is shown to have explosive properties similar to or greater than those of a reference substance (e.g., dinitrobenzene) it is considered explosive.

Oxidizing Properties. A substance with oxidizing properties may, when brought into contact with chemically oxidizable organic or inorganic substances, bring about dangerous reactions causing fire, explosion, or the formation of other hazardous substances. Oxidizing properties are not just related to the presence of oxygen in the molecular structure, but to its potential for acting more generally as an oxidant in an oxidation – reduction reaction. In many cases, specific tests are not necessary to determine whether or not it is oxidizing; examination of its structural formula may provide predictive information. Organic peroxides, for instance, are generally considered as oxidizing, even in the absence of test results. The only doubt is whether they should also be considered as explosive; for this, experimental evidence is needed.

3.1.2. Toxicological Properties

Toxicological properties are of direct interest in assessing the risk of chemical substances. They include a range of effects understood either as general poisoning of the organism or aimed at one or more particular target organs, and whose intensity is generally proportional to the dosage absorbed and to specific effects, which may be independent of dose, e.g., mutagenic and carcinogenic effects. There may also be immediate effects, generally acute, or effects deferred in time, linked to the intake of small, continued doses, or with a period of latency before the effects appear.

The dosage administered and the duration of toxicological tests are inversely proportional with reference to acute, subacute, subchronic, and chronic studies. When studying acute effects, the substance is usually administered to test animals as a single dose, at a level sufficiently high to cause death. When studying subacute effects, lower dosages are used, but for a longer and continuous period (28 d). For subchronic toxicity, even smaller dosages are used,

but over a period corresponding to ca. 10 % of the animal's life; for chronic toxicity the test is extended throughout the animal's life, usually at even lower dosages.

Some basic properties need to be differentiated for assessment of the various types of risk connected with toxicological properties.

Acute toxicity includes those effects which are or may be produced following a single administration. In terms of acute effects, it is common to distinguish between systemic effects (lethal effects following oral, cutaneous, or inhalatory administration) and local effects, such as irritation of the eyes and skin, or corrosive and sensitizing properties.

Each of these effects needs to be examined in slightly more detail.

LD_{50} and LC_{50}. The acute lethal effects are generally evaluated through the LD_{50} (dose killing 50 % of animals tested), valid for oral and cutaneous administration, and the LC_{50} (concentration killing 50 % of animals tested), valid for inhalatory administration. Other routes of administration, e.g., peritoneal, intramuscular, or intravenous, although of some interest, have no direct correlation with the study of chemical substances in relation to personal, domestic, professional, or environmental exposure of humans.

A basic consideration underlies the determination of the oral LD_{50}, cutaneous LD_{50}, and inhalatory LC_{50}: the greater the toxicity of the substance, the smaller the value obtained.

The oral LD_{50} is generally determined by administering a known quantity of the substance to the animal (the rat is taken as a reference) by gastric probe (gavage), possibly dispersed in an appropriate vehicle, depending on the solubility of the substance in water or oil, or a different inert vehicle for insoluble substances. After administration, the animal must be observed for up to 14 d. The LD_{50} is assumed to be the dosage, expressed in milligrams substance per kilogram animal body weight (mg/kg), able to cause, during the observation period, death of 50 % of the test animals. In order to determine the LD_{50}, it is necessary, after preliminary tests, to use groups of animals to which a range of increasing dosages are administered, so that the LD_{50} value may be interpolated. The test should not be limited to establishing the number of animals killed in relation to those tested, but should also be used to show, through appropriate histopathological studies, the organs chiefly affected, and thereby to gather information on the possible mechanism of action of the substance.

The cutaneous LD_{50} is determined by applying the substance, possibly dissolved or dispersed in an inert vehicle, to the shaved skin of the test animal (rat or rabbit) and ensuring that it remains in contact for a period of time (24 h or periods depending on the substance's potential corrosive or irritant effects). As mentioned above, liposoluble substances can readily penetrate through the skin, but this possibility is not significant for nonliposoluble substances. Observation must be continued for 14 d after the removal of the bandage, and subsequent investigations should endeavor to gather any other useful information. The cutaneous LD_{50} is also expressed in mg/kg.

The inhalatory LC_{50} is determined by exposing the animal (rat) for 4 h to an atmosphere in which the substance is uniformly dispersed at a constant concentration. The dosage capable of killing 50 % of the animals is expressed as a concentration, in milligrams per liter of air.

For some substances in the form of dusts, it may be difficult or impossible to disperse a concentration of the substance sufficient to cause fatal effects. In such cases, it is possible only to indicate that the LC_{50} is higher than the maximum concentration tested.

Irritant or Corrosive Properties. The irritant or corrosive properties of a test substance are determined with respect to the skin and the eyes. Two separate tests are conducted on rabbits and provide qualitative or semiquantitative information.

Irritant or corrosive effects on the skin are tested by leaving the substance in contact with the intact, shaved skin of the animal, and observing the effects produced after 4 h; these effects may range from the absence of effects to erythema, the formation of eschar or edema, or tissue corrosion. Corrosion occurs when the whole thickness of cutaneous tissue is destroyed.

Irritant effects on the eyes are tested by instilling a small quantity of the substance into one of the animal's eyes and using the other eye as a reference. The severity of the effects is compared with a reference scale, looking in particular at the effects on the cornea, iris, and conjunctivae.

Sensitizing Properties. Sensitization is an allergic cutaneous or inhalatory reaction occurring when the organism, hypersensitive to a chemical substance, is exposed to the substance, even at very low concentrations. This property may be experimentally evaluated only for skin sensitization; no methods exist for the evaluation on test animals of inhalatory sensitization. The test substance is administered to the animal (guinea pig) by cutaneous injection (induction phase) and after an appropriate period the administration is repeated (challenge phase), observing whether the animal is subject to sensitization phenomena.

Sensitizing effects on the respiratory tract are considered a more severe hazard. However, given the lack of a practicable test, a substance is considered to be sensitizing by inhalation on the basis of epidemiological evidence.

Subacute, Subchronic, and Chronic Toxicity. Medium- or long-term toxicity tests provide essential additional information on the more hidden effects which the substance may cause by exposure to low dosages over a period of time. This type of test can also be conducted via the various routes of exposure (oral, inhalatory, cutaneous) in accord with the main route of human exposure.

A basic aim of these tests is to identify the maximum dose level at which the substance does not present toxic effects. This makes it necessary to test several increasing dose levels, so as to determine the dosage at which toxic effects start to appear. This makes it possible to show whether the toxic phenomenon is dose dependent, i.e., whether the intensity of the effect is proportional to the dose administered.

Tests results in the literature contain a range of notation:

1) NEL (no effect level), expressing the maximum level where there are not even minor effects
2) NOEL (no observed effect level), expressing the maximum level at which no effects have been observed
3) NOAEL (no observed adverse effect level), expressing the maximum level at which no adverse effects have been observed

Between NEL and NOEL, generally the latter is preferred, expressing the actual experimental dose level at which no effects have been observed; NEL is a more theoretical value, situated somewhere above the NOEL, more precisely between the dose level corresponding to the NOEL and the subsequent dose level. For identification of the NOAEL, the problem is to characterize exactly what "adverse effect" means. There is the need to distinguish between severe and minor effects, and the debate is still open, although the internationally recognized guidelines have attempted to define, at least in general terms, an adverse effect.

In any case, the NOEL (or NEL, or NOAEL) should be expressed in milligrams substance per kilogram animal body weight and day ($mg\,kg^{-1}\,d^{-1}$).

If the test is conducted by administering the substance to the animal in its food, the results may also be expressed as ppm in the diet (mg/kg of food). It is possible to convert, within an acceptable approximation, one value to another. In the rat, for instance, this conversion consists of dividing the concentration of the substance in the food by 20. Animals other than the rat, with metabolism closer in some ways to that of humans, are mice, dogs, and monkeys.

Subacute toxicity tests are designed to reproduce repeated exposure over short periods. Subchronic toxicity tests may, in some cases, replace chronic toxicity tests; if subchronic toxicity tests, which last 3 months in the case of rats (i.e., 10 % of their life expectancy), do not reveal major effects at sufficiently high dosages and it is assumed that the chronic toxicity tests are not likely to add relevant information, longer term tests may not be necessary, thereby reducing suffering of the animals and study commitments. Chronic toxicity tests are conducted for the animal's entire life and are designed to reproduce continuous, prolonged exposure.

Various macroscopic (body weight, general condition, food intake) and chemical and clinical measurements (blood, urine, and other specific tests) are also conducted during the main test. Small groups of animals may be killed at predetermined times for histopathological studies which are in any case carried out at the end of the trial with a detailed examination of the individual organs and glands.

The NOEL value is a basic factor in assessing the risks connected with the toxicity of a given substance as a result of medium- or long-term

exposure. The smaller the NOEL value, the more hazardous the substance, with the result that the risks are inversely proportional to the numerical value of this parameter.

Mutagenesis. A substance is considered mutagenic if its adsorption may produce heritable genetic damage or increase their frequency. Assessing a substance's ability to cause mutation in humans, at germinal or somatic level, is very difficult on the basis of current knowledge. Several in vivo and in vitro tests are designed to predict the presence of a potential ability to cause mutations. The comparative simplicity and brevity of these tests, together with the knowledge that many carcinogenic substances are also mutagenic, has promoted increasing interest in a wide-ranging set of mutagenesis tests, each of which can provide specific information on mutagenic potential, bearing in mind that general conclusions cannot be drawn from a single test. It is therefore necessary to carry out a battery of tests to obtain a minimum set of information.

In vivo tests obviously have a greater significance than in vitro tests, although the latter are more sensitive. In vivo studies are usually performed only when a battery of in vitro tests gives positive results.

A number of categories for mutagenic substances have been established for operating purposes and for risk assessment. The EU and other international and national organizations have formulated criteria defining different categories of danger with different levels of mutagenic potential. The allocation of a substance to one or other of these categories provides an overall qualitative assessment of its possible mutagenic potential.

The EU, for instance, has adopted the following classification:

Category 1: substances known to be mutagenic to humans. There is sufficient evidence to establish a causal association between human exposure to a substance and heritable genetic damage.

Category 2: substances which should be regarded as mutagenic to humans. There is sufficient evidence to provide a strong presumption that human exposure to the substance may result in the development of heritable genetic damage, generally on the basis of appropriate animal studies or other relevant information.

Category 3: substances which cause concern, owing to possible mutagenic effects, but in respect of which the available information does not satisfactorily demonstrate heritable genetic damage. There is evidence from mutagenicity studies, but this is insufficient to place the substance in Category 2.

As in the case of other properties of chemical substances, it is possible in mutagenesis to use the structure – activity relationship for an initial assessment of possible mutagenic effects. Extremely reactive substances which are free radical donors are likely to give positive results in mutagenesis tests.

Carcinogenesis. Substances are considered carcinogenic if, after exposure (generally long-term) by inhalation, ingestion, or skin contact, they may produce cancer or increase its frequency. Test protocols have been drawn up at international level for studies of carcinogenic potential of chemical substances.

These long-term tests, conducted using the exposure routes (oral, inhalatory, cutaneous) assumed to be most representative of normal exposure conditions for humans, include detailed histopathological investigations intended to show the number, type, and location of tumors; data need to be statistically processed in comparison with control groups to take account of the natural occurrence of tumors in the particular strain of animals studied, and to establish significance.

Several, substantially similar, international classifications have also been formulated for carcinogens; substances are allocated to different categories in relation to the carcinogenic potential demonstrated.

The following three categories, in decreasing order of risk, have been formulated by the EU:

Category 1: substances known to be carcinogenic to humans. There is sufficient evidence to establish a causal association between human exposure to a substance and the development of cancer.

Category 2: substances which should be regarded as carcinogenic to humans. There is sufficient evidence to provide a strong presumption that human exposure to a substance

may result in the development of cancer, generally on the basis of appropriate long-term animal studies or other relevant information.
Category 3: substances which cause concern owing to possible carcinogenic effects, but in respect of which the available information is not adequate for satisfactory assessment. There is some evidence from animal studies, but this is insufficient to place the substance in Category 2.

Also in the case of carcinogenesis, the fact that a substance belongs to a chemical family whose carcinogenic properties are known should pave the way for detailed specific studies.

Embryotoxicity. Teratogenic effects, i.e., nonheritable birth defects, produced by a chemical substance, are now covered by a broader class of effects, covering all possible effects on reproduction; embryotoxicity seems to be the most appropriate term. Two main subcategories of effects exist: possible effects on human fertility, and developmental toxicity.

Substances are considered dangerous to human fertility if they may produce or increase the frequency of damage either to reproductive functionality or to the reproductive capacity of males and females. Substances are considered toxic to development if they may produce adverse nonheritable effects in the offspring, or increase their frequency. There are also specific test protocols to predict potential embryotoxicity on the basis of animal studies.

The present EU scheme distinguishes three categories for each kind of effect:

1) Substances dangerous to human fertility
   Category 1: substances known to impair fertility in humans. There is sufficient evidence to establish a causal relationship between human exposure and impaired fertility.
   Category 2: substances which should be regarded as impairing fertility in humans. There is sufficient evidence to provide a strong presumption that human exposure may result in impaired fertility on the basis of appropriate animal studies or other relevant information.
   Category 3: substances which may cause concern for human fertility, essentially on the basis of animal studies providing sufficient evidence to cause a strong suspicion of impaired fertility in the absence of toxic effects.
2) Substances dangerous to human development
   Category 1: substances known to cause developmental toxicity in humans. There is sufficient evidence to establish a causal relationship between human exposure and subsequent developmental toxic effects in the progeny.
   Category 2: substances which should be regarded as causing developmental toxicity to humans. There is sufficient evidence to provide a strong presumption that human exposure to the substance may result in developmental toxicity, generally on the basis of animal studies or other relevant information.
   Category 3: substances which cause concern owing to possible developmental toxic effects, generally on the basis of animal studies which provide sufficient evidence to cause a strong suspicion of developmental toxicity in the absence of signs of marked maternal toxicity.

3.1.3. Ecotoxicological Properties

As toxicological properties, studied by tests on animals, are indicators of potential effects in humans, so ecotoxicological properties, studied through tests on natural species, are indicators of the potential effects which a substance may have on the environment. Some species, considered to be "biological indicators" are taken to be representative of the various compartments (water, soil, air).

It is difficult to make a clear distinction between these compartments since air, soil, and water interact. Soil contamination can be assessed directly through toxicity tests on species which live in the soil, e.g., plants and earthworms, but soil may be permeated by water and washed by atmospheric and surface waters. This leads to contamination of the water compartment and aquatic species, both in situ and in watercourses and catchment areas. Diffusion of a substance in the air may lead to precipitation and dissolution in water. Diffusion in the soil or in water

may lead to volatilization and thus to contamination of the air. Direct contamination of rivers or lakes may pave the way for diffusion through the soil. Contamination of soil may contaminate land species (flora and fauna) which may become the prey of birds.

The aim of these tests is to identify sufficiently representative biological indicators of the toxicity of a substance to ecosystems.

Toxicity in Fish. Toxic effects on fish may occur following acute or prolonged exposure to concentrations which may be very low, bearing in mind the ability of aquatic organisms to recycle very high quantities of water and, in many cases, to bioaccumulate toxic substances.

Acute toxicity is evaluated by keeping the fish for 48 – 96 h in water containing a known concentration of the substance. A number of tests have to be conducted in parallel to determine the concentration at which no lethal effect occurs (LC_0), the concentration which is fatal for all tested animals (LC_{100}) and the median concentration which is fatal for 50 % (LC_{50}). The latter is normally calculated by interpolation among the other data or at least between the two extreme values. LC_{50} is expressed in milligrams per liter of water (mg/L).

The smaller the LC_{50} value, the greater the toxicity of the substance for the species tested.

Long-term toxicity in fish is of interest, even if their sensitivity to extraneous substances in water is very high, and the LC_{50} can be taken as a very representative parameter. The LC_{50} and the LC_{100} are often very close numerically. It is important to look for long-term toxicity in substances which are not highly hydrosoluble since ingestion of very small quantities of substances in the aqueous medium may present delayed toxic effects.

The species recommended for the test are:

Brachydanio rerio (zebra fish)
Pimephales promelas (fathead minnow)
Cyprinus carpio (common carp)
Oryzias latipes (red killifish)
Poecilia reticulata (guppy)
Lepomis macrochirus (bluegill)
Salmo gairdneri (rainbow trout)
Leuciscus idus (golden orfe)

On the basis of tests specifically designed to measure long-term toxicity, it is possible to identify an NOEC (no observed effect concentration).

Toxicity in Daphnia. *Daphnia* is a very small crustacean that makes rapid and frequent swimming movements. The *Daphnia* test is designed to detect the immobilization effect, which is taken as an index of the adverse effects caused by the substance.

The substance is dissolved in water, over a range of increasing concentration and the *Daphniae* are exposed for 48 h. The mean effective concentration (EC_{50}) is generally interpolated between the EC_0 and EC_{100}. All data are expressed in mg/L.

As for fish, the lower the EC_{50} value, the greater the acute toxicity of the substance for *Daphnia.* There may be problems with substances which are scarcely hydrosoluble in acute toxicity tests with *Daphnia.*

Effects on reproduction may be evaluated through long-term tests (21 d) on *Daphnia magna*. An NOEC can be identified.

Acute Toxicity in Birds. No acute toxicity tests for birds formulated at international level are yet available, but the literature provides toxicity data for a number of species (quails, ducks, pigeons, chickens).

Administration generally takes place orally through the diet, linking the quantity ingested with body weight. Use is therefore made of the dose level which is lethal for 50 % of the birds in the test (LD_{50}), expressed in mg/kg.

Toxicity in Higher Plants. Plants are biological indicators of major importance. The damage to higher plants by environmental pollution, for instance, is commonly measured. Up to now there have not been any standard methods, given the wide range of species and reactions which may be obtained. The tests available are based on the effect which a given substance may have on the germination and growth of the plants tested, expressed as the EC_{50}.

Herbicides, which by their nature and function, have toxic effects on plants (phytotoxic effects) may be considered as reference substances. In the absence of accepted international methods, the substance can only be generally considered as phytotoxic, moderately phytotoxic, or nonphytotoxic.

Toxicity in Earthworms. The earthworm is an important indicator of soil life following contamination by chemical substances. This information, together with data from toxicity tests on plants, relates therefore to the soil compartment.

The test already defined at international level uses *Eisenia foetida* kept in standard artificial loam soil treated with different concentrations of the substance. After 14 d the soil is examined and the surviving earthworms counted. The result is expressed as LC_{50} with respect to the soil.

Effects on Algae. Algae are important biological indicators. Unfortunately, the available data focus largely on the detection of effects inhibiting algal gowth, while experience has shown that major adverse effects on the environment may result from the abnormal increase of algae as a result of contaminants in the water.

The tests are carried out in water and the EC_{50} is understood as the concentration which inhibits growth by 50 % (also referred to as inhibitory concentration, IC_{50}). It is more convenient to refer to the substance's ability to produce both positive and negative effects, i.e., changes in normal development.

Persistence. The lifetime of a chemical substance may vary as a function of external conditions. A substance which is kept dry in a dark, sterile, oxygen-free atmosphere can remain unchanged indefinitely.

Once introduced in the environment, a substance is subject to different kinds of aggression which, depending on the molecular structure, may produce total or partial degradation of the substance. Such aggression may be divided in two types: biotic and abiotic.

The more rapidly decomposition takes place, biotically or abiotically, the lower the chance of the substance having any residual adverse effect. Nevertheless, unless the substance is totally decomposed to simple compounds (H_2O, CO_2, NO_x, SO_x, PO_x), identification of the degradation products is essential for complete evaluation of residual effects; the degradation products may be more dangerous than the parent molecule.

Biotic degradation is caused by bacteria and microorganisms which use the substance for their own development, causing it to decompose. Organic chemicals are largely formed from carbon and other elements such as hydrogen, oxygen, nitrogen, sulfur, and phosphorus, and they may become involved in the various cycles of use and conversion by bacteria and microorganisms.

The methods formulated for the determination of biodegradation are therefore based largely on determining the quantity of oxygen needed for the oxidation, in due time, of the organic carbon contained in the test substance. Bearing in mind the various environmental conditions in which the substance may be dispersed, the analytical methods set out different conditions for the addition of bacterial inocula, e.g., in activated sludge and sewage, and different operating methods which attempt to ascertain whether the substance can be decomposed under particular conditions by the action of bacteria and microorganisms.

An initial measurement (ready biodegradability) is based on determining the percentage of the substance degrading over 28 d; a substance which degrades by more than 70 % is generally considered biodegradable.

The biochemical oxygen demand (BOD) is related to the chemical oxygen demand (COD), the latter representing the amount of oxygen needed for complete oxidation of the substance by chemical methods. The BOD/COD ratio is a useful factor for assessing to what extent the substance degrades by biotic media.

Abiotic Degradation. The degradation of a chemical substance may also take place abiotically as a result of hydrolysis if it comes in contact with water or moisture and/or as a result of exposure to sunlight (photodegradation).

Decomposition by hydrolysis is usually assessed by determining the concentration of the substance in an aqueous solution as a function of time and pH, to simulate a range of possible acid or basic conditions in the receptor medium. The "half-time" $t_{1/2}$ is calculated as the time for the concentration of the substance to reduce to half the initial concentration.

Photodegradation is ascertained by exposing the substance to light of a given wavelength, under specific environmental conditions, and determining the residual fraction as a function of time. Here too, the aim of the test is the evaluation of $t_{1/2}$.

3.2. Interaction of Parameters

Each of the parameters examined in Section 3.1 gives information on the risk potential for humans and the environment of a chemical substance. Such parameters are, however, not completely independent. Some physicochemical properties, besides being clear indicators of mobility in the various environmental sectors, directly influence the biological behavior.

3.2.1. Hydrosolubility

Hydrosolubility is an important predictor of the behavior of a chemical substance, both in terms of mobility and biological activity. A substance with high hydrosolubility behaves in a largely predictable manner with respect to possible effects on human health and the environment.

A first consequence of high hydrosolubility is, in general, reduced liposolubility; there is almost always a predominance of either lipophilic or hydrophilic groups in a molecule. Only in some types of chemicals (e.g., surfactants) are both types of group represented to such an extent that the substance has hydrosolubility very similar to its liposolubility. Another direct influence is its *n*-octanol/water distribution coefficient log P_{ow}; since high hydrosolubility often corresponds to a low liposolubility, log P_{ow} is negative, resulting in a slight bioaccumulation potential in living organisms.

From the point of view of possible direct effects on humans, a hydrosoluble substance is easily absorbed through the gastrointestinal tract or the pulmonary alveolae; but it is less likely to be absorbed through the skin. In toxicological studies on animal species, it is necessary to concentrate on the oral route and, depending on the volatility (liquids) or particle size (solids), on inhalation. In the initial analysis, the cutaneous route can be ignored for possible systemic effects, though it remains relevant for possible local (corrosive/irritant) effects.

From the environmental aspect, hydrosolubility is fundamental in evaluating possible effects on aquatic species. Experiments on biological indicators such as fish, *Daphnia*, and algae are useless unless the substance has significant hydrosolubility. The same also applies to biotic and abiotic degradation rates.

3.2.2. Liposolubility

As hydrosolubility and liposolubility are generally inversely proportional, it may be expected that a substance with a high liposolubility will behave in a complementary manner to a substance with a high hydrosolubility.

A substance with high liposolubility generally has a low hydrosolubility and a high octanol/water distribution coefficient, with a significant bioaccumulation potential. From the point of view of the effects on humans, the cutaneous route becomes important as a liposoluble substance can readily be absorbed through the skin to produce immediate or subsequent systemic effects, the latter of which are particularly important because of the potential for bioaccumulation.

It is often difficult to evaluate the potential environmental danger of liposoluble substances since, at least as regards aquatic behavior, the tests are performed in aqueous solution without the addition of auxiliary solvents which can affect the hydrosolubility of the substance. Important is the real concentration, not the nominal concentration (this refers to the amount of substance added to the aqueous medium, and it may not be completely dissolved). If the hydrosolubility is negligible, acute toxicity studies in fish, *Daphnia*, and algae are irrelevant. It is wrong to believe that a highly liposoluble substance does not constitute a risk to the aquatic environment; incidents connected with the accidental discharge of petroleum products (crude oil in particular) reinforce this point. On the other hand, the presence of extremely low concentrations in water may constitute a danger for aquatic species, given their bioaccumulation capacity.

3.2.3. Corrosion/Irritation and Oxidizing Power

Chemically aggressive substances such as oxidizing agents can produce local effects on the skin. High reactivity is often synonymous with low persistence; it is particularly important to investigate the acute effects. It seems certain that chemically aggressive substances can influence genotoxicity.

Corrosive substances can present problems in interpretation of results, if they are tested to evaluate possible systemic effects. If a substance produces local effects, this does not mean that it cannot also produce systemic effects. It is precisely through the local effect that the cutaneous barrier is weakened, and the substance can enter the circulation and produce effects on organs or systems remote from the entry site. In spite of this, in order to understand the exact mechanism of action, it is important to distinguish mortality due to the local effect (e.g., perforation of the esophageal tract due to the corrosive effect) from mortality due to the systemic effect.

3.3. Classification Schemes: The European Union System

A chemical substance is characterized by a critical examination of its physicochemical, toxicological, and ecotoxicological properties. However, it is not feasible to require that every time a chemical substance (per se, or mixed with other substances) is handled, transported, or stored, it should be subjected to a comprehensive evaluation. Those who may come into contact with a chemical substance, whether laboratory staff or domestic users, hardly ever have sufficient knowledge to make such an evaluation; this work is therefore entrusted to committees of experts who systematically examine the relevant data in order to draw up concise guidelines. It is vitally important that such guidelines be expressed in clear and understandable terms, so as to provide essential information not only on the risk potential of the substance, but above all on the measures to minimize such risks.

Such ideas are implemented in practice via the danger classification in conjunction with the danger guidelines and information on reducing risk, generally using one or more danger symbols accompanied by standard wording describing, very clearly and in terms comprehensible to all, the risk associated with a possible use of the chemical substance, and the measures to be adopted to prevent exposure and reduce the consequences of accidents.

These classification systems generally tend to cover all possible effects of a substance, and are intended to inform users directly through the labeling. More specialized types of classification exist, aimed at specific end points, e.g., carcinogenesis; these classifications are drawn up by national or international agencies, (e.g., EPA, IARC) and are regularly updated, with the aim of protecting human health, particularly that of exposed workers.

Besides these "specific" classifications, there are the far more widely used "generic" classifications. The two systems that are most widely used are:

1) The UN classification system for the transport of dangerous goods
2) The EU classification system for the introduction of dangerous substances onto the market

3.3.1. Transportation

National or international transport of goods can be by air, rail, road, sea, or inland waterways. For each of these modes of transport, standards authorities have issued international norms regulating the technical details of the transport of dangerous goods. The term "goods" is extremely general, but covers both substances and dangerous preparations. Such norms (e.g., RID for rail transport, ADR for road transport) differ from one another as regards the packaging characteristics and the safety measures to be adopted, because the particular features of each type of transport have to be taken into account. However, as regards the criteria for classifying goods, they all follow the system defined by the UN and published in the "Orange Book" [4].

The Orange Book not only sets out classification criteria for transport, but for each class lists and regularly updates the products officially recognized as belonging to the class; such lists are not exhaustive, but they are supplemented by guideline criteria which increase their scope. Such guideline criteria cover a total of nine danger classes, these are treated in detail in →Transport, Handling, and Storage, Chap. 3.

Each of the nine classes is characterized by a specific danger symbol, applied to the external packaging of the dangerous product being transported (Table 1). The general criteria are then defined in more detail according to the technical demands of each type of transport.

Table 1. Danger symbols for transport of dangerous goods

Class 1

Explosives
Symbol (exploding bomb): black
background: orange

Class 1—*Division 1.4*
(*Except safety explosives*)

Class 1—*Division 1.5*

Explosives
Background: orange; figures: black; numerals must be ca. 30 mm in height and ca. 5 mm wide (for a label measuring 10 × 10 cm)

Class 2

Compressed Nonflammable Gases
Included in Class 2
Symbol (gas cylinder): black; background: green

Class 3

Flammable Liquids
Symbol (flame): black;
background: red

Class 4—*Division 4.1*

Flammable Solids
Symbol (flame): black;
background: white with vertical red stripes

Class 4—*Division 4.2*

Substances Liable to Spontaneous Combustion
Symbol (flame): black;
background: upper half white, lower half red

Class 4—*Division 4.3*

Substances which, in Contact with Water, emit Flammable Gases
Symbol (flame): black;
background: blue

Class 5

Oxidizing Substances; Organic Peroxides
Symbol (flame over circle): black;
background: yellow

Class 6—*Division 6.1*

Poisonous (toxic) Substances
Danger groups: I and II
Symbol (skull and cross-bones): black; background: white

Class 6—*Division 6.1*

Poisonous (toxic) Substances
Danger group: III
The bottom part of the label should bear the inscription:
Harmful stow away from foodstuffs
Symbol (St. Andrew's cross over an ear of wheat): black;
background: white

Class 6—*Division 6.2*

Danger Infectious Substances
The bottom part of the label should bear: INFECTIOUS SUBSTANCE (Optional) and the inscription: **In case of damage or leakage immediately notify public health authority** (optional)
Symbol (three crescents superimposed on a circle) and inscription: black;
background: white

Table 1. (Continued)

Class 7

Radioactive Substances
(a) Category "white"; background: white with one vertical red stripe in the bottom half. Text (mandatory) black in bottom half of label: **Principal radioactive content . . . Activity of contents . . . curies . . .**
Symbol (trefoil): black

Radioactive Substances
(b) Category "yellow"; background: top half yellow, with two or three red stripes in the bottom half white. Text (mandatory) black in bottom half of label: **Principal radioactive content . . . Activity of contents . . . curies . . . Transport index . . .**
Symbol (trefoil): black

Class 8

Corrosives
Symbol (acids spilling from two glass vessels and attacking a hand and a metal): black; background: upper half white, lower half black with white border

The importance of the transport classification is also due to the fact that, in some countries, it is the only classification system and is therefore used, apart from the specific field of transport, as a generic characterization method.

3.3.2. The EU Classification for the Marketing of Dangerous Substances

In the European Union, a system of classification has been developed since the 1970s, covering the labeling to be applied when dangerous substances and preparations are marketed; this includes both professional and domestic use. In the EU, classification norms and labeling for transport are carefully distinguished from those relating to commercial use; e.g., a container or crate intended for transporting tins of paint is labeled according to the norms for transport, whereas the tins of paint are labeled according to the norms for commercial use.

Only in cases involving a common single form of packaging, which is used for both transport and marketing, is it permitted to use a mixed label, consisting of the danger symbol specified by the transport norm and the danger wording and safety instruction wording specified by the marketing norm. EEC Directive 67/548 [5] presents the Community's legal basis for the whole range of norms issued for classification and labeling of dangerous substances and preparations. A group of experts from the Member States is engaged in classifying substances considered to be a priority at the community level; ca. 1500 substances have been officially classified, as reported in Commission Directive 93/73 [6]. Since the basic directive, in an updated version [7], also provides that substances not officially classified should be evaluated by the manufacturer or importer for the purposes of possible classification and labeling (regarded as provisional), the Community system has to clarify the criteria for danger classification in order to provide industry with a clear, unambiguous framework that avoids as far as possible subjective evaluation and differing interpretation of the data. The most up-to-date version (1998) is Commission Directive 98/98/EG [8]. Application of these criteria enables substances to be

allocated to one or more of the following danger categories:

Explosive
Extremely flammable
Readily flammable
Flammable
Oxidizing
Very toxic
Toxic
Harmful
Corrosive
Irritant
Sensitizing
Carcinogenic
Mutagenic
Toxic to reproduction
Hazardous to the environment

A substance can be characterized by a maximum of three danger symbols; one for physicochemical effects, one for biological effects, and one for environmental effects. Figure 2 shows the 10 danger symbols used to identify the 15 danger categories enumerated above.

There is no exact correspondence between the danger categories and the danger symbols, in the sense that in several cases the same symbol is used to characterize more than one danger category (e.g., the skull-and-crossbones used for very toxic, toxic, carcinogenic, mutagenic, and substances toxic to reproduction; or the St. Andrew's cross for harmful substances, irritants, substances producing sensitization on inhalation, Category 3 carcinogens, Category 3 mutagens, and Category 3 substances toxic to reproduction). The symbol thus provides only preliminary generic information on the type of risk; the real danger of the substance is indicated only by the danger indications and risk phrases (R). The safety instructions (S phrases) provide information on the specific precautions to be taken during handling, and first-aid measures in case of accidental contamination or ingestion.

The classification criteria provide a way of utilizing the physicochemical, toxicological, and ecological data on a substance most effectively for assigning danger symbols, risk and safety phrases.

Classification Based on Physicochemical Effects. A substance is classified as explosive if it falls within at least one of the three criteria: shock, friction, or heating. A substance is extremely flammable if it is a liquid with flash point $<0\,^{\circ}C$ and boiling point $\leq 35\,^{\circ}C$, or if it is gaseous and becomes flammable on contact with air at ambient temperature and pressure. A substance is highly flammable if it is a liquid with flash point $0-21\,^{\circ}C$, or if it is a solid capable of igniting and continuing to burn after the source of ignition is removed. A substance is flammable if it is a liquid with flash point $0-55\,^{\circ}C$.

A substance is an oxidizing agent if it responds positively to a specific test designed to check its potential to increase flame propagation velocity in a sample of combustible substance.

Classification Based on Biological Effects. With regard to the lethal acute effects the LD_{50} or LC_{50} values for the appropriate species are used; normally, oral and inhalatory studies on rats and cutaneous studies on rabbits or rats are preferred.

Table 2 gives reference values for danger classification and, within the scope of such a classification, the distinction between the categories of very toxic, toxic, and harmful. For the inhalatory route, two separate sets of reference values are given, depending on the physical state of the substance; in the majority of cases it is practically impossible with powdery solids and high boiling point liquids to prepare suspensions or aerosols of concentrations >5 mg/L.

In order to reduce the number of animals subjected to experimentation, the "fixed-dose" method has recently been developed for the determination of acute toxicity, based on determining not the lethal dose, but the "discriminating dose", i.e., the dose that produces toxic, but not lethal effects. Reference values (mg/kg body weight) are:

Very toxic	<5
Toxic	5
Harmful	50 – 500

As regards long-term effects, a substance is classified as toxic or harmful if repeated or prolonged exposure can produce serious damage (evident functional disturbances or morphological changes of toxicological significance) below specified dose levels. With regard to subchronic toxicity studies, i.e., at 90 d, reference values are:

Figure 2. Symbols used to identify danger categories within the EU classification system

Table 2. Reference values for danger classification based on animal tests

Category	LD_{50}, oral in rat, mg/kg	LD_{50}, dermal in rat or rabbit, mg/kg	LC_{50}, (inhalatory) in rat, mg/L for 4 h	
			Aerosols or particulates	Gases and vapors
Very toxic	≤ 25	≤ 50	≤ 0.25	≤ 0.5
Toxic	25 – 200	50 – 400	0.25 – 1	0.5 – 2
Harmful	200 – 2000	400 – 2000	1 – 5	2 – 20

Oral route, rat: $\leq 50\,\mathrm{mg\,kg^{-1}\,d^{-1}}$
Cutaneous route, rat or rabbit: $\leq 100\,\mathrm{mg\,kg^{-1}\,d^{-1}}$
Inhalatory route, rat: ≤ 0.25 mg/L, 6 h per day.

For subacute toxicity studies, i.e., at 28 d, these values are multiplied by 3.

In a chronic toxicity study, which normally lasts the whole life of the animal (2 years for rats), it is not possible to give individual refer-

ence values, and the evaluation is made case by case.

When an individual reference value is given, it is possible to distinguish between toxic and harmful effects from the dose level at which the toxic effect is observed; if it is slightly below the reference value, the substance is classified as harmful. However, if it is significantly below this value the substance can be classified as toxic. The EU classification reserves the category of very toxic for acute lethal or nonlethal effects.

With regard to local effects (corrosive and irritant) criteria are defined for assigning specific risk wording depending on the type of effect and its severity. For corrosive effects, it is possible to use two different risk wordings, depending on the time it takes a substance to destroy the whole thickness of the cutaneous tissue (between 3 min and 4 h).

A substance is classified as a skin irritant if it produces significant cutaneous inflammation that lasts for at least 24 h following a maximum exposure time of 4 h; such inflammation is regarded as substantial if the scores obtained separately for the occurrence of erythema, eschar, or edema, are greater than the arbitrarily specified limiting values.

With regard to effects on the eyes, a distinction may be made between serious ocular lesions and ocular irritation, depending on the scores obtained for four specific end points:

Opacity of the cornea
Lesion of the iris
Redness of the conjunctiva
Edema of the conjunctiva

It is also possible to classify a substance as a respiratory tract irritant, but since no experimental models are available to evaluate such an effect on laboratory animals, it has to be based on practical observations mainly connected with the professional use of the substance.

For cutaneous sensitization, a reference method (maximization test) exists based on the guinea pig; a positive response in at least 30 % of the animals tested classified the substance as a skin sensitizing agent. There is no experimental model for sensitization by inhalation: in this case tests have to demonstrate reliably that the substance can produce a sensitization reaction in subjects by inhalation, at a frequency greater than that encountered in a control population.

There are three danger categories with regard to carcinogens:

Category 1: substances known to be carcinogenic in humans on the basis of epidemiological studies
Category 2: substances for which there is sufficient evidence of carcinogenesis in experimental animals (as a precaution, such substances are labeled as being in Category 1)
Category 3: substances to be regarded as suspect with regard to possible carcinogenic effects in humans on account of a positive, but inconclusive response in experimental animals

Three danger categories are defined for mutagenesis:

Category 1: substances known to have mutagenic effects in humans. There is sufficient evidence to establish a causal relationship between human exposure and the occurrence of hereditary changes (up to now no substances fulfilling such criterion are known). Moreover, it is extremely difficult to obtain reliable information from studies on the incidence of mutations in the human population, or on a possible increase in frequency.
Category 2: substances for which there is sufficient evidence to say that it is probable that human exposure will lead to the development of hereditary genetic changes. It is necessary to obtain positive results from studies demonstrating the mutagenic effect on mammalian germ cells in vivo or the mutagenic effect on somatic cells, which together clearly demonstrate that the substance, or one of its metabolites, is capable of reaching the germ cells.
Category 3: substances to be considered as suspect with regard to possible mutagenic effects, based on positive test results that demonstrate the mutagenic effect on mammalian somatic cells in vivo, normally supported by positive results of mutagenesis tests in vitro.

For reproductive toxicity, the effects on both fertility and development must be borne in mind. Three danger categories are defined for each of these effects.

1) Effects on fertility:

Category 1: substances capable of reducing human fertility, as demonstrated by epidemiological studies
Category 2: substances capable of affecting human fertility, based on clear evidence from animal studies of reduced fertility in the absence of toxic effects
Category 3: substances that give rise to concern about possible effects on human fertility, based on animal studies demonstrating a decrease in fertility, but not providing sufficient evidence to justify classification in Category 2

2) Effects on development:
Category 1: substances capable of producing harmful developmental effects in human progeny as shown by epidemiological studies
Category 2: substances that can produce harmful developmental effects in human progeny, based on unequivocal results in suitable animal studies, in the absence of maternal toxicity
Category 3: substances that give rise to concern on account of the possible effects on human progeny based on animal studies showing harmful developmental effects in the absence of maternal toxicity, but not providing sufficient evidence to justify classification in Category 2

Classification Based on Environmental Effects. Criteria have been defined at EU level for classifying chemical substances according to their potential environmental hazard. Such criteria essentially refer to the aquatic compartment, at present considered representative of the whole ecosystem. Three parameters are taken into consideration:

Effects on aquatic species
Bioaccumulation capacity
Degradation

Effects on Aquatic Species. Normally these refer to acute effects, even if under some conditions the long-term toxicity data can modify the classification obtained from the acute effects. The species used are:

Fish (LC_{50}, 96 h);
Daphnia (EC_{50}, 48 h);
Algae (IC_{50}, 72 h);

The cut-off toxicity values that define the various subcategories are 1, 10, and 100 mg/L, intended as concentrations in the aquatic medium.

Bioaccumulation Capacity. Normally, a substance is regarded as capable of bioaccumulation if $\log P_{ow} \geq 3$, unless the experimental bioconcentration factor $BCF \leq 100$.

Degradation. A substance is regarded as rapidly degradable if one of the following criteria is satisfied:

1) In a 28-d degradation study, a 70 % degradation is obtained if the test is based on the measurement of dissolved organic carbon, or 60 % of the theoretical maximum value if the test is based on measuring the loss of oxygen or formation of carbon dioxide. These biodegradation levels should be reached within 10 d of the start of degradation, i.e., from the moment at which 10 % of the substance is degraded.
2) In cases in which only the COD and BOD_5 are available, the ratio between the two $BOD_5/COD \geq 0.5$.
3) Data exist, demonstrating that the substance can be degraded (in a biotic or abiotic manner) in an aquatic environment to levels >70 % within 28 d.

3.4. Hazard Identification of Preparations: Calculation Systems for Different Use Categories (EU Approach)

For intentional mixtures of chemical substances (preparations), the problem of hazard identification may be particularly acute given the widespread commercial availability of chemical products, definable as preparations, intended for a wide variety of uses, including domestic environments, where those particularly at risk (children, the elderly) can come into contact with such products and suffer harmful effects. The possibility of human exposure is extremely variable, depending on the intended use of the product. The potential danger of a preparation depends on the danger characteristics of its components and their concentration in the product. Such a danger can essentially be evaluated in two ways:

1) By subjecting the preparation to a series of experimental tests to determine the toxicological and ecotoxicological properties and physical effects, in the same way as for a substance
2) By employing calculation methods that take account of the danger of the various components and their concentrations, normally expressed in wt % (assuming the absence of synergistic or antagonistic phenomena)

Some categories of preparations, because of their intended use, are subject to extremely stringent regulations that require a series of experimental studies to determine the potential danger to humans and, in some cases, to the environment (e.g., for pharmaceuticals and agricultural pesticides). In the majority of cases, however, evaluation of the potential danger is the responsibility of the manufacturer or importer, who is basically free to choose between the experimental and calculation options, with a clear preference for calculation, given the extremely high cost of performing a set of experimental studies. A series of norms has recently been approved at international level that aim to discourage, or at least minimize, experimental studies on animals, for humanitarian reasons.

In the past, a series of specific technical norms has been defined at EU level for particular areas of use (solvents, paints, inks, pesticides), which besides sanctioning the principle of alternative between biological testing and application of a calculation system, define specific mechanisms for establishing the danger classification of the preparation, depending on the nature and proportions of its components. Directive 88/379/EEC [9] is a first step in the unification and rationalization of such systems to arrive at a single evaluation mechanism regardless of the intended use of the product.

3.4.1. Classification Based on Physical and Chemical Effects

No calculation system exists that enables the flammability, explosiveness, and oxidizing power of a preparation to be evaluated, depending on the presence of components exhibiting such properties. The characteristics of the preparation are largely influenced by the chemical nature of the other components. For example, if a liquid preparation contains 20 – 30 % of a flammable substance, it is likely that the preparation will exhibit such characteristics if the other components are nonflammable liquid organic substances, whereas if the preparation contains a large amount of water it is highly probable that this will inhibit flammability. Therefore, in the case of chemical and physical effects, the preparation should be tested as if it were a substance, and eventually be classified according to the experimental results. Not all preparations are to be tested, only those that contain components exhibiting such properties above specific limiting concentrations.

3.4.2. Classification Based on Biological Effects

Given that it is possible to evaluate biological end points by means of experimental tests on the preparation, it is permitted to use a conventional calculation system. Such a system is based on the assumption that the components of the preparation have previously been evaluated and, if considered dangerous, have been correctly classified and labeled. The starting point of the classification of preparations is therefore classification of its components; if a substance is classified in different ways by different manufacturers and importers, such variation will be reflected in preparations containing that substance.

Another basic feature of the calculation system of Directive 88/379 is that each specific effect is evaluated separately, with the possibility of assigning to a substance not just an overall percentage limit, but various limits depending on the specific effects of the substance, and reflected in the risk wording assigned to the latter. For each substance, at the time of its official classification by the group of EU experts it is possible to define specific percentage limits, depending on the various effects of the substance, which are applied when the substance is present in a preparation. Figure 3 shows a file card relating to acrylonitrile, characterized by various types of effect [6]. The card includes identification data, structural formula, chemical name in the various Community languages, classification, labeling, and the various percentage limits

Cas No 107-13-1

No 608-003-00-4

NOTA D

NOTA E

$CH_2 = CH - CN$

ES : acrilonitrilo

DA : acrylonitril

DE : Acrylnitril

EL : ακρυλονιτρίλιο

EN : acrylonitrile

FR : acrylonitrile

IT : acrilonitrile

NL : acrylnitril

PT : acrilonitrilo

Clasificación, Klassificering, Einstufung, Ταξινόμηση, Classification, Classification, Classificazione, Indeling, Classificação

F ; R 11	Carc. Cat.2 ; R 45	T ; R 23/24/25	Xi ; R 38

Etiquetado, Etikettering, Kennzeichnung, Επισήμανση, Labelling, Étiquetage, Etichettatura, Kenmerken, Rotulagem

F

T

R : 45-11-23/24/25-38

S : 53-16-27-44

Límites de concentración, Koncentrationsgrænser, Konzentrationsgrenzwerte, Ορια συγκέντρωσης, Concentration limits, Limites de concentration, Limite di concentrazione, Concentratiegrenzen, Limites de concentração

C ≥ 20 %	T ; R 45-23/24/25-38
1 % ≤ C < 20 %	T ; R 45-23/24/25
0,2 % ≤ C < 1 %	T ; R 45-20/21/22
0,1 % ≤ C < 0,2 %	T ; R 45

Figure 3. Example of a file card published by the EU, defining classification, labeling and concentration limits for a substance (in this case, acrylonitrile) in preparations

Table 3. Concentration limits for substances as components of preparations

Gas-phase preparation, vol %			Substance classification	Solid or liquid preparation, wt %		
T+	T	Xn		T+	T	Xn
		5	R 20, 21, 22, Xn			25
	5	0.5	R 23, 24, 25, T		25	3
1	0.2	0.02	R 26, 27, 28, T+	7	1	0.1
		5	R 40/20, 21, 22, Xn			10
	5	0.5	R 39/23, 24, 25 T		10	1
1	0.2	0.02	R 39/26, 27, 28, T+	10	1	0.1
		5	R 48/20, 21, 22, Xn			10
	5	0.5	R 48/23, 24, 25, T		10	1
		0.2	R 42, 42/43, Xn			1
	0.1		R 45 carc. 1, 2		0.1	
	0.1		R 46 muta. 1		0.1	
		0.1	R 46 muta. 2			0.1
		1	R 40 muta. 3			1
	0.2		R 60 reprotox. 1 R 61		0.5	
	0.2		R 60 reprotox. 2 R 61		0.5	
		1	R 62 reprotox. 3 R 63			5
Gas-phase preparation, vol %			**Substance classification**	**Solid or liquid preparation, wt %**		
C R 35	C R 34	Xi∗		C R 35	C R 34	Xi∗
1	0.2	0.02 (R 37)	R 35, C	10	5	1 (R 36, 38)
	5	0.5 (R 37)	R 34, C		10	5 (R 36, 38)
		5 (R 36, 37 38)	R 36, 37, 38, Xi			20 (R 36, 37, 38)
		0.5 (R 36)	R 41, Xi			5 (R 36)
		5 (R 41)				10 (R 41)
						1 (R 43)
			R 43, Xi			

∗ Applying R . . . given in parenthesis.

to be adopted when the substance is present in a specific concentration in a preparation.

Only a minority of substances have been assigned specific percentage limits, but Directive 88/379 allows the calculation system to be employed, having defined a series of generic limits (established according to the toxicological significance of each effect and independent of the chemical nature of the substance). Table 3 summarizes the percentage limits to be employed, depending on the classification of the components of the preparation. For gaseous preparations, the percentage figures are preferably given by volume rather than by weight.

3.4.3. Classification Based on Environmental Effects

The Directive does not provide for the possibility of classifying a preparation according to its potential environmental danger. A proposal to introduce such criteria is currently under examination. The aim is to define specific percentage limits depending on the risk phrase assigned to the substances. The possibility of determining experimentally the characteristics of the preparation would be limited to the effects on aquatic species, since it is technically impossible to determine log P_{ow} and biodegradability of a preparation. In order to demonstrate experimentally that the preparation is not toxic to aquatic species, it is necessary to carry out tests on all the three species, fish, *Daphnia*, and algae.

4. Risk Assessment

4.1. General Methodology

Risk assessment comprises one or more of the five steps defined in Chapter 2. The fundamental

principle is that risk depends on the inherent hazardous properties of a substance and the extent of exposure to that substance. Consequently, a risk assessment can be concluded at a relatively early stage if it has been demonstrated that a substance exhibits no hazardous properties or that neither humans nor the environment will be exposed to it. The following scheme illustrates the steps and the relationships between them.

Information gathering

Effects	*Exposure*
Hazard identification; dose (concentration) – response (effect) assessment, if appropriate; toxicokinetics	Human exposure assessment (workers, consumers, via the environment, as appropriate); environmental exposure assessment (water, soil, air, as appropriate)

↓

Risk characterization

Human health	*Environment*
	Evaluation of effects data and comparison with exposure data

↓

Outcome of risk assessment

A risk assessment should be amenable to revision in the light of any new information.

4.2. Qualitative and Quantitative Aspects: Toxicological and Ecotoxicological Properties

The definitions in Chapter 2 of risk characterization and risk estimation illustrate the distinction between the qualitative and quantitative aspects of risk assessment. It is sometimes useful to be able to quantify a risk as a numerical probability; equally, it can sometimes be misleading and often impossible. A qualitative characterization of risk, however, should be included in any assessment in which data on intrinsic properties have been compared with data on exposure.

4.2.1. Hazard Identification

The principal toxicological and ecotoxicological properties which should be considered in a hazard identification are described in Chapter 3. If it can be demonstrated that a substance is not inherently hazardous, it is appropriate to conclude that it poses no risk without proceeding to an exposure assessment.

4.2.2. Dose (Concentration) – Response (Effect) Assessment

This step cannot always be separated from hazard identification. Ideally, it should be a quantitative exercise in which an assessor determines the dose or concentration at or below which an adverse effect is unlikely to occur.

In relation to toxicological properties, such a threshold level can be determined for subacute, subchronic, or chronic toxicity, for reproductive toxicity, and for nongenotoxic carcinogenicity. For most other effects, it is usually impossible or inappropriate to determine such a level, and only a qualitative assessment of a substance's inherent capacity to cause such effects can be made.

In relation to ecotoxicity, particularly aquatic toxicity, it is possible to calculate a predicted no-effect concentration (PNEC) by applying an assessment factor to an experimentally derived toxicity value, such as the LC_{50} or NOEC (see Section 3.1.3). An assessment factor is an expression of the degree of uncertainty in extrapolation from test data on a limited number of species to the environment; it should be adjusted according to the extent and quality of the data.

4.2.3. Exposure Assessment

Ideally, exposure assessment should result in the estimation of an exposure level or predicted environmental concentration (PEC). For chemical substances and products already on the market and in use, it should be possible to measure actual exposure levels and concentrations. When conducting a risk assessment before marketing or use, it is usually necessary to predict exposure on the basis of information such as the properties of the substance, its use pattern, and the expected frequency and duration of exposure. Sometimes it is not possible to achieve a quantitative estimation and a qualitative estimation must be made instead.

Exposure assessment should take account of the three principal human populations (workers, consumers, and those exposed indirectly via the environment) and the three principal environmental compartments (water, soil, and air). A risk assessment can be concluded and a substance deemed to pose no risk if there is evidence that neither humans nor the environment

will be exposed to it. If it can be demonstrated that a particular population or compartment is not likely to be exposed to the substance, exposure assessment and risk characterization for that compartment or population are unnecessary.

4.2.4. Risk Characterization

If dose (concentration) – response (effect) assessment and exposure assessment have been conducted quantitatively, it should be possible to compare a human exposure level with a threshold level for toxicological effects or a PEC with a PNEC and to determine an exposure/effect ratio. If only a qualitative assessment of effects or exposure has been made, the comparison will have to be made in more qualitative terms.

Where a threshold level for toxicological effects has been determined, it may be appropriate to apply an assessment factor to that level before comparing it with the exposure level. The assessment factor should take account of the following elements:

1) Nature and severity of the effect
2) Uncertainties in experimental data, including interspecies variations
3) Different susceptibilities in exposed humans
4) Extent to which exposure can be controlled and monitored in different human populations (workers, consumers, those exposed indirectly)

On the basis of the comparison of effects and exposure, whether quantitative or qualitative, the assessor should establish the degree of concern associated with the substance and whether further action is necessary, possibly including recommendations that further information on effects or exposure be acquired so that the risk assessment can be repeated and the risk characterized more precisely. As indicated above, risk estimation (quantification of the likelihood of adverse effects) is not always possible or appropriate.

4.2.5. Physicochemical Properties

As indicated in Section 3.1.1, flammability, explosiveness, and oxidizing potential are the three physicochemical properties which correspond to direct adverse effects. The procedures for assessing risk in relation to those properties are not entirely analogous to those for toxicity and ecotoxicity. The dose (concentration) – response (effect) step is not applicable. The concept of an exposure level is meaningless in relation to physicochemical effects, so exposure assessment amounts to determining the reasonably foreseeable conditions of use; risk characterization then entails evaluation of the likelihood that an adverse effect will occur under those conditions.

4.3. Risk – Benefit Considerations

The outcome of a risk assessment can be that measures should be taken to reduce the risk posed by a substance. The process of deciding precisely which measures are appropriate falls outside the scope of risk assessment. Socioeconomic and political considerations may influence decisions on risk reduction; it is therefore important that risk assessment, based solely on technical data, should be clearly separated from that process.

This applies especially where the withdrawal of a substance or product from the market is being considered. It may be appropriate to weigh the risk which would thus be removed against the benefits which would be forgone. A substance may be necessary to the safe operation of an industrial process for which no satisfactory alternative is available; withdrawal of that substance to remove a toxic risk identified in the risk assessment may create a new risk to worker safety. In other cases, the benefits of the risk reduction itself may be compared with its costs in order to assess whether they are proportionate; if the costs are judged to outweigh the benefits, it may be decided not to reduce the risk or to employ risk reduction measures (e.g., exposure limits) which stop short of withdrawal of the substance.

The allocation of a financial cost to a risk is very difficult and likely to entail subjective judgment. Decisions not to reduce a risk or to take only limited risk reduction measures imply the concept of "acceptable risk."

4.4. Definition of Acceptable Risk

It is extremely difficult, if not impossible, to achieve a general definition of acceptable risk;

acceptability is inherently subjective and variable. In individual cases, it may be useful to quantify the risk (see Section 4.2.4).

Perception of a risk by those subject to it is a highly significant factor. There is evidence that high risk may be considered acceptable if it is perceived as controllable and dependent on personal choice (e.g., car driving, smoking). The risks due to exposure to toxic substances in chemical products, even if quantified as very low, may be considered unacceptable by the public because they are perceived as being beyond its control.

The level of probable damage to humans and the environment by a substance must be viewed in relation to the advantages which its use offers to society. This is difficult, since the groups who have to bear the risk are not the same as those who draw benefit from the product. An abstract definition of the benefit is not as a rule possible, since such a definition varies greatly between individuals, and also depends on sociopolitical conditions. The process of weighing up risks and benefits, which ultimately leads to a definition of the acceptance of a certain risk, is thus often extremely blurred and determined by political motives. It is frequently the case that bodies are entrusted with this task in which relevant, social groups are represented, but it is not feasible for such an assessment to be carried out for every chemical substance. Conventions are, therefore, often established for this evaluation, but the reasons for the final decision are nevertheless frequently of a purely political nature.

Most experience gained in the field of risk – benefit examination has been in pharmaceuticals, since the benefit can usually be very clearly defined, and it is known that harmful side effects cannot be completely avoided. Things are far more problematical in the industrial chemicals field, substances with irreversible effects and without a NOEC being the most difficult to deal with.

A classic example is formaldehyde, which is known to be capable of causing a sensitizing and carcinogenic effect. Because it is indispensable for technical purposes, it is accepted that below certain statutory limits it may continue to be used, even in consumer articles, although adverse effects cannot be precluded, at least in particularly sensitive persons. As a rule, an attempt is made to focus on risks to which humans or the environment are unavoidably exposed. It was calculated, for example, that the occurrence of ethylene oxide as a micropollutant in cosmetic raw materials does not result in a significant increase in this substance in the human metabolism, and that these traces can be regarded as harmless.

Examples in the ecological field are chelating agents such as ethylenediaminetetraacetic acid (EDTA) and nitrilotriacetic acid (NTA), for which it is feared that heavy metals present in sediments in surface waters may be remobilized. An attempt is being made by monitoring programs to ensure that these substances do not accumulate in surface waters in concentrations at which such an effect might arise.

Since many of these effects are highly speculative and elude quantitative assessment, any risk – benefit examination inevitably remains blurred and open to considerations of political expediency. Bans or restrictions on chemical products, which would result in considerable economic consequences, have to a large extent not been imposed.

4.5. Summary of Risk Assessment

The present state of the scientific art – even in toxicology with its rather long tradition, and even more in ecotoxicology – does not allow a real quantified risk characterization which might be used as basis for the definition of an acceptable risk, as practiced in particular in the United States. Attempts to compare risks, identified and characterized as outlined above, with other risks in normal life are seldom successful because society has a different perception of various risks.

A pragmatic approach is appropriate, and should embrace the following elements:

1) Identifying chemicals for which the risk, due to minimal hazardous effects or to minimal exposure, is obviously negligible
2) Refinement of data bases for which this decision cannot be taken on the basis of available data
3) Quantification of risk characterization, as far as possible case by case by expert judgment, for chemicals which show a significant risk potential
4) Decisions on risk reduction measures by political bodies on the basis of acceptability of

risk, taking account of economic and social considerations

5. Risk Management

Risk management should flow logically from risk assessment, but is separate from it. Risk assessment should have characterized the risks, have identified who or what is at risk, and the levels, sources and pathways of exposure. The function of risk management is to reduce the risks and to mitigate the consequences if an adverse event occurs.

It may be possible to eliminate a particular risk, but more usually a residual risk remains, and the challenge is to reduce this to an acceptable level. Science has a role in defining this level, but socioeconomic and other factors are more important, and outside the scope of this article. Those wishing to explore the subject further will find useful starting points in references [10], [11].

There can be no single system of risk management. The key to successful risk management is the selection of the most appropriate set of tools for particular situations and their acceptance by those causing the risks and by those most likely to be affected by them.

Although this article is principally concerned with products control, there are instances where products are only one source of exposure and the risk assessment has to consider the overall impact arising from all sources. Lead is an example of this; there is concern about the effects, particularly on children, of uptake of lead into the body from a range of sources through various pathways. Products control can reduce exposure but cannot eliminate the problem. A much larger range of measures is needed.

In the context of chemicals control, hazard is an inherent property of a chemical but risk depends on the extent of exposure. Exposure is a function of variable, controllable elements (how the chemical is made, how it is incorporated into a product, and how that product is used).

The more important tools for the management of risks from chemical products are dealt with below.

5.1. Supplying Information (Risk Communication)

This is a fundamental requirement in all strategies for managing risks. Unless all those making, taking, or regulating the risks, i.e., the "stakeholders", are aware of the risks, their consequences, and what they can do to reduce these, the strategy will be flawed. Thus:

1) Those creating risks are often best able to take steps to reduce them. They need to inform themselves about the risks associated with their activities and products, and about the possible consequences of the risks. They also need to know what can be done to reduce the risks and to mitigate the consequences.
2) Those who might be affected are more likely to find the risks acceptable if they know what the risks are and what measures have been taken to minimize them. They also need to know what they themselves can do to reduce the risks and the possible consequences.
3) Those responsible for regulating the activities or products which create the risks need to have as much information as possible relating to the risks and about the tools available to reduce them.

Ensuring safety during the manufacture of chemicals, and products containing them, is primarily a matter for industry and the appropriate regulators: safe use in industry and commerce involves management and workers. Outside the workplace, individual behavior becomes the most important factor, and is greatly influenced by personal knowledge, experience, and perception of the risks.

Much environmental, consumer safety, and worker protection legislation requires those producing, importing, and supplying products to give information on the hazardous properties and risks associated with their products, and there is provision for such information to be passed down the manufacturing and supply chain. This is a feature in much EC, Scandinavian, and North American legislation. The most obvious manifestation is the information that appears on labels or is supplied as safety advice with the product. Labels usually carry information of different kinds and give only very basic information on risks, but they may be the

only means of influencing user behavior, which is often important in determining whether a risk becomes manifest. Effective labeling is very important.

Labeling requirements are generally linked to a classification system. These systems cover a range of hazards, e.g., toxicity, irritancy, flammability, effects on the environment, and set criteria to define degrees of hazard, e.g., very toxic, toxic, harmful. The information required on the label, often in the form of standard "risk phrases," follows from this classification. Packaging requirements, which can be important in reducing risk (e.g., through accidental exposure) may also be linked to this classification, as in the EU. Labels are primarily intended to warn and protect the user but, to an increasing extent, include information on environmental safety, in particular on safe disposal.

Labeling is frequently supplemented by data sheets which provide more detailed information. Chemical manufacturers often supply data sheets to their workers, to their customers who process chemicals into products, and to the emergency services. Legislation in the EU on the safe handling of chemicals requires the provision of information in this form. The preparation of international data sheets, principally for worker protection, is a feature of the International Programme for Chemical Safety, a collaborative program involving the World Health Organization, International Labour Office, and United Nations Environment Programme.

Access to information on the hazardous properties of chemicals and on exposure has improved in recent years in most countries. Some countries (e.g., the United States) have legislation requiring such information to be made available. The EC Directive on access to environmental information [12] should ensure greater availability and easier access within the EU to information held by public authorities. Public interest groups have campaigned for greater freedom of access to information, and often act as alternatives to official or industrial sources of information on risks (often seen as less biased or more independent).

5.2. Codes of Practice

Since the way in which products are used can by very important in terms of protecting human health and the environment, codes of good practice are often developed for the more dangerous substances to help ensure safe use. These give guidance on good practice in the use of products and the safe disposal of any wastes. In some countries, demonstration that such codes have been followed is a valid defense in the event of legal proceedings. Failure to comply with a code may lead to prosecution.

5.3. Product Standards

These are usually intended to ensure that a product delivers the required performance, but they can also reduce the impact on human health and the environment, e.g., by specifying maximum levels of dangerous impurities. Conversely, by specifying a performance which can be met only through the use of a dangerous substance, or by failing to revise obsolete standards which encourage the continuing use of dangerous substances, needless risks can be created or perpetuated.

5.4. Control of Supply

This includes control of production and importation. Prevention of entry into the manufacturing and supply chain is among the most powerful and effective forms of product control. Control of production has been used to control the supply and hence the eventual release of chlorofluorocarbons, but control at the point of sale is more common. The latter is widely used as part of approval schemes for pesticides and pharmaceuticals, etc. (e.g., particularly dangerous substances and formulations may be supplied only to professional users). It is also common in EU Member States to implement EU controls on marketing and use, e.g., wood preservatives containing pentachlorophenol cannot be sold to the public in the EU, and the supply for professional use is strictly limited. Lead-based paints may be supplied only for use on specific historic buildings in the United Kingdom.

Attaching conditions to the supply of a chemical rather than banning it outright can be an effective means of control, while allowing appropriate uses. Many of the pesticides available to the public must be supplied in small packs or in diluted form.

Controls on supply are fairly easy to enforce (suppliers can be readily identified, and most are already subject to other forms of control, e.g., inspection by environmental, consumer protection, or labor protection authorities). However, controlling supply has no impact on products already in the control of users. This is why it is often seen as a weaker form of control than control of use (see Section 5.5).

5.5. Control of Use

Although potentially a more powerful form of control, control of use is more difficult than controlling supply. Control of use by an individual may involve changing behavior or interfering with personal rights, preventing the individual from obtaining the product in the first place is a more effective option. But control of use, if it can be enforced effectively, addresses the control of products already in the possession of the user. Many countries have laws which seek to control use by individuals. To help prevent illegal use or disposal, public information and recovery schemes may be introduced to encourage those holding stocks of the product to return them to the supplier or to a body designated for their safe disposal.

Controlling use by industry and business is more feasible. They can be more readily identified, targeted, and the controls enforced by labor or other inspectors.

5.6. Voluntary Agreements

Regulations can be expensive to develop and implement, and often follow rather than force change. By the time regulation takes effect it is likely that industry has adjusted its processes, withdrawn problematic products, and is marketing alternative products. Voluntary agreements sometimes offer a more cost-effective alternative. Paint manufacturers in the United Kingdom agreed to stop using lead driers in paints intended for home use: it is claimed that the voluntary agreements associated with the Toxic Release Inventory in the United States have significantly reduced releases of chemicals to the environment. All the principal manufacturers or suppliers must be parties to such agreements, which must be clear, set targets and timetables, and include systems for monitoring progress. They should be public documents. Even if voluntary agreements are replaced by regulations, they can provide valuable experience which helps avoid problems in implementing the regulations.

5.7. Economic Instruments: Harnessing Market Forces

There is a growing interest in measures to achieve environmental ends through market mechanisms; these can either replace or complement regulation. There are few examples in the field of chemicals control, but tax differentials on motor fuels containing lead additives or to encourage the use of diesel fuels have been adopted in a number of countries.

The attractions of economic instruments are that they highlight that the "polluter is paying," encourage the market to allocate the substance concerned to more essential uses, and provide an incentive to develop substitutes.

Probably the most significant examples of the power of market forces for environmental benefit do not involve government intervention at all. "Green consumer" campaigns can claim many successes, including reducing the use of chlorofluorocarbons in domestic products. Many retailers have recognized the selling power of "environmentally friendly" products and require their suppliers to produce products which exploit this. The ecolabeling schemes now being introduced in the EU and elsewhere will further enhance this trend.

5.8. Integration of Pollution Control and Chemicals Control

Chemicals control deals with problems arising from chemicals as products–pollution control is concerned with chemicals (and other agents) in waste streams. Insofar as the overall problem is from chemicals released to the environment, chemicals control and pollution control complement one another.

In 1991, the OECD adopted a Council Recommendation on "Integrated Pollution Prevention and Control" which commits Member

States to "take into account the effects of activities and substances on the environment as a whole, and the whole commercial and environmental life cycles of substances when assessing the risks they pose and when developing and implementing controls to limit their release." Based on this Recommendation, the Commission of the European Communities is preparing a Union proposal which should lead to greater coordination of the control of waste streams from the chemical and related industries. It should also facilitate the integration of pollution control and chemicals control. Some European countries, e.g., United Kingdom, The Netherlands, Sweden, have already enacted such legislation, but none have gone as far in that direction as the OECD Recommendation implies.

Where total exposure to a substance is important (e.g., lead) and there is a significant contribution from industrial emissions, more effective control of industrial waste streams must clearly be part of the overall risk management strategy.

5.9. Monitoring Effectiveness

The principal requirement is for systems which monitor the effectiveness of the measures taken. It is important that the initial risk assessment be used to define a baseline against which progress can be assessed, and that risk management strategies provide a framework within which progress can be measured. Such a framework should set clear objectives, targets, and timetables which address questions such as:

– Are the measures reducing or containing the risks?
– By how much and to what timescale?
– What are the costs–to all the stakeholders?
– Does more need to be done or can efforts be relaxed?

It may be possible to measure directly whether or not risks have been reduced, e.g., from accident statistics, measures of mortality and morbidity, recovery of species and ecosystems. In many cases, reduction in exposure of the "at-risk group" will be measurable, implying a reduction in the risk, in other cases reduction of inputs to pathways and environmental compartments (air, water, etc.) will be the best that can be achieved. Monitoring programs such as these will answer the first two questions. Normal financial practices, if set up appropriately, can begin to answer the third question, but it is important that all the costs are exposed, including the costs of not taking action, for all those affected. The various stakeholders may differ in their ability to monitor and calculate their costs, and placing a monetary value on environmental damage or improvement remains a difficult problem for economists (some costs are readily monitored while others remain difficult). Nevertheless, cost considerations cannot be avoided.

The final question can only be answered following an assessment of the kinds of information referred to above. The answers represent the outcome of the "feedback loop" which should be an integral part of all risk management strategies.

6. Ecolabeling

6.1. Philosophy

There has been a growing recognition over the past decade that it is not just for public authorities to set rules for environmental behavior. Companies should include the environment in their strategies, over and above the minimum regulatory requirements. Consumers can also play their part by stimulating (by individual choice) the production and marketing of more environmentally friendly products.

The objectives underlying ecolabeling schemes are generally:

1) To provide purchasers with guidance on choosing products with least impact on the environment
2) To encourage the development of products with less impact on the environment, but which still meet the customers' requirements on costs and effectiveness
3) To harness purchasing power to complement regulatory and other controls aimed at protecting and improving the environment

The German "Blue Angel" (Blauer Engel) scheme, set up in 1978, is an early example of a successful ecolabeling scheme, but in the last few years there has been an upsurge in interest in such schemes. Public awareness of environmental issues has been raised, and more and more of

the public want to play a part in preventing or dealing with environmental problems; industry has come to realize that being seen to be "environmentally friendly" helps to sell products, and governments have recognized that consumer purchasing power can be useful in securing environmental improvement. However, the proliferation of labeling schemes and unsubstantiated claims by manufacturers about the "greenness" of their products may confuse, rather than inform, the public. There has also been concern that national schemes might become barriers to international trade. An EC Regulation was agreed in March 1992 [13] and the members of the Nordic Council, Finland, Norway, and Sweden, agreed to set up a joint scheme in November 1989. In North America, the Canadian Government set up its "Environmental Choice Program" in 1988 which, for the first time, adopted a "cradle-to-grave" approach–a concept subsequently adopted by the EU and by the Nordic countries. This approach requires a life-cycle analysis of the product (the EU scheme uses the indicative assessment matrix, Fig. 4).

All existing schemes have restrictions on the types of product covered, e.g., food, drink, and pharmaceuticals are excluded from the EU scheme. Products falling within a scheme are divided into product categories or groups (e.g., washing machines, paints, detergents) and criteria to evaluate the environmental impact of each category are published. Manufacturers who believe that their products meet the criteria may submit them for approval and, if successful, can display the appropriate label or logo on the product and packaging and use it in advertisements and promotional material.

The criteria for award of an ecolabel can be separated into those of a general nature (e.g., legal health and safety requirements) and those that are specific to a particular kind of product. Chemicals per se are unlikely to qualify for any of the product groups in any of the schemes, but they are components of a great many of the products which are eligible for consideration (e.g., adhesives, paints, detergents), and there are many examples where the chemical content can have a bearing on whether or not a product is eligible for the award of an ecolabel. Under the Nordic scheme, no product can be awarded an ecolabel if it contains substances which do not decompose or which bioaccumulate. The Canadian criteria for adhesives prohibit the award of a label to products containing aromatic hydrocarbons, formaldehyde, or chlorinated solvents. The Nordic criteria for detergents limit the content of phosphates, EDTA, NTA, and sodium perborate to less than 0.2 %, and exclude products containing carcinogenic, allergenic, and teratogenic substances.

6.2. The EU Scheme

The scheme was officially launched on June 30, 1993 with the adoption of the first sets of ecological criteria relating to washing machines and dishwashers.

The objectives of the scheme are:

1) To promote the design, production, marketing, and use of products which have a reduced environmental impact throughout their entire life cycle
2) To provide consumers with better information on the environmental impact or products

All consumer products, in principle, are eligible under the scheme, with the exception of food, drink, and pharmaceuticals. Consumers in this context are defined broadly to include business purchases.

The award of the ecolabel means that successful applicants are permitted to use the official logo (Fig. 5) on their approved products. A consumer faced with a choice of, say, washing machines of different makes can pick out, by means of this logo, those which have been shown to achieve a high standard of environmental performance. The same logo is used regardless of the type of product.

Figure 5. The EU official logo for ecolabeling

Environmental fields	Product life-cycle				
	Pre-production	Production	Distribution (including packaging)	Utilisation	Disposal
Waste relevance					
Soil pollution and degradation					
Water contamination					
Air contamination					
Noise					
Consumption of energy					
Consumption of natural resources					
Effects on eco-systems					

Figure 4. Assessment matrix for the EU ecolabeling scheme

The strength of the label is its European dimension. Once approved by one Member State, it can be used throughout the Community. This avoids having to make an application in every country where a national label exists, and thus avoids time-consuming and costly procedures. More importantly perhaps, the success of the EU scheme avoids a proliferation of national schemes, which might confuse the public, serve as barriers to trade and hamper the development of the Single Market.

The scheme should gradually improve the environmental performance and impact of products by giving recognition to those which are the most favorable to the environment. By providing the consumer with information essential to making a sensible purchasing decision, ecolabeling is beginning to make an effective contribution at EU level in the joint interest of consumers and the environment.

7. References

1. Directive 93/67/EEC of the Commission defining the principles for evaluating the risks for man and the environment of substances notified within the meaning of Directive 67/548/EEC of the Council, July 20, 1993.
2. OECD Guidelines for Testing of Chemicals.
3. Directive 92/69/EEC of the Commission July 31, 1992, regarding the 17th adaption to the technical progress of Directive 67/548/EEC of the Council regarding the harmonization of legislative, regulatory, and administrative provisions relating to the classification, packaging and labeling of dangerous substances, July 31, 1992.
4. Transport of Dangerous Goods – Recommendations prepared by the Committee of Experts on the Transport of Dangerous Goods, United Nations, New York.
5. Directive 67/548/EEC of the Council concerning the harmonization of legislative, regulatory and administrative provisions relating to the classification, packaging, and labeling of dangerous substances, June 6, 1967.
6. Directive 93/72/EEC of the Commission regarding the 19th adaption to the technical progress of Directive 67/548/EEC of the Council concerning the harmonization of legislative, regulatory, and administrative provisions relating to the classification, packaging, and labeling of dangerous substances, Sep. 1, 1993.
7. Directive 92/32/EEC of the Council regarding the 7th adaption of Directive 67/548/EEC concerning the harmonization of legislative, regulatory, and administrative provisions relating to the classification, packaging, and labeling of dangerous substances, April 30, 1992.
8. Directive 98/98/EG of the Commission regarding the 25th adaption to the technical progress of Directive 67/548/EEC of the Council concerning the harmonization of legislative, regulatory, and administrative provisions relating to the classification,

packaging, and labeling of dangerous substances, 1998.
9. Directive 88/379/EEC of the Council concerning the harmonization of legislative, regulatory, and administrative provisions of the Member States relating to the classification, packaging, and labeling of dangerous preparations, June 7, 1988.
10. Health and Safety Executive (ed.): *The Tolerability of Risk from Nuclear Power Stations,* HMSO Publications Centre, London 1992.
11. Royal Society: "Risk; Analysis, Perception and Management," report of a Royal Society Study Group, The Royal Society, London 1992.
12. Directive 90/313/EEC of the Council on the freedom of access to information on the environment, June 23, 1990.
13. Regulation 880/92/EEC of the Council on a Community ecolabel award scheme, April 11, 1992.

Chemical Vapor Deposition (CVD) → Thin Films

My Profile Log

WILEY InterScience

HOME
ABOUT US
CONTACT US
HELP

Home / Chemistry / Industrial Chemistry

Ullmann's Encyclopedia of Industrial Chemistry

Production-Integrated Environmental Protection

Standard Article

Jürgen Wiesner[1], Claus Christ[2], Wolfgang Führer[3], Horst Behre[4], Hubertus Cuppen[5], Michael Lumm[6], Franz-Josef Mais[7], Gerhard Schroeder[8], Ferdinand Senge[9], Dieter Stockburger[10], Ludwig Schmidhammer[11], Gregor Lohrengel[12], Lothar Kerker[13], Hans Regner[14], Ulrich Rothe[15], Volkmar Jordan[16], Bernhard Gutsche[17], Thomas Glarner[18], Konrad Stolzenberg[19], Jörg Talbiersky[20], Cornelis van Os[21], Christopher Higman[22], Roberto De Piaggi[23], Motohisa Miyachi[24], Fumihiko Oda[25], Jun Yonamoto[26], Günther Schumacher[27], Wilhelm Tischer[28]

[1]DECHEMA, Frankfurt/Main, Federal Republic of Germany
[2]Hoechst Aktiengesellschaft, Frankfurt/Main, Federal Republic of Germany
[3]Bayer AG, Leverkusen, Federal Republic of Germany
[4]Bayer AG, Leverkusen, Federal Republic of Germany
[5]Bayer AG, Leverkusen, Federal Republic of Germany
[6]Bayer AG, Leverkusen, Federal Republic of Germany
[7]Bayer AG, Leverkusen, Federal Republic of Germany
[8]Bayer AG, Leverkusen, Federal Republic of Germany
[9]Bayer AG, Leverkusen, Federal Republic of Germany
[10]BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany
[11]Wacker-Chemie GmbH, Burghausen, Federal Republic of Germany
[12]Hüls Aktiengesellschaft, Werksgruppe Herne, Herne, Federal Republic of Germany
[13]Hüls Aktiengesellschaft, Marl, Federal Republic of Germany
[14]Hüls Aktiengesellschaft, Marl, Federal Republic of Germany
[15]Kronos International, Inc., Leverkusen, Federal Republic of Germany
[16]Henkel KGaA, Düsseldorf, Federal Republic of Germany
[17]Henkel KGaA, Düsseldorf, Federal Republic of Germany
[18]F. Hoffmann-La Roche AG, Basel, Switzerland
[19]VFT AG, Castrop-Rauxel, Federal Republic of Germany
[20]VFT AG, Castrop-Rauxel, Federal Republic of Germany
[21]Shell Internationale Chemie Maatschappij B.V., The Hague, The Netherlands
[22]Lurgi Öl-Gas-Chemie GmbH, Frankfurt/Main, Federal Republic of Germany
[23]De Nora Permelec S.p.A., Milan, Italy
[24]Mitsubishi Chemical Corporation, Tokyo, Japan
[25]Mitsubishi Chemical Corporation, Tokyo, Japan
[26]Mitsubishi Chemical Corporation, Tokyo, Japan
[27]Boehringer Mannheim GmbH, Werk Penzberg, Penzberg, Federal Republic of Germany
[28]Boehringer Mannheim GmbH, Werk Penzberg, Penzberg, Federal Republic of Germany

Copyright © 2002 by Wiley-VCH Verlag GmbH & Co. KGaA. All rights reserved.
DOI: 10.1002/14356007.b08_213
Article Online Posting Date: June 15, 2000

Recommend to Your Librarian
Save title to My Profile
Email this page
Print this page

BROWSE THIS TITLE
Article Titles A–Z
Topics

SEARCH THIS TITLE
Go
Advanced Product Search
Search All Content
Acronym Finder

Abstract | Full Text: **HTML**

Abstract

The article contains sections titled:

1. Production-Integrated Environmental Protection in the Chemical Industry
1.1. Introduction
1.2. Formation of Residues in Chemical Processes
1.3. Environmental Concepts in the Chemical Industry

1.3.1. The Concept of Integrated Environmental Protection
1.3.2. Environmental Protection in Research and Development
1.3.3. Integrated and Additive Concepts of Environmental Protection
1.3.4. Comparison of Integrated and Additive Environmental Protection
1.4. Limitations of Production-Integrated Environmental Protection
1.4.1. Technical Limitations
1.4.2. Economic Limitations
1.5. Effect of Production-Integrated Environmental Protection
1.6. Costs of Integrated Measures
2. Examples of Production-Integrated Environmental Protection in the Chemical Industry
2.1. Introduction
2.2. Selected Examples
2.2.1. Examples from Hoechst
2.2.1.1. Recovery and Utilization of Residues in the Production of Viscose Staple Fiber
2.2.1.2. Recovery of Methanol and Acetic Acid in Poly(Vinyl Alcohol) Production
2.2.1.3. Acetylation without Contamination of Wastewater
2.2.1.4. Reutilization Plant for Organohalogen Compounds
2.2.1.5. Vacuum Technology for Closed Production Cycles
2.2.1.6. Utilization of Exhaust Gases and Liquid Residues of Chlorination Processes for Production of Clean Hydrochloric Acid
2.2.1.7. Production of Neopentyl Glycol: Higher Yield by Internal Recycling
2.2.1.8. Optimization of Ester Waxoil Production and Recovery of Auxiliary Products
2.2.1.9. Biochemical Production of 7-Aminocephalosporanic Acid
2.2.2. Examples from Bayer
2.2.2.1. Avoidance of Wastewater and Residues in the Production of H Acid (1-Amino-8-hydroxynaphthalene-3,6-disulfonic acid)
2.2.2.2. High-Yield Production of Alkanesulfonates by Means of Membrane Technology
2.2.2.3. Selective Chlorination of Toluene in the *para*-Position
2.2.2.4. Production of Naphthalenedisulfonic Acid with Closed Recycling of Auxiliaries
2.2.2.5. Avoiding Residues in Dye Production by Using Membrane Processes
2.2.2.6. Fuel Replacement in Sewage Sludge Combustion by Utilization of Chlorinated Hydrocarbon Side Products
2.2.3. Examples from BASF
2.2.3.1. Emission Reduction in Industrial Power Plants at Chemical Plant Sites by Means of Optimized Cogeneration
2.2.3.2. Closed-Cycle Wittig Reaction
2.2.4. Integrated Environmental Protection and Energy Saving in the Production of Vinyl Chloride (Example from Wacker Chemie)
2.2.5. Examples from Hüls
2.2.5.1. Integrated Environmental Protection in Cumene Production
2.2.5.2. Production of Acetylene by the Hüls Plasma Arc Process
2.2.6. Low-Residue Process for Titanium Dioxide Production (Example from Kronos International)
2.2.7. Reduction of Waste Production and Energy Consumption in the Production of Fatty Acid Methyl Esters (Example from Henkel)
2.2.8. Integrated Environmental Protection in the Production of Vitamins (Example from F. Hoffmann-La Roche)
2.2.9. Production of Pure Naphthalene without Residues-Replacement of Chemical Purification by Optimized Multiple Crystallization (Example from VFT)
2.2.10. Improvements in the Polypropylene Production Process (Example from Shell)
2.2.11. The Zero-Residue Refinery Using the Shell Gasification Process (Shell - Lurgi Example)
2.2.12. Neutral Salt Splitting with the Use of Hydrogen Depolarized Anodes (HydrinaTechnology, Example from De Nora Permelec)
2.2.13. Ultrapure Isopropanol Purification and Recycling System (Example from Mitsubishi Chemical)
2.2.14. Examples from Boehringer Mannheim
2.2.14.1. Biocatalytic Splitting of Penicillin
2.2.14.2. Production of Diagnostic Reagents by Means of Genetic Engineering: Glucose-6-Phosphate Dehydrogenase and α-Glucosidase
3. Acknowledgement

[Top of Page]

1. Production-Integrated Environmental Protection in the Chemical Industry

Jürgen Wiesner, Claus Christ, Wolfgang Führer, Dieter Stockburger

1.1. Introduction

The business of the chemical industry is to produce and sell chemical products. Profitability is therefore an essential aim of production. However, no chemical process exists that produces only the product desired. Other substances not desired by the producer are also formed in the gas, liquid, or solid state. These are referred to as residues.

Another equally important aim is to reduce the environmental effects of these residues to a level that is acceptable. Therefore, protection of the environment — both as an objective in its own right and as a duty involved in any forward-looking activity — is a constituent part of the managerial policy of the German chemical industry [1]. The European chemical industry has also committed itself to a strategy of environmental protection within the program of Responsible Care. The chemical industry endeavors to operate the production process so as to minimize the entry of residues and waste products into the environment, both quantitatively and with respect to their hazard potential, and to utilize raw materials and energy with maximum efficiency. This growing awareness of the environment follows not only from the principle that the chemical industry is responsible for its actions, but also from the changed conditions governing the introduction of new production methods and the operation of existing plants. These include the following:

1. The demanding international and national decrees and regulations covering environmental protection
2. In chemical companies, the correct allocation of costs required for “cleaning up” a given production unit to comply with regulations (e.g., the purification of waste gas and wastewater and the disposal of waste materials from this unit)
3. Difficulties in the disposal of waste materials due to the shortage of dumping space and the secondary costs of waste disposal plants
4. Costs of raw materials and energy
5. Increased public awareness of the importance of environmental protection
6. In addition, the processes and production plants of chemical technology must meet high standards of workplace protection and operational safety (⟶ Plant and Process Safety)

1.2. Formation of Residues in Chemical Processes

Chemical reactions do not lead exclusively to the desired product but also to residues which may pollute the environment. To understand the reasons for the formation of residues in chemical production processes, and to understand the possibilities for modifying these processes, the general conditions for chemical reactions must be considered. As an example an equilibrium reaction is considered in which a starting material A reacts with a reaction partner B to give a product P:

$$\mathrm{A + B + N \xrightarrow{M,C,H,E} P + P' + P'' + N'}$$

where A is the starting material, B the reaction partner, N a secondary constituent of A and B, N′ the reacted secondary constituent, M the reaction medium, C a catalyst, H an auxiliary, P′ a coupled product, E the energy, P the product, and P″ a side product or byproduct.

The term “residues” denotes all components that take part in the reaction and do not give the desired end product P.

Incomplete Conversion. In equilibrium reactions, components A and B do not react with each other to completion. One of the two substances A or B is generally used in excess to increase the extent of conversion to the other at equilibrium. Occasionally, the component present in excess is the reaction medium for the chemical conversion. The excess amount of A or B remaining at the end of the reaction can contain a higher or lower content of impurities, and must be disposed of in an environmentally friendly manner if it cannot be recycled directly after physical or chemical treatment.

Associated Substances or Impurities in Raw Materials. The appearance of byproducts or residues in chemical reactions can also be due to associated substances or impurities present in the starting materials. This applies especially to the winning of inorganic and organic basic materials from mineral, fossil, or biological raw materials. Examples of associated substances include the following:

1. Hydrogen sulfide in the processing of sour natural gases and mineral oils
2. Sulfur dioxide and dust emissions in the production of acetylene and synthesis gas by the formerly widely used coal-based processes
3. Contamination of wastewater by dissolved decomposition products of lignin (lignosulfonic acids, etc.) in the production of pulp from wood or other vegetable raw materials

Coupled Products. Another characteristic of many syntheses is the formation of coupled products P′, the quantity of which is coupled to the quantity of the main product P.

In inorganic syntheses, for example, the type of coupled product can depend on the form of the raw material component A in the available mineral or ore. For example, when sulfidic nonferrous ores are treated in pyrometallurgical processes, SO_2 is formed as the coupled product, and the coupled product $FeSO_4$ is formed during the digestion of ilmenite ($FeTiO_3$) in the production of the white pigment TiO_2 by the sulfate process. In the production of phosphoric acid from phosphate rock

by the sulfuric acid process the calcium and fluorine content of apatite even leads to two coupled products, namely, gypsum and hydrogen fluoride.

In organic syntheses, starting material A often contains a reactive group that is split off in the form of a coupled product during the reaction. In many cases, halogens or sulfonic acid groups are eliminated, forming acids or salts, and are replaced by other substituents. Also, selective reduction of organic compounds in aqueous solution by metals such as iron or zinc leads to formation of the corresponding metal salts or oxides as coupled products.

Byproducts or Products of Secondary Reactions. Byproducts P″ also reduce the yield of product and must be removed during the course of reaction or during processing of the reaction mixture. They can be formed by continuance of the reaction; e.g., in chlorination processes intended to produce monochlorinated compounds, higher chlorinated products can also be formed.

Reaction Medium. In chemical reactions and separation processes in the liquid phase (e.g., organic solvents or water), solutions or multiphase systems with at least one liquid component are present. The reaction medium or process medium M in these processes is either a solvent, and therefore not a contributor to the reaction, or both a solvent and a reaction partner, and is sometimes a catalyst.

Concentrated mineral acids (e.g., sulfuric acid or oleum) are often used in excess for digestion of inorganic raw materials or for production of reactive intermediates from relatively unreactive organic starting products. That proportion of acid not required for product formation must be either regenerated or disposed of in an environmentally friendly manner. Both routes are costly. Reprocessing is particularly difficult if the acid must be diluted or partially neutralized, or if salts have to be added to the solution in order to recover the product.

When water is used as the reaction medium, wastewater is produced that is sometimes highly contaminated with organic or inorganic substances.

Catalysts can also contribute to the formation of residues. Heterogeneous catalysts must be replaced when their activity is exhausted. In the case of homogeneous catalysts, either the catalyst component is removed along with the used process medium, or an additional processing stage is required to separate the used catalyst from the product.

Auxiliaries can be necessary in some reactions. Like catalysts, these take part in the process without being used up in the chemical reaction. For example, multiphase reactions sometimes require the use of surface-active agents to increase the interfacial surface area or accelerate the transfer of matter. These substances appear as residues or in wastewater when the reaction mixture is processed.

Process Energy. The process energy (electricity, heat) required must also be included in the environmental assessment of chemical processes. Generation of this energy from primary energy carriers also leads to emissions.

1.3. Environmental Concepts in the Chemical Industry

1.3.1. *The Concept of Integrated Environmental Protection*

In Germany, the concept of integrated environmental protection was introduced into the debate on environmental policy at the end of the 1970s [2]. It is being used to an increasing extent and is regarded as a mark of good environmental practice (e.g., in the report on the environment produced by the German Ministry of the Environment [3]). The term is defined differently in various social sectors. The following definitions are given as examples:

German Ministry for the Environment [4], [5]. Environmental protection must be integrated into all phases of the production cycle, i.e.,

1. Reduction or avoidance of harmful materials in the production process
2. Use of internal recycling and increase in energy efficiency in the production process
3. Development of alternative materials
4. Inclusion of the question of disposal in design of the product
5. Establishment of external recycling of production residues

Industrial Management [6]. The following measures should be put into effect:

1. Avoidance of additive end-of-the-pipe techniques where possible — that means process design that avoids or reduces residues
2. Control over the environmental effects of pre-stages and subsequent stages of the process
3. Cooperation with suppliers, consumers, and waste disposal agencies

Matter and energy balances for the process itself and its pre-stages and subsequent steps are taken into consideration when setting up these measures.

Institut für ökologische Wirtschaftsforschung (Institute for Ecological Economic Research), Berlin [7]. Integrated environmental protection techniques either do not cause environmental pollution compared with traditional techniques or do so to a lesser extent. Recycling is regarded as an integrated environmental protection method only if it is an integral part of the production process, i.e., if a closed- cycle operation is used that reduces emissions. If recycling occurs separately from the production process, it is regarded as an additive technique.

Statistisches Bundesamt (German Federal Bureau of Statistics) [8]. The term “integrated environmental protection” is not used in statistics. Instead, distinction is made according to those parts of fixed assets that are concerned with

environmental protection. The corresponding costs of a process change are also included in the statistics.

European Union [9]. The term “integrated techniques” includes

1. Clean production techniques
2. Processes based on materials that result in low environmental pollution
3. Reprocessing and recycling systems

Other Related Terms [10]. The terms “clean production processes,” “clean production,” and “clean technology” are used predominantly in English-speaking countries. *Clean production* goes to the roots of environmental problems, and minimizes both emissions and costs. *Clean technology* is more than a “set of clean production techniques;” rather it is an approach to provide users with services and other benefits enabling the environment to be protected from excessive pollution.

The following terms should also be mentioned: “best available technology” and “nonhazardous production.” However, none of these terms have been clearly defined or distinguished from each other.

Verband der Chemischen Industrie (VCI, Association of the German Chemical Industry) [11]. In integrated environmental protection, the aim is to devise a process that causes as little pollution of air, water, and soil as possible, and in which integrated production methods enable residues to be utilized as far as possible. The technical and economic aims of the production process must nevertheless be fulfilled. This definition is also used by the Swiss chemical industry [12].

1.3.2. *Environmental Protection in Research and Development*

Chemical research is the precursor of production. Its aim is to devise solutions to problems that are valid for the long term. The total time required for research, process development, and planning and designing a plant (i.e., until the start of production) is at least six years. New developments must therefore be planned so that even after this time they are still state of the art. Moreover, they are future-oriented only if they take into account the requirements of environmental protection. This means that the environmental protection technology must also be future-oriented and gives extra impetus to developments in the field of environmental protection. Innovation in production methods is therefore an essential precondition for production-integrated environmental protection.

Improvements in environmental protection require capital investment. The introduction of new processes and new production plants always provides an opportunity for further improvement of environmental parameters. Capital investment must therefore be very carefully planned from an environmental point of view. This is ensured by a system of checklists that accompanies a new process from the research stage to the granting of a license by the authorities. At an early stage, in addition to substance data, environmental data are included in evaluation of the process; these are extended as the process is developed stage by stage [13], [14]. In this way, potential environmental problems can be recognized and dealt with at an early stage. The process must be optimized with respect to the amount of residues produced, which should be as low as possible. The possibility of using production methods that lead to smaller quantities of residues or less environmental pollution must therefore be investigated. The result of this investigation can also be abandonment of a project.

Thus, production-based environmental measures are an integral part of research, development, and planning.

1.3.3. *Integrated and Additive Concepts of Environmental Protection*

The environmental demands placed on the production process can be met with the aid of integrated or additive (end-of-the-pipe [15]) concepts.

Production-Integrated Environmental Protection. According to the definition of the VCI (see Production-Integrated Environmental Protection.), integrated environmental protection is production-linked, and that definition is used in this article. Thus, production-integrated environmental protection means measures taken to reduce, prevent, and utilize residues [12], [16-19].

The reduction and prevention of residues can be achieved by

1. Improving the chemical process with the aid of *new synthesis routes*. For example, in the production of aromatic amines, chemical reduction with iron chips is replaced by catalytic reduction by hydrogen.
2. *Shifting the equilibrium*. The use of more favorable reaction conditions can cause the position of the equilibrium to be shifted so that one of the two components A or B is almost 100 % reacted. This can be achieved by using the second component in excess, by removing the product, or by using more favorable temperature or pressure.
3. *Improving selectivity.* A very effective method of reducing the amounts of residues and improving the yield is to increase the selectivity of the chemical reaction. Examples of this include the following:
   - Improvement of the selectivity of catalysts, e.g., by using catalysts that lower the rate of an undesired side reaction
   - Maintenance of high catalytic activity, e.g., by avoiding contact poisons or by developing simple reaction methods
   - Optimization of reaction conditions, e.g., by utilizing differences in the reaction kinetics of the main reaction and the side reaction, more favorable temperature profiles and residence times, or more suitable reactors
   - Recycling of the side product (if the side reaction is reversible)
4. Developing new catalysts, e.g., in production of polypropylene without generation of wastewater using improved

metal – organic catalysts.
5. Process optimization.
6. Changing the reaction medium. If water is replaced by an organic solvent in syntheses, contamination of wastewater can often be drastically reduced. However, environmentally friendly solvent handling involves not only recovery of the solvent from liquid media but also prevention of losses to the atmosphere during storage, transport, production, and subsequent processing. This can be achieved, for example, by adsorptive recovery of solvents from the gas stream.
7. Using raw materials of higher purity.
8. Replacing or eliminating auxiliaries that have a harmful effect on the environment (e.g., chlorinated hydrocarbons).

Gaseous, liquid, or solid residues whose formation during the production process cannot be avoided, even under optimum operation conditions, can often be reused (see. Fig. 1) by methods such as

1. *Internal utilization*, e.g., of the auxiliaries employed in the process, by processing and recycling directly into the process. In the simplest case, the substance can be used directly after separation from the products or process streams (recovery of volatile components of solvents). In other cases, a physical or chemical processing stage is necessary to remove impurities from the recycled components or convert them into a reusable form. Examples include
   - Recovery of sulfuric acid from waste sulfuric acid by concentration or decomposition and reprocessing
   - Recovery of organic solvents from solvent residues, solvent – product mixtures, aqueous solutions, and production residues
   - Thermal decomposition of residues of chlorination processes to give pure hydrochloric acid

2. *External utilization* of the residue, i.e., as a raw material for the manufacture of other products in a separate production plant.

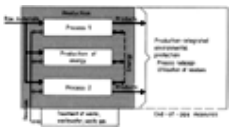

Figure 1. Approaches to environmental protection in chemical production

Reutilization of residues by linking of production processes is not another form of residue disposal, but enables resources to be used as economically as possible. However, this does lead to additional interdependencies, which may decrease flexibility.

Additive Environmental Protection. If the technical and economic possibilities for prevent-ing, avoiding, or reusing residues (as part of a production-integrated environmental protection program) are exhausted, additive environmental technologies must be used:

1. Disposal of waste by land-filling and incineration
2. Techniques for purifying waste gas and wastewater

Of course, not all problems of environmental protection in the chemical industry can be solved by new process concepts based on the principles of prevention and utilization of residues. In many cases, additive processes are still necessary.

In the past, the main emphasis was on additive environmental protection employing conventional purification and disposal methods. Nevertheless, methods of recovery and utilization have long been used in the chemical industry, and can even be found at the beginning of industrial production [19].

In the chemical industry a long tradition exists of reprocessing reaction mixtures (e.g., by solvent extraction or distillation), with recovering of auxiliaries (e.g., organic solvents), and recycling these into the production process (internal recycling). For example, a desired product that is dissolved in a solvent following an extraction process can be recovered only by distilling off the extraction agent. This yields the solvent in a pure form that can be recycled, and such recycling is an inherent part of the production process. It is also a “prevention” (in the sense of the German Federal Pollution Control Act, Bundesimmissionsschutzgesetz) since the residues remain in the production plant [20]. Changes in the production process aimed at increasing the yield or reducing the generation of residues have always been tasks of chemical research. Nevertheless, the technology of integrated environmental protection is becoming increasingly important because current requirements with respect to the prevention and utilization of residues are much more stringent than ever before.

Product-Integrated Environmental Protection. The concept of *product*-integrated environmental protection should be distinguished from that of *production*-integrated environmental protection. The former includes the utilization of product waste (e.g., the recycling of plastics, ⟶ Plastics, Recycling) and the development of environmentally friendly products. These measures are also part of an overall environmental protection strategy of the chemical industry [21]. They form a separate subject that is not discussed here.

1.3.4. *Comparison of Integrated and Additive Environmental Protection*

In additive environmental protection, harmful substances formed in a production process are disposed of at a later stage. There is no saving of raw material; instead, residues are converted into a comparatively environmentally harmless form by the use of additional raw materials and energy.

Additive techniques represent unproductive capital [22]. Moreover, these techniques do not reduce the consumption of raw materials. *Integrated techniques* reduce the consumption of raw materials and the production of waste [23]. However, new integrated techniques can lead to higher energy consumption. Integrated techniques sometimes offer economic advantages, whereas additive techniques with linear emission reduction lead to increased marginal costs or greatly decreased marginal utility [24]. This is shown for three German chemical companies taken as examples. Here, the use of additive environmental protection gave an increase in operating costs (Fig. 2), a reduction of emissions into the atmosphere (Fig. 3) and residual contamination of wastewater (Fig. 4). The extent of purification currently achieved in wastewater treatment plants cannot be improved significantly even at greater expense. The residual contamination of wastewater can be reduced only by techniques of integrated environmental protection. A similar development can also be seen in the case of waste from chemical production (⟶ Waste – Integrated Measures to Reduce Environmental Burdens). The slight reduction in operating costs for environmental protection observed in 1992 and 1993 shows that integrated techniques are to some extent beginning to have an economic effect. However, if the operating costs of environmental protection are expressed as a percentage of value added, a further increase is observed (Fig. 5).

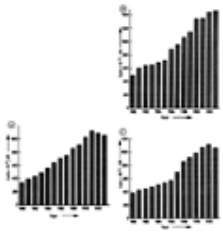

Figure 2. Operating costs for environmental protection

A) Hoechst; B) Bayer; C) BASF

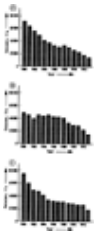

Figure 3. Emissions into the atmosphere

A) Hoechst; B) Bayer; C) BASF

Figure 4. Residual contamination of wastewater determined by COD

A) Hoechst; B) Bayer; C) BASF

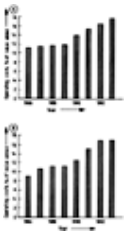

Figure 5. Operating costs of environmental protection in relation to added value

A) Hoechst; B) BASF

In contrast to additive technologies, the costs for operating and managing the new process are sometimes greater in the case of integrated techniques [25], [26]. Additive techniques are widely available and well proven or can be developed and installed in ca. two to four years. Integrated techniques, however, require research, development, and the approval of licensing authorities, which can take six to ten (or even more) years before they are incorporated into the production process [27], since the entire process must often be designed afresh. Administrative barriers of environmental licensing authorities may also hinder the introduction of complex and innovative integrated techniques [22]. On the other hand, additive technologies for the disposal of materials, such as land-filling and waste incineration plants can meet resistance from the public.

1.4. Limitations of Production-Integrated Environmental Protection

1.4.1. *Technical Limitations*

The technology for improving a process by avoiding or reducing residues is not always available. Each process must be researched and developed individually. Additional fundamental knowledge of chemical processes must be acquired, which requires time and money. Even so, the production of residues cannot be prevented completely. The idea of a production process with 100 % yield, i.e., with “zero emission of hazardous materials of any kind [28],” is a vision that can never exist in reality [14], [29].

Also, the utilization of residues (e.g., contaminated solvents or solvent mixtures) may be limited for technical or economic reasons. Some practical examples of this are given below.

Volatile Organic Compounds in Poly (Ethylene Terephthalate) Synthesis [30]. In the production of poly(ethylene terephthalate), PETP, low boiling organics are produced during solvent processing. These consist of a mixture of methanol, methyl esters of aliphatic carboxylic acids, and cyclic ethers. No internal or external method of reusing this material could be found. A process was investigated for hydrolyzing the methyl esters and recovering the methanol.

Waste-Gas Purification by Scrubbing with Glycol Ethers [16]. In the pharmaceutical industry, certain waste-gas streams are purified by scrubbing with glycol ethers. The solvent mixture produced on desorption contains ca. 16 components that are impossible to separate.

Complex Solvent Mixtures [31]. Problems preventing the separation of complex solvent mixtures are summarized below:

1. The azeotropic behavior of solvent mixtures (e.g., methyl acetate – acetone – methanol) can prevent distillative separation
2. The thermal instability of substances dissolved in a solvent mixture can also prevent distillative separation
3. Solvent residues are needed for dissolving partially polymerized organic residues to produce a material of suitable consistency for pumping into an incinerator

1.4.2. *Economic Limitations*

Investment decisions for new production plants or changes in production methods cannot be made without taking into account the existing market situation and the question of future demand for the product.

The quality and profitability of a product must be ensured under given market conditions. In principle, a new production process can be put into practice only if the costs, including the costs of residue disposal, are less than those of the existing process [32-34]. For any investment decision [35],

$$C_{\mathrm{INT}} < C_{\mathrm{ST}} + C_{\mathrm{EOP}}$$

where C_{INT} are the total costs of the new integrated technology; C_{ST} the costs of the standard technology; and C_{EOP} the costs of retrofitting with the additive environmental technology.

If the existing production plant has not yet reached the end of its economic life, the total costs C_{INT} must also include the sunk- cost.

Economic considerations can also limit the utilization of residues. An example is given in the following: In the proposed manufacture of a pharmaceutical product (capital cost: 1.28×10^6 €, ca. \$ 1.6×10^6), a mixture of ethanol, triethylamine, and water is produced. Investigations aimed at producing the ethanol and triethylamine in a purity suitable for utilization showed that recovery of the ethanol and triethylamine would require a special distillation plant consisting of two specially designed columns and a special storage to comply with German regulations for flammable liquids. The capital cost of this plant would be 3.4×10^6 DM (ca. \$ 2.25×10^6). The costs of the recovered solvents would be greater by a factor of 2.8 than the cost of new ethanol and triethylamine plus incineration of the residue mixture in the internal waste incineration plant. The solvent mixture was therefore disposed of by incineration [31].

1.5. Effect of Production-Integrated Environmental Protection

The effect of production-integrated environmental protection is represented by the amounts of residues prevented or utilized. These are compared with the amounts of residues not utilized — i.e., waste products. Figures have been provided by BASF, Bayer, and Hoechst. These are given below:

BASF [36]. In 1987, 750 000 t of residues were utilized in chemical production. Of this, 380 000 t were used for the manufacture of products. The amount of material recycled within a process is considerable (e.g., recycling of raw material not reacted on passing through the reactor or recycling of auxiliaries such as solvents). The quantities certainly exceed the quantities of products sold (1987: 8.3×10^6 t). The quantity of residues utilized by outside companies in 1987 was 110 000 t. These included mainly the following:

1. Calcium carbonate from water treatment and fertilizer manufacture was used as a raw material in the cement and limestone industry
2. Plastic waste was sold to external consumers
3. Chlorinated hydrocarbons were used in the production of hydrochloric acid or as raw materials in the synthesis of perchlorinated alkanes
4. BASF has a waste materials exchange system involving the internal or external supply or sale of byproducts, superseded products, etc.

In the same year, the amount of waste produced was 897 000 t.

Bayer [37]. In the Bayer works, ca. 1.2×10^6 t of production residues were utilized in 1989. Of this material, 80 % was utilized directly, 15 % was recycled to the production operation after chemical conversion, and 5 % was burned to produce energy. The residues were of the following types: waste sulfuric acid 38 %, hydrochloric and other mineral acids 30 %, inorganic residues 15 %, miscellaneous residues 12 %, and flammable residues for energy production 5 %. The amount of waste produced in the same year was 875 000 t.

These figures for BASF and Bayer do not include internally recycled solvents. Therefore, they cannot be compared with the figures given for Hoechst below.

Hoechst [38]. In the Hoechst group in Germany, without special recovery techniques ca. 3.45×10^6 t of potential waste or residues would have been produced in manufacturing operations in 1992. By the introduction of new processes or the improvement of existing processes in earlier years, the formation of ca. 300 000 t of residues and waste materials was prevented. A further 2.68×10^6 t were recycled and utilized, and ca. 470 000 t of residues had to be disposed of as waste. Thus, the fraction of the total residues that were utilized was 85 % (in the solvent residue area: ca. 95 % [30]), and 15 % was disposed of as waste.

The most important residue groups are listed below, together with the recycled amounts in tonnes:

Aliphatic and aromatic solvents by redistillation	779 085
Hydrochloric acid (31 %) by distillative processing	400 746
Hydrochloric acid (100 %) from vinyl chloride production used for oxychlorination	132 600
Chlorinated hydrocarbons by distillation	188 747
Energy-rich residues used as substitute fuels	38 510
Hydrochloric acid (31 %) from incineration of chlorination residues	57 750
Sulfuric acid by distillation reprocessing or as supply to other users	62 534
Carbon disulfide from waste gas	12 320
Na_2SO_4 and $(NH_4)_2SO_4$ from wastewater and spinning baths	106 330
Recovery of acetic acid and acetaldehyde	56 067
Aromatic and aliphatic hydrocarbons	87 791
Activated carbon and carbon residues	86 210
Catalysts and inorganic substances	34 758

1.6. Costs of Integrated Measures

Capital investment in integrated environmental protection as a percentage of total investment for environmental protection is considered by economic experts to be an indicator of the quality of the environmental protection (EP) or an indicator of preventive environmental protection. The following statements have been made based on this indicator [39-42]: A “quantity – quality trade-off ” occurs in the industry.

1. That means that a relative increase in the quantity of capital investment for end-of-pipe instead of production-integrated measures leads to a reduction in the overall quality of environmental protection
2. Capital investment into end-of-pipe technologies does not satisfy the precautionary measures preferred by environmental policy
3. Political support for end-of-pipe measures is short-sighted

All these conclusions are based on an interpretation of data from the Statistisches Bundesamt zur Grundstoff- und Produktionsgüterindustrie (German Federal Bureau of Statistics for the Basic Raw Materials and Capital Goods Industries). This interpretation is as follows [39-42]:

1. Integrated EP capital investment expressed as a percentage of total EP capital investment is ca. 20 %
2. Most capital investment is in end-of-the-pipe and not in integrated technology

However, the figure of 20 % has not been proved for the chemical industry. By using the author's data, the proportion of integrated EP investment in the chemical industry is calculated to be ca. 6 %, but the data cannot be verified in practice.

Several authors [43-45] agree with the above statements. However, these cannot be confirmed because annual statistics from the German Bureau of Statistics do not distinguish between additive and integrated measures [8]. Thus, sufficiently accurate information on such protective measures cannot be obtained from the statistics. This also applies to the proportion of integrated technologies in the total environmental protection capital investments for the Federal Republic of Germany, which was estimated to be 18 % in 1984 [9], [46].

At present, no method exists for calculating the size of capital and operating costs for production-integrated environmental measures. The assessment also fails for new products made by new environmentally-sound processes.

Therefore, figures for integrated EP capital investment and resulting operating costs are not yet available. They are contained in the overall amounts of capital investment and operating costs of the production process. Moreover, the reduction in residues and emissions is a much better indicator of successful environmental protection than the amount of capital hypothetically invested [29].

[Top of Page]

2. Examples of Production-Integrated Environmental Protection in the Chemical Industry

Jürgen Wiesner, Claus Christ, Wolfgang Führer, Horst Behre, Hubertus Cuppen, Michael Lumm, Franz-Josef Mais, Gerhard Schroeder, Ferdinand Senge, Dieter Stockburger, Ludwig Schmidhammer, Gregor Lohrengel, Lothar Kerker, Hans Regner, Ulrich Rothe, Volkmar Jordan,

Bernhard Gutsche, Thomas Glarner, Konrad Stolzenberg, Jörg Talbiersky, Cornelis van Os, Christopher Higman, Roberto De Piaggi, Motohisa Miyachi, Fumihiko Oda, Jun Yonamoto, Günther Schumacher, Wilhelm Tischer

2.1. Introduction

The following examples from the international chemical industry illustrate the concept of production-integrated environmental protection described in Chapter Production-Integrated Environmental Protection in the Chemical Industry. Other publications and company monographs show that the total number of examples of this type in chemistry is actually much larger [12], [17], [18], [30], [31], [47-50].

However, the uneven distribution of these examples with respect to countries shows that so far not all nations place equal emphasis on production-integrated environmental protection in the chemical industry. This observation may be due to the following factors:

1. Differences in legal requirements limiting emissions and residues from production processes
2. The intensity of research required for production-integrated process improvements
3. The close association between production-integrated measures and investment decisions on new production plants or essential modification of processes
4. The usually higher degree of complexity of production-integrated methods compared to conventional processes with additive environmental protection
5. The dependence on the quality of raw materials available when certain components or impurities associated with the raw materials are decisive for emission and residue problems
6. The influence of the operation mode: continuous processes being simpler to optimize than batch processes

The examples given here can be classified in accordance with one or more of the following principles:

1. Prevention or reduction of emissions and residues by modification or optimization of chemical syntheses:
   - Development and use of novel and more selective catalysts Examples: 2.2.2.1, 2.2.2.3, 2.2.7, 2.2.10
   - “Biocatalysis”Examples: 2.2.1.9, 2.2.14.1, 2.2.14.2
   - Kinetic optimization of process conditions Examples: 2.2.5.2, 2.2.7, 2.2.8
   - Increase of product yield by shifting of the equilibrium and product recycling Examples: 2.2.1.7, 2.2.2.2, 2.2.5.1, 2.2.7
   - Nonaqueous synthesis route Examples: 2.2.1.3, 2.2.2.1 (intermediate product), 2.2.10
   - Low-residue digestion of raw material Example: 2.2.6
   - Prevention of byproducts by avoiding associated substances in the raw materials Examples: 2.2.5.2
   - Change to continuous operation Examples: 2.2.1.8, 2.2.2.1, 2.2.7

2. Prevention of emissions by optimized operation techniques
   - Low-emission vacuum technique Example: 2.2.1.5
   - New methods of product separation and purification Examples: 2.2.2.2, 2.2.2.5, 2.2.13

3. Process-integrated recycling of auxiliaries and byproducts Examples: 2.2.1.1, 2.2.1.2, 2.2.1.8, 2.2.2.4
4. Recovery of raw materials or auxiliaries for the production process by reprocessing of residues Examples: 2.2.1.4, 2.2.1.6, 2.2.2.4, 2.2.12, 2.2.13
5. Utilization of residues and coupled products by linking of production processes Examples: 2.2.1.1, 2.2.1.4, 2.2.1.6, 2.2.4, 2.2.11
6. Energy saving and prevention of emissions due to burning of fossil fuels by utilization of process energy, recovery of energy from residues, and higher process efficiency Examples: 2.2.1.3, 2.2.1.8, 2.2.1.9, 2.2.2.6, 2.2.3.1, 2.2.5.2, 2.2.7, 2.2.8, 2.2.9, 2.2.10, 2.2.11, 2.2.14.1, 2.2.14.2

2.2. Selected Examples

2.2.1. *Examples from Hoechst*

2.2.1.1. *Recovery and Utilization of Residues in the Production of Viscose Staple Fiber*

General Features of Viscose Staple Fiber Production (⟶ Cellulose – Viscose Fibers) [51]. Cellulose from wood is the most important starting material for staple fiber production by the viscose process. Cellulose xanthate is obtained by converting cellulose to sodium cellulose followed by treatment with carbon disulfide.

Reaction scheme

1. Production of alkali cellulose
   Cellulose + NaOH ⟶ Alkali cellulose
2. Aging
3. Viscose formation (dissolution of alkali cellulose by addition of carbon disulfide)

$$\text{Alkali cellulose} + CS_2 \rightarrow \text{Cell-O-}\underset{\underset{\text{SNa}}{|}}{\text{C}}\text{=S}$$

Side reaction:

$$2\,CS_2 + 6\,NaOH \rightarrow Na_2CS_3 + Na_2CO_3 + 3\,H_2O + Na_2S$$

4. Ripening
5. Regeneration of cellulose fibers in the H_2SO_4 – Na_2SO_4 – $ZnSO_4$ coagulation bath

$$\text{Cell} - O - CS_2Na + NaHSO_4 \rightarrow Na_2SO_4 + CS_2 + \text{Regenerated cellulose}$$

Side reactions :

$$Na_2CS_3 + H_2SO_4 \rightarrow Na_2SO_4 + H_2S + CS_2$$

$$Na_2S + H_2SO_4 \rightarrow Na_2SO_4 + H_2S$$

$$Na_2CO_3 + H_2SO_4 \rightarrow Na_2SO_4 + H_2O + CO_2$$

Cellulose xanthate is dissolved in aqueous sodium hydroxide solution to give a viscous liquid — the viscose. The sodium hydroxide solu-tion, which is added in excess, also results in the formation of byproducts such as sodium trithiocarbonate and sodium sulfide. In addition sodium carbonate is formed.

The ripening of viscose is stopped after a certain time, depending on the nature of the required end product, and the cellulose is precipitated. This is performed by injecting the filtered spinning solution through fine nozzles (the orifices of the spinneret) into an acid bath. The main reaction occurring in the acid bath is the coagulation of viscose, and CS_2 is formed. In side reactions, salts formed in the preparation of the spinning bath decompose to H_2S, CO_2, and Na_2SO_4 (step 5).

As the strand or slubbing is led out of the spinning bath, subsequent degassing occurs, in which carbon disulfide and hydrogen sulfide dissolved in the acid spinning bath partly evaporate. The spun fibers then pass through several aftertreatment baths. In the course of this, adhering sulfuric acid is washed off. A subsequent degassing of carbon disulfide and hydrogen sulfide also occurs in the wash baths, which are mostly hot.

For production of 1 t of staple fiber, up to 350 kg of CS_2 is required. About 25 % of the CS_2 decomposes to H_2S and CO_2. The schematic of the entire process is shown in Figure 6.

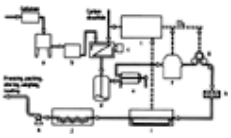

Figure 6. Schematic of viscose staple fiber production

a) Mashing; b) Ripening; c) Sulfidization; d) Dissolving; e) Filtration; f) Spinning; g) Stretching; h) Cutting; i) Washing and aftertreatment; j) Drying; k) Transport; l) CS_2 recovery

The Problem. During the production of staple fibers by the viscose process, toxic *waste gases* containing carbon disulfide and hydrogen sulfide are formed. In the conventional process, those waste-gas streams that contain mainly carbon disulfide were collected in a waste-gas system with associated recovery. These waste gases arise during deaeration and exhaustion in viscose production and during aftertreatment of the viscose staple tow. In the recovery plant, after the separation of the hydrogen sulfide fraction, carbon disulfide was adsorbed on activated carbon. Up to about 65 % of the carbon disulfide used was recovered.

The greater part of the other sulfur-containing waste gases — ca. 60 – 90 kg of sulfur in the form of CS_2 and H_2S per tonne of staple fiber — previously entered the waste air. This part originated mainly from the spinning process and also from the spinning bath components discharged with the fibers, from which the gases are released by diffusion in the later treatment stages. In the original procedure, the waste-gas sources in the process had to be exhausted with very large quantities of air to keep waste-gas concentrations safely below the in-plant threshold limit. The specific quantity of waste gas was therefore extraordinarily high (ca. 235 000 m^3/h at an output of 55 000 t of viscose staple fiber per year). The extracted gases, greatly diluted, reached the open air via a tall stack. As a result, nuisance odors often occurred in the surroundings of viscose staple fiber plants because the threshold odor concentration of hydrogen sulfide, at ca. 0.005 mg/m^3, is extraordinarily low. Disposal of the waste gas failed because of its large volume and low concentration.

Considerable quantities of *sodium sulfate* and *zinc sulfate* are also formed during regeneration of cellulose fibers. Part of the sodium and zinc sulfates is discharged from the spinning bath together with the viscose fibers and enters the wastewater through the aftertreatment of the fibers.

Solution [51-54]. To solve the waste-gas problem, a recycling scheme was developed by Hoechst and Süd-Chemie, Kelheim/Germany, and realized in 1975 – 1979. In this process, 80 % of the viscose waste gases were collected. In the following years, further increase of collection to 90 % was achieved, and the waste-gas volume was simultaneously reduced considerably. In the recycling scheme, sulfur-containing waste gases from viscose staple fiber production (Hoechst) are supplied to a neighboring sulfuric acid plant (Süd-Chemie) where they are used as a raw material or as

combustion air. The sulfur compounds contained in the waste gases are combusted to SO_2, CO_2, and H_2O, and the SO_2 is processed to H_2SO_4. The sulfuric acid produced can be used again directly for the coagulation bath. The new process concept required substantial changes and new developments in both plants. In addition it provided the preconditions for improved recovery of carbon disulfide in the viscose plant.

New Concept of Waste-Gas Capture during Viscose Staple Fiber Production. As a result of the basic chemical reactions in a sulfuric acid plant and the desired conversion of nearly 100 %, the ratio of the quantity of combustion air to sulfur is fixed. If the entire amount of waste gas formed in a conventional viscose staple fiber plant were to be used as combustion air in a sulfuric acid plant, about 30 times as much sulfuric acid would have to be produced as the staple fiber plant consumes. The quantity of lean waste gas from the viscose plant must therefore first be matched to the much lower combustion air requirement of the sulfuric acid plant. This can be achieved by appropriate reduction of the air exhaustion in the viscose staple fiber plant. However, this is possible only if, at the same time, risks to safety as a result of exceeding the in-plant threshold limit value and of the occurrence of explosive gas – air mixtures are avoided.

For this purpose, waste gases from the various process stages are assembled in two separate systems: waste-gas system 1 (“rich gas”) is used for the collection of combustible but not explosive gases (waste-gas concentration >explosive range), and waste-gas system 2 (“lean gas”) is used for the collection of waste gases of low concentration (waste-gas concentration <explosive range). Extensive changes to the viscose staple fiber plant were necessary to minimize waste-gas emission:

1. The quantity of carbon disulfide required in viscose production was reduced by developing a special aging process.
2. Because of the new concept of collection of waste air from viscose production, the spinning machines, stretching and tow washing, and subsequent waste-air treatment, most of the carbon disulfide can be recovered from the waste gas.
3. To recover hydrogen sulfide from the spent spinning bath acids as completely and at as high a concentration as possible, a vacuum degassing plant was constructed. The concentrated waste-gas stream of ca. 200 m^3/h can be utilized directly in the sulfuric acid plant (rich gas).
4. By careful enclosure of the spinning machines and the apparatus in the acid station and by using novel exhaustion systems, the total waste-air stream of 180 000 m^3/h (STP) has been reduced to 10 000 m^3/h (STP). The entire spinning machine waste air can therefore be routed via the waste-gas purification plant (CS_2, H_2S). Lean gas from the acid station is used as combustion air in the sulfuric acid plant.
5. Finally, an extensive waste-air collection and piping system had to be constructed. Corrosion problems in the lines had to be ruled out by selecting suitable materials.
6. To ensure safety, a series of additional measures was required: e.g., a largely flangeless construction and reduced-pressure mode of operation of the waste-gas system to avoid leaks; installation of a process measuring and control technology with emphasis on safety; and construction of an additional exhaustion system on the spinning machines for manual operating procedures.

Hydrogen Sulfide Recovery. Waste air from the production installations (spinning, stretching, tow washing), which is contaminated with CS_2 and H_2S, is freed from H_2S by a multistage wash with sodium hydroxide solution (Fig. 7, a). In the course of this, an NaSH-containing spent lye is formed. Subsequently, CS_2 contained in the waste gas is adsorbed on activated carbon (b) and then desorbed with steam (Supersorbon process). The desorbate water –CS_2 vapor is precondensed at 60 °C (c). The gaseous residue is then condensed at 15 °C, and the condensate is drained off into the separator (e). The aqueous phase obtained, together with the precondensate, is stripped of CS_2 in a column (d). The CS_2 is returned to the production process.

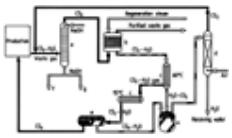

Figure 7. Carbon disulfide recovery in viscose staple fiber production

a) Wash column; b) Adsorption on activated carbon; c) Condenser; d) CS_2 stripper; e) Separator; f) Zinc precipitation; g) Acidification (H_2S used for H_2SO_4 production)

Zinc Recovery. Part of the spent lye from the gas wash (a) is used to precipitate zinc (f) from the stretching bath and from the wastewater of the tow wash. By addition of H_2SO_4 to the precipitated zinc sulfide, zinc sulfate is recovered for the spinning process. During zinc precipitation, a small portion of zinc polysulfide is formed that cannot be converted to zinc sulfate. This portion must be discharged.

Hydrogen Sulfide Utilization. The H_2S content of the remaining spent lye (g) is released by adding H_2SO_4 and utilized in the neighboring sulfuric acid plant together with H_2S from zinc sulfate recovery and waste gas (CS_2, H_2S) from the spinning bath regeneration.

Sodium Sulfate Recovery. In the spinning bath regeneration, the spinning bath is partially evaporated. In the course of this, sodium sulfate is obtained in a marketable form by crystallization. The overall flow sheet of the new process is shown in Figure 8 A.

Figure 8. Residue utilization in viscose staple fiber production

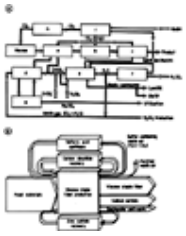

A) a) Spinning; b) Stretching; c) Tow wash; d) Spinning bath regeneration; e) Zn recovery; f) Spent lye degassing; g) Na_2SO_4 crystallization; h) Recovery; i) WashingBWTP = Biological wastewater treatment plant

B) Overall balance of the integrated pollution control solution

Supplementary Engineering Equipment for Process Rearrangement. To be able to utilize the lean gas stream, which is saturated with water vapor, in the sulfuric acid plant, a new wet –dry catalytic process for sulfur combustion was developed jointly by Süd-Chemie and Lurgi. In this process the water vapor from the SO_3 – SO_2 gas mixture is removed after the third catalytic stage by condensation in a new absorption process. The dry SO_3 – SO_2 gas mixture then passes through the fourth catalytic stage. Depending on the degree of dilution of the waste gases (with N_2 and CO_2 as inert constituents), the new process gives a sulfuric acid yield of 99.1 – 99.4 %.

Pollution Control Balance. The overall balance of the integrated pollution control is shown in Figure 8 B. As a result of collecting the waste-air streams, ca. 90 % of the original emissions of sulfur compounds from the plant can be avoided, corresponding to ca. 5000 t/a (calculated as S). Only the waste air arising during manual operations on the spinning machines, which is greatly diluted as a result of the additional exhaustion, must be discharged from the stack as before. Apart from rare plant upsets, the nuisance odor in the surroundings of the plant is a thing of the past.

Moreover, during spinning bath regeneration, about 40 000 t/a of anhydrous sodium sulfate is recovered. This corresponds to about one-half of the amount formed in the overall viscose process (including neutralization steps). The remaining wastewater enters a Biohochreactor (⟶ Biochemical Engineering) for degradation of the organic materials (dissolved cellulose ingredients).

The solution found here to the viscose waste-gas problem dispenses with additional energy requirement in the form of electric power, fuel gas, or natural gas. No new waste materials are produced. The operating costs of the process depend basically on the market price of sulfuric acid. The process scheme has by now been applied internationally in several viscose plants.

2.2.1.2. *Recovery of Methanol and Acetic Acid in Poly(Vinyl Alcohol) Production [16], [55]*

In the production of poly(vinyl alcohol) (Mowiol; ⟶ Polyvinyl Compounds, Others – Poly(Vinyl Alcohol)), a mother liquor containing methanol and methyl acetate is produced. This mother liquor is converted to acetic acid and methanol in a multistage process; these substances are then recycled to the production process.

In the first stage, the mother liquor is subjected to distillation. An azeotropic mixture of methanol and methyl acetate is obtained as overhead product. The methanol – water mixture, which remains in the distillation bottom, is fed to a second distillation stage, and the methanol distilled off is recycled to the production process.

The azeotropic mixture is subjected to an extractive distillation (extraction medium: wastewater from the methanol distillation). The still bottoms so obtained, a water – methanol mixture, are fed to the methanol distillation. Methyl acetate taken from the top of the column is fed to a reactor where it is hydrolyzed to methanol and acetic acid with the aid of an ion-exchange resin. In the subsequent distillation, acetic acid is removed, and the mixture of methanol and unreacted methyl acetate from the extractive distillation is recycled (Fig. 9). Wastewater from this recovery process is treated in a biological wastewater treatment plant.

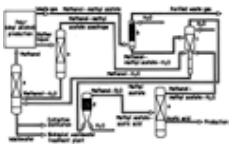

Figure 9. Recovery of methanol and acetic acid

a) Distillation; b) Scrubber; c) Extractive distillation; d) Hydrolysis reactor

2.2.1.3. *Acetylation without Contamination of Wastewater [56]*

The acetylation of 2-naphthylamine-8-sulfonic acid was formerly carried out in an aqueous medium with acetic anhydride. The product, 2-acetaminonaphthalene-8-sulfonic acid, was salted out from the aqueous medium by addition of ammonium sulfate and filtered off.

The mother liquor together with the washwater from this filtration contained 1200 kg of ammonium sulfate (COD 1000 kg) for each 1000 kg of product. This wastewater was treated in a biological wastewater treatment plant (see Figs. 10 and 11). The formation of this highly contaminated wastewater was the reason for replacement of the process by dry acetylation. For this, a flat-bottomed vessel with stirring equipment was installed (Fig. 12, a). Because of the dry process, no contaminated wastewater was formed; thus, treatment of ca. 2000 kg of sewage sludge in the wastewater treatment plant was avoided. In addition, 270 kg of acetic acid was recovered per 1000 kg of product.

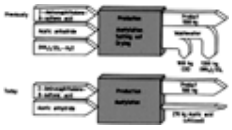

Figure 10. Comparison of previous and current production processes of 2-acetaminonaphthalene-8-sulfonic acid

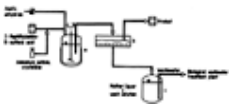

Figure 11. Production of 2-acetaminonaphthalene-8-sulfonic acid by wet acetylation

a) Acetylation and salting out; b) Filtration; c) Separation

Figure 12. Production of 2-acetaminonaphthalene-8-sulfonic acid by dry acetylation

a) Acetylation and drying; b) Filtration; c) Cooling trap; d) Condenser; e) Central waste-air purification plant; f) Separation

2.2.1.4. *Reutilization Plant for Organohalogen Compounds*

Introduction. Organic halogen compounds, especially those containing bromine and fluorine, are intermediates in the synthesis of pharmaceuticals, plant-protection agents, special high-performance plastics, and liquid crystals for optical displays. These intermediates must be produced in highly pure state. They are therefore separated from byproducts formed during the chemical synthesis in later purification stages.

Since the byproducts are often not directly reutilizable, they are disposed of in a waste incineration plant. Up to now only a recovery of the energy content is possible; a chemical reutilization of the byproducts is not. Purification of the flue gas does not lead to reutilizable sustances.

The objective was not only to utilize the calorific values of these byproducts or residues but also to reuse them chemically. This was achieved by thermal oxidation (incineration) with simultaneous neutralization to give potassium halides that are subsequently dissolved in water. The salt solution obtained is used for the recovery of potassium bromide, potassium fluoride, and potassium chloride as raw materials.

Method of Operation (Fig. 13). Liquid organic residues, which contain mainly bromine compounds with some chlorine and fluorine compounds, are fed through jets into a vertical cylindrical oxidation chamber (incinerator, a). The essential feature of the oxidation chamber is a special burner system that causes intensive turbulence of liquid and gaseous residues, fuel, and combustion air. An auxiliary flame produced by a special burner fired with natural gas is necessary to establish and ensure continuous operation of the oxidation process. This burner produces a very intensely radiating flame that reduces the mixing zone within the oxidation chamber. Intensive mixing of the gases ensures complete reaction, which results in comparatively low emissions of NO_x and CO. The high calorific value of the organohalogen residues ensures that the thermal energy necessary for the process is liberated.

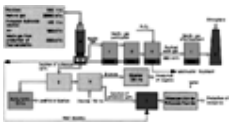

Figure 13. Recovery of bromine from residues containing brominated organics

a) Incineration and salt production (1200 °C); b) Quench cooling (85 °C); c) Scrubber (two stages); d) Electrostatic filter; e) Scrubber (one stage); f) Stack; g) Filtration; h) Bromine recovery; i) Evaporation

Waste air from the production of halogenated aromatics is used to form part of the combustion air, so that the residues present in the waste air (e.g., fluorine compounds) can also be treated in the recovery process.

The residues are converted to hydrogen halides in the oxidation chamber at 1200 °C and a residence time of at least 2 s. The hydrogen halides are then immediately neutralized by the simultaneous spraying of potassium hydroxide solution into the chamber, to form potassium halides. This prevents the formation of free hydrogen halides or of elementary halogens according to the Deacon equation. In the case of bromine, the equation

$$2\,Br_2 + 2\,H_2O \rightleftharpoons 4\,HBr + O_2$$

is shifted completely to the right by immediate neutralization of the hydrobromic acid. The equilibrium reaction is then no longer reversible, and complete conversion of the residues to hydrogen halides and, ultimately, potassium halides occurs.

This also avoids the secondary reaction

$$Br_2 + 2\,KOH \rightleftharpoons KBr + KBrO + H_2O$$

which occurs if the potassium hydroxide solution is added to the waste gas at a later stage, as in conventional exhaust-gas purification.

The potassium salts that are liquid at >800 °C flow partially down the walls to the bottom of the oxidation chamber and are dissolved in the water bath in the quenching vessel (b) directly below. The vapor-phase fraction of the salt forms aerosols

in the quenching vessel on cooling to ca. 85 °C. The concentrations of these aerosols are then reduced in the multistage waste-gas purification equipment (c – e) to levels below the limit values stipulated by Regulation 17 of the German Federal Antipollution Law (17. Verordnung Bundesimmissionsschutzgesetz 17. BImSchG). In this equipment, waste gas is passed successively through two jet scrubbers (c), a condensation electrostatic filter (d), and a packed column fed with H_2O_2 (e).

The liquid level in the quenching vessel is kept constant by adding liquor from the waste-gas scrubbers. Thereby some evaporation of the salt solution occurs.

Insoluble components of the salt liquor (ca. 1 % of the residues treated) are removed by filtration. These components consist mainly of insoluble inorganic salts and metal oxides; they can be dumped or, preferably, reutilized.

Elementary bromine is obtained from the filtered salt solution by reaction with chlorine in a further process step. Potassium chloride and potassium fluoride are recovered by evaporation of the remaining solution (with heat recovery from the quenching vessel) and are internally recycled.

The purified waste gas (see above) then undergoes a final purification by passing through a two-stage waste-gas scrubber in a production plant for inorganics (see Fig. 13).

Advantages of the Process. The formation of dioxins is prevented by the high oxidation temperature (1200 °C), the residence time of 2 s in the reactor, the mainly homogeneous temperature profile in the oxidation chamber, and the quench cooling of the reaction gases from 1200 to 85 °C. Measured concentrations are <0.1 ng/m^3 toxicity equivalents dioxins and furans.

Known alternative methods for treating liquid residues either have disadvantages or offer no advantages:

1. *Hydrogenation process*
   The high fluorine content of the residues leads to problems with construction materials. Elementary bromine is not recovered. Production of the hydrogen required for the process is energy intensive.
2. *Plasma process*
   Since the plasma burner can be used only as a pre-stage before the subsequent oxidation with air or pure oxygen, the consumption of electrical energy, and the cost of the plant are excessively high. Further recovery of the halogens will be similar to that used in the recommended process. The ecological balance of this process is therefore very unfavorable.
3. *Pressure gasification*
   A mixture of CO, CO_2, and hydrogen halides is formed at ca. 1600 °C. The hydrogen halides must be separated by waste-gas purification to recover the halogens. The formation of elementary bromine in the gasification process is problematic and would lead to considerable difficulties in waste-gas purification.
4. *Chemical reverser*
   In a converter tube clad with a special ceramic, a reductive atmosphere with a hydrogen excess at very high temperature is produced by combustion of propane under pressure. The fluid organohalogen compounds are decomposed in this temperature region.

For the substances mentioned above containing high levels of aromatics, the mixture would have to be led through a combustion chamber before entering the reactor to convert the molecules into suitable fragments that are able to react with hydrogen in the cracking process.

The later process steps, such as quenching and gas purification, would be very similar to those in the process initially described. It would be complicated to include the use of waste air as combustion air or the use of heat of reaction in the concept of the chemical reverser.

2.2.1.5. *Vacuum Technology for Closed Production Cycles*

In a chemical production plant, a large number of different machines, fittings, and safety equipment must work together reliably. With suitable engineering design, the amount of residues can be reduced or the residues can be recovered.

In the chemical industry, many processes are operated at reduced pressure (between 2000 Pa and atmospheric pressure). Liquid ring pumps, which are displacement pumps, are widely used for pressure reduction (⟶ Vacuum Technology – Displacement Pumps). The working liquid of a liquid ring pump must perform the functions of a piston, a seal, and a heat absorber.

This working liquid can be water or an organic liquid. Water ring pumps are used in recycling operations to reduce wastewater contamination. If a gas at higher temperature saturated with a solvent is pumped, the same solvent can be used as the working liquid of the pump. In a recycling operation, such a system can be designed for solvent recovery by connecting it to a rectification column. An example of this system is shown in Figure 14, in which isopropanol is the solvent.

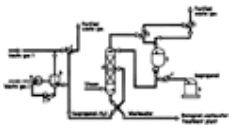

Figure 14. Vacuum plant with liquid ring pump

a) Liquid ring pump; b) Separator; c) Waste-gas condenser; d) Rectification; e) Cooler for working liquid (isopropanol); f) Condenser

Such a system cannot be used for pumping solvent mixtures or low-boiling solvents (e.g., acetone). Nor can water ring

pumps be used because these lead to contamination of the wastewater. Here, sliding vane rotary vacuum pumps (without oil lubrication) can be used (⟶ Vacuum Technology – Sliding Vane Rotary, Rotary Plunger, Rotary Piston Pumps, ⟶ Vacuum Technology – Rotary Lobe Vacuum Pumps). Such a vacuum pump has been developed for pumping solvent mixtures produced during synthetic resin production. It has been put into operation by Herberts in association with the company Provac (see Fig. 15). The following plant design was used: The vapors are drawn under reduced pressure through the precondenser (b) to the Roots vacuum pump (c). The gases that have been heated by compression in the Roots vacuum pump are then cooled in the intermediate condenser (d), and the condensate formed is automatically removed via a condensate trap into a collection vessel (g). The gases leaving the intermediate condenser are fed into two sliding vane rotary vacuum pumps (e) arranged in parallel. The gases from these pumps are fed into a condenser (f) in which more condensate is formed; this also drains into the collection vessel. The residual vapors are then fed to the waste-gas purification system. This method of operation reduces contamination of the wastewater, and a mixture of solvents is recovered. Furthermore, the inclusion of a Roots vacuum pump eliminates the use of a steam ejector pump, thereby avoiding the contamination of wastewater that such a pump would cause.

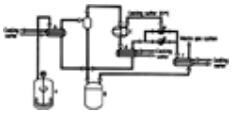

Figure 15. Vacuum plant with nonlubricated sliding vane vacuum pumps to reduce pollution of wastewater

a) Reactor; b) Precondenser; c) Roots vacuum pump; d) Intermediate condenser; e) Sliding vane rotary pump (two stage); f) Final condenser; g) Condensate collection vessel

2.2.1.6. *Utilization of Exhaust Gases and Liquid Residues of Chlorination Processes for Production of Clean Hydrochloric Acid*

When organic compounds are chlorinated on an industrial scale, large quantities of hydrogen chloride are formed. This is contaminated with chlorinated hydrocarbons formed as byproducts of the chlorination process.

A measure of the degree of contamination is given by the content of halogen compounds that can be adsorbed from the acid by activated carbon (AOX content). In nearly all chlorination reactions the chlorine is incompletely consumed, so that a fluctuating proportion of free chlorine usually occurs both in the hydrochloric acid recovered from the residual gases from the chlorination process and in the waste gases.

This free chlorine is as undesirable as other impurities in both the waste gas and the hydrochloric acid, and must be removed for environmental reasons and to improve acid quality.

One method of producing cleaner hydrochloric acid and a waste gas with only traces of organic impurities is to subject the chlorinated residues to thermal treatment before the absorption stage. In this treatment, the impure, hydrogen-chloride-containing gases are heated in suitable combustion chambers to ca. 600 – 1200 °C. The energy is usually provided by the combustion of natural gas or other fuels in an excess of oxygen, the temperature being adjusted to suit the impurities present. The flue gases obtained contain low levels of organic substances as well as some free chlorine. The latter occurs because in the presence of oxygen, free chlorine originally present in the chlorination residues does not react quantitatively to form HCl according to the equation

$$2\,Cl_2 + 2\,H_2O \rightarrow 4\,HCl + O_2$$

Instead, additional free chlorine can even be formed.

In subsequent absorption of the hydrogen chloride, free chlorine partly dissolves in the hydrochloric acid formed and partly remains in the waste gas. Removal of chlorine from both the hydrochloric acid and the waste gas is technically difficult. Air blowing or stripping the hydrochloric acid gives incomplete removal of the chlorine and causes a considerable decrease in the concentration of the acid.

Moreover, additional chlorine-containing and hydrogen-chloride-containing waste-gas streams are formed, and these too must be purified. Like the dechlorination of the original waste gases, this can in principle be performed only by alkaline scrubbing, which transforms the waste-gas problem into a wastewater problem.

A method for removing hydrogen chloride and chlorine from combustion gases with recovery of hydrochloric acid in a multistage gas scrubbing process has been described [57].

In another process for reducing the free chlorine content of flue gases from a combustion plant [58], hydrogen is fed into the quench zone of the combustion apparatus at a rate that depends on the air rate to the combustion chamber. By this means, the chlorine content of a combustion gas has been reduced from 600 to 48 mg/m^3. However, chlorine-free waste gases have not yet been produced from flue gases with such high chlorine contents.

Finally, a process has been developed for disintegrating chlorinated hydrocarbon residues with recovery of hydrochloric acid [59]. The chlorine content in the flue gas from the combustion process is kept as low as possible by minimizing the oxygen content and feeding steam or water into the combustion chamber. However, the hydrochloric acid produced is apparently not completely free of elementary chlorine, because chlorine is sometimes still found in the waste gas after HCl absorption.

Methods are also described for the combustion of liquid and solid residues with the production of hydrochloric acid [60], [61]. In these processes also, the starting material is burned in an excess of oxygen, as already described. Thus, these methods have the same disadvantages as those mentioned, because chlorine is present both in the hydrochloric acid produced and in the waste gas.

A method has been developed for producing hydrochloric acid with low levels of AOX and almost no elementary chlorine, together with a waste gas that contains only traces of these substances and can be released directly into the atmosphere. Here, the hydrogen chloride and chlorine-containing crude gases contaminated with organic compounds are thermally treated by using hydrogen or hydrogen-containing combustion gases with a deficiency of oxygen. The fuel and air rates are adjusted so that no carbon is formed in the combustion zone and an excess of hydrogen is present in the exhaust gas [62].

Although this process enables chlorination residues of the usual composition to be purified so that they can be used to produce pure hydrochloric acid, in extreme cases (e.g., if very high levels of impurities are present in the crude gases or if liquids are to be treated), problems with carbon formation in the combustion chamber can arise. This is undesirable, because both the plant and the products would be fouled.

Novel Process [63], [64] (Fig. 16). A thermal process was needed in which the reaction would proceed in a stable and controllable fashion. This process would also have to be suitable for purifying both crude gases with extremely high levels of impurities and chlorine-containing liquids. These could be converted to hydrogen chloride so that the latter could be directly absorbed without further purification processes, to give chlorine-free hydrochloric acid and a waste gas suitable for immediate discharge to the atmosphere.

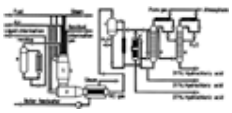

Figure 16. Production of pure hydrochloric acid from residual chlorination gases and liquid chlorination residues

a) Evaporator; b) Two-stage combustion; c) Quench; d) Isothermic absorber; e) Adiabatic absorber

The problem was solved by using thermal treatment of the impurities in a combustion operation with an excess of oxygen, and by following the first combustion stage by a second in which reaction gases from the first stage are subjected to further thermal treatment, in this case in a reducing atmosphere.

The crude gas mixture to be treated contains, in addition to hydrogen chloride and chlorine, considerable quantities of chlorinated hydrocarbons and other organic chlorine compounds. The mixture is subjected to a *two-stage thermal treatment*. It is fed into a *first combustion chamber* supplied with sufficient fuel and air. (Under some circumstances, replacing the combustion air either completely or partially with oxygen can be advantageous in this stage of the combustion process.)

Gaseous fuel (e.g., natural gas or hydrogen) is generally used in the first combustion stage, although liquefied petroleum gas or light or heavy fuel oil can also be used. In the *second combustion stage*, fuels with the lowest possible carbon content are preferred, because carbon formation is more likely here than in the first combustion stage since the process requires a deficit of oxygen.

If high-carbon fuels are used in the first combustion chamber, or if the raw gas contains large quantities of compounds with a high carbon content, the substoichiometric reaction in the second combustion chamber can lead to the formation of large amounts of carbon monoxide according to

$$CO_2 + H_2 \rightarrow CO + H_2O$$

After HCl absorption, the waste gas cannot be discharged directly because of legal restrictions. In this case, CO in the waste gas is converted to carbon dioxide, preferably in an afterburner under controlled conditions or sometimes catalytically.

The combustion chambers generally consist of steel vessels lined with a refractory ceramic material. Since this lining is not usually gastight, the highly corrosive raw gases, which contain hydrogen chloride and especially chlorine, can corrode the interior walls of the steel vessels at the prevailing surface temperatures, so that metallic ions, especially iron ions, appear in the hydrochloric acid formed. If high-purity hydrochloric acid is required (i.e., a product that not only is free of elementary chlorine and organic compounds but also contains only traces of inorganic compounds such as metallic salts, especially iron chlorides), the interior walls of the steel vessels and combustion chambers should be given a corrosion-resistant coating (e.g., enamel) before application of the porous ceramic material.

The crude gas to be treated is fed into the first combustion chamber. If necessary, liquids can first be passed through a vaporizer, and the vapors produced can then be mixed into the rest of the crude gases. If the liquids have high boiling points, feeding steam into the vaporizer can be useful. Combustion can be improved by adding combustion air, fuel, and additional steam or chlorine gas to increase the HCl content of the reaction gases.

The combustion temperature in the first chamber can be 800 – 1600 °C (preferably 1000 –1300 °C); this is maintained by an auxiliary burner. The fuel – air mixture is adjusted so that an oxygen content of 0.1 – 11 vol % (preferably 0.5 – 5 vol %) is measured in the reaction gases at the exit of the first combustion chamber.

The reaction gases pass from the first to the second combustion chamber. There the temperature is maintained at 800 – 1600 °C (preferably 1000 – 1300 °C) by a burner supplied with a fuel –air mixture whose ratio is adjusted so that the reaction gases at the exit of the combustion chamber have a hydrogen content of 1 – 15 vol % (preferably 2 – 5 vol %).

Sometimes, the temperature of the second combustion chamber may have to be decreased by feeding steam into this chamber.

The reaction gases are passed through one or more absorption columns after cooling.

Preferably the pressure in the first combustion chamber should be kept constant automatically by a blower installed downstream.

The hydrochloric acid produced is practically free of elementary chlorine and organic chlorine compounds, and requires no further treatment. Also, the purified gas can be discharged directly into the atmosphere.

2.2.1.7. *Production of Neopentyl Glycol: Higher Yield by Internal Recycling [38], [65], [66]*

In the process for producing neopentyl glycol (2,2-dimethyl-1,3-propanediol) a coupled product is formed and some starting materials remain unreacted (⟶ Alcohols, Polyhydric – 2,2-Dimethyl-1,3-propanediol (Neopentyl Glycol)). The latter are internally recycled to the process with the aid of the coupled products.

The first reaction, in reactor 1, is an aldol condensation in which formaldehyde is reacted with isobutyraldehyde (2-methylpropanal) in an alkaline medium with tri-*n*-propylamine as the catalyst to form hydroxypivalaldehyde (3-hydroxy-2,2-dimethylpropanal).

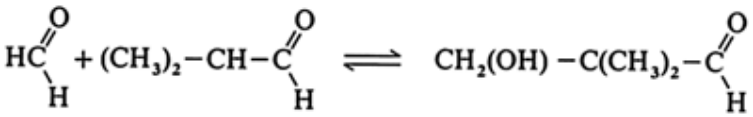

The reaction mixture is transferred to a separating vessel where two layers are formed. The organic phase contains hydroxypivalaldehyde and an equilibrium amount of isobutyraldehyde. The heavier aqueous phase contains hydroxypivalaldehyde together with small amounts of unreacted starting materials (formaldehyde and isobutyraldehyde). The organic phase is hydro-genated in reactor 2 using a nickel contact catalyst. Here, the target product neopentyl glycol is formed from the hydroxypivalaldehyde and, isobutanol is formed from the isobutyraldehyde as the coupled product.

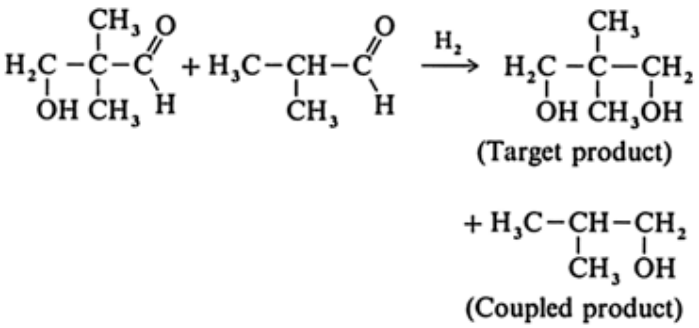

Neopentyl glycol and isobutanol are separated by distillation. Some of the recovered isobutanol is used to extract the aqueous phase formed in the aldol condensation, and the remainder is used to produce synthesis gas. The isobutanol extract contains hydroxypivalaldehyde and starting materials. It is therefore recycled to the aldol condensation reactor. The isobutanol transferred into reactor 1 then enters reactor 2 via the organic phase. This leads to an increase in the amount of isobutanol so that, after separation by distillation, some of the isobutanol can be used to produce synthesis gas. Wastewater from the bottom of the extraction column is fractionated. The low-boiling compounds formed as overhead products are also used to produce synthesis gas (Fig. 17). Wastewater from the fractionating column is treated in the biological wastewater treatment plant.

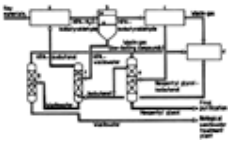

Figure 17. Production of neopentyl glycol — recovery of hydroxypivalaldehyde from wastewater by extraction with isobutanol

a) Reactor 1 (aldol condensation); b) Separating vessel; c) Reactor 2 (hydrogenation); d) Synthesis gas production; e) Distillation; f) Extraction; g) Rectification HPA = Hydroxypivalaldehyde

2.2.1.8. *Optimization of Ester Waxoil Production and Recovery of Auxiliary Products [16], [67]*

Ester waxoils are produced from crude montan wax by oxidation with chromic acid to acid waxes, followed by esterification with aliphatic polyalcohols. The products are used in polishes for automobiles and furniture.

The crude montan wax contains resinous components that are removed by methylene chloride treatment. The methylene chloride is then recovered by distillation of the solution obtained (Fig. 18). In detail, the process consists of mixing the crude wax with methylene chloride (a) and filtering the mixture through a band filter (b). The CH_2Cl_2 is recovered from the filtrate (resin dissolved in CH_2Cl_2) with the aid of a recirculating evaporator (d), separating vessel (i), and condenser (e) and the residual resin can be incinerated (f). The deresinified crude wax is melted (c) and is stripped with steam to remove any remaining CH_2Cl_2 (h).

Figure 18. Methylene chloride recovery in wax production

A) a) Mixing vessel; b) Filter unit; c) Melting vessel; d) Evaporator; e) Condenser; f) Separating vessel; g) Activated carbon filter; h) Stripping column; i) Separator; j) Resin incineration; k) Receiver; l) Stripping column

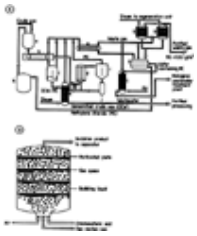

B) Mixing vessel (a) in detail: Continuous wax oxidation

The subsequent oxidation process with chromic acid is optimized. The earlier batchwise process in lead-lined stirred vessels resulted in contamination of water, and the consumption of raw materials and energy was high. This has now been replaced by a continuous process in which montan wax and a chromic acid – sulfuric acid mixture are passed through a sieve tray column. The two components are intensively mixed by compressed air as they enter the column (Fig. 19, a), giving a high conversion of chromic acid. The exothermic reaction heats the mixture to the temperature required, so that external heating of the column is unnecessary. This procedure gives an oxidation product of uniform quality, leads to savings in energy and raw materials, and enables the purification stage to be performed continuously and effectively. The steam saving is 1 t per tonne of wax acid compared with the previous process. The consumption of chromic acid is reduced from 170 % of the crude montan wax (former process) to 115 %. The consumption of electricity for regeneration of the exhausted chromic acid, [i.e., for electrolytic oxidation (i) of chromium(III) to chromium(VI)] is reduced by 2000 kW · h per tonne of acid wax.

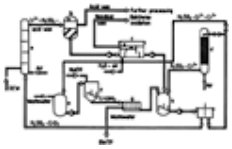

Figure 19. Chromium circulation at the oxidation stage of wax production

a) Continuous oxidation column; b) Separating vessel; c) Flotation; d) Oxidation tower; e, f) Mixing tanks; g) Filter press; h) Mixing tank; i) Electrolysis

DCW = Deresinified crude wax; BWTP = Biological wastewater treatment plant

The exhausted chromic acid [chromium(III) salt solution] produced in the oxidation stage is separated from the acid wax by phase separation (b) and flotation processes (c) and is electrolytically treated (i). The anolyte [chromium(VI) salt] is fed into the oxidation column (a). The catholyte [chromium(III) – chromium(II) salt mixture] is partly fed into an oxidation tower (d) and partly mixed with wastewater (spray and wash water), which contains chromium(VI). This causes the chromium(II) present in the catholyte to be oxidized and the chromium(VI) in the wastewater to be reduced. In both cases, a solution of chromium(III) is formed. The small amount of chromium in the wastewater is precipitated as chromium(III) hydroxide by addition of sodium hydroxide solution (f), and then filtered off in a filter press (g) and brought into solution in the acid conditions in the mixing tank (h). This is followed by the electrolytic oxidation stage to form chromic acid (i). Residual wax from the flotation stage is oxidized later in a separate operation.

Separation of chromic acid residues from the oxidation product is carried out continuously today by a disk separator. Zirconium is the only construction material that can be considered in view of the highly corrosive properties of the medium. A disk separator of zirconium has been developed specifically for this purpose in cooperation with an engineering company. This gives both an improvement in product quality by reducing the chromium content of the oxidation product and a saving of sulfuric acid for the washing processes. The sodium sulfate content of the wastewater is thus reduced by 13 kg per tonne of acid wax.

2.2.1.9. *Biochemical Production of 7-Aminocephalosporanic Acid [68-70]*

Introduction. 7-Aminocephalosporanic acid (7-ACS) is a pharmaceutical chemical obtained by elimination of α-aminoadipic acid from the fermentation product cephalosporin C. 7-Aminocephalosporanic acid is the key product for most semisynthetic cephalosporin antibiotics. It is used by Hoechst to produce Claforan (cefotaxime sodium), cefodizime, Cefquinom, and Cefpirom. Until now, 7-ACS has been produced by a chemical process. A process based on biochemical catalysts has now been developed and is considerably more environmentally friendly.

Chemical Process. The chemical process used hitherto consists of the following stages:

1. Production of the zinc salt of cephalosporin C
2. Reaction of the cephalosporin C zinc salt with trimethylchlorosilane to protect the functional groups (NH_2 and COOH)
3. Productions of the imide chloride from the silylated product by reaction with phosphorus pentachloride at 0 °C
4. Hydrolysis of the imide chloride (see Fig. 20) to yield 7-ACS.

Figure 20. Production of 7-aminocephalosporanic acid

Enzymatic Process. The enzymatic synthesis is carried out as a two-stage process:

1. Cephalosporin C is supplied directly to the enzymatic stage. The side chain of the molecule is oxidatively deaminated by a D-amino acid oxidase (obtained from yeast). Oxygen is consumed, and ketoadipinyl-7-ACS, ammonia, and hydrogen peroxide are formed. Under these reaction conditions, oxidative decarboxylation to form glutaryl-7-ACS occurs spontaneously.
2. In the second stage of the process, the glutaric acid is split off by the enzyme glutaryl amidase, and 7-ACS is liberated (see Fig. 20).

The reactions are carried out at room temperature in aqueous solution. The enzymes are immobilized on a spherical carrier and can be reused many times. The enzyme D-amino acid oxidase is obtained in good yield from the natural host bacterial strain.

In the production of the enzyme glutaryl amidase, only a low concentration was produced by the wild strain *Pseudomonas.* An economically interesting concentration (greater than fiftyfold increase per volume unit) can be produced only by genetic engineering (i.e., cloning the enzyme glutaryl amidase in *Escherichia coli*).

Comparison of the two Processes. The chemical process is inferior to the enzymatic process with respect to the reactants and auxiliaries employed:

1. Use of chlorinated hydrocarbons and toxic auxiliaries
2. Use of phosphorus pentachloride, which is subject to the German decree about emergencies (Störfallverordnung), and of trimethylchlorosilane, which can be handled only in the absence of moisture and is highly flammable
3. Use of heavy-metal salts (i.e., zinc salts)

Use of the enzymatic process leads to the following changes in the amounts of residues produced (per tonne of 7-ACS) compared with the chemical process:

1. Mother liquors requiring incineration: reduction from 29 to 0.3 t
2. Waste-gas emissions: reduction from 7.5 to 1.0 kg
3. Wastewater contamination (COD): increase from 0.1 to 1.7 t
4. Residual zinc by recovery as $Zn(NH_4)PO_4$ and reutilization: reduction from 1.8 to 0 t
5. Distillation residues (from the distillative treatment of the chlorinated hydrocarbons) for incineration as waste: reduction from 2 to 0 t

Furthermore, the chemical process is energy intensive, since it can be carried out only at low temperature.

The fraction of the process costs used for environmental protection is reduced from 21 to 1 %. This fraction includes the cost of the additive environmental protection (waste incineration and purification of wastewater and waste gas). Thus, the absolute environmental protection costs are reduced by 90 % per tonne of 7-ACS.

2.2.2. *Examples from Bayer*

2.2.2.1. *Avoidance of Wastewater and Residues in the Production of H Acid (1-Amino-8-hydroxynaphthalene-3,6-disulfonic acid)*

Introduction. Development of the so-called coal tar dye industry has resulted in extensive knowledge of naphthalene chemistry. The “letter acids” based on naphthalene derivatives are very important, especially for azo dyes. The color properties of the chromophore present depend highly on the substitution pattern of the naphthalene nucleus. The introduction of sulfonate, nitro, amino, and hydroxyl groups in various positions is therefore an important characteristic feature of the production of these intermediates ($\longrightarrow$).

A dye component that is used frequently is H acid (1-amino-8-hydroxynaphthalene-3,6-disulfonic acid). Production of this comparatively complex substitution pattern requires a series of different chemical reactions. The known side reactions and subsequent reactions on sulfonation and nitration of the naphthalene system and high-pressure hydrolysis of its substituents reduce the yield of the target product. If, as a consequence, intermediate isolations are unavoidable during the course of the process to prevent a gradual increase in the amount of byproducts, additional product losses in mother and wash liquors will occur. Management of the often considerable quantities of solid waste and wastewater is therefore an important consideration in the production of H acid.

Conventional Process. The production process for H acid is divided into four basic operations each of which includes several steps:

1. Sulfonation of naphthalene to give the trisubstituted product
2. Nitration of the trisulfonic acid
3. Reduction of the nitro group to T acid
4. High-pressure hydrolysis of the sulfonic acid group in the α-position to give H acid

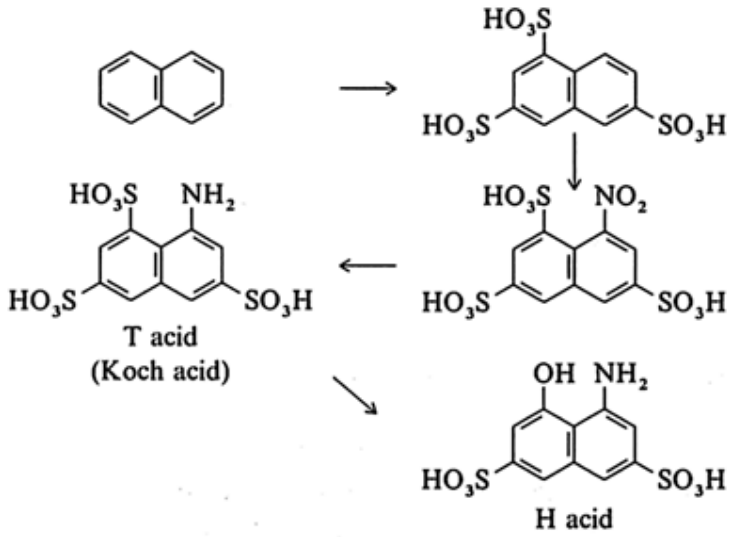

This long reaction chain was distributed over various buildings of the Leverkusen works.

The rule of ARMSTRONG and WYNNE applies to the multiple sulfonation of naphthalene; i.e., the two sulfonic acid groups that enter the ring are in neither the ortho, the para, nor the peri position to each other [71]. Therefore, no further reaction to form naphthalenetetrasulfonic acid occurs after the desired 1,3,6-substitution. However, this pattern can be obtained only via the 1,6- or 2,7-disulfonic acid as intermediates. The desired yield is therefore influenced considerably at the very beginning of the reaction by the selectivity and completeness of the reaction forming the disulfonic acid intermediate stage. A considerable proportion of the naphthalene used is lost at the sulfonation stage according to [72].

The sulfonation product is then nitrated. Reaction conditions can be chosen such that the α-position is preferred [73]. The directing effect of the sulfonic acid groups promotes substitution at the desired position. However, nitration of the naphthalene nucleus has a much greater tendency to lead to side reactions (e.g., oxidations) than is the case with the benzene series, so that when the nitrated mixture is diluted with water, nitrous gases are evolved.

The acid solution is neutralized by adding calcium carbonate, and the precipitated gypsum produced from the excess sulfuric acid is filtered off. The nitrated product remains in the filtrate and is reduced by iron chippings to T acid (1-aminonaphthalene-3,6,8-trisulfonic acid). The remaining calcium ions are removed as calcium carbonate together with the iron oxide by adding sodium carbonate solution, (⟶) and the T acid is isolated as an intermediate by acidification of the filtrate, so that the various byproducts can be removed with the mother liquor.

Naphthalenesulfonic acids are hydrolyzed to hydroxy derivatives under drastic alkaline conditions, and sulfonic acid groups in the α-position react preferentially [74]. Thus, the T acid is converted to H acid by treatment with sodium hydroxide solution under pressure [75]. The H acid is then salted out with sodium chloride. Because of the unfavorable solubility properties, large quantities of sodium chloride must be employed which remain in the mother liquor with the organic byproducts.

In view of the yields of individual stages quoted in the literature, even after many years of production experience, this process gives an overall yield of only 40 % based on naphthalene.

Revised H Acid Process. The poor infrastructure (various stages of the process carried out at different locations), age of the plants, and intensity of emissions, among other factors, have provided the motivation to attempt a systematic revision of the process. This has led to complete redesign of the operation in the factory at Brunsbüttel [76], [77]. Operation in Brunsbüttel began in 1981, and the old plants have been completely replaced.

The basic process pattern is unchanged. The important changes to the process are replacement of several intermediate steps by a continuous system, and replacement of iron reduction by hydrogenation with hydrogen using a nickel catalyst [78]. A modern process control system is used that considerably reduces variation of the process control parameters compared with the earlier in situ manually controlled system. Salting out and isolation of intermediates (e.g., T acid) are omitted. There is now only the isolation of the target product. Waste-air and wastewater emissions that still occur are purified in new plants. Those waste-gas streams from the plant that contain nitrogen oxides, organic compounds, and hydrogen are treated in a combustion plant. The remaining wastewater stream containing organic byproducts from the synthesis that are difficult to degrade biologically is reacted in a wet oxidation plant with oxygen at high pressure. This “wet incineration” proceeds almost to completion, so that subsequent biological wastewater purification is unnecessary.

Ecological Aspects of the Process. In comparison to the earlier process, the modern H acid production process shows considerable improvements in emission levels for equal quantities of the target product.

Results of the revision of the process

1) Starting materials

Naphthalene	– 20 %
Sulfuric acid	– 20 %

Calcium carbonate	– 20 %
Nitric acid	– 20 %
2) Gypsum	– 38 %
3) Iron oxide sludge	–100 %
4) Volume of wastewater	– 70 %
5) Wastewater load	
Sodium chloride	–100 %
COD	– 96 %
Sodium sulfate	+ 12 %

Efficiency. Consumption of the raw materials naphthalene, sulfuric acid in various concentrations, nitric acid, and calcium carbonate is now 20 % lower.

Gypsum and Calcium Carbonate Sludge. Because of their improved quality, gypsum and calcium carbonate sludge are no longer dumped, but are supplied to the cement industry for use as raw materials. Also, the quantity of gypsum produced has been reduced by 38 %.

Iron Oxide Sludge. Attempts were made to utilize the iron oxide sludge produced in the reduction by iron chippings of the nitrated compound to T acid. Use of this sludge as a raw material for the production of iron oxide pigment or in iron smelting was not very successful, so it was disposed of mainly by dumping. The present hydrogenation process produces only water as the byproduct. Catalyst recovered from the process is reutilized.

Volume of Wastewater. The omission of intermediate isolation stages, the change to continuous operation of various intermediate reaction stages, and the introduction of an intermediate evaporation stage led to a reduction by 70 % in the amount of process wastewater that must be treated because of its organic content.

Organic Load of Wastewater (*COD*). The aqueous processing steps, which are the rule in letter acid production, lead to the presence of side products formed from naphthalene in the process wastewater. The biological wastewater treatment is regularly unable to eliminate these materials (elimination rate sometimes <20 % within normal holding periods) satisfactorily.

The efficiency of conversion of raw materials is now considerably improved, and amounts of unavoidable byproducts have been reduced by high-pressure wet oxidation, so that the COD is now reduced by ≥96 %. Wastewater can therefore now be discharged directly into the Elbe.

Salt Content of Wastewater. Because of the considerably smaller amounts of water used in the process, the former technique of isolating H acid by salting out could be omitted. The process no longer requires the addition of sodium chloride, so 4.7 t of sodium chloride per tonne of H acid is no longer used. However, the increase in sodium sulfate content of 12 % (i.e., a total of 3.8 t per tonne of H acid) has to be accepted, but nonetheless the salt load of the wastewater is considerably reduced.

Economic Constraints. The example described is an illustration of the opportunities afforded by modern process technology to improve a production process that has remained the same in principle for 100 years. However, notwithstanding the process improvements achieved, the cost of the new plant and installations presents an increasing problem when viewed against the background of prices in international competition. Especially in Asia, obsolete technology and the associated increase in environmental pollution are tolerated. This results in cost structures that have consequences for the world market when they are widely used to confer market advantages.

The example therefore raises the question to what extent production-integrated environmental protection is hindered by lack of international harmonization. Environmental protection measures require a long-term positive outlook with respect to both market forecasts and official regulations, because the time scales for systematic process modernization, licensing, new plant construction, and an appropriate initial operating phase for refinancing must be taken into account.

2.2.2.2. *High-Yield Production of Alkanesulfonates by Means of Membrane Technology [79], [80]*

Introduction. Alkanesulfonates are anionic surfactants (⟶ Surfactants – Alkanesulfonates). They are produced by Bayer through the sulfochlorination process. Optimum properties are obtained when a mixture of straight-chain alkanes with 14 – 17 carbon atoms is used.

The alkanesulfonyl chloride (Mersol) produced by the sulfochlorination reaction is hydrolyzed with sodium hydroxide solution to give the sodium sulfonate. The hydrolyzed mixture separates on cooling into a product-rich upper phase (“glue”) and a salt-containing lower phase that still contains ca. 2 – 3 % sulfonate and must be disposed of.

Conventional Method of Recovering the Alkanesulfonates. The reaction product of sulfochlorination (obtained in ca. 30 – 33 % yield) is hydrolyzed at ca. 100 °C with aqueous sodium hydroxide, and the mixture obtained is cooled in several stages. Most of the unreacted alkanes separate in the hot phase and are recycled to the production process. The aqueous hydrolyzed solution contains residual alkanes, alkanesulfonates, and sodium chloride produced from reaction of the sulfonylchloride with sodium hydroxide solution.

On cooling to ca. 6 °C, an aqueous phase separates, which contains ca. 2 – 3 % of alkanesulfonates, 6 – 7 % sodium chloride, and traces of alkanes. Further cooling of the lower phase to –6 °C decreases the sulfonate content to ca. 1.5 %.

This lower phase was formerly treated in the central wastewater treatment plant. Apart from salt, it also contains organic substances that give a COD of ca. 20 g/L. Approximately 2.8 m^3 per tonne of sulfonate of the lower phase is produced resulting in a considerable COD in the wastewater.

Membrane Process for Recovery of Alkanesulfonates. Many attempts have been made over the years to reduce the wastewater load — which represents a loss of product — by a number of different methods. These include evaporation, extraction, reverse osmosis, and ultrafiltration. All of these processes have the disadvantage of high equipment cost and high energy requirements, and the space – time yield is low. The first breakthrough came with the development of new types of membrane with a definite separating efficiency and a large surface area, so-called spiral-wound modules.

Nanofiltration is best suited for successful recovery of alkanesulfonates from the lower phase. This process operates with standard spiral-wound modules at moderate pressure and gives recovery efficiencies of ≥90 %.

Suitable nanofiltration membranes based on the spiral-wound module technique are available from various suppliers; they usually consist of a carrier film, a spacer, and the filtration membrane (⟶ Membranes and Membrane Separation Processes – Spiral-Wound Module). After separation, the lower phase is heated preferably to ca. 25 °C, adjusted to a predetermined pH by addition of acid, and pumped at ca. 30 bar through a battery of pressure tubes suitably arranged in parallel and fitted with spiral-wound modules.

The concentrate produced is usually recirculated until the desired concentration is reached and then returned to the production process (e.g., the hydrolysis stage).

The membrane permeate contains <10 % of the initial sulfonate content and is treated in the central biological wastewater treatment plant. The entire process is operated continuously.

Surprisingly, not only do the modules have good retention properties for alkanesulfonates, but the concentrate also contains significantly less salt. The flow sheets for both processes are shown in Figure 21.

Figure 21. Comparison of processes for the production of alkanesulfonates with and without ultrafiltration

Ecological Aspects of the Modified Process. Conventional treatment of the hydrolysis mixture leads to a wastewater stream whose organic load, consisting of lost product and emulsified starting material, is ca. 20 g/L COD. The organic components also contain a certain amount of combined chlorine formed as a result of side reactions with the alkane chains during sulfochlorination.

The introduction of nanofiltration under pressure has led to a reduction in the organic content of the wastewater of up to 95 %, so that the filtrate has a COD of only 0.4 g/L. This decreases the load on the factory's biological sewage treatment plant, especially that due to organically bonded chlorine.

Concluding Comments. Improvement in the efficiency of alkanesulfonate recovery was first made possible by the development of efficient multilayer membranes with long service lives and high retention properties.

A flexible system of switching between modules enables the sulfonate recovery stage to operate for long periods without cleaning or maintenance operations.

The plant was installed in 1992 in the Uerdingen factory and has fulfilled the promise of the early pilot investigations.

2.2.2.3. *Selective Chlorination of Toluene in the para-Position*

Introduction. Monochlorination of toluene — yielding usually a statistical mixture of *ortho*- and *para*-chlorotoluenes — is an important step in the introduction of functional groups into this aromatic hydrocarbon. Monochlorination represents the first stage in a large number of reaction chains leading, for example, to chlorobenzaldehydes or cresols [81]. Applications of such products include plant-protection agents, pharmaceuticals, dyes, antioxidants, and preservatives [82].

Since various product lines follow their own market laws, the demand for *ortho*- and *para*-chlorotoluene fluctuates, *ortho*-chlorotoluene often being the less desired isomer. This situation is a typical problem in coupled production. If it cannot be solved by a pricing policy in accordance with market economics, the only possible solution in the most unfavorable case is to dispose of excess quantities of the less desired isomer (e.g., by incineration).

Conventional Chlorination of Toluene. In the batch chlorination of toluene (e.g., at 50 °C in the presence of $FeCl_3$ catalyst), a crude mixture is formed in which the monochlorotoluene fraction contains small amounts of *meta*-chlorotoluene, with the two isomers *ortho*- and *para*-chlorotoluene in a practically statistical distribution in the ratio *ortho*- : *para*- = 1.92 [83].

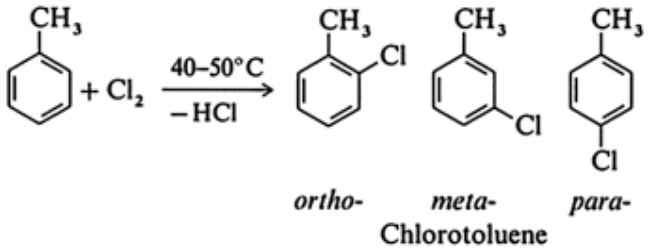

Since the economic importance of *para*-chlorotoluene is considerably greater than that of *ortho*-chlorotoluene, catalytic systems have been developed to increase the *para*-selectivity of the chlorination. Since the early 1930s addition of sulfur or disulfur dichloride [84] to $FeCl_3$ as a cocatalyst has been known to shift the isomer ratio in the monochlorotoluene fraction in favor of *para*-chlorotoluene (i.e., to an *o*- : *p*- ratio of 1.11).

Growth of product lines based on *para*-chorotoluene at the end of the 1970s and early 1980s has led to further development of the cocatalyst (⟶ Chlorinated Hydrocarbons). The use of thianthrenes [85-90], phenoxazines [91-93], or phenothiazines [94-97] instead of sulfur is described in patent applications.

However, these methods usually have significant disadvantages; e.g., the quantities of cocatalysts required are too large (0.1 wt % or more), the *o*- : *p*- ratio can be kept only if antimony compounds are used as the main catalyst, or the reaction temperature is too low (20 °C or less) so that the process is uneconomical.

Exceptions to this include the use of a mixture of $FeCl_3$ and polychlorothianthrene [90] or *N*-substituted phenothiazines (e.g., $FeCl_3$ – *N*-chlorocarbonylphenothiazine) [94]. However, these more economical variations achieve at best an *o*- : *p*- ratio of 0.84.

The market tendencies in recent years (e.g., due to new developments and other changes in the plant-protection field) are increasingly incompatible with this isomer distribution. The demand for *para*-chlorotoluene is increasing while at the same time the demand for *ortho*-chlorotoluene remains static or tends to decrease.

New Cocatalysts in Toluene Chlorination. Intensive research into the development of a catalyst system suitable for industrial use has therefore been started. These investigations have resulted in the discovery of the benzothiazepine [98-101] cocatalysts, whose properties are greatly superior to those of former state-of-the-art products. Their *para*-selectivity in comparison with $FeCl_3$ – sulfur (*o*- : *p*- = 1.11) is much increased (up to *o*- : *p*- = 0.55) by using minimum quantities of cocatalyst (e.g., 30 – 50 ppm) in the presence of $FeCl_3$ as the main catalyst and at industrially favorable temperatures of 40 – 50 °C.

After a relatively short optimization phase to establish the exact composition of the catalyst system, the reaction temperature, and the exact flow rates, etc., the new system could be used in an existing continuous full-scale plant.

Hence, in principle, since the early 1990s, the isomeric ratio in toluene chlorination has been adjustable within the range *o*- : *p*- = 1.11 to *o*- : *p*- = 0.55 according to market conditions; i.e., for each tonne of *para*-chlorotoluene produced, the amount of *ortho*-chlorotoluene can be fixed within the range 1.11 to 0.55 t. A comparison of old and new processes is given in Figure 22.

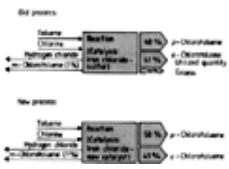

Figure 22. Selective chlorination of toluene in the *para*-position — comparison of old and new process

Ecological Aspects of the Catalyst Change. The decreasing demand for *ortho*-chlorotoluene has limited the possibilities for utilizing the excess material. The effect on cost that would result from incineration of *ortho*-chlorotoluene had to be assessed. With this background, the development of new benzothiazepine cocatalysts relieved the isomer problem at just the right time. The isomer ratio has now been moved in the desired direction and adapted to the demand pattern. No incineration of an undesirable product is necessary.

Concluding Comments. Extensive fundamental research into further development of catalysts has been supported by the use of theoretical chemistry and modern mathematical methods based on random techniques. The plants in which these results have been realized since 1989 operate in the Leverkusen works.

2.2.2.4. *Production of Naphthalenedisulfonic Acid with Closed Recycling of Auxiliaries*

General. Sulfonation of the naphthalene ring system has a long tradition, particularly as an important process step in the synthesis of the numerous “letter acids,” which are important intermediates in dye production. The process parameters (e.g., exclusion of water and temperature control) determine the substitution pattern (positions 1 or 2 in the naphthalene ring system). With an appropriate excess of sulfonating agent, both rings of naphthalene can be substituted, or more highly sulfonated derivatives can be obtained (⟶ Naphthalene Derivatives – Introduction, ⟶ Naphthalene Derivatives – Naphthalenesulfonic Acids). Since undesired isomers and byproducts are formed, sometimes in considerable proportions,

the process steps for workup and product isolation are frequently complex. Product yields relative to the naphthalene used, therefore, turn out to be relatively low.

Conventional Process for Production of Naphthalenedisulfonic Acid. The standard process for sulfonating aromatic hydrocarbons uses sulfuric acid and oleum as reactants. Under these conditions, the reaction vessels can be made out of gray cast iron, and steel can be used for piping and fittings.

$$C_{10}H_8 + 2SO_3 \xrightarrow{H_2SO_4} C_{10}H_8O_6S_2$$

Naphthalene is introduced into sulfuric acid and reacted with added oleum [102]. A temperature of 50 – 60 °C is required for satisfactory space – time yields. The reaction mixture is then added to water in vessels with acid-resistant brick lining. As a result of adding sodium hydroxide solution and sodium sulfate, naphthalenedisulfonic acid precipitates as its disodium salt and can be filtered off. Aside from residual amounts of the target product, considerable proportions of undesired isomeric di- and trisulfonic acid derivatives remain in the filtrate as byproducts. Because of the high saline contamination, a technically satisfactory solution to the problem of obtaining fresh sulfuric acid from these wastewater streams containing sulfuric acid has not yet been found. Likewise, current processes for biological wastewater treatment show only very mediocre elimination performances for this filtrate, so that the effluent of the wastewater treatment plant has high residual organic contamination (see Figs. 23 and 24).

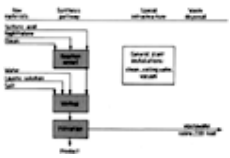

Figure 23. Conventional process for production of naphthalenedisulfonic acid

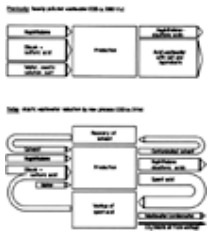

Figure 24. Mass balance of naphthalenedisulfonic acid production

New Process for Naphthalene Sulfonation. Modification of the process aims simultaneously at two goals: increasing the yield with more selective reaction and recycling sulfuric acid from the wastewater. Accordingly, not only must the parameters of the reaction step be changed so that fewer byproducts are formed, but the product must be isolated by a salt-free alternative process to enable sulfuric acid to be recovered.

The desired higher selectivity of the sulfonation reaction cannot be achieved by using sulfuric acid as a reaction medium but rather by dissolving the reactants, naphthalene and sulfur trioxide, in an organic solvent (chlorinated hydrocarbon, CHC) and reacting them at temperatures lower than those of the previous process [103].

Nevertheless, because of the dehydrating action of sulfur trioxide, higher-molecular structures are formed via anhydride bridges (Scheme 25). The anhydride bridges must, therefore, be hydrolyzed in the next step. For this purpose, the reaction mixture is added to water. An aqueous sulfonic acid solution is formed and the solvent is subsequently separated.

Subsequently, instead of the sodium salt, the free sulfonic acid [104] is precipitated by addition of sulfuric acid or recycled mother liquor, followed by cooling. The mother liquor, which remains after filtering off the product, can be supplied directly to existing plants for regeneration of spent sulfuric acid.

Ecological Aspects of the Process Alteration. The previous process is essentially characterized

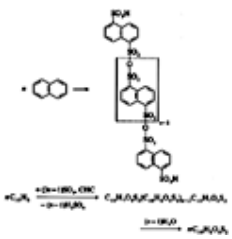

Figure 25. Naphthalene sulfonation

by a linear material flow to obtain the target product coupled with a flow of highly contaminated wastewater that contains the high naphthalene losses due to the formation of byproducts, excess sulfuric acid, and inorganic salt. Two recycled streams are now superimposed on this process [105] (see Fig. 26).

Figure 26. New process for the production of naphthalenedisulfonic acid

The organic solvent is recovered for reuse in the production plant. Excess sulfuric acid mother liquor is conveyed via the internal piping system for spent acid to the existing sulfuric acid plant, where concentration, decomposition to sulfur dioxide, and oxidation to sulfur trioxide occur. Sulfur trioxide is then withdrawn and fed to the sulfonation reaction through piping.

In the new process, process wastewater is, therefore, collected as the distillate from spent acid concentration. Since the volatilized components from evaporation are negligible, the previous wastewater contamination of more than 3000 t/a COD, which exhibits rather unsatisfactory biological degradation, is reduced to a few tonnes per year (e.g., from shower, rinsing, or cleaning water). Thus, for this process, a COD reduction of nearly 99.9 % is achieved (see Fig. 24). Simultaneous alteration of the reaction process markedly improves the yield, so that the proportion of naphthalene that does not react to the target product is now reduced by about a half.

Selection of the solvent however, leads, to considerable additional measures in process design and construction engineering. With regard to solvent power, enthalpy of evaporation, phase separation behavior, etc., low-boiling chlorinated hydrocarbons are best. This, for instance, involves special measures for sealing the tank farm on the water side, as well as monitoring the platform and surface drainage with detectors, so that if leaks occur, an emergency tank can be switched in.

The product is crystallized with vacuum cooling. The associated stripping effect allows maximum recycling of the solvent. Because the solvent must be anhydrous for reuse, the collected solvent streams are dried with sulfuric acid in a pulsed column. Waste air collected from the entire apparatus is supplied to a waste-air incineration plant with downstream washers.

Supplementary Engineering Equipment for Process Rearrangement. The conventional process can, in practice, be run manually with on-site operation. In comparison, the additional infrastructure of the new process with its multiple interconnections, as well as accurate maintenance of reaction parameters, requires considerable supplementary installations. Operation of the process is supported by a new control system developed specifically for this purpose. A high degree of standardization and uniform quality is achieved [106]. The new plant has been in operation in Leverkusen since 1989.

2.2.2.5. *Avoiding Residues in Dye Production by Using Membrane Processes [107-112]*

Introduction. In the production of water-soluble dyes, recovery of the product usually occurs by salting out, filtering on filter presses, redissolving or resuspending and refiltering, standardization, and drying. In this process, large quantities of wastewater and salt are produced. With the introduction of the new technique of pressure permeation, the amounts of salt and organic materials produced are reduced considerably, and a higher yield is obtained.

Conventional Process. The end product of many dye syntheses is an aqueous salt-containing dye solution. Separation of the product dye from byproducts such as salts and residual starting materials is performed conventionally by addition of further large quantities of salt to the dye solution to salt out the product dye, which is then filtered off from the salt solution by means of filter presses. Liquor from the presses contains mainly the salts used in the process together with up to 5 % of the product (Fig. 27).

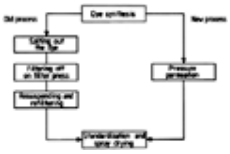

Figure 27. Flow diagrams of the old and new processes for production of a powdered dye

New Process (Fig. 28). An effective new method for reducing wastewater contamination uses pressure permeation. With suitable dyes, the salting-out process and isolation of dyes is replaced by salt removal and concentration of the reaction solution via membranes. The concentrated solution can then be processed further to give a directly marketable product.

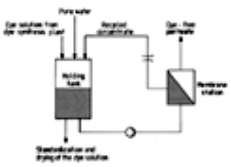

Figure 28. Flow diagram of a pressure permeation plant

Pressure permeation plants use semipermeable membranes that are permeable to water, inorganic salts, and small organic molecules, but quantitatively retain dyes in solution. According to the separating characteristics of the membrane and the resulting operating pressures to be used, separation processes are classified as microfiltration, ultrafiltration, nanofiltration, and reverse osmosis. The membrane surfaces are in the shape of spirally wound, plate-shaped, or tubular modules. The salt-containing synthesis solution passes from the reaction vessel into a holding tank and subsequently through the membranes under pressure. There it separates into a salt-containing, almost dye-free permeate, and the dye concentrate. The dye concentrate is recycled to the holding tank. After further treatment, the permeate can be reused as feedwater or otherwise disposed of as wastewater. Compared with the filtrate of the conventional process the permeate produced from the pressure permeation plant has a COD that is decreased by 80 % and a salt content (only reaction salt) decreased by 90 %. The filtration, washing, and redissolving steps, which are complex and labor intensive, are replaced by a fully automatic pressure permeation stage. Since retention of the dye by the membrane is almost quantitative, the dye yield is increased up to 5 % by using the new pressure permeation technique.

Ecological and Economic Aspects of the Process Change. By use of the pressure permeation technique, wastewater contamination can be decreased considerably and thus the following wastewater treatment simplified. Also, the almost complete removal of salts from dyes marketed in the form of highly concentrated solutions virtually eliminates the necessity of adding auxiliaries for standardization.

In addition, the economics are improved by reducing wastewater costs, while product quality remains the same or better, and the yield is sometimes higher. The application of pressure permeation to other fields is being investigated and developed:

1. Extension of the process to other applications (dye grades and intermediate products)
2. Optimization of the synthesis and production processes (Fig. 29)

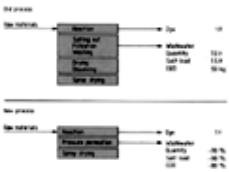

Figure 29. Residue minimization by pressure permeation

To sum up, introduction of the pressure permeation technique has the following economic advantages: (1) increase in yield, and (2) reduction of wastewater treatment costs; and hence an improved profitability with equal or better product quality.

2.2.2.6. *Fuel Replacement in Sewage Sludge Combustion by Utilization of Chlorinated Hydrocarbon Side Products [113-116]*

Introduction. In the chemical industry, much effort is devoted to preventing residue formation completely. Although great success has been achieved in the area of residue prevention, residue formation cannot be completely avoided in the chemical industry. Attention must therefore be directed toward the most feasible methods of utilizing these residues.

Old Processes. *Production of Sewage Sludge.* In the biological wastewater treatment plants of the Leverkusen works, the organic content is degraded by microorganisms in biological stages. In this process, sewage sludge is formed: first a preliminary sediment by chemical-mechanical purification, and then excess activated sludge from the biological wastewater purification stage.

This sludge has until now been treated by thickening, conditioning, and dewatering, and been disposed of as solid filter cake on the works-owned special waste dumps.

Production of Chlorinated Hydrocarbons (CHC). During the production of, e.g., rubber chemicals, quantities of highly chlorinated hydrocarbons result as byproducts, thus far unavoidably. These materials could not be disposed of at facilities on land and were incinerated in special ships on the high seas.

New Methods. *Incineration of Sewage Sludge.* To dispense with disposal of the organic materials in sewage sludge on the works-owned dump, to minimize the amount of water dumped in the sewage sludge filter cake (which contains 60 % water), and to utilize the available dumping space more efficiently, alternative methods of disposing of the sludge were investigated. The following criteria were formulated for comparing the alternative methods:

1. Volume reduction
2. Process safety
3. Environmental compatibility of the residues
4. Cost

Based on these decision criteria, the method of choice for disposing of sewage sludge was incineration.

Incineration of Chlorinated Hydrocarbons. Finding a way of integrating the incineration of highly chlorinated hydrocarbons into the sludge disposal concept was very important. From the political discussion generated by incineration of these liquid residues on the high seas, one expects only incineration on land to be permitted in the future.

Description of the Process (see Fig. 30). The sewage sludge incineration plant, which has a capacity of 80 000 t/a, is divided into the following sections:

1. Multiple-hearth furnace
2. Afterburner chamber
3. Waste-heat boiler for waste-heat utilization
4. Multistage flue gas purification

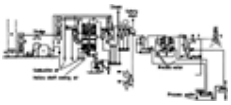

Figure 30. Schematic of sludge incineration

a) Sludge bunker; b) Elevator 1; c) CHC storage tank; d) Elevator 2; e) Multideck furnace; f) Ash removal; g) Transport container; h) Boiler feedwater tank; i) Condensate tank; j) Afterburner chamber; k) Waste-gas cooler; l) Quench tank; m) Rotary scrubber; n) Jet scrubber; o) Demister; p) Induction fan; q) Stack

The sewage sludge is conditioned; i.e., filter aids are added to convert it into a form with good filtering properties. The sludge is dewatered in chamber filter presses, stored temporarily in a bunker (a), and charged via trough feeders into the buckets of a transport plant (b, d). These buckets contain ca. 1.5 t of sludge. The sludge is charged by the buckets into a storage vessel on the top deck of the incinerator and is fed continuously from there into the furnace (e). A feed rate of up to 12 t/h sewage sludge can be achieved.

The cylindrical furnace consists of a steel casing with an interior lining and exterior insulation; it is divided into eight horizontal decks. The diameter of the furnace is 8 m, and its height 12 m.

In the center of the furnace is a vertical hollow rotating shaft with four rabble arms for each deck carrying scraper blades arranged at an angle. The rotary motion of the shaft causes the sewage sludge fed to the top deck to be swept from the inside to the outside, and on the next deck from the outside to the inside, falling through openings at the outer and inner edges of the decks to the deck below. The ash, which is produced on the two lowest decks, moves in the same way. The multideck furnace operates in countercurrent; i.e., the combustion air and hot flue gases pass through the furnace in an upward direction, drying the sludge, while the feed material is transported in the opposite direction. Also, the combustion air cools the ash, the air being thereby preheated, and the hot flue gases dry the sludge to be incinerated.

In this method of operation, three different zones are produced in the multideck furnace. From top to bottom, these are

1. Drying zone from deck 1 to deck 4, temperature 400 – 700 °C
2. Incineration zone in decks 5 and 6, temperature 800 – 1000 °C
3. Cooling zone in decks 7 and 8, 250 – 350 °C

At the side of the sixth deck is an uncooled combustion chamber lined with refractory materials. This is used to provide intense heat for starting up the furnace and, if required, can also be used for supplementary heating during incineration of the sludge.

Exhaust gases from the combustion zone and vapors from the drying zone of the multideck furnace pass into a lined afterburner chamber (j). Liquid organic waste (e.g., chlorinated hydrocarbons from production processes) is used as fuel in the afterburning chamber. To avoid chloride corrosion, the outer wall of the afterburner chamber is connected to the recycling system of the process gas cooler and thereby cooled. The completely burned hot waste gas then passes into the waste-gas cooler (k) where its sensible heat energy is used to produce steam. Combustion gases leave the waste-gas cooler at 250 – 300 °C and are then led to flue gas scrubbing. The ash produced is removed via a rotary valve (f) and fed through a chain trough conveyor (g) to the ash removal system of the multiple-hearth furnace for cooling and agglomeration. The ash produced is disposed of on the works-owned dump.

The hot flue gases at ca. 250 °C are cooled to 70 – 80 °C in the quench tank (l) by spraying with water, and the gaseous substances (e.g., hydrogen chloride and hydrogen fluoride) and dust particles are washed out. Any remaining fine dust and hydrochloric acid are removed from the flue gas by a two-stage rotary scrubber (m).

The flue gas is then treated with sodium hydroxide solution in a jet scrubber (n) to remove sulfur dioxide. The purified flue gas passes through a demister (o) and an induced draft fan (p), and into the stack (q).

A comparison of the flow sheets of the old and new processes is given in Figure 31.

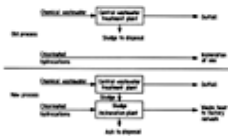

Figure 31. Comparison of old and new wastewater sludge treatment and chlorinated-hydrocarbon incineration processes

Outlook. More than 50 % of the sewage sludge produced can now be incinerated. It is expected that all sewage sludge will be disposed of in this way by 1997.

2.2.3. *Examples from BASF*

2.2.3.1. *Emission Reduction in Industrial Power Plants at Chemical Plant Sites by Means of Optimized Cogeneration [117-119]*

Industrial power stations represent some of the largest individual emitters of a number of pollutants from production sites. The most important of these pollutants are CO_2, CO, SO_2, NO_x, and dust.

The 1972 figures for the BASF Ludwigshafen site, which contains more than 300 production plants, are given below:

Total emissions	113 000 t/a
From energy generation	55 000 t/a
From chemical production	58 000 t/a
Individual emissions	
CO	30 000 t/a
NO_x	30 000 t/a
NH_3	2 000 t/a
C_5H_{12}	2 300 t/a
HCl	500 t/a
Dust	6 300 t/a

The primary task of industrial power stations is to generate heat energy for the production processes. On large sites, this heat energy has customarily been produced in power – heat cogeneration plants, so that part of the electrical energy requirement is also produced on-site. These cogeneration plants are mainly steam power plants with characteristic values of 0.15 –0.25 MW · h_{el}/(MW · h_{th}). These characteristic values on the producer side must be compared with those on the consumer side. The following values are typical:

Sites producing their own basic chemicals	0.4 – 0.8 MW · h_{el}/(MW · h_{th})
Other sites	0.25 – 0.35 MW · h_{el}/(MW · h_{th})

Comparison of these figures shows that cogeneration plants using steam-based processes cannot supply the electricity requirements of the consumer on the basis of power – heat cogeneration. On the other hand, small condensation power plants on industrial sites are generally not competitive with the large installations of electricity supply companies. This follows from the decline in capacity of industry-based power stations in West Germany (Fig. 32), which is due entirely to the closure of condensation plants in the industrial area. The shortfall in electrical energy is made up for by purchasing.

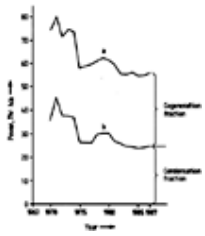

Figure 32. Breakdown of industrial electricity generation by prime mover type

a) Back pressure, tapped back pressure, gas turbines, piston engines; b) Condensation machines

Several methods are used to reduce emissions from industrial power stations:

1. Lowering the energy consumption of production plants
2. Removing pollutants from flue gases of industrial power stations
3. Using alternative fuels
4. Using alternative technologies for energy conversion

An economic and ecological optimum is achieved only by a “tailored” combination of all four methods on any individual site.

Lowering Energy Consumption. Attempts to reduce energy consumption of production plants have been very successful in past years. Here, the pinch analysis method of LINNHOFF has proved very effective (⟶).

Pinch analysis is a method of assessing the works infrastructure of production plants and supply systems with respect to energy consumption, quantities of waste products, quantities of emissions, and capital costs. The fundamental elements of pinch analysis are the composite curves, the grid diagram, and the pinch itself.

In the composite curves (Fig. 33), the cold and hot streams of a production plant are plotted against temperature. When the two curves are shifted horizontally relative to each other until the minimum temperature difference necessary for heat-transfer processes is reached at the point where the curves come closest together (pinch), this gives the theoretical minimum for the heat streams that must be supplied or removed (hot and cold utility targets; ⟶ Pinch Technology – Composite Curves and Energy Targets).

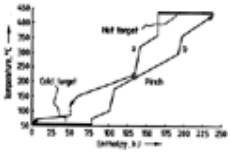

Figure 33. Hot and cold composite curves

a) Hot curve; b) Cold curve

The pinch represents the boundary between the balanced heat sources (system below the pinch) and heat sinks (system above the pinch). In the optimum case, no heat transfer occurs across the pinch. The grid diagram (Fig. 34) is used in the development of economical heat recovery systems using the pinch technology.

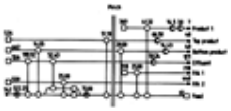

Figure 34. Grid diagram

The result of systematic assessment of the BASF Ludwigshafen site is shown in Figure 35. The amount of sales products increased by 143 % from 1972 to 1992; however, the factory steam consumption decreased slightly. In 1992, only 33 % of the steam requirement was supplied from primary energy. Whereas the fraction of steam produced by combustion of residues remains the same, the largest fraction today is supplied by the utilization of waste heat from chemical plants. Consequently the consumption of primary energy dropped drastically.

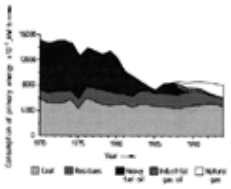

Figure 35. Consumption of primary energy at the Ludwigshafen works

Emission Reduction. The reduction in emissions from power plants was correspondingly drastic, and the more than proportional decrease in the use of the primary energy carrier fuel oil has had an additional effect.

At chemical sites producing sulfuric acid, further reduction in SO_2 emissions of power plants is possible using the Wellman – Lord process (⟶ Sulfur) in which pure SO_2 is recovered and recycled to the production process. The power plant thus supplies starting materials to the chemical production plant, and becomes an integral part of it. In the Wellman – Lord process, SO_2 is absorbed in aqueous sodium sulfite solution with formation of sodium bisulfite:

$$SO_2 + Na_2SO_3 + H_2O \rightarrow 2\,NaHSO_3$$

The SO_2 is then desorbed by heat in a later stage.

Other additive purification processes carried out on exhaust gases led to the decrease in atmospheric emissions at the Ludwigshafen site shown in Figure 36. Emissions of individual components for the year 1992 are shown in Table 1.

Table 1. Atmospheric emissions at the Ludwigshafen site in 1992

	Power plants, t/a	Total, t/a
CO	160	12 500
NO_x	240	6 000
SO_2	1350	2 300
NH_3	0	600
C_5H_{12}	0	100
HCl	26	140
Dust	200	900
Total emission	3500	24 000

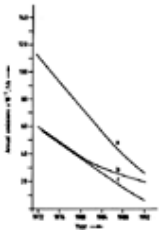

Figure 36. Atmospheric emissions from the BASF Ludwigshafen works (excluding CO_2)

a) Total emissions; b) Emissions from power plants; c) Emissions from production plants

Use of Alternative Technologies for Energy Conversion. A new cogeneration technology using natural gas as the energy carrier has also led to a reduction in CO_2 emission. Figure 37 shows the principal arrangements of a conventional steam turbine power plant (A), a gas turbine power plant (B), and a combined gas and steam turbine power plant (C). The range of characteristic values [$MW \cdot h_{el}/(MW \cdot h_{th})$] is also shown in each case. Industrial power stations are fundamentally heat controlled. In a modern cogeneration process using gas turbines in this way ca. three to four times the amount of electrical energy is produced compared with the steam turbine process.

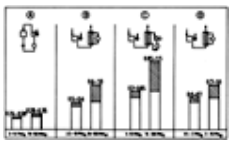

Figure 37. Characteristic values for typical cogeneration processes

A) Combined heating and power stations with steam turbines; B) Combined heating and power stations with gas turbines; C) Combination plant (without supplementary firing); D) Block-type thermal power station with gas engines

So production plants can meet their total demand of electric power and even sell a considerable amount to the public. This also leads to an overall reduction of emissions.

2.2.3.2. *Closed-Cycle Wittig Reaction [120-123]*

The Wittig reaction is widely used in the production of vitamins, carotenoids, pharmaceuticals, and antibiotics. It proceeds, e.g., according to the following scheme:

$$R^1-CH_2-Cl + P(C_6H_5)_3 \rightarrow R^1-CH_2-P^+(C_6H_5)_3Cl^-$$

$$R^1-CH_2-P^+(C_6H_5)_3Cl^- \xrightarrow[-HCl]{Base} R^1-CH=P(C_6H_5)_3$$

$$R^1-CH=P(C_6H_5)_3 + R^2-CHO \rightarrow R^1-CH=CH-R^2 + O=P(C_6H_5)_3$$

The Wittig reaction is characterized by mild reaction conditions and high yields. It therefore meets the criteria for clean (smooth) chemistry. However, its more widespread use is hindered by the fact that equimolar amounts of triphenylphosphine (TPP) must be used, and the inactive triphenylphosphine oxide (TPPO) formed in the reaction is today still disposed of as waste.

The active triphenylphosphine is synthesized according to the overall equation

$$PCl_3 + 3\,C_6H_5Cl + 6\,Na \rightarrow P(C_6H_5)_3 + 6\,NaCl$$

Since TPPO is not easily degraded and phosphorus must not be present in wastewater, the only possible disposal route is as follows (Fig. 38). The TPPO-containing residues are incinerated to form P_2O_5, which is removed in a scrubber with the aid of dilute sodium hydroxide solution. Phosphorus is then precipitated as calcium phosphate and filtered off, and the filter cake is disposed of on a hazardous waste dump. This disposal method has the following disadvantages:

1. The filter cake uses valuable dumping space
2. The P_2O_5 aerosols have a poisoning effect on the DENOX catalysts of the incineration plant
3. The P_2O_5 aerosols cause clogging of cloth filters

Figure 38. Flow sheet for disposal of TPPO

Attempts were therefore made to convert inactive TPPO back to active TPP, which could then be used again in the Wittig

reaction. A route along the following lines was ultimately found to be successful (Fig. 39):

Figure 39. Flow sheet for treatment of TPPO

The TPPO-containing residues are dissolved in chlorobenzene and reacted with phosgene gas to form triphenylphosphine chloride, with liberation of CO_2. In the second step, finely divided aluminum powder is added with mixing. The TPP · $AlCl_3$ complex formed is then decomposed hydrolytically in a third reaction step, giving the desired TPP.

A flow diagram of the process is shown in Figure 40. The TPPO-containing residues from various operations are collected. Residues from other companies can also be included at this point. Pure TPPO is recovered by distillation and dissolved in chlorobenzene.

Figure 40. Schematic of the treatment process of TPPO

Chlorination is carried out in a stirred-tank reactor inside a gastight chamber. Phosgene is also generated in this chamber and fed immediately to the reactor.

Dehalogenation with aluminum powder is carried out in the next stirred-tank reactor.

The TPP · $AlCl_3$ complex is then hydrolytically decomposed with water in another stirred tank in which the pH is adjusted with HCl so that water-soluble $Al(OH)Cl_2$ is formed. The TPP is dissolved in the chlorobenzene layer. Traces of TPP are recovered from the aqueous phase by extraction with chlorobenzene. In a subsequent distillation stage, chlorobenzene is distilled off and recycled. The TPP is then purified by distillation. The aqueous phase from the extraction stage is fed to a sewage treatment plant, which operates in the weakly alkaline range, so that aluminum is precipitated as $Al(OH)_3$ and acts as a flocculation agent in the desired manner.

The ecological advantages of the new process can be seen from a comparison of the ecological balances (see Fig. 41):

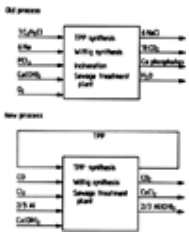

Figure 41. Ecological balances of old and new processes

This comparison shows that the new process results in the complete absence of phosphorus emissions, a reduction of the chloride load of the wastewater to one-third of its former value, and a reduction of CO_2 emission to only ca. 5 % of its former value. The new low-emission process for the production of TPP enables the Wittig reaction to be used much more widely in industrial syntheses.

2.2.4. *Integrated Environmental Protection and Energy Saving in the Production of Vinyl Chloride (Example from Wacker Chemie)*

Introduction. The production of vinyl chloride monomer (VCM), starting material for poly(vinyl chloride) (PVC) has a long tradition. Raw materials for VCM production are ethylene, chlorine, air (oxygen), and sometimes external hydrogen chloride (HCl) from other chlorination processes (⟶ Chlorinated Hydrocarbons – Production, ⟶ Chlorinated Hydrocarbons – Vinyl Chloride from Acetylene, ⟶ Chlorinated Hydrocarbons – Vinyl Chloride from 1,2-Dichloroethane). The main production steps are shown in Figure 42. This integrated process contains two internal material cycles, namely, the closed-loop recycling of nonconverted 1,2-dichloroethane (EDC) from the pyrolysis step (so-called recycle EDC) and the closed- or partly open-loop recycling of HCl formed in the cracking step back to the oxychlorination unit. With totally closed-loop recycling of the crack HCl — or if HCl must be exported due to quality reasons or the exported quantity made up by HCl from external sources — the combined VCM process runs “HCl balanced.” However, if crack HCl is exported

without compensation, the combined VCM process becomes “HCl unbalanced.”

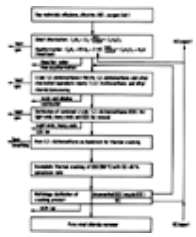

Figure 42. Main steps in production of vinyl chloride monomer

Differences in the processes used by individual VCM producers are due mainly to the different technologies applied in direct chlorination; these include high- or low-temperature direct chlorination (both processes are operated in the homogeneous liquid phase in the presence of oxygen as substitution inhibitor), and oxychlorination (fixed- or fluidized-bed process in the heterogeneous gas phase). Differences in these steps determine the energy requirements and the design of equipment for processing and purifying the intermediate products. The latter is as complex as pure product isolation because the reaction products formed in EDC synthesis and EDC cracking consist of multicomponent mixtures.

Conventional Production Process. Figure 43 shows the overall process flow and mass balance of the old process in operation since mid-1976. Target products in addition to VCM are hydrogen chloride for silicon chemistry and 20 % hydrochloric acid for internal site demand. The combined VCM process was nearly “HCl balanced” because the HCl flow exported to silicon chemistry was balanced by imported HCl from a chlorolysis process (tetrachloroethylene production plant). The amount of byproducts was all in all 4.4 wt % related to VCM production. According to JENSEN et al. [124], ca. 64 % of these residues are heavy ends and EDC – VCM tars which were incinerated internally or partly externally (depending on the uptake capacity of the rotary kiln incinerator) with heat recovery but without utilization of their considerable chlorine content. This resulted in a high salt load in the flue gas scrubber effluent. The light ends, ca. 36 % of the total residue amount, were used as secondary raw material for the production of tetrachloroethylene by a perchlorination process. Because this perchlorination process also produced hexachlorobenzene- (HCB) and hexachlorobutadiene- (HCBD) containing wastes — 3600 t of light ends form 345 t of those solid wastes — which had to be disposed at great expense on an underground deposition site, this way of utilizing light ends finally was only a transposed residue recycling without a really closed loop.

Figure 43. VCM production according to the HCl-balanced old process

a) Direct chlorination; b) Acidic wash; c) Recycle EDC chlorination and dechlorination; d) VCM cracking; e) Oxychlorination; f) Acidic and alkaline wash; g) EDC distillation; h) Light ends and H_2O removal; i) Heavy ends and tar removal; j) Chlorination of vent gas from oxychlorination; k) Vent gas incinerator with HCl recovery (Figures are given in 10^3 t/a unless otherwise specified)

The large amounts of toxic vent gases [containing, besides partly cancerogenic chlorohydrocarbons (CHCs), compounds such as carbon monoxide, chlorine, HCl, and hydrocarbons] are produced predominantly during oxychlorination with air. Further emissions are due to various process vent gases and tank breathing. Despite the large gas volume, these vent gases have been treated since 1977 by thermal incineration in an integrated HCl recovery plant (first plant in Europe, Fig. 44) because catalytic incineration, which consumes much less energy, was not practicable due to the high chlorine content of these vent gases (chlorine acts as a strong catalyst poison). Based on the low calorific value of these vents, considerable amounts of natural gas had to be used as auxiliary fuel to guarantee complete carbon combustion at 950 °C and 0.6 – 0.8 s residence time. Most of the waste heat of the flue gas was utilized for the generation of steam at 20 bar by a waste-heat boiler. The mostly organically bonded chlorine in the vents was recovered as 20 % hydrochloric acid in the two-stage flue gas scrubber. To satisfy the site demand of 20 % hydrochloric acid, additional crack HCl had to be introduced into the vent gas incinerator.

Figure 44. Vent gas incineration with HCl recovery

a) Combustion chamber; b) Waste-heat boiler; c) Boiler feedwater preheater; d) Wash column; e) Quenching vessel; f) Demister

Wastewaters produced in the different process stages are (1) reaction water of the oxychlorination process; (2) $FeCl_3$-containing wastewater from the acidic wash decanter of the direct chlorination; (3) slightly alkaline wastewater from azeotropic drying of wet crude EDC in the light ends column; (4) sodium-hypochlorite-containing spent caustic from the flue gas scrubber of the incineration plant; (5) strongly alkaline wastewater from the VCM wash decanter and VCM drying traps; and (6) wastewater from the acidic and alkaline wash decanter of the oxychlorination unit with high COD. Additional surficial wastewater is collected in plant and tank farm pits. This wastewater is more or less strongly contaminated with CHCs. An exception is the spent caustic from the incineration plant. The chlorinated hydrocarbons dissolved in wastewater were recovered by steam stripping and recycled back to the process. Three steam strippers were in operation: one for the surficial wastewater collected in several pits; a second stripper for $FeCl_3$-containing wastewater from the acidic wash decanter of direct chlorination (which has a low COD); and a third stripper (oxy stripper) for acidic reaction water from oxychlorination (which has a high COD), for the COD-rich wastewater from the acidic or alkaline wash decanter of the oxychlorination, for wastewater from the azetropic drying of the wet crude EDC, and for wastewater from the VCM wash decanter and drying traps. The effluents of strippers 1 and 2 were led into a chemical – mechanical sewage treatment

plant ($FeCl_3$ is a desired flocculant in the neutral or slightly alkaline range); the alkaline salt-containing oxy stripper effluent was led into a one-stage biological sewage treatment plant. The effluent from the oxy stripper was still contaminated with a relatively high COD and adsorbable organic halogen (AOX) load.

Besides this high energy-consuming wastewater pretreatment, the combined VCM production process has a high steam consumption in the various processing steps that are necessary for thermal separation of the diverse intermediate material mixtures. Despite steam credit for the integrated heat recovery in the central vent gas incineration plant and for recovery of reaction heat from the exothermic oxychlorination reaction (generation of steam at 20 bar) and the exothermic chlorination of oxychlorination vent gas (generation of steam at 1.5 bar), the specific steam consumption amounts to 1.74 t per tonne of VCM with 55 % conversion rate at EDC cracking.

The consumption of electrical energy for pumps, for the refrigeration unit, for the air compressor, and for the combustion air and flue gas suction fan of the vent gas incineration plant was 136 kW · h per tonne of VCM.

Large amounts of natural gas are necessary for firing of the EDC cracking furnace, and for the generation of radiant heat as heat flux for the endothermic cracking process, and as auxiliary fuel for the vent gas incineration plant. Thereby, most of the sensible heat of the superheated reaction product leaving the cracking furnace and of the waste heat of the flue gases used for firing the cracking furnace was lost to the environment, resulting in a specific natural gas consumption of 954 kW · h per tonne of VCM.

Despite the close integration of the VCM process with chlorine-producing and chlorine- or HCl-consuming plants, the old VCM production process was not quite satisfactory in economical or ecological terms.

New Process for Integrated VCM Production. The process changeover aimed at the following objects:

1. Increase of yields, i.e., minimization of formation of byproducts by more selective reactions
2. Utilization of all CHC-containing residues produced in the integrated process with chlorine recycling and heat recovery in a totally closed material recycling loop
3. Reduction of wastewater flow with simultaneous minimization of COD, AOX, and salt load
4. Energy savings in the distillation processing of the particular material streams, and in the EDC cracking unit
5. Energy savings in the vent gas incineration unit by import of external CHC-containing vent gases with high calorific value and by improving the thermal efficiency of the vent gas incineration unit.

Figure 45 depicts the overall process flow and mass balance of the new process built at the end of 1991. The new process became “HCl unbalanced” by crack HCl export to silicon chemistry after special purification [125] and simultaneous shutdown of the perchlorination process for economical and ecological reasons. All residues containing chloroorganic compounds are thermally treated under chlorine and heat recovery in the new HCl generation plant (Fig. 46) [126].

Figure 45. VCM production according to the HCl-unbalanced new process

a) Direct chlorination; b) Acidic wash; c) Benzene removal; d) Recycle EDC chlorinaton and dechlorination; e) VCM cracking; f) Oxychlorination; g) Acidic and alkaline wash; h) EDC distillation; i) Light ends and H_2O removal; j) Heavy ends and tar removal; k) Chlorination of vent gas from oxychlorination; l) Byproduct recycling (BPR) plant (oxygen incinerator); m) Vent gas incinerator with HCl recovery (Figures are given in 10^3 t/a unless otherwise specified)

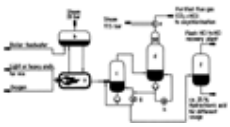

Figure 46. HCl-generating plant

a) Reactor (combustion chamber); b) Steam drum; c) Quenching vessel; d) Washer; e) Heater; f) Flash drum; g) Cooler; h) Cooler (brine at –15 °C)

The process is based on the following chemical reaction:

$$C_xH_yO_zCl + \left(x + \frac{y-1}{4} - \frac{z}{2}\right) O_2 \rightarrow x\,CO_2 + HCl + \left(\frac{y-1}{2}\right) H_2O$$

$$\Delta H^R_{25\,°C} = -12\,560 \text{ to } -14\,650 \text{ J/kg}$$

Because of an H : Cl atomic ratio of >1.3 : 1 in the residues to be incinerated, the reaction is autothermal and needs no auxiliary fuel.

With a 4 – 6 % oxygen excess related to stoichiometry — required for optimum carbon combustion, i.e., to avoid CO formation — the temperture dependent Deacon equilibrium reaction is shifted to the right:

$$4\,HCl + O_2 \rightleftharpoons 2\,Cl_2 + 2H_2O$$

The following reaction products are formed:

1. Flue gas with the following composition is fed to the oxychlorination unit without inert gas separation:

CO_2	56.8 vol %
Cl_2	1.1 vol %
HCl	32.6 vol %
O_2	3.2 vol %
N_2	6.3 vol %

2. 25 % hydrochloric acid for general usage
3. Flash HCl which is sent to the HCl recovery plant in the gaseous state

The chlorine recovery rate as gaseous HCl is ca. 90 %. The steam production is ca. 4.4 t per tonne of residue, which amounts to ca. 85 % heat recovery related to the calorific value of the residues.

The CHC residue mix is converted to gaseous HCl by reaction with oxygen at a pressure of 8 bar gauge and at an extremely high temperature (up to 2500 °C in the flame). The HCl formed is recycled to the oxychlorination unit without expensive compression or inert gas separation steps [126-128]. Use of pure oxygen instead of air has the advantages of an extremely high flame temperature and very long residence times in the combustion chamber. This guarantees (1) quantitative conversion of CHC to CO_2, H_2O, and HCl (in addition to traces of chlorine according to the Deacon equilibrium reaction), and (2) complete destruction of polychlorinated biphenyls (PCB) and polychlorinated dibenzodioxins and dibenzofurans (PCDD, PCDF) being present in trace amounts in the heavy ends and EDC – VCM tars [129] and also preventing their de novo synthesis (destruction rate for PCB: 99.99999 %, for PCDD/PCDF: 99.992 %). By operating this HCl generation plant, the material recycling loop is totally closed. To save auxiliary fuel (natural gas) and satisfy the internal site demand for 20 % hydrochloric acid, part of the liquid light ends and fuel-rich CHC-containing vent gas from other plants are incinerated in addition to the VCM vent gases in the central HCl recovery plant with heat recovery. By installation of a boiler feedwater preheater [130], the thermal efficiency of the vent gas incinerator was strongly enhanced and the water vapor content in the purified flue gas drastically reduced. Even though liquid light ends were incinerated in addition, the PCDD – PCDF content in the purified flue gas is less than 0.1 ng/m^3 (STP) calculated as international toxicity equivalents (ITEQ). Besides, the so-called fugitive emissions were extensively reduced by technical measures such as canned pumps, grooved gaskets, emission-free sampling devices, and sucking off open drains and wastewater collection pits in the direction of the vent gas incinerator. Thus, the total output of toxic compounds into air could be reduced by 90 % in combination with technical measures for increasing the on-stream factor of the HCl recovery plant (e.g., by improving burner control and selecting more resistant construction material). Improvements in the EDC washing unit [131] led to a drastic reduction of the COD load of the wastewater. Process improvements in the oxy stripper led to steam savings and a nearly quantitative hydrolysis of chlorinated ethanol to glycol and glycolaldehydes, followed by essential reduction of wastewater AOX load. The overall stripping efficiency was increased drastically. Surficial wastewater is collected in a 200-m^3 basin and purified in an adsorber (Fig. 47). The new process saves energy (steam consumption only in the desorption step) and is very efficient in terms of AOX reduction and elimination [132]. All these improvements resulted in a higher overall purification efficiency with a reduction of the total wastewater flow of about 25 % (see below):

Old process

Alkaline stripper for COD-containing wastewater

Acidic stripper for inorganic wastewater ($FeCl_3$ containing) with low COD

Neutral stripper for surficial wastewater

Wastewater flow:	235 577 t/a
CHC recovery:	2 538 t/a
CHC composition:	46.7 wt % 1,2-$C_2H_4Cl_2$
	49.2 wt % $CHCl_3$
	2.0 wt % C_2H_5Cl

Remainder: other chlorinated C_1 – C_4 hydrocarbons

Purification efficiency:	99.75 %

New Process

Improved alkaline stripper for COD-containing wastewater

Improved acidic stripper for inorganic wastewater ($FeCl_3$ containing) with low COD

Adsorptive purification of surficial wastewater at pH≈1

Wastewater flow:	178 780 t/a
CHC recovery:	1 156 t/a
CHC composition:	75.2 wt % 1,2-$C_2H_4Cl_2$
	22.0 wt % $CHCl_3$

	1.6 wt % C_2H_5Cl
Remainder: other chlorinated $C_1 – C_4$ hydrocarbons	
Purification efficiency:	99.95 %

Simultaneously, the biological sewage treatment plant was extended by a second stage that additionally treats municipal wastewaters (improving the AOX elimination rate) and by a biosludge (fluidized-bed) incinerator with flue gas purification. The CHC of the return cooling water and steam condensate of the VCM plant is measured by on-line analyzers (semiconductor sensor, FID) to recognize small leakages very quickly. In connection with operating instructions for directives in case of a leak, CHC input into the main outfall ditch because of leakages is drastically decreased. All these measures contribute to keeping the AOX content in the effluent of the biological sewage treatment plant distinctly below the existing regulatory emission levels (max. 20 g AOX per tonne of pure EDC or max. 1 mg AOX/L, or max. 5 g EDC per tonne of pure EDC or 2.5 mg EDC/L, both as monthly averages).

Production-integrated improvements have been achieved by long-term research in the field of chemical catalysis and radical chain pyrolysis. The improvements are more selective catalysts for oxychlorination with increased activity [133] (fewer byproducts, higher HCl conversion, lower degree of ethylene oxidation rate), an improved catalyst for chlorination of the oxychlorination vent gases [134] (nearly quantitative ethylene conversion by increased activity and, because of increased selectivity, much lower formation of chlorinated ethanols that interfere with the VCM process); and a totally changed heat flux system for EDC pyrolysis [135] which, by using the so-called reverse heat flux technique, reduces the amount of byproducts and coke produced even at cracking depths around 60 %. By chlorination of a small portion of recycle EDC a benzene concentration of <500 ppm in the feed EDC could be achieved [136]. This influences the cracking kinetics favorably and less coke and VCM tar were produced. Chloroprene removal from total recycle EDC was distinctively improved [137]. Further ressource saving and byproduct lowering effects were obtained by improving acetylene removal from cracking HCl by selective hydrogenation [138] and by amending the reaction parameters of direct chlorination [139].

With the pinch technology (⟶), important energy savings were obtained. At first, all material streams had to be summed up. Based on this, hot process streams with a technically utilizable temperature level were coupled with heat-absorbing process steps [135], [140]. Also the steam consumption of EDC distillation and EDC vaporization into the cracking furnace could be reduced further [131], [135], [137]. In combination with an open heat pump for reboiling the especially energy-consuming heavy ends column (so-called vapor thermocompression) [141], [142], the specific steam consumption of the VCM process has been lowered to 0.25 t per tonne of VCM including steam production in the oxychlorination unit and in the HCl recovery and HCl generation plant. Further steam could be saved as a side effect of the thermocompression in the VCM column, in the liquid feed EDC preheat before entering the cracking furnace, and in the EDC vaporization into the cracking furnace. The increased electrical power consumption for vapor recompression (ca. 45 kW · h per tonne of VCM) is more than compensated by the overall energy savings. These overall energy savings reduce the consumption of natural gas by a total amount of 24.5×10^6 m^3 (STP)/a, resulting in less emission of CO_2 (51 940 t/a) and $(NO)_x$ (21.9 t/a).

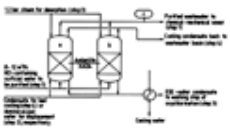

Figure 47. Flow sheet of an adsorber plant (specific steam consumption: 42 kg per cubic meter of wastewater, 70 % less than a steam stripper with feed-bottom heat exchanger)

a) Working adsorber, 2 m^3; b) Standby adsorber, 2 m^3; c) FID analyzer for breakthrough determination

Ecological and Economical Evaluation of the New Process. By the economically and ecologically significant utilization of residues from the integrated VCM production process, an amount of chlorine is saved that corresponds to the flow of recycled HCl. In addition, HCl or salt-containing wastewaters are avoided. Due to the combination of the HCl recovery and generation plant and the AOX pretreatment of wastewater with recycling of CHC removed from wastewater, the new process for manufacturing EDC and VCM is nearly emission and residue free. The HCl generation plant makes the VCM process totally self-sufficient in terms of residue utilization. In combination with measures for process optimization that save ressources and energy, the specific input data and the environmental pollution balance could be demonstrably improved in relation to the old process (Table 2). The potential for the formation of chlorinated dibenzodioxins and -furans has been reduced. Remaining amounts are destroyed almost quantitatively. A shift of dioxin emissions to residues and wastewater is therefore avoided. These successfully operating production-integrated improvements for environmental protection are fundamentally based on the combined material balance with closed-loop material recycling. High investment costs were necessary for the chemical engineering of the new integrated process, including new equipment, and for proper design of the operation and reaction control parameters supported by a modern computerized process control system with video terminals. Extensive research and development work over years required additional capital costs for thorough investigation of each process step.

Table 2. Comparison of specific input and output data in VCM production (Target products are each 225 000 t/a VCM, 30 000 t/a HCl for export, and 25 800 t/a 20 % hydrochloric acid for total site demand)

Substance or energy	Old process	New process	Change, %
Input			
Ethylene, kg per tonne VCM	478	471	–1.5
Chlorine[a], kg per tonne VCM	628.8	592.4	–5.8

Electrical power, kW · h per tonne VCM	136	188	+38.2
Steam, t per tonne VCM	1.74	0.25	–85.6
Natural gas, kW · h per tonne VCM	954	835	–12.5
20 % Hydrochloric acid, kg per tonne VCM	6	3	–50.0
Caustic[b], kg per tonne VCM	40	15	–62.5
Process water[c], m^3 per tonne VCM	0.86	0.56	–34.9
Output			
Coke, g per tonne VCM	90	45	–50.0
Wastewater, m^3 per tonne VCM	1.05	0.77	–26.7
COD, kg per tonne VCM	15	1.3	–91.3
AOX, kg per tonne VCM	0.35	0.14	–60.0
NaCl[b], kg per tonne VCM	50	15	–70.0
Vent gas purified, m^3 (STP) per tonne VCM	1220	979	–19.8
Residue (heavy ends, tar), kg per tonne VCM	28	0	–100.0
CO_2[d], kg per tonne VCM	776.7	588.7	–24.2
$(NO)_x$ as NO_2[d], g per tonne VCM	585.9	437.9	–25.3
∑ Toxic compounds (air), kg per tonne VCM	4.9	0.5	–89.8

[a] Including HCl export – import and production of 20 % hydrochloric acid.
[b] Neutralization of oxychlorination reaction water, wastewater of EDC acidic wash, surficial wastewater, or acidic wastewater from flue gas scrubber of heavy ends – tar incineration.
[c] Demineralized water, however, without boiler feedwater for internal and external steam production.
[d] Integrated vent gas and liquid residue incineration, external residue incineration (old processes), cracking furnace firing, and electrical power and steam production in a thermal power station required for VCM production.

2.2.5. *Examples from Hüls*

2.2.5.1. *Integrated Environmental Protection in Cumene Production* (⟶ Hydrocarbons – Cumene)

Introduction. Hüls has produced cumene since 1954 at its site in Marl. The production capacity is 150 000 t/a. Cumene is used mainly in the production of acetone and phenol, and also in the production of cumene sulfonate, which is a raw material for detergents. It is produced by Friedel – Crafts alkylation of benzene by propene, catalyzed by aluminum chloride, which does not give cumene exclusively, but also produces di- and triisopropylbenzene.

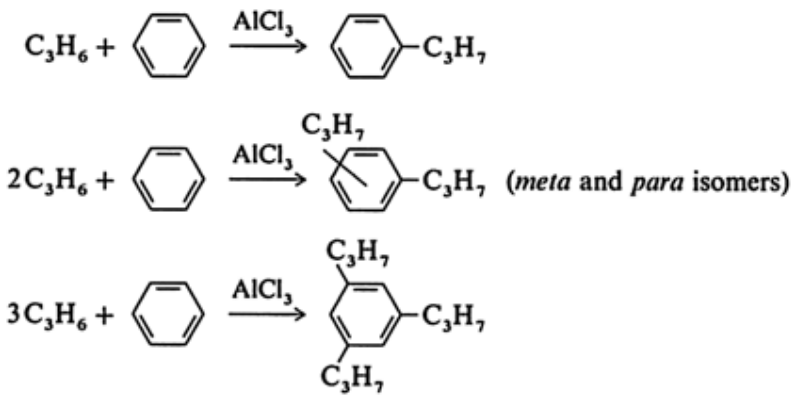

Other side reactions also occur to a small extent. Whereas the polyalkylated isopropylbenzenes can be converted to the target product cumene by using aluminum chloride as the catalyst, other byproducts, which include a high proportion of ethylbenzene, must be removed from the reaction mixture.

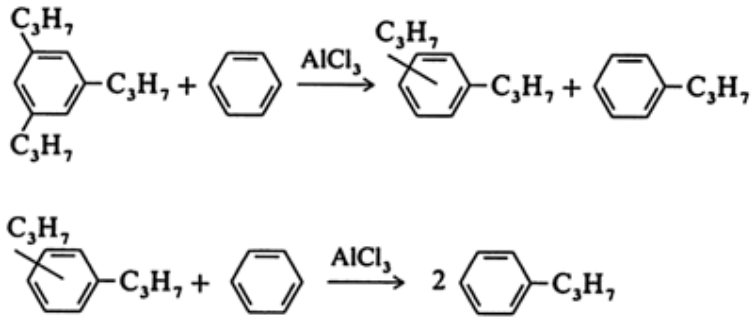

Conventional Method of Producing Cumene (Fig. 48). In the past, the Friedel – Crafts alkylation of benzene was carried out together with the conversion of the polyalkylated isopropylbenzenes in a single reactor (a). The catalyst was produced directly in the reaction mixture from aluminum chippings and HCl. It formed a second separate organic phase in the

reactor. The water present in the propene was simply removed mechanically, whereas the benzene used was dried over sodium hydroxide.

After removal from the reactor, the catalyst was decomposed by addition of water, and the liberated aluminum chloride was extracted (b) into the aqueous phase. However, the amount of water required was determined not by the extraction parameters but by the decrease in temperature required for the water washing. Thus, a high excess of cold water was added to the hot reaction mixture to obtain the desired mixing temperature.

Subsequent distillative processing (c) led to the following

1. The target product cumene
2. Polyalkylated isopropylbenzenes, which were returned to the reactor
3. Low-boiling products, which were used as motor gasoline
4. Medium-boiling products, which were processed within the Hüls organization to produce synthesis gas
5. High-boiling products, which were used first as solvents in the production of styrene and then as raw materials for the production of synthesis gas

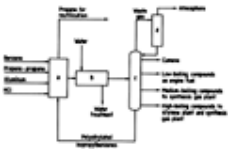

Figure 48. Cumene production — old process

a) Reactor; b) Extraction; c) Distillation; d) Purification

New Process for Cumene Production [143] (Fig. 49). The new process for the production of cumene was introduced between 1986 and 1988. Its most important features were an increase in the selectivity of cumene formation and suppression of undesired side reactions that do not lead to monoalkylated or polyalkylated isopropylbenzenes. This was achieved by the following modifications to the process:

1. Catalyst production in a separate reactor (a)
2. Avoidance of a second organic catalyst phase in the alkylation reactors (b)
3. Separation of the alkylation of benzene by propene from the reaction of the recycled polyalkylated isopropylbenzene with benzene
4. Complete dehydration of the starting materials benzene and propene by distillation

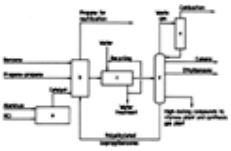

Figure 49. Cumene production — new process

a) Catalyst production; b) Reactors (several stages); c) Extraction; d) Distillation; e) Purification

An improvement in the distillative processing of the reaction mixture enabled ethylbenzene to be isolated in a pure state and used as starting material in styrene production.

Because of the total dehydration of the starting materials, catalyst consumption and the associated contamination of wastewater by inorganic substances were reduced. The amount of wastewater could be lowered by separating the cooling of the reaction mixture from aqueous extraction of the catalyst. Since the low temperature necessary for extraction could now be adjusted by indirect cooling, the amount of wastewater produced is determined only by the operating parameters for the extraction. The water produced in the extraction can be recycled many times.

Due to the introduction of heat recovery in the distillation process, the energy consumption of the entire process was decreased. Whereas formerly the heat of condensation was removed without utilization, it is now used to produce process steam and hot water, which are fed to the supply networks of Hüls at Marl.

Results. A comparison of the old cumene process (1980) with the new (1990) shows that the amount of wastewater is now 5 % of the former amount, the organics content of the wastewater after passing through the water treatment plant is 1 % of its former value, and the amount of residues has been reduced to 50 % of the earlier figure. Energy consumption, despite additional requirements for distillative processing of the starting materials, has been reduced by 60 %. The gaseous emissions of organic substances, which were already very low, have been prevented completely by passing the waste gases through an existing incinerator.

2.2.5.2. *Production of Acetylene by the Hüls Plasma Arc Process (* ⟶ Acetylene – Production from Liquid Hydrocarbons (Plasma Arc Process)*; ⟶)*

Production and Use of Acetylene. As a result of its strongly unsaturated character, acetylene reacts readily with many compounds. During the early development of industrial organic chemistry, it was one of the most important starting materials for a large number of products. Since the 1940s (United States) and the 1950s (Europe), it has been largely replaced by ethylene and other olefins. Bulk products still produced from acetylene include butanediol and vinyl compounds (vinyl chloride, vinyl ethers, and vinyl esters).

An important feature of acetylene production is the high energy requirement due to the formation of a triple bond. In the Hüls plasma arc process, the energy is supplied by an electric arc. Hydrocarbons are fed to the electric arc reactor in

gaseous form and are pyrolyzed at high temperature (>2000 K). At these temperatures, acetylene is formed in preference to other hydrocarbon compounds. To suppress further reaction to carbon and hydrogen, the hot crack gas must be cooled rapidly (a few microseconds) to low temperatures.

In the electric arc reactor, which has a power consumption of 8 – 10 MW, ca. 1 t/h acetylene is produced from ca. 2 t/h hydrocarbons. The composition of a typical reaction gas mixture is given below.

C_2H_2	16 vol %
H_2	60 vol %
CH_4	10 vol %
C_2H_4	8 vol %
C_3 compounds	2 vol %
C_4 compounds	1.5 vol %
Higher alkynes	1.5 vol %
Residues	1 vol %
Carbon black	0.4 kg per kilogram C_2H_2

The hot reaction gas with this composition is processed in the following main steps:

1. Cooling
2. Removal of carbon black, higher-boiling hydrocarbons (aromatics), and polyunsaturated compounds
3. Purification of acetylene and byproducts such as hydrogen and ethylene

Hüls has produced acetylene since 1939 by electric arc technology. Until acetylene was superseded by ethylene as starting material for poly(vinyl chloride) manufacture in 1992, the production capacity was 12×10^4 t/a using 18 electric arc groups. After this changeover, the supply of the remaining consumers has been satisfied by using a new method of processing the crack gas that was developed jointly with the company Linde. This method meets the stricter present-day ecological and economic requirements. This was achieved mainly by techniques of production-integrated environmental protection.

Old Production Process. The relevant steps of the old process are shown in the flow diagram in Figure 50. Hot crack gas from the electric arc furnace was quenched to ca. 200 °C with water to enable most of the carbon black to be removed in cyclones. The quench water condensed in a further scrubbing step with water at ca. 50 °C, and the remaining carbon black was removed at the same time. After mechanical removal of carbon black the heated wash water was cooled by direct contact with air in a cooling tower and recycled to the cooler. Excess water was fed to the wastewater treatment plant. Both carbon black fractions recovered were used as fuel.

Figure 50. Flow diagram of the old Hüls acetylene process

The high-boiling, mainly aromatic compounds were washed out with special oil fractions and then recovered by distillation.

The crude gas so obtained was compressed; acetylene was absorbed at increased pressure in water and then desorbed as crude acetylene for further purification by decompression. The wash water was fed into the recycling system, some of it being cooled by direct contact with air. The remaining hydrocarbons in the crude gas were recovered by cryogenic processes involving gas liquefaction.

The polyunsaturated and hence energy-rich compounds recovered in the acetylene purification stage were thermally treated, and the polymerizate so obtained could be used as fuel without problems.

The main reasons for the high emissions and high wastewater contamination were the large, partly open water recycling systems. The emissions consisted of hydrocarbons and HCN formed in the electric arc reactor from the reaction of the nitrogen present in the starting materials.

Of the electrical energy supplied, ca. 50 % was utilized for chemical reactions. The rest (thermal energy) was not used.

New Production Process. The flow diagram of the new plant is given in Figure 51. The main difference between the new and the old processes is that the hot reaction gas from the electric arc furnace is quenched with water only to the temperature level at which it is cool enough for no further reaction of acetylene to carbon black and hydrogen to occur on subsequent cooling in a heat exchanger.

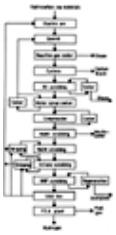

Figure 51. Flow diagram of the new Hüls acetylene process

The thermal energy of the hot reaction gases is used to produce hot steam in a heat exchanger of special design and operational features that prevent the formation of deposits by the hot carbon black present in the gas.

The carbon black is recovered mainly in cyclone separators, and the gas is further cooled by washing with oil, which also removes the high-boiling aromatics. Heavy fuel oil is used as the washing medium, and this, together with the remaining carbon black, is used to produce synthesis gas. The “dry carbon black” from the cyclone separators is used as fuel in the power station.

The reaction gas is cooled to ambient temperature in a water spray column, and the quench water that condenses is fed back in a closed cycle to the electric arc reactor. The gas is then compressed and subjected to a sequence of scrubbing operations with selective solvents [i.e., methanol, octane, and *N*-methyl-2-pyrrolidone (NMP) in that order]. The solvents are regenerated by low-pressure stripping with gas fractions from the production process. This removes mainly the polyunsaturated hydrocarbons.

The acetylene-free gas stream from the NMP wash is first passed through a cryogenic separating stage (cold box) which removes most of the remaining hydrocarbons. These are used first as stripping gas in the regeneration stages of the washing processes described above and are then recycled to the electric arc process. The raw hydrogen stream from the cold box is processed in the pressure swing adaptation (PSA) plant to pure hydrogen and waste gas for heating purposes.

The formation of HCN is almost completely prevented by using virtually nitrogen-free hydrocarbon raw materials. Small residual amounts of HCN are effectively removed by scrubbing with NaOH.

Improvements Resulting from the New Process. The new process is characterized by a high degree of heat recovery and the use of closed cycles. This gives the following improvements:

1. Increase in the thermal efficiency from ca. 50 % to the present figure of 75 %
2. Drastic reduction in the amount of wastewater to be treated in the wastewater treatment plant
3. Prevention of emissions of hydrocarbons and HCN
4. Utilization of the residual carbon black and polyunsaturated hydrocarbons that are recovered in the scrubbing processes

Considerable economic benefits have been obtained:

1. Lowering of energy consumption and wastewater costs
2. Lower personnel requirements
3. Improvement in product purity

2.2.6. *Low-Residue Process for Titanium Dioxide Production (Example from Kronos International)*

General. Titanium dioxide pigment is an important industrial intermediate for multiple utilization in various areas (paints and coatings, printing inks, plastics, man-made fibers, paper, etc.) [144]. The raw materials for pigment production contain titanium dioxide associated with oxides of other metals, mainly iron. The only way to isolate titanium dioxide from this oxide mixture in pigment-grade purity is by chemical processing.

The first commercial process, the sulfate process, uses concentrated sulfuric acid as reagent. After use, the acid cannot readily be recycled because of its content of metals and water taken up during the process. Therefore, this acid (so-called spent acid) was typically discarded as waste. Because of the large acid quantities involved, the (conventional) sulfate process became increasingly obsolete and had to be improved to a modern version [145].

An alternative to the sulfate process is the chloride process for titanium dioxide pigment manufacture, using chlorine as reagent [146], [147]. Since regeneration and recycling of chlorine after use have always been an integral part of this process, it never was impaired by a problem such as the disposal of spent acid from the sulfate process. Thus, the chloride process has become more and more successful and has already gained equal or even greater importance than the sulfate process, although it was developed considerably later.

Conventional (Sulfate) Process (⟶ Pigments, Inorganic – Sulfate Method). Raw material is either an ilmenite ($FeTiO_3$) concentrate from beneficiation of naturally occurring ilmenite or a slag obtained from ilmenite by electrosmelting and tapping off the major part of the iron as pig iron for further use. The titaniferous material is digested with concentrated sulfuric acid, leading to a solution of sulfates of most of the metals contained in the ore, e.g.:

$$FeTiO_3 + 2\,H_2SO_4 \rightarrow TiOSO_4 + FeSO_4 + 2\,H_2O$$

The insoluble solids (unreacted ore, gangue) are separated, and part of the iron(II) sulfate is crystallized and recovered (this part of the process generally being omitted when starting from slag). Subsequently, titanium is selectively precipitated by hydrolysis

$$TiOSO_4 + 2\,H_2O \rightarrow TiO(OH)_2 + H_2SO_4$$

Subsequent filtration yields the spent acid filtrate and a cake of aqueous titanium oxide. Thorough washing of the cake, calcination, grinding, surface treatment in suspension, filtration, drying, and micronizing finally lead to the titanium dioxide pigment product. These process steps result in some slightly acidic effluents.

The spent acid obtained by the filtration step after hydrolysis is diluted sulfuric acid, 18 –23 wt % H_2SO_4, containing dissolved sulfates of iron and other metals from the raw material. In the conventional process, ca. 6 – 9 t of spent acid is generated per tonne of pigment, depending on the raw material used. The spent acid was discharged into rivers, coastal waters, or the high seas [148]. Formerly, the recovered iron(II) sulfate (when using ilmenite as raw material) was redissolved in the spent acid and discarded as well.

Modern Sulfate Process. The sulfate process was thoroughly optimized to reduce the quantity of spent acid and increase its H_2SO_4 concentration. Nevertheless, a substantial amount of spent acid remained; its low H_2SO_4 concentration and its metal sulfate content prevented direct recycling to digestion. Further improvement of the conventional sulfate process by alterations proved impossible, and attempts to commercialize a novel sulfate process with direct recycling of all the spent acid failed. Therefore, further processing of the spent acid was mandatory. In some countries, the spent acid is concentrated somewhat and then used for fertilizer production. In other countries, especially those that have few natural gypsum deposits, the spent acid is reacted with calcium compounds to produce gypsum for use (e.g., in plasterboard manufacture).

Another solution to the spent acid problem is improving its quality by concentration and purification to the extent that it can be completely recycled to the digestion step in pigment production. This alternative is adopted, for example, by all producers in Finland, Germany, and Spain. Although individual plants may differ in some details, the multiple-step reprocessing and recycling of spent acid generally consist of [146], [149-151]:

1. Concentrating the spent acid up to 60 –75 wt % H_2SO_4 where the dissolved metal sulfates have a minimum solubility and are mostly crystallized.
2. Separation of the solid metal sulfates (so-called filter salt) from the acid by filtration. Sometimes, the acid filtrate can be recycled directly to digestion; otherwise, further concentration up to 80 – 90 % H_2SO_4 is required before recycling.

The separated metal sulfates can be processed further by thermal decomposition to generate metal oxides (mainly iron oxide) and sulfur-dioxide-containing gas, which is then processed to virgin sulfuric acid as in conventional pyrite roasting – sulfuric acid plants.

When ilmenite is used as raw material for pigment production, the spent acid contains large amounts of dissolved iron(II) sulfate. In this case, spent acid treatment starts with process steps permitting the recovery of a major part of the iron(II) sulfate before further concentration. This iron(II) sulfate or copper as, $FeSO_4 \cdot 7\,H_2O$, is marketed jointly with the copperas recovered during titanium dioxide production. In Germany, for example, none of the copperas produced is discarded any longer; rather, it is utilized, as such or after further processing, for multiple environmental purposes (water treatment, effluent or sludge treatment in sewage plants).

Chloride Process. The chloride process (see Fig. 52) was developed and commercialized as an alternative to the sulfate process. The processing of titaniferous raw materials (ilmenite, slag, or rutile) occurs via the metal chlorides, by using chlorine as reactant and carbon as a reducing reagent, e.g.:

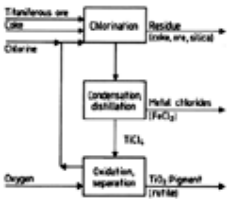

Figure 52. Flow diagram of the chloride process for production of titanium dioxide pigment

$$TiO_2 + FeO + 3\,Cl_2 + 2\,C \rightarrow TiCl_4 + FeCl_2 + CO + CO_2$$

This exothermic reaction takes place at ca. 1000 °C. The reaction products are gases, and after fractional condensation, solids separation, and distillation, sufficiently pure, liquid $TiCl_4$ is produced. The $TiCl_4$ is then reacted with pure oxygen to generate titanium dioxide pigment:

$$TiCl_4 + O_2 \rightarrow TiO_2 + 2\,Cl_2$$

Chlorine is separated from the solid pigment and directly recycled to the first step, chlorination. The pigment may be surface treated in suspension and further processed in the same way as sulfate pigment.

Ecological Aspects of Modern Production Processes. *Sulfate Process.* The conventional sulfate process is characterized by a linear flow of sulfuric acid through the process. Some H_2SO_4 ends up in the copperas byproduct, but the main part remains separate from the pigment end product, as a used reagent with deteriorated quality in terms of concentration and purity. Formerly, this large stream was discarded as waste (see Fig. 53). Now, the spent acid is recovered, and complex acid concentration and filter salt treatment plants are added to recycle the acid entirely. Hence, the modern sulfate process plant for titanium dioxide pigment manufacture is characterized by a closed sulfuric acid cycle that completely withholds spent acid from the environment (see Fig. 54).

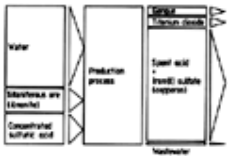

Figure 53. Mass flow in the conventional sulfate process

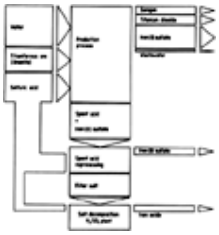

Figure 54. Mass flow in the modern sulfate process

This production-integrated recovery could not be achieved by process redesign; new, separate process units had to be added to the essentially unchanged conventional manufacturing plant. Thus, not only did closing of the acid cycle require substantial investment, but energy requirements increased considerably as well. A substantial amount of fossil fuel is required as such — for thermal decomposition of the filter salt — or indirectly in the form of valuable steam for water evaporation in acid concentration. Parts of these additional energy requirements cannot be recovered but are lost to the environment via cooling towers and off-gases.

Chloride Process. In contrast to the sulfate process, the chloride process has the great benefit that the reagent chlorine is recycled directly without major treatment, this recycling being an integral part and an inherent characteristic of the process right from the beginning. In past years, the chloride process has increasingly been used in construction of new titanium dioxide pigment factories or even for replacing obsolete sulfate process-based plants to cope with the spent acid problem.

Process Engineering Equipment Aspects. Spent acid treatment and concentration as part of the modern sulfate process were possible only because of the availability of improved construction materials and equipment items permitting the safe handling of this highly corrosive substance.

Chloride process technology, plant design, and equipment are entirely different from sulfate process technology. Its commercialization required considerable development and engineering efforts. Safe operation is possible because of high-technology equipment and sophisticated control technology.

2.2.7. *Reduction of Waste Production and Energy Consumption in the Production of Fatty Acid Methyl Esters (Example from Henkel)*

Fats and Oils: The Raw Materials of Oleochemistry. Fats and oils are triglycerides (i.e., fatty acid esters of glycerol). They are the starting materials for the production of fatty acid methyl esters, which are important intermediates in the production of fatty alcohols and surfactants [152] by the oleochemical route, which has great ecological benefits [153]. The fatty acid methyl esters are produced either by the esterification of fatty acids after hydrolysis of the triglycerides or by direct transesterification with methanol. The overall transesterification reaction is as follows:

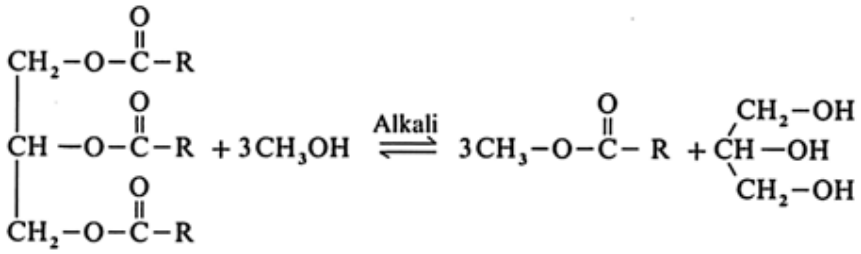

This is actually the result of three consecutive reactions because the fatty acid groups of the triglyceride are rearranged sequentially. Transesterification can be carried out at a relatively low temperature (60 °C) in the presence of an alkaline catalyst (e.g., an alkali-metal alcoxide or hydroxide).

The fat or oil raw material is not a pure triglyceride since significant amounts of other components are also present. Quantitatively, the most significant of these are free (unesterified) fatty acids (1 – 5 %), which neutralize the transesterification catalyst so that the consumption of catalyst can sometimes be very high. The catalyst dissolves in the coupled product glycerol and must be removed as a salt at the end of the process.

Minimization of catalyst consumption is desirable for environmental reasons; therefore, the free fatty acids contained in the oil must be neutralized.

Low-Pressure Transesterification — Old Batch Process. In the old batch transesterification process, neutralized oil is used. The oil is treated with aqueous sodium hydroxide to neutralize the free fatty acids (see Fig. 55). The aqueous phase, which contains a high proportion of the soaps formed (soap stock), is then removed by means of separators, and the remaining soap residues are removed from the oil by washing with water several times (usually twice). These soap residues are then separated [154], [155]; the wash water and the aqueous phase from the first stage, which has a high soap content, are bulked together and treated with sulfuric acid. The soaps are thereby split to form fatty acids and sodium sulfate. The fatty acids can be separated as an organic phase by decantation. The water, which contains sodium sulfate and organic impurities, must be treated in a wastewater treatment plant. The fatty acids produced are a low-value coupled product.

Figure 55. Flow diagram of alkaline refining

The oil, which is free of fatty acids, is then dried and transesterified (Fig. 56). The batch transesterification process represents the usual state of technology [156]. The process is operated in as simple a way as possible.

Figure 56. Low-pressure transesterification — batch process

Sodium hydroxide dissolved in methanol is used as the catalyst for methanolysis. A high proportion of the sodium hydroxide is converted to sodium methoxide in the solution [157], liberating water:

$$NaOH + CH_3-OH \rightleftharpoons NaOCH_3 + H_2O$$

The dried oil is mixed with methanol and the catalyst solution, and the reaction mixture is pumped into a stirred tank. The reaction is completed there at 80 °C as a batch operation. The methanol that evaporates is completely condensed and returned to the reactor.

At the end of the reaction, excess methanol is partly distilled off. The mixture of fatty acid methyl ester and glycerol is pumped into a settling tank to allow phase separation. After standing for several hours, the lower phase (glycerol phase) is separated. The upper ester phase is removed and temporarily stored before further processing (e.g., hydrogenation to fatty alcohols).

Although excess methanol has been distilled out of the reactor, the glycerol phase still contains methanol, which must be removed before further glycerol processing. Removal of residual methanol is carried out continuously. The glycerol is cooled again and treated with acid in a stirred vessel in a batch process. The sodium soaps formed from the catalyst (see reaction equation), which are almost completely extracted from the mixture by the glycerol,

$$NaOH + CH_3\text{-}O\text{-}\overset{\overset{\displaystyle O}{\|}}{C}\text{-}R \rightarrow NaO\text{-}\overset{\overset{\displaystyle O}{\|}}{C}\text{-}R + CH_3OH$$

are decomposed, forming fatty acid and sodium salt:

$$NaO\text{-}\overset{\overset{\displaystyle O}{\|}}{C}\text{-}R + \text{acid} \rightarrow H\text{-}O\text{-}\overset{\overset{\displaystyle O}{\|}}{C}\text{-}R + \text{salt}$$

The fatty acids can be separated after a short time as the upper phase.

The old batch transesterification process has the following disadvantages:

1. Transesterification is carried out in only one step, so that a large excess of methanol is necessary to give a high conversion efficiency.
2. The catalyst used is sodium hydroxide dissolved in methanol. Water is formed in the course of alcoxide formation.

In the presence of water, the sodium hydroxide present reacts very quickly with fatty acid methyl esters to form sodium soaps. Under these conditions, the soaps formed are not effective catalysts.

For this reason, large quantities of catalyst must be used. This leads to product loss and to large quantities of fatty acids in the crude glycerol after soap splitting. Furthermore, the crude glycerol then has a high salt content. This amount of salt must be removed in subsequent processing and disposed of as waste.

New Method of Low-Pressure Transesterification. The modified process has the following aims:

1. Continuous operation
2. Utilization of the free fatty acids contained in the oil
3. Prevention or minimization of waste products and residues
4. Optimum utilization of the raw materials used, especially reduction in the amount of catalyst and the amount of excess methanol required
5. Heat recovery and optimization of energy consumption

Neutralization of Oils by Pre-Esterification. The free fatty acids present in the raw material are esterified with methanol. The proton-catalyzed esterification can be carried out homogeneously [158]. However, the result is that some of the acidic catalyst is also esterified, and the ester and catalyst must then be separated by a water wash. This results in wastewater that sometimes contains toxic substances [159].

To avoid this problem, a heterogeneous catalyst [157], [160-162] is used (see Fig. 57) consisting of a strongly acid macroporous ionexchange resin arranged in a fixed-bed reactor. The reaction temperature is between 60 and 80 °C.

Figure 57. Flow diagram of pre-esterification

In pre-esterification, not only are the fatty acids in the crude oil esterified, but so are the fatty acids produced on splitting the soaps formed in the transesterification process (see Fig. 58). Recycling leads to an increased yield of methyl ester. After pre-esterification, excess methanol is separated and removed together with the reaction water.

Figure 58. Low-pressure transesterification — continuous process

The neutralized oil (mixture of triglycerides and methyl esters) is transesterified after addition of methanol and catalyst. Transesterification leads to the formation of glycerol and fatty acid methyl esters. A marked solubility gap exists in the three-component system glycerol – methyl ester – methanol, so that a glycerol phase and an ester phase are formed [162]. This two-phase system gives the possibility of a multistage process. By removing the product glycerol (GLY, see Eq. 1), a high yield of methyl ester (ME) with a reduced consumption of methanol (MeOH) [163] is achieved

$$K = \frac{[\mathrm{ME}]^3 \times [\mathrm{GLY}]}{[\mathrm{MeOH}]^3 \times [\mathrm{TRI}]} \tag{1}$$

Transesterification is carried out in two stages, the glycerol phase formed being removed after each stage.

The reaction is carried out in tubular reactors with a flow rate that ensures good mixing of the two-phase reaction mixture at the beginning and end of the reaction [164]. Since the amount of back mixing is low in these tubular reactors, short reaction times with low catalyst concentrations can be achieved at ca. 80 – 90 °C.

With a new catalyst (alkali-metal oxide), smaller catalyst concentrations can be used than is the case with sodium hydroxide. With suitable reactor design and choice of catalyst, the salt content of the glycerol produced is decreased by 50 %, and smaller quantities of fatty acids are formed in the splitting reaction.

Immediately after transesterification, excess methanol contained in both liquid phases is removed with the aid of a multistage falling-film evaporator. The multistage system enables heat from the hot methyl ester and glycerol to be utilized in the initial evaporation stages.

The recovered methanol at atmospheric pressure is fed to the methanol distillation stage as a vapor.

The soaps in the glycerol from which the methanol has been removed are continuously split with acid. Consumption of acid is minimized by pH control. Concentrated acid is used to minimize the dilution of glycerol with water.

The ester phase, from which the methanol has been removed, still contains undissolved, highly dispersed residual glycerol, which is removed after cooling with the aid of coalescence separators [165]. The use of such equipment avoids the need to wash the esters with water [152]. A highly concentrated crude glycerol is thus obtained, which leads to reduced energy consumption in glycerol processing.

Ecological Aspects of the New Low-Pressure Transesterification Process with Pre-Esterification. In the earlier process, refined oil was used. The waste streams produced in the oil refining process are completely avoided by use of preesterification. Moreover, the waste fatty acid produced in the low-pressure transesterification process can be reutilized.

The reaction is catalyzed by a macroporous cation exchanger. This catalyst has a very long service life and can be disposed of without problems.

By carrying out the transesterification reaction in several stages, the process can be carried out with a lesser excess of methanol. This leads to an energy saving in the methanol recovery and methanol purification processes. The use of a continuous process enables heat to be recovered during the recovery of methanol from the ester and glycerol.

Impure methanol is fed into the methanol purification stage as a vapor, so that the energy consumption in methanol distillation is very low.

The use of an alkali-metal alcoxide in combination with the new reactor design has the advantage of lower catalyst consumption. Since the catalyst is separated along with the byproduct glycerol, this means that the amounts of salt produced in a subsequent glycerol processing stage are greatly reduced.

During the process, crude glycerol is diluted as little as possible (use of coalescence separators) to give reduced energy consumption in glycerol processing and to minimize the amount of wastewater produced.

2.2.8. *Integrated Environmental Protection in the Production of Vitamins (Example from F. Hoffmann-La Roche)*

Introduction. “Etinol” (chemical formula see below) is an important intermediate in the production of vitamins and food dyes, and is produced in a major plant. By using the conventional synthesis route, the raw material consumption of the presented intermediate step is very high, and the byproducts formed pollute environmental compartments such as water and air. To produce 1 kg etinol in a multipurpose apparatus, ca. 3 kg raw material is required, so that ca. two-thirds of the raw materials are lost as waste, some in chemically changed form. This situation is unsatisfactory both ecologically and economically.

In principle, two possibilities exist for reducing the amount of waste: to recover the useful raw materials from waste streams and recycle these to the process, or preferably to modify the process so that raw materials are more efficiently utilized.

Original Process for the Production of “Etinol” in a Multipurpose Plant. The reaction consists of the addition of acetylene to a vinyl ketone using lithium as the process auxiliary according to a very simple-looking overall equation:

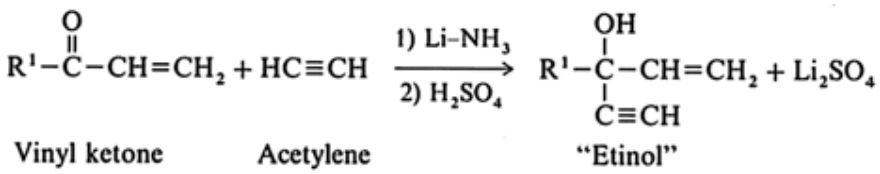

However, the reaction must be performed in three complex stages (see below):

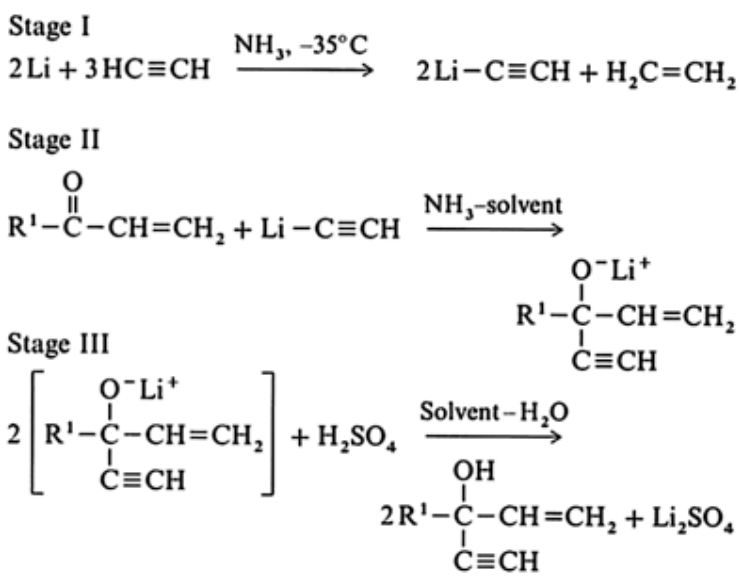

In stage I, metallic lithium is dissolved in liquid ammonia at its boiling point (ca. –33 °C). Acetylene is then fed to this solution, first forming dilithium acetylide (“lithium carbide”). Lithium carbide then reacts further with acetylene to form monolithium acetylide and ethylene. In stage II, the vinyl ketone, diluted with an organic solvent, is added to the lithium acetylide, whereupon “lithium etinolate” is formed. The ammonia that served as a solvent in stages I and II of the reaction is then distilled off from the reaction mixture. In stage III the desired product is obtained from the lithium etinolate by hydrolysis with dilute sulfuric acid.

The former industrial method of carrying out this process is shown schematically in Figure 59. Acetylide formation and reaction of the acetylide with the ketone occurred in that order in the same reaction vessel. Ammonia was then removed in a distillation vessel specially designed for that purpose. The resins produced in stage II of the reaction were removed from the product by a wastewater stripper in the production plant. These resins, along with the wastewater that also contained valuable lithium, were sent to the wastewater treatment plant.

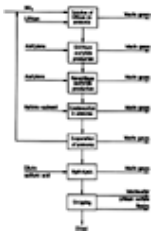

Figure 59. Etinol production — flow diagram of original process

As already mentioned, this process had disadvantages, and great efforts were made to improve it. The production of etinol was described as early as 1948 as the most expensive and difficult part of this vitamin synthesis, and the situation had hardly changed by the early 1980s.

Attempts were made at first to replace the lithium needed in the condensation by cheaper metals such as sodium, potassium, magnesium, or calcium. However, all of these attempts failed because either the yields obtained with the alternative methods were poor or other disadvantages had to be tolerated, such as the very long times required to dissolve the metals. Attempted improvements to the method, such as recovering the expensive lithium from the wastewater in the form of a salt, were also unsuccessful because of the high resin content of the wastewater.

New Process for the Production of “Etinol” (Fig. 60). *Improvements Developed in the Laboratory.* Laboratory investigations showed that the vinyl ketone used is unstable in liquid ammonia. Therefore, under the conditions of ethynylation in the absence of lithium acetylide or when it was present in only substoichiometric amounts, resins were produced almost exclusively from the vinyl ketone even though the solution was diluted with an organic solvent. Therefore, experiments were made in which the addition was carried out in an organic solvent instead of ammonia. The laboratory results were overwhelming, the yields obtained were 95 % instead of 80 %, and the lithium excess could be reduced from 20 to 5 %. Production of the undesired resins almost ceased [166].

Figure 60. Etinol production — flow diagram of new process

After these laboratory investigations, pilot plant trials were carried out on a modified process. After formation of the monolithium acetylide in stage I, ammonia is distilled off and replaced by solvent I. The distillative removal of ammonia must be carried out at a controlled heating rate so that the monolithium acetylide –ammonia complex, which is not very stable, does not decompose. The distillation is interrupted at exactly 5 °C. In order to facilitate drying of the solvent, a further solvent change is carried out after condensation, since solvent II can easily be dried and is also used in the following step of the process [167].

As a result of the much reduced amounts of resin, a lithium recovery method could be developed. It became apparent that an aqueous solution of lithium hydroxide was especially suitable for this recycling method, provided organic impurities were excluded. The “lithium etinolate” solution is hydrolyzed in stage III with water instead of sulfuric acid. The mixture so produced is in two phases, which are separated in a decantation apparatus. The aqueous solution of lithium hydroxide, which still contains part of the product in solution, is concentrated in a column. The water – product mixture obtained as overhead product is recycled to the hydrolysis stage, and the 10 % solution of lithium hydroxide that collects at the bottom of the column is returned to the lithium manufacturer for reprocessing. Only the organic phase is subsequently neutralized with dilute sulfuric acid, and high-boiling components are removed in a stripper column. The consumption of sulfuric acid is considerably lower than in the original process since lithium is obtained not as a salt but as the hydroxide. The amount of sulfuric acid is only that required to neutralize ammonia liberated from the monolithium acetylide – ammonia complex.

Process Changes: Optimization. In the original process, considerable quantities of acetylene and ammonia had to be disposed of as waste products. After the drastic reduction in the amount of resin produced, the consumption of these raw materials could be carefully scrutinized. Improvements can be achieved by:

1. Lowering of acetylene consumption by a modified reaction design
2. Recirculation of acetylene and ammonia by means of an absorption – desorption system
3. Utilization of the calorific value of the unused acetylene and of the ethylene produced in the process

The acetylene balance (percentage of total consumption) in the original process is approximately as follows:

Consumption in the reaction	30 %
Waste gas as ethylene:	20 %
Waste gas as acetylene:	50 %

The important 50 % loss in the form of acetylene can be recovered and returned to the reaction cycle. All of the waste gases from the steps up to the condensation stage are fed to the central waste-gas recovery system. The gases are first passed through the acetylene absorption plant where most of the acetylene is removed from the waste-gas stream by supercooled ammonia with a low acetylene content. Because the other components of the waste gas are less soluble than acetylene in liquid ammonia they pass through the absorption column. The acetylene-rich ammonia from the bottom of the column goes to a central ammonia – acetylene buffer tank. The gases leaving the acetylene absorption column, together with waste gases from the neutralization stage and from all processes of stage III, are passed through the ammonia absorption column, where the waste-gas stream saturated with ammonia is scrubbed with water. The aqueous ammonia solution in the lower part of the column is retained in the absorption recirculation system long enough for the concentration of ammonia to reach ca. 15 %. The ammonia –water mixture is separated by a pressurized rectification column [168]. The remaining waste gases after ammonia absorption — mainly ethylene and nitrogen with small amounts of acetylene — are fed to the boiler plant for burning as fuel.

Acetylene recovered by the absorption process is reused for lithium acetylide production in stage I of the process. For this, the acetylene-containing ammonia from the central ammonia –acetylene buffer tank is fed into a desorption column where the acetylene is driven off from the ammonia by heating and is obtained as overhead product. Ammonia from the bottom of the column is reused for dissolving metallic lithium or absorbing acetylene.

The idea of burning the high-energy waste gases acetylene and ethylene in an environmentally friendly process to produce steam is certainly attractive. However, serious safety problems must be solved before such a process can be considered. Particular attention must be paid to the safety aspects of acetylene, which is explosive over a wide range of concentrations in air (lower flammability limit 2.3 vol %, upper flammability limit 82 vol %, ⟶ Acetylene – General Safety Factors and Safety Measures) and which can decompose at moderately elevated pressure even in the absence of oxygen, generating large amounts of heat and high pressure. Investigations have shown that self-decomposition can be suppressed by diluting the acetylene with natural gas, even when the gas mixture is ignited. Since the boiler plant runs on natural gas, this principle has been used in combustion of the waste gas. Waste gases from the production plant are diluted to 40 % by injection of a stream of natural gas after they have passed through the absorber; they are then compressed to the pressure required by the boiler plant. Moreover, the production plant itself is separated from the combustion plant by a system of flame arrestors [169], [170].

Energy Considerations. Both the process of dissolving metallic lithium and the condensation of lithium acetylide with ketone are exothermic. Since these reactions are carried out at the boiling point of ammonia (–33 °C), or in any case below 0 °C, a cooling system for very low temperatures is inevitable. The cooling medium used is identical to the reaction solvent (in this case, ammonia). Thus, not only can cooling be carried out without using chlorofluorohydrocarbons, but also any accidental mixing of the cooling ammonia with the ammonia of the process will not lead to problems, since it is identical with the solvent used. The pressure difference guarantees that in case of a leak the only possible flow will be of coolant into the reaction solution. Also, the strong smell of ammonia facilitates leak detection, so that amounts of coolant lost by leakage are very small.

The refrigeration process uses compressors and an expander system. The proximity of the refrigeration machinery enables the compressed hot ammonia to be used as a source of heat in the etinol plant. Thus, the acetylene desorption column used to drive off dissolved acetylene is heated by waste heat from the refrigeration machines. Only the excess heat is removed by a cooling tower, and this requires a minimum amount of cooling water.

Ethylene and acetylene are diluted with natural gas in the gas injection equipment and fed to the boiler plant to utilize the energy they contain. The resins can also be used for energy production with the aid of suitable incineration plants. Thus, the high calorific values of these byproducts and wastes can be used to produce energy in the form of steam and electricity, thereby avoiding the environmental pollution that would result from the burning of fossil fuels.

Environmental Protection Balance. The material flow diagrams for the old process and for the new improved process are shown in Figure 61.

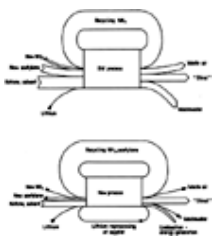

Figure 61. Comparison of mass flow of old and new processes

The presence of lithium salt in the wastewater is avoided. Also, the waste-air streams conform with legal requirements and are all used either in the process or for energy production. Increased recycling of ammonia and acetylene leads to reduction in the consumption of ammonia to 25 % of its former value; acetylene consumption is reduced by 50 %. Thus, the stoichiometric excess of acetylene used in stage I of the reaction is reduced to ca. 12 %.

The increased extent of closed material cycles is remarkable. Not only ammonia and acetylene, but also lithium in the form of lithium hydroxide solution, are recycled. The efforts toward a better environmental compatibility have resulted in a new process that is more complex than the original “straight-through” process. The effects are beneficial for both the ecology and the economy. The performance of new installations is optimized by ensuring that individual process steps harmonize with each other. This applies not only within the identical batch plants used for the different stages, but also to energy utilization in the boiler plant. Operators of these plants are required to have an increased level of understanding. Improved training and education can ensure that consumption of materials is minimized and utilization of equipment and raw materials is maximized.

2.2.9. *Production of Pure Naphthalene without Residues — Replacement of Chemical Purification by Optimized Multiple Crystallization (Example from VFT)*

Formation and Uses of Naphthalene. In the production of coke for use in blast furnaces, coal tar is produced as a liquid byproduct. Raw materials for the chemical industry, metallurgy, and energy production are obtained from this complex mixture of chemical compounds with aromatic character (⟶ Tar and Pitch) [171].

The most important constituent of coal tar with respect to quantity is naphthalene, which is present at a concentration of 10 %. Pure naphthalene (up to 99.9 %), obtained by a refining process, is of great economic importance as a basic chemical for the production of, e.g., phthalic anhydride-based plasticizers, dyes, plant-protection agents, and special solvents (⟶ Naphthalene and Hydronaphthalenes – Physical and Chemical Properties, ⟶ Naphthalene and Hydronaphthalenes – Production, ⟶ Naphthalene and Hydronaphthalenes – Uses) [172].

Chemical Refining. The conventional method of purifying naphthalene uses several stages of physical and chemical treatment:

1. Coal tar distillation
2. Crystallization of the naphthalene fraction
3. Treatment of the crystallized product with concentrated sulfuric acid and formaldehyde, followed by neutralization with sodium hydroxide solution
4. Redistillation to remove dark-colored secondary constituents (Fig. 62)

Figure 62. Flow diagram of old process of chemical naphthalene refining by treatment with sulfuric acid and sodium hydroxide solution

This refining process also produces highviscosity naphthalene-containing acidic resins amounting to ca. 8 % of crude naphthalene. These must be disposed of by dumping.

Also, based on the original weight of coal tar, naphthalene-containing waste sulfuric acid (7 %) and naphthalene-containing dilute sodium hydroxide solution (ca. 18 %) are produced. These waste streams can be neutralized by mixing them together and are then sent to a sewage plant. However, because the material has large amounts of naphthalene-containing oils, sludge deposits are formed that require very complex treatment.

Physical Refining Process (Solution of the Environmental Problem). VFT realized the advantage of preventing the production of waste materials by removing the impurities in naphthalene simply by repeated application of a physical separation process. High selectivity is required to enable the residual oil to be used as a raw material in later process stages without further treatment (Fig. 63).

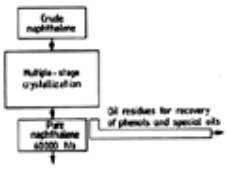

Figure 63. Flow diagram of new process of physical refining by countercurrent crystallization

To achieve this desirable situation, a new process had to be developed. The phase diagram of naphthalene and the main impurity benzo[*b*]thiophene is given in Figure 64. It shows that with a starting mixture of 90 % naphthalene, a single-stage crystallization gives only a 6 % increase in concentration with a very low yield. On the other hand, a multistage crystallization process can give naphthalene in good yields and high purities that meet market requirements (Fig. 65).

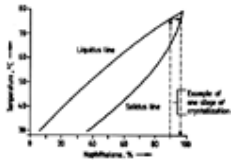

Figure 64. Phase diagram of benzo[*b*]thiophene – naphthalene

Figure 65. Schematic of five-stage crystallization

Process Details and Plant Technology. As an illustrative example, a detailed description of stage 2 of the process is given. The crude naphthalene, together with the residual oil from stage 3 and the crystalline product (C 1) from stage 1, is cooled in the crystallizer at controlled temperature. This causes a layer of crystals with an increased naphthalene concentration (in accordance with the phase equilibrium) to form on the cooled surface of the inner wall of the crystallizer.

The depleted residual melt is recycled to stage 1, and is there separated into a residual oil and the crystallized product (C 1). Benzo-[*b*]thiophene and other impurities in the crude naphthalene remain in the residual oil, which can be processed in adjoining production plants.

Naphthalene shows a marked tendency to undergo heterogeneous seed formation on cooling of the melt [173]. A crystallizer was therefore developed in which the molten naphthalene fraction runs down the interior of vertical tubes as a falling film [174]. During the crystallization phase, the depleted naphthalene oil that runs off the inside of the tube is collected in a vessel and recycled by a pump to the falling-film distributor at the top of the crystallizer. The outside surfaces of the crystallization tubes are covered with a falling film of water, which acts as a cooling or heating medium. The temperature of the water is lowered (crystallization phase) or raised (melting phase) in accordance with a predetermined function.

At the end of the crystallization phase, the liquid residual oil is pumped off. In the next stage of the process, the falling film of water is heated under temperature-controlled conditions so as to give partial melting of the crystal layer on the inside of the tube. This causes melting of that fraction of material that contains a higher proportion of impurities (in accordance with the phase diagram). These fractions are collected separately in the vessel under the crystallization tubes and are distributed to various crystallization stages according to their purity, as shown in Figure 65.

A crystallizer containing >1000 tubes, all 12 m in length, was designed and built to give an economically viable production capacity. Careful attention was given to the design of the crystallizer, the equipment for providing heat energy at the correct rate, and precise process control.

This multistage crystallization process is controlled by a central control system. In this way, the energy and mass flow rates can be continuously balanced in the various crystallization stages. Determination of the process parameters and control of the production program are based on the analytically determined starting concentrations at each stage and the desired separation factors.

Since crystallization is a purely physical separation process, neither solid waste nor wastewater is produced. The comparatively small flows of waste air due only to changes in the contents of the plant are taken to the central waste-gas combustion plant. All impurities in the crude naphthalene are concentrated in the residual oil from stage 1 and can be economically utilized in further processing stages. This naphthalene purification process was installed in the Castrop-Rauxel works of VFT, and has produced >60 000 t/a of pure naphthalene consistently and without problems for more than 10 years.

Benefits of the Plant. The shortage of dumping space has led to a drastic increase in dumping costs, so that production of pure naphthalene by chemical refining is no longer economically possible. Development of this residue-free environmentally friendly process has made a decisive contribution to the stabilization of coal tar utilization in Germany.

2.2.10. *Improvements in the Polypropylene Production Process (Example from Shell)*

Introduction (⟶ Polyolefins – Raw Materials, ⟶ Polyolefins – Industrial Processes). The discovery by ZIEGLER and NATTA of the catalyst system named after them in the 1950s opened the way to the large scale production of polypropylene. Shell companies were among the first to realize the potential of this process and started a development program that culminated in the start-up of Shell's first two polypropylene units in Carrington, United Kingdom (1962) and Pernis, The Netherlands (1963). These plants were very small (capacity 5×10^3 t/a) compared to present-day standards and used first-generation Ziegler – Natta catalysts. These catalysts had a low activity and a low selectivity, making it necessary to remove the catalyst residues as well as the atactic polymer by de-ashing. This process made use of a hydrocarbon diluent, which resulted in relatively low system pressures.

In 1964, Shell Research discovered a catalyst with improved selectivity, which led to a significant rationalization of the production process [175]. The removal of atactic polymer could be omitted because polymer of the desired isotacticity could be made directly in the reactor. The above developments formed the basis for a continuous improvement in plant design, with production capacities steadily rising to 100×10^3 t/a in the late 1970s.

Major Developments since 1980. During the 1980s, two fundamental changes in the polypropylene process led to a

breakthrough in terms of improved process economics, energy efficiency, and reduced environmental impact. The first of these changes was the conversion from a diluent process to a process in which liquid propene acts as both reactant and “diluent” for the reactor slurry. This mode of operation, which is called LIPP (Liquid Propylene Process) leads to a considerable increase in the pressure at which such a process is operated.

The second breakthrough was the use of a high-activity catalyst. This proprietary catalyst called “Shell High Activity Catalyst” (SHAC) was developed by Shell Research and achieved a hundredfold improvement in product yield with respect to the quantity of titanium catalyst employed, leading to a titanium residue in the polymer on the order of 1 to 2 µg/g [176]. As a result, removal of the catalyst residues could be omitted, leading to further process simplification. An impression of the simplification in terms of unit operations between the “diluent” process and new “LIPP” process can be gained from Figures 66 and 67.

Figure 66. Flow diagram of the diluent – slurry process

Figure 67. Flow diagram of the bulk polypropylene process (LIPP)

Capital Cost. Clearly, from the above, the investment cost per tonne of capacity for the LIPP – SHAC process will be appreciably lower because fewer equipment items are required in the streamlined process.

Energy Consumption. Without the need for de-ashing and diluent distillation, and with a more efficient operation due to the higher pressure level in the reactor, major reductions (ca. 50 % expressed in terms of fuel equivalent per tonne of polymer) in the energy consumption of the modern polypropylene plant have been achieved. Figure 68 shows the energy consumption per tonne of polypropylene on changing from a diluent to a LIPP process.

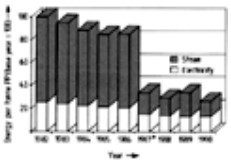

Figure 68. Energy consumption of polypropylene (PP) plants

* Conversion to LIPP process, 1987.

Environmental Impact. With regard to environmental impact, four areas can be identified in which clear progress is evident:

1. *Metal salts* The metal salts that originated from the de-ashing step required disposal. Since de-ashing is no longer needed, their quantity has been diminished by a factor of >100. The only source of metal salts remaining is equipment flushing where small amounts are found.
2. *Diluent* Although extensive precautions are taken to avoid loss of diluent from the plant, this can never be totally prevented. Apart from the fugitive losses of diluent, a loss occurs in the heavy ends fraction of the diluent recovery operation. Since the LIPP process does not use a diluent except for some flushing, the amount of diluent lost is drastically reduced.
3. *Alcohol* In the diluent process, an alcohol is used to deactivate the catalysts. Although the alcohol is recovered from the aqueous stream, a small residual amount will be contained in the wastewater. Since no alcohol is required in the new process, this discharge no longer exists.
4. *Propene* The overall streamlining of the process has resulted in a considerable reduction in fugitive emission of propene from the plant.

The relative improvements in the above-mentioned fields are summarized in Figure 69.

Figure 69. Environmental discharges from polypropylene production processes

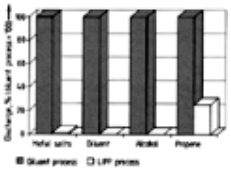

Developments Developments are mainly in the field of catalyst improvement, both of conventional Ziegler – Natta catalysts and of single-site catalyst. These developments will mainly influence the properties of the polymer product rather than the manufacturing process.

2.2.11. ***The Zero-Residue Refinery Using the Shell Gasification Process (Shell – Lurgi Example) (see also, ⟶ Oil Refining)***

Introduction. Refineries worldwide are being subjected to increasing pressures, both legislatively and economically. The effects of the Clean Air Act (CAA) in the United States of America are well documented [177]. In Europe, increasing environmental legislation is making itself felt both in refinery operation and in terms of product quality. The former is governed by the requirements of the EC Large Combustion Plant Directive or more stringent national regulations. The latter is dictated by EC Directive EN 228 requiring a reduction of sulfur in diesel motor fuels to 0.05 % S by 1996 and in heating oils to 0.1 % S by 1999.

Simultaneously, a medium-term pressure is occurring to reduce the low-sulfur crude intake of refineries because of declining availability and an increasing cost differential compared with high-sulfur crudes.

Various schemes including either coking or residue gasification have been proposed to cope with these pressures. Realization that coke disposal is becoming increasingly difficult has led to wider application of residue gasification (⟶ Gas Production – Partial Oxidation of Hydrocarbons, ⟶ Gas Production – Hydrogenating Gasification). Four existing refineries in Germany already employ the Shell Gasification Process (SGP) for residue utilization. Worldwide, five plants are currently (mid-1990s) under construction, including a major revamp at Shell's own Pernis refinery near Rotterdam.

SGP in the Refinery Flow Sheet. Before describing the gasification and gas treating technology, the way they fit into the overall picture of the refinery flow sheet should be reviewed. Figure 70 provides a simplified refinery flow sheet after implementation of a combined hydrocracker (HCU) and SGP project. These two units, together with the gas-treating section downstream of the SGP, are shaded in the diagram.

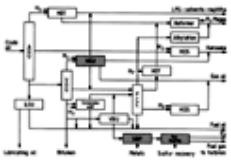

Figure 70. Shell gasification process in the refinery environment

ADU = Atmospheric distillation unit; LOU = Lube oil unit; HDT = Hydrotreating; VDU = Vacuum distillation unit; HCU = Hydrocracker unit; VBU = Visbreaker unit; FCC = Fluid catalytic cracking; HDS = Hydrodesulfurization; SGP = Shell gasification process; Mogas = motor gasoline; LPG = Liquefied petroleum gas

The main environmental contaminants, sulfur and heavy metals, tend to remain in the bottom product of distillation processes. Their path through the refinery is from the crude oil intake via the atmospheric distillation unit (ADU), through the vacuum distillation (VDU) to the visbreaker unit (VBU), a mild thermal cracking process with subsequent distillation facility. In the original flow scheme, prior to implementing the new units all the residue from the visbreaker together with its load of contaminants, typically 2 – 4 % sulfur and up to 400 µg/g vanadium, was sold into the heavy fuel oil pool. This material is largely used in oil-fired power stations, where the restrictions of the Large Combustion Plant Directive are now placing restraints on its use, particularly to reduce atmospheric sulfur emissions.

Reductions in the allowable sulfur content of diesel and heating oil both require the use of more hydrogen in the hydrodesulfurization (HDS) and hydrotreating (HDT) units in the gas oil fractions. The new hydrocracker, whose main purpose is to convert middle distillates into lighter fractions, is also a large consumer of hydrogen.

Reduction in allowable benzene levels in motor gasoline (Mogas), which the CAA requires and which may yet become a requirement in Western Europe, requires a reduction of operating severity on the reformer, which reduces the amount of byproduct hydrogen available from this unit. Environmental pressures as expressed in legislation to reduce sulfur emissions to the environment from refinery products, especially when combined with a reduction of potentially carcinogenic material from gasoline, have the following results:

1. A surplus of high-sulfur fuel oil
2. A large increase in the internal hydrogen demand of the refinery
3. A substantial decrease in the amount of internally generated hydrogen from the reformer

The overall technological concept provides for utilizing the surplus high-sulfur fuel oil to manufacture the hydrogen required to cover the ever-increasing internal hydrogen deficit. It in cludes the construction of a new hydrocracker and the residue gasification unit for the production of hydrogen and sulfur-free fuel gas, which will be fed to a gas turbine that forms the heart of a combined-cycle cogeneration plant. Additionally, it includes new amine treating and Claus units.

The feedstock for the SGP unit is a heavy, vacuum-flashed cracked residue from an existing visbreaker unit. This material is typically characterized by high viscosity and high metal content.

The product hydrogen would be aimed primarily at supplying the hydrocracker. Sufficient flexibility can, however, be built into the gas treatment facilities to accommodate alternative scenarios in the overall hydrogen refinery balance and to produce methanol for methyl-*tert*-butyl ether (MTBE) production or other uses.

Process Description. *General.* When looking at the market possibilities, the sales volume of the product should be matched with the total volume of residue available. Of those products of immediate use in the refinery that can be produced from residues, hydrogen is an obvious choice (⟶ Gas Production – Hydrogen Production Based on Heavy Residues). Advanced processing schemes consume large quantities of hydrogen. Whether the hydrogen demand will match the quantities of residue available will of course depend on overall refinery configuration. Methanol, either as a direct fuel additive or as a feedstock for MTBE production, can account for utilization of some additional residues. Ammonia prices are too low to make it attractive, and there is no marketing synergy as might be the case with methanol. Speciality chemicals do not generate sufficient volumes to make any impact on the residue problem, although a sidestream unit on a base plant might offer possibilities for a high margin investment. The market for electrical power is acquiring increasing importance as another possibility for absorbing the quantities of residue available [178].

The block flow diagram (Fig. 71) illustrates the plant arrangement for the simultaneous production of all three main products indicated above — power, hydrogen, and methanol — together with the possibility of manufacturing pure carbon monoxide as well. Although, in practice, a plant would rarely be required for all four products, two product schemes are by no means uncommon.

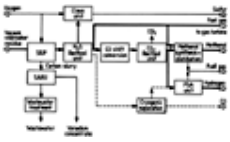

Figure 71. Block flow diagram of an SGP-based methanol and hydrogen plant

SGP = Shell gasification process; SARU = Soot ash removal unit; PSA = pressure swing adsorption

Gas Production. The feedstock is gasified in the SGP reactor (a) with oxygen to produce raw synthesis gas at about 1300 °C and 60 bar (Fig. 72). This gas is a mixture of hydrogen and carbon monoxide, which also contains CO_2, H_2S, and COS together with some free carbon and ash from the feedstock.

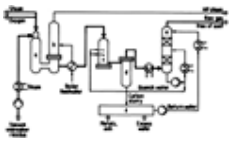

Figure 72. Flow diagram of an SGP unit based on residual oil

a) Reactor; b) Waste-heat exchanger; c) Quench pipe; d) Carbon separator; e) Scrubber; f) Soot ash removal

Recovery of the sensible heat in this gas is an integral feature of the SGP process. Primary heat recovery takes place in a waste-heat exchanger (b) generating high-pressure saturated steam in which the reactor effluent is cooled to about 340 °C. The waste-heat exchanger is of a special design developed specifically for these operating conditions. Part of the steam thus generated is used for feedstock and oxidant preheating; the remainder is superheated for use in the CO shift conversion and in steam turbine drives. Steam conditions are chosen to be optimized with the steam turbine part of the combined cycle.

Particulates are removed from the gas by means of a two-stage water wash. Carbon formed in the partial oxidation reactor is removed from the system as a carbon slurry together with the ash and the process condensate. This slurry is subsequently processed in the ash removal unit (f) described below. The product synthesis gas leaves the scrubber at a temperature of about 40 °C and is essentially free of carbon. It is then suitable for treatment with any commercial desulfurization solvent.

Gas Treatment (⟶ Gas Production – Gas Treating). Selection of the actual process to be used for desulfurization depends on the application and on site-specific conditions. The very high purity required for methanol synthesis gas (S<0.1 ppm) practically mandates the Rectisol process, which uses cold methanol as a solvent. For power applications, the high H_2S – CO_2 selectivity of the Purisol solvent *N*-methyl-2-pyrrolidone (NMP) makes it a good choice. Where low investment costs have priority over reduced operating costs methyldiethanolamine (MDEA) may be used. In all cases, the desulfurized gas is suitable for firing in a gas turbine as part of an integrated gasification combined cycle (IGCC) power plant (⟶ Gas Production – Combined Cycle Power System Based on Coal).

The remaining gas is then subjected to CO shift conversion using steam to convert the CO to CO_2 and hydrogen. Hydrogen purification to >99 % can be effected by pressure swing adsorption (PSA) as shown in Figure 71, either with or without prior CO_2 removal. An alternative scheme of CO_2 removal followed by methanation produces a 98 % pure H_2 product with <10 ppm carbon oxides.

A side stream of gas is bypassed around the shift unit to achieve the correct H_2 : CO ratio for the methanol synthesis process. Methanol synthesis itself operates at about 80 bar using the Lurgi Low Pressure Methanol Process followed by a distillation step to achieve the required product quality specification.

The provision for carbon monoxide production is shown by the dotted line in Figure 71. CO_2 is removed to a greater extent than for hydrogen or methanol production before the gas enters the cold box of the cryogenic unit. This produces a CO product that meets the specification for acetic acid production and a >97 % hydrogen stream at about 22 bar. Depending on the overall hydrogen scheme, this may be recompressed and fed through the PSA for final cleanup or used raw in the

refinery.

Side-Stream Units. Wash water from particulate removal is treated in the soot ash removal unit described in more detail below. Carbon and ash are separated from the water by filtration. The filter cake is worked up by controlled oxidation to a salable vanadium concentrate. The bulk of the water is recycled to the water wash.

When using a physical wash such as Rectisol or Purisol, the regeneration system of the desulfurization solvent can be designed to operate selectively so that it produces a gas of sufficiently high H_2S content to allow it to be processed to elemental sulfur in the sulfur recovery unit. In many instances, the Lurgi Oxygen Claus technology (⟶ Sulfur) can be used at this location, either to de-bottleneck existing units or for new plants.

Process Refinements. Handling the carbon slurry is an important aspect of partial oxidation processes. The traditional approach to handling the carbon slurry was to contact it with a hydrocarbon to retain the carbon, which could thus be recycled to the reactor. The early Shell units used fuel oil as the extraction medium (⟶ Gas Production). This allowed atmospheric operation, providing a plant with low investment costs and, at least initially, low operating costs as well. The economics of this approach deteriorated with heavier feedstocks. The pelletizing oil had to be purchased at a premium to the main feedstock, and with higher ash contents, 100 % carbon recycle became nearly impossible since much of the ash was recycled as well.

The logical development of the process was to use naphtha as extraction medium. Soot is extracted from water with naphtha in the form of pellets. These naphtha – soot pellets are sieved off from the water and then mixed in with the main feedstock at whatever temperature is required to achieve the desired viscosity. The naphtha is then distilled off from the feed and recycled to the extraction stage leaving the soot behind in the feed. The use of naphtha as an intermediate allowed the use of heavier, more viscous feedstocks than in the case of pelletizing with fuel oil. Also, improvement in the separation of carbon and ash allowed 100 % recycle. Nonetheless, an ash buildup factor of about 3 : 1 can be observed under 100 % carbon recycle conditions. The plant is costlier than fuel oil pelletizing in both investment and operation.

These facts have led to the review and finally the development of an alternative approach and in particular the possibilities of recovering vanadium for metallurgical use. Figure 73 illustrates the principles of the soot ash removal unit. The carbon slurry from the SGP unit is flashed to atmospheric pressure in the slurry tank (a). The slurry is then filtered on an automatic filter (b) to recover a filter cake with about 80 % residual moisture and a clear water filtrate. The filter cake is subjected to a controlled oxidation process in a multiple-hearth furnace (c). This type of furnace, which is well established in many industries and specifically in the vanadium industry, allows combustion of the carbon to occur under conditions where the vanadium oxides neither melt nor corrode. This is not an easy task if one considers the problems of burning a high-vanadium fuel oil in a conventional boiler. The product is a vanadium concentrate, which contains about 75 % V_2O_5. Compared to the old naphtha extraction-based recycle system, the new once-through process consists of only two processing steps, which are not integrated with the gasification section of the unit. It is cheaper both in investment and in operating costs.

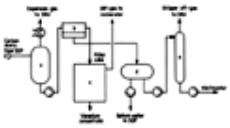

Figure 73. Flow diagram of a soot ash removal unit

a) Slurry tank; b) Filter; c) Multiple-hearth furnace; d) Return water vessel; e) Wastewater stripperSGP = Shell gasification process; SRU = Soot removal unit

An additional benefit of this development is increased feedstock flexibility, both for the SGP itself and for the refinery as a whole. The carbon recycle currently practised also recycles some of the ash, so that the charge pump, burner, and reactor system are exposed to a higher content of ash than that present in the fresh feed. Elimination of the recycle thus makes it possible to use feeds with considerably higher ash contents than previously. Current experience of >1000 µg/g vanadium at the reactor inlet would be directly applicable to residues of this quality. This is certainly an important feature when reviewing the possibility of introducing heavier (and cheaper) crudes into the refinery.

Environmental Aspects. This article has concentrated on the SGP as part of an overall refinery modernization project. Evaluation of the environmental impact can, however, be performed only in the context of such a plant. The data in Tables 3 and 4 are taken from [179] which describes the project under construction in the Shell refinery in The Netherlands.

Table 3. Refinery mass balance, 10^6 t/a

Mass balance	Before 1993	After 1997
Feedstock		
High-sulfur crude	11.3	15.0
Low-sulfur crude	4.5	2.8
Other feedstocks	2.1	0.1
Total intake	17.9	17.9
Production		
White product make	12.8	14.0
Fuel oil make	3.5	2.1
Sulfur	0.11	0.30

Total output	16.4	16.4

Table 4. Refinery and global emissions, 10^3 t/a

Emissions	Before 1993	After 1997
Refinery emissions		
SO_2	35	24
NO_x	12	7
Particulates	4.0	1.9
CO_2	5 000	6 100
Global emissions (refinery + products)		
SO_2	348	206
CO_2	56 600	56 800

The refinery mass balance of Table 3 shows a decisive shift to high-sulfur crude in the future increasing the total intake of high-sulfur crudes from 63 to 84 % of total feed. Nonetheless, the refinery SO_2 emissions and the sulfur content of products are both reduced by one-third, as can be seen in Table 4. Similar reductions can be seen for NO_x and particulate emissons. CO_2 emissions from the refinery increase, but because of the lighter product slate, global CO_2 emissions are hardly affected. All this is achieved with an increase in the high-value white product from 78 to 85 % of total output and a reduction of 40 % in fuel oil output.

The technology described has been applied in a number of refineries as a means of manufacturing chemicals from low-value feedstocks. With increasing environmental awareness and continued refinement of the technology, it is now finding a strategic place in the planning of low- or even zero-residue refineries [180-182].

2.2.12. *Neutral Salt Splitting with the Use of Hydrogen Depolarized Anodes (HydrinaTechnology, Example from De Nora Permelec)*

General. Worldwide, more than 3×10^6 t of sodium sulfate is obtained per year as a byproduct of many industrial processes: only a portion of this production can be commercialized after crystallization and calcination, e.g., for the detergent and glass industries, which, however, have been progressively reducing their consumption in the last ten years. The remaining portion of sodium sulfate is usually diluted and led into the bulk of effluents or disposed of.

However, disposal will become increasingly difficult, if not impossible, in the near future due to tighter environmental regulations. In any case, it will certainly become increasingly expensive and for this reason, at least, the recovery by electrolysis of sodium sulfate — and in general of other neutral salts, such as sodium nitrate and sodium chloride, which are also generated in some chemical processes — is expected to have a bright future.

The overall economics of the recovery of byproduct salts by electrolysis benefits from the saving of the disposal cost as well as from the production of caustic and acid, which may be recycled to upstream plants. The success of this operation depends on a number of conditions including

1. Substantial savings of money, which otherwise would have been required for disposal of byproduct salts. In the specific case of sodium sulfate, electrolysis becomes economically interesting when the disposal cost of the salt is in the order of $ 100 per tonne, with a caustic soda price in the range of $ 200 per tonne.
2. Electric energy consumption per tonne of produced caustic soda, which is close to that of membrane chlor – alkali electrolysis (i.e., 2200 – 2500 kW · h/t).
3. Simple and reliable integration with the plants where the byproduct salts are generated. This integration may require auxiliary sections designed to eliminate impurities, to concentrate byproduct salt solutions when these are excessively dilute, (e.g., <10 wt %), and, in some cases, to concentrate the acid produced.
4. Negligible requirements of manpower for both operation and maintenance.

Structure of the Electrolyzers (see also ⟶ Electrochemistry). The following discussion is based on sodium sulfate, which is the byproduct salt of greatest concern [183], [184]. However, many of the conclusions also apply to the electrolysis of other salts such as potassium sulfate and sodium and potassium nitrate.

Industrial electrolyzers are made of rows of electrically connected elementary cells, preferably of the bipolar type. Many elementary cell structures have been described in the literature. The most important from the industrial point of view are:

1. Elementary cells comprising two compartments separated by a cation-exchange membrane (MC), as illustrated in Figure 74, to produce caustic soda and a mixture of sulfuric acid and unconverted sodium sulfate, hydrogen, and oxygen.
2. Elementary cells comprising three compartments separated by two membranes, namely a cation-exchange (MC) and an anionexchange (MA) membrane (see Fig. 75). In this case the electrolysis products are caustic soda,

sulfuric acid, hydrogen, and oxygen.

3. Elementary cells comprising three compartments separated by two cation-exchange membranes as illustrated in Figure 76. In this case, the products are again caustic soda, a mixture of sulfuric acid and unconverted sodium sulfate, hydrogen, and oxygen, as in the case of the two-compartment process. The additional anodic compartment acts as a buffer section so that the solution to be electrolyzed is separated from the anode. This scheme is particularly useful to prevent:
   - Anodic side reactions, as in the presence of certain oxidizable anions, such as chlorate [185]
   - Corrosion of the anodes due to the high concentration of sulfuric acid and unconverted sodium sulfate at elevated temperature
   - Decay of the membranes caused by large contents of impurities in the feed solutions [186]

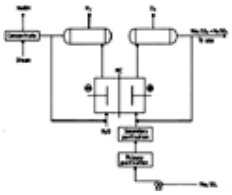

Figure 74. Flow diagram of sodium sulfate electrolysis in two-compartment electrolyzers (production of acid sodium sulfate, oxygen-evolving anodes)

MC = Cation-exchange membrane

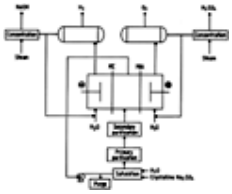

Figure 75. Flow diagram of sodium sulfate electrolysis in three-compartment electrolyzers (production of pure sulfuric acid, oxygen-evolving anodes)

MC = Cation-exchange membrane; MA = Anion-exchange membrane

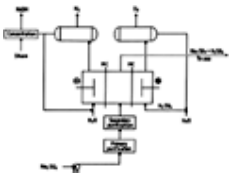

Figure 76. Flow diagram of sodium sulfate electrolysis in three-compartment electrolyzers (buffer compartment)

MC = Cation-exchange membrane

A problem affecting all three schemes is the high cell voltage, which involves high energy consumption. The high cell voltage results from the energy required to perform the following reaction at an industrially useful rate:

$$Na_2SO_4 + 3\,H_2O \rightarrow 2\,NaOH + H_2SO_4 + H_2 + 0.5\,O_2$$

Energy consumption may be substantially reduced by replacing the oxygen-evolving anode with a hydrogen depolarized anode [186], [187]. The modified schemes are shown in Figures 77 and 78.

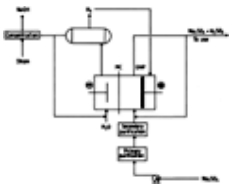

Figure 77. Production of acid sodium sulfate in the Hydrina membrane electrolyzer

MC = Cation-exchange membrane; DAP = Depolarized anode package

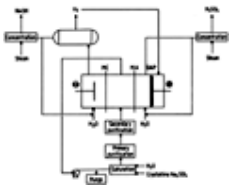

Figure 78. Production of pure sulfuric acid in the Hydrina membrane electrolyzer

MC = Cation-exchange membrane; MA = Anion-exchange membrane; DAP = Depolarized anode package

The energy consumption is decreased remarkably according to the following new overall reaction:

$$Na_2SO_4 + 2\,H_2O \rightarrow 2\,NaOH + H_2SO_4$$

Figures 77 and 78 demonstrate that the hydrogen consumed at the depolarized anode is the same hydrogen evolved at the corresponding cathode. This hydrogen need only be washed in a suitable scrubber to eliminate entrained catholyte mists. The hydrogen thus obtained has a high purity, which allows for long-term stability of the catalyst of the depolarized anode.

A conventional structure used for the hydrogen depolarized anode consists of an electrically conductive sheet with controlled porosity and hydrophobicity, coated with a catalytic layer. Under optimum operating conditions the electrolyte partially fills the pores, with the gas –liquid meniscus in the area of the sheet containing the catalyst.

Operating conditions causing percolation of the electrolyte through the pores or loss of gas in the electrolyte should be avoided. For this reason, anodes of this type allow rather small pressure differentials between the electrolytic compartment and the hydrogen chamber. The height of these anodes is thus limited. Thus, cell designs with segmented compartments have been proposed [188]. Besides this limitation, conventional electrodes are affected by problems caused by the presence of electrolyte in the pores in direct contact with catalyst. Therefore, both crystallization phenomena are experienced, which result in destruction of the porous sheet structure and poisoning of the catalyst when the electrolyte contains heavy metals, even in trace amounts.

These problems are overcome by using a depolarized anode made of a porous sheet containing a catalyst, bonded or pressed against a cation-exchange membrane [189-191]. The assembly formed by the membrane and the porous catalytic sheet is in turn bonded to or pressed against a suitable porous current collector, such as an expanded metal or wire mesh. The dimensions of the voids in the expanded metal or wire mesh are preferably small in order to allow simultaneously good mechanical support of the membrane – porous sheet assembly and homogeneous current distribution.

The current collector is made of a material that resists the corrosive action of sulfuric acid, which may contact the collector in the event of a discontinuity in the membrane, such as a pinhole or crack, formed during fabrication, handling, or operation. Suitable materials capable of operating at 60 – 70 °C are nickel alloys. With optimized assemblies (DAP, De Nora Permelec, Milan, Italy), overall anodic overvoltages of 0.1 – 0.2 V at 3000 A/m^2 are obtained.

In the bipolar arrangement of elementary cells the following overall reaction occurs in the separation wall between the depolarized anode and the cathode:

$$H_2O \rightarrow H^+ + OH^-$$

which is the same reaction occurring in bipolar membranes for electrodialysis [192]. In particular, the overall reaction results from an anodic partial reaction:

$$H_2 - 2e^- \rightarrow 2H^+$$

which takes place in the acid environment of the ionic groups of the cation-exchange membrane and a cathodic partial reaction:

$$2H_2O + 2e^- \rightarrow H_2 + 2OH^-$$

which occurs in the alkaline environment of the catholyte made of caustic soda.

When the cathode is provided with an electrocatalytic coating, the overall voltage across the bipolar separation wall between two adjacent cells has the remarkably low value of 1.2 – 1.3 V at 3000 A/m^2, thus equal to or even lower than that of bipolar membranes at 1000 A/m^2 [193].

The outstanding performance at high current densities largely counterbalances the intrinsically higher cost of the structure comprising the hydrogen depolarized anode, bipolar wall, and activated cathode as compared with bipolar membranes. Further advantages of the assembly are the long service life due to the mechanical strength, a low sensitivity to impurities contained in the feed electrolyte, and high quality of the products.

Operating Conditions. Voltage vs. current relationships for three types of cells provided with DAP, electrocatalytic cathodes, electrolytic compartments having 3-mm gaps, and turbulence promoters are given below:

Voltage (V) – current density (CD) relationships for three internal structures of the elementary cell for sodium sulfate electrolysis:

1. Two-compartment, cation-exchange membrane V = 0.96 + 0.59×CD (kA/m^2)
2. Three-compartment, cation- and anionexchange membranes V = 0.97 + 1.02×CD (kA/m^2)
3. Three-compartment, two cation-exchange membranes (buffer compartment) V = 0.98 + 0.89×CD (kA/m^2)

The three relationships apply to current densities up to 3000 A/m^2. The DAP assembly comprises a hydrogen depolarized anode and a Nafion 117 membrane; the electrolytic compartments are limited by Nafion 324 and Selemion AAV ionexchange membranes. The relationships further apply to operating temperatures of 60 – 70 °C and to 13 – 18 % caustic soda, 200 – 300 g/L sodium sulfate, and 10 % sulfuric acid.

The current efficiency for the production of caustic soda (“cathodic efficiency”) is a complex function of the following factors:

1. *Caustic soda concentration*
   Other conditions being the same, the cathodic current efficiency decreases as the caustic soda concentration increases to more than 15 – 18 %.
2. *Molar ratio (MR) between sulfuric acid and sodium sulfate at constant sodium sulfate concentration*
   The second dissociation constant of sulfuric acid is rather low, in the range of 0.01. Consequently, in a solution

containing both sulfuric acid and sodium sulfate at MR <1, substantially all the sulfuric acid reacts with the stoichiometric amount of sodium sulfate to give sodium bisulfate (buffer action). Hence, the actual concentration of free protons (H^+) is directly proportional to the actual concentration of sodium bisulfate and inversely proportional to that of the unreacted sodium sulfate. This type of dependence indicates that the actual concentration of free protons should increase quickly when MR exceeds a certain critical value (ca. 0.5). At higher MR values the current transported by the protons becomes significant at the expense of that transported by the sodium ions, and the cathodic efficiency shows a sharp decrease.

3. *Sodium sulfate concentration at constant MR* At MR values <1, the proton concentration, which is a function of the ratio MR to (1 – MR), is expected to remain constant at constant molar ratios, even if the concentrations of both sulfuric acid and sodium sulfate are greatly changed. Consequently, the transport number of the protons decreases as the concentration of sodium sulfate (i.e., the concentration of sodium ions) is increased. At the same time, the cathodic current efficiency increases.

This behavior has also been discussed in terms of alkaline state (to the left of the critical point of MR) and acid state (to the right of the critical point) of the cation-exchange membrane [187], [194]. This difference in internal states of the membrane is connected to the presence in the conductive channels of the membrane of a population of either hydroxyl ions (back migration from the catholyte when the electric current is substantially transported by sodium ions; low MR values) or protons (significant portion of the current transported by protons; high MR values). The model has remarkable practical consequences, particularly in terms of the sensitivity of the membrane with respect to impurities contained in the feed solutions.

The above considerations also apply to other salts that derive from weak acids, such as phosphates, acetates, and in general the sodium salts of many organic acids. The behavior of salts deriving from strong acids, such as nitrates, chlorates, and perchlorates, is substantially different because these salts are not able to perform any buffer action. Thus, at a given cathodic current efficiency, the allowed concentration of acid is significantly lower than that typical of sodium sulfate solutions. The concentration of acid can, however, be increased by the addition of a “dead load” of sodium sulfate, which creates the necessary buffer action [195].

With two-compartment cells the cathodic efficiency also represents the total current efficiency of the process. In the three-compartment cells, the current efficiency for sulfuric acid production (anodic efficiency) must also be taken into consideration. The overall current efficiency of electrolysis is represented by the lower of the two efficiencies. If, for example, the anodic efficiency is lower than the cathodic, acidity builds up in the sodium sulfate compartment during operation so that the cathodic current efficiency progressively decreases; when it reaches the same value as the anodic, stationary process conditions are maintained. Therefore, the process is provided with a self-regulating mechanism. Alternatively, if the accumulation of acidity in the sodium sulfate compartment cannot be allowed, then a portion of the caustic soda produced can be fed into the sodium sulfate loop. The quantity of caustic soda available for use outside the electrolysis plant again represents the overall current efficiency, which still remains a function of the anodic efficiency. The anodic and cathodic current efficiencies can be made independent when the control of acidity in the sodium sulfate compartment is performed by means of sodium carbonate addition.

Auxiliary Sections. The electrolysis plant may contain auxiliary units necessary for functional integration with the upstream plants where sodium sulfate is generated and the electrolysis products are fed back.

Treatment of Sodium Sulfate Feed. Impurities normally present in the sodium sulfate feed are Ca, Mg, Al, and Si. These impurities penetrate into the cation-exchange membranes separating the sodium sulfate and the caustic soda compartments; they react with the internal alkaline environment and precipitate, causing a voltage increase and a loss of cathodic efficiency.

Internal precipitation is less likely when the caustic soda contained in the cathodic compartment is diluted. With caustic soda concentrations in the range of 10 – 15 %, the total concentration of impurities may reach a maximum value in the order of 1 – 5 µg/g, depending on the type of membrane [183]. If the membrane is operated in the acid state [187], [194], this limit of concentration may be increased to 20 – 30 µg/g [187], [196]. Obviously, this is a compromise, because the advantage of operating under safe conditions for the membrane (even in the presence of relatively high impurity concentrations) is counterbalanced by the lower cathodic current efficiency typical of the acid state of the cationexchange membranes.

These general considerations must, in any case, take into account both the kind of electrolysis and the structure of the cells (two or three electrolyte compartments). In the first case, the sulfate solution is strongly acid and high levels of impurities are allowable. Besides, the electrolysis cell is inserted in an open cycle where the accumulation of impurities is negligible.

In the second case, the cell is inserted in a substantially closed loop. This situation leads to the accumulation of impurities. In this case, a chemical purification unit must be introduced, which, in its simplest form, may be quite similar to the primary treatment of chlor – alkali plants. The treatment is based on additions of caustic soda and sodium carbonate followed by filtration, and it allows the concentration of impurities, such as calcium and magnesium, to be kept in the range of 1 – 5 µg/g. The additional investment required for chemical treatment is rather limited as are the operational costs.

Chemical treatment may be replaced by crystallization or ion-exchange units. However, their investment and operating costs are substantial and negatively influence the economics of the entire process. In addition, crystallization and ion exchange are delicate operations that may require specialized personnel.

Concentration of Sodium Sulfate Solutions. A concentration unit is always necessary for both two- and three-compartment cells when the feed consists of dilute sodium sulfate solutions, because the cell voltage increases rapidly when the sodium sulfate concentration is <10 %. The methods available are reverse osmosis, electrodialysis, and evaporation. With the

present average European prices of steam and electric energy, evaporation seems to be the best method for concentrating up to 20 % diluted feed solutions having a salt content around 3 – 5 %.

Even if the feed solutions are concentrated (e.g., to ca. 20 %), evaporation is still required when electrolysis is carried out in three-compartment cells. In this case, evaporation is directed to keep constant the amount of water in the electrolysis section. Water is let in with the feed solution and only partially migrates through the cationic and anionic membranes to form caustic soda and sulfuric acid.

Obviously, evaporation is not required when the makeup consists of solid salts. The evaporation unit is expensive with respect to both investment and operating costs, and its installation should be the object of careful economic analysis.

As a purely indicative example, an electrolysis plant based on three-compartment cells is considered, having a capacity of 3000 t/a of sodium sulfate (100 % basis) fed to the cells as a 20 % solution: The investment for a doubleeffect evaporation unit represents about 10 – 15 % of the total investment for the complete plant, comprising cells, rectifier, pretreatment unit, evaporation, and construction. The cost of steam amounts to about 50 % of the total electric energy cost.

Clearly, the energy consumption, which is already substantial for the 20 % feed solution, increases sharply as the feed solution becomes more dilute.

Before beginning an economic analysis to define the most convenient trade-off between evaporation and electrolysis cost, the upstream plants where sodium sulfate is generated should be inspected carefully to determine whether the dilute solutions are formed as such or are the result of mixing with other process wastewaters. Internal dilution is often a consequence of the old attitude toward byproduct salt solutions, which were diluted as much as possible within the battery limits of the plants to reach the limit allowed for discharge.

Sulfuric Acid Concentration. In three-compartment electrolysis, 10 to 15 wt % of pure sulfuric acid is produced with an anodic efficiency in the range of 60 – 75 %. Often these solutions are too dilute to be fed directly to upstream plants. This requires concentration by evaporation, which, however, may be limited to a maximum concentration of 75 wt %. A unit of this type with an annual capacity of ca. 2000 t of acid (100 % basis) involves an investment that amounts to about 15 – 20 % of the total investment required for the complete plant. The steam consumption for a double-effect unit, at current European prices for electric energy and steam, corresponds to about 30 % of the electrical energy cost of electrolysis.

Conclusions. The Hydrina technology developed by De Nora Permelec for low-cost splitting of neutral salts has the following advantages:

1. Easy integration of the electrolysis process into the production cycle, thereby forming a “closed loop”
2. Transformation of byproducts into valuable chemicals such as caustic soda and acid
3. Elimination of any environmental problems deriving from the disposal of byproduct salts

2.2.13. *Ultrapure Isopropanol Purification and Recycling System (Example from Mitsubishi Chemical)*

Background of Isopropanol Recycling System. In the electronics industry, ultrafine pattern processing technology for greater-capacity devices has been developed. This technology requires complete exclusion of impurities from the manufacturing lines, i.e., a highly sophisticated clean room technique, and highly purified water and solvents among other measures.

One of the solvents used in ultrafine pattern processing technology is ultrapure isopropanol (IPA), which is widely used in vapor dryers for drying the electronic devices and liquid crystal display substrates that have been washed with ultrapure water [197-199]. At present, the large amount of used IPA — containing impurities such as water, metal ions, and particles — is discarded after use. When using IPA, the following serious problems arise:

1. Increase in production costs through the use of large quantities of ultrapure IPA that are not recycled
2. Latent danger of using a large amount of flammable liquid
3. Environmental problems because of the large quantities of waste IPA that are disposed of

To solve these problems, the new IPA purification and recycling system has been developed, which can reproduce ultrapure IPA that meets the specifications of the electronics industry from the waste IPA. This system is composed of a high-performance dehydration unit based on pervaporation membrane separation; a metal-ion elimination unit based on distillation; and a particle elimination unit based on microfiltration. These units remove each impuritiy efficiently from waste IPA, and purified IPA can be recycled to manufacturing processes of the electronics industry.

Pervaporation Membrane Separation. Pervaporation (PV) is a membrane process to separate the components of liquid mixtures (⟶ Membranes and Membrane Separation Processes – Principle of Pervaporation); a liquid mixture is fed to one side of a membrane, permeating as vapor to the opposite side to which a vacuum or sweeping gas is typically applied (see Fig. 79). This nonequilibrium process is regarded as a suitable alternative process for the separation of organic liquid mixtures such as azeotropic mixtures, mixtures having close boiling points, mixtures of structural isomers, and of heat-sensitive organic compounds. Pervaporation has been commercialized for the separation of water from concentrated alcohol solution, primarily ethanol (⟶ Ethanol) and IPA. Other applications, such as separations of small amounts of organic solvents from contaminated waters and separations of purely organic mixtures, are under development.

The most widely accepted model for describing the transport mechanism through the pervaporation membrane is a modified solution – diffusion model, where a permeating component is first dissolved or adsorbed in the membrane and then transported through it by a diffusion process. In this process, separation of the mixture is achieved by differences in the solubility and diffusibility of the individual components.

Performances of PV membranes are represented by parameters such as separation factor, flux of permeates, and service life. The separation factor of a membrane is a measure of its permeation selectivity (permselectivity) and is defined as the ratio of the concentration of components in the permeate mixture to that in the feed mixture. The component flux is the amount of a component permeating per unit time and unit area, and is given by the product of the permeability coefficient of the membrane and the driving force. The driving force is the gradient in the chemical potential of the components between the feed and the permeate side of the membrane. These values are influenced by operating variables such as temperature, composition of each component in the feed mixture, and permeate side pressures (see Fig. 80).

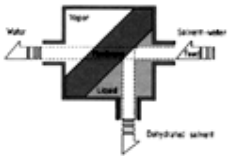

Figure 79. Flow sheet of the pervaporation process

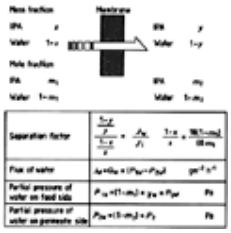

Figure 80. Mass-transfer model for pervaporation of an IPA – water mixture

J_W = Flux of water; J_i = Flux of IPA; Q_W = Permeation coefficient of water; γ_W = Activity coefficient of water; P_{0W} = Saturated vapor pressure of water at operating temperature; P_{1W} = Partial pressure of water on feed side; P_{2W} = Partial pressure of water on permeate side; P_2 = Total pressure on permeate side

Separation of IPA and Water by Pervaporation. In the PV membrane separation unit, the water permselective membrane is used for dehydration. Water in the feed mixture can permeate through the membrane; the IPA, in contrast, effectively cannot. Consequently, the residue IPA on the feed side is dehydrated.

The water concentration of the feed side decreases as dehydration progresses and lowers the chemical potential of water on the feed side. This means that a greater degree of dehydration requires a higher operating temperature to increase the driving force. The low heat resistance of conventional membranes has up to now prevented PV separation processes at high temperature [200], [201]. In contrast, the novel heat-resistant membrane used in the IPA – water system enables the operation to be performed at higher temperature, achieving not only the high degree of dehydration but also the large water flux and increasing total system performance (see Fig. 81).

The permeate water vapor is continuously condensed in a cooler, stored in a tank, and discharged. Due to the superior permselectivity of the membrane, the amount of IPA in the permeate mixture becomes very small; therefore, the IPA load of the wastewater decreases significantly.

The membrane is in the form of a hollow fiber (see Fig. 82), which has the advantage of reduced outer dimensions together with a large membrane area. The membrane module consists only of the hollow fiber bundle and the module housing. Such a simple structure can avoid difficulties encountered with other membrane module designs, such as sealing of flat seat type and spiral-wound type membranes and furthermore can reduce not only the volume and the weight of the modules, but also the total system size.

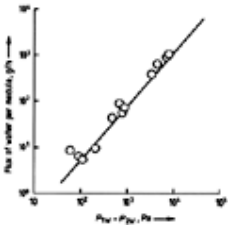

Figure 81. Relationship between difference in partial pressure of water and flux of water

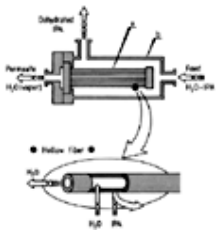

Figure 82. Structure of the pervaporation module

a) Fiber bundle; b) Module housing

Distillation Unit and Filtration Unit. In this system, purification and recycling of IPA is performed by the combination of the PV membrane dehydration unit, the distillation unit, and the microfiltration unit. In the distillation unit, impurities, which are difficult to separate in the PV membrane separation unit such as dissolved metal ions and high-boiling impurities, are completely eliminated.

In the last process step, fine particles are removed by the microfiltration unit. In the manufacture of highly integrated electronic devices, particles from the solvents used in these processes must be removed to improve product yields and suppress wafer contamination defects. For example, particles with >0.05-µm diameter should be removed to the extent of less than 10 particles per milliliter from solvents used in 16-Mbit level production lines [202]. Accordingly, the level of the microfiltration unit affects total system performance; therefore, the unit should be equipped with an appropriate filtration membrane, although only a few membranes with sufficient performance are available [203]. In the solvent a very low level of dissolved metals and low total organic carbon (TOC) is desired. Moreover, high chemical resistance of the filtration membrane is also needed.

The system can vary the required capacity by a change in its composition. For example, if the 90 % IPA makeup solution could be purified up to 99.9 % IPA concentration with a feed rate of 12 kg/h, then more than 10 kg/h of product would be obtained. The number of particles with diameter of >0.2 µm is less than 10 particles per milliliter product. The metal ion content of the product is ≤1 ppb (i.e., 1 µL/m^3) for the following elements:

Al	Ca	Cu	Mg	Pb
As	Cd	Fe	Mn	Sn
B	Co	K	Na	Sr
Ba	Cr	Li	Ni	Zn

These results were achieved by using S 111 and M 111 type purification/recycling systems; the dimensions (mm) of the systems are 600, 2300 width, 1700 height and 800 depth, 2300 width, 1460 height respectively, and the masses of both systems are 500 kg. A schematic of the entire unit is shown in Figure 83.

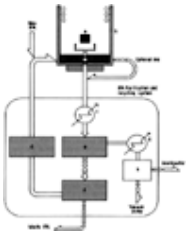

Figure 83. Typical application of the IPA purification and recycling system

a) Electronic device to be dried; b) Vapor dryer; c) Heater; d) Microfiltration unit; e) Pervaporation unit; f) Evaporation unit; g) Condenser; h) Tank

2.2.14. *Examples from Boehringer Mannheim*

2.2.14.1. *Biocatalytic Splitting of Penicillin [204]*

Introduction. The technique of biocatalysis has been used for thousands of years in the production of various food products including bread (“sourdough”); the fermentation of sugar from fruit and fruit pulp to produce alcoholic beverages; the processing of milk to produce cheese, yoghurt, and other products; the production of vinegar and sauerkraut; and many other products.

These processes utilize the activity of biological catalysts (enzymes), which are formed by bacteria, yeasts, and molds. Each individual enzyme performs a different biocatalytic function.

A particularly successful and technically elegant biocatalytic process is the conversion of natural penicillin into a product that is necessary for the production of chemically modified penicillins. This product, known as 6-aminopenicillanic acid, is formed by splitting off the side chain of penicillin.

6-Aminopenicillanic acid is treated to obtain the so-called semisynthetic penicillins, of which the most well known are ampicillin and amoxycillin (see below).

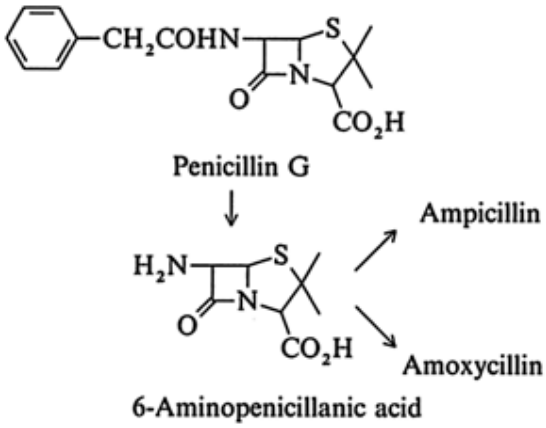

The enormous importance of penicillins in combating infections is shown by the fact that >4000 t/a of 6-aminopenicillanic acid alone is produced.

Today, by use of genetic engineering, the active enzymes can be obtained relatively simply and selectively in pure form.

By immobilizing these in a porous matrix, they can be converted into a form with very convenient properties for technical handling purposes (⟶ Immobilized Biocatalysts – Methods of Enzyme Immobilization).

In the biocatalytic process for splitting penicillin described here, genetically engineered penicillin amidase is used, which is bonded to polymer particles the size of sand grains.

This process has almost completely superseded an older chemical process, over which it has a number of advantages including a better environmental balance:

1. Lower production costs
2. Energy savings
3. Fewer safety problems
4. Fewer disposal problems
5. Improved product quality

Comparison of the Environmental Balance of the Chemical and Bioctalytic Processes. The reason for the *reduction in production costs* is immediately apparent from the reaction sequence shown in Figure 84. Whereas the chemical process requires three intermediate stages, the biocatalytic process occurs in a single synthesis step in water at room temperature with ammonia addition for neutralization.

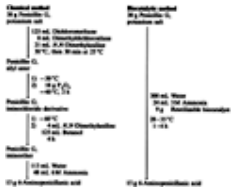

Figure 84. Reaction scheme for chemical and biocatalytic splitting of penicillin

The advantages of the biocatalytic process are impressive if the conversion of 1000 t penicillin is considered; this yields ca. 500 t product or slightly less than one-eighth of annual world production (see below).

Chemical method	**Biocatalytic method**
Consumption of reagents	
1000 t Penicillin G, potassium salt	1000 t Penicillin G, potassium salt
800 t *N,N*-Dimethylaniline	
600 t Phosphorus pentachloride	
300 t Dimethyldichlorosilane	
160 t Ammonia	45 t Ammonia
	0.5 – 1 t Biocatalyst
Consumption of solvents (recoverable, partly for disposal)	
4200 m^3 Dichloromethane	10 000 m^3 Water
4200 m^3 Butanol	
Product isolation and purifiction	
Acetone	Acetone
Ammonium bicarbonate	Butyl acetate
	Hydrochloric acid
Energy consumption (estimated from cooling costs)	
6×10^6 kW · h	3×10^6 kW · h

The potential *energy savings* can be estimated from the much reduced cooling costs. Whereas the chemical process requires a temperature of –50 °C, the biocatalytic process is carried out at 28 – 35 °C. Thus, expensive cooling is replaced by comparatively cheap heating.

The absence of highly reactive chemicals in the biocatalytic process reduces the complexity of *safety measures*, which would otherwise be considerable. Phosphorus pentachloride especially poses a problem, because it must be stored and handled with complete exclusion of moisture since it is strongly hygroscopic and decomposes with the liberation of hydrogen chloride gas.

In the biocatalytic process, waste gas purification is not necessary, and no *disposal problems* involving environmentally harmful chlorine-containing solvents or byproducts from the synthesis process occur.

In the bioctalytic process, the *product quality requirements* are considerably easier to comply with. In the old process, *N,N*-dimethylaniline had to be removed from the reaction product, because this toxic substance may be present only below its toxicity threshold (i.e., in extremely small traces).

The absence of significant quantities of environmentally problematic substances and the large contribution to energy savings reflect the enormous potential of biocatalysis for solving and preventing environmental problems.

2.2.14.2. *Production of Diagnostic Reagents by Means of Genetic Engineering: Glucose-6-Phosphate Dehydrogenase and α-Glucosidase [204]*

Introduction. Diagnostic reagents are important tools for medical investigations (⟶). For some of these (e.g., the testing of blood and other bodily fluids), chemical reactions are no longer used; biochemical methods based on enzymes are preferred.

Glucose-6-phosphate dehydrogenase (G6PDH) (E.C. 1.1.1.49) is used in diagnostics to determine the glucose content of serum (⟶), and *α-glucosidase* is used in a test combination for the determination of amylase content, an important parameter in determining the functional capability of the human pancreas. High-purity enzymes in kilogram quantities are required for the enormous number of laboratory investigations conducted.

Conventional biochemical processes require up to 1000 t of biological raw materials to produce 1 kg of an enzyme suitable for bioanalysis.

The efficiency of enzyme production by biotechnological methods can be much improved by genetic engineering (⟶ Genetic Engineering). Reprogramming (cloning) the genetic information of microorganisms that can be grown by simple means (bacteria or yeasts) enables them to produce certain proteins (e.g., enzymes) in excess. By means of recombinant DNA technology, up to 1000 copies per cell can encode the desired protein. In this way, up to 50 % of the total protein in the cell can consist of the desired product. By culturing the genetically engineered microorganisms under special conditions in a medium containing special nutrients, the overproduction of the cell can be utilized technologically. At the same time, large savings in energy and raw materials are made.

Reduction of Environmental Pollution by Production of Glucose-6-phosphate Dehydrogenase and α-Glucosidase by Genetic Engineering. *Production of Glucose-6-Phosphate Dehydrogenase.* The bacterial strain *Leuconostoc* sp., used in the production of G6PDH by the conventional biochemical method, has a volume activity of the desired enzyme of 5.5 U/mL. By cloning the appropriate gene in *Escherichia coli*, an increase in the genetic enzyme yield (genetic expression) to 500 U/mL is obtained even on the 5-mL scale. Optimization of the fermentation process enables a further increase by a factor of 1000 to be achieved, i.e., a genetic expression of 4500 –5000. The savings in raw materials and energy during fermentation alone resulting from using *E. coli* instead of *Leuconostoc* are listed in Table 5.

Table 5. Prevention of environmental pollution by use of recombinant organisms in the producton of G6PDH during fermentation

Quantity	*Leuconostoc*	Recombinant *E. coli*
Fermentation volume, m^3	600	1
Salts and nutrients, kg	64 000	160
Pure water consumption, m^3	120	1
Cooling water consumption, m^3	15 000	30
Amount of O_2-free water required, m^3	740	
Amount of ice water required, m^3	4 800	
Energy requirements for steam production, kW · h	16 000	270
Compressed air, m^3	114 000	570
Wastewater PE* value	300 000	300

* PE = population equivalent (i.e., organic load due to a member of the population over 24 h).

A comparison of the conventional culture process with that using recombinant *E. coli* shows that only 0.1 % of the raw materials and energy of the conventional process is required for the recombinant process. This is true for both the fermentation process and the downstreaming procedure (Table 6).

Table 6. Prevention of environmental pollution by use of recombinant organisms in the production of G6PDH during workup

Quantity	*Leuconostoc*	Recombinant *E. coli*
Biomass, kg	22 000	200
Wastewater, m^3	1 200	0.2
Water, m^3	300	10
Ice water, m^3	4 000	50
Pure water, m^3	300	10

Steam, t	180	10
Electricity, kW · h	4 000	100
Ammonium sulfate, kg	13 000	200

To provide a given quantity of this enzyme, only 200 kg biomass is required when using recombinant *E. coli* instead of 22 000 kg. Savings of a similar order of magnitude are obtained in the areas of energy consumption, steam, water, pure water, and of course, electricity.

The recombinant process gives a saving of ca. 95 – 98 % in raw materials and energy. This leads to a reduction in the load on the sewage treatment plant which is illustrated by so-called population equivalents. A population equivalent (PE) is defined as the organic load due to a member of the population over 24 h. Treatment of this organic load requires 60 g of oxygen. In the production of G6PDH by recombinant organisms, the organic load produced by the fermentation process is equivalent to 0.1 % of that produced by the conventional process. This means that the amount of organic waste to be treated, instead of being equivalent to that from a large town (300 000 PE), is only equivalent to that from several large families (300 PE).

Production of α-Glucosidase. α -Glucosidase can also be produced very effectively using recombinant yeast instead of normal yeast strains (Table 7).

Table 7. Prevention of environmental pollution by use of recombinant organisms in the production of α-glucosidase (1988 consumption figures)

Quantity	Production of α-glucosidase	
	From new yeast	From recombinant yeast
Yeast consumption, t	236	10
Wastewater from yeast production plant, t (ca. 9 t wastewater/t yeast)	ca. 2 000	ca. 90
Quantity of yeast residues for disposal, t	440	12
Ammonium sulfate, t	1 100	25
Potassium phosphate, t	25	0.5
Aluminum gel, t	90	
Filter aids, t	133	5
Solid waste, t	540	18
Liquid waste, t (ammonium sulfate + potassium phosphate as fertilizers)	1 125	25.5
Deionized water, m^3	3 700	50
Ice water, m^3	52 000	2 000
Electric power, kW	44 500	9 000
Steam, t	220	50

The treatment of residues from processes based on genetic engineering (wastewater, decomposed biomass) must be carried out in accordance with the safety requirements of genetic engineering law and genetic engineering safety regulations. The genetically modified microorganisms (*E. coli* and yeast) used in the processes described here are classified in the lowest safety group S 1. The wastewater (fermentation liquors) can therefore be transferred directly to a biological sewage treatment plant. Solid residues (decomposed biomass) can be composted without further treatment.

When using recombinant yeast strains, the amount of biomass to be disposed of is much smaller than with normal yeast strains because the fermentation volume is smaller (i.e., 1.2 t instead of 440 t). The water and energy consumption in the α-glucosidase production process involving genetic engineering is ca. 10 – 20 % of that of the conventional process (Table 7).

About 100 diagnostic products are produced by Boehringer Mannheim in this environmentally friendly way using genetic engineering techniques.

[Top of Page]

3. Acknowledgement

The entire topic was coordinated by Jürgen Wiesner.

[Top of Page]

References

1. Verband der Chemischen Industrie e.V. (eds.): *Umwelt-Leitlinien*, Frankfurt/Main, Feb. 1988.
2. C. Lange: *Umweltschutz und Unternehmensplanung –Die betriebliche Anpassung an den Einsatz umweltpolitischer Instrumente (neue betriebswirtschaftliche Forschung)* vol. **8**, Verlag Dr. Th. Gabler, Wiesbaden 1978.
3. Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit: *Umweltbericht 1990*, Bundesanzeiger no. **145 a**, Köln 1990.
4. R. Vieregge: "Integrierter Umweltschutz aus Sicht der Umweltpolitik" in H. Kreikebaum (ed.): *Integrierter Umweltschutz – Herausforderung an das Innovationsmanagement*, Verlag Dr. Th. Gabler, Wiesbaden 1990, p. 87.
5. K. Töpfer, *Trend Z. soziale Marktwirtschaft* **39** (1989) 44.
6. H. Strebel: "Integrierter Umweltschutz – Merkmale, Voraussetzungen, Chancen," in H. Kreikebaum (ed.): *Integrierter Umweltschutz – Herausforderung an das Innovationsmanagement*, Verlag Dr. Th. Gabler, Wiesbaden 1990, p. 3.
7. R. Antes: "Umweltschutzinnovationen als Chancen des aktiven Umweltschutzes für Unternehmen im sozialen Wandel," *Schriftenreihe des Institutes für ökologische Wirtschaftsforschung*, no. 16, Berlin 1988.
8. Statistisches Bundesamt: Comments on the ecological-economic studies, personal communication, July 27, 1989.
9. Kommission der Europäischen Gemeinschaften (eds.): *Panorama der EG-Industrie*, 1990, p. 149.
10. "The Perspective on the International Chemical Industry,": *Chemical Insight* **534** (May 1994) 1.
11. Verband der Chemischen Industrie e.V. (eds.): *Umweltbericht 1988/89*, Frankfurt/Main 1989, p. 29.
12. J. Wiesner: *Produktionsintegrierter Umweltschutz in der Chemischen Industrie*, SATW, Zürich, VDI-GVC, Düsseldorf, DECHEMA, Frankfurt/Main 1990.
13. K. Holoubek, J. Geywitz: "Die Integration des Umweltschutzes in das operative Betriebsgeschehen," in U. Steger (ed.): "Umwelt-Auditing – Ein neues Instrument der Risikovorsorge," *Frankfurter Allgemeine Zeitung*, Verlag Bereich Wirtschaftsbücher, Frankfurt/Main 1991, p. 97.
14. H. Wunderlich: "Umweltschutz und Sicherheit in der Produktion," in: *Die Bayer-Umweltperspektive*, Bayer, Leverkusen 1987, p. 44.
15. M. Faber, G. Stephan, P. Michaelis: *Umdenken in der Abfallwirtschaft – Vermeiden, Verwerten, Beseitigen*, 2nd ed., Springer Verlag, Berlin 1989.
16. C. Christ: "Produktionsintegrierter Umweltschutz am Beispiel der Chemischen Industrie," *VDI Ber.* **899** (1991) 79.
17. G. Scharfe, B. Seweko: "Produktionsintegrierter Umweltschutz – Das Beispiel Bayer," *Chem. Ind.*, Sonderausgabe Nordrhein-Westfalen 1991, p. 17.
18. K.-G. Malle: "From 'End of the Pipe' Philosophy to an Integrated Approach. Examples for the Evaluation of Environmental Protection in a Chemical Company," *Kem. Kemi* **20** (1993) no. 1, 23.
19. D. Becher: "Vermeiden, Vermindern, Verwerten –integrierter Umweltschutz in der Produktion," in: *Die Bayer-Umweltperspektive II*, Bayer, Leverkusen 1991, p. 34.
20. K. Hansmann, *Neue Zeitschrift für Verwaltungsrecht* **9** (1990) 409.
21. W. Hilger: "Chemie im Dialog: Herausforderung und Perspektiven," VCI-Dokumentation der Veranstaltung am 4. Juli 1991 im Wissenschaftszentrum Bonn, Frankfurt/Main, Oct. 1991.
22. H. Meffert, M. Kirchgeorg: *Marktorientiertes Umweltschutzmanagement*, J. B. Metzler Verlag and C. E. Poeschel Verlag, Stuttgart 1992.
23. J. Senn: *Ökologie-orientierte Unternehmensführung*, Verlag Peter Lang, Frankfurt/Main 1986.
24. E.-H. Rohe: "Umweltschutz heute in einem Chemiekonzern — Aufgaben, Probleme, Fortschritte," in: *Die Bayer-Umweltperspektive*, Bayer, Leverkusen 1987, p. 30
25. U. Steger: *Umweltmanagement – Erfahrungen und Instrumente einer umweltorientierten Unternehmensstrategie*, Verlag Dr. Th. Gabler, Wiesbaden 1988.
26. U. Steger: "Integrierter Umweltschutz als Gegenstand eines Umweltmanagements," in H. Kreikebaum (ed.): *Integrierter Umweltschutz – Herausforderung an das Innovationsmanagement*, Verlag Dr. Th. Gabler, Wiesbaden 1990, p. 33.
27. R. F. Nolte: "Neue Verfahren der Produktions- und Umwelttechnik: BMFT-Förderprogramm, Märkte und Erfolgsfaktoren, 1988," in Der Rat von Sachverständigen für Umweltfragen (eds.): *Abfallwirtschaft-Sondergutachten 1990*, J. B. Metzler Verlag and C. E. Poeschel Verlag, Stuttgart 1991, p. 207, Tz 711.
28. U. Ullrich: "Ökologische Produktionskonzepte – Skizzen für ein naturverträgliches Leben, Arbeiten und Wirtschaften," in E. Schmidt (ed.): "Ökologische Produktionskonzepte – Kriterien, Instrumente, Akteure," *Schriftenreihe des Instituts für ökologische Wirtschaftsforschung*, no. 23, Berlin 1989.
29. C. Christ: "Umweltschutzmanagement in der Chemischen Industrie am Beispiel der Hoechst AG," in D. Adam (ed.): *Schriften zur Unternehmensführung Nr. 48 "Umweltschutzmanagement in der Produktion,"* Verlag Dr. Th. Gabler, Wiesbaden 1993, p. 57.
30. C. Christ: "Produktionsintegrierte Umweltschutzmaßnahmen im Lösemittelbereich," *Chem. Tech*. (Heidelberg) 19 (1990) no. 9, 42.
31. C. Christ: "Produktionsintegrierter Umweltschutz – Chancen und Grenzen," *Chem. Ing. Tech.* **64** (1992) no. 10, 889.
32. M. Schneider: *Umweltschutz bei Bayer*, press conference, Leverkusen, Sep. 8, 1993.
33. H. H. Hulpke: "Umweltschutz als gesamtunternehmerische Aufgabe: Aktives und vorbildhaftes Umsetzen," *Schweiz. Maschinenmarkt* **35** (1989) 132.

34. H. H. Hulpke: “Naturwissenschaftliche, technische sowie wirtschaftliche Aspekte und Konsequenzen des Umweltschutzes,” in: *Wirtschaft und Umwelt*, C. F. Müller Juristischer Verlag, Heidelberg 1986, p. 45.
35. H. Kreikebaum: “Integrierter Umweltschutz durch strategische Planungs- und Controlling-Instrumente,” in U. Steger (ed.): *Handbuch des Umweltmanagement*, Verlag C. H. Beck, München 1992, p. 257.
36. H. I. Joschek: “Reststoffe – Vermeidung, Verminderung, Verwertung, Entsorgung in der BASF,” *WLB Wasser Luft Boden* **11/12** (1989) 8.
37. Bayer AG: *Umweltbericht*, Leverkusen 1993.
38. C. Christ: “Umweltschonende industrielle Arbeitsmethoden – Verfahrensverbesserungen und Stoffkreisläufe,” in: *Kreislaufwirtschaft statt Abfallwirtschaft*, Universitätsverlag, Ulm 1994, p. 123.
39. K. Zimmermann: *Präventive Umweltpolitik und technologische Anpassung*, Internationales Institut für Umwelt und Gesellschaft am Wissenschaftszentrum Berlin für Sozialforschung, Berlin 1985.
40. K. Zimmermann, *Konjunkturpolitik* **34** (1988) 327.
41. K. Zimmermann: “Technologische Modernisierung der Produktion – eine Variante präventiver Umweltpolitik,” in U. E. Simonis (ed.): *Präventive Umweltpolitik*, Wissenschaftszentrum Berlin für Sozialforschung, Forschungsschwerpunkte Technik, Arbeit Umwelt, Campus Verlag, Frankfurt/Main 1988, p. 205.
42. K. Zimmermann: *Umweltpolitik und integrierte Technologien, der Quantität-Qualitäts-Trade-off*, Wissenschaftszentrum Berlin für Sozialforschung, Berlin 1989.
43. W. Hopfenbeck: *Umweltorientiertes Management und Marketing*, Verlag Moderne Industrie, Landsberg/Lech 1990.
44. H. Kreikebaum: “Innovationsmanagement bei aktivem Umweltschutz in der Chemischen Industrie – Bericht aus dem Forschungsprojekt,” in G. R. Wagner (ed.): *Unternehmung und ökologische Umwelt*, Verlag Franz Vahlen, München 1990, p. 113.
45. U. E. Simonis: “Ökologische Modernisierung der Wirtschaft – Option und Restriktionen,” in G. R. Wagner (ed.): *Unternehmung und ökologische Umwelt*, Verlag Franz Vahlen, München 1990, p. 29.
46. ifo Institut für Wirtschaftsforschung: *ifo – Spiegel der Wirtschaft*, München 1990.
47. M. Zlokarnik: “Produktionsintegrierter Umweltschutz in der chemischen Industrie,” in: *Produktions- und produktintegrierter sowie additiver und sanierender Umweltschutz*, vol. 2, Springer Verlag, Heidelberg (in press).
48. D. Becher: “Vermeiden – Vermindern – Verwerten, Integrierter Umweltschutz in der Produktion,” in: *Die Bayer-Umweltperspektive II*, Bayer, Leverkusen 1991.
49. C. Christ: “Vermeidung von Abfällen in der chemischen Industrie,” *Entsorgungspraxis Spezial* **1** (1990) 26.
50. *Ullmann*, 4th ed., **6**, 155 ff.
51. *Winnacker-Küchler*, 4th ed., **6**, 713 ff.
52. M. Schmitt: “Zellwolleabgase als Rohstoff zur Schwefelsäureherstellung,” *Chem. Ind*. (Düsseldorf)104 (1981) no. 5, 286 ff.
53. R. Peichl: “Das Kelheimer Modell – ein neuartiges Recycling-Verfahren,” *Chem. Tech*. (Heidelberg)11 (1982) no. 1, 21 ff.
54. J. Dörges, K. Dietrich, B. Sülzer, J. Wiesner: “Schwefelsäure aus den Abgasen der Zellwolleproduktion,” in: *Produktionsintegrierter Umweltschutz in der chemischen Industrie*, DECHEMA, Frankfurt/Main 1990.
55. Les Usines de Melle, US 2 936 321, 1960 (J. Mercier).
56. C. Christ: “Umweltschonende industrielle Arbeitsmethoden – Verfahrensverbesserungen und Stoffkreisläufe,” in Bundesverband der Deutschen Industrie e.V. (eds.): *Industrieller Umweltschutz*, Tagungsband zum Seminar in Zusammenarbeit mit der Universität Leipzig vom 5. bis 10. April 1992 in Leipzig, Köln, Mai 1993, p. 210.
57. Hoechst, DE-OS 36 11 886, 1987 (K. Schuchardt, H. Scholz, H. W. Neuß, H. Adam).
58. E. I. du Pont de Nemours, GB 2 040 897, 1980 (L. J. Harwell).
59. Knapsack, GB 1 070 515, 1967 (A. Czekay, A. Jacobowsky, E. Kottenmeier).
60. Hoechst, US 3 980 758, 1976 (R. Krumböck, W. Kühn).
61. ICI Australia, US 4 464 351, 1984 (V. Vasak, J. Sencar, K. Mok).
62. Hoechst, EP-A 0 362 666, 1990 (W. Freyer, T. Olffers).
63. Hoechst, EP-A 0 464 758, 1992; DE-OS 4 021 408, 1992 (W. Freyer, T. Olffers, A. Riedel).
64. C. Christ: “Integrierter Umweltschutz – Strategie zur Abfallminderung und -vermeidung,” *Chem. Ing. Tech.* **64** (1992) 431.
65. P. Lappe, H. Springer, J. Weber: “Neopentylglykol als aktuelle Schlüsselsubstanz,” *Chem. Z.* **113** (1989) 293.
66. Hoechst, EP 0 278 016 B 1, 1989 (N. Breitkopf *et al.*).
67. P. Sckuhr: “Umweltschutz eine ständige Herausforderung in der Verfahrensentwicklung,” in: *125 Jahre Hoechst AG – Wissenschaftliches Symposium*, May 19 – 20, 1988, Frankfurt/Main.
68. K. Matsumoto: “Production of 6-APA, 7-ACA, and 7-ADCA by Immobilized Penicillin and Cephalosporin Amidase,” in A. Tanaka (ed.): *Industrial Application of Immobilized Biocatalysts*, Marcel Dekker, New York 1993, p. 67.
69. U. Giesecke, F. Wedekind, W. Tischer: “Biocatalytic 7-Aminocephalosporanic Acid Production,” in: *Dechema Biotechnology Conferences* **5,** Part B (1992) 609.
70. Hoechst, DE 4 136 389, 1991 (W. Aretz, K. Koller, G. Rieß).
71. *Houben-Weyl*, 4th ed., vol. **9**, p. 475.
72. *Houben-Weyl*, 4th ed., vol. **9**, p. 476.
73. *Houben-Weyl*, 4th ed., vol. **10/1**, p. 492.
74. *Houben-Weyl*, 4th ed., vol. **6/1 c**, p. 211.
75. *Houben-Weyl*, 4th ed., vol. **6/1 c**, p. 221.

76. D. Becher: *Die Bayer-Umweltperspektive II*, press conference, Sep. 16 – 18, 1991, Leverkusen, p. 41 ff.
77. Bayer AG: *Umweltbericht*, Leverkusen 1993, p. 46 ff.
78. Bayer, DE-OS 2 703 076, DE-OS 2 747 714 ,1977 (O. Barth *et al.*).
79. Bayer, DE-OS 4 212 086, 1992 (W. Ludwig *et al.*).
80. D. I. Klukas, Leuna Werke, paper presented at the Membran-Kolloquium, Aachen, March 11, 1992.
81. *Handbuch Chlorchemie II*, Umweltbundesamt Text 42/92, pp. 101 – 138.
82. Umweltbundesamt (BUA) (ed.): “Chlortoluole,” Stoffbericht 38.
83. *Ullmann*, 4th ed., **9**, 512 – 513.
84. Dow Chem., US 1 946 040, 1931 (W. C. Stoesser).
85. Hooker Chem. & Plas., US 4 024 198, 1975 (H. E. Buckholtz).
86. Hooker Chem. & Plas., US 4 031 142, 1975 (J. C. Graham).
87. Hooker Chem. & Plas., US 4 031 147, 1975 (J. C. Graham).
88. Hooker Chem. & Plas., US 4 069 264, 1976 (H. C. Lin).
89. Hooker Chem. & Plas., US 4 069 263, 1977 (H. C. Lin).
90. Hooker Chem. & Plas., US 4 250 122, 1979 (H. C. Lin).
91. Ihara Chem. Ind., US 4 289 916, 1980 (Y. Nakayama).
92. Hodogaya Chem., EP 63 384, 1981 (R. Hattori).
93. Hoechst, EP 173 222, 1984 (H. Wolfram).
94. Atochem, EP 126 669, 1983 (R. Commandeur).
95. Ihara Chem. Ind., JP 60 125 251, 1983 (J. Kiji).
96. Ihara Chem. Ind., JP 60 136 576, 1983 (J. Kiji).
97. Ihara Chem. Ind., JP 61 171 476, 1985 (J. Kiji).
98. Bayer, US 4 851 596, 1987 (F. J. Mais).
99. Bayer, US 4 925 994, 1988 (F. J. Mais).
100. Bayer, US 4 990 707, 1988 (F. J. Mais).
101. Bayer, US 5 105 036, 1990 (F. J. Mais).
102. *Houben-Weyl*, 4th ed., vol. **9**, p. 477.
103. *Houben-Weyl*, 4th ed., vol. **9**, p. 502.
104. *Houben-Weyl*, 4th ed., vol. **9**, p. 436.
105. M. Bueb, M. Finzenhagen, T. Mann, K. Müller, *Korresp. Abwasser* **37** (1990) 542 – 558.
106. H. Emde, G. Randau, G. Scharfe in: *Produktionsintegrierter Umweltschutz in der chemischen Industrie*, DECHEMA, Frankfurt/Main 1990, pp. 25 – 26.
107. Bayer: *Bayer-Umweltperspektive II*, press conference, Leverkusen 1991, p. 44.
108. Bayer: “Erfolge, Projekte, Ausblick,” *Umweltschutz 2/91*, p. 87.
109. D. Tegtmeyer: “Möglichkeiten und Chancen einer membrantechnischen Abwasserbehandlung in der Textilfärberei,” *Melliand Textilber.* **74** (1993) no. 2, 148 – 151.
110. R. Rautenbach, A. Gröschl: “Umkehrosmose organisch-wäßriger Lösungen,” *F & S Filtrieren Separieren* **4** (1990) no. 4, 231 – 239.
111. T. Janssen, W. Samhaber: “Industrielle Nanofiltration zur Vorbehandlung von chemischen Abwässern,” Sandoz report, Basel 1993.
112. P. Meyer, W Samhaber, C. Hamelbeck: “Vorreinigung von problematischem Chemieabwasser durch Kombination von Adsorption, Membrantrennung und Oxidation,” Lecture held at an Uhde-Symposium at IFAT, Munich, May 25, 1990.
113. VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen: “Kombinierte Verbrennung von Klärschlamm und energiereichem hochchloriertem Kohlenwasserstoff GVC,” Tagung Entsorgung von Sonderabfällen durch Verbrennung, Baden-Baden 1989.
114. Abfallverbrennung am Beispiel des Bayerwerkes Leverkusen, Ein Teil des umweltgerechten Gesamtkonzeptes, Umwelt und Technik, 6/1988.
115. E.-H. Rohe in: *Die Bayer-Umweltperspektive II*, press conference, Leverkusen, Sep. 16 – 18, 1991, pp. 27 – 28.
116. Inauguration of the new Wastewater and Sewage Sludge Incineration Plant, December 2, 1988, in Leverkusen, documentation of lectures, p. 13.15.
117. U. Kaier, “Chancen und Grenzen der Kraft-Wärme-Kopplung,” *VDI Ber.* **808**, 1990.
118. B. Linnhoff, “Pinch Analysis, a State-of-the-Art Overview,” *Transl Chem E.* **71 Part A** 9/1993.
119. U. Neumann, *Chem. Tech. (Heidelberg)* **12** (1983) no. 7, 21 – 24.
120. BASF, DE 4 326 952, 1993.
121. BASF, DE 4 326 953, 1993.
122. F. Hoffmann-LaRoche, EP 5746.
123. F. Hoffmann-LaRoche, EP 5747.
124. S. Jensen *et al.*: “On the Chemistry of EDC Tar and its Biological Significance in the Sea,” *Proc. R. Soc. London B* **189** (1975) 333 – 346. Links
125. Wacker, DE 3 817 938, 1988 (L. Schmidhammer *et al.*).
126. BASF, US 4 233 280, 1980 (G. Devroe, J. Langens).

127. Stauffer Chem., *Hydrocarbon Process.* **60** (1981) no. 11, 170.
128. L. Schmidhammer: "Rohstoffrecycling bei der Vinylchloridherstellung," *NIU Chem.* **4** (1993) no. 16, 48 – 50.
129. K. Olie *et al.*, *Chemosphere* **12** (1983) 627 – 636. Links
130. Wacker, EP 0 177 013, 1985 (G. Dummer, L. Schmidhammer, P. Hirschmann, G. Stettner).
131. Wacker, EP 0 180 998, 1985 (G. Dummer *et al.*).
132. G. Dummer, L. Schmidhammer: "Neues Verfahren zur Entfernung von aliphatischen Chlorkohlenwasserstoffen aus Abwässern mittels Adsorberharzen," *Chem. Ing. Tech.* **56** (1984) no. 3, 242 – 243 (Synopsis); *VDI Ber.* **545** (1983) 885 – 901.
133. Degussa, Wacker, EP 0 461 431, 1991 (K. Deller, H. Krause, L. Schmidhammer, W. Dafinger).
134. Wacker, US 4 754 088, 1988 (L. Schmidhammer, P. Hirschmann, H. Patsch, R. Straßer).
135. Wacker, EP 0 225 617, 1986 (L. Schmidhammer *et al.*).
136. Wacker, EP 0 452 909, 1991 (L. Schmidhammer, K. Haselwarter, H. Klaus, K.-P. Mohr).
137. Wacker, EP 0 002 501, 1978 (L. Schmidhammer, H. Frey).
138. Wacker, EP 0 052 271, 1981 (G. Dummer, L. Schmidhammer, R. Straßer).
139. Wacker, DE 2 803 285, 1978 (L. Schmidhammer, E. Selbertinger).
140. Wacker, EP 0 180 995, 1985 (L. Schmidhammer *et al.*).
141. Wacker, EP 0 180 925, 1985 (L. Schmidhammer *et al.*).
142. G. Dummer, L. Schmidhammer: "Großtechnische Nutzung der Thermokompression bei der Destillation von 1,2-Dichlorethan," *Chem. Ing. Tech.* **63** (1991) no. 1, 74 – 75 (Synopsis 1914).
143. G. Lohrengel, *Gewässerschutz Wasser Abwasser* **135** (1993) 249.
144. G. Buxbaum (ed.): *Industrial Inorganic Pigments*, VCH Verlagsgesellschaft, Weinheim, Germany 1993, pp. 43 – 71.
145. T. Fryer, *Processing (Sutton, Engl.)* (1989) July, 17 – 21.
146. U. Rothe, K. Velleman, H. Wagner, *Congr. FATIPEC* **21** (1992) no. 2, 167 – 186.
147. F. Wagener, J. Wiesner in: *Produktionsintegrierter Umweltschutz in der chemischen Industrie*, DECHEMA, Frankfurt/Main 1990, pp. 91 – 95.
148. K.- G. Malle, *Chem. unserer Zeit* **21** (1987) no. 1, 9 – 16. Links
149. U. Rothe in: *Sulphur 88*, Preprints of British Sulphur's 13th International Conference, Wien 1988, pp. 87 – 95.
150. D. Hody, P. Bansal, *Sulzer Tech. Rev.* **69** (1987) no. 1, 25 – 28.
151. H. Thumm in: *Recycling International, 4th International Recycling Congress*, EF-Verlag, Berlin 1984, pp. 1074 – 1077.
152. G. Dieckelmann, H. J. Heinz: *The Basics of Industrial Oleochemistry*, P. Pomp, Essen 1989.
153. H. J. Richtler, J. Knaut: *Fat. Sci. Technol.* **93** (1991) no. 1, 1 – 13.
154. H. Pardun, *Fette Seifen Anstrichm.* **85** (1983) no. 1, 1 – 18.
155. J. Baltes: *Gewinnung und Verarbeitung von Nahrungsfetten*, Parey, Berlin 1975.
156. A. Davidsohn, *Seifen Öle Fette Wachse* **114** (1988) no. 15, 595 – 600.
157. A. Zellner: Dissertation, Universität GH Duisburg 1989.
158. K. Yoshiharu, O. Toshio, US 4 164 506, 1978.
159. H. Druckrey *et al.*, *Z. Krebsforsch.* **74** (1970) 241 –270. Links
160. Henkel, DE-OS 3 501 761, 1986 (L. Jeromin, E. Peukert, G. Wollmann).
161. Y. M. Choo, K. Y. Cheah, A. N. Ma, A. Halim: "Conversion of Crude Palm Kernel Oil into its Methyl," Proc. World Conference on Oleochemicals 1990 (Publ. 1991, pp. 292 – 295).
162. B. Gutsche, V. Jordan, L. Jeromin: *Chem. Ing. Tech.* **64** (1992) 774.
163. B. Freedman, E. H. Pryde, T. L. Mounts, *JAOCS J. Am. Oil Chem. Soc.* **61** (1984) no. 10, 1638 – 1643.
164. Henkel, DE-OS 3 932 514, 1991 (G. Aßmann *et al.*).
165. Henkel, DE-OS 3 911 538, 1991 (G. Aßmann *et al.*).
166. H. Wiederkehr: "Reduktion der Umweltbelastung durch Prozeßoptimierung in der chemischen Produktion," *Chimia* **40** (1986) no. 9, 323 – 330. Links
167. F. Hoffmann-La Roche, CH 642 936, 1984.
168. J. Jeisy: Vortrag auf dem Jahrestreffen der Verfahrens-Ingenieure, Sep. 29, 1982, Basel.
169. J. Jeisy *et al.*: "Beseitigung acetylenhaltiger Abgase," *Chem.-Ing.-Tech.* **55** (1983) no. 5, 388 – 389. Links
170. G. H. Lenske, J. Findling: "Ein Gasstrahler zur Förderung von Kopfgasen der Rohöldestillation ins Heizgasnetz," *Erdöl Kohle Erdgas Petrochem.* **35** (1982) no. 2.
171. H.-G. Franck, J. W. Stadelhofer: *Industrielle Aromatenchemie*, Springer Verlag, Berlin 1987, p. 36.
172. H.-G. Franck, J. W. Stadelhofer: *Industrielle Aromatenchemie*, Springer Verlag, Berlin 1987, pp. 324 – 344.
173. S. Rittner, R. Steiner, *Chem. Ing. Tech.* **57** (1985) no. 2, 91 – 102.
174. A. Papp, K. Saxer, *vt Verfahrenstech.* **15** (1981) no. 3.
175. Shell, US 3 562 239, 1966.
176. Shell, US 4 414 132 and 4 329 253, 1979.
177. *The Clean Air Act and the Refining Industry*, UOP, Sep. 1991.
178. C. Higman, G. Grünfelder: "Clean Power Generation from Heavy Residues," *Proc Inst Mech Eng, IMech E Conf* 1990, no. 13.

179. P. Ladeur, H. Bijwaard: “Shell plans $ 2.2-billion Renovation of Dutch Refinery,” *Oil Gas J.* **91** (1993) no. 17, 64 – 67. Links
180. E. Stöldt: “Concepts for Zero Residue Refineries,” Interpec China '91, Beijing, Sep. 1991.
181. H. Heurich, C. Higman: “Partial Oxidation in the Refinery Hydrogen Management Scheme,” AIChE Spring Meeting, Houston, 1993.
182. C. Higman, M. Eppinger: “The Zero-Residue Refinery using the Shell Gasification Process,” Achema 1994, Frankfurt/Main.
183. M. J. Niksa: “Acid/Base Recovery from Sodium Sulphate,” 5th International Forum on Electrolysis in the Chemical Industry, Electrosynthesis Inc., Nov. 10 – 15, 1991.
184. H. V. Plessen *et al.*, *Chem. Ing. Tech.* **61** (1989) no. 12, 933 – 940.
185. Olin, US 5 084 148, 1992.
186. Hoechst, US 4 613 416, 1986.
187. J. Jorisse, K. H. Simmrock, *J. Appl. Electrochem.* **21** (1991) 869 – 876.
188. E. Hillrichs, K. Lohrberg, *Dechema Monogr.* **125** (1992) 221 – 232.
189. General Electric, US 4 561 945, 1985.
190. De Nora Permelec, WO 93/0046, 1993.
191. U.S. Dept. of Energy, US 4 876 115, 1989.
192. Chem. Systems Inc.: Bipolar Membranes, Report no. 86–9, USA, Dec. 1987.
193. K. N. Mani *et al.*, *Desalination* **68** (1988) 149 – 166. Links
194. B. Pieper: “Natriumsulfat-Elektrolyse in der Membranzelle,” Reihe 3, VDI Verlag, Düsseldorf 1987.
195. G. Cowley et al.: “Electrochemical Generation of Chloric Acid at High Current Efficiencies,” 5th International Forum on Electrolysis in the Chemical Industry, Electrosynthesis Inc., Nov. 10 – 15, 1991.
196. Bayer, US 5 071 522, 1991.
197. N. Kimura: *Senjo-Sekkei*, Kindai-Hensyu Ltd., Tokyo 1991, pp. 8 – 12.
198. H. Matsuzaki: *Senjo-Sekkei*, Kindai-Hensyu Ltd., Tokyo 1987.
199. T. Ohmi et al.: Nikkei Microdevices, May 1988, p. 101.
200. N. Miyake, Y. Matsuo: Outline of Membrane Operation Technology, Fuji Techno System Ltd., Tokyo 1991, p. 933.
201. Y. Sugimoto in , [200] p. 683.
202. Y. Hiratsuka, H. Sato, N. Hashimoto, T. Shinoda: *Senjo-Sekkei*, Kindai-Hensyu Ltd., Tokyo 1989, pp. 2 – 12.
203. M. Okada, J. Yonamoto, *Ultraclean Technology* **3** (1991) no. 2, 203.
204. Boehringer Mannheim (ed.): *Umweltschutz durch Biotechnik*, Mannheim 1990.

[Top of Page]

About Wiley InterScience | **About Wiley** | **Privacy** | **Terms & Conditions**
Copyright © 1999-2007**John Wiley & Sons, Inc.** All Rights Reserved.

Plant and Process Safety

Volker Pilz, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 1, Sections 3.1 – 3.3)

Herbert Bender, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 2.1)

Michael Müller, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.2.1, Section 3.5.3 in part)

Dietrich Conrad, Berlin, Federal Republic of Germany (Section 2.2.2)

Claus-Diether Walther, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.2.2)

Werner Berthold, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany(Section 2.2.3)

Martin Glor, Ciba-Geigy AG, Basel, Switzerland (Section 2.2.3)

Peter-Andreas Wandrey, Bundesanstalt für Materialforschung und –prüfung, Berlin, Federal Republic of Germany (Section 2.2.4)

Karl-Heinz Mix, Bayer AG, Leverkusen, Federal Republic of Germany (Section 2.2.4)

Jörg Steinbach, Schering AG, Berlin, Federal Republic of Germany(Section 2.3)

Albert Eberz, Bayer AG, Leverkusen, Federal Republic of Germany(Section 2.3)

Francis Stoessel, Ciba-Geigy AG, Basel, Switzerland(Section 2.3)

Hans Hagen, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 3.4.1, Section 5.1 in part)

Helmut Schacke, Bayer AG, Leverkusen, Federal Republic of Germany (Section 3.4.2)

Richard Viard, Bayer AG, Leverkusen, Federal Republic of Germany (Section 3.4.2)

Bernd Schrörs, Bayer AG, Leverkusen, Federal Republic of Germany (Section 3.4.3)

Stephan Weidlich, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Section 3.4.3)

Stefan Drees, Bayer AG, Leverkusen, Federal Republic of Germany (Section 3.5.1)

Günter Hesse, Bayer AG, Brunsbüttel, Federal Republic of Germany (Section 3.5.2)

Hans Förster, Physikalisch-Technische Bundesanstalt, Braunschweig, Federal Republic of Germany (Section 3.5.3 in part)

Klaus Bartels, Berufsgenossenschaft der chemischen Industrie, Heidelberg, Federal Republic ofGermany (Section 4.1)

Ulrich Widmer, Sandoz, Basel, Switzerland (Section 4.2.1)

Adrian Geiger, Sandoz, Basel, Switzerland (Section 4.2.1)

Klaus Noha, Hoechst Aktiengesellschaft, Frankfurt, Federal Republic of Germany (Section 4.2.2)

Edmund Müller, BASF Aktiengesellschaft, Ludwigshafen, Federal Republic of Germany (Section 4.3)

Dieter Grenner, Bayer AG, Dormagen, Federal Republic of Germany (Sections 4.4, 4.5)

Nikolaus Schulz, Bayer AG, Dormagen, Federal Republic of Germany (Sections 4.4, 4.5)

Jürgen Zimmermann, Bayer AG, Dormagen, Federal Republic of Germany (Sections 4.4, 4.5)

Jürgen Harbordt, Bayer AG, Leverkusen, Federal Republic of Germany (Section 5.1 in part)

Matthias Walper, Bayer AG, Brunsbüttel, Federal Republic of Germany (Sections 5.2, 5.3)

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_311

1. Safety Problems in Chemical Plants . . . 5
1.1. Types and Sources of Hazards . . 5
1.1.1. Fundamentals . . . 5
1.1.2. Causes and Effects . . . 6
1.2. Requirements for Safe Chemical Plants . . . 7
1.3. Definitions . . . 8
1.4. Magnitude of Hazard Potentials and Risks . . . 8
1.4.1. Risk According to Previous Experience . . . 8
1.4.2. Hazard Potential from Release of a Volatile Substance . . . 9
1.4.3. Hazard Potential from an Explosion 11
1.5. Conclusions . . . 12
2. Determination and Evaluation of Hazardous Properties of Substances (Safety Ratings) . . . 13
2.1. Harmful Effects of Substances . . 13
2.2. Ratings of Flammable and Explosive Substances . . . 16
2.2.1. Flammability Ratings . . . 16
2.2.1.1. Introduction . . . 16
2.2.1.2. Combustibility/Flammability of Chemicals . . . 16
2.2.1.3. Oxidizing Properties of Chemicals . 19
2.2.1.4. Flammability of Construction Materials . . . 19
2.2.2. Explosion Data for Gas Mixtures . 20
2.2.2.1. Introduction . . . 20
2.2.2.2. Explosion Limits; Limiting Oxygen Concentration . . . 20
2.2.2.3. Maximum Explosion Pressure and Maximum Rate of Pressure Rise . . 23
2.2.2.4. Pressure Limit of Stability for Unstable Gases . . . 25
2.2.2.5. Ignition Temperature, Ignition Energies . . . 26
2.2.3. Explosion Indices of Dust-Air Mixtures . . . 28
2.2.3.1. Introduction . . . 28
2.2.3.2. Ignition Sensitivity of Dust Clouds 28
2.2.3.3. Dust Explosibility . . . 31
2.2.4. Characterization of Explosive Condensed Substances . . . 33
2.2.4.1. Introduction . . . 33
2.2.4.2. Chemical Nature and Definition of an Explosion . . . 33
2.2.4.3. Grouping of Explosive Substances . 34
2.2.4.4. Explosion Mechanisms . . . 34
2.2.4.5. Testing of Explosive Substances . . 35
2.2.4.6. Conclusions . . . 36
2.3. Exothermic and Pressure-Inducing Chemical Reactions . . . 36
2.3.1. Introduction . . . 36
2.3.1.1. Exothermic Reactions: Runaway Potential . . . 36
2.3.1.2. Causes and Consequences of Overpressure-Inducing Exothermic Reactions . . . 38
2.3.1.3. Processes Hazard Assessment and Safety Evaluation for Exothermic Reactions . . . 39
2.3.1.4. Hazard Characteristics of Exothermic Processes due to Process Design 39
2.3.2. Methods of Investigation and their Systematic Application . . . 41
2.3.2.1. Screening Methods . . . 41
2.3.2.2. Thermal Aging and Heat Accumulation Storage Tests . . . 42
2.3.2.3. Calorimetry . . . 43
2.3.2.4. Systematic Testing . . . 44
2.3.3. Assessment Criteria . . . 45
2.3.3.1. Runaway Scenario . . . 45
2.3.3.2. Severity: Adiabatic Temperature Increase . . . 45
2.3.3.3. Probability and Kinetics of a Runaway . . . 46
2.3.3.4. Assessment of Criticality . . . 46
2.3.4. Measures . . . 48
3. Development, Design, and Construction of Safe Plants . . . 48
3.1. Objectives, Regulations, and Concerns . . . 48
3.2. Procedure for Designing and Constructing Safe Plants . . . 50
3.3. Methodological Aids . . . 52
3.4. Safe Processing: Strategies . . . 57
3.4.1. Fire Protection . . . 57
3.4.1.1. Fire Prevention . . . 57
3.4.1.2. Limitation of Fire Effects . . . 58
3.4.2. Explosion Prevention and Protection 60
3.4.2.1. Introduction . . . 60
3.4.2.2. Hazard Identification . . . 61
3.4.2.3. Explosion Risk . . . 61
3.4.2.4. Explosion Risk Reduction . . . 68
3.4.2.5. Legal Aspects . . . 70
3.4.3. Safety Techniques Based on Process Control . . . 71
3.4.3.1. Integration of PCE into the Safety Concept . . . 71
3.4.3.2. Classification of PCE Systems . . . 72
3.4.3.3. Requirements for PCE Equipment for Process Plant Safety and Design Principles . . . 74
3.4.3.4. Operation of PCE Safety System . . 77

3.4.3.5. Summary 78
3.5. Special Safety Equipment 78
3.5.1. Pressure-Relief Devices 78
3.5.1.1. Introduction 78
3.5.1.2. Safety Valves 79
3.5.1.3. Bursting Disks 81
3.5.1.4. Sizing of Safety Valves 81
3.5.2. Blowdown Systems 83
3.5.2.1. Procedure 83
3.5.2.2. Alternatives 83
3.5.3. Flame Arresters and Explosion Barriers 84
3.5.3.1. Introduction 84
3.5.3.2. Flame Arresters for Mixtures of Vapors (Gases) and Air 84
3.5.3.3. Flame Arresters for Dust – Air Mixtures 86
4. Safe Plant Operation 87
4.1. Safe Handling of Chemicals 87
4.1.1. Introduction 87
4.1.2. Normal Operation 87
4.1.3. Sampling 88
4.1.4. Cleaning of Vessels 89
4.1.5. Safety Practices when Working with Hazardous Substances Under Abnormal Operation 89
4.2. Safety in Batch and Continuous Processes 90
4.2.1. Safety in Batch Processes 90
4.2.1.1. Introduction 90
4.2.1.2. The Operating Manual 90
4.2.1.3. The Human Aspect of Safety 93
4.2.1.4. Internal Organization and Policies . 94
4.2.1.5. Safety in Production Practice 95
4.2.2. Safety in Continuous Processes . . . 96
4.2.2.1. Introduction 96
4.2.2.2. Analogies with Batch Processes . . 96
4.2.2.3. Special Features of Continuous Processes 97
4.3. Technical Inspection 99
4.4. Maintenance 101
4.4.1. Actions During Plant Design and Construction 101
4.4.2. Actions During Plant Operation . . 103
4.4.2.1. Maintenance Activity 103
4.4.2.2. Performance of Maintenance 107
4.4.3. Continuing Plant Development (Analysis of Weak Points) 110
4.5. Modification of Plants 110
4.5.1. Reasons for Modifications 110
4.5.2. Procedure for Modification 111
4.5.3. Summary of Plant Maintenance and Process Modification 112
5. Hazard Control 112
5.1. Means of Limiting Accident Impacts 112
5.1.1. Retention Systems, Catch Wells . . 113
5.1.2. Rapid-Closing Valves, Emergency Compartmentalization Systems . . . 114
5.1.3. Emergency Drain and Collection Systems 114
5.1.4. Blowoff and Disposal Systems . . . 114
5.1.5. Spray-Curtain (Drench) Systems . . 114
5.1.6. Partial and Complete Containment . 114
5.1.7. Fire Protection; Retention of Water Contaminated by Fire Fighting . . . 115
5.2. Hazard Control Plans at Plant Level and Beyond 115
5.2.1. Hazard Control Plans 115
5.2.1.1. Plant Hazard Control Plan 116
5.2.1.2. Works Hazard Control Plan 116
5.2.1.3. Required Contents of the Plan . . . 116
5.2.2. Off-Battery Hazard Control Plans . 117
5.3. Public Awareness and Responsible Care 117
6. References 118

The entire topic was coordinated by Volker Pilz

1. Safety Problems in Chemical Plants

1.1. Types and Sources of Hazards

1.1.1. Fundamentals

Large-scale chemical production is one of the most important industrial activities. Its products play key roles in human nutrition, health, quality of life, and welfare. Nevertheless, many people regard chemical production as dangerous, even though chemical processes are associated with accident frequencies which are much lower than in other industries – and very few such accidents have anything to do with chemistry [2].

Despite the good accident statistics of chemical plants, grave accidents have occurred, endangering human life and the environment, and causing considerable damage. The names

Flixborough (1974) [3], Seveso (1976) [4], Bhopal (1984) [5], and Basel (1986) [6] call spectacular events to mind.

Three types of event are traditionally associated with the chemical industry:

Releases and spills (Seveso, Bhopal)
Fires (Basel)
Explosions (Flixborough)

"Explosion" as used here refers to an event in which large amounts of energy are released very rapidly (ultimately, in the form of expansion energy), producing a pressure wave that is (at least) audible and propagates at high velocity, decaying over distance.

Damage caused by the above mentioned three types of event comes about through two distinct processes:

Direct action of chemicals on humans and the environment
Indirect action of liberated energy

In the Bhopal accident, many people died after inhaling a dangerous gas that had escaped from a tank storage facility [5]. Most of the damage from the Basel warehouse fire was due to fire fighting water carrying starting material, decomposition products, and combustion products out of the stored material, harming the ecosystem of the Rhine [6]. The consequences of the Flixborough explosion occurred largely through the destructive action of the pressure wave, initiated by ignition of a cloud of escaped flammable gases [3].

Generally, chemical reactions underlie such harmful effects. When substances act directly on living matter, a chemical reaction may alter or destroy cellular tissue (toxic effect). In an explosion, energy is liberated by an exothermic chemical reaction.

It is obvious that the dangers arising from chemical plants have much to do with the nature of the substances being processed, the way they are treated in the plant, and their tendency to take part in chemical reactions under these conditions.

A conflict arises between process needs and chemical safety. On the one hand, chemistry requires reactive substances; on the other hand, this reactivity of the substances is a key aspect of the danger they pose.

A chemical process is impossible without substances that show hazardous properties and effects. The substances must therefore be reliably contained in the process equipment, and their reactivity must be governed so that uncontrolled chemical reactions cannot take place.

1.1.2. Causes and Effects

Releases and Spills. An uncontrolled release from a closed facility presupposes an undesired opening of the containment (e.g., of a valve), or damage to the walls of vessels, process equipment, fittings, or piping (e.g., as a consequence of corrosion, seal failure, or rupture).

If liquids and solids are released, people outside the plant are, as a rule, not directly endangered (or at worst after some time delay) and they are thus more easily protected. The material, on the other hand, can penetrate into the soil, contaminating the groundwater or passing through sewers into surface waters, posing a danger to the environment. Precautions inside the facility include erection of barriers (liquidproof floors, seals on sewers, and retention systems; see Section 5.1).

If the substances are gaseous or finely dispersed (aerosols), they are transported by the motion of the ambient air after their release, so that they can be spread over large areas, diluted by convection and diffusion, separated from the air by sorption and possibly sedimentation, and gradually degraded through chemical reactions with the environment.

Above certain concentration levels in air, pollutants can harm people, fauna, and flora in downstream regions. The amount of damage depends on the exposure time (dose). Usually, such a drifting gas or aerosol cloud causes a short-term burden on individuals exposed to varying concentrations. Toxicology must provide information about these harmful effects (see Section 2.1).

Explosions. An explosion involves the rapid, almost instantaneous release of heat and expansion energy resulting in the formation of large gas volumes and a (destructive) pressure wave, which propagates at high velocity. At a fixed point, the wave causes a sudden rise in pressure

followed by a pressure drop, all within a fraction of a second.

The pressure wave arises because the volumes of gas formed cannot escape quickly enough by mass motion. The destructive action of the wave results from the pressure force and the rapid transfer of momentum to objects in the region of wave propagation.

Energy-supplying processes in an explosion can be chemical or purely physical. There are three different types of chemical explosion. The best-known is the instantaneous decomposition of sensitive chemical compounds to gaseous species occupying a large volume, because it figures in the use of explosives. The process is initiated by external excitation. The volume required for the gaseous products is markedly increased by the heat of decomposition evolved in the process. Decomposition can be excited mechanically by shear forces (friction) or impact; thermally by heating; or by the chemically generated shock wave from an igniter (see Section 2.2.4).

The second chemical explosion process is also fairly well known: rapid oxidation of gases or finely dispersed particles in oxidizing gases (see Sections 2.2.2, 2.2.3). A highly exothermic reaction (usually combustion) heats the gases to high temperature so quickly that the pressure in confined systems generally increases by a factor of 6 – 10 (depending on the situation); in unconfined systems, the increase in volume can bring about a pressure wave with a peak overpressure of up to 100 kPa, with a positive pressure phase lasting for a few to a few hundred milliseconds (depending on the quantity of material that takes part in the spontaneous reaction, how well the fuel gas is mixed with air, how much turbulence there is to accelerate the combustion process, and how well the reacting mixture is confined). This type of explosion can be prevented "classical" anti-explosion measures (see Section 3.4.2).

The third type of chemical explosion, less well known, at least among the general public, is the "thermal runaway" explosion, initiated by a homogeneous exothermic reaction that goes out of control.

If the heat of an exothermic reaction cannot be removed, it further raises the temperature of the system. The reaction velocity increases, more heat is produced per unit time, and the system continues to heat up more and more rapidly until the maximum possible reaction velocity has been reached. The accompanying temperature rise can lead to a dangerous pressure buildup in two ways: by raising the vapor pressure of reactants and solvents, or by initiating another reaction (e.g., a decomposition) that liberates gases (see Section 2.3). If the pressure increase leads to overstressing of the reactor, it may burst and the expansion of its contents may generate a pressure wave.

In principle, an explosion can also be brought about by a purely physical process. A sudden liberation of expansion energy is possible after gases have been compressed or when vaporization is abruptly initiated, e.g., if a hot, condensed substance is quickly mixed with another substance that vaporizes readily. So far, this phenomenon does not seem to have played a major role in chemical plant explosions, and it is not discussed here; but its possibility must be considered in relation to plant safety.

Fires. Fires in chemical plants can occur if flammable substances are released and come in contact with an ignition source. Section 2.2.1 deals with the substance properties and ambient conditions that must be present for fires to start; Section 3.4 discusses measures to control fires.

The effects of fires relate to the action of heat (energy evolved mainly as thermal radiation) and the release of pollutants (combustion gases and decomposition products).

1.2. Requirements for Safe Chemical Plants

The phenomenological discussion in Section 1.1 showed that most of the dangers in the operation of chemical plants have to do with the reactivities of the substances present. The principal hazards are caused by releases and spills of substances outside the plant, and by uncontrolled chemical reactions between substances. The danger is greater, the larger the quantities of substances and energy released.

Plant and process safety efforts must have as their essential goals:

1) To minimize the quantities of substances (Section 3.1)
2) To control the potential risks that remain

Objective (2) requires:

Reliably confining hazardous substances in process equipment that can withstand the anticipated stresses (due to pressure, temperature, corrosive attack, etc.).

Insuring that process parameters (e.g., pressure, temperature, concentration) do not take on values such that the substances can undergo uncontrolled reactions.

Some help is gained from the fact that substances, as a rule, react spontaneously with one another only if they have been suitably prepared (e.g., by treatment to increase their surface area); mixed and concentrated; and excited (e.g., by heat).

Experimental laboratory tests are performed, e.g., to determine the temperature and concentration conditions under which reactions can get so far out of control that they can lead to an explosion, or to establish the initiation conditions that can cause dangerous reactions.

A point of special interest is where the regions begin in which uncontrolled reactions cannot occur. The characteristic (critical) parameters found in such experiments are called safety ratings. They include such figures as the ignition temperature and the explosion limits. Safety ratings form the basis for the safety measures in chemical plants. All of Section 2 concerns how safety ratings are determined; the account is broken down by the event classes of Section 1.1.

At the planning and design stage, plant and process safety requires taking steps so that critical concentrations, temperatures, and pressures are not reached. This is achieved by appropriate process design and process control engineering (Chap. 3). Safety also entails preventing the occurrence of potential ignition sources that could cause an undesired reaction (e.g., hot surfaces or sparks generated by mechanical or electrical equipment); great care needs to be taken in the layout of machinery and process equipment (Section 3.4).

Critical process parameters and hazardous potential ignition sources must not occur in the plant as a result of process upsets or human error. These become an additional concern of plant safety, requiring a painstaking cause-and-effect analysis of all possible errors and malfunctions, and the institution of measures to prevent or neutralize situations that could lead to an unsafe condition. Such measures may be technical or organizational; the second group includes operating instructions, inspections, etc. Chapters 3 and 4 present a fuller treatment of these points.

The complete absence of all possible hazards – absolute safety – is not possible, for several reasons.

1) First, it cannot be ruled out that several safety measures will fail simultaneously, so that a potential hazard may become an actual one
2) People make errors from time to time, and they can misjudge things, assess them wrongly, even fail to notice them at all

The fact that absolute safety is impossible is important; two inferences follow from it. Knowledge has to be continually advanced, and precautions have to be taken to avert risks in case of failure (Chap. 5).

1.3. Definitions

The fact that absolute safety is not possible leads to two key questions:

1) What does plant safety mean in the first place?
2) How safe is safe enough?

These intimately related questions can be answered in two different ways. In the first approach, safety is defined in terms of risk; a plant is said to be safe if the risk created by it is acceptable.

Here "risk" means the possibility of harm, defined in terms of the probability of the harm during the lifetime of the plant (or the frequency of the harm) and its anticipated severity.

In the second approach, a plant is said to be safe if it complies with the appropriate regulations and codes. Then it is also said to be safe enough.

While the first method is based on a probabilistic concept, risk, the second employs a deterministic principle. In both cases, what is safe is clearly a matter of convention.

The German standard DIN 31 000, Part 2, defines safety as "a state of affairs in which the risk is not greater than the greatest acceptable risk due to the technical process or condition under consideration" [7].

The standard states that this risk is generally not quantifiable, since only in rare cases can it

be expressed as R, the product of a frequency F and a measure of severity S:

$$R = FS \tag{1}$$

The standard treats danger as the diametric opposite of safety, where the risk of a process is greater than the acceptable limiting risk (Fig. 1).

A useful notion in plant and process safety is the hazard potential, a measure of the greatest harm that can occur in the worst possible event in a plant or plant subdivision. It is reasonable to use this concept in assessing safety measures in a plant: the greater the hazard potential, the more and better safety measures are needed to lower the probability of occurrence of the undesired event to the point that the level of risk is at or below the acceptable risk level.

Safety measures may include intrinsic measures and conditions [8], which insure a priori that a hazard potential can become real only in the event of a relatively improbable combination of multiple independent failures.

Where product quality considerations make it necessary to design the process and process control system so as to prevent an exothermic reaction getting out of control, safety or protective measures can be built up on the basis of this intrinsic safety in order to lower the risk to the acceptable level.

1.4. Magnitude of Hazard Potentials and Risks

The risks created by chemical plants are often overestimated if their principles are not known.

1.4.1. Risk According to Previous Experience

An anticipated value for the risk posed by chemical plants to employees or uninvolved third parties can be derived in a relatively simple way by statistical analysis of historical data.

Consider the risk of death incurred by a chemical worker due to a typical chemical accident (poisoning, chemical burn, explosion). In the Federal Republic of Germany, this risk can be determined by analysis of the annual reports of the mutual accident insurance association of chemical industry, the so-called *Berufsgenossenschaft der chemischen Industrie* [2]. When the number of persons per year suffering death from poisoning, chemical burn, fire, or explosion is divided by the total number of persons employed in the chemical industry, the annual individual lethal risk averaged over the period 1983 – 1992 is ca. 7×10^{-6} a^{-1}, i.e., statistically, 7 persons in 10^6 die every year owing to an on-the-job chemical accident. This risk is a factor of 20 less than the risk of dying in a traffic accident in Germany (ca. 1.5×10^{-4} a^{-1} [9]) and is comparable to the risk of drowning (ca. 8×10^{-6} a^{-1} [9]).

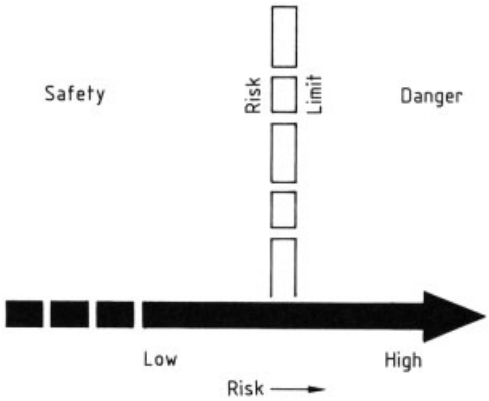

Figure 1. Risk chart [7]

Figure 2 illustrates the situation. The vertical axis shows the probability of any employee dying in the next year from one of the causes cited (individual lethal risk). For comparison, values for a range of risks, [9–12], are also plotted. In terms of a statistical average, a chemical worker is much more likely to be killed in a traffic or domestic accident than in an on-the-job accident. The worker is also more likely to be a victim of violent crime than to die in an industrial chemical accident. His on-the-job risk is low compared with other risks threatening his life.

Those living nearby and others outside the chemical plant are even safer from chemical effects, because the effects of the infrequent incidents in chemical plants fall off quite rapidly with distance. It can be assumed that this risk is, at most, of the same order as the risks due to natural catastrophes [10]. In Germany, the past 50 years have seen no identifiable serious per-

sonal injuries or deaths outside a chemical plant site resulting from accidents inside. This shows that the German chemical industry, like those in many other industrialized countries, operates very safely.

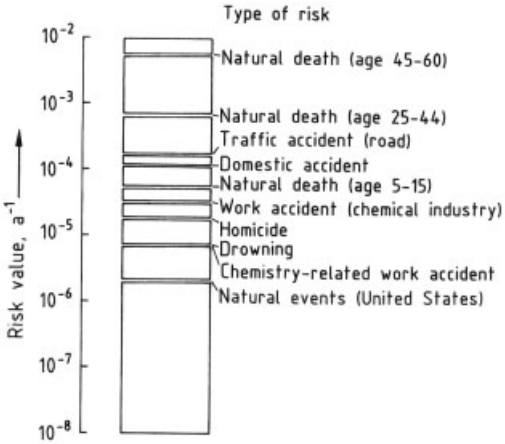

Figure 2. Comparison of lethal risks to humans (per person per year) in the Federal Republic of Germany [2], [9–11]

However, it is true that a low risk may well conceal high hazard potentials, when the probability of occurrence is low. It is therefore advantageous to consider the size of hazard potential in chemical plants.

Analysis of hazard potentials is broken down by type of accident:

Spills and releases
Explosions

1.4.2. Hazard Potential from Release of a Volatile Substance

When a volatile substance posing a health hazard is released, the hazard arises because the liberated gas is taken up by air currents, and propagates in the atmosphere as it is steadily diluted. The size of the impact area, the region in which concentration levels of the substance rise above a critical value φ_{H}, is one measure of the hazard potential; ultimately, however, the hazard potential is governed by the number of persons who stay in this impact area without protection.

To assess the hazard potential, it is necessary to determine the greatest mass of gas that can be released (in accordance with the model).

Consider a vessel that contains a mass M_{L} of a substance in the liquid state under its vapor pressure p_{v} at a temperature T_{α}, above the boiling point T_{b}. The mass of vapor in the vessel is M_{v}. If the containment fails, the contents of the vessel expand to ambient pressure; the liquid continues to vaporize and cool until the remaining liquid reaches its boiling point T_{b} at atmospheric pressure.

Because this process is very fast, it approximates to an adiabatic expansion. The mass of substance vaporized plus the original mass of vapor present are released into the atmosphere; this is the quantity that determines the hazard potential. A material and energy balance, with some simplifications, gives, for the mass vaporized:

$$\Delta M_{\mathrm{v}} = M_{\mathrm{F}(\alpha)} \left[1 - \exp\left(-c_{\mathrm{L}} \Delta T / \Delta h_{\mathrm{v}}\right)\right] \tag{2}$$

where $\Delta T = T_{\alpha} - T_{\mathrm{b}}$, c_{L} is the specific heat of the liquid, and Δh_{v} is the heat of vaporization.

If the temperature change is small, i.e., $(c_{\mathrm{L}}/\Delta h_{\mathrm{v}})\Delta T \ll 1$, a series expansion leads to the approximation:

$$\Delta M_{\mathrm{v}} \approx M_{\mathrm{F}(\alpha)} \left(c_{\mathrm{L}} \Delta T / \Delta h_{\mathrm{v}}\right) \tag{3}$$

A typical value for $c_{\mathrm{L}}/\Delta h_{\mathrm{v}}$ is $5 \times 10^{-3}\,\mathrm{K}^{-1}$. For ΔT, a value of ca. 50 °C, corresponding to an overpressure of 0.5 – 1.0 MPa, is not unusual for the storage of liquefied gases. Calculation shows that with these assumptions 20 – 30 % of the liquid contents of the vessel vaporizes spontaneously. A small additional amount vaporizes because of heat supplied from the surroundings, especially if liquid escapes from the vessel in finely divided form entrained with the vapor. In any case, this simple analysis shows that only a fraction of liquefied mass under its vapor pressure in a vessel can generally be spontaneously released into the atmosphere on failure of confinement, even when substances are kept under high pressure well above their boiling point. Only if the contents are superheated by some hundreds of degrees will the released fraction approach 100 %.

To evaluate the acute hazard created by such a process, atmospheric propagation and dilution of the gas cloud have to be calculated, but these processes show an extremely strong dependence on the motion and turbulence of the atmosphere. Results can differ by orders of magnitude; what

is more, initial dilution (e.g., dilution due to momentum of released vapor), height of release, and gas density all play roles. Exact prediction is impossible, but a rough estimate is possible [13]: under unfavorable conditions, ca. 1000 m^3 (reduced to standard temperature and pressure) or several tonnes of gas have to be released to establish maximum concentrations of ca. 10 ppm out to a range of several kilometers from the point of release. Under propagation conditions of higher probability, the concentrations drop below such levels at a range of less than 1 km [13], [19].

As a rule of thumb, a hazard to humans due to a volatile toxic gas release at a range of more than a few hundred meters from the source presupposes catastrophic failure of a vessel, and the vessel must contain several tonnes of the gas in releasable form (compressed or liquefied). Such apparatus can be found in chemical plants, but the kind of catastrophic damage under discussion is extremely improbable because such vessels are designed with safety factors against stress, and, they are subject to regular inspection.

From the safety standpoint, it is still important to assess hazard potentials relative to one another and to know how they can be influenced by process parameters.

The following argument can be of help. The simple equations describing gas propagation show [13] that the maximum concentration in a gas cloud (plume) at distance x downwind of the source is directly proportional to the quantity released M_{rel} and inversely proportional to the nth power of x:

$$\varphi = \text{const}\ (M_{\mathrm{rel}}/x^n) \tag{4}$$

where $2 < n < 3$ in general.

In the case of a release near the surface, the downwind impact area, the region swept over by a concentration higher than φ_{H}, is commonly cigar-shaped. It becomes broader and shorter, the more turbulent the atmosphere.

For a given state of atmospheric turbulence (i.e., fixed n and constant of proportionality), the (maximum) width of the impact region is always proportional to its length. Therefore, the impact region, as a measure of the hazard potential (HP), is proportional to the square of the distance x from the source to the point where the concentration falls below φ_{H}:

$$HP \sim x^2 = \text{const}\ (M_{\mathrm{rel}}/\varphi_{\mathrm{H}})^{2/n} \tag{5}$$

For the case under discussion (spontaneous vaporization and escape of a substance posing a health hazard and having a critical concentration level φ_{H}) the hazard potential is:

$$\mathrm{HP} = K \left\{ \frac{c_{\mathrm{L}} \Delta T}{\Delta h_{\mathrm{v}}} \frac{M_{\mathrm{F}(\alpha)}}{\varphi_{\mathrm{H}}} \right\}^{2/n} \tag{6}$$

Nevertheless, this is not a well-defined quantity, because K and n depend on the ambient conditions (state of motion of the atmosphere) and can only be derived from the known gas dispersion formula if a special case is assumed [13], [19].

The hazard potential is also a function of the quantity of substance, its properties, and its thermodynamic state inside the vessel. The hazard potential increases with the mass of substance present and its superheating (above boiling point), i.e., with pressure.

From this, it is possible to devise ways to reduce the hazard potential, even to the point of inherent safety.

1.4.3. Hazard Potential from an Explosion

The hazard potential of an explosion is due to the pressure wave, and is determined by the size of the (ideally circular) area around the explosion center in which there is significant damage, e.g., collapsing walls or broken windows. Window glass typically breaks at a peak overpressure in the 1 kPa range and masonry walls begin to fail at overpressures ≥ 30 kPa [14], [15]. Brasie and Simpson speak of serious damage at peak overpressures > 15 kPa [15].

The task of evaluating the hazard potential can thus be reduced to determining at what distance from the explosion center the peak pressure just falls below a critical value.

Suppose the mass M of substance being handled is such that its sudden decomposition liberates an enthalpy of reaction Δh_{r}. The total energy evolved is:

$$E = M \Delta h_{\mathrm{r}} \tag{7}$$

Hopkinson and Cranz independently found that for two distinct "charges" (releasing energies E_1 and E_2) to produce equal effects, the radii (distances R_1 and R_2 from the source point)

must be related as the cube roots of the energies [16]:

$$\frac{R_1}{R_2} = \frac{E_1^{1/3}}{E_2^{1/3}} \tag{8}$$

This result led SACHS to postulate that the dimensionless peak overpressure and the dimensionless momentum of a pressure wave (the quantities that govern the destructive action of the wave) can each be represented as a unique function of a dimensionless distance from the origin [16], [17]. The dimensionless peak overpressure $p*$ is obtained by dividing by the ambient pressure; the dimensionless range $R*$, from an expression involving the energy and ambient pressure:

$$p^* = \frac{p_\mathrm{p} - p_0}{p_0} \tag{9}$$

$$R^* = \frac{R_\mathrm{p}}{(E/p_0)^{1/3}} \tag{10}$$

The dimensional formula corresponding to Equation (10) is:

$$\frac{\mathrm{m}}{\left[\left(\frac{\mathrm{Nm \cdot m^2}}{\mathrm{N}}\right)^{1/3}\right]} = \frac{\mathrm{m}}{\mathrm{m}}$$

A diagram presented by BAKER and co-workers (Fig. 3) shows that these parameters make possible a relatively accurate dimensionless representation of the characteristic pressure-wave parameters, at least for many explosives [16].

For a provisional estimate, consider 1 t of a typical explosive, TNT (Δh_r = 4230 kJ/kg). The energy liberated in an explosion is:

$$E = (4230\mathrm{kJ/kg}) \times 1000\mathrm{kg} = 4.23 \times 10^6\,\mathrm{kJ}$$
$$= 4.23 \times 10^9\,\mathrm{Nm}$$

If p_0 = 100 kPa, Equation (10) yields:

$$R_\mathrm{p} = \left(\frac{E}{p_0}\right)^{1/3} R^* = \left(\frac{4.23 \times 10^9\,\mathrm{Nm}}{10^5\,\mathrm{N/m^2}}\right)^{1/3} R^*$$
$$R_\mathrm{p} = \left(42300\mathrm{m}^3\right)^{1/3} R^*$$
$$R_\mathrm{p} = (35\mathrm{m})\, R^* \tag{11}$$

Figure 3 yields the following dimensionless ranges:

Δp, kPa	$R*$
30	ca. 1.5
15	ca. 2
1	ca. 20

Thus, Equation (11) implies that an explosion corresponding to the detonation of 1 t of TNT would break windowpanes (Δp=1 kPa) out to a range of several hundred meters (R_p = (35 m)×$R*$ = 700 m) and cause serious damage (Δp = 15 kPa) in a circle radius 50 – 100 m (R_p = 70 m).

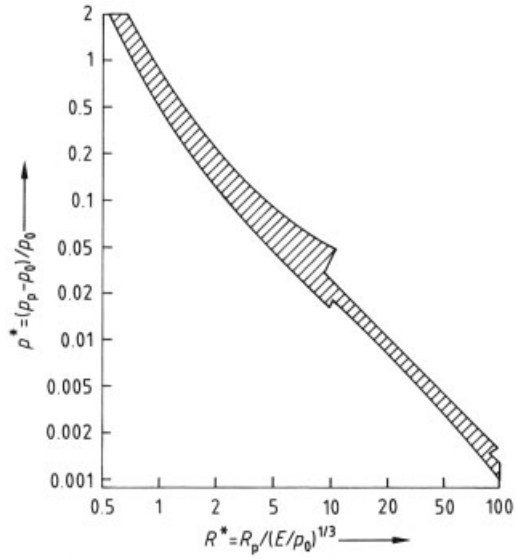

Figure 3. Dimensionless peak overpressure in an explosion pressure wave as a function of dimensionless distance from the source [16]
The expression for the dimensionless distance involves the explosion energy and the initial pressure

Explosives and similar materials are not commonly used in chemical plants. Reactions that can lead to explosion-like phenomena (Sections 1.1.2) are governed by other laws and differ strongly in kinetics, enthalpies of reaction, and efficiencies (in terms of energy conversion). For this reason, no direct comparison with the explosion just described is possible.

Even so, a number of workers have sought correlations [4], [18]. MARSHALL [1.3, p. 256], working with event analyses for explosions of unconfined gas clouds, arrives at the conclusion that for less than ca. 1 t of fuel gas forming an unconfined explosive cloud, peak overpressures

on ignition are negligibly low. He estimates that for an unconfined cloud with $> 10\,t$ fuel gas, about 5 % of the liberated energy (of combustion) can contribute to a TNT-equivalent pressure wave, but with significantly lower pressures ($< 100\,kPa$) in the near region.

Hence, 2 t fuel gas reacting to completion would produce the same effect as 1 t TNT (with allowance for the fact that the enthalpy of combustion of a fuel gas is a factor of 10 greater than the energy of TNT decomposition).

In a real unconfined gas cloud formed through damage to a vessel or in a plant, only a fraction of the fuel gas is ever in a region where the concentration is high enough for sudden combustion – in the worst case. It must therefore be supposed that the effect produced by the explosion of 1 t TNT can be matched only by the spontaneous release and ignition of ca. 10 t fuel gas. This could not happen without serious damage to a large vessel.

The hazard potential of an explosion is directly proportional to the area within which a certain level of damage occurs. According to Equations (8) or (10) and (7):

$$HP \sim R^2 \sim E^{2/3} = (M\,\Delta h_r)^{2/3} \tag{13}$$

i.e., the hazard potential increase is less than directly proportional to the energy release and the mass involved.

1.5. Conclusions

The analyses given here show, as does experience, that chemical plants may occasionally represent high hazard potentials, especially when volatile substances of a flammable or health-endangering nature are processed in large quantities, and when exothermic (decomposition and combustion) reactions are possible.

It is also clear that, while hazards can be identified and analyzed, the magnitude of hazard potentials cannot be quantified precisely. For the design of safe processes and plants, a qualitative order-of-magnitude estimate, a relative comparison with other hazard potentials, and an identification of key independent variables is sufficient, because such considerations give a point of reference such that hazard potentials can be reduced in the most efficient way, and safety measures and equipment can be designed to handle the remaining potentials (Chap. 3).

Experience has led to the low risk values cited for chemical plants and shown that major chemical plant accidents have been infrequent, especially in view of the scope of industrial activity in this field. The advanced chemical industry has obviously had much success in controlling its hazard potentials.

2. Determination and Evaluation of Hazardous Properties of Substances (Safety Ratings)

2.1. Harmful Effects of Substances

According to the EC classification criteria [20], substances are described in terms of 15 hazard-related characteristics, which can be broken down into four groups:

1) *Acute Toxicity:*
   Very toxic substances
   Toxic substances
   Harmful substances
   Corrosive substances
   Irritant substances
2) *Specific Toxic Properties:*
   Sensitizing substances
   Carcinogens
   Substances with effects on reproduction
   Substances with heritable effects
3) *Physicochemical Properties:*
   Extremely flammable substances
   Highly flammable substances
   Flammable substances
   Oxidizing substances
   Explosive substances
4) *Environmental Impacts:*
   Substances harmful to the environment

While acute and specific toxic properties can cause direct harm to health, physicochemical properties lead to indirect harm. Substances can take any of three routes into the body:

Oral: by mouth directly into the stomach
Dermal: through the skin
Inhalation: via the respiratory organs

The actions of a substance can differ greatly depending on the route of exposure. Because of the acidic conditions in some regions of the gastrointestinal tract (pH 1 – 5), hydrolyzable substances can be broken down after oral intake. Chemical reactions can lead to both more toxic substances (metabolic activation) and less toxic ones (detoxification). While acidic compounds are preferentially absorbed in the stomach, basic and lipophilic substances are preferentially taken up in the intestinal tract. Chemicals not resorbed in either the stomach or the intestines can be excreted directly without realizing their toxic potential (e.g., mercury, barium sulfate).

The skin protects the body against external effects. It performs very effectively against substances readily soluble in water. Fat-soluble (lipophilic) substances, on the other hand, commonly diffuse well through the skin into the body. Dermal resorption is extremely effective for polar molecules with lipophilic and hydrophilic groups (e.g., many organic solvents such as dimethyl formamide, dimethyl sulfoxide, ethyl acetate, toluene, chloroform). Corrosive chemicals resorbed through the skin (e.g., phenol) are taken up very quickly and effectively by this route.

Highly corrosive substances are those that destroy the skin (necrosis) if allowed to act for a maximum of 3 min; corrosive substances have the same effect within 4 h. Examples of highly corrosive substances are inorganic acids and alkalies. In general, the corrosive action of alkaline substances is stronger than that of acids.

Irritant substances are characterized by inflammation of the skin after a maximum of 4 h exposure (e.g., dilute acids and alkalies, weak organic acids, acrylates, alcohols, amines). While the irritant action of substances is reversible, corrosive action implies irreversible injury. The mucosa of the eyes and respiratory passages are particularly sensitive to irritation.

Highly water-soluble substances, when taken up by inhalation, are generally absorbed by the mucosa in the upper part of the trachea, and do not reach the deeper layers of the lung. Because many receptors are present, irritation reactions such as coughing and sneezing may be initiated by agents such as ammonia, hydrogen chloride, sulfur dioxide, and vapors of acids and alkalies.

Compounds that are less water-soluble can readily penetrate into the bronchi. Partial diffusion through the thin bronchial tissue is possible and can lead to lung injury (e.g., chlorine, bromine, iodine, ozone, phosphorus chlorides).

If the tissue of the lung suffers local corrosion, liquid can seep into the alveoli (pulmonary edema) and the oxygen transfer required for life can be hindered for hours to days after exposure (latent action; e.g., phosgene, nitrogen dioxide, ozone, some isocyanates). Lipophilic compounds, in particular, can cross over via the alveoli when respired air is exchanged with the blood.

Substances with systemic action, after inhalation or oral or dermal exposure, are distributed throughout the body by the circulatory system. In this way they can reach all organs.

If injury takes place only at the point of exposure, the substance is said to have a local action.

With regard to acute toxic effects, substances are classed as very toxic, toxic, or harmful. This description is based on the median lethal dose LD_{50}: the dose at which half of the experimental animals die. Toxicity is reported as oral, dermal, or by inhalation. Oral and dermal LD_{50} values are given in milligrams of the substance per kilogram of the experimental animal for one-time administration; LD_{50} values by inhalation are given in milligrams of the substance per liter of respired air (exposure time 4 h). Because oral toxicity values are more common, they alone are discussed here.

The lethal dose represents only the end point of toxic effects. The toxic action, however, begins to appear even when much smaller quantities are present in the body. The dose at which no biologically relevant effect can be detected is called the no adverse effect level (NOAEL).

Table 1 shows how substances are classified in accordance with EU guidelines [20]. Substances are arranged by acute toxicity properties; the table also gives danger symbols and indications along with the "R phrases" (which specify the principal risk more closely). In some non-European countries, the boundaries between classes differ slightly from the values shown.

Table 1. LD_{50} values (oral)

Attribute	LD_{50}, mg/kg	Symbol	R phrases
Very toxic	<25	T+	26, 27, 28
Toxic	25 – 200	T	23, 24, 25
Harmful	200 – 2000	Xn	20, 21, 22

Substances with LD_{50} (oral) values over 2 g/kg do not, according to past experience, require special precautions; these are no longer classed as hazardous substances.

Sensitizing substances are those that evoke an allergic reaction in a large number of individuals, either on the skin (skin or contact allergy; e.g., formaldehyde, glutaraldehyde, alkyl acrylate) or in the respiratory passages (respiratory allergy; e.g., phthalic anhydride, isocyanates). Sensitizing reactions of chemicals cannot be distinguished from allergic reactions to agents such as flour, pollen, and animal hair.

Substances that threaten the ability to reproduce are termed reproductive toxins or substances with effects on reproduction. Such effects are classified as:

Developmental Injuries (symbol R_E). These effects include birth defects (teratogenic effects) and embryo- or fetotoxic effects such as low weight and growth or developmental disorders of the unborn child.

Impaired Fertility (symbol R_F). The EU classification criteria [20] divide substances with effects on reproduction into three categories:

Category 1. Substances known to have effects on reproduction in humans.

Category 2. Substances that should be regarded as having effects on reproduction on the basis of animal studies.

Category 3. Substances suspected of having effects on reproduction.

The labeling requirements in countries of the European Union are summarized in Table 2 [20].

Table 2. Classification of reproductively toxic substances according to the EU guidelines [20]

	Category 1	Category 2	Category 3
Hazard symbol	T	T	Xn
Classification	toxic	toxic	harmful
R phrase	60 (R_F), 61 (R_E)	60 (R_F), 61 (R_E)	62 (R_F), 63 (R_E)

Some countries have regulations different from the EU scheme. For example, the MAK Commission in Germany classifies substances with respect to the MAK value (*maximale Arbeitsplatzkonzentration* = maximum occupational exposure concentration) [21]:

Pregnancy group A. A risk of fetal injury has been demonstrated. Injury cannot be ruled out, even if the MAK value is complied with.

Pregnancy group B. A risk of fetal injury cannot be ruled out, even if the MAK value is complied with.

Pregnancy group C. A risk of fetal injury need not be feared if the MAK value is complied with.

If a health hazard is posed, not by one-time exposure to the substance, but by quantities administered over a longer period, the terms subacute, subchronic, and chronic are employed, depending on the exposure time. Carcinogenicity is merely one specific chronic effect.

Carcinogenic substances can initiate tumors. A tumor is an uncontrolled cell growth that penetrates surrounding cells in infiltrative fashion and destroys them. The latency period (time between exposure and externally detectable cancer growth) is typically 20 – 40 a for chemical carcinogens. According to the EU criteria, carcinogenic substances are classified as follows [20]:

Category 1. Known to be carcinogenic in humans.

Category 2. Unambiguously carcinogenic in animal trials.

Category 3. Suspected of being carcinogenic.

Labeling requirements under the EU scheme [20] are listed in Table 3.

Table 3. Classification of carcinogenic substances according to the EU guidelines [20]

	Category 1	Category 2	Category 3
Hazard symbol	T	T	Xn
Classification	toxic	toxic	harmful
R phrase	45 (gases, vapors) 49 (dusts)	45 (gases, vapors) 49 (dusts)	40

The best-known substances in Category 1 are asbestos, benzene, benzidine, β-naphtylamine, and vinyl chloride; in Category 2, acrylonitrile, butadiene, and *N*-nitrosamines.

The classification of the German MAK Commission [21], [261] is very similar to the EU scheme (Table 4).

Table 4. Comparison of classifications of carcinogenic substances

	MAK Commission	EU Classification
Letter codes	K	C
Human carcinogens	1	Category 1
Animal carcinogens	2	Category 2
Suspected carcinogens (classification is temporary)	B	Category 3
Carcinogens without genetoxic effect	4	–
Carcinogens with genetoxic effect	5	–

MAK = maximum workplace concentration (Germany), *GefStoffV* = hazardous materials regulation

From the standpoint of toxicological action, substances causing heritable genetic damage do not differ fundamentally from carcinogens; the injury they cause, however, is to the genetic information of the germ cells rather than to that of the body cells. The classification again includes Categories 1 – 3. No Category 1 chemicals causing heritable genetic damage are known.

The classification of hazardous substances on the basis of thermal properties employs the flash point (Table 5). Substances are characterized as oxidizing if they can sustain combustion with the exclusion of oxygen. The symbol O is used in their labeling (e.g., permanganates, bromates, hypochlorites, chlorites, nitrates).

Table 5. Thermal properties according to EC guidelines [20]

Property	Flashpoint, °C
Extremely flammable∗	< 0
Highly flammable	< 21
Flammable	21

∗ *bp* must be < 35 °C.

Substances with a greater explosive tendency than dinitrobenzene (reference substance) are described as explosive, and must be labeled with the symbol E (e.g., picrates, pentaerythritol tetranitrate, many organic peroxides).

In order to assess the workplace, the concentration of hazardous substances must be determined and compared with the applicable workplace concentration limits (see § 18 of the *German Hazardous Materials Regulation* [22]). The most important limits for workplace air are the German MAKs [21] and the American TLVs (threshold limit values) [23]. If these limits are maintained, no damage to health is to be feared, given a daily exposure time of 8 h and a weekly exposure time of 40 h for healthy adult workers over a working lifetime (50 a). The OEL values (occupational exposure levels) set by the EU Commission [24] are also determined primarily on scientific grounds. Because no health hazard is present if these limits are satisfied, they are less than the NOAELs mentioned earlier. The limits cited are defined as the maximum permissible concentrations averaged over one working day (shift averages). For this reason, exposure peaks must be limited. Substances are classified in exposure peak groups according to their toxicological action profiles. According to, say, the provisions of the German MAK Commission, maximum exposure levels are 2 – 10 times the MAK. Substance-specific STELs (short-term exposure levels) are also established in the United States. For a few substances, "ceiling values" are also stated; it is prohibited to exceed these at any time [23].

Because NOAELs cannot be established for carcinogens and substances with heritable effects, some countries have not established medically and toxicologically justified limits for them. In Germany, for example, TRK values (*Technische Richtkonzentration* = technical guideline concentration) can be set for carcinogens, instead of MAKs [25]. A TRK represents the exposure level that can be reached at the state of the art. If the TRK is observed, acute health hazards are eliminated; the cancer risk is minimized, but not completely eliminated. These TRKs are also 8h or shift averages.

For important industrial chemicals not governed by limits, the chemical industry has created internal guidelines on its own initiative. The results are called PELs (permissible exposure limits) in the United States and ARWs (*Arbeitsplatzrichtwerte* = workplace guideline values) in Germany [26]. All workplace exposure levels are now summarized in Germany in a single guideline [27].

With regard to protective practices, technical measures always take priority over personal protective measures (see, e.g., § 19 of the *GefStoffV* [22]). Only when the use of enclosed apparatus is not reasonable can other measures consistent with the protective objective be used, e.g., the application of suction at the source in sampling, filling, and draining operations. The use of respiratory protection is permitted only temporarily,

for short-term activities, and employment as a long-term measure is not permitted.

In order to prevent dermal intoxication, suitable body protection must always be worn when working with chemicals. Polar compounds, in particular, commonly diffuse very rapidly through ordinary glove materials; breakdown times reported by manufacturers must always be observed. For further details, see [28].

In the preparation of the safety analysis pursuant to the *Störfallverordnung* (The Seveso Directive of the EU), the hazards due to product release in an emergency must be determined. With the aid of generally accepted propagation calculations (e.g., VDI 3783), the concentration in the region near a plant is calculated on the basis of a plausible leakage rate. "Accident rating values" are employed in assessing the health risk. These limits, worked out by the German chemical industry, represent concentrations such that there is no immediate danger to life, given a maximum exposure time of 0.5 h. Temporary impairment of health cannot, however, be completely ruled out.

2.2. Ratings of Flammable and Explosive Substances

2.2.1. Flammability Ratings

2.2.1.1. Introduction

A variety of data and experimental techniques can be used to characterize flammability and combustibility. Requirements as to fire behavior and fire resistance apply not just to chemicals themselves, but also to the materials and structures used in buildings where chemicals are handled.

2.2.1.2. Combustibility/Flammability of Chemicals

Combustion is a reaction that takes place in the gas phase and involves oxygen or another oxidizing agent present in the ambient gas phase. The flame phenomenon is characteristic of the reaction for liquids, gases, and most solids; constituents previously vaporized or carbonized and mixed with oxygen by diffusion oxidize in the flame. Any residue in, e.g., coke-like form, is heated to incandescence in this process and then reacts with oxygen present at the surface of the solid and diffusing into its pores. Substances such as pure carbon or iron, which do not vaporize sufficiently at the temperatures under consideration, react by glowing [29].

The speed of a self-sustaining combustion process depends on the rate of reaction, the heat liberated in the reaction, and the balance of reaction heat, heat loss to the surroundings, and heat required for melting/vaporization.

Combustibility of Liquids. The safety parameters flash point and ignition point are used to characterize the flammability of liquids.

The flash point is the temperature of a liquid, tested in a certain apparatus and under certain conditions, at which the gas phase above the liquid can burst into flame when an ignition source is brought near. The flash point is measured in various test methods [30], [31], e.g., DIN 51 755, DIN 51 758, and DIN 53 213 for Germany; GOST 12.1.044-84 for the Russian Federation; BS 2000 Part 34 and Part 170 for the United Kingdom; and ASTM D 56-87, ASTM D 3278-89, and ASTM D 93-90 for the United States. All test setups and specifications allow for the fact that the flash point is a rough measure because of the tolerated uncertainty of measurement and the variation of material properties (e.g., viscosity).

The two basic types of apparatus are the open and closed cups. Methods differ in the access of oxygen and in the quality of mixing or accumulation of fuel over the liquid. Methods can also be classified as thermodynamic equilibrium or nonequilibrium, and a given test substance may show different flash points in these two kinds of test.

In spite of national and international differences, all the methods yield data sufficiently reliable for safety assessments.

Figure 4 shows one setup for flash point determination.

If the temperature of a liquid exceeds the flash point, flammable mixtures with air may form over the liquid. As a rule, combustion of this mixture does not result in self-sustaining combustion of the liquid; this is possible only when its temperature is raised (by a few degrees celsius) to the combustion point. Even then, it

is necessary that the temperature rise does not cause any further inert constituents (e.g., water in spirits of very low alcohol content) to go into the gas phase.

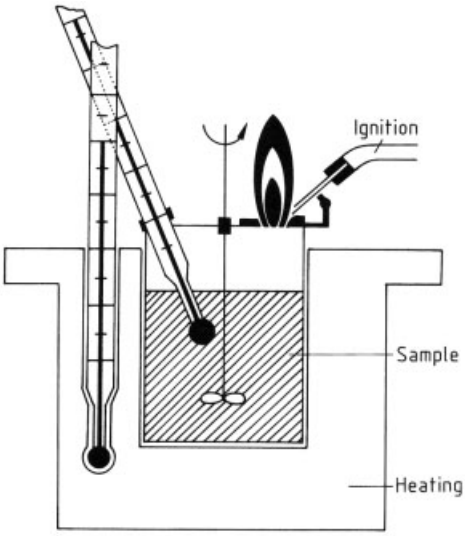

Figure 4. Flash point test equipment
Examples of flash point values (°C): Diethyl ether, gasoline − 20; ethanol 12; rum (40 vol %) − 27; diesel fuel 55; glycerine 160

Internationally, the flash point is the basis for placing substances in hazard classes with regard to ignitability. Under international transportation law [30], a liquid is classed as flammable if it has a flash point $< 60.5\,°C$ (closed cup) or 65.6 °C (open cup) or if it is handled at a temperature higher than the flash point. Self-sustaining combustion is tested in a separate procedure [32]. If a substance with flash point $> 35\,°C$ does not continue to burn after ignition in this test, it is not placed in Class 3 (flammable liquids); this rule takes into account the reduced risk of fire.

According to the German regulations (e.g., the VbF (Verordnung brennbarer Flüssigkeiten) or Flammable Liquids Regulation [31]), a substance having flash point $\leq 100\,°C$ is classified as combustible. Liquids miscible (extinguishable) with water make up Hazard Class B; those not miscible with water, Hazard Class A. Details of the classification are as follows:

Hazard Class A
Liquids with flash point $\leq 100\,°C$ which do not have the water solubility characteristics of Hazard Class B

A I: flash point $< 21\,°C$
A II: flash point 21 – 55 °C
A III: flash point 55 – 100 °C

Hazard Class B
Liquids with flash point $< 21\,°C$ which are soluble in water at 15 °C or whose combustible components are soluble in water at 15 °C

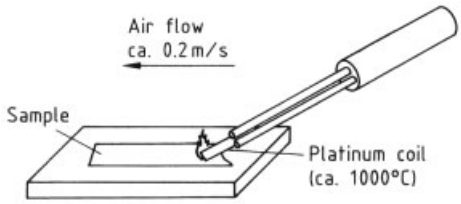

Type of reaction	Class	Reference product
No spreading of fire		
no ignition	BZ 1	table salt
brief ignition and rapid extinction	BZ 2	tartaric acid
localized combustion or glowing with practically no spreading	BZ 3	D-lactose
Fire spreads		
glowing without sparks (smoldering) or slow decomposition without flame	BZ 4	protonic acids; tobacco
burning with flame or spark generation	BZ 5	sulfur
very rapid combustion with flame propagation or rapid decomposition without flame	BZ 6	black powder

Figure 5. Flammability index test equipment

Combustibility of Solids (Dusts). The combustibility/flammability of solids is characterized by the flammability index [33] and the burning rate [33], [34] of an uncompacted layer of dust. The flammability index gives a qualitative description of how such a bed of dust behaves when a strong ignition source (igniting flame, incandescent coil) is applied to it (Fig. 5). A substance with flammability index > 3 tends to keep burning after local ignition; a burning-rate test must be performed on such a substance. For flammability index ≤ 3, the substance generally shows no tendency to rapid flame propagation. On the basis of a burning-rate test (flame path 100 mm, time limit 45 s),

solids are classified as highly flammable or not highly flammable, but these terms are misleading because they both relate to the rate at which a fire spreads once started, not the tendency of a substance to catch fire.

Beside ignition from an outside source, another concern with dusts is the spontaneous flammability in air or other gaseous oxidizing medium due to self-heating. VDI 2263 describes the Grewer accelerated test (Fig. 6) and the heat-aging tests used to characterize the spontaneous flammability of dusts. Volumes of the dust ranging from 5 to 1000 mL are subjected to a programmed variation in temperature in the presence of (usually flowing) oxidant, and the behavior of the dust sample is recorded. Maximum permissible storage and surface temperatures can be derived from the results with the aid of mathematical models and appropriate safety margins. Similar test methods, with other volumes of dust tested in other containers (square, cylindrical), are used in Transportation Class 4.2 testing under international transportation law [30], and in spontaneous flammability testing pursuant to the *Chemikaliengesetz* (Chemicals Act) [34].

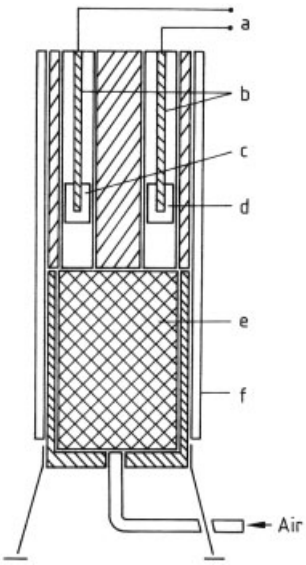

Figure 6. Grewer test apparatus for testing spontaneous flammability of dusts (VDE 2263)
a) Temperature recording; b) Thermocouples; c) Test substance (8-cm^3 wire basket); d) Reference substance; e) Packing; f) Temperature-controlled housing

Applicability of Flammability Data. The combustibility measures described up to this point are based on standard test methods. The data cannot be extended to practical conditions unless possible variations in the ambient conditions are taken into account. The flash point, for example, depends heavily on pressure as well as oxidizing medium. Changes in these lead to changes in fire behavior. The temperature of spontaneous ignition can also fall markedly below the ignition temperature of DIN 51 794 if organic liquids (including heat-transfer media) penetrate into open-pored insulants such as rock wool.

Assessment of dusts is likewise based on the experimental conditions (air as oxidizing medium, adequate heat removal from combustion zone). If the partial pressure of oxygen is increased, a stronger oxidizing agent is employed, the heat-removal conditions are made worse (storage in large drums, silos, etc.), or the ambient temperature is higher, an increased risk of fire must be expected.

In the case of solids or ointments containing solvents, it has to be decided case by case whether the flash point or the burning rate/flammability index better describes the situation.

2.2.1.3. Oxidizing Properties of Chemicals

Oxidizing substances in contact with combustible substances, give off oxygen or other oxidizing agents so as to increase the danger of fire or the vigor with which the combustibles burn. As a rule, oxidizing substances are not themselves combustible. In the pan-European market area, testing is done pursuant to Directive 84/449/EEC Annex A 17 [34]. Various mixtures of the test substance with cellulose powder are prepared and the burning rates are determined; the results are compared with the burning rate of a reference mixture of barium nitrate and cellulose. A substance is classified as oxidizing if the burning rate of a test mixture is greater than that of the reference mixture. As a rule, incorrect predictions are made if this test procedure is applied to liquids or to melting and highly flammable (combustible) substances, so its use with these substances is ruled out. A comparable test method is employed internationally for the

transportation of hazardous substances [30]. In some instances, the two procedures give different classifications for the same substance.

Oxidizing properties make a chemical a hazardous substance, and safety regulations apply to its storage and transportation.

Table 6. Classification and test method for building materials [35]

Building material class	Building inspection designation	Test method
Class A	noncombustible	
A 1 ∗		furnace test 750 °C
A 2		Brandschacht
		smoke density to ASTM D 2843-70 and DIN E 53 436/37
		toxicity to DIN 53436
		calorific potential to DIN 51 900 Part 2 and heat release to DIN 4102
		Part 8 or furnace test 750 °C
Class B	combustible	
B 1 ∗∗	low flammability	Brandschacht and small burner test
		special case of floor coverings, radiant panel test (NBSIR 75-950) including evaluation of smoke density
B 2	moderately flammable	small burner test
		special case of textile floor coverings (DIN 54 332)
B 3	highly flammable	no tests, no compliance with *B 2*

∗ *Class A 2* requirements must also be satisfied.
∗∗ *Class A 2* requirements must also be satisfied.

2.2.1.4. Flammability of Construction Materials

For fire precautions to be effective, they must take into account not only the combustibility of chemicals, but also the fire behavior of construction materials and structures, floor coverings, and other items within the production facility.

In Germany, structures and construction materials are tested as set forth in DIN 4102 [35]. Evaluated along with the fire endurance are the flammability, the burning rate, the toxicity of the combustion gases, and (for structures) the mechanical strength. The fire endurance is stated in minutes, for example F 30, F 60, F 90. DIN 4102 requires a fire endurance of F 90 for firewalls. Construction materials are classified as combustible or noncombustible according to the test results (Table 6).

Outside the Federal Republic of Germany, standards for assessing the combustibility of construction materials and other substances vary from country to country. A comprehensive survey of national regulations can be found in [35].

2.2.2. Explosion Data for Gas Mixtures

(General References: [36–40])

2.2.2.1. Introduction

In any operation on combustible and/or unstable gases, explosive systems may be formed and explosions initiated. This danger can be effectively countered only if the safety characteristics of the gases involved are well known. These characteristics are not constants in the physical sense, for they are influenced by the method used to determine them. There is no generally accepted method for estimating and calculating them, although for a few characteristics empirical formulas exist for specific groups of substances [41]. In other words, recourse must be had to the experimental methods described in the following sections. Laboratory values must be of such a quality that they can serve as a basis for explosion-protection measures in full-scale plants. This is possible only if the method of determination is standardized or meets other requirements derived from experience. The link between parameters and explosion-protection measures must be taken into account, even in bench-scale determination; this link essentially has to do with the type of plant and the process parameters, regardless of whether explosion protection is achieved through precautions or constructional measures.

2.2.2.2. Explosion Limits; Limiting Oxygen Concentration

A gas explosion is a rapid chemical reaction (oxidation) in the gas phase, which propagates through the explosive mixture in a self-sustaining process. Because the reaction proceeds rapidly, the heat released leads rapidly to high temperatures, and can result in the buildup of high pressures, especially in enclosed vessels.

A gas explosion requires the presence of an explosive mixture consisting of a finely dispersed fuel in an oxidizer (oxidizing medium) and the simultaneous presence of an effective ignition source. Fuels include all combustible gases (e.g., hydrogen, methane, carbon monoxide, ethylene) as well as combustible vapors such as those of the volatile hydrocarbons. The oxidizer is commonly atmospheric oxygen, but there are other oxidizing gases, the most important being chlorine, fluorine, ozone, nitrous oxide, nitric oxide, and nitrogen dioxide. Inert gases such as nitrogen and carbon dioxide play an indirect role in explosion reactions.

Any gaseous system containing mixed fuel and oxidizer is explosive only within a certain concentration range, the "explosion range", governed by the properties of the substances involved and by pressure and temperature. In practice, there is often only one fuel and air is the oxidizer; here the explosion range can be characterized simply by stating the concentration limits of the fuel in air. These bounds, the lower explosion limit (LEL) and the upper explosion limit (UEL) have been tabulated for a number of combustible gases and vapors [42], [43].

Nonexplosive mixtures in the range below the LEL are said to be "too lean" and those above the UEL are said to be "too rich."

The explosion limit itself is defined quite generally as the fuel concentration such that an explosive reaction initiated by an outside source just fails to propagate through the mixture in a self-sustained fashion. In practice, the explosion limit is recognized by the fact that the visible flame (present in virtually all gas explosions and directly tied to the reaction zone) just fails to detach from the ignition source; in other words, the velocity of flame propagation becomes zero. This flame propagation velocity or combustion rate can take values from a few centimeters per second up to several kilometers per second in the case of detonations. It is not an important parameter, except in certain technical systems with precisely defined boundary conditions (e.g., the internal combustion engine).

One basic explosion-protection measure is to prevent the formation of explosive mixtures, i.e., to keep the mixture composition outside the explosion range (see also Section 3.4.2). This requires an accurate knowledge of the explosion limits for the system.

Any apparatus for determining explosion limits must satisfy definite minimum requirements, since the limits relate not just to thermal conditions, but also to reaction kinetics.

A set-up for measuring the explosion limits of gases should therefore comprise a heated cylindrical glass vessel (ca. 1 L) with inner diameter ≥ 6 cm, set with its axis vertical, and an ignition source located on-axis in the bottom third (Fig. 7). The boundary of the ignition vessel is thus sufficiently far from the ignition point and has only a slight deactivating (explosion-hindering) action; the propagation or detachment of the flame can be observed visually and is not hindered by the burned gases, which are free to escape upward.

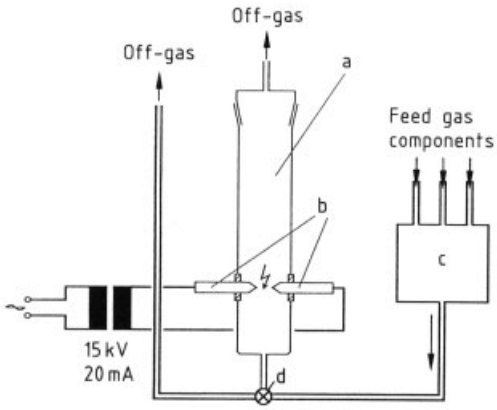

Figure 7. Apparatus for determining the explosion limits of gases
a) Ignition vessel; b) Ignition electrodes; c) Mixture preparation; d) Three-way valve

The criterion for a self-sustaining progressive combustion reaction (explosion) is the visually observed detachment of a flame from the ignition source (Fig. 8, [44]). Each ignition trial must be done in a stationary mixture, since movement (e.g., flow) of the unburned mixture, especially at the explosion limits, leads to additional transport of heat and free radicals out of the ignition volume, and this can interfere with the formation of the explosion reaction.

A proven way of delivering energy to initiate the reaction is with an induction spark (e.g., that generated with a 15 kV transformer) lasting ca. 0.2 – 0.5 s; the spark jumps a gap of 5 mm between two pointed rod electrodes ca. 3 mm diameter.

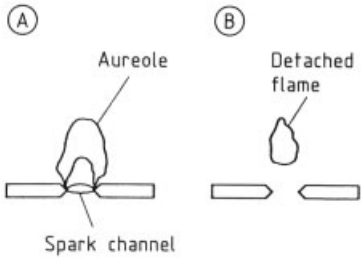

Figure 8. Flame detachment criterion
A) No flame detachment; B) Flame detachment

To determine the explosion limits accurately, the preparation of the mixture and its transfer into the test vessel must introduce as little error as possible. Calibrated devices must be used to meter the volume flow rates (of gases) and mass flow rates (of totally vaporized liquid mixtures). Temperature control, especially of the ignition vessel, should be as gradient-free as possible.

Gas-metering pumps with volumetric delivery have been found satisfactory for preparing pure gas mixtures. The test vessel must be adequately purged with the test gas mixture before every ignition trial, to remove burned or unburned residues and so that the vessel wall is sufficiently conditioned. To insure constant mixture composition, it is useful to avoid interrupting the flow of mixture during the trial; a three-way valve can be used to bypass the mixture around the vessel.

In pure gas mixtures, the explosion limits are generally stated in percentage by volume of fuel gas relative to the total mixture (vol %). This is appropriate when the mixture is prepared with a metering device with volumetric delivery. Otherwise, the limits should be based on the molar ratio (mol %).

If care is taken in operating the mixture preparation and explosion devices, explosion limits can be determined to an accuracy of ca. ± 0.1 vol % [44].

When the UEL is being determined, it should be kept in mind that some substances, such as acetylene and ethylene oxide, can decompose explosively without the presence of oxidizer; this can result in an observable flame detachment or pressure rise not attributable to an oxidative gas explosion, and it may not be possible to state an UEL. Because the effects of explosive decomposition (e.g., the maximum pressure) may actually surpass those of an oxidative gas explosion, a UEL of 100 vol % is reported and the decomposability of the substance is noted [42].

In the case of an explosive gas mixture containing several fuel species, if the concentrations and the explosion limits of the individual fuels are known, the explosion limit of the fuel mixture can be calculated, provided certain mixing rules (e.g., that of Le Chatelier [45]) are known to hold (see also [46]). As a rule, this gives satisfactory results only for the LEL.

If explosion limits must be determined for an elevated initial pressure, the tests must be carried out in a closed, pressure-tight apparatus, such as that used for measuring the maximum explosion pressure and the maximum rate of pressure rise (see Section 2.2.2.3).

Along with tables of LELs and UELs for combustible gases and vapors in air [42], [43], a triangular plot (explosion diagram) can also be used for systems of three or more components. The triangular plot displays the LEL, UEL, limiting oxygen concentration $[O_{2max}]$, and characteristic concentration ratios needed to avoid explosive mixtures. Figure 9 presents such a diagram for a hydrocarbon/nitrogen/oxygen system; the explosive region is shaded. Each concentration (fuel on the base of the triangle, inert gas on the right leg, oxygen on the left leg) is shown increasing in the counterclockwise direction. In systems with more components, those whose concentration ratio is held constant can be plotted on one axis; for example, air ($[N_2]/[O_2]=3.76$) can be plotted as one species if CO_2 is introduced as an additional (inert) component. The so-called air line, joining 0 vol % N_2 (point B) and 21 vol % O_2 connects all mixture compositions in which the concentration ratio $[N_2]/[O_2]$ is constant at 3.76. Points L and U where the boundary of the explosive region intersects this line give the lower and upper explosion limits in air. The tangent to the explosive region boundary parallel to the N_2-axis cuts the O_2-axis at point D, which defines an O_2 concentration below which no explosion can take place because of the deficiency of oxygen, regardless of how much fuel or N_2 is present (partial inertization). This O_2 concentration (point D) is called the limiting oxygen concentration $[O_{2max}]$; it is a key parameter in the design of inert-gas blanketing (including that for plant-scale equipment), and the oxygen level must

never be permitted to exceed it. A detailed explanation of the triangular diagram has appeared in [47]. Limiting oxygen concentrations for some substances are listed in [48].

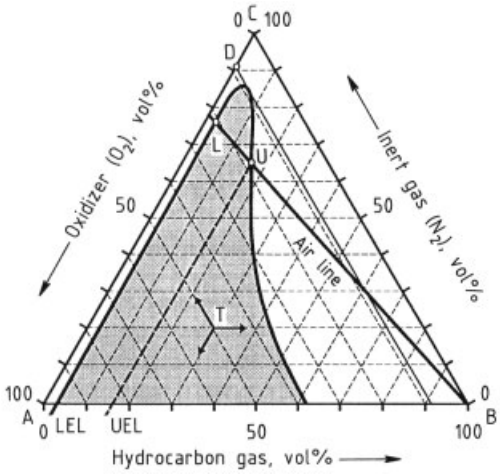

Figure 9. Triangular plot for a hydrocarbon – oxygen – nitrogen system

In practice, inerting is often best achieved by monitoring just one component, commonly the oxygen level at partial inertization, because ratio monitoring is generally more involved and expensive.

Since the explosion reaction depends on pressure and temperature, so do the explosion limits. Generally as pressure increases and temperature rises, the explosion range becomes larger. This change is much more strongly marked near the UEL than at the LEL, as Figure 10 illustrates for ethylene. The limiting oxygen concentration decreases as the pressure increases. The explosion limits and the limiting oxygen concentration vary similarly with temperature [49]. An exception is the pressure dependence of the explosion limits of hydrogen; with increasing pressure the explosion range for hydrogen does not become wider but narrower [50]. A similar behavior was found for carbon monoxide.

In the design of explosion-protection measures, it must be kept in mind that the safety-relevant parameters employed are those that apply to the conditions (pressure, temperature) of the particular application.

At present, explosion limits can be calculated only by fitting experimental data, because simulation models do not offer an adequate (substance-specific) description of the physicochemical conditions at the explosion limit.

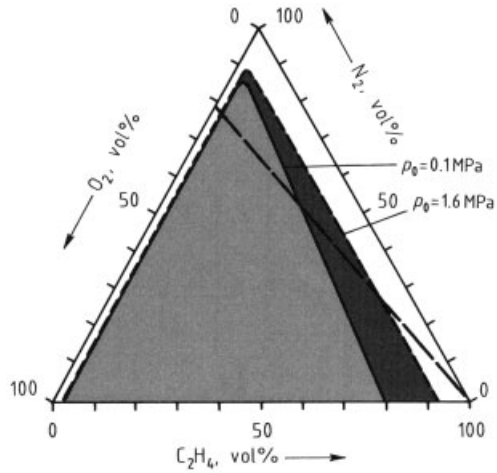

Figure 10. Pressure dependence of the explosion range

2.2.2.3. Maximum Explosion Pressure and Maximum Rate of Pressure Rise

In chemical process engineering, it is frequently not possible to avoid working with reactive gas systems in closed equipment. Avoiding potential ignition sources is often not sufficient to insure safe working conditions. The hazardous effects of an explosion on the equipment and its surroundings, including people, can be averted only through design measures [48]. The elimination of ignition sources is necessary, if only for economic reasons, as a supplementary measure. Before protective practices are planned, the following question must be elucidated: can the gas system become capable of detonation (occurrence of shock waves) as a result of the operating conditions (pressure, temperature, composition) or the geometry of the apparatus (long vessels, piping) ? If so, the discussion that follows does not apply. If the occurrence of detonations can be ruled out with confidence, it is possible to design vessels and piping such that they can withstand a deflagrative explosion. The principle of pressure-resistant or pressure-shock-resistant design is employed in such cases.

In pressure-resistant design, the anticipated explosion pressure must not exceed the design pressure; in shock-resistant design, the pressure is allowed to go above the design pressure, and

deformations are allowed, provided the system remains tight (deformed parts must be replaced after such an event).

For either of these design approaches, the maximum anticipated explosion pressure generated by the gas system under the design operating conditions must be known. In simple systems, a theoretical calculation is possible, but in other cases the maximum explosion pressure has to be determined experimentally.

If economics dictates that pressure-resistant or shock-resistant design not be used, pressure relief is still a design option for drastically reducing the pressure loads. The design of rupture disks and explosion doors requires knowledge of (among other parameters) the maximum rate of pressure rise in the anticipated explosion. This cannot be calculated theoretically and must be determined experimentally for every application [51–53].

When pressure relief is employed for protection – an approach calling for a high degree of experience – it must not be forgotten that the release of products into the atmosphere may infringe air pollution control regulations. Both the maximum explosion pressure $p_{\max}$ and the maximum rate of pressure rise $(\mathrm{d}p/\mathrm{d}t)_{\max}$, are measured in a single laboratory trial. There is no single prescribed method for determining these parameters. The requirement that bench-scale measurements must be suitable for the design of much larger equipment is best met by using a spherical autoclave with the ignition source at the center of the explosion vessel. This choice stems from the knowledge that the flame front of a gas explosion propagates through the mixture as a sphere, centered at the point of initiation [54]. To allow the reactions to proceed in virtually adiabatic fashion (i.e., to minimize heat loss through the autoclave wall in comparison with the total heat of reaction), the distance between the igniter and the wall must be ≥ 10 cm. Otherwise, excessive heat loss causes the measured values to be too low and yields poor estimates of the safety ratings [55]. An autoclave with a pressure rating of ca. 100 MPa at 200 °C covers the pressure and temperature ranges of practical interest in such studies. Proven types of initiators are wire igniters and surface-discharge plugs.

Energies, which depend not only on the electrical parameters, but also on the gas and the pressure, are 50 – 80 J [56].

The following instrumentation chain is recommended for recording the time dependence of pressure during reactions:

1) Pressure transducer with appropriate range; pressure-sensing membrane must make a tight seal in the autoclave wall
2) Amplifier adapted to the pressure pickup
3) Digital storage oscilloscope or computer with measuring card

For undistorted transmission of the pressure signal, the measuring chain must have a rise-time short in comparison with the pressure rise-time in the explosions under study.

Other internal features of the autoclave are a thermocouple to monitor the initial temperature and a mixer to homogenize the gas mixture. The mixer is dispensable if gas mixtures are prepared in a separate mixing vessel. Generally, the individual components are charged into the evacuated vessel in accordance with their partial pressures; the quantities are governed by the mixture composition and the initial pressure.

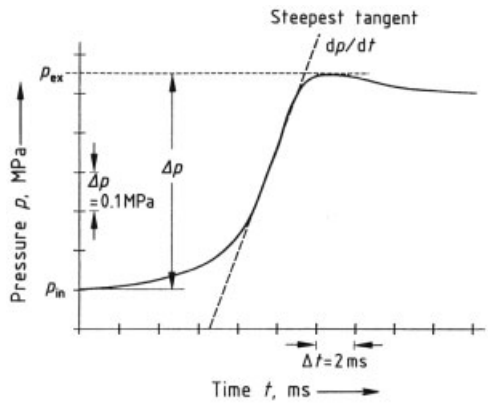

Figure 11. Time dependence of pressure in a gas explosion p_{in} = initial pressure; Δp = explosion overpressure; p_{ex} = explosion pressure; $\mathrm{d}p/\mathrm{d}t$ = rate of pressure rise

Initiation trials are performed in still gas mixtures (mixer turned off) with various fuel concentrations. Analysis of each separate run with a given fuel-gas content yields an explosion pressure value (maximum on the $p-t$ curve) and a rate of pressure rise (slope of the steepest tangent to this curve, Fig. 11). Analysis of the digitally stored curves can conveniently be done by computer. Values from individual trials, plotted against the fuel-gas content, yield two curves (Fig. 12) whose maxima are the param-

eters sought, p_{max} and $(dp/dt)_{max}$ for the system under study for the specified initial pressure and temperature conditions. The maximum explosion pressure thus obtained remains virtually unchanged when the volume is increased.

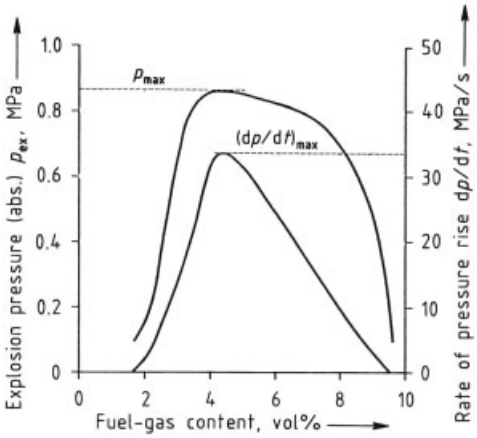

Figure 12. Illustrative curves of the explosion pressure and the rate of pressure rise for a fuel – air mixture, plotted versus the fuel-gas content
Tests took place in a still gas mixture in a 14-L spherical autoclave.

If the maximum pressure is divided by the initial pressure p_{in} the result is called the maximum pressure increase factor:

$$F_{max} = p_{max}/p_{in} \tag{14}$$

Typical values for explosive mixtures with air are $F_{max} = 5 - 10$. F_{max} is more or less independent of the initial pressure, but decreases with rising temperature.

The value found for the maximum rate of pressure rise cannot be extended directly to larger volumes. The only practical way to get an acceptable value for volumes larger than laboratory autoclaves is to apply the so- called cube law, but it must not be forgotten that the validity of the law is restricted; it can be used only for volumes that do not differ too greatly in their geometry. Such a transformation is not possible if the turbulence characteristics differ, e.g., when the laboratory value is determined in the still gas mixture, as it commonly is. This must be taken into consideration in practical situations, especially when the pressure is to be relieved with blast pipes; full-scale tests may be necessary.

According to the cubic law, the maximum rate of pressure rise multiplied by the cube root of the test volume is a constant K_G:

$$K_G = (dp/dt)_{max,V_L} V_L^{1/3} \tag{15}$$

where V_L is the volume of the laboratory autoclave. The maximum rate of pressure rise can be calculated for a different volume V:

$$(dp/dt)_{max,V} = K_G V^{-1/3} \tag{16}$$

The pressure increase factor and K_G value can now serve as a basis for analyzing design practices, regardless of the fuel content in the explosive mixture in the actual plant. (Note that the spherical autoclave can be replaced by a different ignition vessel if the new vessel has the same volume/area ratio as the sphere. For example, a cylindrical vessel can be employed if its diameter and height are equal.)

2.2.2.4. Pressure Limit of Stability for Unstable Gases

There are gaseous compounds whose thermodynamic and/or chemical properties enable them to decompose explosively, without any air or oxygen, if an ignition source is present. A thermodynamically unstable compound can be prepared by direct synthesis from the elements only if energy (specifically, the free enthalpy of formation) is supplied. If such a compound (e.g., acetylene, nitrous oxide, ethylene) decomposes to its elements, energy equal to the free enthalpy of formation is released [57]. Thermodynamically stable compounds with a negative free enthalpy of formation can also decompose to compounds of lower molecular mass if the sum of the free enthalpies of formation of the decomposition products is more negative than the free enthalpy of formation of the starting substance (e.g., ethylene oxide, tetrafluoroethylene). The features responsible for the ability of thermodynamically stable compounds to decompose are “weak bonds” between atoms in a molecule, where the molecule can dissociate if its vibrations are strongly excited (e.g., by absorption of external energy). Organic molecules with multiple bonds between carbon atoms are susceptible because of the diminished binding forces at these sites. Thus there are two factors, one thermodynamic and one structural that suggest

an ability to decompose. A gas with both kinds of instability (e.g., acetylene) is most likely to decompose.

Fortunately, an unstable gas does not decompose spontaneously. The free enthalpy of activation ΔG^{A}, must be supplied to initiate the reaction. Only when the energy threshold ΔG^{A} has been overcome (Fig. 13) the reaction can proceed, with the release of the free enthalpy of reaction ΔG_{r} [58]. Such reactions become dangerous if the enthalpy of reaction is evolved so rapidly that it can no longer be absorbed by the surroundings, but remains in the system in some form. The reaction products can be heated up so that the pressure rises (in the same way as in a gas explosion involving oxygen). The energy threshold ΔG^{A} is overcome more easily the higher the pressure and/or temperature of the gas. For safe handling of gases tend to decompose, it is urgently necessary to know the pressure and temperature limits such that the energy threshold cannot be exceeded, even by a strong ignition source. These limits can be determined only by experiment.

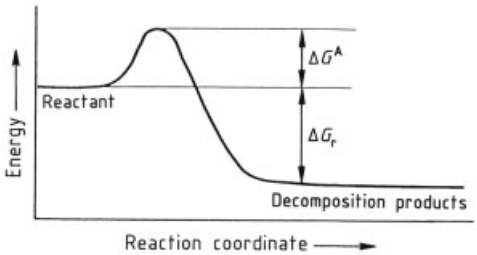

Figure 13. Energy diagram of a decomposition reaction

The apparatus described in Section 2.2.2.3 is well suited to studies of this kind. By performing a series of ignition trials, the researcher seeks to determine the pressure, as a function of temperature, at which it just becomes impossible to initiate decomposition. The result is the pressure limit of stability as a function of temperature. To provide at least approximate statistical support for the limit, the test is replicated five times. Figure 14 shows the pressure limit of stability for ethylene at 100 – 200 °C [59].

To deal with problems relating to the safe transportation of unstable gases, studies of the kind just described are run, up to 70 °C. Pressure limits of stability at even higher temperatures are of interest in view of operations in chemical plants. It is also often desirable to find out how the pressure limit of stability can be modified by the addition of a stable gas (not necessarily inert) or what amount of foreign gas must be added to render the system stable under the operating conditions. All these questions can be answered by experiments with the laboratory apparatus described.

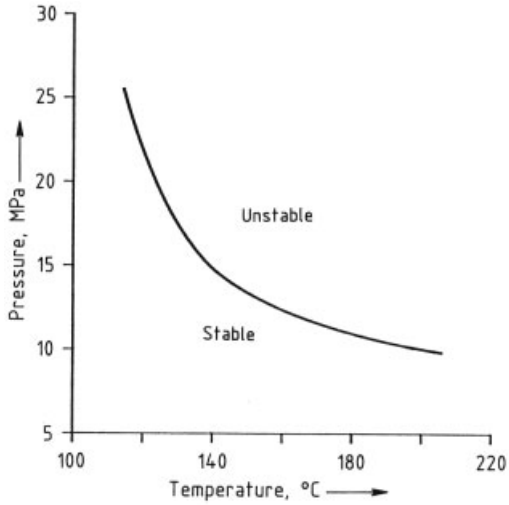

Figure 14. Pressure limit of stability for ethylene as a function of temperature, ignition power 1 kW

In practice, an adequate margin of safety relative to the pressure limit of stability must always be maintained when working with unstable gases.

2.2.2.5. Ignition Temperature, Ignition Energies

Explosive systems are characterized by certain parameters (LEL, UEL, $[\mathrm{O}_{2\max}]$, $p_{\max}$, $(\mathrm{d}p/\mathrm{d}t)_{\max}$, K_{G}) that can be employed in the design of explosion-protection measures. Two approaches to explosion protection are to prevent explosive mixtures from forming, and to insure safe handling of such mixtures by explosion-proof design with or without pressure relief. In a similar way, the sensitivity of these systems to ignition by certain ignition sources is described in terms of parameters that can be used in safety design (elimination of ignition sources).

It is also important in practice to find a parameter that characterizes fuels with regard to the propagation of explosions, for example, through

narrow gaps. The propagation of an explosion from a part of the plant featuring explosion-resistant design into upstream or downstream sections that are not so designed, or that are not vented, must be effectively prevented.

One such parameter is the ignition temperature, which makes it possible to assess the effectiveness of a hot surface (e.g., a motor winding) as an ignition source. Others are the minimum ignition energy, used for evaluating electrostatic (capacitive) discharges, and the minimum ignition current, for evaluating electrical (inductive) sparks. Finally, the ability of exploding gas mixtures to propagate through a narrow gap is characterized by the maximum experimental safe gap (MESG), which can be used in the design of flame arresters or "pressure-resistant" enclosed electrical equipment (see EN 50 018).

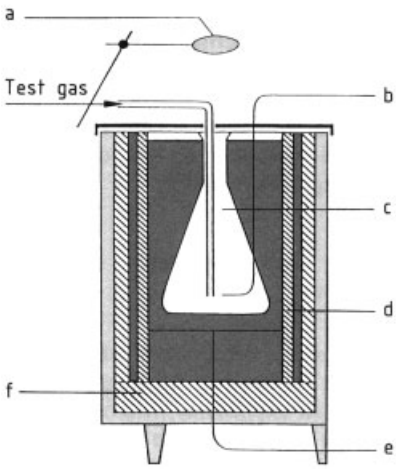

Figure 15. Apparatus for determining the ignition temperature (DIN 51 794)
a) Observation mirror; b) Thermocouple, FeCo; c) 200-mL Erlenmeyer flask; d) Heating; e) Resistance thermometer (Pt-100); f) Thermal insulation

The ignition temperature T_I is the lowest temperature of a hot surface at which a fuel – air mixture of the most easily ignitable composition can just be stimulated to burn with a visible flame. Two crucial points for determination of T_I are the geometry of the test vessel, which must provide a temperature field as closed as possible, and the formation of steep fuel concentration gradients when the fuel is admitted to the oxidizing gas phase already present in the vessel. The standard method of IEC 79-4 (embodied in DIN 51 794 and ASTM D 2155) takes these points into account. The glass ignition vessel (a 250-mL Erlenmeyer flask) is charged with air or another oxidizer. The fuel is admitted to the vessel through an inlet tube (for gaseous substances) or dropwise (for liquids; Fig. 15). The criterion for ignition is a flame forming in the vessel, which can be observed with mirrors.

On the basis of the ignition temperatures in air, fuels are placed in temperature classes (EN 50 014/5.78, DIN VDE 0165; Table 7). Ignition temperatures and temperature classes have been tabulated [42]. The maximum surface temperature T_{max} associated with a given temperature class must not be exceeded by explosion-resistant electrical (and other) equipment belonging to the class (EN 50 014, pr EN 1127).

Table 7. Temperature classes according to EN 50 014

Ignition temperature∗, °C	Temperature class	Maximum surface temperature, °C
$450 < T_I \leq 600$	T 1	450
$300 < T_I \leq 450$	T 2	300
$200 < T_I \leq 300$	T 3	200
$135 < T_I \leq 200$	T 4	135
$100 < T_I \leq 135$	T 5	100
$85 < T_I \leq 100$	T 6	85

∗ Measured according to IEC 79-4 (e.g., DIN 51 794).

When determining the ignition temperatures T_I of fuel mixtures, it must be kept in mind that there is no linear relationship between the concentration ratio (or mole fraction) of the fuels and the ignition temperature for the mixture [46], [60]. T_I values for fuel mixtures must therefore be determined for every individual case. Ignition temperatures for a given fuel in oxidizing media other than air may sometimes be quite low. This is especially so for pure oxygen and chlorine, in which T_I may be lowered by as much as hundreds of kelvins: T_I (toluene, air) = 535 °C, T_I (toluene, Cl_2) = 175 °C. This effect with pure O_2 can be accounted for by the lack of a diluting inert gas; in other cases, e.g., Cl_2, the ignition temperature is lower because the Cl_2 molecule has a much smaller dissociation energy than O_2. Also, T_I generally decreases with increasing when initial pressure [60].

The minimum ignition current, used as a measure of the igniting ability of electrical (inductive) sparks, is a fuel-specific parameter. It is the smallest current flowing in a circuit of a certain inductance, such that the spark produced

when the current flow is interrupted (in an explosive mixture of optimal ignitability) just ignites the mixture. If the minimum ignition current is known, electrical circuits can be designed so that no igniting sparks are created by switching, or on failures such as a short- circuit. This design approach is employed in explosion-resistant electrical equipment rated "intrinsically safe" (i; see EN 50 020).

The minimum ignition energy is the smallest amount of energy, stored in a capacitor, that is just sufficient, when discharged across a spark gap, to ignite the most ignitable explosive mixture. It is a fuel-specific parameter. As a measure of the effectiveness of capacitive sparks to ignite mixtures, it is particularly useful in the evaluation of electrostatic ignition sources.

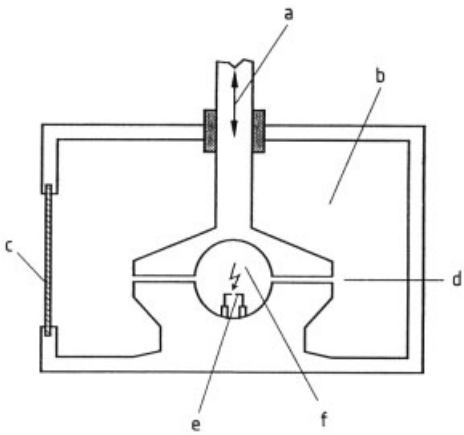

Figure 16. Apparatus for determining the maximum experimental safe gap
a) Gap adjustment; b) Outer chamber; c) Observation window; d) Gap; e) Ignition source; f) Inner chamber

The minimum ignition current, the minimum ignition energy, and the maximum experimental safe gap (MESG) make it possible to rank fuels with allowance for the ability of explosive mixtures to propagate initiating action through a narrow gap.

The apparatus for determining the MESG is shown in Figure 16. It consists of two chambers connected by an annular gap of constant length (25 mm) and variable width. The two chambers and the gap are filled with an explosive mixture of the most sensitive composition. After ignition in the inner chamber (ca. 20 cm^3), the presence or absence of a flame breaking through into the outer chamber is determined visually (cf. IEC 79-1 A).

On the basis of the MESG, fuels are classed into explosion groups (EN 50 014, DIN VDE 0165): I (CH_4; only relevant for hazards due to firedamp); II A (e.g., C_2H_6); II B (e.g., C_2H_4); II C (H_2, C_2H_2, CS_2).

This knowledge makes it possible to design objects such as flame arresters, which prevent the transmission of explosions between sections of the plant, or to build explosion-resistant electrical equipment with the "flame-proof enclosure" (d) rating (EN 50018). A lamp switch housing so rated is constructed so that ignition of the mixture inside the housing does not propagate outside it; this requires the housing also to be explosion-pressure resistant.

The great advantage of being able to rank fuels is that a knowledge of the explosion group is enough for the selection of suitable explosion-resistant equipment for ignition-source elimination or explosion decoupling of certain plant sections; there is no need for costly experimental studies.

When a gas mixture contains species belonging to different explosion groups, it is necessary to determine whether the properties of the sensitive components (e.g., H_2 or other Group II C substances) govern the explosion propagation behavior of the mixture as a whole.

2.2.3. Explosion Indices of Dust-Air Mixtures

2.2.3.1. Introduction

All substances that burn in the solid state may "explode" when finely dispersed in air in the form of a dust cloud. The violence of such a dust explosion, which is best characterized by its pressure – time diagram, is similar to that of the explosion of a homogeneous gas – air mixture. In the latter, fuel dispersion takes place on a molecular level whereas in dust – air mixtures the size of the dispersed particles is typically 10 – 200 µm. Historically, dust explosions have been associated with coal mining and grain milling. Further industrialization has led to dust explosions in the chemical, pharmaceutical, metallurgical, food, and wood processing industries. All

these industries have supported research to investigate measures to prevent dust explosions, or to limit their effects [61–66].

The probability of a dust explosion is highest for very fine dusts (particle diameter < 63 µm). The upper size limit for particles forming an explosible dust cloud is ca. 400 µm. Handling or processing of coarser product may lead to the accumulation of fines (e.g., by abrasion) and thus to the formation of an explosible dust cloud.

Dust explosions may occur within apparatus in which combustible powders are handled, transported, or processed, and in which dust clouds at explosible concentrations are difficult to avoid. They may also occur in rooms with deposits of combustible powders on the floor or other surfaces when this powder is taken up in the blast wave of a primary explosion, or by some other shock wave, in the presence of an ignition source. Layers of deposited powder as thin as 0.2 mm may be sufficient to fill a room homogeneously with an explosible dust cloud.

In contrast to gas – air mixtures, which are sufficiently identified by the nature and concentration of the flammable gas, dust – air mixtures need more parameters for their exact description. Owing to gravity the particles do not stay long in suspension; the large particles settle rather quickly. Therefore, to maintain a dust cloud for a certain amount of time, and to improve homogeneity, the particles should continue to move around. A dust cloud is not normally a stationary system, but maintains a degree of turbulence which is one of the additional parameters. Others are particle size distribution (or median of the mass distribution in the case of rather monodisperse powder), water content of solid particles, and dust concentration. Ignition and explosion characteristics of dust clouds may change drastically when a flammable gas or vapor is also present (hybrid mixture).

Because of the complexity of dust – air mixtures, it is necessary to standardize methods to determine the explosion indices which describe ignition and explosion behavior.

2.2.3.2. Ignition Sensitivity of Dust Clouds

Minimum Ignition Temperature (MIT). The MIT is the lowest temperature of a hot surface at which a dust – air mixture is ignited on contact. For determination of the MIT the Godberg –Greenwald furnace or the BAM oven are used [65–67]. The Godberg – Greenwald furnace consists of an electrically heated vertical tube (Fig. 17). A dust sample is dispersed by a blast of air from a chamber containing pressurized air, and flows through the heated tube. The temperature of the tube is varied until flames are seen to emerge from the heated tube. The Godberg – Greenwald furnace is in the process of becoming an IEC standard.

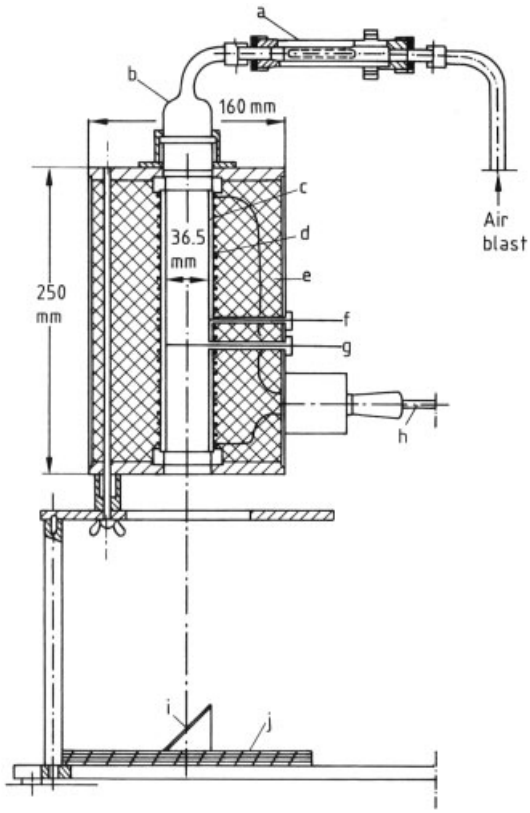

Figure 17. Godbert – Greenwald furnace
a) Dust sample chamber; b) Glass adaptor; c) Ceramic tube; d) Heating coil (1 kW); e) Thermal insulation; f) Control thermocouple; g) Measurement thermocouple; h) Power; i) Mirror; j) Heat-resistant plate

The BAM oven consists of an electrically heated horizontal pipe (Fig. 18). Close to the center of the pipe there is a cap shaped impact plate to which the dust sample is blown from the outside with air. Starting at 600 °C, the temperature is gradually decreased until ignition no longer occurs.

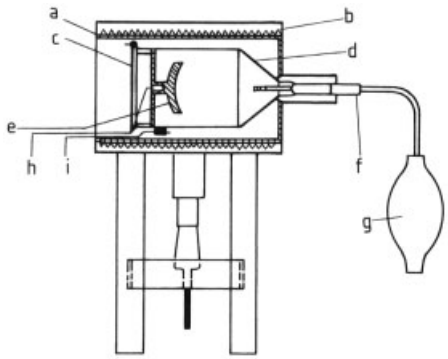

Figure 18. BAM oven
a) Oven; b) Heating coil (1.5 kW); c) Flap; d) Test chamber; e) Deflecting surface; f) Air inlet pipe; g) Rubber bulb; h) Measurement thermocouple; i) Control thermocouple

Minimum Ignition Energy (MIE). The MIE is the lowest value of the energy stored in a capacitor which, when released in a spark discharge, is just sufficient to ignite the most readily ignitable dust – air mixture at atmospheric pressure and room temperature. No internationally accepted test method exists for determination of the MIE of dispersed powder, but an IEC standard is being prepared [68]. The dust is dispersed in a modified Hartmann apparatus (Fig. 19), a 1-m^3 vessel (Fig. 20) [66], [67], or a 20-L apparatus (Fig. 21) [66], [67]. The spark gap must be adjusted to 6 mm and the discharge circuit should meet special requirements. The inductance of the circuit should be 1 – 2 mH, except when data are to be used for assessment of electrostatic ignition hazards, when it should be $\leq 25\,\mu$H. On varying the ignition delay time

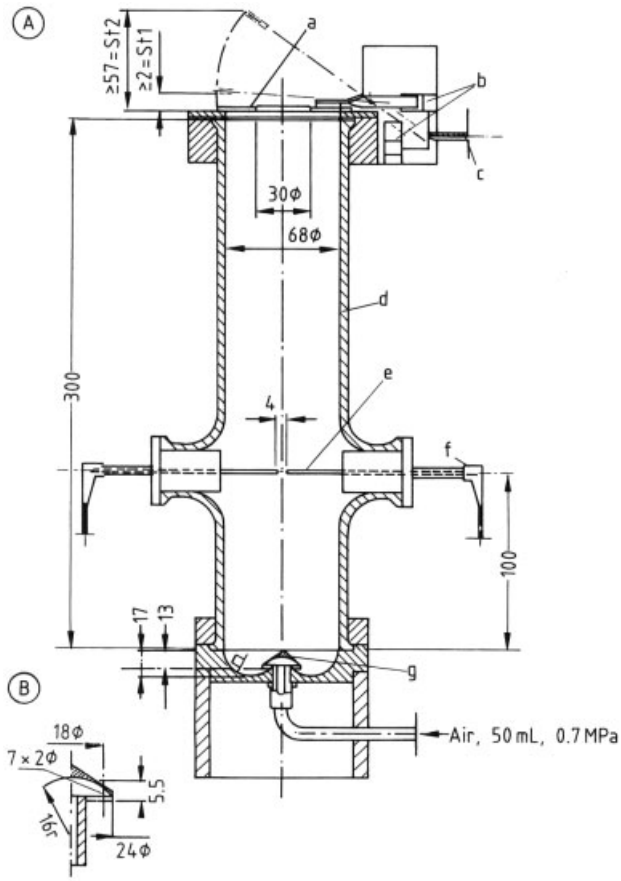

Figure 19. Modified Hartmann apparatus (A); Dust dispersion device (B)
a) Cover (Al, 17.2 g); b) Induction coil; c) Connection for digital indicator; d) Pyrex tube; e) Electrodes; f) Connection to transformer; g) Mushroom-type dust dispersion nozzleDimensions given in millimeters.

(between dust dispersion and activation of the ignition source) in the modified Hartmann apparatus, as well as the dust concentration (in all types of enclosure) the energy is always decreased by the same factor until no ignition occurs in a defined number of successive trials. The actual MIE lies between this lowest energy and the next higher energy at which ignition persists. Both values are usually reported.

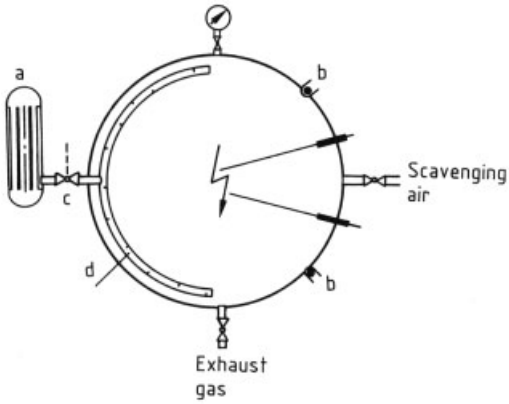

Figure 20. 1-m^3 Vessel
a) Dust container; b) Pressure sensor; c) Detonator-activated valve; d) Semiannular dispersion nozzle

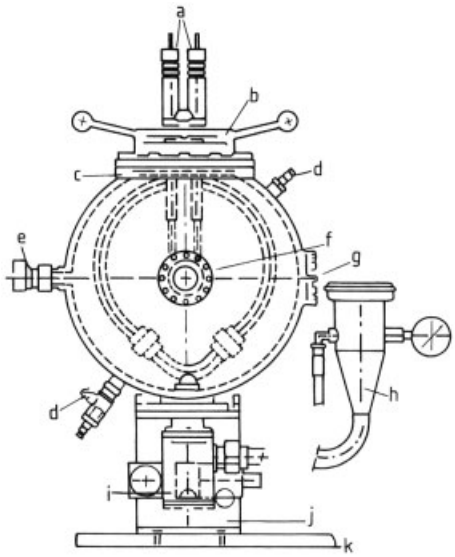

Figure 21. 20-L Apparatus
a) Ignition leads; b) Bayonet closure; c) Flange; d) Thermostatically controlled heating/cooling; e) Vent; f) Measuring flange; g) Sight glass; h) Dust storage chamber; i) Outlet valve; j) Support; k) Base plate

The MIE of fine organic powders is typically ca. 1 – 1000 mJ. It decreases with increasing temperature.

2.2.3.3. Dust Explosibility

The fact that a dust cloud is not normally stationary has implications for the course of the flame front after ignition. From studies of flammable gases, it is known that any disturbance of the flame front results in an increase of the reaction interface and thus leads to an increase of the rate of pressure rise. The same is true for combustible dusts. The rate of pressure rise measured in any test apparatus therefore depends on the dust dispersion procedure and the exact moment of ignition. Knowledge of the rate of pressure rise is most important when designing explosion venting or explosion suppression. It is necessary to use a procedure for the generation of the dust cloud which can easily be reproduced, and which represents the worst turbulence characteristics that may occur in practice.

The ISO dust dispersion procedure, developed in Germany in the 1960s, meets these requirements. This procedure is used worldwide, in combination with 1-m^3 or 20-L explosion chambers. In combination with the 1-m^3 vessel, it has become an ISO standard [69]. In Germany, this dust dispersion procedure in combination with the 1-m^3 vessel or the 20-L apparatus is recommended in the VDI Guideline 2263 [67]. At CEN level, work is in progress on standardization of dust explosion indices on this basis. The dust sample is kept at high air pressure (2 MPa) in a storage container connected to the explosion chamber. For generation of the dust cloud, the powder is released into the explosion chamber via a rapid-action valve and a specially designed dust dispersion system. After a well-defined interval (ignition delay time) a chemical ignitor, located at the center of the explosion chamber, is activated, and the pressure within the explosion chamber is recorded as a function of time.

For many years the Hartmann apparatus has been used for dust explosibility tests (e.g., the closed Hartmann apparatus and the modified Hartmann apparatus). The latter (Fig. 22) has a hinged top cover. When an explosion occurs, the cover is lifted to a certain degree, depending on the violence of the explosion (i.e., the pres-

sure rise per unit time). The closed Hartmann apparatus is almost obsolete, but the modified Hartmann apparatus is still useful as a screening instrument. However, for reliable dust explosion indices (lower explosion limit, LEL, limiting oxygen concentration LOC, maximum explosion pressure p_{max}, maximum rate of pressure rise $(dp/dt)_{max}$, or the constant K_{max}, required for the design of safety measures such as inerting, explosion venting [70] or explosion suppression [71], it is necessary to use results from tests performed in accordance with the ISO standard test procedure in the 20-L apparatus or the 1-m^3 vessel.

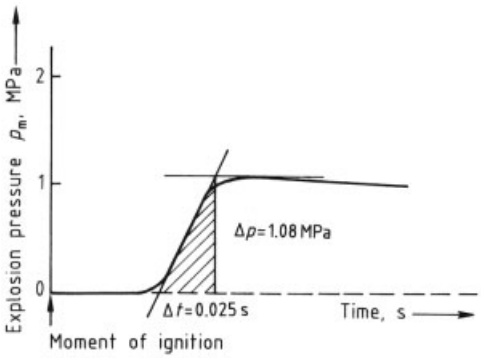

Figure 22. Explosion pressure in an enclosure as a function of time

$$\left(\frac{dp}{dt}\right)_m = \frac{1.08\,\text{MPa}}{0.025\,\text{s}} = 43.2\,\text{MPa/s}$$

Explosion Limits (EL). The explosion limits describe the range of dust concentrations in air within which an explosion is possible. Usually, only the lower explosion limit (LEL) is determined. Owing to the rather fast settling of dust particles, the upper explosion limit (UEL) is usually not significant. It is important to note that the LEL loses its importance if deposits of fine powder are present which may become dispersed in an uncontrollable way. For the determination of the LEL of powders the 1-m^3 vessel or the 20-L apparatus are used, the former in combination with chemical ignitors of total energy 10 kJ, the latter with chemical ignitors of total energy 1, 2, or 10 kJ (the exact value is still a matter of debate). Starting with an explosible concentration, the concentration is gradually reduced until there is no longer an explosion. The dust concentration (g/m^3) at which ignition just fails over a defined number of successive trials is stated as the LEL.

The LEL can also be calculated from the heat content of the dust [72–74]. For organic powders it is typically ca. 15 – 60 g/m^3. The LEL decreases with increasing temperature and increases proportionally with increasing absolute pressure.

Limiting Oxygen Concentration (LOC). The LOC is the experimentally determined maximum residual oxygen concentration in a mixture of air and inert gas at which a dust explosion does not occur. It depends on the dust and the type of inert gas. For determination of the LOC of powders, the 1-m^3 vessel or the 20-L apparatus are used (the former in combination with chemical ignitors of total energy 10 kJ, the latter with chemical ignitors of total energy 1, 2, or 10 kJ). Starting from an oxygen concentration, at which explosion occurs, the oxygen concentration is gradually reduced, varying the dust concentration until no explosion occurs. The oxygen concentration (vol %) at which no explosion occurs over a defined number of successive trials is stated as the LOC.

For organic powders and nitrogen as inert gas, the LOC is usually 9 – 14 vol %; for fine metal powders values down to 5 vol % are reported [66]. It decreases with increasing temperature; the pressure has only a minor influence on the LOC measured in vol %.

Maximum Explosion Pressure (p_{max}) and constant K_{max}. p_{max} is the maximum explosion pressure of a combustible powder, in a closed system at optimum concentration. In spherical or cubic enclosures with the ignition source located at the center, p_{max} is independent of volume. K_{max} is the rate of pressure rise per unit time, measured in a 1-m^3 enclosure at optimum concentration. The dependence of the rate of pressure rise on the volume V is known as the cubic law (see also Section 2.2.2.3):

$$K_{max} = \left(\frac{dp}{dt}\right)_{max} V^{1/3} \tag{17}$$

This relationship is used to calculate K_{max} from results obtained in test apparatus of any volume.

Both, p_{max} and K_{max} can be determined in the 1-m^3 vessel or the 20-L apparatus with chemical ignitors of total energy 10 kJ located at the

center. The dust cloud is generated according to the ISO standard procedure and ignited after a well-defined ignition delay time. The explosion pressure p_m is measured as a function of time (Fig. 22). Such explosion tests are performed over a wide range of dust concentration until no further increase in either p_m or $(dp/dt)_m$ is observed. From a plot of p_m and $(dp/dt)_m$ against dust concentration, the final quantities p_{max} and $(dp/dt)_{max}$ are determined (Fig. 23).

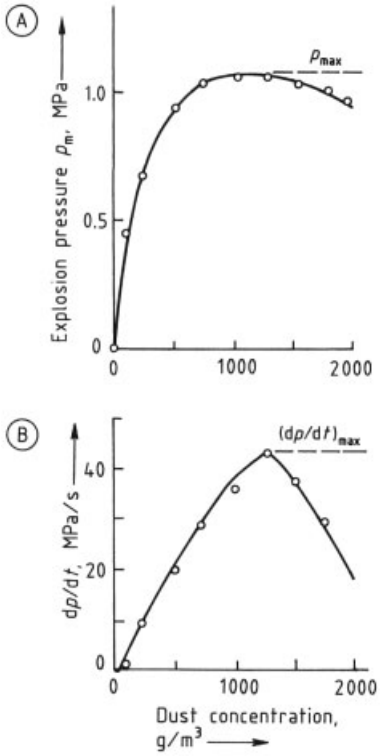

Figure 23. Explosion pressure p_m (A) and $(dp/dt)_m$ (B) as a function of dust concentration
Maximum explosion pressure 1.08 MPa; maximum rate of pressure rise 43.2 MPa/s.

Fine organic powders usually have p_{max} ca. 0.6 – 0.9 MPa, and K_{max} ca. 5 – 30 MPa m s^{-1}. Very fine aluminum powder may have p_{max} up to 1.3 MPa and $K_{max} > 100$ MPa m s^{-1}. Substances are classified in accordance with their explosion violence, as expressed by K_{max}, tabulated as follows:

K_{max}, MPa m s^{-1}	Dust explosion class
<0 – 20	St 1
20.1 – 30	St 2
>30	St 3

Determination of p_{max} and K_{max} by the standard procedure described is the decisive test of whether a powder can be ignited at all in the form of a dust cloud.

p_{max} increases proportionally with increasing absolute pressure at least up to 1.1 MPa initial absolute pressure. For K_{max} this is true only up to initial pressure 0.2 – 0.4 MPa; p_{max} decreases with increasing temperature, and the effect of temperature on K_{max} cannot generally be predicted.

2.2.4. Characterization of Explosive Condensed Substances

2.2.4.1. Introduction

Explosions can occur in the gaseous phase as well as in the liquid or solid phase. Three types of explosions have been defined in Chapter 1.

- Explosions due to spontaneous decomposition
- Explosions due to spontaneous combustion
- Thermal explosions

Gas-phase explosions (due to spontaneous combustion) and the thermal explosions are described in Sections 2.2.2, 2.2.3, and 2.3. This section discusses condensed explosive substances which are likely to undergo explosions due to spontaneous decomposition.

2.2.4.2. Chemical Nature and Definition of an Explosion

A condensed-phase explosion is defined as an exothermic chemical reaction, during which gases and vapors are formed at such a high rate as to have destructive effects on their surroundings. Fragments, projectiles, ground and air shock waves are produced by the high temperature, pressures, and rates of pressure rise. The extent of damage depends on the type and the mass of the explosive substance as well as on the nature of the surroundings.

An explosive reaction can be initiated only in substances which have a high positive heat of formation or a high negative heat of decomposition. Such substances usually have a defined chemical structure. Being organic in nature and possessing, in most cases, reactive groups

with available oxygen, they can undergo intramolecular oxidation of the combustible part. Examples of oxidizing reactive groups are nitro, nitroso, and peroxy groups, but organic compounds containing azide, acetylenic, or diazonium groups may also be powerfully explosive substances. Explosive compositions may also be produced by mixing inorganic oxidizing substances with combustible materials; well-known examples are: mixtures of potassium nitrate, sulfur, and carbon (black powder); ammonium nitrate and fuel oil (ANFO); nitration mixtures containing nitric acid; and oxidation mixtures with hydrogen peroxide. Bretherick [75] and King [76] give guidance for identifying substances or mixtures as potentially explosive because of the presence of reactive groups. When a substance or mixture contains reactive groups, it depends on the molecular mass and the type and number of reactive groups, or on the oxygen balance, whether and with what intensity explosive properties are exhibited. An important difference between gas-phase and condensed-phase explosions is that in the latter the participation of oxygen from the air is not necessary.

The power of an explosion is closely related to the dynamics of gas production and energy release, as well as to the amount of gas produced and the total heat of decomposition. The energy release when a substance explodes is difficult to determine because it depends on the resulting explosion products. There may be considerable uncertainty in reported heats of decomposition of explosive substances. A slow reaction at low temperature, due to a low degree of adiabatic efficiency under the test conditions, may result in a less energetic decomposition measurement. It is known [77] that the decomposition energy measured can be an underestimate of the energy actually available. On the other hand, computer programs such as CHETAH (chemical thermodynamic and hazard evaluation [78]) can be used to predict the maximum heat of decomposition for any substance or mixture of known chemical formula and structure. As a rough guide, substances with decomposition energy > 500 J/g may have explosive properties, and those > 800 J/g may be detonable. Although this is often a good qualitative prediction of explosivity, reliable assessment of an explosive substance necessarily requires experimentally determined data.

2.2.4.3. Grouping of Explosive Substances

Explosive substances are grouped according to their intended use and/or the type of hazard connected with handling them.

When substances, because of their energy content, are used to produce explosive, initiating, propellant, or pyrotechnic effects, they are called explosives. They may be used, e.g., for blasting in rocks, coal, or salt, as initiating explosives in blasting caps (detonators), as propellant powders in ammunition, or as components of pyrotechnic articles.

All explosives are regulated internationally with respect to handling, storage, and transport. Except for international transport, this is not so for explosive substances used for chemical, scientific, or technical purposes.

According to the type of associated hazard, explosive substances (especially explosives and articles with explosive properties) are divided into groups representing:

1) The hazard of a mass explosion (mass detonation)
2) The hazard of explosions (detonations) together with the production of fragments, without the hazard of mass explosion
3) The hazard of a mass fire
4) Ordinary fire hazard

This grouping is the basis for all modern regulations and safety measures during transport and storage (e.g., safe distances).

2.2.4.4. Explosion Mechanisms

Commonly, "explosion" is used collectively for the three distinguishable mechanisms by which a substance can explode. The thermal explosion proceeds, usually in the liquid phase and more or less homogeneously, as a temperature-controlled (Arrhenius law) self-accelerating reaction (see Section 2.3). When, during the homogeneous decomposition, a temperature gradient develops, a deflagration may start at the hottest spot. It passes through the substance as a reaction front in which the heat of reaction and the reaction products are liberated. The deflagration is propagated by conductive, radiative, or convective heat transfer into the unre-

acted materials. If a substance is liable to deflagrate, this reaction can be initiated locally, e.g., by flame, heat, impact, or friction. The deflagration velocity increases with the energy content, temperature, and porosity of the substance, and exponentially with pressure. Typical values are 0.003 – 100 m/s.

The pressure dependence of the reaction velocity explains why deflagration of a small quantity of material may proceed slowly, whereas under confinement by pressure buildup, or in even larger masses by self- confinement, it occurs with explosive violence. When the speed of the gaseous deflagration products reaches sonic velocity, a shock wave develops which propagates supersonically into the unreacted substance; a deflagration to detonation transition (DDT) has taken place. Compression combined with strong heating initiates chemical reaction of the substance. Typical detonation velocities are 1000 –8000 m/s; detonation pressures reach values up to several thousand mega pascales. Detonation velocity increases with the energy content of the substance and its density. The detonation reaction may be initiated locally by heat, impact, friction, DDT, or a shock wave from other detonating substances, leading to mass explosion, and producing disastrous damage. Primary (initiating) explosives give rise to direct immediate detonation by flame, impact, or friction in quantities of a few milligrams. Secondary explosives may be detonated by primary explosives or by DDT in larger quantities. The nature of detonation is further described in the thermohydrodynamic theory of detonation [79–81].

2.2.4.5. Testing of Explosive Substances

To determine the risks associated with handling an explosive substance, experimental investigation into its individual explosive properties is imperative. Explosive properties refer to the mechanisms by which an explosive reaction can proceed, to the types of stresses and the ease with which explosions can be initiated (sensitivity), and to the power of the explosion once it takes place. Understanding the initiation, propagation, and the possibility of terminating explosive reactions, and recognition of any destructive potential form the basis for safe handling of an explosive substance [82]. The sensitivity of an explosive substance is of special importance since, if it is too high, it may preclude some or all modes of technical handling.

For several decades, explosive properties have been described by test data determined by standardized test methods [83], [84]. The latest and most comprehensive collection of tests, covering all relevant properties together with evaluation criteria for classifying explosive substances for the purpose of transport, is given in the UN Test Manual [85]. In Part I, the Manual contains test methods and criteria, together with a classification flowchart, for deciding whether a substance or mixture is an explosive substance or article of Class 1 of the UN Recommendations on the Transport of Dangerous Goods [86], and which division of Class 1 best reflects its risks during transport. Part II of the Manual presents tests methods, criteria, and a flowchart for evaluating organic peroxides and self-reactive substances (Divisions 5.2 and 4.1 of the UN Recommendations) for explosive properties.

The Manual includes ca. 70 test methods on laboratory and field scale for the substances (packed as for transport). The test methods are related to the properties to be identified:

– Detonability by shock wave, including sensitivity to detonation shock of variable strength, detonation velocity, and sensitization by cavitation (gas bubbles)
– Deflagration after ignition in an open vessel or under confinement, determining the linear deflagration velocity or the rate of pressure rise
– DDT after ignition under confinement
– Thermal sensitivity to heating under variable defined degrees of confinement, the sensitivity being characterized by the limiting diameter of a pressure-relief vent, or by the maximum pressure and rate of pressure rise
– Sensitivity to mechanical stresses (impact and friction), determination of the sensitivity limits
– Explosive power after initiation with a blasting cap or thermal decomposition, measuring the work performed or the specific energy
– Thermal stability of explosives on storage at 75 °C, or of organic peroxides and self-reactive substances on storage under adiabatic, isothermal, or heat-accumulation

conditions, determining the self-accelerating decomposition temperature (SADT) related to the packed substances

The European Community has provisionally agreed on three methods for evaluating a substance as explosive or not. These screening methods include testing the sensitivity of substances to impact, friction, and heating under partial confinement [87]. The EC regulates the classification, packaging, and labeling of such substances for all modes of handling except transport.

2.2.4.6. Conclusions

Use of standardized test methods resulting in criteria for assessing the safety of chemical procedures and installations involving explosive substances should take special note of certain factors limiting their applicability.

It requires considerable experience to perform the tests, to modify them if necessary, and to evaluate observations correctly. Each test result should be assessed for plausibility in connection with all other results. Sensitivity does not depend on the energy content of the substance, as does the explosive power. To increase the reproducibility of test results, standardized test procedures often reflect idealized conditions. The heating rate, degree of filling of the apparatus, and the strength of confinement may greatly influence the results of thermal tests. Problems of scaling must also be addressed.

Any test result corresponds to the conditions under which it was obtained, and its validity for substances in a particular plant should be thoroughly checked. Special care should be taken before neglecting a given detonability of a substance. With respect to production in a chemical plant (excluding the explosives industry), it is imperative that initiation of a detonable substance under the conditions of handling should be impossible.

2.3. Exothermic and Pressure-Inducing Chemical Reactions

2.3.1. Introduction

The overwhelming majority of chemical reactions are accompanied by heat release, i.e., the overall change in enthalpy between the starting materials and products is negative (exothermic).

The heat produced heats up the material itself, the container (e.g., the reaction vessel), and/or the surroundings. According to the law of conservation of energy, these heat flows equilibrate in a heat balance (Fig. 24). The degree of heat dissipation depends on the heat capacity of the system itself and the heat removal capacity provided by the reaction vessel design and/or the latent heat of phase transitions. If the heat production rate of a chemical process exceeds the heat removal capacity of the system, and if there are no self-stabilizing boundary conditions, e.g., reaching of a boiling point, self-acceleration will occur, resulting in a "runaway" reaction. As runaway reactions are rather hazardous, it is necessary to understand the underlying mechanism, as well as the systematic experimental test procedures to assess the hazard potential for a runaway, and to know the design criteria for safe processes in order to prevent runaway scenarios from occurring.

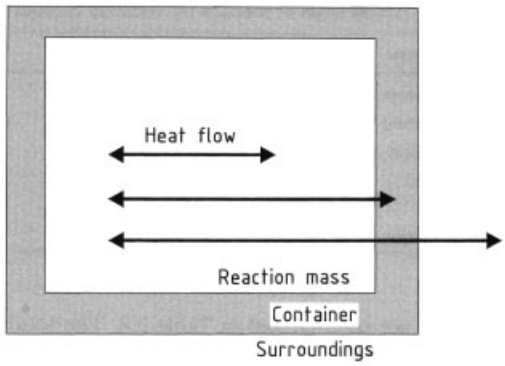

Figure 24. Schematic representation of heat balance [88]

2.3.1.1. Exothermic Reactions: Runaway Potential

To assess the runaway potential of a chemical system it is necessary to understand the basic physicochemistry and its interaction with the transport phenomena of heat, mass, and momentum, which determine the behavior of the system, including the reactor. The hazardous character of a chemical reaction is determined by the

Table 8. Typical heats of reaction for common chemical processes [89]

Reaction type	ΔH, kJ/mol	Reaction type	ΔH, kJ/mol
Neutralization (HCl)	− 55	Hydrogenation (nitroaromatics)	−560
Neutralization (H_2SO_4)	−105	Amination	−120
Diazotization	− 65	Combustion (hydrocarbons)	−900
Sulfonation	−105	Diazo-decomposition	−140
Nitration	−130	Nitro-decomposition	−400
Epoxidation	− 96		

overall heat of reaction (thermodynamics), and the rate of heat production (strongly influenced by the kinetics). Typical values for heats of reaction are summarized in Table 8. The total heat output of a real system can only be influenced by the concentration of the reacting materials. In the sense of a primary method to prevent runaway reactions it can be stated that the more dilute a system the smaller is its runaway potential. The disadvantage of this rule, besides its effect on the economics of the process, is the environmental consequence of a higher demand for auxiliary materials and energy consumption for their recovery. Consequently, an inherent runaway potential has to be accepted, but must be adequately controlled, either kinetically (e.g., by choice of process temperature), or technically (by reactor design).

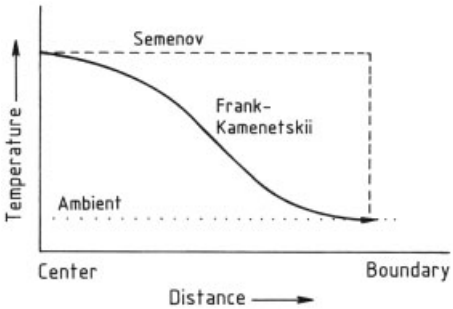

Figure 25. Semenov (uniform) and Frank-Kamenetskii (nonuniform) temperature profiles

As already mentioned, control of chemical heat release depends on the ability of the system to dissipate this energy. Mathematically this is formulated as a heat balance, including an accumulation term as well as the heat generation and dissipation terms. Appropriate description of the dissipation terms depends strongly on the heat transport mechanism (conductive or convective, free or forced). For a uniform temperature distribution within the reaction mass, which is usually the case in well-mixed gases or Newtonian liquid, heat transfer is governed by the boundary, e.g., across a reactor jacket. This situation was originally discussed by Semenov [90].

The other extreme, nonuniform temperature distribution and heat loss governed by conduction through the bulk, is a good model for large, unstirred masses, solids, and powders and was first discussed by Frank-Kamenetskii [91]. Both cases are shown in Figure 25. For more detailed discussion of self-heat models see [92].

For phenomenological discussion of runaway reaction systems the Semenov model [90] is the more adequate. It assumes a chemical reaction $A \rightarrow B$ with an indefinite amount of initial material, so that its consumption may be neglected. It is further assumed that the temperature dependence follows the Arrhenius relationship. Based on this pseudo-zero-order kinetic approach, the heat production rate $\dot{Q}_R$ is governed by an exponential dependence on temperature:

$$\dot{Q}_R = V\,(-\Delta_R H)\,k_\infty c_{A,\infty} \exp\left(-\frac{E}{RT}\right) \tag{18}$$

The rate of heat removal $\dot{Q}_C$, assuming the coolant has Newtonian properties, depends linearly on the driving temperature difference between the uniform reaction mass temperature and ambient (jacket) temperature:

$$\dot{Q}_C = U\,A\,(T_R - T_A) \tag{19}$$

The controllability of the heat production rate can best be explained by plotting the two heat flows (Eq. 18 and 19) as a function of temperature (Fig. 26). Three different cases can be discussed: with two, one, or no intersections. These three cases can be obtained either by varying the ambient temperature (Fig. 26) or by varying the slope of the heat removal lines (equal to the product of heat transfer area and overall heat transfer coefficient). The intersections represent pseudo-steady-state conditions. In case

1 (low T_A), small deviations from the steady state, represented by the lower of the two intersections, automatically result in a return to the origin, as can easily be deduced from a comparison of the relative magnitudes of the two heat flow terms. This operating point is rated "stable". With respect to the upper steady state, once a temperature deviation occurs, the original operating conditions are never reached again. In the case of a temperature decrease, the process quickly approaches the lower steady state; for a temperature increase, the heat production rate always exceeds the heat removal capacity of the system. This leads to an unhindered self-acceleration of the reaction rate and thereby of the heat production rate (thermal explosion or runaway reaction). The same is true for all operating conditions of case 3 (high T_A). Case 2 (medium T_A) represents the limiting case of the first occurrence of an unstable operating point, characterized by the equality of rates and their temperature derivatives. From these characteristics, it can be shown mathematically that, as long as the following relation holds:

$$\frac{RT_R^2}{E} > T_R - T_A \tag{20}$$

runaway conditions need not be expected. This may be regarded as a first rule of thumb for the safety assessment of chemical processes. More sophisticated design criteria, presented in 2.3.1.4 and Section 2.3.3.1, extend the stability discussion of chemical processes from this pseudo-steady-state approach to a discussion of dynamic stability.

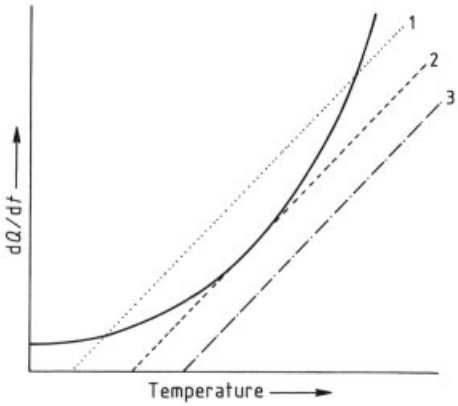

Figure 26. Semenov plot
1) Low T_A; 2) Medium T_A; 3) High T_A

2.3.1.2. Causes and Consequences of Overpressure-Inducing Exothermic Reactions

If the defined operating conditions do not include extensive overpressure, as is common for a number of petrochemical processes and hydrogenations, the sudden and unexpected occurrence of overpressure must be regarded as hazardous. The discussion of causes and consequences of a pressure increase must begin with a distinction between effects related to gas-generating chemistry and those related to process design.

For chemistry-related effects, the prime cause of overpressure is the formation of noncondensable gases. These may be reaction products of the desired process or of secondary reactions initiated at the elevated temperatures due to a runaway. For the first category, the basis of a suitable safety assessment is a fundamental understanding of the chemical process. Many common chemical reactions produce large amounts of gases, e.g., reductions with $LiAlH_4$ or substitutions of NH_2 groups by halogens via diazonium intermediates. Less obvious processes are those where the gas is formed in parallel reactions, e.g., the formation of gaseous ethylene oxide from 2- chloro-1-ethanol in the presence of a strong base. Provided that the reaction has been properly identified, these kind of gas-generating processes are controllable by conventional means, as their generation rate is kinetically controlled.

The most important class of gas-generating processes are decomposition reactions. These are initiated, e.g., when a runaway reaction raises the operating temperature so that the decomposition reaction overwhelms the desired reaction, or when cooling medium comes into contact with the reaction mixture, owing to a jacket or pipe rupture. Many of such undesired reactions stoichiometrically produce more than 1 mol gas per 1 mol initial material, which accelerates the pressure increase dramatically. In these cases, the process cannot be controlled and only mitigating measures can help reduce the consequences. It is therefore important to assess the process in all conceivable fault conditions, in order to initiate appropriate process or plant modifications to make the process inherently safe,

or to define suitable preventive measures as the second best choice.

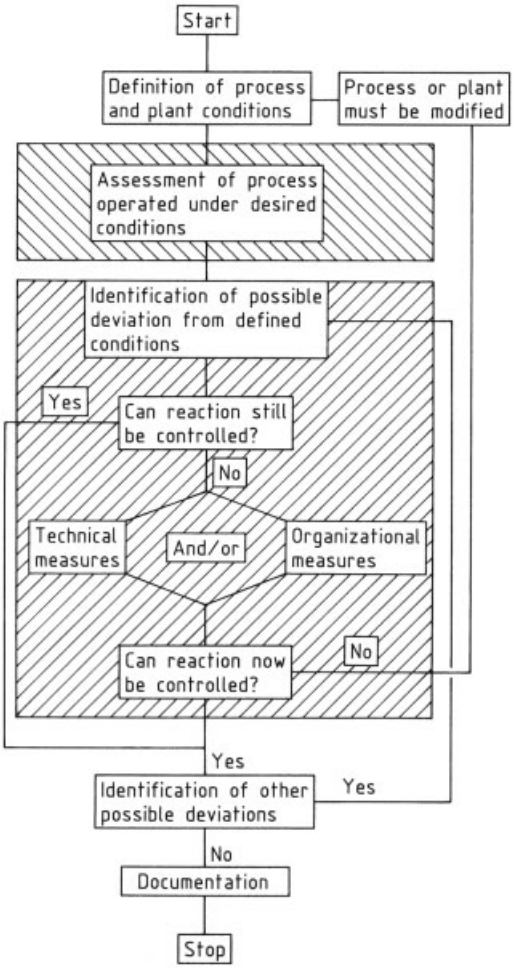

Figure 27. Process hazard and safety evaluation (PHASE)

2.3.1.3. Processes Hazard Assessment and Safety Evaluation for Exothermic Reactions

Process hazard assessment and safety evaluation (PHASE) [93] is a mandatory procedure to assure safe manufacturing of chemicals. The PHASE procedure has two characteristics:

1) It has to follow predefined systematic pathways to optimize its efficiency to identify as many hazardous situations as possible, and to assure the implementation of recommended safety measures
2) It is not performed once and for all, but should be used repeatedly during the lifetime of a plant or process

PHASE has to be applied to differing degrees, depending on the status of development of a process. Preliminary assessments should be performed on laboratory scale. The first fully comprehensive application of PHASE is recommended before the first run, on pilot plant scale. Later it should be repeated whenever a process scale-up or process/plant modifications are intended.

PHASE is a stepwise iterative procedure (Fig. 27). It starts with an assessment of all chemicals, together with operating and plant conditions, for the desired process. If hazardous situations are identified at this stage, process or plant modifications have to be initiated. Experimental methods and evaluation criteria for this purpose are the subject of Sections 2.3.2 and 2.3.3. If the desired process is rated as safe, possible fault conditions and process deviations have to be identified and evaluated. If the implementation of technical or organizational preventive measures cannot assure safe operation, further process or plant modifications become mandatory, and assessment of the normal operating conditions has to be repeated, accounting for the changed parameters. This is the iterative part of PHASE. For the second step in PHASE, tools such as HAZOP, fault trees, or FMEA are recommended (see Section 3.3).

2.3.1.4. Hazard Characteristics of Exothermic Processes due to Process Design

Chemical reaction engineering principles form the basis of an understanding of reaction hazards originating from process design. Basically three different designs can be distinguished: continuous, semicontinuous, and batch. A detailed discussion of reactor design principles is given in the standard textbooks [94], [95]. This section reviews those characteristics which have a direct impact on the hazard assessment of exothermic processes.

Design and safety criteria depend strongly on whether the system is homogeneous or heterogeneous. Heterogeneous processes depend on a much larger number of parameters (e.g., particle size distribution, surface area, and porosity for

solid – liquid systems, or mass transfer parameters for liquid – liquid systems). For a detailed discussion of heterogeneous processes see [94], [95]. The following discussion is for homogeneous systems only.

Continuous processes are performed either in continuously stirred tank reactors (CSTR) or plug-flow tube reactors (PFTR). Phenomenologically, the longitudinal spatial coordinate of the PFTR is equivalent to the time coordinate of a batch process. Therefore, with respect to continuous processes, the following discussion focuses on CSTR. Continuous processes have two major advantages: the reaction rate can be controlled by the reactant feed stream; the reaction products and the solvent immediately dilute the initial materials entering the reactor. Both characteristics influence the effective concentration of reacting materials directly and reduce the thermal potential. The problem areas with CSTR are the correct design of the start-up procedure and the avoidance of dynamic instabilities. Additionally (but not of primary importance to hazard assessment), continuous processes require a larger reaction volume than discontinuous processes to achieve the same space – time yield. It is recommended to start up a CSTR as a semicontinuous process, up to the point where the expected stationary extent of reaction is achieved, and then to start the second feed stream. To avoid ignition/extinction and oscillatory phenomena, design criteria are available in [94], [95], which can easily be applied, provided the thermal kinetics of the overall process have been determined appropriately, e.g., by reaction calorimetry.

The true batch process represents the other extreme of reactor design. Batch reactors are often preferred in the fine chemicals and pharmaceutical industries because of the smaller amounts of substance handled. From a hazard evaluation viewpoint, batch processes are the most difficult, as all reacting materials are charged initially, so that the total hazard potential is present in the reactor right from the beginning. Depending on plant or process design, batch reactors (BR) can be operated either isothermally, under computer control, or isoperibolically, i.e., with a coolant kept at constant temperature (Fig. 28). In the first case, the assessment must focus on the maximum heat rate developed, ideally at $t = 0$. If the coolant temperature can be rapidly lowered by the control system, quasi-isothermal conditions can be achieved. On the other hand, sensitivity analyses by SEMENOV and others suggest, that the cooling temperature difference should not be too great [96]. There are four engineering numbers, which can be used to give guidance on safe operability: the adiabatic temperature increase ΔT_{ad}; the thermal reaction number B; the dimensionless reaction rate Da (Damkoehler number); and the modified Stanton number St (dimensionless cooling capacity of the system):

$$\Delta T_{\mathrm{ad}} = \frac{(-\Delta_{\mathrm{R}} H)\, c_{\mathrm{A0}}}{(-\nu_{\mathrm{A}})\, r c_{\mathrm{P}}} \tag{21}$$

$$B = \frac{E \Delta T_{\mathrm{ad}}}{R T^2} \tag{22}$$

where ν_{A} is the stoichiometric coefficient of the limiting component.

$$Da = \frac{(-\nu_{\mathrm{A}})\, r_0 t}{c_{\mathrm{A0}}} \tag{23}$$

$$St = \frac{U A t}{V r c_{\mathrm{P}}} \tag{24}$$

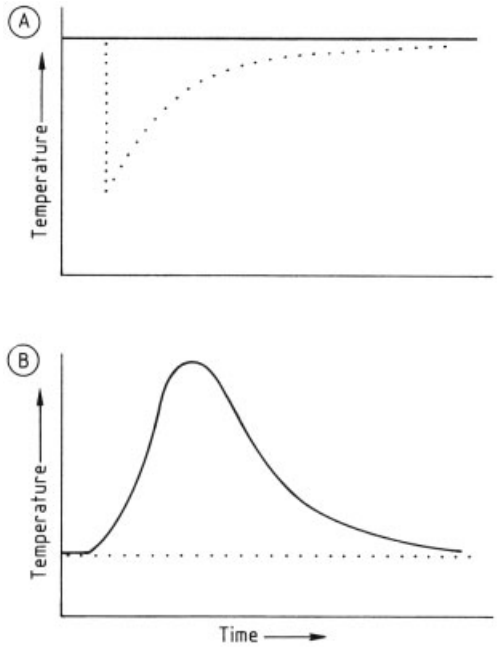

Figure 28. Batch reactor operating modes
(A) Isothermal; B) Isoperibolic
Full lines, internal temperature; dotted lines, jacket temperature.

If $\Delta T_{\mathrm{ad}} < 50\,\mathrm{K}$ and $B < 5$, isothermal batch process can be operated safely. Such processes

may be regarded as inherently safe, acknowledging that possible plant hazards are excluded from this assessment.

Under isoperibolic operating conditions, the hazard assessment focuses on the maximum temperature difference, which, under these conditions, occurs some time into the reaction. Additionally it has to be assessed whether or not the maximum reaction temperature exceeds a limiting value, which represents the temperature level at which undesired reactions become dominant. It is also necessary to assess the sensitivity of this operating point to the normal variability of other operating parameters, e.g., coolant temperature or overall heat transfer coefficient. If *Da/St* $<<1$ and thereby $(T_{\max} - T_{\mathrm{A}}) < 0.15\,\Delta T_{\mathrm{ad}}$, such batch processes can be regarded as safe.

Semicontinuous processes are performed by charging one reacting component initially and feeding the second at a constant rate. Processes performed in these semi-batch reactors (SBR) combine the advantages of CSTR, i.e., control of the heat production rate by the feed rate, and of BR, i.e., the relatively small reactor volumes. These positive factors become effective only if the problem inherent to SBR design – the risk of reactant accumulation – can be prevented. Accumulation occurs if the rate of addition exceeds the rate of reaction. A SBR can be operated safely if *Da/St* >1. Provided this is assured by the process design, accumulation of the added component does not pose a hazard [97].

2.3.2. Methods of Investigation and their Systematic Application

It is the aim of safety studies either to determine whether or not a chemical process may be carried out under predefined conditions (e.g., optimal conditions with regard to yield and time), or to establish safe operating conditions with respect to temperature, pressure, concentrations, and time. For this purpose, it is necessary to make experimental investigations by appropriate test methods. Theoretical considerations may be helpful (e.g., estimation of the hazard potential of a process from the chemical equation and molecular structures), but they cannot replace experiments. Several screening methods and more detailed test methods are used in the chemical industry to obtain information necessary for the assessment of exothermic or pressurizing chemical processes [98].

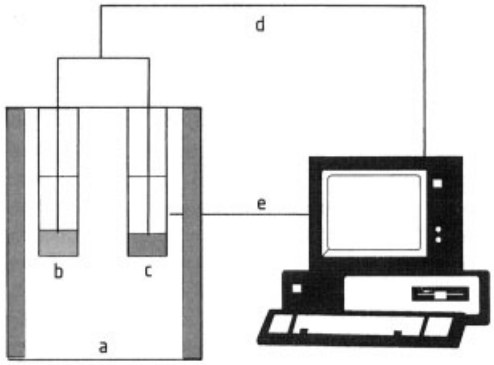

Figure 29. Differential thermal analysis (DTA)
a) Oven; b) Sample; c) Reference; d) Temperature difference signal; e) Oven temperature signal

2.3.2.1. Screening Methods

Thermal Analysis. Differential thermal analysis (DTA) is a common method for the investigation of chemical reaction hazards. A sample of the material to be tested, which may be a pure compound, a reaction mixture, a residue, or a waste sample, is exposed to thermal stress. For this purpose, the sample and a thermally inert reference material (e.g., Al_2O_3) is placed in an oven. Thermal stress can be applied either dynamically by ramping up the oven temperature at a fixed rate, or isothermally in thermal aging at elevated temperatures. The difference between sample and reference temperature is recorded as a function of oven temperature or time. The recorded signal is converted to a heat flow by a calibration function. For screening purposes, linear heating rates of 1 – 10 K/min, from room temperature up to 300 – 500 °C, are widely used. To suppress the endothermic effect of vaporization and catalytic effects of the vessel material, sealed glass ampoules are often used. A more sophisticated method is differential scanning calorimetry (DSC), where the heat flux between the sample and its surroundings is recorded directly. The resulting thermogram shows exothermic and/or endothermic transitions, originating from physical or chemical effects. Integration of the signal peaks provide

the heat content of the observed transitions. The measurable beginning of the transition is called the onset temperature (Figs. 29, 30).

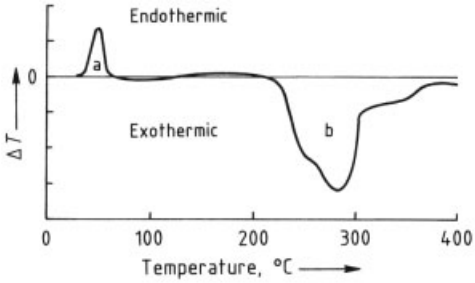

Figure 30. Typical thermogram of a solid
a) Melting (endothermic); b) Decomposition (exothermic)
Heating rate 3 K/min

The advantages of DTA/DSC are the use of small sample masses and the short time (1 – 3 h) required for a run. Onset temperature, peak area, and peak shape give hints on the hazard potential of a sample. It is not permitted to perform a direct scale-up of the measured values to the true chemical process or unit operation conditions. The recorded onset temperature depends critically on the measuring conditions, e.g., heating rate and sensitivity of the apparatus. Furthermore, there is no analytical control on the extent of reaction. So the area of a peak due to a chemical reaction cannot be interpreted as equivalent to the molar heat of a defined reaction. As the sample mass cannot be stirred during the experiment, the measuring conditions are not representative for heterogeneous systems. In particular, thermograms of such samples must be interpreted carefully.

Other Screening Methods. Isoperibolic calorimetry on the scale of several grams is another screening method for testing the thermostability of chemical systems, and may be used as an alternative or in addition to DTA/DSC. The calorimeter consists of a jacketed sample vessel (the jacket temperature may be programmed). The temperature difference between the sample mass and the jacket is measured and recorded as a function of jacket temperature or time. The recorded curves are evaluated similarly to DTA/DSC curves. The use of a larger sample mass (compared with DTA/DSC) and the option of stirring allows the investigation of heterogeneous systems with more success. The low mechanical stability of the sample vessel may be a disadvantage.

For a comprehensive assessment of the hazard potential of chemical systems, not only the thermal behavior, but also pressure effects in a closed vessel, or the rate of material loss in an open system, are of interest. Both effects can be measured with the help of these screening tools as a function of the sample temperature (programmed) for sample masses of ca. 50 – 1000 mg. For the interpretation of results, restrictions similar to those for thermograms apply. The results must not be scaled up directly to the true process and plant conditions.

2.3.2.2. Thermal Aging and Heat Accumulation Storage Tests

Data from thermal aging tests can be used directly for plant-scale assessment, provided the test conditions are representative of actual process conditions (e.g., quality of solid or liquid sample identical with the actual plant material, temperature and testing time identical with the process temperature and residence time). By varying the sample amount it is possible to extrapolate the experimental results for larger quantities. A sample vessel of defined size (e.g., a glass cylinder 100 – 500 mL) is stored in a drying oven at fixed temperature. Temperature sensors inside the sample detect the onset of an exothermic reaction. It is possible to eliminate the dependence of the exothermic onset on the sample mass by storing the test sample under adiabatic conditions, using so-called heat accumulation storage tests (open or under pressure). For adiabatic tests, the sample is stored in a Dewar flask with an insulating lid. A higher degree of adiabaticity is achieved by reducing heat losses to the surroundings. For this purpose, the Dewar flask is placed in a controllable oven, thereby assuring an ambient temperature equal to the sample temperature (Fig. 31).

The heat accumulation storage test provides the sample temperature – time profile under adiabatic conditions, starting at a fixed storage temperature. In evaluating the curves, it is necessary to account for the "thermal inertia" of the experimental system. Thermal inertia expresses the fact that the heat produced by the chemical reaction system is partially absorbed

by the sample container and internal components (temperature sensor, stirrer, etc.). From a heat balance viewpoint, the extent of this effect depends on the ratio of the heat capacity of the complete measuring system to that of the reaction mixture, the so- called Φ-factor:

$$\Phi = \frac{m_R c_{PR} + m_A c_{PA}}{m_R c_{PR}} \tag{25}$$

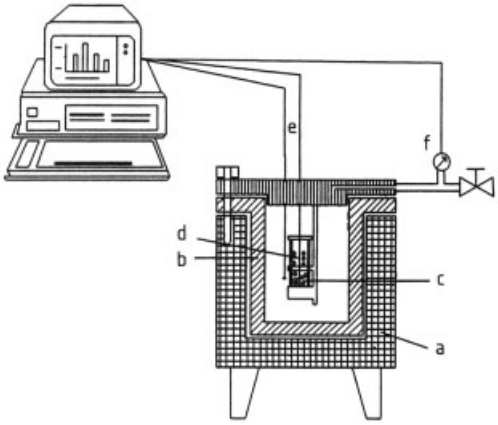

Figure 31. Adiabatic Dewar test setup
a) Oven; b) Autoclave; c) Dewar; d) Sample; e) Temperature signal; f) Pressure signal

Generally, Φ approaches unity with increasing mass of the reaction mixture. This is important as this is both demand and justification for model based mathematical correction to data obtained in small experimental tools. Figure 32 shows typical temperature – time curves for the decomposition of a substance, starting the test at different initial temperatures.

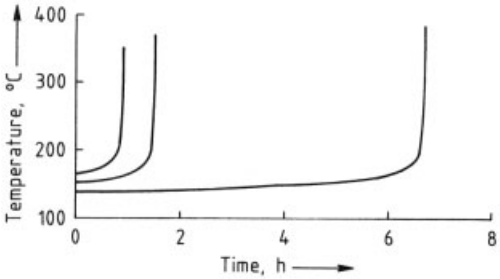

Figure 32. Typical recording of decomposition of a substance from adiabatic Dewar experiments

The handling of toxic reaction products and the repair problems are the disadvantages of the thermal aging and heat accumulation storage tests when a decomposition reaction occurs. The problem can be reduced by using smaller sample masses (< 10 g) in metallic bombs with high pressure stability. As such apparatus has a high heat-loss factor, it is necessary to use an efficient oven with accurate temperature control to minimize this heat loss.

With modern equipment, it is not necessary to predefine a fixed starting temperature for the adiabatic test. The measuring system detects the temperature at which a threshold value of the heat production rate (e.g., 1 – 2 W/kg) is exceeded, by a "search-and-wait" procedure. Evaluation of the test run yields the temperature and pressure rates as a function of temperature, as well as a thermal kinetic interpretation of the exothermic reaction under adiabatic conditions, and the adiabatic temperature increase. The temperature/pressure rate – temperature data of such an experiment (Fig. 33) can be used for designing a rupture disk or a safety relief valve.

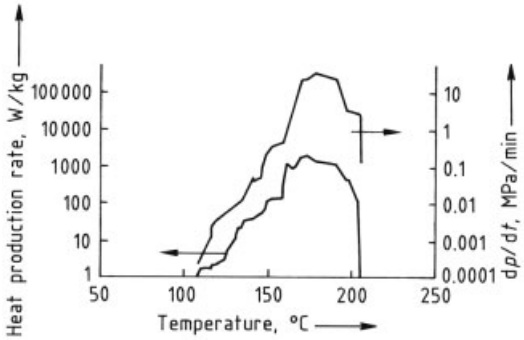

Figure 33. Experimental self-heat and pressure rate data from an adiabatic study of an exothermic reaction

2.3.2.3. Calorimetry

The methods described usually provide information on the thermal and/or adiabatic runaway potential of batch systems. They are not suitable for studying reactions under their predefined, normal or upset operating conditions. For this purpose, reaction calorimetry is the recommended method. In a reaction calorimeter, batch, semi-batch, and continuous processes can be studied, and complex reaction procedures can be simulated (e.g., pH control, temperature or pressure control, reflux, distillation). Moreover, the effect of deviating operating conditions can be tested

(e.g., by varying the amounts of reaction components, catalyst, or solvent, temperature, loss of agitation, simulation of adiabatic conditions).

The equipment consists of a stirred glass reactor with a cooling jacket. For the investigation of reactions under pressure, special test vessels are also available. The reactor lid accommodates several measuring instruments and input/output devices. The experiment is controlled by a data processing/acquisition unit. The calorimeter can be operated isothermally (reaction temperature kept constant throughout the run), in an isoperibolic mode (jacket temperature is kept constant), in a free-programmable mode (linear temperature ramps for the internal or the jacket temperature can be defined), or in an adiabatic mode. Calorimetric evaluation is based on temperature measurements and calibration runs, to characterize the heat transfer. The evaluation yields the overall heat production rate over time of the reaction system as investigated, as well as the heat of reaction. Special evaluation techniques allow determination of the thermal accumulation potential in the reactor due to unreacted material.

Evaluation can be based on heat-flow or heat-balance algorithms. In the first case, the heat flux between the inside of the reactor and the surrounding jacket is determined by measuring the temperature difference between the reaction mass and the coolant. This kind of evaluation may be significantly in error if the overall heat transfer coefficient changes markedly during the reaction in a nonlinear way. This is especially the case for polymerization reactions, as they are usually accompanied by major changes in viscosity. This systematic error can be eliminated by running the calorimeter in the isoperibolic mode, and evaluating the heat balance for the jacket independently. However, this evaluation method is rather complex.

If the calorimetric experiment is evaluated according to the heat-balance algorithm, the liquid flow rate of the coolant and its heat capacity must be well defined and measured; otherwise, significant errors may occur.

If reactions are investigated under reflux or distillation, it is necessary that all heat flows related to the vapor phase are accounted for appropriately. Figure 34 shows a typical apparatus for reaction calorimetry. The information obtainable from reaction calorimetry is of great use, not only for process safety, but also for process development and optimization.

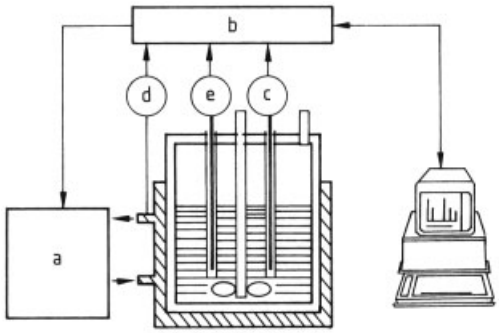

Figure 34. Typical reaction calorimetry setup
a) Thermostat; b) Controlling unit; c) Reaction temperature; d) Jacket temperature; e) Quality measurement

2.3.2.4. Systematic Testing

Some 25 different experimental techniques are commercially available for the investigation and characterization of exothermic reactions. It has to be emphasized that there is no comprehensive method; it is sensible to apply more than one method to obtain the necessary PHASE data. On the other hand, PHASE demands a systematic procedure, so a certain ranking of the different techniques is required. This can be achieved by considering the life cycle of a chemical process. Researchers usually carry out their experiments on the 1-g scale; thorough experimental hazard characterization is not required, as possible consequences of events are limited to the fume cupboard, provided occupational health and safety recommendations have been strictly followed.

Process development on the 1-kg scale represents the first milestone of the process life cycle, where the preliminary PHASE should be performed, based on screening test data. Before the first run of a process in a pilot plant, the first comprehensive PHASE is required. This can only be performed reasonably if it is based on data which are not subject to large model-based, scale-up corrections; reaction calorimetry and Dewar tests are the recommended techniques [99].

For intended plant or process modifications, selection of the appropriate experimental test methods depends on each individual case, and no general recommendations can be given.

2.3.3. Assessment Criteria

The incident risk is defined as the product of incident severity and its probability of occurrence. Hence, risk assessment includes both severity and probability. In the following, a generalized procedure for risk assessment of runaway reactions with special emphasis on evaluation criteria is presented.

2.3.3.1. Runaway Scenario

It is assumed that a batch reactor is being operated under normal operating conditions, and a cooling failure occurs. Provided that unconverted material is still present in the reactor, the temperature increases, owing to the heat output related to completion of the reaction. This temperature increase is proportional to the amount of unreacted material. The temperature reached at the end of this period is called the maximum temperature of the synthesis reaction (MTSR). At this level, a secondary decomposition reaction may be initiated (Fig. 35). The heat produced by this undesired reaction leads to a further increase in temperature [100].

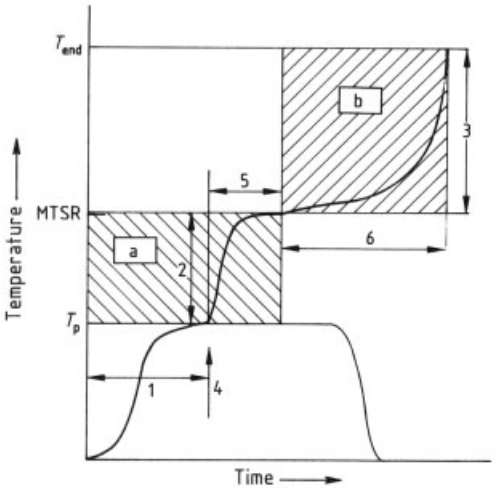

Figure 35. Runaway scenario
a) Runaway of desired process; b) Runaway of undesired process
1 = time to cooling failure; 2 = ΔT_{ad} due to unreacted material (for desired reaction only); 3 = ΔT_{ad} due to decomposition; 4 = $T_p - T_{initial}$; 5 = time to reach MTSR; 6 = adiabatic induction period

The following questions help to characterize the runaway scenario and to provide data for risk assessment:

1) Can the process temperature be controlled by the cooling system?
2) What temperature can be reached after runaway of the desired reaction?
3) What temperature can be reached due to runaway of the decomposition reaction?
4) Which moment of a cooling failure will have the worst consequences?
5) How fast is runaway of the desired reaction?
6) How fast is runaway of the decomposition at MTSR ?

This type of scenario should be worked out for the different possible process deviations, e.g., charging errors or higher accumulation due to loss of agitation. It allows assessment of the sensitivity of the process to these deviations.

2.3.3.2. Severity: Adiabatic Temperature Increase

Most reactions in the fine chemicals industry are exothermic. In the case of a cooling failure, or if heat removal is not sufficient to compensate for the heat production, the temperature increases proportionally to the heat of reaction. Thus, the reaction energy is a direct measure of the severity of a runaway, i.e., the destructive potential. It is usually expressed as the adiabatic temperature increase ΔT_{ad}.

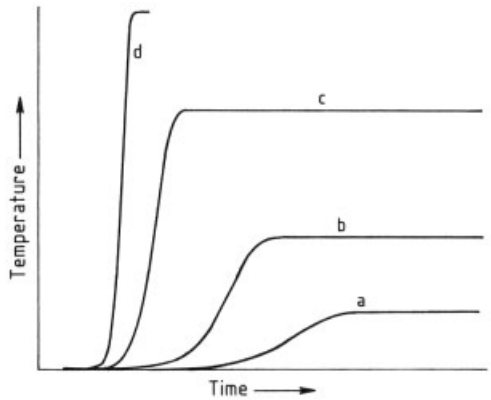

Figure 36. Adiabatic runaway temperature profiles
a) – d) increasing rates of temperature increase with increasing energy

Independent of its meaning for characterization of the hazard potential and determination of attainable temperature levels, the adiabatic temperature increase assists estimation of the dynamics of a runaway. As a general rule, high energies result in fast runaway or thermal explosion, while lower energies ($\Delta T_{\mathrm{ad}} < 100\,\mathrm{K}$) result in slower temperature increase rates (Fig. 36), given the same activation energy and initial heat release rate [89].

The heats of common industrial synthesis reactions are often of the order of -100 kJ/mol, whereas decomposition reactions may reach far higher levels, e.g., -400 kJ/mol for mononitrated aromatic compounds. In many cases, the desired reactions are not themselves inherently dangerous, but decomposition reactions may lead to dramatic effects. The consequences of a runaway event can be manifold. One possibility is solvent evaporation with the subsequent possibility of a vapor cloud explosion, if the boiling point of the system is reached. Another possibility is formation of gaseous products from a decomposition reaction, which leads to a pressure increase and the risk of vessel rupture. In practice, three levels – low, medium, and high – are sufficient for assessment the severity:

Low:	$\Delta T_{\mathrm{ad}} < 50\,\mathrm{K}$, no pressure buildup
Medium:	$50\,\mathrm{K} < \Delta T_{\mathrm{ad}} < 200\,\mathrm{K}$
High:	$\Delta T_{\mathrm{ad}} > 200\,\mathrm{K}$

2.3.3.3. Probability and Kinetics of a Runaway

There is no direct quantitative measure for the probability of occurrence of an incident, or in the case of thermal process safety, for the occurrence of a runaway reaction. But, if a runaway, anticipated to occur within several minutes, is compared with one expected to occur over several hours, it is obvious that in the first case there is very little time to take preventive measures, whereas in the second case there is time to regain control of the reaction. The probability of runaway is higher with a fast than with a slow temperature increase. While quantification of probabilities is not easy, at least a comparison is feasible.

To assess the probability of occurrence of a decomposition reaction, it is necessary to characterize its kinetics. The concept of time to explosion or $\mathrm{TMR_{ad}}$ (time to maximum rate under adiabatic conditions) is of great help [98]. It can be estimated by the following equation:

$$\mathrm{TMR_{ad}} = \frac{c_{\mathrm{p}} R T_0^2}{q_0 E_{\mathrm{a}}}\ (\mathrm{s}) \tag{26}$$

where c_{p} is the heat capacity of reaction mixture ($\mathrm{J\,kg^{-1}\,K^{-1}}$); R is the gas constant, $8.31431\,\mathrm{J\,mol^{-1}\,K^{-1}}$; T_0 is the initial temperature of the runaway (K); q_0 is the heat release rate at T_0 (W/kg); and E_{a} is the activation energy (J/mol).

This equation was derived for zero-order reaction kinetics, but it can be used for reactions of higher order, provided the influence of concentration on the reaction rate can be neglected. This approximation is particularly valid for fast, highly exothermic reactions.

The isothermal mode of DSC provides an easy way to measure the kinetic parameters in the $\mathrm{TMR_{ad}}$ equation [89]. A set of isothermal experiments is run at different temperatures. The natural logarithm of maximum heat release rate, determined on each thermogram, is plotted against the reciprocal temperature in an Arrhenius diagram (Fig. 37). Runaway curves and the corresponding time to maximum rate can also be measured by adiabatic calorimetry.

In practice, three levels are sufficient for the assessment of the probability. For discontinuous chemical reactions on an industrial scale, a probability can be considered low if the time to maximum rate of a runaway reaction under adiabatic conditions is > 24 h. The probability becomes high if the time to maximum rate is < 8 h (a working shift).

These timescales are only orders of magnitude, and depend on a lot of organizational and plant design factors, e.g., degree of automation, operator training, frequency of electrical power failures, reactor size. This scaling of probabilities is valid only if safety measures in proportion to the known severity are implemented.

2.3.3.4. Assessment of Criticality

For reactions presenting a thermal potential, criticality ranking can be based on the relative values of four temperatures.

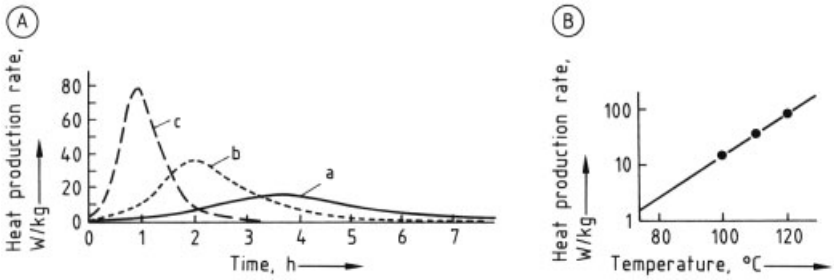

Figure 37. Determination of TMR_{ad} by differential scanning calorimetry (DSC)
A) Isothermal DTA traces: a) $T = 100\,°C$; b) $T = 110\,°C$; c) $T = 120\,°C$
B) Pseudo-Arrhenius diagram for evaluation of the isothermal runs

– $T_{process}$ (process temperature) is the initial temperature in the cooling failure scenario. In nonisothermal processes, it is defined as the temperature level at which a cooling failure produces the most severe consequences (worst case).
– MTSR (maximum temperature of synthesis reaction) depends essentially on the degree of accumulation of unconverted reactants, and therefore depends strongly on the process design [101].
– ADT 24 (temperature at which TMR_{ad} = 24 h) is defined by the thermal stability of the reaction mixture.
– MTT (maximum temperature of technical reasons), in an open system, is the boiling point. For a closed system, it is the temperature at which the pressure reaches the maximum permissible value, i.e., the set pressure of a safety valve or rupture disk.

These four temperatures permit classification of the scenarios into five different classes, from the least critical (1) to the most critical (5) [102] (Fig. 38). They are defined as follows:

1) In the case of loss of control of the desired reaction, the boiling point cannot be reached and the decomposition reaction cannot be triggered.
2) The situation is very similar to scenario (1), but if the reaction mass is kept under heat accumulation conditions for a longer time, the decomposition reaction can be triggered and the boiling point of the system can be reached.
3) In the case of loss of control of the desired reaction, the boiling point of the system is reached, but the decomposition reaction cannot be triggered. The safety of the process depends on the heat production rate of the desired reaction at the boiling point [103].
4) In the case of loss of control of the desired reaction, the boiling point is reached, and the decomposition reaction can, theoretically, be triggered. The safety of the process depends on the heat production rate of both the desired and decomposition reactions at the boiling point. Evaporative cooling may serve as a safety barrier in open systems.
5) In the case of a loss of control of the desired reaction, the decomposition reaction is triggered and the boiling point is reached during runaway of the decomposition reaction. It is very unlikely that evaporative cooling can serve as a safety barrier. The heat production rate of decomposition at the boiling point determines the thermal safety of the process. It is the most critical of all scenarios.

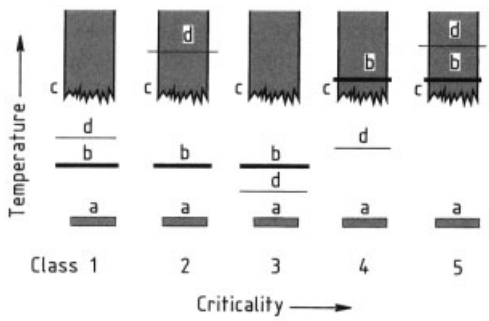

Figure 38. Assessment of criticality
a = T_P; b = MTSR; c = range where decomposition becomes relevant; d = *bp*

2.3.4. Measures

Three categories of measures, with different applications, are distinguished:

1) Hazard-preventing measures, which suppress the potential, e.g., dilution, change of synthesis route. These measures render the process inherently safe [101].
2) Hazard- controlling measures, e.g., change from batch to semibatch or continuous operation, interlocks in the process control system (Section 3.4.3) etc. These measures render the process fail-safe.
3) Consequence-mitigating measures, e.g., dumping, quenching, venting (Sections 3.5, 5.1).

Depending on the criticality, different measures can be applied to prevent or control the runaway or to mitigate its consequences:

1) No special measures are required, but the reaction mass should not be kept under heat accumulation conditions for long periods. Evaporative cooling serves as an additional safety barrier.
2) No special measures are required, but the reaction mass should not be kept under heat accumulation conditions for long periods.
3) The measure of choice is to use the evaporative cooling to keep the reaction mass under control. A distillation system must be designed for that purpose, and has to work effectively, even in cases of utility failures. A backup cooling system, dumping of the reaction mass, or quenching can also be used. These measures must be designed adequately and must be available immediately after the failure occurs.
4) Similar to (3). The same measures apply, but the additional heat production rate due to the secondary reaction has also to be taken into account.
5) In this case, the boiling point is very unlikely to serve as a safety barrier. Therefore only quenching or dumping can be used. Since, in most cases, the decomposition reactions release very high energies, particular attention has to be paid to the design of safety measures. It is worth considering an alternative process design, in order to reduce the severity or at least the probability. The following possibilities should be considered: reduction of concentration, change from batch to semi-batch, optimizing semibatch operating conditions to minimize accumulation [89], [104], changing to continuous operation, etc.

3. Development, Design, and Construction of Safe Plants

3.1. Objectives, Regulations, and Concerns

The goal of all plant safety effort is to eliminate or reduce the possible hazards described in Chapters 1 and 2. According to Section 1.3, this means limiting the risk created by the facility to an acceptable level. At the same time, the relevant regulations and engineering codes must be complied with. In many areas, these implicitly establish the required level of safety (or acceptable risk).

All the industrialized countries have adopted regulations to protect workers, uninvolved third parties, and the environment [105]. These regulations employ various combinations of the following approaches:

1) Setting standards for the quality of engineering facilities and their safe operation
2) Establishing approval or licensing procedures for the erection and operation of plants and for substantial modifications to them
3) Requiring regular inspections of the plant and its technical equipment by the operators, government agencies, and (sometimes) independent third parties
4) Assigning civil liability and legal responsibility for damage caused

Because there was initially little experience relating to the hazards of technology, the first regulations were written for individual types of apparatus regarded as dangerous (steam boilers, pressure vessels, machinery with moving parts) and aimed chiefly to insure occupational safety and to protect third parties. Later, regulations for other types of equipment were added, attempts were made to extend protection to new areas such as surface waters, and new regulatory principles such as the principle of precautionary action were introduced. It is not unusual that regulations are created by many independent bodies,

with very different jurisdictions. As a result, in Germany (for example) there is a rather complicated system of regulations for chemical plants, with overlaps and multiple rules, and the density and the degree of detailing of regulation differ widely from one regulation to another.

The safe design and safe operation of chemical plants in Germany are governed by regulations stemming from the following fields of law [106]:

1) Building and construction law
2) Occupational safety law (including the Equipment Safety Act, regulations defining facilities subject to supervision, and the Accident Prevention Regulations)
3) Hazardous substances law (including the Chemicals Act and the Hazardous Materials Regulation)
4) Water pollution control law (including the Regulation on Facilities with Water-Contaminating Substances)
5) Air pollution control law (including the Major Hazards Directive [107])

This complicated situation is not likely to be fundamentally changed by harmonization efforts in the European Single Market, for two reasons: First, under the EEC Treaty (Article 100 a) [108], only requirements on the quality of technical devices are subject to harmonization. Second, regulations in the European Union (EU) continue to be created, adopted, and enacted into law by the member countries according to the old pattern (e.g., for specific types of facilities).

As a rule, the regulatory system is structured vertically [109]. Legislation at high level sets forth the scope and objectives as well as the solution approaches in general form, while engineering codes and standards at lower levels are largely concerned with technical details.

Regulations, especially at sublegislative levels, are designed deterministically in that the required safety is deemed to be attained if the plant satisfies certain requirements on design, sizing and equipment. An exception can be found in a Dutch regulation under which a chemical plant, to be approved, must operate at a risk below a value described in terms of fatal accident probabilities [110].

Plant builders and operators view plant safety in terms of the types of hazard arising from the process. For them and their dealings with the authorities, it would be simpler if there were a single consistent set of regulations, broken down by hazard type (releases and spills, fires, explosions) and – instead of setting fixed technical solutions for protective tasks – such a code would use fixed solutions merely as examples of ways to achieve the stated objective.

This approach would have the advantage that the required safety could be achieved by the most advantageous technical resources in each instance. In addition to that, the "plant" orientated view would no longer require multiple specific regulations covering occupational safety and health, air and water pollution control, since one criterion for the plant (e.g., tightness) would simultaneously take care of related problems in the other areas.

The main safety tasks for the process designer and the designer and builder of process plants are:

1) To identify and correctly assess all types of hazard
2) To take appropriate steps to reduce and control these hazards

The key hazard types are those listed in Section 1.1:

Releases and spills
Fires
Explosions

The safety tasks can be broken down into the process itself and into the safe design and operation of the technical facility required for the process. The specific tasks can then be listed as follows:

1) Achieve safe process design by: a) identifying all types of hazards; b) assessing their hazard potentials; c) minimizing the hazard potentials; d) deactivating the hazard potentials
2) Achieve safe plant design and operation by: a) systematically analyzing danger sources; b) evaluating their probabilities of occurrence (usually qualitatively); c) minimizing sources of trouble and error; d) employing fault-tolerant design

In detail (see also Section 1.2), this means:

1 a) For all substances to be processed, safety ratings and toxicologically and ecologically relevant data must be acquired (see Chapter 2). A comparison of process parameters and design data with safety ratings, step by step through the process, reveals where danger sources exist, or may arise.
1 b) The magnitude of each hazard potential is evaluated through analyses, e.g., those described in Sections 1.4.2 and 1.4.3.
1 c) It must be determined whether the hazard potentials can be reduced through suitable process design.

The practices recommended here lead to inherent safety [111], [112]. To achieve this goal the planner should replace hazardous substances by less hazardous ones wherever possible. Large inventories should be avoided as far as possible, e.g., by using process steps that can be carried out quickly in a small volume, by introducing continuous operations in place of batch operations, and by eliminating large buffer volumes.

1 d) Any hazard potentials that remain must be deactivated in such a way that they cannot manifest themselves in the process. Lowering the temperature far below critical values or diluting substances and handling them in solution (to lower the vapor pressure) are the approved methods. They also help make the process inherently safe [111], [112].

If the plant is to be conceived for a process that has been safety optimized in this way, two analytical tasks, followed by two design tasks, must still be performed:

2 a) The system plant must be systematically searched for danger sources, i.e., possible defects and failures that can activate the deactivated hazard potential.
2 b) When possible faults are identified, their frequencies or probabilities of occurrence must be evaluated so that appropriate safety measures can be taken.

The last two tasks can now be undertaken with an eye to the magnitude of each hazard potential (task 1 b) and a (qualitative) measure of the probability of a defect that would activate the hazard potential:

2 c) All possibilities for minimizing sources of trouble and error must be exhausted [112].
2 d) As far as possible, the facility must be designed and equipped so that faults are "forgiven" without resulting in harm. One way to accomplish this is to use redundant (multiple) safety devices [112] (see Section 3.4.3).

In carrying out these tasks, some of the steps (2 a – 2 d) may have to be done more than once, recursively, because any change in the system due to the new measures can introduce new danger sources.

For this reason, and because the development of a process and the associated plant is done step by step, with concomitant advances in understanding a "holistic" procedure, segmented by time, technical specialty, and logical relationships, must be adopted in the development, design, construction, and operation of a chemical plant.

Examples of such methods are presented in the Section 3.2.

3.2. Procedure for Designing and Constructing Safe Plants

Chemical processes and their associated technical facilities are developed in steps. Process development in the laboratory is followed by testing of the process on a pilot scale before the project goes through the various planning stages (preliminary, draft, and detailed design). The planning process culminates in the purchase of equipment and erection of the plant. After a test and commissioning phase, the facility is put into production. The left-hand side of Figure 39 shows these stages [113].

Each phase involves questions as to the safety of the process and the plant, each of which must be answered immediately or, at the latest, before the process goes on to the next phase. The most expedient way of creating a safe plant is thus to plan for safety studies. At each step in process and plant development, safety analyses must be done (ideally integrated into the development) in order to pose the right questions and immediately seek solutions to the problems identified.

These ideas are the basis for the procedure chosen by Bayer, a large German chemical company (Fig. 39). Safety analyses are broken down into four sections, each concluding with a certificate prepared by safety experts. This certificate

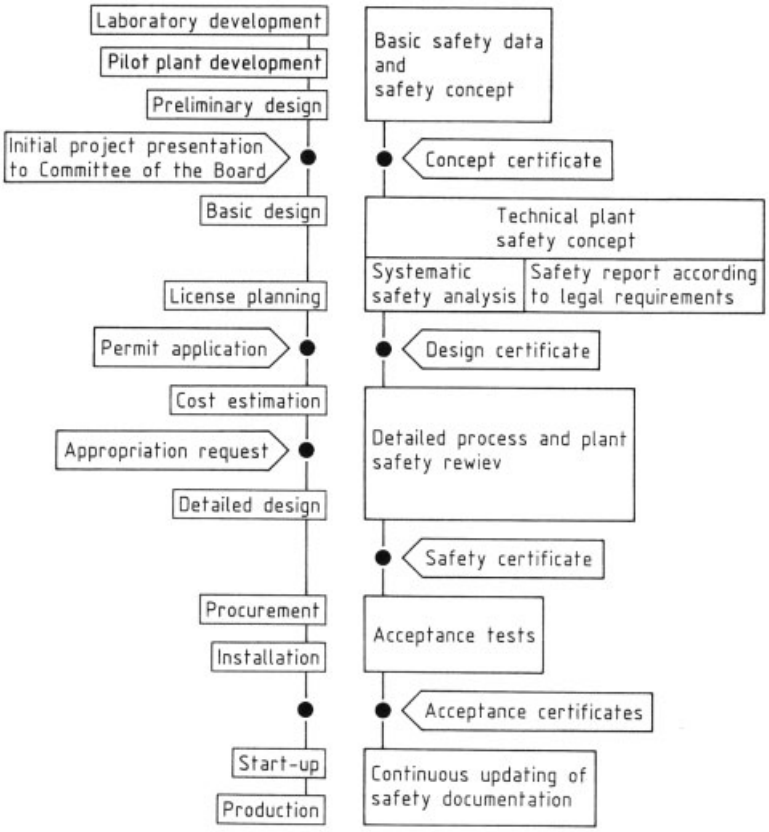

Figure 39. Technical safety audit scheme for new plants or changes to existing plants (Bayer) [113]

states that the required studies have been performed, and the correct conclusions drawn from them.

The four phases of safety analysis can be listed as follows:

A 1) To create safety principles by: compiling and determining safety, toxicological, and ecological data; identifying sources of danger in the process; examining possible safety solutions; establishing the safety concept for the process
A 2) Defining the safety concept for the plant by: performing systematic analysis; identifying technical protective measures
A 3) Performing a detailed safety analysis by: analyzing all plausible forms of trouble as to cause, effect, and corrective measures; adopting the final detailed safety concept
A 4) Conducting the safety acceptance of the plant by: doing a nominal/actual comparison (concept/implementation); carrying out functional tests

This procedure is organized in an obvious way, and has the advantage that a completion certificate must be prepared at the end of each phase, before the next phase begins. The project is not released unless the planning certificate is ready, and the plant does not go on stream unless there is a safety acceptance certificate. These rules are set forth in an internal directive [114].

Similar procedures are common in other companies. They are not always so formal, but they embody the same principles and use the same resources.

BASF, for example, uses a three-phase approach for these studies [115]. Phase 1 generates a basic safety concept with provisional solutions for the main possible hazards in the process. Phase 2 involves elaboration of the safety concept, so that it can be submitted to the author-

ities along with the approval application (safety analysis, safety report). Detailed safety verification takes place in phase 3.

Hoechst utilizes the principle that certain safety analysis and design results, once created, become important documents that must accompany the process and the plant throughout their lives [116]. A substance data report and transmission report have been developed for this purpose; they function to gather key safety information and communicate it within the company; a checklist is used in the safety assessment of production facilities.

A number of publications on the performance of safety analyses have been issued by ESCIS, the Expert Commission on Safety in the Swiss Chemical Industry [117], [118]. These stress principles and organization for this kind of work, establish key points to be analyzed at various times, and include methodological aids.

Essential in all these procedures is that nothing is overlooked and the proper specialists are brought in at the proper time. All such procedures can therefore be said to regulate three things – who does what and when.

The methods employed by Degussa [119] and by Lurgi, a plant construction company [120], assign great importance to these aspects.

Over the plant's economic life, all the results of safety work are used again and again for guidance, e.g., during maintenance and plant modifications, or when new personnel are being trained. Close attention must be paid to clear and complete documentation of the safety analyses; the recommended form for this is a concise, uniform, computerized database keyed to process, process step, plant, and plant section, and covering all danger sources with causes, effects, and corrective measures (including brief justifications). It can be used directly in the writing of safety reports pursuant to the Major Hazards Regulation (Seveso Directive) [107] and can easily be consulted from the plant operating personnel whenever necessary.

By way of summary: the design and construction of safe plants calls for a highly structured and organized procedure clearly setting forth what has to be done by whom and in what way, and focusing on the creation and routing of documents. This is called a safety management system.

In accordance with this principle, it has recently been suggested in both the United States [121] and Germany [122] that "safety management" should be implemented on the model of quality management following the well known international standards ISO 9000 ff. This idea makes sense, provided the procedures recommended for quality assurance are reinterpreted for this new purpose. This tends toward the procedures discussed above, which are better suited to the legal situation, at least in Germany.

In order that nothing should be overlooked in the analyses, it is desirable to provide methodological aids to the specialists as they deal with various problems. Section 3.3 concerns such aids.

3.3. Methodological Aids

The safety concept for a chemical plant must be complete, with an answer ready for any safety question. All necessary questions which can help reveal potential sources of danger must be posed ahead of time.

The problem of a comprehensive and complete analysis comes up twice in the course of process and plant development.

First, for the process (task 1 a of Section 3.1) the safety figures for the various substances (toxicity, flammability, explosibility) must be determined. Second, for the plant (task 2 a of Section 3.1) defects and malfunctions in the system that can activate the hazard potentials must be identified.

Appropriate safety engineering involves assessing the hazards as to both possible scope (task 1 b of Section 3.1) and probability of occurrence (task 2 b of Section 3.1).

Methodological aids available for use in these tasks [123], [124] are all characterized by clear and easily understood structures and systematic procedures.

The specific methods differ in:

Objective of the analysis
Functional principle employed
Basic knowledge required
Appropriate time of use
Aids needed
Results achievable
Documentation
Cost/benefit ratio

If we go back to the four chief tasks of safety analysis

1) Completely identifying hazards due to substances (task 1 a)
2) Assessing hazards as to potential scope (determining the hazard potential, task 1 b)
3) Identifying danger sources in the plant (identifying defects and possiblities for malfunctions, task 2 a)
4) Evaluating the possible hazards with respect to probability of occurrence (task 2 b)

the methods can be placed in four groups, two relating to the identification of problems and two to their assessment.

Table 9 summarizes the most important methods. The literature refers to many more methods than those in the table. In most cases, the additional names do not refer to formal methods, e.g., "safety audit", "what-if-method", and "preliminary hazard analysis", as defined in [124]; or they are synonyms for known methods: "human error analysis", "action error analysis", and "human reliability analysis", all concerned with human errors and using checklists, keywords, tables, or event trees to study their effects; or they are combinations of two methods, e.g., "cause – consequence analysis", a combination of incident sequence analysis and fault tree analysis [123–125].

The methods of Table 9 increase in difficulty and cost from top to bottom; with the exception of consequence analysis, they are also arranged roughly in chronological order to use in process and plant development. In accordance with this ordering, the amount of input information required, the level of complexity, and the amount of special knowledge required also increase from top to bottom. Therefore assessment methods must generally be carried out by specialists, while the identification methods should be among the tools used by every chemist and engineer in the process industries, particularly the chemical industry.

The identification methods used in safety analysis are summarized below.

Checklists and Relationship Charts. A checklist enumerates points that, according to experience, are associated with hazards in the handling of substances and mixtures of substances, or in the performance of a technical process. It can be as detailed as desired, and must be suitable for the kind of analysis being carried out. For example, checklists can be developed and used for thorough analysis of all safety-relevant material properties. If the concern is whether explosions can be initiated by undesired reactions, the checklist comprises such questions as:

– Are the substances stable?
– What happens if foreign substances are admitted?
– What happens if the process parameters (temperature, concentration) change?

The formal method is to work through such lists of questions and to determine whether:

1) A hazard is possible (yes or no)
2) Further study is needed
3) Safety measures are called for

The results of the analysis can be documented directly on the checklist or on a summary form reflecting its organization.

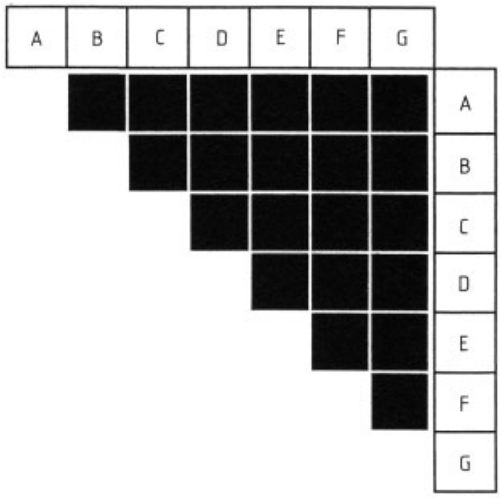

Figure 40. Reaction matrix showing all possible pairings of substances (tests for unwanted reactions)

Checklists can be used to insure the completeness of the safety concept in later phases of a plant project, e.g., in the design phase, to insure that possible events such as power outage, cooling outage, or stirrer outage are included in the safety concept. The creation and use of checklists is also customary in plant operation. They have the obvious advantage that they can

Table 9. Safety analysis methods

Task	Aim	Working principle	Method
1) Identification of hazards	1) To complete the safety concept	1) Memory-jogging	1) Checklists
			2) Relationship charts
		2) Use of search aids and tables	3) Failure modes and effects analysis
			4) Action error analysis
			5) Hazard and operability studies
2) Evaluation of hazards according to probability of occurrence	2) To optimize safety systems with regard to reliability and availability	3) Representation of interconnections of failures in graphical form and evaluation of probability	6) Incident sequence analysis (inductive)
			7) Fault tree analysis (deductive)
3) Evaluation of hazards according to consequences	3) To minimize the hazard potential and devise optimum protective measures	4) Mathematical analysis of physicochemical processes	8) Hazard consequence analysis

be adapted to any problem; their drawback is that things not included may not always be recognized and dealt with.

The same limitation applies to the second "memory-jogging" method, the use of compatibility charts [124] or relationship charts [125]. One use for this technique is to check the compatibility of all the substances that can come into contact in a process. If, for example, there are substances A to G, then half of a two-dimensional matrix (Fig. 40) gives all possible pairings of substances.

By systematically scanning over the interaction fields, it is possible to check whether mixing two substances creates a hazard, and whether further experimental study is called for. The disadvantage of this technique is obvious: mixtures of three or more components do not appear. Caution is therefore indicated, and matrices are more often used simply to illustrate safety situations, e.g., causes and effects or effects and corrective actions [117].

Hazard and Operability Studies (HAZOP) and Similar Methods. The second group of problem-identifying methods includes:

- Failure modes and effects analysis [123], [124]
- Action error analysis [123], [124]
- HAZOP studies (called PAAG methods in Germany) [127], [128]

These methods are so similar that they can be described together. All can be characterized as "deviation analysis" in that they look for possible hazards that can arise in the process, in the plant, or in plant operation if an error occurs, or the state or sequence of actions deviates from the prescribed state of sequence. (The assumption is that there are no hazards in the prescribed state or sequence.)

In this analysis of deviations, similar forms are employed for all three cases. As a rule, the form covers the following aspects:

- Deviation (error, failure)
- Cause
- Effect
- Corrective action

In many analysis, the following are also included:

- Error detection (how?)
- Frequency (often/seldom)
- Severity of effects

The third trait common to all three methods is the use of search aids. It is here and in the objects of study that the methods differ.

Failure modes and effects analysis (FMEA) is oriented to items of equipment and machinery, each having a certain function in the plant. It employs the working hypothesis that this function is not performed, i.e., each piece of equipment is examined for the effect its failure has, and what corrective action may be required.

The search aids in action error analysis and HAZOP studies are keywords describing deviations from nominal conditions or a nominal sequence. Keywords used to characterize human errors include:

- Too early

– Too late
– Not
– Wrong object
– Wrong sequence

Typical keywords for the HAZOP method [128] include:

– More
– Less
– No
– Different from

Much has been written about this and the other methods [123–128]; Table 10 presents a few examples of analyses by all three methods for a stirred-tank reactor.

All that remains is to indicate the proper time for applying these methods. In the design of a new plant, they should come in at the detailed safety design phase, when most of the piping and instrumentation diagrams are ready. These methods are also useful in the inspection of plants already built and on stream.

In the author's view, the possibility of creating suitable keywords for all parameters, functions, and action sequences in a plant makes the HAZOP method so flexible that it can cover the full range of applications of the other techniques.

Just two identification methods are therefore of any practical importance in the development of a complete safety concept: checklist methods and the HAZOP method.

Assessment of Hazards by Consequences. The concept of the hazard potential was introduced in Section 1.3 as a way of evaluating the magnitude or severity of a safety problem. Section 1.4 described how the hazard potential can be determined and what difficulties and imponderable factors must be taken into account.

If the scope and quality of safety practices are to be suited to the hazard potential so that a generally comparable and acceptable risk level is achieved, it is necessary to get at least a rough idea of the magnitude of the hazard potential (see also Section 3.1).

On the other hand, if effect-limiting safety devices are to be custom designed, e.g., an emergency pressure-relief system or a scrubber to handle off-gases in case of an emergency, it is necessary to model the physical and chemical processes taking place during the accident and to investigate their effects. Protective systems can be tailormade, and hazard potentials can be minimized (tasks 1 b and 1 c of Section 3.1).

Evaluation of Hazards by Probability of Occurrence: Incident Sequence Analysis and Fault Tree Analysis. Both methods in this class examine links between faults, represent them graphically, and (in principle) can assign probabilities to them. Incident sequence analysis starts with a single fault, and observes how it may develop. The working hypothesis is that every safety measure can succeed or fail with a certain probability.

Figure 41 shows a simple event tree for a temperature rise in an exothermic semibatch chemical reaction where three safety measures (heating shut-down, feed cut-off, and emergency cooling) have been instituted to prevent the reaction getting out of control.

The figure illustrates the three-stage safety concept. The probability of the undesired event can be calculated if the probability of the initial event and the failure probabilities of the several safety measures are known.

More careful analysis of the sequence of events reveals that the model is far too simple for the process under discussion. Whether the safety actions succeed or fail depends heavily on the intensity and rate of temperature rise, as well as the dynamic response of each action, and these are quantities that cannot be simply described in probability terms. Therefore, the only utility of event tree analysis for such processes is that it can make relationships easy to understand in simple cases.

Similar limitations apply to fault tree analysis, at least when used for typical processes in chemical systems (Fig. 42). The analysis goes in the opposite direction to that in incident sequence analysis: deductively from an undesired event. Faults are connected from top to bottom in chains that can lead to the top event. There are two main kinds of link:

1) The blocking AND junction: more than one event must occur in order to open the path upward
2) The passing OR junction: one event from a set of several events is sufficient to lead to the next higher event

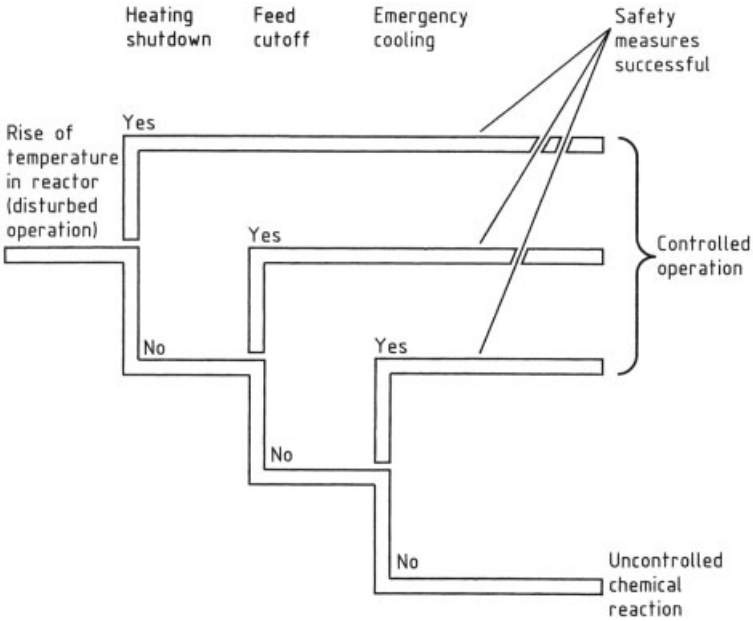

Figure 41. Event tree diagram showing success or failure of safety measures

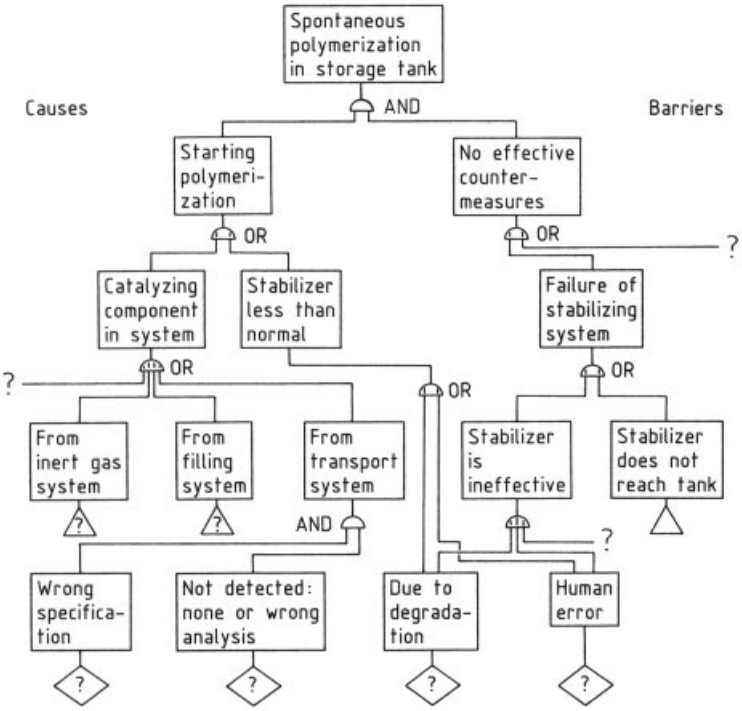

Figure 42. Fault tree diagram

Table 10. Examples of safety analysis methods applied to a stirred-tank reactor

Method	Deviation	Cause	Consequences	Safety measure
Failure modes and effects analysis (failure of equipment)	stirrer: does not turn, turns wrong way:	no electricity, motor defective, wrongly installed	heat transfer impeded, solids precipitate	revolution counter, feed shut-off, control unit
	steam valve: does not close	jammed	temperature increase in reactor	feed shut-off
Action error analysis (human error)	"too late": steam off, cooling water in	operator not paying attention, overworked	temperature increase, danger of uncontrolled reaction	alarm, feed shut-off, emergency cooling
	"not": steam off	operator not paying attention		automatic shut-off
	"incorrect": product	mix-up	unknown reaction	labeling analysis
Hazard and operability studies (deviations in functions and process variables)	"more": steam, heat	pressure reduction fails, no cooling	temperature increase, danger of uncontrolled reaction	steam shut-off by hand, feed shut-off, emergency cooling
	"none": cooling water, stirring	valve does not open, driving mechanism defective	heat dissipation impeded, temperature increase	feed shut-off, emergency cooling
	"reverse": direction of stirrer	wrongly installed	solids precipitate	direction indicator, control unit

Both fault tree analysis and event tree analysis pose stringent requirements on correct application and interpretation. Both call for experts. The methods have been fully described in the literature [123–126] and have been standardized in Germany [129], [130]; with respect to chemical safety, however, they have so far not played a significant role, for the reasons discussed earlier. Their main area of application is in the assessment and optimization of complicated technical systems in which the components display a yes/no failure behavior with assignable probabilities, and where the failure rates are known (e.g., in aviation and space flight). For further discussion see [123], [124], [131].

It can be maintained that these methods do not provide a satisfactory basis for assessing hazards in chemical processes as to probability or frequency of occurrence. The process variables in the analyses are also functions of time and intensity, and there is not a sufficiently large number of comparable states for statistical analysis. Estimates of probabilities therefore have to be established qualitatively or semiquantitatively on the basis of experience and the number and quality of safety measures.

For the same reasons, attempts to evaluate risks from chemical plants on the basis of occurrence probabilities of single events and process and component failure rates have not yielded useful results [131]; but this is not a fatal drawback because the risk due to chemical production (Section 1.4.1) can easily be derived from historical data of long-time experience and has proved low in the past.

3.4. Safe Processing: Strategies

3.4.1. Fire Protection

3.4.1.1. Fire Prevention

The key factors governing the occurrence and impact of fires are the quantity of combustible substances and possible ignition sources present in the plant. Both factors must be determined and taken into consideration for effective design of fire-protection measures.

First, it is necessary to ascertain the form and quantity of combustible, flammable, and explosive substances located inside battery limits. As well as products, auxiliary substances, and packaging materials, thus evaluation must also include energies, construction materials, and combustible parts of the plant.

Next, these combustible substances are rated according to their key fire parameters (see Section 2.2.1). The analysis must include the state of aggregation in which these substances are held in the plant under normal or process conditions, as well as the extent to which a possible fire, by raising the temperature, can initiate further critical reactions. A crucial point is whether fire loads

are unprotected or are protected from the direct action of ignition sources by noncombustible packaging or enclosures, e.g., steel tanks or reactors [132–134].

The first important measure to prevent fires is to minimize the fire loads. In industrial plants, this can be done only to a limited extent, so possible ignition sources must also be identified, eliminated, or isolated from the existing combustible substances by bulkheads. Some examples of ignition sources are [135]:

- Open flames
- Sparks
- Compensating currents
- Spontaneous ignition
- Hot surfaces
- Electrostatic charges
- Chemical reactions
- Vibrations, shock waves
- Thermal manifestations of mechanical energy (e.g., friction)
- Thermal manifestations of electrical energy (e.g., overheating)

Ignition sources present during maintenance and repair operations also merit special attention. Sparks from welding, burning, and abrasive cutting have repeatedly caused fires. It is therefore desirable to prescribe the necessary safety precautions in safety certificates (see Section 4.4).

Smoking must be expressly forbidden in plants and warehouses. The creation of special smoking areas is preferable to concealed smoking in no-smoking areas. Appropriate containers, e.g., self-extinguishing wastepaper baskets, must be furnished for the disposal (daily if possible) of combustible wastes.

To avoid ignition by lightning, all buildings must have lightning protection. Potential compensation between large-area metal structures should be kept in mind, as should the use of overvoltage protection.

Critical effects can also follow from reactions between combustible substance and a fire-extinguishing agent. If incompatibilities lead to restrictions on extinguishing agents, these must be identified and integrated into the planning of organizational and technical fire-fighting procedures. For example, the reaction of fire-fighting water with a combustible substance may release not only flammable gases (from potassium, alkali metal compounds, etc.) but also corrosive vapors (HCl from chlorosilanes, acetyl chloride, etc.). Ignition by static discharge has caused accidents where CO_2 or medium-expansion foam has been improperly used [133], [136].

3.4.1.2. Limitation of Fire Effects

The principal effects of fire have to do with the toxic action of the smoke, and the high temperatures. To minimize harm to humans and animals, the environment, and property, a fire-protection concept must be developed for every industrial facility. The following sections identify the measures that must be included in such a concept.

Design und Construction. Access to the plant must be guaranteed by linking it to roads, streets, and fire service entries. Staircases – their number and placement dictated by the permissible lengths of escape and rescue routes – must give access to interior areas of the building. Escape and rescue routes simultaneously serve as lines of attack for firefighters, and must therefore be built to remain safely usable for an extended time. In modern industrial buildings, this is achieved by erecting massive stair towers (turrets) or administrative service wings, which are more or less independent of the supporting structure in the production building proper [137–139].

The availability of fire-fighting water is insured by a main supply designed to serve the needs of the entire developed area, and a plant supply designed to meet the special needs of the facility.

Fireproofing of the supporting beams of the building should be suited to the possible thermal loads in the anticipated fire. Simple approximations can be used for rough estimates (e.g., DIN 18 230), or more exact thermal-balance calculations can be performed. Regardless of the result, the use of noncombustible structural materials, or at least materials that burn only with difficulty or are difficult to ignite, is preferable.

An effective way to reduce temperature stress is to build in smoke and heat exhausts. Smoke extraction devices, which have to be opened mechanically, serve chiefly to remove cold smoke, and thus protect escape and rescue

routes; heat exhaust systems are intended to remove hot fumes, and thus reduce the thermal stress on the supporting members of the building. Materials that fuse in a short time can be employed effectively and economically in these devices [140], [141].

The subdivision of the plant building into fire compartments or fire-fighting compartments should prevent unacceptable spreading of the fire. This is accomplished by walls and ceilings rated for a specified fire endurance (commonly 90 min). Openings in these members must be closed off appropriately; this includes doors, windows, ventilation ducts, cable and pipe penetrations, openings for tracked conveying systems, and other weak points [139], [142].

When a building is compartmentalized in this way, care should be taken to separate areas allocated to uses, e.g., administration, production, technical zones, and storage. Storage of starting materials and end products in production areas should be avoided.

Retention of contaminated fire-fighting water must form part of the design concept for fire protection. Retention is implemented, appropriate to the purpose of the building, the water contamination class of the substances, and the protective measures employed. If facilities are available for intercepting and removing contaminated fire-fighting water, normal water-retention facilities can be erected for more than one plant or storage area [143].

Fire-Protection Equipment. When fire compartments must be very large, for reasons dictated by production aspects, or when very large quantities of combustible substances or high fire risks are present, fire-protection equipment must be installed. The selection of such equipment is governed by the fire risk. The fundamental distinction is between room-protective equipment, which covers the entire enclosed region or fire compartment, and local protective equipment, whose action is restricted to a particular unit [144], [145]:

Early fire-detection devices are connected to a continuously staffed central office (that of the fire-fighting service if possible) and give a prompt alarm to the personnel assigned to damage control, so that the time allowed for the fire to develop undisturbed is cut short. Depending on the combustile material present, smoke, heat, and flame detectors can be used.

Automatic fire-extinguishing systems involve more than mere early detection. They are connected to a continuously staffed central office, and not only send an alarm to fire-fighting personnel, but also take immediate fire-fighting action. A properly designed and constructed system can suppress a nascent fire or block its progress, so that the fire-fighting service has a better chance of success.

The design of the fire-extinguishing system must include the proper extinguishing agents carefully matched to the type and quantity of combustible substances. Restrictions on extinguishing agents, e.g., because of critical reactions of stored substances with water, must be observed.

A sprinkler system consists of fixed piping with spray nozzles, closed off by heat-sensitive elements (liquid-filled glass vessels or fusible closures, with triggering temperatures ca. 50 – 260 °C). If a fire starts, the elements above the fire site experience a thermal load, which opens the outlets of the spray nozzles so that extinguishing agent is supplied to the fire site. Because only the nozzles above the actual site of the fire open, damage by the extinguishing agent is limited. Applications include piece-goods storage areas and production areas with large quantities of combustible substances. Water sprinkler systems are suitable mainly for Class A solid fuels. Sprinkler systems with a mixture of water and foaming medium can be employed with Class A solid fuels and with combustible liquids (Class B).

A water-spray or deluge system consists of a fixed pipe grid with open nozzles. In contrast to the sprinkler system, extinguishing agent is delivered through all the nozzles at once, over the entire area of the system. The extinguishing action is stronger than that of sprinklers, but the possibility of secondary damage is greater. This type of system is generally used only where the fire spreading rate is extremely high, e.g., transformer installations or warehouses containing substances that burn very rapidly or explosively. Water and mixtures of water with foaming media can be used.

Carbon dioxide fire-extinguishing systems deliver CO_2 through piping systems with spray or "snow" nozzles. The high concentration of ex-

tinguishing agent required means that the compartment must be isolated by bulkheads in case of fire, so systems of this type are preferred for smaller spaces.

Their use is essentially restricted to combustible liquids and electrical equipment. Carbon dioxide does not penetrate to a great depth, and so there are problems in putting out solid (smoldering) fires by this method. CO_2 is useless against metal and carbon fires, because the high temperatures that may occur in such fires can break down CO_2 to carbon monoxide and oxygen, and the liberated oxygen can promote the development of the fire. Certain metals such as magnesium can reduce CO_2 to carbon, e.g.:

$$CO_2 + 2\,Mg \longrightarrow 2\,MgO + C$$

so these metals can continue to burn, even after CO_2 has displaced oxygen.

The toxicity of the extinguishing agent means that special measures must be taken for personnel safety (redundant warning systems). CO_2 fire-fighting installations are increasingly used for local protection, e.g., in laboratory fume hoods or computer centers.

Powder extinguishing systems can be used primarily as local protective devices for special applications, e.g., to extinguish compressed gases or organometallic liquids. This type of system has the advantage that its use is largely independent of temperature (−50 to +60 °C). Powder extinguishing agents are based on sodium hydrogen carbonate, potassium hydrogen carbonate, potassium sulfate, and other chemicals. Sodium chloride or potassium chloride powder can be used against metal fires.

Halon extinguishing systems can be used only with special approval because of current bans on these halogenated hydrocarbons; their application is limited to military uses, and on aircraft, spacecraft, and ships.

Semistationary, nonautomatic fire-extinguishing systems have recently come into increasing use alongside automatic systems. In such a device, the extinguishing agent is delivered through fixed piping from mobile supply units operated by the fire-fighting service. This kind of system is effective only when it is carefully matched to the fire service's equipment; it is used chiefly where the plant has its own fire-fighting unit. The same extinguishing media can be used as in automatic systems.

Dry or wet standpipes are almost always required because they markedly improve the performance of the fire-fighting service.

Fire-Fighting Methods. Industrial fire-fighting includes both the training of employees to carry out immediate response, and the installation of fire extinguishers or wall hydrants to deliver suitable extinguishing agents.

Establishing a properly trained in-plant fire-fighting service improves efficiency; training and equipment should be similar to those of the municipal fire services. Professional industrial fire-fighting services differ from in-plant services in that they are recognized by government agencies, and are trained, certified, and equipped to the same level as municipal fire departments. They represent the highest level of fire protection in industrial plants. Their value stems from local knowledge of the personnel, special equipment tailored to the plant, and the fact that they can work continually to provide high-quality fire-prevention and fire-protection services.

An essential aspect of industrial fire precautions is continuous firefighter training, together with combined exercises with plant personnel [145].

3.4.2. Explosion Prevention and Protection

3.4.2.1. Introduction

Flammable substances (gases, liquids, and solids), if in the form of gas, mist, or dust and mixed with a gaseous oxidizer (most commonly atmospheric oxygen, but other substances such as pure oxygen and chlorine can serve), can form an explosive mixture. The explosion of such systems is generally a fast, exothermic oxidation – a gas-phase chain reaction with chain branching [146]. Special aspects are presented by substances that are not only oxidizable in the gas phase, but can experience exothermic, explosive decomposition reactions without other reactants; examples are ethylene, ethylene oxide, and acetylene.

A gas-phase explosion is initiated by an ignition source, delivering sufficient energy and having a suitable energy distribution. The reaction initiated propagates spontaneously through

the mixture. Such combustion reactions are accompanied by the release of a large quantity of energy with increases in temperature and pressure (and often also by the formation of dangerous reaction products). The hazards associated with an explosion are thus governed by three factors :

Occurrence of an explosive mixture
Presence of an effective ignition source
Effects of an explosion

These factors provide a logical structure for explosion prevention and protection in the following order:

Hazard identification
Risk assessment
Identification of countermeasures, risk reduction

3.4.2.2. Hazard Identification

(see also Sections 2.2.2, 2.2.3)

In gas-phase explosions, it is not the flammable material on its own that represents the potential hazard, but its contact or mixing with an oxidizer. The first step in identifying the hazard is therefore to define the combustion properties of the substances. The result indicates whether, and under what conditions, the substances can give rise to an explosive gas mixture. These conditions are characterized by data such as flashpoint, lower and upper explosion limits, and limiting oxygen concentration.

The second step is to determine the (minimum) requirements for the explosion hazard to be activated, i.e., the ignition characteristics of the system. These include the minimum ignition temperature for a dust layer, the ignition temperature of a flammable gas or liquid, the minimum ignition temperature of a dust cloud, and the minimum ignition energy.

The third step, focusing on the behavior of the explosive system after ignition, provides information on the expected physical explosion effects. Of interest are the heat of combustion, the explosion pressure, the rate of pressure rise, and the maximum experimental safe gap.

After the characteristics of the substances are known, this information has to be linked to the dimensions and parameters of the real chemical process and plant. Since substance data are obtained under standard laboratory conditions, it is important to know how they depend on factors such as temperature, pressure, volume, and volume of the process and plant under consideration. The basic transformation rules are qualitatively summarized in Table 11 [147].

It is necessary to compare substance properties with parameters of the chemical process and plant, not only for normal operating conditions, but also for failure states of the process and of the plant. Such analysis is particularly important for identification of process and plant conditions that may lead to the release of energy (i.e., creation of an ignition source) into the chemical system being handled (see the discussion of ignition hazards in Section 3.4.2.3).

3.4.2.3. Explosion Risk

The necessary condition for an explosion in the gas phase is the simultaneous occurrence of an explosive mixture and an effective ignition source. The likelihood of an explosion may therefore be considered as the product of the probability of the occurrence of an explosive mixture and the probability of the presence of an effective ignition source. The risk is defined as the combination of the (relative) frequency of an event occurring and the expected extent of damage; this means that explosion effects must also be considered.

Likelihood of Explosive Mixtures. Even when the chemical and physical relationships are understood in detail and years of practical experience have been accumulated, the likelihood that explosive mixtures may arise cannot be quantitatively evaluated, and it is doubtful whether such an evaluation would be of any use. It is more to the point to define a few qualitative categories that take in the limiting cases. In these categories, the probability of the formation or presence of a hazardous explosive mixture is described verbally [148]. The following zone definitions have been worked out in the course of European harmonization of explosion prevention and protection regulations [149]:

1) The explosive atmosphere (i.e., mixture under normal conditions) is present continuously, for an extended time, or frequently

Table 11. Transformation rules

Substance characteristics ∗	Real plant parameters			
	Temperature	Pressure	Dispersion	Volume
T_f (flashpoint)		+		
LEL (lower explosion limit)	−	−	−	0
UEL (upper explosion limit)	+	+	(+)	0
LOC (limiting oxidator concentration)	−	−	−	0
MIT (minimum ignition temperature)		−	−	−
MIE (minimum ignition energy)	−	−	−	0
p_{max} (maximum explosion pressure)	−	+	0	0
$(dp/dt)_{max}$ (maximum rate of explosion pressure rise)	+	+	+	−
MESG (maximum experimental safe gap)	−	−	(−)	(0)

∗ Laboratory data
+, −, 0 indicate parallel, inverse, or neutral effects; i.e., increase in value of parameter results in increase, decrease, or no change in substance property compared with laboratory data.

(i.e., for most of the time in the process under consideration). This zone is, as a rule, limited to the interior of apparatus: Zone 0 for flammable gases, vapors, and mists; Zone 20 for flammable dusts.

2) The explosive atmosphere may be present occasionally during normal operation: Zone 1 for flammable gases, vapors, and mists; Zone 21 for flammable dusts.
3) The explosive atmosphere is present only rarely and for a short time only (e.g., in the case of an infrequent malfunction or upset): Zone 2 for flammable gases, vapors and mists; Zone 22 for flammable dusts. (Zone 22 also covers deposits of dust that can lead to an explosive atmosphere in the infrequent event of the dust becoming agitated in air.)
4) The "zone-free area" where the occurrence of explosive mixtures is not anticipated, even in infrequent upsets (nonhazardous area).

In countries whose codes are based on United States practice, the classification differs slightly, but the rating arrived at is similar [150].

A mixture of flammable substances with air is explosive only if the concentration of flammable substances lies between the lower and upper explosion limits (see Sections 2.2.2 and 2.2.3 for definitions). With flammable liquids, if the processing temperature is above the flashpoint, an explosive mixture can be produced above the free liquid surface. If a flammable liquid is atomized (to form a mist or aerosol), the degree of dispersion can be sufficient to generate an explosion hazard, even at temperatures below the lower explosion point. The probability of occurrence of explosive mixtures for a given process in a given plant is not constant for all points in the system:

1) Inside the apparatus, this likelihood depends on the substances present, the prevailing process conditions, and the possibility of upsets
2) Flammable gases, vapors, liquids, or dusts can enter the outside atmosphere at openings and leaks (release points) and form explosive mixtures with air (explosive atmosphere)

The frequency of a release in the vicinity of apparatus depends on the procedures used; it is important to retain flammable substances in closed vessels, not just on grounds of explosion prevention, but also for reasons of exposure and environmental protection. Accordingly, portions of the plant containing flammable substances, as well as equipment and piping associated with these plant sections, are designed to be technically tight in relation to the mechanical, chemical, and thermal stresses anticipated under the intended operating conditions. The following examples illustrate the differences between various forms of release [151]:

– Continuous sources (continuous release) include the permanent free surface of a flammable liquid, open to the atmosphere (open vessel or separator, immersion bath, conveyor belts carrying products moist with solvent) at or above the flashpoint.
– Primary sources (occasional release) include expected leaks from flanges, shaft seals, apparatus and connections opened in normal

operation, venting to the atmosphere during filling or as a result of overpressure, occasional upsets such as the failure of an unmonitored cooling system, sampling, and drainage.
– Secondary sources (infrequent release) include cracks in apparatus walls or piping, defective gaskets in flange joints, defective shaft seals, and emergency pressure-relief devices (e.g., safety valves).

Dilution in the air causes the concentration of flammable substances to decline with distance from the release point. The size of an explosion hazard region thus depends on the rate of release and the propagation conditions. The rate of release depends in turn on the geometry of the source and the substance and process parameters, all of which control the mechanism of release (gas release, flash evaporation, spray jet, pond evaporation).

A central problem is to determine the maximum anticipated leak rate [151]. Catastrophic leaks, e.g., a large leak in a pipe or complete pipe rupture, do not fall within the scope of current explosion regulations.

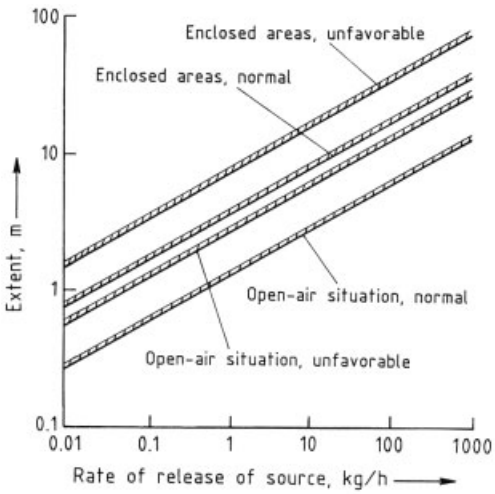

Figure 43. Hypothetical extent of an explosive cloud as a function of release rate of flammable substances for different ventilation conditions
Enclosed areas: normal, $C/f = 5$; unfavorable, $C/f = 0.2$
Open-air situation: normal, $C/f = 100$; unfavorable, $C/f = 4$
Extent $= 2r_z$ (see Fig. 44 for definitions); assumed typical LEL ca. 50 g/m^3

Other expected malfunctions generally lead to leaks up to ca. 0.01 – 1 kg/h. The failure of a seal (leak size in the mm^2 range) results in a release rate up to ca. 10 kg/h for gases or 100 kg/h for liquids, but in liquids it may be that only a fraction of this amount contributes to the explosion hazard.

The main factors governing propagation of released substances are the density of the fluid, the momentum of the release, and the climatic conditions and topography of the surroundings (see Section 1.4.2).

Ventilation conditions can contribute greatly to reducing explosion hazards:

– *Local ventilation.* If the atmosphere is exhausted from the release point (e.g., around the margins of open vessels), flammable substances can be kept from spreading into the compartment.
– *Dilution.* The concentration of flammable substances is lowered by dilution in air so that the concentration may fall below the lower explosion limit.

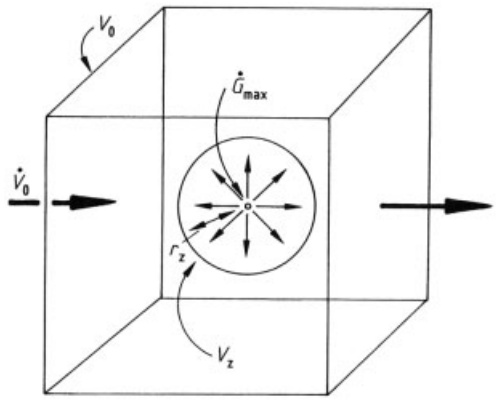

Figure 44. Ventilation model
V_0 volume considered; $\dot{G}_{\max}$ release rate of flammables (mass/time); $\dot{V}_0$ fresh airflow (volume/time); $C = \dot{V}_0/V_0$ air change (1/time); f ventilation quality factor (1 – 5); $V_z = f\dot{V}_{\min}/C$ hypothetical residual sphere of explosive mixture with radius r_z on dilution by fresh airflow; $\dot{V}_{\min} = \dot{G}_{\max}/k \cdot \mathrm{LEL}$ required minimum fresh airflow to dilute below LEL with safety factor k

In buildings above ground with no special air intake or exhaust openings, an airflow of at least 1 h^{-1} is normally maintained. In typical chemical plants with natural ventilation, the temperature differences commonly present between different process apparatus units produce an airflow $> 3\,\mathrm{h}^{-1}$. Below ground, ventilation conditions are generally less favorable; most flammable substances are denser than air, so they tend

to collect in places such as drains or underfloor spaces.

Artificial ventilation moves larger amounts of air, increasing dilution. Air movement can be designed in an artificial ventilation system, with allowance for explosive mixtures with density different from that of air. In an outdoor installation, air movement produces a much more rapid turnover.

Figure 43, based on a simple ventilation model (Fig. 44) [148], [151], shows the characteristic size of an explosive cloud as a function of release rate for various airflows. For a given release rate, an explosive cloud in a closed building is a factor of 3 – 6 larger than in an outdoor plant.

Initially, a hazardous area is created around any potential release source. Figure 45 shows the size of the hazardous areas around a stirred tank opened for charging in a ventilated room. The areas where an explosion hazard exists are identified in the ground plan. Where these areas overlap, the zone of higher probability prevails. For clarity, the hazardous areas have been smoothed to simple geometric forms containing the actual regions. The result is an area classification document. Figure 46 shows such a document for a simplified facility in a compartment with artificial ventilation. In practice, especially in smaller compartments, it is normal to extend, say, Zone 2 out to the boundaries of the room [151], [152].

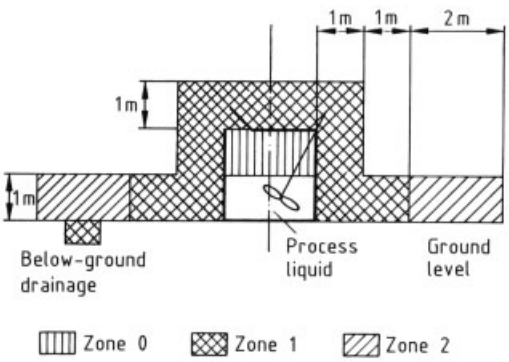

Figure 45. Type, shape, and extent of hazardous areas for a stirred vessel opened during normal operation [151]

Ignition Hazards. In view of the various ways energy can be introduced into an explosive system, 13 types of ignition sources have been defined [148]:

Hot surfaces
Flames, hot gases
Mechanical sparks
Electrical installations
Transient currents
Lightning
Static electricity
Electromagnetic waves (high frequency)
Electromagnetic waves ($3\times10^{11} - 3\times10^{15}$ Hz)
Ionizing radiation
Ultrasonics
Adiabatic compression, shock waves
Chemical reactions

The conditions for such ignition sources to become effective can be generalized in only a few cases. The effectiveness of ignition sources involved in process operation or the working of apparatus depends strongly on the properties of the substances present (see Section 3.4.2.2).

Hot surfaces may be present in normal operation (heated piping) or may arise through malfunctions (increased friction between moving parts). A large hot surface can initiate an explosion if its temperature exceeds the ignition temperature of the fuel that is present. With regard to small (cm^2 scale) hot surfaces, studies have shown that these must be at temperatures much higher than the ignition temperature before they can initiate an explosion [153].

Flames and hot gases are among the most effective ignition sources.

Visible sparks produced mechanically (friction, impact, and grinding) can be attributed to the combustion of metal particles heated to 1000 °C on separation from solid materials. The effectiveness of such sparks as ignition sources depends on the material as well as the ignition temperature of the fuel. For most optimal mixtures of common flammable gases and vapors with air, sparks from friction against steel are effective ignition sources; however, these sparks are effective on dust – air mixtures only if the minimum ignition energy and ignition temperature are low. At relative speeds of ca. 1 m/s, visible frictional sparks no longer occur (Fig. 47). Impact sparks may be effective especially with the combination of light metal and non-stainless steel.

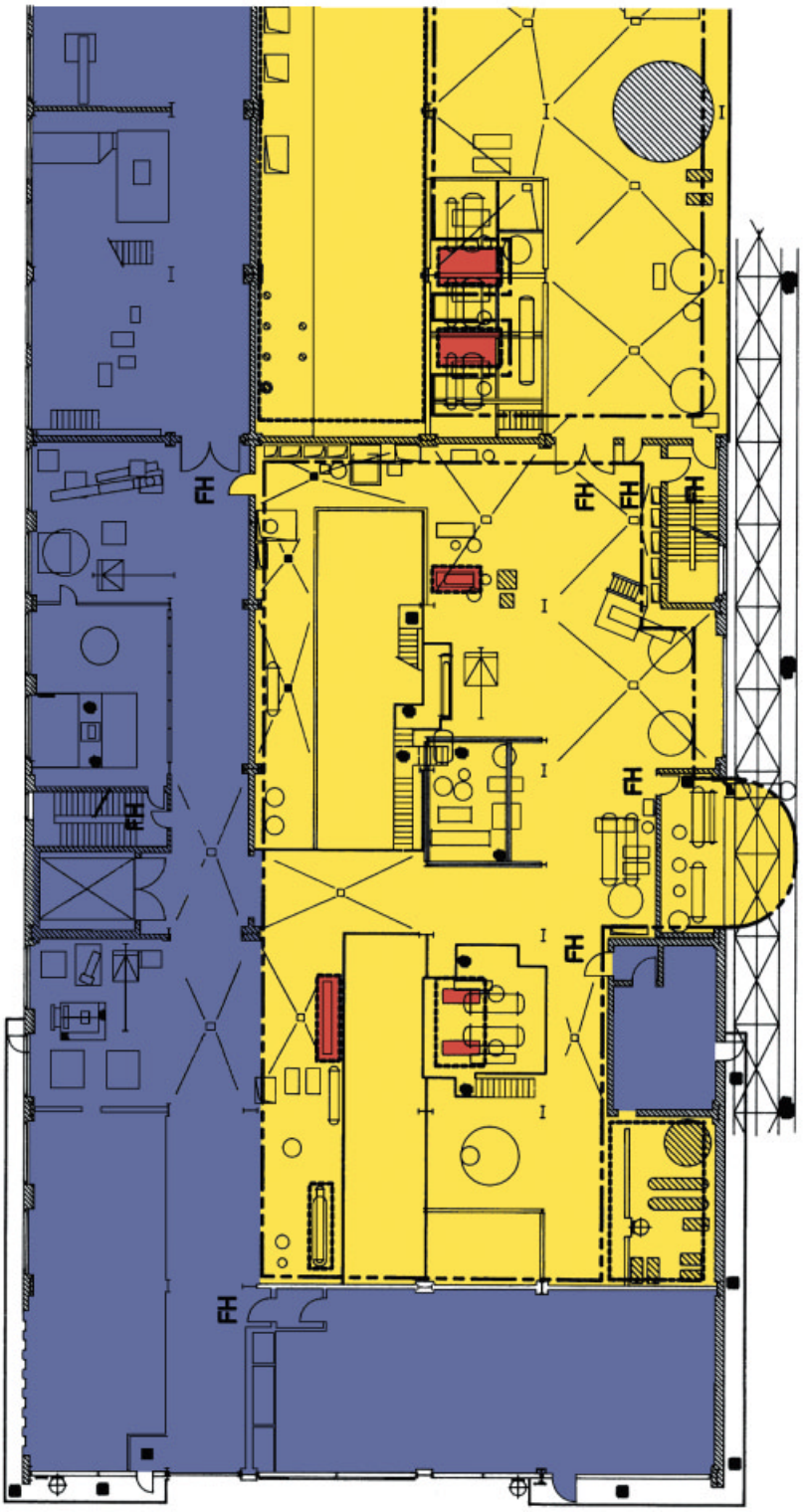

Figure 46. Plan view of the area classification of an in-house chemical plant
Zone 1 enclosed by boundary; Zone 2 by boundary.

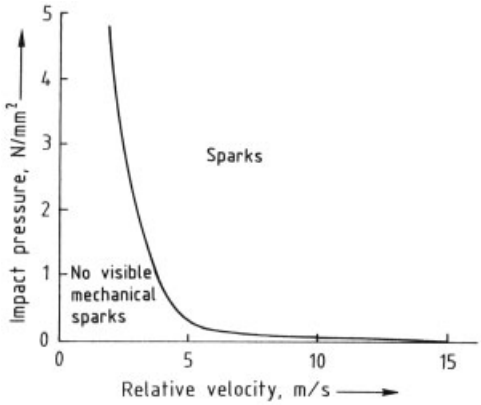

Figure 47. Generation of mechanical sparks
Typical curve for steel construction material [153]

Electrical sparks and hot surfaces may occur as ignition sources in electrical installations. Explosive mixtures can be ignited, even at low voltages. Electrical equipment is considered a priori safe if none of the following values is exceeded: voltage 1.2 V, current 0.1 A, energy 20 μJ, power 25 mW. The use of electrical devices in hazardous areas is strictly (and internationally) regulated (Fig. 48), and as a rule such equipment is rated on the basis of type approval tests or comparable technical documentation, e.g., a manufacturer's certificate [155].

Static electricity as an ignition source is a typical example of how a realistic assessment demands an analysis of the interaction of substance properties and process/plant parameters [156], [157]. Electrostatic discharges occur primarily with substances of very low electrical conductivity; discharges result from fast separation processes at interfaces. Solids with surface resistances $< 10^9\ \Omega$ and liquids with conductivities $> 10^8$ S/m cannot be charged to dangerously high potentials, provided they are grounded. In the case of a liquid capable of holding a charge, charging (e.g., in stirring) can be significantly increased if the liquid is a multiphase system. Dusts often acquire an electrostatic charge during handling. Gases do not become charged, but solid or liquid contaminants or solid or liquid species formed by condensation in flowing gases can lead to an electrostatic charge.

If, say, a pipe wall is sufficiently highly charged, the dielectric strength of the ambient air is exceeded and a gas discharge ensues. The effectiveness of the discharge as an ignition source depends on the form of the discharge as well as the sensitivity of the explosive mixture.

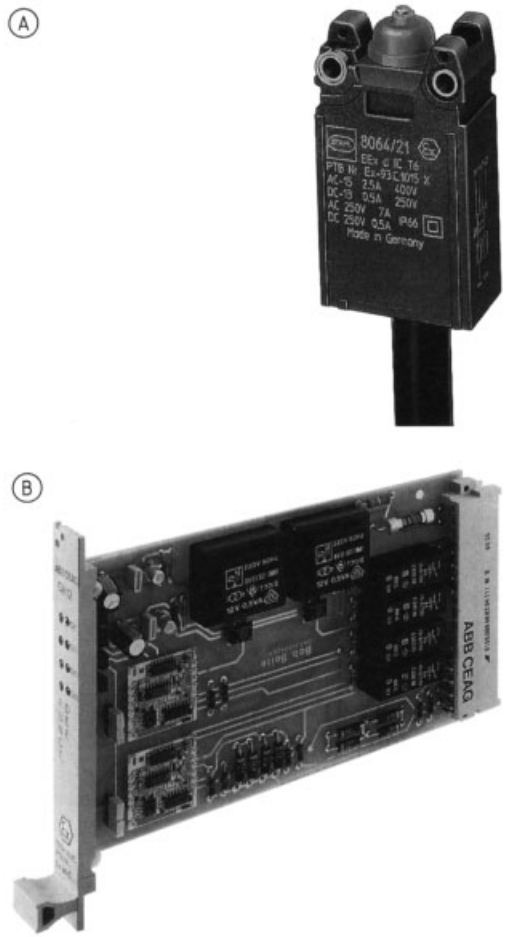

Figure 48. Explosion-proof electrical equipment, tested and certified, with the required explosion-resistant marking and data of testing house (with kind permission of the manufacturers STAHL and ABB CEAG)
A) Limit switch, flame-resistant enclosure "d", suitable for use in zone 1, explosion hazard due to products of Group II C and Temperature Class T6 (e.g., CS_2);
B) Electronic device for PCI, intrinsic safety "i(a)", suitable for use even with zone 0, explosion hazard due to products of Group II C and Temperature Class T 6

Spark discharges take place when charged, conductive objects are brought close together and a discharge channel forms between these "electrodes." Such discharges are effective ignition sources if the energy stored capacitively is greater than the minimum ignition energy of the mixture, as is commonly the case with large objects. Conductive substances (resistivity

$< 10^4\ \Omega \cdot \mathrm{m}$), if isolated, represent an increased ignition hazard.

Brush discharges can occur from charged nonconductive substances. The counterelectrode can be a conductive, grounded piece of equipment or a (grounded) person. The effectiveness of bruhs discharges as ignition sources depends on the charged area; they can ignite gaseous mixtures with a minimum ignition energy of up to ca. 3 mJ (dust – air mixtures cannot generally be ignited by brush discharges).

At pointed electrodes, brush discharges give way to corona discharges; this weakest form of discharge, in its pure form, is not usually an effective ignition source.

If a grounded, conductive body has a thin, chargeable coating, charge can be drained off, and the danger of ignition by brush discharge may be greatly reduced. If rapid charging takes place, as in the pneumatic conveying of dusts, the charge density in the coating can become so high that propagating brush discharges with high energy content can occur (Fig. 49).

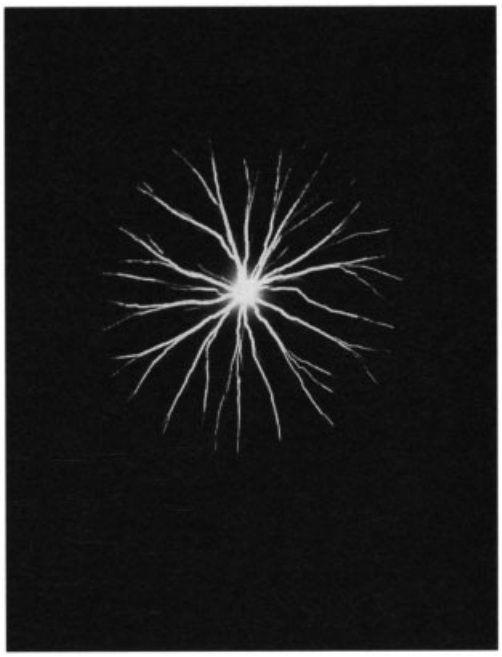

Figure 49. Propagating brush discharge from a chargeable surface [154]

Chemical reactions, by producing heat, can raise the temperature of any substances present, and thus become ignition sources. In highly reactive systems, this can happen through spontaneous exothermic reactions; such systems are said to be auto-igniting or unstable. Self-heating does not occur in less reactive systems under normal operating conditions. If heat removal is blocked, e.g., by (oxidation-sensitive) solid deposits, the temperature can rise to the point that the conditions for ignition are attained.

Other conceivable types of ignition source are less important in practice and cannot be discussed here.

Ignition sources can be categorized in a similar way to explosion hazards (the rating of explosion prevention measures is discussed later in this section).

Explosion Effects. Structures that may rupture in an explosion present a danger to the surroundings, owing to overpressure, underpressure, flames, and expelled fragments. In the field of explosion protection, design data for apparatus are compared with the anticipated maximum explosion pressure. If the possible explosion effects exceed the design limits of the apparatus, and product releases or flying debris from ruptured equipment cannot be ruled out, substantial consequences have to be reckoned with. In general, severe effects will not extend beyond a few tens of meters (see Section 1.4.3).

Rating of Explosion Prevention Measures. In the chemical industry, a distinction is made between cases involving appreciable impacts (i.e., severe personal injury or environmental harm) and those where damage to plant and equipment is the most that can be expected. If no appreciable impacts are anticipated, then economic considerations alone (plant availability, product loss, damage) dictate that an analysis be done to determine how tolerable is the occurrence probability of an explosion. If appreciable impacts cannot be eliminated, the occurrence probability of events must be made low enough that such events are regarded as positively prevented. The following commonly accepted rating system should be applied to such cases:

A) Where an explosive atmosphere is anticipated most of the time (Zone 0 or Zone 20), the occurrence of an ignition source is never tolerable, not even under extreme fault conditions.

B) Where an explosive atmosphere occurs occasionally (Zone 1 or Zone 21) the only ignition sources that are tolerable are those that result from a very infrequent malfunction, e.g., during run-up of a mill rotor (mechanical sparks) or a short- circuit (electrical sparks).
C) Where an explosive atmosphere can occur only infrequently (Zone 2 or Zone 22) continuous ignition sources or those present for long periods or for most of the time, cannot be tolerated. Sources occurring only as a result of a (simple) upset, e.g., a hot surface due to abnormal overheating of a heat exchanger, can be tolerated.
D) In an area where an explosive atmosphere cannot occur at any time, even under extreme upset conditions (i.e., nonhazardous area), the question of possible ignition sources is irrelevant. Even continuous ignition sources remain ineffective, and can be tolerated (e.g., the flame in an off-gas flare system).

For these four categories, Table 12 uses keywords to summarize complementary measures to restrict the formation of explosive mixtures and to avoid ignition sources.

The prerequisite for such a rating (categories B and C) is that the occurrence of an explosive atmosphere and the occurrence of an ignition source are mutually independent; i.e., both do not result from a common malfunction in the process or plant (common mode failure).

The explosion rating of a system states what zone exists for the situation under consideration, and in what zone the ignition sources found would still be tolerable. This safety balance can take different forms, depending on the specific question posed:

1) The ignition source appears to be tolerable in the ascertained category of occurrence of explosive mixtures. The system properties are so favorable that the probability of occurrence of an explosion is sufficiently low. Such a system is said to be intrinsically safe. No further precautions are necessary [158].
2) There is a gap between the category in which the ignition source would still be tolerable and the ascertained zone. Precautions are necessary, but they need only complement the existing intrinsic safety. Three cases can be observed:
   a) For a reduction by one level (e.g., explosion hazard lowered from Zone 0 to Zone 1), a "simple" measure is sufficient
   b) For reduction by two levels (e.g., from Zone 0 to Zone 2), the measure must satisfy extensive requirements
   c) For a three-level reduction (e.g., from Zone 0 to "zone-free"), the reliability of the measure(s) taken must be enhanced even more, e.g., by redundancy

3.4.2.4. Explosion Risk Reduction

There are two options for lowering the explosion risk: measures to prevent an explosion, or measures to limit its effects.

If explosion prevention is chosen, measures are instituted to reduce the explosion hazard (preferable) or the ignition hazard; combinations of both approaches are possible. Limitation of the explosion hazard establishes a zone. If the limitation of the ignition danger meets the requirements for this zone, prevention of an explosion is insured.

Reduction of Explosion Hazard. The avoidance or restriction of explosive mixtures should always be the first priority. Especially outside closed systems, this rule also applies on grounds of occupational safety and health, as well as environmental protection. A simple way is to replace flammable substances with nonflammable ones (e.g., a flammable solvent by water).

Inside apparatus and equipment, the formation of explosive mixtures can be prevented by:

1) Keeping the concentrations of flammable substances in the gas phase outside the explosion range. Gas systems can be operated at overpressure, to prevent entry of air. Suitable purging is needed when such a system is placed in service [159]. In waste-gas streams, the total concentration of flammable components can be held above the upper explosive limit, e.g., by adding flammable gases (enrichment). Conversely, a stream can be diluted with air to reduce the concentration of flammable gases below the lower explosion limit. With liquids, the formation of

Table 12. Keywords for explosion risk assessment

Category	Occurrence of explosive mixtures	Keyword	Tolerance of ignition sources	Keyword
A	continuously/for long periods/frequently	permanently	not even in very rare situations	never
B	occasionally in normal operation	occasionally	not even in rare situations	rarely
C	in rare situations and with short duration	rarely	not in normal operation	occasionally
D	not even in very rare situations	never	in normal operation	permanently

explosive mixtures can be avoided by keeping the temperature sufficiently far below the flashpoint. Substances with low flashpoints can often be replaced by substances whose flashpoints are a safe margin above the ambient and process temperatures.

2) Displacing or replacing air (or other gaseous oxidizing agent) with an inert gas until the oxidizer concentration drops below a case-specific critical level (see Section 2.2.2). A dryer for flammable dusts can be operated with combustion off-gases no longer having oxidizing properties, or a solvent tank can be purged with nitrogen.

In the vicinity of process equipment, the occurrence of explosive mixtures can be largely controlled by:

- Using closed systems to avoid release of flammable substances during normal operation. Examples include gas equilization lines for storage tanks, and the use of lock systems for tank filling and emptying.
- Installing apparatus that has more integrity in terms of possible leaks. Examples include: welded piping to avoid flange joints; O-ring seals instead of stuffing boxes; and materials less susceptible to fracture, such as metal instead of glass.
- "Good housekeeping," i.e., the immediate removal of all flammable liquids and dusts released in the vicinity of plant equipment.
- Creating suitable ventilation to insure the instant dilution of any gas mixtures. This practice can be established throughout the plant (e.g., open-air plant). Selective efforts can also be made to avoid formation of explosive mixtures in areas where sparks can occur, e.g., the installation of an open cage on stirrers and pumps insures that vapors escaping at the shaft seal are immediately diluted with air.

Reduction of Ignition Hazard. The ignition hazard from each type of ignition source listed in Section 3.4.2.3 can be reduced by specific practices. Hot surfaces can be avoided by matching the cooling capacity to the energy dissipated in the apparatus, and providing overtemperature protection. If necessary, insulation can be installed as a form of shielding.

Possible ignition sources involving frictional heat or sparks can be dealt with by suitable design:

- Use of adequately sized components with sufficiently long service lives
- Avoidance of improper material combinations, use of materials with favorable dry friction properties
- Monitoring of temperatures, and of lubrication and cooling systems
- Maintenance of adequate clearance between moving parts

Electrical ignition sources can be avoided by limiting electrical characteristics such as energy and power so as not to produce sparks capable of initiating an explosion ("intrinsic" safety, as in instrumentation circuits) or by preventing extreme conditions and overload in equipment that is free of ignition sources under normal conditions ("enhanced" safety, as in electric power distribution and lighting).

The chief way of reducing ignition hazards due to electrostatic discharges is to use (grounded) conductive construction materials, and to ground conductive objects, liquids, and personal (resistance to ground $< 10^6\ \Omega$). The use of objects made of chargeable, nonconductive substances is generally permitted in Zone 2, but not in Zones 0 and 1. Ignition hazards can be avoided by limiting the dimensions of such objects, depending on the zone, and the ignition sensitivity of substances present; e.g., chargeable surfaces can be kept below $100\,\mathrm{cm}^2$ for substances classified in explosion groups II A and II B in Zone 1. Similarly, the thickness of

chargeable coatings on conductive substrates can be restricted (e.g., < 2 mm for substances belonging to explosion groups II A and II B) and ignition thus avoided.

When working with chargeable liquids, dangerous charging can be avoided by limitations on filling rates (e.g., < 7 m/s for hydrocarbons in a pipe with nominal diameter 50 mm) and by systematic under-level filling.

Ignition hazards due to chemical reactions are avoided by excluding pyrophoric systems from hazardous areas, and maintaining safe process conditions (e.g., layer thickness, temperature, residence time, and thermal load in drying processes).

Limitation of Explosion Effects. Design practices can insure that, if an explosion does take place in an apparatus, it does not have appreciable impacts. Maximum possible pressures generated by an explosion can be determined to acceptable accuracy by thermodynamic calculations, or measured experimentally.

Explosion-pressure resistant design consists in designing apparatus to withstand the anticipated explosion pressure, generally less than 10 times the initial pressure for fuel – air mixtures. If the safety factor relative to yield strength is smaller, the term used is explosion pressure shock resistant design. In piping where detonations are possible, the standard design for PN 10 includes significant hidden reserves that are adequate for nominal diameters ≤ 200 mm; for nominal diameters > 200 mm, the piping should be designed to PN 16 at least. In interconnected vessels, an explosion in one portion can raise the pressure in the remainder; in this way, a (consequent) explosion can be induced, with a higher initial pressure. Explosion protection in such a case may entail decoupling of volumes.

Pressure-relief devices (rupture disks or explosion valves) open as the explosion pressure rises and reaches their response pressure. They insure that the apparatus experiences only a reduced explosion pressure. Such systems are designed especially for protection against dust explosions, where the rate of pressure rise is generally fairly low [160], [153]. Naturally, pressure relief must not create new hazards (see Section 3.5.1).

In an explosion suppression system, a complex of detectors respond at a very early stage in the explosion, and an extinguishing agent (e.g., ammonium-phosphate-based powder) is injected on a scale of milliseconds in order to quench the explosion. The protected apparatus can then be designed for a lower pressure, e.g., 200 kPa. The amount of extinguishing agent needed depends on the violence with which the substance explodes, and the volume of apparatus to be protected (typically 50 kg per 25 m^3 for St 1 dusts; see also Section 2.2.3).

Isolation ("decoupling") in the context of explosion protection means separation of an explosion-resistant section (where an explosion can take place without danger) and a non-explosion-resistant section (where explosions must be positively prevented). A number of devices have been created to prevent flame-front breakthrough; these employ several principles:

- Fast mechanical closure and lock-type features
- Extinguishing of flames in narrow gaps
- Use of flame-extinguishing agents (as in explosion suppression)
- Flame arresting by high counterflow velocity
- Flame arresting by liquid barriers

Equipment selection must take into account substance properties; key aspects include sensitivity to ignition and propagation velocity (see Section 3.5.3). Narrow gaps are also employed in the design of explosion-proof equipment, e.g., fans, gas pumps for recycle systems, submerged pumps, and internal combustion engines [161]. This principle, when used in electrical equipment design, results in the rating "flame-proof enclosure."

3.4.2.5. Legal Aspects

In the domain of European legislation, EU directives set the framework for explosion protection at continental and national levels. The principal directives are ATEX 100 a [162], ATEX 118 a [163] and portions of the Machinery Directive [164].

The ATEX 100 a directive, aimed at manufacturers of "devices and protective systems intended for use in explosion hazard regions," governs the quality and certification of such devices and systems. (For electrical equipment intended

for use in Zones 0 and 1, type testing and a certificate of conformity issued by an authorized agency have long been required [165], according to earlier harmonized European standards.) For Zone 2 equipment, the manufacturer's declaration of conformity is sufficient. Similar rules apply for equipment to be used in areas where there is a dust explosion hazard (this equipment is still classified according to earlier zone definitions with just two types, i.e., Zone 10 and Zone 11, substantially comparable to the "new" zones 20 and 21 or zone 22, respectively).

The ATEX 100 a directive, in contrast to its predecessor, now includes equipment and devices as well as protective systems "which are intended to halt incipient explosions immediately and/or to limit the effective range of explosion flames and explosion pressures," and makes such equipment and systems subject to certification. The pertinent standards are being drawn up or are under revision by the European standards institutions CEN (e.g., committees TC 305 and TC 114) and CENELEC (e.g., TC 31), partly in collaboration with international standards bodies such as ISO and IEC.

The ATEX 118 a directive (in preparation) applies to "the minimum requirements for improving the safety and health protection of workers potentially at risk from explosive atmospheres." It is aimed at plant and process operation and represents a synthesis of earlier national regulations, reflecting the basic principles of explosion protection.

3.4.3. Safety Techniques Based on Process Control

Safety objectives in chemical process plants can be achieved with systems based on process engineering and process control engineering (PCE). Technical or organizational measures, or a combination, can be employed. Alternative solutions that are equivalent from the safety standpoint can be found, so that the most appropriate solution from the process engineering standpoint can be selected.

Safety concepts for specific safety objectives set down in technical regulations must take first priority.

This section deals with PCE equipment for safety in process plants. The discussion is based on NAMUR Recommendation NE 31 [166], VDI/VDE 2180 [167], and provisional standard DIN V 19250 [168], with some points from the book *Praxis der Sicherheitstechnik* (Practical Safety Engineering), vol. 1, *Anlagensicherung mit Mitteln der PLT-Technik* (Safety Techniques Based on Process Control Engineering) [169].

The PCE safety measures suggested in this section are tailored to the risk component that would arise from a process plant in the absence of such measures. This way of assessing risks is only qualitative, and must be carried out by case. Other procedures leading to a comparable safety level are possible.

As used in this section, "safety objective" refers to preventing injuries to persons, major environmental damage, and major equipment damage to property. Safety precautions for electrical installations and equipment, occupational safety practices, or measures taken to safeguard machinery are excluded.

3.4.3.1. Integration of PCE into the Safety Concept

As a rule, safety requirements under DIN V 19250 and the measures required are more stringent the greater the risk that must be covered. The risk must be reduced at least to the acceptable limit (DIN VDE 31000 [170]; see Section 1.3) by non-PCE or PCE safety measures. Non-PCE safety measures and equipment may include elimination of ignition sources, isolation to control explosions, pressure-proof design, safety valves, interception spaces, and retention systems.

PCE safety measures commonly reduce only a portion of the risk resulting from a unit. The risk on which the following discussion is based is just the portion to be covered by PCE safety measures. The remainder can be dealt with by non-PCE safety measures.

Safety tasks allotted to PCE vary widely in character and significance. Similarly diverse are the safety requirements on PCE equipment and practice (technical; nontechnical, e.g., organizational).

The use of PCE equipment for plant safety, the tasks assigned, and their performance are among the matters to be decided in a general

safety review. The following are determined in the safety conference:

- The safety objectives (personnel, environment, property)
- The PCE equipment to be used for plant safety
- The classification of PCE equipment by task:
  - Operating systems
  - Monitoring systems
  - Safety systems
  - Damage minimizing systems
- The required technical and organizational measures (e.g., cycle of functional testing)

The safety review is the basis for the task-driven planning, construction, and operation of PCE safety systems, at a reasonable expense and with clearly defined functional scope.

The selection of measures for maximum simplicity and direct impact generally leads to a safe and economic solution. PCE safety measures find use when other approaches are not applicable, inadequate, or uneconomic, given a comparable level of risk reduction. The economic comparison takes in not only one-time investment costs, but also recurring maintenance expenses.

The requirements for PCE safety and monitoring systems are derived from this definition. Inadvertent triggering of the PCE safety and monitoring systems during any phase of the process must not lead to an unacceptable fault condition of the plant.

3.4.3.2. Classification of PCE Systems

In chemical process plants, PCE systems are classified as operating systems, monitoring systems, and safety systems (Fig. 50), in relation to the ranges of a process variable:

- Specified operation, subdivided into
  - Normal operating range
  - Admissible error range
- Nonspecified operation – nonadmissible error range

Region 1 of Figure 50 corresponds to the situation in which factors inherent in the process keep the process variable from reaching the nonadmissible error range. A monitoring system is sufficient. The process variable is brought into the normal operating range by automatic or (after a status signal) manual action. Region 2 represents the case where the process variable exceeds the limit for the nonpermissible range. Because another protective device is present (safety valve, rupture disk, rapid-opening or rapid- closing valve), a PCE device connected ahead of it to signal or limit the increase in the process variable is classified under monitoring system. In region 3, the PCE device keeps the process variable from entering the nonpermissible range, and is therefore classed as a safety system.

PCE Operating System. PCE operating systems are used in specified operation of the plant in its normal operating range. These devices implement the automation functions required for production: measurement and control of all variables relevant to operation, including related functions such as logging and report generation. High-level control algorithms, complex control sequences, automated recipe processing, and optimization strategies are increasingly employed. Many binary, digital, and analog signals have to be processed if all these tasks are to be performed. Because the functions of PCE operating systems are called on continuously or frequently in operation, these devices are subject to plausibility checking by plant personnel, so that failures and malfunctions can be detected immediately.

PCE Monitoring System. When the plant is in specified operation, PCE monitoring systems responds to conditions in which one or more process variables are outside the normal operating range, but there is no safety reason to discontinue operation; i.e., these monitoring devices respond at the boundary between normal operating and admissible error ranges of process variables.

Acceptable fault conditions of the plant are reported, in order to evoke heightened awareness or direct action by operating personnel; the monitoring system may even initiate action itself to bring process variables back into the normal operating range.

Also included are PCE devices connected ahead of PCE or non-PCE safety systems in order, if possible, to prevent them from responding.

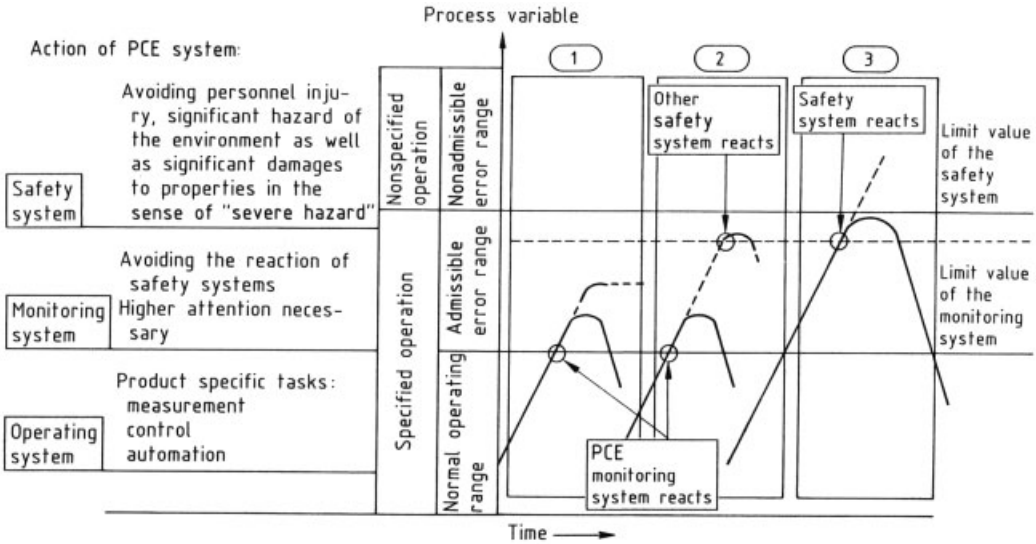

Figure 50. Functions of PCE systems

PCE Safety Systems. In contrast to the functions of PCE operating and monitoring systems, PCE safety devices have the function of preventing nonadmissible error state of the plant.

A PCE safety system is necessary if its absence would allow the plant to reach states that could lead directly to personal injury, major environmental damage, or major equipment damage.

The task of a PCE safety system is usually to monitor a process safety variable and take one of the following actions if the variable goes outside the admissible error range:

1) Initiate a control process
2) Notify the operating personnel (who are always present) so that the proper action can be initiated ahead of time

The functions of PCE safety systems always take priority over those of PCE operating and monitoring systems, and must be executed at a level near the process level with minimum complexity.

The functions of PCE safety systems, in contrast to those of PCE operating equipment, are called on extremely seldom, because the probability of occurrence of the undesired event is low, and because PCE operating, monitoring, and safety systems are often in a staggered configuration (Fig. 50).

Because they are called on so infrequently, it may be desirable to allow sharing of PCE safety components, such as actuators, by PCE operating systems, to enhance availability and facilitate plausibility checking. Such common components must meet the requirements for PCE safety equipment.

PCE safety systems employed for prevention of major equipment damage are designed solely on economic grounds and are not discussed here.

PCE Damage Minimizing Systems. PCE damage minimizing systems come into action: when the plant is in nonspecified operation; when an undesired event occurs. They limit the impacts on persons and the environment, limiting the extent of harm under these extremely rare circumstances.

If PCE equipment is used to detect the undesired event, what is monitored is not a process variable (e.g., pressure, temperature), but some other quantity such as the concentration of gases released in the atmosphere. Effectors put into action do not influence the process, but have to do with the threatened region outside the vessels, apparatus, and piping (e.g., initiation of a water curtain to prevent the spread of ammonia). PCE damage control devices are often combined with non-PCE damage minimizing devices and organizational measures.

3.4.3.3. Requirements for PCE Equipment for Process Plant Safety and Design Principles

PCE Monitoring Systems. No special requirements apply to PCE monitoring devices; they are treated like operating equipment.

PCE Safety Systems. If the interdisciplinary safety review establishes the need to employ a PCE safety system, the following procedure is employed [168]:

1) Qualitative estimation of risk
2) Setting requirements for PCE safety system
3) Definition of technical and organizational measures

In particular, the following steps are identified and documented:

- Task statement, safety problem
- Function of PCE safety system
- Technical design (principle)
- Nature and frequency of scheduled function testing
- Other organizational measures (e.g., scheduled maintenance; see Section 4.4)

Two points must be considered in the selection of PCE safety devices: the minimum risk reduction aimed for; and the safety-related availability of PCE safety systems.

Fundamental Requirement. PCE safety systems must be designed and operated in such a way that if one passive fault occurs, the protective function is still performed.

In the design of PCE safety systems, the safety-relevant availability must therefore be chosen such that the risk is reduced to a residual level lower than the limiting risk, even if one passive fault occurs.

The safety-relevant availability of PCE safety systems depends on:

- The failure rate due to passive faults
- The mean time to detect and remedy passive faults
- The degree of redundancy of the PCE safety system

The safety-relevant availability can thus be enhanced by the following practices, or a combination of them:

- Redundancy: the presence of more operable technical means than are required to fulfill the intended basic functions (e.g., special safety or reliability requirements)
- Homogeneous redundancy: redundant design of a device or parts thereof such that the redundant channels are identically structured, and operate by identical physical processes
- Heterogeneous redundancy or diversity: redundant design of a device or parts thereof such that the redundant channels operate by different physical processes or are differently structured
- Fail-safe quality: ability of a safety device to hold a process safety variable in a safe state, or to modify the variable directly into another safe state, on the occurrence of certain faults in the safety device
- Self-monitoring: a protective device is self-monitoring if, aside from certain faults, it is so constructed that all other faults are detected by the self-monitoring, and thus the safe state is achieved

To meet the requirements on PCE safety systems after risk estimation, distinction is made between high and low risk application [166].

Low Risk Application

Requirement: One passive fault must be detected and remedied within a time interval in which no violation of specified operation is anticipated.

Response: Single-channel PCE safety device with:

1) Short fault-detection time (e.g., high frequency of function testing, continuous plausibility checking); or
2) Low probability of passive faults of the PCE safety device.

A reduced risk may be present if non-PCE safety measures of a technical or organizational nature are employed.

High Risk Application

Requirement: One passive fault must not impair the ability of the PCE safety device to carry out the safety function. Independently of the process behavior, such a fault must be detected and remedied within a time interval in which the simultaneous occurrence of a second, independent fault is not anticipated.

Response: Redundancy of PCE safety device. In general, one-out of-two or (for simultaneously high production availability) two-out of-three design combined with regular functional checking is sufficient. Diversity (e.g., in acquisition of measurements) does not automatically enhance safety over homogeneous redundancy. It is a supplementary measure for preventing possible systematic faults.

Fail-safe or self-monitoring systems are equivalent to redundant devices.

Within a PCE safety system, combinations of these measures are possible depending on the particular availability requirements; e.g., acquisition of measured values may be redundant, control fail-safe, and actuator single- channel.

There are several ways to estimate risk. Quantitative evaluation is not possible for the chemical production plants under consideration here (see also Sections 1.3 and 3.3), because the wide variety of processes and the comparatively short lives of plants mean that adequate statistical material cannot be obtained. This holds equally for the safety relevant availability of PCE safety devices.

Accordingly, the qualitative scheme of Table 13 is used in the design of PCE safety systems. It involves measurements of the risk to be dealt with as well as the safety-relevant availability of the PCE safety systems (the availability of the single- channel device when there is redundancy).

Table 13. Grading of PCE safety systems by risk level and safety-related availability of PCE safety system (availability of single-channel device when there is redundancy)

Safety availability	Risk to be covered	
	Lower	Higher
Higher	I	II
Lower	II	II

Damage Minimizing Systems. Because the undesired event is expected to occur extremely seldom, damage minimizing systems are commonly single- channel units, and must be tested for function in regular intervals.

Principles for Design and Construction of PCE Safety Systems. A number of important principles must be considered in the design and construction of PCE safety systems:

- Proven, reliable hardware and installation methods must be employed.
- The PCE safety device must be simple in construction. Fault effects (e.g., secondary or sequential faults in the PCE safety device) should, if possible, be limited by suitable barriers to fault propagation: high-impedance decoupling, short- circuit strength, galvanic isolation, etc.
- Harmful effects due to environment and products, e.g., vibration, impact, static strain forces, thermal action, corrosion, contamination, wear, and electromagnetic effects (including those resulting from lightning, ripple content in power supply, grid malfunctions, grid noise, etc.), must be taken into account.
- Fail-safe properties of equipment must be utilized (actuator with spring return to safe position, closed circuit to reset, etc.).
- When operating and monitoring systems are shared with PCE safety systems: the safety function must take priority over other functions; and the shared elements must be rated as for the safety device.
- The measurement of process safety variables, processing operations, and the implementation of the safety function must be done with accuracy and speed suited to the safety problem.
- The measurement ranges of process safety variables must be chosen so as to insure adequate resolution. Limiting values must be far enough from range end points so that resolution is still guaranteed if measurement errors are within tolerance.
- The correct setting of limiting values must be protected against inadvertent change.
- As a rule, automatic reactivation after initiation of the safety function should be disabled.
- Whenever possible, process safety variables should be selected such that they can be measured directly, simply, and by a proven method. Indirect derivation of process safety variables by combining measurement signals should be employed only when direct measurement is impossible or unreliable.
- It is desirable to record process safety variables.
- Analog process safety variables should be displayed, together with the limiting values, in the monitoring (control) room, or at local

control panels. In this way, operating personnel can do plausibility checking, so that fault detection times are kept short and the limiting value setting can be checked easily.
- The design of PCE safety systems must also take account of maintenance and start-up needs. Ease of inspection and accessibility of all components of the PCE safety system are important, even in the planning phase.
- Manual overrides can be provided to allow inspection or repair of PCE safety systems during plant operation.
- In isolated cases, redundant PCE safety systems should be examined to determine whether fire hazard or the possibility of mechanical damage necessitates: a split construction; or a protected and/or separate power supply and spatial separation of cable runs for both channels.

Use of Programmable Electronic Systems (PES). Most PCE safety systems are hardwired. The main arguments for this are:

1) The setup must be simple and straightforward.
2) The PCE safety device is seldom, or never, modified while in service.
3) Only a few of the PCE devices in a process plant are safety devices.

In special cases (more complex protective devices), if the use of PES is more economic, the following principles apply:

1) Certified systems can be used for both single-channel and multichannel safety devices in the applications, and under the conditions, set forth in the test certificate.
2) Noncertified systems may be used for safety measures only as part of a multichannel safety device, and then for at most one channel. The other channels must be either hardwired (as required for PCE safety devices) or constructed with certified systems. The following minimum requirements apply to the noncertified system:
   - The system must be proven in former applications
   - Safety and nonsafety functions must be implemented separately whenever possible (in software and, if applicable, in hardware) to avoid influencing the safety section.
   - Along with the standard system monitoring diagnostics (e.g., "watchdog" circuits), which return outputs to the safe state in malfunctions or failures, further system-specific requirements (fan monitoring, climate control, etc., as instructed by the manufacturer) are to be implemented.
   - Shutdowns may be implemented only via binary outputs. Resetting analog outputs to 0/4 mA is not acceptable as the sole shutdown action.
   - Reports of a safety function response must be distinct from reports of operating and monitoring equipment (e.g., different colors).
   - Application software must be created only by trained personnel, written in an easy-to-follow form (structured, modular software), and easy to test.
   - The safety portion of the application program must be subjected to initial examination by the persons involved, and by a specialist, along with the planning documentation.
   - When major changes are made in software, even in the non-safety-relevant portion, a functional test of the safety-relevant functions should be performed.
   - Scheduled functional testing should also include the current software status (application software and firmware), including software documentation.

Labeling. All important components of the PCE safety system must be labeled as such in the documentation, locally in the instrument and the control room.

Testing Before Commissioning. Before the PCE safety system is first placed in service, a test must be performed to determine whether its design and function comply with the provisions of the safety review. At this time, the documentation of the PCE safety system and the test instructions for periodic testing must be available.

Testing must be done in such a way that proper functioning is demonstrated in the interplay of all components. The first functional test must be documented in writing.

3.4.3.4. Operation of PCE Safety System

Organizational Measures. Organizational measures are required for the operation of the PCE safety systems. They fall into four groups:

Continuous monitoring (surveillance)
Inspection (functional testing)
Maintenance
Repair

Continuous Monitoring. Malfunctions of PCE safety systems must be detected by regularly observing process variables, and checking them for plausibility. Such observations are done by qualified personnel.

Outwardly noticeable defects or damage to the PCE safety equipment or their installations are identified by regular visual inspections, and must be immediately repaired.

Functional Testing. Functional tests are needed in order to reveal passive failure. Test instructions must be prepared, summarizing the nature and extent of periodic testing. These instructions must contain information on the nominal state and nominal performance of the safety device, along with a description of the properties and functions to be tested. This includes, in particular, statements of limiting values and measurement ranges, and other specified features to be inspected (such as actuating times for valves, time lags for trigger signals, and other properties important for the performance of the safety task). The testing procedure must be described in test instructions (e.g., checklists) that the checking personnel can understand. The limiting values must be documented in writing by the operations manager.

The test cycle is set in the general safety review. Differences in availability may dictate that some parts of a protective device be tested more often than others.

If no comparable experience exists, an appropriately short testing interval should be set at first. If the tests reveal sufficient safety-related availability, the test interval can be lengthened as the time in service increases.

By analogy with pertinent technical guidelines, inspection of the entire PCE safety system (from sensor to actuator) must take place at least once a year. The operations manager is responsible for seeing that functional tests are performed.

It is desirable that testing be done under conditions corresponding to the demand case and with the least possible modification of the PCE safety system. Whenever modifications are necessary to perform the check, special care should be taken to restore the PCE safety system to its proper state. If manual overrides are installed in multichannel PCE safety systems, only one channel at a time may be bridged.

Care must be taken that neither safety nor availability is substantially impaired during the test. The method used should start with the assumption that faults may be present in the safety device; thus suitable (e.g., organizational) measures should be taken to insure that the plant remains in specified operation.

Further details on testing, especially testing methods for PCE safety devices, are explained in VDI/VDE 2180, Sheet 4.

In addition, tests should be performed after long shutdowns and repairs to the device.

Maintenance. When service conditions are severe, or in the case of certain measurement techniques (e.g., process analytical instruments), scheduled maintenance may be necessary. Work schedules must be prepared summarizing the nature and extent of periodic maintenance, and the required permits from plant management. Maintenance operations are performed by qualified personnel in accordance with these work schedules, and the work must be documented.

Repair. Repair of PCE safety devices must be performed without delay by qualified personnel whenever defects are found in the devices, and there is no alternative that will maintain the level of safety.

Documentation. Testing, maintenance, and repair of PCE safety and damange control devices must be documented. In particular, documentation of functional tests must include at least the following information:

- Identification of test objective
- Results of test, with detailed information on faults corrected
- Date of test
- Signature of tester
- Signature of operations manager

The signatures of the tester (inspector) and operations manager confirm the release and ac-

ceptance of the functional PCE safety or damage control device. The test report must be retained for at least five years so that performance of specified tests can be proved.

Fault Analysis. To improve the reliability of protective devices, any faults discovered must be carefully analyzed; longitudinal documentation of test results reveals weak points. If the same causes are seen to produce faults again, this indicates a weak point and calls for improvement or testing at shorter intervals.

Decommissioning, Restart, and Change of Limiting Values. *Decommissioning and Short-Time Overrides.* If a PCE safety device must be temporarily taken out of service or bypassed, e.g., during plant start-up, this must be documented in writing. Other technical or organizational measures must be put into effect to insure safety while the device is out of service or bypassed. The decommissioning or override status must be clearly identifiable.

Appropriately labeled technical devices, e.g., key switches, can be provided when a safety device has to be bypassed repeatedly. A clear signal must be given to indicate bypassing; an automatic interlocking involving the bypassed PCE safety device must be released if necessary.

There are three options for restart, and these must likewise be documented:

1) Restart without functional testing (e.g., after overriding for calibration)
2) Restart with partial or special testing (e.g., after replacement of a unit, cable, or data line, or after correction of a malfunction). A complete test (as under 3) should be performed as soon as possible.
3) Restart with full functional testing, as provided in the testing specification (e.g., after prolonged interruption of service)

Restart after Initiation of Safety Function. Automatic restart after initiation of the safety function is to be "disabled," i.e., restarting must be prevented when the plant is placed in operation again. This applies particularly when the process safety variable has already returned to the normal operating range. Until the plant state has not been checked portions of the plant affected by the initiation of PCE safety devices cannot be returned to service by manual action. PCE devices used for restoring the PCE safety device to service are considered PCE operating equipment.

Setting Limiting Values. Limiting values for PCE safety systems may be changed only on the instructions of the operations manager, who must also determine the extent to which this action affects the results of the safety review.

A report must be prepared to document the change in limiting value (time of change, authority, person performing the change), and the correct setting of the new limiting value must be verified.

3.4.3.5. Summary

In the technical safety concept, non-PCE and PCE safety measures are employed to reduce the risk arising from a chemical process plant to a value below the limiting risk. The risk to be countered is evaluated qualitatively, and the result is used in "grading" protective measures by scope. Technical measures are implemented as safety systems.

The task of a PCE safety system is to reduce the risk at least to the limiting risk, possibly in cooperation with non-PCE safety systems or organizational measures. From the risk assessment and the associated requirements, safety requirements on PCE safety systems are first determined, and from these are derived the technical and nontechnical measures needed to fulfill the safety function. In a second step, graded measures are devised to meet the requirements. The requirements are more stringent the greater the risk to be covered by the PCE safety system.

3.5. Special Safety Equipment

3.5.1. Pressure-Relief Devices

3.5.1.1. Introduction

Liquids and gases in chemical plants are frequently held in vessels that can be closed. A variety of mechanisms (Section 3.5.1.4) can cause overpressure to develop in such vessels. Initially, the vessel walls withstand this overpressure, but if it exceeds some limit, the "maximum allowable working pressure," which depends on vessel

shape, wall thickness, and material properties, the vessel bursts and the fluid contained is released, unless some sort of relief device limits the rising pressure to the maximum allowable working pressure.

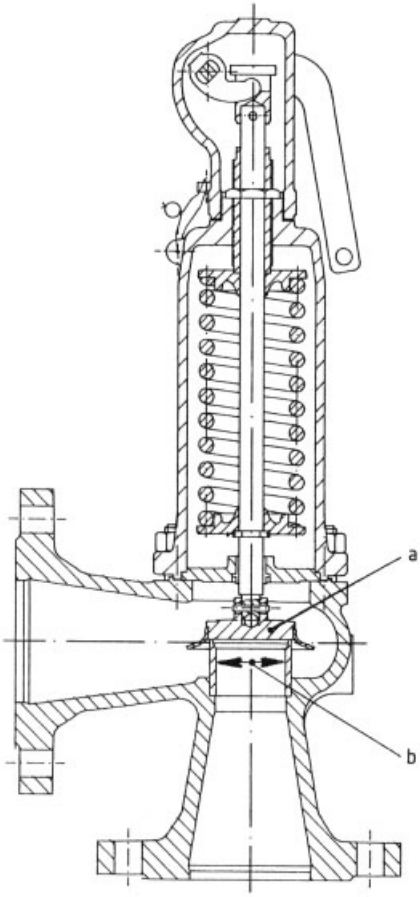

Figure 51. Safety valve
a) Valve disk; b) Flow area

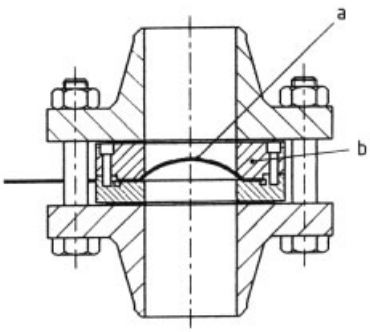

Figure 52. Bursting disk
a) Disk; b) Holder

Classical devices that act automatically to prevent excessive pressure in pressurized compartments are the safety valve and the bursting disk (Figs. 51 and 52) [171–176].

Vents on tanks are operating equipment, not actual safety devices. Not until modern plants were built did their function change as artificial ventilating and gas-displacement systems were installed on tanks.

3.5.1.2. Safety Valves

A safety valve opens automatically when the set pressure p_0 of the valve is exceeded. The valve disk is lifted by an amount h and makes the flow area available for pressure relief. In general, the full lift is attained within a pressure rise of no more than 10 % above the set pressure. The disk reseats when the pressure has dropped below the set pressure. In this way, the amount of vessel contents released is only that needed for instantaneous pressure limitation.

The "opening characteristic" of a safety valve is described by the lift of the valve disk as a function of the pressure p in the vessel. This characteristic depends on the interaction of opening and closing forces on the disk. Opening forces are the pressure force on the disk and, after the disk begins to lift, the fluid dynamic forces exerted by the escaping fluid. These fluid dynamic forces are a function of the lift and strongly influenced by the geometry of the disk, which is designed to achieve a suitable opening characteristic. Closing forces are a constant weight or a lift-depending spring force and the back pressure on the valve disk, which is the sum of an external "superimposed" pressure and the pressure generated by the escaping fluid in the discharge line after the valve opens.

Figure 53 illustrates the interaction of opening and closing forces. Where the opening forces, as a function of disk lift at constant pressure, increase more slowly than the spring force, the valve can open only when the pressure rises further. In this range (A – B in Fig. 53), it opens continuously. If the opening forces increase more rapidly than the closing force (to the right of B), the valve opens abruptly along the constant-pressure curve, up to the next equilibrium point D or the full lift limit C. Similarly, as the pressure declines, the valve reacts by closing

continuously from C (or D) to E and on to F. It then closes abruptly (F – G) because the remaining opening forces at constant pressure and decreasing lift (to the left of F) are smaller than the closing force exerted by the spring. The pressure increase to achieve the full lift between A and C (or D) is the "opening pressure difference." The pressure drop between A and G is called the "reseating pressure difference," or "blowdown."

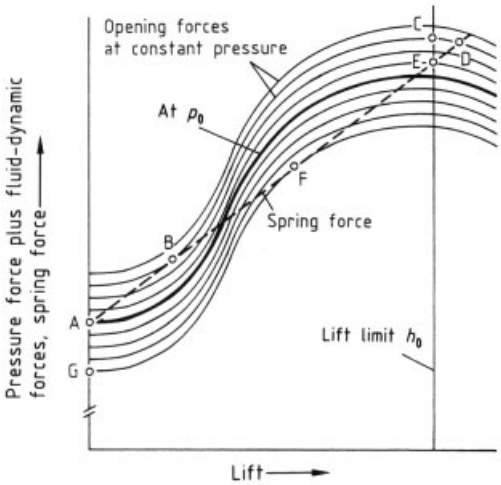

Figure 53. Lift – force – pressure diagram
See text for details

Specially designed safety valves have supplementary loads, which act as a function of pressure to reinforce the closing force until the set pressure is reached, or reinforce the opening force when the set pressure is exceeded. These loads can completely replace weights or springs. In the bellows-type safety valve, a bellows between the disk and housing largely eliminates the effect of back pressure on the discharge side.

Safety valves are classified by opening characteristic and loading principle. The *full-lift safety valve* opens almost abruptly to full lift within a pressure rise of $\leq 5\,\%$ above the set pressure, thereby attaining maximum discharge capacity very quickly. It also closes abruptly. The *proportional safety valve*, in contrast, opens continuously (though not necessarily linearly) as the pressure increases above the set pressure; it also closes gradually. There are no special requirements on the opening characteristic of the *standard safety valve*, and its performance falls between those of the full-lift and proportional types. With *weight-loaded* and *spring-loaded safety valves*, the set pressure must be significantly higher than the normal working pressure in the protected compartment so the valve remains tight during normal operation. The working and set pressures can be much closer together in *supplementary loaded* and in *controlled safety valves*, because the (controlled) closing and opening forces at the setpoint can change almost discontinuously. *Pilot-operated safety valves* have to be classed as controlled safety valves because of their construction and functional principles.

The full-lift safety valve is preferred when large volumes of gas have to be discharged suddenly. The proportional type adapts better to a fluctuating mass flow, and is well suited to low volume flow rates such as those resulting from thermal expansion, and also to variable flows of mostly incompressible fluids (liquids). Supplementary loaded and controlled safety valves are used chiefly in power plants, where large mass flow rates have to be handled and the generally constant service conditions make it possible to run the plant continuously at the limit of its design pressure. Because it permits good sealing of the interior space against the surroundings, the bellows valve is preferred with fluids that must never be released to the environment. It is suitable where back pressure in the discharge line is high or fluctuating, because it restricts the back-pressure effect.

The mechanical reliability and discharge capacity of safety valves are commonly determined by type testing. The discharge capacity is characterized by the certified coefficient of discharge α_w, which holds only for a range of set pressures established by type testing. Normally $\alpha_w = 0.6 – 0.8$ for full-lift valves and gaseous discharge fluids; for proportional valves and liquids $\alpha_w = 0.1 – 0.35$. Against the actual discharge performance, α_w includes a safety margin of ca. 10 %. Unauthorized changes made on a safety valve, affect the basis of the type testing and the discharge capacity is no longer certified. In particular, the installation of an incorrect spring can keep the valve from fully opening, even though the set pressure is correct, so that the required discharge capacity is not available.

For proper functioning, a safety valve must have the correct set pressure, and must be able

to discharge the maximum possible mass flow on opening. Spring-loaded full-lift valves tend to go into self-excited oscillation (chattering) when there are abrupt pressure changes close to the set pressure; the valve opens and closes at short intervals, and the discharge capacity is markedly reduced. Chattering of the valve disk damages the seat and possibly the entire valve, and tightness is impaired. To prevent chattering, the pressure drop in the feed line during discharging must not exceed 3 % of the set pressure. Chattering can also occur if the safety valve is too large for the mass flow rate it is to discharge. If chattering cannot be eliminated, the safety valve can be equipped with a vibration damper. A design modification to limit the full lift is also possible in oversized valves.

The discharge capacity is also reduced if the back pressure in the discharge line is too high. It should therefore not exceed 15 % of the set pressure, or 50 % for bellows safety valves.

3.5.1.3. Bursting Disks

A bursting disk is a closure element that fails when its "bursting pressure" is reached, leaving open the flow area of a discharge line. The design is governed by the construction materials.

Disks can be classified as follows:

Membrane-type disks are loaded by membrane stresses. They fail when the yield point is reached. *Brittle plates* are loaded by bending stresses. They fracture when the breaking stress is reached. The yield point and breaking strength are functions of temperature, so the bursting pressure varies with service temperature. *Shells of revolution* fail by buckling when a stability limit, governed by the modulus of elasticity, is reached (reverse-acting disks). The stability limit is less temperature dependent than the yield point and breaking strength. The bursting of the shell after buckling can be aided by special devices, e.g., a fixed knife in the form of a cross, located behind the bursting disk. The serviceability of bursting disks is determined by bursting-pressure tests on samples.

A bursting disk causes only a slight increase in the flow resistance of the discharge line; it opens spontaneously, and frees a larger flow area than does a safety valve of the same nominal diameter. It is thus particularly well suited to rapidly rising pressures and large volume flow rates. Bursting disks are less sensitive than safety valves to gumming and polymerizing media. They have the disadvantage that they do not reclose, and if a disk fails a large quantity of product can be released. Depending on the design, bursting disks are often sensitive to underpressure in the protected compartment and back pressure on the discharge side.

Membranes are employed chiefly for large flow areas and low overpressures. Panels are used when the corrosiveness or other properties of the medium dictate the use of chemically resistant materials, e.g., graphite, which are not appropriate for the fabrication of thin shells.

3.5.1.4. Sizing of Safety Valves

The protection of a plant section with a pressure-relief device such as a safety valve involves five steps:

1) Definition of scenario
2) Calculation of relief cross section
3) Valve selection
4) Determination of mass flow capacity
5) Impact analysis

The first step in defining the scenario is a systematic safety analysis of the plant section under consideration, e.g., in a HAZOP study. These scenario can be divided into two groups:

1) An energy or material flow is supplied from outside (physical action)
2) The energy or material flow is generated in the system (chemical reaction)

A simple systems analysis makes it possible to break the first group down further:

1) The system is heated. Pressure builds up as a result of thermal expansion, rising vapor pressure, or evolution of dissolved substances. The mass flow rate to be relieved depends on the maximum quantity of heat per unit time that can be supplied to the system.
2) A substance flows from a higher-pressure system into the system under consideration. The mass flow rate to be relieved depends on the rate at which mass can be delivered to the system.

3) A flow of matter is supplied to the system by a pressurizing device, e.g., a pump. The rate at which mass must be discharged depends on the delivery capacity of the pressurizer.

The second group includes systems in which an unacceptable rise in pressure would result if a reaction went to completion. The pressure rise can result from the increase in vapor pressure as the temperature increases, and/or from the liberation of gas, e.g., by decomposition of the condensed phase. Possible causes and mechanisms of such reactions are described further in Section 2.3, while Section 2.3.2 lists methods by which the progress of an exothermic reaction can be determined experimentally. As a rule, studies of this kind are essential in the design of a pressure-relief device. Useful findings include the rate of pressure rise and the rate of heat production. This forms the basis for determining the rate of mass discharge, and hence for designing the safety valve, with the help of fundamental hydrodynamic equations [171], [172], [177].

If the required minimum cross section of the valve is known, a valve is selected such that its smallest cross section is larger than the required value. The cross section gives the maximum rate of discharge through the valve, which makes it possible to calculate the pressure drops on the high-pressure and discharge sides of the valve. These are compared with the limits discussed in Section 3.5.1.2, which must not be exceeded.

It must also be determined whether two-phase discharge can occur. Correlations useful for this purpose have been published by DIERS (Design Institute of Emergency Relief Systems) [178].

Figure 54 illustrates what happens in a pressure-relief event with two-phase flow [179]. When the pressure in the gas space of a vessel is relieved, single-phase discharge occurs first; the pressure drops rapidly. The liquid cannot immediately react to this change of state with a corresponding drop in temperature, and becomes superheated. To restore thermodynamic equilibrium, intensified formation of vapor bubbles begins (after a boiling delay time). The growing and ascending bubbles both displace the liquid and entrain it, setting the contents of the vessel in upward motion. If the free surface of the liquid reaches the discharge opening, a two-phase flow results.

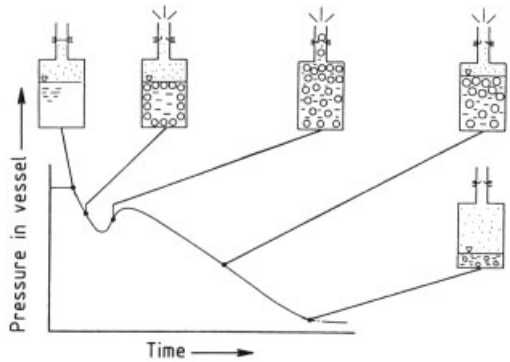

Figure 54. Pressure variation and behavior of vessel contents when pressure on the vapor space above a single-component mixture is relieved

This process depends on both the properties of the reactor contents (viscosity, surface, tension, foaming tendency) and the reactor geometry, fill level, and discharge capacity of the valve [178]. Thus low surface tension and high viscosity promote two-phase flow, while low fill level and a pressure-relief opening small in comparison with the vessel cross section help suppress two-phase flow.

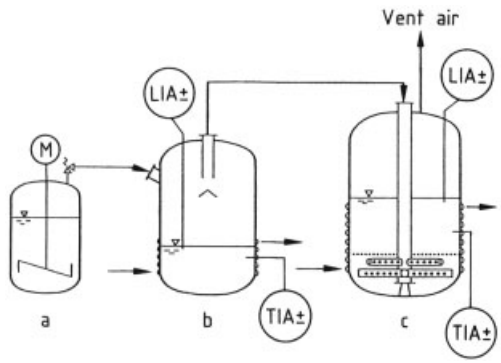

Figure 55. Retention system with liquid separator and direct condenser
a) Reactor; b) Liquid separator with retention volume and cooling jacket; c) Direct- contact condenser with dip tube and cooling jacket

The development of two-phase flow has the consequences that more liquid is removed, more total mass is removed, and the mass flow rate is increased substantially. On the other hand, the gas/vapor mass flow rate that has to be discharged for safe pressure relief is smaller than in single-phase flow. This means that a valve designed for single-phase discharge can be un-

dersized for safe pressure relief under two-phase discharge conditions, and a new test must be performed for this case (the pressure drops on the pressure and discharge sides of the valve must also be remeasured).

The occurrence of two-phase flow can result in markedly higher inertial forces and momentum transfer.

An iterative procedure is usually followed until all constraints are met. A thorough discussion of models to describe two-phase flow has appeared in [180].

3.5.2. Blowdown Systems

3.5.2.1. Procedure

The substances discharged for pressure relief must be handled in such a way that they do not cause personal injury or environmental damage [181]. In general, it is necessary first to determine [178] whether the discharge flow is single-phase, gas/vapor, or two-phase. In the two-phase case, it is expedient to separate the liquid phase [182] and retain it in order to suppress any further reactions. When separation is accomplished in separate vessels, e.g., impingement separators or cyclones, the reaction is stopped by a large quantity of cold water, containing a reaction inhibitor if necessary. On grounds of space and cost, it is desirable to use one apparatus for both functions, if possible (Fig. 55) [183].

The safe removal of substances in gas/vapor form must also be insured. Three approaches to gas/vapor handling are possible, depending on the total quantity of the substances discharged and their properties [184]:

Discharge into the atmosphere
Retention by treatment systems
Retention in closed recovery systems

Substances may be released only if it has been demonstrated for the particular case, e.g., by a propagation calculation [185], that the impact of the released substances on humans and the environment stays within acceptable limits, otherwise, retention in treatment systems is necessary. Industrial treatment systems include thermal cleanup systems, flare systems, scrubbers, and dip-tube and other condensation units.

When systems of these kinds are employed, the first point to check is that the risk potential is reduced to an adequate degree, since both untreated residues and newly formed degradation products are released into the environment. The second point is whether the risk potential is merely shifted; a scrubber may clean up the exhaust air, but then generate highly contaminated wastewater. Finally, it must be demonstrated that the disposal systems are not harmed by the reaction forces and momentum transfer occurring during discharge.

If no satisfactory answers are obtained to these three questions, the substances must be delivered to a closed retention system [183], [186]. In the simplest case, they can be captured in a static vessel, or in systems that can be blown up, but devices of this type may quickly run into practical limits because the volumes of gas to be intercepted are often very large.

One way to minimize the volume needed is to dissolve or condense the discharged substances, or to fix them chemically. Often, this can be done by introducing the gases or vapors into a water-filled vessel [187]. The degree of condensation or absorption can be enhanced, and the necessary volume minimized, by spraying the substances into the vessel through nozzles or distributor pipes [188], [189]. Other variables affecting the quality of condensation or absorption are the temperature of the liquid in the vessel, the fill level, and the pressure in the retention system, which should be as high as possible, consistent with the requirements of the pressure-relief device (Fig. 55).

3.5.2.2. Alternatives

The safe design of a pressure-relief system combined with suitable facilities for handling the discharge can be a time-consuming iterative process. When costly retention or treatment units appear necessary, economic and environmental factors dictate that alternative pressure-relief arrangements must also be evaluated.

For example, pressure relief can be dispensed with if the vessel is constructed so that neither physical nor chemical processes can generate a pressure higher than its maximum allowable working pressure.

In addition, no pressure-relief device is needed if appropriate instrumentation and control [190] (Section 3.4.3) can insure that an unacceptable pressure rise cannot occur. These features must reliably control the reaction, or provision must be made to inject an emergency suppressor or emergency coolant to stop the reaction immediately [191].

Which solution is preferable can be decided only when the specific constraints of the application are known.

3.5.3. Flame Arresters and Explosion Barriers [192–201]

3.5.3.1. Introduction

The initiation and propagation of explosions in (chemical) plants where explosive mixtures are present can lead to significant harm to persons, the environment, and property. One feasible, and often economic, way to control and prevent such explosions in closed systems is to separate parts of the plant with flame arresters. These block the propagation of an explosion, and help limit and control its impacts. This type of isolation can be applied to mixtures of flammable gases/vapors and oxidants as well as explosive dust – oxidant mixtures.

3.5.3.2. Flame Arresters for Mixtures of Vapors (Gases) and Air

In normal plant operation, flame arresters permit the free passage of explosive mixtures. If ignition occurs, however, they block the passage of the flame, preventing the ignition from being transmitted to the protected side. The extinguishing action in present-day flame arresters, regardless of design, is based on one or more of three mechanisms:

Flame Quenching in Narrow Channels. Through intimate contact with the cold walls of a filter element comprising many narrow channels, heat and free radicals are withdrawn from the combustion process, and the flame is extinguished. This mechanism is employed in dry flame screens (e.g., the crimped-ribbon arrester shown in Fig. 56).

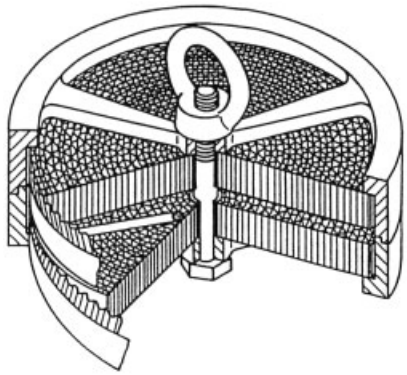

Figure 56. Cutaway drawing of a crimped-ribbon-type flame arrester [201]

Blocking by Liquids. There are two main types of liquid-type flame arresters, the liquid seal in the form of a siphon inserted in a liquid- conveying pipe, and the hydraulic arrester in which a stream of gas passes through a dip tube and is divided into noncoalescing bubbles. The siphon, by virtue of the configuration, is continuously filled with the conveyed medium (Fig. 57). Water is the most common barrier liquid in the hydraulic flame arrester.

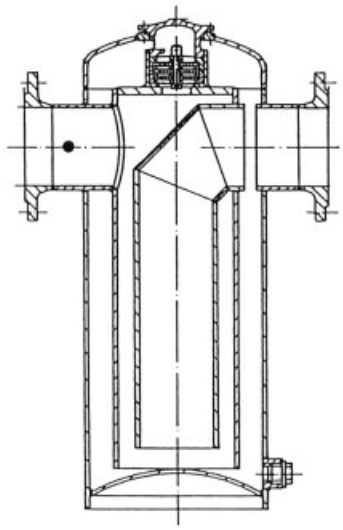

Figure 57. Wet-type flame arrester (siphon arrester) [201]

Dynamic Blocking. The geometry of the flame arrester includes a narrow clearance, in which the flow velocity is always higher than the

turbulent flame velocity (combustion velocity) in the explosive mixture. Upstream propagation of the reaction is thus impossible.

Use of Flame Arresters in Practice. The selection of an arrester is governed by a variety of constraints. Only exact knowledge of the system to be protected can provide the basis for choosing suitable and economical units.

First, it must be established what parts of the plant are to be safeguarded against what effects, i.e.,

> What are the potential ignition sources?
> Is ignition due to external or internal sources?

This is crucial in fixing the location, installation direction, and layout for the arrester. Two applications must be considered for the layout: deflagration and detonation (i.e., coupled combustion and shock waves with rapid pressure rise); the type of protection chosen will depend on which installation is used. Under some circumstances, a run of a few meters of pipe is enough to allow detonation, so that detonation-type arresters are commonly selected for installation in piping. Deflagration-type arresters can be employed where run-up distances are very short, e.g., where tank vents to the atmosphere are to be protected.

In a second step, the construction of the flame arrester is considered. The properties of the medium being transported are important: contamination with solids, solidification temperatures (heatability, frost protection), and condensation or solidification of impurities from the stream.

Quenching-type flame arresters in some services may become contaminated and suffer corrosion; these devices need regular inspection and cleaning. It makes sense to install several such flame arresters in parallel so that one at a time can be bypassed. If even a few of the passages in an arrester become enlarged, or if the flame filter is incorrectly inserted in the holder, the unit may fail.

In the case of pipework not continuously filled with liquid, the wet types of arrester are desirable on grounds of maintenance and availability; these can be obtained in a variety of designs.

Hydraulic arresters are often used to protect contaminated exhaust air streams. Safe operation depends on correct sizing and design, as well as the use of suitable instrumentation for continuous monitoring of important parameters. An automatic alarm should be provided, together with automatic initiation of emergency response when limiting values are violated. Safety-relevant parameters of this type of arrester include a sufficiently high liquid level over the end of the dip tube, limitation of maximum air flow rate, monitoring of liquid temperature (frost protection), and monitoring of temperature in the gas space of the arrester. As a rule, buildup of contaminants and condensing species from the air stream necessitates continuous cycling of liquid in and out of the arrester. To prevent the formation of a continuous explosion channel by the bursting of explosive gas bubbles in the ascending cloud of bubbles – which would allow an ignition to get through the arrester – it is important to have a hydraulic arrester designed for maximum air flow rate and minimum dip depth.

For the correct design of the flame arrester, it is essential to know the maximum experimentally safe gap for the substance being handled. This parameter serves as the basis for classifying substances into Explosion Groups II A, II B, and II C. In Germany, for cases covered by the Flammable Liquids Regulation (VbF), flame arresters must be type-tested (type-approved) for the appropriate explosion group, and for one of the two layouts (deflagration or detonation). Type approval tests are done with appropriate mixtures of fuel and air. Any change in oxidant (e.g., oxygen or chlorine instead of air) and any increase in oxygen partial pressure can lead to a change in performance, which can be assessed only after further experimental tests. Where the VbF does not apply, Germany does not yet have a type-testing requirement.

The quenching of flame in an arrester presupposes adequate removal of heat from the reaction or flame zone. In some arrester designs, handling flowing explosive mixtures, heat removal may be hindered so that a flame can stabilize on the arrester. The filter element and piping may be heated to the ignition point of the flowing mixture, and the flame can be propagated through the arrester. As a rule, the only way to insure adequate heat removal is by using end-of-line ar-

resters. Experimental demonstration of the "endurance burning safety" (ability to prevent flame transmission for an unlimited time) is required. Any flame arrester that does not have this property must be expected to fail after stabilizing the flame for a certain time unless countermeasures are taken. Therefore continuous temperature monitoring (Fig. 58) is required, along with initiation of emergency functions (e.g., cutting off the gas stream).

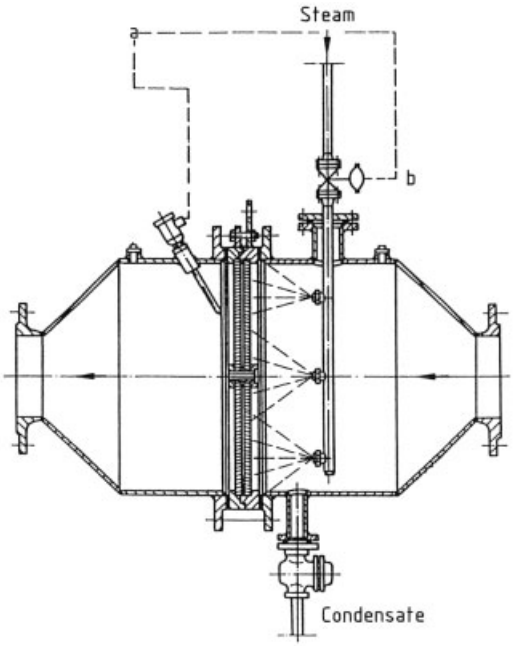

Figure 58. Flame arrester with temperature monitoring and steam injection [201]
a) Electronics; b) Solenoid valve

The use of a flame arrester leads to quenching of the flame at the arrester. However, the pressure generated in the explosion must be safely accomodated through design features (pressure-proof design) in conjunction with adequate provisions for pressure relief.

Safety Concepts. The use of flame arresters to separate plant sections that are not proof against explosion pressure shocks or explosion pressures is just one of a range of concepts applicable to explosion protection.

As is usual in explosion protection, all protective actions are based on risk analyses, which call for countermeasures with appropriate availability, depending on the probability of occurrence of explosive atmospheres (classified in zones) and potential ignition sources. Even when flame arresters are used, the zone and the ignition probability must still be analyzed. As a result of this interplay of occurrence probabilities redundancy of flame arresters is necessary. The important factor is the number of independent protective measures (Table 14).

Table 14. Number of measures to prevent flame transmission

Ignition sources anticipated in plant	Number of precautionary measures		
	Zone 0	Zone 1	Zone 2
Normal operation (e.g., burner flame)	3	2	1
Failures that are likely to occur	2	1	0
Failures that are not likely to occur	1	0	0

German regulations mandate the use of flame arresters to protect non-pressure-proof plant sections in storage facilities and filling points for flammable liquids. Further details can be found in the Flammable Liquids Regulation [196] and the Engineering Code on Flammable Liquids (TRbF).

3.5.3.3. Flame Arresters for Dust – Air Mixtures

The flame arresters that have been discussed are not designed for use in plant areas and piping where there is a danger of dust explosion. Other approaches must therefore be employed in such cases. Among the means used for isolation are star-wheel feeders, active barriers with extinguishing media, rapid- closing valves (explosion valves) and doors, and explosion vents. These methods are described more fully in VDI Guideline 2263 [200].

The use of star-wheel feeders requires prior testing to determine the ability to prevent flame propagation, as well as to withstand the explosion pressure. In order that no smoldering product is transported into downstream parts of the plant, the star wheel must be automatically cut off (e.g., with a pressure monitor) in the event of an explosion.

When active barriers and rapid- closing doors are used, injection of extinguishing agent into the flame front or closing of the doors is initiated in a few milliseconds by a flame detector. Specialists must perform tests to determine

the effectiveness (speed from detection to response) in the real plant geometry, and the ability to withstand pressure. Explosion valves, acting without auxiliary sources of energy, close at the higher flow velocities occurring during explosive pressure relief, automatically blocking off pipe cross sections, and preventing the explosion from propagating.

When explosion vents are used, the formation of long flames and the release of large amounts of product, possibly of environmental relevance, must be anticipated.

4. Safe Plant Operation

4.1. Safe Handling of Chemicals

4.1.1. Introduction

Substances and formulations in the form of raw materials, intermediates, and finished products are employed in many branches of industry. Some substances and formulations have dangerous properties (see also Sections 2.1 and 2.2); these are referred to as hazardous substances.

When hazardous substances are present, safety precautions must be taken so that concentrations stay below the limiting values in the workplace air, so that employees do not come into direct contact with hazardous substances, and so that fires and explosions do not occur.

Measures must be taken to prevent upsets in operation. If upsets come about anyway, the threat to employees must be limited.

Systematic safety analyses should be performed to determine the measures required [202]. Hazard sources and the conditions for their activation are systematically and comprehensively examined so that the hazard potentials can be ascertained. The knowledge gained makes it possible to identify the necessary measures case by case.

Points to be considered in establishing safety practices include provisions of legislation, engineering codes, industry and factory standards, and safety information relating to specific processes.

Technical measures, including the selection of processes with the lowest possible hazard potential, should be favored over administrative measures whenever possible. Preference should go to practices that safeguard the employee or mitigate risks, regardless of employee behavior.

Despite many technical measures, ranging up to "enclosed" plants, contact of employees with hazardous substances cannot be completely prevented. Normal operation entails filling, transferring, and emptying media, taking samples, and cleaning and inspecting vessels. In abnormal operation, for example, substances can get into work areas by leakage and when pressure-relief devices operate.

Selected practices for the safe handling of chemicals are discussed in what follows. These should prevent the release of hazardous substances and threats to employees and the environment.

4.1.2. Normal Operation

Transferring, Filling, and Emptying. In the chemical industry, the handling of liquids is part of daily routine. Liquids must be drained from tanks, drums, and reactors, or masterbatch tanks and reactors must be filled with liquids. It is often necessary to meter exact quantities in specified times [203].

Preventing Hazardous Concentrations of Substances in the Gas Phase. Some ways to prevent hazardous concentrations are:

1) Transferring liquids in systems that are enclosed or can be equalized (e.g., fixed piping from storage tank to plant); to be employed especially when transferring carcinogenic, highly toxic, and toxic liquids
2) Using gas-displacement devices for pumping
3) Extracting vapors at the point of escape
4) Providing adequate ventilation at the workplace

Avoiding Escape of Liquids. While being transferred, liquids can escape by splashing, dripping, spilling, or overflowing. Some ways to prevent the escape of liquids are:

1) Restricting transfer operations
2) Selecting and properly installing suitable filling devices
3) Using appropriate designs for piping disconnections

4) Employing process-specific practices as well as instrumentation and control, examples being:
   Emptying into calibrated vessels
   Emptying through volumetric meters
   Running drain lines to installed scales

Avoiding Electrostatic Charging. Liquids can become electrostatically charged while flowing next to walls or when being sprayed, filtered, or stirred. Charges transported along with the liquid can charge vessels to the extent that static electricity is discharged at the liquid surface; this can result in the ignition of flammable gas – air or vapor – air mixtures. The amount of charging depends strongly on the conductivity, the flow velocity, and the quantity of liquid transported.

A high level of charging must always be expected when interfaces are present in nonconductive liquids, e.g., when immiscible liquids are transported, or when liquid mists or vapors are generated.

Liquids with conductivities $> 10^4$ pS/m do not acquire a charge while flowing.

Avoiding Mix-up of Liquids and of Containers. Dangerous reactions may occur if substances are mistaken for others. Mix-up can be avoided by:

1) Design practices; connections for vessels containing different liquids can differ in diameter from one liquid to another or be made with incompatible flange constructions (grooves, lugs)
2) Unambiguous labeling; vessels, piping, and connections for individual liquids can be labeled, e.g., with substance names, colors, or numbers
3) Identity checking before liquids are drained

Avoiding, Restricting, or Assisting Manual Transport Operations. Ways of avoiding transport operations include:

1) Fixed piping between storage tanks and consumer (reactor, tap); the preconditions for this is that pipe runs be as short as possible, i.e., all parts of the plant are arranged compactly
2) Mechanization of operations

Ways of restricting transport operations include:

1) Purchase of liquids in larger vessels, e.g., 1000-L containers instead of 200-L drums
2) Suitable placement of vessels and other measures to minimize the need for transfers

If manual transfers are unavoidable, off-the-shelf or plant-designed equipment can make them easier and thus less dangerous.

4.1.3. Sampling

Samples are required for the following reasons:

1) Determining identity and quality of feedstocks
2) Monitoring and controlling chemical reactions and other processes in the plant
3) Assessing quality of intermediate and end products

Employees conducting sampling operation may be endangered by hazardous substances and the sampling conditions [204].

Safe Design of Sampling Devices. The following points must be considered in the design of sampling stations:

1) The sample should be withdrawn at a point in the plant where the pressure and temperature are as low as possible.
2) The cross-sectional area of the sampling device must be kept as small as possible.
3) Sampling must be designed such that large quantities of hazardous substances cannot escape from a plant or part of a plant, e.g., because of a malfunctioning or damaged sampling device.
4) The inevitable pre-sample liquid must be returned to the closed system, as far as possible. Two points merit special attention: the sampling point should be designed such that there is no pre-sample flow, or the unavoidable pre-sample flow is safely returned to the processing system.

When the sample is transferred into the sample container, splashing, vaporization, dripping, overflowing, and escape of hot liquids must be prevented. It may be necessary to take steps to prevent static charging, such as:

If possible, avoid sampling from stirred vessels through the manhole (especially if the vessel contains flammable substances). Otherwise, turn off stirrer, sprayers, and other charge-producing components before sampling and wait for a "calming" time.

Do not allow to spray flammable, chargeable liquids into open sample containers, especially if partial vaporization of the liquid is anticipated.

If possible, conduct sampling of flammable, chargable liquids in a closed system only, e.g., in a nitrogen-purged sample vessel.

Limit the outflow velocity of the liquid; rule-of-thumb ca. 1 m/s.

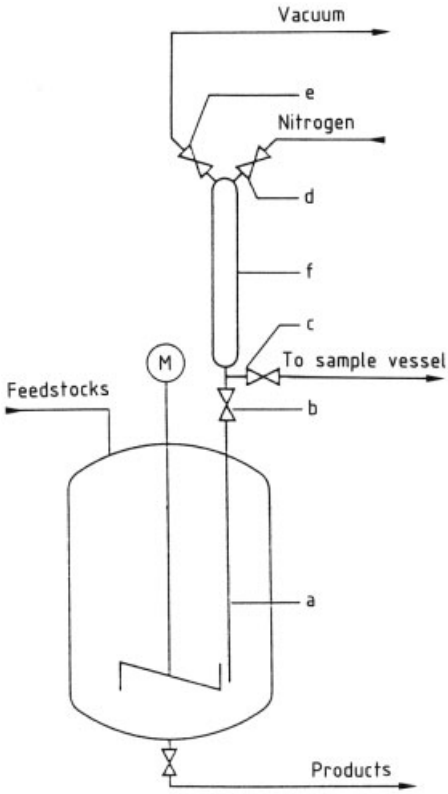

Figure 59. Sampling with vacuum
a) Dip tube; b) Stopcock; c), d), e) Fittings; f) Container

Sampling with Vacuum. Figure 59 shows a setup for taking samples with no pre-sample flow. The procedure is shown by the numbers in the diagram:

1) Open ball valve (b)
2) Blow dip tube (a) empty through fitting (d) (e.g., with nitrogen)
3) Close fitting (d)
4) Apply vacuum to fitting (e) in order to draw product into container (f)
5) Close ball valve (b)
6) Close fitting (e)
7) Apply nitrogen pressure through fitting (c) to transfer sample into suitable vessel
8) Close fitting (c)

Nonreturn valves in the vacuum line prevent the working fluids in the vacuum generator from getting into the product vessel.

4.1.4. Cleaning of Vessels

Cleaning methods should be designed so that worker health is not endangered by residues, contaminants, or cleaning agents [205]. Before cleaning is begun, the status of the vessels to be cleaned must be determined. It must be known what substances, with what hazardous properties, are or have been present (solid deposits, liquids, vapors, gases).

Hazards can arise from:

1) Reaction of the cleaning liquid with residues and contaminants
2) Flammable cleaning agents
3) Static charging when liquids are sprayed under high pressure
4) Materials unstable with respect to the substances, pressures, temperatures, and mechanical stresses present

The proper precautions must be ascertained. Cleaning must be performed by trained personnel under expert supervision, in accord with written information. The cleaning method must be described in an instruction manual.

Whenever possible, closed cleaning systems should be used.

4.1.5. Safety Practices when Working with Hazardous Substances Under Abnormal Operation

Safety practices must prevent employees being exposed to excessive concentrations of hazardous substances, even in abnormal or upset operation. Pressure vessels and pressure piping

must be designed, constructed, and operated in such a way that personnel and third parties are not endangered. The action of pressure-relief devices must not threaten employees or third parties. The quantity and nature of the substances present may make it necessary to conduct releases to the atmosphere or to provide supplementary features, e.g., scrubbers, flare systems, or blowdown tanks.

Leaks represent a further source of danger to personnel and the environment.

Seal systems installed on shafts of pumps, drives, mixers, stirrers, etc. must be selected in accord with the media present and the regulatory requirements [206].

Along with proper design, the effectiveness of the seal systems must be tested regularly, and any leaks must be detected. When media are under pressure, splash deflectors must be used.

Technical protective practices must also be coordinated with administrative and personnel practices, especially for the case of abnormal operation.

4.2. Safety in Batch and Continuous Processes

4.2.1. Safety in Batch Processes [207–210]

4.2.1.1. Introduction

Safe production imposes stringent requirements on the owners and operators of chemical plants, who must observe and implement a body of legislation, rules, and regulations. What is more, the manufacturer of chemical products must make compromises between productivity, quality, safety, and environmental protection, but safety must take priority.

The principles of safety and environmental protection are established by management in accordance with the pertinent laws. Every chemical process involves a special combination of chemicals, equipment, and process conditions.

Safe operation makes it essential to study each process separately. The written set of instructions, the operating manual, occupies a central place in the safety effort.

4.2.1.2. The Operating Manual

Principles. The operating manual is a comprehensive document that fully describes a particular chemical or physical process used in making a product. It must guarantee that the process can be carried on in a safe, environmentally benign, economical, quality-oriented way.

A fundamental requirement is that the manual must describe in full scope and depth the steps that must be performed in the process. It must be unambiguous in that the employee at the workplace knows at all times precisely what actions are to be taken, in what order (e.g., numbering of steps). In extreme cases, every detail of a manipulation may have to be described.

While the manual must be exact and complete, it must also be easily understood. It may be expedient to write separate SOPs (standard operating procedures) when complicated and extended actions must be repeated for each batch (e.g., preparation of a hydrogenator).

Aids. The operating manual is based on a wide range of documents and other information sources (Fig. 60). For reasons of bulk, it cannot include all this information down to the last detail; a vital task of the author is to set priorities and to pick out specific points about a process and set them apart from what is obvious. Operators of multipurpose plants face a special challenge, for they must diagnose continuously varying conditions in the same apparatus and decide what process-specific information and actions are appropriate for each case.

Electronic data processing systems are nearly universal today, and the use of standard forms makes it possible to prepare the operating manual in a consistent way and present it in an easily understood fashion. Modifications and additions can be done quickly (e.g., if production is to be shifted to other plants).

The author of the manual must harmonize process data with plant-specific features while observing all guidelines on safety, environmental protection, occupational health, etc. Many documents and other information sources can be used as aids in describing the process, but it is not enough merely to have these to hand. They must be correctly interpreted and applied. A plant manager often finds it difficult to read documents written by specialists in other disci-

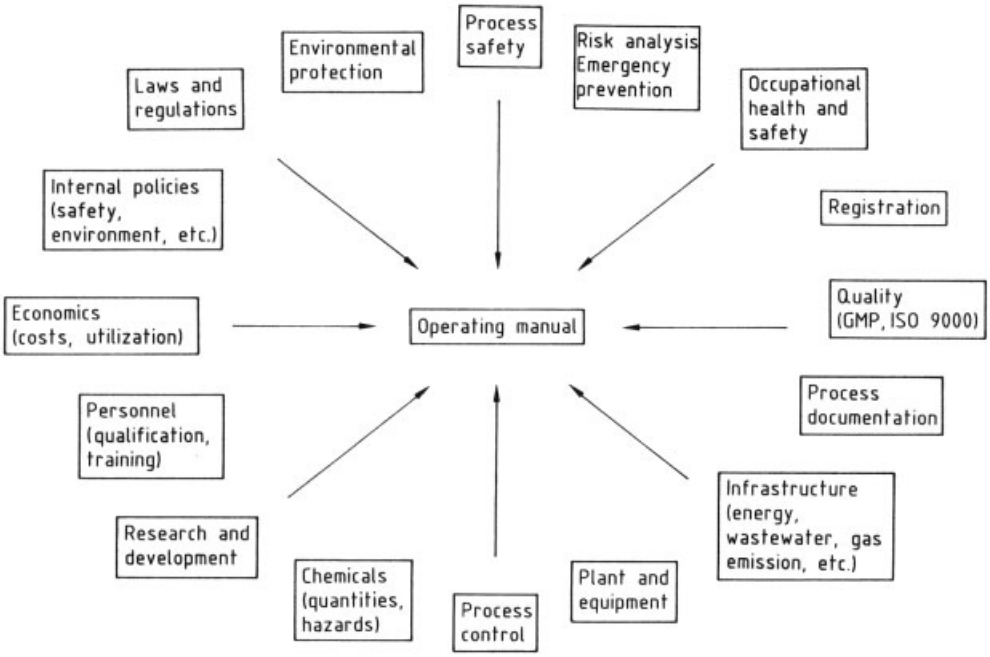

Figure 60. Operating manual: documents and information sources

plines (e.g., some chemists find technical plant flowsheets incomprehensible). Close cooperation with these specialists is needed to insure safe operation. The following information must be checked for completeness:

- Correctly acquired and interpreted material and process safety data
- Knowledge about the plant (cooling capacities, construction materials, plant status, etc.)
- Problems of high-technology facilities (automation, measurement and control, process control engineering systems)
- Limitations of technical safeguards (safety equipment)
- Utilization of infrastructure (energy supply, alarm systems, emergency systems, gas emission, wastewater, etc.)
- Correct and complete risk analysis (process and plant) and transfer of results into practice

For some plants or chemical processes, further points may have to be added to this list.

A good operating manual is never the product of one individual; it can be created only based on effective collaboration among specialists. As soon as the definitive operating manual is in being, the involved departments involved should apply their special knowledge to prepare opinions on the process. It is desirable to set up a fixed routing (toxicology/occupational health, safety, environment, registration, etc.) and have each activity signed off on the manual.

Contents and Organization. The organization of the operating manual is dictated by company policy, and there is no universal rule. The following three-part breakdown has been tested in practice:

1) Information section: intended chiefly for management; a quick overview of the process, presenting all safety-relevant aspects
2) Operations section: focused on the operating procedures; containing all information needed for the practical conduct of the process in the plant
3) Appendixes: data, details, and special information on the process; can be used to gain a deeper understanding of the process

A possible table of contents reflecting this overall breakdown is shown below.

Overall Organization of Operating Manual
Information Section

A. Changes from Precedent Operating Procedures
B. Chemical Scheme of Process
C. Materials (Chemicals, Yield)
D. Environmental and Safety Aspects
  D.1. Safety Data
  D.2. Summary of Risk Analysis
  D.3. Waste Disposal
  D.4. Environmental Impacts (Ecogram)
E. Plant and Equipment
E.1. Flowsheet (Streams in Apparatus)
E.2. Apparatus List
F. Material Data
G. Safety and Occupational Health
  G.1. General Provisions
  G.2. Process-Specific Provisions
H. Response to Accidents, Fire Alarms, Power Outages
Operations Section
I. Process Performance
  I.1. Step-by-Step Description
  I.2. Process Instructions
  I.3. Process Deviations, Remarks
J. Intermediate and End Product Inspection
K. Packaging and Storage
L. Equipment Cleaning on Product Changeover
Appendixes
M. Safety Tests, Wastewater Analyses, Operating Instructions for Individual Apparatus, Distribution List

Safety-relevant items are discussed as follows.

D.2, *Summary of Risk Analysis* gives a quick overview of the risks involved in the chemical process and the apparatus used. Known major risks are described along with key safety practices for the process. These include administrative practices, special technical safeguards, and important limit values (temperature, pressure, etc.).

G.1, *Safety and Occupational Health: General Provisions* lists general directives and rules to be followed in the chemical process. These may relate to reporting of deviations from the operating specifications, grounding, inert gas blanketing, etc. The instructions and rules must be formulated so that they can be observed in practice.

G.2, *Safety and Occupational Health: Process-Specific Provisions* gives detailed information and practices derived from the risk analysis. It may include special characteristics of equipment, special protective measures, handling of hazardous chemicals, increased risks, temperature limits, etc. This section is crucial for the operating personnel, telling how this process differs from others. It is especially important in multipurpose plants where the hazard potential varies continuously.

H, *Response to Accidents, Fire Alarms, Power Outages* is one of the hardest sections to write because problems often occur where they are least expected. With the help of the risk analysis, the steps in the process are analyzed, device by device; the analysis suggests what problems may occur and need to be considered in the planning of emergency response practices. Critical quantities are temperatures, times, and mixing; these conditions have to be determined separately for a variety of potential events. Practices are defined for each step in the process so that any kind of deviation can be responded to and the process returned to a safe condition.

I, *Process Performance* is the heart of the operating manual. The sequence of operations is detailed and described in chronological order. Sequential numbers are assigned to the process steps, which must be formulated in a clear and unambiguous way. The process conditions to be maintained and the quantities of chemicals to be used are written down as nominal values; the person controlling the process must acknowledge these in the form of actual values. When key reactants are present, or when process conditions are especially safety-relevant, it is desirable to confirm maintenance of the nominal values by having a second person (possibly a supervisor) take readings and sign.

In multipurpose plants, the base settings of the apparatus take on central significance for safety. The frequent changes of conditions in a given apparatus mean that adjustments must be made as a function of process conditions (temperature limits, stack gas downstream of scrubber, or stack gas cleanup unit, etc.), and some apparatus may have to be modified (cartridge filters, charging hoppers, etc.). Before a piece of equipment is placed in service, these base values must be set and checked.

The operations to be performed may be identical for a number of apparatus items (e.g., charging of a centrifuge). In such cases, reference may

be made to separate operating instructions or unit operation instructions (SOPs). When this is done, all that has to be set down step-by-step are the parameters specific to the process (e.g., speed of a centrifuge).

I.3, *Process Deviations, Remarks* is a form, to be completed on site for each batch. It helps in documenting irregularities in the process. The plant manager thus has a powerful tool for early detection of safety-relevant excursions; a shift log serves the same purpose.

L, *Equipment Cleaning on Product Changeover* is very important. From the safety standpoint, it is essential to describe the individual process steps chronologically and in detail, precisely as in I, *Process Performance*. It is a mistake to think that cleaning with water is completely harmless; if, for example, warm water is run into a reactor thought to be empty, solvent residues in the vapor phase may reach an explosive concentration. The mechanical removal of tenacious crusts with various tools can be critical if solvent residues are still present. It is therefore important that all safety principles are observed, even when cleaning the equipment.

Appendixes give standard operating procedures (SOPs), checklists, piping and electrical diagrams, etc. The exact distribution list is also important, because such a list is the only way to make certain that changes are made in all existing copies of the operating manual.

Introduction of a New Operating Manual. Careful, comprehensive documentation is just one important condition for safety in chemical production. Equally important is that employees properly understand the instructions. Often, the use of technical language and foreign words in the operating manual places it at too high a linguistic level for the workers. It is also remarkable to see the wide range of interpretations different people can put on imprecisely formulated directions! Another problem arises when some employees are not fluent in the language of the instructions, and thus fail wholly or in part to understand an exact description. Any new operating manual must be presented and explained, one step at a time, by oral instruction. Actions that are difficult to describe in speech or writing must be demonstrated in the plant. The objective of such training is not simply that the staff understand the instructions; as comprehension grows, so do acceptance and safety awareness.

The instruction program must end in a test to ascertain whether the operating manual has been understood. This can take the form of a game or competition, or can be administered as a moderated discussion. Another proven method is to have the employees make a display, illustrating and explaining the steps in the process.

4.2.1.3. The Human Aspect of Safety

Responsibility. The chemical process employee works in an environment with a substantial hazard potential. Every day brings the need to work with harmful substances, dangerous reactions, and large quantities of chemicals. Each employee thus bears a high degree of responsibility for co-workers, the environment, and the plant. Even small mistakes can lead to grave accidents. In multipurpose plants, the situation is even more critical: Processes differing in their demands on work methods and safety are performed in succession. Close concentration and an ability to adapt quickly to new conditions are required.

One of the most urgent tasks for management is to provide each employee with the conditions for safe work and the resources for performing the job. Occupational safety becomes a task for management, to insure that workers are safety-motivated, and to aid them in carrying out their demanding tasks with discipline and self-confidence – and without accidents.

Training. The training of plant operating personnel is a continuing task for plant managers. Thorough training greatly enhances on-the-job safety by helping workers to deal better with the pressure of responsibility, and to react calmly and correctly in stress situations. Training must be related to practice in the plant; an ideal example is discussion of the operating manual before production is started.

Special attention must be focused on the training of new employees, who possess good technical knowledge, but not the specific skills for their jobs. During their first months, these employees must be given technical orientation and support so they can perform in accordance with safety rules.

A particularly important point in training is how to respond in abnormal situations. In the context of the operating manual, responses must be explained in the clearest possible way. Practical and theoretical emergency training (e.g., with simulator programs) help establish a routine for dealing with abnormal conditions. It is important to instill in the worker the self- confidence that will lead to calm, considered actions in a real emergency.

This training is documented and evaluated. Feedback, including worker self-assessment, helps reveal gaps so that the appropriate steps can be taken.

Motivation. People are not inherently motivated to work safely; external motivation must be provided. Repeated use of convenient, but unsafe, work methods, not immediately leading to an accident, gives rise to bad habits. Administrative measures (e.g., checkups) and technical precautions can evoke safe working methods, and make unsafe behavior difficult or impossible. Correct behavior should be associated with all possible advantages: recognition, easier job performance, higher qualifications, etc.

Cooperation. In the course of time, process workers accumulate much practical experience. It is essential for plant management to make use of this knowledge by integrating the workers in the safety process. It is an important distinction whether the wearing of protective clothing and respirators is ordered or self-prescribed by a group within the plant working on a safety problem. Safety awareness, and above all the acceptance of key safety practices, can be markedly enhanced in this way.

4.2.1.4. Internal Organization and Policies

The processes that follow one another in a multipurpose facility call for intensified administrative policies. Without clear direction and clear structures, an operation where products are changed over, say, weekly slides into chaos, and an accident becomes inevitable. In what follows, some points are listed that should help avoid such problems.

Weekly Program. The weekly program provides a quick survey of planning for the next week. It yields information on personnel allocations, production program, maintenance activities, and special events. The weekly program is presented orally to the operating personnel and is posted in the plant. Every worker can then make the optimal preparations for coming assignments. Surprises and improvisation are largely eliminated.

Weekly Conference. All employees take part in a brief conference every week. This is an occasion for comments on the preceding week and the coming one, and is also very important for exchanging safety information. Absentees must be identified and informed later.

Shift Change. A formalized shift change must be instituted, so that the chemical process can be carried on without pause. The procedure should include explicit allocation of tasks, provided in writing if necessary (e.g., in a shift log).

Responsibility. Clear structures in work allocation is essential. The chain of command (hierarchical levels) must be preserved. Unless a situation arises calling for prompt intervention, the plant manager should issue instructions to process personnel only through supervisors and team leaders.

This avoids, the confusion generated by instructions and counterinstructions, and by indistinct areas of responsibility.

Monitoring of Administrative Practices. Procedures on paper are not enough; they must be carried out. Spot checks should be done to find out whether selected points in the operating manual are being complied with. Above all, misunderstandings and gradual safety-relevant changes in work procedures must be detected promptly.

Information Wall. An information or notice wall can be used to promote safety awareness, but this can fulfill its purpose only if it is actually read, so it must be prominently located. Attractive presentation (generous use of color, graphics, etc.) makes it more likely that employees take an interest in the information, but no information wall can replace the oral communication of important instructions.

4.2.1.5. Safety in Production Practice

Before the Start of Production. The saying "as you make your bed, so you must lie on it" takes on great significance when preparations are being made for a new process. The decisions made at this stage must lead to safe, accident-free operation later. Serious preparation is particularly important in multipurpose plants where chemical processes change frequently, so that the plant must continually be modified to fit changed process conditions and operations. Plant personnel must also adapt. There must be a distinct "breathing space" between finished and following production. Overlap allows the possibility of confusion, and increases risk.

Central importance attaches to the production conference with all employees concerned. The operating manual is reviewed step by step and supplementary instructions are given by management. This production conference is an intensified form of the conference discussed under the heading *Introduction of a New Operating Manual* at the end of Section 4.2.1.2. Its timing has much to do with its success; it should be held immediately before the start of production, so that employees can concentrate on the new process without distraction, and get themselves mentally set for it.

The preparations for a production changeover are extensive, but identical in their main features. It is worth standardizing these procedures; ideally with a checklist.

Checklist for Production Preparation

1) Information for leadership
   - Scope and timing of production
   - Study of process documentation
   - Leadership conference
2) Operating manual
   - Check operating manual (validity, changes)
   - Make available report forms
   - Post necessary documents (e.g., chemical data sheets) in plant
3) Feedstocks
   - Verify availability of reactants required
   - Verify quality of reactants
4) Training
   - Conduct process-specific training
   - Discuss process with all employees concerned
5) Plant/facilities
   - Check maintenance completed
   - Carry out functional tests
   - Install additional safety features
   - Make piping hookups according to diagrams
   - Perform leak tests
   - Check cleanness of equipment
6) Trial run
   - Carry out and document trial run with solvent and prepare report
7) Analysis
   - Prepare analytical sampling
8) Administrative
   - Order necessary placards and labels
   - Prepare folders and forms for process documentation
9) Wastes
   - Check disposal system
10) Storage
    - Prepare packages/containers
    - Set aside storage capacity
11) Safety
    - Verify completeness of safety documentation
    - Consider safety and occupational health checks

During Production. In a correctly prepared start-up, the management concentrates on inspections, monitoring work procedures, and verifying process safety. This task includes:

Surveillance of Action on the Shop Floor. It is obvious that a safety problem exists when the author of an operating manual is convinced that the instructions are being strictly followed while, for whatever reason, something entirely different is going on. The comment, "That's what it says in the book, but we've always done it differently," must not be tolerated. Discrepancies should be identified and immediately corrected.

Detection of Irregularities. A permanent task for both leadership and operating personnel is to identify process irregularities promptly; an important aid is a form for reporting process excursions. Prompt response to irregularities is a key element in accident prevention.

Assessment of Process Modifications. If deviations from the process instructions are necessary (e.g., because of defective apparatus), the plant manager must consider carefully the safety

consequences of such a change. If a distillation is carried out in a different piece of equipment under the same conditions, but the reactor jacket offers worse heat transfer, or the stirrer is less effective, the operation may last longer than specified. This type of change can lead to a dangerous condition.

Feedback. To obtain feedback on the practicability of an operating manual, it is essential to make contact with the workers at the workplace. An amazing amount of information and suggestions can be conveyed in a short, relaxed conversation.

Working Conditions. Noise, heat, uncomfortable posture, and other conditions can strongly affect human performance; working conditions should continually be examined and improved.

At Termination of Production. When a process is terminated, extensive process documentation is generated: production reports, analysis reports and certificates, printouts, etc., as well as a report summarizing all vital aspects of the process. This documentation yields information on potentially safety-relevant deviations in process chemistry and technology.

Process Documentation

1) Title/production stage
   Product, reactant, duration
2) Operating manual
   Applicable operating manual, deviations from valid procedure
3) Production
   Technical and chemical feasibility
4) Yields
   Comparison with earlier productions
5) Analysis/quality
   Remarks on product and reactants
6) Modifications/remodelling
   Summary comment
7) Economic aspects
   Calculation
8) Documentation
   Identification of changes, defects, and missing or outdated documents (e.g., safety tests); all changes made in flowsheets and other schemes in the course of repairs
9) Conference notes on termination of production
   Proposals for improving safety and process conditions

Good process documentation guarantees continuity of production with regard to work procedures and safety. Reading these documents when the same process is resumed is an absolute must.

Finally, the operating manual must be revised on the basis of the process documentation and, if necessary, rewritten. This insures that the manual reflects the reality of the plant and is ready to use for a new production at any time.

4.2.2. Safety in Continuous Processes

4.2.2.1. Introduction

A continuously operating production plant usually makes a single product. The steps involved can include both chemical reactions and physical separation operations. The quantity of material involved in both cases can be extraordinarily large, and special care is required, particularly when the substances have dangerous properties (flammable, explosive, toxic, or environmentally harmful). Continuous production plants are highly automated, and are seldom run without process control systems. Apart from start-up and shut-down modes, they operate according to fixed, specified process parameters. Intervention by operating personnel is necessary only during indicated or infrequent deviations from normal operating conditions. Because the process does not change, the danger of neglecting to readjust safety-relevant process variables on a change of product is not crucial. Typical continuous processes are:

– Refinery processes
– Cracking processes for olefin production
– Polymerizations
– Processes for making organic and inorganic bulk products

4.2.2.2. Analogies with Batch Processes

While the objectives, procedures, and plant sizes in the batch case are much different from those in the continuous case, quite a few of the basic statements in Section 4.2.1 also apply to continuous processes.

For example, exact knowledge of the process steps is important for process safety, in particular the acceptable deviations from the specified process parameters. The safety parameters and their acceptable ranges, established before the process plant is designed, must be set forth in an operating manual, and embodied in safe procedures for use by the operating personnel. In the context of risk or safety analysis (Section 3.3), the conditions for the occurrence of conceivable upsets must be identified, the consequences must be analyzed in terms of occurrence probability and severity, and appropriate safety practices must be instituted and covered in detail by bulletins and training of operating personnel. These studies must include not only the failure of technical devices (compressors, pumps, mills, and centrifuges) along with malfunctions of instrumentation and control devices, but also errors made by operating personnel, and partial and total power outages.

There must be a plant emergency plan, prescribing how workers should respond to a variety of abnormal conditions and accidents. A higher-order emergency and hazard- control plan dealing with major accidents, i.e., those that can pose a serious danger to employees, the immediate neighborhood of the plant, or the environment, must establish internal report routing and hazard- control measures. This plan must also designate those accidents in which external assistance, e.g., the public fire brigade or disaster response teams, must be requested, and what public institutions, e.g., the police and regulatory authorities, are to be informed, immediately or later.

Before production is started, operating personnel must receive training and orientation based on:

Process documentation:
- Process flowsheet
- Chemical reactions
- Physical steps in the process
- Parameter values to be maintained
- Safety rules
- Functioning of the process monitoring system

Safety information:
- Working with dangerous substances
- Use of personal protective equipment
- Periodic activities in the plant

Operating and safety instructions:
- Proper operation of apparatus and machinery
- Repair procedures
- Response to abnormal events

This information must be presented to all employees before they take up work in a plant, and at regular intervals thereafter. It is essential to test whether the knowledge presented has been understood and whether it can be promptly and correctly put into practice in stress situations. When the plant is being shut down on purpose, it is recommended that the shut-down process be combined with a simulated emergency.

Another vital activity in continuous plants is the keeping of a shift log. This helps inform later shifts about all important events in the plant. It is documentary in character, and must contain the following items:

– Bulletins from management
– Process upsets
– Deviations from normal operation
– Defects found in the plant
– Repairs needed and those completed
– Availability of standby equipment and machinery
– Warnings from supplier and customer plants

4.2.2.3. Special Features of Continuous Processes

Usually, large quantities of materials are involved in continuous processes. Large quantities, especially of flammable gases or toxic liquids or solids, represent high hazard potentials that must be taken into consideration when starting up the process for the first time or restarting after a temporary halt, during operation, and in the shut-down phases.

A continuous process has the advantage of inherent safety. While reactors have large dimensions and high throughputs, the quantities of products directly involved in the reaction and the holdup in the reactor are small; the hazard potential due to the reaction is minimal, especially for strongly exothermic reactions.

Continuous plants must be carefully inspected before they are first brought on stream:

1) All installation work, including subordinate work such as steam-trace lines and thermal insulation, has been completed
2) Scaffolding has been removed from the plant
3) Craft workers and assistants not actually needed in starting up the process must have left the plant
4) The entire system (logically subdivided into functional units) has been pressurized with gases (air or nitrogen) or liquids (water or harmless products) and checked for leaks
5) Auxiliary loops, coolers and chillers, special product or scrubber circuits, cleaning loops, and the wastewater and waste air cleanup units have been put through successful test runs
6) All instrumentation and control equipment, particularly their safety-relevant and fail-safe functions have been subjected to functional testing by qualified specialists. These functional checks must include not only the operation of alarms and safety interlocks on partial or total power outage, but also the emergency shut-down systems for plant sections or the entire plant

Almost all large continuous plants are now run with computer-assisted process control systems. Operating personnel must be thoroughly trained in the use of these systems so that they can take the proper risk-reducing actions in case of trouble, even in stress situations. Safety-related circuits are commonly hard-wired in parallel with the process control system. Switch points (temperatures, pressures, quantities, flow rates) are then fixed; they may be changed only after plant management and/or a safety expert has performed a safety analysis and issued a written directive. Where safety circuits are integrated into the process control system, technical or administrative policies must insure that safetyrelevant switch points cannot be altered by operating personnel on their own authority.

When starting up a continuous chemical process, special attention should be focused on "priming" the reaction. The pressure, temperature, quantities, and concentrations of reactants or catalysts must meet the criteria specified for the start-up phase. To avoid dangerous conditions due to delayed reactions and the associated buildup of unreacted feeds, the reaction start should be precisely monitored for rates of change (e.g., temperature or pressure rise) or with real-time analytical methods. If irregularities are seen, the reaction start should be interrupted, the fault identified and remedied, and the plant readied for another attempt. Many process control systems have sequence controls programmed to halt start-up automatically if an upset occurs.

During on-stream operation, the plant personnel continuously observe the process parameters, comparing them with the acceptable nominal ranges. Analyses of intermediate and end products not only help keep the commerical product within specification, but also provide information for use in maintaining the process. In particular, buildup of undesired byproducts due to recycles within a process or in closed- cycle scrubber units must elicit a high level of operator attention, and lead to countermeasures if appropriate.

Important process parameters that are recorded or stored by the process control system must be retained for up to ten years (depending on legislative provisions). These can be used to demonstrate that the process was run properly, or to prove compliance with emission limits and other air and water pollution control standards during regulatory audits; they are also indispensable in post-accident reconstruction.

As a rule, a continuous plant is shut down once a year for cleaning, inspection, and major repairs. After the process is stopped and before these operations are started, the plant must be carefully drained, purged, and cleaned according to the operating manual. For especially hazardous work (entry into vessels and tight spaces, or welding in plants that process flammable or explosive substances) plant management must establish safety practices that must be specified in detail in written work permisssions given to the employees performing the jobs. Exact instructions for the performance of repairs must be given (see Section 4.4).

When several craft teams are working at the same time, a coordinator must be designated to prevent them from endangering each other. The coordinator plans the work in time and space and is authorized to issue instructions to the craft teams.

Large continuous plants where explosive, flammable, acutely toxic, or environmentally harmful substances are handled in such quan-

tities as to create a high latent hazard potential present a special concern. Such a plant should have a rapid-response service, organized at each key level of the hierarchy, which can provide expert advice to emergency workers, to minimize the consequences of an accident.

4.3. Technical Inspection [211–217]

Introduction. In what follows, the term "technical inspection" denotes the inspection of technical devices by independent third parties; its purpose is to find out whether the equipment meets requirements set forth in legislation, regulations, or standards. In this way, unacceptable risks to employees working with the equipment – and to third parties and the environment – are prevented.

Plant hardware subject to technical inspection includes pressurized systems (vessels, piping); storage facilities for flammable substances or water pollutants; elevating, hoisting, and conveying equipment for the transportation of persons or freight; electrical equipment; nuclear facilities; means of transport such as automobiles and trucks, railroads, ships, and aircraft; power plants; petrochemical plants; and chemical plants.

Experts. As a rule, inspection is performed by an expert recognized by the responsible authority. Conditions for recognition include not only personal integrity, but also appropriate specialist qualifications (through education and continuing education), and it must be determined that the expert is not subject to instruction or restraint having to do with the inspection activity, i.e., independent.

The specialist qualification is generally a natural science or engineering degree. In many cases, a specified period of employment in the field is also required.

The expert may also be required to take part in an exchange of experience, which has two purposes: to assure consistency of technical inspection, and to identify crucial points of concern for a certain type of technical equipment. The results of such an exchange can also provide a basis for technical progress, and serve as input for standards development.

The role of the expert in technical inspection has long been established in Germany. The high point of the independent expert tradition was the field of law dealing with facilities "subject to surveillance." The experts recognized in this area are usually comprised in Technical Inspection Boards (TÜVs), which have in turn combined (along with some industrial firms that support their own inspection services, staffed by recognized experts) in the Association of Technical Inspection Boards.

Official recognition as an expert used to be limited to an individual on the German system. At European level, organizations will be entrusted with the performance of safety tests in the context of equipment manufacturing. Such organizations, after gaining accreditation at national level, can apply to the European Commission for "notification". These inspection organizations or "notified bodies" are to be engaged by manufacturers, especially for certain conformity tests pursuant to EU directives. The requirements on notified bodies are set forth in the Directives and relate chiefly to independence, personnel and hardware resources, and employee training. While expert inspection in Germany has been handled on a regional basis, notification as a testing service is valid throughout Europe, i.e., a laboratory accredited by a European Union member country and notified by the Commission can also do the pertinent conformity tests in the other member countries.

German Federal Water Quality and Pollution Control Act (Wasserhaushaltsgesetz) is following a similar path. Where tests by experts used to be prescribed, jurisdiction is shifting to inspection organizations, whose recognition in one Federal State is valid in all states.

Legal Background. The purpose of technical inspection is to make certain that the equipment inspected satisfies the quality and operational requirements set forth in regulations and standards.

Regulations are based on legislation that designates goods for protection and identifies technical devices to be regulated. The legislature thus authorizes regulatory bodies to set requirements, e.g., quality, operation, and inspection. Administrative actions such as permits and approvals are also authorized in this way.

A regulation commonly defines the scope of application, defines exceptions, states what technical devices are covered, establishes the scope and intervals of inspection, designates the persons or institutions that perform the inspections, and states what agencies have surveillance authority. The requirements set forth in such regulations are then implemented in instructions to the agencies or in engineering codes developed by private or public standardization bodies.

The harmonized European standards promise to be of great importance in this respect. These standards implement the basic safety requirements stated in European directives. Although their application is voluntary, they allow the presumption that a technical device complying with the harmonized European standards also fulfills the essential safety requirements of the pertinent directives.

Expert Inspection. Inspection and testing by the expert usually takes place in two phases. In the first, the existence and completeness of all necessary documents for the technical device are verified (official approvals, type approvals, type test certificates, certificates covering the inspection and use of materials, technical drawings and plans, operational descriptions, and reports from individual tests already performed). The second phase comprises the actual testing of physical facts.

The process of expert inspection can also be broken down in time. The first point examined is whether the design being inspected is suitable for the intended service. In the second phase, it is determined whether the technical device as fabricated complies with the submitted and examined design.

The next step is to verify that the finished technical device is operated in such a way that no stresses are imposed on it other than those considered in the first phase. An integrated system analysis may be called for here, especially if interaction with other technical devices or within the system as a whole must be taken into account. Finally, after a set inspection interval has passed, so- called periodic tests are conducted. These verify whether changes in the technical device due to its service, e.g., those resulting from wear, suggest that the stresses experienced in the next inspection cycle may no longer be tolerated. Another question at this stage is whether there have been safety-relevant changes within the system in which the component is integrated.

In addition to inspections before and during service, technical devices may also have to be inspected after they are taken out of service, if significant danger can result from improper operation.

When technical devices or their service conditions are modified in such a way that there may be safety-relevant effects, a new expert inspection generally has to be performed.

Finally, the responsible agency can instruct the expert to assess departures from the regulations or to investigate accidents.

The methods used in technical inspection range from simple visual examination to complex, specialized testing methods such as nondestructive ultrasonic tests or electron microscopy in connection with accident investigations. Such tests are generally performed, and the results interpreted, by specialists acting in support of the expert, who evaluates the results and cites them in the final report.

Integrated Plant Inspection. The interdisciplinary efforts of specialists in assessing safety problems is particularly important in the inspection of facilities subject to approval under the German Federal Pollution Control Act (*Bundesimmissionsschutzgesetz*). "Whole-plant inspection" means a comprehensive system inspection and monitoring of a plant with regard to safety functions.

This procedure takes in material properties, process conditions, interaction between apparatus and other plant components, interaction between human beings and hardware, administrative policies, and the effect of external safety-relevant factors. Inspections in such a program are concerned with whether the operation complies with legislative requirements (pollution control, occupational safety and health, water quality, soil conservation), pertinent regulations, and recommendations.

These inspection and surveillance tasks concern all those concerned. The three pillars of whole-plant inspection are thus:

1) Supervision of the plant as part of the operator's responsibility
2) Inspection and surveillance by public agencies

3) Safety inspections by experts

Such a procedure can also serve as a model for the inspection of plants, as proposed in a draft European Council Directive for avoiding the dangerous consequences of serious accidents with hazardous substances.

4.4. Maintenance

A plant operating in a trouble-free manner is in its safest condition. However, because technical components are subject to mechanical and process-related wear, upsets can occur. As a result, maintenance is necessary to keep the plant in a safe condition or restore it to such a condition. On the other hand, maintenance can create new dangers in the plant.

This maintenance-related risk can be minimized only by good safety practices in the planning and execution of maintenance work.

In the past, mistakes in the preparation or execution of maintenance were a significant cause of accidents in the chemical industry. The gas explosion at Flixborough in 1974 was traced back to a makeshift repair procedure [218], and many smaller and less spectacular events have resulted from maintenance errors [219], [220]. In a representative time interval, 30 % of all events occurred during maintenance [220].

Therefore, the value of safe maintenance in the chemical industry cannot be overstated. This is not surprising if the cost of maintenance is analyzed. It amounts to 3 – 5 % of the plant replacement value per year; in large companies, this can add up to more than DM 10^9 a year [221], [222]. These costs arise from a large number of small operations, and are made up mainly of personnel costs (including those incurred by vendors). A high level of administrative and logistical expenditure is called for. Maintenance actions, especially repair work, are only partially recurrent; many repair jobs must be planned and carried out individually.

These points imply the following three fundamental goals:

1) Plants should be designed and constructed in such a way that they are easy to maintain, i.e., easy to repair (good cleaning and draining capabilities, proper access, etc.); require little maintenance; and are made up of the most reliable and long-lived components possible
2) The proportion of unforeseen repairs caused by damage should be reduced in favor of scheduled (including preventive) operations
3) Safe procedures must be adopted and enforced for the performance of maintenance work, to reduce the risk arising from maintenance

4.4.1. Actions During Plant Design and Construction

The task ensuring safety during maintenance work is to create the technical and administrative conditions for safe performance of the work. This requires a systematic procedure (Fig. 61).

A good maintenance strategy begins with the design of a low-maintenance, easy-to-maintain plant. Every intervention in the plant can lead to danger during maintenance work and on restarting. The achievement of long life and high reliability at this stage is also economically important.

Systematic safety analysis (e.g., HAZOP) can identify the consequences of malfunctions and their possible causes at the design stage. Two types of event must be distinguished:

– Events that follow immediately from the technical failure and the resulting deviation of the process parameters from the design conditions
– Events that can occur during a maintenance operation

The objective in event analysis is to minimize hazards:

– To achieve the highest possible reliability in terms of failures, and to limit the consequences of failures
– To minimize the repair risk, i.e., to reduce the number of maintenance actions and to perform them as safe as possible

Selection Criteria for Apparatus and Machinery. High reliability is obtained by designing and selecting components on the basis of their requirements. Key criteria are:

Construction materials

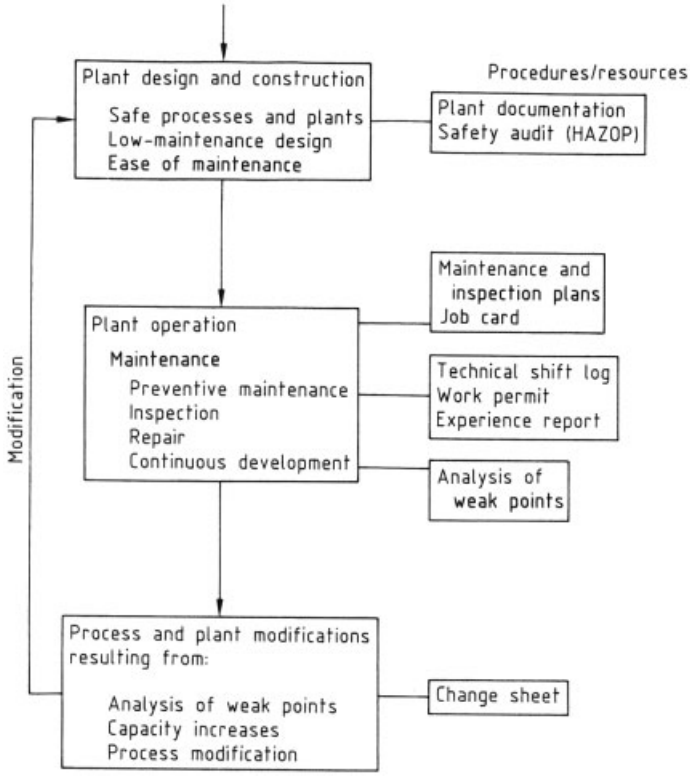

Figure 61. Maintenance strategy: central significance in a chemical plant

Functional principle
Robustness

Secondary points are hardware safeguards against improper operating conditions, e.g., flow monitors on pumps, which help to prevent damage.

Because a component may fail anyway, the consequence of such a failure must be analyzed. Even if the probability of occurrence is low, an event (failure mode) with serious consequences can influence the above criteria and make it necessary to seek a different solution.

Once these criteria are set, quality assurance systems must be instituted to monitor and guarantee compliance with the standards during the subsequent fabrication and installation of apparatus and machinery.

Design Criteria. Because maintenance and repair are inevitable, danger arising from them can be most effectively minimized by good installation and design practices. The criteria at this stage include clear layout (accessibility), easy and safe disassembly, and draining and pressure-relief capabilities. In a chemical plant, special importance attaches to draining. Some maintenance jobs can be made safer by the use of assembly aids, local spot-vent systems, etc.

Aids for Monitoring Plants and Performing Safe Maintenance. The documents needed for reliable operation and safe maintenance should be created at the design stage. These documents can usually be assembled from documents required for plant design and construction, e.g., manufacturers' data sheets, internal experience reports, and piping and instrumentation diagrams.

Periodic inspections are supported by maintenance and inspection schedules, which must be present when the plant is commissioned. Test methods involving no opening of equipment or piping in the plant (e.g., ultrasonic inspection) are preferred. Where safety-relevant process control equipment is not redundant, self-

monitoring is recommended. Inspections, including functional tests and plausibility checks, help todiscover concealed defects and to identify needed repairs. In some situations, automatic monitoring devices such as vibration monitors can also be used.

Anticipated, i.e., recurrent repairs must be thought out ahead of time. Detailed job instructions, supported by extracts from technical drawings (job cards, Figs. 62, 63), cut the number of mistakes during a repair or preparation for one.

Other documents needed later for the proper execution of maintenance tasks include: piping and instrumentation diagrams with piping and seal specifications; piping and fitting lists; apparatus specifications; measurement station data sheets (to ensure compliance with standards); and functional diagrams of the process control system with lists of setpoints, limits, and alarm settings (unambiguous transfer of process parameters to process control system).

4.4.2. Actions During Plant Operation

Figure 64 illustrates the maintenance strategy for this phase. The maintenance task is to keep the plant in its nominal condition so that requirements on product quality, safety, and availability are always met. According to DIN 31 051, maintenance is defined in the following way:

1) Preventive maintenance: monitoring and keeping plant in nominal state
2) Inspection: determination and assessment of actual state
3) Repair: restoration of nominal state

There is a fundamental distinction between scheduled maintenance and breakdown maintenance. An attempt is made to minimize the proportion of breakdown maintenance caused by unforeseen failures. Scheduled maintenance is particularly concerned with components having safety and quality relevance; it includes upkeep and inspection as well as measures to obviate repairs or cut their cost [223], [224].

Preventive maintenance and inspection are usually done according to predetermined schedules and are thus systematic (timing and scope set in advance). Maintenance and inspection schedules are used for this purpose. Inspections are scheduled on the basis of the anticipated wear of the components involved, known either from manufacturers' data or from operating experience.

Preventive maintenance and inspection are done both while the plant is on stream, nominal/actual comparisons being made continuously and documented during regular rounds, and during scheduled shutdowns. Maintenance and inspection schedules are a reliable and efficient way to organize the latter. Unforeseen failures of plant equipment occur despite scheduled maintenance, and these must be covered by administrative policies (usually operating procedures) providing for later repairs. This "breakdown" maintenance is also supported by repair instructions and aids rendering the work more reliable and faster.

4.4.2.1. Maintenance Activity

The procedures for scheduled and breakdown maintenance are similar (Fig. 64). The key step in minimizing risk of breakdown is early detection of a deviation from nominal conditions, which can be apparent in:

– Obvious defects: drips, noise, etc.
– Concealed defects, e.g., gradual slippage of process parameters

Regular rounds by shift personnel under a fixed plan can aid detection of the problem and thus contribute to minimizing risks. The maintenance and inspection schedule for every technical component contains information on:

– Test criteria applied in maintenance and inspection
– Methods practiced
– Timing

The action ends with a finding, documented in a report signed by the inspector.

If problems are identified, but cannot be remedied immediately, they are recorded in the plant shift log. The description should be oriented to the concrete observation, and should not contain speculation about the cause. The proper remedy cannot be established until the malfunction or the damage has been jointly assessed by plant management, the shift supervisor, the shop supervisor, and in some cases a specialist. A meeting must be held to determine how and with

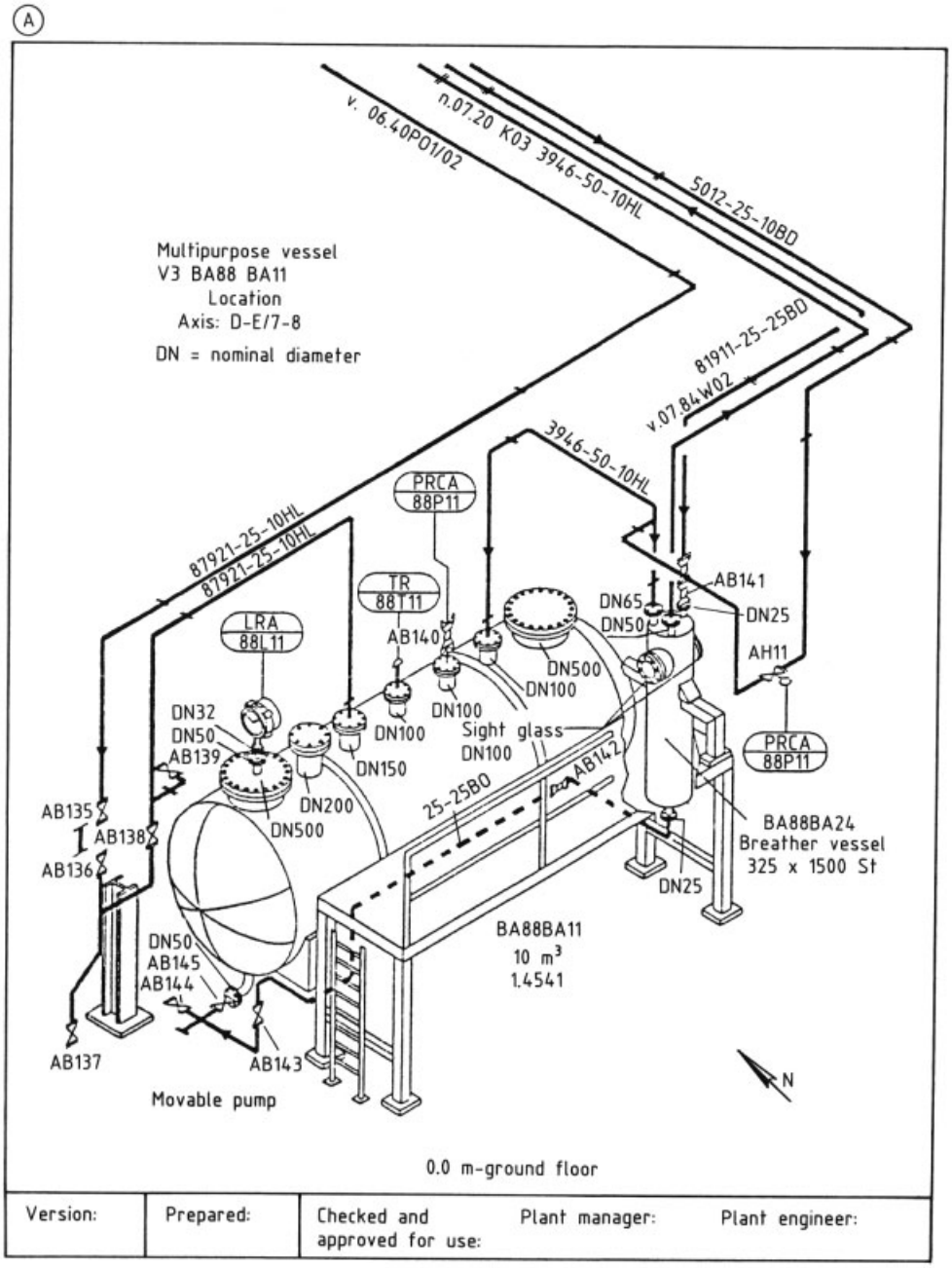

Figure 62. A) Example of a job card

(B)

		Multipurpose vessel				
MID Plant	B577 Building	22 Works	SIS Plant compl.	V 3 Unit	BA88 Unit sect.	BA11 Techn. device
Shop Job No.		Short text			Recurrent Job No.	Job No.

Work permit	Documentation	No. in shift log	Exhibitor	Tel.No.	Date
Yes ☐ No ☐	Yes ☐ No ☐				

Job instructions:

Inspections:		Medium	Pressure, bar IR	RDDR	AR	RUDR		Acceptance test requested:	
Pressure test	☐							Safety department	☐
Leak test	☐							Voluntary	☐
Intl. exam.	☐							Acceptance test performed:	
Endoscopy	☐							Safety department	☐
Ultrasonic wall thickness meas.	☐							Material testing	☐
— — — —	☐							SK	☐
Cl.No. — — — —	☐								

No.	DN	PN	Material	Bolts	Remarks

Dept.	X	Crafts	Week MO,TU,WE, TH,FR,SA,SU	Week	Week	Week	Week
Maintenance		Plant crafts					
		Measurement					
		Pool					
		PCS B					
Production		Operating					
Contractor		Central assembly					
		Piping fabrication					
		Steel construction					
		Erection/cranes					
		Apparatus fabrication					
Civil		Scaffolding					
		Insulation					
		Painting					
Engineering		PAT 3.2					

Prepared:
Name: Date:

Figure 63. B) Example of a job card

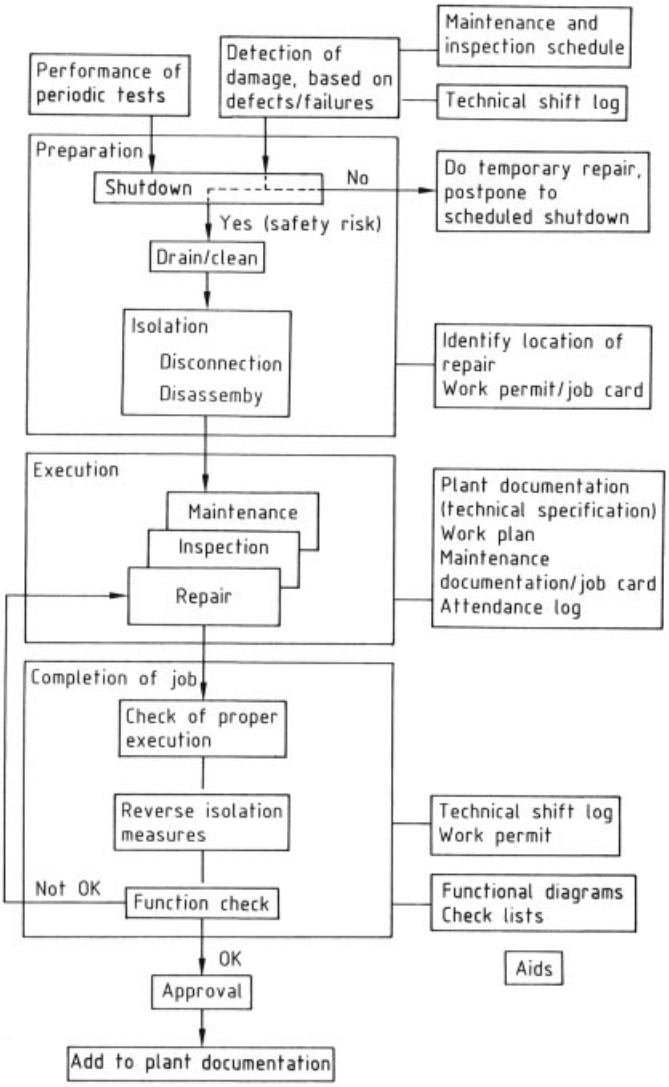

Figure 64. Maintenance activity chart

what precautions the repair work is performed (Section 4.2.2.2).

A further aid to proper reporting of technical problems by operating personnel to maintenance personnel is the technical shift log, which simultaneously serves as a tool for evaluating damage; the shift personnel enter the following information:

– Trouble during the shift
– Exact location (i.e., identification of component)
– Nature of trouble or malfunction observed
– Time of discovery

The log should also show what maintenance action was taken; the person responsible must sign it.

The purpose of the technical shift log is to state clearly what was done, and who was responsible, to prevent ill- considered and uncoordinated repair efforts. With the help of the technical shift log – including the signature of the responsible person – it is always possible to verify that necessary action has been done properly.

The technical shift log performs two other functions:

– Identification of temporary and makeshift repairs
– Evaluation in terms of recurrent problems (analysis of weak points)
– Computer support is recommended so that:
– Temporary repairs are not left in place permanently
– Statistical analysis can be done and appropriate countermeasures instituted to extend service life and reduce risk by cutting the number of repair projects [225]

Safety procedures must be followed whether the maintenance job is:

– Anticipated when preventive maintenance, inspection, and repairs are to be carried out during a scheduled shut-down (scheduled maintenance); or
– Unplanned, i.e., caused by a failure (breakdown maintenance)

In either case, the procedures must be designed to cope with the risks occurring at every stage.

If a problem arises, e.g., failure of a safety device, it may directly affect plant safety. This aspect is not pursued here; the proper response is always set out in the context of the safety analysis.

As a rule, the plant or sections of it must be put in a safe condition by a shut-down procedure when a problem occurs; this must not be confused with the "emergency off" or "panic" button for a section of the plant. While planned or "checklist" shutdown of the plant is a frequent procedure, the failure of a technical component required for shutdown may lead to the isolation of specific apparatus or sections of the plant, or to overriding the safety interlocks. Examples include bypassing a leaking agitated vessel and releasing an interlocked fitting that normally cannot be opened except at a certain point in the reaction.

Shutdown is not always necessary. In many instances, countermeasures can be taken that do not lead directly to a solution of the problem. There are three possibilities:

1) Postponing the proper repair (e.g., temporary plugging of leaks)
2) Using planned fallbacks (e.g., putting standby systems on stream)
3) Taking alternative measures (e.g., blowing a vessel down instead of pumping it dry)

Alternative actions are temporary modifications; such makeshifts may be more desirable than immediate shutdown of the plant, e.g., in a large continuous process where operation can be maintained until the next planned shutdown. Any modification must be checked first by a systematic safety analysis.

4.4.2.2. Performance of Maintenance

Scheduled and breakdown maintenance differ, especially in the documentation existing at the start of production, which describes the specific work to be done and takes the form of instructions:

– Scheduled maintenance is covered by preventive maintenance and inspection schedules describing the nature and scope of the work. Repair instructions (e.g., job cards) are generally used. Plans for shutdown maintenance are of this type, especially in continuous process plants.
– For breakdown maintenance, the problem and the corrective action taken are recorded in the technical shift log. Special repair instructions are not always available, and the documentation is then limited to the written plant documentation, including updates.

There are essentially two further possibilities:

1) Maintenance work while the plant is on stream
2) Maintenance work during a shutdown

In what follows, only maintenance during shutdown is discussed. There is no fundamental difference between total and partial shutdown. The only routine actions that should be taken in an operating plant are those that raise no safety concerns.

Maintenance can be broken down into three phases:

Preparation
Execution
Post- completion measures

When danger can arise during the performance of this work, all workers involved must follow special safety practices, which are written up in a "work permit". The purpose of this is twofold: to make certain that the proper safety procedures are followed; and to ensure that responsibility for the components of the work are clearly defined by signature.

Work Permit **No.**

Equipment: **Polymerization reactor 1 (R 3.02)**

Description of work (incl. to whom assigned)
Reattachment of submerged product feed pipe by fitters

1. **Isolation of plant section/component:**
   Depressurize all product lines connected; open pipe unions and close off with blank flanges; disconnect N_2 lines and close off with blank flanges. Disconnect external steam tracing. Disconnect power to agitator motor. Release interlock on bottom drain valve. Completely empty vessel through R 2.03, and leave valve open.
2. **Actions on isolated plant section/component:**
   Using high-pressure cleaning unit, clean vessel with water through manhole. Close bottom drain valve and fill with water. Drain both cleaning waters through R 2.04 for analysis. Have specialist team perform radiometric measurement of level from outside pipes. Suspend ventilation device in vessel. Close bottom drain valve.
   Date: Time: Signature:
3. **Action before beginning work:**
   Instruct fitters that no welding or driving is permitted.
   Instructions given to: Signature:
4. **Occupational safety measures:**
   Introduce conductors safely into tank; post lookout at manhole; ventilation unit; protective clothing for fitters in vessel; only ordinary work clothing (note: hardhat, protective goggles, safety shoes, . . .).
5. **Actions after completion of work:**
6. **Management approval (by plant manager) for performance of work:**
   From 1 March 1993, 10 : 30 am, to 1 March 1993, 4 : 30 pm
   Date: Time: Signature:
7. **Signature of shop supervisor responsible for compliance with items 4 and 5**
   Date: Time: Signature:
8. **Approval to proceed on site by engineering supervisor of plant**
   Date: Time: Signature:
9. **After completion of work: approval by plant supervisor**
10. **All- clear and technical acceptance**
    All clear:
    Date: Time: Signature of shop supervisor:
    Technical acceptance:
    Date: Time Signature:
11. **Waiver of stated procedures (see 1 and 2)**
12. **Acceptance before restart**

Preparatory steps include isolating single components or whole sections of the plant. This requires proper shutdown of the section concerned to a safe condition. Where redundant components exist for the malfunctioning components, plant management and/or the set operating instructions may allow these to be placed on stream. Other temporary actions such as:

– Blowing down instead of pumping dry
– Changing the preset process control switching or alarm values
– Interfering with interlocks

are permitted only after a special safety check, unless they are provided for in special operating procedures.

After shutdown, the critical danger points are release of substances by opening up the system to drain it (danger of poisoning or fire/explosion), and the possible introduction of undesired substances during cleaning.

Ideally, a plant should have emptying devices and holding tanks for the contents of the equipment, so the plant does not have to be opened for drainage. This concept is desirable for substances with high hazard potentials, but the cost may be substantial. Regardless of whether fixed holding tanks are installed, emptying by gravity should be possible.

Normally it is necessary to drain only certain portions of the plant. Such areas can be safely separated from other, product- containing sections by closing some fittings and inserting blanks. Residual quantities of product remain in the drained parts of the plant, however, especially in fittings, filters, and pumps. The individual draining features of these must be used to free them of product. Drainage is frequently aided by nitrogen pressure. Subsequent cleaning is not required with highly volatile substances. The vented air from such an operation must be considered; a vent system may be necessary, along with downstream treatment of the gases.

When cleaning is required, the agents used must be compatible with the product and, physiologically, as innocuous as possible. If inert media cannot be used, the following become particularly important:

– Medium adequately diluted in the shortest possible time

– Agent reactive only with the small quantity of product found in the drained segment of the plant
– Cleaning liquid safely separated from sections of the plant containing larger amounts of product

Failure to observe these conditions has frequently led to events, often because, in a nonroutine cleaning operation, the flow direction, pressure, and valve configuration were different from the normal conditions, and so were interpreted wrongly. All connections to other sections of the plant must therefore be examined, including those not ordinarily filled with product (e.g., off-gas piping).

When maintenance work calls for disconnection or disassembly of a plant section, personal injury hazards can arise if the repair site is not rendered safe, e.g., if the electric power to drives is not disconnected or product-filled pipes are not safely closed with a blank flange after disassembly of a subsystem. The latter point is crucial when equipment such as pumps and fittings are disassembled, because the safety measures are carried out right at the disassembly site; in larger maintenance operations, entire piping systems are generally made product-free.

Engineering codes, accident prevention regulations, and special safety instructions aid in the proper execution of these operations. To avoid misunderstandings at the site, particularly in a complex installation, the defective item must be identified, and its isolation must be indicated visually. The area can be further marked with a tag stating which craft has jurisdiction and what kind of work is being done.

The key requirement for safe repairs is up-to-date plant documentation. On the basis of the technical specifications (complete and unambiguous description of performance data, media, construction materials, and design data), it must be ensured that parts and materials used (e.g., greases) are such that no danger arises when the affected part of the plant is on stream.

Hazards in this sense are leaks due to improper materials (e.g., gaskets) or incorrect temperature/pressure design; unwanted reactions due to the catalytic action of construction materials or other media present; and parameters outside safe ranges owing due to the installation of unsuitable equipment (e.g., a pump with too large an impeller generating an excessive pressure). Similar hazards must be considered when replacing or repairing process control system components. A typical mistake involving an instrument is the use of the wrong range.

Just as a repair or inspection site must be labeled to prevent confusion, replacement parts or newly delivered repaired components must be labeled so they can be recognized as equipment meeting the specifications. Often there is no outward difference between two parts, e.g., when only the materials inside the device differ.

These practices are intended to ensure that the on-site activity can be finished in the shortest possible time, without interruption or risk. Safety and economy are not in conflict.

Virtually every maintenance activity comprises many single activities performed by different crafts. Because economics does not permit craft specialists to be employed especially for all such tasks, workers unfamiliar with the special hazards of the production plant have to be brought in; the requirement of up-to-date plant documentation must be supplemented by explicit instructions to the craftspeople at the work site, and a clear task description (see the job card in Fig. 62).

It is necessary to establish both the plant documentation needed for maintenance, and the scheduling of various maintenance activities in a plant area. Only by matching and skillfully coordinating maintenance activities to the rhythm of the process they can be prevented from threatening each other. An aid in this respect is a working plan, which describes the sequence of tasks and their duration from their starting times (e.g., a bar chart). In an area where welding, driving, and caulking work is done, for example, explosive mixtures are not permitted. Production not connected to the maintenance action may have to be interrupted. The same holds when flammable or explosive substances are present in the course of maintenance, e.g., when a protective coating is applied to a surface.

Some kinds of work also require a second person to be present, even if the job requires only one craft worker. This is the case when a hazard such as the escape of toxic or explosive mixtures cannot be positively ruled out, or when work has to be done inside vessels.

Maintenance jobs can extend over more than a day, so interruptions cannot always be avoided.

The area of the plant where repairs are being done should be marked off with signs throughout the job, especially to prevent an inadvertent return to service before the work is complete.

Approval to restart can be given only by persons whose functions include this task, e.g., the responsible plant engineer. Only a single person should be made responsible for maintenance. This person is the contact point throughout maintenance, otherwise there is a danger that someone who does not have an overview of the work may make a wrong decision.

To avoid the danger of premature startup, interruptions of the work must be recorded, e.g., in an attendance log, and completion must be documented in the technical shift log. For work that requires a permit, the permit includes a space for approval after completion of the job.

A maintenance action cannot be considered finished until all the unit activities have been checked and accepted by the coordinator of the overall action. This includes both inspection for correct execution of a task, and functional testing. Proper execution means both the use of the components provided (correct temperature and pressure ratings, material, etc.) and their functionally correct installation. Functional testing is then done to make certain that all technical components work together in the way necessary for the process. The simplest example is leak testing of a repaired section of the plant; a very complicated one is the comprehensive testing of a process control system.

Where preparatory action has been taken, e.g., isolation of sections of the plant, these must be reversed in the same way as temporary modifications. Approval can then be given for production. In some cases, the plant is restarted a step at a time, extra functional tests with reduced throughput or enlarged crew being done first. Such precautions are needed especially when extensive maintenance work has been carried out.

4.4.3. Continuing Plant Development (Analysis of Weak Points)

The elimination of problems by repair is a necessity, but it also offers the plant operators a chance to analyze recurring problems, and to use the results in improving the system. Analysis of maintenance actions for continuing development presupposes that the technology and the operating conditions are adequately documented (this includes unambiguous nomenclature in the plant). Entries in the technical shift log aid in evaluating the frequency of problems. If plant documentation is computerized, the cost of such a "weak point analysis" can be substantially reduced. The goal is to use technically feasible and economically acceptable means to optimize sections of the plant, i.e., to minimize the frequency and/or severity of trouble. Before such an analysis, the possibility of human error causing the problem must be eliminated.

This analysis of weak points must not be limited to the technology. Often, the technical defect can be eliminated more simply by changing the process conditions. The weak point found may not be the cause, but rather the effect, of defects in the process, e.g., process conditions or product specifications. In such cases, the analysis of weak points becomes a process analysis (Section 4.5).

The causes of problems may well be hard to identify. A final statement of causes often has to be based on the vigilance of the operating personnel on the spot, as well as the years they have spent learning the details. The plant manager must therefore make a continuing effort to motivate the personnel to be observant.

4.5. Modification of Plants

4.5.1. Reasons for Modifications

While the objective of maintenance is to keep the plant in its nominal condition, modifications to the process and plant involve changing the documented nominal state of either, or both. A variety of factors can lead to process and plant modifications:

- Increases in capacity
- Change in process to lower manufacturing costs and/or improve product quality, or to modify the product
- Measures to eradicate weak points
- Upgrading to the state of the art (technology or safety)

Any change in the nominal state means that the corresponding parts of the process and plant documentation must be revised or rewritten.

In this respect, modification differs in scope, but not in basic execution, from a new design project. The reason it is so important to treat a modification like a new design is that there is no way to know ahead of time how widely the modification will ramify itself. Even minor changes, e.g., replacing one apparatus by another, or raising the temperature in a unit operation, can cause a violation of the design or process control parameters. The systematic methods already described for evaluating new plants must be applied here too, because they cover the entire field of possible hazards. First, all areas of concern, e.g., explosion hazards, are studied to determine whether the proposed change affects them at all. Because a change often relates to only a few danger spots, the number of areas of concern quickly becomes small, as does the number of aspects to be considered; detailed analysis is needed for only a few aspects in most cases. Knowledge of the existing safety concept is essential for assessment of all new practices. Any change represents a "compatible" expansion of the whole concept.

4.5.2. Procedure for Modification

A suitable procedure must be established if modifications are to be carried out correctly. As soon as a major change becomes necessary in the plant, e.g., because a new unit operation is being introduced, the project is implemented in the same way as any other design project, right up to the construction of a new plant (Section 3.2).

In the case of minor changes, the amount of investigative effort required is generally small, so a simplified procedure is useful in such cases. The principle remains the same, but the effort invested in safety analysis and documentation updating is matched to the scope of the planned change. A useful aid is the change sheet, which has two functions: to describe the objective and the implementation of the planned change; and to insure that necessary safety questions have been asked and properly answered. The change sheet also documents the work and its approval, and is used in identifying the parts of the plant documentation affected by the change (Section 4.4.1). A signature is required for each approval and each inspection of each individual step in the process. The work permit already mentioned complements this task by listing actions to be taken during its execution.

The items contained in the change sheet are:

1) Identification of what is to be changed (including nature of change and other ordering criteria; suitable tools for this purpose are job cards, piping and instrumentation diagrams, and operating manuals)
2) Purpose of the change
3) Description of the change
4) Identification of essential danger spots with the help of a checklist
5) Discussion with specialist departments on the danger spots relevant to each
6) Inquiry into necessary updating of safety analysis
7) List of plant and process documents affected, with possibility of monitoring progress of the change; testing steps after completion (e.g., acceptance test, test run)
8) Questions as to possible effects on product quality

Important consultations with the specialist departments and other safety information must be recorded, and the safety analysis must be attached.

The following fictitious examples show how small changes can have wide-ranging consequences.

1) An additive used in a thermally sensitive mixture is purchased from another vendor; the composition appears to be only slightly different, but it lowers the safe temperature limit. The mixture is to be subjected to distillation at high temperature. The result is thermal decomposition of the bottoms, faster heat evolution, and a rise in pressure in the reboiler.

Corrective action: Perform safety study in the lab with the new feedstocks; lower the temperature set point.

1) A pump delivering acid from a storage tank must fill a vessel more rapidly. On the basis of the characteristic curve, the old pump was selected so that the vent line of the vessel cannot be completely flooded. As a result, releases of substances are prevented without a level-limiting feature having to be installed. The new pump, with a larger impeller, attains a higher pressure. It is now possible to flood

the vent completely so that acid is released. The danger is not noticed as long as overfilling does not occur. The inherent safety formerly present no longer exists.

Corrective action: Select pump according to proper safety documentation; see change sheet.

4.5.3. Summary of Plant Maintenance and Process Modification

The systematic procedure described, with the associated aids, contributes to plant safety from design to operation and process upgrades. By itself, however, it is not sufficient to guarantee plant safety. While it cannot replace the knowledge of experienced workers, it offers a guide and support for daily work. The reasons for the increasing volume of technical documentation and administrative support are:

– Plants themselves have become far more complex (process control engineering and other advances)
– Maintenance work is being done by a larger number of specialized teams
– There is increased public interest in readily understandable safety practices

5. Hazard Control

5.1. Means of Limiting Accident Impacts

The responsible care of every plant-operating company to take active measures to protect employees, third parties, and the environment against potential harm means that the operator must take responsibility for technical and administrative precautions to limit the impacts of minor and major accidents in the facility. Legal and regulatory provisions independently require limitation of the consequences of accidents in plants; e.g., the "Seveso Directive" [226], [227] and the implementing regulations in EC member countries.

Action to limit accident impacts is also mandated by other regulations, e.g., the German Pressure Regulation [228]. The release of substances from pressure-relief features of a plant (safety valves, rupture disks, emergency blowdown devices) must present "no hazard" to employees, third parties, and the environment [228], [229].

Whether such a release is permissible must be assessed in view of the properties of the substances, the location of the plant, the propagation behavior of the substances, the duration of the release, and the occurrence of hazardous concentrations. If a release is not possible without hazard, retention or disposal equipment must be provided to limit the consequences.

German water law also requires that plants be designed, built, and operated in such a way that water pollutants cannot escape. Substances that escape anyway must be retained quickly and reliably so that contamination of receiving waters does not become a concern [230], [231].

The basic idea of a two-barrier system is contained in these provisions. The enclosing shell of a storage tank, e.g., for fuel oil, or the pressure-resistant construction of a reactor in the chemical industry, is the first barrier; it must safely withstand the mechanical, thermal, and chemical stresses imposed on it in normal operation, and in the event of accidents.

If a release occurs in spite of the preventive measures, a second barrier must limit the consequences of the release in such a way that employees and third parties, as well as the neighborhood of the plant and the environment, cannot be threatened, or that such threat is held to an acceptable level.

The chemical industry uses a variety of means to limit the impacts of accidents. As a rule, these also have functions relating to occupational safety, pollution control, and water and soil conservation. In addition to retaining the product in liquid, solid, or gaseous form in case of an upset in the plant, they also provide for retention of contaminated wastewater or cooling water, and the retention of water contaminated by fire fighting.

The necessity of retention systems and the way in which they are implemented depends chiefly on the properties of the substances being handled, or those produced in the plant during an accident, and on the actions of such substances on persons and the environment – not just in terms of the immediate effect of exposure, but also the effects produced by a reaction (fire, explosion) of such substances. The hazard

potential of a substance is assessed on the basis of its physical, chemical, and biological properties. The water-polluting effect of a substance is generally evaluated from its assignment to a pollutant class [232], [233], based on the acute toxicity of the substance to mammals, bacteria, and fish (sometimes to algae and Daphnia), its degradability, long-term effects, and distribution.

The principal ways to limit the impacts of potential accidents are:

– Retention systems, catch wells
– Quick-closing valves, emergency compartmentalization systems
– Emergency drain and collection systems
– Blowoff and disposal systems
– Spray-curtain systems
– Partial and complete containment
– Fire protection
– Retention of water contaminated by fire fighting

The choice of the means to limit accident consequences must be based on the operating conditions (e.g., pressure, temperature, substance properties), the type of plant, its location, and the situation in its vicinity.

5.1.1. Retention Systems, Catch Wells

The best-known retention system for liquids is the catch pan or well. In Germany, the Engineering Code for Flammable Liquids (TRbF) prescribes this way of retaining any liquid that leaks from tanks located in the well [234], [235].

Catch wells not only prevent harm to soils and receiving waters by flammable liquids, but also limit the spread of fires and emissions. If the liquid surface is completely covered (e.g., with a low-expansion fire-fighting foam), fires and emissions can be almost entirely suppressed.

In addition to product retention, a catch pan or well can also hold back contaminated wastewater and water contaminated by fire fighting.

Special requirements apply to the design and selection of material of chemically resistant, leakproof retention systems such as catch wells, pans, and baseplates. These standards are chiefly dictated by the stresses occurring, for example:

– Mechanical stress due to loads, impacts, friction, building subsidence, internal stresses, and contraction
– Thermal stresses due to ambient temperature cycles (day – night, summer – winter) or hot media
– Chemical stresses dependent on water pollutant class, intermittent attack, penetration, or aggressive properties of retained substances relative to construction materials

The resistance to chemical attack needed in a retention system, and hence the material chosen, generally depend on the water pollutant classification and chemical nature of the substances present. Chemical stresses are classed by severity:

1) Low load: no water pollutants present in normal operation; retention of water contaminated by fire fighting
2) Moderate load: leaks with action limited in time (nonrecurring or recurring but brief)
3) High load: intermittent contact with water-polluting liquids; regular contact with heavily laden cleaning water

The duration of exposure must be determined for each component. If the loads are low or moderate, the design can be based on a fairly low-probability event, occurring once and of limited duration. Infrastructural administrative or technical measures, e.g., provision for prompt detection and remedy of leaks, may also be taken [236–239].

If required by the load assessment, retention systems can be rendered tight and chemically resistant by special coatings (resins, paints, etc.) or plastic or metal foil liners.

A large industrial plant can set up a further barrier for retention of chemicals and contaminated fire-fighting water by connecting production equipment or storage tanks to biosewers leading to the plant's wastewater treatment system.

A special type of retention system involves double-wall construction of vessels, pipes, etc. The retention space is attached to the containing barrier (in the case of internally jacketed vessels, placed inside it) in such a way that the space between the two walls catches any leaks and can be fitted with an alarm [239–242].

Double-wall design is not always safer than single-wall construction. The reasons have to do with fabrication cost and additional stresses

(e.g., those due to constrained expansion at penetrations and supports) as well as stability against denting and buckling. Inspection and maintenance of a double-wall component is difficult or impossible.

5.1.2. Rapid-Closing Valves, Emergency Compartmentalization Systems

In plants containing large amounts of products where the release of a substance would represent a high hazard potential, it may be useful to subdivide the plant with rapid- closing valves and fittings. These can be actuated after the detection of a release, e.g., by a gas warning system [243], [244], and after the scope of the release has been evaluated. They can be activated manually or automatically. The effect is to isolate the area of the leak from the rest of the facility. The quantity released at a leak can thus be limited to a value estimated in advance. The use of a rapid- closing valve system presupposes detailed advance planning and design of the plant sections, knowledge of the amounts of products contained, the way in which a leak is detected, etc. The analysis may also include the emergency shut-down of the entire plant if dangerous conditions in the plant can be created by blocking individual sections [245], [246].

5.1.3. Emergency Drain and Collection Systems

The product in a plant or in an isolated portion of the plant may be drained or released into a retention system (e.g., a catch tank or blowdown tank permanently installed in the plant) if a leak or other hazardous condition arises (e.g., fire). The retention system may be fitted with a separator (e.g., a cyclone) if the properties of the substance and the system pressure demand separation of the phases of the substance. It is also useful to configure the system in such a way that the plant drains by gravity. A pressure-relief system may also be needed to reduce the pressure of the gas phase [247], [248].

5.1.4. Blowoff and Disposal Systems

In plants handling gases, flares, scrubbers, absorbers, and condensers are useful for pressure relief and gas disposal. Systems of this type function not only as emergency pressure-relief systems, but are also used for the disposal of gases and vapors arising during normal operation, or when the plant is up or shut down, e.g., where large amounts of low-quality gas are produced for a short time. If these gases or vapors are flammable, the plant design must also meet the safety standards for explosion protection; other safeguards against spontaneous ignition and fire may also be necessary, e.g., with absorption on activated- carbon filters [249], [250].

5.1.5. Spray-Curtain (Drench) Systems

Spray curtains are often employed in plants with critical gases or vapors, where a release at an unpredictable location cannot be ruled out and may lead to a hazard to employees, third parties, or the environment. The plant section in question, or the whole plant, is surrounded by a pipe system fitted with closely spaced spray nozzles. If a pollutant escapes, a high-pressure mixture of water and steam issues from the nozzles, surrounding the plant in a dense water mist that can condense or capture the released substances. Neutralizing agents can also be added to the water; e.g., ammonia for phosgene plants. Due to the high-pressure and thermodynamic effects the effectiveness of these systems can be maintained, even in windy weather.

Mobile equipment such as articulated booms and water cannons operated by the plant fire fighting service can be used to create spray curtains in smaller plants, or as a supplementary measure in larger ones [251].

5.1.6. Partial and Complete Containment

A technique well known from the nuclear industry is to place the entire plant inside a pressure-tight containment, enclosed on all sides, but with personnel entry points. This approach has also been discussed for chemical plants and portions of plants where very dangerous substances are handled.

This approach may be correct from the environmental protection standpoint, but in practice there are limits, since some chemical plants

cannot be remotely controlled and cannot be run without operating personnel.

For reasons of occupational safety and health, a chemical plant containment cannot be hermetically sealed from the outside world while the plant is in normal operation. Entry for operation and maintenance must be possible, and continuous ventilation is required, both for waste heat removal and on occupational health grounds. A release would have to be detected very promptly in order for all openings to be shut and the stack-gas cleanup system turned on so that retention could be achieved.

A further problem arises with flammable substances. Even a small cloud of an explosive mixture released because of a malfunction is diluted much more slowly in the containment than in an open plant, so it remains flammable over a long time. Containment also exacerbates the effect of an explosion; because there are no pressure-relief openings, the pressure wave and heat of an explosion affect humans and structures many times more than in the open case.

A large fire in a containment plant inevitably causes a huge amount of heat to accumulate in the building. Thermal stresses on the structure and plant equipment result, possibly causing their complete failure. Mobile fire fighting is far more difficult in such a facility.

It is possible that an initially minor event in a containment plant may lead to substantial personal injury and a dangerous release of substances.

A variety of plant subdivisions, where special process or design conditions bring about a severe hazard potential for the environment, are already being built with double walls or enclosed in boxes or compartments. Aside from this form of enclosure and double-walled construction, whole-plant containment cannot be recommended for general use. Complete containment is an option only in rare cases, e.g., where there are no flammable or explosive substances; even in such cases it is necessary to strike a careful balance between advantages and disadvantages (pollution control, industrial safety, fire and explosion protection, plant reliability, and safe maintenance practices) [251–253].

5.1.7. Fire Protection; Retention of Water Contaminated by Fire Fighting

Effects on employees, third parties, and the environment must also be expected when fire occurs in the plant. These can result from the fire itself (heat, combustion gases, etc.), or from the extinguishing agents (e.g., the Sandoz accident at Basel). Fire protection legislation requires that the fire fighting service seek to extinguish a fire virtually always and under all circumstances. From the environmental protection standpoint, it may be preferable not to extinguish a fire that has already passed a certain size, but to let it burn at a high temperature while keeping it from spreading to the surroundings; the amount of environmentally harmful combustion gases generated is smaller, and no contaminated extinguishing agent is produced.

There must, however, always be fire protection measures to prevent the start and growth of a fire, or to safeguard the surroundings from the effects of fire (Section 3.4.1). Fixed facilities for the retention of water contaminated by fire fighting must be provided, depending on the substances involved, their water pollution classifications, the construction of the first barrier, and other factors [254].

Mobile retention systems can also be used as a barrier. A number of options, e.g., mobile folding vessels and rapid-set-up tanks, are now on the market. Even simple devices such as magnetic plates and compressed-air bags to close off sewer intakes, have their place in contaminated fire-fighting water retention.

5.2. Hazard Control Plans at Plant Level and Beyond

5.2.1. Hazard Control Plans

In what follows, plans established by a plant or works to control hazards inside the battery limits are referred to as plant alarm and hazard control plans. In Europe, plants subject to the "Seveso Directive" (implemented in Germany as the Industrial Emergencies Regulation) are required to create an alarm and hazard control plan [255], [256].

An alarm and hazard control plan describes administrative and technical actions to minimize the impacts of accidents; relates to various levels

of organization (e.g., constituent plants and the larger facility to which they belong) and is prepared with an eye to the immediate surroundings of each unit.

The actions prescribed include:

Management
Personnel
Technical resources
Communications
Documentation
Procedures and jurisdictions

In facilities containing a large number of varied units, alarm and hazard control plans are prepared at "plant" (smaller unit) and "works" (larger unit) levels [257].

5.2.1.1. Plant Hazard Control Plan

The plant alarm and hazard control plan sets forth the plant safety organization. It contains information necessary to plan and specify the emergency response.

The coordinated administrative measures taken in response to an accident are such as to afford the highest possible level of protection to humans and facilities.

The hazard control plan identifies all potential threats within the plant (e.g., those due to fire, explosion, or other accident) and external to the plant (e.g., trouble in nearby plants, flooding), lists safety equipment in the plant, and prescribes the desired actions of plant personnel in a dangerous situation.

The plant hazard control plan is aimed primarily at the people in charge (e.g., plant managers and engineers, laboratory and workshop managers), who must ensure that their personnel are properly instructed as to how to act in a particular hazard situation. The plan is a document for use in plant personnel instruction and drills, as provided in the Industrial Emergencies Regulation. The plant hazard control plan serves as a guide for people in charge, giving instructions for actions in a particular hazard situation. It lays down guidelines both for measure within the plant and for measures by technical task forces and service groups (e.g., fire safety, occupational safety, plant security, maintenance). The plan is also an informational aid and guide to emergency action for management, and provides a pattern for the allocation of technical resources.

5.2.1.2. Works Hazard Control Plan

The alarm and hazard control plan of the works relates to a higher organizational level than the plant, and documents the emergency response organization adopted by the works when an event occurs.

It defines, for all personnel involved:

The essential tasks to be performed
The way in which efforts are coordinated

It also establishes duties relating to information and prescribes the forms of cooperation with:

External organizations
Government agencies

5.2.1.3. Required Contents of the Plan

Many suggestions have appeared concerning what alarm and hazard control plans should contain [258], [259]. In the German state of North Rhine – Westphalia, a working group, including independent experts as well as agency and industry representatives, has developed a model plant alarm and hazard control plan [257], which can serve as a pattern for plant or works plans. Major headings in the model plan are:

Introduction. Mailing address, scope, revision (update) status

General Information on the Facility and its Surroundings. Orientation for emergency workers on:

1) Access routes and staging areas
2) Information on the work force
3) Danger sources, major hazards
4) Safety equipment and plans in place

Alarm Plan. The routing of alarms in the event of an accident is summarized, as is mandatory internal and external reporting.

Hazard Control by Internal Activities. This lists departments in the works (works manager, plant, works fire service, central safety office, industrial medicine, safety engineer, etc.) and defines their jurisdictions and assignments.

Warnings. Options for informing persons in the works or plant about possible hazards are discussed, along with ways of delivering warnings and other information to those living nearby.

Instructions for Particular Events. In view of local circumstances, precautions and responses are described for a variety of events (flood, smog, power outage, product release, need for outside assistance, etc.).

Notification of Government Agencies and the Public. This describes the internal location, where all information required to be reported externally can be retrieved.

Resources and Technical Specialists. A table lists all current technical documents, devices, equipment, and technical specialists, and records their availability and contact information.

Telephone List. Telephone numbers that may be vital for government agencies (for inquiries within the works) and for responsible works management personnel (e.g., for reports to agencies) must be listed here.

Definitions, Regulations, Conventions. This lists concepts and procedures defined in regulatory documents; key passages may be quoted verbatim.

Appendixes. These include maps, keyword index, and other items.

5.2.2. Off-Battery Hazard Control Plans

An example of an off-battery hazard control plan is the special accident response plan, created jointly by the disaster control agency and the plant operator. Such a plan relates to a plant or works in which, because of the production process, storage, and other features, hazards to nearby inhabitants may occur and disasters cannot be ruled out. The special accident response plan repeats portions of the works alarm and hazard control plan that are vital for response by the agency; these include:

1) Description of the facility
2) Routing of reports and alarms
3) Immediate response to accident/disaster
   - Identification of threat
   - Identification of threatened area
   - Measurement of pollutant levels
   - Warning/notification of populace
4) Subsequent actions

The disaster response plan of a municipality covers all events that may figure in a disaster. It is not restricted to any one plant or works.

5.3. Public Awareness and Responsible Care

Responsible Care. The chemical industry's worldwide responsible care program obliges companies to strive for steady improvements in health, safety, and environmental protection. It mandates that corporations seek a stronger dialog with the public, to enhance public knowledge of product and manufacturing safety.

The centerpiece of an active environmental philosophy, as the Association of the German Chemical Industry made binding on its member companies in 1992, consists of ten Guiding Principles and six Management Practice Codes intended to implement them. CEFIC (the European Chemical Industry Council) has published similar guiding principles for member societies [260].

Member companies are bound by the following guiding principles:

1) To acknowledge the public interest in chemical products and the activities of manufacturers, and to respect this interest
2) To develop and manufacture only those products that can be safely produced, transported, used, and disposed of
3) In the planning of new products and production processes and the improvement of existing products and processes, to assign a high priority to health, safety, and environmental aspects
4) In any chemical-related health and environmental activity, to provide full information to government agencies, employees, customers, and the public, and to recommend suitable protective practices
5) To advise customers on the safe use and transportation of chemical products, as well as their safe disposal
6) To operate production facilities so as to safeguard the environment as well as the health and safety of employees and the public

7) To conduct research to increase knowledge about the possible health, safety, and environmental impacts of products, production processes, and waste products
8) To cooperate with other companies in solving problems occasioned in the past by the handling and disposal of hazardous substances
9) To cooperate with the government and government agencies in devising responsible legislation, procedures, and standards that enhance the safety and protection of the public, workers, and the environment
10) To promote the principles and practices of the responsible care program by sharing experience with all the other companies manufacturing, using, transporting, and disposing of chemical products, offering all kinds of support to them

The six Management Practice Codes deal with:

Information and emergency response
Process and plant safety
Environmental protection
Transportation safety
Occupational health and safety
Product responsibility

CEFIC is promoting and developing the responsible care concept throughout Europe.

Public Awareness. Article 13 of the Seveso Directive for Europe in general, and § 11a of the Industrial Emergencies Regulation in Germany, mandates that the operators of accident-relevant plants provide information on safety practices and proper accident response to persons who may be affected by an accident as well as to the public at large. This information must be made available in suitable form, and without having to be solicited [255].

The objective is to create a bulletin, understandable to the general public, containing the following information:

– Name of plant operator and location of plant
– Name and title of the person giving information
– Brief description of the nature and purpose of the plant
– Identification of substances or preparations that can cause an accident, along with their essential hazardous attributes
– Nature of dangers in an accident, including possible impacts on humans and the environment
– Nature of warning given to affected persons and follow-up
– Correct behavior and actions of affected persons in the event of an accident
– Pertinent safety practices
– Internal and external hazard control plans
– Coordination of plans between plant operator, municipality, and agencies responsible for hazard control

In areas of high industrial concentration, a joint bulletin may be issued by the operators of plants where emergencies may occur.

6. References

General References

1. "Das Sicherheitskonzept für die chemische Technik," *Dechema Monogr.* **88** (1980) no. 1818 – 1835. BASF (eds.): *Sicherheit in der Chemie,* Verlag Wissenschaft und Politik, Berend von Nottbeck, Köln 1979. Bayer AG (eds.): *Sichere Chemietechnik,* Leverkusen 1982. F. P. Lees: *Loss Prevention in the Process Industries,* **vols. 1 and 2,** Butterworths, London 1980. G. L. Wells: *Safety in Process Plant Design,* J. Wiley & Sons, New York – Toronto 1980. S. Lange: *Ermittlung und Bewertung industrieller Risiken,* Springer Verlag, Berlin – Heidelberg 1984. "Fortschritte der Sicherheitstechnik I," *Dechema Monogr.* **107** (1987). D. A. Crowl, J. F. Louvar: *Chemical Process Safety: Fundamentals with Applications,* Prentice Hall Inc., New Jersey 1990. J. Whiston: *Safety in Chemical Production,* Blackwell Sci. Publ., Oxford 1991. H. Pohle: *Chemische Industrie – Umweltschutz, Arbeitsschutz, Anlagensicherheit,* VCH, Weinheim 1991. G. Strohrmann: *Anlagensicherung mit Mitteln der MSR-Technik,* 2nd ed., R. Oldenburg Verlag, München – Wien 1983. DECHEMA (ed.): "Anlagensicherung mit Mitteln der MSR-Technik," *Praxis der Sicherheitstechnik,* **vol. 1,** Frankfurt/Main 1988. NAMUR (eds.): *NAMUR-Empfehlung: Anlagensicherung mit Mitteln der Prozeßleittechnik,* Erstausgabe

30. 10. 92, Vertrieb: NAMUR-Geschäftsstelle. Berufsgenossenschaft der chemischen Industrie (BG Chemie) (eds.): *Ratgeber Anlagensicherheit; Grundlagen und Anwendungshilfen zur Anlagensicherheit,* Verlag Kluge, Berlin 1992.

Specific References

2. Berufsgenossenschaft der chemischen Industrie (BG Chemie), Heidelberg, annual reports 1983 – 1992.
3. Her Majesty's Stationery Office (HMSO), *The Flixborough Disaster – Report of the Court of Inquiry,* London 1975.
4. V. C. Marschall: *Major Chemical Hazards (Ellis Horwood series in chemical engineering),* Ellis Horwood, Chichester 1987.
5. H. Steen et al.: *Das Bhopal-Unglück 1984,* Umweltbundesamt (eds.), Berlin 1987.
6. Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit (eds.): "Rhein-Bericht, Bericht der Bundesregierung über die Verunreinigung des Rheins durch die Brandkatastrophe bei der Sandoz AG/Basel," Bonn, Feb. 12, 1987.
7. DIN/VDE 31 100,part 2 (Deutsche Norm), Begriffe der Sicherheitstechnik, Beuth Verlag, Berlin 1987.
8. H. Schacke et al.: "Redundanz im Staubexplosionsschutz? – Konzept komplementärer Schutzmaßnahmen," *Staub Reinhalt. Luft* **53** (1993) 453 – 459.
9. Statistisches Bundesamt (eds.): *Statistisches Jahrbuch 1993 für die Bundesrepublik Deutschland,* Wiesbaden 1993.
10. T. A. Jäger: "Zur Sicherheitsproblematik technologischer Entwicklungen," *Qualität Zuverlässigkeit* **19** (1974) no. 1, 2 – 9.
11. Polizeiliche Kriminalstatistik des Bundeskriminalamtes, Wiesbaden (für die Jahre 1985 bis 1991).
12. Amtliche Mitteilungen der Bundesanstalt für Arbeitsschutz, no. 2, Dortmund, April 1993.
13. V. Pilz: "Grundlagen für die Vorhersage der Auswirkungen von Störfällen," *VFDB 2.* (1981) no. 3, 116 – 125.
14. H. Giesbrecht et al.: "Analyse der potentiellen Explosionswirkung von kurzzeitig in die Atmosphäre freigesetzten Brenngasmengen," *Chem. Ing. Tech.* **52** (1980) no. 2, 114 – 122 (part 1); **53** (1981) no. 1, 1 – 10 (part 2).
15. W. C. Brasie, D. W. Simpson: "Guidelines for Estimating Damage Explosion," Reprint 21 A, Symp. Loss Prevention Process Ind., *63rd Nat. Meeting AIChE,* St. Louis, Mo, 1968.
16. W. E. Baker et al.: *A Short Course on Explosion Hazards Evaluation,* Southwest Research Institute, San Antonio, Tx, 1976.
17. W. E. Baker: *Explosions in Air,* Wilfried Baker Engineering, San Antonio 1973, (2nd Printing 1983).
18. R. A. Strehlow, W. E. Baker: *The Characterization and Evaluation of Accidental Explosions,* NASA CR 134 779, Urbana 1975.
19. VDI-Richtlinie, VDI 3783, Blatt 1, Dispersion of Emissions by Accidental Releases, Beuth Verlag, Berlin 1987.
20. Directive 92/32/EEC, ABIEG Nr. L 154 vom 5. 6. 1992, 7. Amendment to Framework Directive 67/548/EEC.
21. Deutsche Forschungsgemeinschaft, Senatskommission zur Prüfung gesundheitsschädlicher Arbeitsstoffe, 31. Mitteilung, Sep. 1, 1995.
22. Verordnung zum Schutz vor gefährlichen Stoffen (Gefahrstoffverordnung), Sep. 19, 1994, BGBl. I, p. 2557.
23. American Conference of Governmental Industrial Hygienists, Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices, ACGIH Technical Information Office, Cincinatti 1995.
24. Framework Directive 80/1107/EEC, Amendment Directive 88/642/EEC, Directive for Occupation Exposure Levels: 91/322/EWG.
25. Technische Regel für Gefahrstoffe (TRGs) 102, "Technische Richtkonzentration für gefährliche Stoffe," BArbBl. (1994) no. 2, p. 71.
26. Arbeitsplatzrichtwerte, BArbBl. no. 3/1991, p. 69;
27. Technische Regel für Gefahrstoffe (TRGS) 900, "Grenzwerte," BArbBl (1995) no. 4, p. 47.
28. H. Bender: *Sicherer Umgang mit Gefahrstoffen,* VCH Verlagsgesellschaft, Weinheim 1995.
29. H. H. Freytag: *Handbuch der Raumexplosionen,* Verlag Chemie, Weinheim 1965.
30. United Nations: *Recommendations on the Transport of Dangerous Goods,* 8th rev. ed., New York 1993.
31. VbF, Verordnung über Anlagen zur Lagerung, Abfüllung und Beförderung brennbarer Flüssigkeiten zu Lande mit Technischen Regeln für brennbare Flüssigkeiten (TRbF), C. Heymanns Verlag, Köln.

32. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* 2nd rev. ed., Deutscher Eichverlag, Braunschweig 1980.
33. VDI-Kommission Reinhaltung der Luft, VDI 2263, Blatt 1, Untersuchungsmethoden zur Ermittlung von sicherheitstechnischen Kenngrößen von Stäuben, Düsseldorf 1990.
34. Richtlinie 84/449/EWG, Prüfmethoden für physikalisch chemische Eigenschaften, Anhang A, veröffentlicht im Amtsblatt der Europäischen Gemeinschaft L 251/89.
35. J. Troitzsch: *International Plastics Flammability Handbook,* 2nd ed., Hanser Verlag, München 1990.
36. W. Jost: *Explosions- und Verbrennungsvorgänge in Gasen,* Springer Verlag, Berlin 1939.
37. B. Lewis, G. v. Elbe: *Combustion, Flames and Explosions of Gases,* 2nd ed., Academic Press, New York 1961.
38. H. H. Freytag: *Handbuch der Raumexplosionen,* Verlag Chemie, Weinheim 1965.
39. D. A. Franck-Kamenetzkii: *Stoff- und Wärmeübergang in der chemischen Kinetik,* Springer Verlag, Berlin 1959.
40. W. Bartknecht: *Explosionsschutz – Grundlagen und Anwendung,* Springer Verlag, Berlin 1993.
41. F. Funk: "Berechnung sicherheitstechnischer Kennzahlen," *Chem. Tech.* (Leipzig) 29 (1977) no. 9, 494 –497.
42. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* Deutscher Eichverlag, Braunschweig 1980/89 mit Nachträgen.
43. H. F. Coward, G. W. Jones: "Limits of Inflammability of Gases and Vapors," *Bull. U.S. Bur. Mines* **279** (1952) no. 503.
44. DIN 51 649,part 1, Bestimmung der Explosionsgrenzen von Gasen und Gasgemischen in Luft, Beuth Verlag, Berlin 1986. pr EN 1839, Determination of Explosion Limits of Gases, Vapours and their Mixtures.
45. H. Le Chatelier, *Ann. Mines.* **(Ser. 8) 19** (1891) 388 –395.
46. W. Berthold, U. Löffler: *Lexikon Sicherheitstechnischer Begriffe in der Chemie,* Verlag Chemie, Weinheim 1981.
47. D. Conrad: "Inertisierung explosionsfähiger Gassysteme als Maßnahme des primären Explosionsschutzes," *2. Int. Kolloquium der IVSS,* Frankfurt 1973.
48. Hauptverband der gewerblichen Berufsgenossenschaften, *Explosionsschutz-Richtlinien (EX-RL),* C. Heymanns Verlag, Köln 1990.
49. M. Müller, H. Schacke, C.-D. Walther: Explosion Characteristics of Carbon Monoxide at High Temperatures, Loss Prevention and Safety Promotion in the Process Industries, *Proc. Int. Symp. 6th,* Oslo, June 1989, vol IV, p. 104/104-12, European Federation of Chemical Engineering, 1989.
50. D. Conrad, R. Kaulbars: "Druckabhängigkeit der Explosionsgrenzen von Wasserstoff," *Chem. Ing. Tech.* **67** (1995) no. 2, 185 – 188.
51. H.-J. Heinrich: "Bemessung von Druckentlastungsöffnungen zum Schutz explosionsgefährdeter Anlagen in der chemischen Industrie," *Chem. Ing. Tech.* **38** (1966) 1125 – 1133.
52. H.-J. Heinrich: "Zur Bemessung von Druckentlastungsöffnungen bei Gas- und Staubexplosionen," Wissenschaftliche Berichte aus der Bundesanstalt für Materialprüfung (BAM), Berlin 1973, pp. 277 –280.
53. S. Crescitelli, G. Russo, V. Tufano, *J. Occup. Accid.* **2** (1979) 125 – 133.
54. W. Jost: *Probleme der Flammenfortpflanzung,* report no. 226, Deutsche Versuchsanstalt für Luft- und Raumfahrt e.V., Oberpfaffenhofen 1962, pp. 3 – 24.
55. J. Nagy et al.: "Explosion Development in Closed Vessels," *Rep. Invest. U.S. Bur. Mines* **7507** (1971) 1 – 50.
56. D. Conrad, S. Dietlen: "Untersuchungen zur Zerfallsfähigkeit von Distickstoffoxid," *BAM-Forschungsbericht,* no. 89, Verlag für neue Wissenschaft, Bremerhafen 1983.
57. H. R. Christen: *Thermodynamik und Kinetik chemischer Reaktionen,* Diesterweg u. Salle, Frankfurt 1974.
58. D. Conrad: "Vermeidung von Gefahren beim Umgang mit zerfallfähigen Gasen," *2. Sicherheitstechnische Vortragsveranstaltung über Fragen des Explosionsschutzes,* PTB Braunschweig, March 1983.
59. D. Conrad, R. Kaulbars: "Untersuchungen zur chemischen Instabilität von Äthylen," *Chem. Ing. Tech.* **47** (1975) 265.
60. D. Oberhagemann: "Zündtemperaturen von Ein- und Mehrkomponentensystemen," *VDI-Fortschrittber.* , Reihe 3, no. 185, VDI Verlag, Düsseldorf 1989.
61. P. Field: *Dust Explosions,* Elsevier Sci. Publ., Amsterdam 1982.

62. J. Cross, D. Farner: *Dust Explosions,* Plenum Press, New York 1982.
63. W. E. Baker et al.: *Explosion Hazards and Evaluation,* Elsevier Sci. Publ., Amsterdam 1983.
64. M. Glor: *Electrostatic Hazards in Powder Handling,* Research Studies Press, Letchworth, 1983.
65. R. K. Eckhoff: *Dust Explosions in the Process Industries,* Butterworth-Heinemann, Oxford 1991.
66. W. Bartknecht: *Explosionsschutz, Grundlagen und Anwendung,* Springer Verlag, Berlin 1993.
67. VDI-Guideline 2263, part 1, Test Methods for the Determination of the Safety Characteristics of Dusts, Beuth Verlag, Berlin 1990.
68. W. Berthold: "Bestimmung der Mindestzündenergie von Staub/Luft-Gemischen," *VDI Fortschrittber.* **134, Reihe 3,** VDI-Verlag, Düsseldorf 1987.
69. ISO 6184/1- 1985 (E) Explosion Protection Systems, part 1, Determination of Explosion Indices of Combustible Dusts in Air.
70. VDI-Guideline 3673, part 1, Pressure Release of Dust Explosions, Beuth Verlag, Berlin 1979 (Draft 1992).
71. ISO 6184/4- 1985 (E) Explosion Protection Systems, part 4, Determination of Explosion Suppression Systems.
72. J. Zehr: "Anleitungen zu den Berechnungen über die Zündgrenzwerte und die maximalen Explosionsdrücke," *VDI-Ber.* **19** (1975) 62.
73. J. Schönewald: "Vereinfachte Methode zur Berechnung der unteren Zündgrenze von Staub-Luft-Gemischen," *Staub Reinhalt. Luft* **31** (1971) no. 9, 376.
74. DIN 55 990,part 6, Pulverlacke, Berechnung der unteren Zündgrenze.
75. L. Bretherick: *Handbook of Reactive Chemical Hazards,* 4th ed., Butterworths, London 1990.
76. R. King: *Safety in the Process Industries,* Butterworths-Heinemann, London 1990.
77. T. Grewer, O. Klais: *Exotherme Zersetzung – Untersuchung der charakteristischen Stoffeigenschaften,* VDI-Verlag, Düsseldorf 1988.
78. W. H. Seaton, E. Freedman, D. N. Treweek: "CHETAH: The ASTM Chemical Thermodynamic and Energy Release Potential Evaluation Program," *ASTM Data Ser.* **DS 51** (1974).
79. M. A. Cook: *The Science of High Explosives,* Robert E. Krieger Publ. Corp., Huntington, New York, 1971.
80. J. Taylor: *Detonation in Condensed Explosives,* Clarendon Press, Oxford 1952.
81. K. K. Andreyev, A. F. Belyayev: *Theorie der Explosivstoffe* (Translation from Russian into German by Svenska Nationalkommittén för Mekanik, Sektionen för Detonik och Förbränning, Stockholm 1964).
82. T. M. Groothuizen, J. W. Hartgerink, H. J. Pasman in C. H. Buschmann (ed.): "Phenomenology, Test Methods and Case Histories of Explosions in Liquids and Solids," *First International Loss Prevention Symposium Hague/Delft, May 28 – 30, 1974,* Elsevier Sci. Publ. Corp., Amsterdam – London – New York 1974.
83. H. Koenen, K. H. Ide, K.-H. Swart: "Sicherheitstechnische Kenndaten explosionsfähiger Stoffe," *Explosivstoffe* **9** (1961) 4, 30, 195.
84. K.-H. Swart, P.-A. Wandrey, K. H. Ide, E. Haeuseler: "Sicherheitstechnische Kenndaten explosionsfähiger Stoffe," *Explosivstoffe* **13** (1965) 339.
85. Recommendations on the Transport of Dangerous Goods – Test and Criteria, 3rd ed., United Nations, New York 1995.
86. Recommendations on the Transport of Dangerous Goods, 9th ed., United Nations, New York 1995.
87. EG A 14 Explosionsgefahr, Richtlinie 92/69/EWG, Amtsblatt der Europäischen Gemeinschaft L 383 A, 12/95.
88. J. Barton, R. Rogers: Guide on Chemical Reaction Hazards, Institute of Chemical Engineers (IChmE), Rugby 1993.
89. ESCIS-Booklet 8, 1993: Thermal Process Safety Data, Assessment Criteria, Measures (available from SUVA, CH-6001 Lucerne).
90. N. N. Smenov, *Z. Phys. Chem.* **48** (1928).
91. D. A. Frank-Kamenetskii, *Dokl. Akad. Nauk SSSR* **18** (1938) 1.
92. P. C. Bowes: *Self-heating: Evaluating and Controlling the Hazard,* Her Majesty's Stationery Office (HMSO), London 1984.
93. N. Gibson, Conference on Chemical Reaction Hazards, London 1993.
94. O. Levenspiel, *Chemical Reaction Engineering,* J. Wiley & Sons, New York 1972.
95. K. R. Westerterp, W. P. M. van Swaaij, A. A. C. M. Beenackers: *Chemical Reactor Design and Operation,* J. Wiley & Sons, New York 1984.
96. J. Steinbach, Habilitationsschrift, TU Berlin 1994.

97. P. Hugo, *Chem. Ing. Tech.* **52** (1980) 712.
98. T. Grewer et al., *Dechema Monogr.* **111** (1987) 41 – 65.
99. J. Steinbach: *Chemische Sicherheitstechnik,* 1st ed., VCH Verlagsgesellschaft, Weinheim 1995.
100. J. Steinbach, *Dechema Monogr.* **107** (1986) 133 – 152.
101. D. I. Townsend, J. C. Tou, *Thermochim. Acta* **37** (1980) 1.
102. J. Wiss, G. Killé, F. Stoessel, *Chimia* **47** (1993) 417 –423.
103. F. Stoessel, *Chem. Eng. Prog.* **89** (1993) no. 10, 68 – 75.
104. P. Hugo, J. Steinbach, F. Stoessel, *Chem. Eng. Sci.* **43** (1988) no. 8, 2147 – 2152.
105. DECHEMA: *Safety Standards and Regulations for Chemical Plant in Europe: A Comparison,* Frankfurt/Main 1988 (Praxis der Sicherheitstechnik, vol. 2).
106. V. Pilz: "Der Technische Ausschuß für Anlagensicherheit – Wegbereiter für ein ganzheitliches Regelwerk zur Sicherheitstechnik aus immissionsschutzrechtlicher Sicht," *UMWELT* (1992) no. 11, 415 – 416.
107. European Council Directive on the Major Accident Hazards of Certain Industrial Activities (the "Seveso Directive"), 82/501/EEC, June 24 1982, European Community, Brussels.
108. "EWG-VERTRAG, Grundlage der Europäischen Gemeinschaft," Europa-Union Verlag, Bonn 1988.
109. P. Marburger: "Die Regeln der Technik im Recht," C. Heymanns Verlag, Köln 1979.
110. J. M. Ale: "Risk Analysis and Risk Policy in the Netherlands and the EEC," *J. Loss Prev. Process Ind.* **4** (1991) 58 – 64.
111. T. A. Kletz: "Inherently Safer Plants: An Update," *Plant/Oper. Prog.* **10** (1991) no. 2, 81 – 84.
112. V. Pilz: "Integrierte Sicherheit bei verfahrenstechnischen Anlagen", *UMWELT* **19** (1989) no. 5, D 27 – D 30.
113. V. Pilz: "Bayer's Procedure for the Design and Operation of Safe Chemical Plants," paper presented at the *XIII World Congress on Occupational Safety and Health,* New Delhi, India, April 1993.
114. Bayer: "Directive Process and Plant Safety," 2nd ed., Leverkusen 1989.
115. BASF: *Sicherheitsbetrachtungen bei der Planung von Chemieanlagen,* 2nd ed., Ludwigshafen 1989.
116. Hoechst: *Richtlinie zur Sicherheitsüberprüfung von Verfahren und Produktionsanlagen (Überprüfungsrichtlinie),* Nov. 1989, Frankfurt.
117. "Einführung in die Risikoanalyse," *Schriftenreihe der Expertenkommission für Sicherheit in der chemischen Industrie der Schweiz (ESCIS),* 2nd ed., **no. 4,** Basel 1986.
118. "Behelf für die Durchführung von Sicherheits-überprüfungen (Safety Audits)," *Schriftenreihe der Expertenkommission für Sicherheit in der chemischen Industrie der Schweiz (ESCIS),* no. 9, Basel 1991.
119. K. A. Ruppert: "Sicherheitsanalytische Vorgehensweise für Alt- und Neuanlagen," *Chem. Ing. Tech.* **62** (1990) no. 11, 916 – 927.
120. Lurgi: *Sicherheitsleitfaden,* Frankfurt, Nov. 1990.
121. A. W. Bickum, A. E. Barnette: "Common Documentation Systems for ISO 9000 and Process Safety Management," *Preprints Int. Process Safety Management Conference and Workshop (AIChE-CCPS),* San Francisco, Sept. 1993.
122. H. W. Adams: "Erhöhung der Sicherheit durch Qualitätssicherung bei Planung, Bau und Betrieb von chemischen Produktionsanlagen," *Forschungsbericht 10 409 221,* Umweltbundesamt, Texte 18/93.
123. V. Pilz: "Safety Analyses for the Systematic Checking of Chemical Plant and Processes – Methods, Benefit and Limitations," *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) 65 – 74; *Chem. Ing. Tech.* **57** (1985) no. 4, 289 – 307.
124. "Guidelines for Hazard Evaluation Procedures," 2nd ed., Center for Chemical Process Safety (CCPS) of AIChE, New York 1992.
125. G. L. Wells: *Safety in Process Plant Design,* J. Wiley & Sons, New York 1980.
126. F. P. Lees: *Loss Prevention in the Process Industries,* **vol. 1,** Butterworths, London 1980, chaps. 8 and 9.
127. K. Bartels et al.: *Risikobegrenzung in der Chemie, PAAG-Verfahren (HAZOP),* BG Chemie/IVSS, Heidelberg 1990.
128. H. G. Lawley: "Operability Studies and Hazard Analysis," *Chem. Eng. Progr.* **70** (1974) no. 4, 45 –56.
129. DIN 25 419,Ereignisablaufanalyse, Beuth Verlag, Berlin, Nov. 1985.
130. DIN 25 424,Fehlerbaumanalyse, Beuth Verlag, Berlin, Sep. 1981.

131. V. Pilz: "Risk Analysis for Chemical Production Processes? – Some Remarks on Meaningful Applications of Available Methods and Limitations Thereof," *Angew. Systemanalyse* **2** (1981) no. 4, 175 – 178.
132. H. Dembeck: *Chemie-ABC für Feuerwehr- und Sicherheitsfachkräfte,* Kohlhammer Deutscher Gemeindeverlag, Stuttgart 1981.
133. A. Rempe, G. Rodewald: *Brandlehre,* Kohlhammer Deutscher Gemeindeverlag, Stuttgart 1988.
134. J. Troitzsch, *Brandverhalten von Kunststoffen,* Hanser Verlag, München 1982.
135. S. Bussenius, *Brand- und Explosionsschutz in der Industrie,* Staatsverlag der DDR, Berlin 1982.
136. F. Kaufhold, A. Rempe: *Feuerlöschmittel,* Kohlhammer Deutscher Gemeindeverlag, Stuttgart 1990.
137. *Vereinigung zur Förderung des deutschen Brandschutzes, VFDB* **2** (1986).
138. W. M. Predtetschenski, A. I. Milinski: *Personenströme in Gebäuden,* Verlagsgesellschaft R. Müller, Köln 1971.
139. K. Klingsohr, *Vorbeugender baulicher Brandschutz,* Kohlhammer Deutscher Gemeindeverlag, Stuttgart 1986.
140. H. P. Morgan, *Rauchschutzmethoden in ein- oder mehrgeschossigen, geschlossenen Einkaufszentren: Eine Entwurfsübersicht,* Colt, Kleve 1988.
141. Richtlinien des Verbandes der Sachversicherer, Köln, zu: Brandmeldeanlagen; Rauch- und Wärmeabzugsanlagen; Sprinkleranlagen; CO2-Feuerlöschanlagen.
142. Fire Resistance Tests, European standards EN 1363 – EN 1366, draft 1994.
143. Verband der Sachversicherer: *VdS-Fachtagung 1990: Lagerung von gefährlichen Stoffen,* Köln 1990.
144. Seminarberichte des Instituts für Baustoffe, Massivbau und Brandschutz, Brandschutz im Industriebau bzw. Brandschutz – Forschung und Praxis, Braunschweig 1987 – 1992.
145. J. Steinmetz, E. Merz: *Umweltschutz und Gefahrenabwehr als betriebliches Gesamtkonzept,* Boorberg Verlag, Stuttgart 1992.
146. H. H. Freytag: *Handbuch der Raumexplosion,* Verlag Chemie, Weinheim 1965.
147. H. Schacke: "Fact-finding and Basic Data/Fire and Explosion, Safety in Chemical Production," *Proceedings of the First IUPAC Workshop on Safety in Chemical Production,* Blackwell Scientific Publications, Oxford 1991, pp. 21 – 26.
148. Berufgenossenschaft der chemischen Industrie: *Explosionsschutz-Richtlinien (EX-RL),* issue 9/94, Druckerei Winter, Heidelberg 1994.
149. prEN 1127-1 Safety of Machinery, Fire and Explosions, Part 1, Explosion Prevention and Protection/Draft Standard CEN TC 114 WG 16, Apr. 1995.
150. NFPA 497 A-1986 Standard-National Electrical Code (USA).
151. prEN 50 145 Electrical Apparatus for Potentially Explosive Gas Atmospheres/Classification of Hazardous Areas, draft, March 1993.IEC-SC 31 J Revision of publication 79-10 transforming the publication in IEC Standard: Electrical apparatus for explosive atmospheres/Area classification and installation requirements/Classification of hazardous areas, to be issued as a DIS, 1995.
152. A. W. Cox, F. P. Lees, M. I. Ang: *Classification of Hazardous Locations,* Institution of Chemical Engineers, Rugby, UK 1990.
153. W. Bartknecht: *Explosionsschutz – Grundlagen und Anwendung,* Springer Verlag, Heidelberg 1993.
154. K. Ritter: "Mechanisch erzeugte Funken als Zündquellen," *VDI-Ber.* **494** (1984) 129 – 144.
155. DIN VDE 0165/02.91, Errichten elektrischer Anlagen in explosionsgefährdeten Bereichen, Beuth Verlag, Berlin 1991.
156. G. Lüttgens, M. Glor: *Elektrostatische Aufladungen begreifen und sicher beherrschen,* expert Verlag, Ehningen 1993. G. Lüttgens, M. Glor: *Understanding and Controlling Static Electricity,* expert Verlag, Ehningen 1989.
157. Berufgenossenschaft der Chemischen Industrie: Richtlinien für die Vermeidung von Zündgefahren infolge elektrostatischer Aufladungen, issue 10/89, ZH 1/200, C. Heymanns Verlag, Köln 1989.
158. R. Viard, H. Schacke, C.-D. Walther: "Structurized Explosion Prevention and Protection – An Adequate and Economical Way to Satisfy Safety Requirements," *Loss. Prev. Saf. Promot. Process Ind. Int. Symp.* **8th** (1995) 527 – 538.
159. Berufsgenossenschaft der chemischen Industrie: Unfallverhütungsvorschrift Gase (VBG 61), Jedermann-Verlag Dr. O. Pfeffer, Heidelberg 1995.
160. VDI 3673, Pressure Venting of Dust Explosions, part 1, Beuth Verlag, Berlin 1995.

161. Safety of Machinery – Industrial Trucks – Operation in Potentially Explosive Atmospheres, CEN TC 150 WG 7, draft standard, May 1994.
162. Directive 94/9/EC of the European Parliament and the Council.
163. Draft Proposal for a Directive Concerning the Minimum Requirements for Improving the Safety and Health Protection of Workers Potentially at Risk from Explosive Atmospheres, EU-Commission, Doc. No. 427/4/93.
164. Directives 89/392/EEC and 91/368/EEC of the European Parliament and the Council.
165. Directives 76/117/EEC, 79/196/EEC, 82/130/EEC, and 90/487/EEC of the European Parliament and the Council.
166. NAMUR (eds.): NAMUR-Empfehlung "Anlagensicherung mit Mitteln der Prozeßleittechnik," NE 31, NAMUR-Geschäftsstelle Bayer AG, Leverkusen, Jan. 1993.
167. VDI/VDE-Richtlinie 2180, Sicherung von Anlagen der Verfahrenstechnik mit Mitteln der Meß-, Steuerungs- und Regelungstechnik, "Einführung, Begriffe, Erklärungen," Blatt 1, April 86; "Berechnungsmethoden für Zuverlässigkeitskenngrößen von Sicherungseinrichtungen," Blatt 2, April 86; "Klassifizierung von Meß-, Steuerungs- und Regelungseinrichtungen," Blatt 3, Dec. 84; "Ausführung und Prüfung von Schutzeinrichtungen," Blatt 4, July 88; "Bauliche und installationstechnische Maßnahmen zur Funktionssicherung von Meß-, Steuerungs- und Regelungseinrichtungen in Ausnahmezuständen," Blatt 5, Dec. 84.
168. DIN V 19 250,Grundlegende Sicherheitsbetrachtungen für MSR-Schutzeinrichtungen, Mai 1994.
169. DECHEMA: "Praxis der Sicherheitstechnik, ein Leitfaden für Planung, Bau und Betrieb chemischer Produktionsanlagen," *Anlagensicherung mit Mitteln der MSR-Technik,* vol. 1, Frankfurt/Main 1988.
170. DIN VDE 31 000,Allgemeine Leitsätze für das sicherheitsgerechte Gestalten technischer Erzeugnisse, Begriffe der Sicherheitstechnik, part 2, Dec. 1987: Grundbegriffe.
171. AD-Merkblatt A 1, Berstsicherungen, C. Heimanns Verlag, Köln 1992.
172. AD-Merkblatt A 2, Sicherheitsventile, C. Heimanns Verlag, Köln 1993.
173. DIN 3320,part 1, Sicherheitsventile – Begriffe Größenbemessung Kennzeichnung, 1984.
174. ISO 4126,part 1, Safety Valves, draft 1994.
175. ISO 4126,part 5, Controlled Safety Pressure Relief Systems, draft 1994.
176. ISO 6718,Bursting Discs and Bursting Disc Devices, 1991.
177. R. Bozòki: *Überdrucksicherungen für Behälter und Rohrleitungen,* Verlag TÜV Rheinland, Köln 1986.
178. C. M. Sheppard: "Diers Bubbly Disengagement Correlation Extended to Horizontal Cylinders and Spheres," *J. Loss Prev. Ind.* **7** (1994) no. 1, 3 – 5.
179. E. Molter: "Druckentlastung von Gas-/Dampf-Flüssigkeitsgemischen," PhD Thesis, Universität Dortmund 1991.
180. F. Mayinger: *Strömung und Wärmeübergang in Gas/Flüssigkeitsgemischen,* Springer Verlag, Berlin 1982.
181. Hauptverband der gewerblichen Berufsgenossenschaften, ZH 1/621 Technische Regeln Druckbehälter, TRB 404, 6.1 – 6.3, C. Heimanns Verlag, Köln, Apr. 1989, p. 27.
182. S. Muschelknautz: *Mechanische Phasentrennung bei Entspannungsverdampfung, VDI-Forschrittsber.* **72** (1990) Reihe 15, Umwelttechnik.
183. R. Spatz, E. Molter, H. Schoft: "Auffangsysteme zur Entsorgung der aus Druckentlastungseinrichtungen abgeblasenen Stoffströme," *Chem. Ing. Tech.* **63** (1991) no. 3, 233 – 236.
184. Technischer Ausschuß für Anlagensicherheit beim Bundesminister für Umwelt, Naturschutz und Reaktorsicherheit, Gesellschaft für Reaktorsicherheit GSR, Köln. Sicheres Rückhalten von gesundheitsgefährlichen Stoffen aus Druckentlastungseinrichtungen, TAA-GS-06.
185. VDI-Richtlinie, VDI 3783, Ausbreitung von störfallbedingten Freisetzungen – Sicherheitsanalyse, Beuth Verlag, Köln.
186. V. Pilz: "Auslegung, Einsatz und Wirkungsweise von Blow-Down-Systemen als Sicherheitseinrichtungen von Chemieanlagen," *Chem. Ing. Tech.* **49** (1977) no. 11, 873 – 883.
187. K. Beher, A. Steiff, P.-M. Weinspach: "Direktkondensation inertgashaltiger Dämpfe mit Hilfe von Tauchvorlagen," *Chem. Ing. Tech.* **62** (1990) no. 2, 136 – 137.
188. K. Herrmann, H.-G. Schecker, H. Schoft: "Direktkondensation notentspannter Dampf/Gas-Gemische in einem

Strahlkondensator," *Chem. Ing. Tech.* **64** (1992) no. 2, 183 – 184.
189. H. Schoft, R. Spatz: "Direktkondensation in Strahlapparaten zur Niederschlagung entlasteter Dampf/Gas-Gemische," *Chem. Ing. Tech.* **65** (1993) no. 6, 735 – 739.
190. AD-Merkblatt A 6, C. Heymanns Verlag, Köln.
191. P. Walzel, H. Schoft: "Einmischen von Notabstoppern in chemische Reaktoren," *Chem. Ing. Tech.* **65** (1993) no. 4, 447 – 449.
192. K. Schampel: *Flammendurchschlagsicherungen,* Expert Verlag, Ehningen 1988.
193. D. Lietze, *J. Occup. Accid.* **5** (1983) 17 – 37.
194. G. G. Börger, W. Hüning, H. U. Nentwig, M. Schweitzer, *Chem. Ing. Tech.* **55** (1983) 396.
195. H. Förster, *MariChem93 Conference Papers,* Sect. 8.1, Gastech RAI Ltd., London 1993.
196. VbF, Verordnung über Anlagen zur Lagerung, Abfüllung und Beförderung brennbarer Flüssigkeiten zu Lande mit Technischen Regeln für brennbare Flüssigkeiten (TRbF), C. Heymanns Verlag, Köln 1994.
197. H. H. Freytag: *Handbuch der Raumexplosionen,* Verlag Chemie, Weinheim 1965.
198. H. Förster, H. Leinemann: *Neuester Stand der Explosionsschutzmaßnahmen mit flammendurchschlagsicheren Einrichtungen,* Unterlagen zum Lehrgang der Technischen Akademie Esslingen, 1990.
199. K. Nabert, G. Schön: *Sicherheitstechnische Kennzahlen brennbarer Gase und Dämpfe,* 2nd ed., Deutscher Eichverlag, Braunschweig 1980.
200. VDI-Richtlinie 2263, VDI Kommission Reinhaltung der Luft, Düsseldorf, Nov. 1986.
201. Braunschweiger Flammenfilter GmbH, Braunschweig, catalog.
202. V. Pilz, *Chem. Ing. Tech.* **57** (1985) no. 4, 289 – 307; *Ger. Chem. Eng. (Engl. Transl.)* **9** (1986) 65 – 74.
203. Berufsgenossenschaft der chemischen Industrie: Technisches Merkblatt T 025, Sicherer Umgang mit Flüssigkeiten, part 1: Umfüllen.
204. Berufsgenossenschaft der chemischen Industrie: Technisches Merkblatt T 026, Sicherer Umgang mit Flüssigkeiten, part 2: Probenahme.
205. Berufsgenossenschaft der chemischen Industrie: Technisches Merkblatt T 006, Reinigen von Behältern.
206. Berufsgenossenschaft der chemischen Industrie, Verein Deutscher Sicherheitsingenieure: Ratgeber Anlagensicherheit, Verlag Kluge, Berlin.
207. Berufsgenossenschaft der chemischen Industrie: *Betriebsanweisung für den Umgang mit Gefahrstoffen,* Merkblatt A 010, Jedermann-Verlag Dr. Otto Pfeffer, Heidelberg 1992.
208. T. A. Kletz: "Managing Risk in Chemical Manufracture," *Proc. Int. Conf.,* Cambridge, Jul. 13 – 16, 1992, R. Soc. Chem. 1992, 92 – 105.
209. R. Leicht, K. A. Ruppert: "Sicherheitstechnische Dokumentation in Chemieanlagen," *Tech. Überwach.* **34** (1993) no. 1, 22 – 25.
210. Y. Sakami: "Safety Management at Chemical Plants in Japan," *Proceedings of Second IUPAC-Workshop on Safety in Chemical Production,* Yokohama, May 31 – June 4, 1993 (in preparation).
211. W. E. Hoffmann: *Unabhängig und neutral. Die TÜV und ihr Verband VdTÜV,* Wirtschaftsverlag, Wiesbaden 1986.
212. G. Wiesenack: *Wesen und Geschichte der Technischen Überwachungsvereine,* C. Heymanns Verlag, Köln 1971.
213. W. E. Hoffmann: *Erfahrungsaustausch in der Technischen Überwachung, Technische Eigenüberwachung in der Chemie,* Verlag Wissenschaft und Politik, Köln 1982.
214. H. Pohle: *Chemische Industrie Umweltschutz, Arbeitsschutz, Anlagensicherheit,* VCH Verlagsgesellschaft, Weinheim 1991.
215. W. Schoch: *Selbstverwaltung der Wirtschaft auf dem Gebiet der Technischen Überwachung, Sicherheit in der Chemie,* Verlag Wissenschaft und Politik, Köln 1979.
216. Vergleich des Prüf- und Sachverständigenwesens in Großbritannien, Frankreich und Deutschland am Beispiel der Druckgeräte, Arthur D. Little International, Wiesbaden 1992.
217. DECHEMA: *Safety Standards and Regulations for Chemical Plant in Europe: A Comparison,* Frankfurt/Main 1988.
218. V. C. Marshall, *Major Chemical Hazards,* Ellis Harwood, Chichester 1987.
219. N. M. Azarkin: "Unflagging Attention to the Preparation of Chemical Equipment for Maintenance Operations, in which there is Hazard of Fire," *Sov. Chem. Ind.* (Engl. Transl.) 20 (1988) no. 7, 59 – 61.
220. *Dangerous Maintenance, Health and Safety Executive,* Library and Information Services, Sheffield 1987.

221. K. Schwab, "Praxisbericht, Vorbeugende Instandhaltung (VI) und Qualitätssicherung (QS)", *Konferenz Sicherheits- und Instandhaltungsmanagement in der chemischen Industrie,* VCI, Darmstadt 1993.
222. U. Olsen, "Inspektion und Wartung," in *Die zustandsorientierte Maschineninstandhaltung in der Praxis, 5. Instandhaltungs-Forum,* Reihe Praxiswissen für Ingenieure – Instandhaltung, TÜV Rheinland, Köln 1989, pp. 95 – 123.
223. H.-H. Rauschhofer: "Vorbeugende Instandhaltung ist ein Garant auch für Sicherheit," *Maschinenmarkt/MM Industriejournal* **87** (1981) no. 37, 737 –740.
224. S. Pradham: "Apply Reliability Centered Maintenance to Sealless Pumps," *Hydrocarbon Process.* **72** (1993) no. 1, 43 – 47.
225. N. Hering, Verbesserung der Voraussetzungen zum Optimieren von Sicherheit und Wirtschaftlichkeit im Industrieanlagenbetrieb, *VDI-Z.* **134** (1992) 22 –28.
226. Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaft, Luxemburg: Richtlinie des Rates über die Gefahren schwerer Unfälle bei bestimmten Industrietätigkeiten (82/501/EWG) (mit 1. und 2. Änderungsrichtlinie), ABl. Nr. L 230 vom 05. 08. 1982 (Abl. Nr. L 85 vom 28. 03. 1987, Abl. Nr. L 336 vom 07. 12. 1988).
227. Amt für amtliche Veröffentlichungen der Europäischen Gemeinschaft, Luxemburg: Vorschlag für eine Richtlinie des Rates zur Abwehr der Gefahren schwerer Unfälle mit gefährlichen Stoffen vom 26. 01. 1994, Katalog-Nr. CB-CO-94-016-DE-C.
228. Verordnung über Druckbehälter, Druckgasbehäl-ter und Füllanlagen (Druck-behälterverordnung –DruckbehV)/Technische Regeln zur Druckbehälterverordnung (z. B. TRB 404 Nr. 6, TRB 600 Nr. 3.4), C. Heymanns Verlag, Köln.
229. J. Steinmetz, E. Merz: *Umweltschutz und Gefahrenabwehr als betriebliches Gesamtkonzept,* Boorberg Verlag, Stuttgart 1992.
230. § 19 g ff Wasserhaushaltsgesetz (WHG), *Bundesgesetzblatt,* 1986, part 1, pp. 1536 ff.
231. Verordnung über Anlagen zum Umgang mit wassergefährdenden Stoffen und über Fachbetriebe (VAwS) – (des jeweiligen Bundeslandes, z. B. NW, Gesetz und Verordnungsblatt für das Land Nordrhein-Westfalen – no. 56, pp. 676 ff).
232. Katalog wassergefährdender Stoffe Bekanntmachung des BMI, March 1, 1985 – U III 6523074BGemeinsames Ministerialblatt, 36. Jahrgang, no. 11, pp. 175 ff und deren Fortschreibungen.
233. Allgemeine Verwaltungsvorschrift v. 09. 03. 1990 über die nähere Bestimmung wassergefährdender Stoffe und ihre Einstufung entsprechend ihrer Gefährlichkeit – VwV wassergefährdender Stoffe (VwVwS), Gemeinsames Ministerialblatt, 41. Jahrgang, no. 8, pp. 114 ff.
234. Verordnung über Anlagen zur Lagerung, Abfüllung und Beförderung brennbarer Flüssigkeiten zu Lande (Verordnung über brennbare Flüssigkeiten – VbF) und Allgemeine Verwaltungsvorschrift mit den zugehörigen Technischen Regeln für brennbare Flüssigkeiten (TRbF), C. Heymanns Verlag, Berlin – Köln.
235. Verband der Technischen Überwachungs-Vereine e.V., VdTÜV-Merkblatt "Auffangräume für die Lagerung brennbarer und nichtbrennbarer wassergefährdender Flüssigkeiten," Essen 1992.
236. Verband der Chemischen Industrie: Sicherheitskonzept für Anlagen zum Umgang mit wassergefährdenden Stoffen, Frankfurt/Main 1987.
237. B. Wittke: "Lager- und Abfüllanlagen für Gefahrstoffe," *VDI-Ber.* no. 726 (1989).
238. V. Papenhausen: "Standardisierung von Ableitflächen," *VDI-Ber.* no. 869 (1991).
239. DIN Deutsches Institut für Normung e.V.:Richtlinie des DAfStB – Unbeschichtete Betonbauteile beim Umgang mit wassergefährdenden Stoffen, Berlin 1994.
240. DIN 6608 Teil 2,Liegende Behälter (Tanks) aus Stahl, doppelwandig, für die unterirdische Lagerung wassergefährdender, brennbarer und nichtbrennbarer Flüssigkeiten, Beuth Verlag, Berlin.
241. TRbF 501, Richtlinie/Bau- und Prüfgrundsätze für Leckanzeigegeräte für Behälter.
242. TRbF 502, Richtlinie/Bau- und Prüfgrundsätze für Leckanzeigegeräte für doppelwandige Rohrleitungen.
243. Sicherheitsregeln für Anforderungen an Eigenschaften ortsfester Gaswarneinrichtungen für den Explosionsschutz, ZH 1/8. Hauptverband der gewerblichen Berufsgenossenschaften, Fachausschuß Chemie, C. Heymanns Verlag, Köln.

244. Umgebungsüberwachung von verfahrenstechnischen Freiluftanlagen durch Gasdetektoren, Abschlußbericht des F+E-Vorhabens Nr. 104 09 215 des Umweltbundesamtes vom 12. 08. 1991.
245. S. Muschelknautz, K. Nießer: "Den Schaden begrenzen," *Chem. Ind.* no. 9 (1993) 31 – 33.
246. Sicherheitstechnische Hinweise und Anforderungen an Abschott- und Entlastungssysteme aus Sicht der Störfall-Verordnung (Entwurf), Ausarbeitung eines Arbeitskreises bei der Landesanstalt für Immissionsschutz NRW, Essen.
247. V. Pilz: "Auslegung, Einsatz und Wirkungsweise von Blow-Down-Systemen als Sicherheitseinrichtung bei Chemieanlagen," *Chem. Ing. Tech.* **49** (1977) no. 11, 873 – 883.
248. Leitfaden Rückhaltung von gefährlichen Stoffen aus Druckentlastungseinrichtungen, Abschlußbericht eines Arbeitskreises des Technischen Ausschusses für Anlagensicherheit beim BMU, TAA-GS-06.
249. J. Seeger, E. Marx: "Hochfackeln zur Verbrennung von Industriegasen," *Gas Wärme Int.* **29** (1980) nos. 2/3, 91 – 98.
250. Hauptverband der gewerblichen Berufsgenossenschaften, Fachausschuß Chemie, Sicherheitsregeln für Anlagen zum Entfernen von Gasen und Dämpfen organischer Lösemittel aus der Abluft nach dem Adsorptionsverfahren, ZH 1/595, C. Heymanns Verlag, Köln.
251. Landesanstalt für Immissionsschutz NRW, LIS-Info Anlagensicherheit Nr. 2, Anforderungen an phosgenführende Anlagenteile, Essen.
252. V. Pilz: internal report, Bayer AG, Leverkusen, Mai 1992.
253. DOW Öffentlichkeitsarbeit: Die Herstellung von Polymeric-MDI . . . ein sicherer chemischer Prozeß.
254. Richtlinie zur Bemessung von Löschwasser-Rückhalteanlagen beim Lagern wassergefährdender Stoffe (LöRüRL), Ministerialblatt für das Land Nordrhein-Westfalen, no. 71 vom 20. 11. 1992, pp. 1719 ff
255. "Seveso-Richtlinie" 82/501/EWG, 1982, überarbeiteter Entwurf von 1992 in H. J. Uth: *Störfall-Verordnung,* Kommentar, 2nd ed., Bundesanzeiger Verlagsgesellschaft, Köln 1994.
256. Bekanntmachung der Neufassung der 12. Verordnung zur Durchführung des Bundes-Immissionsschutzgesetzes (Störfall-Verordnung), Sep. 20, 1991, BGBl. I, p. 1891.
257. NRW Arbeitskreis "Alarm- und Gefahrenabwehrplan": Betrieblicher Alarm- und Gefahrenabwehrplan, Stand 23. 08. 1993, Düsseldorf.
258. Leitlinie zur Erstellung betrieblicher Alarm- und Gefahrenabwehrpläne, VCI, Frankfurt/Main, Mar. 1988.
259. W. Steuer: Fachveranstaltung "Der chemische Unfall," Haus der Technik, Essen 1991.
260. Responsible Care, A Chemical Industry Commitment to Improve Performance in Health, Safety and the Environment, CEFIC Schrift 5/1993.
261. DFG MAK- und BAT-Werte-Liste, 2001, Wiley-VCH Weinheim

Plant Design → Chemical Plant Design and Construction

Occupational Health and Safety

Leopold W. Miksche, Bayer AG, Leverkusen, Federal Republic of Germany (Chap. 1, Section 4.6)

Gunter Meyer, IG-Chemie-Papier-Keramik, Hannover, Federal Republic of Germany (Section 2.1)

Jan J. Kolk, Arnhem, The Netherlands (Sections 2.2, 2.3 in part, 4.1, 4.2, 4.5)

Philip Noordam, The Hague, The Netherlands (Sections 2.2, 2.3 in part)

Donald W. Lamb, Miles Inc., Pittsburgh, United States (Sections 2.4, 3.10, 4.8)

Dan Bosatra, Dow Europe S. A., Horgen, Switzerland (Section 3.1–3.9)

Niko Kiesselbach, Bayer AG, Leverkusen, Federal Republic of Germany (Section 4.3)

Joachim Gruber, Bayer AG, Leverkusen, Federal Republic of Germany (Section 4.4)

Rainer Dosch, Bayer AG, Leverkusen, Federal Republic of Germany (Section 4.7)

1. Introduction 2
1.1. History 2
1.2. Differentiation of Tasks 3
1.3. Hazards in the Chemical Industry 3
1.4. General Outline of Goals 4
1.5. Experts and their Training 4
2. Legal Background 5
2.1. Laws and Regulations in Germany 5
2.1.1. State Legislation and Controls . . . 5
2.1.2. Trade Associations 10
2.1.3. Company Occupational Safety System 11
2.1.4. Role of the Social Partner and Associations 14
2.1.5. Future Developments 15
2.2. Legal Requirements in Other European Countries 15
2.3. European Legislation 19
2.4. Regulations in the United States . 19
3. Industrial Hygiene 20
3.1. Introduction and History 20
3.2. Requirements for an Effective Program 22
3.3. Objectives and Standards 22
3.4. Organization and Staff Management 22
3.5. Anticipation of Health Hazards . 23
3.6. Occupational Exposure Assessment 24
3.6.1. Introduction 24
3.6.2. Characterization and Identification 24
3.6.3. Preliminary Evaluation and Prioritization (Qualitative Assessment) . . 25
3.6.4. Occupational Exposure Limits . . . 26
3.6.5. Quantitative Assessment – Occupational Hygiene Monitoring . 27
3.6.6. Monitoring Planning and Strategy Definition 27
3.6.7. Sampling and Analysis; Data Quality Assurance 28
3.6.8. Sampling and Analysis Execution . 29
3.6.9. Data Interpretation and Reporting; Record Keeping 29
3.7. Exposure Control Methods 30
3.7.1. General Concepts 30
3.7.2. Personal Protective Equipment . . . 30
3.8. Physical Agents 31
3.9. Communication and Training . . 32
3.10. Industrial Hygiene in the United States 32
4. Occupational Medicine 33
4.1. Health Surveillance 33
4.1.1. Monitoring 34
4.1.2. Periodical Medical Examination . . 36
4.1.3. Evaluation and Interpretation of Data 37
4.1.4. Reporting Health Surveillance Results 38
4.1.5. Corrective Actions 38
4.2. Biological Monitoring 39
4.2.1. Basic Considerations 39
4.2.2. Prerequisites 40
4.2.3. Biological Media 40
4.2.4. Biological Occupational Limits . . . 41
4.2.5. Interpretation of Biological Monitoring Data 41
4.3. Occupational Epidemiology 43
4.3.1. Introduction 43
4.3.2. Basic Requirements – Data Collection 44
4.3.2.1. Data on Exposed Individuals in the Study Group 44

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_431

4.3.2.2. Technical Data on the Occupational Environment 44
4.3.3. Descriptive and Analytical Study Types 45
4.3.4. Usefulness of Occupational Epidemiological Studies 46
4.4. Ergonomics 47
4.4.1. Introduction; Definitions 47
4.4.2. Scope of Ergonomics 48
4.4.3. Stress Factors 48
4.4.4. Assessment and Testing Criteria . . 48
4.4.5. Objectives 50
4.5. Evaluation of Health Risks 50
4.5.1. Basic Considerations 50
4.5.2. Difficulties in Interpretation 51
4.5.3. Dose/Effect/Polymorphism 52
4.5.4. Hazard Identification 53
4.5.5. Risk Estimation 54
4.5.6. Risk Evaluation 54
4.5.7. Risk Limitation 54
4.6. Occupational Diseases and Work-Related Illnesses 54
4.6.1. Introduction; Definitions 54
4.6.2. Occupational Diseases 54
4.6.3. Prevention 56
4.6.4. Compensation for Occupational Diseases 56
4.6.5. Trends 57
4.6.6. Work-Related Illnesses 57
4.7. First Aid 57
4.8. Occupational Medicine in the United States 59
5. References 60

The entire topic was coordinated by Leopold W. Miksche

1. Introduction [1–3]

Work has always involved certain health hazards and risks of accidents. With progressing industrialization the pattern of potential threats to human health at work has been changing. At present, in a highly industrialized society, health protection and accident prevention at work are issues of high ethical and socioeconomic priority.

1.1. History

Work-related health impairments are not a discovery of modern times.

Signs of dust disease of the lungs have been discovered in prehistoric corpses. Scripts from ancient Egypt noted that farmworkers suffered from intense heat and dust on dry soil, whereas working in the muddy grounds along the Nile induced severe rheumatic disorders and chronic arthritis, which still can be seen in mummies. Reports and documentation of occupation-related diseases can be traced throughout ancient history. In ancient Greece Aristoteles and Plato described certain diseases found in specific professions such as "tanners, porters of heavy loads, and couriers." Around 400 B.C. Hippocrates noted: "there are many crafts and arts which cause a number of illnesses and adverse health conditions for those who work in these specific professions."

He also requested that physicians investigate thoroughly the professional history of their patients. Signs of pneumoconiosis and arthrosis of the knee in miners were described by him, as well as the toxicity of lead in lead miners.

Silicosis and diseases caused by lead and mercury or by certain melting procedures in gold mining are reported most frequently in the early European literature.

Georgios Agricola described miners diseases in the early 1500s, and Paracelsus wrote a book on occupational diseases in Germany at the same time.

The Italian physician Bernardino Ramazzini (1633 – 1714) has been called the father of occupational medicine. In *De Morbis Artificum Diatriba* he described different diseases of workmen and their treatment, and stressed the importance of prevention of occupational diseases and accidents to make work "a love rather than a curse." His book includes information on afflictions caused by inhaling noxious gases and dusts, as well as such "modern" issues as illnesses resulting from disorderly motion and improper posture of the body.

With increasing industrialization plant physicians began to take care of workers and their health in the early 1800s. In the second half of the 1800s legislation on health protection and

technical inspection of factories was established, for example, in England and Germany.

Some important events in the history of occupational health are summarized in the following:

B.C.

1500 Description of back pain and deformities in quarrymen (Papyros Ebers).

400 Hippocrates (460 – 377 B.C.) observed an occupational disease (colic) in slaves working in lead mines.

40 Pliny the Elder wrote about lead, mercury, sulfur, and zinc as occupational hazards. He also noted that in dusty conditions workers tied bladders over their mouths to breathe through, and that workers who used asbestos to make textiles became ill.

A.D.

250 Galen identified the danger of acid mists in copper mines.

1473 Ellenbog identified dangers of metal fumes in a report on gold miners. He also discussed carbon monoxide, lead, mercury, and nitric acid, and provided instructions in preventative measures.

1500 Paracelsus (1493 – 1541) wrote a book on occupational diseases, identifying diseases of miners, which he attributed to vapors from the metals. He advised that the vapors be avoided.

1500s Agricola (1494 – 1555) published *De Re Metallica*, a series of books on the hazards of mining. He recommended the use of ventilation in mines and of protective masks to limit exposure to dusts.

1700s *De Morbis Artificum*, a book dealing, among others, with mercury, lead, and the pathology of silicosis, was published by Bernardino Ramazzini, the "father of occupational medicine". He advised physicians to ask patients what their trade was, and to study work environments to learn about occupational diseases.

1775 Percival Pott (1714 – 1788), a London surgeon, identified cancer of the scrotum in chimney sweeps as the first occupational cancer.

1833 The English Child Labour Laws were passed, to limit childrens' working hours, and to provide a mechanism for inspections in certain industries.

1884 In Germany the law on accident insurance for workers was passed.

1925 The first official list of occupational diseases was passed in Germany.

1.2. Differentiation of Tasks

Historically health protection and provision of safety at the workplace has been taken care off mostly by physicians.

They tried to detect work-related health impairments and to discover the causes for these diseases in order to prevent them by changing work hygiene or other conditions at the workplace. In the 1800s close cooperation between physicians and technicians was necessary to improve production processes and working technology to prevent diseases and work-related accidents.

Government inspectors began supervising factories and mines in the second half of the 1800s. At the turn of the century, and specially in the first half of the 1900s, additional groups of specialists developed for specific tasks in the field of health and safety at work.

Nowadays there is a separation of certain tasks covered by different specialists (see Section 1.5). The occupational health physician is responsible for health protection, prevention of occupational illnesses and diseases, accident prevention, and medical surveillance of the workforce. Provision and control of technically safe workplaces, machinery, and working conditions and the technical side of accident prevention is dealt with by safety engineers in most countries. In a number of countries a third group is involved in health and safety at the workplace: the industrial hygienists. Their task is the maintenance of hygienic conditions at the worksite and surveillance of the working environment to identify chemical and physical hazards and give advice on their control and minimization.

In Germany the profession of industrial hygienist has so far not become established. These tasks are taken care off by occupational physicians and safety engineers in close cooperation.

In countries where the occupational health system is not completely separate from general health care, provision of general health care and preventive measures for non-work-related diseases are partially covered by occupational physicians.

1.3. Hazards in the Chemical Industry

Throughout the development of the chemical industry the number of available chemical compounds has grown tremendously. This has led in turn to a broader range of potential health risks caused by chemical compounds. In earlier centuries mostly dusts, fumes, and heavy metals were described as disease-causing agents. Since the mid-1800s organic compounds gathered increasing importance as causative agents for work-related illnesses and diseases.

The specific hazards to human health by chemicals are derived from their general toxicity and from specific properties such as mutagenicity, teratogenicity, carcinogenicity.

The major organ systems affected by chemicals are the eyes (irritation), the skin (chemical burns, allergy), the respiratory system (dust diseases, allergy, irritation), the central and peripheral nervous system (neurotoxicity), the liver, and the kidneys.

The second group of hazardous factors are physical agents such as noise, vibration, heat, dust, radioactive radiation, high pressure, low pressure, and magnetic fields.

The third group of hazardous agents are biological agents, which are gaining in importance with the development of biotechnology and gene technology. Whereas in former times pathogenic bacteria, viruses, and fungi have been known as risk factors in the health- care industry and the pharmaceutical industry during production of vaccines and antibiotics, nowadays microorganisms used in genetic engineering are causing concern as potential health risks.

Developments in the last decades have also attached greater importance to possible health impairments caused by physical conditions such as lifting of heavy loads, monotonous repetitive motions, and improper posture of the body.

Diseases caused by these conditions were already described by Ramazzini. But while in former times humans had to adapt to fit the working conditions – as is demonstrated by the employment of children in certain mining industries – it is nowadays the general philosophy to adjust equipment, machinery, and work-place conditions to the physiological capacity of the human body.

Besides chemical, physical, and biological agents, other factors that have come to be regarded as probable hazards to health in the working population in the last few decades include mental stress and socioeconomic conditions.

The question of occupational disease caused by stress at work is under serious consideration.

1.4. General Outline of Goals

The joint efforts of all parties involved in occupational health and safety at present can be summarized in the following:

1) Prevention of occupational diseases, health impairments, and illnesses caused by influences at work
2) Prevention of work-related accidents
3) Early detection of potential new hazards for human health in the working environment
4) Elimination of identified health hazards and accident sources
5) Adaptation of working conditions, tools, and working environment to the physiological capability of the workers
6) Assistance in providing a socioeconomic work environment to prevent effects on mental health

The means to reach the above-listed goals are:

1) Thorough inspection of the working environment and provision of necessary improvements
2) Provision of a safe technical environment
3) Regular health surveillance and preplacement health examinations
4) Provision of a high standard of hygiene at the worksite
5) Provision of health-supporting socioeconomic conditions

To reach these goals of health and safety provision at the workplace close cooperation of the professional specialists with the employer, employees, and employee representatives is essential.

1.5. Experts and their Training

The tasks of health protection and safety at the workplace are carried out by experts whose professional training and field of expertise varies from country to country.

Occupational Health Physicians. In some countries occupational medicine is a state-licenced speciality of the medical profession. This means that physicians receive additional training in occupational health after several years of medical practice.

In several European countries, training in occupational medicine takes two years, including several months of theoretical study at a university or a special academy. In some countries this additional training leads to certification in occupational medicine, which in Germany includes a state exam.

In the United States a specific training program and residency of two or three years – depending on the university – has been established, leading to the degree of master in public health (MPH). This degree is normally achieved by physicians but can also be acquired by other medical professionals.

In the United States and Great Britain, special training for nurses leads to the degree of occupational health nurse; this is unknown in other countries. The occupational health nurse performs part of the occupational health service in the plant.

Safety Experts. Safety experts normally hold a technical degree, mostly in engineering or chemistry. The special additional education as safety expert or safety engineer differs widely in different countries. Whereas in some countries it is a special academic training with a masters degree or the degree of safety engineer, other countries require only a minor amount of theoretical training parallel to practical experience on the job.

Industrial Hygienists. The profession of industrial hygienist is not yet established in all European countries. It was first developed in the United States and Great Britain, where industrial hygienists hold a specific academic degree in that field. In other countries, such as Germany, industrial hygiene activities are taken care of partially by occupational health physicians or by safety engineers who then are mostly either chemical engineers or chemists by profession.

Worldwide there is wide variability in provision of health and safety services in industry as well as in training and experience of the health and safety specialists.

Many countries are still understaffed and are developing their occupational health and safety systems, with the advise and assistance of more advanced countries. The future goal according to WHO and ILO (International Labour Office) is to provide occupational health and safety services to the entire working population.

2. Legal Background [4–8]

2.1. Laws and Regulations in Germany

The Federal Republic of Germany is a confederation of states. The responsibilities of the government are divided between the Federal Government and the states (Länder) as described below.

Laws, decrees, and guidelines are passed at a national level, the states contributing through the Bundesrat. The states are responsible for the enforcement of laws and for supervising observance of government regulations via their trade supervisory office. In addition to the central government and the states, the legal accident insurance institutions are also involved in the industry-wide occupational safety system. Statutory accident insurance is a branch of social security and is a compulsory insurance. The legal basis of the accident insurance is anchored in the book on social welfare legislation and in the “Reichsversicherungsordnung” (RVO).

The statutory accident insurance institutions are the trade associations. Trade associations are organized according to the specific trade. They have the right to pass accident prevention regulations (UVV).

These regulations serve the trade-specific concretization of governmental laws and decrees.

The relationship between the central government and states on the one hand and the trade associations on the other hand correspond to the principle of subsidiarity. To control the observance of accident prevention regulations, the trade associations have a technical control service (TAD) which assumes control and consultation functions in the factories parallel to the governmental trade supervisory office.

2.1.1. State Legislation and Controls

The enforcement and control of federal laws on occupational safety are the responsibility of the Länder. The State Ministry of Labor (Landesarbeitsministerium) is the highest state authority and is superordinate to the state institute for occupational safety and industrial medicine and the trade supervisory offices. In the application and further development of occupational

safety regulations, close cooperation is required between the participants of the industry-wide occupational safety system and also between the industry-wide and the company occupational safety system.

To guarantee smooth cooperation between the governmental trade supervisory office and the trade associations, the Federal Minister of Labor has passed an appropriate administrative regulation. Apart from the central government, states, and trade associations, private organizations also have important functions in occupational safety.

In Germany, there is a long tradition that private organizations fulfill by means of definitions (e.g., "generally recognized rules of technology") the requirements of federal laws and decrees. These include the standards of the German Institute of Standardization (DIN), guidelines of the Association of German Engineers (VDI), regulations of the Association of German Electrical Engineers (VDE), etc. This is intended to guarantee that occupational safety laws are linked to the further development of technology.

The German Institute of Standardization plays a special role because as far as standardization within the European Community is concerned, it is the direct partner of the European standardization institute CEN/CENELEC.

Private organizations are also active in supervising industrial safety, e.g., the Technical Control Board (TÜV) in the inspection of plant.

The Most Important Laws. The government regulations pertaining to occupational safety can be divided into six areas:

Workplaces, Including Industrial Hygiene. As early as 1891, the first general basic demands on the workplace were laid down in the industrial code. With the workplace decree of 1975, obligatory regulations for the construction and equiping of workplaces were drawn up for all companies.

The individual control areas of the workplace regulations are summarized in the workplace guidelines:

Ventilation
Room temperature
View outside
Artificial lighting
Safety lighting
Floors
Transparent walls
Doors and gates
Glass doors, doors with glass inserts
Protection against falling doors and gates
Automatic doors
Protection against falling objects
Fire extinguishing equipment
Roads
Escalators
Seating
Pause rooms
Rest rooms
Changing rooms
Washrooms
Opportunities for washing outside of designated washrooms
Toilets
Medical room
First-aid equipment
Tempory accommodation for building sites
Washrooms for building sites

Machines, Equipment, and Technical Facilities. Apart from the regulations governing factory premises, a book of rules also exists for the equipping of the premises with machines, instruments, and technical systems.

In particular, the law passed in 1968 governing technical working devices (Equipment Safety Law), the decrees on systems that require monitoring, the Accident Regulations, and the Accident Prevention Regulations should be mentioned here.

The range of the Equipment Safety Law is very wide. It covers not only technical working devices which are used in industrial, manual, and agricultural concerns as well as in the administrative sector, but also machines and equipment used in the private sector, e.g., for the household, leisure activities, recreation, and sports.

The manufacturer and importer are required by this law to put only such equipment into circulation which is safely built. A safety label (GS: tested safety) has been introduced as an orientation aid for the buyer. All machines and instruments carrying this test label have been tested for safety by an officially recognized testing institute.

The most important regulations for facilities that require monitoring are shown in Figure 1.

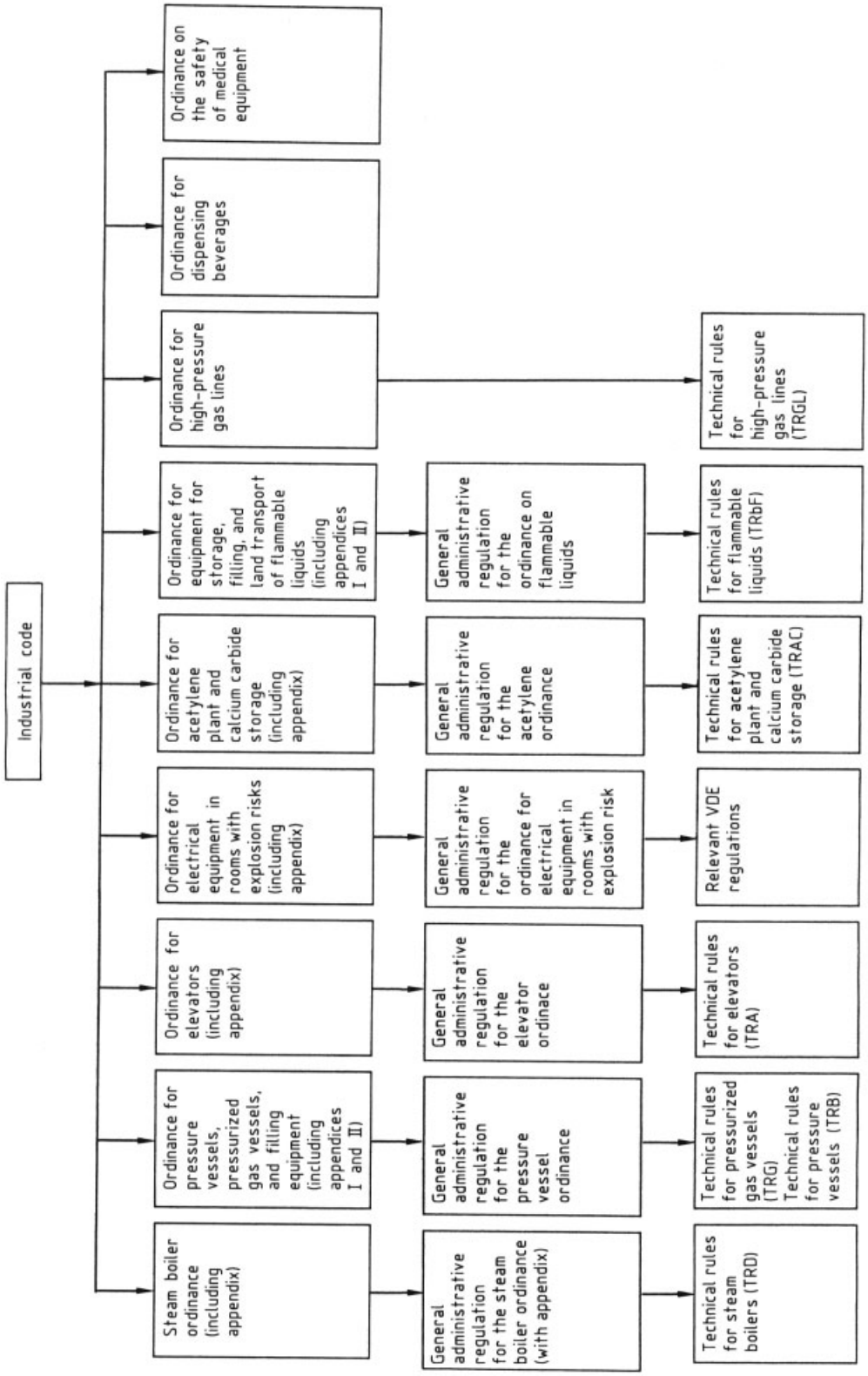

Figure 1. Important regulations for facilities that require monitoring

The emergence of the regulations for facilities that require monitoring can be understood from the historical development. On examination of complex processing plants today, it is clear that these regulations often cover parts of the plant only.

Therefore, occupational safety in processing plants requires supplementary regulations which take the system behavior into account. Based on the federal emissions protection law, the decree pertaining to industrial accidents (12th BImSchV) covers not only environmental protection requirements, but also occupational safety measures in the case of disruption of normal operation.

This is significant because the operating personnel can be directly affected by the onset, control, and possibly also by the elimination of the results of a disturbance.

Accident prevention regulations occupy the fourth part of the area "machines, equipment, and technical facilities". They contain safety requirements for special machines, equipment, and technical facilities which are used in certain branches of industry.

Hazardous Substances. A modern industrial society is inconceivable without the use of chemicals. Today, it is estimated that more than 100 000 chemicals are used worldwide to manufacture products.

Some of these chemicals are classified as dangerous. Therefore, they may be used and put into circulation only after certain safety precautions are taken.

The Chemicals Law (Chemikaliengesetz) and the decree on dangerous substances (Gefahrstoffverordnung) are the most important regulations.

The Chemicals Law covers not only occupational safety, but also environmental protection and general health protection. Nevertheless, occupational safety is the main focus of the law because in most cases the potential hazard of a substance is first observed at the workplace.

The main emphasis of the Chemicals Law is a testing obligation and a compulsory notification for each new substance introduced to the market. Apart from containing special regulations for introduction into the market, labelling, packing, storage, and passing on of information on occupational safety and environmental protection to users of chemicals, the decree on dangerous substances primarily regulates the handling of chemicals within the company.

This decree demands, in particular, a compulsory inquiry (how and which chemicals are handled within the company), compulsory control (checking the maintenance of stipulated air and biological threshold values, etc.), protective measures to be taken, and prohibition or limitation of use.

To further concretize the decree on dangerous substances, the Federal Ministry of Labor has set up a sociopolitical committee which has the task of elaborating the technical regulations for dangerous substances (TRGS). This committee for dangerous substances (AGS) has passed technical regulations for the following areas:

1) Technical and organizational safety measures (e.g., asbestos: demolition, renovation, and maintenance work)
2) Technical guide concentrations (threshold values for carcinogenic substances)
3) Substitutes and substitute processes (e.g., substitutes for cleansing agents containing chlorinated hydrocarbons).

Regulation of Working Hours. Regulations pertaining to working hours are contained in the working hours code, trading and industrial code, the law regulating the closing time of shops, and the special regulations for the working hours of professional drivers.

The regular daily working hours must not exceed 10 h. Working on Sundays and holidays is basically forbidden. However, there are a series of exceptions:

1) Work that must be done without delay in emergencies or in the public interest
2) Work that is required to prevent the spoiling of raw materials or the failure of production
3) Work that is required to prevent excessive damage when an unforeseeable necessity arises

Protection of Certain Groups of People. The law protecting mothers protects the working mother and the fetus from dangers and health impairments. It protects the mother from excessive demands at the workplace, financial losses, and loss of her job due to pregnancy and delivery.

The law for the protection of children and young persons applies to the employment of per-

sons under 18 years of age. The minimum working age has been fixed at 15 years.

Apart from the laws protecting mothers and young persons, there are also regulations for the disabled and for working at home.

Occupational Safety Organization in the Company. The employer is responsible for the enforcement of occupational safety in the company. He is obliged by law to take all measures required for the protection of his employees in all the occupational safety areas mentioned above, including occupational safety organization in the company. Here, the Workers' Protection Law, the rules for the prevention of accidents, and the labor management act are of special importance.

The Workers' Protection Law requires that the companies engage doctors and qualified personnel for occupational safety. The qualified personnel and the company doctors have to support the employer with regard to occupational safety and accident prevention in all questions of health protection and working safety, including the designing of humane working conditions.

The qualification requirements, profile of duties, and working hours are further concretized in the rules for the prevention of accidents, "Safety Engineers and other Specialists for Occupational Safety" (VBG 122) and "Company Doctors" (VBG 123).

Responsibilities of the Federal Government and the Länder. At the federal level, the Federal Ministry of Labor is in overall charge of occupational safety. In this ministry, drafts of new laws and decrees are elaborated and EC guidelines are translated into German law.

Furthermore, via lower authorities, this ministry supports research and model projects and provides counseling and information.

Supplementary focal points are taken on by other ministries, e.g., the program "Work and Technology" at the Federal Ministry of Research and Technology.

Institutions and Control Functions. The Federal Ministry of Labor is supported by the Federal Institution of Industrial Safety and the Federal Institution of Industrial Medicine.

The duties of the Federal Institution of Industrial Safety include:

1) Cooperation with state officials and with statutory accident insurance institutions.
2) Observation and analysis of occupational safety, health situation, and working conditions in companies and offices.
3) Development of solutions to problems by the application of knowledge of technical safety, industrial medicine, ergonomics, and other industrial sciences. For this purpose, the Federal Institution of Industrial Safety conducts research or awards research grants.
4) Promotion of the application of acquired knowledge through publications, assistance with the defining of rules, development of materials for training and further education, model-like implementation and counseling, exhibitions and professional meetings.

The Federal Institution of Industrial Medicine also belongs to the Federal Ministry of Labor. It emerged from the Central Institute of Industrial Medicine of the former German Democratic Republic and has the following duties:

1) To observe and evaluate the effects of working conditions on health and make contributions to prevention.
2) To conduct and support research on subjects in industrial medicine. The main emphasis is placed on biotechnology and genetic engineering, allergies, stress, hazardous substances, noise, vibration, radiation, dust, and carcinogenic, mutagenic, and teratogenic substances.
3) To collect and disseminate knowledge of industrial medicine as well as to draw up concepts for the care given by company doctors and for health promotion in the company.

The trade supervisory authority that belongs to each state ministry is responsible for technical occupational safety.

This trade supervisory authority is equipped with all the required official powers, especially the right to inspect and check plants at any time. The trade supervisory authority checks that laws, decrees, and other regulations are complied with. Legal occupational safety covers the following:

1) Machines, instruments, plants
2) Hazardous substances, radiation
3) Working hours
4) Social occupational safety (protection of special groups of people)

5) Accident prevention and safety technology

Each trade supervisory office is responsible for all the companies in its area – independent of the branch of industry – and supervises not only the observation of government safety regulations, but also environmental protection. It is also responsible for the granting of exemptions.

The trade medical service also comes under the state ministries and is responsible for medical occupational safety. It is concerned with preventive health protection, such as evaluation of the workplace and medical checkups.

With the help of medical inquiries, this service should elaborate information on possible health hazards caused by the type of work, substances, etc. In addition, it is involved in proceedings for the acknowledgement of occupational diseases.

2.1.2. Trade Associations

Origin, Structure, and Organization. The fundamental activities of trade associations are specified by the Accident Insurance Law of 1884. These are:

1) The private liability of the employer for his employees is replaced by the public law system of statutory accident insurance.
2) Employers in the same or related branches of industry join together to form trade associations. The demands of accident victims are henceforth directed towards the appropriate trade association.
3) The benefits provided by law are granted independently of who caused the accident. The injured person is entitled to this benefit even if he caused the accident himself by thoughtlessness, negligence, or unlawful action. He is denied the benefit only if he intentionally caused the accident.
4) The trade associations are obliged to pursue accident prevention by all suitable means. The employer, who remains fully responsible for occupational safety in his company, receives professional advice and help from his trade association.

The common basis for the social, health, and accident insurance was created by the " Reichsversicherungsordnung" (RVO) of 1911.

At present, the legal accident insurances are divided into one marine, 21 agricultural, and 34 industrial trade associations.

The statutory accident insurance institutions are the federal government, the states, the commune and insurance associations, certain municipalities (cities with their own accident insurance), and the fire department accident fund. The accident insurance institutions are public corporations. They have the right of self-administration, i.e., they carry out the duties assigned to them by law on the responsibility of their own honorary autonomous organs, but under state supervision.

The autonomous organs are the assembly of representatives and the board of directors, elected every six years. They are made up of equal numbers of representatives of the insured persons and of the employers. In the case of agricultural trade associations, representatives of the insured persons and of the employers and of the self-employed make up one-third each.

The assembly of representatives establishes the rules and other autonomous rights of the insurance institutions (e.g., accident prevention regulations). The expenditure of the trade associations is covered by the companies. The employees make no contribution.

This regulation of the fees is based on the fact that the statutory accident insurance is primarily meant to replace the civil liability of the employer for his employees with regard to occupational diseases and to accidents at the workplace and on the road to work.

Further duties of the trade associations include prevention and rehabilitation.

Accident Prevention Regulations. The regulations for the prevention of accidents are the rules of the particular statutory insurer passed according to the RVO and confirmed by the Federal Ministry of Labor.

The employers united in the trade associations and their employees are obliged to observe the accident prevention regulations.

At present, more than 150 regulations for the prevention of accidents exist with different contents. These regulations must be adapted by the individual trade associations to each sphere of responsibility and then passed. Thus, more than 1800 individual accident prevention regulations exist. These regulations control facili-

ties, orders and measures taken for the prevention of accidents, rules of action for employees, requirements for medical examinations, and demands made on the company safety organization. They are supplemented by numerous guidelines, safety rules, leaflets, and brochures issued by the trade associations to facilitate enforcement within the company.

The trade associations have their own umbrella organizations for the areas of industry and agriculture. In the civil service, there is also a head organization: the statutory accident insurance institution.

The main organization of the industrial trade associations looks after common interests and primary duties of the trade associations. These include:

1) Occupational safety, industrial medicine, and first aid
2) Medical and professional rehabilitation
3) Research, testing, and counseling
4) Training and further education of the members of the trade associations
5) Provision of information material
6) Implementation of interstate social insurance agreements

The main organization looks after the common duties of the trade associations in the region through the individual state organizations.

Duties and Control Functions. Each trade association has its own technical control service which supervises the observation of accident prevention regulations. In addition, it counsels companies with regard to industrial safety and health protection. For instance, it can give valuable information on occupational safety prior to the procurement of new machines or the use of new processes. Moreover, the members of the technical control service also work on the further development of the book of technical safety rules in Germany and in the EC. In addition, they advise manufacturers on the safety and ergonomic development of working facilities through the expert committees of the trade associations. The trade associations are also active in the area of occupational medical counseling. Companies that do not have their own doctor can fall back on the doctors of the trade associations' medical service.

In addition, qualified safety personnel, safety representatives, employers and executives, works and staff council members, and company workers are trained in occupational safety and health protection by the trade associations (see Section 2.1.3).

2.1.3. Company Occupational Safety System

The employer is responsible for ensuring that occupational safety regulations are observed. His facilities, measures, and instructions must provide safe and healthy working conditions in the company.

Management must ensure that, as part of their authority to issue directives, all executives and senior employees must regard the practice of occupational safety and health protection as one of their most important executive duties.

In spite of this transfer of duties, the employer alone takes the entire responsibility.

The Specialists. Legal regulations (e.g., workers' protection law) and the rules of the trade associations require companies to provide doctors who are specialists in industrial medicine. The requirements are stipulated in the accident prevention regulation: "Company Doctors" VBG 123 (regulations of the main organization of the trade associations). Company doctors counsel the employer on all matters relating to occupational and health protection:

1) Planning, implementation, and maintenance of production facilities and of social and sanitary facilities
2) Procurement of technical working devices and the introduction of working methods and substances
3) Selection and testing of protective equipment
4) Occupational physiological, occupational psychological and other ergonomic and occupational hygienic questions, which include working rhythm, working hours and pauses, design of workplace, course of work, and working surroundings

Their duties also comprise organization of first aid in the company and questions of changing the workplace and integration of the handicapped into the working process.

Company doctors also perform medical checkups and provide medical care for employees.

The working hours of company doctors are regulated by the trade associations and the public accident insurance institutions in their accident prevention regulations. They depend on the potential hazard of the particular branch and the number of employees.

Company doctors can be employees of the company in question, employees of an industry-wide occupational medical service, or self-employed. The type of medical care is stipulated in agreement with the employer and the works council.

The employer is also required by the workers' protection law to hire qualified personnel for occupational safety. These are engineers, master craftsmen, or technicians with additional training in occupational safety. Apart from counseling duties, similar to those of the company doctors (see above), they are responsible for the technical safety supervision of production facilities and working devices and processes, and for observing occupational safety and accident prevention in practice. Like the doctors, they must inspect the workplaces at regular intervals, report shortcomings to the employer, and recommend measures for their elimination.

Similar to the doctors, the qualified safety personnel has no authority to issue directives. In their work, however, they are free of directives, i.e., not subject to the opinions of company executives when they make demands or have complaints in the interest of occupational safety.

The working hours of the qualified safety personnel are stipulated in the Accident Prevention Regulation: "Safety Engineers and other Qualified Personnel for Occupational Safety" (VBG 122).

Like the company doctors, the qualified safety personnel can be employees of the company in question, employees of an industry-wide technical safety service, or self-employed.

The work of the qualified safety personnel is supplemented by the safety representatives in the company.

In companies with more than 20 employees, safety representatives are appointed by the employer in accordance with the RVO. In their field of activity, they have to pay attention to safety at the workplace. The number of safety representatives depends on the branch, number of employees, and hazard potential of the particular company.

Apart from company doctors, qualified personnel for occupational safety, and safety representatives, other occupational safety experts can also work in the company, e.g., representatives for radiation protection and representatives for biological safety. The experts within the company also include recognized specialists who are authorized to check plants that require monitoring (§ 24 c of the industrial code).

The company representatives for environmental protection (waste, wastewater, and transport of hazardous materials) are also to be taken into account because their duties can partially overlap with occupational safety.

Employees have to support all measures that serve occupational safety. They are obliged to comply with directions of the employer for the purpose of occupational safety. Furthermore, they have to use the personal protective equipment placed at their disposal. Employees must use company equipment only for purposes that are in accordance with the regulations and they need not follow directions that are adverse to safety.

Before taking up employment and with each assignment to a new workplace, the employee is to be informed about the accident and health hazards involved in his work and about the measures and equipment available to avert these risks. These safety instructions can be delivered verbally or in writing. However, the mere handing over of leaflets – without additional explanations – is not sufficient.

If an employee discovers that some equipment is not perfectly safe, he must correct the defect immediately. If this does not belong to his duties or if he does not have the necessary knowledge, he should report the defect to his superior immediately.

If an employee makes use of his right to lodge a complaint about inadequate safety measures with his employer, no disadvantages should accrue to him therefrom. After checking the complaint, the employer must inform the employee about the results and, if necessary, immediately rectify unsafe conditions.

The employee can also lodge a complaint with the works council, either immediately or

after he has unsuccessfully complained to the employer. The works council must check the complaint and, in case it is justified, urge the employer to take remedial action.

Duties of the Works Council. To combat accident and health hazards in the company, all those who can contribute are required to work together. In this connection, the employees have a special role. The German Industrial Relations Act regulates the right of information, participation, and codetermination of employees. The workforce elects and is represented by the works council. The main concern of the Industrial Relations Act is to involve the works council (and the workforce) in all occupational and health protection decisions taken by the employer. With respect to occupational safety, the works council has the mandate to control and organize, the right of codetermination, the obligation to support the authorities, and the right of information and participation.

The works council should ensure that not only the laws, decrees, and accident prevention regulations passed in favor of the employee, but also the negotiated industrial agreements and employment agreements (see Section 2.1.4) are observed and enforced.

The mandate to monitor applies to the occupational safety measures to be taken by the employer and company executives as well as to those to be taken by employees. The works council has the right and the duty to point out shortcomings in occupational safety and accident prevention to the employer or responsible company executives and to press for their elimination with all means at its disposal. The supervisory mandate also includes the obligation to urge the employees to comply with the regulations passed for their protection during the time they spend in the company.

The supervisory mandate is supplemented by the organizational mandate. The works council is required to propose measures that serve the company and the workforce. It can deal with all matters that serve the protection from accidents and health hazards. It can demand from the employer that appropriate measures are taken to eliminate hazards.

According to the Labor Management Act, the works council has the right of codetermination in the settlement of matters of occupational safety and accident prevention. This means that the consent of the works council to the safety measure proposed by the employer is required. It is not sufficient for the works council to simply take note of the intended occupational safety measure. In accordance with the Workers' Protection Law, the appointment of qualified safety personnel and company doctors is also subject to this codetermination.

Based on their obligation to support the industrial safety authorities (trade supervisory authority and trade association), the works council has the duty to support the authorities in their work in the company by providing suggestions, advice, and information.

On the other hand, the authorities are obliged to involve the works council in all company inspections, accident investigations, and in meetings on subjects of occupational safety and to hold discussions, if necessary. Apart from the control services, the works council can also take advice from the qualified safety personnel and from the company doctors. In agreement with the employer, the works council can also consult other experts outside the company. The works council can do justice to its mandate to supervise, organize, codetermine, and support only if it is adequately informed. For this reason, according to the Labor Management Act, the works council is granted an extensive right of information. This right of information is based on the principle that the employer and the works council should have available the same level of information on occupational safety and accident prevention in order to fulfill their duties.

Although the works council has the right to control and participate, the employer carries the sole responsibility for occupational safety. The works council cannot independently enforce safety measures.

The Occupational Safety Committee. All experts for occupational safety in the company are obliged to cooperate. Therefore, the employer is responsible for forming a suitable occupational safety organization within the company.

The central body of the company occupational safety organization is the occupational safety committee. Occupational safety committees are to be formed in all companies in which company doctors and qualified safety personnel

are appointed. The following are members of the committee: the employer or his representative, two members selected by the works council or staff council, company doctors, qualified safety personnel, and safety representatives.

The occupational safety committee is required to discuss matters of concern for occupational safety. It meets at least once every three months. Apart from routine duties, such as dealing with accident statistics or the general supervisory function and coordination, the occupational safety committee should take into account the conditions in the particular company and concentrate its efforts on, e.g.,

1) Counseling and recommendations for safety programs in the company
2) Discussing suggestions for putting into effect safety investments of the company
3) Regular evaluation of the frequency and severity of occupational diseases
4) Analysis of the results of technical safety controls of work processes and the introduction of new working methods or new substances
5) Discussions of proposals for company participation in industry-wide accident prevention measures

In companies with more than three safety representatives and an existing occupational safety committee, it is also required that the safety representatives meet regularly.

In these meetings, the technical safety problems of the individual areas are to be discussed. The safety representatives for the occupational safety committee are to be named.

In smaller companies which have no qualified safety personnel and company doctors, a safety committee is to be formed if more than three safety representatives are appointed. With the participation of the works council, the employer should meet at least once a month with the safety representatives or the safety committee.

2.1.4. Role of the Social Partner and Associations

The function of the unions and the employers' associations are not regulated by law. However, some of their powers are legally described, especially the right to negotiate industrial agreements and to send representatives to certain occupational safety bodies.

Occupational safety and health protection of workers have always been a goal of the unions. Progress has been made in some areas, e.g., a reduction of the number of accidents at the workplace or on the road to work and an increase in the recognition of industrial diseases. However, new problems arise at many workplaces due to job reorganization (e.g., due to stress, performance demands, and new substances).

To attain their occupational safety and health protection goals, the unions work on the further development of the books of rules for occupational safety in both parliamentary and extraparliamentary bodies (occupational safety bodies). Their efforts are concentrated in the sociopolitical bodies of the Federal Ministry of Labor (e.g., committee for hazardous substances), in the advisory councils of the Federal Institution of Industrial Safety and the Federal Institution of Industrial Medicine, and in the aid program "Work and Technology" of the Federal Ministry for Education, Science, Research, and Technology. Moreover, occupational safety councils have been founded in some states with the participation of the unions. However, the most extensive field of action of the unions is the self-government of the trade associations, which comprise equal numbers of representatives of the employer and of the employees. In this connection, unions can be represented by working members or by full-time officials.

Via the autonomous organs (assembly of representatives and the board) and their committees (e.g., accident prevention committee, pension committee), the unions can directly influence the book of rules of the trade associations, help to plan the essential features of the activity of the technical control service (information, counseling, control), and assist in the assessment of occupational diseases and accidents at the workplace or on the road to work.

The representatives of the employer who work in the autonomous bodies of the trade associations can also be the representatives of employers' associations. The resulting common planning platform in occupational safety of trade associations is supplemented by agreements or industrial contracts between the social partners.

These contracts can contain regulations which go beyond the book of rules of the govern-

ment or the trade associations or concretize these rules in a branch- or company-specific manner. For the employees, industrial contracts of this type can be completed only be the trade associations. For the employers, however, both the employers' associations (joint agreements) as well as each individual employer (company or house agreements) may negotiate. The workman's guild and the guild associations are also entitled to complete contracts by virtue of special legal regulations. Where occupational safety was a part of joint industrial agreements in the past decades, it usually applied to the classical areas of working hours, protection of special groups of people, and special allowances.

On the other hand, company agreements contain detailed regulations regarding, e.g., the participation of employees and members of the works council in the enforcement of the Workers' Protection Law and the decree on dangerous substances, or the role of the safety representative.

2.1.5. Future Developments

As a result of the formation of the European Common Market in 1 January 1993, Germany is required to translate European guidelines into German law. Of special significance is the guideline 89/391/EEC: "EEC guideline for the enforcement of measures to improve the safety and health protection of the employee at work". At present, it is being translated into a basic German occupational safety law. The political consultations are not yet completed.

The important points of this EC guideline or basic occupational safety law will be:

1) Extension of the understanding of occupational safety including health protection in the company and, thus, improvement of prevention
2) All workplaces and work processes are to be subjected to risk assessment
3) Employees are to be involved in the shaping of working conditions
4) The control, counseling, and compensation mandate of the trade associations will be extended to include prevention of all health hazards due to work

Due to the "centralization" of the occupational safety legislation in the EC, adaptation of the accident prevention regulations of the trade associations will become necessary. National standards (DIN, VDI, VDE) for machines, equipment, and facilities will be replaced gradually by European standards by CEN/CENELEC.

In Germany, the statutory health insurance companies will also play a new role in future occupational safety and health protection in the company. As a branch of social security, the health insurance companies have been assigned duties in health protection at the workplace in accordance with Section 4, § 20 of the social code. These duties are:

1) To counsel and inform insured persons about health hazards in general and about the prevention of diseases in particular
2) To pursue the causes of health hazards and to work on their elimination
3) To assist in the prevention of health hazards caused by work and to cooperate with the statutory accident insurance institutions (the trade associations)

2.2. Legal Requirements in Other European Countries

Most European countries have specific regulations to protect workers from harm due to work. For some of these countries the regulations are described in more detail below (situation as of December 1994).

United Kingdom. The principal legislative framework is the Health and Safety at Work Act 1974. This sets out general duties of employers and others and is supplemented by specific regulations with more detailed requirements. In the case of chemicals the principal regulations are:

1) Notification of New Substances Regulations 1993 (NONS)
2) Chemicals (Hazard Information and Packaging) Regulations 1993 (CHIP)
3) Control of Substances Hazardous to Health Regulations 1994 (COSHH)

Strictly speaking, NONS and CHIP, both of which are derived from EC directives, deal with the supply of hazardous substances. COSHH, by

contrast, deals with the use of hazardous substances, and with assessing and managing the health risk to employees. If the assessment indicates that there may be risks to health, COSHH requires the employer to:

1) Prevent (by substitution or elimination) or adequately control exposure
2) Make sure that precautions are used and controls maintained
3) Monitor, if necessary, exposure and/or carry out health surveillance
4) Inform, instruct, and train employees as necessary

COSHH contains specific provisions to deal with carcinogens; these implement the EC Carcinogens Directive.

Other regulations deal with more specific risks such as from lead, asbestos, and flammable liquids.

A list of occupational exposure limits (OELs) for use with COSHH is published annually. The list contains two types of OEL. Maximum exposure limits (MEL) are maximum concentrations of airborne substances, averaged over a reference period, to which employees may be exposed by inhalation under any circumstances. The occupational exposure standard (OES) is the concentration of an airborne substance, averaged over a reference period, at which, according to current knowledge, there is no evidence that it is likely to be injurious to employees if they are exposed daily by inhalation to that concentration.

Denmark. The general regulation of chemicals at the workplace is laid down in Order No. 540 of Substances and Materials, which came into force in 1983. (In practice "materials" are considered to be the same as "preparations").

The order establishes some very general provisions for working with chemicals. More specifically it also defines the substances and preparations that are to be considered as hazardous. This Danish definition for workplace purposes is broader than the EC definition, laid down in the 1967 Substances Directive and in the 1988 Preparation Directive.

Included in the definition of "hazardous" are:

1) All substances and preparations classified as dangerous according to the EC directives
2) All substances on the Danish list of OELs
3) All substances on a recommended list (an annex to the list of the OELs) of volatile organic solvents and preparations containing more than 0.5 % of such a substance
4) All substances on a list of agents considered to be carcinogenic and preparations containing more than 0.1 % of such a substance; the list consists of EC carcinogens and IARC carcinogens (Groups 1, 2 A, and 2 B)

The extension compared to the EC definition is primarily for organic solvents. For carcinogens the difference to the EC definition is only small but was substantial until 5 years ago.

For the substances and preparations classified as hazardous specific provisions are given:

1) Substitution/replacement by less dangerous substances if technically feasible
2) Safety data sheets prepared by the supplier, including obligatory information on specific Danish restrictions for use
3) Safety data sheets prepared by the employer (more detailed than those from the supplier)

In more detailed governmental orders, specific rules are laid down for minimization of exposure and for specific types of preparations such as paints.

A specific order prohibits employers ordering persons under 18 years of age to work with preparations containing more than 0.5 % of a volatile solvent.

In general, legislation is similar to the latest EC directives.

Before use of a new substance or new preparation at the workplace the supplier must notify the Product Register at the Danish Working Environment Service. Among other information the notification must contain a full quantitative and qualitative description of the substance/preparation and the amount expected to be sold per annum in Denmark.

Finland. Finnish occupational chemical safety regulations are based on two framework acts: the Labor Protection Act (299/58, amendments) and the Chemicals Act (744/89, amendments). The Act on the Supervision of Labor Protection and Appeal Procedure in matters concerning Labor Protection (131/73, amendments) lays down the rights and powers of the

labor protection authorities. In addition to these, the Occupational Health Care Act (743/78) regulates the occupational health care services which are to be provided to workers.

The Chemicals Act prescribes the requirements concerning chemical substances and preparations as products. The purpose of the Act is to prevent damage to health and environment. Being a framework act, it contains quite general provisions on the duties of manufacturers, importers, and suppliers of chemicals.

With some minor exceptions the regulations concerning chemical substances and preparations as well as safety data sheets are identical with the corresponding EU directives. Under Article 12 of the Preparations Directive, Finnish manufacturers and importers of chemicals hazardous to health are obliged to send fairly detailed information to the Ministry of Labor, which maintains a registry of chemical products. The EU directives regarding restrictions on the marketing and use of certain dangerous substances and preparations have also been implemented in Finnish legislation.

Protection of workers against the risks related to chemical agents is covered by the Labor Protection Act. It also contains general provisions on protection of workers against other hazardous factors at the workplace. More detailed regulations are given in decisions of the Council of State and Ministry of Labor. Detailed regulations have been issued, e.g., on benzene (355/82), lead (1154/93), asbestos (886/87, 1413/91), carcinogenic substances (1182/92, 838/93), and teratogenic substances (1043/91, 1044/91).

The most important regulations concerning exposure to chemicals, however, are contained in the Council of State Decision on the Protection of Workers from the risks related to exposure to chemical agents (920/92, 727/93). These regulations cover widely the obligations of the employer, from assessment of exposure and risk to preventive and protective measures, as well as instructions and information to be given to the employees. The limit values for workplace air contaminants are also specified. Basically, there are two types of limit values. The Council of State issues "binding limit values for workplace air contaminants" which are directly legally binding (e.g., the values for asbestos, benzene, lead, and vinyl chloride). The other type of limit values are the "indicative limit values of workplace air contaminants" (HTP values). The philosophy of the indicative limit values is very close to the concept of the EU indicative limit values. Employers have to take the indicative limit values into account when assessing the quality of workplace environment, exposure of workers, and the meaning of the measurement results.

Spain. Spanish law concerning health and safety at work derives from a number of different sources with varying levels of importance. The Constitution of 1978 requires the public authorities to safeguard workers' safety and health. International obligations such as ILO conventions are included in national law. The Spanish Parliament has adopted a number of specific acts dealing with health and safety, in particular the Workers' Statute of 1980, which states that workers have a general right to be protected by safety rules and the responsibility to perceive the rules adopted by the employer. There are also elaborate administrative regulations establishing the Labor Inspectorate and laying down detailed and technical health and safety rules.

The general principles as laid down in the law are worked out in regulations. The most important of the regulations concerning health and safety at work is the general Ordinance Concerning Safety and Health at Work of 1971, being a tool for the Ministry of Labor to make rules, to initiate research, to give technical guidance, to lay down the functions of the Labor Inspectorate, to establish provincial occupational safety and health councils, to lay general duties on both employers and workers, to set out the functions of occupational safety and health committees and of safety officers, and to make provision for administrative fines.

At the same time there are a number of other regulations which set out specific technical standards, e.g., on protection against lead, asbestos and ionising radiations.

New legislation is designed to bring Spain into line with EC legislation, especially the Framework Directive. Until now provisions have been adopted to comply with the directives on vinyl chloride, lead, asbestos, and banning of certain specified agents and activities. This new law will probably be adopted in 1995.

France. The French laws concerning the risks of chemicals is mainly based on EC directives. For the workplace this means:

1) Risk classification of substances and preparations
2) Notification of new chemicals and preparations
3) Material safety data sheets
4) General prevention rules for workplaces
5) Quality control procedures for workplace air
6) Prevention of cancer risks

There are specific rules and protection measures for women and younger persons. Basis for taking prevention measures is the risk assessment of dangerous substances at the workplace.

Occupational exposure limits (not binding) give indications for protection of workers. At the moment about 520 indicative limit values are available.

The Netherlands. The Working Environment Act (*Arbeidsomstandighedenwet*), adopted in 1980, deals with safety, health, and well-being during work and is meant as a framework law. It comprises basic requirements and responsibilities of employers and employees to be elaborated in decrees and regulations for special topics.

With respect to chemical substances the following regulations are in force:

Relevant decrees are the *Decree on Industry and Workshops* (1938) and the *Decree on Commercial and Noncommercial Services* (1990). For substances a central point is that effective measures must be taken to prevent harm, damage, or nuisance to workers exposed to chemicals during their work. The hierarchy of preventive measures to be taken reads:

1) Prevention of all emissions at source
2) Ventilation
3) Separation of workers and source
4) Use of personal protective equipment

The measures to be taken have to be effective with respect to health protection and prevention of nuisance, but feasibility aspects may also be taken into account. Only in the ultimate case may personal protective equipment be used.

The Minister of Social Affairs has the right to set legal limit values for exposure during work. In March 1994 an initial list of 96 substances with legally binding limit values was published. All these chemicals have passed the Dutch three-step procedure for setting limit values. Additionally, legally binding limit values for 15 carcinogenic chemicals have been covered by a specific decree on carcinogenic substances. In addition to the binding values, administrative occupational exposure limits (MAC values) are in force, adopted from the United States (ACGIH), Germany, Sweden, and the United Kingdom.

The *Decree on Specific Harmful Chemical Compounds* (1991) prohibits the production, storage, and use of certain aromatic amines (2-naphthylamine, 4-aminodiphenylamine, benzidine, and their salts, 4-nitrodiphenylamine). Also, preparations containing more than 0.1 wt % fall under this regulation. It is an implementation of EC Directive 88/642.

The *Decree on Carcinogenic Substances and Processes* (1994) is an implementation of EC Directive 90/394 on the protection of workers from harmful effects due to exposure to carcinogenic chemicals. It covers all chemicals that fall within the criteria for carcinogenicity in EC Directives 67/548 and 88/379.

The decree covers:

1) Registration and evaluation of carcinogenic substances
2) Restrictive working conditions to minimize exposure
3) Setting legal limit values for carcinogenic substances (so far 15 have been published)
4) Medical surveillance, including environmental and biological monitoring and medical investigations
5) Keeping personal log-books of workers involved, which are to be kept for at least 40 years

Moreover, a number of acts have since been published, all of which are appended to the Working Conditions Law. The *Propane Sultone Act* (1976) prohibits the use, production, and storage of this suspect carcinogen.

The *Asbestos Act* (1977) is an implementation of EC Directive 83/477. This Act has been amended several times; the 1993 version prohibits the use and production of all types of asbestos. In situations where significant exposure occurs (e.g., demolition) strict rules and standards apply.

The *Benzene Act* (1977) is an implementation of ILO Treaty 136 of 1971 and restricts the use of this solvent. The *Vinyl Chloride Act* (1987) implements EC Directive 78/610. The *Lead Act* (1988) includes biological limit values for exposure.

2.3. European Legislation

Due to the increasing mobility of the labor force throughout Europe, it is becoming more and more desirable to harmonize regulations and rules concerning working conditions.

As in other matters, the EU Member States try to come to a joint standpoint by mutual discussions. Directives related to working conditions are drafted by the Directorate General V (DG V, Health and Safety) in Luxemburg. These drafts are elaborated in the relevant EC Council working group, leading to a common position. After adoption by the Council of Ministers the new EC legislation must be implemented in national legislation within a certain time period. Since EC workplace legislation concerns minimum requirements (Art. 118 A of the EC Treaty), each Member State has the right to set stricter regulation than specified in EC directives.

Since the 1970s a number of directives on working conditions have been adopted. With respect to chemical substances the following directives, dealing with the protection of workers against health risks of exposure to chemical agents during work, are noteworthy:

Directive 78/610 – vinyl chloride
Directive 80/1107 – a framework directive for chemical, physical and biological agents with general principles of protective measures
Directive 82/605 – exposure to metallic lead and its compounds
Directive 83/477, amended by 91/382 – handling of asbestos
Directive 88/642 – binding and indicative limit values
Directive 88/364 – relating to the banning of certain agents
Directive 90/394 – the carcinogens directive
Directive 91/322 – establishing indicative limit values for 27 substances

Since 91/322 a multistep procedure has been followed to develop additional indicative limit values. As a first step the Scientific Expert Group (SEG) of DG V establishes a health-based recommended value; in the next step the socioeconomic and technical feasibility of the SEG value is investigated by the Commission, the Member States, and the tripartite Advisory Committee on Safety, Hygiene, and Health Protection at Work. In the third and last step the committee of Member State representatives votes on the Commission proposal for a definite indicative limit value.

As of December 1994 a number of SEG recommendations has been published. The next steps in the procedure, however, have stagnated seriously.

A Commission proposal [COM (93) 155 final – SYN 459] for a council directive on the protection of the health and safety of workers from the risks related to chemical agents at work is under discussion in the Council working group.

100 A-Directives. In addition to the specific 118 A-workplace directives, a number of 100 A Common Market Directives also have important implications for workers.

Most prominent is Council Directive 67/548/EEC of 27 June 1967 relating to the classification, packaging and labelling of dangerous substances, as amended by Council Directive 92/32/EEC of 30 April 1992. The sixth amendment (Directive 79/831/EEC) concerns notification of new substances. In the seventh amendment (1993) the risk assessment paragraph has been worked out extensively.

Further Council Directive of 26 June 1978 relating to the classification, packaging, and labelling of dangerous preparations (pesticides), and Council Directive 88/379/EEC of 7 June 1988 relating to the classification, packaging, and labelling of dangerous preparations, as amended by Council Directive 90/492/EEC, defining and laying down a system of specific information on dangerous substances and preparations, in the form of safety data sheets, are principally intended to enable industrial users to take the measures necessary to ensure the protection of the safety and health of workers.

2.4. Regulations in the United States

Industrial safety and health in the early stages in America changed from work-related accident

and illness being the worker's responsibility to a team approach of worker and management. Some industrialists realized early that unhealthy or unsafe work conditions cost money, with lowered work quality, interrupted work schedules, and compensation payments to injured or sick workers. Others refused to recognize the inherent hazards of their business, and this required a series of regulatory requirements for adequate controls to be introduced.

The Occupational Safety and Health Act (OSHA) of 1970 authorized the Department of Labor to set and enforce mandatory occupational and health standards for businesses. The act provided for research and assistance with the discovery of latent data to establish any connection between diseases and the work environment. While other laws gave the authority to regulate food additives, drugs, and pesticides for public exposure, no employer was previously required to assure worker health in production operations. OSHA is unique because it does not have authority to pre-approve products entering the market, unlike FDA and EPA. Furthermore, OSHA cannot require mandatory testing of products.

The Act created the National Institute of Occupational Safety and Health (NIOSH) under the Department of Health and Human Services, with responsibility for research and recommendations for reports on employee exposure and the authorization of medical examinations and tests where desirable to protect employees.

OSHA health standards typically set maximum limits on employee exposure and prescribe procedures or equipment to achieve this level. The Environmental Protection Agency (EPA) under the Toxic Substances Control Act (TSCA) requires disclosure of safety information and testing to develop data for health and environmental effects for chemicals intended for manufacture, import, or distribution. Testing is initiated by the finding that a chemical presents a potential risk or that humans will be exposed to substantial quantities. With the different regulatory agencies involved (federal and state), there is potential for inconsistency in government standards.

The Interagency Testing Committee, now the Regulatory Council, had the goal of securing agreement on requiring chemical-specific testing and common policies. Agencies have also achieved collaboration in the National Toxicology Program (NTP) to coordinate numerous federal efforts to improve methods and performance of animal studies on selected chemicals, funded by public taxation.

Another joint effort is the Agency for Toxic Substances and Disease Registry (ATSDR) under the Public Health Service, where the EPA publishes toxicological profiles of chemicals determined by the collaborators to be the most significant potential threat to human health.

In summary, the Federal Government's Occupational and Health Administration, in cooperation with state agencies, sets and enforces standards for production and process workers. Other agencies (EPA, FDA, and Consumer Product Safety Commission) have various laws on packaging, transportation labels, and pre-marketing requirements to protect consumer health and the environment. To combine efforts, OSHA has published a standard on Hazard Communications (1986), which includes requirements for material safety data sheets and labels for hazardous products.

3. Industrial Hygiene [9–28]

3.1. Introduction and History

Introduction. Industrial Hygiene (IH) is a term given in the United States primarily to the discipline broadly concerning the recognition, evaluation, and control of the chemical, physical, and biological stressors that may cause discomfort to or impair the health or well-being of employees in the workplace. Such stressors may result from exposure to:

1) Chemical gases, vapors, fumes, particulate matters, etc.
2) Physical agents such as noise, vibration, heat, and cold
3) Ionizing and non-ionizing radiation
4) Biological materials or microorganisms
5) Ergonomic stress

In a number of European countries this discipline is defined more appropriately as Occupational Hygiene (OH), and this term, originating from the United Kingdom, is generally used in these Sections. In other countries (e.g., Germany) the term industrial hygienist or occupa-

tional hygienist is less well-known, and the general functions, defined above, are generally performed by specialized safety or medical personnel.

The main objective of occupational hygiene is the prevention of undesirable effects to health by assessing and adequately controlling harmful exposures to chemical, physical, and biological agents in the workplace. While safety is more broadly dedicated to the prevention of personal injuries and loss of assets and equipment resulting from sudden events and failures, OH focuses on potential hazardous exposures arising from short- and long-term working conditions.

Occupational medicine, also concerned with health protection, monitors for any early signs of illness by means of a health surveillance program on individual workers. Proper interaction and communication between occupational hygiene and occupational medicine guarantees a comprehensive and truly preventive occupational health program.

History. Occupational diseases were identified more than two millennia ago by the Greek physician, Hippocrates, who described lead toxicity in mining (400 b.c.). It is only since the early 1900s, however, that the dimensions of the problem of health and safety of workers have raised general interest and stimulated legislative actions in some industrialized countries.

During the long intermediate period, occupational health remained dormant, with the exception of a few, far-sighted persons who studied and recorded the problem in now famous treatises. Among them are Agricola (1500s), who in his *De Re Metallica* described a lung disease caused by the inhalation of silica (silicosis); Paracelsus (1567), the father of modern toxicology; and Ramazzini (1700s), who first outlined preventive measures for various occupational hazards in *De Morbis Artificum*.

Despite the enactment of some well-meaning legislation and the establishment of public health services in the United States and many European countries, working conditions improved little, and statistics continued to show high incidences of both injury and illness at work until the end of World War II. It was during this period that the first occupational hygiene services and associations were formed in the United States. With the exception of the United Kingdom, occupational hygiene in Europe came later, and a leading role in this field was played by occupational physicians from public services, universities, and industry.

The explosive technical progress in all scientific fields since the 1960s and the enormous development of different industrial activities and services (extractive, chemical, pharmaceutical, agrochemical, automotive, transport, nuclear, electronic, etc.) have required a new and more specialized effort in the increasingly complex field of health and safety at work.

A strong safety culture has been established by industry during this period. This resulted in programs which broadly incorporated health and environmental protection, and aimed at protecting key assets (people, property, and technology). Workplace conditions have progressively improved, as demonstrated by positive trends in safety records for occupational injuries and illnesses. New factors have more recently emerged which have modified public awareness of health and environmental matters, changed priorities and expectations on quality of life, and, in general, reduced trust in industry and its ethics.

The scientific knowledge of the effects of chemical, physical, and biological agents on human health gained by toxicologists and health professionals has increased dramatically, together with the analytical ability to measure smaller and smaller amounts of substances in any media (air, water, soil, food, biological fluids, etc.). Scientific and social standards of acceptability for both workplace and environmental pollutants have continuously been lowered. In addition public perception of the risk arising from chemicals changed after recent industrial disasters which damaged and threatened entire populations (Seveso, Bhopal, etc.). Regulations have become increasingly stringent and sometimes bureaucratic and complex. Industry has recognized these concerns and is working to rebuild confidence and trust with more transparent and coherent environmental, health, and safety (EH & S) programs. In particular, the chemical industry has relaunched its commitments by voluntary worldwide implementation of the Responsible Care program, with excellent and measurable performance in integrated EH & S as the main objective.

In this new climate the role and function of cost-effective occupational hygiene services can

significantly contribute to the goal of protecting and promoting the health and well-being of workers and the community.

3.2. Requirements for an Effective Program

There are no consolidated and internationally accepted requirements for establishing an effective occupational hygiene (OH) program, and different approaches have been adopted in various countries, appropriate to national legislation, cultural background, and needs. Attempts to harmonize OH practices are being made by international associations and standardization bodies [e.g., ILO (International Labour Office), WHO (World Health Organization), EC, etc.] but the quality of service varies considerably, particularly in Europe.

Public OH services exist in many countries. They are usually part of the governmental health and safety institutions, which may have an advisory role and/or assure compliance with legislation. Other services are part of academia and may act as consultants for the government and for companies.

Most large companies have in the last decade established an internal OH service, usually as part of the environmental, health, and safety organization. Private consultancies have developed in some countries, and also represent a resource for medium-sized and small companies that cannot afford an in-house service.

Experience gained from well-run and successful services, particularly in multinational chemical and petrochemical companies, indicates that a number of key basic elements and systems should be considered and implemented. The philosophy and detailed practical guidelines for each element are described in the following sections of this chapter.

3.3. Objectives and Standards

Independent of the nature and size of the OH service, the first and most important factors to be defined with full agreement and committed support of top management, is the definition of mission of the service, together with long- and short-term objectives.

The mission statement together with well-defined objectives will clearly indicate the core values the service is aiming at and provide directions, means, and ethical provisions which will be adopted to achieve the mission.

In addition, the OH service should define the level of standards strived for in the various program elements. In some cases the reference standard can simply be national legislation and practices; in other cases, for large multinational companies, stringent internal requirements are set to guarantee high-quality and consistent working conditions, independent of the local country's economic and social structure, legislation, authorities, and verification of compliance.

Corporate codes of conduct and guidelines should also be issued to harmonize practices and to help implement the programs. Periodic review and adjustment of the objective should also be instituted, together with an auditing system to measure the performance of the service against the targets.

3.4. Organization and Staff Management

The second essential element for an OH service is an adequate organization, sufficiently staffed and well managed, in which allocation of responsibilities is clearly defined. It is clearly specified in all legislations that the health and safety of employees is the responsibility of the employer, and not of the safety, medical, or OH services. The role of the service is to provide management and employees with technical and scientific opinions, expert advice, and authoritative input for implementation and interpretation in the OH field. This is a clear advisory function in which the occupational hygienist should maintain his professional independence and ethics.

Typically, an occupational hygienist has a university degree in science (e.g., chemistry, biology) or engineering, or a specific degree in occupational hygiene or environmental health, plus postgraduate specialization in industrial hygiene/occupational hygiene and accumulated practical experience in this field. This broad, basic knowledge in a variety of disciplines, which should also include elements of toxicology, occupational medicine, safety, environmental protection, ergonomics, analytical chemistry, and

statistics, allows the occupational hygienist to perform well in a complex workplace environment.

Codes of ethics for this profession have been established by some national and international associations [e.g., AIHA (American Occupational Hygiene Association), IOH (Institute of Occupational Hygiene), and IOHA (International Occupational Health Association)]. In addition, there are well-known, highly reputed, professional certification schemes in the United States and the United Kingdom [ABIH (American Board of Industrial Hygiene) and BERBOH (British Examining and Registration Board in Occupational Hygiene)], with the aim of elevating the level of professionalism and guaranteeing high standards. These independent institutions provide certification by examination at various levels of competence and specialization. The ABIH also requires periodic re-training and qualification in order to keep the certification.

It is important that the professional OH service not become disconnected from workplace practices or be regarded as a consultative body. Continuous links with management can be established by implementing a so- called OH unit contact concept. Well-trained persons can be nominated as OH contact in the production unit or plant where they work. In this function they carry out specific routine tasks on a daily basis under the technical assistance of the professional OH staff. These tasks can include sampling and analytical activities, workplace observation, and self inspection. Good cooperation and balanced approach between OH contacts and the professional OH organization is instrumental in the effectiveness of the program. A similar approach with a part-time person, eventually trained and supported by external services, can be adopted by small to medium-sized companies, for which the hiring of a professional OH is economically infeasible.

Continuous communication and good teamwork is needed between the two main health professionals – the industrial hygienist and the occupational physician. This is particularly important in areas where knowledge of each other's findings and data is essential for a truly preventive occupational health program (work history; job class; exposure data; biological monitoring; first aid, acute and chronic cases; health surveillance trends, etc.). Good communication and cooperation should also be sought with other related expert functions, such as safety and environmental protection due to the overlapping and often conflicting fields of interest (e.g., communication and training of employees, personal protective equipment, job safety analysis and working procedures, auditing, ergonomics, accident investigation, air pollution, and emergency responses).

3.5. Anticipation of Health Hazards

The early anticipation of health problems is probably the most cost-effective practice of occupational hygiene. It consists of identifying and correcting sources of exposure or potential health-affecting situations early in the design stage of new facilities and when modifying existing ones. It is cheaper to make changes at the design stage than to retrofit processes when in operation. Although the concept is very straightforward, it is not easy for the engineering/construction group to foresee and minimize all possible factors adversely affecting the project (e.g., process yields, quality, performance, reliability, environment, health, and safety). If the occupational hygienist is involved, for example, at the research and development stage, he can review the properties and intended use of new products and processes. He can help determine additional toxicological testing if necessary, in order to make a preliminary hazard assessment.

Similar assistance should be provided when a material new to a facility is introduced or handled. In this case one should also search for existing occupational exposure limits (OELs) and the availability of adequate sampling and analytical methods in case quantitative evaluation is required. Advice on ergonomically designed equipment, workplaces and procedures, requirements for appropriate hygiene and facility decontamination, and selection and testing of machinery based on specifications driven by health hazard considerations, are other examples of anticipation. Major maintenance work or shut-down operations, which currently represent the conditions with the highest risk level, are also cases where a program should be specially designed to prevent major exposure problems during execution. Consultation of OH be-

fore acquiring an existing facility should also be considered by large companies.

3.6. Occupational Exposure Assessment

3.6.1. Introduction

Research and development (R & D) activities, manufacturing operations and related services (e.g., maintenance, utilities), transport and distribution, end-use and disposal of products are all stages at which, in spite of technological progress, the personnel involved may be exposed to workplace pollutants. This section describes the most traditional and fundamental function of an occupational hygienist: the assessment of exposure to chemical, physical, and biological agents. The more accurate the assessment, the more valid the determination of related health risks that the occupational hygienist and other health professionals can then make.

Appropriate risk assessment allows the risk manager to take the most effective measures to control hazardous exposure and ultimately assure a safe and healthy work environment throughout the working life of all employees. Exposure assessment can be carried out by various methods, which often reflect the different legislative emphasis and OH traditions. The strategy for an exposure assessment described here is a general and simplified one which reflects current accepted good practice and is founded on a preventive philosophy. This strategy is outlined in Figure 2 as a cyclic sequence of steps which are broken down into subsets and described in detail in the following sections. The entire process starts again when a re-assessment is needed due to significant process and procedure modifications, or simply because of ageing of the unit. It is good practice to consider re-assessment every four to five years if exposure conditions have not changed prior to that time.

3.6.2. Characterization and Identification

The first step in occupational exposure assessment is collecting and evaluating the relevant information, documents, and reports that allow full recognition and characterization of the factors and conditions leading to hazardous exposures. This is the preparatory stage upon which the entire assessment is built. The various factors that influence the degree and frequency of exposures depend on the type of work and working environment, how the personnel operate, and the nature of the agents. All these intimately interconnected aspects, once properly described and examined, provide the basic information for the assessment process.

To properly conduct this step, the occupational hygienist should become very familiar with the unit/plant under investigation, and spend sufficient time collecting information, studying, holding discussions with supervisors and workers, and observing what happens in the process. Naturally, previous OH data and experience, if available, are a fundamental source of information. The information, answers, and factors that generally have to be gathered during this preparatory phase are listed in the following:

Work environment and process:	Type of process and technology, chemistry and operating conditions
	Type of production and variation (continuous, batch, etc.)
	Layout and building characteristics
	Equipment and machinery
	Degree and type of environmental controls
	General and local ventilation systems
	Main emission sources and propagation
	Seasonal variation
	Maintenance requirements
	Hygiene facilities
Workforce and human factors:	Number of people, age, sex
	Shift system
	Awareness and training
	Personal protective equipment
	Workclothes
	Food and drink consumption
	Routine operations and tasks
	Infrequent and critical tasks
	Typical length and frequency
	"Dirty jobs"
	Special groups at risk
	Contractors
Agents and their characteristics :	List of chemical substances (raw materials, intermediates, byproducts, products, etc.)
	Physical and warning properties
	Noise, vibration, heat stress, and other physical agents
	Toxicological profile and main health adverse effects
	Acute, chronic, carcinogen, mutagen, reproductive
	Immediate or delayed
	Reversible and irreversible
	Asphyxiant, irritant, allergenic, etc.
	Biological agents
	Route of exposure (inhalation, ingestion, skin contact)
	Applicable occupational exposure limits (OELs)

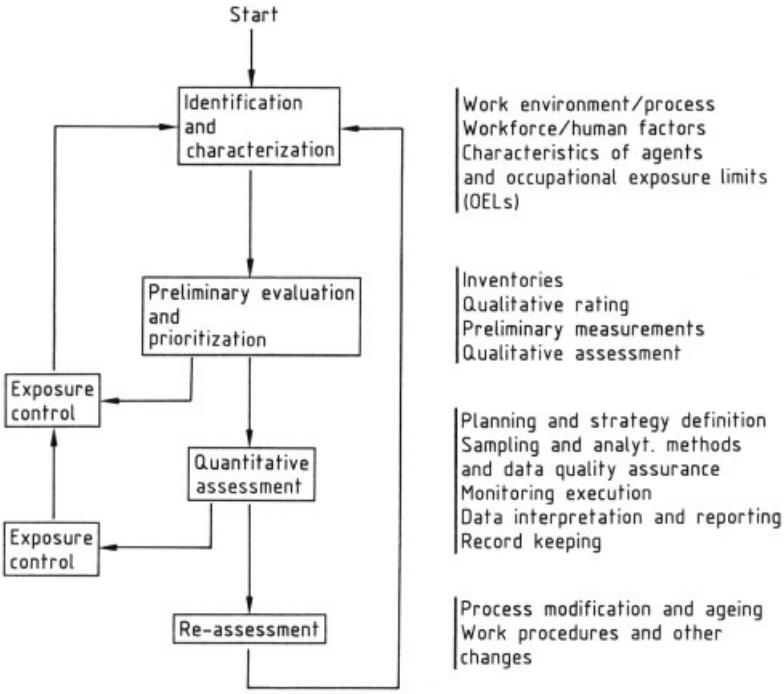

Figure 2. Outlines for an occupational exposure assessment

A list of typical areas, operations, and conditions where experience has shown that higher risk of exposure can occur is as follows:

Warehouse	Sampling raw materials Fork lift trucks
Unloading/loading stations	Ship, barge, rail car, tank truck unloading/loading operations Handling bags/sacks/drums
Laboratories	Handling incoming samples Storage and handling of reagents/solvents Lab waste disposal/glassware washing
Process area	Tank vents, pumps, centrifuges, filters, dryers, extruders, grinders, mill furnaces, process sampling, adding raw materials, filling line, packaging and bagging operation, waste handling
Maintenance and mechanical shop	Decontamination and line breakup Welding, soldering, cutting, grinding, machinery degreasing, polishing, sand blasting, vessel entry, and confined space operation

The list is not exhaustive and refers to chemical production facilities, but can be extrapolated to other operations.

3.6.3. Preliminary Evaluation and Prioritization (Qualitative Assessment)

By using a systematic approach and documenting the preliminary evaluation as described in recent publications or recommended by some codes of practice based or legislation (Germany, United Kingdom, EC, etc.), the preliminary evaluation becomes a valid qualitative assessment of health risks. In addition, it can constitute the basis for a system for prioritizing potential OH concerns so that resources can effectively be directed to those conditions that pose the greatest risk. This prioritization scheme is determined by the qualitative assessment and monitoring results, but can also be used to establish priorities in other important areas, such as exposure control, communication, and training of employees.

Preliminary assessment consists of estimating the nature and degree of exposure, combining considerations of type and extent of health effects, and finally deriving a sound judgment of the overall related risk. The degree of exposure is estimated by reviewing the information collected during basic characterization with one or more of the following tools: previous monitor-

ing data, experience from similar conditions, exposure modeling techniques, semi-quantitative field monitoring, and spot measurements. Qualitative ratings or indices can be established for typical exposure categories for tasks performed by homogeneous exposure groups: 0 = none or negligible, 1 = low and infrequent, 2 = moderate and infrequent, 3 = high and frequent.

Similarly, toxicological properties, related health effects, OELs, or other parameters can be used to qualitatively classify the hazards of agents: 0 = minimal/low effects, 1 = moderate/reversible, 2 = moderate/high, 3 = very high/ irreversible. Lack of OELs does not prevent assessment but makes it more complicated. A useful approach is to assign to the agent a more stringent priority category where doubts exist.

By combining the degree of exposure ranking and the hazard classification of agents, it is possible to assign an overall qualitative risk index and prioritization to the various homogeneous exposure groups or working areas. In most modern workplaces, which are characterized by an enclosed and fully automated process, the assessment normally indicates negligible risk and low OH priority; few are normally judged unhealthy and require control (Fig. 2); some others cannot be properly evaluated with a qualitative method and require further, quantitative assessment. The above rating refers to routine, normal operating, or planned conditions and cannot be used for emergency and unplanned operations.

3.6.4. Occupational Exposure Limits

Occupational exposure limits (OELs) are the most traditional and widely used criteria to assess the acceptability of workplace conditions. The various types of OELs for chemical and physical agents are established by means of different approaches, purposes, definitions, and updating systems by various national and international regulatory bodies. Most are guidelines for evaluating health hazards and assisting in other preventive activities (adequacy of control, design standard, selection of PPE, etc.). Others have legal status and serve for demonstration of compliance. OELs are based on current knowledge of toxicological and health effects of the substance on the working population, and on the best judgment of health scientists and occupational health professionals.

In general, an OEL for a chemical substance refers to the airborne concentration to which most workers can be repeatedly exposed by inhalation for a full working day and working lifetime without adverse short- or long-term health effects. However, OELs do not guarantee complete protection of all workers and cannot be used as an absolute dividing line between safe and harmful concentrations, but only as a guideline.

OELs have a long tradition worldwide, and the first recommendations were published in Germany by Lehmann in 1889. A systematic development of OELs started in the United States in the mid-1940s with the ACGIH-TLV system which has remained the most widely used and highly reputed system to today. Germany started an independent MAK system in 1965, and after this other nations started their own national schemes (Sweden, United Kingdom, France, The Netherlands, Denmark). In 1977 the International Labour Office (ILO) in Geneva published its "air quality limits" for the workplace, followed in 1980 by the WHO with "safe limits" for occupational exposure. The ILO and WHO OELs are health-based limits established solely by considerations of the health of the working population. Most other OELs, especially those with regulatory compliance status, incorporate other technical considerations (technical feasibility, analytical methods) as well as socioeconomic factors.

In light of the differences that exist between OELs from country to country in Europe, which is also influencing the principle of free movement of labor, the EC Commission began in the 1980s to tackle the problem by offering a scheme to harmonize OELs in the Member States and to coordinate the activities of the national bodies. More than 1500 of the most important chemical substances now have an official OEL established by combining the activities of the various agencies around the world, and many hundreds more are set internally by international enterprises. Although these limits are relevant for many key occupational health concerns, the very large number of chemicals on the market (100 000) represents a complex challenge for both authorities and companies.

When no OEL has been established for an existing chemical (particularly common in the case of process intermediates) or for a novel substance, assessment becomes more complicated, and assistance from other health professionals (occupational physicians and toxicologists) is necessary. A program to obtain a minimum set of toxicological data with the aim of establishing a preliminary OEL or internal guideline is required.

Different categories of OELs are established to take into account the toxicological properties and related health effects of the substances. The most common category is the TWA concentration for a defined period of work (normally 8 h). A defined number of excursions above the TWA are allowed per shift, taking into account the short-term effects of the substance (Short-Term Exposure Limits, STELs). For certain substances, where exposure can result in acute health effects, there is a ceiling limit that describes the concentration beyond which workers should not be exposed under any circumstances.

For a limited number of chemical substances, a biological exposure limit has been established. This provides criteria for carring out and evaluating a biological monitoring program which assesses the overall internal exposure to the chemical by analysis of biological samples. It should complement the traditional OH airmonitoring program.

3.6.5. Quantitative Assessment – Occupational Hygiene Monitoring

The preliminary assessment step indicates, in order of priority, which exposure conditions, related agents, groups of workers, and homogeneous job groups need to be assessed by an appropriate monitoring program.

The objective of this section (see also Fig. 2) is to analyze the elements and provide guidance on how to plan, conduct, document, and maintain records of an investigation evaluating exposure of workers to chemicals.

The technique for evaluating exposure in the workplace has changed in many ways over the years. Work situations in which workers are clearly and grossly exposed have declined; in most cases, the levels of exposure are quite low. A quantitative assessment of very low levels of complex chemical mixtures is a challenge requiring professionalism, experience, and a carefully thought out working model. Before the start of any monitoring, the preparation of written plans for appropriate monitoring (protocol) is the first step to assure the required level of quality, reliability, and effectiveness of the entire program.

3.6.6. Monitoring Planning and Strategy Definition

There are many reasons for conducting an OH monitoring program besides evaluating personnel exposure: to demonstrate regulatory compliance, verification of applied control, selection of PPE, continuous air monitoring for leak detection and air emissions, etc. Each requires a different strategy. The first elements of a monitoring protocol should state clearly the purpose of the program and define the most appropriate strategy to accomplish these objectives. A detailed and in-depth review of information and data collected during the preliminary assessment, specifically related to the agents and workplace prioritized for the monitoring, is then required. Defining a monitoring strategy simply means answering the basic questions what, when, how, how long, and how frequently sampling should be carried out.

Typical workplaces are complex, and concentrations of environmental agents vary temporally and spatially within and between workdays. Workers move through these variations in concentration in variable patterns, and they also influence the concentration and the degree of workplace contamination by the tasks that they perform. The exposure routes by which workers can absorb a substance must be carefully studied in relation to its physical, chemical and toxicological properties. Inhalation is usually the primary mode of exposure, and this depends on the vapor pressure of the substance, temperature and pressure of the process, etc. Absorption via skin contact can also be significant, especially if the substance readily penetrates the skin in toxic amounts. Exposure via ingestion is generally negligible, but it is necessary to watch for

potential intake via food, beverage consumption, and smoking.

The toxicological properties of the substance and its potential health effects should be reviewed in detail: how it exerts its acute properties; the type of chronic toxicity; the target organs; the carcinogenic, reproductive, and neurotoxic potential; the reversibility or irreversibility of the effects; and the biological half-life, etc. A careful review of the documentation of the OEL of the substance is necessary in order to understand the rationale and the safety factors adopted.

There are two major categories of sampling strategy: traditional and statistical. In practice there is a third category that results from specific regulatory requirements, where approach, methods, procedure, etc. are indicated by the legislation. The traditional diagnostic approach is based on past experience and professional judgment of the industrial hygienist. It is normally an effective combination of the most common types of sampling: personal, task, and stationary sampling. TWA personal sampling is performed by taking samples with miniaturized devices in the breathing zone of a worker for the entire shift (ca. 8 h). This is the preferred and most effective technique to evaluate exposure.

Task-related sampling is carried out, for example, at the workers' breathing zone during potentially critical operations and tasks. Stationary sampling, carried out in strategically fixed locations of the process, is generally useful in identifying the source of exposure and can be used to estimate worker exposure if other sampling is impractical. Continuous air monitors are generally sophisticated stationary sampling systems equipped with alarms, which are installed to measure one or more chemical agents in various selected parts of the process area. They can be used as leak detectors and for monitoring background concentrations.

There are various statistical approaches reported in the literature and specifically recommended for regulatory compliance. The selection and application of the most appropriate and powerful statistical model, backed up by sound professional judgment, can result in enhancement of the overall quality and credibility of the program.

Frequency of sampling is dictated by statistical considerations, and largely depends on the variability of the conditions being sampled. The more variable the conditions are, the more samples are required. However, 5 – 8 sampling repetitions is a good rule of thumb, valid for most conditions. Different shifts should be sampled for the same homogeneous exposure group. Work load, seasonal, and evironmental conditions should be considered. The conditions during sampling should be representative of the conditions that one intends to evaluate.

Estimation of potential for skin contact can be made by measuring surface or item contamination by wipe-tests. Choosing an adequate analytical method to match the selected sampling strategy is the next step, and often requires help from analytical chemists.

The final important elements of a monitoring protocol include a careful timetable and cost estimate for the entire program. Extensive monitoring can become extremely expensive and may not be necessary and/or add very little valuable information to the assessment. A proper balance between needs and scope of monitoring and cost should be the goal.

3.6.7. Sampling and Analysis; Data Quality Assurance

Sampling equipment and analytical techniques for air monitoring have undergone marked changes over the past decades, with improved detection sensitivity, miniaturization, and automation.

For the most important chemical agents, specific and reference methods are published by the National Institute of Occupational Safety and Health (NIOSH), and DFG (Deutsche Forschungsgemeinschaft). These methods generally describe sampling, analytical procedure, and instrumentation and specify the important parameters that have been determined in laboratory and field tests to satisfy good practice for trace monitoring. Generally the range of applicability, stability, recovery, effect of humidity, interferences, limits of detection, accuracy, and precision are indicated. When a reference-validated method is available and its field of application satisfies the needs of the planned monitoring strategy, the normal quality assurance tests, such as instrument calibration, blanks, and known standard samples should be run to verify

that the sampling and analytical conditions chosen correspond to those of the validated method. If no method is available, a plan for its development should be designed in cooperation with analytical chemists. Air monitoring techniques, which are highly developed because OELs are generally based on inhalation exposure, can be conducted by:

1) Collecting contaminated air in glass bulbs, plastic bags, etc. for later analysis
2) Instruments and techniques that directly measure the pollutants in air
3) Removal of the chemical agent from the air for later analysis

Direct-measurement techniques range from relatively simple colorimetric detectors to highly sophisticated portable infrared detectors or gas chromatographs (GC). They are extremely useful as preventive tools when equipped with alarm systems, or in diagnostic evaluation.

The most widely used techniques fall into the third category, where a volume of air is drawn by a pump (active system) or diffusion (passive) into an appropriate collecting medium, either solid (e.g., charcoal tubes) or liquid (e.g., solvent). This is generally used for vapors and gases, which are then extracted and analyzed. Similarly, filters or membranes are used to remove airborne particulate matter (dust and aerosols) from pumped contaminated air. When a specific fraction of the particles (e.g., inhalable fraction) is to be measured, an appropriate size-selecting device is placed before the membrane.

Common analytical techniques include GC, Atomic Absorption (AA), HPLC, and GC – MS. Classification and counting under the phase-contrast microscope is carried out for airborne fibers (e.g., asbestos).

3.6.8. Sampling and Analysis Execution

Good communication and arrangements should exist with the unit and analytical lab supervision prior to the sampling and analytical execution in order to take into account production schedule changes, etc. This is particularly important when monitoring specific and infrequent tasks. Sampling preparation often requires calibration of sampling devices and instrumentation, and labelling of tubes, badges, vials, and bags.

During sampling execution, it is advisable to use predefined sampling record forms, in which the key information and observations noted during sampling are written down (humidity, setup conditions, use or not of personal protective equipment, etc.). Handling, temporary storage, and transportation of collected sample to the laboratory should be done according to planned procedure to minimize errors and prevent invalidation of results.

3.6.9. Data Interpretation and Reporting; Record Keeping

Once the analytical results have been converted into appropriate units (e.g., ppm), they must be grouped into meaningful categories (TWA personal data of each homogeneous exposure group, the STEL results of similar operations, etc.) and then studied by adding the information noted during sampling. Appropriate statistical tools are then used to describe the central tendency of samples (e.g., arithmetic mean) and their distribution (range, geometric standard deviation, etc.). Ultimately, the best estimate of exposure is compared with the OEL to determine the acceptability of the assessed exposure conditions.

Another possible interpretation is that it is not possible to make a decision with the data available, and further monitoring is necessary. If the sampling strategy includes a specific statistical analysis, additional and more precise conclusions can be drawn (e.g., chance of exposure above the OEL).

The main elements of a quantitative assessment should be documented in an appropriate survey report. This should include purpose, information on the process/unit and homogeneous exposure groups, OELs and criteria of evaluation, sampling and analysis, data quality assurance, tables of results, conclusions, and recommendations.

Because much of the information is already available in the monitoring protocol, proper use should be made of existing documentation and/or references to avoid duplication of efforts. The recommendations, if related to exposure control methods, should be made in consultation with management and workers' representatives if necessary.

The OH survey report is a document that must be completed in accordance with the retention requirements of the company and/or the general legislative retention time of 30 – 40 years. Other OH documents that should be kept are: chemical and physical inventories with qualitative assessment; work histories; homogeneous exposure group and other OH documents that, together with health surveillance and biological monitoring data, can serve future epidemiologial or legal needs.

3.7. Exposure Control Methods

3.7.1. General Concepts

The most important function of a sound occupational health program is controlling and reducing exposure to hazardous agents to acceptable levels as part of risk management. The entire function consists of the selection and application of the most appropriate control measure and subsequent verification of its effectiveness.

Even if he is not directly responsible, the specific knowledge and experience of an occupational hygienist, particularly in certain control methods [e.g., industrial ventilation, personal protective equipment (PPE)], can contribute significantly to the success of the entire process. When conclusions drawn from qualitative and quantitative assessments (see Fig. 2) indicate that reduction of exposure is necessary, a number of factors should be considered by management in selecting the most appropriate method. Such factors should include urgency, cost-effectiveness, reliability, maintenance, and comfort of employees. Methods of exposure control can be divided into different categories. The order in which these categories are listed below corresponds to the hierarchical order of the risk manager's options; he should make his choice in line with good OH practice and in the spirit of current legislation.

1) Substitution (e.g., replacement of a toxic substance)
2) Containment (e.g., total enclosure)
3) Removal at source (e.g., local exhaust ventilation)
4) Administrative/procedural (e.g., reduced length of exposure)
5) Personal protective equipment (e.g., dust mask)

The above approach is based on the fact that elimination of exposure should be considered first, while PPE should be adopted as the last line of defense when it is neither possible nor practical to use any of the other methods. Using a material with better toxicological properties or different emission characteristics is clearly the best method of controlling exposure, although this is generally quite difficult because it requires extensive research and investment.

Control by containment means enclosing the process so that the characteristic emission of the substance cannot reach the worker. An automated feed system in processes and glove boxes in labs are typical examples. Removal at source normally implies the use of a local exhaust ventilation (LEV). The principles behind control of exposure by LEV are as follows:

1) Design the maximum degree of enclosure around the operation that still allows the work to be done efficiently. Decide on the necessary capture velocity to entrain emissions at the remaining enclosure opening (e.g., 0.5 m/s for vapors released into still air).
2) Calculate the necessary volume flow rate to achieve the required capture velocity.
3) Design, install, test, and maintain the system.

3.7.2. Personal Protective Equipment

Where it is not feasible or practical to adopt one of the other methods, including procedural measures, personal protective equipment (PPE) must be considered. Management has the responsibility to ensure that correct PPE is available and used, while workers must ensure its proper use. PPE includes a variety of devices for head, eyes, face, hands, ears, body, feet and lungs, providing protection against a wide variety of workplace health and safety hazards (chemical, physical, mechanical, biological, electrical, falls, fire, etc.). A sound program for the proper use of PPE taking into account its limitations, should be established and should address in general the following points:

1) Identification of health and safety conditions and hazards

2) Selection of the most appropriate PPE types
3) Procedures for use, storage, cleaning, maintenance, and disposal
4) Emergency procedure and use of PPE
5) Medical fitness and physical stresses
6) Education, training, and retraining of wearers
7) Inspections
8) Documentation

The Commission of the European Community (CEC) has promulgated a directive (89/656/EEC) which lays down minimum requirements for PPE used by workers at work. This directive, together with related CEC brochures and, above all, the standardization of technical specifications for PPE in CEN standards will harmonize manufacturing, selection and proper use of PPE in the EU countries, which up to now was regulated by national standardization schemes.

The safety officer normally administers the entire program. However, OH skills and knowledge are particularly significant when adequate respiratory protection and skin protection from chemicals are to be assured during routine tasks, as well as during special emergency operations.

The definition of the most suitable types of respirators derive from the analysis of the hazards and the assessment of risks to personnel, and then comparing merits and limitations of the various types and classes available (air-purifying respiration, air-supplying respirators, tight or loose fitting face piece, negative or positive pressure mode, escape only, etc.)

Consideration of the nominal protection factor (PF) assigned to a certain class of respirator, the required level of protection, the specific face-fit of the wearer, etc. are very important for an appropriate selection and are vital in critical conditions when personnel with facial hair may also be involved.

The protection factor is the ratio of ambient concentration outside the respirator to that inside the respirator facepiece. Multiplying the PF by the OEL of the substance under consideration gives the maximum use concentration.

The selection of adequate protective clothing as a barrier against chemical penetration also requires specialized knowledge. Interpretation and judgment of extensive and often conflicting data from suppliers and literature on chemical permeation testing is needed, together with considerations of comfort and hand sensitivity.

3.8. Physical Agents

Various physical agents may be encountered in the workplace. Some, such as noise and vibrations, are very diffuse and affect many workers; others, such as compression and decompression, are typical of work environment where the atmosphere is not at normal pressure and temperature (e.g., divers).

Other physical agents include:

1) Stress due to heat and cold.
2) Sources of intense optical radiation, such as ultraviolet and infrared lights and lasers, which may produce eye and skin damage.
3) Electromagnetic fields, for which certain chronic health effects are still under investigation.
4) Ionizing radiation, typical of the nuclear industry and in various other applications (e.g., medicine, gauging equipment). Exposure to ionizing radiation causes damage to living cells.

Each physical agent has its own characteristics, mode of entry, health effects, etc., and therefore requires a specific approach and technology if it is to be assessed and controlled in the workplace. Specialized literature, international guidelines, and legislative requirements for physical agents have been established in some countries.

Occupational exposure to *noise* is by far the most diffuse, most widely studied, and has the greatest impact of all physical agents. It has been estimated that $> 13 \times 10^6$ workers in the EU states are exposed to excessive noise for most of their working time. Although for many years noise at work has been known to cause hearing loss, it was only during the last three decades that major scientific studies have allowed full understanding and quantification of relationship between noise exposure and risk to the human auditory system.

Noise, or an unwanted sound, is described as oscillations in pressure above and below atmospheric pressure which evoke an auditory sensation. Prolonged exposure to sufficiently high levels of noise can produce permanent damage

to the hearing mechanism. Other nonauditory effects (e.g., physiological) can be attributed to noise exposure at lower levels. Such effects are less well-known and also occur in many human leisure activities. The level of exposure to noise for which there is international consensus that the risk of hearing damage is limited and at which most legislations set the limit for 8 h of exposure is 85 dB (A) as a time weighted average. Impulse noise should also not exceed 120 dB at any time.

Assessment of noise exposure is the first important element of a complete hearing conservation program. The other integral steps are: technical measures and abatement techniques; administrative measures (reducing time of exposure, area delimitation, noise specifications, etc.); selection and use of hearing protection (various types of ear muffs and ear plugs); education or training of personnel; and preliminary and routine audiometric testing to check the degree of hearing impairment.

There are two basic methods of assessing noise exposure. The first consists of fixed-position measuring of noisy areas or equipment by a sound level meter which can also measure impulse noise, and calculating the total noise exposure, taking into account the time spent by the worker in each measured area. In the second method, the worker has a personal audiometer, capable of continuously monitoring noise levels throughout the working shift. The use of both techniques at appropriate intervals can be considered the most effective approach for assessing noise exposure.

3.9. Communication and Training

The final essential element of an OH program is providing information, training, and education in the field of IH and hazard communication for personnel at all levels. This is a basic requirement of all health and safety legislation. Regular communication of accomplishments, performance, new relevant information, issues, concerns, and regulatory development to management is needed to maintain high awareness and support.

Potentially more exposed personnel who operate equipment and handle materials can contribute with personal control and attention in minimizing exposure if they understand the nature of the hazard and how work procedure affects emission and contamination.

OH training is normally done in cooperation with the supervisor and other departments (e.g., safety and medical) as part of the initial routine health and safety training. Topics that should be included are:

1) Physical, toxicological, and warning properties of substances and their possible consequences and health effects upon exposures
2) Ergonomic factors and recommendations
3) Results and interpretation of occupational exposure assessment
4) The proper use and limitations of PPE

Many specific sources of such information are available at international and national level. The most appropriate approach and method to provide such training should be designed according to the background and level of education of the people it is intended for.

Verification of the effectiveness of this program, for example, by means of a questionnaire, should be considered. Training records should be made according to legislative requirements and practices.

3.10. Industrial Hygiene in the United States

Although some education programs existed previously, the first industrial hygiene section was organized by the Public Health Association in 1914.

In 1915 the U.S. Public Health Service organized a division of Industrial Hygiene and Sanitation. The *Journal of Industrial Hygiene* was first established in 1919 and in 1925 became the *American Industrial Hygiene Association Journal*. The present Harvard School of Public Health had a Department of Industrial Hygiene in 1922. Chemists, engineers, physicians, toxicologists, and other specialized graduates have applied their knowledge to the growing scientific profession of industrial hygiene. The American Hygiene Association was organized in 1939. Since then improvements in plant design, machine design, and plant layout have seen the worker come to expect a safe workplace, in

part because of the activities of the industrial hygienist.

With progressing industrial and workplace safety, today's industrial hygienist is still involved with control of dust hazards, skin irritants, noise levels, personal protective equipment, classification of hazards, and training. However, with additional data, better monitoring methods, and more precise measurements, chronic occupational problems are now of greater importance. Information has been developed on chemical exposure limits in workplace air by several organizations such as the National Institute for Occupational Safety and Health (NIOSH) with recommended exposure limits (RELs). The Occupational Safety and Health Administration has proposed permissible exposure limits (PELs) for several hundred chemicals. If accepted by the legal process, these could become regulatory standards. As with all OSHA standards, the PELs are being strongly challenged in the courts.

The American Conference of Government Industrial Hygienists (ACGIH) have developed guidelines or recommendations called threshold limit values (TLVs). These recommendations and guidelines all assist the industrial hygienist with the control of potential health hazards. In addition, for different working conditions, these values are expressed as time weighted averages (TWAs) and as threshold limit values- ceiling (TLV-C) for exposures not to be exceeded even instantaneously.

All of these guides and values represent levels to which nearly all workers can be exposed without adverse health effects. With the help of data automation, the industrial hygienist sets up sampling monitoring and analytical programs to measure and correct exposure conditions. Physical agents such as temperature, stress, radiation, vibration, and pressure variations are more recent areas of study where practicing industrial hygienists can continue their strategy of occupational safety.

Today's industrial hygienist must draw on the knowledge of a broad range of disciplines in the medical, biological, and technological fields to prevent work strain from reducing performance or becoming pathological. The primary objective is to provide a workplace free of recognized hazards.

4. Occupational Medicine

4.1. Health Surveillance [29–32]

One of the main goals of occupational health care is to protect workers from harmful effects due to work or the working environment.

The various means that can be used to achieve this goal can be summarized under the heading of health surveillance.

In general terms health surveillance comprises four sequential steps:

1) Monitoring/measuring
2) Evalution
3) Interpretation
4) Corrective actions

The first step in health surveillance consists of measuring parameters of exposure and uptake or effects. Secondly, the results must be evaluated and interpreted with respect to the question of whether any health impairment exists or is to be expected.

When it is concluded from the monitoring results that inadmissible exposure or uptake is occurring, corrective measures should be taken to prevent further health risks. Which measures have to be taken depends on the particular situation and is decided by a case-to- case approach.

Health surveillance in the framework of occupational health care may be aimed at:

1) Specific occupational branches, such as health care services, construction and building industry, chemical industry
2) Special occupational cohorts, such as busdrivers, railway personnel, operators in the process industry
3) Particular occupational diseases, such as hepatitis in nurses and skin conditions in bricklayers
4) Occupational health problems due to exposure to particular biological, physical, or chemical agents

The discussion in this chapter is restricted to examinations dealing with potential hazards from occupational exposure to chemicals.

4.1.1. Monitoring

Monitoring is "the repetitive and continued observation, measurement and evaluation of health and/or environmental or technical data for defined purposes, according to pre-arranged schedules in space and time and using comparable methodologies for sensing and data collection" [30]. This description is still valid.

The most essential criteria for reliable monitoring procedures are:

1) The activity should serve a defined objective
2) Continuous or repeated observation
3) Comparison of the data observed with a reference level [31]

Monitoring implies a multidisciplinary activity in testing, watching, and observing situations and/or persons. In the occupational setting this means examinations of the working environment and/or the workers.

Which mode of monitoring will be preferred depends primarily on the goal of the examination and a number of factors, which are discussed below.

Monitoring Strategy. Before any health surveillance action is started, a decision has to be made about the nature and design of a prearranged schedule, adopted to meet a defined objective. This decision depends mainly on the answers to a number of questions to be posed, certain criteria to be met, and the availability of appropriate monitoring methods.

The most important and crucial questions deal with the aims and means of a health surveillance program:

Why carry out an examination? What are the objectives and what do we want to know? Is exposure of workers actually taking place?

The answer to these key questions generally implies the answers to the next questions.

Who must be included in the surveillance program? What is the actual population at risk? Which criteria are valid for assessing this risk group?

What will be measured and monitored?

The answer to this question presupposes some knowledge about the nature of the exposure.

What is the most suitable way and the most relevant place to monitor and which parameters need to be measured?

The answer to this question presupposes fairly good knowledge of the biological fate of a chemical in the human body.

Which sampling procedures and analytical methods are available and suitable in a particular situation?

When and how often must monitoring be performed?

In some countries detailed guidelines or strict rules are given for the methods and frequency with which health surveillance is to be performed. In Germany this is regulated by the "Berufsgenossenschaftliche Grundsätze für Arbeitsmedizinische Vorsorgeuntersuchungen" [31]. Guidelines are given in the United Kingdom by the Health and Safety Executive (HSE) and in the United States by the Occupational Safety and Health Administration (OSHA).

Monitoring Criteria. Apart from the above mentioned questions, some criteria must be met for proper design of a health surveillance program.

In 1975 WHO issued criteria for periodical medical examination that are still valid. The most important of these criteria are:

1) It should not involve undue expenditure of time, equipment, or staff.
   In other words, the cost and discomfort for the individual and for society should be in balance with the health benefits to be expected.
2) It should not be inconvenient to workers.
   A liver biopsy might be considered to be the best diagnostic means in a particular situation. But, as this operation could be fatal, one is very reluctant to perform it for the medical surveillance of healthy people.
   This criterion can also play a role in the choice of biological medium. Thus taking a blood sample or collecting a 24-h urine sample could be experienced as difficult and inconvenient.
3) It should be an early reversible predictor before overt health effects are impending. Surveillance activities which can be classified under biological monitoring and biological effect monitoring meet this criterion.
4) There should be a quantitative relationship with exposure and with health risk.

For most effects caused by chemicals, a dose – response relationship exists. Measuring a particular parameter is only significant when this dose – response relation and/or dose – effect relationship can be expressed numerically. Preferably, the effect measured should have a steep dose – response curve which enables the intensity of effects to be more easily distinguished.

For proper evaluation and interpretation, valid reference data are essential.

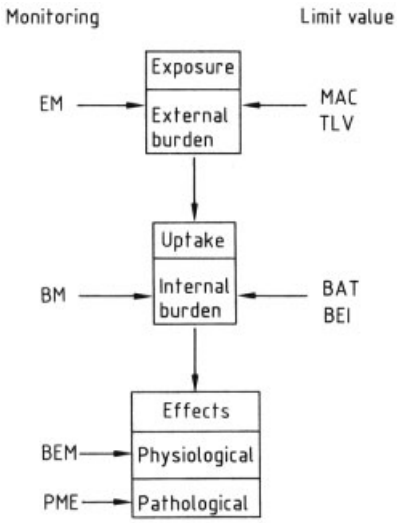

Figure 3. Instruments for health evaluations of workers
EM = Environmental monitoring; BM = Biological monitoring; BEM = Biological effect monitoring; PME = Periodical medical examination
MAC = Maximum allowed concentration; TLV = Threshold limit value; BAT = Biologische Arbeitsstofftoleranz; BEI = Biological exposure indices

A very important criterion is that the tests should have high validity and effectivity.

Thus the aim is the highest possible specificity and sensitivity. Low sensitivity leads to an unsatisfactory number of false negatives (positives that wrongly score negative). The higher the specificity of a test is, the lower is the number of false-positives.

An essential condition for performing meaningful health surveillance programs for workers is that the management responsible is prepared to improve working condition if this is indicated. Also employees must be willing to improve their working habits and hygiene when necessary. If these requirements are not met, any health surveillance program is completely redundant and a waste of time for everyone involved.

Modes of Monitoring. The following modes of health surveillance can be distinguished:

Estimation of the health risk:
- Environmental monitoring (EM),
- Biological monitoring (BM),

Estimation of the state of health:
- Biological effect monitoring (BEM)
- Periodical medical examinations (PME).

In Figure 3 a scheme is given of these "instruments" for health evaluations.

These different monitoring methods can be exemplified on the basis of lead exposure.

When workers are exposed to lead fumes or dust, the following can be investigated:

1) The concentration of the fumes or dust in the working area (environmental monitoring, EM).
2) The content of lead in blood or urine of the exposed workers (biological monitoring, BM).
3) Indications of biological changes in blood production by measuring the content of δ-aminolevulinic acid dehydrogenase (δ-ALAD) or zinc protoporphyrine (ZPP) in the blood, or δ-aminolevulinic acid in the urine. This method is known as biological effect monitoring (BEM).
4) Indications of impaired health by physical medical examination of symptoms and signs of specific adverse health effects: anemia, lead line on the gums, abdominal cramps, obstipation, wrist drop, impaired renal function, and blood changes.

Environmental monitoring (EM) in this context means measuring exposures in the air at the workplace. These measurements are generally specific for a defined chemical. EM is used to determine the intensity of exposure, expressed as the concentration of a single chemical or group of chemically related compounds in the working environment.

However, exposure to a chemical or group of chemicals, does not necessarily mean that uptake into the human body will happen or has already taken place. Whether, in what way, and to what extent uptake results from exposure depends on a number of factors.

Air sampling close to the mouth and nose of an individual gives the best reflection of the actual nature and intensity of the chemicals inhaled during work. This type of EM is called personal air sampling (see Section 3.6.6).

In biological monitoring (BM) a chemical compound or one of its metabolites is determined in a biological medium, mostly urine or blood. BM can be regarded as a person-related, specific measurement of so- called biomarkers, which reflect the internal burden of a compound in the human body, irrespective of the route of entry.

Therefore, BM is a relevant method of health surveillance when individual screening is the goal. However, when it is intended to check or to control the working environment, environmental monitoring is the first method of choice. BM is discussed in more detail in Section 4.2.

4.1.2. Periodical Medical Examination

Periodical medical examination (PME) is the periodically repeated medical examination of individuals for health effects that may be related with their work or working circumstances. The main aim of these examinations is to detect early, reversible (pre-)clinical signs of health impairment.

PME is individual-related and aimed at the detection of nonspecific anatomical and/or functional disturbances of organs or tissues, irrespective of the cause.

Besides occupation, a number of intrinsic and nonoccupational extrinsic factors, including eating, drinking and smoking habits, may influence the nature and intensity of health effects. This means that in most cases it only can be estimated or inferred indirectly from occupational histories that a specific chemical is the causative agent.

A periodical medical examination implies the drafting of a medical history, a physical examination and, when indicated, additional tests.

Elements of a periodical medical examination are:

Medical History
- current state of health
- previous ailments
- medication
- life style

Physical examination

Additional examinations
- blood or urine tests
- lung function
- neurological/electrophysiological
- chromosomes
- sperm

Example 1. The presence in blood of a particular quantity of carbon monoxide bound to hemoglobin (CO-Hb), may be caused by the exposure and uptake of carbon monoxide itself, but can also be due to exposure and uptake of the solvent dichloromethane. Heavy smoking has to be considered as a confounding factor.

Example 2. A large number of pesticides, chiefly organophosphates, act in the body by blocking the enzyme acetylcholinesterase, which regulates nerve – muscle action by hydrolyzing acetylcholine; thus, excessive amounts of this muscle stimulator accumulate. However, the clincal symptoms of inhibition of this enzyme do not indicate which particular acetylcholinesterase inhibitor is the causal agent.

Example 3. Headache may be caused by a large number of etiological factors, unrelated to any occupational activity. But even when a relationship with occupation cannot be ruled out, it is often difficult to determine whether this symptom is caused by the exposure and uptake of a neurotoxic organic solvent or is due to mental stress at work.

Example 4. A disorder of liver function may be of bacterial, viral, or chemical origin. But even when it is obvious that it has been induced by a chemical agent, it may have been the result of a lifelong excessive drinking of alcoholic beverages during leisure hours, the use of particular pharmaceuticals, or exposure to and uptake of chlorinated hydrocarbons during work.

In the early days periodical medical examination of individual workers consisted almost entirely of detection of clinically apparent symptoms. Nowadays, increasingly subtle diagnostic methods are used, enabling the physician to detect early reversible effects which may be considered to be not yet adverse or pathological.

A number of early effects can be regarded as forming part of the physiological defense mechanism of the human organism. Zielhuis and Henderson [32] use the term biological effect

monitoring for this kind of examination and defined it as: "the measurement and assessment of early biological effects, irrespective of their significance to health, in exposed workers and health risk compared to an appropriate reference."

However, the line between physiological and pathological effects is often hard to draw because the distinction may be indistinct.

For some effects it is still unclear whether they should be considered adverse or harmful, or neither. Examples of effects that are still under discussion between experts are enzyme induction (see Section 4.5) and some electrophysiological phenomena in nerves or muscles.

Apart from whether a particular effect can be regarded as a biological or pathological response of the human organism, there is another crucial criterion for deciding whether a specified effect is tolerable or not. This criterion is the reversibility of an effect after exposure to the causative chemical agent ceases.

Sometimes it is a matter of academic or political dispute whether a particular person-related measurement is called BM or BEM. Although this does not change measurement method, it can have a major impact on the establishment of biological limit values and the setting of standards.

For example, the determination of carbon monoxide – hemoglobin complexes in the peripheral blood may be regarded as an indication of the total body burden of CO (which is BM), but also as a parameter of BEM because it reflects the proportion of hemoglobin that is blocked and not available for oxygen transport from the lungs to the tissues.

4.1.3. Evaluation and Interpretation of Data

The most important step in the entire health surveillance (HS) process is the evaluation and interpretation of the test results.

For interpretation of the monitoring results, the data must be compared with reference data. Therefore, it should be clear before starting any health surveillance program which references can and should be used.

The availability of valid reference data is an inherent part of the considerations during the structuring of a surveillance scheme.

Evaluation. Before the test results of a health surveillance examination can be interpreted properly, it should be determined whether the collected data are appropriate and reliable with respect to health.

It is essential that relevant and proper methods be used with appropriate sensitivity and specificity, and that they be performed with correct sampling, storage, and analysis. For this reason some method of quality control for accuracy and precision should be incorporated [e.g., good laboratory practice (GLP) or round robin tests].

For intra- and interindividual comparison with previous test results, adjustment of data may be indicated. Thus, data from urine samples have to be adjusted for specific gravity or creatinine excretion.

Another consideration is whether – and if so which – intrinsic and extrinsic factors could have confounded the results. These factors can include smoking, drinking, diet, medication, and others.

For correct evaluation and interpretation proper choice of reference data is essential. The reference data used depend on the objectives of the HS program and the conclusions to be achieved: Do we have to make comparisons with figures for the general national or local population? Will we draw conclusions for groups of workers or individuals? Do we want a cross-sectional impression of the current overall health state of a group of workers or a longitudinal view of the progression of specific parameters throughout the years for individuals or groups of workers? In longitudinal studies, the current data can be compared with results from previous examinations. For individual workers, reference data from the pre-employment examination may be of value. When no reliable reference data are available, the best estimate should be made from epidemiological and animal studies (see also Section 4.5).

Reference data for physical (pre-)clinical symptoms and signs of impaired health depend on the knowledge and skills of the physician who performed the periodical medical examinations. Data from laboratory tests for estimating the state of health depend on the methods used and the specificity and their sensitivity.

As HS is meant to determine preclinical reversible signs in healthy people before these

must be considered adverse, in general the laboratory tests have to be more sensitive and the reference data more subtle than those used in hospital for patients with impaired health.

For a number of chemicals, occupational threshold limit values for exposure are set. For a limited number of chemicals biological limit values for uptake have also been established. In the United States biological exposure indices (BEI) are used, and in Germany Biologische Arbeitsstofftoleranzwerte (BAT values). For details, see Sections 3.6 and 4.2.

To ensure that all essentials are considered, the use of a checklist for the evaluation of health surveillance data is helpful:

Methods and sampling
- appropriate (sensitivity, specificity) ?
- correctly performed?

Adjustment for correct interpretation

Confounding factors?
- intrinsic (age, enzyme deficiencies, etc.)
- extrinsic (life-style, pharmaceuticals, etc.)

Reference data available?
- individual
- group
- general population

Occupational limit values available?
- guideline/advisory
- legal/obligatory

Interpretation. Once the data have been evaluated it has to be established whether the HS results indicate impairment of health.

Universally fixed "normal" values do not exist, because geographical, ethnic, age, and gender differences may be apparent. Even between individuals within a quite homogeneous group of people and, even within the same individual, fluctuations can be found, due to the range of physiological variability. Differences in basic body burden of individuals may be due to extrinsic factors such as life-style (eating, smoking, etc.) and leisure activities.

With respect to pre- clinical reversible biological effects, it has to be considered whether these are medically acceptable or not. This is completely a matter of professional knowledge and skills. The final decision however, on what is tolerable, is not only a medical but also a political issue and will include a balance between technical and economical feasiblity and acceptability by society. This can differ from country to country and between social groups.

4.1.4. Reporting Health Surveillance Results

After completion of the health surveillance program, it has to be decided to whom, when, and which information about the results will be given.

The workers involved as well as the responsible management must be provided with the relevant information.

With respect to BM and BEM data, it should be emphasized that these monitoring methods are person-related. This implies that individual workers have the right to be informed of their own personal test results. To enable the worker to put these data into perspective and to understand the uses and limitations of BM, some basic information on the dose – effect relation, reference data, etc., should be provided as well. The individual data must be considered as confidential and only accessible to the worker and physician involved. They are only to be provided to workers' representives or management with the explicit consent of the individual worker.

The employer is given an anonymized compilation of the group data, to allow him to judge the working conditions and the effectiveness of preventive measures already taken.

4.1.5. Corrective Actions

The results of health surveillance indicate whether corrective preventive actions are necessary or advisable.

These actions may be restricted to individuals in a group who showed aberrant and unwanted signs of chemical overload or a health threat, compared with the group mean and/or the average in a group of nonexposed people.

It must be investigated why these individuals differ from the others: are they at risk due to higher exposure or uptake, or are they subject to individual hypersensitivity?

Individuals with abnormal HS results should be temporarily withdrawn from their work until the figures are back to normal. Sometimes, permanent withdrawal from specific work conditions has to be considered; for instance, after (pre-)clinical signs or symptoms of asthma due

to isocyanate exposure or when skin conditions due to a susceptibility for sensitizing agents have been diagnosed.

Moreover, certain medical conditions unrelated to work may be incompatible with certain job activities. Restrictive and obstructive lung diseases prohibit exposure to irritating gases, vapors, and dusts. Exposure to a number of chemicals is contraindicated for workers with functional liver and kidney disorders.

When the group mean of the examined workers appears too high compared with nonexposed people, or higher than the group mean in a previous examination, measures at the job or working environment level should be considered.

A critical survey of current work systems and practises can often lead to improvements which are simple and cheap to achieve and lead to reduction of emissions, exposures, and uptake.

As has been mentioned before, TLV values may be not always protect every worker from adverse health effects under all circumstances. Therefore, for each group of workers and for each specific working condition, a critical appraisal should be made about the time-weighted average concentration of chemicals that can be considered to be safe in a defined situation. This holds too for chemicals that are not included in official lists of TLV values.

A reduction of emission can be achieved by substitution of potentially toxic chemicals with less hazardous ones, or modification of manufacturing processes, technology, and installations. Also changes in procedures and organization can improve the working conditions.

The intensity of uptake can be reduced by personal protective equipment (see Section 3.7.2).

4.2. Biological Monitoring [33–38]

Biological monitoring (BM) is one of the methods for the health surveillance of individual workers.

BM can be described as: "Measurement and assessment of workplace agents or their metabolites either in tissues, secreta, excreta or any combination of these, to evaluate exposure and health risk, compared to an appropriate risk" [33].

4.2.1. Basic Considerations

BM is a person-related specific way of monitoring to estimate the health risk due to exposure to a particular chemical or group of chemicals. The aim is to obtain a measure of the internal burden of a defined chemical in the body of an individual after absorption, irrespective of the route, time, or duration of intake. BM is especially important for chemicals that are readily absorbed through intact skin, such as benzene, acrylamide, dimethylformamide, dinitro-*o*- cresol, and methyl parathion.

With respect to health surveillance, BM is far more relevant than environmental monitoring, which only reflects the possible exposure by inhalation during work.

When appropriate methods are available, BM is always the method of choice for the health surveillance of individual workers, because it reflects the individual uptake by all routes and is almost always substance-specific.

Thanks to steady progress in knowledge about the biological fate of chemicals in the human body and improved specific and sensitive analytical methods, there is a growing number of chemicals for which BM methods have become available.

Although a specific response is measured by means of BM, it is not a reflection of the way the chemical has entered the body. Mercury in the urine can be caused by ingestion of metallic mercury, inhalation of mercury vapor, or transdermal and transmucosal absorption of an organomercury compound. Carbon monoxide, determined in the blood, might be due to smoking, gas fumes in the home due to poor ventilation, occupational exposure to carbon monoxide in a blast-furnace plant, or stripping paints with methylene chloride. Hippuric acid in the urine might be due to the metabolism of toluene taken up during work or to digestion of various food constituents.

As BM always reflects a compilation of past exposures, it is generally not suitable for determining peak exposures during work. It also will give no indication of contact of skin or mucous membranes with chemicals that are not absorbed through them, although they may exhibit local effects.

Moreover BM results must be evaluated with care. Figures that exceed certain reference val-

ues may not necessarily be considered as harmful to health and therefore unacceptable. And levels below reference values may not always be acceptable for every individual or in all circumstances.

4.2.2. Prerequisites

BM can only be used properly when the following conditions have been met:

1) Before any measuring strategy can be designed the aims of health surveillance must be defined. For example: Are we interested in current or past exposures? Do we want individual figures or group means? Only when these conditions have been met is the design of a proper and sensible BM scheme possible. In general this means that in a BM program adapted to the conditions chosen, the most suitable parameter should be measured in the most suitable biological medium in the right way at the right moment in the right population.
   In general, determination of concentrations in blood better mimics current exposure, whereas urine samples give a better indication of past exposures.
2) As the parameters to be measured are intended to the way the human body deals with the chemical involved, relevant knowledge of pharmacokinetics, uptake, distribution and the route and rate of elimination is essential.
3) To consider possible health effects additional information about a quantifiable relationship between the parameters to be measured and apparent clinical or biological effects is needed.
4) Appropriate and reliable analytical methods must be available, and must meet the requirements for sensitivity, specificity and reproducibility.
5) The considerations for a proper BM surveillance scheme are the same as for other methods of health surveillance (see Section 4.1).

4.2.3. Biological Media

In principle, various media can be used:

1) Body fluids such as whole blood, plasma, and serum
2) Body tissues, e.g., hairs, nails, teeth, fat tissues, placenta
3) Excreta: urine, sweat, saliva, feces, expiratory air, mother milk.

However, the choice of biological medium depends upon a number of considerations, including aims, feasibility, availability of proper analytical methods, specificity, sensitivity, and cost.

Urine and blood are still the prevalent media in common practice. The advantages and disadvantages of the most frequently used media are discussed below:

Urine is the most suitable and easy obtainable test medium for water-soluble chemicals or metabolites that are excreted in the urine. It can give an impression of the average uptake in the past.

A drawback of the determination of chemicals in the urine is that the quantity and composition of the urine varies over a 24-h period, as does the content of the chemical to be determined. Sampling over a 24-h period overcomes this disadvantage to some extent, but it is technically difficult to achieve. For correct interpretation of the analysis of one single sample a correction for specific gravity or creatinine excretion is indicated.

Blood is the most important means of transport and distribution for chemicals in the body. Biomarkers determined in blood reflect the circulating quantity of the chemical itself or of a metabolite. Based on the half-life of a chemical in the body, the data give an impression of recent uptakes, and sometimes peak exposures can even be distinguished.

Some chemicals can be found in the red blood cells (e.g., chromium) whilst others are bound to plasma proteins. Chemicals such as carbon monoxide, hydrogen sulfide, cyanides, and nitrites form complexes with hemoglobin.

Sampling of blood by venopuncture may often act as an emotional barrier for people to cooperate.

Expiratory air is especially useful for volatile compounds which are metabolized slowly, are poorly soluble in body fluids, and are largely exhaled. Sampling and analysis of expiratory air is often simple. The interpretation is complicated by inter-individual differences in respiratory patterns.

Analysis of chemicals in *faeces* is suitable for chemicals that are taken up orally or are excreted into the intestines via the biliary system. The interpretation with respect to exposure or contact and to health is difficult or even impossible.

For most chemicals, determination in hair, nails, saliva, mother milk, etc. does not provide additional useful information with respect to health risks for exposed people. Determination of chemicals in mother milk may give an indication of the amount that could be taken up by breast-fed babies.

4.2.4. Biological Occupational Limits

For a number of chemicals, reference values are available for some relevant biomarkers. Until now, only in Germany and the United States are lists of this type of limit value published. The figures may differ between these two countries due to differences in definition.

In the United States biological exposure indices (BEI) are published by the American Conference of Governmental Industrial Hygienists [34]. These indices are intended as guidelines for evaluation of potential health hazards. They "represent the levels of determinants which are most likely to be observed in specimens collected from a healthy worker who has been exposed to the same extent as a worker with inhalation exposure to the TLV."

The German BAT value (tolerated biological value for occupational exposure to chemicals) refers to the maximal concentration of a chemical or metabolite which, based on current knowledge, does not impair the health of individual workers [35].

For the current values refer to the lists mentioned in the literature reference list. In Table 1 parameters for BM and BEM are given for some chemicals.

Except for inorganic lead, no biological exposure limits are set as legal standards. Until now they only serve as guidelines and recommended reference data in the health surveillance of workers.

For inorganic lead, a CEC Directive was adopted in 1982, which all Member States are committed to incorporate in the national legislation.

The Directive sets action levels and limits both for lead in air and for biological indicators.

Based on the CEC Directive, a resolution on lead ("Loodbesluit", 1988) came into force in The Netherlands with the same action levels and limits. However, the Dutch Expert Commission for Occupational Standards recommended a biological acceptable limit of 400 µg lead/L blood for men and 300 µg lead/L blood for women.

For health surveillance of workers exposed to lead, it is recommended to start with a simple determination of ZPP (zinc protoporphyrin) in blood with a hematofluorometer. When the ZPP content of blood exceeds 20 µg/g hemoglobin, additionally PbB should be determined.

Action levels:
PbB > 400 µg/L $\longrightarrow$ information of workers
PbB > 500 µg/L $\longrightarrow$ full application of the directive, including obligation for periodical medical examination
Limit values:
PbB max. 700 µg/L
PbB 700 – 800 µg/L is acceptable when:
ALAU > 20 mg/g creatinine
ZPP > 20 µg/g hemoglobine or,
ALAD < 6 Eur. Units

where PbB = lead in blood, ALAU = δ-amino-levulinic acid in urine, ZPP = zinc protoporphyrin in blood, ALAD = δ-aminolevulinic acid dehydrogenase in blood.

4.2.5. Interpretation of Biological Monitoring Data

From the point of view of health, BM is much more relevant than EM. Even when the relation between BM and EM results is fairly well known on a group basis, this relationship may differ significantly on an individual basis.

Therefore, BM programs should be performed under the supervision of an (occupational) physician. Interpretation of BM data for groups or individuals should be based on medical-toxicological expertise. The physicians involved must have the expertise to evaluate the data with respect to intrinsic and extrinsic individual factors which might explain interindividual variability, in order to come to the correct conclusion about the state of health of the individuals under surveillance.

Table 1. Biological monitoring and biological effect monitoring for some chemicals

Compound	Biological monitoring		BEI∗∗	BAT∗∗	Biological effect monitoring	
	Parameter	Medium∗			Parameter	Medium∗
Acetone	acetone	U				
	formic acid	U				
	acetone	B				
	acetone	E				
Acrylonitrile	thiocyanate	B				
	acrylonitrile	B				
Aniline	*p*-aminophenol	U	X		methemoglobin	B
	aniline	U		X		
Antimony	antimony	U				
Arsenic	arsenic	U				
Benzene	benzene	B			leucocytes	B
	benzene	E	X		hemoglobin	B
	phenol	B	X			
Bromine (salts)	bromide	U				
Butylglycidyl ether	butoxyacetic acid	U				
Cadmium	cadmium	U	X		β-microglobulin	U
	cadmium	B	X			
Chloroform	chloroform	E				
	chloroform	B				
Chromium	chromium	U	X			
Cyanides	thiocyanate	U			cyanohemoglobin	B
	thiocyanate	B				
	cyanide	B				
2-Ethoxyethanol	2-ethoxyacetic acid	U		X		
Ethylbenzene	mandelic acid	U	X			
	ethylbenzene	E	X			
Ethylene glycol	oxalic acid	U				
Fluorides	fluoride	U	X	X		
Formaldehyde	formic acid	U				
	formaldehyde	B				
n-Hexane	2,5-hexanedion	U	X	X	electromyography	
Hydrogen cyanide	thiocyanate	U			cyanohemoglobin	B
	thiocyanate	B				
	cyanide	B				
Isopropylbenzene	2-phenyl-2-propanol	U				
Lead	lead	U	X	X	zinc protoporphyrine	B
	lead	B	X		δ-aminolevulinic acid	U
Manganese	manganese	U				
	manganese	B				
Mercury	mercury	U	X	X	renal functions	
	mercury	B	X			
Methanol	methanol	U	X	X		
	formic acid	U	X			
Methylchloroform	methylchloroform	E	X			
	trichloroacetic acid	U	X			
	trichloroethanol	U	X			
	trichloroethanol	B	X			
Methylene chloride	methylene chloride	B		X	CO – hemoglobin	B
Methyl ethyl ketone	methyl ethyl ketone	U	X			
Monochlorobenzene	4-chlorocatechol	U	X	X		
	p-chlorophenol	U	X	X		
Nitrates					methemoglobin	B
Nitrites					methemoglobin	B

Table 1. (Continued)

Compound	Biological monitoring		BEI∗∗	BAT∗∗	Biological effect monitoring	
	Parameter	Medium∗			Parameter	Medium∗
Nitrobenzene	*p*-nitrophenol	U	X		methemoglobin	B
	aniline	B		X	methemoglobin	B
Organophosphates					cholinesterase	B
Parathion	*p*-nitrophenol	U	X	X	cholinesterase	B
Pentachlorophenol	pentachlorophenol	U	X			
Phenol	phenol	U	X	X		
Pyrene	hydroxypyrene	U				
	1-pyrenol	U				
Styrene	mandelic acid	U	X	X		
	styrene	B	X			
	styrene	E				
	hippuric acid	U				
	phenylglyoxylic acid	U	X			
Tetrachloroethene	tetrachloroethene	E	X			
	tetrachloroethene	B	X	X		
	trichloroacetic acid	U	X			
Toluene	hippuric acid	U	X			
	toluene	E	X			
	toluene	B	X	X		
1,1,1-Trichloroethane	trichloroacetic acid + trichloroethanol	U				
Trichloroethene	trichloroacetic acid	U		X		
	trichloroethene	E				
	trichloroethanol	B	X	X		
Xylene	methylhippuric acid	U	X	X		
	xylene	B		X		
Carbon monoxide	carbon monoxide	E	X		CO−hemoglobin	B
	carbon monoxide	B	X	X		
Carbon disulfide	thiothiazolidine-4-carboxylic acid	U	X			

∗ B = Blood; U = Urine; E = Expired air. ∗∗ BEI = Biological exposure index (ACGIH); BAT = Biologische Arbeitsstofftoleranz.

Among other things, inter-individual differences in enzyme configurations can influence the possible health effects of chemicals. For instance, due to a difference in biotransformation, people with inherited low acetylating enzyme profiles ("slow acetylators") are at greater risk of intoxication with methemoglobin forming compounds such as aniline or nitrobenzene than those who have large amounts of the acetylating enzyme.

Note, that BM cannot always replace EM, and sometimes a combination of both methods is indicated. Especially when the test results from BM exceed the reference data or recommended biological occupational limits, EM is indicated to estimate whether air levels at the work place are responsible, or other work-related factors such as poor hygiene at work, or endogenic or nonoccupational exogenic factors must be considered.

This question is particularly important for the decision whether corrective changes in the working conditions must be made or not. BM may also serve as a check for the effectivity of measures taken to reduce the exposure of workers.

4.3. Occupational Epidemiology (General references: [39–43])

4.3.1. Introduction

Going back to ancient Greek roots, epidemiology can be defined as the science (*logos*) of what comes upon (*epi*) the people (*demos*).

In modern terms it should be understood as the study of the occurrence and patterns of disease in groups of people. Whereas daily occupational medical routine looks at individual persons, occupational expidemiology is concerned with groups or cohorts. The basic requirements

for any analytical epidemiological study are careful collection of relevant data, appropriate statistical analysis, and qualified interpretation of the results. The type of study chosen to answer a particular question should not only suit the size of the group under investigation but also the quality and quantity of data that can be obtained. Although occupational epidemiology can be regarded as one of several scientific instruments for the assessment of a particular occupational health hazard, it must be remembered that no epidemiological study can prove or disprove cause and effect relationships. How close to the "truth" a particular study will come largely depends on the quality of the study (i.e., the quality of the data), on careful performance of the study, and on expert interpretation of the results. The results should be interpreted in the context of the outcome of similar studies. Successful occupational epidemiological studies demonstrate the magnitude of risk associated with a particular level of exposure to a particular agent at the workplace.

4.3.2. Basic Requirements – Data Collection

4.3.2.1. Data on Exposed Individuals in the Study Group

For occupational cohort studies, data are needed to identify individuals in the cohort (e.g., full name, sex, ethnic origin, date of birth, personnel or social security number).

Also needed are data that connect these individuals to the cohort, such as date of entry into the company or plant where the individual is employed during the period of study, and where applicable, also the date of leaving employment.

For mortality studies the date of death or the vital status on the closing date of the study should be known. Also the cause of death must be known and it should be stated from which source (e.g., death certificate, autopsy report, etc.) this information was obtained.

For morbidity studies (concerning persons suffering from a disease) the date on which the key diagnosis was made should be known, as well as its diagnostic certainty (e.g., in the case of cancer whether the diagnosis was made by clinical means only or by histology of surgical material).

If other, confounding factors influence the occurrence of a disease (e.g., smoking in the case of lung cancer) then they must also be considered.

During data collection, it is essential to link personal data to occupational or work-environment data if a study on a specified occupational hazard is to make any sense. For subsequent statistical evaluation and publication, the data can be anonymized without any loss of quality to the study.

In many countries restrictive laws on medical confidentiality and/or data protection severely hamper the conduction of sound epidemiological studies.

Data may be very difficult to obtain when studies deal with exposures in the past because personnel files or occupational medical records may have been destroyed. In case of foreign labor many cases may be lost to the study because of workers having returned home.

In many countries death certificates are not available for occupational epidemiological studies; therefore, data on cause of death must be obtained by the best available means. If the study population is compared with the general population of the country by using the official death statistics, which are based on death certificates, then this differing source of information must be taken into consideration, as it can influence the result of the study.

For morbidity studies (e.g., the occurrence of certain types of cancers) the obstacles can even be greater. Fully functioning cancer registries, as well as national population registries, are often lacking. Some cancer registries contain little or no information on the occupational history of the cases or only include the last occupation. Even if all cancer cases were known in a particular population under study the comparison with another (e.g., the general population) would be difficult if not impossible. Thus it is not only important to very carefully collect data but also to validate the source from which they are obtained.

4.3.2.2. Technical Data on the Occupational Environment

The occupational environmental exposure of an individual worker stems from the local environment (the workplace), the activity performed

(the job), and the factors present such as chemicals or physical agents. It is thus essential that information on these items be available. For studies concerned with past exposure (retrospective studies) this may be very difficult. For example, the time between the beginning of exposure to a bladder carcinogen such as 2-naphthylamine and the development of a bladder tumor (the latency period) may be as long as 20–40 years. Quantitative exposure data may not be available after such a long period, or may never have existed. The only source of information may be the detailed occupational history taken from those persons who are still available for questioning by the occupational physician in charge of the study. In some cases an intelligent guess can be made with the help of the industrial hygienist about the exposure level.

For studies concerned with future exposure (prospective studies) occupational exposure recording systems should be established in production plants where a potential health hazard is present or suspected and should be linkable to the individuals working in that particular plant or production site. Two types of exposure data can be used: exposure information relating to the workplace or to the job. For some chemicals biological monitoring yields a third type of exposure data specific for individual workers.

Information on the Workplace. Descriptions of industrial processes and working conditions should be available and should be updated when changes occur. Lists of chemicals or raw materials should be available in an accessible and comparable form, indicating substances present by means of appropriately defined workplace codes (e.g., by the CAS code).

Records should be kept on the quantities used and produced. Exposure measurements at the workplace (ambient monitoring) should be based on standardized sampling and analytical methods, backed by quality control programs to confirm the validity of the data produced. Results of personal monitoring should be stored with cross-references to the individual concerned to permit dose – effect and/or case studies to be performed.

Information on the Job. Job exposure information can be obtained from data representative of the activity performed (job) as opposed to that of a particular workplace. For this an appropriate description of the job is necessary (job exposure profile), which should be exposure orientated. Wherever possible the job description should be complemented with job-specific exposure measurements. Problems arise when a worker cannot be assigned to one specific workplace (e.g., working in different parts of the production or doing maintenance work), when mixtures of several chemical substances are handled, or production and products change very frequently.

Biological monitoring gives information on individual exposure (inhalatory and skin absorption), and helps to identify exposed workers.

4.3.3. Descriptive and Analytical Study Types

Descriptive studies are concerned with the observation of the distribution and the progress of a disease in groups of exposed persons. They are limited to the determination of incidence (number of new cases occurring per unit time relative to the number of persons at risk), prevalence (the number of existing cases per unit population relative to the number of persons in the population at a certain point of time), or mortality rates for certain diseases.

Analytical studies are concerned with a hypothesis about causal (etiological) factors of the disease, developed from a descriptive study or from individual occupational medical observations. The purpose of an analytical study is to determine an association between a suspected causal or risk factor (e.g., a carcinogen) and an effect (e.g., the occurrence of a specific cancer). Under favorable circumstances the strength of this association can be quantified (the "power" of the study).

The several types of analytical studies are described below.

Cross-Sectional Studies. Cross-sectional (also called prevalence) studies concentrate on a defined population at a particular point in time. They are set up to determine prevalences of disease in the exposed and nonexposed groups. Cross-sectional studies are relatively simple and quick to carry out and may be used for preliminary testing of a hypothesis. But since they

merely examine new cases or drop outs they give rise to a selection effect.

Case Control Studies. In case control studies two groups of individuals are selected: the cases (suffering from the disease) and the controls (not suffering from the disease). The two groups are then compared with respect to present or past characteristics thought to be of possible relevance to the etiology of the disease. This type of study is always retrospective, i.e., the effects are already present at the beginning of the study. The conditions of exposure must therefore be reconstructed from the past.

A case control study has the advantage of being easily carried out in a relatively short time and does not need large numbers. Indeed, for the study of rare diseases this may be the only useful method. The great disadvantage, however, is the difficulty of collecting past data for both cases and controls. Interviewing cases in the knowledge that they are suffering from the disease may bias data collection, and it may be difficult to find appropriate controls.

Cohort Studies. The cohort study compares the incidence of a disease in two or more groups (cohorts) which are observed in parallel for a given period of time. An exposed group is composed of individuals with known exposure to a suspected risk factor. The control group is not exposed, but otherwise comparable for as many criteria as possible.

This type of study is also called longitudinal. It can be carried out in a prospective way, i.e., starting with apparently healthy individuals and observing them for a defined period. It can also be carried out in retrospective (historical) way, i.e., a study date is set from which the observation is retrospectively carried out to a starting date. For both types of cohort studies standardized rates of mortality or morbidity are used and compared with standardized rates of a suitable reference population.

Prospective studies have the great advantage of being more reliable and of better validity than retrospective studies, but they may take a very long time and be very costly. Retrospective studies can yield a much quicker result and cost less but often are less reliable and of lower validity. The "power" of cohort studies tends to be dependent on fairly large numbers, particularly if the disease searched for has a low incidence.

Intervention studies are concerned with the observation of an effect on a population, e.g., by reducing a certain exposure, or by altering any other relevant factor.

4.3.4. Usefulness of Occupational Epidemiological Studies

An occupational epidemiological study may be initiated to describe the health status of a particular work population. It may also be used to analyze possible associations of a disease with the place of work or the work environment. It could be used to answer specific questions resulting from findings in animal toxicological studies (mutagenicity, carcinogenicity, etc.), or to assess the validity of existing or planned limit or threshold values for certain substances at the place of work or in the work environment. Under favorable circumstances occupational epidemiological studies can even act as a monitor for the efficiency of protective measures. Occupational medical questions that might be answered by epidemiology are:

What is the morbidity or mortality for workers exposed to a particular substance or agent?

Is there an elevated risk of morbidity or mortality due to a particular disease?

What working history is found for workers who suffer from a particular disease?

What is the significance of certain abnormal laboratory or biological monitoring findings in certain workers, with, for example, elevated urine or blood levels of a particular substance or its metabolites?

The interpretation of an occupational epidemiological study eventually has to be compared with experimental animal and laboratory data, which often may give more precise and reproducible data on dose and effect. Since epidemiology is a descriptive science it concerns itself with humans and helps to evaluate possible effects at exposure levels that are actually experienced in the work environment.

But any occupational epidemiological study can only be as good as the basic data used for it. The interpretation of an epidemiological study is rarely a simple task and should be made within

the epidemiological framework, taking into consideration the findings of similar studies. Therefore, so- called negative studies should also be published whenever possible.

4.4. Ergonomics

4.4.1. Introduction; Definitions

The physiology of work is based on the principles, experiences, and findings of applied physiology. It is the systematic study of the interactions between humans and their work. In particular, the effects of various forms of human work on individual body functions or on the entire human organism are analyzed. The factors studied include time, duration, hardness, difficulties, amount, and monotony of work as well as influence of the surroundings. These effects are then compared with the prerequisites for the performance of the working person.

The performance prerequisites can be differentiated according to the capabilities as structural characteristics of the person:

1) Prerequisites for physical performance, general condition, constitution, acquired physical and mental prerequisites, such as knowledge, experience, skills, and degree of practice and training
2) Prerequisites relating to work, such as powers of vision and hearing, dexterity, etc.
3) Willingness to work as driving and guiding characteristics: advancement orientation, intro- and extroversion, perseverance, satisfaction of needs, striving for a sense of achievement, varying physiological readiness to work as a result of the circadian rhythm, etc. [44], [45]

The goal of the physiology of work is to acquire scientific principles, findings, and rules in order to adapt the various forms of human work to the capabilities of the working person in the sense of preventing strain-induced health damage.

Ergonomics is regarded as an interdisciplinary core area of industrial science. A broad definition is given in [46]: "Industrial science deals with the analysis and organization of work systems and working facilities. The starting point and the goal of this study are the individual and social relationships of the working person to the other elements of the work system."

Therefore, industrial science is the science of:

1) Human work, especially from the standpoint of cooperation with others and interaction with working facilities or objects
2) The prerequisites and conditions under which the work is carried out
3) The consequences and effects of work on people, their behavior, and hence their productivity
4) The factors by which work, its conditions, and effects can be influenced in a humane fashion

Thus, the organization of work according to knowledge of industrial science includes all measures by which the system human and work can be humanely influenced.

These diverse and multifaceted questions can be solved only by the cooperation of the various scientific fields, by using findings based on human work. These fields include:

1) Medicine, especially physiological, hygienic, and pathological aspects
2) Social sciences, especially psychology, sociology, and education
3) Technical sciences
4) Economics and jurisprudence

The terms ergonomics and industrial science are sometimes used synonymously because clear and generally recognized boundaries do not exist, and opinions on the definition of "ergonomics" differ.

However, it can be stated that ergonomics is an integral part of industrial science without completely encompassing the entire field. Numerous problems of social science and sociology and a series of medical areas are not addressed by ergonomics. On the whole, industrial science is regarded as a pure science and ergonomics is more of an applied science.

Ergonomics is increasingly becoming a human- centered systematic science. The results are based on humans and the constructive consequences for technology are also derived from people [47]. The term ergonomics is derived from the Greek *ergon* (work, act) and *nomos* (law, rule, teaching).

The term ergonomics has spread internationally, largely due to the journals *Ergonomics* (since 1957) and *Applied Ergonomics* (since 1969) [48], [49].

The International Ergonomics Association (IEA) was founded in 1959 and represents about 20 national associations [50]. Today, ergonomics includes various branches of industrial science, e.g., anatomic, physiologic, psychologic, medical, and scientific aspects.

Ergonomics has increasingly opened up different spheres of activity and areas of application, including: household articles, traffic including safety aspects, sport, and leisure activities. Only areas of ergonomics that are relevant to industrial medicine are discussed here.

4.4.2. Scope of Ergonomics

The following focal points of ergonomics are worth mentioning:

1) Analysis, measurement, and assessment of human – machine/human – work systems and the determination of the limits of overworking and underworking. In this connection, the stress – strain concept plays an important role.
2) Organization of human – machine/human – work systems with the purpose of adapting the work to the person. This area includes the ergonomic planning and construction of the workplace, working facilities, working surroundings, and hours and structure of work.
3) In an inter-individually and intra-individually varying range (variable time-dependent prerequisites for performance of an individual person), the adaptation of the person to the work is also possible by means of education, practice, habit, training, and rehabilitation measures.

4.4.3. Stress Factors

The stress – strain concept of ergonomics mentioned above includes a large number of influencing factors which act on the working person during the work under investigation. The following factors can contribute to the burden of the working person.

Workplace:	Spatial design, visual conditions, acting forces (e.g., when lifting and carrying loads)
Working facilities:	Indicators and signals, screens, working chairs, handles and actuators, working aids and tools
Work surroundings:	Sounds, vibrations, light (e.g., lighting), and air (e.g., ventilation)
Working hours:	Shift system, flexibility, regulation of breaks
Work structure:	Work content (e.g., enrichment of work) distribution of work (e.g., group work)

The above sequence corresponds to experience gained in production in the processing industries, predominantly in mechanical engineering, the motor industry, electrical engineering, and precision engineering. However, it can also be applied to other areas of work, such as the chemical industry. It has proved to be a suitable basis for examining and developing the main aspects for the ergonomic organization of work [50].

4.4.4. Assessment and Testing Criteria

According to [51], prior to a detailed examination, the following criteria and questions can be used for the assessment of working conditions as a prerequisite for the organization of work.

1st	Criterion:	Feasibility
	Question:	Can the planned work be carried out with the designated persons?
2nd	Criterion:	Endurability
	Question:	Is the work on a permanent basis endurable for the designated persons?
3rd	Criterion:	Reasonableness
	Question:	Are the work and the expected working conditions reasonable?
4th	Criterion:	Satisfaction
	Question:	Will the persons designated for the work be satisfied with the work and the working conditions to be expected [51] ?

An examination of the above criteria is helpful in assessing whether all persons designated for the work are capable of carrying out the specific task planned.

For instance, a certain task cannot be performed if it involves work that must be performed beyond arms' reach. If the strength required exeeds that of the designated persons, this

can be a reason for a task being not performable [50].

The criterion bearableness includes the term "strain". Here, the effect of a specific task on humans is examined. For instance, a given stress does not produce the same strain in all persons.

The demand made on the particular worker depends on individual capabilities, which can vary widely due to, e.g., age, constitution, sex, personal abilities, and skills. Consequently, strain is the individual, psychophysical reaction of the organism to a stress [52]. For the assessment of individual strain, evaluation criteria from industrial medicine are of considerable importance. This applies in particular to handicapped employees. ROHMERT describes strain as a varying degree of exhaustion of individually different capabilities by a given stress [53].

Strain can be technically measured by physiological means, e.g., cardiac rate, changes in blood pressure, body temperature, electromyogram, electrooculogram, or by psychophysiological means, e.g., flicker fusion frequency or selection reaction test. Psychological and sociological strain can also be detected by widening the stress – strain concept, e.g., by means of a questionnaire, interview, strain scaling, and group observation.

This can be explained with the help of a few examples:

Stress by lifting and carrying of weights will lead to strain of the musculo-skeletal and the cardio-vascular systems. The intensity of strain depends on, e.g., posture, individual muscular power, quality of cartilage and intervertebral tissue, the axial conditions of the strained section of the skeleton, e.g., post traumatic faulty alignment, individual patterns of movement, movements required by the specific task (e.g., frequent overhead movements), and on the individual cardio-vascular performance.

The thermoregulation system is strained by the stress resulting from the action of heat when emission of body heat is obstructed, e.g., by working in fully protective clothes, or heat is supplied to the body from the outside. Other factors to be taken into account include the heaviness of the work, working time under the action of heat, the interval between and duration of cooling breaks, and the provision of special clothing. Moreover, the intensity of strain in this example depends on the degree of acclimatization to the working conditions.

A control and supervisory task can require high attention and concentration, especially if the product quality essentially depends on this work. The intensity of strain can depend on a series of individual performance prerequisites, e.g., reactions, experience, routine with regard to the control and supervisory task, and motivation (e.g., avoidance of standstill periods and achievement of the best possible quality).

In the sense of motivation or inhibiton, certain factors can be linked to individual abilities and skills. These factors include working conditions resulting from the subordination or dependency of workers, conditions resulting from the organization of the company, working atmosphere, wages and social benefits, and social and company systems of standardization, control, and sanctioning.

Negative social relationships can represent an additional stress which, because of the resulting strain, can possibly lead to the overtaxing of the individual [52], [54].

In the case of screen work, strain can depend on the amount of work, arrangement and design of the keyboard, the office chair, the quality of the screen, and glare in the field of vision. Here, the individual factors can be further differentiated from the standpoint of ergonomics. For instance, increased demand on the eyes due to impeded focusing (accommodation) can be caused by low resolution of the characters, insufficient contrast between the characters and the background, instability of the characters, reflections on the glass surface of the display, and unsuitable design of characters and typeface [55].

The spatial design of a computer-terminal workplace must take into account the variations of body size. To optimize the design of the workplace, a graphic scale model of the workplace is used into which the schematic depiction of human figures with the different expected body sizes can be inserted. This method is known as somatography [56]. In the meantime, somatographic models and computer aided methods are available, making perspective representations of the spatial situation possible [50].

In many cases, ergonomic findings have already been embodied in law, e.g., the decree and guidelines on workplaces, acknowledged rules of technology in numerous clauses of DIN and

VDI, guidelines of the EC Council on the minimum regulations with regard to safety and health protection, e.g., for working with visual display units and for manual handling of loads (see Sections 2.1, 2.2, 2.3, and 2.4).

The application of ergonomic rules in industrial medicine requires ergonomic assessment of the working conditions and medical assessment of the working person.

To assess working conditions, numerous methods for the analysis of work and stress can be used. These include ergonomic test lists for the evaluation of the state of organization [54], questionnaires for the ergonomic description of working conditions [57], and representations in text books and monographs [47,54,55,58–65]. Moreover, the legally embodied regulations mentioned above must be considered. To assess the working person, industrial medical criteria are to be used. These must be supplemented by the application of biomechanical, neurophysiological, and functional anatomic principles as well as physiological and psychological findings.

4.4.5. Objectives

In health protection at the workplace, ergonomics supports the concept of prevention and helps avoid work-related illness and disease [66].

Attempts should be made to replace corrective ergonomics (subsequent consideration of ergonomic findings and industrial scientific evaluation criteria when, e.g., ergonomic findings have been violated in the planning process) by the concept of prospective ergonomics.

In this concept, Laurig [66] summarizes two meanings of the term "prospective":

1) The possible prediction of future effects of working conditions and their assessment (e.g., the prediction of the limits for the endurability of the effects of vibration)
2) The application of ergonomic findings to the improvement of work productivity (e.g., the design of lighting conditions to improve quality)

In principle, for the application of the concept of prospective ergonomics, it is true that both ergonomic and industrial-medical organizational objectives must always be established in connection with the definition of production engineering planning objectives. In this process, the regulations of public occupational safety and of the trade association must be taken into account (see Sections 2.1, 2.2, 2.3, and 2.4).

Consequently, this priority can also result in restrictions on decisions preferred from a production engineering standpoint if these cannot be made compatible with the criterion of harmlessness.

Therefore, ergonomic findings should always be taken into account in the planning of human – machine/human – work systems. Only in this way can it be guaranteed that a workplace need not to be changed after it has been set up [50].

4.5. Evaluation of Health Risks [67–72]

The American National Academy of Sciences has defined health evaluation of persons exposed to chemical compounds as:

"The scientific activity of evaluating the toxic properties of a chemical and the condition of human exposure to it in order both to ascertain the likelihood that exposed humans will be adversely affected, and to characterize the nature of the effect they may experience" (1982).

Before the health risk of a particular task or exposure can be evaluated, terms such as "health" and "risk" must be defined, to arrive at a reliable balance and conclusion.

4.5.1. Basic Considerations

The state of health of an individual is in no way a static entity. It is a dynamic balance between the inborn potential of the human body and external, environmental factors (i.e., the balance between capacity and burden of an individual). Every biological alteration may disturb or change the balance. For instance, age is a biological factor that continuously influences the capacity of an individual.

A second problem in the evaluation of health risks is the meaning of the phrase "risk". Risk is a combination of the chance and nature of health impairment due to exogenic factors, caused by

a particular action or event. It is difficult to outline a universally applicable framework for the phrases "risk" and "safety", because they are largely determined for individual cases.

Since there is no life without risk, which kind and degree of risk is accepted or tolerated as inherent to normal life, and therefore not be considered as risk, differs from person to person and situation to situation. Each person will, intentially or intuitively, create his own framework of safety that embodies the acceptance of a particular chance and extent of disability or harm.

Thus inter-individual evaluation of a particular hazard may vary in nature and intensity, according to whether it is an activity performed by oneself or another person. Smokers accept the risk of dying early from lung cancer, whilst passive smokers do not accept harmful effects from inhaling cigarette smoke produced by other persons.

A sensible definition of health with respect to medical checks on individuals might be: "the absence of objectively and/or subjectively perceivable signs that are considered to be harmful".

However, when using this operational definition, another problem is introduced: the question to what extent a particular finding must be considered harmful or not.

Also, from a medical point of view, not every response of the body due to a particular external physical and chemical agent should by definition be considered as harmful or impairing to health. The human body responds in one way or another to all types of endured burden. The nature and intensity of the burden is largely decisive for the nature and intensity of the response.

Increased physical burden on performing strenuous physical work, for instance, induces increased oxygen demand of the muscular tissue. This is compensated for by faster and more intense transport of oxygen from the lungs to the tissues. This compensation mechanism includes the intensivation of respiration and increased heartbeat. As soon as the burden becomes too high, these mechanisms become insufficient, and functional decompensation follows. The incidence of decompensation must be regarded as medically unacceptable and harmful, unless it is readily reversible when exposure ceases, and reconvalescence can be expected within a short time.

The working capacity of the human body mainly depends on the intrinsic, inborn properties but also on other, external stresses.

This principle of individual capacity and load also holds for chemical exposures. Contact with and exposure to chemicals may cause local and/or systemic effects.

Local effects can be caused by physical contact of the skin, eyes, or mucous membranes with certain chemicals. Contact with acids and alkalis, in particular, may cause such local effects, but also other compounds, such as aldehydes (e.g., formaldehyde) and ketones may cause local harm.

Systemic effects can result from uptake of a compound via the lungs, the digestive tract, or through the intact skin and mucous membranes.

The nature and extent of uptake depend mainly on compound-related properties such as volatility and solubility and the exposure circumstances. Some examples that illustrate this are:

1) During spraying of the volatile soil and space fumigant bromomethane large amounts can be inhaled if proper respiratory protection devices are not used. But even when inhalation is not possible, large amounts of liquid and vapor can be absorbed through unprotected skin.
2) The oily liquid aniline with only minor volatility, is easily absorbed through intact skin.
3) Dimethylformamide is a high-boiling bipolar solvent with hepatotoxic (livertoxic) potential. The percutaneous absorption rate is fast and high. Thus, exposure to DMF vapors far below the threshold limit values do not guarantee the absence of harmful effects when handling this solvent.

4.5.2. Difficulties in Interpretation

The systemic response, which may or may not be clinically apparent, can also be regarded as a physiological response as long as the extent of uptake of the chemical does not exceed the compensatory capacity of the human body. The most important compensation mechanisms for chemical agents are detoxification and elimination.

A sharp distinction between biological response and harmful, pathological effects cannot

always be made. Therefore, interpretation differences between experts concerning evaluation of data are possible.

For some phenomena, such as chromosome aberration, the biological composition of semen, and some electrophysiological responses of muscle tissue and peripheral nerves, insufficient knowledge is available about normal inter- and intra-individual biological variations. An increase in the number of chromosome aberrations and a decrease of the number of healthy spermatozoa in the semen in response to the uptake of a particular chemical must not à priori be considered harmful to health.

Increased production of an enzyme that interacts with a particular chemical may follow prolonged exposure to that chemical. This enzyme induction can be regarded as a physiological response to a particular chemical burden. In some cases enzyme induction leads to better and faster detoxification and is an efficient defense mechanism against a chemical burden. However, there is still difference of opinion between experts whether enzyme induction should be regarded as an unacceptable effect or a normal response of the human organism to exogenic chemical stresses.

To arrive at a consensus, it is imperative to attain uniformity in the judgement of health risks. When this cannot be realized, a compromise should be accepted by all parties than can be expressed in a operational limitation between physiology and pathology.

Another difference of opinion between experts in the assessment of health risks might arise about the significance of impairment of health. Also, consensus should be sought about the nature and intensity of impairment that will be considered harmful.

4.5.3. Dose/Effect/Polymorphism

The intrinsic harmful properties of a chemical are determined by its inherent properties. Most effects on the human organism due to exposure to chemical compounds satisfy the principle of dose-dependency of effect. This holds that for almost every response to the uptake of a chemical, a no effect level and/or a no adverse effect level exists. It implies that in principle every chemical, under defined conditions, in a defined quantity and route of administration, and in some or all exposed individuals may cause harmful effects. This principle was discovered and published by the alchemist PARACELSUS as early as the 1500s.

An exception to this rule concerns genotoxic chemicals. For a proper preventive policy, when no contradicting data are available, it must be assumed that all concentrations of these particular chemicals, even in molecular quantities, may initiate a carcinogenic or mutagenic process after entering the cell.

However, in reality this appears to be too stringent. Firstly, it may only be valid for chemicals that have not been detoxified before reaching the organ cells. And secondly, living species have a great capacity for repairing DNA damage in cells.

The toxicological risk for an individual of contact with a chemical depends on (1) the intrinsic properties of the chemical, (2) the impairing potential (toxicity), (3) the exposure conditions, and (4) a number of person-related factors. In particular factors such as age, gender, and life-style (smoking, alcohol consumption, eating habits) may influence both the nature and intensity of an effect. This also holds for genetically determined differences in the availability of particular enzymes or in tissue susceptibility.

It is well known that a fairly large part of the global population has insufficient amounts of the enzyme *N*-acetyltransferase to guarantee optimal detoxification of aromatic amines via noncarcinogenic metabolic pathways.

Another example is the congenital deficiency of glucose-6-phosphate dehydrogenase (G 6 PD) in some ethnic groups. This enzyme is necessary for the metabolic reduction of organic peroxides. G 6 PD deficiency can lead to serious damage to the membranes of red blood cells, followed by extensive internal hemorrhage, after uptake of peroxide-forming chemicals, because these peroxides are reduced too slowly.

Although a number of inter-individual differences in biochemical capacity can be well explained, the cause of a number of these differences is still unclear. Therefore, in every estimate of health risks, the possibility should be reckoned with that a small number of a population will respond to exposure to a particular chemical, quantitatively or qualitatively, in a manner other than the majority.

Especially in setting exposure limits for chemicals, the existence of particular risk factors or risk groups, should always be considered.

The estimate of health risks of a particular activity or exposure to a particular harmful agent during work passes through the following phases [68]:

1) Hazard identification
2) Risk estimation
3) Risk evaluation
4) Risk limitation

4.5.4. Hazard Identification

The initial phase of a health risk assessment includes inventarization and description of all components of an activity to establish qualitatively whether there is a health risk due to conditions and influences at work.

A detailed study should be made of the nature of work and of the workplace. The inventory should include a full list of chemicals to be used, their physical and chemical characteristics (e.g., volatility, solubility, flammability, explosivity) and toxicity.

Toxicity. Information on toxicity can be derived from

1) Human experience with the compound
2) Results from animal experiments

Data on the state of health of humans who have been exposed to a particular compound in the past must be carefully collected and evaluated to obtain relevant information. Human experimentation is restricted, stringently regulated, and rarely acceptable. So, almost all human effects due to chemical exposure are only observational and descriptive.

The aim of this kind of investigation is to obtain knowledge about dose – effect relationships. For this reason, knowledge as accurate as possible concerning the extent of past exposures is essential. Unfortunately, this condition causes major problems because in the past measurements have not been performed at all, not on a regular basis, or by means of an unreliable or insensitive method (see Section 4.3).

As reliable data on previous exposures are often not available for estimating the toxicity, the assessment must be based on the results of animal experiments.

Animal experiments have the advantage that the exposure to a single compound can be controlled with respect to nature, dose, and duration. Thus, confounding influences, such as simultaneous exposures to other chemicals and life-style factors, can be avoided.

A major drawback, however, may be differences in susceptibility between human and animal species. Therefore, the use of an animal species that best mimics human metabolism is recommended. However, in many cases the most elementary knowledge for making a reliable choice is lacking.

Due to metabolic differences between species and sometimes in different strains of the same species, fundamental differences in detoxification and elimination mechanisms may be possible.

This can be illustrated by the following examples:

1) The human carcinogen β-naphthylamine does not give a carcinogenic response in assays with rats and rabbits
2) Mice respond with malignant liver and lung tumors in long-term inhalation assays with dichloromethane, whereas in rats only males show benign fibro-adenoma in mammary glands
3) B 6 C 3 F 1 and C 3 H mice strains are highly susceptible to spontaneous liver carcinoma. In males hepatic tumor incidence is ca. 30 % and in females ca. 6 %. With the same treatment in mice of the C 57 BL/6 J strain, induction of liver tumors does not occur [72].

Until now no generally accepted method for bridging inter- and intra-species differences is available. Only when a good qualitative and quantitative understanding of the toxicokinetics of a compound exists with respect to the metabolic pathway can a valid translation from animal to human be made. In case of doubt, uncertainty in the extrapolation can be compensated by using safety factors or uncertainty factors. The magnitude of these factors can vary from zero to several thousand, depending on the intended extent of certainty that no harmful effects will occur. The choice of magnitude of a safety factor is arbitrary, but is always related to the nature, quantity, and quality of the available

data. The greater the knowledge gap, the greater is the safety factor.

4.5.5. Risk Estimation

The qualitative inventory of risk factors is followed by a quantification of these factors. For this phase a clear understanding of the dose–effect relationship is indispensible. As sufficient relevant and valid data are often lacking, the risk estimation is often an educated guess or "guestimate".

A second activity during this phase is the assessment of the nature of the exposure. Is it a single, short-term or a long-lasting, chronic exposure? What is the magnitude of the exposure?

Also the nature of interaction between simultaneously acting chemicals should be determined, as well as the influence of other environmental conditions and individual- or grouprelated factors. For this reason, the nature of the environmental air at the working place and the magnitude of the concentrations should be estimated as accurately as possible. For details on analysis and monitoring strategies, see Section 3.3.

4.5.6. Risk Evaluation

The third phase of a risk assessment procedure deals with the compilation of the data collected during the two previous phases. This compilation may lead to a conclusion about the potential health damage of a defined population due to a defined activity and about the chance that the damage will actually happen. It is a matter of political considerations whether this estimated risk is acceptable or not. In a number of situations it can be estimated whether the risk is in compliance with actual or optional limit values.

4.5.7. Risk Limitation

Based on knowledge of the likelihood, the nature, and seriousness of potential health risks a program can be designed to prevent harmful health effects completely or to reduce the chance of health risks as far as possible.

The nature and the extent of preventive measures depend on the technological possibilities on the one hand, and on political, social, and economical considerations on the other. Tightening the limits of risk acceptance always results in more expensive preventative programs. Therefore, a balance between technical and economical feasability is indicated. Which precautionary measures can be taken is discussed in Section 4.1.

4.6. Occupational Diseases and Work-Related Illnesses [73–76]

4.6.1. Introduction; Definitions

A variety of factors at the workplace can effect human health and cause health impairment, illness, and disease.

More than 4000 years ago the ancient Egyptians described certain illnesses linked to specific jobs or professions (see Chap. 1). Since then knowledge about the negative effects of factors from the working environment on human health has increased tremendously. The continuous advance in medical science and the ongoing improvement in diagnostic techniques has fostered the identification of occupational diseases and work-related illnesses.

An occupational disease is "a disease caused by any physical factor, chemical substance, or biological agent encountered in the course of work under a contract of employment" [74]. It must be established by scientific means that the disease can be caused by the factor under consideration and the worker must have been exposed seriously to this factor during his work. The exposure generally must persist over a long period of time, and the adverse health condition develop as a chronic disease.

The term work-related illnesses covers all situations of ill health that are triggered or aggravated by influences at work. They may be acute and curable, or chronic. In practice, the term is used to characterize illnesses that are not covered by the term occupational disease.

4.6.2. Occupational Diseases

Occupational diseases are classified in various subgroups.

Each disease caused by a specific agent exhibits characteristic symptoms which are the basis for its identification.

The largest group of causative agents are chemical substances. Chemicals that are known to be capable of causing occupational disease include the following elements and their compounds

Arsenic
Beryllium
Cadmium
Chromium VI
Cobalt
Fluorine
Lead
Manganese
Mercury
Nickel
Phosphorus
Vanadium
Zinc

and the following compounds

Aromatic amines
Benzene and homologes
Carbon disulfide
Carbon monoxide
Halogenated aliphatic hydrocarbons
Halogenated aromatic hydrocarbons
Hydrogen sulfide
Nitro derivatives of aromatic hydrocarbons
Phenols, their homologues, and halogenated derivatives

Occupational lung diseases caused by dust or fiber exposure are regarded as a special group. Occupational diseases caused by dusts and fibers are:

Asbestos	asbestosis
Quartz	silicosis
Certain metal dusts	pneumoconiosis

Certain physical specifications are required (specific fiber size for asbestos, crystalline dust for quartz).

Here the disease is caused either by chemical effects, mechanical impact, and in some cases allergic reactions, or by a combination of these effects.

Another specific group of chemically induced illnesses are asthmatic disorders and allergic skin effects. The most prominent causative agents for occupational allergies and their target organs are summarized in the following:

Chromium	skin
Animal hair	respiratory tract
Flour	skin, respiratory tract
Isocyanates	skin, respiratory tract
Nickel	skin
Penicillin	skin, respiratory tract

Most compounds can cause both skin and respiratory sensitization.

With allergic occupational diseases, early detection of the cause – effect relationship and consequent avoidance of any further exposure to the allergen is of major importance for the further course of the disease. The severity and irreversibility of the health impairment is directly correlated with the duration of exposure after the onset of allergy.

Since the first description of scrotal skin cancer in chimney sweeps by Pott in 1775 a number of chemicals have been identified that can cause occupational cancer.

The connection between exposure to chemical agents at the workplace and cancer development in humans has been definitely established for the following substances:

4-Aminodiphenyl	bladder cancer
Arsenic	skin cancer
Asbestos	lung cancer, mesothelioma
Benzene	leukemia
Benzidine	bladder cancer
Bischloromethyl ether	lung cancer
Calcium chromate	lung cancer
Coal tars and pitches	skin cancer
2-Naphthylamine	bladder cancer
Vinyl chloride	liver cancer (hemangiosarcoma)

Occupational diseases can also be caused by physical agents such as noise, ionizing radiation, high pressure, heat, and vibration. The related diseases are:

Ionizing radiation	bone marrow damage, skin changes
Noise	hearing deterioration
Overpressure	divers' disease
Vibration	white finger syndrome

Another group of occupational diseases can be caused by biological agents. This group includes infectious diseases and parasitic disorders acquired during work with infectious agents

or in areas where infection is enhanced by climate or other conditions. Bacteria, viruses, and parasites can cause diseases in this group.

For most occupational diseases no specific therapy is available except for those caused by biological agents, where therapy with antibiotics can often cure the disease.

Some of the diseases continue to progress even after the exposure to the causative agent has been stopped. Classical examples for this are some of the pulmonary diseases caused by dust, such as beryllosis and silicosis.

Occupational cancer normally develops after a latency period of 15 – 40 years after first exposure to the causative agent. Workers with known exposure to human carcinogens should therefore be observed even after retirement in order to detect the cancer as early as possible.

In most European countries and North America, as well as in some other countries throughout the world, occupational diseases are compiled in a national list, which is generally the basis for possible compensation [73], [74].

4.6.3. Prevention

The prevention of occupational diseases is one of the most important tasks in occupational health and safety.

All groups of professional experts must cooperate with employers and employees to avoid adverse conditions at work that could cause occupational diseases. Of special importance is the medical surveillance (see Section 4.1) and the control and improvement of hygiene at the workplace.

Technical means and substitution of known hazardous chemicals (e.g., carcinogens) as well as providance and proper use of personal protective equipment are important tools in prevention of occupational diseases.

4.6.4. Compensation for Occupational Diseases

In the states throughout Europe and North America, as well as in an increasing number of countries in other parts of the world, recognition of occupational diseases and compensation for resulting disability is regulated by national laws.

An important basis for compensation is the respective national list of occupational diseases. Within the European Union a proposal for a European list of occupational diseases has been issued, and attempts to harmonize the legal basis for occupational diseases throughout the European Union will continue [75].

The World Health Organisation (WHO) has also issued a list of occupational diseases, which can be used by states that have not yet established a national list. The existing national lists of occupational diseases show large variations in classification, scope, and number of diseases listed.

The basic condition for the recognition of a disease as occupational is

1) Proof of sufficient exposure to the causative agent
2) Scientific proof that the disease in question can be caused by that agent

The mode of compensation differs from one country to another. In some countries standardized rules govern the amount of compensation.

In Germany, for example, compensation for an occupational disease is only granted if it causes a disability of 20 % or more. Below that margin the occupational disease is recognized as such but without compensation [73].

In the United States compensation is granted under diversified rules in the different states, based upon state legislation. The U.S. federal list of occupational diseases is modified at state level. Thus a wide variety of conditions governs the question of compensation for occupational diseases [76].

In several states, including the United States, Germany and other European countries, diseases can be recognized and compensated as occupational even if they are not contained in the national list, provided that:

1) The disease is causally related to a hazard at the workplace
2) The claimant has been exposed to the causative agent by reason of his employment
3) The incidence of the disease is substantially greater in that specific industry/occupation than in the general population

By this provision new occupational diseases can be discovered and compensated, and later included in the national list.

The most important tool for discovery of new occupational diseases is epidemiology (see Section 4.3).

Compensation for occupational diseases and related disability is provided in different countries under the national legislation either by a specific insurance fund, by government funds, or directly from the employer. Depending upon the respective legislation it can be payed as a lump sum or provided as a monthly pension.

For several occupational diseases compensation is granted only if the employee is not able to continue work in his former profession. Under German legislation this is the case of irritative and allergic disorders of the pulmonary tract and of the skin as well as for the newly established occupational diseases of the cervical and lumbar spine.

Under these circumstances the statutory accident insurance institutions in Germany (Berufsgenossenschaft) provide the financial means for training in a new profession.

This possibility is also available early on even if the occupational disease is not yet compensable but is expected to progress if work under the causative conditions continues.

4.6.5. Trends

The continuing development in medical diagnostics and epidemiology is providing new information on the possible cause – effect relationship for illnesses or diseases so far not recognized as occupational disease.

Areas under research and discussion are certain orthopedic diseases like chronic diseases of the spine; repetitive motion trauma (RMT), e.g., tendovaginitis; neurological disorders in connection with a number of organic chemicals (mainly solvents); and stress-related health impairments, including myocardial infarction triggered by stress at work. Further research and epidemiological studies will help to clarify these questions.

4.6.6. Work-Related Illnesses

Whereas occupational diseases mostly develop under chronic influence of the causative agent and are of persistant nature, work-related illnesses can be caused by acute or chronic exposures and are often fully reversible with adequate therapy and cessation of exposure. The scope of illnesses covered under this term is under discussion and covers acute infections as well as acute effects by chemical agents and acute disorders of the musculo-skeletal system, or health impairments related to shift work. Aggravation of preexisting ailments by influences at work is also covered by this term.

So far there is no systematic categorization or compensation. In some instances repeated acute illnesses of the same kind can lead to chronic occupational disease and in turn to compensation. Again the prevention of adverse influences at work is the best way to avoid adverse health effects. Frequent workplace inspections and evaluation of working conditions by the occupational physician and the industrial hygienist are of major importance for prevention, as is occupational medical surveillance for early detection of work-related illnesses.

4.7. First Aid

Basic Principles. In the case of occupational accidents, the medical care given must follow the same principles as at the communal level. The individual therapy steps taken to prevent life-threatening emergencies must be adapted to the particular requirements and must be well coordinated. The prerequisite for this is a well-functioning rescue chain consisting of the following four links:

1) Initial care at the site of the accident
2) Observation during transport
3) Initial care in the clinic
4) Final care in the clinic

This section discusses the measures necessary at the accident site.

Basic Requirements in the Companies. In Germany, the accident insurance institutions (trade associations) demand certain precautionary measures from the employers. Producing companies must train 10 % of their workers in first aid. In the case of shift work, one person trained in first aid should be present on each shift. In clerical work, 5 % of employees should

be trained in first aid. This training is conducted by organizations authorized by the trade associations. The basic course includes heart – lung resuscitation. A refresher course must be taken after two years at the latest.

Companies must have alarm systems to call the rescue units immediately and lead them to the site.

In the case of contact with chemical products, the victims must be showered immediately. Companies must be equipped with the necessary emergency showers and eye-washing devices. First aid for accident victims includes the application of an emergency dressing. Depending on the number of employees, first aid boxes of various sizes must be kept ready in suitable containers that are quickly reached and readily accessible.

Rescue of Accident Victims. When freeing workers who are trapped or have fallen, the actual rescue measures are taken by the fire fighters. The fire brigades must have available all the technical facilities required for rescuing accident victims. Workers contaminated by chemical substances must be removed from the danger zone as quickly as possible. In case of a gas cloud, for example, they must be removed at right angles to the wind direction. The rescuers may be coworkers or firemen who have arrived at the site; they must protect themselves from contact with the substances.

The rescue equipment includes not only breathing masks and compressed air masks, but also protective clothing. Exhalation of toxic substances, such as hydrogen cyanide, carbon monoxide, or solvents, can be dangerous for the rescuer during mouth-to-nose respiration. Accident victims must be immediately freed of clothing contaminated with the product. The rescuers must wear gloves to protect themselves from skin contact. Victims should be undressed under running water, if possible. Contaminated clothes must be safely disposed of.

Maintenance of Vital Functions. The vital functions of each accident victim must be checked before he is transported. In case of respiratory arrest and/or circulatory arrest, reanimation measures must be taken immediately by the colleagues trained in first aid, with the support of medically trained firemen.

The doctor on emergency call continues the life-saving therapy, determines when the patient can be moved, and accompanies the victim to the hospital.

Product Removal. In case of skin contact, chemical substances must be removed immediately. Every second counts, especially if the victim has been burned with, e.g., hot phenol or hydrofluoric acid.

Workers must be informed about how and where they can reach the next washing facility. The primary measure is washing with copious water for several minutes, possibly with the additional use of soap. In all chemical companies which work with hazardous substances, shower facilities must be installed at certain key positions, where the accident victims can shower immediately. This helps to reduce the damage caused by all products, especially if water-soluble substances are involved. However, even sparingly water soluble products can be washed off by this purely mechanical measure.

For certain products, special neutralizing agents should be made available:

Hydrofluoric acid	calcium gluconate gel, calcium gluconate solution
Aromatic nitro and amino compounds	poly(ethylene glycol) 400
Phenols, cresols	mixture of ethanol and poly(ethylene glycol) 300 (1 : 2)
Chromium	ascorbic acid

In case of exposure to such products, a thorough aftertreatment/neutralization must be carried out in the company-owned accident department.

For neutralization, acetic acid (0.3 %) can be used for alkaline substances, and sodium acetate solution (0.3 %) for acids. Repeated testing with pH indicator paper is required to ensure that neutralization is complete.

The hair, finger nails, and ears deserve special attention. Hardened plastics must be mechanically removed.

The eyes are severely endangered in chemical accidents. The eyes must be rinsed directly at the site of the accident, and rinsing continued during transport.

Water should generally be used as the rinsing agent because of the wide range of products in most companies. In the company accident de-

partment, this preliminary treatment of eye burns is continued after the application of anesthetizing eye drops. A special neutralizing solution can be used.

Treatment. In treating victims of chemical accidents, symptomatic measures must generally be taken. Special antidotes exist only for very few chemicals. The following antidotes are permitted in Germany:

Chemical substance	*Antidote*
Acrylonitrile	*N*-acetylcysteine
Aromatic nitro and amino compounds	vitamin C and toluidine blue
Hydrogencyanide and cyanides	sodium thiosulfate and 4-dimethyl-aminophenol
Carbamates	atropine
Chromic acid, dichromates	vitamin C
Hydrofluoric acid	calcium gluconate, magnesium sulfate
Methanol	ethanol and folic acid
Phosphoric acid esters	atropine and toxogonin
Irritant gases	glucocorticoids
Hydrogen sulfide	4-dimethylaminophenol

Toxic pulmonary edema caused by inhalation of irritant gases such as phosgene must be treated directly in the inital stage. In addition to the administration of glucocorticoids, positive pressure ventilation improves the success of the treatment. The clinical latent phase, i.e., the interval between the accident and the appearance of clinical symptoms, is of varying length. In case of individual pulmonary irritants, it can last up to 24 h.

Toxic pulmonary edema can be radiologically detected in the early phase, between 6 and 8 h after the inhalative trauma, before clinical symptoms will appear. The duration of the latent phase mainly depends on the dose inhaled. The higher the dose, the shorter is the latent phase.

Prevention. First aid measures must be well organized in each company. Furthermore, the rescue chain must function smoothly. It must be constantly checked whether the necessary precautions have been taken in the company. Regular practice with the simulation of possible accidents improves the effectivity. The antidotes for certain types of poisoning should be kept ready in special therapy containers to minimize the time between the accident and the start of treatment. Regular schooling of employees, efficient training of medical staff, and technically perfect equipment help to minimize damage.

4.8. Occupational Medicine in the United States

Occupational medical programs in the United States function similarly to those in Europe with the exception that U.S. programs extend beyond the workplace conditions. Overall general health and wellness programs are included in addition to health examinations and biological monitoring.

Biological exposure indices provide a tool for assessing workers exposure to chemicals, in addition to workplace air monitoring. Biological monitoring can be required by regulatory standards or recommended guidelines in the evaluation of potential health hazard. Biological exposure indices (BEIs) provided by the American Conference of Government Industrial Hygienists are reference values usually measured in exhaled air, urine, and blood.

Biological monitoring is subject to normal biological variability and requires multiple samples over a period of time to reduce the effects of variable factors. Interpretation of biological monitoring should be with professional judgement and compared to workplace air monitoring and nonwork activities. Legal considerations are necessary when designing such monitoring programs in the United States. Health risk assessment is increasingly important and the basis for recent developments of major regulations and legislative proposals in the United States.

The American Industrial Health Council (AIHC) has made fundamental recommendations for preserving the scientific credibility of risk assessments to improve the role of science in decision-making for protection of workers' health and public health.

AIHC's proposals include having an active process to update current existing guidelines with full weight-of-the-evidence to incorporate new data and methodology. Also, the legislation should promote input from all interested parties including the general public, and should lead to meaningful communications to managers, workers, and the public.

Government-wide risk assessment practices should be considered under the Office of Sci-

ence and Technology Policy to avoid different regulatory agencies having different interpretations from a single data base. The International Program on Chemical Safety (IPCS) is also discussing issues related to risk assessment harmonization. IPCS is coordinating the attainment of a common frame for risk assessment and is working to improve risk assessment methodology in the areas of reproductive/development toxicity, genotoxicity, and carcinogenicity (see Section 4.5).

Epidemiological studies have a special relationship to risk assessment. These studies are normally developed to investigate associations between health effects and a contributing factor. Comparisons between exposed and nonexposed groups can establish associations between exposure and health effects (see Section 4.3).

However, many epidemiological studies do not have the adequately quantified exposures which are necessary for risk assessment methods.

Retroactive studies have limitations because of confounding exposures and numerous health effects. Animal studies can be evaluated for similarity in response, but special attention must be given to species susceptibility, response of target organ mechanisms, and dose response.

Well-designed epidemiological studies are always preferable to animal studies for the risk assessor, due to the advantage of observing dose – response results in humans rather than extrapolating animal results to humans.

In general, for prevention of occupational diseases (risk management) the recommended concentrations for threshold chemical effects use safety factors, while zero threshold chemicals require establishing acceptable concentrations with risk assessment methods.

5. References

General References

1. A. H. Murke, H. Rodegra: "Geschichte der Arbeitsmedizin," in J. Konietzko, H. Dupuis (eds.): *Handbuch der Arbeitsmedizin,* (I-1), ecomed Verlag, Landsberg 1988.
2. W. N. Rom: *Environmental and Occupational Medicine,* Little, Brown and Company, Boston 1983.
3. D. J. Hansen: *The Work Environment,* vol. 1, Lewis Publisher, Chelsea, Mich. 1992.

General References

4. Bundesminister für Arbeit und Sozialordnung (ed.): *Übersicht über das Recht der Arbeit,* Bonn 1994.
5. Arbeitskammer des Saarlandes (eds.): *Das Arbeitssicherheitsgesetz,* Saarbrücken 1992.
6. Arbeitskammer des Saarlandes (eds.): *Arbeitsschutzvorschriften, Unfallverhütungsvorschriften, Vorsorgeuntersuchungen,* Saarbrücken 1993.
7. Bundesanstalt für Arbeitsschutz (eds.): *Arbeit menschlicher machen,* Dortmund 1992.
8. BASF: *Mehr Erfolg durch Sicherheit, sichere Betriebsführung – Leitfaden für Führungskräfte,* Ludwigshafen 1994.

General References

9. *The Industrial Environment. Its Evaluation and Control,* NIOSH Publication, U.S. Govt Printing Office, Washington 1973.
10. R. S. Brief: *Basic Industrial Hygiene – A Training Manual,* AIHA Publication, Fairfax, VA, 1975.
11. R. W. Allen et al.: *Industrial Hygiene,* Prentice Hall, Englewood Cliffs, N.J., 1976.
12. L. Parmeggiani: *Encycl. Occup. Health Saf.,* 3rd ed., ILO, Geneva 1983.
13. G. D. Clayton, F. E. Clayton (eds.): "General Principles," *Patty's Industrial Hygiene and Toxicology,* **vol. I,** Wiley-Interscience, New York 1978.
14. J. T. Garrett, L. V. Cralley, L. J. Cralley: *Industrial Hygiene Management,* Wiley & Sons, Chichester 1988.
15. Leidel, Busch, Lynch: *Occupational Exposure Sampling Strategy Manual,* NIOSH Publication, U.S. Govt Printing Office, Washington 1977.
16. R. R. Langner et al.: "Two Methods Establishing Industrial Hygiene Priorities," *Am. Ind. Hyg. Assoc. J.* **40** (1979) 1039 – 1045.
17. N. Hawkins, S. Norwood, J. Rock: *A Strategy for Occupational Exposure Assessment,* AIHA, Fairfax, VA, 1991.
18. BOSH Technical Guide, N. 11: Sampling Strategies for Airborne Contaminants in the Workplace, 1993.
19. CEFIC: *Report on Occupational Exposure Limits and Monitoring Strategy,* Brussels.
20. CEFIC: *Symposium on Occupational Exposure Limits and Harmonisation in the Setting and Control of OELs for the Protection of Workers,* Brussels 1984.

21. L. V. Cralley, L. J. Cralley: *Industrial Hygiene Aspects of Plant Operations,* MacMillan, New York 1987.
22. ACGIH: *Documentation of the Treshold Limit Values,* 5th ed., Cincinnati, 1986.
23. ACGIH: *Threshold Limit Values 1993 – 1994,* Cincinnati.
24. *Occupational Exposure Limits Worldwide,* AIHA Publication, Akron, OH, 1987.
25. *NIOSH Manual of Analytical Methods,* 3rd ed., Cincinnati 1984.
26. E. H. Berger et al.: *Noise and Hearing Conservation Manual,* 4th ed., AIHA, Fairfax, VA, 1986.
27. ACGIH: *An Industrial Ventilation – A Manual of Recommended Practices,* 20th ed., Cincinnati 1988.
28. K. Hallenbeck, K. M. Cunnigham: *Quantitative Risk Assessment for Environmental and Occupational Health,* Lewis Publishers, London 1987.
29. P. H. L. Williams, J. L. Burson: *Industrial Toxicology,* Van Nostrand Reinhold, New York 1985.
30. WHO: Early Detection of Health Impairment in Occupational Exposure to Health Hazard, *Tech. Rep. Ser.* **571** (1975).
31. BG Chemie: *Berufsgenossenschaftliche Grundsätze für Arbeitsmedizinische Vorsorge-Untersuchungen,* Gentner Verlag, Stuttgart 1971.
32. R. L. Zielhuis, P. Th. Henderson: "Definitions of Monitoring Activities and their Relevance for the Practice of Occupational Health," *Int. Arch. Occup. Environ. Health* **57** (1986) 249 – 257.
33. A. Berlin, R. E. Yodaiken, B. A. Henman (eds.): *Assessment of Toxic Agents at the Workplace. Roles of Ambient and Biological Monitoring,* Nÿhoff, Boston 1984.
34. ACGIH: *Threshold Limit Values and Biological Exposure Indices,* Cincinnati 1991/1992.
35. DFG: *MAK- und BAT-Werte Liste 1992,* VCH Verlagsgesellschaft, Weinheim 1992.
36. R. Roi et al.: *Occupational Health Guidelines for Chemical Risk,* CEC, Luxemburg 1983.
37. L. Allessio et al.: *Human Biological Monitoring of Industrial Chemicals Series,* CEC, Luxemburg 1983.
38. DFG: "Analysen in biologischem Material," *Analytische Methoden zur Prüfung gesundheitsschädlicher Arbeitsstoffe,* **vol. 2,** loose-leaf book, VCH Verlagsgesellschaft, Weinheim.
39. J. M. Last (ed.): *A Dictionary of Epidemiology,* Oxford University Press, Oxford 1988.
40. M. Alderson: *An Introduction to Epidemiology,* Mac Millan Press, London 1976.
41. R. R. Monson: *Occupational Epidemiology,* CRC Press, Boston 1990.
42. M. Alderson: *Occupational Cancer,* Butterworths, London 1986.
43. R. Doll, R. Peto: *The Causes of Cancer,* Oxford University Press, Oxford 1981.
44. W. Hacker: *Arbeitspsychologie. Psychische Regulation von Arbeitstätigkeiten,* Huber Verlag, Bern 1986.
45. R. Tielsch: "Ansatz, Kriterien und Ergebnisse zur Ermittlung subjektiver Belastung und Beanspruchung an industriellen Arbeitsplätzen," *Fortschr. Ber. VDI Reihe 17* (1987).
46. Gesellschaft für Arbeitswissenschaft: Denkschrift "Arbeitswissenschaft in der Gesetzgebung," Rationalisierungskuratorium der deutschen Wirtschaft, 2nd ed., Europäische Verlagsanstalt, Frankfurt/Main 1974.
47. H. Schmidtke: *Handbuch der Ergonomie,* 2nd ed., Hanser Verlag, München 1989.
48. *Ergonomics,* an international journal of research and practice in human factors and ergonomics, Taylor & Francis, London – Washington.
49. *Appl. Ergonomics,* journal on the technology of mans' relations with machines, environment and work systems, Butterworth Heinemann, Oxford.
50. W. Laurig: "Einführung in die Ergonomie," in J. Konietzko, H. Dupuis (eds.): *Handbuch der Arbeitsmedizin,* 6th ed., ecomed Verlagsgesellschaft, Landsberg 1991.
51. W. Rohmert: "Aufgaben und Inhalt der Arbeitswissenschaft," *Die berufsbildende Schule* **24** (1972) 3 – 14.
52. K. H. Norpoth: *Einführung in die Arbeitsmedizin,* ecomed Verlagsgesellschaft, Landsberg 1991.
53. W. Rohmert: "Das Belastungs-Beanspruchungs-Konzept," *Zentralbl. Arbeitswissensch.* **38** (1984) 193 – 199.
54. REFA: *Methodenlehre der Betriebsorganisation. Grundlagen der Arbeitsgestaltung,* Hanser Verlag, München 1991.
55. E. Grandjean: *Physiologische Arbeitsgestaltung,* 4th ed., ecomed Verlagsgesellschaft, Landsberg 1991.

56. P. Jenik: "Maschinen menschlich konstruiert: Somatographie als arbeitswissenschaftliche Methode für Konstrukteure," *Maschinenmarkt/MM-Industriejournal* **78** (1972) 87 – 90.
57. W. Laurig, K. Wieland, H.-D. Mecheln: "Arbeitsplätze für Behinderte – I. Dokumentation technischer Arbeitshilfen," Forschungsbericht no. 233 der Bundesanstalt für Arbeitsschutz und Unfallforschung Dortmund, Wirtschaftsverlag NW, Bremerhaven 1980.
58. W. Laurig: *Grundzüge der Ergonomie, Erkenntnisse und Prinzipien,* 4th ed., Beuth Verlag, Berlin 1992.
59. W. Rohmert, J. Rutenfranz: *Praktische Arbeitsphysiologie,* 3rd ed., Thieme Verlag, Stuttgart 1983.
60. J. Rutenfranz, P. Knauth: *Schichtarbeit und Nachtarbeit: Probleme – Formen – Empfehlungen,* 2nd ed., Bayerisches Staatsministerium für Arbeit und Sozialordnung, München 1987.
61. K. Ruppe, K. Nienerowski: *Methoden der Arbeitshygiene. Erkennen – Messen – Bewerten – Gestalten,* Verlag Volk und Gesundheit, Berlin 1988.
62. H. Schmidtke, J. Jastrzebska-Fraczek, H. Ruhmann: *Ergonomische Prüfung von technischen Komponenten, Umweltfaktoren und Arbeitsaufgaben: Daten und Methoden,* Hanser Verlag, München 1989.
63. K. Wieland, W. Laurig: "Arbeitsplätze für Behinderte: Handbuch technischer Arbeitshilfen zur Arbeitsgestaltung," Forschungsbericht no. 375 der Bundesanstalt für Arbeitsschutz und Unfallforschung Dortmund, 4th ed., Wirtschaftsverlag NW, Bremerhaven 1987.
64. K. Landau, W. Rohmert: *Recent Developments in Job Analysis,* Taylor & Francis, London 1989.
65. W. Rohmert, K. Landau: *Das arbeitswissenschaftliche Erhebungsverfahren zur Tätigkeitsanalyse (AET),* Huber Verlag, Bern 1979.
66. W. Laurig: *Prospektive Ergonomie – Utopie oder Wirklichkeit?* Arbeitswissenschaft des AGV Metall, Köln, no. 8, Arbeitgeberverband der Metallindustrie, Köln 1984.
67. BG Chemie: *Toxikologische Bewertungen,* Springer Verlag, Berlin 1970.
68. ECETOC: *Risk Assessment of Occupational Carcinogenes,* Monograph no. 3, Brussels 1982.
69. ECETOC: Joint Assessment Commodity Chemicals Reports,
70. ECETOC: Monographs Enquiries ECETOC, Av. E. van Nieuwenhuyse 4, Box 6, B 1160 Brussels, Belgium
71. Arbeidsinspectie "Health based Recommended Occupational Exposure Limit" reports. Edited by the Dutch Expert Committee for Occupational Standards. Ed. Min. Sociale Zaken en Werkgelegenheid/Dir. Generaal van de Arbeid, P.O.Box 90 804, 2509 LV The Hague, Netherlands.
72. T. R. Fox, T. L. Goldsworthy: "Molecular Analysis of the H-ras Gene. An Understanding of the Mouse Liver Tumor Development," *CIIT-Activities* **13** (1993) 1 – 2.
73. E. Perlebach, G. Mehrtens: *Die Berufskrankheitenverordnung (BeKV),* Ergänzbare Sammlung der Vorschriften, Merkblätter und Materialien, Erich Schmidt Verlag, Berlin.
74. Institute of Occupational Health: *Act on Occupational Diseases and Ordinance on Occupational Diseases,* Helsinki 1989.
75. Empfehlungen der Kommission betreffend die Annahme einer Europäischen Liste der Berufskrankheiten. Empfehlung der Kommission vom 22.05.90 – (90/326/EWG), Amtsblatt der Europäischen Gemeinschaften Nr. L 160/39
76. U.S. Department of Labor: *Recordkeeping Guidelines for Occupational Injuries and Illnesses,* Washington 1986.

Transport, Handling, and Storage

Clemens Stroetmann, Bad Honnef, Federal Republic of Germany (Chap. 1)

Emile Berson, Ministère des Transports, Paris, France (Chap. 2)

Nadia Berson, Paris, France (Chap. 2)

Detlef Rennoch, Bundesanstalt für Materialforschung und -prüfung, Berlin, Federal Republic of Germany (Chap. 3)

Bernhard Droste, Bundesanstalt für Materialforschung und -prüfung, Berlin, Federal Republic of Germany (Chap. 4)

Wieger Johannes Visser, Nederlandse Spoorwegen, Utrecht, The Netherlands (Sections 5.1 – 5.2.1)

Frits Wybenga, International Standards Coordinator, Office of Hazardous Materials Safety, Washington D.C. 20590, United States (Section 5.2.2)

Yoshio Yasogawa, Nippon Kaiji Kentei Kyokai, Tokyo, Japan (Section 5.2.3)

Heinz W. Hübner, Bundesanstalt für Materialforschung und -prüfung, Berlin, Federal Republic of Germany (Section 6.1)

Kurt-Joseph Doktor, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 6.2.1.1, 6.2.1.2, 6.3, 6.4)

Klaus Jochem Döhrn, Germanischer Lloyd, Hamburg, Federal Republic of Germany (Section 6.2.1.3)

Josef Hawighorst, Bayer AG, Leverkusen, Federal Republic of Germany (Sections 6.2.1.1, 6.2.1.2)

Annette Rolle, Bundesanstalt für Materialforschung und -prüfung, Berlin, Federal Republic of Germany (Section 6.2.2)

Hans Klusacek, Bayer AG, Dormagen, Federal Republic of Germany (Section 6.3)

Hans Aschenbrenner, Bayer AG, Leverkusen, Federal Republic of Germany (Section 6.4.2)

Hans-Peter Lühr, Technische Universität, Institut für wassergefährdende Stoffe, Berlin, Federal Republic of Germany (Chap. 7)

Heinrich Freiherr von Lersner, Umweltbundesamt, Berlin, Federal Republic of Germany (Chap. 8)

1. Introduction 3

2. Safety 5

2.1. Risk 5

2.2. Transport Safety 6

2.3. Storage Safety 7

2.4. Security 8

3. Dangerous Goods and Substances 9

3.1. Transport Classification of Dangerous Goods 11

3.1.1. Explosive Substances and Articles (Class 1) 12

3.1.2. Gases (Class 2) 13

3.1.3. Flammable Liquids (Class 3) 14

3.1.4. Flammable Solids, Substances Liable to Spontaneous Combustion, and Substances which in Contact with Water Emit Flammable Gases (Class 4) 15

3.1.4.1. Flammable Solids [Division 4.1 (a)] 15

3.1.4.2. Self-Reactive and Related Substances [Division 4.1 (b)] 15

3.1.4.3. Desensitized Explosives [Division 4.1 (c)] 16

3.1.4.4. Substances Liable to Spontaneous Combustion [Division 4.2] 16

3.1.4.5. Substances which in Contact with Water Emit Flammable Gases (Division 4.3) 16

© 2005 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
10.1002/14356007.b08_497

3.1.5. Oxidizing Substances and Organic Peroxides (Class 5) 17
3.1.6. Toxic and Infectious Substances (Class 6) 17
3.1.6.1. Toxic Substances (Division 6.1) . . 17
3.1.6.2. Infectious Substances (Division 6.2) 18
3.1.7. Radioactive Material (Class 7) . . . 19
3.1.8. Corrosive Substances (Class 8) . . . 19
3.1.9. Miscellaneous Dangerous Substances and Articles (Class 9) 19
3.1.10. Estimation of Major Hazards when there are Several Hazardous Properties 20
3.1.11. Seawater Pollutants and Inland Water Pollutants 20
3.2. Storage of Dangerous Goods . . . 21
3.2.1. Explosive Substances (Class 1) . . . 21
3.2.2. Flammable Liquids (Class 3) 22
3.2.3. Oxidizing Substances (Class 5.1) . 22
3.2.4. Toxic Substances (Class 6.1) 22
3.2.5. Radioactive Material (Class 7) . . . 23
4. Safety Measures 23
4.1. Technical Safety Measures 24
4.1.1. Design and Equipment of Tanks and Packages 25
4.1.2. Fire Protection Measures 26
4.2. Quality Assurance Measures . . . 28
4.3. Safety Distances 28
4.4. Emergency Response 30
5. Regulations and Standards 30
5.1. World Transport Regulations . . . 30
5.1.1. UN Recommendations on the Transport of Dangerous Goods 30
5.1.2. The IMDG-Code 31
5.1.3. The ICAO Technical Instructions . 31
5.2. Regional Transport Regulations . 32
5.2.1. Europe 32
5.2.2. United States 33
5.2.3. Japan 34
6. Containment Technology 35
6.1. Transport of Piece Goods 35
6.1.1. UN Packagings, "Packagings by Description", Intermediate Bulk Containers 35
6.1.1.1. UN Packagings 35
6.1.1.2. Packagings by Description 37
6.1.1.3. Intermediate Bulk Containers (IBC) 37
6.1.2. Packagings for Radioactive Material 39
6.1.3. Consignments, Means of Loading, and Loading Security 41
6.2. Bulk Transport 41
6.2.1. Bulk Transport of Liquids and Gases 42
6.2.1.1. Tank Containers 42
6.2.1.2. Tank Trucks and Railroad Tank Cars 43
6.2.1.3. Marine Tankers 44
6.2.2. Solid Bulk Cargo 48
6.2.2.1. Solid Bulk in Road and Rail Transport 48
6.2.2.2. Solid Bulk in Marine Transport . . . 50
6.3. Cargo Storage 50
6.4. Bulk Storage 52
6.4.1. Storage Concept 52
6.4.2. Tanks 55
7. Pipelines 56
7.1. Types of Pipeline 56
7.2. Requirements for Pipeline Systems 56
7.3. Design and Operation 57
7.3.1. Routing 57
7.3.2. Pipeline Design 57
7.3.3. Materials 58
7.3.4. Pipe Laying 58
7.3.5. Operation 59
7.4. Incidents 59
7.5. Economic Aspects 60
8. Future Developments 60
8.1. International Harmonization of Dangerous Goods Regulations . . 61
8.2. Transport Avoidance and Regulatory Measures 62
9. References 62

The entire topic was coordinated by Bernd Schulz-Forberg

1. Introduction

At present, there is a lack of comprehensive data on the transport of dangerous goods, in terms of transport volume, transport capacity, carriers, dangerous goods (including dangerous wastes), accident frequency and causes, nature and extent of releases of material because of accidents, leakages, and recovery, etc. The same is true for the area of storage and transport-dependent intermediate storage of dangerous materials and for comparative surveys, e.g., strategies for transport avoidance and traffic deflection as well as site contingency plans. The EC Green Paper on the Effect of Traffic on the Environment makes the following points:

1) An enormous increase is expected in total carriage, average distances, frequencies of consignments, and speed [1]
2) From 1990 to 2010, a 42 % increase of the carriage of goods by road and a 33 % increase is expected by rail [2, 3]
3) Transport of dangerous goods involves extensive risks, especially potential effects on the environment [4]
4) A package of measures for the reduction or avoidance of risk during the transport of dangerous goods has been prepared [5], which sets out the environmental, traffic policy, and economic dimensions of these trends

Comparable forecasts for individual Member States, e.g., Germany, predict a large increase both of goods traffic volume and of transport capacity over the next two decades [6], the proportion of dangerous goods in rail traffic being estimated as ca. 12.4 – 13.8 % and in road traffic as 9.1 – 9.6 % [7]. A breakdown of shipments of dangerous goods by transport sector shows that ca. 27.4 % go by inland waterways, 25 % by road transport (long-distance and short hauls across frontiers), ca. 24 % by sea, and 23.5 % by rail [8].

A forecast of the traffic volume and its distribution, which shows the percentage change compared with 1988, assumes a growth in long-distance road transport of goods of 72.9 % by the year 2010; in inland waterway transport 35.3 %, but in rail transport (less polluting than by road) a reduction of 7.3 %.

Moves toward economic integration of Europe should influence the scale of macroeconomic activity in the short, medium, and long term; this should result in a rise in the demand for transport services. With the increase of freight traffic volume, the number of shipments of dangerous goods inevitably rises, and with it the risk of accidents and release of hazardous materials. There is a consequent rise in other external costs (e.g., destruction of the landscape, pollution, and noise) which differ from each other according to the mode of carriage.

The following global goals can be derived from the forecasts outlined above:

1) Increase of transport safety, including transport avoidance strategies
2) A consistent emphasis on precautionary traffic measures.

A multitude of detailed regulations, constantly updated and amended, are aimed at minimizing the special risks of the transport of dangerous goods. However, most regulations – aside from the UN Recommendations – are specific to a particular mode of transport. In some cases, they have different legal status, different member or signatory states, different geographic areas of application, and to some extent different content. Above all, implementation of regulations in different countries may depend on national interest, the external environment, and geographic region.

However, these regulations do not address the long-term transport policy and environmental challenges of goods outlined above, and are often incomprehensible to the nonexpert and difficult to enforce. Even the UN Recommendations fail to deal with the necessity of transport avoidance and the transfer of consignments of dangerous goods to safer or environmentally beneficial means of transport. The basic issue of classification of dangerous materials and goods is still far from being resolved. The worldwide harmonization of dangerous goods regulations is to be encouraged, but it must not lead to a downgrading of safety and environmental standards, especially under the influence of commercial competition. Undifferentiated harmonization can threaten national instruments, in some cases tried and true, for improving transport safety and guaranteeing precautionary environmental protection, e.g., traffic deflection regulations and the allocation of special routes for certain dangerous goods.

It is also necessary to eliminate distortions of competition between carriers, frequently at the expense of safety and environmental considerations, and to resolve conflicts of aims, e.g., between the freedom of service-rendering and legitimate safety interests, or between interests of transport and environmental policy (assistance or avoidance strategies for transport).

The development of a sufficiently differentiated, yet nationally and internationally coordinated safety culture oriented toward prevention is therefore essential.

2. Safety

Although modern technology has overcome many natural dangers (predators, famines, diseases), it has introducted dangers that were unknown in the past. For example, it is estimated that there are about 100 000 dangerous substances, increasing by 1000 every year.

Furthermore, while relying more and more on technology, people are increasingly less prepared to tolerate the risk that this involves. The problem of safety is therefore difficult and ongoing.

2.1. Risk

Risk is an uncertain, possible event, and, as regards safety, often involves damage.

Damage occurs when a harmful agent, finding a medium favorable to its production and/or propagation, reaches a receptive site, i.e., a target.

Safety objectives are designed to prevent the formation of harmful agents, to confine them so as to prevent their dispersal, or to ensure that receptive sites are beyond their reach. Safety measures often combine these objectives. Elaboration of these measures involves analysis of the various parameters and their evolution over time. This risk identification requires a constant questioning approach, e.g., what can a harmful agent do? What can be done to prevent it?

All possibilities should be explored: normal operation, incident, and accident, up to a possible serious catastrophe; in each case measures have to be devised to prevent the situation becoming worse and to reduce damage. Such a method is effective only if it covers all the stages (e.g., in an industrial installation the design, execution, and operation) and involves all factors in these stages.

Risk identification should be supplemented by an evaluation of the probability that the risk will occur. A risk is fully defined only by the conjunction of its detrimental effects and its probability; this enables risks to be compared. Statistical methods enable such an evaluation to be made, and although these methods are still imperfect and limited, especially in new fields, they are improving all the time.

One method that applies particularly to fixed installations is that based on experimental feedback. Starting from an analysis of incidents by the method of event trees, it evaluates the probability that such events may degenerate into a serious accident, and by adding these probabilities over a period of time and referring the result to the number of installations of the same type, it can sometimes be shown that a serious accident becomes probable after a certain period. The importance of this method is that it is able to show that the safety of a system has become compromised, even though no accident has occurred and those responsible may honestly believe that the system is operating satisfactorily.

How can these principles be applied to the transport and storage of dangerous materials? First, this field concerns public safety, on account of the extent and potential consequences of accidents; in all developed countries, it is regulated. Regulations are generally national or local as regards production and storage or national or international as regards transport.

All regulations include an identification of dangerous materials and their classification according to the nature of the danger they present; classification consists of a list of products, drawn up on the basis of experience, and often supplemented by criteria and tests for new materials. These lists and criteria are not necessarily the same for storage and transport (e.g., carcinogenicity or toxicity by bioaccumulation are of little importance in connection with transport). Listing of a product means that the safety regulations applicable to the material and to the activity in question must be observed.

The regulations almost always include the following provisions:

1) Declaration and marking of materials: transport document (manifest) or storage declaration, and labeling of containers
2) Confinement: regulations covering the construction and testing of containers and tanks, safety systems, and warning systems
3) Operations: safety instructions for normal operation and in an accident
4) Environmental factors: distances of storage sites from places of habitation, regulations governing the construction and movement of transport vehicles

2.2. Transport Safety

Transport is an important stage in the life cycle of a dangerous material, which extends from its production to its use. Dangerous materials are scarcely handled at all during transport, except during loading and unloading; the materials are packed and not normally in direct contact with potential targets (humans, animals, the environment).

Accidents due solely to the material itself are therefore fairly rare; the risk comes from the medium in which the material is found, namely traffic. Those responsible for transport safety have to deal specifically with traffic accidents and the consequences arising from the presence of dangerous materials, attempting to reduce the probability of:

Traffic accidents
Escape of material in the event of an accident
Contact between the unconfined material and a target

The best way of preventing a transport accident is to avoid transport. This is why transport of some extremely dangerous materials (e.g. nondesensitized nitroglycerine) is prohibited. Manufacturing industries have to adapt their production methods to these demands.

Measures intended to reduce the probability of escape of material involve the use of the strongest possible containers (crates, boxes, tanks), not forgetting that, except in certain cases (e.g., containers for certain radioactive materials), they are unable to withstand major accidents, and that, if there is a fire, containers that are too strong can cause high pressures, resulting in explosions. Particularly for large containers, such as tanks, it is better to concentrate on their ability to deform without bursting when subjected to violent shock, rather than their overall strength. The ductility of the material and the amount of energy it absorbs on deformation are important safety factors, as is the provision of structures which distribute the deformation (rings, collars, baffles, etc.).

A compromise has to be reached in the solution of these complex problems, depending on the hazard level of the material, type of transport, and industrial technology, which is reflected in national and international transport regulations that lay down provisions and tests for containers.

In addition tanks, large bulk containers, and gas cylinders are subjected to periodic checks. Also, limits are placed on the amount of material per container in order to reduce the amount of material that can escape in the event of an accident. Some materials are not allowed to be transported in bulk. In limiting the amount of material carried per vehicle, it has to be remembered that this increases the number of vehicles transporting this material, and thus the probability of an accident. This type of provision is restricted almost exclusively to explosives.

Materials that can react with one another, causing a new additional danger, e.g., fire, explosion, release of toxic gas, must not be packed together. Similarly, it is prohibited to transport in the same vehicle materials (explosives and organic peroxides) that can cause an explosion with other dangerous materials, since dispersal of these materials by an explosion exacerbates their effects and hinders their recovery.

In order to reduce the probability of a traffic accident, the safety of vehicles is continually being improved, e.g. by improving braking (ABS system, heavy-duty brakes) or by lowering the center of gravity of tanks. Attempts are also being made to improve driver behavior by awareness and reflex training.

To reduce the probability of traffic accidents, and to prevent contact between unconfined material and a target, is more problematic, sometimes contradictory. For example, in the case of land transport there may be a choice between rail and road. It might seem preferable to chose rail since the overall probability of an accident is very low, but railway lines in general pass through built-up areas, whereas bypasses on trunk roads often avoid these.

Furthermore, railway stations, with their points crossings and intersections (i.e., in populated areas) involve the greatest probability of an accident. The amounts of dangerous materials transported in a train are much greater than those in a lorry, and much more varied, which gives rise to the possibility of much more serious accumulations and dangerous synergistic effects. The decisions taken by some countries in favor of rail transport are often politically or commercially motivated. Modes of transport can be compared only from case to case, between a specified starting point and destination.

The same applies when choosing one particular route over another. Local authorities are sometimes inclined to prohibit movement of certain dangerous materials within their built-up areas. Such a decision, which shifts the danger elsewhere, should be taken only on the basis of a comparative analysis of the risk of transport subject or not subject to such a prohibition. The decision should specify a recommended alternative route for transporting dangerous goods. The choice of such a route is complex, and should take account of the environment, avoid possible proximity of industrial activities capable of exacerbating a potential road accident, and avoid populated areas, taking account of daily journeys and travel (the commercial and business center of a large town contains many people during the day, whereas the suburbs are almost empty; the reverse is true at night).

Other types of more general prohibition or restriction exist, often decided at government level and concerning the days, or even hours, when road traffic is very heavy and/or the probability of accidents is very high (beginning and end of holidays, weekends, festivals, etc.). All these measures tend to channel the movement of dangerous goods into periods and zones where the probability and consequences of an accident are a minimum.

Another way to tackle the consequences of an accident is to improve the speed and efficiency of the emergency services. Studies are being carried out on the possibilities of satellite surveillance and warning.

2.3. Storage Safety

With regard to the production or storage of dangerous goods the drafting of the safety measures should take account of the agent, medium, and target. Identification of the agent appears simple. However, with regard to production as well as storage, when the length or conditions of storage can affect stability, all the possible forms and transformations of a material have to be examined. The major accident in a chemical factory in Seveso led to a European Community Regulation, the so- called Seveso Directive. The product, which was distributed over a large area by the accident, was an extremely toxic dioxin. No measures had been envisaged to control this compound, since it was not the end product of the process, but was formed only as an intermediate and converted immediately into a less dangerous chemical. No one had considered that an accident resulting in escape could occur during the brief existence of the dioxin. This graphically illustrates the need for extreme care in identifying the danger.

Unlike transport, the medium is geographically fixed, which simplifies matters, though its study should take into account all the possible variations: time of day; seasonal climatic conditions; wind speeds and directions, as well as all the normal, incidental, and unplanned operating conditions in the installation itself, so as to evaluate the forseeable consequences and formulate appropriate safety measures.

These measures vary, depending on the nature of the goods that are stored and the industrial activity, but always aim to reduce the risk of escape of dangerous products. They include restricting or splitting up amounts of materials, strengthening containers, verifying that they are in good condition by periodic checks, using devices to contain leaks, etc. It should also be ensured that the installation can withstand foreseeable natural catastrophes, including earthquakes, tornadoes,